Molecular Diagnosis

Rainer Spang

Courses in Practical DNA Microarray Analysis

NGFN
Nationales
Genomforschungsnetz

Questions in medical research:

Basic Research:

Which role plays gene A in disease B ?

Clinical Routine:

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

Yesterday the focus was on basic research questions

We have investigated genes

- Differentially expressed genes

- Coexpressed genes (clustering)

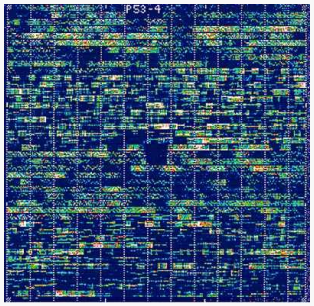
Today it will be on patients

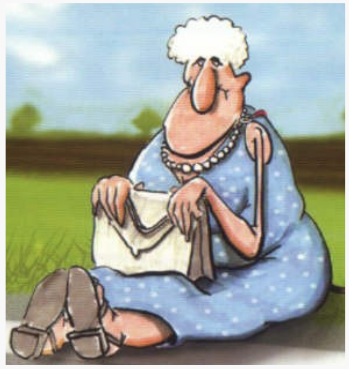
- Molecular diagnosis

- Predicting survival / therapy response

Personalized Medicine

DNA Chip of Ms. Smith





Ms. Smith

genome:~/SIBC/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

3%

Expression profile of Ms. Smith

The expression profile …

genome:~/S/BC/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

3%

… a list of 30,000 numbers

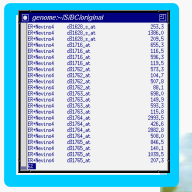
… that are all properties of Ms. Smith

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

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

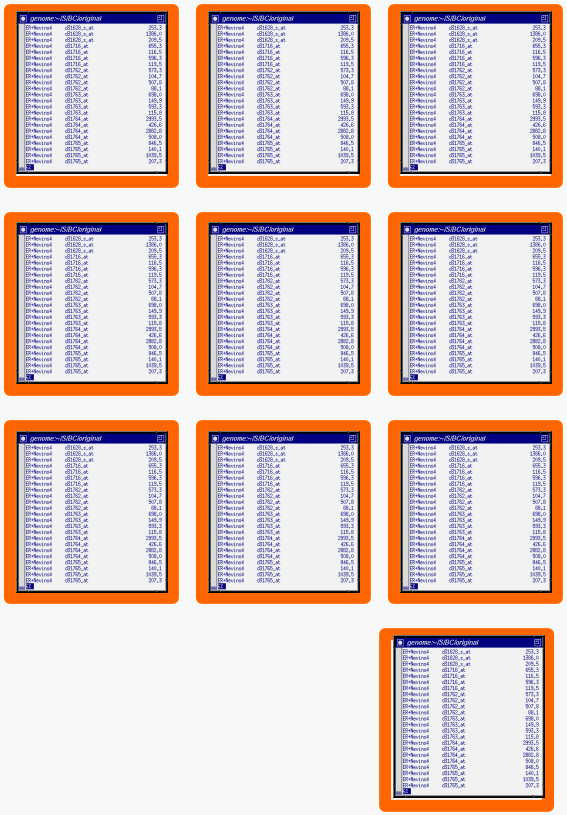
How can these numbers *tell us (predict)* whether Ms. Smith has tumor type A or tumor type B ?

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

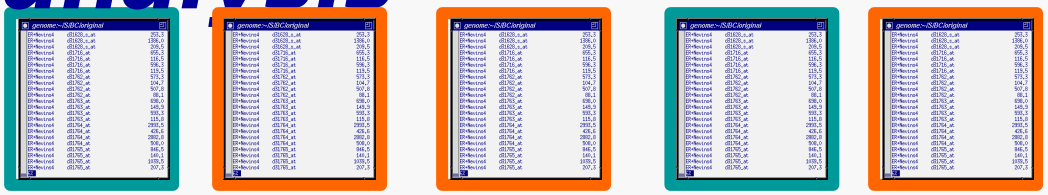


?

Ms. Smith

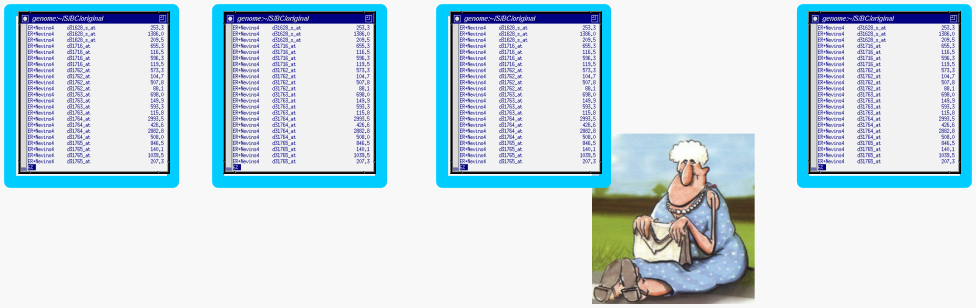


The setup for predictive data analysis



There are patients with known outcome

- the trainings samples -

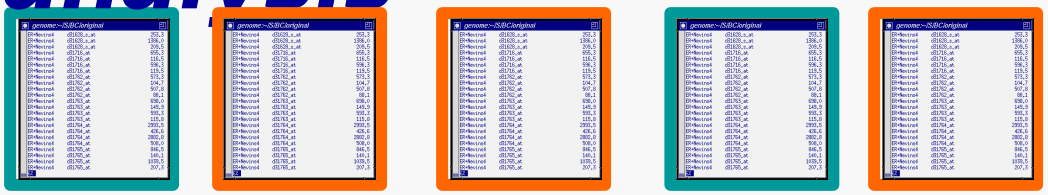


Ms. Smith

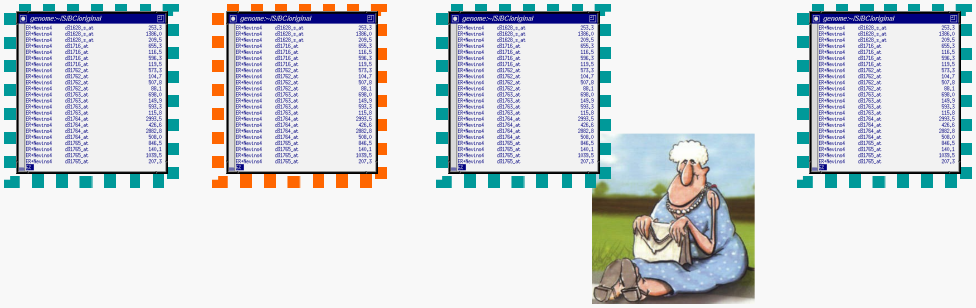
There are patients with unknown outcome

- the „new“ samples -

The challenge of predictive data analysis



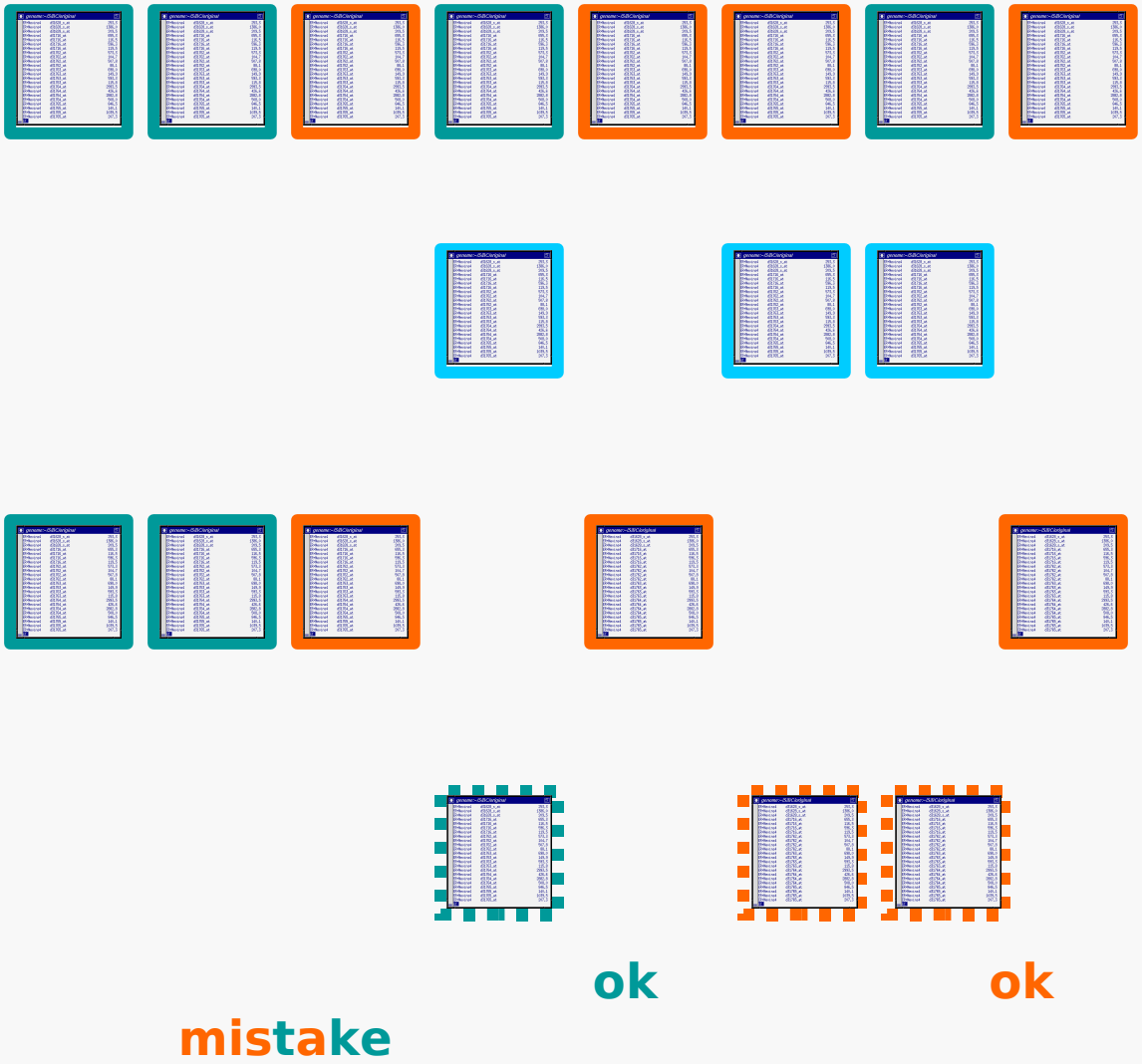
Use the trainings samples …



Ms. Smith

… 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

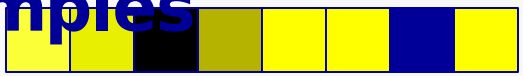
Predict the test samples and compare the predicted outcome to the true outcome

We will proceed in 4 Steps

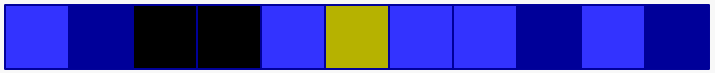
- **Prediction with 1 gene**
- **Prediction with 2 genes**
- **Prediction with a small number of genes**
- **Prediction with the microarray**

Prediction with 1 gene

Color coded expression levels of trainings samples



A



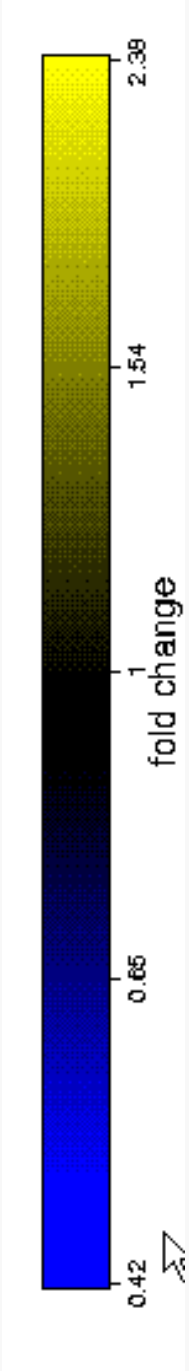
B

Ms. Smith → **type A**

Ms. Smith → **type B**

Ms. Smith → **borderline**

Which color shade is a good decision boundary?



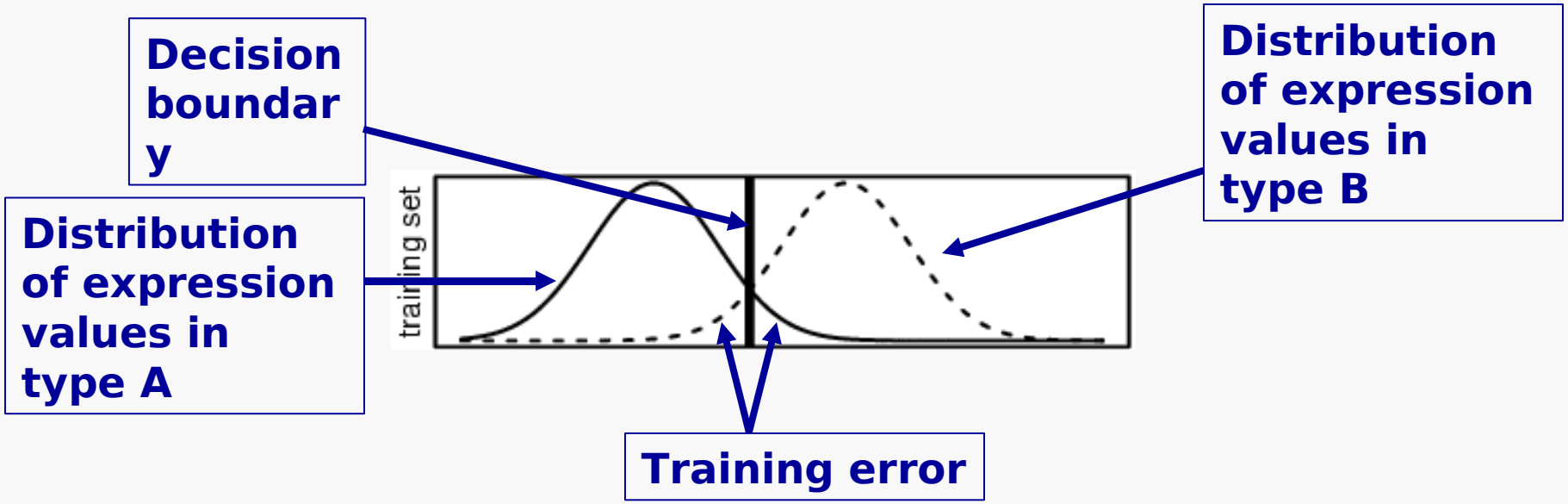
Approach:

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

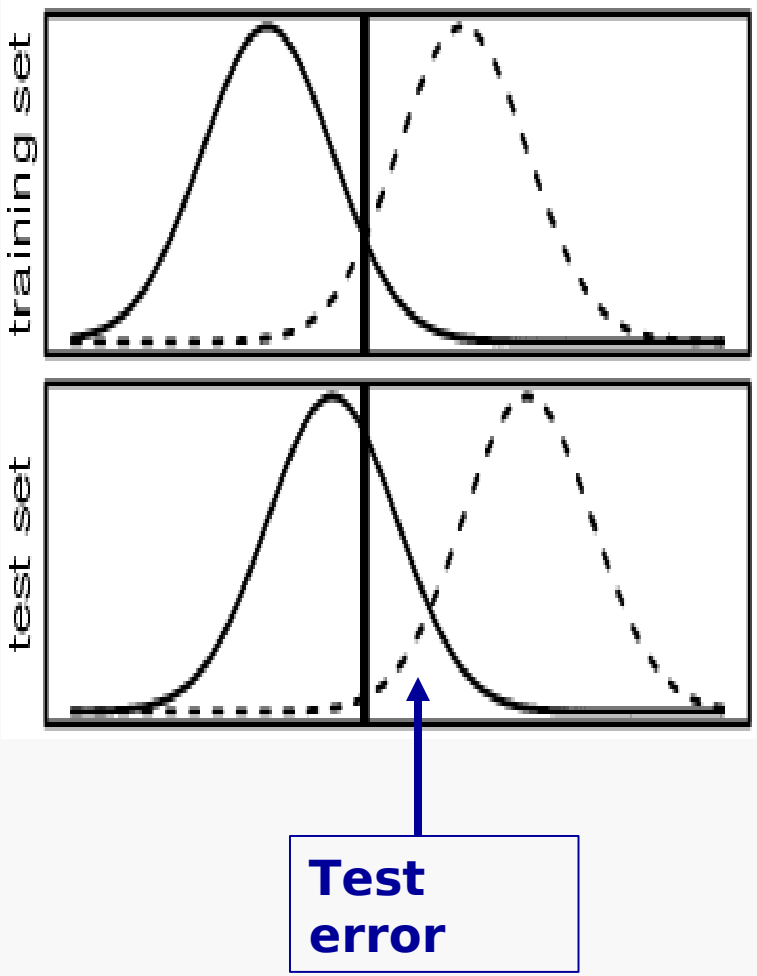
„ Smallest *training error* “

Zero training error is not possible!

A more schematic illustration:



What about the test samples?



The decision boundary was chosen to minimize the trainings 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

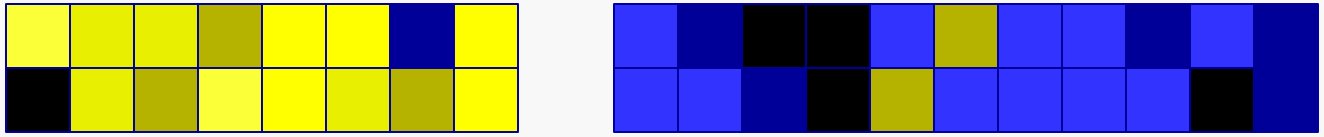
Test errors are in average bigger then training errors

This phenomenon is called *overfitting*

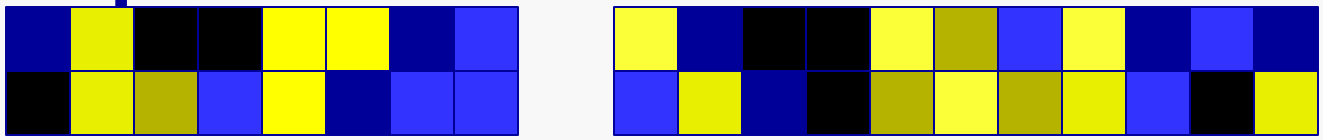
Prediction with 1 gene

The gene is differentially expressed

Prediction with 2

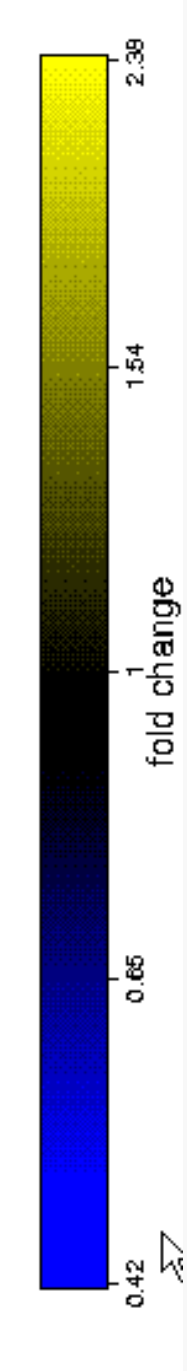


Both genes are differentially expressed



These genes are not differentially expressed.

Can they be of any use?

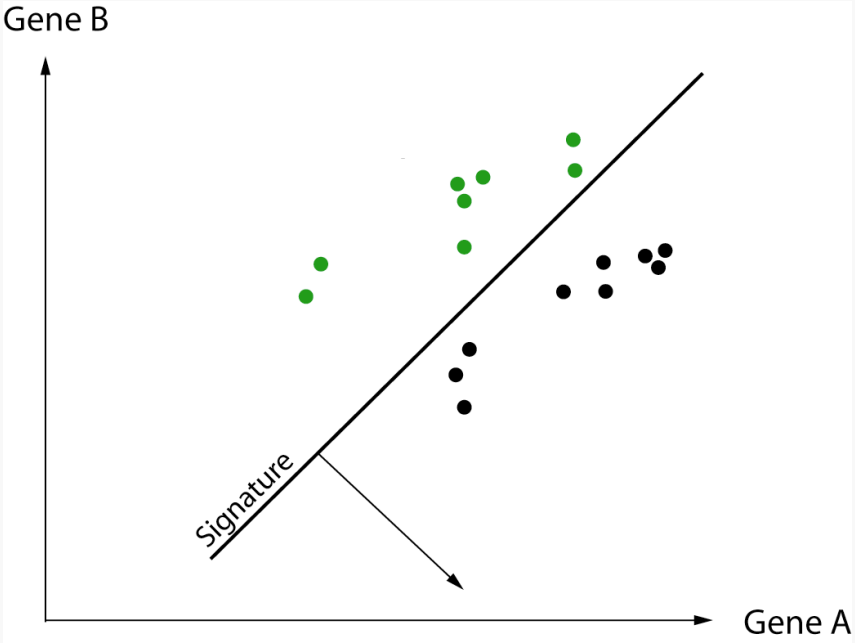


Interacting genes

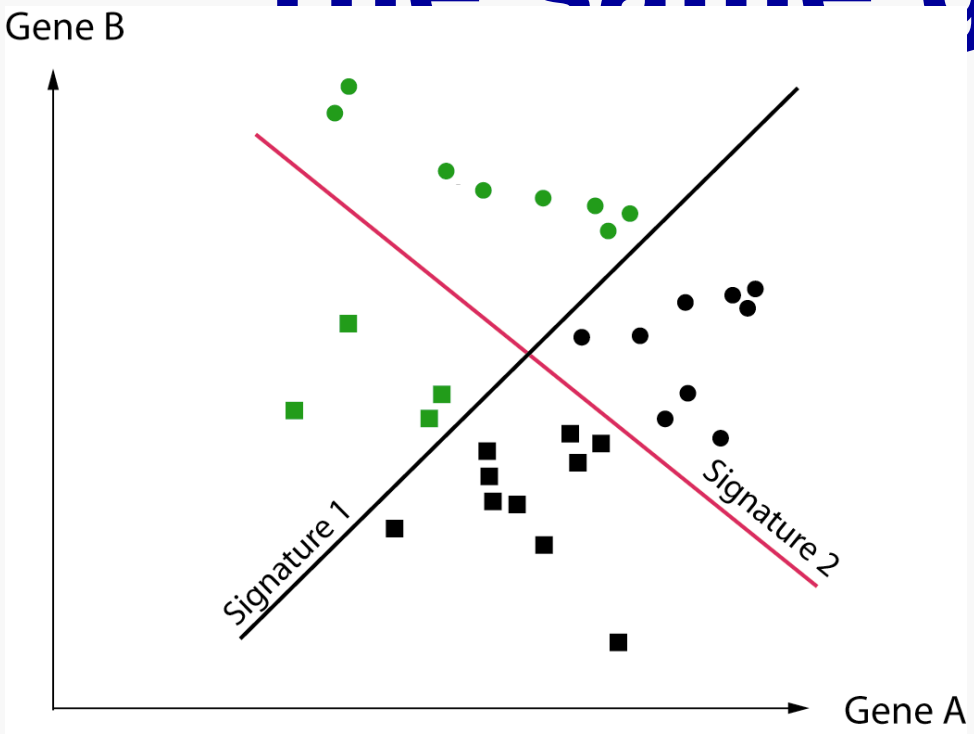
Assume protein A binds to protein B and inhibits it

The clinical phenotype is caused by active protein A

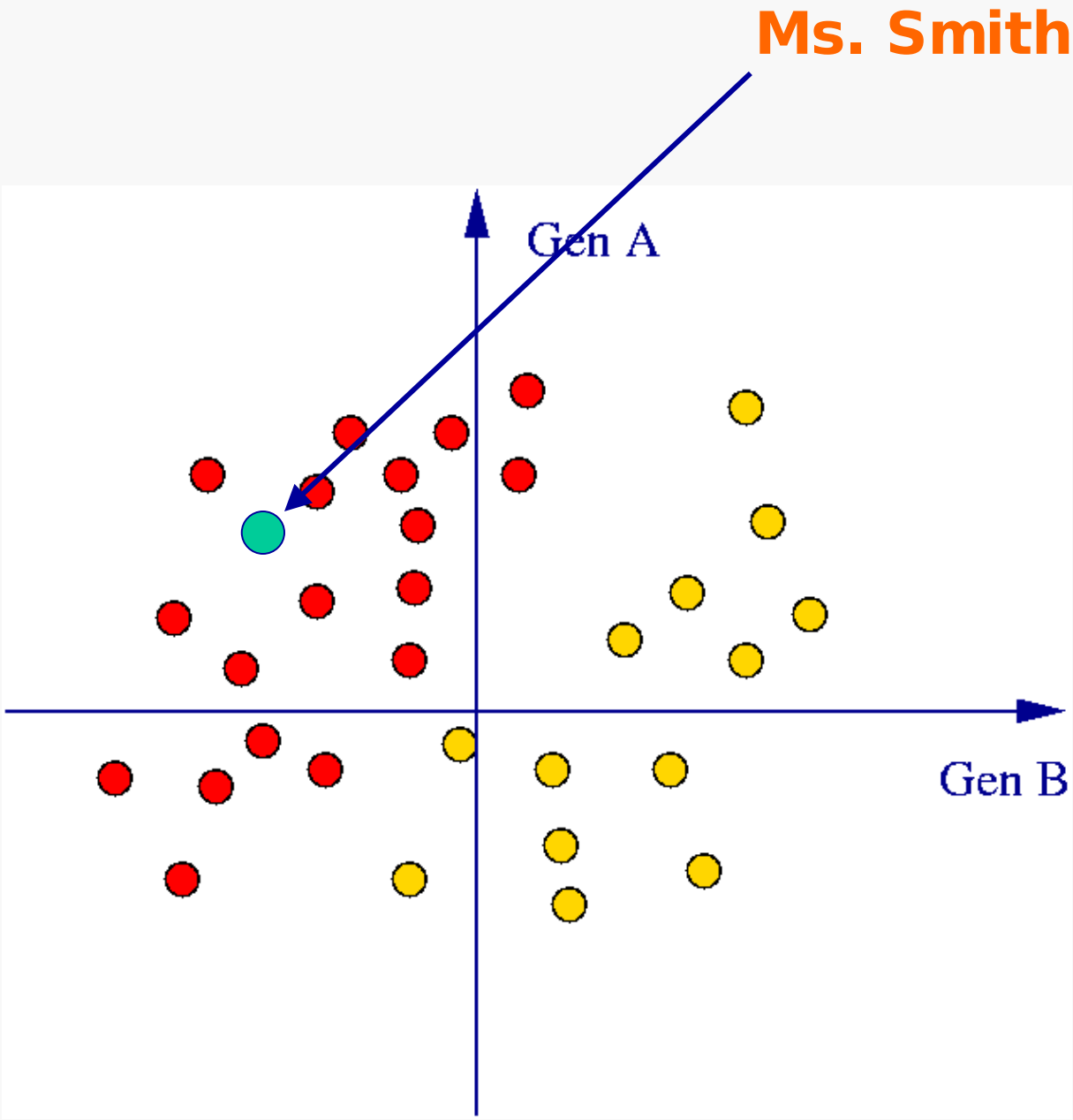
Predictive information is in expression of A minus expression of B

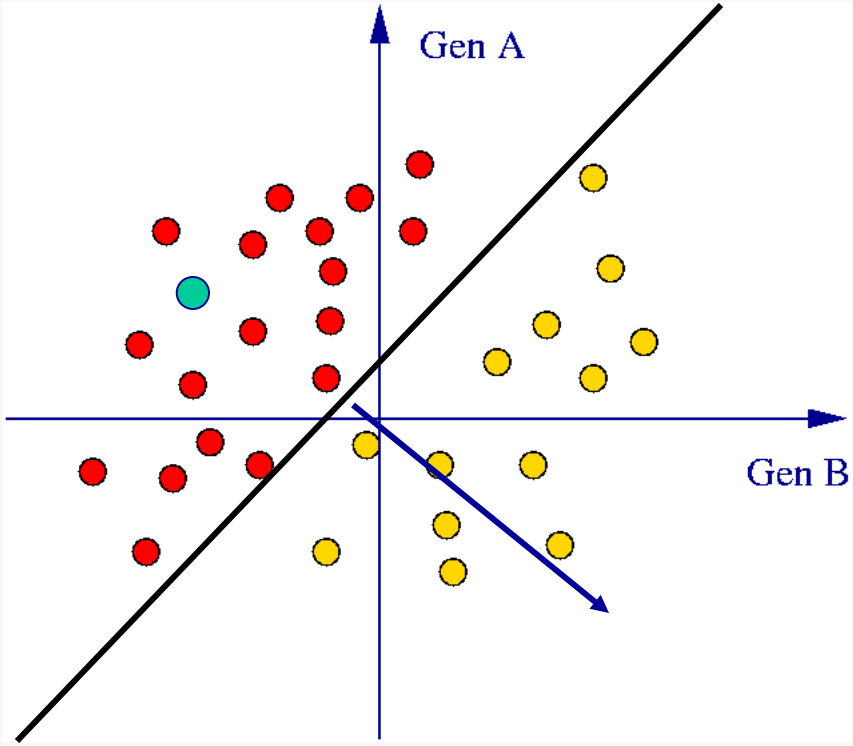


Two different signatures based on the same genes



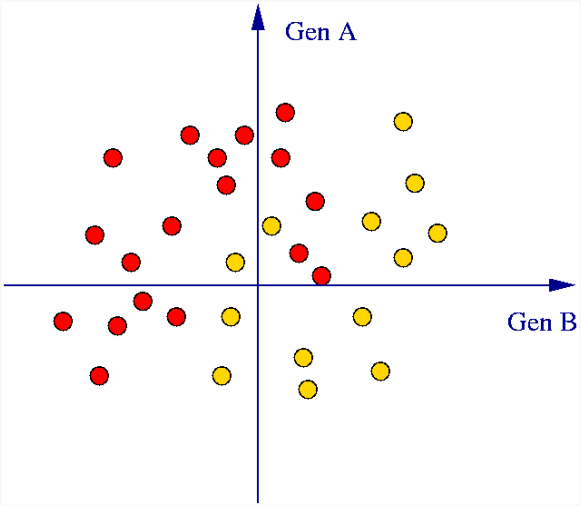
Calling signature genes markers for a certain disease is misleading!


Ms. Smith
Gen A
Gen B



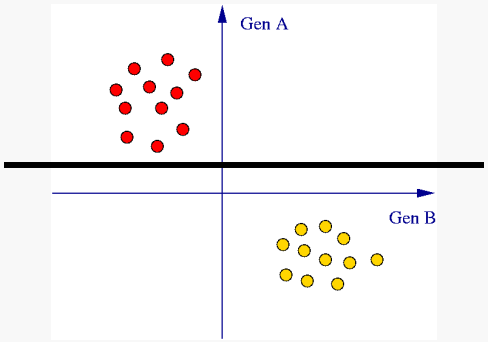
A decision boundary can be defined by a weighted sum (linear combination) of expression values

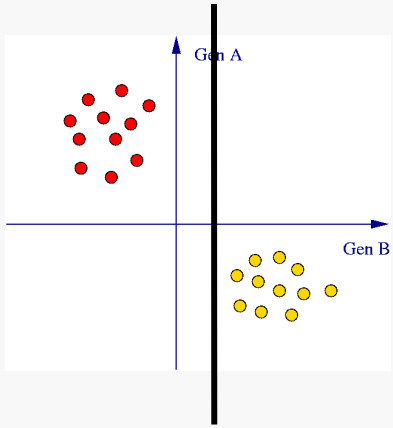
→ Separating Signature



Problem 1:

No separating line



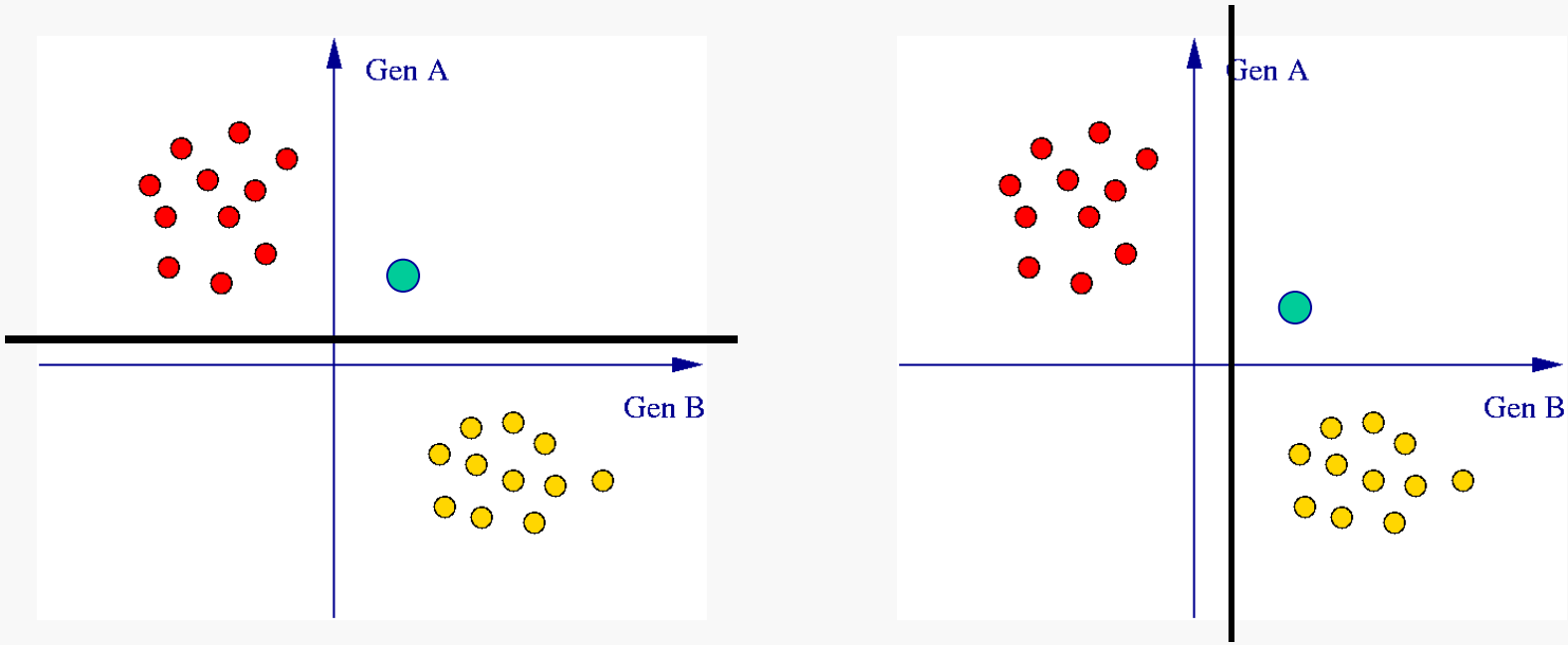


Problem 2:

To many separating lines

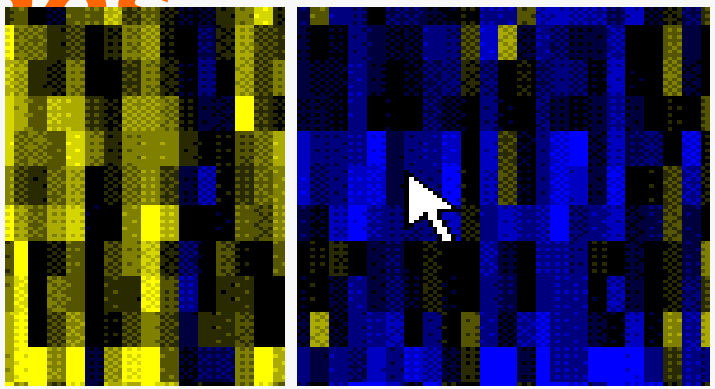
Why is this a problem?

What about Ms. Smith ?

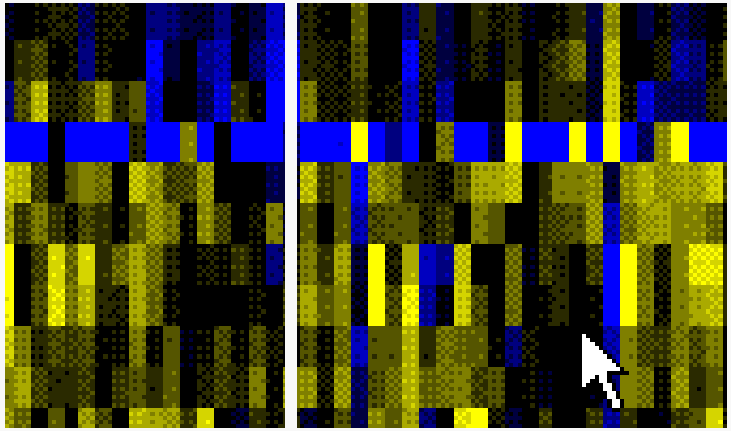


This problem is also related to overfitting … more soon

Prediction with more genes



with differentially expressed genes ...



... or with multivariate signatures

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

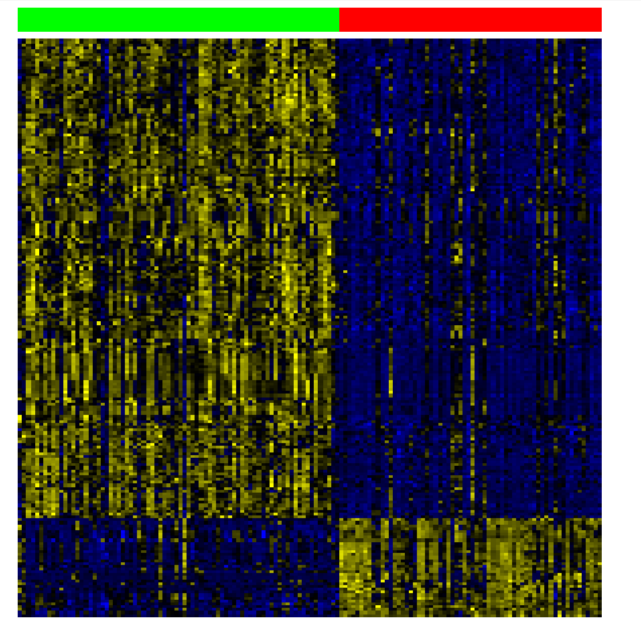
Finding the needle in the haystack

A common myth:

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

The avalanche

Aggressive lymphomas with and without a MYC-breakpoint



MYC-neg MYC-pos

Verbundprojekt maligne Lymphome

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

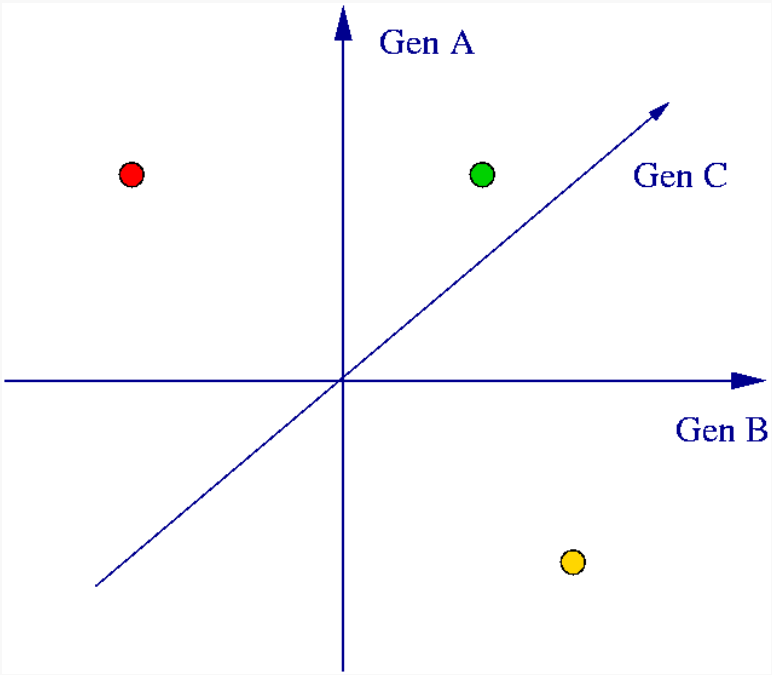
Prediction with 30,000 genes

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

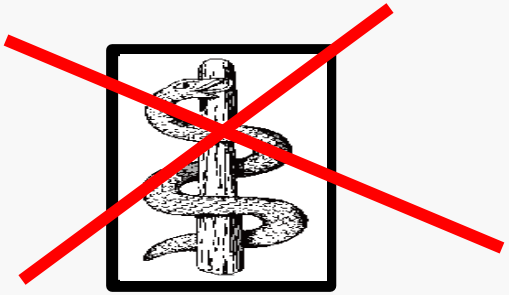


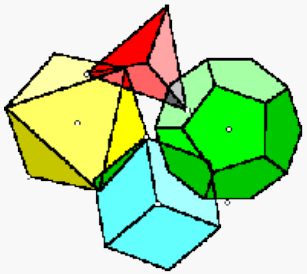
OK! If all points fall onto one line it does not always work. However, for measured values this is very unlikely and never happens in praxis.

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.

More important consequences of understanding 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 !!!

More later …

Strategies for finding meaningful signatures ?

Later we will discuss 2 possible approaches

4. **Gene selection followed by discriminant analysis (QDA,LDA,DLDA), and the PAM program**
5. **Support Vector Machines**
6. **Random Forests**

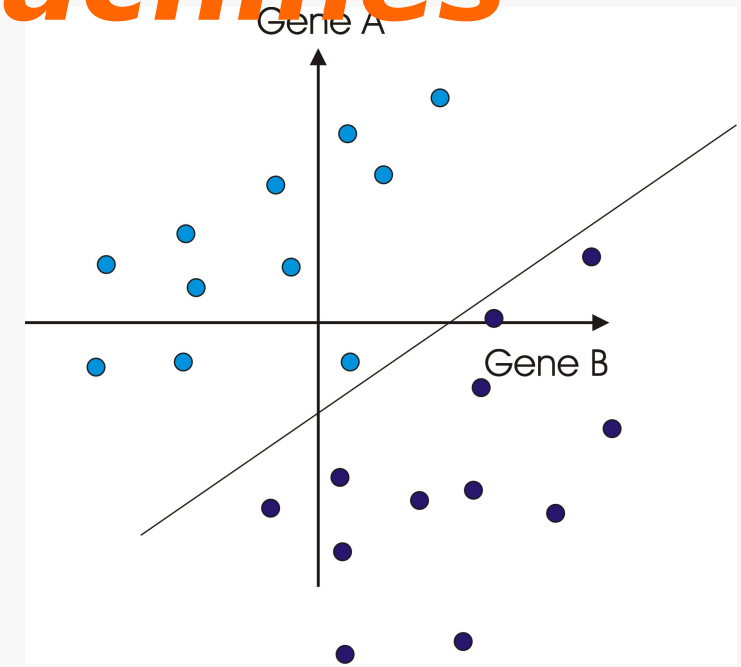
What is the basis for this methods?

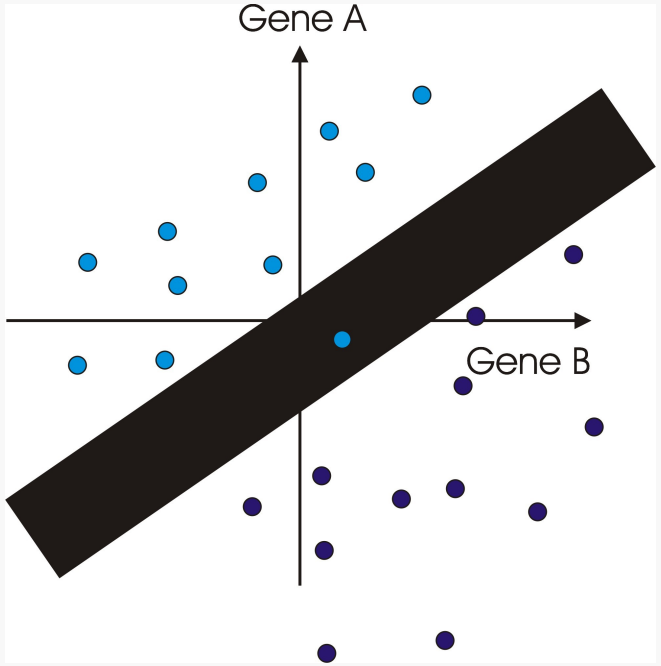
Gene selection

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





Fat planes: **With an infinitely thin plane the data can always be separated correctly, but not necessarily with a fat one.**

Again if a large margin separation exists, chances are good that we found something relevant.

Regularization

Both gene selection and Support Vector Machines confine the set of a priori possible signatures. However, using different strategies.

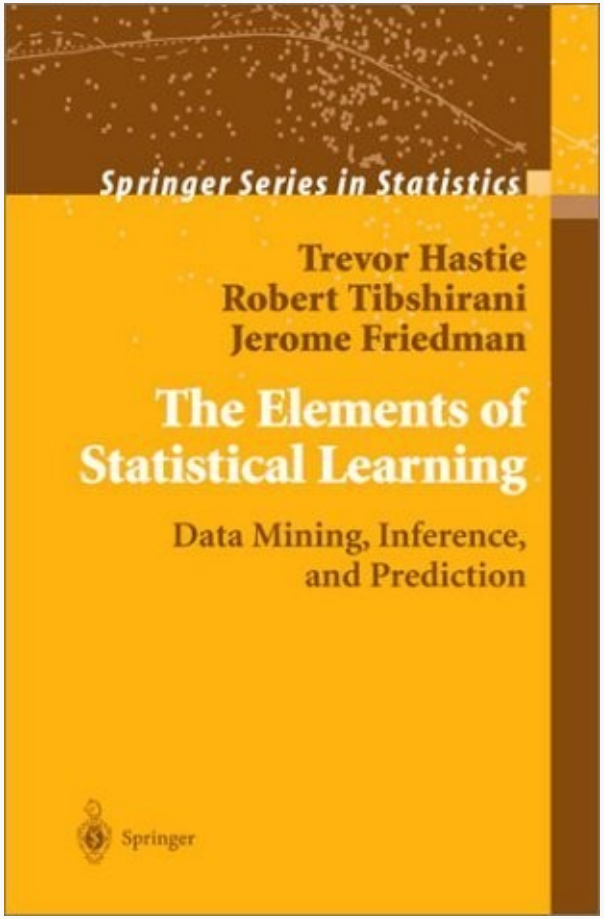
Gene selection wants a small number of genes in the signature *- sparse model -*

SVMs want some minimal distance between data points and the separating plane

- large margin models -

There is more than you could do …

Learning Theory



Ridge regression, LASSO, Kernel based methods, additive models, classification trees, bagging, boosting, neural nets, relevance vector machines, nearest-neighbors, transduction etc. etc.

Questions?

Coffee