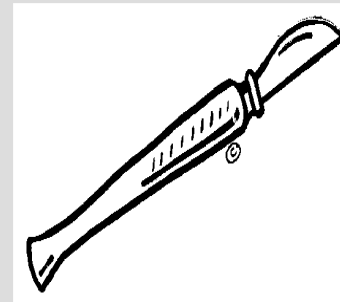# Diagnosis using computers

One disease

Three therapies

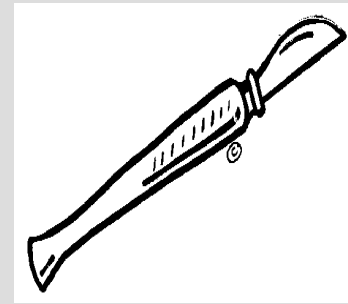# Clinical Studies

In average

75%   55%   35%

Success

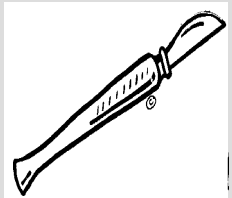# Three subtypes of the disease

A

B

C

	A	B	C
💊	**100%**	60%	65%
💉	40%	40%	**85%**
🔪	10%	**90%**	5%

A — 100%

B — 90%

C — 85%

91,7%
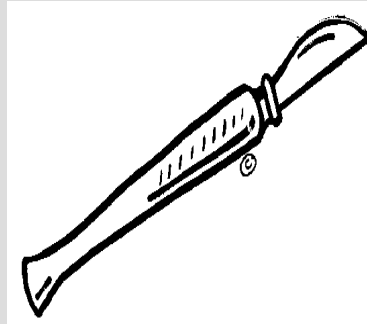
Therapeutic success improved because of the refined diagnosis

91,7%

75%
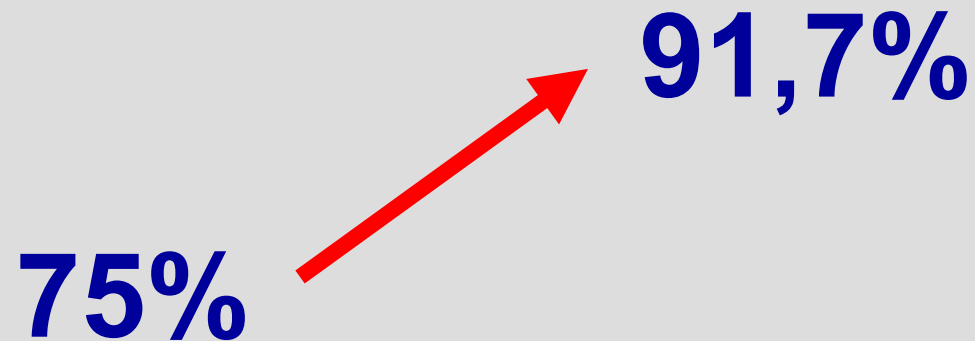
Without developing any new therapies

A higher resolution of dividing a disease into subtypes improves therapeutic success rates
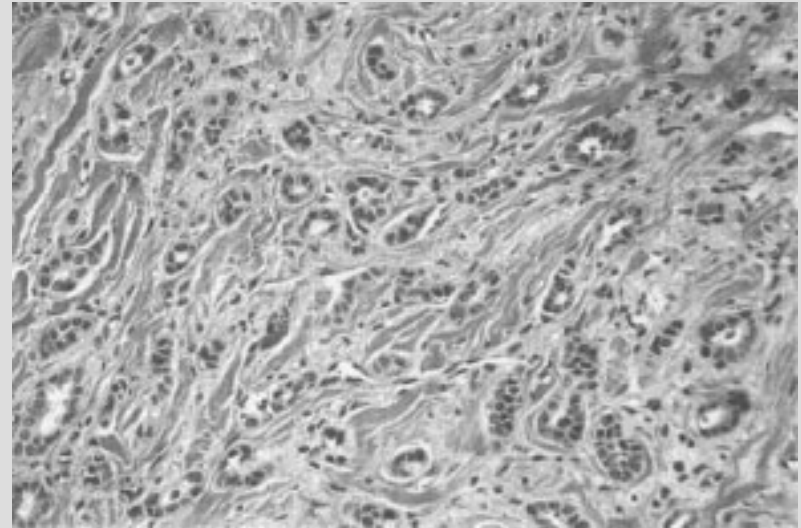
How do we obtain a higher resolution of diagnosis that is clinically relevant?
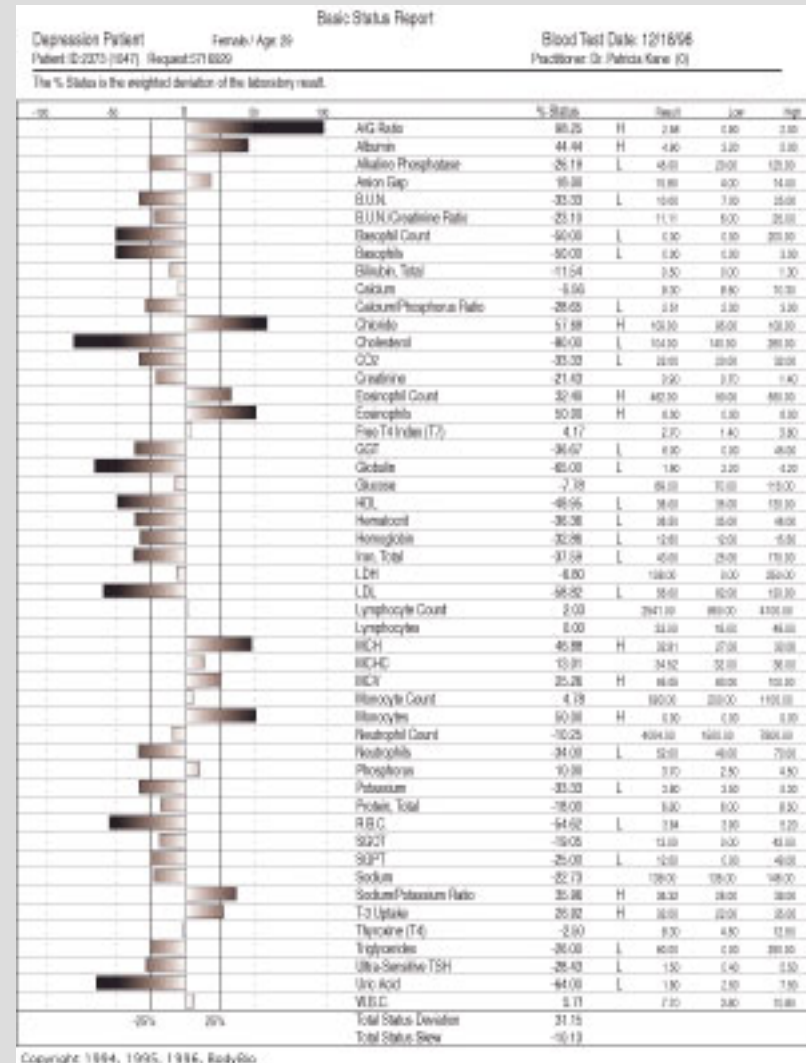
Looking at cells from outside

The microscope

Details of Metabolism:
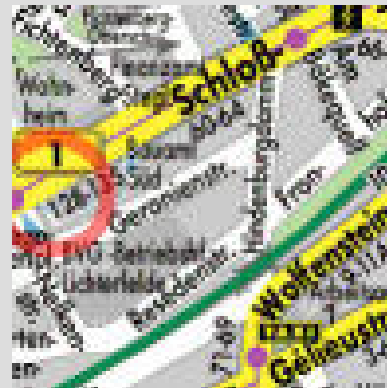
The hemogram

Diagnostics crabwise

- **Deregulation of metabolism causes disease**

- **Occasionally, they also lead to characteristic changes in tissue morphology or the hemogram.**
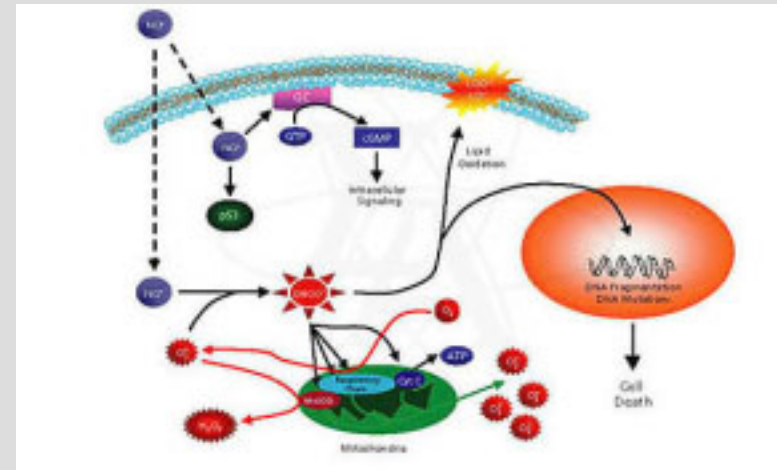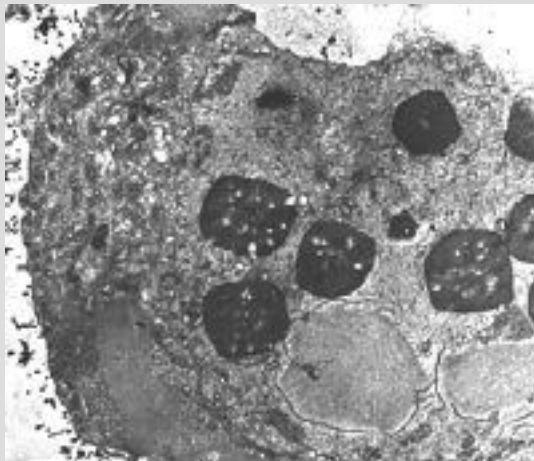
Diagnostics based on details

- **A small number of genetic variations, transcription levels, and protein expression levels are routinely measured in single assays.**
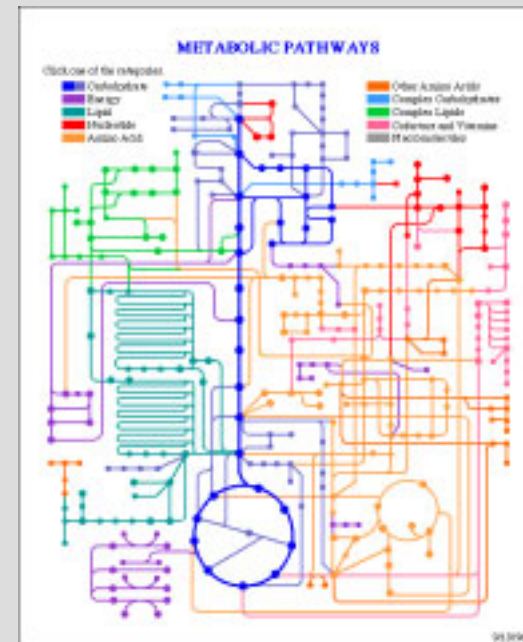
Desirable

- **Looking into cells and not onto cells**
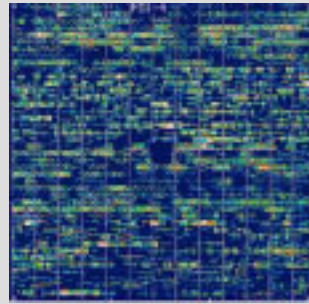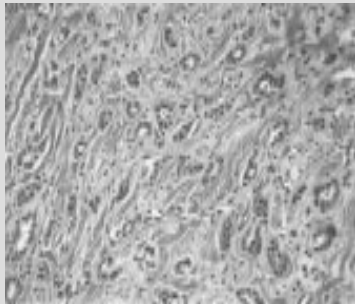- **A protocol of what is going on in the cells**

In addition desirable

- **A patients metabolism in a bird's eye view**

DNA Chip

Tissue

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

5%

Expression

profile

Ok, what is the problem ?

Morphological differences and differences in single assay measurements are the basis of classical diagnosis

A
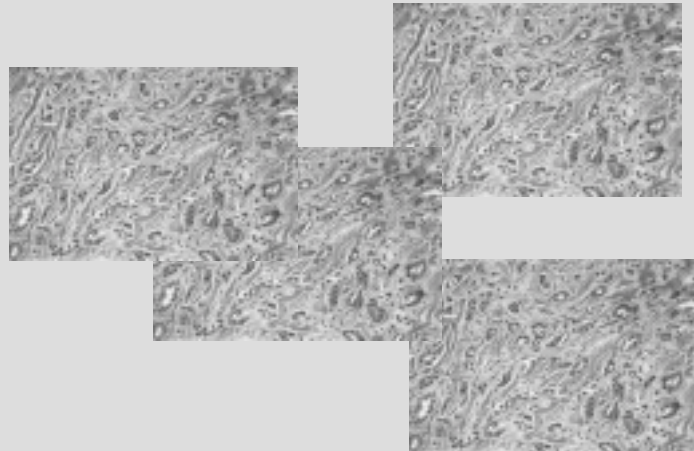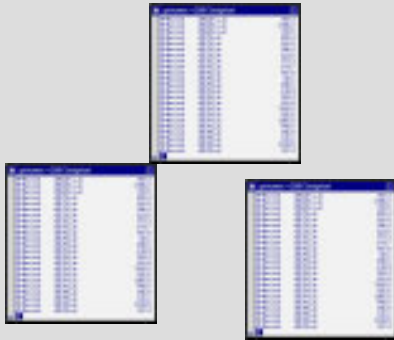
B
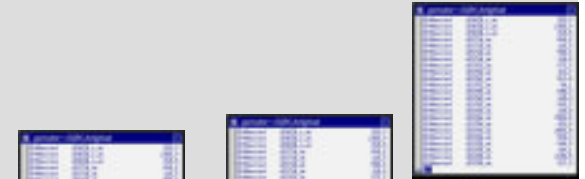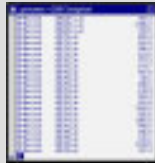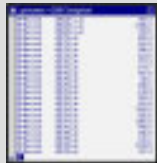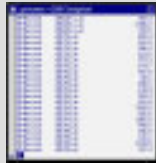
What about differences in the profiles?

Do they exist?
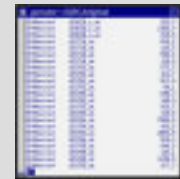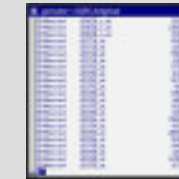
What do they mean?

A

B

Are there any differences between the gene expression profiles of type A patients and type B patients?

30.000 genes are a lot. That's to complex to start with

Let's start with considering only two genes:

gene A und gene B

In this situation we can see that ...
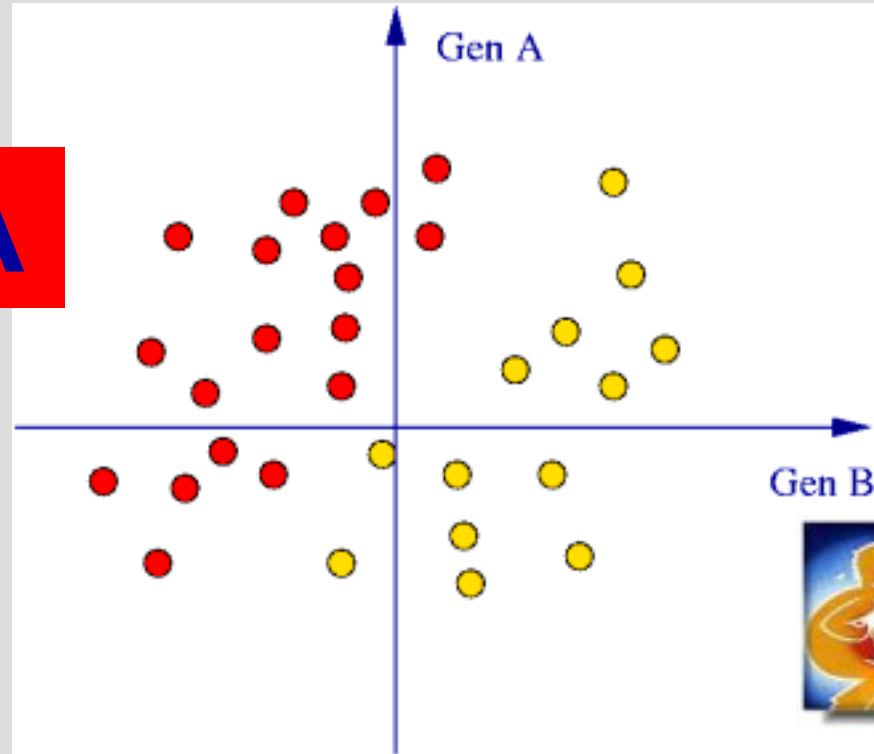
... there is a difference.
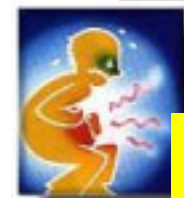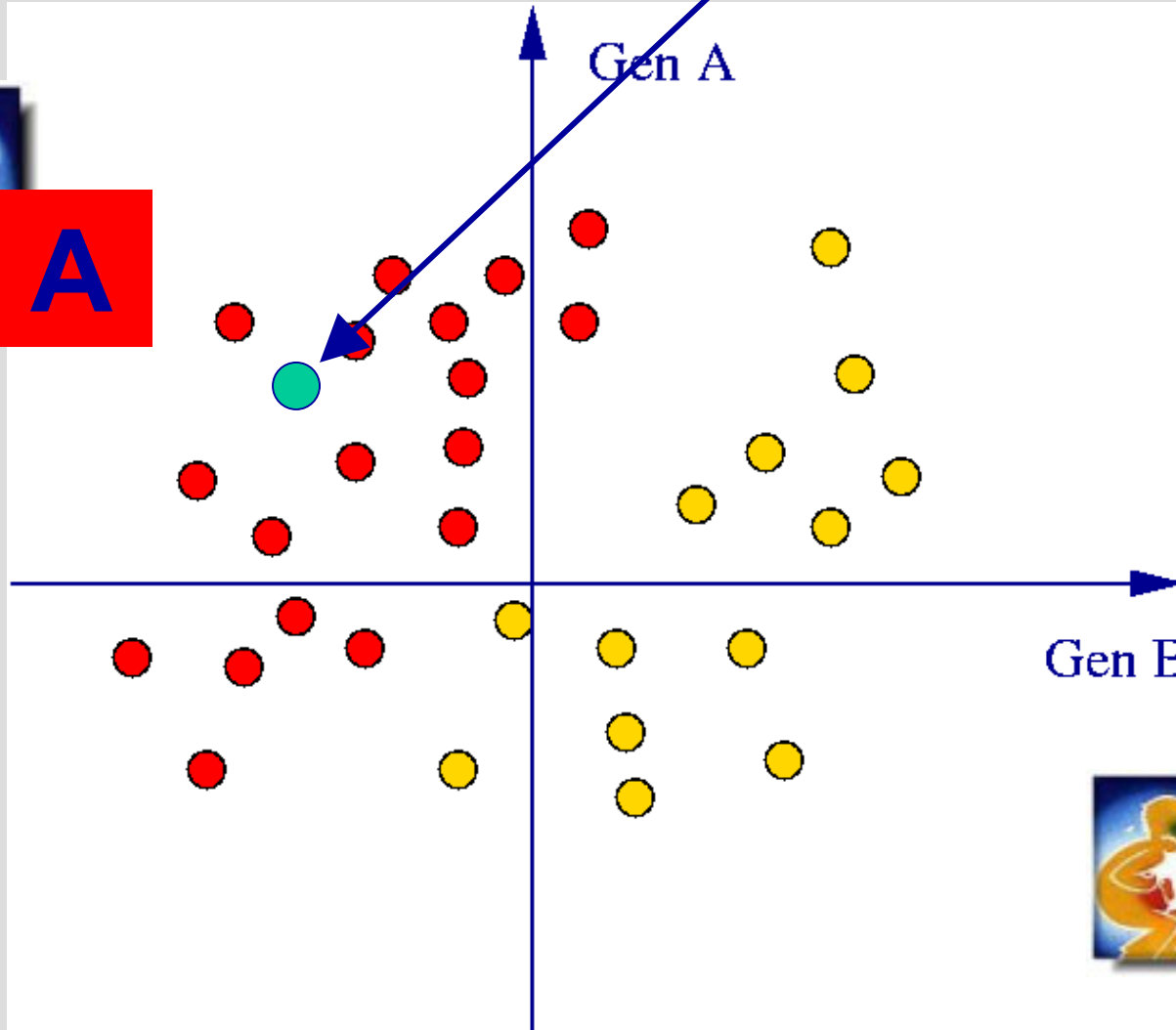
The new patient
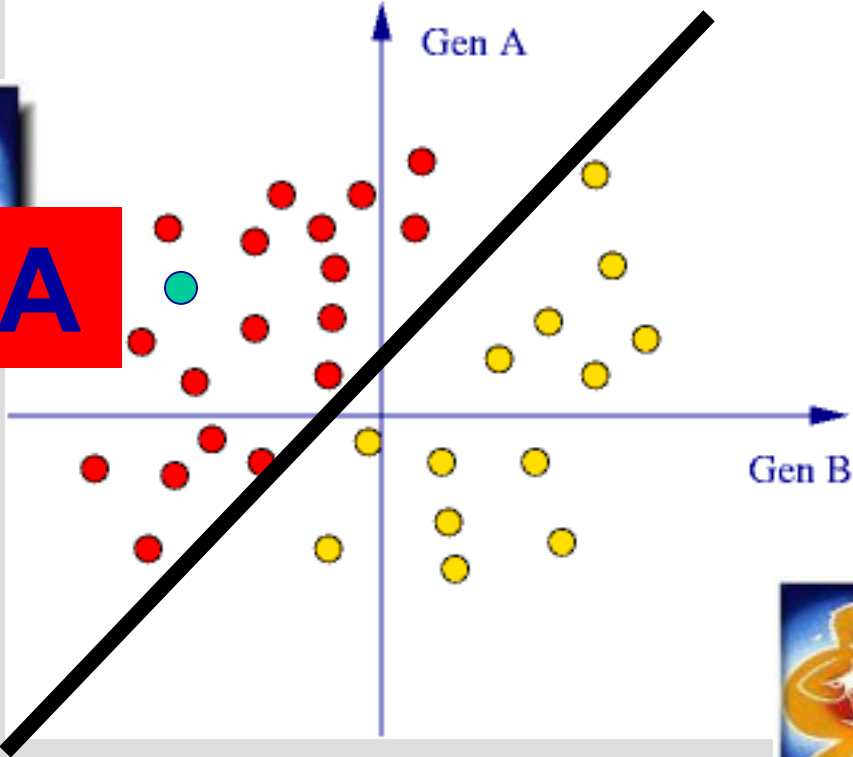
Here everything is clear.
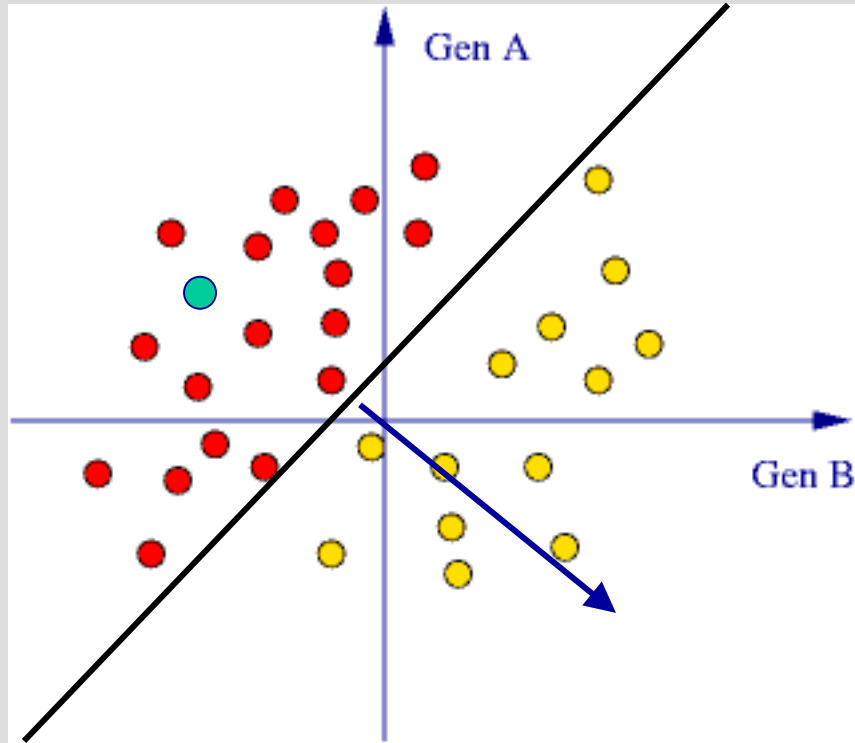
The normal vector of the separating line can be used as a signature

.... the separating line is not unique

What exactly do we mean if we talk about signatures?

$$x_1, \ldots, x_{30000}: \quad \text{expression levels}$$

$f(x_1, \ldots, x_{30000}):$ Mapping that assigns one number

                           to the expression levels

High values of f indicate class 1
low values class 2

Example:

$$f(x_1, \ldots, x_{30000}) = x_1$$ **gene 1 is the signature**

Or a normal vector is the signature:

$$f(x_1, \ldots, x_{30000}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
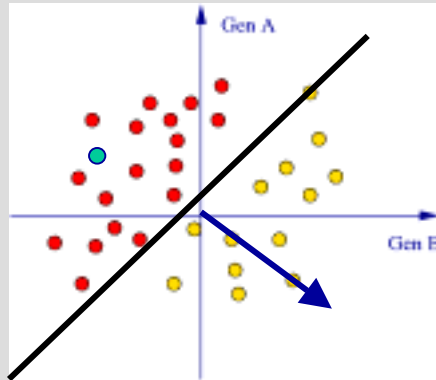
if x1 and x2 are the two genes in the Diagram

Using all genes yields:

$$f(x_1, \ldots, x_{30000}) = \beta_0 + \sum_{i=1}^{30000} \beta_i x_i$$

Or you choose a very complicated signature
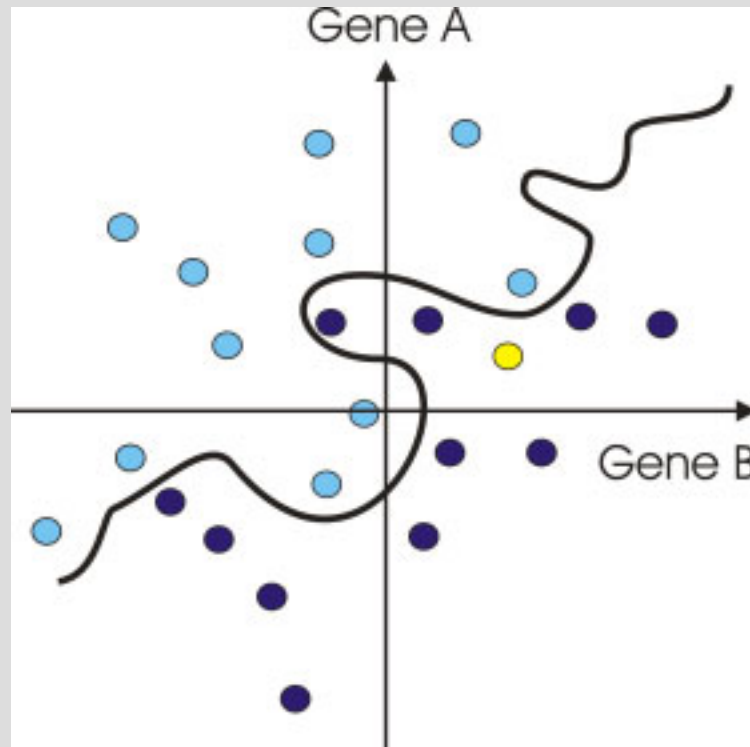
$$f(x_1, \dots, x_{30000}) = \text{complicated}$$
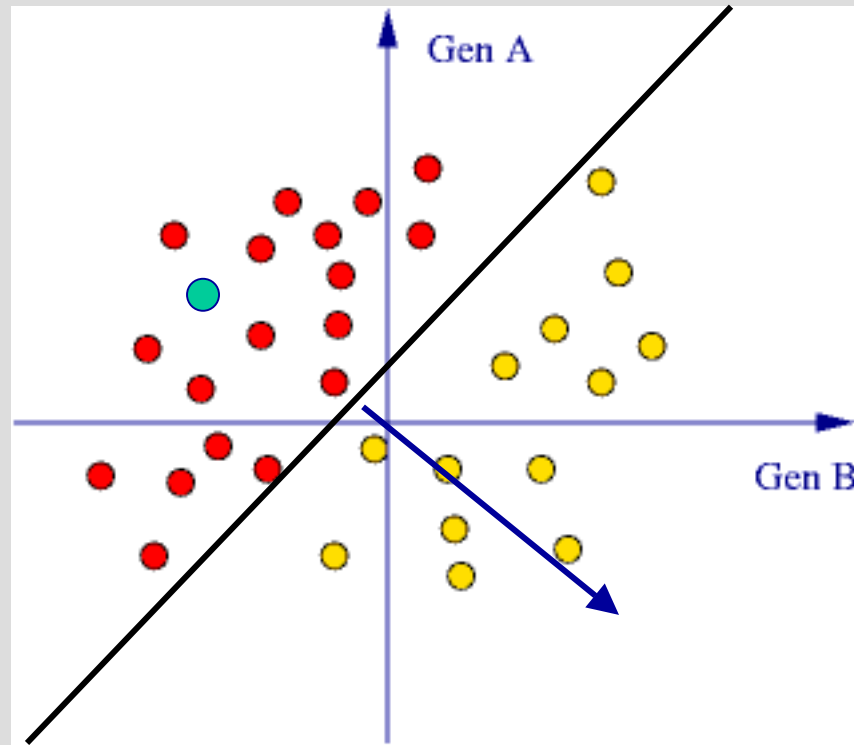
Unfortunately, expression data is different.

What can go wrong?

There is no separating straight line
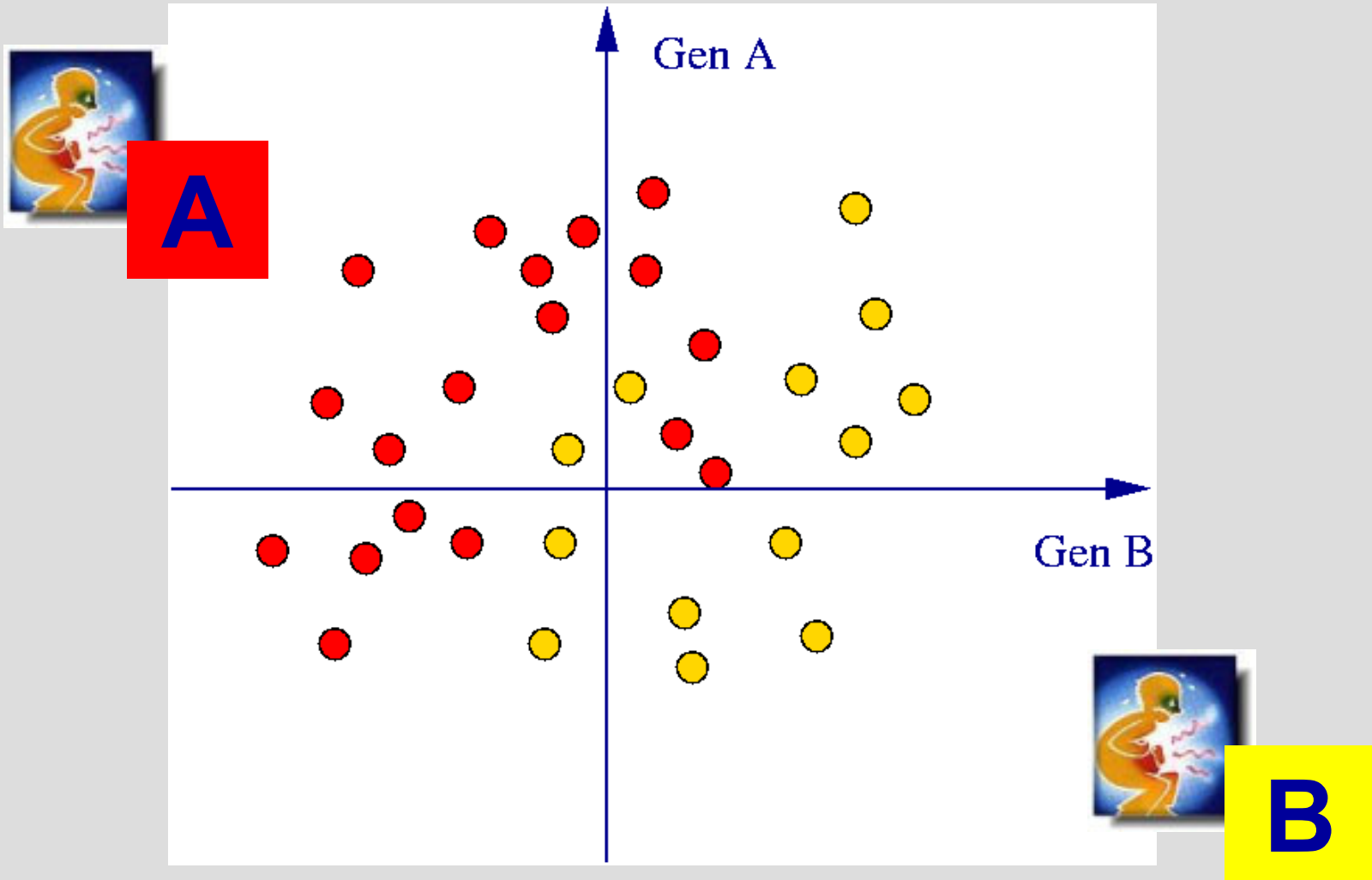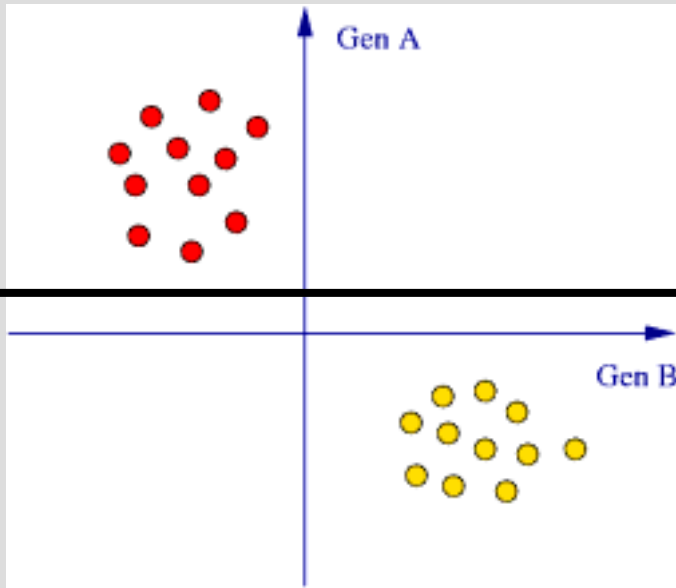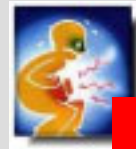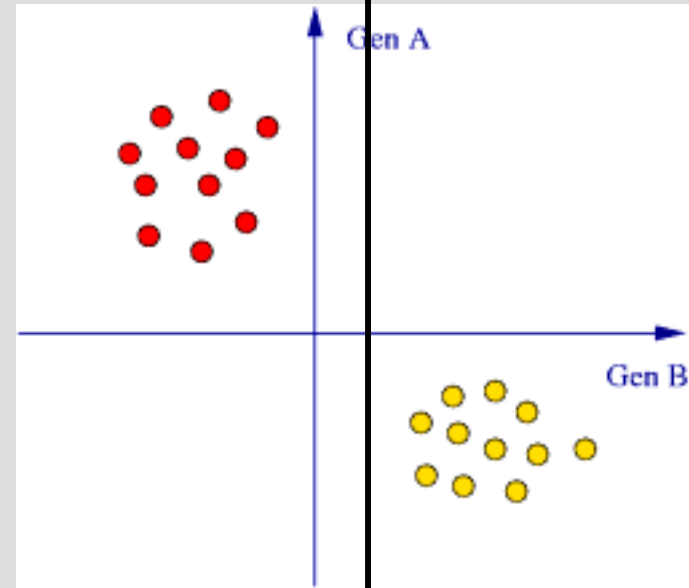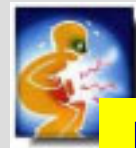
New patient ?
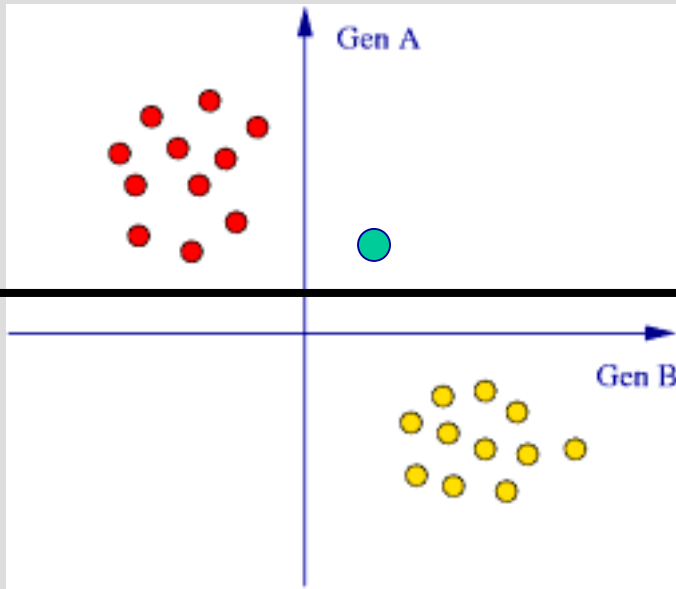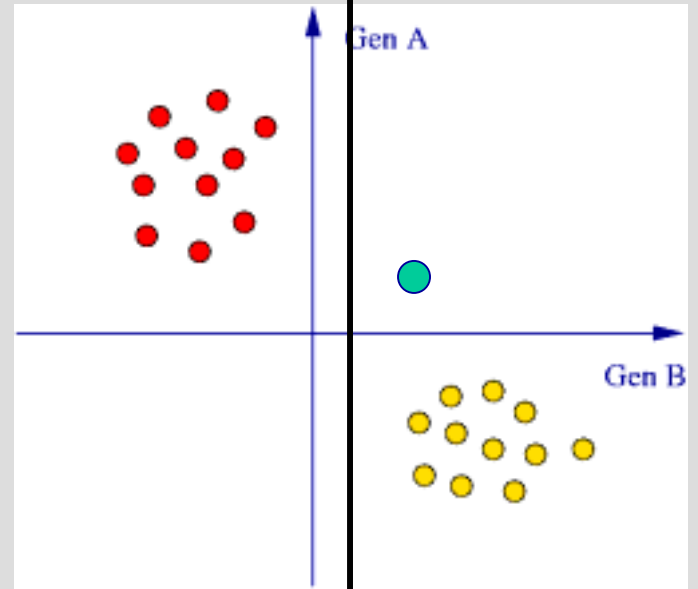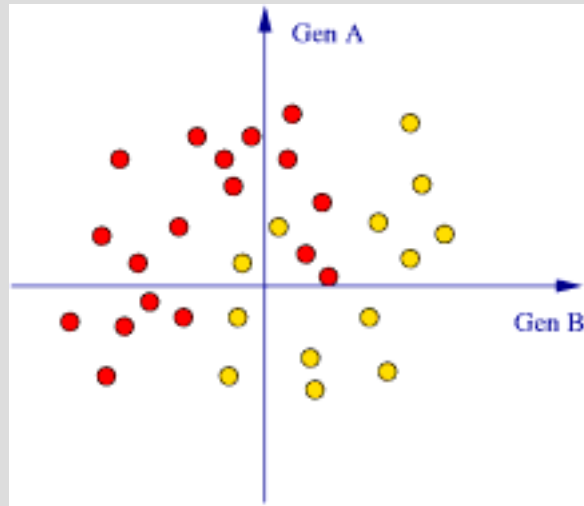
Problem 1:

No separating line
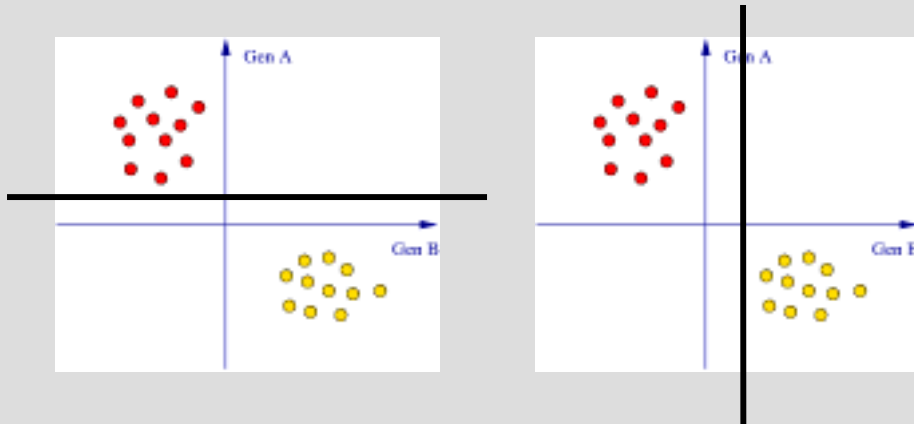
Problem 2:

To many separating lines

In praxis we look at thousands of genes, generally more genes than patients
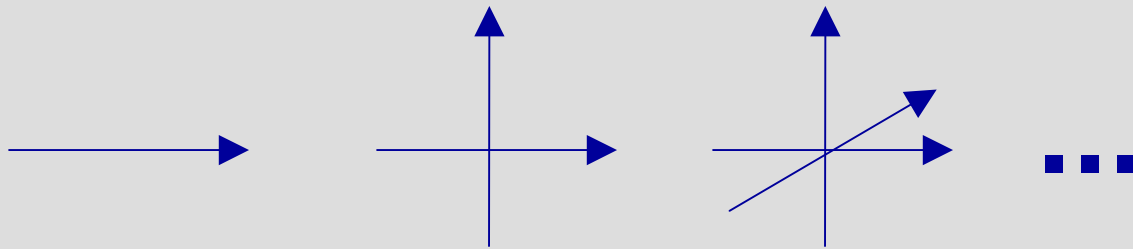
An in 30000 dimensional spaces different laws apply

1
30000

2

3

...
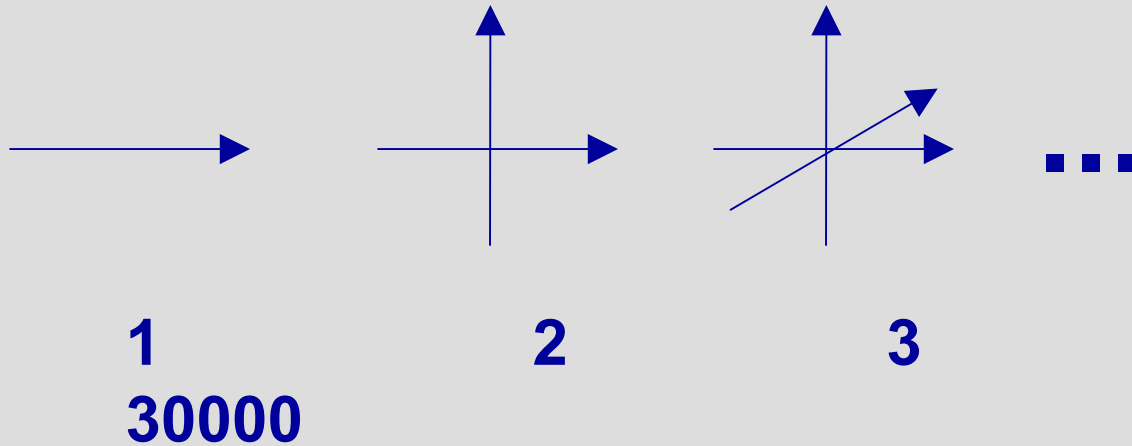
- **Problem 1 never exists!**
- **Problem 2 exists almost always!**

Spent a minute thinking about this in three dimensions

Ok, there are three genes, two patients with known diagnosis, one patient of unknown diagnosis, and separating planes instead of lines
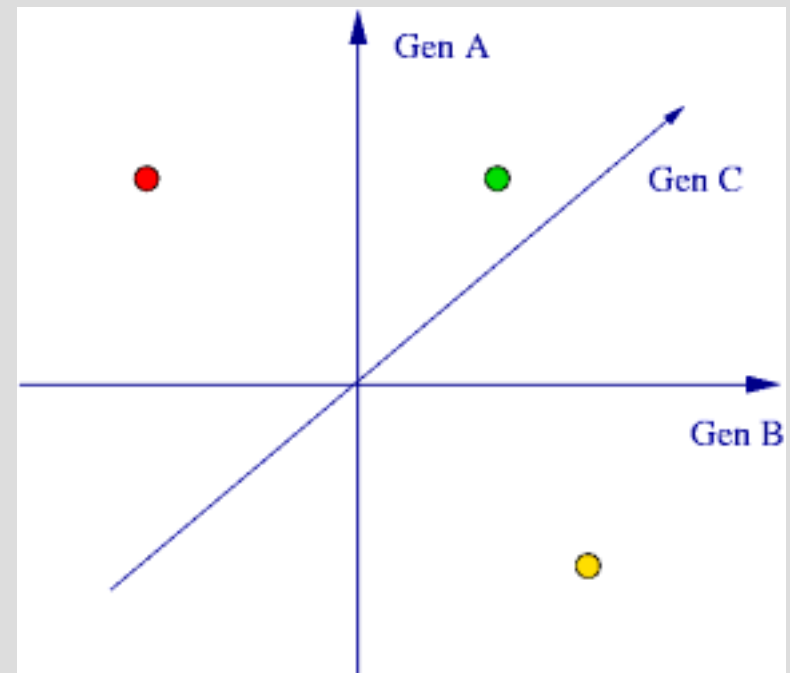
OK! If all points fall onto one line it does not always work. However, for measured values this is very unlikely and never happens in praxis.

With more gene than patients the following problem exists:
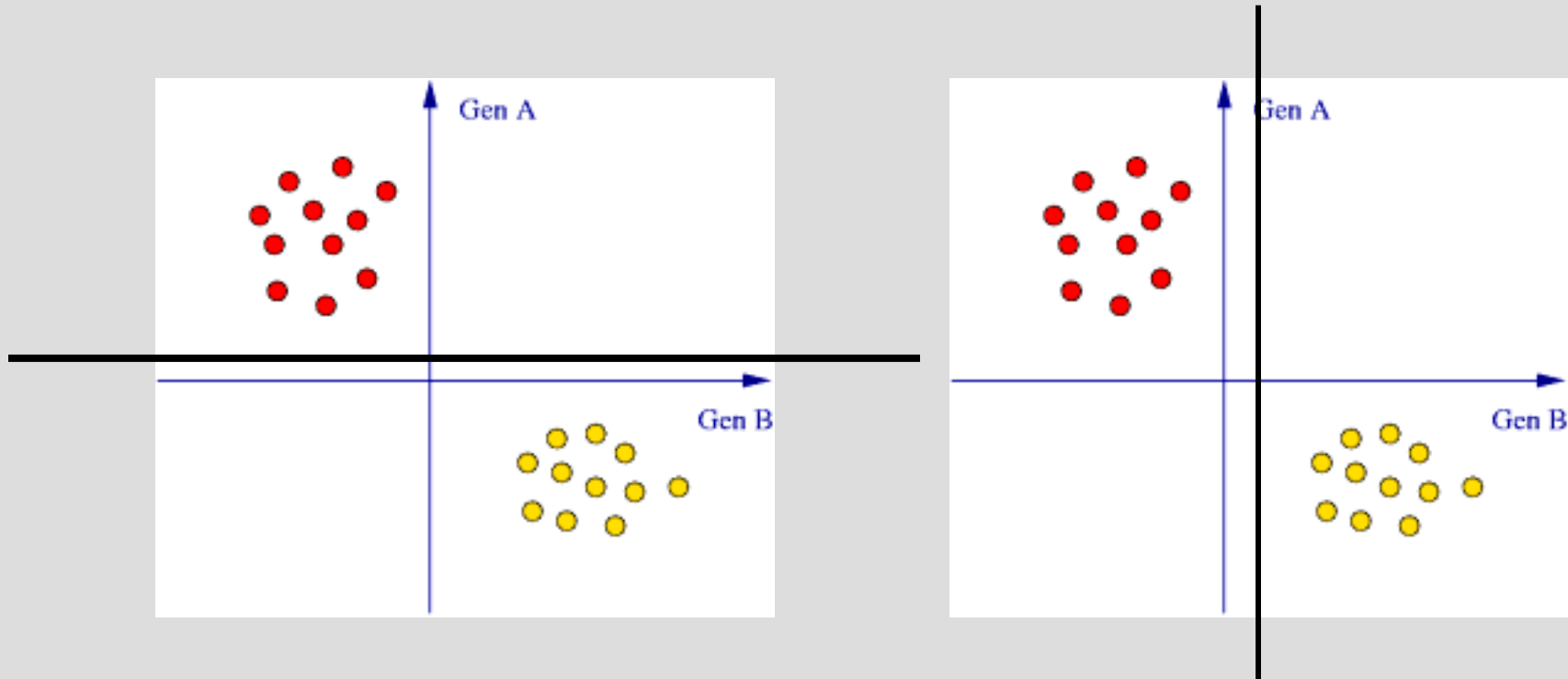
Hence for microarray data it always exists

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.
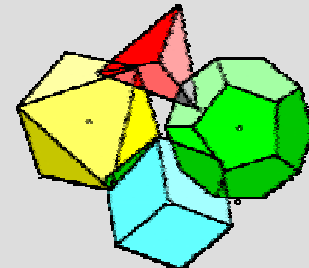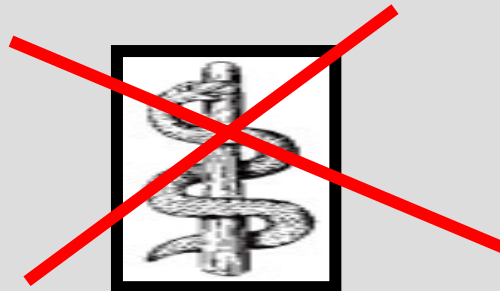
Whenever you have expression profiles from two groups of patients, you will find differences in their genes expression ...
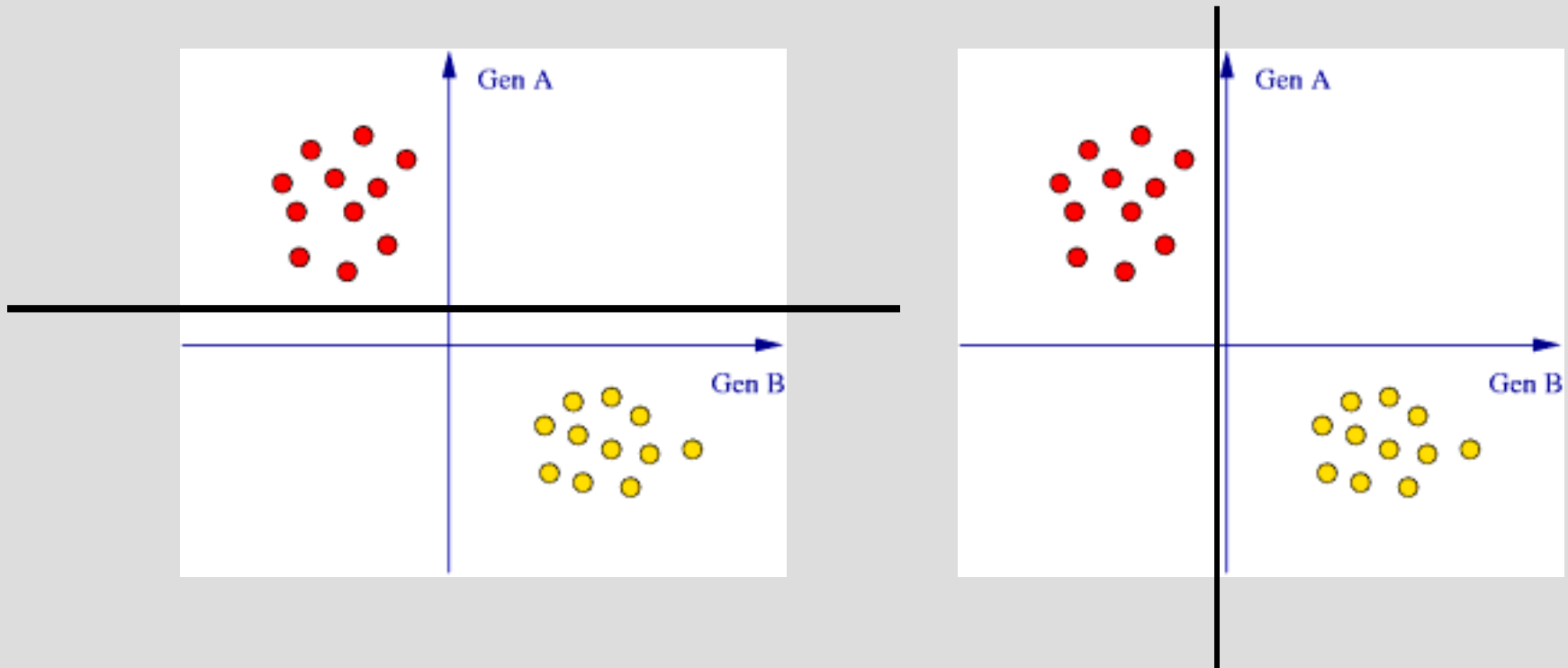
... no matter how the groups are defined .

There is a guarantee that you find a signature:

- which separates malignant from benign tumors

- but also

- Müllers from Schmidts

- or using an arbitrary order of patients odd numbers from even numbers

In summary:

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

... in most cases it means nothing.

Wait! Believe me!

There are meaningful differences in gene expression. And these must be reflected on the chips.

Ok,OK...

On the one hand we know that there are completely meaningless signatures and on the other hand we know that there must be real disorder in the gene expression of certain genes in diseased tissues

How can the two cases be distinguished?

What are characteristics of meaningless signatures?

They come in large numbers

Parameters have high variances

Under-determined models

We have searched in a huge set of possible signatures

No regularization

They refect details and not essentials

Overfitting

Under-determined models

They come in large numbers

Parameters have high variances
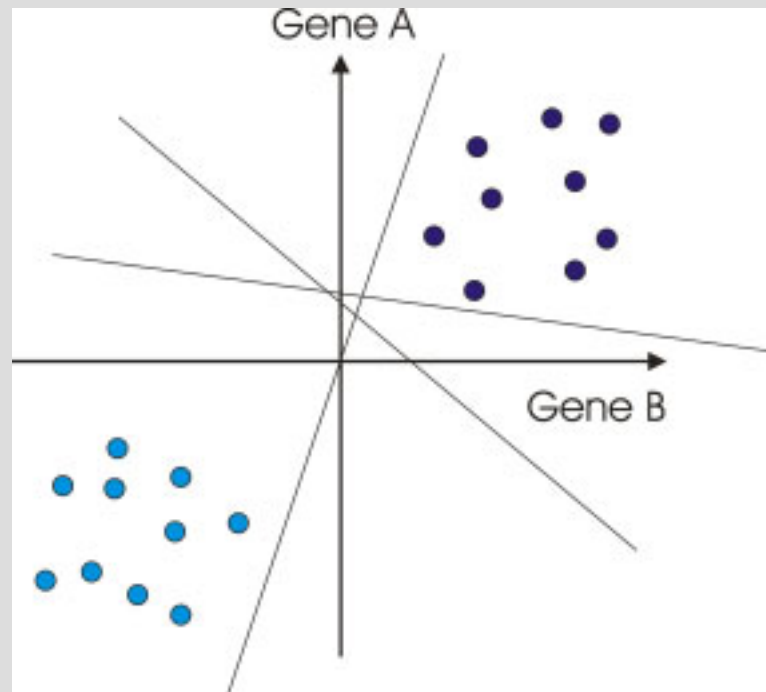
No regularization

We have searched in a huge set of possible signatures

When considering all possible separating planes there must always be one that fits perfectly, even in the case of no regulatory disorder

Overfitting

They reflect details and not essentials
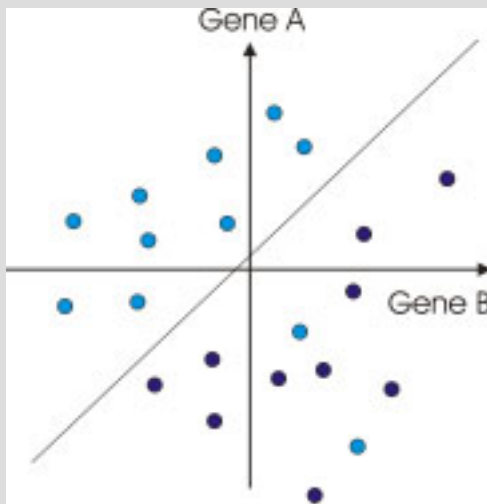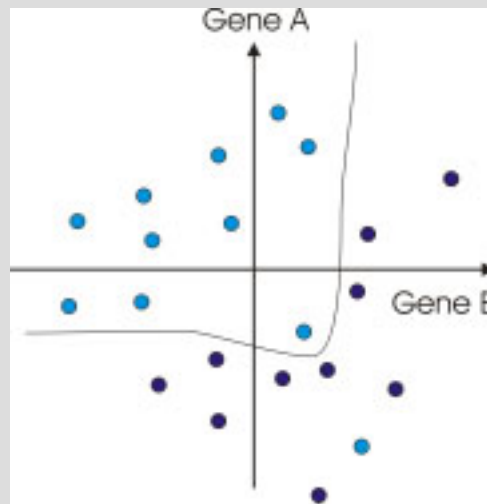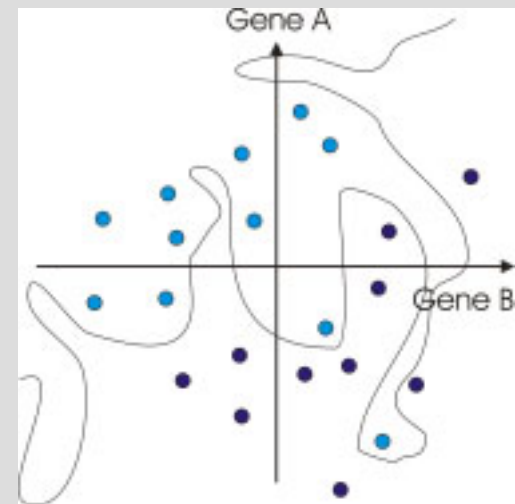
2 errors   1 error   no errors

Signatures do not need to be perfect

Examples for sets of possible signature

- All quadratic planes

- All linear planes

- All linear planes depending on at most 20 genes

- All linear planes depending on a given set of 20 genes

High probability for finding a fitting signature

Low probability that a signature is meaningful

Low probability for finding a fitting signature

High probability that a signature is meaningful

What are strategies for finding meaningful signatures?

Later we will discuss 2 possible approaches

1. Gene selection followed by linear discriminant analysis, and the PAM program

2. Support Vector Machines

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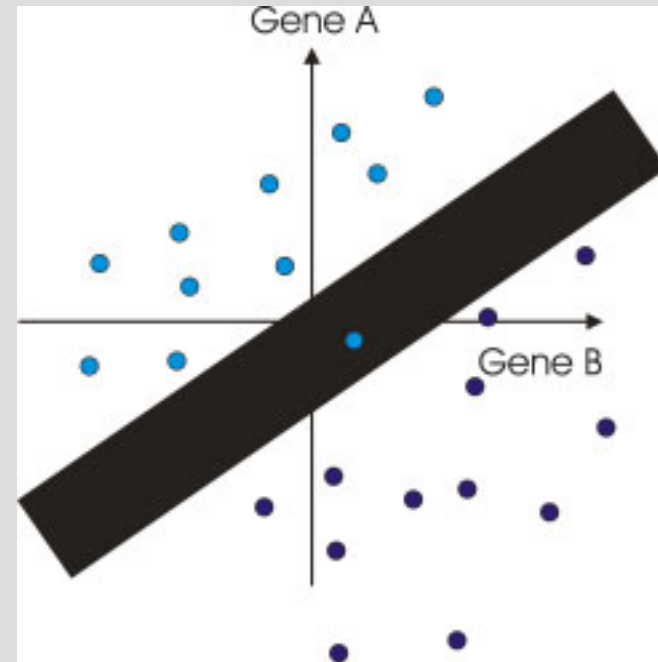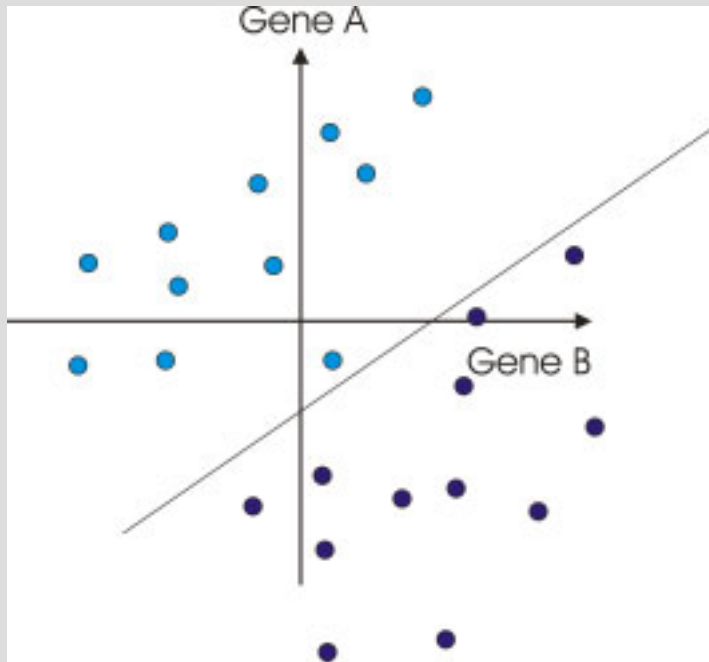
Fat planes: With an infinitely thin plane the data can always be separated correctly, but not necessarily with a fat one.

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

Large Margin Classifiers

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.
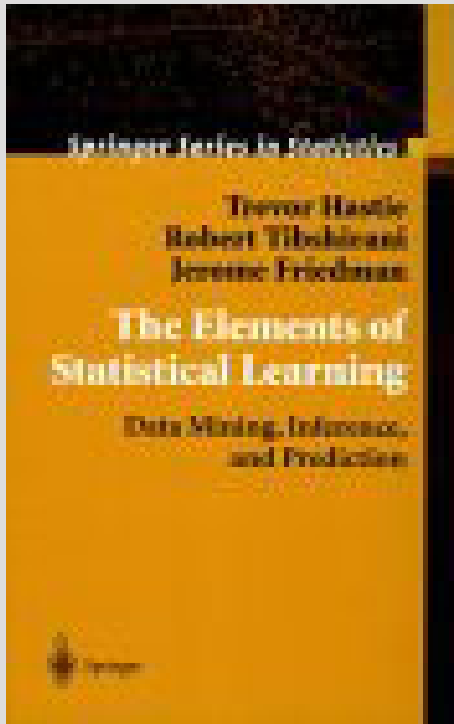
**The Elements of
Statistical Learning**

**Hastie, T. Tibshirani,
R. Friedman, J**
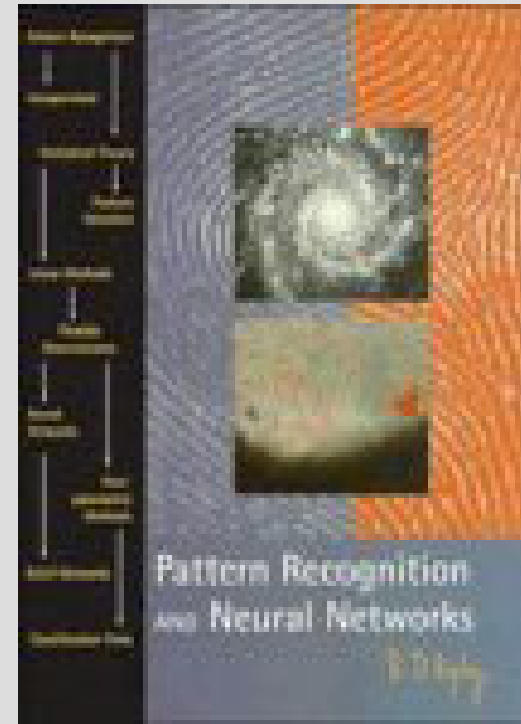
**Pattern
Recognition and
Neural Networks**

Brian D. Ripley

Questions

Coffee

Learning Methods

Setup

We have 200 patient profiles and 30000 genes on the chip

Patients can be divided into two groups according to some clinical or pathological criterion. There are 100 patients in each group.

The group distinction is not derived from the expression data

Problem: Can we reconstruct the group assignments from the expression profiles?

Consider a single gene first

$a_1, \ldots, a_{100}$ expression levels in group a

$b_1, \ldots, b_{100}$ expression levels in group b

$$\overline{a} = \frac{1}{100}\left(a_1 + \ldots + a_{100}\right)$$

$$\overline{b} = \frac{1}{100}\left(b_1 + \ldots + b_{100}\right)$$

c expression level of a patient
      with unknown diagnosis

Compare $|c - \overline{a}|$ and $|c - \overline{b}|$

Diagnosis: a if $|c - \overline{a}| < |c - \overline{b}|$

b if $|c - \overline{a}| \geq |c - \overline{b}|$

Both groups are summarized by the mean gene expression in this

Diagnosis is according to the closest mean

Consider two genes:

$a_{1,1}, \ldots, a_{1,100}, \; a_{2,1}, \ldots, a_{2,100}$ group a

$b_{1,1}, \ldots, b_{1,100}, \; b_{2,1}, \ldots, b_{2,100}$ group b

$\bar{a} = (\bar{a}_1, \bar{a}_2)$

$\bar{b} = (\bar{b}_1, \bar{b}_2)$

$c = (c_1, c_2)$ Patient without diagnosis

Compare : $d_a = (\bar{a}_1 - c_1)^2 + (\bar{a}_2 - c_2)^2$ and

$\qquad d_b = (\bar{b}_1 - c_1)^2 + (\bar{b}_2 - c_2)^2$

Diagnosia : a if $d_a < d_b$

$\qquad\qquad$ b else

Many (N) genes:

Nearest Centroid Method

(Plain Vanilla)
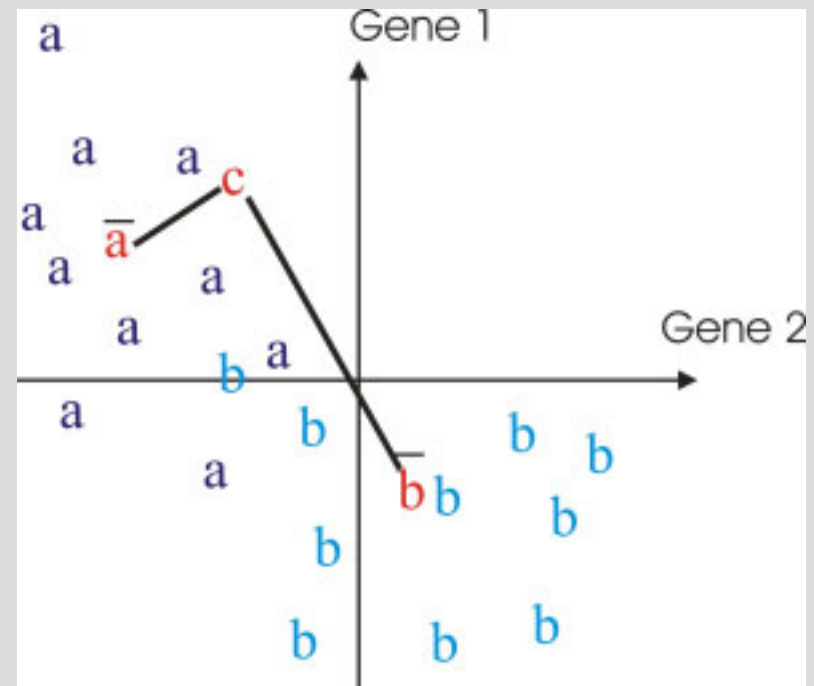
$a_{i,j}$ — Gene i in Patient j from group a

$b_{i,j}$ — Gene i in Patient j from group b

$\bar{a} = (\bar{a}_1, \ldots, \bar{a}_N)$

$\bar{b} = (\bar{b}_1, \cdots, \bar{b}_N)$

$c_1, \ldots, c_N$ — Patient without diagnosis

Compare distances to the centroids :

$$d_a = \sum_{i=1}^{N} (\bar{a}_i - c_i)^2$$

$$d_b = \sum_{i=1}^{N} (\bar{b}_i - c_i)^2$$

Diagnosis : a if $d_a < d_b$

b else

Patient groups are modelled separately by centroids

Diagnosis is according to the nearest centroid in euclidean distance

$a_{i,j}$ gene i in patient j from group a

$b_{i,j}$ gene i in patient j from group b
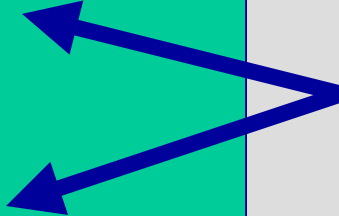
$$d_a = \sum_{i=1}^{N} (\overline{a}_i - c_i)^2$$

$$d_b = \sum_{i=1}^{N} (\overline{b}_i - c_i)^2$$

Diagnosis : a if $d_a < d_b$

            b else

All N genes contribute equally to the diagnosis ...

... that is a problem
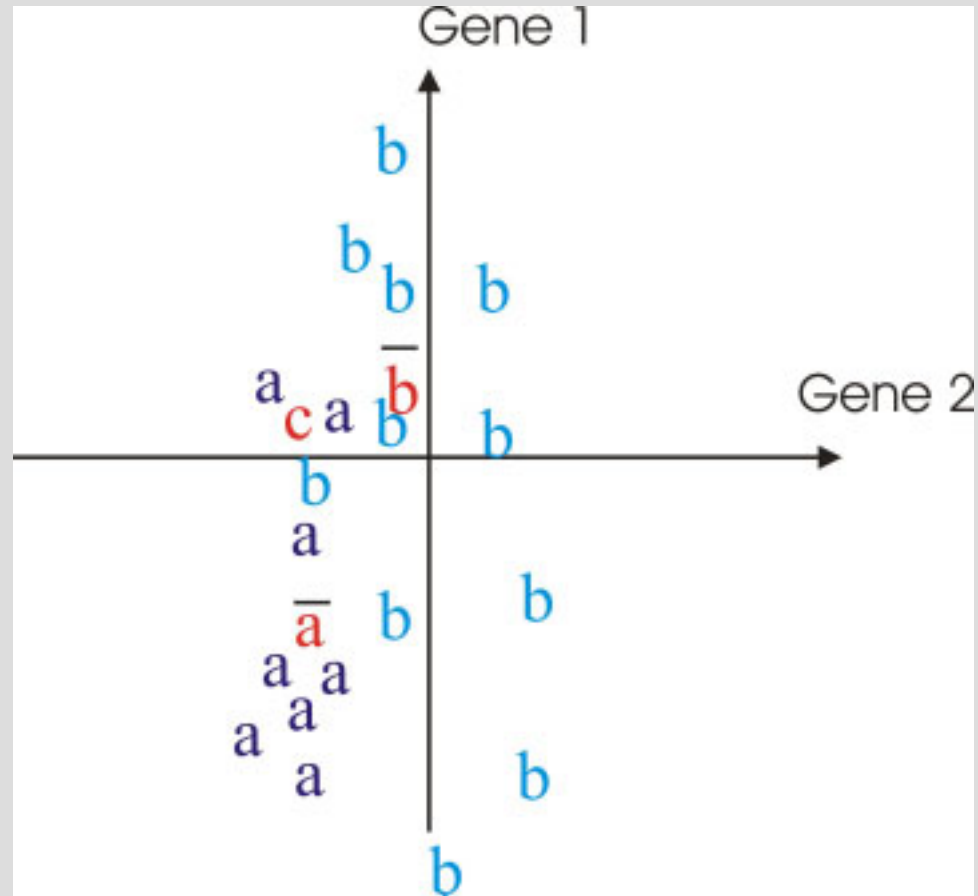
Genes with a small „variance" should get more weight than genes with high variance

$$d_a = \sum_{i=1}^{N} w_i(\overline{a}_i - c_i)^2 \qquad d_b = \sum_{i=1}^{N} w_i(\overline{b}_i - c_i)^2$$
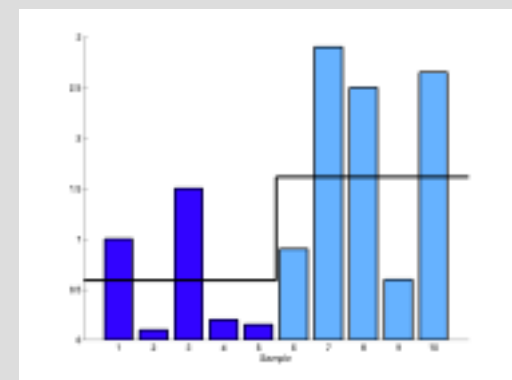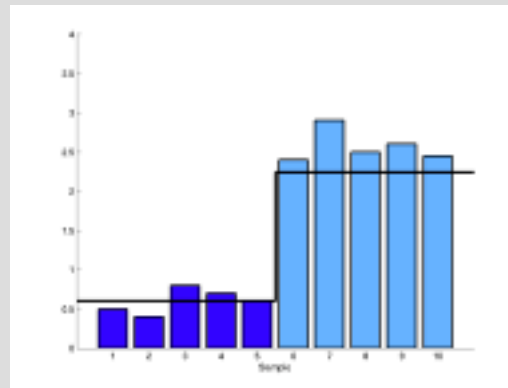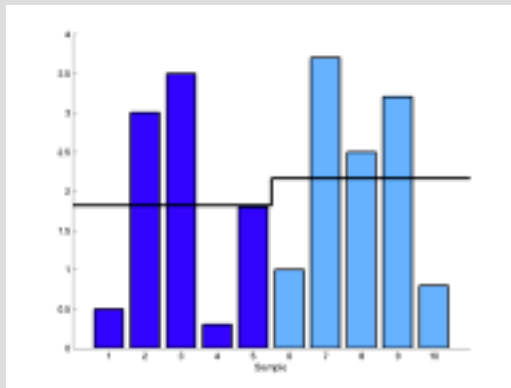
Use the pooled within class variance ... instead of the overall variance

The variances need to be estimated

$$\sigma_i^2 = \frac{1}{n-2} \sum_{j=1}^{n/2} (a_{i,j} - \bar{a}_i)^2 + (b_{i,j} - \bar{b}_i)^2$$

pooled in class variance

In our case :

$n = 200$

→ **SAM**

$$w_i = \frac{1}{(\sigma_i + \sigma_0)^2}$$

$$\sigma_0^2 = \text{median}(\sigma_1^2, \ldots, \sigma_N^2)$$

The estimated variance is not the true variance. It can be higher or lower. If a small variance is underestimated σ_i^2

can be very small and w_i

is unnaturally high.

While this is a rare event for a fixed gene it happens quite often if we are looking for 30000 genes

Is c an a or a b?
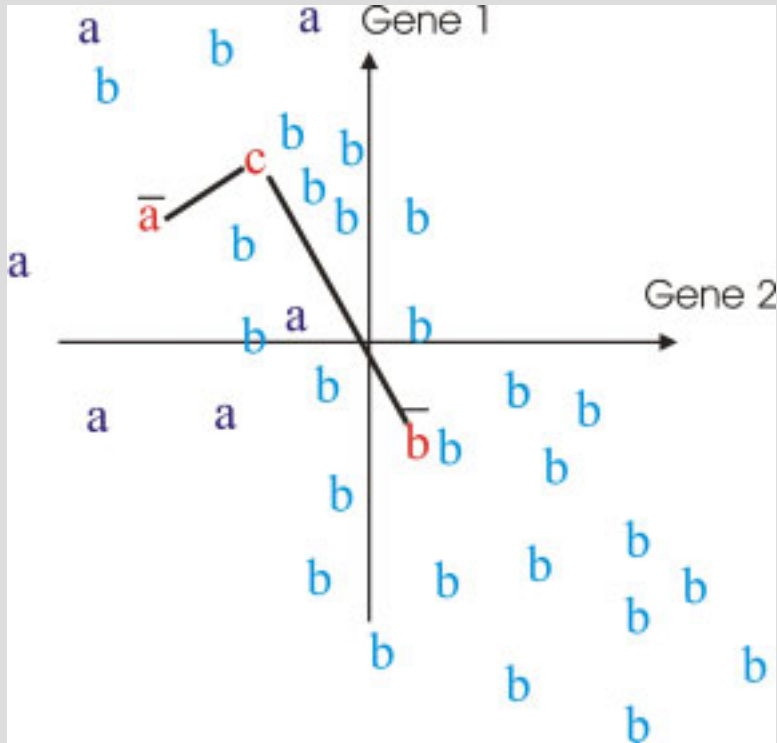
Is closer to the a centroid but there much more b than a samples

If this reflects the true population, than c should be classified as b

Baseline correction

π_a = relative size of group a

    i.e. relative frequency of type a
samples in the study, or expert
knowledge

$\pi_b = 1 - \pi_a$

$$d_a(c) = \sum_{i=1}^{N} \frac{(\overline{a}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2\log \pi_a$$

$$d_b(c) = \sum_{i=1}^{N} \frac{(\overline{b}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2\log \pi_b$$

Discriminant Score

distance to the centroid

baseline correction

$$d_a(c) = \sum_{i=1}^{N} \frac{(\overline{a}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2\log \pi_a$$

$$d_b(c) = \sum_{i=1}^{N} \frac{(\overline{b}_i - c_i)^2}{(\sigma_i + \sigma_0)^2} - 2\log \pi_b$$

pooled within class variance

variance regularization parameter

Classification probabilities
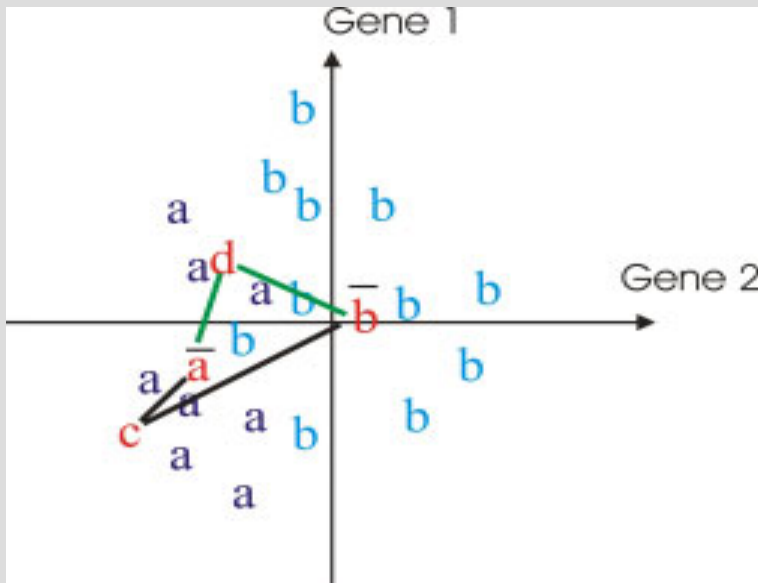
Both c and d are diagnosed as group a

But for d that was a close decision

$$\text{Prob}\left[Group(c) = a\right] = \frac{e^{-\frac{1}{2}d_a(c)}}{e^{-\frac{1}{2}d_a(c)} + e^{-\frac{1}{2}d_b(c)}},$$

$$\text{Prob}\left[Group(c) = b\right] = 1 - \text{Prob}\left[Group(c) = a\right]$$

Putting things into context

$$d_a(c) = d_b(c)$$ is a linear plane

We are still using all the 30000 genes

→ Overfitting problem

The plane is not necessarily optimal in terms of separation

This might be an advantage or a disadvantage

There is already some regularization going on

Variable selection

30000 genes are to many

They may cause overfitting

They introduce noise ... there weights are low ... but if there are many ...

They can not all matter

→ Choose genes:
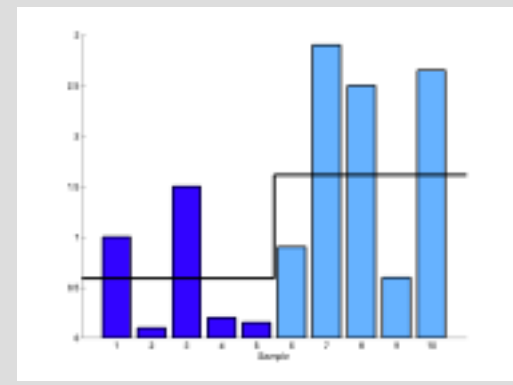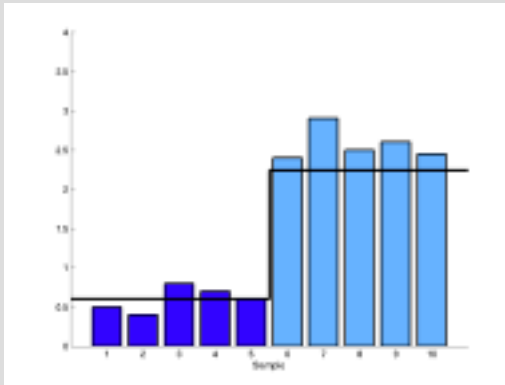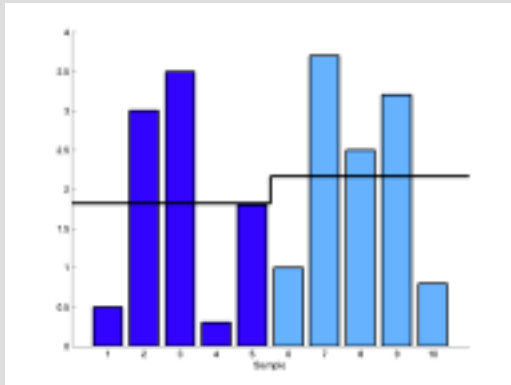
Choose the genes with the highest weights

regularized t-score a la SAM

Hard thresholding vs. soft tresholding

Lets say we pick the top 100 genes

Gene Nr. 100 is in but gene Nr. 101 is not,

however, both genes are almost equally informative

If you want to get rid of genes you can chop them off or slowly push them out
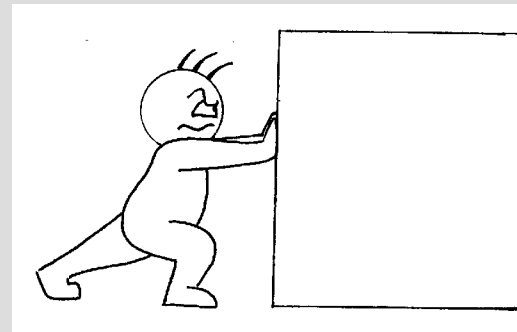
The shrunken centroid method and the PAM program

Tibshirani et al 2002
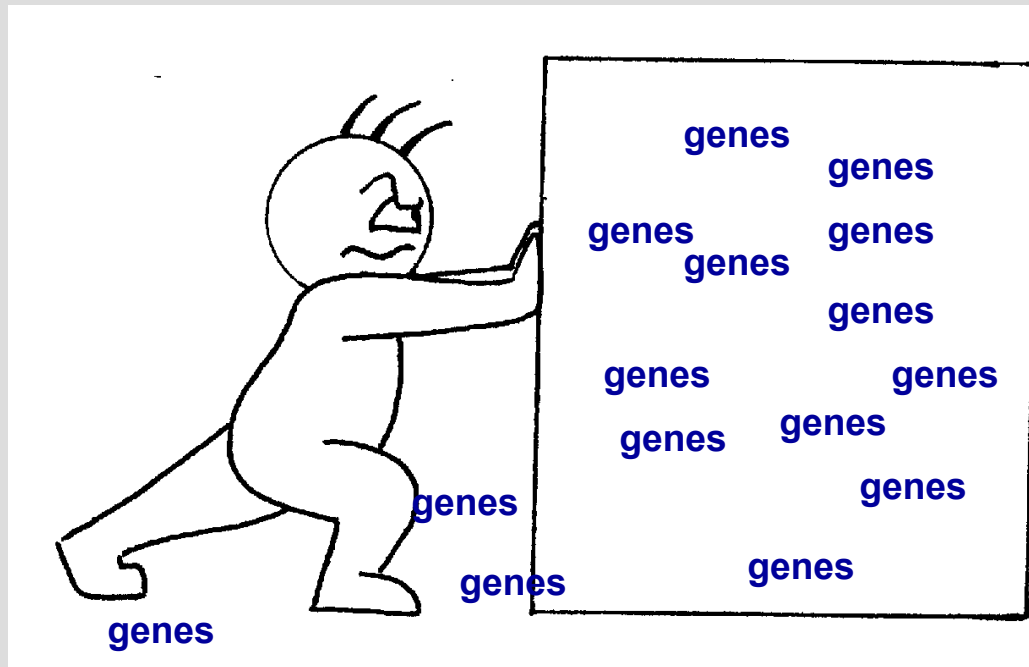
Idea

Genes with high weights are influential for diagnosis

Genes with lower weights are less influential for diagnosis

Genes that are excluded can not be influential for diagnosis at all

Before you exclude a gene totally from analysis make

it continously less influential for the diagnosis

How? By centroid shrinkage!

genes genes
genes genes
genes
genes
genes genes
genes genes
genes
genes
genes genes
genes
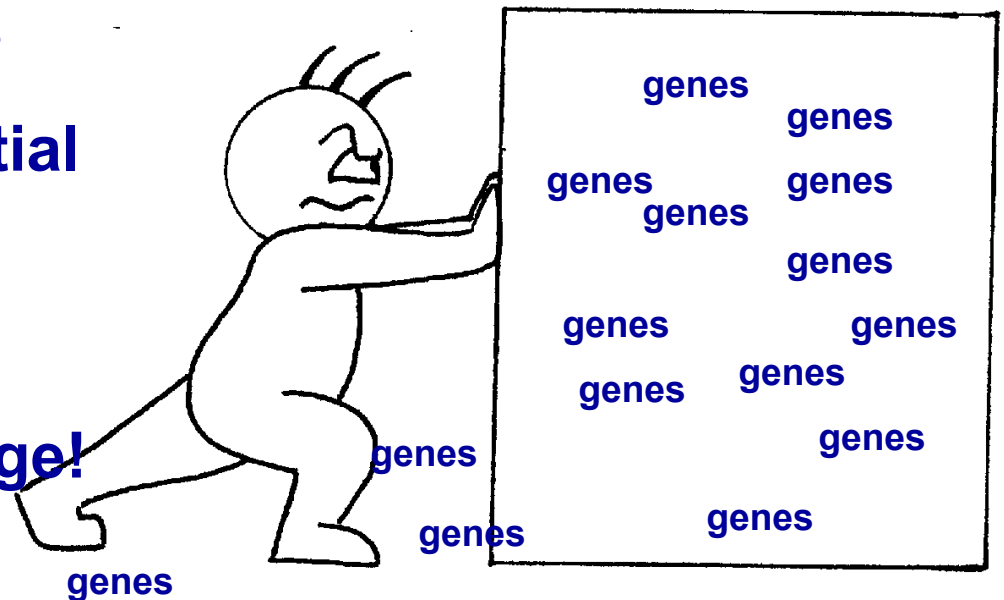
Centroid shrinkage
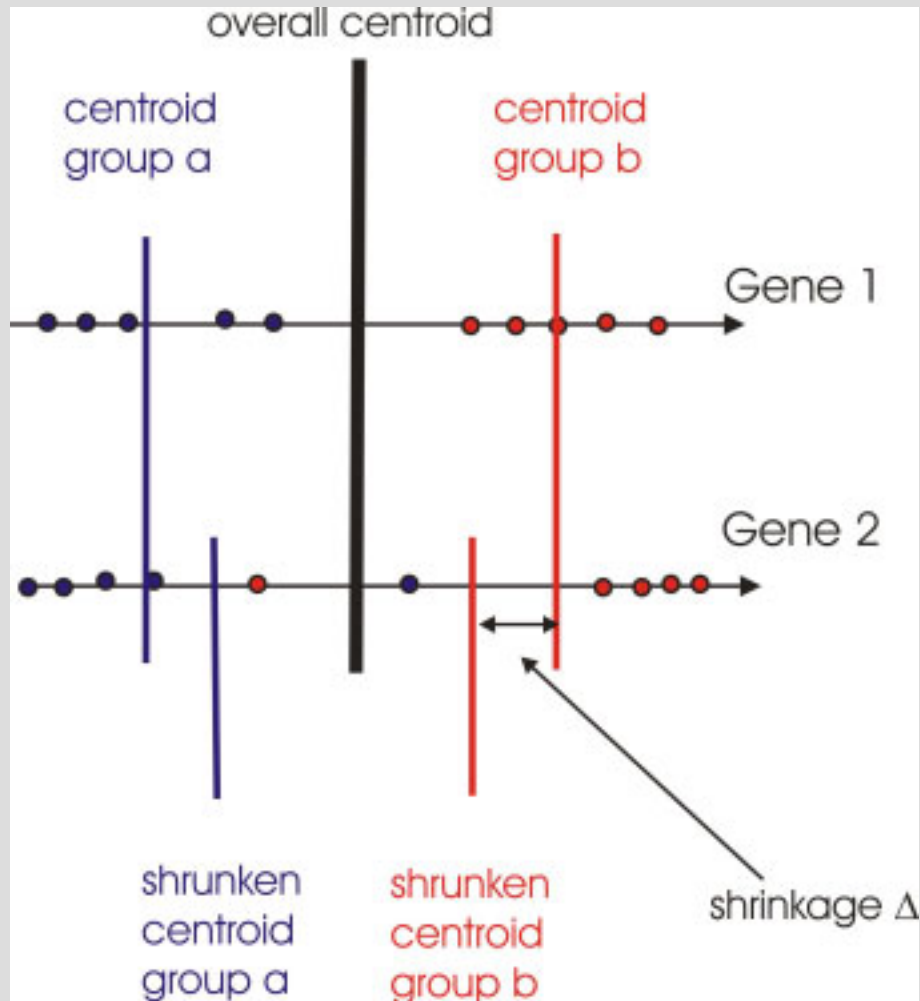
Notation

$\overline{a}_i$ mean of gene i in group a

$\overline{b}_i$ mean of gene i in group b

$\overline{x}_i$ mean of gene i using all data

Let

$$D_{i,a} = \frac{\overline{a}_i - \overline{x}_i}{m_a(\sigma_i + \sigma_0)}, \qquad m_a = \sqrt{1/n_a + 1/n}$$

$$D_{i,b} = \cdots$$

or

$$\overline{a}_i = \overline{x}_i + m_a(\sigma_i + \sigma_0)D_{i,a}$$

$$\overline{b}_i = \cdots$$

group centroid　　　　**overall centroid**

scaling factor

$$\bar{a}_i = \bar{x}_i + m_a (\sigma_i + \sigma_0) \, D_{i,a}$$

offset

$$\bar{a}_i = \bar{x}_i + m_a (\sigma_i + \sigma_0) \, D'_{i,a}$$

shrunken offset

$$D'_{i,a} = sign(D_{i,a}) \, (\left| D_{i,a} \right| - \Delta)_+$$

shrinkage parameter

$$(\cdots)_+ = \text{truncation at zero}$$

Ok, the same in words for those who do not like formulae

Gene by gene, we shrink the group centroids towards the overall centroids standardized by the within-class standard deviations until the group centroids fall onto the coverall centroid ... then the gene is excluded.

When a group centroid moves towards the overall centroid the corresponding gene becomes continuously less influential for diagnosis until it is finally excluded

The amount of shrinkage is controlled by Delta

Little shrinkage many genes are still contributing to the centroids

High shrinkage only few genes are still in the analysis

The amount of shrinkage can be determined by

cross validation … we will discuss this later

Estrogen Receptor Status

- **7000 genes**
- **49 breast tumors**
- **25 ER+**
- **24 ER-**
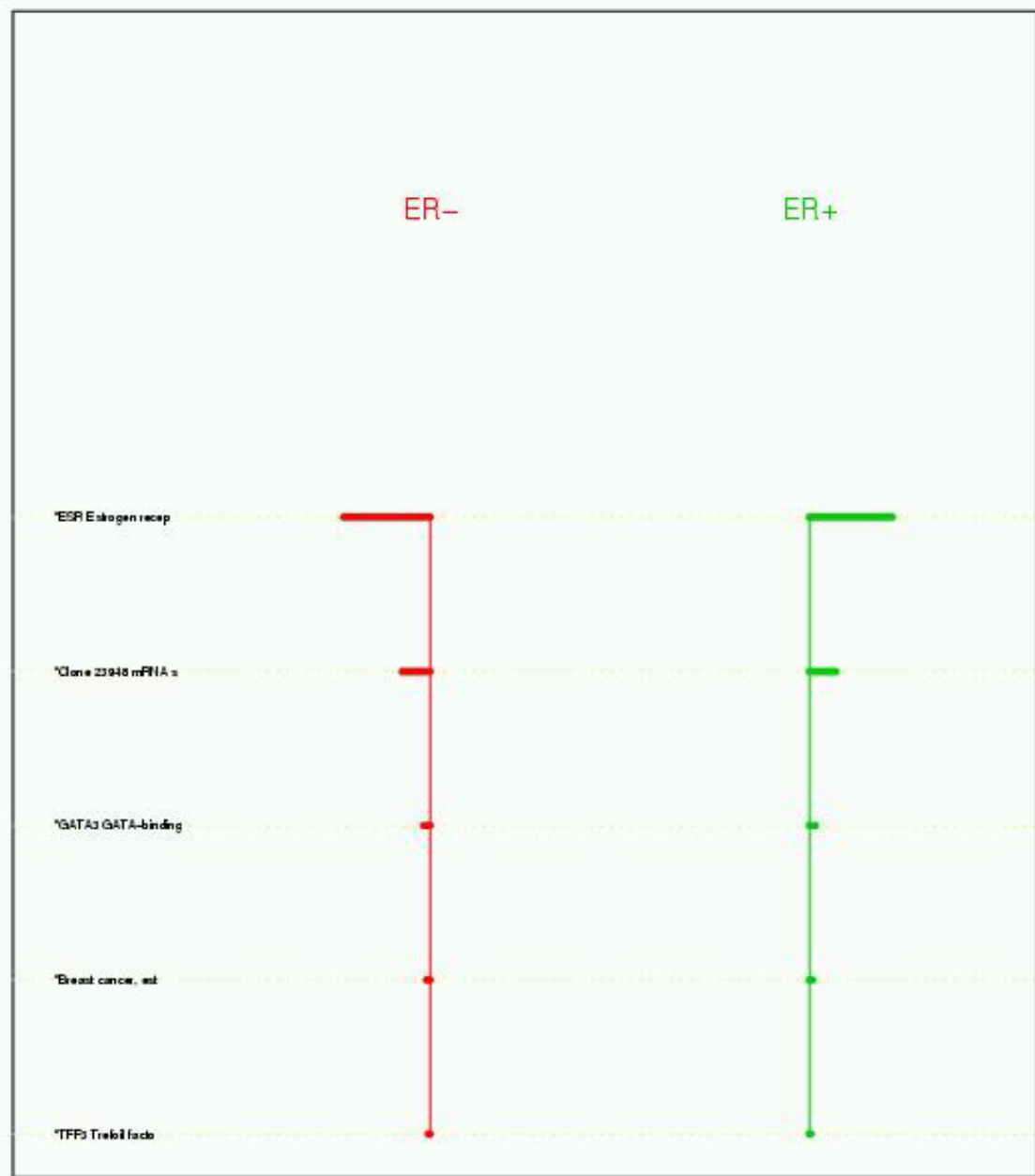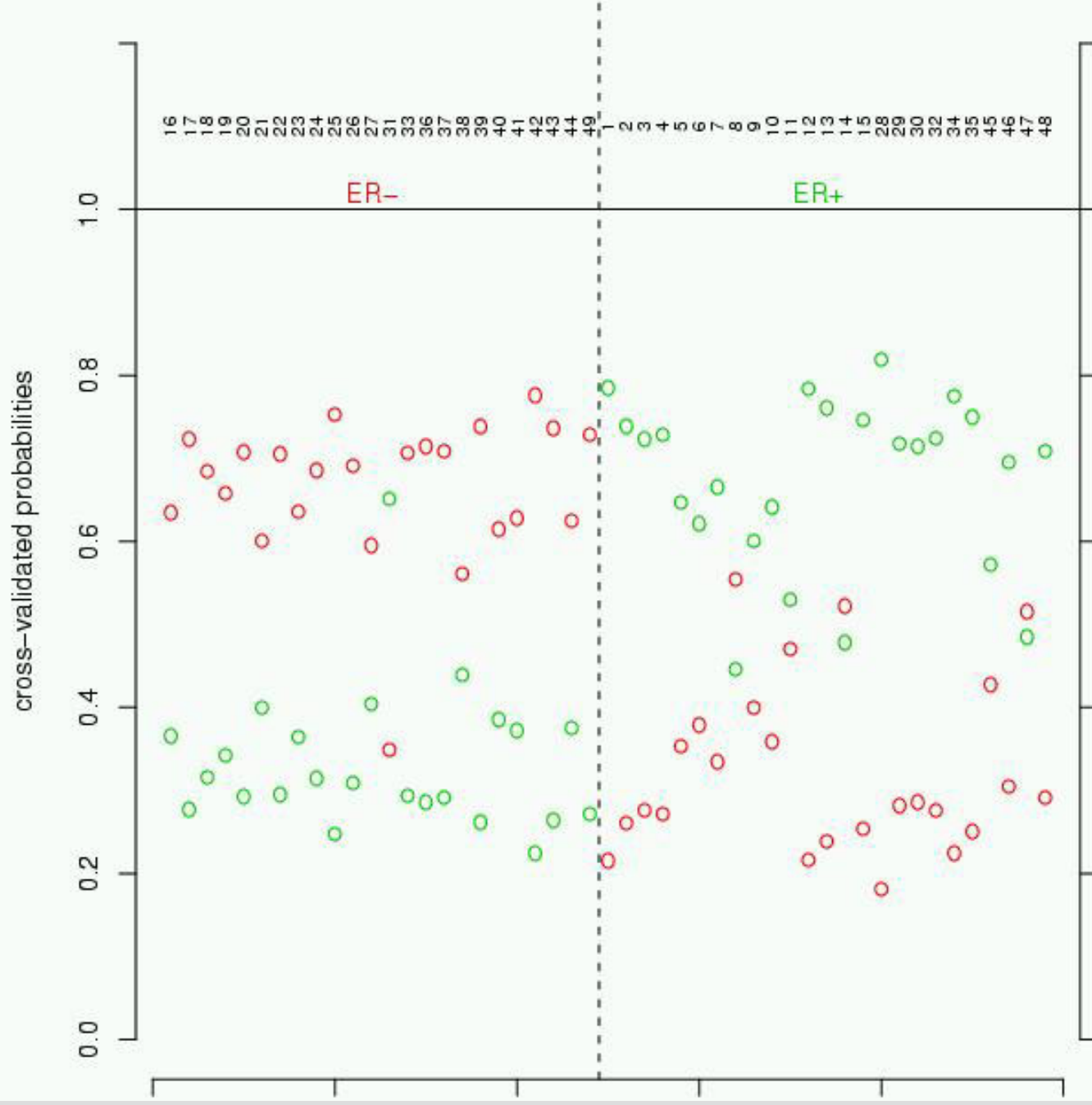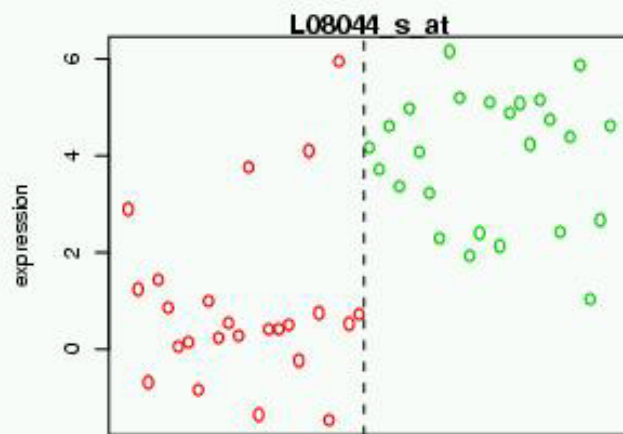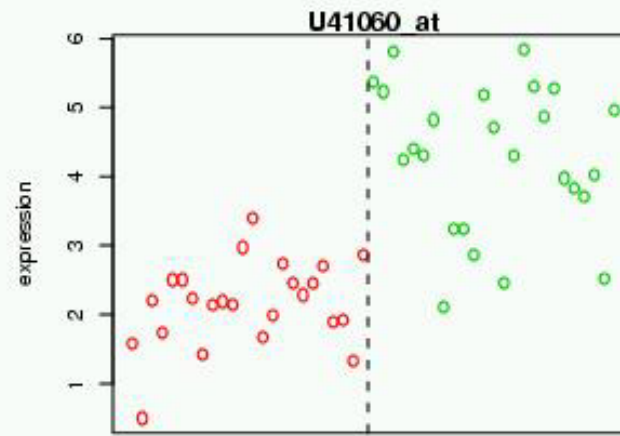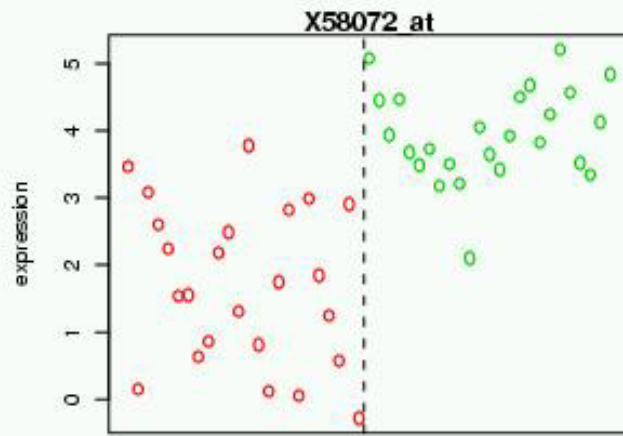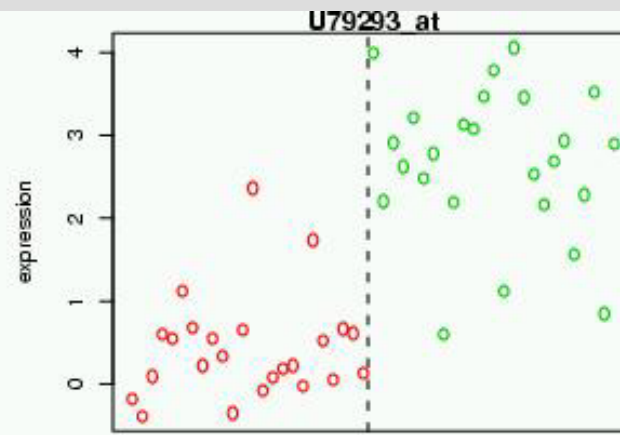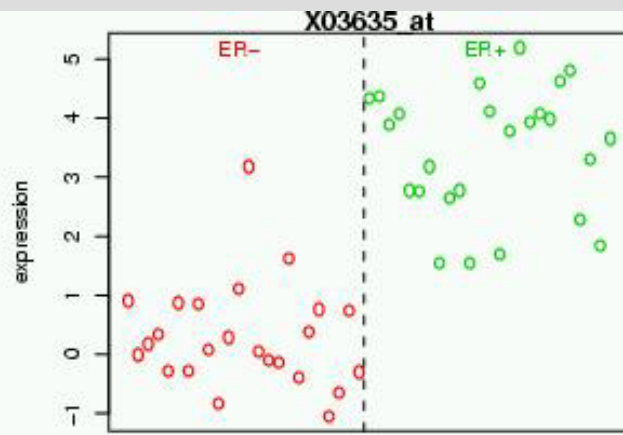
Imagine we have a study with 30000 genes 29998 of them with no biological significance and the 2 below
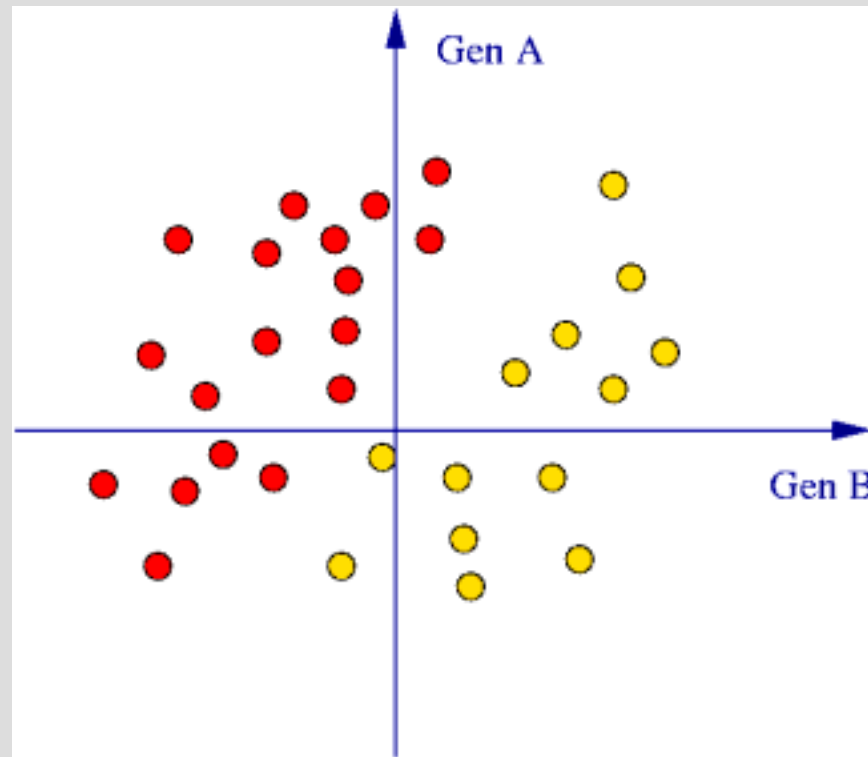
What would PAM do?

What would PAM do?
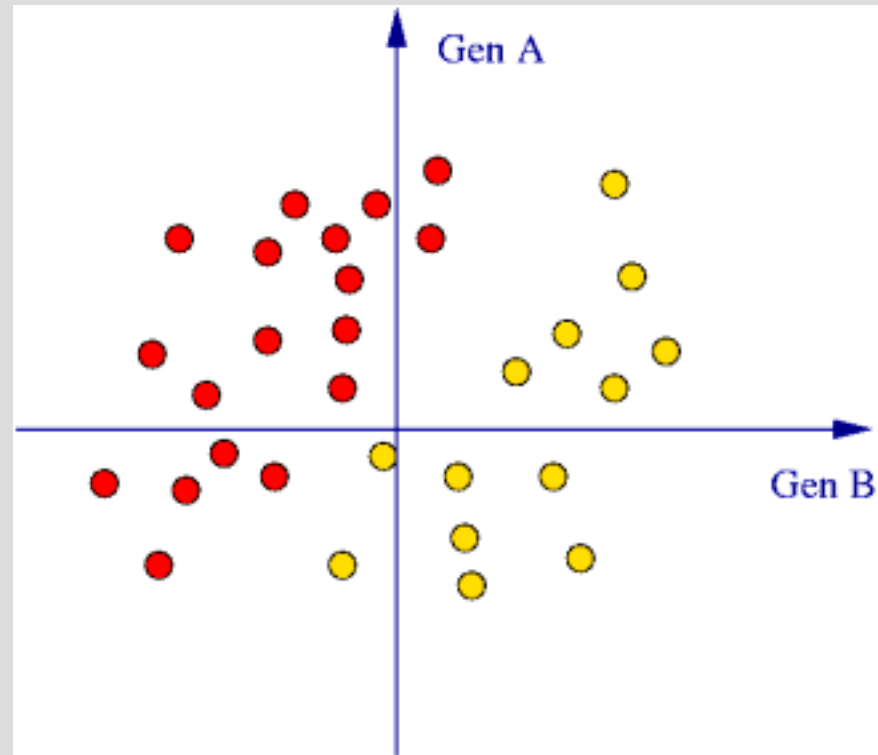
Fail

Pam would not find these two genes because their group centroids are to near to the overall centroid

Each of them is a poor classifier, together they are a good one

This is both a bug and a feature of PAM

Again, there is regularization going on

PAM does not find everything, but what it finds has a good chance to be of importance

- PAM does variable selection by screening one gene after another

- The centroids are the signatures

- So when we decide whether a gene should go into a signature we only look at this single gene and decide

- Interaction of genes is unimportant for the selection

- We combine consistently up and down regulated genes into signatures

Devices of regularization used by PAM

-Gene selection

-Shrinkage

-Gene selection by screening (no wrapping)

-The weight of a gene only depends on the gene and not on its interaction with others

-Use of a baseline depending on the population size of the groups ... more information in addition to the expression data

Questions

Coffee

What did we learn so far, and what didn't we?

-The high dimensional data leads to overfitting problems

-There are meaningful signatures and those that mean nothing

-Regularization (PAM,SVM,...) helps finding meaningful signatures
...

 -... but if I have found one there is still no guarantee

-The patients in my data display differences in a signature between group a and b ... but does this apply to a new patient too ?

- Is the signature predictive? Can it be used for diagnosis?

Problems:

1. How much regularization is good?

2. If I have found a signature, how do I know whether it is meaningful and predictive or not?

Model Selection & Model Assessment
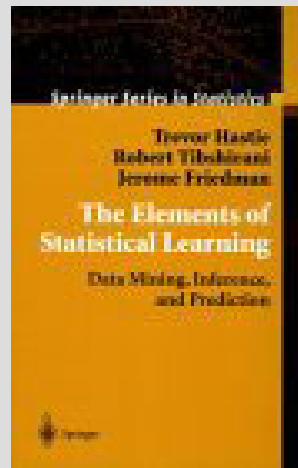
Chapter 7

Cross-Validation and Bootstrap

We only discuss Cross-Validation

Test and Training Data

150   **50**
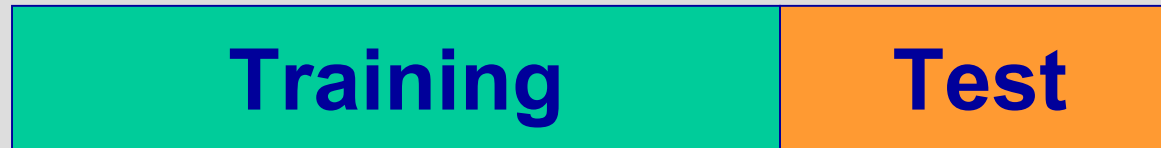
Training	Test

Split your profiles randomly into a training set and a test set

Train your model only using the data in the training set

(define centroids, calculate normal vectors for large margin separators, ...)

Apply the model to the test data ...

The setup

$x = (x_1, \ldots, x_n)$ profile

$y \in \{a, b\}$ class asignement

$g(x) = y$ true class of x

$\hat{g}(x) = \hat{y}$ predicted class of x

$\hat{p}_y(x)$ estimated probability that x is of class y

(PAM, logistic regression, logistic discrimination etc.)

For example :

$f(x)$ signature, c cutoff

$\hat{g}(x) = a$ if $f(x) > c$

$\hat{g}(x) = b$ if $f(x) \leq c$

Trainings and Test Data

Trainings data :

$$x_j^{train} = (x_{1,j}^{train}, \ldots, x_{n,j}^{train})$$ a trainings profile

$$y_j^{train}$$ its true class (used when fitting the model)

$$\hat{y}_j^{train}$$ its predicted class

$$\hat{p}_y(x_j^{train})$$ estimated probability that x_j^{train} is of class y

Test data :

$$x_j^{test} = (x_{1,j}^{test}, \ldots, x_{n,j}^{test})$$ a test profile

$$y_j^{test}$$ its true class (NOT used when fitting the model)

$$\hat{y}_j^{test}$$ its predicted class

$$\hat{p}_y(x_j^{test})$$ estimated probability that x_j^{test} is of class y

Errors & Deviances

Notation : Indicator function

$I(y = \hat{y}) = 1$ if $y = \hat{y}$

$I(y = \hat{y}) = 0$ if $y \neq \hat{y}$

Trainings Error :

$err^{train} = \sum_{\text{trainings sample}} I(y_j^{train} \neq \hat{y}_j^{train})$ Number of misclassifications in the trainings set

Trainings Deviance :

$dev^{train} = -\dfrac{2}{N^{train}} \sum_{\text{traings sample}} I(y = a) \log \hat{p}_a(x_j^{train}) + I(y = b) \log \hat{p}_b(x_j^{train})$

Test Error :

$err^{test} = ...$

Rest Deviance :

$dev^{test} = ...$

The deviance is a continuous probabilistic error measure

The bias variance trade off
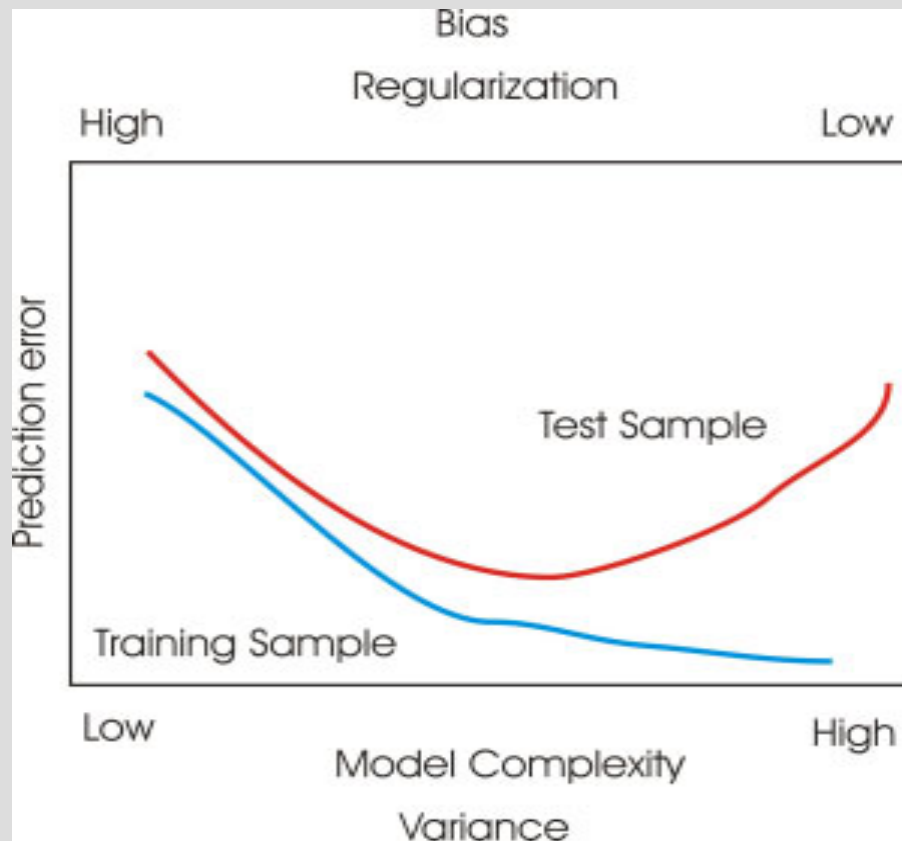
Model Complexity:

-max number of genes

-shrinkage parameter

-minimal margin

-etc

Small round blue cell tumors
4 classes
(Data: Khan et al. 2001)
(Analysis (PAM): Hastie et al 2002)
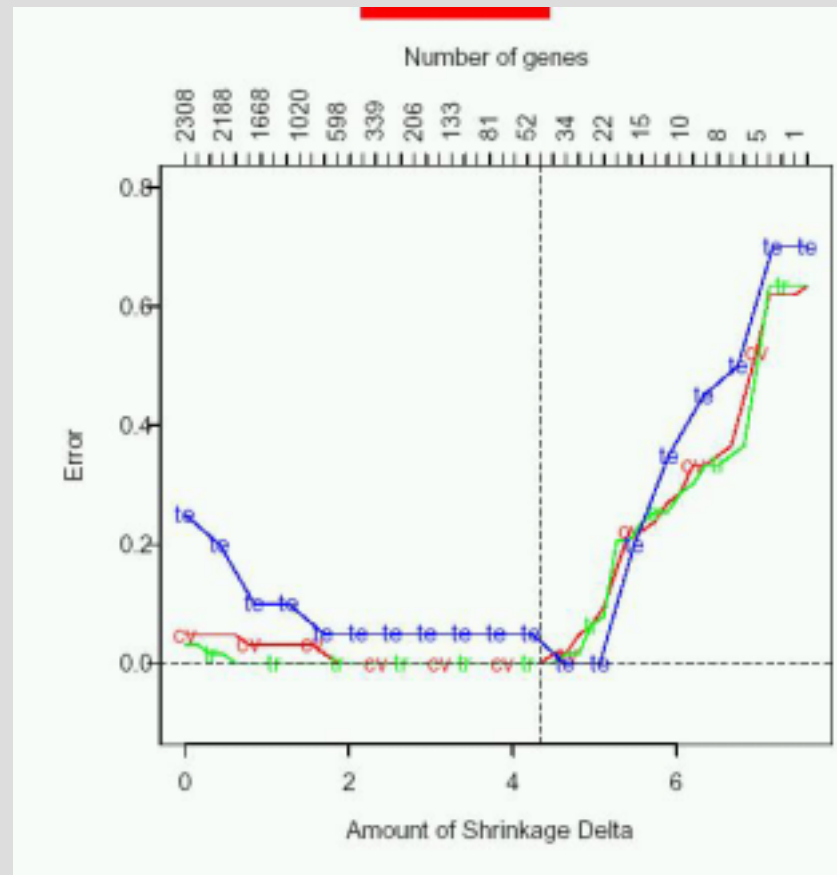
How come?

Population mean:

Genes have a certain mean expression and correlation in the population
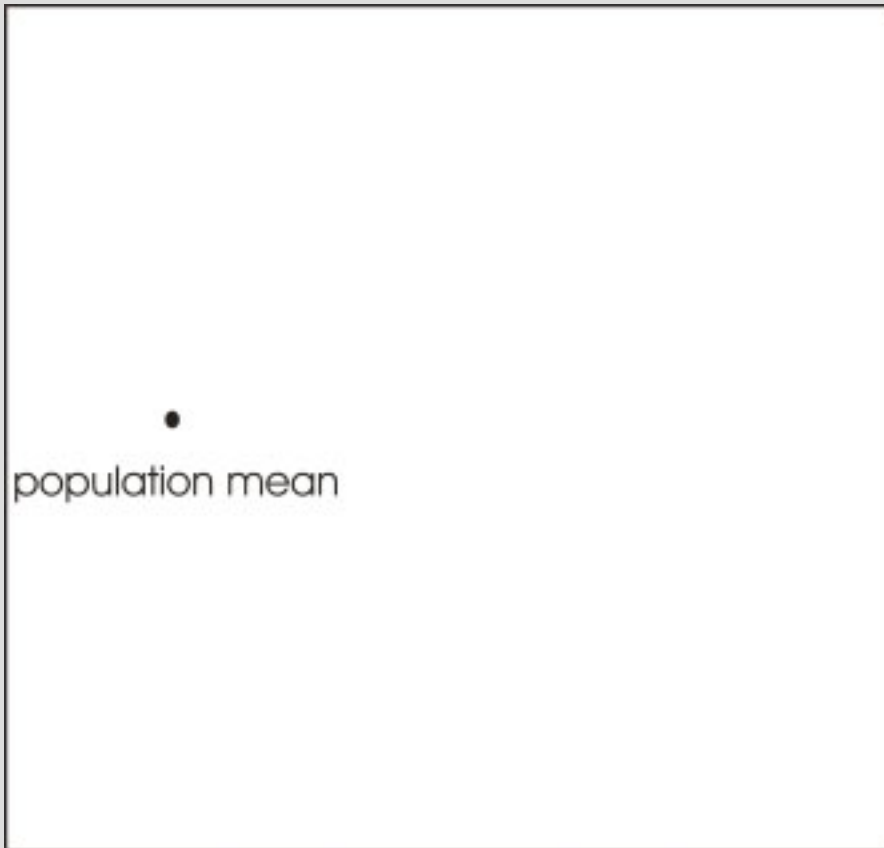
Sample mean:

We observe average expression and empirical correlation
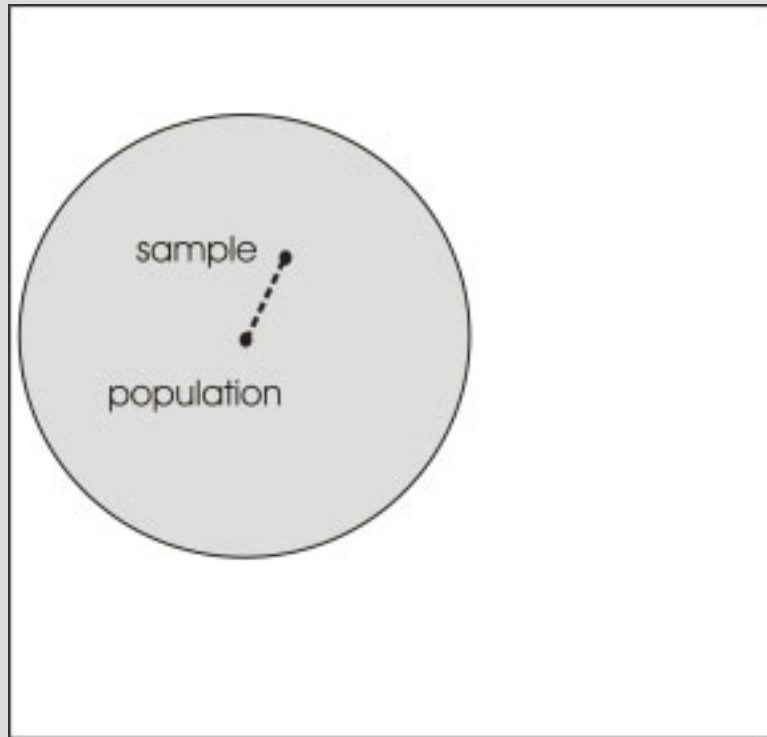
Fitted model:
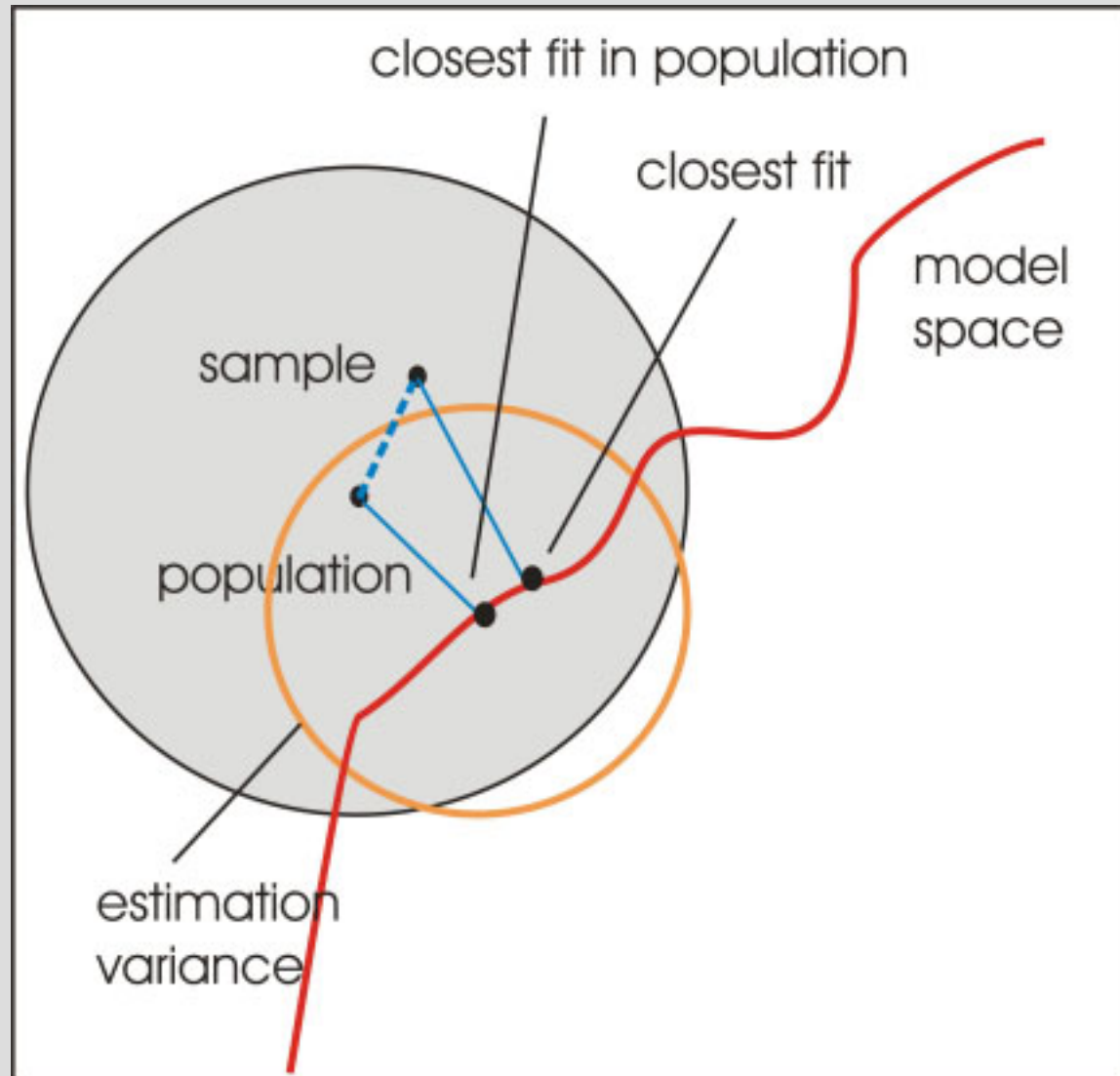
Regularization
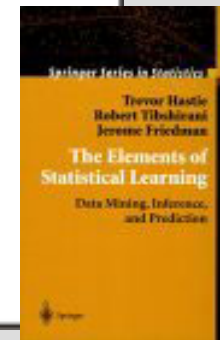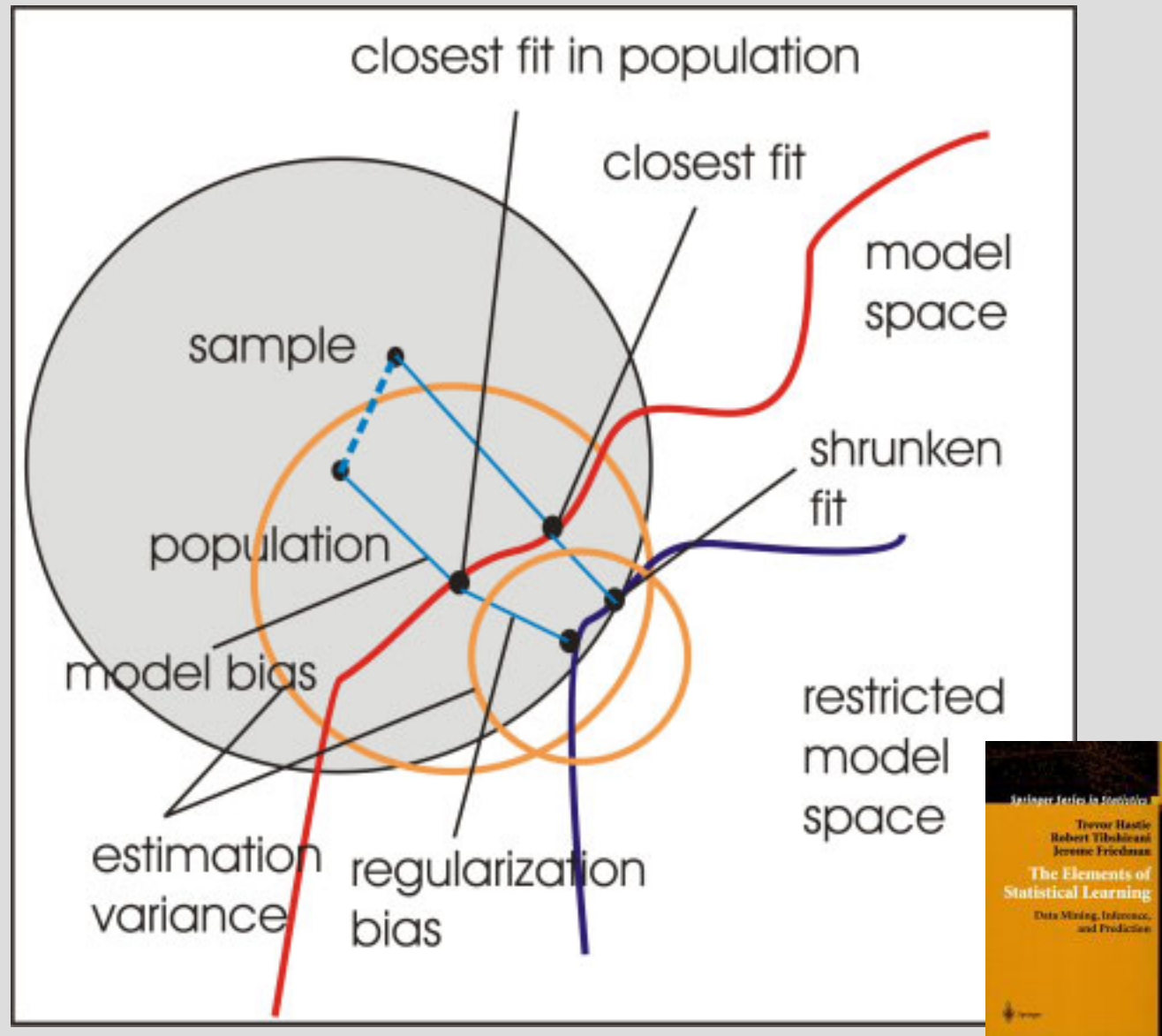
Bias-Variance-Trade-Off in PAM and in general
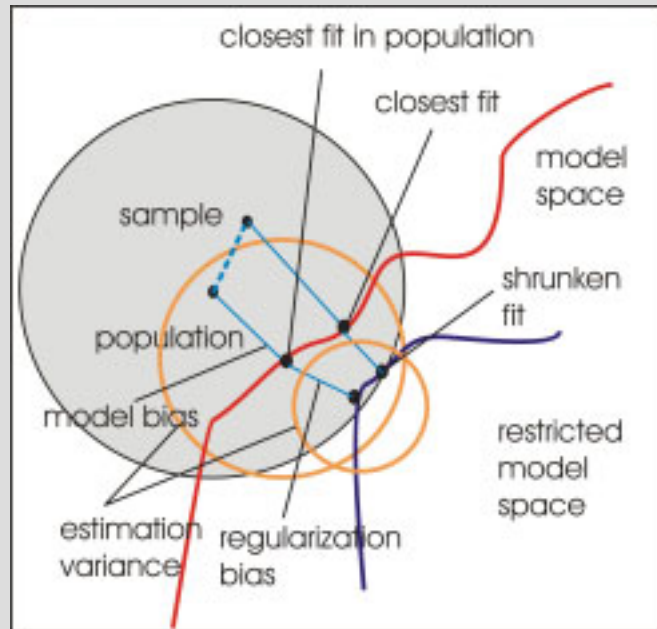
A lot of shrinkage:

Poor fit & low variance

Little shrinkage

Good fit & high variance

How much shrinkage should I use?

Model Selection with separate data

100   50

50 Training	Selection	Test

Split of some samples for Model Selection

Train the model on the training data with different choices for the regularization parameter

Apply it to the selection data and optimize this parameter (Model Selection)

Test how good you are doing on the test data (Model Assessment)

10 Fold Cross-Validation

...

Train	Train	Select	Train	Train

Train	Train	Train	Select	Train

...

Chop up the training data (don't touch the test data) into 10 sets

Train on 9 of them and predict the other

Iterate, leave every set out once

Select a model according to the prediction error (deviance)

Leave one out Cross-Validation

...

1

| Train | Train | Select | Train | Train |

| Train | Train | Train | Select | Train |

1

...

Essentially the same

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

Good for small training sets

Model Assessment

How well did I do?

Can I use my signature for clinical diagnosis?

How well will it perform?

How does it compare to traditional methods?

The most important thing:

Don't fool yourself! (... and others)

This guy (and others) thought for some time he could predict the nodal status of a breast tumor from a profile taken from the primary tumor!

... there are significant differences. But not good enough for prediction

(West et al PNAS 2001)

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.

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

(select genes, define centroids, calculate normal vectors for large margin separators,perform 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 ...

don't even think of optimizing this procedure

The selection bias

- You can not select 20 genes using all your data and then with this 20 genes split test and training data and evaluate your method.

-**There is a difference between a model that restricts signatures to depend on only 20 genes and a data set that only contains 20 genes**

-**Your model assessment will look much better than it should**

FAQ

How many patients do we need?

Do we need to replicate patient profiles?

Do we need to consult a bioinformatics expert?

When on do we need to contact him/her?

Where do we find him/her?