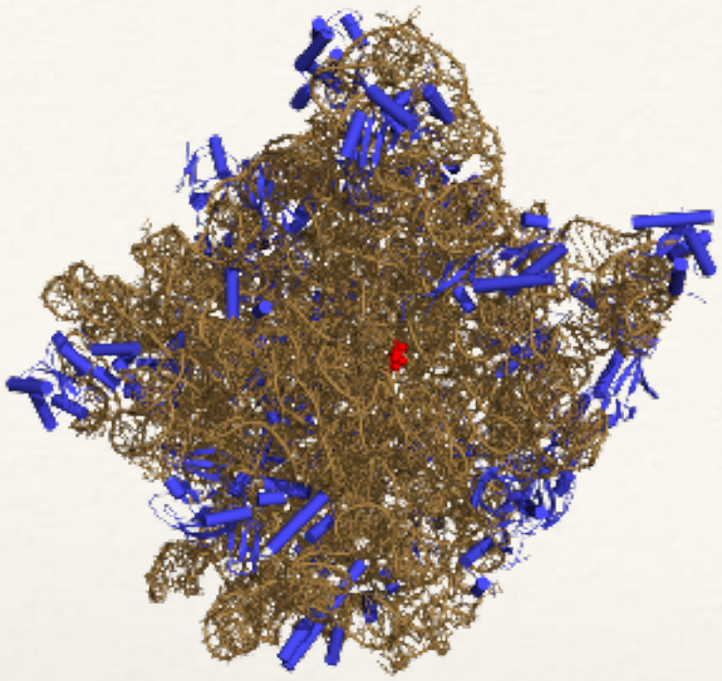
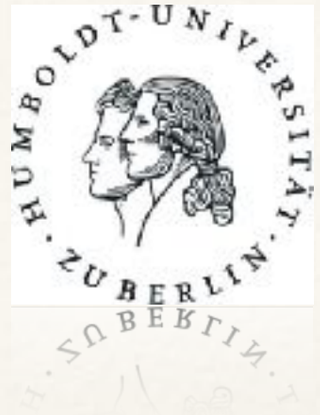


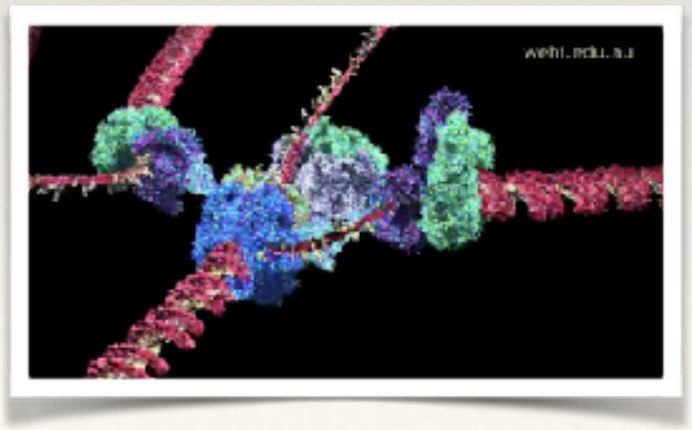


Humboldt-Universität zu Berlin

Gene Expression Analysis

Grundlagen der Bioinformatik
SS 2017

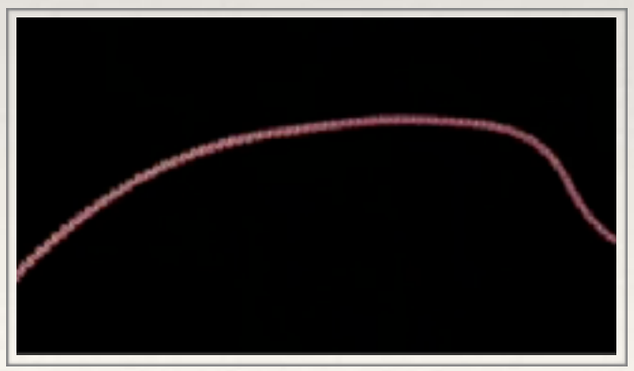
Lecture 7
16.06.2017

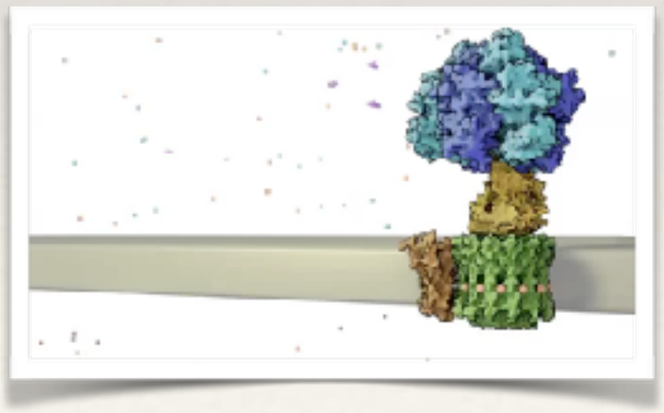


"DNA Repair" by Tom Ellenberger, Washington University School of Medicine in St. Louis. - Biomedical Beat, Cool Image Gallery. Licensed under Public Domain via Commons - https://commons.wikimedia.org/wiki/File:DNA_Repair.jpg#/media/File:DNA_Repair.jpg

Recap: Proteins & mRNA

- Cellular worker-units
- DNA -> mRNA -> Amino-acids -> Protein
- Abundance mRNA ~ Gene-activity
- Connected to phenotypes e.g. cancer





Recap: Microarrays

Structure

- Single-stranded DNA on glass-slides
- cDNA-Hybridization
- Laser-illumination

Recap: Microarrays

Structure

- Single-stranded DNA on glass-slides
- cDNA-Hybridization
- Laser-illumination

Data-Analysis

- Biological & technical errors / biases
- Discretize, visualize and correct errors and biases

Normal distribution assumption

Lecture 7

Gene Expression Analysis

Make heterogeneous data great again

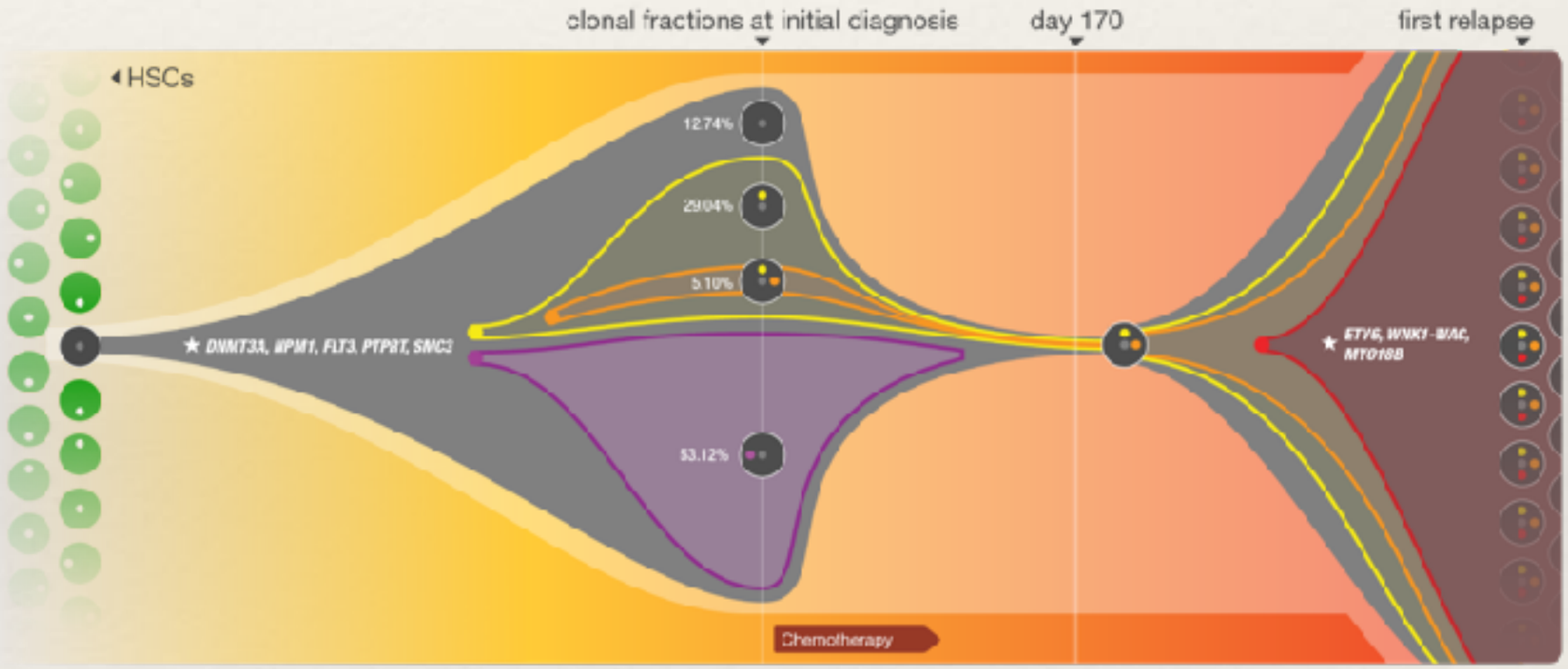
Structure

- Differential expression
  - Fold-change
  - T-test
- Clustering
- Databases

Differential gene expression - Etiology

Identify causes and evolution of e.g. cancer (etiology)

Adapt treatment



Example:
Understand development of cancer

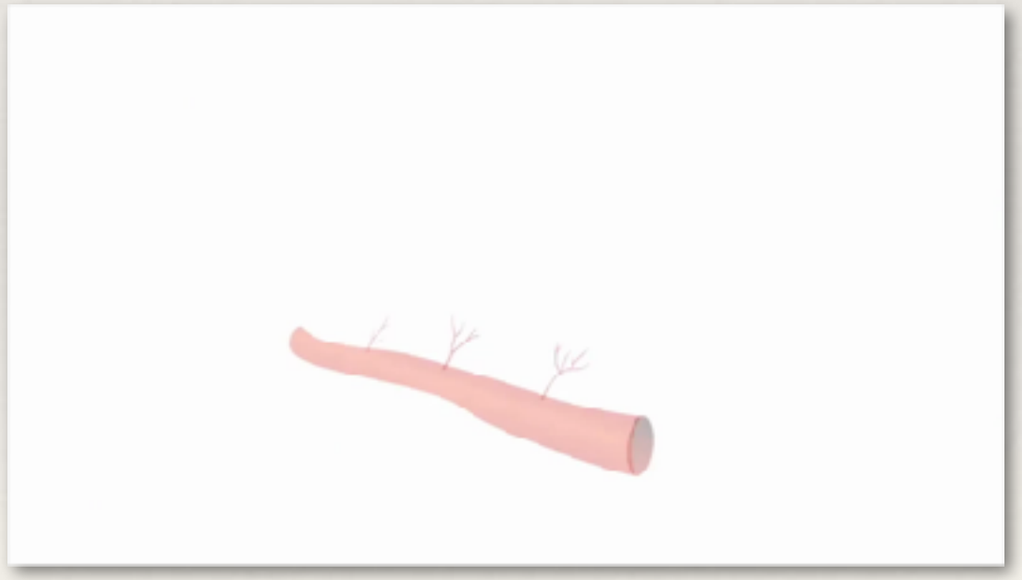
Differential gene expression - Biomarker

Ovarian cancer antigen CA125: a prospective clinical assessment of its role as a tumour marker.

P. A. Canney, M. Moore, P. M. Wilkinson, and R. D. James

Find early-presence-marker of cancer

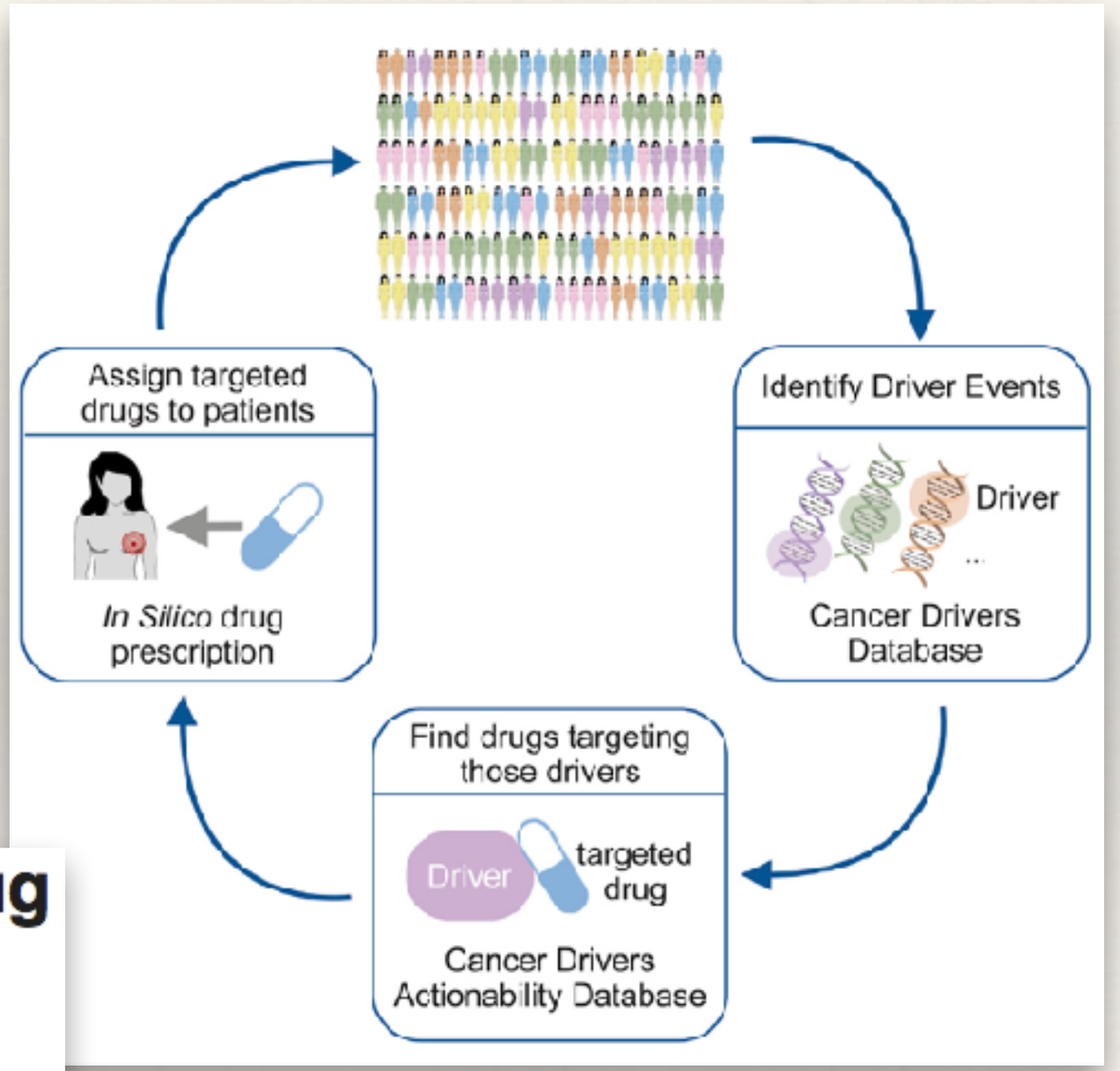
Find marker for e.g. drug-response



Example: Increasesed angiogenesis signals

Differential gene expression - Personalized medicine

- Sequence patient
- Determine similarity to known cases
- Administer best drug
  - And avoid side-effects!

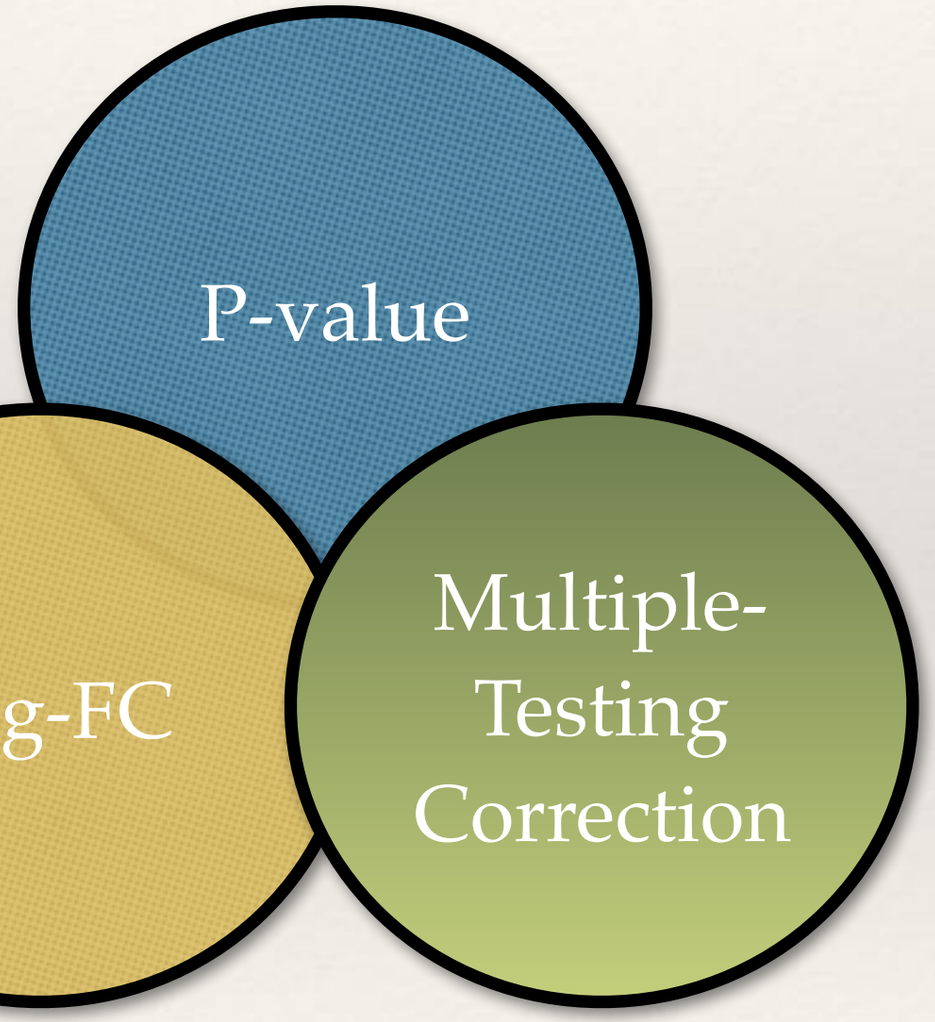


'Milestone' prostate cancer drug

By James Gallagher
Health editor, BBC News website

29 October 2015 | Health

Basic concept differential expression



Problem definition

We **have**:

$N_1,...,N_m$: normale samples

$T_1,...,T_n$: tumor samples

We **look for**: genes with significant differences between N and T

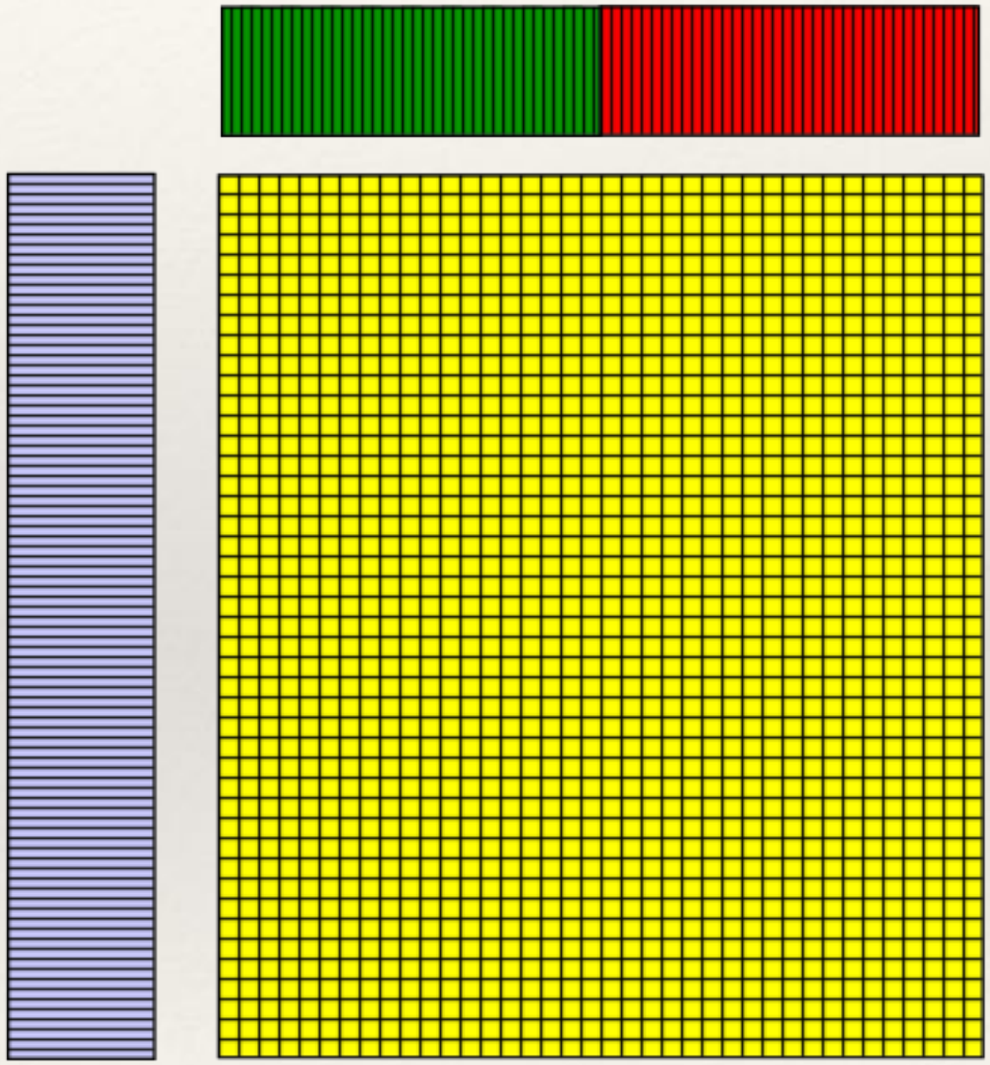
Compare values of gene X from group N with those of group T

$N = \{n_1,...,n_m\}$

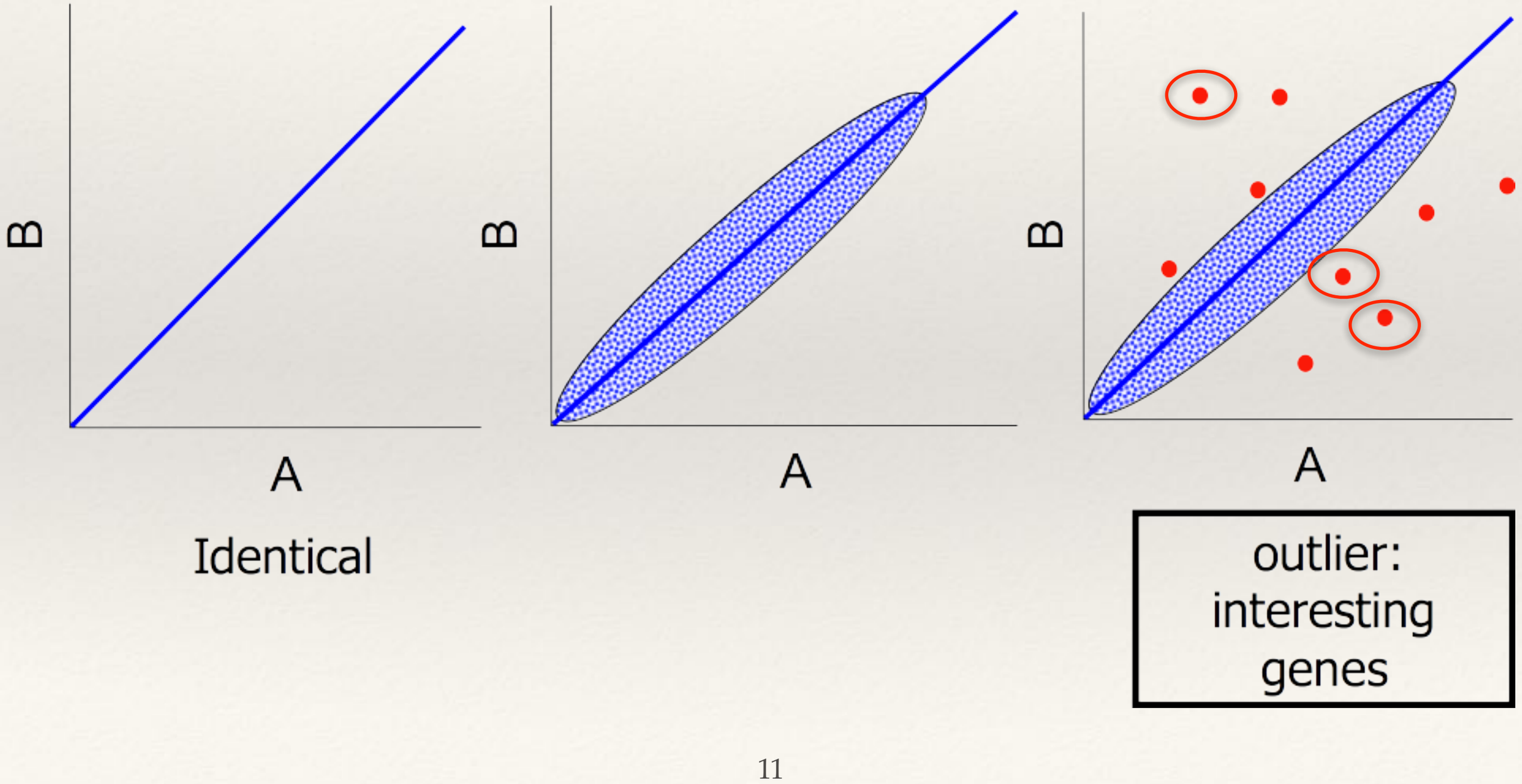
$T = \{t_1,...,t_n\}$

many methods, here:

- Fold change
- t-test



Scatterplot vs. differential expression



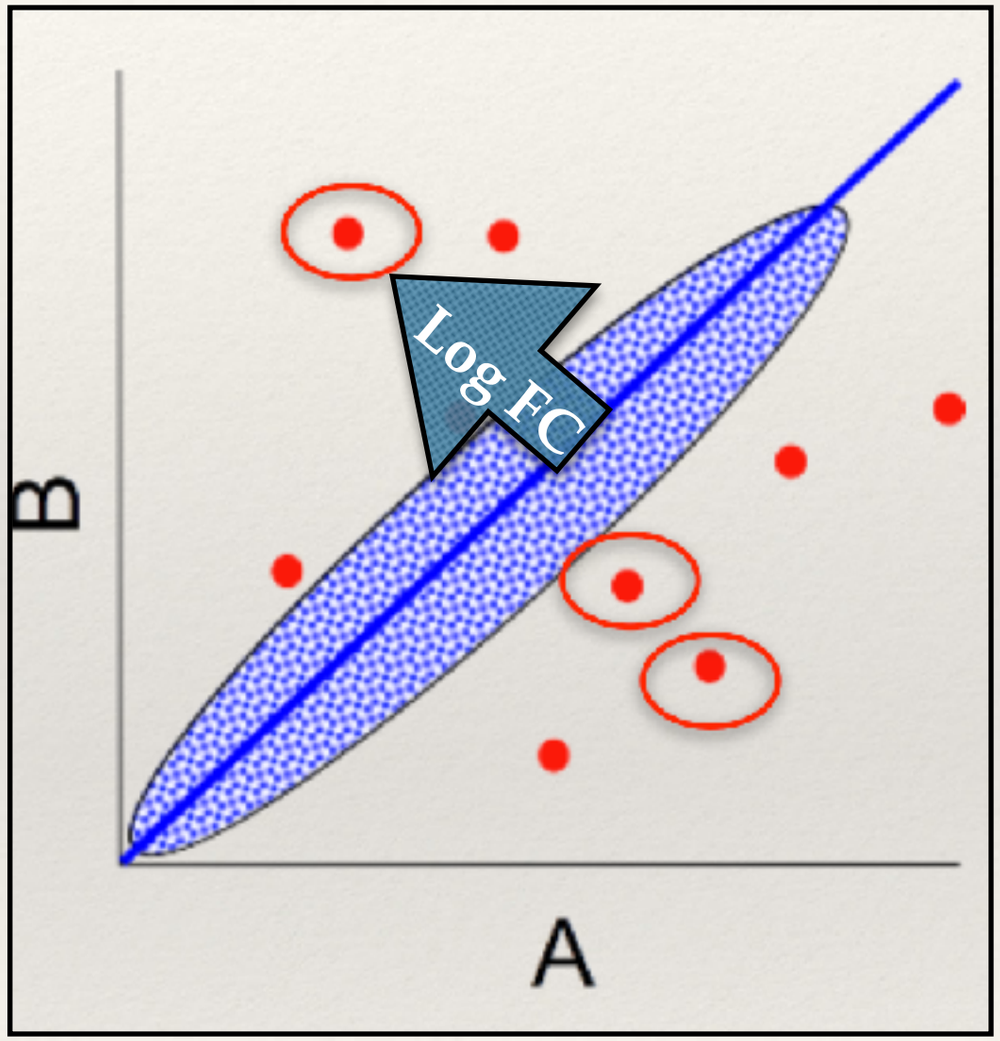
Fold-Change

$$FC = \log_2\left(\frac{mean(T)}{mean(N)}\right) = \log_2(mean(T)) - \log_2(mean(N))$$

Thresholds (common but arbitrary)

- $|FC| < 1$ not interesting
- $|FC| > 2$ very interesting

	mean(tumor)	mean(normal)	mean(t) / mean(n)	FC
gene a	16	1	16	4
gene b	0.0625	1	0.0625	-4
gene c	10	10	1	0
gene d	200	1	200	7.65



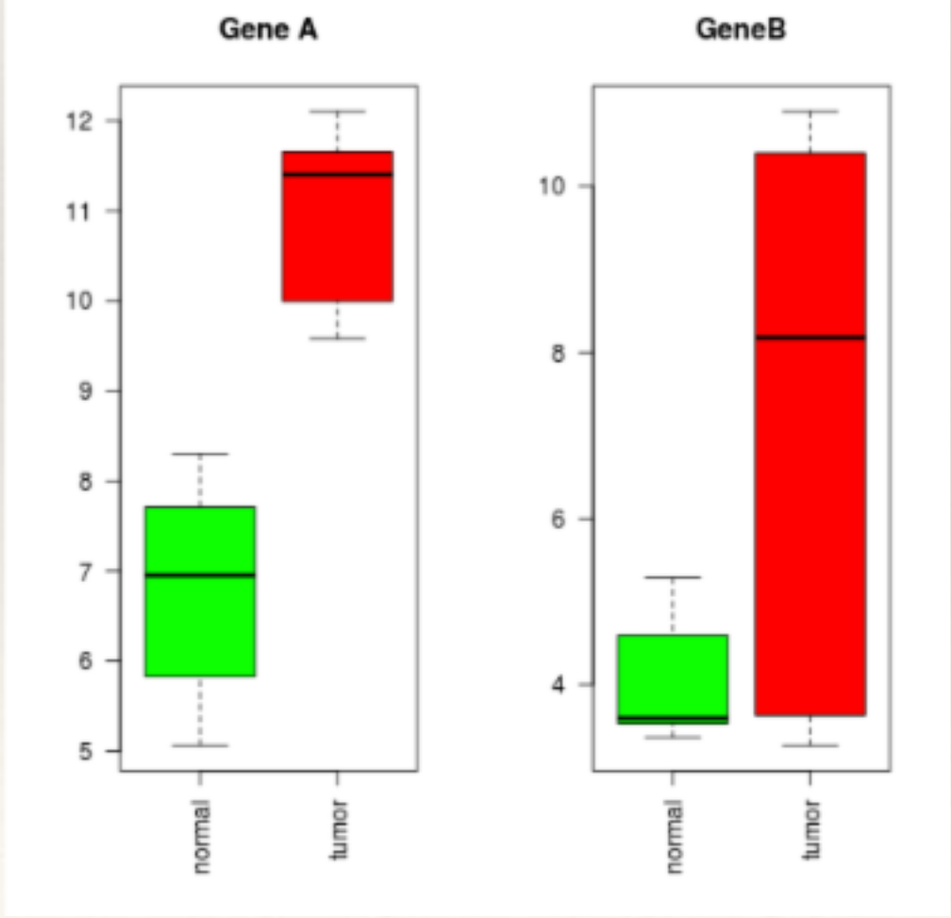
Identification differential expression

Gene expression matrix:

Gene	N1	N2	N3	N4	N5	N6	N7	T1	T2	T3	T4	T5	T6	T7	FC
A	5.06	5.22	8.3	8.03	6.95	6.43	7.39	10.1	9.89	11.7	11.6	11.4	9.58	12.1	**-4.14**
B	3.58	4.14	3.49	3.37	5.29	5.06	3.6	3.7	10.9	10.3	3.57	10.5	8.18	3.27	**-3.13**

High abs(FC) for Gene A and Gene B

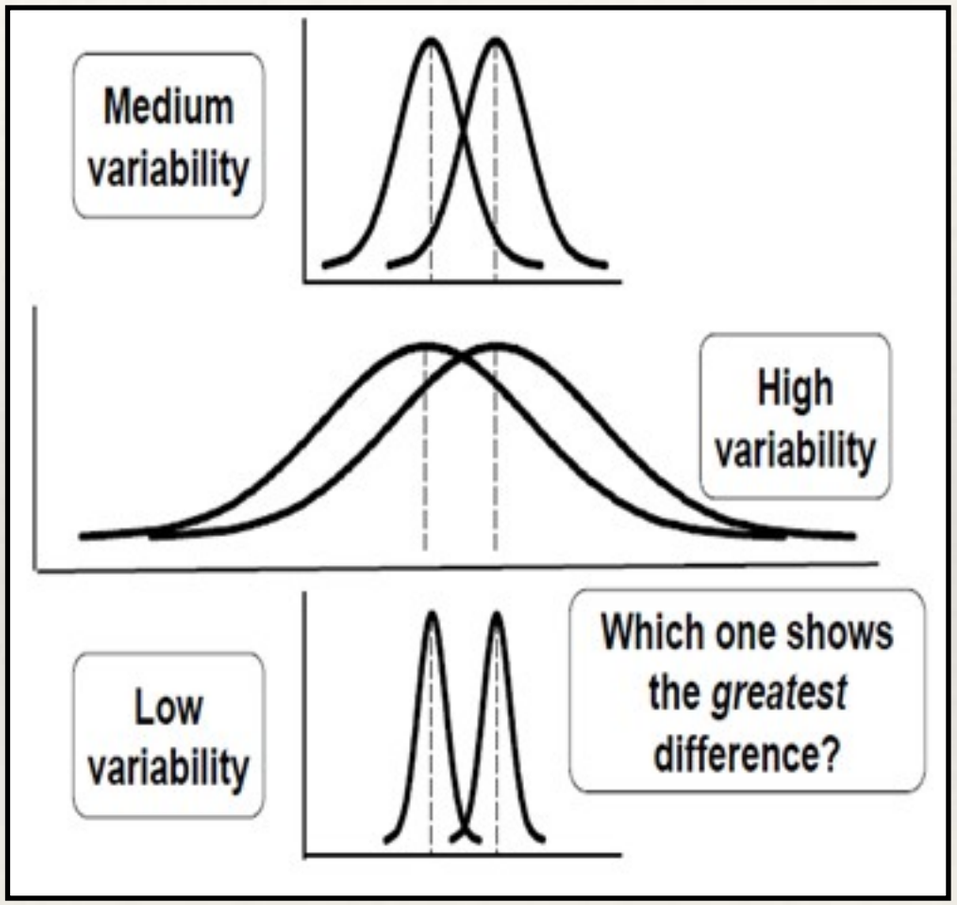
But: variance very high in the tumor samples of Gene B



Log FC differential expression

- Identify differentially expressed genes

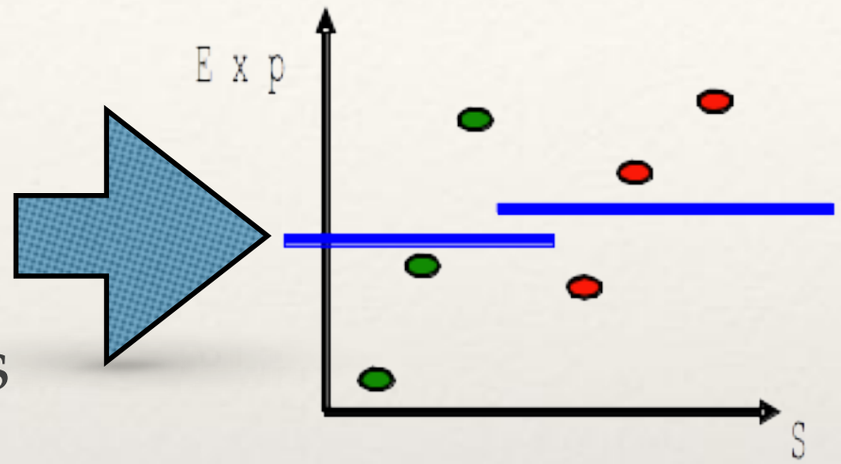
- Fold-change problematic

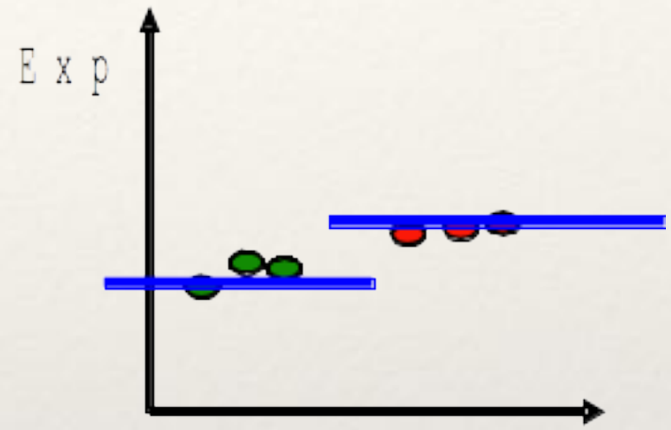


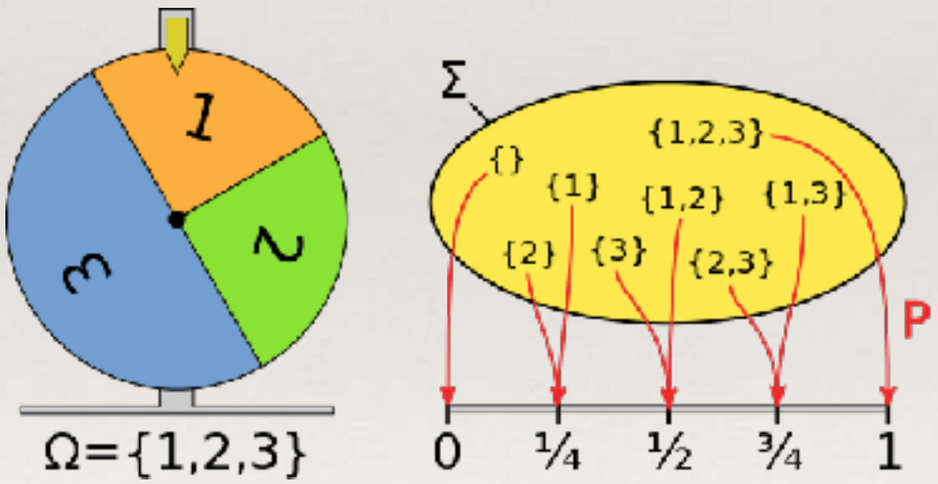
Same FC but different likelihood to be dif. exp.

Probability meassure

Meassure likelihood for truely dif. exp genes to show these distributions

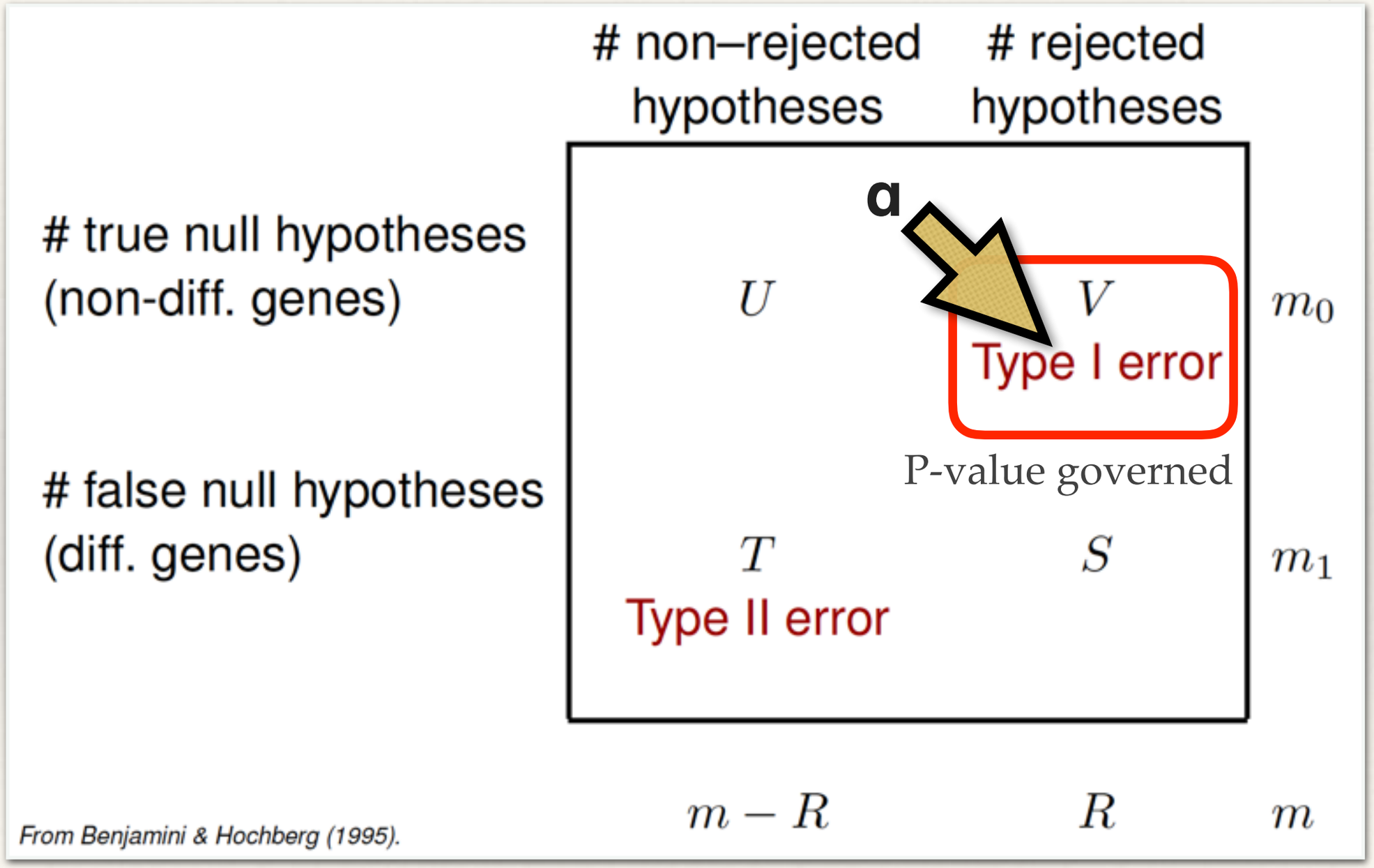






Example probability meassure

P-value & statistical error types



Alpha:= likelihood for type 1 error

Statistical Hypothesis testing

1. Formulate: null and alternative hypothesis
2. Select a significance level alpha
3. Sample population/ cohort
4. Calculate test statistic
5. P-value-based decision

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

Requires known variance + mean

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Central Limit Theorem

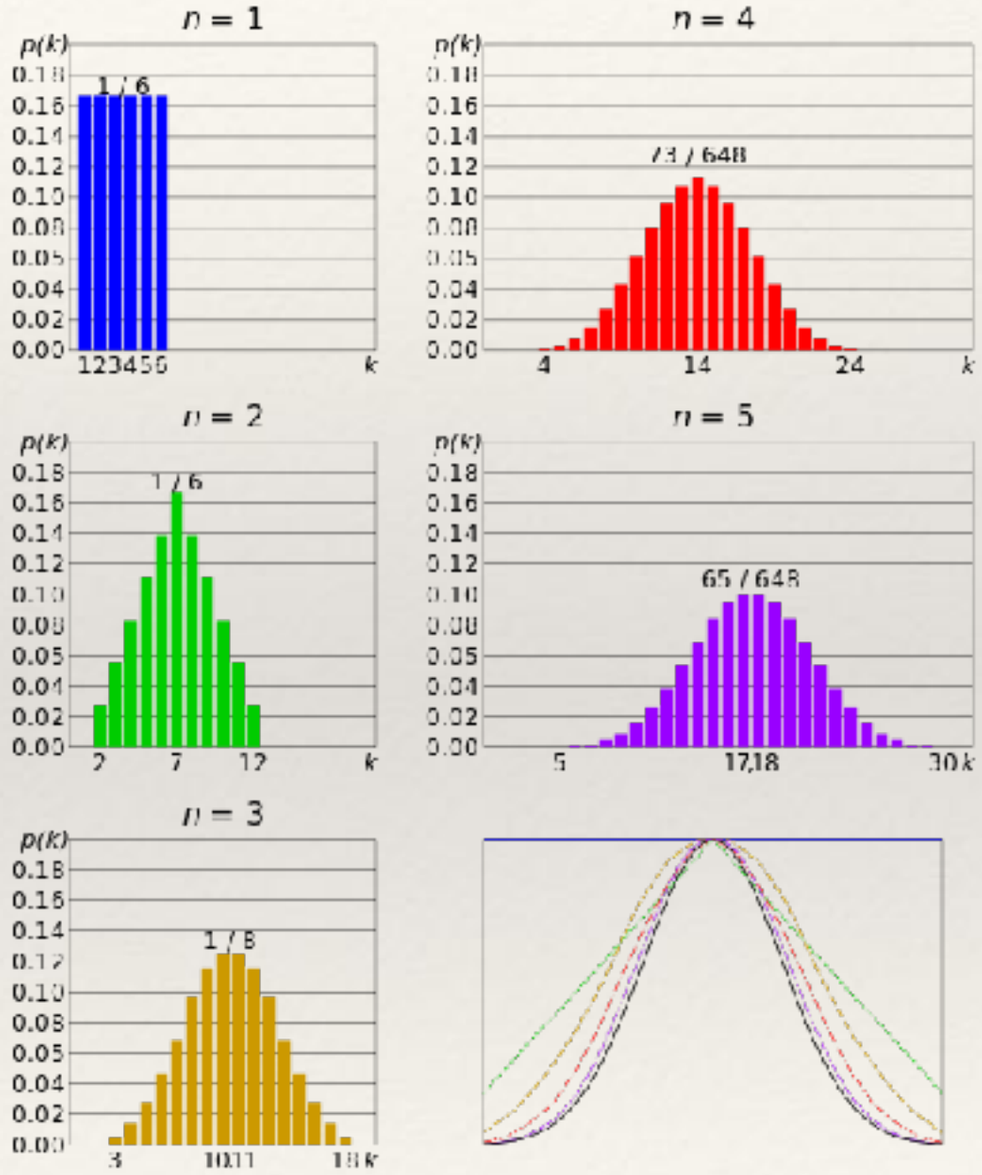
Assume normal distribution for **mean-probabilities**

(empirically expected value)

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

The probability distribution of the mean of i.i.d. random variables tends to the normal distribution

- i.i.d. = independent and identically distributed



Likelihood sum of n 6-sided dice

Student's t-test

- Compare mean & variance of cohorts
- Equal variances
- Probality to be dif. exp. follows t-probability meassure

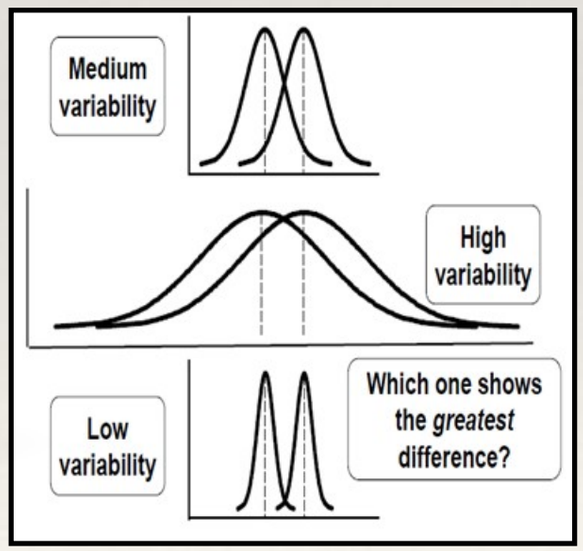
$$T = \frac{\bar{X} - \bar{Y} - \omega_0}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

t-value calculation

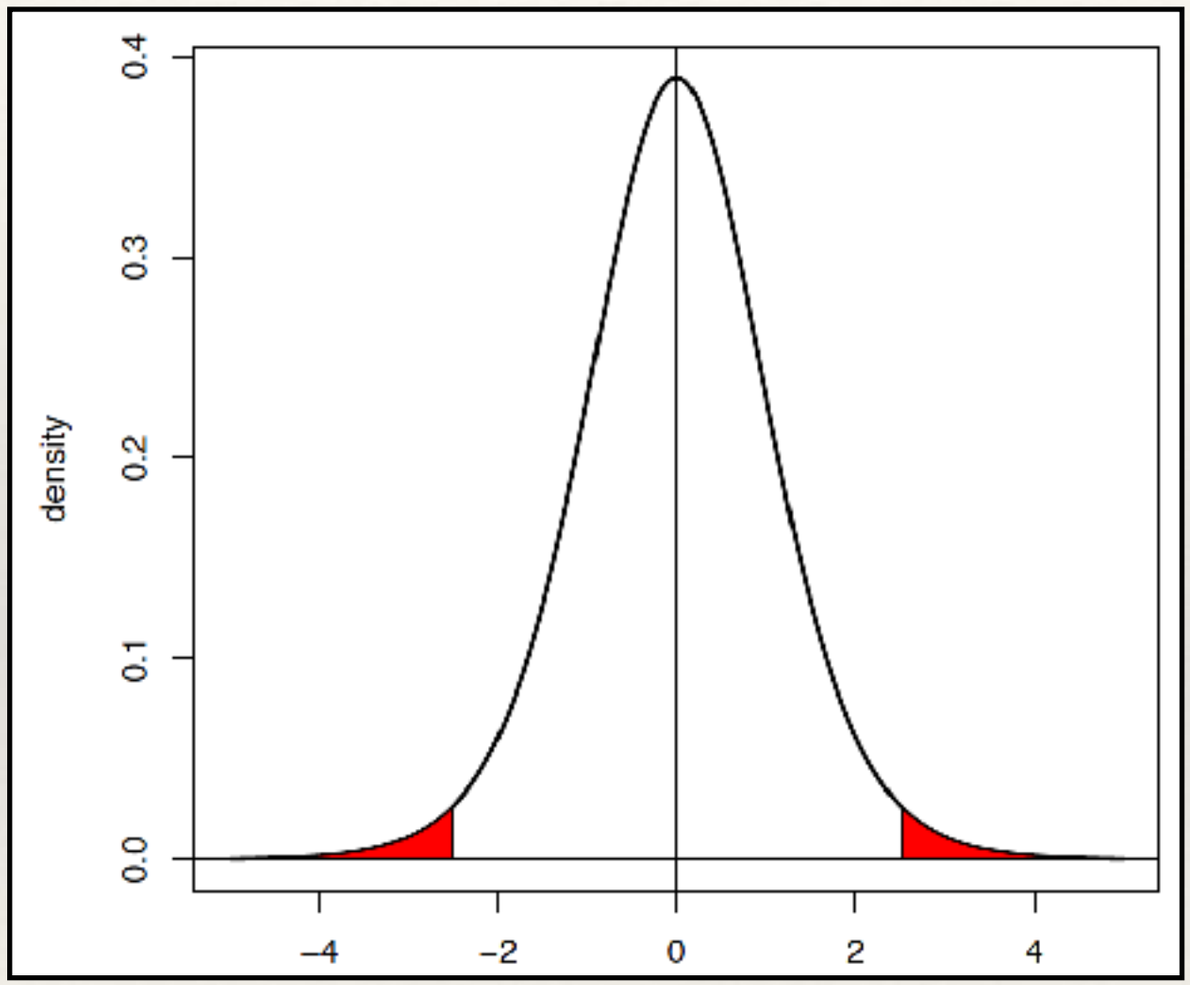
in general omega_0 = 0

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0$$

Test on rejection of equality

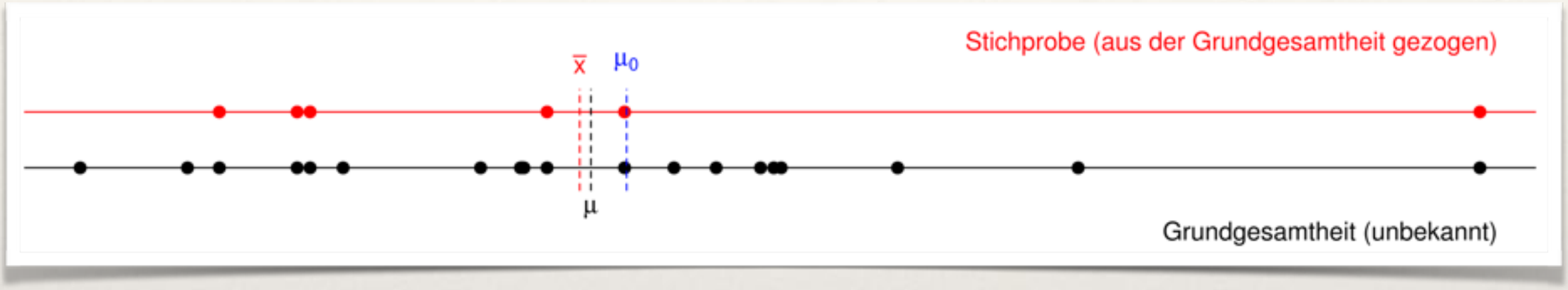


Problem variance & sub-sampling



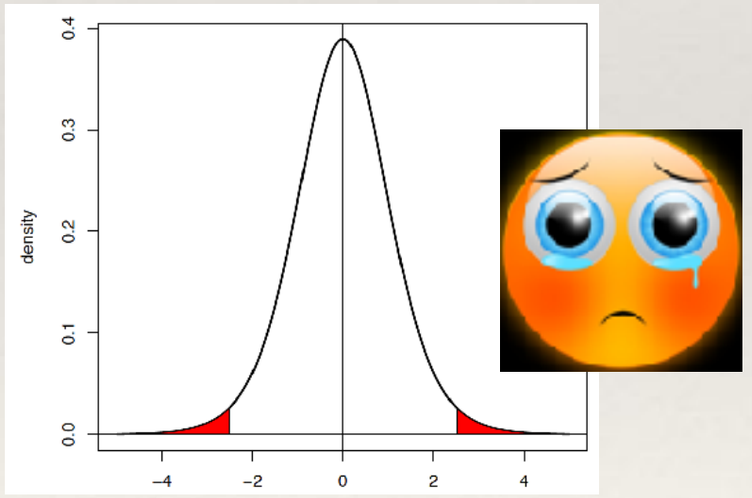
Why not use assume normal distribution for **variance**?

Problem variance & sub-sampling



u = true mean, x = empirical mean

Variance of dif. exp. not normal-distributed for sub-sampled data



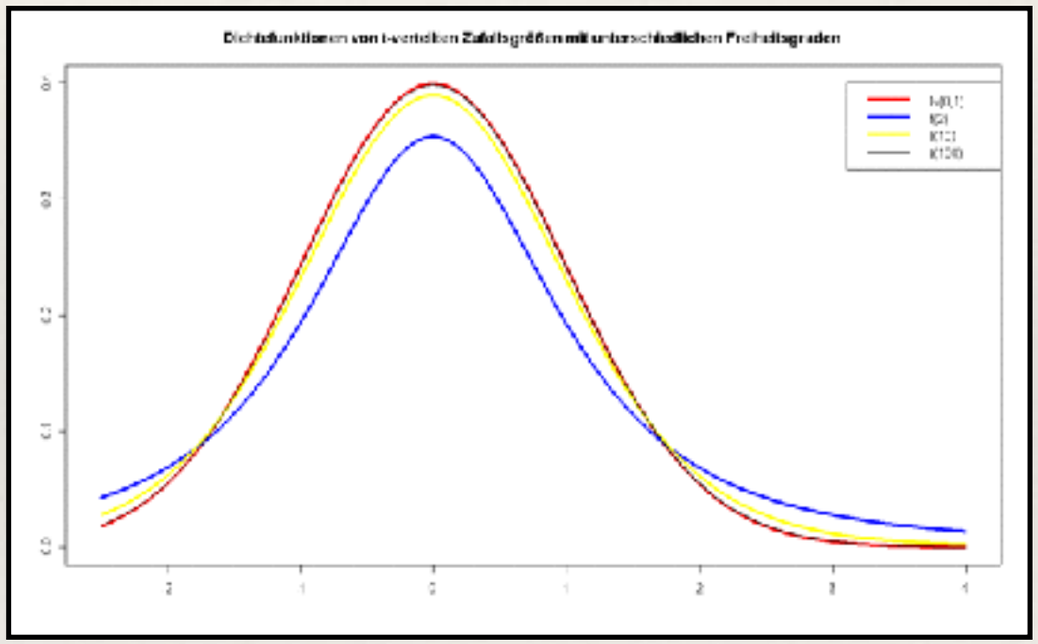
Bummer for normal distribution

T-distribution

Definition

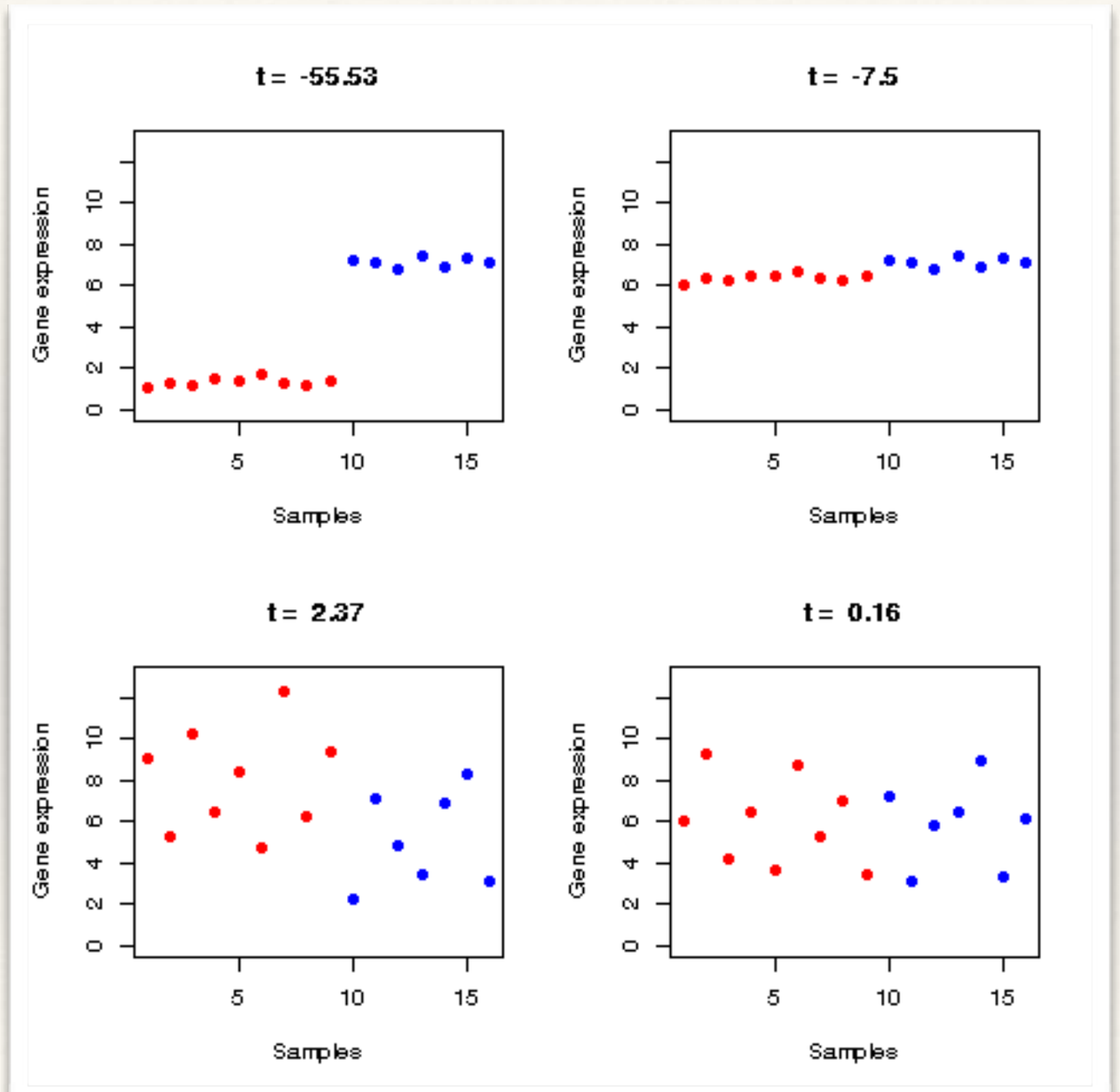
- Variance of sub-samples follows t-distribution

- Thus, apply *t-test* and not *normal-test*



Probability density function
t-distribution

Example T-statistic



T-test statistic

- Retrieve t-values from *test statistics*
- Based on |cohorts 1| (n) and |cohort 2| (m)

p-value = 1- value

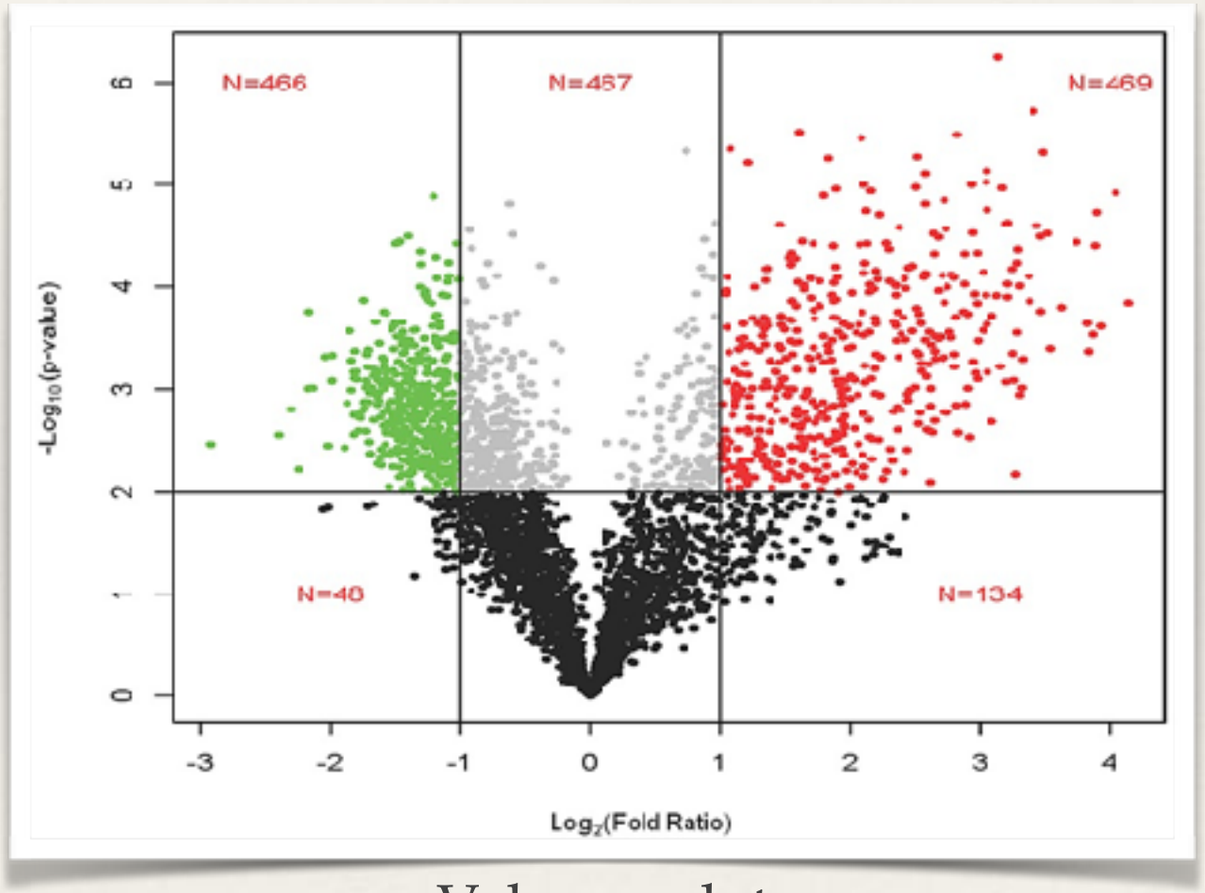
# m + # n - 2	0,75	0,875	0,90	0,95
1	1,000	2,414	3,078	6,314
2	0,816	1,604	1,886	2,920
3	0,765	1,423	1,638	2,353
4	0,741	1,344	1,533	2,132
5	0,727	1,301	1,476	2,015
6	0,718	1,273	1,440	1,943
7	0,711	1,254	1,415	1,895
8	0,706	1,240	1,397	1,860
9	0,703	1,230	1,383	1,833
10	0,700	1,221	1,372	1,812

t-statistic

P-value acquisition

Volcano Plot

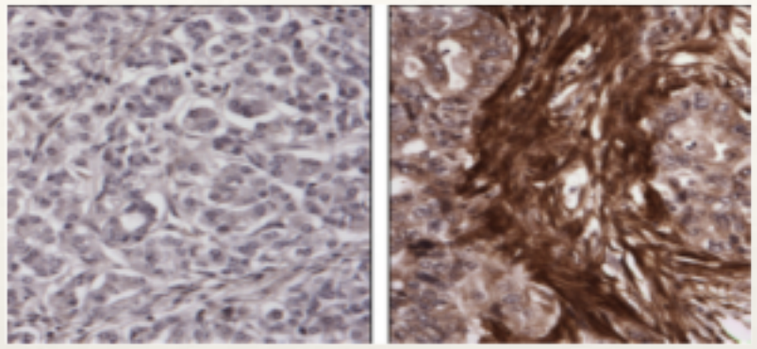
- Combines log-FC and p-value (here as negative log 10)
- Discretizes two-parameter cut-off
- Identifies dif. exp. genes



Volcano plot
Note higher (right) and lower (left) expression

Example hypothesis testing

N = {3.58, 4.14, 3.49, 3.37, 5.29, 5.06, 3.6}



T = {3.7, 10.9, 10.3, 3.57, 10.5, 8.18, 3.27}

Hypothesis $H_0: m_N - m_T = 0$ $H_1: m_N - m_T \neq 0$

Significance level $\alpha = 0.05$

Test statistic

$$t = \frac{mean(T) - mean(N)}{\sqrt{\frac{sd(T)^2}{m} + \frac{sd(N)^2}{n}}} = -2.27$$

(Critical value |T| = 2.45)

p-Value p-value = 0.06 → We cannot reject H_0, gene B ist not significantly differentially expressed!

Multiple Testing Problem

Thousands of hypotheses are tested simultaneously

- Increased chance of false positives
- 10,000 genes á chip, 10k * 0.01 = 100 have a p-value < 0.01 by chance
- Multiple testing methods allow to assess the statistical significance of findings

Corrected **P**-values := **Q**-values

Multiple Testing Problem

Approach 1: FWER

Family–wise error rate (FWER) is defined as the probability of at least one Type I error (false positive) among the genesselected as significant

$$FWER = Pr(V > 0)$$

	# non–rejected hypotheses	# rejected hypotheses	
# true null hypotheses (non-diff. genes)	U	V Type I error	m_0
# false null hypotheses (diff. genes)	T Type II error	S	m_1
	$m - R$	R	m

From Benjamini & Hochberg (1995).

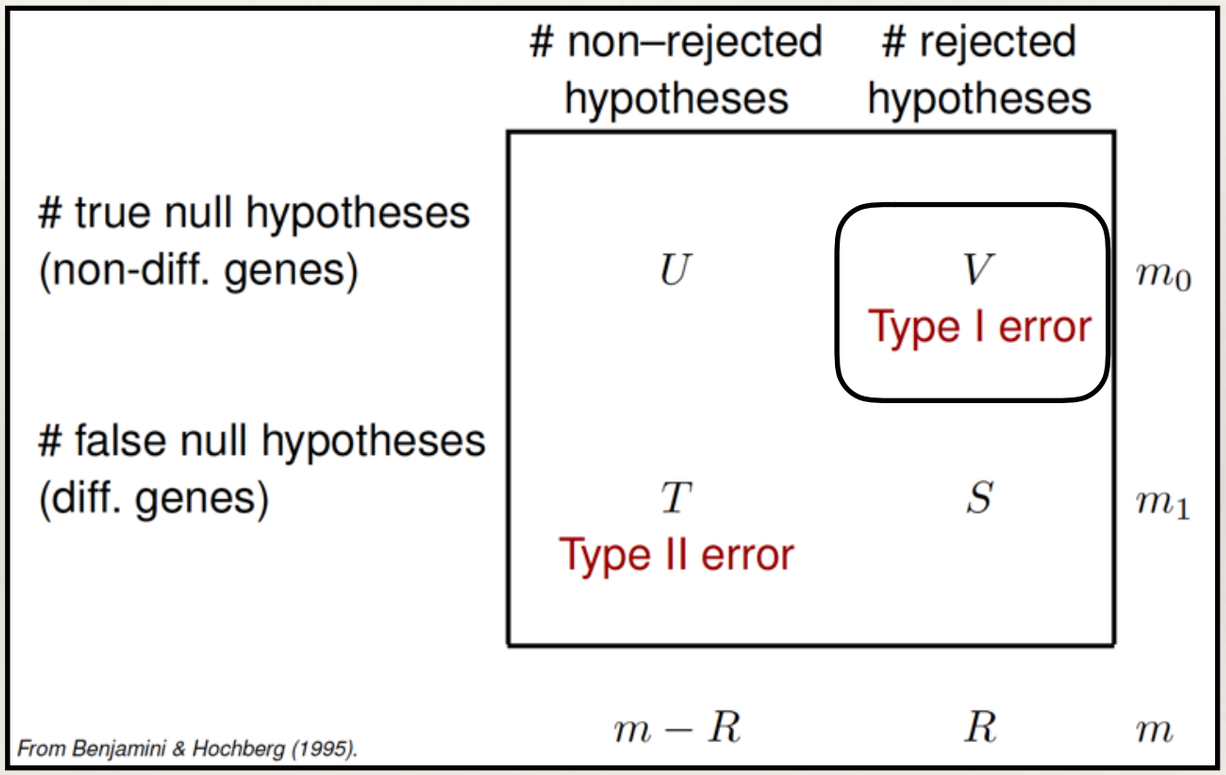
Multiple Testing Problem

Approach 2: FDR

False discovery rate (FDR), the expected ***proportion of true null hypotheses*** *rejected in the total number of rejections*

$$FDR = E(Q),$$

$$Q = \begin{cases} V/R, & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$



From Benjamini & Hochberg (1995).

Bonferoni-Correction

Q-value = P-value * # P-values

- Adjusted p-value is smaller than the pre-chosen significance value, probe is differentially expressed
- Very conservative (many failures to reject a false H0), rarely used
- Bonferoni assumes independence between the tests (usually wrong)
- Appropriate when a single false positive in a set of tests would be a problem (e.g., drug development)

Benjamini-Hochberger

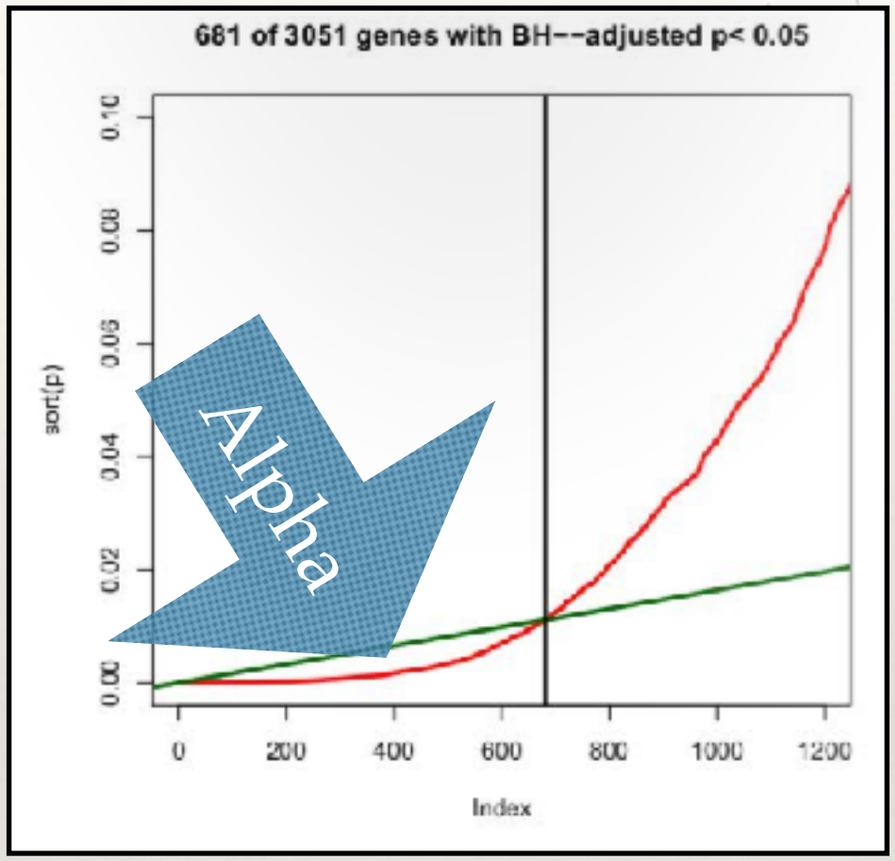
1. Choose α (e.g. α=0.05)

2. Sort p-values from small to large

3. Correct p-values: BH(p_i)

   i=1,…,m = pi * m/i

4. BH (p) = significant if BH(p) ≤ α



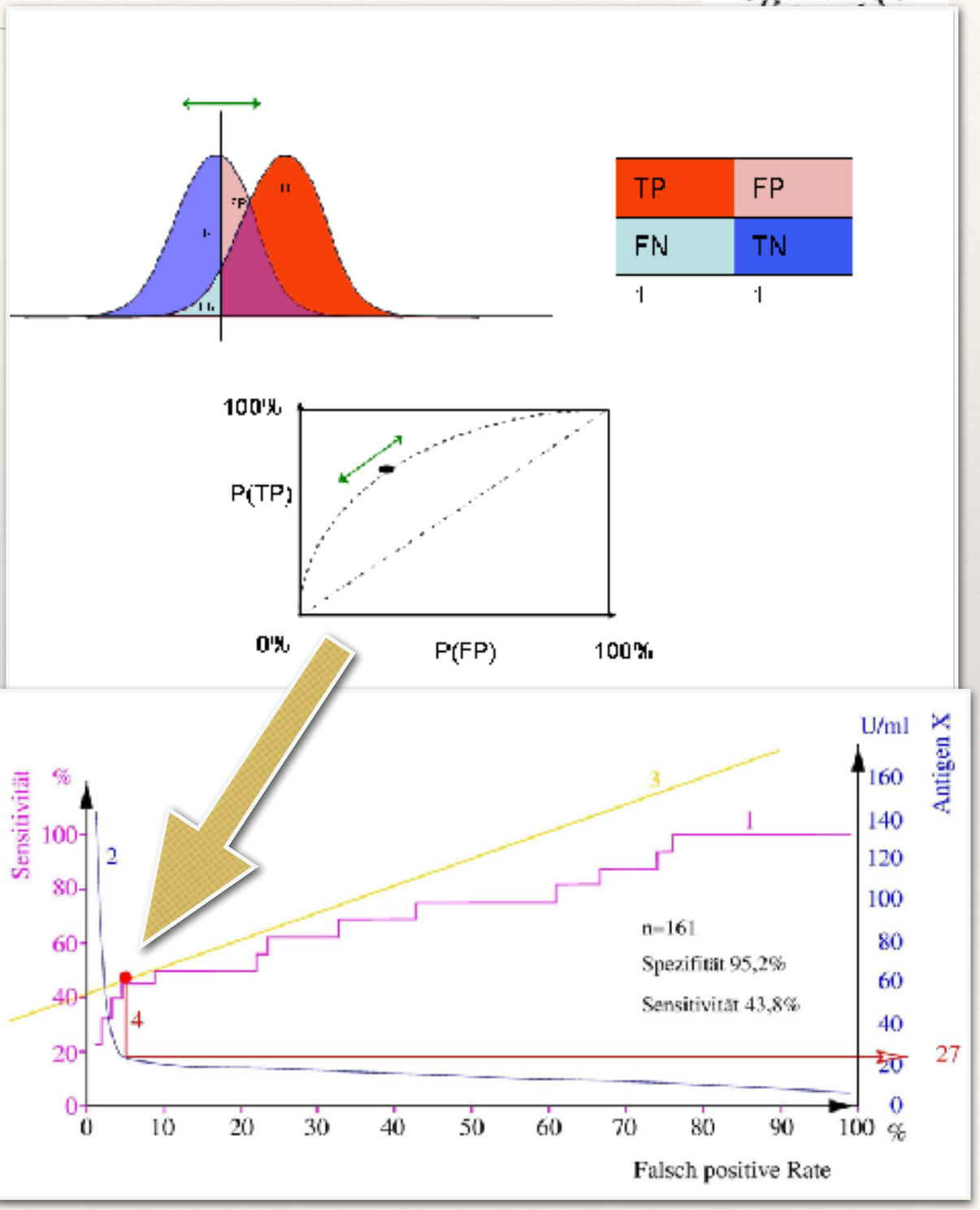
Area under curve holds 5% of p-values

Genes	p-value	rank	BH(p)	Significant? (α=0.05)
Gene A	0.00001	1	0.00001*1000/1=0.01	yes
Gene B	0.0004	2	0.0004*1000/2=0.20	no
Gene C	0.01	3	0.01*1000/3=3.3 → 1.0	no

ROC-curve

Receiver Operating Characteristic-curve

- Determine optimal e.g. q-values
- Trained on goldstandard
- Estimation of (future) sensitivity and specificity



Linear regression

- Model data

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Data

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Dependent variable

- Predict e.g. cancer-risk

- Identify correlated parameters B

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Correlated features

Linear regression

Linear regression equation

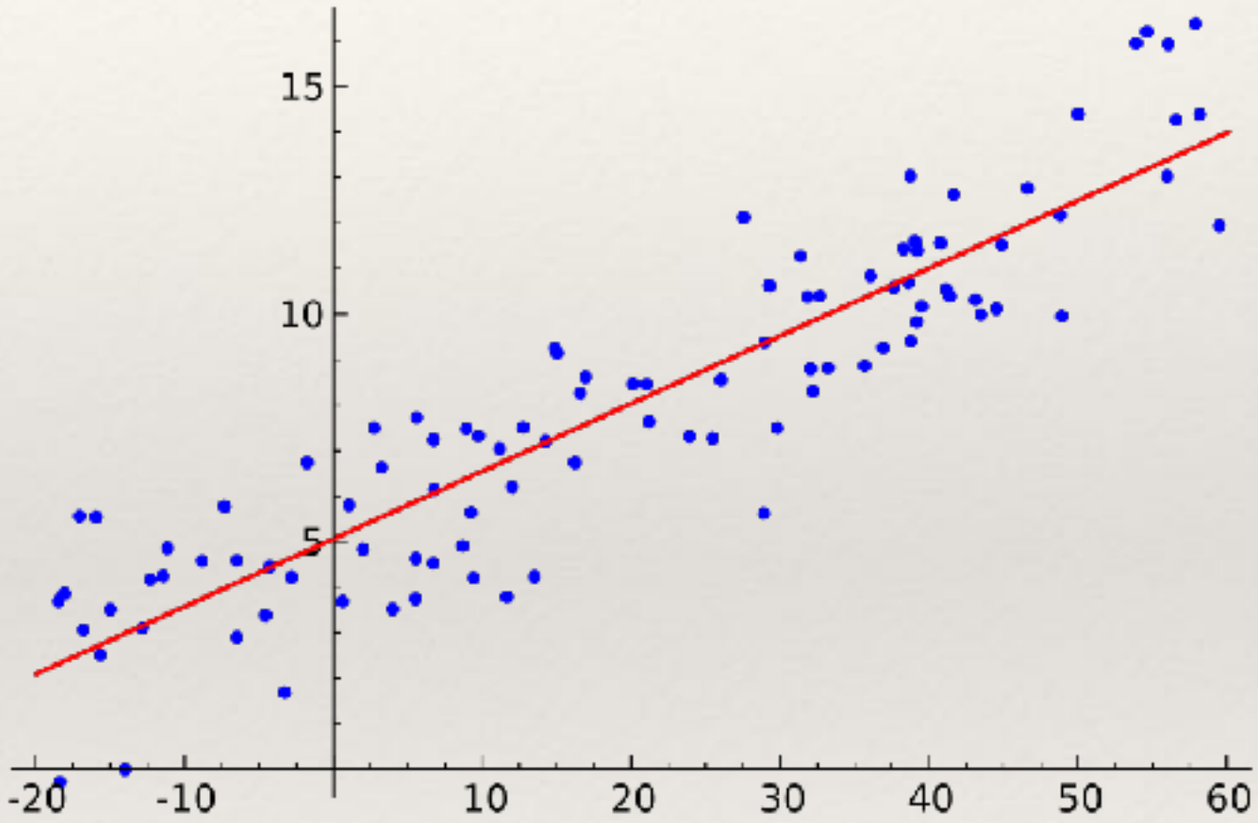
(without error)

$$\hat{Y} = bX + a$$

predicted values of *Y*

b = slope = rate of predicted ↑/↓ for Y scores for each unit increase in *X*

Y-intercept = level of *Y* when X is 0



$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Y (effect) = X (data) * B (linear parameters)

Robust Multi-array Average

- Abreviated RMA

- Utilized match & mismatch probes

1. Corrected, log 2 data
2. Rank expression
3. Replace ranked expression-values by mean
4. Linear (regression) expression-model

RMA example

		Background-Corrected and Log-Transformed Perfect-Match Intensity		
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
1	1	7	9	19
1	2	3	5	14
1	3	2	6	11
2	1	4	8	8
2	2	10	11	16
2	3	12	10	15

Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
1	1	7	9	**14**
1	2	3	5	14
1	3	2	6	11
2	1	4	8	8
2	2	10	**14**	16
2	3	**14**	10	15

RMA example

		Background-Corrected and Log-Transformed Perfect-Match Intensity		
Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
1	1	7	9	*14*
1	2	3	5	14
1	3	2	6	11
2	1	4	8	8
2	2	**12**	*14*	**12**
2	3	*14*	**12**	15

2 … n

Probe Set	Probe	GeneChip 1	GeneChip 2	GeneChip3
1	1	*10.33333*	*10.33333*	*14*
1	2	*6.66667*	**5**	*8.66667*
1	3	**5**	*6.66667*	*6.66667*
2	1	*8.66667*	*8.66667*	**5**
2	2	*12*	*14*	*12*
2	3	*14*	*12*	*10.33333*

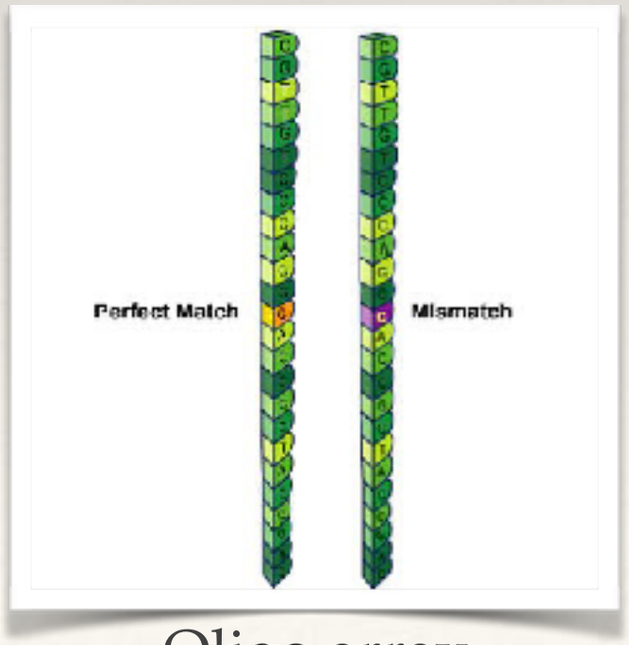
Linear (regression) model

$$Y_{ij}=m_i+a_j+e_{ij}$$

- Y_ij = corrected (single) probe's value
- m_i = probe **set** value
- a_j = (single) probe's affinity
- e_ij = error term
- i = Sample
- j = Probe

Normalized single probe expression

Note distinction between perfect and mismatch probes

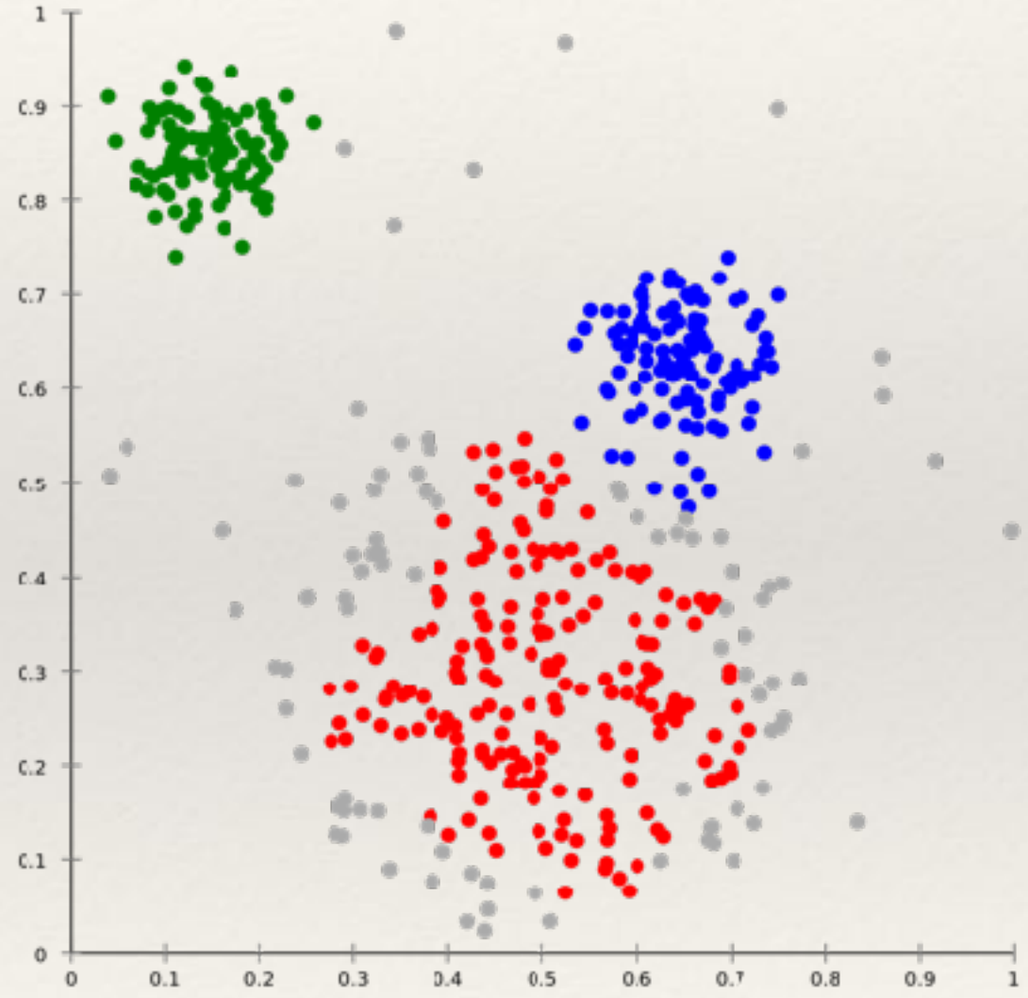


Oligo array

Clustering

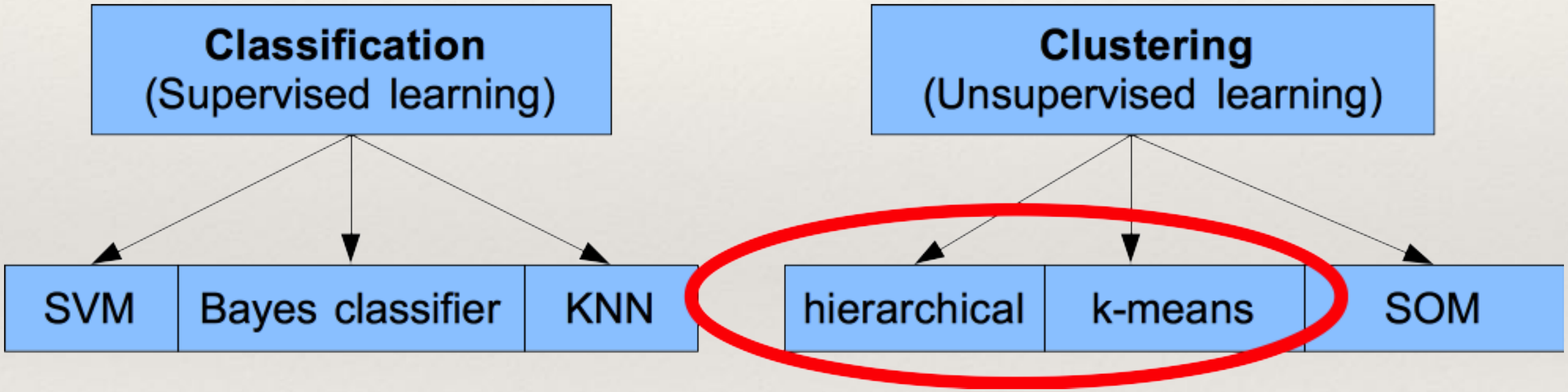
- Identify subgroups
- Quality control
- Similarity-based

Distance metric critical



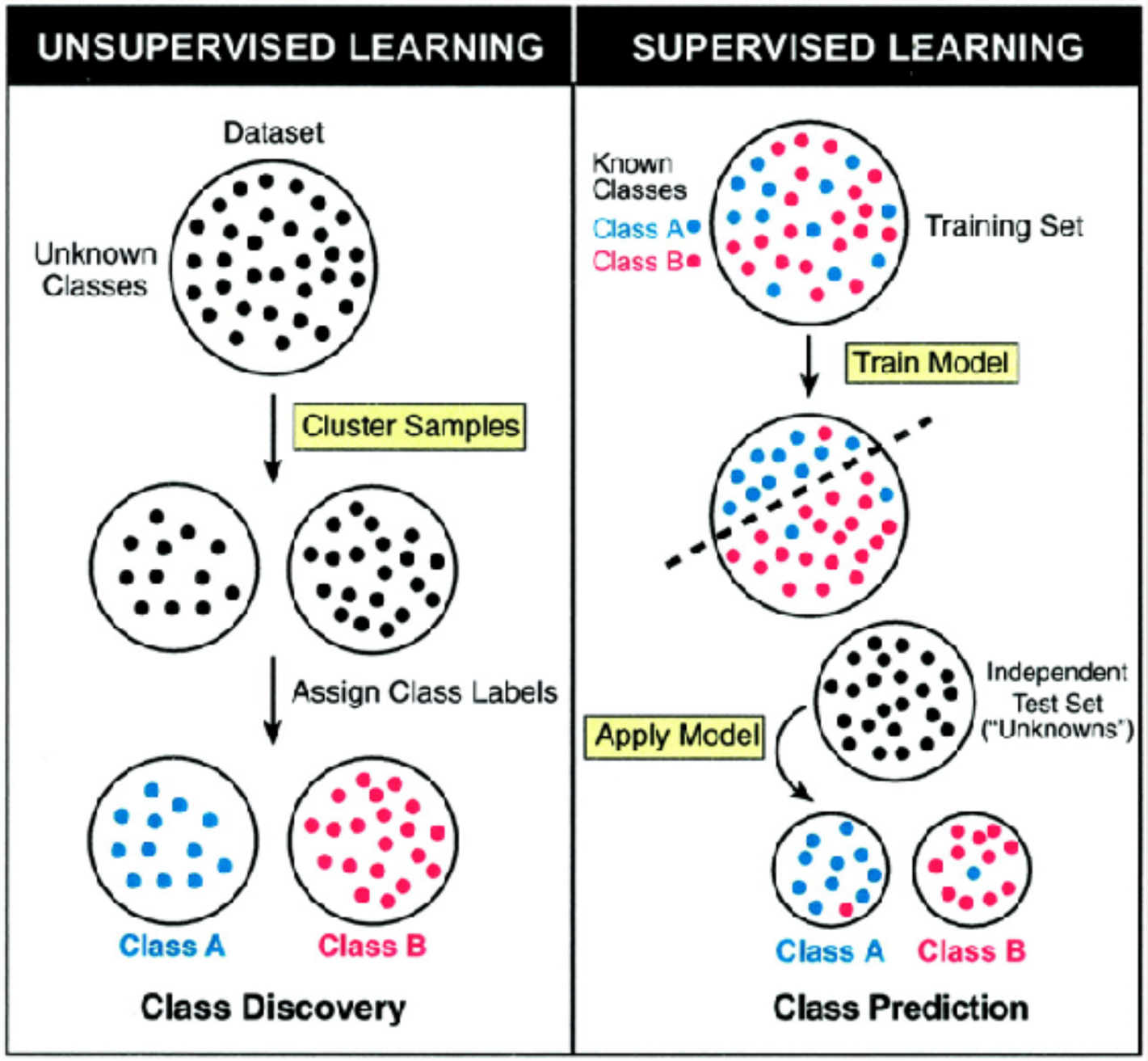
Colors == spacial-clustering

Overview Clustering

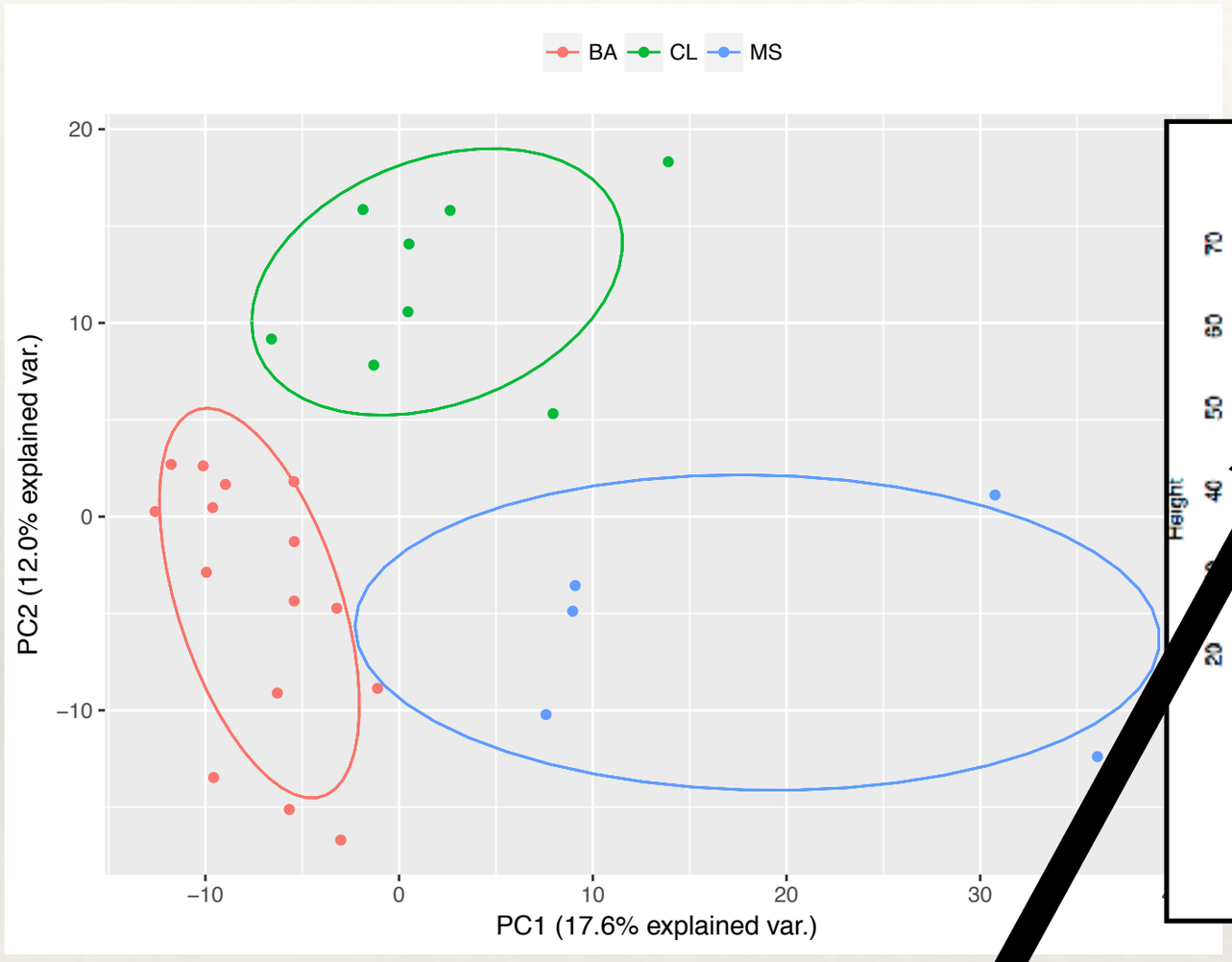


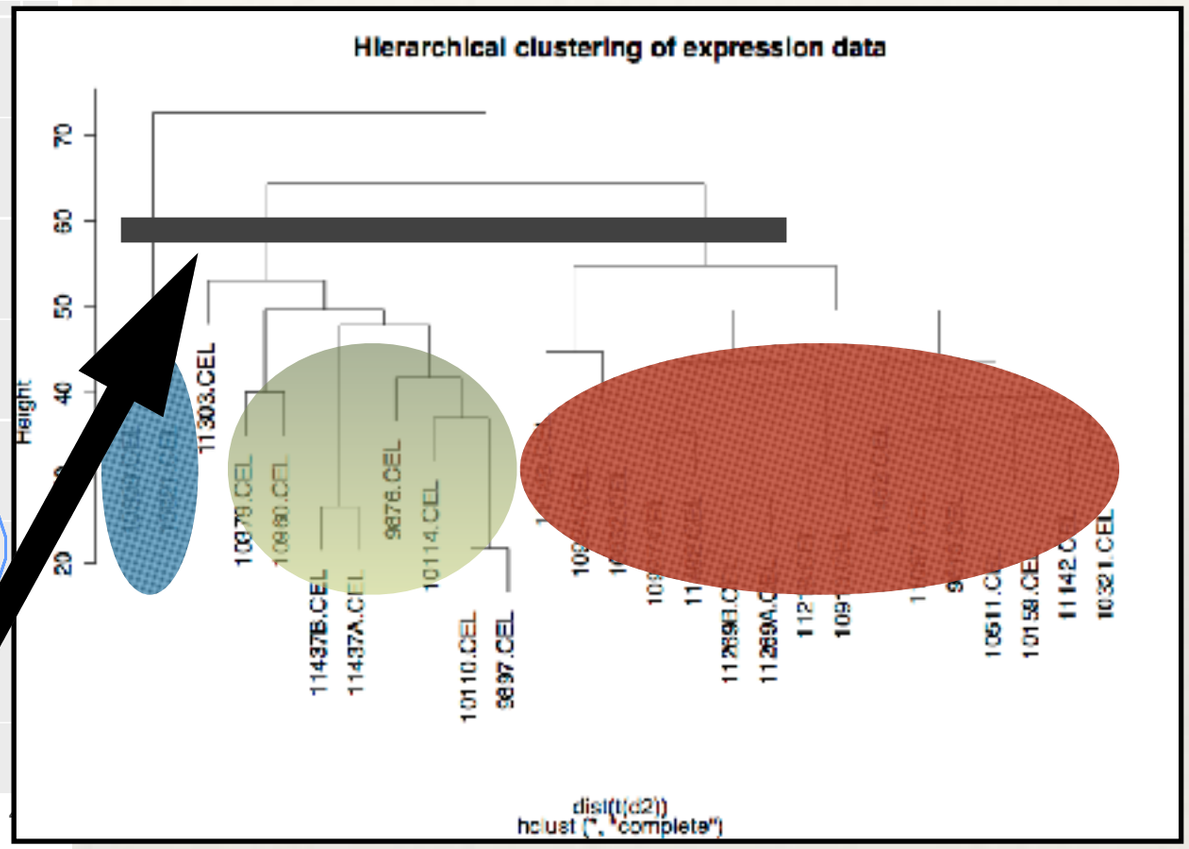
Today's topic

Unsupervised vs. supervised



Example Clustering





- Colors := hierarchical tree-cut

Hierarchie pair-wise similarity based

Hierarchical Clustering

1. Choose distance metric

   Euclidean

   Pearson, etc.

2. Compute similarity matrix S

Hierarchical Clustering

1. Choose distance metric

   Euclidean

   Pearson, etc.

2. Compute similarity matrix S

3. While |S|>1:

   Determine pair (X,Y) with minimal distance

   Compute new value Z = avg (X,Y),

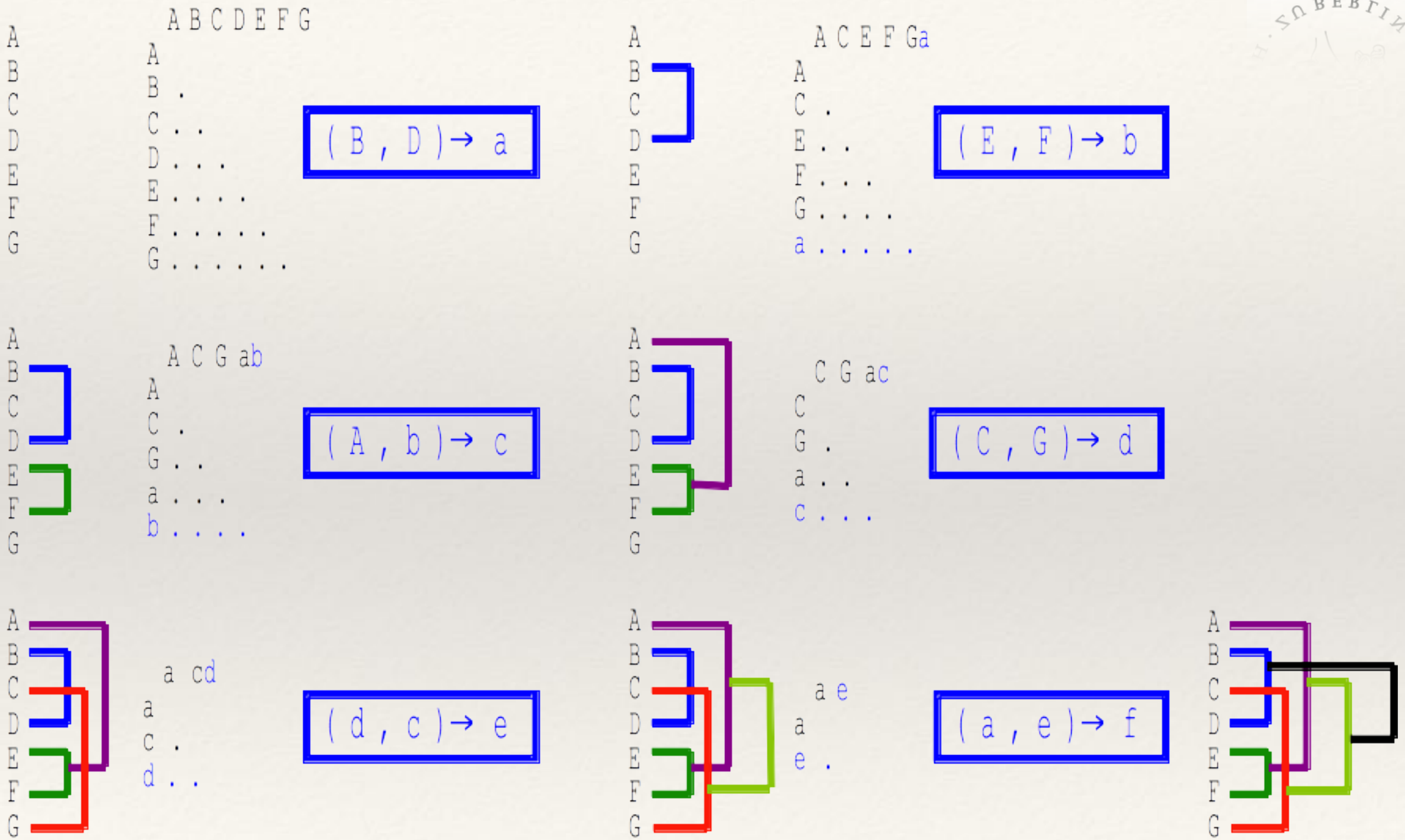
   (single, average, or complete linkage)

   Delete X and Y in S, insert Z in S

   Compute new distances of Z to all elements in S

   Visualize X and Y as pair

Example hierarchical clustering



Distance metrics

- Define ‚distance' i.e. **which** data (dots) are merged
- Linear vs. non-linear distances
- Differ especially w.r.t. outlier-sensitivity

- Euclidian $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
- Squared $\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
- Manhattan $\|a-b\|_1 = \sum_i |a_i - b_i|$
- Maximum $\|a-b\|_\infty = \max_i |a_i - b_i|$
- Mahalinobis $\sqrt{(a-b)^\top S^{-1} (a-b)}$
  - S = Correlation matrix

Linkage Rules

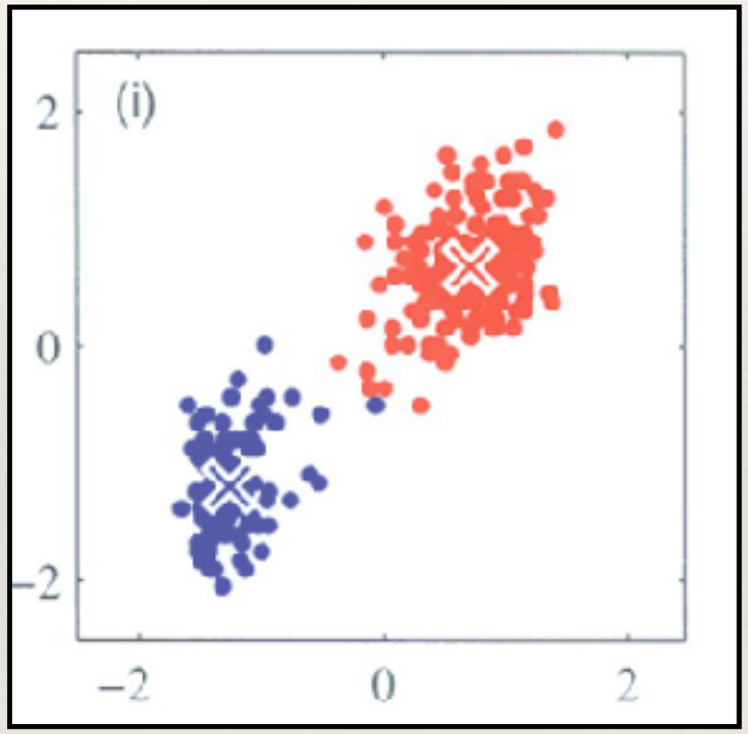
- Define **how** to cluster data (dots)
- Represent desired ‚definition' of a cluster
  - E.g. ‚mean'-linkage will generall yield more balanced clusters

- Single $\min\{ d(a,b) : a \in A, b \in B \}$
- Complete $\max\{ d(a,b) : a \in A, b \in B \}$
- Average $\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$
- Cluster-centers $\|c_s - c_t\|$
  - c = cluster-centroids

K-Means clustering

- Partitions n observations into k clusters

- Minimize the distance of the n data points from their respective cluster centres.

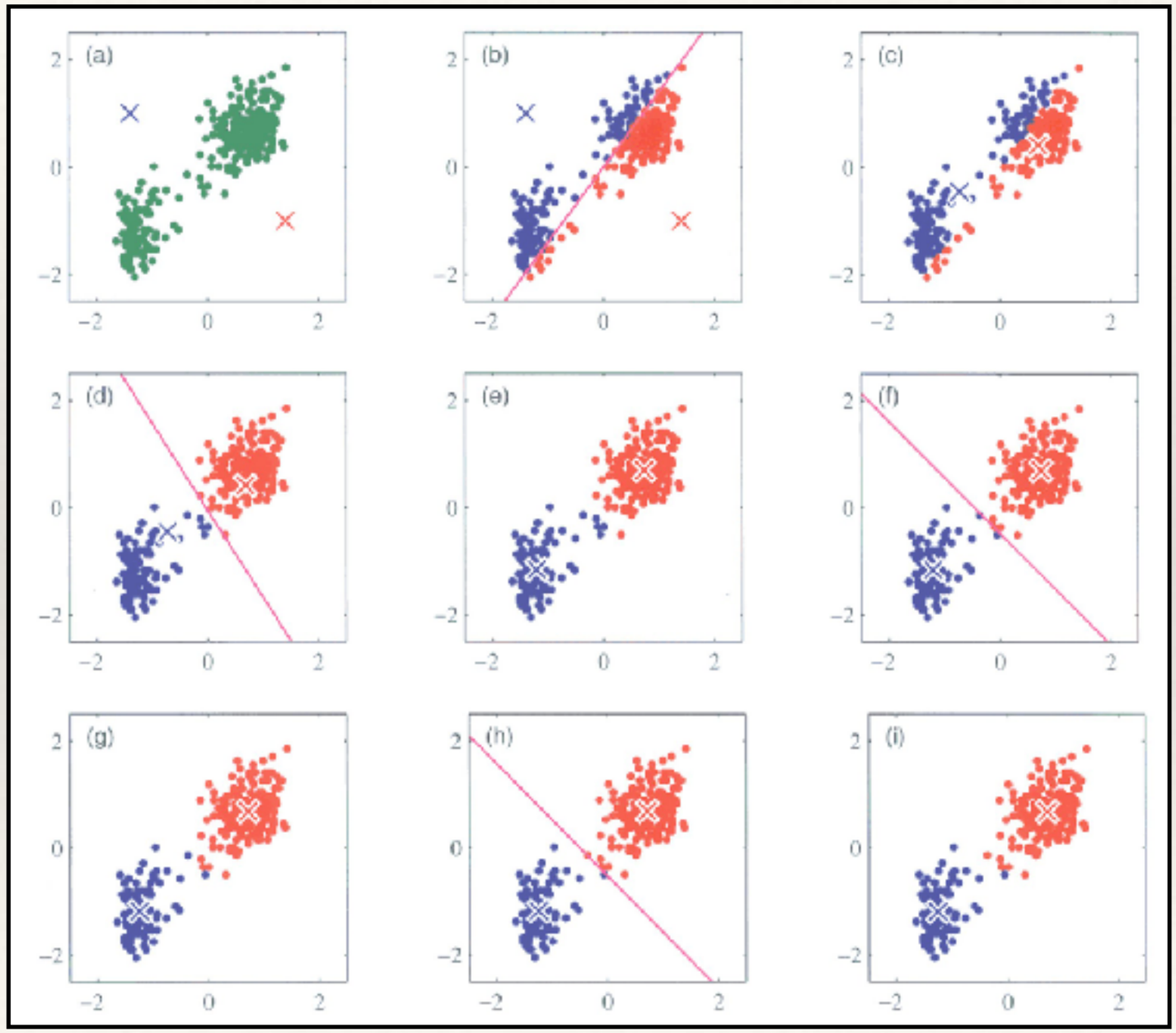


Cluster on proximity of k-centers

Difference hierarchical clustering: No pair-wise clustering

K-Means clustering

1. Choose k random cluster centers μ1,...μk
2. Assign for each point x in dataset S the closest cluster center
3. Compute a new center μi for every cluster Ci
4. Repeat 2-3. until cluster centers do not change



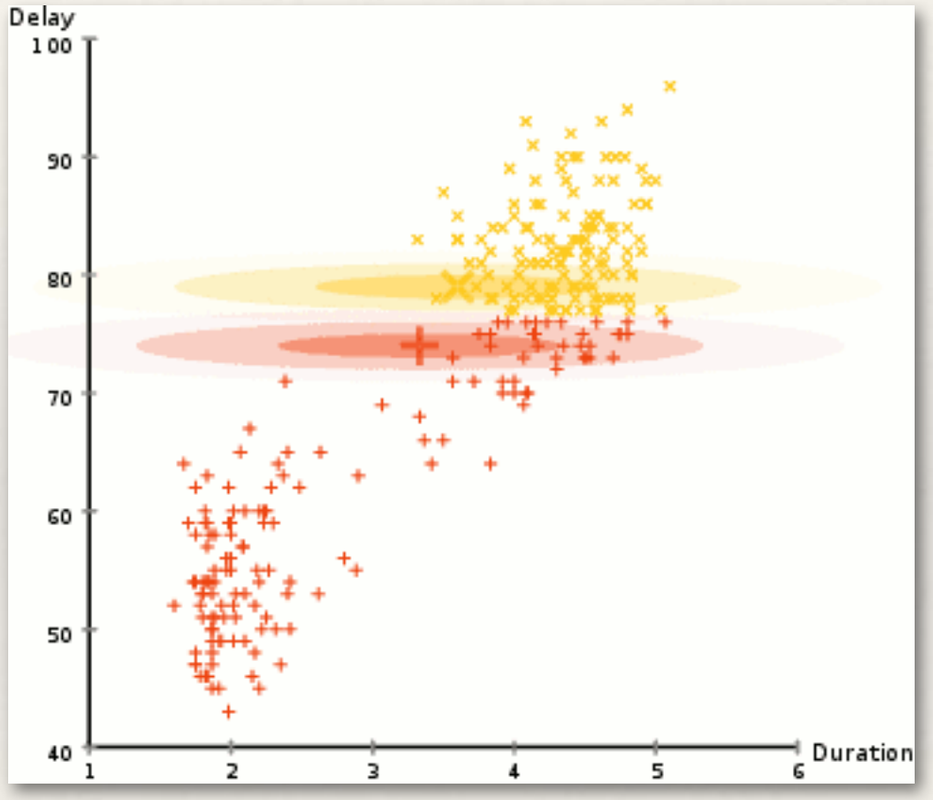
Cluster-centers critical

Maximum-likelihood

- Find optimal cluster-centers

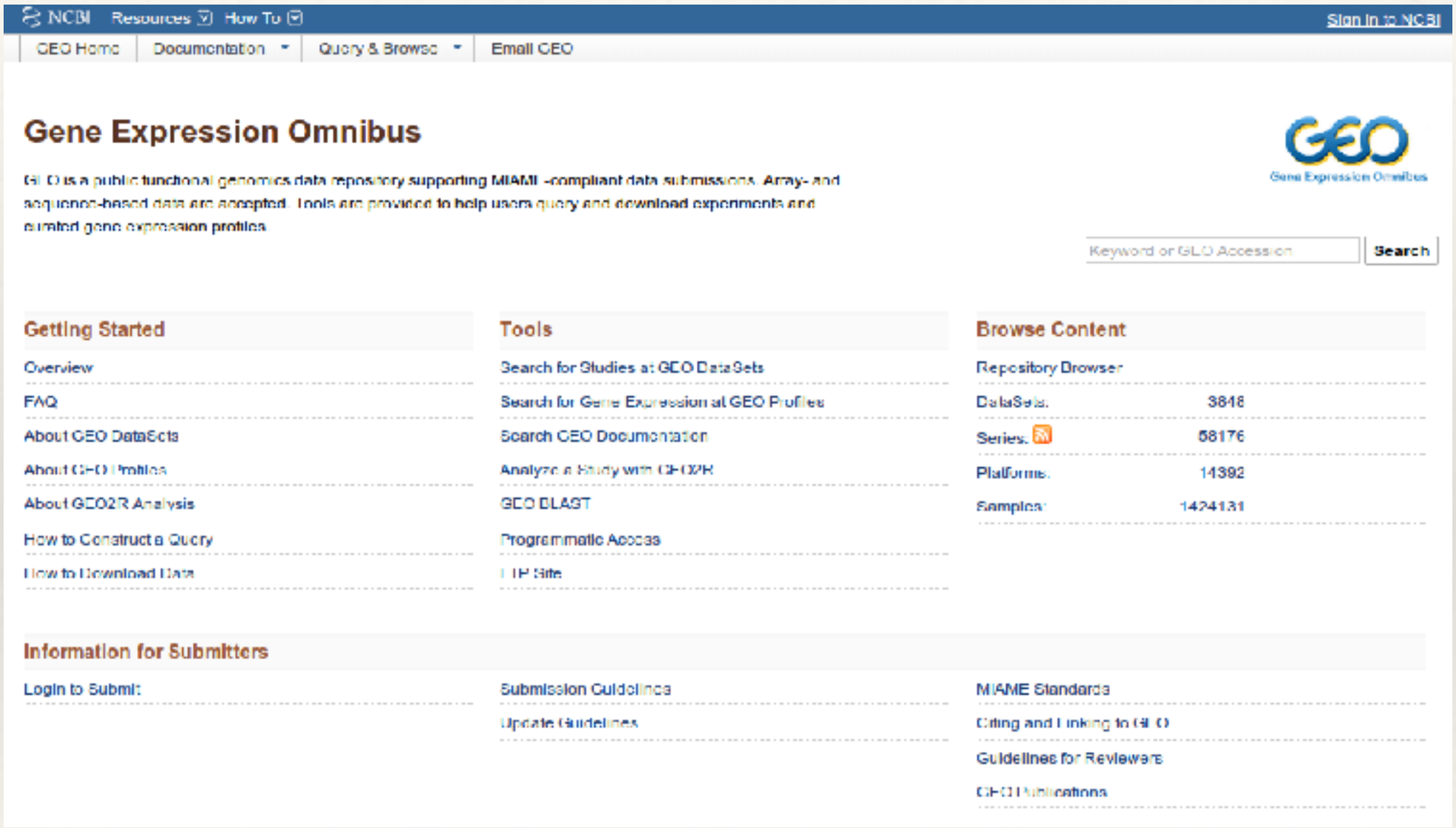
- **Convergence not assured**

- Initialization and number of centers (centroids) critical



Maximum likelihood centroids

Databases - GEO

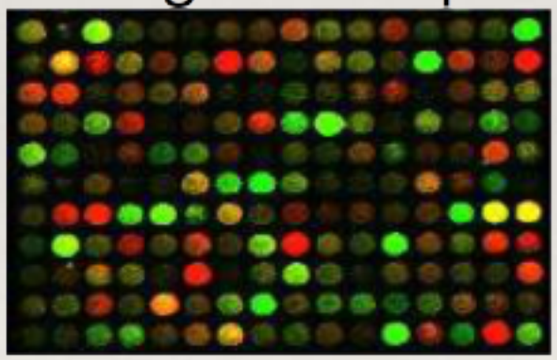


Databases - GEO

NCBI public repository http://www.ncbi.nlm.nih.gov/geo/
archives microarray, NGS, and other high-throughput
genomics data submitted by the research community

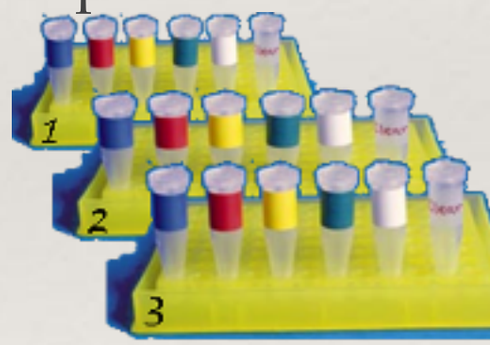
GPL
(GEO platform)
platform description

GSM
(GEO sample)
raw-processed
intensities from a
single or chip



GSE
(GEO series)
grouping of chip data,
a single experiment

GDS
(GEO dataset)
grouping of
experiments



submitted by experimentalist

curated by NCBI

MIAME checklist

1. Raw data present
2. Processed data present
3. Sample annotation present (e.g. experimental factors, values & protocols)
4. Experimental design explained (e.g. what samples are replicates and why)
5. Annotation of the array (e.g., gene identifiers & genomic coordinates)
6. Laboratory and data processing protocols (e.g. normalisation method)

Take-home messages

Differential expression

- Combination Log-FC and P-values (Volcano plot)

- T-test identifies **significantly differentially expressed** genes

- Multiple-testing correction

Take-home messages

Clustering

- Identifies subgroups
- Depends on **distance metric** & **linkage function**
- GEO databases offer public expression data