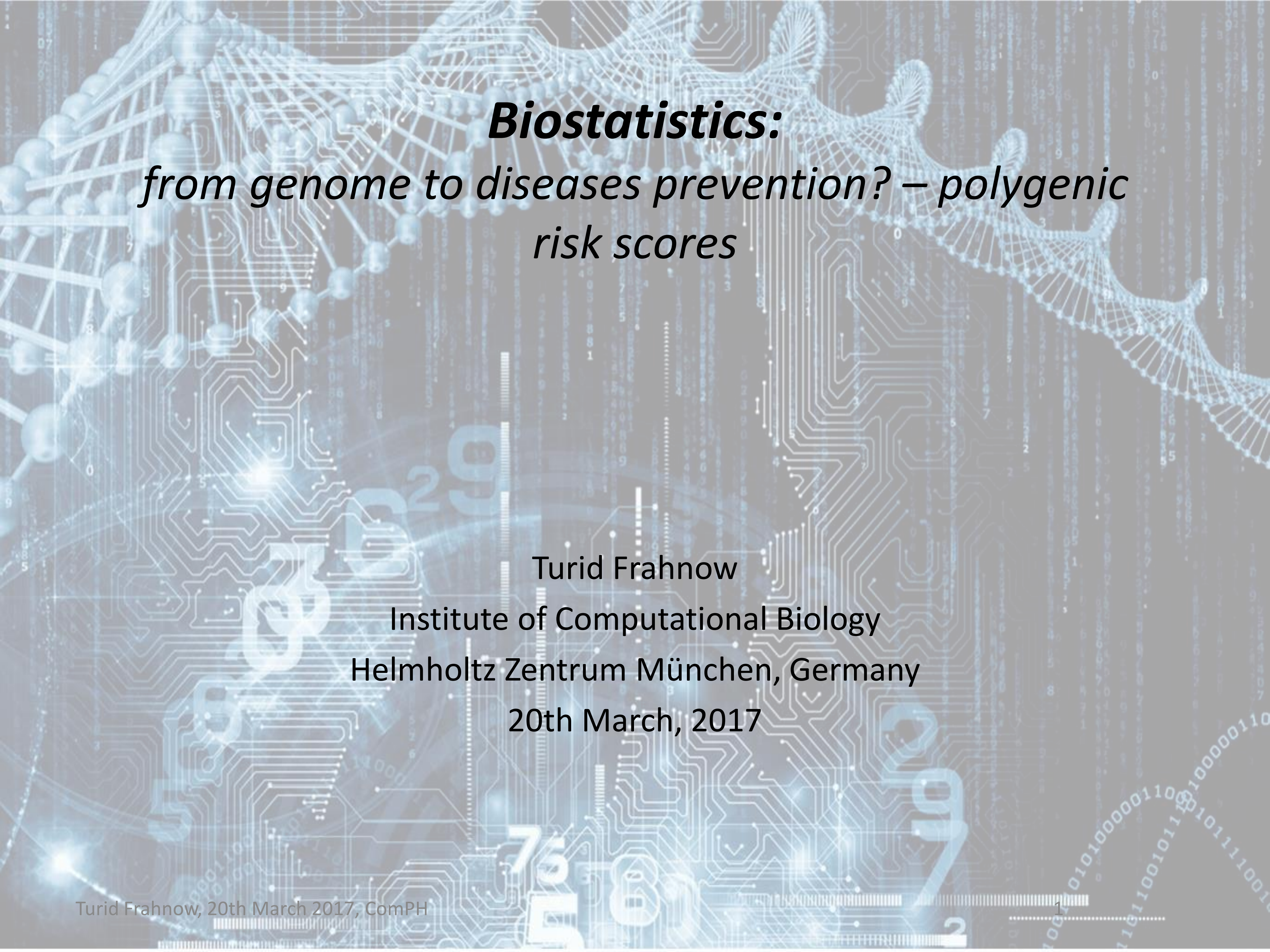


Biostatistics:

from genome to diseases prevention? – polygenic risk scores

Turid Frahnow

Institute of Computational Biology

Helmholtz Zentrum München, Germany

20th March, 2017

Schedule

- Biostatistics in general
  - What is it?
  - Why we use it?
  - Applications

- Biostatistics in the context of polygenic risk scores
  - Biological background
  - What is it?
  - Why we use it?
  - Problems

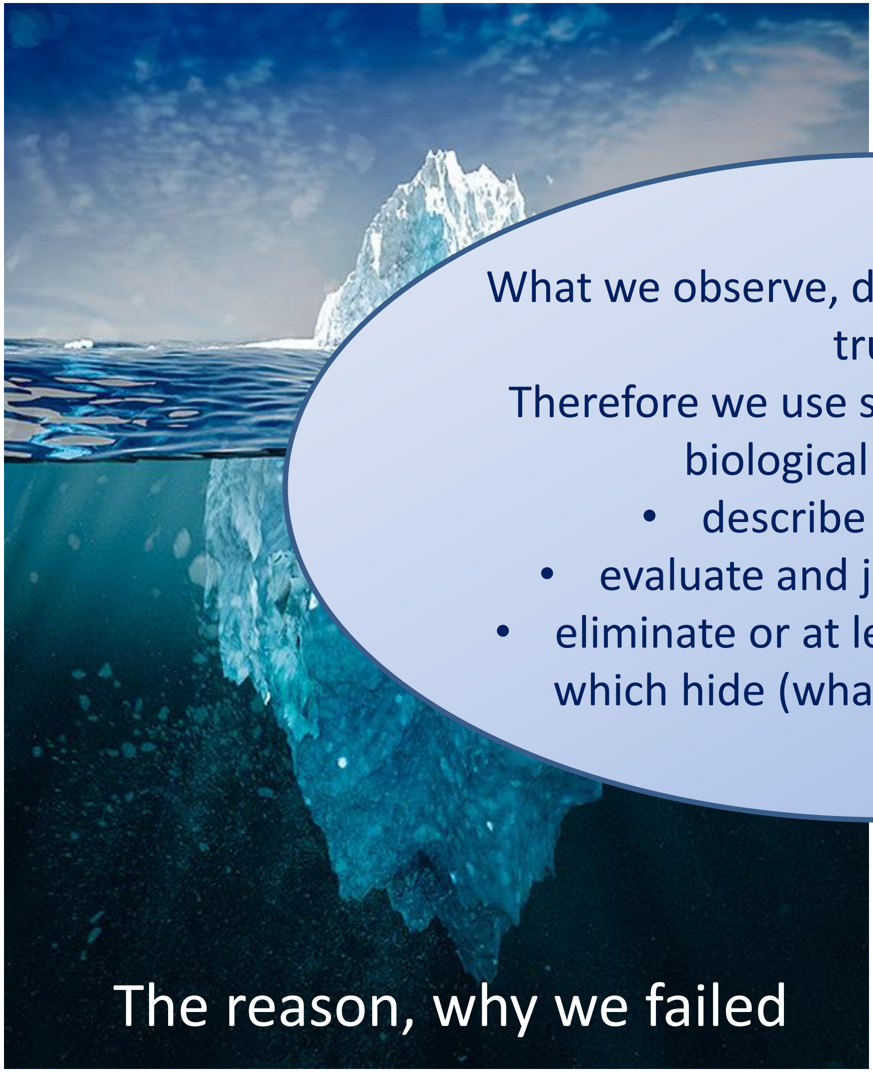
First of all…

Please interrupt me (immediately),
if you don't understand something!

Biostatistics in general

Motivation

Our observation / data



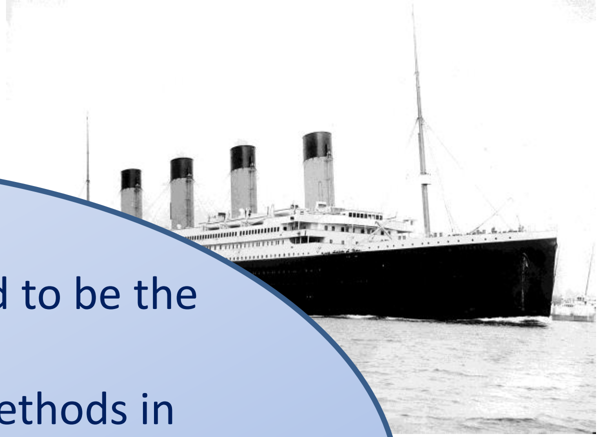
The reason, why we failed

What we observe, doesn‘t need to be the truth!

Therefore we use statistical methods in biological contexts to

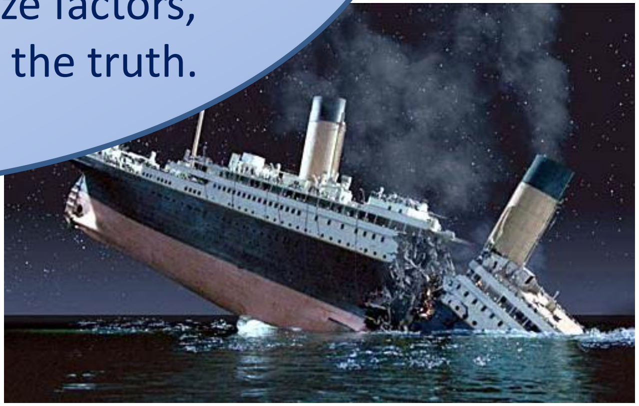
- describe observations,
- evaluate and judge observations,
- eliminate or at least minimize factors, which hide (what might be) the truth.

Our decision/conclusion



RMS Titanic

lt ☹



Biostatistics - „measurement of life“

= medical statistics, biometry …

- Definition:

  „Biostatistics is the branch of statistics responsible for the proper interpretation of scientific data generated in biology, medicine, public health and other health or natural sciences.“

- Or shorter: The **application** of statistics to biological/natural sciences.

- And important: It is much more, than using a t-test.

- (Should be) always interdisciplinary !

It‘s about understanding….

Biological Problem >

Experimental design >

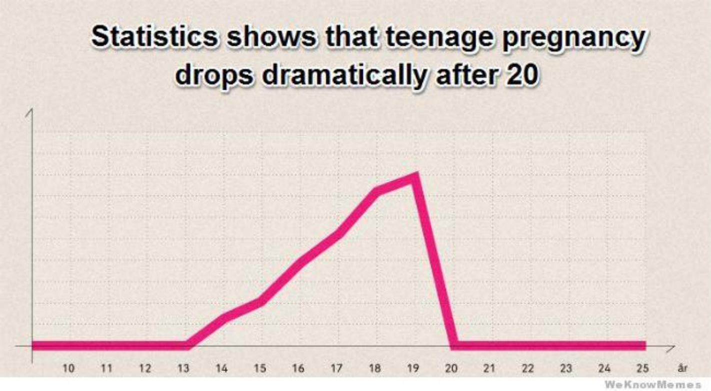
Experimental techniques>

Data structure>

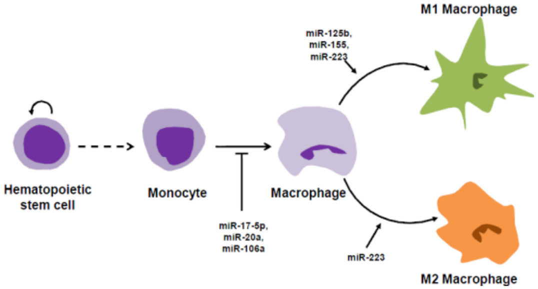
(Transformation >)

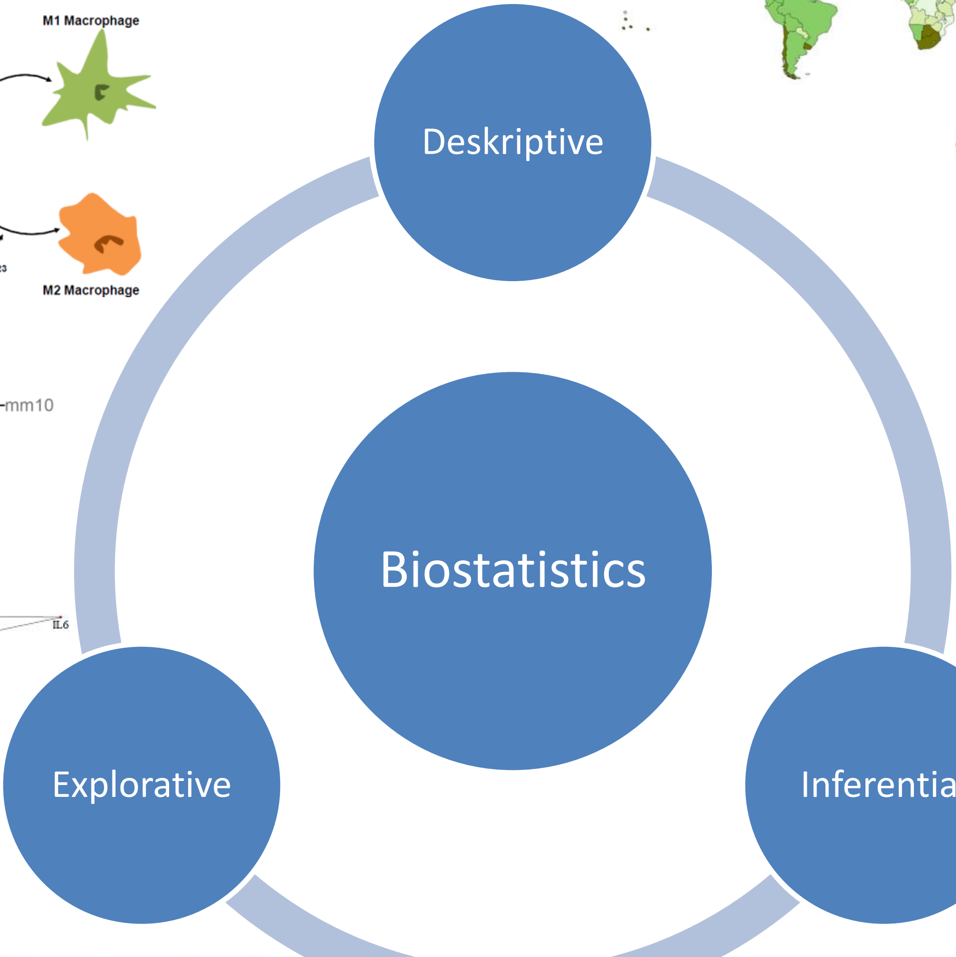
Results >

Interpretation



Applications



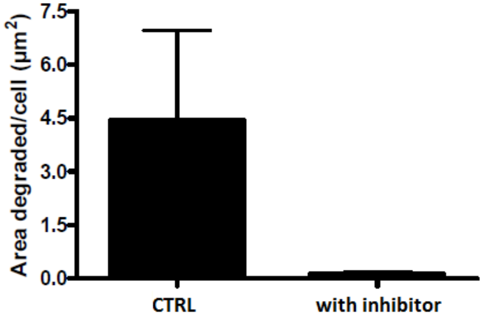


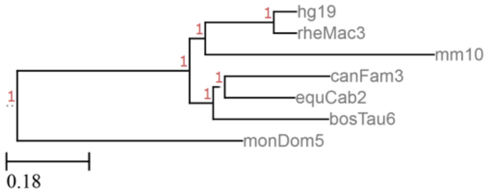
Deskriptive

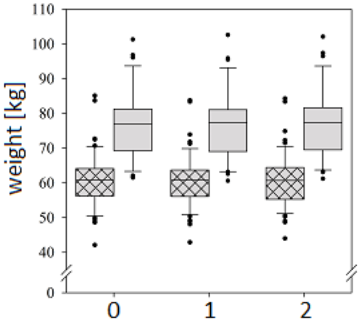
Biostatistics

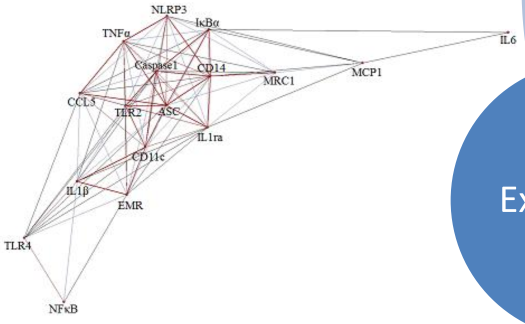
Explorative

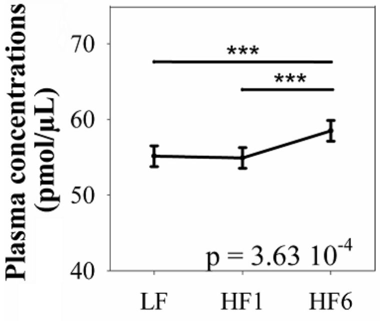
Inferential

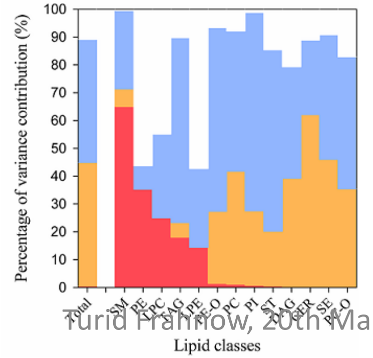


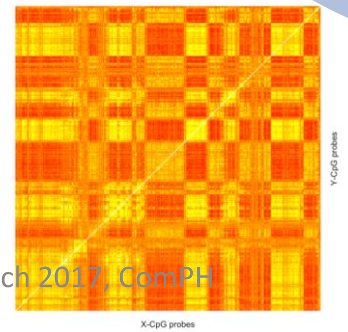


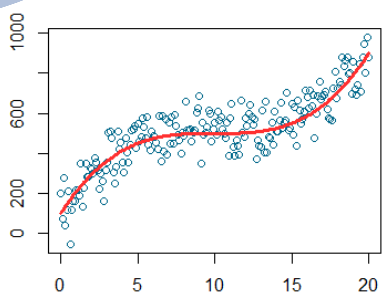








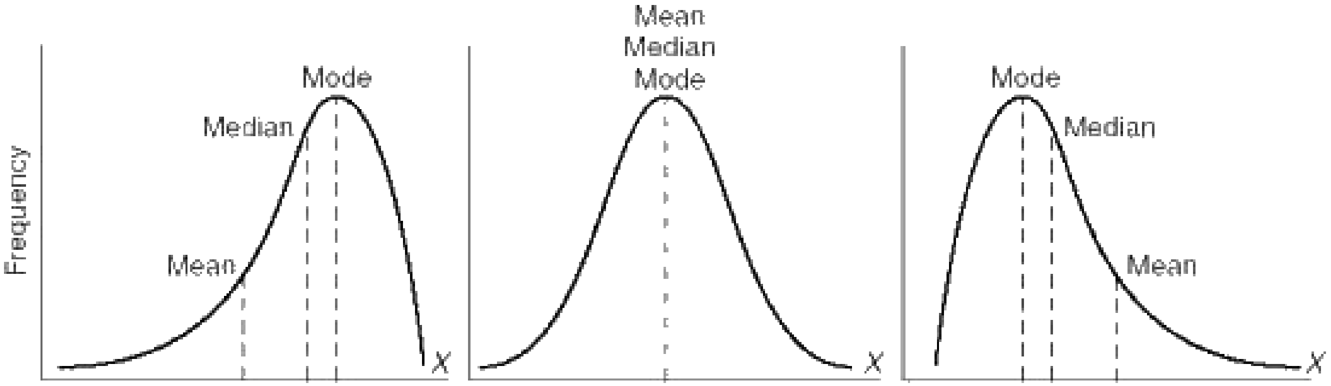




Pierre-Charles-Alexandre Louis (1787-1872) was a pioneer of the “numerical method” in medicine.

Descriptive statistics

The 3 m’s of central tendency



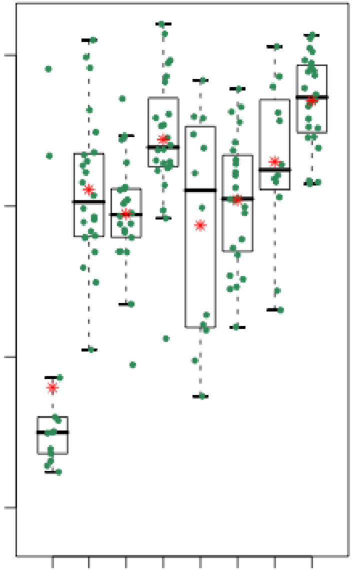
- **Mean** = the sum of observations divided by the number of observations

- **Median** = the observation in the center when all observations are ordered from smallest to largest

- **Mode** = the most frequently occuring value among all observations

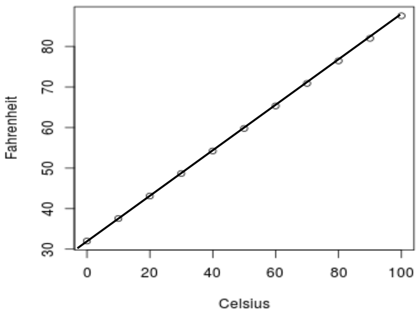
Measures of Spread/Variation

- **Range** = Max – Min
- **Standard deviation** = amount of variability of the observations (range of 68% of the observations around the mean)
- **Standard error (of mean)** = amount of variability in the population mean
- **Interquartile range** = range between 25% and 75% quartile (range of 50% of the observations around the median)
- **Quartile coefficient of dispersion,**
  coefficient of variation,
  …

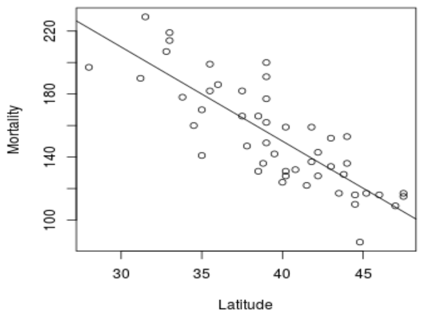


Associations

- **Correlation** = measurement of association



deterministic



statistical

- **Odds ratios** = the odds of exposure in the group with disease divided by the odds of exposure in the control group (Case-Control design)

- **Relative risk** = the ratio of the incidence of disease in the treated group divided by the corresponding incidence of disease in the placebo group (Cohort design)

		Disease	
		Yes	No
Test	+	A	B
	−	C	D

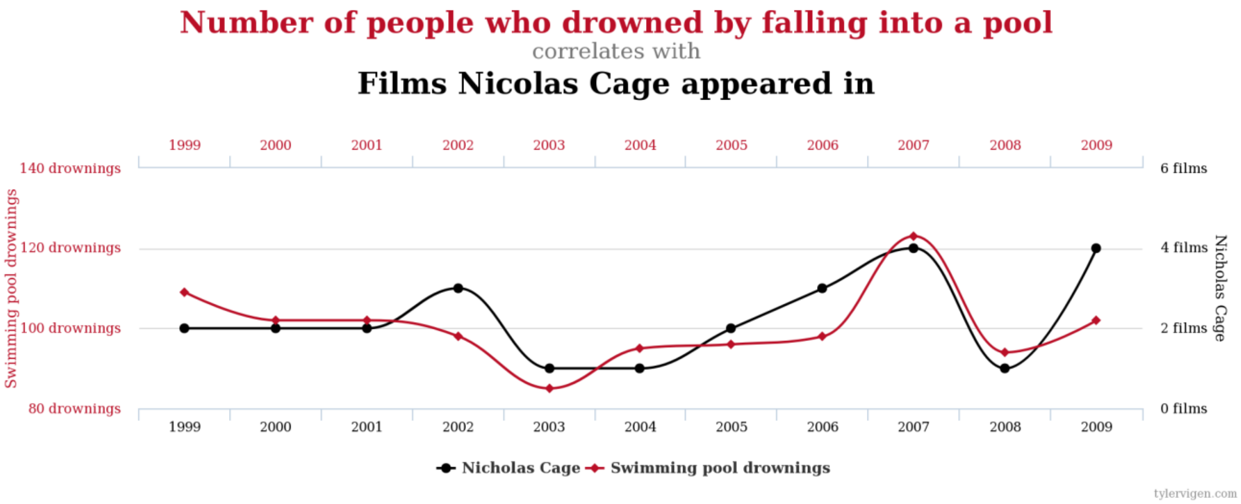
Odds Ratio

$$\frac{A/C}{B/D} = \frac{A \times D}{B \times C}$$

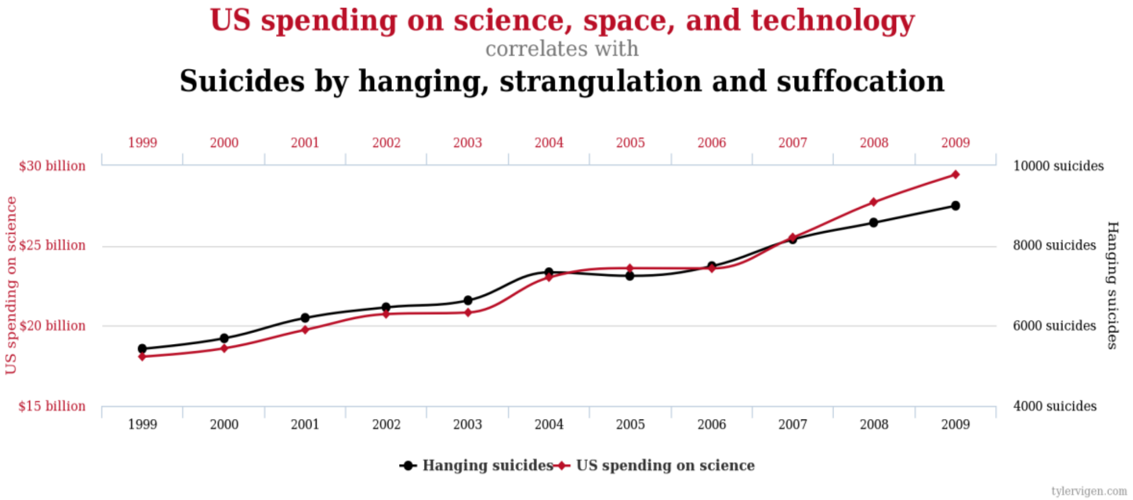
Relative Risk

$$\frac{A/(A+B)}{C/(C+D)}$$

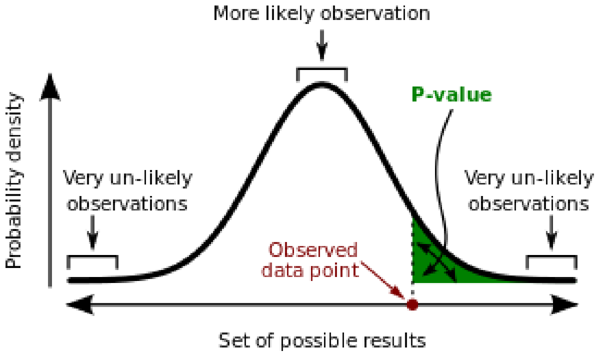
Correlation doesn‘t imply causation!



r ≈ 0.6660



r ≈ 0.9978



Inferential and explorative statistics

Measures of Quality

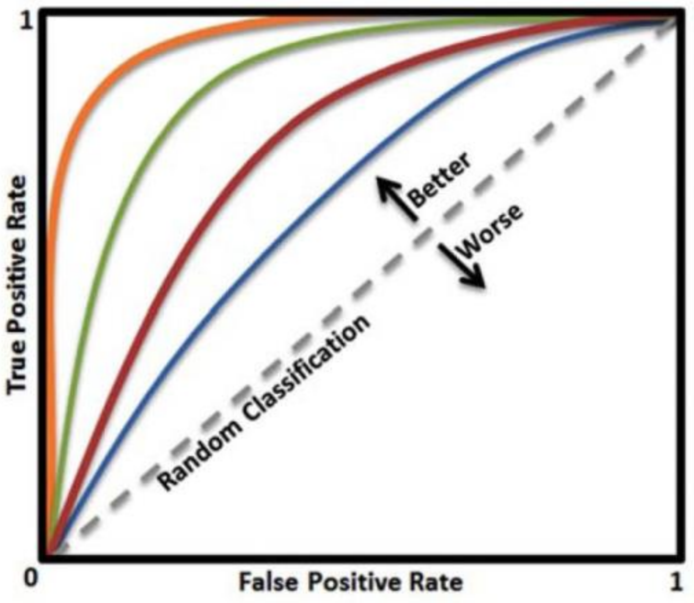
Estimates

- **Confidence intervals** = a range of values within which there is a high probability (95% by convention) that the true population value can be found.

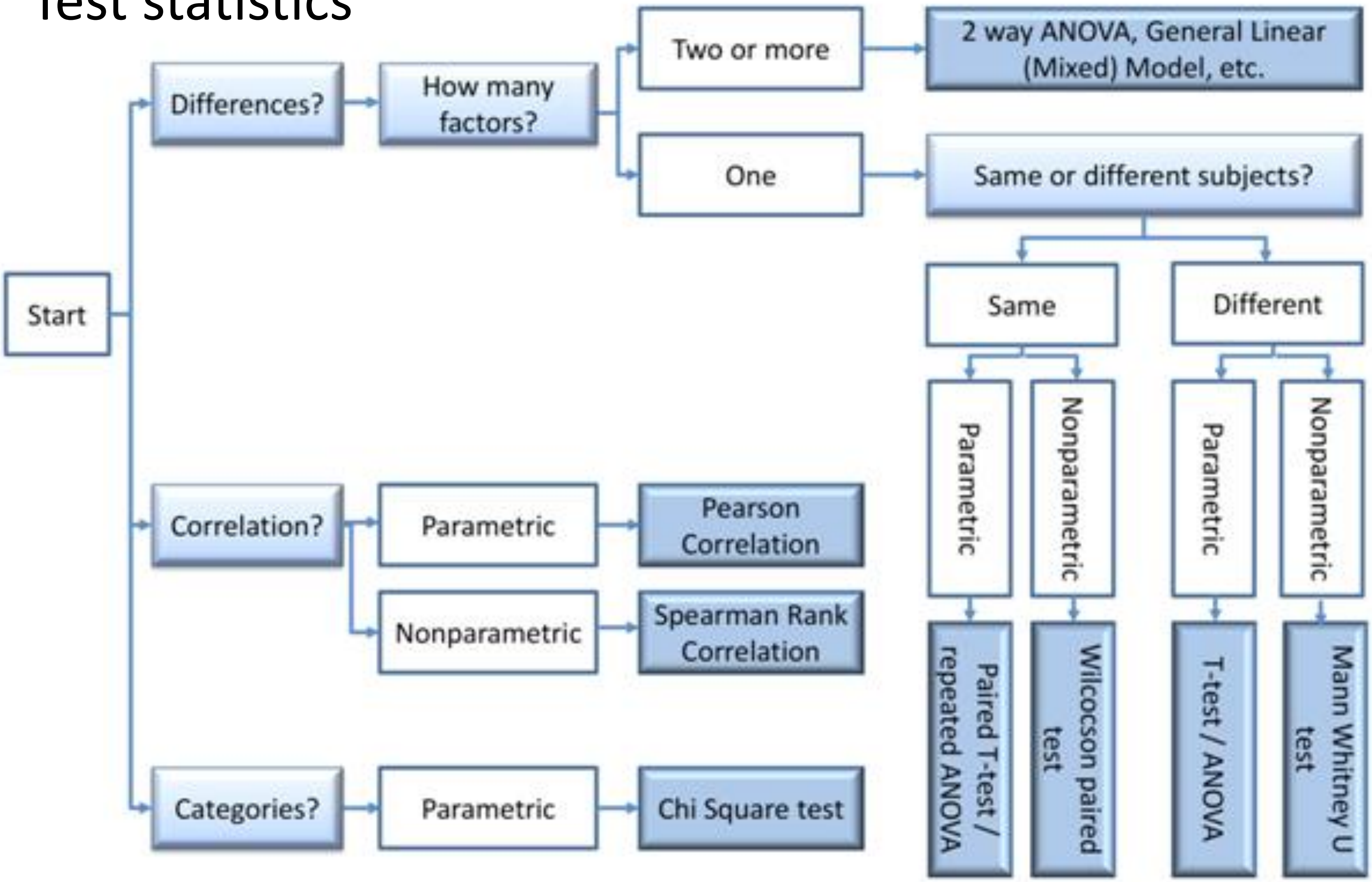
Tests

- **Sensitivity** = the ability of the test to identify correctly those who have the disease (true-positives).

- **Specificity** = the ability of the test to identify correctly those who do not have the disease (true-negatives).



Test statistics



Start

Differences?

How many factors?

Two or more

2 way ANOVA, General Linear (Mixed) Model, etc.

One

Same or different subjects?

Same

Parametric

Paired T-test / repeated ANOVA

Nonparametric

Wilcocson paired test

Different

Parametric

T-test / ANOVA

Nonparametric

Mann Whitney U test

Correlation?

Parametric

Pearson Correlation

Nonparametric

Spearman Rank Correlation

Categories?

Parametric

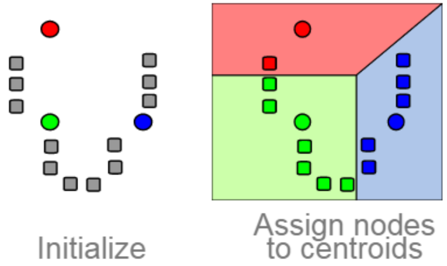
Chi Square test

The P-value is not the probability that the null hypothesis is true given the data.
Absence of evidence is not evidence of absence.

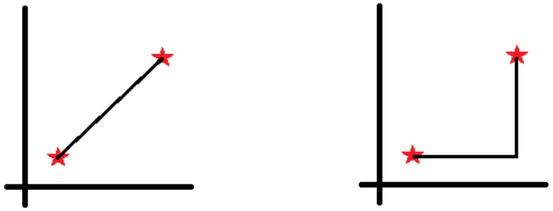
Classification – Cluster analysis

- Grouping of similar observations by distance in p-dimensional system (p = number of variables)

- Different strategies
  - K-means: partition of n observations in k cluster by choosing centroids



  - Hierarchical
    e.g. „Bottom up“: each observation starts in its own cluster, the pair of “closest” clusters are merges



euclidian

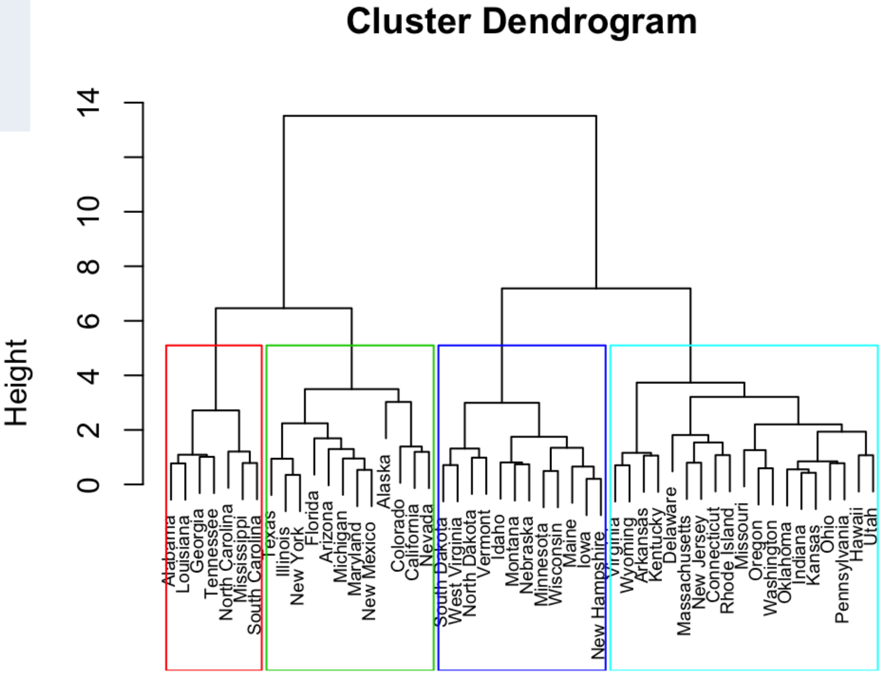
manhattan „taxi cab“

Cluster analysis - example

	Murder	Assault	UrbanPop	Rape	…
Alabama	1,24256408	0,7828393	-0,5209066	-0,00341647	
Alaska	0,50786248	1,1068225	-1,2117642	2,48420294	
Arizona	0,07163341	1,4788032	0,9989801	1,04287839	
Arkansas	0,23234938	0,230868	-1,0735927	-0,1849166	
California	0,27826823	1,2628144	1,7589234	2,06782029	
Colorado	0,02571456	0,3988593	0,8608085	1,86496721	
…					

e.g. euclidian distance

Das Bild kann zurzeit nicht angezeigt werden.



Classification – Principal component analysis (PCA)

- Dimension reduction

- Linear orthogonal transformation:

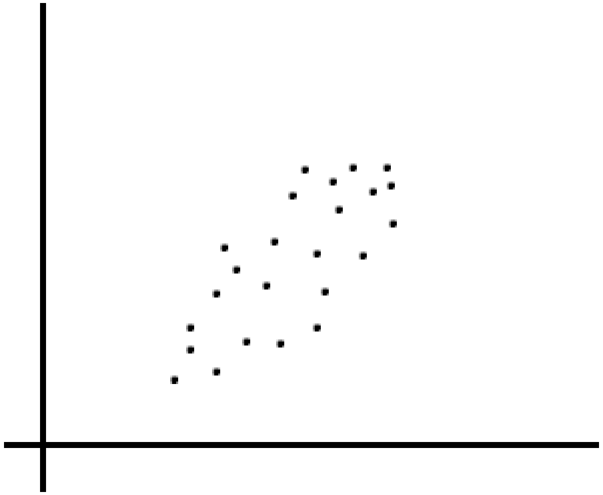
  n possibly correlated variables -> m linearly uncorrelated factors(n > m)
  = orthogonal basis set

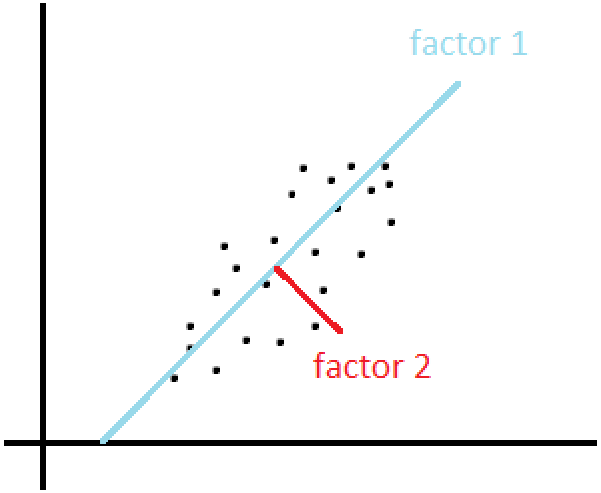
- PCA can be thought of as fitting an n-dimensional ellipsoid to the data.

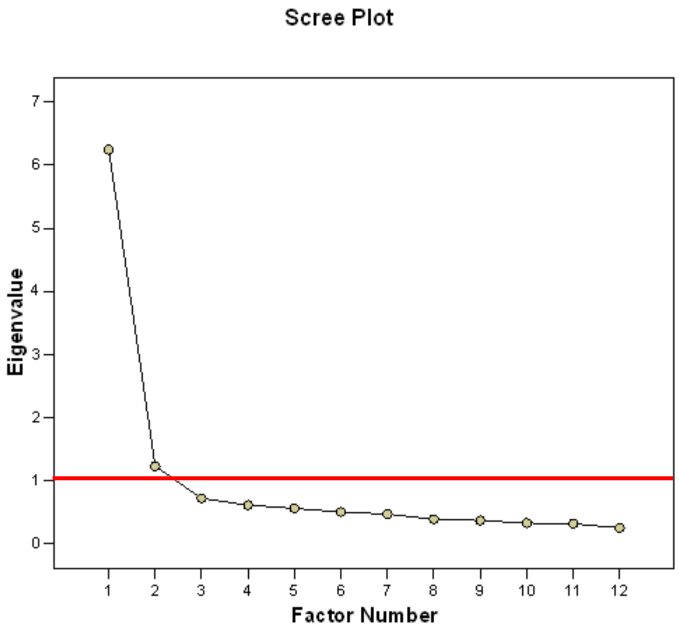
- Largest possible variance explained by first component/factor

- Different strategies to decide how many components are needed:
  - Eigenvalue > 1 (Kaiser rule)
  - variance explained > 90 %
  - Scree-plot „elbow“ (Cattell rule)

PCA - Example

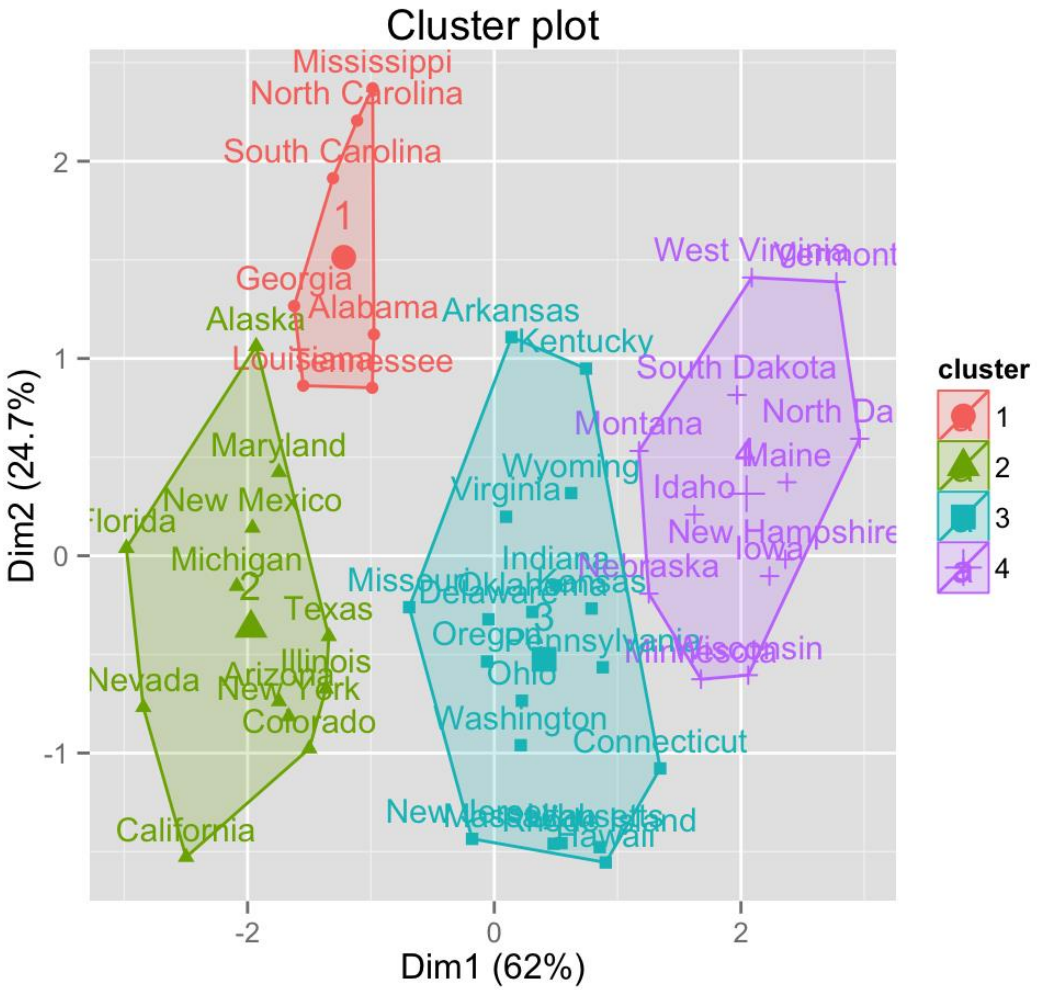






The factor with an eigenvalue of 1 accounts for as much variance as a single variable.

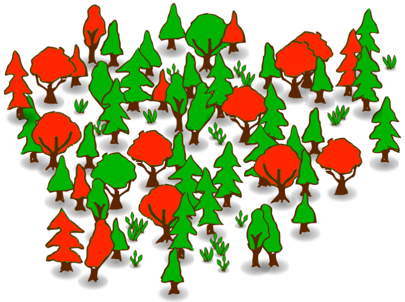
A combined approach – HCPC Analysis

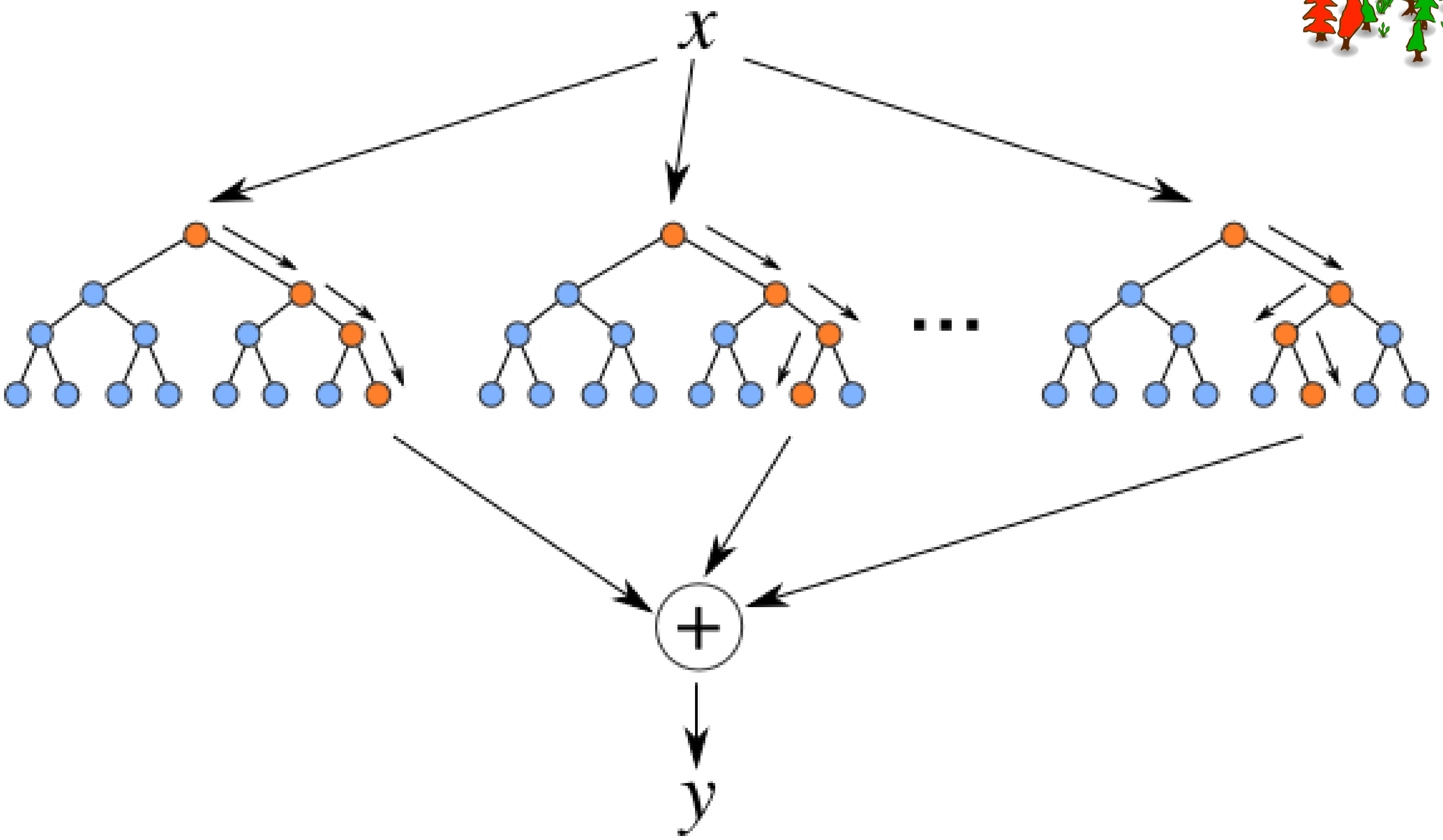


Classification - Random Forrest

- Construction of multiple decision trees (up to several thousands)
- To rank the importance of variables
- „Tree bagging“ - repeatedly fitting trees to random subset
- Output: Mode of classes or mean of prediction
- While the predictions of a single tree are highly sensitive to noise, the average of many (uncorrelated) trees is not.

Classification - Random Forrest



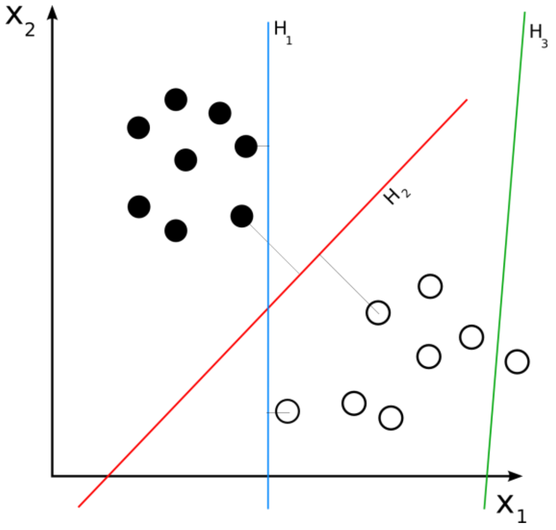


Support vector machines

- Large margin classifier

- Strategy: Find a hyperplane (p-1 dimensional space) which
  1) Separates the two classes,
  2) Maximizes the margin (distance to closest data points).

- linearly separable or up-transformation



H_2 is the correct solution

Regression / modeling – generalized linear model

- Many subtypes – here just an introduction – to study relationships.

- $g(y) = X\beta + \varepsilon$ observation y and predictor(s) X

- Goal: Prediction of future observations, assessment of effects of predictors on observation or general description of data.

- (Pseudo) coefficient of determination to judge, how well the model fit the data

- In the generalized case, y is continous. What if y is binary (health/disease)?
  - (binary) logistic regression
  - Death/suvival processes

Regression / modeling – overfitting

- We want to explain the data in the simplest way

- Unnecessary predictors will add noise to the estimation of other quantities that we are interested in

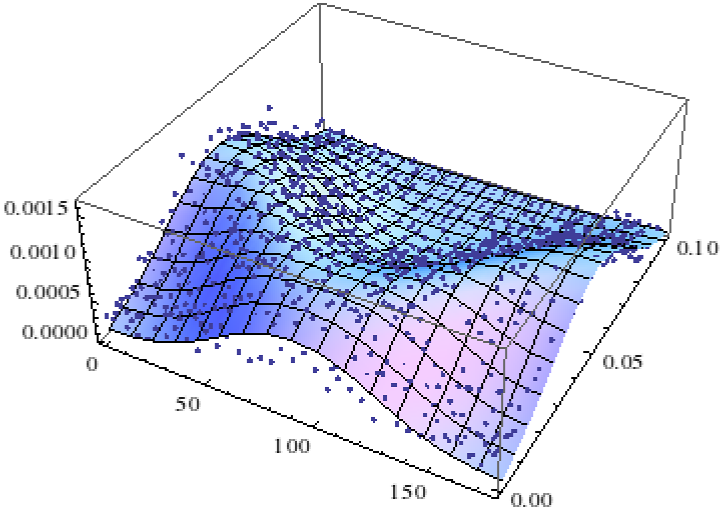
- Collinearity is caused by having too many variables trying to do the same job

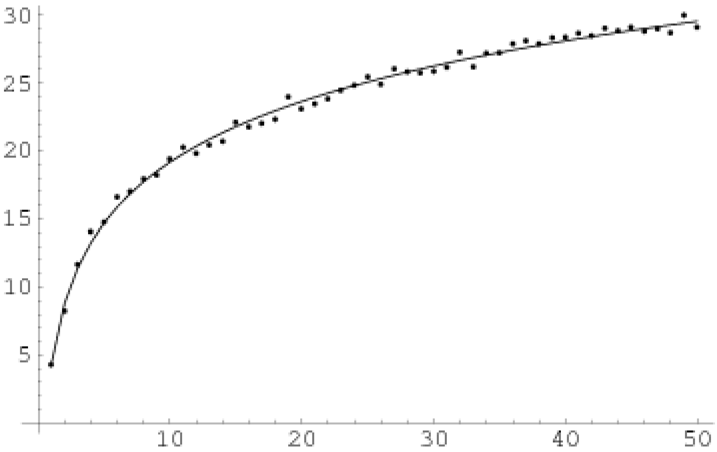
- Procedures: Backward Elimination/Forward Selection

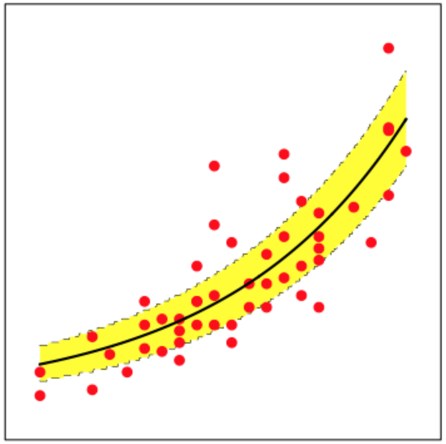
- Decision to keep / drop variables based on
  - Hypothesis tests
  - Information criteria (AIC, BIC)

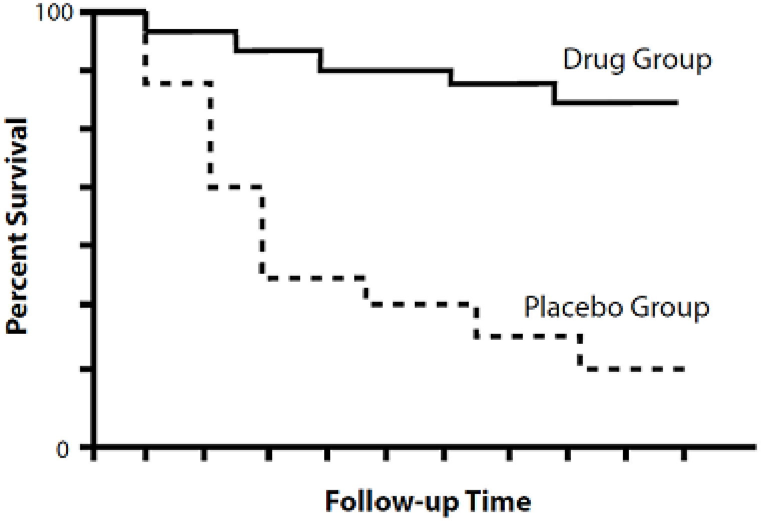
- All procedures are heuristics, so try out!

Regression / modeling – different examples









Regression / modeling – Lasso regrularization

- lasso (least absolute shrinkage and selection operator)

$$\min_{\alpha,\beta} \frac{1}{N} \sum_{i=1}^{N} f(x_i, y_i, \alpha, \beta) \text{ to subject } \|\beta\|_1 \leq t$$

- Performs both variable selection (can set $\beta_i = 0$) and regularization

- Enhance the prediction accuracy and interpretability

Known problem - Multiple Testing

- Consider a case where you have 1000 hypotheses to test, and a significance level of $\alpha = 0.05$

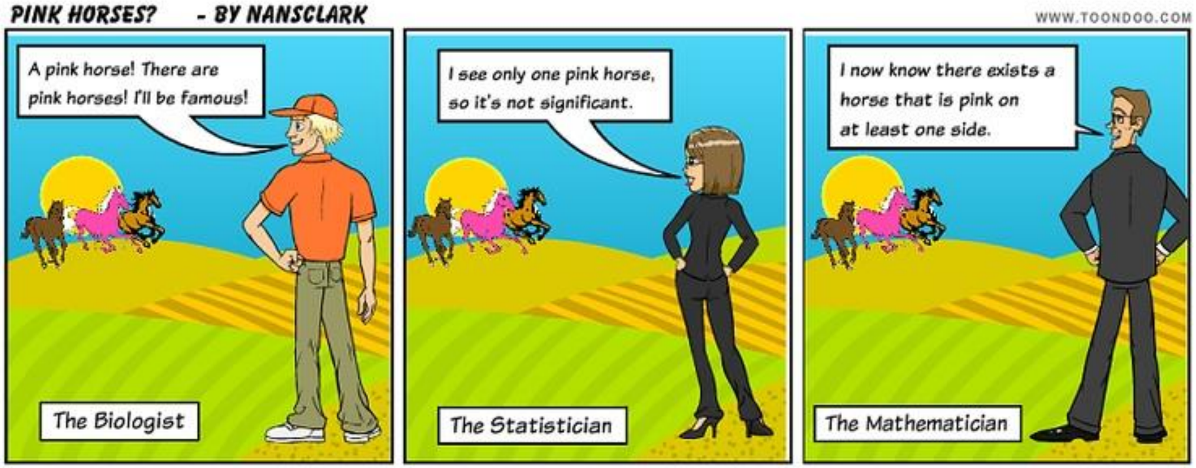
⇓

50 significant hypotheses just by chance

- Different strategies to adjust α (for n hypotheses)
  - Bonferroni: $\alpha_{new} = \alpha/n$
  - Benjamini-Hochberg: control of the false discovery rate
  - Holm
  - …

Biostatistics in the context of polygenic risk scores

Different languages…


PINK HORSES? - BY NANSCLARK
WWW.TOONDOO.COM
A pink horse! There are pink horses! I'll be famous!
The Biologist
I see only one pink horse, so it's not significant.
The Statistician
I now know there exists a horse that is pink on at least one side.
The Mathematician

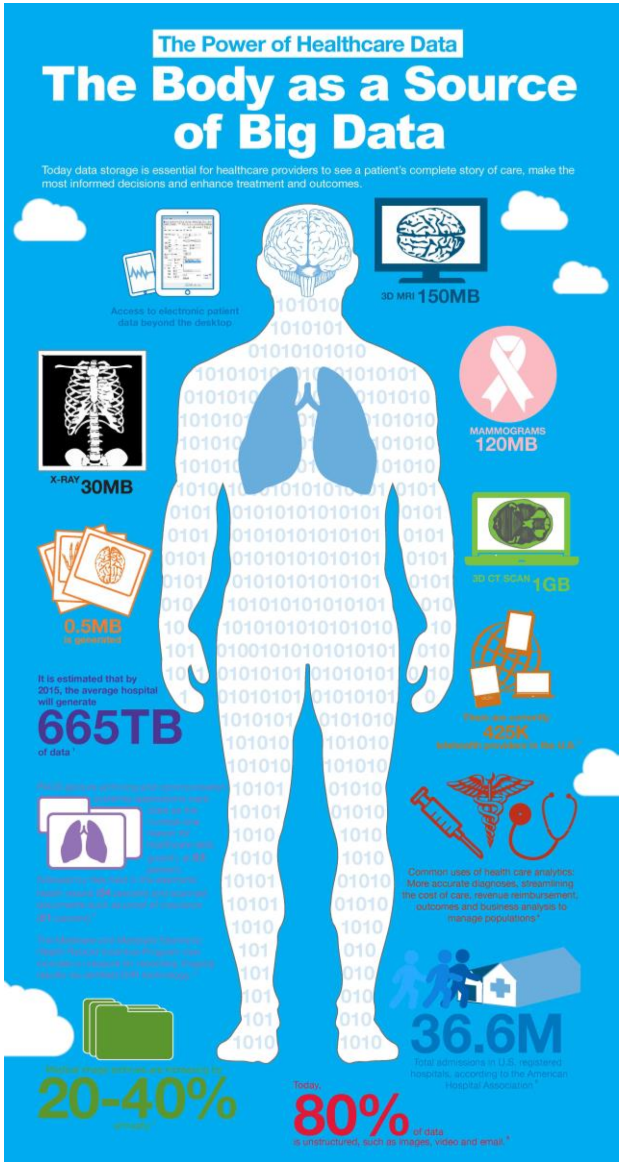
Nevertheless, it is not all about significance – especially for big data.

21st Century – The Century of big data

- 90% of the data in the world today has been created in the last two years.

- Challenge:
  Traditional methods (and data processing software) are inadequate to deal with them

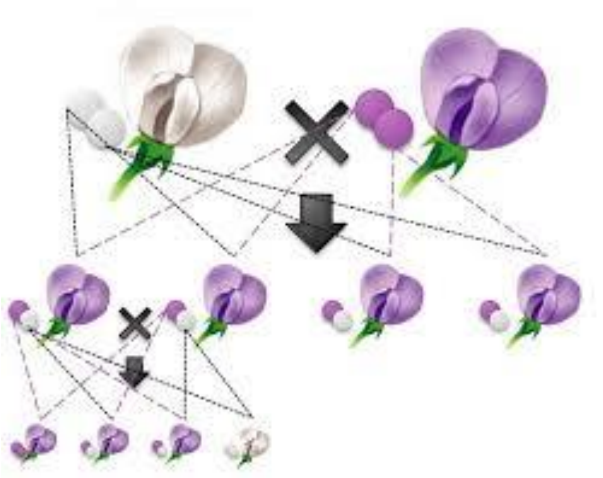
- Where do we find big data?
  e.g. in (epi)genetics…
  - Single nucleotide polymorphisms (SNPs)
  - Next generation sequencing
  - Histone modifications/DNA methylation

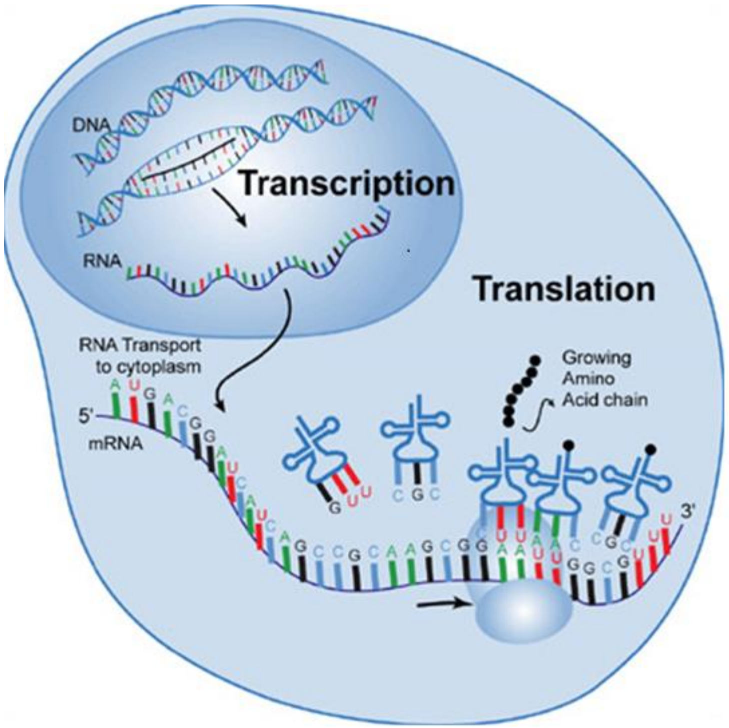

The Power of Healthcare Data
The Body as a Source of Big Data
Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.
Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 30MB
MAMMOGRAMS 120MB
3D CT SCAN 1GB
0.5MB is generated
It is estimated that by 2015, the average hospital will generate 665TB of data
There are currently 425K telehealth providers in the U.S.
Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations
36.6M
Total admissions in U.S. registered hospitals, according to the American Hospital Association
Medical image archives are increasing by 20-40% annually
Today, 80% of data is unstructured, such as images, video and email.

The central concept of gene-environment interactions

- It all starts with…peas and a monk (G. Mendel)

- Phenotype = Genotype + Environment

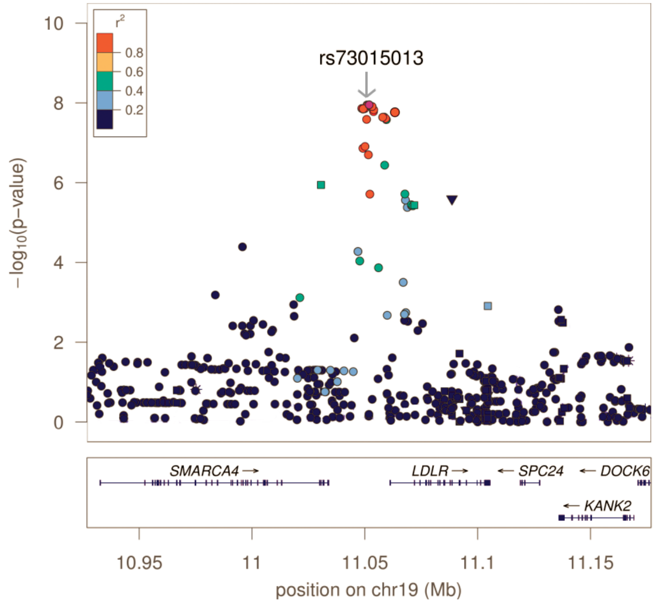




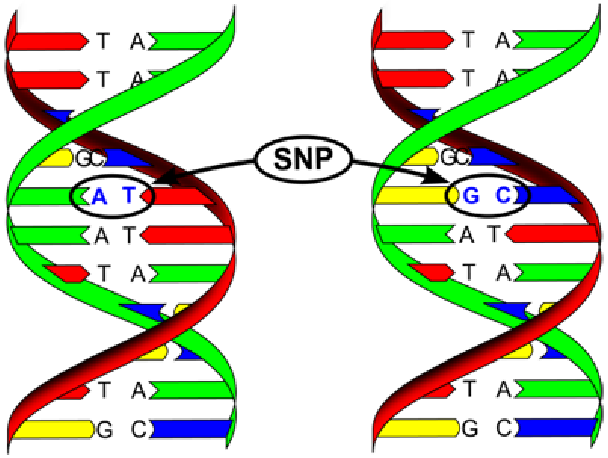
- 100 years later:
  Transcription + Translation and their regulation as key players of gene-environment interactions

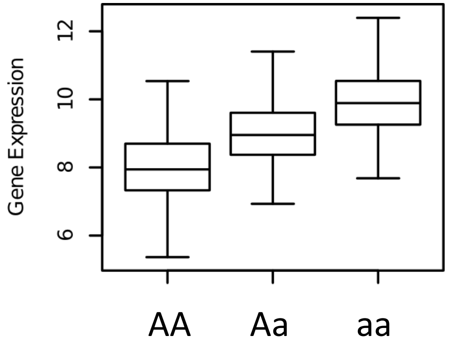
- DNA -> mRNA -> Protein

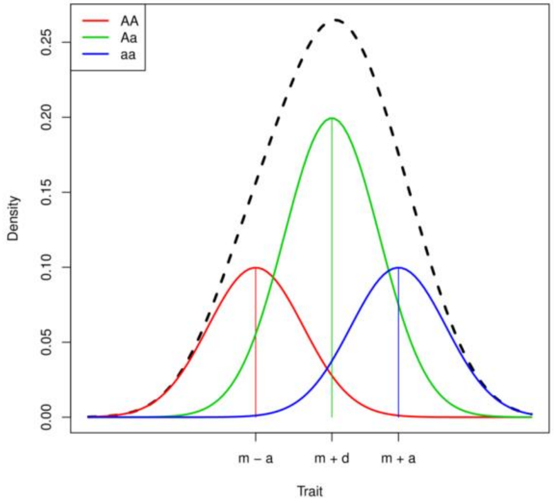
SNPs – Single nucleotide polymorphisms



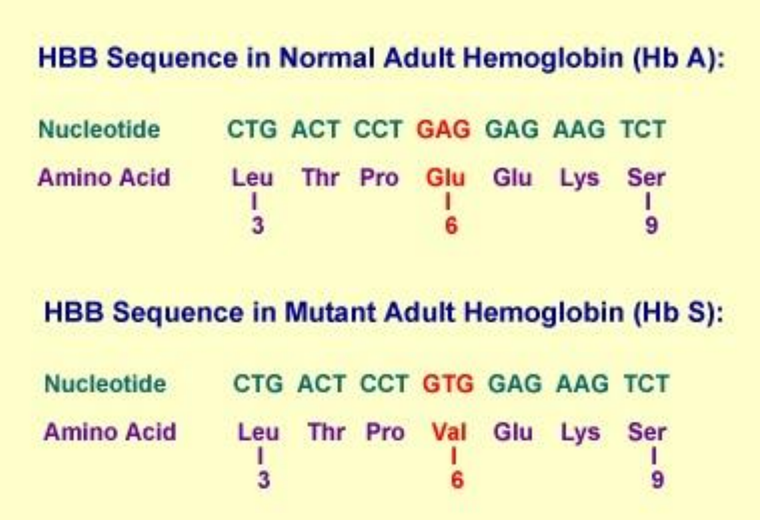
Kettunen et al. (2012) Nat. Genet.







SNPs – Example (Sickle-cell anaemia)

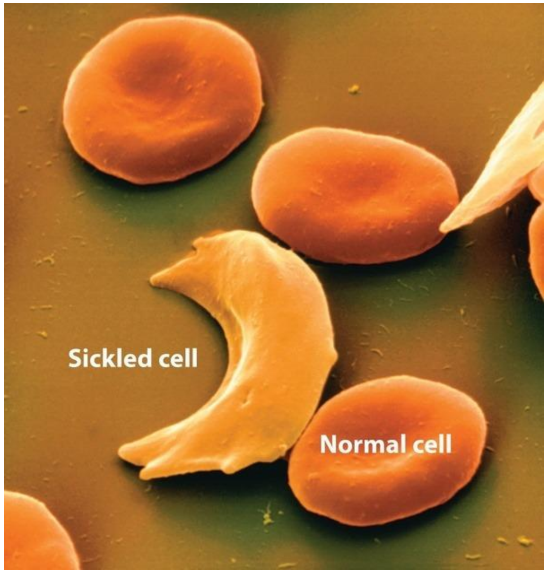


HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu 3	Thr	Pro	Glu 6	Glu	Lys	Ser 9

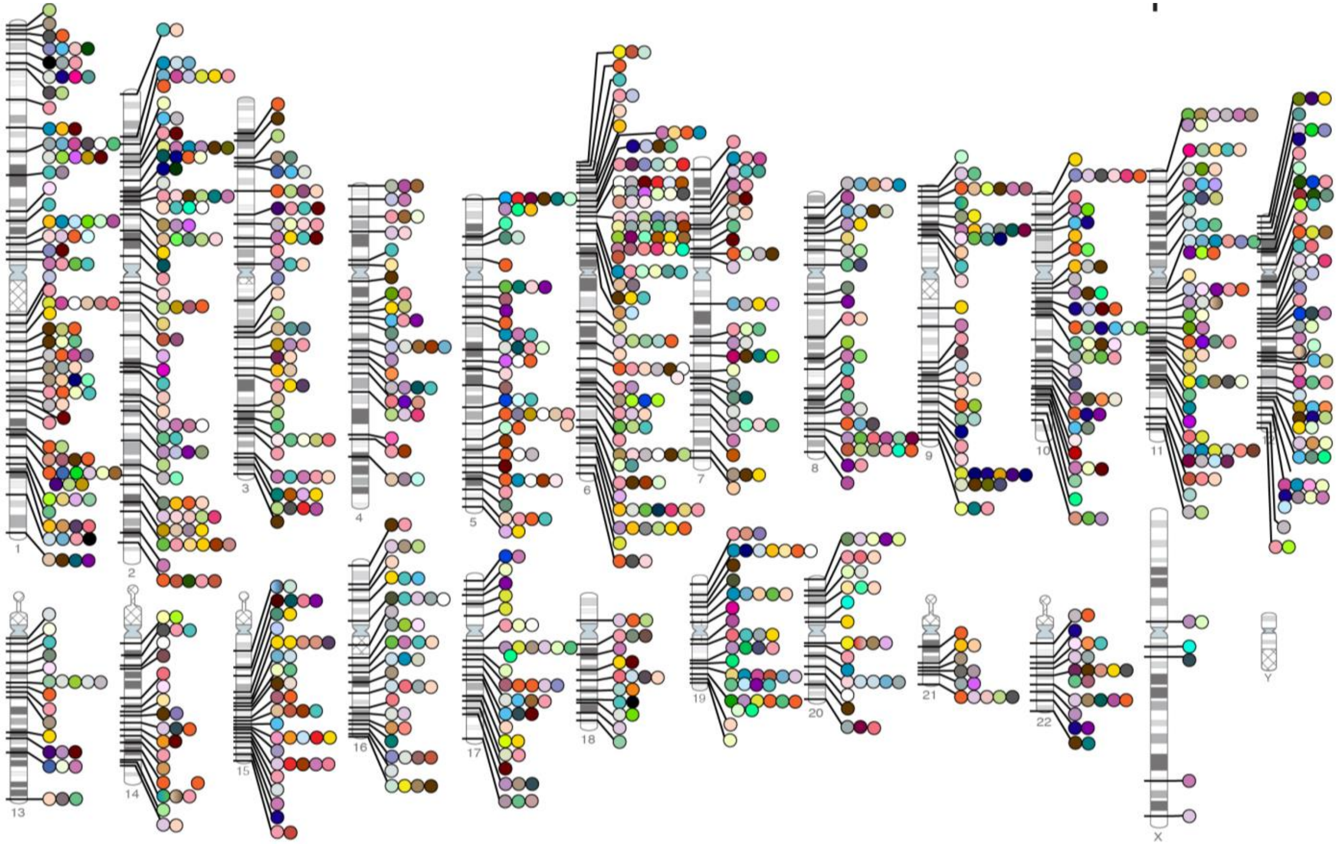
HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu 3	Thr	Pro	Val 6	Glu	Lys	Ser 9



attacks of pain ("sickle-cell crisis")
anemia
swelling in the hands and feet
bacterial infections
stroke
higher risk of infection

Each Disease, another SNP…



Each color is another disease/disorder

What is a risk score?

- In mathematical terms: A RS is a classifier!

In statistical learning, we want to infer to unknown relationship f(·) in:

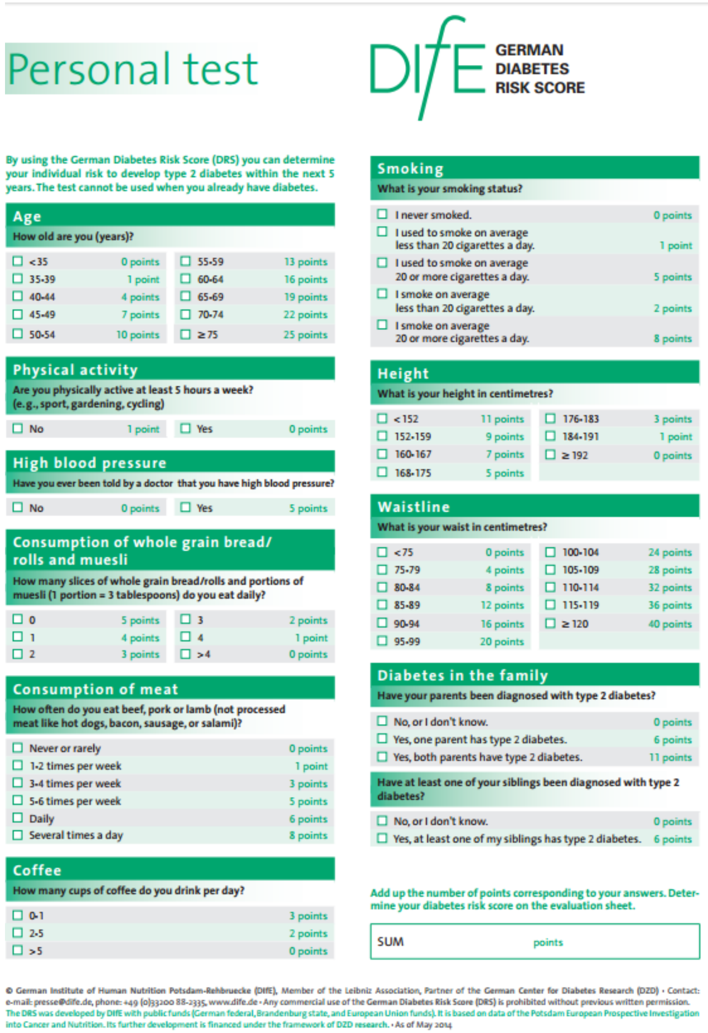
$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\epsilon}$$

where **y** [sometimes g(y)] is the outcome, **X** is your data matrix, and **ε** is the error.

In classification, y is binary ➔ yes/no, 1/0, true/false, healthy/sick

f(X) can have many forms, as we will see in the following

The easiest way of risk prediction



Personal test

DIfE GERMAN DIABETES RISK SCORE

By using the German Diabetes Risk Score (DRS) you can determine your individual risk to develop type 2 diabetes within the next 5 years. The test cannot be used when you already have diabetes.

Age

How old are you (years)?

□ <35	0 points	□ 55-59	13 points
□ 35-39	1 point	□ 60-64	16 points
□ 40-44	4 points	□ 65-69	19 points
□ 45-49	7 points	□ 70-74	22 points
□ 50-54	10 points	□ ≥ 75	25 points

Physical activity

Are you physically active at least 5 hours a week? (e.g., sport, gardening, cycling)

□ No	1 point	□ Yes	0 points

High blood pressure

Have you ever been told by a doctor that you have high blood pressure?

□ No	0 points	□ Yes	5 points

Consumption of whole grain bread/ rolls and muesli

How many slices of whole grain bread/rolls and portions of muesli (1 portion = 3 tablespoons) do you eat daily?

□ 0	5 points	□ 3	2 points
□ 1	4 points	□ 4	1 point
□ 2	3 points	□ >4	0 points

Consumption of meat

How often do you eat beef, pork or lamb (not processed meat like hot dogs, bacon, sausage, or salami)?

□ Never or rarely	0 points
□ 1-2 times per week	1 point
□ 3-4 times per week	3 points
□ 5-6 times per week	5 points
□ Daily	6 points
□ Several times a day	8 points

!

Coffee

How many cups of coffee do you drink per day?

□ 0-1	3 points
□ 2-5	2 points
□ >5	0 points

Smoking

What is your smoking status?

□ I never smoked.	0 points
□ I used to smoke on average less than 20 cigarettes a day.	1 point
□ I used to smoke on average 20 or more cigarettes a day.	5 points
□ I smoke on average less than 20 cigarettes a day.	2 points
□ I smoke on average 20 or more cigarettes a day.	8 points

Height

What is your height in centimetres?

□ < 152	11 points	□ 176-183	3 points
□ 152-159	9 points	□ 184-191	1 point
□ 160-167	7 points	□ ≥ 192	0 points
□ 168-175	5 points		

Waistline

What is your waist in centimetres?

□ <75	0 points	□ 100-104	24 points
□ 75-79	4 points	□ 105-109	28 points
□ 80-84	8 points	□ 110-114	32 points
□ 85-89	12 points	□ 115-119	36 points
□ 90-94	16 points	□ ≥ 120	40 points
□ 95-99	20 points		

Diabetes in the family

Have your parents been diagnosed with type 2 diabetes?

□ No, or I don't know.	0 points
□ Yes, one parent has type 2 diabetes.	6 points
□ Yes, both parents have type 2 diabetes.	11 points

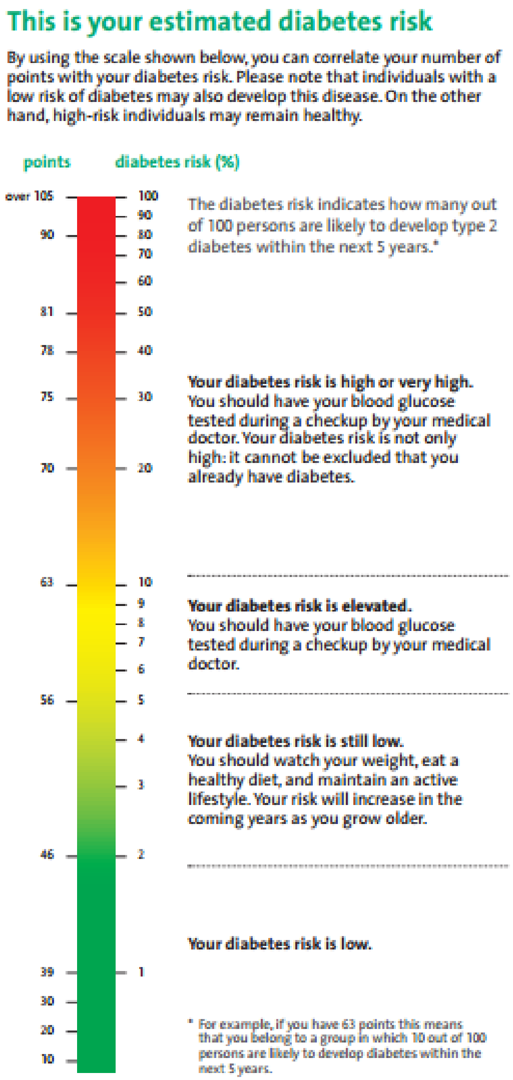
Have at least one of your siblings been diagnosed with type 2 diabetes?

□ No, or I don't know.	0 points
□ Yes, at least one of my siblings has type 2 diabetes.	6 points

Add up the number of points corresponding to your answers. Determine your diabetes risk score on the evaluation sheet.

SUM	points

© German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Member of the Leibniz Association, Partner of the German Center for Diabetes Research (DZD) • Contact: e-mail: presse@dife.de, phone: +49 (0)33200 88-2335, www.dife.de • Any commercial use of the German Diabetes Risk Score (DRS) is prohibited without previous written permission. The DRS was developed by DIfE with public funds (German federal, Brandenburg state, and European Union funds). It is based on data of the Potsdam European Prospective Investigation into Cancer and Nutrition. Its further development is financed under the framework of DZD research. • As of May 2014



This is your estimated diabetes risk

By using the scale shown below, you can correlate your number of points with your diabetes risk. Please note that individuals with a low risk of diabetes may also develop this disease. On the other hand, high-risk individuals may remain healthy.

The diabetes risk indicates how many out of 100 persons are likely to develop type 2 diabetes within the next 5 years.*

Your diabetes risk is high or very high. You should have your blood glucose tested during a checkup by your medical doctor. Your diabetes risk is not only high: it cannot be excluded that you already have diabetes.

Your diabetes risk is elevated. You should have your blood glucose tested during a checkup by your medical doctor.

Your diabetes risk is still low. You should watch your weight, eat a healthy diet, and maintain an active lifestyle. Your risk will increase in the coming years as you grow older.

Your diabetes risk is low.

* For example, if you have 63 points this means that you belong to a group in which 10 out of 100 persons are likely to develop diabetes within the next 5 years.

If it‘s not diabetes, then it‘s dementia…

What is a polygenic risk score?

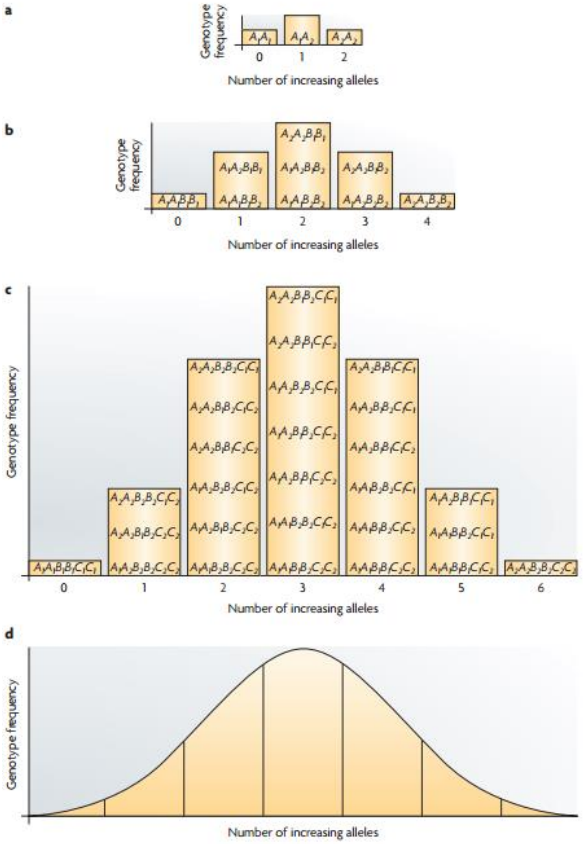
- Adding genetic informations to a risk score.
  “polygenic” = “more than one gene”

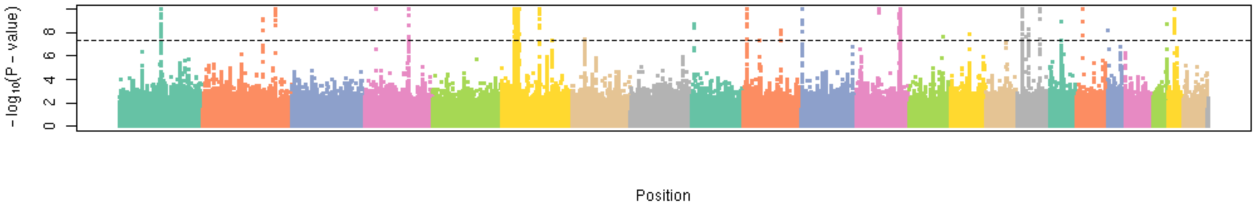
- Diseases/disorders (especially their risk) are quantitative traits!

- What does this mean? → additive genetic effects

- Identifying the variants, that are causing the disease.

- GWAS = genome-wide association studies.





Aim of polygenic risk scores

We want to predict patients at high risk for a disease based on clinical **and** genetic data

What is the benefit?

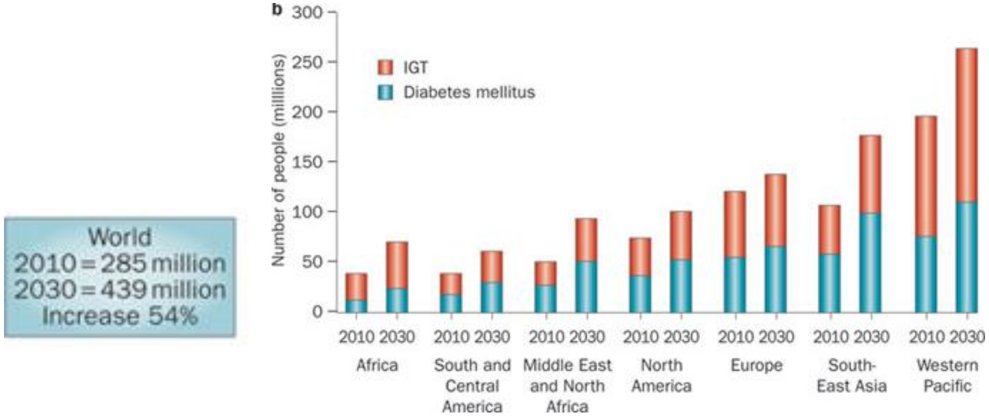
If we are able to identify patients at high risk, we are able to…

- individualize the therapy,
- start therapy quiet early or (in best case) prevent the disease outbreak,
- (maybe) understand the mechanisms behind the disease/disorder,
- reduce the economic burden, e.g

$174 billion in 2007
$245 billion in 2012

only in the U.S. and only Type II Diabetes
(source: ADA)

Chen, L. *et al.* (2011) *Nat. Rev. Endocrinol.*



Incidents vs. prevalence

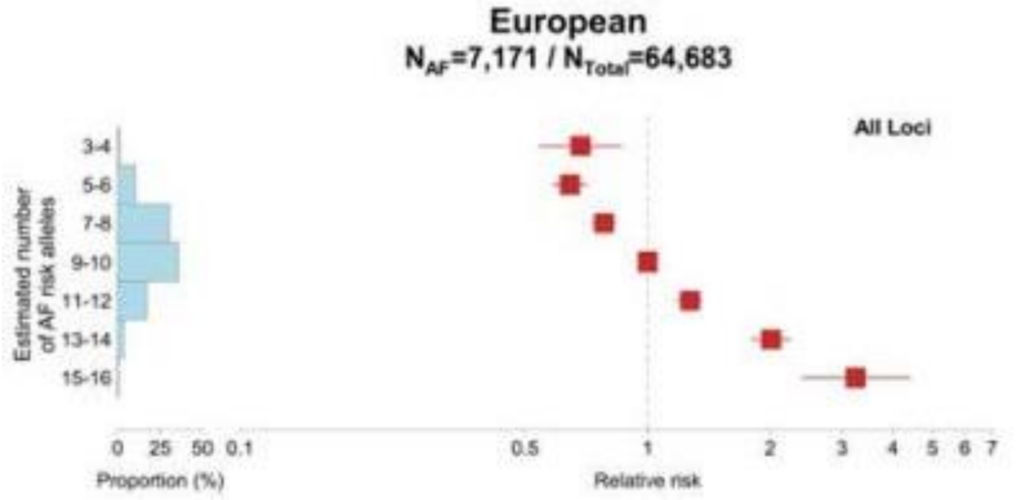
- **Incidence** The number of new events (e.g. death or a particular disease) that occur during a specified period of time in a population at risk for developing the events.

- **Prevalence** The number of persons in the population affected by a disease at a specific time divided by the number of persons in the population at the time.

The easiest way of a PRS

- Sum up the risk allells

- Example: Atrial Fibrillation



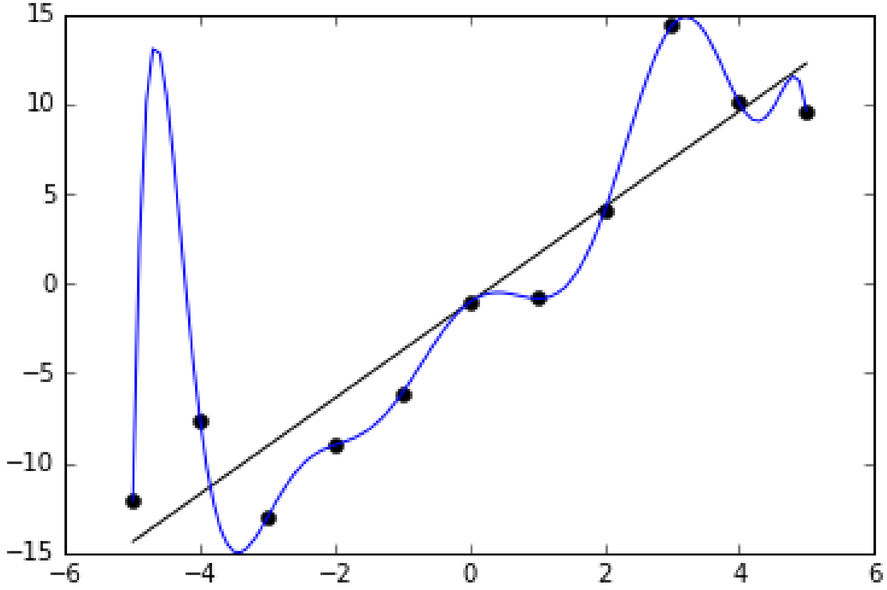
- Problem: Quiet different for each outcome/disease and subpopulation (sex/ethnicity/…).

Problems – Overfitting

- Multicolinearity = high intercorrelation between the predictors
  -> the information of one predictor might be contained in another one

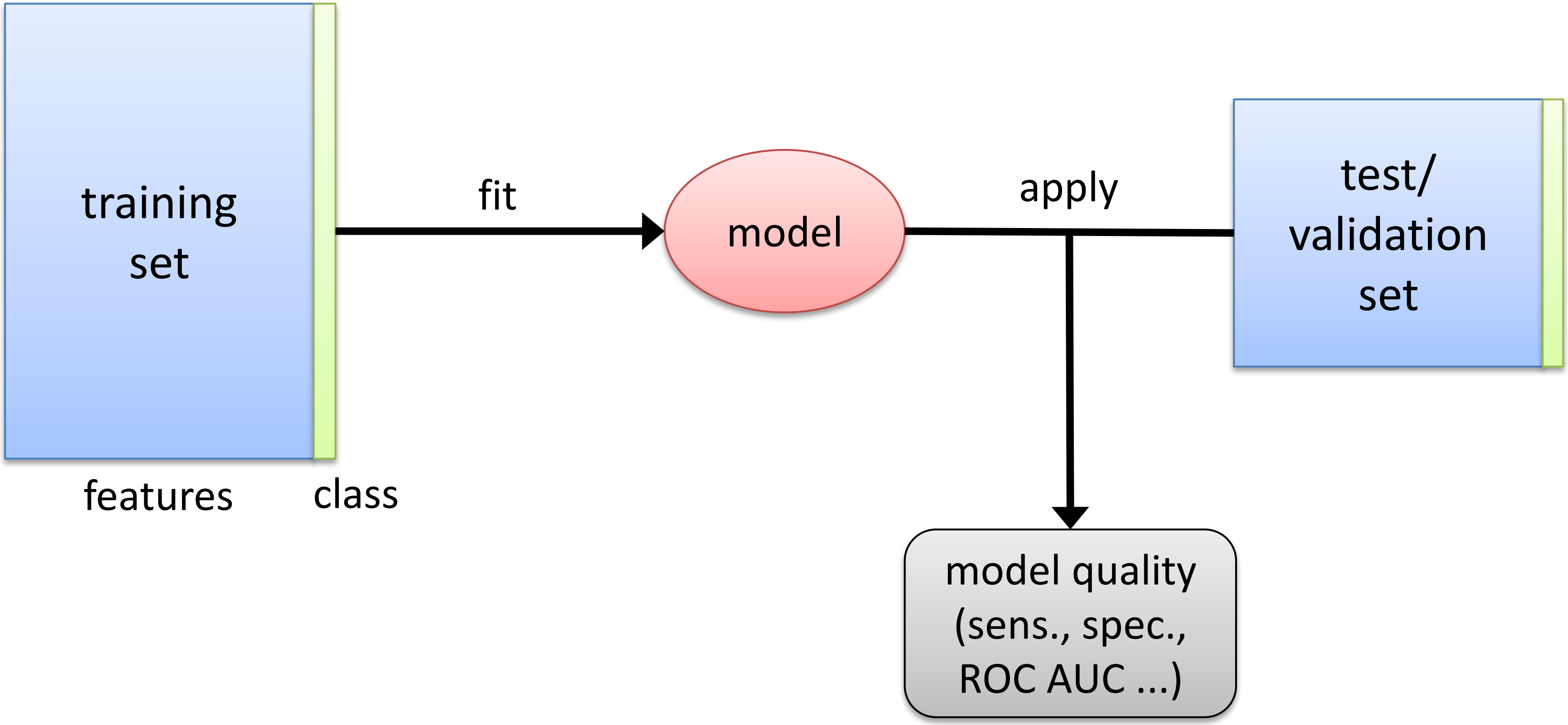
- These classical statistical techniques were developed for low dimensional data (n >> p).

- The blue curve fits the data points too well.



Validation

Ideal case:



if the model learned a pattern that does not generalize, the quality on the test set will be low

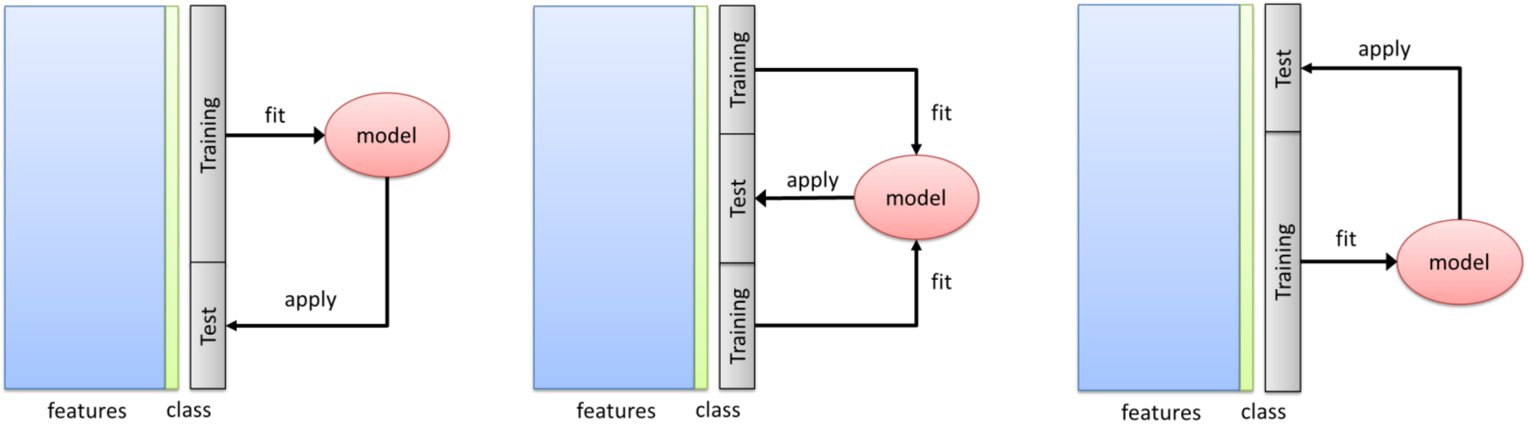
Cross validation

Main motivation for cross-validation:

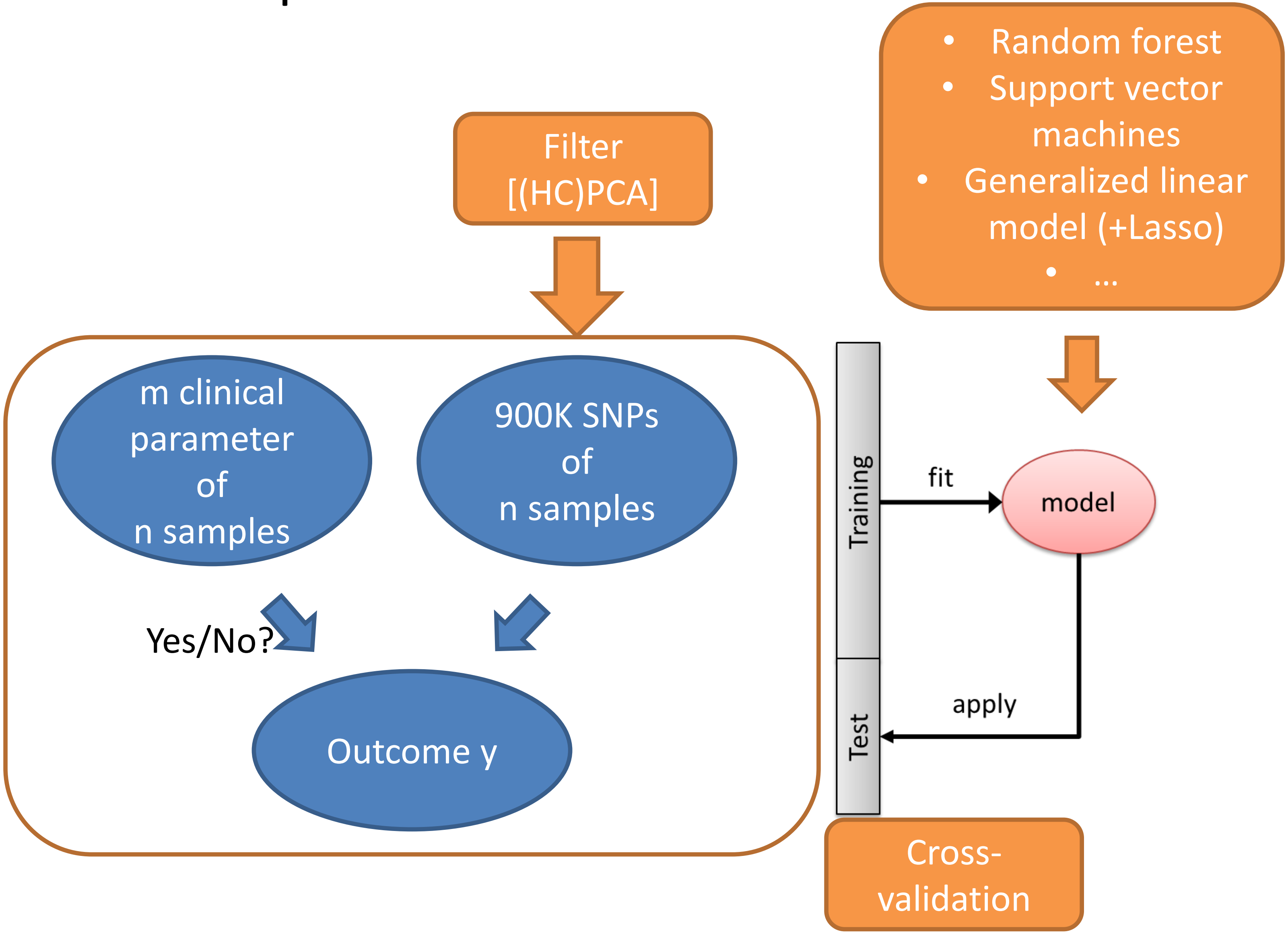
In reality, we can often not afford to leave out a substantial part of the data for validation.

Can we perform validation just using the training data?

Example: 3-fold cross-validation



Scheme of risk prediction



Thank you for your attention!