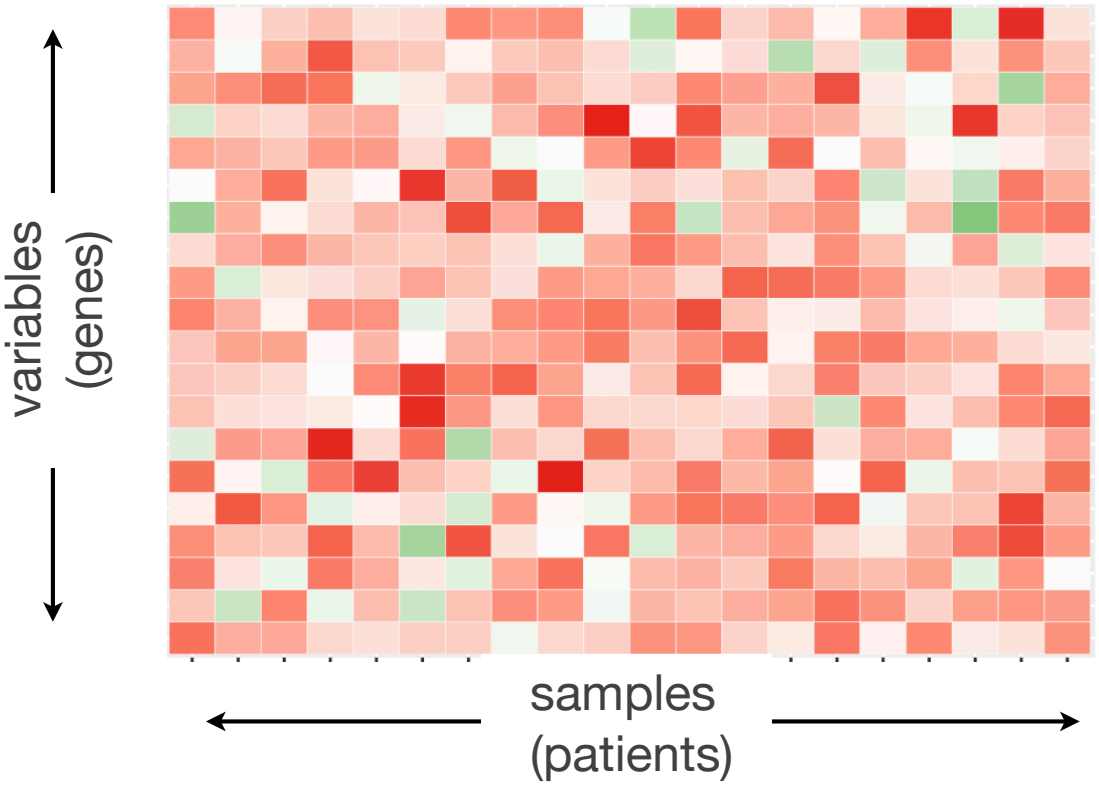
3. finding structure in the data

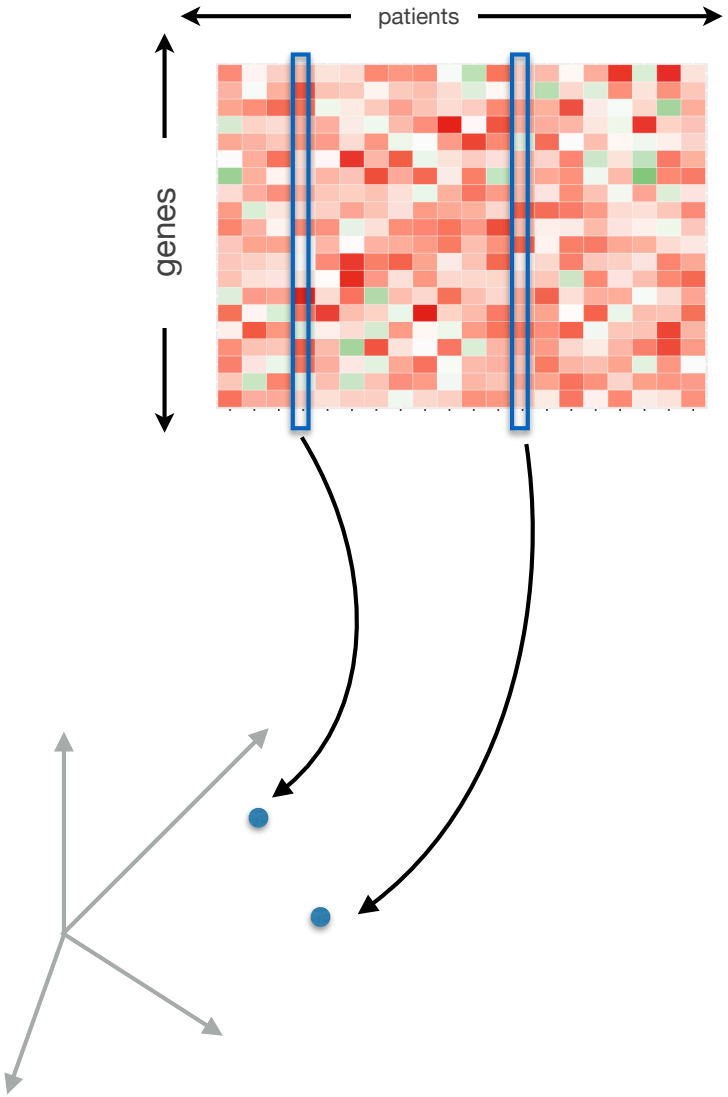
Finding structure in the data

Medizinische Fakultät Heidelberg

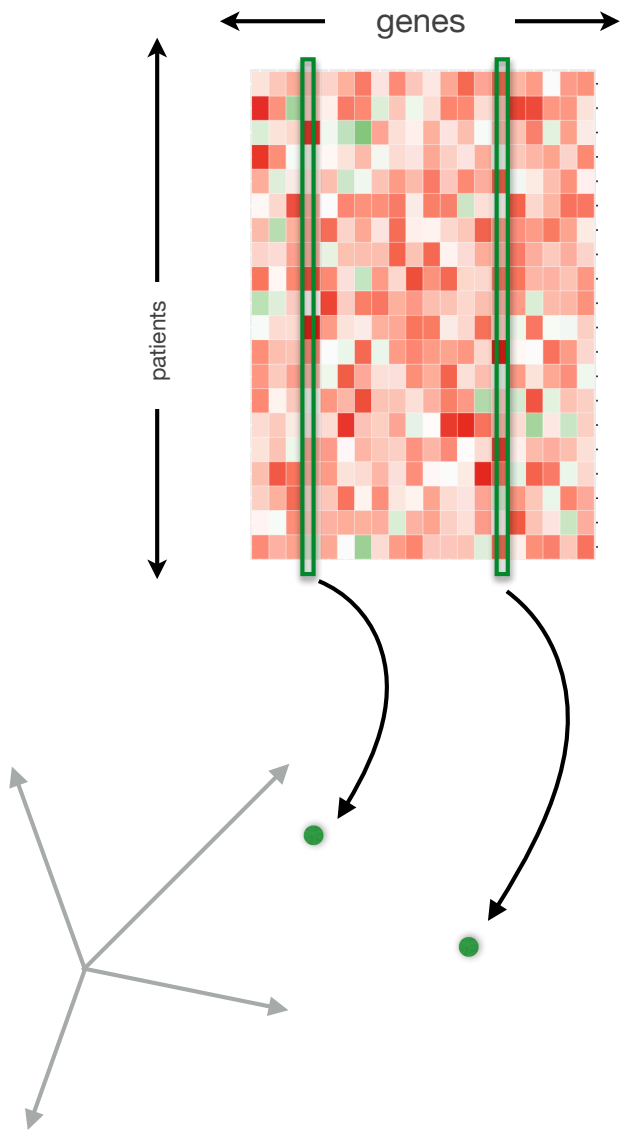
- So far, we have looked at individual variables (distribution, mean, spread,…) and the relationship between **variables**
- We now want to find a structure between **samples** (groups!)

  - group similar samples
    → **clustering**
  - show variables with high variance
    → **principal component analysis**





Points = **patients** in 20,000 dimensional **gene space**



Points = **genes** in 500 dimensional **patient space**

Distances / dissimilarities

Medizinische Fakultät Heidelberg

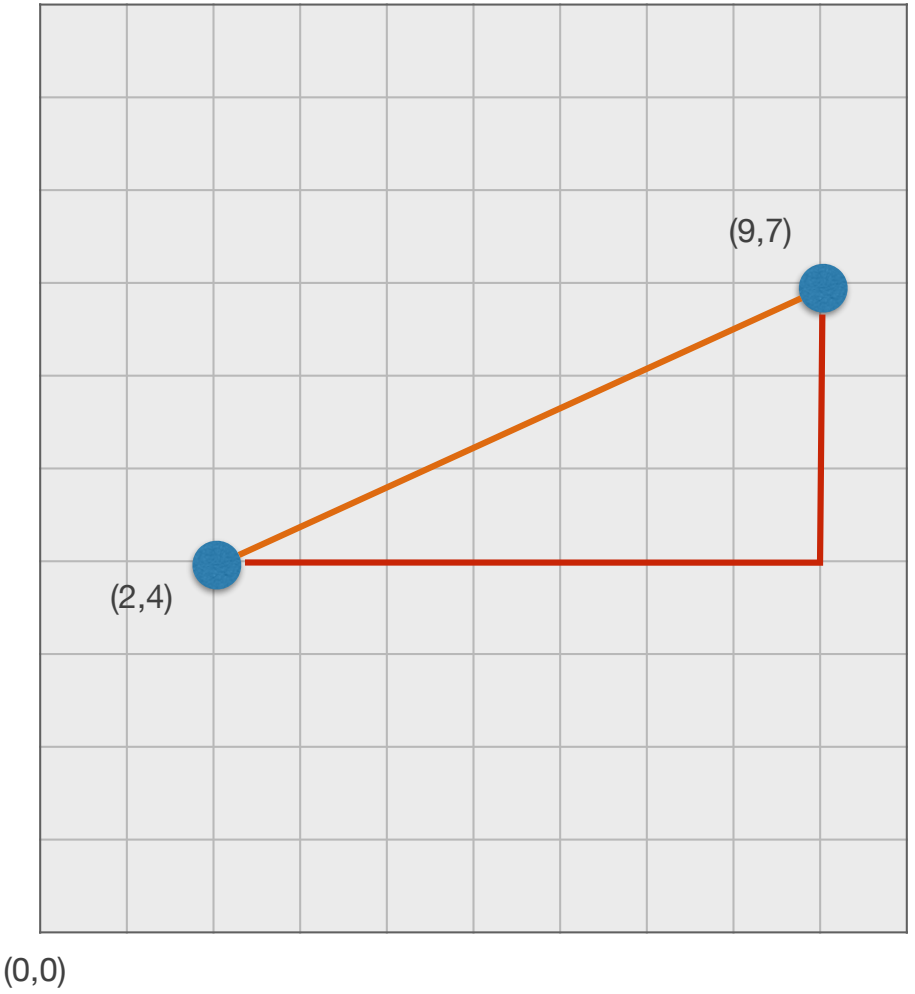
- Distances should be
  - positively defined : *d(a,b) >= 0*
  - *d(a,a) = 0*
  - d(a,c) < d(a,b) + d(b,c)
- Several distances can be defined in n-dimensional space
  - **Euclidean distance**

    $d^2 = (9-2)^2+(7-4)^2 = 58$
  - **Manhattan distance**

    d = 7 + 3 = 10
  - **Correlation distance** (when dimension >2)

$$d = \frac{1}{2}(1 - r)$$

$$r = \sum_{i=1,2} \frac{(x_i - \bar{x})(y_i - \bar{y})}{sd(x)sd(y)}$$



Can be generalized to any number of dimensions!

Distances / dissimilarities in any dimensions

Medizinische Fakultät Heidelberg

- Euclidean distance

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan distance

$$d_{Manhattan}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- Correlation distance

$$r = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{sd(x)sd(y)} \qquad d_{corr}(x, y) = \frac{1}{2}(1 - r)$$

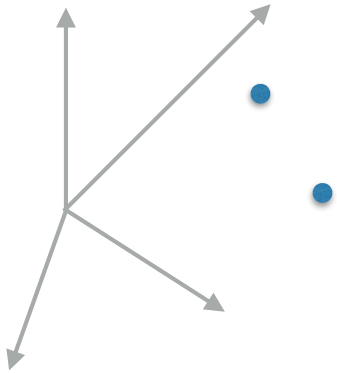
- Maximum distance

$$d_{Max}(x, y) = max\,|x_i - y_i|$$

Example

Medizinische Fakultät Heidelberg

	variable 1	variable 2	variable 3	variable 4
x	-0.7152422	-0.7526890	-0.9385387	-1.0525133
y	-0.4371595	0.3311792	-2.0142105	0.2119804

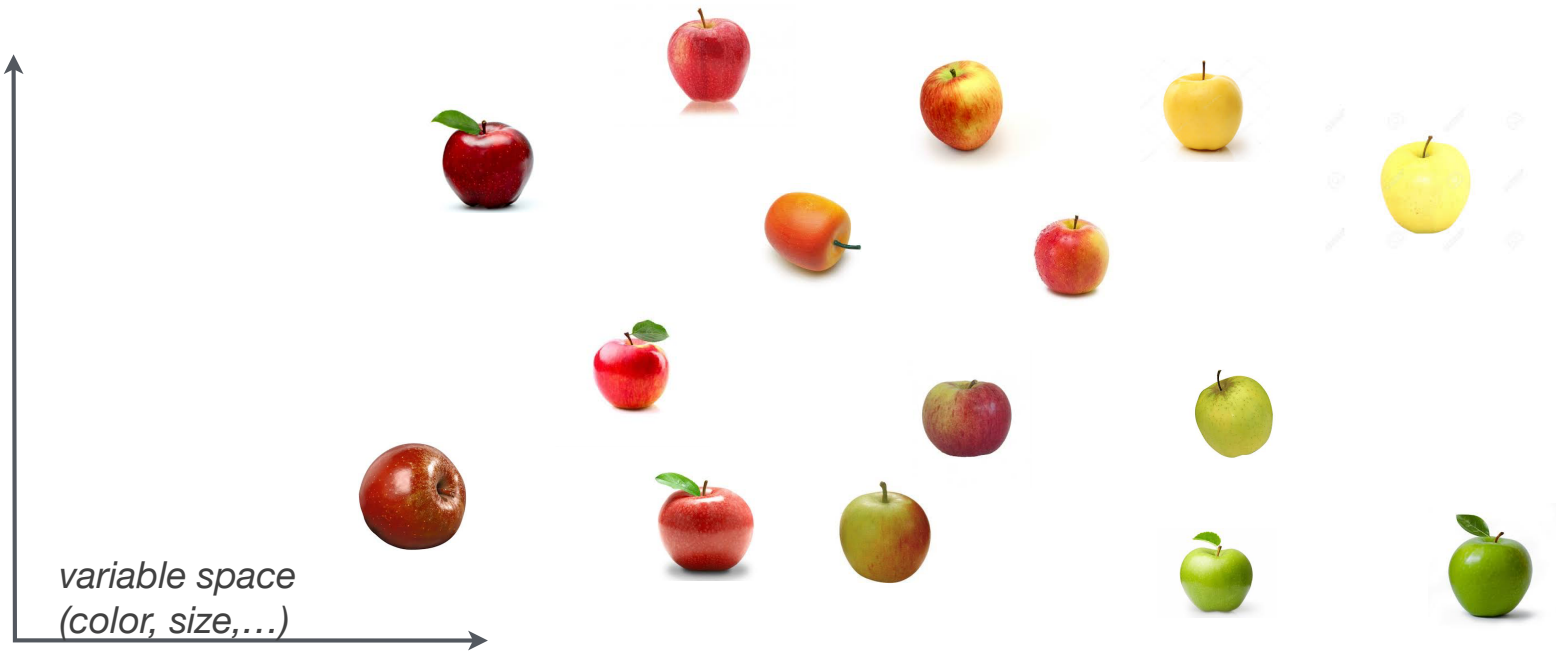


Distance (x,y)	Value
euclidean	2.00
manhattan	3.70
maximum	1.26
correlation distance	0.5(1-0.16) = 0.42

Clustering data

Medizinische Fakultät Heidelberg

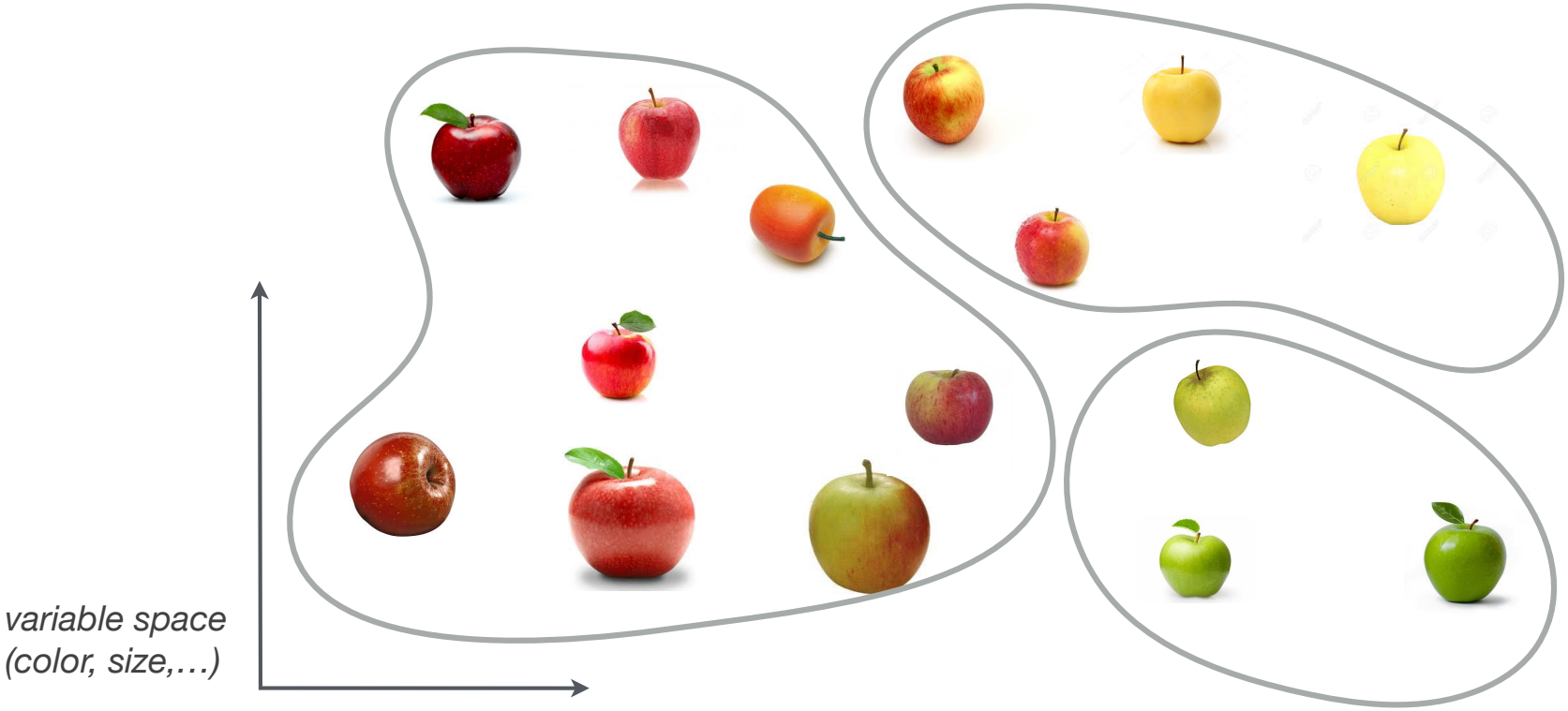
- Idea: group objects (= samples) which are **close to each other** in the variable space
- this implies the definition of a **distance between objects**
- **unsupervised approach:** we do **NOT** know to which group which sample belongs!



Flat clustering

Medizinische Fakultät Heidelberg

- **Partition samples** into k-groups
- Issues
  - how to define a **distance**?
  - optimal **number *k* of groups**?
  - how to deal with **ambiguous** samples?



k-means clustering

Medizinische Fakultät Heidelberg

- **Objects** in a d-dimensional space
  - points in a 2d euclidean space
  - patients in a 20000d space of gene expression, …
- Distance between object is the **euclidean metric**
- Define an **expected number *k* of clusters**
  - try out different numbers k
  - use external knowledge
- See demo here
  http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

k-means

k-means

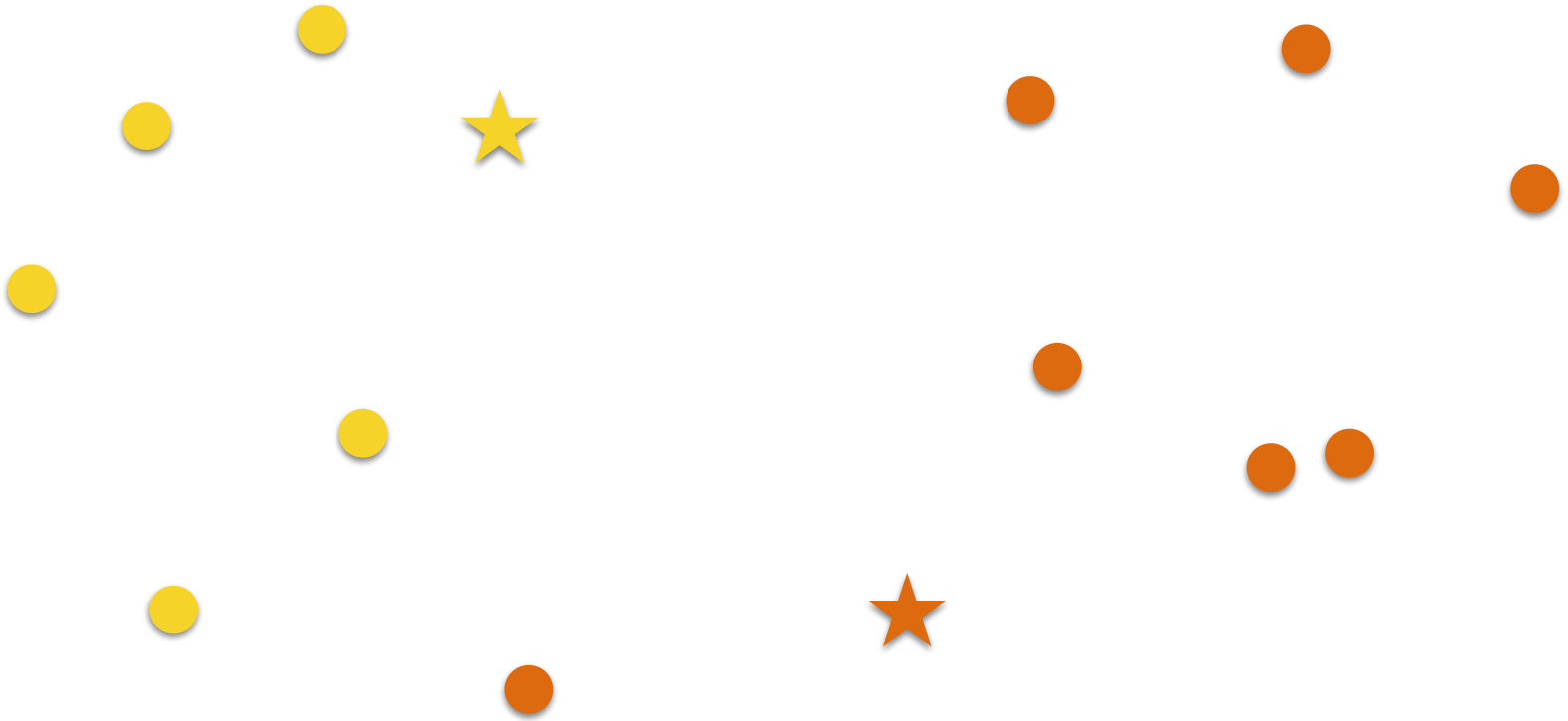
Medizinische Fakultät Heidelberg

- **Step 1:** define k random centers (here: k=2)
- assign each point to the closest one of these centers

k-means

Medizinische Fakultät Heidelberg

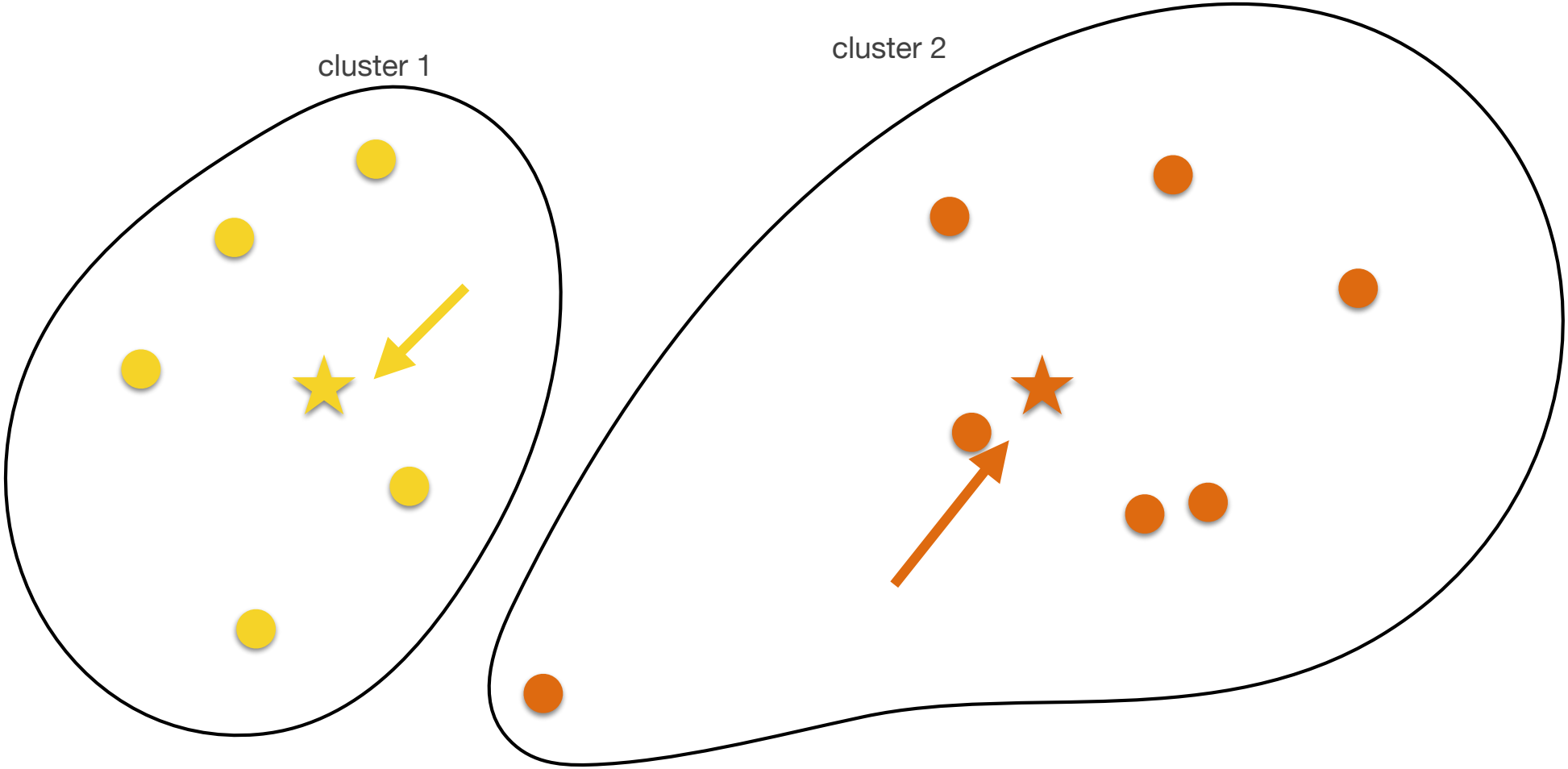
- **Step 1:** define k random centers (here: k=2)
- assign each point to the closest one of these centers



k-means

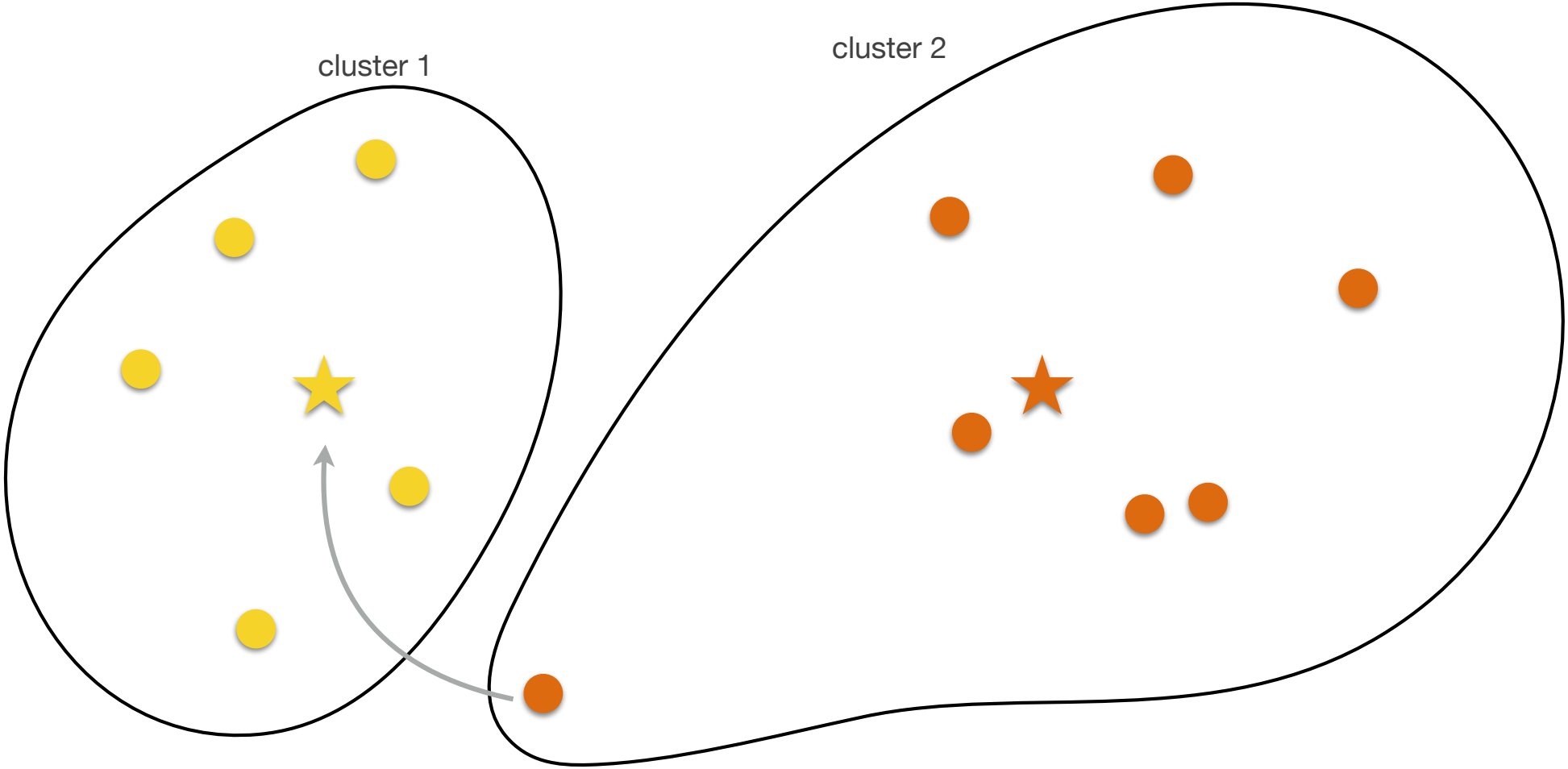
Medizinische Fakultät Heidelberg

- **Step 2:** determine the **center of gravity** of these 2 groups of objects
- assign each point to the closest one of these centers



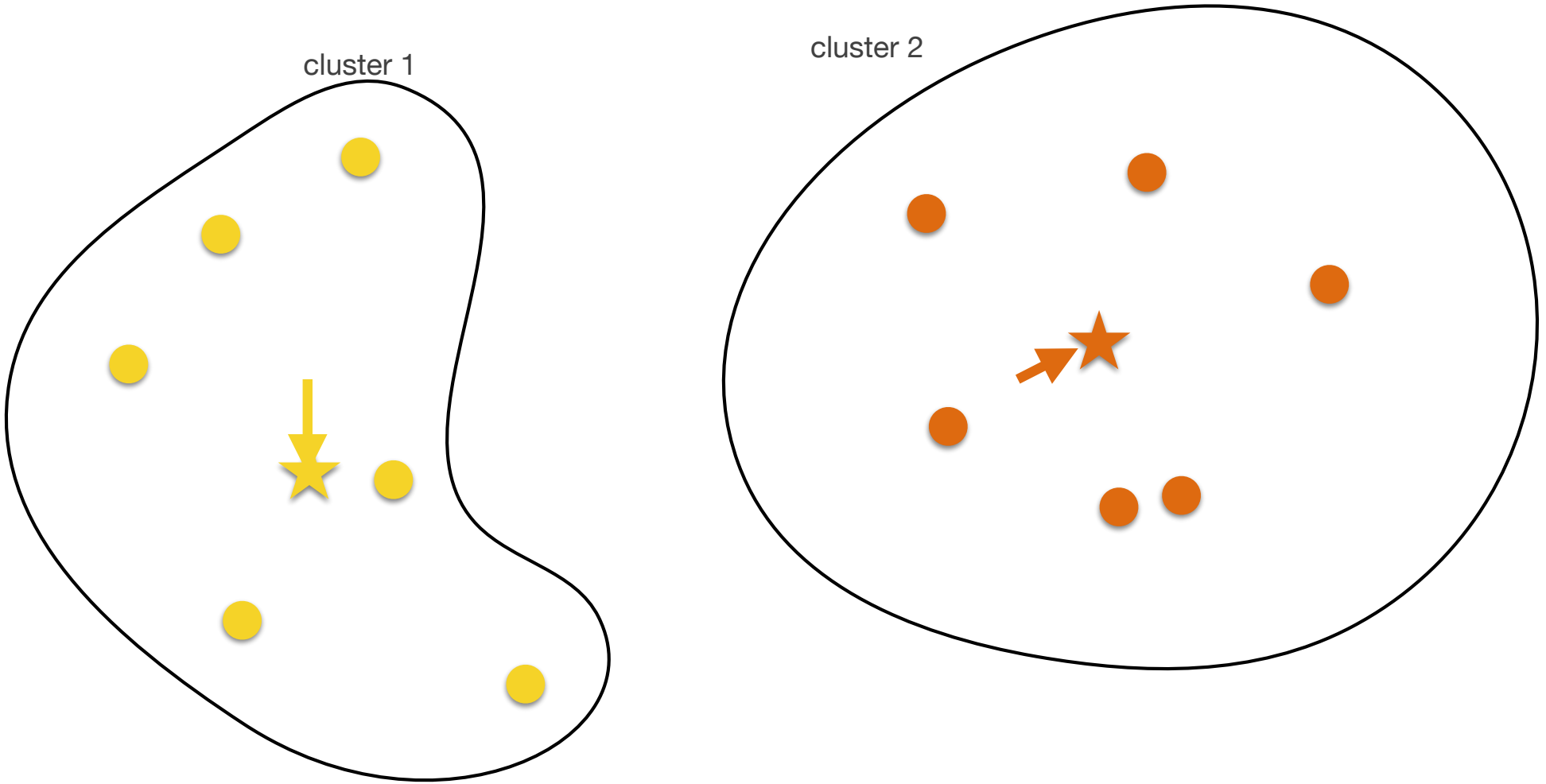
k-means

- **Step 3:** assign again each object to the updated centers
- assign each point to the closest one of these centers



k-means

- **Step 4:** update the centers according to the new assignment, and start over



k-means

Medizinische Fakultät Heidelberg

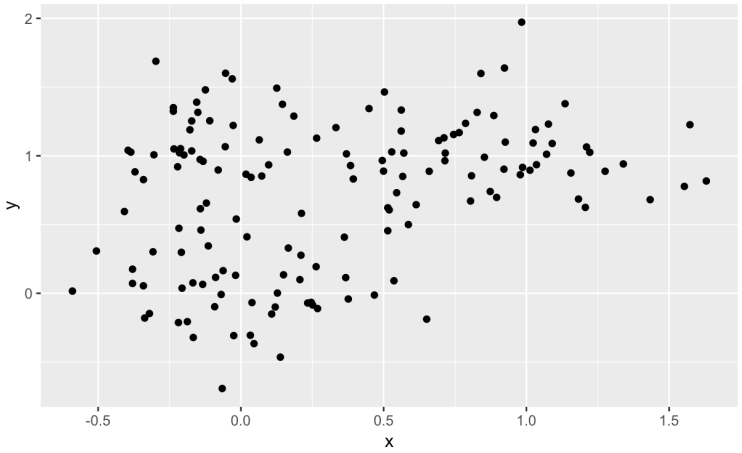
- When to stop:
  - when no object chances its cluster assignment anymore
  - or: when a certain number of max iterations are reached

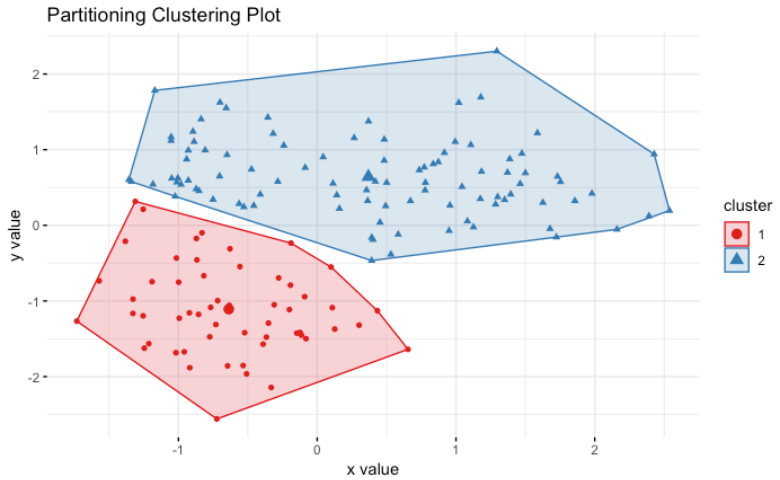
- Process should be repeated over a number of random initial conditions!
- Verify the consistency of the clustering obtained using several random conditions
  - identical = very strong cluster structure
  - fluctuating = no clear cluster structure

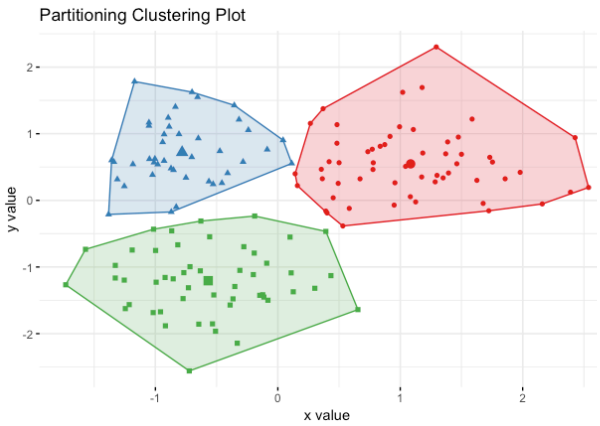
- **Clusters will be identified, even if there is no group structure in the data!**

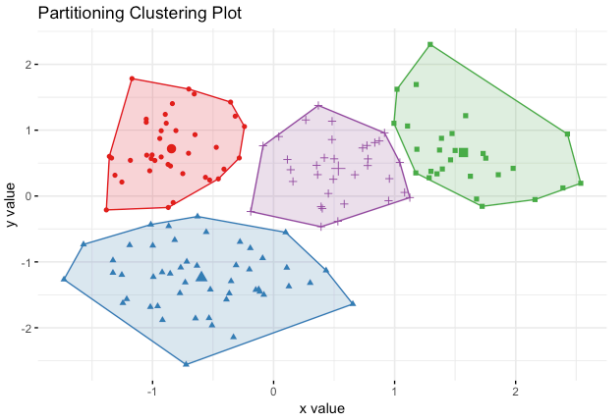
k-means
choosing k

Medizinische Fakultät Heidelberg









Which of these clusterings is better and why?

k-means
choosing k

Medizinische Fakultät Heidelberg

- **Fundamental idea**

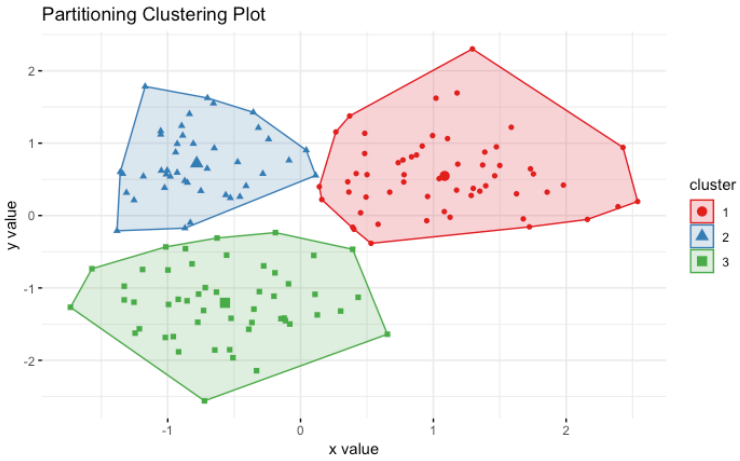
  in a good clustering, the **distance whithin clusters** should be much smaller that the **distance between clusters**

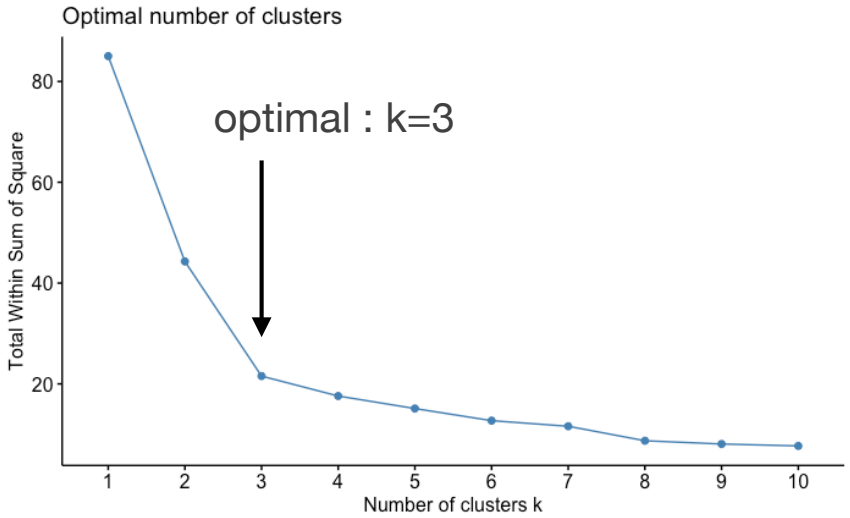
- Several methods implement this principle
  - **elbow** method:
  - **silhouette** method
  - (many other methods exist to define optimal cluster number…)

k-means
elbow method

Medizinische Fakultät Heidelberg

- Sum all **pairwise distances squared** between members of the same cluster
  (within square distance = WSS)
- If each point is its own cluster, then WSS = 0
- WSS decreases with increasing k
- when optimal number of clusters is reached, adding additional clusters does not improve the WSS that much
  → kink (or “elbow”) in the curve





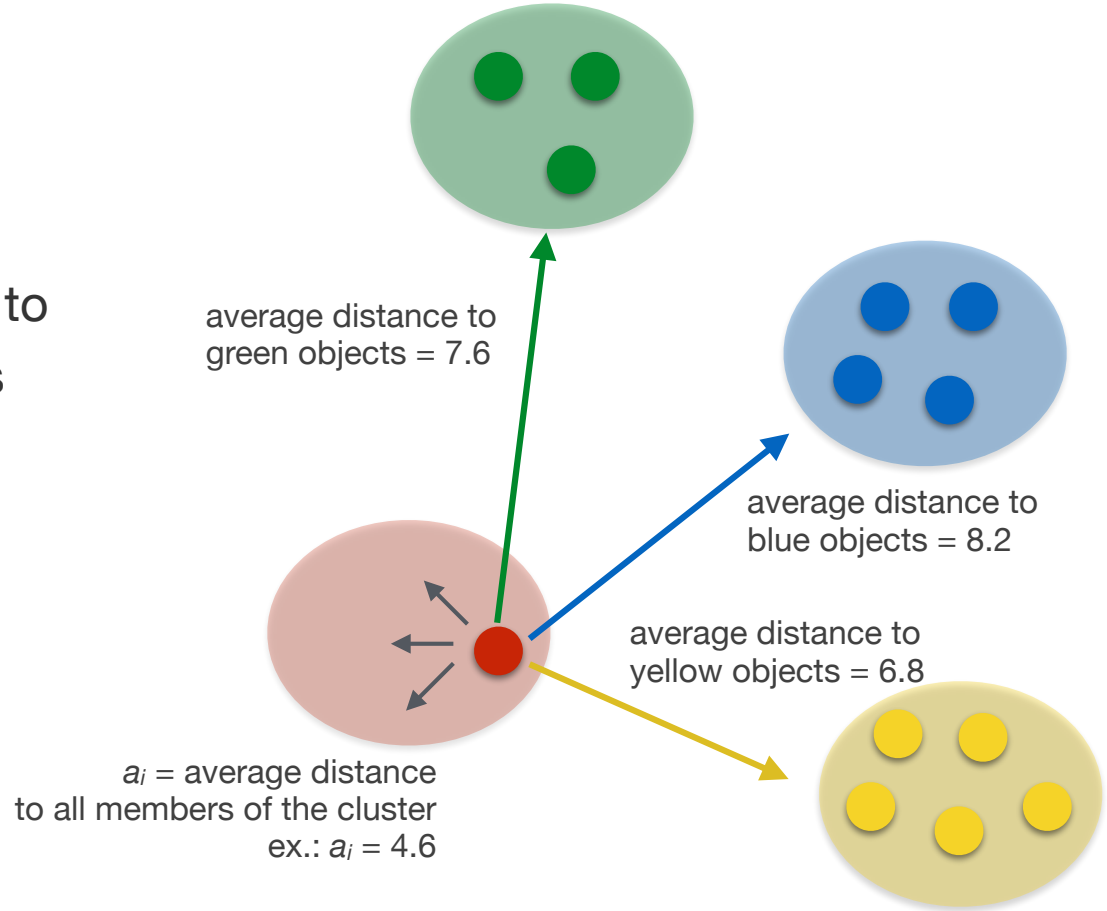
k-means
silhouette method

Medizinische Fakultät Heidelberg

- for each element i, compute
  - mean distance a_i to all members of his cluster
  - smallest average distance b_i to members of all other clusters
- Define silhouette value as

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

- Properties
  - s_i ~ 1 : object very well clustered
  - s_i ~ 0 : ambiguous object
  - s_i <0 : wrongly assigned



b_i = min(7.6, 8.2, 6.8) = 6.8

s_i = (6.8 - 4.6)/6.8 = 0.32

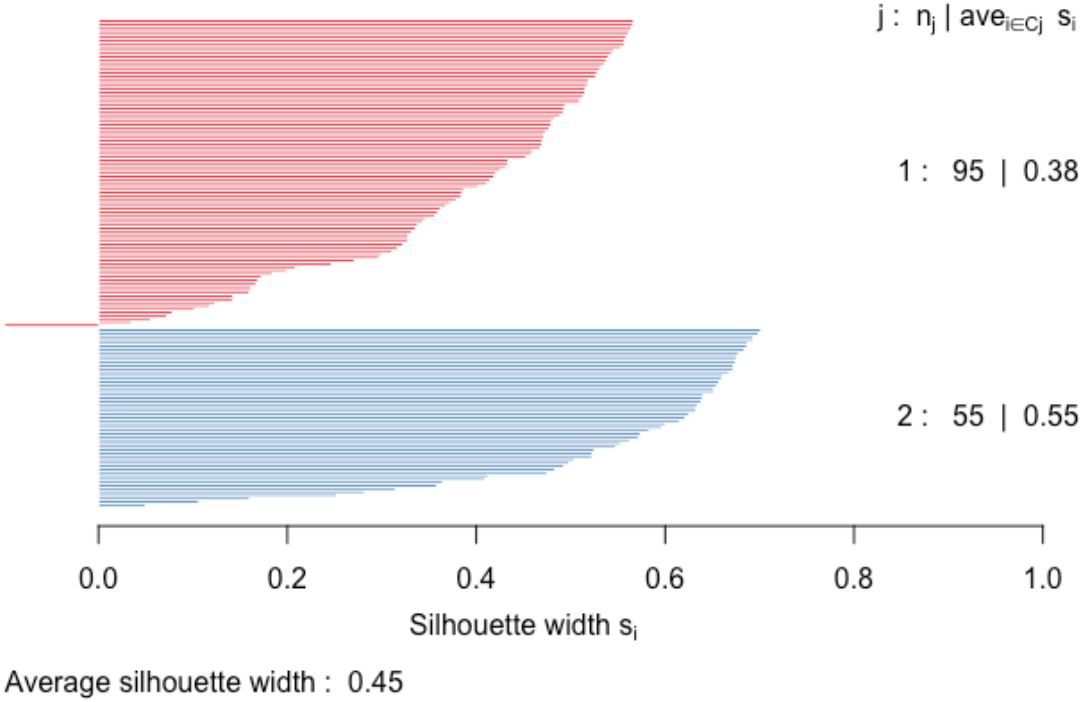
k-means
silhouette method

Medizinische Fakultät Heidelberg

average value for each cluster
(the larger the better)

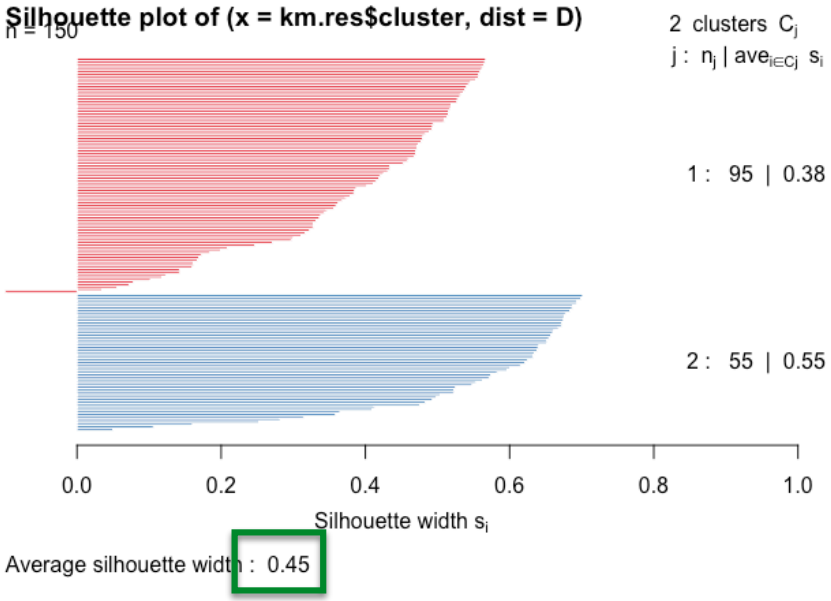
silhouette value for each object

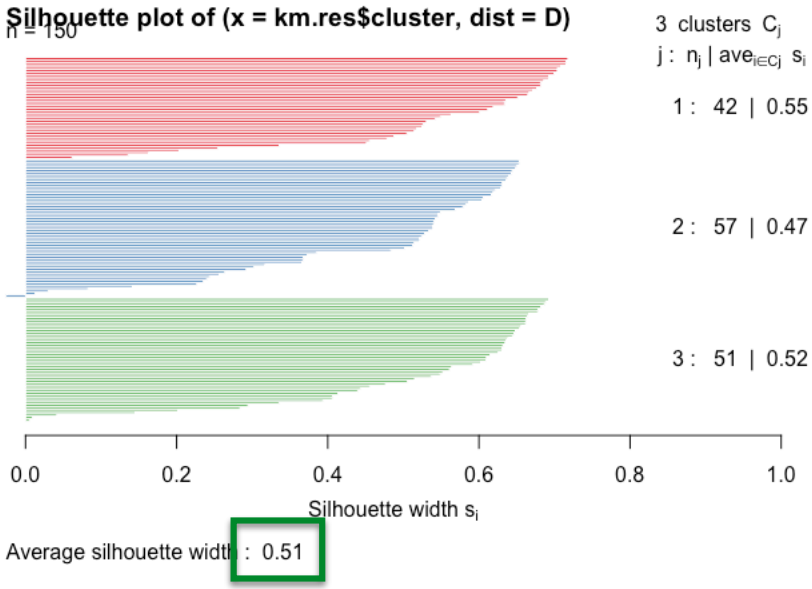
negative values:
object is assigned to the wrong cluster
(closer to some other cluster than
to the members of his cluster…)

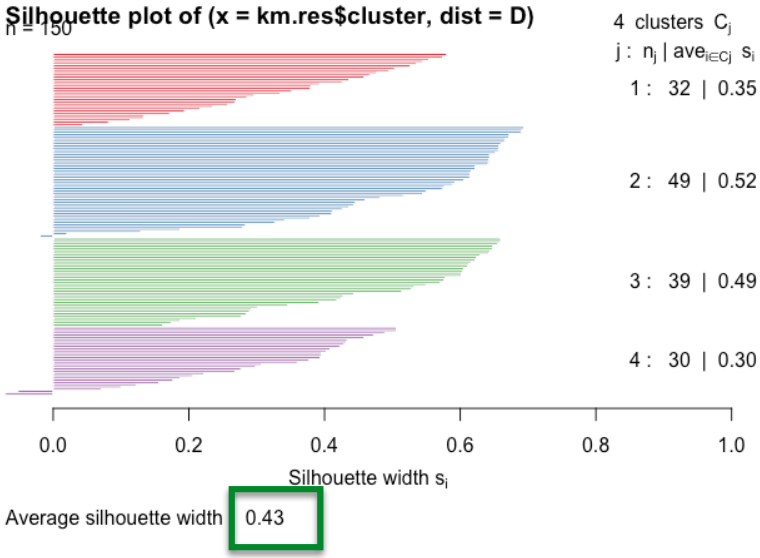


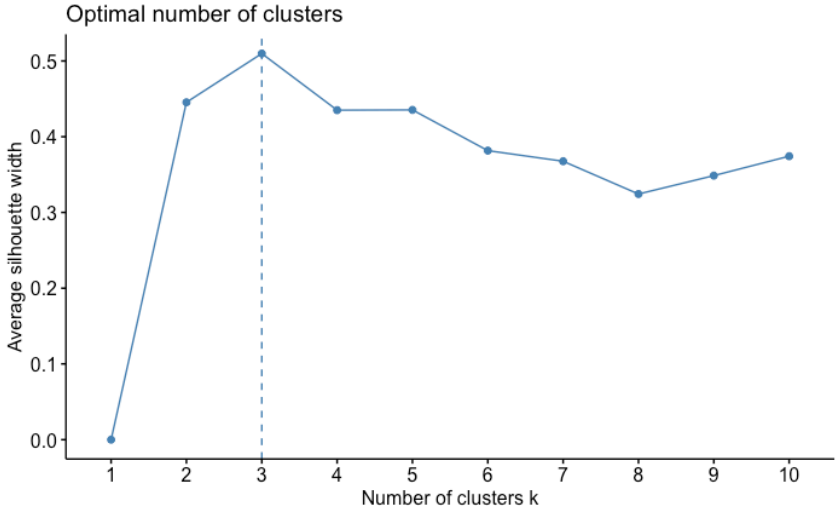
Average silhouette width : 0.45

k-means
silhouette method





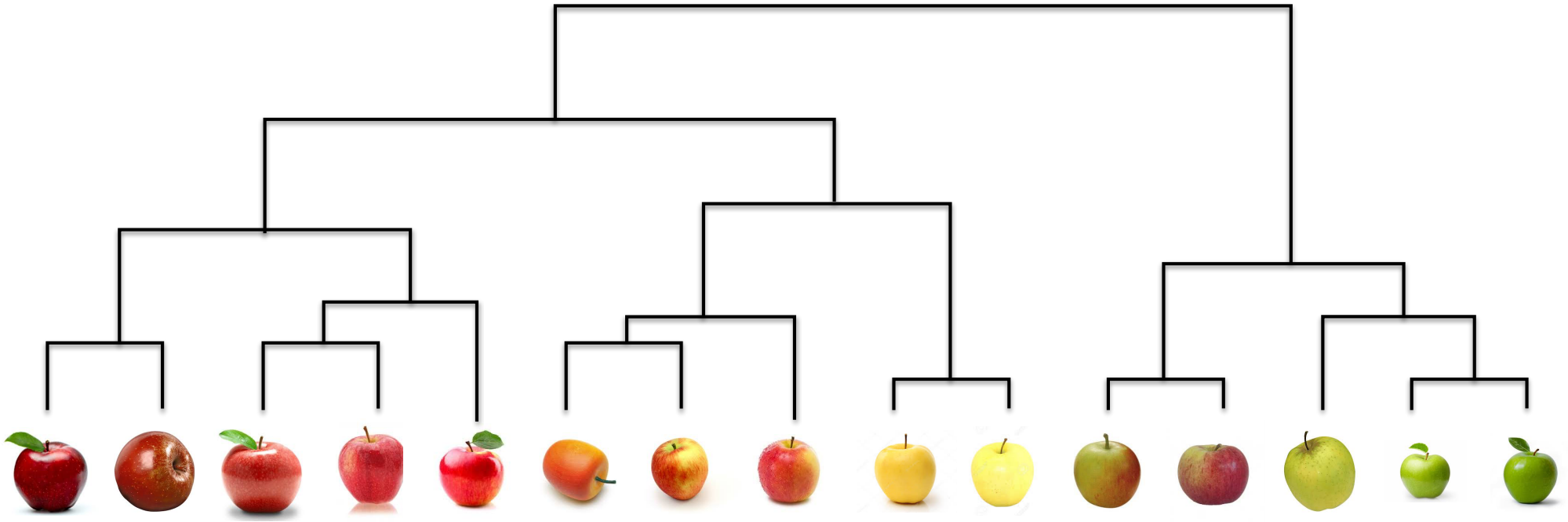




Hierarchical clustering

Medizinische Fakultät Heidelberg

- Bottom-up agglomeration: group iteratively objects from most similars to less similar groups
- tree-like representation of the relationships/distances between objects



Hierarchical clustering

Medizinische Fakultät Heidelberg

- Starting point: **distance / similarity matrix** between all objects
- **Agglomerative** method (“bottom up”)
  - Step 1 : find closest pair (x,y)
  - Step 2 : join objects to form a cluster (x,y)
  - Step 3 : replace x and y in the distance matrix by cluster (x,y) and recompute all distances
  - Step 4 : Repeat from Step1
  - Stop when all objects have been merged!

	a	b	c	d	e
a	0	7.1	2.7	5.1	3.8
b		0	3.2	2	1.1
c			0	4.3	4.1
d				0	2.5
e					0

smallest distance

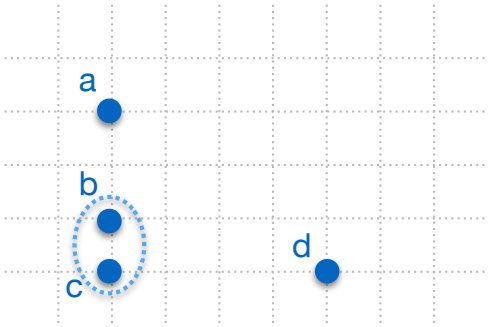
→ **join (b,e)**

	a	c	d	**(b,e)**
a	0	2.7	5.1	?
c		0	4.3	?
d			0	?
(b,e)				0

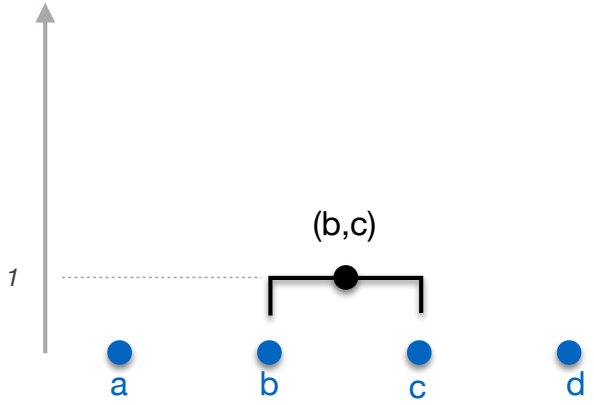
- Distance between single objects: based on distance measure
- Distance between object and cluster?
  → several possible definitions
  (= “linkage methods”)

Hierarchical clustering

(single-linkage)



	a	b	c	d
a	0	2	3	5
b		0	**1**	4.1
c			0	4
d				0



Linkage methods

Medizinische Fakultät Heidelberg

- Several possible definitions of the distance between two clusters of objects (or an object and a cluster)
- **single-linkage**
  d(A,B) = **minimal** distance between all elements of A and B

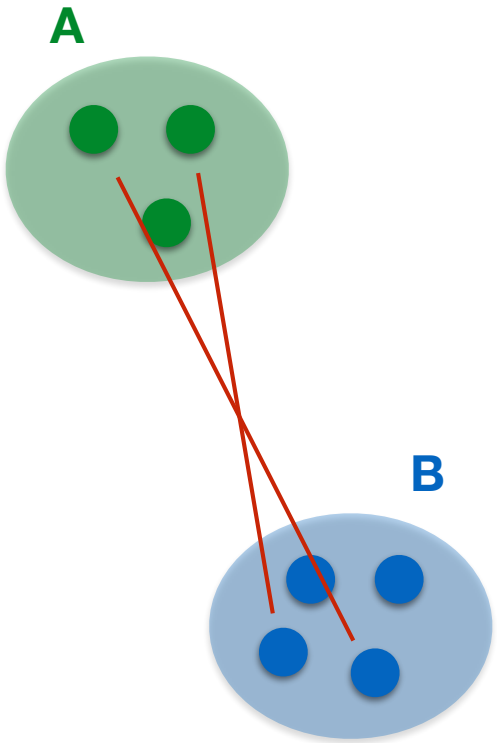
$$d_{\text{single linkage}} = \min_{i,j}(d(a_i, b_j))$$

- **complete-linkage**
  d(A,B) = **maximal** distance between all elements of A and B

$$d_{\text{complete linkage}} = \max_{i,j}(d(a_i, b_j))$$

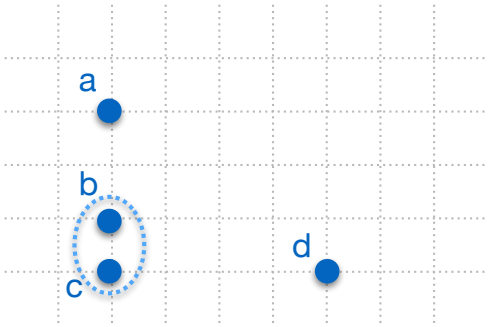
- **average linkage (UPGMA)**
  d(A,B) = **average** of all pairwise distances between elements of A and B

$$d_{\text{average linkage}} = \frac{1}{|A||B|}\sum_i \sum_j d(a_i, b_j)$$

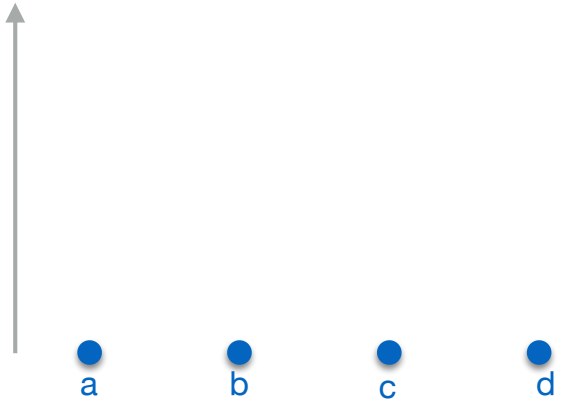


Hierarchical clustering

(single-linkage)

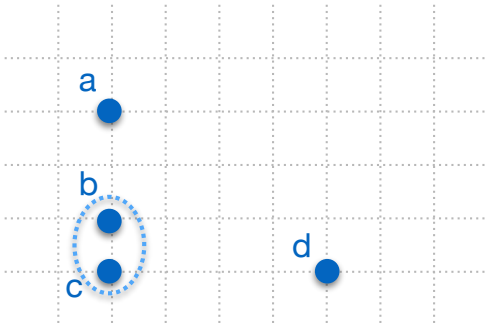


	a	b	c	d
a	0	2	3	5
b		0	**1**	4.1
c			0	4
d				0

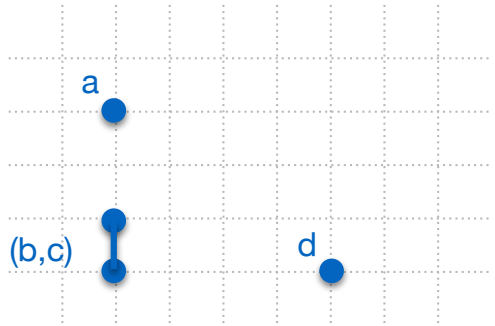


Hierarchical clustering

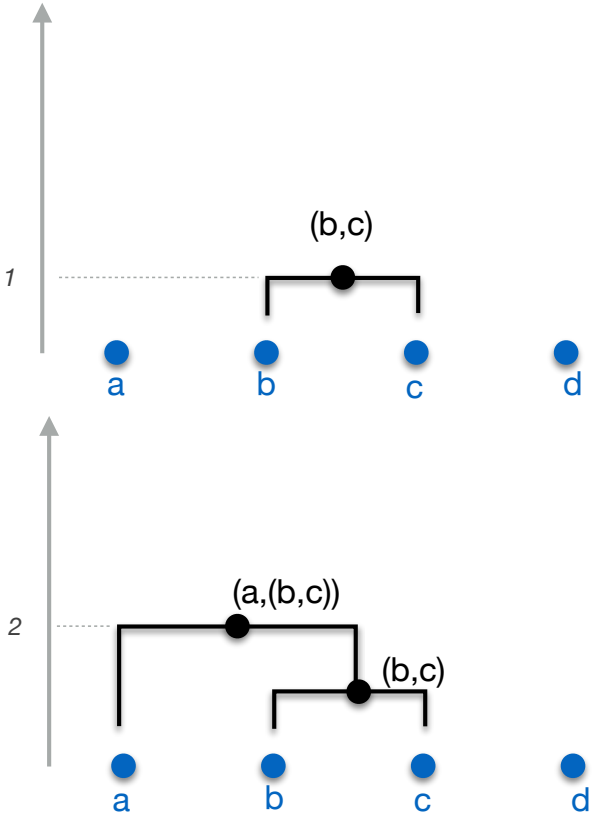
(single-linkage)



	a	b	c	d
a	0	2	3	5
b		0	**1**	4.1
c			0	4
d				0

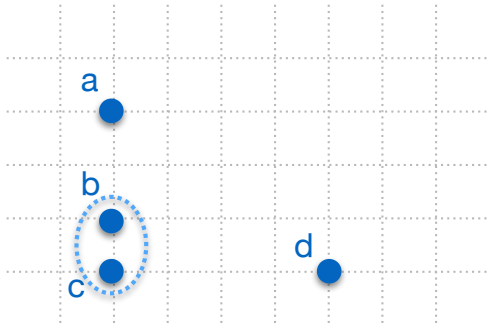


	a	(b,c)	d
a	0	**2**	5
(b,c)		0	**4**
d			0

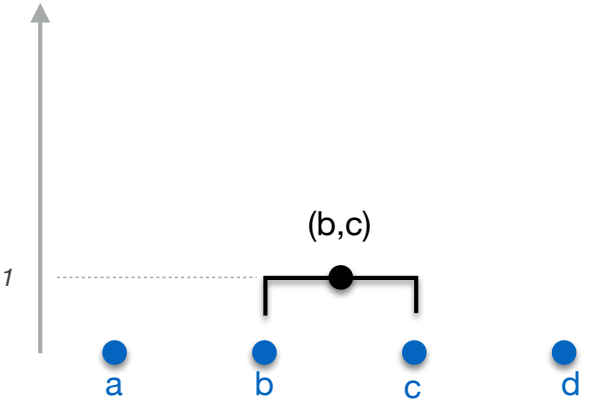


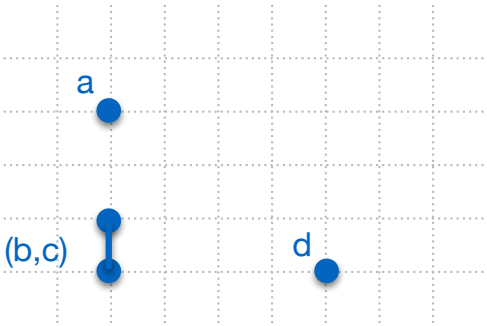
Hierarchical clustering

(single-linkage)

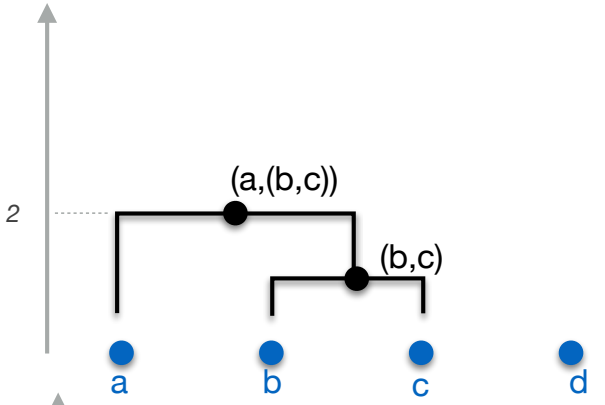


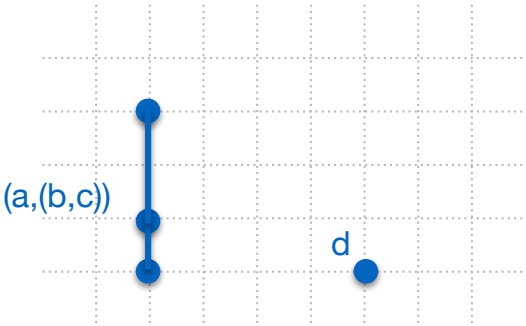
	a	b	c	d
a	0	2	3	5
b		0	**1**	4.1
c			0	4
d				0



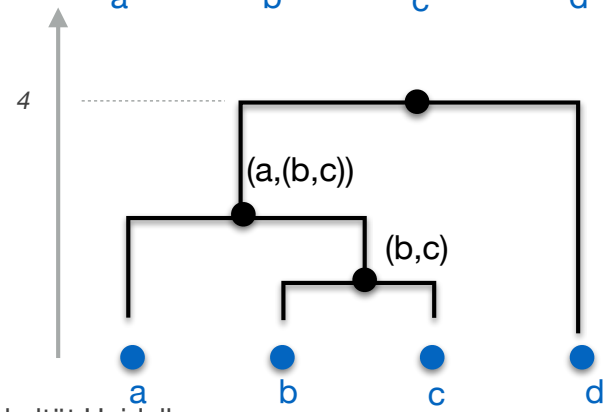


	a	(b,c)	d
a	0	**2**	5
(b,c)		0	**4**
d			0



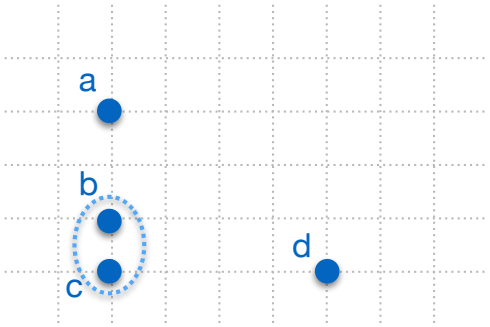


	(a,(b,c))	d
(a,(b,c))	0	**4**
d		0

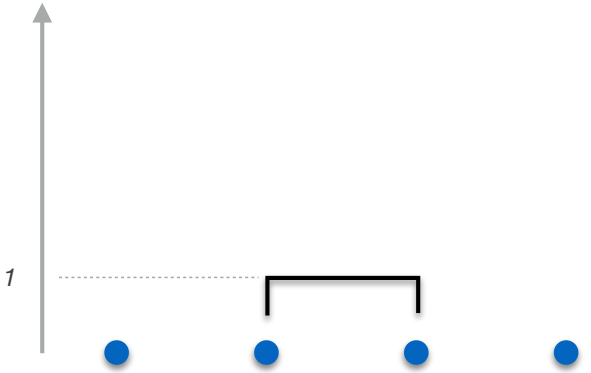


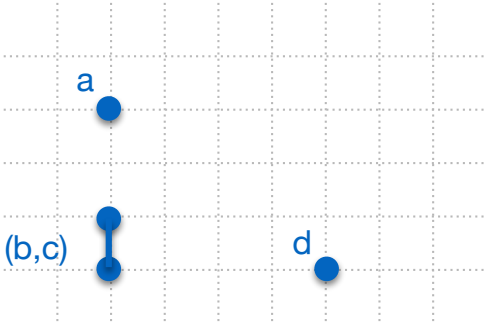
Hierarchical clustering

(complete-linkage)

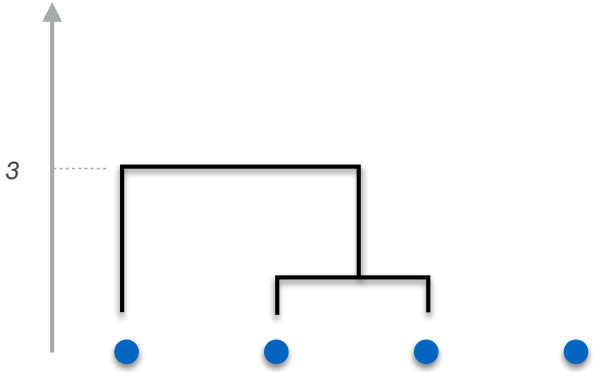


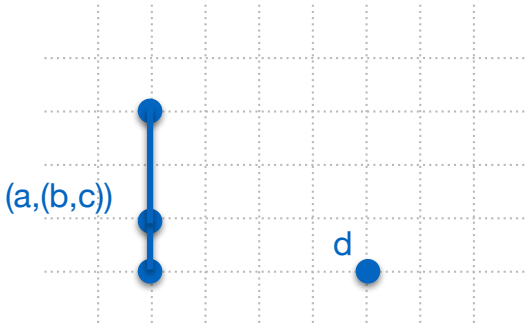
	a	b	c	d
a	0	2	3	5
b		0	**1**	4.1
c			0	4
d				0



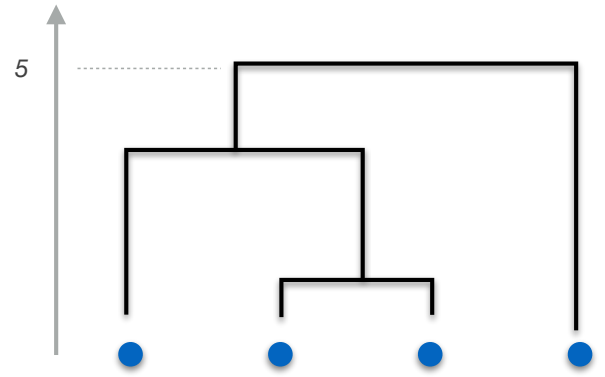


	a	(b,c)	d
a	0	**3**	5
(b,c)		0	**4.1**
d			0



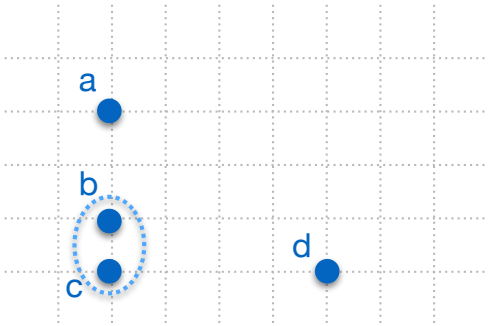


	(a,(b,c))	d
(a,(b,c))	0	**5**
d		0

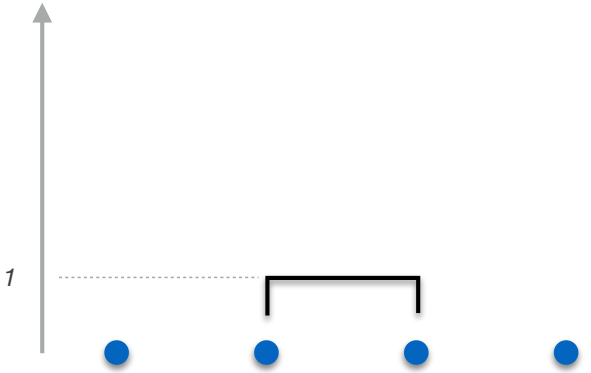


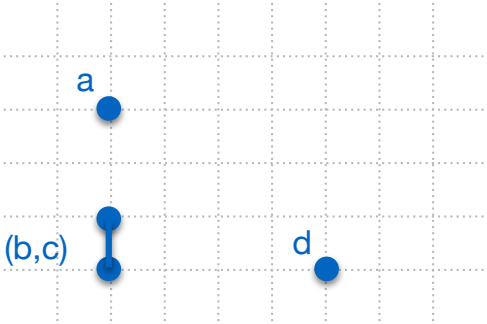
Hierarchical clustering

(average-linkage)

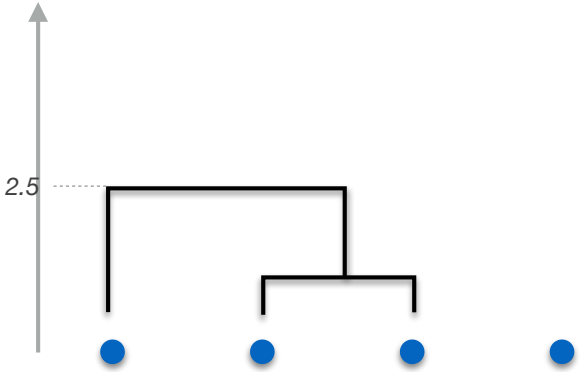


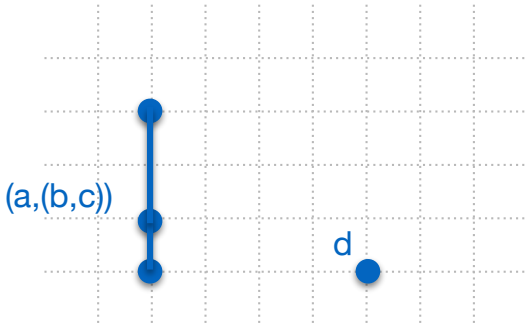
	a	b	c	d
a	0	2	3	5
b		0	**1**	4.1
c			0	4
d				0



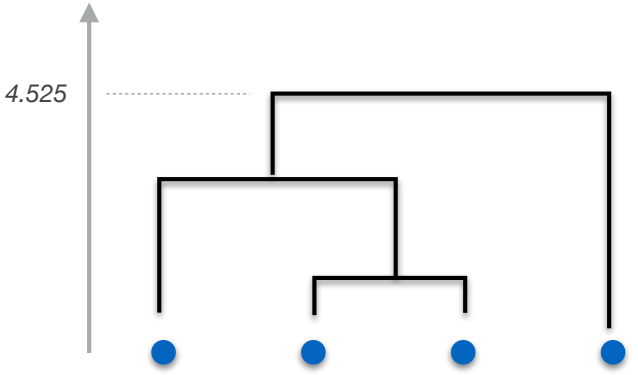


	a	(b,c)	d
a	0	**2.5**	5
(b,c)		0	**4.05**
d			0



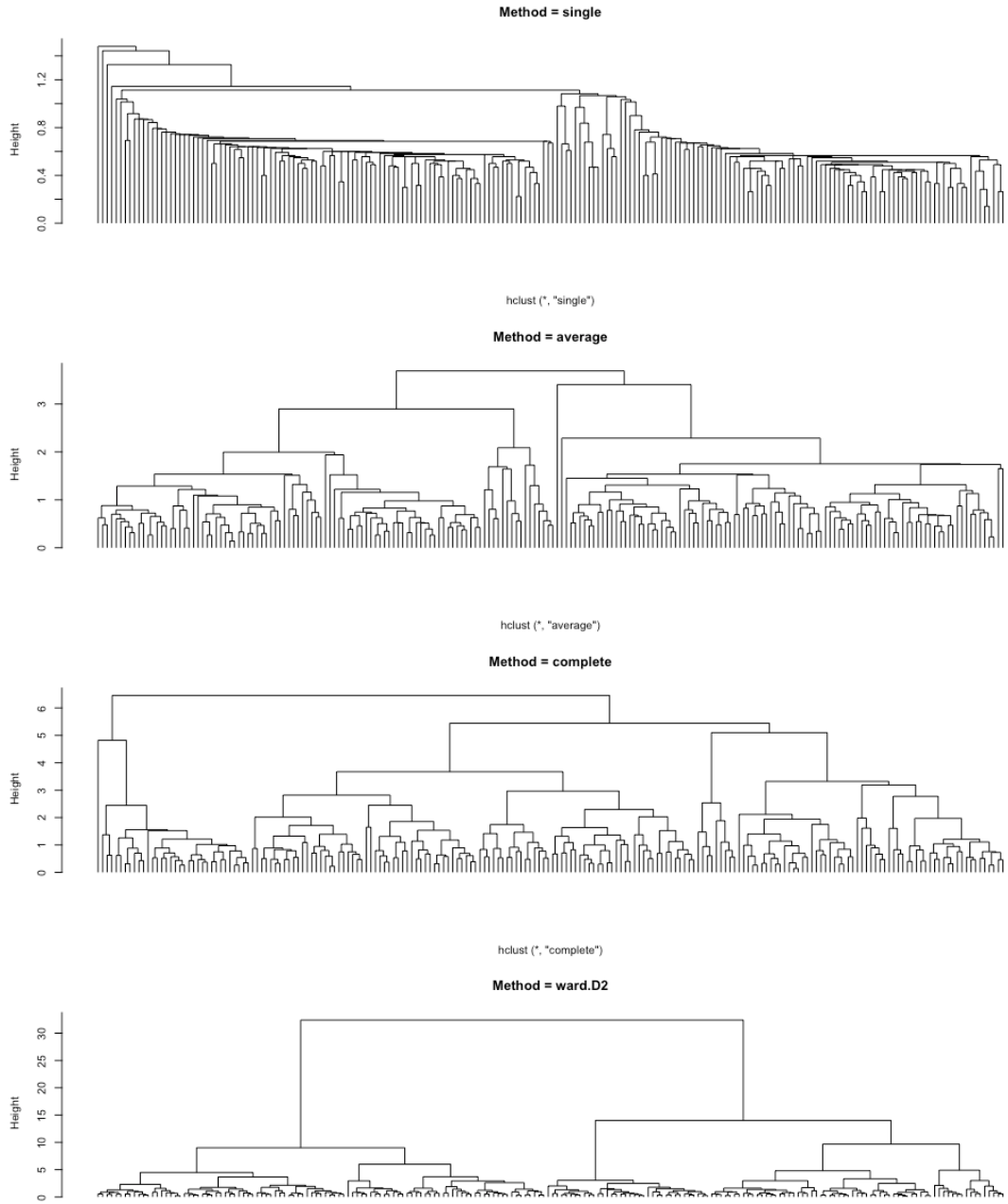


	(a,(b,c))	d
(a,(b,c))	0	**4.525**
d		0

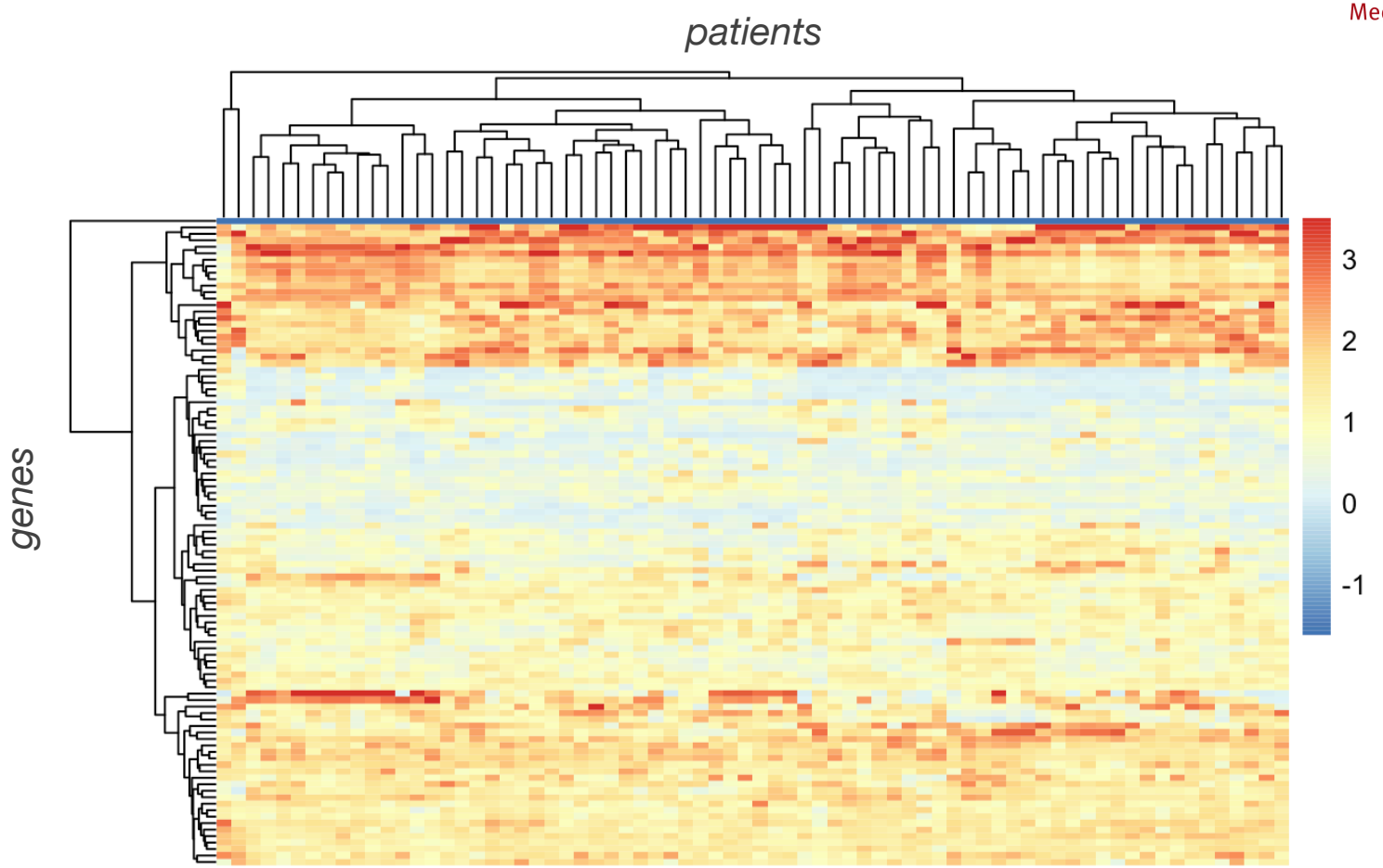


Effect of linkage

Medizinische Fakultät Heidelberg

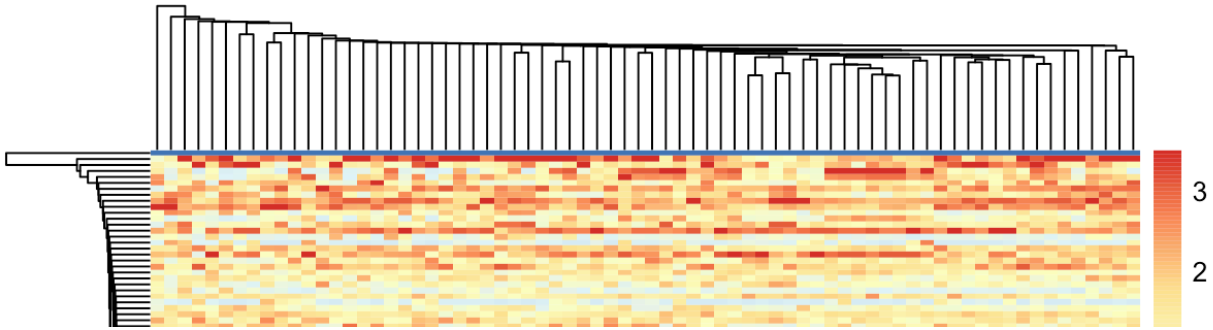


Real world example

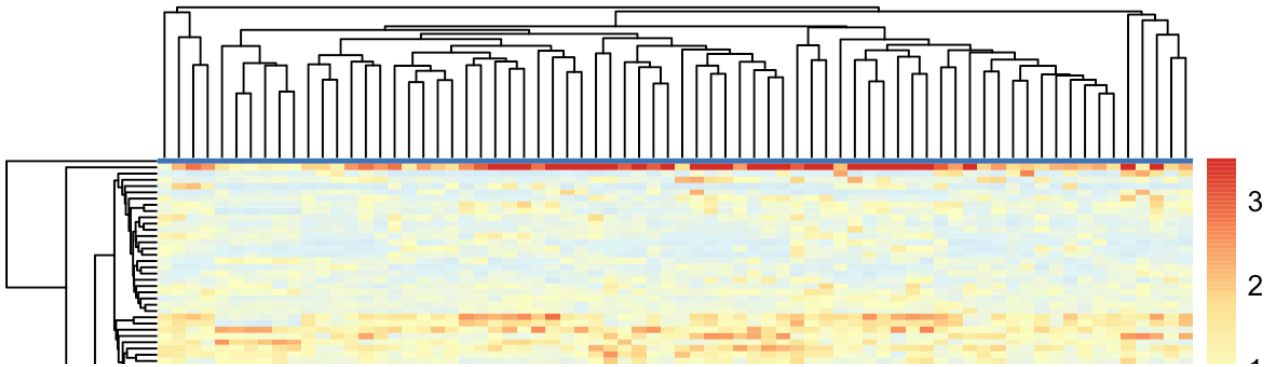


- expression data from 72 AML / ALL patients (Golub et al. 1999)
- clustering of rows and columns: **complete linkage**
- distance between rows / columns: pearson correlation coefficient

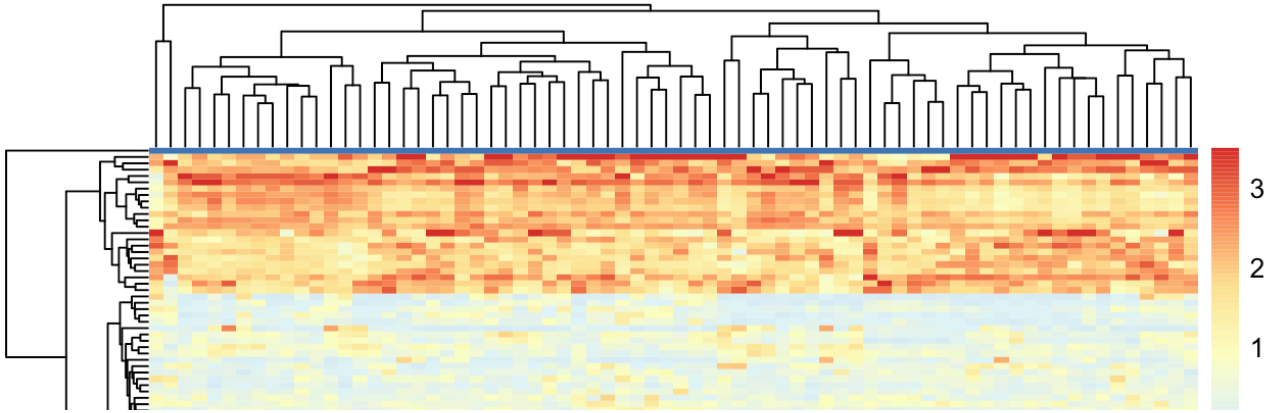
Effect of linkage



single linkage



average linkage

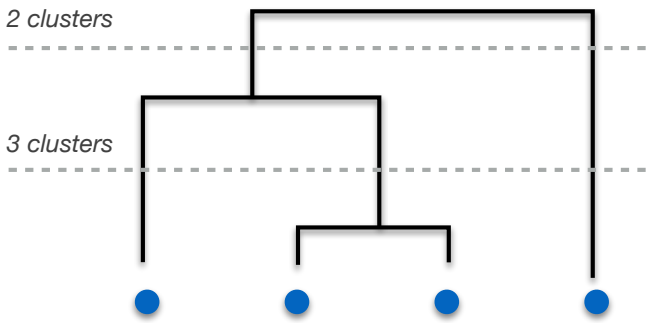


complete linkage

Hierarchical clustering

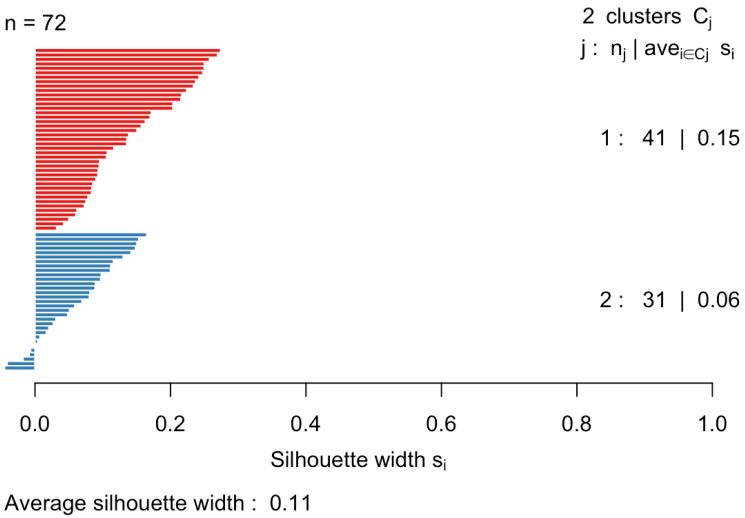
Medizinische Fakultät Heidelberg

- Dendrogram (= rooted tree diagram)
- height of a branch = distance between the clusters it joins
- **Different linkage methods** can lead to **different tree topologies**!
- **Different similarity measures** can lead to **different tree topologies**!
- Tree does not yet define clusters: tree can be cut at different branch height to define clusters
- Where to cut? How many clusters?
  → see criteria discussed previously! (elbowplot, silhouette,…)

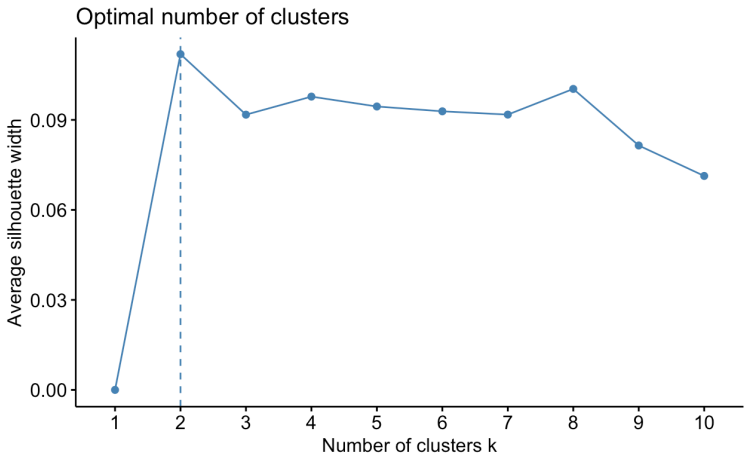


Real world example
Clustering patients

optimal number of cluster: 2
(however, very low silhouette value…)



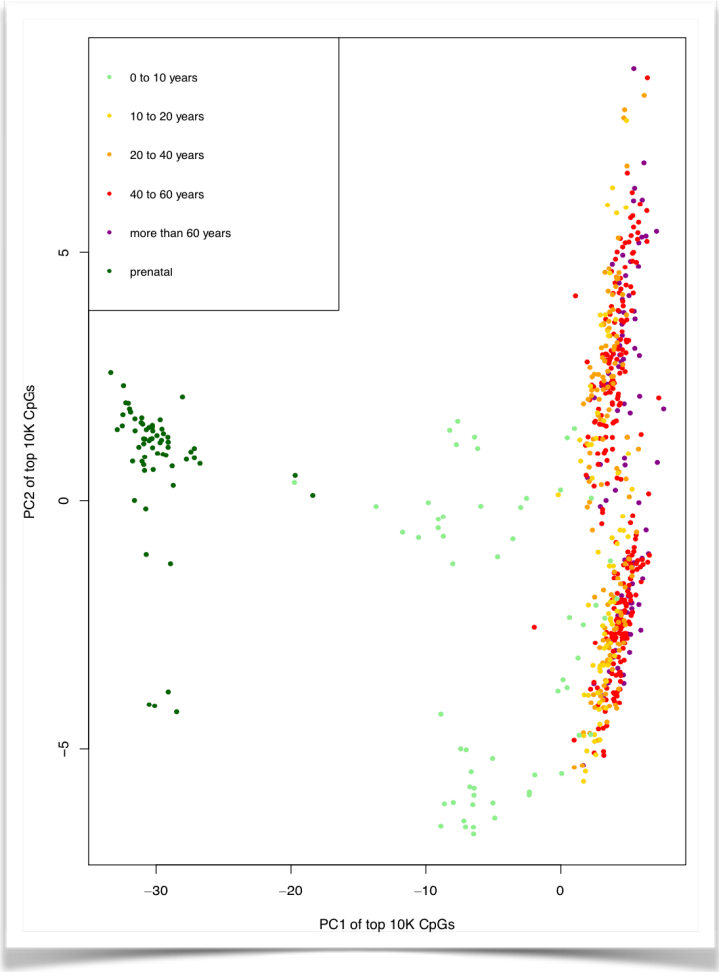
Average silhouette width : 0.11



Reducing dimensions

Medizinische Fakultät Heidelberg

- Dataset have a very high dimensionality (e.g. number of genes)
- Need to reduce this large number of dimensions to a smaller number of relevant variables
- Relevant variables = variables which carry most of the information (or variance) of a dataset
- Goal:
  - identify **directions** in the data corresponding to **biological effects**



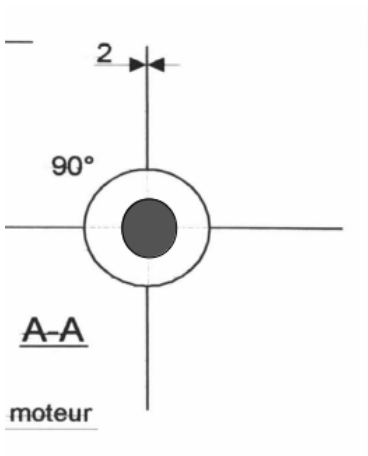
Example of DNA methylation of blood samples in patient cohort (Jana Dalhoff)
data matrix : 400.000 CpG positions / 250 patients

3d → 2d

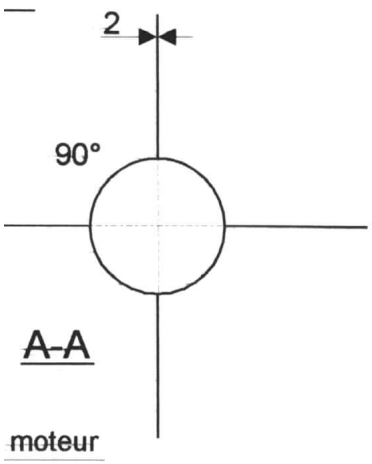
Medizinische Fakultät Heidelberg

- How to draw (= 2d) a rocket (= 3d) ?

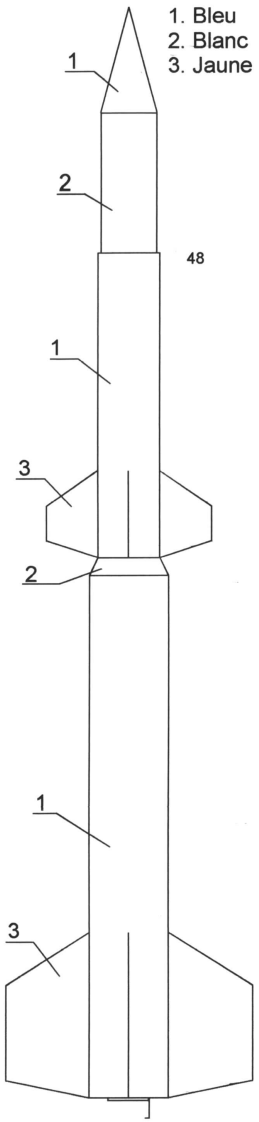
from above



from below



from the side

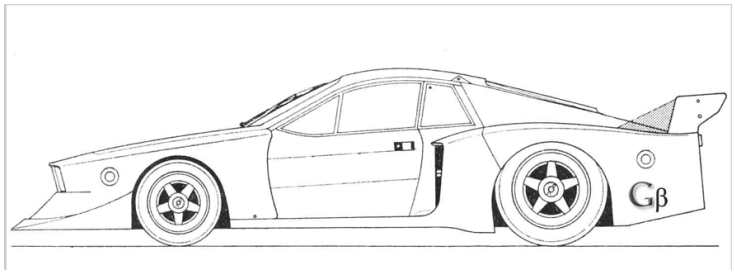


3d → 2d

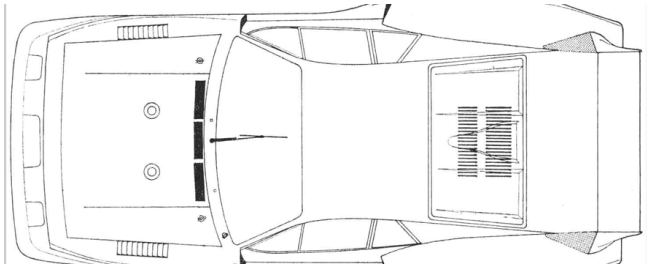
Medizinische Fakultät Heidelberg

- How to draw (= 2d) a car (= 3d) ?

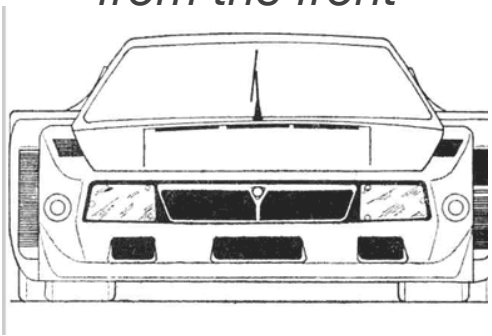
from the side



from above

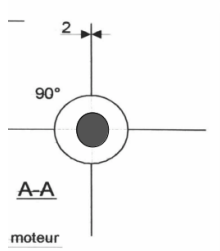


from the front

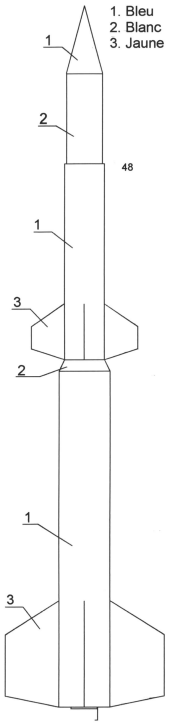


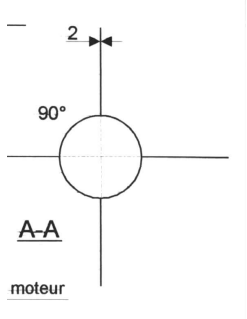
How to choose?

10% information



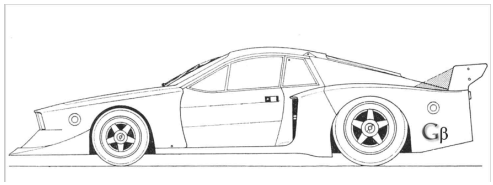
80% information



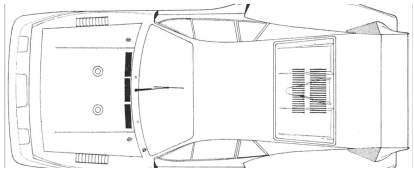


10% information

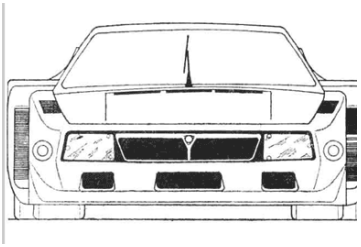
45% information



30% information



25% information

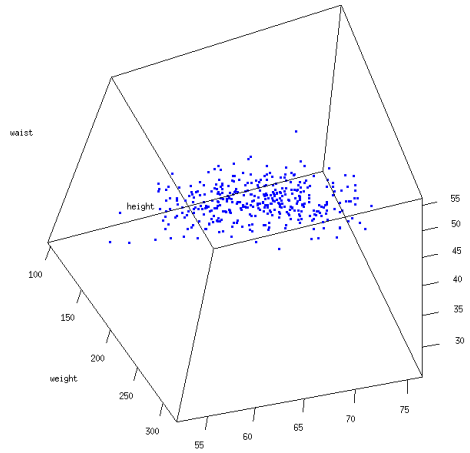


This is principal component analysis!

Principal Component Analysis

Medizinische Fakultät Heidelberg

- Idea of principle component analysis: **reduce a multi-dimensional dataset to a lower number of informative dimensions** (“Principal Components”)
- Example diabetes dataset with chol / age / height / weight / waist
- each patient is a dot in a 5d space
  → how to choose new coordinates?



id	chol	age	height	weight	waist
1000	203	46	62	121	29
1001	165	29	64	218	46
1002	228	58	61	256	49
1003	78	67	67	119	33
1005	249	64	68	183	44
1008	248	34	71	190	36
1011	195	30	69	191	46
1015	227	37	59	170	34
1016	177	45	69	166	34
1022	263	55	63	202	45
1024	242	60	65	156	39
1029	215	38	58	195	42
1030	238	27	60	170	35
1031	183	40	59	165	37
1035	191	36	69	183	36
1036	213	33	65	157	37
1037	255	50	65	183	37

k=17 patients, n=5 variables

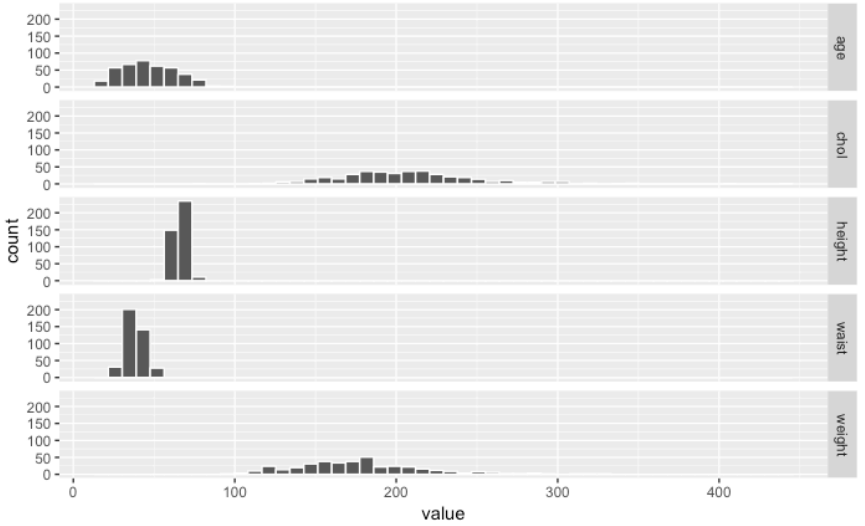
Preparing the dataset

Medizinische Fakultät Heidelberg

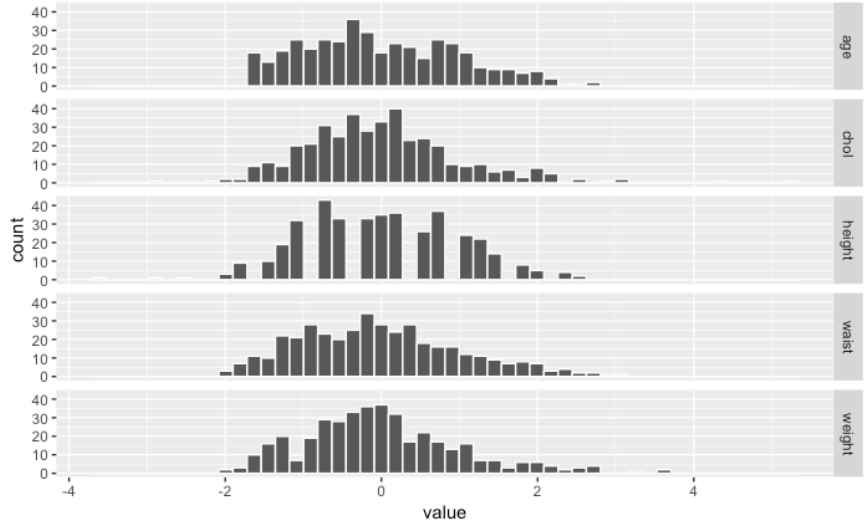
- Each variable has a totally different **range of variation**
- **Solution:** make variables dimensionless using a z-transformation
  → centering and scaling

$$Z = \frac{x - \bar{x}}{\sigma_x}$$

Raw data

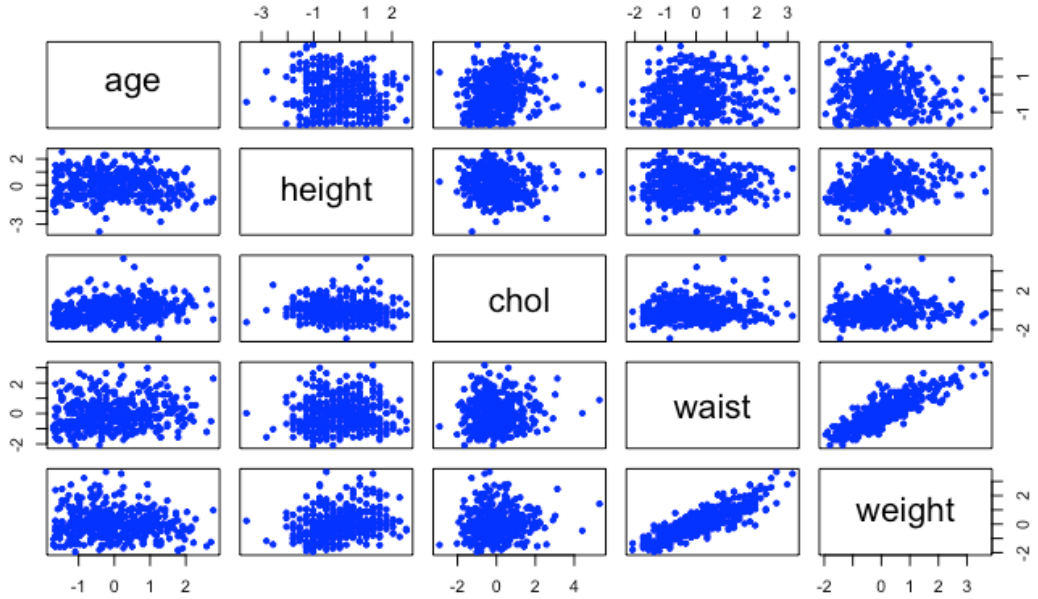


z-transformed data



Correlation structure

Medizinische Fakultät Heidelberg



	age	height	chol	waist	weight
age	1.000	-0.095	0.240	0.153	-0.063
height	-0.095	1.000	-0.059	0.057	0.253
chol	0.240	-0.059	1.000	0.112	0.059
waist	0.153	0.057	0.112	1.000	**0.850**
weight	-0.063	0.253	0.059	**0.850**	1.000

- Some variables are highly correlated : knowing one gives a lot of information about the second
- Some are unrelated : no information gain from one over the other

Correlation structure

Medizinische Fakultät Heidelberg

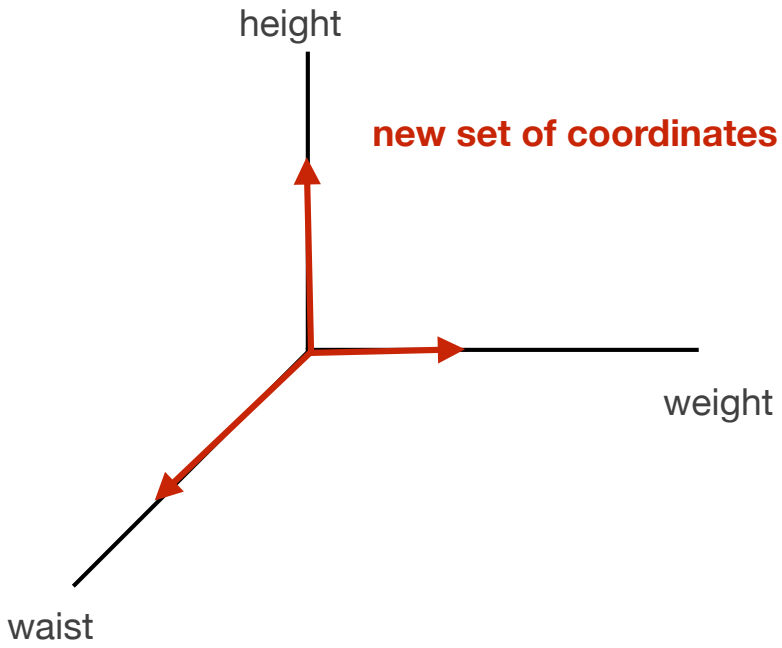
- if two variables are **strongly correlated**, they are partly redundant: knowing the variation of one, you have information about how the second variables changes
  → having 2 variables does not add much information w.r.t. a single variable

- if two variables have **little correlation**, each variable carries information not contained in the other
  → we need to keep these 2 variables to have information about the dataset

The more diagonal a correlation matrix is, the more information is revealed by the variables

	age	height	chol	waist	weight
age	1.000	-0.095	0.240	0.153	-0.063
height	-0.095	1.000	-0.059	0.057	0.253
chol	0.240	-0.059	1.000	0.112	0.059
waist	0.153	0.057	0.112	1.000	**0.850**
weight	-0.063	0.253	0.059	**0.850**	1.000

Defining new variables



- We want to express the dataset in a new set of coordinates
- for each of these coordinate systems, the correlation matrix will change
- **Goal**: find the rotation that makes the correlation matrix diagonal

Hello, matrix diagonalization !

Matrix diagonalization

Medizinische Fakultät Heidelberg

- For some square matrices X, we can find vectors v such that

$$Xv = \lambda v$$

eigenvectors

eigenvalues

- Property: **symmetric matrices** (such as the correlation matrix) are **always diagonalizable**

  → we can find n eigenvalues and n eigenvectors

Matrix diagonalization

Medizinische Fakultät Heidelberg

1. Consider the correlation matrix *A*

	age	height	chol	waist	weight
age	1.00000000	-0.09479919	0.23990232	0.15255761	-0.06269027
height	-0.09479919	1.00000000	-0.05853973	0.05661532	0.25298143
chol	0.23990232	-0.05853973	1.00000000	0.11245805	0.05932074
waist	0.15255761	0.05661532	0.11245805	1.00000000	0.84955930
weight	-0.06269027	0.25298143	0.05932074	0.84955930	1.00000000

2. Determine its *n* eigenvalues and *n* eigenvectors and build the n x n matrix *V* from all the n eigenvectors

```
$values
[1] 1.9201374 1.3081302 0.9011191 0.7635241 0.1070892

$vectors
           [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -0.0782536  0.66340112 -0.1957637  0.70124582 -0.15396822
[2,] -0.2139712 -0.41884235 -0.8557896  0.16410053  0.13957979
[3,] -0.1338086  0.59835882 -0.3893703 -0.68726253  0.01108988
[4,] -0.6768556  0.08069999  0.2542954  0.06898591  0.68258976
[5,] -0.6870622 -0.14115337  0.1141285 -0.06508820 -0.70054229
```

3. Compute $S = V' \cdot A \cdot V$

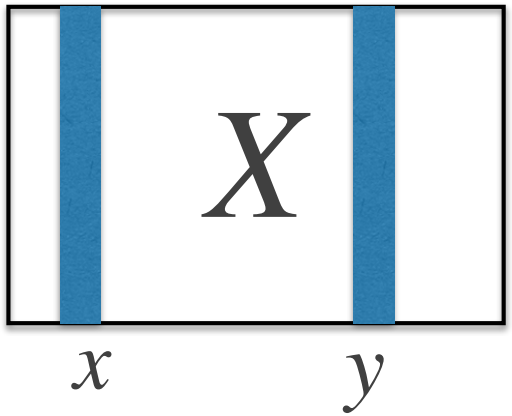
	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.92	0.000	0.000	0.000	0.000
[2,]	0.00	1.308	0.000	0.000	0.000
[3,]	0.00	0.000	0.901	0.000	0.000
[4,]	0.00	0.000	0.000	0.764	0.000
[5,]	0.00	0.000	0.000	0.000	0.107

New coordinate system

Medizinische Fakultät Heidelberg

- Correlation / covariance

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}X_c' \cdot X_c$$



$$cor(x,y) = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} = \frac{1}{N}X_{cs}' \cdot X_{cs}$$

Z-transformation

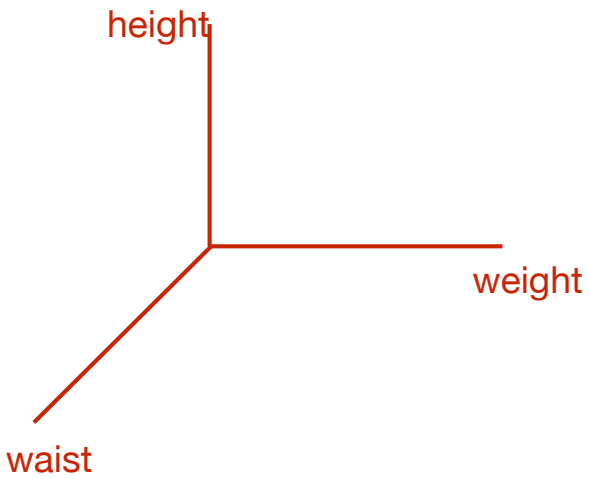
New coordinate system

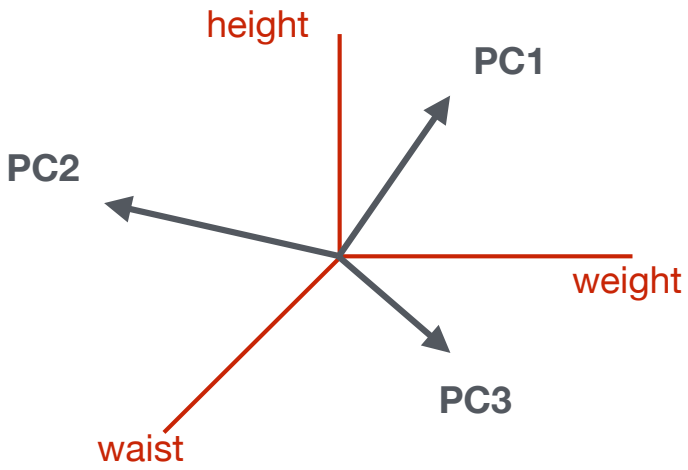
Medizinische Fakultät Heidelberg

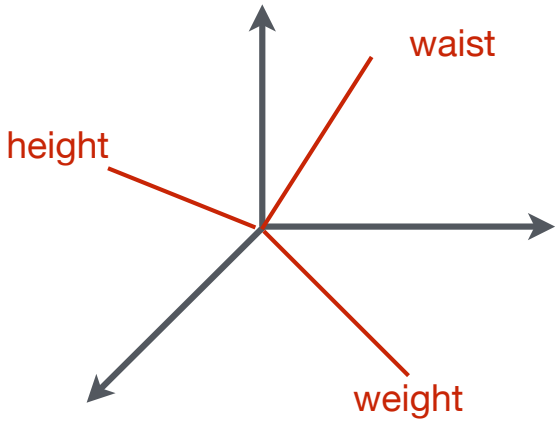
$$S = V' \cdot A \cdot V = V' \cdot \left(\frac{1}{N} X'_{cs} \cdot X_{cs} \right) \cdot V$$

$$S = \frac{1}{N} (X_{cs} \cdot V)' \cdot (X_{cs} \cdot V)$$

- *V* is the **rotation matrix** transforming the initial variables into new variables called principal components



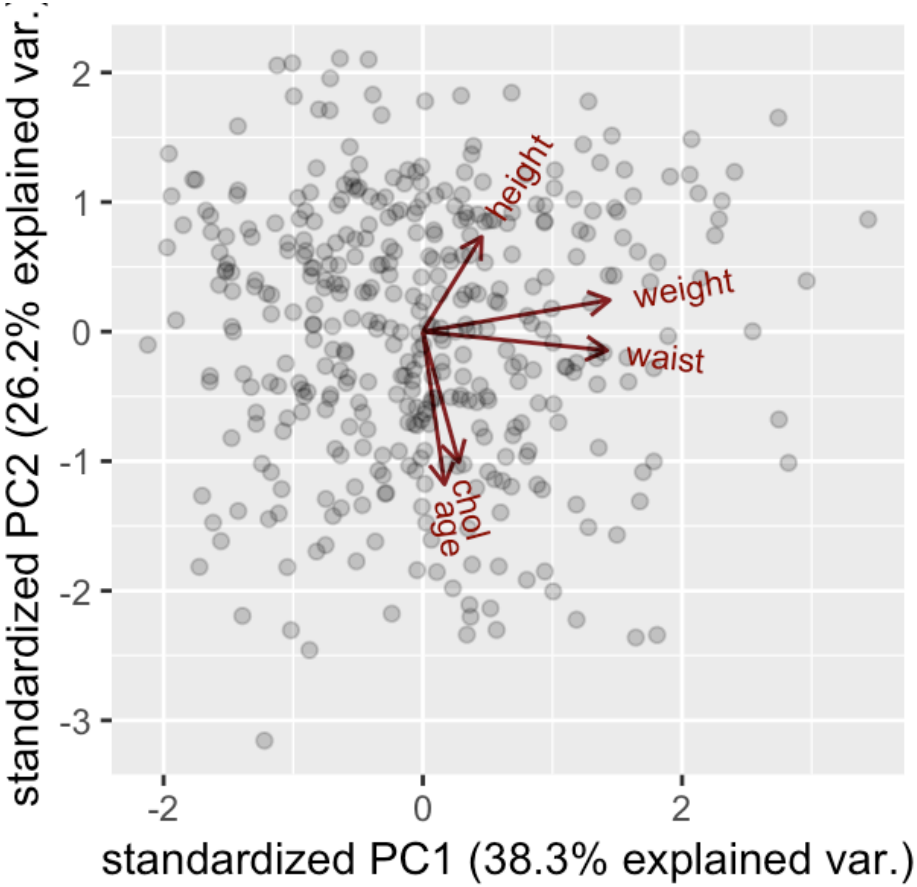




PCA biplot

Medizinische Fakultät Heidelberg

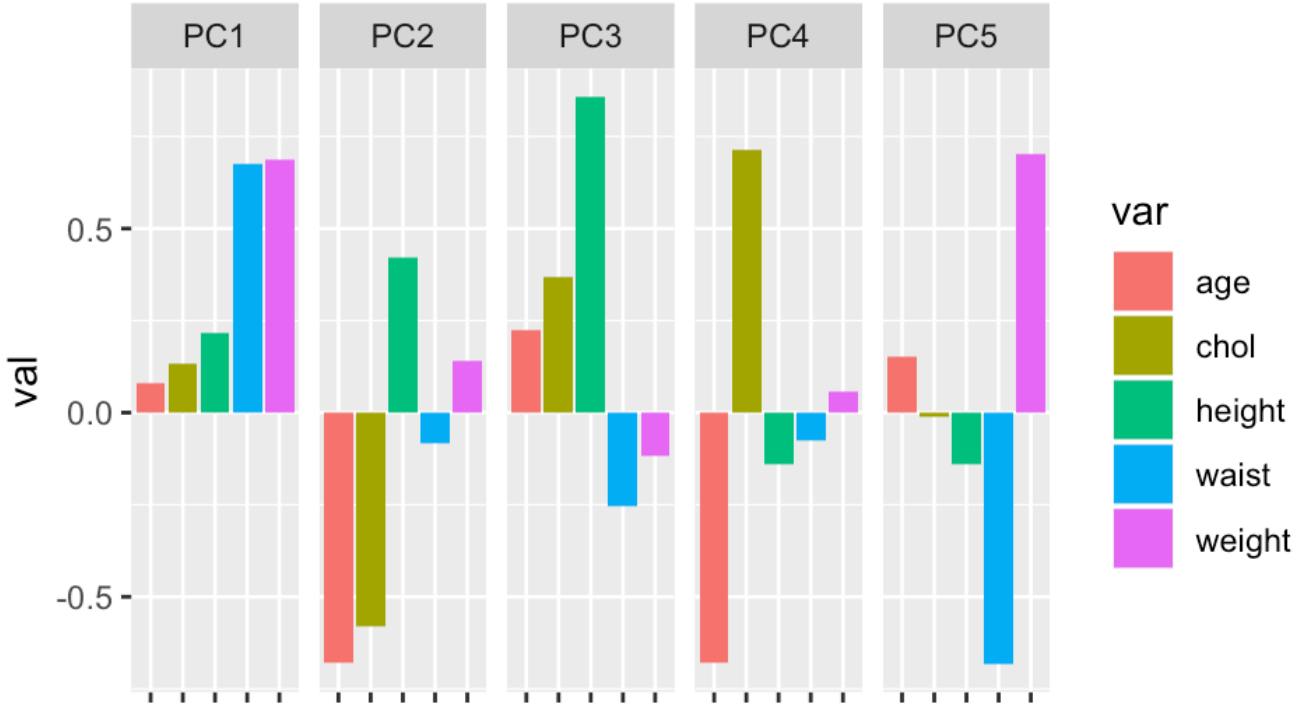
- each **dot** is a sample / patient
- new coordinate system is (PC1,PC2)
- **Red arrows** indicate the contribution of each “old" coordinate to the PCs



Principal components

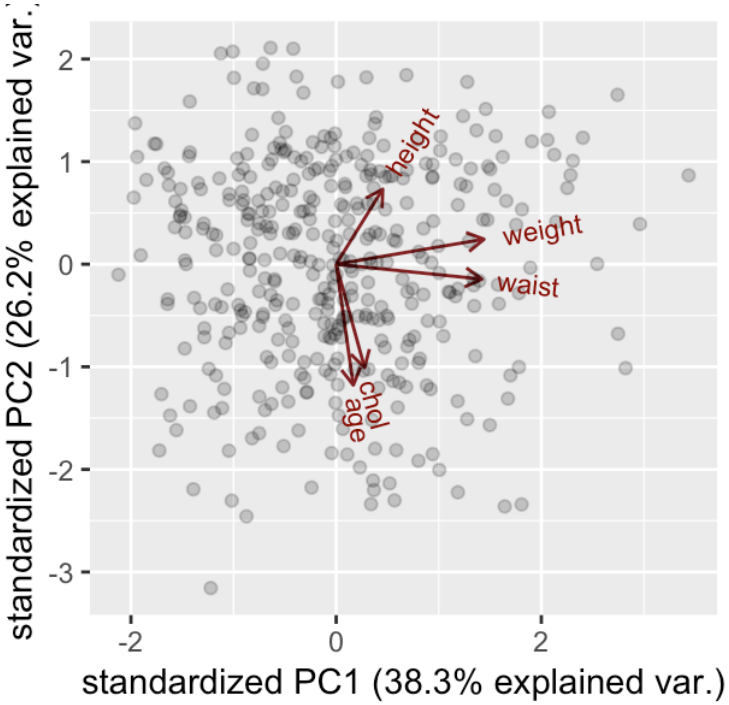
Medizinische Fakultät Heidelberg

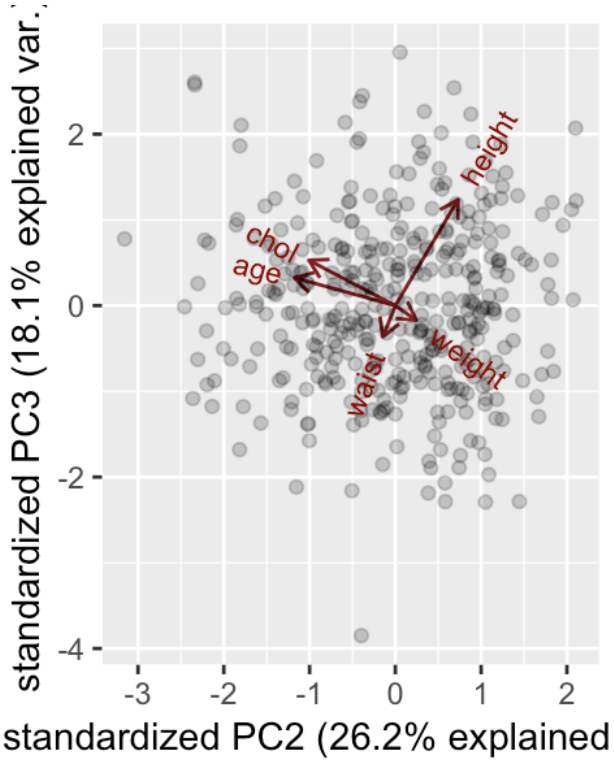
$$PC_i = \alpha_i \cdot \text{age} + \beta_i \cdot \text{chol} + \gamma_i \cdot \text{height} + \delta_i \cdot \text{waist} + \epsilon_i \cdot \text{weight}$$

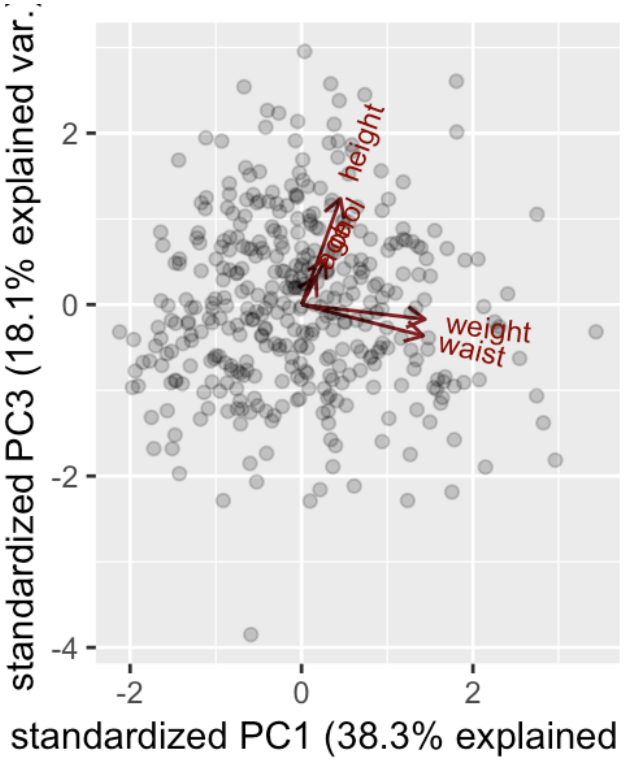


- contribution of each variable to the principal components (coefficients are called "*loadings*")
- some variables contribute in the same direction to some PCs (e.g. waist and height for PC1), but opposite to others (PC5)

Principal components



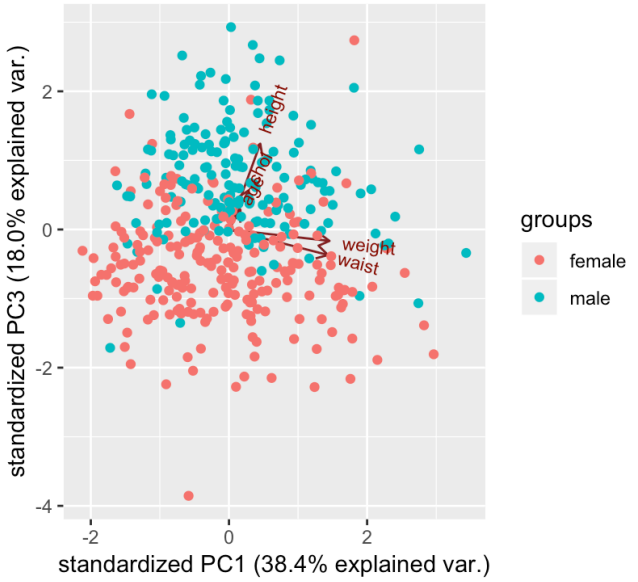




Identifying interesting PCs

Medizinische Fakultät Heidelberg

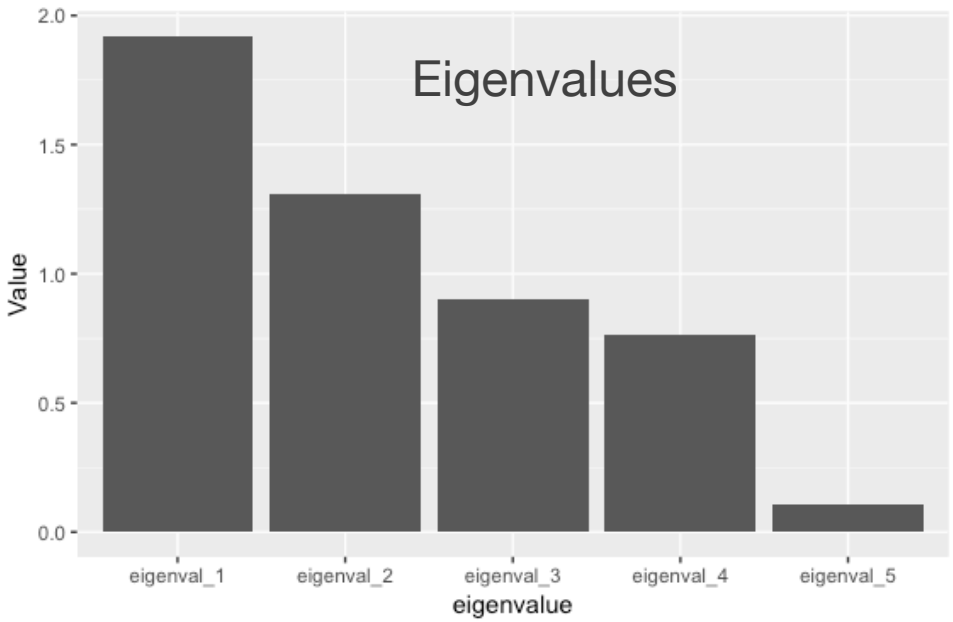
- PC plots can highlight a new group structure
- Example: **PC3** seems very associated to gender
- indicates that a combination of height and cholesterol does separate men /women

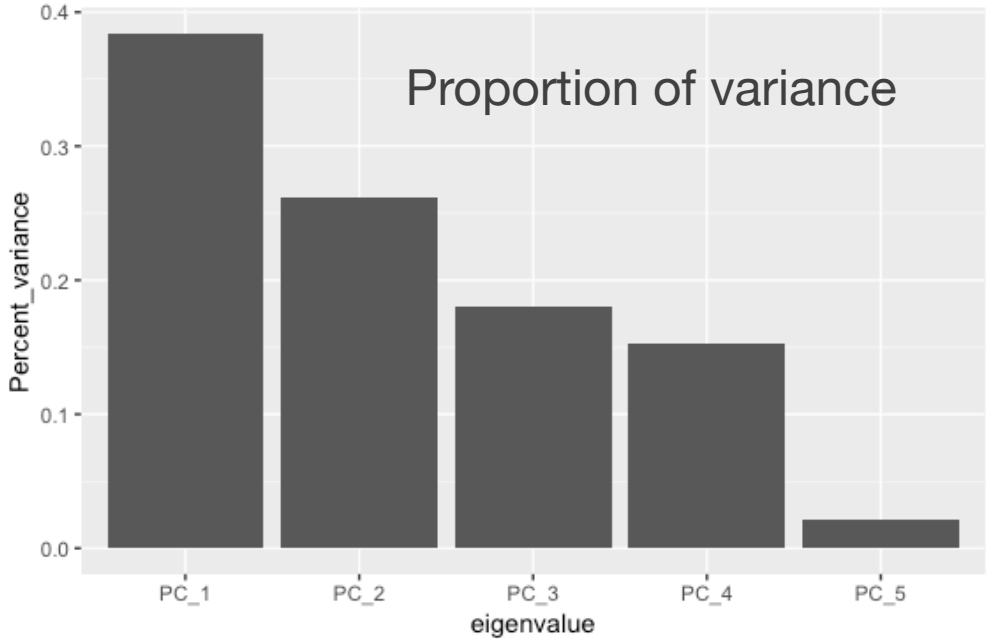


Number of PCs?

Medizinische Fakultät Heidelberg

- Each PC explains some part of the total variance of the dataset
- This amount is proportional to the corresponding **eigenvalue**
- PCs are ordered by **decreasing eigenvalue** (hence variance)





Considering PC1 & PC2 explains 63% of the total vairance

Choosing the number of PCs

- several criteria to select the optimal subset of PCs, without loosing too much information

- **Proportion of variance:**
  keep PCs such that the cumulative variance is above threshold

$$\sum_{i=1}^{k} \frac{\lambda_i}{\sum \lambda_i} \geq \text{var}_{min}$$

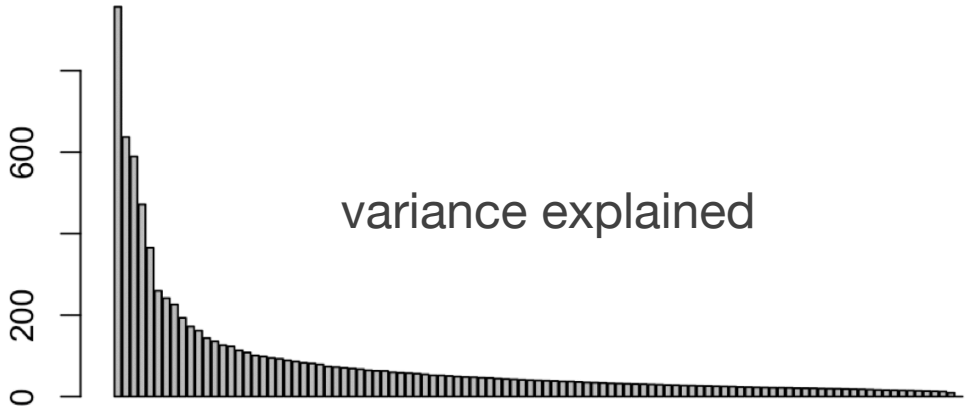
- **Average eigenvalue criteria:**
  keep PCs which have eigenvalue larger than
  - mean eigenvalue (Kaiser rule) or
  - 70% of mean eigenvalue (Jottclife rule)

$$\lambda_i \geq \bar{\lambda}$$

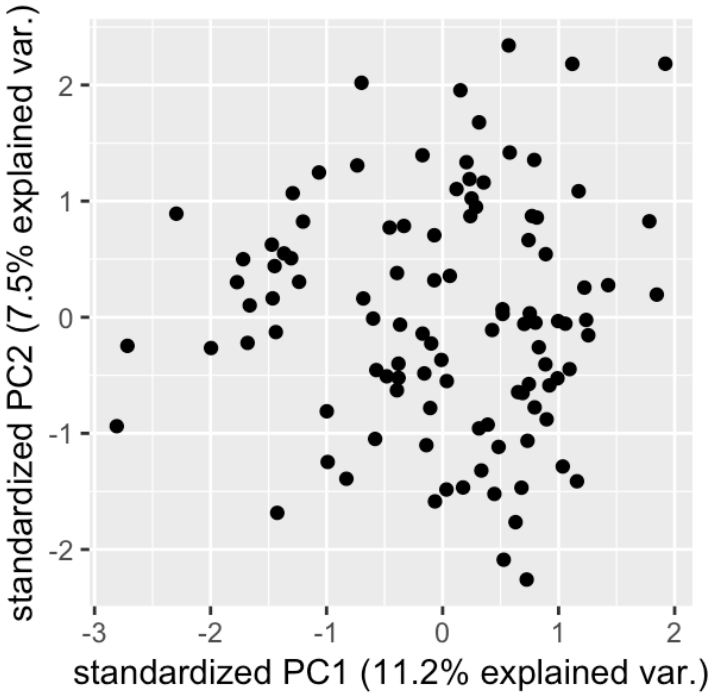
Application to gene expression

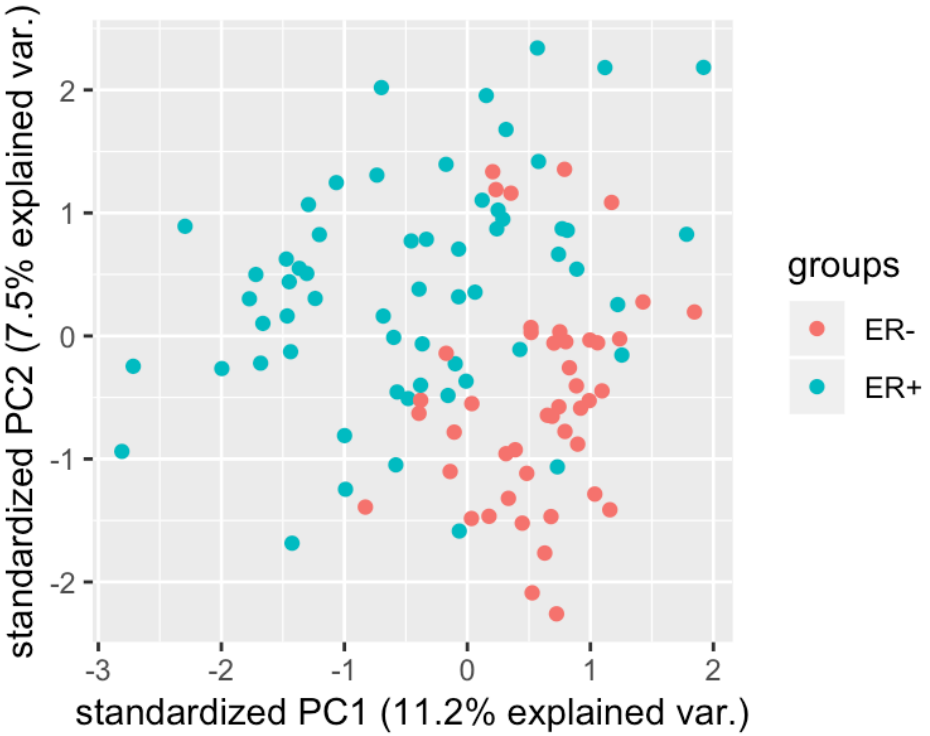
Medizinische Fakultät Heidelberg

- Gene expression dataset of **breast cancer patients**
- 2 groups: ER+ and ER- patients [M. Ringner, Nature Biotech. (2008)]
- Dimension: k = 105 patients / n = 8534 genes (here: n >> k)
- pre-processing:
  - **scale** the gene expression across patients
  - **center** the gene expression across patients
- How many principal components do we get?
  → **k-1** (this has to do with the rank of the data matrix)



Application to gene expression





- PC1 separates ER+ from ER- patients