



UNIVERSITY OF TARTU
Institute of Computer Science



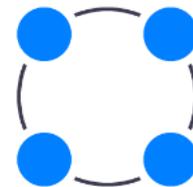
Data Preprocessing

Unsupervised learning

Elena Sügis

elena.sugis@ut.ee

Machine learning, SPB2016



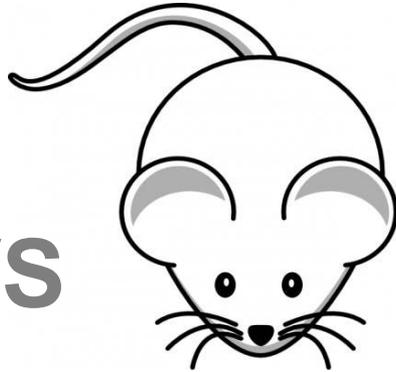
ИНСТИТУТ
БИОИНФОРМАТИКИ



Questions we ask



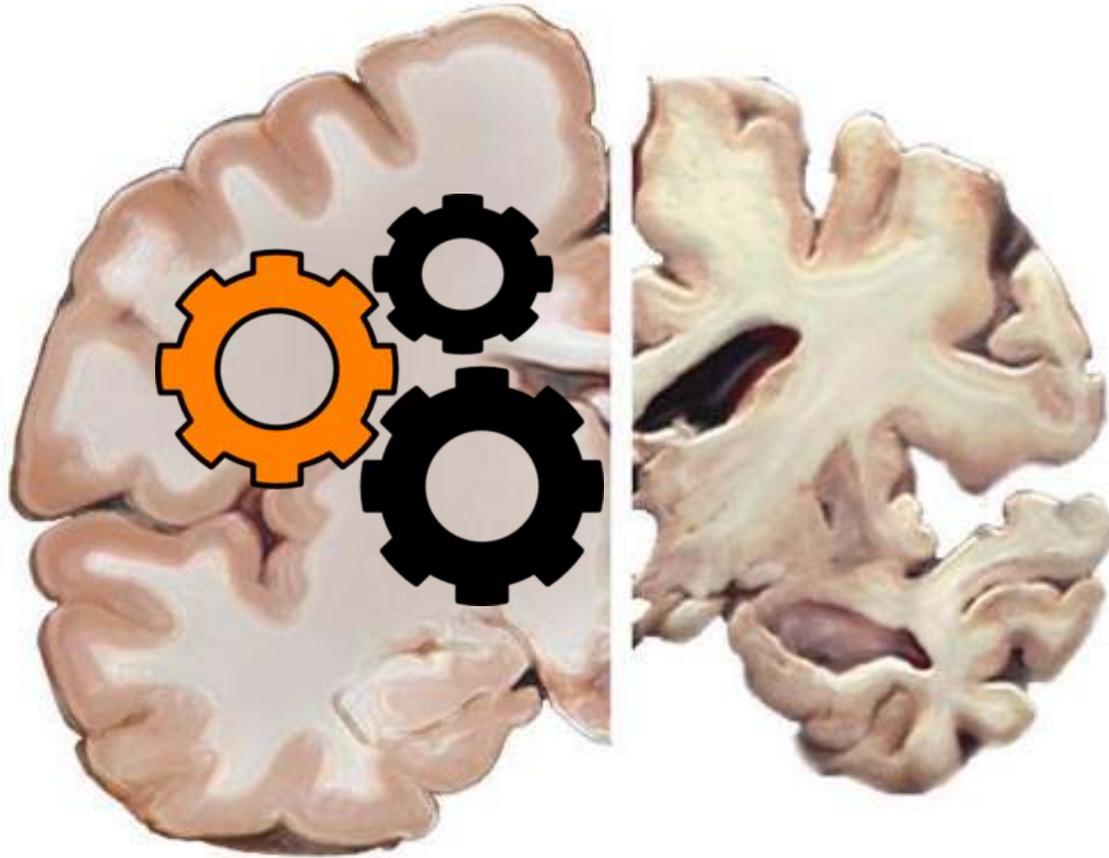
VS



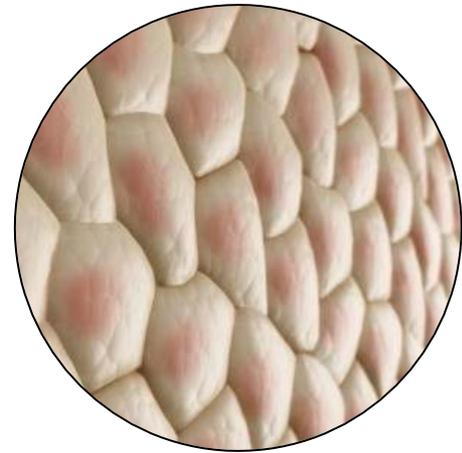
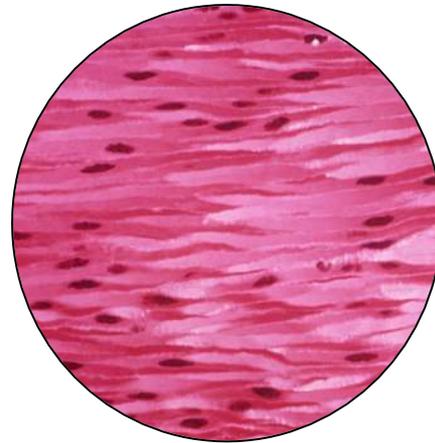
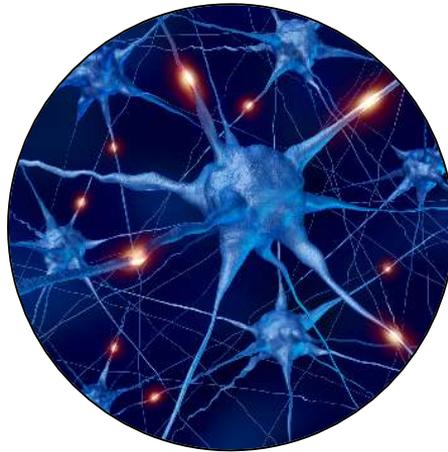
VS



Questions we ask



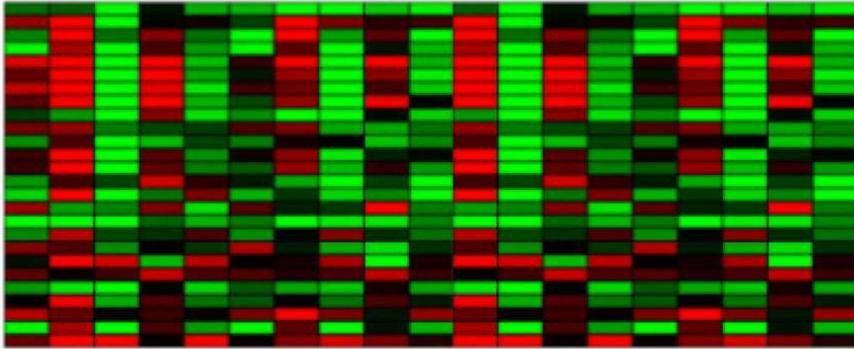
Questions we ask



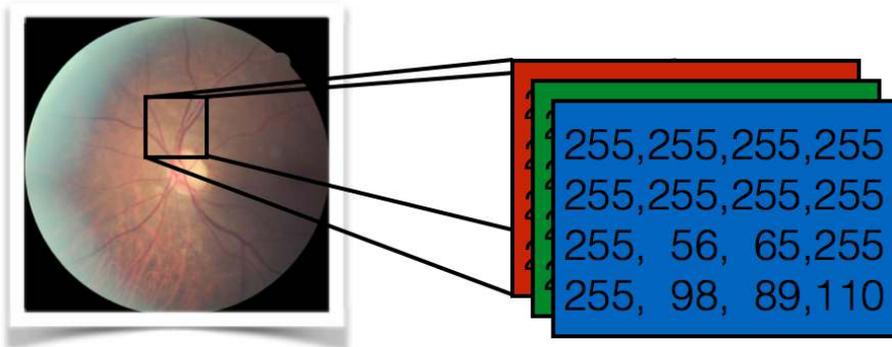
Experiments



Data comes in different forms



$x = (0.5, 0.9, 0.7, -0.3, \dots)$



$x = (255, 255, 255, 255, \dots)$

Diagnose: asthma

bite *the* **Diagnose** *low* *cancer* *not* **asthma**

$x = (0, \dots, 0, 1, 0, 0, 0, 1, \dots)$

Microsoft Excel ribbon showing Home, Layout, Tables, Charts, SmartArt, Formulas, Data, and Review tabs. The Font section includes options for Arial, size 10, bold, italic, underline, and color. The Alignment section includes options for text alignment, wrap text, and merge. The Number section includes options for number format (General), percentage, and decimal places. The Format section includes options for Normal, Bad, Neutral, and Calculation styles.

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	C21-FI	APS1-316	APS1-502	APS1-312	APS1-11_2	APS1-03_2	APS1-404	APS1-360	APS1-26_2	APS1-514	APS1-352	C355-SI_2	APS1-362	APS1-01_2	APS1-354	APS1-509	APS1-20_2	C APS1-5	APS1-2
2	-0,4887	0,234767	0,22172	0,228849	-0,06724	0,198947	0,399174	-0,17976	0,120272	0,508874	0,23948	0,18876	-0,49479	0,403385	-0,02036	0,2073	0,01052	0,655229	0,350
3	-0,56255	1,900956	0,729879	0,360465	0,809556	0,511848	-0,66508	-0,6931	-0,69132	0,122348	-0,24677	0,0434	-0,80776	0,912349	0,626272	0,575758	1,131717	0,646509	0,593
4	0,526629	0,37935	0,693528	1,077765	0,692455	0,83677	0,940985	0,737664	0,722068	0,923591	0,697278	0,728161	0,459127	0,997002	0,719225	0,694275	0,58619	0,823099	0,537
5	0,826536	0,592142	0,179319	2,114469	1,34637	0,601974	1,131415	0,809367	-0,29039	1,568368	0,358889	-0,18146	-0,96937	1,060401	0,829092	0,150945	-0,71767	-0,59129	-1,09
6	-1,04294	-0,9215	0,860857	0,57745	-0,61676	-0,98568	0,152973	-1,03148	2,196609	-1,05021	-0,59313	-0,71867	-1,06424	-0,93653	-0,92463	-0,2697	-1,07338	-0,53509	-1,05
7	-0,86355	1,293565	0,632034	0,211295	-1,27675	1,477679	-0,77714	0,301728	-0,58079	-0,59545	-0,83298	-0,56856	-1,2272	0,266086	-1,62381	-0,71642	-0,31974	-1,82247	0,647
8	-0,23357	0,83206	-1,52215	-0,09742	-1,06264	1,642972	-0,5063	0,606325	0,070995	0,222972	-0,06913	0,420075	-1,53497	0,198994	-2,20022	-0,74523	-0,68012	-1,0979	0,659
9	-1,3041	0,097798	-0,97237	-0,66157	-1,20775	0,829845	-0,97955	-0,34582	-0,85999	-1,04579	-0,24335	-0,14158	-1,102	-0,21181	-1,62612	-0,93486	-0,75502	-1,10553	-0,14
10	-0,87564	0,630128	0,369398	0,028096	-0,96322	1,161014	-1,31319	-0,1592	-0,39759	-0,55606	-0,69929	-0,62458	-1,23338	0,005355	-2,30668	-0,80058	-1,04211	-1,89696	0,702
11	-1,6572	0,374363	0,065398	0,371467	-0,866728	1,934384	-0,64302	0,574107	0,024587	0,377852	-0,56434	0,306704	-0,94488	-0,29084	-0,83979	-0,4002	0,13761	-1,05651	1,036
12	-1,27014	1,844753	-0,21737	0,305616	-0,10246	0,846505	-1,01227	-0,46867	0,345629	0,150133	-0,34854	0,32053	-0,91744	0,427385	-0,86404	-0,84993	-0,54509	-1,23573	0,280
13	-1,06418	-0,23232	-0,73055	-0,15014	-0,73944	-0,56059	-1,28617	-0,9307	-0,56149	-0,56702	-0,42032	-0,00241	-0,66564	-0,24882	0,002075	-0,30149	-1,13682	-1,22597	-0,31
14	0,1685	0,923524															-0,59357	-1,93058	-1,15
15	-1,99211	-0,1565															-1,0553	-1,31462	1,088
16	-1,15695	1,192555															-0,92405	-0,96535	0,532
17	-0,77825	0,88967															-0,41446	-1,72813	0,003
18	-0,75095	0,771783															-1,32835	-1,89204	-0,41
19	0,124064	1,289183															-0,67996	-2,11821	0,49
20	-1,05107	1,007949															-0,59897	-1,01041	0,329
21	-1,68098	0,24404															-0,90058	-1,23607	0,411
22	-0,61673	0,936892															-0,69683	-0,93969	0,905
23	-0,98689	0,553994															-0,17849	-0,20448	1,609
24	-0,58859	0,350248															-0,70385	-1,99204	0,95
25	-1,32352	1,591618															-0,45297	-0,70038	2,088
26	-1,15175	0,780826															-0,9364	-1,03009	0,872
27	-0,48428	-0,18117															-1,15534	-1,60816	1,442
28	-1,29442	0,879724	0,463348	-0,12957	-0,04748	1,085606	-0,24867	3,394451	-0,16635	0,944466	-0,34051	0,898699	-0,93194	0,551689	0,241056	-0,37739	-0,9511	-1,78407	0,749
29	-1,4557	-0,09856	-0,36621	-0,21392	-0,25299	1,736314	-0,07791	-0,41333	0,154922	0,055765	-0,45575	-0,2442	0,290909	-0,03705	-1,51206	-1,33489	-0,52094	-1,5145	1,319
30	-0,90803	1,029048	-0,68845	0,074348	-0,33582	2,164151	-0,45578	0,972013	0,714796	1,124594	0,26288	-0,25571	-0,55467	-0,23539	-0,58071	-1,08873	-0,84031	-0,042	0,610
31	-0,87354	1,094595	-0,03171	-0,56182	-0,59634	-0,33306	-0,73437	-0,8253	1,183227	0,355806	-0,38504	0,283831	-0,73996	0,52729	-1,88813	-1,00116	-1,21085	-2,16812	0,385
32	-0,82	1,261128	0,243196	0,517289	0,750895	1,194548	-0,12374	-0,11973	0,76713	1,125359	-0,70457	0,047693	-1,04306	-0,07699	-1,19814	-0,86195	-1,00517	-1,74549	1,215
33	-0,78018	1,615431	-0,28837	0,924789	-0,56636	0,892182	-1,37028	-0,02633	0,293597	1,069584	-0,92696	-0,08003	0,556214	1,860747	-1,58945	-1,12683	-1,36817	-1,13405	1,557
34	-0,08889	-0,14304	-0,65432	-0,94164	-0,42861	2,349137	1,430127	-0,76511	-0,58829	0,870025	-0,00874	0,908636	-0,18801	-0,35026	-1,48148	-1,24546	-0,8854	-1,16535	-0,21
35	-0,53756	0,768457	-0,89423	0,007531	-0,98362	2,166296	-0,93857	2,122845	0,136294	0,274948	0,585357	0,217324	-0,43374	0,833201	-0,62019	-0,00437	-0,82566	-1,43261	-0,20
36	-0,91029	1,137358	-0,78256	0,157272	-1,32158	0,807747	-1,40975	-0,10943	-1,61007	-1,59391	-1,21212	-0,85131	-0,99906	-0,14809	-1,65978	-1,39265	-1,16014	-1,91198	0,117
37	-1,06434	0,920069	0,485924	-0,39182	-1,38788	0,875248	-0,10139	0,293085	-0,65016	0,805495	-0,55107	0,461659	-0,99427	0,037787	-1,73893	-1,6779	-0,60676	-1,43584	1,282
38	-1,2704	0,575787	-0,33616	-0,13808	-0,81116	-0,18085	-0,85878	-1,30527	1,49447	-0,85285	2,201395	-0,98317	-0,10382	-1,65102	-1,17816	-0,3144	-0,4824	0,690	
39	-0,06869	1,451572	0,330493	0,424922	-0,35006	1,557744	-0,84059	0,551154	-0,32348	-0,52607	-0,6899	-1,18485	-0,13469	0,461043	1,438855	-0,69172	-0,91765	-0,9455	-0,12
40	-0,21291	1,085538	0,721041	1,026748	-0,91361	2,11202	-0,84595	0,393947	-1,27669	-0,24826	0,070477	0,184991	-0,14065	0,788965	-0,31057	-0,56817	-0,9869	-1,88387	0,296
41	-1,31421	-0,04399	-0,87947	0,124897	-1,7042	1,25426	-1,14679	-0,59666	-0,42305	1,155636	-0,81311	-0,40797	-0,10097	1,149502	-1,11011	-0,48103	-1,39916	-1,39502	-0,20
42	1,395162	0,887155	0,782284	-1,11207	-1,268	-0,16978	-0,34473	2,488763	-0,2545	-0,2045	-0,37378	-0,32574	0,993644	1,113215	-1,3905	-0,72315	-0,84154	-1,40338	0,506
43	0,205675	1,798406	0,161273	0,873045	-0,74173	1,999119	-0,21271	0,950538	-0,59416	0,075442	-0,81113	0,798829	-0,51307	0,589372	-1,60255	-0,40974	4,715026	1,14736	0,125
44	0,912575	0,440841	0,041677	0,358948	0,497338	-0,41173	0,28027	0,740028	-0,28786	0,937606	1,373673	1,108055	-0,13597	1,216988	-0,74516	-0,84187	-0,17839	-1,22191	-0,00
45	0,255077	1,441045	0,753418	0,560932	-0,74819	3,24302	0,455104	2,07394	-0,15657	0,279524	0,186881	0,038683	0,00062	0,372195	-0,68122	0,289968	-0,76888	-0,36963	-0,14
46	0,7266	0,990771	0,693198	3,311101	-0,86564	0,099075	-0,3591	0,098547	-1,02008	1,371354	1,551675	-0,30888	-0,62109	0,285789	-0,80572	-0,70176	-0,59881	-1,28081	0,368
47	-0,30744	0,374103	-0,28648	-0,33953	-0,26302	0,127071	-0,65487	0,379004	-0,71526	-0,08089	-0,58992	0,066524	-1,13557	-0,27846	-1,57127	-1,4171	-0,61321	-1,34249	0,655
48	-0,06002	1,756581	0,722590	0,452448	-1,02287	1,571101	-0,19671	0,54311	-1,08292	-0,41464	-0,57415	-0,40354	-0,08513	0,414661	-1,29962	-0,57578	-0,86634	-1,13518	0,513
49	-1,45918	1,226267	0,29335	-0,81014	-1,11937	0,140541	-0,16811	-0,11347	-0,89466	-0,26755	-0,04944	0,392566	-0,74152	0,526024	-1,71912	-0,56683	-0,76126	-1,19997	0,504

Data ≠ Knowledge

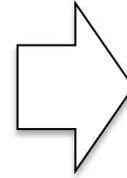
Simple data analysis pipeline



high quality data



machine learning
method



awesome result

Simple data analysis pipeline



poor quality data



machine learning
method



not so
awesome result

Clean

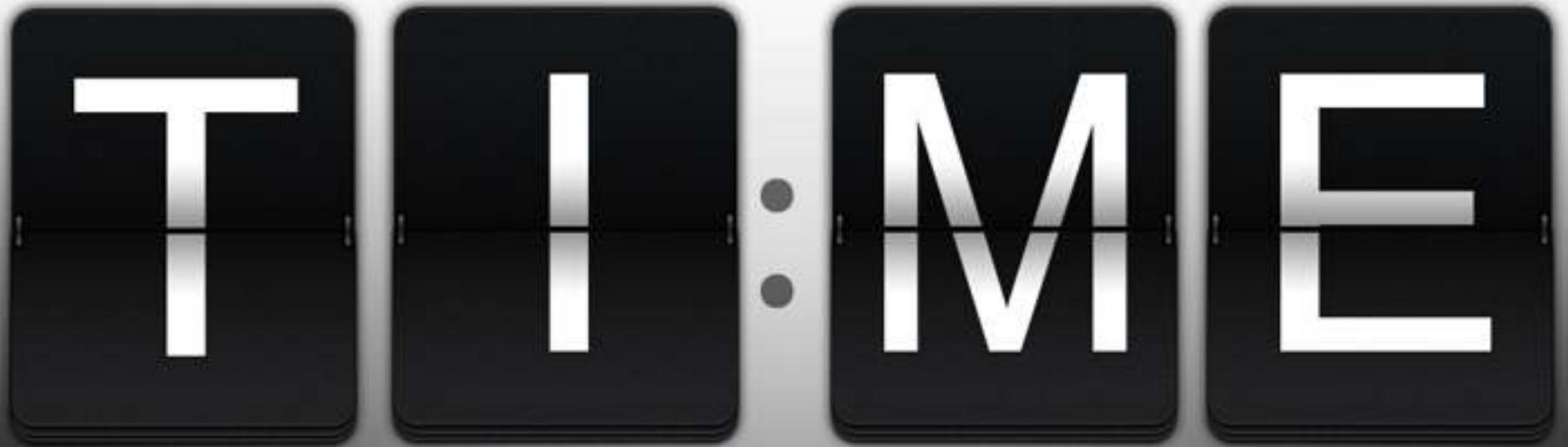


Massage your data



80 %

T I M E



Interpretation
validation

Import
data

Data
analysis

Handle
outliers

Normalize/
Standardize

Impute missing
values



?

Interpretation
validation

Import
data

Data
analysis

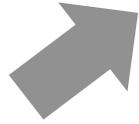
Summarize/
plot raw data



Handle
outliers

Impute missing
values

Normalize/
Standardize



Meet your data



H001		H002		H003		H004		H005	
Min.	:-18.057485	Min.	:-14.8885	Min.	:-19.497	Min.	:-21.1075	Min.	:-21.0675
1st Qu.:	0.004001	1st Qu.:	-2.5086	1st Qu.:	-2.823	1st Qu.:	-3.1494	1st Qu.:	-3.3773
Median :	3.036367	Median :	1.0200	Median :	1.219	Median :	1.0470	Median :	1.3058
Mean :	1.406008	Mean :	-0.2383	Mean :	-0.195	Mean :	-0.3506	Mean :	-0.1845
3rd Qu.:	4.752667	3rd Qu.:	3.4390	3rd Qu.:	3.567	3rd Qu.:	3.3603	3rd Qu.:	4.6048
Max. :	7.243000	Max. :	8.0437	Max. :	8.777	Max. :	9.0990	Max. :	8.3813
NA's :	:5	NA's :	:5	NA's :	:4	NA's :	:3	NA's :	:4

Interpretation
validation

Import
data

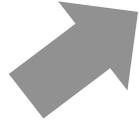
Data
analysis

Summarize/
plot raw data

Handle
outliers

Impute missing
values

Normalize/
Standardize



Missing Values

Origins:

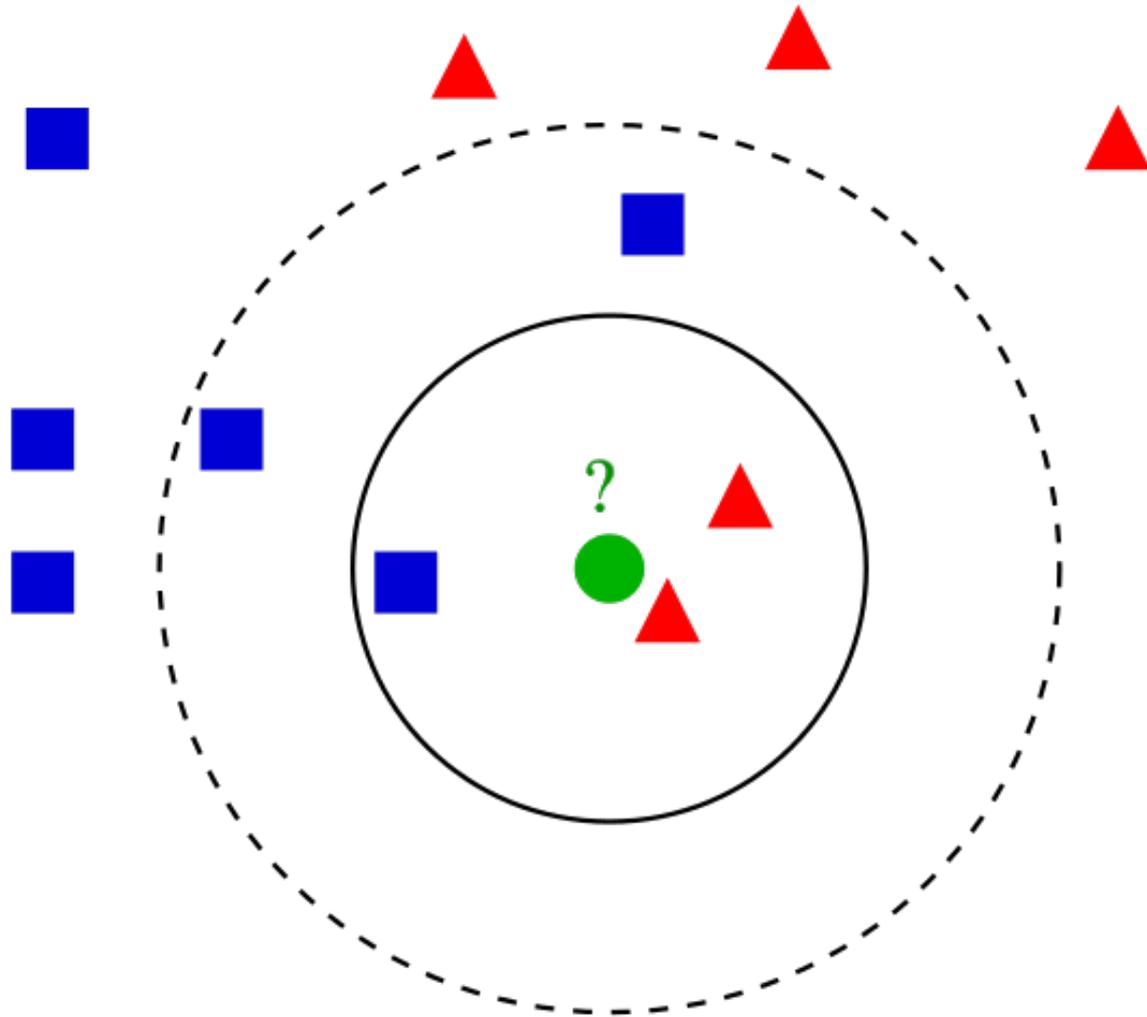
- Malfunctioning measurement equipment
- Very low intensity signal
- Deleted due to inconsistency with other recorded data
- Data removed/not entered by mistake

Missing Values

How to deal with them:

- Filter out
- Replace missing values by 0
- Replace by the mean, median value
- K nearest neighbor imputation (KNN imputation)
- Expectation—Maximization (EM) based imputations

k-nearest neighbors

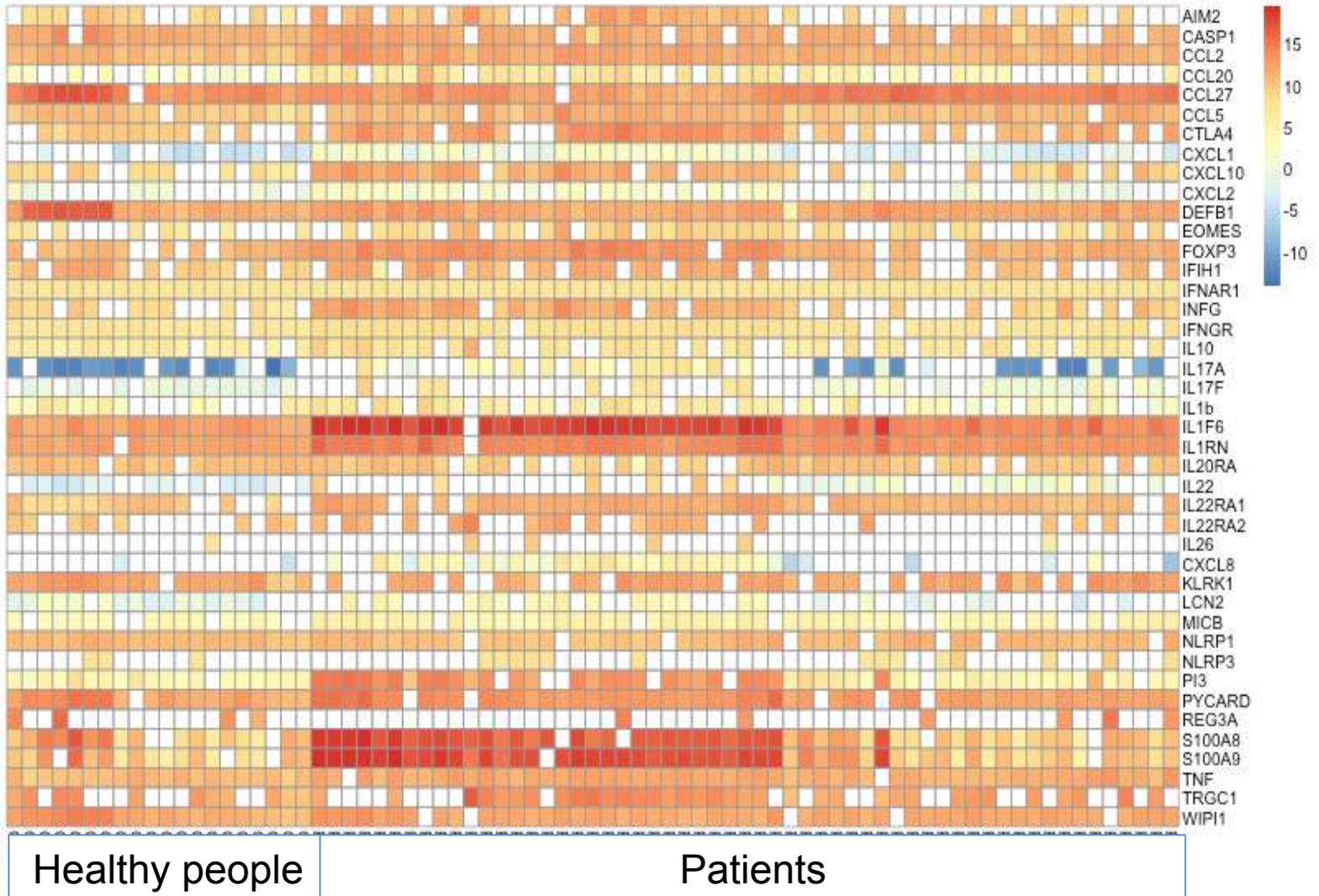


KNN

- We are given a gene expression matrix M
- Let $X=(x_1, x_2, \dots, x_i, \dots, x_n)$ be a vector in the matrix M with a missing value at x_i at the dimension i
- Find in the gene expression data matrix matrix vectors X_1, X_2, \dots, X_k , such that they are the k closest vectors to X in M (with a chosen distance measure) among the vectors that do not have a missing value at dimension i
- Replace the missing value x_i with the mean (or median) of $X_{1_i}, X_{2_i}, \dots, X_{k_i}$, i.e., mean (median) of the values at dimension i of vectors X_1, X_2, \dots, X_k

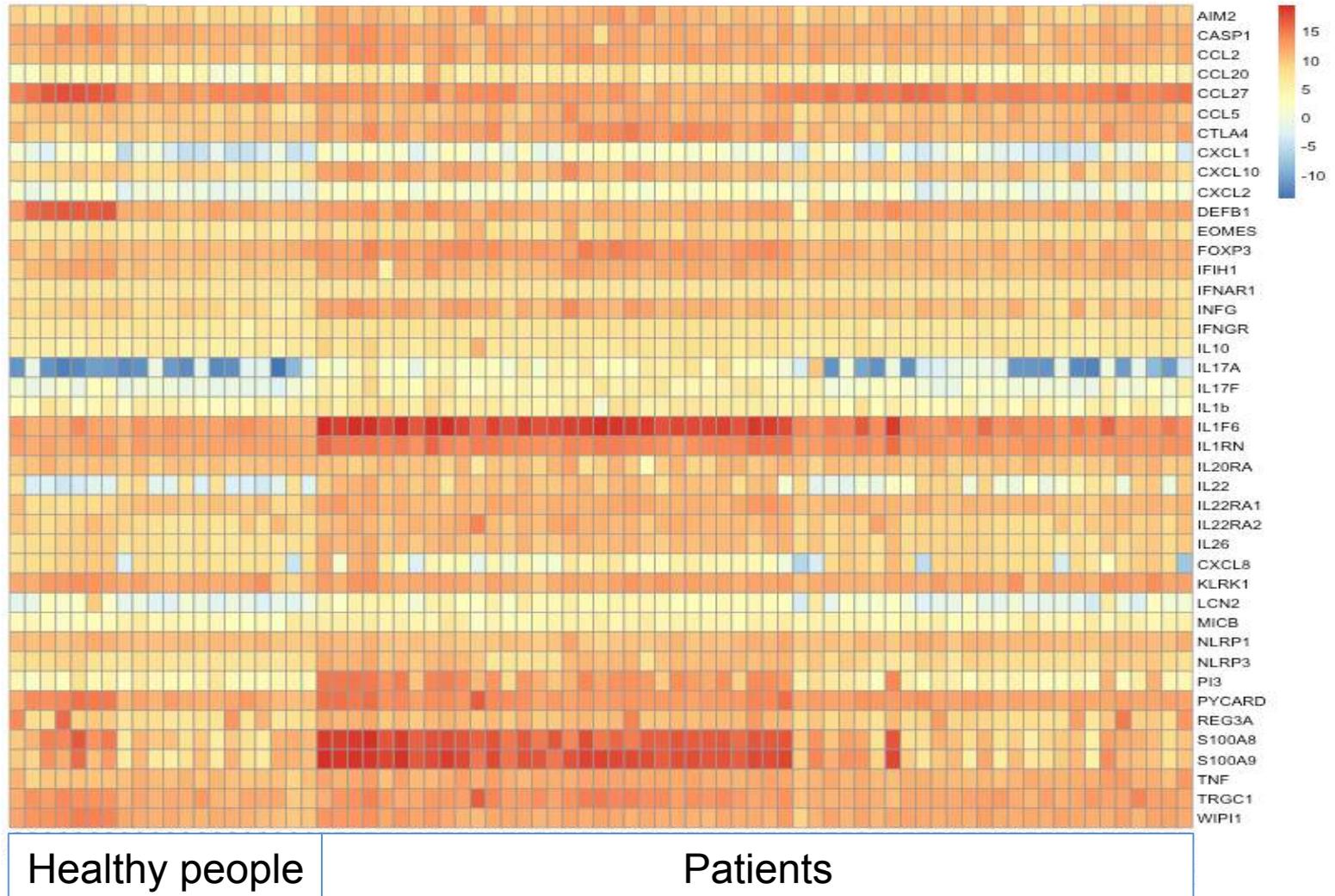
KNN

Gene expression matrix



Imputed missing values

Gene expression matrix



Interpretation
validation

Import
data

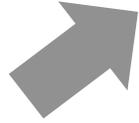
Data
analysis

Summarize/
plot raw data

Handle
outliers

Impute missing
values

Normalize/
Standardize



Technical vs Biological



Normalization & Standardization

Objective:

adjust measurements so that they can be appropriately compared among samples

Key ideas:

- Remove technological biases
- Make samples comparable

Methods:

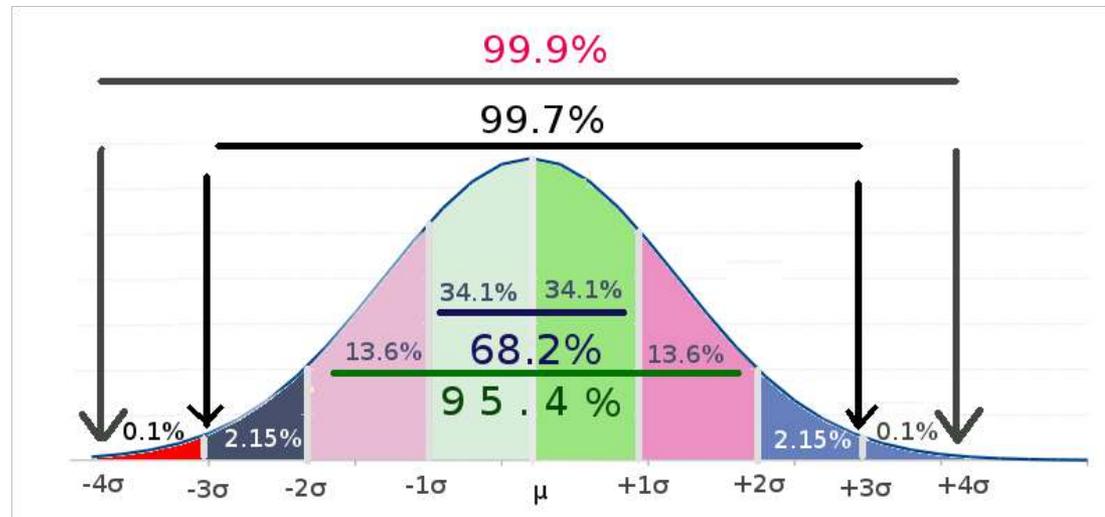
- Z-scores (centering and scaling)
- Logarithmization
- Quantile normalization
- Linear model based normalization

Z-scores

Centering a variable is subtracting the mean of the variable from each data point so that the new variable's mean is 0.

Scaling a variable is multiplying each data point by a constant in order to alter the range of the data.

$$z = \frac{x - \mu}{\sigma}$$

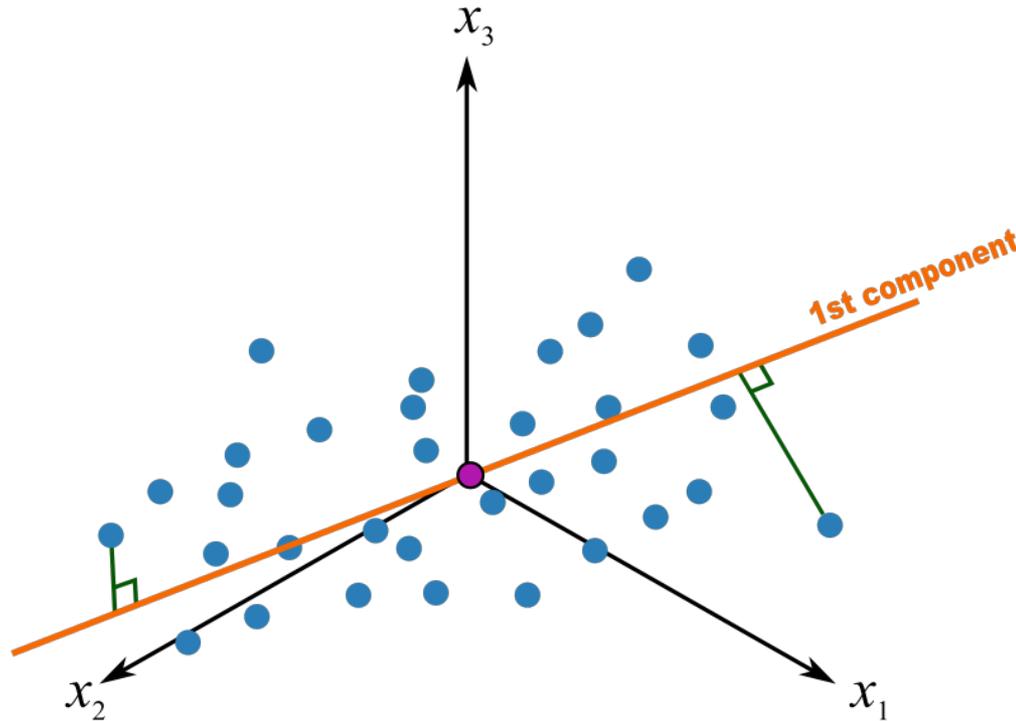


where:

μ is the mean of the population.

σ is the standard deviation of the population.

Principal Component Analysis

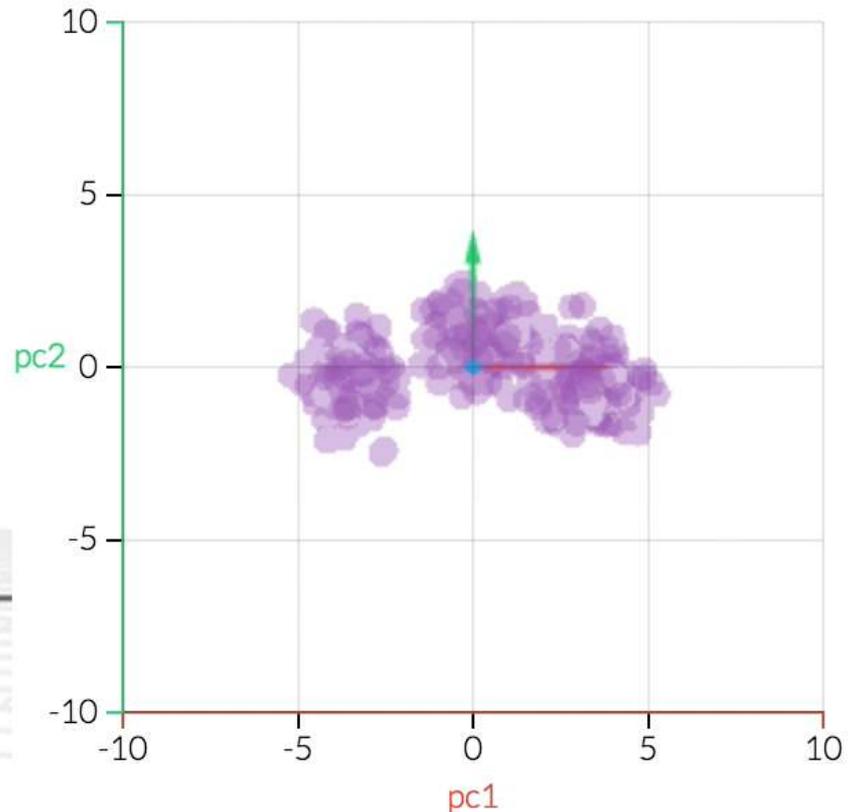
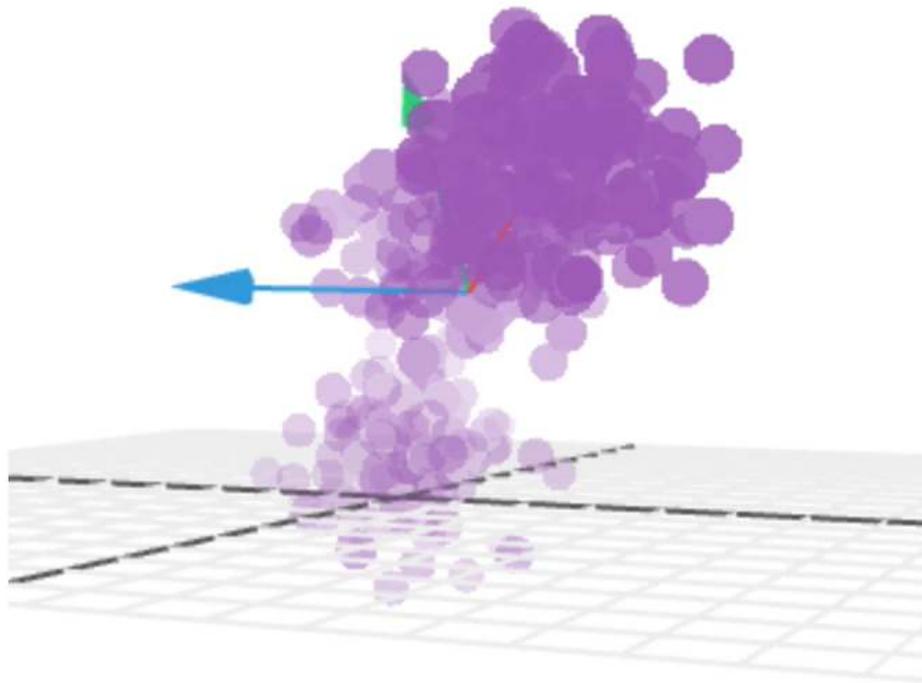


transforms the data by a **linear projection** onto a **lower-dimensional space** that **preserves** as much **data variation** as possible

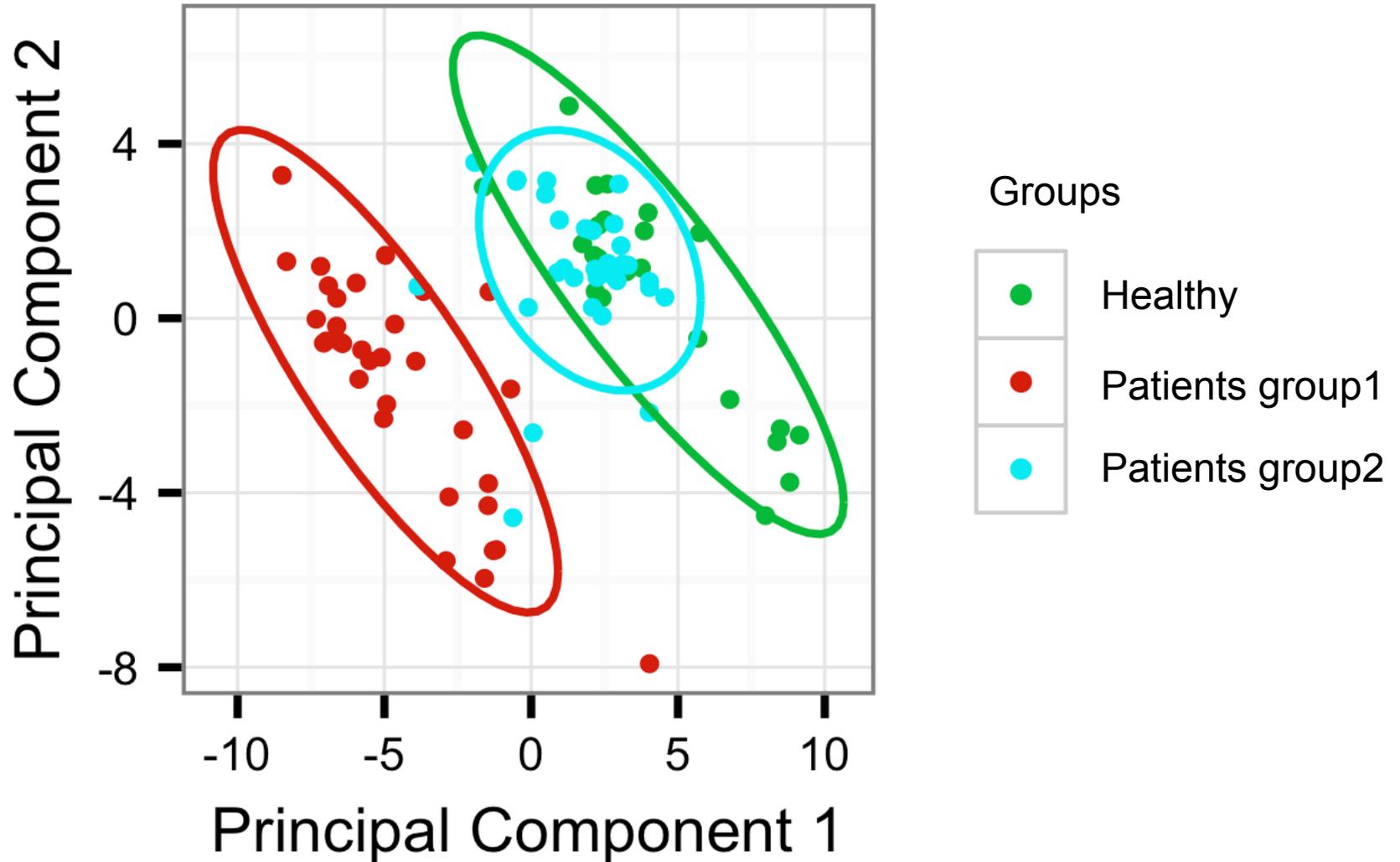
Principal Component Analysis

Objective:

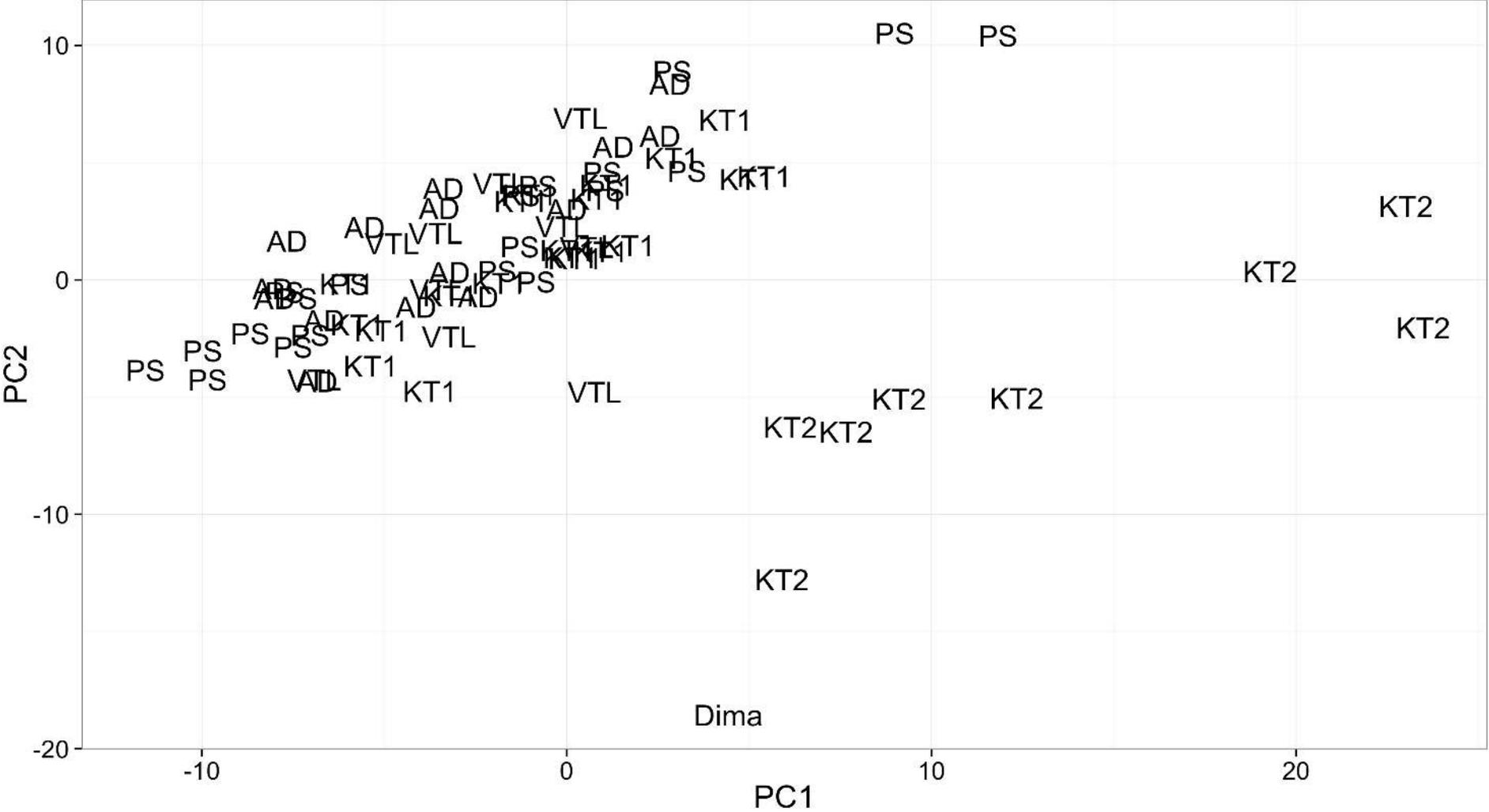
Reduce dimensionality while preserving as much variance as possible



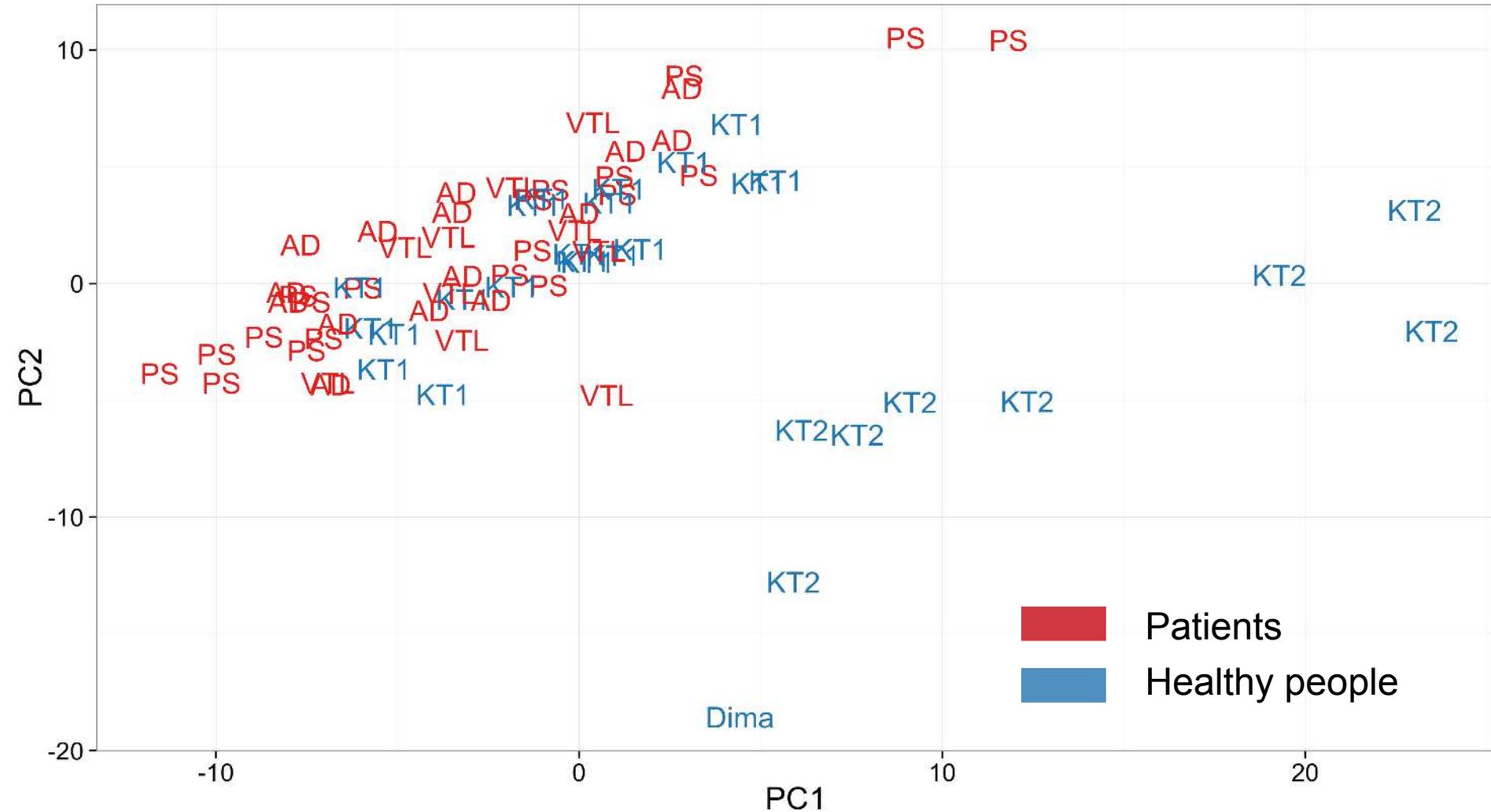
Visualize normalized data



Visual inspection after normalization



Visual Inspection. PCA



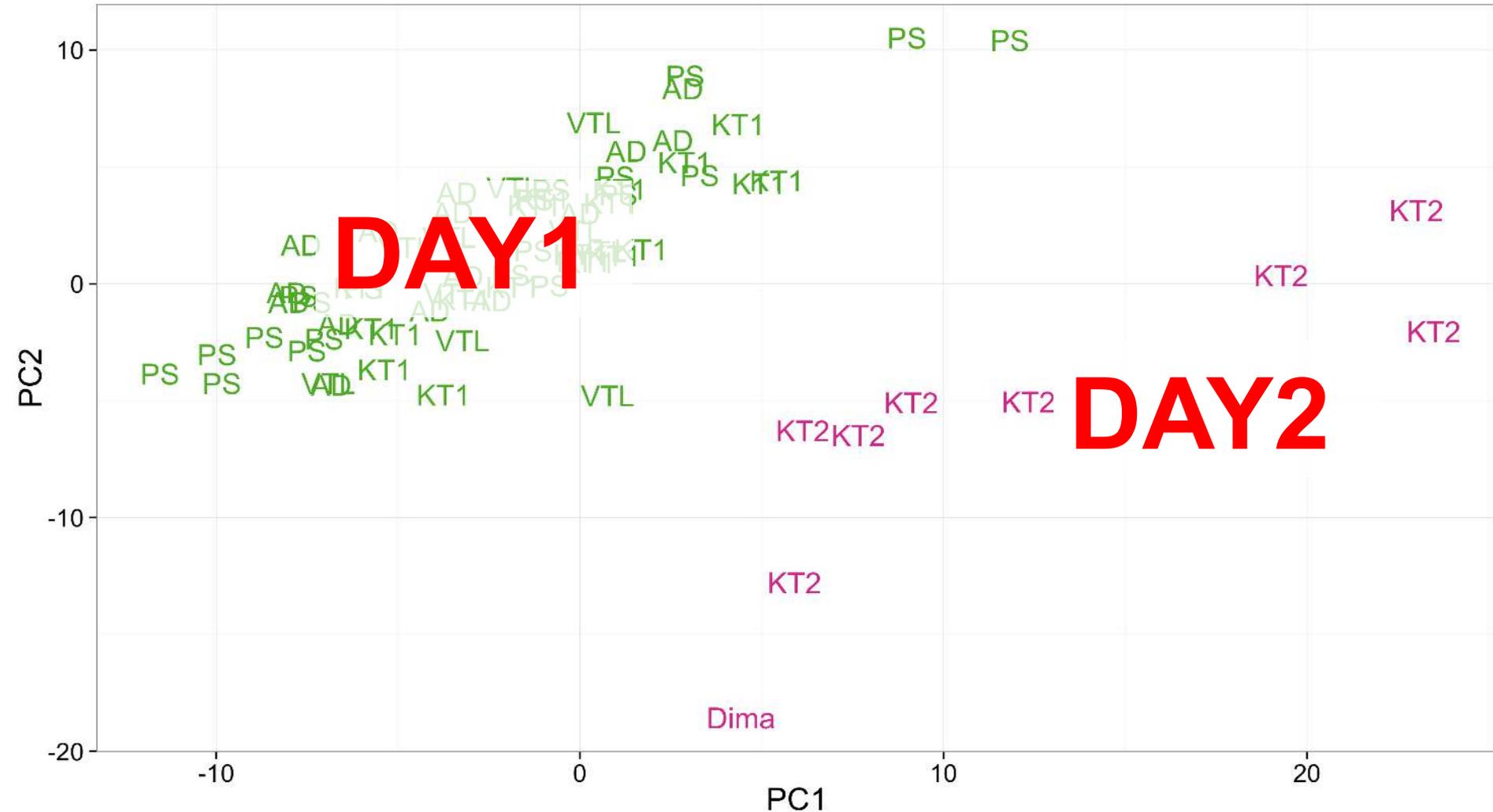
Highlight groups

Arrrgh!!!

**Why aren't you
together ?!?!**



Visual Inspection. PCA



Color by experiment/dataset/day

Batch Effects

are technical sources of variation that have been added to the samples during handling. They are unrelated to the biological or scientific variables in a study.

Measurements are affected by:

- Laboratory conditions
- Reagent lots
- Personnel differences

Major problem :

might be **correlated with** an **outcome of interest** and lead to **incorrect conclusions**

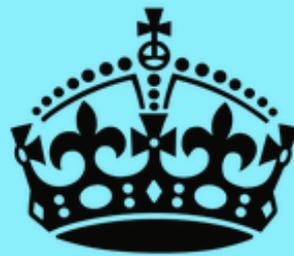
Fighting The Batch Effects

Experimental design solutions:

- Shorter experiment time
- Equally distributed samples between multiple laboratories and across different processing times, etc.
- Provide info about changes in personnel, reagents, storage and laboratories

Statistical solutions:

- **ComBat**
- **SVA**(Surrogate variable analysis, SVD+linear models)
- PAMR (Mean-centering)
- DWD (Distance-weighted discrimination based on SVM)
- Ratio_G (Geometric ratio-based)



**KEEP
CALM
AND
MESSAGE
YOUR DATA**

Interpretation
validation

Import
data

Data
analysis

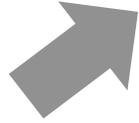
Summarize/
plot raw data



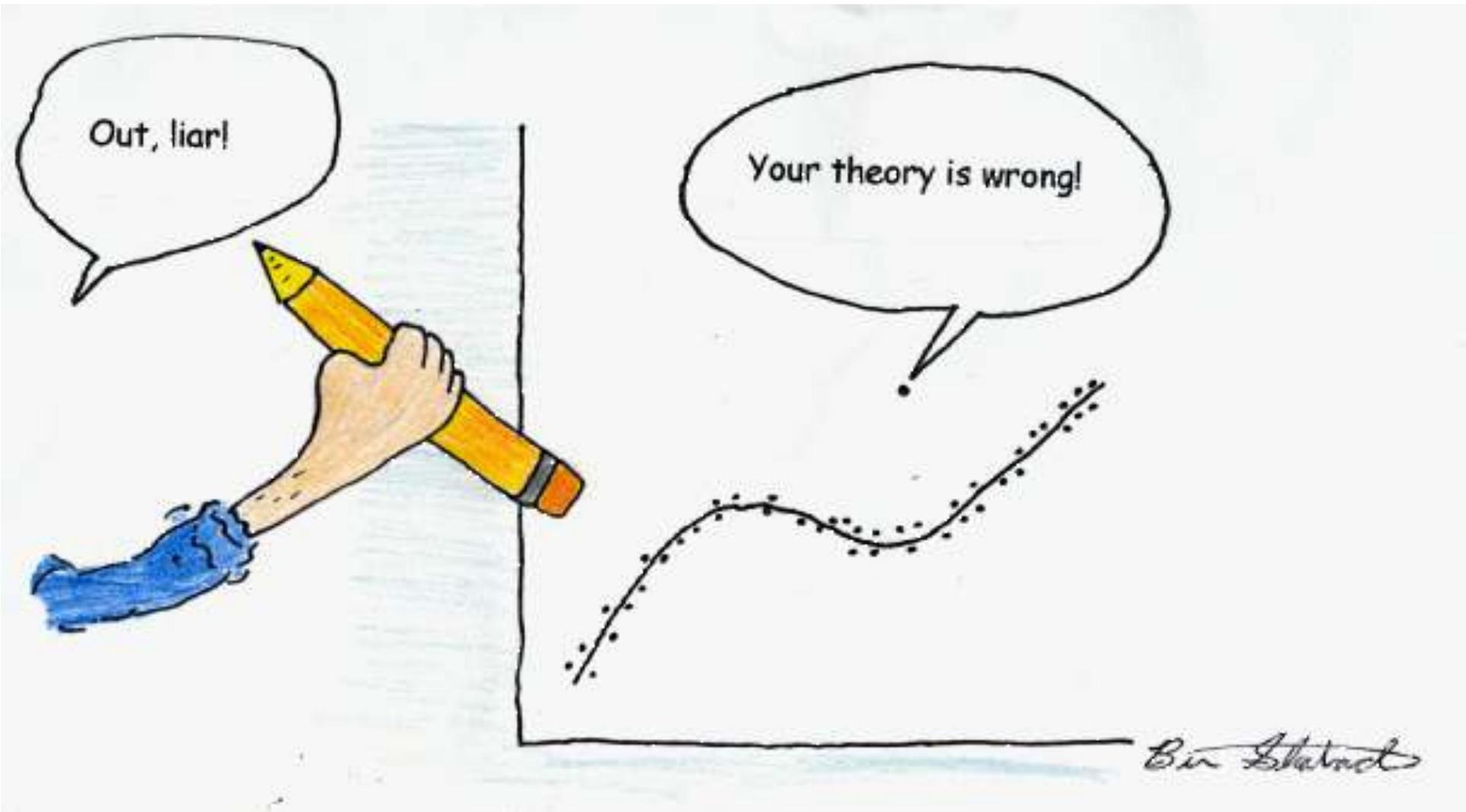
Handle
outliers

Impute missing
values

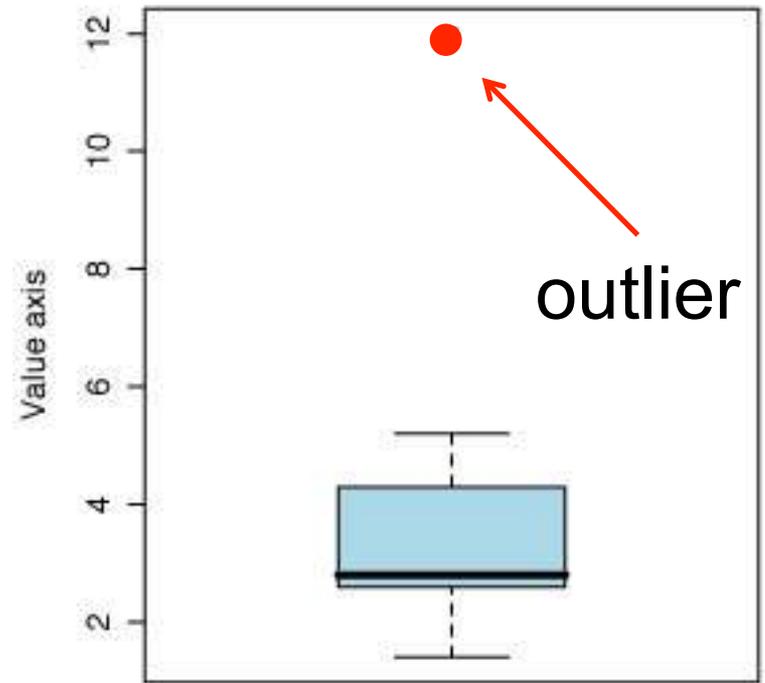
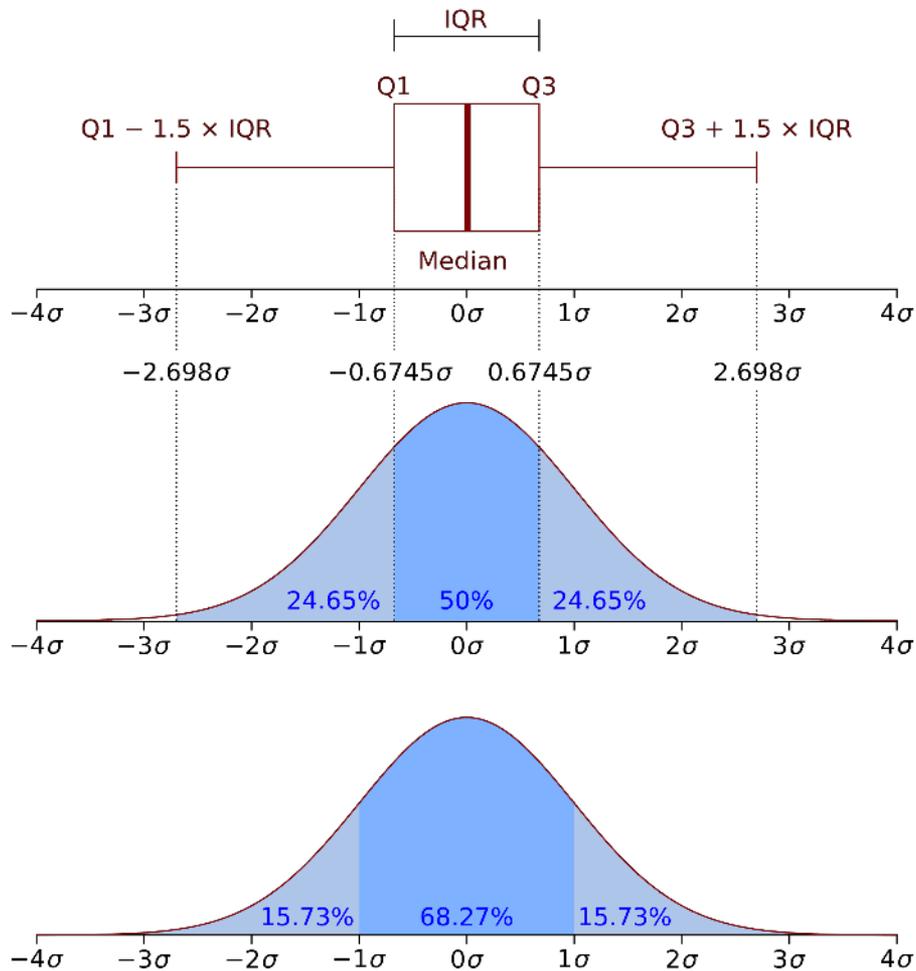
Normalize/
Standardize



Outliers Detection



Interquartile range



Single sample



Interpretation
validation

Import
data

Data
analysis

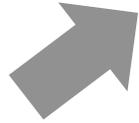
Summarize/
plot raw data



Handle
outliers

Impute missing
values

Normalize/
Standardize



**IF YOU TORTURE
THE DATA
LONG ENOUGH
IT WILL CONFESS
TO ANYTHING**

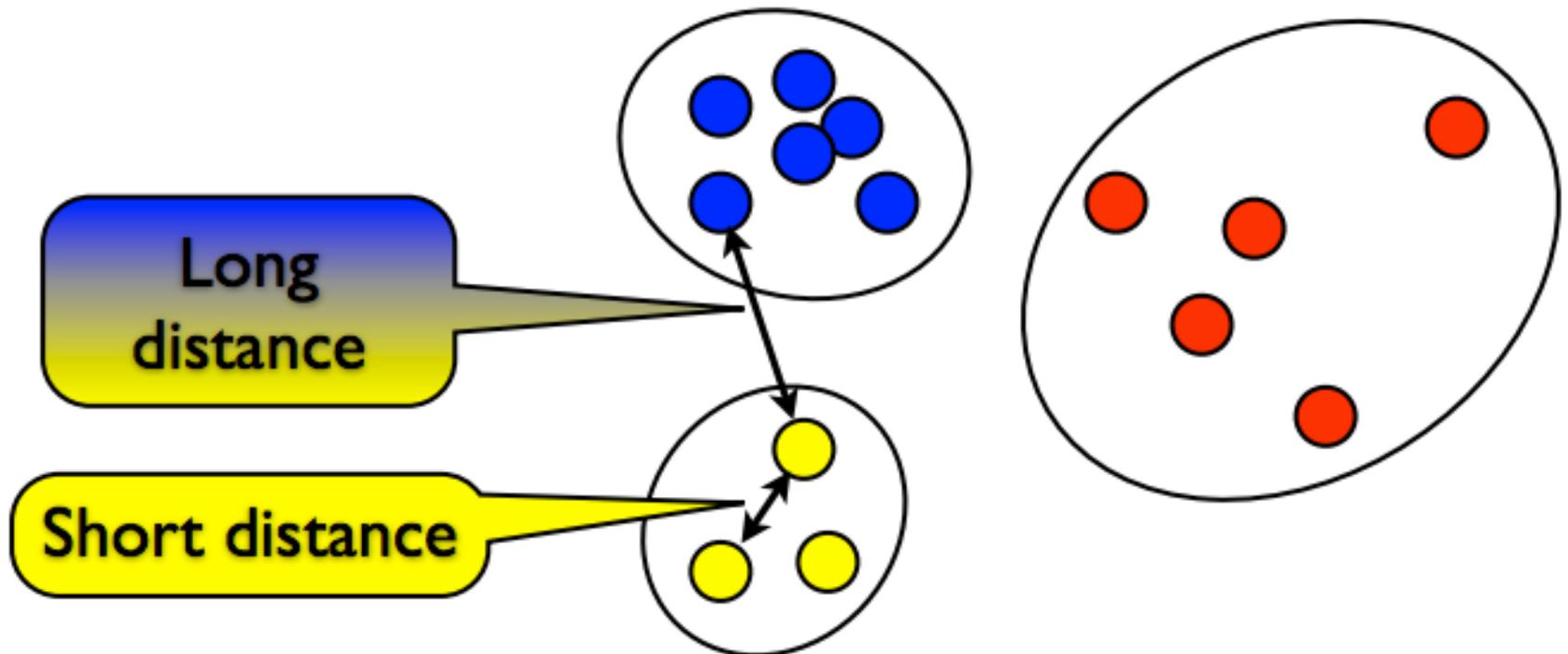
Ronald Coase, Economist, Nobel Prize winner

What is cluster analysis?

Clustering is **finding groups** of objects such that:

similar (or related) to the objects in **the same group** and

different from (or unrelated) to the objects in **other groups**



Properties

- Classes/labels for each instance are derived only from the data
- For that reason, cluster analysis is referred to as **unsupervised classification**

Why to cluster biological data?

- **Intuition building**

Finding hidden internal structure of the high-dimensional data

- Hypothesis generation

Finding and characterizing similar groups of objects in the data

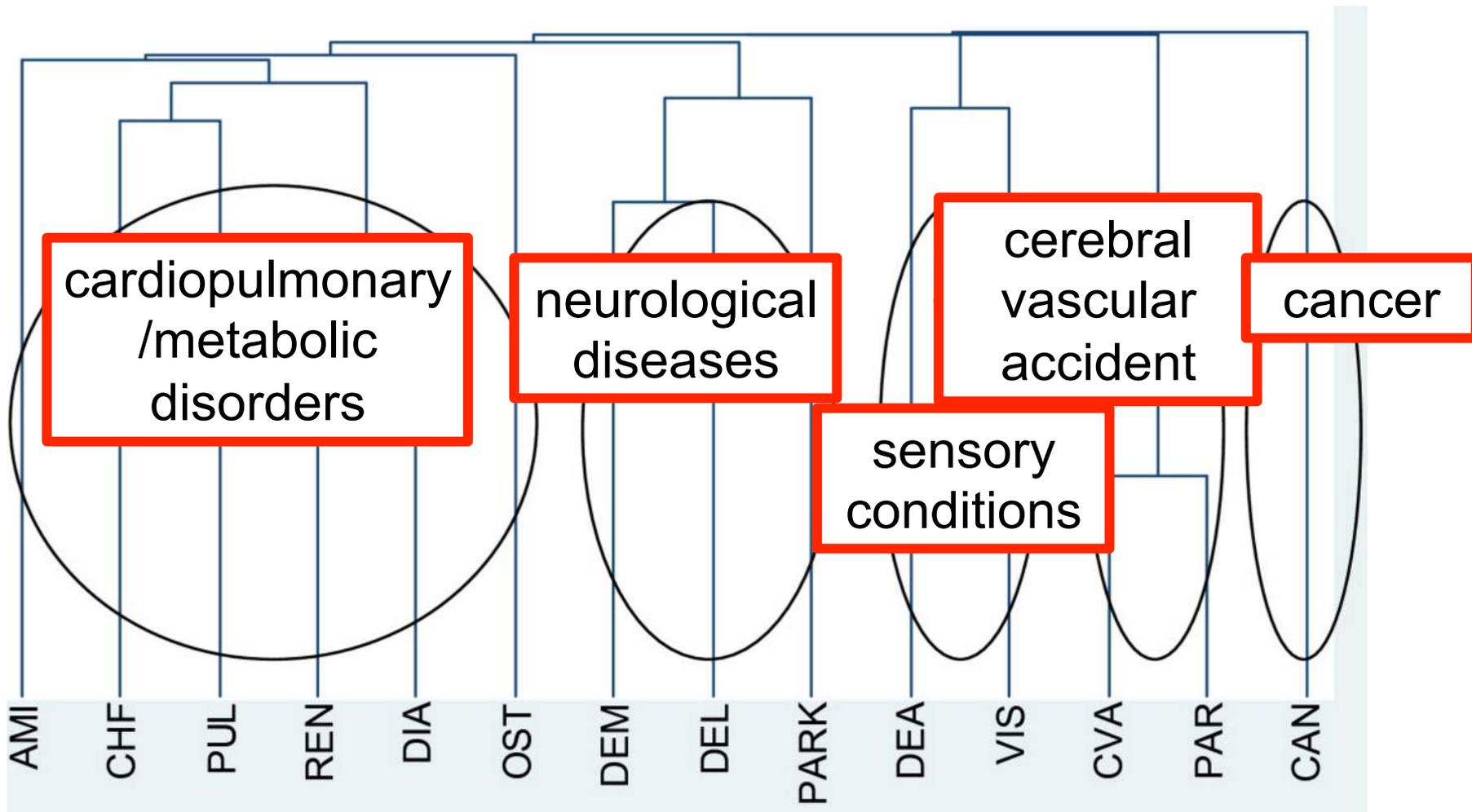
- Knowledge discovery in data

Ex. Underlying rules, reoccurring patterns, topics, etc.

- Summarizing / compressing large data

- Data visualization

Intuition building



Why to cluster biological data?

- **Intuition building**

Finding hidden internal structure of the high-dimensional data

- **Hypothesis generation**

Finding and characterizing similar groups of objects in the data

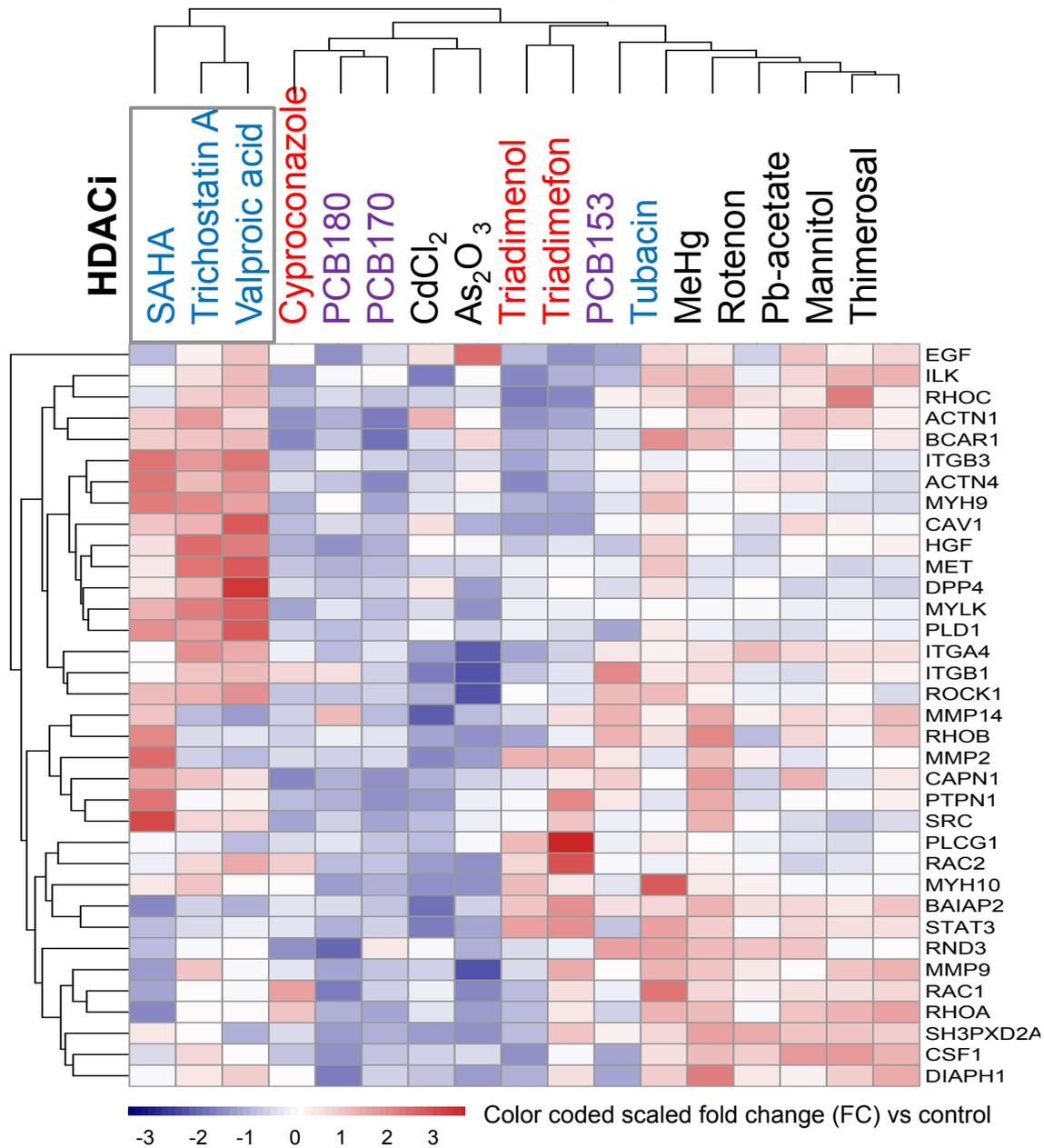
- Knowledge discovery in data

Ex. Underlying rules, reoccurring patterns, topics, etc.

- Summarizing / compressing large data

- Data visualization

Hypothesis generation



Why to cluster biological data?

- **Intuition building**

Finding hidden internal structure of the high-dimensional data

- **Hypothesis generation**

Finding and characterizing similar groups of objects in the data

- **Knowledge discovery in data**

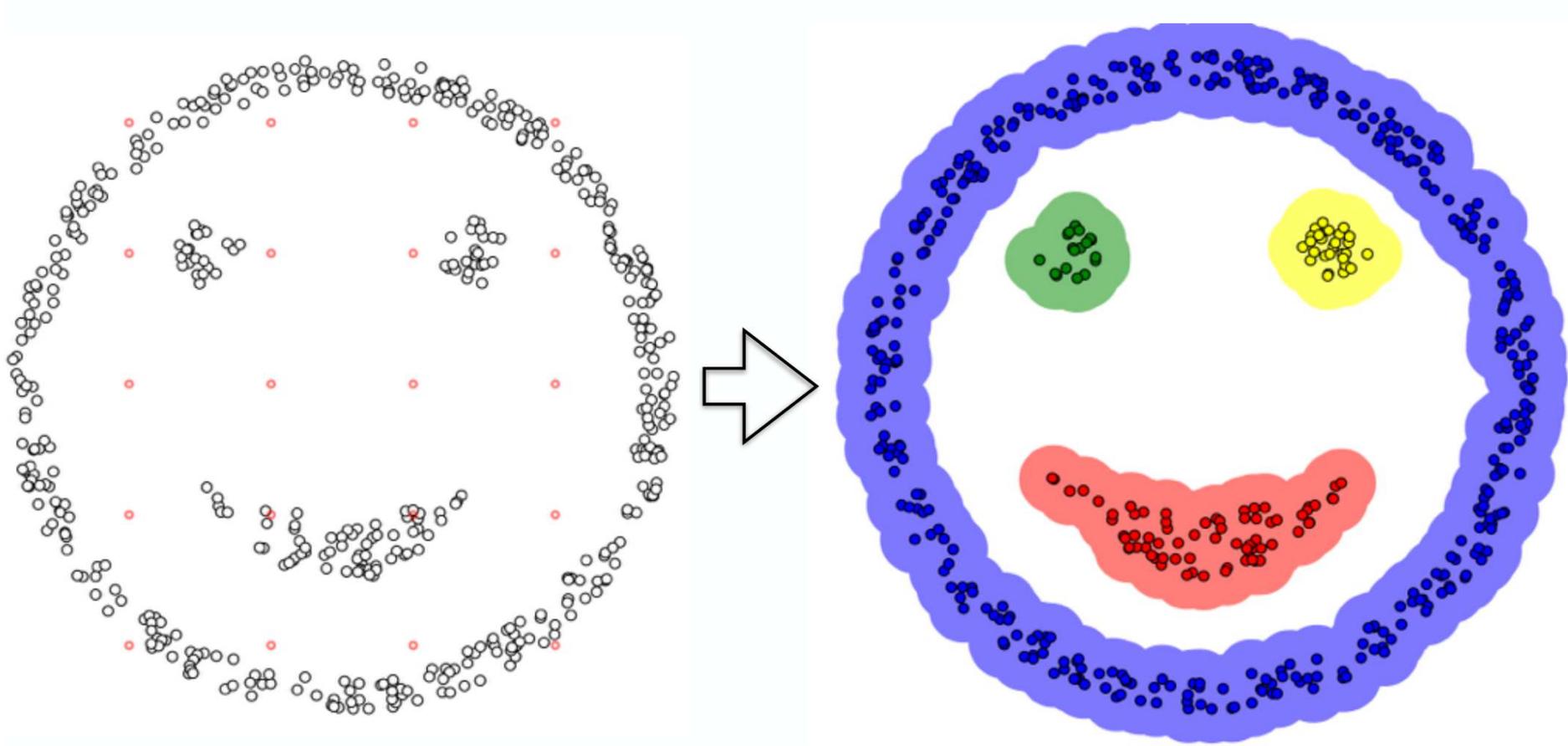
Ex. Underlying rules, reoccurring patterns, topics, etc.

- Summarizing / compressing large data

- Data visualization

Knowledge discovery in data

Ex. Underlying rules, reoccurring patterns, topics, etc.



Why to cluster biological data?

- **Intuition building**

Finding hidden internal structure of the high-dimensional data

- **Hypothesis generation**

Finding and characterizing similar groups of objects in the data

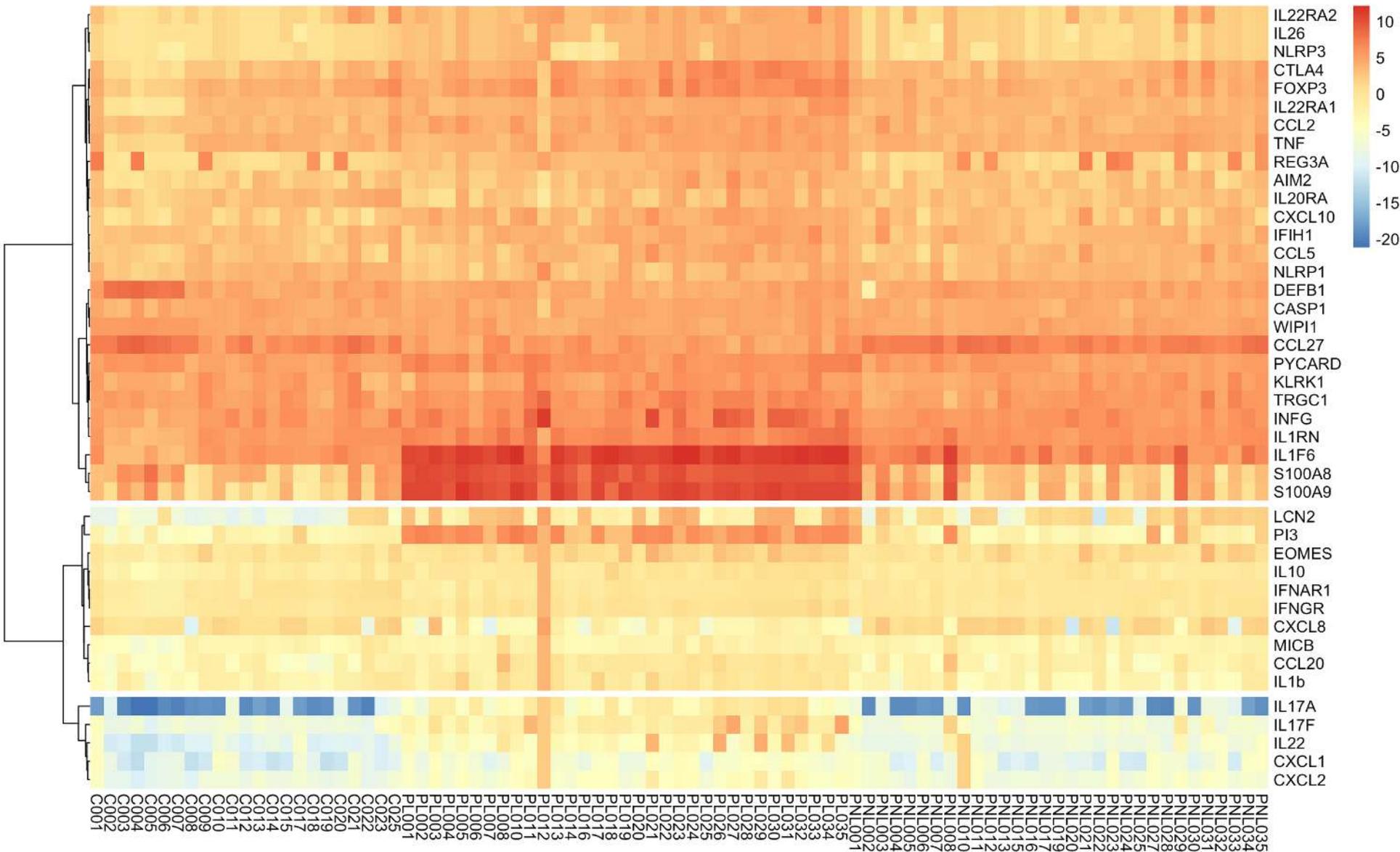
- **Knowledge discovery in data**

Ex. Underlying rules, reoccurring patterns, topics, etc.

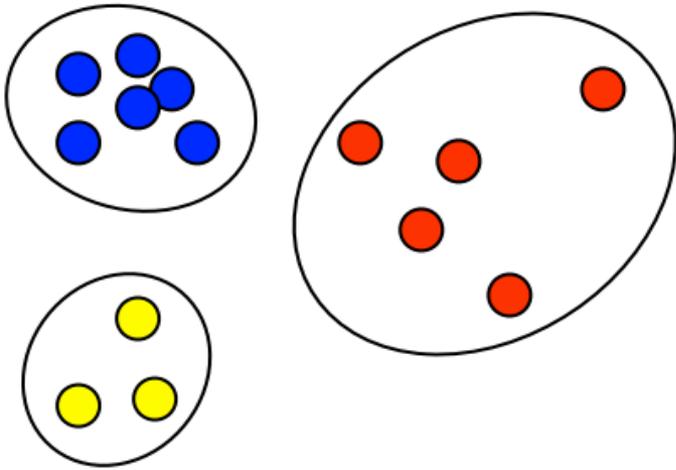
- **Summarizing / compressing large data**

- **Data visualization**

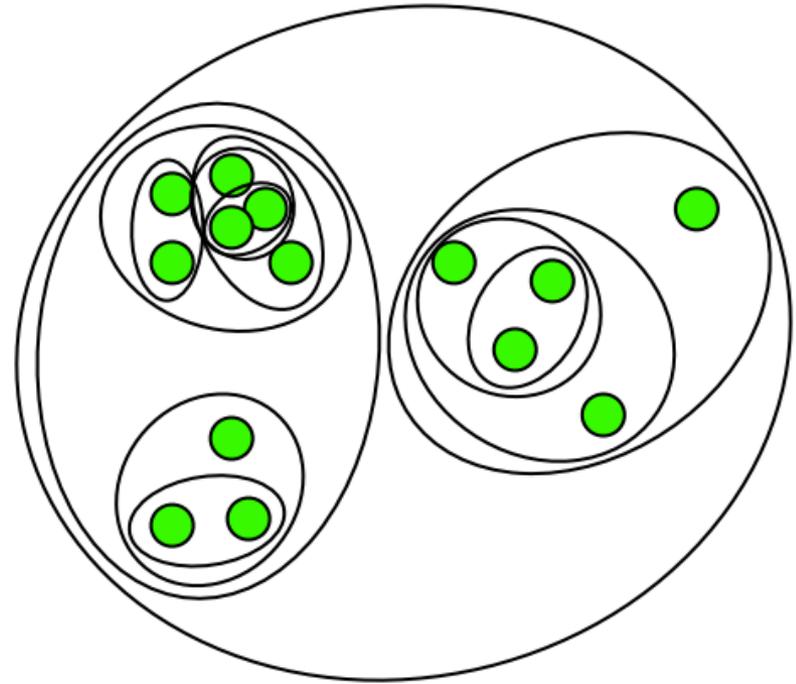
Summarizing/compressing the data



Partitional vs Hierarchical

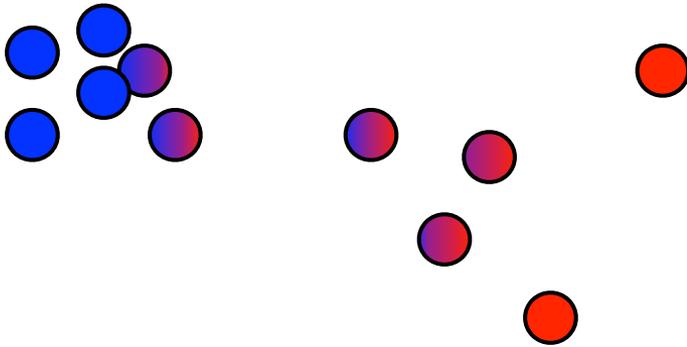


Each sample(point) is assigned to a unique cluster

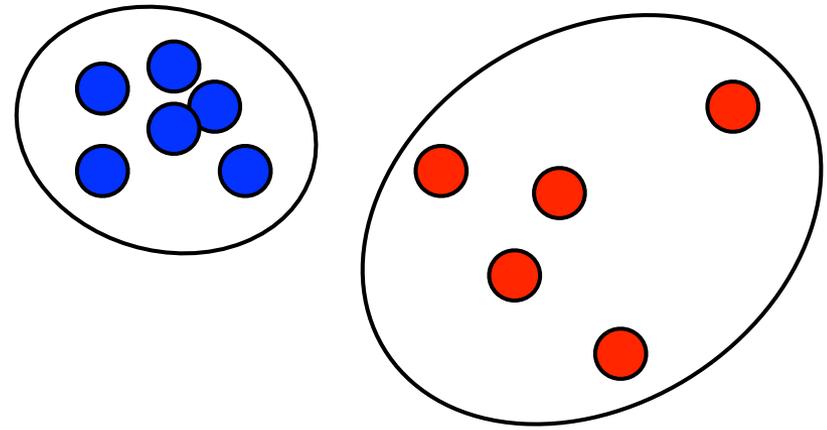


Creates a nested and hierarchical set of partitions/clusters

Fuzzy vs Non-Fuzzy

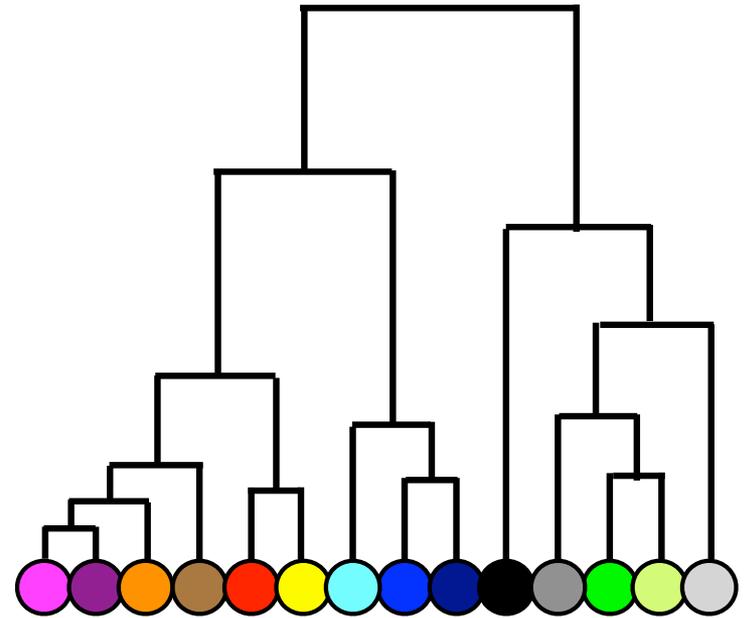
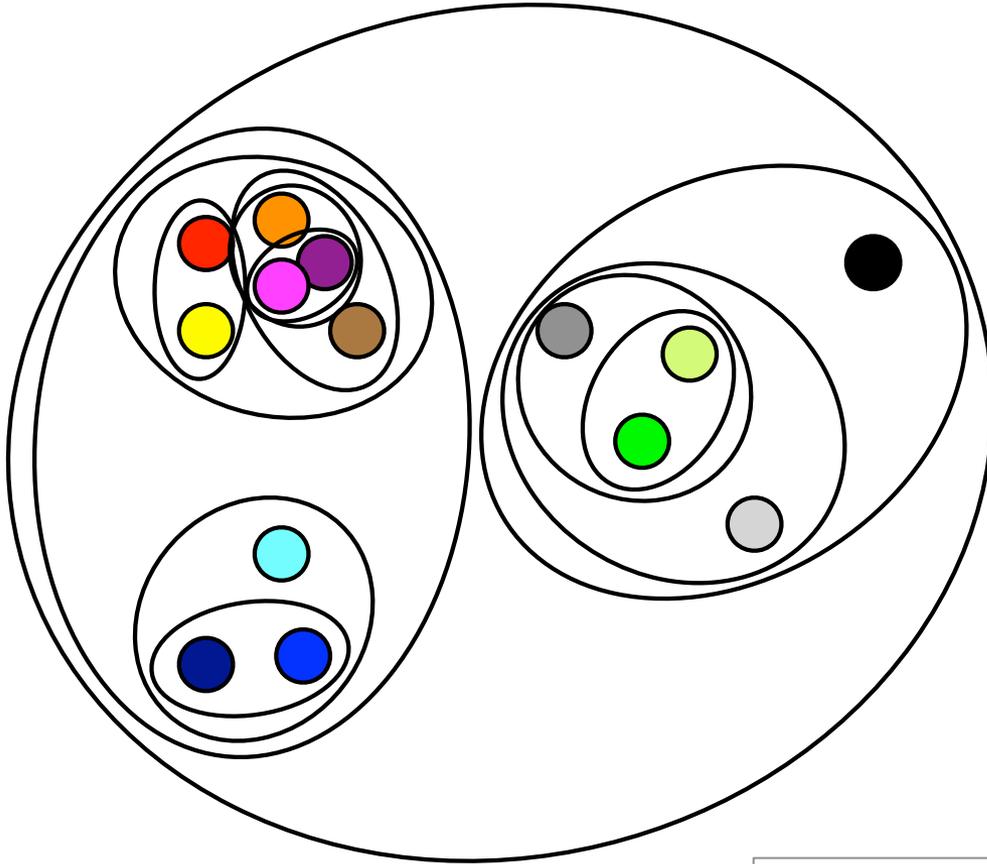


Each object belongs to each cluster with some weight



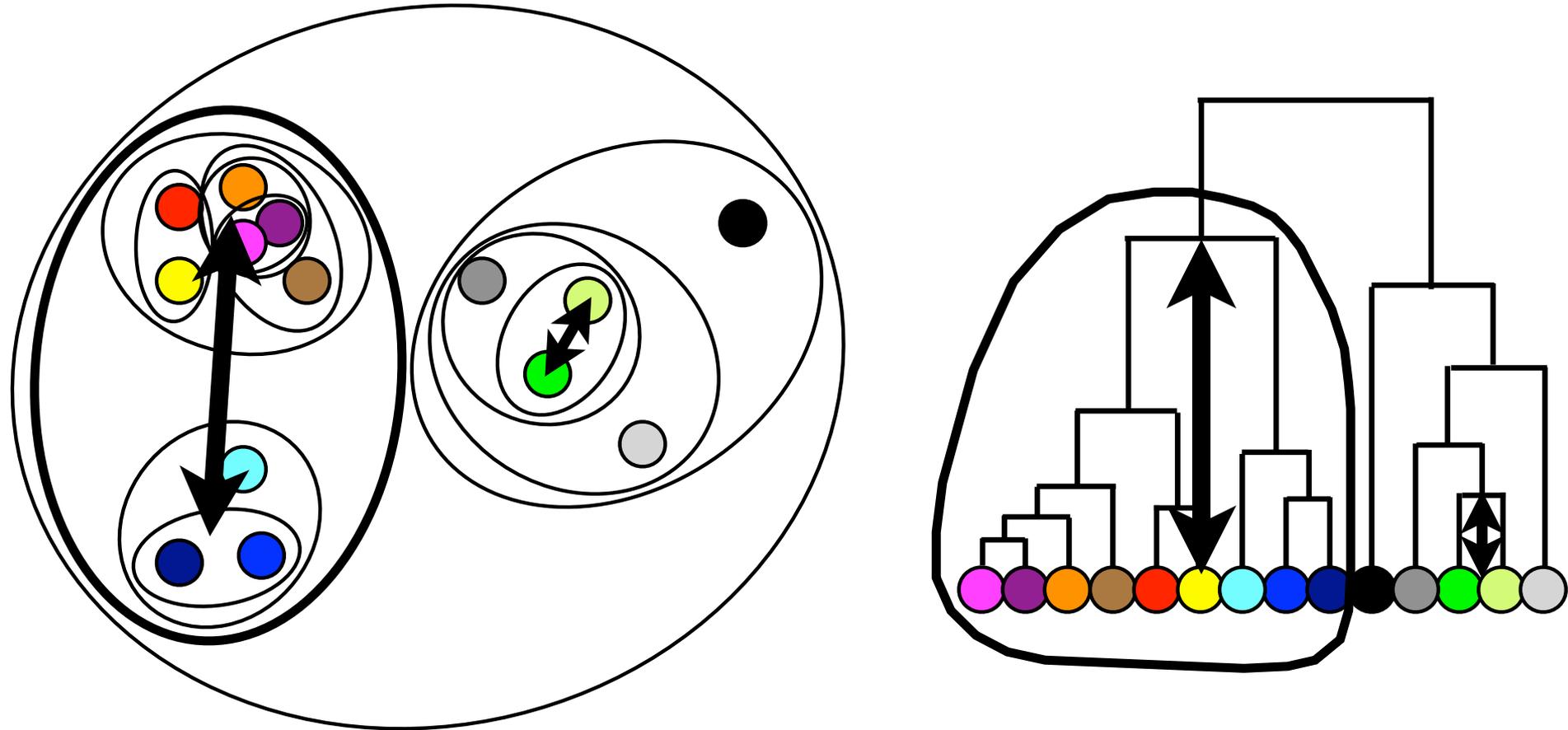
Each object belongs to exactly one cluster

Hierarchical clustering



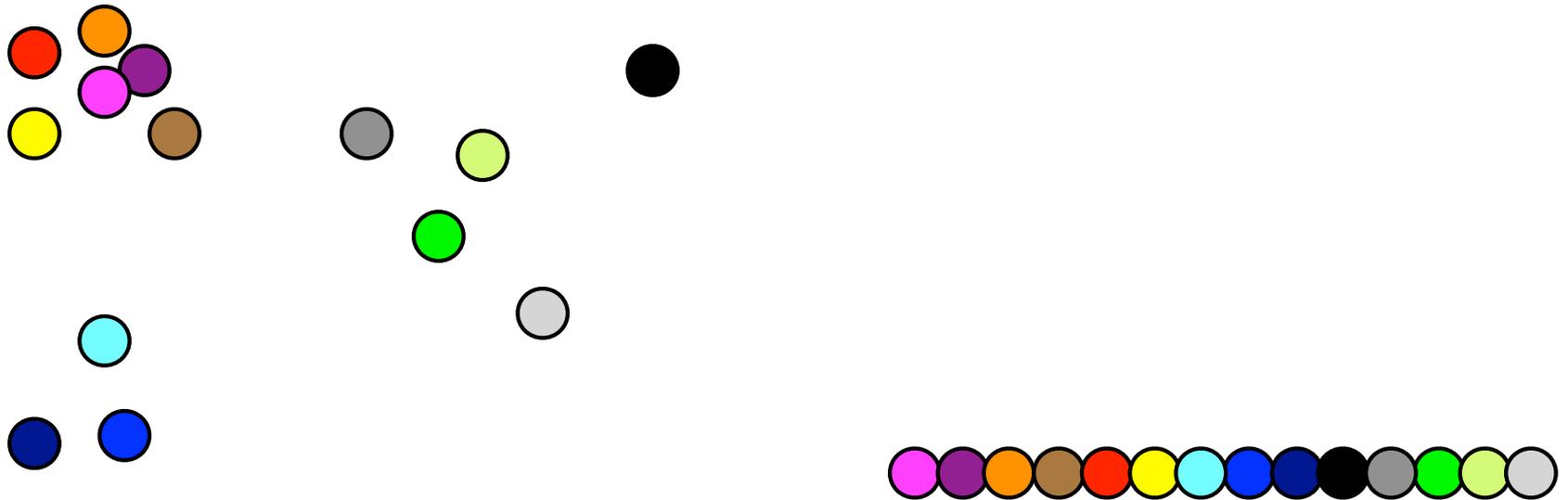
Hierarchical clustering is usually depicted as a dendrogram (tree)

Hierarchical clustering



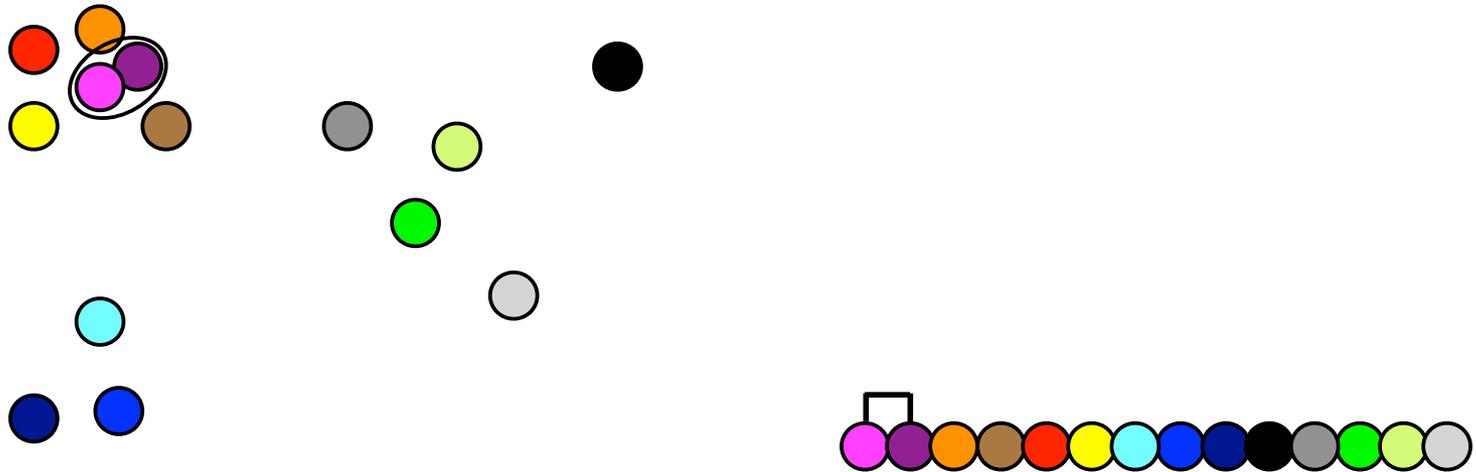
- Each subtree corresponds to a cluster
- Height of branching shows distance

Hierarchical clustering



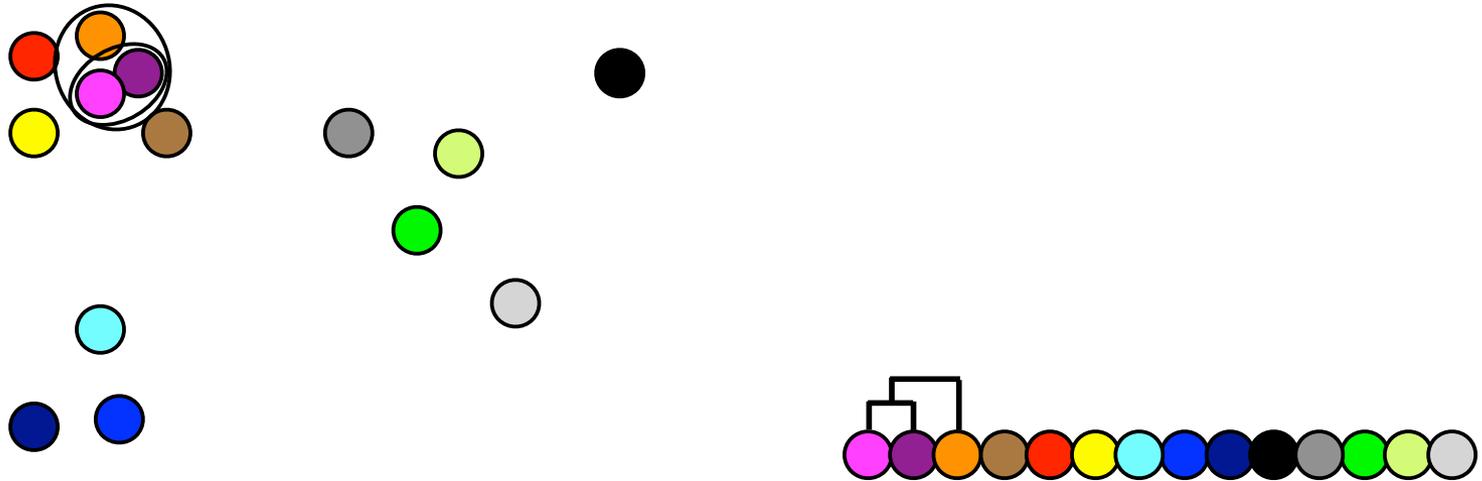
Algorithm for Agglomerative Hierarchical Clustering:
Join the two closest objects

Hierarchical clustering



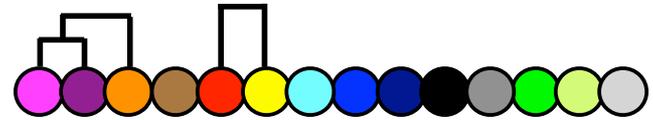
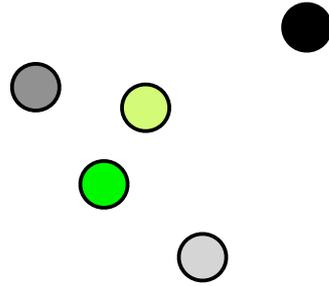
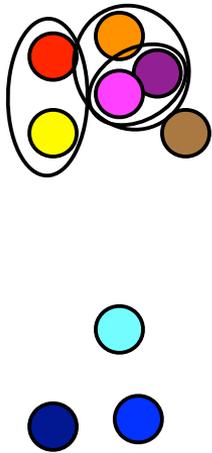
Join the two closest objects

Hierarchical clustering



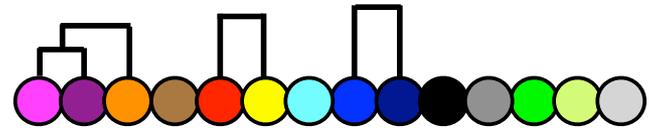
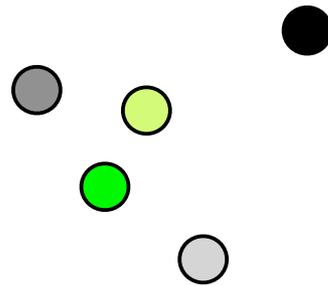
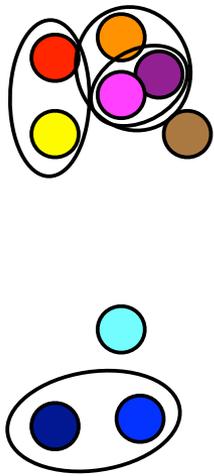
Keep joining the closest pairs

Hierarchical clustering



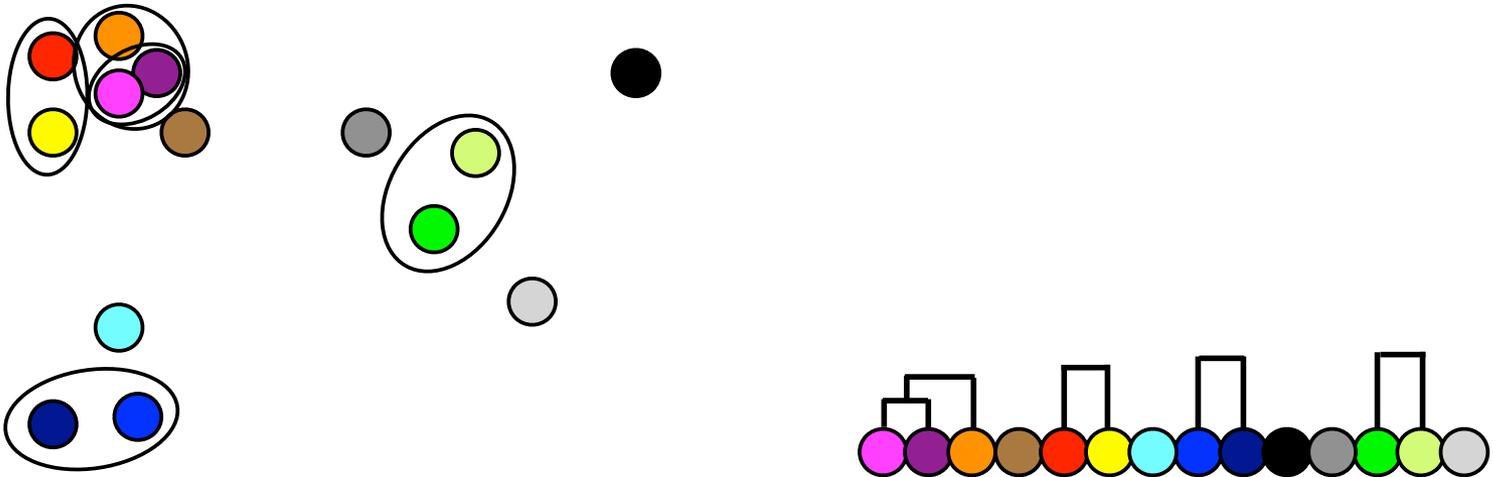
Keep joining the closest pairs

Hierarchical clustering



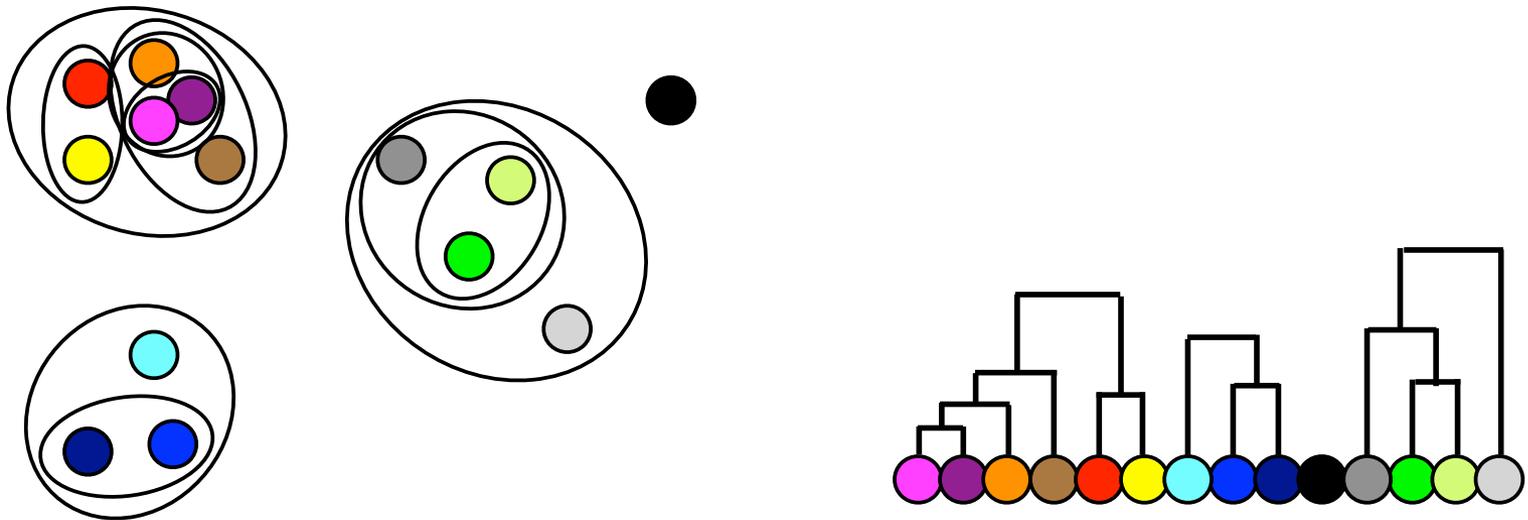
Keep joining the closest pairs

Hierarchical clustering



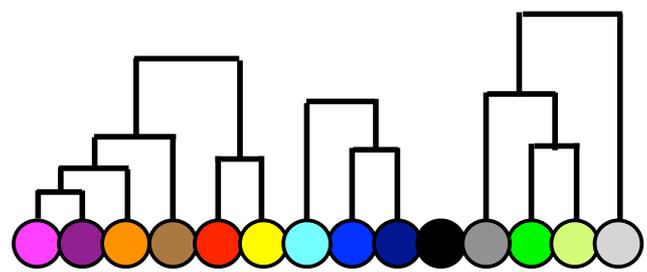
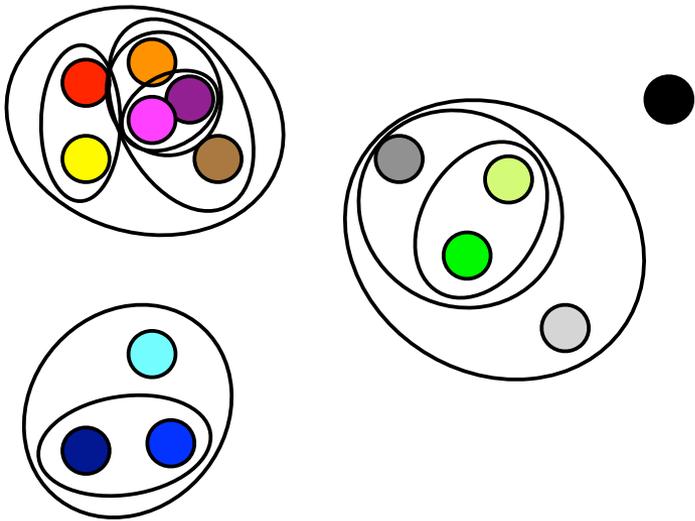
Keep joining the closest pairs

Hierarchical clustering



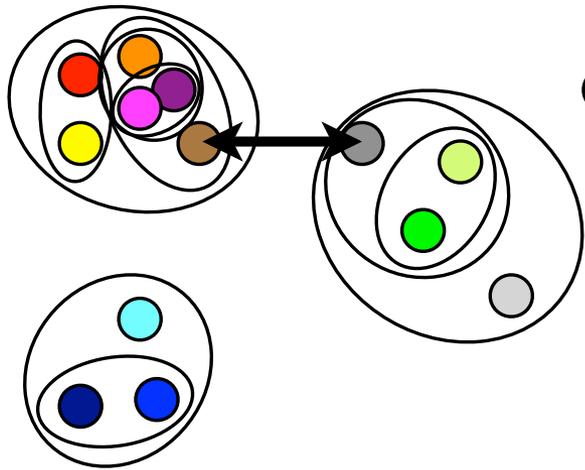
After 10 steps we have 4 clusters left

Q: Which clusters do we merge next?



Hierarchical clustering

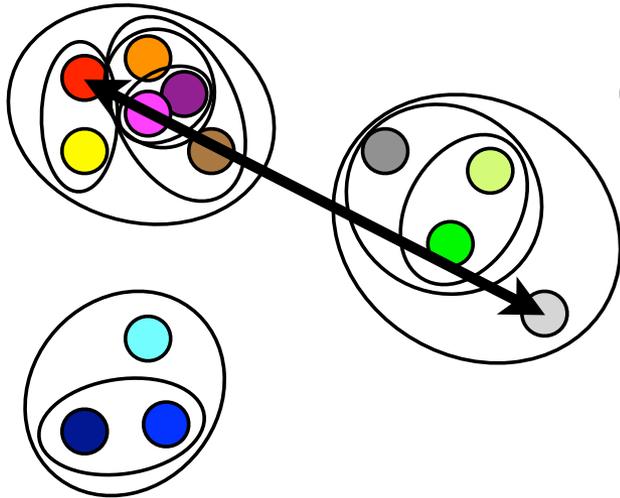
Several ways to measure distance between clusters:



- • Single linkage(MIN)

Hierarchical clustering

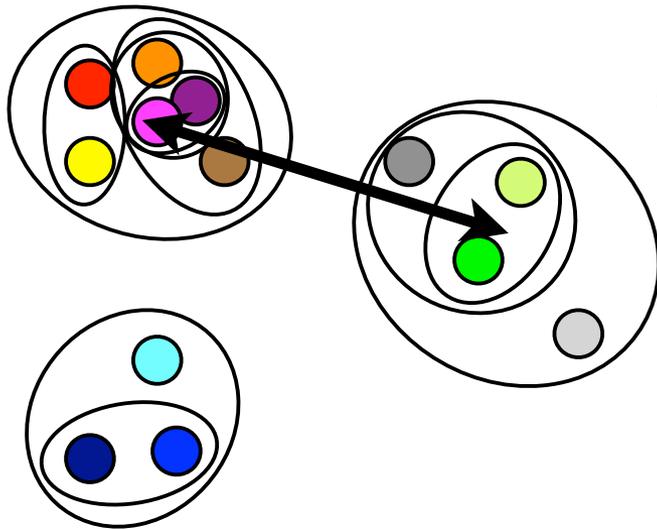
Several ways to measure distance between clusters:



- Single linkage(MIN)
- Complete linkage(MAX)

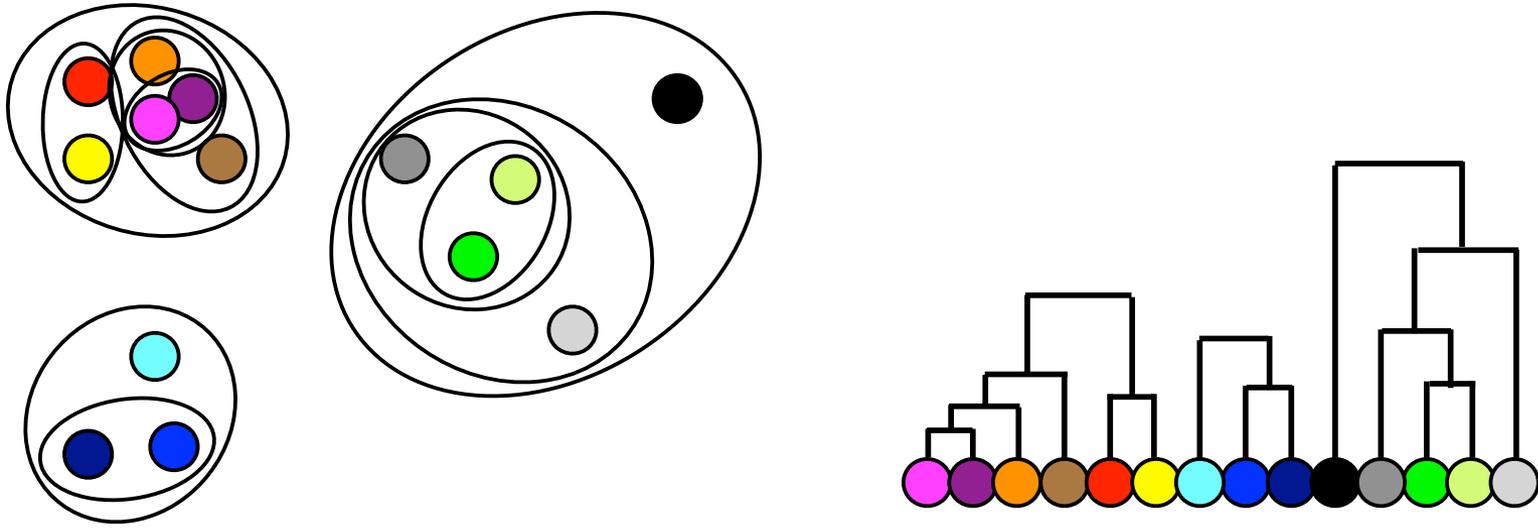
Hierarchical clustering

Several ways to measure distance between clusters:



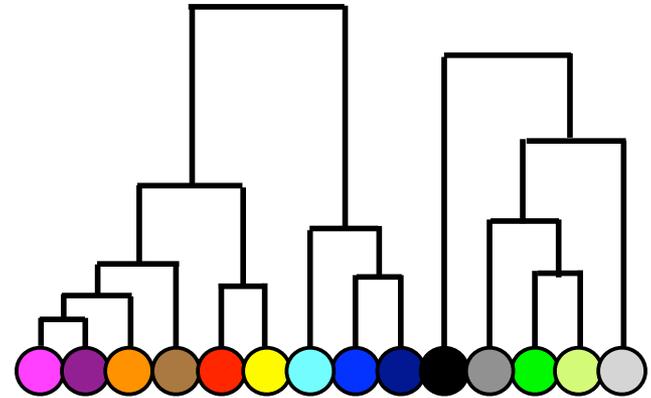
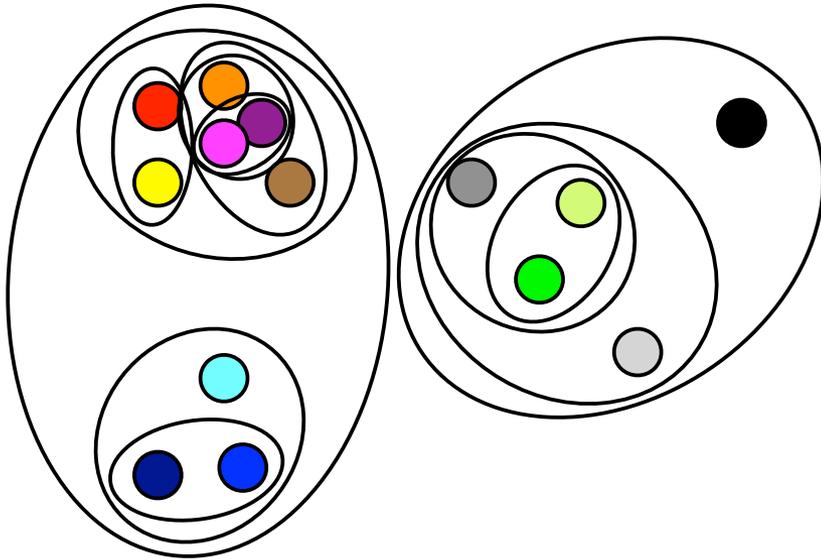
- Single linkage (MIN)
- Complete linkage (MAX)
- Average linkage
 - Weighted
 - Unweighted ...
- Ward's method

Hierarchical clustering



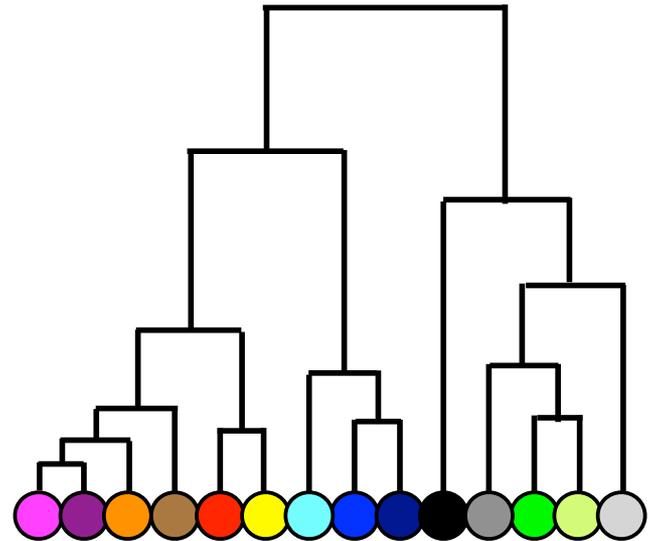
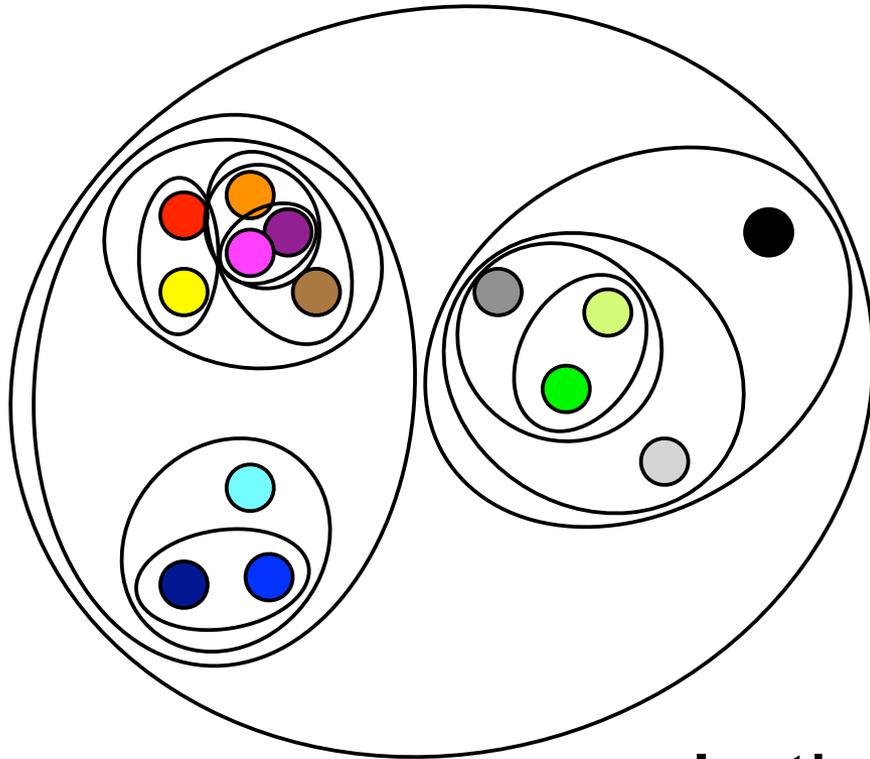
In this example and at this stage we have the same result as in partitional clustering

Hierarchical clustering



In the final step the two remaining clusters are joined into a single cluster

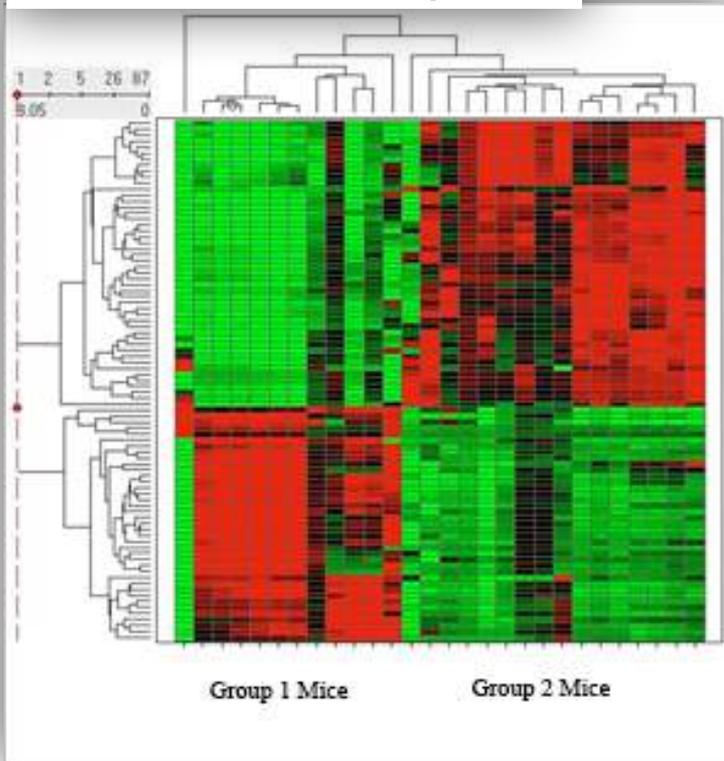
Hierarchical clustering



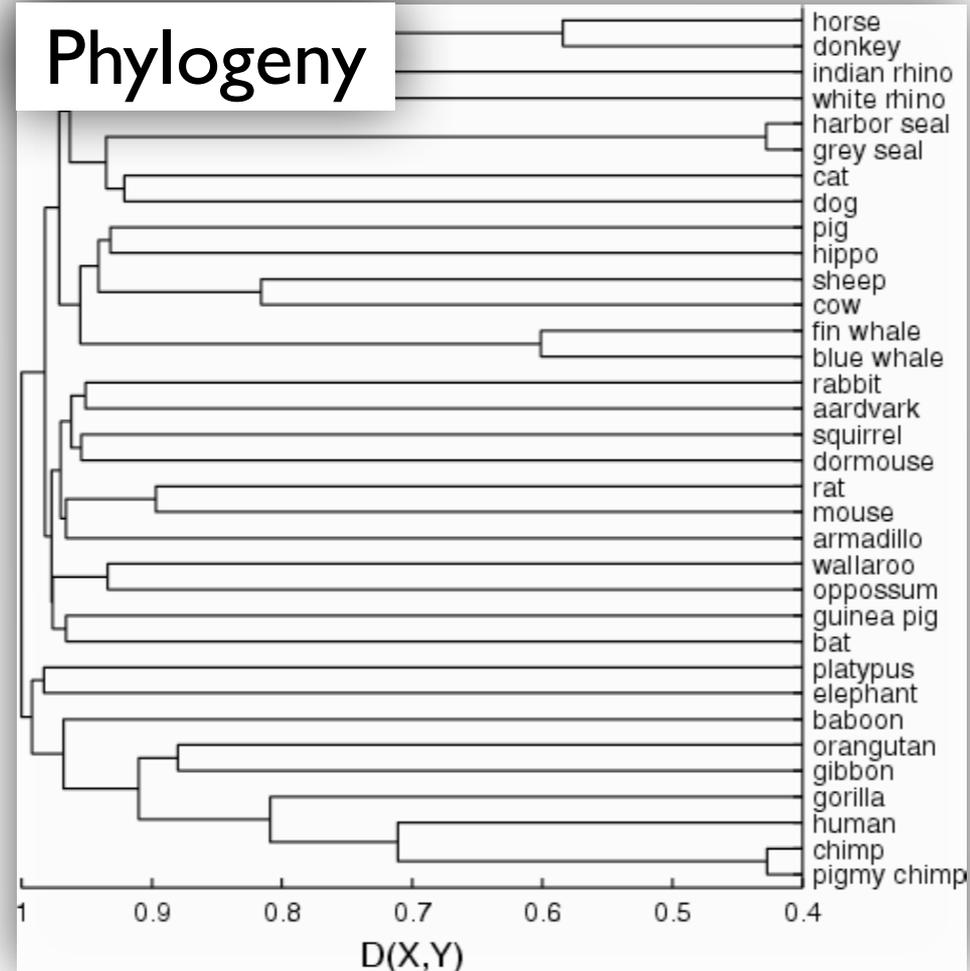
In the final step the two remaining clusters are joined into a single cluster

Examples of Hierarchical Clustering in Bioinformatics

Gene expression clustering



Phylogeny



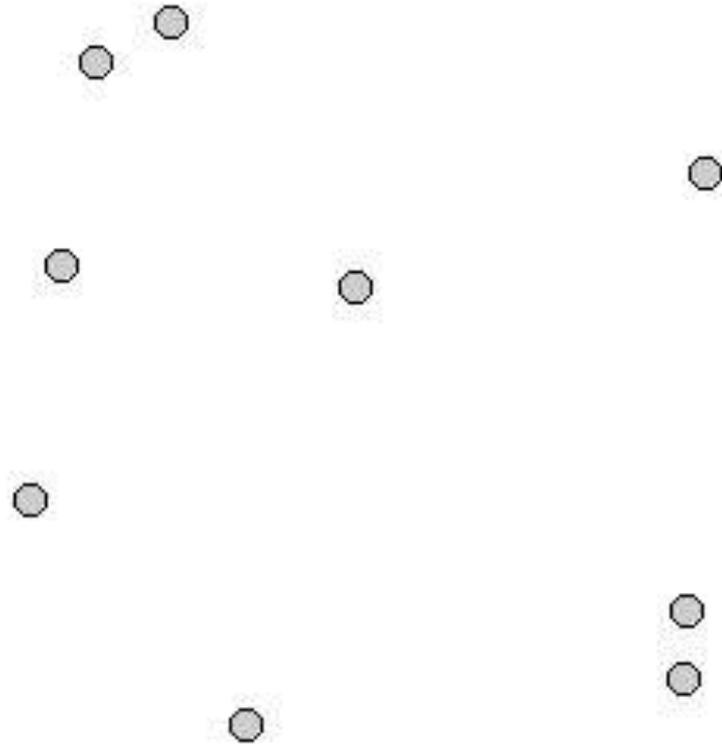
K-means clustering

- Partitional, non-fuzzy
- Partitions the data into K clusters
- K is given by the user

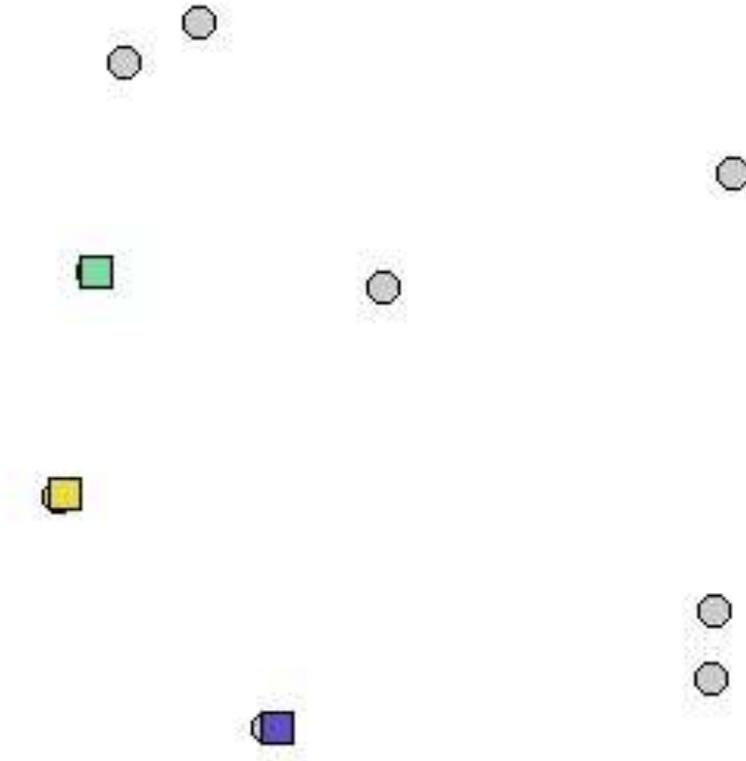
Algorithm:

- Choose K initial centers for the clusters
- Assign each object to its closest center
- Recalculate cluster centers
- Repeat until converges

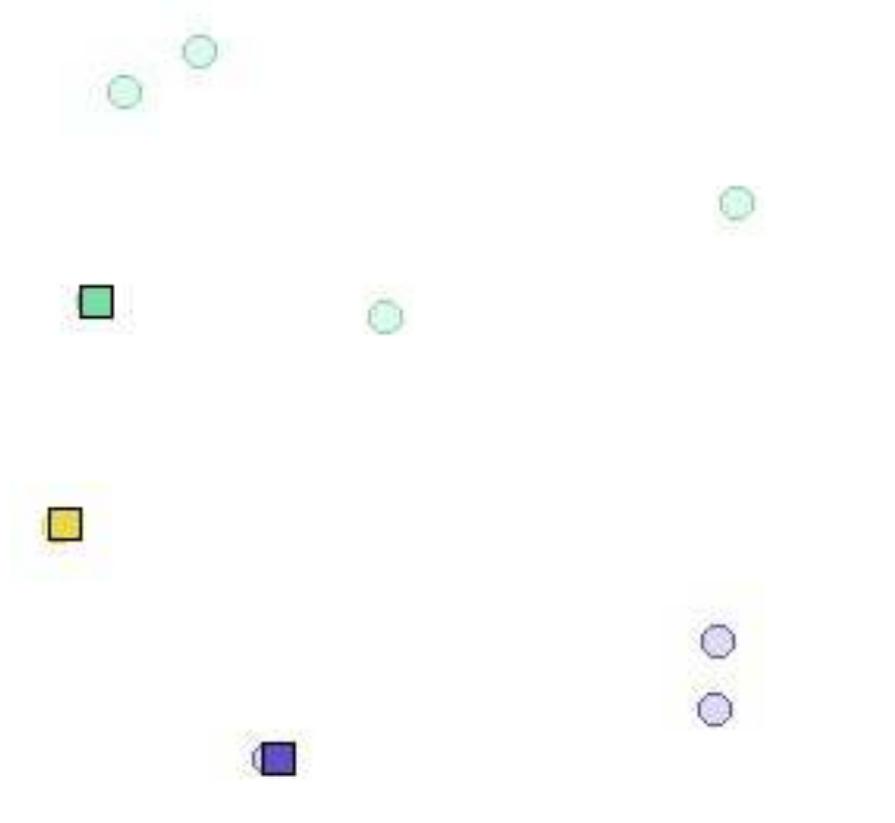
K-means (1)



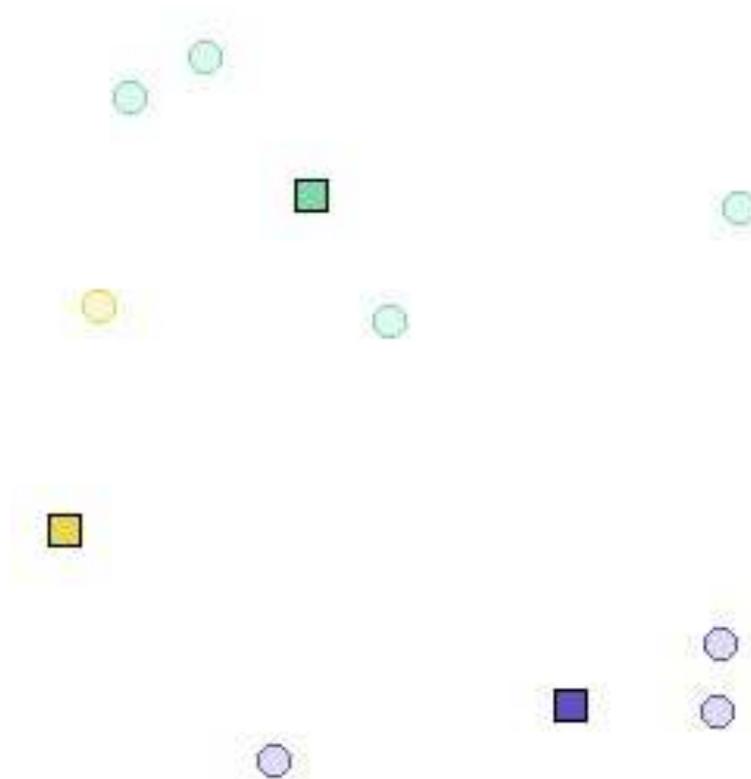
K-means (2)



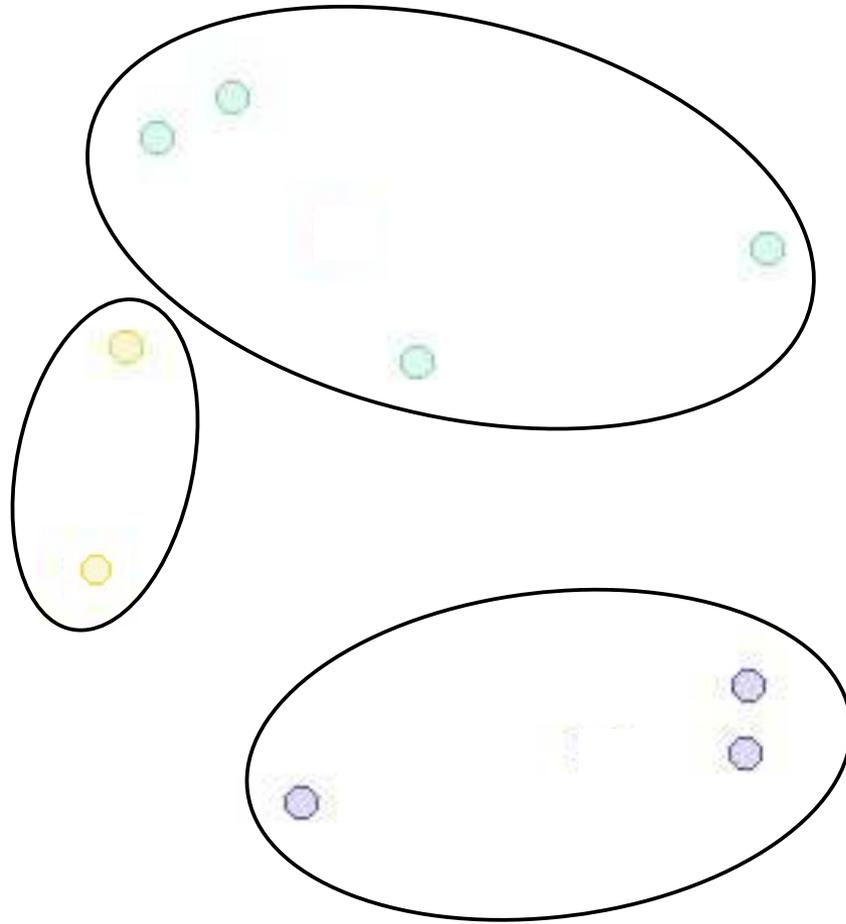
K-means (3)



K-means (4)

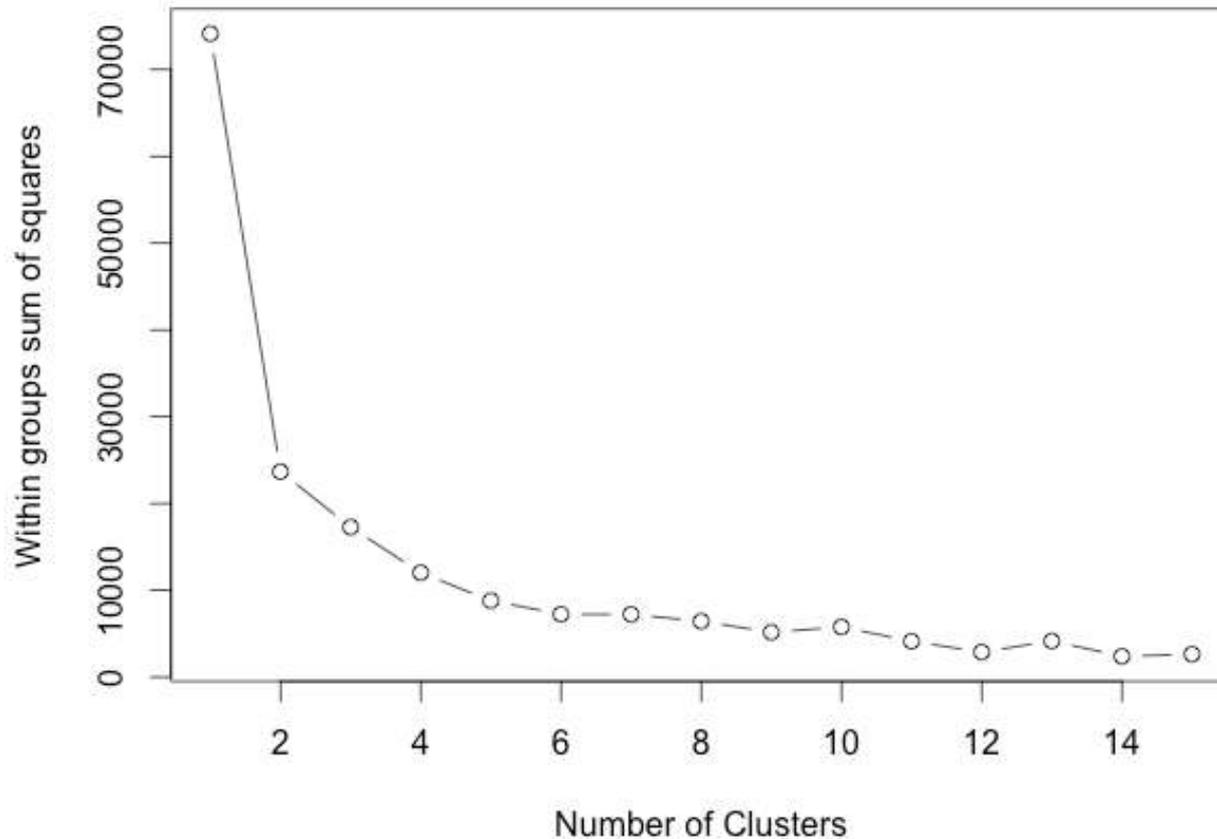


K-means (5)



Elbow method

Estimate the number of clusters



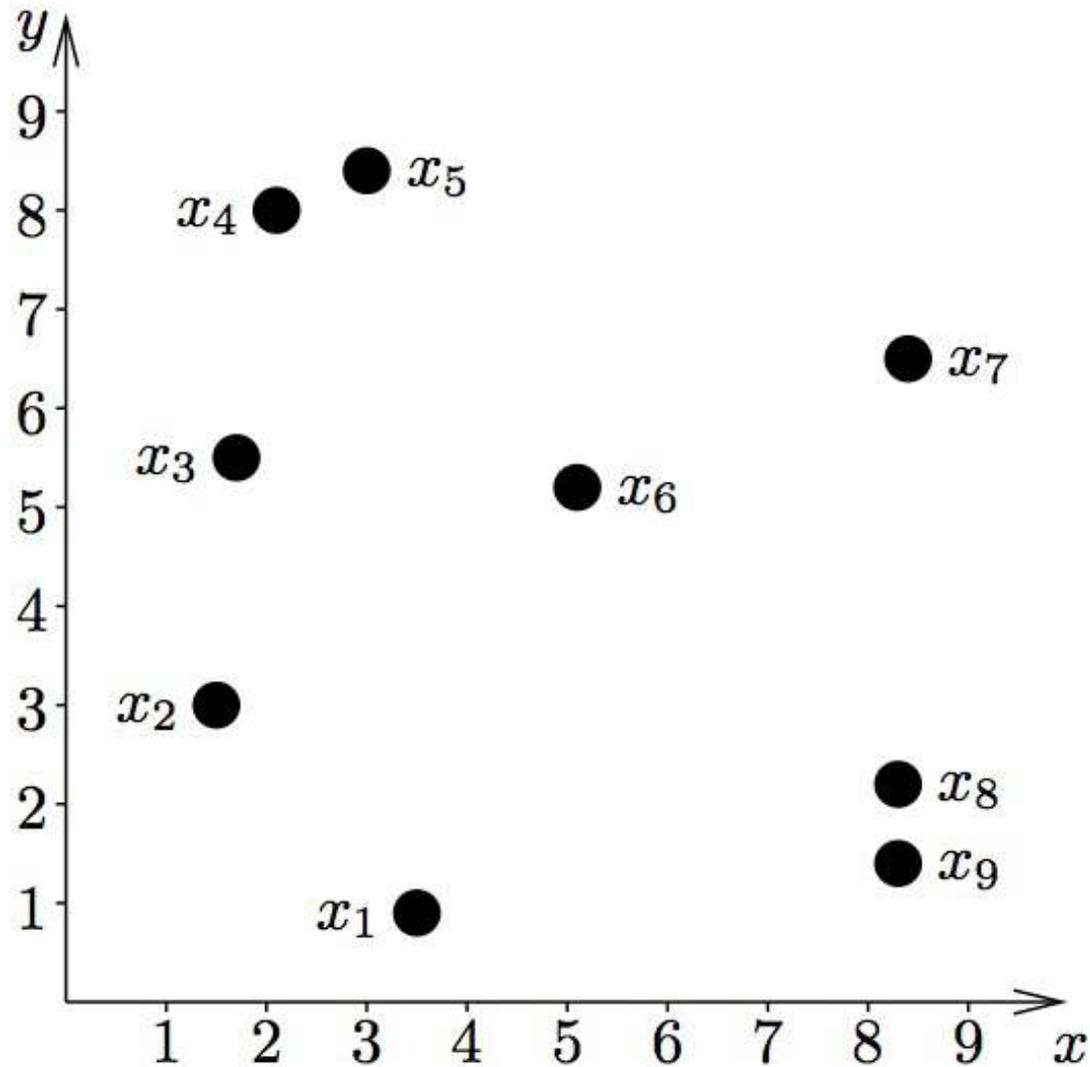
K-means clustering summary

- One of the fastest clustering algorithms
- Therefore very widely used
- Sensitive to the choice of initial centers
 - many algorithms to choose initial centers cleverly
- Assumes that the mean can be calculated
 - can be used on vector data
 - cannot be used on sequences
(what is the mean of A and T?)

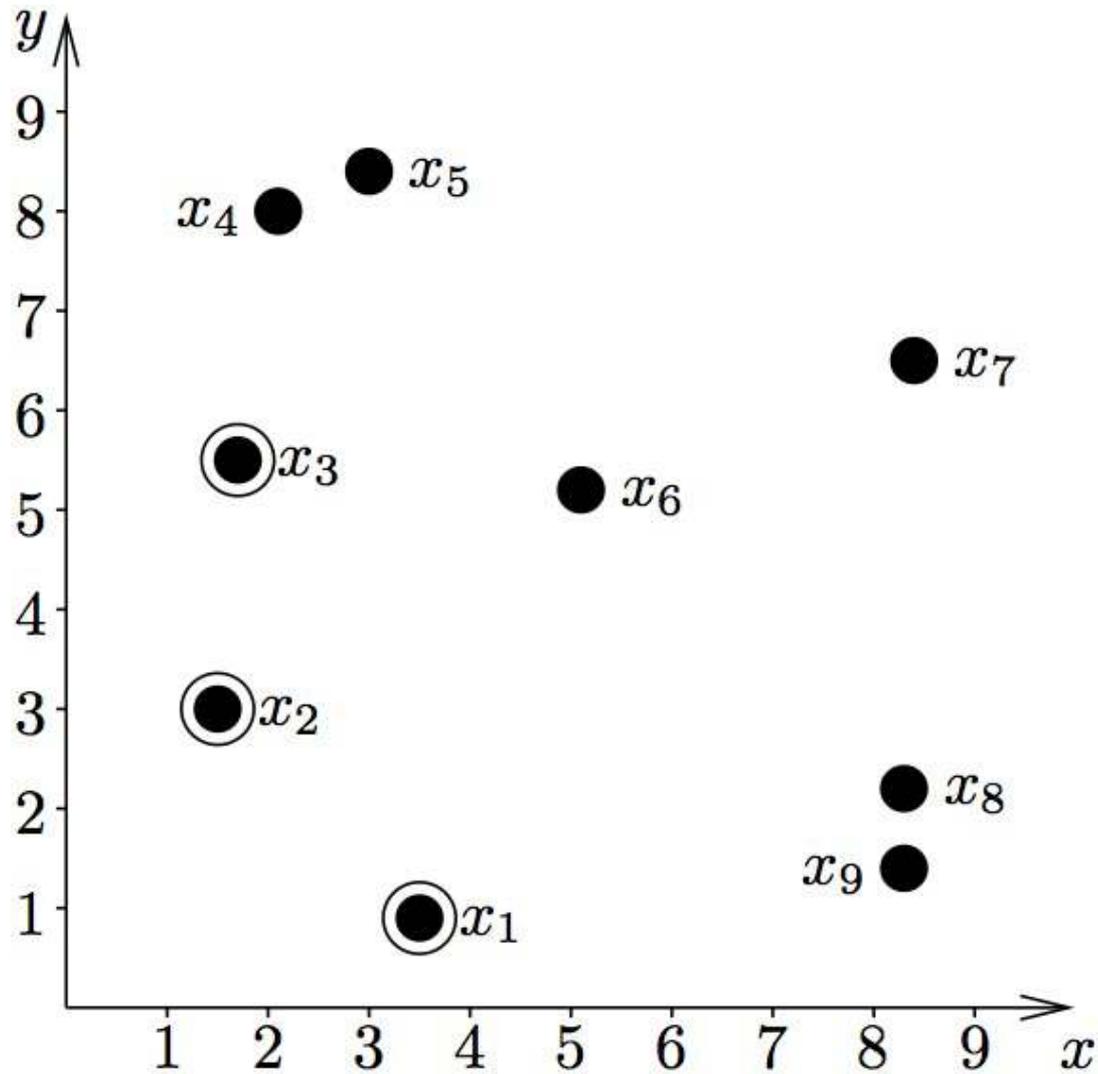
K-medoids clustering

- The same as K-means, except that the center is required to be at an object
- Medoid - an object which has minimal total distance to all other objects in its cluster
- Can be used on more complex data, with any distance measure
- Slower than K-means

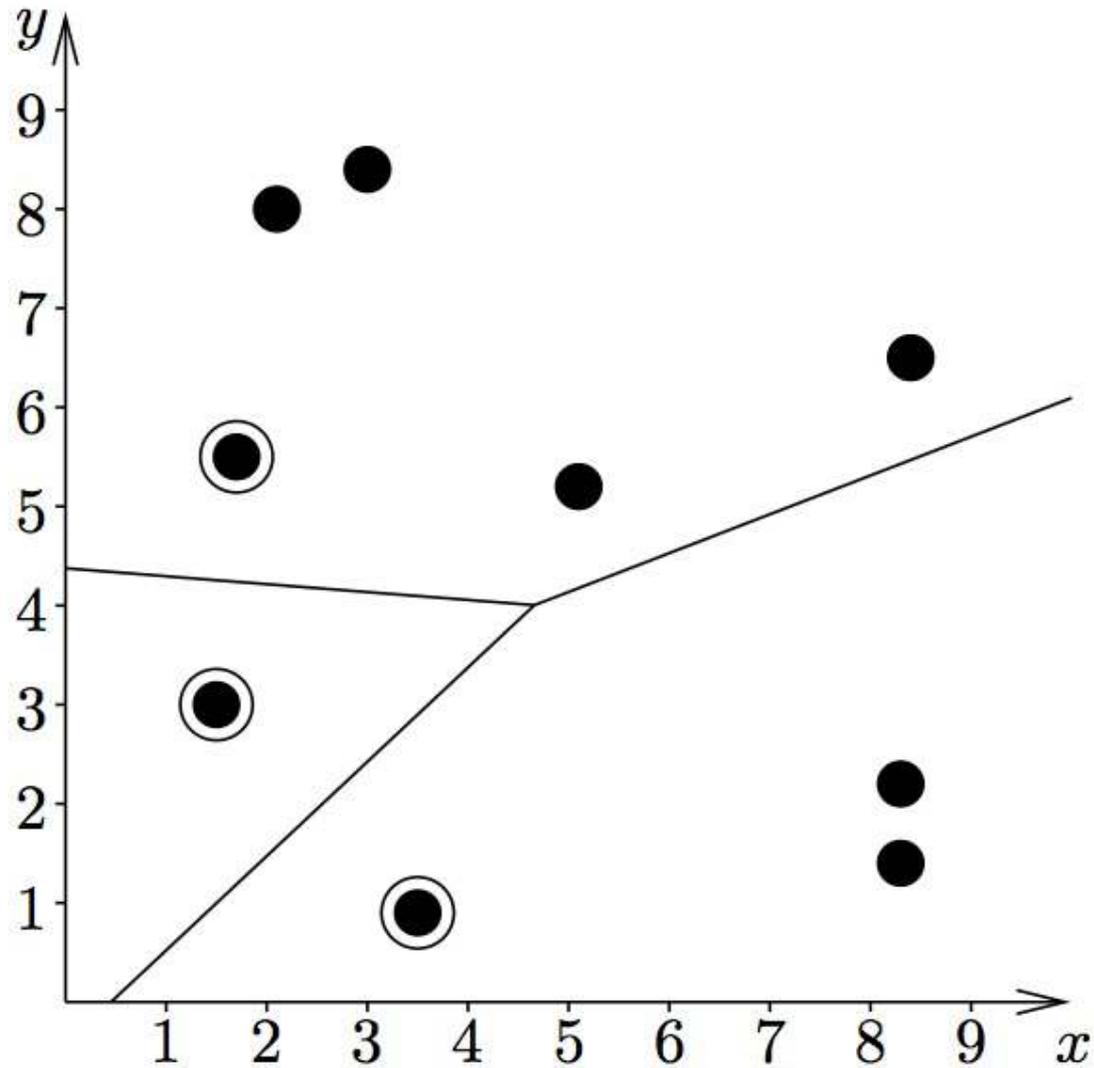
K-medoids (1)



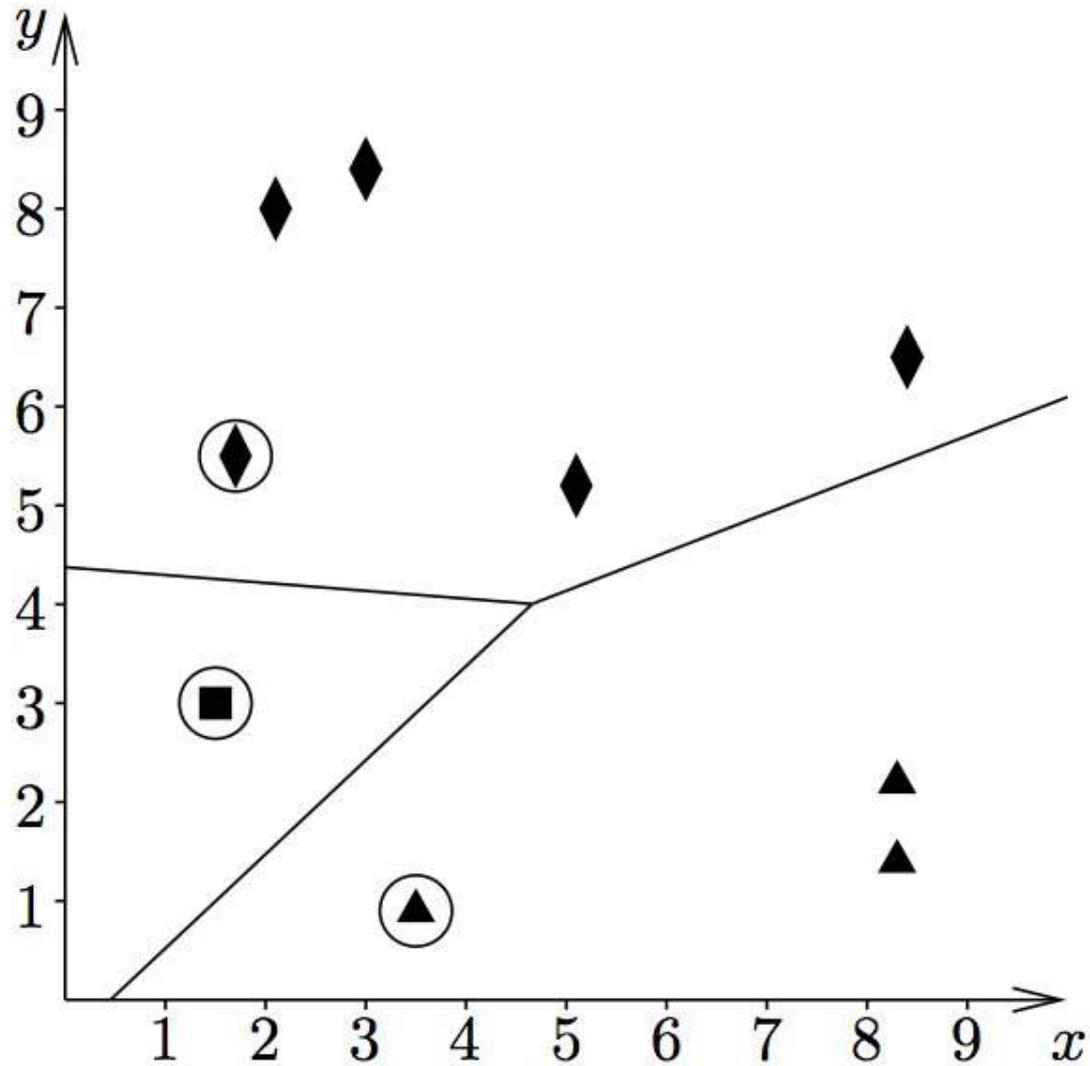
K-medoids (2)



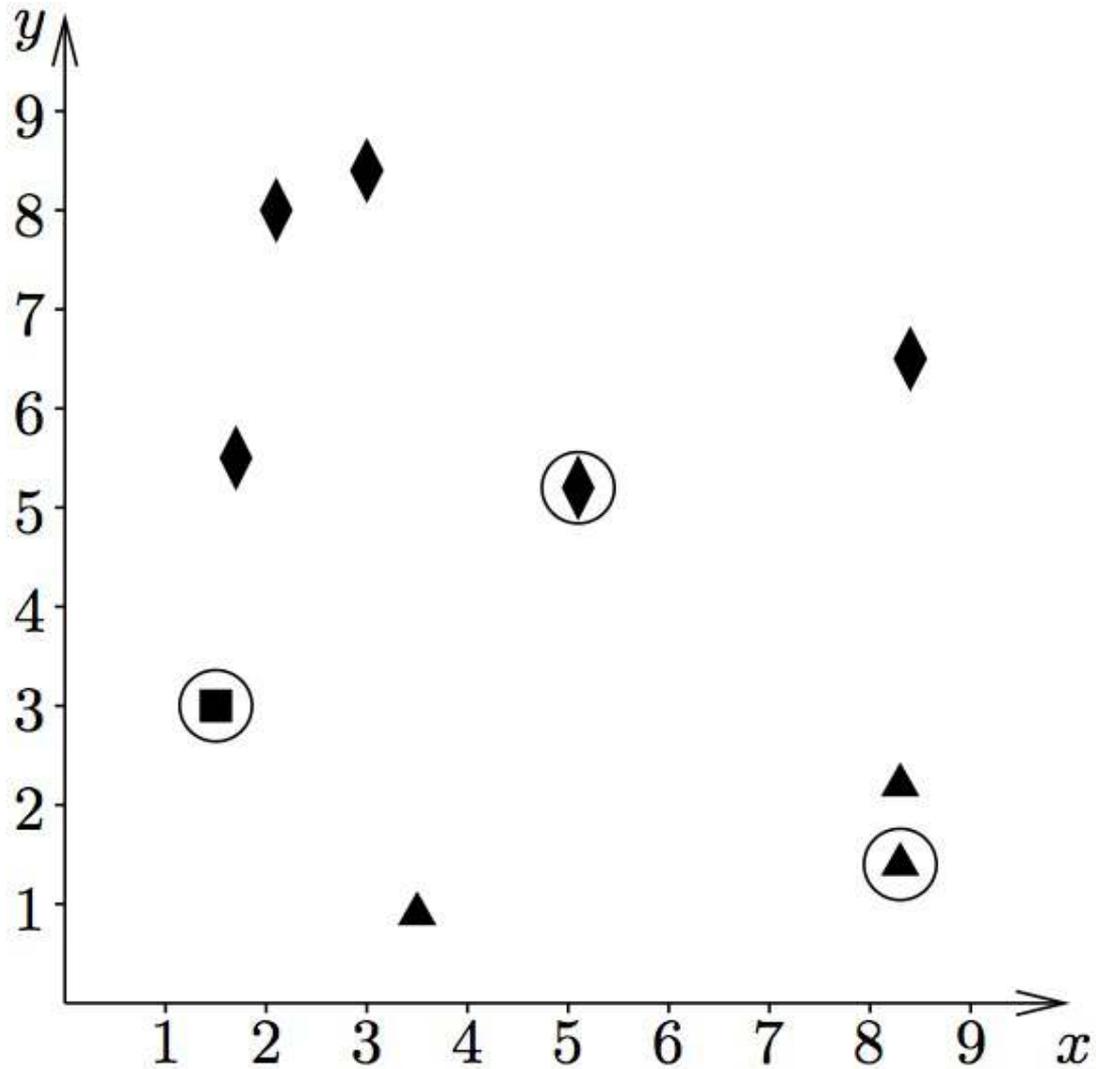
K-medoids (3)



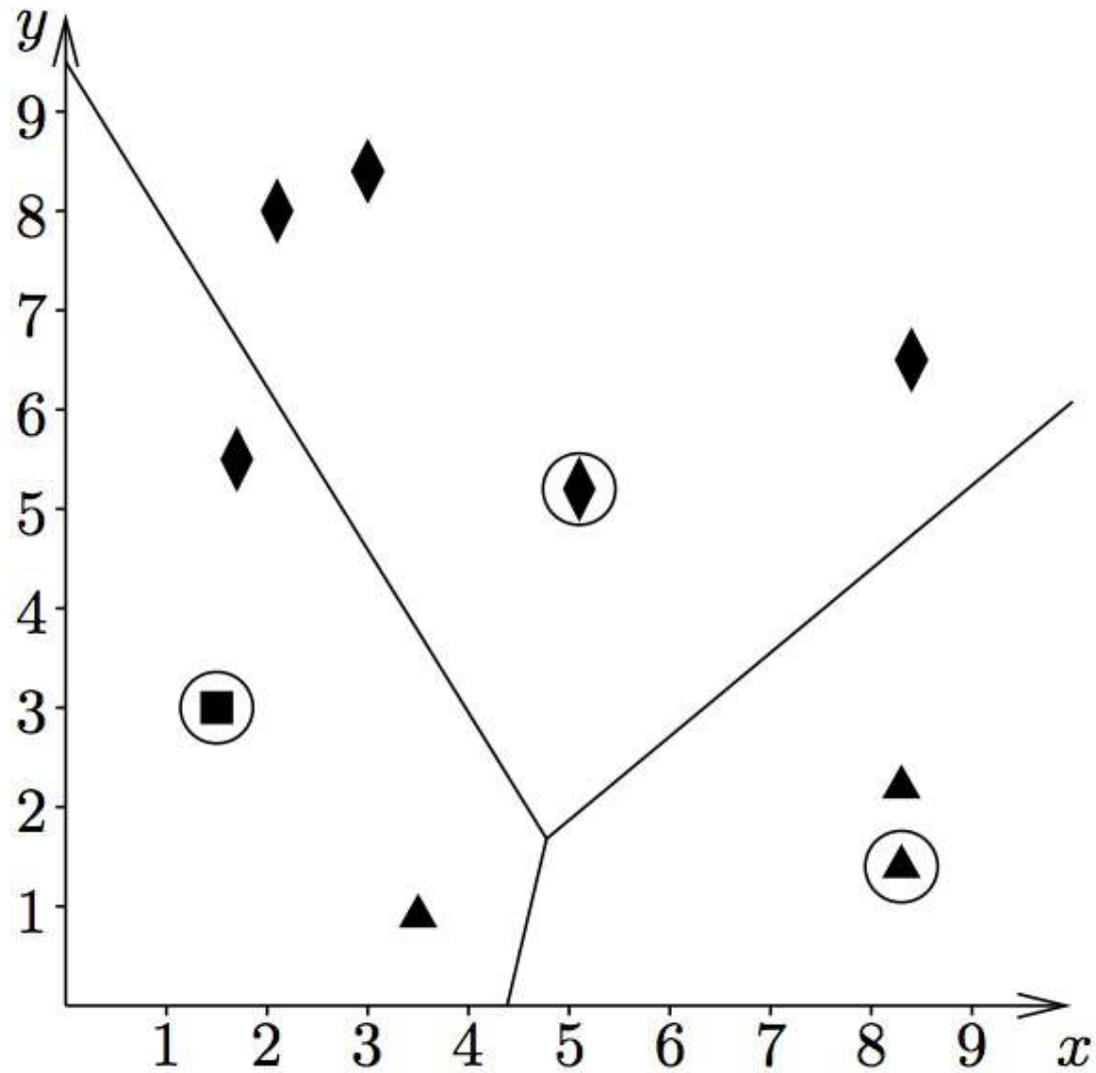
K-medoids (4)



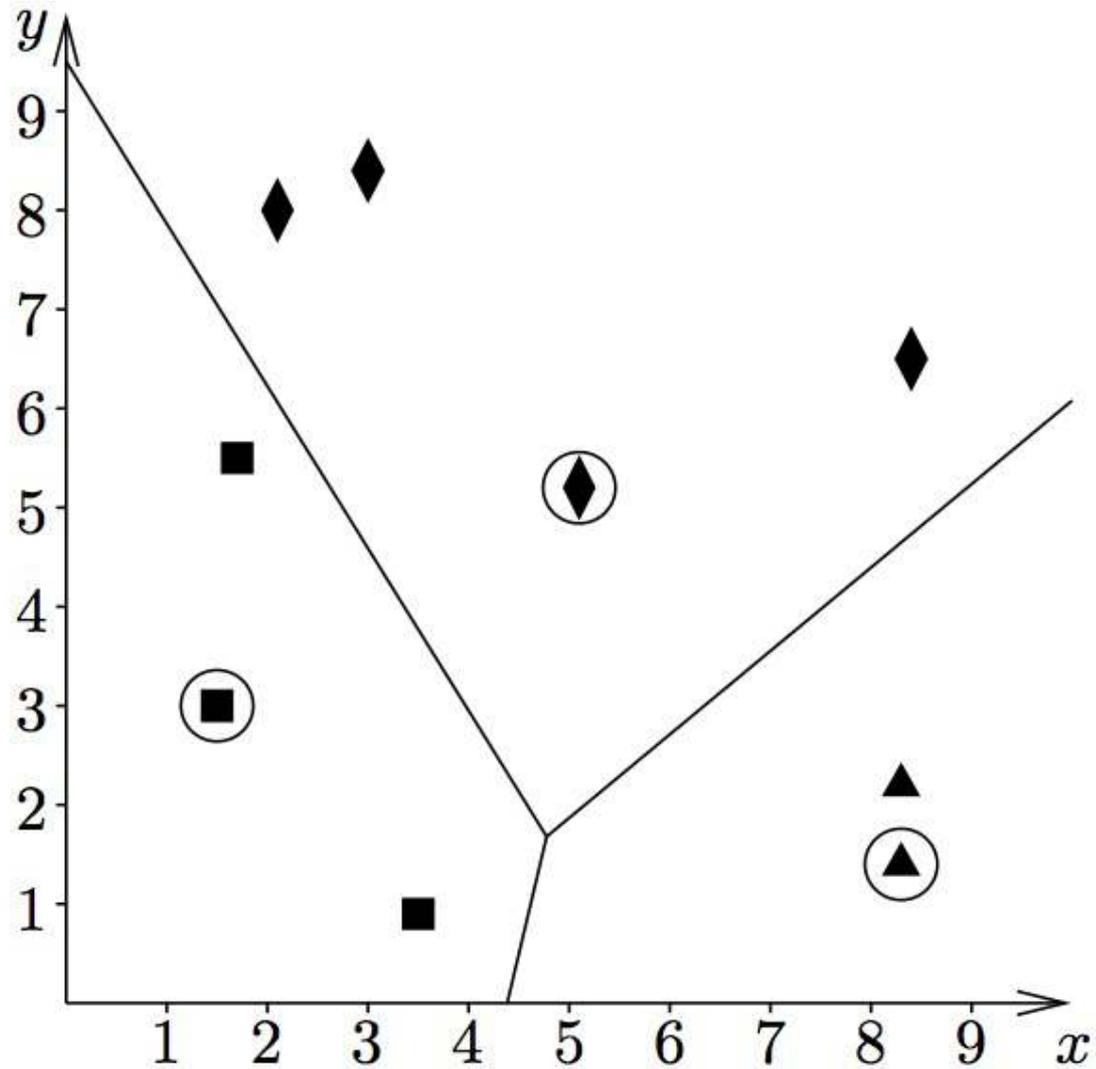
K-medoids (5)



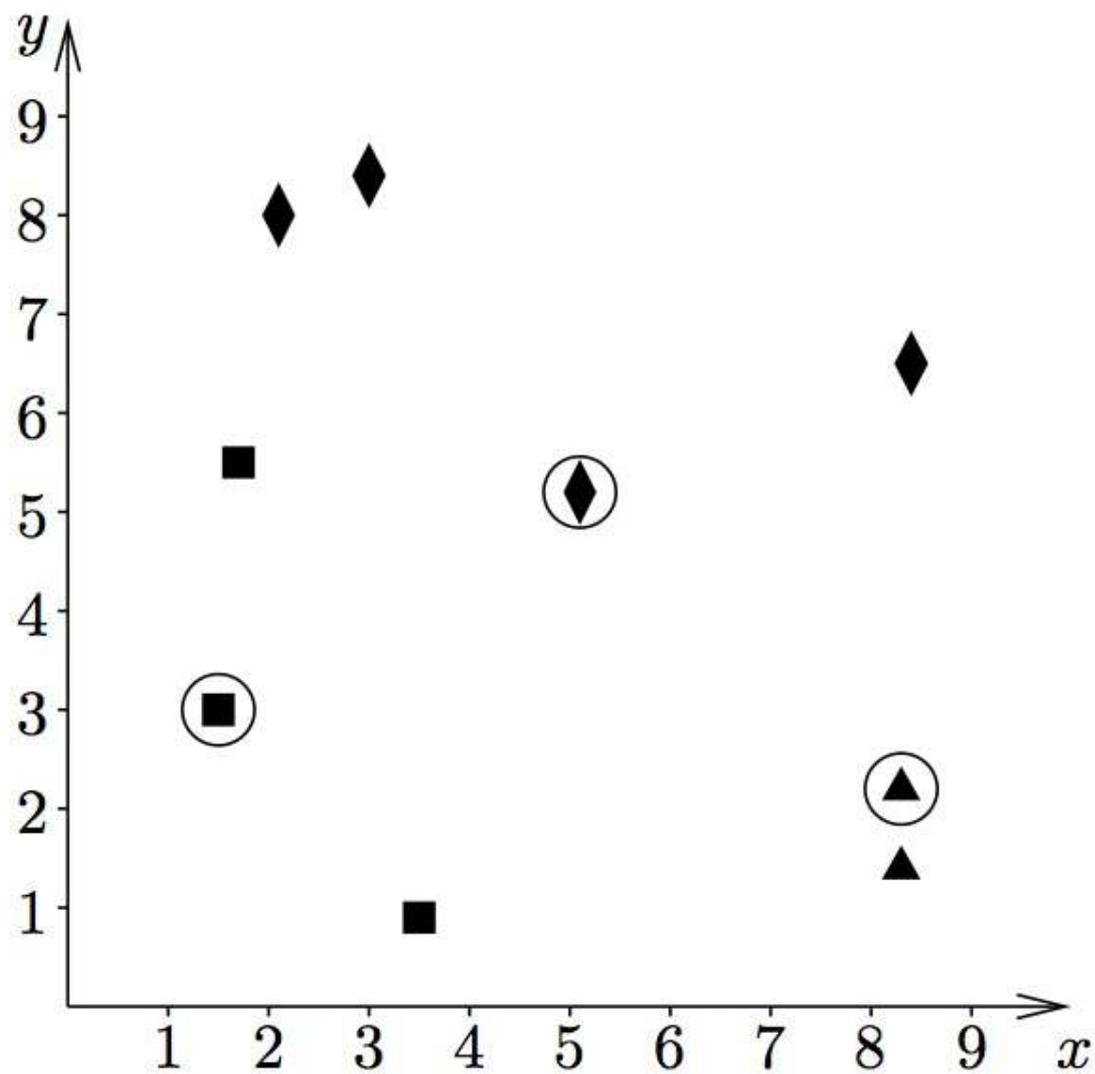
K-medoids (6)



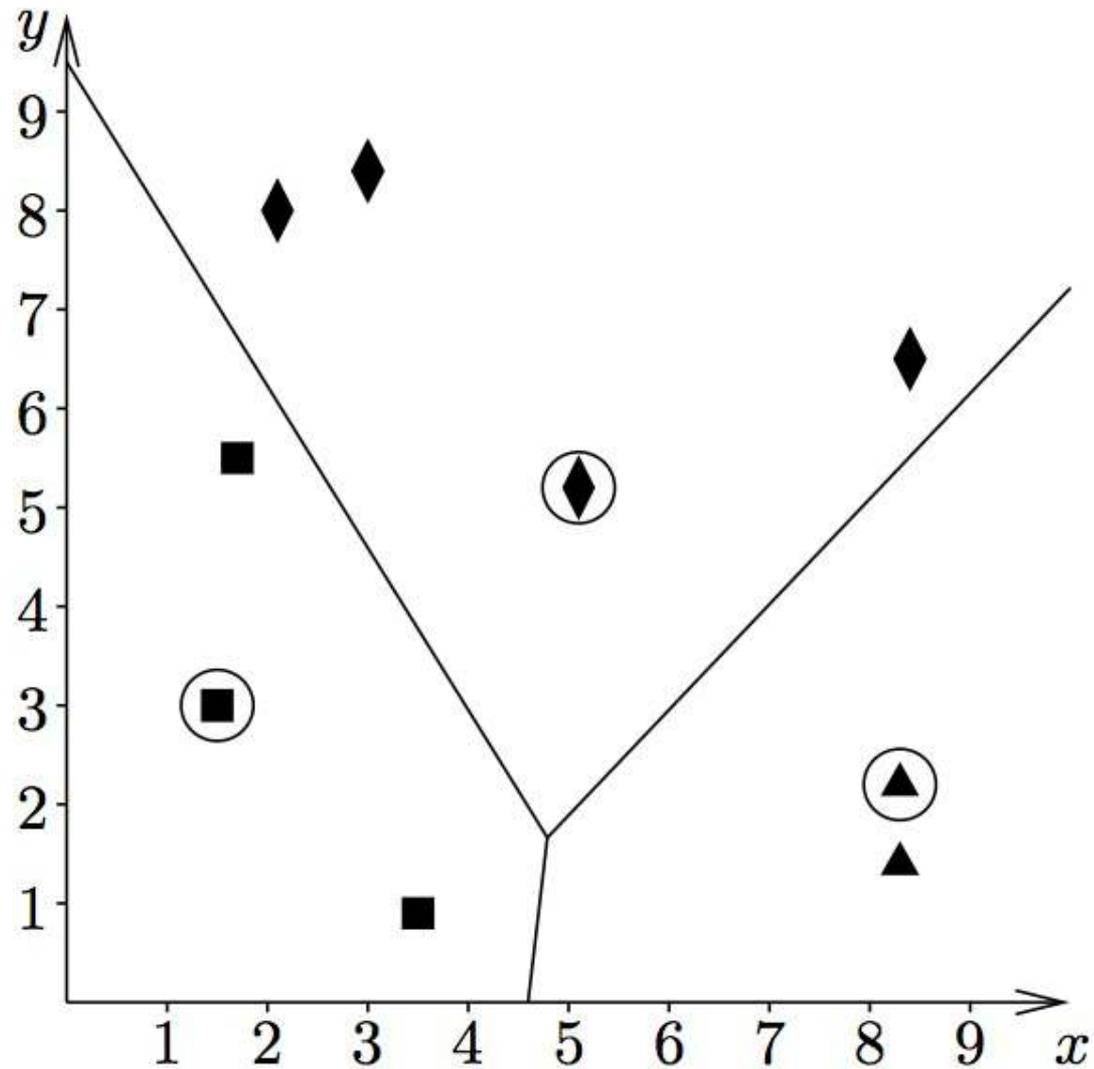
K-medoids (7)



K-medoids (8)

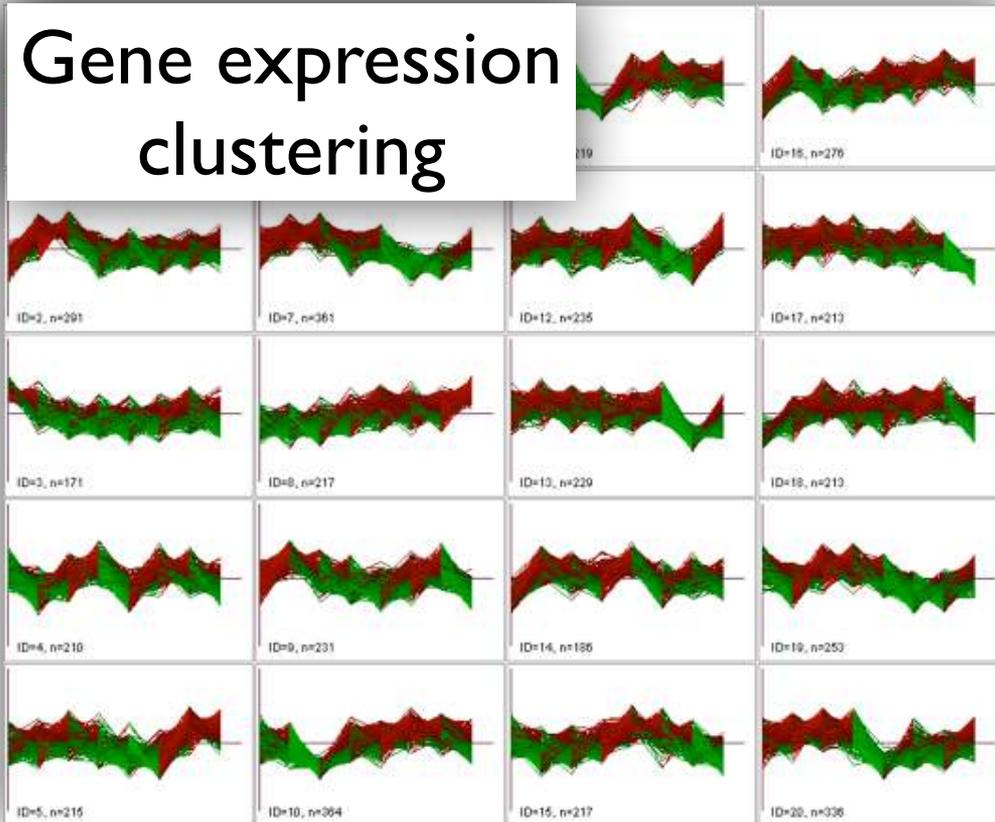


K-medoids (9)

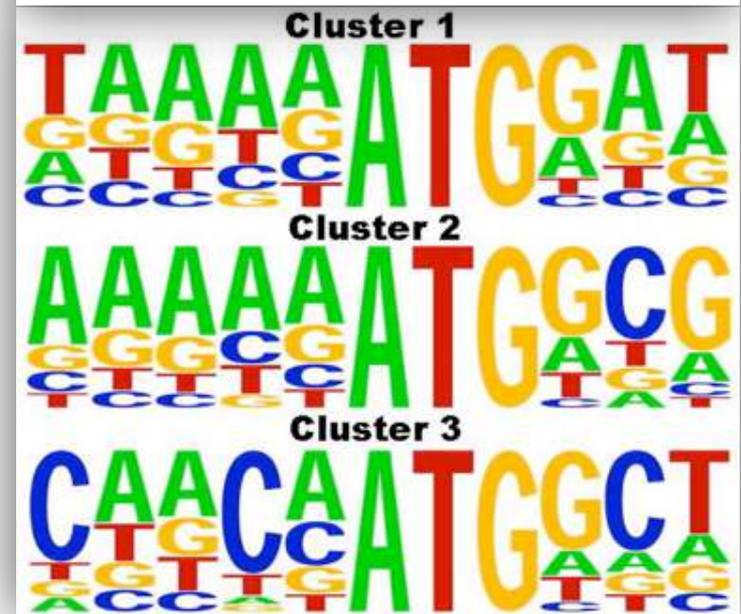


Examples of K-means and K-medoids in Bioinformatics

Gene expression clustering



Sequence clustering



Distance measures

Distance of vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$

- Euclidean distance $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
- Manhattan distance $d(x, y) = \sum_{i=1}^n |x_i - y_i|$
- Correlation distance $d(x, y) = 1 - r(x, y)$ $r(x, y)$ is Pearson correlation coefficient

Distance of sequences ACCTTG and TACCTG

- Hamming distance $\begin{array}{c} \underline{\mathbf{A}}\underline{\mathbf{C}}\underline{\mathbf{C}}\underline{\mathbf{T}}\underline{\mathbf{T}}\underline{\mathbf{G}} \\ \underline{\mathbf{T}}\underline{\mathbf{A}}\underline{\mathbf{C}}\underline{\mathbf{C}}\underline{\mathbf{T}}\underline{\mathbf{G}} \end{array} \Rightarrow 3$
- Levenshtein distance $\begin{array}{c} \underline{\mathbf{\cdot}}\underline{\mathbf{A}}\underline{\mathbf{C}}\underline{\mathbf{C}}\underline{\mathbf{T}}\underline{\mathbf{T}}\underline{\mathbf{G}} \\ \underline{\mathbf{T}}\underline{\mathbf{A}}\underline{\mathbf{C}}\underline{\mathbf{C}}\underline{\mathbf{\cdot}}\underline{\mathbf{T}}\underline{\mathbf{G}} \end{array} \Rightarrow 2$



+1!
LEVEL
UP

Interpretation
validation

Import
data

Data
analysis

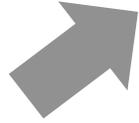
Summarize/
plot raw data



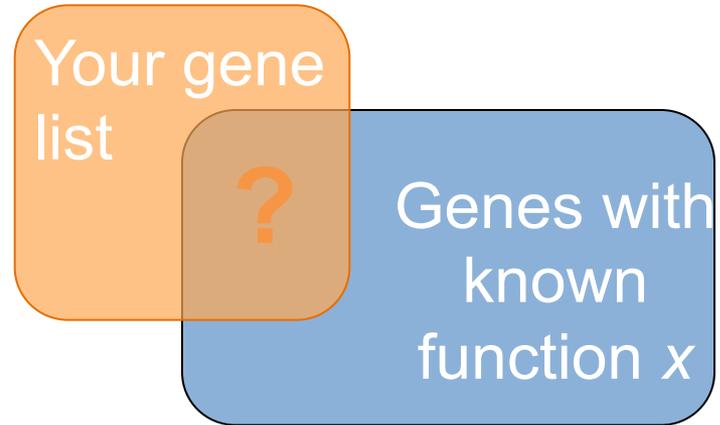
Handle
outliers

Impute missing
values

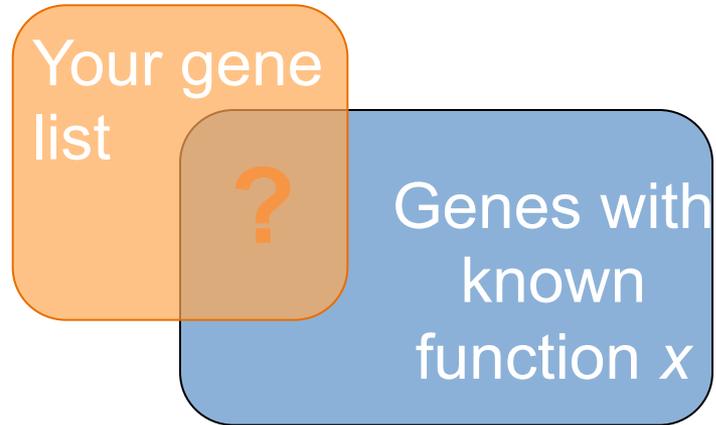
Normalize/
Standardize



Functional enrichment statistics

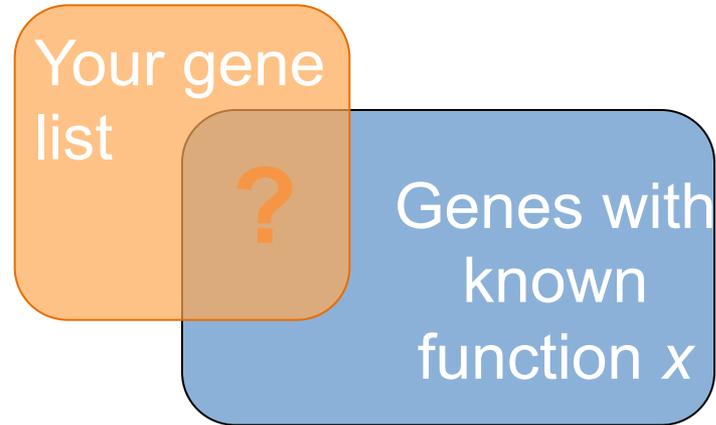


Functional enrichment statistics



Does your gene list includes **more** genes with function x than expected by random chance?

Functional enrichment statistics



Does your gene list includes **more** genes with function x than expected by random chance?

$$p = \sum_{i=k_{\pi}}^{\min(n, K_{\pi})} \frac{\binom{K_{\pi}}{i} \binom{N-K_{\pi}}{n-i}}{\binom{N}{n}}$$

g:Profiler toolset

<http://biit.cs.ut.ee/gprofiler>

The logo for g:Profiler, featuring the text "g:Profiler" in a blue and orange font, with a stylized bar chart graphic to the right.

[g:GOST](#) Gene Group Functional Profiling

[g:Cocoa](#) Compact Compare of Annotations

[g:Convert](#) Gene ID Converter

[g:Sorter](#) Expression Similarity Search

[g:Orth](#) Orthology search

[g:SNPense](#) Convert rsID

[Welcome!](#)

[Contact](#)

[FAQ](#)

[R / APIs](#)

[Beta](#)

[Archive](#)

J. Reimand, M. Kull, H. Peterson, J. Hansen, J. Vilo: *g:Profiler - a web-based toolset for functional profiling of gene lists from large-scale experiments* (2007) NAR 35 W193-W200

Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, Jaak Vilo: *g:Profiler -- a web server for functional interpretation of gene lists (2016 update)* Nucleic Acids Research 2016; doi: 10.1093/nar/gkw199

Reading the output

- >> **g:Convert**
Gene ID Converter
- >> **g:Orth**
Orthology Search
- >> **g:Sorter**
Expression Similarity Search
- >> **g:Cocoa**
Compact Compare of Annotations
- >> **Static URL**
Come back later

Significantly enriched network of BioGRID interactions found. >> [Show network](#)

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value	
BP	negative regulation of biological process	GO:0048519	4572	25	13	4.52e-02	
BP	negative regulation of cellular process	GO:0048523	4200	25	13	1.68e-02	
BP	embryo development	GO:0009790	1014	12	5	4.59e-02	
BP	embryonic morphogenesis	GO:0048598	593	8	4	2.95e-02	

Statistics

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	ncRNA metabolic process	GO:0034660	475	50	10	8.05e-04
BP	trRNA metabolic process	GO:0006399	199	50	6	3.02e-02

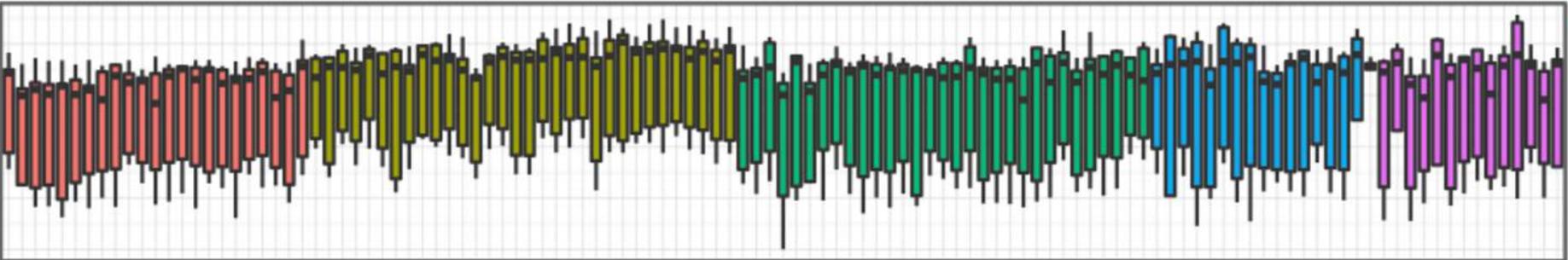
Your genes
50

GO:0034660
ncRNA metabolic process
475 genes

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	cell fate commitment involved in formation of primary germ layer	GO:0060795	34	2	2	6.95e-03
BP	endodermal cell fate commitment	GO:0001711	16	2	2	1.49e-03
BP	endodermal cell fate specification	GO:0032743	9	2	2	4.46e-04
BP	negative regulation of nucleobase-containing compound metabolism	GO:0048999	1422	25	8	4.59e-02
BP	anatomical structure formation involved in morphogenesis	GO:0032743	1160	31	9	9.11e-03
BP	dosage compensation	GO:0032743	16	15	1	3.27e-02
BP	dosage compensation by inactivation of X chromosome	GO:0032743	8	16	2	3.27e-02
BP	negative regulation of RNA biosynthetic process	GO:0032743	1252	20	7	4.59e-02
BP	negative regulation of nucleic acid-templated transcription	GO:0032743	1236	20	7	4.23e-02

Cluster annotation

n: 8



Legionellosis
chemokine-mediated signaling pathway
positive regulation of immune system process
response to other organism
Influenza A

NOD-like receptor signaling pathway
chemokine-mediated signaling pathway
positive regulation of immune system process
regulation of cell proliferation

Arsoebiasis
Salmonella infection

Rheumatoid arthritis
positive regulation of leukocyte chemotaxis
positive regulation of response to stimulus

Chemokine signaling pathway

TNF signaling pathway
Rheumatoid arthritis
positive regulation of leukocyte chemotaxis
positive regulation of response to stimulus

Chemokine receptors bind chemokines
response to lipopolysaccharide
Cytokine-cytokine receptor interaction

RIG-I-like receptor signaling pathway
defense response

GOsummaries

Practice time!

