UNIVERSITY OF TARTU
1632
Institute of Computer Science

Data Preprocessing Unsupervised learning

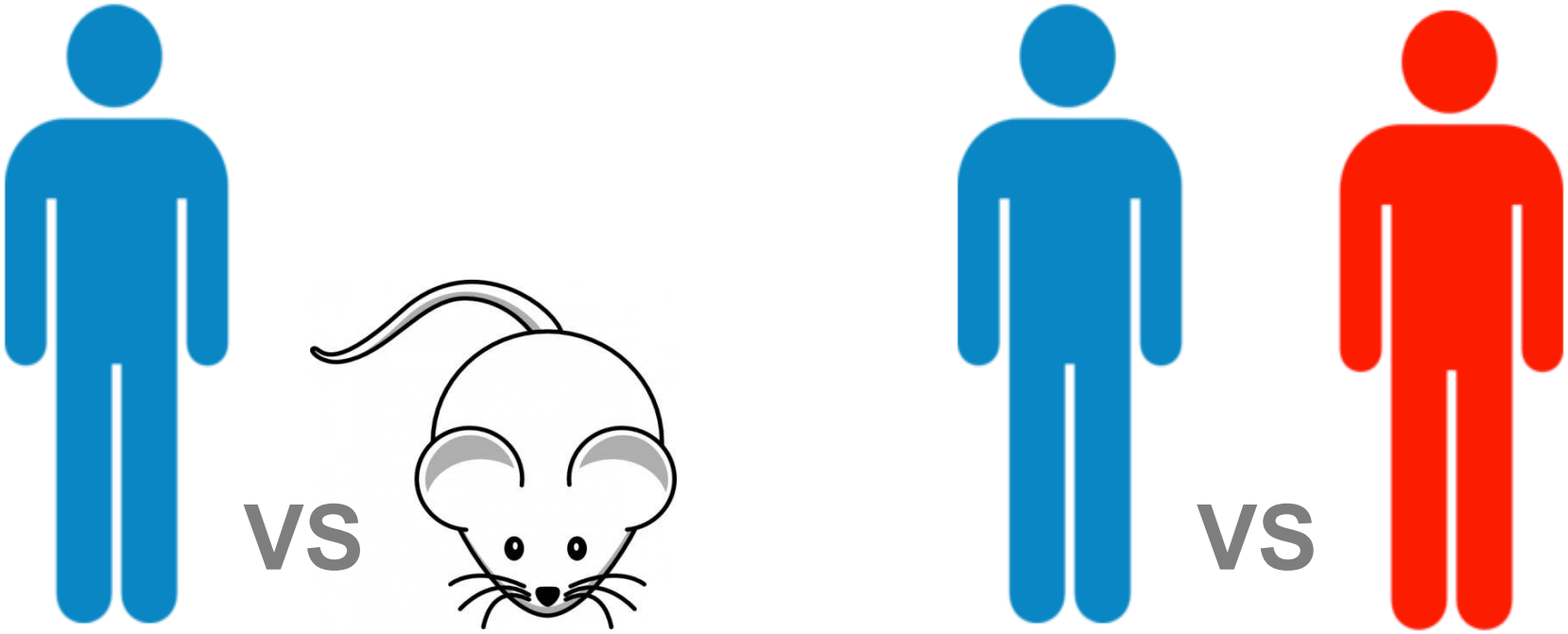
Elena Sügis

elena.sugis@ut.ee

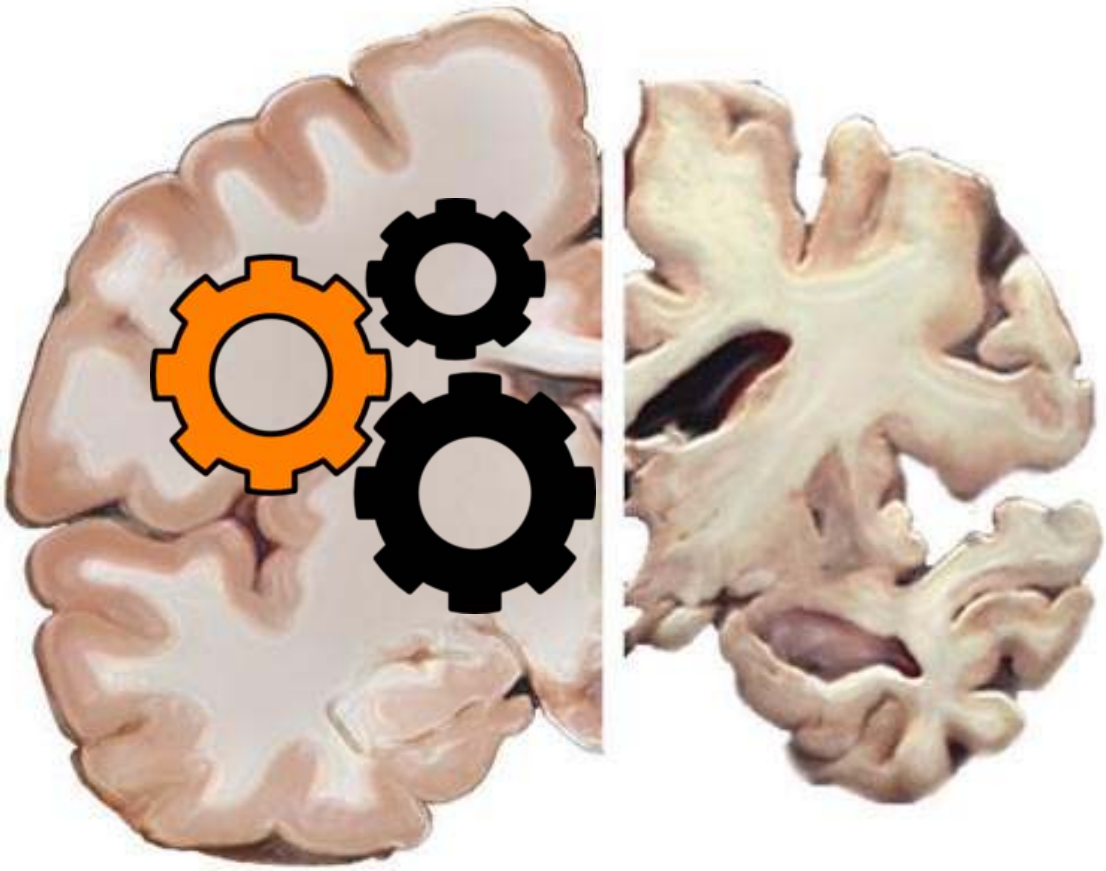
Machine learning, SPB2016

ИНСТИТУТ БИОИНФОРМАТИКИ

Questions we ask

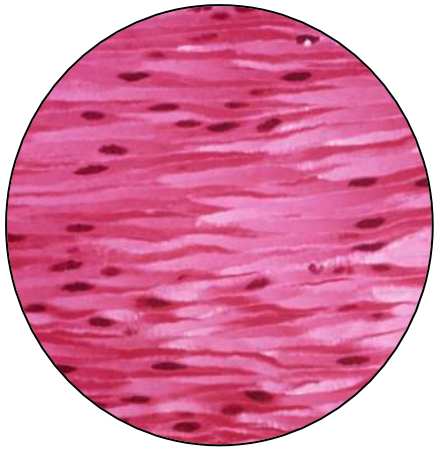


Questions we ask



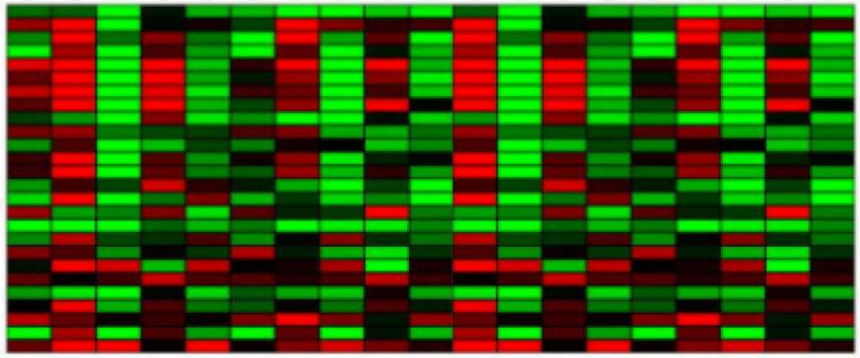
Questions we ask



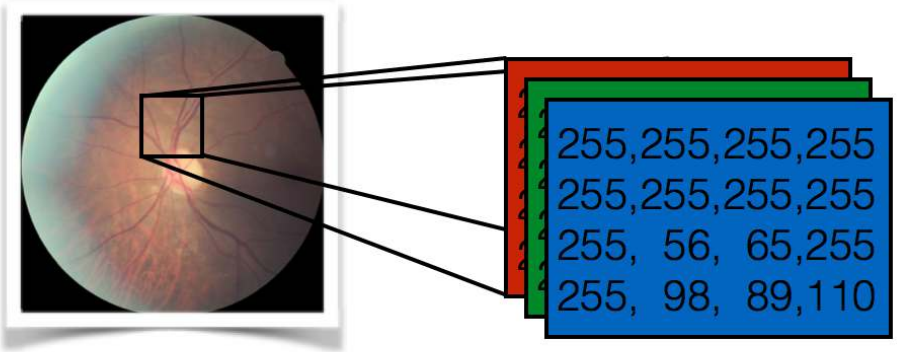



Experiments

Data comes in different forms



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

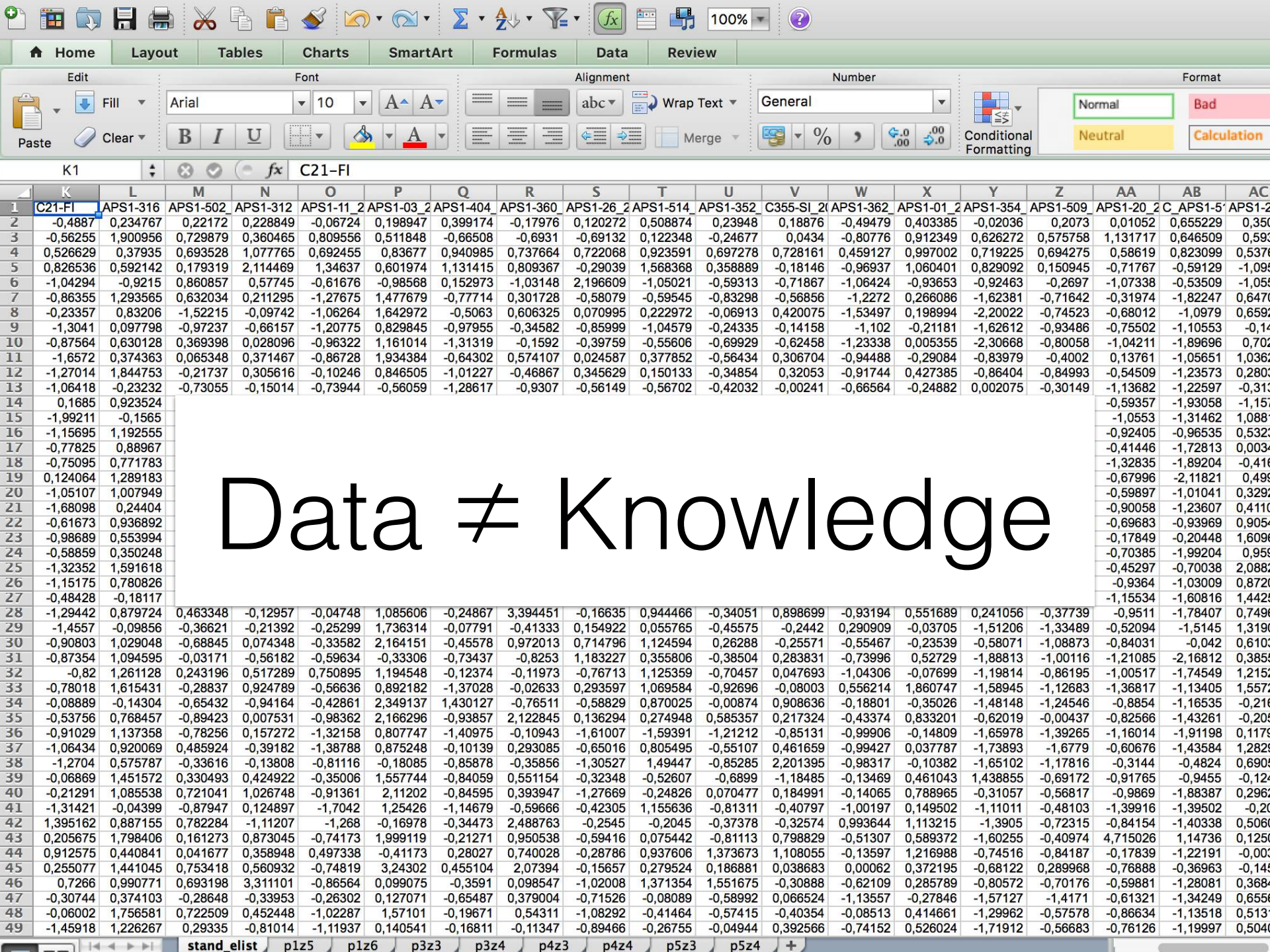


$x = (255, 255, 255, 255, \ldots)$

Diagnose: asthma

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

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


Home
Layout
Tables
Charts
SmartArt
Formulas
Data
Review
Edit
Font
Alignment
Number
Format
Paste
Fill
Clear
Arial
10
Wrap Text
Merge
General
Conditional Formatting
Normal
Bad
Neutral
Calculation
K1
C21-FI
Data ≠ Knowledge
stand_elist
p1z5
p1z6
p3z3
p3z4
p4z3
p4z4
p5z3
p5z4

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 %

Import data

?

Impute missing values

Normalize/ Standardize

Handle outliers

Data analysis

Interpretation validation

Import data

Summarize/ plot raw data

Impute missing values

Normalize/ Standardize

Handle outliers

Data analysis

Interpretation validation

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

Import data

Summarize/ plot raw data

Impute missing values

Normalize/ Standardize

Handle outliers

Data analysis

Interpretation validation

Missing Values

Origins:

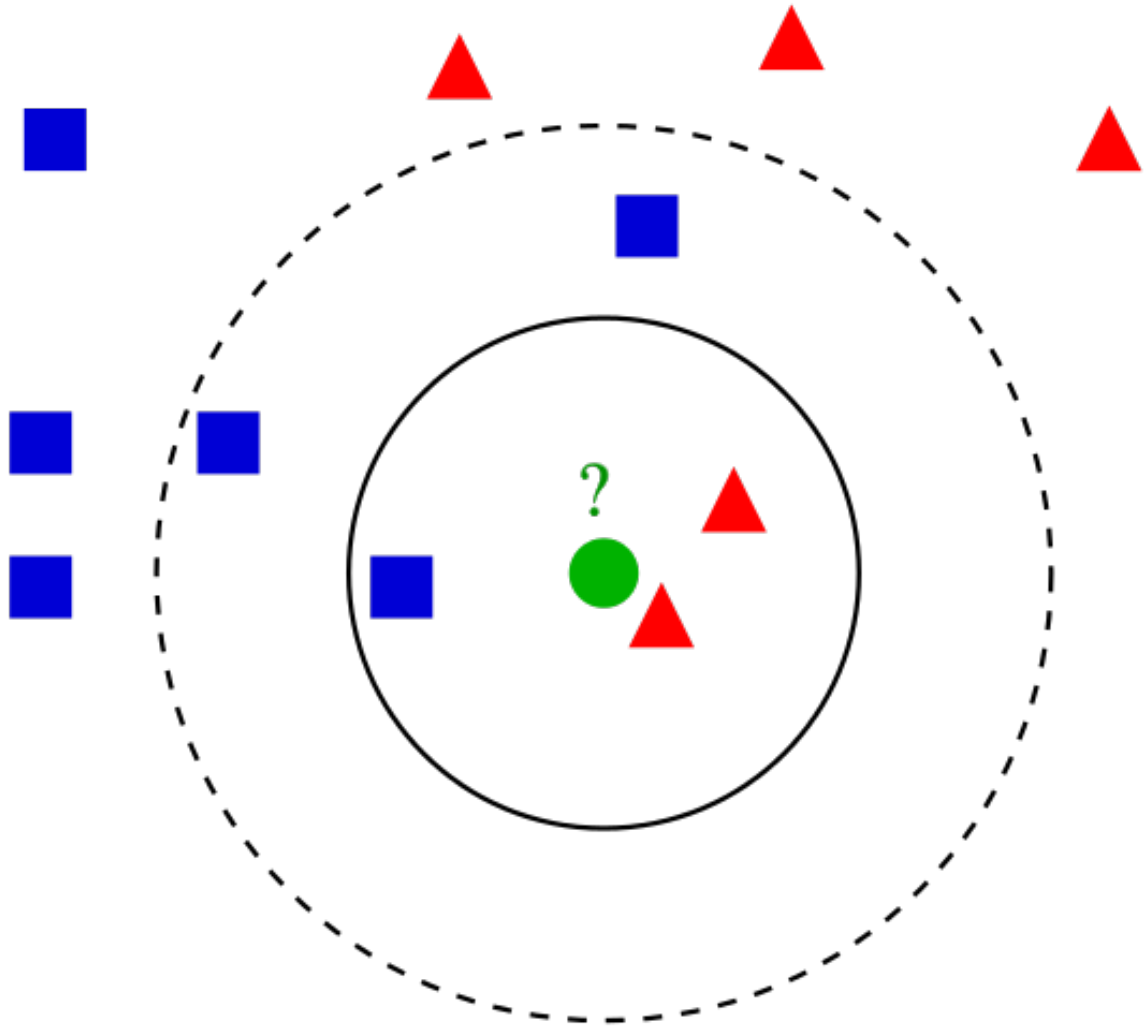
- Malfunctioning measurement equipment
- Very low intensity signal
- Deleted due to inconsistency with other recorded data
- Data removed/not entered by mistake

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

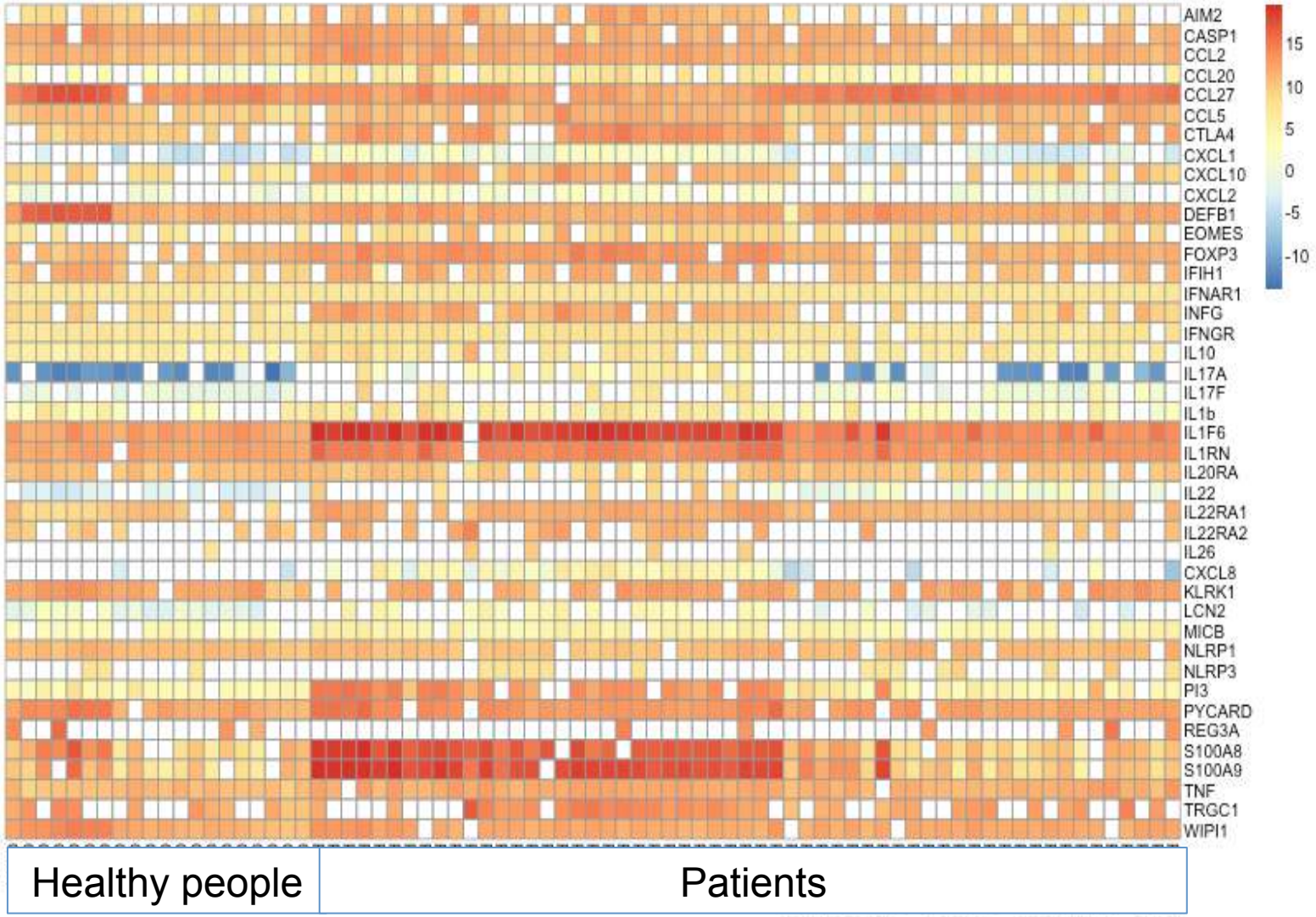
- We are given a gene expression matrix M
- Let $X=(x_1, x_2, \ldots, x_i, \ldots, x_n)$ be a vector in the matrix M with a missing value at x_i at the dimension i

- Find in the gene expression data matrix matrix vectors X1 , X2 , …, Xk , 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 $X1_i, X2_i, \ldots, Xk_i$, i.e., mean (median) of the values at dimension i of vectors X1 , X2 , …, Xk

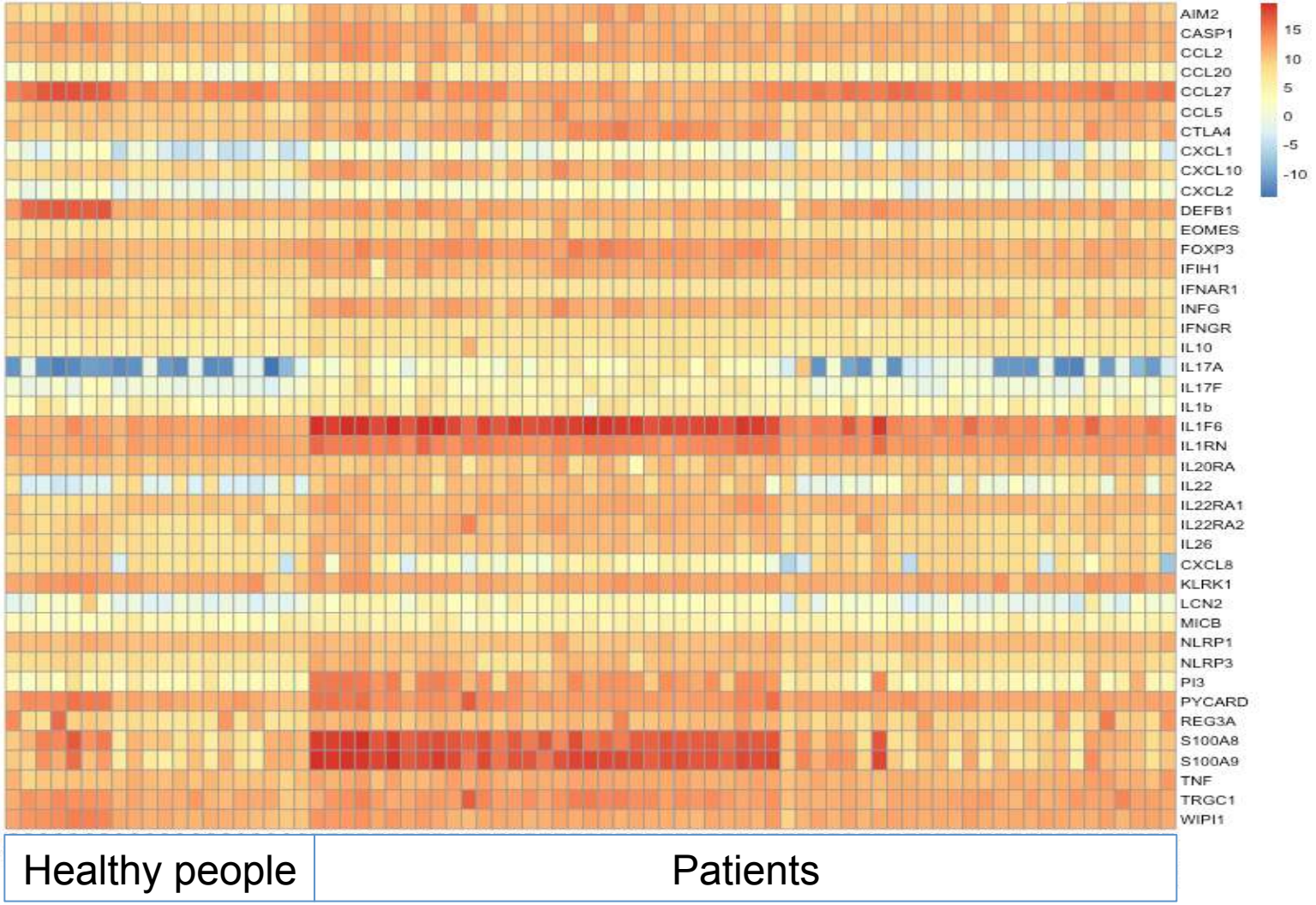
KNN

Gene expression matrix



Imputed missing values

Gene expression matrix



Import data

Summarize/ plot raw data

Impute missing values

Normalize/ Standardize

Handle outliers

Data analysis

Interpretation validation

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

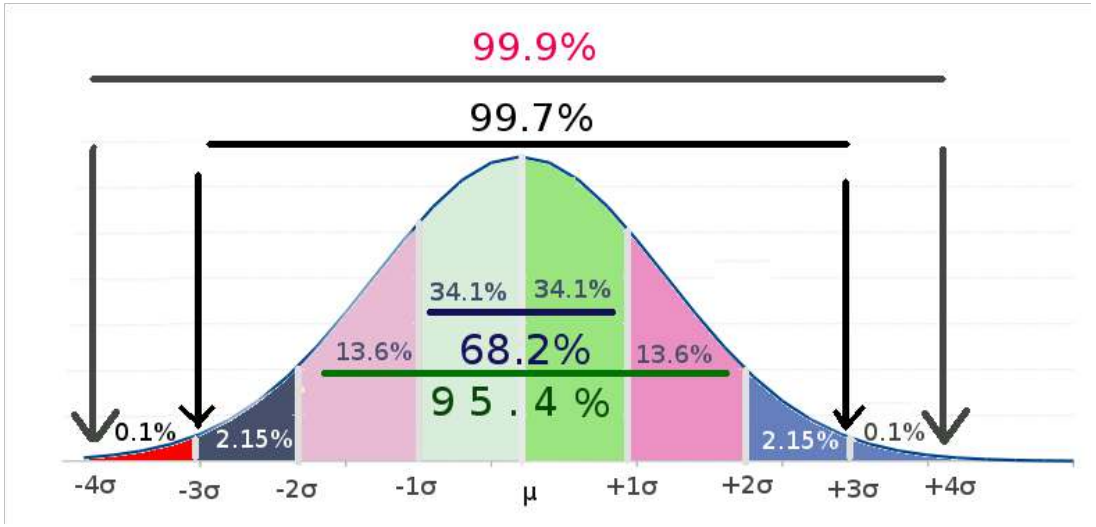
Methods:

- 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.

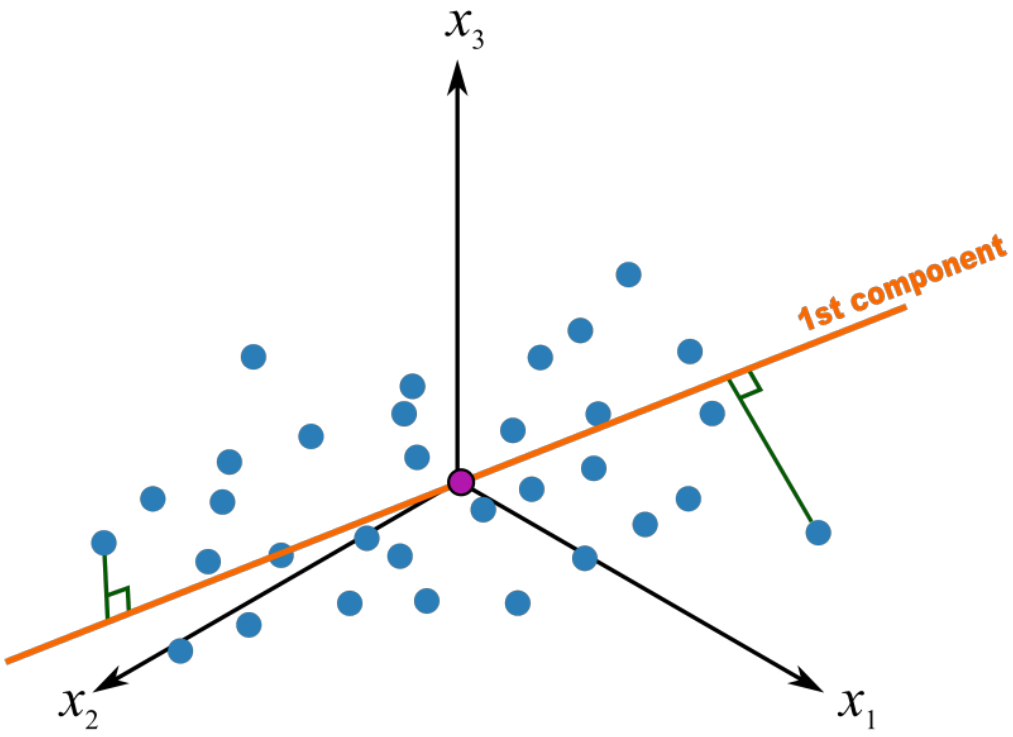
$$z = \frac{x - \mu}{\sigma}$$



where:

μ is the mean of the population.
σ is the standard deviation of the population.

Principal Component Analysis

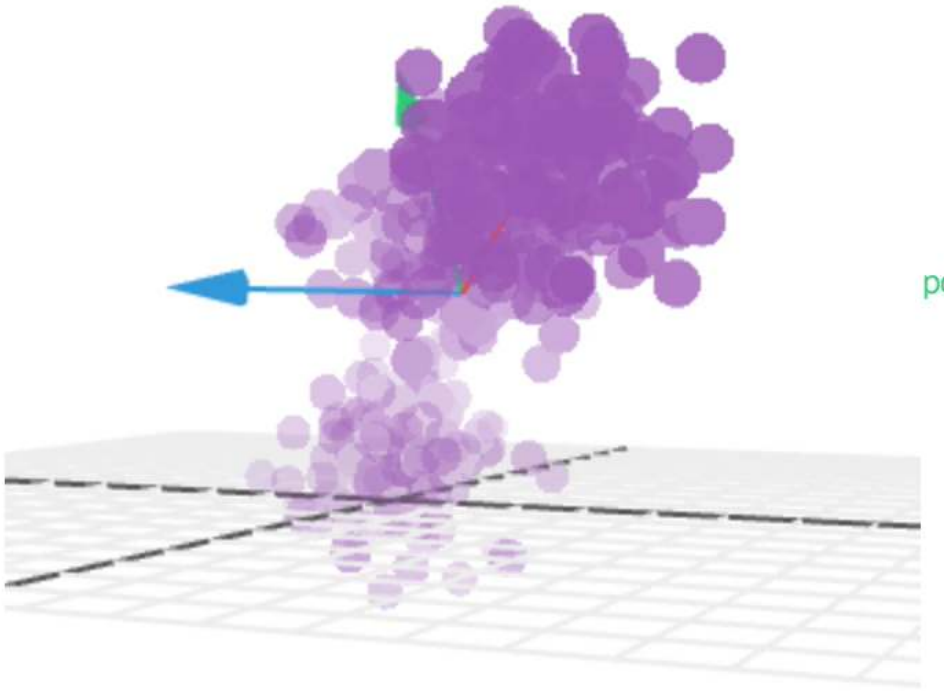


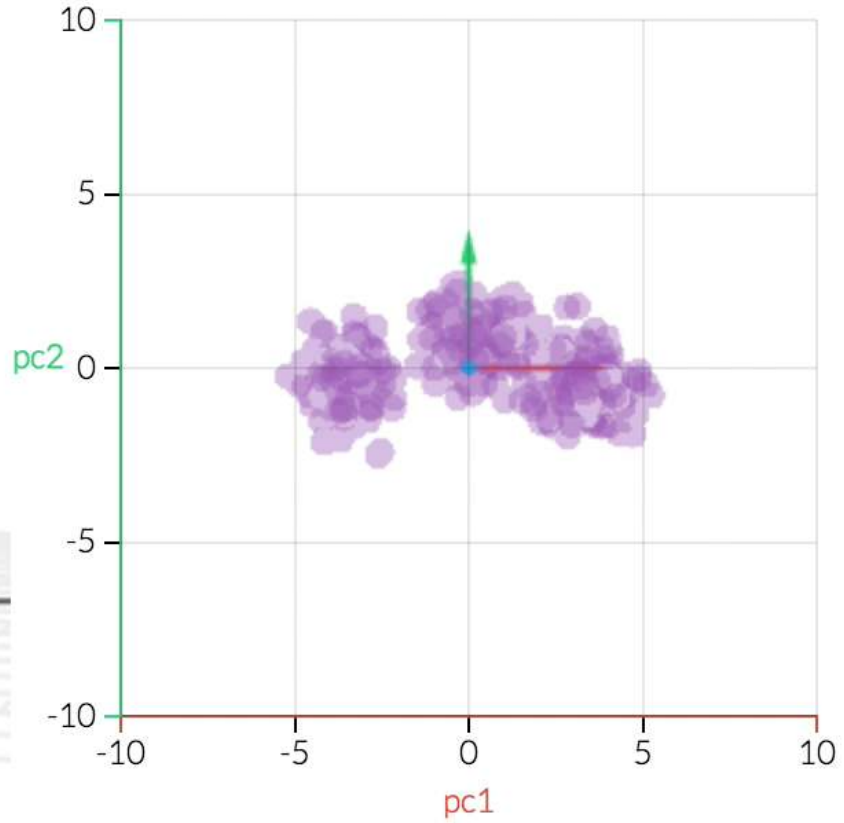
transforms the data by a linear projection onto a lower-dimensional space that preserves as much data variation as possible

Principal Component Analysis

Objective:

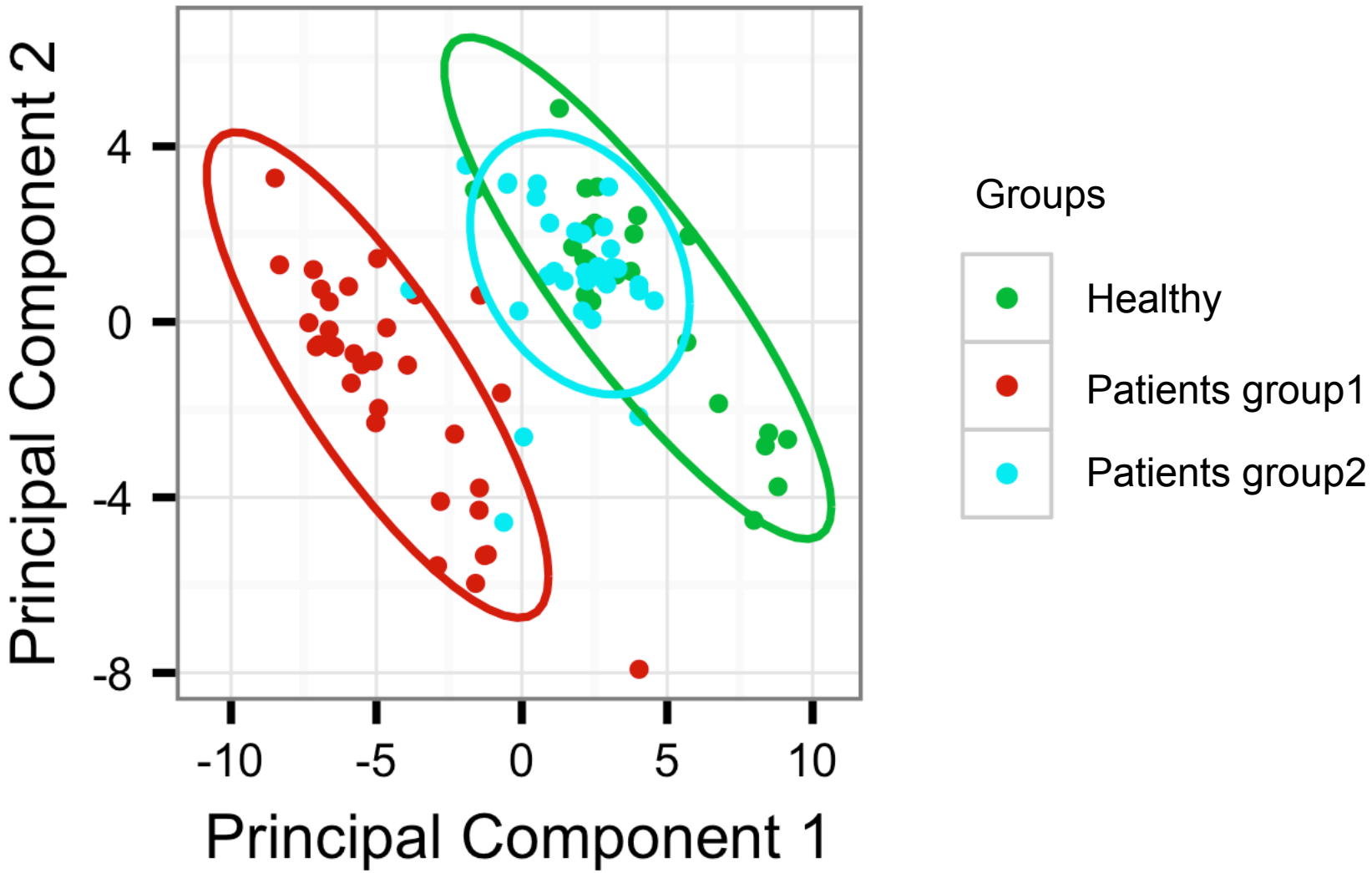
Reduce dimensionality while preserving as much variance as possible



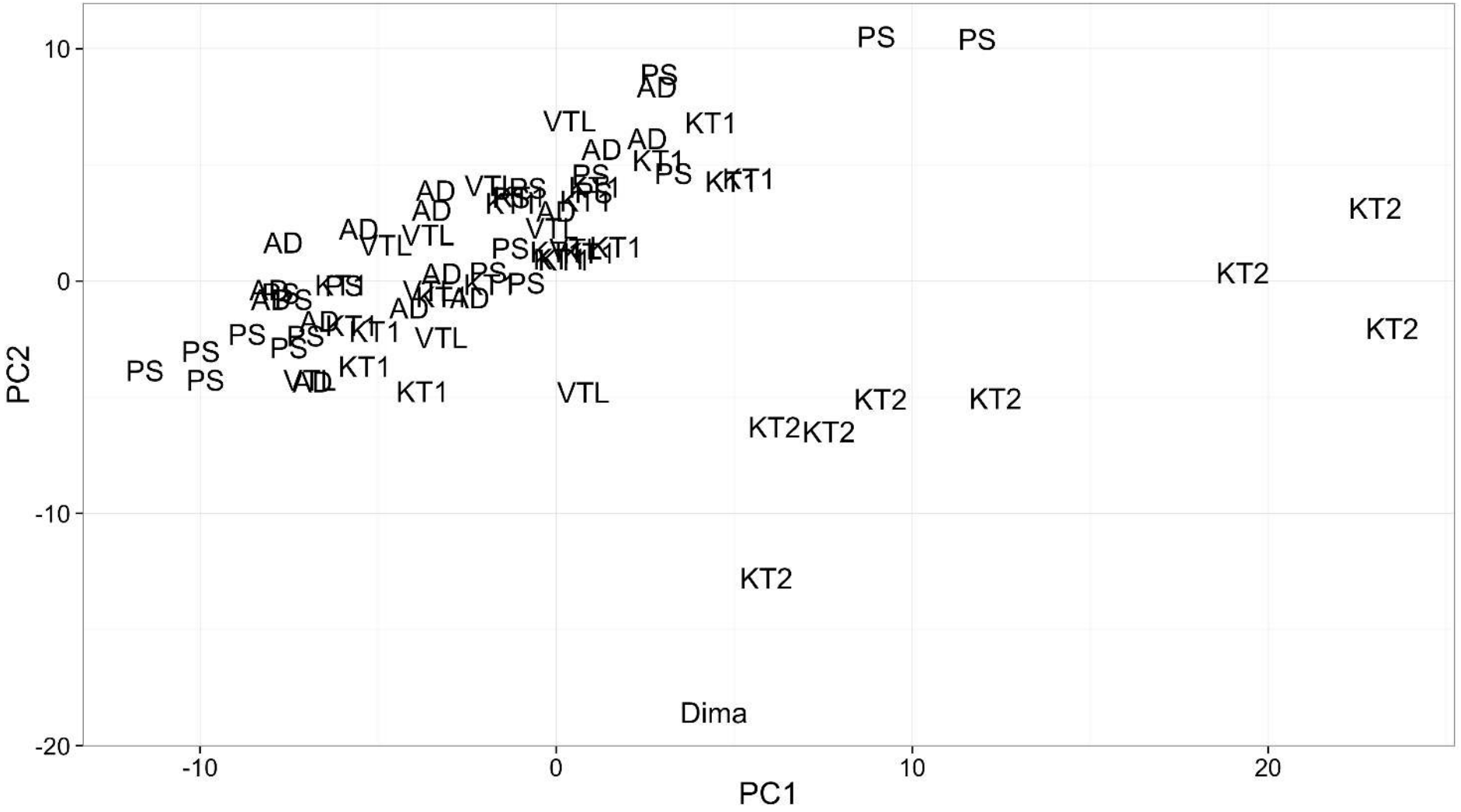


http://setosa.io/ev/principal-component-analysis/

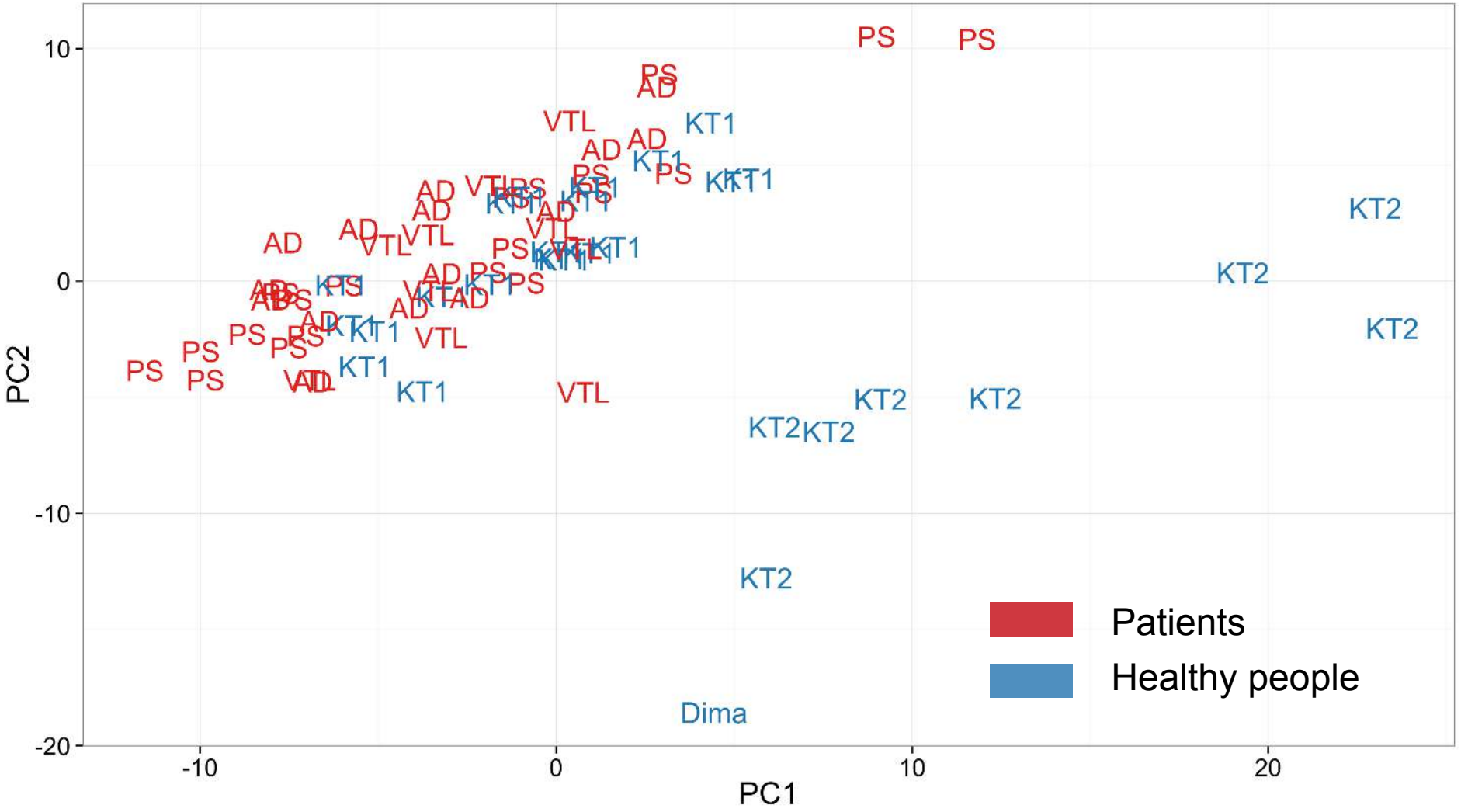
Visualize normalized data



Visual inspection after normalization



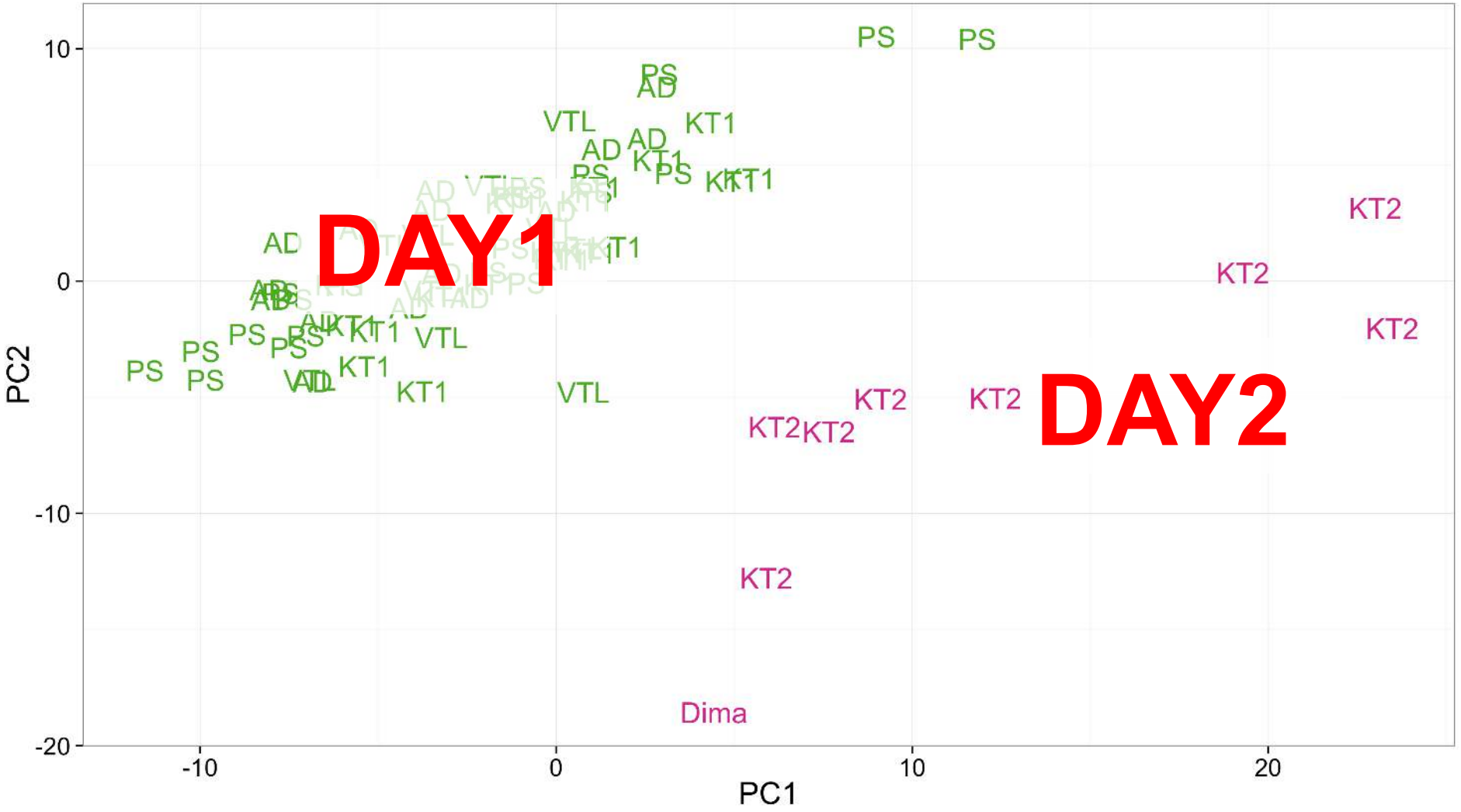
Visual Inspection. PCA



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)

J.T. Leek, Nature Reviews Genetics 11, 733-739 (October 2010,)
Chao Chen, PlosOne, 2011

KEEP CALM AND MASSAGE YOUR DATA

Interpretation validation

Import data

Data analysis

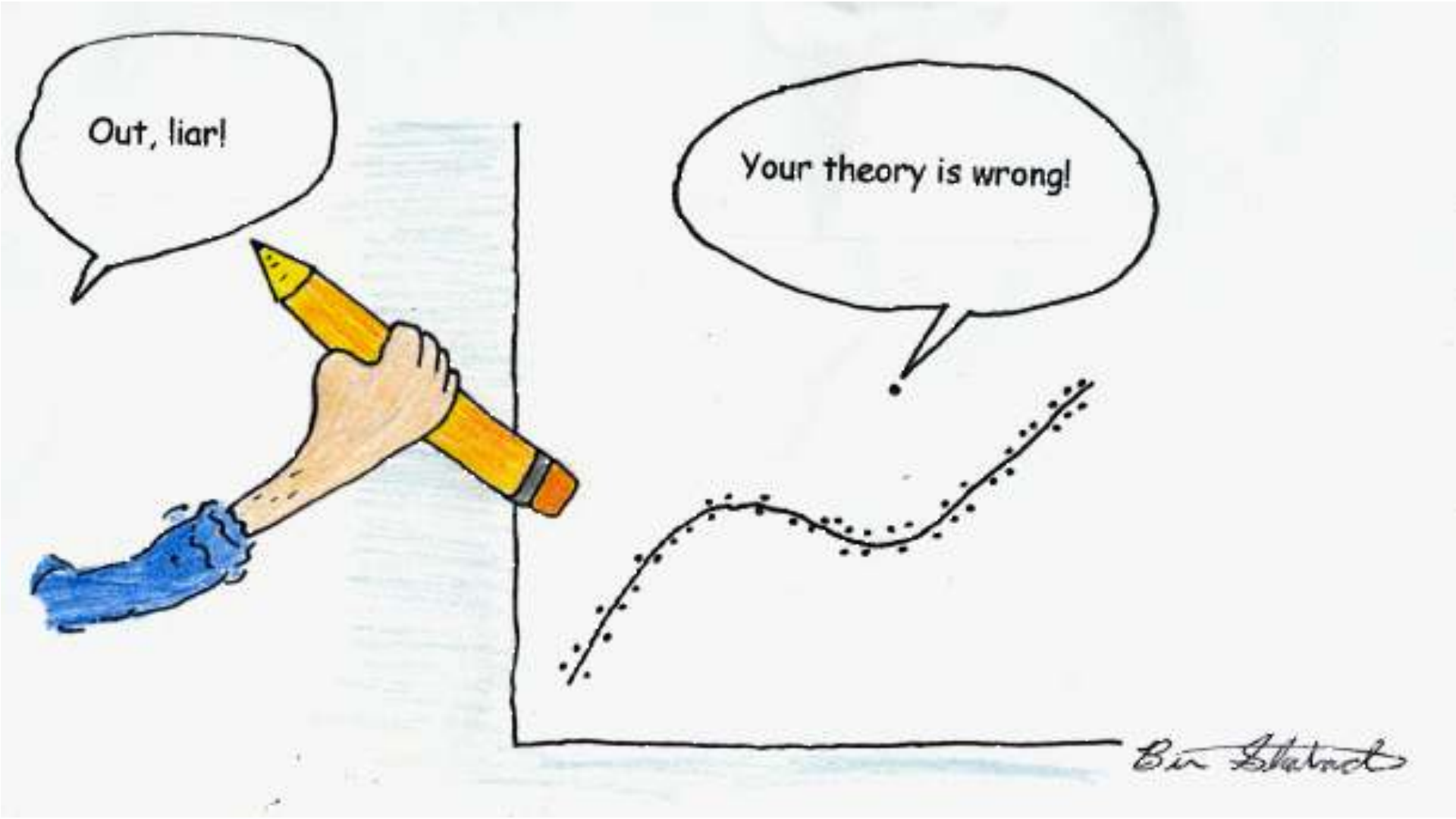
Summarize/ plot raw data

Handle outliers

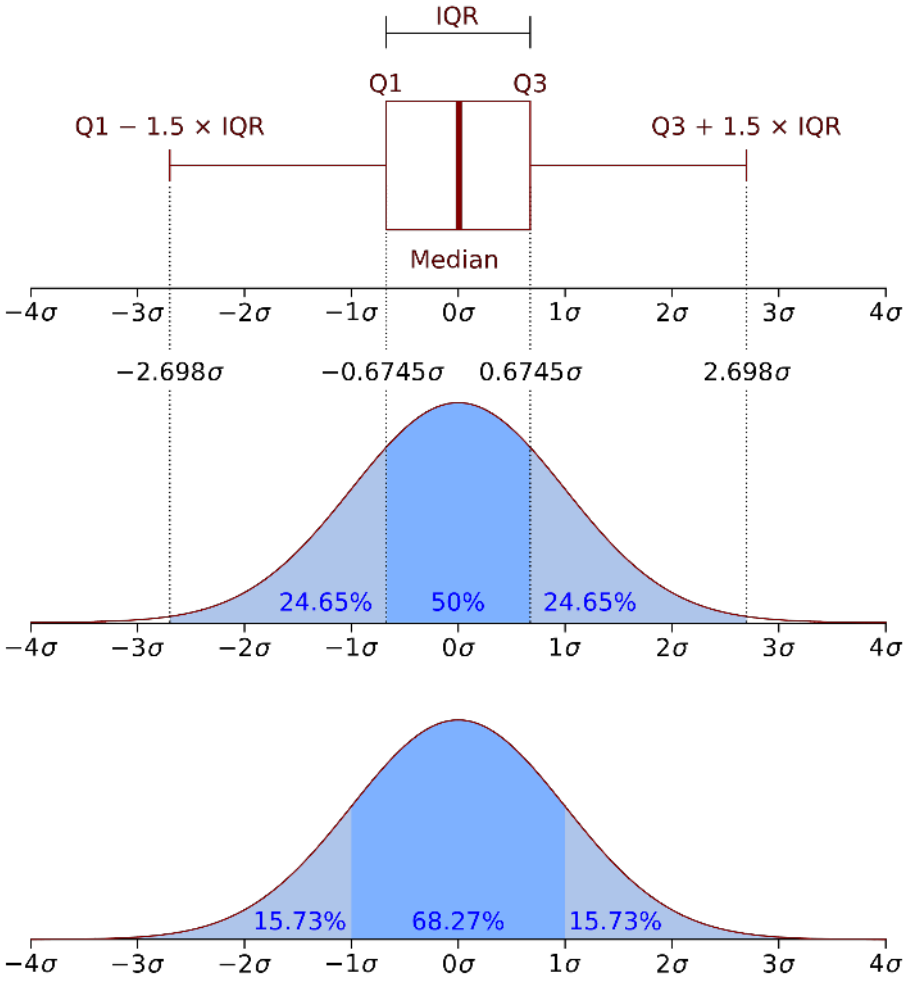
Impute missing values

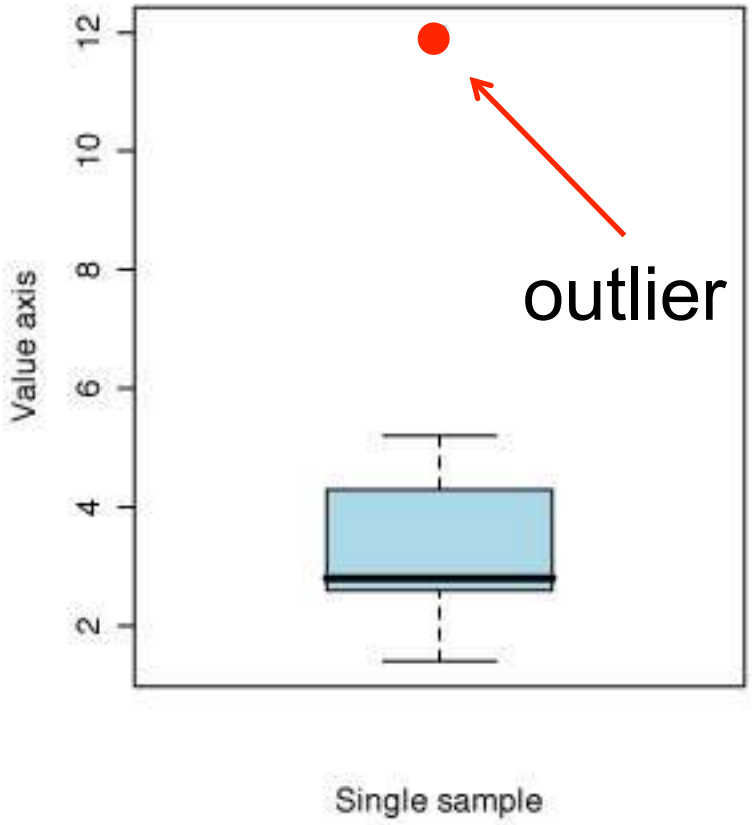
Normalize/ Standardize

Outliers Detection



Interquartile range






LEVEL
UP!

Interpretation validation

Import data

Summarize/ plot raw data

Impute missing values

Normalize/ Standardize

Handle outliers

Data analysis

IF YOU TORTURE

THE DATA

LONG ENOUGH

IT WILL CONFESS

TO ANYTHING

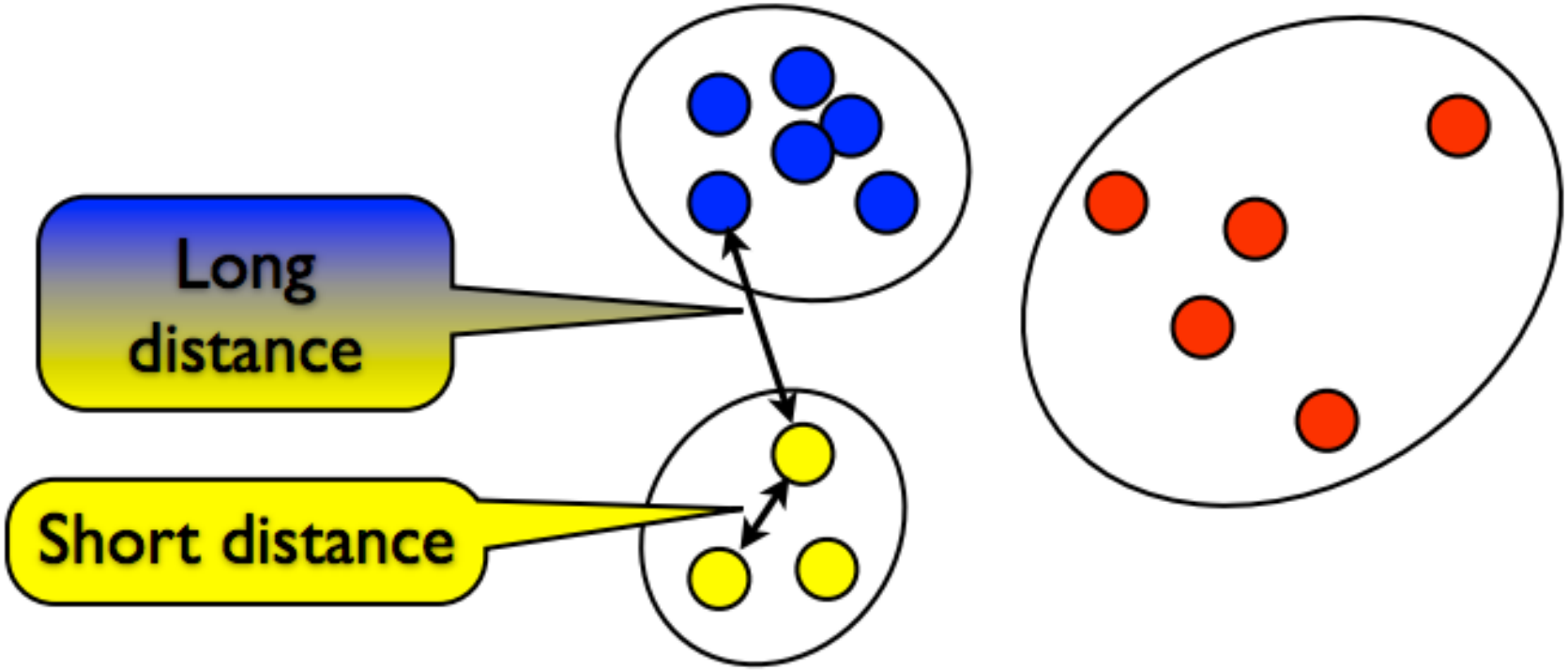
Ronald Coase, Economist, Nobel Prize winner

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

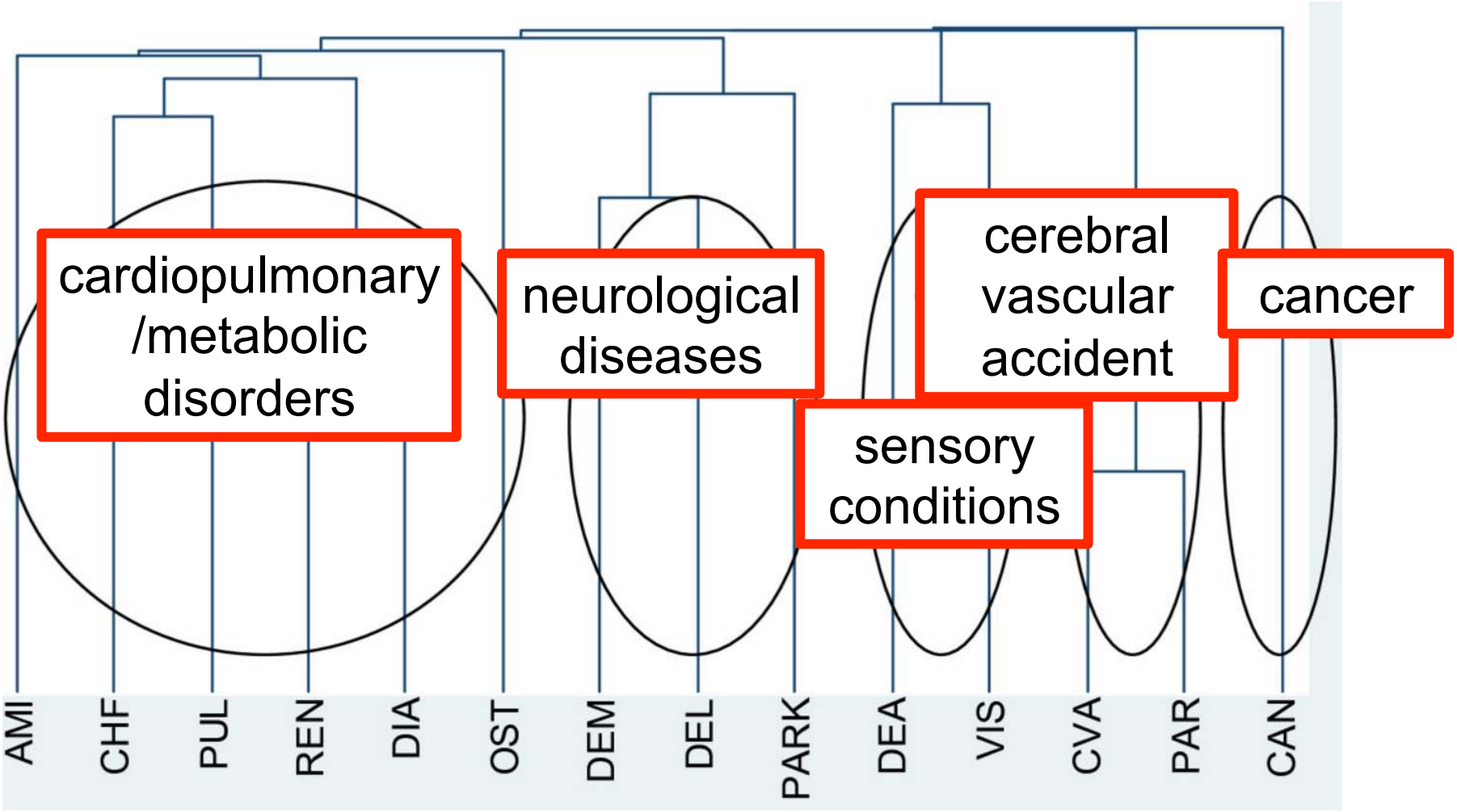
  Finding and characterizing similar groups of objects in the data

- Knowledge discovery in data

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

- Summarizing / compressing large data
- Data visualization

Intuition building



http://bmcgeriatr.biomedcentral.com/articles/10.1186/1471-2318-11-45

Why to cluster biological data?

- Intuition building

  Finding hidden internal structure of the high-dimensional data

- Hypothesis generation

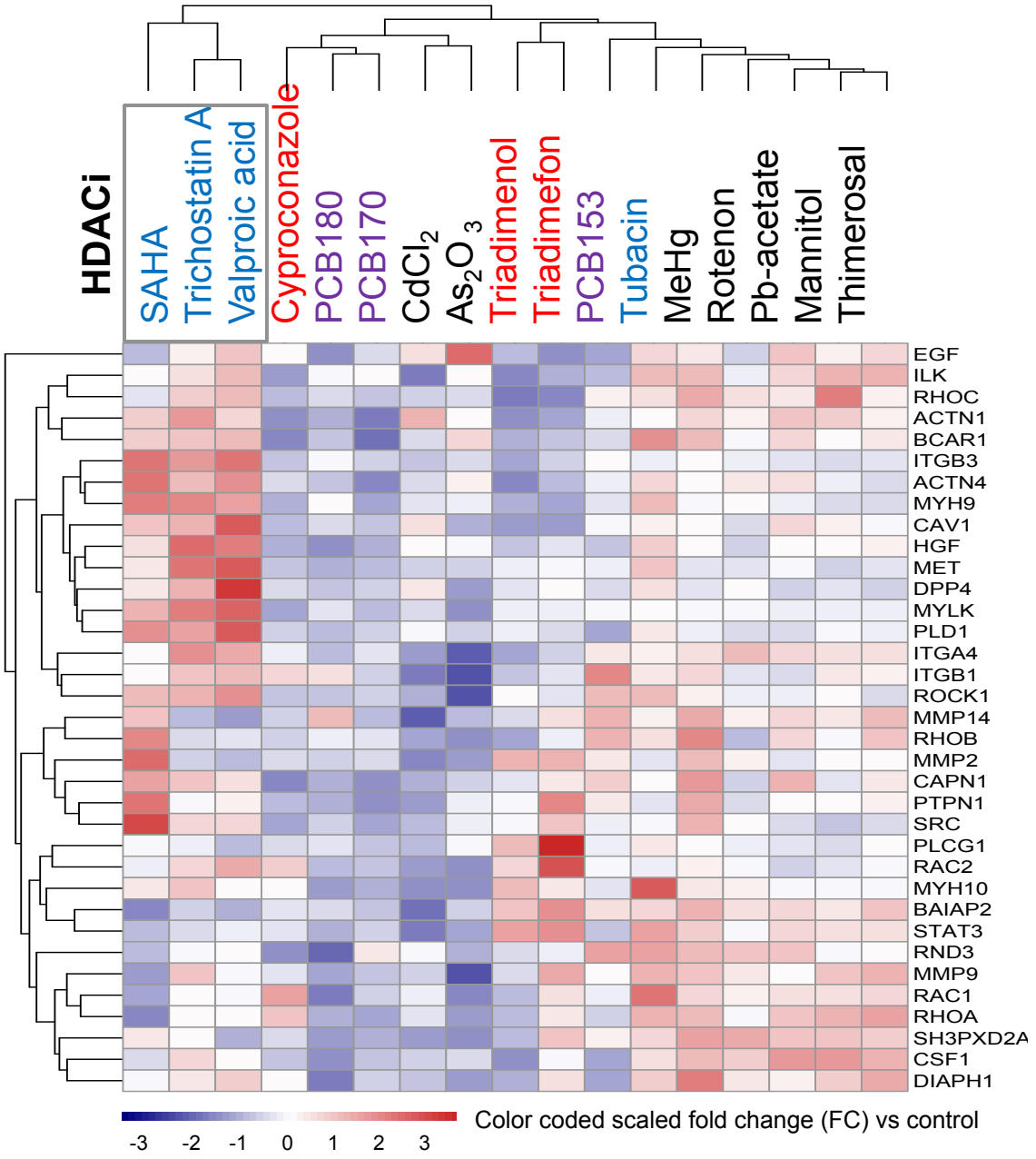
  Finding and characterizing similar groups of objects in the data

- 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

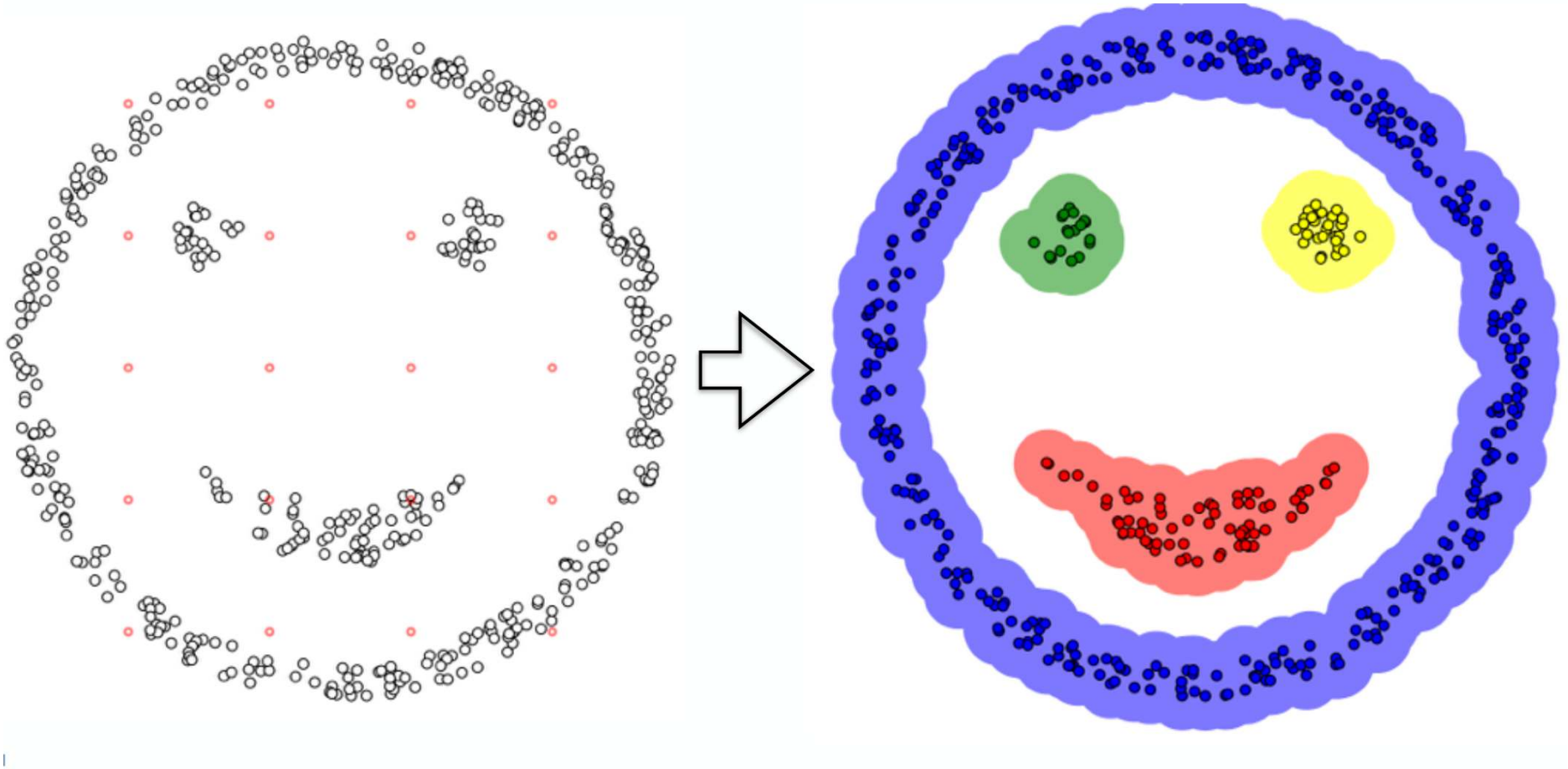
- Knowledge discovery in data

  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

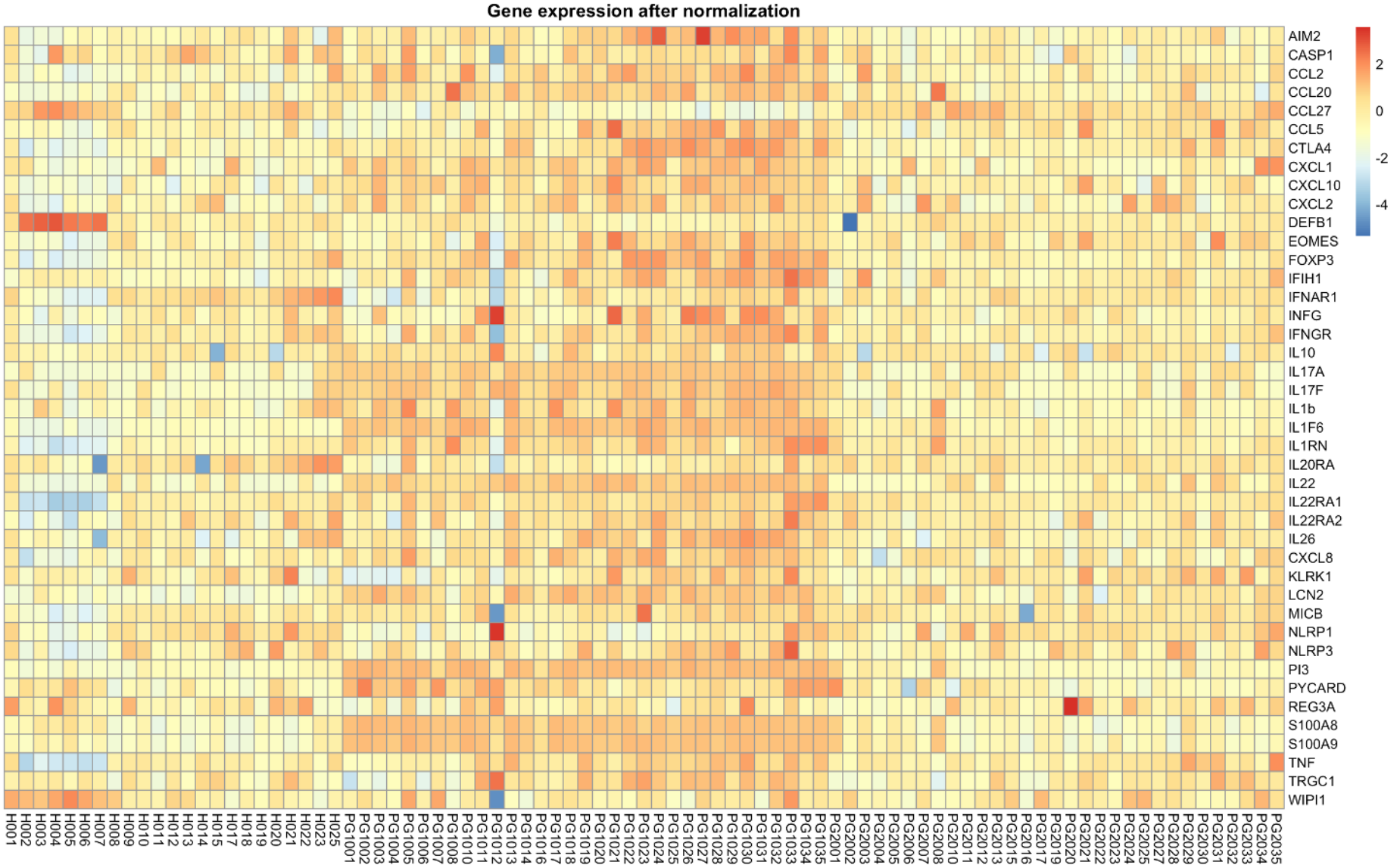
  Finding and characterizing similar groups of objects in the data

- Knowledge discovery in data

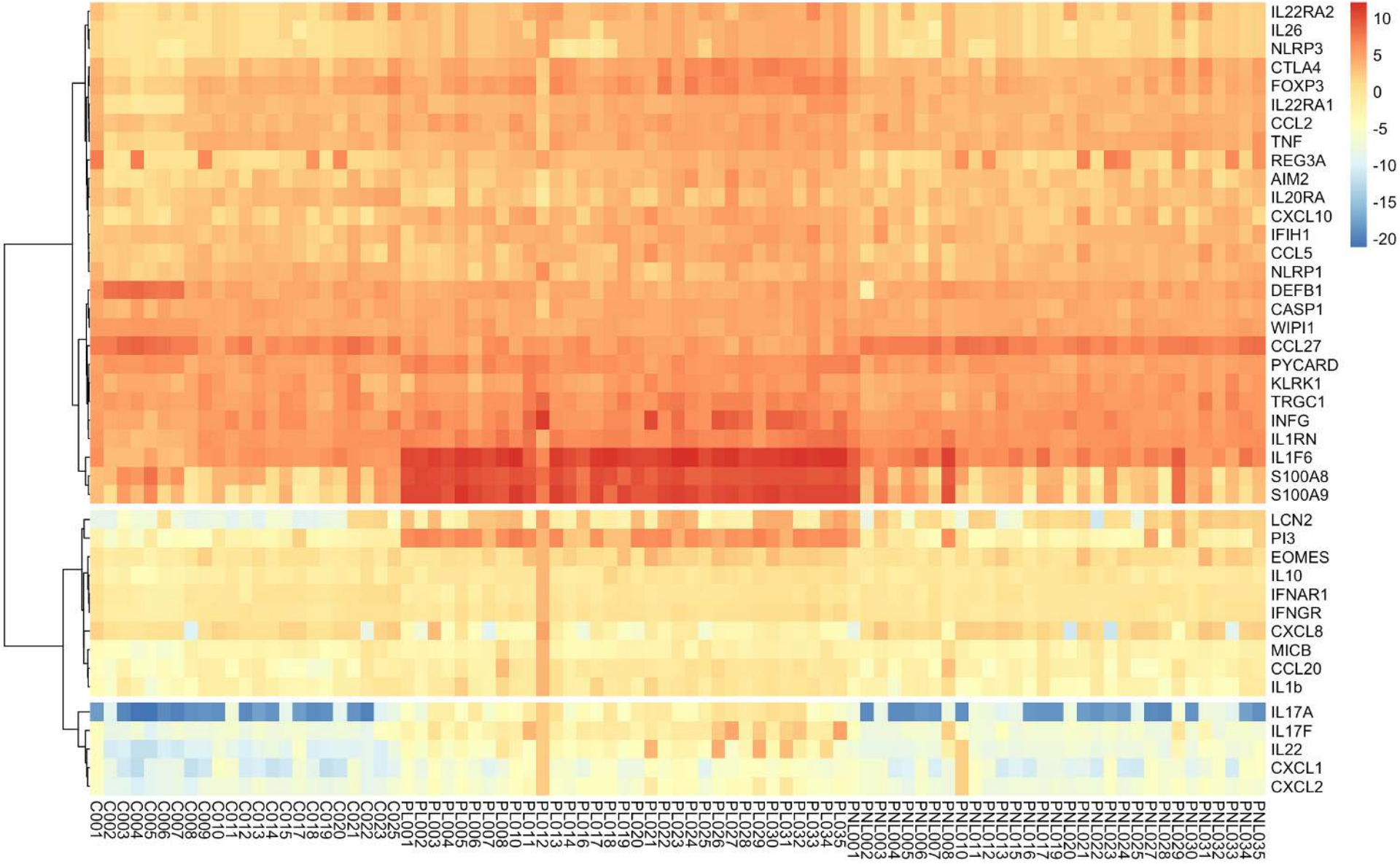
  Ex. Underlying rules, reoccurring patterns, topics, etc.

- Summarizing / compressing large data
- Data visualization

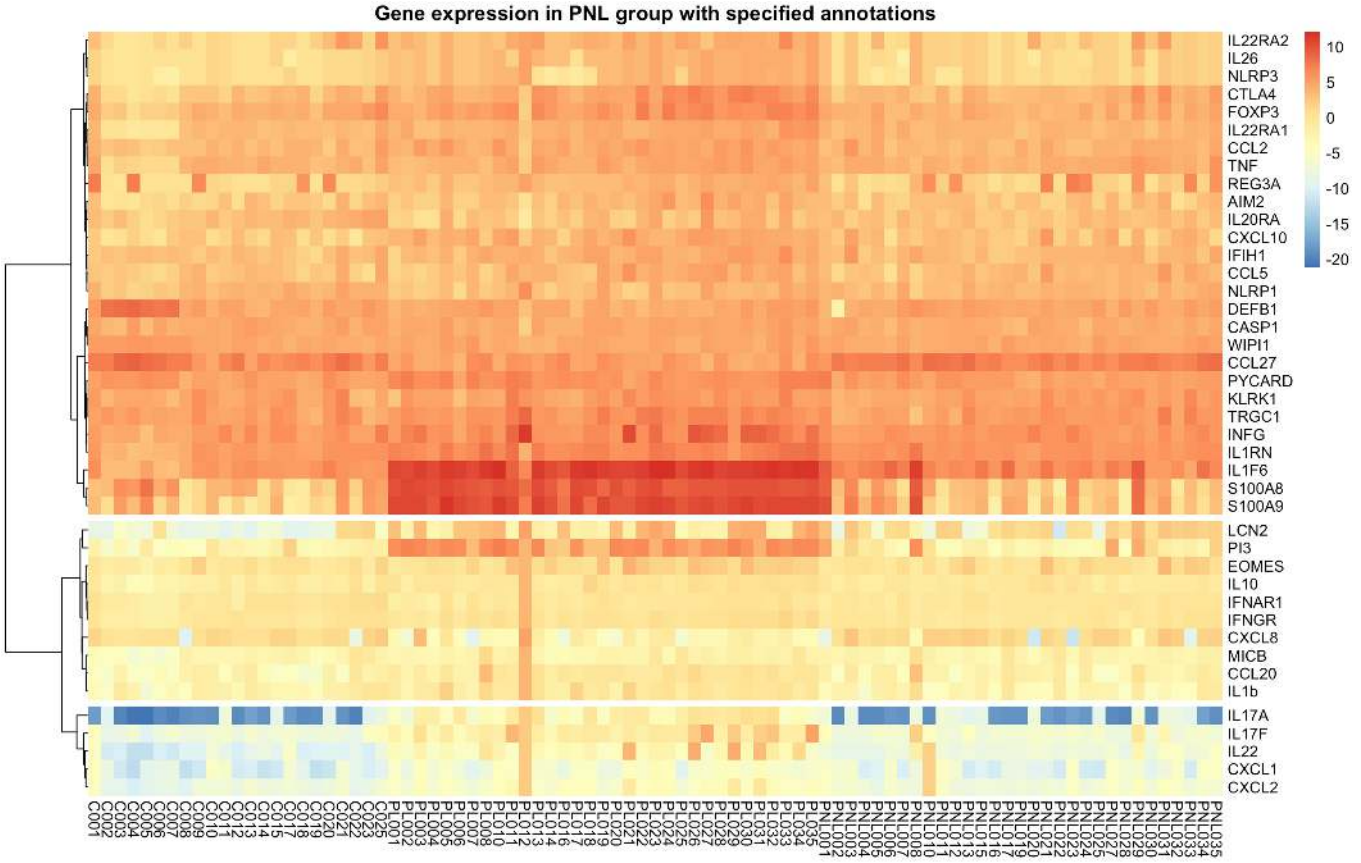
Summarizing/compressing the data

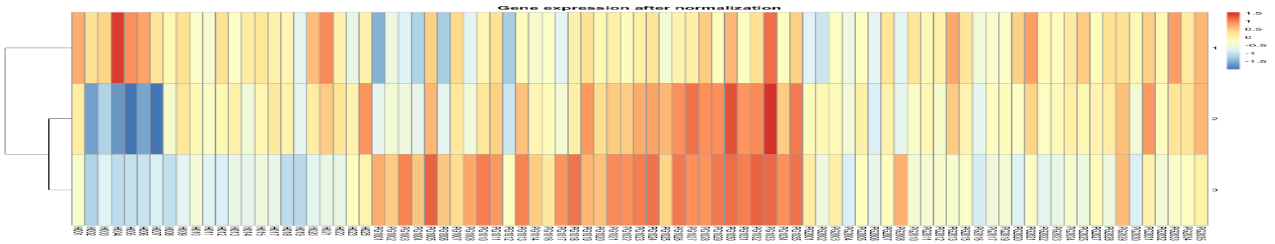


Summarizing/compressing the data

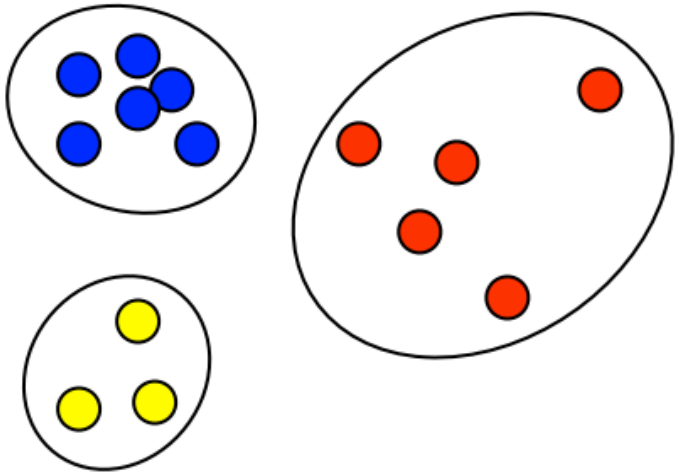


Summarizing/compressing the data

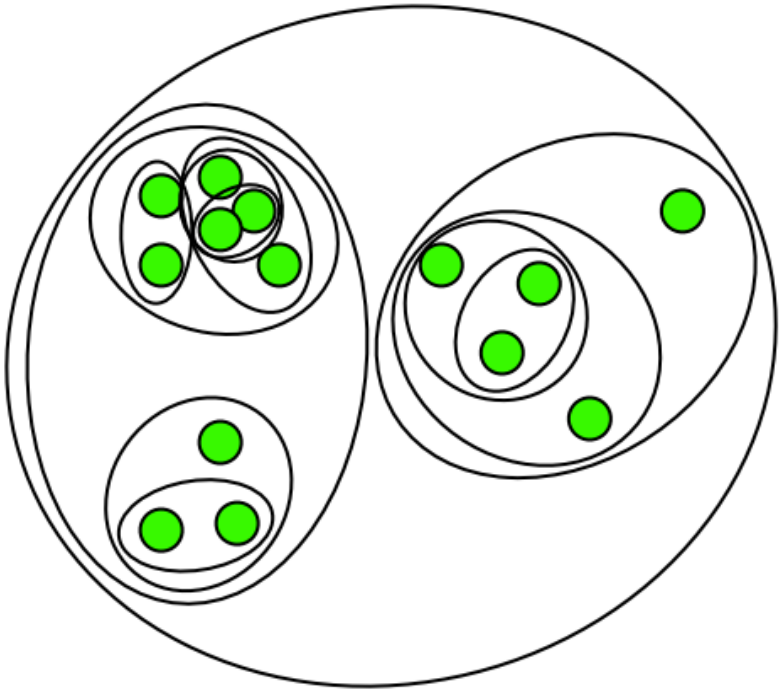




Partitional vs Hierarchical

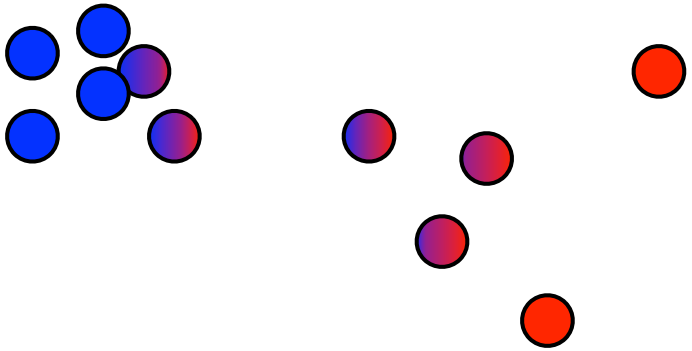


Each sample(point) is assigned to a unique cluster

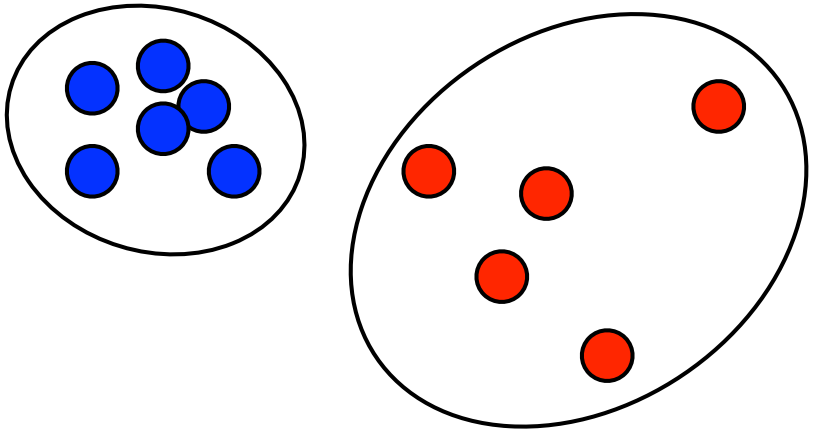


Creates a nested and hierarchical set of partitions/clusters

Fuzzy vs Non-Fuzzy

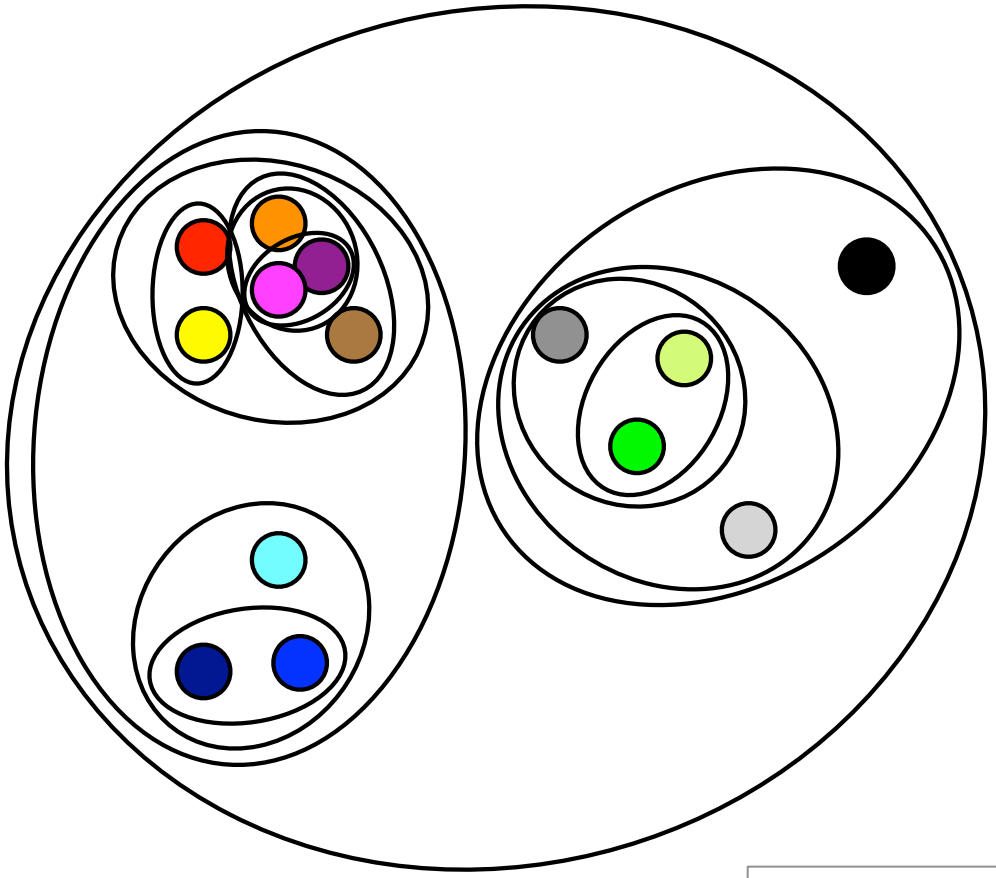


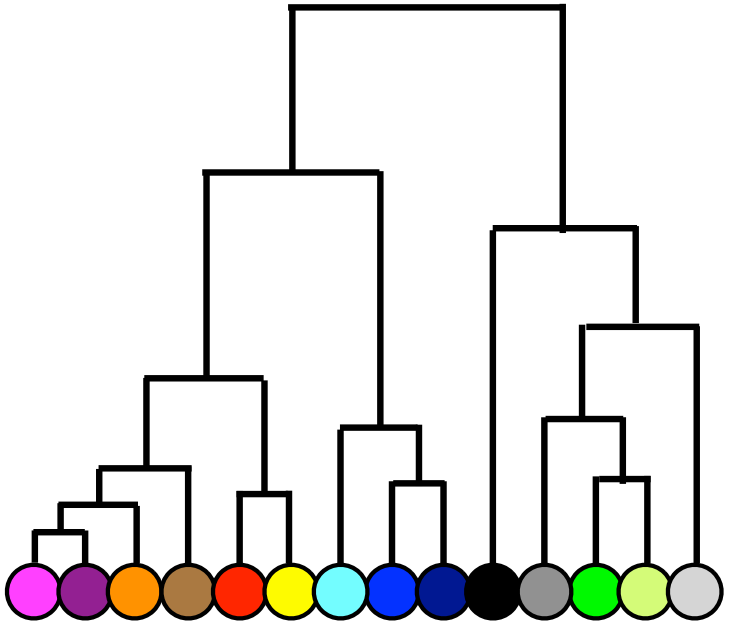
Each object belongs to each cluster with some weight



Each object belongs to exactly one cluster

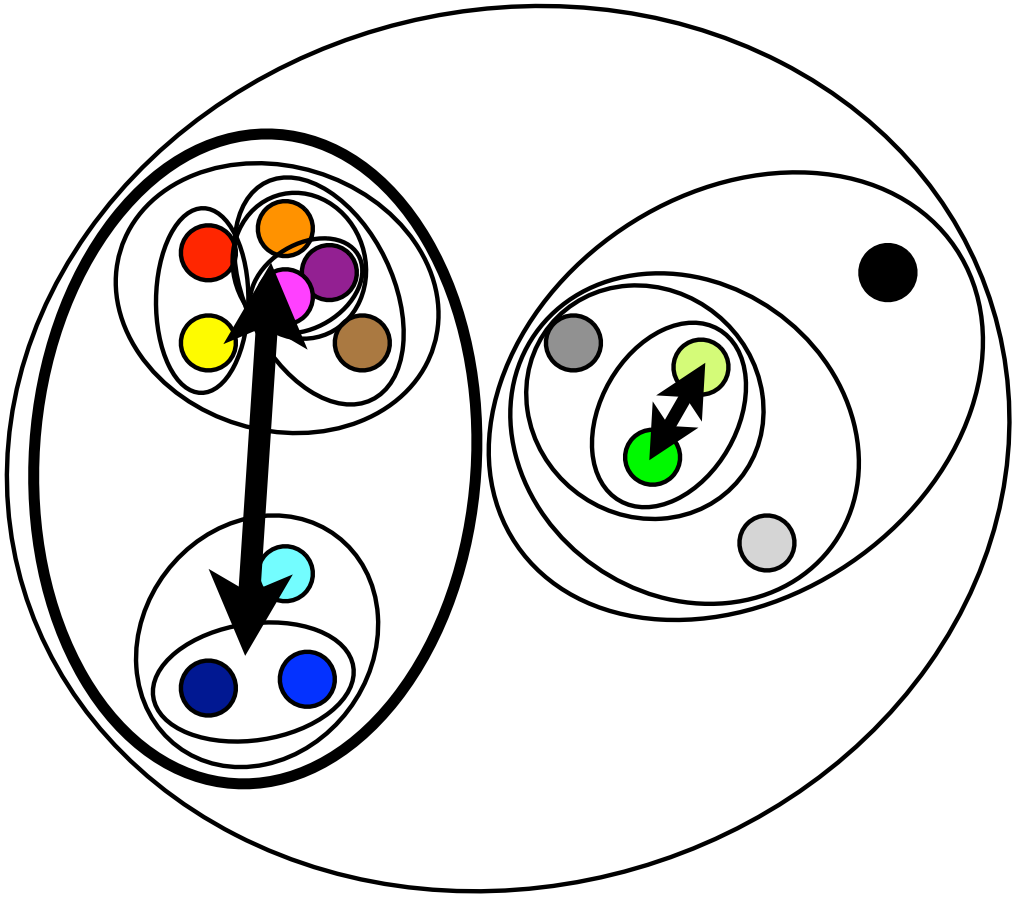
Hierarchical clustering

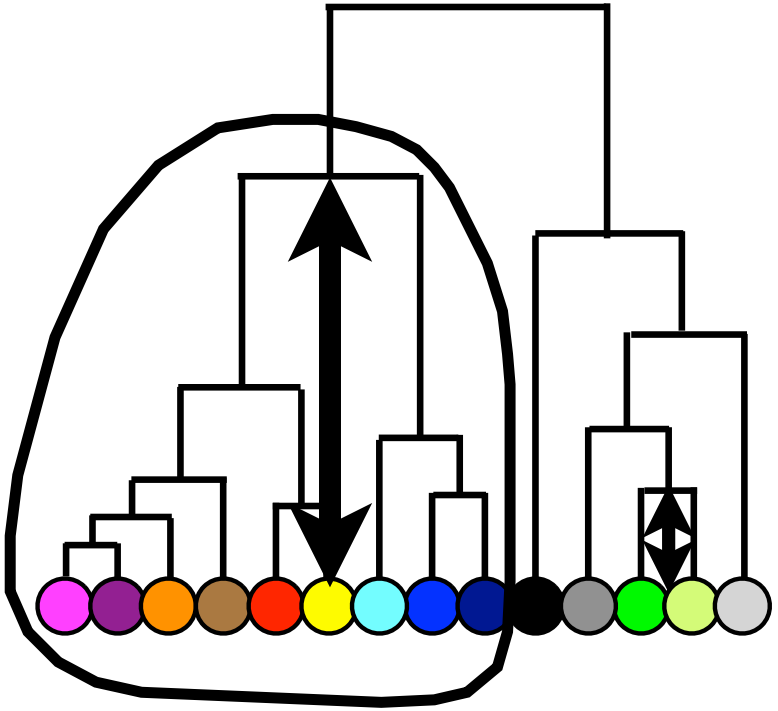




Hierarchical clustering is usually depicted as a dendrogram (tree)

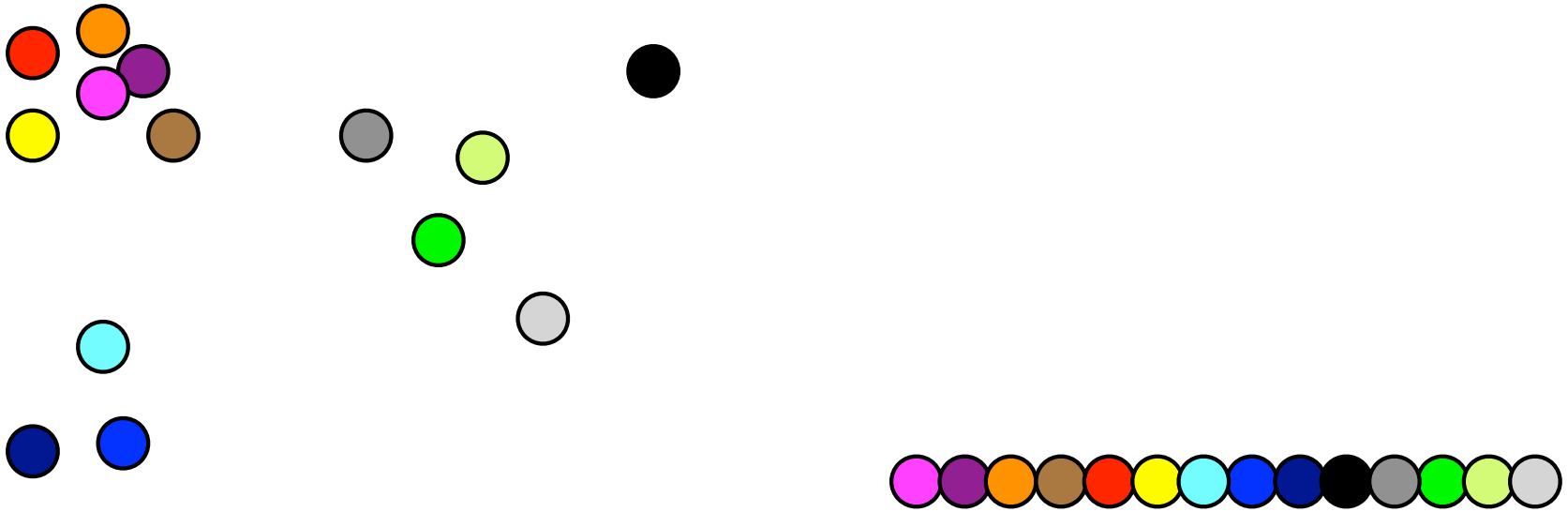
Hierarchical clustering





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

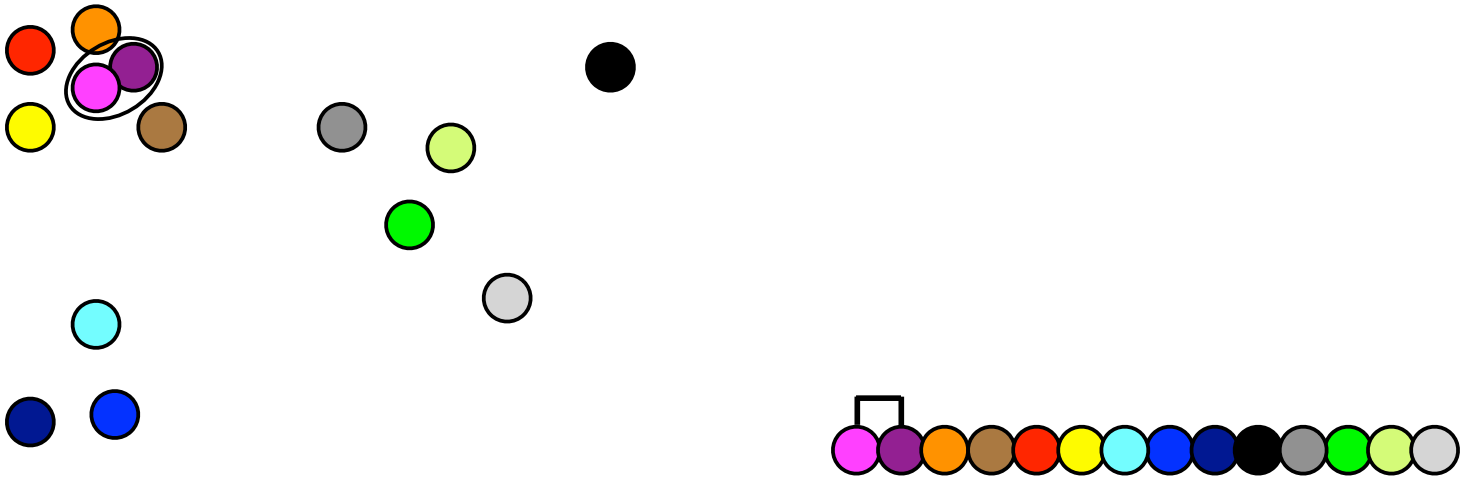
Hierarchical clustering



Algorithm for Agglomerative Hierarchical Clustering:

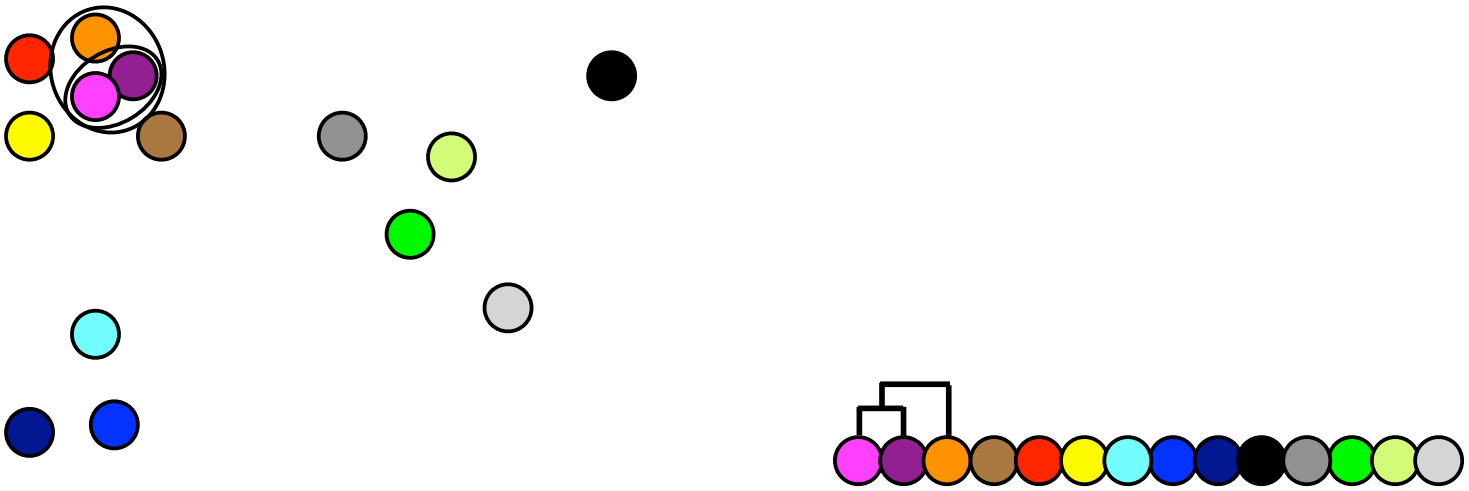
Join the two closest objects

Hierarchical clustering



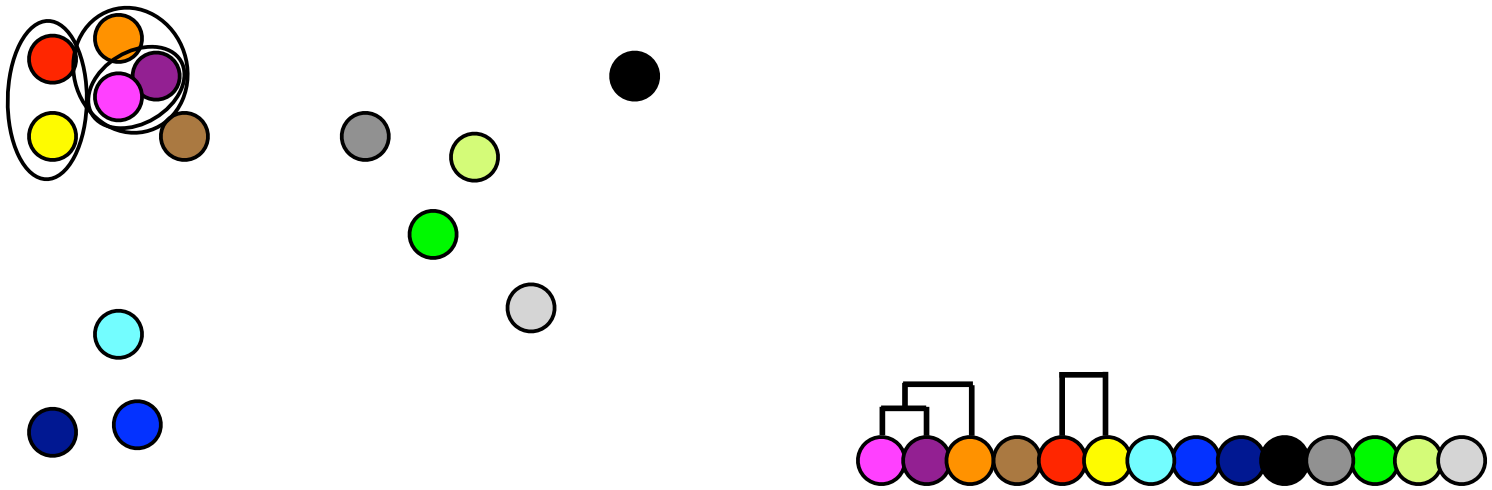
Join the two closest objects

Hierarchical clustering



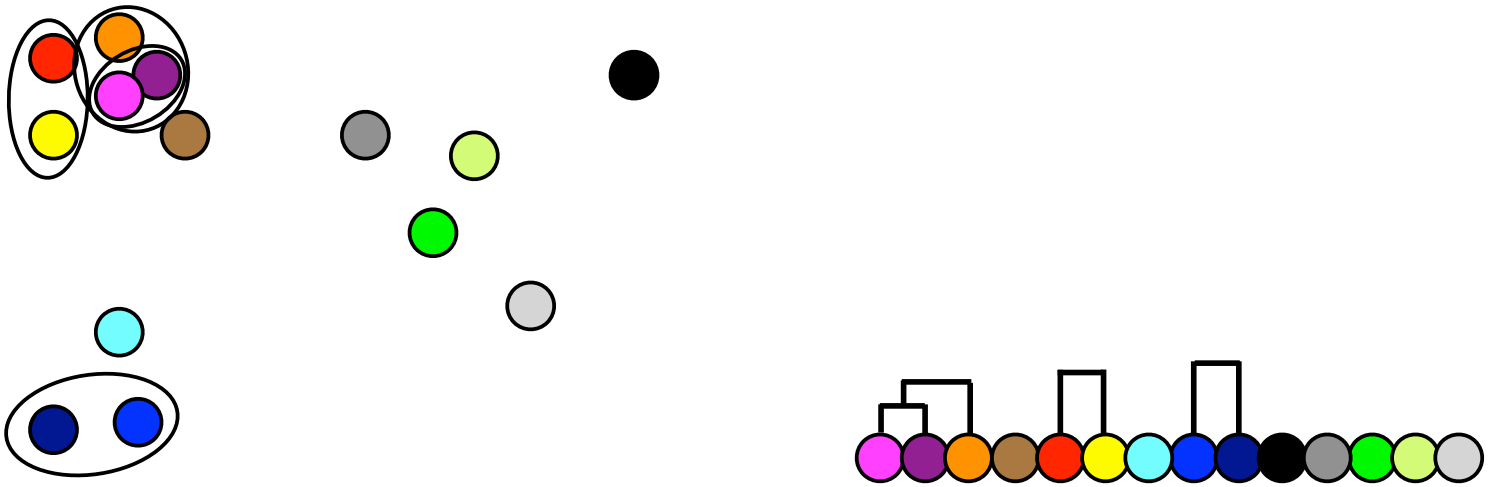
Keep joining the closest pairs

Hierarchical clustering



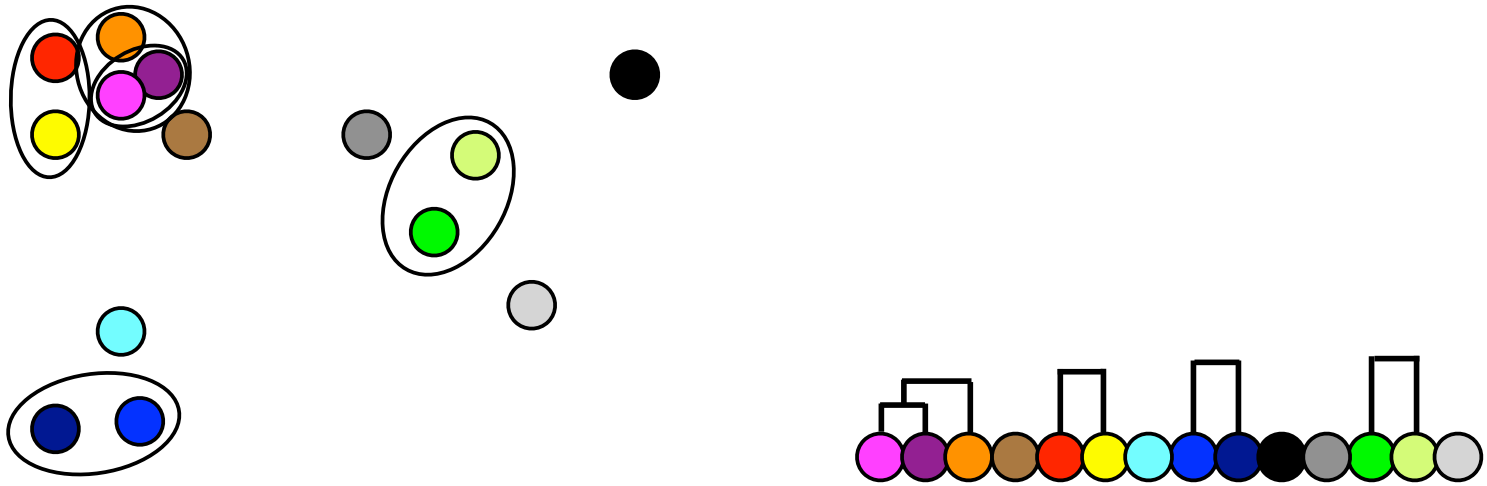
Keep joining the closest pairs

Hierarchical clustering



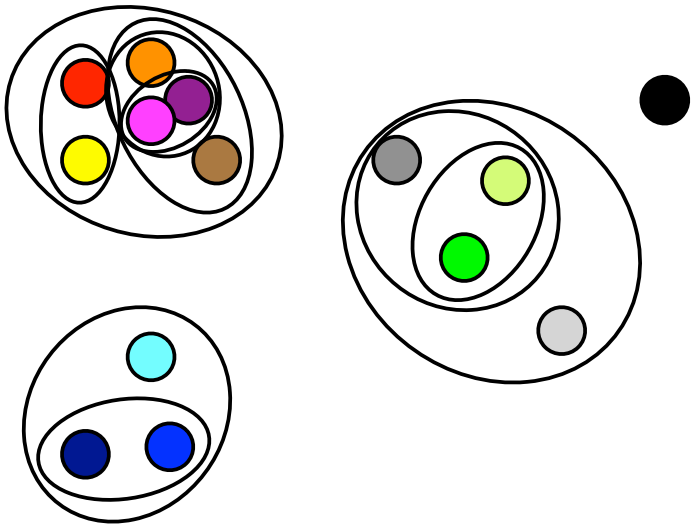
Keep joining the closest pairs

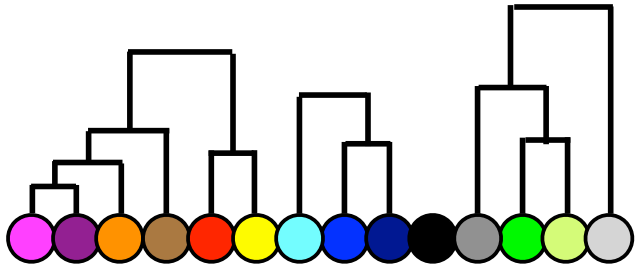
Hierarchical clustering



Keep joining the closest pairs

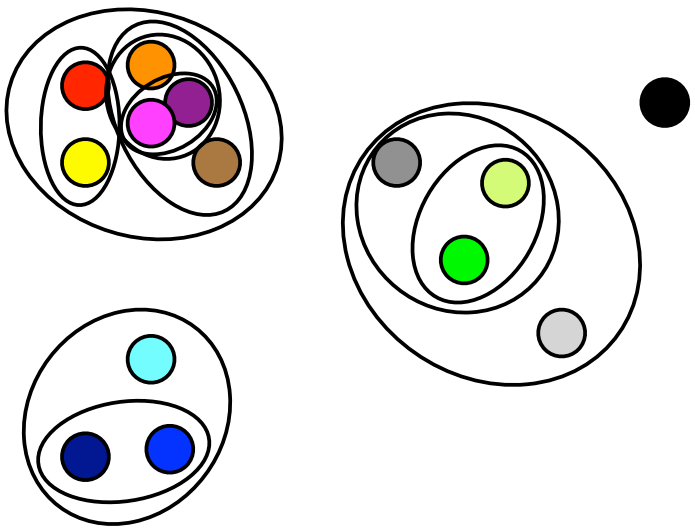
Hierarchical clustering

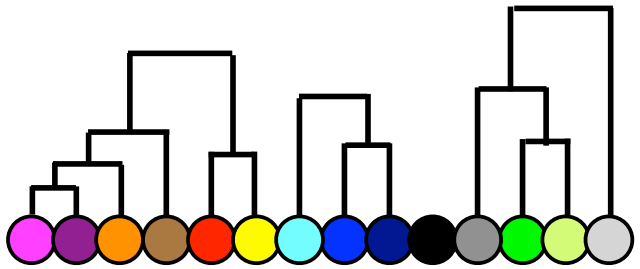




After 10 steps we have 4 clusters left

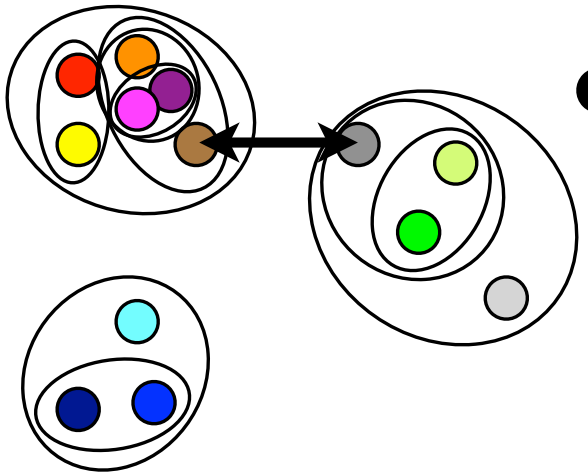
Q: Which clusters do we merge next?





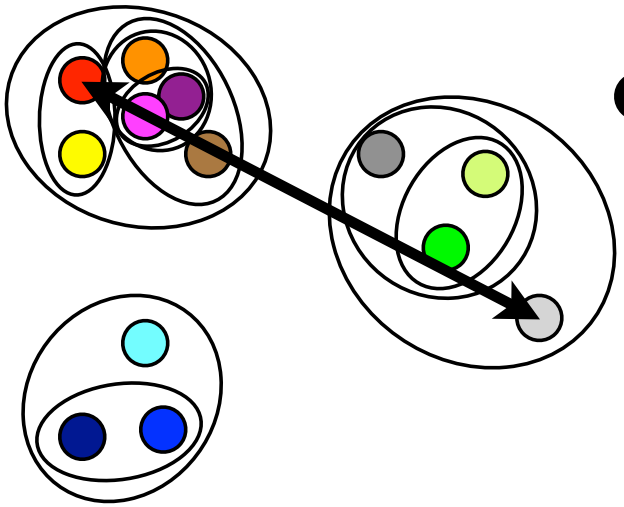
Hierarchical clustering

Several ways to measure distance between clusters:



- Single linkage(MIN)

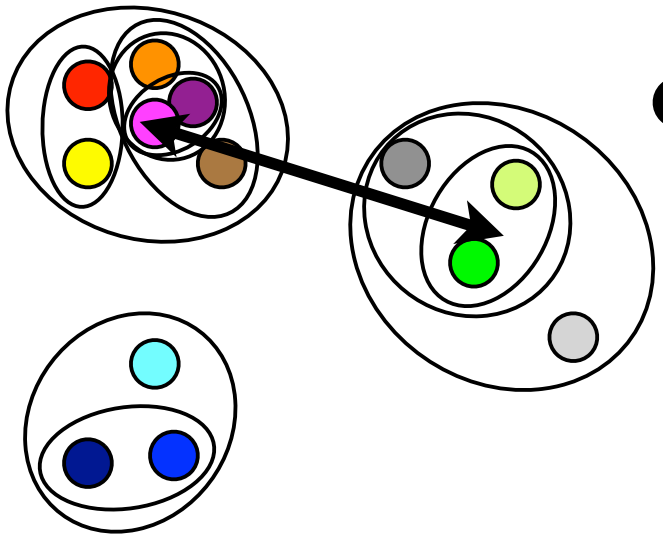
Hierarchical clustering



Several ways to measure distance between clusters:

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

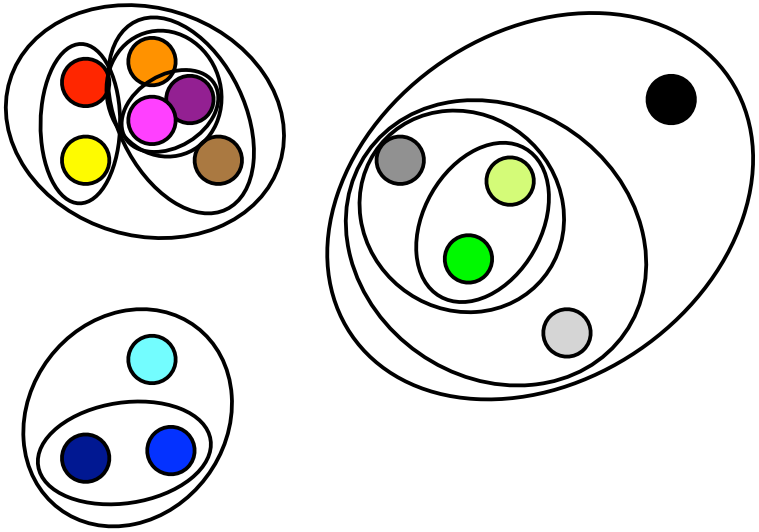
Hierarchical clustering

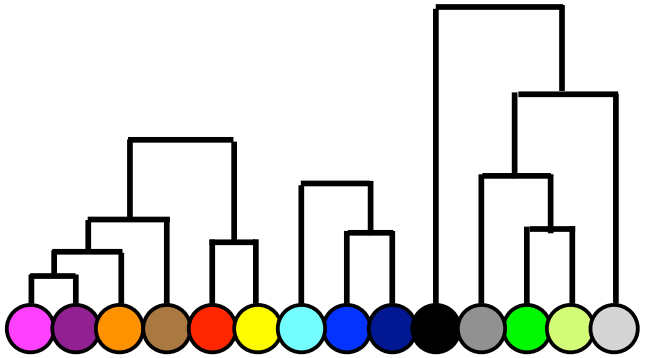


Several ways to measure distance between clusters:

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

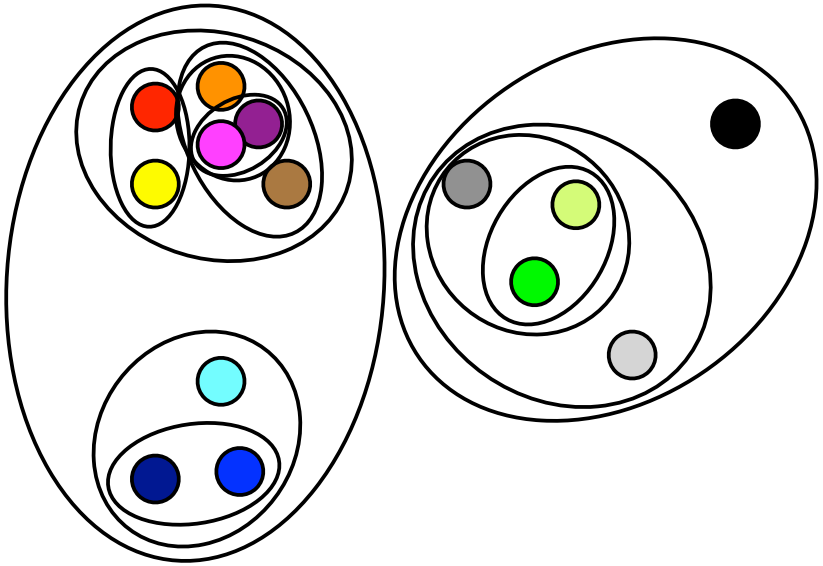
Hierarchical clustering

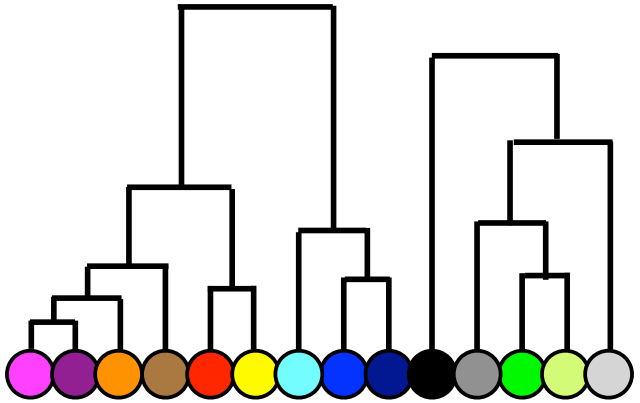




In this example and at this stage we have the same result as in partitional clustering

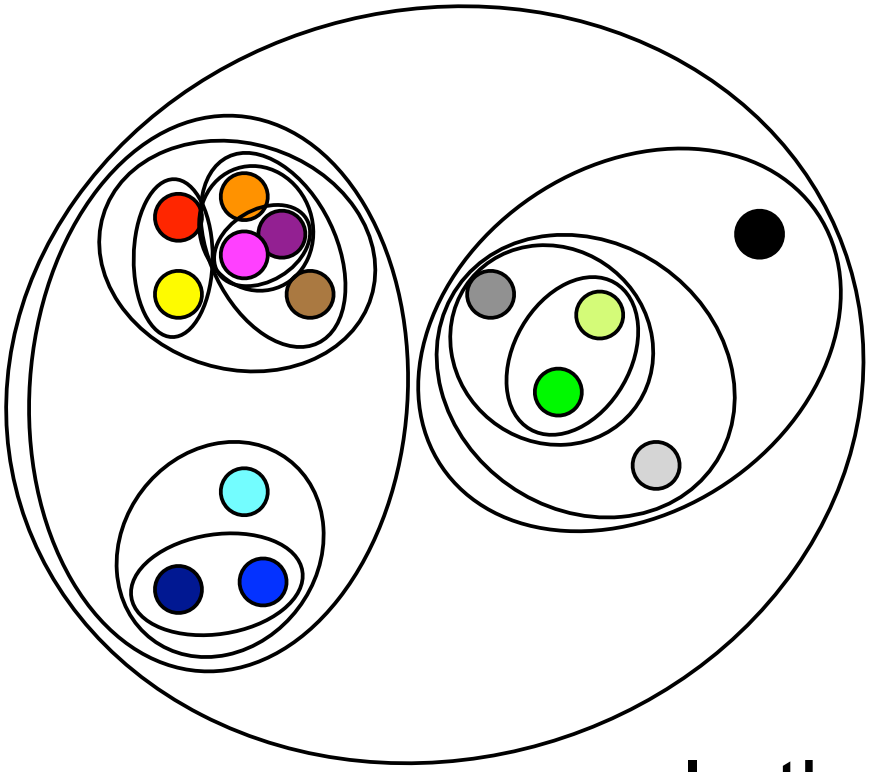
Hierarchical clustering

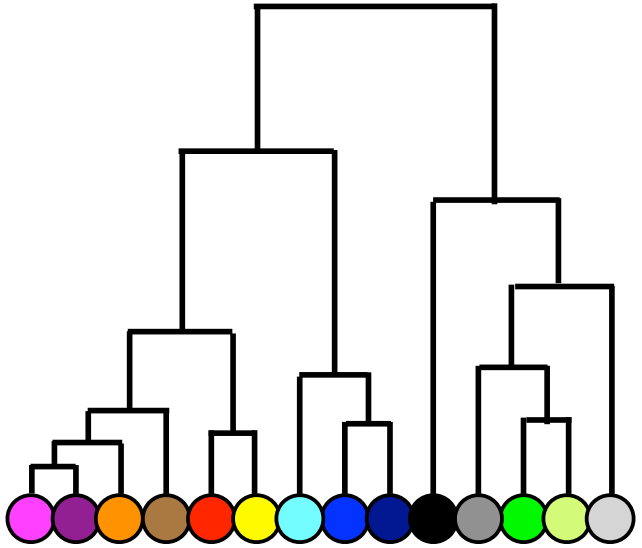




In the final step the two remaining clusters are joined into a single cluster

Hierarchical clustering

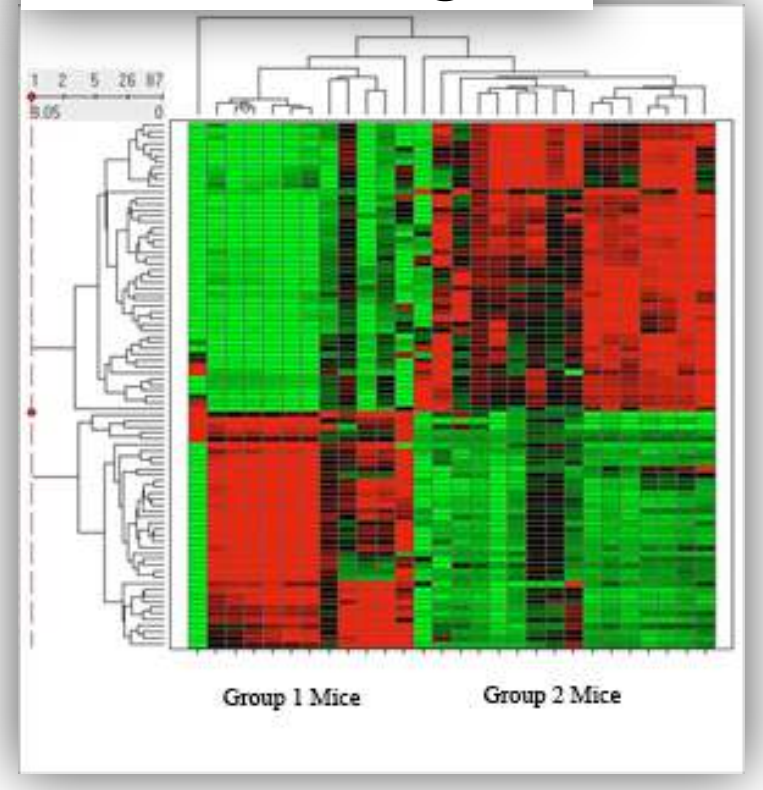




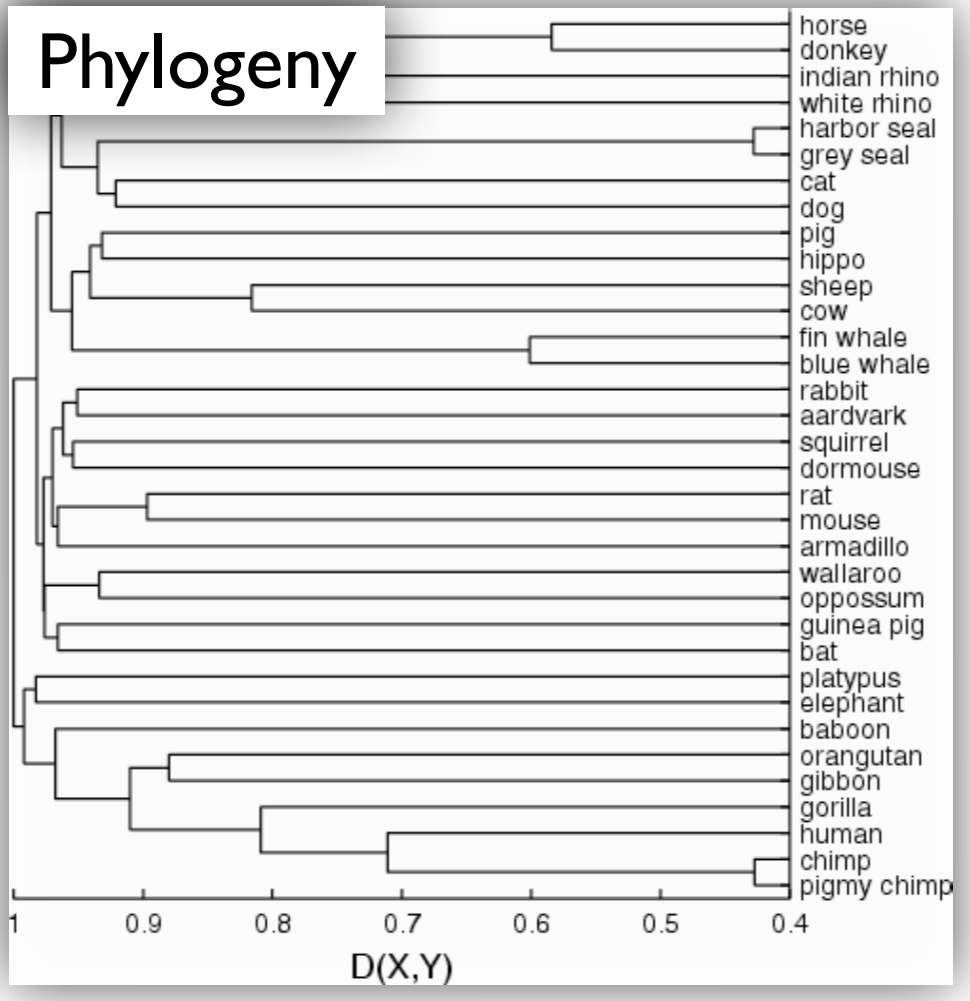
In the final step the two remaining clusters are joined into a single cluster

Examples of Hierarchical Clustering in Bioinformatics

Gene expression clustering



Phylogeny



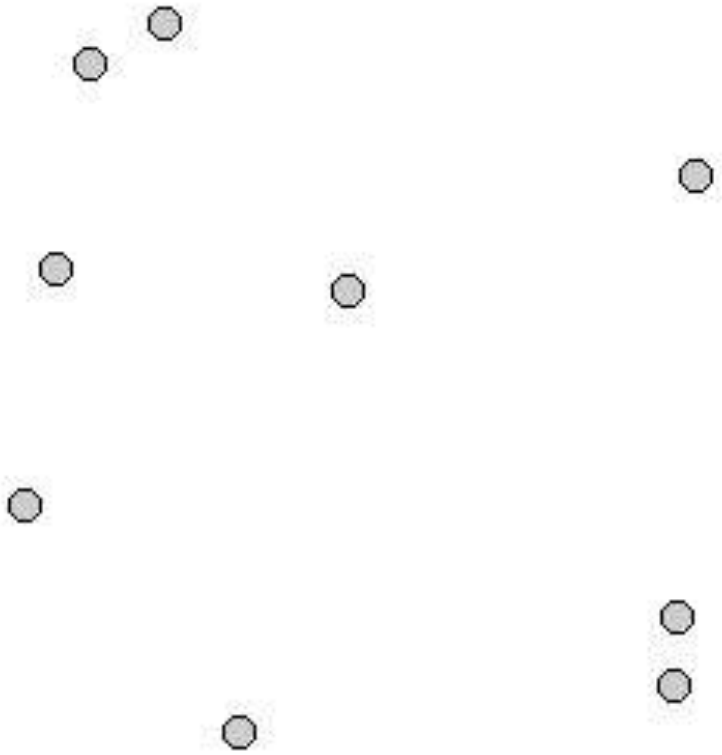
K-means clustering

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

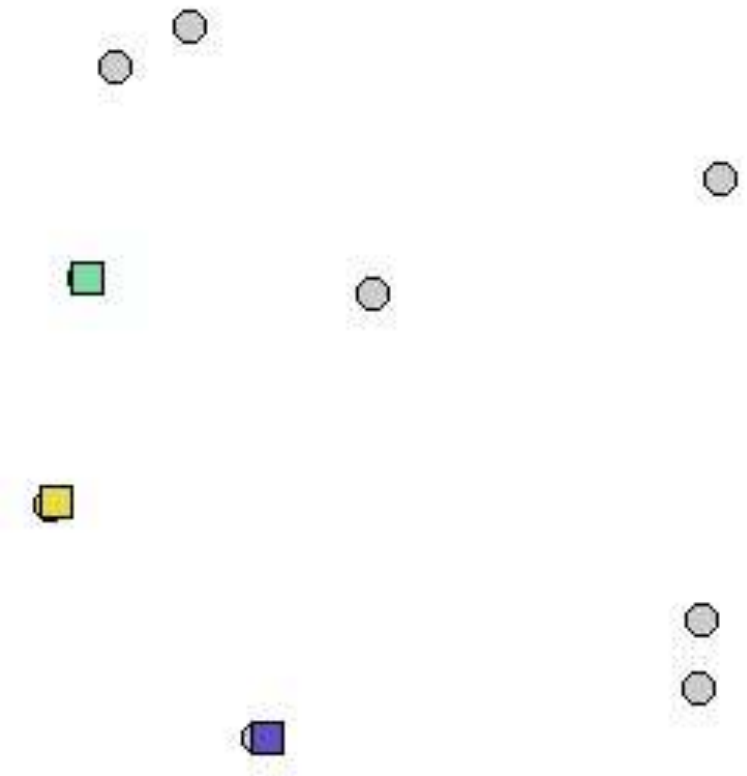
Algorithm:

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

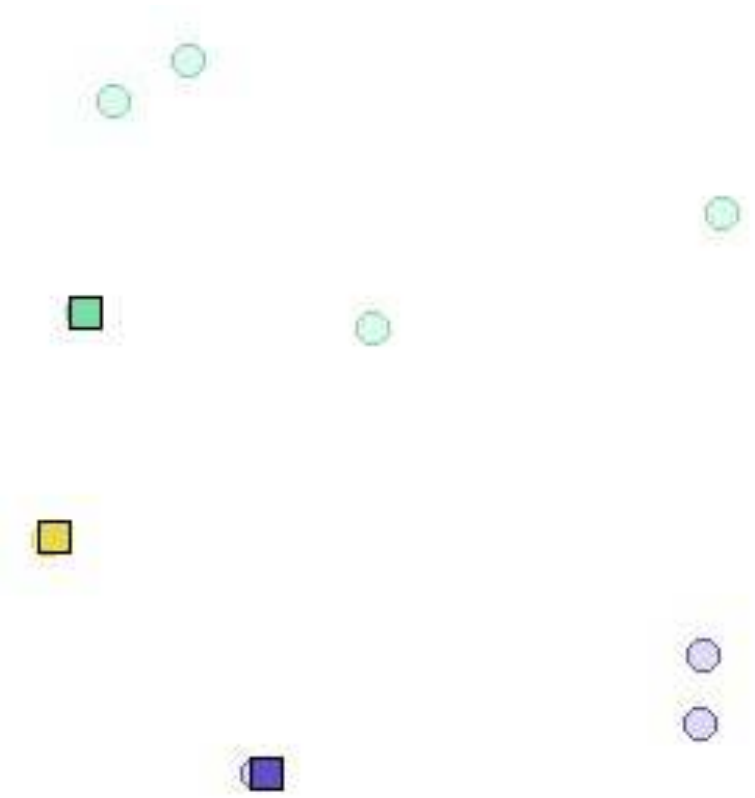
K-means (1)



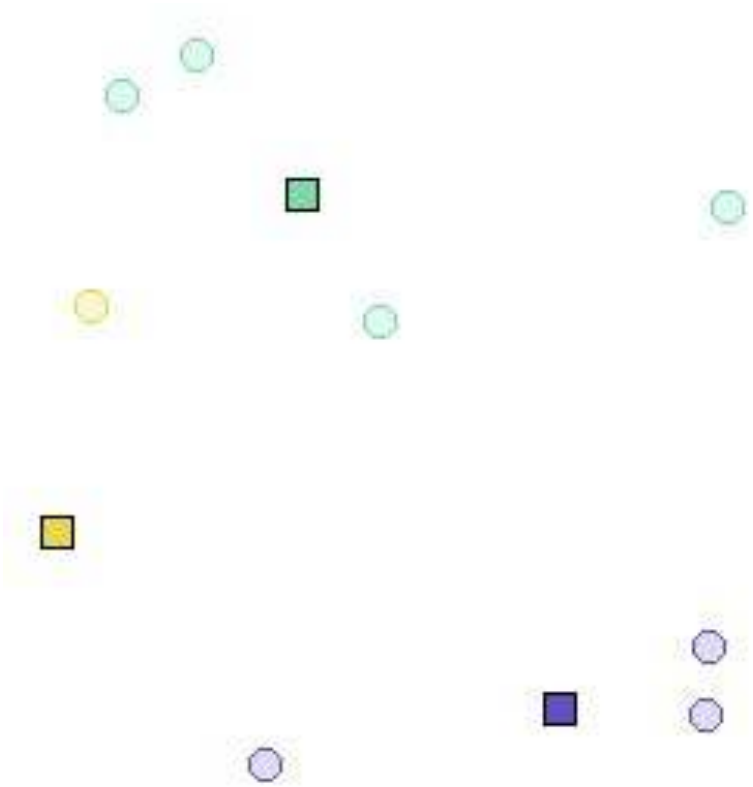
K-means (2)



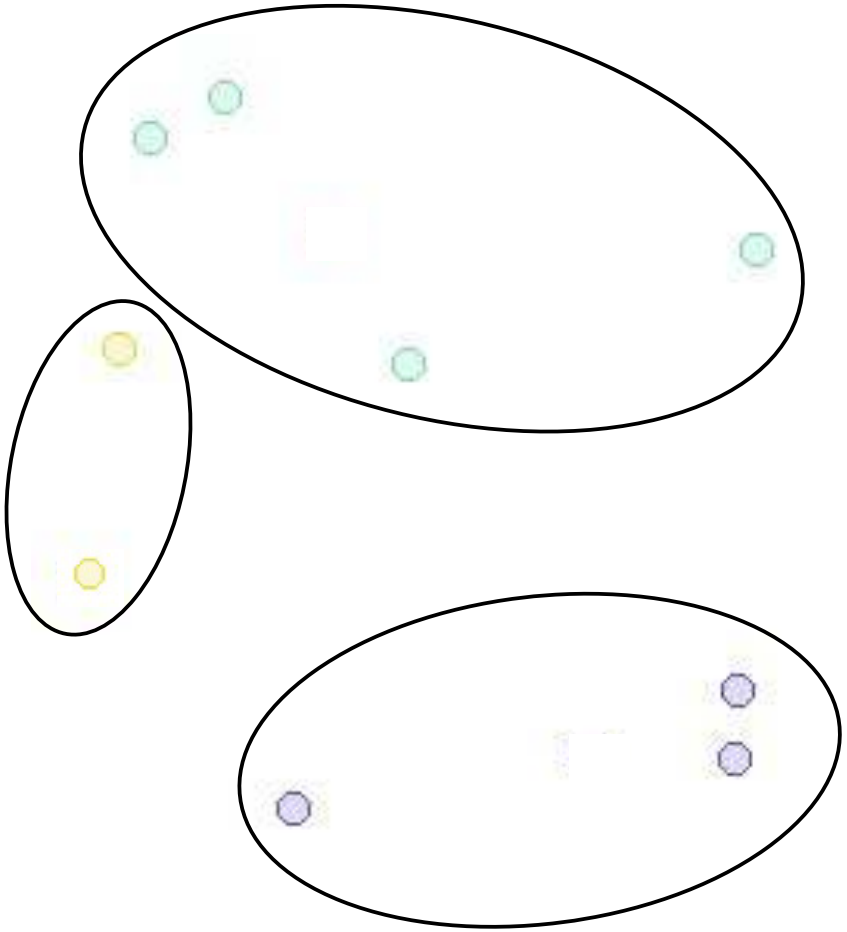
K-means (3)



K-means (4)

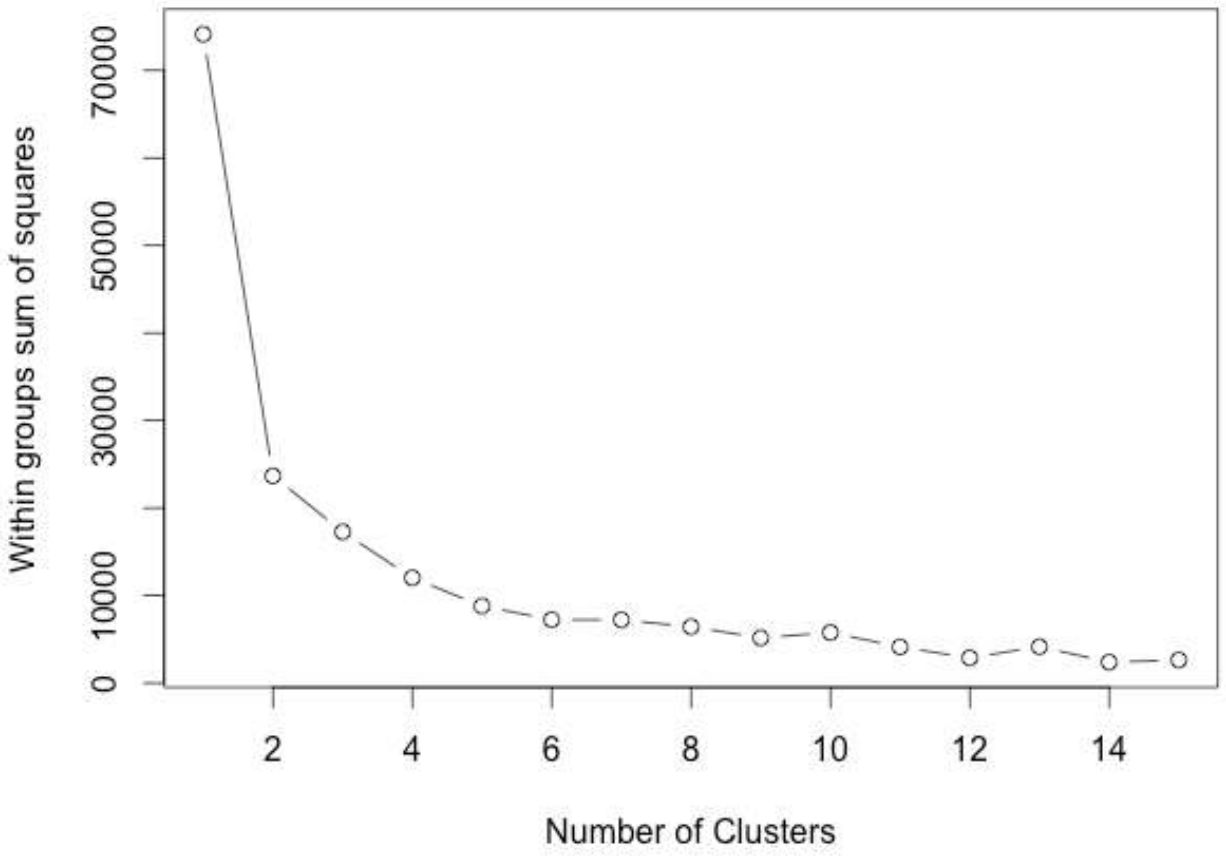


K-means (5)



Elbow method

Estimate the number of clusters



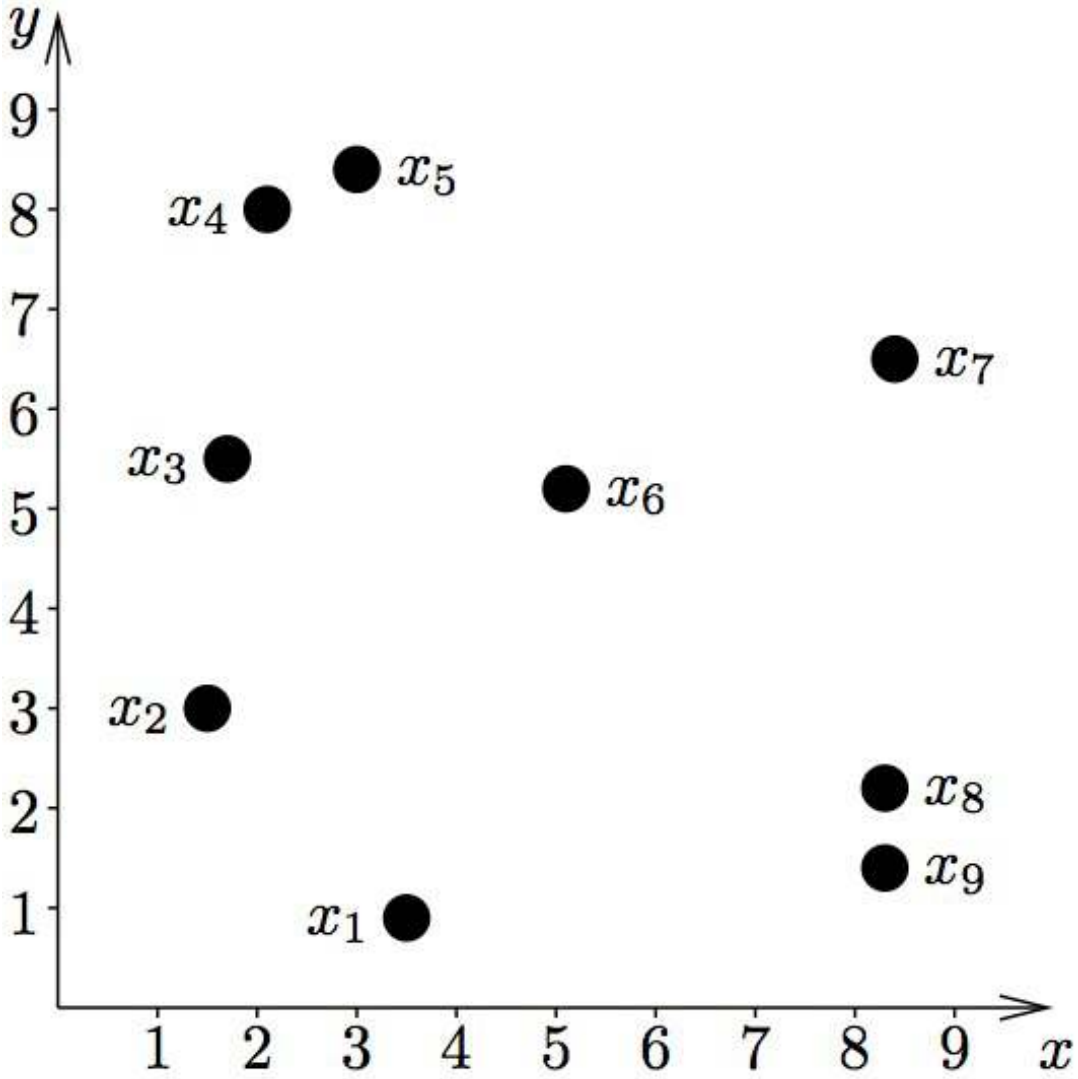
K-means clustering summary

- 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?)

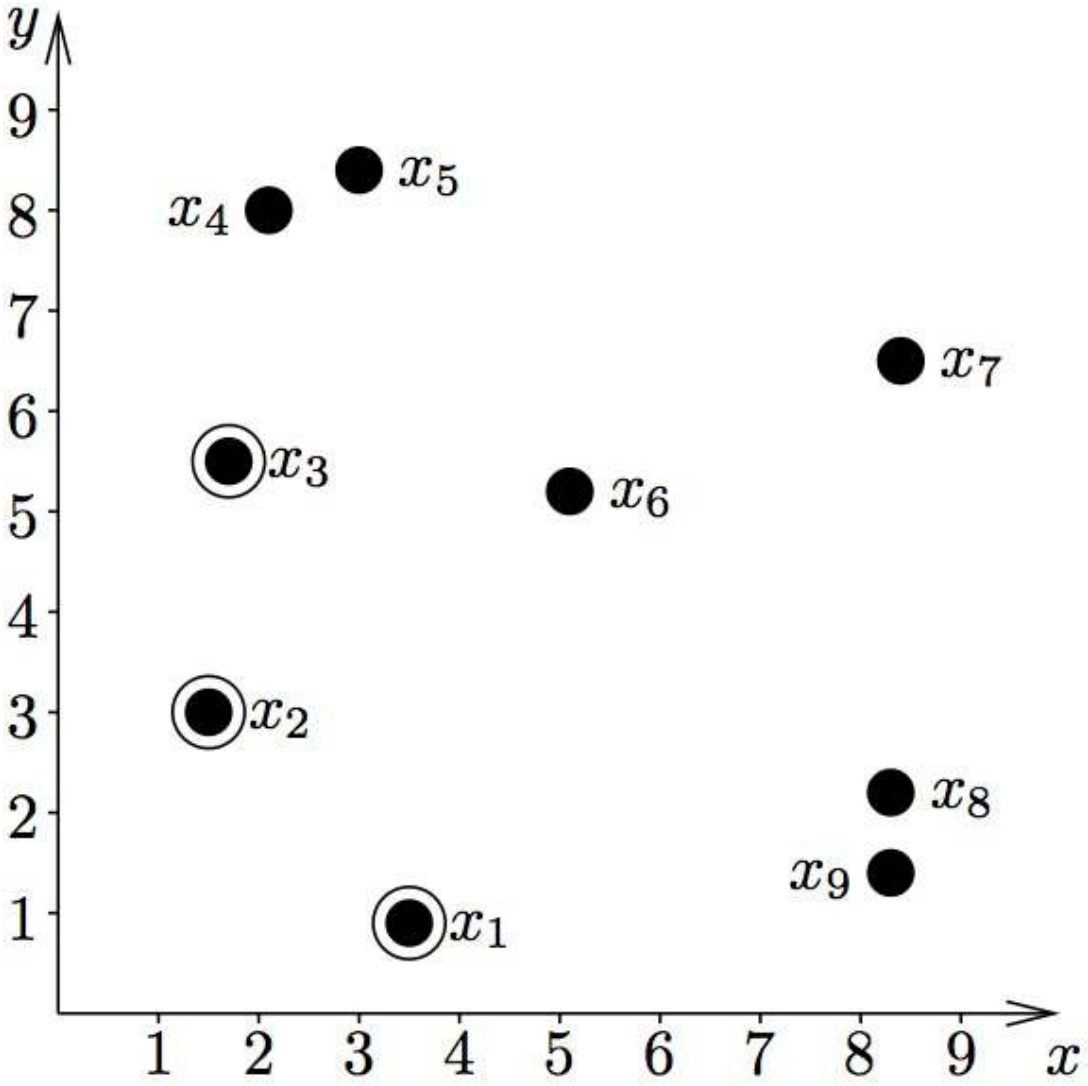
K-medoids clustering

- 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

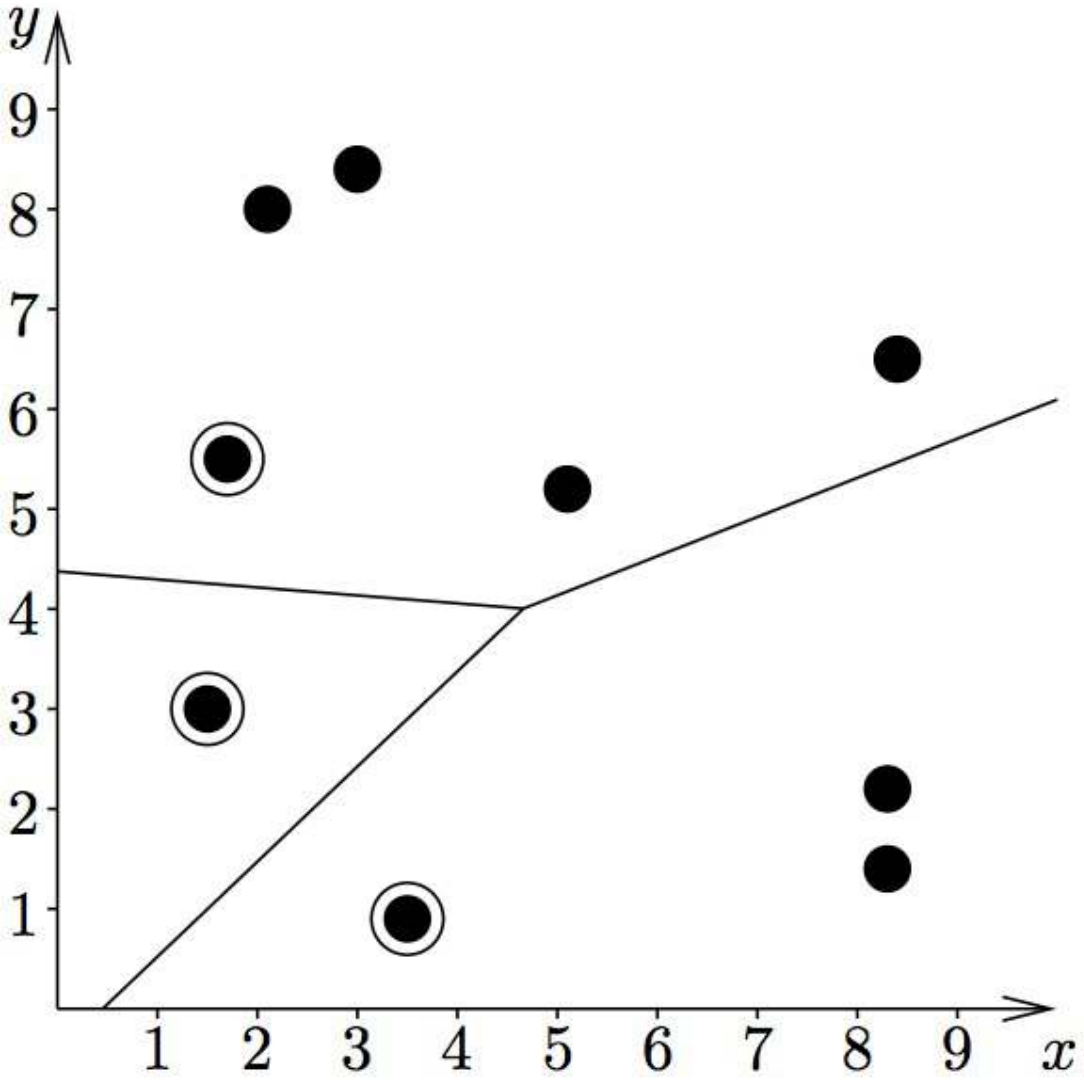
K-medoids (1)



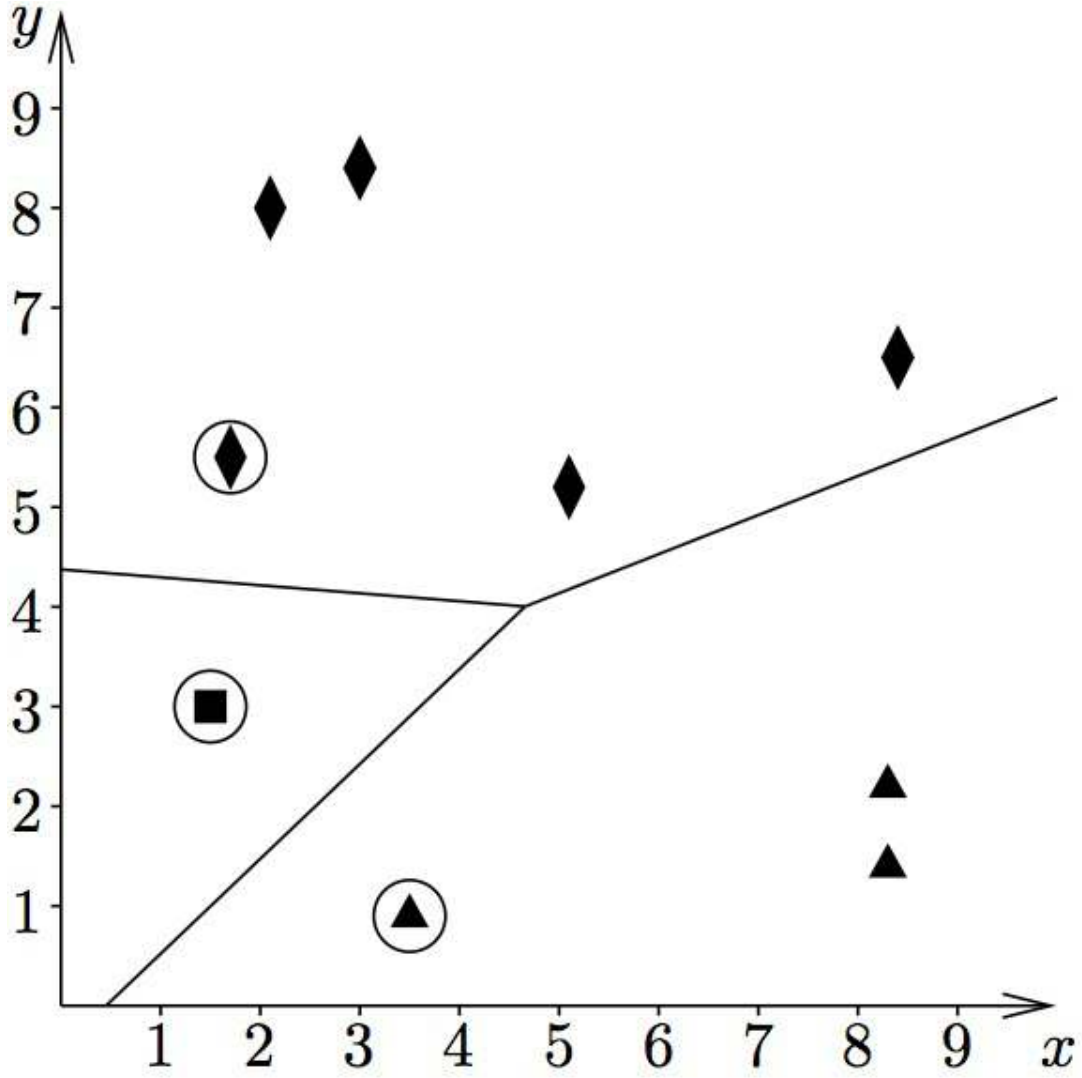
K-medoids (2)



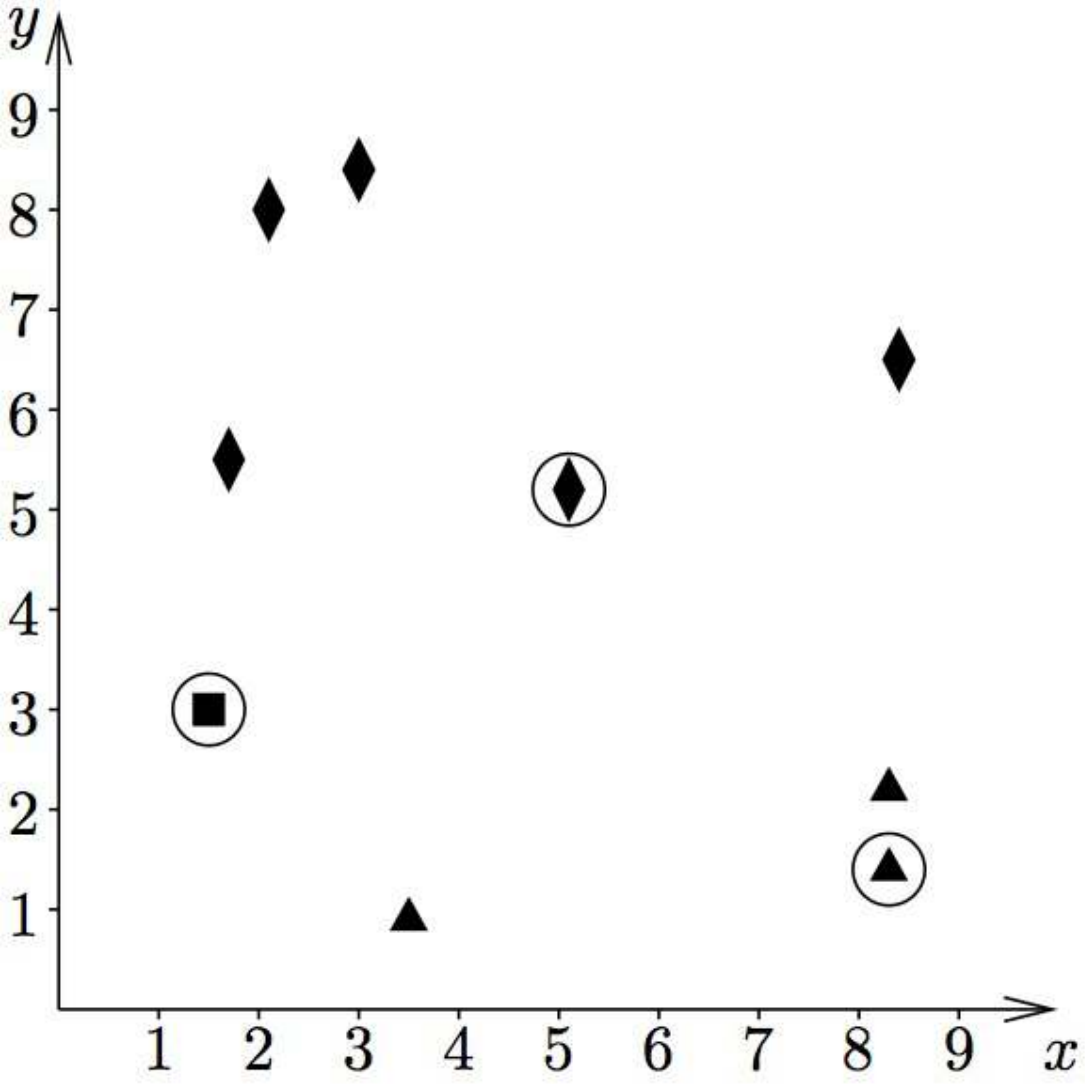
K-medoids (3)



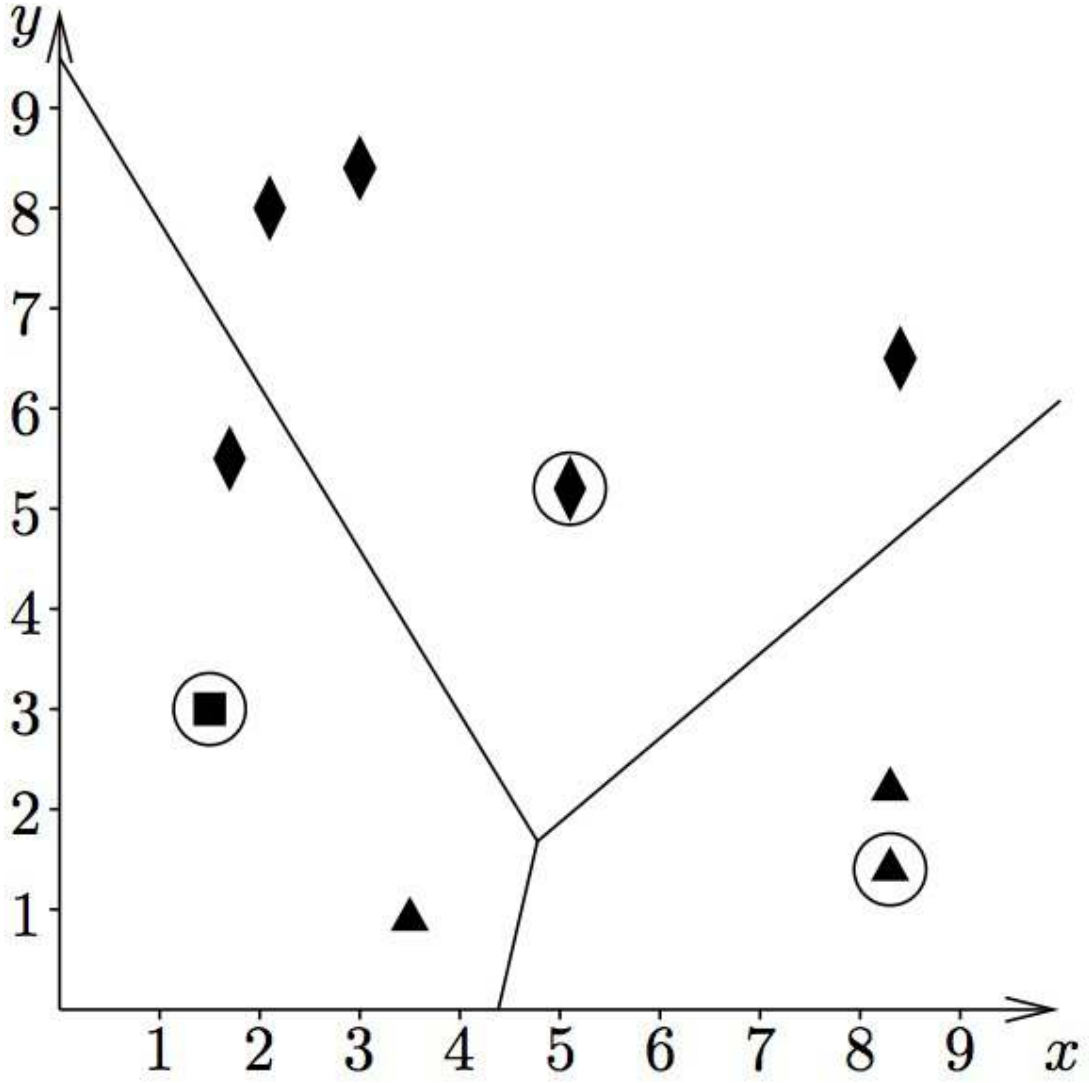
K-medoids (4)



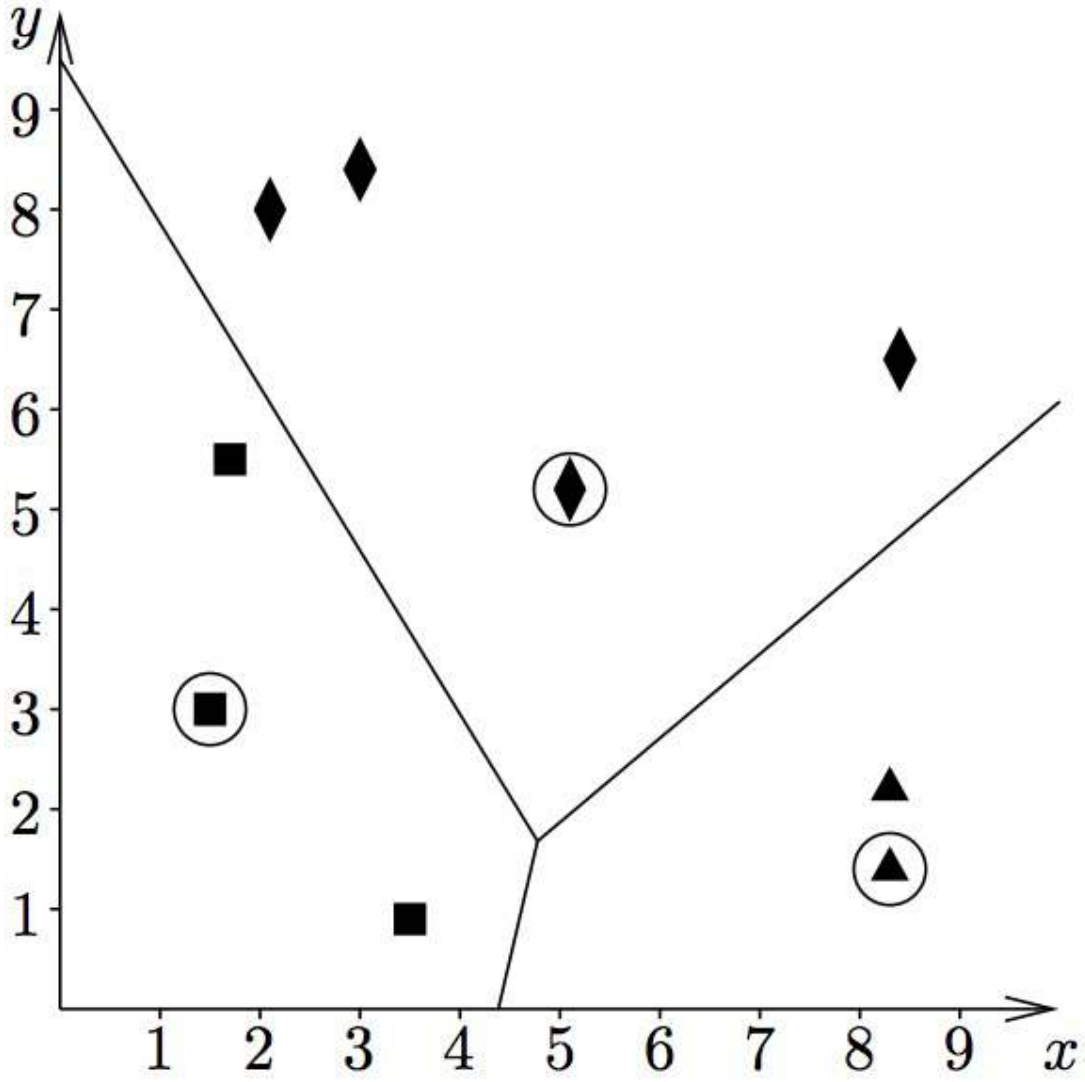
K-medoids (5)



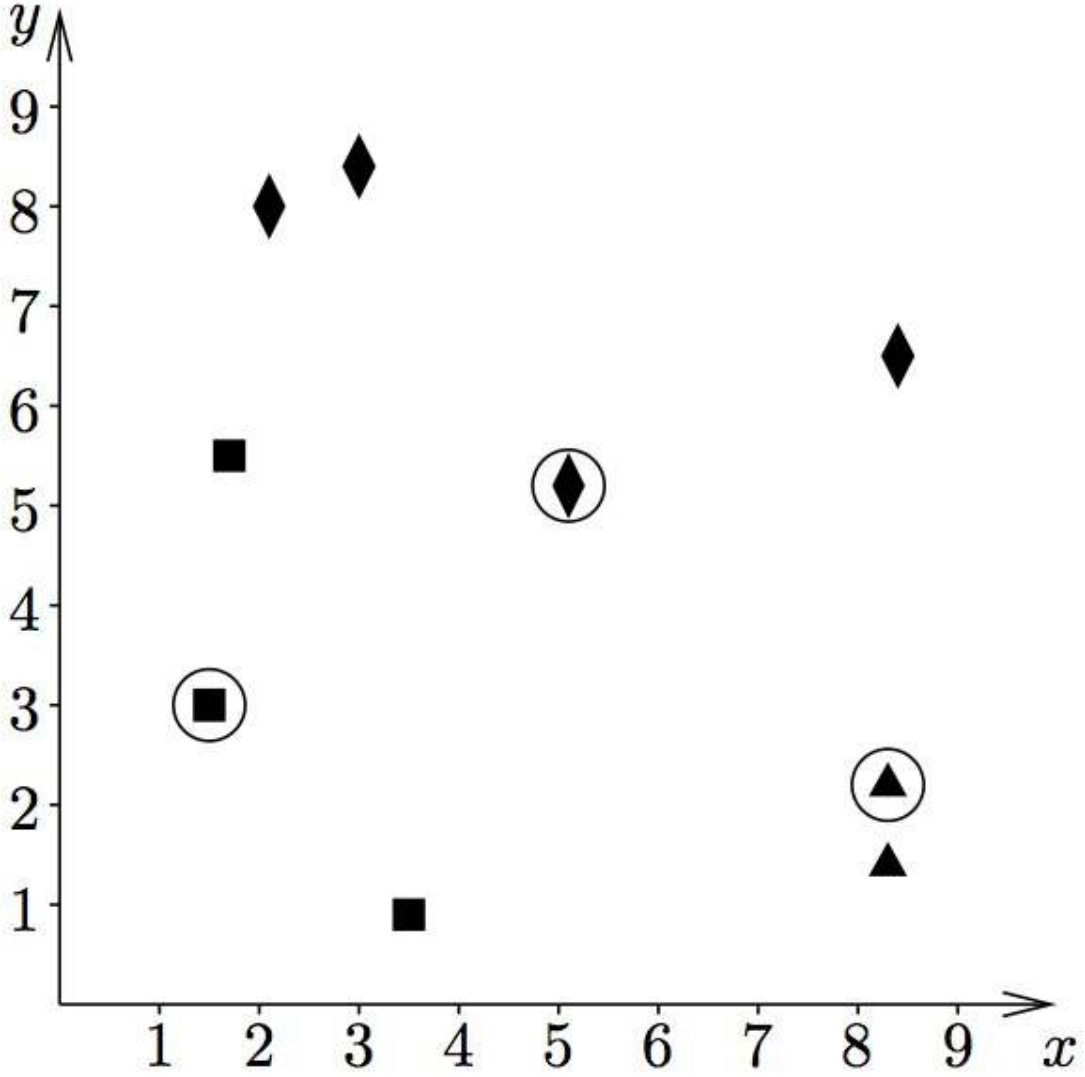
K-medoids (6)



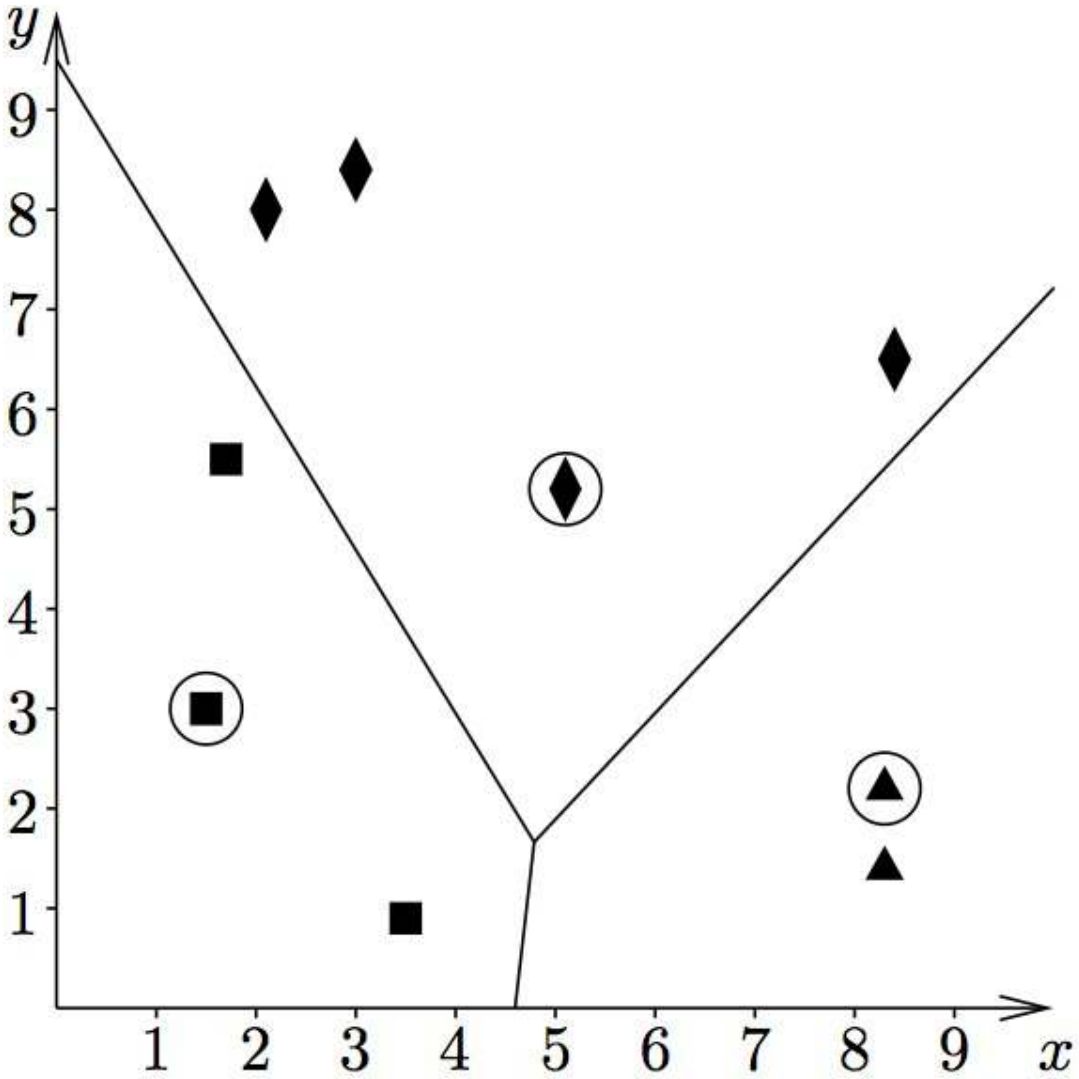
K-medoids (7)



K-medoids (8)

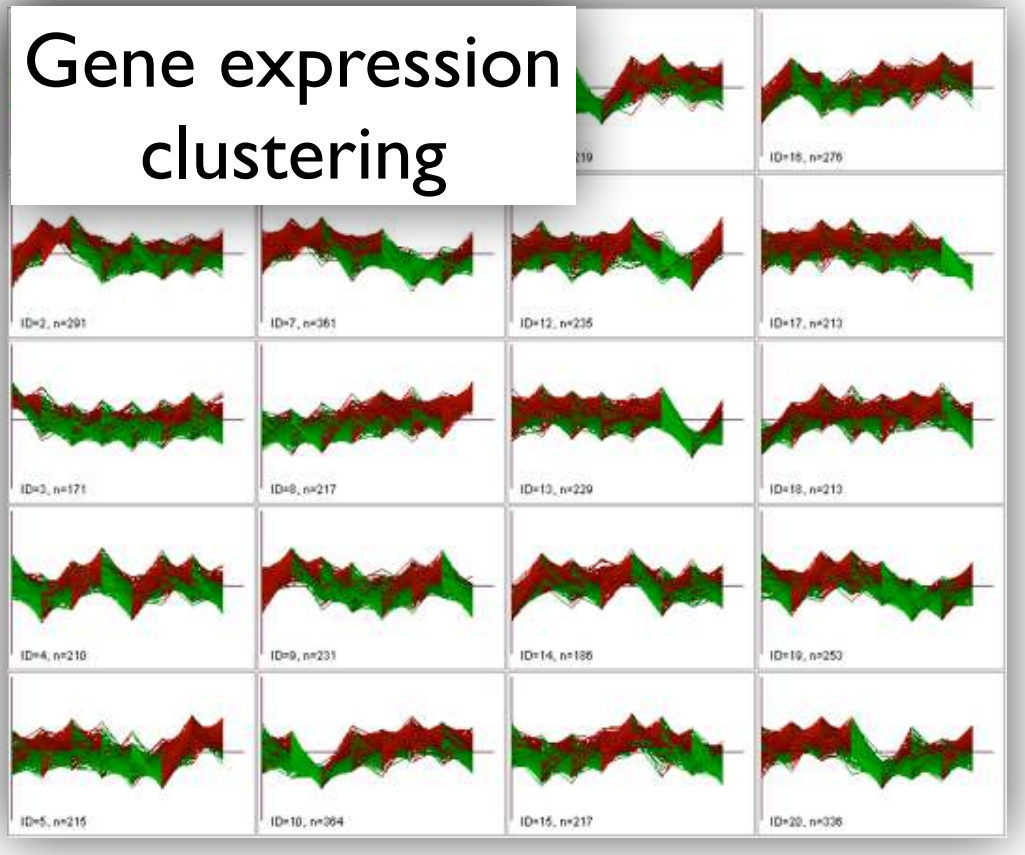


K-medoids (9)

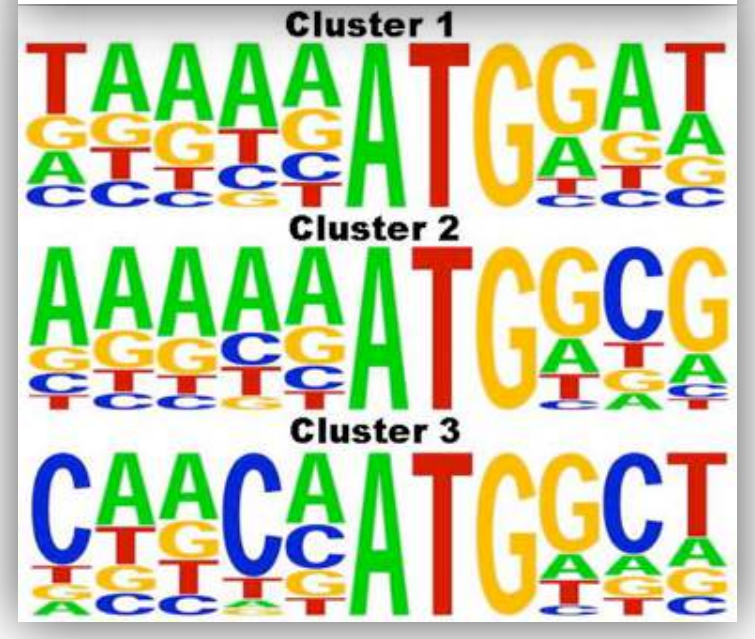


Examples of K-means and K-medoids in Bioinformatics

Gene expression clustering



Sequence clustering



Distance measures

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

- 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

```
ACCTTG
TACCTG
```
=> 3

- Levenshtein distance

```
.ACCTTG
TACC.TG
```
=> 2


+1!
LEVEL
UP

Interpretation validation

Import data

Summarize/ plot raw data

Impute missing values

Normalize/ Standardize

Handle outliers

Data analysis

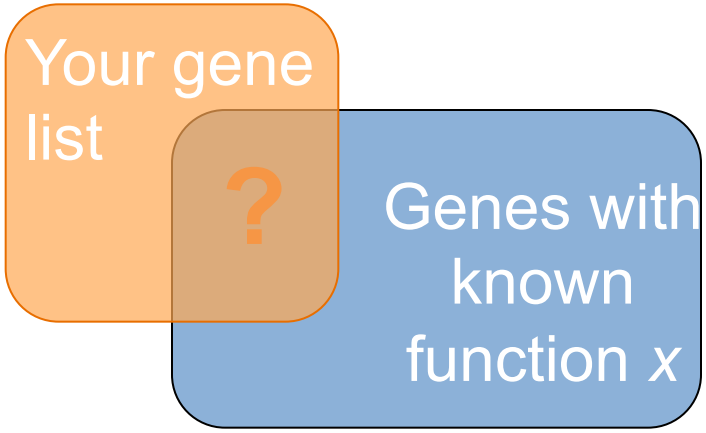
Put it into words & Discover

Gene ontology

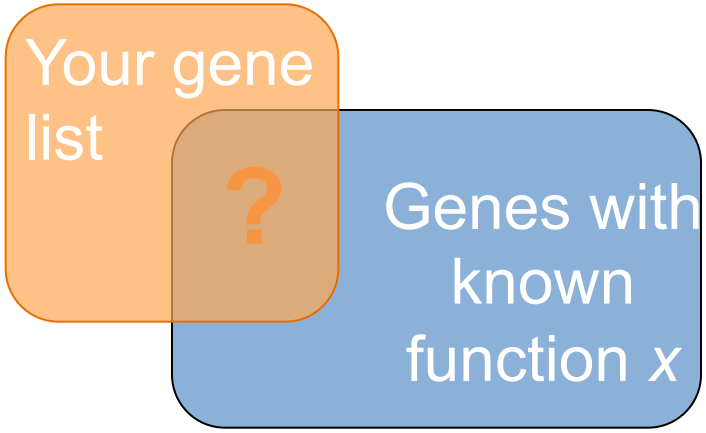
What found genes are doing

- Molecular Function - elemental activity or task
- Biological Process - broad objective or goal
- Cellular Component - location or complex

Functional enrichment statistics

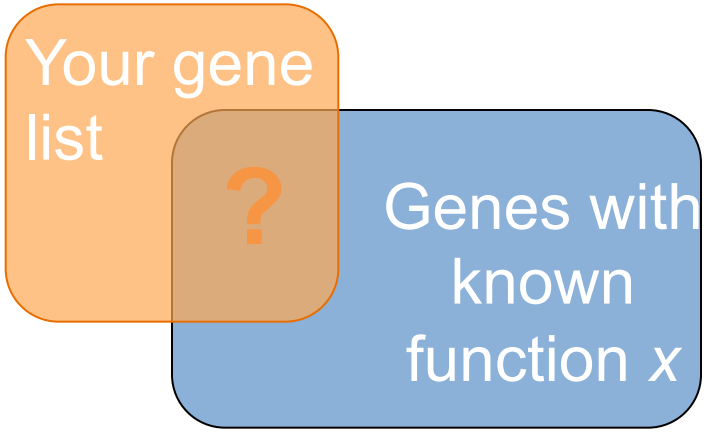


Functional enrichment statistics



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?

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

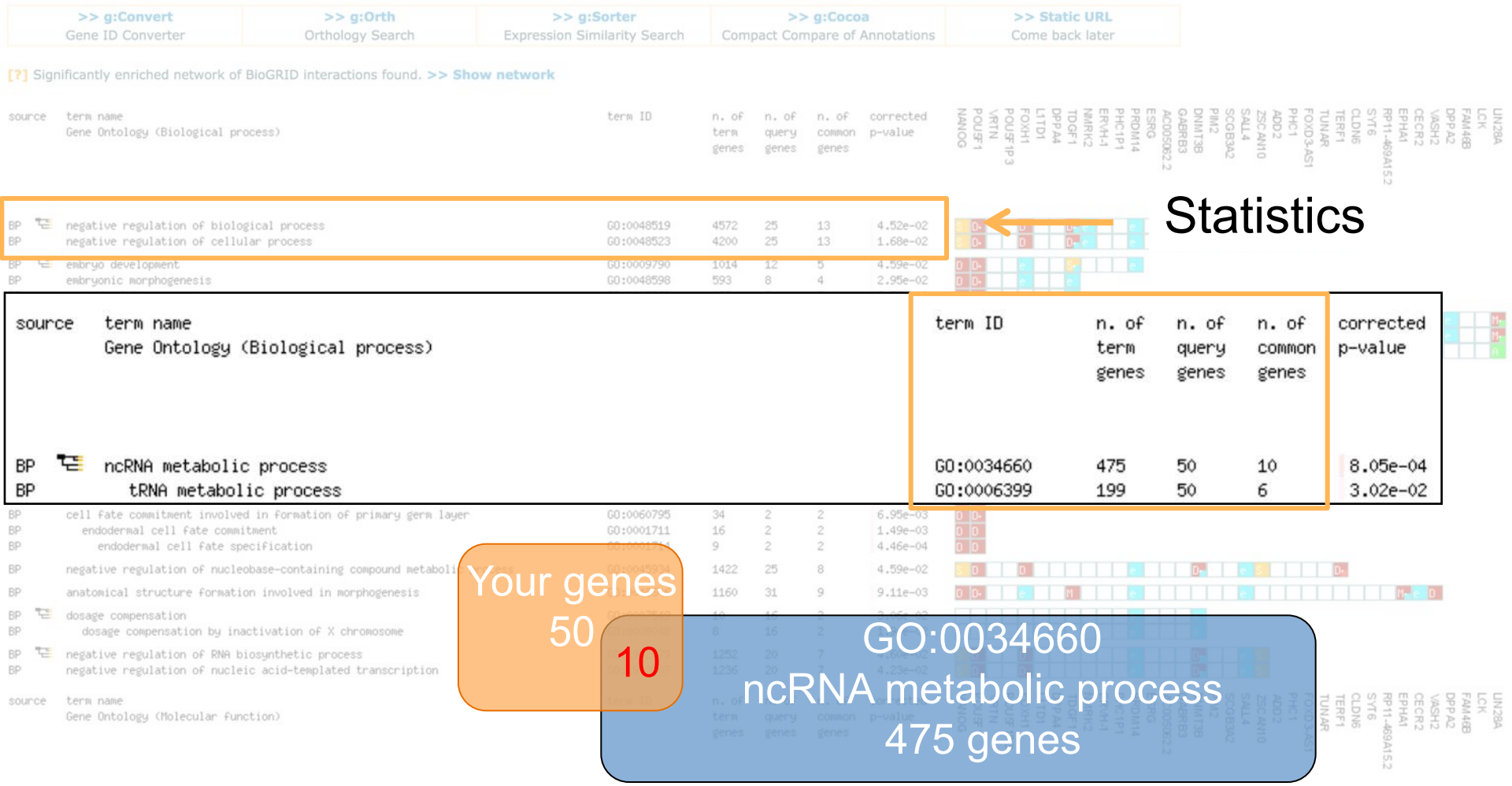
g:Profiler toolset
http://biit.cs.ut.ee/gprofiler

Welcome!	Contact	FAQ	R / APIs	Beta	Archive

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

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

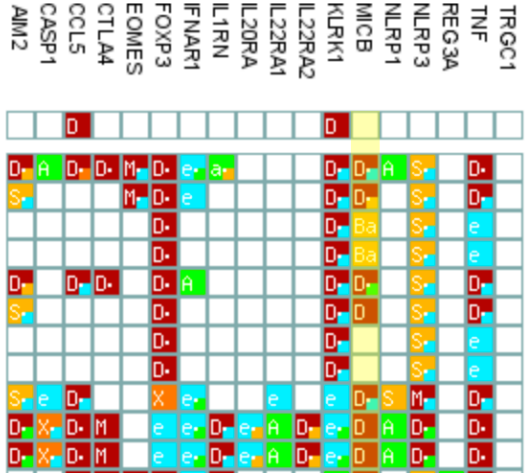


Functional annotations & Significance

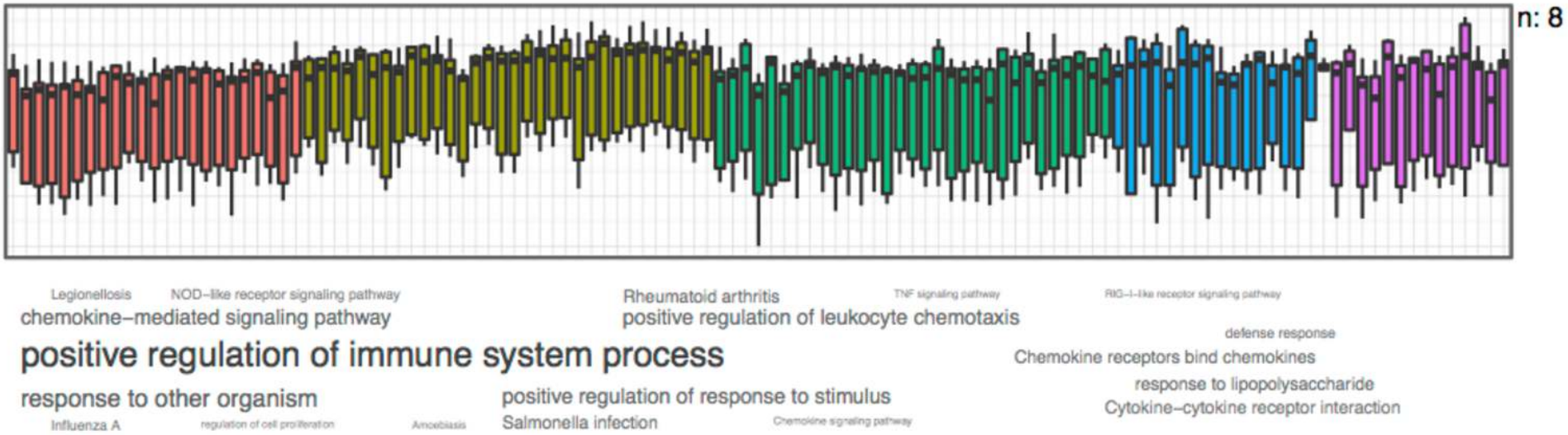
$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

statistical significance of having drawn a sample consisting of a specific number of k successes out of n total draws from a population of size N containing K successes.

source	term name Gene Ontology (Biological process)	n. of term genes	n. of query genes	n. of common genes	corrected p-value
BP	regulation of natural killer cell chemotaxis	8	18	2	4.65e-02
BP	immune system process	2587	18	13	4.21e-05
BP	immune effector process	759	18	8	3.57e-04
BP	leukocyte mediated immunity	285	18	5	1.02e-02
BP	lymphocyte mediated immunity	216	18	5	2.63e-03
BP	regulation of immune system process	1513	18	9	5.85e-03
BP	regulation of immune effector process	410	18	6	2.75e-03
BP	regulation of leukocyte mediated immunity	146	18	4	1.82e-02
BP	regulation of lymphocyte mediated immunity	107	18	4	5.28e-03
BP	multi-organism process	2373	18	11	2.82e-03
BP	signaling	6166	18	15	2.25e-02
BP	single organism signaling	6157	18	15	2.20e-02



Cluster annotation



GOsummaries

https://www.bioconductor.org/packages/release/bioc/html/GOsummaries.html

Practice time!