Universität Rostock Traditio et Innovatio

Universitätsmedizin Rostock

Integrative Bioinformatics and Systems Biology

Dr. Mohamed Hamed

Lecture 11

Introduction to Machine Learning I

Prerequisites

- Basic mathematics
- Basic programming skills
- Linear algebra
- Basic knowledge in statistics

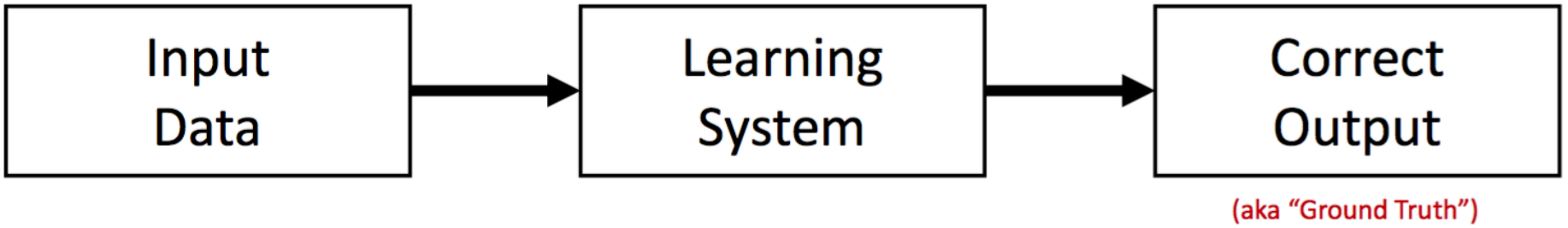
Machine Learning Definition

- Machine learning concerns **How to construct computer programs that automatically improve with experience**. "Tom Mitchel"

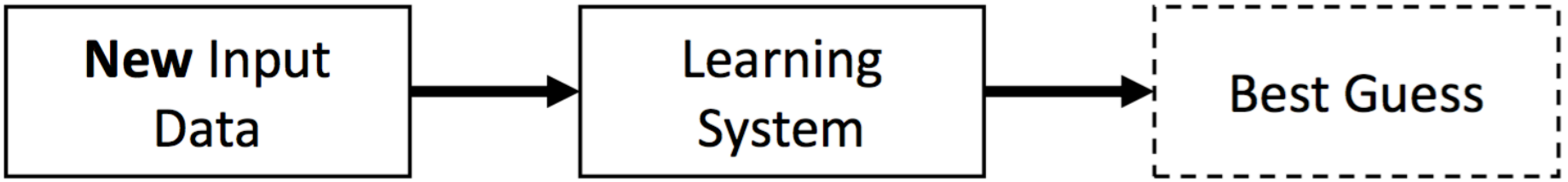
- Machine Learning studies **models that can learn to make predictions from data instead of using static instructions**

Machine Learning

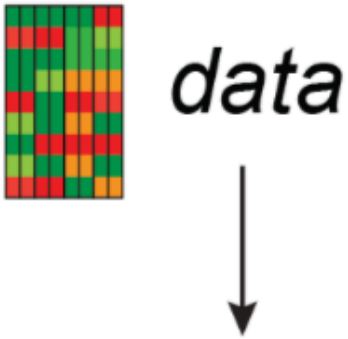
Training Stage:



Production Stage:



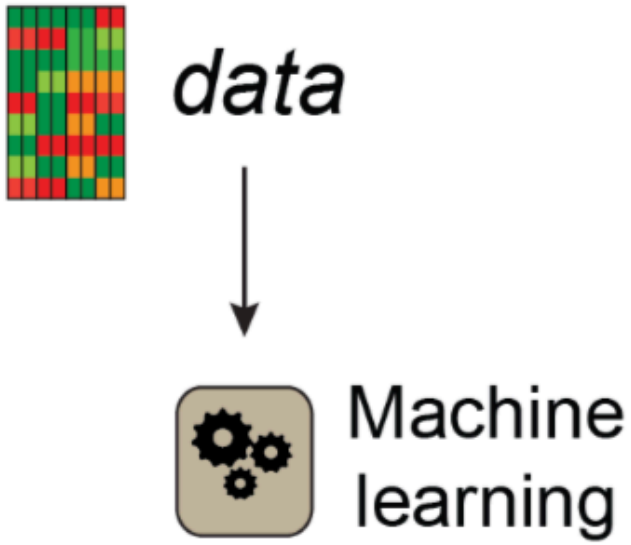
Machine Learning



Learning from examples

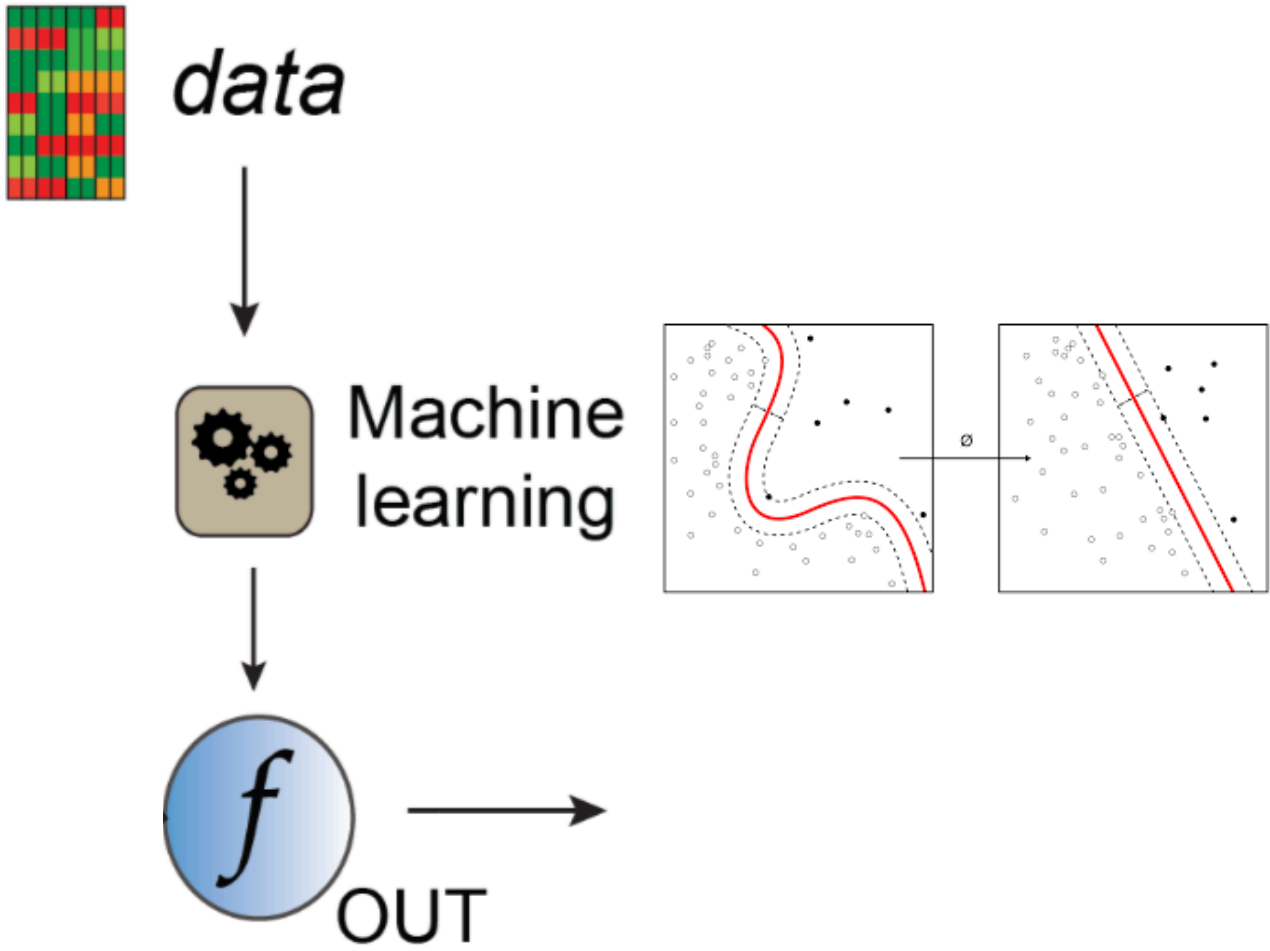
Machine Learning

Learning from examples



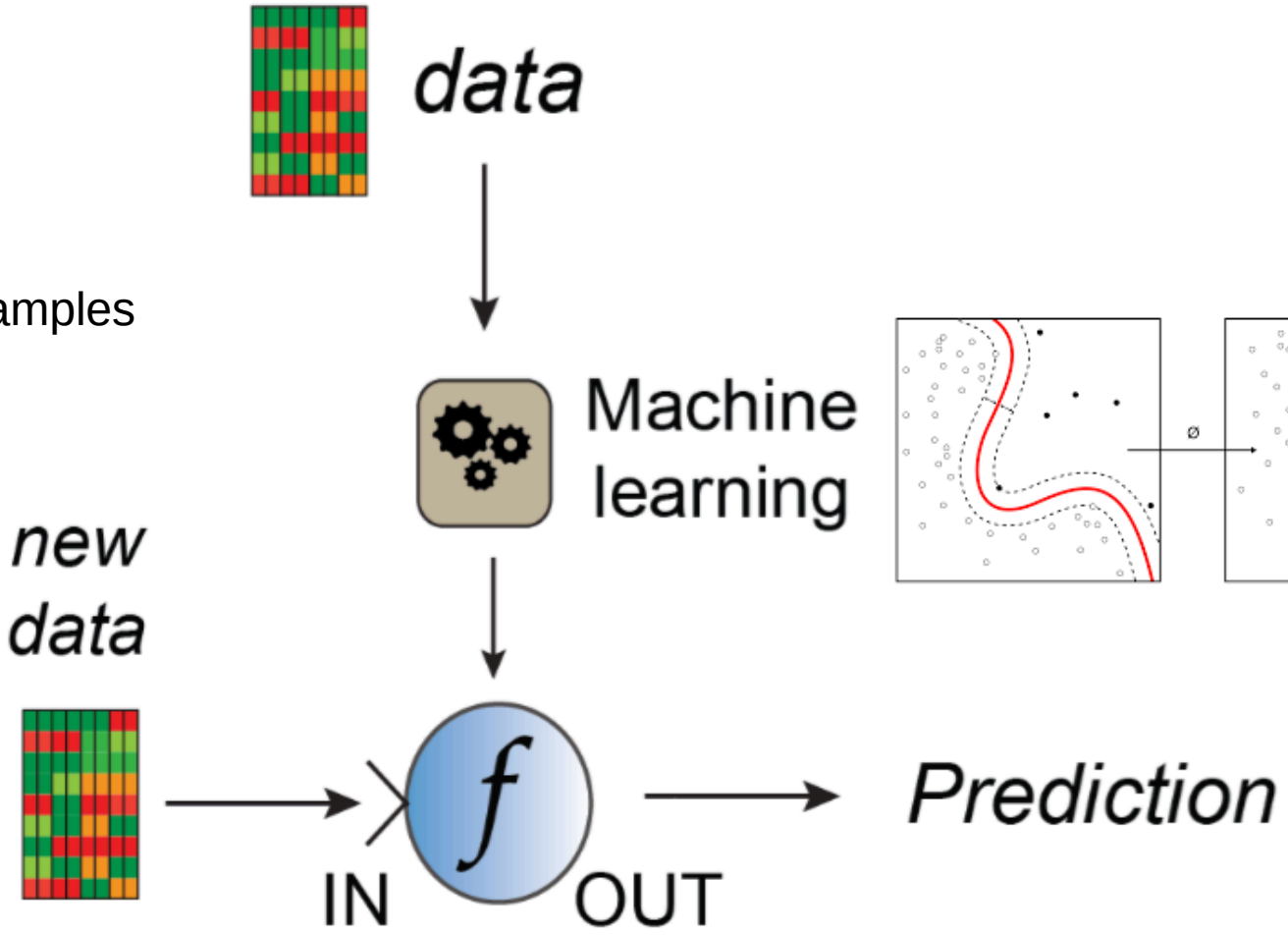
Machine Learning

Learning from examples



Machine Learning

Learning from examples



Machine Learning

General Approach:

Given **training** data $T_D = \{y, \mathbf{x}\} = (y,x)_1 \ldots (y,x)_N$,

function space {f} and a
constraint on these functions

Teach a machine to learn the **mapping** y = f(x)

```
trained_model <- model(data, known_quantity)
predicted_quantity <- trained_model(new_data)
```

Examples for applications

- **Medical**: Predicted whether a patient will have a second heart attack

  Data: demographic, diet, clinical measurements
- **Business/Economics**: Predict the price of stock 6 months from now.
  - Data: company performance, economic data
- **Vision :** Identify hand-written ZIP codes
  - Data: Model hand-written digits
- **Medical**: Amount of glucose in the blood of a diabetic
  - Data: Infrared absorption spectrum of blood sample
- **Medical**: Risk factors for prostate cancer

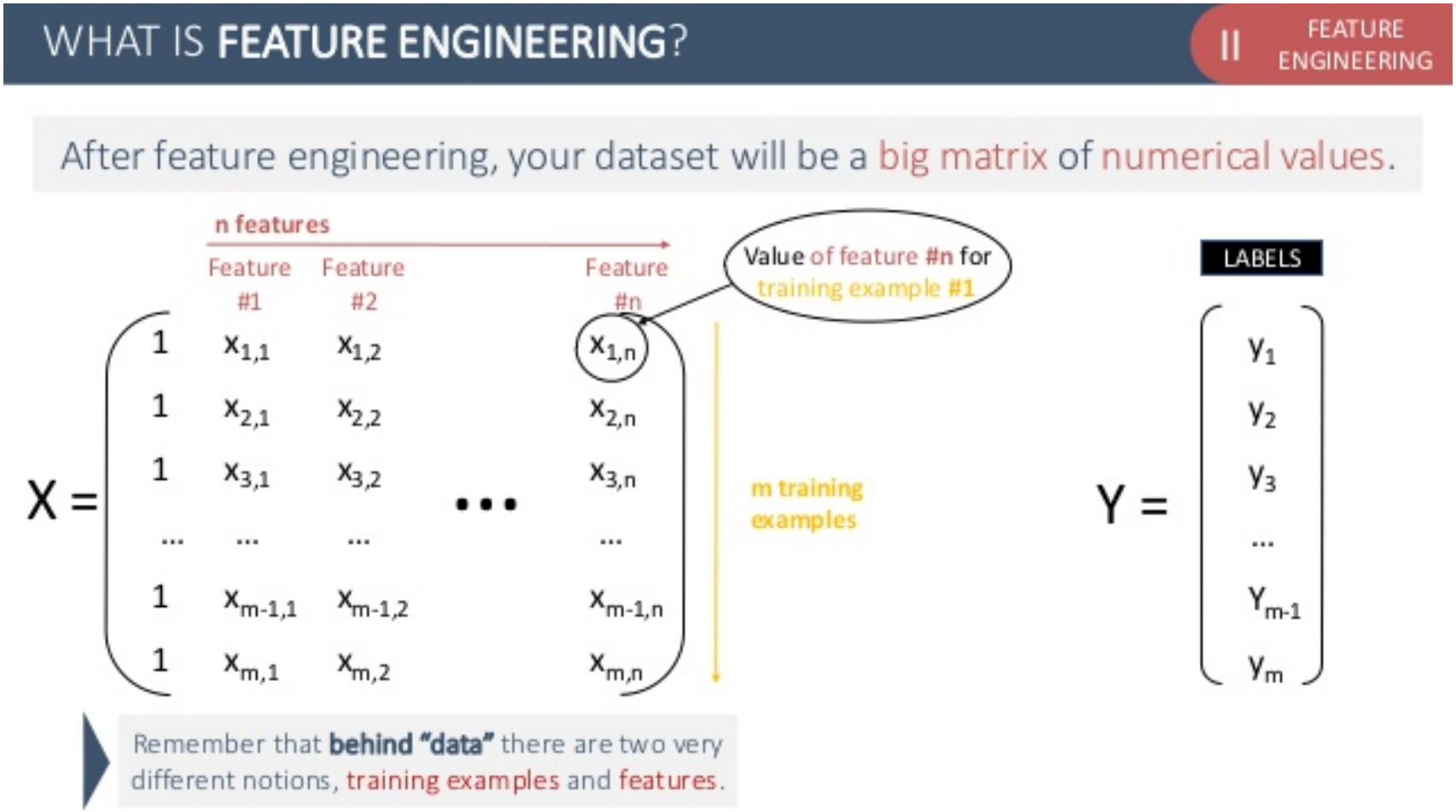
  Data: Clinical, demographic

Data Types

- Two basically different types of data
  - Quantitative (numerical): e.g. stock price
  - Categorical (discrete, often binary): cancer/no cancer
- Data are predicted
  - on the basis of a set of features (e.g. diet or clinical measurements)
  - from a set of (observed) training data on these features
  - For a set of objects (e.g. people).
  - Input features for the problems are also called predictors
  - or independent variables
- Outputs are also called responses or dependent variables
  The prediction model is called a learner or estimator

- Supervised learning: learn on outcomes for observed features
- Unsupervised learning: no output values available

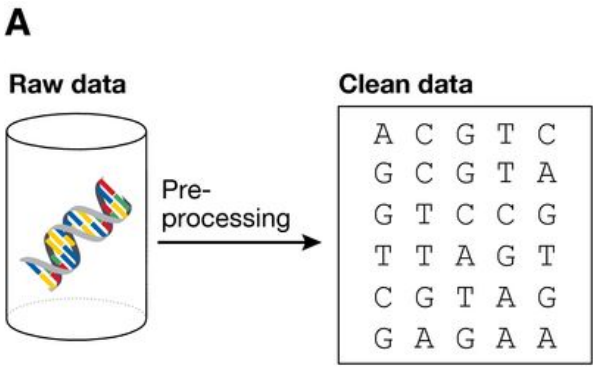
Matrix representation



Machine Learning workflow

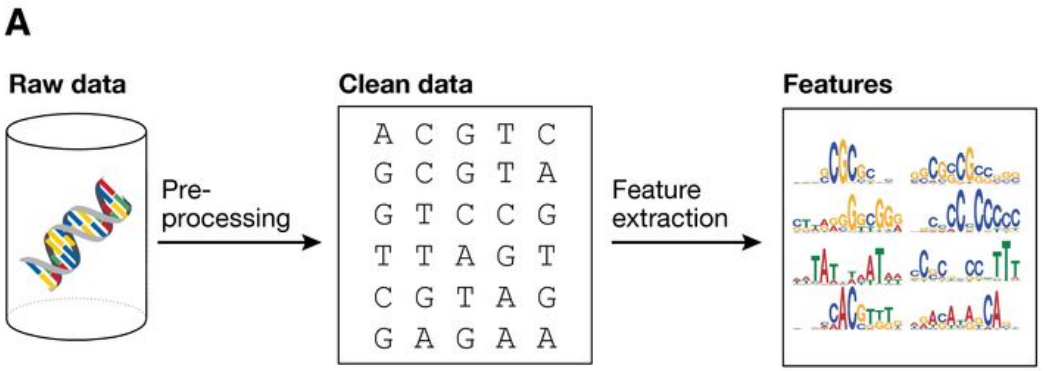
Christof Angermueller et al. Mol Syst Biol 2016;12:878

Machine Learning workflow



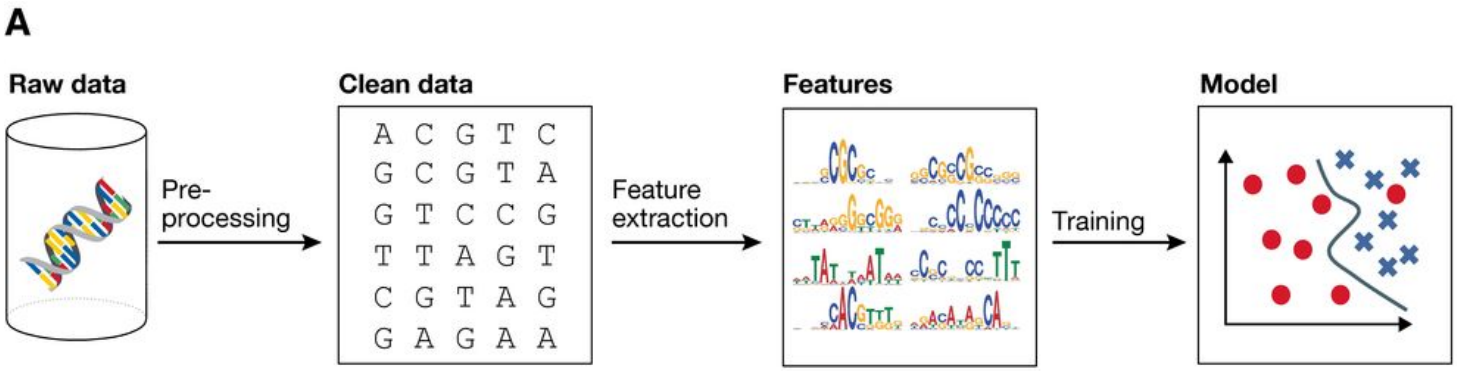
Christof Angermueller et al. Mol Syst Biol 2016;12:878

Machine Learning workflow



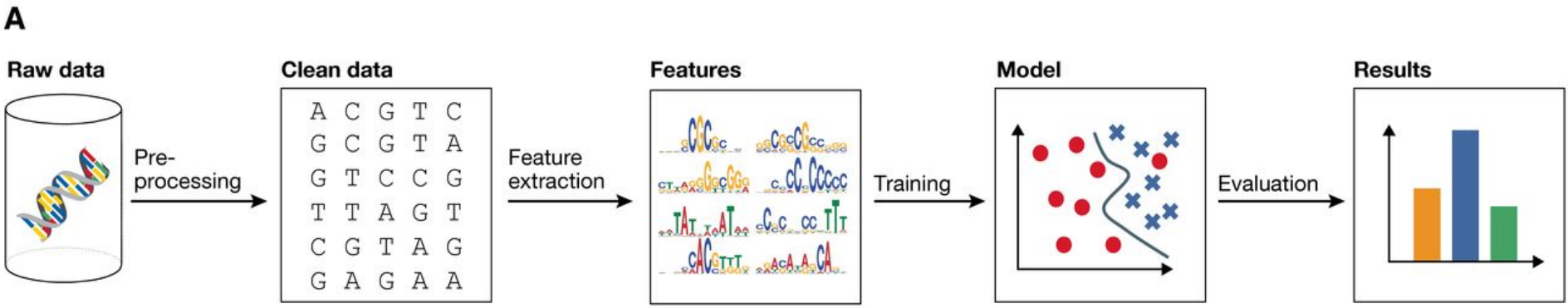
Christof Angermueller et al. Mol Syst Biol 2016;12:878

Machine Learning workflow



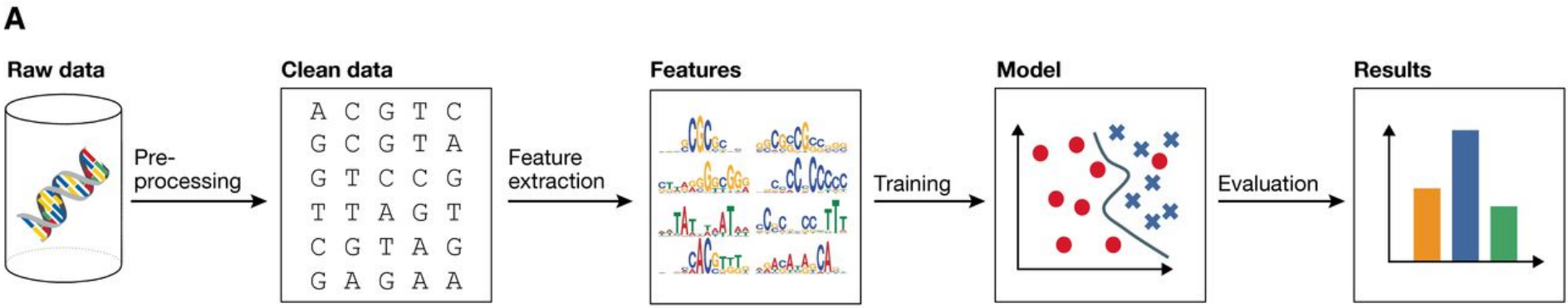
Christof Angermueller et al. Mol Syst Biol 2016;12:878

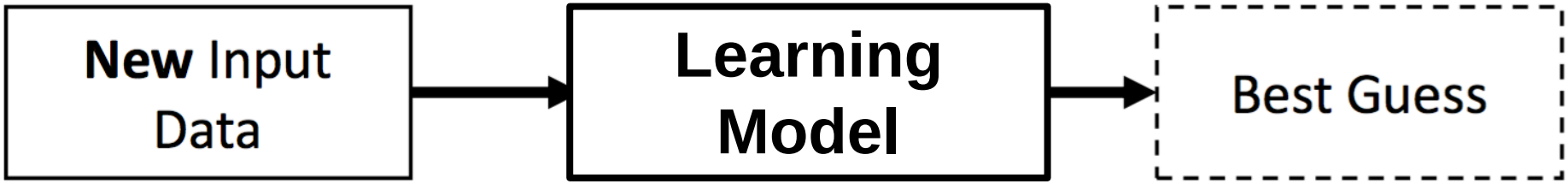
Machine Learning workflow



Christof Angermueller et al. Mol Syst Biol 2016;12:878

Machine Learning workflow





Christof Angermueller et al. Mol Syst Biol 2016;12:878

Machine Learning Types

Supervised learning

Training your machine to learn a function
by showing couples of input and corresponding output (target)
→ *Classification and Regression*

Unsupervised learning

Training your machine to learn structure or relationships
by presenting to it a set of inputs
→ *Clustering and Dimensionality reduction*

Machine Learning Types

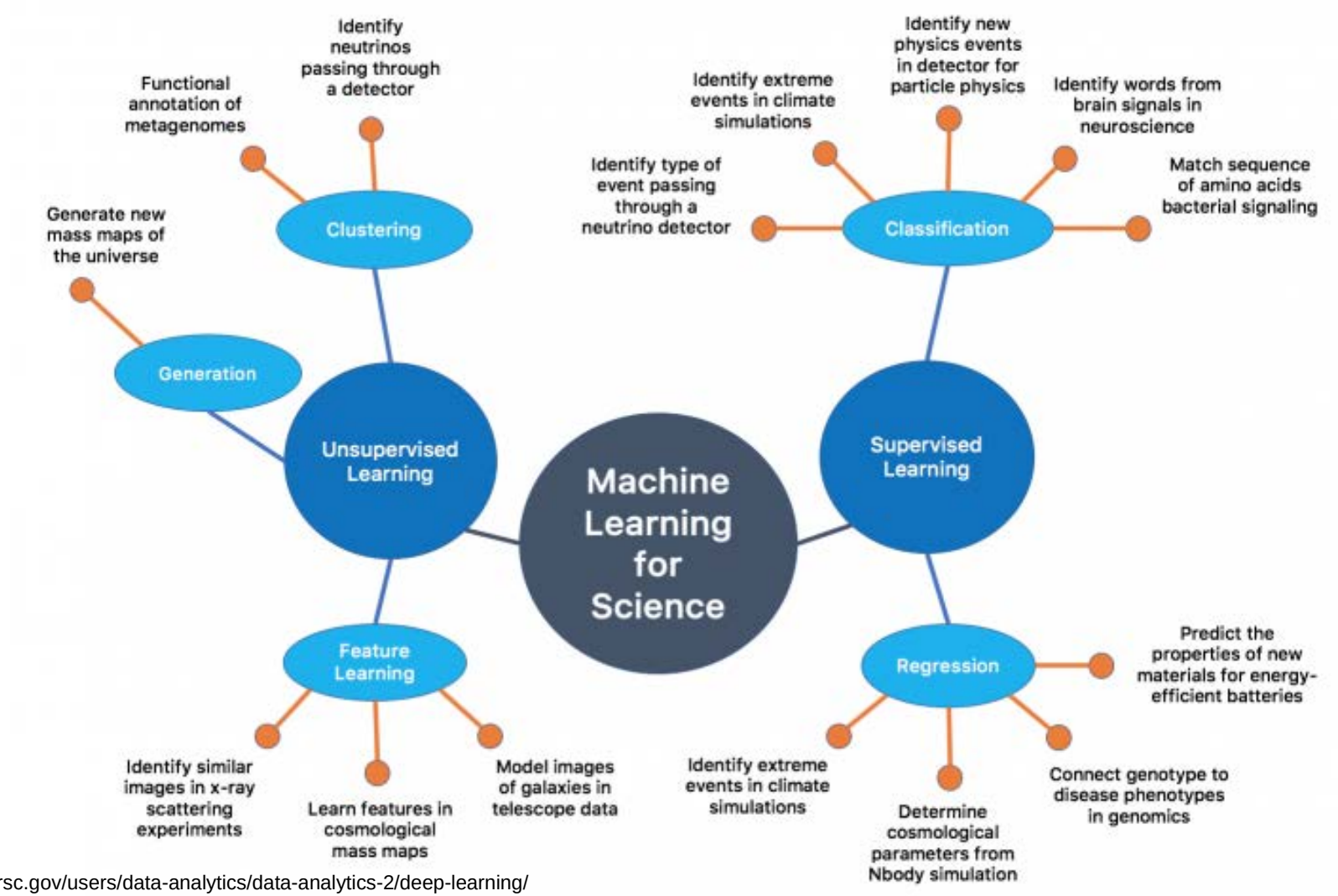
Supervised learning

Training your machine to learn a function
by showing couples of input and corresponding output (target)
→ *Classification and Regression*

Unsupervised learning

Training your machine to learn structure or relationships
by presenting to it a set of inputs
→ *Clustering and Dimensionality reduction*

Application Examples for ML Types



http://www.nersc.gov/users/data-analytics/data-analytics-2/deep-learning/

Gene Expression Matrix

	Control samples				Study samples			
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
Gene 1	12	11	15	13	1	2	4	5
Gene 2	7	6	4	0.5	23	21	23	22
Gene 3	5	2	5	6	12.4	14	12	15.5
.....	18	15	15	20	3	2	3	4
.....	19	18	11	17	11	14	12	17
Gene n	5	13	14	22	12	11	23	10

T (Gene Expression Matrix)

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	Type
Sample1	12	11	15	13	1	2	4	N
Sample2	7	6	4	0.5	23	21	23	N
Sample3	5	2	5	6	12.4	14	12	N
Sample4	18	15	15	20	3	2	3	T
Sample5	19	18	11	17	11	14	12	T
Sample6	5	13	14	22	12	11	23	T

Supervised learning : Classification Task

Label/ outcom

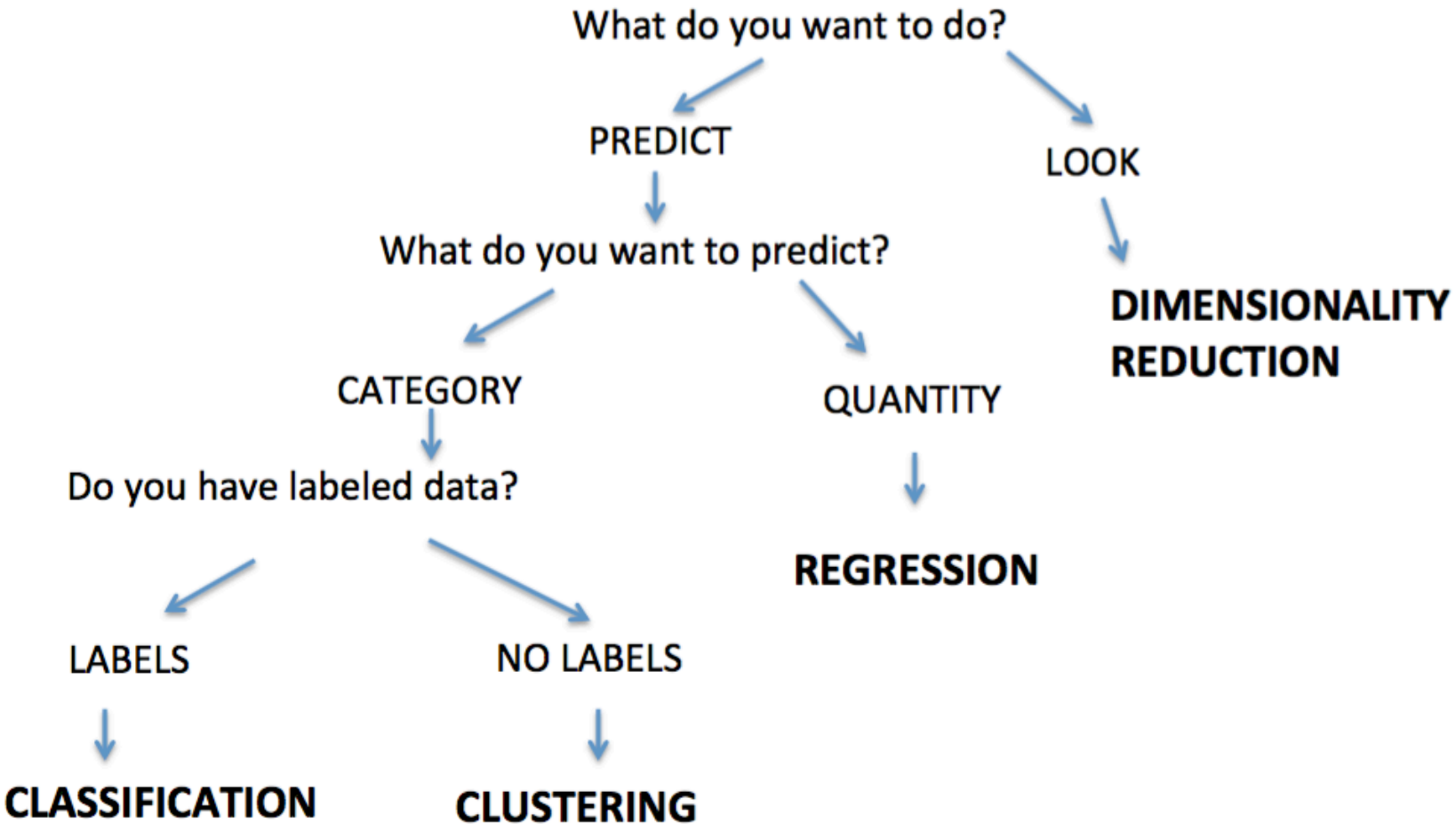
	gene1	gene2	gene3	gene4	gene5	gene6	gene7	Type
Sample1	12	11	15	13	1	2	4	N
Sample2	7	6	4	0.5	23	21	23	N
Sample3	5	2	5	6	12.4	14	12	N
Sample4	18	15	15	20	3	2	3	T
Sample5	19	18	11	17	11	14	12	T
Sample6	5	13	14	22	12	11	23	T

Supervised learning : Regression Task

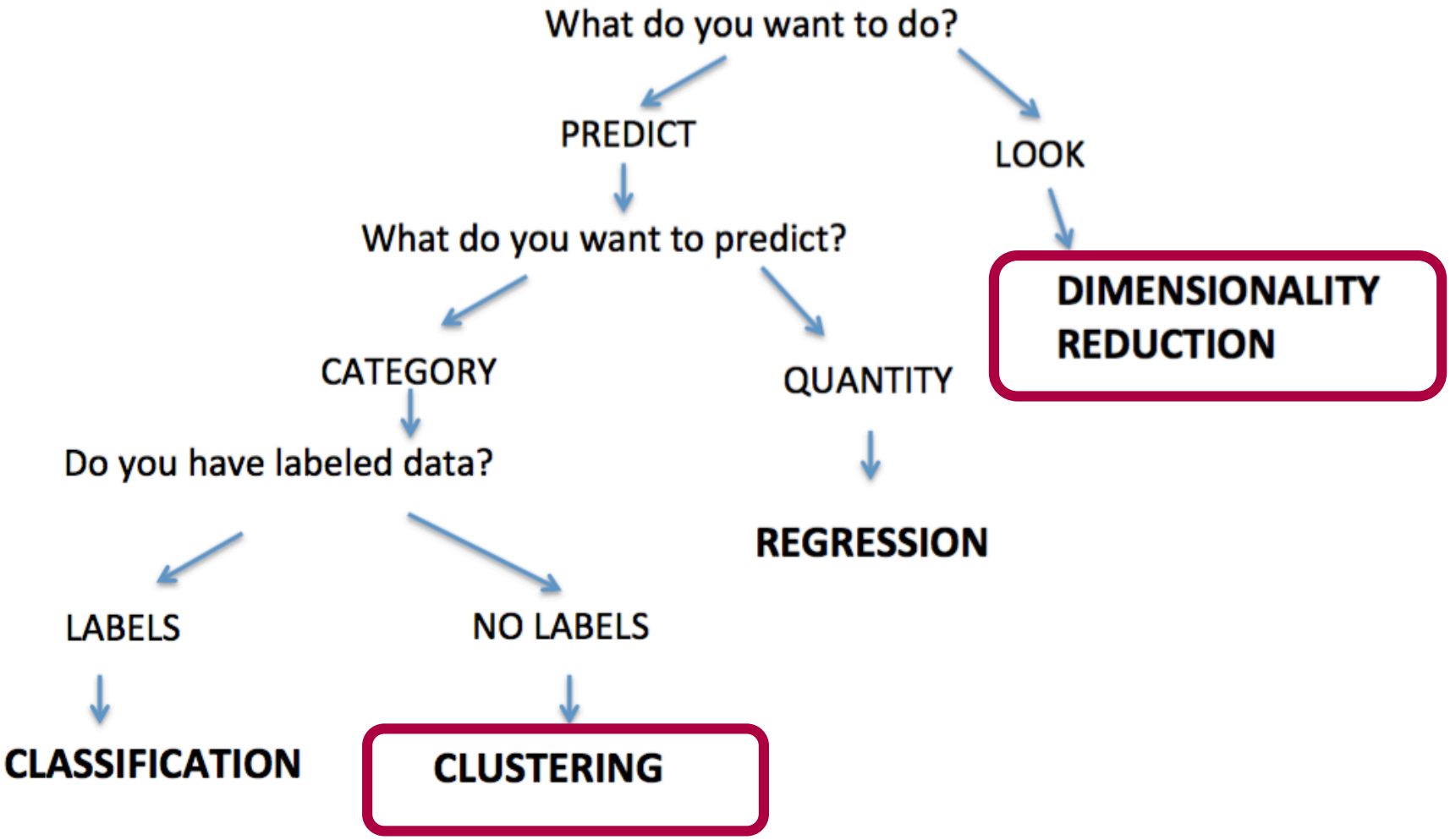
Label/ outcom

	gene1	gene2	gene3	gene4	gene5	gene6	gene7	OSR
Sample1	12	11	15	13	1	2	4	0.6
Sample2	7	6	4	0.5	23	21	23	0.7
Sample3	5	2	5	6	12.4	14	12	0.5
Sample4	18	15	15	20	3	2	3	0.8
Sample5	19	18	11	17	11	14	12	0.8
Sample6 **.....** **.....** **.....** **.....**	5	13	14	22	12	11	23	0.4

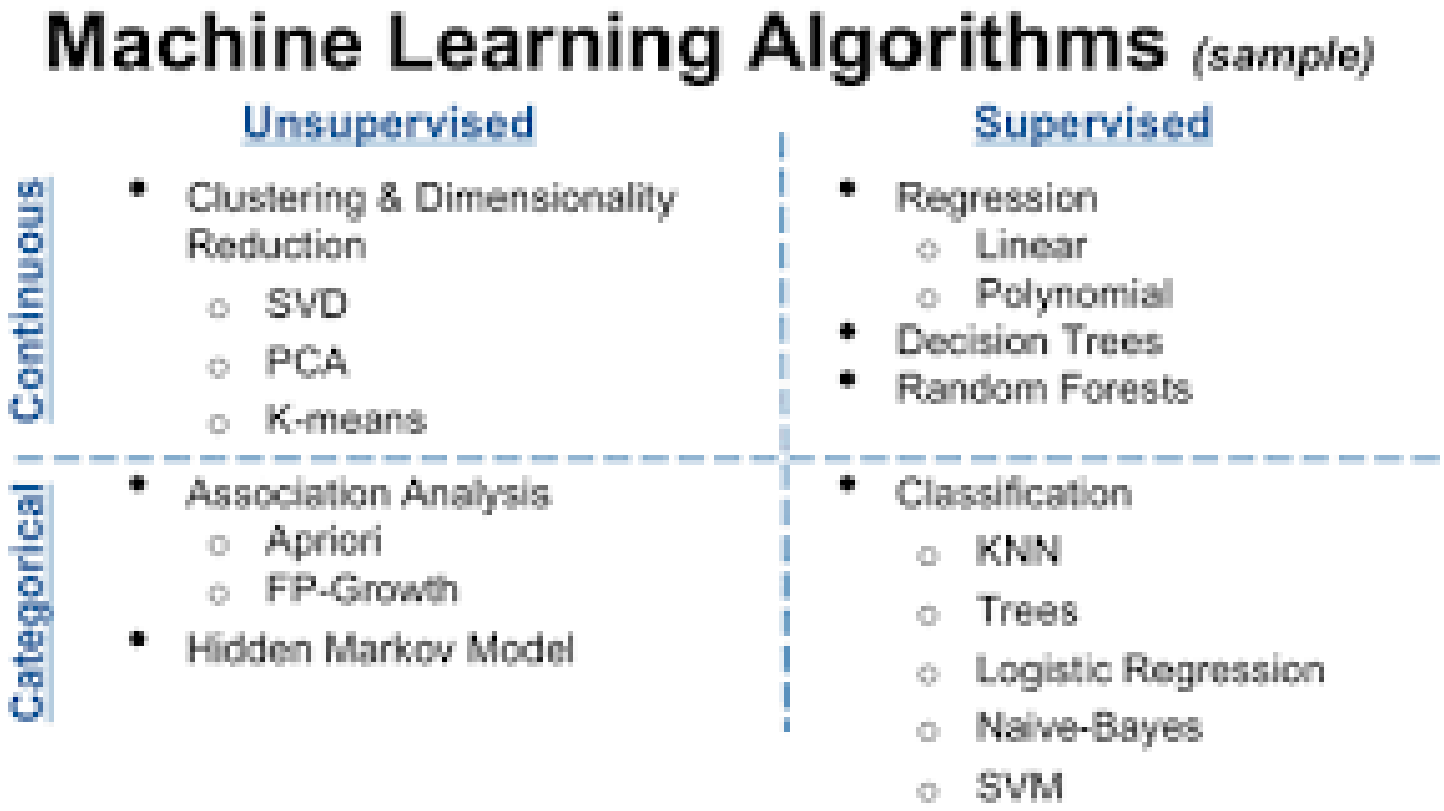
ML types : based on the research problem



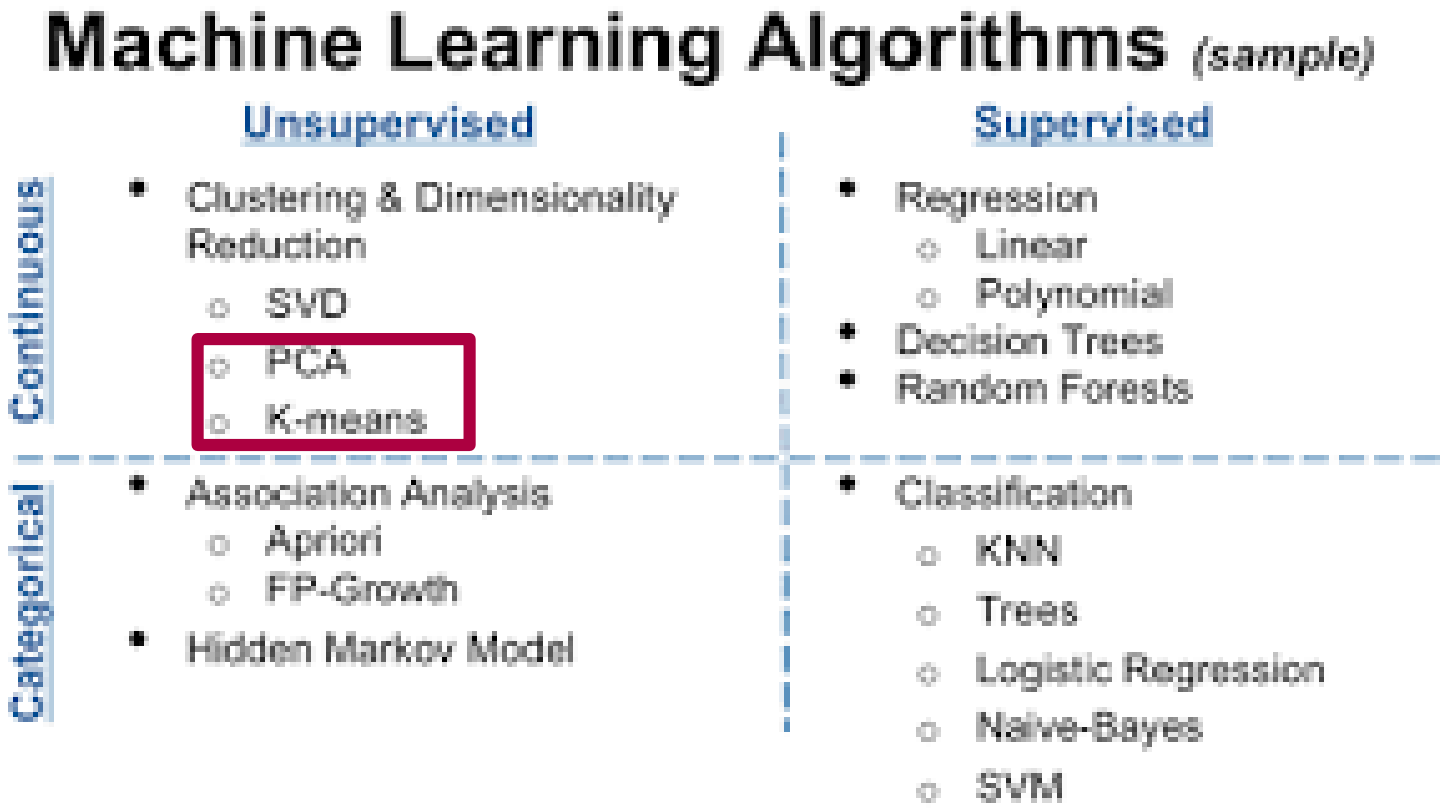
ML types : based on the research problem



Quick outlook



Quick outlook



Unsupervised Machine Learning
1- Clustering

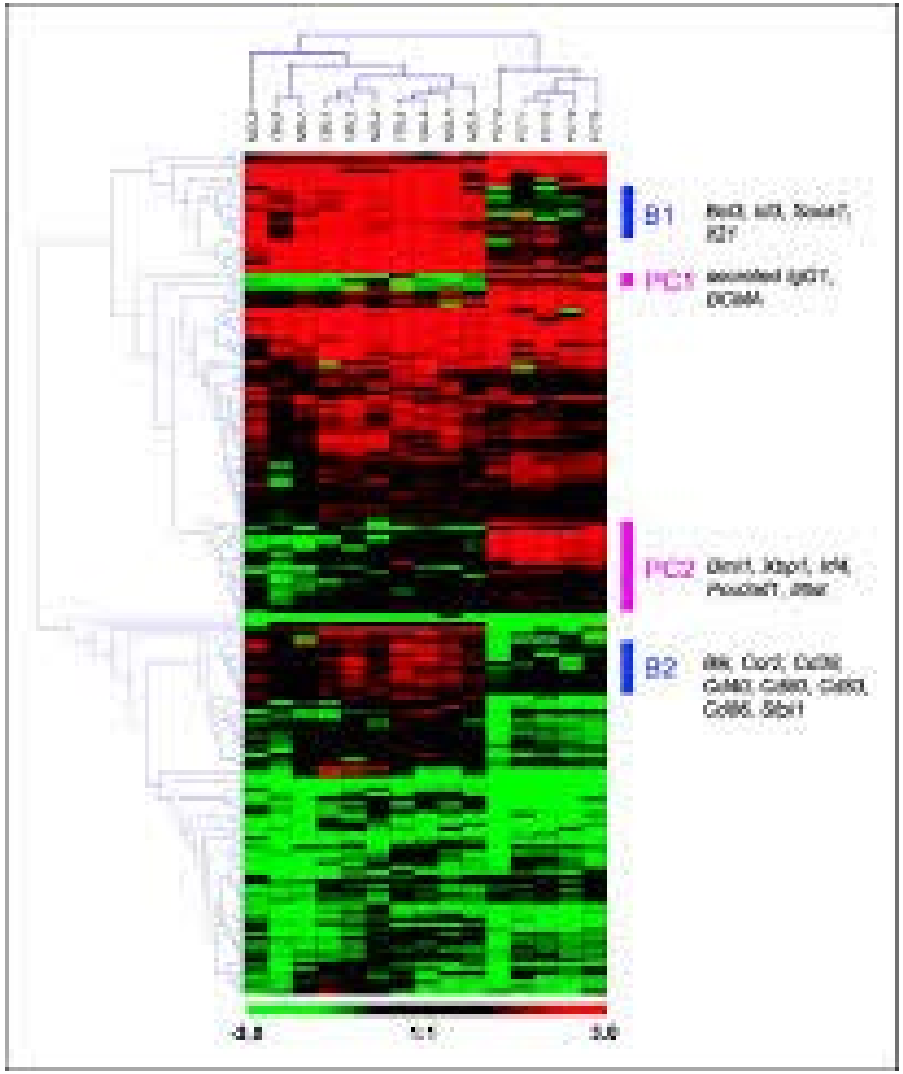
Clustering Gene expressions

Predict:

Which genes show similar expression over the samples

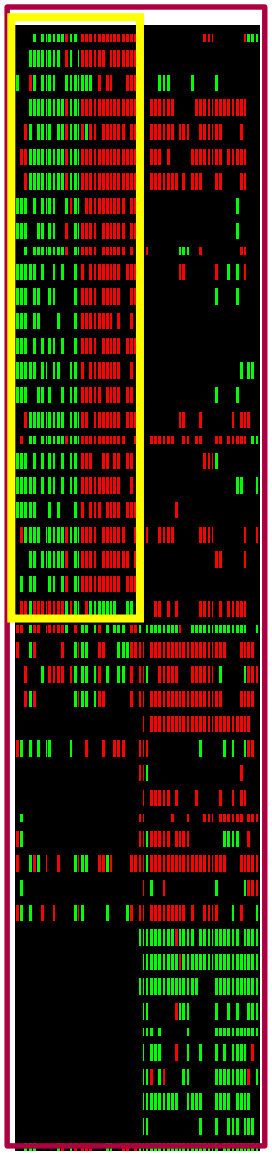
Which samples show similar expression over the genes (unsupervised learning problem)

Which genes are highly over or under expressed in certain cancers (supervised learning problem)

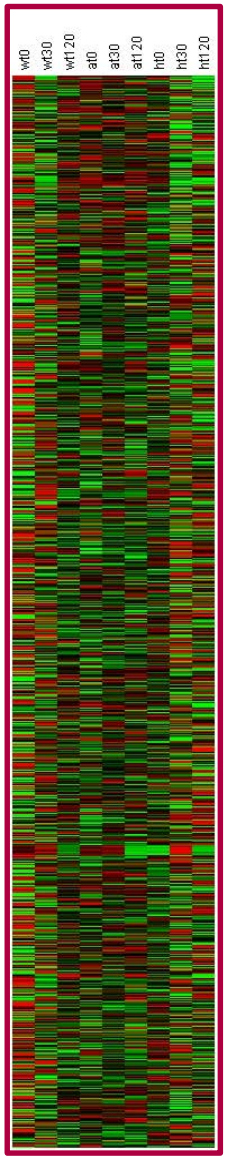


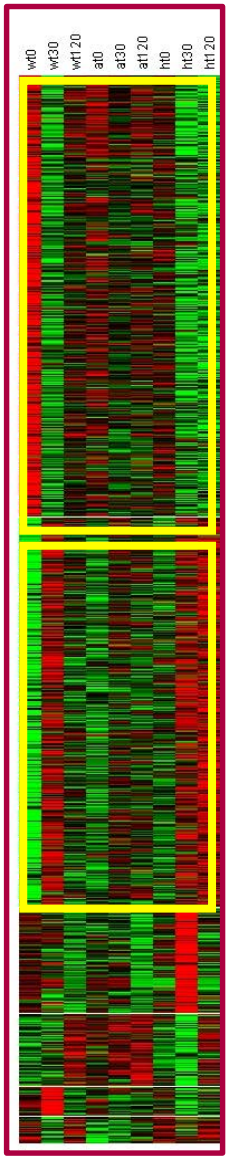
Bi-clustering

clustering



Bi cluster I





Cluster I

Cluster II

Cluster III

Homogeneity
Separation

Bi-clustering

Bioinformatics. 2006 May 15;22(10):1282-3. Epub 2006 Mar 21.

BicAT: a biclustering analysis toolbox.

Barkow S[1], Bleuler S, Prelic A, Zimmermann P, Zitzler E.

Author information

Abstract

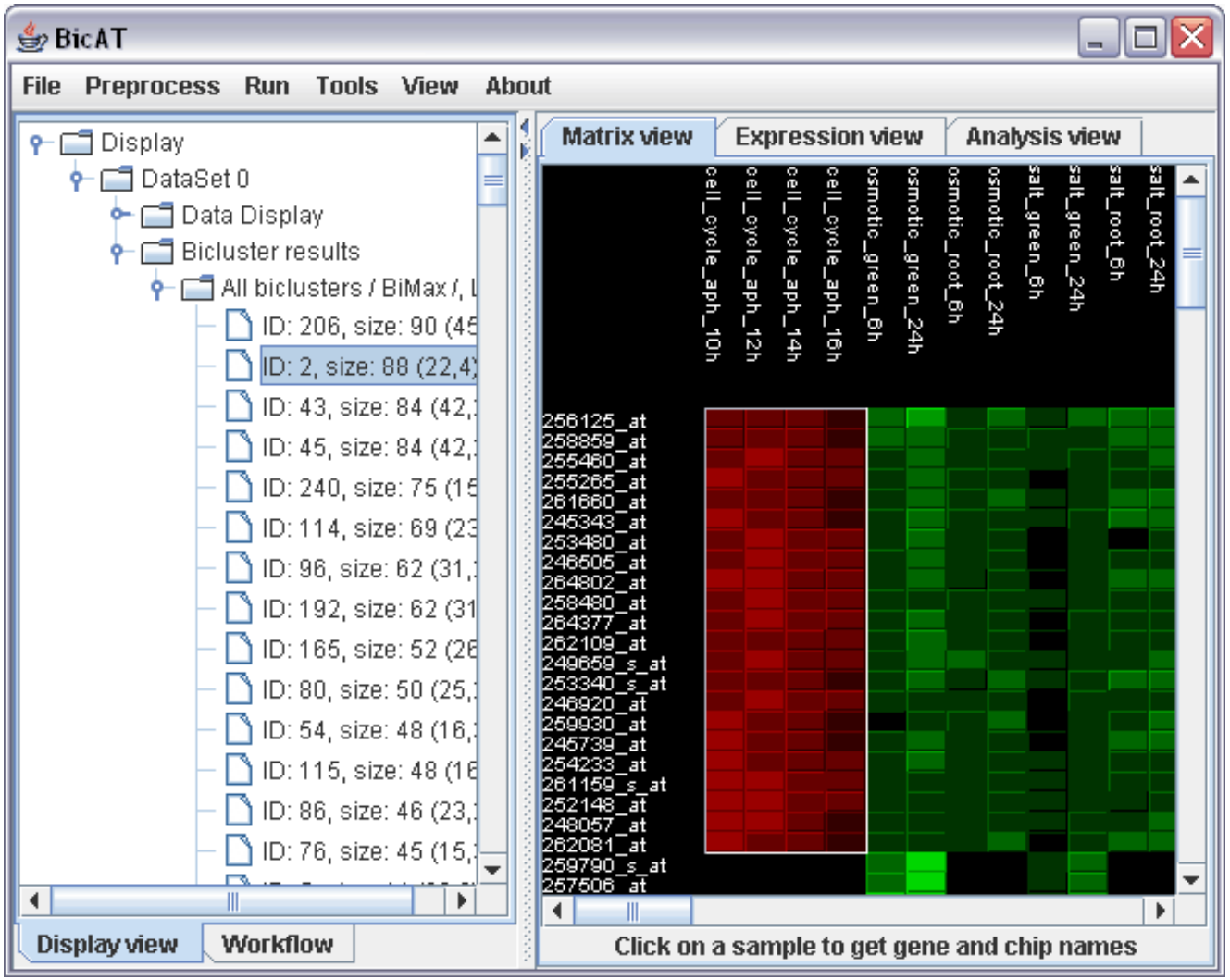
SUMMARY: Besides classical clustering methods such as hierarchical clustering, in recent years biclustering has become a popular approach to analyze biological data sets, e.g. gene expression data. The Biclustering Analysis Toolbox (BicAT) is a software platform for clustering-based data analysis that integrates various biclustering and clustering techniques in terms of a common graphical user interface. Furthermore, BicAT provides different facilities for data preparation, inspection and postprocessing such as discretization, filtering of biclusters according to specific criteria or gene pair analysis for constructing gene interconnection graphs. The possibility to use different biclustering algorithms inside a single graphical tool allows the user to compare clustering results and choose the algorithm that best fits a specific biological scenario. The toolbox is described in the context of gene expression analysis, but is also applicable to other types of data, e.g. data from proteomics or synthetic lethal experiments.

AVAILABILITY: The BicAT toolbox is freely available at http://www.tik.ee.ethz.ch/sop/bicat and runs on all operating systems. The Java source code of the program and a developer's guide is provided on the website as well. Therefore, users may modify the program and add further algorithms or extensions.

BiCAT bi-clustering algorithms

Bi-clustering is a special case of Clustering.
But on certain set of samples and genes.

Could be used for building the Association map in Radiogenomics



[BiCat : ETH Zürich]
No further maintenance nor development

WGCN: an R package for weighted correlation network analysis

- Useful package for clustering and it provides nice visualization
- https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/

WGCNA: an R package for weighted correlation network analysis

Peter Langfelder and Steve Horvath
with help of many other contributors

Semel Institute for Neuroscience and Human Behavior, UC Los Angeles (PL),
Dept. of Human Genetics and Dept. of Biostatistics, UC Los Angeles (SH)

Peter (dot) Langfelder (at) gmail (dot) com, SHorvath (at) mednet (dot) ucla (dot) edu

BMC Bioinformatics, 2008 9:559 (link opens in a new tab/window)

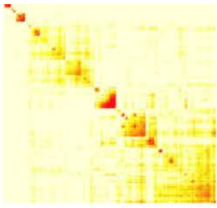
WGCN:

-detects clusters/ co-expression modules/network modules.

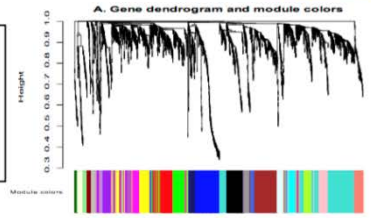
-It provides nice visualization and

-further association analysis for each module

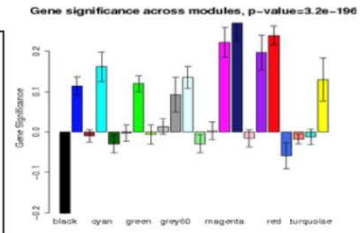
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



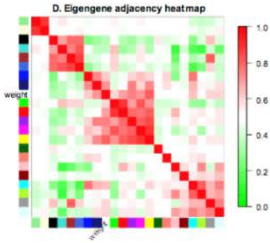
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



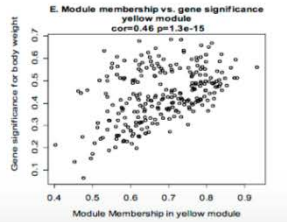
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing



Unsupervised Learning

1. Unsupervised learning
    1. Clustering
        1. Subgroups within the data
        2. Distance based
    1. Dimensionality reduction
        1. Pattern identification in features
        2. Facilitate visualisation
        3. Pre-processing before supervised

Clustering based on distance metrics

- Cluster analysis is the task of partitioning the dataset into subsets, so that: the points in each subset are more similar to each other than those from different subsets

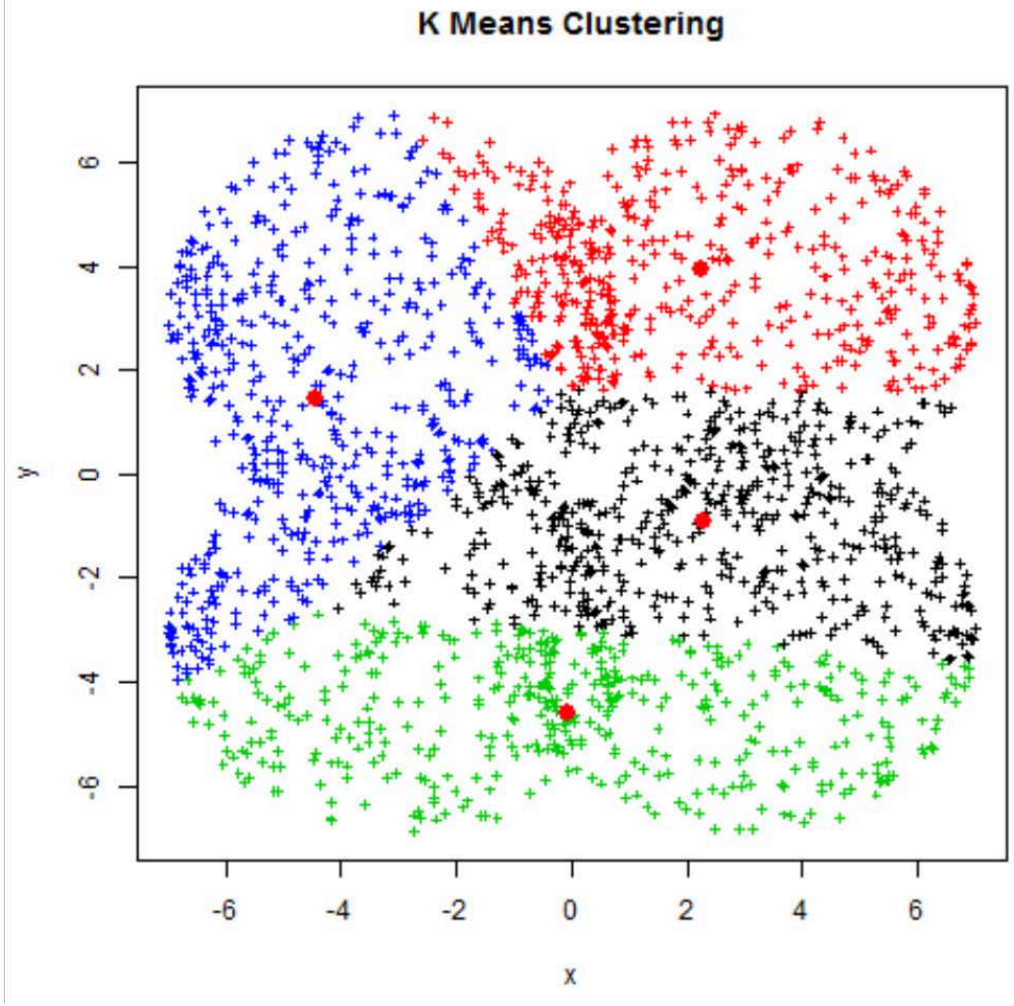
- Based on distance metrics:

Euclidean (L2 norm) $\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$.

Manhattan (L1 norm) $\sum_{i=1}^{n} |p_i - q_i|$

Minkowski $\left(\sum_{i=1}^{n} |p_i - q_i|^c\right)^{1/c}$

K-Means algorithm

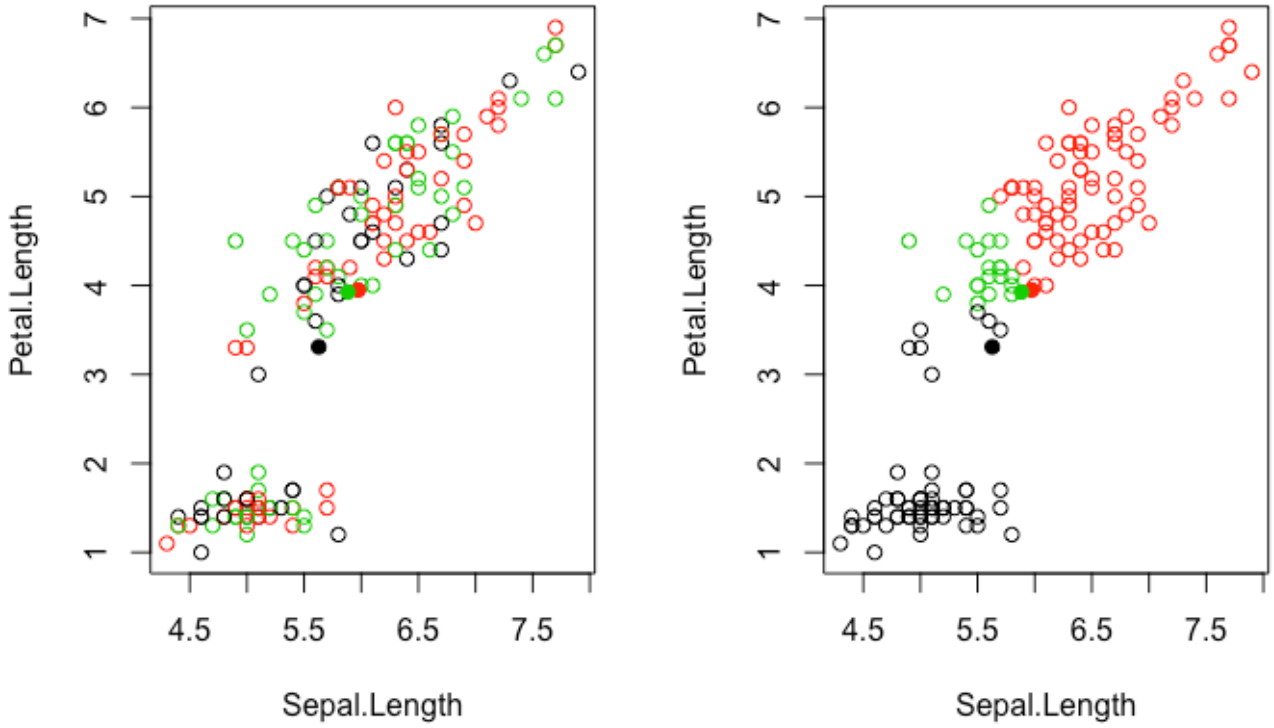


Algorithm

- Choose k centroids randomly.
- Calculate the distance from each point in the dataset to be classified to each centroid.
- Assign each point to the nearest centroid.
- Calculate the centroids of the resulting clusters.
- Repeat until the centroids don't move too much.

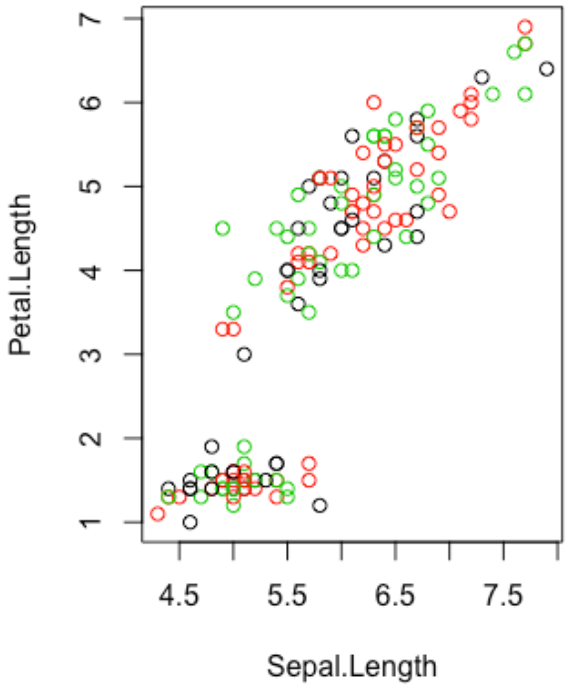
K-Means algorithm

K-means clustering aims at partitioning n observations into a fixed number of k clusters. Detecting homogeneous clusters

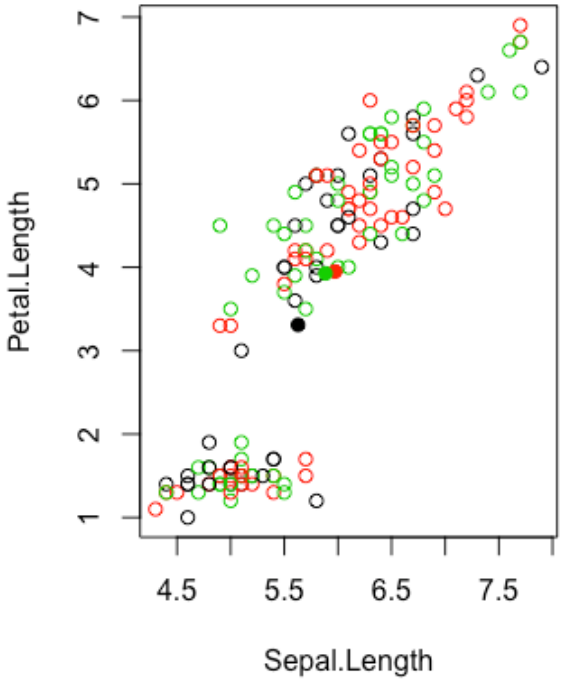


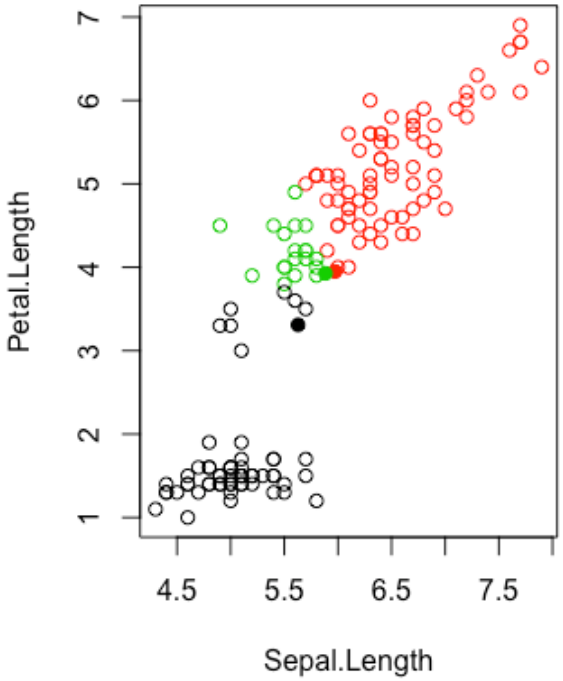
K-Means algorithm

Initialization



Iteration





2- Hierachical clustering (HCL)

Algorithm :

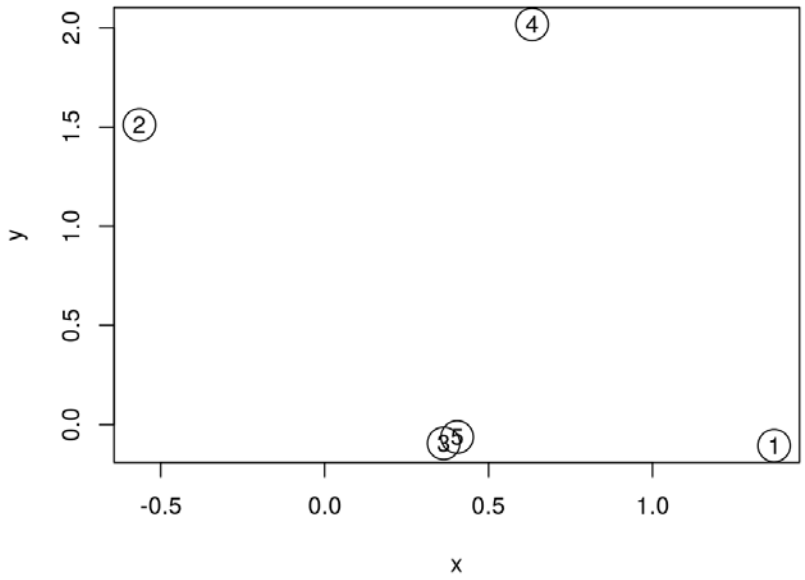
1. Initialization:
   - Assign each of the n points its own cluster

2. Iteration:
   - Find two nearest cluster, join them, leading to n-1 cluster
   - Continue merging cluster process, until a single cluster is left

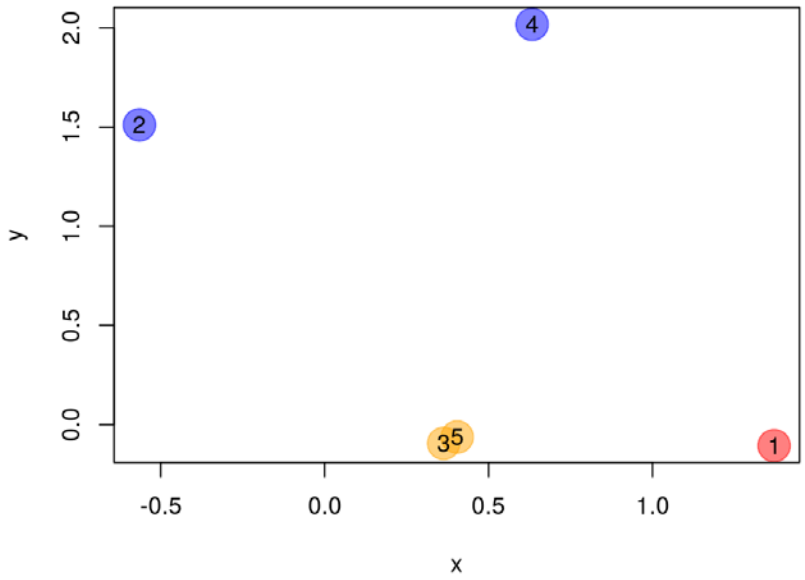
3. Termination:
   - All observations are grouped within a single cluster

2- Hierachical clustering (HCL)

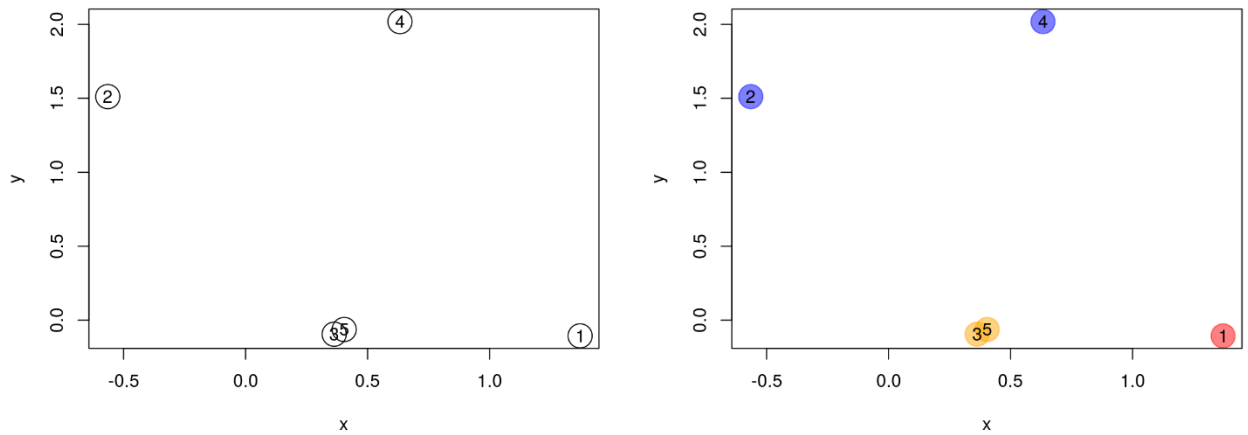
Initialisation:
- Numbers are the clusters

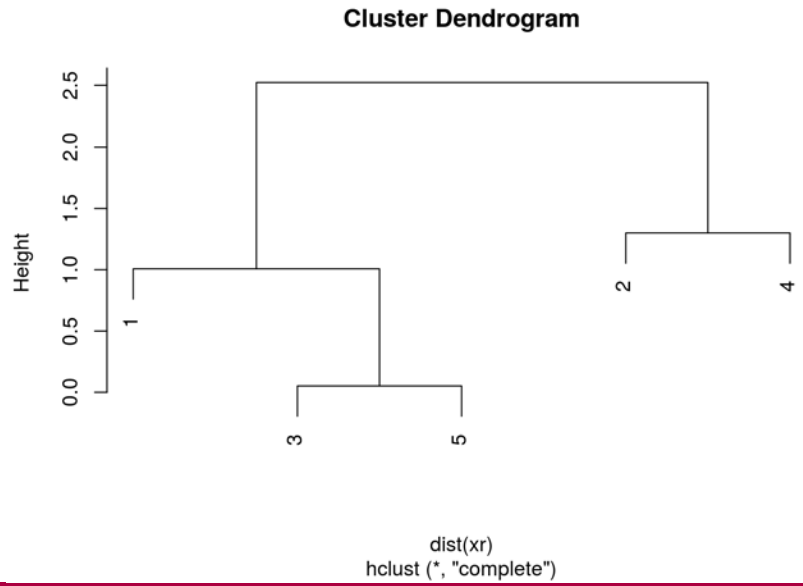


First iteration:
- Colours are the clusters



Hierachical clustering (HCL) - Dendrogram

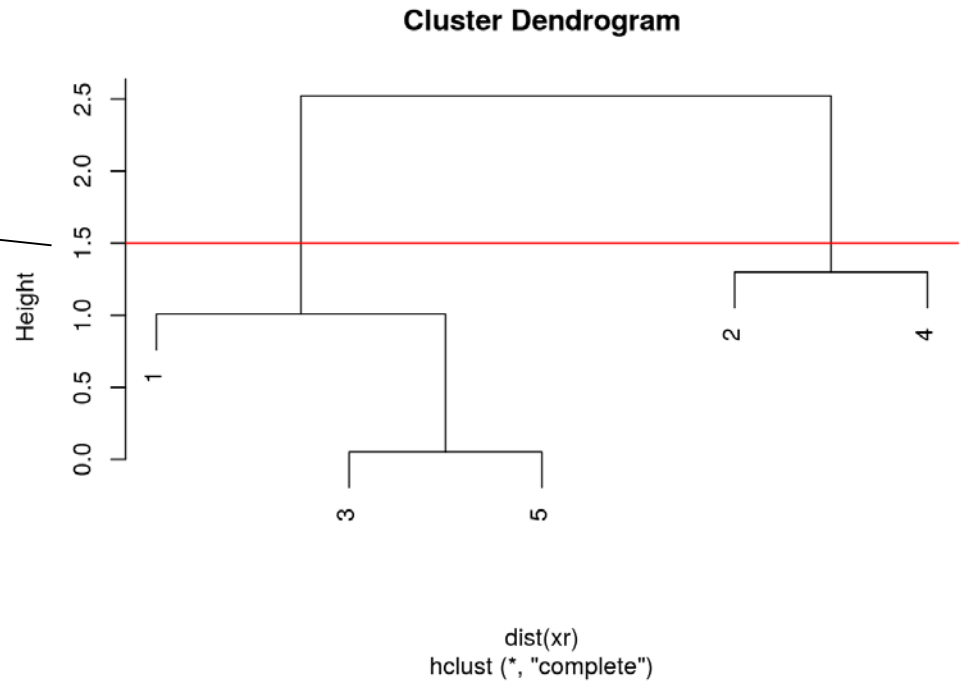




Hierachical clustering – Defining clusters

Cut the tree (dendrogram) at a specific height to defined the clusters

Cut at distance 1.5



- Results : 2 clusters

Unsupervised Machine Learning
2- Dimensionality reduction

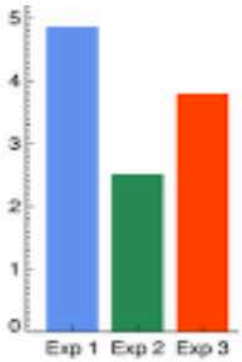
Curse of Dimensionality

Point:

x_1

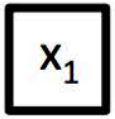
1 dimension

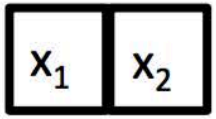
Representation of the space:



Curse of Dimensionality

Point:

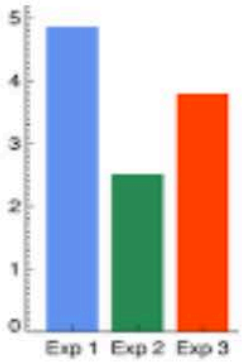


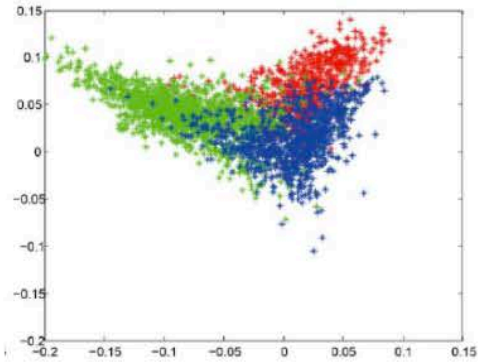


1 dimension

2 dimensions

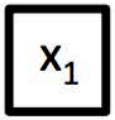
Representation of the space:



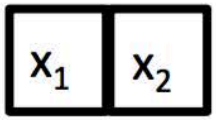


Curse of Dimensionality

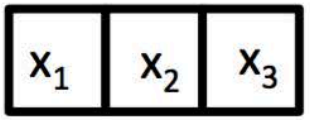
Point:



1 dimension

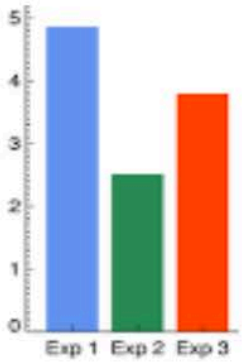


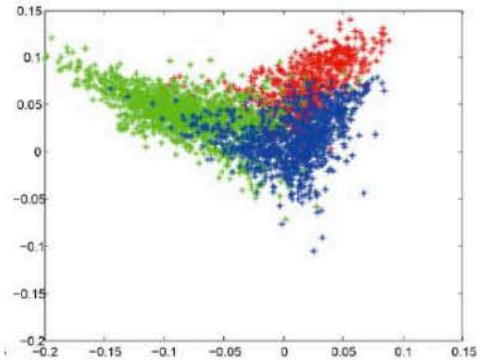
2 dimensions

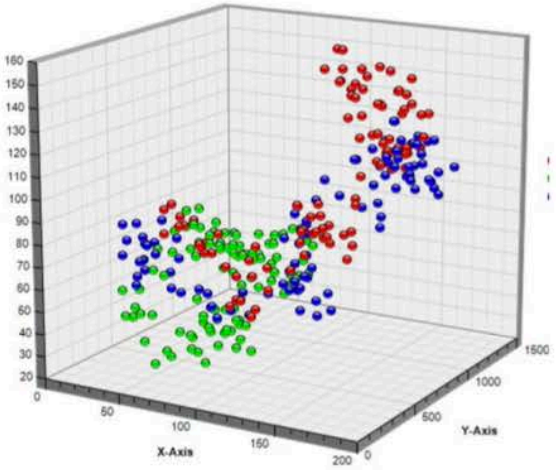


3 dimensions

Representation of the space:

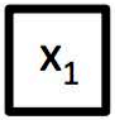




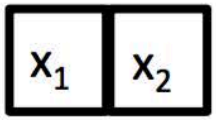


Curse of Dimensionality

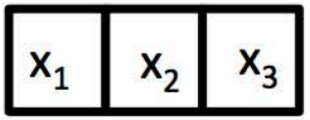
Point:



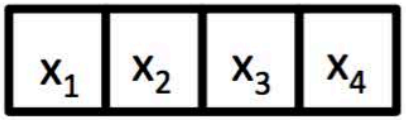
1 dimension



2 dimensions

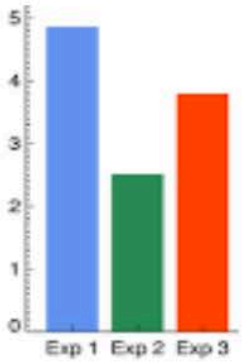


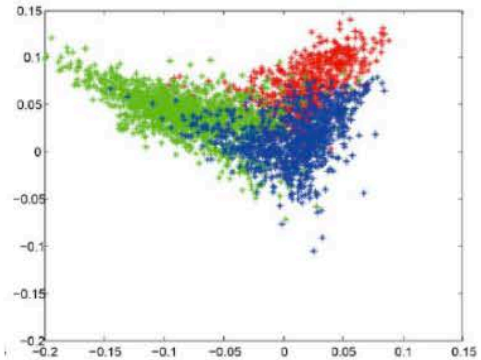
3 dimensions

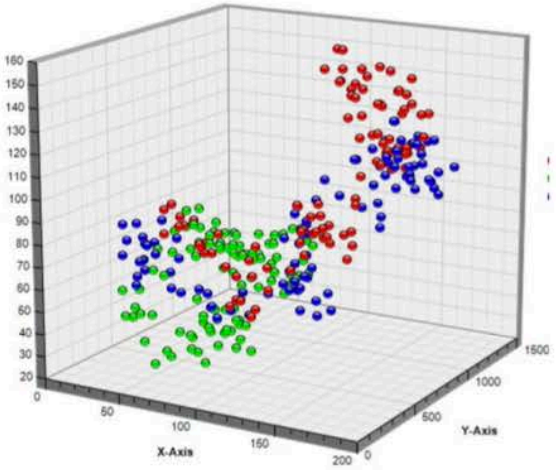


4 dimensions

Representation of the space:





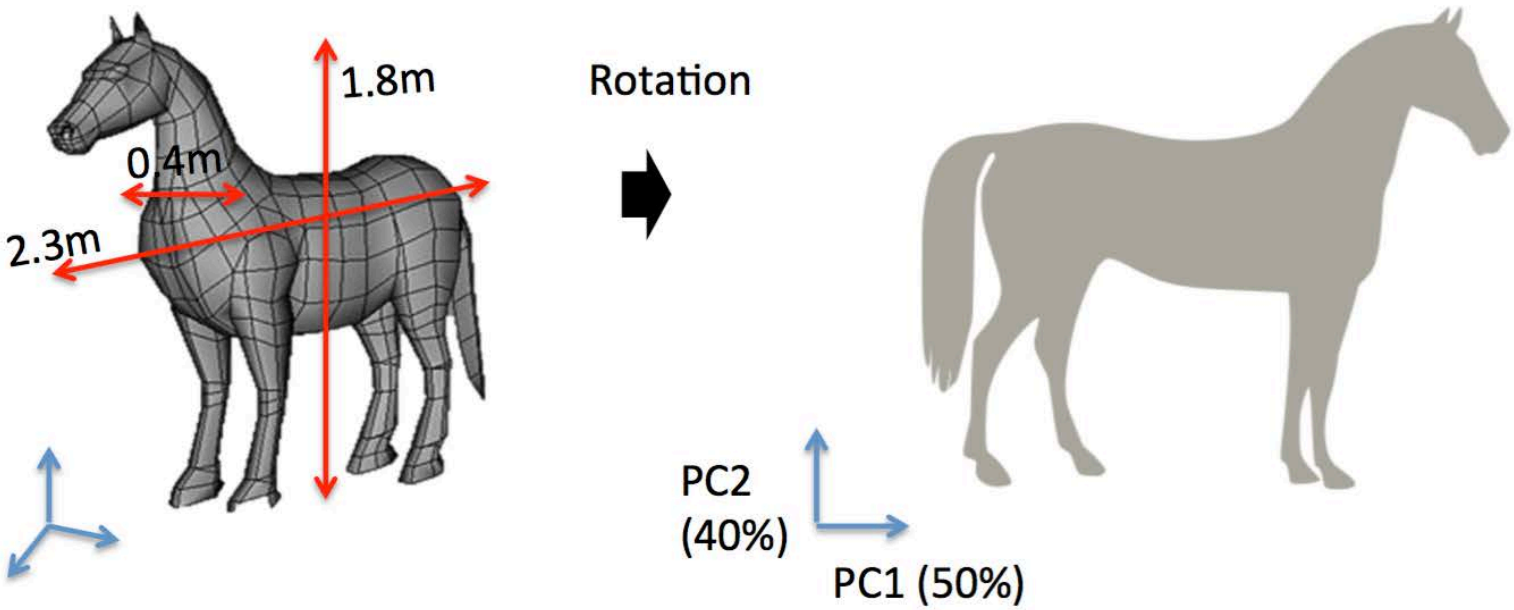


?

Principle component analysis (PCA) rotation in multi dimensional space

Orthogonal linear transformation of the data to a new coordinate system such that the greatest variance comes to lie on the first coordinate (first PCA component) and the second greatest variance on the second coordinate, so on.

Principal components = Eigenvectors of covariance matrix
Amount of contributed variance = Eigenvalues



Principle component analysis: rotation in multi dimensional space

- Useful technique for exploratory data analysis
- visualize the variation present in your dataset with many variables.

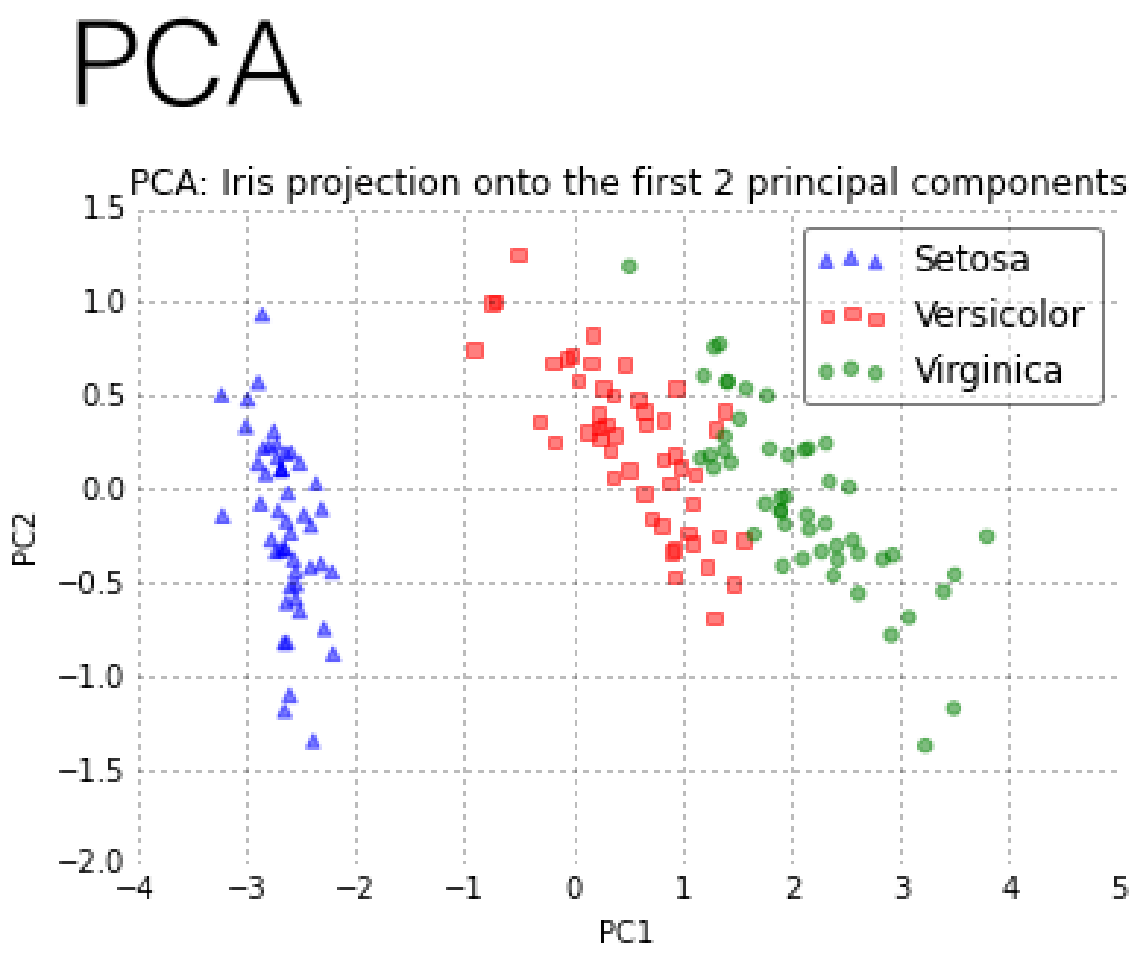
- Low dimensional (2D or 3D) representation of a high dimensional data set.

- Detect structure in features

- Pre-process for other ML algorithms

- Aids in visualization

PCA plot


PCA
PCA: Iris projection onto the first 2 principal components
Setosa
Versicolor
Virginica
PC1
PC2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−4
−3
−2
−1
0
1
2
3
4
5

Heatmaps and Clustering

Heatmap :

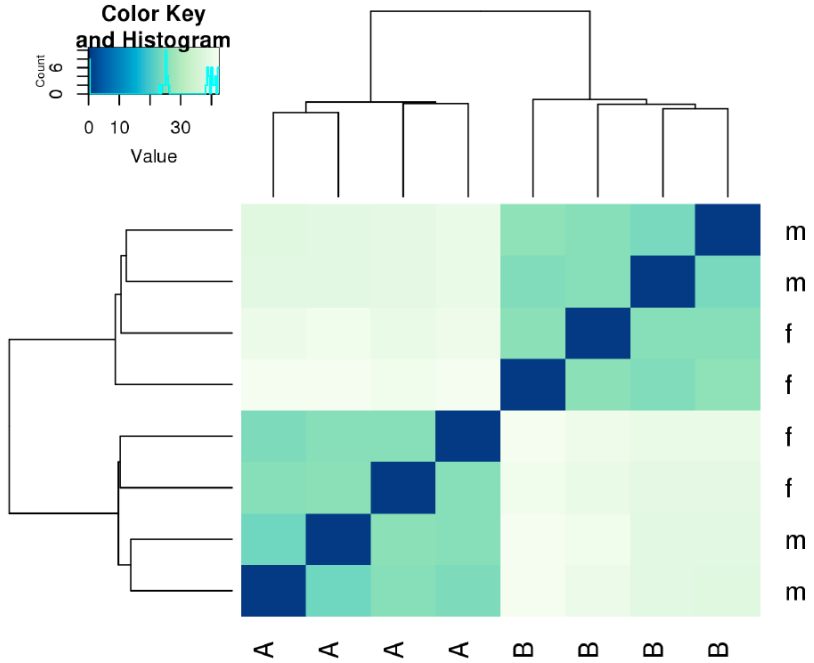
- visualizes the clustered pairwise distances between samples / genes via a false color representation

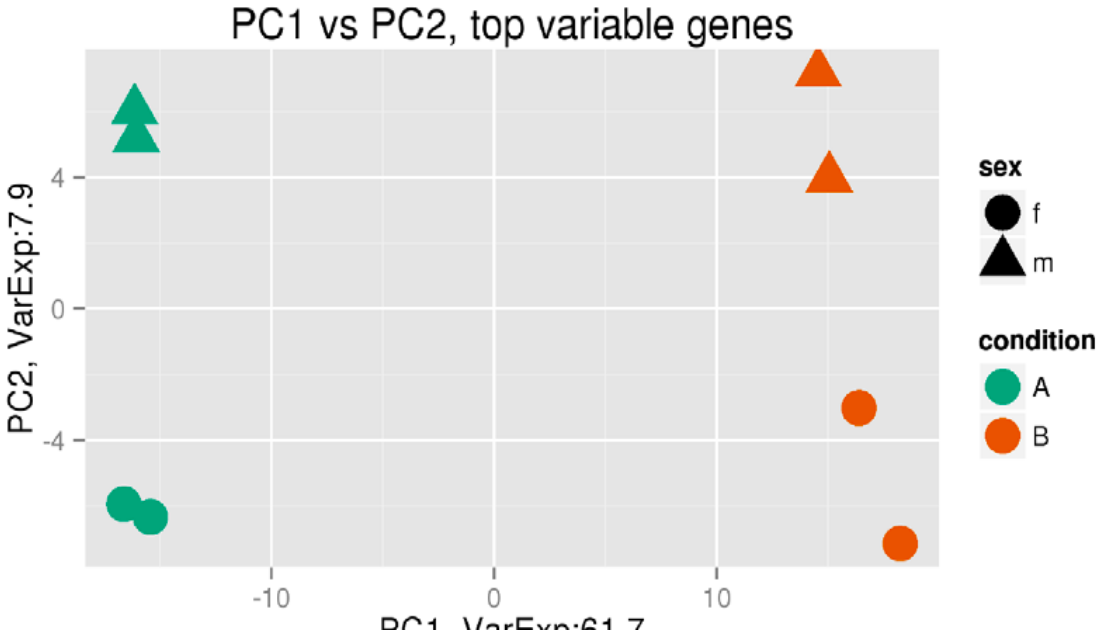
Hierarchical clustering:

- starts with as many clusters as there are samples
- successively merges samples that are close to each other
- merging process is visualized as a tree like graphic –a dendogramm

Heatmaps vs PCA

- Heatmaps: visualizes the clustered pairwise distances between samples / genes via a false color representation





YOUR TURN
START WITH THE TUTORIALS