



Machine Learning for Biological Use Cases

Prof. Valentina Boeva

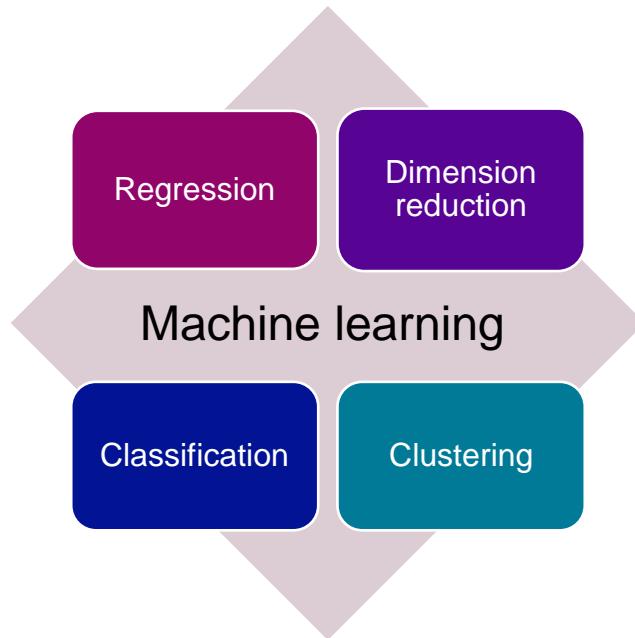
ETH Zürich, Dept. of Computer Science,
Institute for Machine Learning



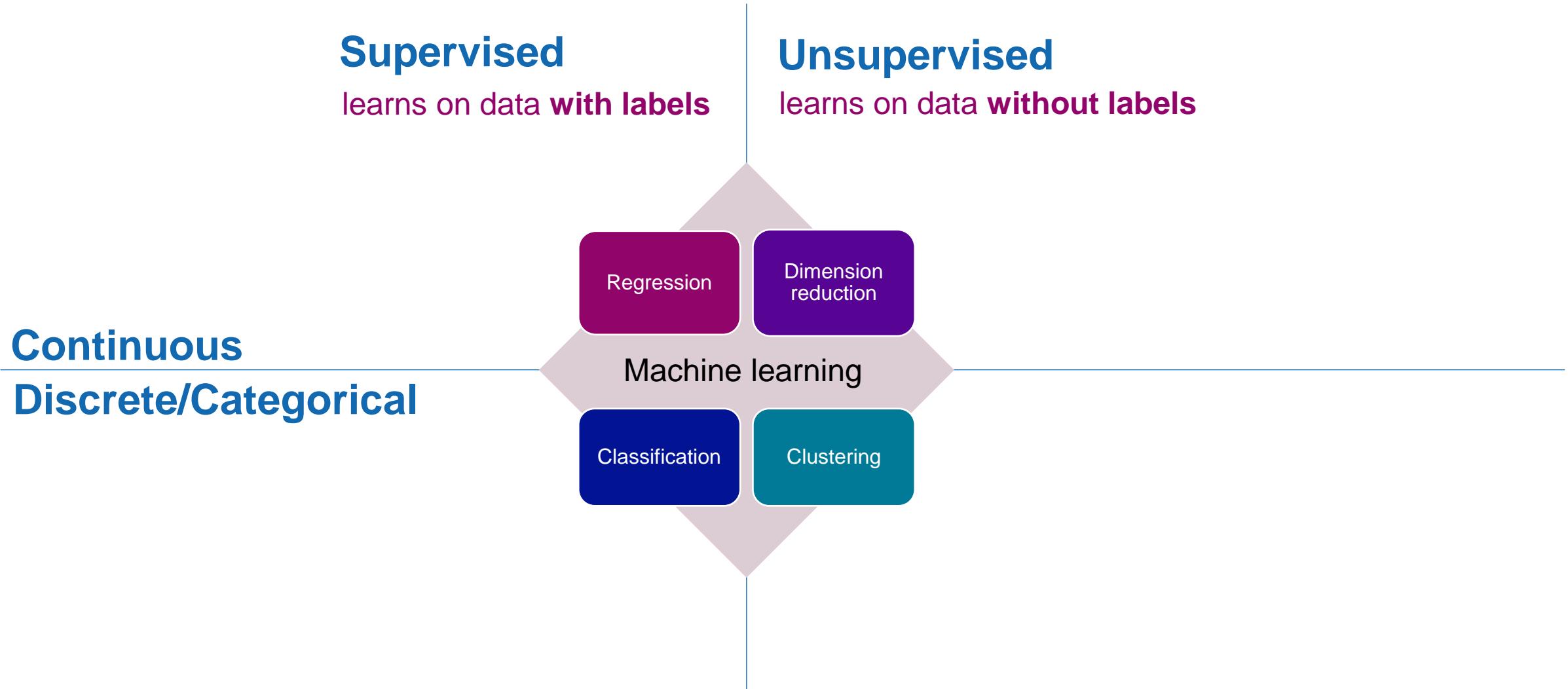
Machine learning



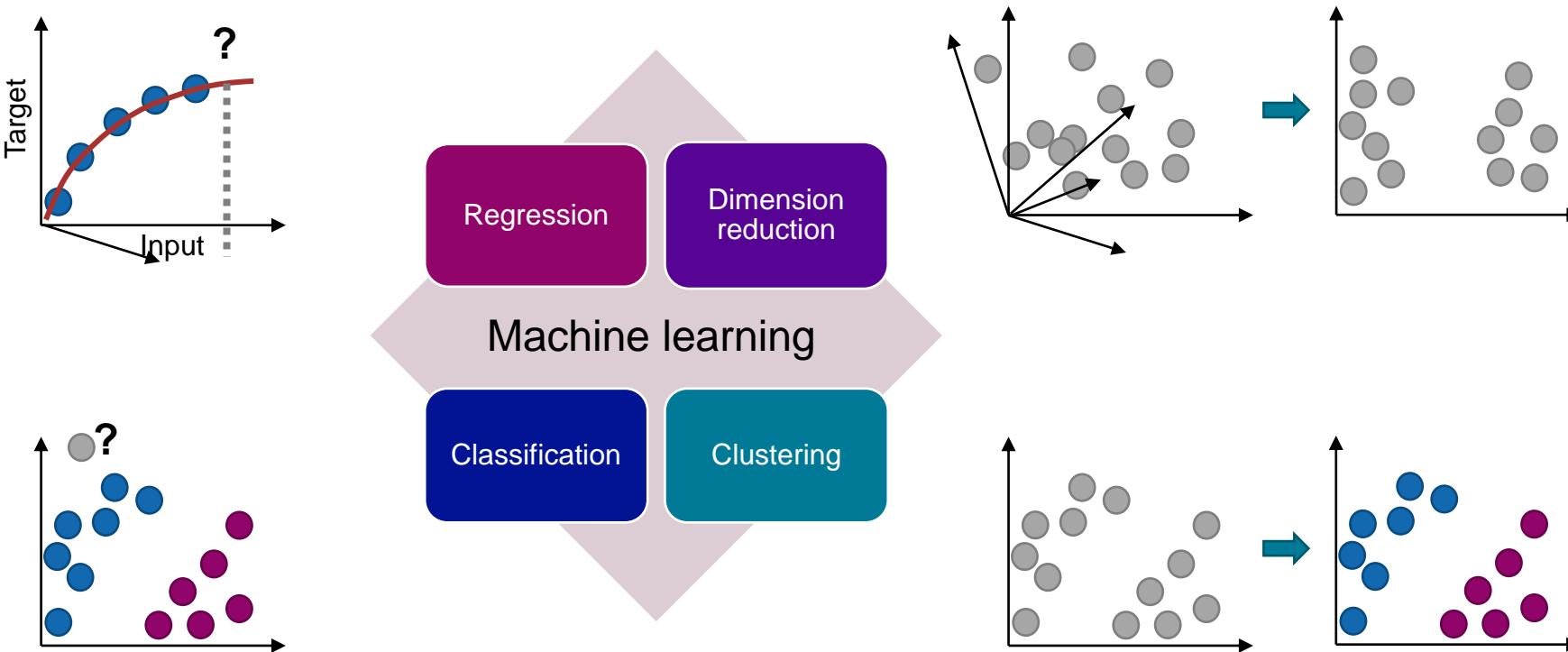
Map of classical machine learning methods



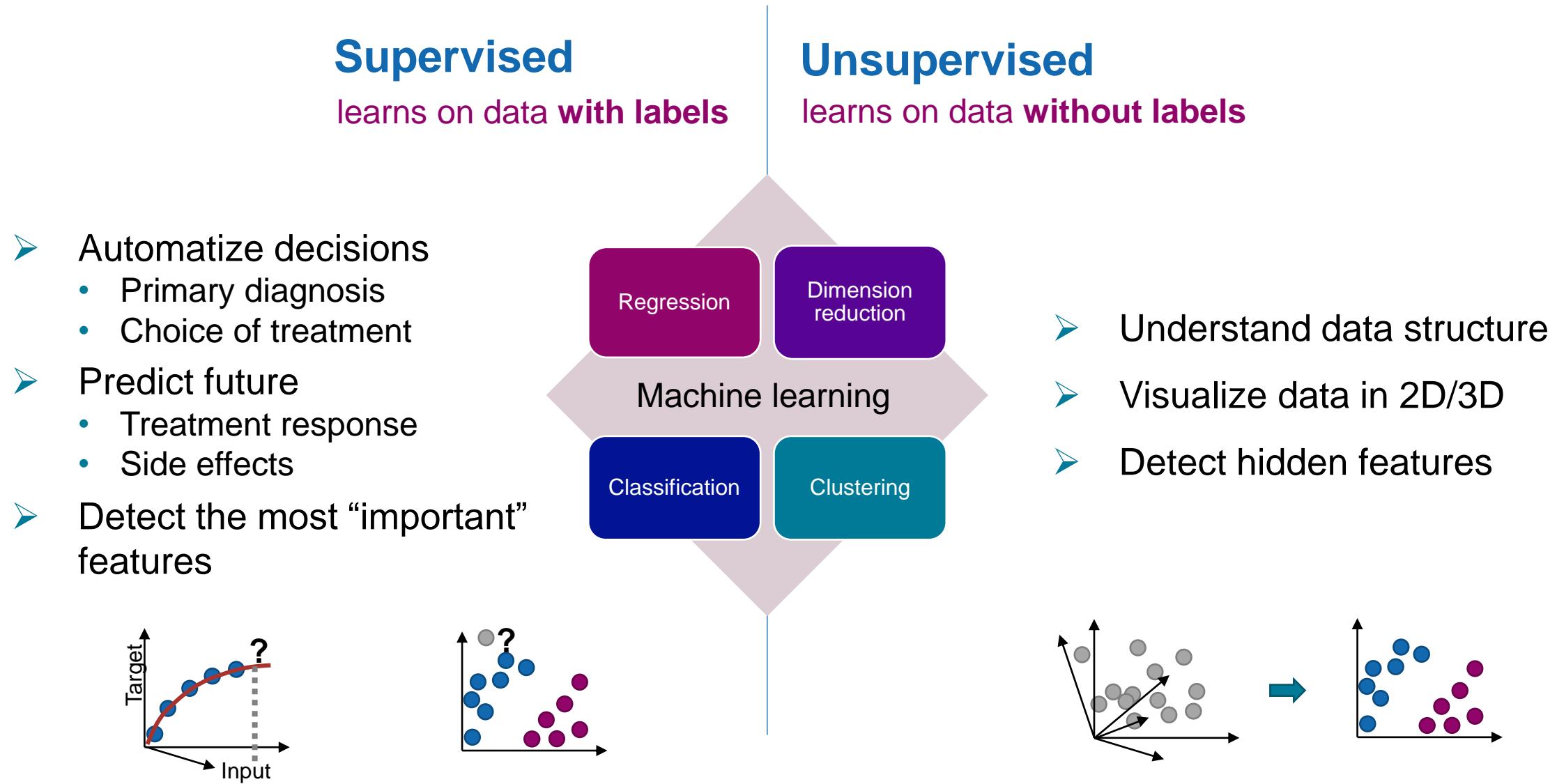
Map of classical machine learning methods



Map of classical machine learning methods



Map of classical machine learning methods



Types of input data in molecular biology and genetics

Type of data:

- **Genomic:**
 - Variants common in general population (SNPs)
 - Rare variants
 - Single nucleotide variants, SNV (a.k.a. mutations)
 - Structural aberrations (e.g., translocations, amplifications)
 - Copy number profiles
- **Transcriptomic:**
 - RNA-seq (or expression microarrays) – bulk and single cell
 - mi-RNA
- **Epigenetic:**
 - DNA methylation (sequencing or methylation arrays)
 - Histone modifications (ChIP-seq) and open chromatin (ATAC-seq, DHS-seq)
- **Proteomic:**
 - RPPA (bulk), CyTOF (single cell)

Context:

- Common diseases (Alzheimer, asthma, hypertension,...)
- Genetic syndromes (Down syndrome, CHARGE syndrome, ...)
- Cancer

Type of samples:

- Blood samples
- Saliva samples
- Tissue samples
- Maternal blood samples

Omics data are high dimensional

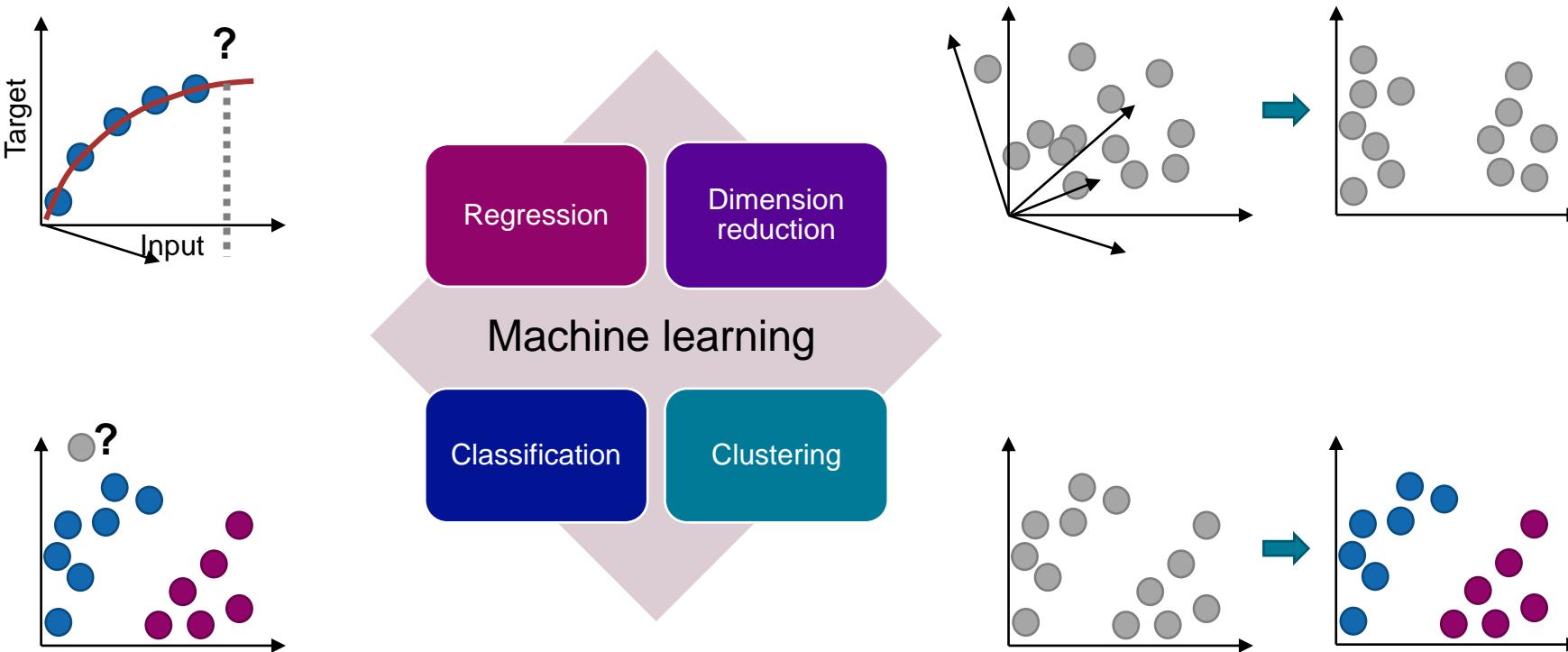
| SNPs | Mutations (SNVs) | Structural variants | Copy number alterations | mRNA expression | miRNA expression | DNA methylation |
|---------|------------------|---------------------|-------------------------|-----------------|------------------|-----------------|
| 3M-100M | ~10-10K | ~1-1000 | ~10-25K | ~25K | ~1000 | 27K-28M |

+ sometimes these data is complemented with proteomics data (expression of hundreds of proteins)

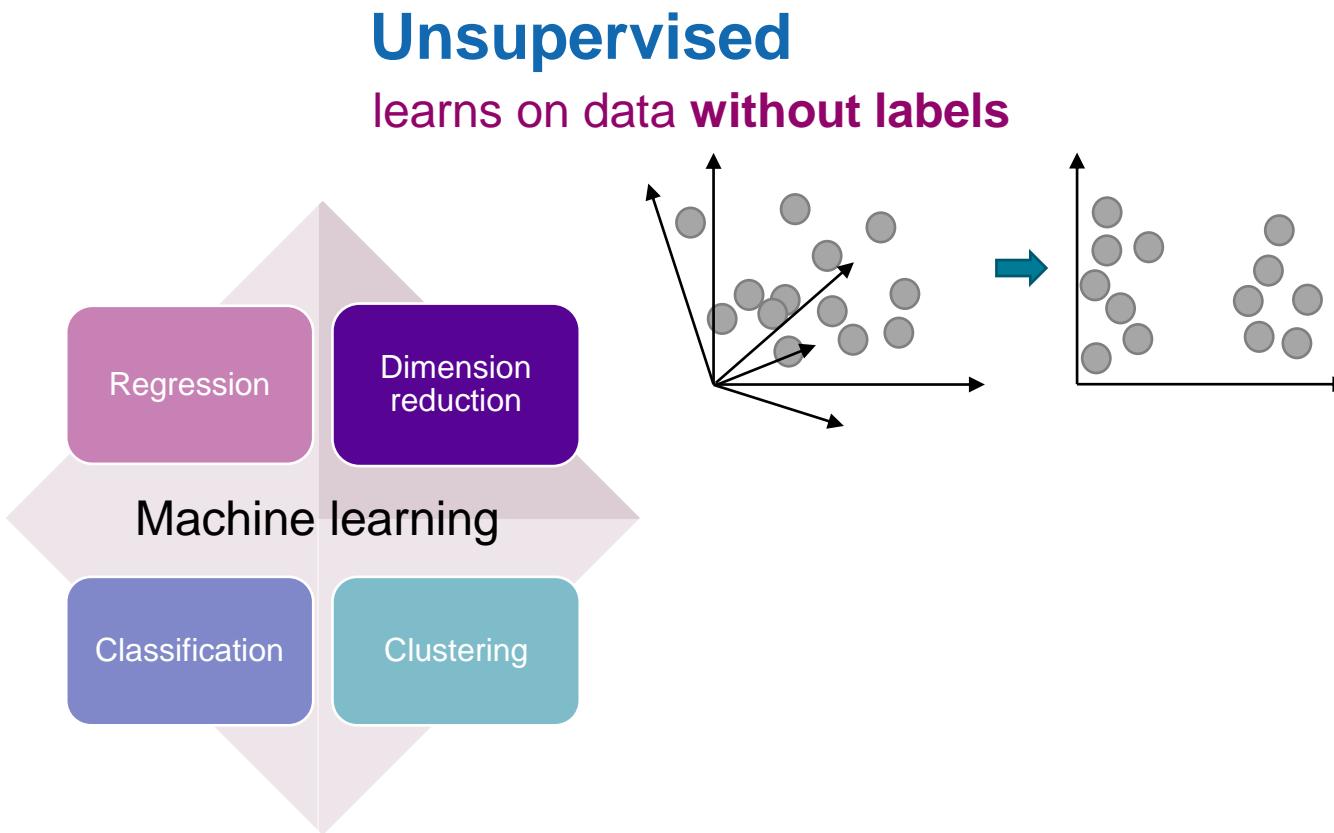
Full –omics dataset millions of observations per patient:

Great challenge to avoid over-fitting and perform feature selection!

Map of classical machine learning methods

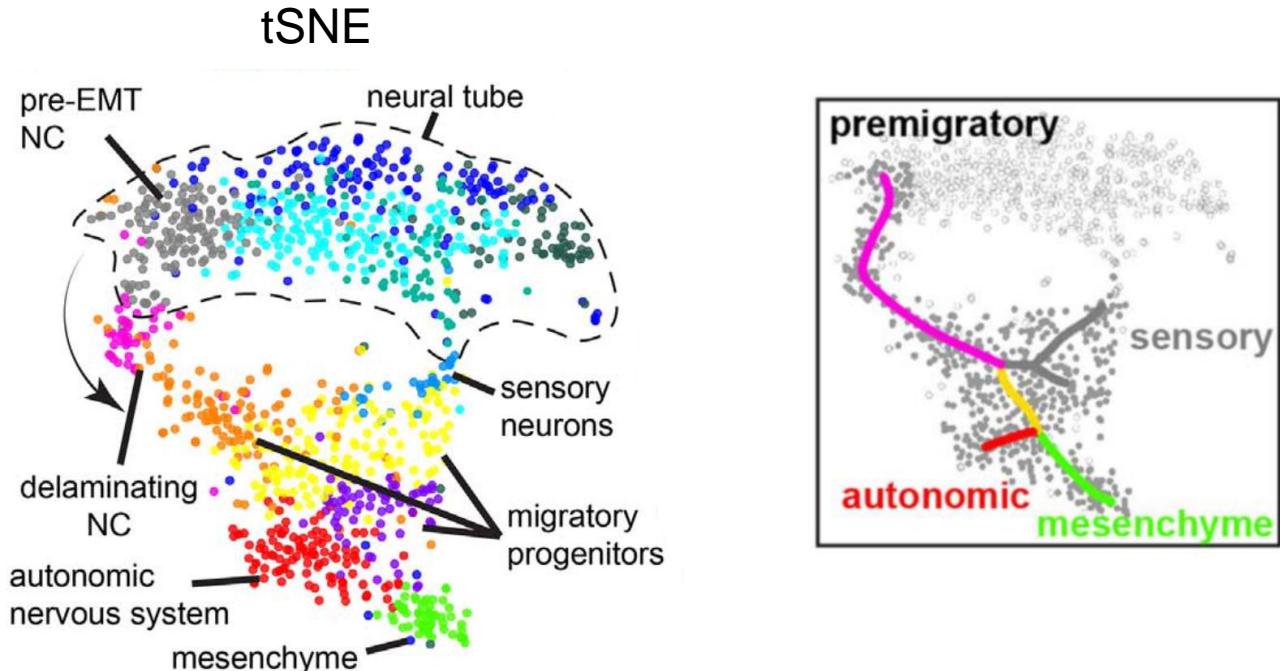


Map of machine learning methods



Examples: Dimension reduction and clustering

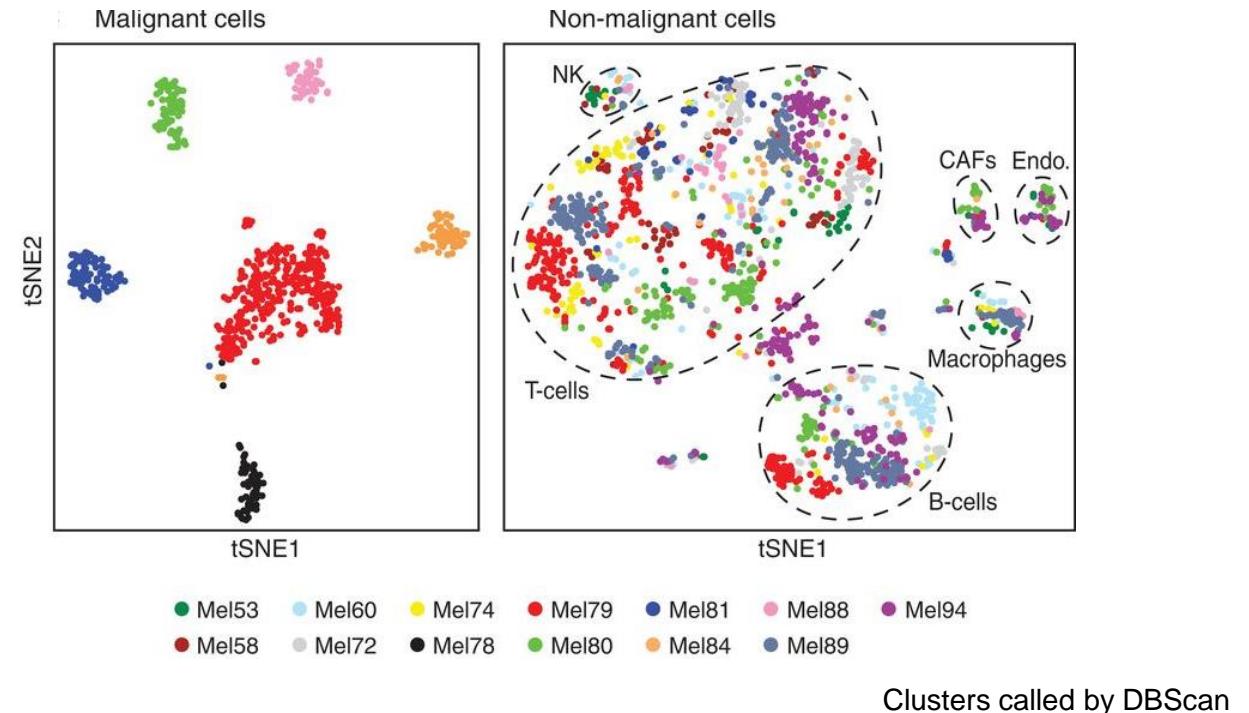
- Single cell transcriptomics:
development, cell type
heterogeneity and cancer
- Bulk transcriptomics and
epigenetics (cancer)
- Bulk transcriptomics: Effect of
mutations



Spatiotemporal structure of cell fate decisions in murine neural crest. Soldatov et al., *Science* 364, 971 (2019)

Examples: Dimension reduction and clustering

- Single cell transcriptomics:
development, cell type
heterogeneity and cancer
- Bulk transcriptomics and
epigenetics (cancer)
- Bulk transcriptomics: Effect of
mutations

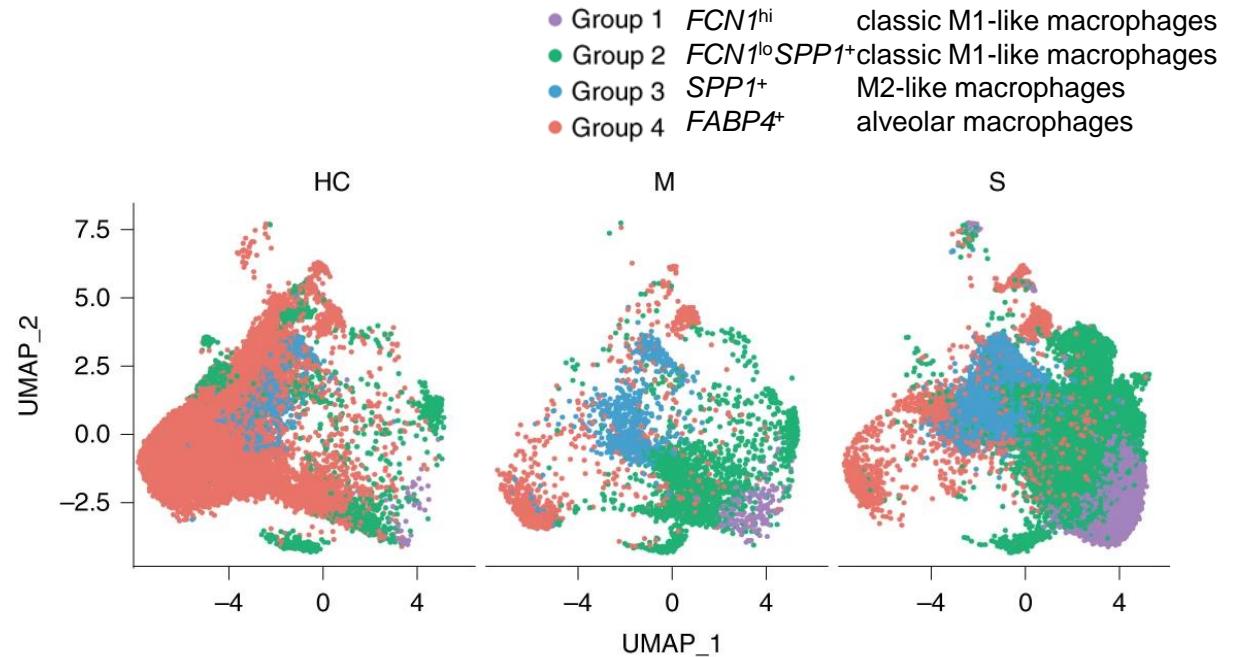


Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Tirosh et al. *Science*. 2016 Apr 8;352(6282):189-96.

Examples: Dimension reduction and clustering

- Single cell transcriptomics:
development, cell type
heterogeneity and cancer
- Bulk transcriptomics and
epigenetics (cancer)
- Bulk transcriptomics: Effect of
mutations

Bronchoalveolar lavage fluid macrophages from patients with varying severity of COVID-19 and from healthy people:

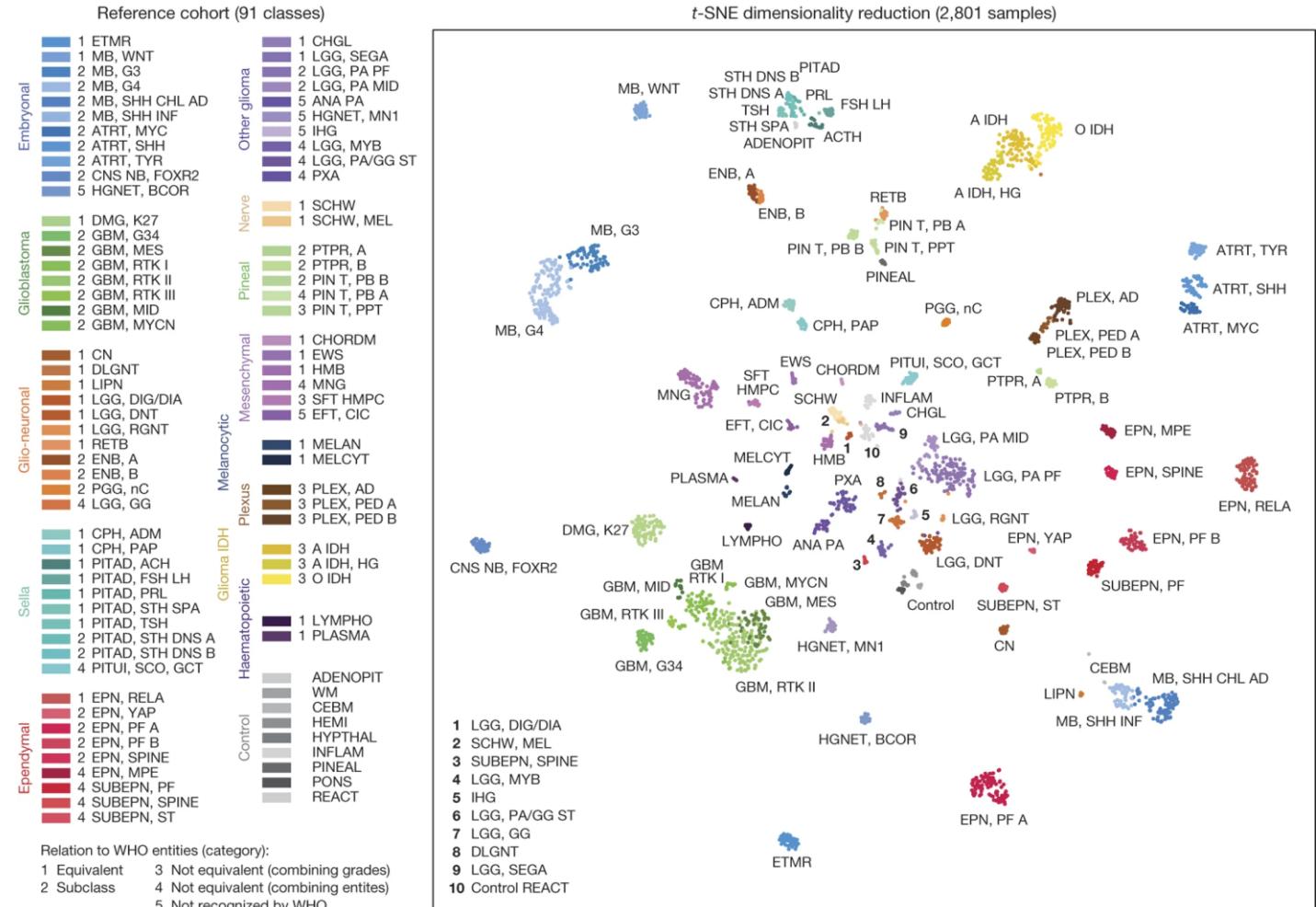


Severe cases: Presence of proinflammatory monocyte-derived macrophages

Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Liao et al., *Nature Medicine*. 2020

Examples: Dimension reduction and clustering

- Single cell transcriptomics: development, cell type heterogeneity and cancer



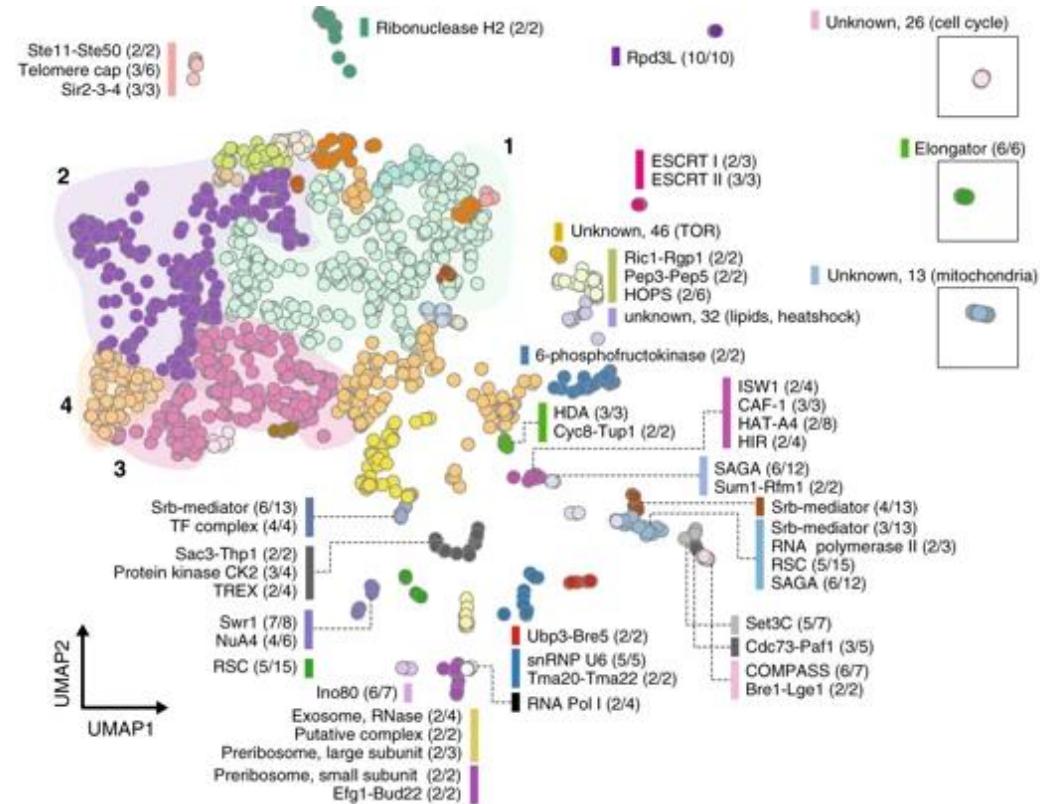
DNA methylation-based classification of central nervous system tumours. Capper et al. *Nature* 2018

- Bulk transcriptomics: Effect of mutations

Examples: Dimension reduction and clustering

- Single cell transcriptomics:
development, cell type
heterogeneity and cancer
- Bulk transcriptomics and
epigenetics (cancer)
- Bulk transcriptomics: Effect of
mutations

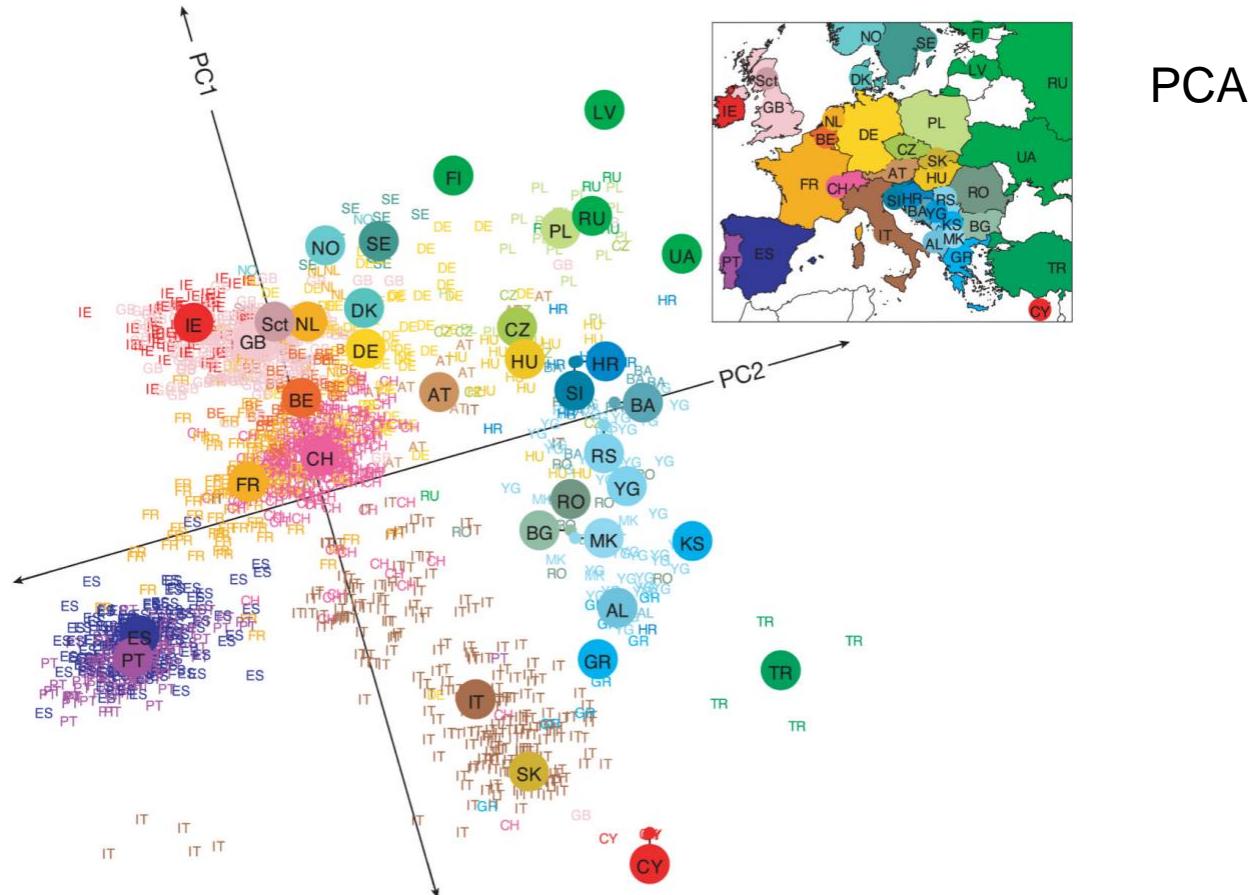
Transcript profiles of 1484 single gene deletions of *Saccharomyces cerevisiae* (baker's yeast)



Dimensionality reduction by UMAP to visualize physical and genetic interactions. Michael W. Dorrity et al., *Nature Comm.* 2020

Examples: Dimension reduction and clustering

- Population genetics

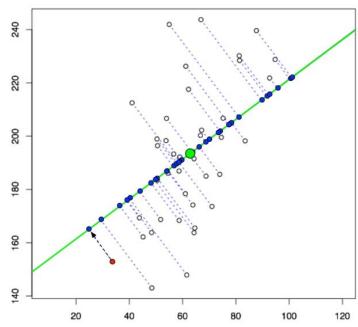


«Genes mirror geography within Europe» Novembre et al, *Nature*, 2008

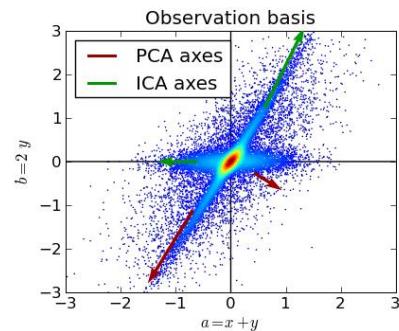
Methods for dimension reduction

Most widely used methods:

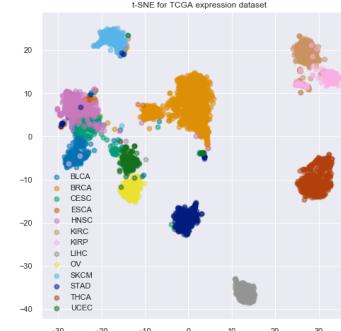
- Principal component analysis (PCA)
- Independent component analysis (ICA)
- t-distributed Stochastic Neighbor Embedding (tSNE) – check!
- Uniform Manifold Approximation and Projection (UMAP) – check!



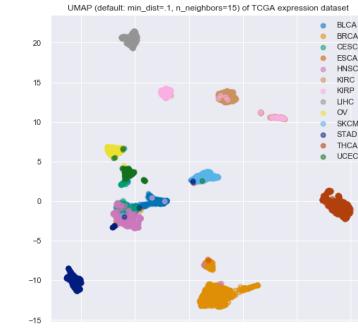
PCA



ICA



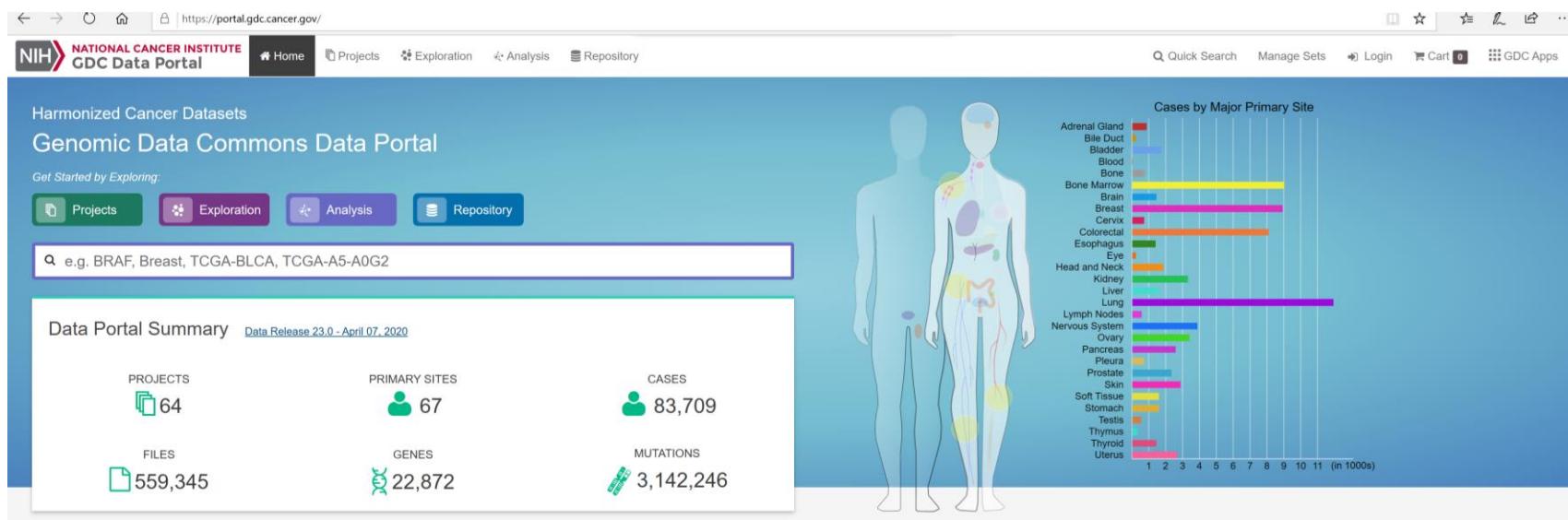
t-SNE



UMAP

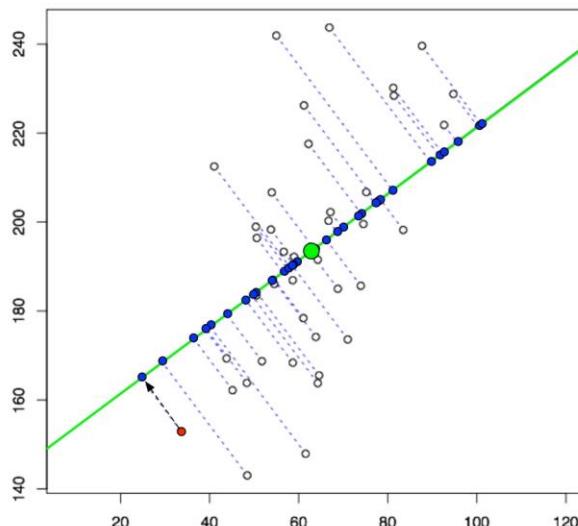
Hands-on: gene expression data from several cancer types

- <https://github.com/BoevaLab/Teaching>
- Input: The Cancer Genome Atlas (TCGA) mRNA expression data



Principal component analysis (PCA)

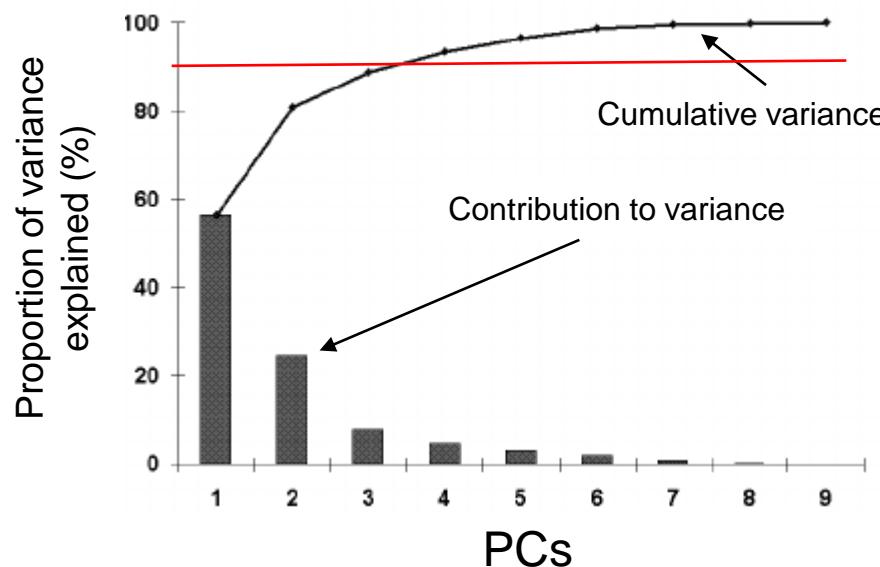
- PCA: an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.



Given n points in \mathbb{R}^p , principal components analysis consists of choosing a dimension $k < p$ and then finding the affine space of dimension k with the property that the squared distance of the points to their orthogonal projection onto the space is minimized.

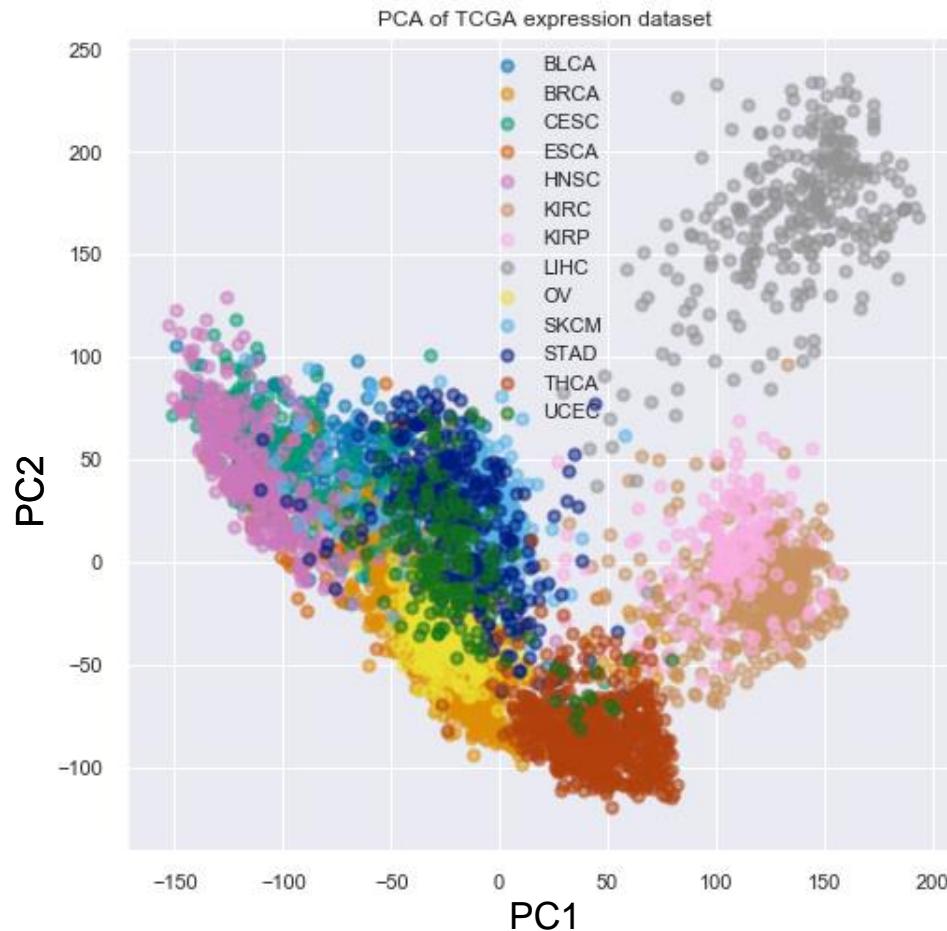
Principal component analysis (PCA)

- PCA is a deterministic method, the only parameter one can choose is k for how many principal components to keep.
- Choosing k based on the proportion of the variance explained.



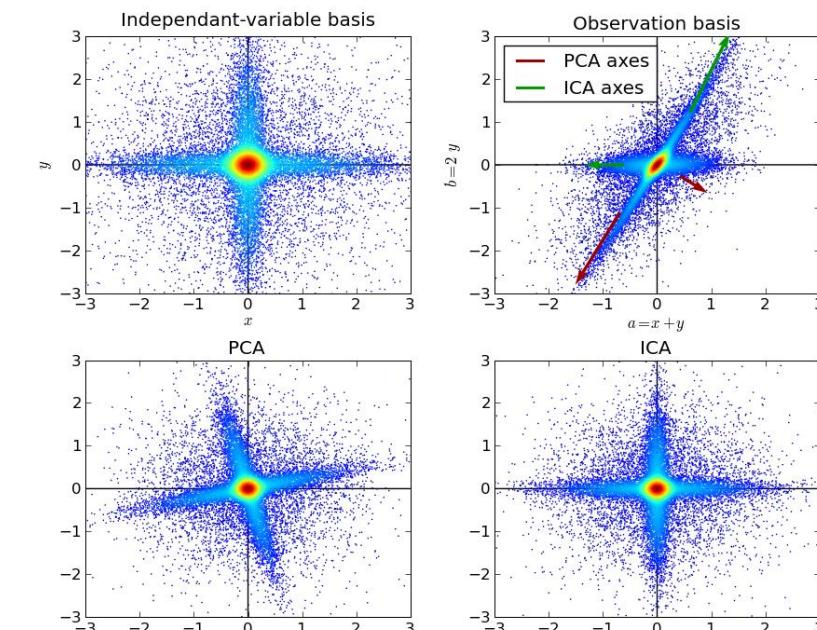
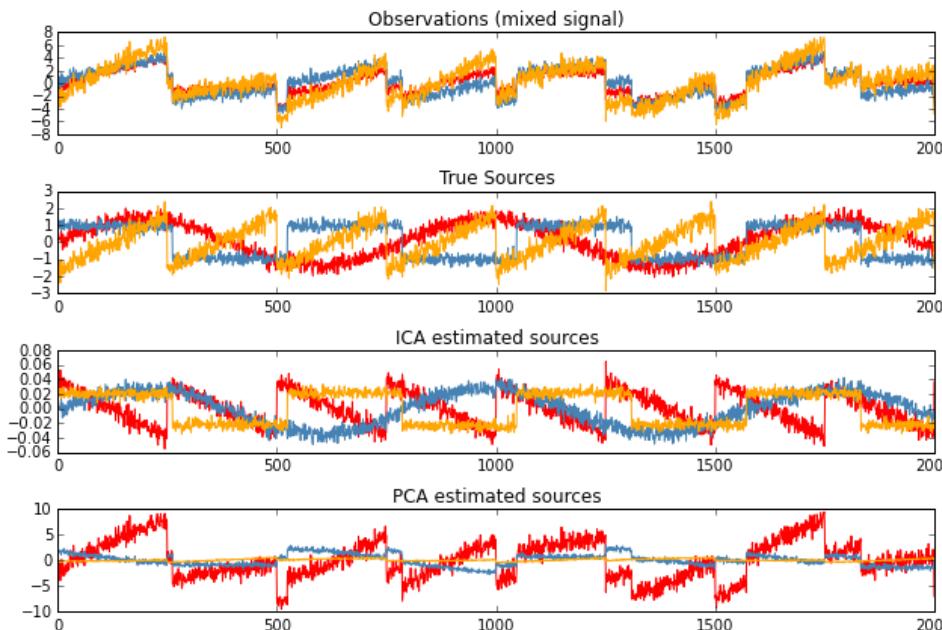
$Var_explained_k = \frac{\lambda_k}{\sum_i \lambda_i}$, where λ_k is k^{th} eigenvalue.

Let's go to the Jupiter Notebook to see the result of PCA on our toy data set



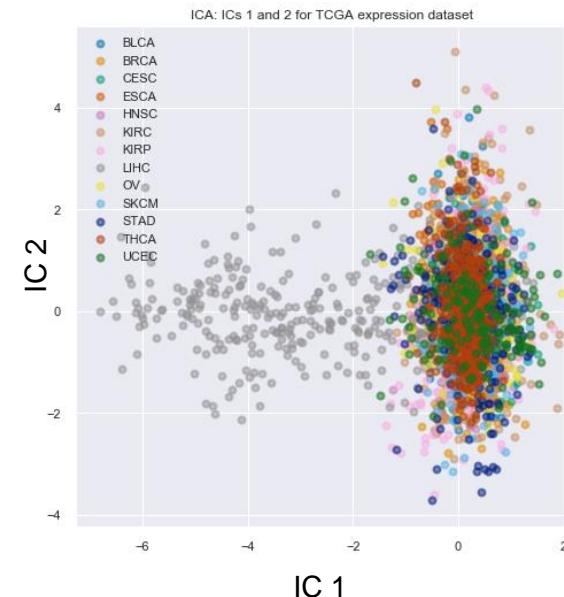
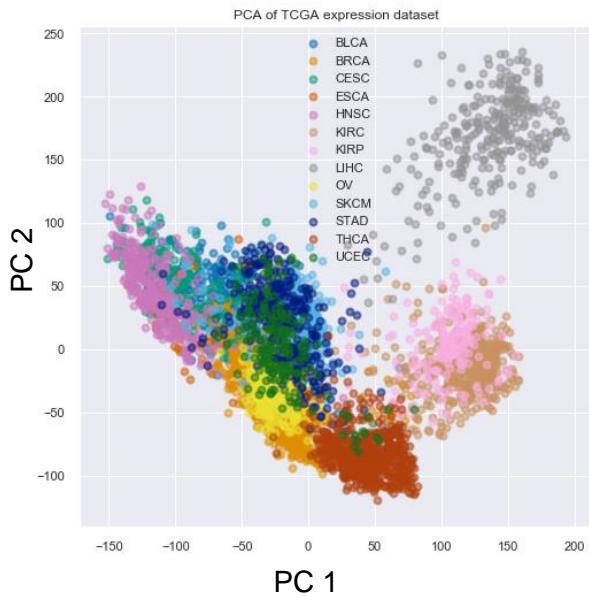
Independent component analysis (ICA)

- ICA: a method for separating a multivariate signal into additive subcomponents: assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other.



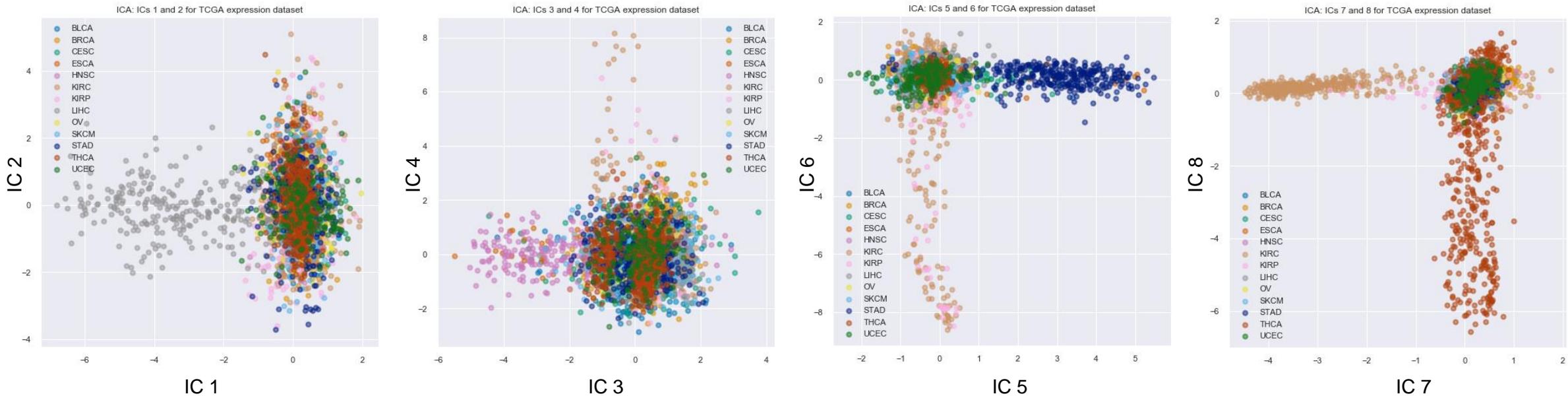
Independent component analysis (ICA)

- ICA: a method for separating a multivariate signal into additive subcomponents: assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other.



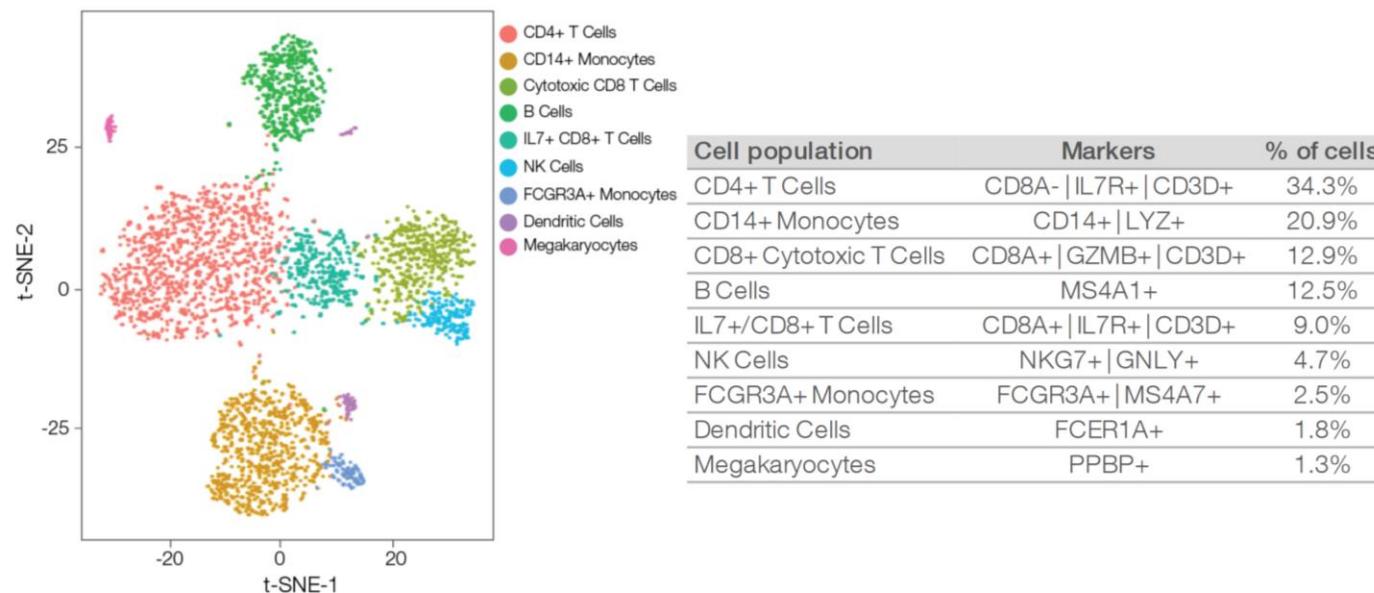
Independent component analysis (ICA)

- ICA: a method for separating a multivariate signal into additive subcomponents: assuming that the subcomponents are non-Gaussian signals and that they are statistically independent from each other.



t-distributed Stochastic Neighbor Embedding (tSNE)

- tSNE: nonlinear dimensionality reduction technique, converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.



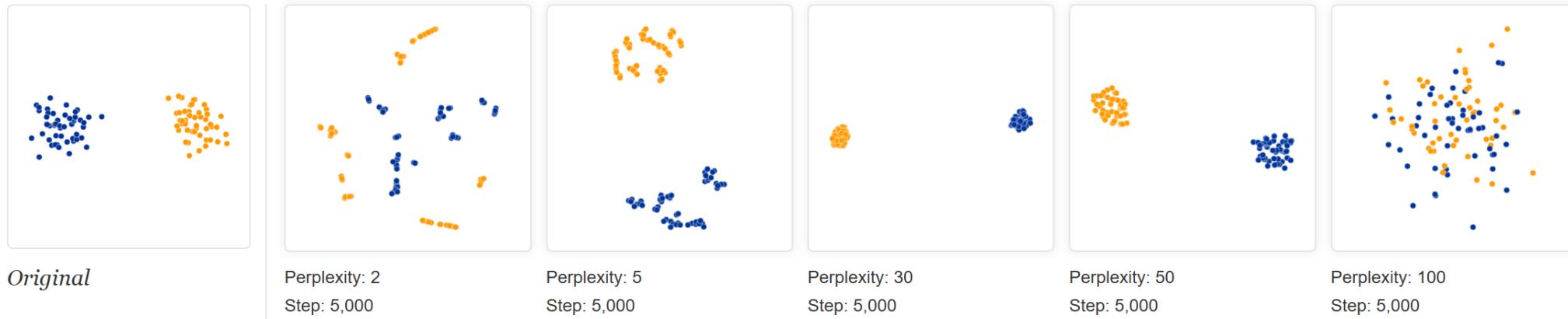
- t-SNE has a cost function that is not convex, i.e., with different initializations we can get different results.

[First introduced by van der Maaten & Hinton paper from 2008.](#)

t-distributed Stochastic Neighbor Embedding (tSNE)

tSNE method's the most important parameter:

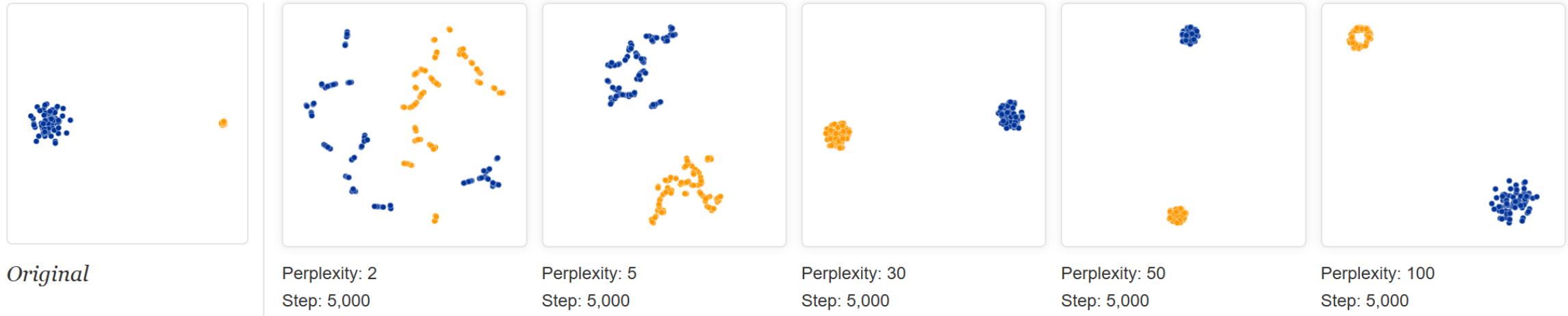
- **Perplexity**, scikit-learn recommended range: [5, 50], default: 30



t-distributed Stochastic Neighbor Embedding (tSNE)

tSNE method's particularities:

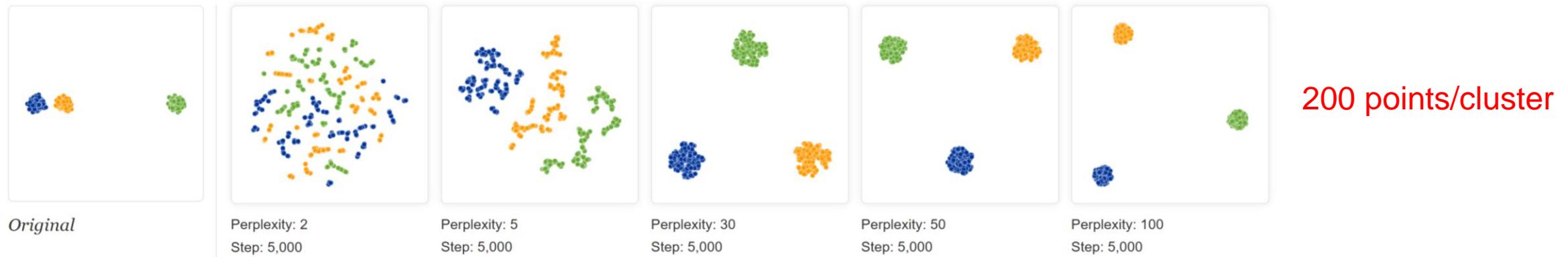
- Cluster sizes in a t-SNE plot mean nothing



t-distributed Stochastic Neighbor Embedding (tSNE)

tSNE method's particularities:

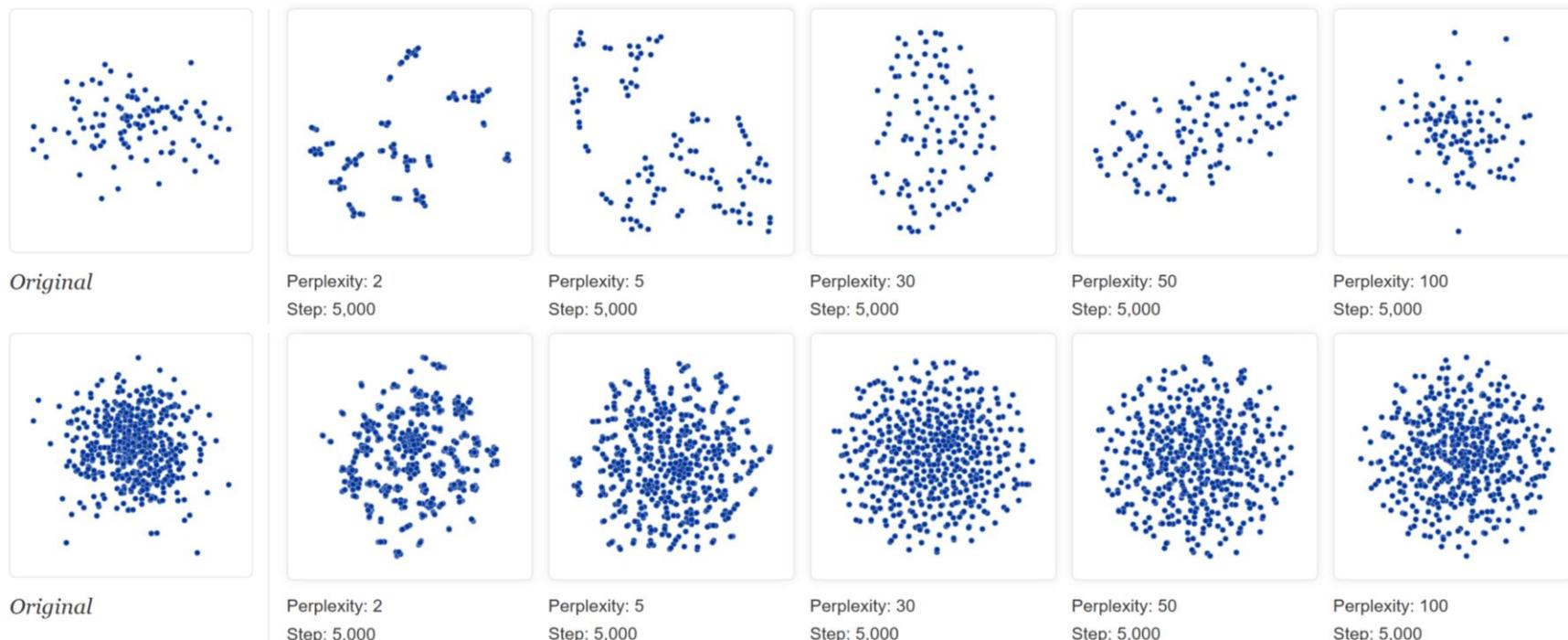
- Cluster sizes in a t-SNE plot mean nothing
- Distances between clusters might not mean anything



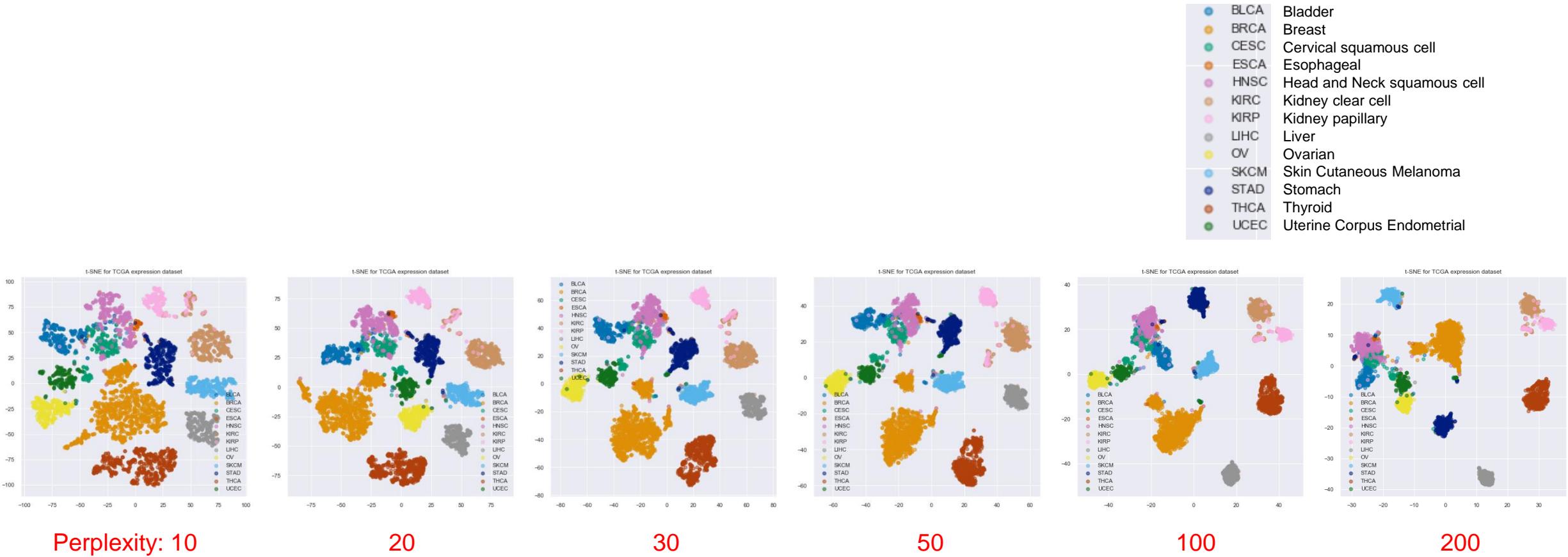
t-distributed Stochastic Neighbor Embedding (tSNE)

tSNE method's particularities:

- Cluster sizes in a t-SNE plot mean nothing
- Distances between clusters might not mean anything
- Random noise does not always look random, and sometimes one can see some shapes

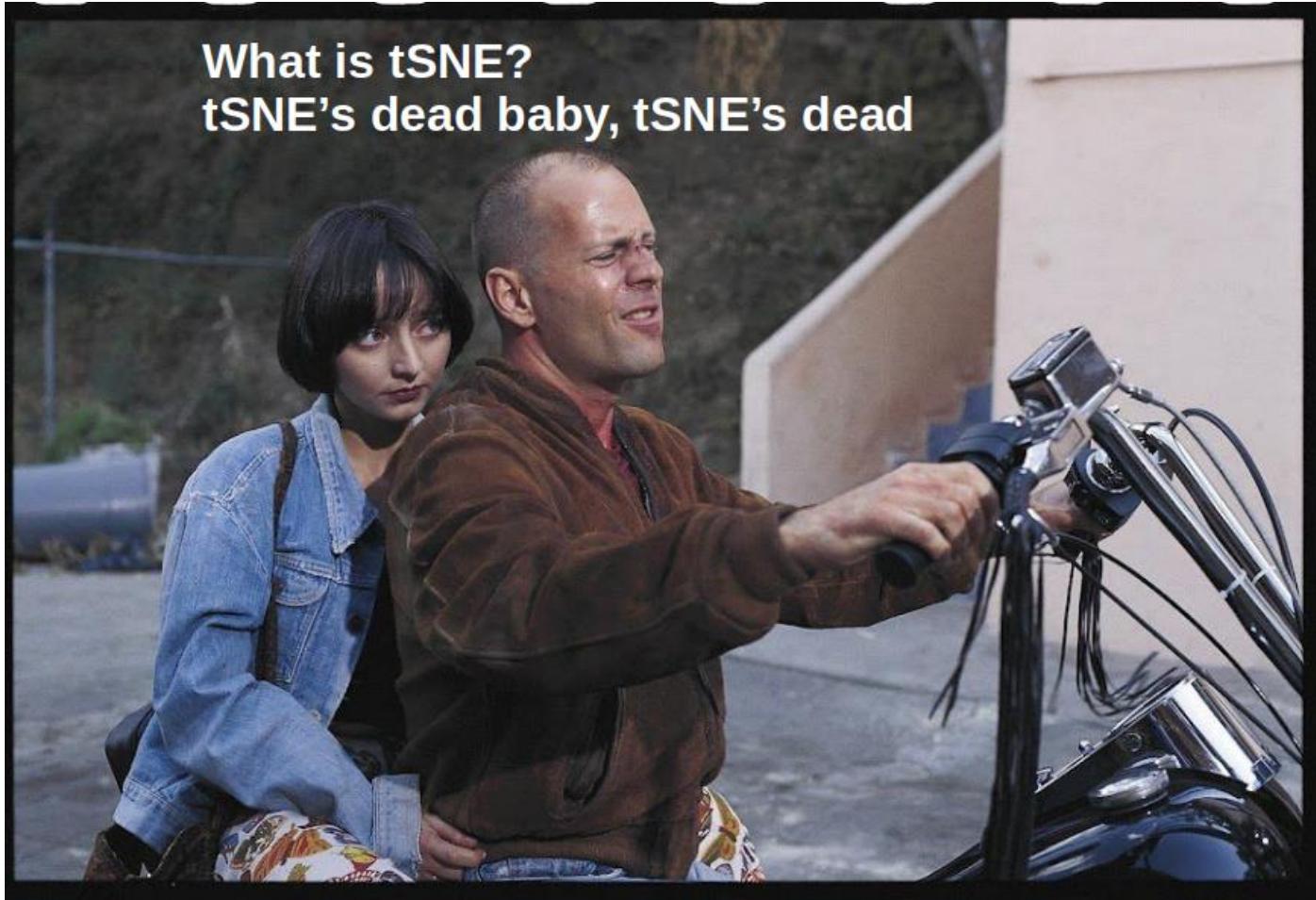


What would you choose as the best value of perplexity in this example?



Cancer type abbreviations: <https://gdc.cancer.gov/resources-tcga-users/tcgacode-tables/tcgastudy-abbreviations>

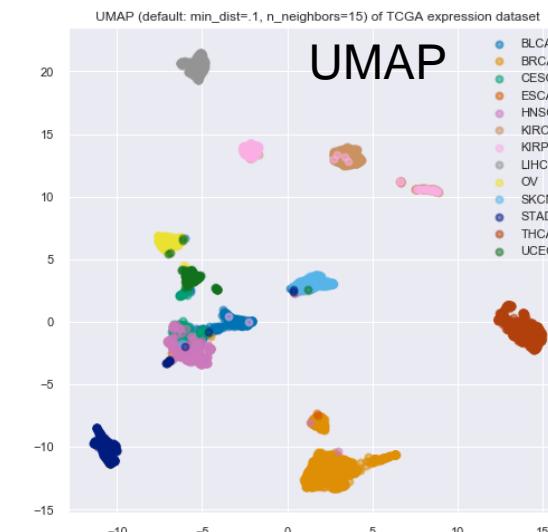
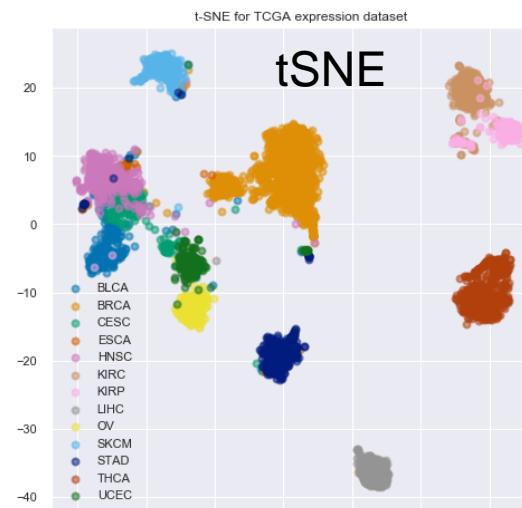
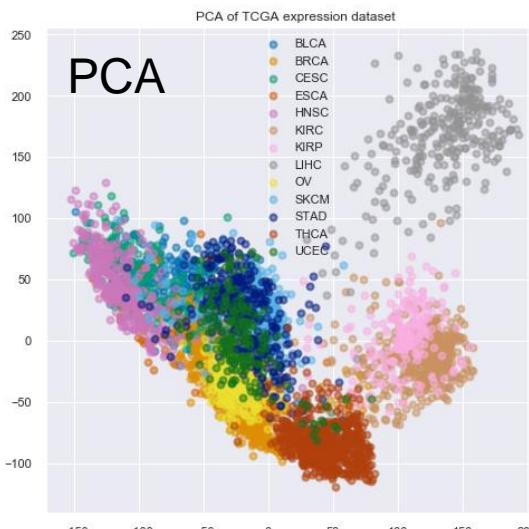
Uniform Manifold Approximation and Projection (UMAP)



Uniform Manifold Approximation and Projection (UMAP)

UMAP: nonlinear dimensionality reduction technique. Idea is similar to tSNE, but

- Much faster
- Not limited to the first 2-3 dimensions
- Uses binary cross-entropy as a cost function instead of the KL-divergence
- Preserves global structure
- Uses the number of nearest neighbors instead of perplexity



First introduced by [McInnes, L, Healy, J, ArXiv e-prints 1802.03426, 2018](#)

Uniform Manifold Approximation and Projection (UMAP)

UMAP's the most important parameters:

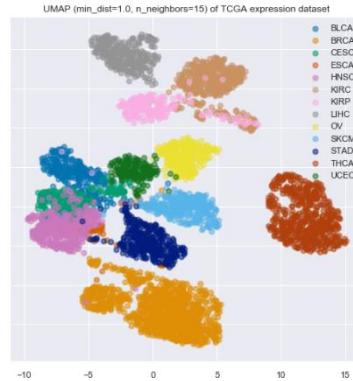
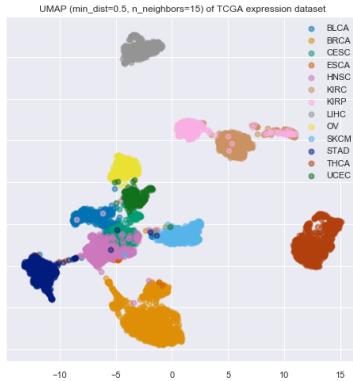
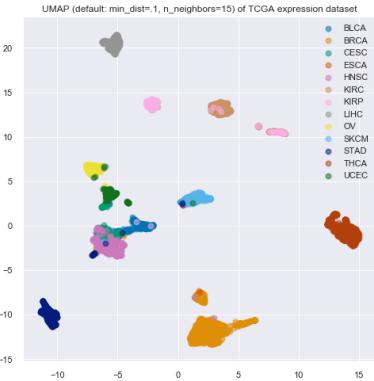
min_dist, recommended range: [0.1, 1], default: 0.1

- Controls how tightly the embedding is allowed compress points together. Larger values ensure embedded points are more evenly distributed, while smaller values allow the algorithm to optimise more accurately with regard to local structure. Sensible values are in the range 0.001 to 0.5, with 0.1 being a reasonable default.

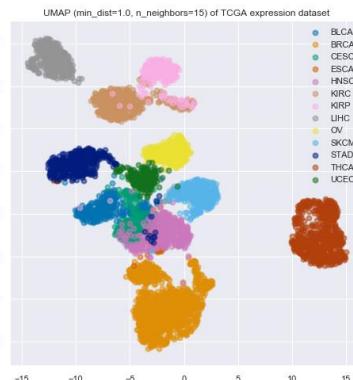
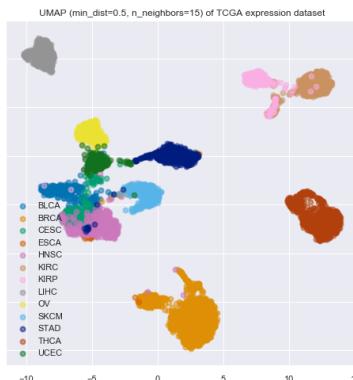
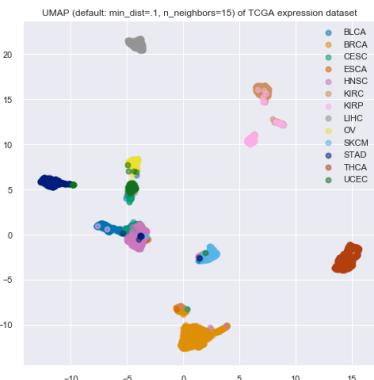
n_neighbors, recommended range: [2, 100], default: 15

- Determines the number of neighboring points used in local approximations of manifold structure. Larger values will result in more global structure being preserved at the loss of detailed local structure. In general, this parameter should often be in the range 5 to 50, with a choice of 10 to 15 being a sensible default.

What would you choose as the best value of `min_dist` and `n_neighbors` for this example?



`n_neighbors: 15`



`n_neighbors: 50`

| | |
|-------------------|-----------------------------|
| <code>BLCA</code> | Bladder |
| <code>BRCA</code> | Breast |
| <code>CESC</code> | Cervical squamous cell |
| <code>ESCA</code> | Esophageal |
| <code>HNSC</code> | Head and Neck squamous cell |
| <code>KIRC</code> | Kidney clear cell |
| <code>KIRP</code> | Kidney papillary |
| <code>LIHC</code> | Liver |
| <code>OV</code> | Ovarian |
| <code>SKCM</code> | Skin Cutaneous Melanoma |
| <code>STAD</code> | Stomach |
| <code>THCA</code> | Thyroid |
| <code>UCEC</code> | Uterine Corpus Endometrial |

`Min_dist: 0.1`

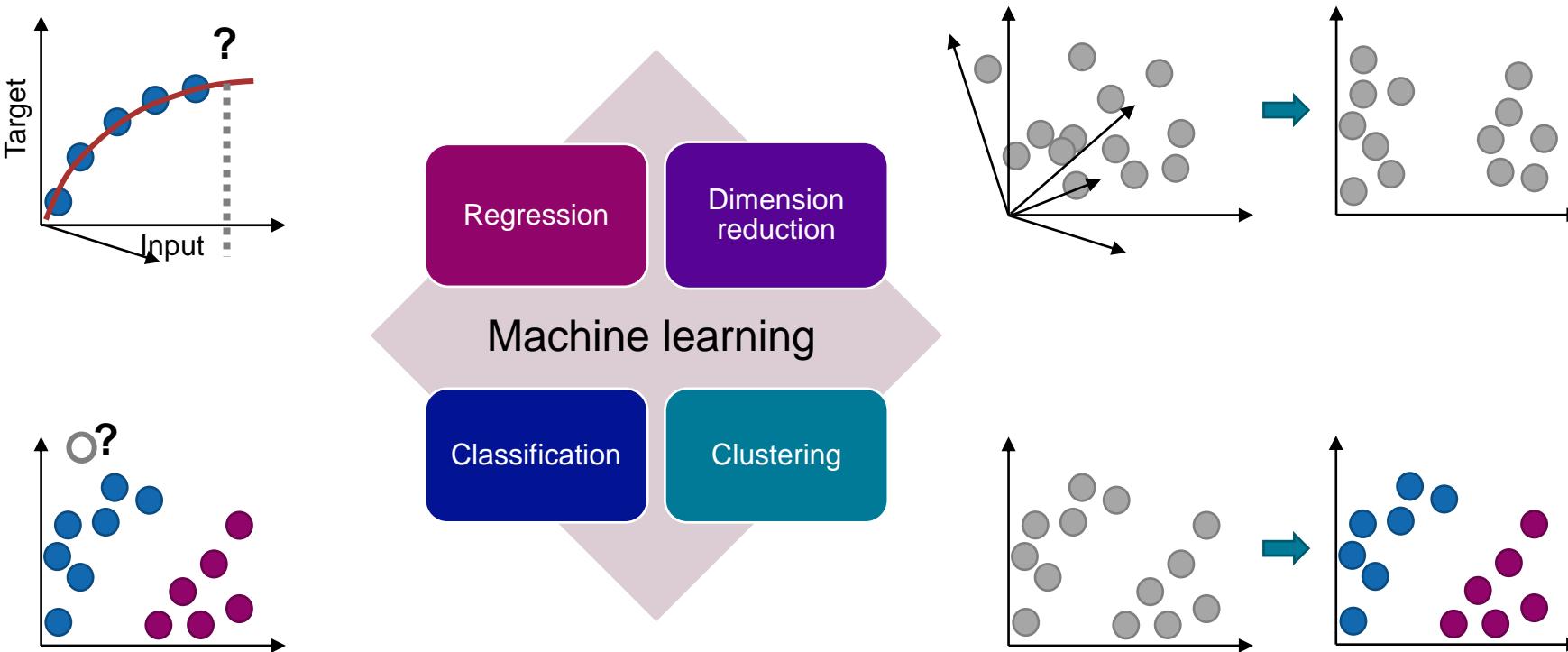
`0.5`

`1`

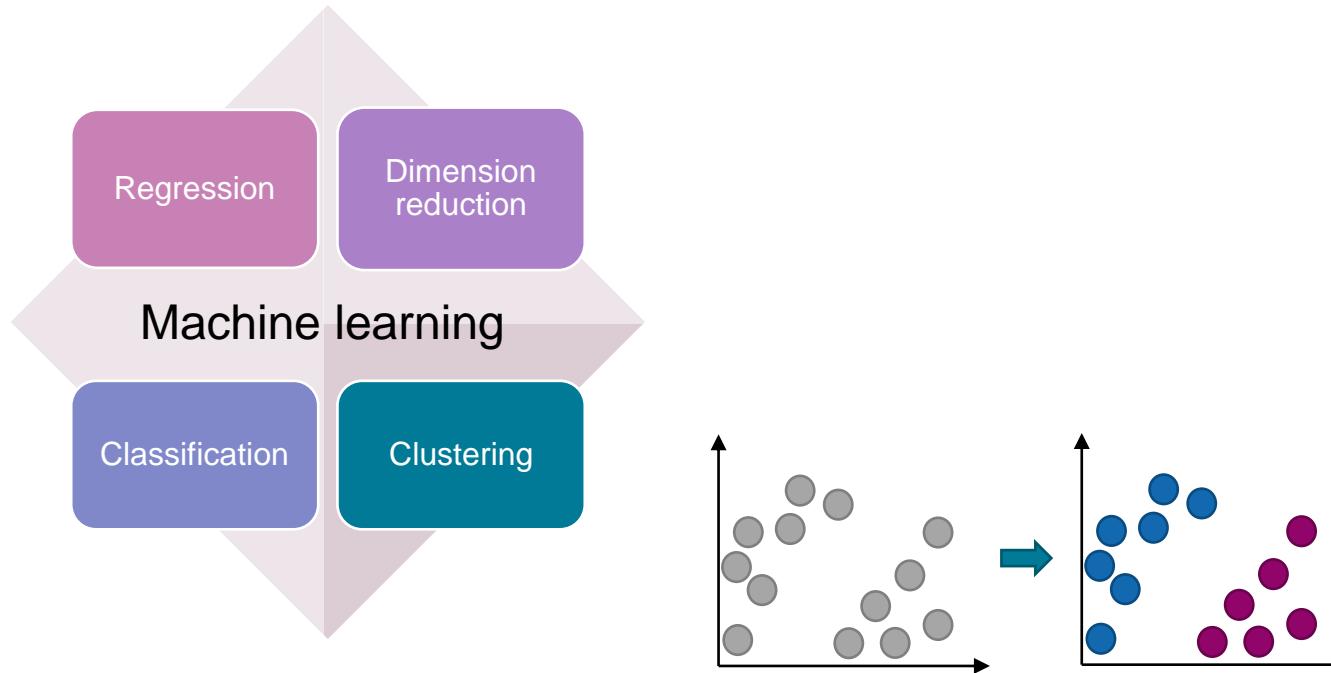
Take home message: dimensionality reduction

- One should often try different methods with different parameters to choose the one that fits the best to our expectations from the data
- Random noise does not always look random, and sometimes one can see shapes
- Projections on the first n principal components can be used as input (instead of the original X) to other dimension reduction methods such as tSNE to reduce execution time.

Map of classical machine learning methods

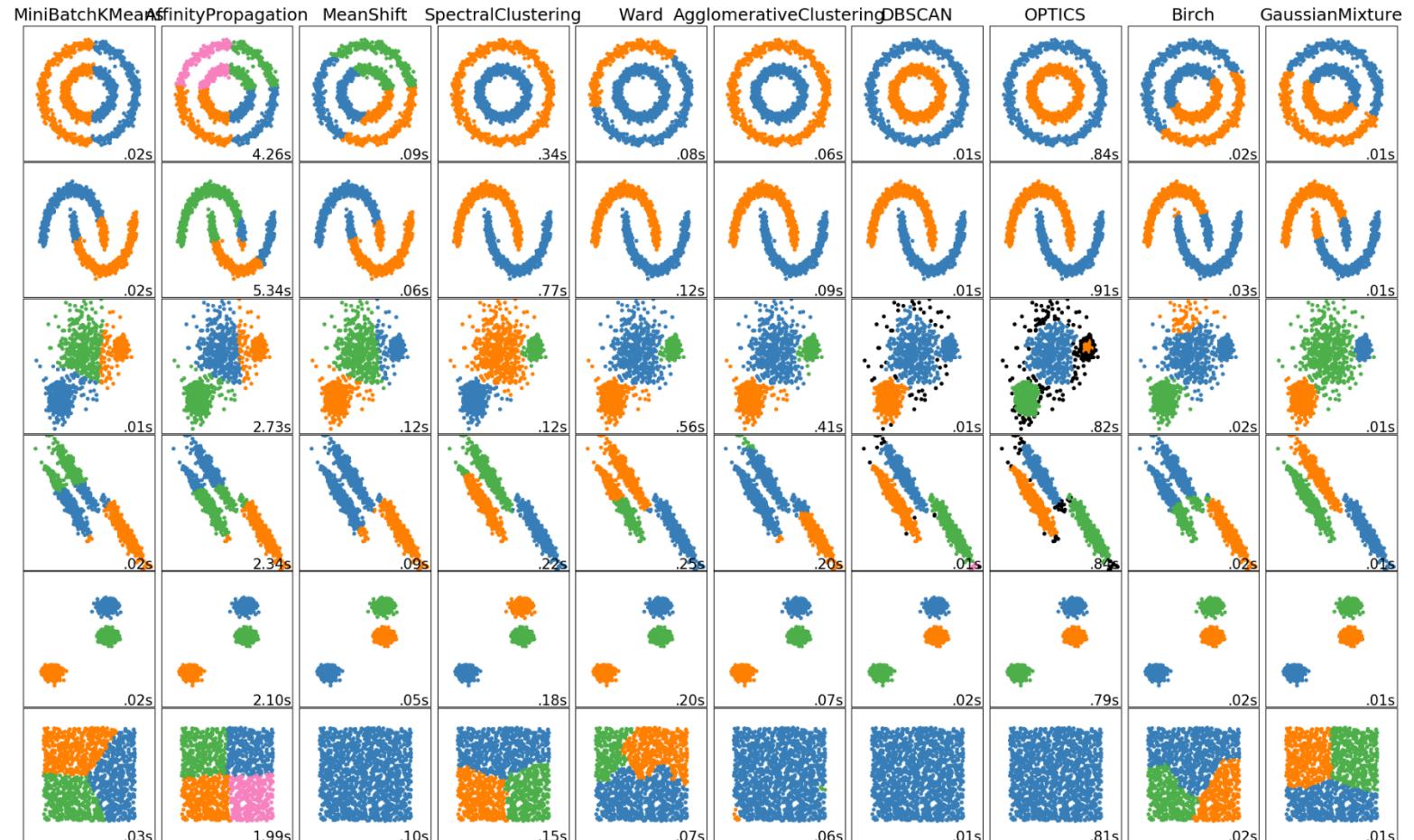


Map of classical machine learning methods



Clustering methods

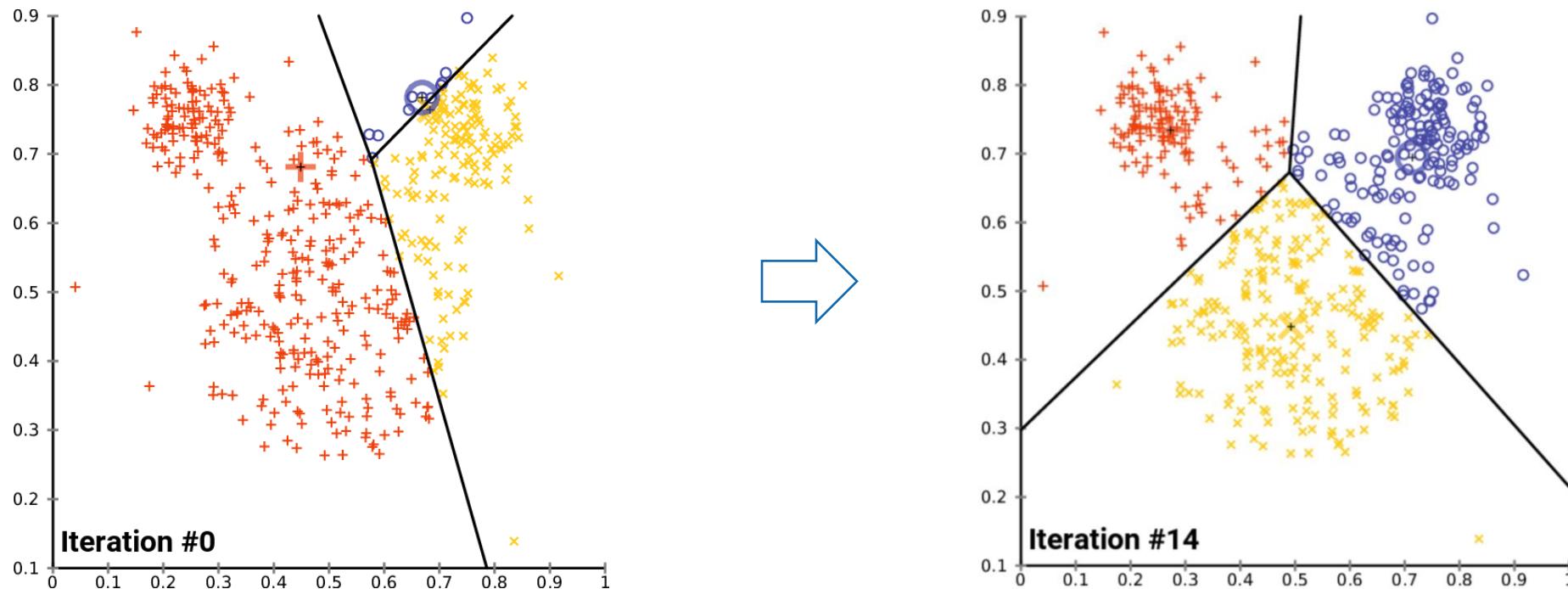
- K-means
- Gaussian mixture models
- Spectral clustering
- Hierarchical clustering



From <https://scikit-learn.org/stable/modules/clustering.html>

K-means

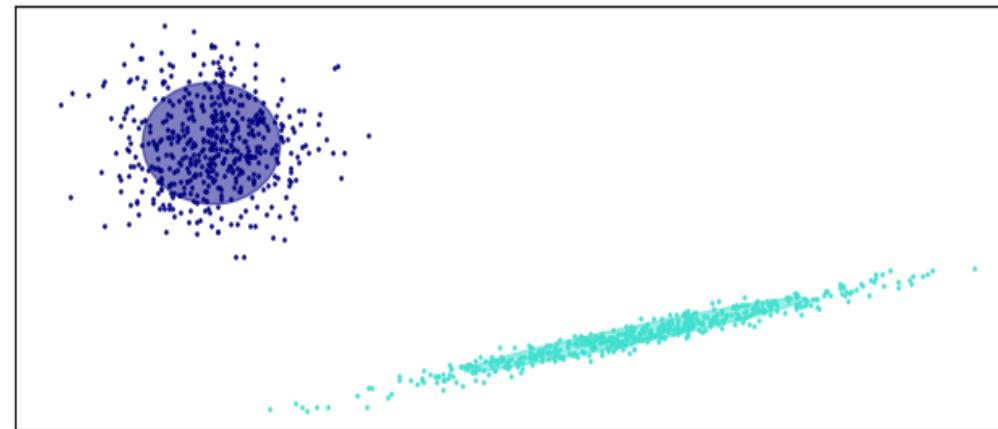
- k-means: aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers)



- k-means clustering tends to find clusters of comparable spatial extent

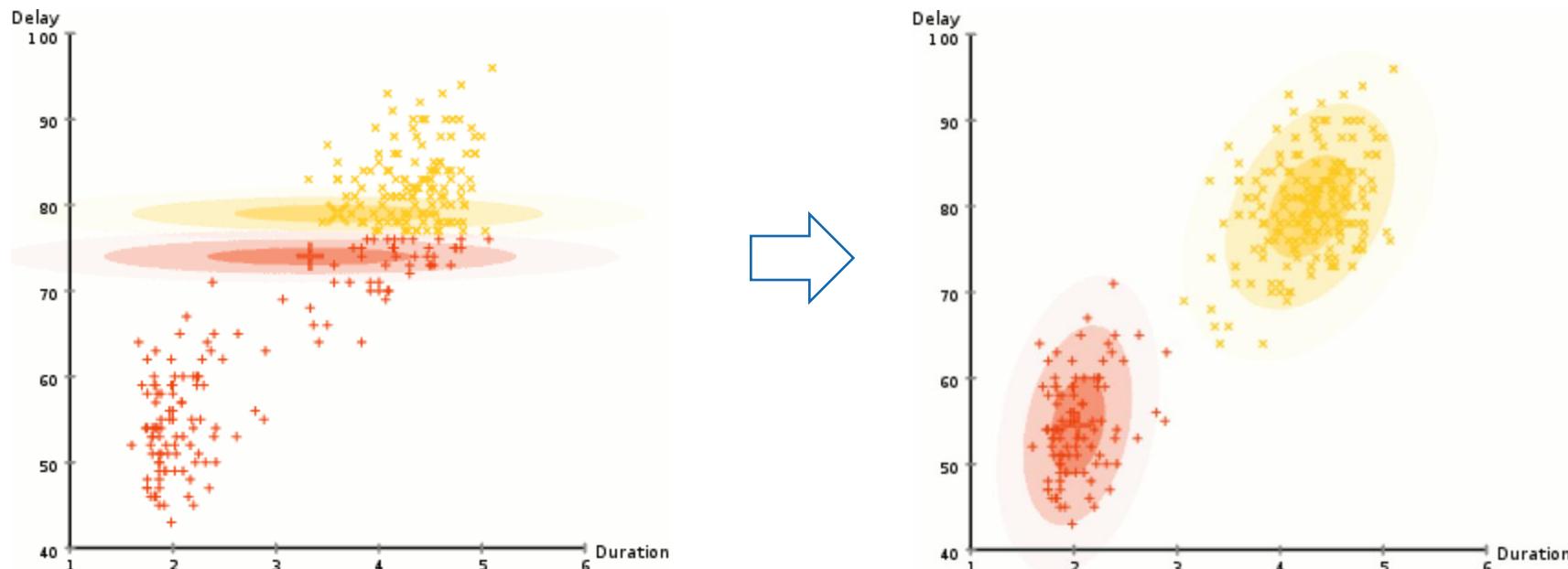
Gaussian mixture models (GMM)

GMM: probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters



Gaussian mixture models (GMM)

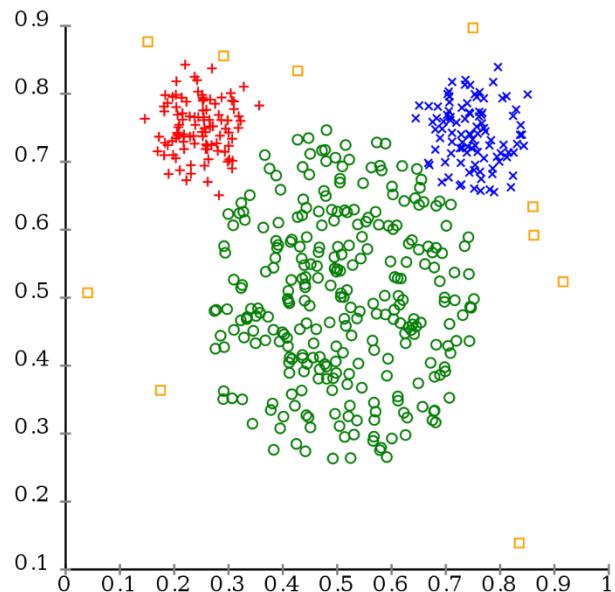
GMM: probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters



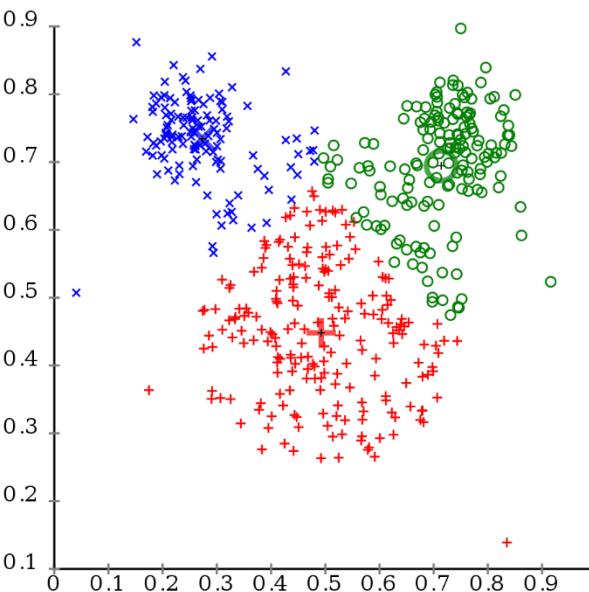
k-means clustering tends to find clusters of comparable spatial extent, while the GMM expectation-maximization mechanism allows clusters to have different shapes.

Different cluster analysis results on "mouse" data set:

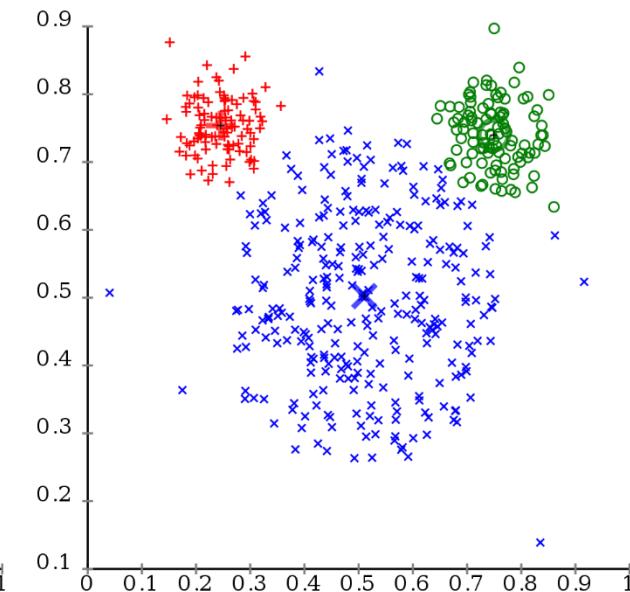
Original Data



k-Means Clustering



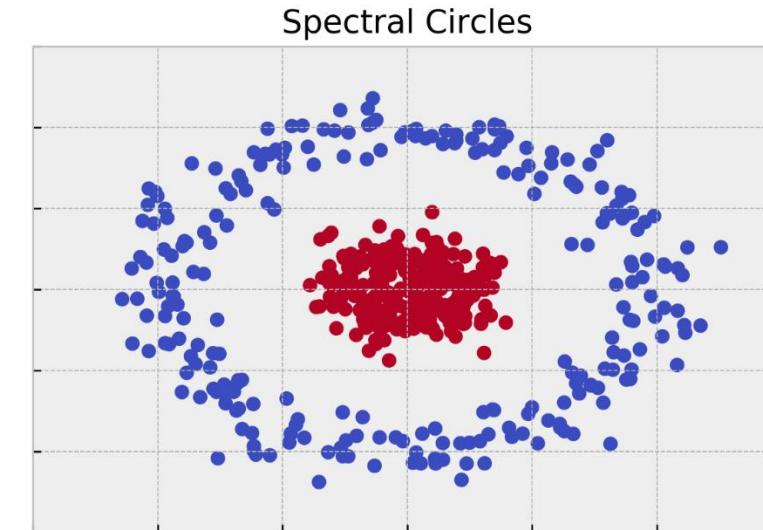
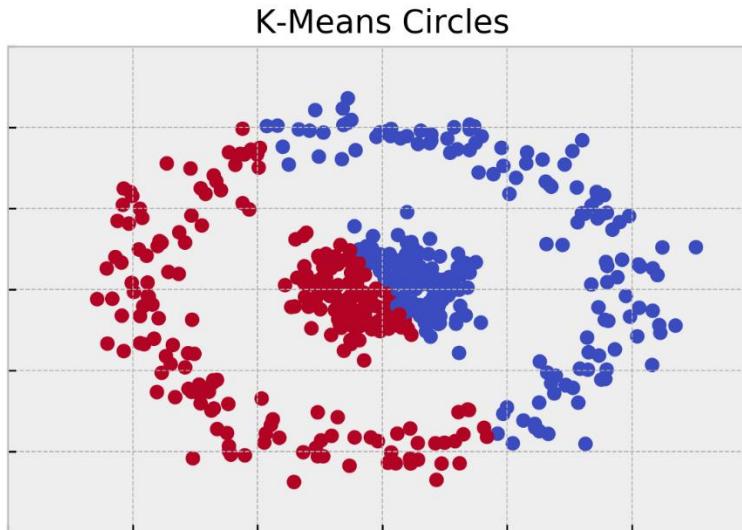
EM Clustering



Also, GMMs support mixed membership

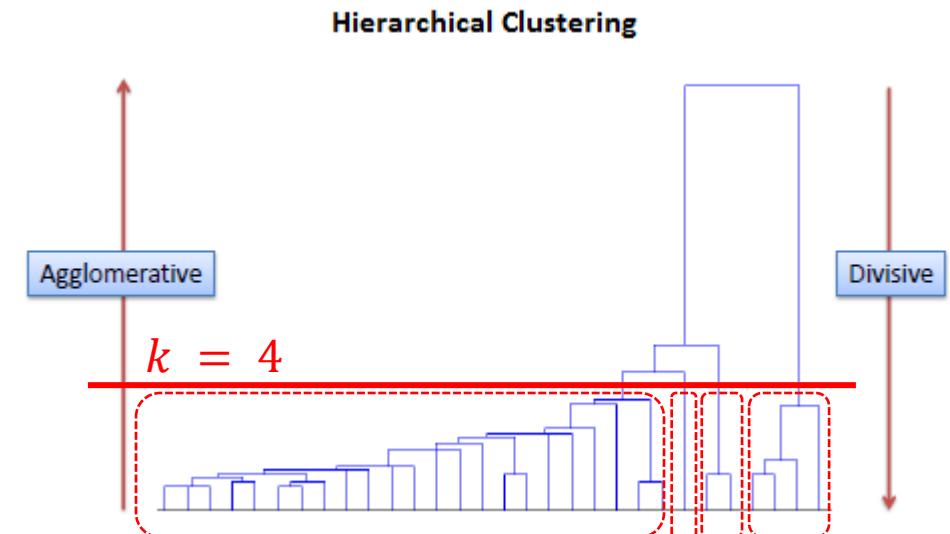
Spectral clustering: relies on the assumption that “close” points should belong in the same cluster

- Spectral clustering: uses a standard clustering method (e.g., k-means) on relevant eigenvectors of a Laplacian matrix L of symmetric data similarity matrix A . – can also use k-nearest neighbors graphs for construction of A .
- $L := D - A$, where D is the diagonal matrix, such as $D_{ii} = \sum_j A_{ij}$.



Hierarchical clustering

- Hierarchical clustering: family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram).
- **Agglomerative approach:** This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive approach:** This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

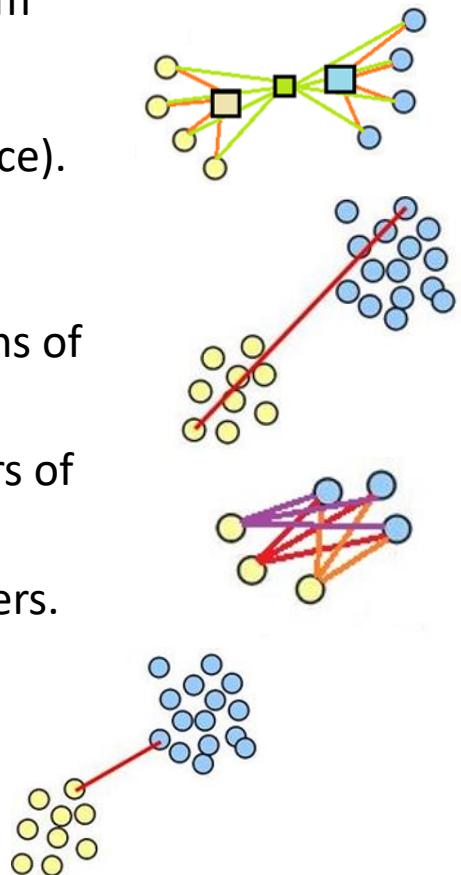


From https://www.saedsayad.com/clustering_hierarchical.htm

Hierarchical clustering: important parameters

You choose:

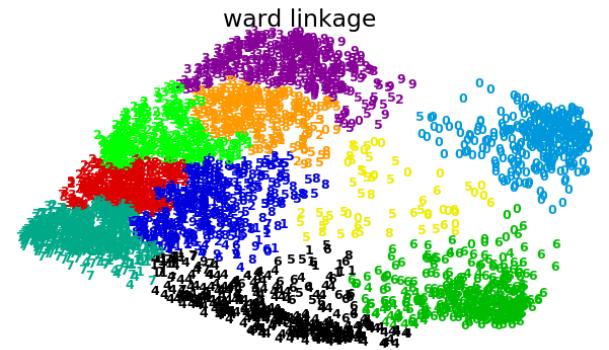
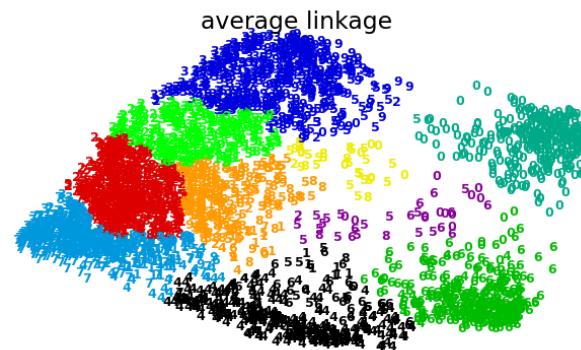
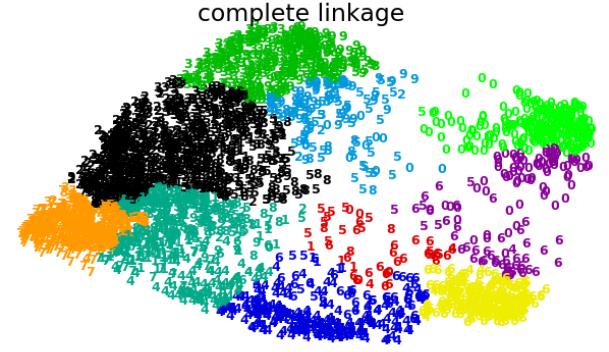
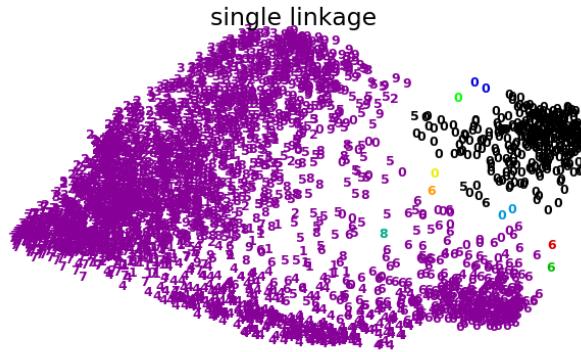
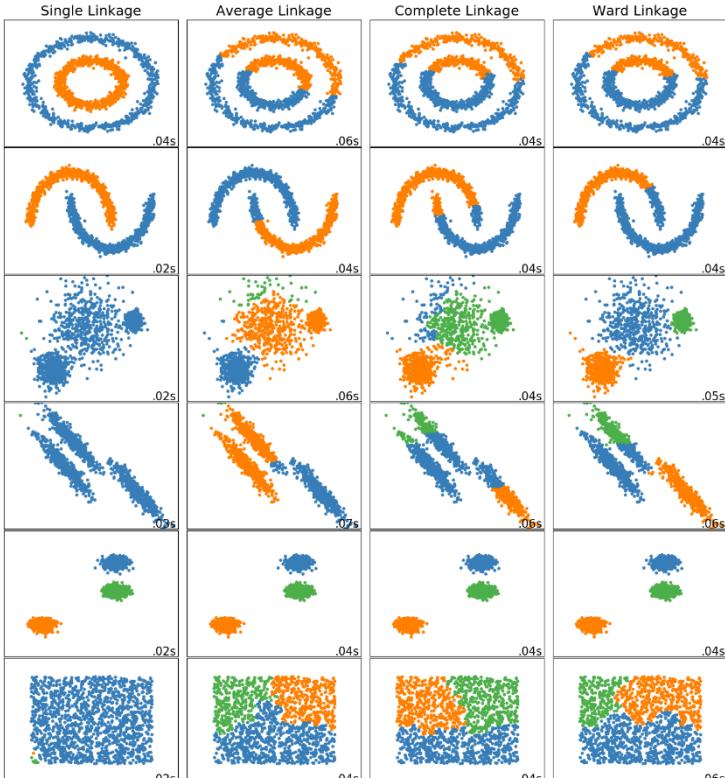
- **Distance metric** (between observations): Euclidean, Squared Euclidean, Manhattan, Maximum
- **Linkage criterium** (distance between sets of observations):
 - “**Ward**” minimizes the sum of squared differences within all clusters (within-cluster variance). It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.
 - “**Maximum**” or “complete linkage” minimizes the maximum distance between observations of pairs of clusters.
 - “**Average linkage**” minimizes the average of the distances between all observations of pairs of clusters.
 - “**Single linkage**” minimizes the distance between the closest observations of pairs of clusters.



<https://blog.tdwi.eu/hierarchical-clustering-in-python/>

Hierarchical clustering: important parameters

- **Linkage criterium** (distance between sets of observations):

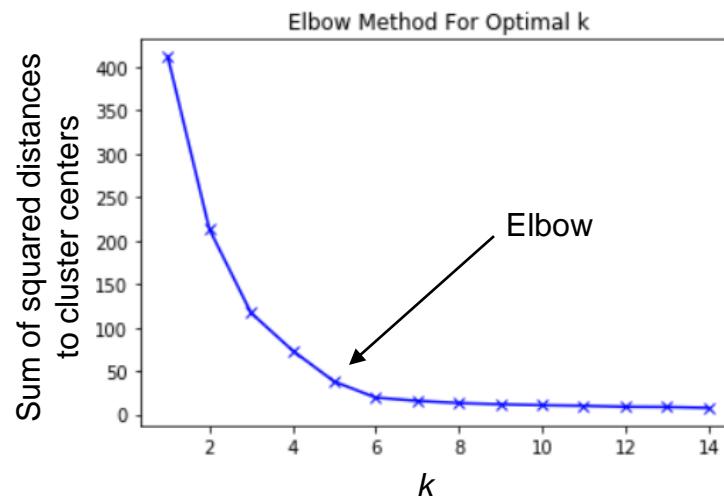


“Ward” gives the most regular cluster sizes

From <https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>
https://scikit-learn.org/stable/auto_examples/cluster/plot_digits_linkage.html

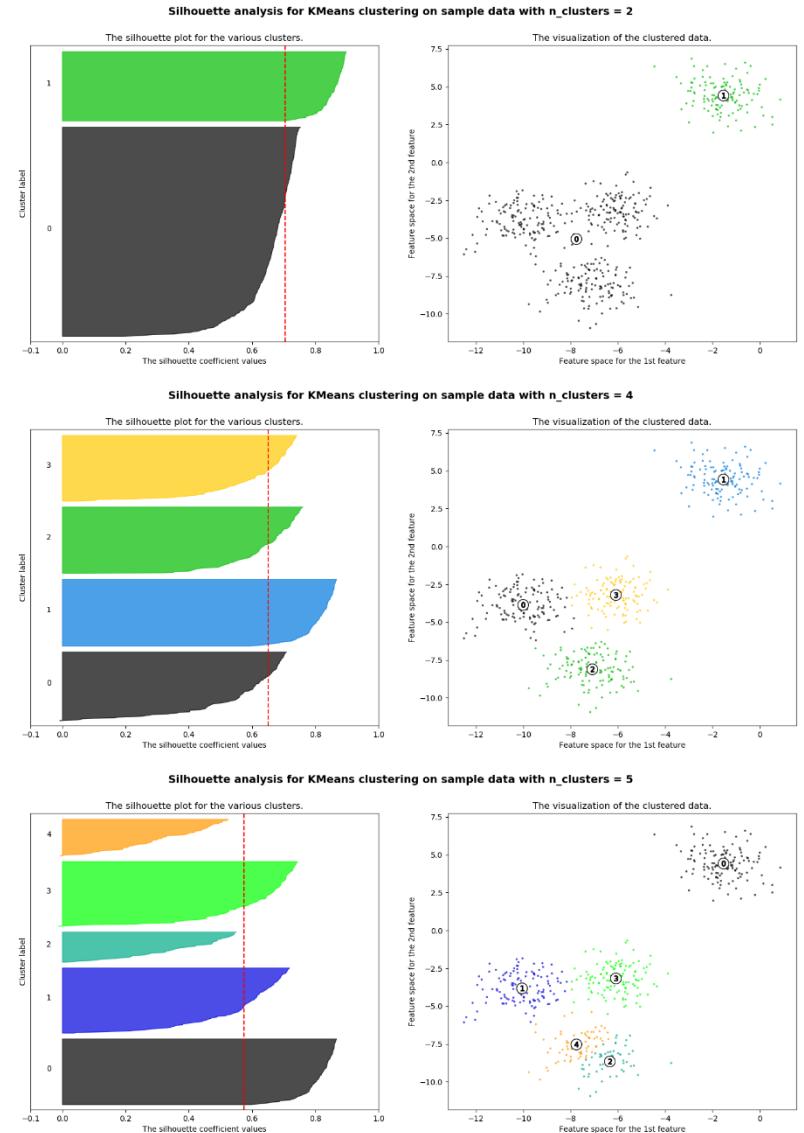
How to choose the best number of clusters k?

- Elbow method (for sum of squared distances to cluster centers)
 - How find the “elbow”? By eye or <https://github.com/arvkevi/kneed>
- Silhouette analysis
- BIC and AIC



How to choose the best number of clusters k?

- Elbow method
- Silhouette analysis
 - Plots silhouette score: a measure of how close each point in one cluster is to points in the neighboring clusters, in $[-1, 1]$.
 - Values near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.



From https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

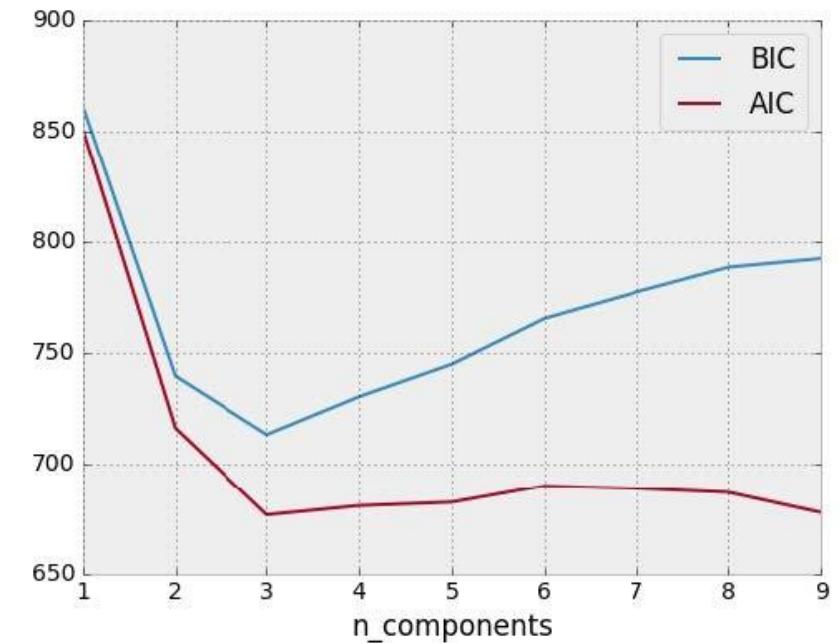
How to choose the best number of clusters k?

- Elbow method
- Silhouette analysis
- Akaike information criterion (AIC) or Bayesian information criterion (BIC)

$$\mathbf{AIC} = 2k - 2 \ln(\hat{L})$$

$$\mathbf{BIC} = k \ln(n) - 2 \ln(\hat{L}).$$

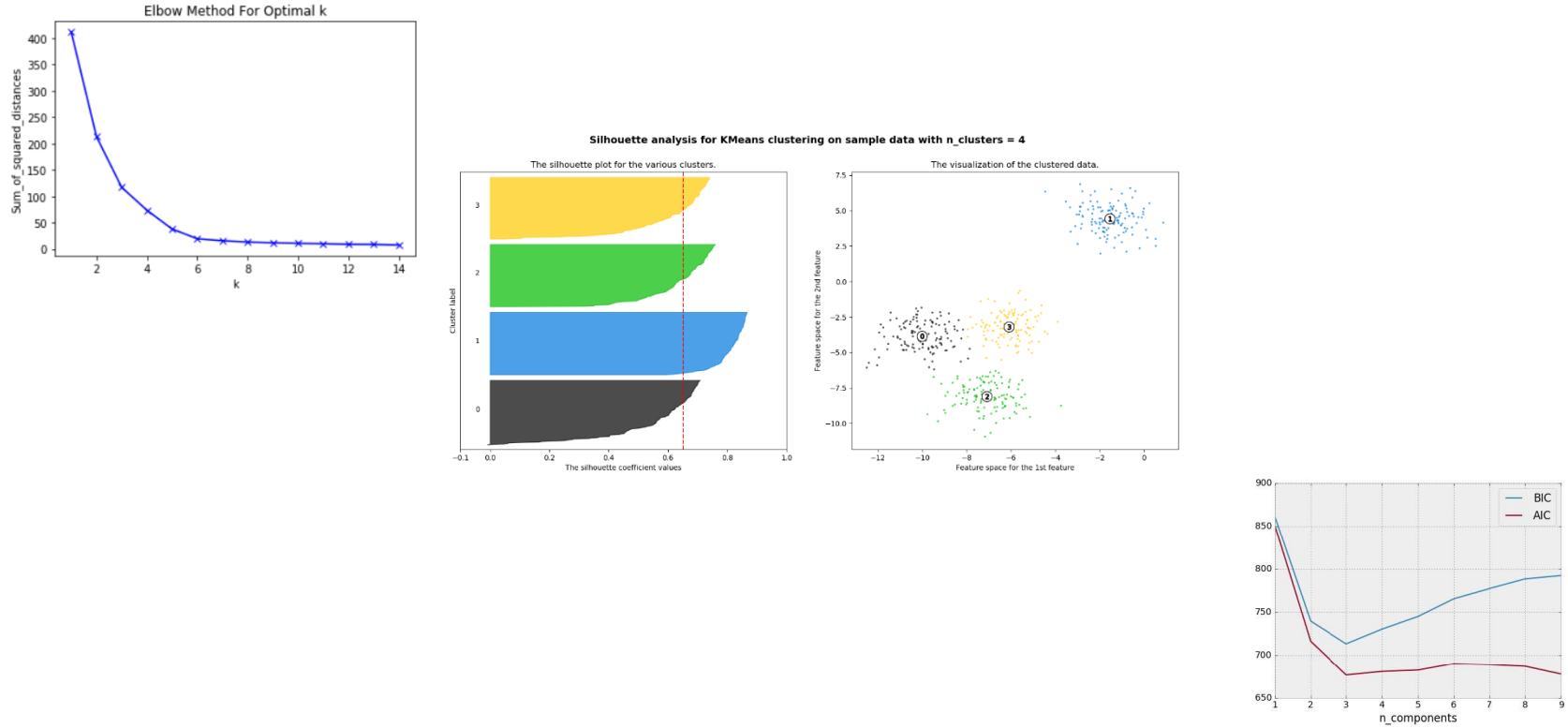
- The BIC generally penalizes free parameters more strongly than the AIC



From <https://sites.northwestern.edu/msia/2016/12/08/k-means-shouldnt-be-our-only-choice/>

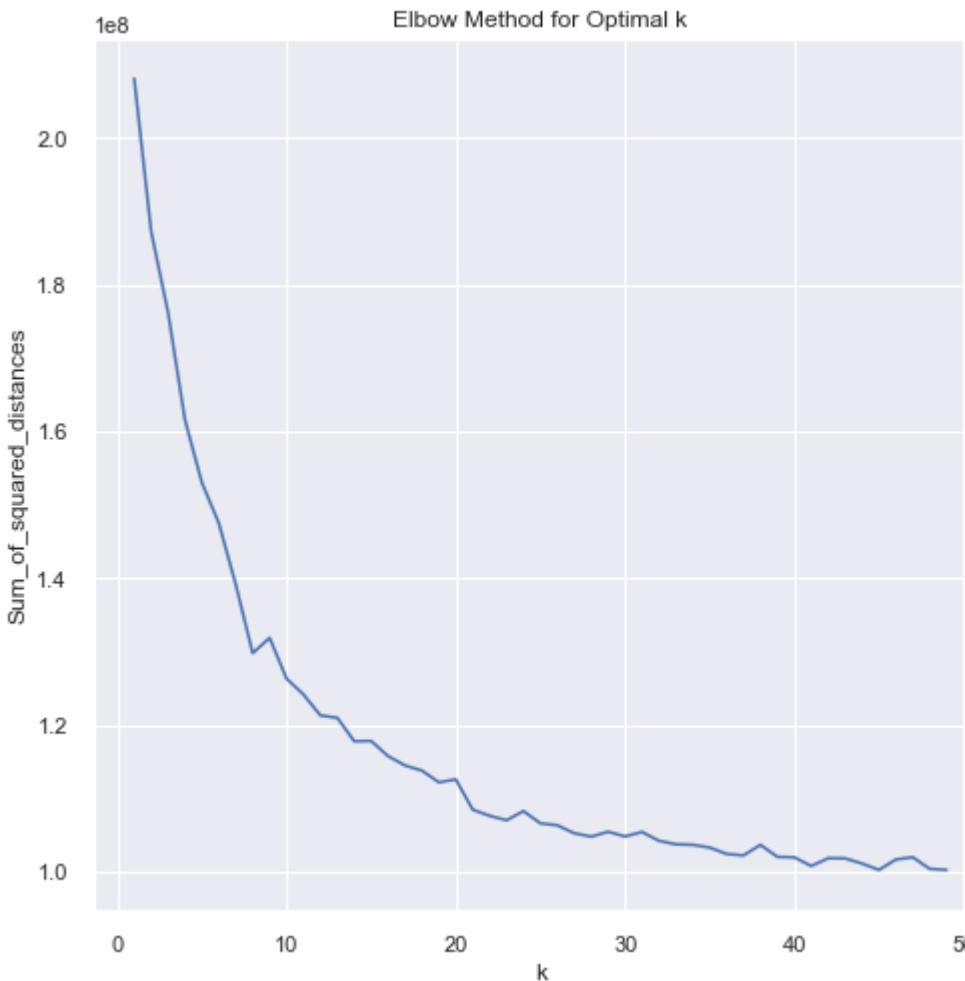
How to choose the best number of clusters k?

- Elbow method
- Silhouette analysis
- BIC and AIC



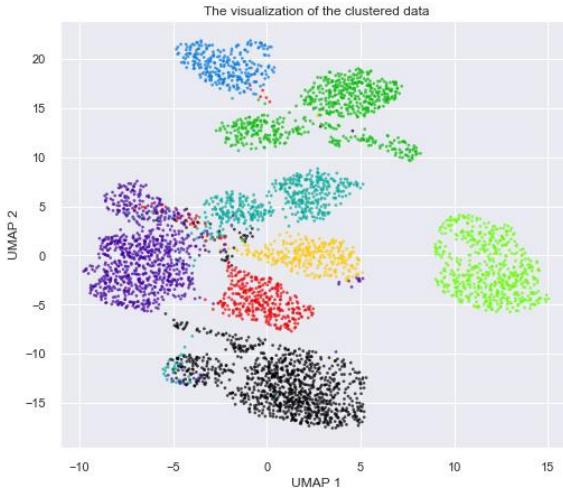
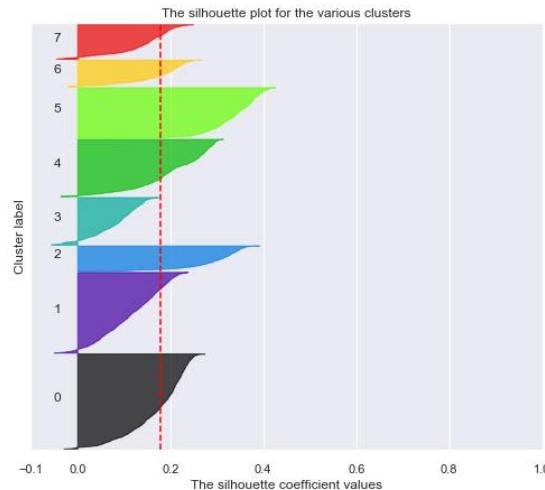
Let's go to our hands on exercise to see how it works!

Which value of k would you choose?
(Mini-batch k-means clustering method)



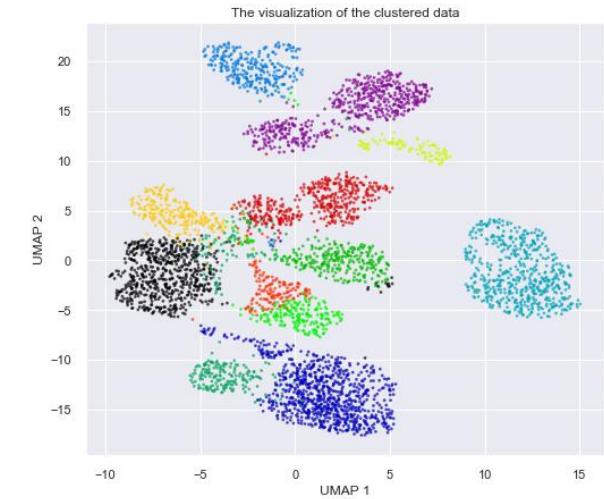
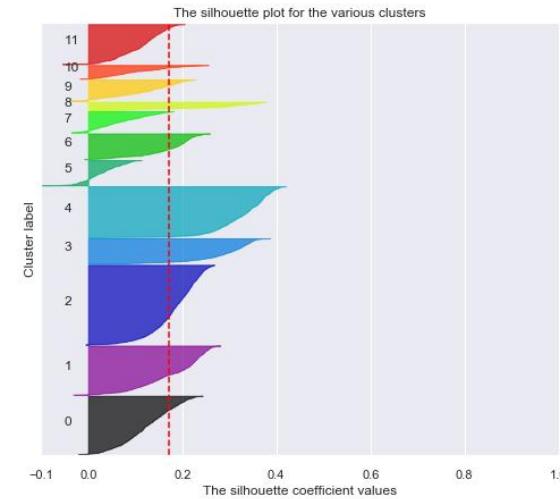
Which value of k would you choose?

Silhouette analysis for KMeans clustering on sample data with n_clusters = 8



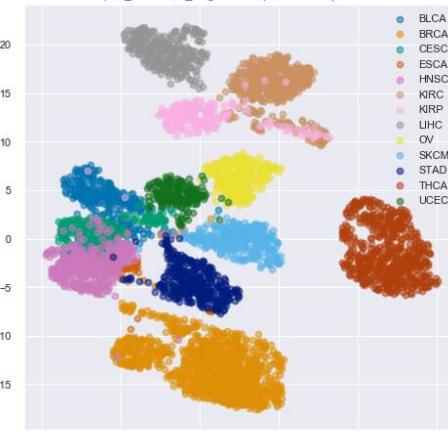
$k = 8$

Silhouette analysis for KMeans clustering on sample data with n_clusters = 12



$k = 12$

UMAP (min_dist=1.0, n_neighbors=15) of TCGA expression dataset

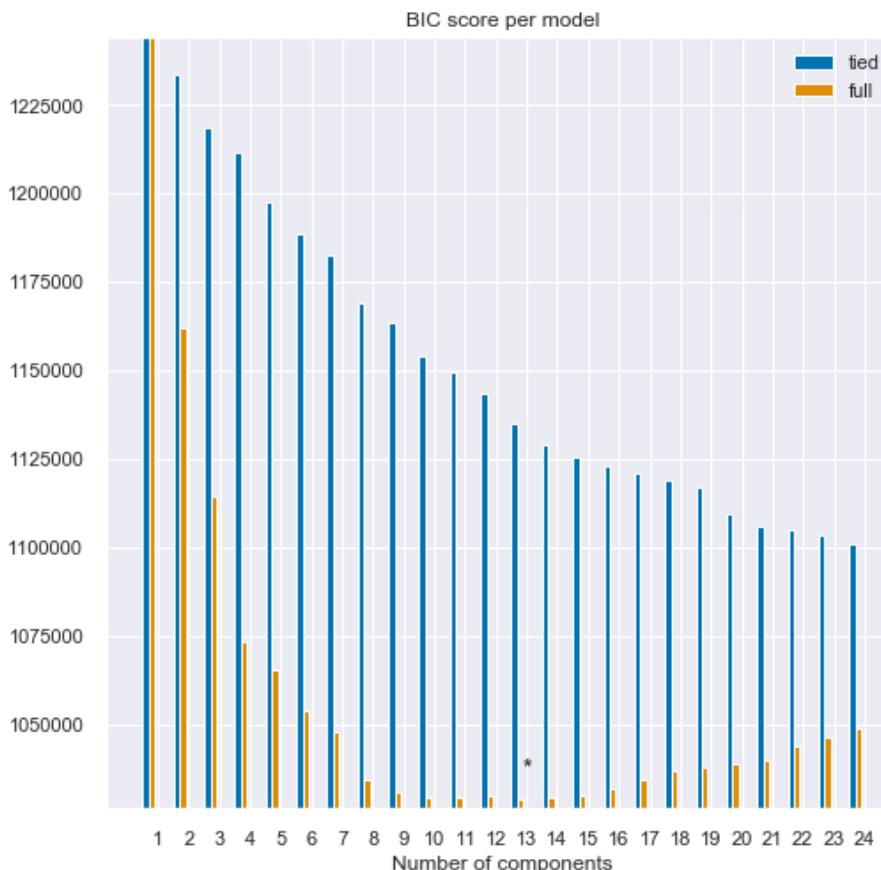


True types

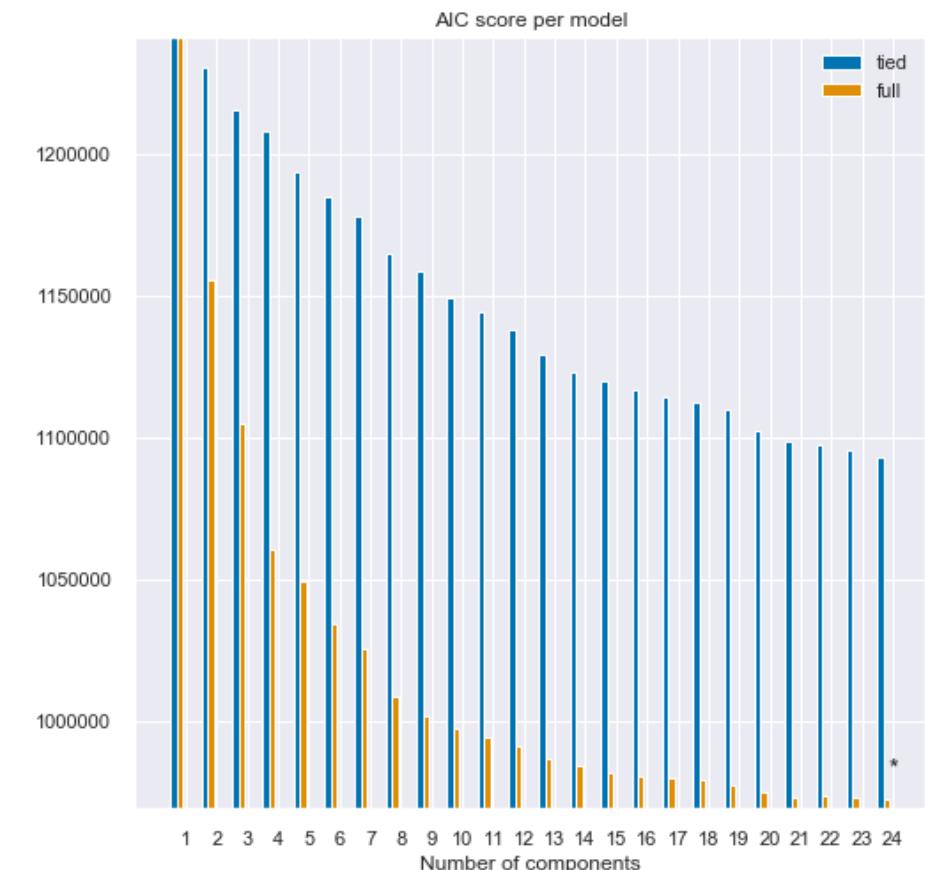
| | |
|------|-----------------------------|
| BLCA | Bladder |
| BRCA | Breast |
| CESC | Cervical squamous cell |
| ESCA | Esophageal |
| HNSC | Head and Neck squamous cell |
| KIRC | Kidney clear cell |
| KIRP | Kidney papillary |
| LIHC | Liver |
| OV | Ovarian |
| SKCM | Skin Cutaneous Melanoma |
| STAD | Stomach |
| THCA | Thyroid |
| UCEC | Uterine Corpus Endometrial |

Which value of k would you choose? (Gaussian mixture model)

BIC



AIC



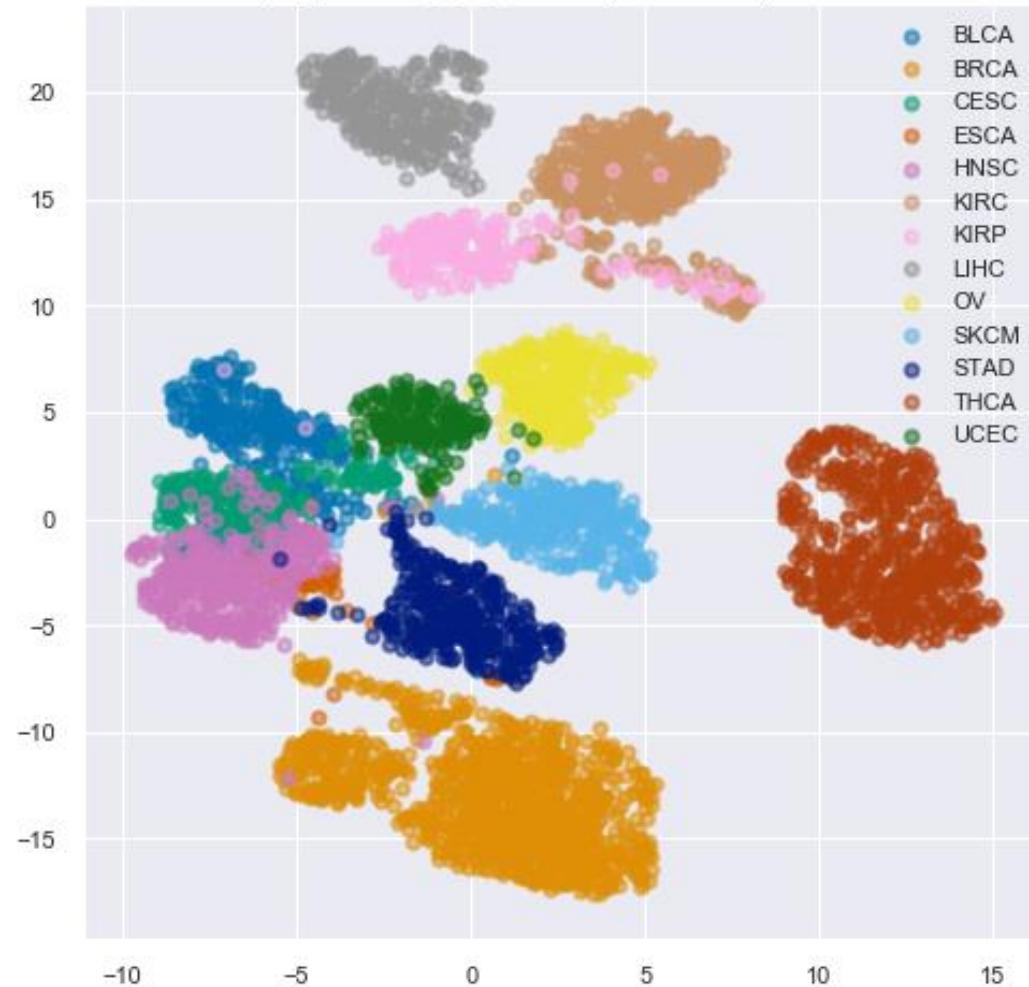
Which value of k would you choose? (Gaussian mixture model)

Best solution according to BIC: $k=13$



True labels

UMAP (min_dist=1.0, n_neighbors=15) of TCGA expression dataset



NMI and ARI allows to see which clustering model is the best on our data set

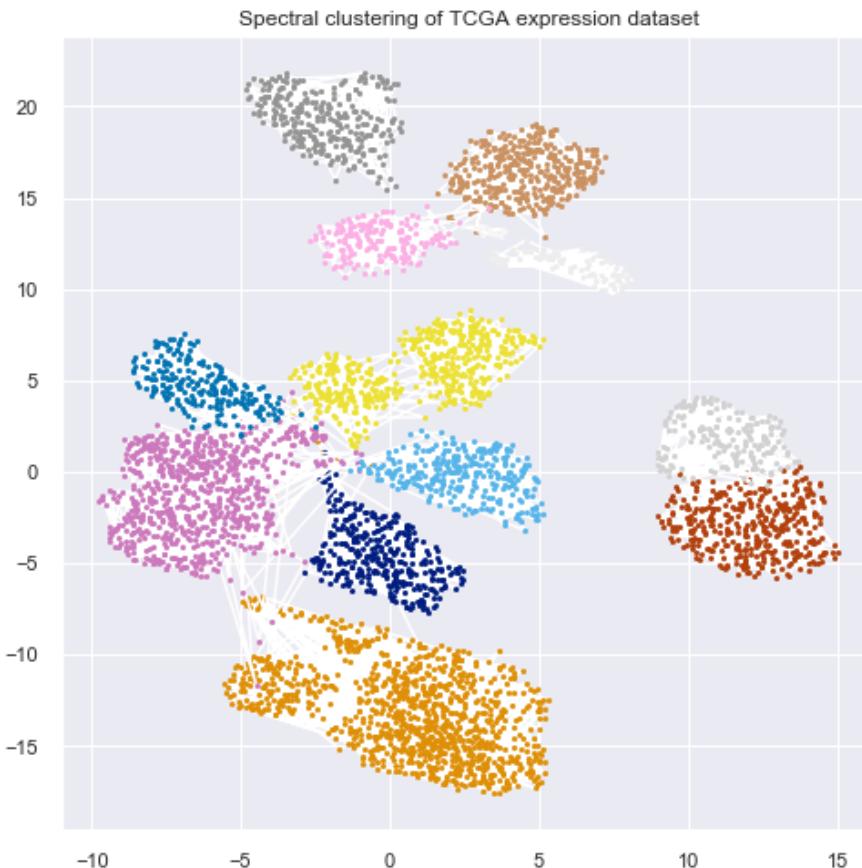
When true labels are available (rarely the case though):

NMI = Normalized Mutual Information
ARI = Adjusted rand index } measure mutual dependence between the true and predicted labels
(can be also used as a measure of similarity between two clusterings)

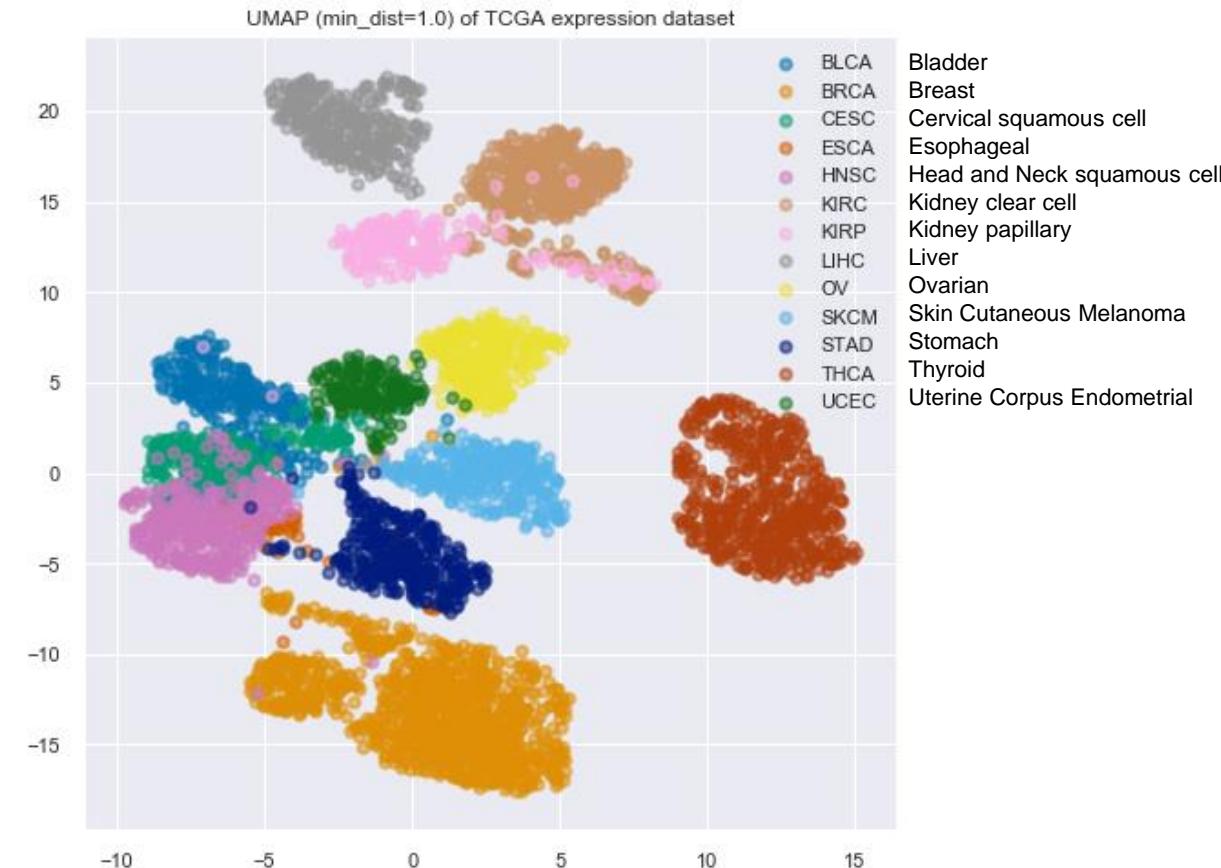
| Method | k | ARI | NMI |
|--|-----------|---------------|---------------|
| Mini-batch k-means | 8 | 0.7479 | 0.8102 |
| Mini-batch k-means | 19 | 0.7426 | 0.8179 |
| Gaussian mixture model | 13 | 0.7600 | 0.8479 |
| Spectral clustering | 13 | 0.8148 | 0.8671 |
| Hierarchical clustering (linkage='ward') | 13 | 0.7394 | 0.8379 |

Spectral clustering result (k=13)

Best solution according to NMI



True labels



Take home message: clustering

- The choice of the clustering method should be advised by data structure
 - Visualize the data first
 - E.g., ellipsoids => GMM; sophisticated connected components => structural clustering
- Choosing the number of clusters can be done using:
 - Elbow method
 - Silhouette method
 - BIC or AIC
- Different methods for the estimation of the number of clusters provide different results
 - E.g., BIC is more conservative than AIC

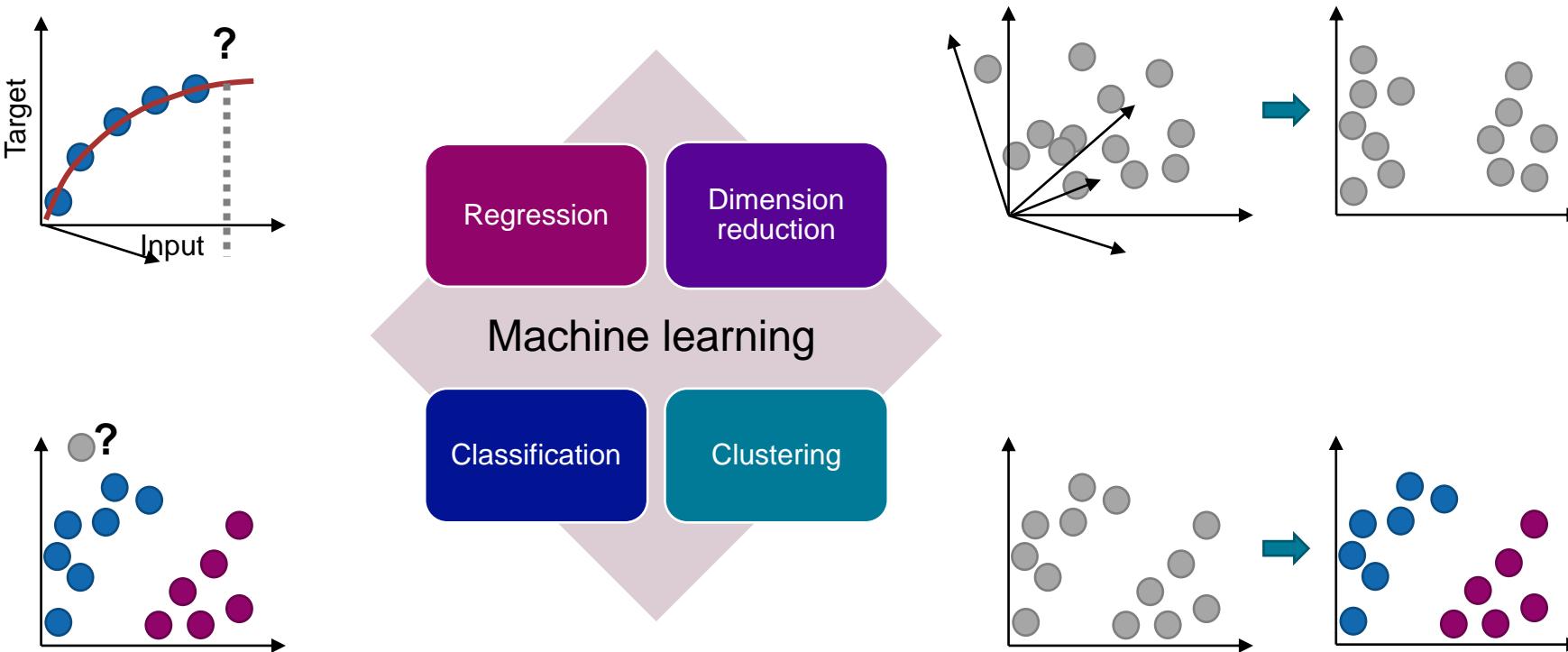
What we did not cover today

- Graph-based clustering methods (Leiden and Louvain algorithms)
- Topic models/LDA
- Autoencoder-based methods for dimensionality reduction
- Integration of different data types (e.g., clusters of clusters)

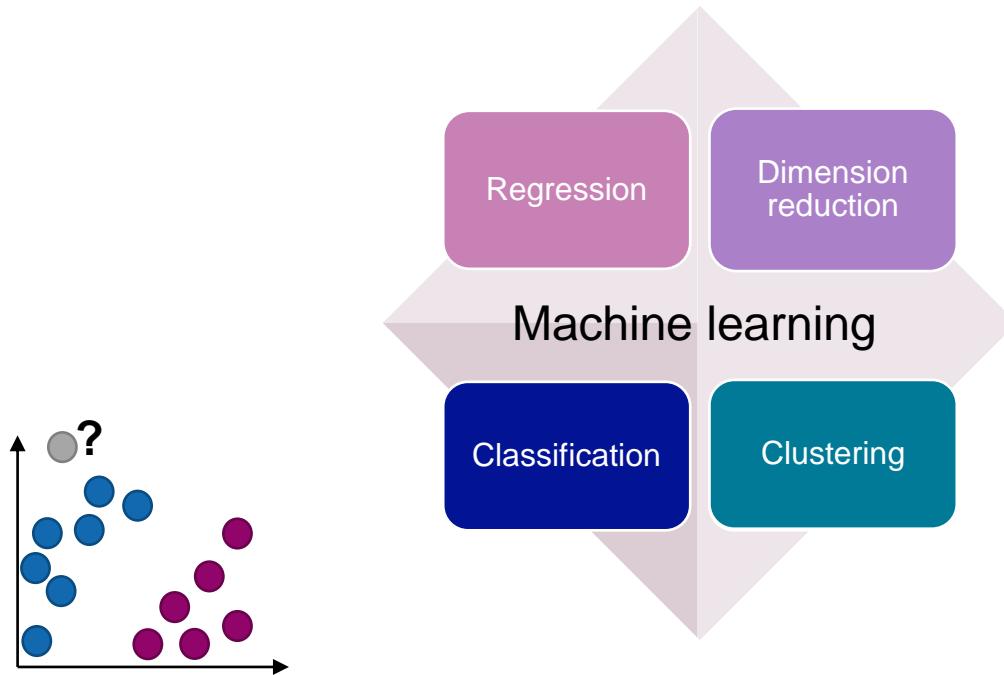
Take home message

- Dimensionality reduction can be used as a step prior to clustering
- One usually tries different dimensionality reduction techniques to choose the one that fits the expectations
- The choice of clustering method should match the data structure

Map of classical machine learning methods

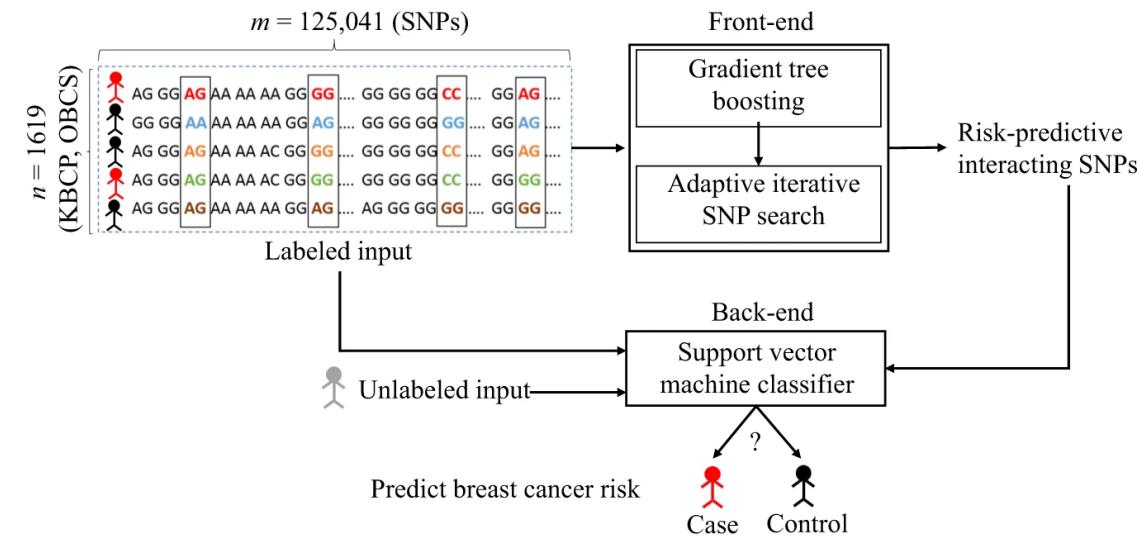


Map of classical machine learning methods



Classification: Biological examples

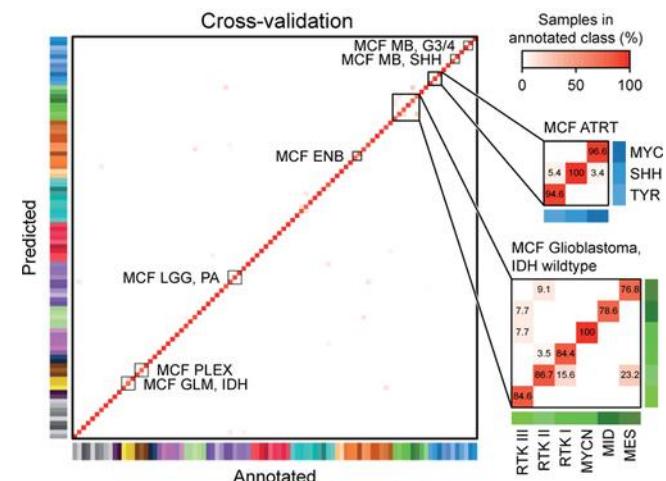
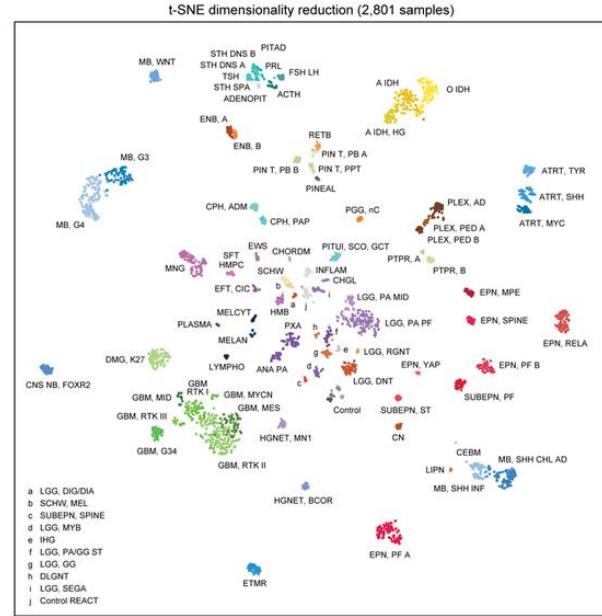
- Prediction of risk groups
 - **Machine learning identifies interacting genetic variants contributing to breast cancer risk [...] Behravan et al., *Sci Rep.* 2018**
- Primary diagnosis
 - **DNA methylation-based classification of central nervous system tumours. Capper et al. *Nature* 2018**
- Patient stratification / choice of treatment
 - **Genomic Amplifications and Distal 6q Loss: Novel Markers for Poor Survival in High-risk Neuroblastoma Patients. Depuydt et al., *J Natl Cancer Inst.* 2018**



Proposed breast cancer risk prediction approach using identified risk-predictive interacting SNPs

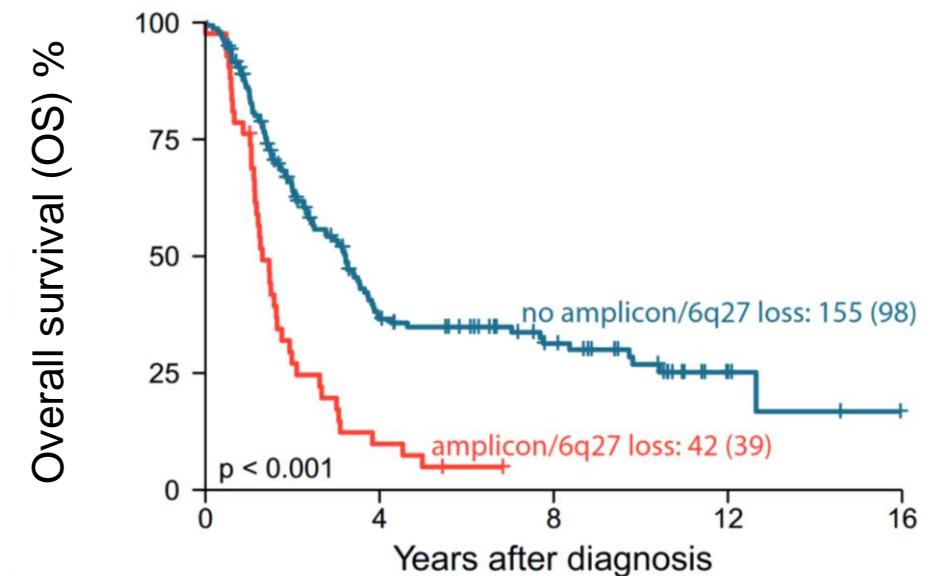
Classification: Biological examples

- Prediction of risk groups
 - Machine learning identifies interacting genetic variants contributing to breast cancer risk [...] Behravan et al., *Sci Rep.* 2018
- Primary diagnosis
 - DNA methylation-based classification of central nervous system tumours. Capper et al. *Nature* 2018
- Patient stratification / choice of treatment
 - Genomic Amplifications and Distal 6q Loss: Novel Markers for Poor Survival in High-risk Neuroblastoma Patients. Depuydt et al., *J Natl Cancer Inst.* 2018



Classification: Biological examples

- Prediction of risk groups
 - Machine learning identifies interacting genetic variants contributing to breast cancer risk [...] Behravan et al., *Sci Rep.* 2018
- Primary diagnosis
 - DNA methylation-based classification of central nervous system tumours. Capper et al. *Nature* 2018
- Patient stratification / choice of treatment
 - Genomic Amplifications and Distal 6q Loss: Novel Markers for Poor Survival in High-risk Neuroblastoma Patients. Depuydt et al., *J Natl Cancer Inst.* 2018

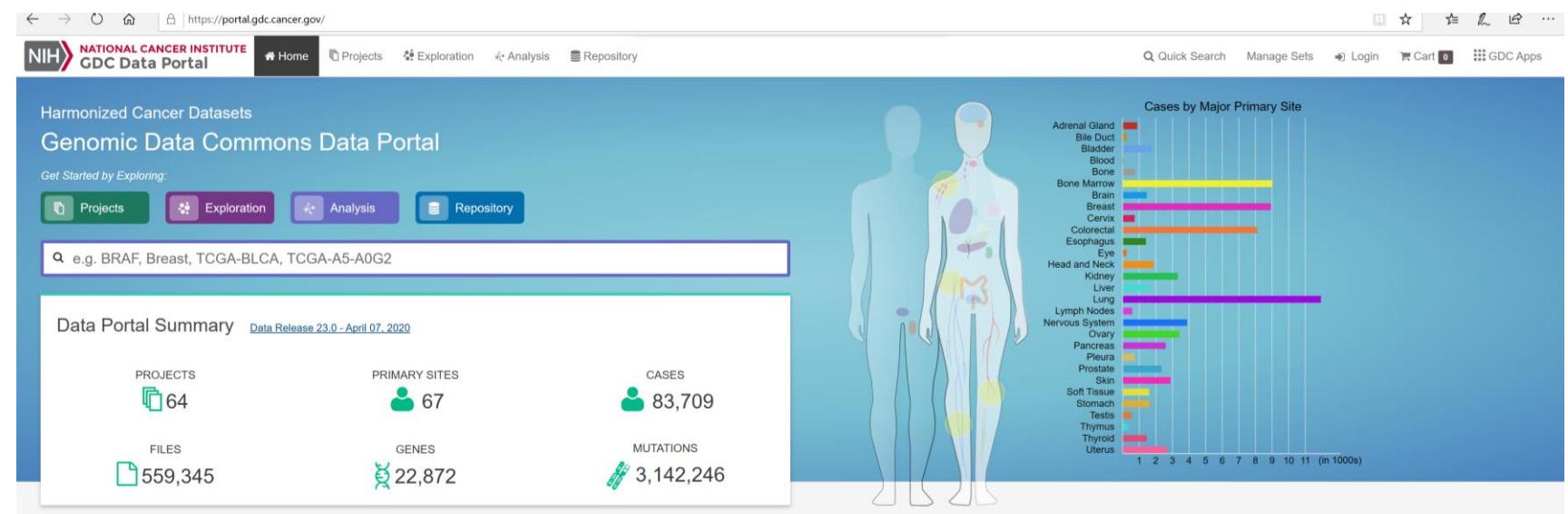


Survival differences in high-risk neuroblastoma patients with and without amplicon/6q27 aberrations.

Hands-on:

1. Primary cancer diagnosis from gene expression
2. Breast cancer patients' stratification based on gene expression

- **Input:** The Cancer Genome Atlas (TCGA) mRNA expression data

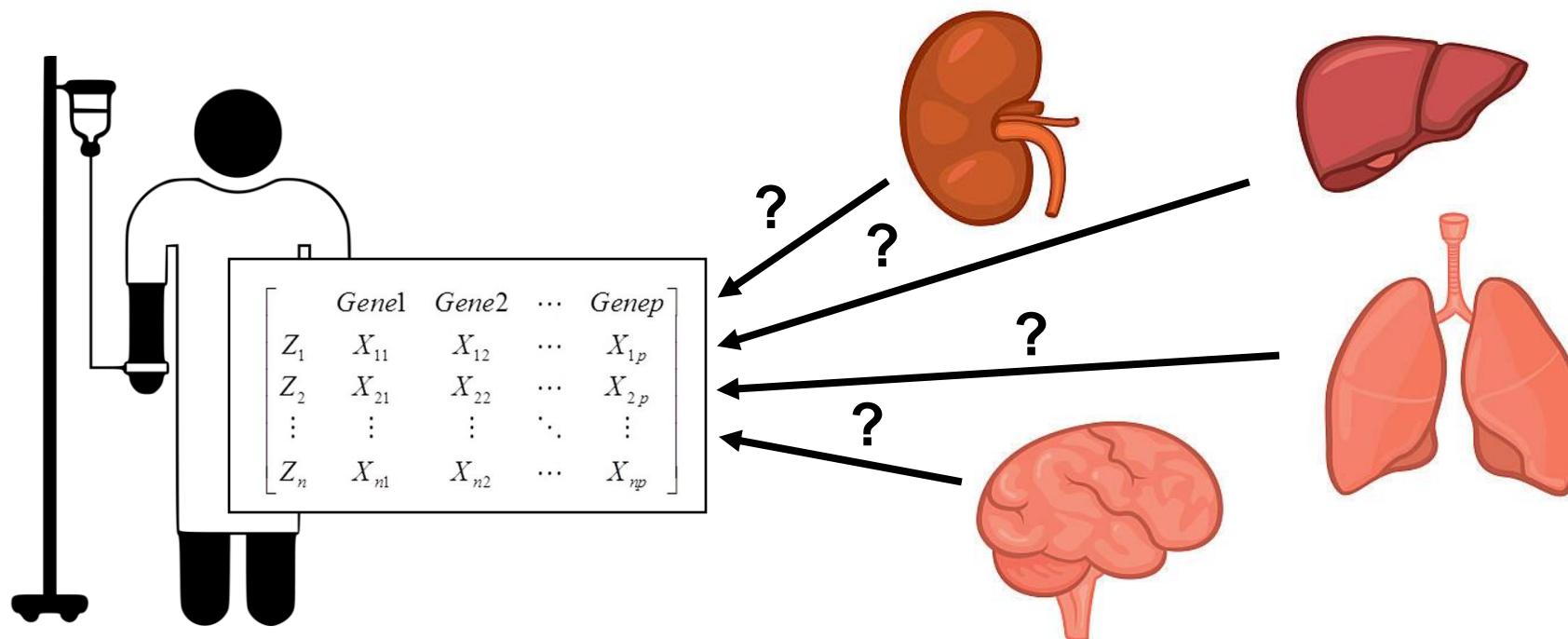


Hands-on:

1. Primary cancer diagnosis from gene expression
2. Breast cancer patients' stratification based on gene expression

- **TASK 1:**

Given mRNA expression, predict cancer type



Hands-on:

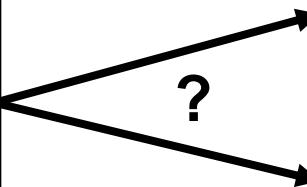
1. Primary cancer diagnosis from gene expression
2. Breast cancer patients' stratification based on gene expression

- **TASK 2:**

Given mRNA expression (and clinical data: stage, age), stratify patients according to good and bad prognosis


$$\begin{bmatrix} & Gene1 & Gene2 & \cdots & Genep \\ Z_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Z_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

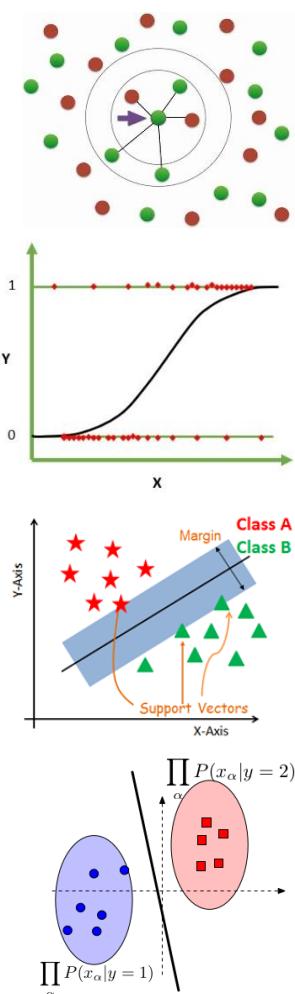
+ clinical stage + age



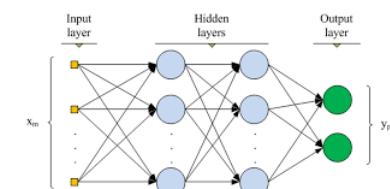
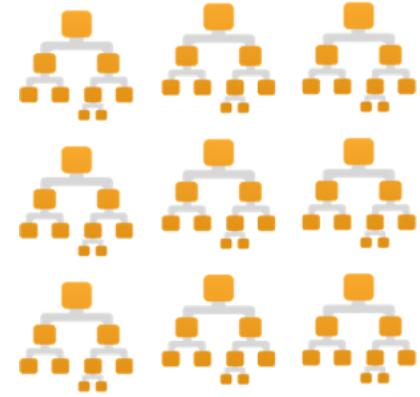
Aggressive treatment to be applied

Standard approaches for classification

- k Nearest Neighbors (k-NN)
- Logistic Regression
- Logistic Regression with L1+L2 (Elastic Net) penalty
- Support Vector Machines (SVM)
- Naive Bayes (Gaussian)

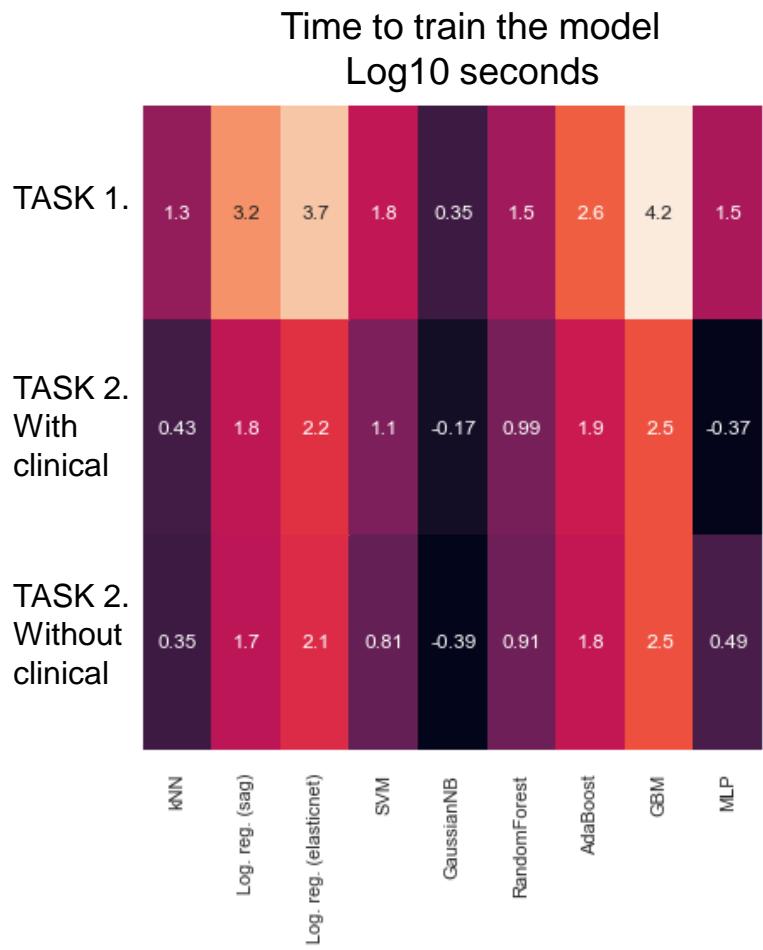


- Random Forest
- AdaBoost
- Gradient Tree Boosting (gradient boosting machine, GBM)
- Multi-layer perceptron (MLP)

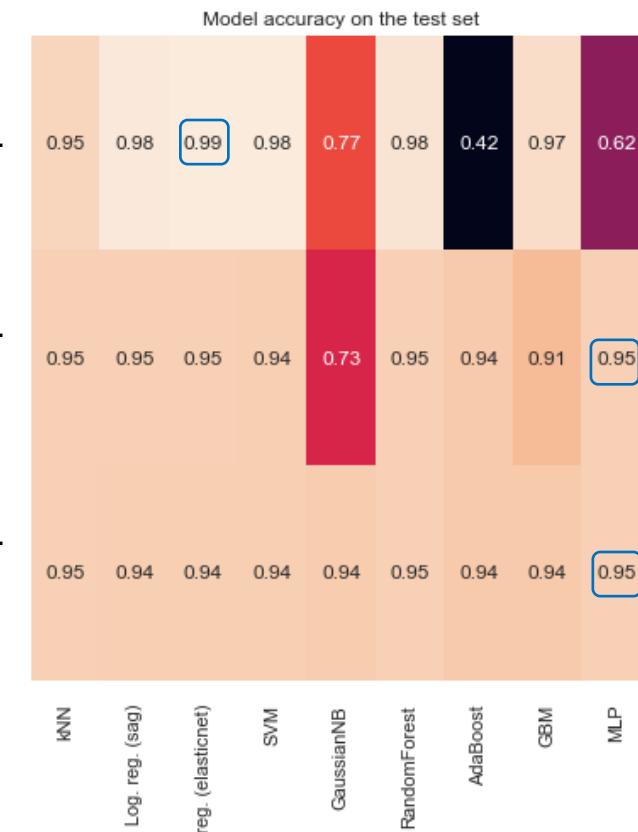


Hands-on:

1. Primary cancer diagnosis from gene expression
2. Breast cancer patients' stratification based on gene expression

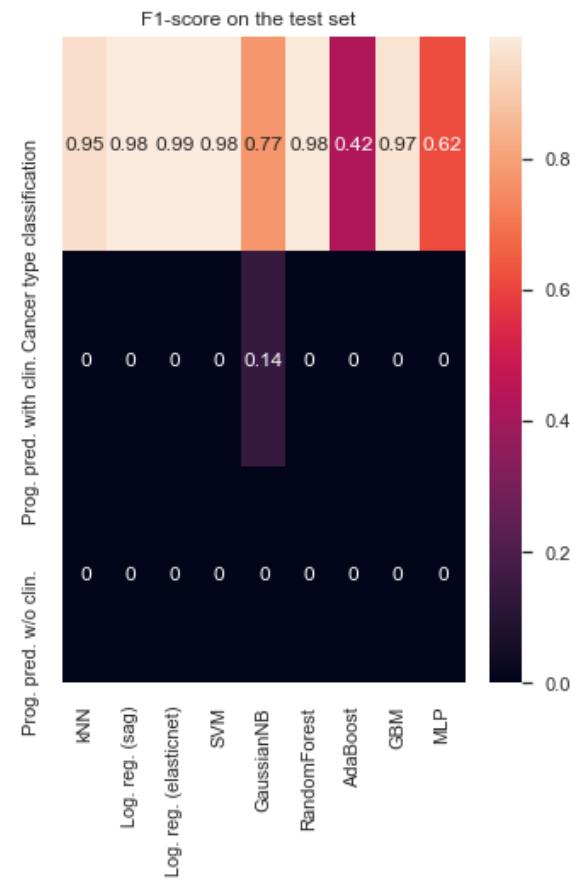


Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
Fraction predicted correctly



F1 score: harmonic mean of recall and precision

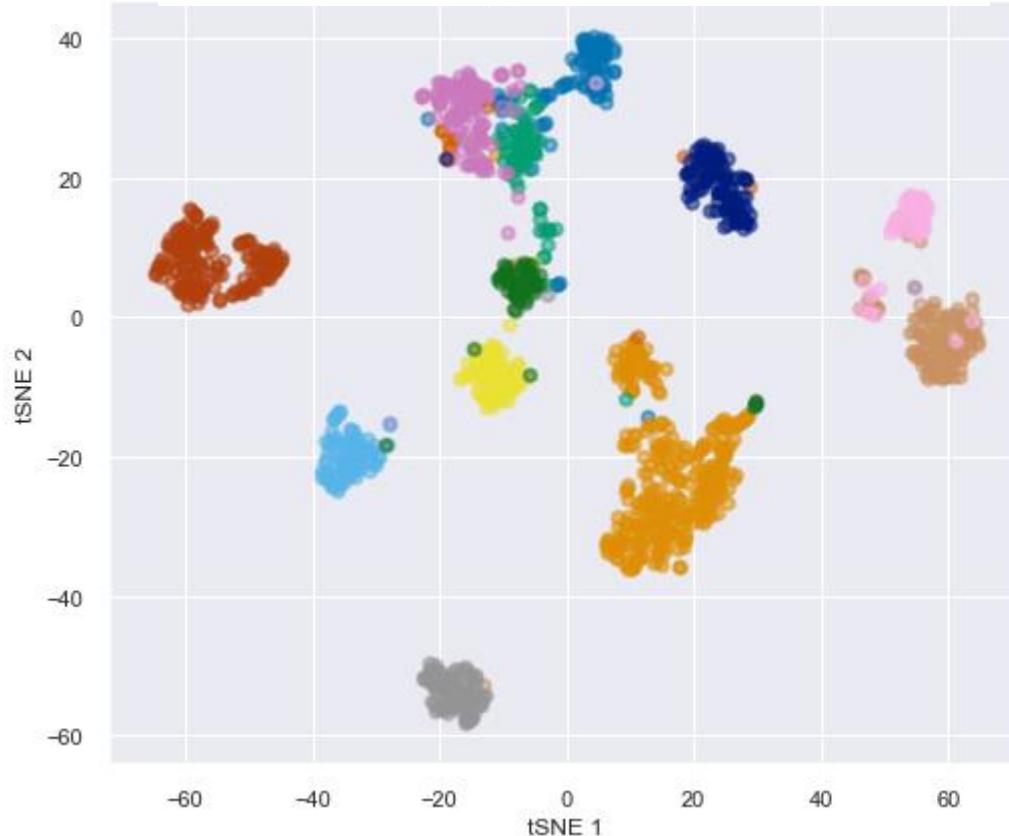
$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * (precision * recall)}{precision + recall}$$



Hands-on:

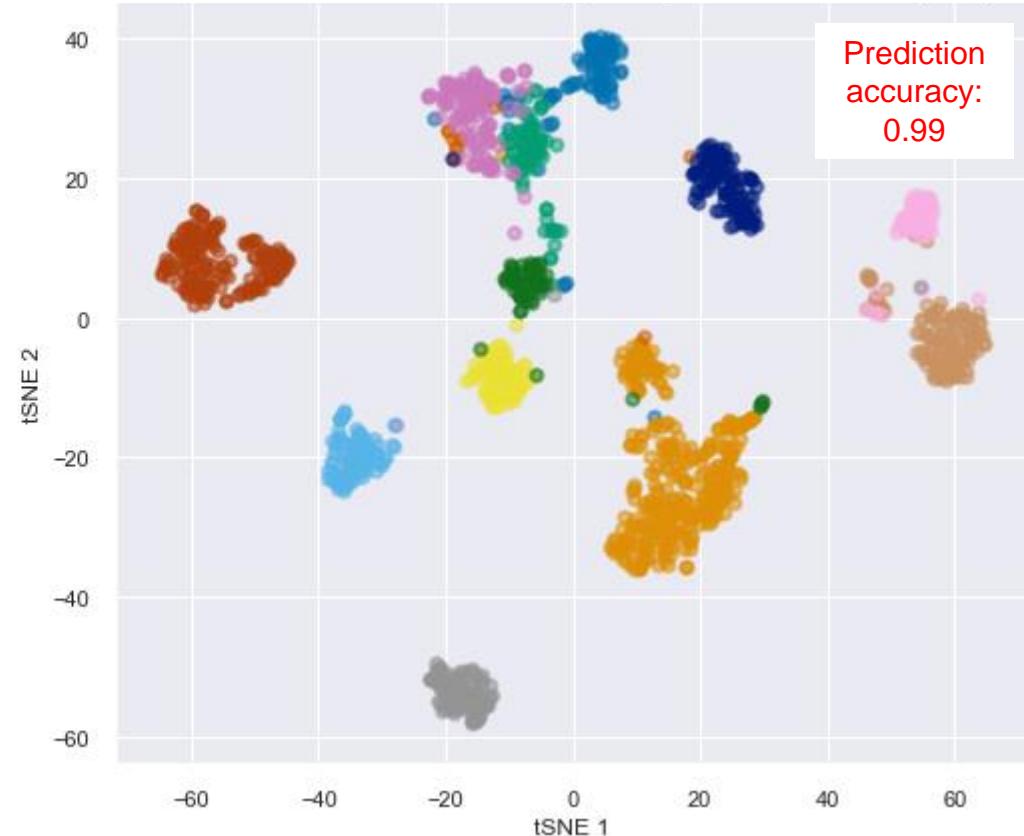
1. Primary cancer diagnosis from gene expression
2. Breast cancer patients' stratification based on gene expression

True values, tumor classification task



Validation set (colors correspond to true cancer types)

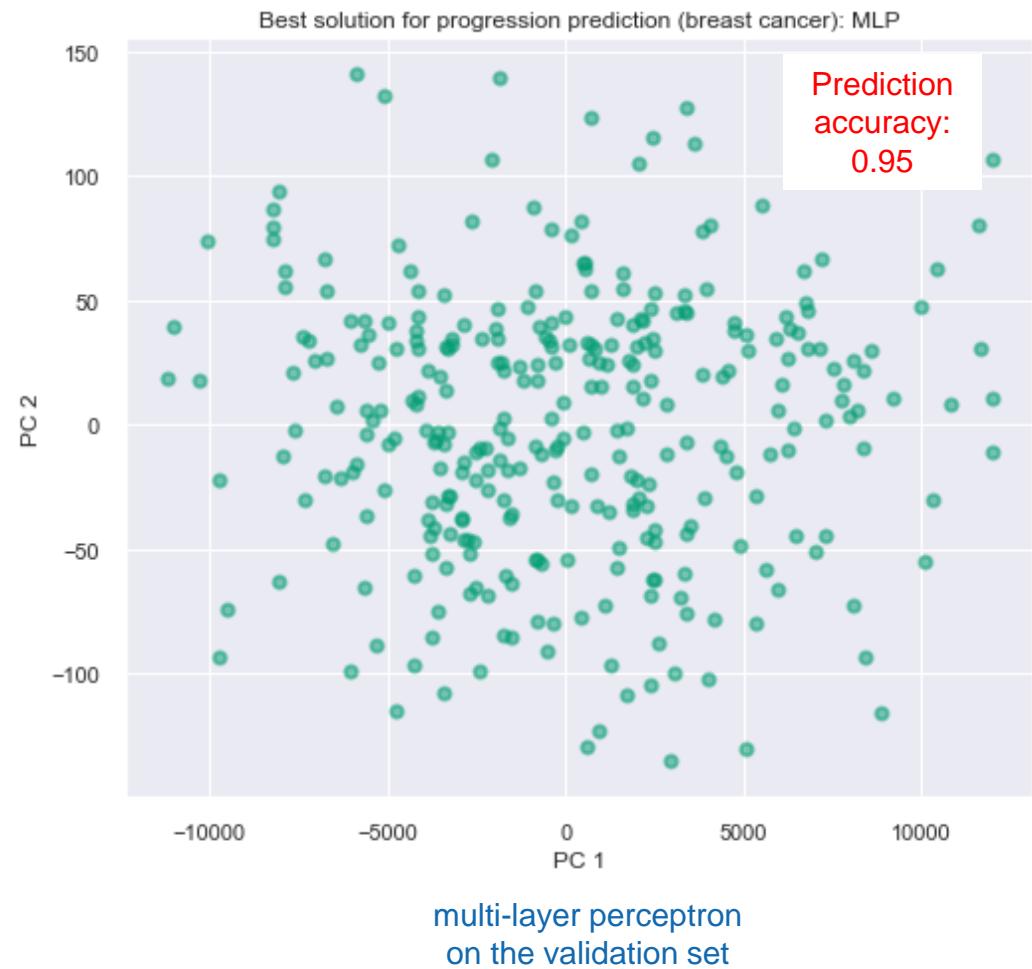
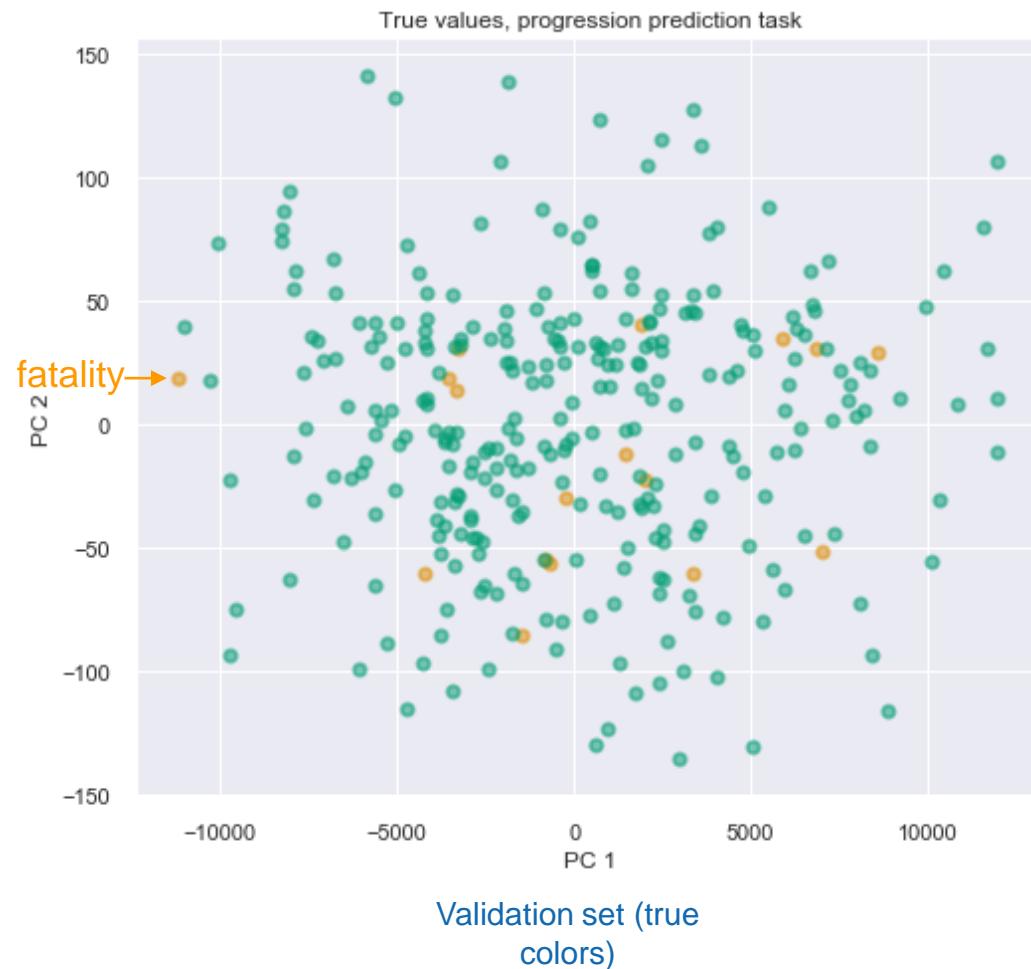
Best solution: Elastic net logistic regression



Elastic net on the validation set (colors correspond to predictions)

Hands-on:

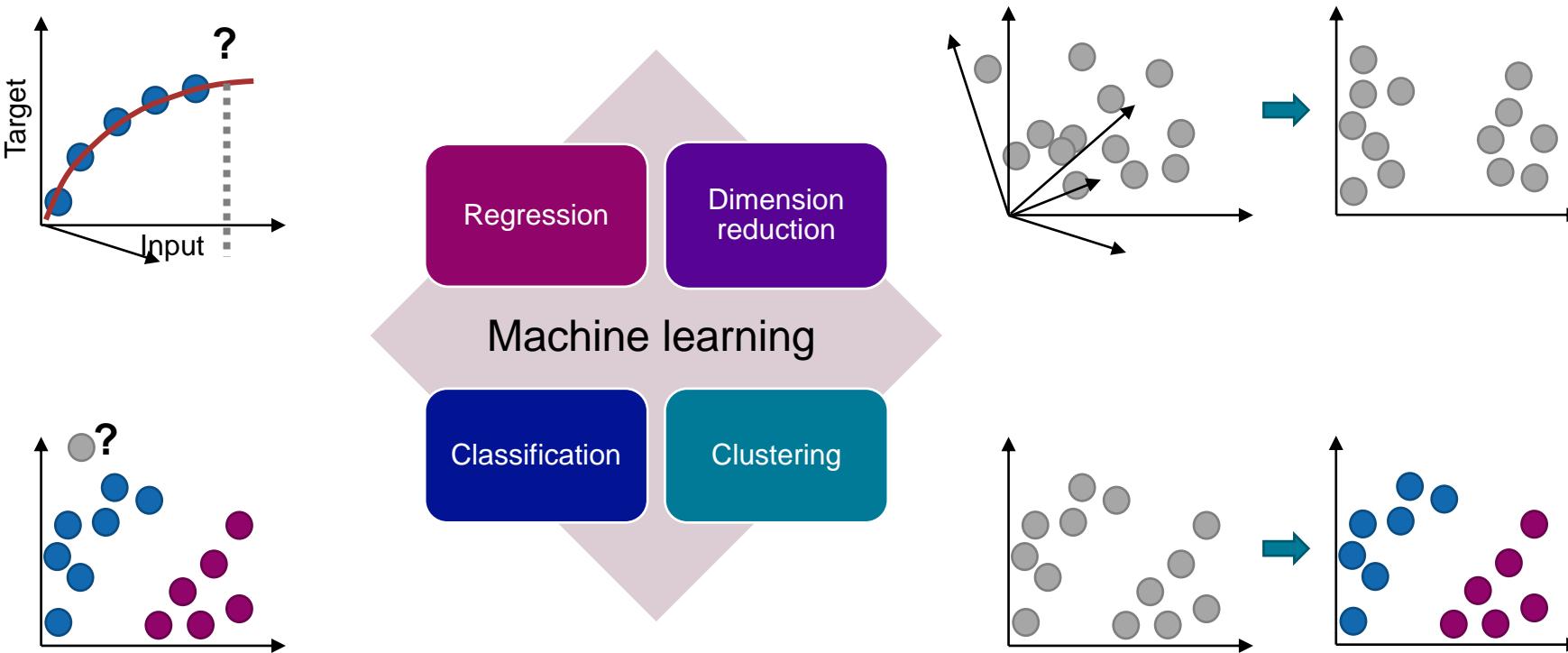
1. Primary cancer diagnosis from gene expression
2. Breast cancer patients' stratification based on gene expression



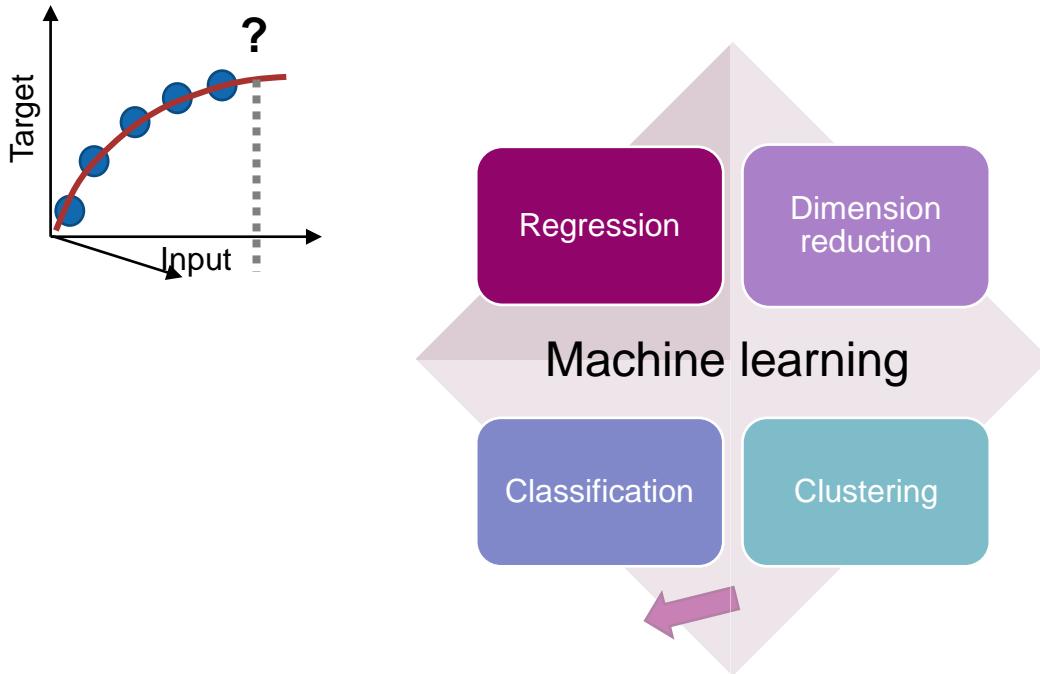
Take home message: Classification

- Linear and non-linear models can provide similar prediction accuracy (TASK1)
- Classification on imbalanced groups with low information content may fail (TASK2)
 - Study your data first
 - Check data summary
 - Visualize your data
 - Use the right evaluation metrics (e.g., precision and recall)
 - Consider redesigning your task

Map of classical machine learning methods

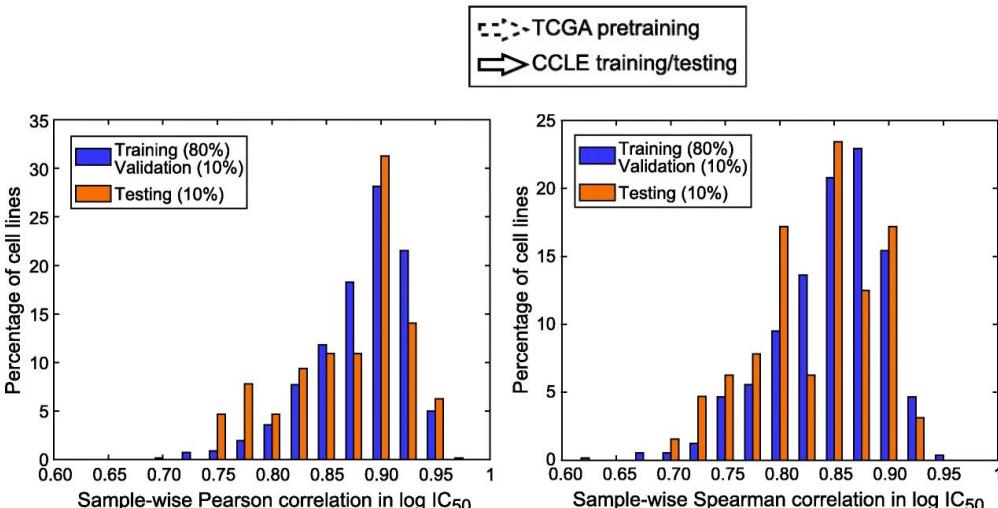
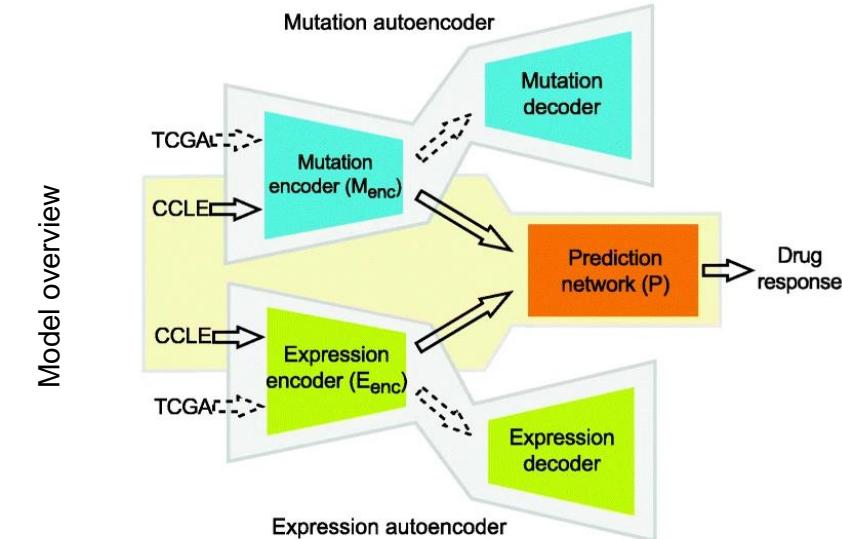


Map of classical machine learning methods



Regression: Biological examples

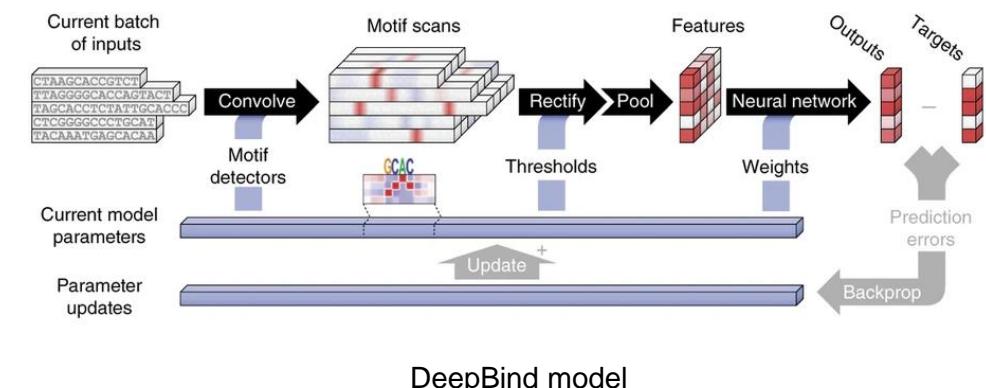
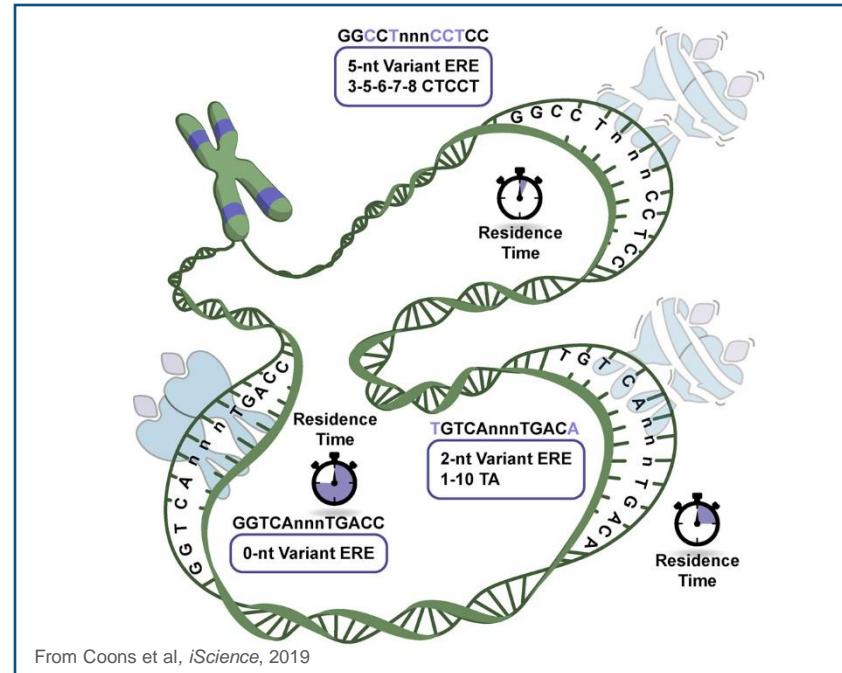
- Prediction of treatment efficiency / drug response
 - Predicting drug response of tumors from integrated genomic profiles by deep neural networks. Chiu et al., *BMC Med. Genomic*, 2019
- Prediction of molecular/cellular properties (e.g., protein-DNA binding affinities)
 - Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Alipanahi et al., *Nature Biotech.* 2015



Sample-wise Pearson and Spearman correlation between imputed and predicted IC_{50} data of CCLE samples

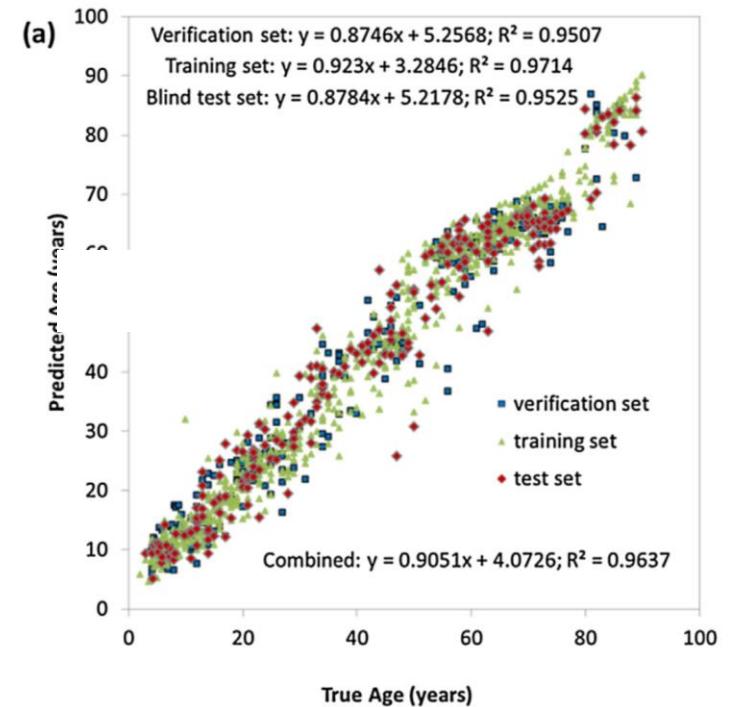
Regression: Biological examples

- Prediction of treatment efficiency / drug response
 - Predicting drug response of tumors from integrated genomic profiles by deep neural networks. Chiu et al., *BMC Med. Genomic*, 2019
- Prediction of molecular/cellular properties (e.g., protein-DNA binding affinities)
 - Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Alipanahi et al., *Nature Biotech.* 2015



Regression: Biological examples

- Age prediction from DNA methylation (e.g., for forensics)
 - **DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing.** Vidaki et al. *Forensic Sci Int Genet.* 2017



Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

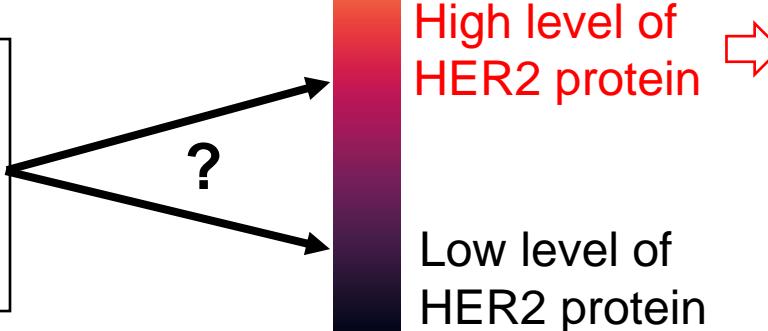
1. Concentration of HER2 (coded by the *ERBB2* gene)
2. Concentration of p53 (coded by the *TP53* gene)

<https://github.com/BoevaLab/Teaching>

TASK 1:


$$\begin{bmatrix} Gene1 & Gene2 & \cdots & Genep \\ Z_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Z_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

+ clinical stage + age



Treatment with HER2 inhibitors: Herceptin (trastuzumab) or Tykerb (lapatinib)

The **HER2** protein is coded by the *ERBB2* gene, frequently amplified in human breast cancer

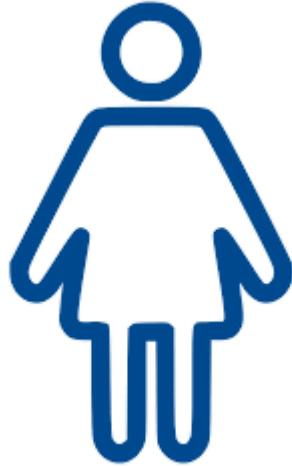
Target variable (y): Reverse Phase Protein Array (*RPPA*) value of HER2 presence in tumor cells

Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

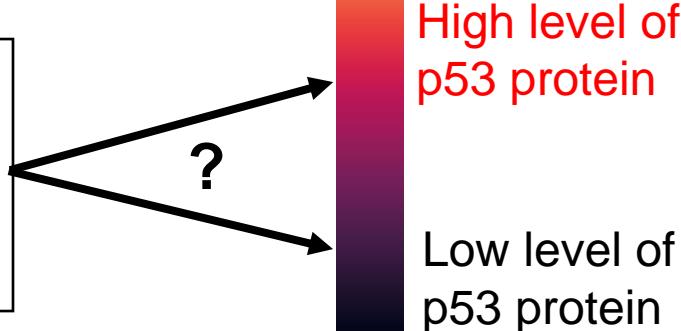
1. Concentration of HER2 (coded by the *ERBB2* gene)
2. Concentration of p53 (coded by the *TP53* gene)

<https://github.com/BoevaLab/Teaching>

TASK 2:


$$\begin{bmatrix} Gene1 & Gene2 & \cdots & Genep \\ Z_1 & X_{11} & X_{12} & \cdots & X_{1p} \\ Z_2 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_n & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

+ clinical stage + age



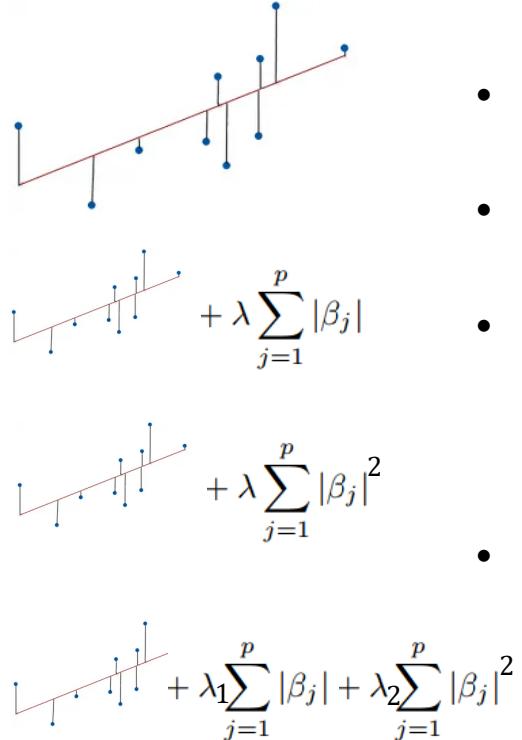
Standard chemotherapy, e.g., cisplatin

The p53 protein is coded by the *TP53 gene*, frequently deleted, mutated or repressed in human cancers

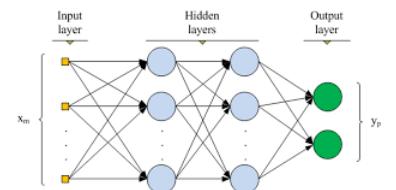
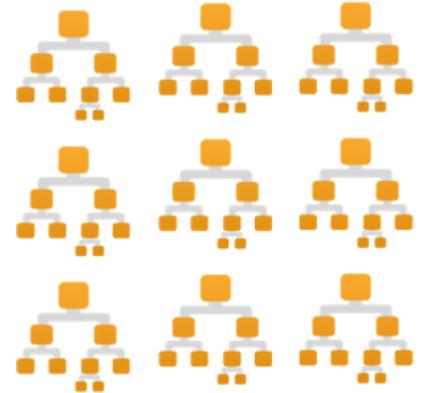
Target variable (y): Reverse Phase Protein Array (*RPPA*) value of p53 presence in tumor cells

Classic regression models

- Ordinary Least Squares
- Lasso (L1 penalty on model coefficients)
- Ridge (L2 penalty on model coefficients)
- Elastic Net (L1 and L2 penalty on model coefficients)

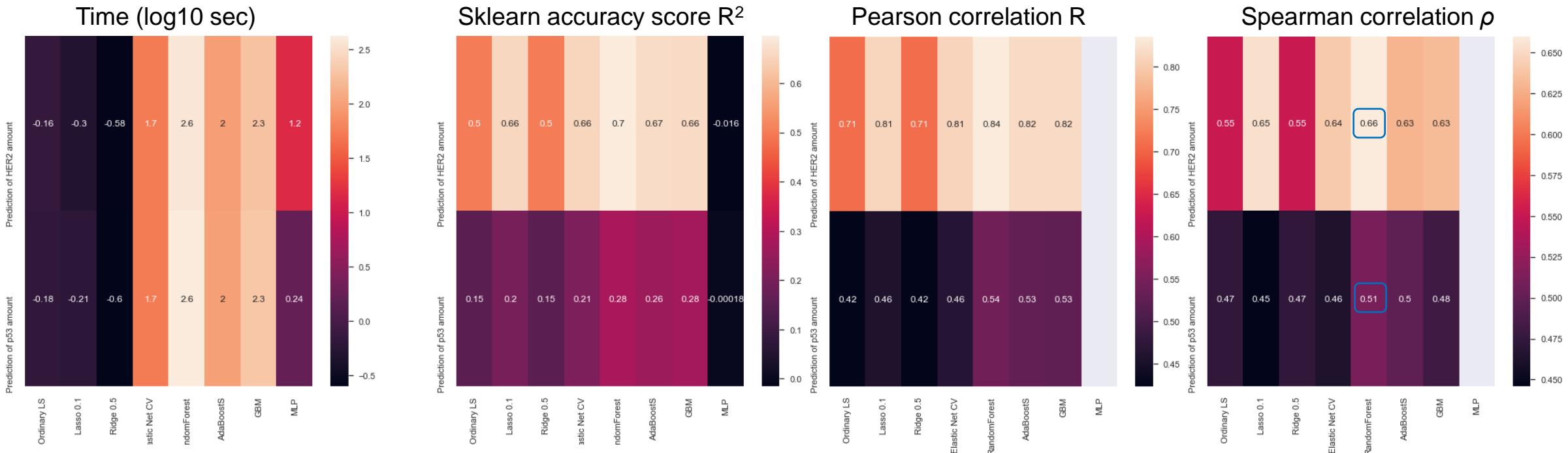


- Random Forest
- AdaBoost
- Gradient Tree Boosting (gradient boosting machine, GBM)
- Multi-layer perceptron (MLP)



Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

- Concentration of HER2 (coded by the *ERBB2* gene)
- Concentration of p53 (coded by the *TP53* gene)



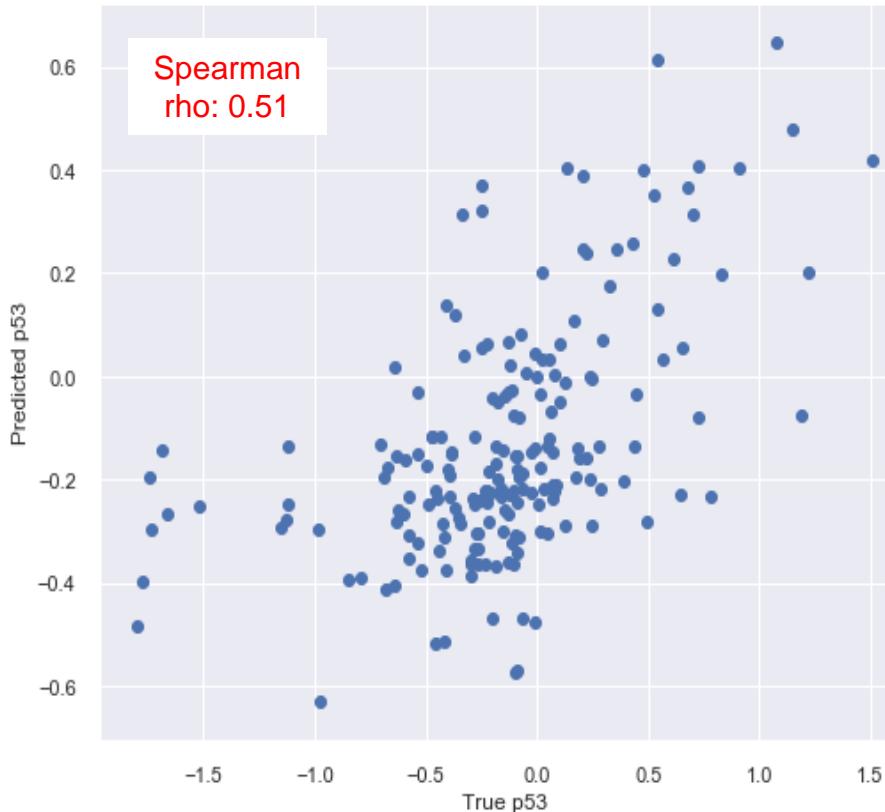
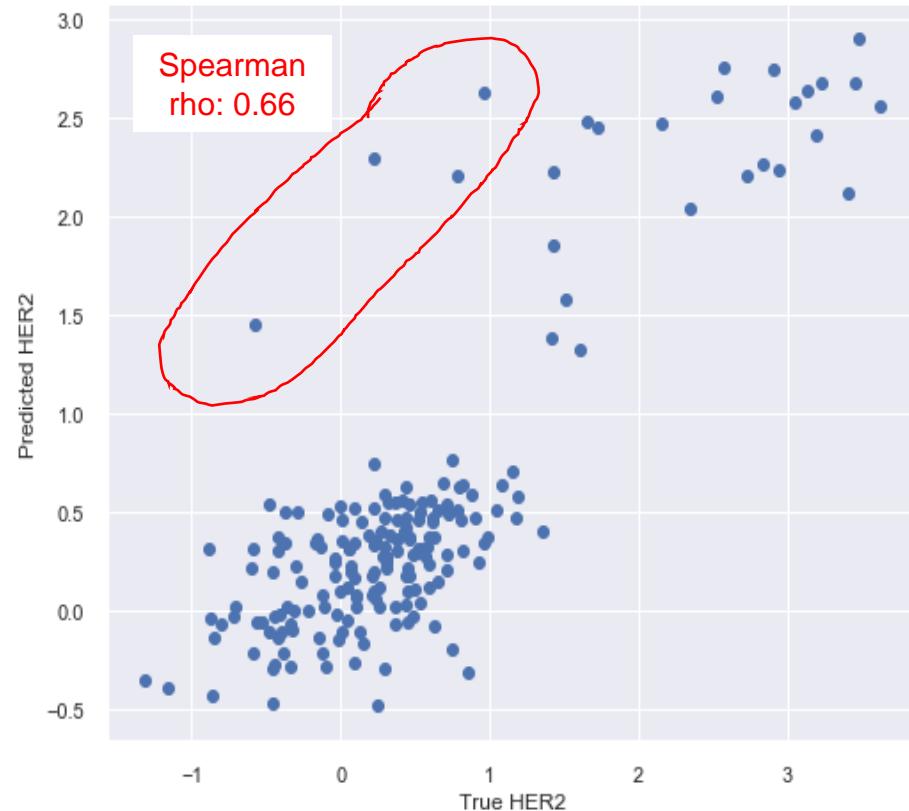
$$\text{Coefficient of determination } R^2 = \left(1 - \frac{\text{RSS}}{\sum(y_i - \bar{y})^2}\right)$$

$$\text{Spearman } \rho (\text{HER2}, \text{ERBB2}) = 0.63$$

$$\text{Spearman } \rho (\text{p53}, \text{TP53}) = 0.27$$

Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

1. Concentration of HER2 (coded by the *ERBB2* gene)
2. Concentration of p53 (coded by the *TP53* gene)

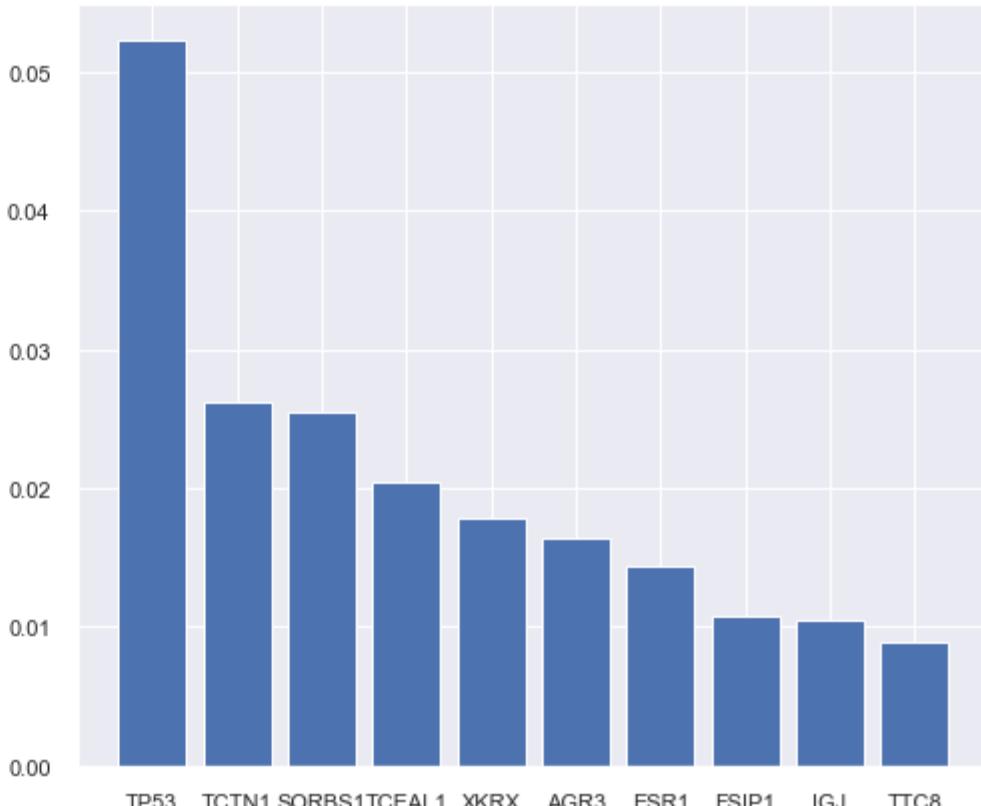
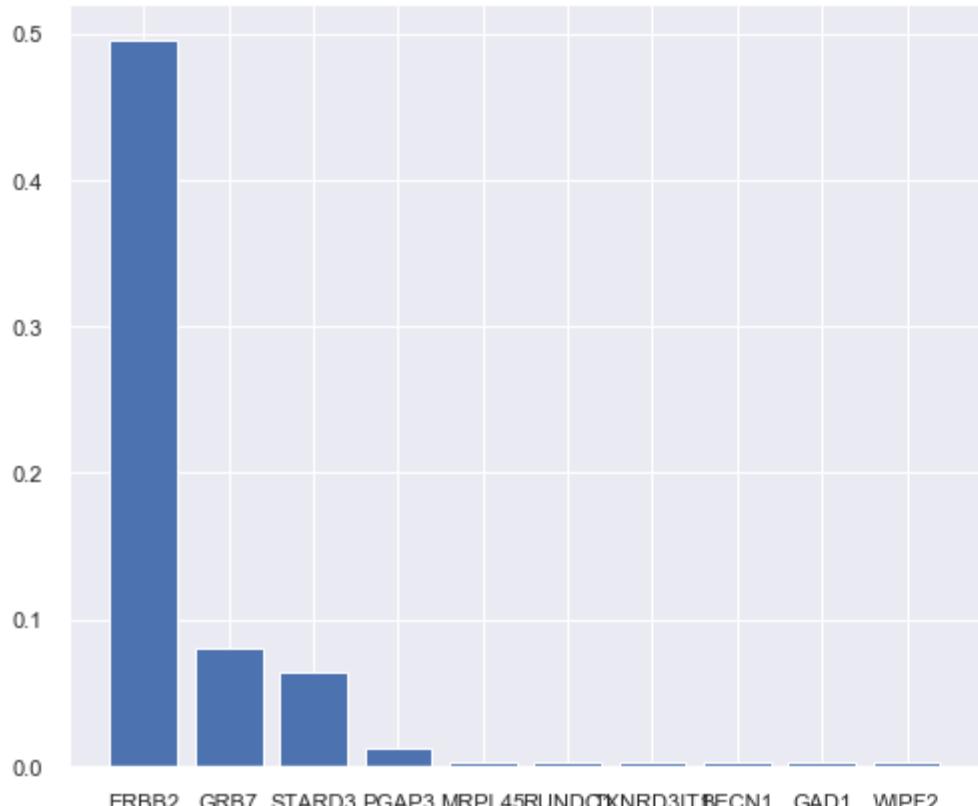


Random Forest predictions

Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

1. Concentration of HER2 (coded by the *ERBB2* gene)
2. Concentration of p53 (coded by the *TP53* gene)

FEATURE IMPORTANCE:

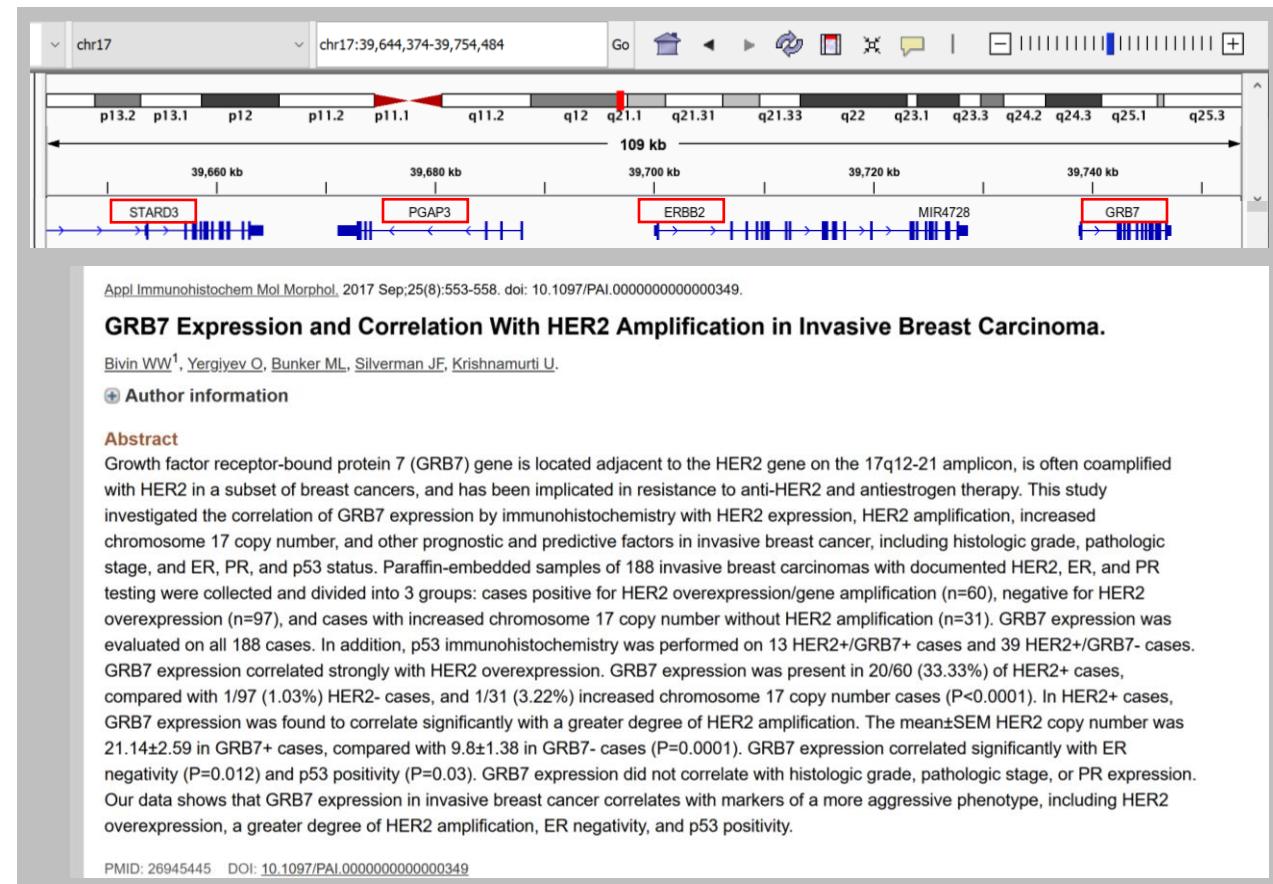
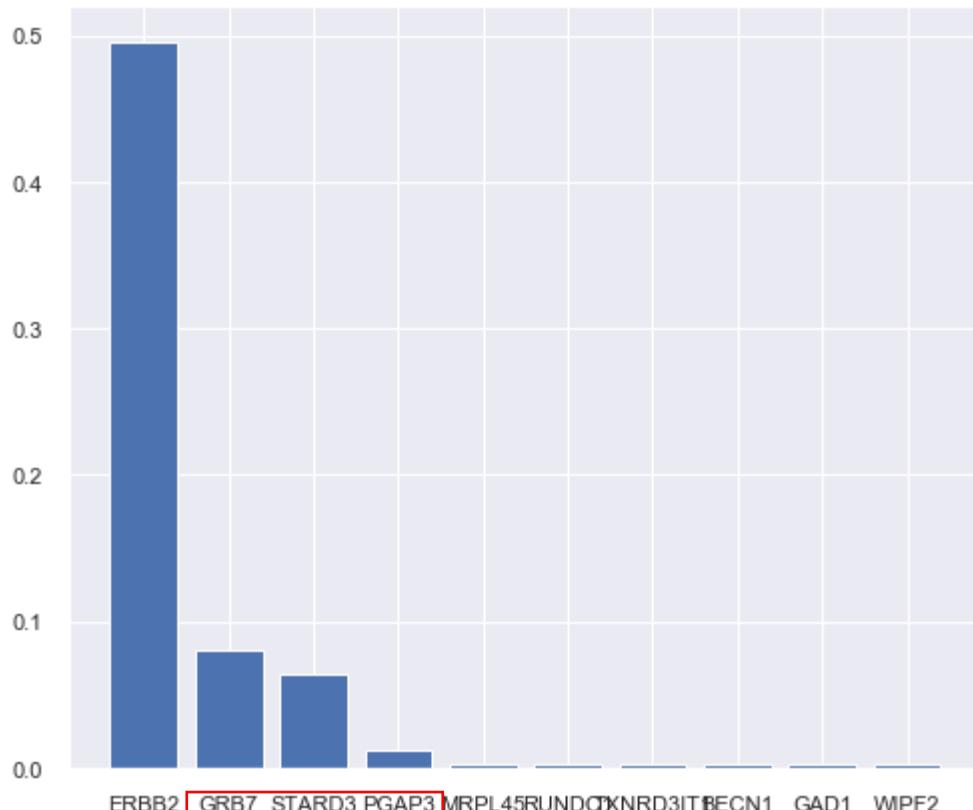


Random Forest predictions

Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

1. Concentration of HER2 (coded by the *ERBB2* gene)
2. Concentration of p53 (coded by the *TP53* gene)

FEATURE IMPORTANCE:



Hands-on: Prediction of protein concentration based on mRNA data (Breast cancer samples)

1. Concentration of HER2 (coded by the *ERBB2* gene)
2. Concentration of p53 (coded by the *TP53* gene)

FEATURE IMPORTANCE:



Review

The p53 Pathway in Glioblastoma

Ying Zhang^{1,†}, Collin Dube^{1,†}, Myron Gibert Jr.^{1,†}, Nichola Cruickshanks¹, Baomin Maeve Coughlan¹, Yanzhi Yang¹, Initha Setiady¹, Ciana Deveau¹, Karim Saoud¹, Cassandra Grello¹, Madison Oxford¹, Fang Yuan¹ and Roger Abounader^{1,2,3,*}

Meng et al. *Journal of Translational Medicine* 2014, **12**:288
http://www.translational-medicine.com/content/12/1/288



RESEARCH

Open Access

Expression and prognostic significance of TCTN1 in human glioblastoma

Delong Meng^{1†}, Yuanyuan Chen^{1†}, Yingjie Zhao¹, Jingkun Wang¹, Dapeng Yun¹, Song Yang², Juxiang Chen³, Hongyan Chen¹ and Daru Lu^{1*}

"ARF deletion [from p53 pathway] is correlated with overexpression of tectonic family member 1 (TCTN1), a protein involved in a diverse range of cellular processes, including promotion of GBM cell proliferation".

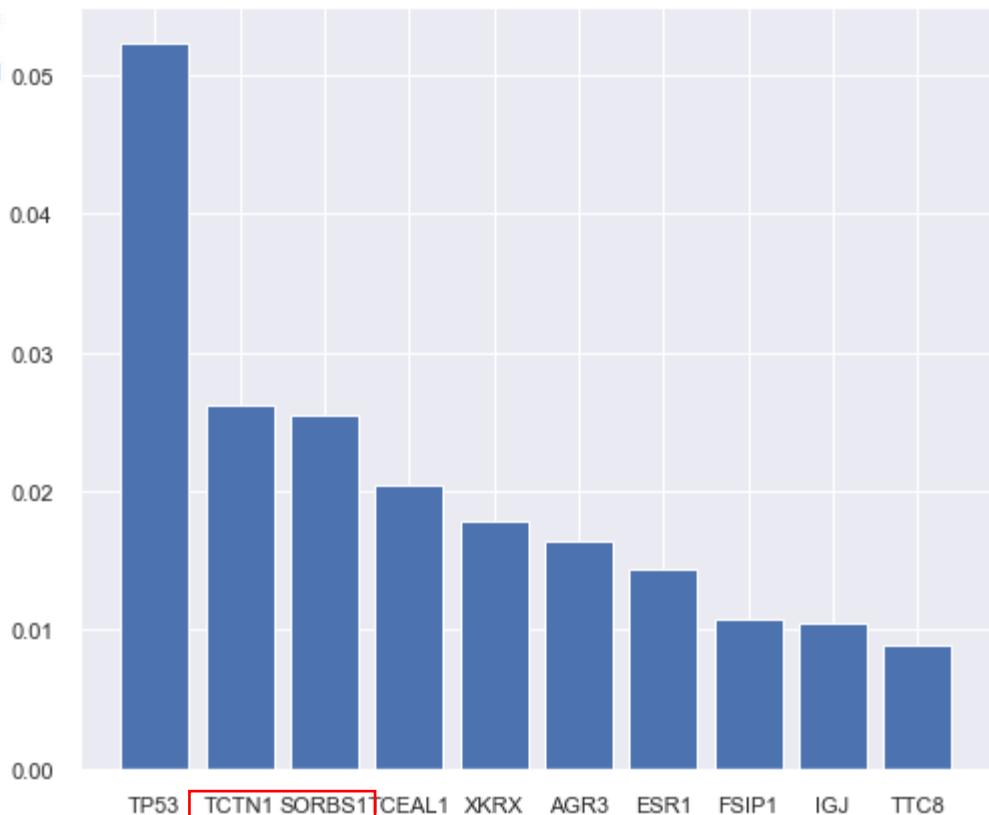
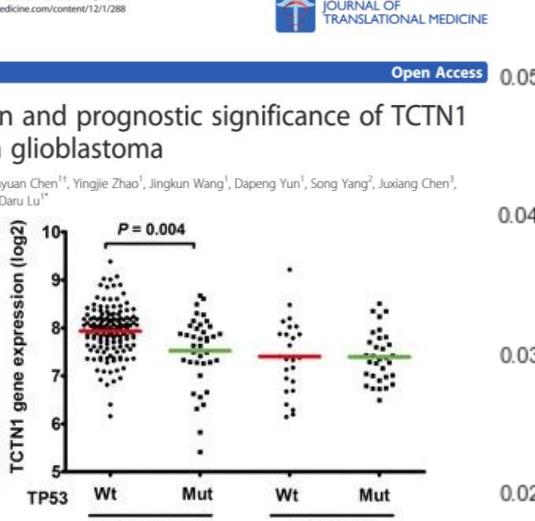
www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 6), pp: 9108-9122

Research Paper

SORBS1 suppresses tumor metastasis and improves the sensitivity of cancer to chemotherapy drug

Lele Song^{1,2}, Renxu Chang^{1,2}, Cheng Dai^{1,2}, Yanjun Wu^{1,2}, Jingyu Guo^{1,2}, Meiyang Qi¹, Wu Zhou³, Lixing Zhan¹

"Silencing of SORBS1 [...] attenuates chemical drug sensitivity especially that to cisplatin, by inhibition of p53 in **breast cancer cells**."

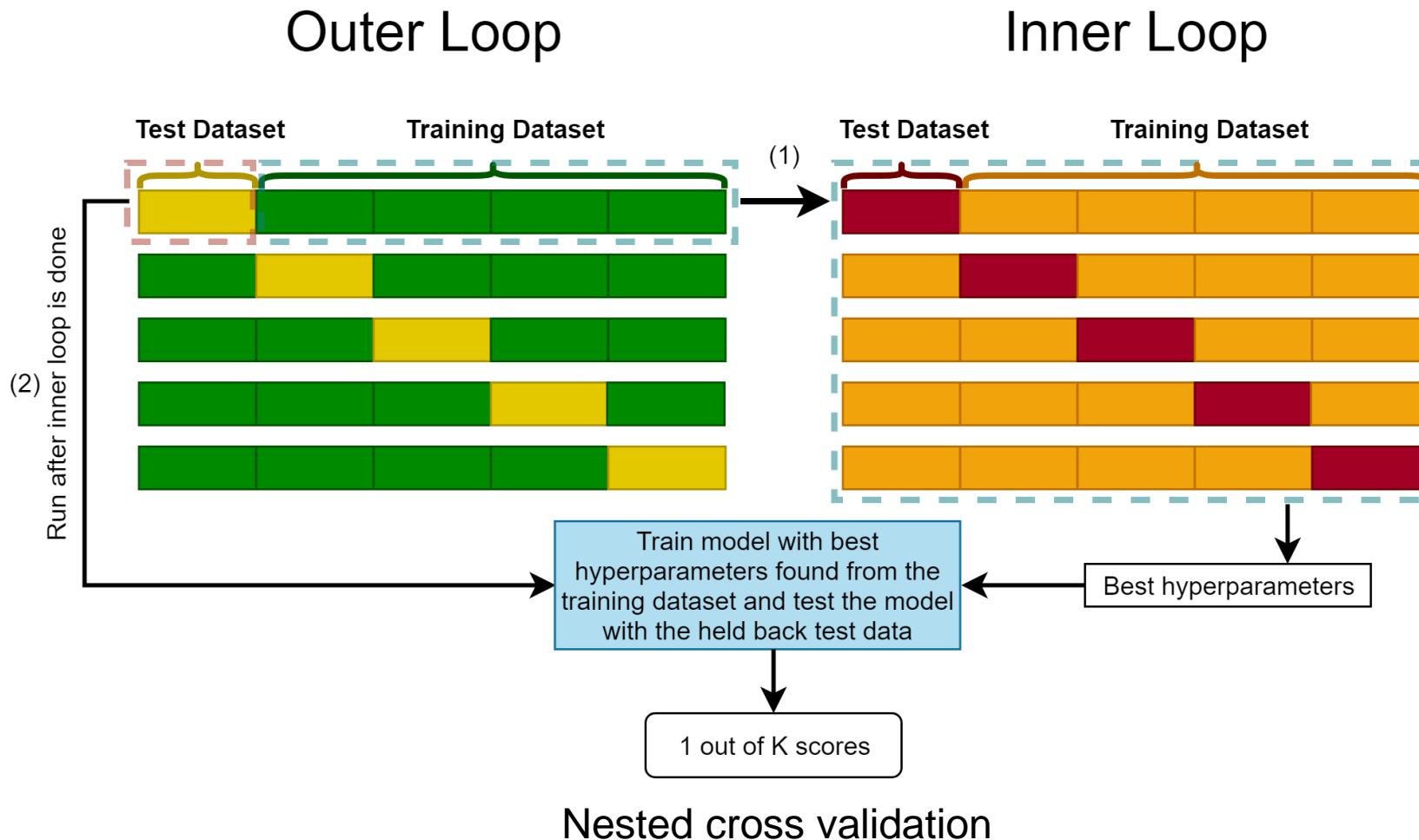


Take home message: Regression

- Regularized methods generally work better
 - Regularization may prevent over-fitting and select “important” features
- Regularized linear methods may provide accuracy similar to non-linear methods
- Neural networks do not always win
- Checking the feature importance may provide insights into biological mechanisms

Selection of hyperparameters via cross validation

What we did not do today, but in real life we should do it:



Take home message

- Classification and regression are extremely widely used in biology and medicine to automatize decisions of clinicians (diagnosis, choice of treatment) and predict treatment response and side effects
- The accuracy of predictions depends a lot on the information present in the data rather than on the ML method used
 - In our hands-on exercises the difference in accuracy between linear and non-linear methods varied between 0.5% and 15%
- There is no method that works the best in any situation
- Cross validation should be always applied to select the best hyperparameters
- In real life, one should compare a model built on omics data (+ clinical) with a model built using clinical variables only

Thank you for your attention!

Professor Valentina Boeva
valentina.boeva@inf.ethz.ch

ETH Zurich
Dept of Computer Science
CAB F51.2
Universitatstrasse 6
8092 Zurich, Switzerland

www.boevalab.com

