Supervised Classification

Francisco García García, fgarcia@cipf.es

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Outline

1. Introduction
2. Algorithms in Babelomics 5.0
3. Error Estimation of Prediction
4. Feature selection
5. Exercises on Babelomics

Outline

1. **Introduction**
2. Algorithms in Babelomics 5.0
3. Error Estimation of Prediction
4. Feature selection
5. Exercises on Babelomics

1 Brief History

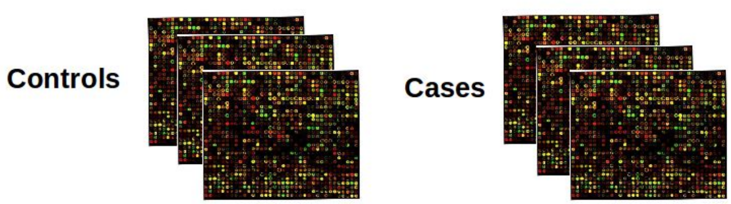
- **1956**, John McCarthy coined the term “Artificial Intelligence” and defined it as “the science and engineering of making intelligent machines.”
- **60-70's**, the mathematical background of some of these algorithms/methods were developed.
- **70-80's**, with the apparition of computers first predictors were developed for many different areas:
  - handwriting recognition
  - weather prediction
  - face recognition
  - speech recognition

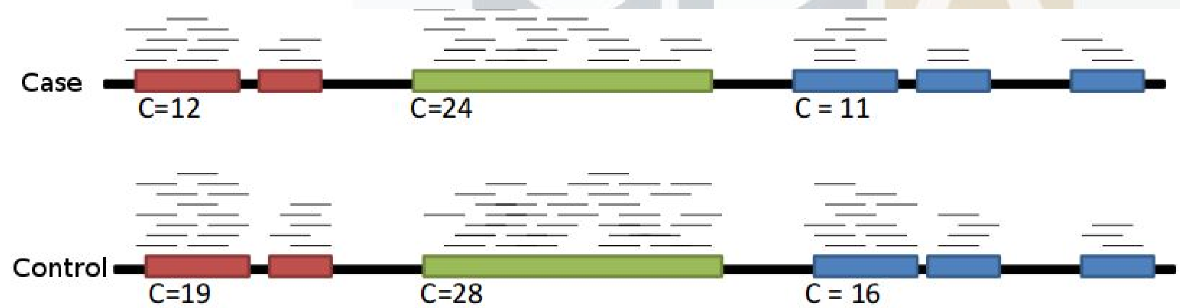
1 Brief History

- **90's,** predictors begun to be used in Bioinformatics and Computational Biology:
  - gene prediction: sequence based gene annotation
  - genome annotation: sequence based TFBS, exons, ... annotation
  - protein structure prediction
- **In late 90's**, DNA microarray technology was developed:
- **In early 2000**, two questions arose:
  - could biological samples be classified according to gene expression?
  - and, could we use computers to help us classifying these samples?

1 Brief History

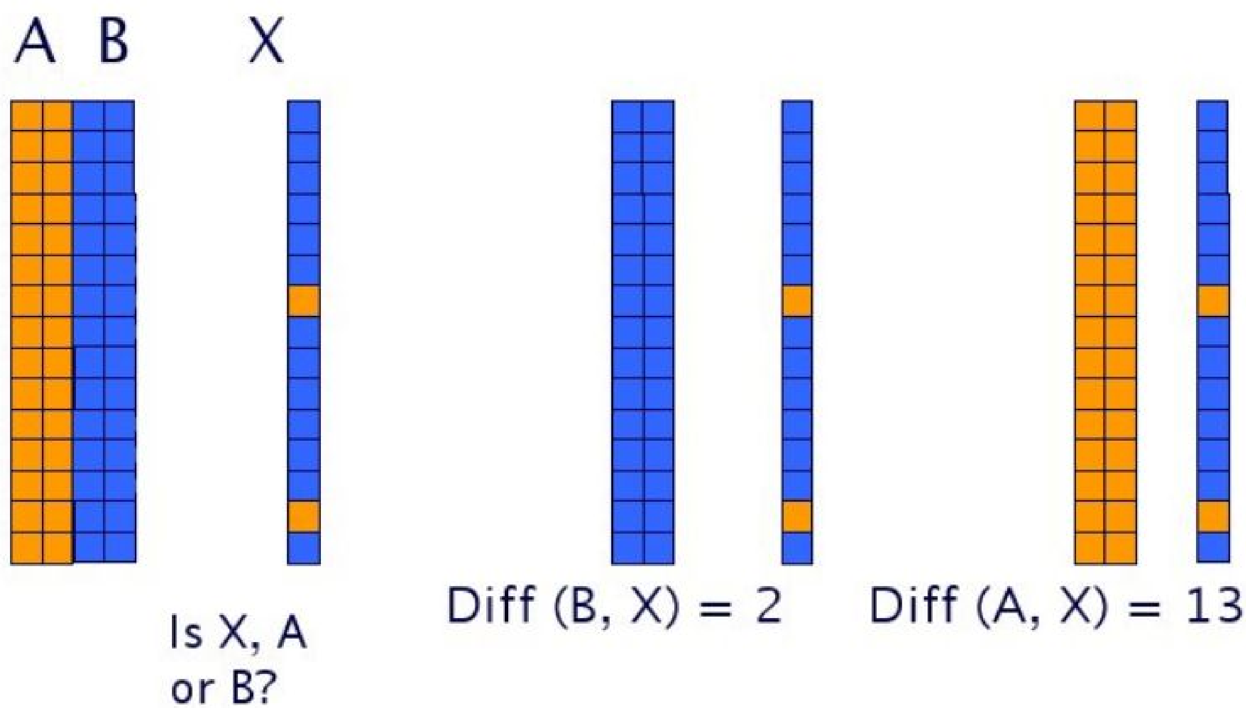
- **2002**, appears the first paper applying predictors method to DNA microarray data, van't Veer et al. (http://www.ncbi.nlm.nih.gov/pubmed/11823860)
- Since that moment hundreds of papers, applications and new methods have been developed which are also used for NGS data (e.g. RNAseq).





1 Introduction

- Set of algorithms/methods that are going to allow the computer to **learn** a specific **labelled** problem and then be able to predict or classify new unlabelled samples is called supervised classification.
- Predictors/classifiers are a subset of methods of the Artificial Intelligence area.
- Predictors need to be trained.



1 Introduction

Known Class Data (*Babelomics format*)

Class Label

GeneIDs / ProbeIDs

#VARIABLE *tumor* CATEGORICAL{c1,c2,c3} VALUES{c1,c1,c2,c2,c3,c3}

#NAMES	GSM26878	GSM26883	GSM26886	GSM26887	GSM26903	GSM26910
1007_s_at	11.08578155	11.04457022	11.02479206	11.00837346	11.04430518	11.01921026
1053_at	7.787503325	8.010263804	7.872064511	7.711140759	7.703846348	7.509845931
117_at	7.487539205	7.526590226	7.442468793	7.394731634	7.450764725	7.558967177
121_at	9.589979282	9.516503297	9.610811352	9.282059896	8.323068371	8.664237594
1255_g_at	5.000099854	5.127166256	4.952998877	4.881038876	4.948734762	5.087888404
1294_at	8.358097049	8.403219181	8.255863646	7.947778797	8.328705461	8.230633848
1316_at	7.187245349	6.652952654	6.445444909	6.463659189	6.399722565	6.404821127
1320_at	5.645994428	5.765206267	5.772052661	5.609287091	5.621417391	5.723352308
1405_i_at	7.138444163	7.490198393	7.382302176	7.379200666	7.541671446	6.493521779
1431_at	4.697298725	4.722480562	4.795825627	4.703361751	4.701914661	4.904298823
1438_at	7.430761532	8.112797873	7.578819384	7.699611607	7.496504531	7.776384116
1487_at	7.646126117	7.544048497	8.754540699	8.476873549	9.084035203	9.028724488
1494_f_at	7.498031252	7.679595836	7.662561072	7.201093115	7.426192546	7.669669586
1598_g_at	10.31770877	10.92530764	10.50092321	9.630201704	10.23473332	10.49766918
160020_at	8.529411037	8.738065073	8.617216353	8.445386532	8.425365655	8.76023381
1729_at	9.607320487	8.171988017	8.73040537	8.978602862	9.156752025	8.033237589
1773_at	6.216319215	6.441555855	6.165785507	6.325464779	6.121753223	6.229420354
177_at	6.535525364	6.453887146	6.519400663	6.333366799	6.385077422	6.407541976

arrays

Select train data

Unknown Class Data

which is the class label?

1007_s_at	11.28578155
1053_at	7.787503325
117_at	7.487539205
121_at	9.489979282
1255_g_at	5.000099854
1294_at	8.358097049
1316_at	7.187245349
1320_at	5.645994428
1405_i_at	7.138444163
1431_at	4.697298725
1438_at	7.430761532
1487_at	8.646126117
1494_f_at	7.498031252
1598_g_at	10.31770877
160020_at	8.529411037
1729_at	9.107320487
1773_at	6.216319215
177_at	6.535525364

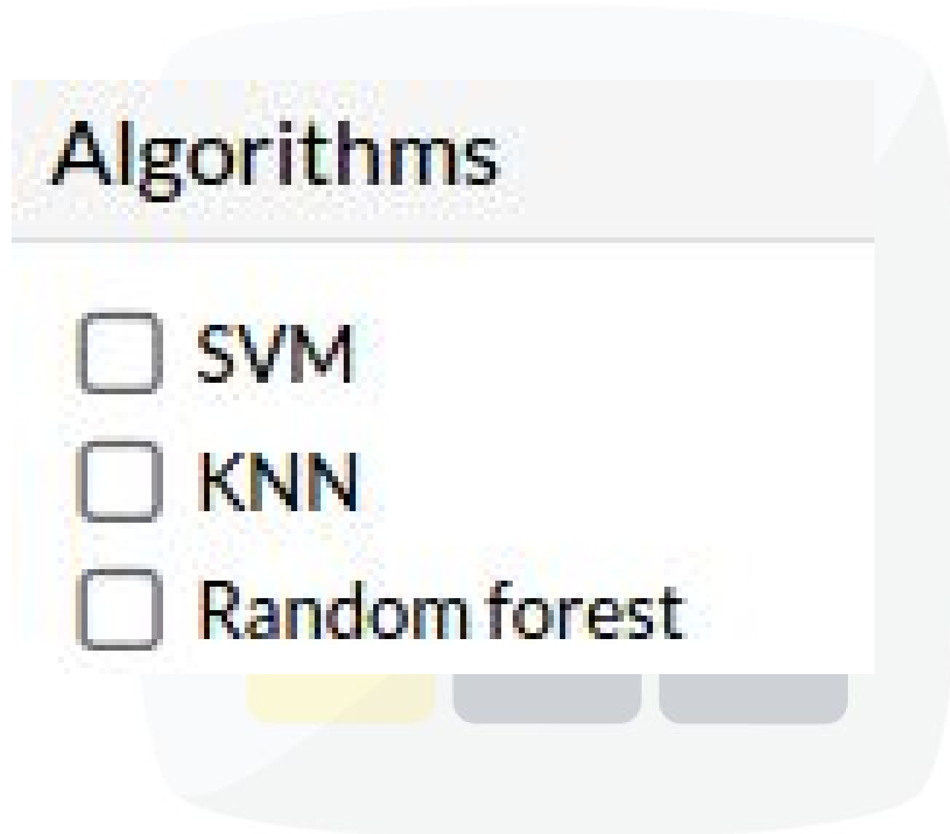
Select test data (Optional)

Outline

1. Introduction
2. **Algorithms in Babelomics 5.0**
3. Error Estimation of Prediction
4. Feature selection
5. Exercises on Babelomics

2 Algorithms in Babelomics 5.0

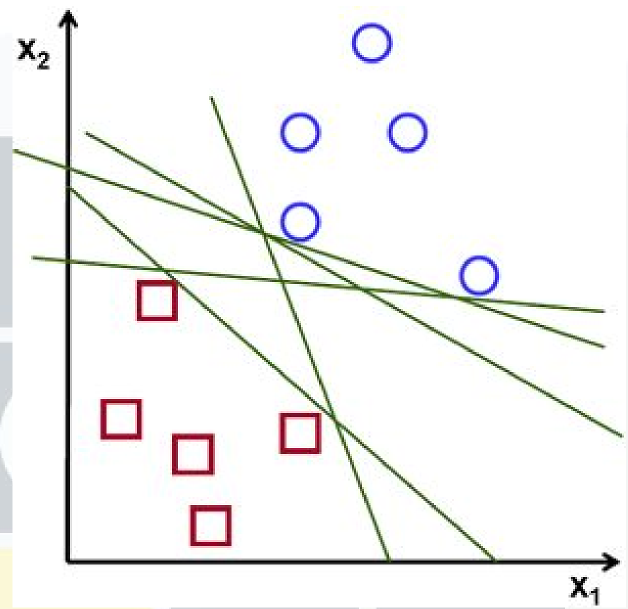
- Support Vector Machine (SVM)
- k-Nearest Neighbors (KNN)
- Random Forest



2 SVM

A **Support Vector Machine (SVM)** is a discriminative classifier formally defined by a separating hyperplane.

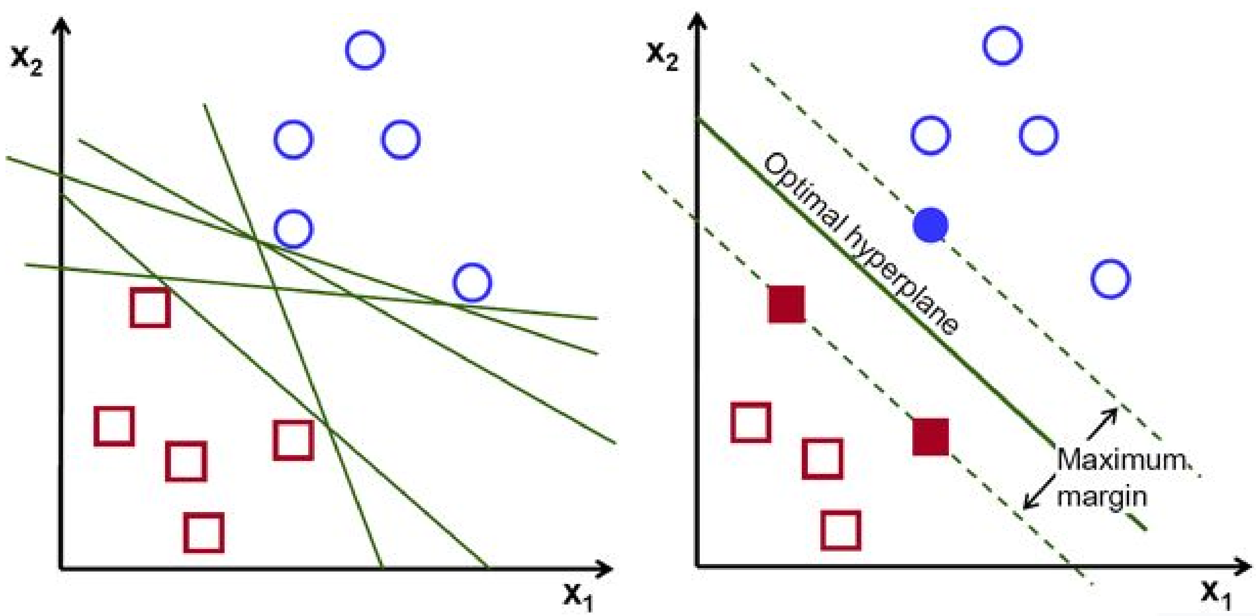
In the picture you can see that there exists multiple lines that offer a solution to the problem.



Is any of them better than the others?

2 SVM

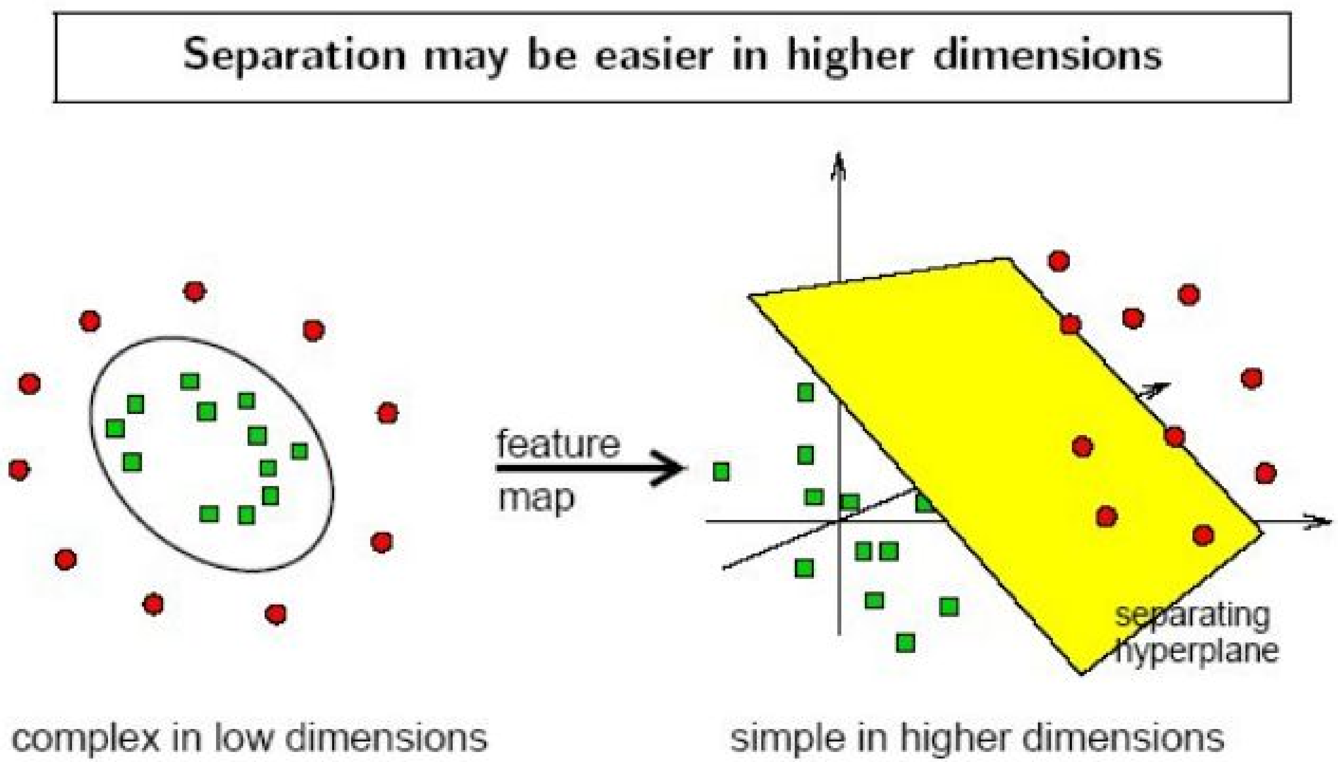
A special property of SVMs is that maximizes the margin between the decision hyperplane and the training examples.



2 SVM

But many times the data does not have a linear solution.

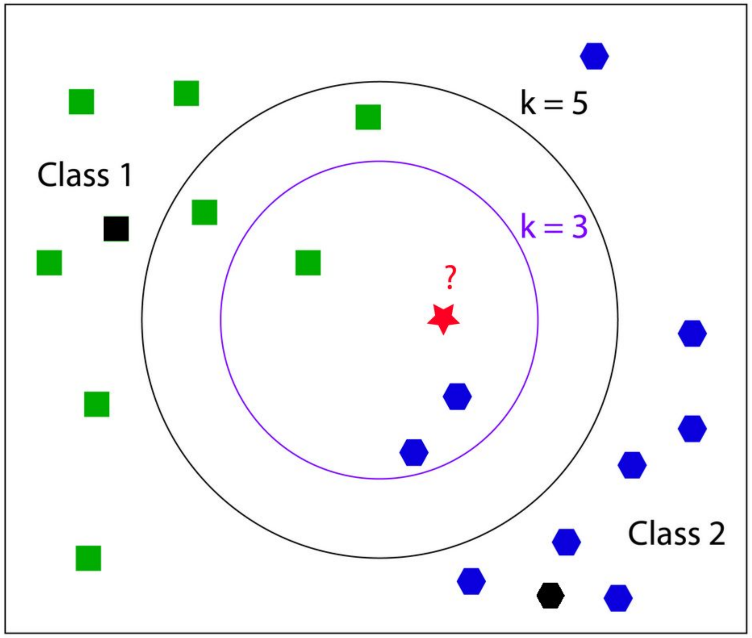
Then we can use a **kernel trick** and map the data into a higher dimension space.



2 KNN

k-Nearest Neighbor (KNN) is a distance based prediction method.

In order to make KNN more robust we are going to look for the K nearest neighbours instead of only the nearest.

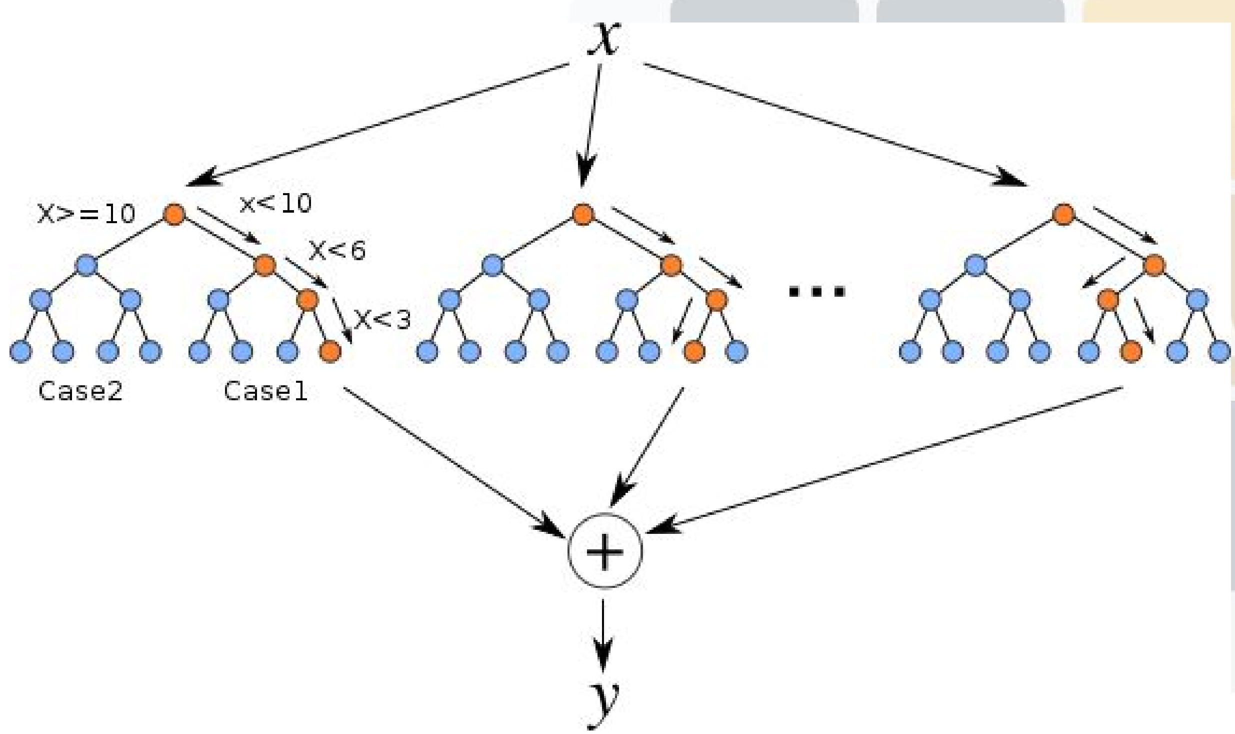


Predictor tool test automatically with **K=1..20**

2 Random Forest

Random Forest builds many trees using a subset of the available input variables and their values.

The forest chooses the classification having the most votes over all the trees in the forest.



Outline

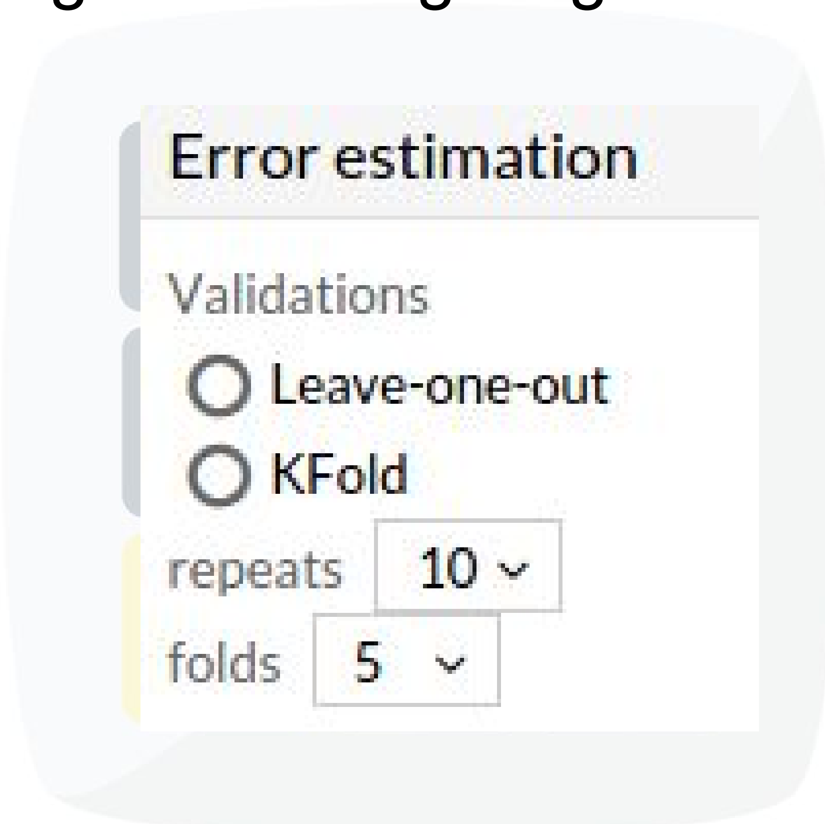
1. Introduction
2. Algorithms in Babelomics 5.0
3. **Error Estimation of Prediction**
4. Feature selection
5. Exercises on Babelomics

3 Error Estimation of Prediction

It is not a simple task, we have to estimate the error that the predictor will have in future gene expression data.

This estimation can only be done during the training stage.

- Leaving-one-out cross-validation
- k-fold cross-validation

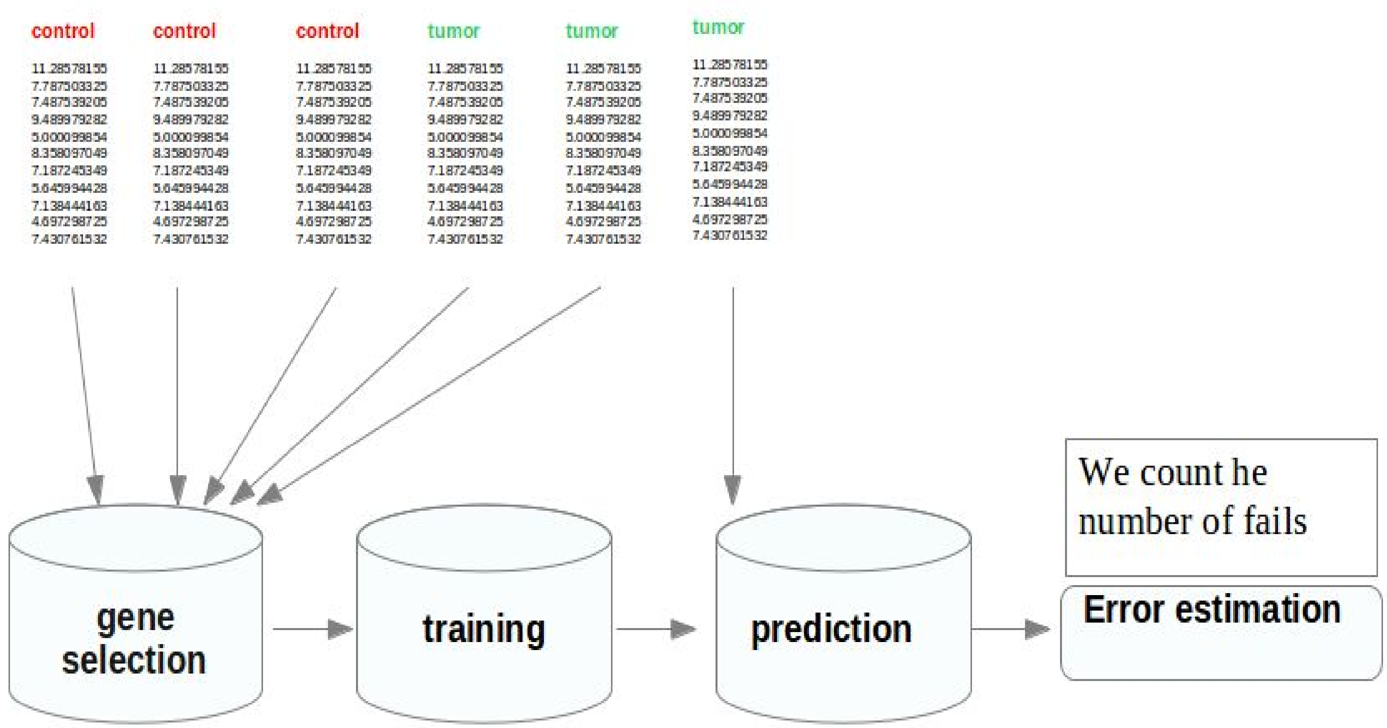


3 Leaving-one-out cross-validation

We select one sample to use as a test set and the rest as a training set.

k = number of arrays   i.e.: k=6

Repeat k times changing the array to be used in prediction



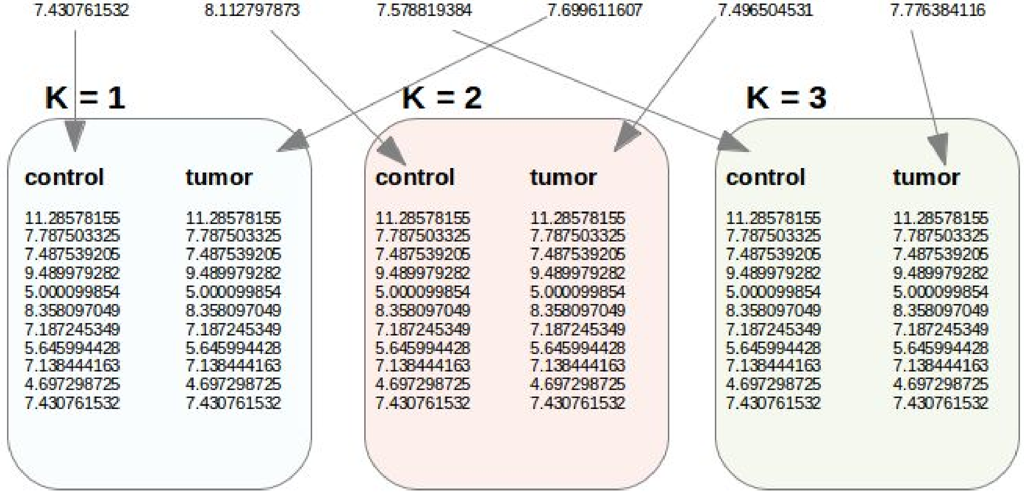
3 k-fold cross-validation

With this method we are going to split the data in k partitions and we will use (k-1) partitions to train the the other to test the predictor.

We split arrays into k partitions of equal size, i.e.: k=3

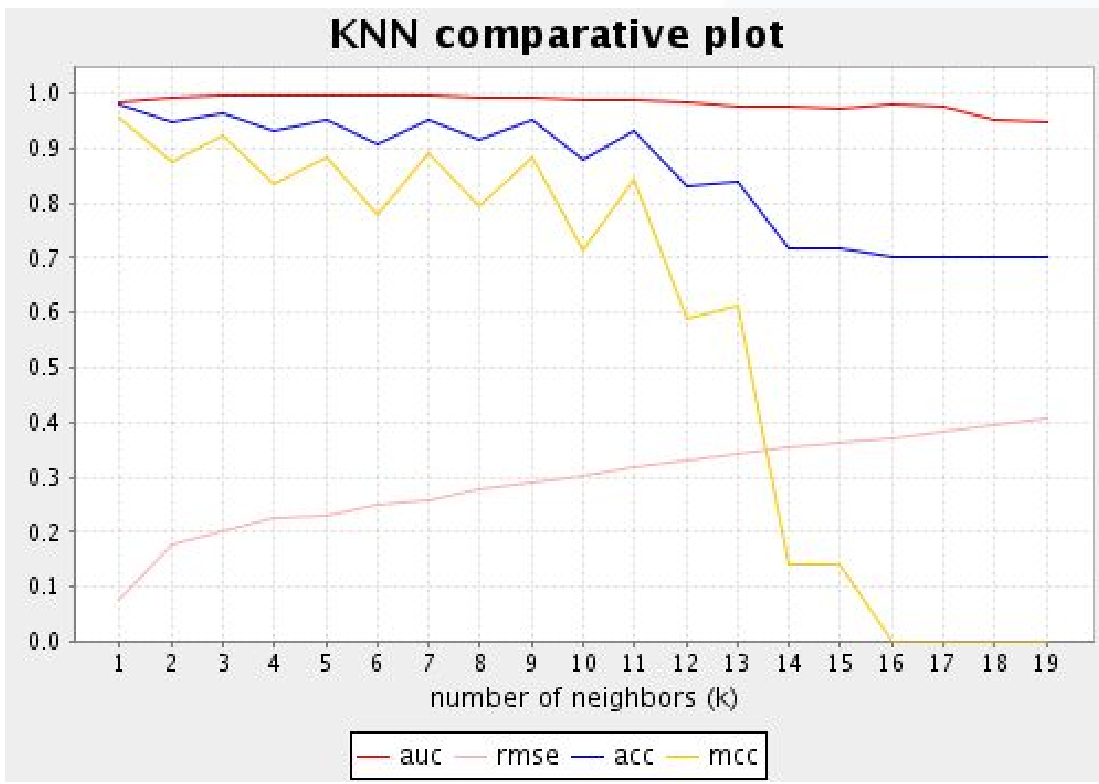
#VARIABLE **PHEN** CATEGORICAL{control,tumor} VALUES{control,control,control,tumor,tumor,tumor}

#NAMES	GSM26878	GSM26883	GSM26886	GSM26887	GSM26903	GSM26910
1007_s_at	11.08578155	11.04457022	11.02479206	11.00837346	11.04430518	11.01921026
1053_at	7.787503325	8.010263804	7.872064511	7.711140759	7.703846348	7.509845931
117_at	7.487539205	7.526590226	7.442468793	7.394731634	7.450764725	7.558967177
121_at	9.589979282	9.516503297	9.610811352	9.282059896	8.323068371	8.664237594
1255_g_at	5.000099854	5.127166256	4.952998877	4.881038876	4.948734762	5.087888404
1294_at	8.358097049	8.403219181	8.255863646	7.947778797	8.328705461	8.230633848
1316_at	7.187245349	6.652952654	6.445444909	6.463659189	6.399722565	6.404821127
1320_at	5.645994428	5.765206267	5.772052661	5.609287091	5.621417391	5.723352308
1405_i_at	7.138444163	7.490198393	7.382302176	7.379200666	7.541671446	6.493521779
1431_at	4.697298725	4.722480562	4.795825627	4.703361751	4.701914661	4.904298823
1438_at	7.430761532	8.112797873	7.578819384	7.699611607	7.496504531	7.776384116



3 Error Estimation of Metrics

- Accuracy (ACC)
- Area Under ROC (AUC)
- Matthews correlation coefficient (MCC)
- Root Mean Square Error (RMSE)



3 Error Estimation of Metrics

Accuracy (ACC)

$$ACC = (TP + TN) / (P + N)$$

TP (True Positive)	FP (False Positive)	P* (Predicted Positive)
FN (False Negative)	TN (True Negative)	N* (Predicted Negative)
P (Total Positive)	N (Total Negative)	D (Total Documents)

Acc = 92%

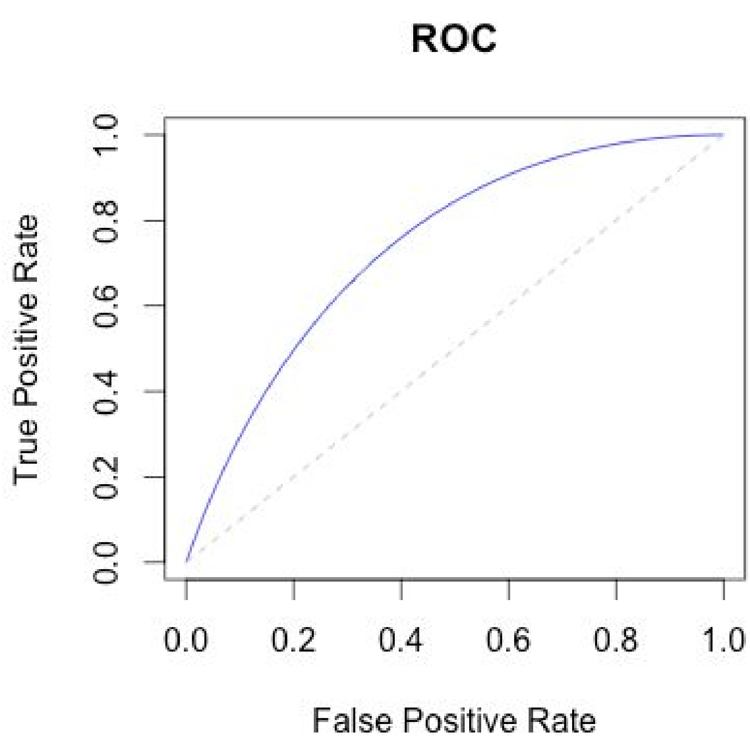
46	4
4	46

90	0
8	2

3 Error Estimation of Metrics

Area Under ROC (AUC)

The best classification has the largest area under the curve.



3 Error Estimation of Metrics

Matthews correlation coefficient (MCC)

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{[TP + FP][TP + FN][TN + FP][TN + FN]}}$$

TP (True Positive)	FP (False Positive)	P* (Predicted Positive)
FN (False Negative)	TN (True Negative)	N* (Predicted Negative)
P (Total Positive)	N (Total Negative)	D (Total Documents)

3 Error Estimation of Metrics

The **Root-Mean-Square Error (RMSE)** is the square root of the average value of the square of the residual (actual - predicted)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Outline

1. Introduction
2. Algorithms in Babelomics 5.0
3. Error Estimation of Prediction
4. **Feature selection**
5. Exercises on Babelomics

4 Feature selection

- Finding genes that discriminate classes.
- Genes that change randomly within classes are not useful for classifying.
- Genes that do not change do not bear any information.

#VARIABLE *tumor* CATEGORICAL{c1,c2,c3} VALUES{c1,c1,c2,c2,c3,c3}

#NAMES	GSM26878	GSM26883	GSM26886	GSM26887	GSM26903	GSM26910
1007_s_at	11.08578155	11.04457022	11.02479206	11.00837346	11.04430518	11.01921026
1053_at	7.787503325	8.010263804	7.872064511	7.711140759	7.703846348	7.509845931
117_at	7.487539205	7.526590226	7.442468793	7.394731634	7.450764725	7.558967177
121_at	9.589979282	9.516503297	9.610811352	9.282059896	8.323068371	8.664237594
1255_g_at	5.000099854	5.127166256	4.952998877	4.881038876	4.948734762	5.087888404
1294_at	8.358097049	8.403219181	8.255863646	7.947778797	8.328705461	8.230633848
1316_at	7.187245349	6.652952654	6.445444909	6.463659189	6.399722565	6.404821127
1320_at	5.645994428	5.765206267	5.772052661	5.609287091	5.621417391	5.723352308
1405_i_at	7.138444163	7.490198393	7.382302176	7.379200666	7.541671446	6.493521779
1431_at	4.697298725	4.722480562	4.795825627	4.703361751	4.701914661	4.904298823
1438_at	7.430761532	8.112797873	7.578819384	7.699611607	7.496504531	7.776384116
1487_at	7.646126117	7.544048497	8.754540699	8.476873549	9.084035203	9.028724488
1494_f_at	7.498031252	7.679595836	7.662561072	7.201093115	7.426192546	7.669669586
1598_g_at	10.31770877	10.92530764	10.50092321	9.630201704	10.23473332	10.49766918
160020_at	8.529411037	8.738065073	8.617216353	8.445386532	8.425365655	8.76023381
1729_at	9.607320487	8.171988017	8.73040537	8.978602862	9.156752025	8.033237589
1773_at	6.216319215	6.441555855	6.165785507	6.325464779	6.121753223	6.229420354
177_at	6.535525364	6.453887146	6.519400663	6.333366799	6.385077422	6.407541976

4 Feature selection

#VARIABLE ***tumor*** **CATEGORICAL{c1,c2,c3} VALUES{c1,c1,c2,c2,c3,c3}**

#NAMES	GSM26878	GSM26883	GSM26886	GSM26887	GSM26903	GSM26910
1007_s_at	11.08578155	11.04457022	11.02479206	11.00837346	11.04430518	11.01921026
1053_at	7.787503325	8.010263804	7.872064511	7.711140759	7.703846348	7.509845931
117_at	7.487539205	7.526590226	7.442468793	7.394731634	7.450764725	7.558967177
121_at	9.589979282	9.516503297	9.610811352	9.282059896	8.323068371	8.664237594
1255_g_at	5.000099854	5.127166256	4.952998877	4.881038876	4.948734762	5.087888404
1294_at	8.358097049	8.403219181	8.255863646	7.947778797	8.328705461	8.230633848
1316_at	7.187245349	6.652952654	6.445444909	6.463659189	6.399722565	6.404821127
1320_at	5.645994428	5.765206267	5.772052661	5.609287091	5.621417391	5.723352308
1405_i_at	7.138444163	7.490198393	7.382302176	7.379200666	7.541671446	6.493521779
1431_at	4.697298725	4.722480562	4.795825627	4.703361751	4.701914661	4.904298823
1438_at	7.430761532	8.112797873	7.578819384	7.699611607	7.496504531	7.776384116
1487_at	7.646126117	7.544048497	8.754540699	8.476873549	9.084035203	9.028724488
1494_f_at	7.498031252	7.679595836	7.662561072	7.201093115	7.426192546	7.669669586
1598_g_at	10.31770877	10.92530764	10.50092321	9.630201704	10.23473332	10.49766918
160020_at	8.529411037	8.738065073	8.617216353	8.445386532	8.425365655	8.76023381
1729_at	9.607320487	8.171988017	8.73040537	8.978602862	9.156752025	8.033237589
1773_at	6.216319215	6.441555855	6.165785507	6.325464779	6.121753223	6.229420354
177_at	6.535525364	6.453887146	6.519400663	6.333366799	6.385077422	6.407541976

unknown class data

1007_s_at	11.28578155
1053_at	7.787503325
117_at	7.487539205
121_at	9.489979282
1255_g_at	5.000099854
1294_at	8.358097049
1316_at	7.187245349
1320_at	5.645994428
1405_i_at	7.138444163
1431_at	4.697298725
1438_at	7.430761532
1487_at	8.646126117
1494_f_at	7.498031252
1598_g_at	10.31770877
160020_at	8.529411037
1729_at	9.107320487
1773_at	6.216319215
177_at	6.535525364

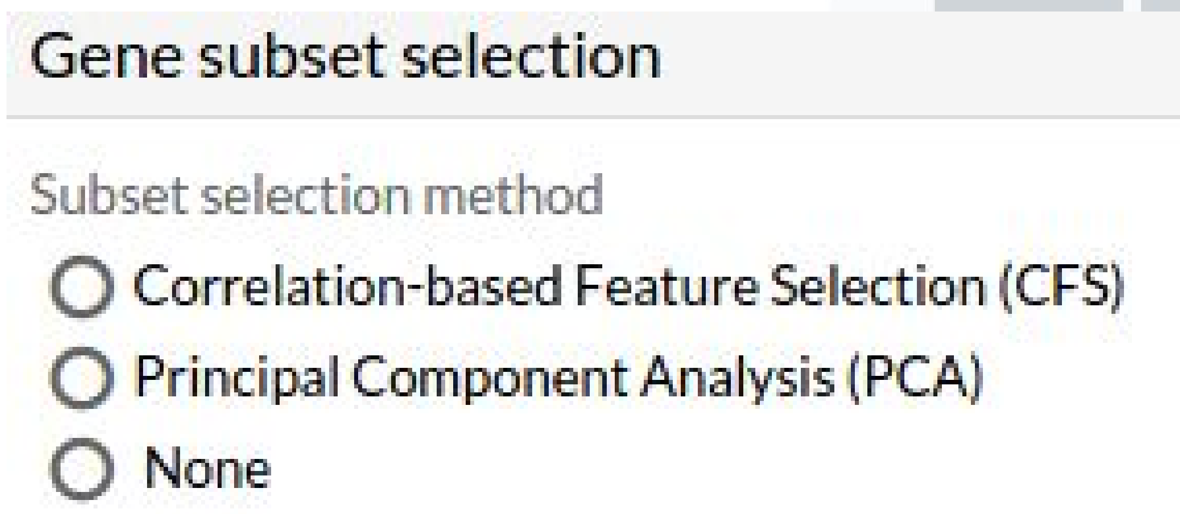
class 2?

4 Feature selection

- Feature selection is the technique of selecting a subset of relevant features for building robust learning models.
- Decreases analysis time.
- Increases the accuracy.

4 Feature selection

- Correlation-based Feature Selection (CFS)
- Principal Components Analysis (PCA)

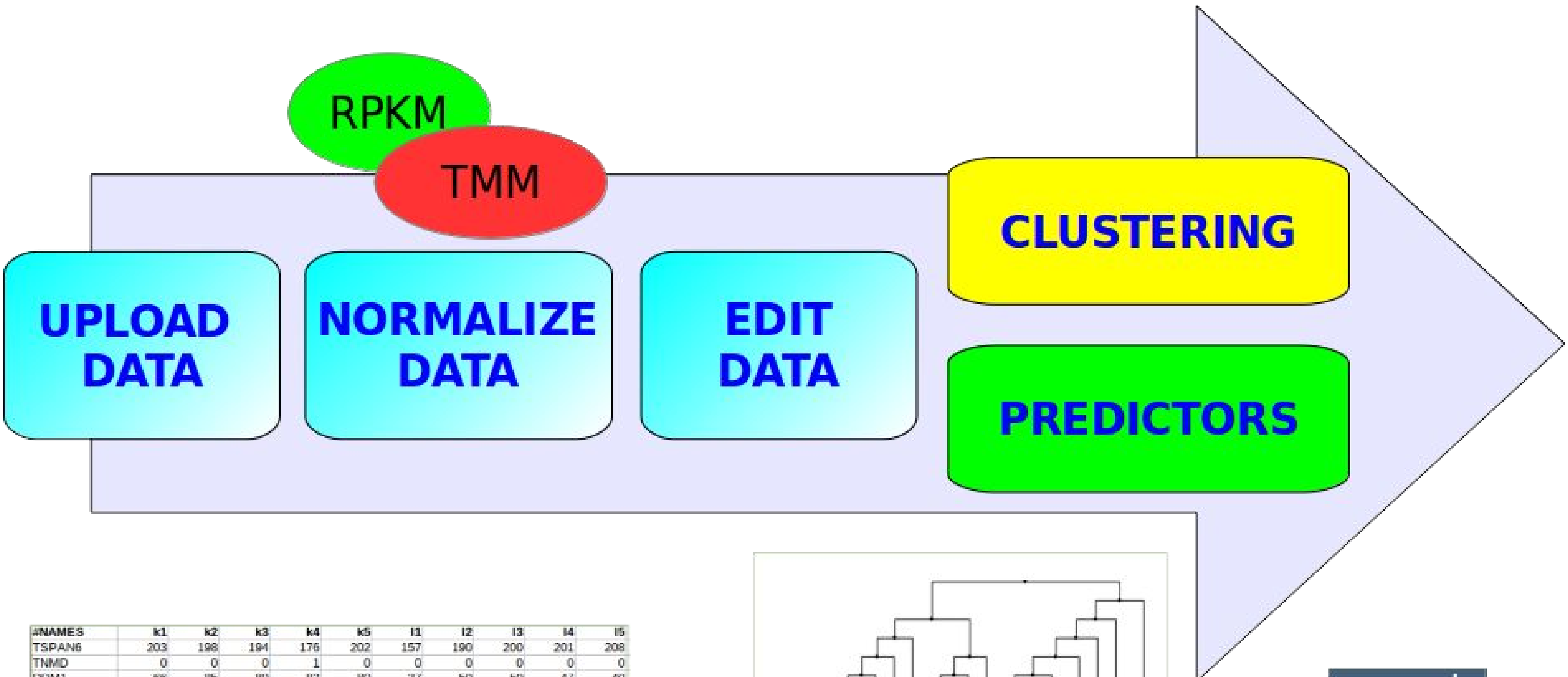


4 Any questions?

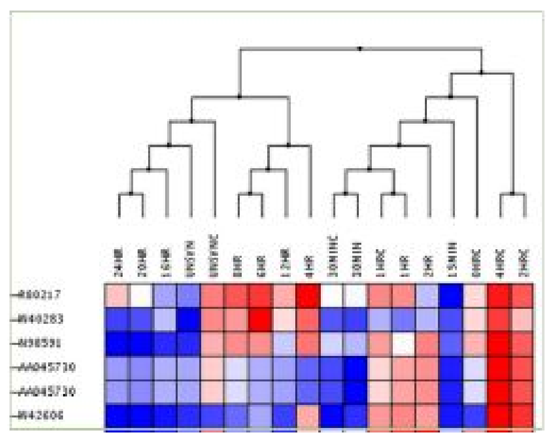
Outline

1. Introduction
2. Algorithms in Babelomics 5.0
3. Error Estimation of Prediction
4. Feature selection
5. **Exercises on Babelomics**

5 Exercises on Babelomics



#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPG1	34	43	41	31	46	6	7	7	8	7



5 Exercises on Babelomics

1. **Go to Babelomics**

2. **Worked example + exercises**