Clustering

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: http://krcadinac.com

Clustering

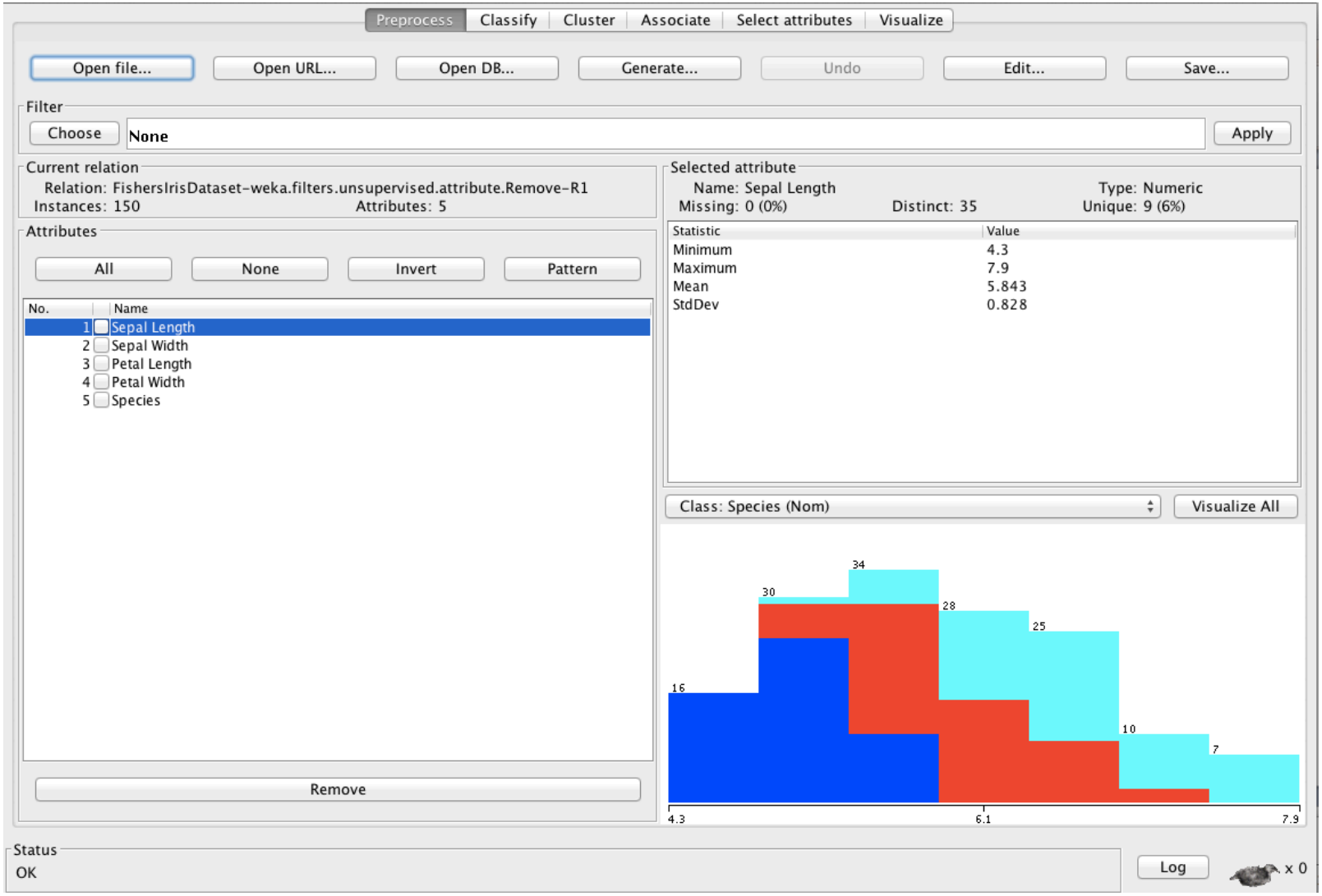
Clustering belongs to a group of techniques of unsupervised learning. It enables grouping instances into groups, where we know which are the possible groups in *advance*.

These groups are called **clusters**.

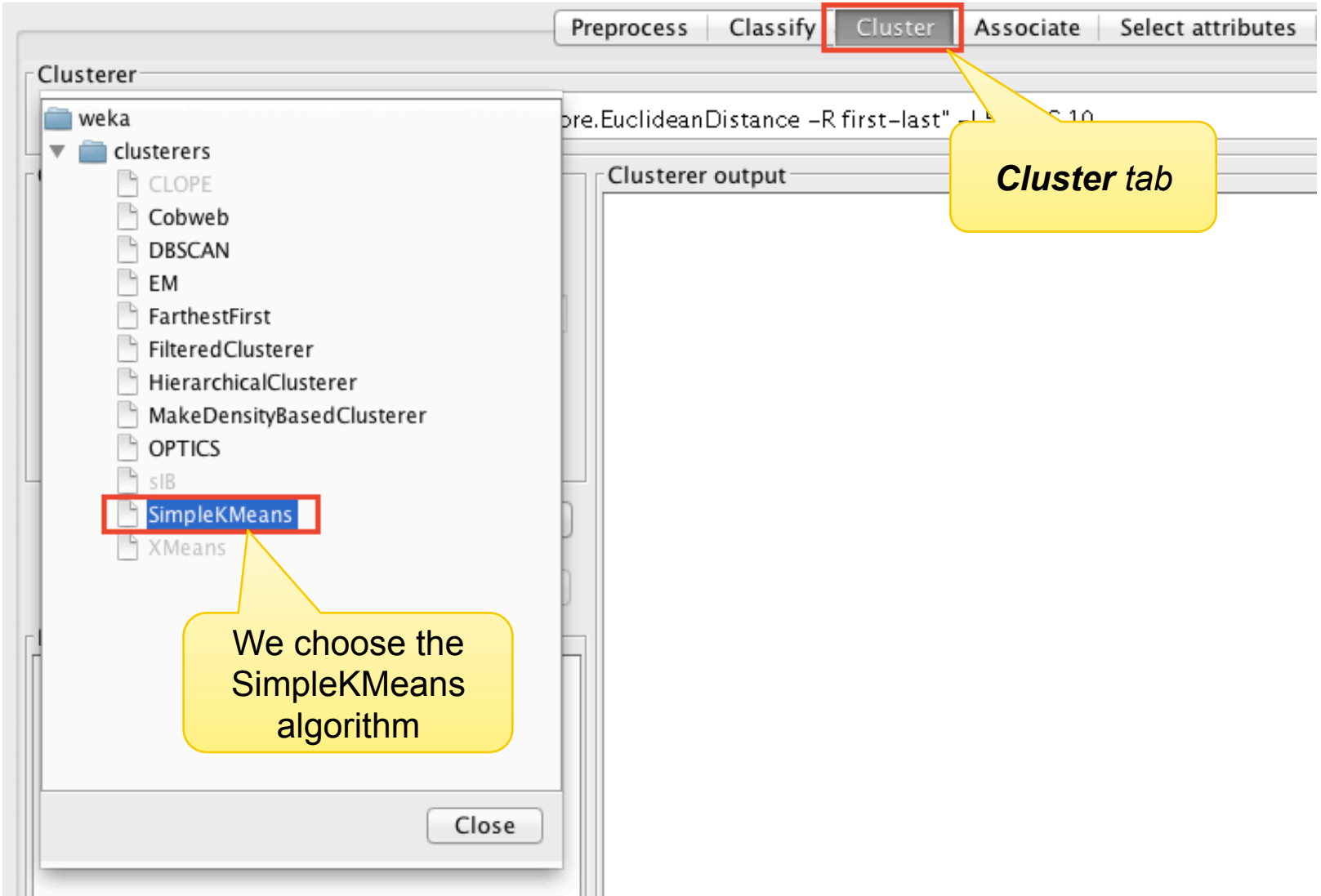
As the result of clustering each instance is being added a new attribute – the cluster to which it belongs. The clustering is said to be successful if the final clusters make sense, if they could be given meaningful names.

K-Means algorithm in Weka

FishersIrisDataset.arff



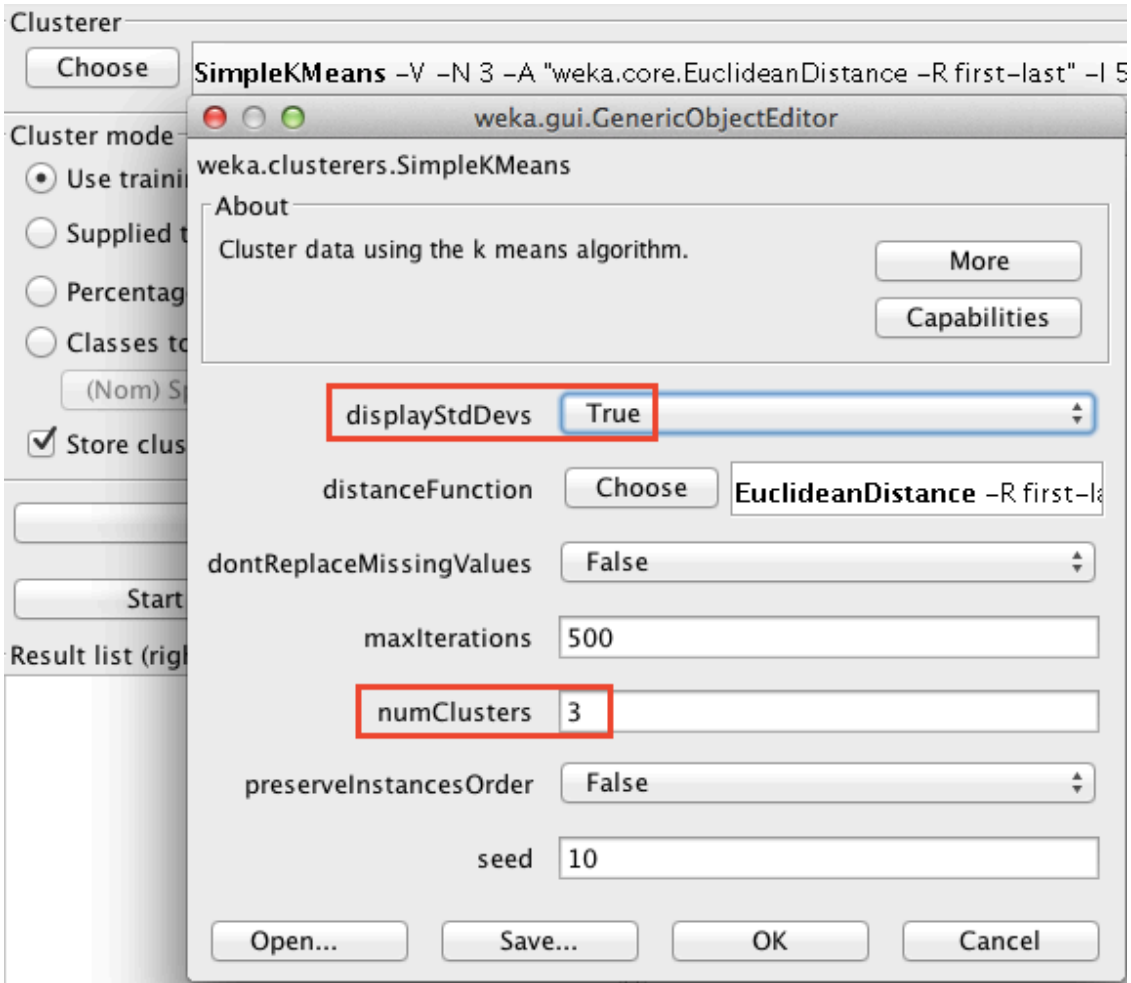
Choosing the clustering algorithm



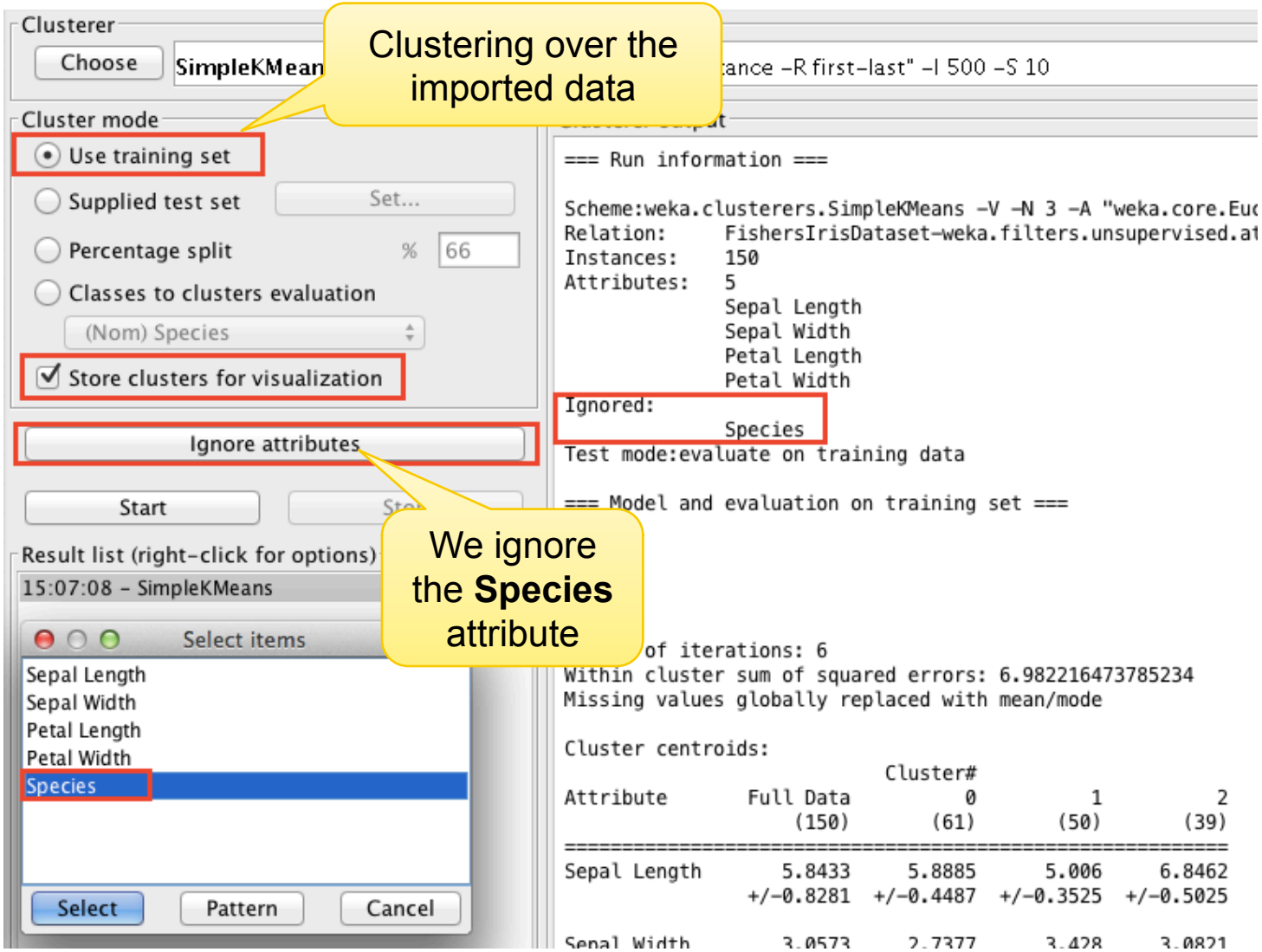
Parameter settings

numClusters – the number of desired clusters;
we set it to 3 because we have 3 kinds

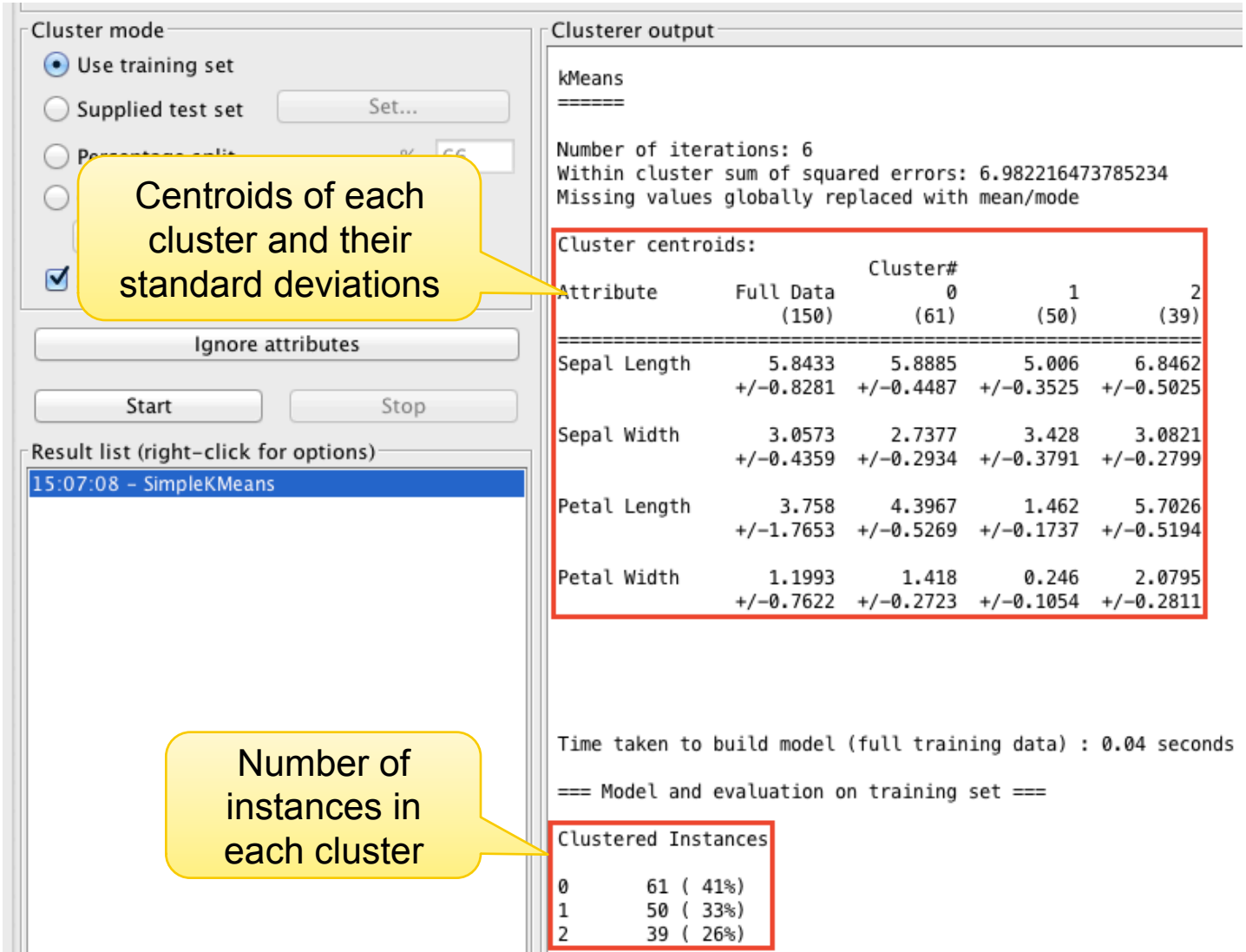
displayStdDevs – if true, the standard deviation will be displayed



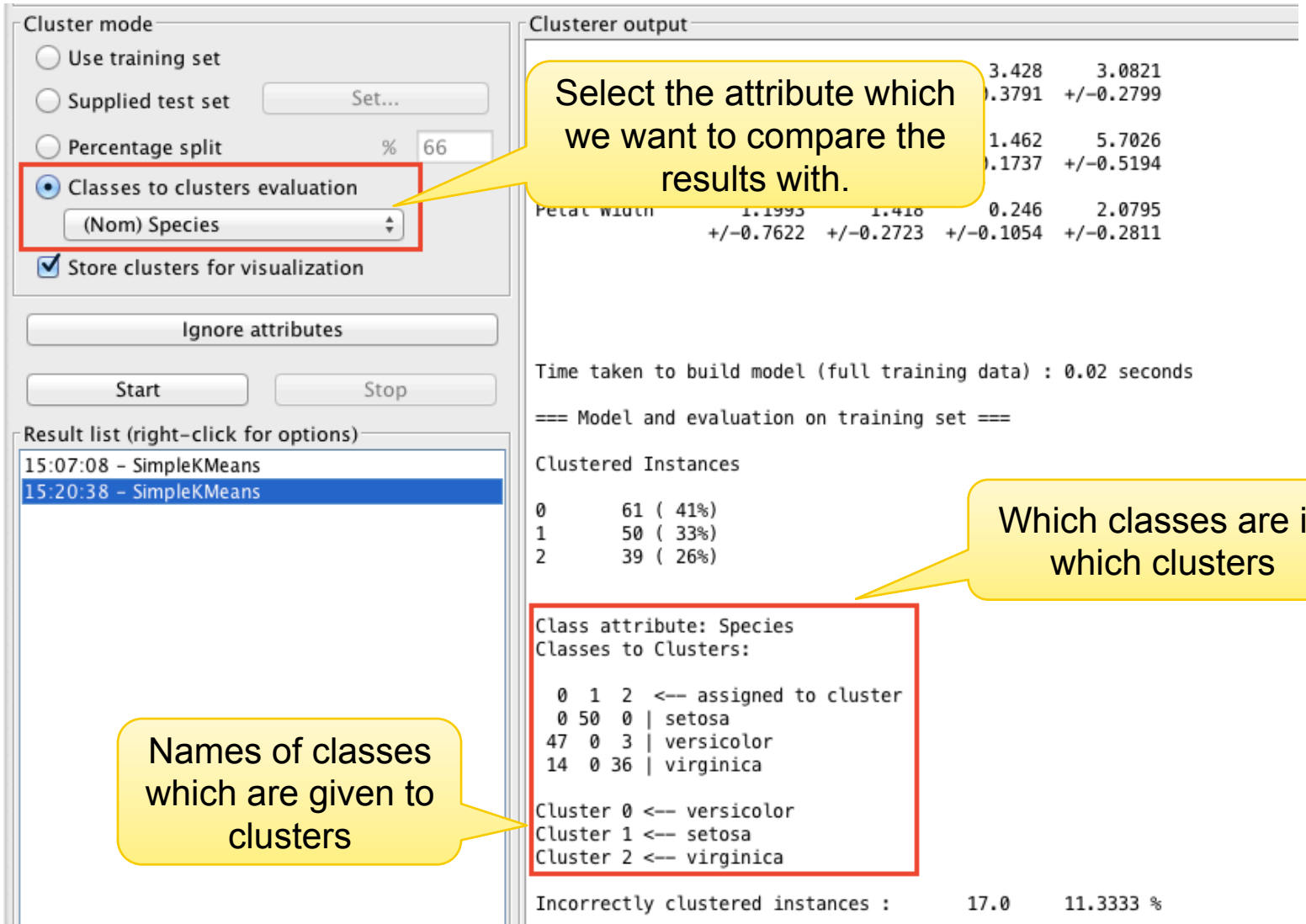
Running the Clustering



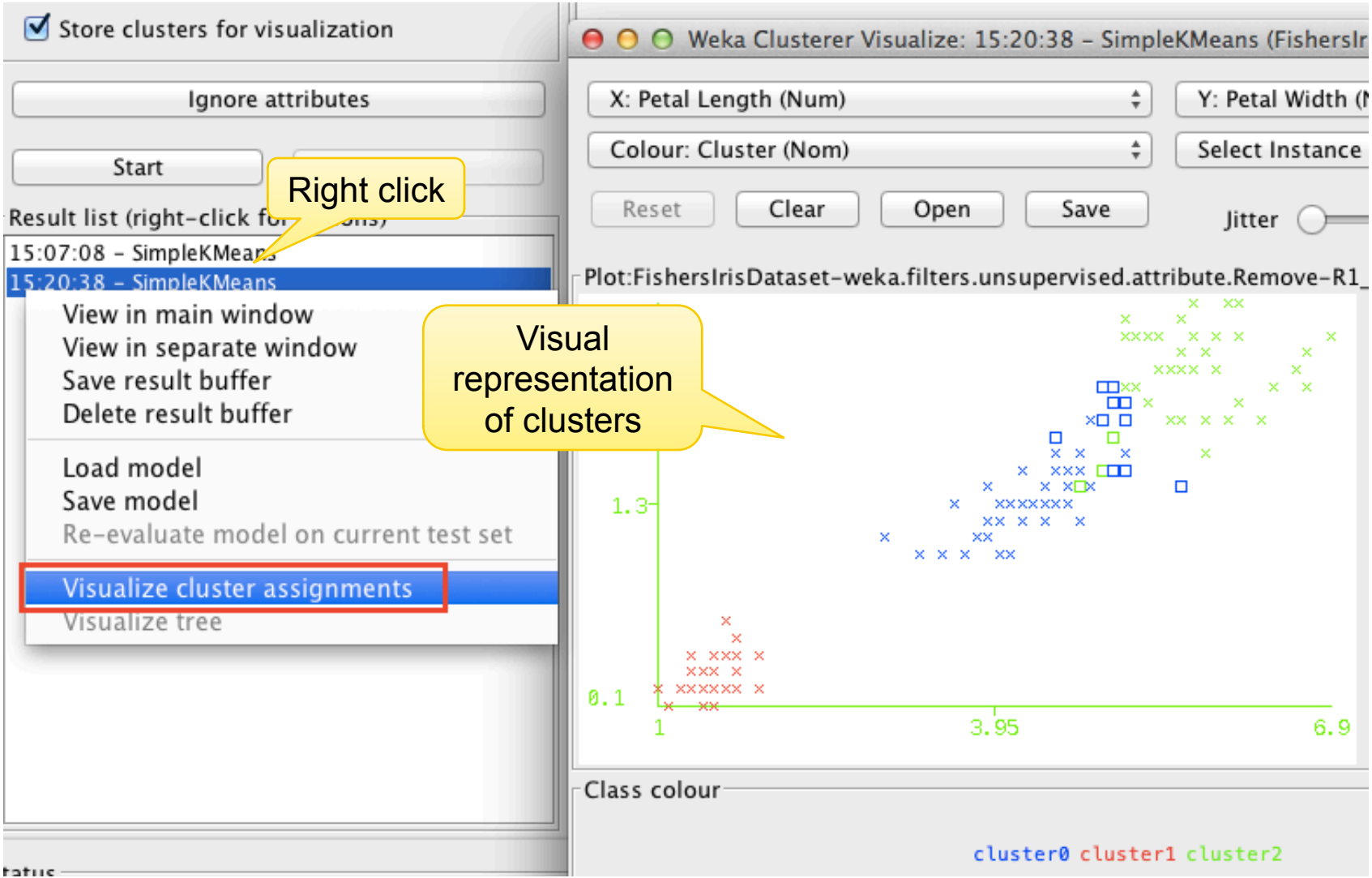
Results of Clustering



Evaluation of Results

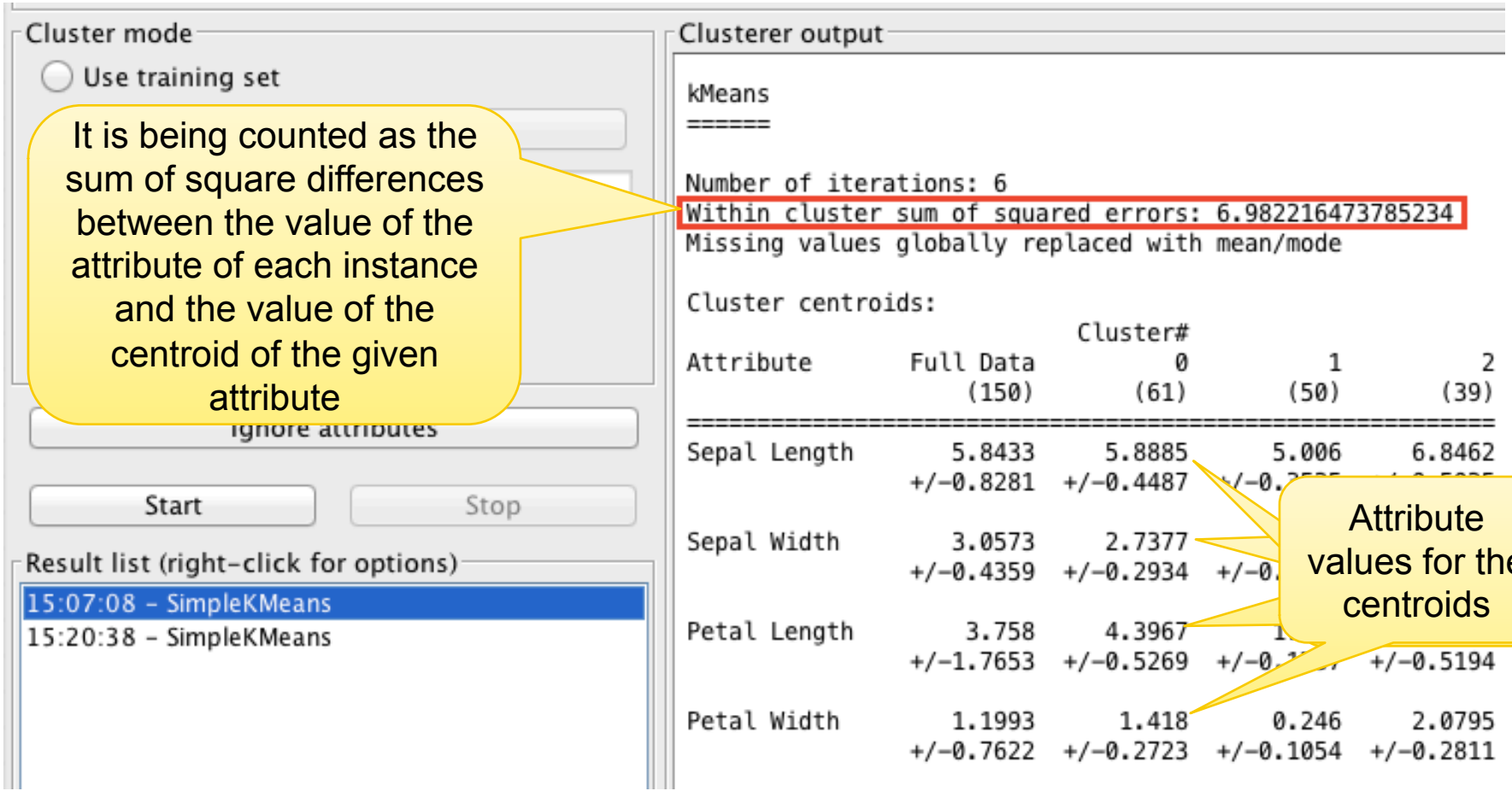


Visualization of Clusters

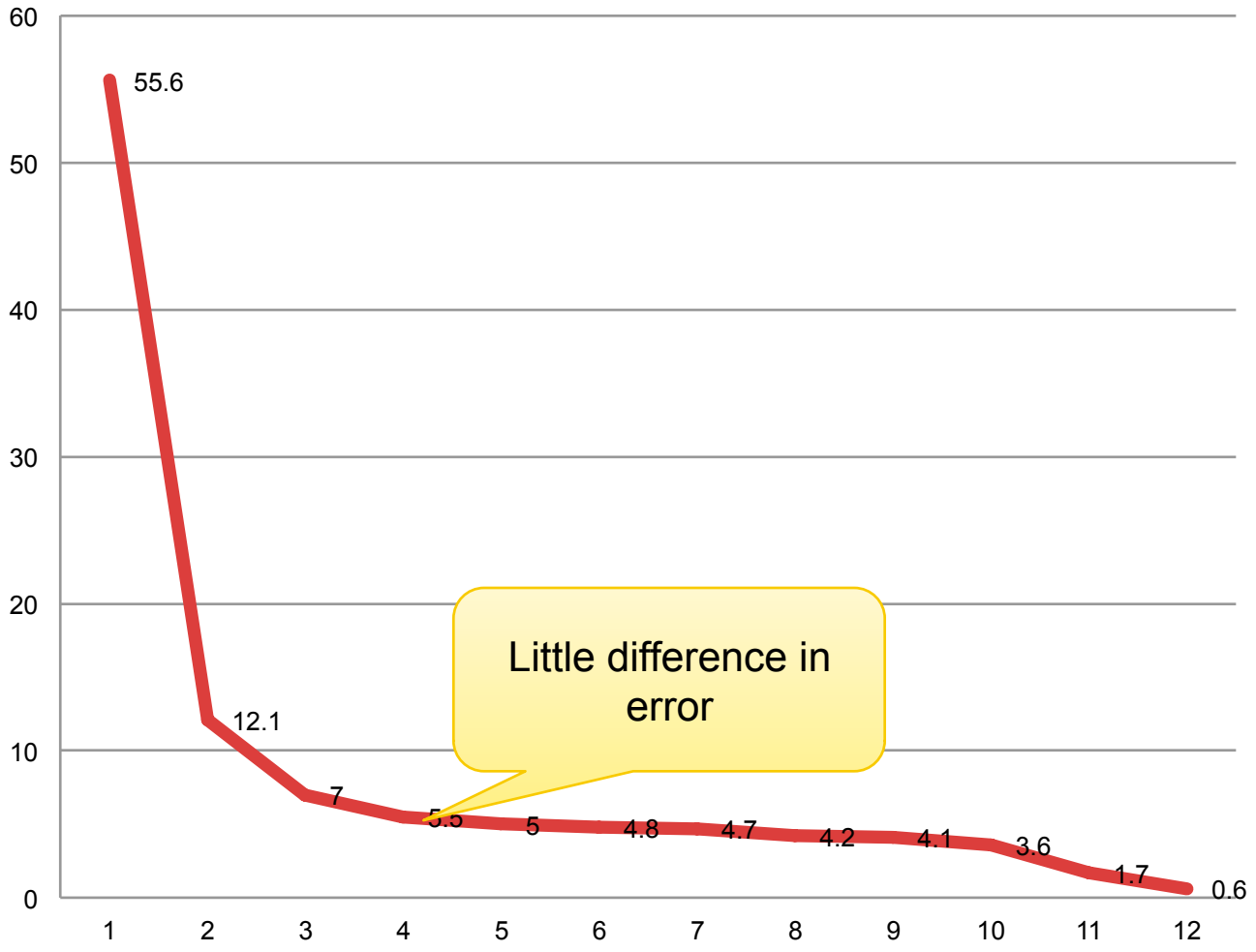


Was clustering successful?

Within cluster sum of squared error gives us the assessment of quality

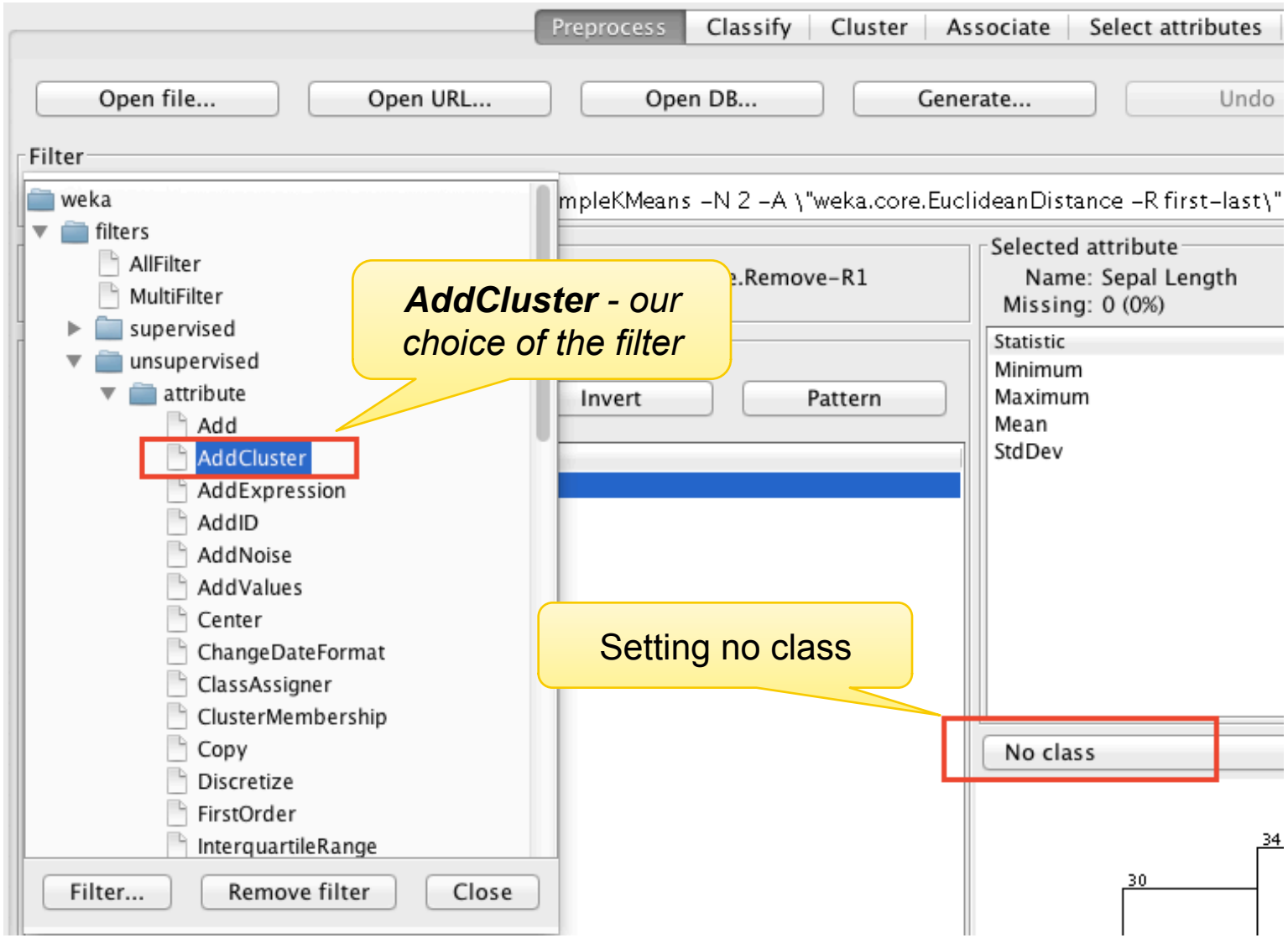


How to figure out the number of clusters?

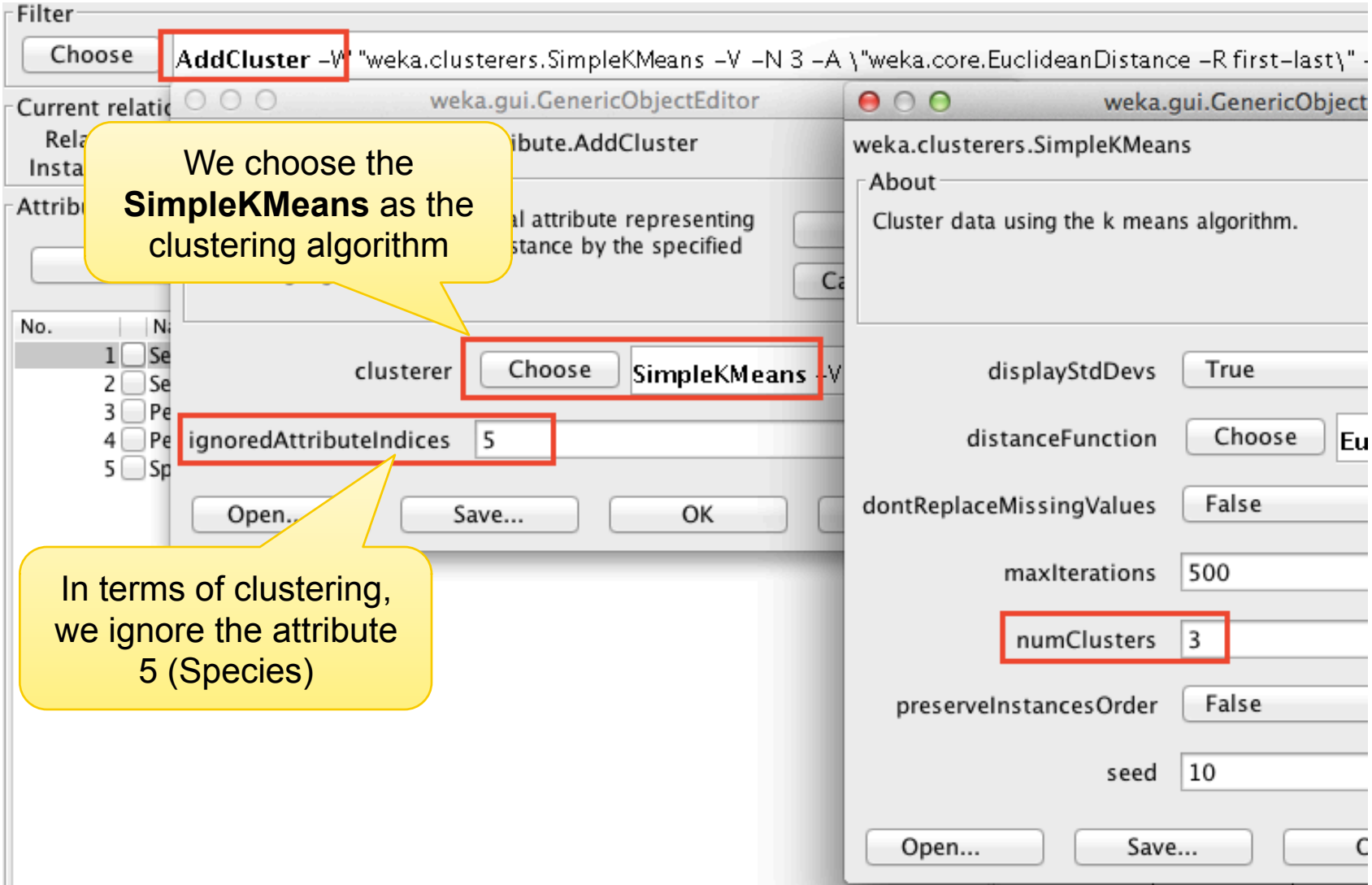


Clusters	Errors
1	55.6
2	12.1
3	7.0
4	5.5
5	5.0
6	4.8
7	4.7
8	4.2
9	4.1
10	3.6
20	1.7
50	0.6

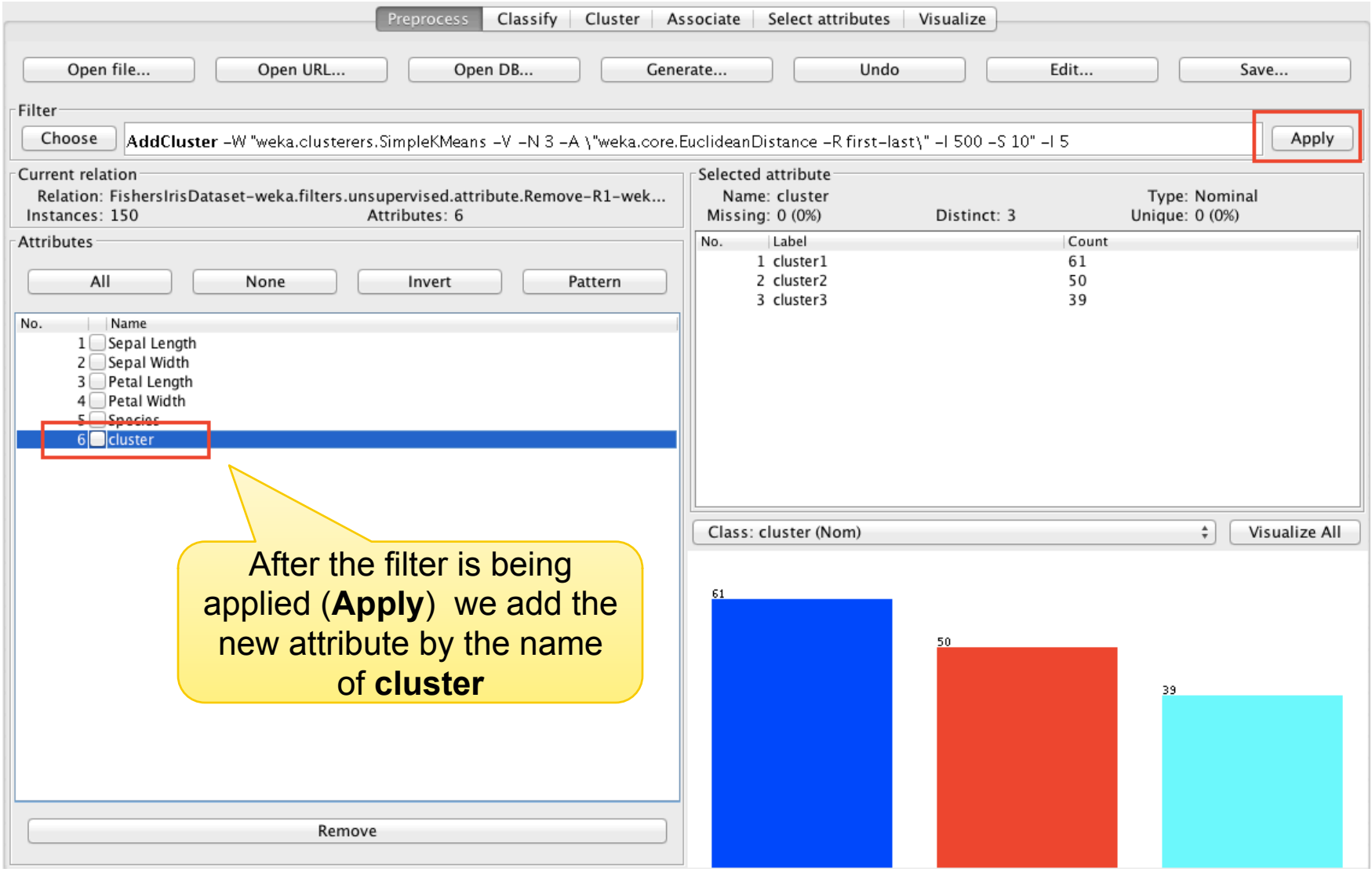
Using Clusters for Classification



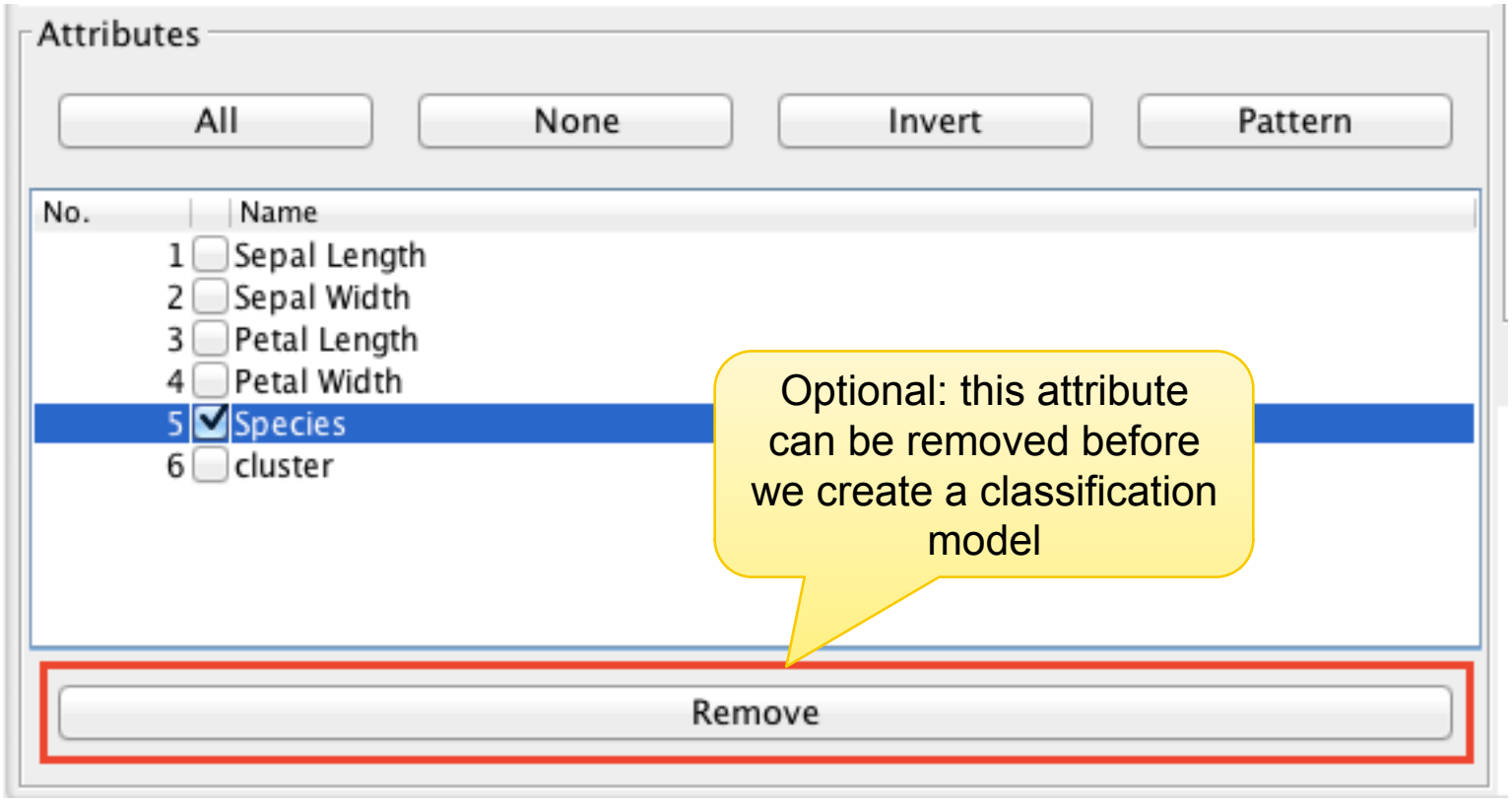
Using Clusters for Classification



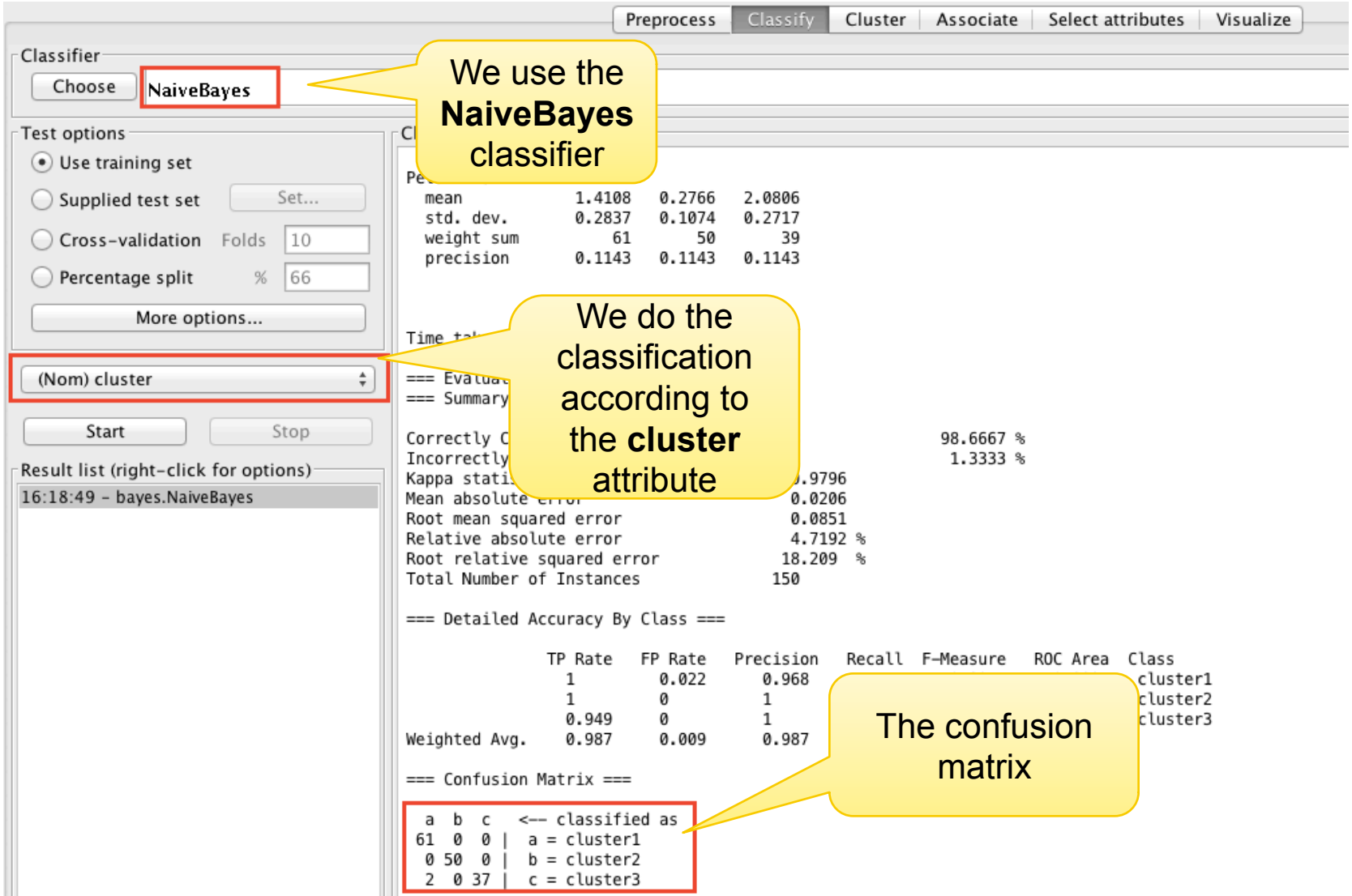
Using Clusters for Classification



Using Clusters for Classification



Using Clusters for Classification



Expectation Maximization (EM)

The EM algorithm consists of 2 key steps:

- **E (expectation) step** – calculation of the cluster probabilities; in this step we assume that we know the values of all the model parameters;
- **M (maximization) step** – calculation of the model parameters; we aim to "maximize" the likelihood of the model given the available data

These steps are repeated until the algorithm starts to converge

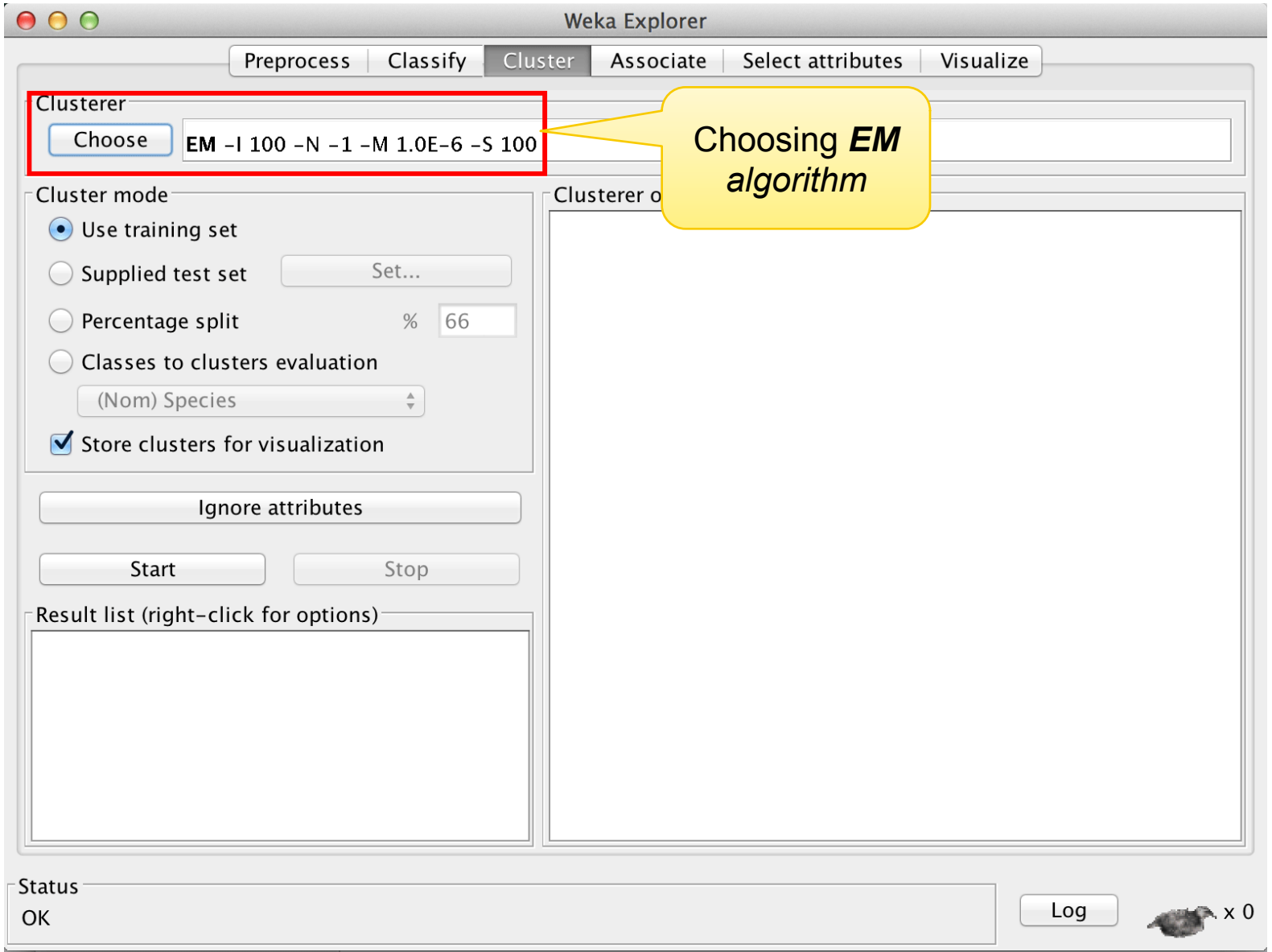
Expectation Maximization (EM)

To solve the described problem, we can apply a procedure similar to the one used for the K means algorithm:

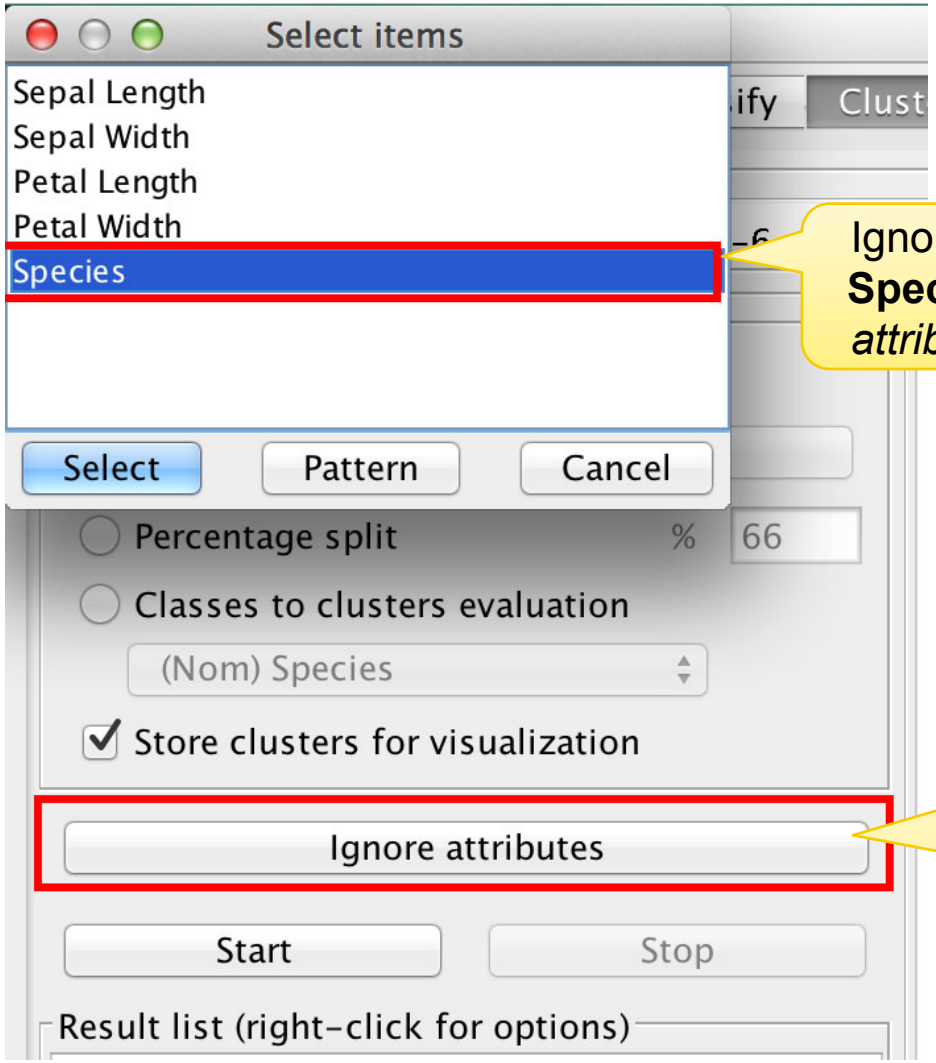
1. start by defining the number of clusters (k) and randomly choosing the model parameters (μ_i, σ_i, p_i, i =1,k)
2. for the given parameter values, compute, for each instance, the probability of belonging to each of the k clusters
3. use the computed probabilities to re-estimate the parameter values

Repeat steps 2) and 3) until the parameter values start to converge

Using EM in Weka



Ignoring Class Attribute



Recommendations and credits

Weka Tutorials and Assignments @ The Technology Forge

- http://www.technologyforge.net/WekaTutorials/

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- https://www.youtube.com/user/WekaMOOC/

(Anonymous) survey for your comments and suggestions:
http://goo.gl/cqdp3l

ANY QUESTIONS?

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: http://krcadinac.com