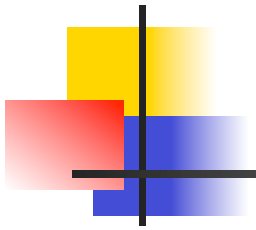


Prof. Pietro Ducange

Students' Tutor and Practical Classes

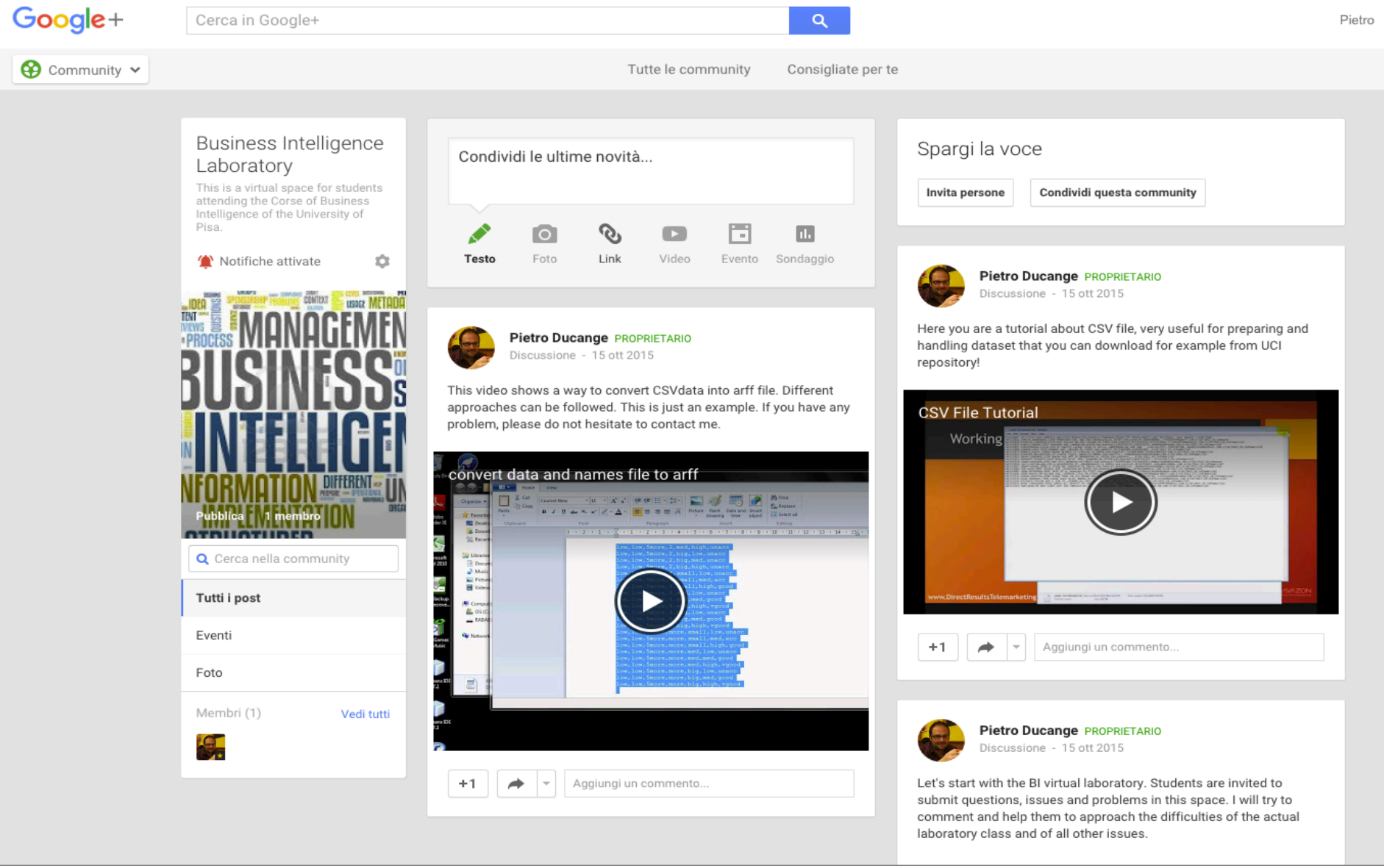
Course of Business Intelligence 2016

http://www.iet.unipi.it/p.ducange/esercitazioniBI/

Email: p.ducange@iet.unipi.it

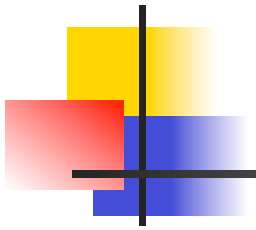
Office: Dipartimento di Ingegneria dell'Informazione (Room 221)

Tutoring Hours: send an email for reserving your time slot



A G+ community of the BI Laboratory is available for students at:

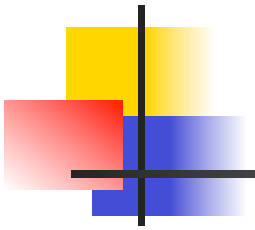
http://bit.ly/2dFxLUx



WEKA
Waikato Environment for Knowledge Analysis

An Introduction

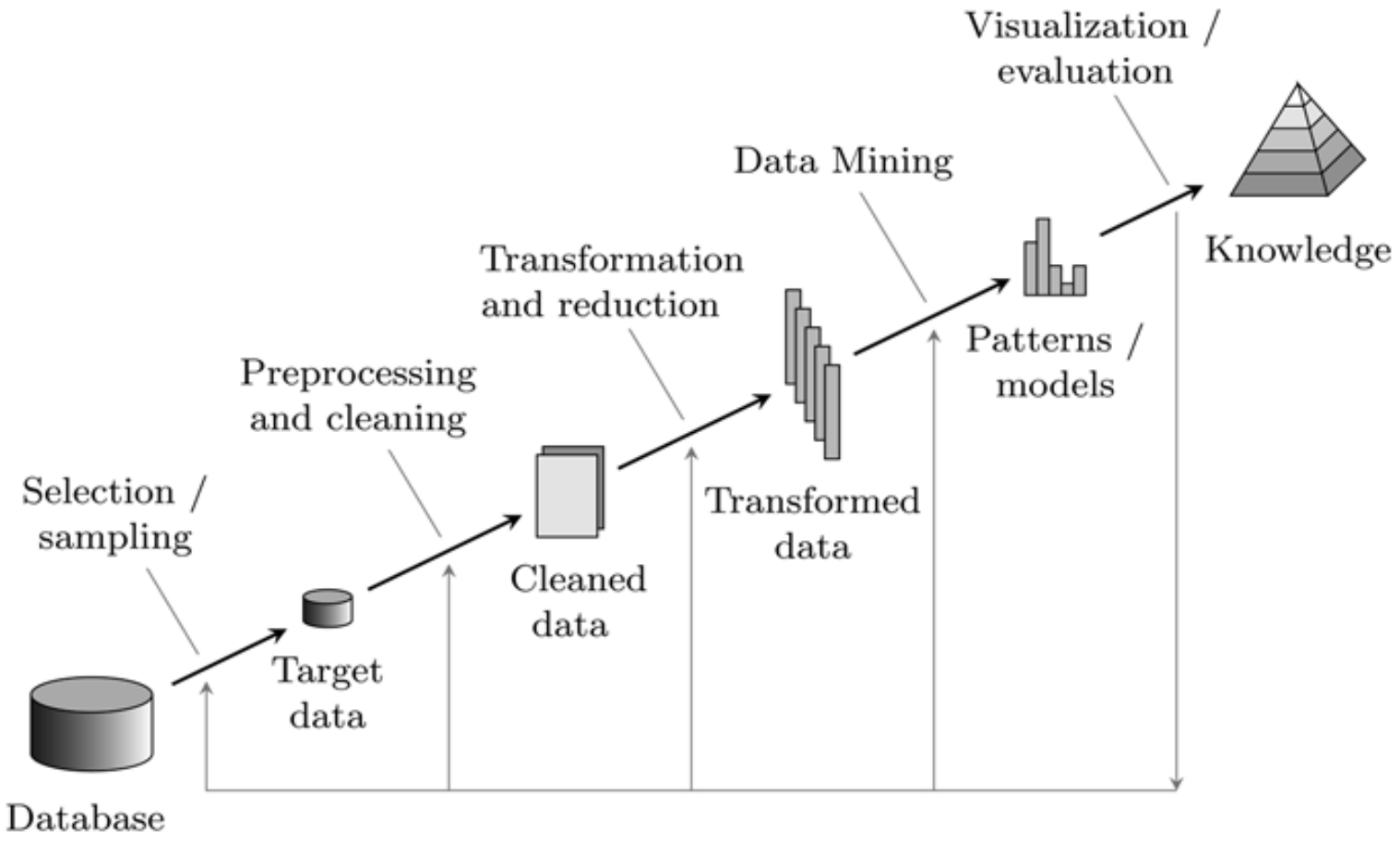
Introduction



- The WEKA workbench is a collection of state-of-the-art machine learning algorithms and data preprocessing tools.

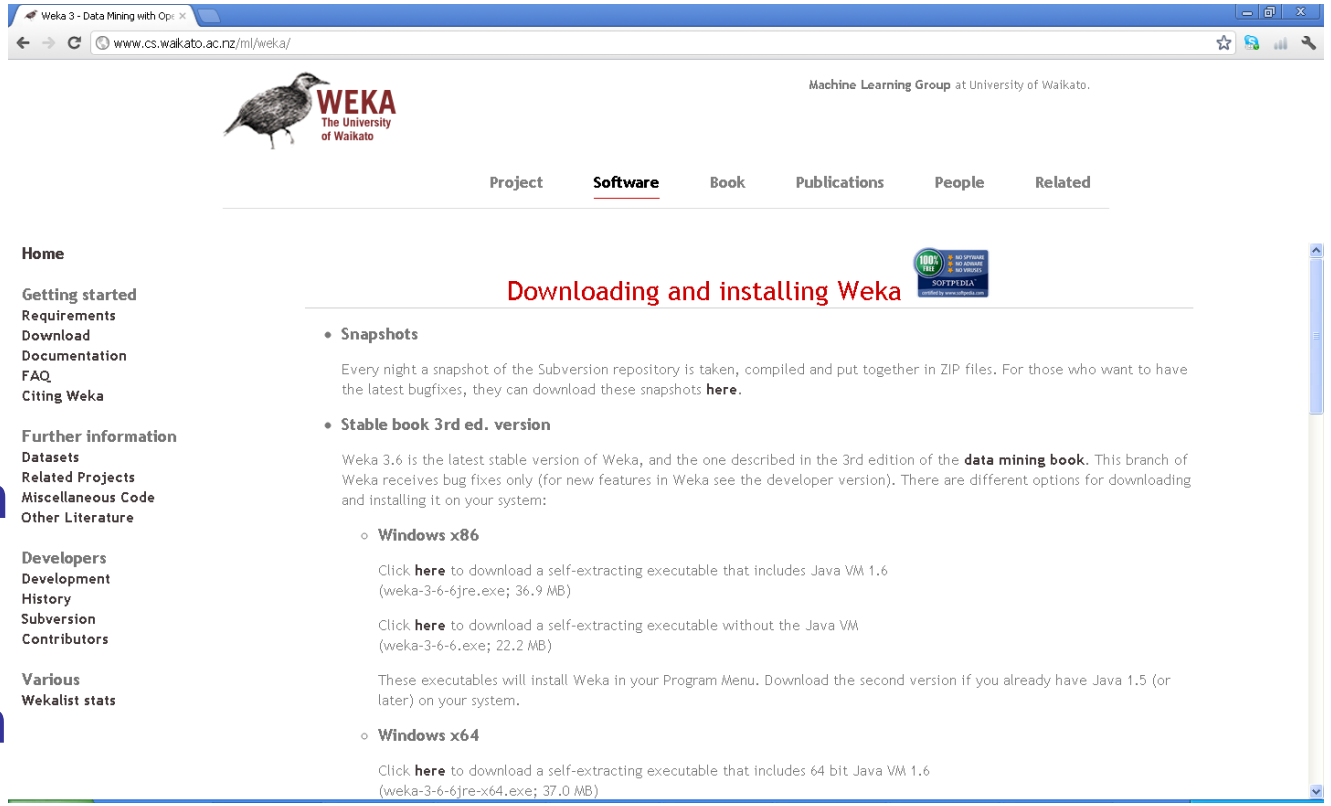
- It provides extensive support for the whole process of experimental data mining: preparing the input data, evaluate learning scheme, visualizing the input data and the results

The knowledge extraction process

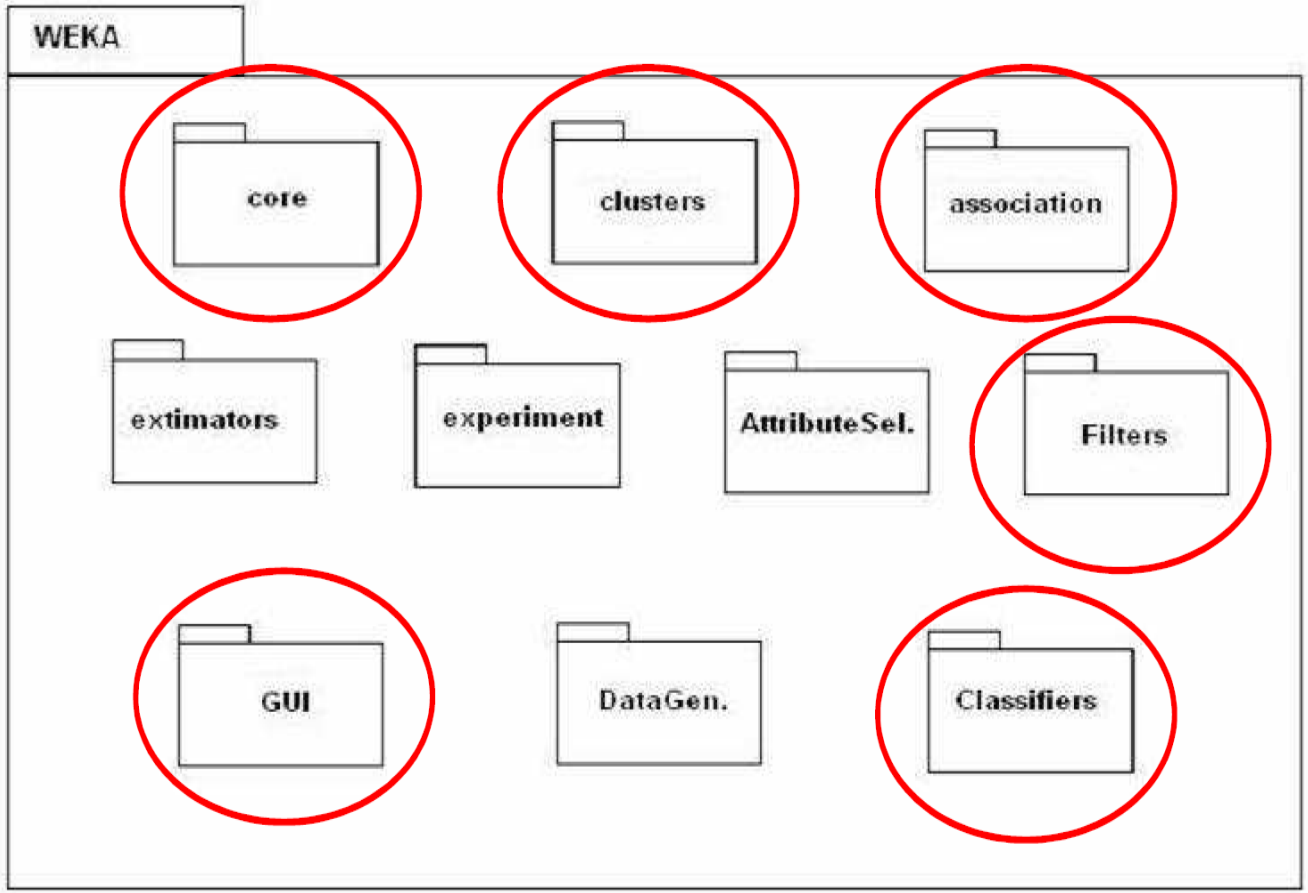


How do you get it?

- Weka is available from www.cs.waikato.ac.nz/ml/weka
- It is possible to download either a platform-specific installer or an executable Java jar file
- Weka Version 3-6-10



WEKA Architecture



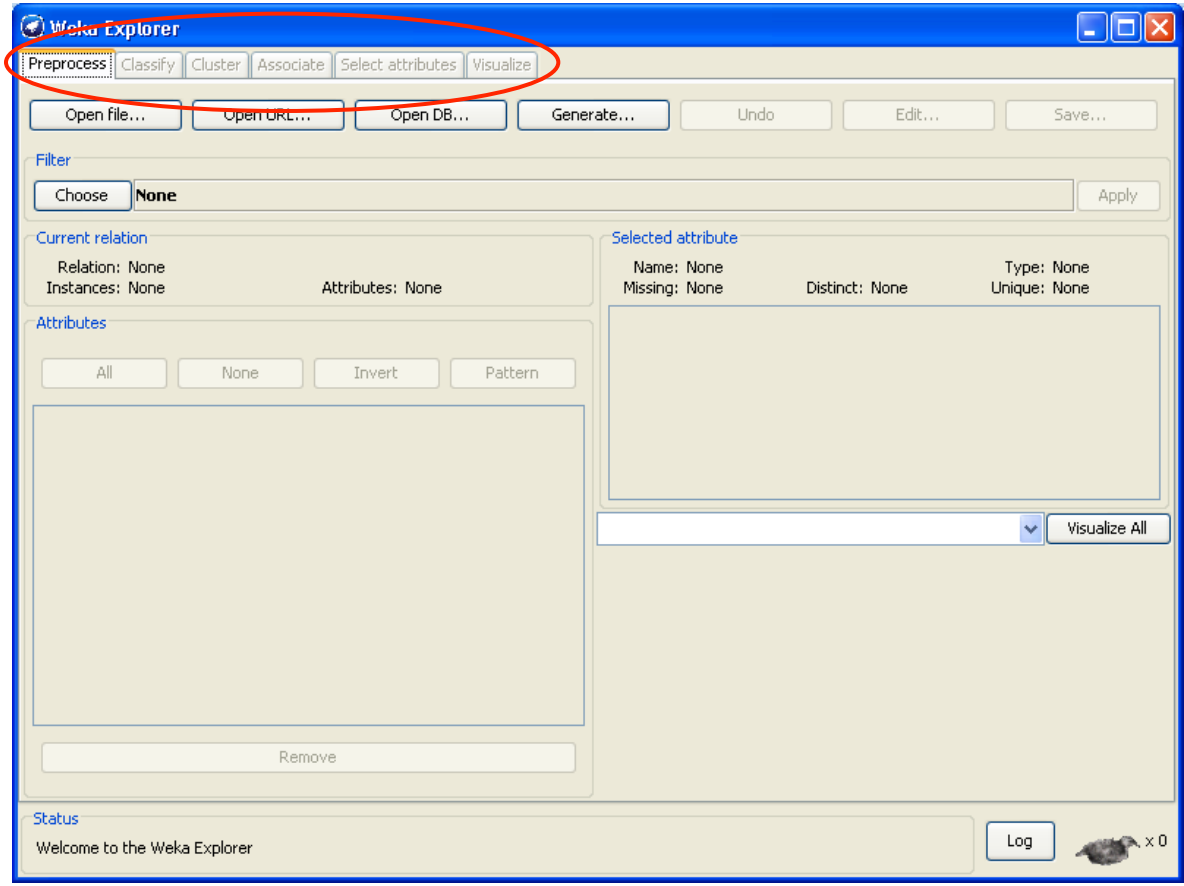
The Weka GUI chooser

- In addition to the command line interface (SimpleCLI), there are three graphical user interfaces: Explorer, Experimenter, KnowledgeFlow

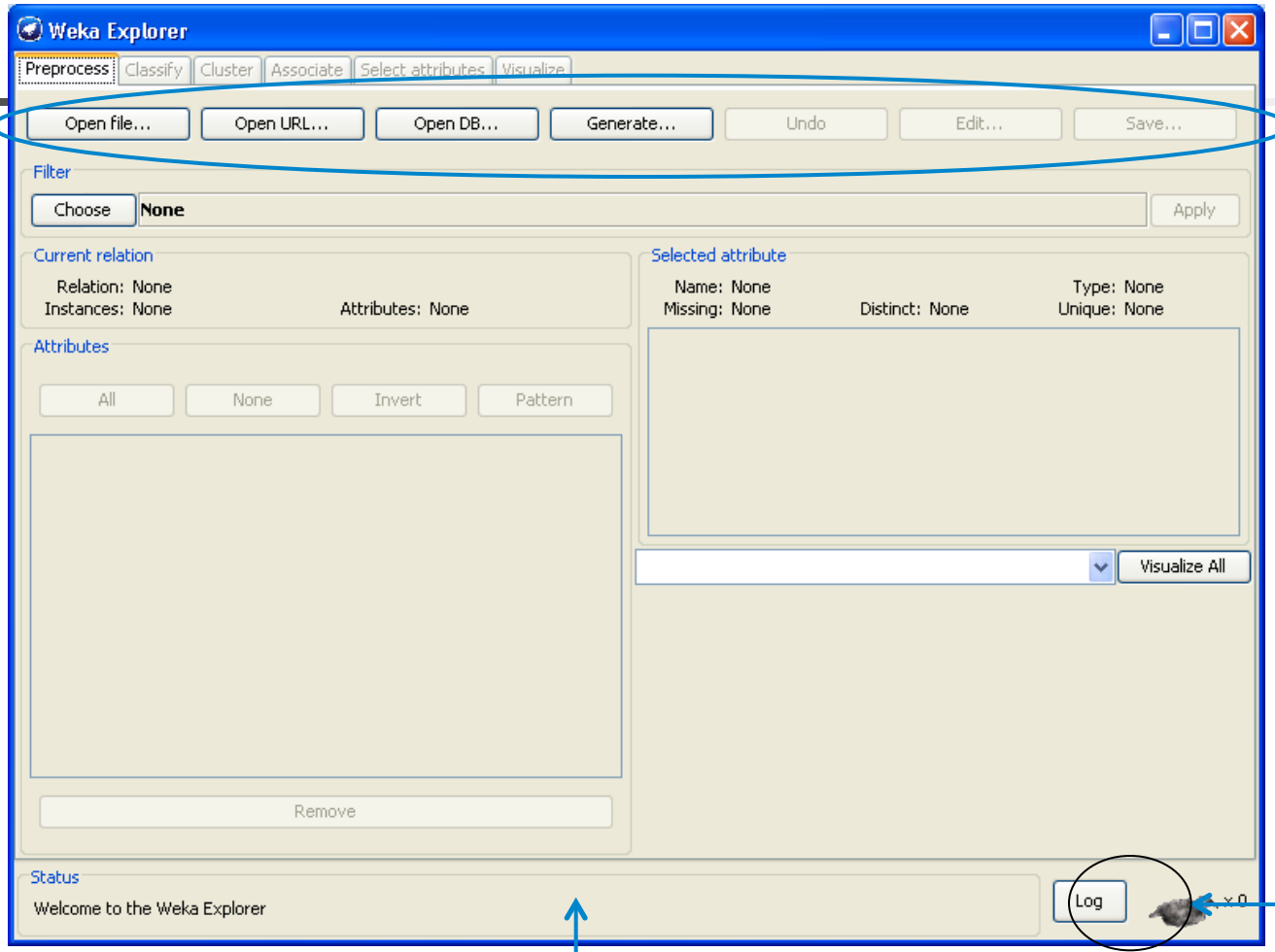
- The menu consists of four sections:
  - Program
    - LogWindow: Opens a log window
    - Exit: Closes WEKA.
  - Visualization
    - Plot: for plotting a 2D plot of a dataset.
    - ROC: displays a previously saved ROC curve.
    - TreeVisualizer: for displaying directed graphs, e.g., a decision tree.
    - GraphVisualizer: Visualizes XML BIF or DOT format graphs, e.g.,for Bayesian networks.
    - BoundaryVisualizer: for visualizing classifier decision boundaries in two dimensions.
  - Tools
    - ArffViewer An MDI application for viewing ARFF files in spread-sheet format.
    - SqlViewer Represents an SQL worksheet, for querying databases via JDBC.
    - Bayes net editor An application for editing, visualizing and learning Bayes nets.
  - Help Online resources

Explorer (1)

- Preprocess: Choose and modify the data being acted on.
- Classify: Train and test learning schemes that classify or perform regression.
- Cluster: Learn clusters for the data.
- Associate: Learn association rules for the data.
- Select attributes. Select the most relevant attributes in the data.
- Visualize: View an interactive 2D plot of the data.



Explorer (2)



Load data

Status Box
→ Memory information
→ Run garbage collector

Status Icon
→ Memory information
→ Run garbage collector

Loading data

- **Open File:** Brings up a dialog box allowing you to browse for the data file on the local file system. You can read files in a variety of formats: WEKA's ARFF format, CSV format, etc.
- **Open URL**: Asks for a Uniform Resource Locator address for where the data is stored.
- **Open DB:** Reads data from a database.
- **Generate:** Enables you to generate artificial data from a variety of DataGenerators.

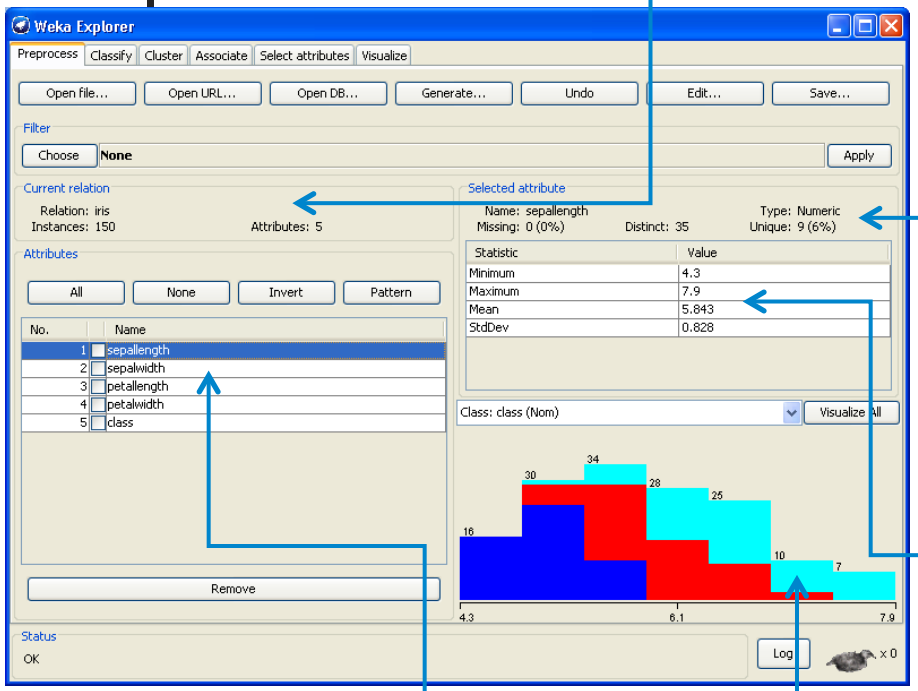
The .ARFF format

```
%ARFF file for the weather data
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
%
% 14 instances
%
@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

- ARFF files have two distinct sections: Header and Data.
- The **Header** section contains the name of the relation, a list of the attributes (the columns in the data), and their types.
- The **Data** section starts with the @data declarations and contains all the instances of the dataset .
- The ARFF format has 5 types of attributes:
  - Numeric
  - Nominal (represented by the set of values they can take on, enclosed in curly braces)
  - String
  - Date
  - Relational-valued (the value is a separate set of instances , the attribute is defined with a name and the type relational, followed by a nested attribute block. The key **@end** *AttributeName* is used to end the nested block of attributes)

Information on the data



Name of the relation, number of instances, number of attributes.

Information on the selected attribute:
→ Name and type of attribute
→ Number of instances in the data for which this attribute is missing
→ Number of different values that the data contains for this attribute
→ Number of instances in the data having a value for this attribute that no other instances have.

→ Nominal attributes: the list consists of each possible value for the attribute along with the number of instances that have that value.
→ Numeric attributes: the list gives four statistics describing the distribution of values in the data

List of attributes with three columns
→ A number that identifies the attribute
→ Selection tick boxes, that allow you to select attributes
→ Name of the attribute

Coloured histogram, colour-coded according to the attribute chosen as the Class using the box above the histogram

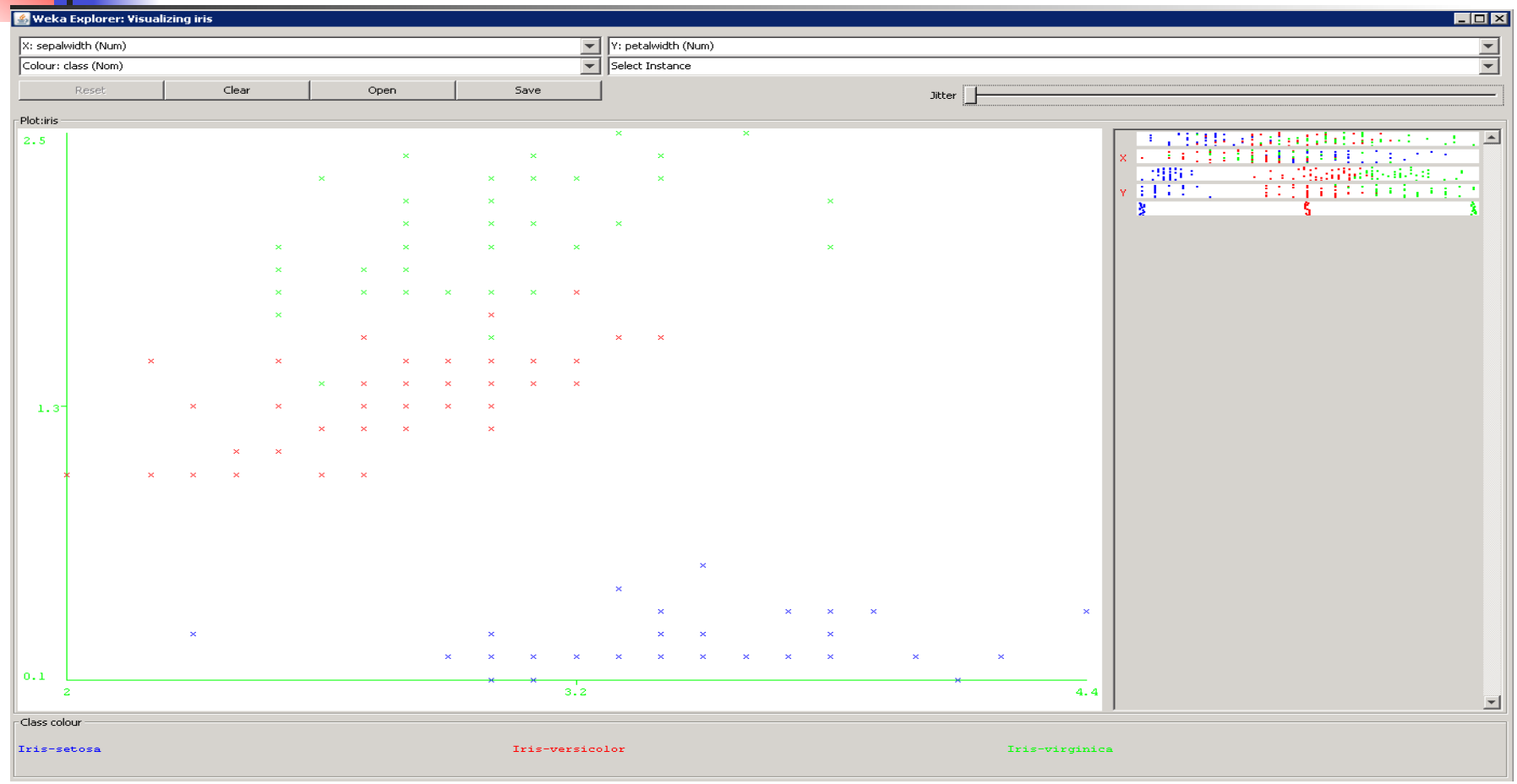
Converters to ARFF (i)

- ARFF files are not the only format one can load, but all files that can be converted with Weka's "core converters". The following formats are currently supported:
- C4.5
- CSV
- libsvm
- binary serialized instances
- XRFF
- text files in folders

If Weka cannot load the data, it tries to interpret it as a ARFF. If that fails, it pops up a box from which the user can select the converter.

Visualizing Data (Scatter Plot)

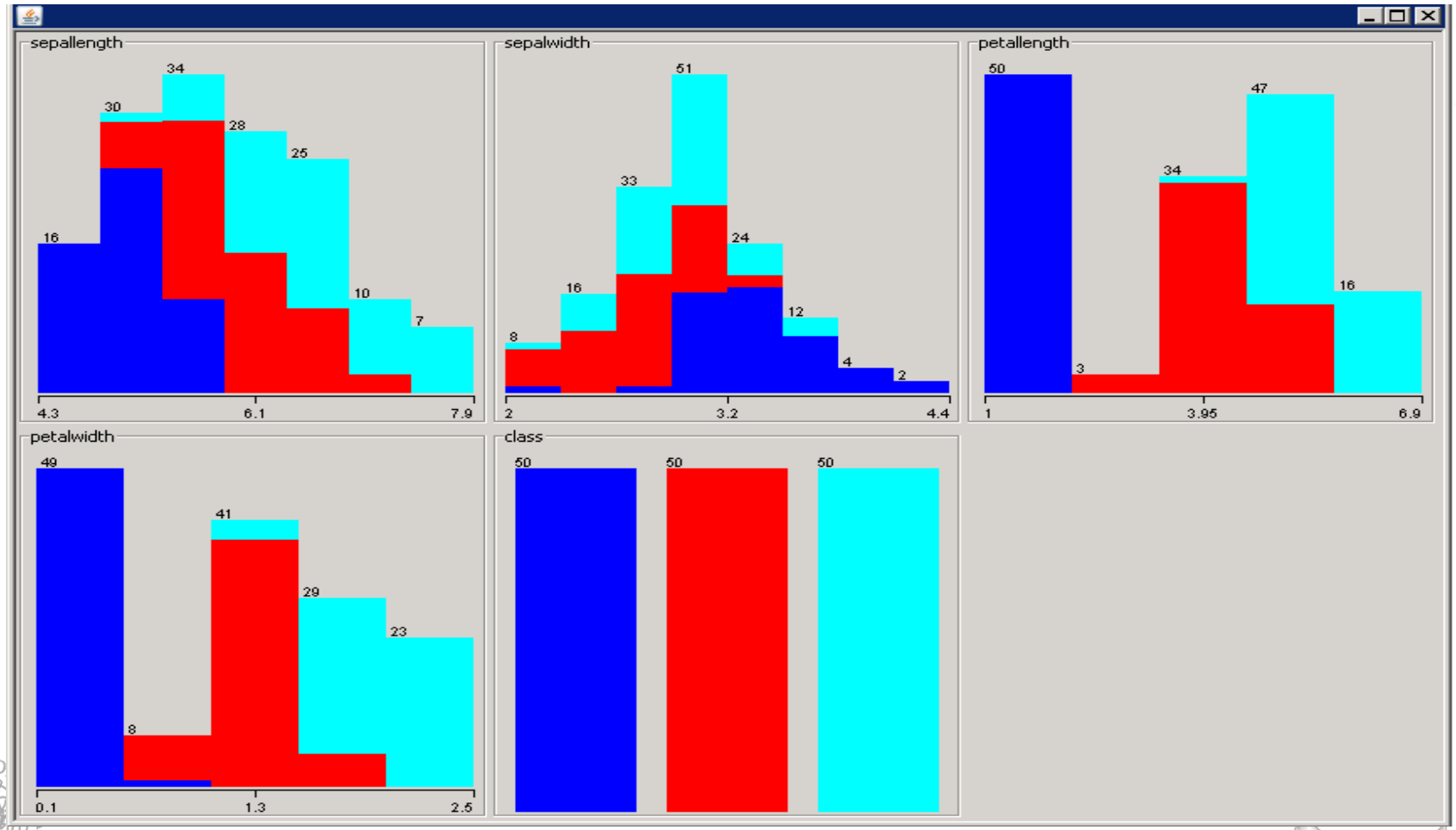
Select the **Visualize Section**



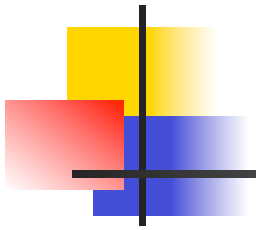
The jitter function just adds artificial random noise to the coordinates of the plotted points in order to spread the data out a bit (so that you can see points that might have been obscured by others.

Visualizing Data (Histograms)

Click **Visualize All**



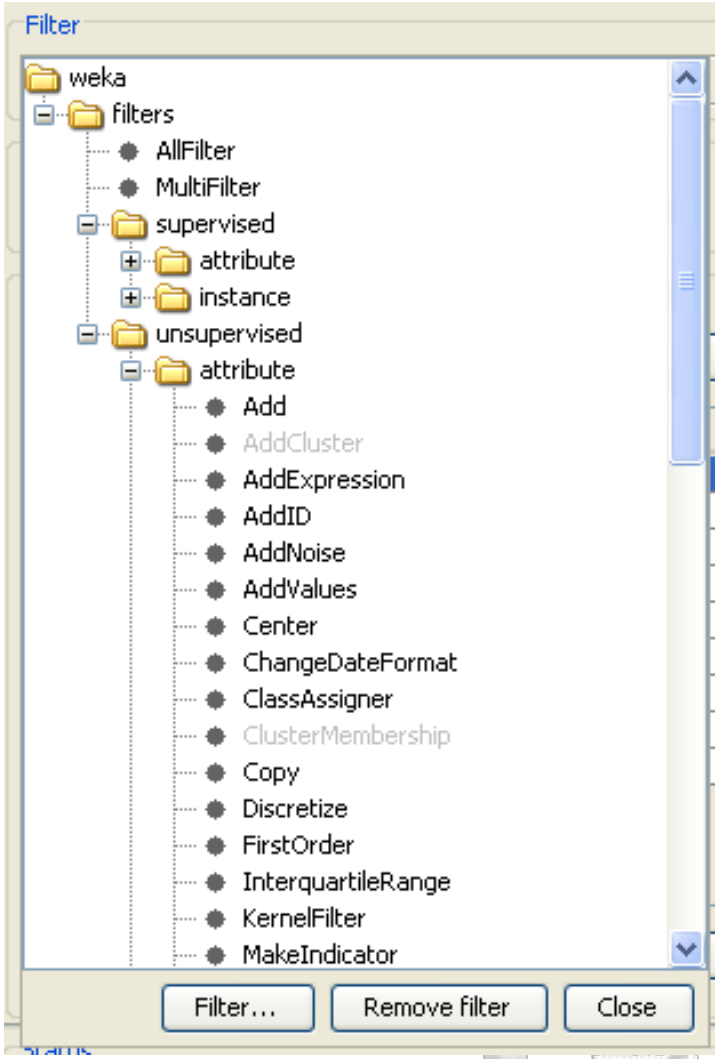
Working with filters (1)



- The pre-process section allows filters to be defined, they transform the input dataset in some way.
- The Filter box is used to set up the filters that are required.
- At the left of the Filter box there is a Choose button. By clicking this button it is possible to select one of the filters in WEKA.
- Once a filter has been selected, its name and options are shown in the field next to the Choose button.

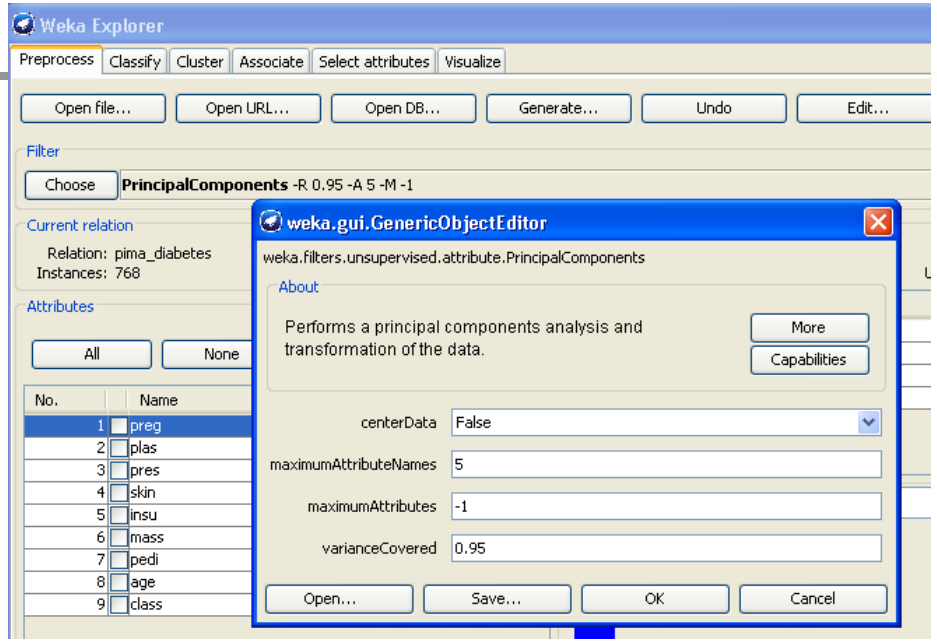
Working with filters (2)

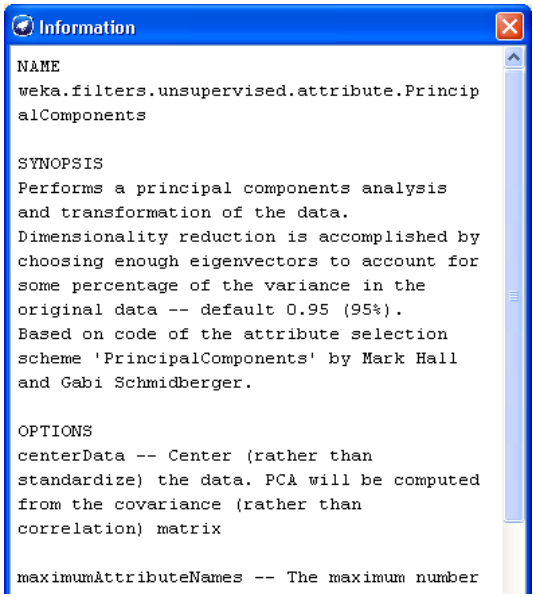
- There are two kind of filters, supervised and unsupervised. Within each type of filtering there is a further distinction between attributes filters, which work on the attributes of the datasets, and instance filters, which work on the instances.



Working with filters (3)

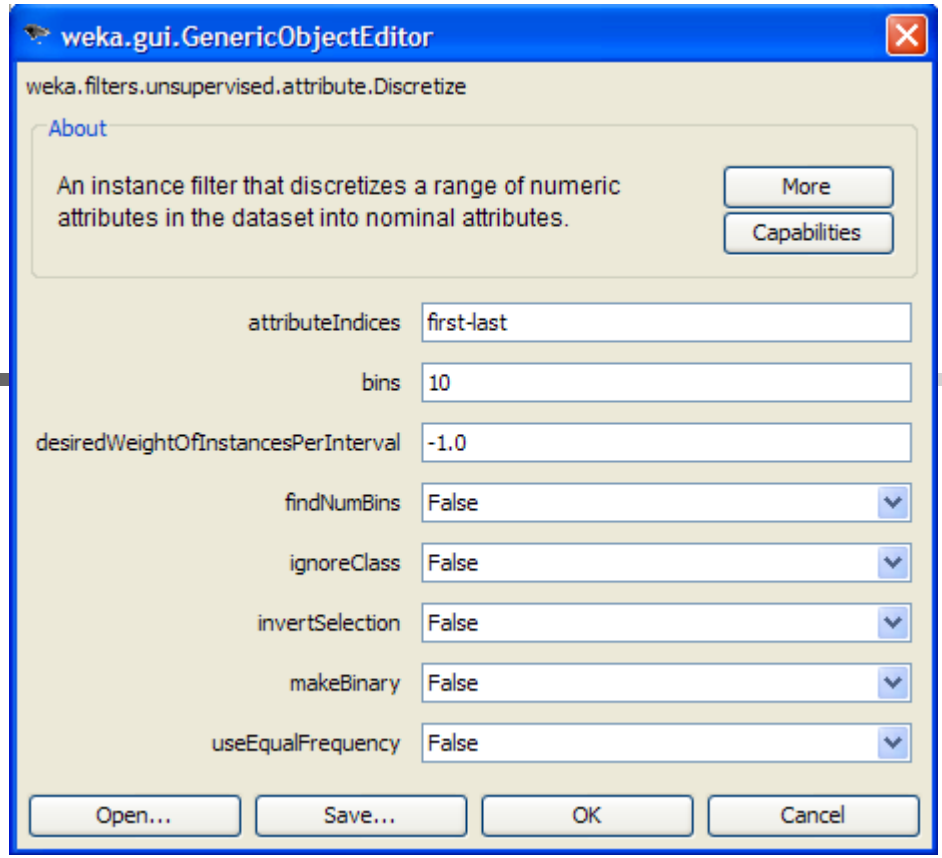
- When a filter is selected, its name and its parameters appears in the line beside the Choose button
- Click that line to get a generic object editor to specify the filter properties.



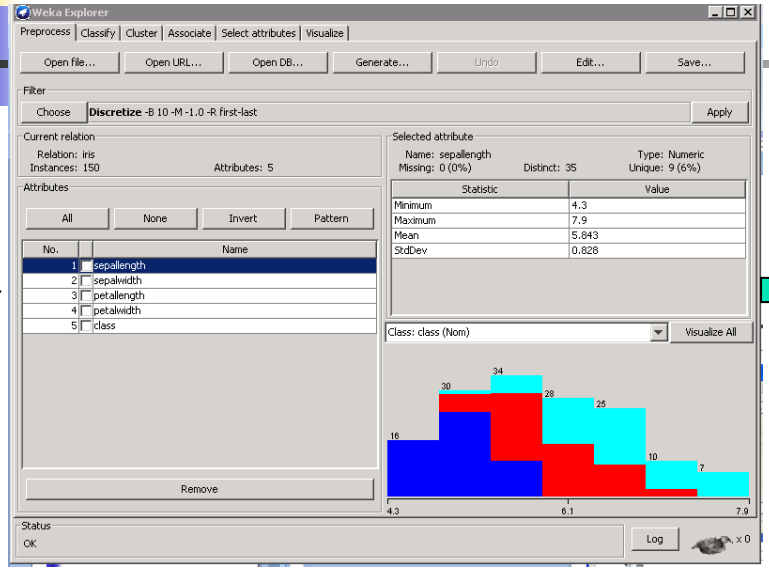


- Click the More button to get Information about the filter: a brief summary of the filter behavior and the description of its parameters

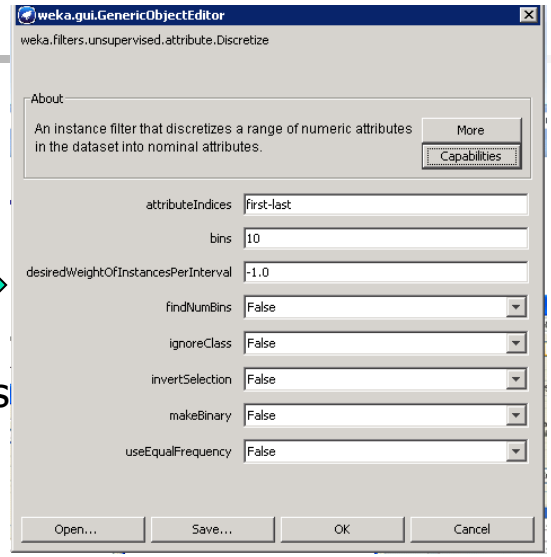
Unsupervised Filters: Discretize



Unsupervised Filters: Discretize

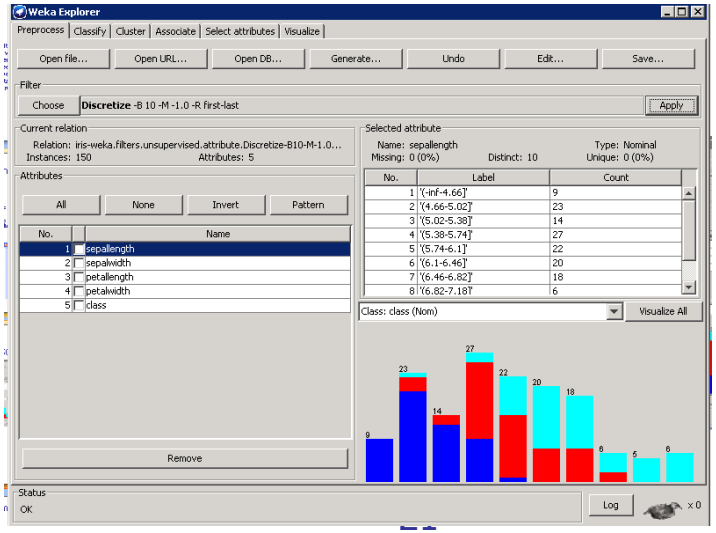


Load Data



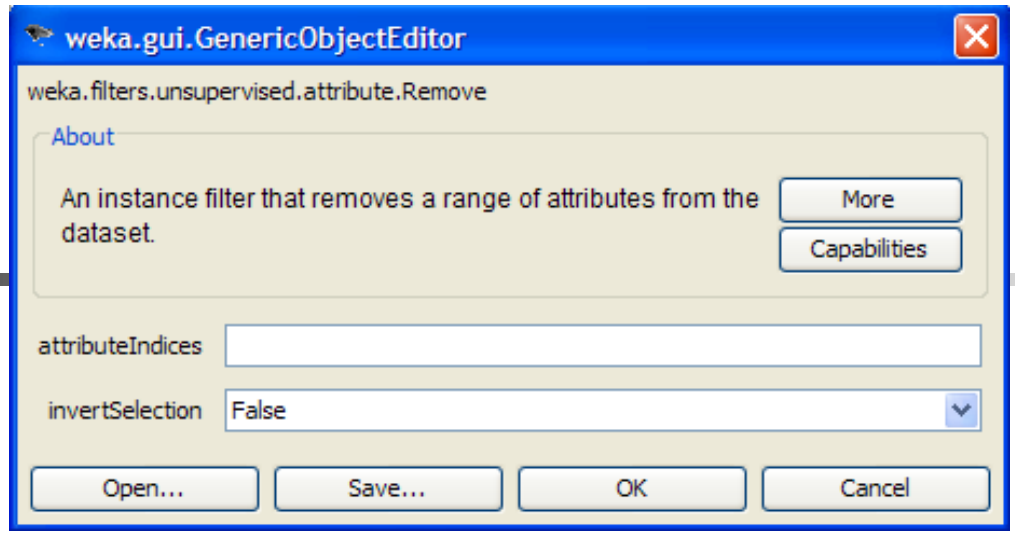
Set Parameters

Apply Algorithm

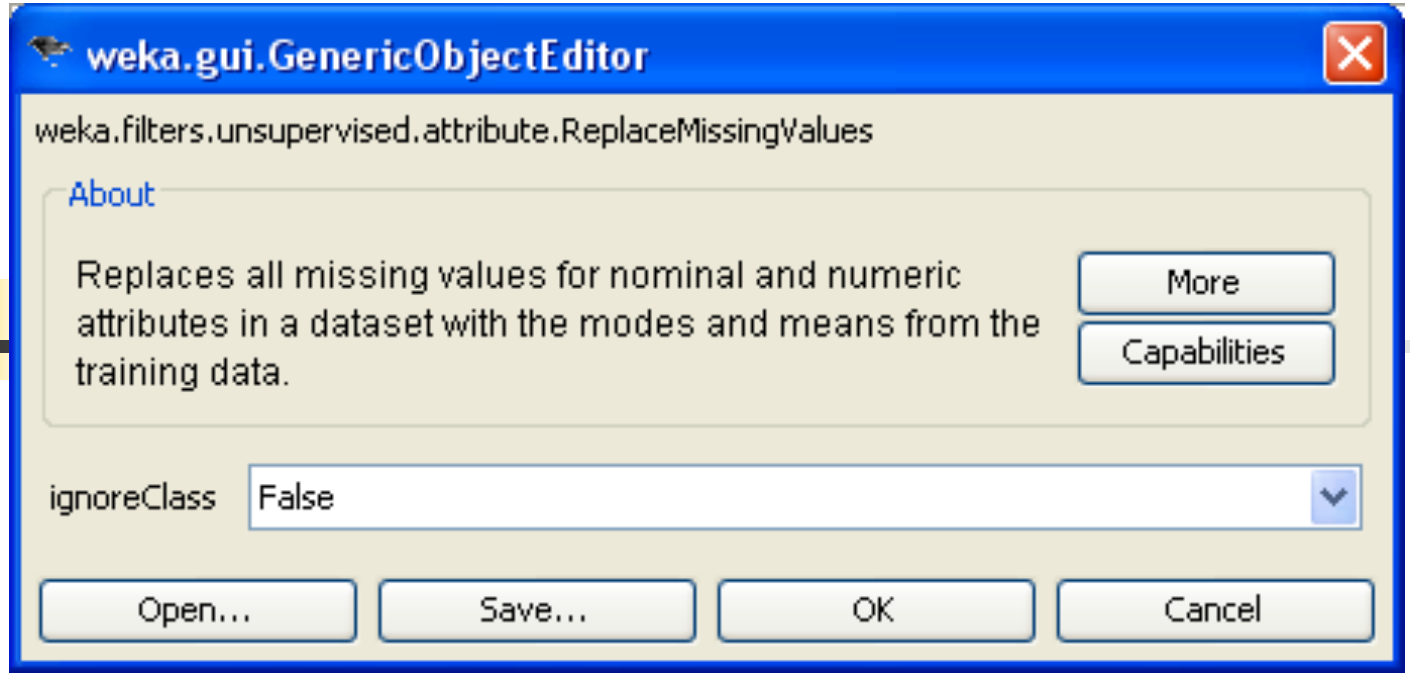


Save or Use Data

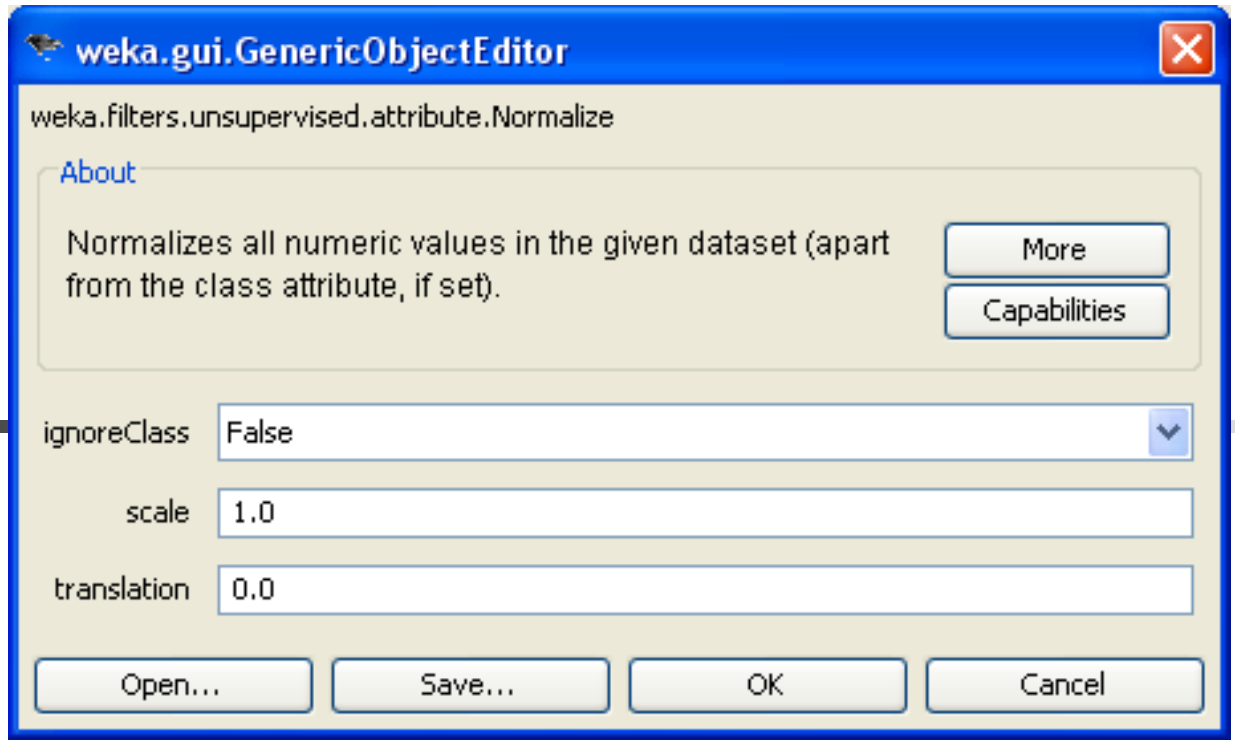
Unsupervised Filters: Attribute Remove



Unsupervised Filters: Replace Missing Values

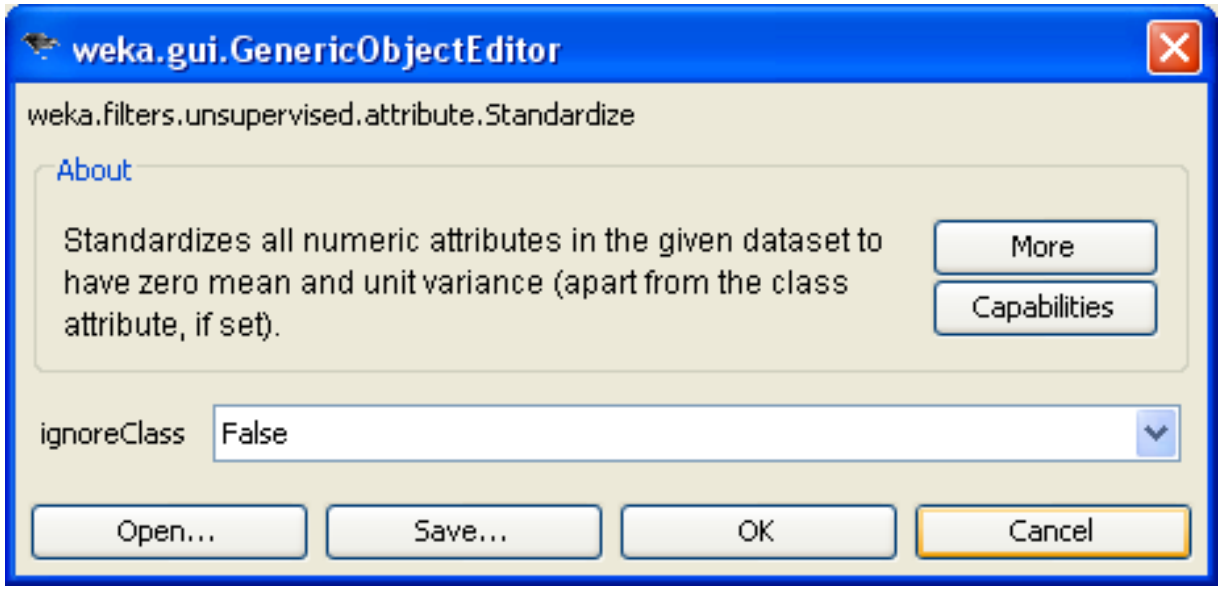


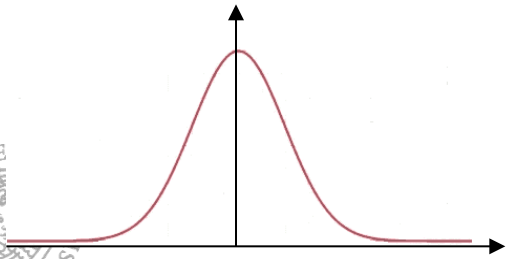
Unsupervised Filters: Normalize



$$v' = \frac{v - \min A}{MAXA - \min A} * scale + translation$$

Unsupervised Filters: Standardize



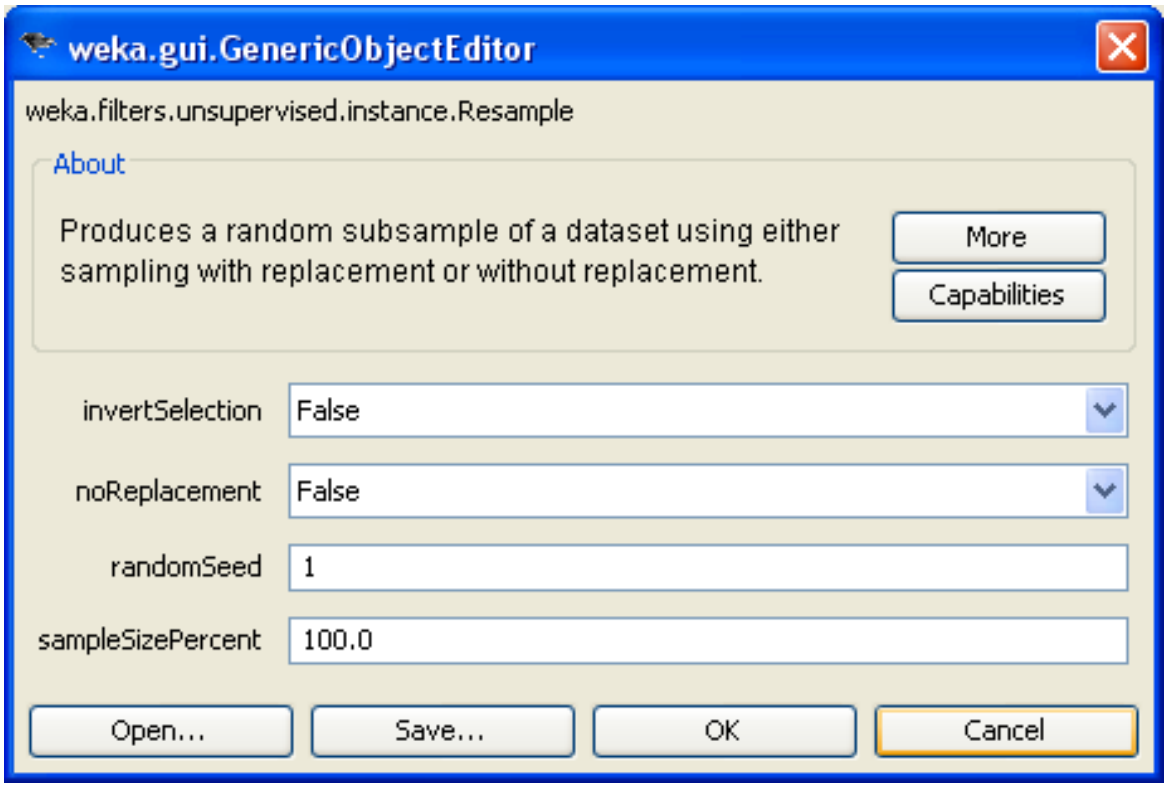


$$v' = \frac{v - \mu}{\sigma}$$

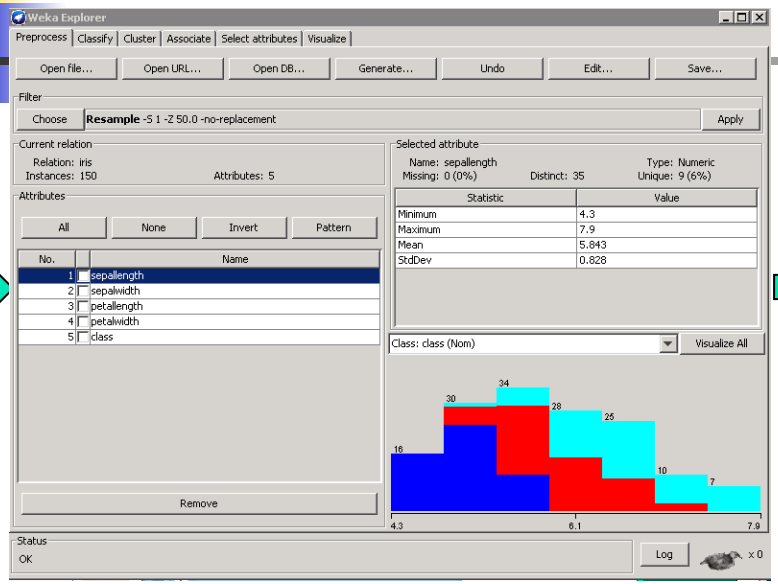
μ= mean values

σ= standard deviation

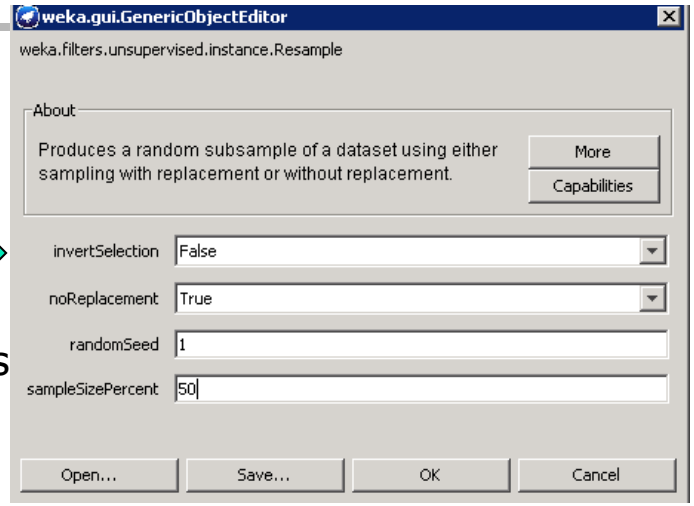
Unsupervised Filters: Resample



Undersampling Instances(Unsupervised)

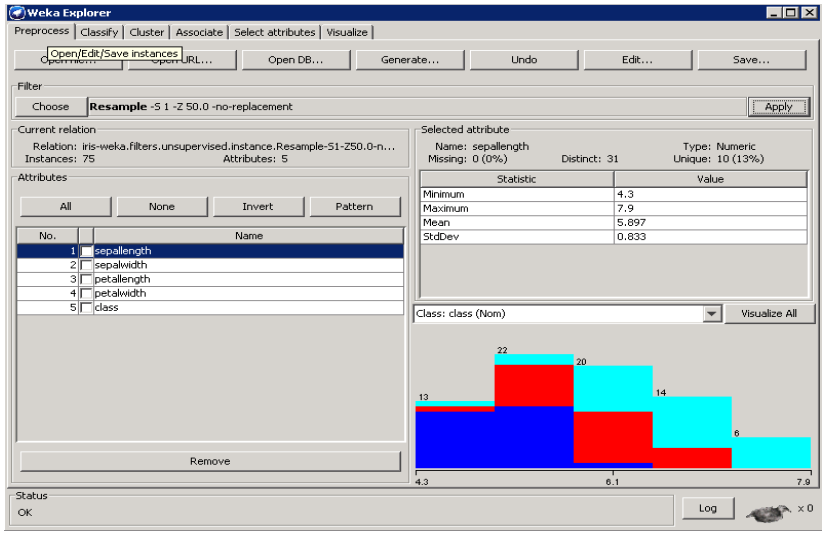


Load Data



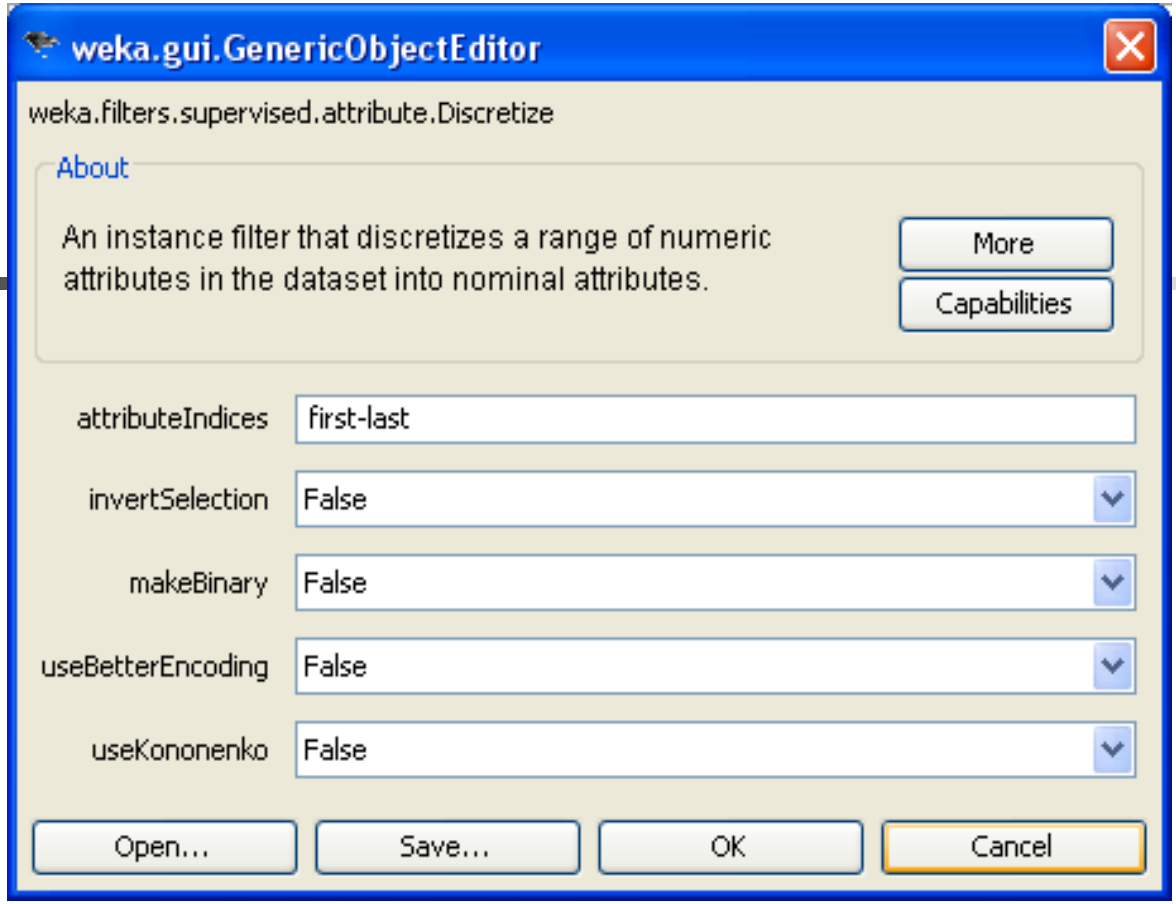
Set Parameters

Apply Algorithm



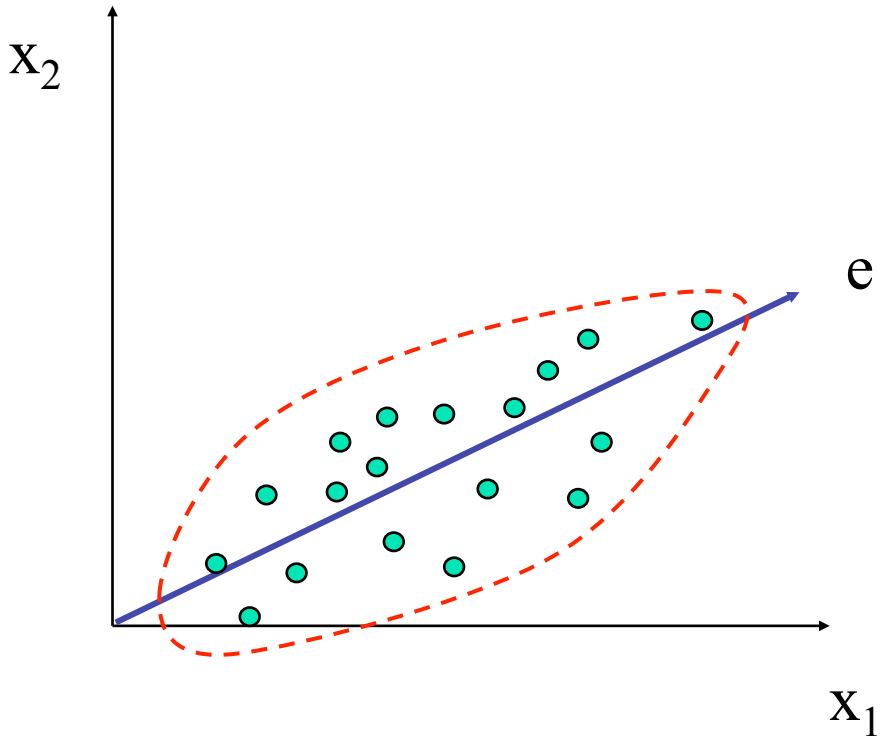
Save or Use Data

Unsupervised Filters: Discretize with Fayyad & Irani's Algorithm



Principal Component Analysis (PCA)

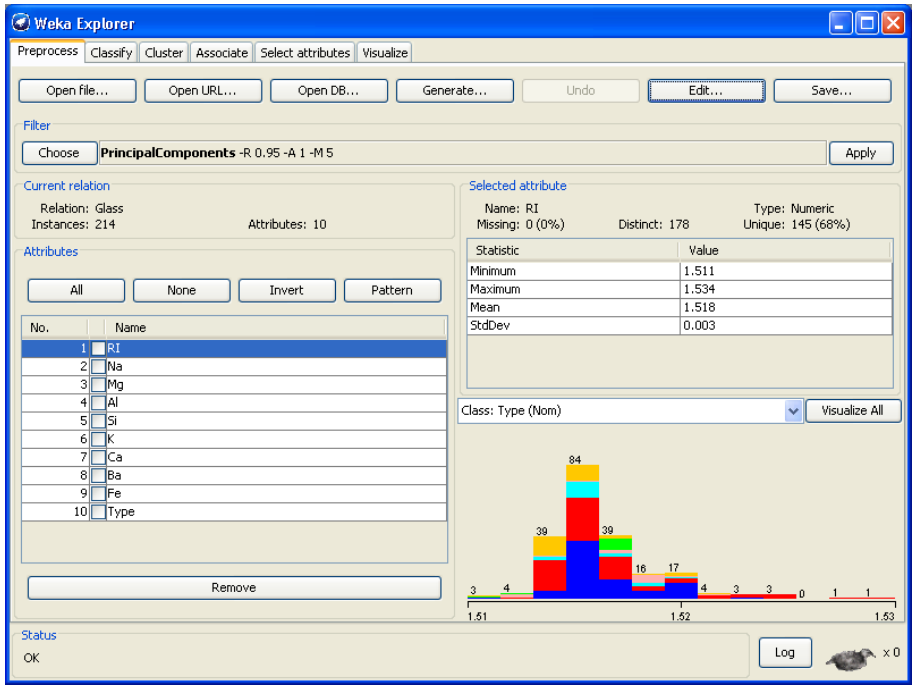
- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space

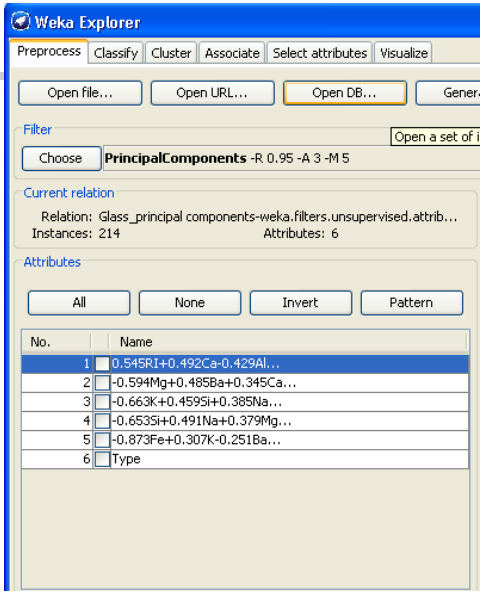


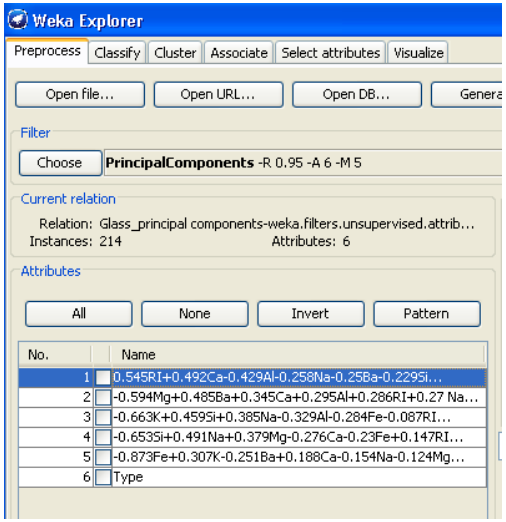
Principal Component Analysis (Steps)

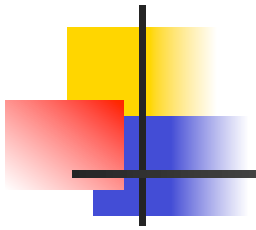
- Given *N* data vectors from *n*-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the *k* principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Works for numeric data only

Principal Component Analysis (PCA)



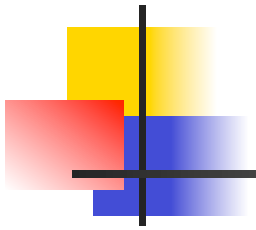






Command Line Filtering

- The **weka.filters** package is concerned with classes that transform datasets by removing or adding attributes, resampling the dataset, removing examples and so on.
- All filters offer the options -*i* for specifying the input dataset, and -*o* for specifying the output dataset. All others including specific parameters can be found via -*h*



Command Line Filtering

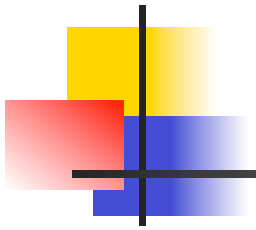
- In Unix based operating systems the classpath can be set by typing the following command:

```
export CLASSPATH=$CLASSPATH:/CompletePathOfweka/weka.jar
```

- For Windows OS:

1. In the Control Panel click on System (or right click on My Computer and select Properties) and then go to the Advanced tab. There you will find a button called Environment Variables, click it.
2. Enter the following name for the variable CLASSPATH
3. Add this value /CompletePathOfweka/weka.jar, where CompletePathOfweka is your own path in which weka.jar file is located.

Check on the web instructions on how set Environmental Variables in your specific WIN OS. You will find also videos on you tube.



Command Line Filtering: Examples

1) Write on the console, after adding weka.jar to the CLASSPATH

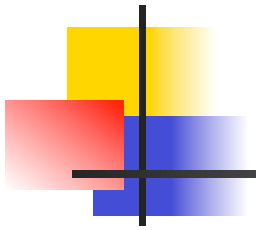
java weka.filters.unsupervised.attribute.PrincipalComponents

-I yourPahtOfDataset/iris.arff -o iris-PC.arff -c last

2) Resample creates a stratified subsample of the given dataset. This means that overall class distributions are approximately retained within the sample. A bias towards uniform class distribution can be specified via -B.

java weka.filters.supervised.instance.Resample -i yourPahtOfDataset/soybean.arff -o soybean-5%.arff -c last -Z 5

java weka.filters.supervised.instance.Resample -i yourPahtOfDataset/soybean.arff -o soybean-uniform-5%.arff -c last -Z 5 -B 1



For additional information on

1) how to use WEKA via command line

2) Set the CLASSPATH

please check the WEKA manual, where examples are provided.