

**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](mailto:p.ducange@iet.unipi.it)**

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

**Tutoring Hours: send an email for reserving your time slot**



## Business Intelligence Laboratory

This is a virtual space for students attending the Course of Business Intelligence of the University of Pisa.

Notifiche attivate



Pubblica 1 membro

Cerca nella community

Tutti i post

Eventi

Foto

Membri (1)

Vedi tutti



Condividi le ultime novità...



Testo



Foto



Link



Video



Evento



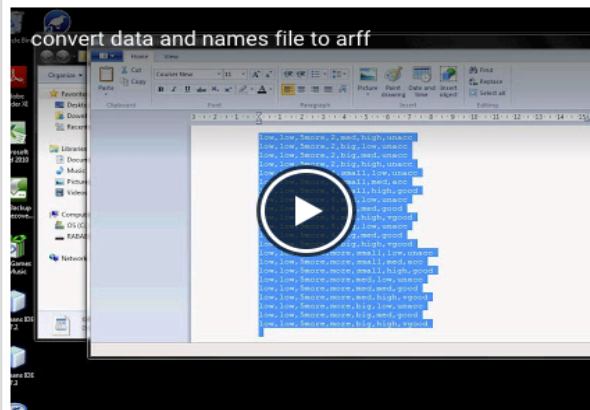
Sondaggio



**Pietro Ducange** PROPRIETARIO

Discussione - 15 ott 2015

This video shows a way to convert CSV data into arff file. Different approaches can be followed. This is just an example. If you have any problem, please do not hesitate to contact me.



+1



Aggiungi un commento...

Spargi la voce

Invita persone

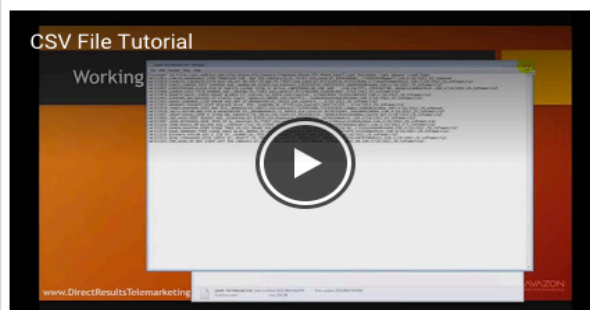
Condividi questa community



**Pietro Ducange** PROPRIETARIO

Discussione - 15 ott 2015

Here you are a tutorial about CSV file, very useful for preparing and handling dataset that you can download for example from UCI repository!



+1



Aggiungi un commento...



**Pietro Ducange** PROPRIETARIO

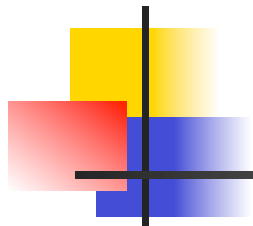
Discussione - 15 ott 2015

Let's start with the BI virtual laboratory. Students are invited to submit questions, issues and problems in this space. I will try to comment and help them to approach the difficulties of the actual laboratory class and of all other issues.

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

<http://bit.ly/2dFxLUx>





# **WEKA**

## **Waikato Environment for Knowledge Analysis**

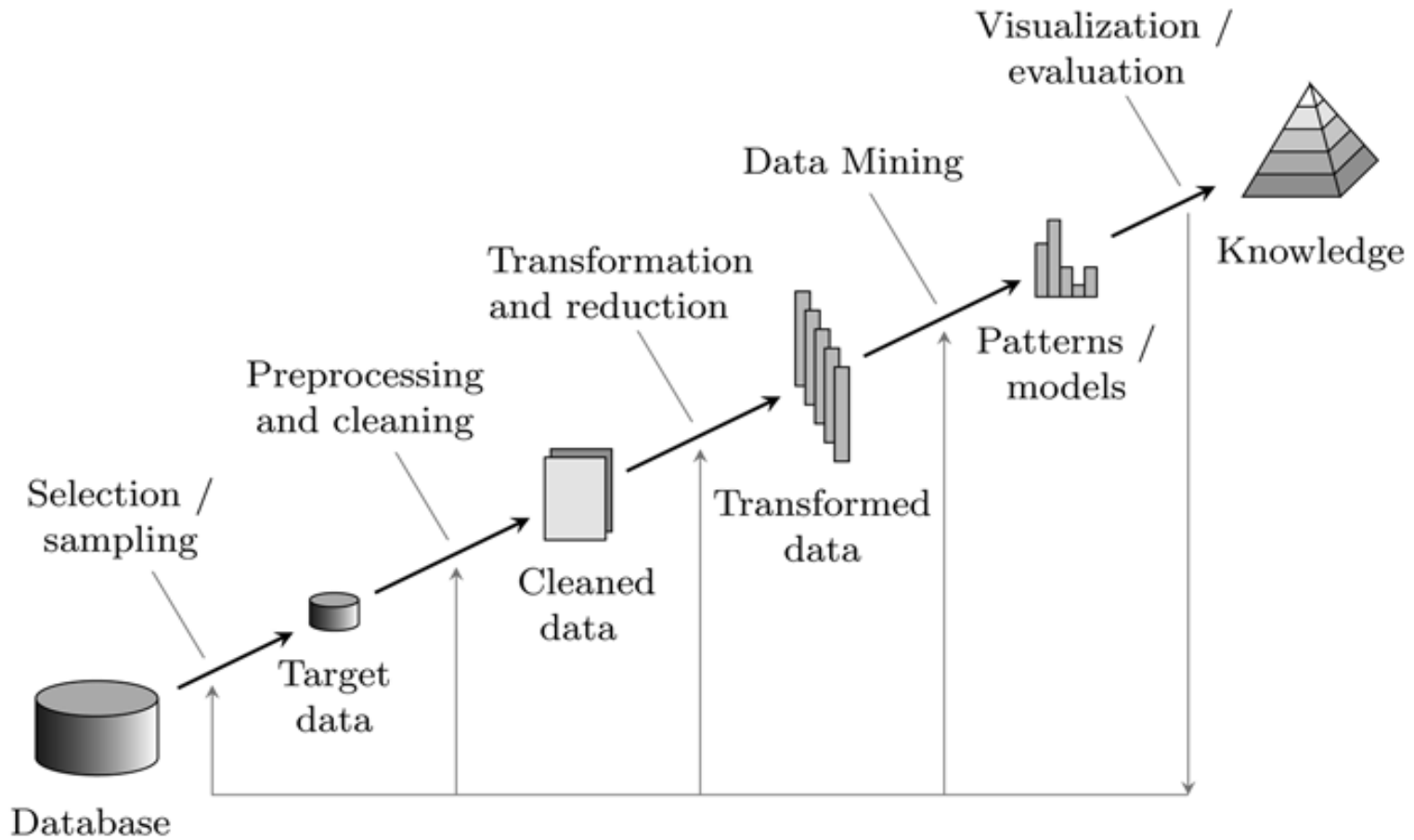
### **An Introduction**



# 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](http://www.cs.waikato.ac.nz/ml/weka)

- It is possible to download either a platform-specific installer or an executable Java jar file

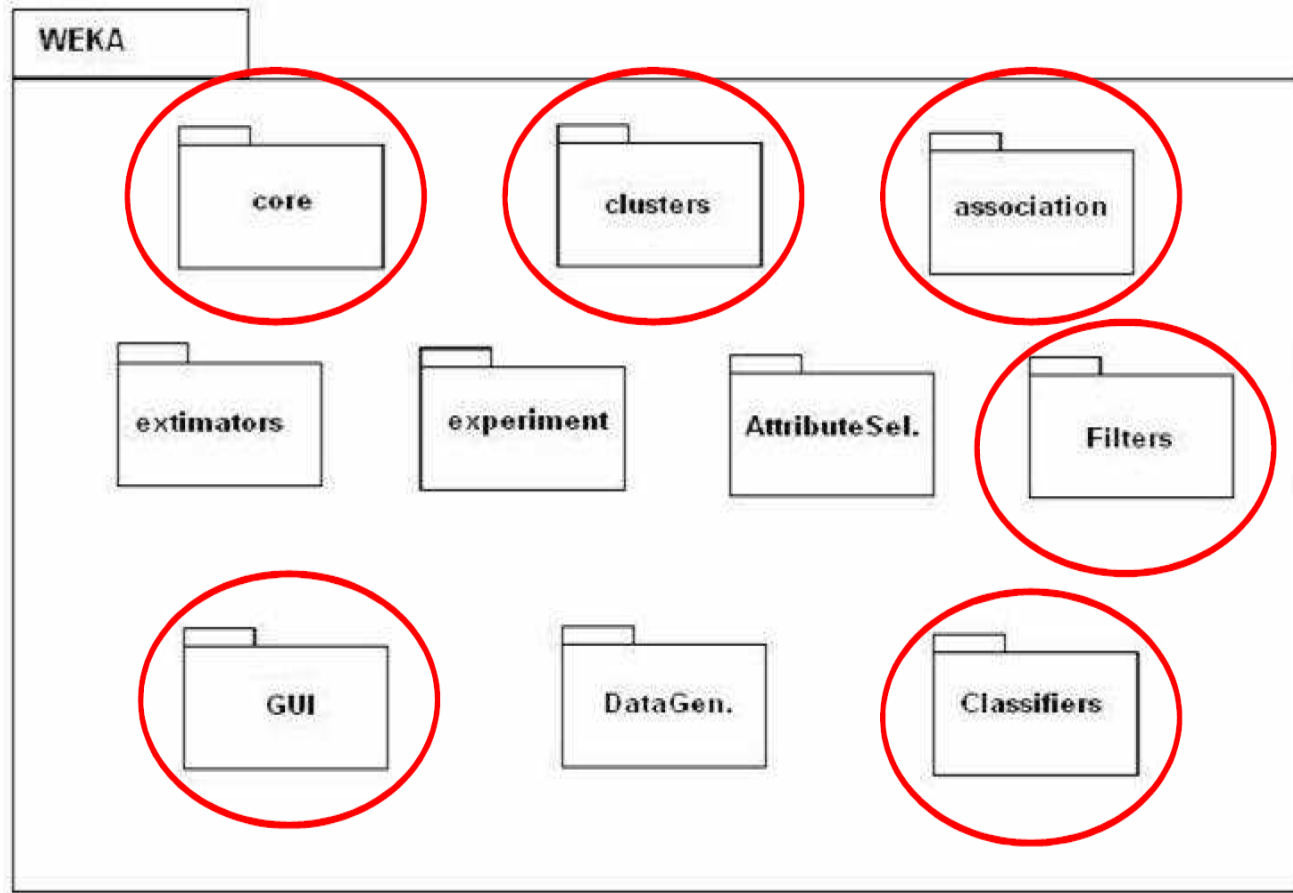
- Weka Version 3-6-10

The screenshot shows a web browser window displaying the Weka website. The URL in the address bar is [www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/). The page features the Weka logo (a bird) and the text 'WEKA The University of Waikato'. A navigation menu includes 'Project', 'Software' (which is underlined), 'Book', 'Publications', 'People', and 'Related'. On the left side, there is a sidebar with links for 'Home', 'Getting started', 'Requirements', 'Download', 'Documentation', 'FAQ', 'Citing Weka', 'Further information', 'Datasets', 'Related Projects', 'Miscellaneous Code', 'Other Literature', 'Developers', 'Development', 'History', 'Subversion', 'Contributors', 'Various', and 'Wekalist stats'. The main content area is titled 'Downloading and installing Weka' and contains the following information:

- **Snapshots**  
Every night a snapshot of the Subversion repository is taken, compiled and put together in ZIP files. For those who want to have the latest bugfixes, they can download these snapshots [here](#).
- **Stable book 3rd ed. version**  
Weka 3.6 is the latest stable version of Weka, and the one described in the 3rd edition of the **data mining book**. This branch of Weka receives bug fixes only (for new features in Weka see the developer version). There are different options for downloading and installing it on your system:
  - **Windows x86**  
Click [here](#) to download a self-extracting executable that includes Java VM 1.6 (weka-3-6-6jre.exe; 36.9 MB)  
Click [here](#) to download a self-extracting executable without the Java VM (weka-3-6-6.exe; 22.2 MB)  
These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.5 (or later) on your system.
  - **Windows x64**  
Click [here](#) to download a self-extracting executable that includes 64 bit Java VM 1.6 (weka-3-6-6jre-x64.exe; 37.0 MB)



# 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.

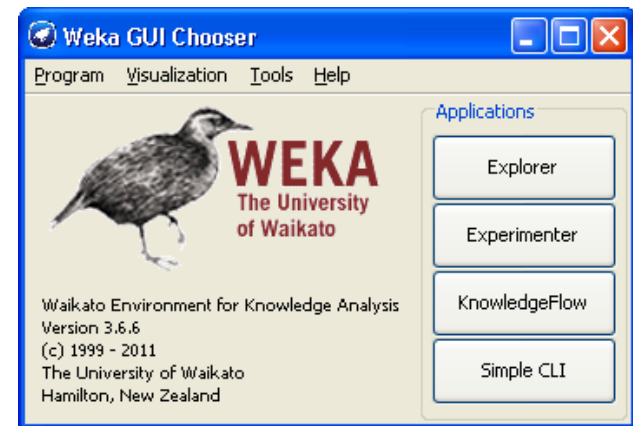
- Visualization

- ❖ 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.

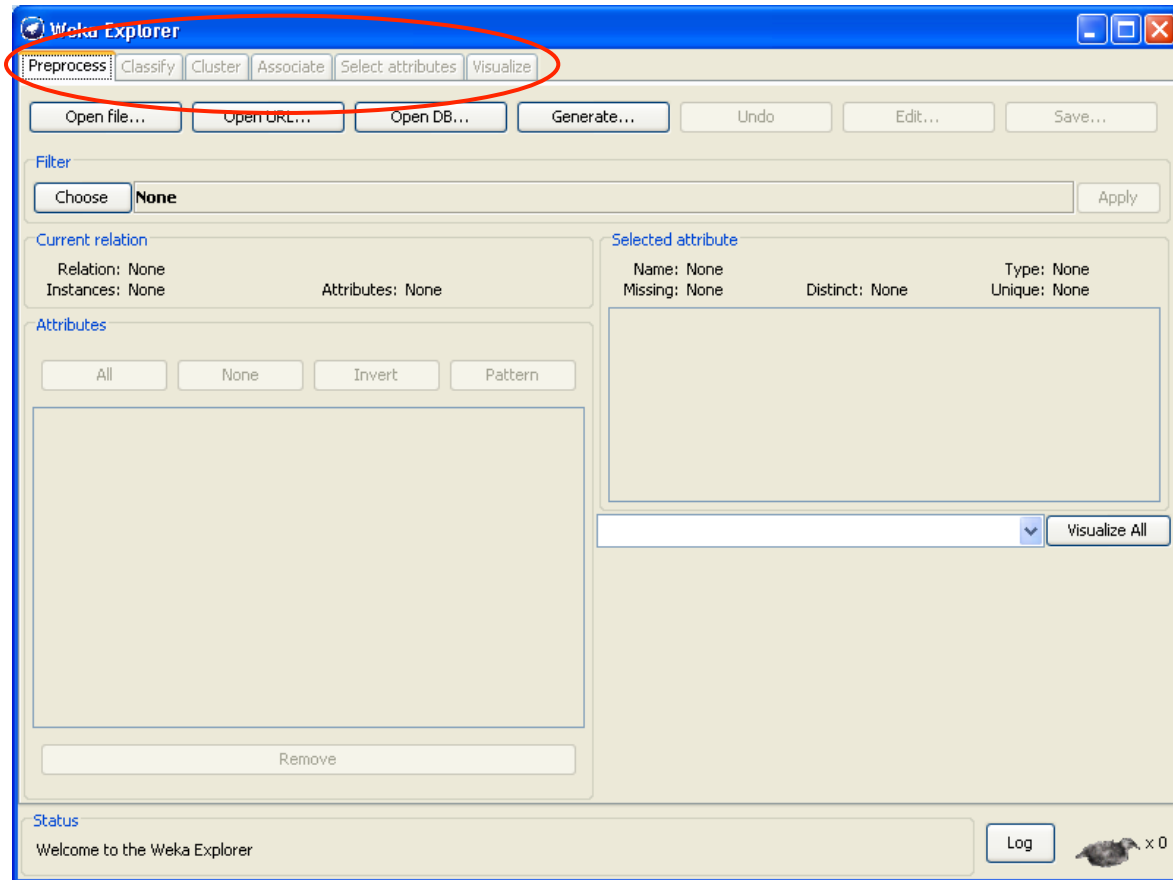
- Help Online resources



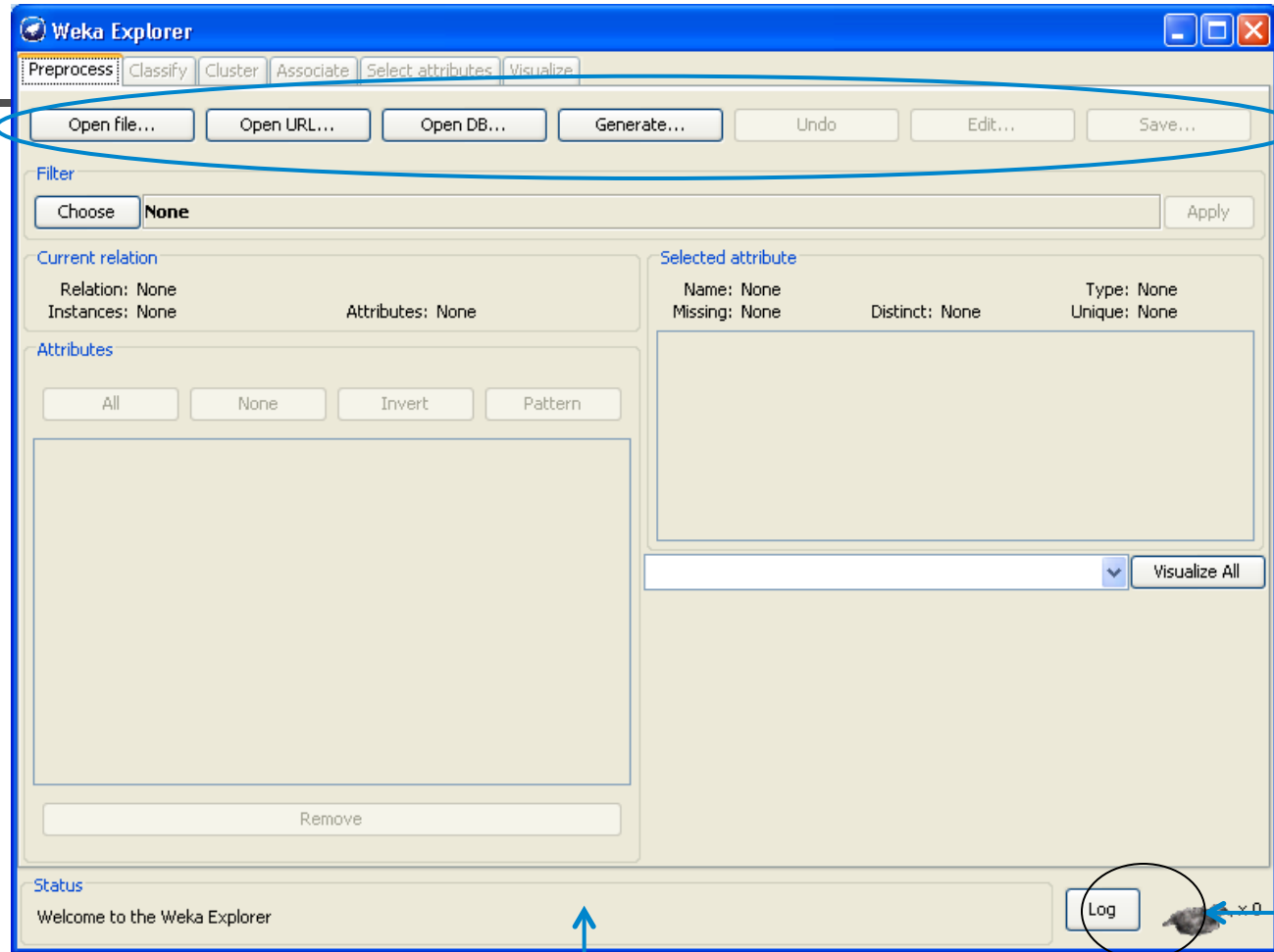


# Explorer (1)

- 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

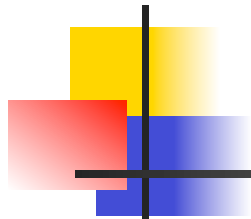


# Loading data

- **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

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

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

The screenshot shows the Weka Explorer interface. At the top, there are menu options: Preprocess, Classify, Cluster, Associate, Select attributes, Visualize. Below that are buttons for Open file..., Open URL..., Open DB..., Generate..., Undo, Edit..., and Save... A Filter section has a 'Choose' button and 'None' selected. The 'Current relation' section shows 'Relation: iris' and 'Instances: 150'. The 'Attributes' section shows a list of 5 attributes: sepalength, sepalwidth, petalength, petalwidth, and class. The 'Selected attribute' section shows 'Name: sepalength', 'Missing: 0 (0%)', 'Distinct: 35', and 'Type: Numeric Unique: 9 (6%)'. A table of statistics for the selected attribute is shown: Minimum 4.3, Maximum 7.9, Mean 5.843, StdDev 0.828. Below this is a histogram with a 'Class: class (Nom)' dropdown and a 'Visualize All' button. The histogram has four bars with values 16, 30, 34, and 25, and a fifth bar with value 10. The x-axis is labeled with 4.3, 6.1, and 7.9. The status bar at the bottom says 'OK'.



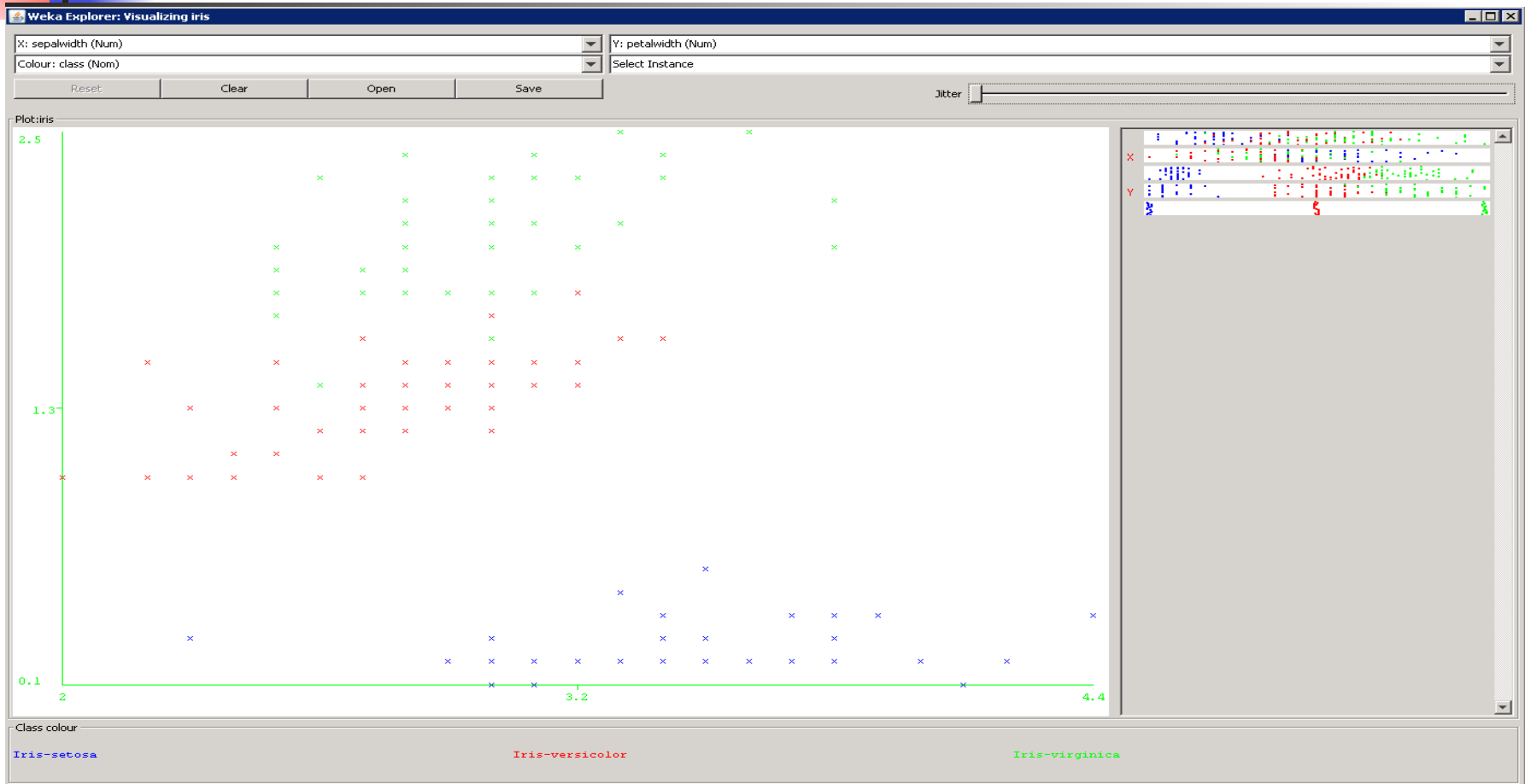
# Converters to ARFF (i)

- 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

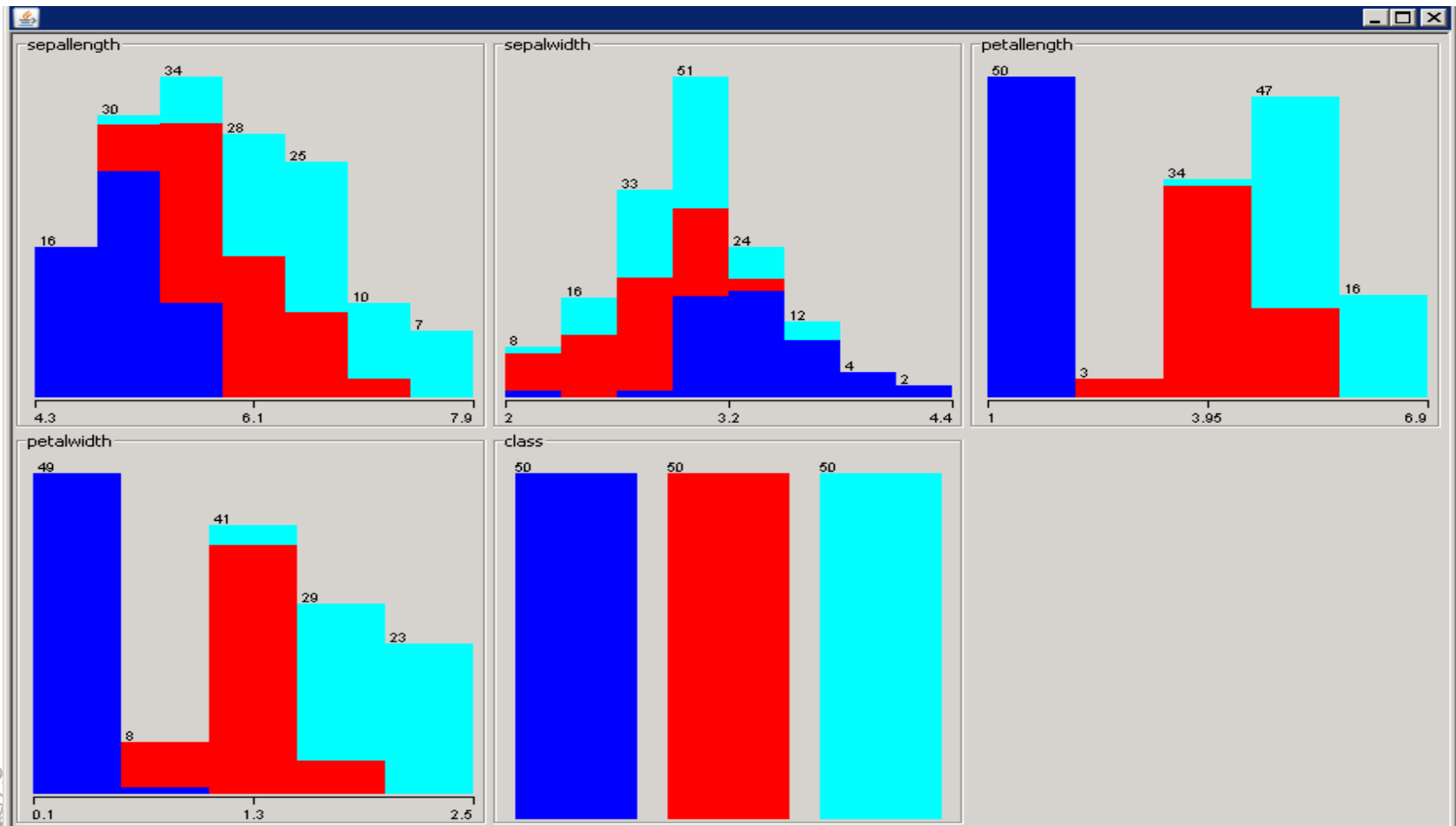


*The jitter function just adds artificial random noise to the coordinates of the plotted points in order to spread the data out a bit (so that you can see points that might have been obscured by others).*



# Visualizing Data (Histograms)

Click **Visualize All**





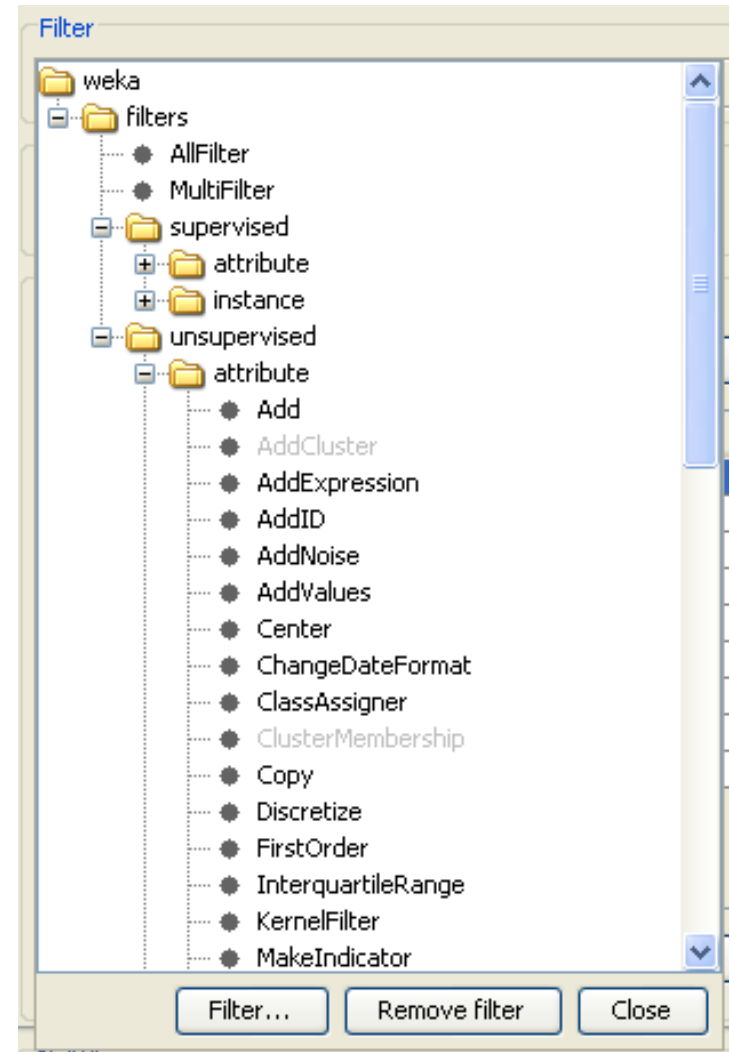
# Working with filters (1)

- 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.



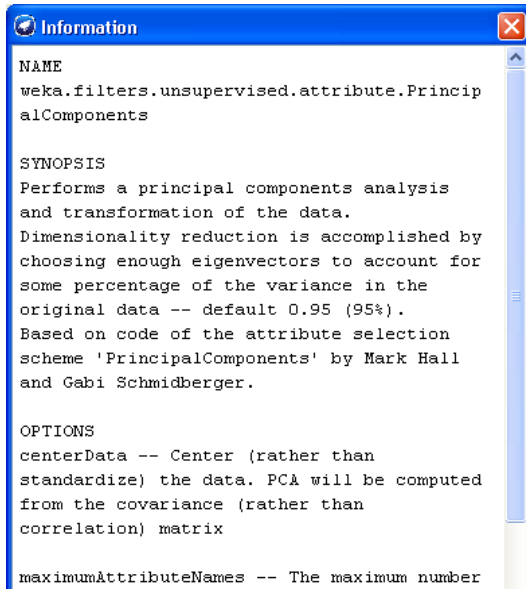
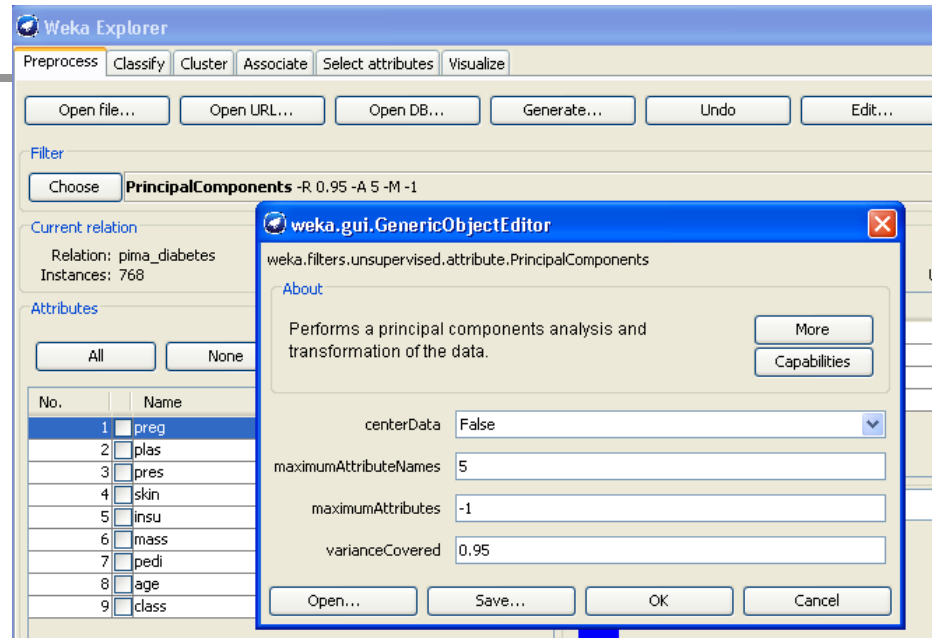
## Working with filters (2)

- 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)

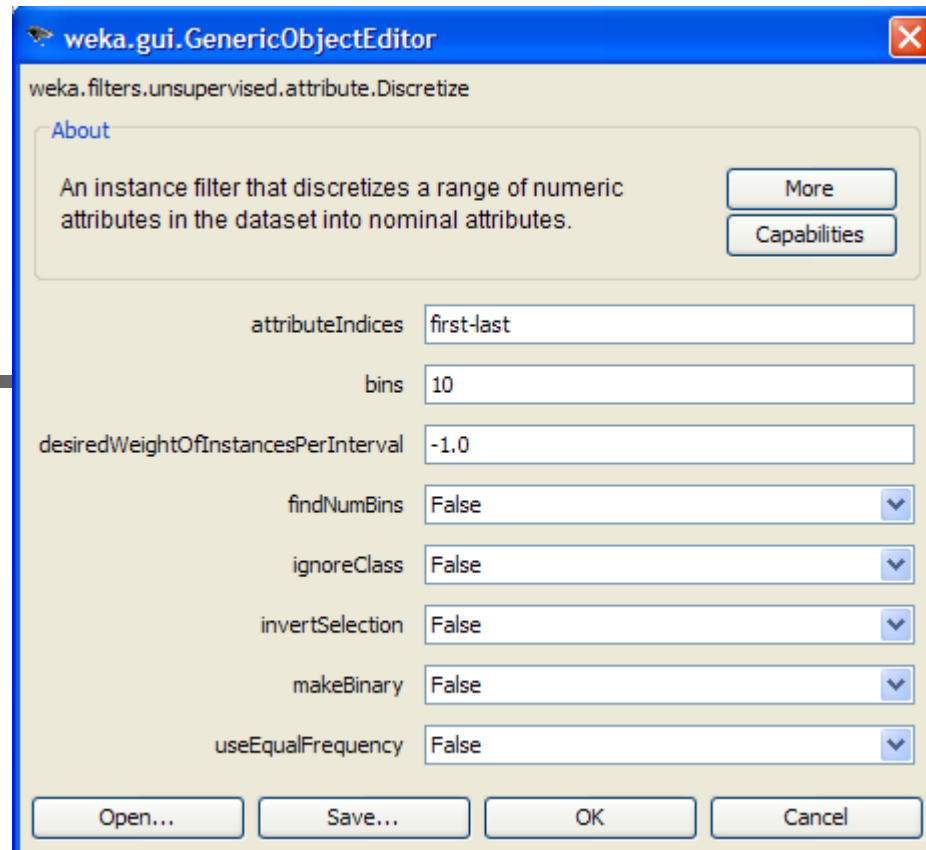
- When a filter is selected, its name and its parameters appears in the line beside the Choose button
- Click that line to get a generic object editor to specify the filter properties.



- 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

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Filter: Choose **Discretize -B 10 -M -1.0 -R first-last** Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Selected attribute: Name: sepalength Missing: 0 (0%) Distinct: 35 Type: Numeric Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

Set Parameters

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About: An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices: first-last

bins: 10

desiredWeightOfInstancesPerInterval: -1.0

findNumBins: False

ignoreClass: False

invertSelection: False

makeBinary: False

useEqualFrequency: False

Open... Save... OK Cancel

Apply Algorithm

Save or Use Data

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Filter: Choose **Discretize -B 10 -M -1.0 -R first-last** Apply

Current relation: Relation: iris-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0... Instances: 150 Attributes: 5

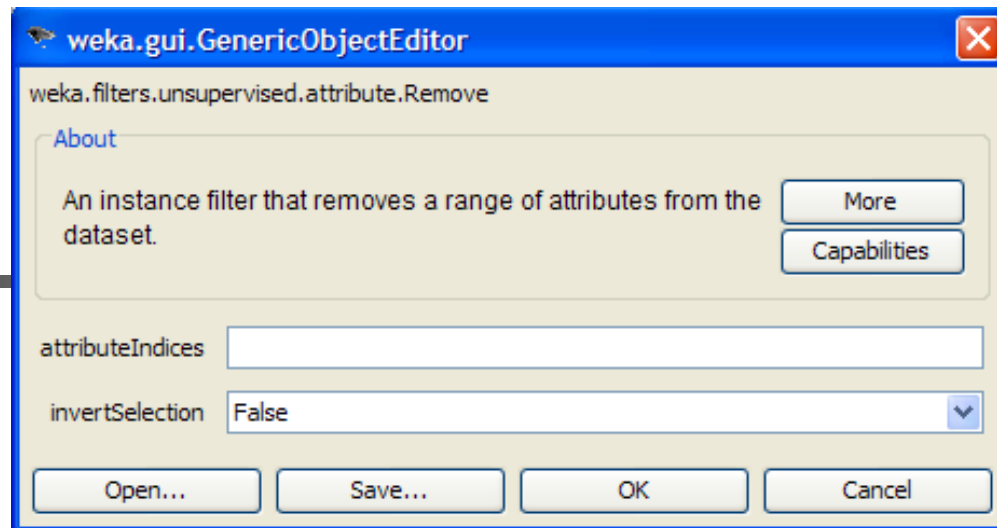
Selected attribute: Name: sepalength Missing: 0 (0%) Distinct: 10 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	(-inf-4.66]	9
2	(4.66-5.02]	23
3	(5.02-5.38]	14
4	(5.38-5.74]	27
5	(5.74-6.1]	22
6	(6.1-6.46]	20
7	(6.46-6.82]	18
8	(6.82-7.18]	6

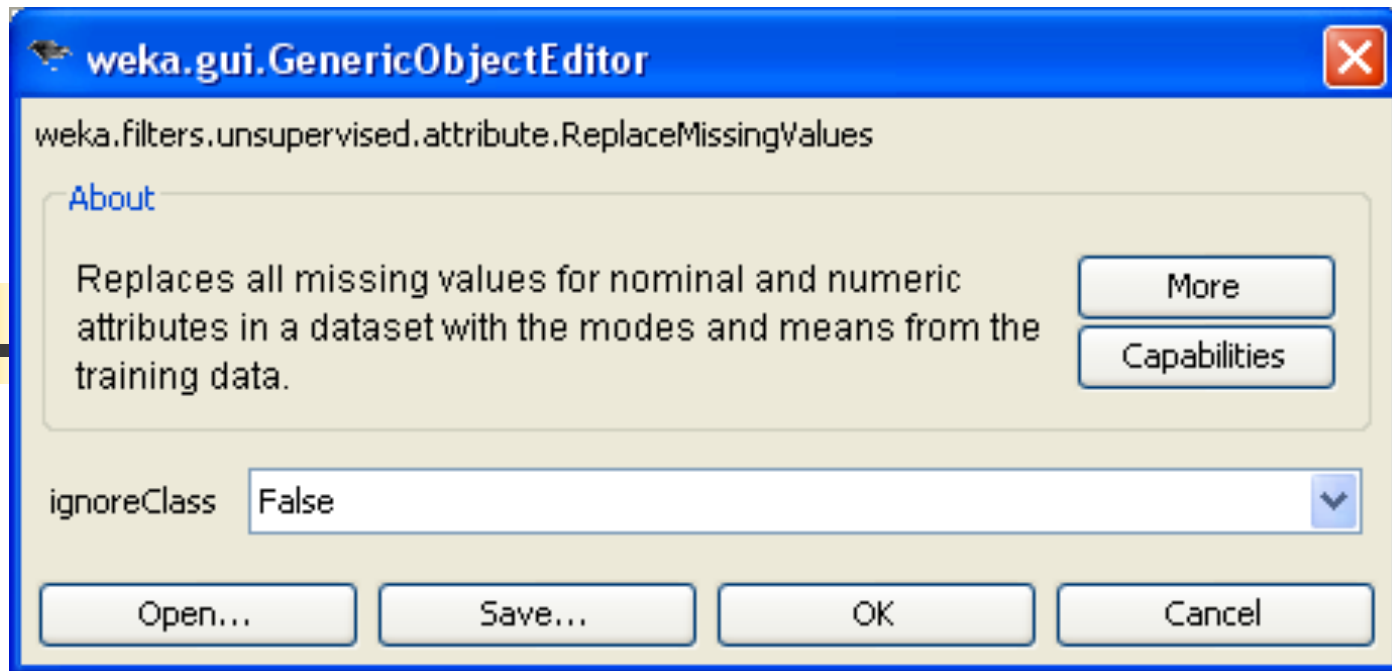
Class: class (Nom) Visualize All



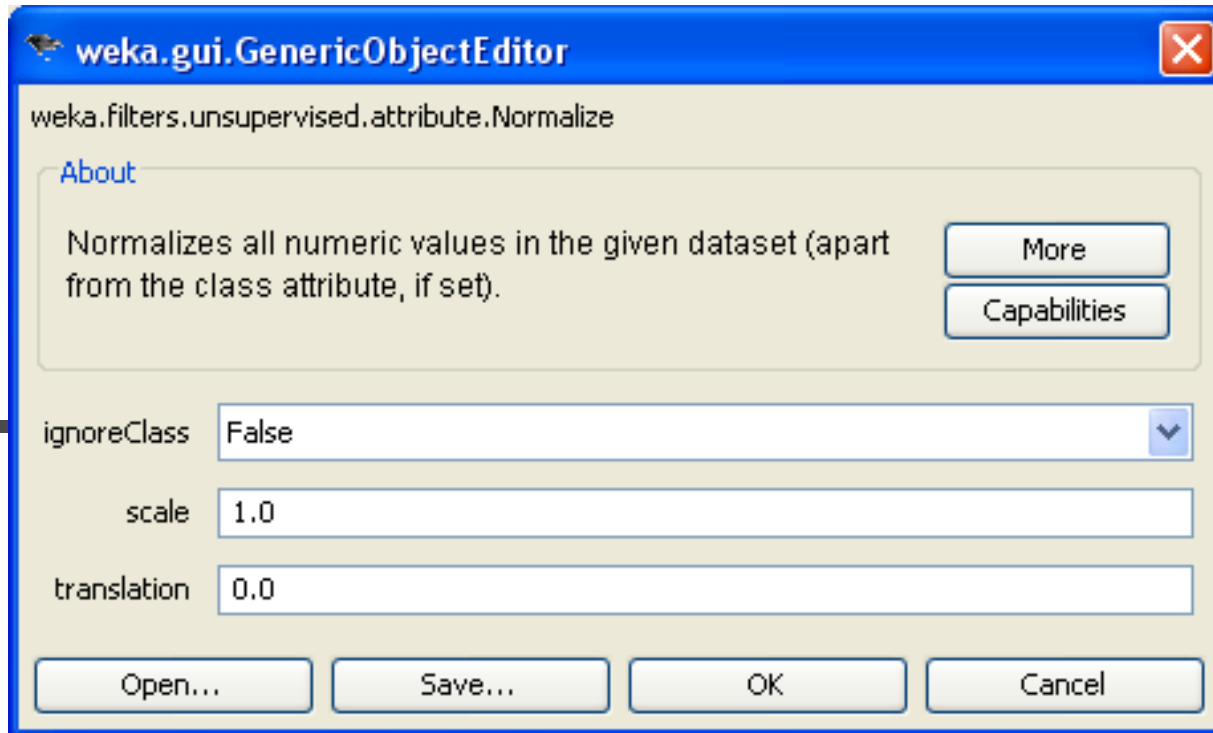
# Unsupervised Filters: Attribute Remove



# Unsupervised Filters: Replace Missing Values



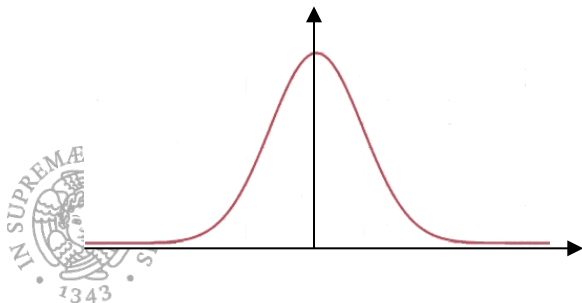
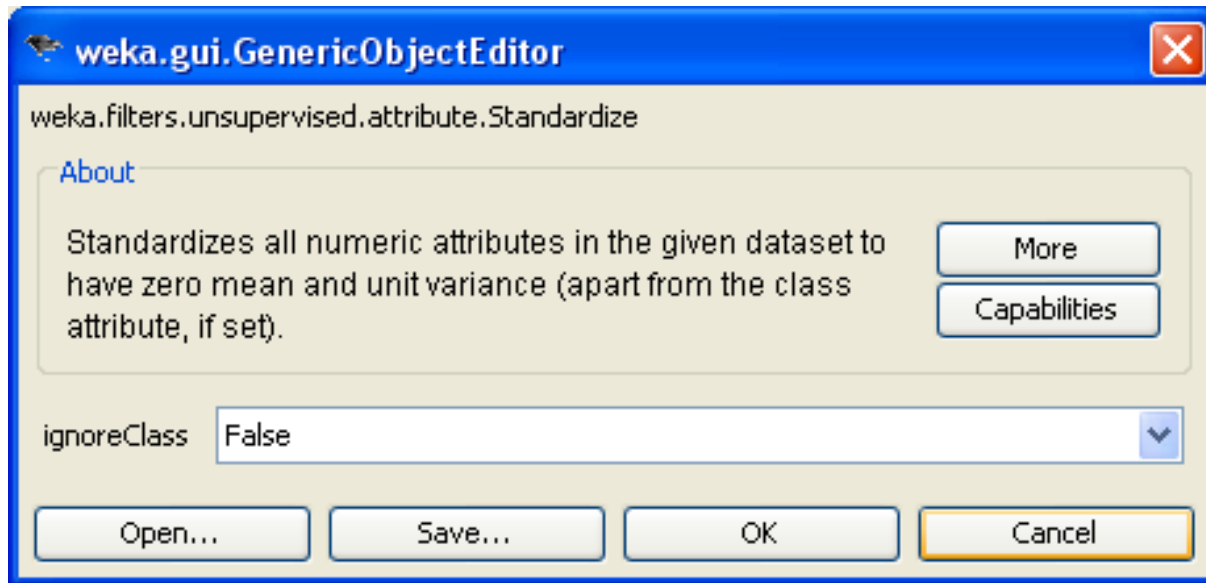
# Unsupervised Filters: Normalize



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



# Unsupervised Filters: Standardize



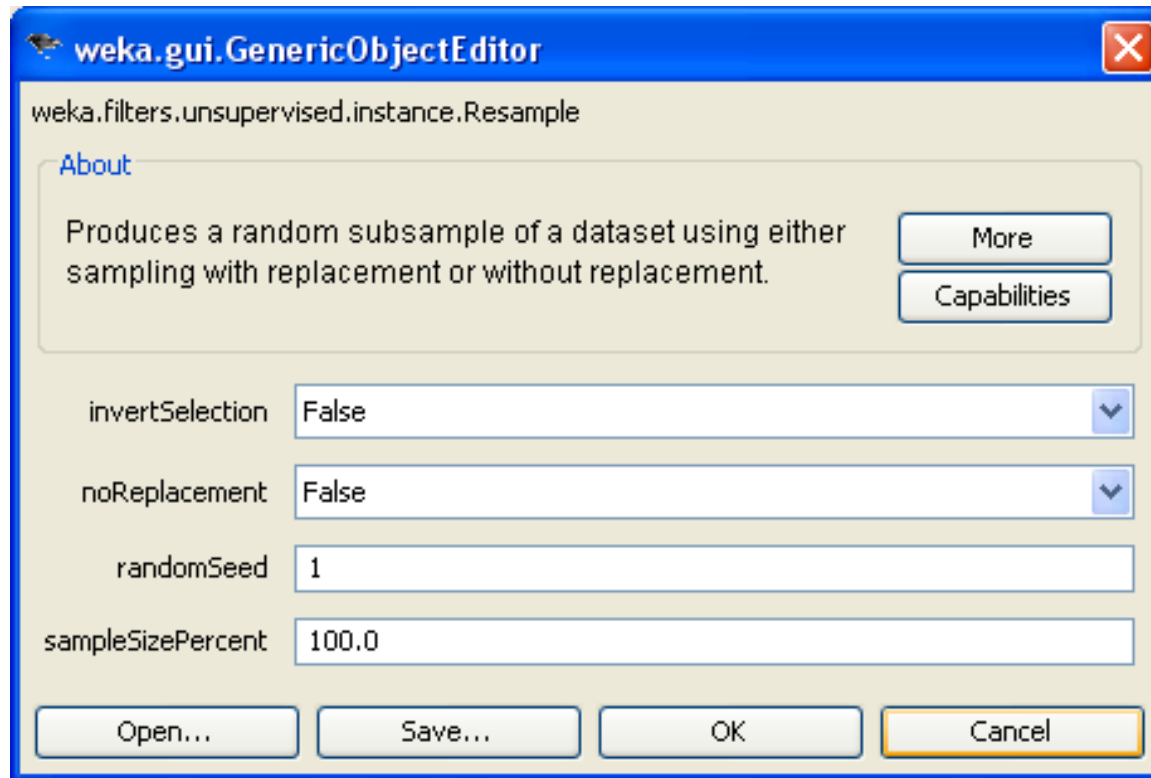
$$v' = \frac{v - \mu}{\sigma}$$

$\mu$  = mean values

$\sigma$  = standard deviation



# Unsupervised Filters: Resample



# Undersampling Instances(Unsupervised)

Load Data

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Filter: Choose **Resample** -S 1 -Z 50.0 -no-replacement

Current relation: Relation: iris, Instances: 150, Attributes: 5

Selected attribute: Name: sepal.length, Missing: 0 (0%), Distinct: 35, Type: Numeric, Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

Set Parameters

weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.Resample

About: Produces a random subsample of a dataset using either sampling with replacement or without replacement.

invertSelection: False

noReplacement: True

randomSeed: 1

sampleSizePercent: 50

Apply Algorithm

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Filter: Choose **Resample** -S 1 -Z 50.0 -no-replacement

Current relation: Relation: iris-weka.filters.unsupervised.instance.Resample-51-Z50.0-n..., Instances: 75, Attributes: 5

Selected attribute: Name: sepal.length, Missing: 0 (0%), Distinct: 31, Type: Numeric, Unique: 10 (13%)

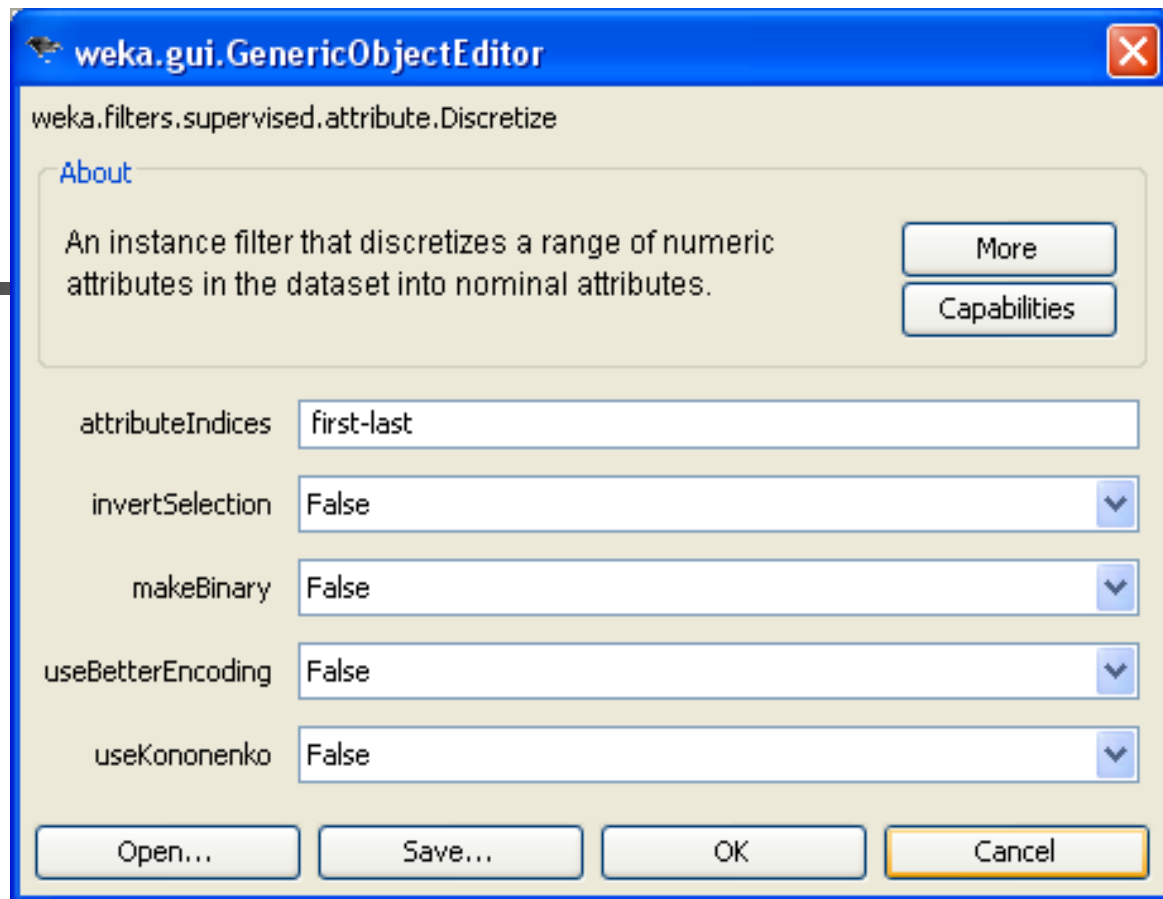
Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.897
StdDev	0.833

Class: class (Nom)

Save or Use Data

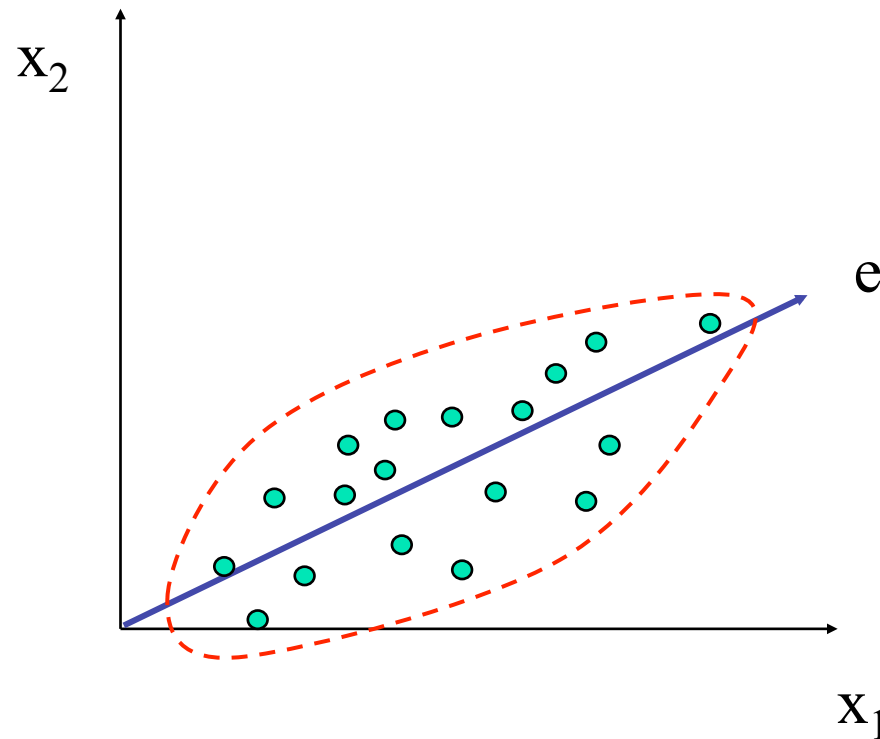


# Unsupervised Filters: Discretize with Fayyad & Irani's Algorithm



# 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



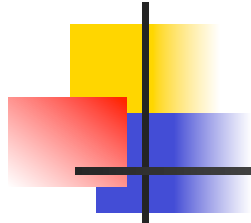
# Principal Component Analysis (Steps)

- 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)



**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **PrincipalComponents -R 0.95 -A 1 -M 5** Apply

Current relation: Relation: Glass Instances: 214 Attributes: 10

Attributes: All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	RI
<input type="checkbox"/>	Na
<input type="checkbox"/>	Mg
<input type="checkbox"/>	Al
<input type="checkbox"/>	Si
<input type="checkbox"/>	K
<input type="checkbox"/>	Ca
<input type="checkbox"/>	Ba
<input type="checkbox"/>	Fe
<input type="checkbox"/>	Type

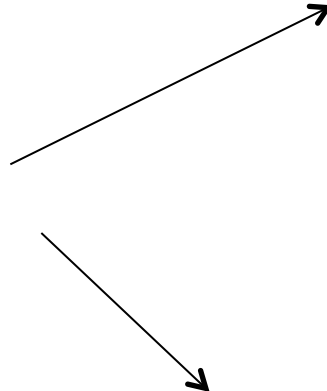
Selected attribute: Name: RI Missing: 0 (0%) Distinct: 178 Type: Numeric Unique: 145 (68%)

Statistic	Value
Minimum	1.511
Maximum	1.534
Mean	1.518
StdDev	0.003

Class: Type (Nom) Visualize All

3 4 39 84 39 16 17 4 3 3 0 1 1

Status: OK Log x 0



**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate...

Filter: Choose **PrincipalComponents -R 0.95 -A 3 -M 5** Open a set of i

Current relation: Relation: Glass\_principal components-weka.filters.unsupervised.attrib... Instances: 214 Attributes: 6

Attributes: All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	0.545RI+0.492Ca-0.429Al...
<input type="checkbox"/>	-0.594Mg+0.485Ba+0.345Ca...
<input type="checkbox"/>	-0.663K+0.459Si+0.385Na...
<input type="checkbox"/>	-0.653Si+0.491Na+0.379Mg...
<input type="checkbox"/>	-0.873Fe+0.307K-0.251Ba...
<input type="checkbox"/>	Type

**Weka Explorer**

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate...

Filter: Choose **PrincipalComponents -R 0.95 -A 6 -M 5**

Current relation: Relation: Glass\_principal components-weka.filters.unsupervised.attrib... Instances: 214 Attributes: 6

Attributes: All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	0.545RI+0.492Ca-0.429Al-0.258Na-0.258Ba-0.229Si...
<input type="checkbox"/>	-0.594Mg+0.485Ba+0.345Ca+0.295Al+0.286RI+0.27 Na...
<input type="checkbox"/>	-0.663K+0.459Si+0.385Na-0.329Al-0.284Fe-0.087RI...
<input type="checkbox"/>	-0.653Si+0.491Na+0.379Mg-0.276Ca-0.23Fe+0.147RI...
<input type="checkbox"/>	-0.873Fe+0.307K-0.251Ba+0.188Ca-0.154Na-0.124Mg...
<input type="checkbox"/>	Type



# 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

---

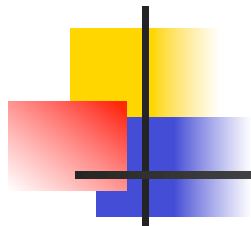
1) Write on the console, after adding weka.jar to the CLASSPATH

```
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.

