Lab 4

More Weka features

Useful information

Alex Becheru

irlab@becheru.net

irlab.becheru.net

What is this lab about?

- selecting the best attributes

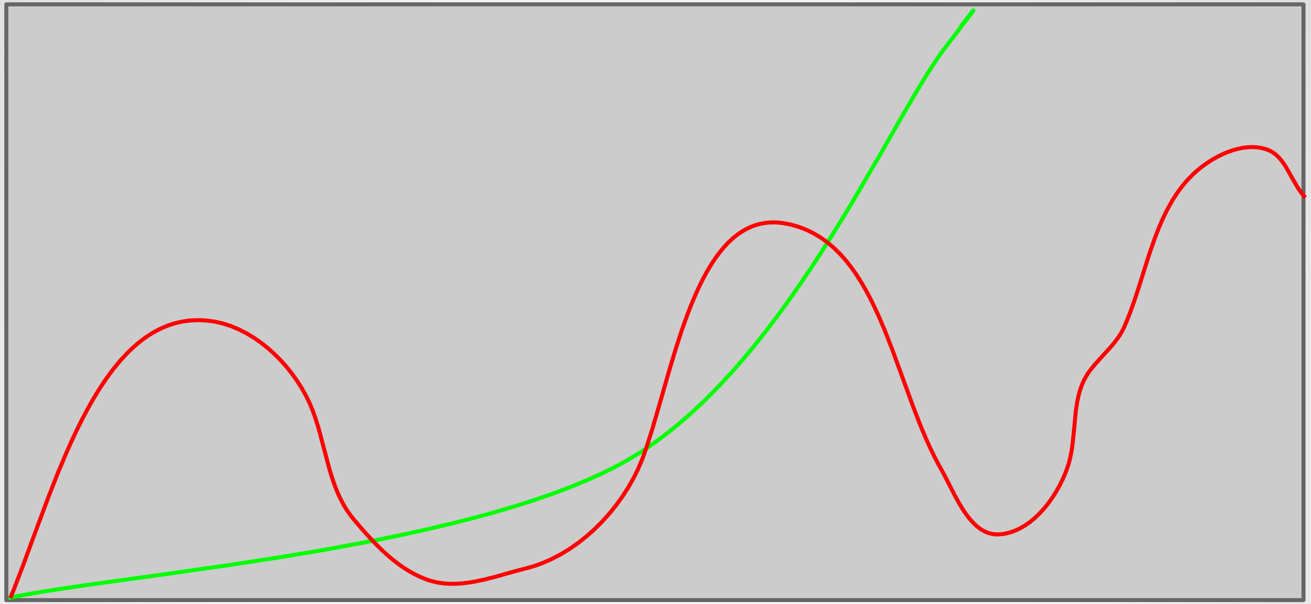
Question 1

Do attributes have effect on the IR algorithms?

Yes attributes have a big impact on the algorithms’ results

Attributes:

- may be redundant
- introduce useless information
- introduce little new knowledge



red line: an attribute that presents large variation of values
green line: an actual good attribute that can be easily predicted

In this case the red attribute will affect the results of the predictive function making it’s value to jump up and down.

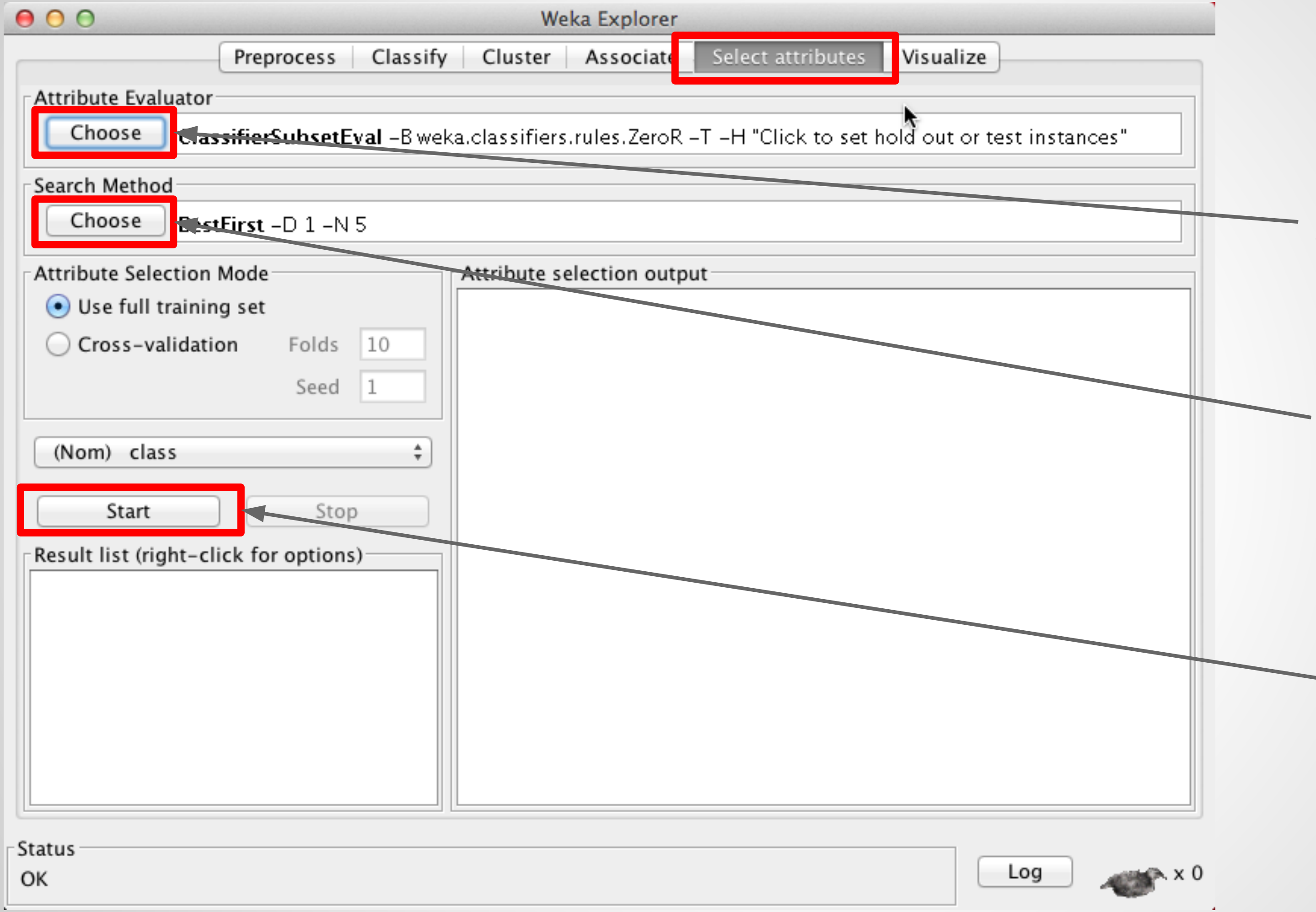
Question 2

How do I choose the best combination of attributes in such a way that my IR algorithms perform at their best?

Select attributes

There are two ways to analyse attributes:

- determine the best subset of attributes to be used
- individually rank each attribute



Attribute Evaluator

ClassifierSubsetEval -B weka.classifiers.rules.ZeroR -T -H "Click to set hold out or test instances"

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) class

Result list (right-click for options)

Attribute selection output

Status

OK

Choose the method to evaluate attributes

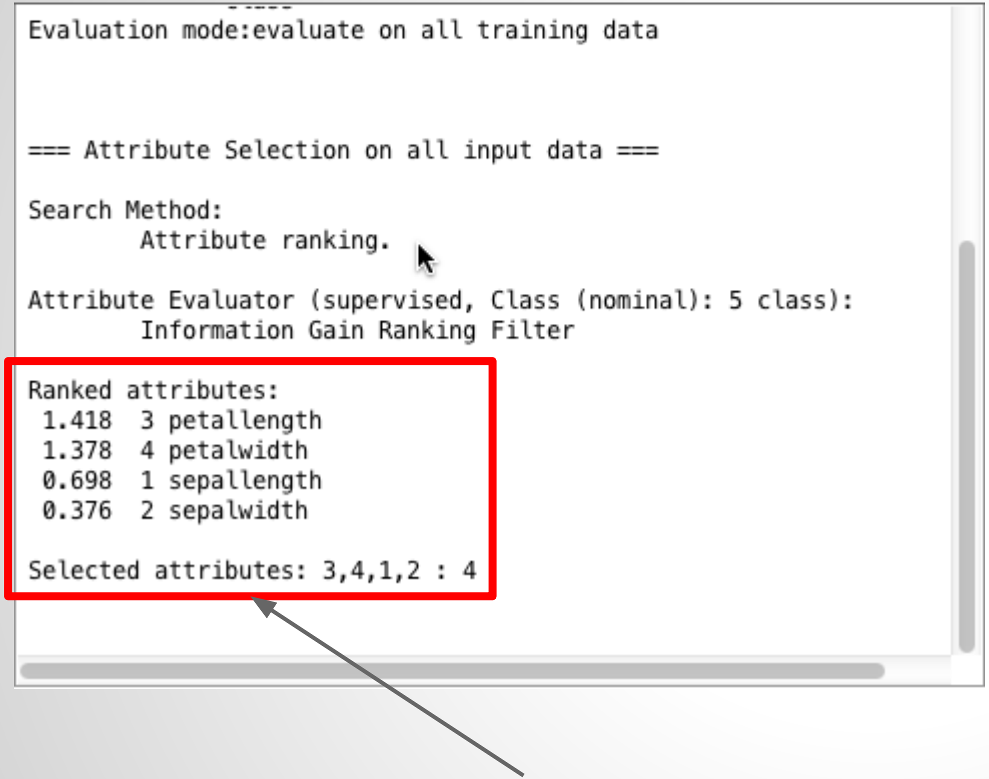
Choose the method to search the attribute space

Start comparison of attributes

Ex.1 Rank attributes

- open iris.data
- go to select attributes panel
- choose InfoGainAttributeEval attribute evaluator
- use Ranker T method for search
- remove attributes one at a time, and see the impact it has on the J48(trees) classifier

Attributes removal effects on J48 classifier



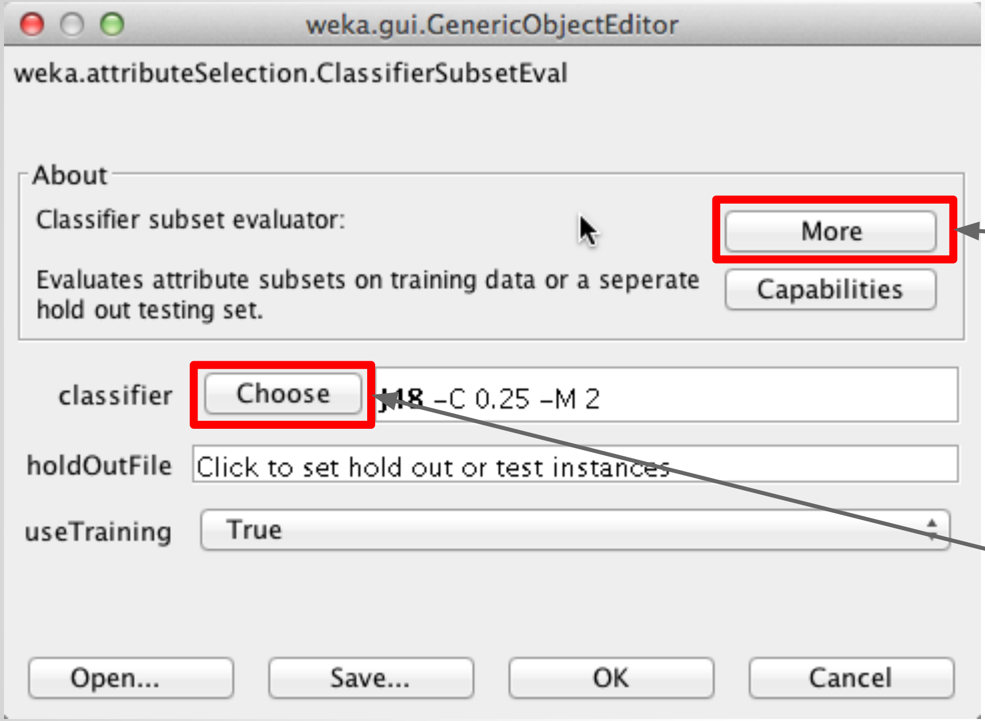
```
Evaluation mode:evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 class):
        Information Gain Ranking Filter

Ranked attributes:
 1.418   3 petallength
 1.378   4 petalwidth
 0.698   1 sepallength
 0.376   2 sepalwidth

Selected attributes: 3,4,1,2 : 4
```

The information gain of each attribute ranked

J48 initial: 96%
J48 - sepalwidth: 96%
J48 - sepalwidth: sepallength: 96%
J48 with only petallength: 94%

Select the best subset of attributes

- open iris.data
- go to select attributes panel
- choose ClassifierSubsetEval attribute evaluator
- use BestFirst search method
- determine the best subset to of attributes to be used with the J48 classifier



weka.gui.GenericObjectEditor

weka.attributeSelection.ClassifierSubsetEval

About

Classifier subset evaluator:

Evaluates attribute subsets on training data or a seperate hold out testing set.

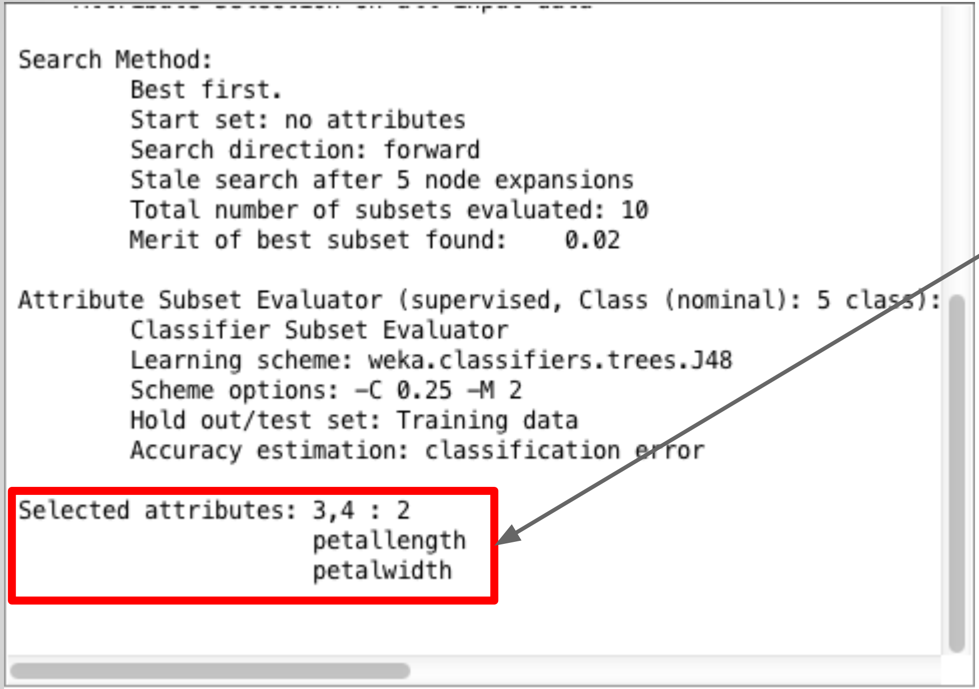
classifier: J48 -C 0.25 -M 2

holdOutFile: Click to set hold out or test instances

useTraining: True

Information about the evaluator

Choose the algorithm to be tested



```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 10
        Merit of best subset found:    0.02

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
        Classifier Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Hold out/test set: Training data
        Accuracy estimation: classification error

Selected attributes: 3,4 : 2
                     petallength
                     petalwidth
```

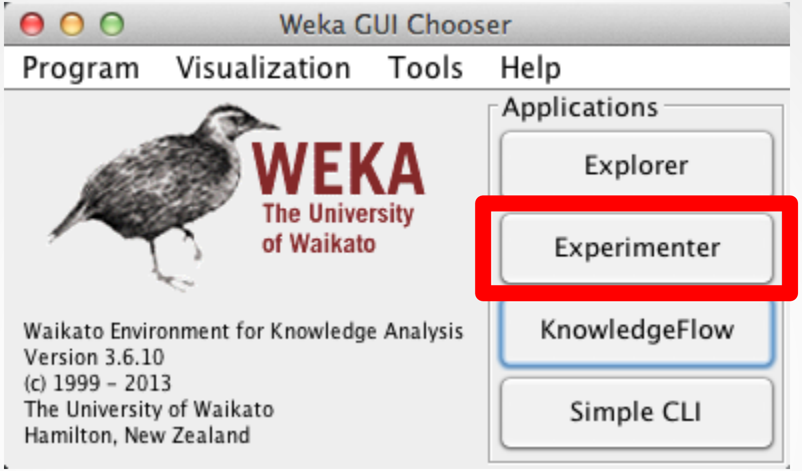
The smallest subset of attributes with the greatest value

Ex. 3

- open vote.arff
- determine the best subset of attributes to be used with j48
- rank the attributes information gain

Question 3?

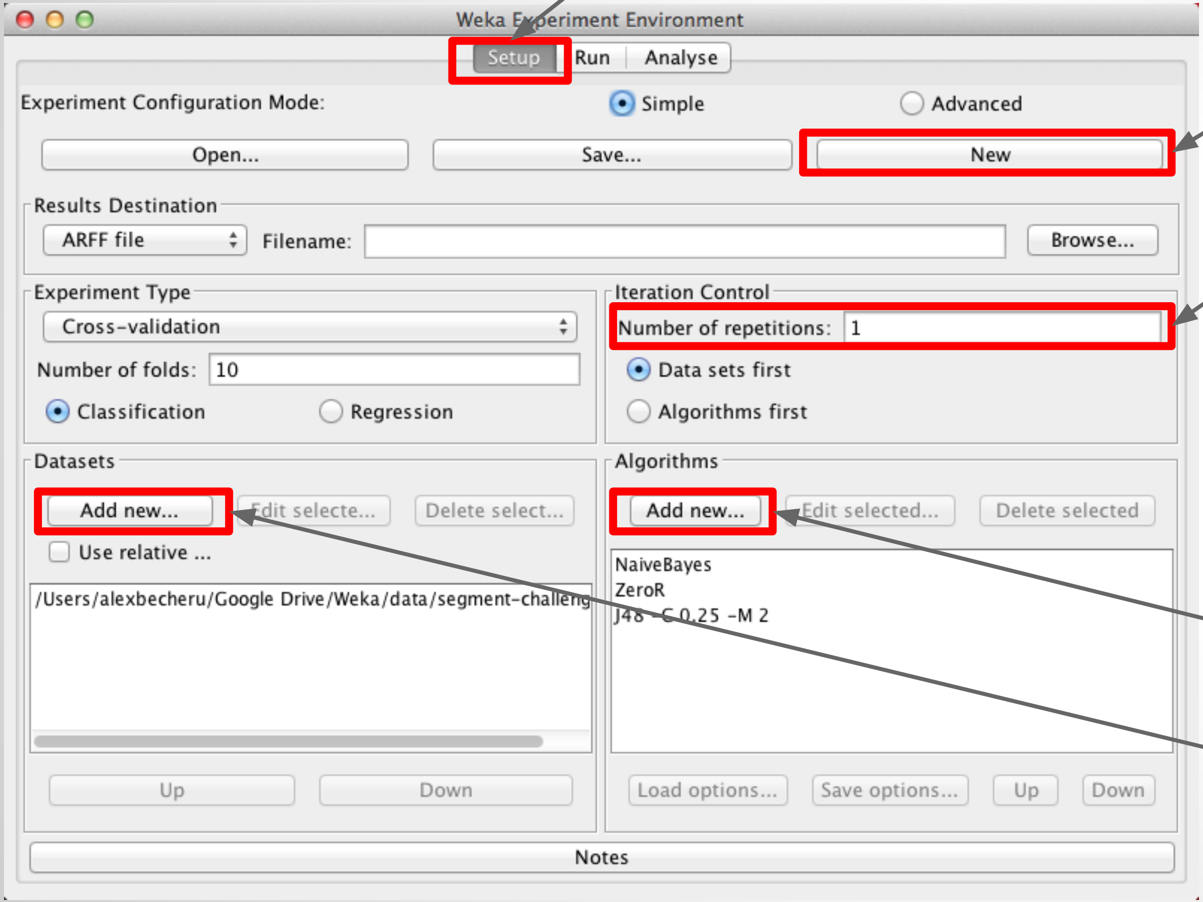
How can I compare different algorithms at the same time?



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

You can use the Experimenter environment in Weka to easily compare different algorithms.

1. Setup panel



Weka Experiment Environment

Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination

ARFF file Filename:

Experiment Type

Cross-validation

Number of folds: 10

● Classification ○ Regression

Iteration Control

Number of repetitions: 1

● Data sets first

○ Algorithms first

Datasets

□ Use relative ...

/Users/alexbecheru/Google Drive/Weka/data/segment-challeng

Algorithms

NaiveBayes
ZeroR
J48 -C 0.25 -M 2

2. Start new experiment

Determine how many time the experiment takes place

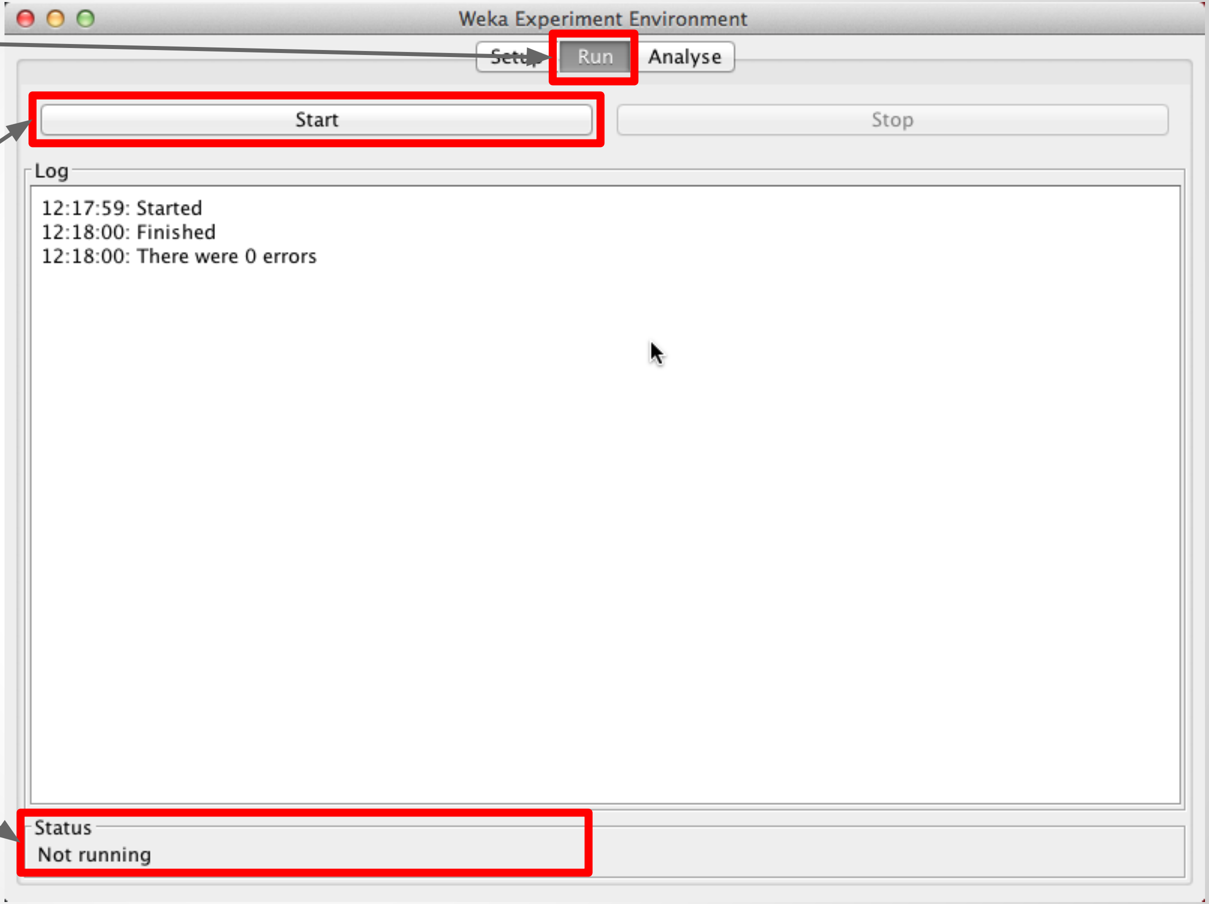
3. Add algorithms

4. Add datasets

5. Run panel

6. Start the experiment

Live status



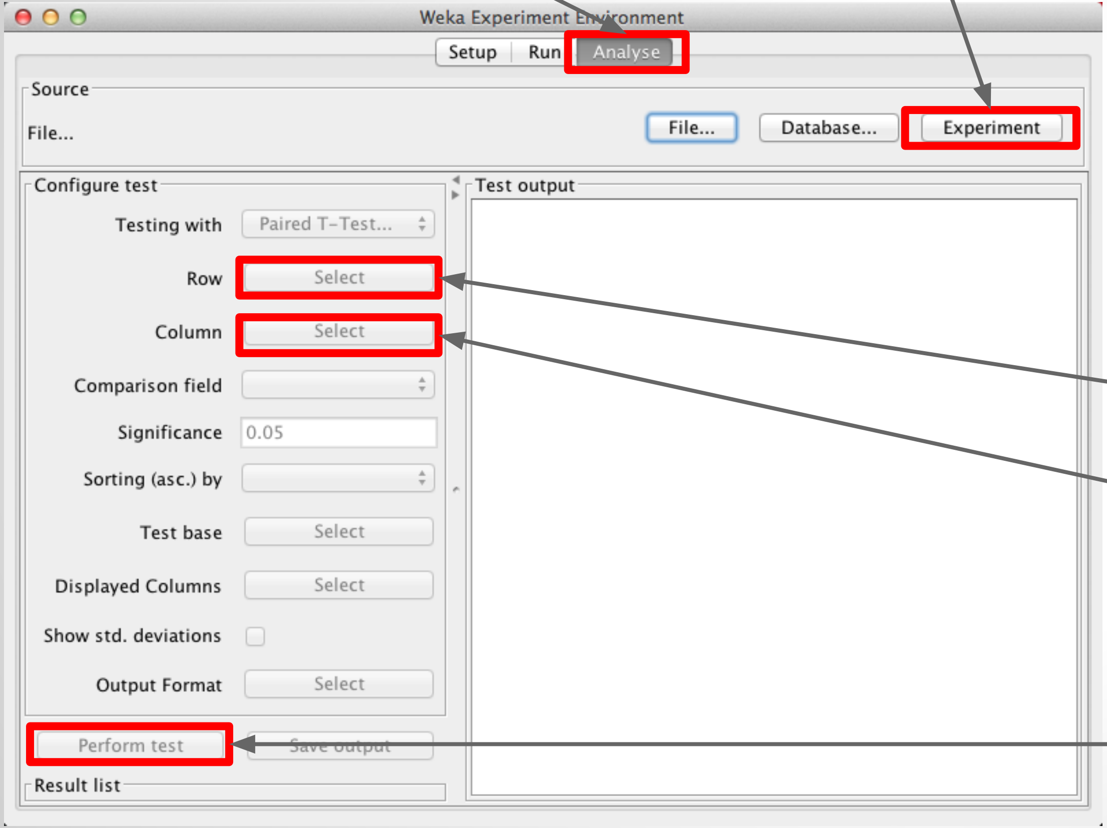
```
12:17:59: Started
12:18:00: Finished
12:18:00: There were 0 errors
```

Status

Not running

7. Analyse panel

8. Get data

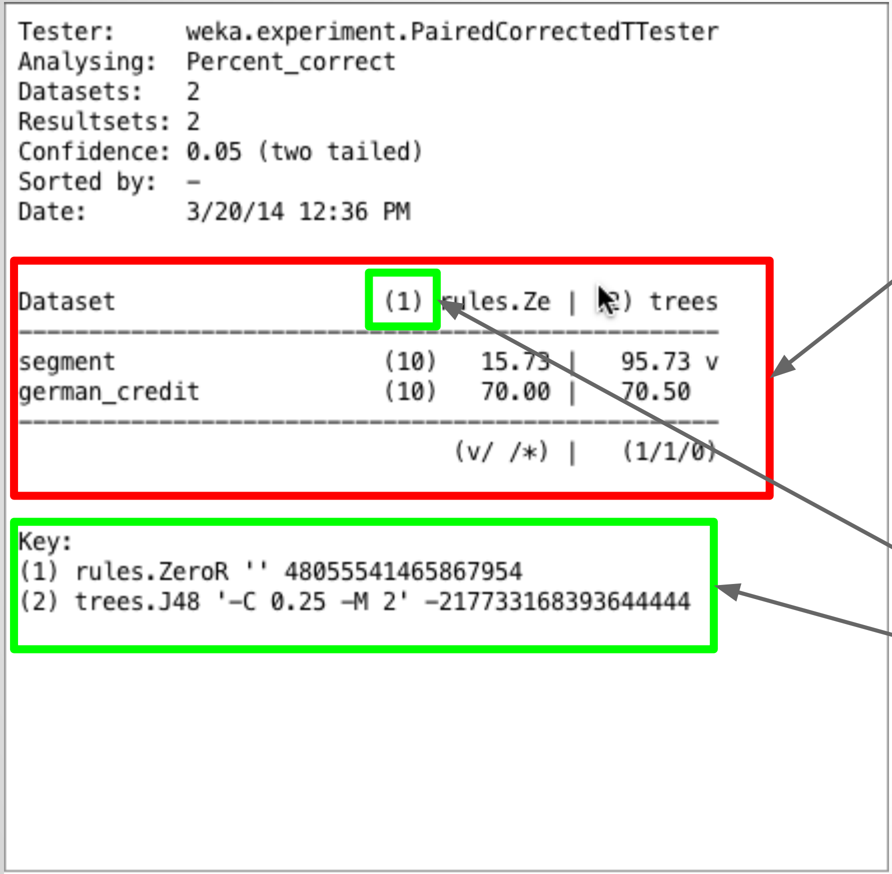


The results will be shown in a table where the lines represent each dataset, and the columns represent the percentage of correct classification for each algorithm

8. Select “datasets” for row

9. Select “Percent_corect”

10. Start the test



```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   2
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       3/20/14 12:36 PM

Dataset                   (1) rules.Ze | (2) trees
--------------------------------------------------
segment                   (10)   15.73 |   95.73 v
german_credit             (10)   70.00 |   70.50
--------------------------------------------------
                               (v/ /*) |   (1/1/0)

Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
```

Results table

Algorithm index number

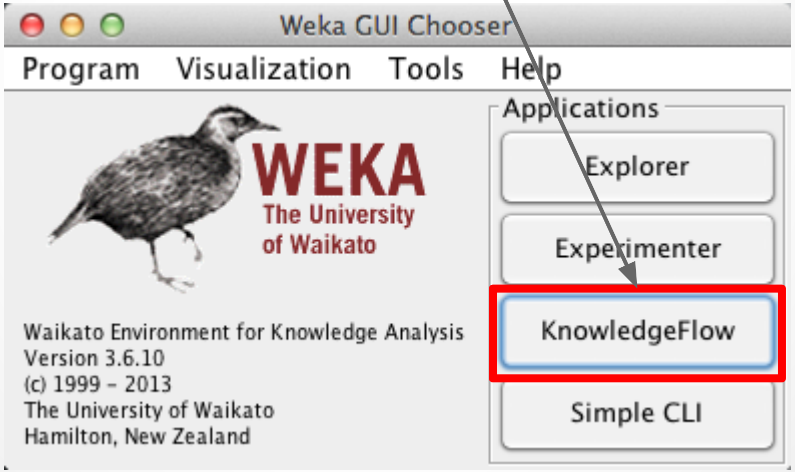
Ex. 4

compare J48 algorithm and ZeroR on the labor.
arff and glass.arff

Question 4?

How can I make Weka more configurable?

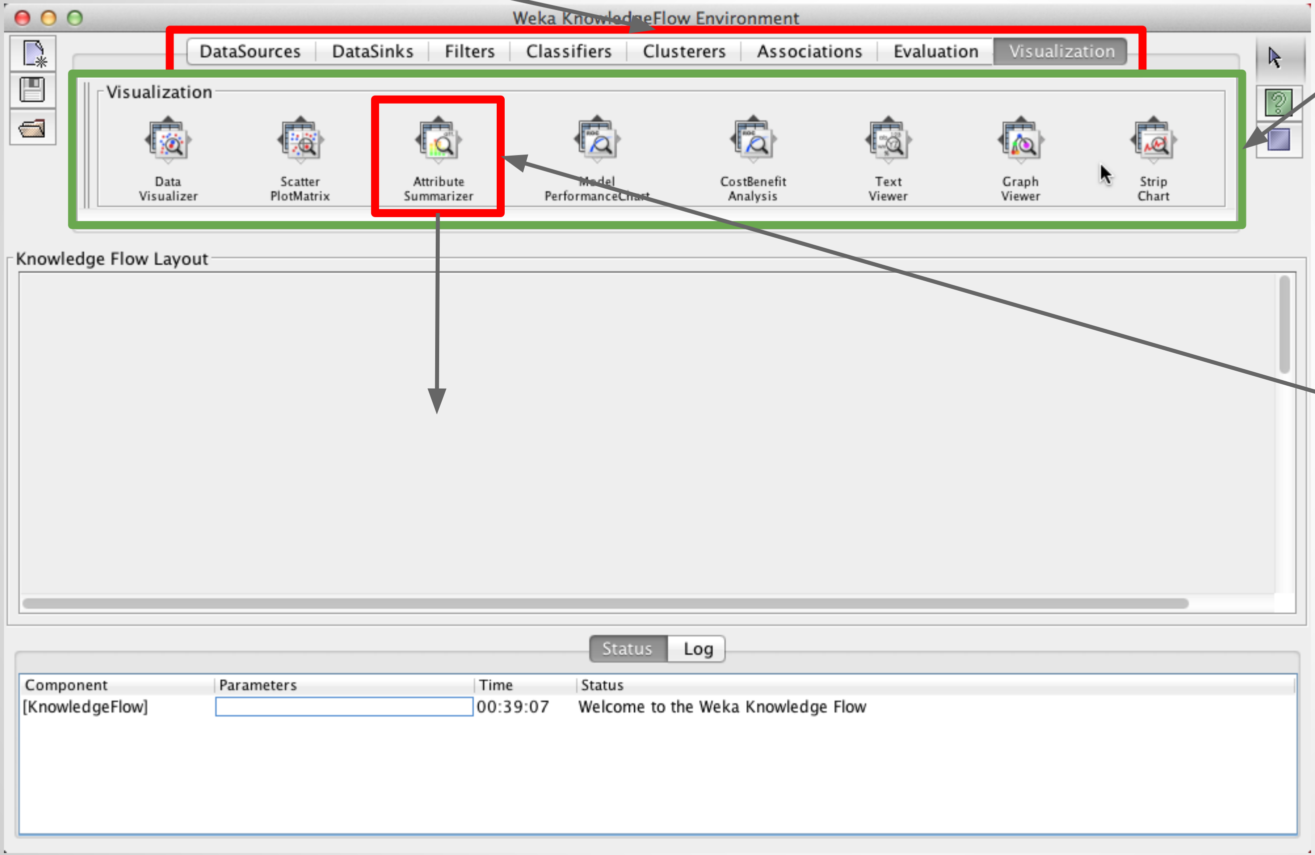
You can easily configure Weka with the use of the KnowledgeFlow



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Control panel

Options for each panel

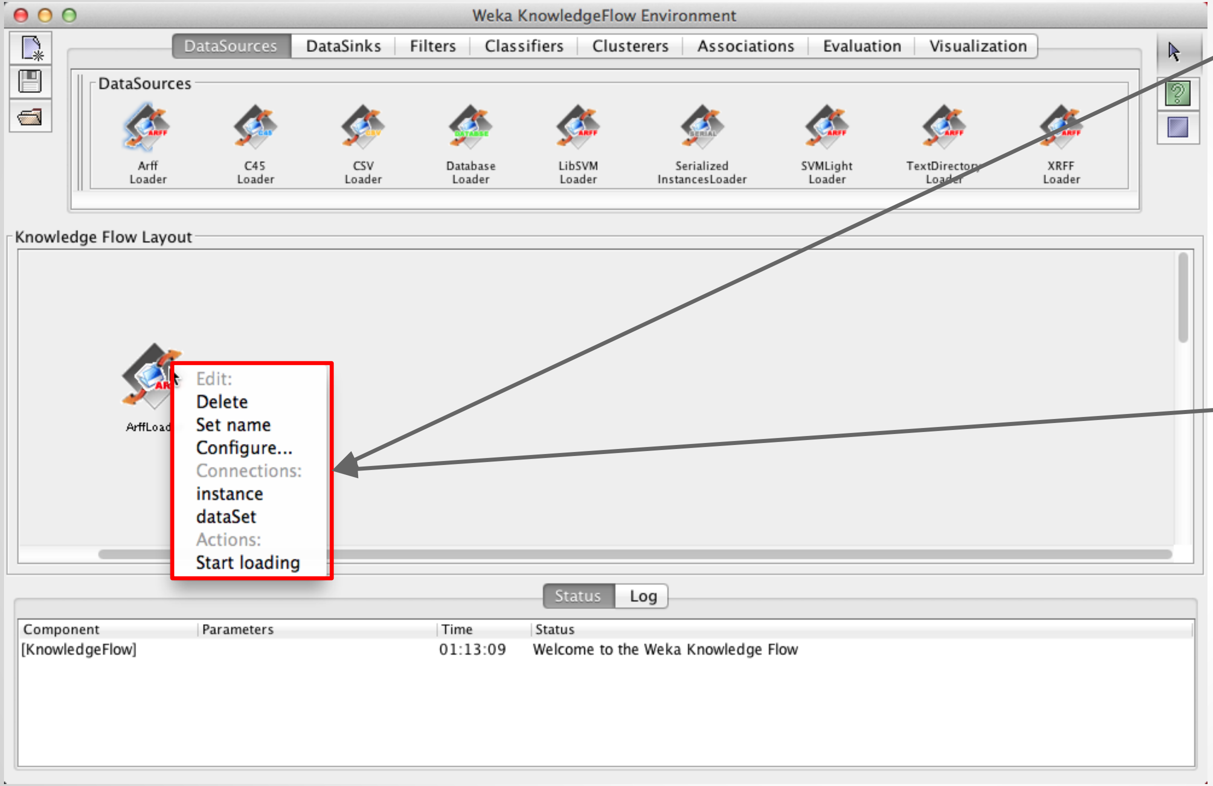


Component	Parameters	Time	Status
[KnowledgeFlow]		00:39:07	Welcome to the Weka Knowledge Flow

You import each options in the knowledge flow by clicking on it and then clicking on the Knowledge Flow Layout

1. Import arff file

- go to DataSources panel
- choose arff loader
- put the arff loader on the knowledge flow
- open diabetes.arff



Component	Parameters	Time	Status
[KnowledgeFlow]		01:13:09	Welcome to the Weka Knowledge Flow

Right click on the arff loader to get specific options.

- Configure: choose the data set.
- dataSet: export data
- Start loading: start the knowledge flow

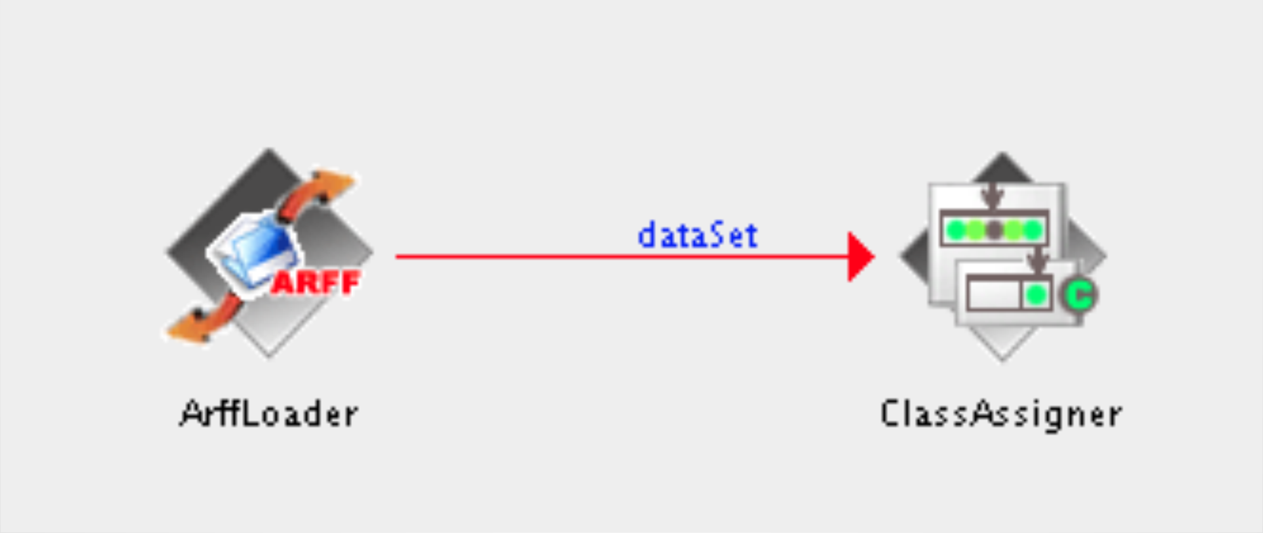
2. Set the class attribute

- go to Evaluation panel
- choose the “Class assigner” option
- put the “Class assigner” to the Knowledge Flow
- set the class of by configuring “Class assigner”

You create a link between them by:

- right clicking the ArffLoader and and choosing dataSet
- left click on the ClassAssigner

This way you take the data set and assign a class attribute to it



3. Creating a training set and a testing set

Out of the data set you have to create a subset of data to train and a subset to test the classification

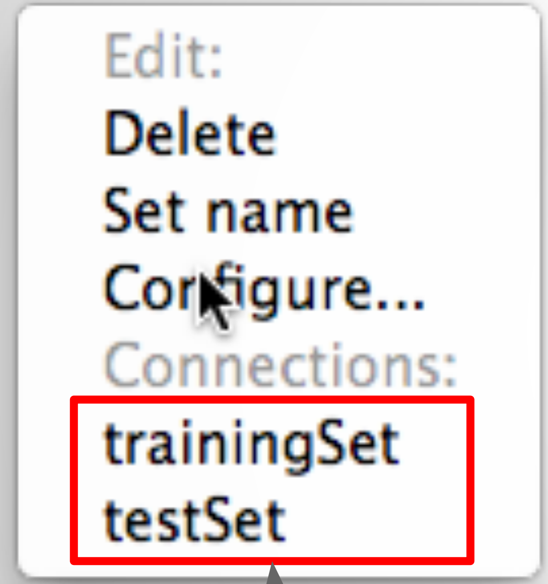
- go to Visualisation panel
- select CrossValidationFoldMaker
- put CrossValidationFoldMaker on the Knowledge flow
- link the ClassAssigner with the CrossValidationFoldMaker

Questions? Remarks?

4. Use an algorithm

- go to Classifiers panel
- choose NaiveBayes classifier
- put NaiveBayes Classifier on the Knowledge Flow

link the NaiveBayes with the CrossValidationFoldMaker



You link the CrossValidationFoldMaker with the algorithm by:

- choosing the trainingSet in the CrossValidationFoldMaker and linking it with the algorithm
- choosing the testSet in the CrossValidationFoldMaker and linking it with the algorithm

You get this options by right clicking on CrossValidationFoldMaker

5. Evaluate the algorithm

- go to Evaluation panel
- choose “Classifier Performance Evaluator”
- put the “Classifier Performance Evaluator” on the Knowledge Flow
- link the “Classifier Performance Evaluator” with the algorithm (batchClassifier option)

Ex. 5

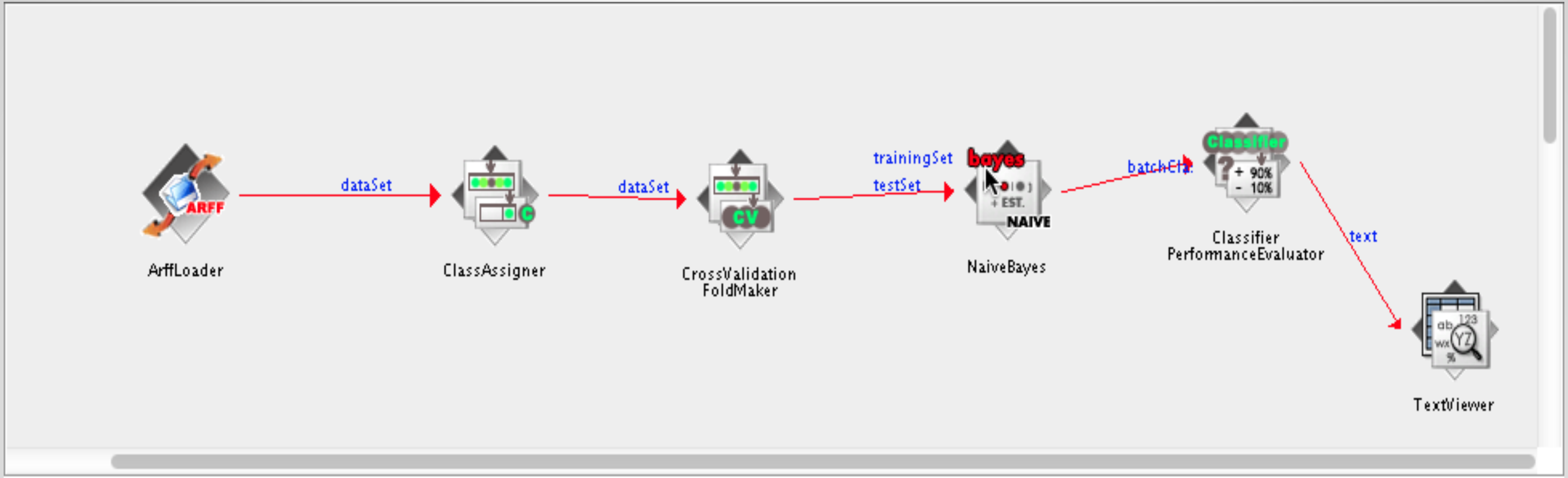
Create a knowledge flow with:

-iris.arff

-NaiveBayes algorithm

6. Visualise results

- go to Visualisation panel
- choose “Text Viewer”
- put “Text Viewer” on the Knowledge Flow
- link “Text Viewer with”Classifier Performance Evaluator (text option)



- start the knowledge flow: go to ArffLoader and choose “Start Loading”
- view the results: go at “TextViewer” and choose “Show Results”