AN INTRODUCTION TO WEKA

JOHN KUNDERT-GIBBS

WHAT IS WEKA?

(HINT: NOT A BIRD!)

WHAT IS WEKA?

- A data mining suite with a sizable collection of supervised and unsupervised algorithms
- Freeware!
- Methods to statistically compare the success of different algorithms
- An easy GUI mode
- Java-Based
- Command-line access as well
- http://www.cs.waikato.ac.nz/ml/weka/

MAJOR INTERFACE ELEMENTS

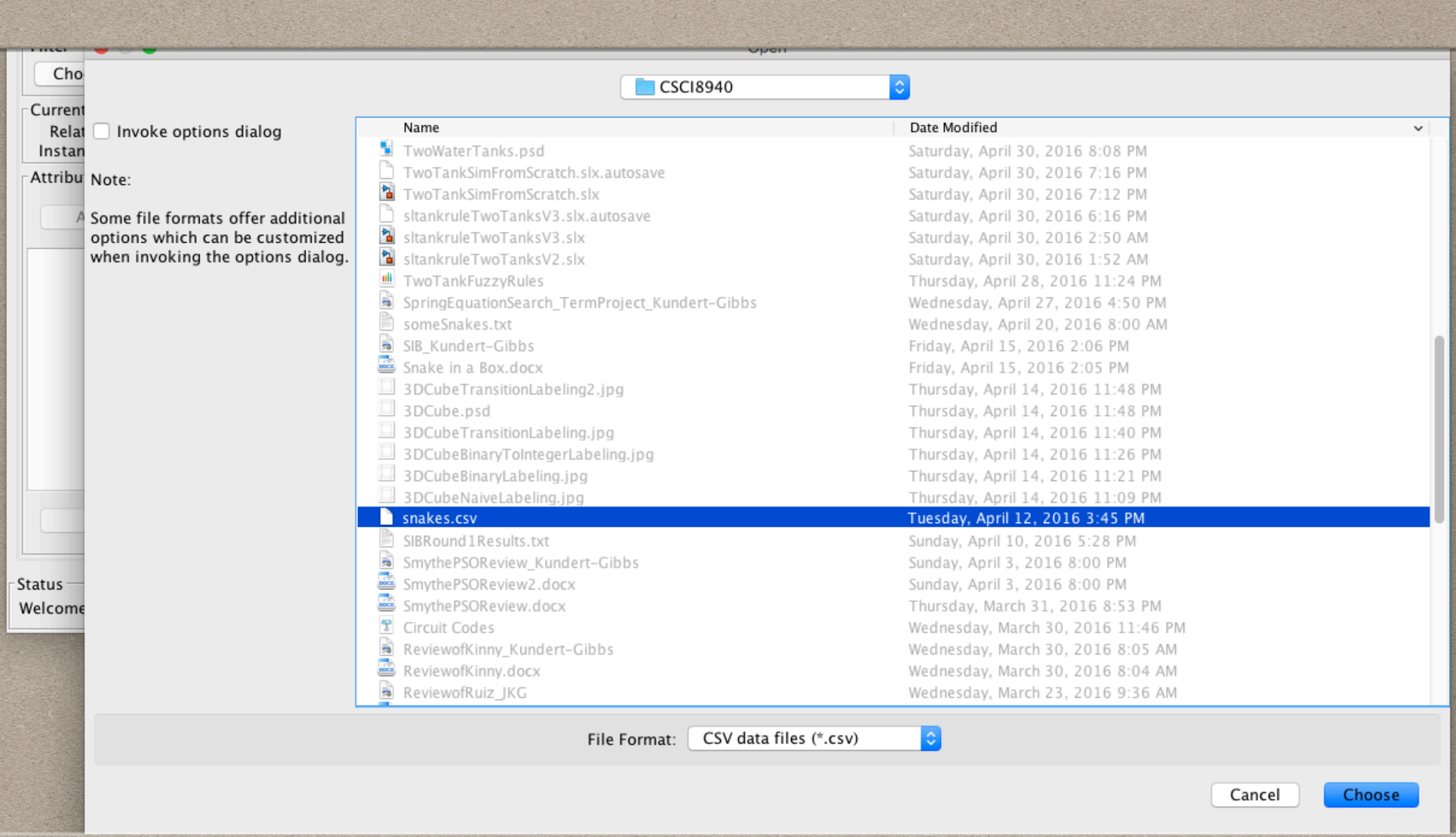
- Explorer
- Experimenter
- KnowledgeFlow
- Simple CLI

ARFF (ATTRIBUTE-RELATION FILE FORMAT)

- Weka's native data representation/file format
- Three elements
  - @relation
  - @attribute
  - @data

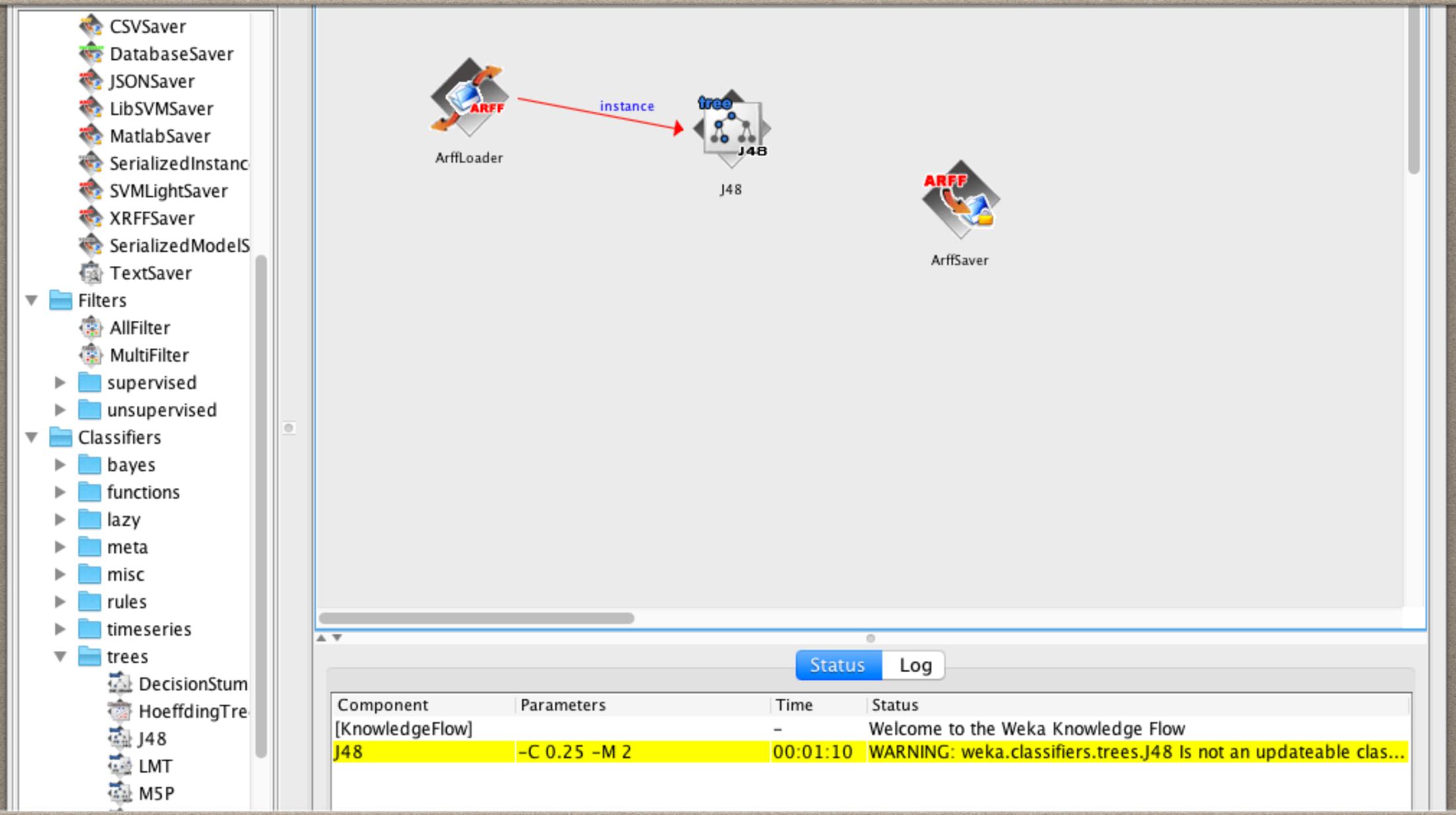
ARFF EXAMPLE

```
@Relation iris

@Attribute sepallength    Real
@Attribute sepalwidth     Real
@Attribute petallength    short, medium, long
@Attribute petalwidth     Real
@Attribute class     {Iris-setosa,Iris-versicolor,Iris-virginica}

@Data
5.1,3.5,medium,0.2,Iris-setosa
4.9,3.0,medium,0.2,Iris-setosa
4.7,3.2,medium,0.2,Iris-setosa
4.6,3.1,long,0.2,Iris-setosa
```



OTHER FILE FORMATS

E.G., .CSV



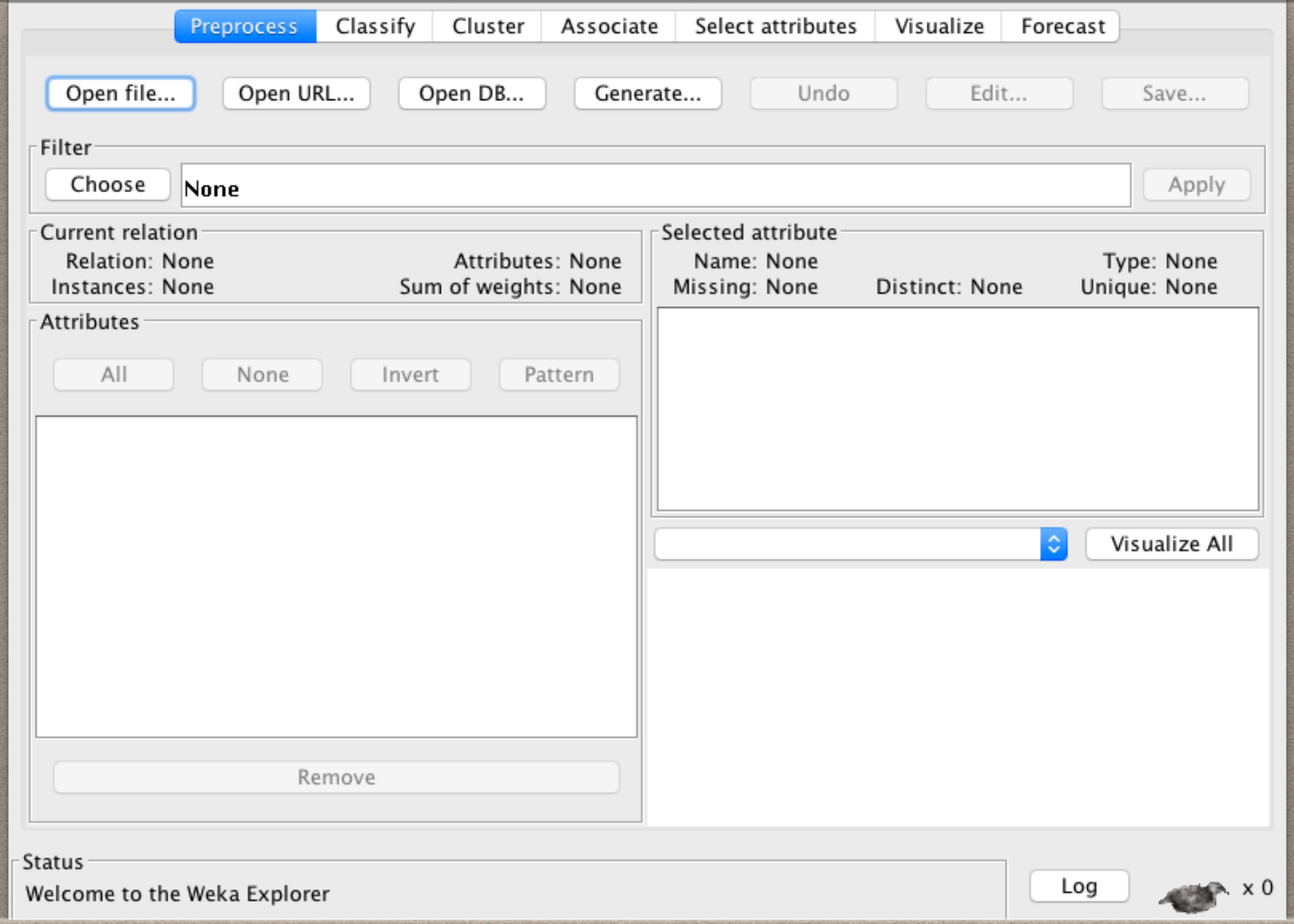
KNOWLEDGEFLOW

QUICK OVERVIEW

```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
          java <classname> <args> [ > file]
          break
          kill
          capabilities <classname> <args>
          cls
          history
          exit
          help <command>
```

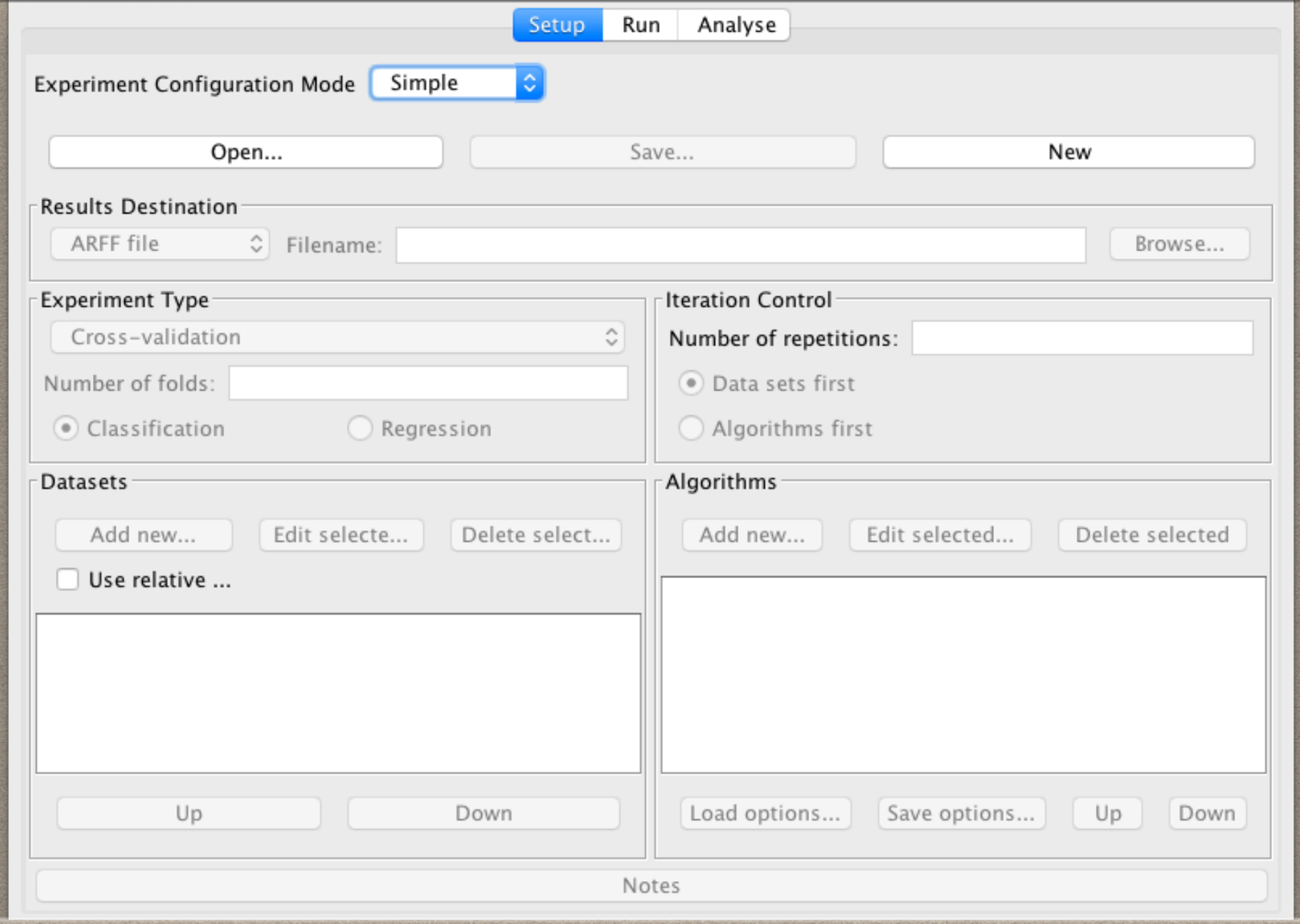
SIMPLE CLI

SIMPLE COMMAND LINE INTERFACE
QUICK OVERVIEW



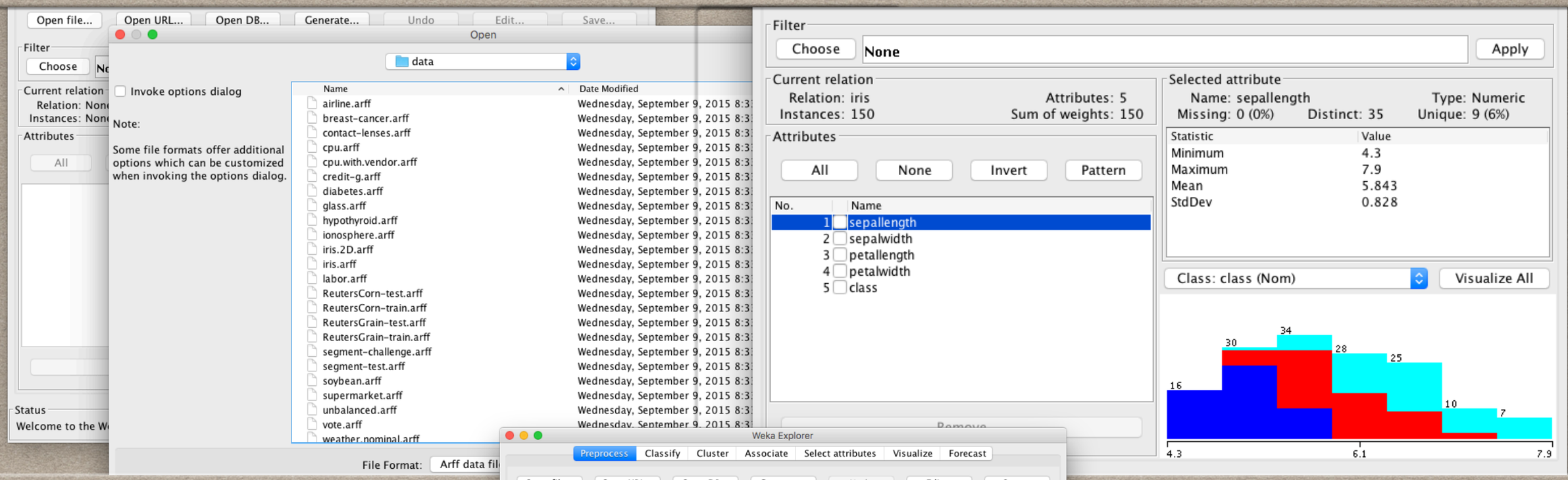
EXPLORER

THE MAIN WEKA INTERFACE



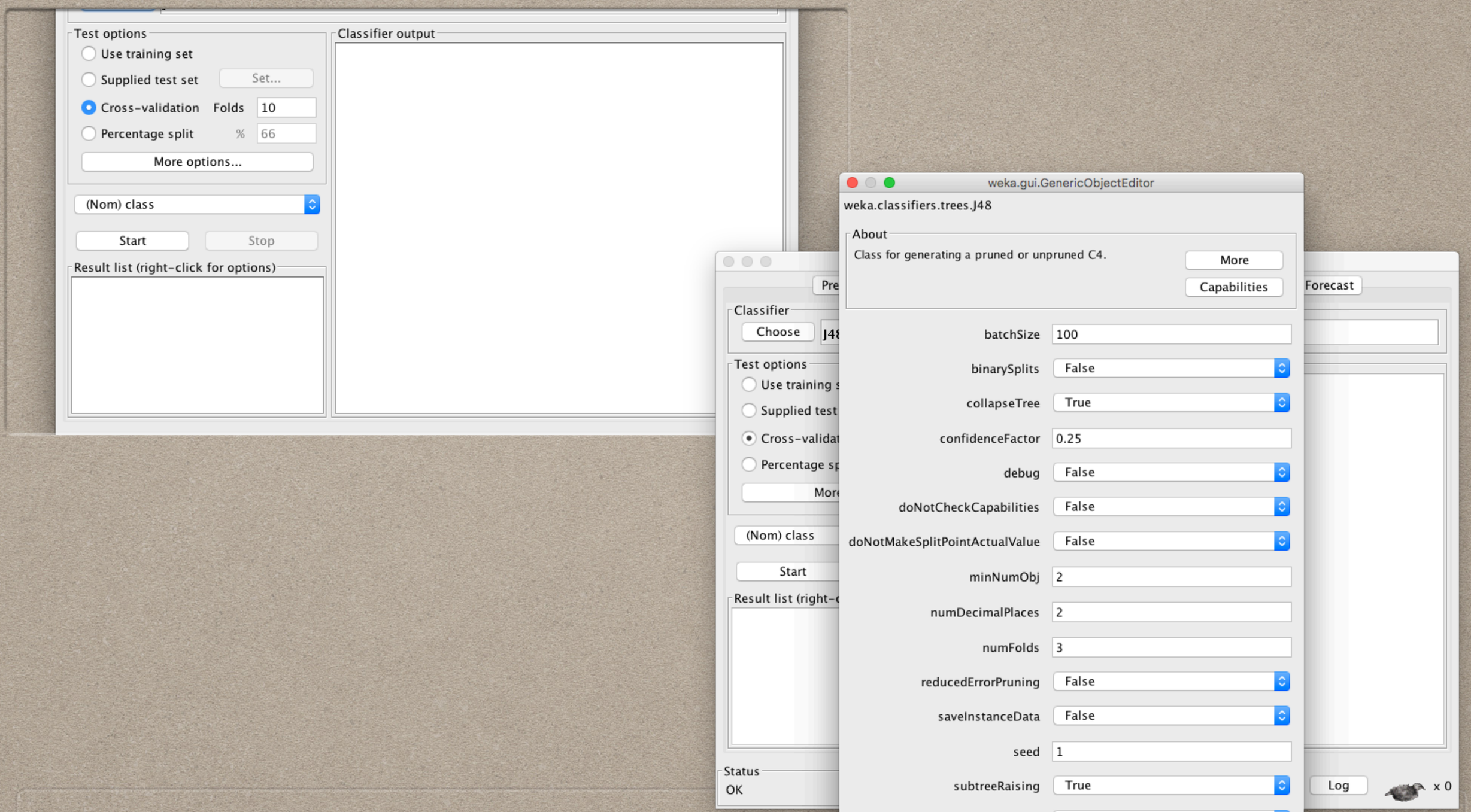
EXPERIMENTER

RUN TESTS BETWEEN ML ALGORITHMS



EXPLORER

DATA FILE



EXPLORER

CLASSIFY

Weka Explorer

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

Classifier

Choose: J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds: 10
- Percentage split — %: 66

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

23:13:03 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Coverage of cases (0.95 level)          96.6667 %
Mean rel. region size (0.95 level)      33.7778 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.980    0.000    1.000      0.980    0.990      0.985    0.990     0.987     Iris-setosa
               0.940    0.030    0.940      0.940    0.940      0.910    0.952     0.880     Iris-versicolor
               0.960    0.030    0.941      0.960    0.950      0.925    0.961     0.905     Iris-virginica
Weighted Avg.  0.960    0.020    0.960      0.960    0.960      0.940    0.968     0.924

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

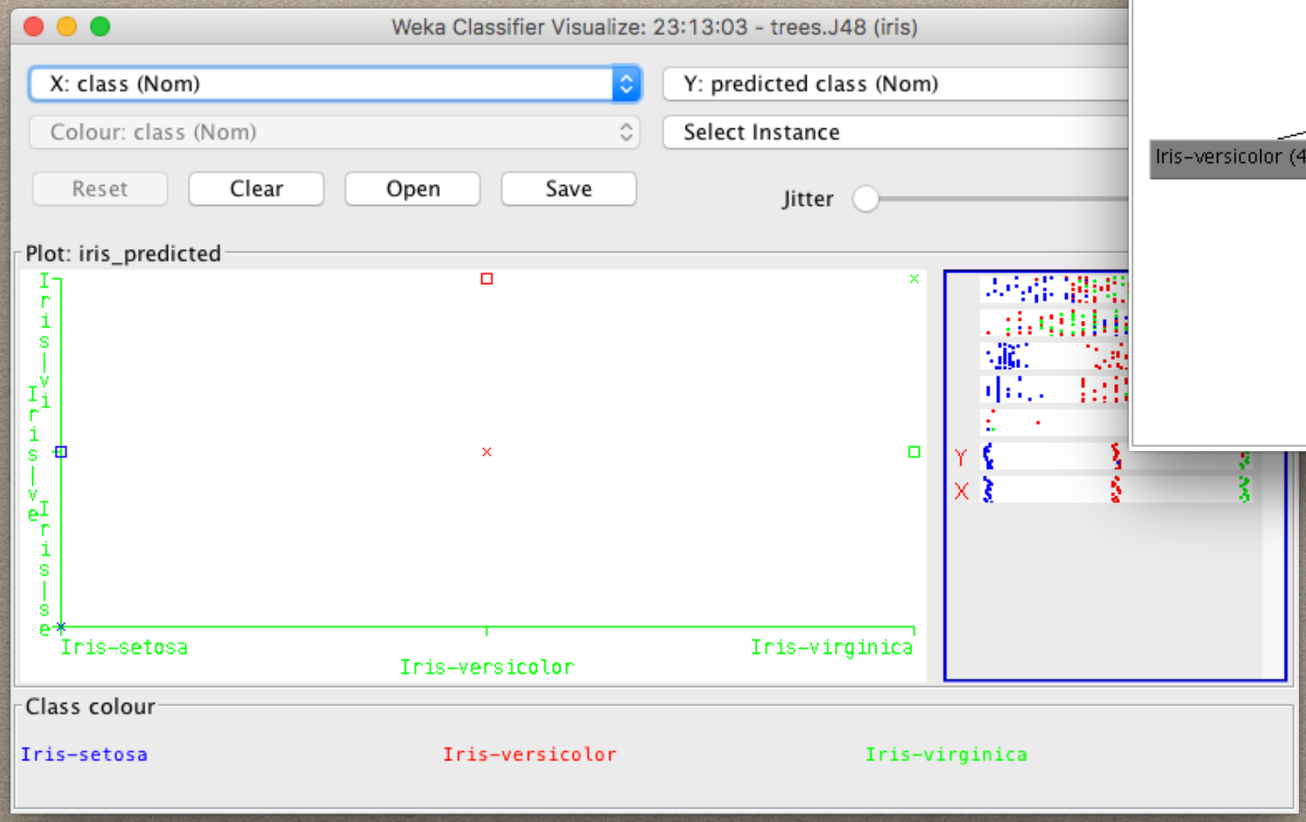
Status: OK

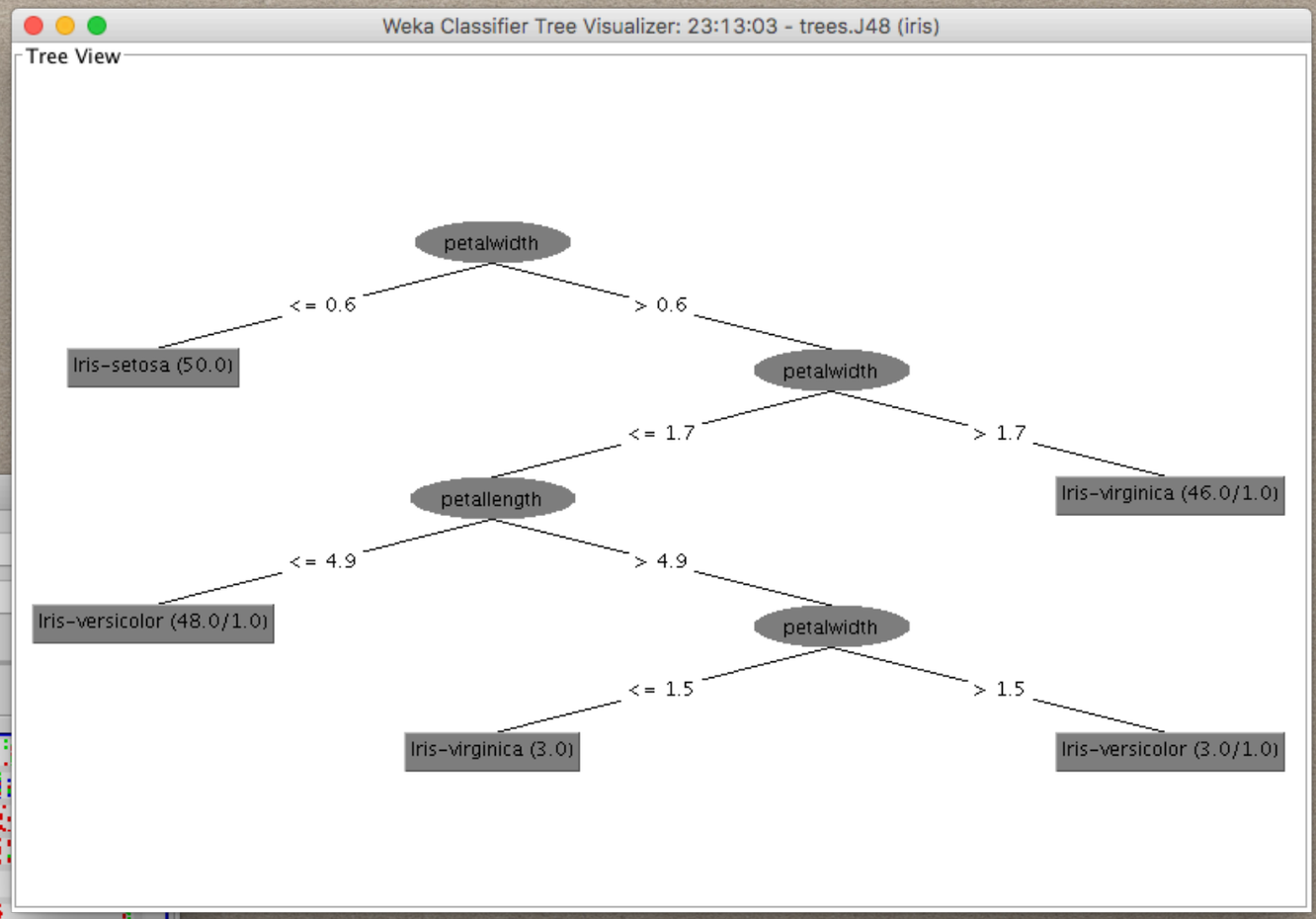
Log | x 0


```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```


EXPLORER

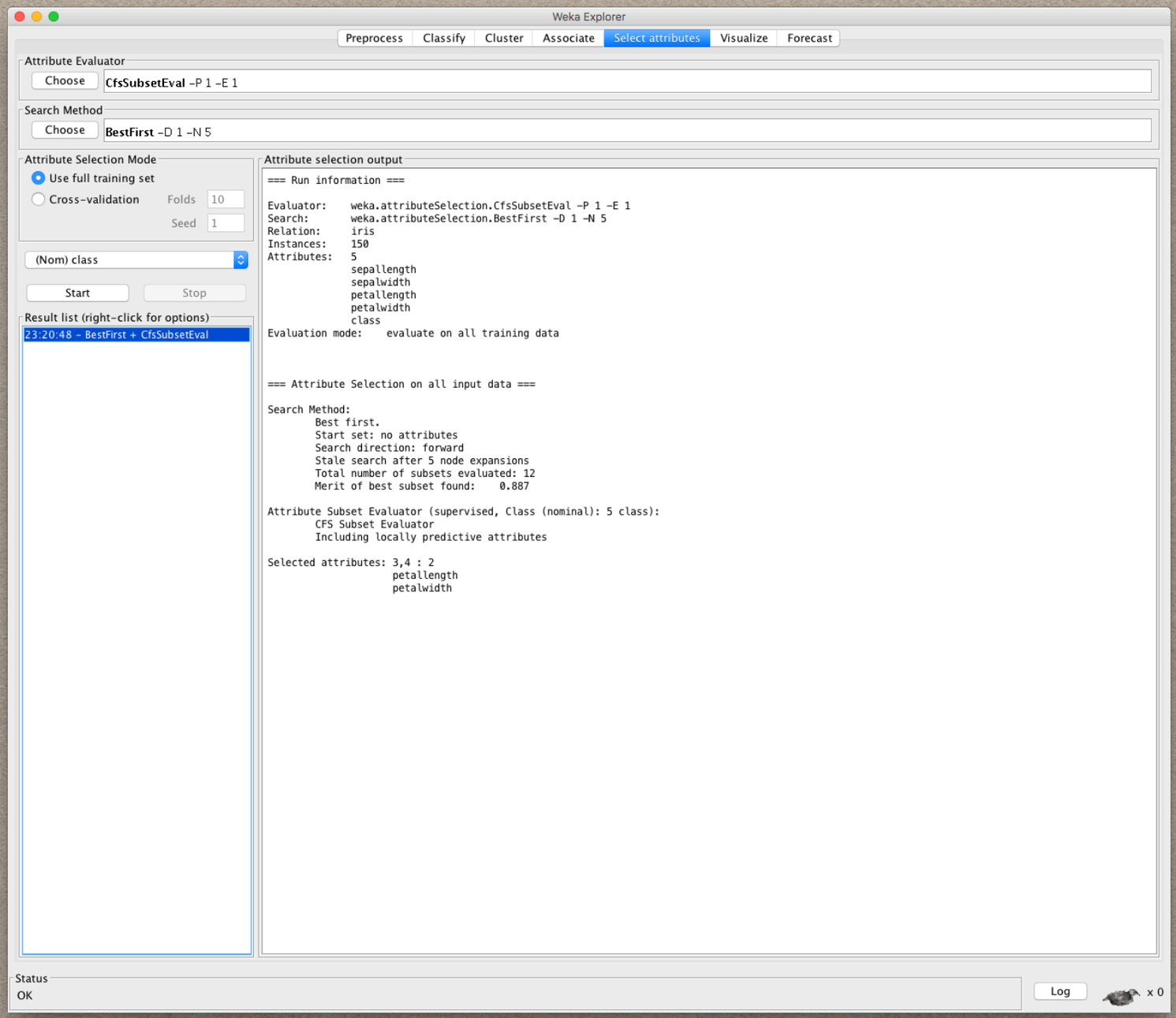
RESULTS





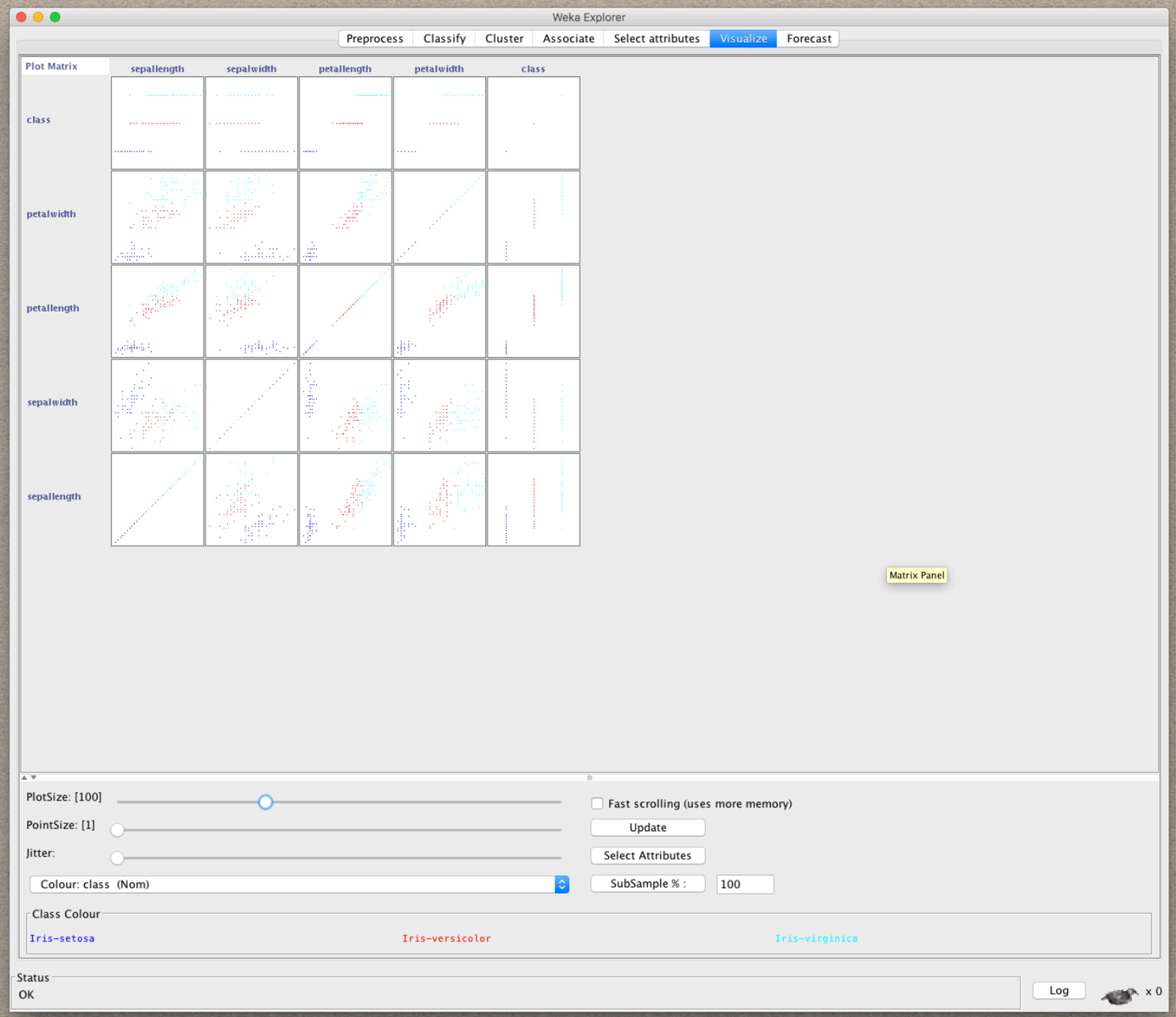
EXPLORER

VISUALIZE



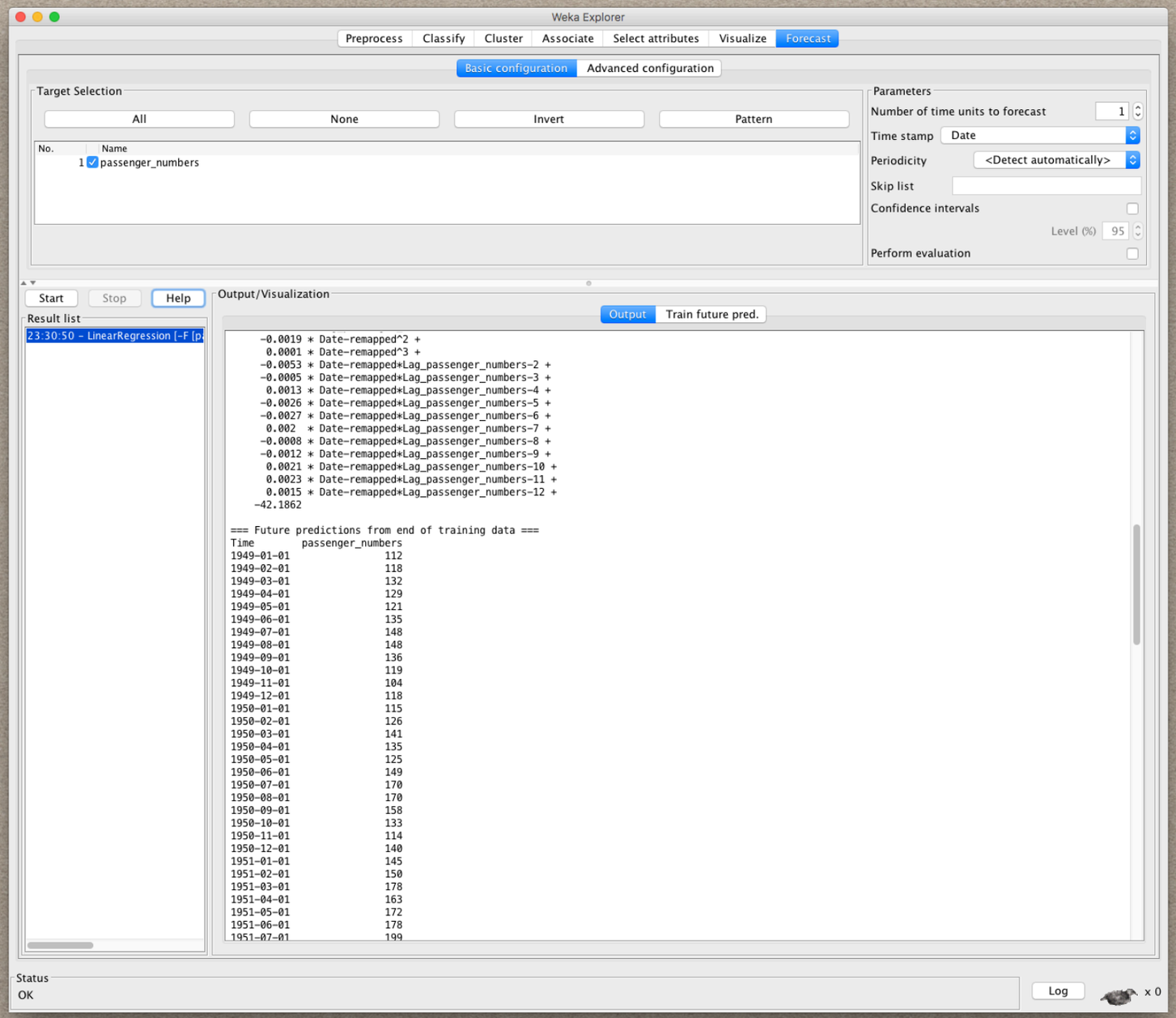
EXPLORER

ATTRIBUTE SELECTION



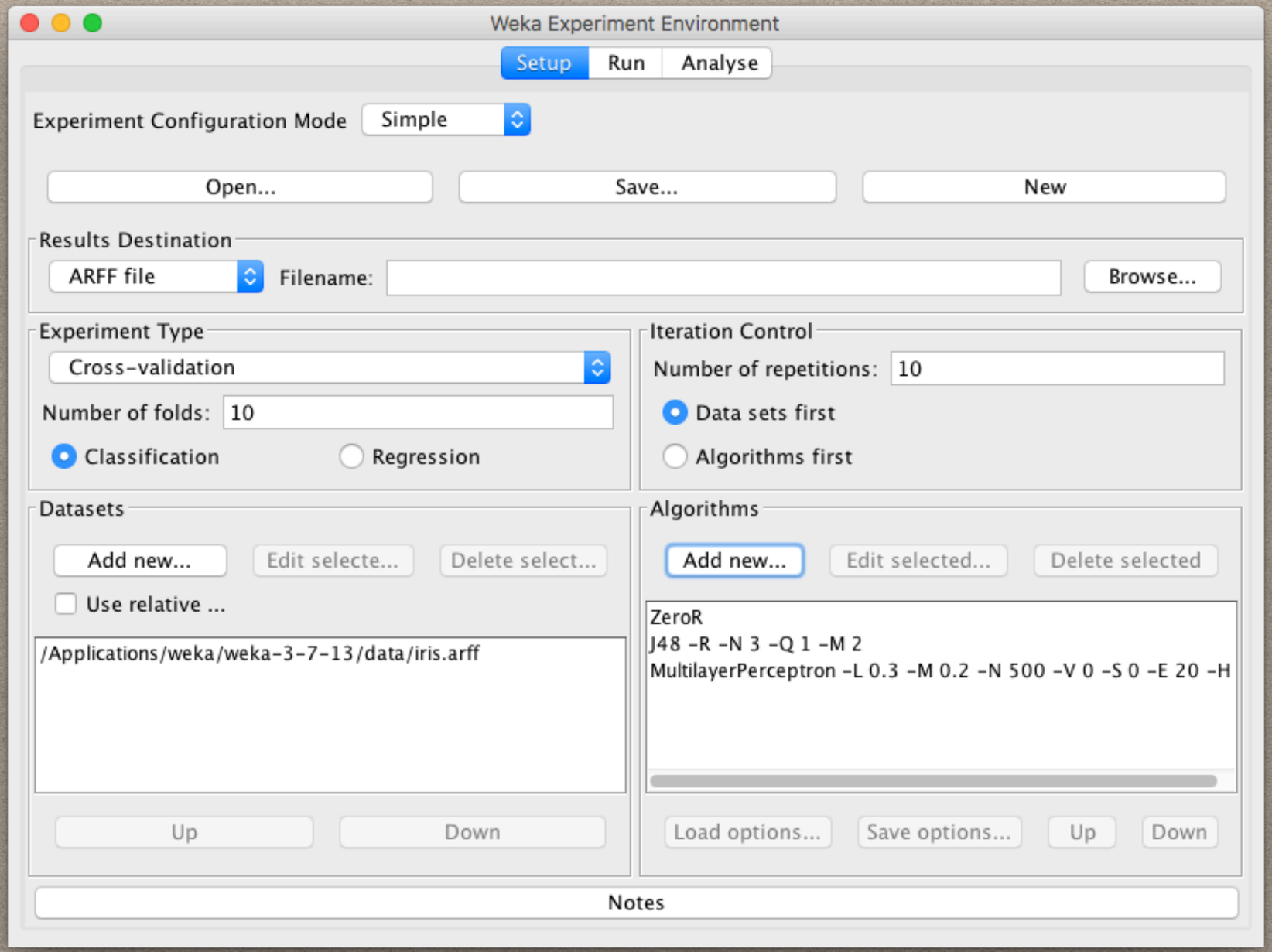
EXPLORER

VISUALIZE ATTRIBUTES



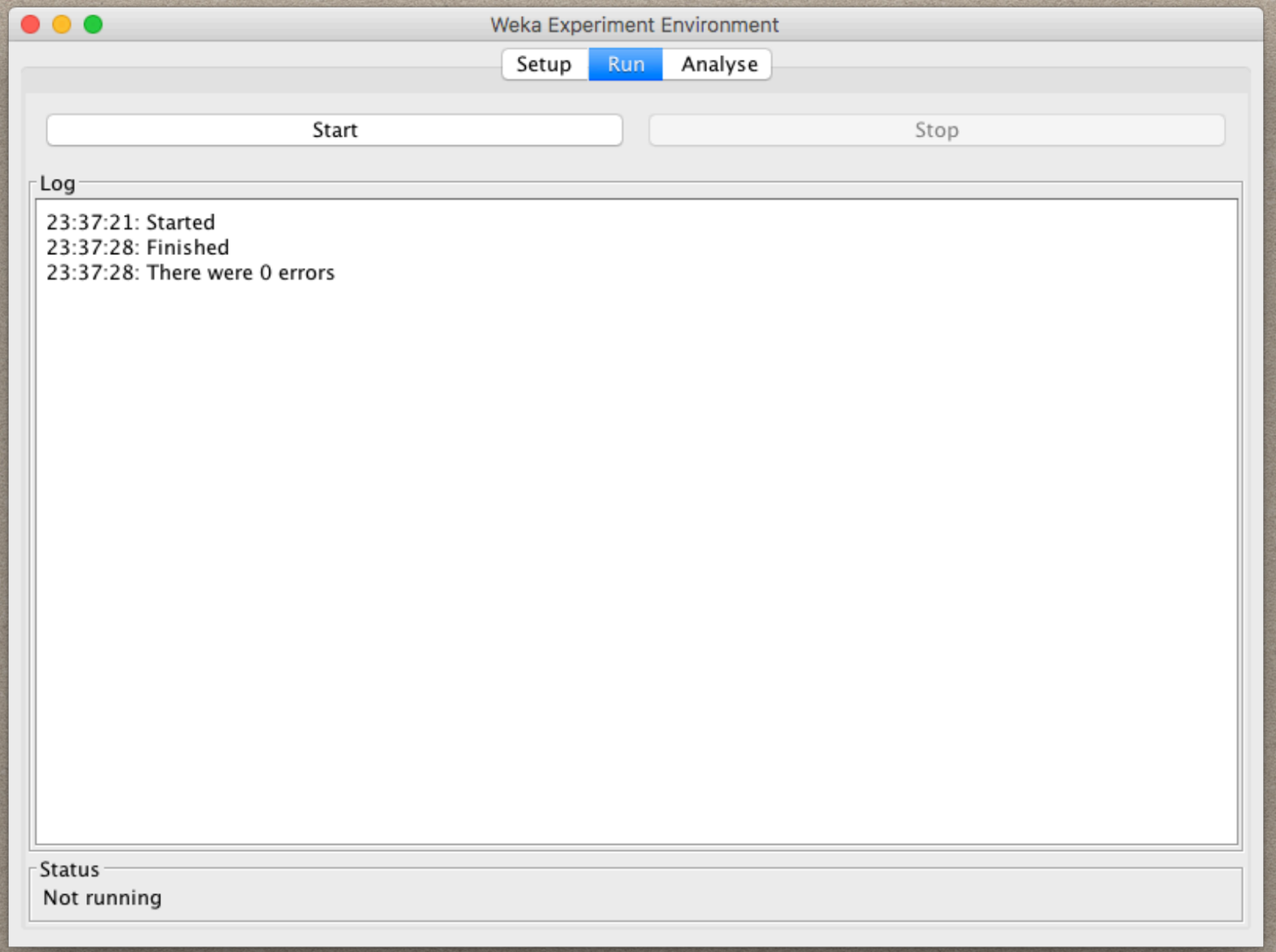
EXPLORER

FORECASTING



EXPERIMENTER

SETUP



EXPERIMENTER

RUN

Weka Experiment Environment

Setup | Run | Analyse

Source

Got 300 results

File... | Database... | Experiment

Actions

Perform test | Save output | Open Explorer...

Configure test

Testing with: Paired T-Tester (correc...

Select rows and cols: Rows | Cols | Swap

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations:

Output Format: Select

Result list

23:37:39 - Available resultsets

23:37:59 - Percent_correct - rules.ZeroR '' 4805554146586

Test output

```
Tester:     weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMa
Analysing:  Percent_correct
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       1/31/17 11:37 PM


Dataset                   (1) rules.Ze | (2) trees (3) funct
----------------------------------------------------------------
iris                     (100)   33.33 |   93.87 v   96.93 v
----------------------------------------------------------------
                               (v/ /*) |   (1/0/0)   (1/0/0)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-R -N 3 -Q 1 -M 2' -217733168393644444
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
```

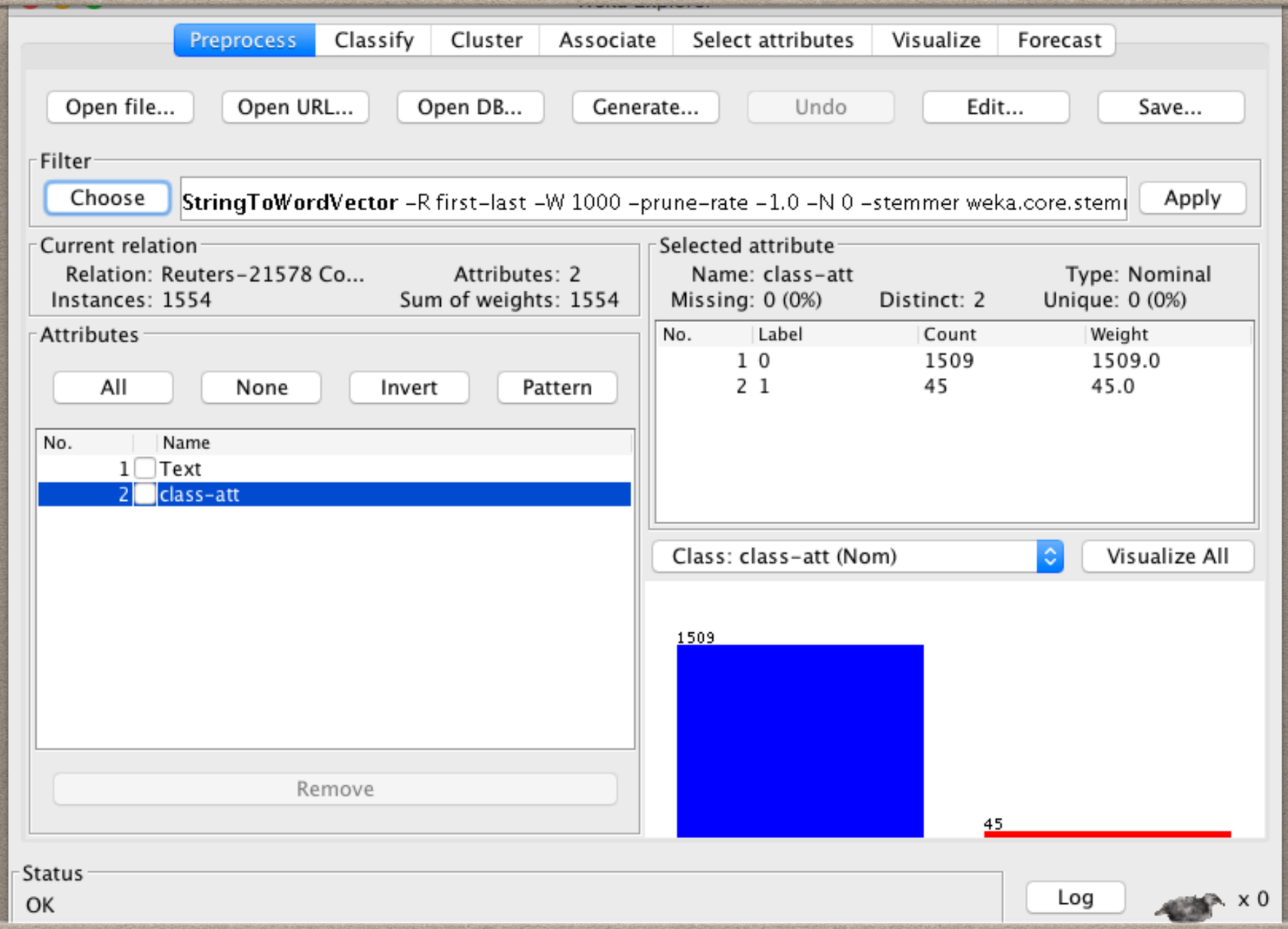
```
Dataset                   (2) trees.J4 | (1) rules (3) funct
----------------------------------------------------------------
iris                     (100)   93.87 |   33.33 *   96.93
----------------------------------------------------------------
                               (v/ /*) |   (0/0/1)   (0/1/0)


Key:
(1) rules.ZeroR '' 48055541465867954
(2) trees.J48 '-R -N 3 -Q 1 -M 2' -217733168393644444
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
```

EXPERIMENTER

ANALYZE

SUPERVISED LEARNING METHODS

- Decision Trees
- Bayes (Naive and Nets)
- Logistic
- Regression
- Support Vector Machine
- Multilayer Perceptron
- K Nearest Neighbors
- Boosting
- Bagging
- etc!



TEXT

STRINGTOWORDVECTOR

DEMO: J48 AND MULTILAYER PERCEPTRON

QUESTIONS?

THANK YOU!