

WEKA

Introduction

# WEKA software

- ◉ WEKA homepage

- > <http://www.cs.waikato.ac.nz/ml/weka/>



# WEKA software

- ◉ Data mining software written in Java (distributed under the GNU Public License).
- ◉ Used for research, education, and applications.
- ◉ Comprehensive set of data preprocessing tools, learning algorithms and evaluation methods.
- ◉ Graphical user interfaces – including data visualization.
- ◉ Environment for comparing learning algorithms.
- ◉ Book and development versions.

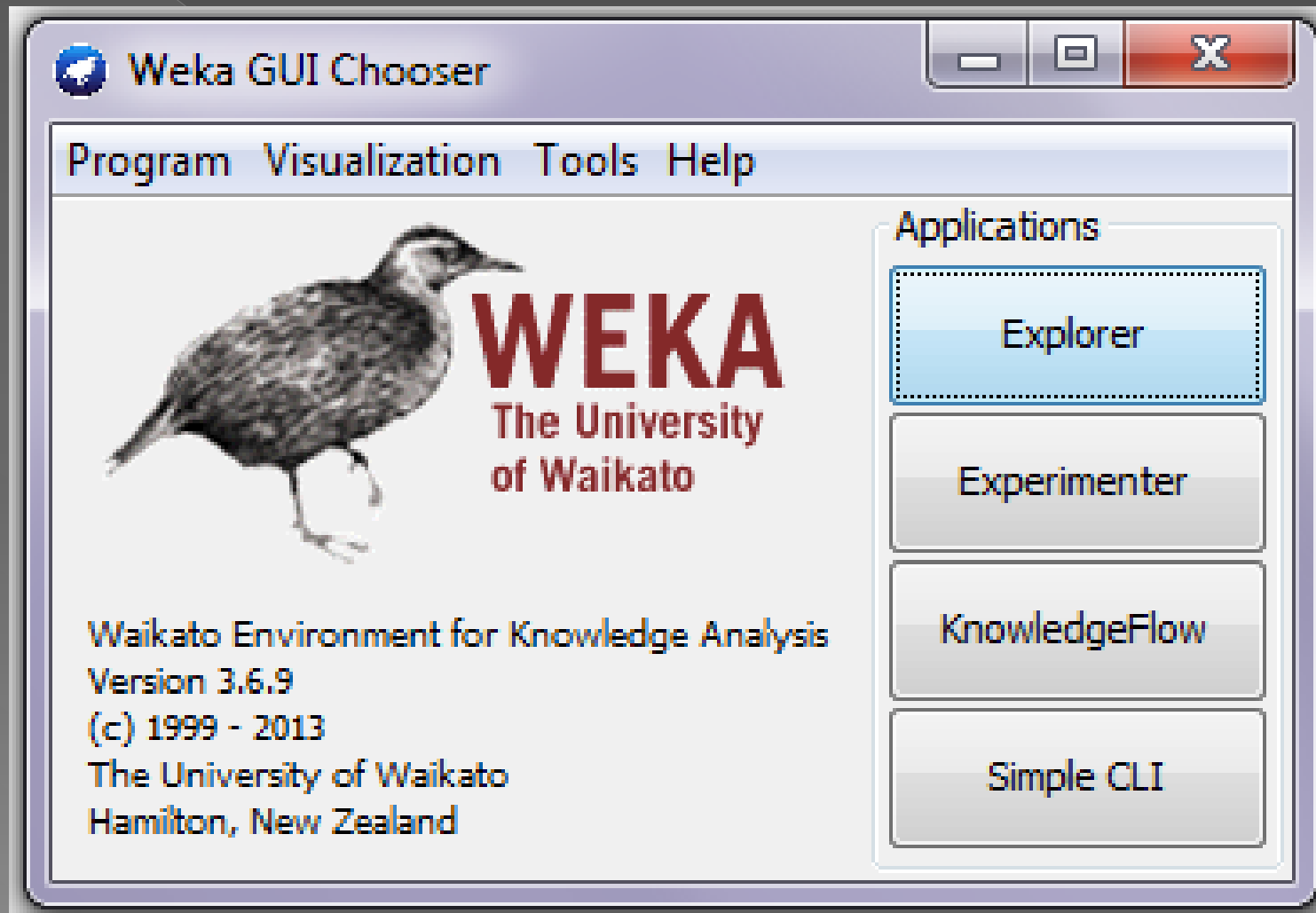
# ARFF format file

- ◉ @relation data-relation-name
- ◉ @attribute attribute1 real
- ◉ @attribute attribute2 { nominal1, nominal2 }
- ◉ @attribute class { class1, class2 }
  
- ◉ @data
- ◉ 1, nominal1, class1
- ◉ 2, nominal2, class2
- ◉ 3, ?, class1
- ◉ 4, nominal1, class2

# WEKA GUI Chooser

- The Weka GUI Chooser (class `weka.gui.GUIChooser`) provides a starting point for launching Weka's main GUI applications and supporting tools. If one prefers a MDI ("multiple document interface") appearance, then this is provided by an alternative launcher called "Main" (class `weka.gui.Main`).
- The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.

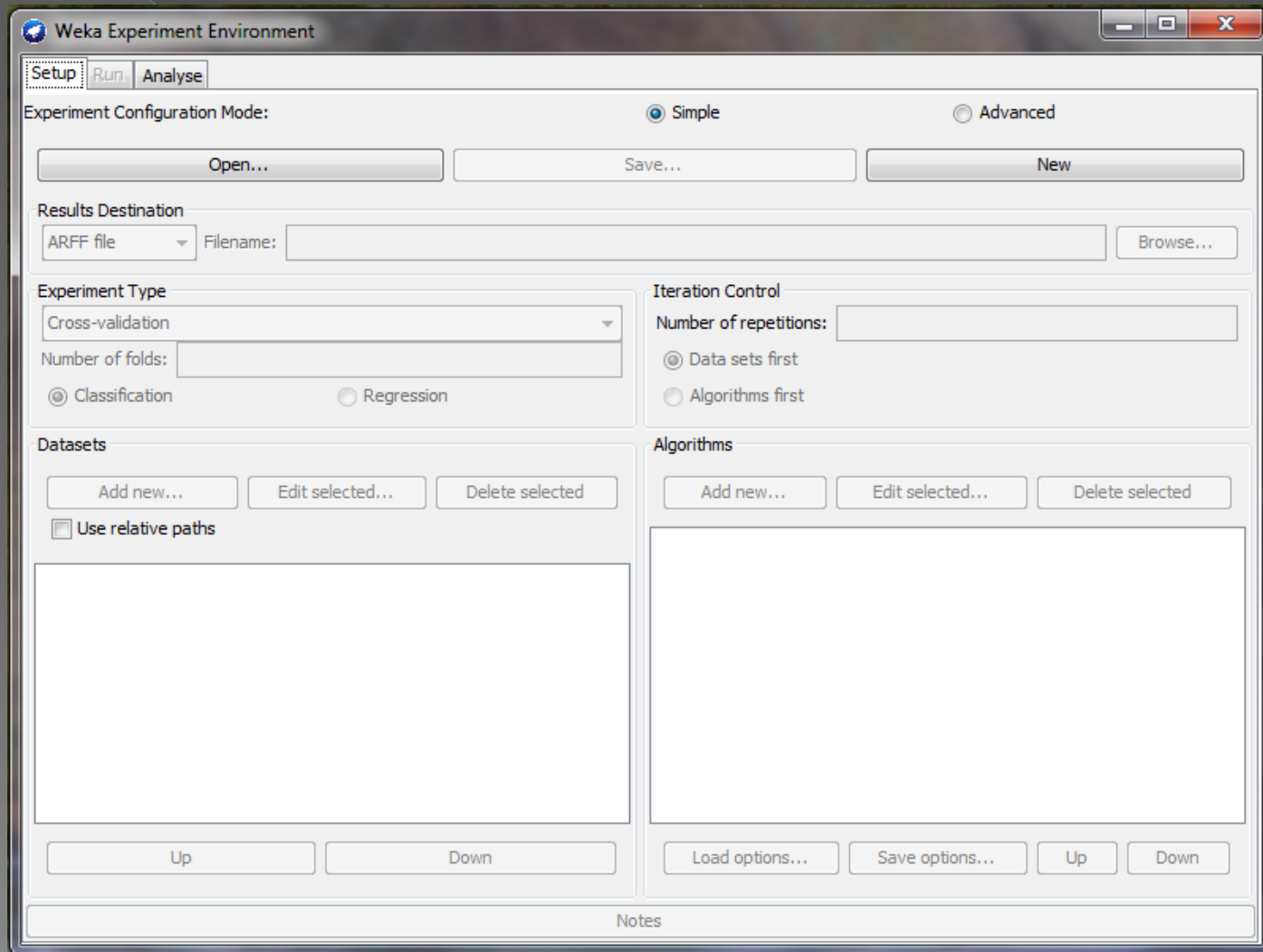
# WEKA GUI Chooser



# WEKA Experimenter

- An environment for performing experiments and conducting statistical tests between learning schemes.
- The Weka Experiment Environment enables the user to create, run, modify, and analyse experiments in a more convenient manner than is possible when processing the schemes individually. For example, the user can create an experiment that runs several schemes against a series of datasets and then analyse the results to determine if one of the schemes is (statistically) better than the other schemes.

# WEKA Experimenter





# WEKA Knowledge Flow

- This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
- The KnowledgeFlow presents a data-flow inspired interface to WEKA. The user can select WEKA components from a tool bar, place them on a layout canvas and connect them together in order to form a knowledge flow for processing and analyzing data. At present, all of WEKA's classifiers, filters, clusterers, loaders and savers are available in the KnowledgeFlow along with some extra tools.

# WEKA Knowledge Flow

Weka KnowledgeFlow Environment

DataSources | DataSinks | Filters | Classifiers | Clusterers | Associations | Evaluation | Visualization

DataSources

- Arff Loader
- C45 Loader
- CSV Loader
- Database Loader
- LibSVM Loader
- Serialized InstancesLoader
- SVMLight Loader
- TextDirectory Loader
- XRFF Loader

Knowledge Flow Layout

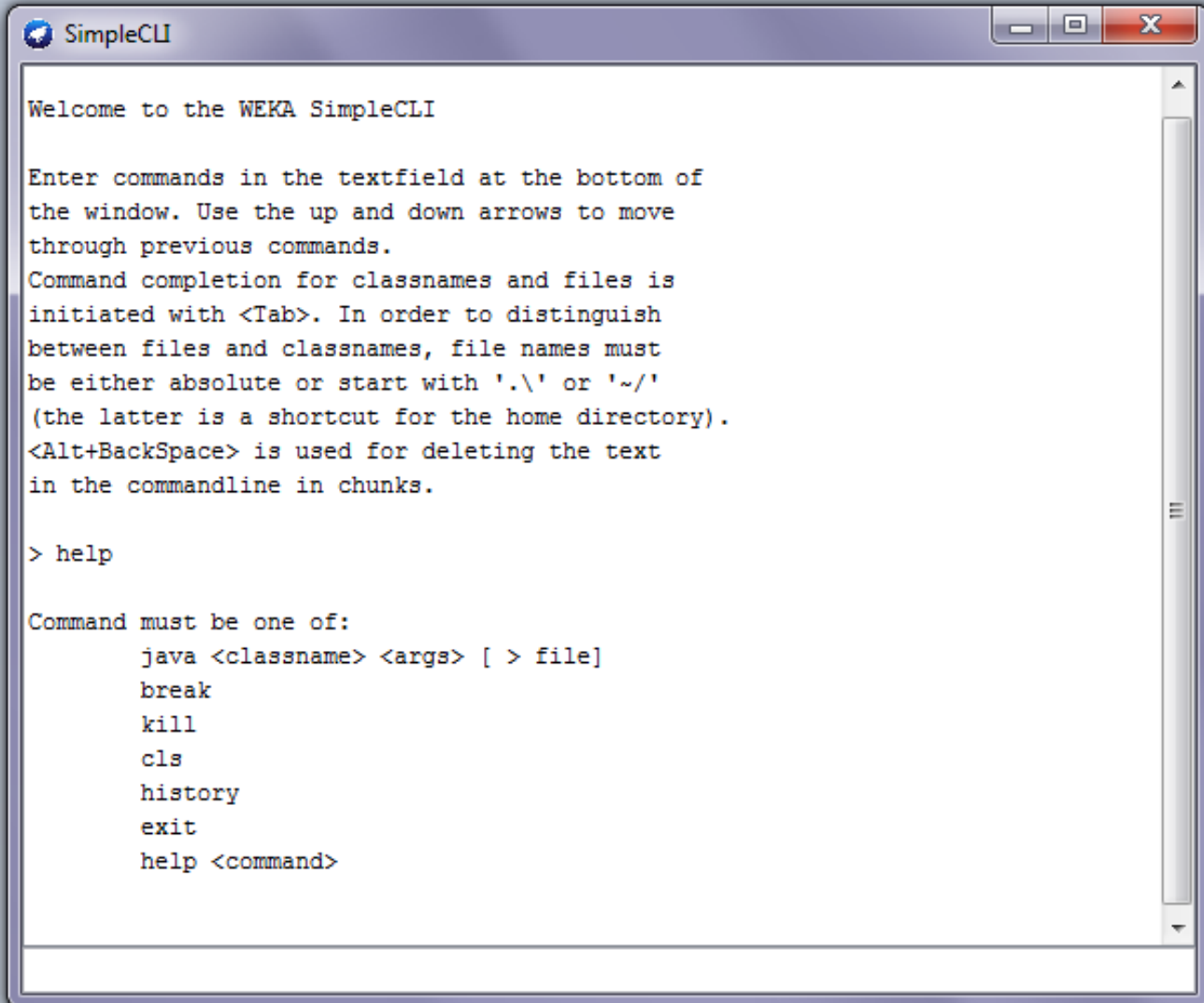
Status Log

Component	Parameters	Time	Status
[KnowledgeFlow]		0:0:13	Welcome to the Weka Knowledge Flow

# WEKA Simple CLI

- Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.
- The Simple CLI provides full access to all Weka classes, i.e., classifiers, filters, clusterers, etc., but without the hassle of the CLASSPATH (it facilitates the one, with which Weka was started).
- It offers a simple Weka shell with separated commandline and output.

# WEKA Simple CLI



```
SimpleCLI

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.

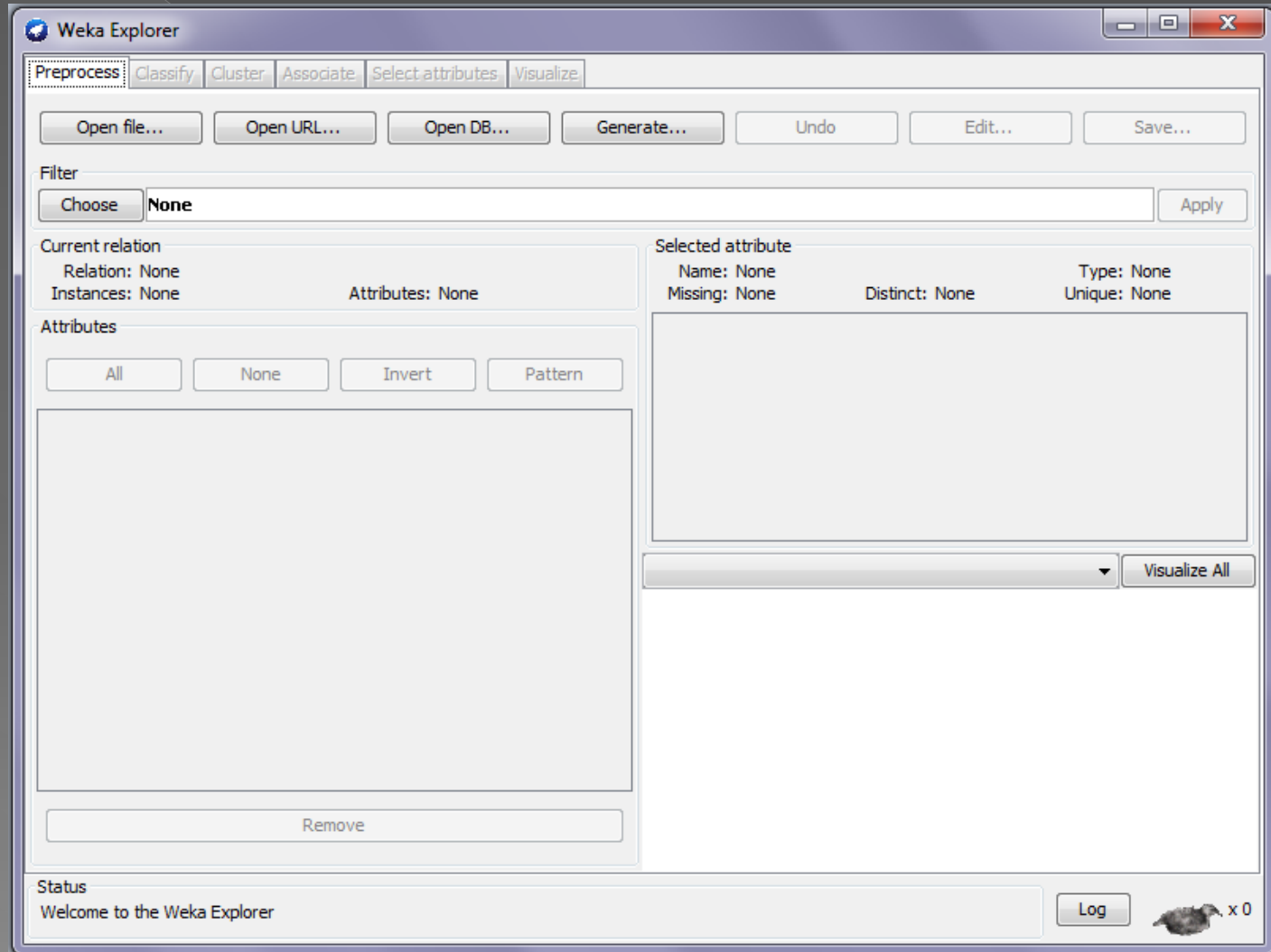
> help

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

# WEKA Explorer

- An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail).
- At the very top of the window, just below the title bar, is a row of tabs. When the Explorer is first started only the first tab is active; the others are greyed out. This is because it is necessary to open (and potentially pre-process) a data set before starting to explore the data.

# WEKA Explorer



# WEKA Explorer

- The tabs are as follows:
  - > **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.

# WEKA Explorer

Weka Explorer

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

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

Filter: Choose **None** Apply

Current relation  
Relation: churn-data  
Instances: 3333 Attributes: 21

Attributes

All None Invert Pattern

No.	Name
<input checked="" type="checkbox"/>	state
<input type="checkbox"/>	account_length
<input type="checkbox"/>	area_code
<input type="checkbox"/>	phone_number
<input type="checkbox"/>	international_plan
<input type="checkbox"/>	voice_mail_plan
<input type="checkbox"/>	number_vmail_messages
<input type="checkbox"/>	total_day_minutes
<input type="checkbox"/>	total_day_calls
<input type="checkbox"/>	total_day_charge
<input type="checkbox"/>	total_eve_minutes
<input type="checkbox"/>	total_eve_calls
<input type="checkbox"/>	total_eve_charge

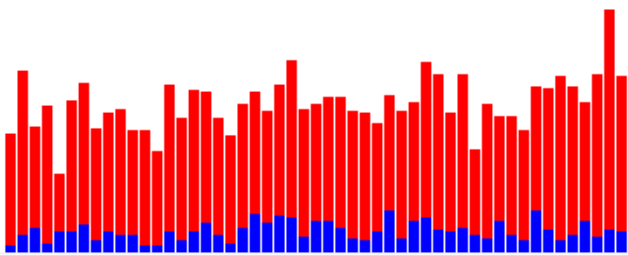
Remove

Selected attribute

Name: state  
Missing: 0 (0%)  
Distinct: 51  
Type: Nominal  
Unique: 0 (0%)

No.	Label	Count
1	AK	52
2	AL	80
3	AR	55
4	AZ	64
5	CA	34
6	CO	66
7	CT	74

Class: churn (Nom) Visualize All



Status: OK Log x 0



# WEKA Explorer - Classification

- Selecting a Classifier

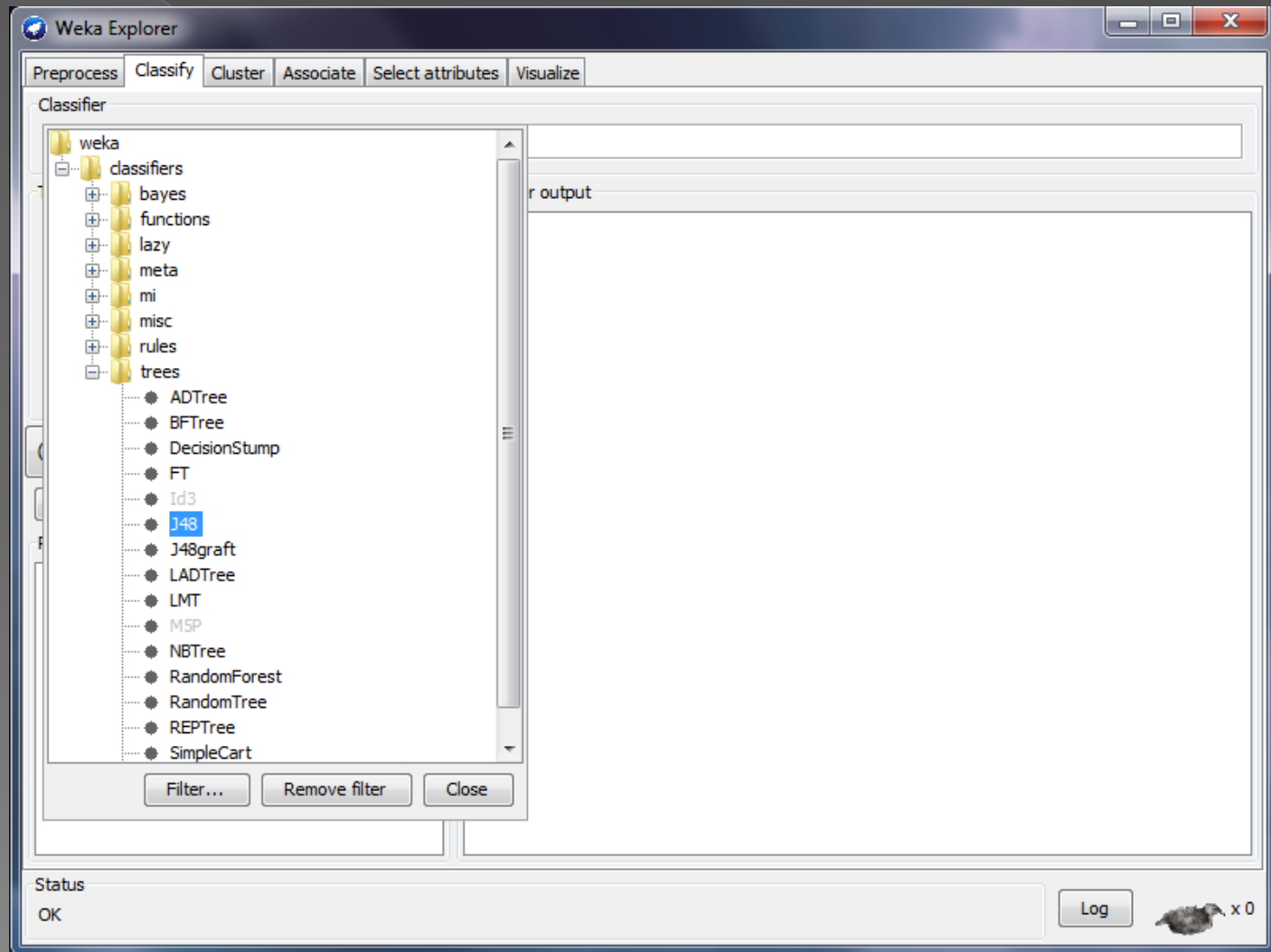
At the top of the classify section is the Classifier box. This box has a text field that gives the name of the currently selected classifier, and its options. The Choose button allows you to choose one of the classifiers that are available in WEKA.

- There are four test modes available:

- > **Use training set:** The classifier is evaluated on how well it predicts the class of the instances it was trained on.
- > **Supplied test set:** The classifier is evaluated on how well it predicts the class of a set of instances loaded from a file. Clicking the Set... Button brings up a dialog allowing you to choose the file to test on.
- > **Cross-validation:** The classifier is evaluated by cross-validation, using the number of folds that are entered in the Folds text field.
- > **Percentage split:** The classifier is evaluated on how well it predicts a certain percentage of the data which is held out for testing. The amount of data held out depends on the value entered in the % field.

- Once the classifier, test options and class have all been set, the learning process is started by clicking on the Start button. While the classifier is busy being trained, the little bird moves around.

# WEKA Explorer - Classification



# WEKA Explorer - Classification

The screenshot shows the Weka Explorer application window. The 'Classify' tab is active, and the 'J48 -U -M 2' classifier is selected. The 'Test options' section shows 'Cross-validation' with 10 folds. The 'Classifier output' pane displays the following results:

```
--- Summary ---
Correctly Classified Instances      3068      92.0492 %
Incorrectly Classified Instances    265       7.9508 %
Kappa statistic                    0.6639
Mean absolute error                 0.095
Root mean squared error             0.265
Relative absolute error             38.3215 %
Root relative squared error         75.2804 %
Total Number of Instances          3333

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
	0.671	0.037	0.753	0.671	0.71	0.829
	0.963	0.329	0.945	0.963	0.954	0.829
Weighted Avg.	0.92	0.287	0.917	0.92	0.919	0.829

```

=== Confusion Matrix ===
      a    b  <-- classified as
324  159 |  a = True
106 2744 |  b = False

```

The 'Result list' shows a single entry: '17:08:50 - trees.J48'. The 'Status' bar at the bottom indicates 'OK'.

# WEKA Explorer - Classification

- When training is complete, several things happen. The Classifier output area to the right of the display is filled with text describing the results of training and testing. A new entry appears in the Result list box.
- The text in the Classifier output area is split into several sections:
  - > **Run information:** A list of information giving the learning scheme options, relation name, instances, attributes and test mode that were involved in the process.
  - > **Classifier model (full training set):** A textual representation of the classification model that was produced on the full training data.
  - > **Summary:** A list of statistics summarizing how accurately the classifier was able to predict the true class of the instances under the chosen test mode.
  - > **Detailed Accuracy By Class:** A more detailed per-class break down of the classifier's prediction accuracy.
  - > **Confusion Matrix:** Shows how many instances have been assigned to each class. Elements show the number of test examples whose actual class is the row and whose predicted class is the column.

# WEKA Manual

- ◉ More can be found in WEKA Manual:
  - > <http://www.cs.waikato.ac.nz/ml/weka/documentation.html>

