

# Machine Learning with WEKA

Eibe Frank

Department of Computer Science,  
University of Waikato, New Zealand

- WEKA: A Machine Learning Toolkit
- The Explorer
  - Classification and Regression
  - Clustering
  - Association Rules
  - Attribute Selection
  - Data Visualization
- The Experimenter
- The Knowledge Flow GUI
- Conclusions

## WEKA: the bird



*Copyright: Martin Kramer (mkramer@wxs.nl)*

2/4/2004

University of Waikato

2

## WEKA: the software

- Machine learning/data mining software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications
- Complements “Data Mining” by Witten & Frank
- Main features:
  - ◆ Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods
  - ◆ Graphical user interfaces (incl. data visualization)
  - ◆ Environment for comparing learning algorithms

2/4/2004

University of Waikato

3

## WEKA: versions

- There are several versions of WEKA:
  - ◆ WEKA 3.0: “book version” compatible with description in data mining book
  - ◆ WEKA 3.2: “GUI version” adds graphical user interfaces (book version is command-line only)
  - ◆ WEKA 3.3: “development version” with lots of improvements
- This talk is based on the latest snapshot of WEKA 3.3 (soon to be WEKA 3.4)

2/4/2004

University of Waikato

4

## WEKA only deals with “flat” files

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```



Flat file in  
ARFF format

2/4/2004

University of Waikato

5

## WEKA only deals with “flat” files

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal, atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

numeric attribute

nominal attribute



2/4/2004

University of Waikato

6

Weka GUI Chooser  
Waikato Environment for Knowledge Analysis  
(c) 1999 – 2003 University of Waikato New Zealand

GUI  
Simple CLI  
Explorer  
Experimenter  
KnowledgeFlow

2/4/2004 University of Waikato 7

Weka GUI Chooser  
Waikato Environment for Knowledge Analysis  
(c) 1999 – 2003 University of Waikato New Zealand

GUI  
Simple CLI  
Explorer  
Experimenter  
KnowledgeFlow

2/4/2004 University of Waikato 8

2/4/2004 University of Waikato 9

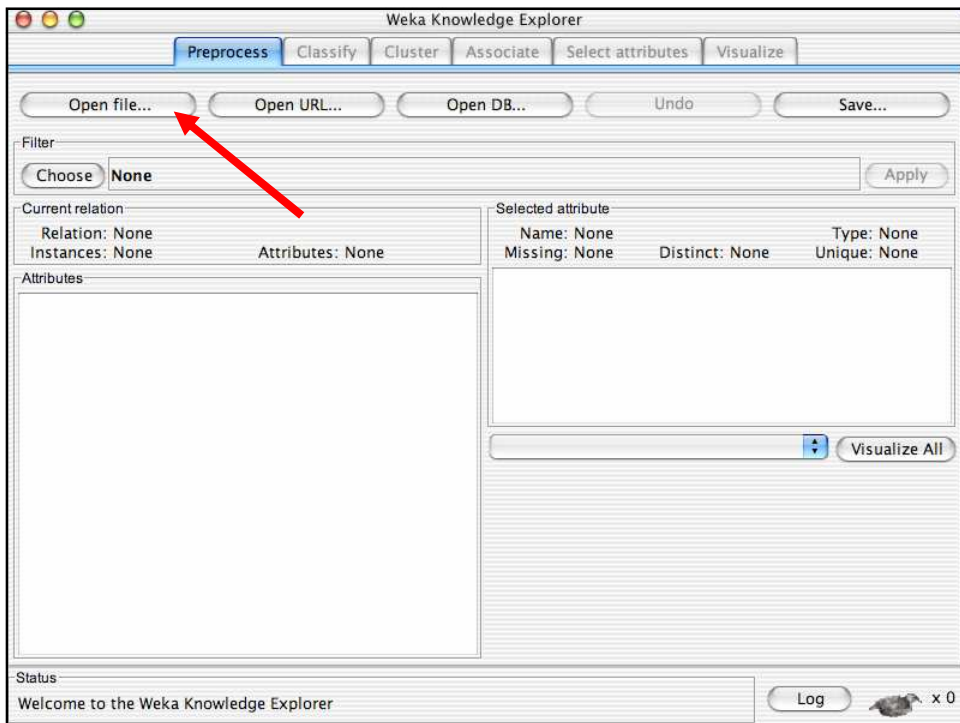
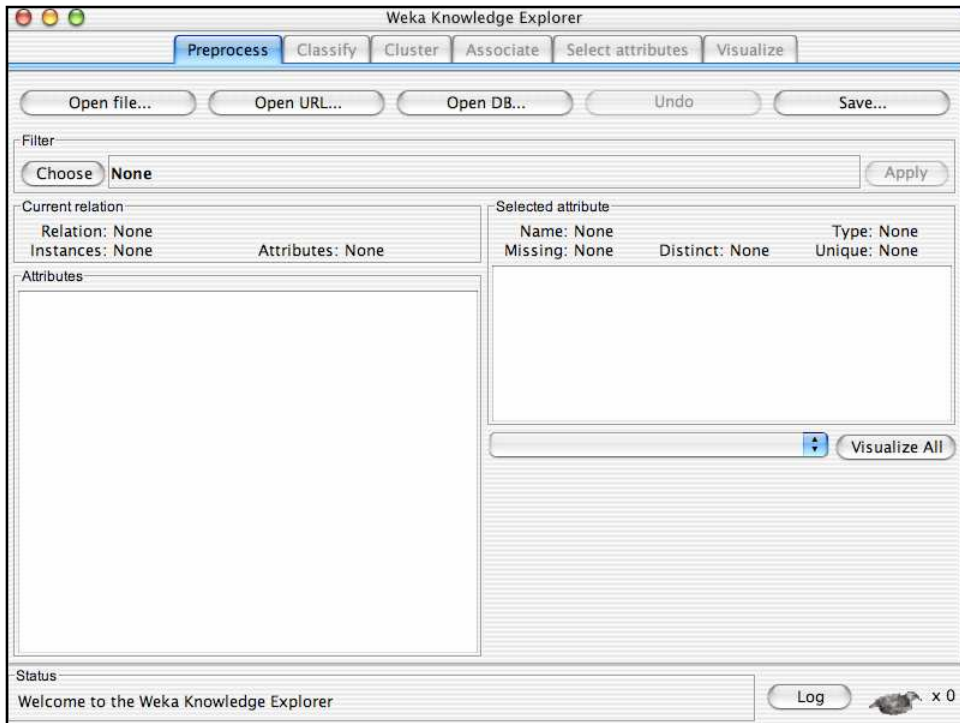
## Explorer: pre-processing the data

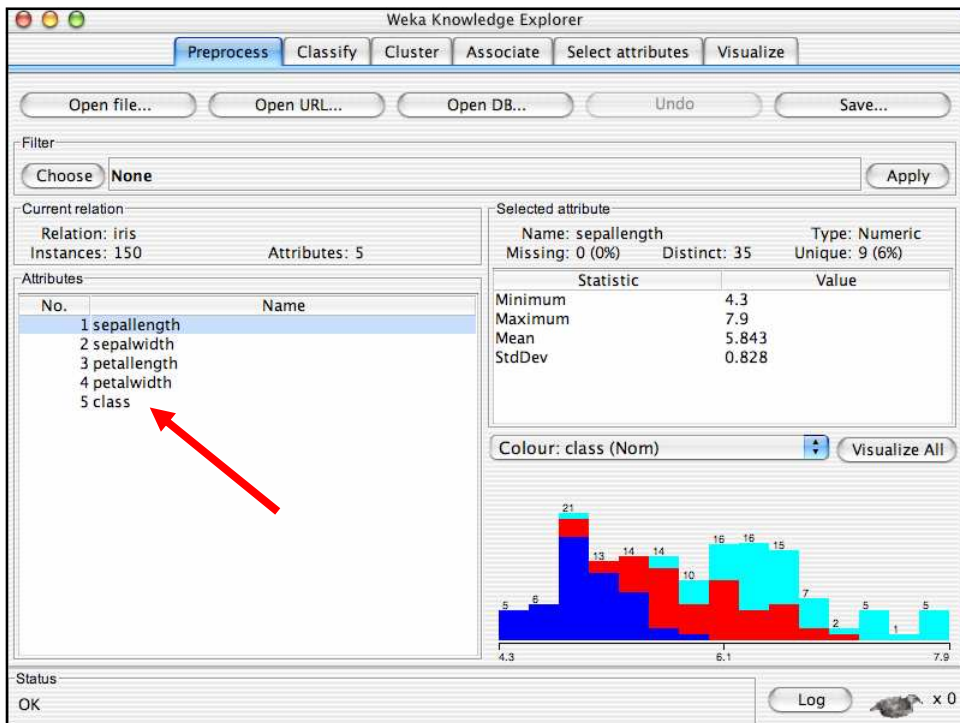
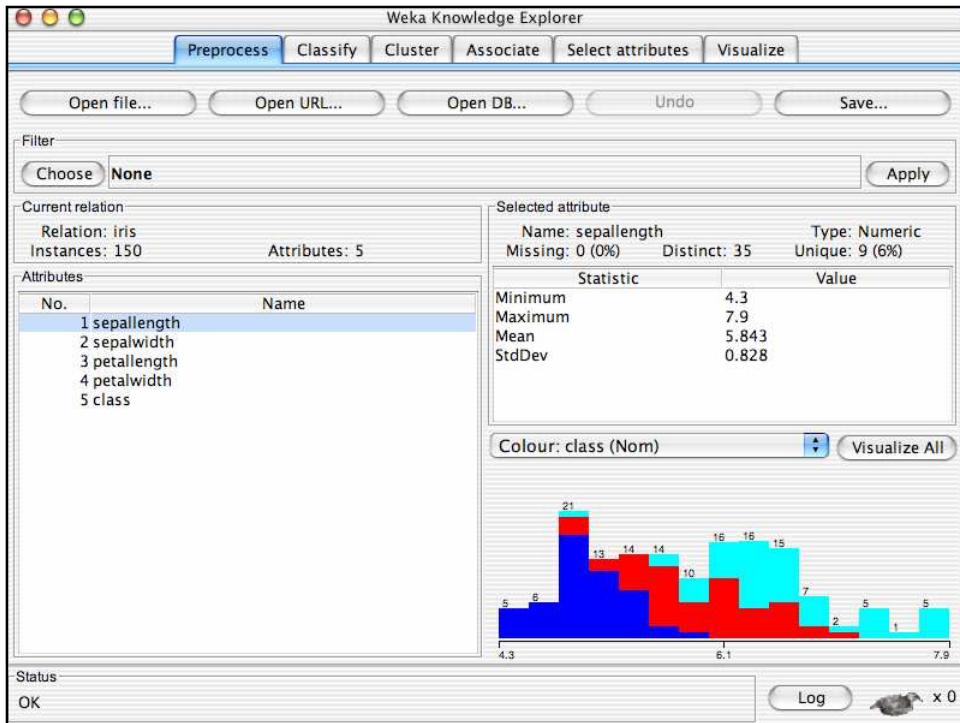
- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - ◆ Discretization, normalization, resampling, attribute selection, transforming and combining attributes, ...

2/4/2004

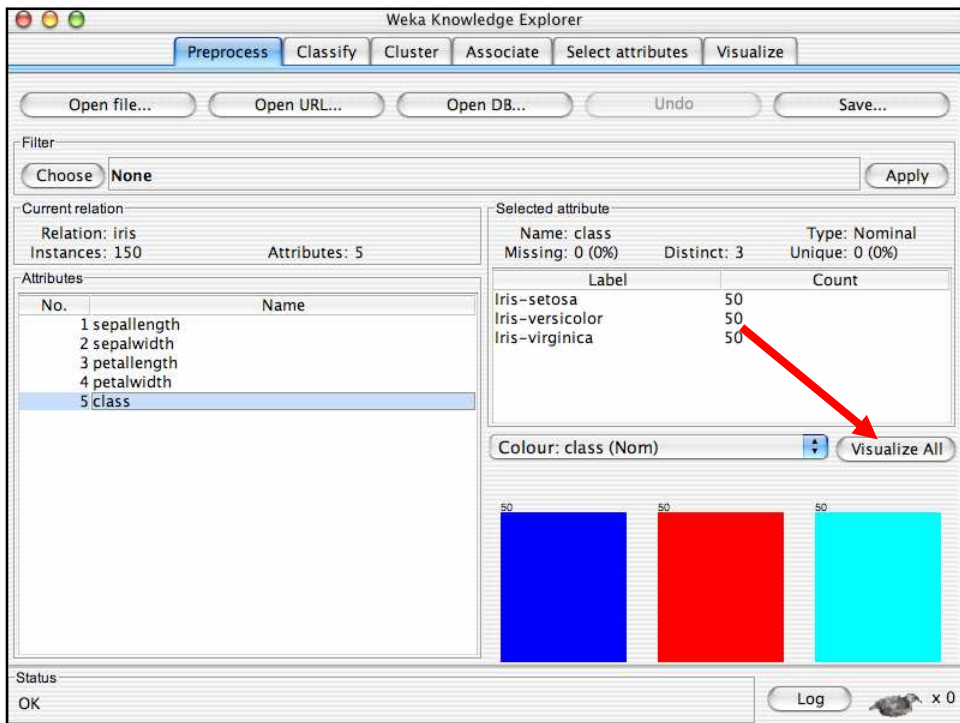
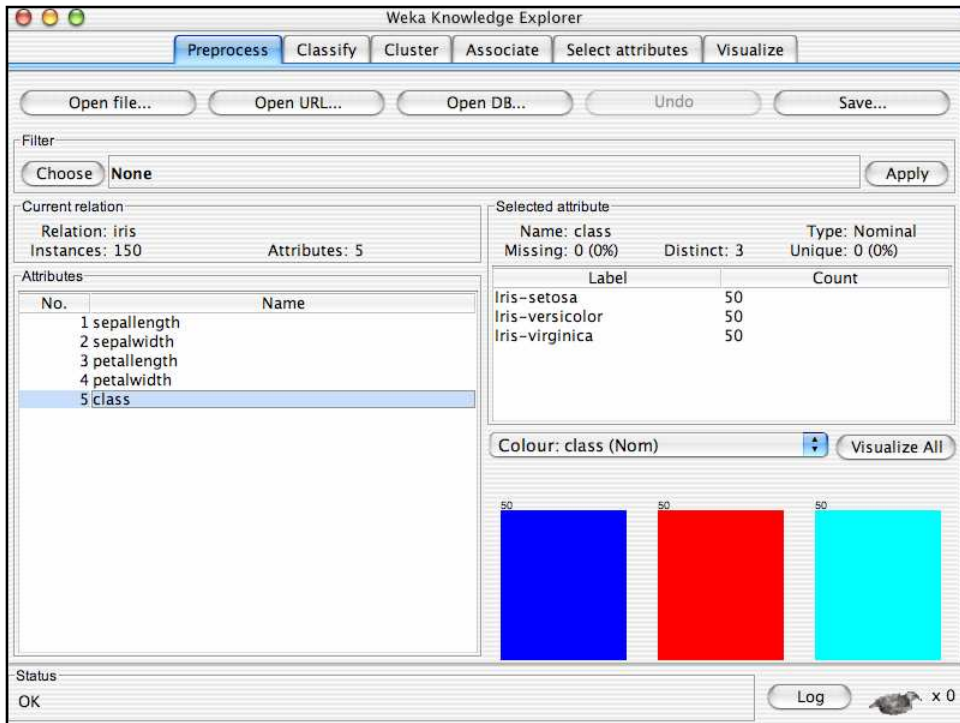
University of Waikato

10

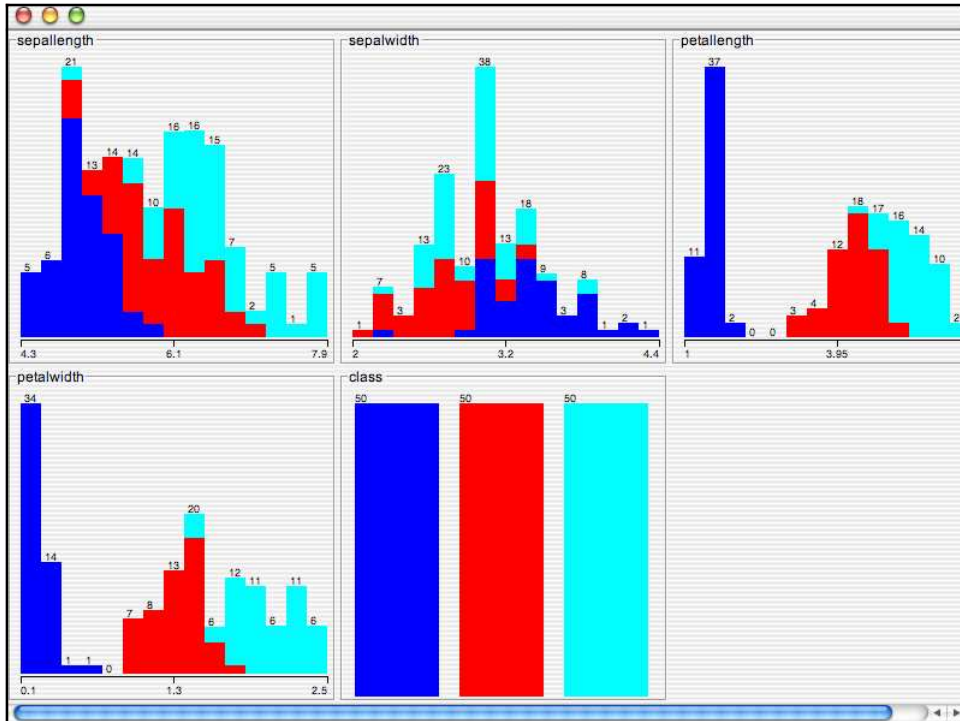












Weka Knowledge Explorer

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

Open file... | Open URL... | Open DB... | Undo | Save...

Filter: Choose **None** Apply

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

Selected attribute:  
Name: petal length  
Missing: 0 (0%)  
Distinct: 43  
Type: Numeric  
Unique: 10 (7%)

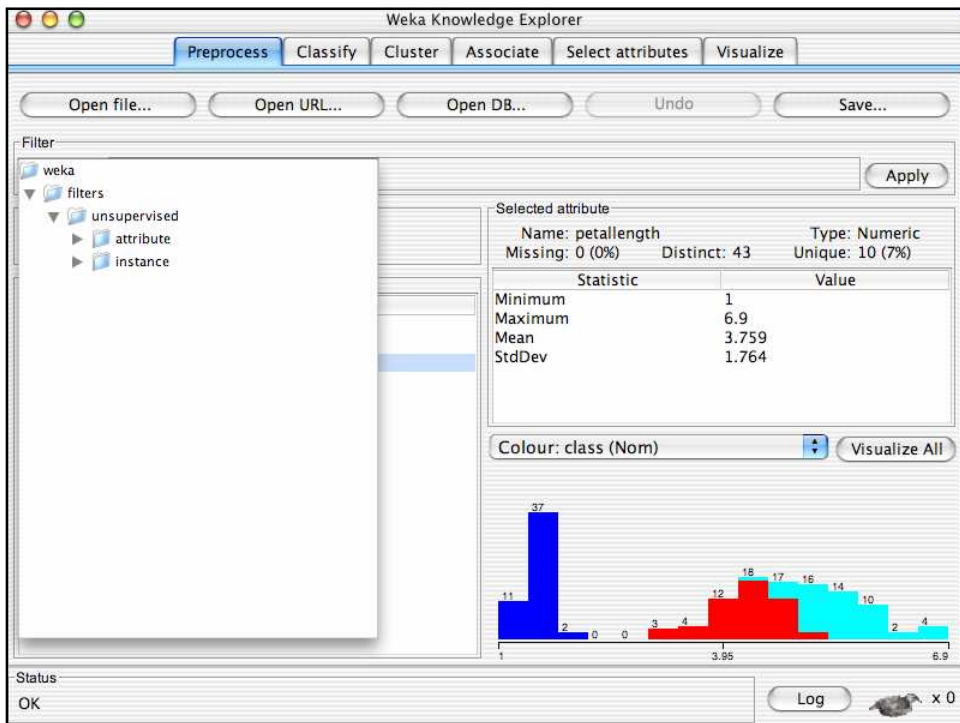
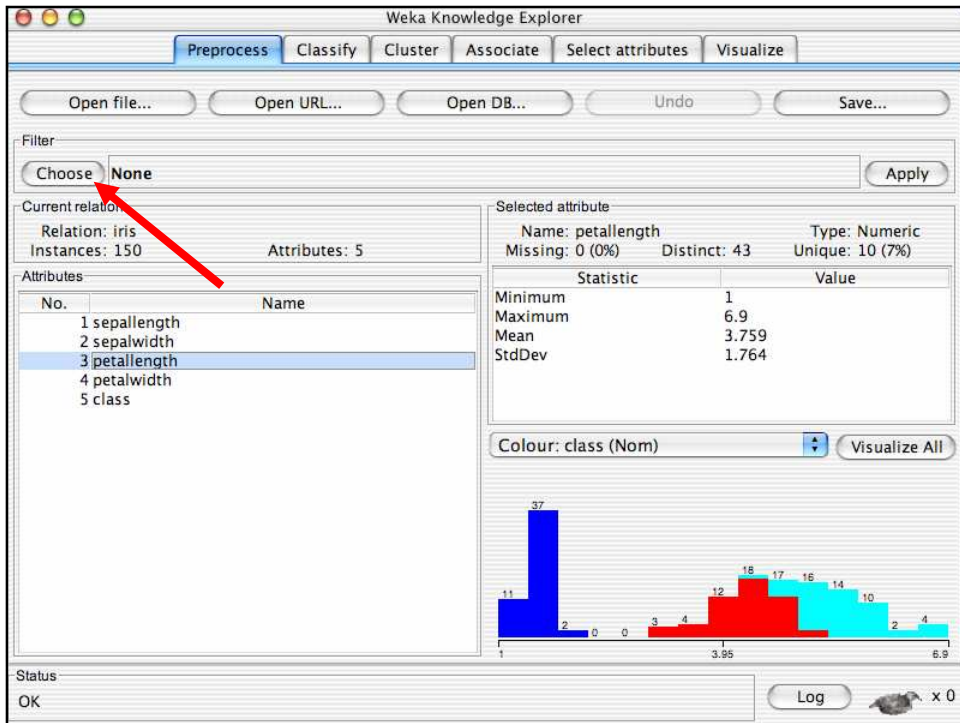
Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

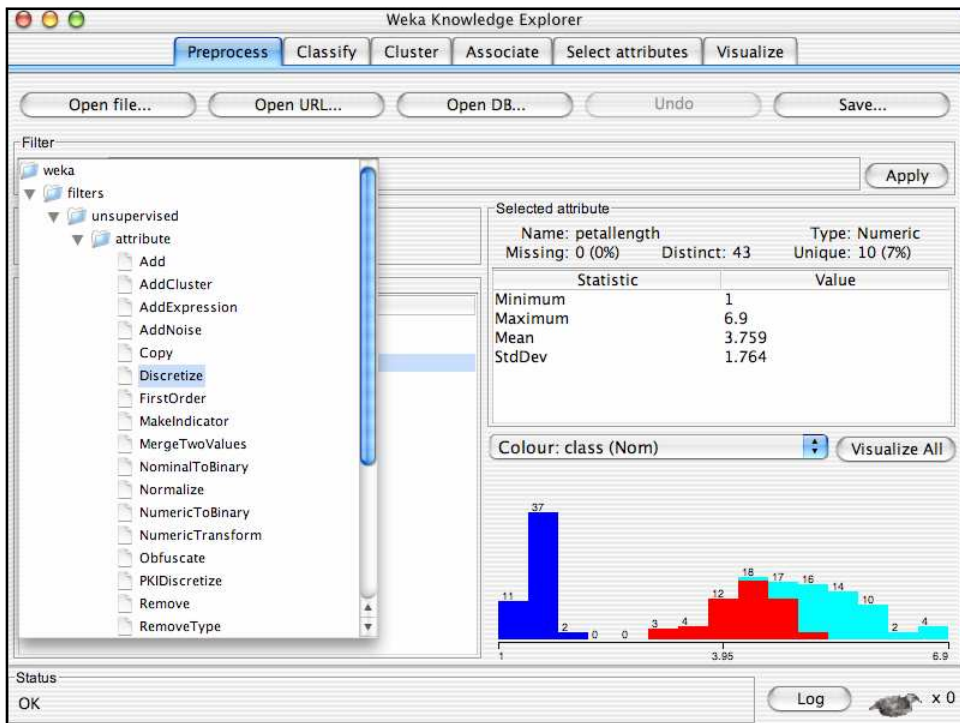
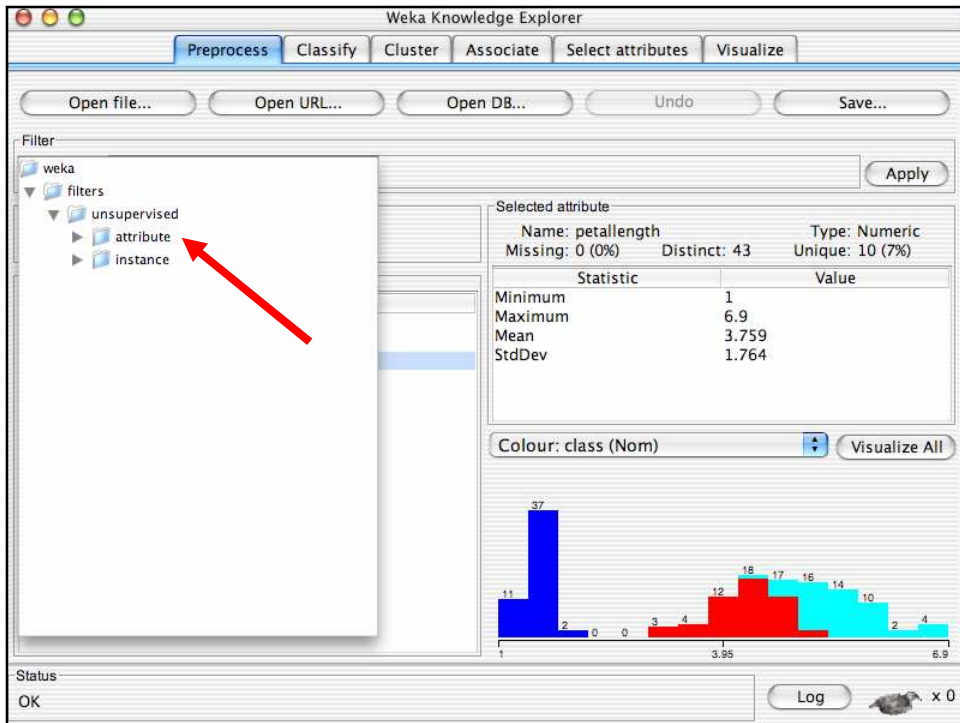
Attributes:

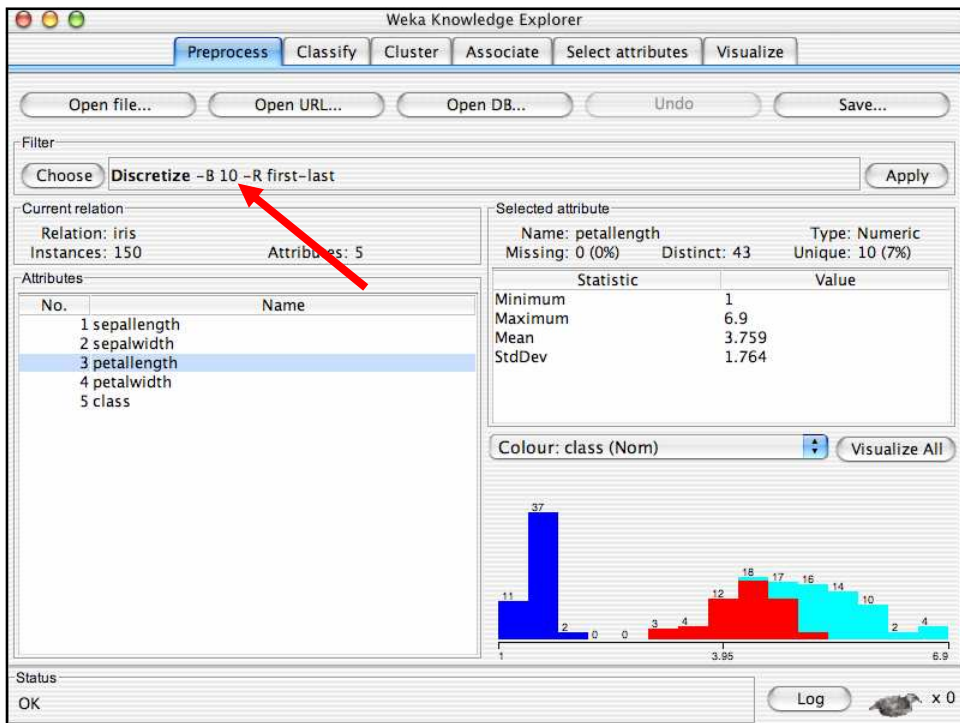
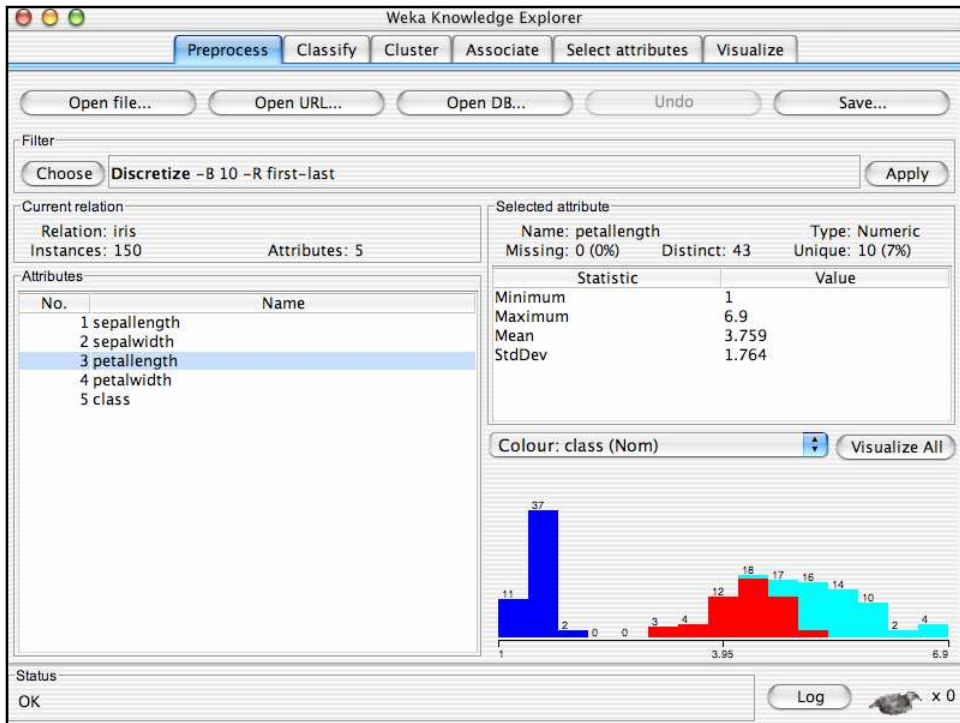
No.	Name
1	sepal length
2	sepal width
3	petal length
4	petal width
5	class

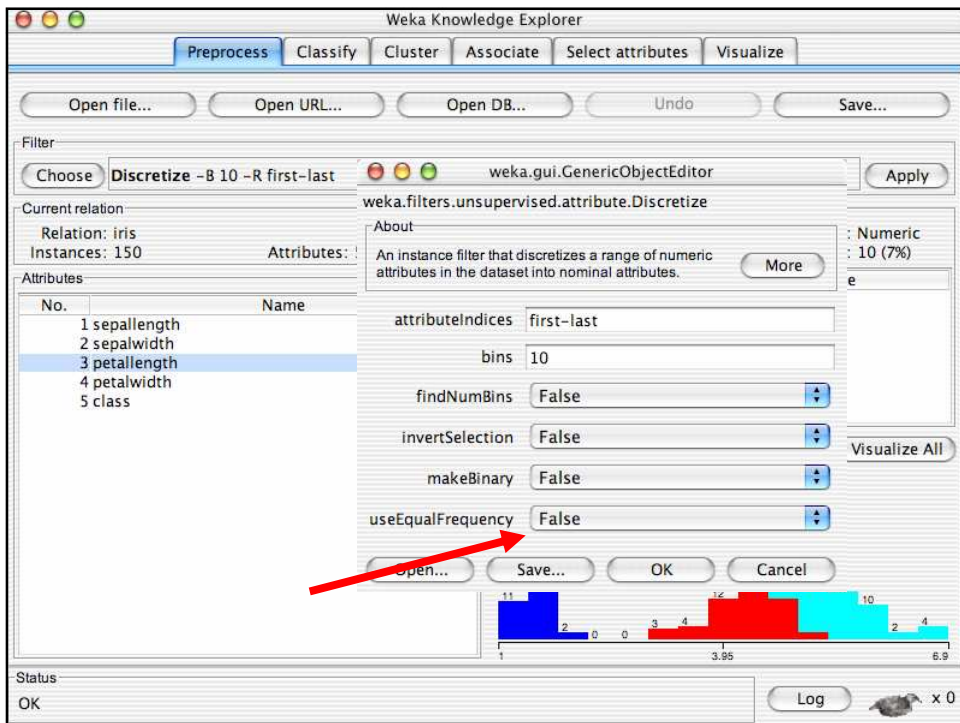
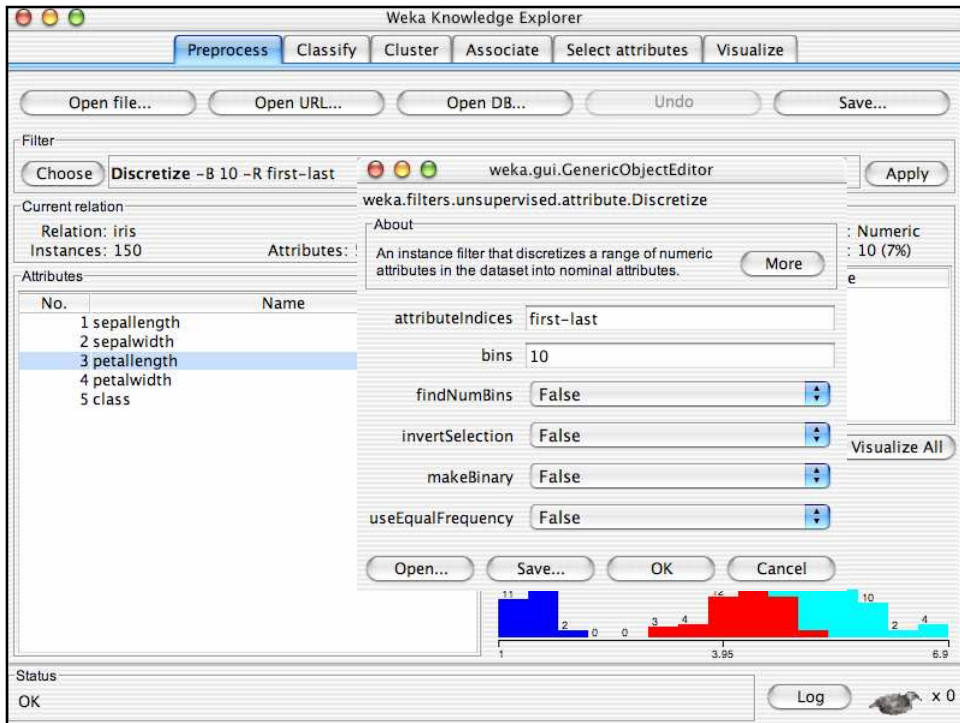
Colour: class (Nom) Visualize All

Status: OK Log x 0

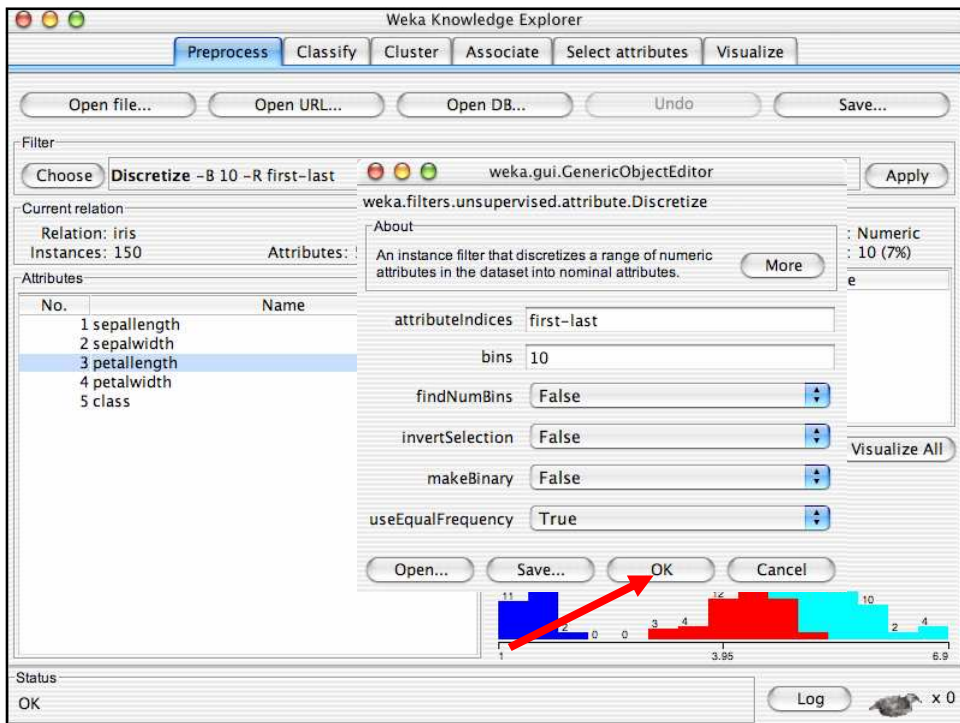
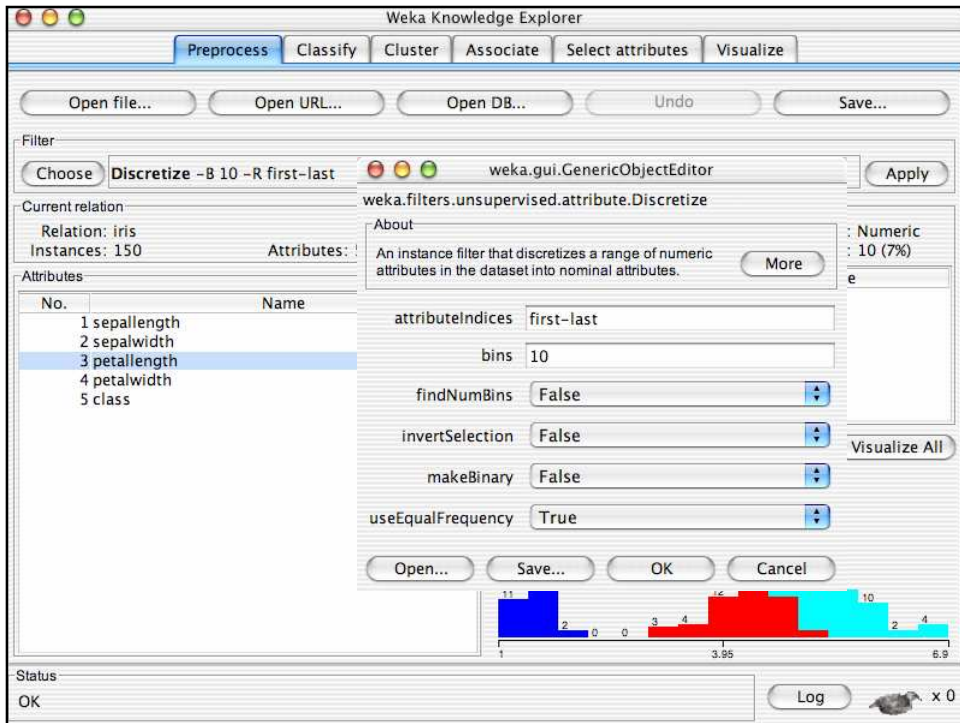


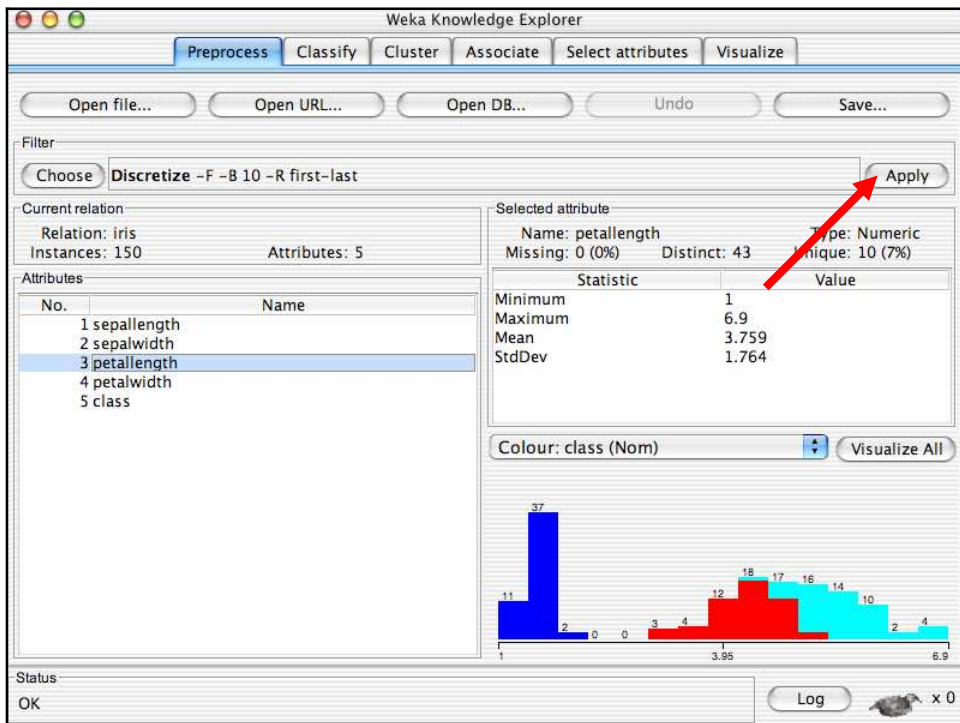
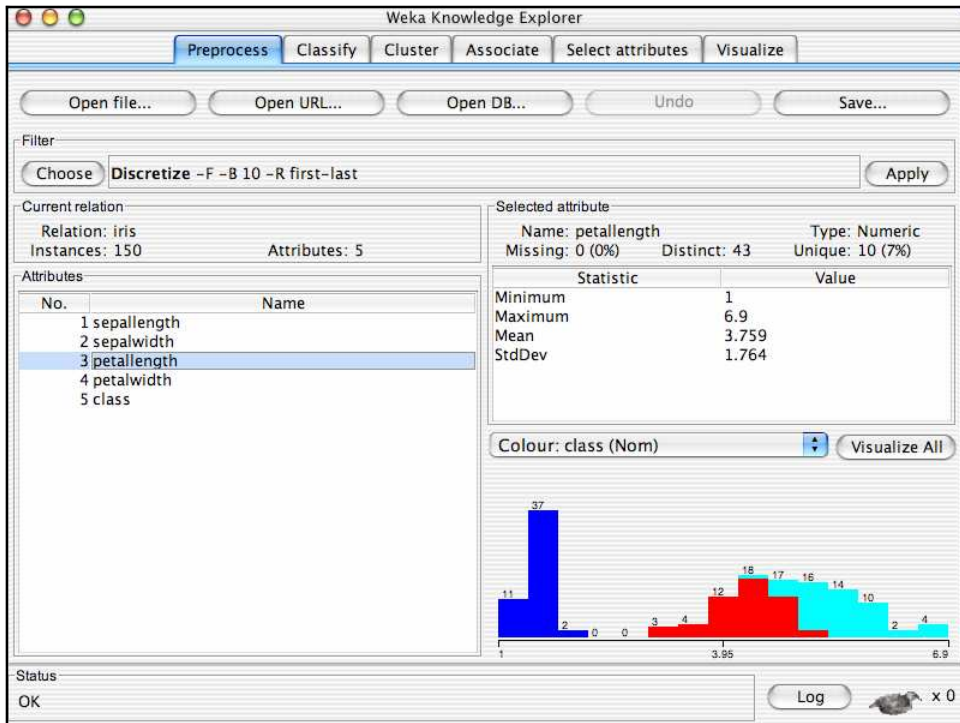




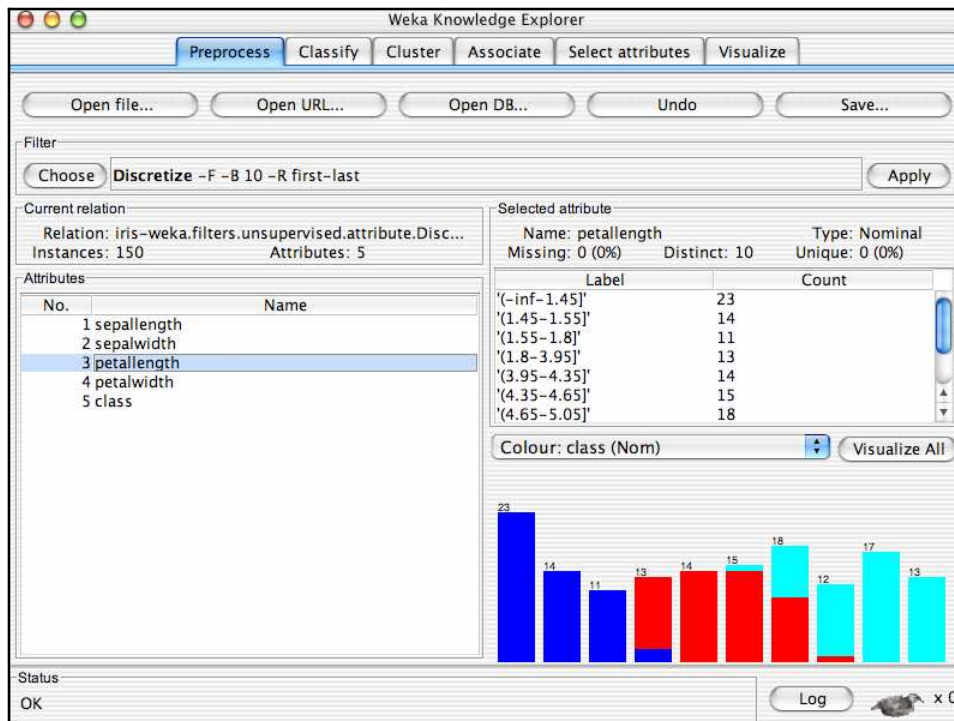






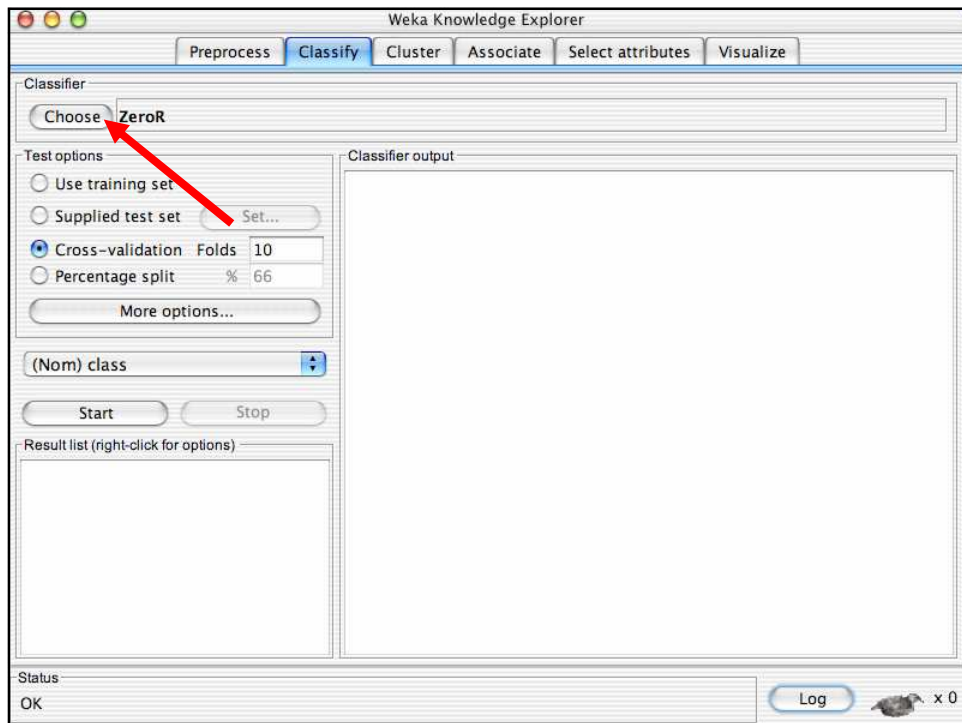
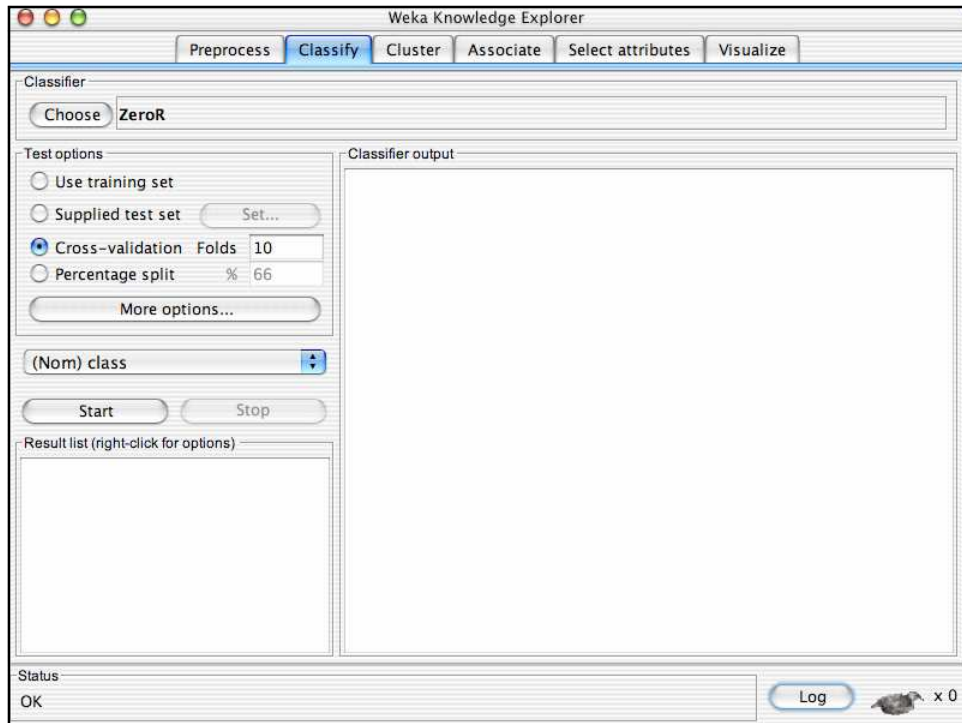


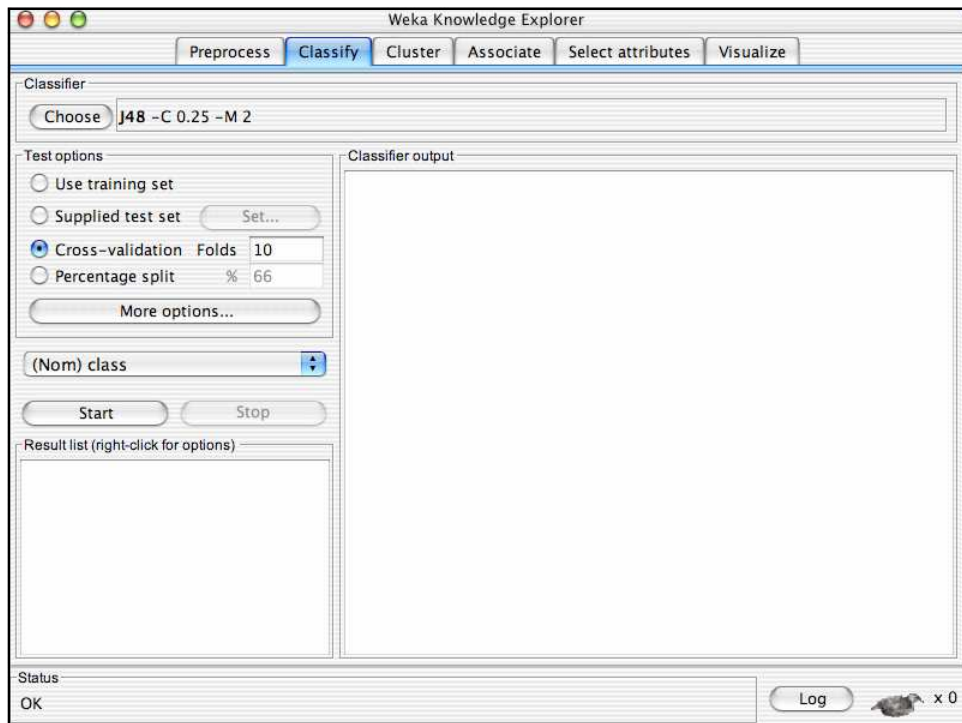
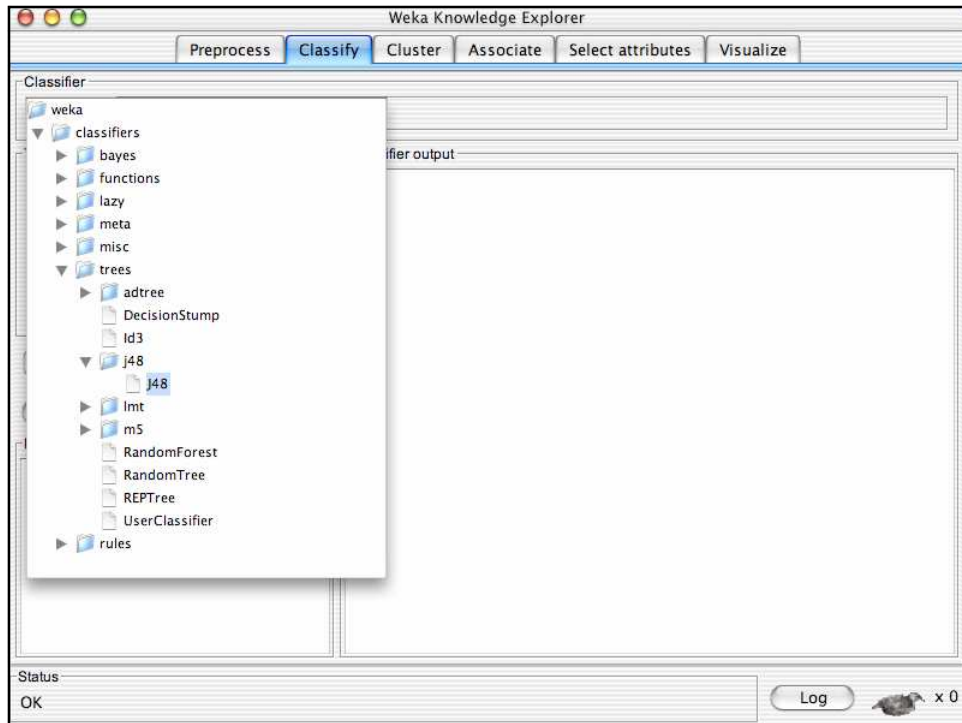


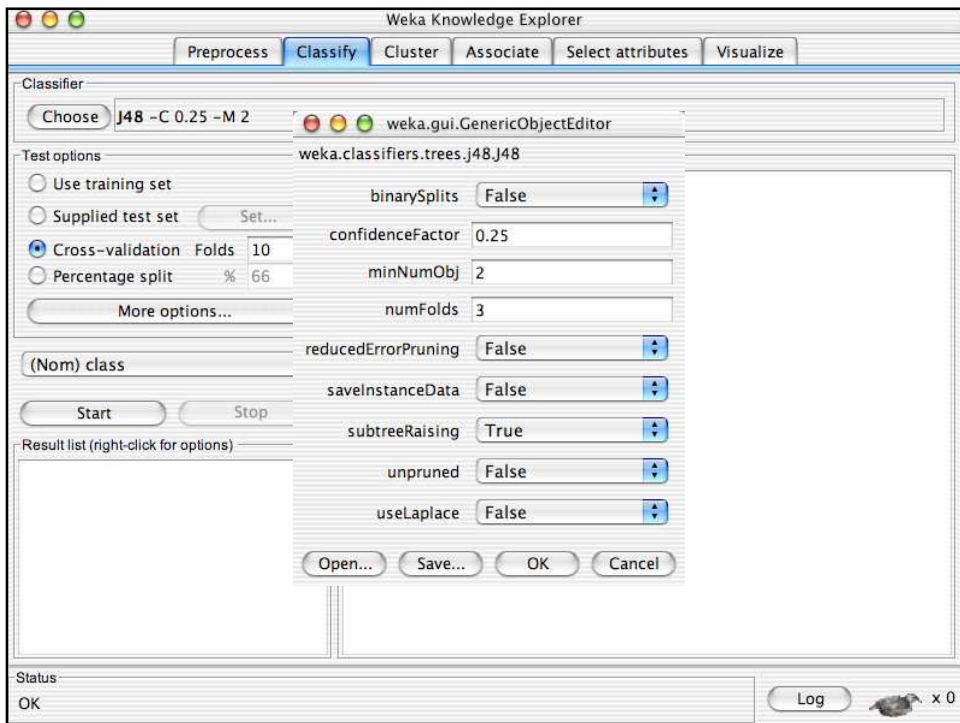
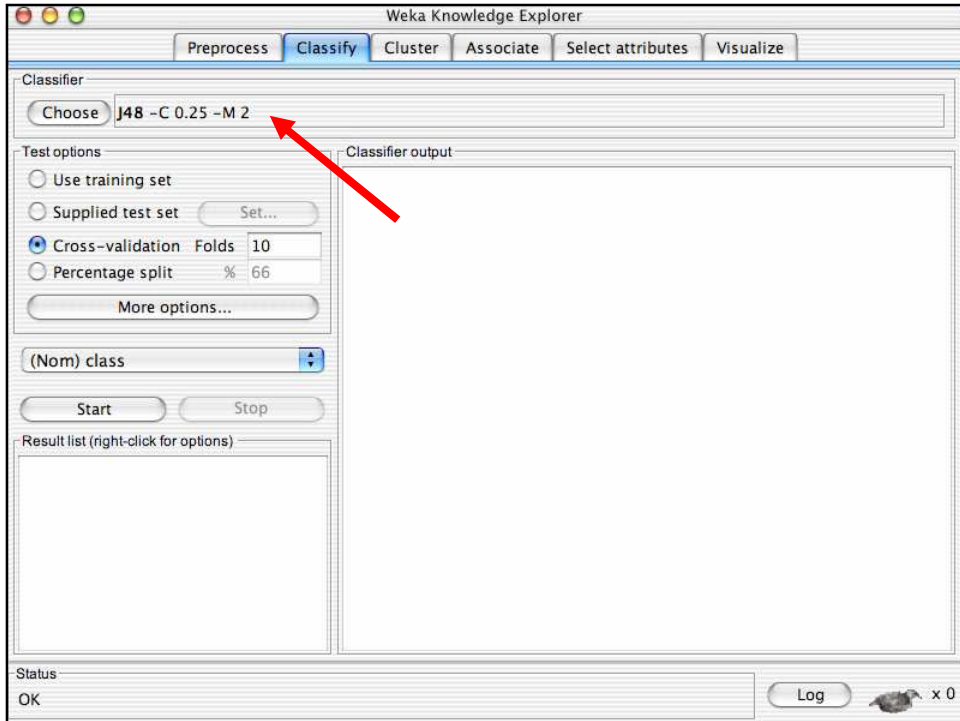


## Explorer: building “classifiers”

- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - ◆ Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes’ nets, ...
- “Meta”-classifiers include:
  - ◆ Bagging, boosting, stacking, error-correcting output codes, locally weighted learning, ...









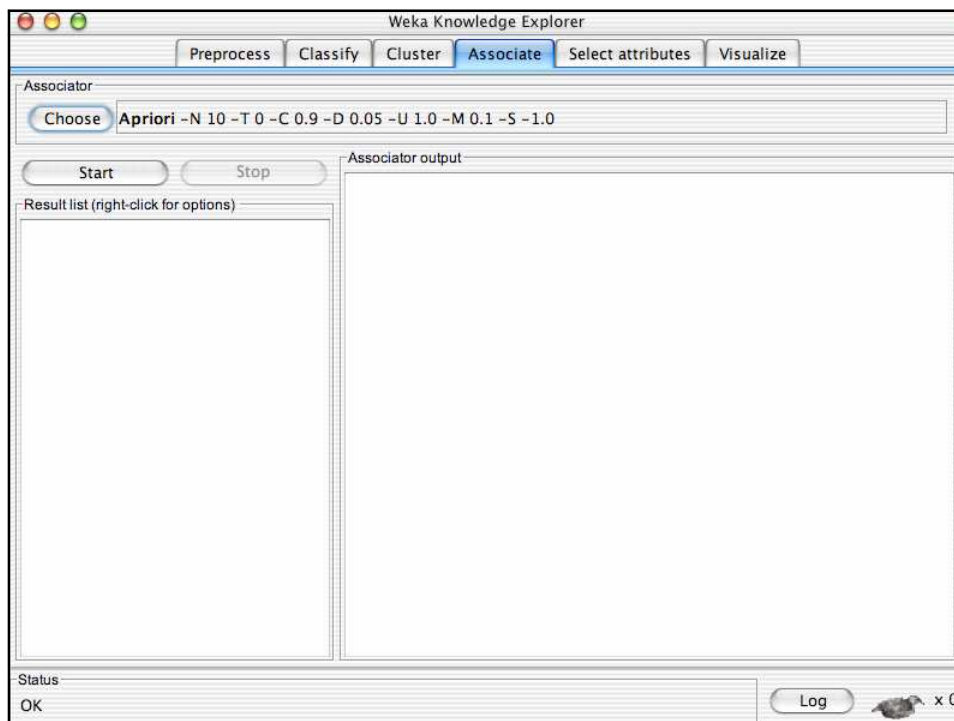
## Explorer: finding associations

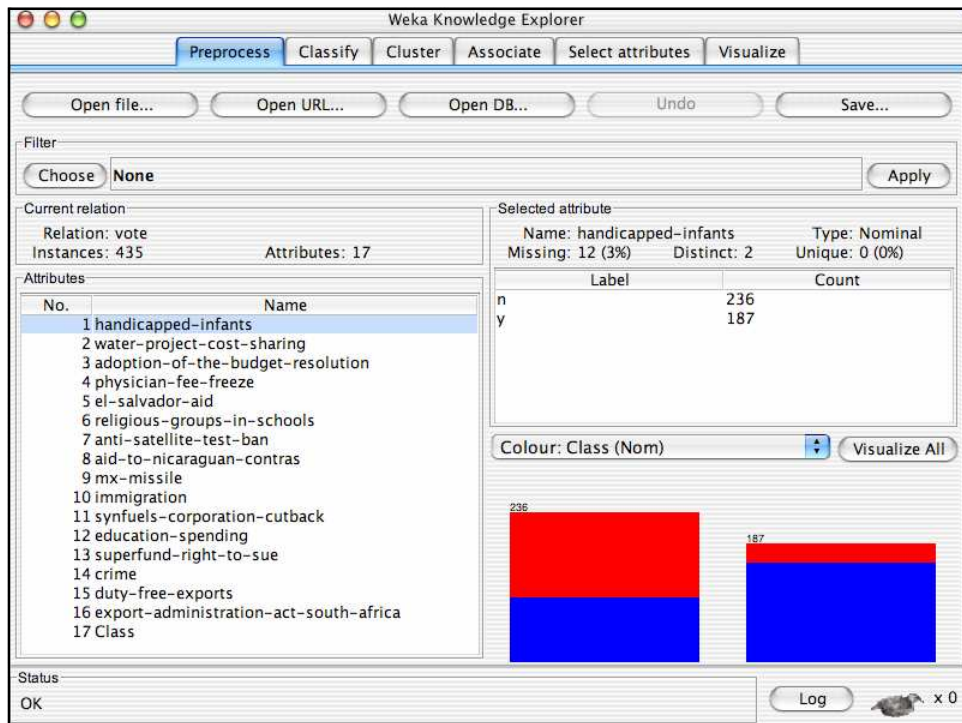
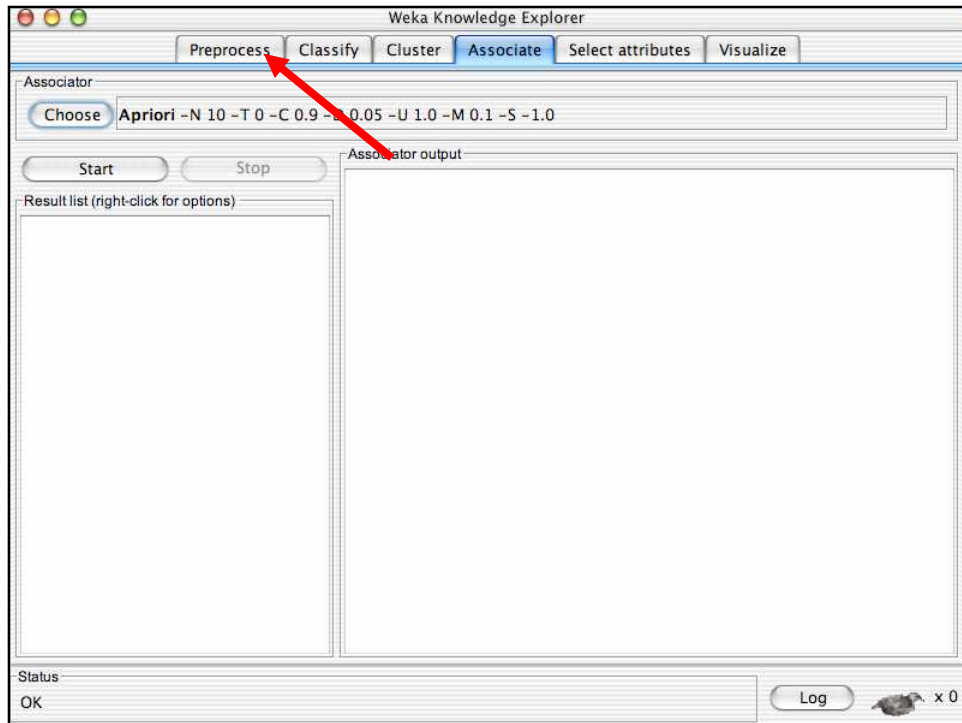
- WEKA contains an implementation of the Apriori algorithm for learning association rules
  - ◆ Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - ◆ milk, butter  $\Rightarrow$  bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that have a given minimum support and exceed a given confidence

2/4/2004

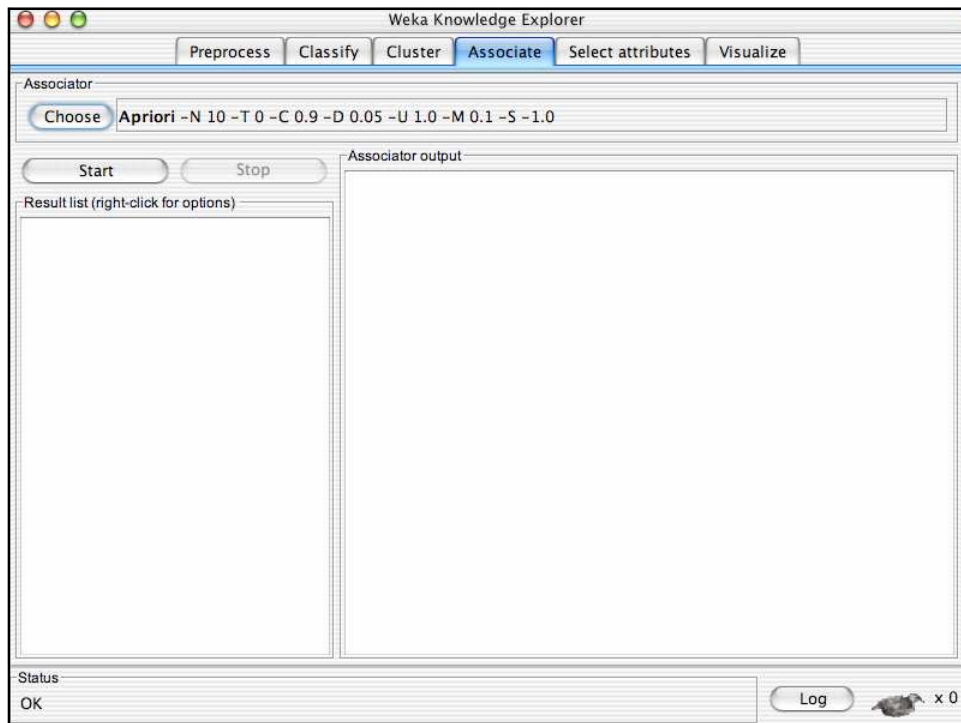
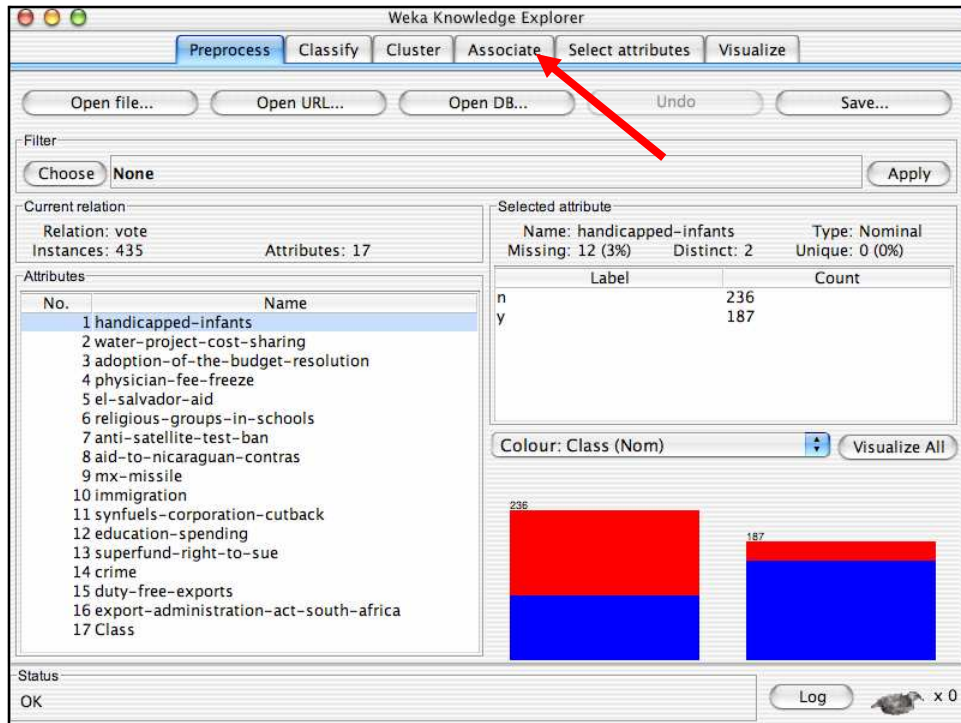
University of Waikato

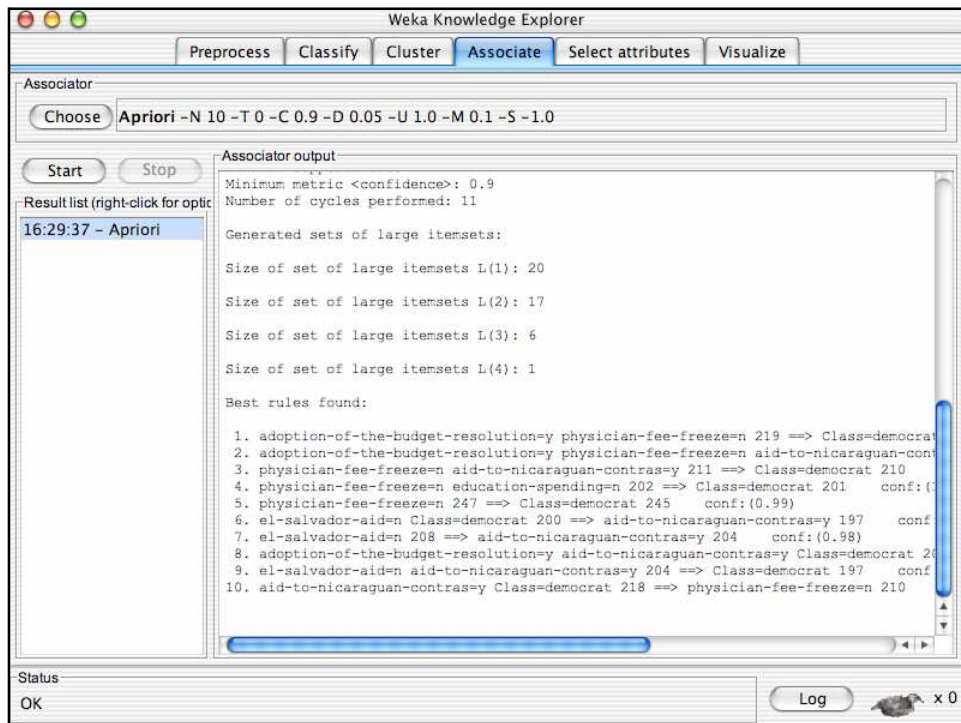
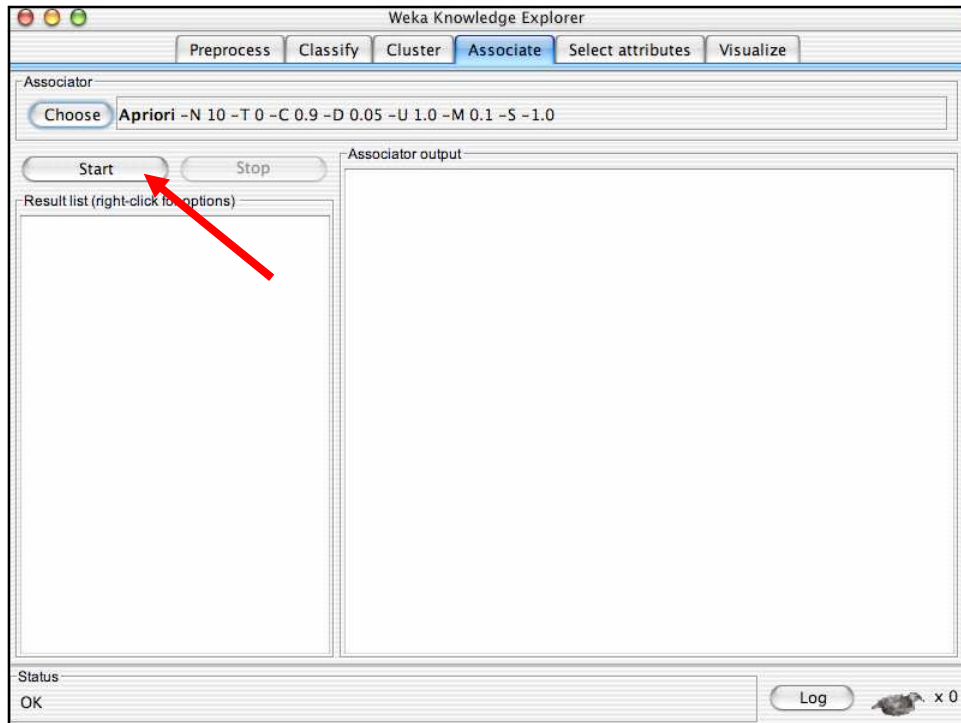
108











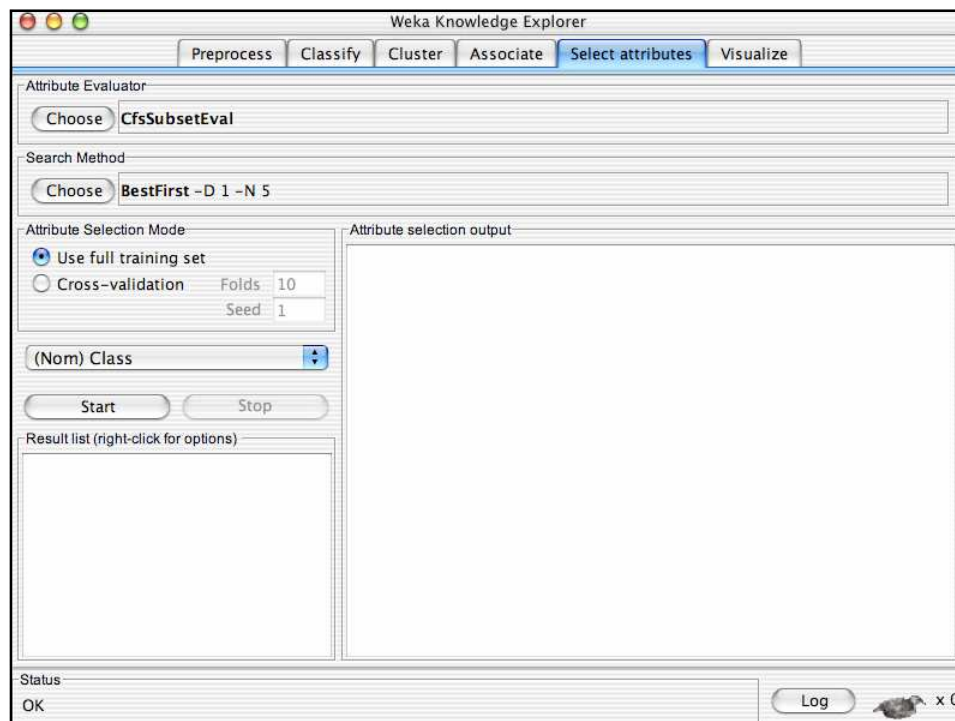
## Explorer: attribute selection

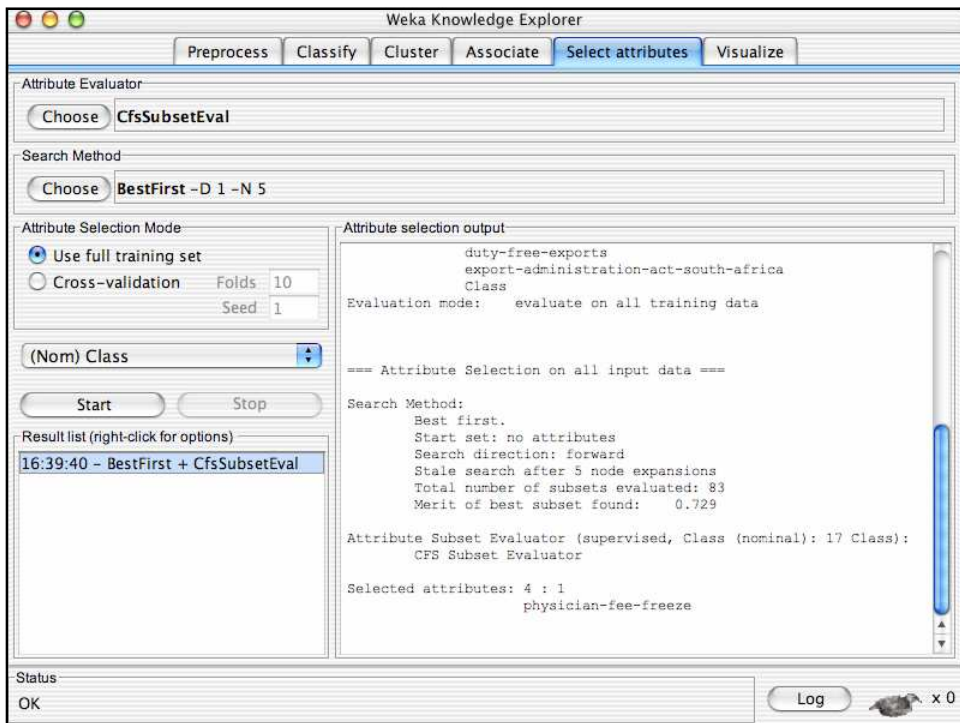
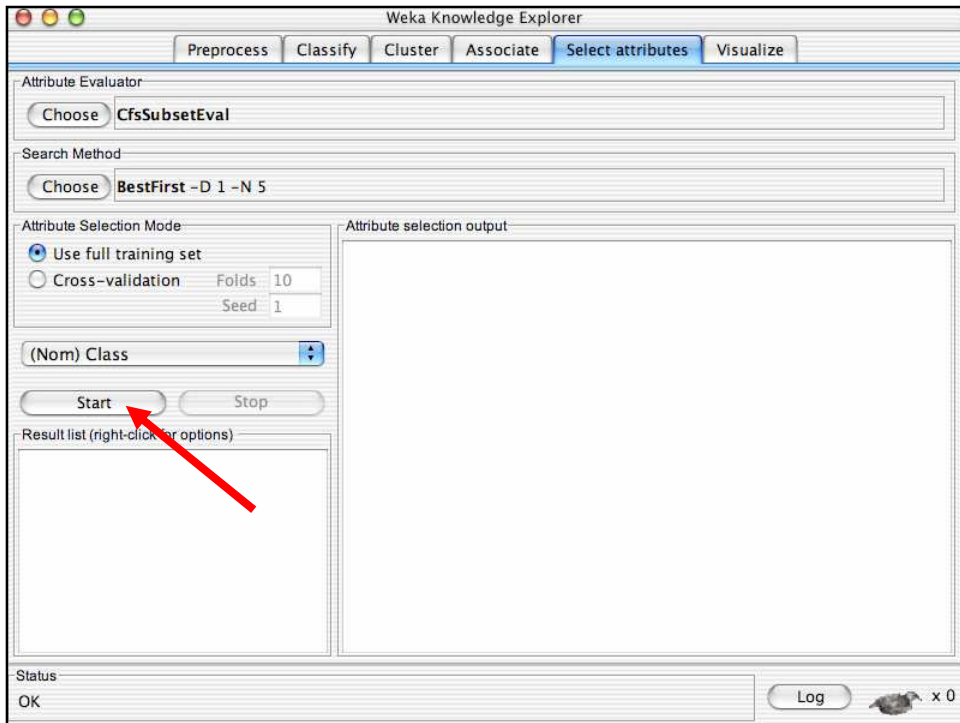
- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
  - ◆ A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
  - ◆ An evaluation method: correlation-based, wrapper, information gain, chi-squared, ...
- Very flexible: WEKA allows (almost) arbitrary combinations of these two

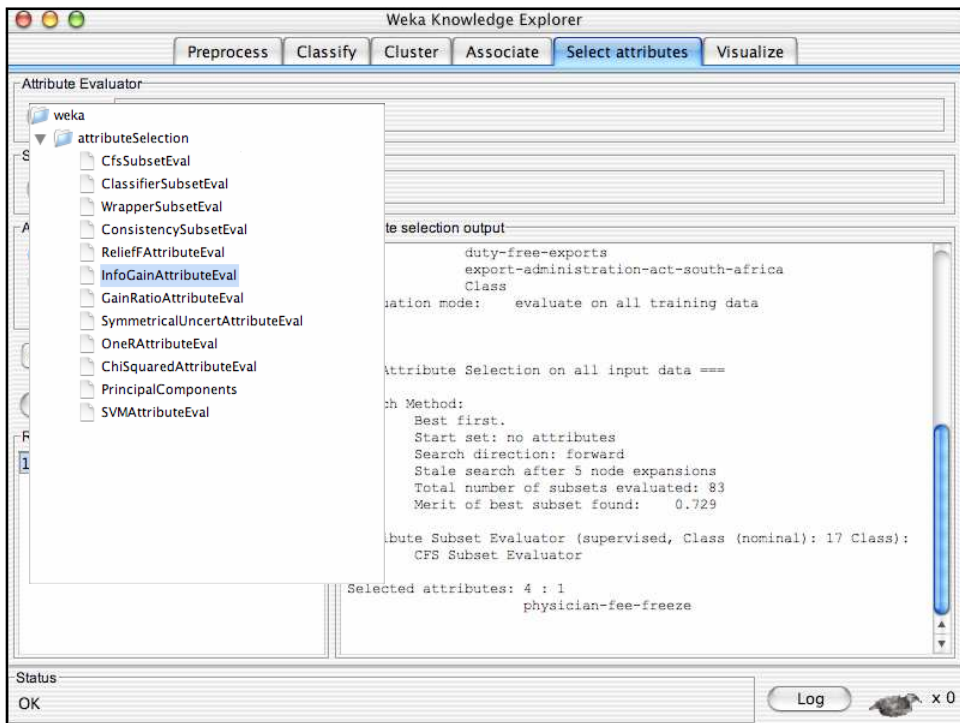
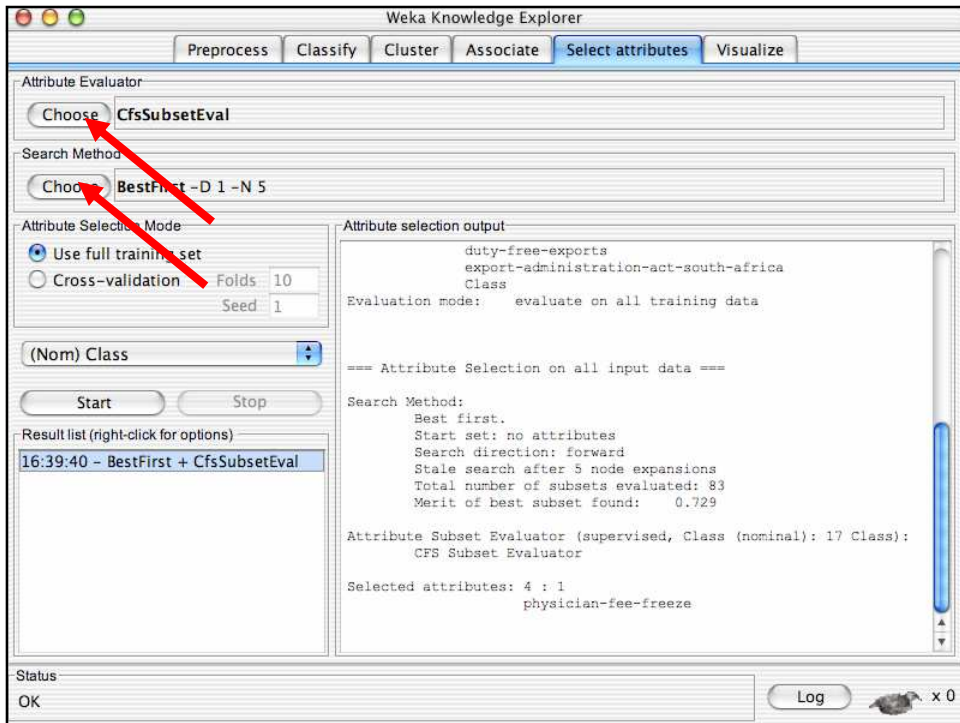
2/4/2004

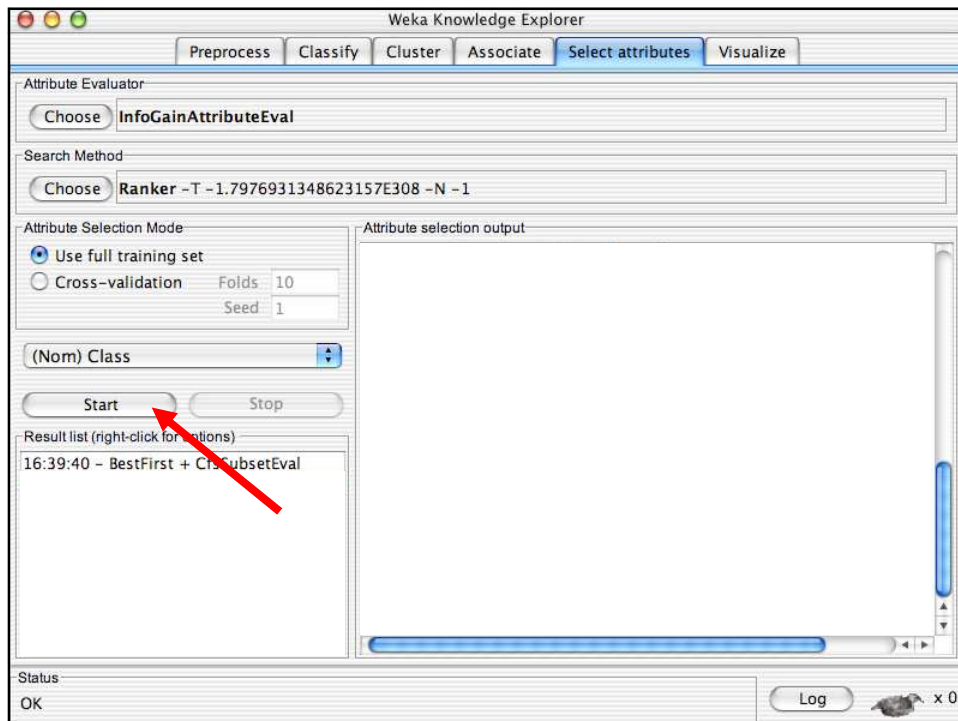
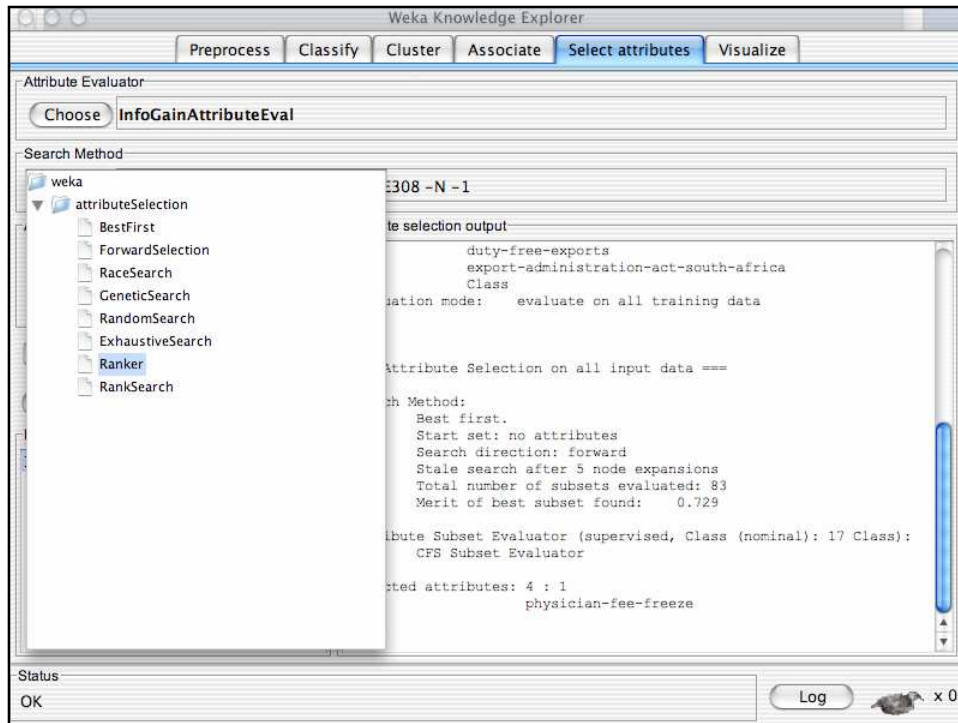
University of Waikato

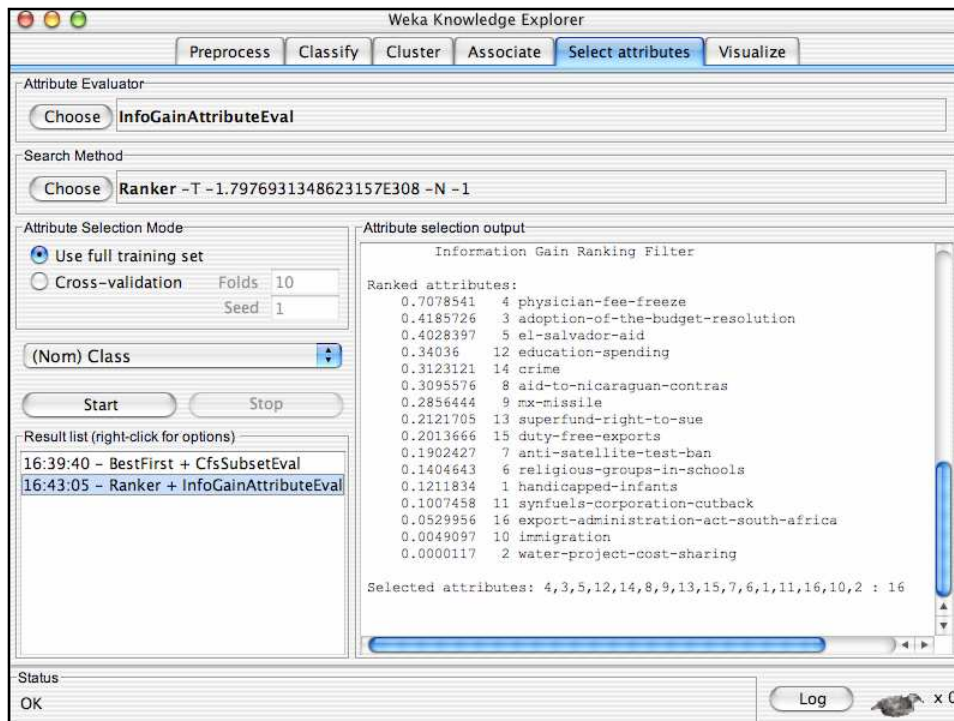
116







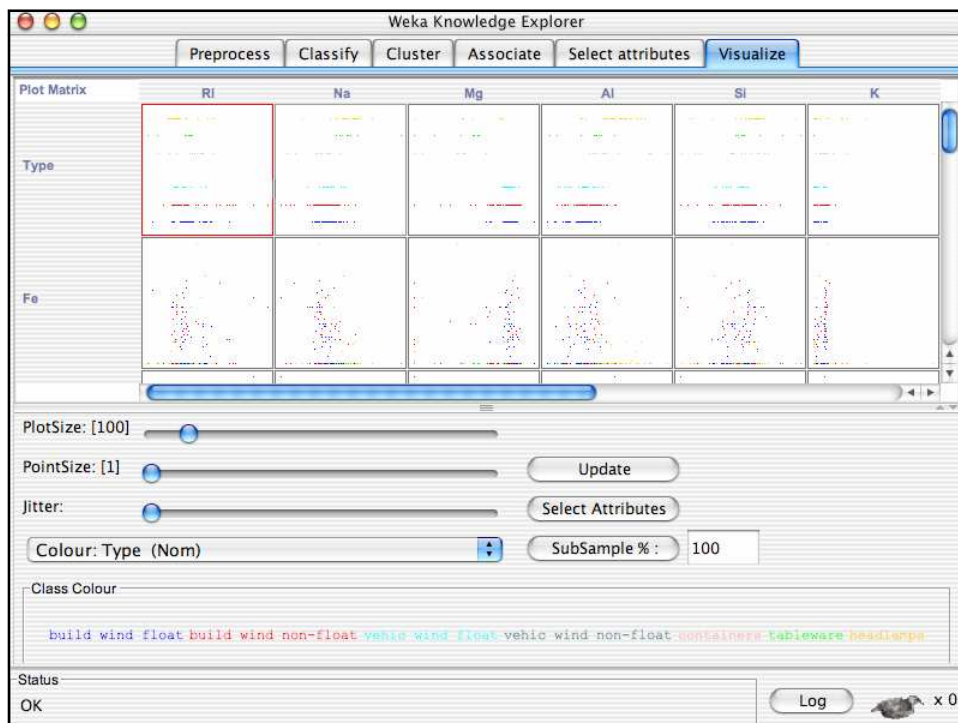
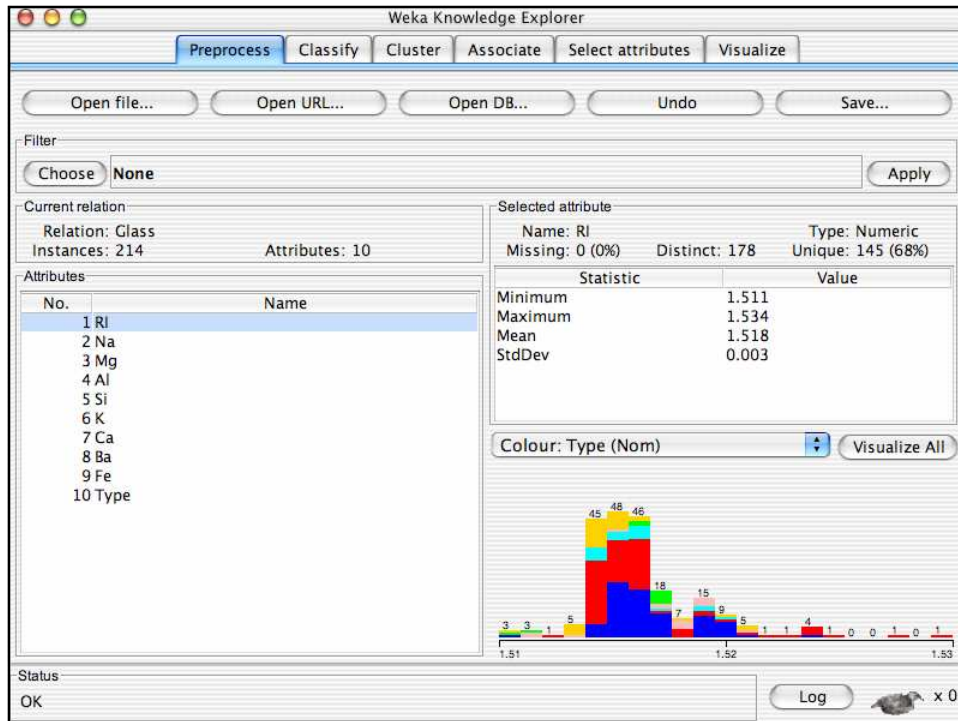


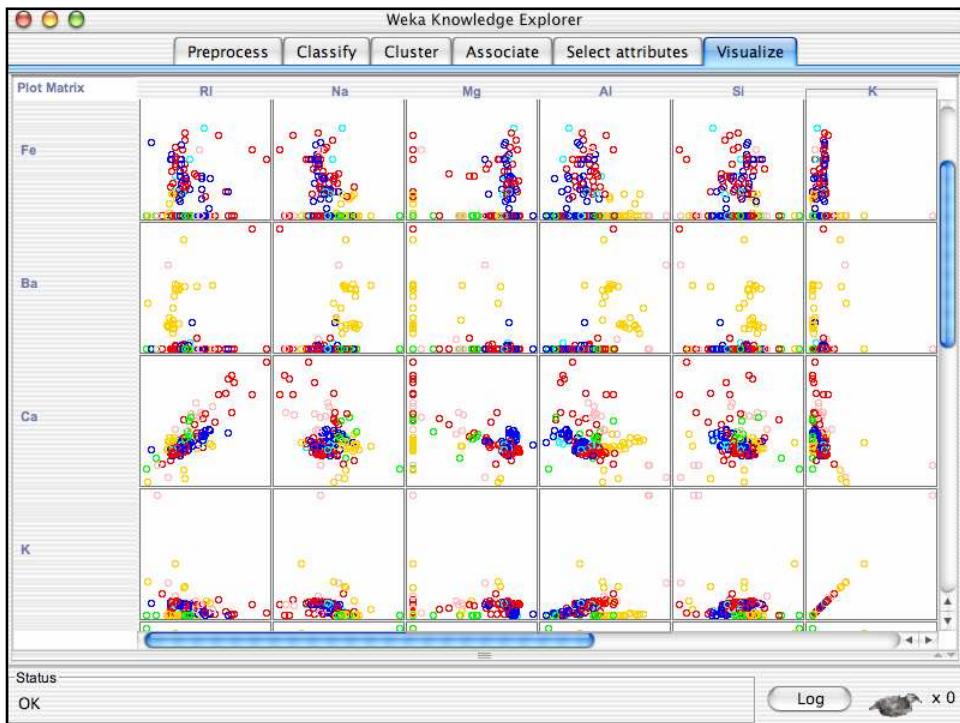
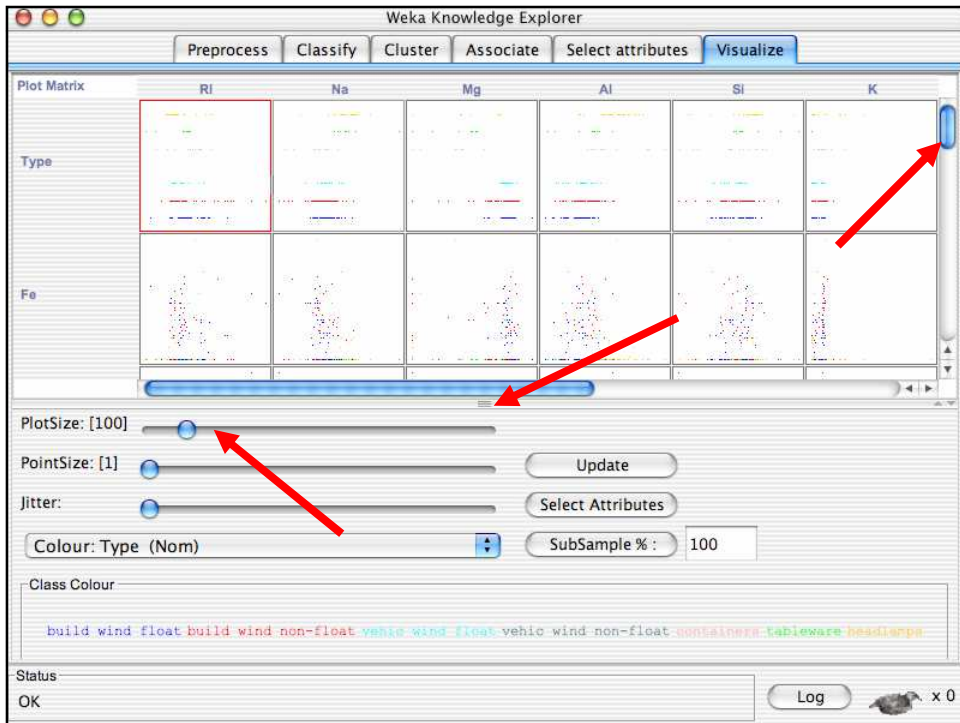


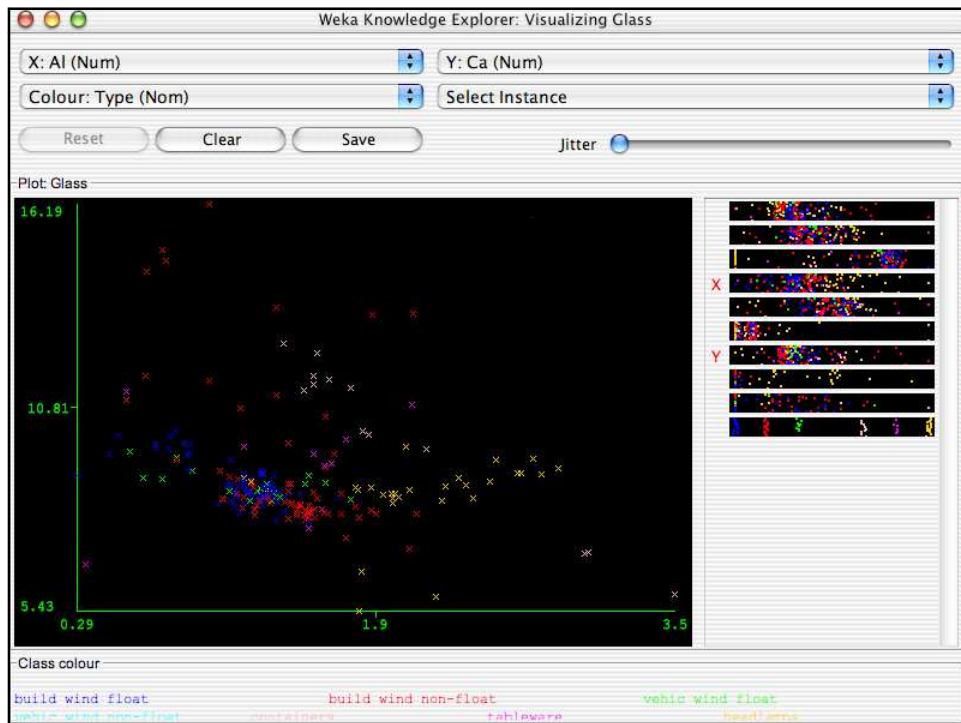
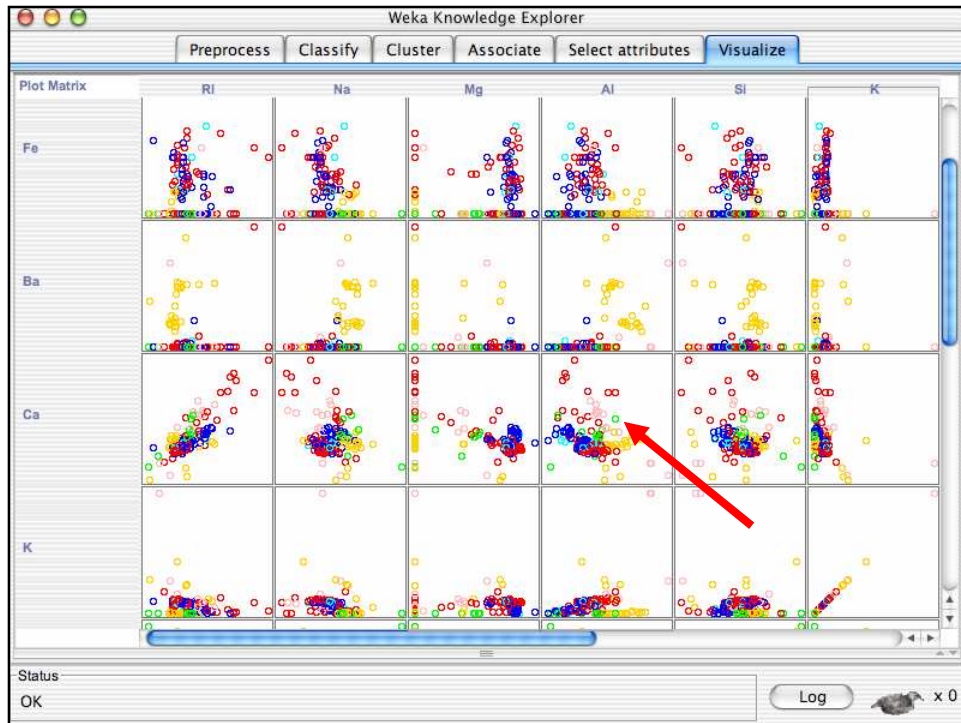
## Explorer: data visualization

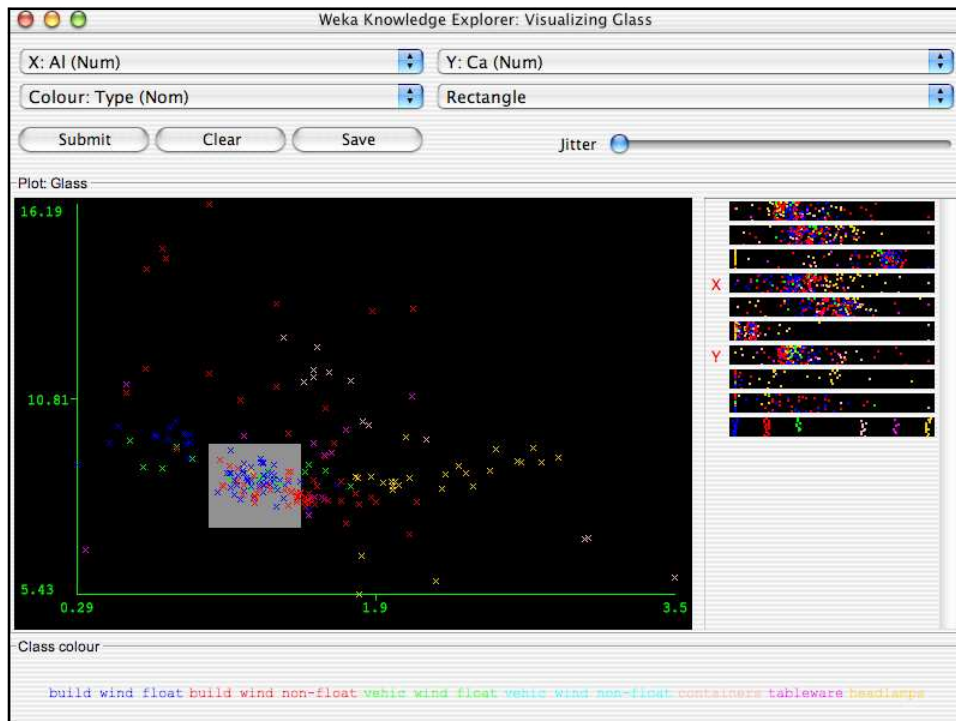
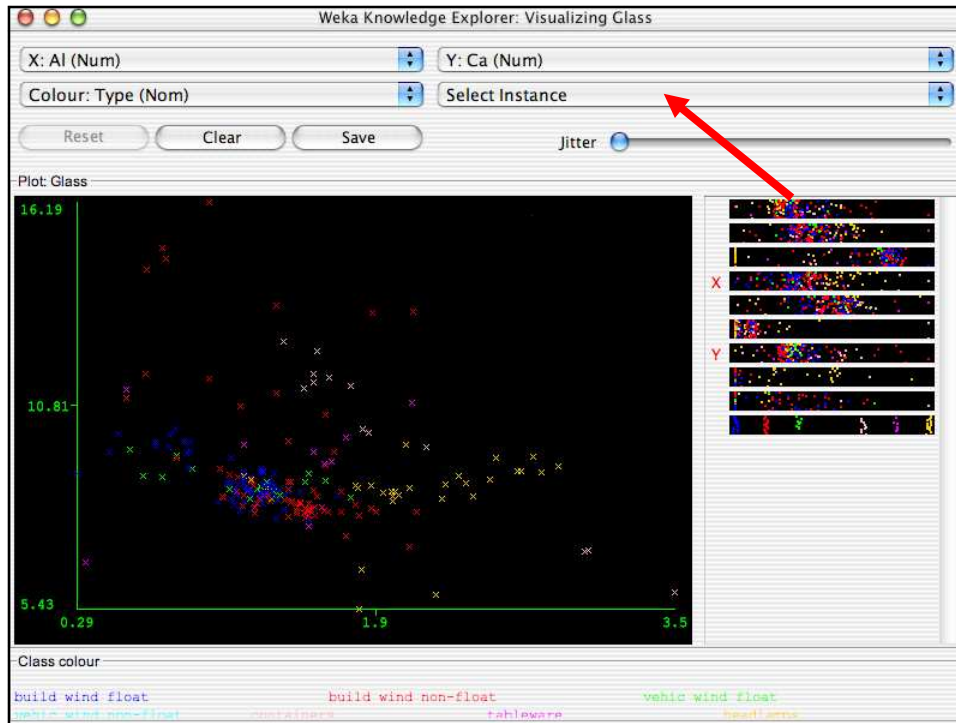
- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
  - ◆ To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- “Jitter” option to deal with nominal attributes (and to detect “hidden” data points)
- “Zoom-in” function

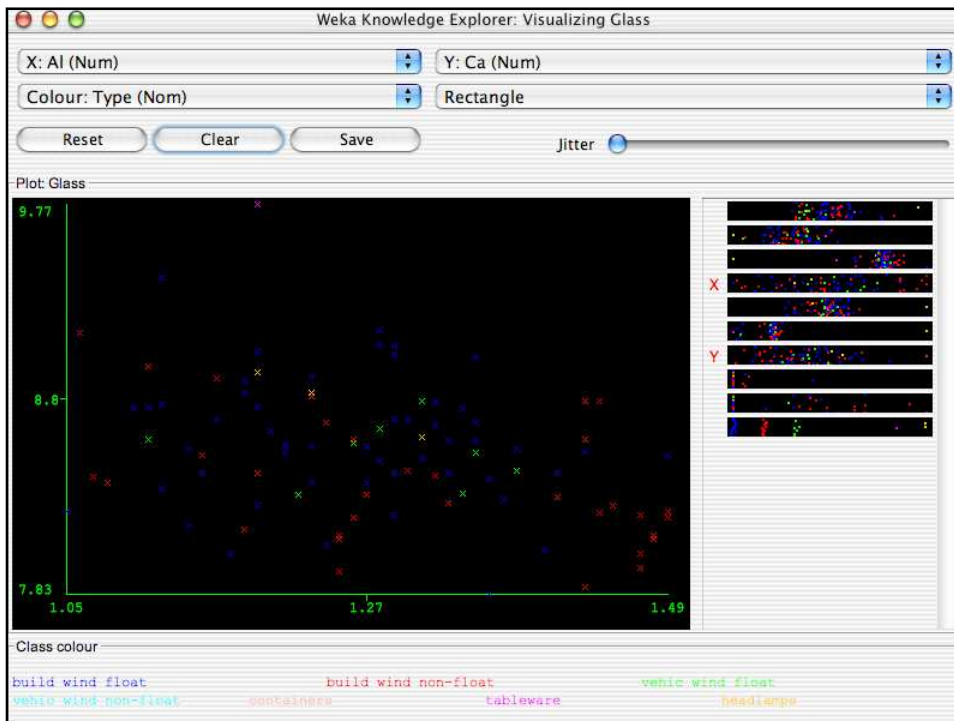
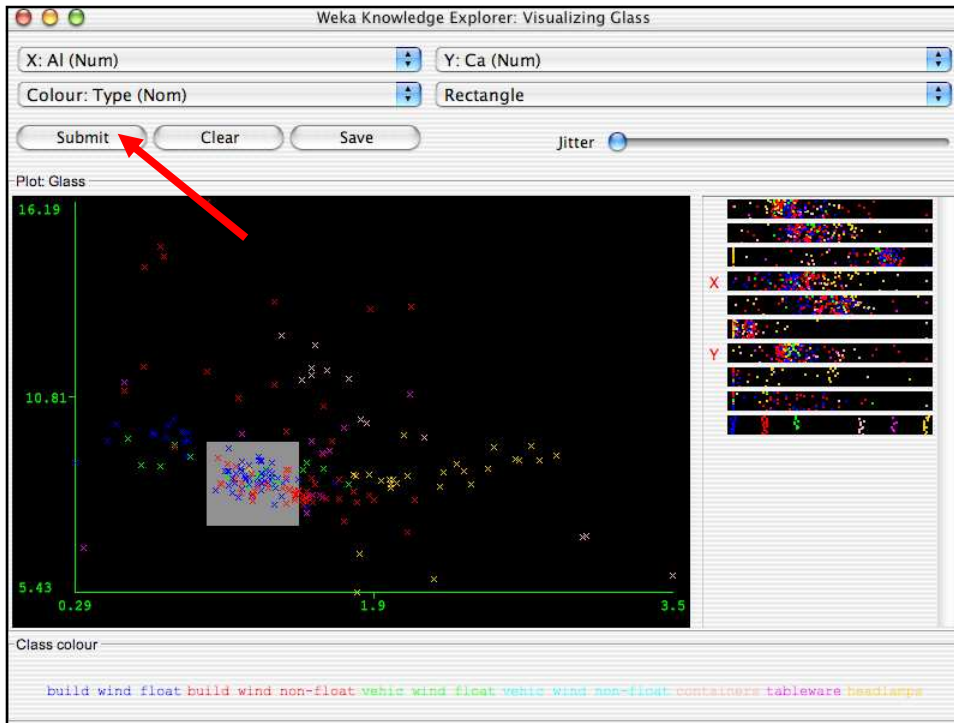














Weka GUI Chooser  
Waikato Environment for Knowledge Analysis  
(c) 1999 – 2003 University of Waikato New Zealand

GUI  
Simple CLI Explorer  
Experimenter KnowledgeFlow

2/4/2004 University of Waikato 136

Weka GUI Chooser  
Waikato Environment for Knowledge Analysis  
(c) 1999 – 2003 University of Waikato New Zealand

GUI  
Simple CLI Explorer  
Experimenter KnowledgeFlow

2/4/2004 University of Waikato 137

## Performing experiments

- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing built in!

2/4/2004

University of Waikato

138

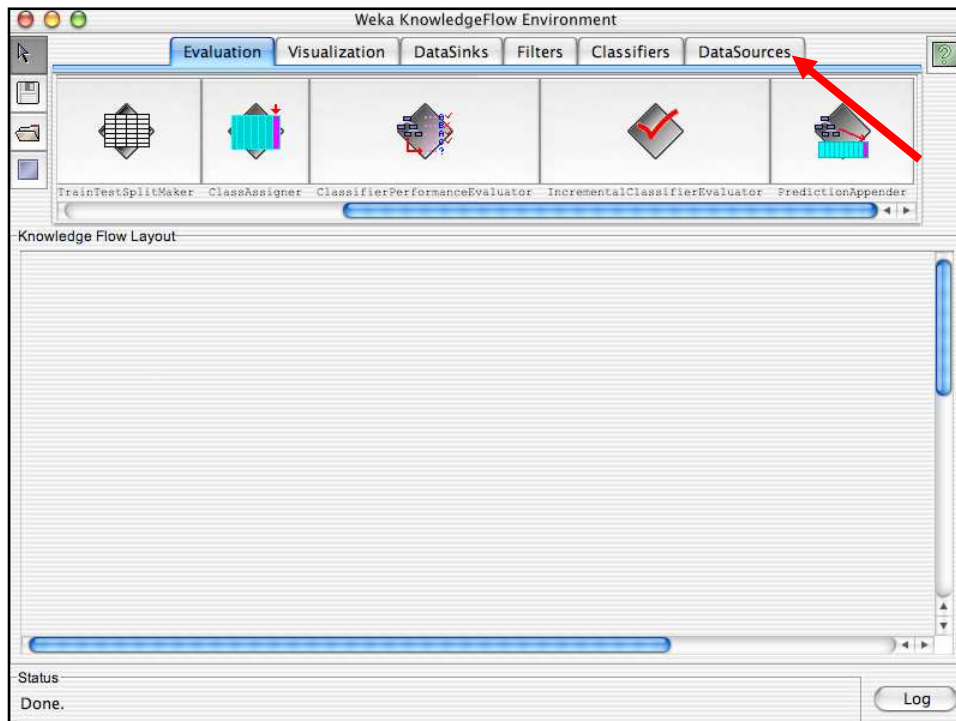
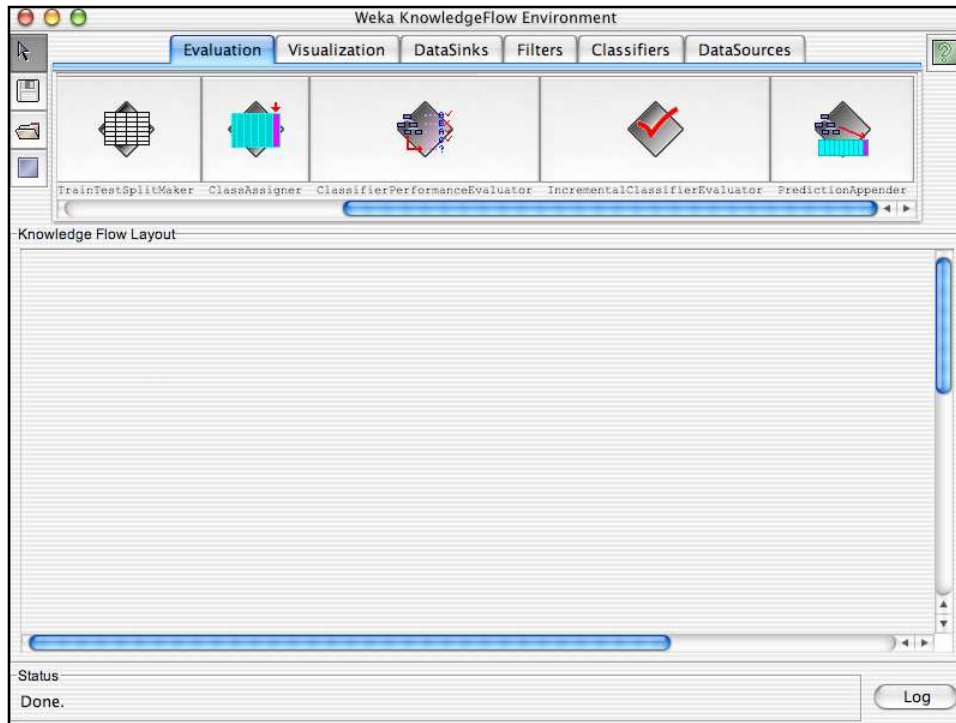
## The Knowledge Flow GUI

- New graphical user interface for WEKA
- Java-Beans-based interface for setting up and running machine learning experiments
- Data sources, classifiers, etc. are beans and can be connected graphically
- Data “flows” through components: e.g., “data source” -> “filter” -> “classifier” -> “evaluator”
- Layouts can be saved and loaded again later

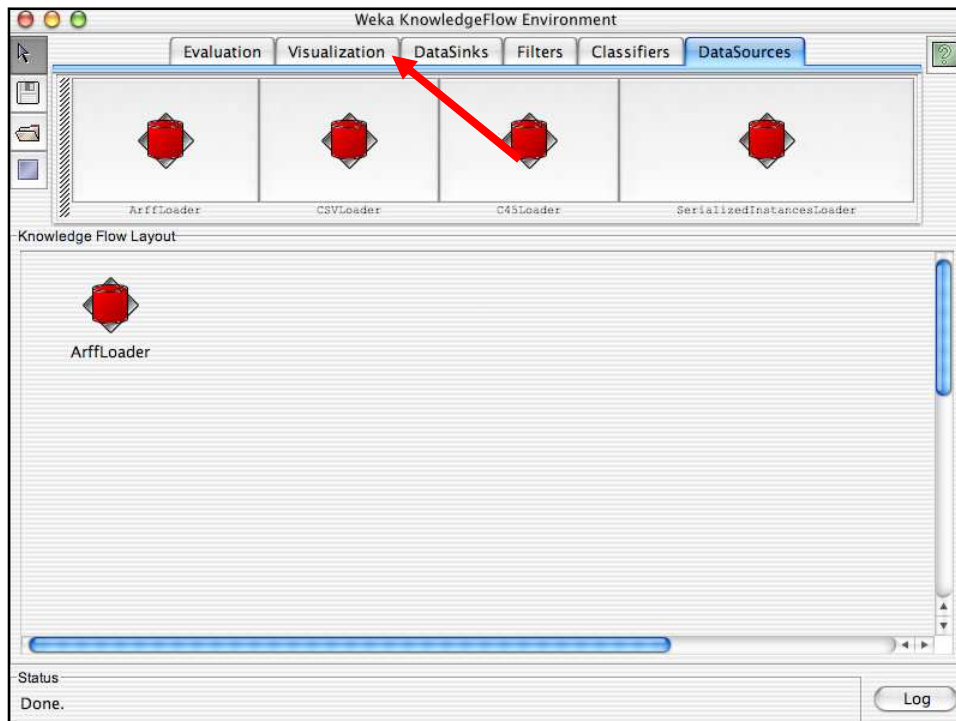
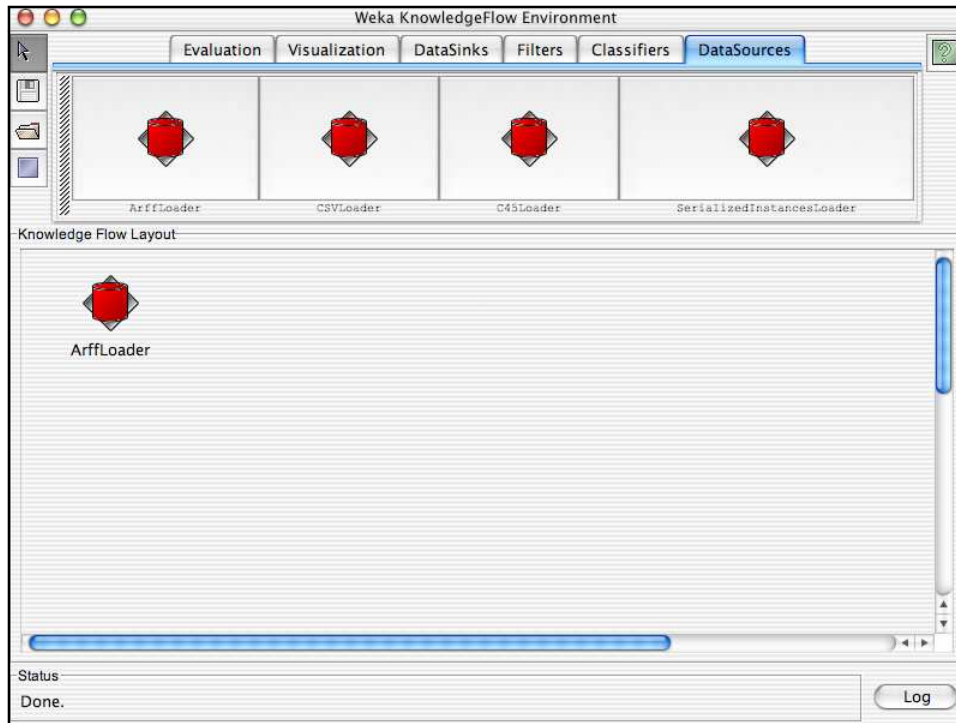
2/4/2004

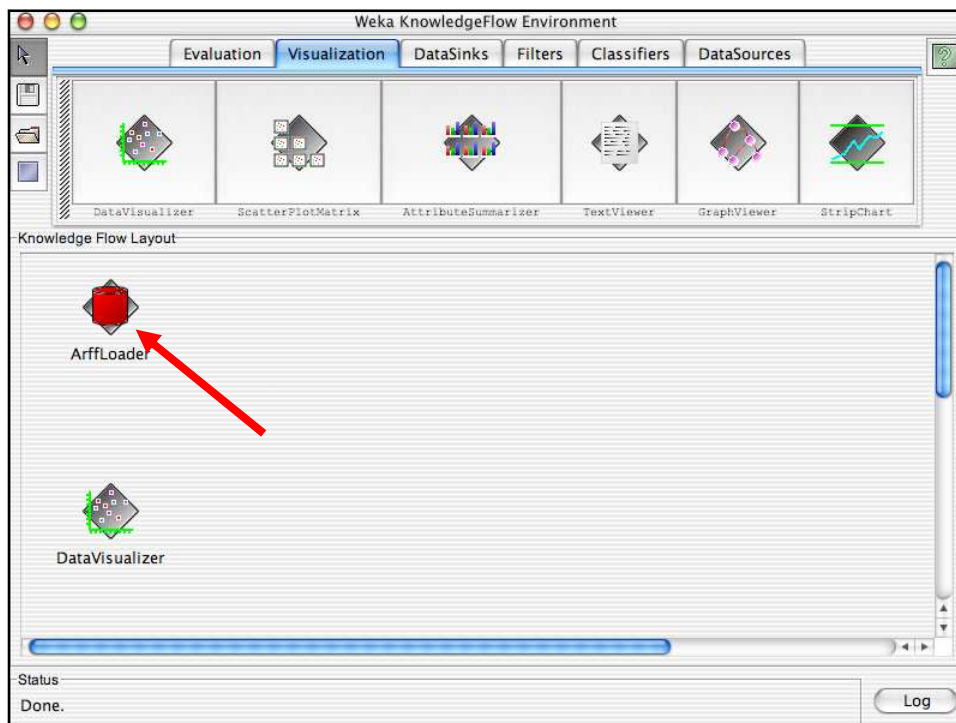
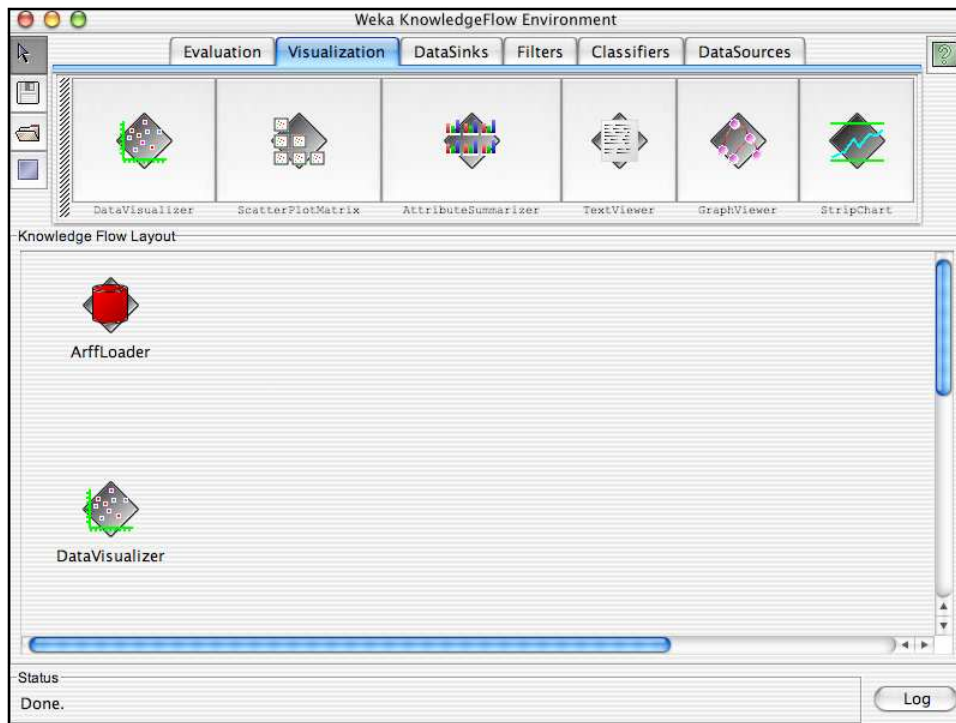
University of Waikato

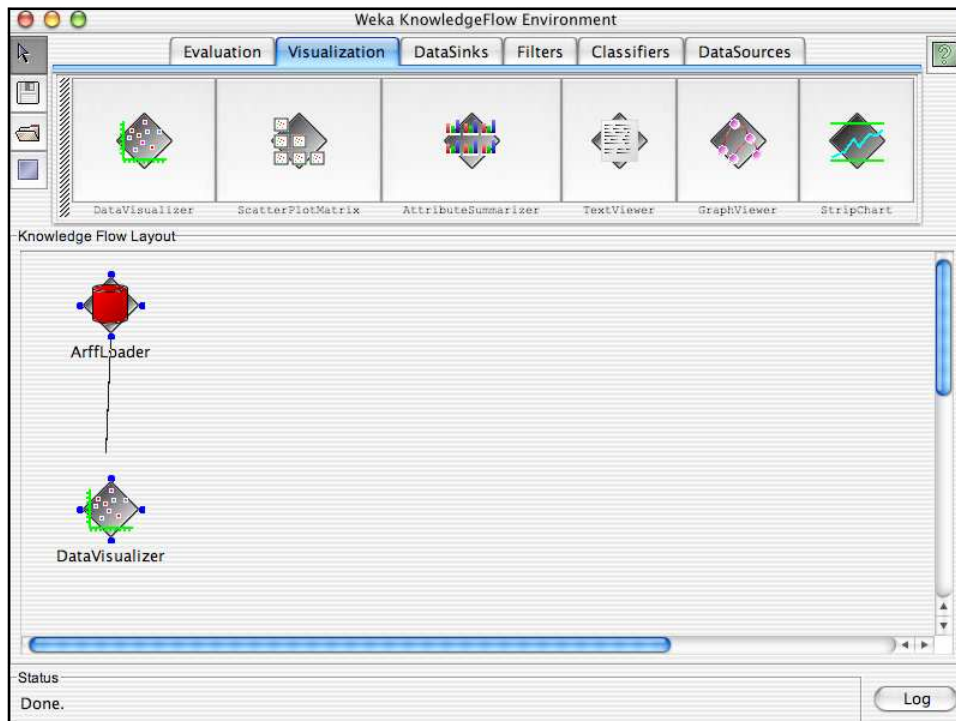
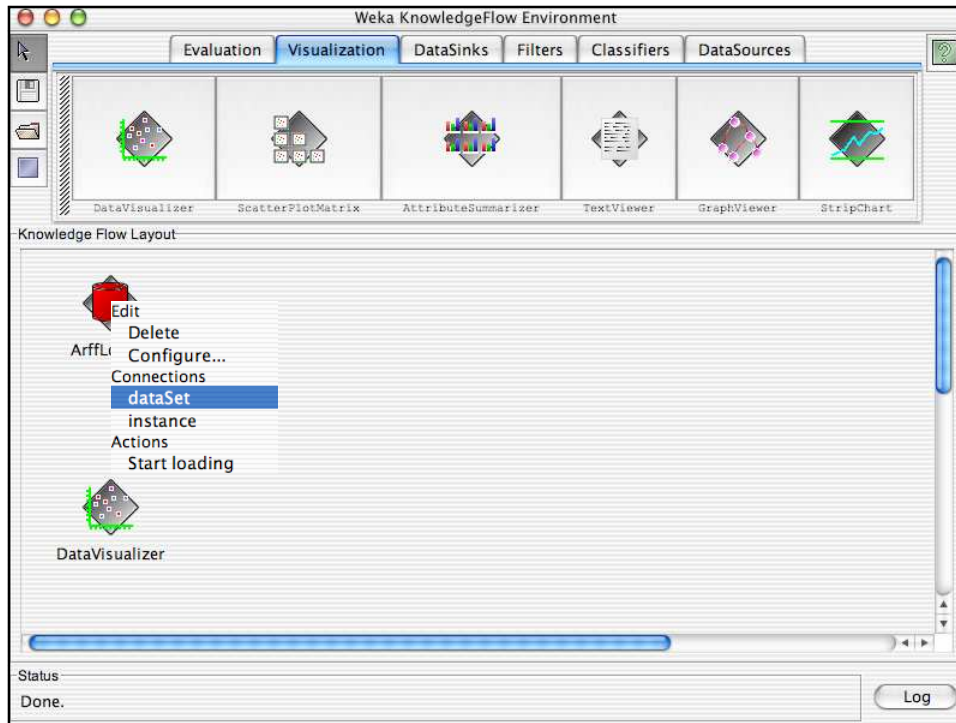
152

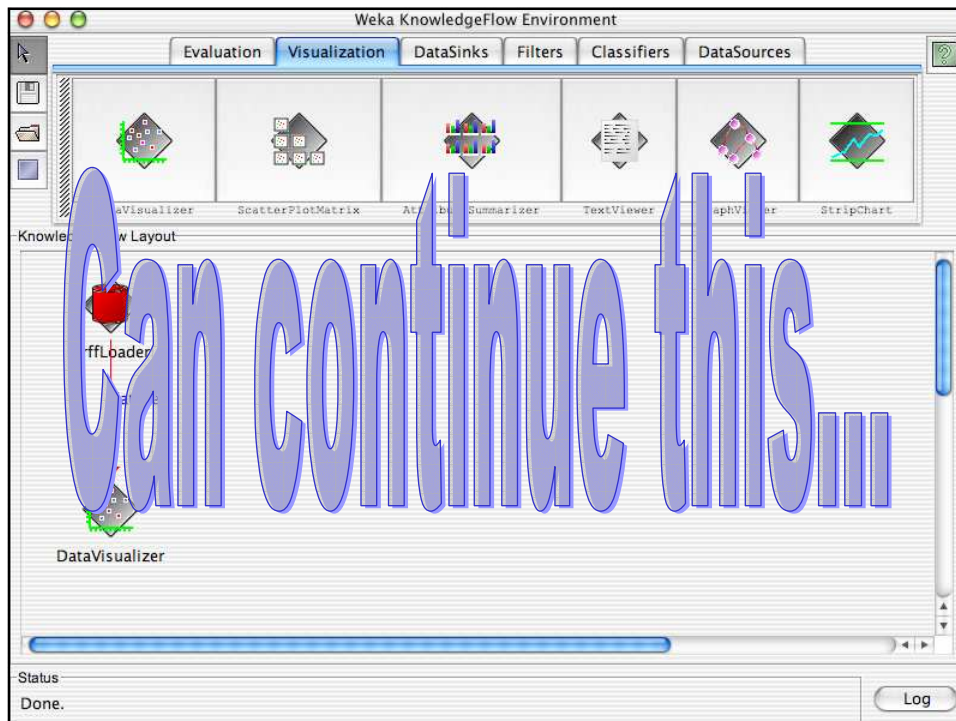
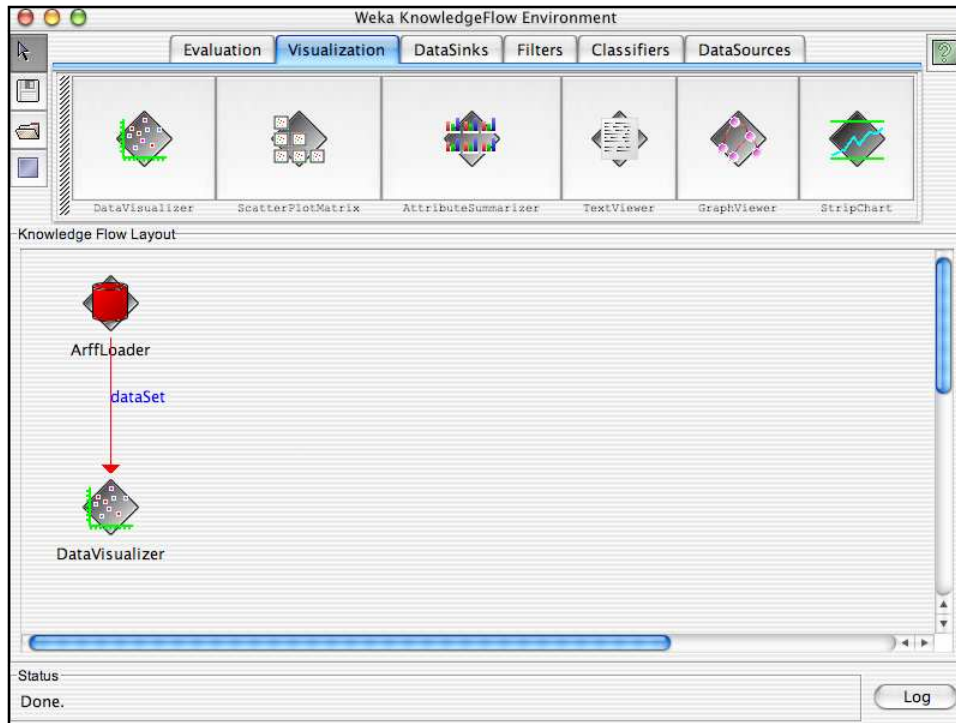


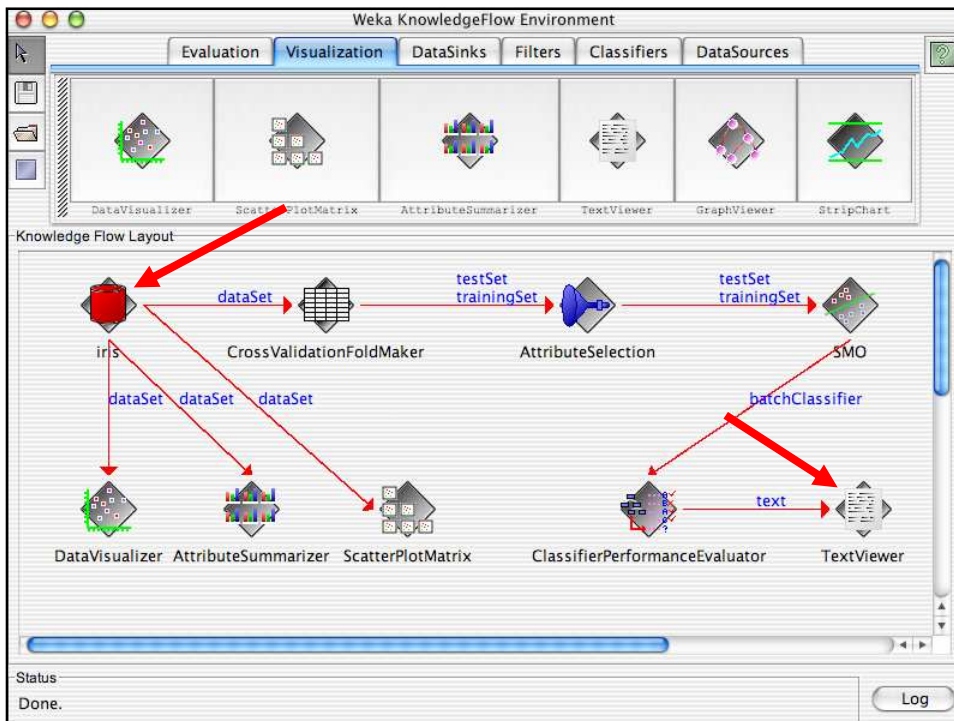
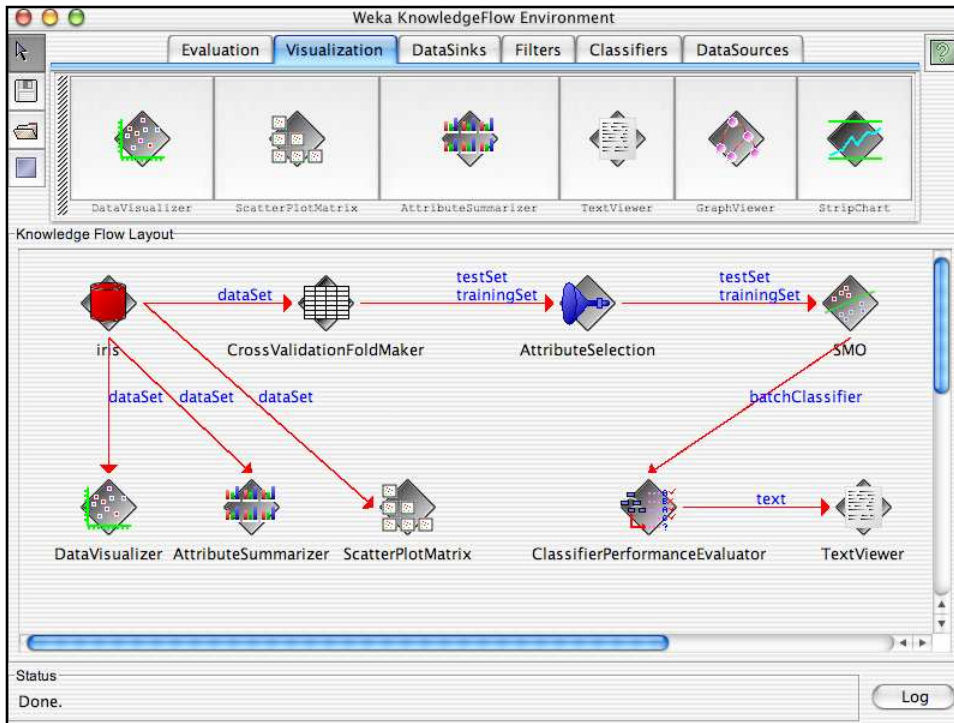


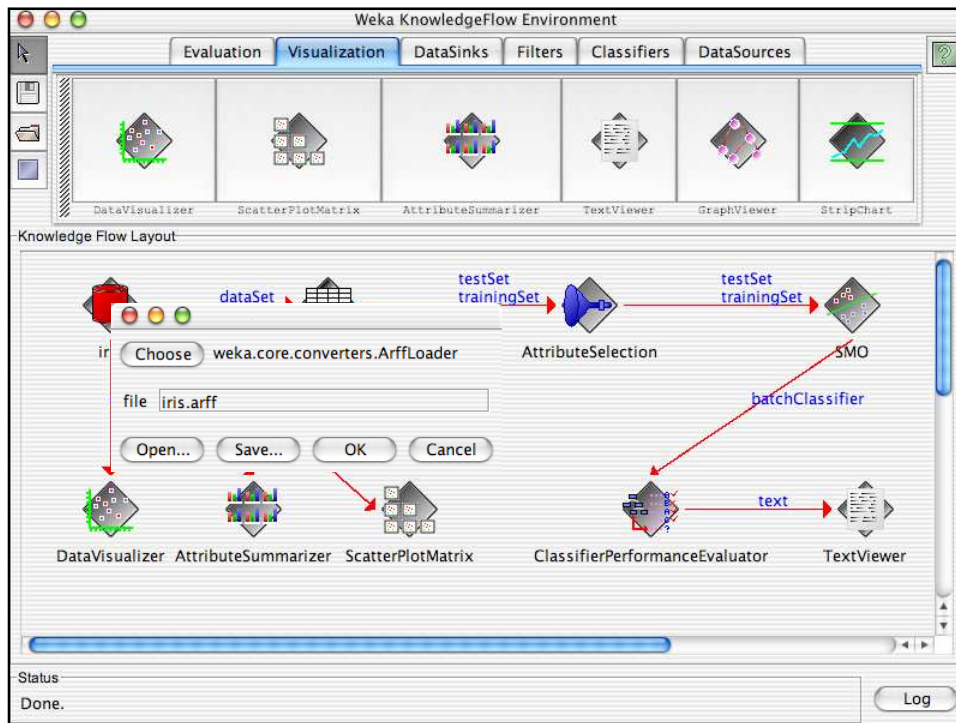
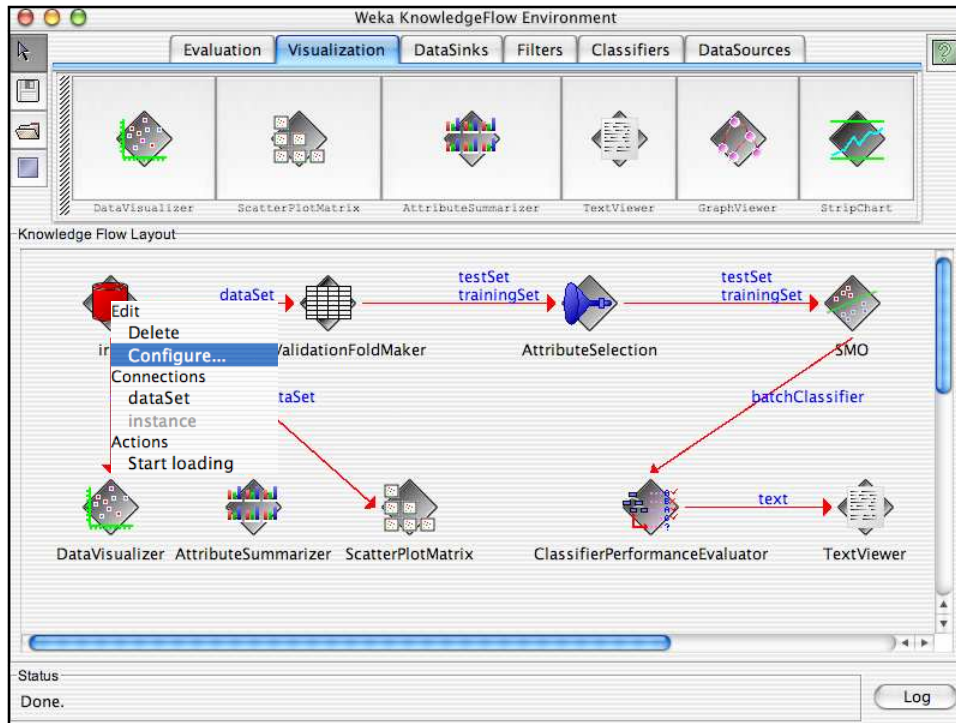




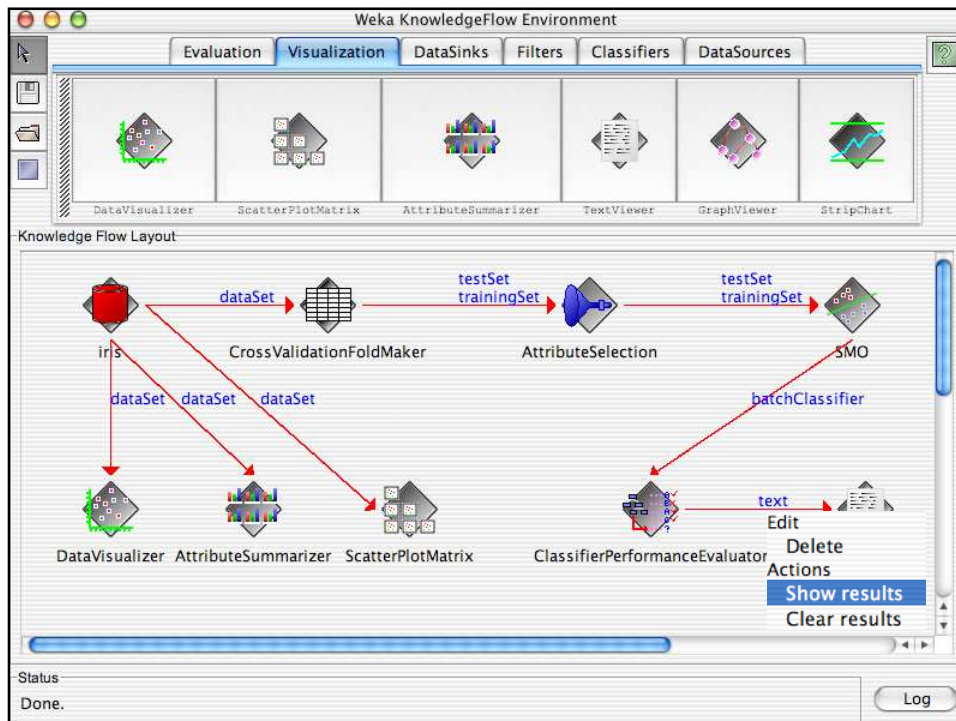
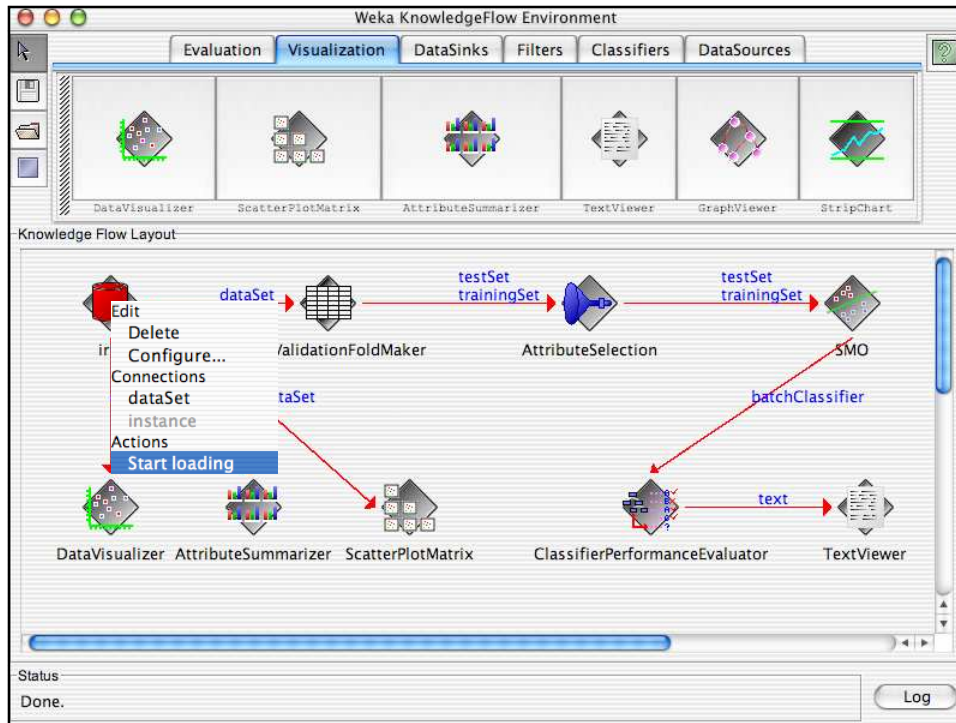




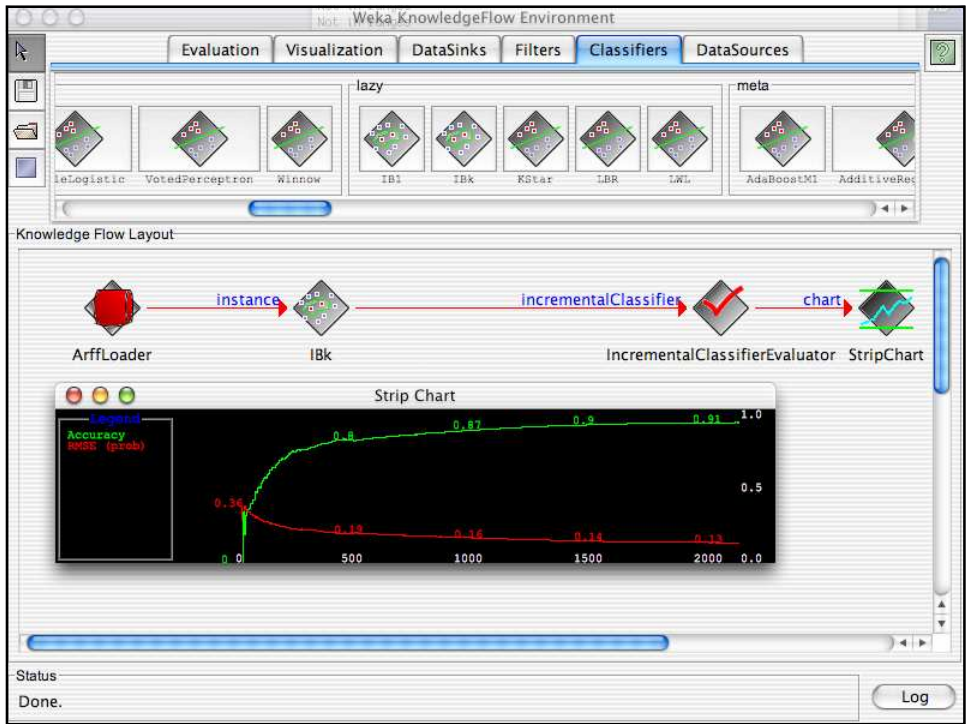
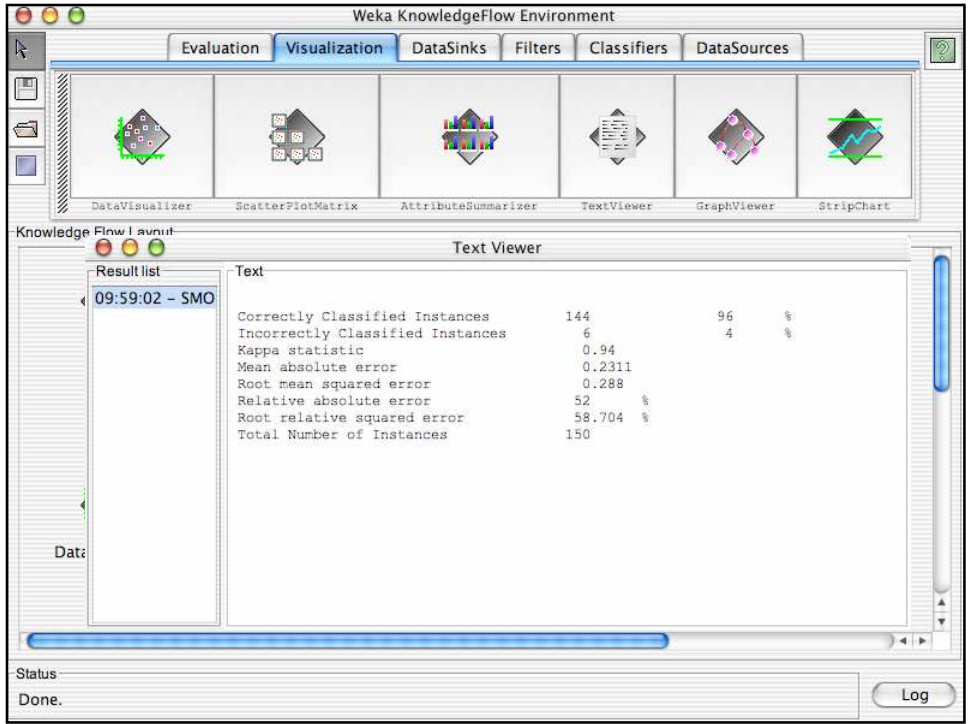












Weka GUI Chooser  
Waikato Environment for Knowledge Analysis  
(c) 1999 – 2003  
University of Waikato  
New Zealand

GUI  
Simple CLI Explorer  
Experimenter KnowledgeFlow

2/4/2004 University of Waikato 171

Weka GUI Chooser  
Waikato Environment for Knowledge Analysis  
(c) 1999 – 2003  
University of Waikato  
New Zealand

GUI  
Simple CLI Explorer  
Experimenter KnowledgeFlow

2/4/2004 University of Waikato 172

## Conclusion: try it yourself!

- WEKA is available at  
<http://www.cs.waikato.ac.nz/ml/weka>
- Also has a list of projects based on WEKA
- WEKA contributors:

Abdelaziz Mahoui, Alexander K. Seewald, Ashraf M. Kibriya, Bernhard Pfahringer , Brent Martin, Peter Flach, Eibe Frank ,Gabi Schmidberger ,Ian H. Witten , J. Lindgren, Janice Boughton, Jason Wells, Len Trigg, Lucio de Souza Coelho, Malcolm Ware, Mark Hall ,Remco Bouckaert , Richard Kirkby, Shane Butler, Shane Legg, Stuart Inglis, Sylvain Roy, Tony Voyle, Xin Xu, Yong Wang, Zihai Wang