Machine Learning with WEKA

Eibe Frank

Department of Computer Science,
University of Waikato, New Zealand

WEKA: the bird

Copyright: Martin Kramer (mkramer@wxs.nl)

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

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)

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
...
```

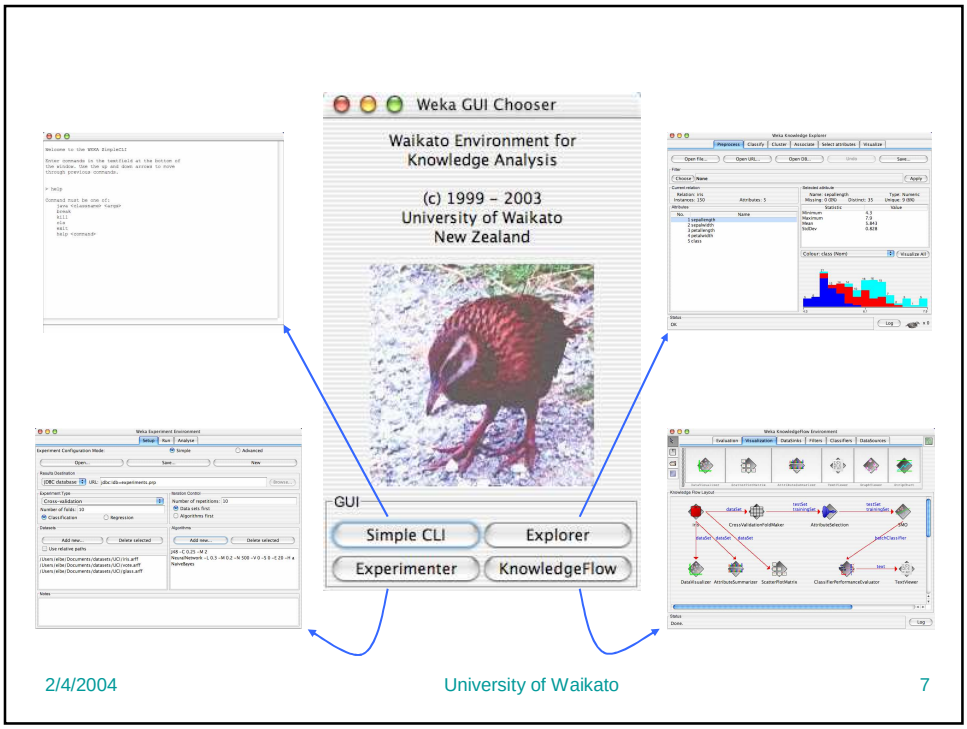
Flat file in ARFF format

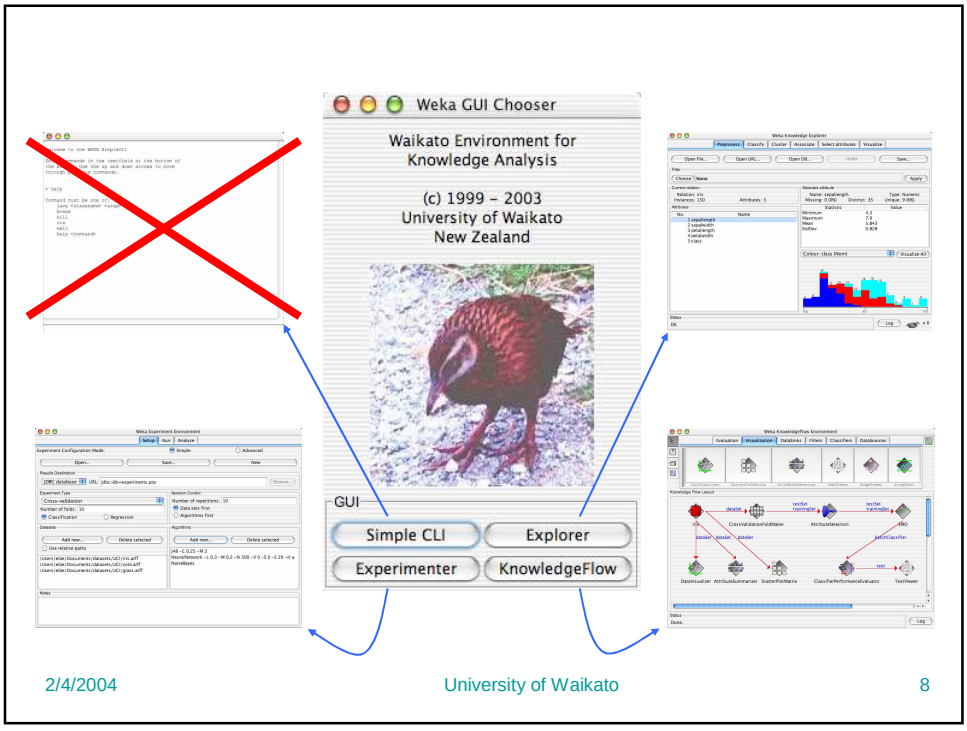
WEKA only deals with "flat" files

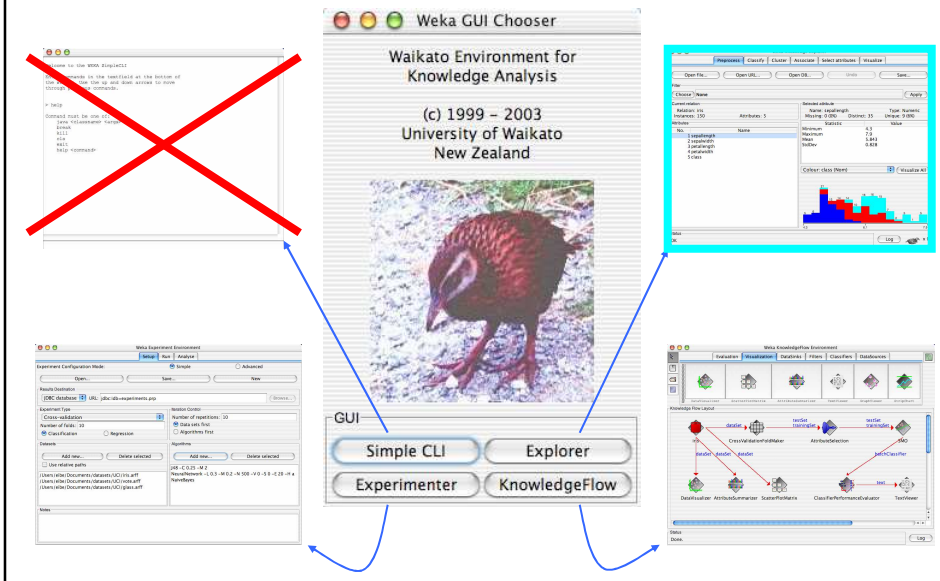
```
@relation heart-disease-simplified

@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
...
```

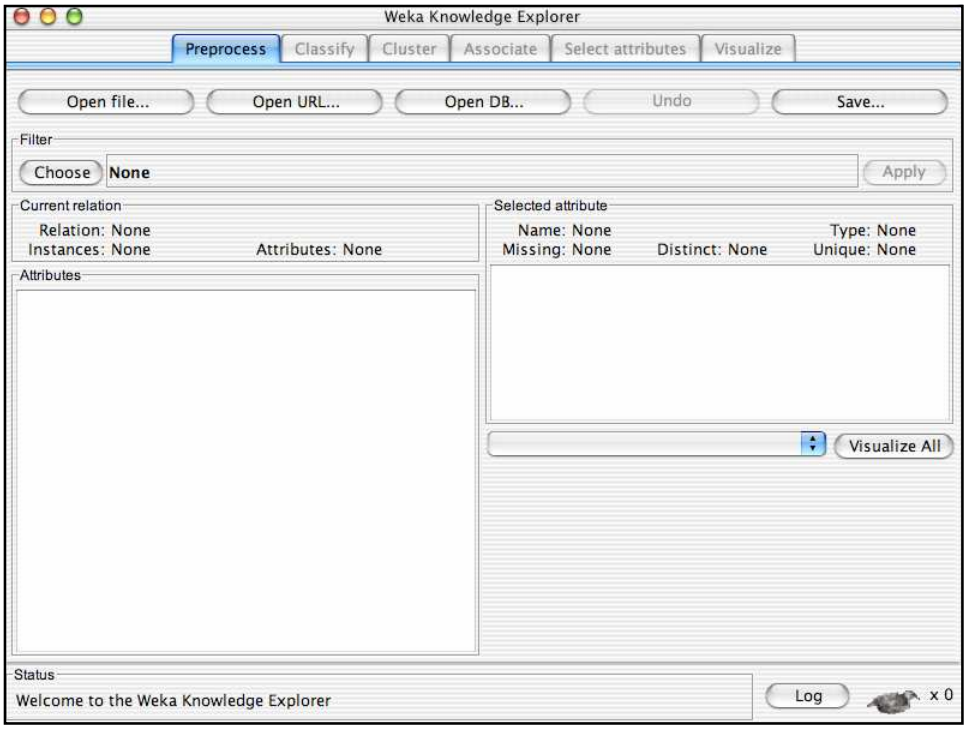

Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow
2/4/2004
University of Waikato
7

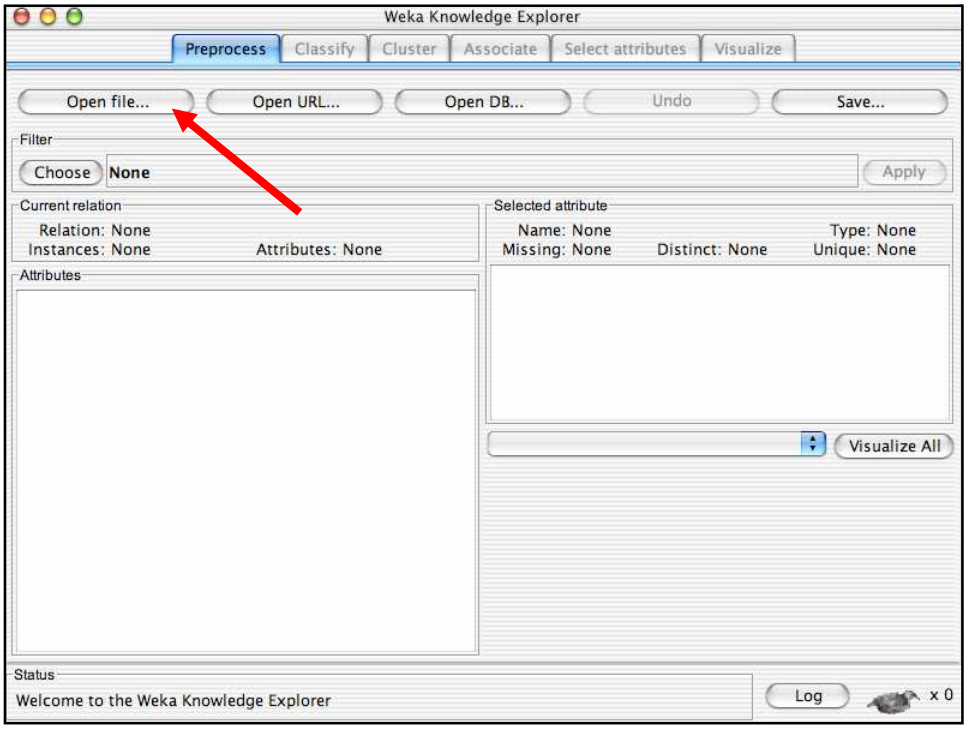

Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow
2/4/2004
University of Waikato
8

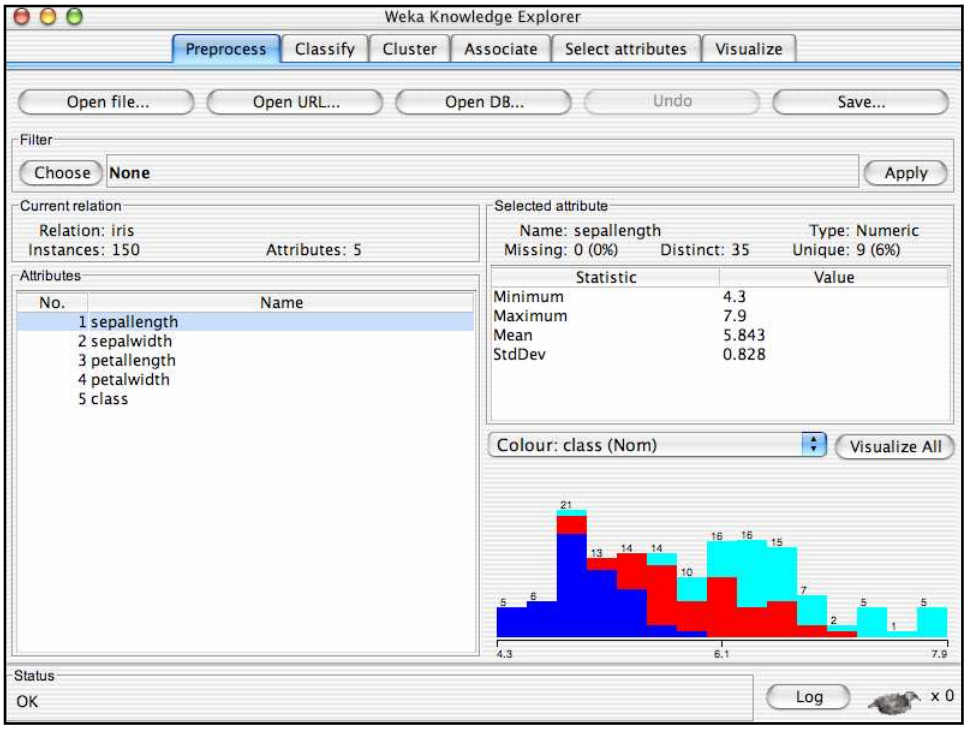


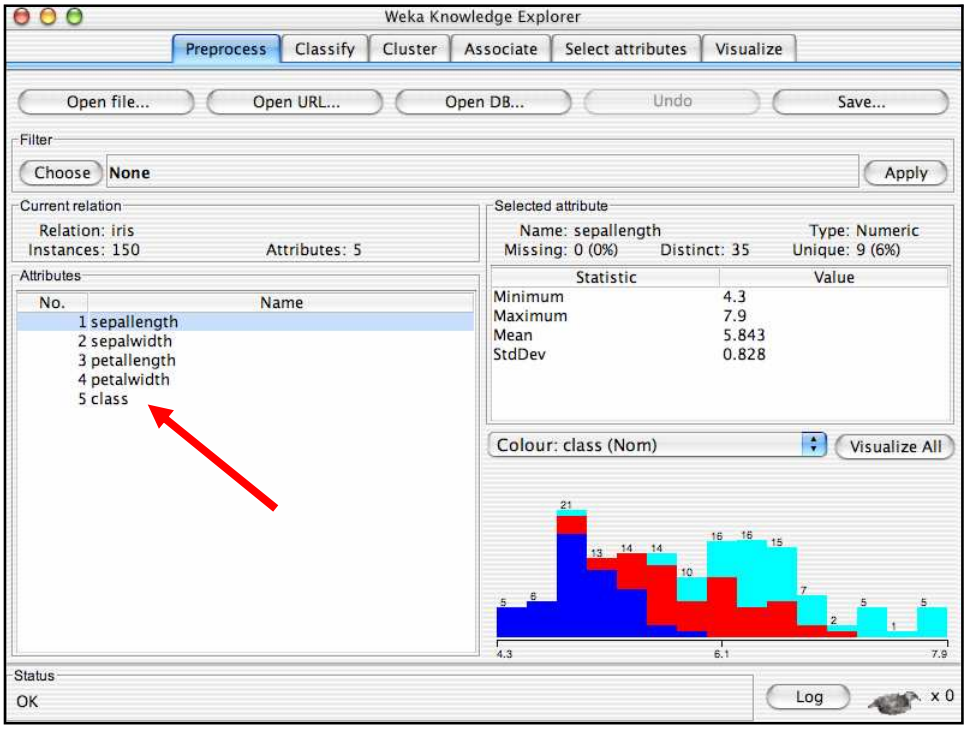
Explorer: pre-processing the data

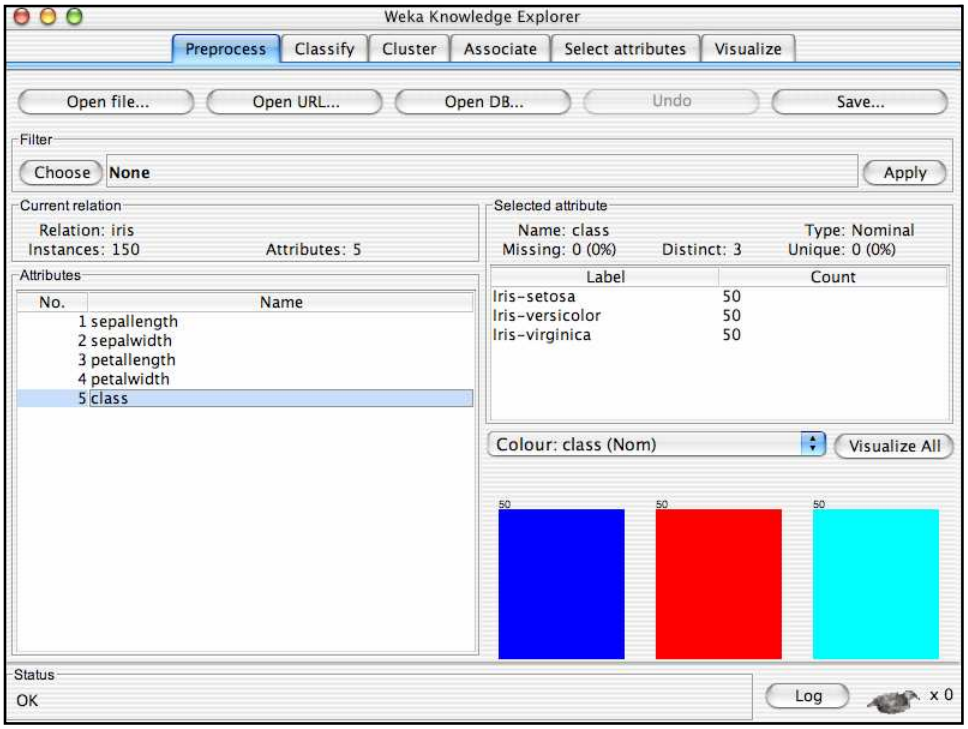
- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary
- Data can also be read from a URL or from an SQL database (using JDBC)
- Pre-processing tools in WEKA are called “filters”
- WEKA contains filters for:
  - Discretization, normalization, resampling, attribute selection, transforming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

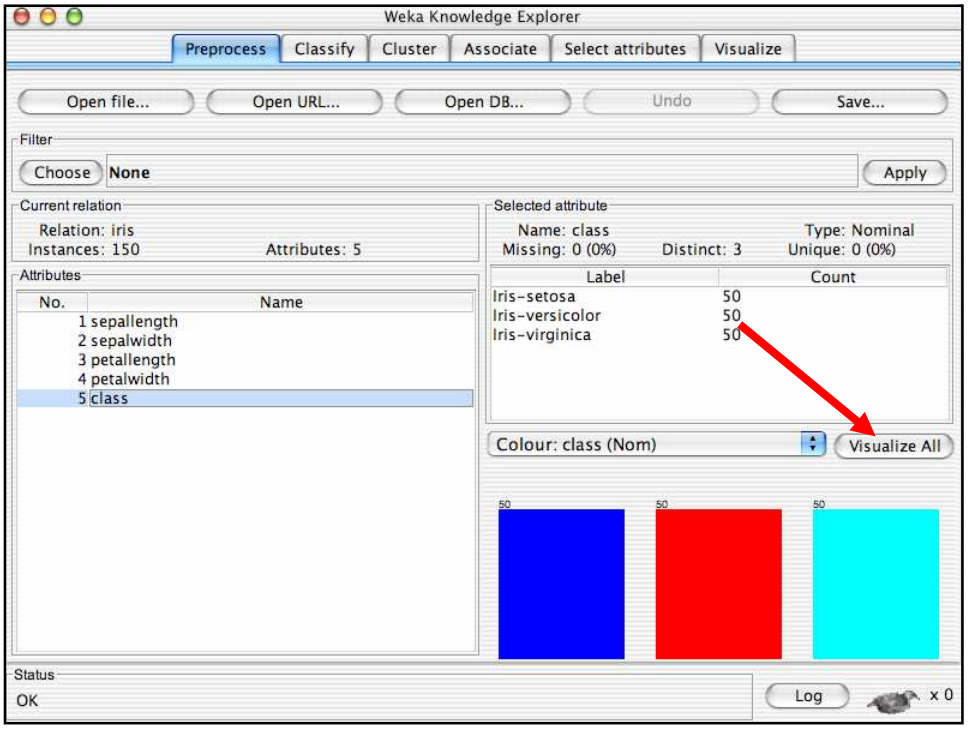

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None
Attributes
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0

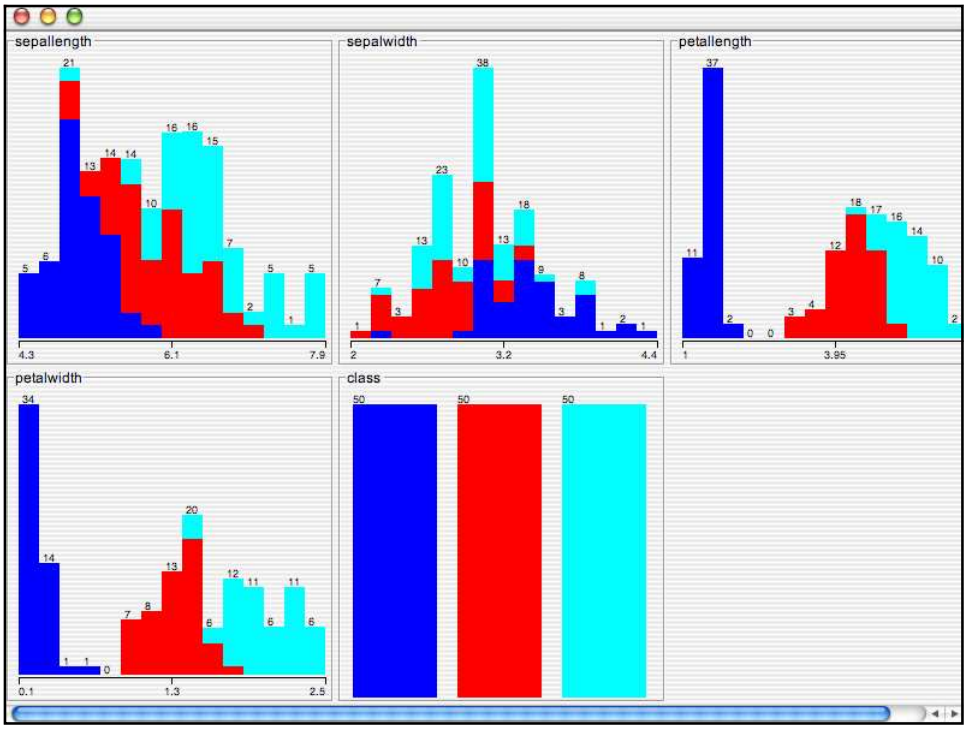

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None
Attributes
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0

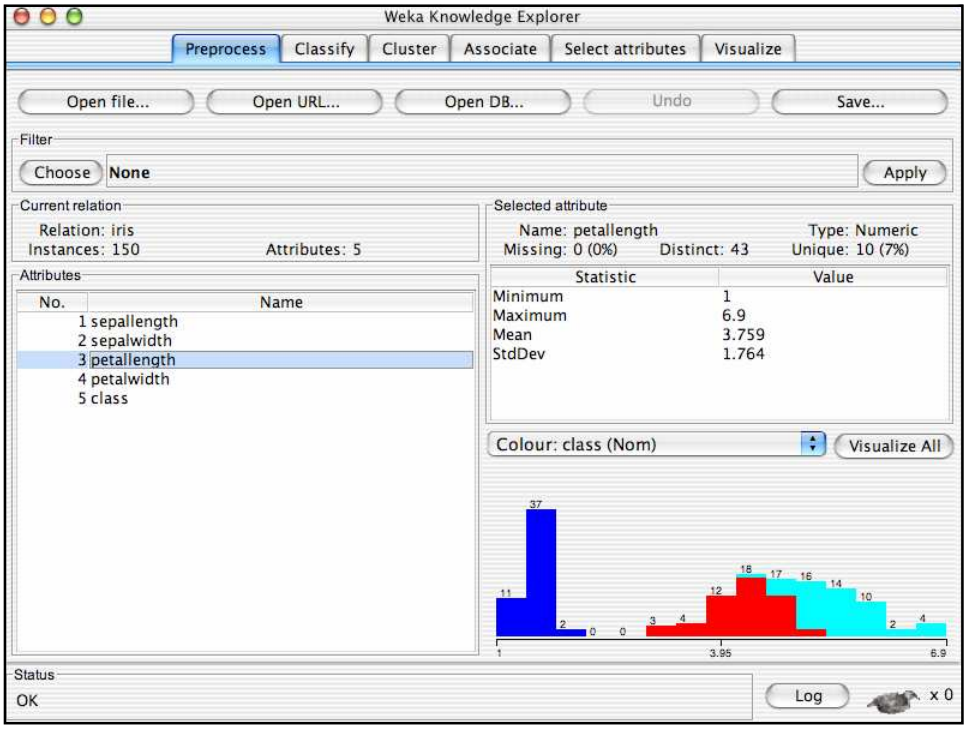

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Colour: class (Nom)
Visualize All
4.3
6.1
7.9
Status
OK
Log
x 0

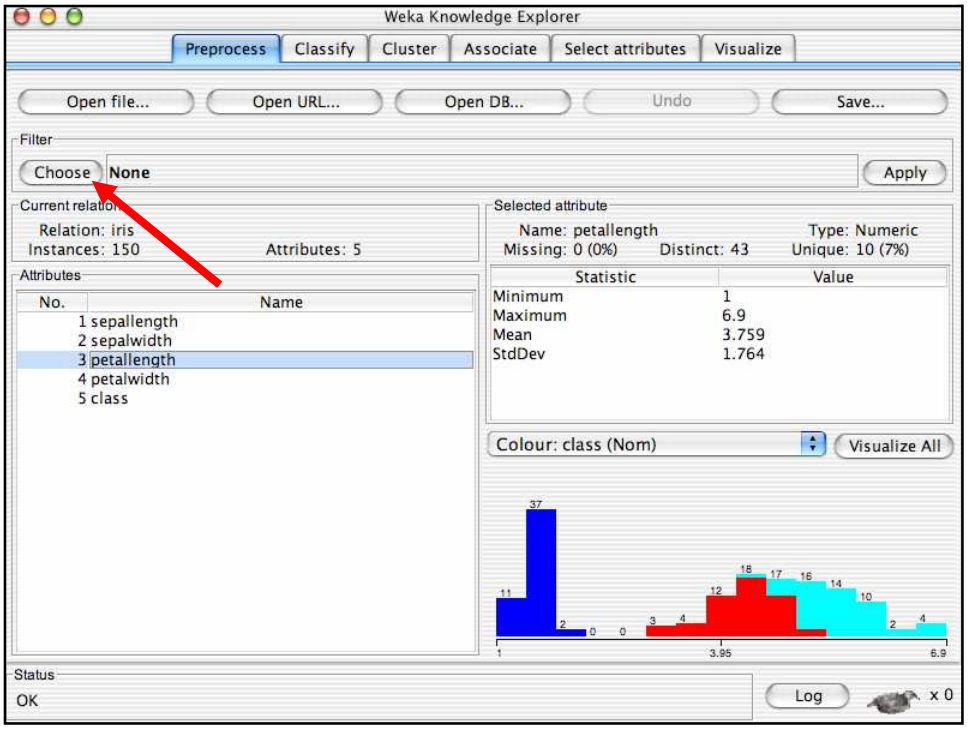

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: sepallength
Type: Numeric
Missing: 0 (0%)
Distinct: 35
Unique: 9 (6%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
StdDev
0.828
Colour: class (Nom)
Visualize All
4.3
6.1
7.9
Status
OK
Log
x 0

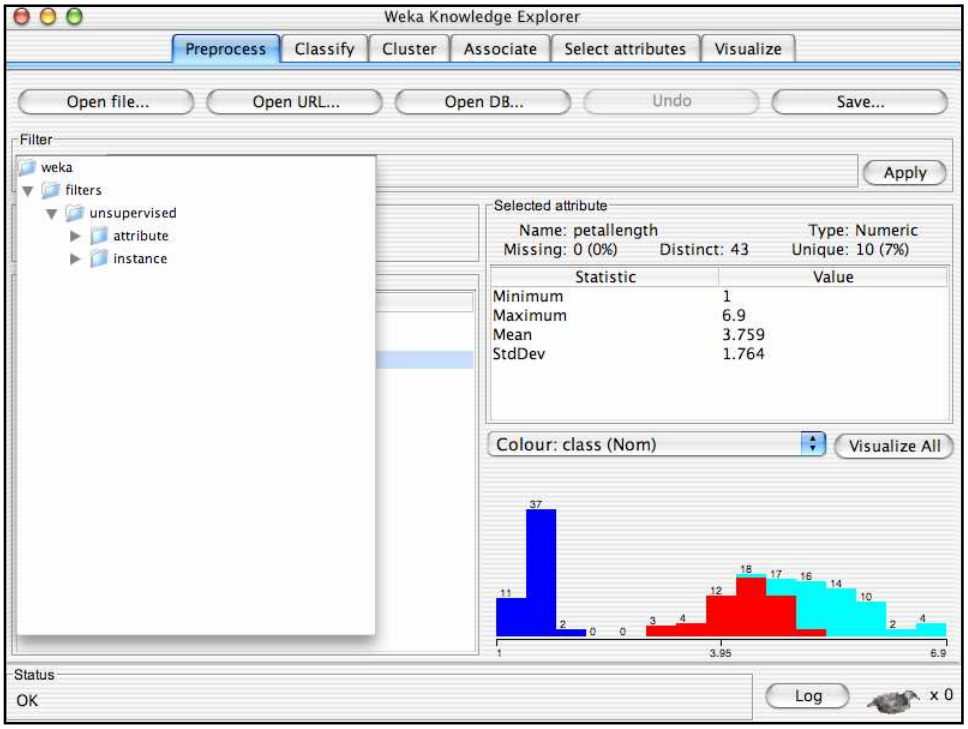

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Label
Count
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
Colour: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0

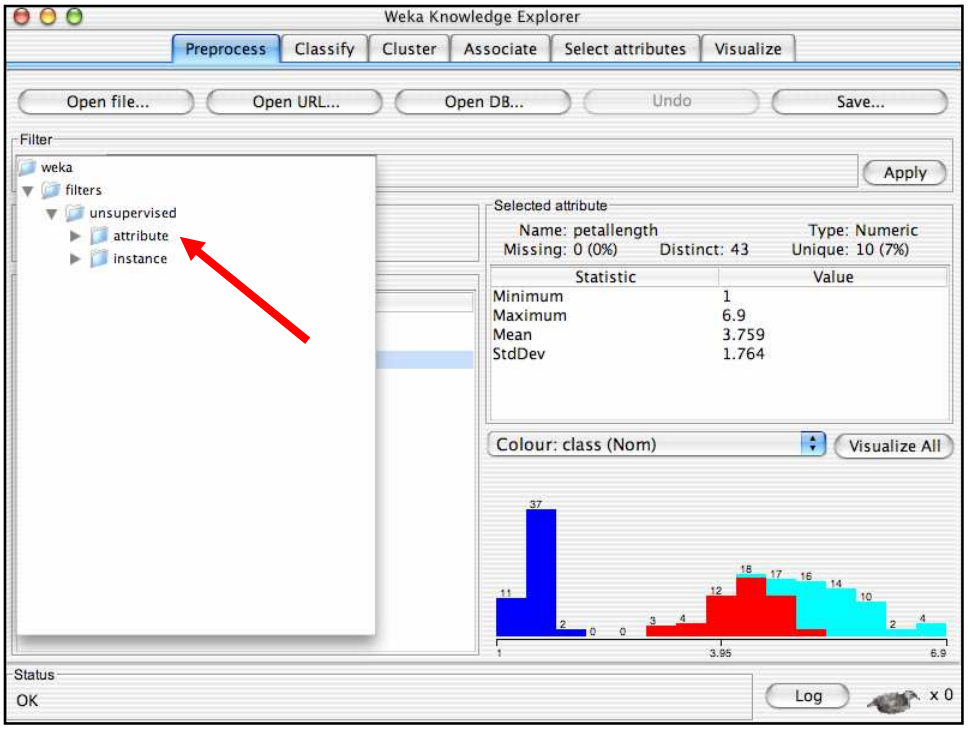

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Label
Count
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
Colour: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0

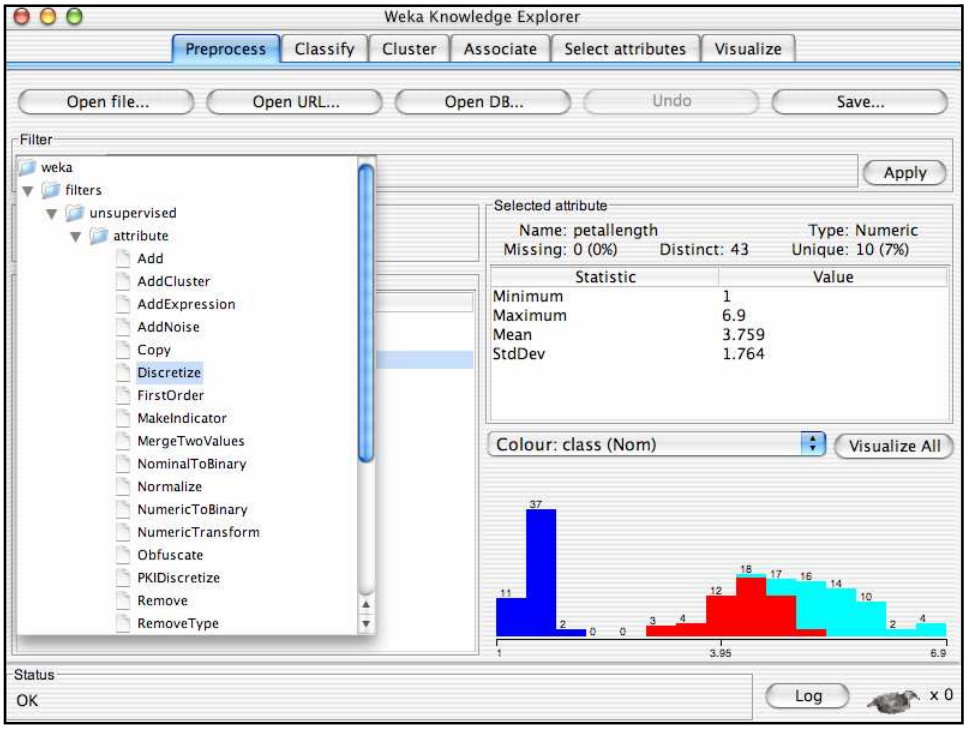

sepallength
sepalwidth
petallength
petalwidth
class

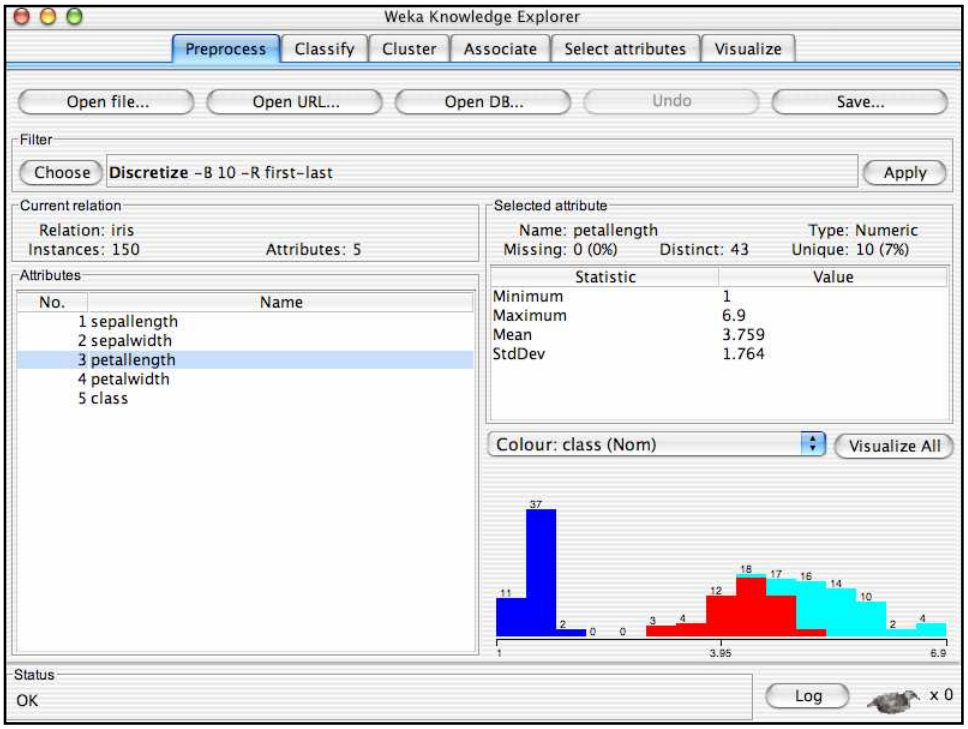

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

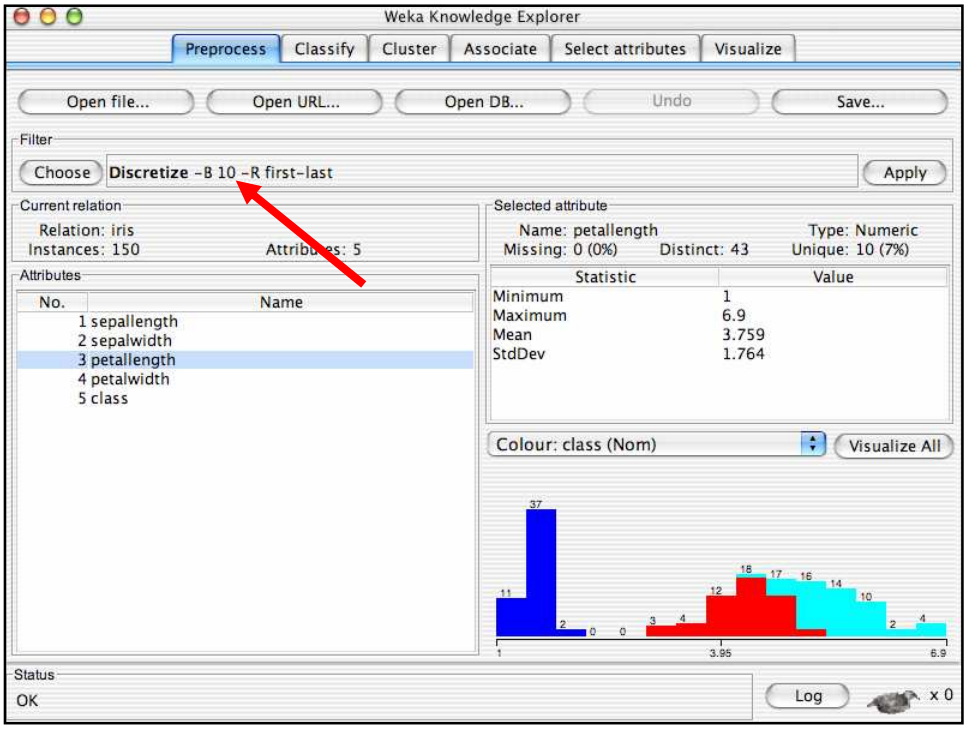

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

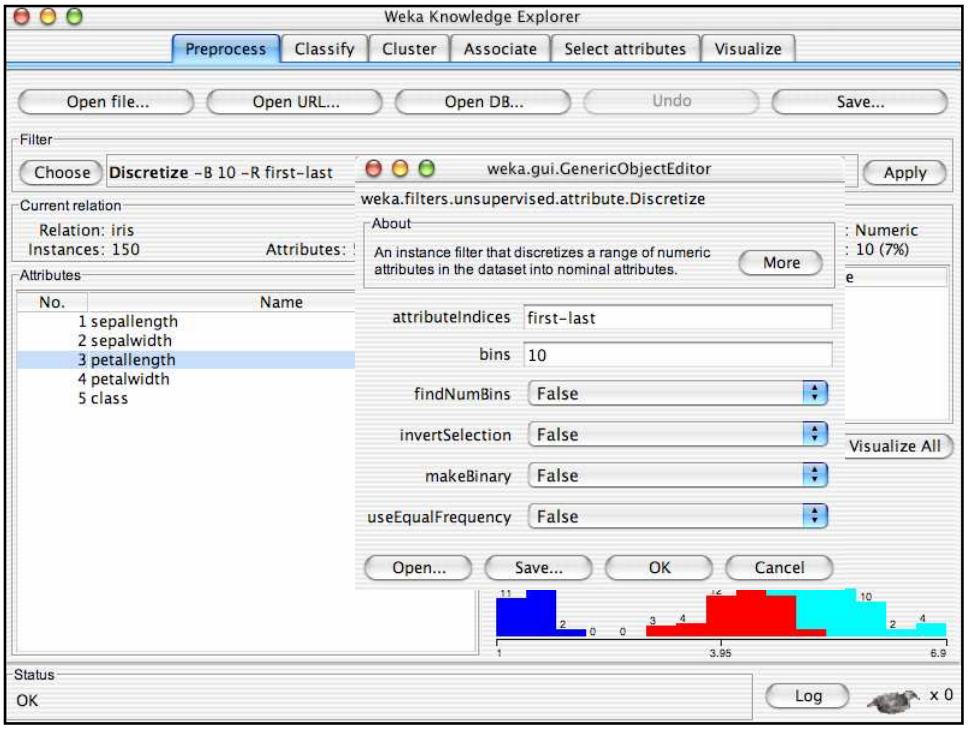

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
instance
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

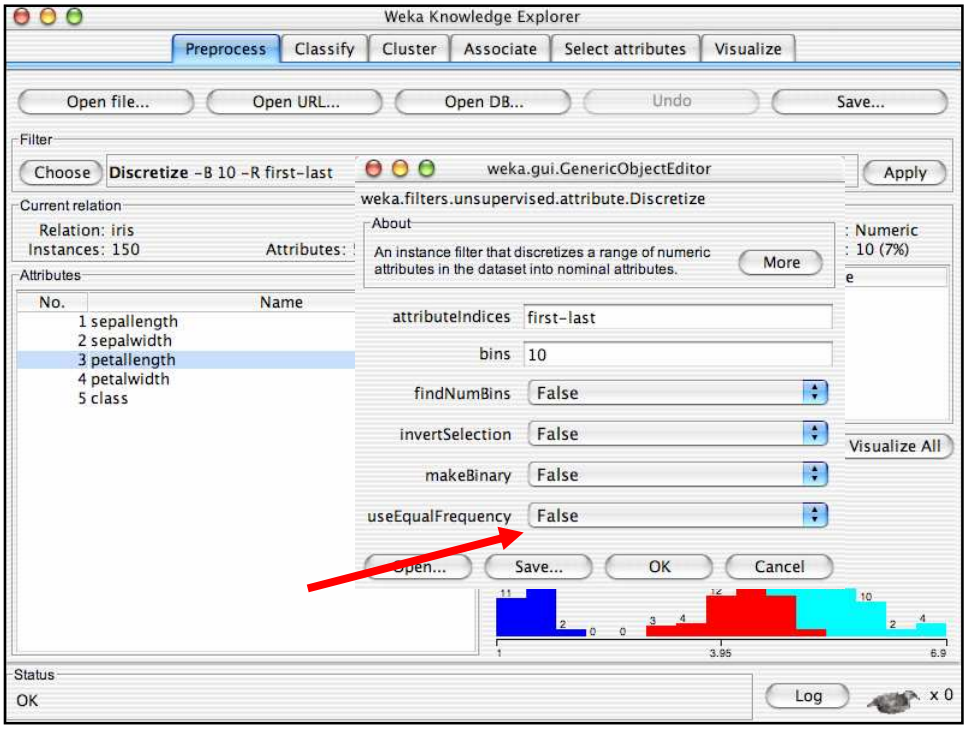

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
instance
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

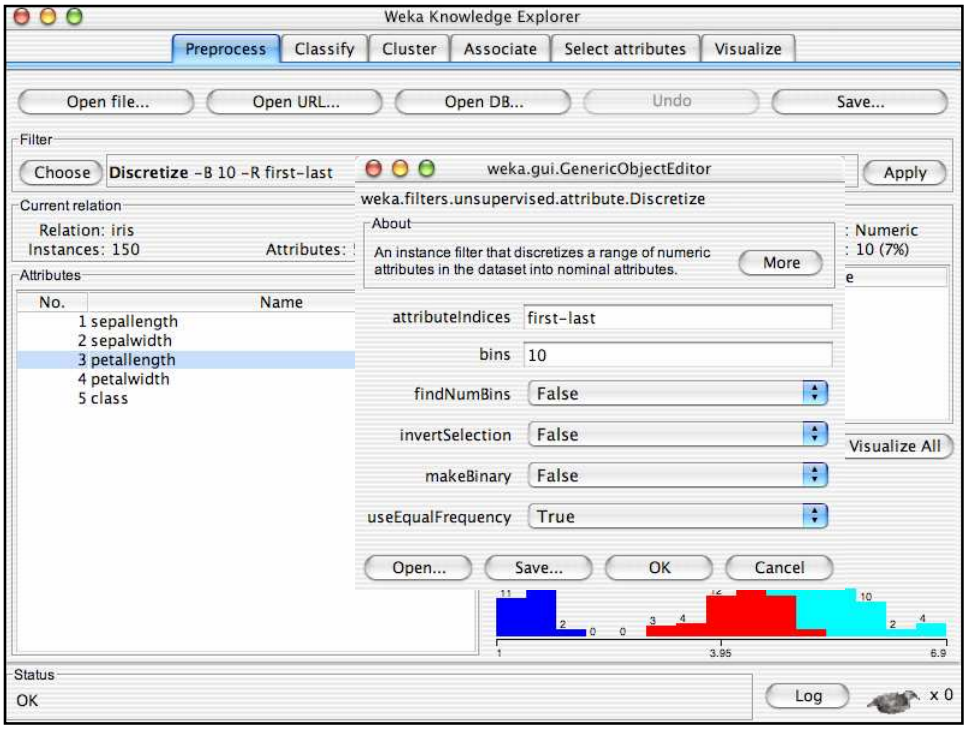

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
Add
AddCluster
AddExpression
AddNoise
Copy
Discretize
FirstOrder
MakeIndicator
MergeTwoValues
NominalToBinary
Normalize
NumericToBinary
NumericTransform
Obfuscate
PKIDiscretize
Remove
RemoveType
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
Status
OK
Log
x 0

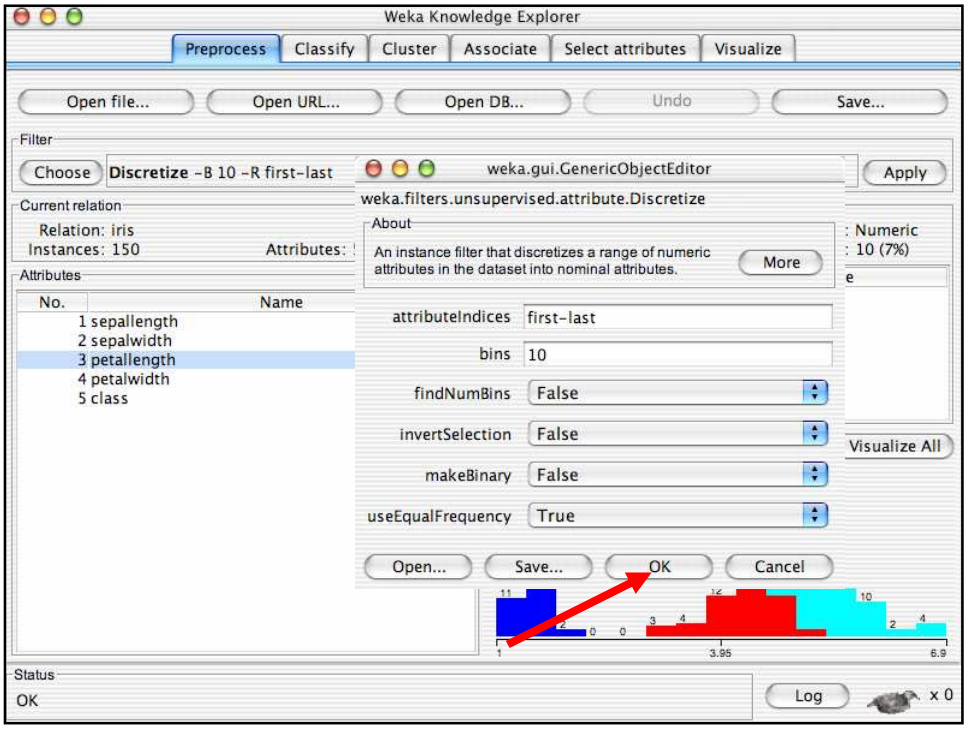

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

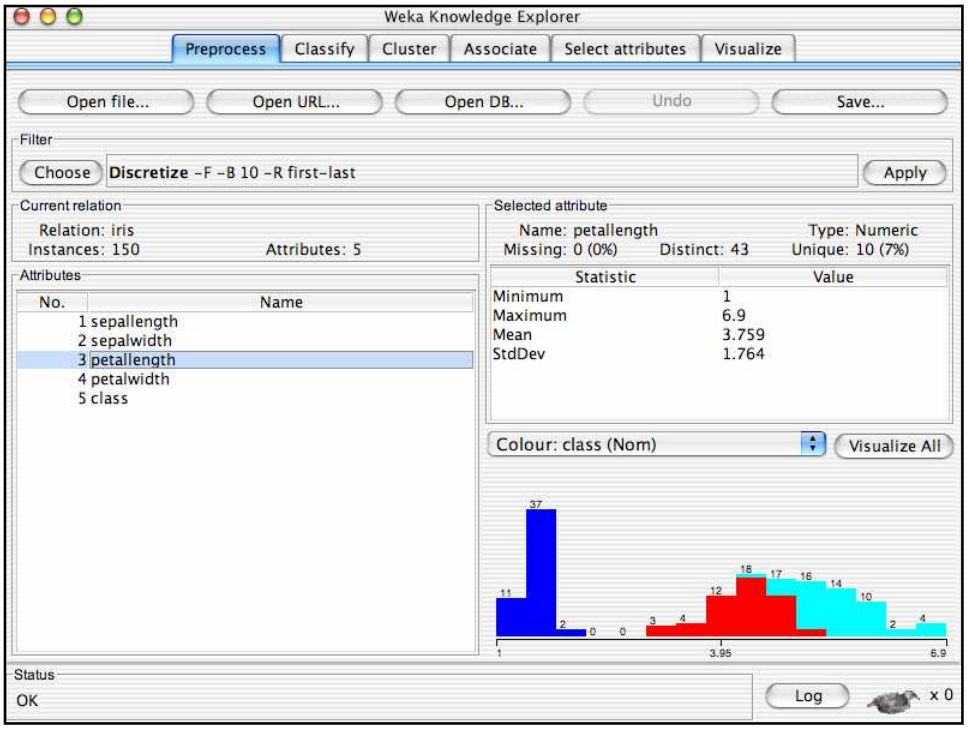

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

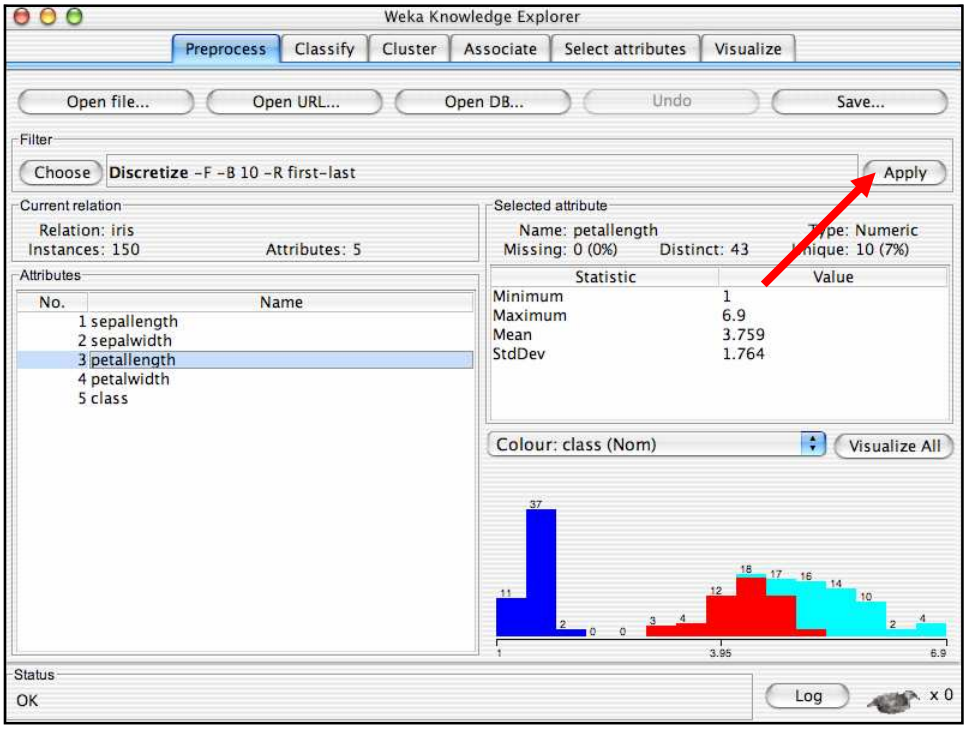

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
weka.gui.GenericObjectEditor
Apply
weka.filters.unsupervised.attribute.Discretize
Current relation
Relation: iris
Instances: 150
Attributes:
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
: Numeric
: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Visualize All
Open...
Save...
OK
Cancel
Status
OK
Log
x 0

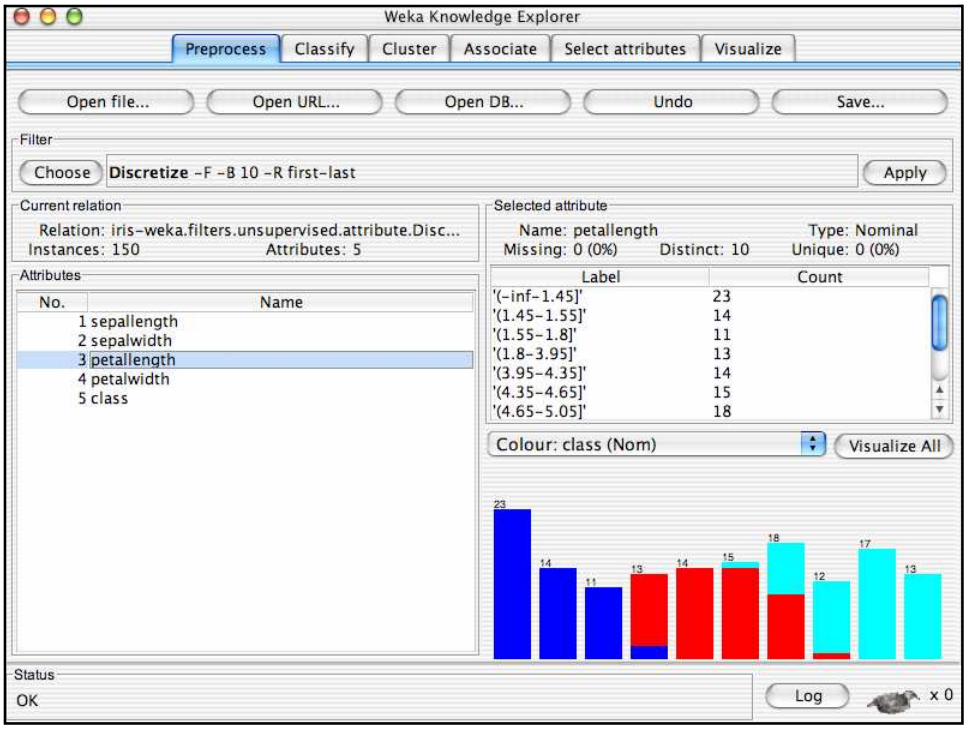

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
weka.gui.GenericObjectEditor
Apply
weka.filters.unsupervised.attribute.Discretize
Current relation
Relation: iris
Instances: 150
Attributes:
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
: Numeric
: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Visualize All
Save...
OK
Cancel
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
weka.gui.GenericObjectEditor
Apply
weka.filters.unsupervised.attribute.Discretize
Current relation
Relation: iris
Instances: 150
Attributes:
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
: Numeric
: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Visualize All
Open...
Save...
OK
Cancel
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
weka.gui.GenericObjectEditor
Apply
weka.filters.unsupervised.attribute.Discretize
Current relation
Relation: iris
Instances: 150
Attributes:
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
: Numeric
: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Visualize All
Open...
Save...
OK
Cancel
Status
OK
Log
x 0

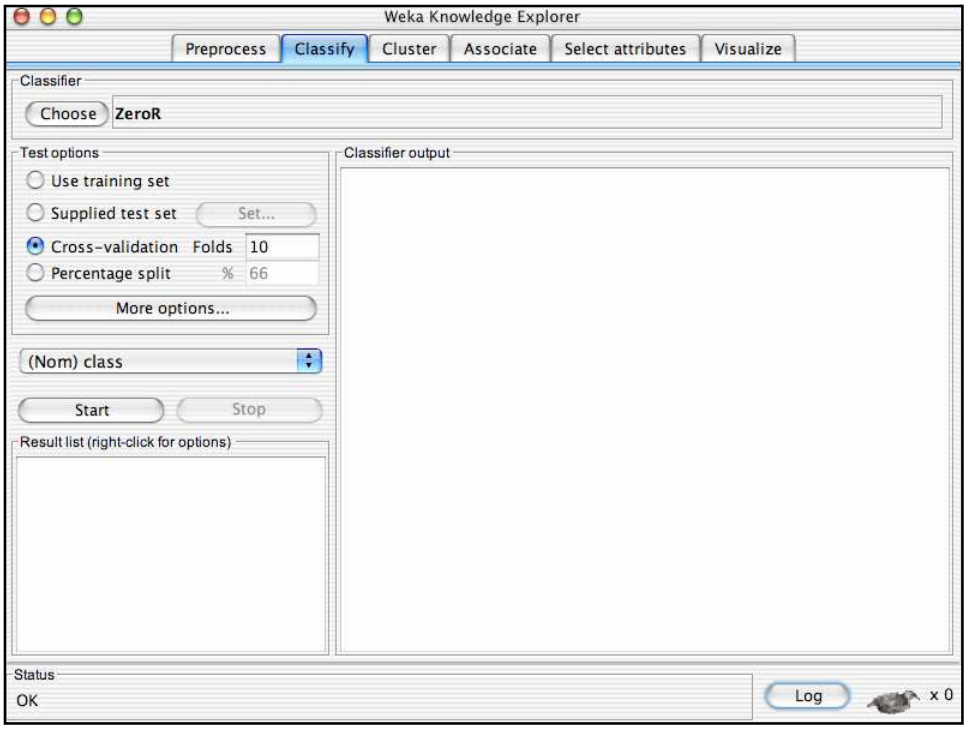

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -F -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

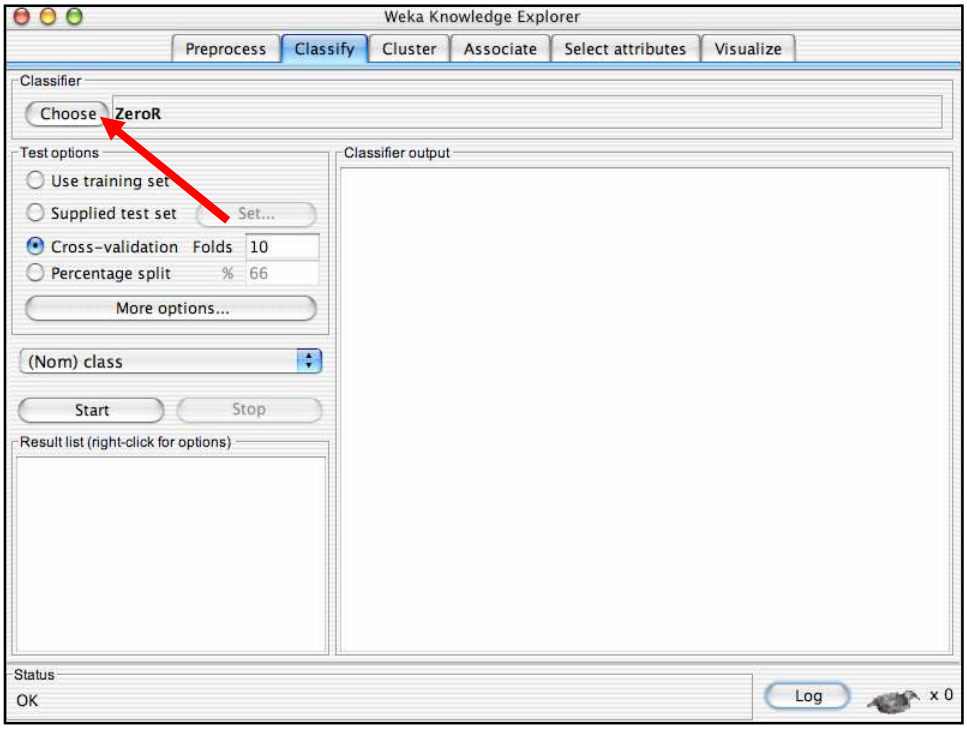

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -F -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Missing: 0 (0%)
Distinct: 43
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

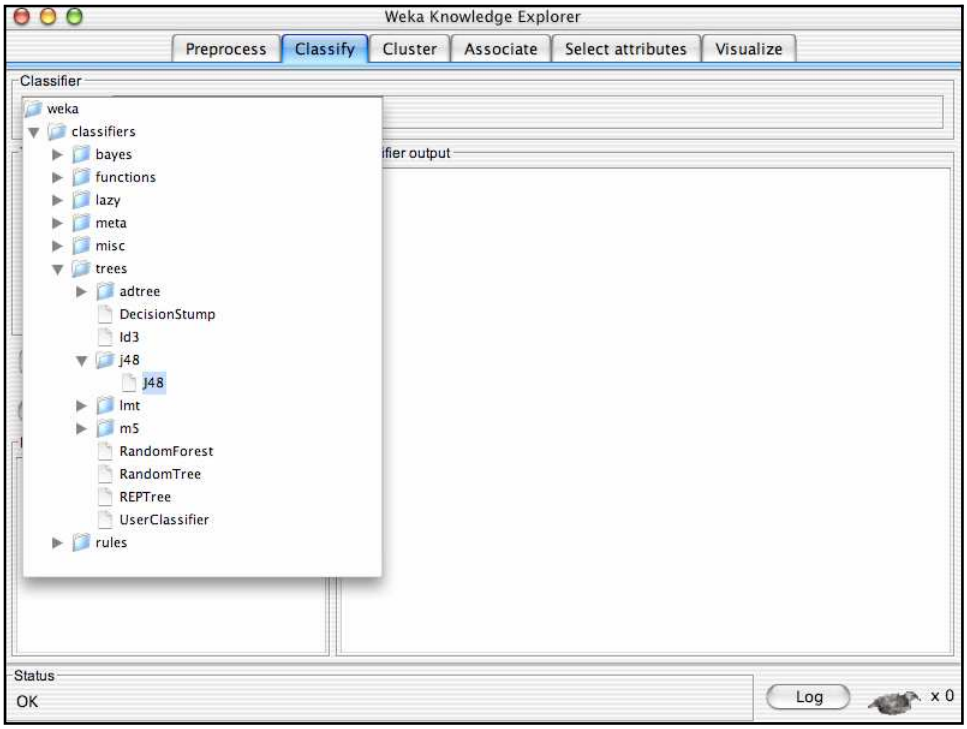


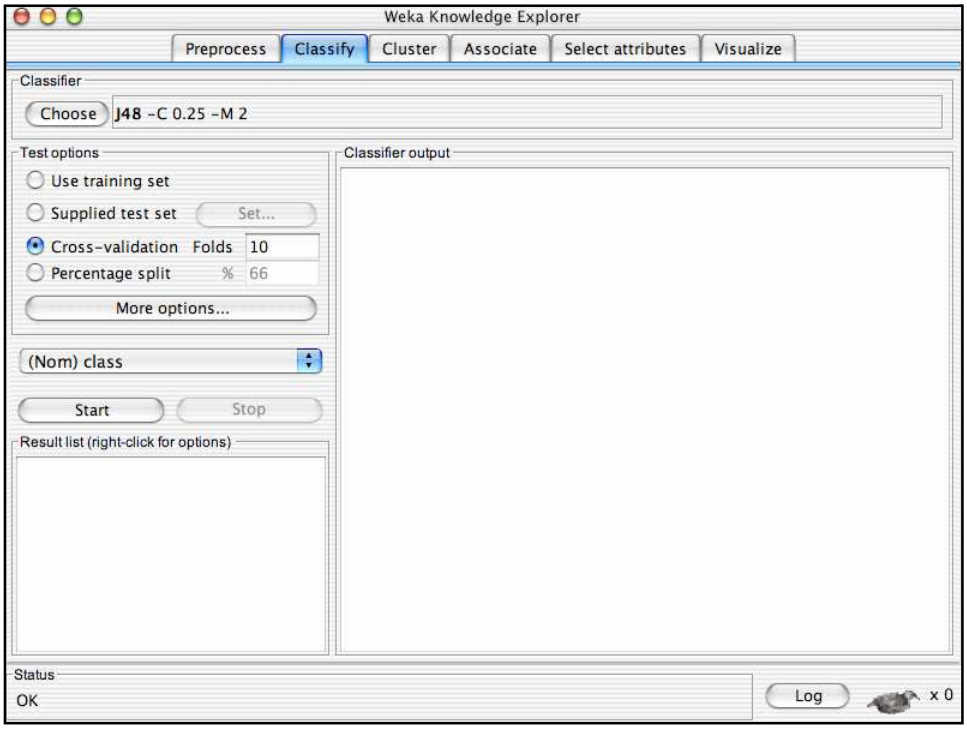
Explorer: building "classifiers"

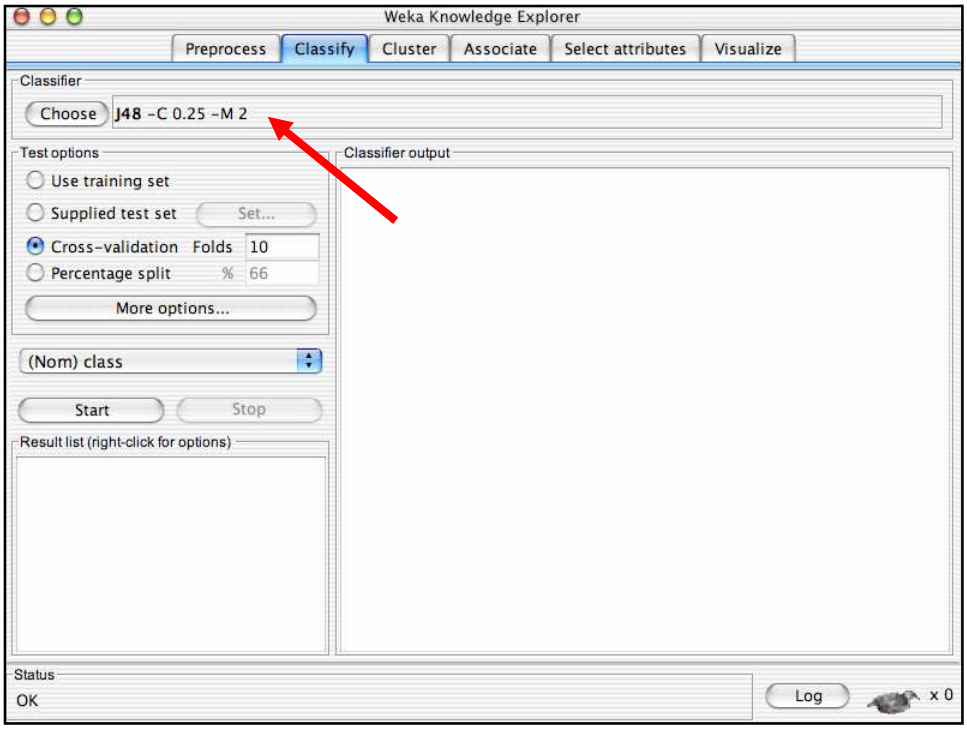
- Classifiers in WEKA are models for predicting nominal or numeric quantities
- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, …
- "Meta"-classifiers include:
  - Bagging,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

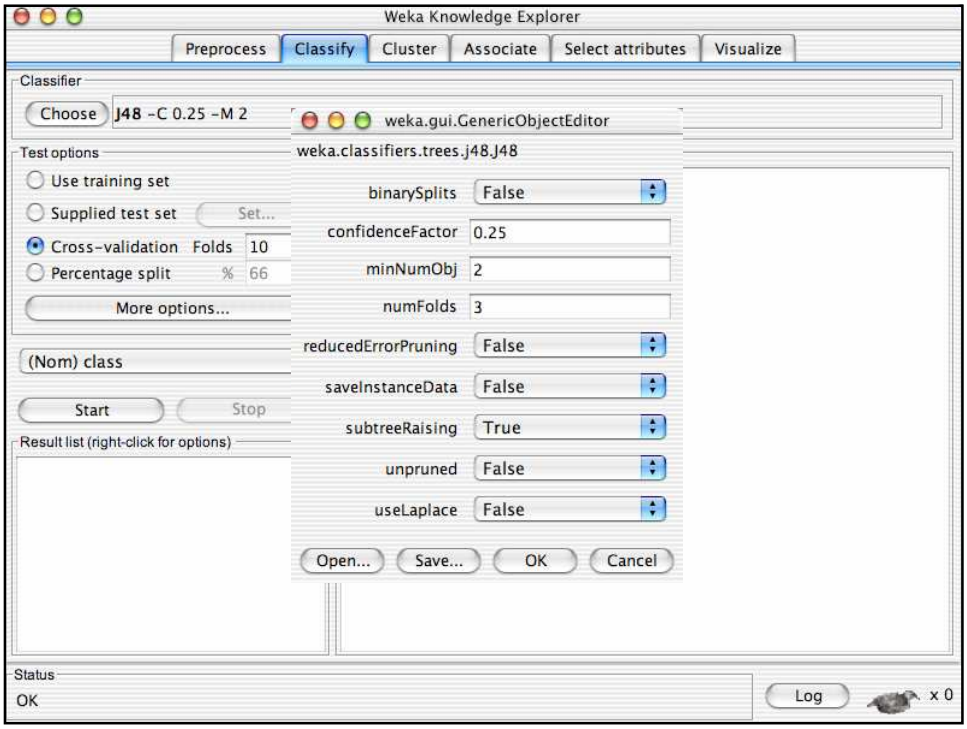

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

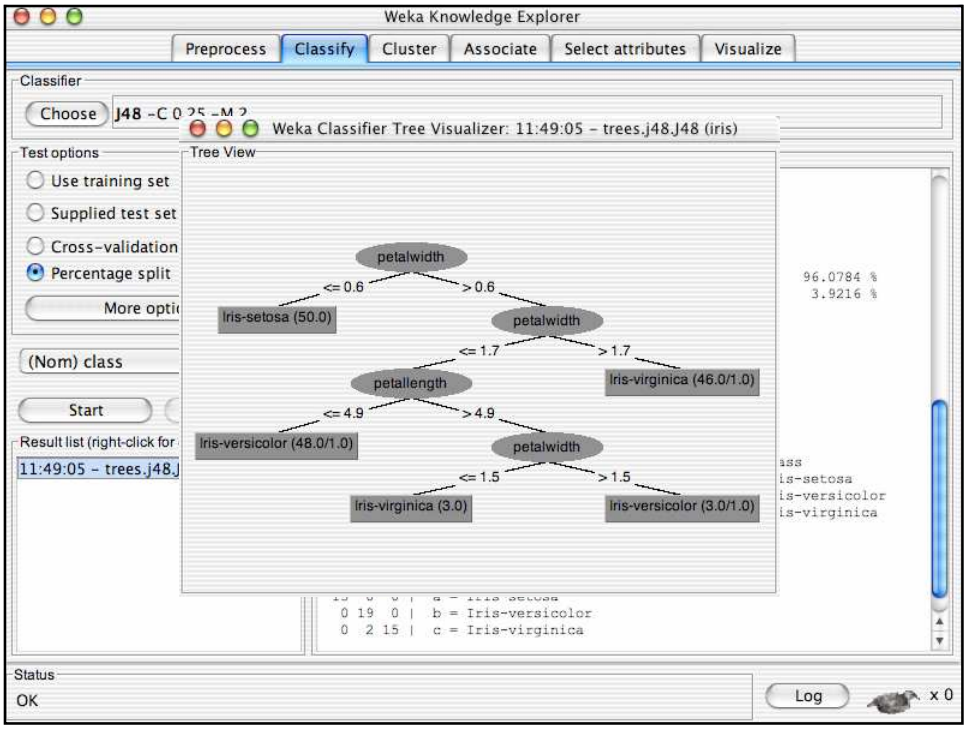

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
trees
adtree
DecisionStump
Id3
j48
J48
lmt
m5
RandomForest
RandomTree
REPTree
UserClassifier
rules
ifier output
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
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...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
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...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
weka.gui.GenericObjectEditor
weka.classifiers.trees.j48.J48
binarySplits
False
confidenceFactor
0.25
minNumObj
2
numFolds
3
reducedErrorPruning
False
saveInstanceData
False
subtreeRaising
True
unpruned
False
useLaplace
False
Open...
Save...
OK
Cancel
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)
Status
OK
Log
x 0

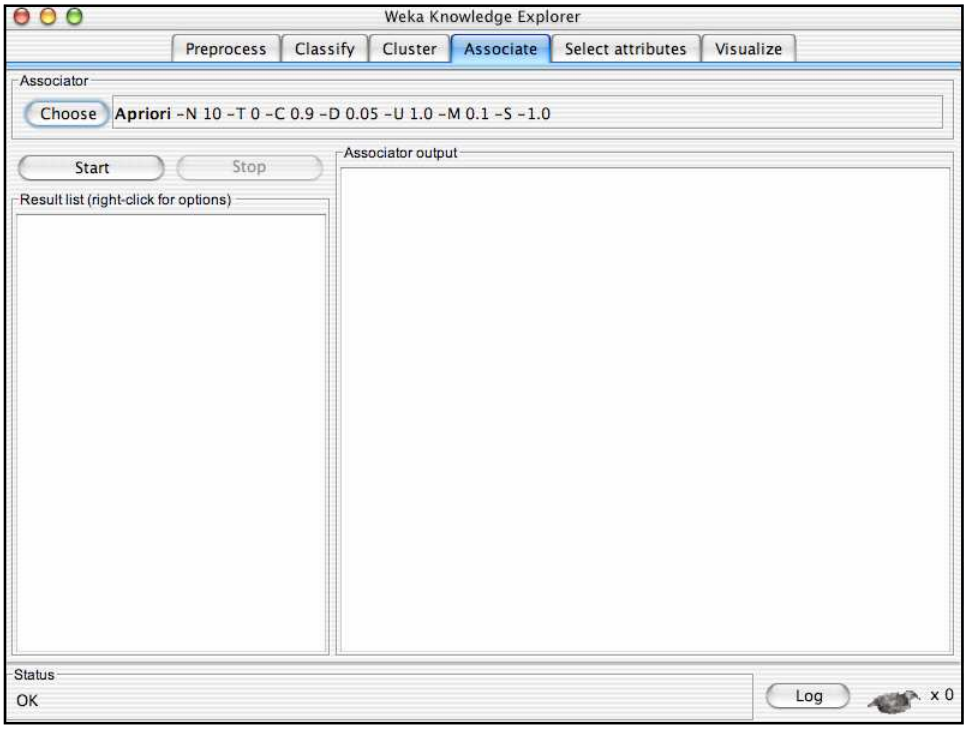


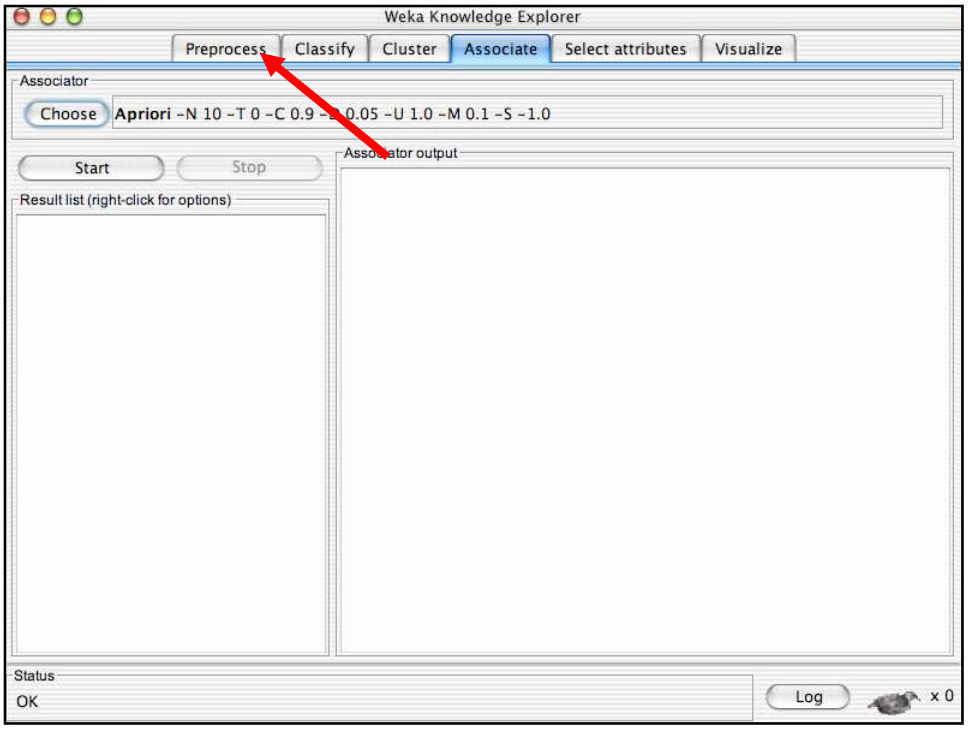
Explorer: clustering data

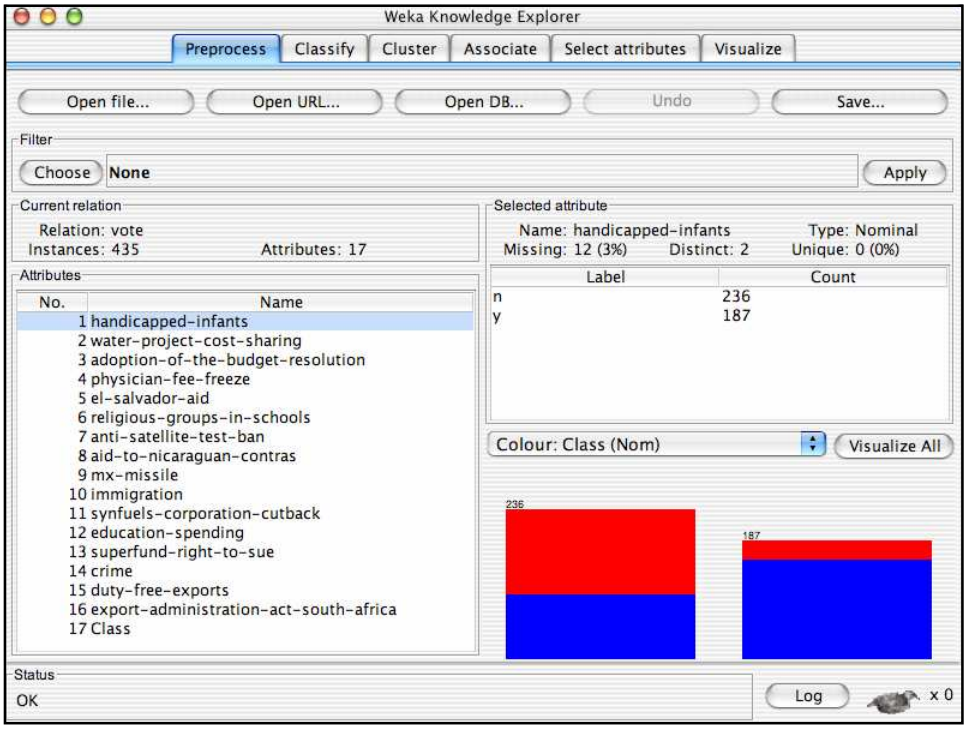
- WEKA contains “clusterers” for finding groups of similar instances in a dataset
- Implemented schemes are:
  - *k*-Means, EM, Cobweb, *X*-means, FarthestFirst
- Clusters can be visualized and compared to “true” clusters (if given)
- Evaluation based on loglikelihood if clustering scheme produces a probability distribution

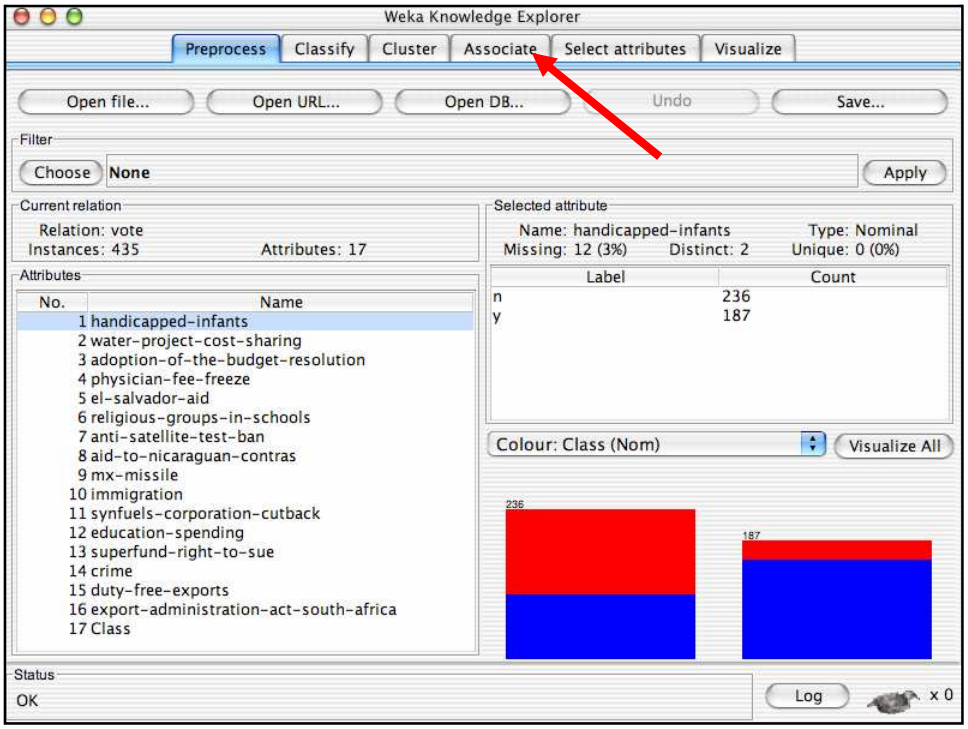
Explorer: finding associations

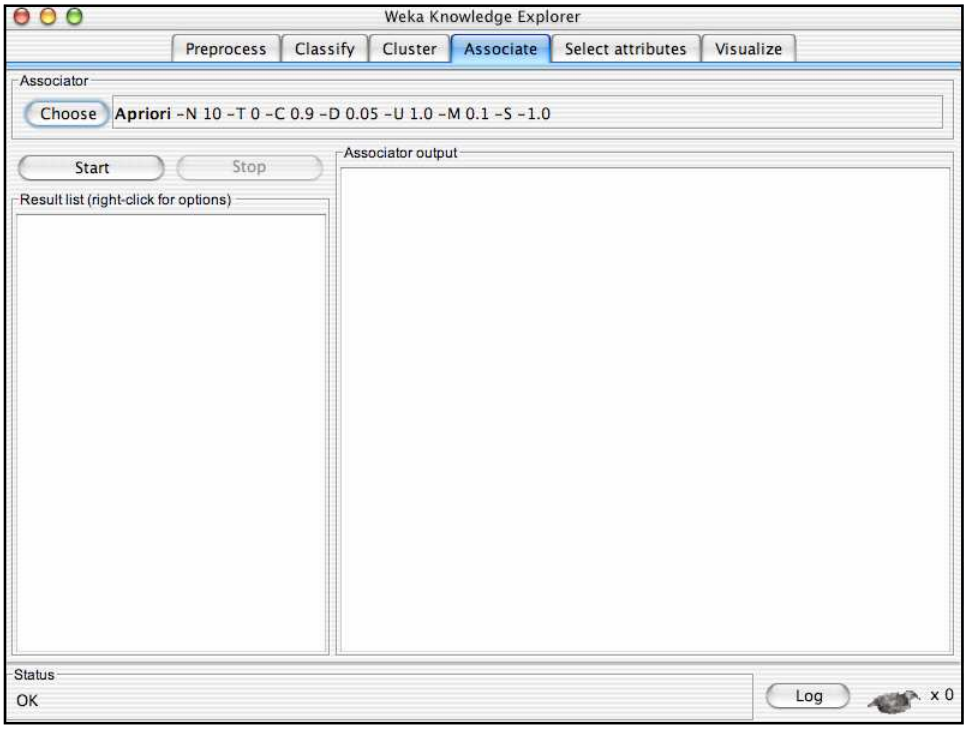
- WEKA contains an implementation of the Apriori algorithm for learning association rules
  - Works only with discrete data
- Can identify statistical dependencies between groups of attributes:
  - milk, butter ⇒ bread, eggs (with confidence 0.9 and support 2000)
- Apriori can compute all rules that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

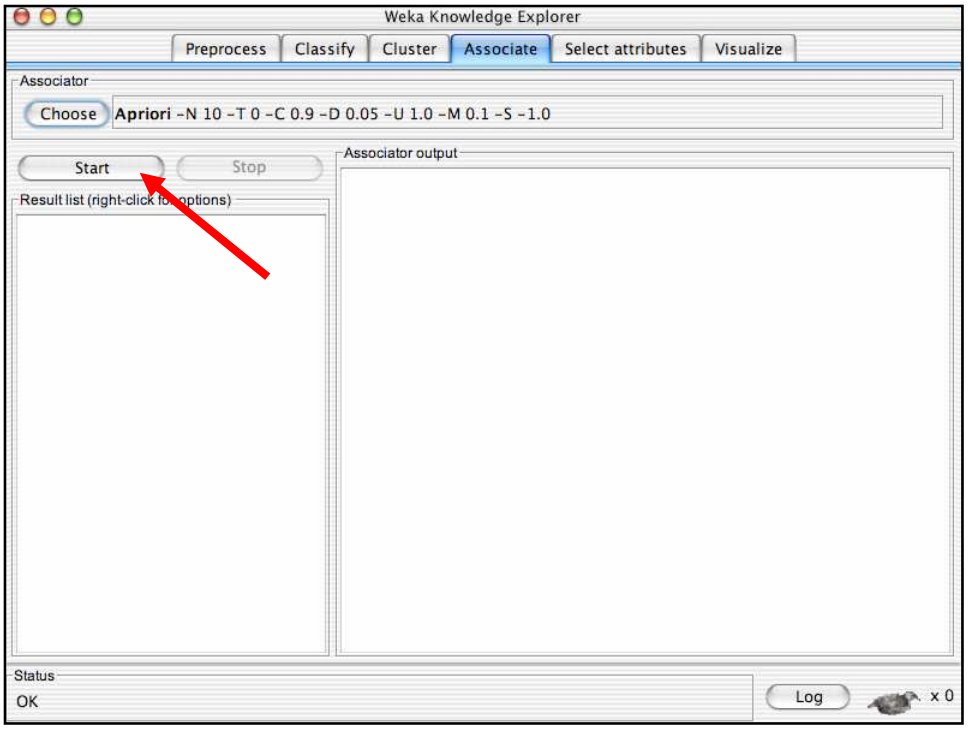


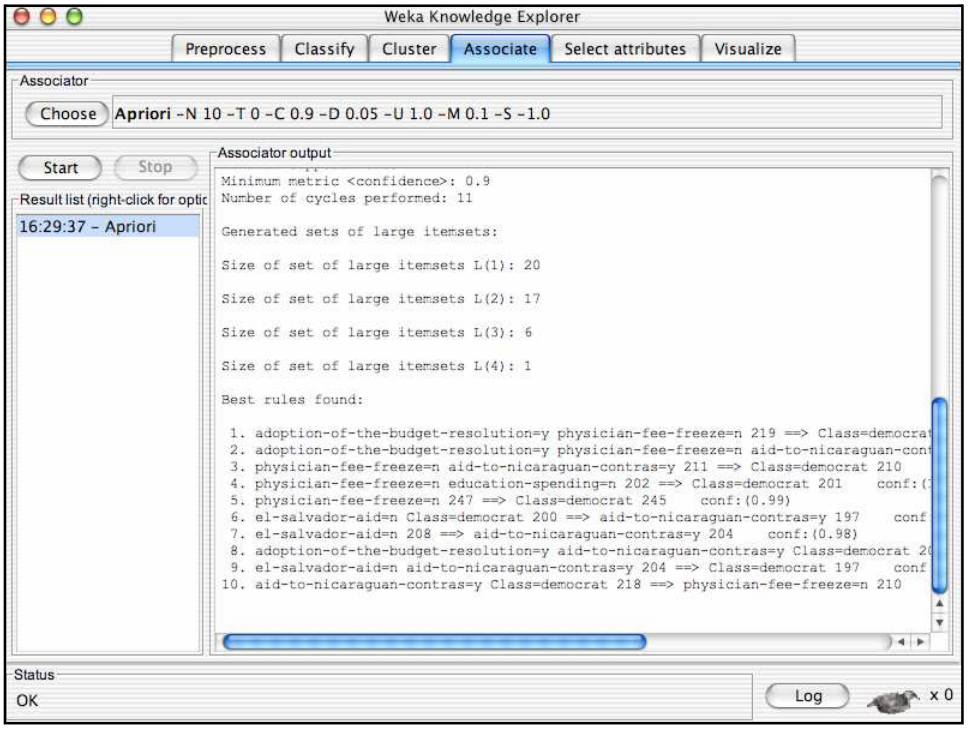

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Start
Stop
Result list (right-click for options)
Associator output
Status
OK
Log
x 0


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: vote
Instances: 435
Attributes: 17
Selected attribute
Name: handicapped-infants
Type: Nominal
Missing: 12 (3%)
Distinct: 2
Unique: 0 (0%)
Attributes
No.
Name
1 handicapped-infants
2 water-project-cost-sharing
3 adoption-of-the-budget-resolution
4 physician-fee-freeze
5 el-salvador-aid
6 religious-groups-in-schools
7 anti-satellite-test-ban
8 aid-to-nicaraguan-contras
9 mx-missile
10 immigration
11 synfuels-corporation-cutback
12 education-spending
13 superfund-right-to-sue
14 crime
15 duty-free-exports
16 export-administration-act-south-africa
17 Class
Label
Count
n
236
y
187
Colour: Class (Nom)
Visualize All
236
187
Status
OK
Log
x 0


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: vote
Instances: 435
Attributes: 17
Selected attribute
Name: handicapped-infants
Type: Nominal
Missing: 12 (3%)
Distinct: 2
Unique: 0 (0%)
Attributes
No.
Name
1 handicapped-infants
2 water-project-cost-sharing
3 adoption-of-the-budget-resolution
4 physician-fee-freeze
5 el-salvador-aid
6 religious-groups-in-schools
7 anti-satellite-test-ban
8 aid-to-nicaraguan-contras
9 mx-missile
10 immigration
11 synfuels-corporation-cutback
12 education-spending
13 superfund-right-to-sue
14 crime
15 duty-free-exports
16 export-administration-act-south-africa
17 Class
Label
Count
n
236
y
187
Colour: Class (Nom)
Visualize All
236
187
Status
OK
Log
x 0

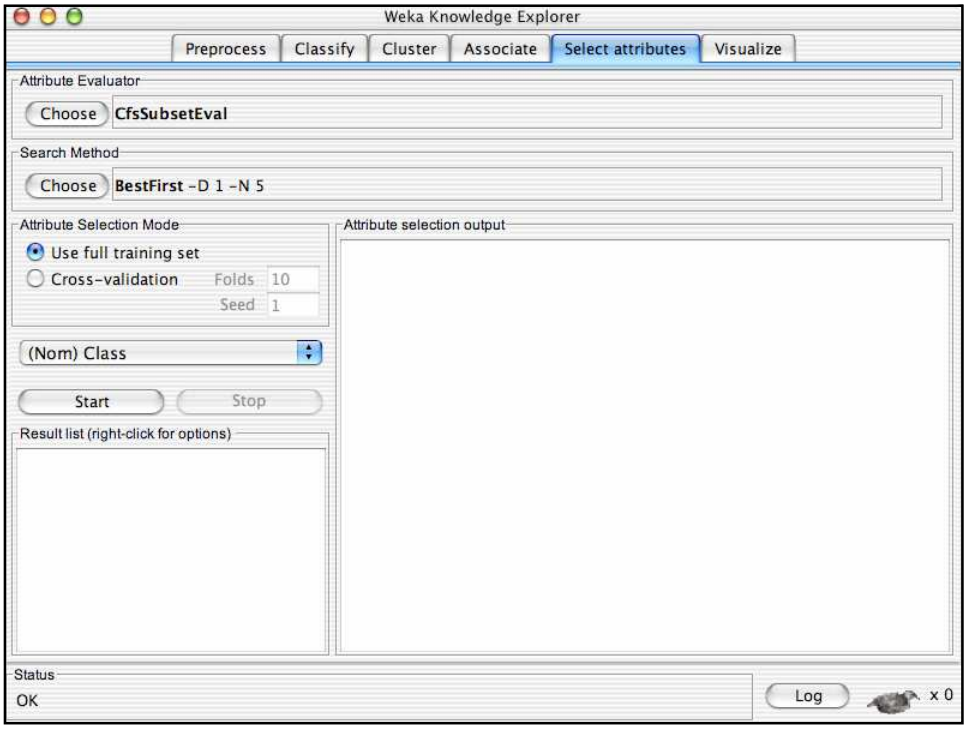

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Start
Stop
Result list (right-click for options)
Associator output
Status
OK
Log
x 0

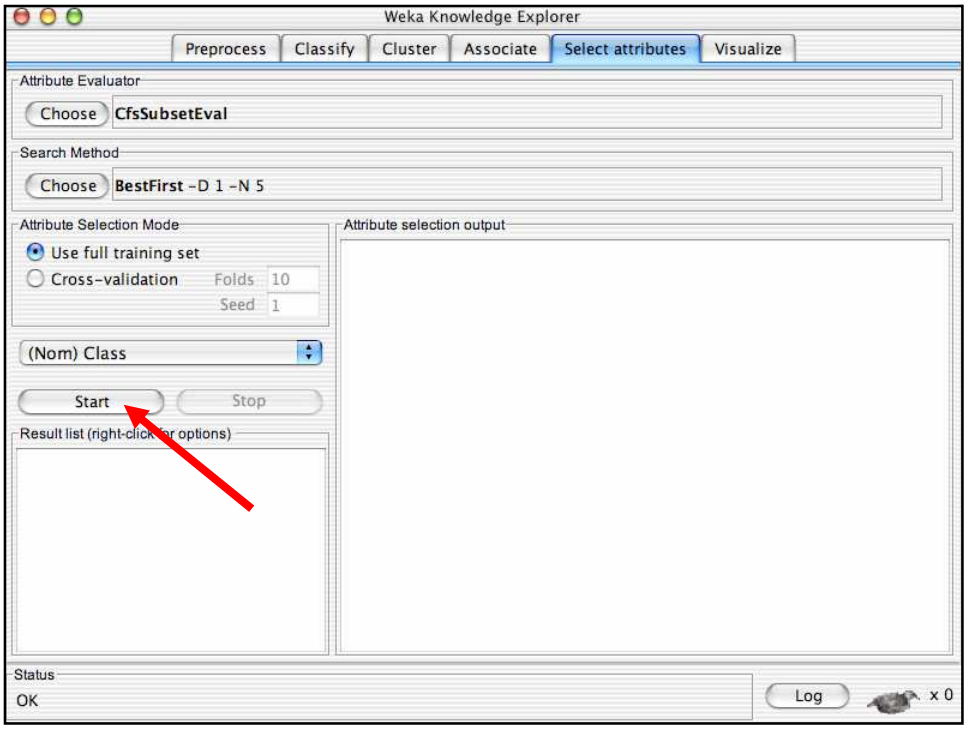

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Start
Stop
Result list (right-click for options)
Associator output
Status
OK
Log
x 0

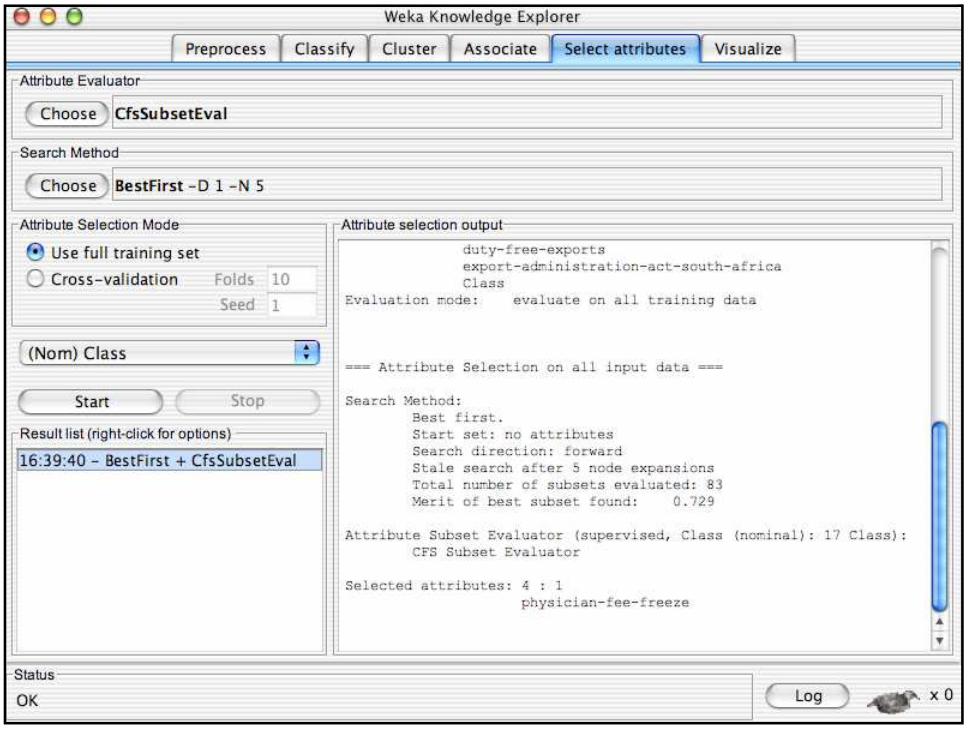

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
Start
Stop
Result list (right-click for optic
16:29:37 - Apriori
Associator output
Minimum metric <confidence>: 0.9
Number of cycles performed: 11
Generated sets of large itemsets:
Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 17
Size of set of large itemsets L(3): 6
Size of set of large itemsets L(4): 1
Best rules found:
1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat
2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-cont
3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210
4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 conf:(
5. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99)
6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 conf
7. el-salvador-aid=n 208 ==> aid-to-nicaraguan-contras=y 204 conf:(0.98)
8. adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y Class=democrat 2
9. el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 conf
10. aid-to-nicaraguan-contras=y Class=democrat 218 ==> physician-fee-freeze=n 210
Status
OK
Log
x 0

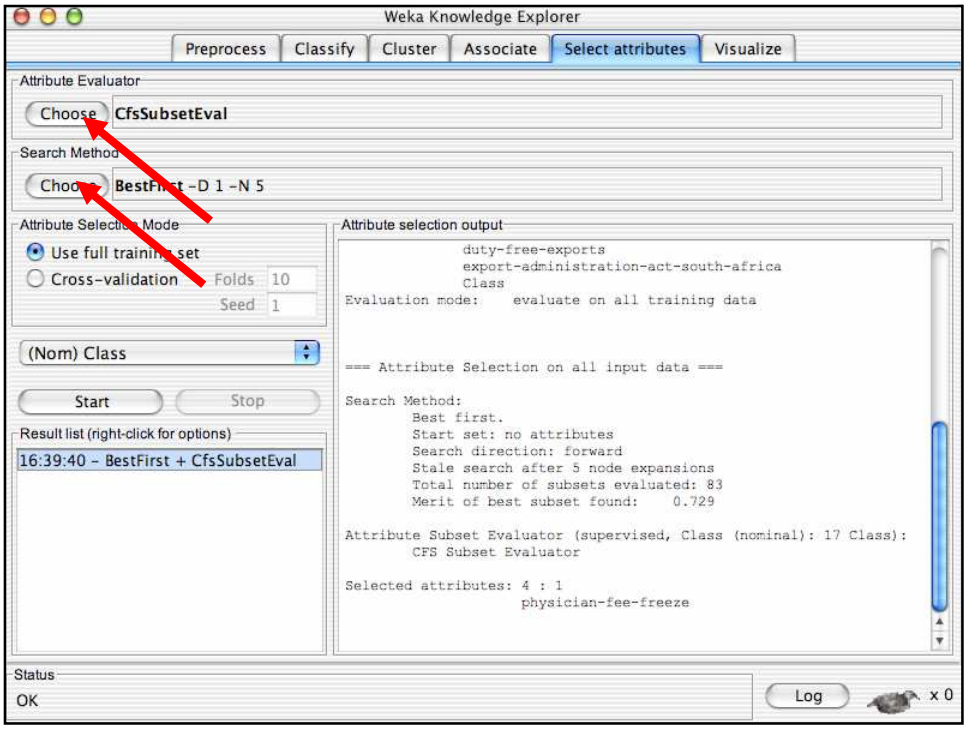
Explorer: attribute selection

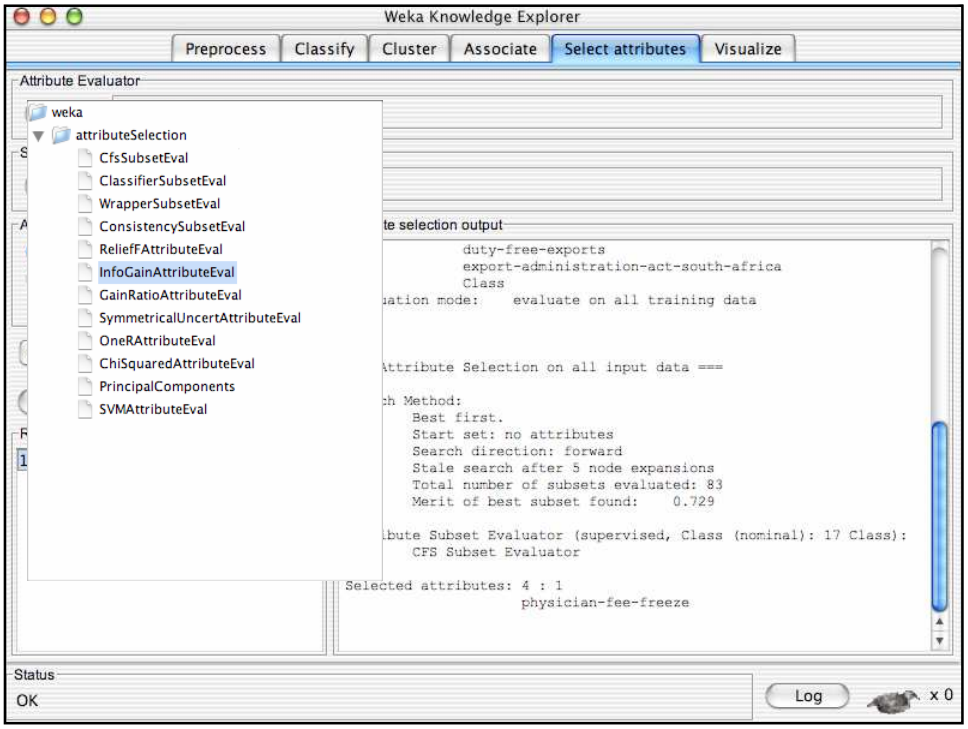
- Panel that can be used to investigate which (subsets of) attributes are the most predictive ones
- Attribute selection methods contain two parts:
  - A search method: best-first, forward selection, random, exhaustive, genetic algorithm, ranking
  - An evaluation method: correlation-based, wrapper,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

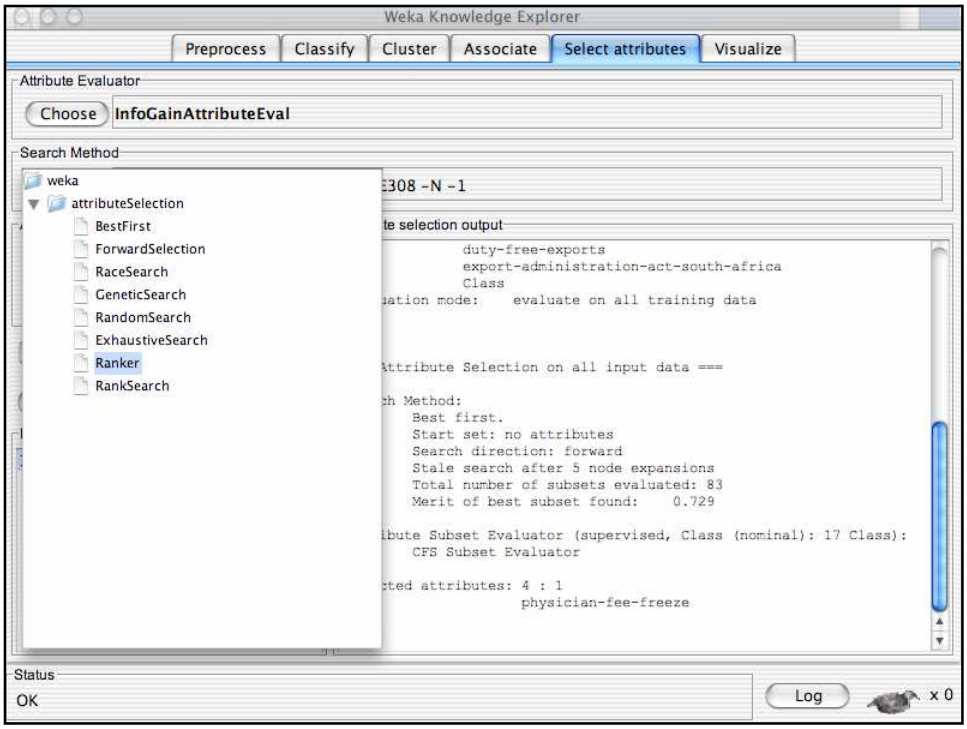


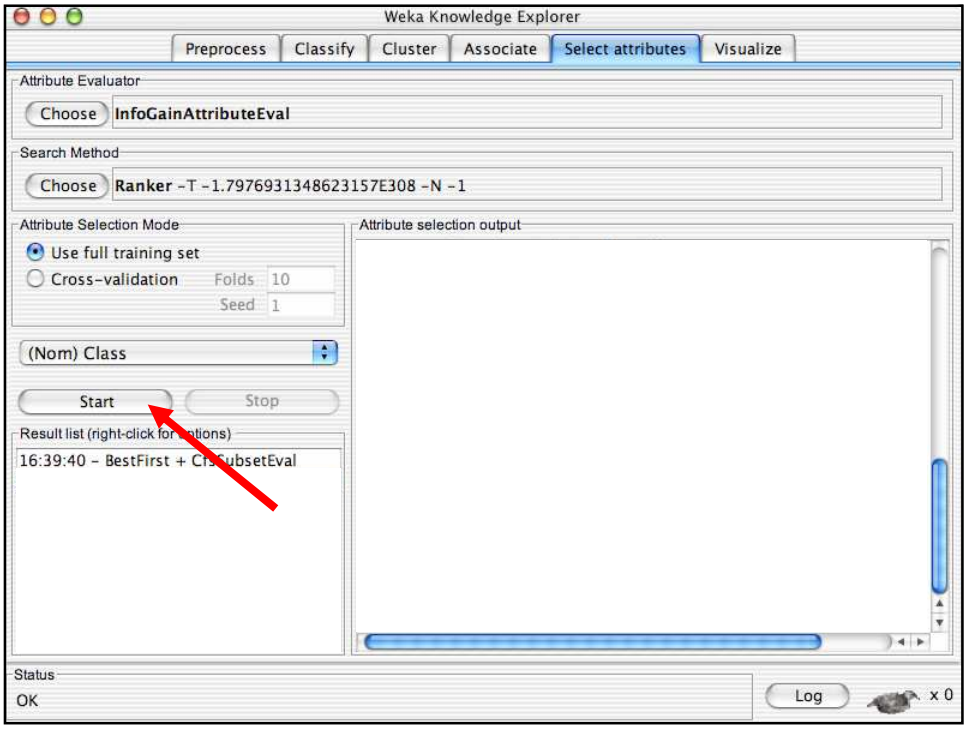

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval
Search Method
Choose
BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
(Nom) Class
Start
Stop
Result list (right-click for options)
Attribute selection output
Status
OK
Log
x 0

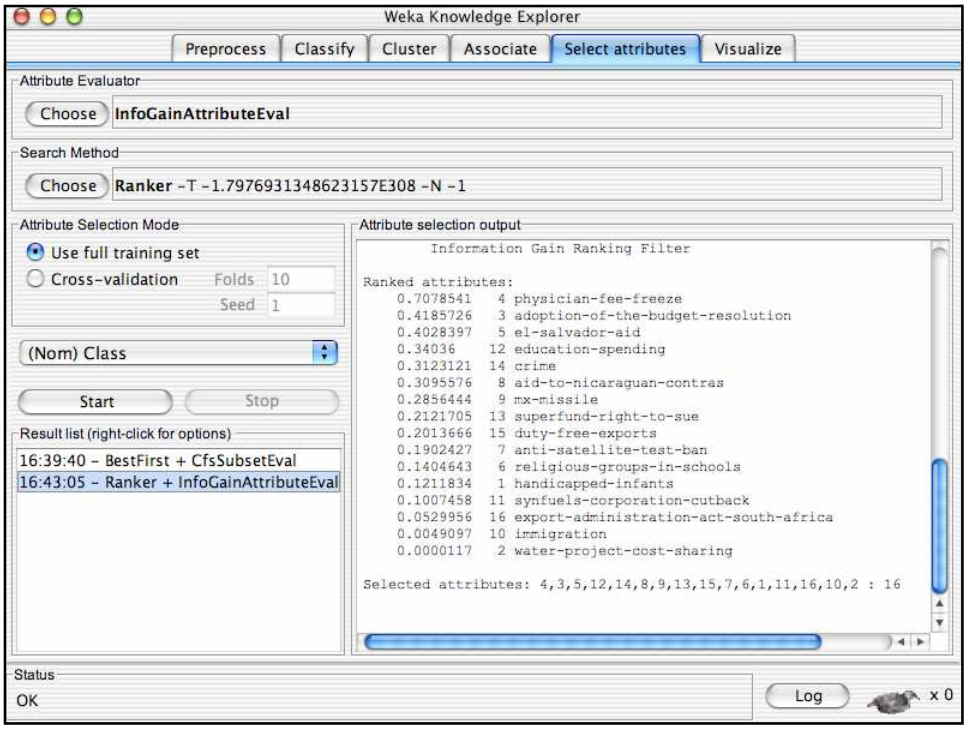

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval
Search Method
Choose
BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
Attribute selection output
duty-free-exports
export-administration-act-south-africa
Class
Evaluation mode: evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729
Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator
Selected attributes: 4 : 1
physician-fee-freeze
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CfsSubsetEval
Search Method
Choose
BestFirst -D 1 -N 5
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
Attribute selection output
duty-free-exports
export-administration-act-south-africa
Class
Evaluation mode: evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729
Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
CFS Subset Evaluator
Selected attributes: 4 : 1
physician-fee-freeze
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
weka
attributeSelection
CfsSubsetEval
ClassifierSubsetEval
WrapperSubsetEval
ConsistencySubsetEval
ReliefFAttributeEval
InfoGainAttributeEval
GainRatioAttributeEval
SymmetricalUncertAttributeEval
OneRAttributeEval
ChiSquaredAttributeEval
PrincipalComponents
SVMAttributeEval
Status
OK
Log
x 0

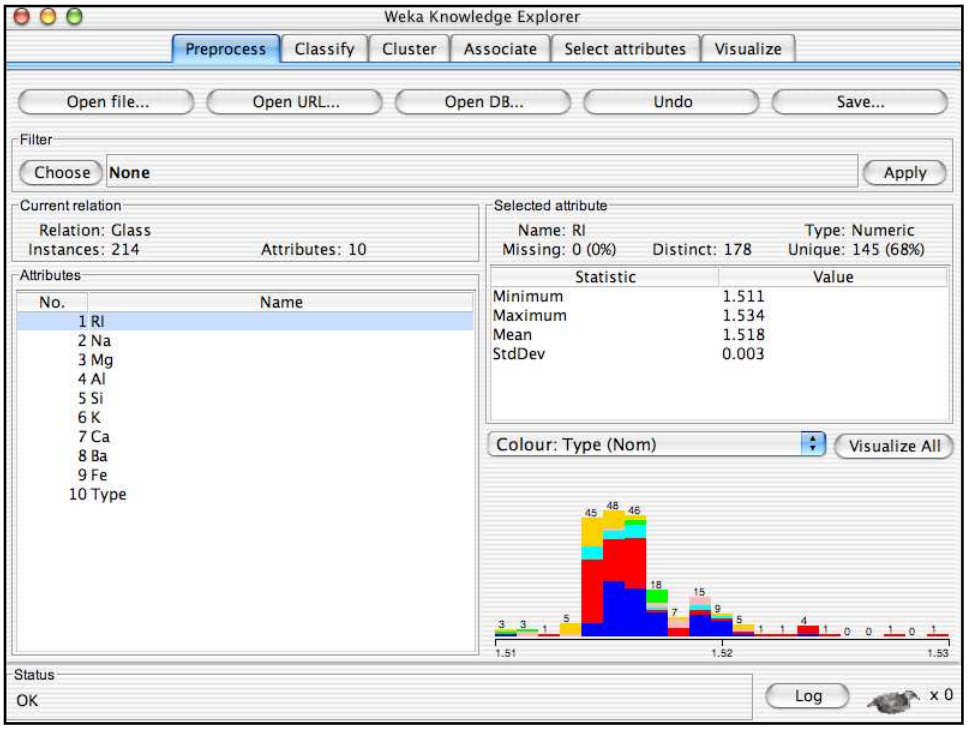

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
weka
attributeSelection
BestFirst
ForwardSelection
RaceSearch
GeneticSearch
RandomSearch
ExhaustiveSearch
Ranker
RankSearch
duty-free-exports
export-administration-act-south-africa
Class
evaluate on all training data
Attribute Selection on all input data
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 83
Merit of best subset found: 0.729
CFS Subset Evaluator
physician-fee-freeze
Status
OK
Log
x 0

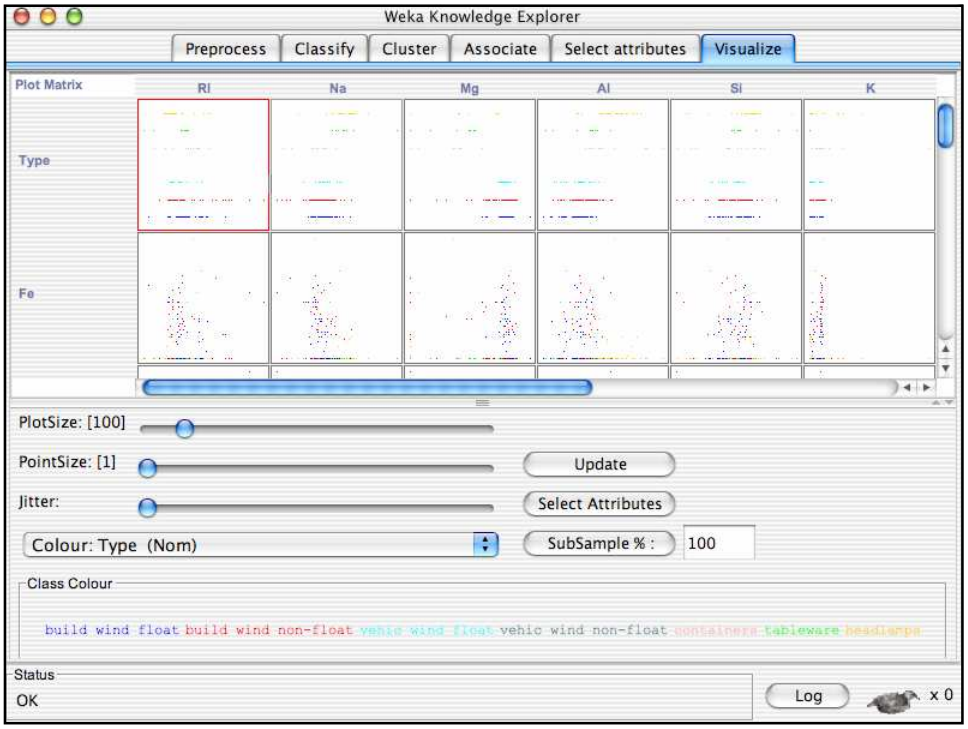

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) Class
Start
Stop
Result list (right-click for options)
16:39:40 - BestFirst + CfsSubsetEval
Attribute selection output
Status
OK
Log
x 0

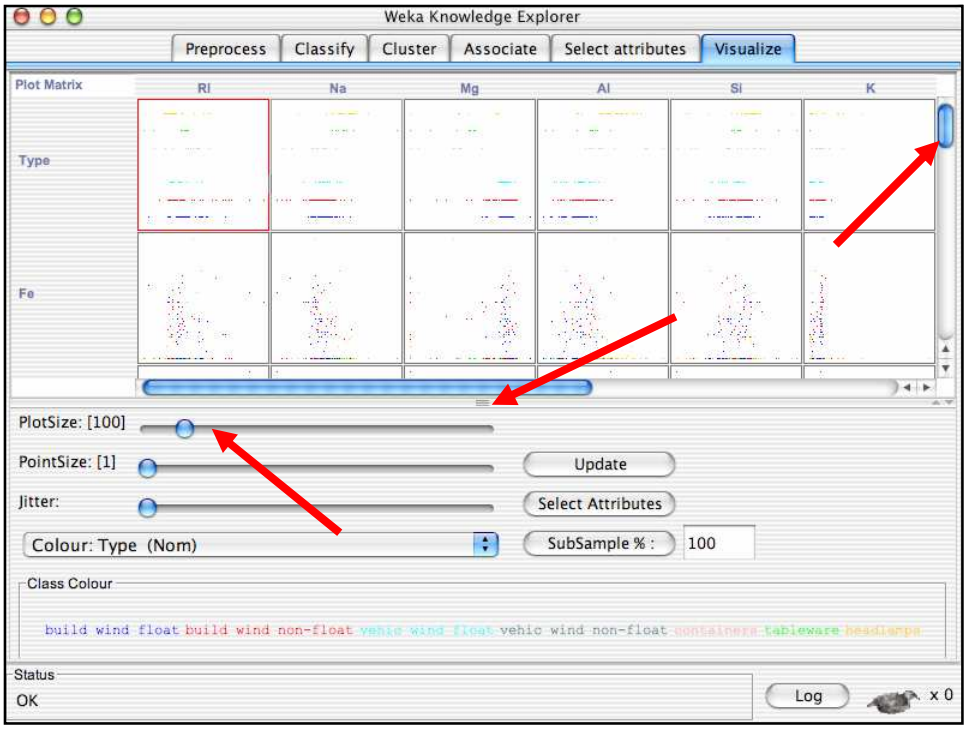


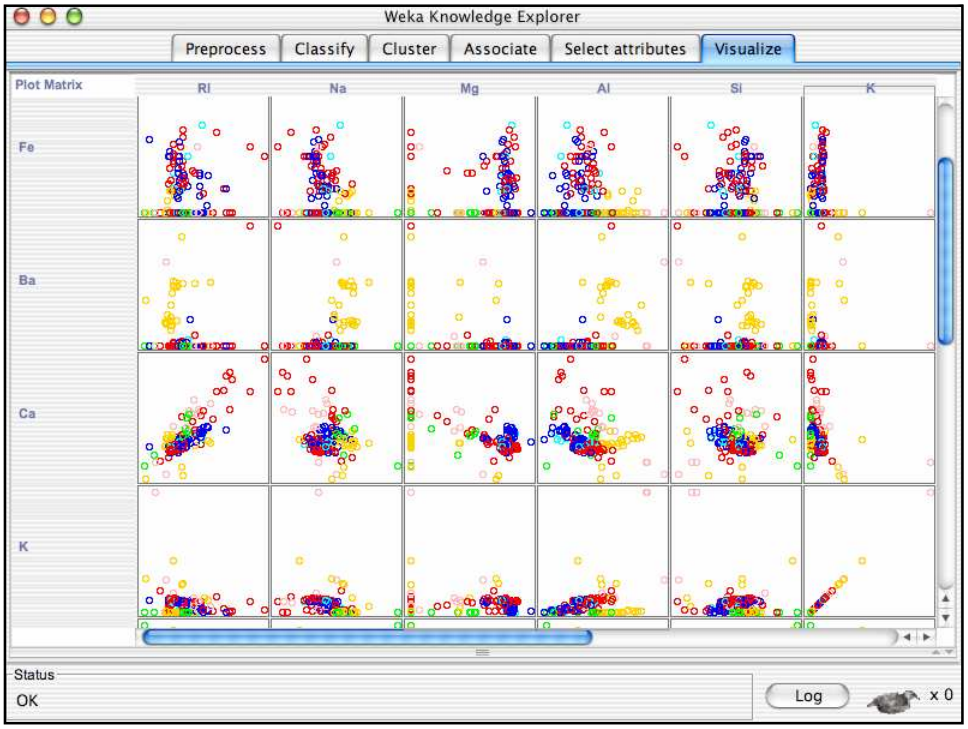
Explorer: data visualization

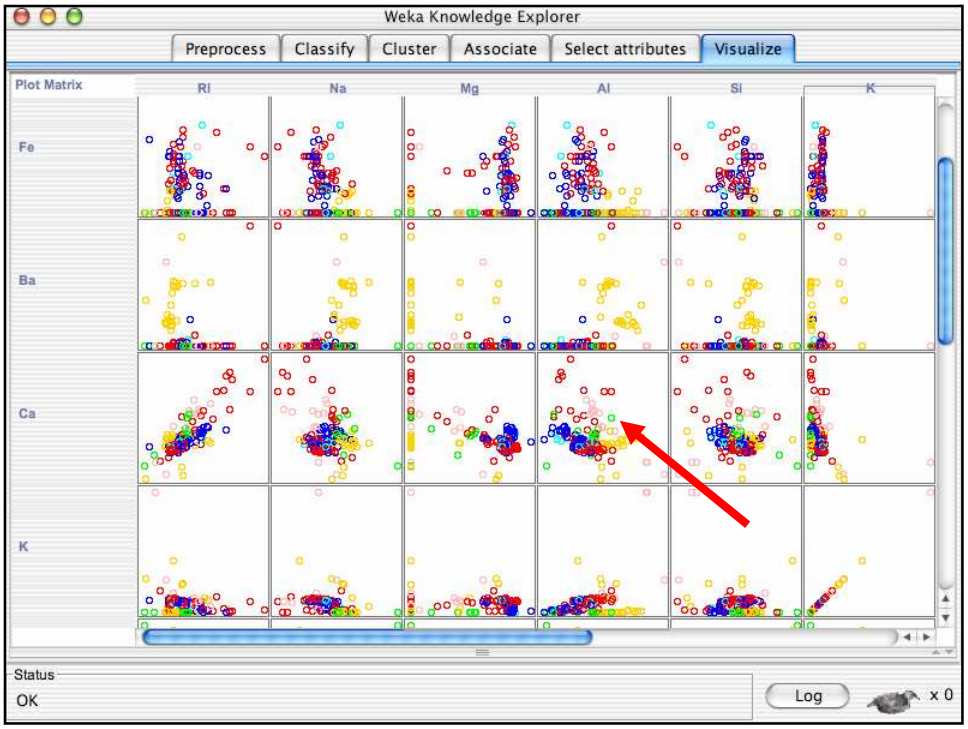
- Visualization very useful in practice: e.g. helps to determine difficulty of the learning problem
- WEKA can visualize single attributes (1-d) and pairs of attributes (2-d)
  - To do: rotating 3-d visualizations (Xgobi-style)
- Color-coded class values
- "Jitter" option to deal with nominal attributes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

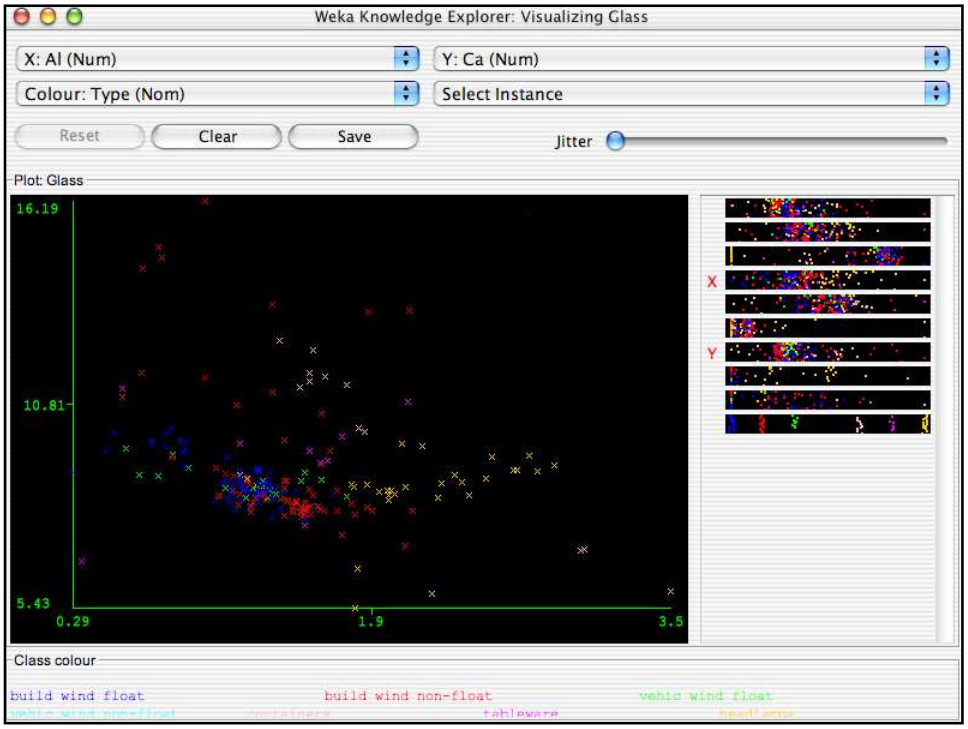

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: Glass
Instances: 214
Attributes: 10
Selected attribute
Name: RI
Type: Numeric
Missing: 0 (0%)
Distinct: 178
Unique: 145 (68%)
Attributes
No.
Name
1 RI
2 Na
3 Mg
4 Al
5 Si
6 K
7 Ca
8 Ba
9 Fe
10 Type
Statistic
Value
Minimum
1.511
Maximum
1.534
Mean
1.518
StdDev
0.003
Colour: Type (Nom)
Visualize All
1.51
1.52
1.53
Status
OK
Log
x 0

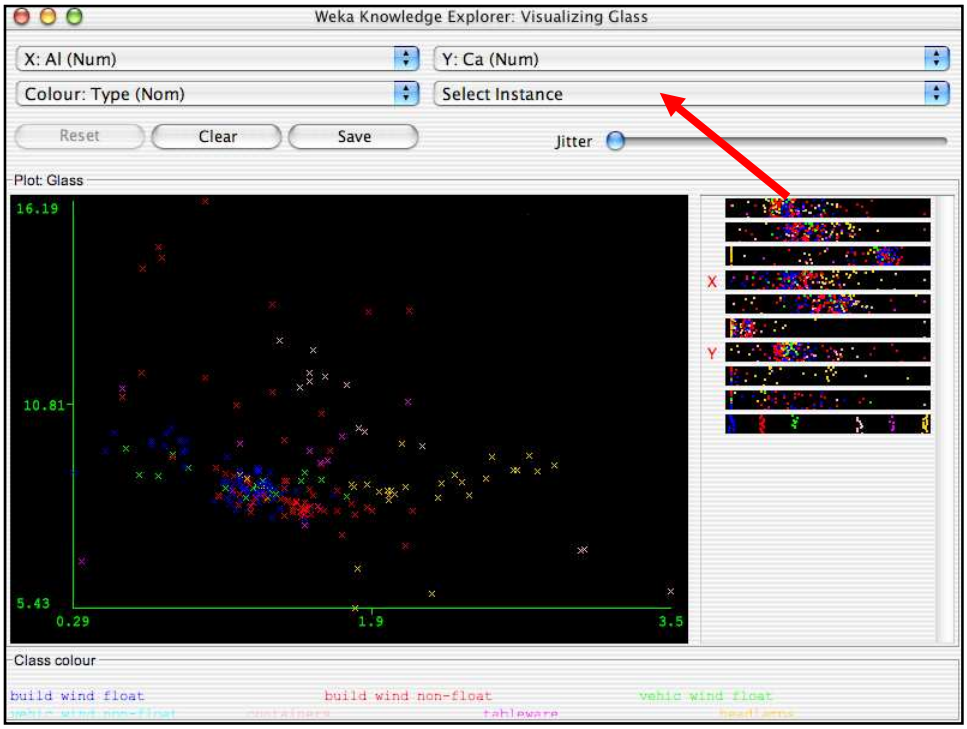

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Type
Fe
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: Type (Nom)
SubSample % :
100
Class Colour
Status
OK
Log
x 0

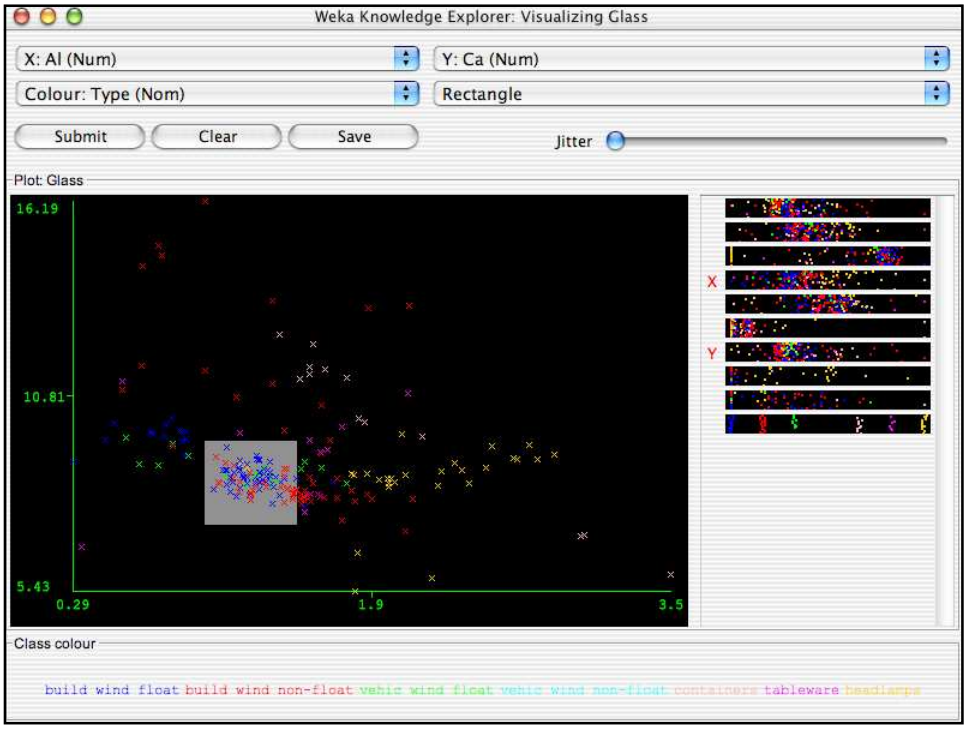

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Type
Fe
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: Type (Nom)
SubSample % :
100
Class Colour
build wind float build wind non-float vehic wind float vehic wind non-float containers tableware headlamps
Status
OK
Log
x 0

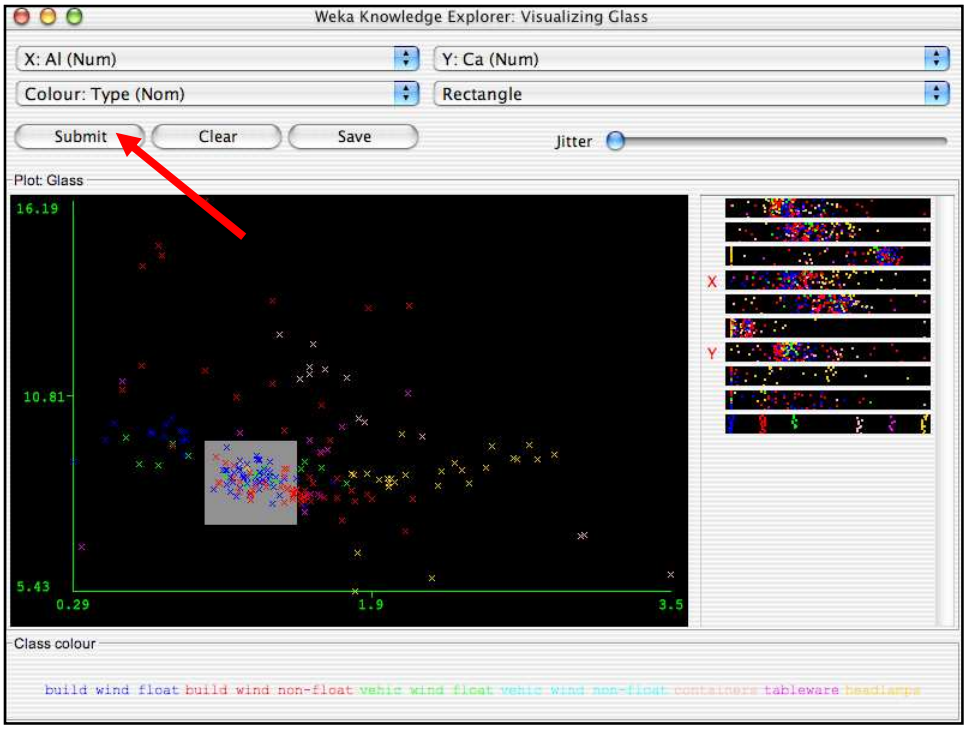

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Fe
Ba
Ca
K
Status
OK
Log
x 0

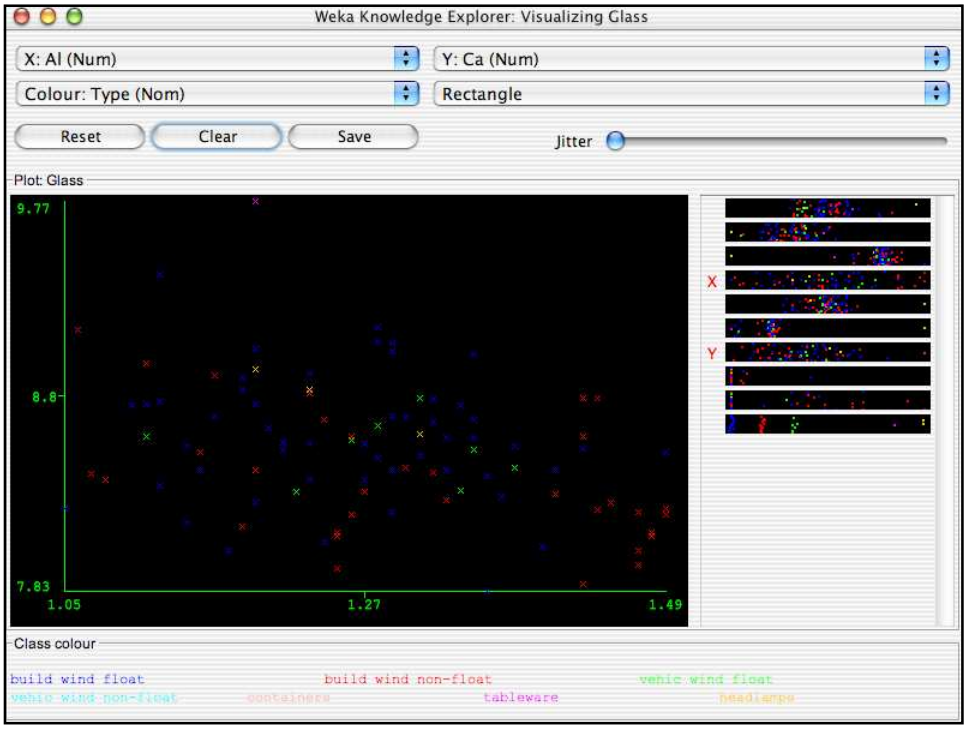

Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
RI
Na
Mg
Al
Si
K
Fe
Ba
Ca
K
Status
OK
Log
x 0

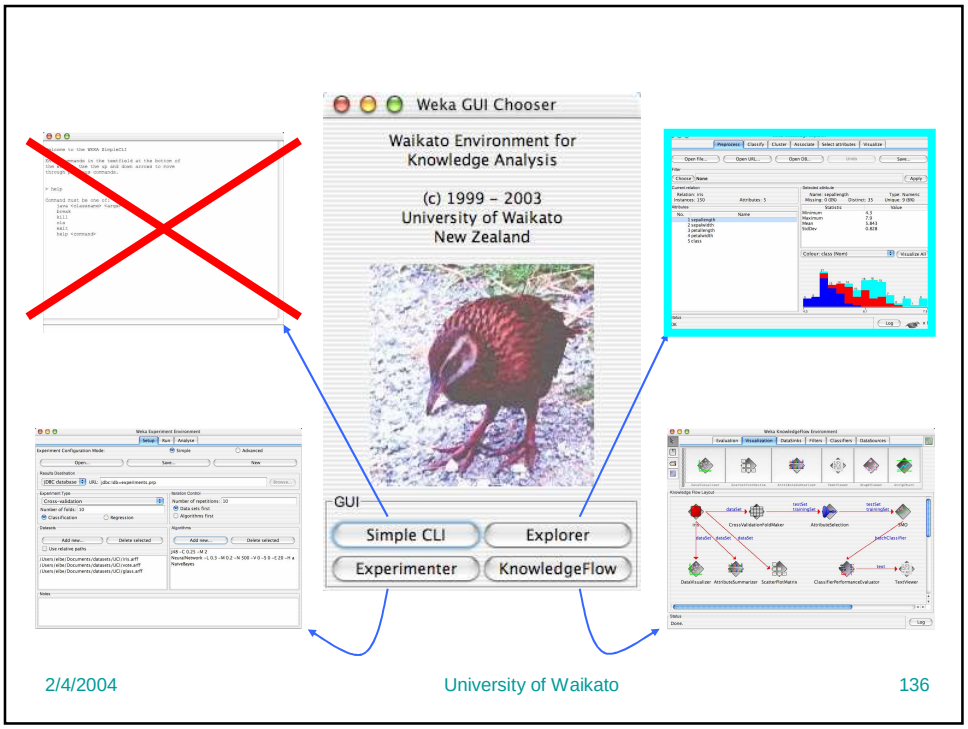

Weka Knowledge Explorer: Visualizing Glass
X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: Glass
16.19
10.81
5.43
0.29
1.9
3.5
X
Y
Class colour
build wind float
build wind non-float
vehic wind float

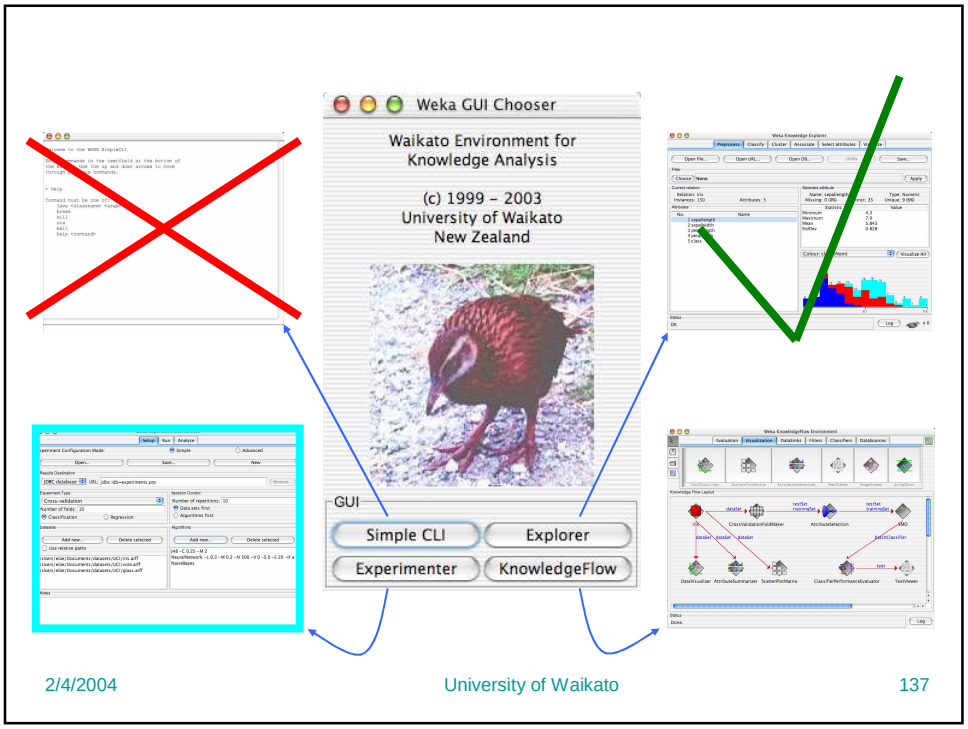

Weka Knowledge Explorer: Visualizing Glass
X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: Glass
16.19
10.81
5.43
0.29
1.9
3.5
X
Y
Class colour
build wind float
build wind non-float
vehic wind float


Weka Knowledge Explorer: Visualizing Glass
X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Rectangle
Submit
Clear
Save
Jitter
Plot: Glass
16.19
10.81
5.43
0.29
1.9
3.5
X
Y
Class colour
build wind float
build wind non-float
vehic wind float
vehic wind non-float
containers
tableware
headlamps


Weka Knowledge Explorer: Visualizing Glass
X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Rectangle
Submit
Clear
Save
Jitter
Plot: Glass
16.19
10.81
5.43
0.29
1.9
3.5
X
Y
Class colour
build wind float
build wind non-float
vehic wind float
vehic wind non-float
containers
tableware
headlamps


Weka Knowledge Explorer: Visualizing Glass
X: Al (Num)
Y: Ca (Num)
Colour: Type (Nom)
Rectangle
Reset
Clear
Save
Jitter
Plot: Glass
9.77
8.8
7.83
1.05
1.27
1.49
X
Y
Class colour
build wind float
build wind non-float
vehic wind float
vehic wind non-float
containers
tableware
headlamps


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

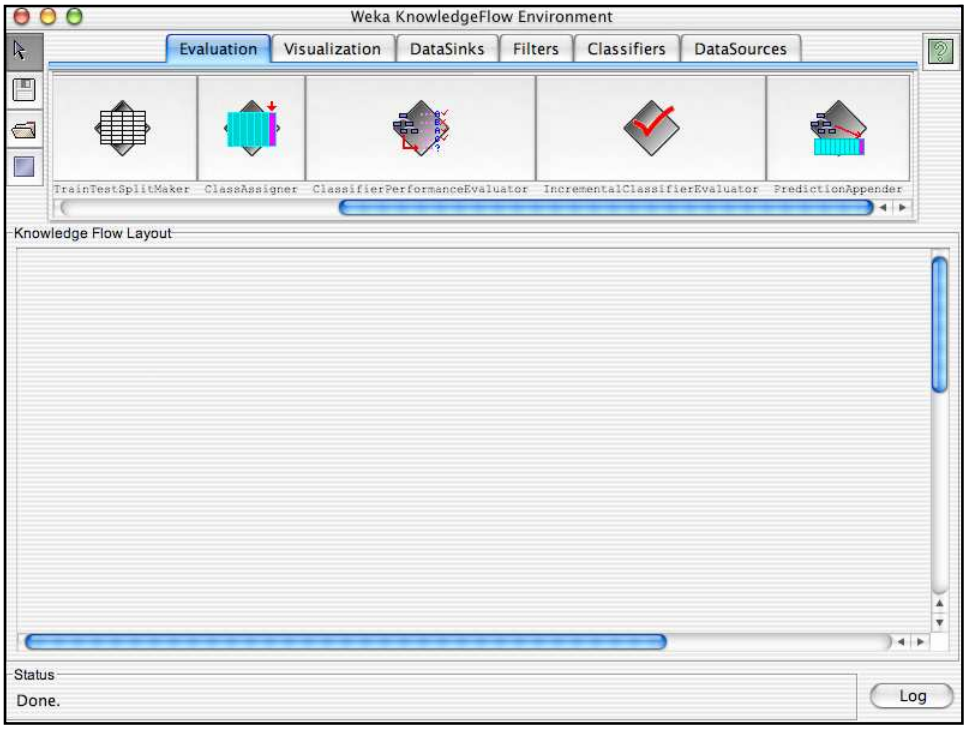

Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow
2/4/2004
University of Waikato
137

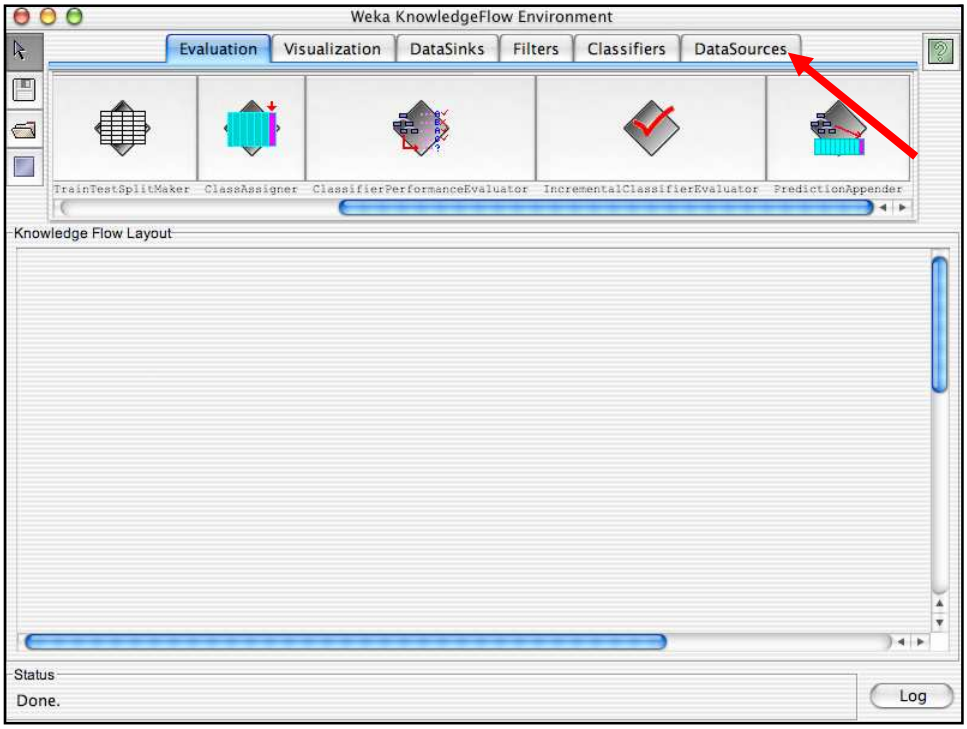
Performing experiments

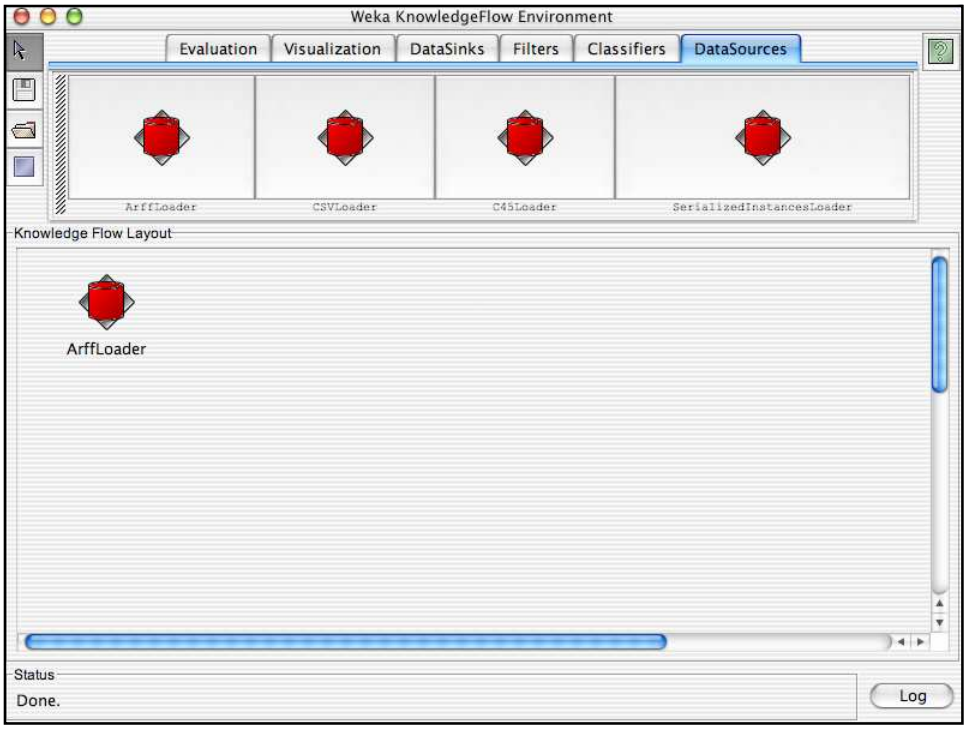
- Experimenter makes it easy to compare the performance of different learning schemes
- For classification and regression problems
- Results can be written into file or database
- Evaluation options: cross-validation, learning curve, hold-out
- Can also iterate over different parameter settings
- Significance-testing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

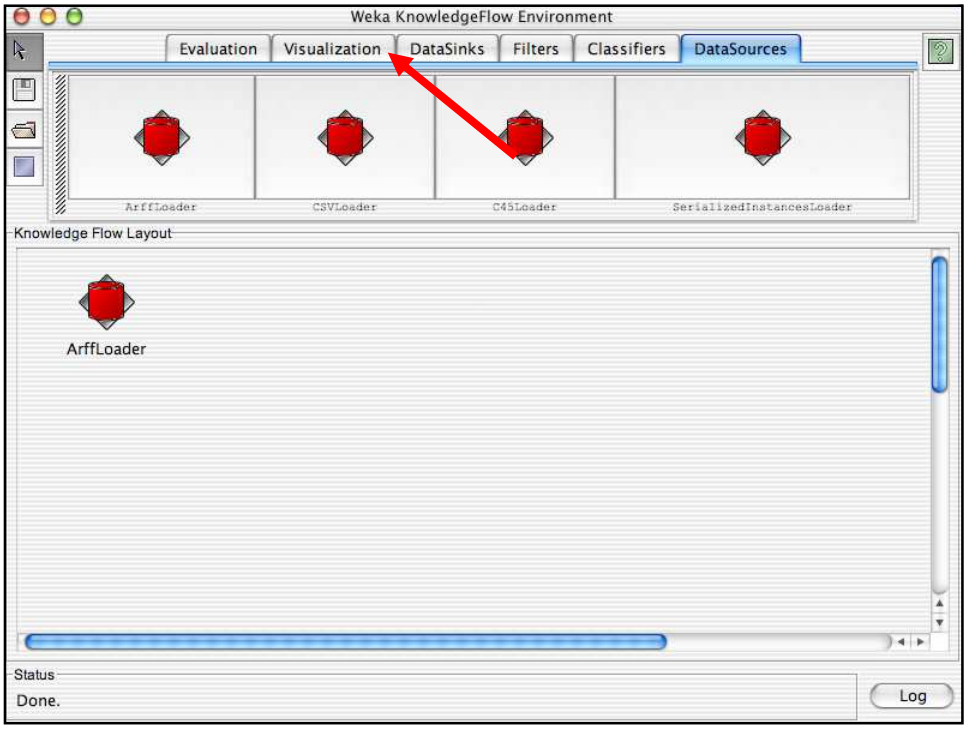
The Knowledge Flow GUI

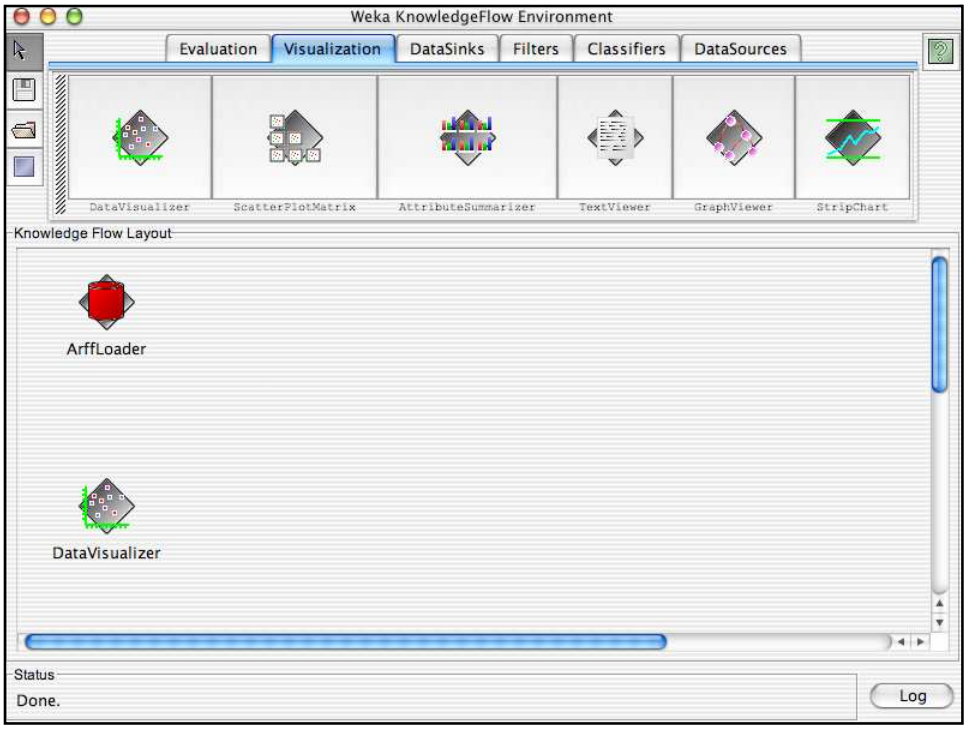
- New graphical user interface for WEKA
- Java-Beans-based interface for setting up and running machine learning experiments
- Data sources, classifiers, etc. are beans and can be connected graphically
- Data "flows" through components: e.g., "data source" -> "filter" -> "classifier" -> "evaluator"
-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

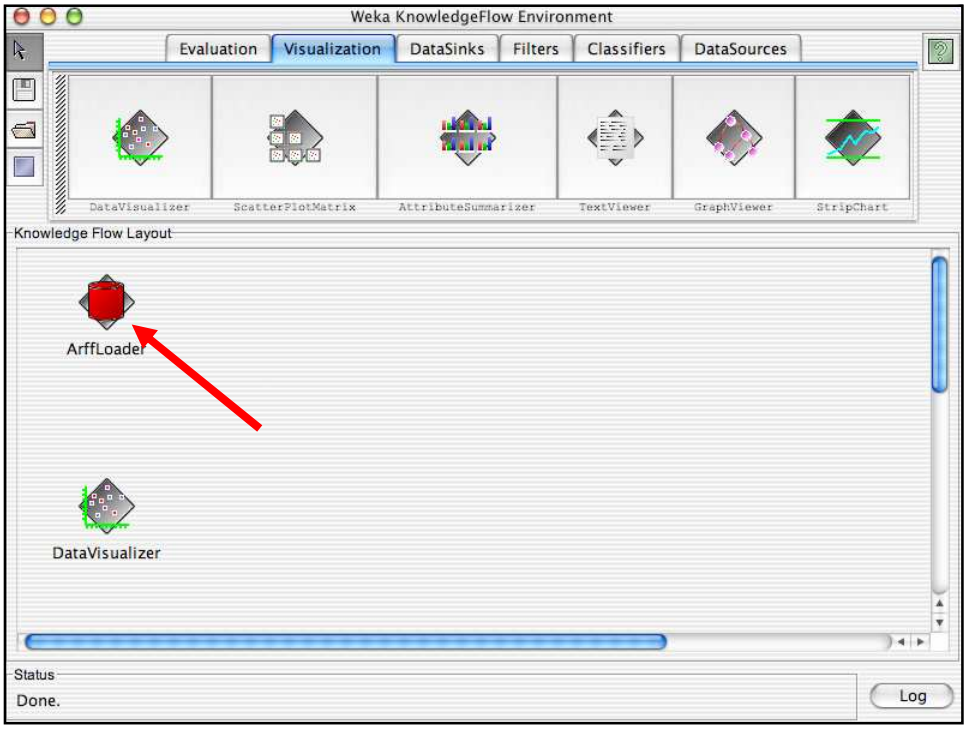

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
TrainTestSplitMaker
ClassAssigner
ClassifierPerformanceEvaluator
IncrementalClassifierEvaluator
PredictionAppender
Knowledge Flow Layout
Status
Done.
Log

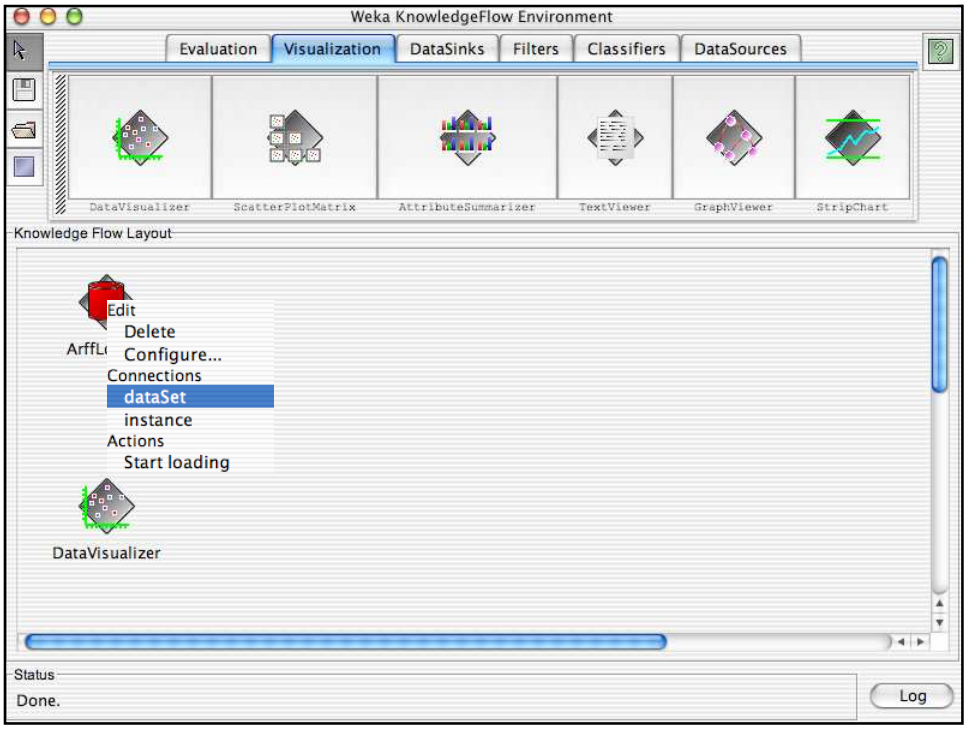

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
TrainTestSplitMaker
ClassAssigner
ClassifierPerformanceEvaluator
IncrementalClassifierEvaluator
PredictionAppender
Knowledge Flow Layout
Status
Done.
Log

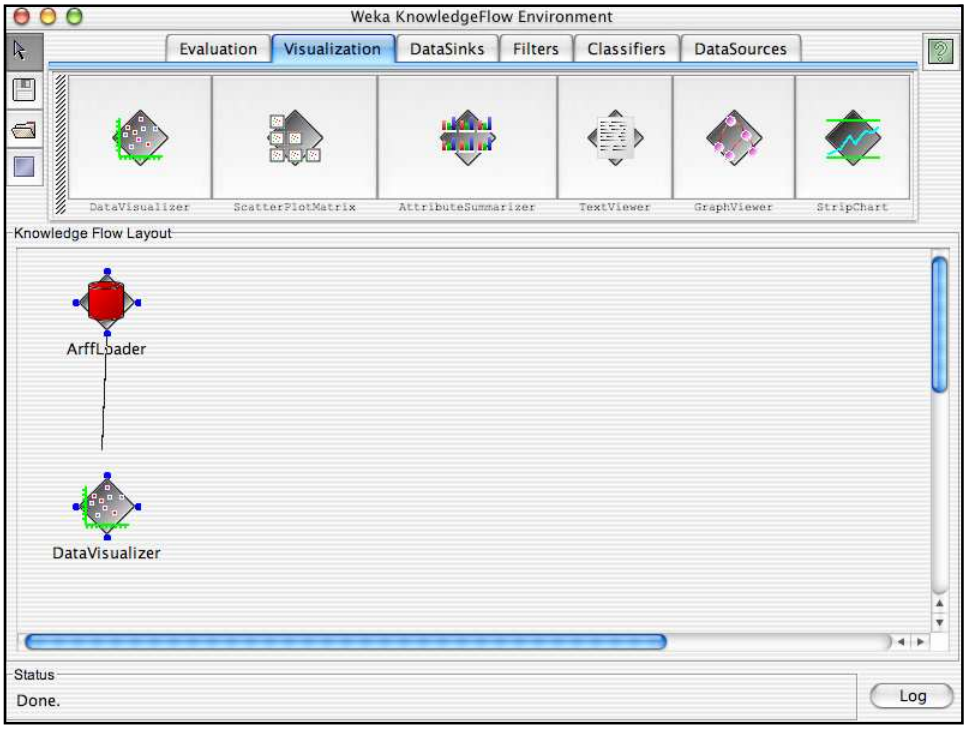

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
ArffLoader
CSVLoader
C45Loader
SerializedInstancesLoader
Knowledge Flow Layout
ArffLoader
Status
Done.
Log

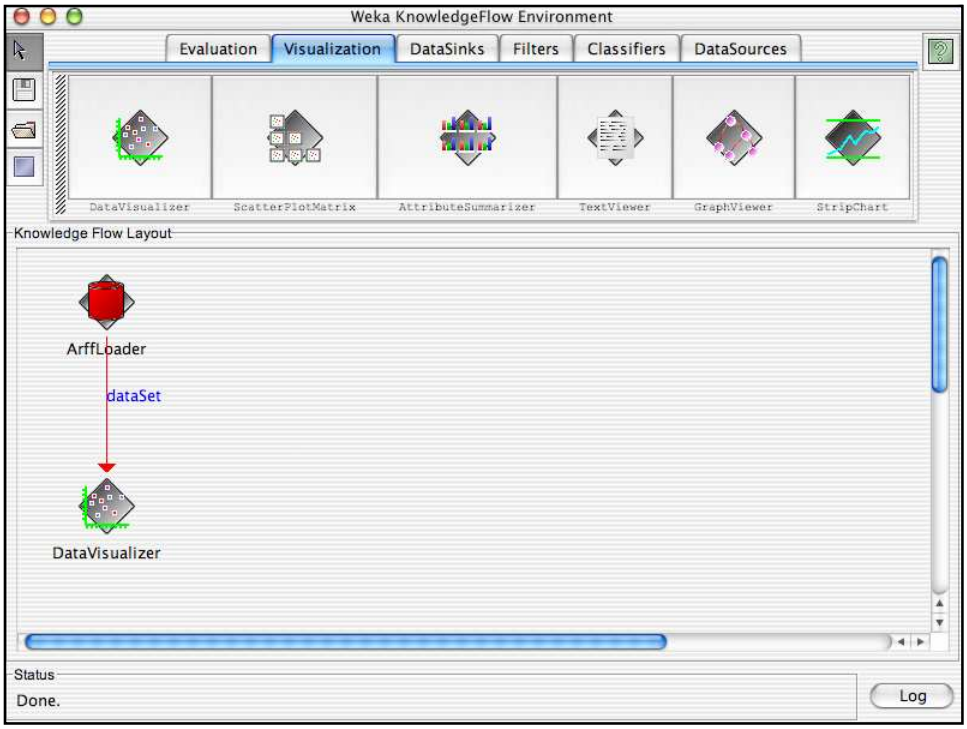

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
ArffLoader
CSVLoader
C45Loader
SerializedInstancesLoader
Knowledge Flow Layout
ArffLoader
Status
Done.
Log

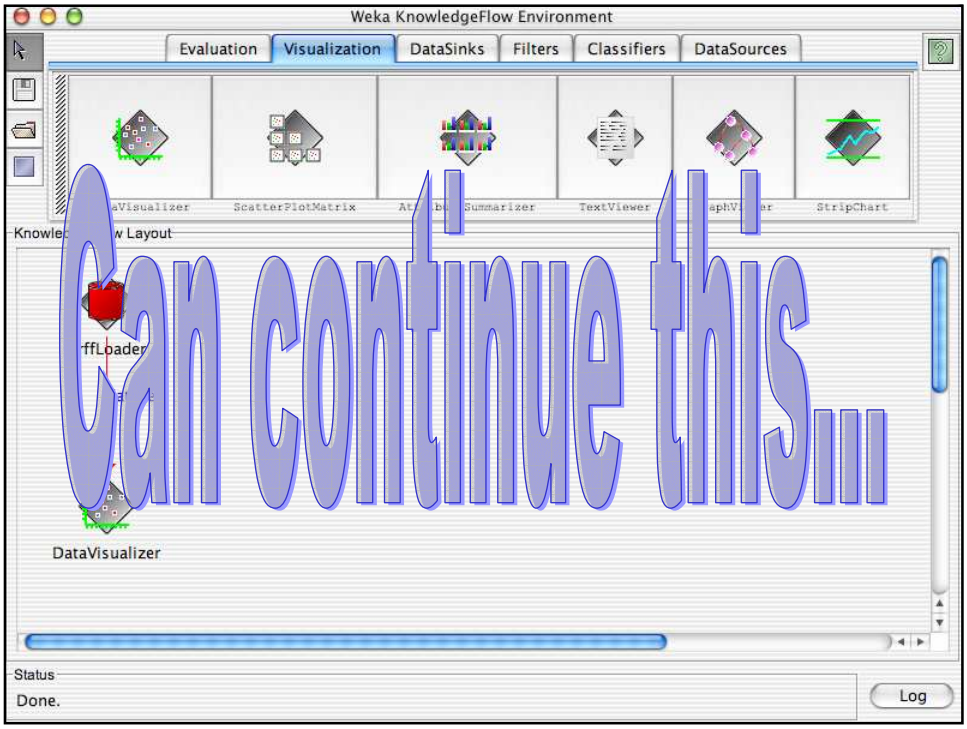

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
ArffLoader
DataVisualizer
Status
Done.
Log

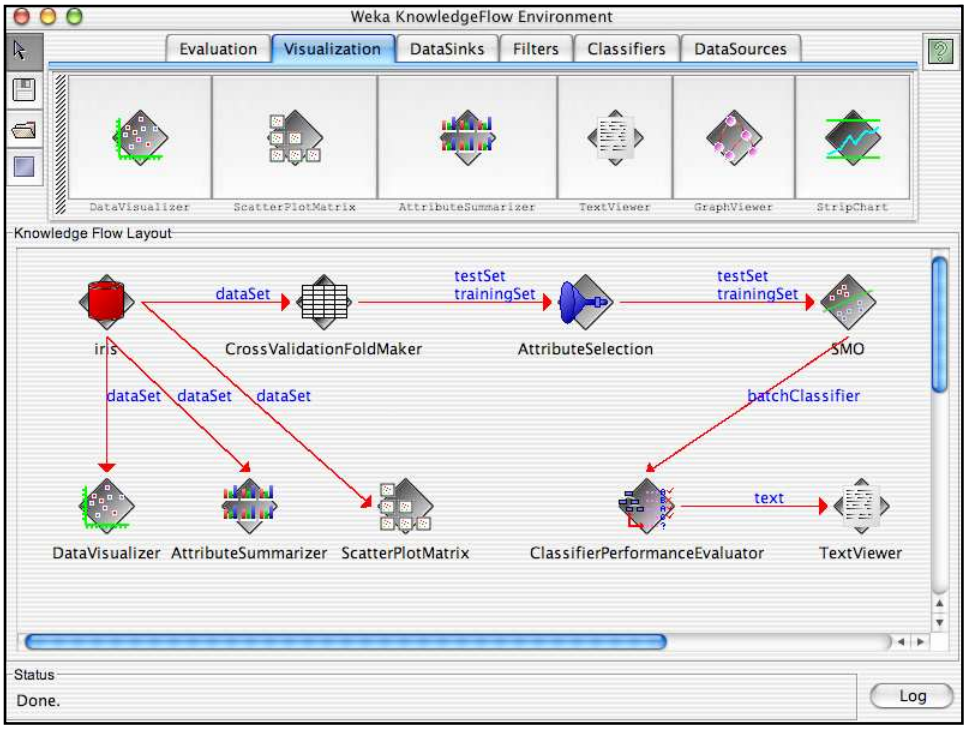

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
ArffLoader
DataVisualizer
Status
Done.
Log

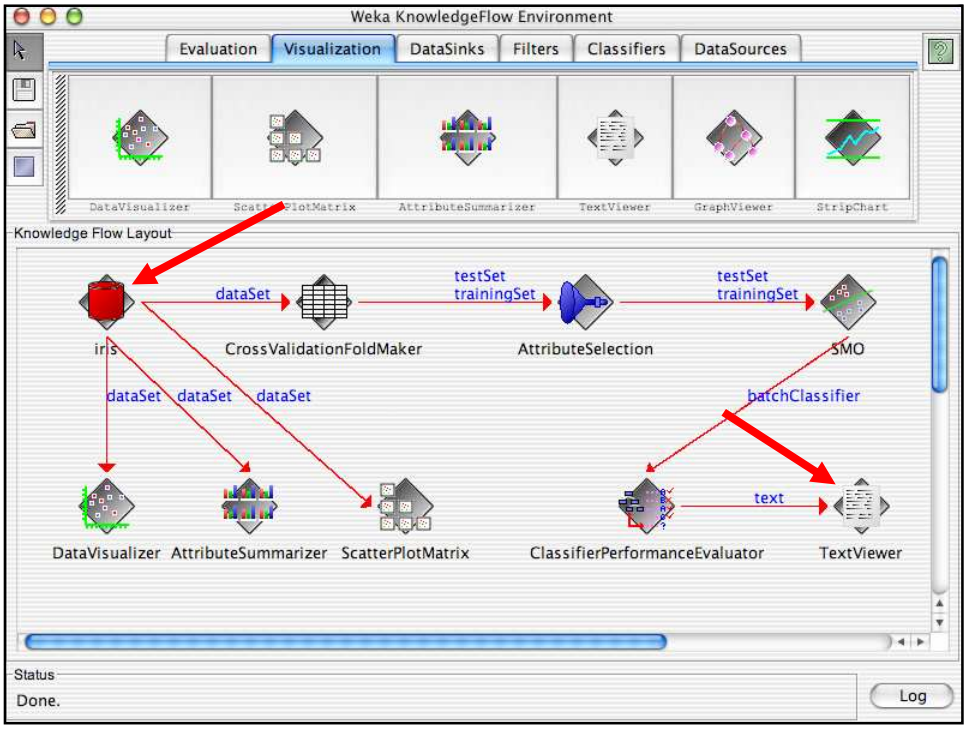

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
Edit
Delete
Configure...
Connections
dataSet
instance
Actions
Start loading
DataVisualizer
Status
Done.
Log

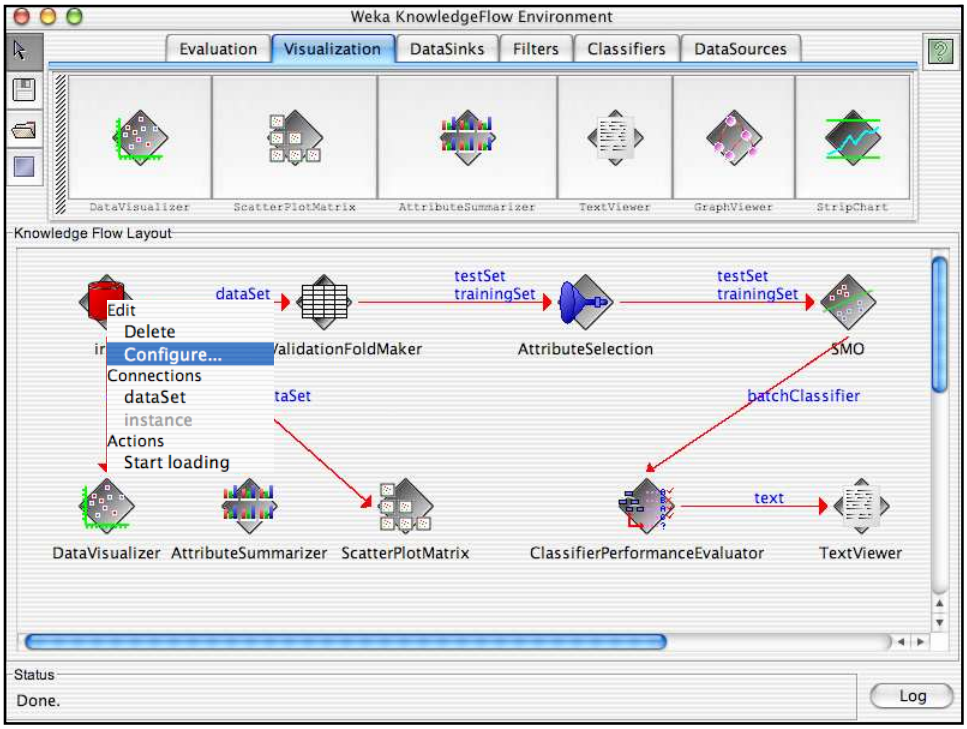

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
ArffLoader
DataVisualizer
Status
Done.
Log

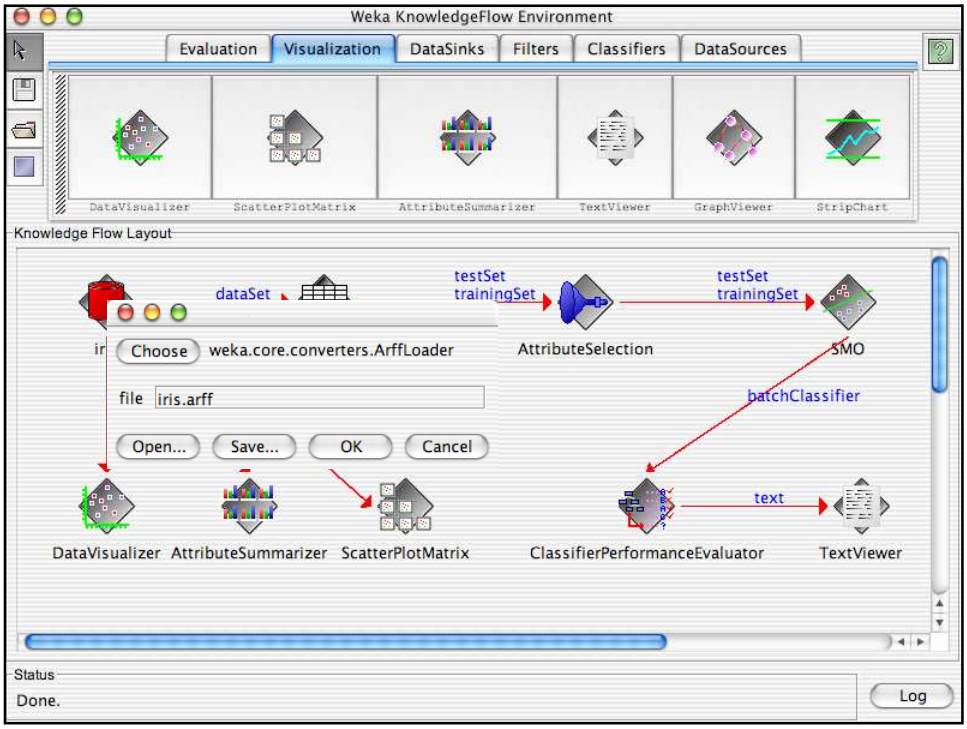

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
ArffLoader
dataSet
DataVisualizer
Status
Done.
Log

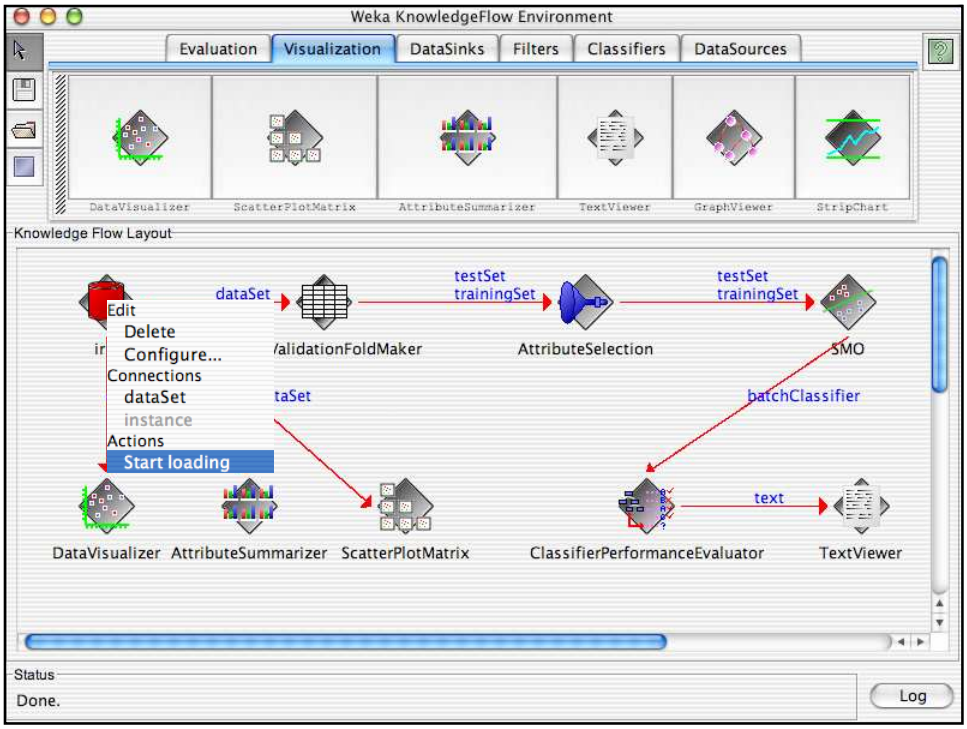

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
ScatterPlotMatrix
TextViewer
StripChart
Can continue this...
DataVisualizer
Status
Done.
Log

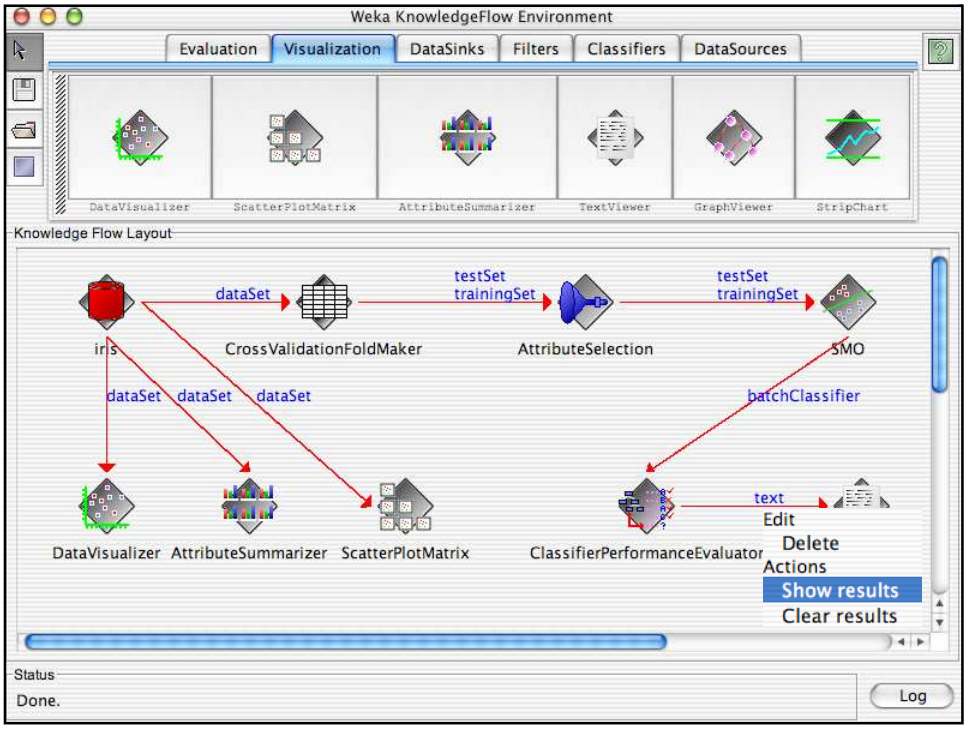

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
iris
dataSet
CrossValidationFoldMaker
testSet
trainingSet
AttributeSelection
testSet
trainingSet
SMO
dataSet
dataSet
dataSet
batchClassifier
DataVisualizer
AttributeSummarizer
ScatterPlotMatrix
ClassifierPerformanceEvaluator
text
TextViewer
Status
Done.
Log

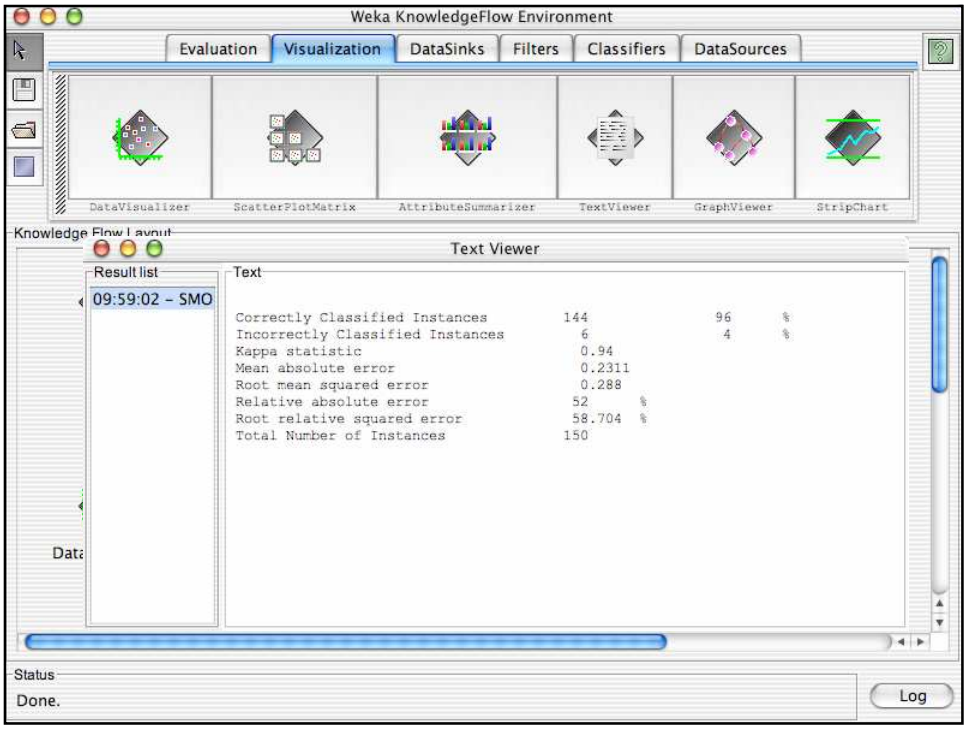

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
iris
dataSet
CrossValidationFoldMaker
testSet
trainingSet
AttributeSelection
testSet
trainingSet
SMO
dataSet
dataSet
dataSet
batchClassifier
DataVisualizer
AttributeSummarizer
ScatterPlotMatrix
ClassifierPerformanceEvaluator
text
TextViewer
Status
Done.
Log

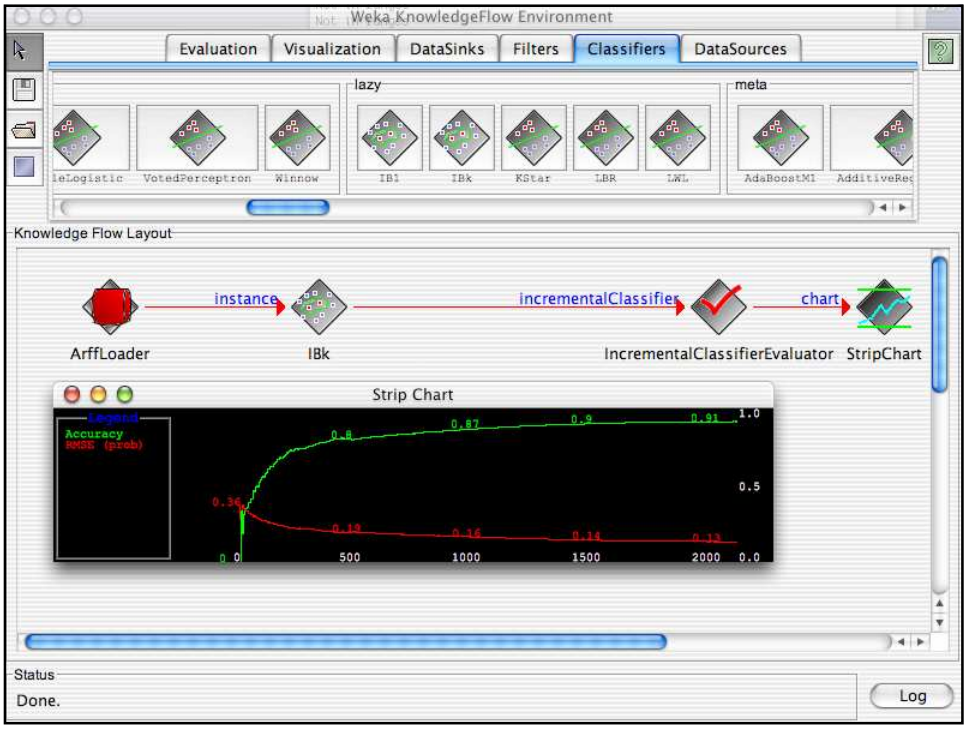

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
Edit
Delete
Configure...
Connections
dataSet
instance
Actions
Start loading
dataSet
testSet
trainingSet
AttributeSelection
SMO
batchClassifier
text
ClassifierPerformanceEvaluator
Status
Done.
Log

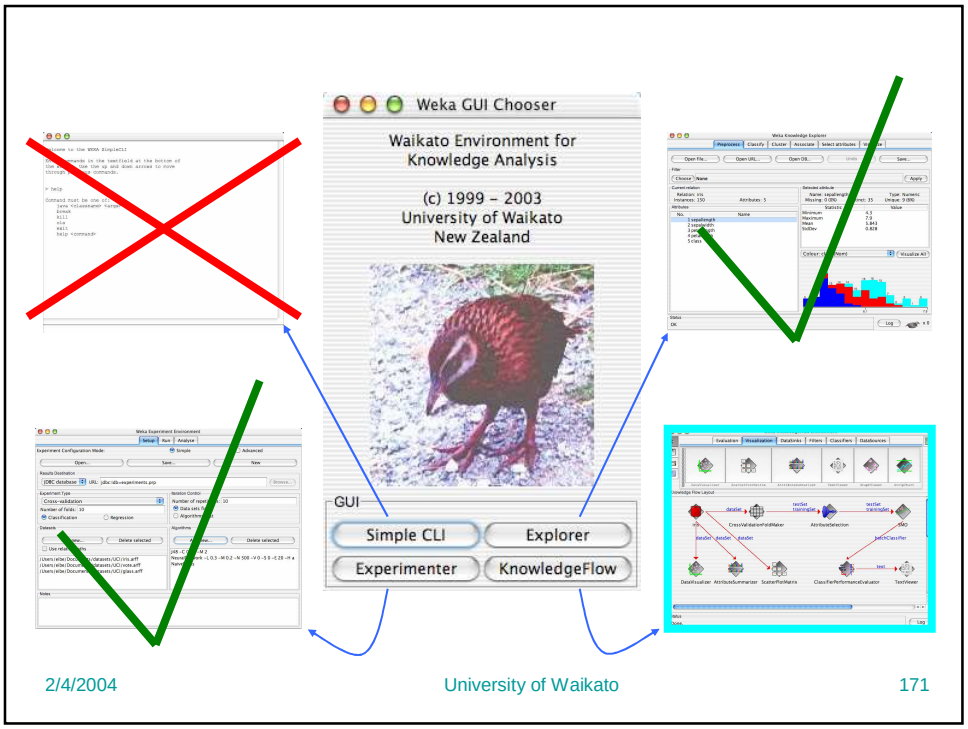

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
Knowledge Flow Layout
dataSet
testSet
trainingSet
Choose
weka.core.converters.ArffLoader
file
iris.arff
Open...
Save...
OK
Cancel
AttributeSelection
SMO
batchClassifier
text
DataVisualizer
AttributeSummarizer
ScatterPlotMatrix
ClassifierPerformanceEvaluator
TextViewer
Status
Done.
Log

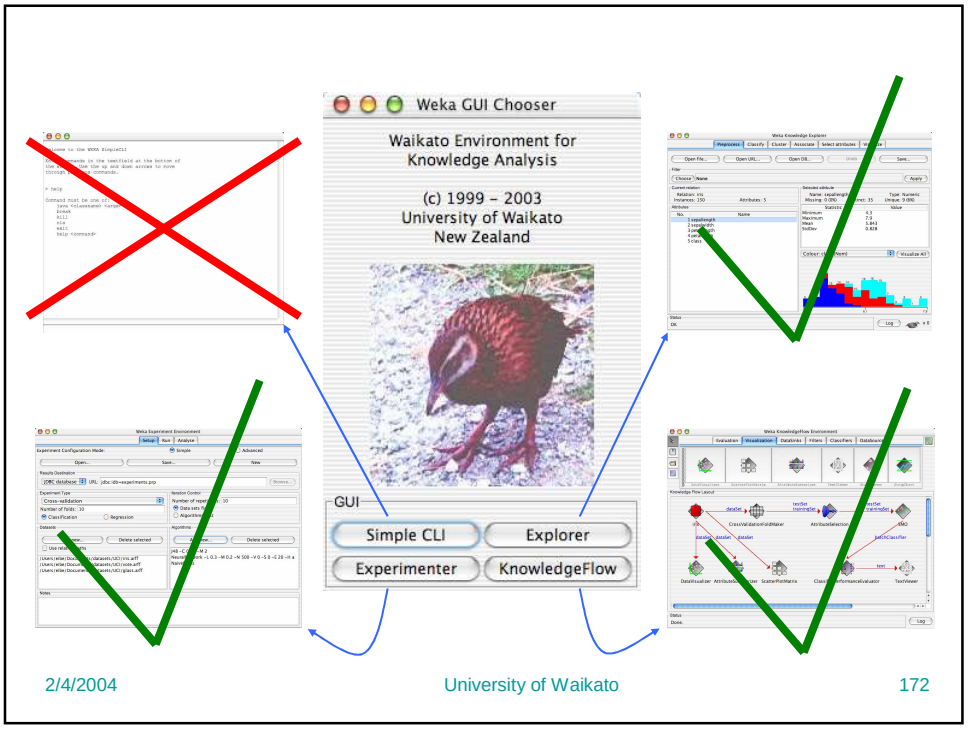

Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
Edit
Delete
Configure...
Connections
dataSet
instance
Actions
Start loading
dataSet
testSet
trainingSet
testSet
trainingSet
AttributeSelection
SMO
batchClassifier
text
DataVisualizer
AttributeSummarizer
ScatterPlotMatrix
ClassifierPerformanceEvaluator
TextViewer
Status
Done.
Log


Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
dataSet
testSet
trainingSet
testSet
trainingSet
CrossValidationFoldMaker
AttributeSelection
SMO
dataSet
dataSet
dataSet
batchClassifier
text
Edit
Delete
Actions
Show results
Clear results
DataVisualizer
AttributeSummarizer
ScatterPlotMatrix
ClassifierPerformanceEvaluator
Status
Done.
Log


Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
DataVisualizer
ScatterPlotMatrix
AttributeSummarizer
TextViewer
GraphViewer
StripChart
Knowledge Flow Layout
Text Viewer
Result list
09:59:02 - SMO
Text
Correctly Classified Instances 144 96 %
Incorrectly Classified Instances 6 4 %
Kappa statistic 0.94
Mean absolute error 0.2311
Root mean squared error 0.288
Relative absolute error 52 %
Root relative squared error 58.704 %
Total Number of Instances 150
Status
Done.
Log


Weka KnowledgeFlow Environment
Evaluation
Visualization
DataSinks
Filters
Classifiers
DataSources
lazy
meta
IB1
IBk
KStar
LBR
LWL
AdaBoostM1
Knowledge Flow Layout
instance
incrementalClassifier
chart
ArffLoader
IBk
IncrementalClassifierEvaluator
StripChart
Strip Chart
Legend
Accuracy
RMSE (prob)
Status
Done.
Log


Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow


Weka GUI Chooser
Waikato Environment for
Knowledge Analysis
(c) 1999 – 2003
University of Waikato
New Zealand
GUI
Simple CLI
Explorer
Experimenter
KnowledgeFlow

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, Zhihai Wang