An Introduction to WEKA

Lecture by Limsoon Wong

Slides prepared by Dong Difeng

Lecture at National Yang Ming University, June 2006

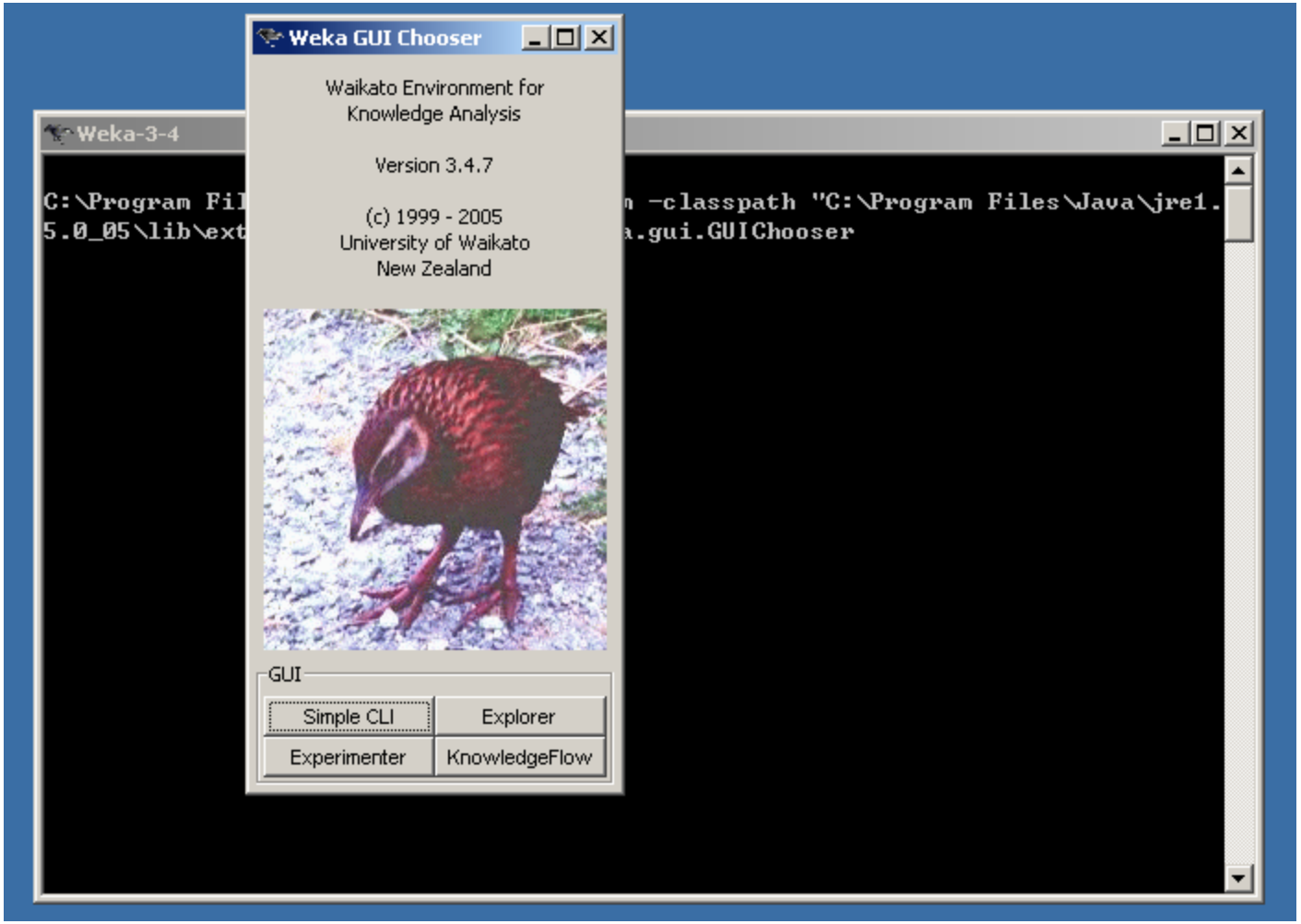
Outline

- **What is WEKA**
- **Knowledge Flow**
- **Explorer**
- **Why Knowledge Flow**
- **Cross Validation**
- **Reference**

What is WEKA

- **Developed at Univ of Waikato in New Zealand**
- **A collection of state-of-art machine learning algorithms and data preprocessing tools**
- **Provide implementation of**
  - Regression
  - Classification
  - Clustering
  - Association rules
  - Feature selection

What is WEKA



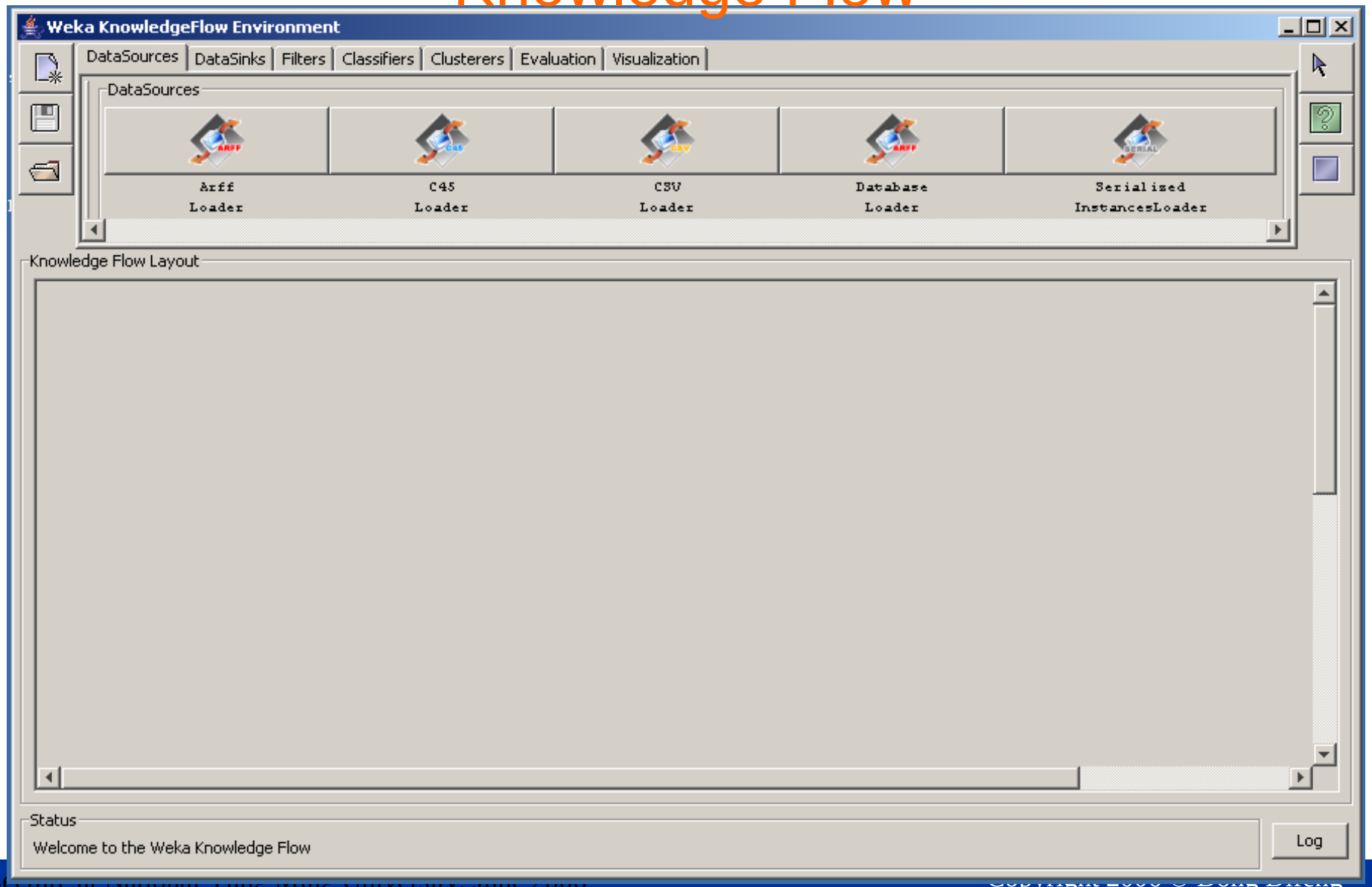
Outline

- What is WEKA
- Knowledge Flow
- Explorer
- Why Knowledge Flow
- Cross Validation
- Reference

Knowledge Flow

- **Experiment 1:**
  - Type: Classification
  - Feature selection: GainRatio; Ranker (top 3)
  - Algorithm: ID3
  - Training: Weather_nominal.arff
  - Test: Weather_nominal.arff

Knowledge Flow

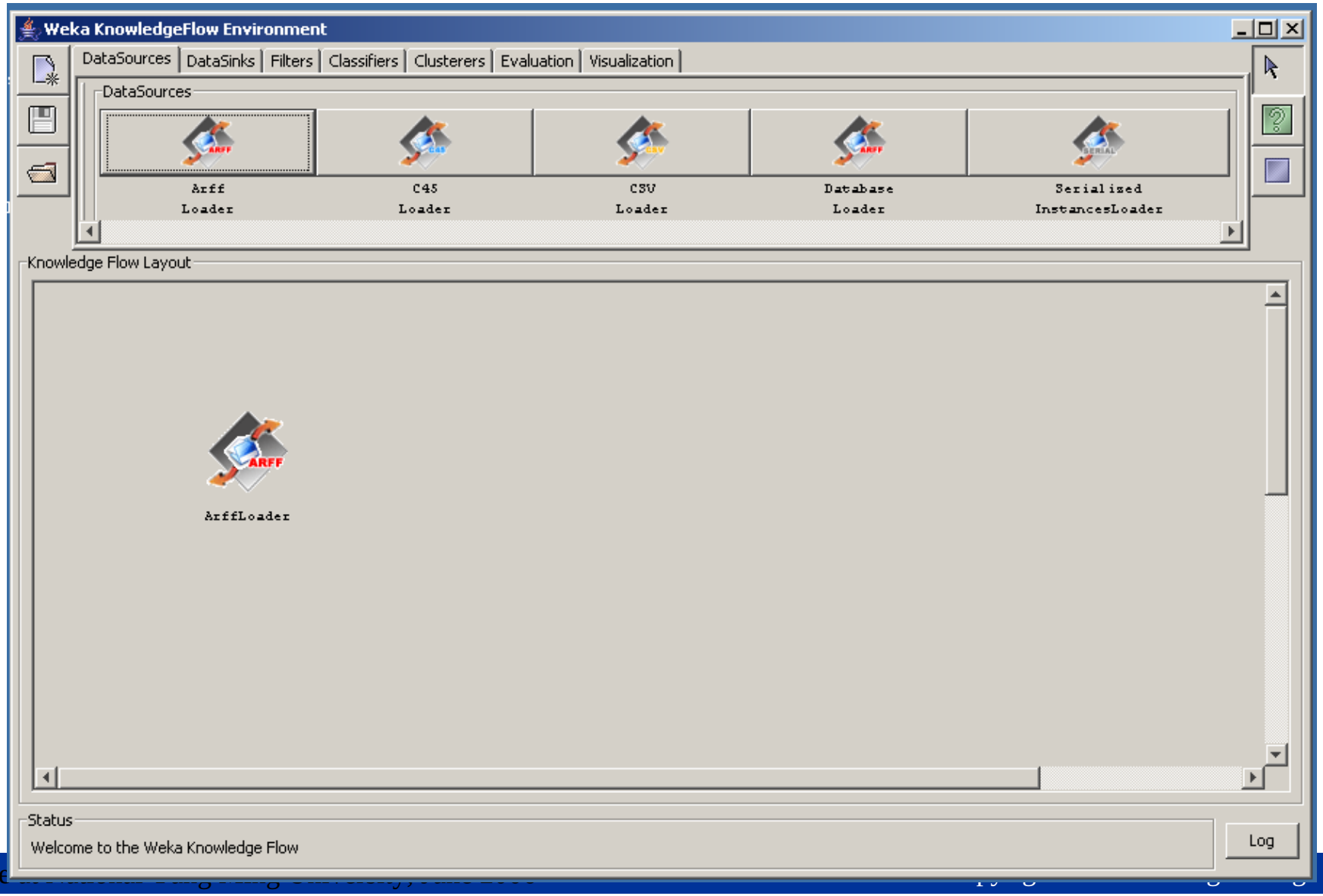


Knowledge Flow

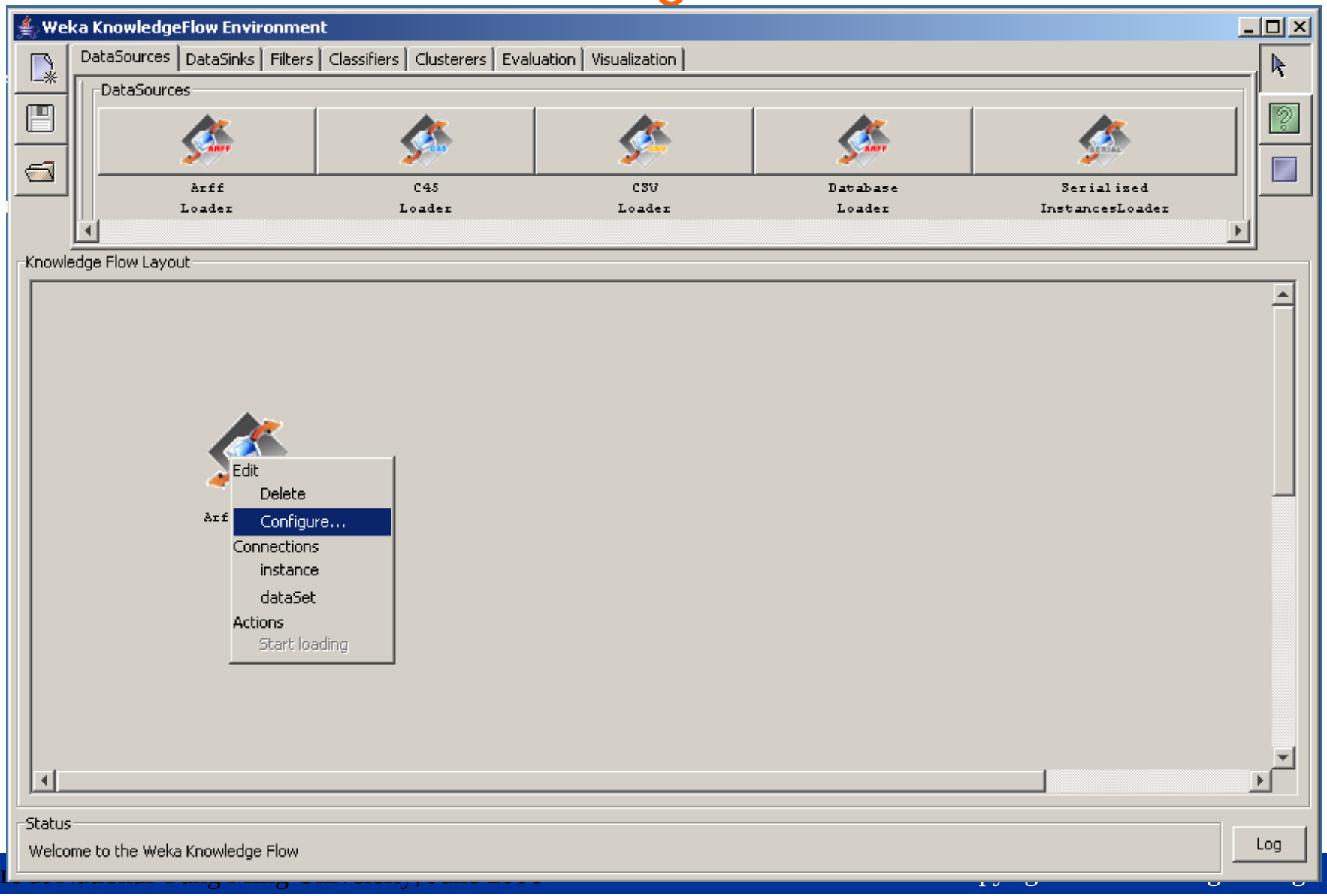
- **Source file (.ARFF)**

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no

```

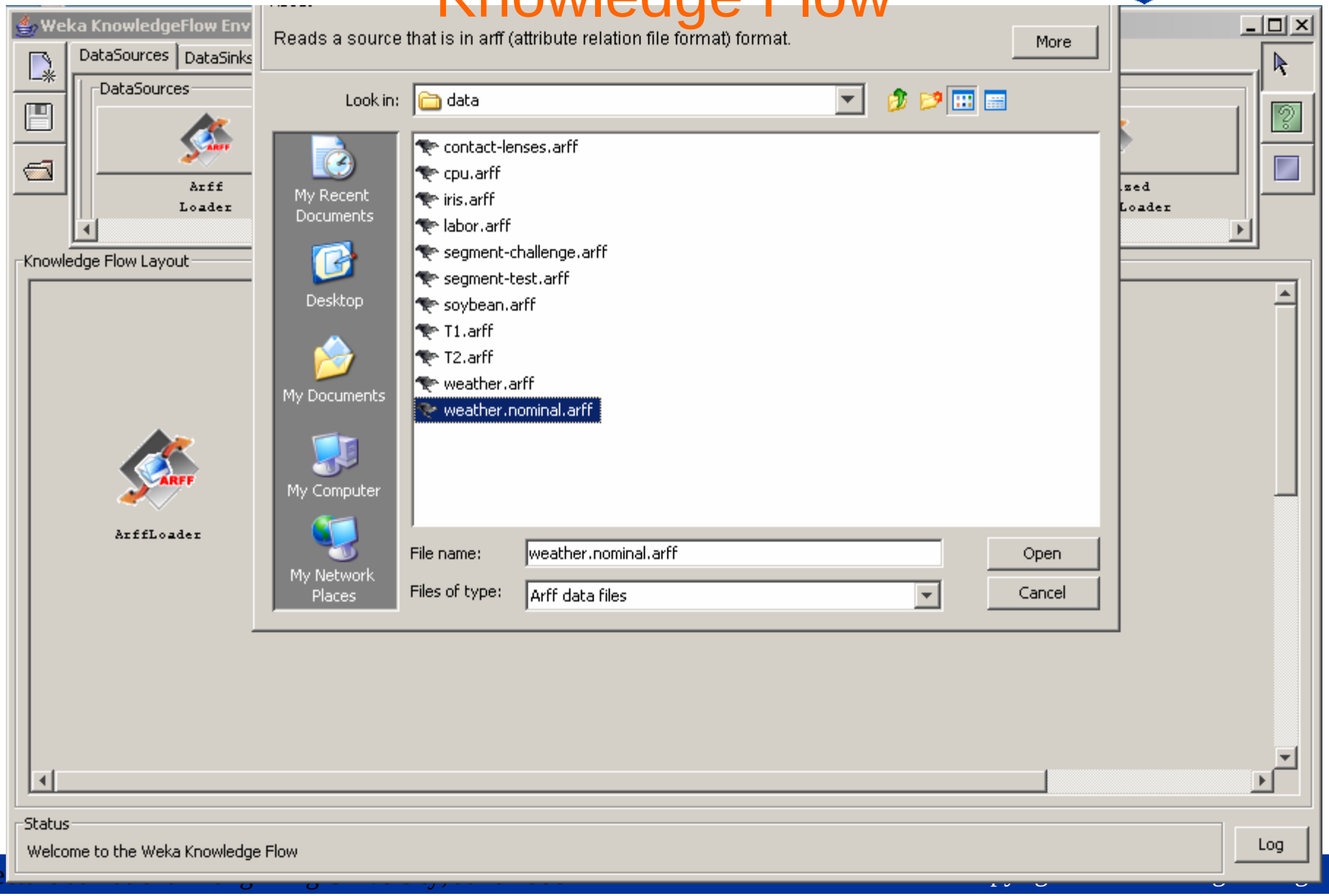
Knowledge Flow



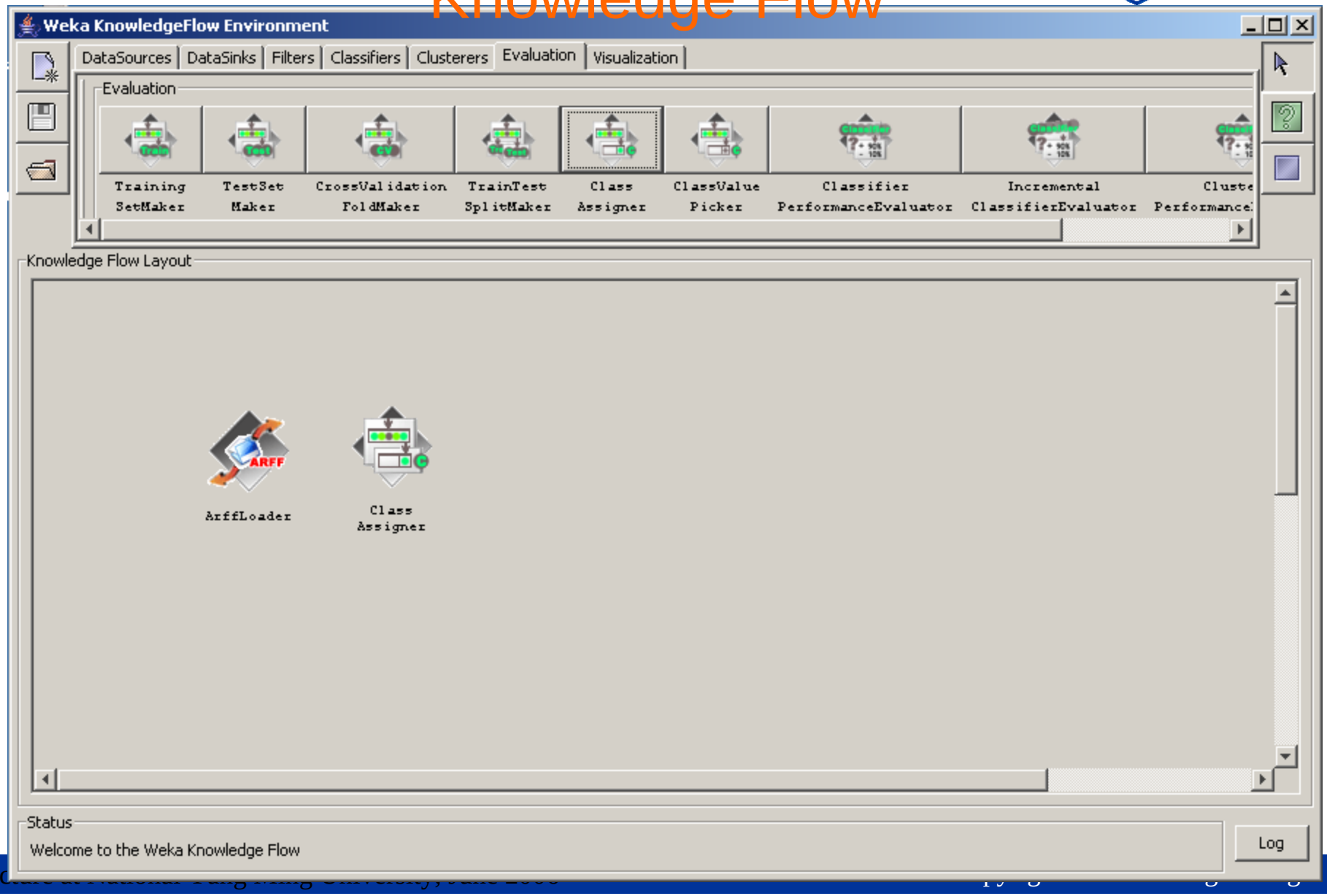
Knowledge Flow



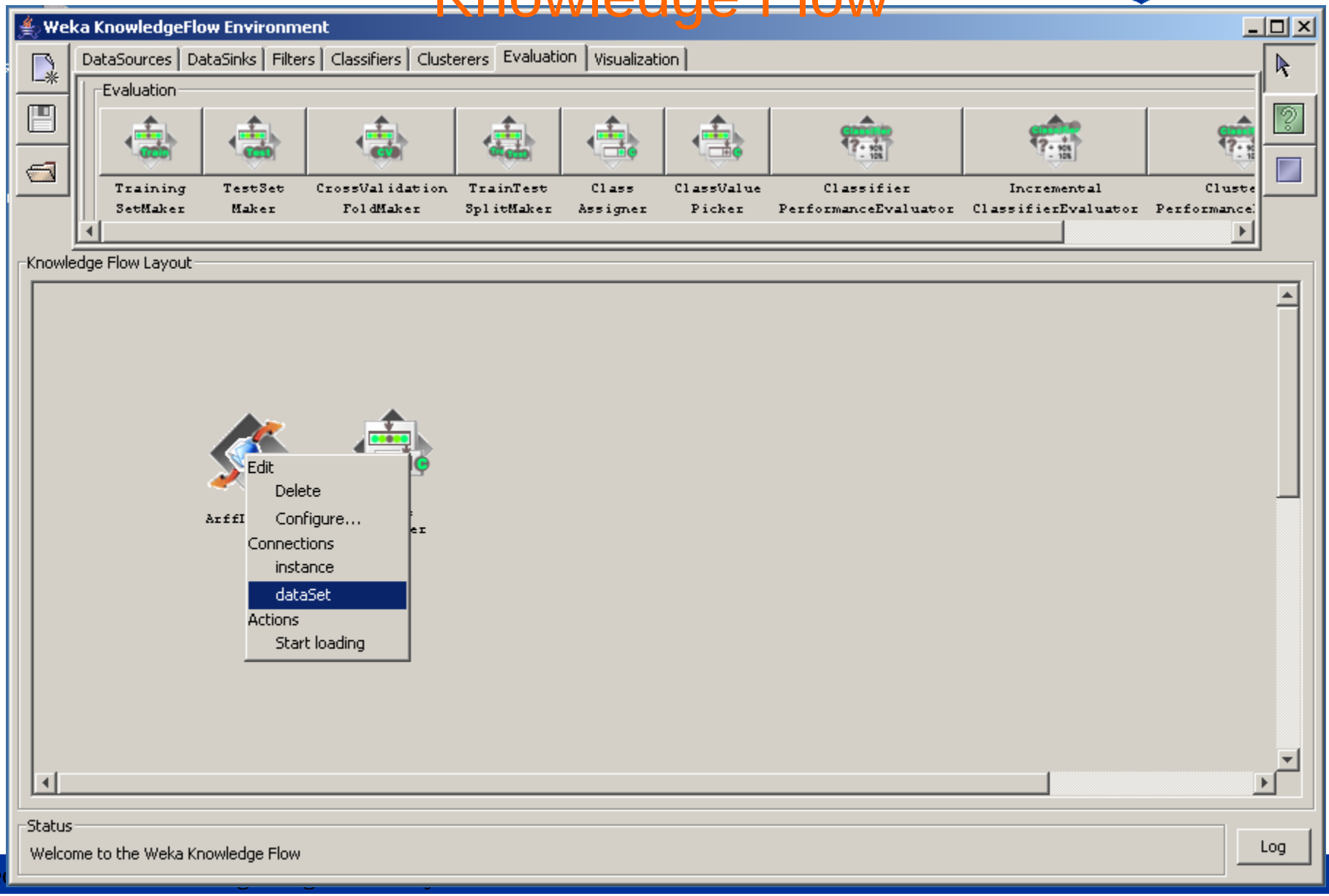
Knowledge Flow



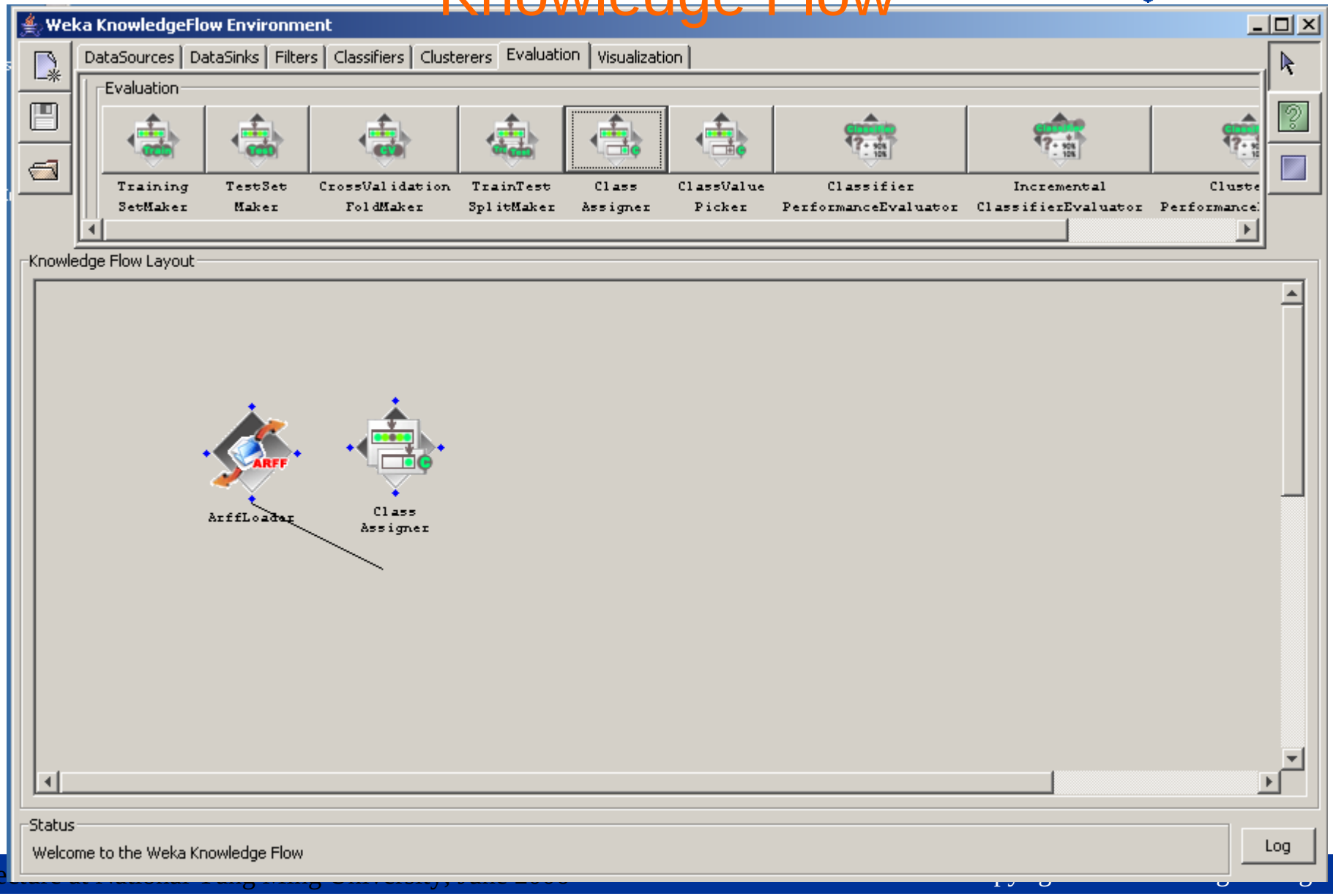
Knowledge Flow



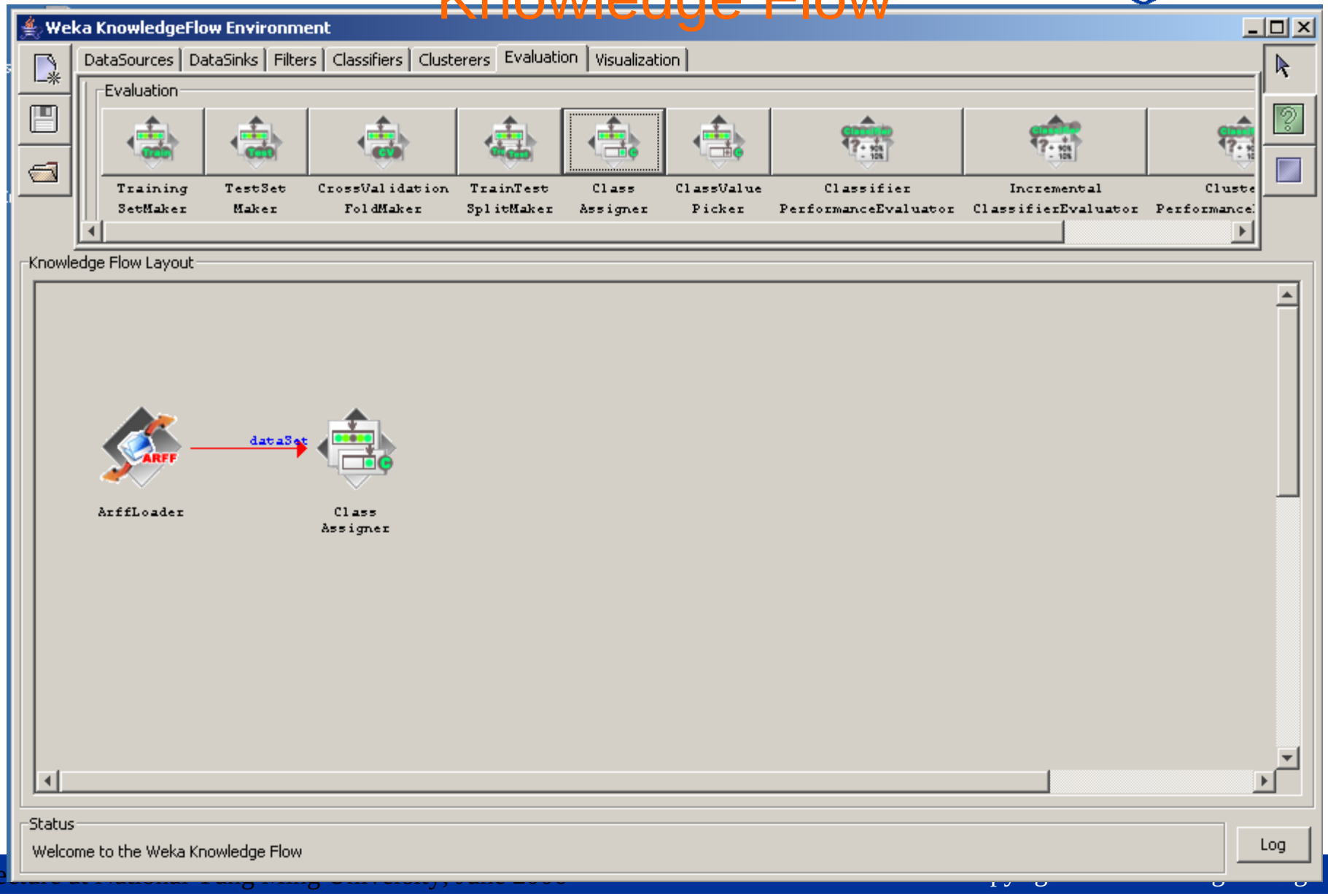
Knowledge Flow



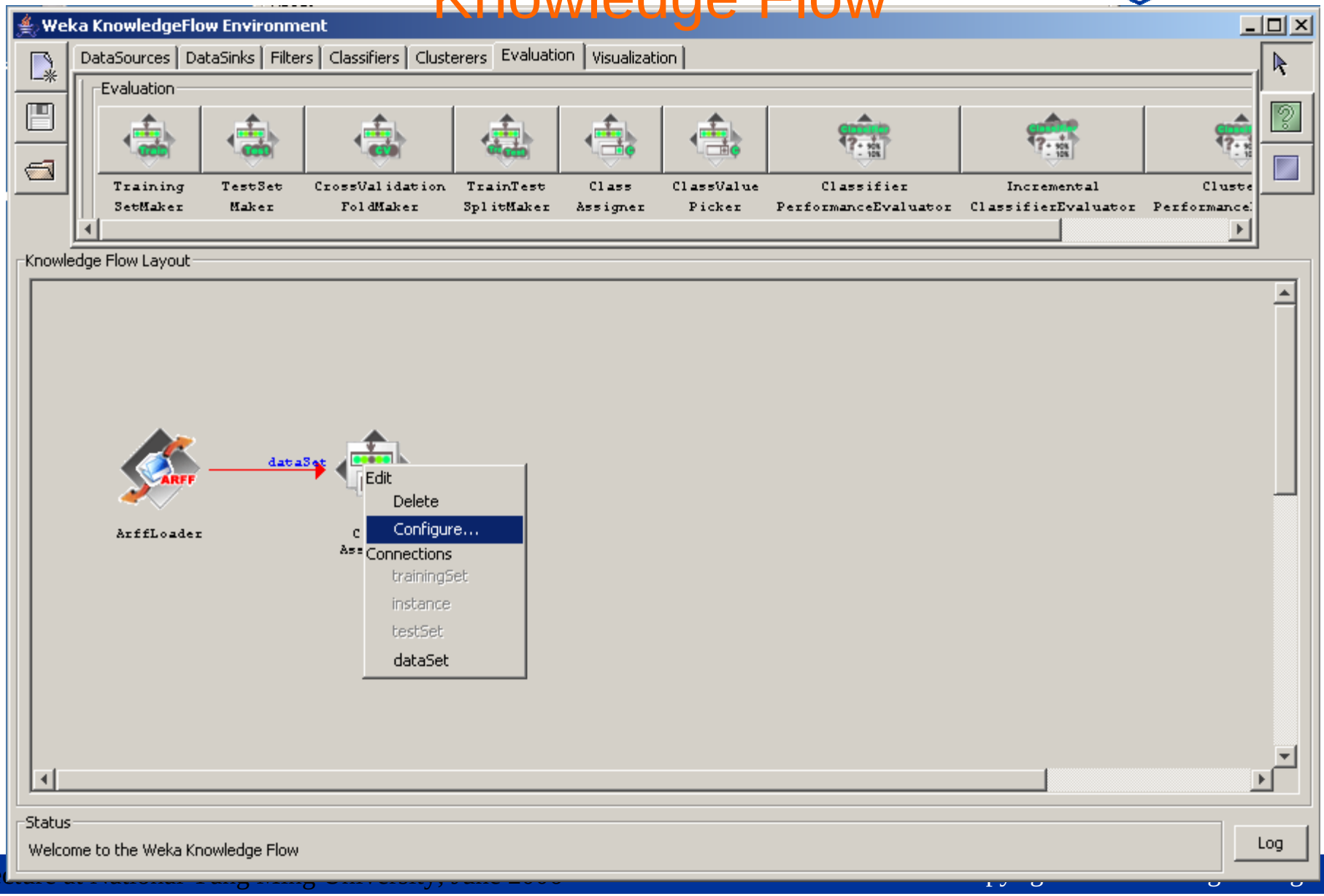
Knowledge Flow



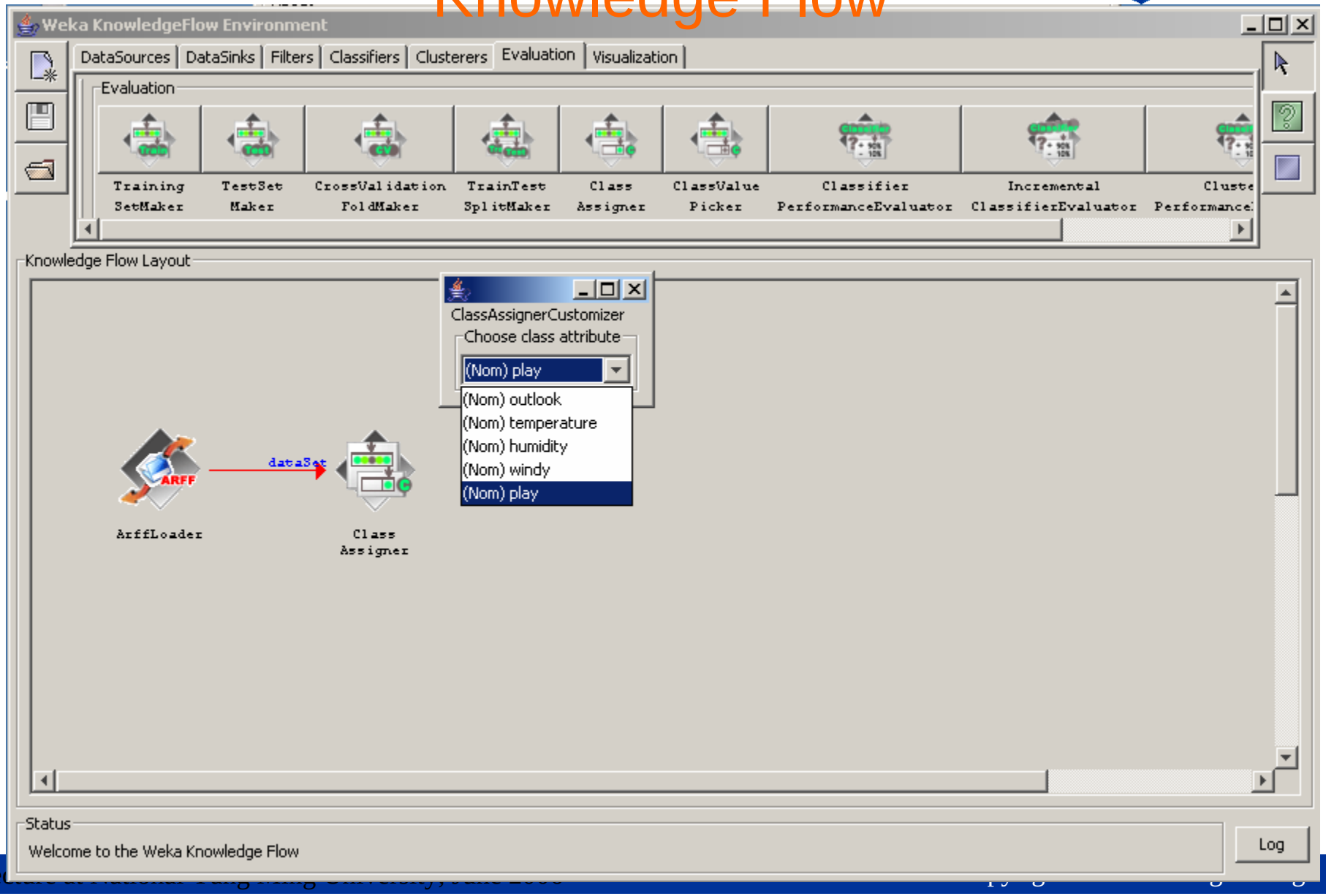
Knowledge Flow



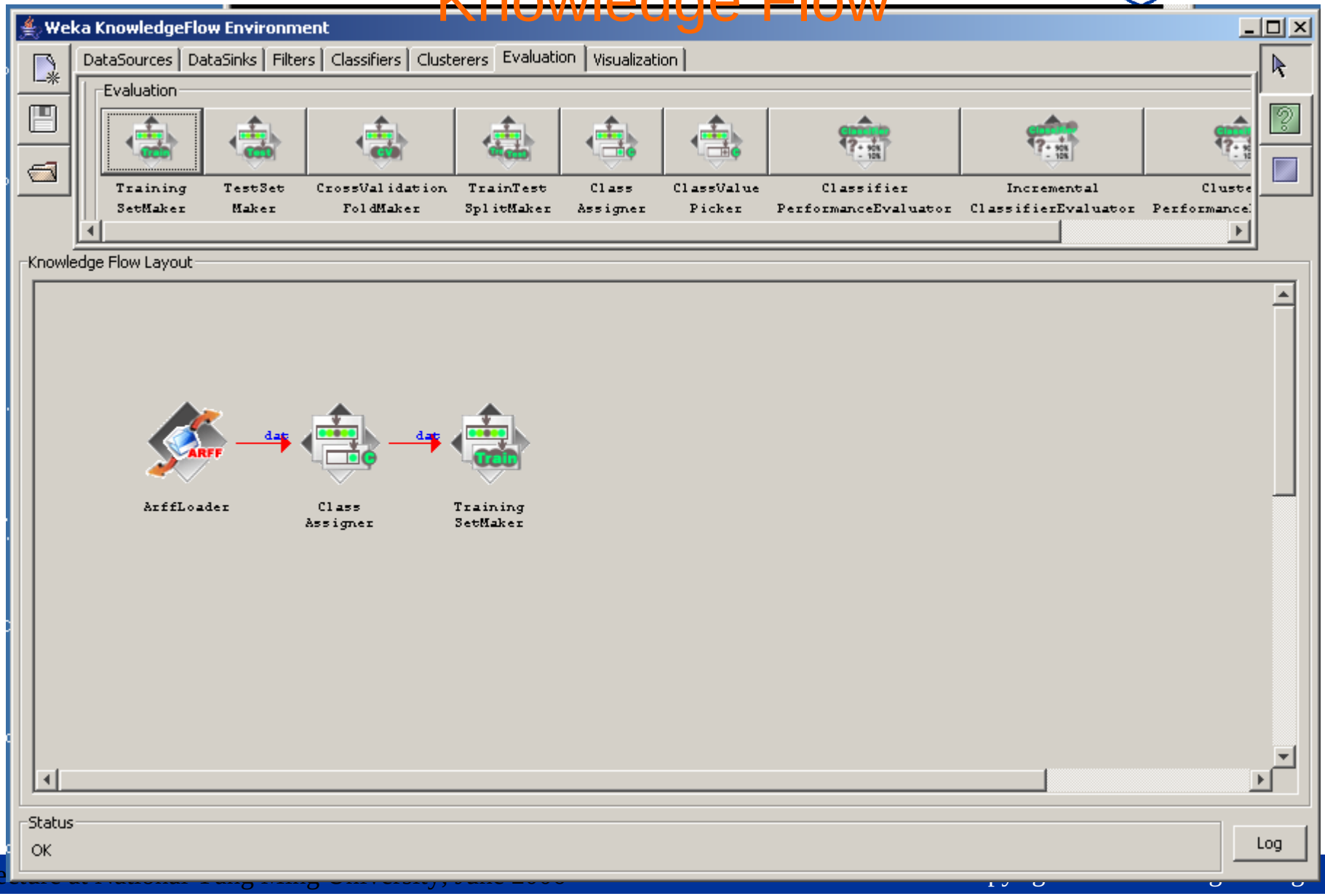
Knowledge Flow



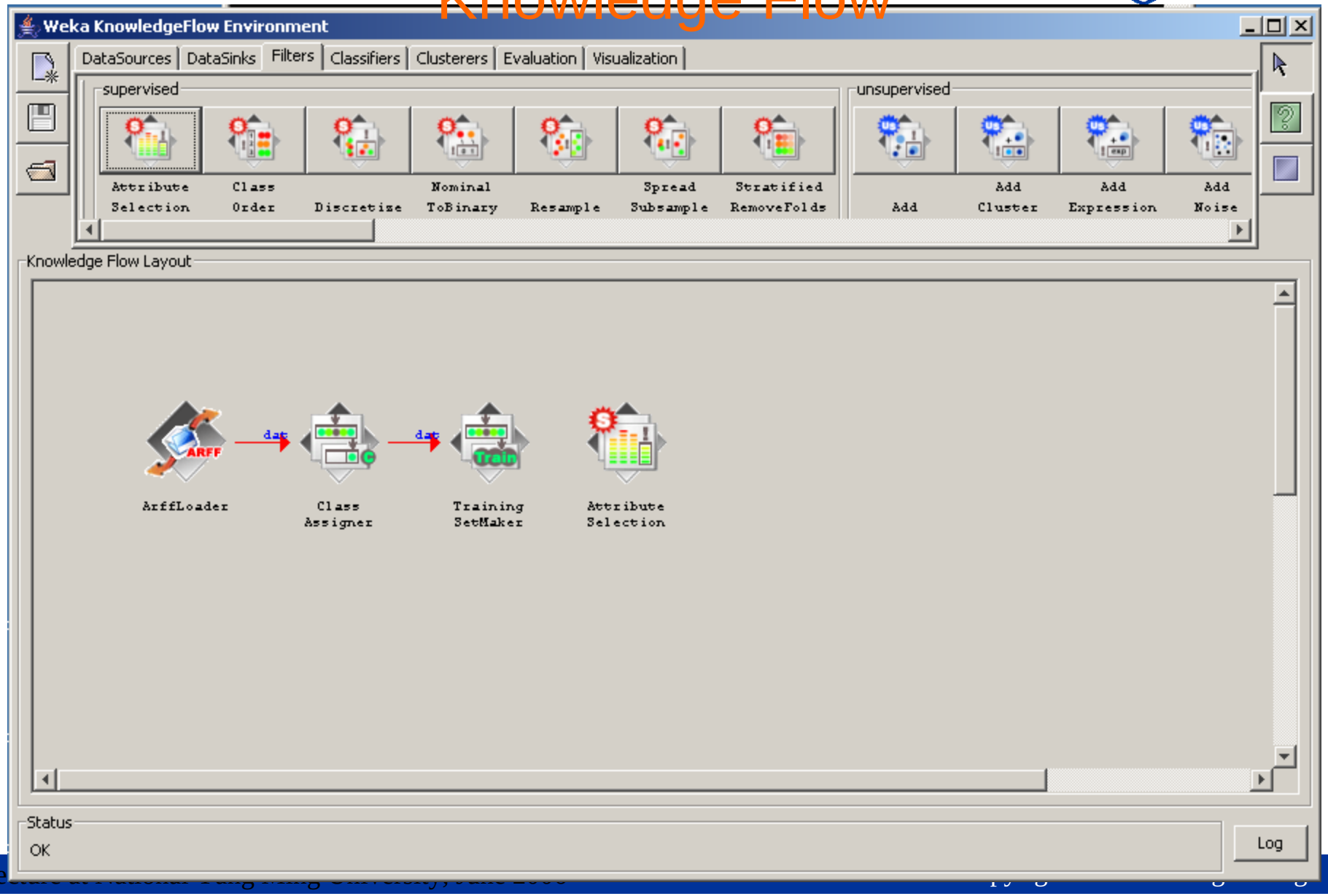
Knowledge Flow



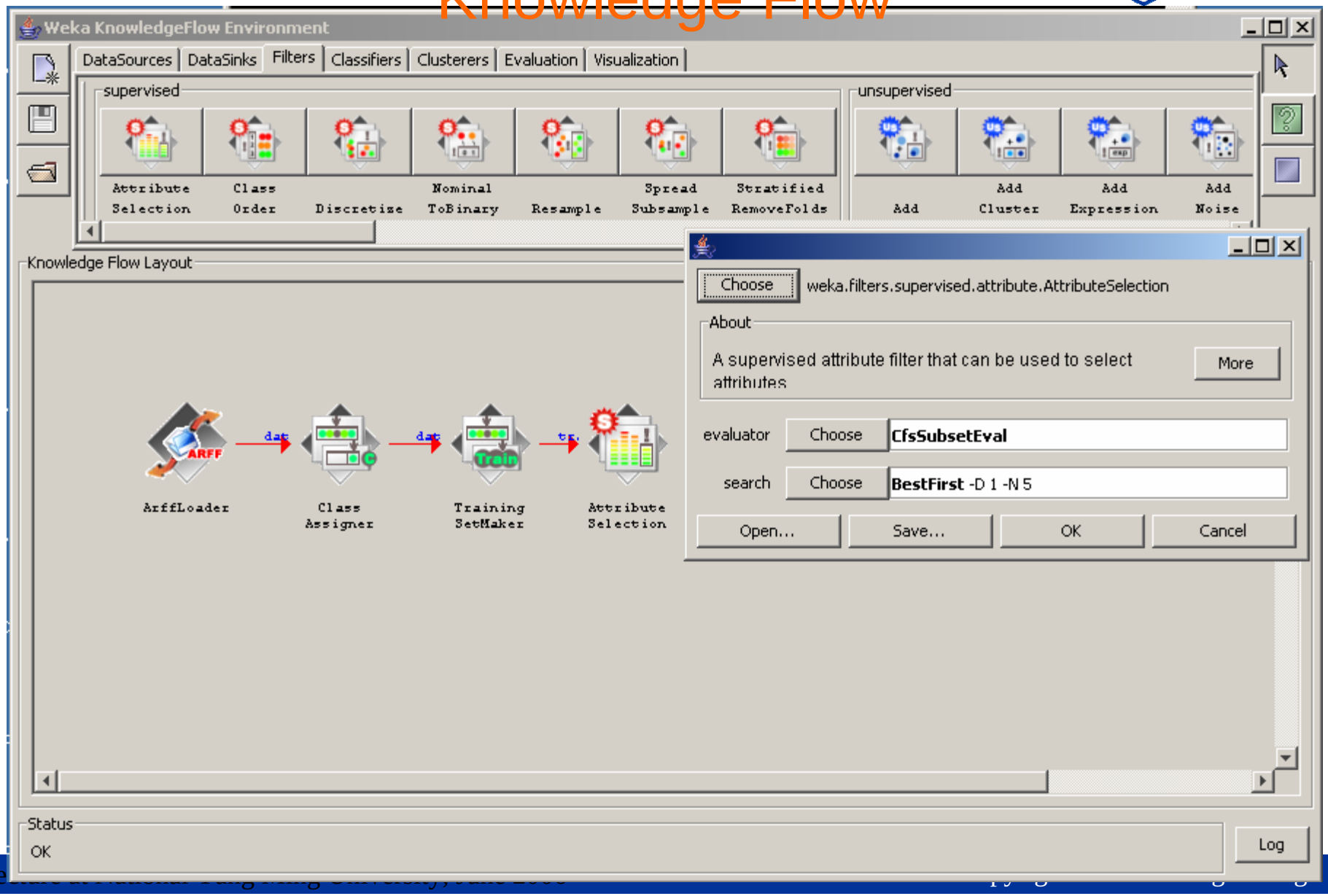
Knowledge Flow



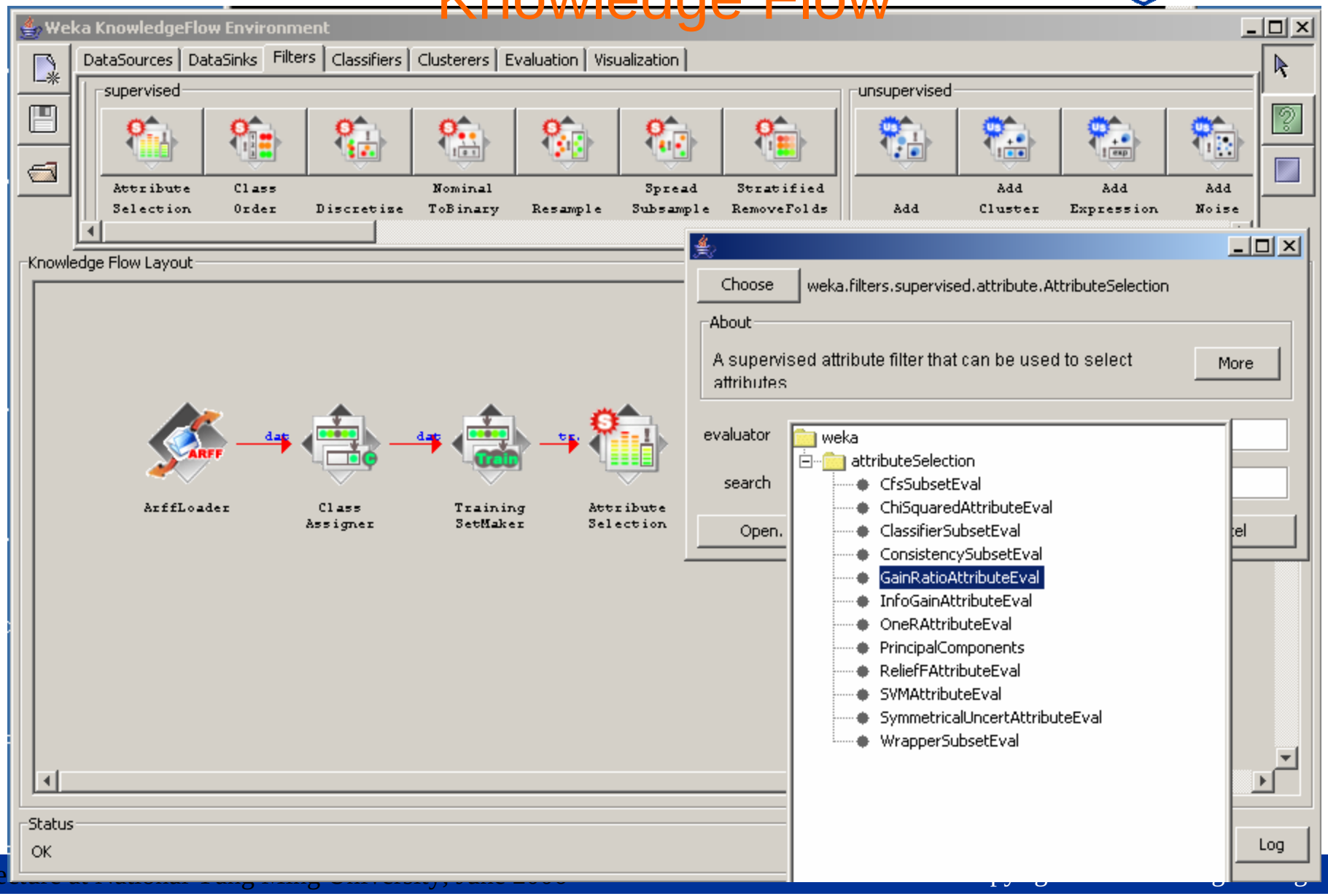
Knowledge Flow



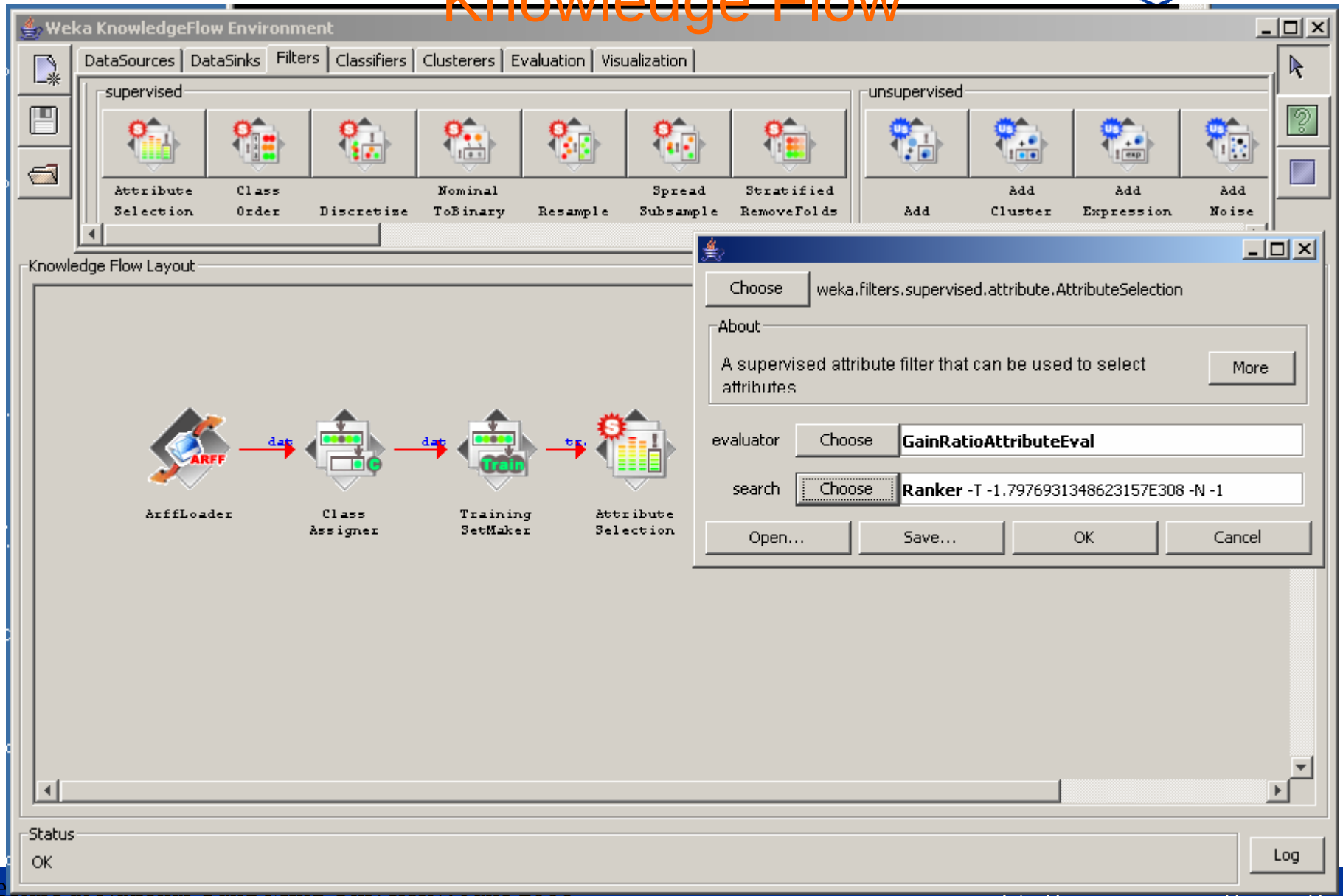
Knowledge Flow



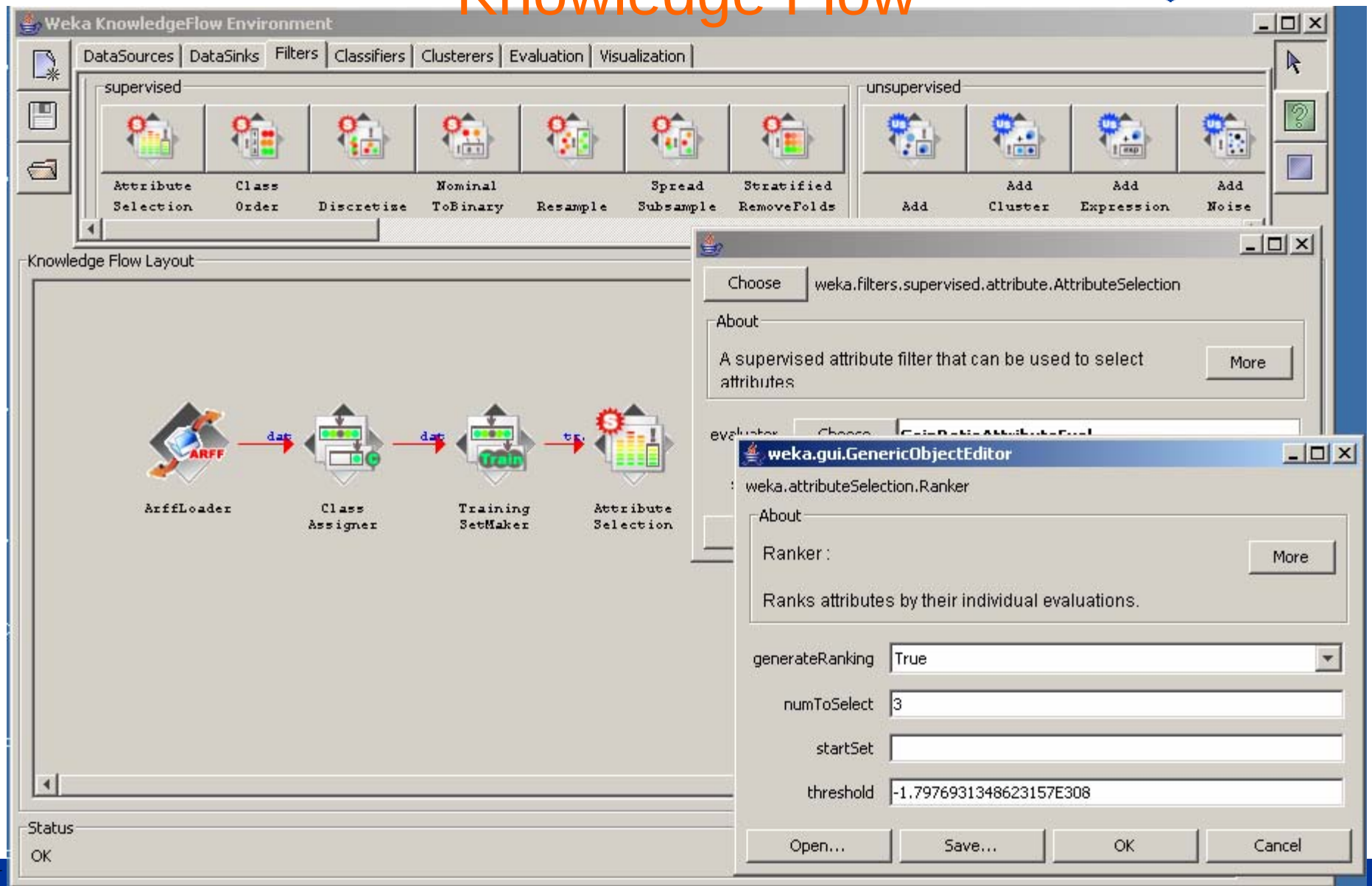
Knowledge Flow



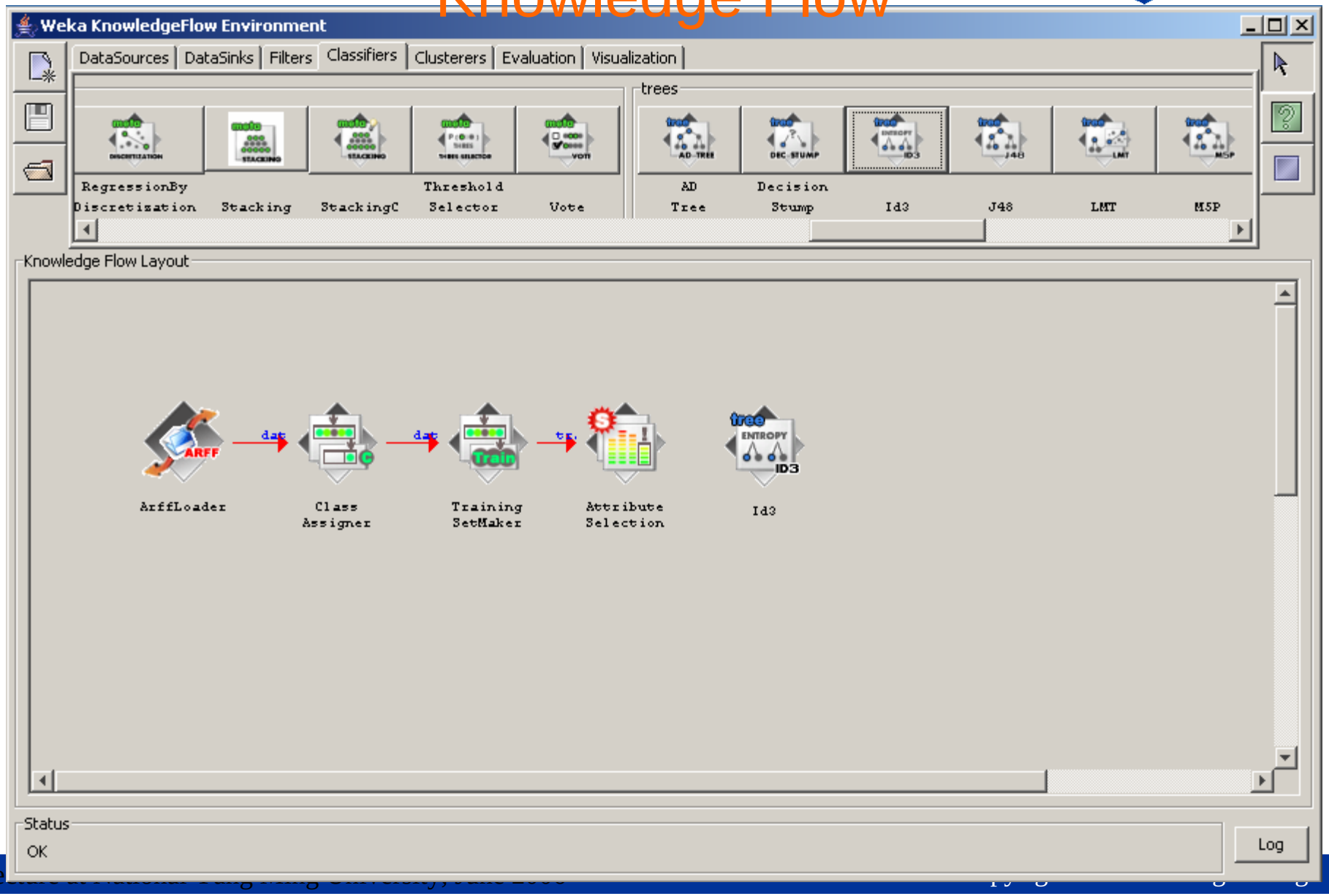
Knowledge Flow



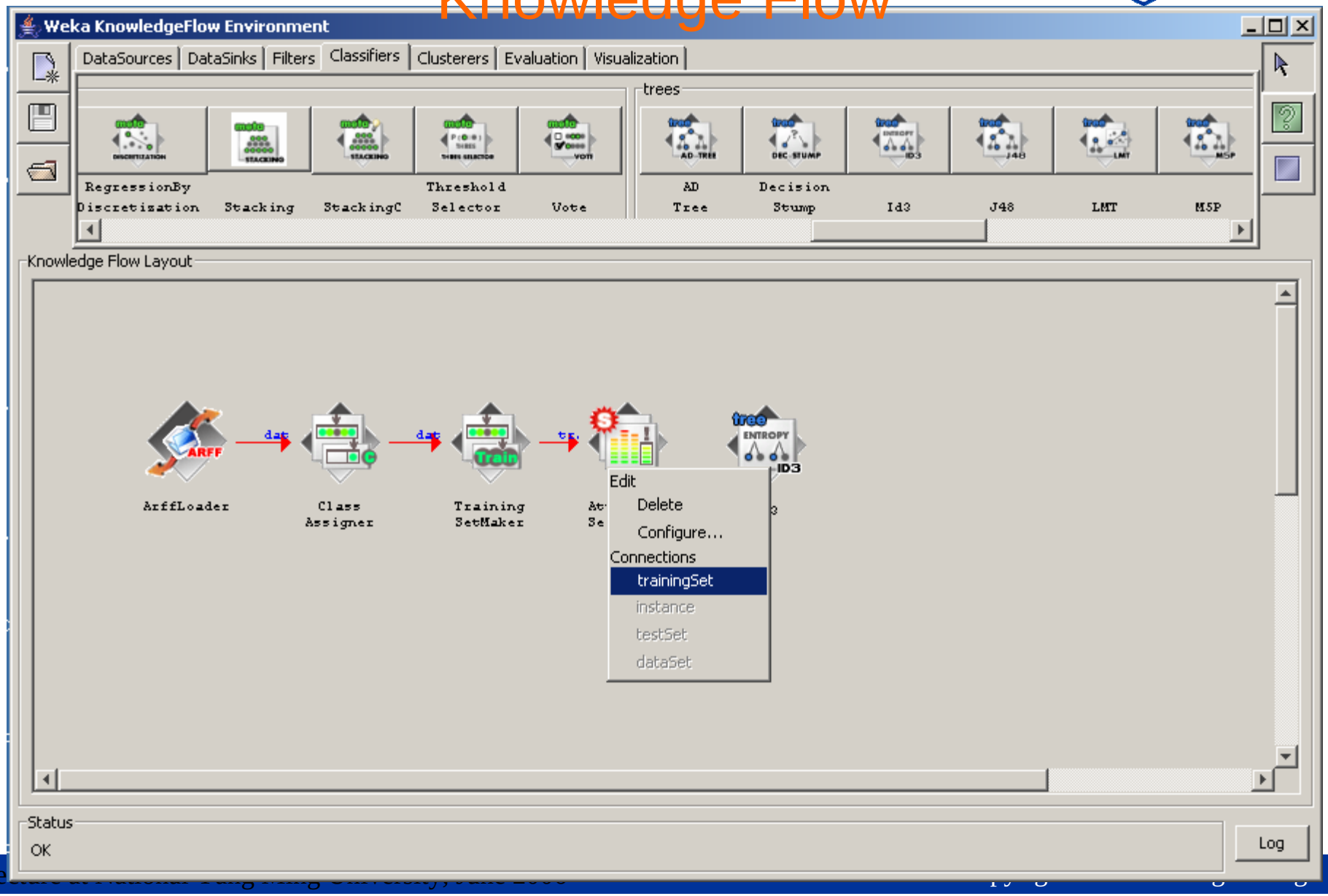
Knowledge Flow



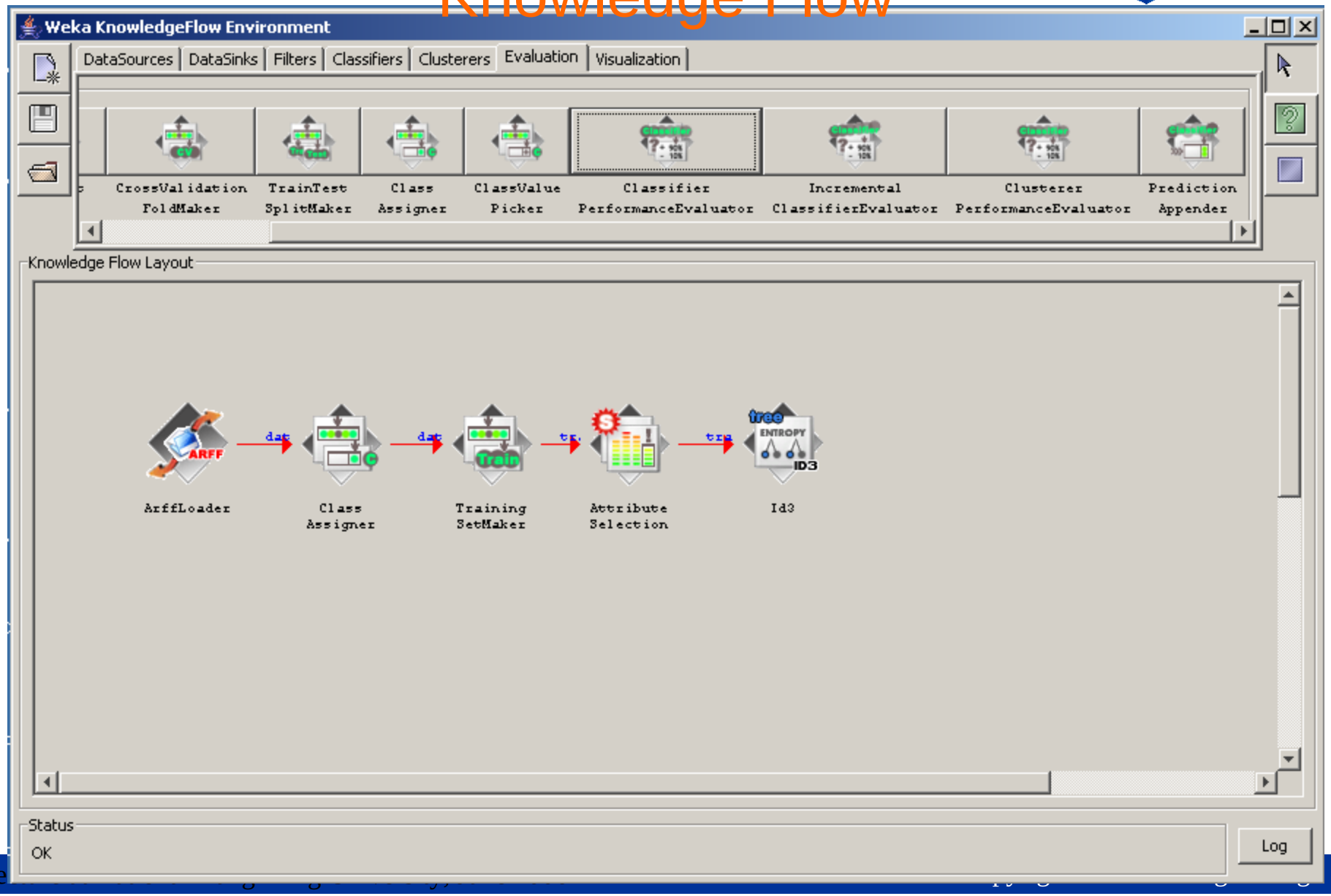
Knowledge Flow



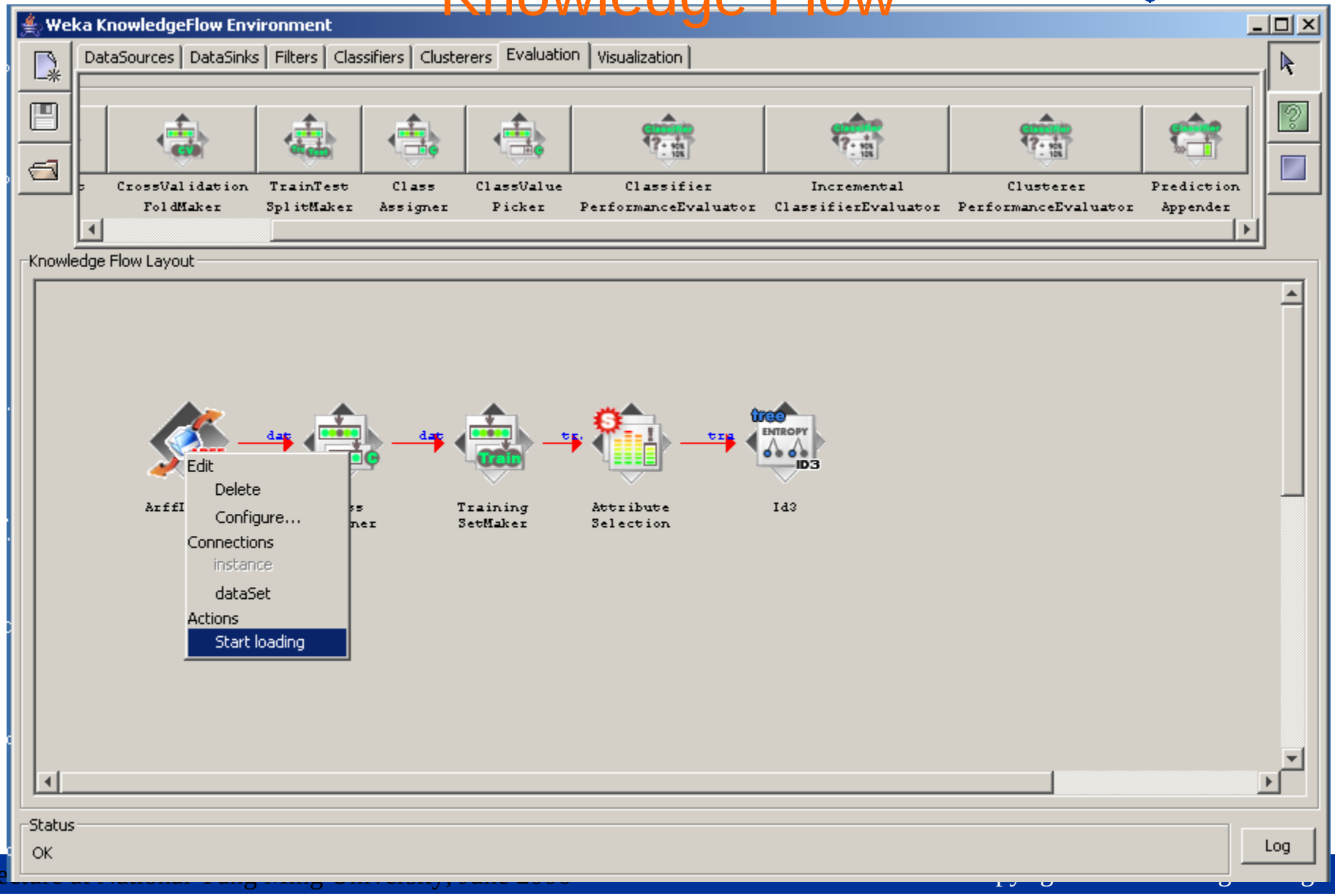
Knowledge Flow



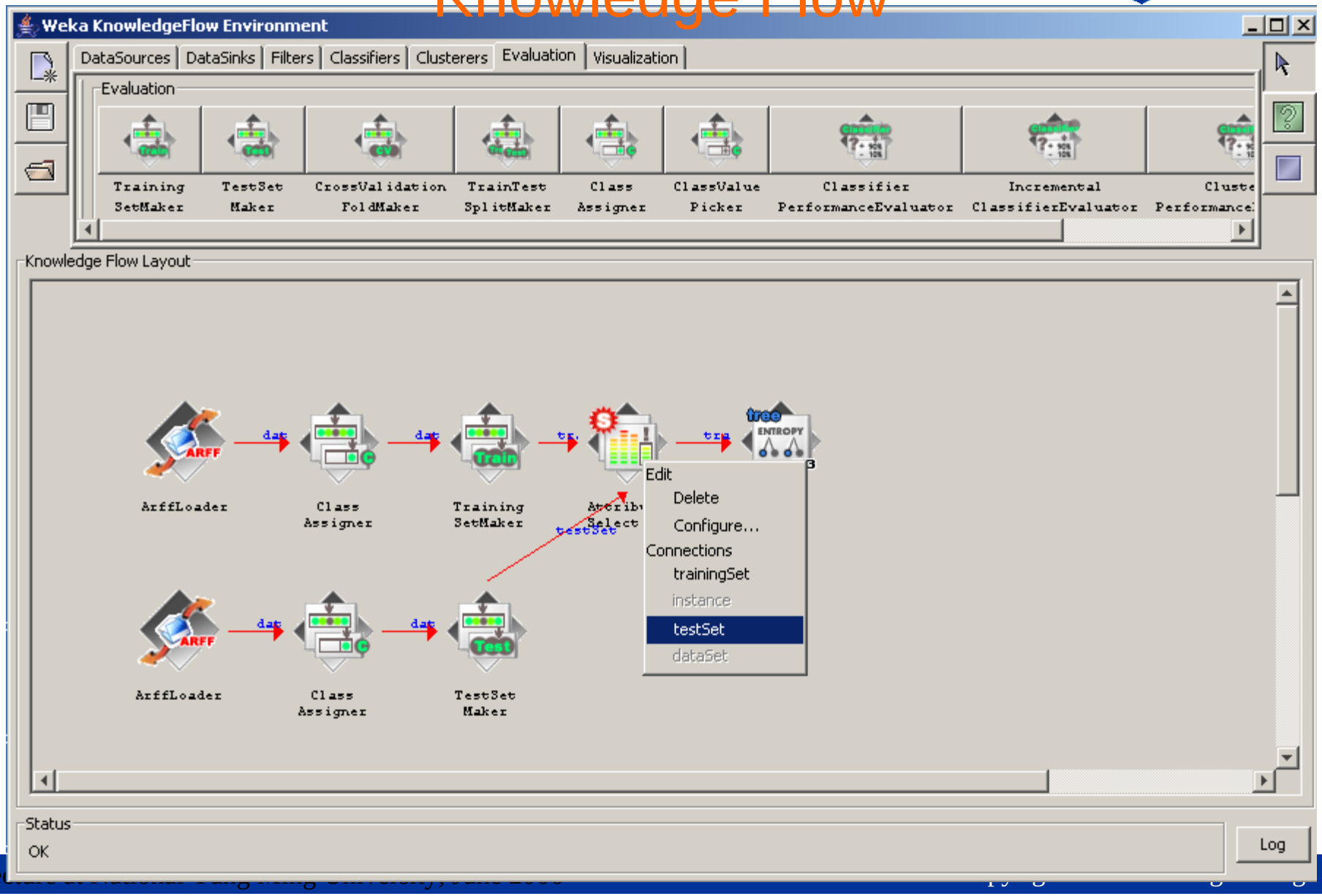
Knowledge Flow



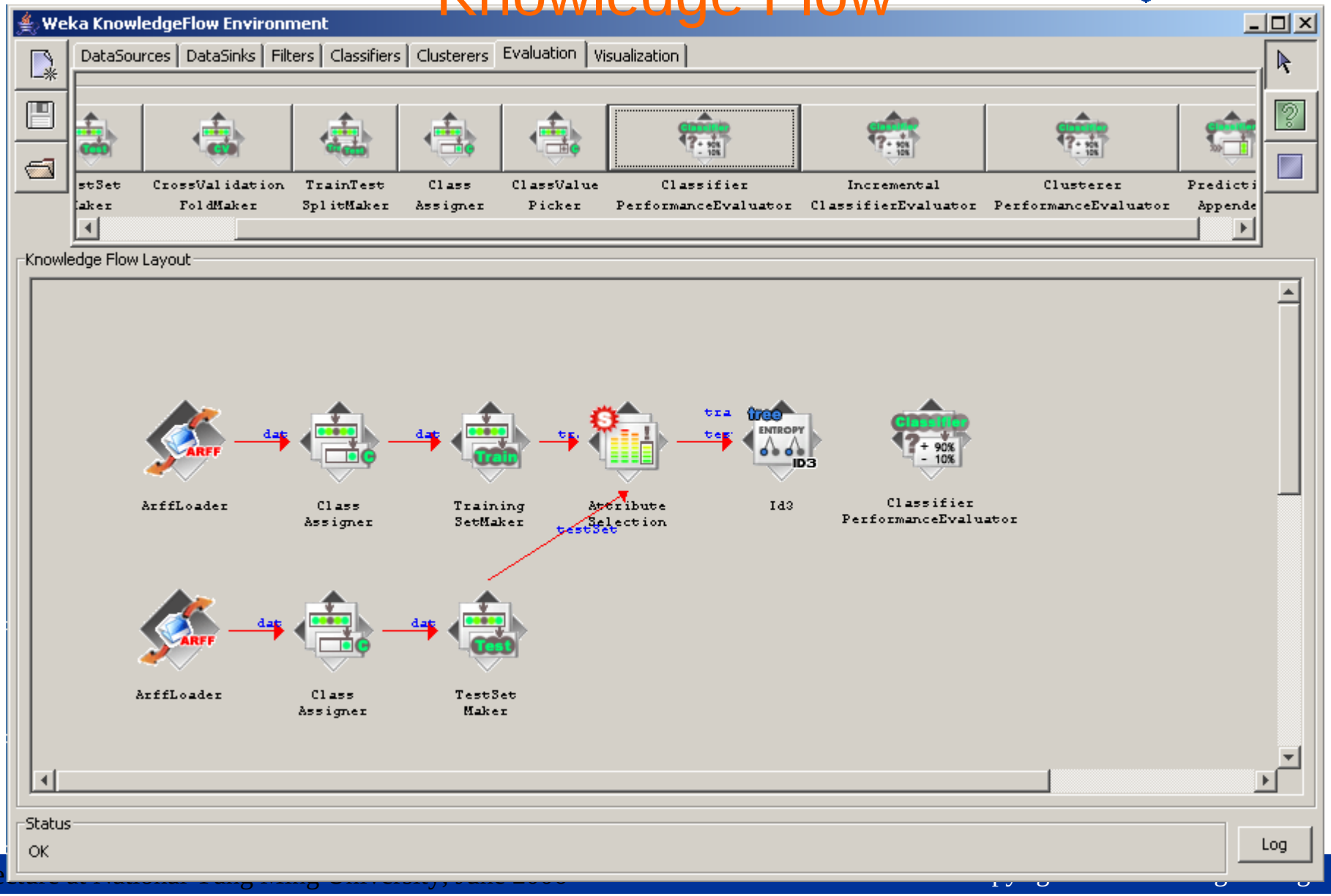
Knowledge Flow



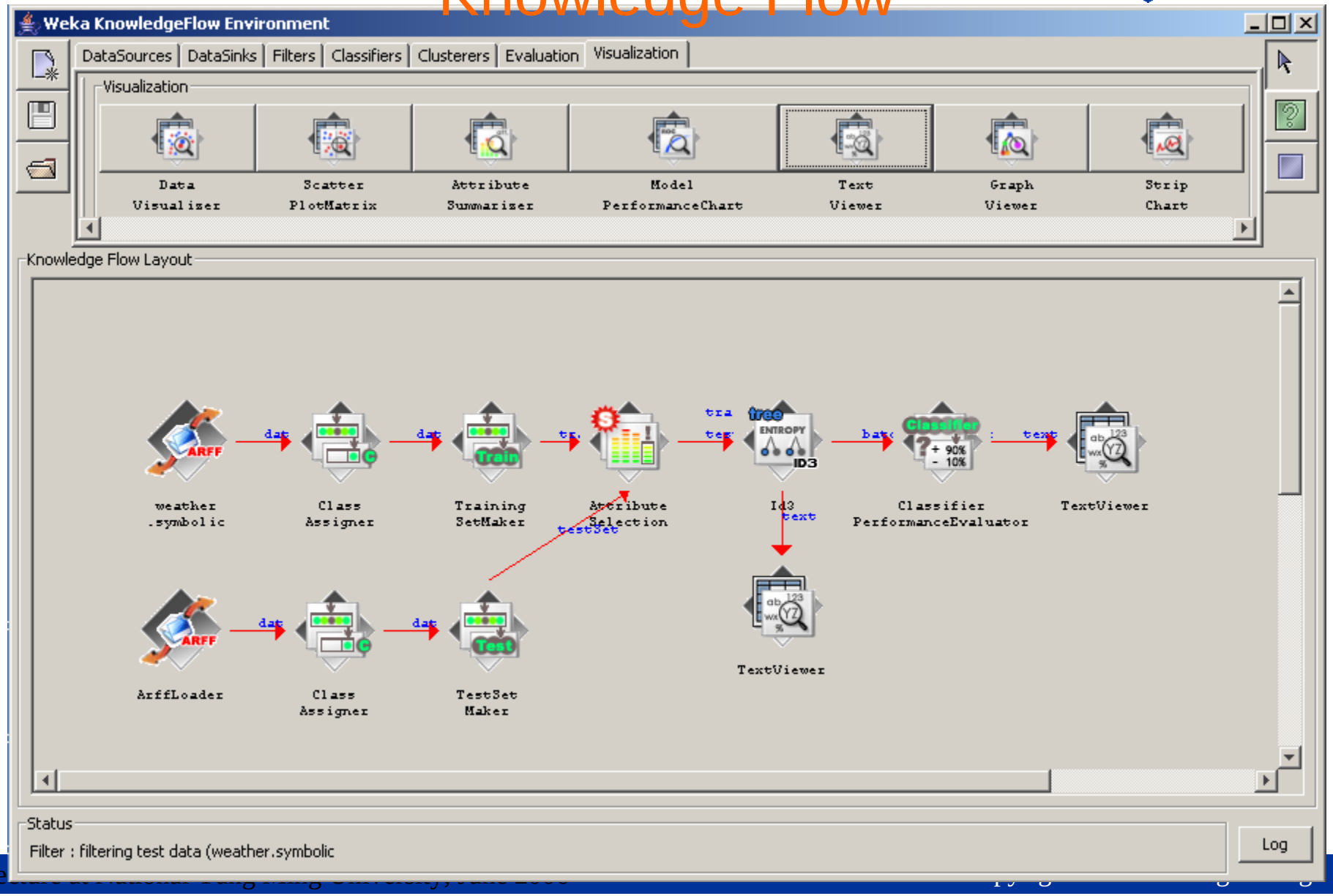
Knowledge Flow



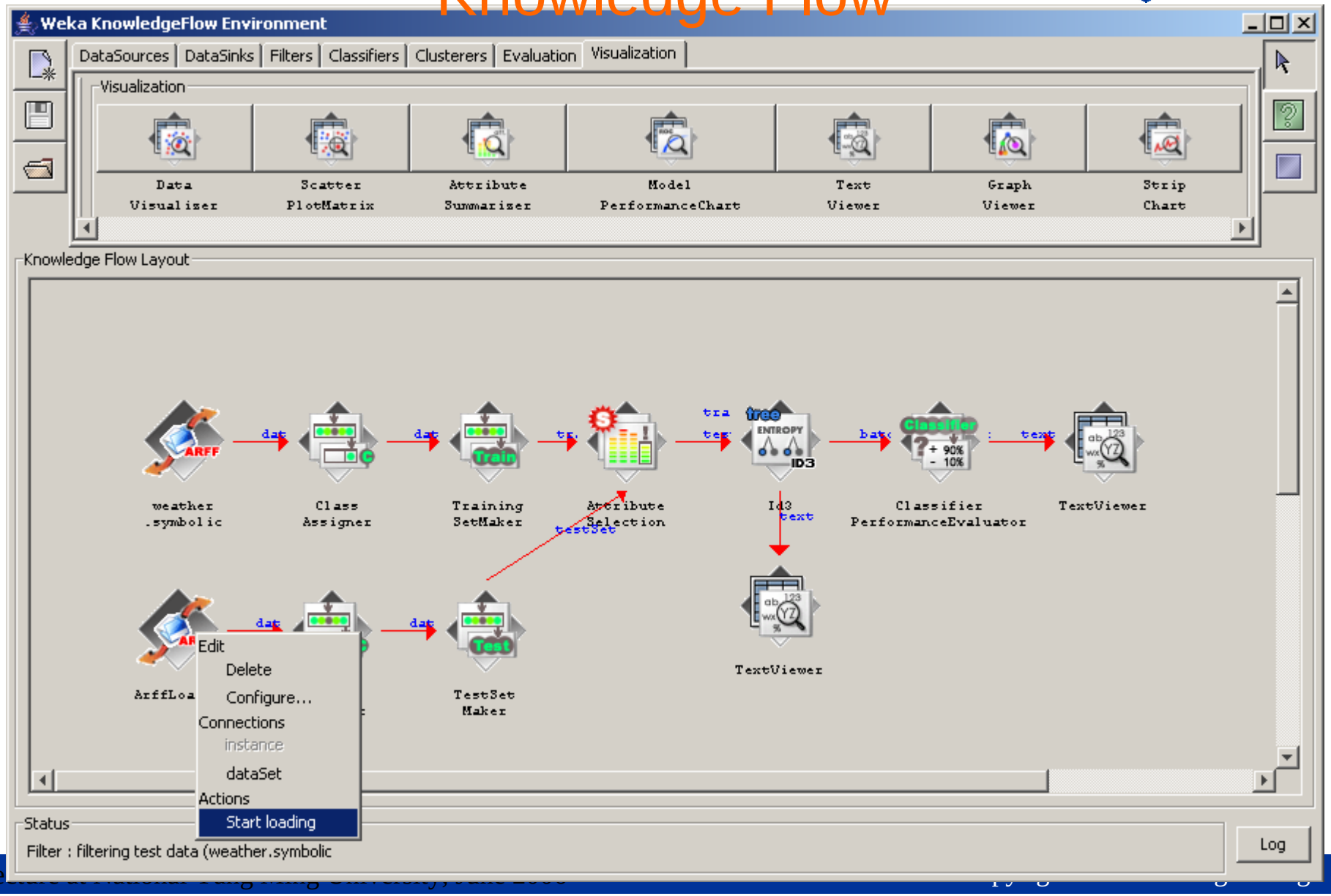
Knowledge Flow



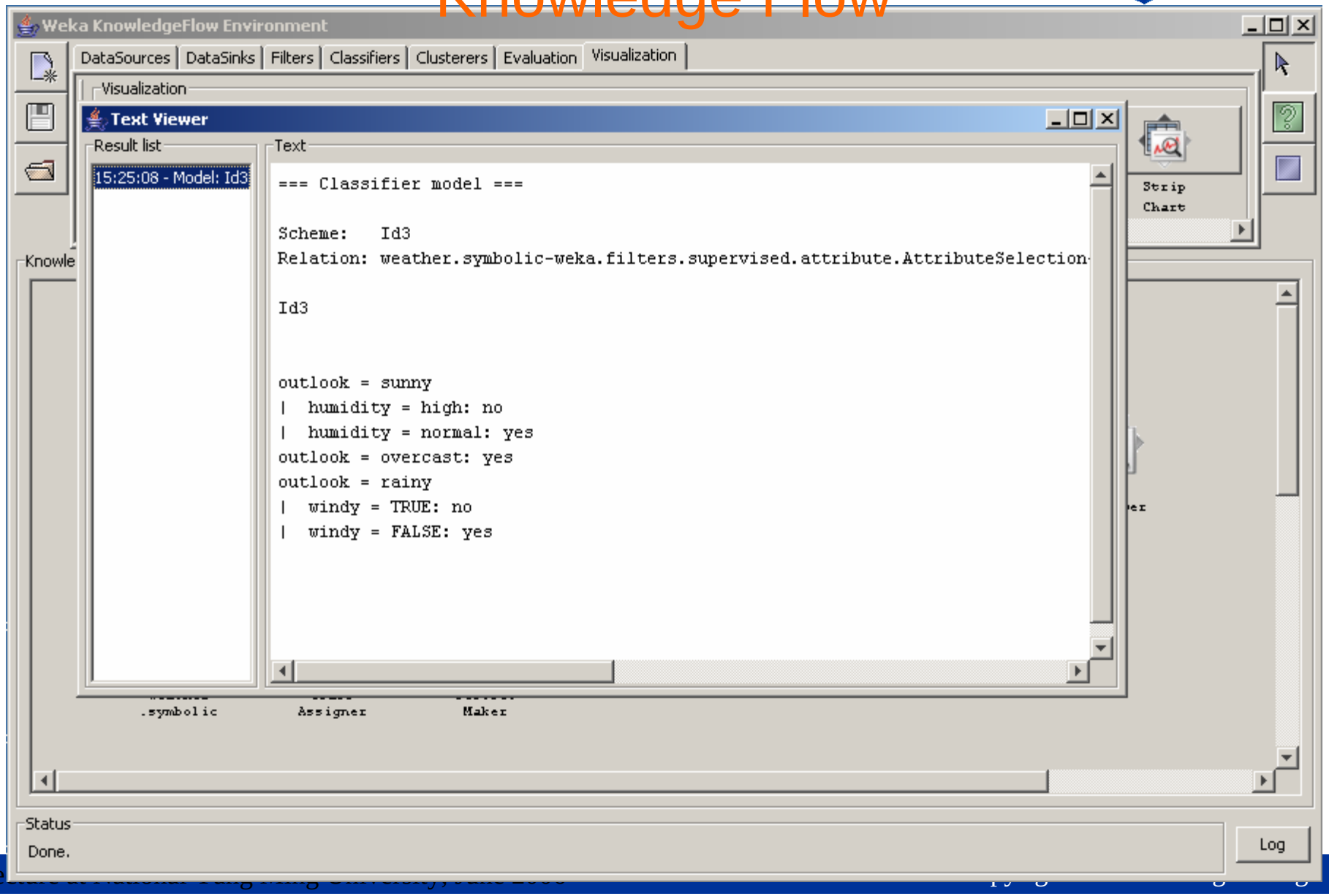
Knowledge Flow



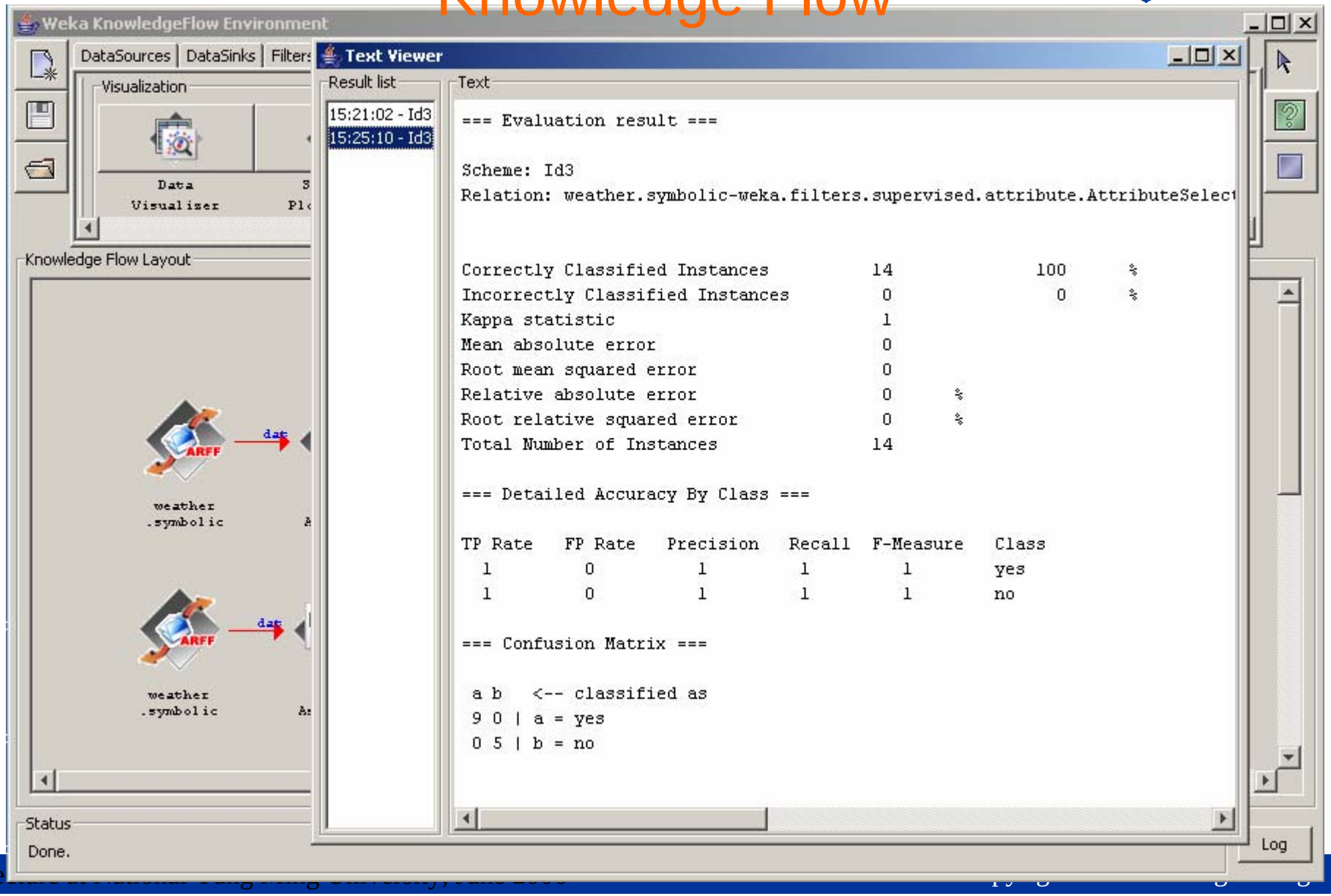
Knowledge Flow



Knowledge Flow



Knowledge Flow



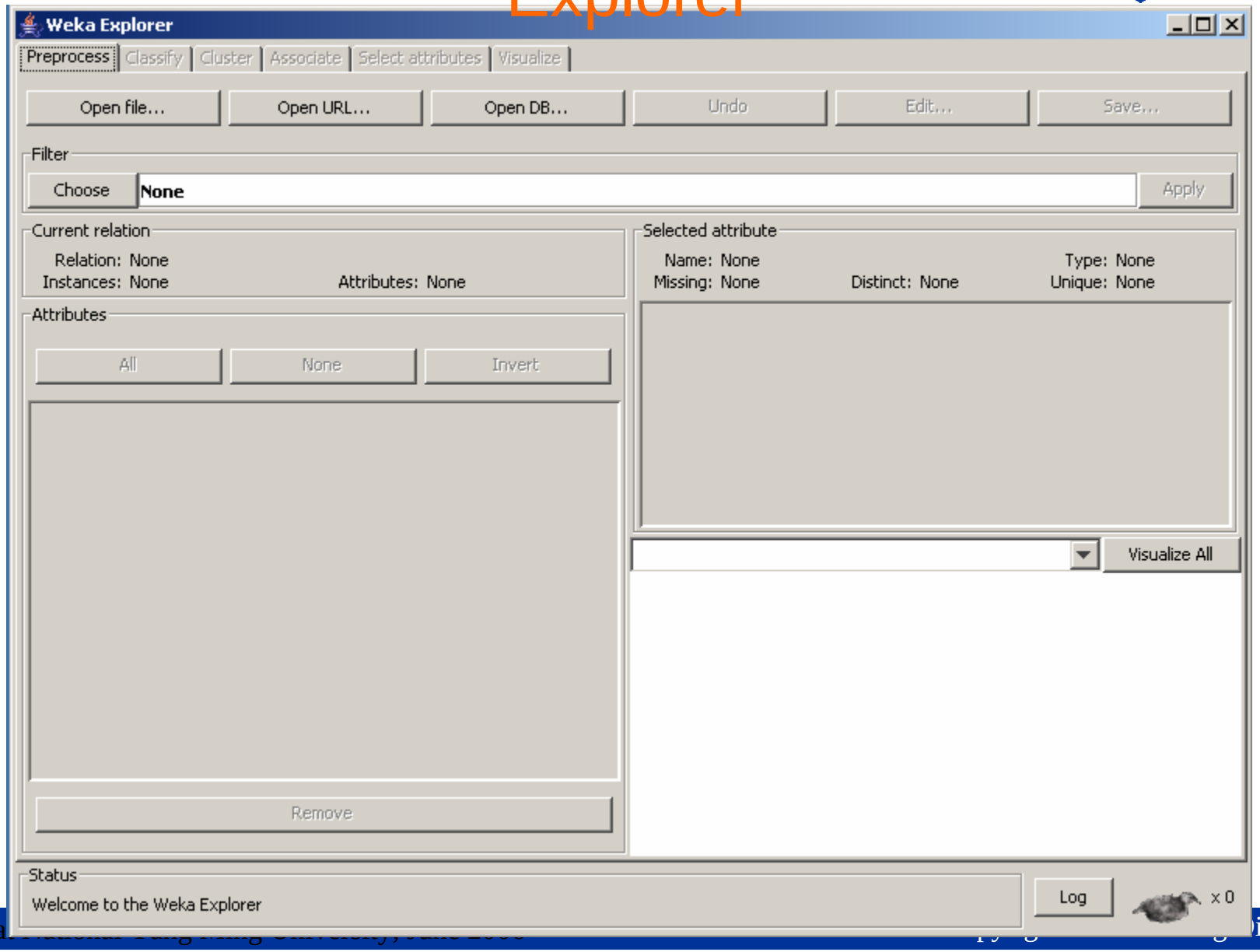
Outline

- What is WEKA
- Knowledge Flow
- Explorer
- Why Knowledge Flow
- Cross Validation
- Reference

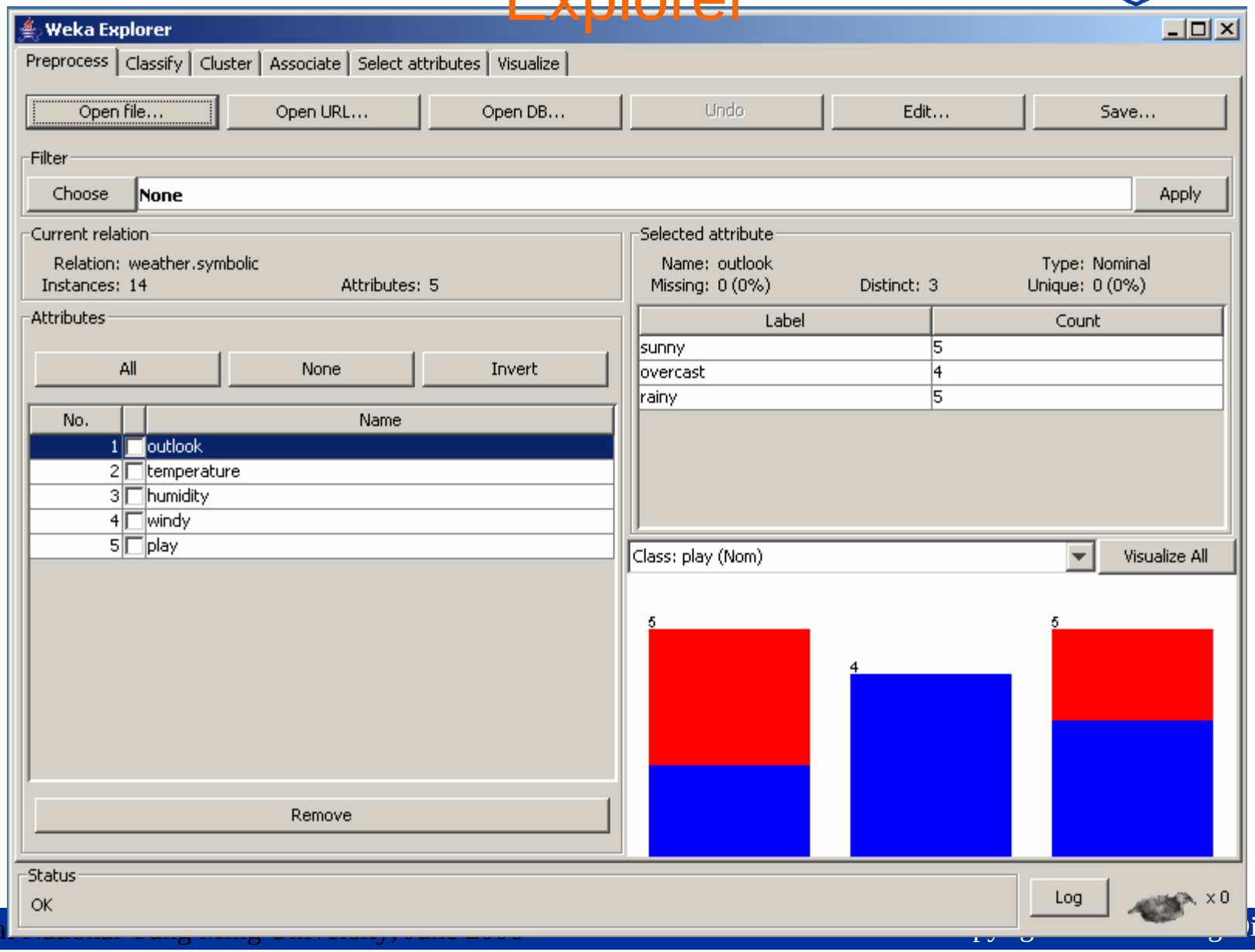
Explorer

- **Do the same experiment…**
- **Experiment 1:**
  - Type: Classification
  - Feature selection: GainRatio; Ranker top 3
  - Algorithm: ID3
  - Training: Weather_nominal.arff
  - Test: Weather_nominal.arff

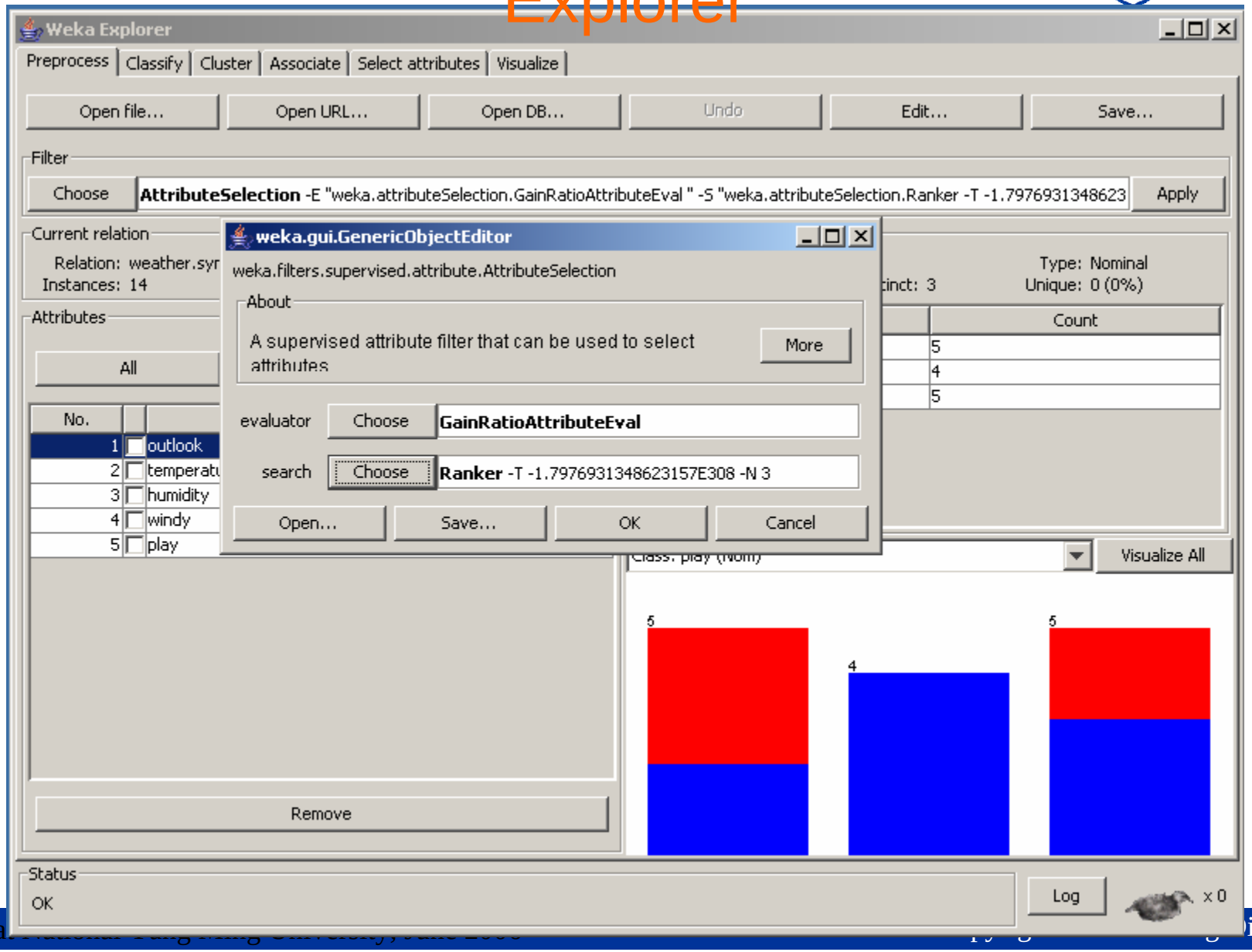
Explorer


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

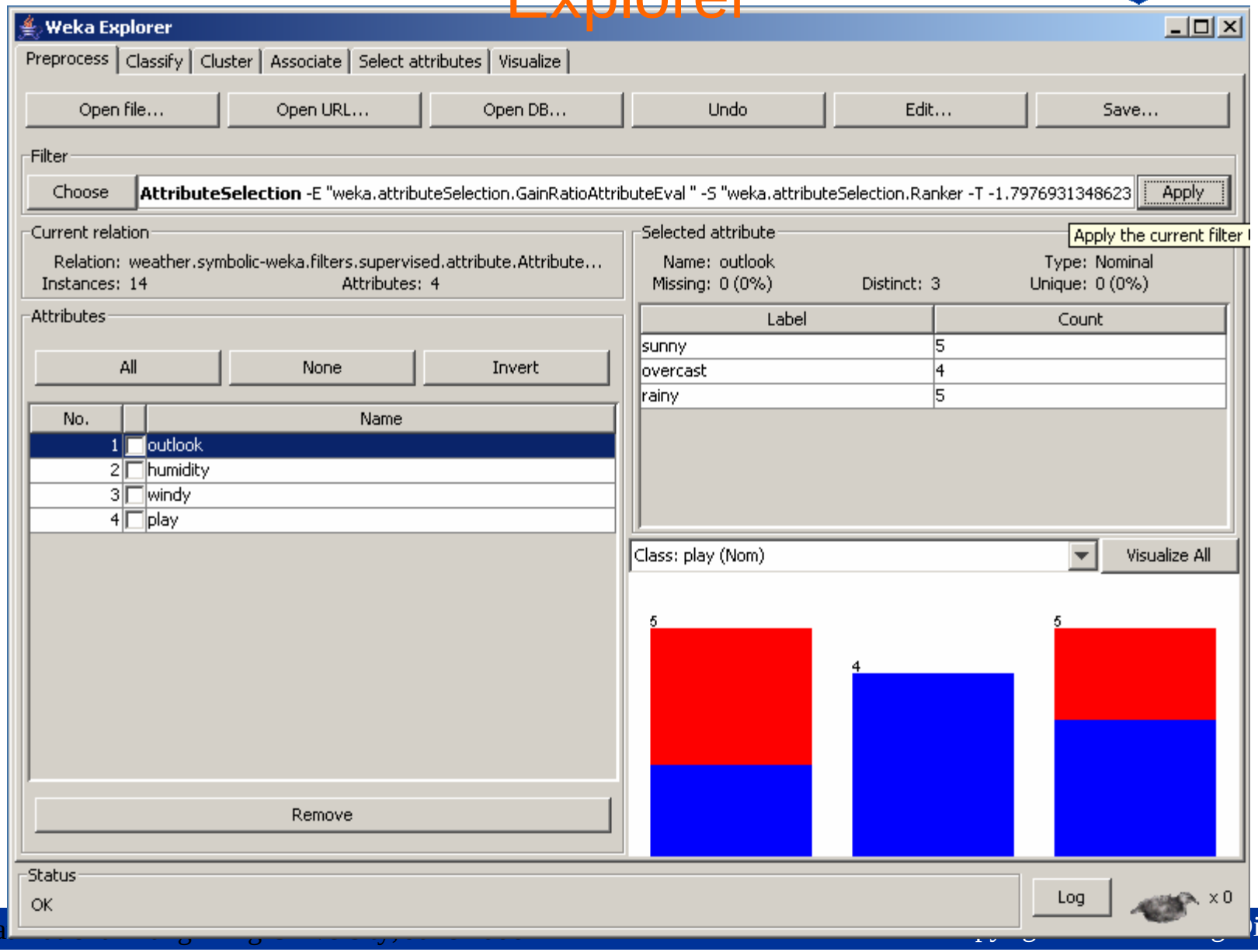
Explorer


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Attributes
All
None
Invert
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Label
Count
sunny 5
overcast 4
rainy 5
Class: play (Nom)
Visualize All
5
4
5
Status
OK
Log
x 0

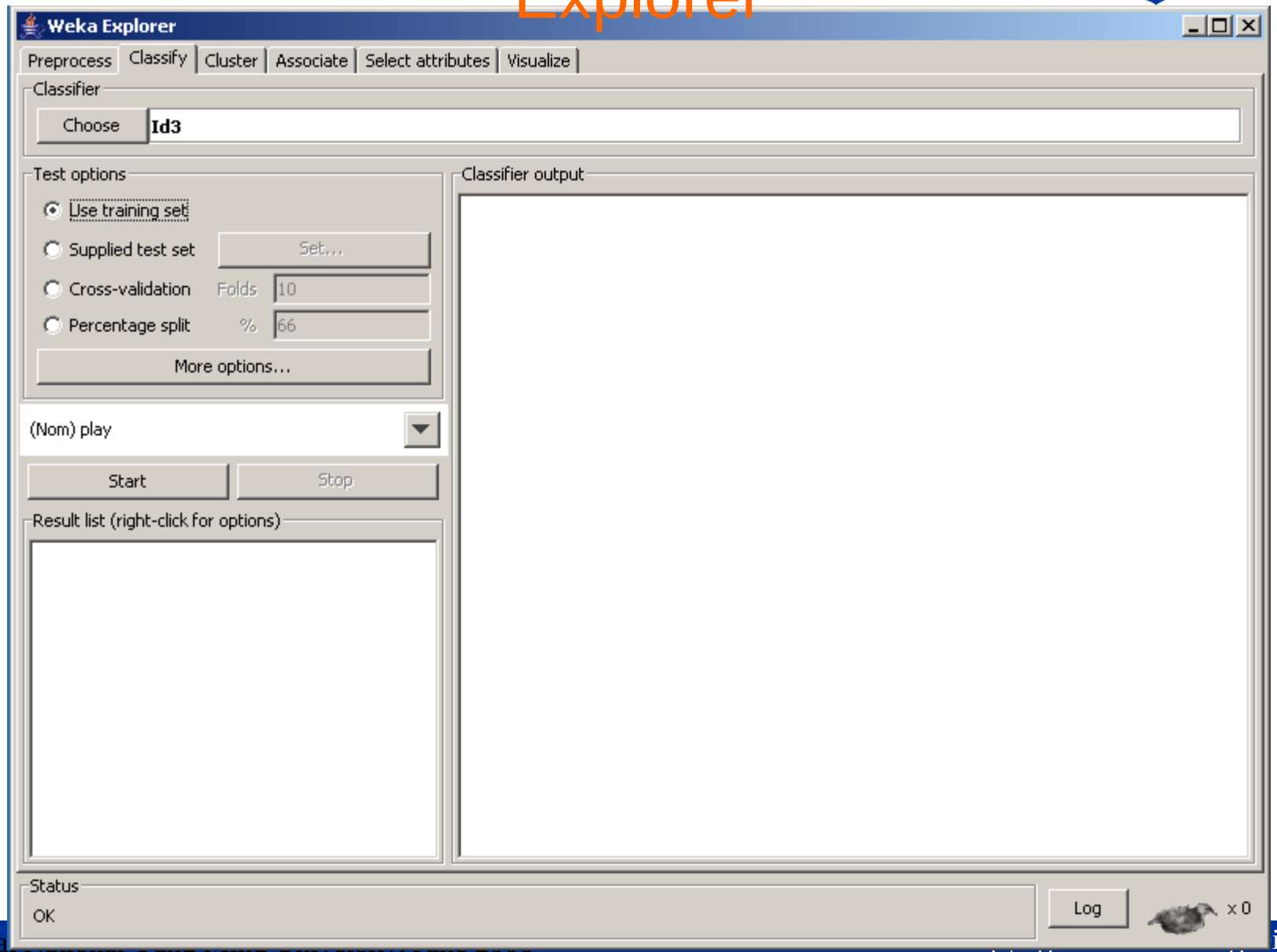
Explorer


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
AttributeSelection -E "weka.attributeSelection.GainRatioAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623
Apply
Current relation
Relation: weather.syr
Instances: 14
Attributes
All
No.
1 outlook
2 temperatu
3 humidity
4 windy
5 play
Remove
weka.gui.GenericObjectEditor
weka.filters.supervised.attribute.AttributeSelection
About
A supervised attribute filter that can be used to select attributes
More
evaluator
Choose
GainRatioAttributeEval
search
Choose
Ranker -T -1.7976931348623157E308 -N 3
Open...
Save...
OK
Cancel
Type: Nominal
Unique: 0 (0%)
Count
5
4
5
Visualize All
Status
OK
Log
x 0

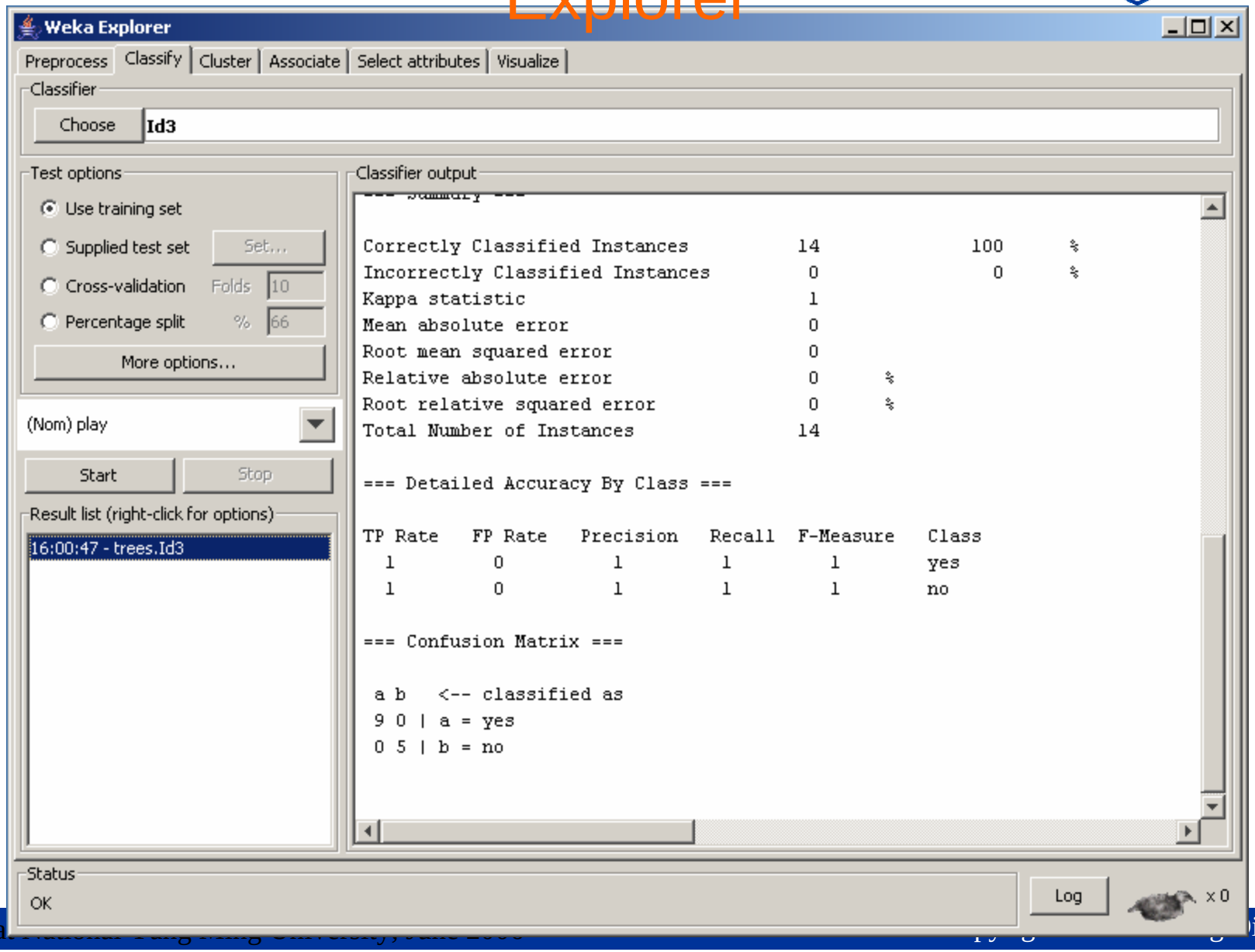
Explorer


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Edit...
Save...
Filter
Choose
AttributeSelection -E "weka.attributeSelection.GainRatioAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623
Apply
Current relation
Relation: weather.symbolic-weka.filters.supervised.attribute.Attribute...
Instances: 14
Attributes: 4
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Label
Count
sunny
5
overcast
4
rainy
5
Attributes
All
None
Invert
No.
Name
1 outlook
2 humidity
3 windy
4 play
Remove
Class: play (Nom)
Visualize All
5
4
5
Status
OK
Log
x 0

Explorer


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Id3
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) play
Start
Stop
Result list (right-click for options)
Status
OK
Log
x 0

Explorer


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Id3
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
16:00:47 - trees.Id3
Classifier output
Correctly Classified Instances 14 100 %
Incorrectly Classified Instances 0 0 %
Kappa statistic 1
Mean absolute error 0
Root mean squared error 0
Relative absolute error 0 %
Root relative squared error 0 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure Class
1 0 1 1 1 yes
1 0 1 1 1 no
=== Confusion Matrix ===
a b <-- classified as
9 0 | a = yes
0 5 | b = no
Status
OK
Log
x 0

Outline

- What is WEKA
- Knowledge Flow
- Explorer
- Why Knowledge Flow
- Cross Validation
- Reference

Why Knowledge Flow

- **There are some jobs we cannot do in explorer …**
  - Combine feature selection
  - Build more complicated systems
- **KF describes the process more clearly**
  - Never regard the training and test data to be separate in the previous example in explorer
- **KF help us to access some mid-process info of the machine learning method**
  - Cross Validation

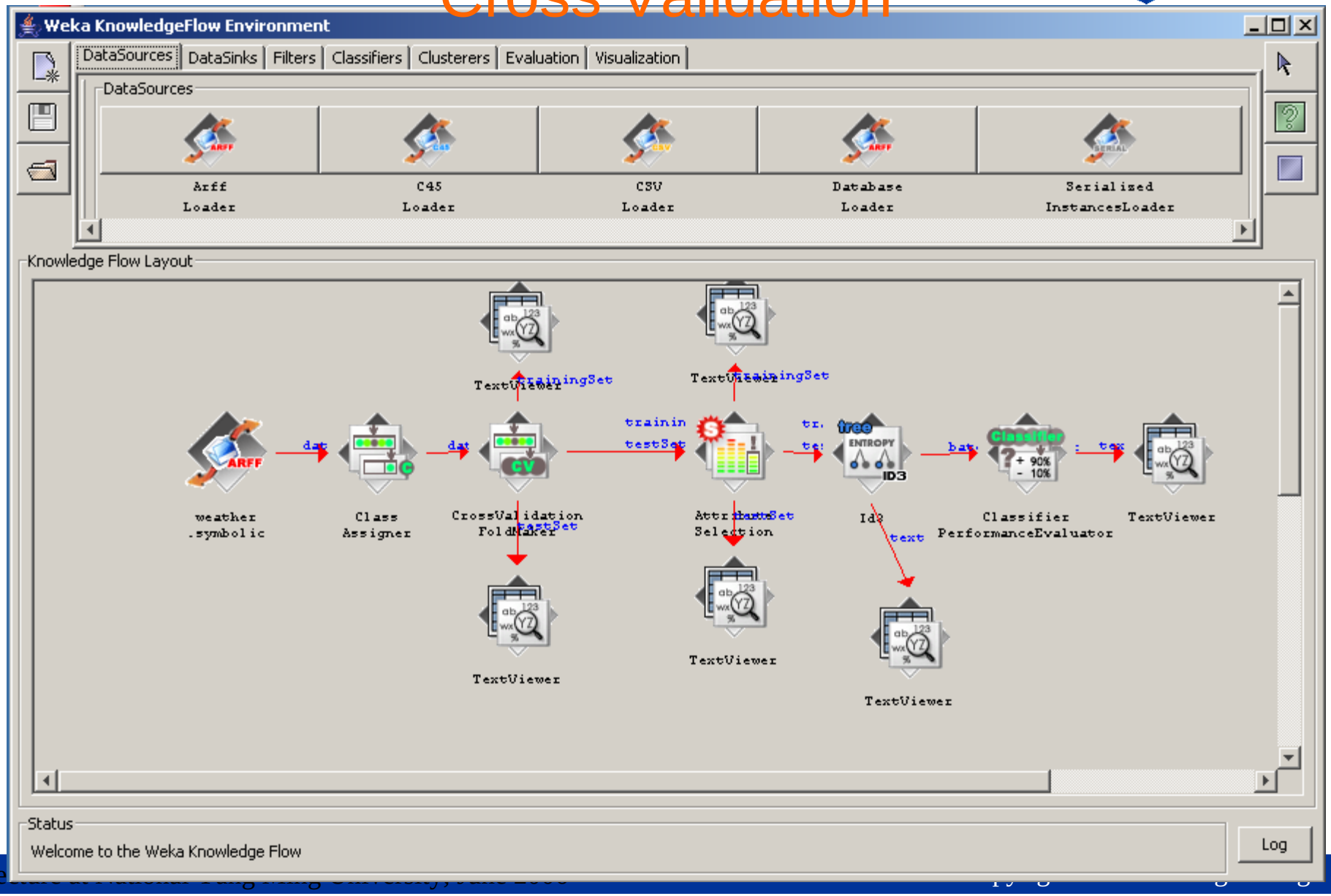
Outline

- What is WEKA
- Knowledge Flow
- Explorer
- Why Knowledge Flow
- Cross Validation
- Reference

Cross Validation

- **Experiment 2:**
  - Type: Classification
  - Feature selection: GainRatio; Ranker top 3
  - Algorithm: ID3
  - Training: Weather_nominal.arff (CV)
  - Test: Weather_nominal.arff (CV)
  - CV type: 3-folder CV

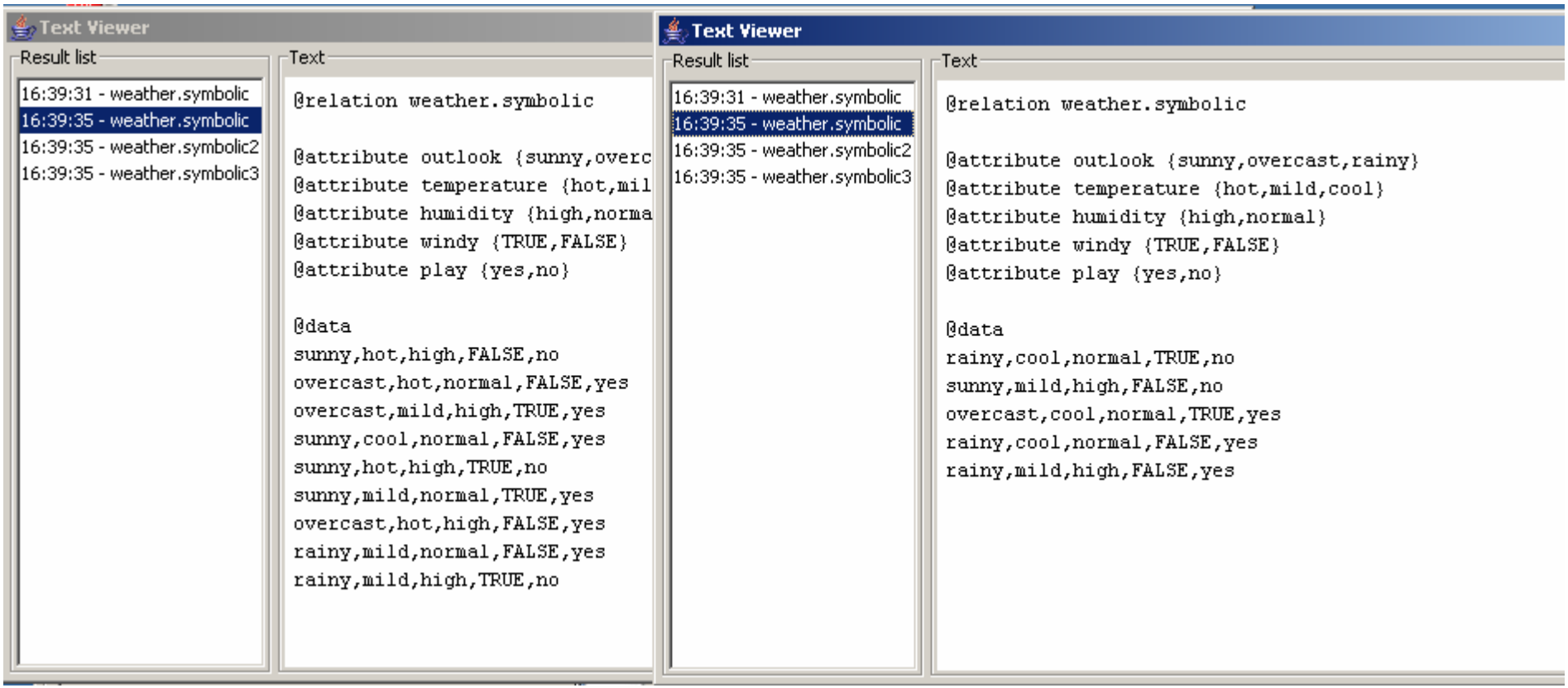
Cross Validation



Cross Validation

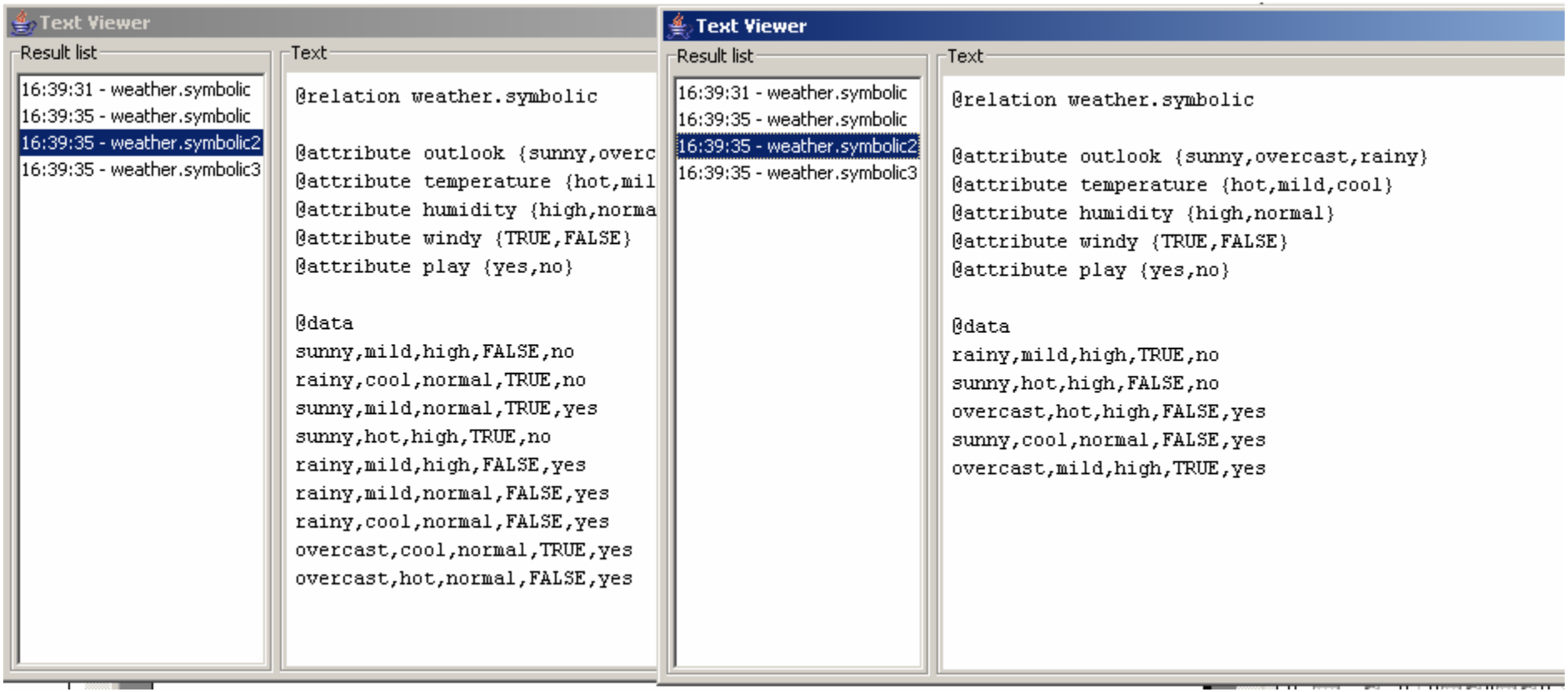
What do we view in this case?

Text1 VS. Text2 (1)



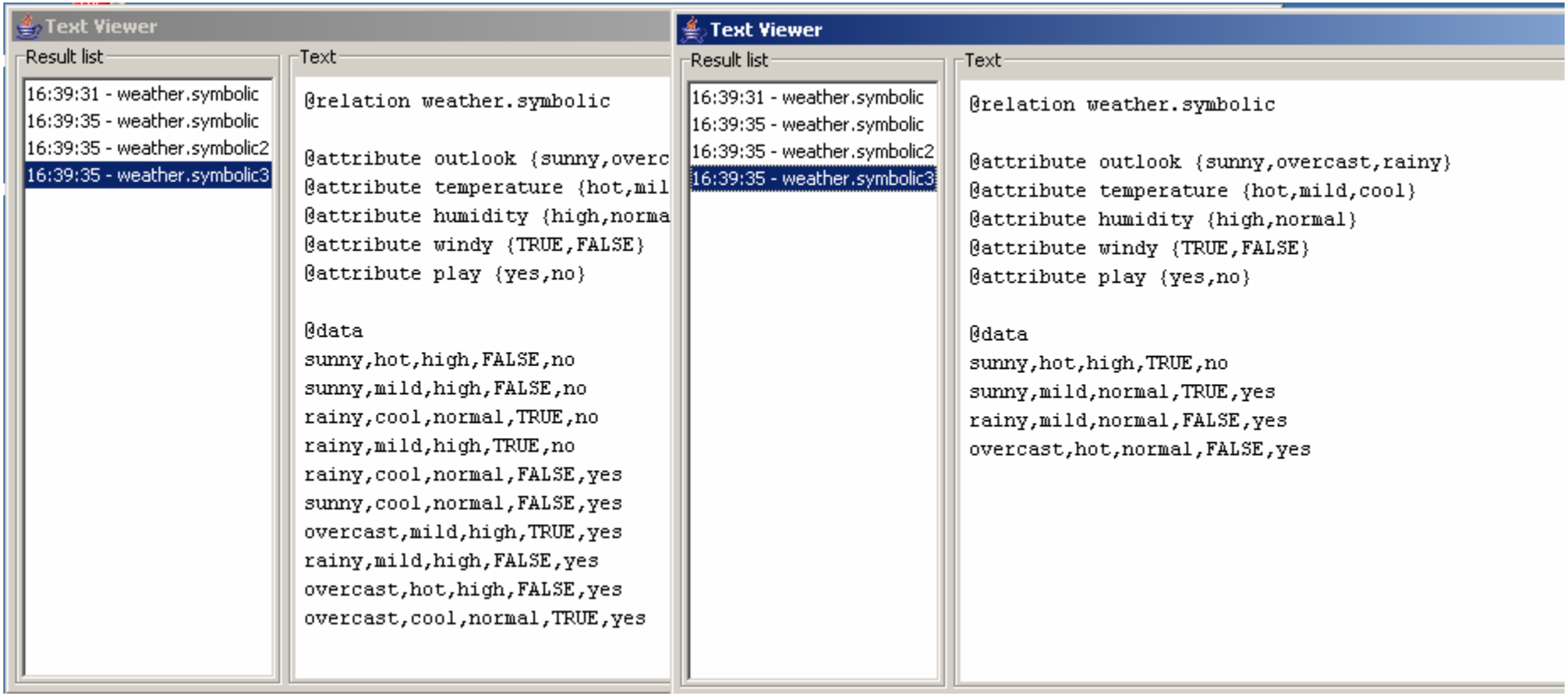
Cross Validation

Text1 VS. Text2 (2)



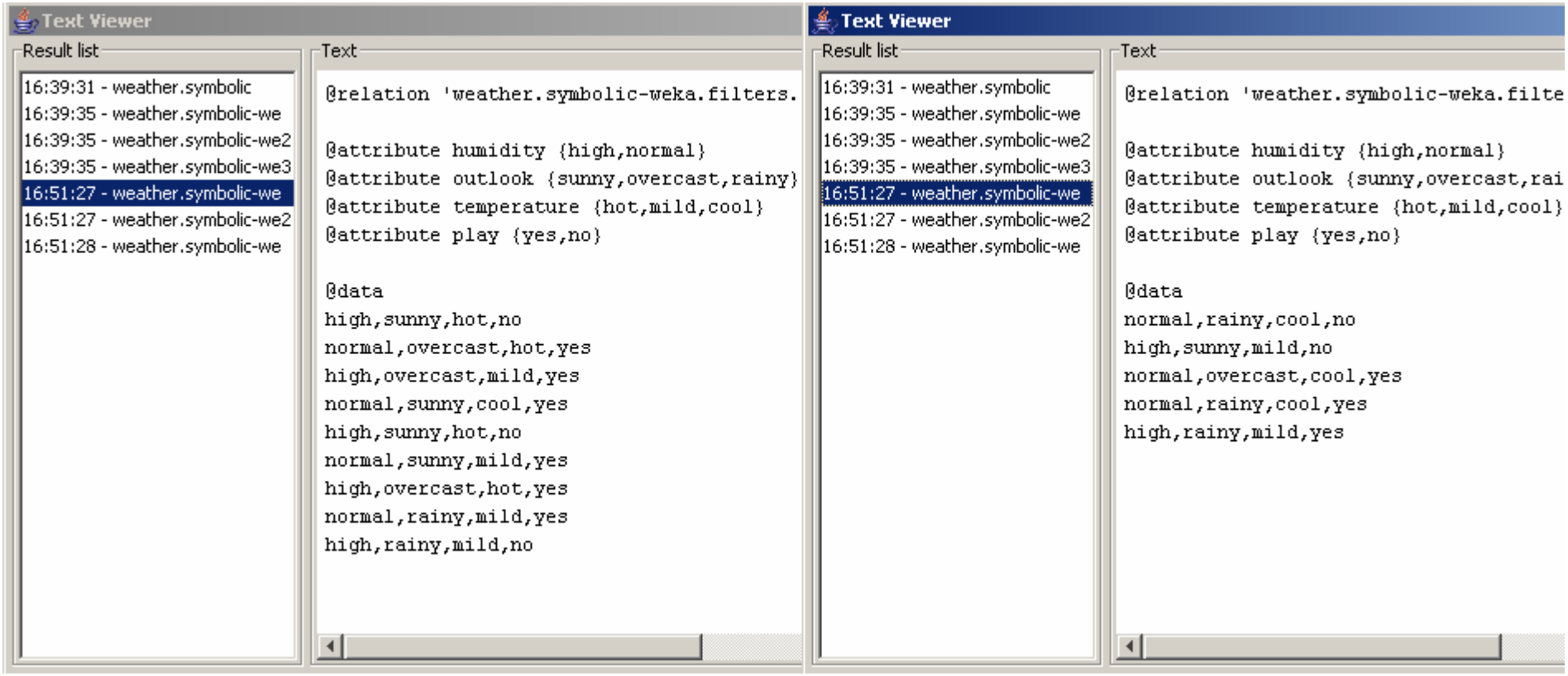
Cross Validation

Text1 VS. Text2 (3)



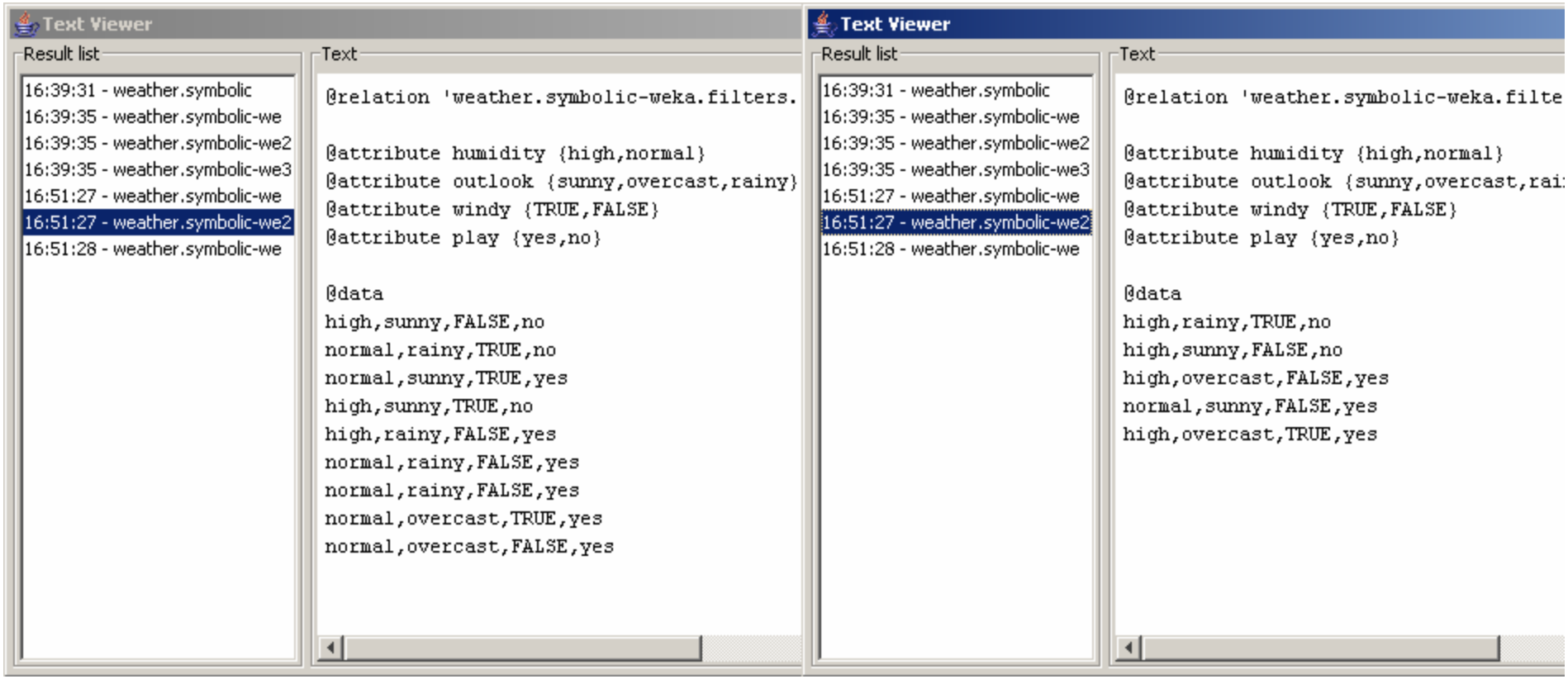
Cross Validation

Text3 VS. Text4 (1)



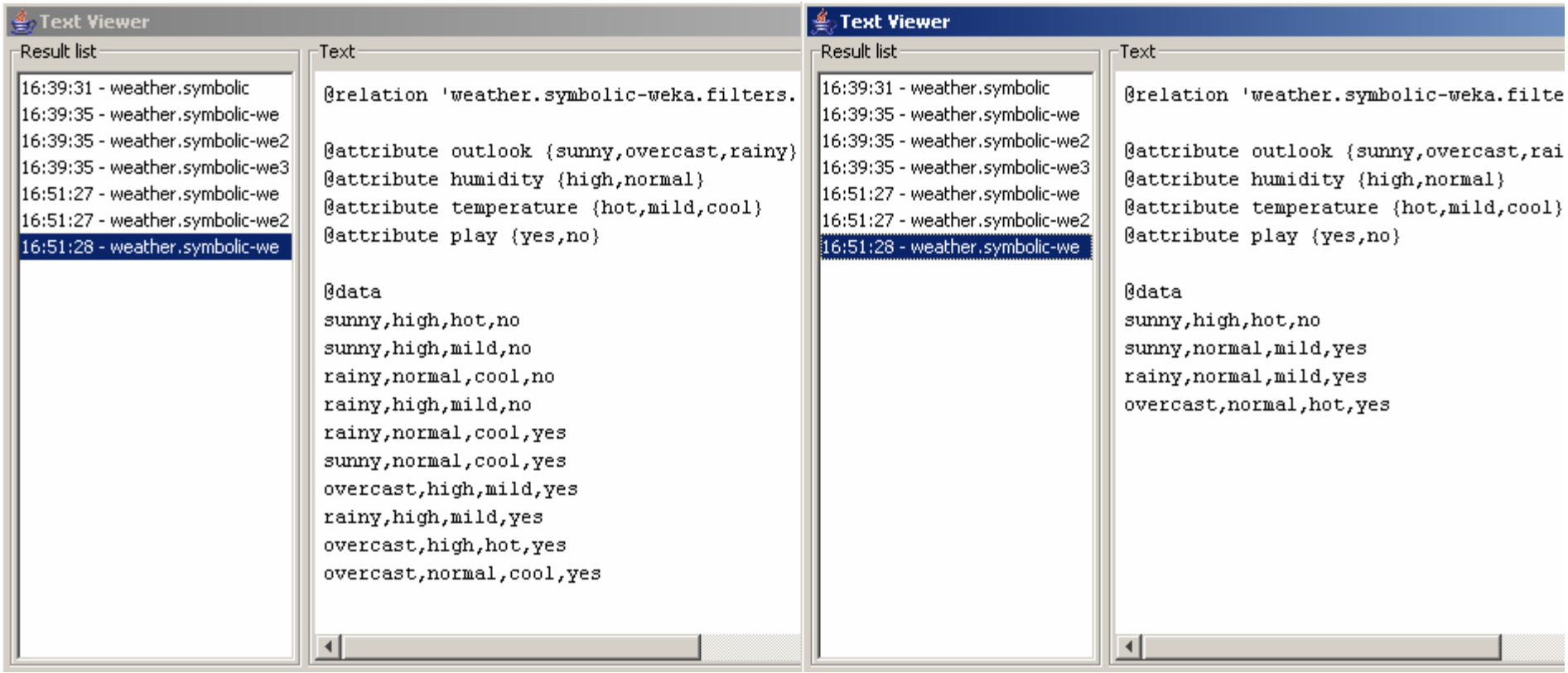
Cross Validation

Text3 VS. Text4 (2)



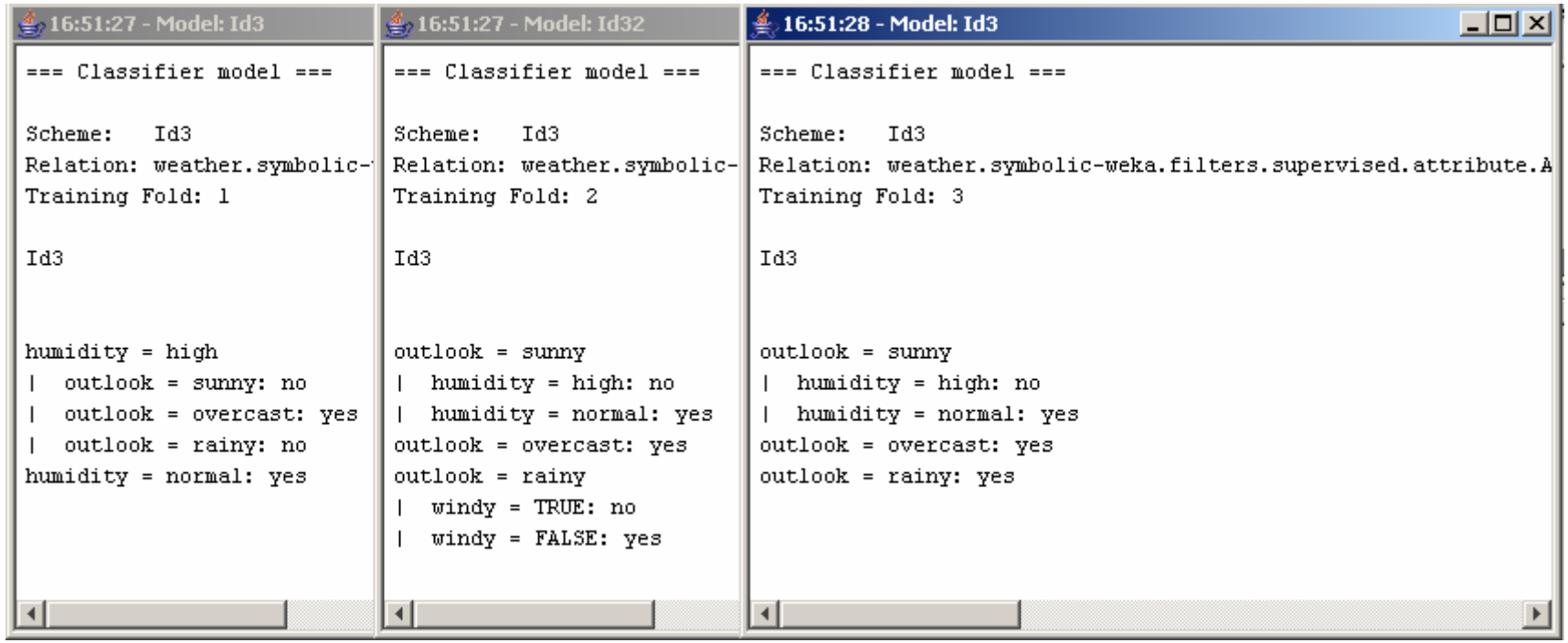
Cross Validation

Text3 VS. Text4 (3)



Cross Validation

Trees



Cross Validation

Evaluation of result

Text Viewer

Result list

16:39:35 - Id3
16:51:28 - Id3

Text

```
Incorrectly Classified Instances         2               14.2857 %
Kappa statistic                          0.6889
Mean absolute error                      0.1786
Root mean squared error                  0.4009
Relative absolute error                 37.234  %
Root relative squared error             82.7516 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.889     0.2         0.889     0.889     0.889     yes
  0.8       0.111       0.8       0.8       0.8       no

=== Confusion Matrix ===

 a b   <-- classified as
 8 1 | a = yes
 1 4 | b = no
```

Cross Validation

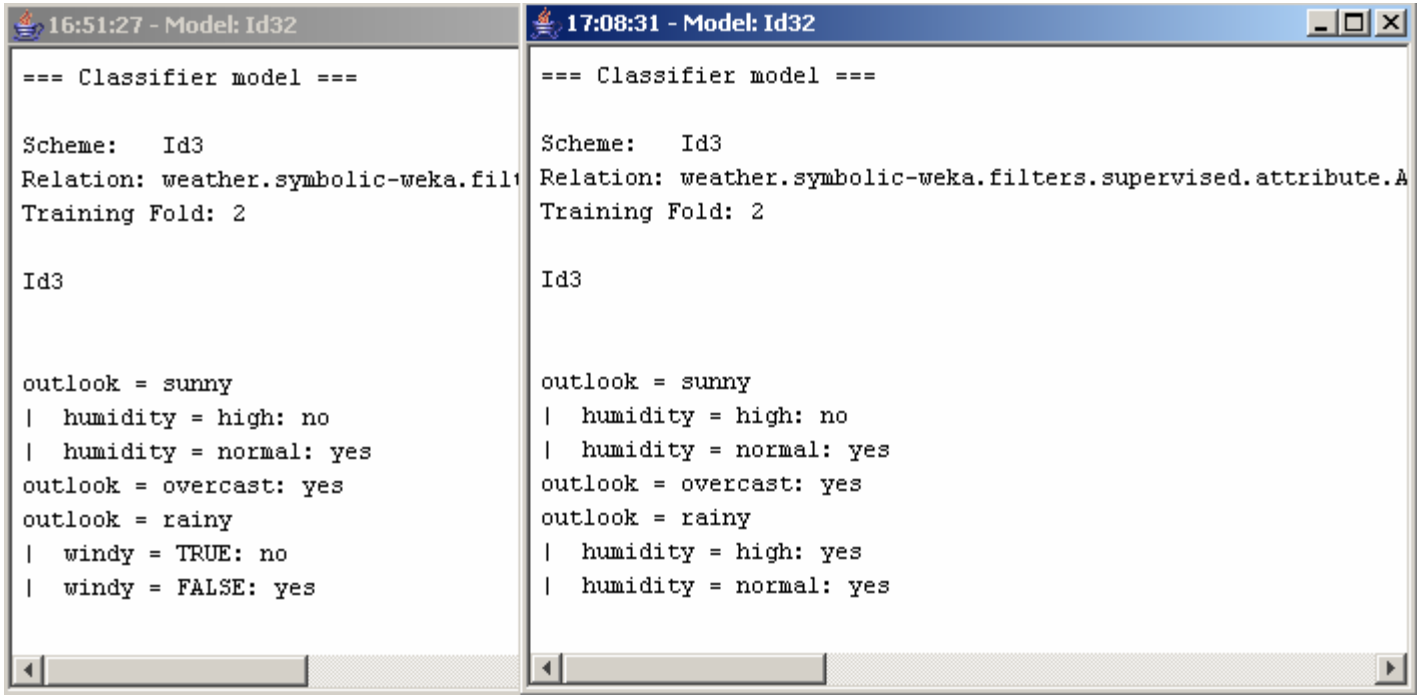
- **Conclusion:**
  - Source data are separated into several folders for cross validation
  - Feature selection is done for each training folder (only training) folder separately
  - Different trees are build in different cases
  - The evaluation of classification is by overall results

Cross Validation

- **Experiment 3:**
  - Type: Classification
  - Feature selection: GainRatio; Ranker top 2
  - Algorithm: ID3
  - Training: Weather_nominal.arff
  - Test: Weather_nominal.arff

Cross Validation

Ranker top 3 VS. Ranker top 2



Cross Validation

- **Conclusion:**
  - Attribute “windy” was ignored. In this case, the classifier only consider the attribute that was kept

Reference

- http://www.cs.waikato.ac.nz/~ml/

- **Ian H. Witten, Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition)**