

# An Introduction to WEKA

**Lecture by Limsoon Wong**  
**Slides prepared by Dong Difeng**





## Outline

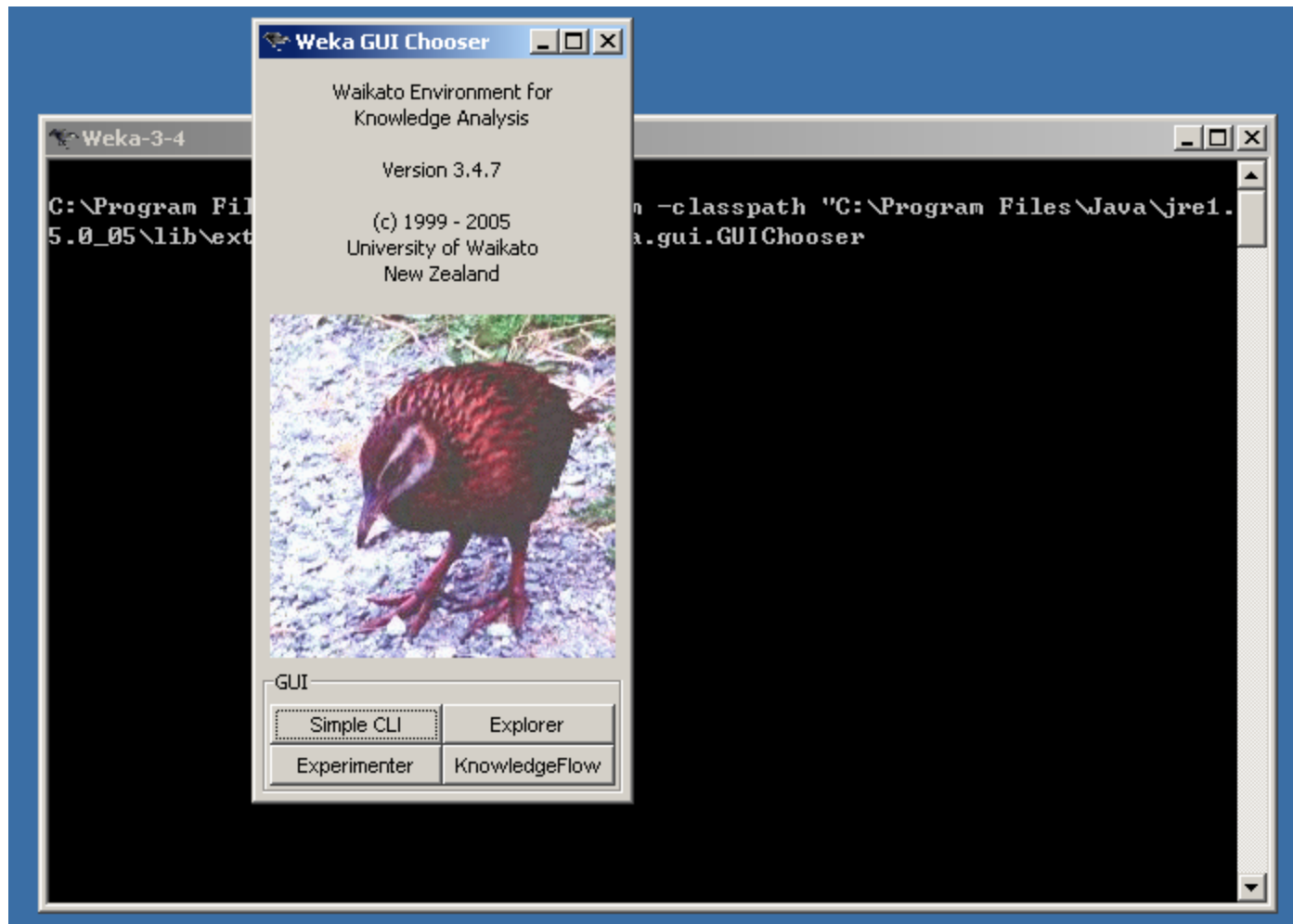
- **What is WEKA**
- Knowledge Flow
- Explorer
- Why Knowledge Flow
- Cross Validation
- Reference



## 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

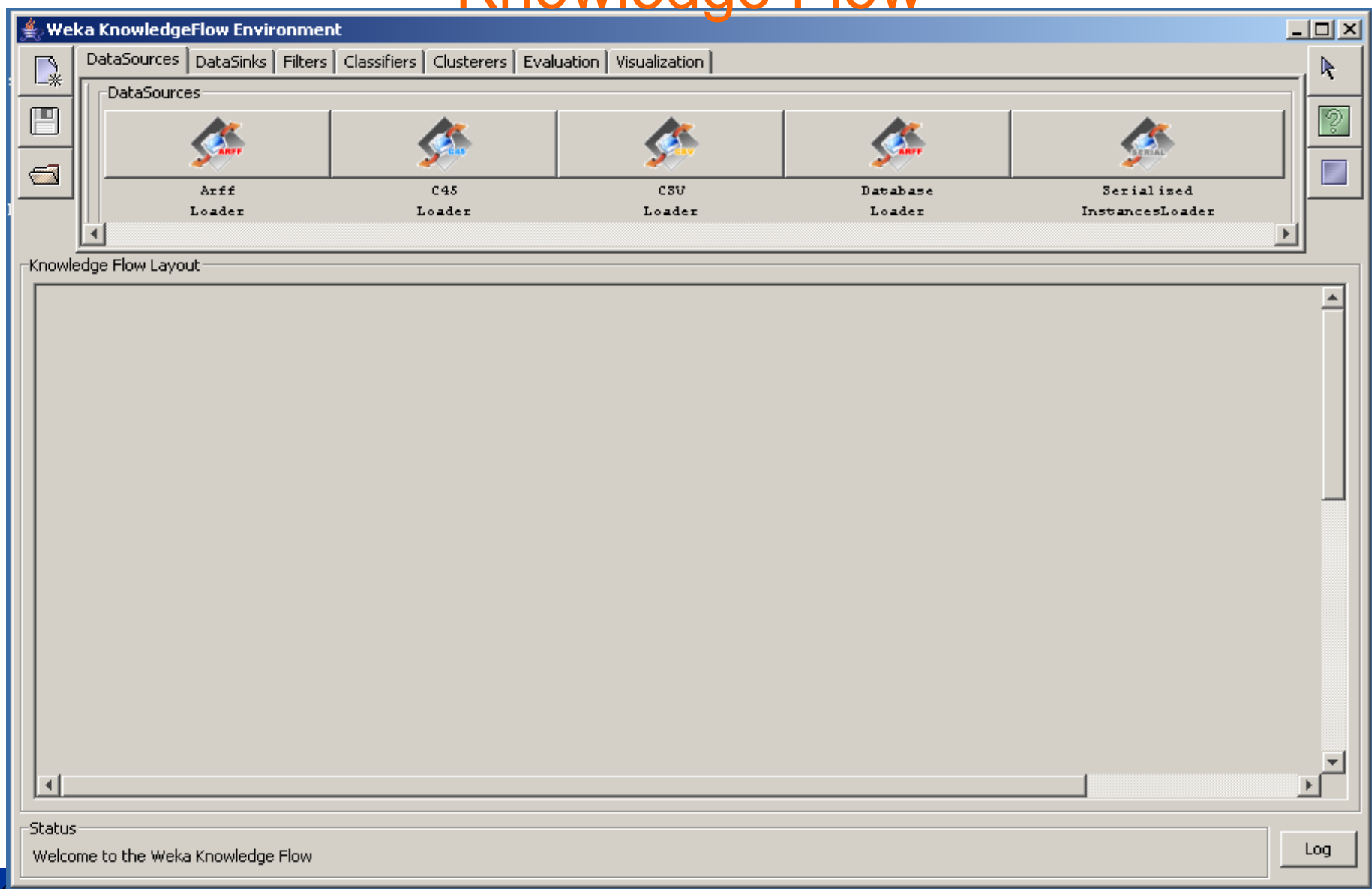
- What is WEKA
- **Knowledge Flow**
- Explorer
- Why Knowledge Flow
- Cross Validation
- Reference



# Knowledge Flow

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

# Knowledge Flow



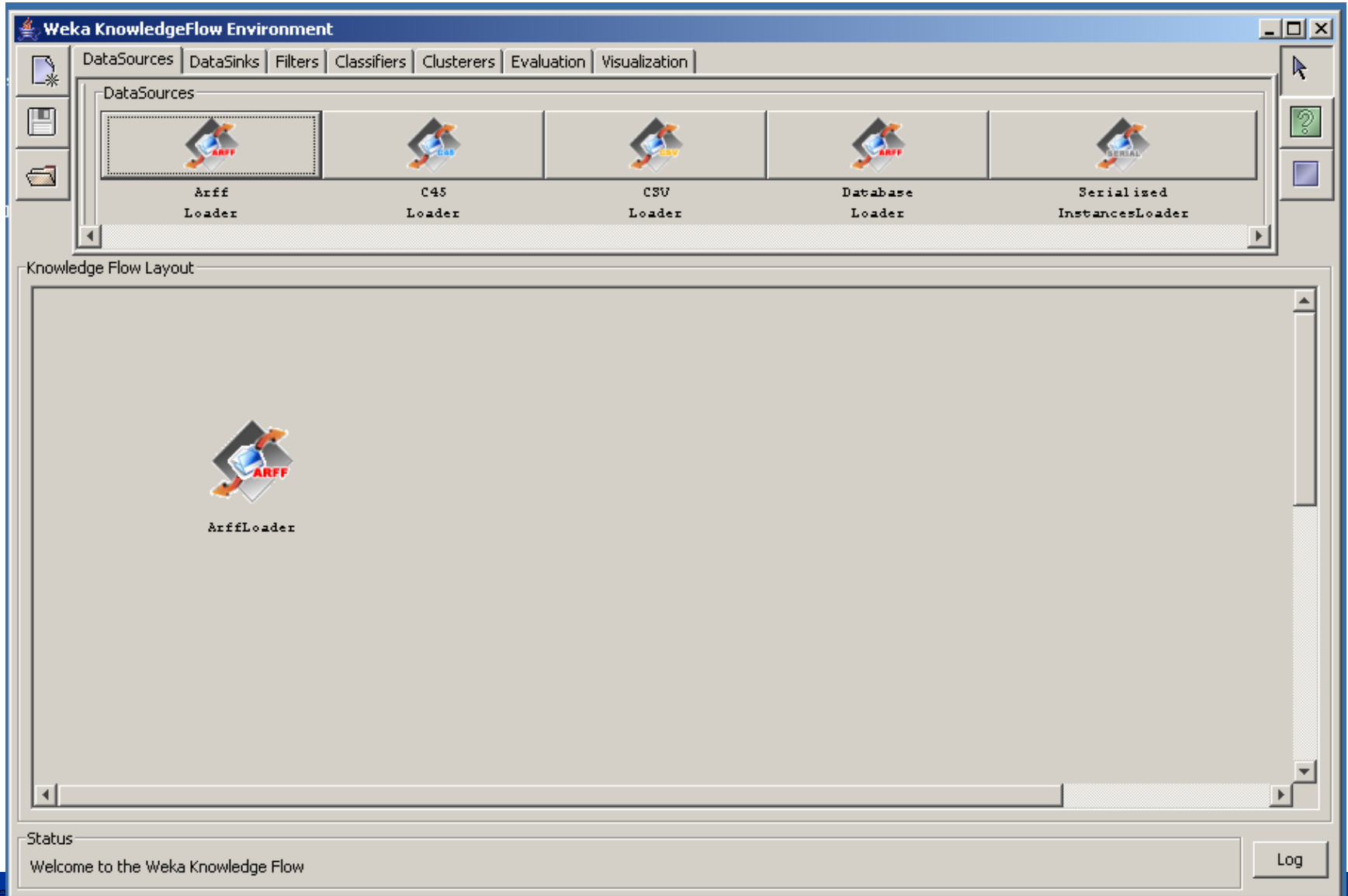
# Knowledge Flow

- Source file (.ARFF)

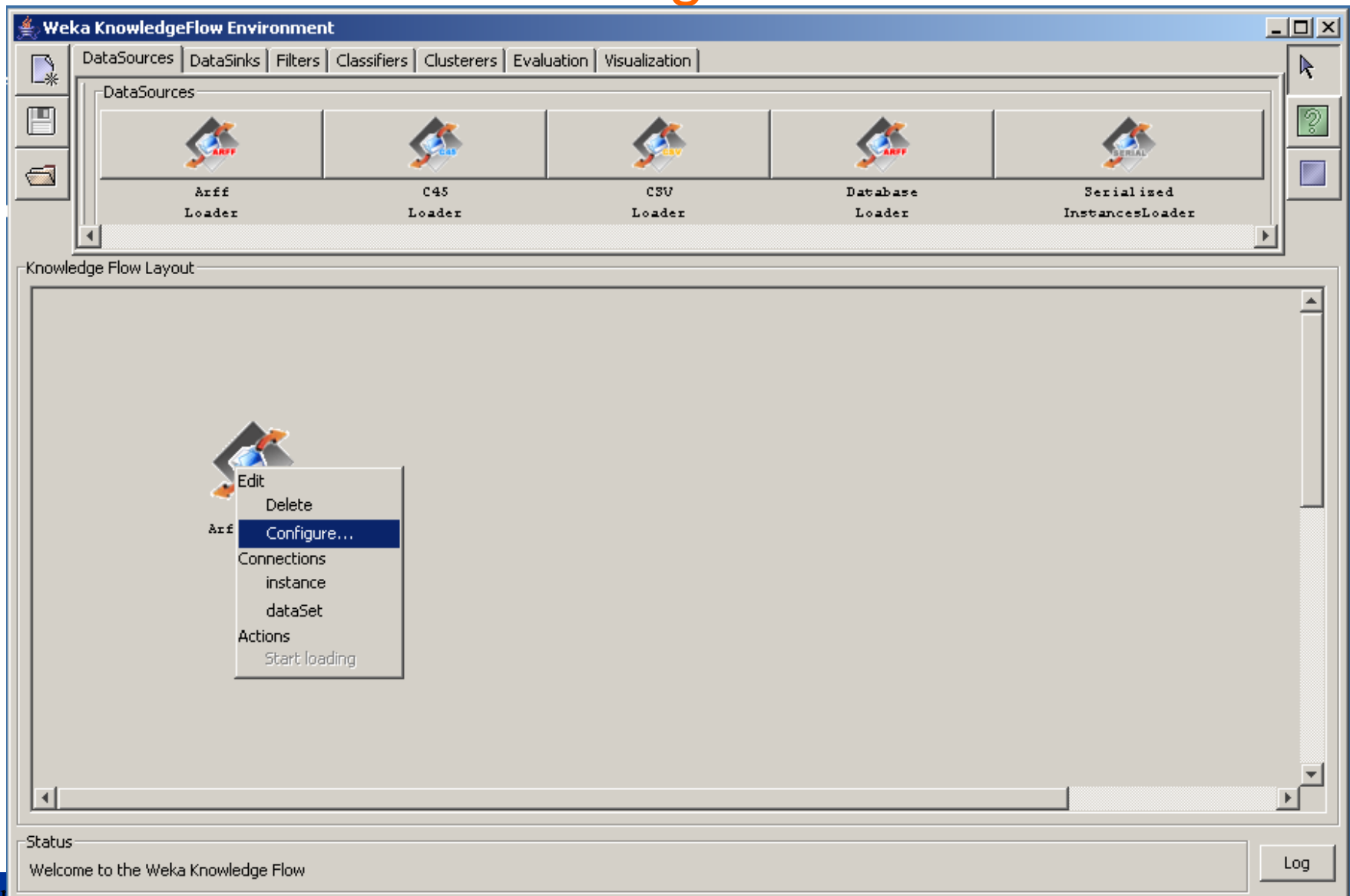
```
▶ 1 @relation weather.symbolic
  2
  3 @attribute outlook {sunny, overcast, rainy}
  4 @attribute temperature {hot, mild, cool}
  5 @attribute humidity {high, normal}
  6 @attribute windy {TRUE, FALSE}
  7 @attribute play {yes, no}
  8
  9 @data
10 sunny,hot,high,FALSE,no
11 sunny,hot,high,TRUE,no
12 overcast,hot,high,FALSE,yes
13 rainy,mild,high,FALSE,yes
14 rainy,cool,normal,FALSE,yes
15 rainy,cool,normal,TRUE,no
16 overcast,cool,normal,TRUE,yes
17 sunny,mild,high,FALSE,no
18 sunny,cool,normal,FALSE,yes
19 rainy,mild,normal,FALSE,yes
20 sunny,mild,normal,TRUE,yes
21 overcast,mild,high,TRUE,yes
22 overcast,hot,normal,FALSE,yes
23 rainy,mild,high,TRUE,no
24
```



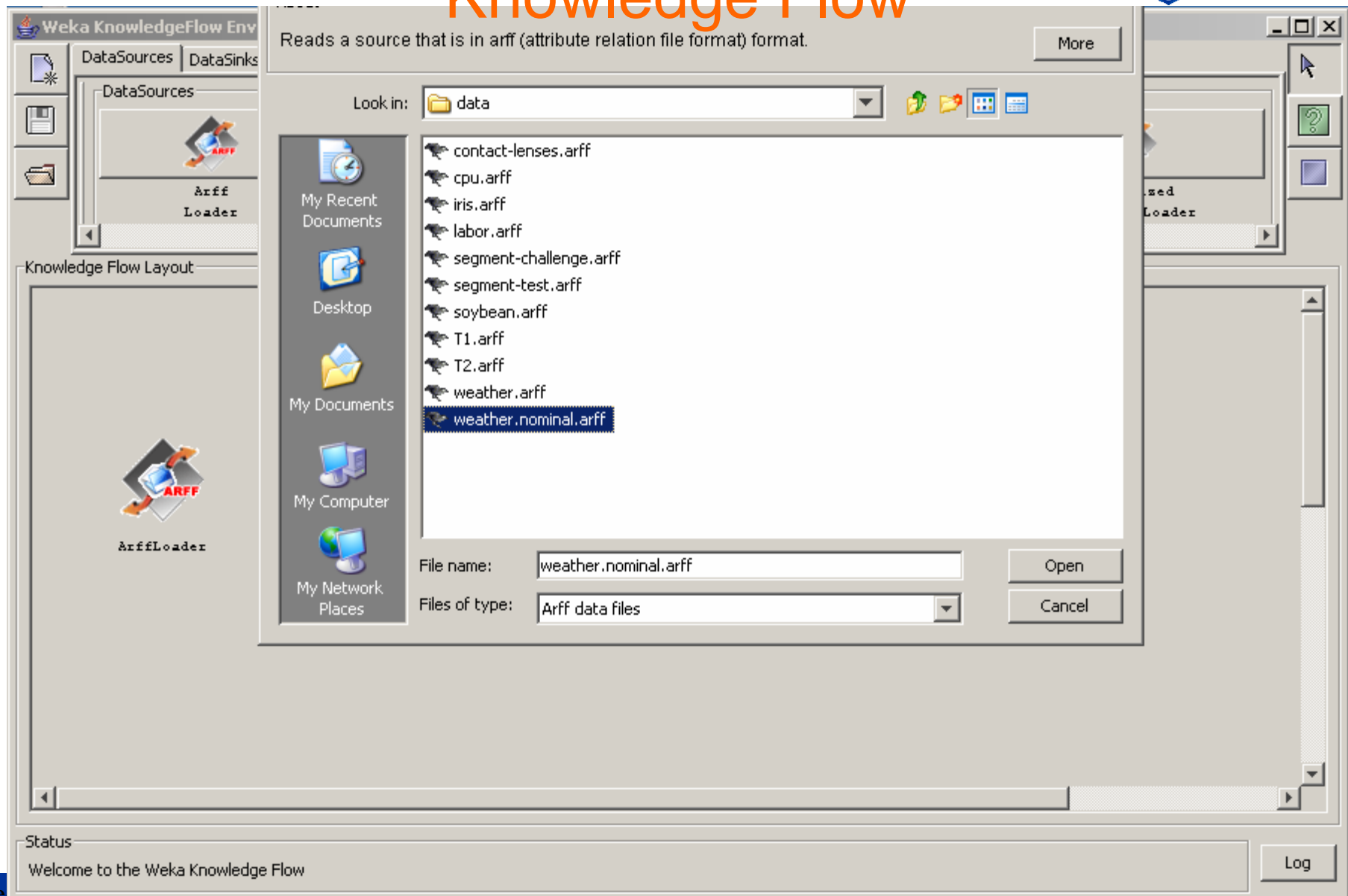
# Knowledge Flow



# Knowledge Flow



# Knowledge Flow



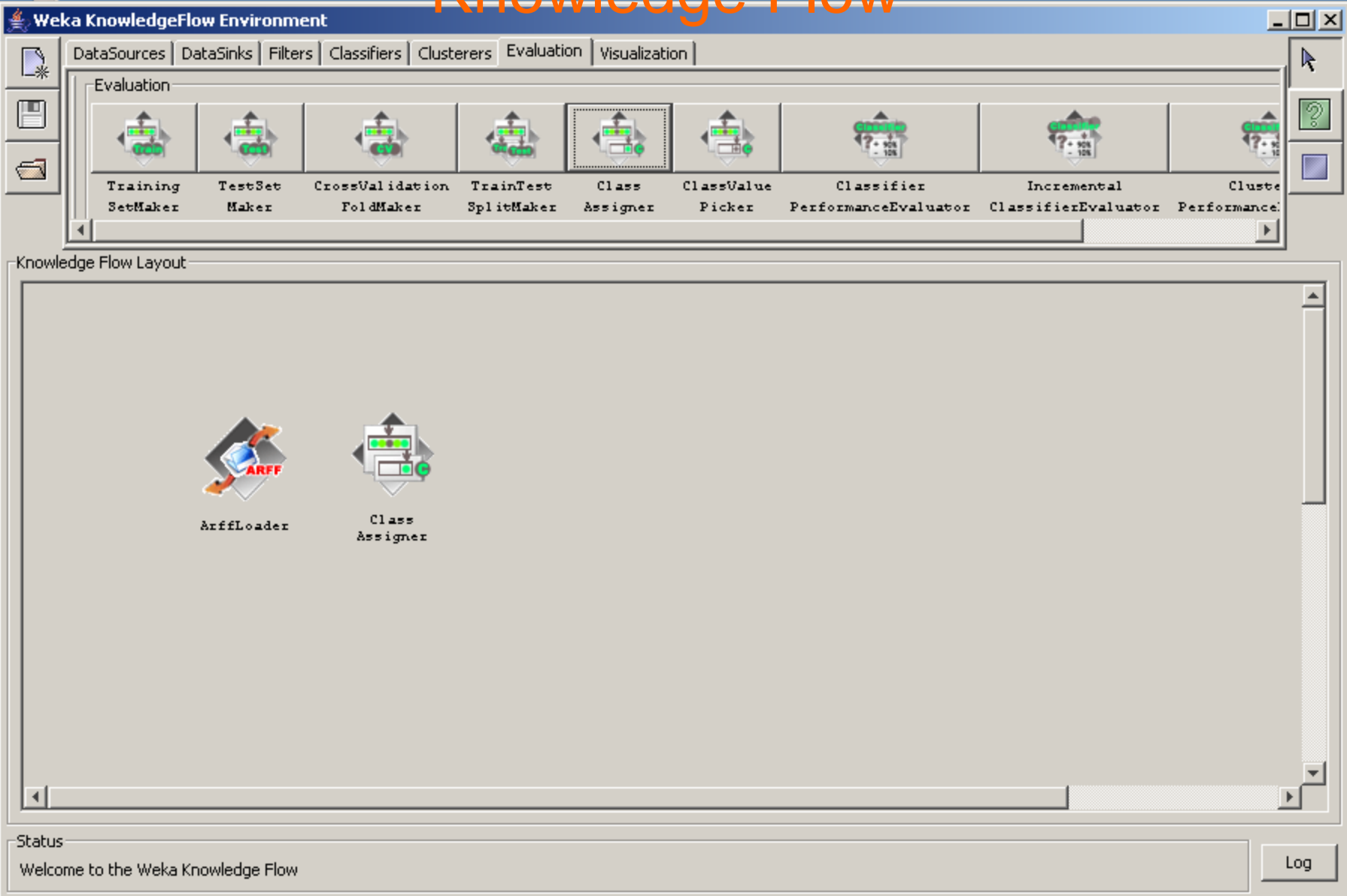
The screenshot displays the Weka KnowledgeFlow environment. A central window titled "Reads a source that is in arff (attribute relation file format) format." is open, showing a file selection dialog. The "Look in:" field is set to "data". The file list contains the following files:

- contact-lenses.arff
- cpu.arff
- iris.arff
- labor.arff
- segment-challenge.arff
- segment-test.arff
- soybean.arff
- T1.arff
- T2.arff
- weather.arff
- weather.nominal.arff** (highlighted)

The "File name:" field contains "weather.nominal.arff" and the "Files of type:" dropdown is set to "Arff data files". The "Open" and "Cancel" buttons are visible at the bottom right of the dialog.

In the background, the Weka KnowledgeFlow Env interface is visible, showing a "DataSources" panel with an "Arff Loader" component and a "Knowledge Flow Layout" area with an "ArffLoader" component. The status bar at the bottom reads "Welcome to the Weka Knowledge Flow" and includes a "Log" button.

# Knowledge Flow



Weka KnowledgeFlow Environment

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

Evaluation

Training SetMaker | TestSet Maker | CrossValidation FoldMaker | TrainTest SplitMaker | **Class Assigner** | ClassValue Picker | Classifier PerformanceEvaluator | Incremental ClassifierEvaluator | Cluste Performance

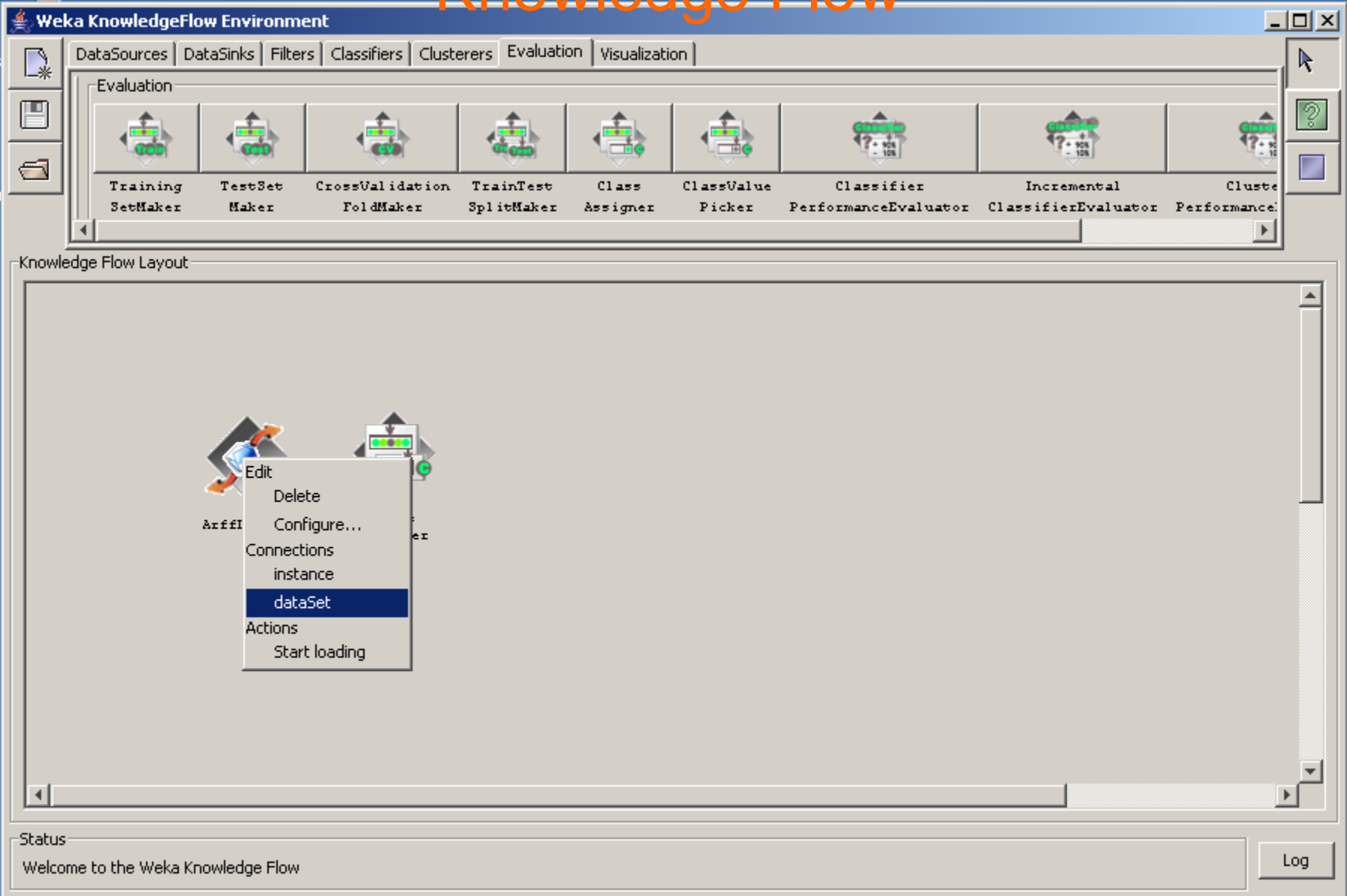
Knowledge Flow Layout

ArffLoader | Class Assigner

Status  
 Welcome to the Weka Knowledge Flow

Log

# Knowledge Flow



Weka KnowledgeFlow Environment

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

Evaluation

Training SetMaker | TestSet Maker | CrossValidation FoldMaker | TrainTest SplitMaker | Class Assigner | ClassValue Picker | Classifier PerformanceEvaluator | Incremental ClassifierEvaluator | Cluste Performance

Knowledge Flow Layout

- Edit
- Delete
- Configure...
- Connections instance
- dataSet**
- Actions
- Start loading

Status  
Welcome to the Weka Knowledge Flow










Log

# Knowledge Flow

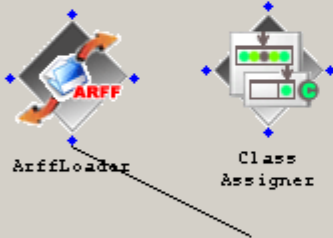
**Weka KnowledgeFlow Environment**

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

Evaluation

								
Training Set Maker	Test Set Maker	Cross Validation Fold Maker	Train Test Split Maker	Class Assigner	Class Value Picker	Classifier Performance Evaluator	Incremental Classifier Evaluator	Cluster Performance Evaluator

Knowledge Flow Layout



```

graph LR
  ArffLoader[ArffLoader] --> ClassAssigner[Class Assigner]
  
```

Status

Welcome to the Weka Knowledge Flow

Log

# Knowledge Flow

**Weka KnowledgeFlow Environment**

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

Evaluation

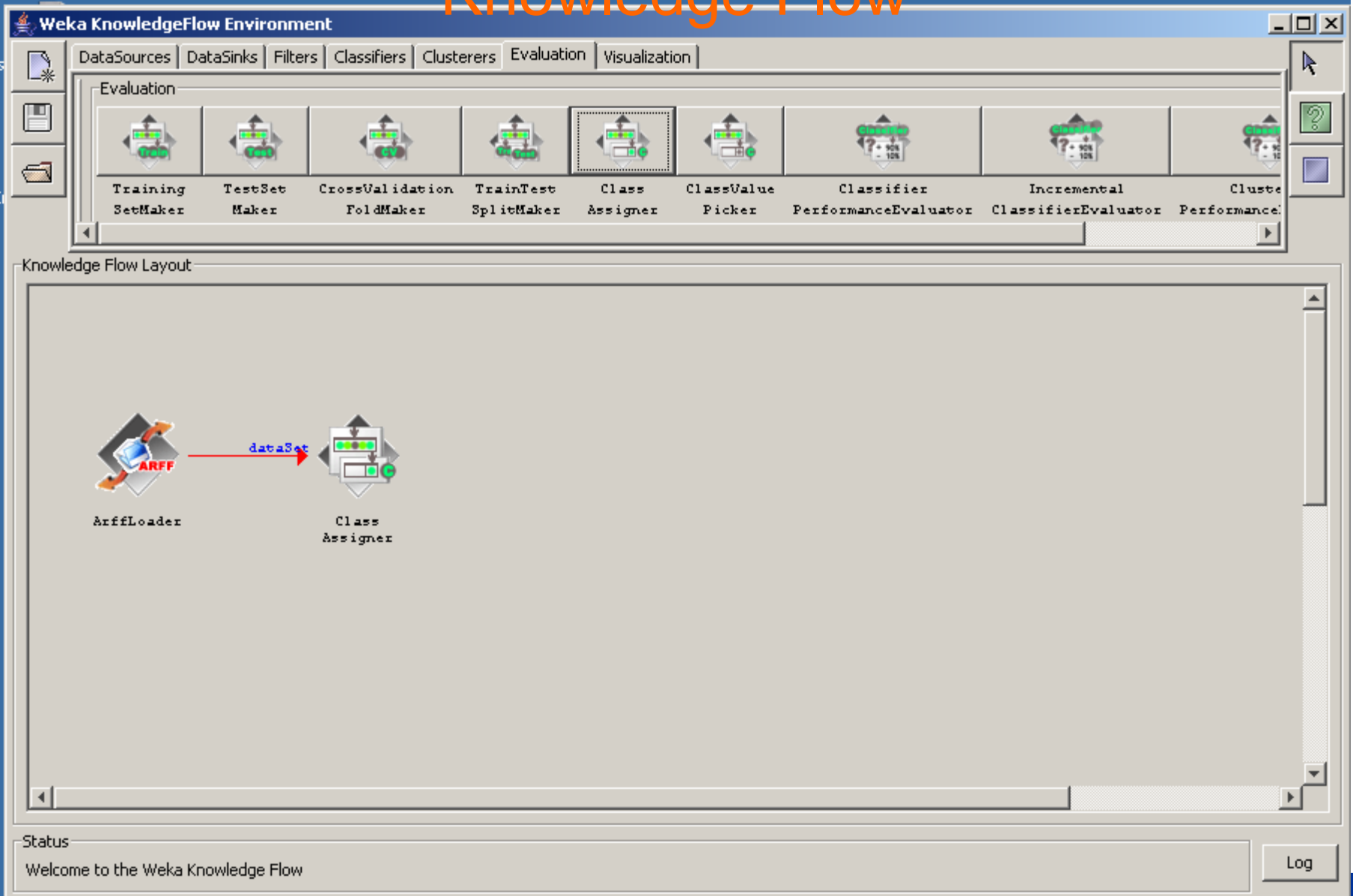
Training SetMaker | TestSet Maker | CrossValidation FoldMaker | TrainTest SplitMaker | **Class Assigner** | ClassValue Picker | Classifier PerformanceEvaluator | Incremental ClassifierEvaluator | Cluste Performance

Knowledge Flow Layout

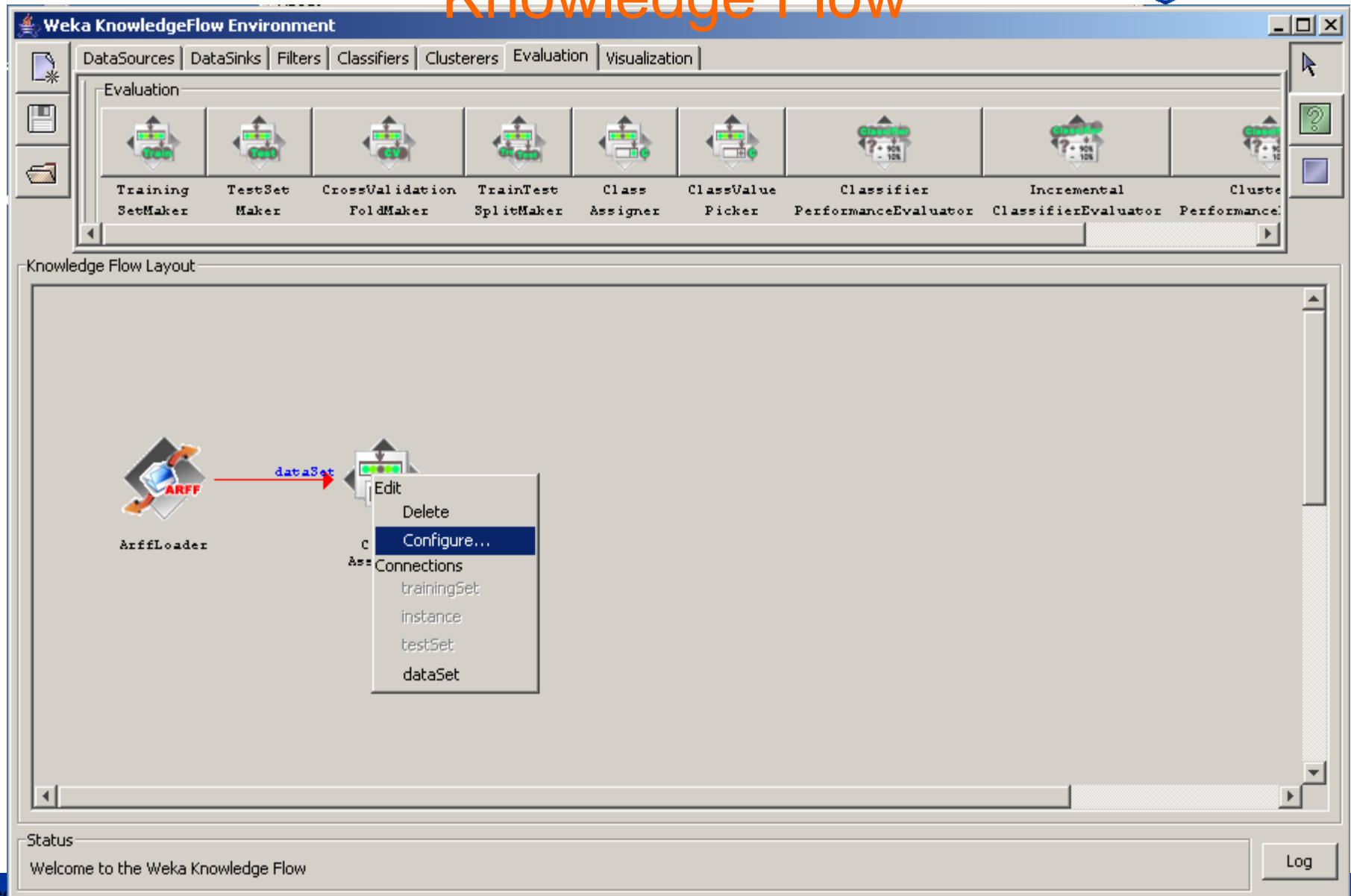
ArffLoader → dataSet → Class Assigner

Status  
 Welcome to the Weka Knowledge Flow

Log



# Knowledge Flow



The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusterers, Evaluation, and Visualization. The 'Evaluation' tab is active, showing a toolbar with icons for Training Set Maker, Test Set Maker, Cross Validation Fold Maker, Train Test Split Maker, Class Assigner, Class Value Picker, Classifier Performance Evaluator, Incremental Classifier Evaluator, and Cluster Performance Evaluator. The main workspace, titled 'Knowledge Flow Layout', contains an 'ArffLoader' node connected to a 'Filter' node by a red arrow labeled 'dataSet'. A context menu is open over the 'Filter' node, listing options: Edit, Delete, Configure..., Connections, trainingSet, instance, testSet, and dataSet. The status bar at the bottom reads 'Welcome to the Weka Knowledge Flow' and includes a 'Log' button.



# Knowledge Flow

Weka KnowledgeFlow Environment

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

Evaluation

Training SetMaker | TestSet Maker | CrossValidation FoldMaker | TrainTest SplitMaker | Class Assigner | ClassValue Picker | Classifier PerformanceEvaluator | Incremental ClassifierEvaluator | Cluste Performance

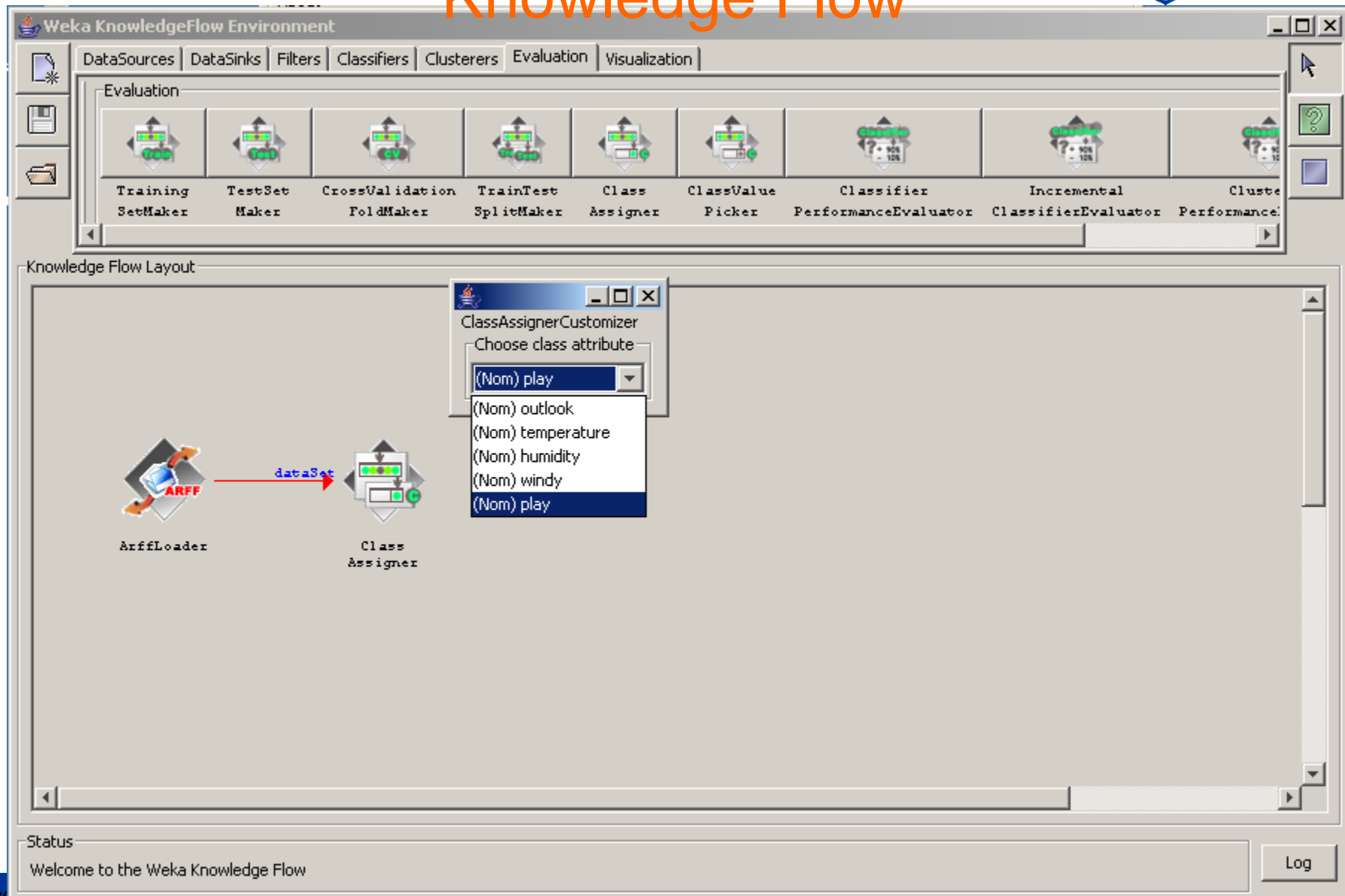
Knowledge Flow Layout

ArffLoader → dataSet → Class Assigner

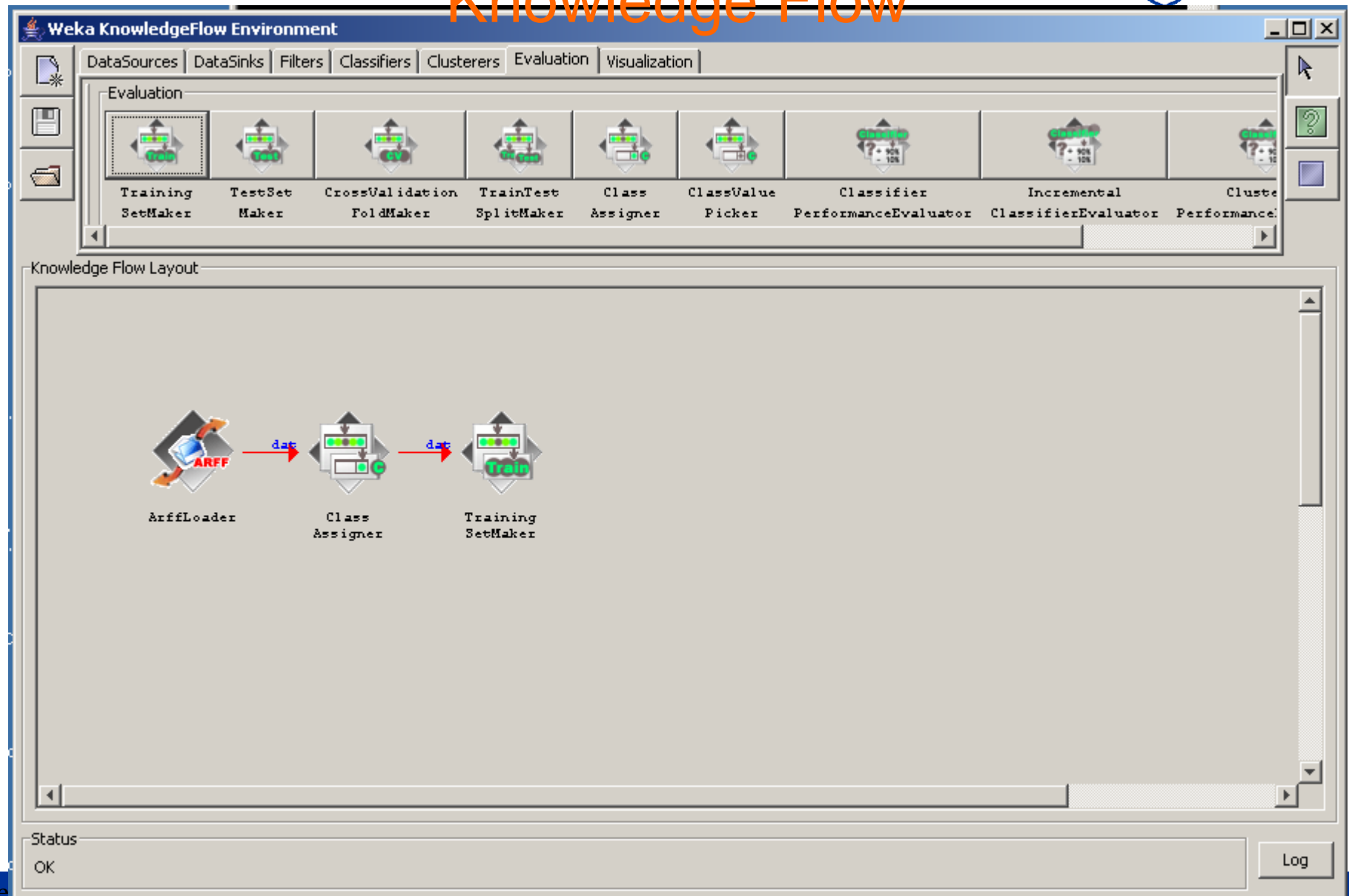
ClassAssignerCustomizer  
 Choose class attribute  
 (Nom) play  
 (Nom) outlook  
 (Nom) temperature  
 (Nom) humidity  
 (Nom) windy  
 (Nom) play

Status  
 Welcome to the Weka Knowledge Flow

Log

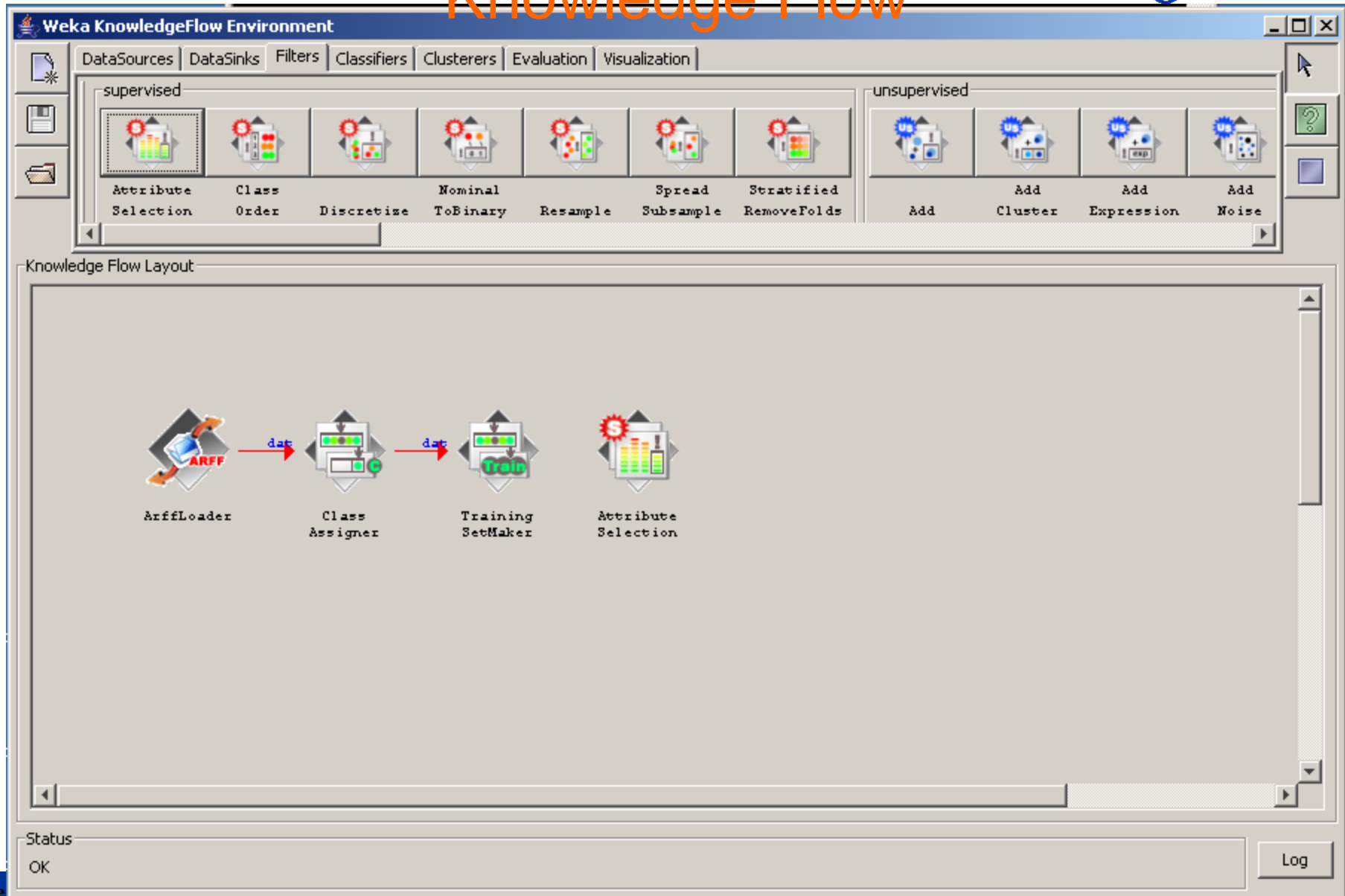


# Knowledge Flow



The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusterers, Evaluation, and Visualization. The 'Evaluation' tab is active, showing a palette of components including Training SetMaker, TestSet Maker, CrossValidation FoldMaker, TrainTest SplitMaker, Class Assigner, ClassValue Picker, Classifier Performance Evaluator, Incremental Classifier Evaluator, and Clustering Performance Evaluator. Below this is the 'Knowledge Flow Layout' area, which contains a workflow diagram. The diagram shows three components connected by arrows: 'ArffLoader' (with an 'ARFF' icon) outputs a 'dat' file to 'Class Assigner', which in turn outputs a 'dat' file to 'Training SetMaker'. The 'Training SetMaker' component has a 'Train' icon. At the bottom of the window, there is a 'Status' field showing 'OK' and a 'Log' button.

# Knowledge Flow



The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusterers, Evaluation, and Visualization. Below these tabs are two main categories of filters: supervised and unsupervised. The supervised filters include Attribute Selection, Class Order, Discretize, Nominal To Binary, Resample, Spread Subsample, and Stratified Remove Folds. The unsupervised filters include Add, Add Cluster, Add Expression, and Add Noise. The main workspace, titled 'Knowledge Flow Layout', shows a workflow with four nodes: ArffLoader, Class Assigner, Training SetMaker, and Attribute Selection. ArffLoader is connected to Class Assigner, and Class Assigner is connected to Training SetMaker, both connections labeled 'dat'. Attribute Selection is positioned to the right of Training SetMaker. The status bar at the bottom shows 'Status OK' and a 'Log' button.

# Knowledge Flow

Weka KnowledgeFlow Environment

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

supervised

Attribute Selection | Class Order | Discretize | Nominal ToBinary | Resample | Spread Subsample | Stratified RemoveFolds

unsupervised

Add | Add Cluster | Add Expression | Add Noise

Knowledge Flow Layout

```

  graph LR
    A[ArffLoader] -- dat --> B[Class Assigner]
    B -- dat --> C[Training SetMaker]
    C -- tr --> D[Attribute Selection]
  
```

ArffLoader → Class Assigner → Training SetMaker → Attribute Selection

Choose weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes More

evaluator Choose **CfsSubsetEval**

search Choose **BestFirst -D 1 -N 5**

Open... Save... OK Cancel

Status  
OK Log

# Knowledge Flow

Weka KnowledgeFlow Environment

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

supervised

Attribute Selection | Class Order | Discretize | Nominal To Binary | Resample | Spread Subsample | Stratified Remove Folds

unsupervised

Add | Add Cluster | Add Expression | Add Noise

Knowledge Flow Layout

```

  graph LR
    ARFF[ArffLoader] -- dat --> CA[Class Assigner]
    CA -- dat --> TSM[Training SetMaker]
    TSM -- tr --> AS[Attribute Selection]
  
```

Choose weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes More

evaluator

search

Open.

- weka
  - attributeSelection
    - CfsSubsetEval
    - ChiSquaredAttributeEval
    - ClassifierSubsetEval
    - ConsistencySubsetEval
    - GainRatioAttributeEval
    - InfoGainAttributeEval
    - OneRAttributeEval
    - PrincipalComponents
    - ReliefFAttributeEval
    - SVMAttributeEval
    - SymmetricalUncertAttributeEval
    - WrapperSubsetEval

Status

OK

Log

# Knowledge Flow

Weka KnowledgeFlow Environment

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

supervised

Attribute Selection | Class Order | Discretize | Nominal To Binary | Resample | Spread Subsample | Stratified Remove Folds

unsupervised

Add | Add Cluster | Add Expression | Add Noise

Knowledge Flow Layout

```

  graph LR
    A[ArffLoader] -- dat --> B[Class Assigner]
    B -- dat --> C[Training SetMaker]
    C -- ts --> D[Attribute Selection]
  
```

ArffLoader → Class Assigner → Training SetMaker → Attribute Selection

Choose weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes More

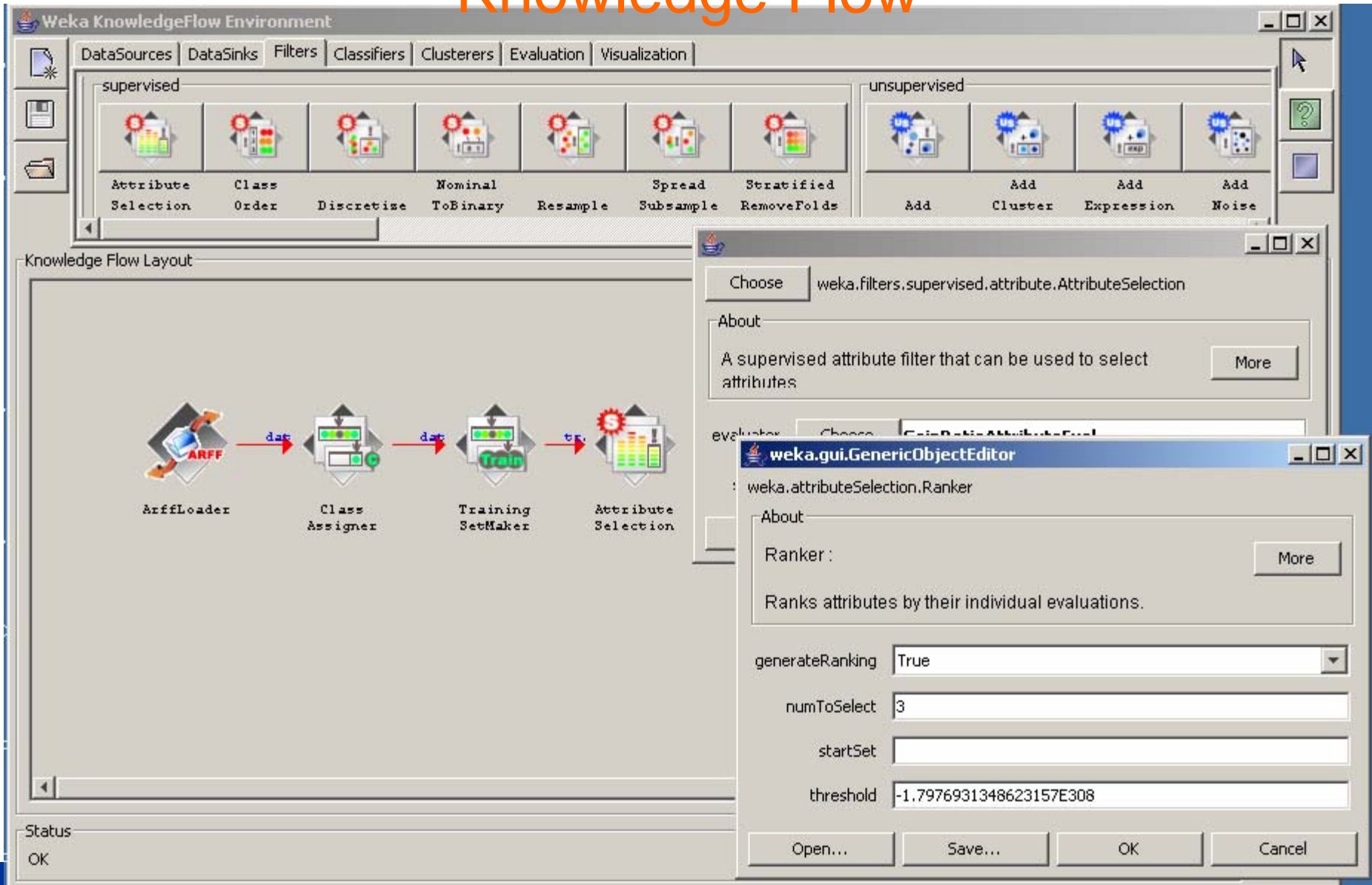
evaluator Choose **GainRatioAttributeEval**

search Choose **Ranker -T -1.7976931348623157E308 -N -1**

Open... Save... OK Cancel

Status  
OK Log

# Knowledge Flow



The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusterers, Evaluation, and Visualization. Below these are two main sections: supervised and unsupervised. The supervised section contains icons for Attribute Selection, Class Order, Discretize, Nominal To Binary, Resample, Spread Subsample, and Stratified Remove Folds. The unsupervised section contains icons for Add, Add Cluster, Add Expression, and Add Noise.

The Knowledge Flow Layout area shows a sequence of four nodes connected by arrows: ArffLoader (data) → Class Assigner (data) → Training SetMaker (train) → Attribute Selection (S). The status bar at the bottom indicates "OK".

Two dialog boxes are open over the interface:

- The first dialog, titled "Choose weka.filters.supervised.attribute.AttributeSelection", shows the "About" section: "A supervised attribute filter that can be used to select attributes." It includes a "More" button.
- The second dialog, titled "weka.gui.GenericObjectEditor", is for the "Attribute Selection" node. It shows the "About" section: "Ranker : Ranks attributes by their individual evaluations." It includes a "More" button and several configuration fields:
  - generateRanking: True
  - numToSelect: 3
  - startSet: (empty)
  - threshold: -1.7976931348623157E308

# Knowledge Flow

**Weka KnowledgeFlow Environment**

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

Discretisation    Stacking    StackingC    Threshold Selector    Vote    trees  
 AD Tree    Decision Stump    ID3    J48    LMT    MSP

Knowledge Flow Layout

```

    graph LR
      ArffLoader[ArffLoader] -- dat --> ClassAssigner[Class Assigner]
      ClassAssigner -- dat --> TrainingSetMaker[Training SetMaker]
      TrainingSetMaker -- tr --> AttributeSelection[Attribute Selection]
      AttributeSelection --> ID3[ID3]
  
```

Status  
OK

Log



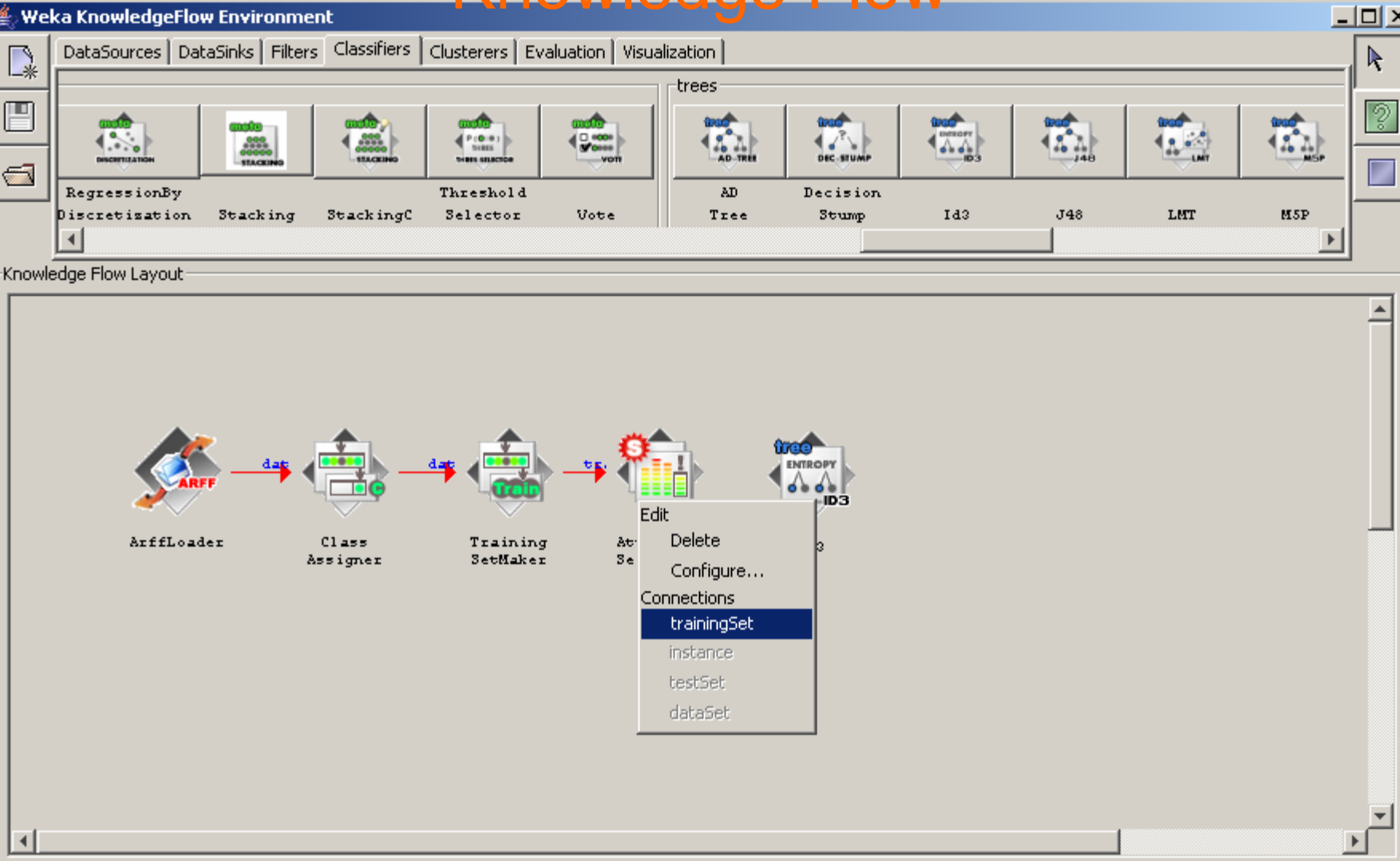
# Knowledge Flow

**Weka KnowledgeFlow Environment**

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

RegressionBy Discretisation | Stacking | StackingC | Threshold Selector | Vote | trees | AD Tree | Decision Stump | Id3 | J48 | LMT | M5P

Knowledge Flow Layout



```

    graph LR
      ArffLoader[ArffLoader] -- dat --> ClassAssigner[Class Assigner]
      ClassAssigner -- dat --> TrainingSetMaker[Training SetMaker]
      TrainingSetMaker -- tr. --> Id3[Id3]
  
```

Status  
 OK

Log


# Knowledge Flow

**Weka KnowledgeFlow Environment**

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

CrossValidation FoldMaker | 
 TrainTest SplitMaker | 
 Class Assigner | 
 ClassValue Picker | 
 Classifier PerformanceEvaluator | 
 Incremental ClassifierEvaluator | 
 Clusterer PerformanceEvaluator | 
 Prediction Appender

Knowledge Flow Layout



```

    graph LR
      ArffLoader[ArffLoader ARFF] -- ".dat" --> ClassAssigner[Class Assigner]
      ClassAssigner -- ".dat" --> TrainingSetMaker[Training SetMaker]
      TrainingSetMaker -- ".tr" --> AttributeSelection[Attribute Selection]
      AttributeSelection -- ".tra" --> Id3[Id3 tree ENTROPY ID3]
  
```

ArffLoader    Class Assigner    Training SetMaker    Attribute Selection    Id3

Status  
OK

Log

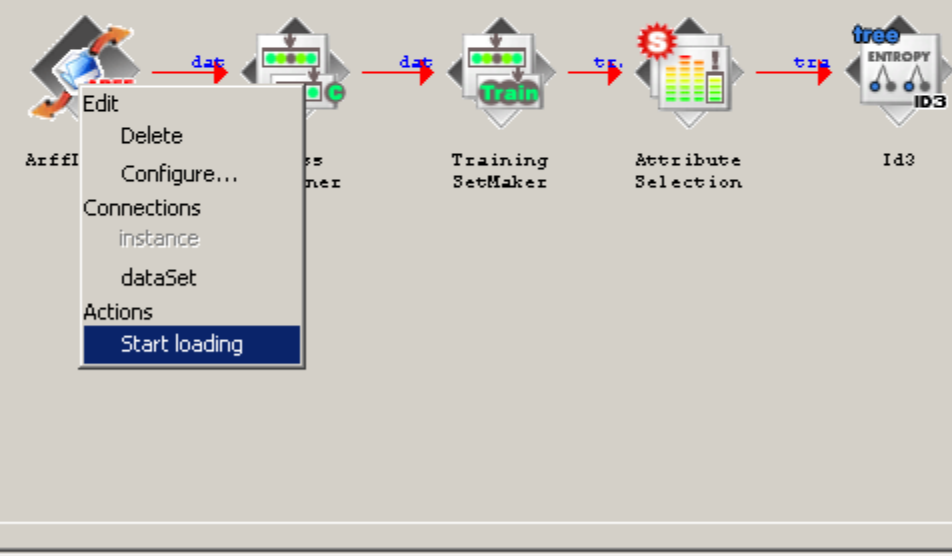
# Knowledge Flow

**Weka KnowledgeFlow Environment**

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

CrossValidation FoldMaker | TrainTest SplitMaker | Class Assigner | ClassValue Picker | Classifier PerformanceEvaluator | Incremental ClassifierEvaluator | Clusterer PerformanceEvaluator | Prediction Appender

Knowledge Flow Layout



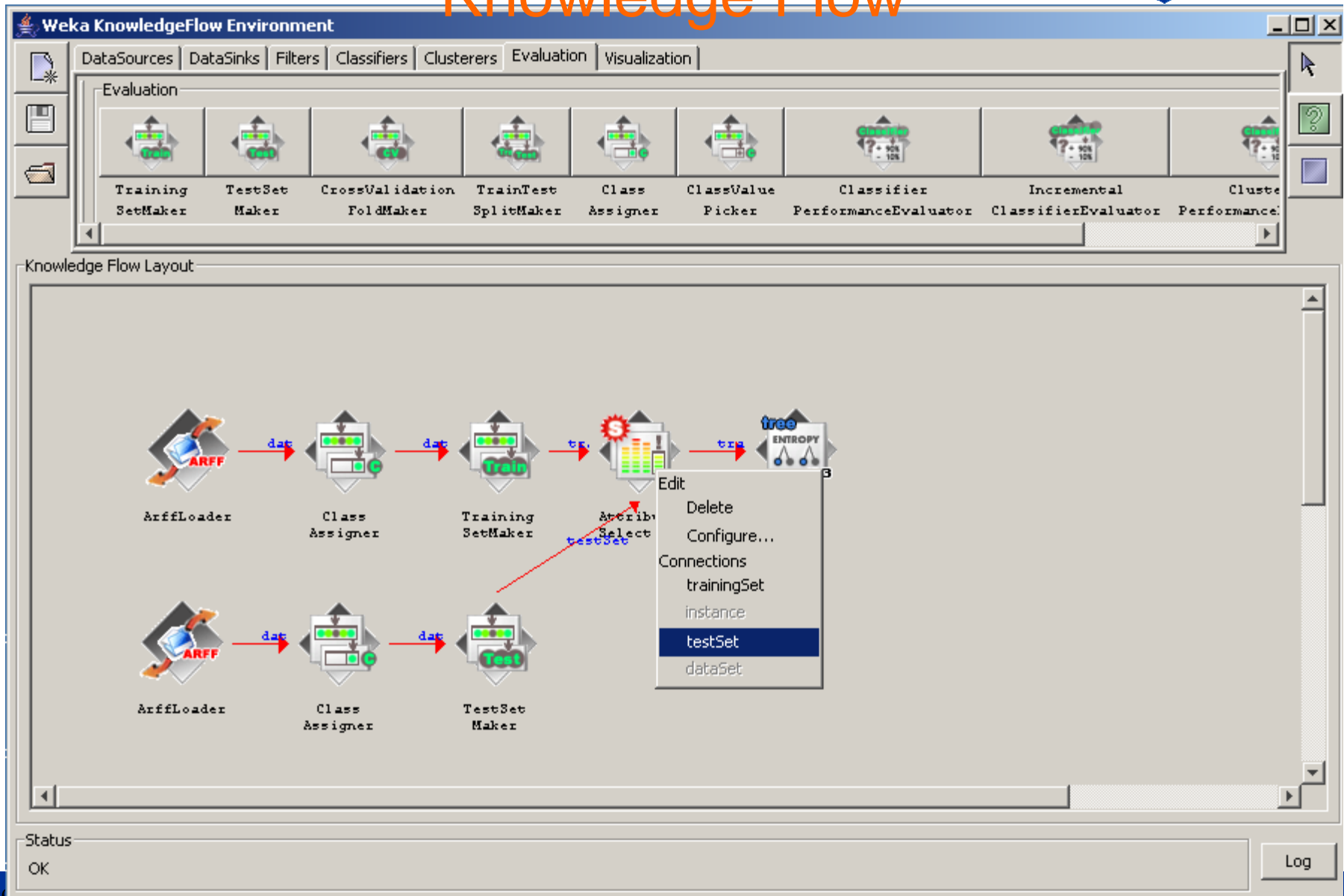
```

graph LR
  ArffFile[ArffFile] -- dat --> TrainingSetMaker[Training SetMaker]
  TrainingSetMaker -- dat --> AttributeSelection[Attribute Selection]
  AttributeSelection -- tr --> Id3[tree ENTROPY ID3]
  
```

Status  
OK

Log

# Knowledge Flow



The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there is a menu bar with categories: DataSources, DataSinks, Filters, Classifiers, Clusterers, Evaluation, and Visualization. Below this is a toolbar with icons for various operations. The main workspace, titled "Knowledge Flow Layout", contains a workflow diagram. The diagram shows two parallel paths starting from "ArffLoader" nodes. The top path consists of "ArffLoader" (ARFF) connected to "Class Assigner", then "Training SetMaker", and finally "Entropy" (tree). The bottom path consists of "ArffLoader" (ARFF) connected to "Class Assigner", then "TestSet Maker", and finally "Entropy" (tree). A context menu is open over the "Entropy" node in the top path, with the "testSet" option selected. The menu items are: Edit, Delete, Configure..., Connections, trainingSet, instance, testSet, and dataSet. The status bar at the bottom left shows "Status OK", and there is a "Log" button at the bottom right.

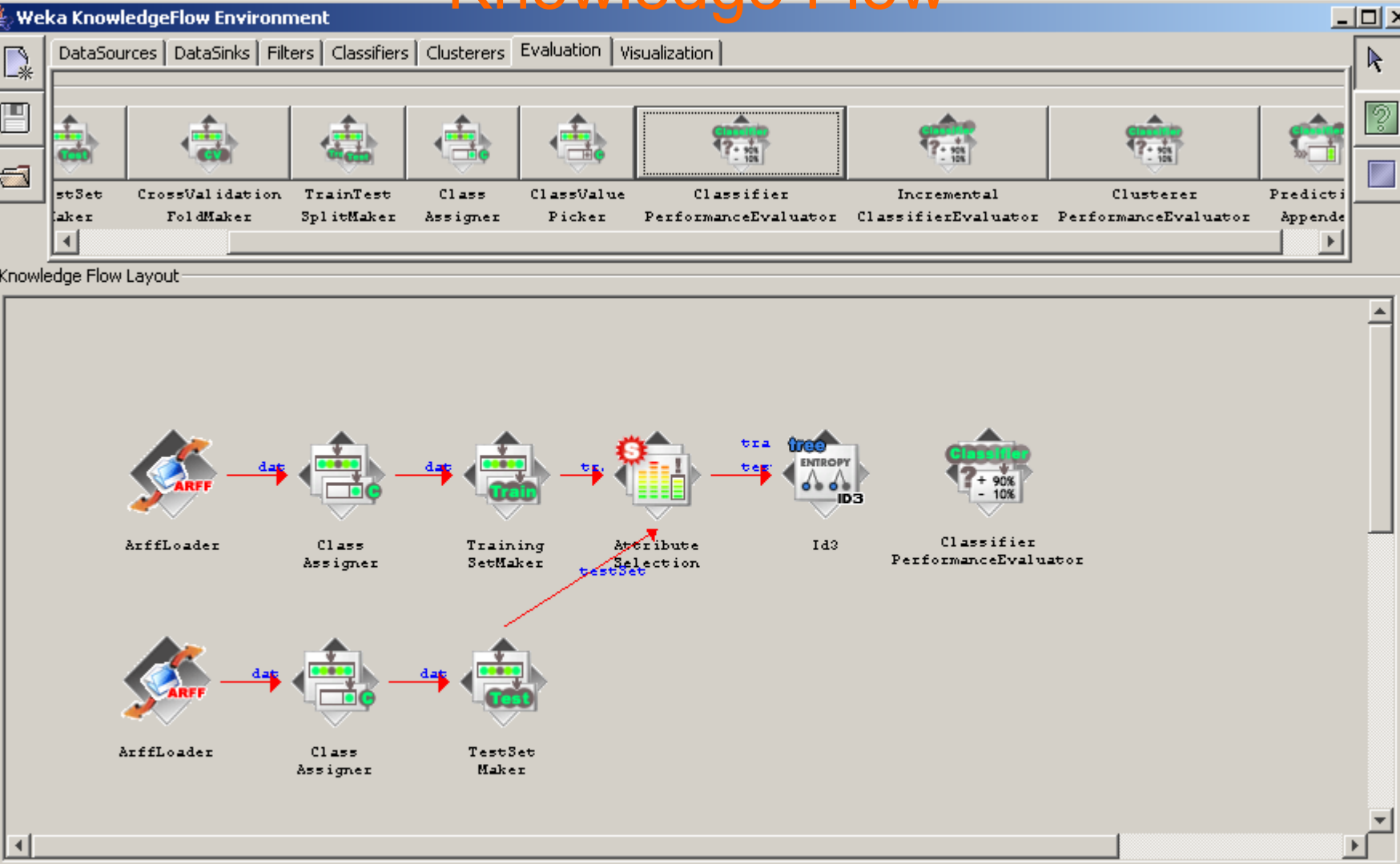
# Knowledge Flow

**Weka KnowledgeFlow Environment**

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

TestSetMaker | CrossValidationFoldMaker | TrainTestSplitMaker | ClassAssigner | ClassValuePicker | ClassifierPerformanceEvaluator | IncrementalClassifierEvaluator | ClustererPerformanceEvaluator | PredictionAppender

Knowledge Flow Layout



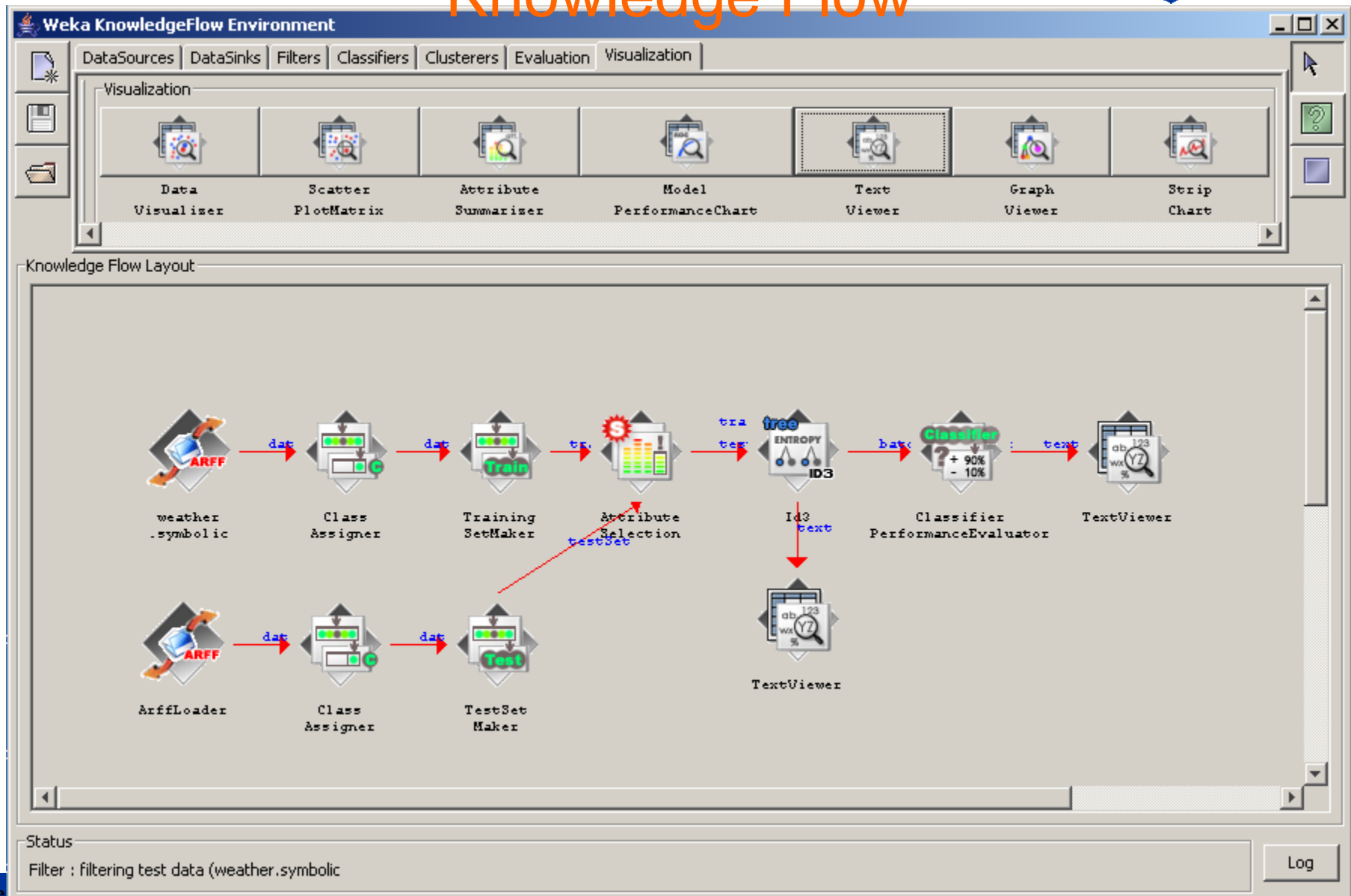
```

graph LR
    subgraph Top_Path
        A1[ArffLoader] -- dat --> B1[Class Assigner]
        B1 -- dat --> C1[Training SetMaker]
        C1 -- tr --> D1[Attribute Selection]
        D1 -- tra --> E1[Id3]
        E1 --> F1[Classifier Performance Evaluator]
    end
    subgraph Bottom_Path
        A2[ArffLoader] -- dat --> B2[Class Assigner]
        B2 -- dat --> C2[Test Set Maker]
    end
    C2 -- testSet --> D1
  
```

Status  
OK

Log

# Knowledge Flow



# Knowledge Flow

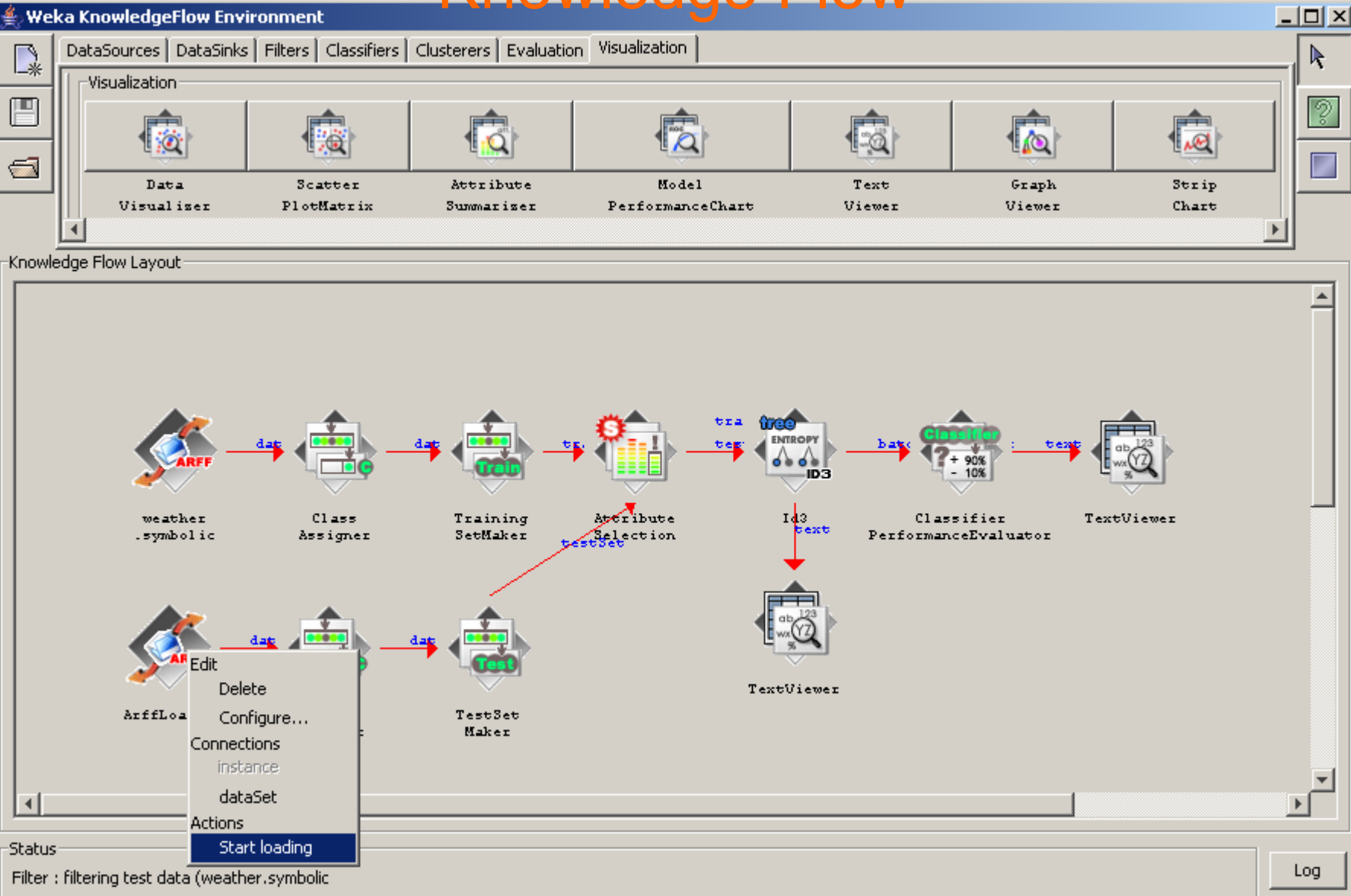
**Weka KnowledgeFlow Environment**

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

Visualization

Data Visualiser | Scatter PlotMatrix | Attribute Summariser | Model PerformanceChart | Text Viewer | Graph Viewer | Strip Chart

Knowledge Flow Layout



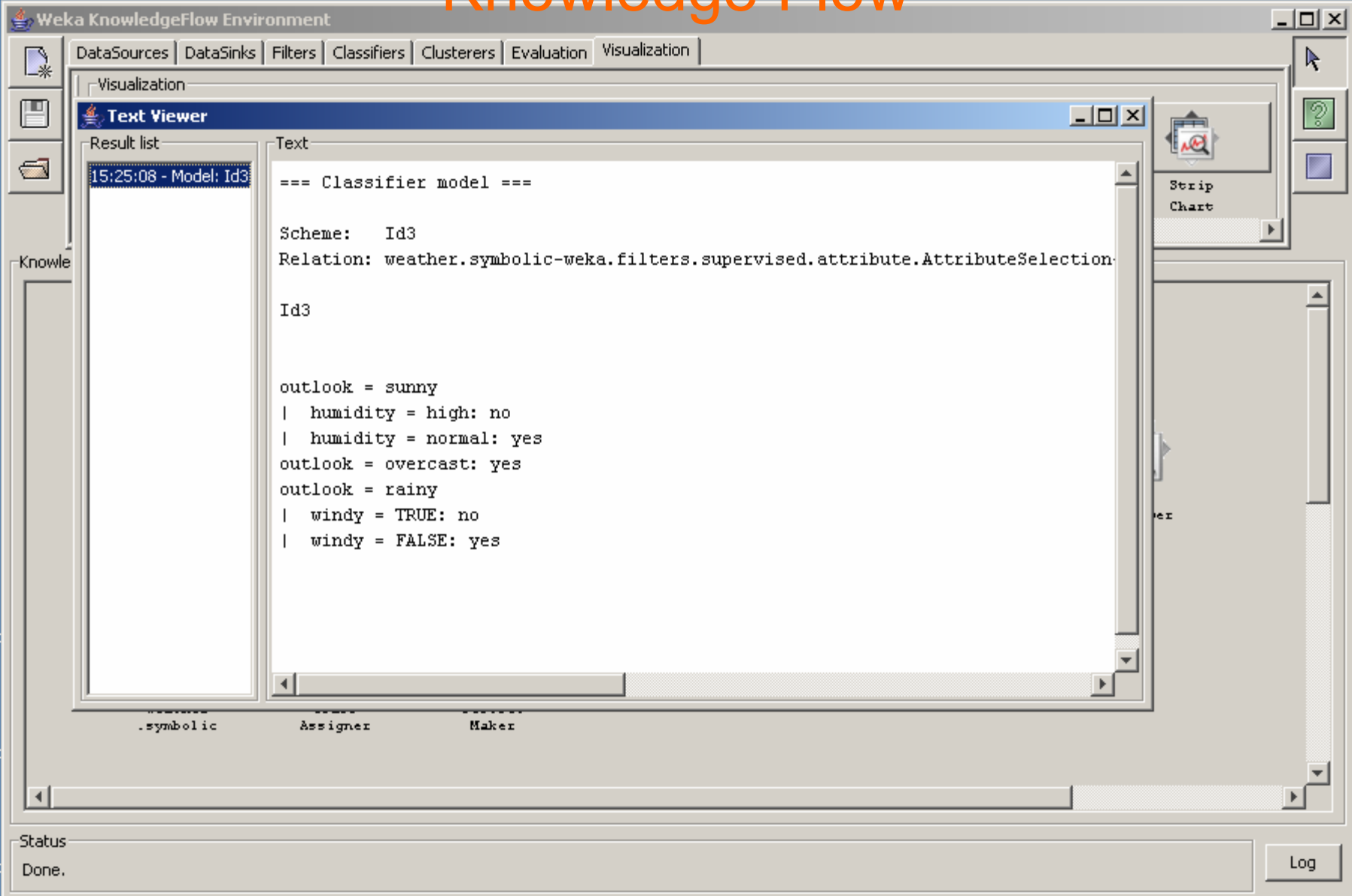
```

graph LR
    A[weather.symbolic] -- dat --> B[Class Assigner]
    B -- dat --> C[Training SetMaker]
    C -- tr --> D[Attribute Selection]
    D -- tra --> E[tree ID3]
    E -- bat --> F[Classifier PerformanceEvaluator]
    F -- text --> G[TextViewer]
    A -- dat --> H[TestSet Maker]
    H -- text --> I[TextViewer]
  
```

Status: Filter : filtering test data (weather.symbolic)

Log

# Knowledge Flow



The screenshot shows the Weka KnowledgeFlow Environment interface. The main window is titled "Text Viewer" and displays the output of a classifier model. The output is as follows:

```

15:25:08 - Model: Id3

=== Classifier model ===

Scheme:  Id3
Relation: weather.symbolic-weka.filters.supervised.attribute.AttributeSelection

Id3

outlook = sunny
|  humidity = high: no
|  humidity = normal: yes
outlook = overcast: yes
outlook = rainy
|  windy = TRUE: no
|  windy = FALSE: yes
  
```

The interface includes a top menu bar with options: DataSources, DataSinks, Filters, Classifiers, Clusterers, Evaluation, and Visualization. A left sidebar contains icons for file operations and a "Knowle" section. The bottom status bar shows "Status Done." and a "Log" button.



# Knowledge Flow

Weka KnowledgeFlow Environment

DataSources | DataSinks | Filters | **Text Viewer**

Visualization

Data Visualizer

Knowledge Flow Layout

weather .symbolic

weather .symbolic

Status: Done.

Result list

15:21:02 - Id3  
 15:25:10 - Id3

Text

```

=== Evaluation result ===

Scheme: Id3
Relation: weather.symbolic-weka.filters.supervised.attribute.AttributeSelect

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
  
```

Log



## Outline

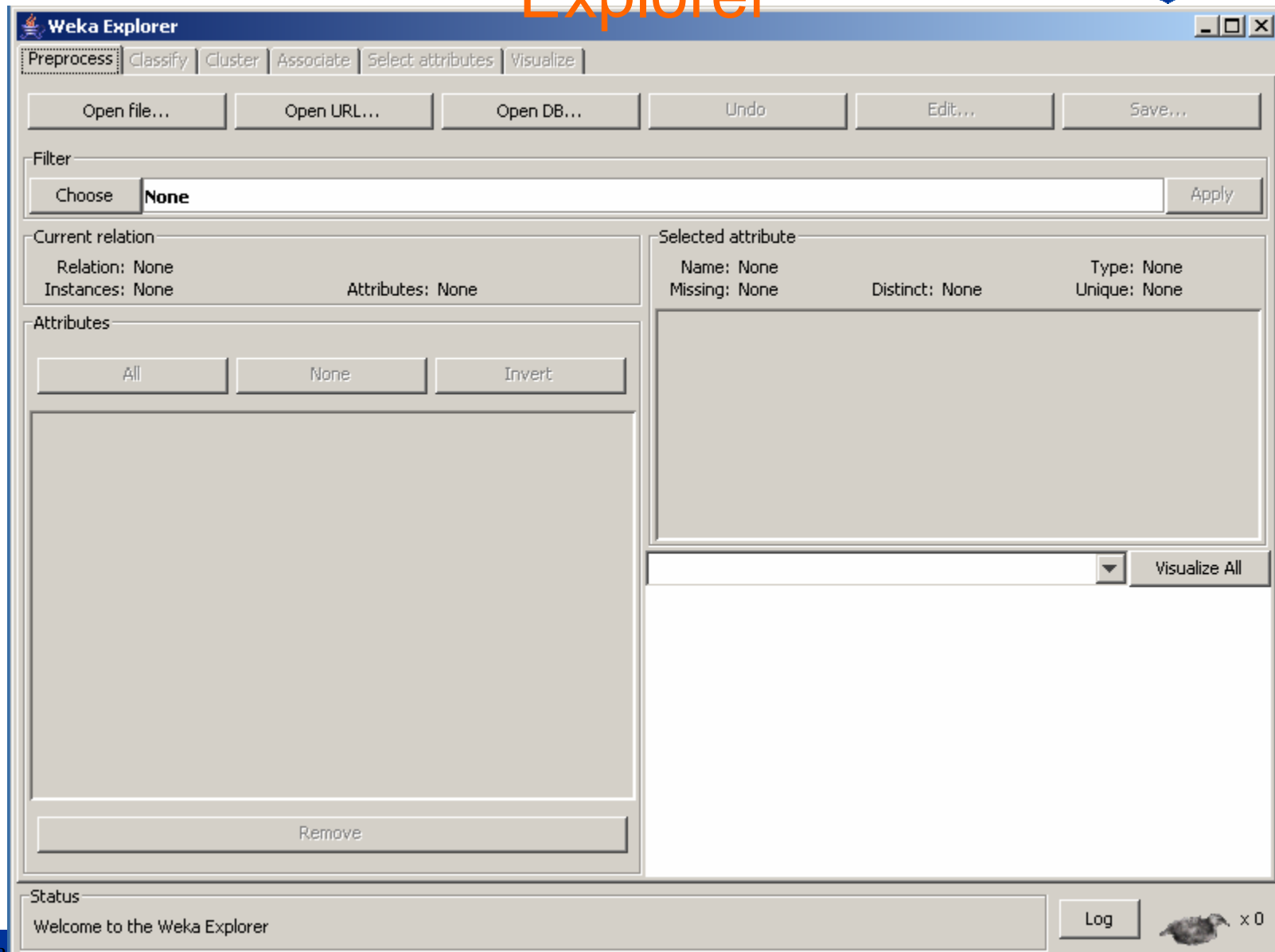
- What is WEKA
- Knowledge Flow
- **Explorer**
- Why Knowledge Flow
- Cross Validation
- Reference



## 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



# 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

Selected attribute: Name: outlook Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

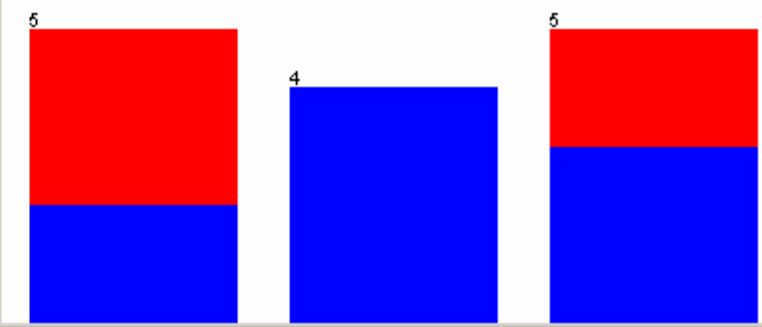
Label	Count
sunny	5
overcast	4
rainy	5


Attributes: All | None | Invert

No.	Name
<input checked="" type="checkbox"/> 1	outlook
<input type="checkbox"/> 2	temperature
<input type="checkbox"/> 3	humidity
<input type="checkbox"/> 4	windy
<input type="checkbox"/> 5	play

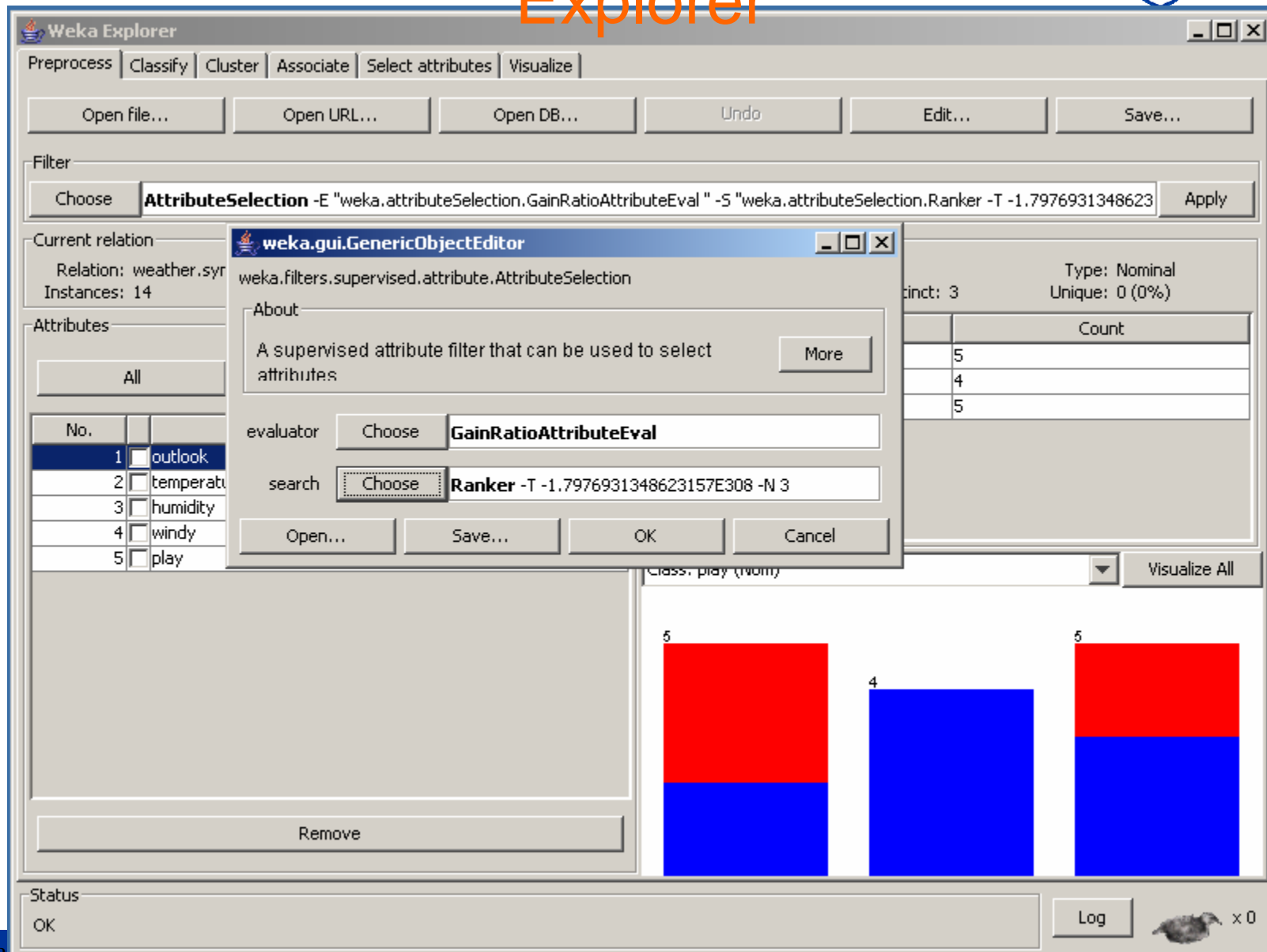
Remove

Class: play (Nom) Visualize All



Status: OK Log  x 0

# Explorer



The screenshot shows the Weka Explorer application window. The 'Filter' tab is active, displaying the configuration for the 'AttributeSelection' filter. The filter is set to use 'GainRatioAttributeEval' as the evaluator and 'Ranker -T -1.7976931348623' as the search method. The 'Current relation' is 'weather.syr' with 14 instances. The 'Attributes' list shows 'outlook', 'temperature', 'humidity', 'windy', and 'play'. The 'outlook' attribute is selected. A 'weka.gui.GenericObjectEditor' dialog box is open, showing the configuration for the 'GainRatioAttributeEval' evaluator and the 'Ranker' search method. The 'Ranker' search method is set to '-T -1.7976931348623157E308 -N 3'. The 'Visualize All' button is visible, and a bar chart is displayed below it, showing the distribution of the 'play' class (nominal) across the 'outlook' attribute values. The bar chart has three bars: the first bar has a total height of 5 (2 blue, 3 red), the second bar has a total height of 4 (all blue), and the third bar has a total height of 5 (3 blue, 2 red).

**Attribute Selection Filter Configuration:**

- Filter: **AttributeSelection** -E "weka.attributeSelection.GainRatioAttributeEval" -S "weka.attributeSelection.Ranker -T -1.7976931348623" Apply
- Current relation: weather.syr (Instances: 14)
- Attributes: All
- Attributes list:
 

No.	Attribute	Selected
1	outlook	<input checked="" type="checkbox"/>
2	temperature	<input type="checkbox"/>
3	humidity	<input type="checkbox"/>
4	windy	<input type="checkbox"/>
5	play	<input type="checkbox"/>
- Filter evaluator: **GainRatioAttributeEval**
- Search: **Ranker -T -1.7976931348623157E308 -N 3**

**Bar Chart Visualization (Class: play (nom)):**

Attribute Value	Count
5	5
4	4
5	5

# 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%)

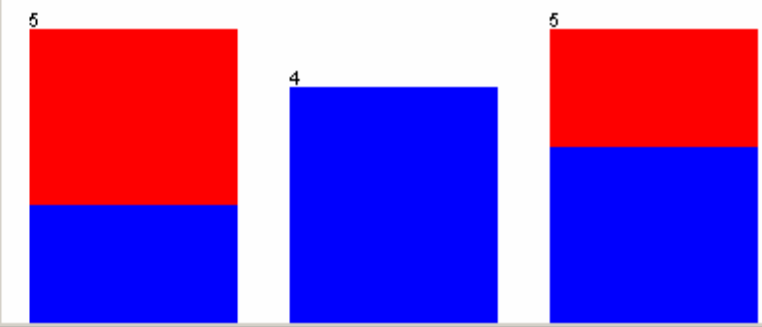
Attributes: All | None | Invert

No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> humidity
3	<input type="checkbox"/> windy
4	<input type="checkbox"/> play

Remove

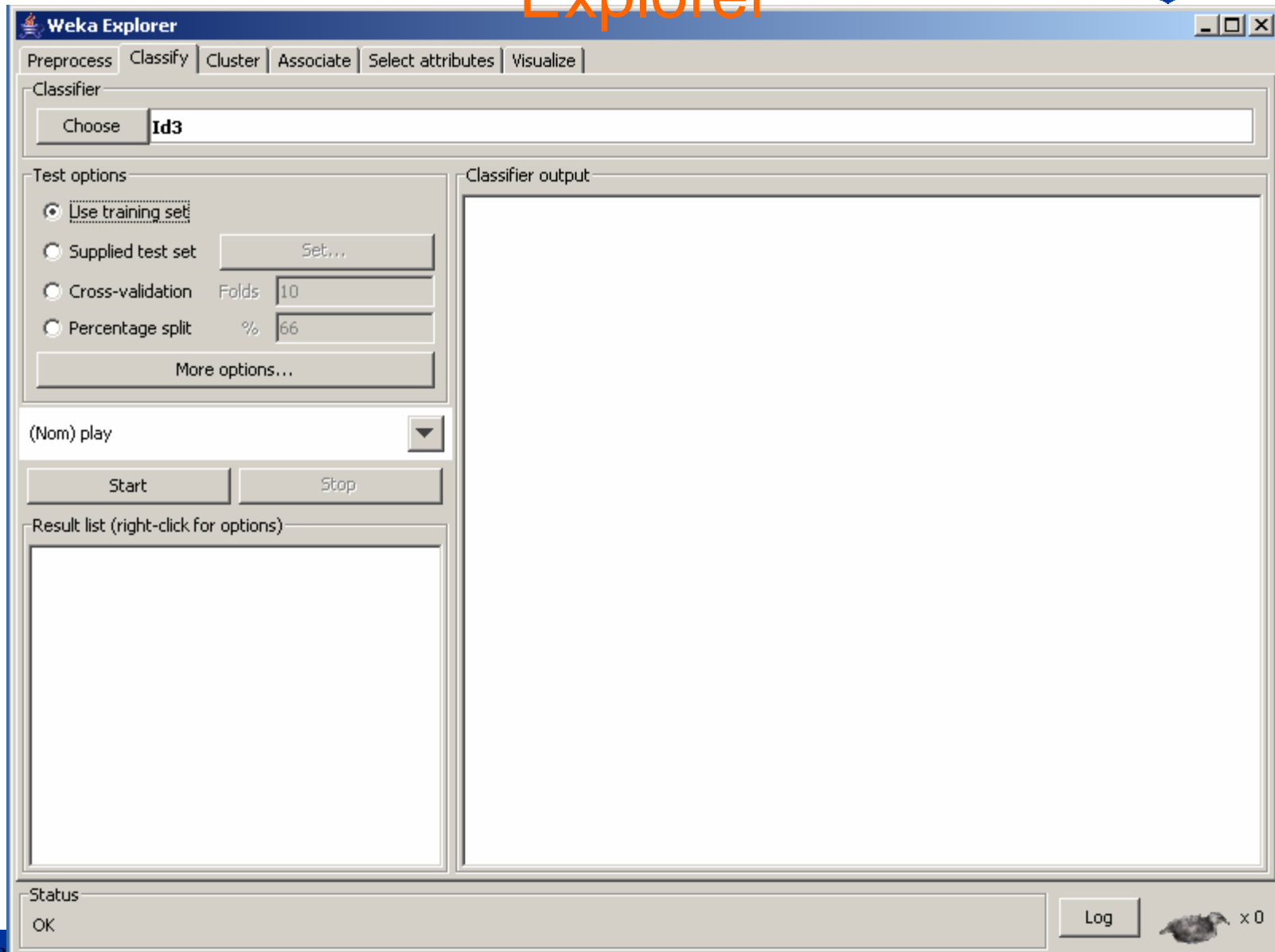
Label	Count
sunny	5
overcast	4
rainy	5

Class: play (Nom) **Visualize All**



Status: OK | Log | x 0

# Explorer





# Explorer

**Weka Explorer**

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

Classifier: Choose **Id3**

Test options:

- Use training set
- Supplied test set
- Cross-validation Folds:
- Percentage split %:

(Nom) play

Result list (right-click for options):

16:00:47 - trees.Id3

Classifier output:

```


Summary
-----
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   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

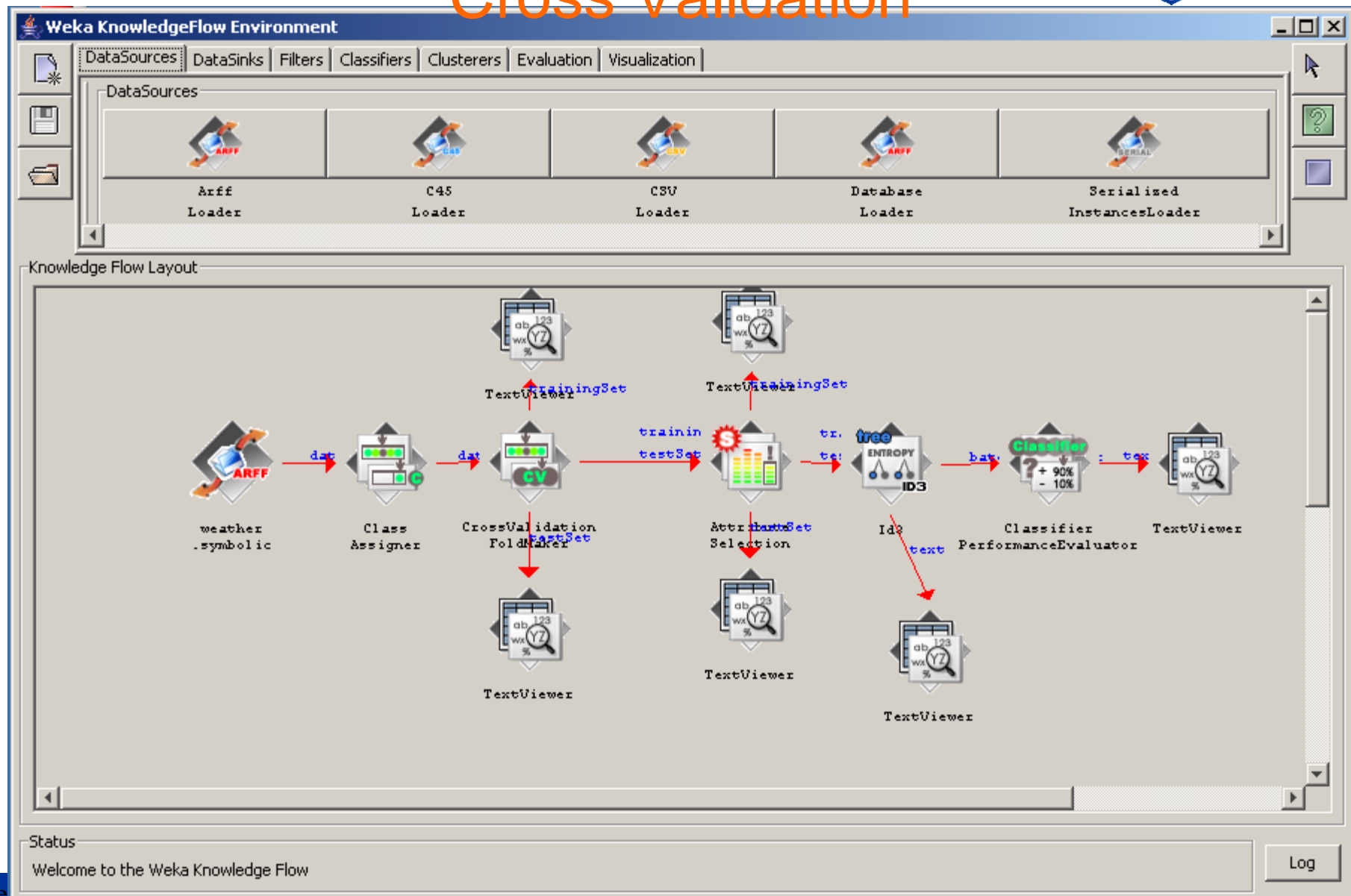
- What is WEKA
- Knowledge Flow
- Explorer
- Why Knowledge Flow
- **Cross Validation**
- Reference



# 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

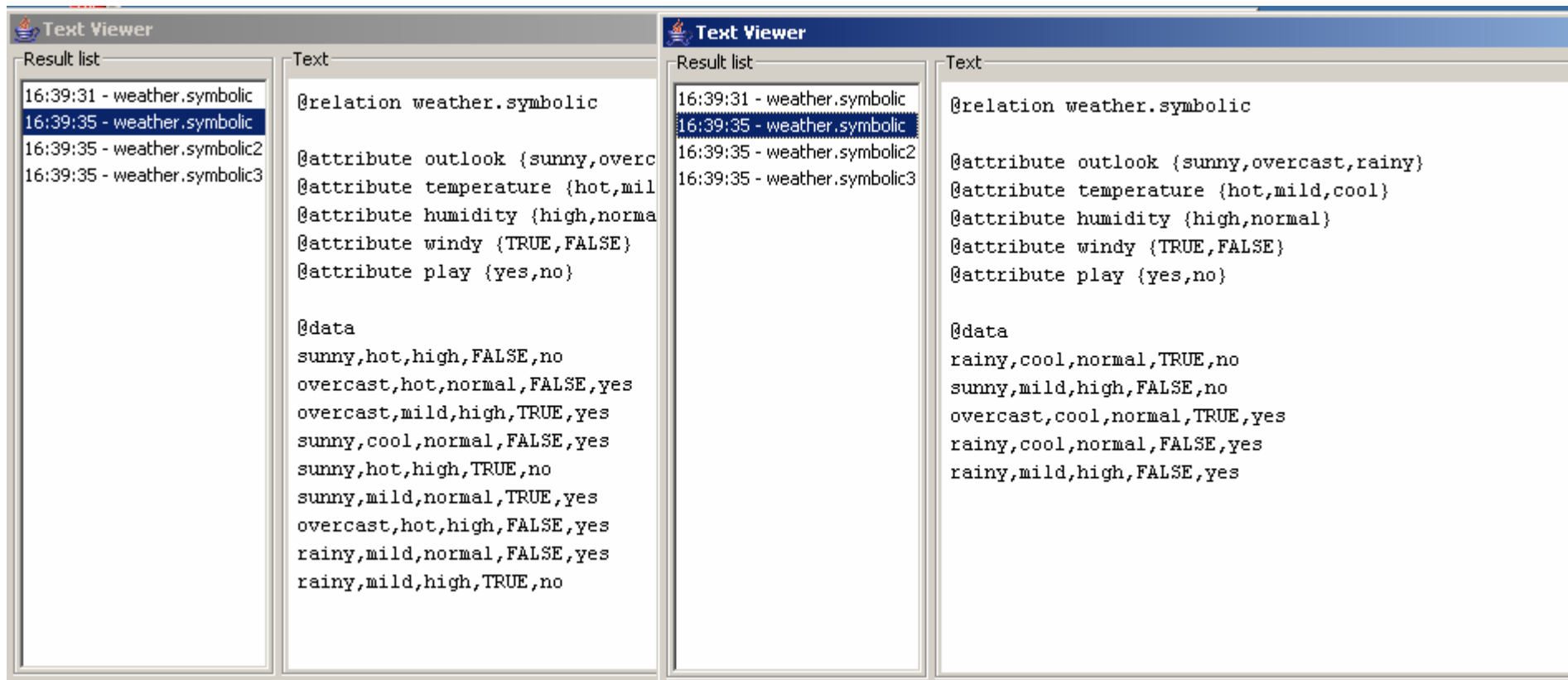
# Cross Validation



## Cross Validation

What do we view in this case?

Text1 VS. Text2 (1)



The image displays two side-by-side screenshots of a 'Text Viewer' application window, illustrating the results of a cross-validation process. Both windows show a 'Result list' and a 'Text' area.

**Left Window (Text1):**

- Result list:**
  - 16:39:31 - weather.symbolic
  - 16:39:35 - weather.symbolic (Selected)
  - 16:39:35 - weather.symbolic2
  - 16:39:35 - weather.symbolic3
- Text:**

```
@relation weather.symbolic

@attribute outlook {sunny,overc
@attribute temperature {hot,mil
@attribute humidity {high,norma
@attribute windy {TRUE,FALSE}
@attribute play {yes,no)

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

**Right Window (Text2):**

- Result list:**
  - 16:39:31 - weather.symbolic
  - 16:39:35 - weather.symbolic (Selected)
  - 16:39:35 - weather.symbolic2
  - 16:39:35 - weather.symbolic3
- Text:**

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

# Cross Validation

## Text1 VS. Text2 (2)

Text Viewer	Text	Text Viewer	Text
Result list 16:39:31 - weather.symbolic 16:39:35 - weather.symbolic 16:39:35 - weather.symbolic2 16:39:35 - weather.symbolic3	<pre> @relation weather.symbolic  @attribute outlook {sunny,overc @attribute temperature {hot,mil @attribute humidity {high,norma @attribute windy {TRUE,FALSE} @attribute play {yes,no}  @data sunny,mild,high,FALSE,no rainy,cool,normal,TRUE,no sunny,mild,normal,TRUE,yes sunny,hot,high,TRUE,no rainy,mild,high,FALSE,yes rainy,mild,normal,FALSE,yes rainy,cool,normal,FALSE,yes overcast,cool,normal,TRUE,yes overcast,hot,normal,FALSE,yes           </pre>	Result list 16:39:31 - weather.symbolic 16:39:35 - weather.symbolic 16:39:35 - weather.symbolic2 16:39:35 - weather.symbolic3	<pre> @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 rainy,mild,high,TRUE,no sunny,hot,high,FALSE,no overcast,hot,high,FALSE,yes sunny,cool,normal,FALSE,yes overcast,mild,high,TRUE,yes           </pre>



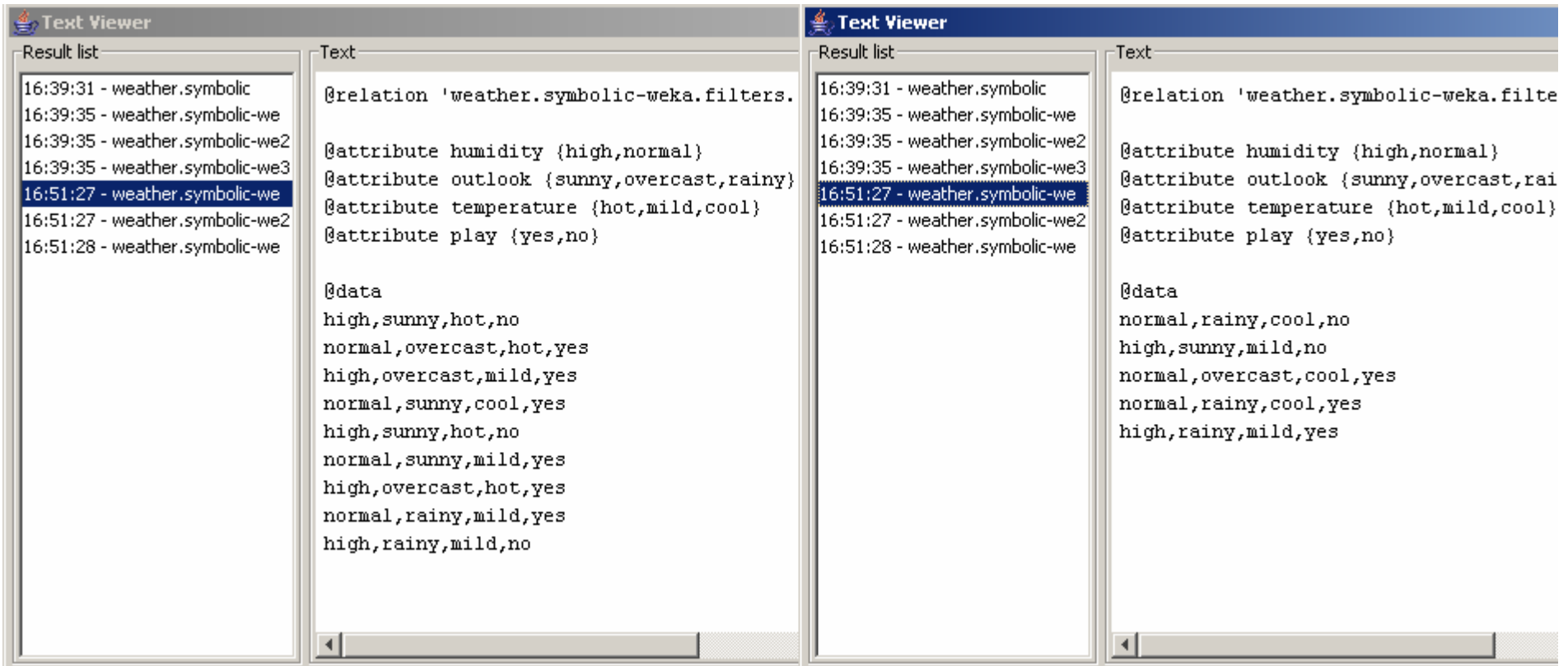
# Cross Validation

## Text1 VS. Text2 (3)

Text Viewer	Text	Text Viewer	Text
Result list 16:39:31 - weather.symbolic 16:39:35 - weather.symbolic 16:39:35 - weather.symbolic2 <b>16:39:35 - weather.symbolic3</b>	<pre> @relation weather.symbolic  @attribute outlook {sunny,overc @attribute temperature {hot,mil @attribute humidity {high,norma @attribute windy {TRUE,FALSE} @attribute play {yes,no}  @data sunny,hot,high,FALSE,no sunny,mild,high,FALSE,no rainy,cool,normal,TRUE,no rainy,mild,high,TRUE,no rainy,cool,normal,FALSE,yes sunny,cool,normal,FALSE,yes overcast,mild,high,TRUE,yes rainy,mild,high,FALSE,yes overcast,hot,high,FALSE,yes overcast,cool,normal,TRUE,yes           </pre>	Result list 16:39:31 - weather.symbolic 16:39:35 - weather.symbolic 16:39:35 - weather.symbolic2 <b>16:39:35 - weather.symbolic3</b>	<pre> @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,TRUE,no sunny,mild,normal,TRUE,yes rainy,mild,normal,FALSE,yes overcast,hot,normal,FALSE,yes           </pre>

# Cross Validation

## Text3 VS. Text4 (1)



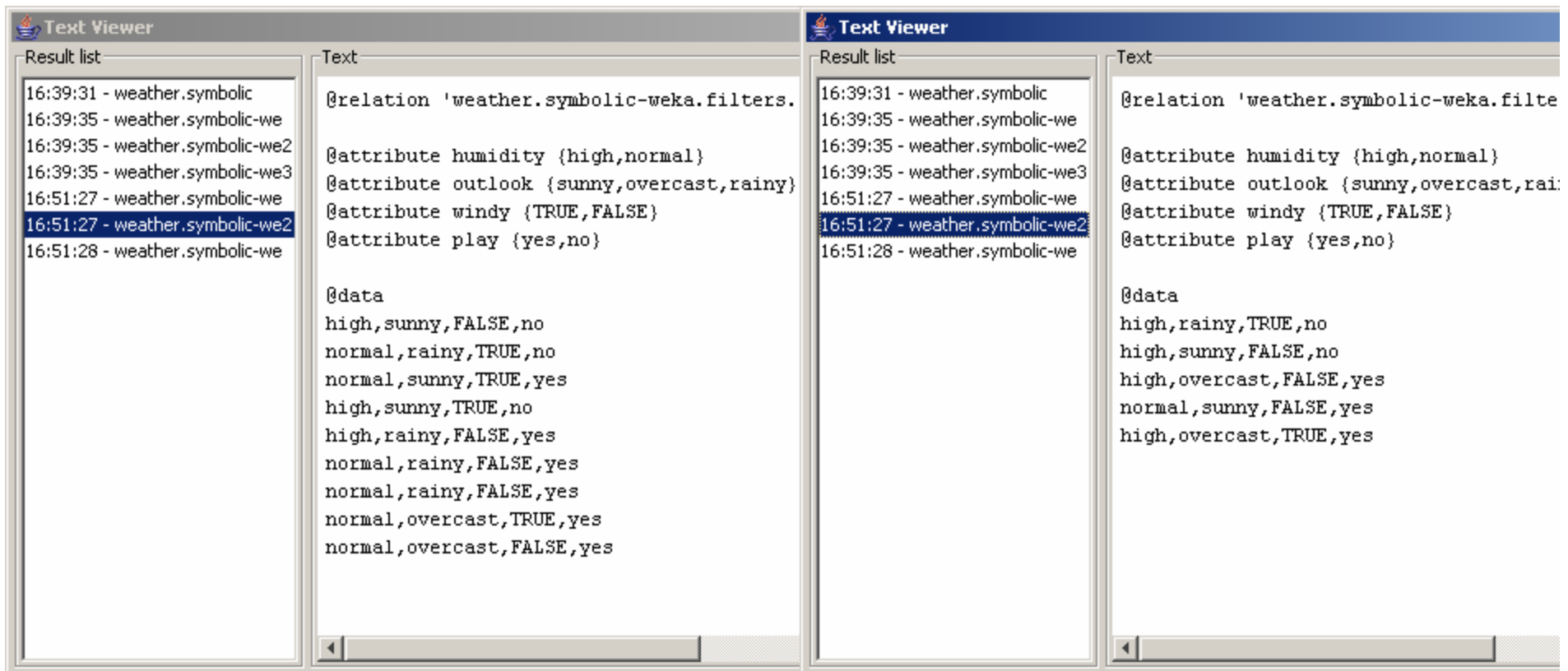
Result list	Text
16:39:31 - weather.symbolic	@relation 'weather.symbolic-weka.filters.
16:39:35 - weather.symbolic-we	
16:39:35 - weather.symbolic-we2	@attribute humidity {high,normal}
16:39:35 - weather.symbolic-we3	@attribute outlook {sunny,overcast,rainy}
16:51:27 - weather.symbolic-we	@attribute temperature {hot,mild,cool}
16:51:27 - weather.symbolic-we2	@attribute play {yes,no}
16:51:28 - weather.symbolic-we	
	@data
	high,sunny,hot,no
	normal,overcast,hot,yes
	high,overcast,mild,yes
	normal,sunny,cool,yes
	high,sunny,hot,no
	normal,sunny,mild,yes
	high,overcast,hot,yes
	normal,rainy,mild,yes
	high,rainy,mild,no

Result list	Text
16:39:31 - weather.symbolic	@relation 'weather.symbolic-weka.filte
16:39:35 - weather.symbolic-we	
16:39:35 - weather.symbolic-we2	@attribute humidity {high,normal}
16:39:35 - weather.symbolic-we3	@attribute outlook {sunny,overcast,rai
16:51:27 - weather.symbolic-we	@attribute temperature {hot,mild,cool}
16:51:27 - weather.symbolic-we2	@attribute play {yes,no}
16:51:28 - weather.symbolic-we	
	@data
	normal,rainy,cool,no
	high,sunny,mild,no
	normal,overcast,cool,yes
	normal,rainy,cool,yes
	high,rainy,mild,yes

# Cross Validation

## Text3 VS. Text4 (2)



Result list	Text
16:39:31 - weather.symbolic	@relation 'weather.symbolic-weka.filters.
16:39:35 - weather.symbolic-we	@attribute humidity {high,normal}
16:39:35 - weather.symbolic-we2	@attribute outlook {sunny,overcast,rainy}
16:39:35 - weather.symbolic-we3	@attribute windy {TRUE,FALSE}
16:51:27 - weather.symbolic-we	@attribute play {yes,no}
<b>16:51:27 - weather.symbolic-we2</b>	@data
16:51:28 - weather.symbolic-we	high,sunny,FALSE,no
	normal,rainy,TRUE,no
	normal,sunny,TRUE,yes
	high,sunny,TRUE,no
	high,rainy,FALSE,yes
	normal,rainy,FALSE,yes
	normal,rainy,FALSE,yes
	normal,overcast,TRUE,yes
	normal,overcast,FALSE,yes

Result list	Text
16:39:31 - weather.symbolic	@relation 'weather.symbolic-weka.filte
16:39:35 - weather.symbolic-we	@attribute humidity {high,normal}
16:39:35 - weather.symbolic-we2	@attribute outlook {sunny,overcast,rainy}
16:39:35 - weather.symbolic-we3	@attribute windy {TRUE,FALSE}
16:51:27 - weather.symbolic-we	@attribute play {yes,no}
<b>16:51:27 - weather.symbolic-we2</b>	@data
16:51:28 - weather.symbolic-we	high,rainy,TRUE,no
	high,sunny,FALSE,no
	high,overcast,FALSE,yes
	normal,sunny,FALSE,yes
	high,overcast,TRUE,yes

# Cross Validation

## Text3 VS. Text4 (3)

Text Viewer	Text	Text Viewer	Text
Result list 16:39:31 - weather.symbolic 16:39:35 - weather.symbolic-we 16:39:35 - weather.symbolic-we2 16:39:35 - weather.symbolic-we3 16:51:27 - weather.symbolic-we 16:51:27 - weather.symbolic-we2 16:51:28 - weather.symbolic-we	<pre> @relation 'weather.symbolic-weka.filters.  @attribute outlook {sunny,overcast,rainy} @attribute humidity {high,normal} @attribute temperature {hot,mild,cool} @attribute play {yes,no}  @data sunny,high,hot,no sunny,high,mild,no rainy,normal,cool,no rainy,high,mild,no rainy,normal,cool,yes sunny,normal,cool,yes overcast,high,mild,yes rainy,high,mild,yes overcast,high,hot,yes overcast,normal,cool,yes           </pre>	Result list 16:39:31 - weather.symbolic 16:39:35 - weather.symbolic-we 16:39:35 - weather.symbolic-we2 16:39:35 - weather.symbolic-we3 16:51:27 - weather.symbolic-we 16:51:27 - weather.symbolic-we2 16:51:28 - weather.symbolic-we	<pre> @relation 'weather.symbolic-weka.filte @attribute outlook {sunny,overcast,rainy} @attribute humidity {high,normal} @attribute temperature {hot,mild,cool} @attribute play {yes,no}  @data sunny,high,hot,no sunny,normal,mild,yes rainy,normal,mild,yes overcast,normal,hot,yes           </pre>

# Cross Validation

## Trees

16:51:27 - Model: Id3	16:51:27 - Model: Id32	16:51:28 - Model: Id3
<pre> === Classifier model ===  Scheme: Id3 Relation: weather.symbolic- Training Fold: 1  Id3  humidity = high   outlook = sunny: no   outlook = overcast: yes   outlook = rainy: no humidity = normal: yes </pre>	<pre> === Classifier model ===  Scheme: Id3 Relation: weather.symbolic- Training Fold: 2  Id3  outlook = sunny   humidity = high: no   humidity = normal: yes outlook = overcast: yes outlook = rainy   windy = TRUE: no   windy = FALSE: yes </pre>	<pre> === Classifier model ===  Scheme: Id3 Relation: weather.symbolic-weka.filters.supervised.attribute.A Training Fold: 3  Id3  outlook = sunny   humidity = high: no   humidity = normal: yes outlook = overcast: yes outlook = rainy: yes </pre>

# 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



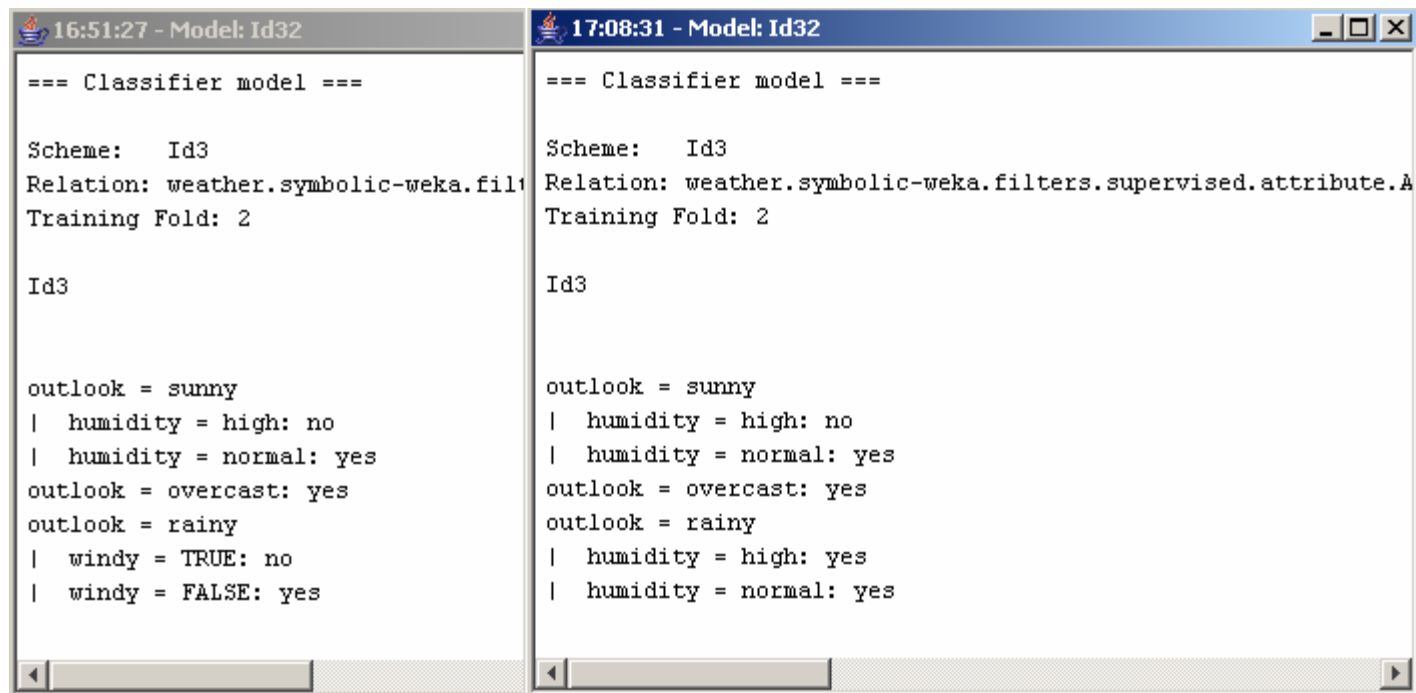
# Cross Validation

- **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



```
16:51:27 - Model: Id32
=== Classifier model ===

Scheme: Id3
Relation: weather.symbolic-weka.filters.supervised.attribute.A
Training Fold: 2

Id3

outlook = sunny
| humidity = high: no
| humidity = normal: yes
outlook = overcast: yes
outlook = rainy
| windy = TRUE: no
| windy = FALSE: yes

17:08:31 - Model: Id32
=== Classifier model ===

Scheme: Id3
Relation: weather.symbolic-weka.filters.supervised.attribute.A
Training Fold: 2

Id3

outlook = sunny
| humidity = high: no
| humidity = normal: yes
outlook = overcast: yes
outlook = rainy
| humidity = high: yes
| humidity = normal: yes
```



## 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)*