

WEKA: Evaluation. Knowledge flow

Lab 4

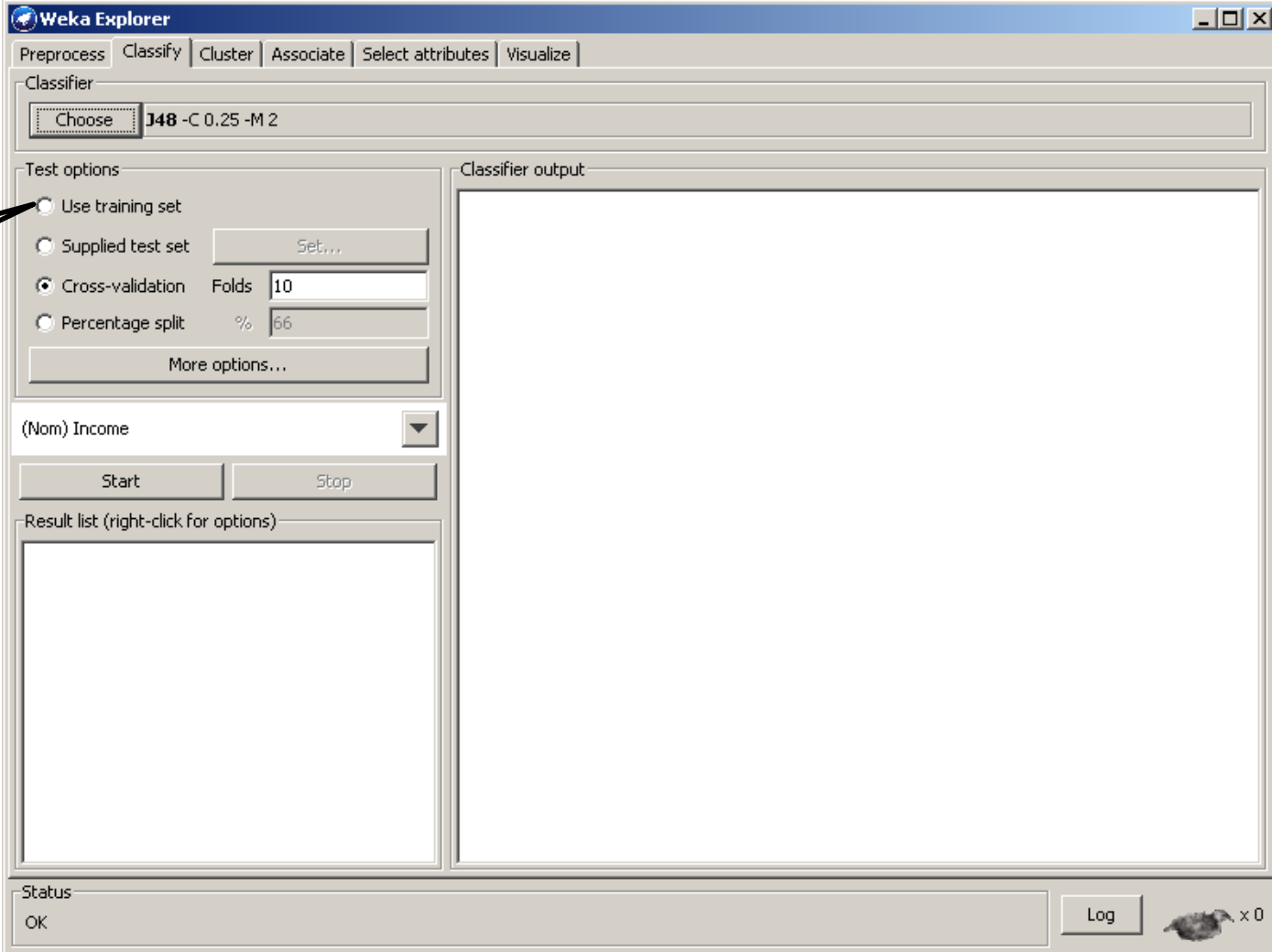
Lab outline

- Evaluation metrics in WEKA Explorer
- Knowledge flow interface
- Generating ROC curves in Knowledge flow interface

WEKA: evaluation metrics

- Open WEKA
- Open file “adult_income.arff”

Evaluation options



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier is set to 'J48 -C 0.25 -M 2'. The 'Test options' section is highlighted with a callout box containing a question mark. The 'Test options' section includes four radio buttons: 'Use training set', 'Supplied test set', 'Cross-validation', and 'Percentage split'. The 'Cross-validation' option is selected, with 'Folds' set to 10 and 'Percentage split' set to 66. Below the radio buttons is a 'More options...' button. The 'Classifier output' area is empty. The 'Result list' area is also empty. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

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

Classifier

Choose J48 -C 0.25 -M 2


Test options

- Use training set
- Supplied test set
- Cross-validation Folds 10
- Percentage split % 66

(Nom) Income

Result list (right-click for options)

Status

OK  x 0

The best possible accuracy

The screenshot shows the Weka Explorer interface. The classifier selected is J48 -C 0.25 -M 2. The test options are set to 'Use training set'. The classifier output window displays the following summary:

```
=== Evaluation on training set ===  
=== Summary ===  
Correctly Classified Instances      27427      84.2327 %  
Incorrectly Classified Instances    5134      15.7673 %  
Kappa statistic                    0.5393  
Mean absolute error                 0.235  
Root mean squared error             0.3424  
Relative absolute error             64.2706 %  
Root relative squared error         80.0764 %  
Total Number of Instances          32561
```

The value 84.2327 % is circled in red. Below the summary is a detailed accuracy by class table:

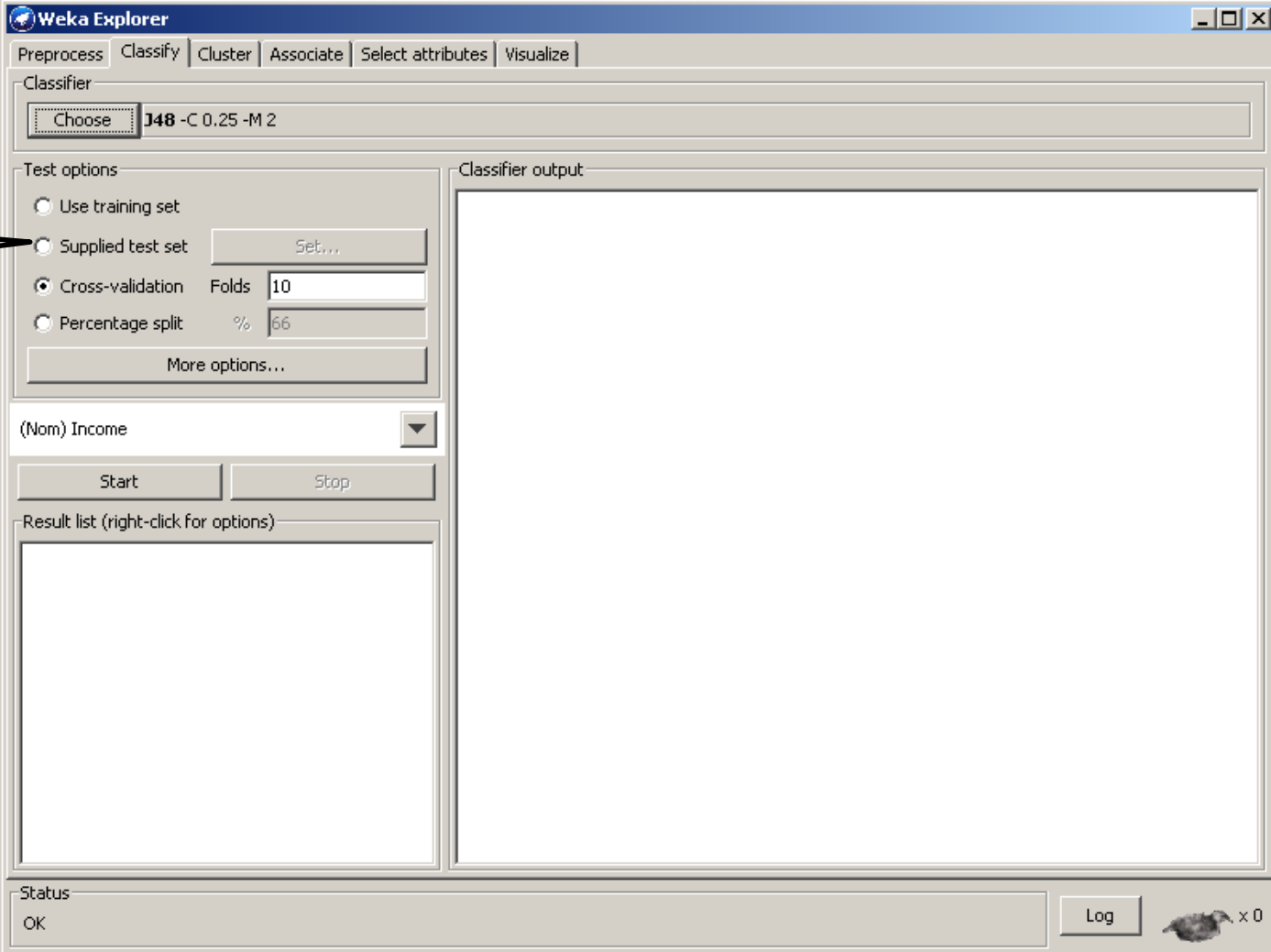
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.926	0.421	0.874	0.926	0.899	0.854	<=50K
	0.579	0.074	0.712	0.579	0.639	0.854	>50K
Weighted Avg.	0.842	0.338	0.835	0.842	0.836	0.854	

Below the table is the confusion matrix:

```
=== Confusion Matrix ===  
  
a    b  <-- classified as  
22888 1832 | a = <=50K  
3302  4539 | b = >50K
```

The status bar at the bottom shows 'OK' and a 'Log' button.

Evaluation options

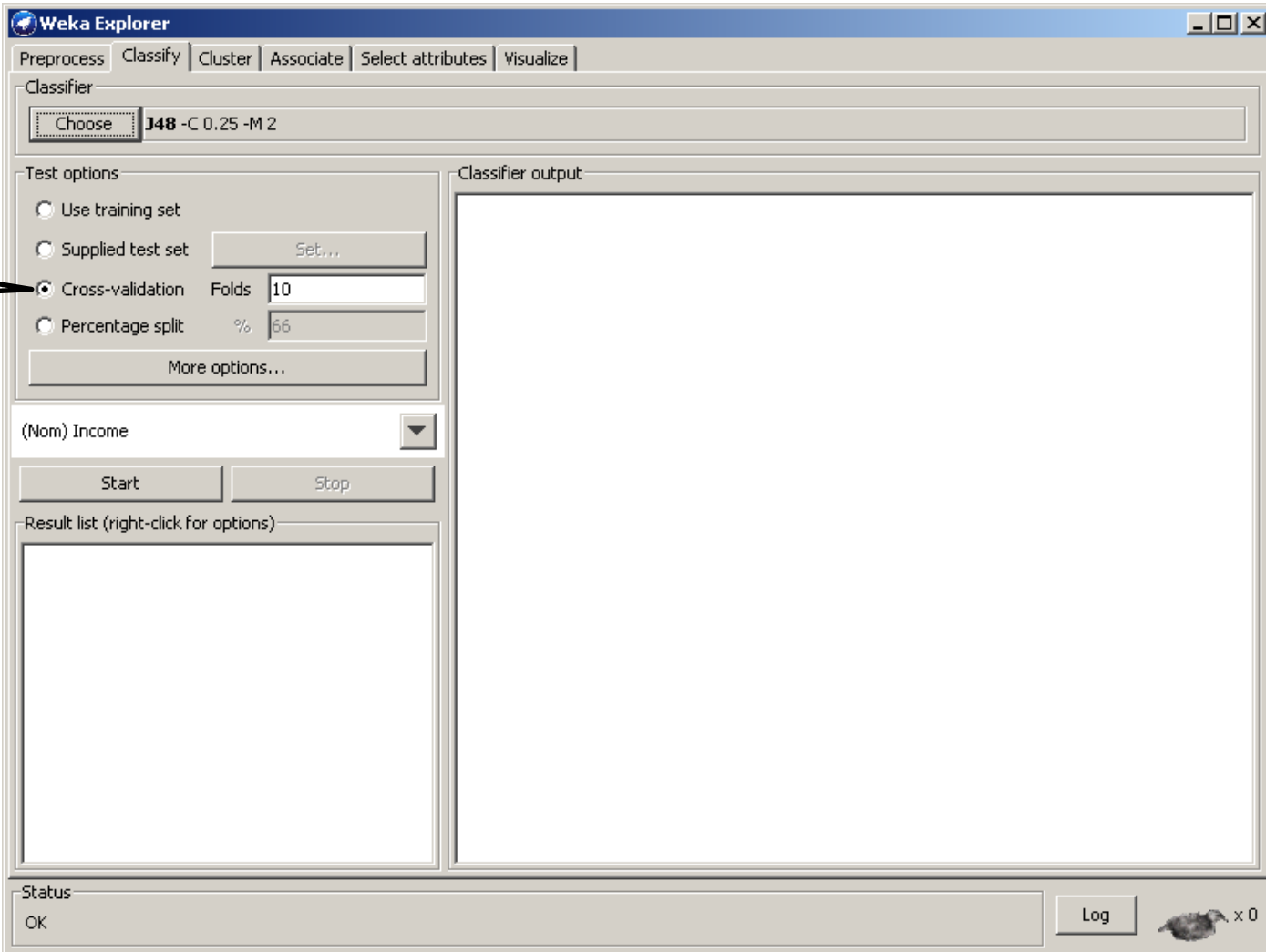


The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier is set to 'J48 -C 0.25 -M 2'. The 'Test options' section is highlighted with a callout box containing a question mark. The 'Test options' section includes the following options:

- Use training set
- Supplied test set (with a 'Set...' button)
- Cross-validation (with 'Folds' set to 10)
- Percentage split (with '%' set to 66)

Below the 'Test options' section, there is a dropdown menu for '(Nom) Income', 'Start' and 'Stop' buttons, and a 'Result list (right-click for options)' area. The status bar at the bottom shows 'OK' and a 'Log' button.

Evaluation options



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. Under the 'Test options' section, the 'Cross-validation' radio button is selected, with 'Folds' set to 10. A callout box with a question mark points to this radio button. Other options include 'Use training set', 'Supplied test set' (with a 'Set...' button), and 'Percentage split' (with a '%' sign and '66'). Below the test options are 'Start' and 'Stop' buttons. The 'Classifier output' area is empty. The 'Result list' area is also empty. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

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

Classifier

Choose J48 -C 0.25 -M 2


Test options

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

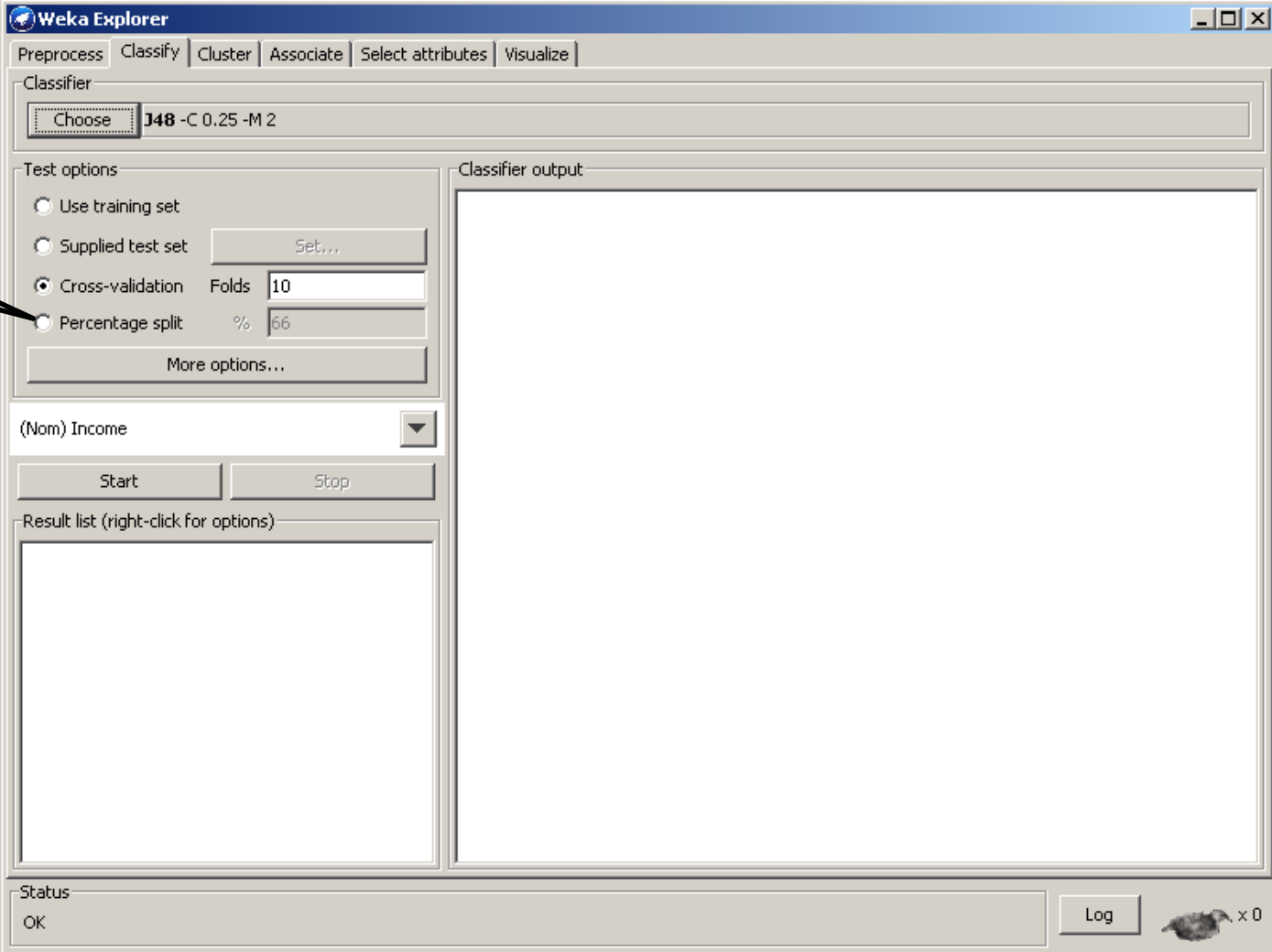
(Nom) Income

Result list (right-click for options)

Status

OK  x 0

Evaluation options



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. The 'Test options' section is highlighted with a callout box containing a question mark. The options are:

- Use training set
- Supplied test set (with a 'Set...' button)
- Cross-validation (with 'Folds' set to 10)
- Percentage split (with '%' set to 66)

Below these options is a 'More options...' button. The 'Classifier output' area is empty. The 'Result list' area is also empty. The status bar at the bottom shows 'OK' and a 'Log' button.

Build classifier: output



```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431           16.6795 %
Kappa statistic                    0.5107
Mean absolute error                 0.2409
Root mean squared error            0.3505
Relative absolute error            65.8833 %
Root relative squared error        81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.922   0.445    0.867     0.922   0.893     0.841    <=50K
                0.555   0.078    0.692     0.555   0.616     0.841    >50K
Weighted Avg.   0.833   0.357    0.825     0.833   0.827     0.841

=== Confusion Matrix ===

  a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K
```

Build classifier: output



```
Classifier output

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      27130      83.3205 %
Incorrectly Classified Instances    5431      16.6795 %
Kappa statistic                     0.5107
Mean absolute error                  0.2409
Root mean squared error              0.3505
Relative absolute error              65.8833 %
Root relative squared error          81.9782 %
Total Number of Instances           32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.922   0.445    0.867     0.922   0.893     0.841    <=50K
                0.555   0.078    0.692     0.555   0.616     0.841    >50K
Weighted Avg.   0.833   0.357    0.825     0.833   0.827     0.841

=== Confusion Matrix ===

  a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K
```

Vs. 84.23%

Build classifier: output

Your prediction is better than random prediction by 51%

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431            16.6795 %
Kappa statistic                     0.5107
Mean absolute error                  0.2409
Root mean squared error              0.3505
Relative absolute error              65.8833 %
Root relative squared error          81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.922   0.445    0.867     0.922    0.893     0.841    <=50K
                0.555   0.078    0.692     0.555    0.616     0.841    >50K
Weighted Avg.   0.833   0.357    0.825     0.833    0.827     0.841

=== Confusion Matrix ===

  a    b  <-- classified as
22782 1938 |   a = <=50K
 3493 4348 |   b = >50K
```

Build classifier: output

Some
per/instance
metrics

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431            16.6795 %
Kappa statistic                     0.5107
Mean absolute error                  0.2409
Root mean squared error              0.3505
Relative absolute error              65.8833 %
Root relative squared error          81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.922   0.445    0.867     0.922   0.893     0.841   <=50K
                0.555   0.078    0.692     0.555   0.616     0.841   >50K
Weighted Avg.   0.833   0.357    0.825     0.833   0.827     0.841

=== Confusion Matrix ===

  a    b  <-- classified as
22782 1938 |   a = <=50K
 3493 4348 |   b = >50K
```

Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

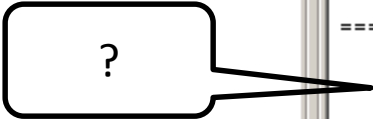
Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431            16.6795 %
Kappa statistic                     0.5107
Mean absolute error                  0.2409
Root mean squared error              0.3505
Relative absolute error              65.8833 %
Root relative squared error          81.9782 %
Total Number of Instances           32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.922   0.445    0.867     0.922    0.893     0.841    <=50K
                0.555   0.078    0.692     0.555    0.616     0.841    >50K
Weighted Avg.   0.833   0.357    0.825     0.833    0.827     0.841

=== Confusion Matrix ===

  a    b  <-- classified as
22782 1938 |   a = <=50K
 3493 4348 |   b = >50K
```



Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431           16.6795 %
Kappa statistic                    0.5107
Mean absolute error                 0.2409
Root mean squared error             0.3505
Relative absolute error             65.8833 %
Root relative squared error         81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.922	0.445	0.867	0.922	0.893	0.841	<=50K
	0.555	0.078	0.692	0.555	0.616	0.841	>50K
Weighted Avg.	0.833	0.357	0.825	0.833	0.827	0.841	

```

=== Confusion Matrix ===
      a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K

```

TPos/Pos

Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431            16.6795 %
Kappa statistic                     0.5107
Mean absolute error                  0.2409
Root mean squared error              0.3505
Relative absolute error              65.8833 %
Root relative squared error          81.9782 %
Total Number of Instances           32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.922   0.445    0.867     0.922    0.893     0.841    <=50K
                0.555   0.078    0.692     0.555    0.616     0.841    >50K
Weighted Avg.   0.833   0.357    0.825     0.833    0.827     0.841

=== Confusion Matrix ===

      a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K
```

FPos/Neg

Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431           16.6795 %
Kappa statistic                    0.5107
Mean absolute error                 0.2409
Root mean squared error             0.3505
Relative absolute error             65.8833 %
Root relative squared error        81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
                0.922     0.445     0.867       0.922     0.893       0.841     <=50K
                0.555     0.078     0.692       0.555     0.616       0.841     >50K
Weighted Avg.   0.833     0.357     0.825       0.833     0.827       0.841

=== Confusion Matrix ===

      a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K
```

$TPos / (TPos + Fpos)$

Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431            16.6795 %
Kappa statistic                    0.5107
Mean absolute error                 0.2409
Root mean squared error            0.3505
Relative absolute error             65.8833 %
Root relative squared error        81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
                0.922     0.445     0.867     0.922     0.893     0.841     <=50K
                0.555     0.078     0.692     0.555     0.616     0.841     >50K
Weighted Avg.   0.833     0.357     0.825     0.833     0.827     0.841

=== Confusion Matrix ===

      a      b  <-- classified as
22782 1938 |   a = <=50K
 3493 4348 |   b = >50K
```

TP Rate

Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431           16.6795 %
Kappa statistic                    0.5107
Mean absolute error                 0.2409
Root mean squared error             0.3505
Relative absolute error             65.8833 %
Root relative squared error         81.9782 %
Total Number of Instances          32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
<=50K          0.922    0.445    0.867     0.922    0.893     0.841    <=50K
>50K           0.555    0.078    0.692     0.555    0.616     0.841    >50K
Weighted Avg.   0.833    0.357    0.825     0.833    0.827     0.841

=== Confusion Matrix ===

      a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K
```

$2 * \text{precision} * \text{recall}$
precision + recall

Build classifier: output

```
Classifier output

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      27130           83.3205 %
Incorrectly Classified Instances    5431            16.6795 %
Kappa statistic                     0.5107
Mean absolute error                  0.2409
Root mean squared error              0.3505
Relative absolute error              65.8833 %
Root relative squared error          81.9782 %
Total Number of Instances           32561

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                -----  -----  -
                0.922    0.443    0.867     0.922    0.893     0.841    <=50K
                0.555    0.078    0.692     0.555    0.616     0.841    >50K
Weighted Avg.   0.833    0.357    0.825     0.833    0.827     0.841

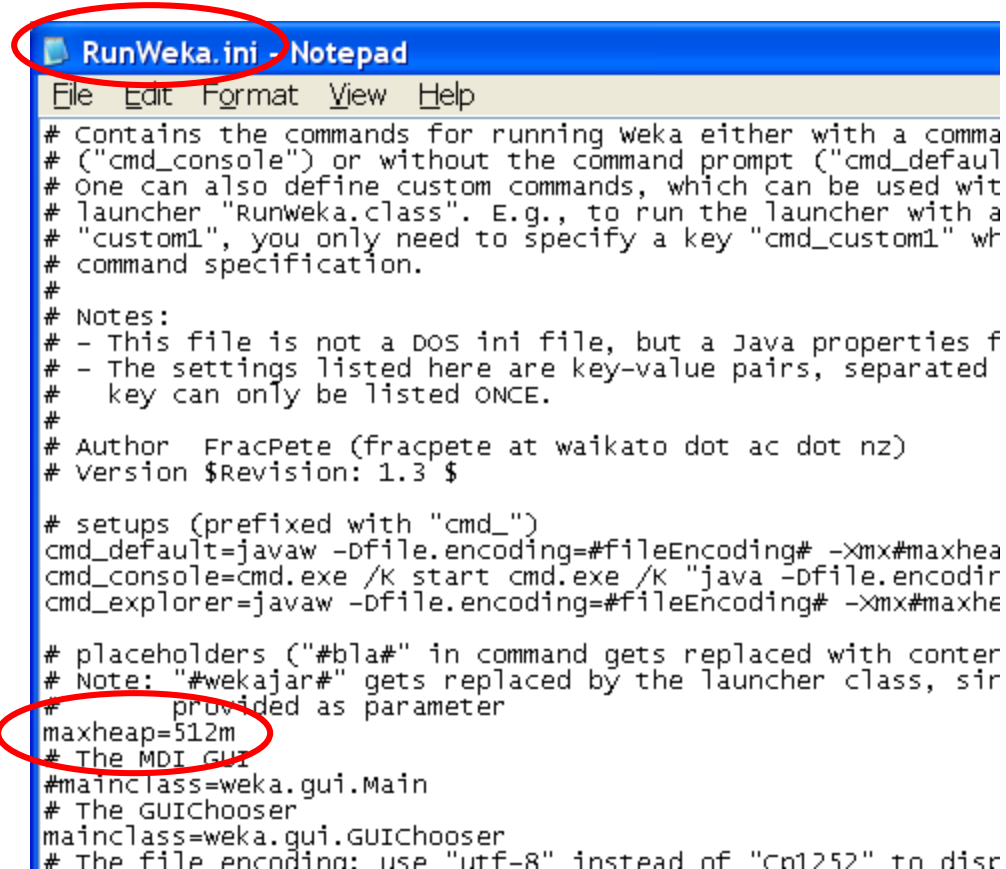
=== Confusion Matrix ===

      a    b  <-- classified as
22782 1938 |    a = <=50K
 3493 4348 |    b = >50K
```

Area under the ROC curve

WEKA: dealing with large datasets

- Increase java heap space
- Still might get “Out of memory” exception



```
RunWeka.ini - Notepad
File Edit Format View Help
# Contains the commands for running weka either with a comma
# ("cmd_console") or without the command prompt ("cmd_defaul
# One can also define custom commands, which can be used wit
# launcher "Runweka.class". E.g., to run the launcher with a
# "custom1", you only need to specify a key "cmd_custom1" wr
# command specification.
#
# Notes:
# - This file is not a DOS ini file, but a Java properties f
# - The settings listed here are key-value pairs, separated
#   key can only be listed ONCE.
#
# Author  FracPete (fracpete at waikato dot ac dot nz)
# Version $Revision: 1.3 $
#
# setups (prefixed with "cmd_")
cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxhea
cmd_console=cmd.exe /K start cmd.exe /K "java -Dfile.encodir
cmd_explorer=javaw -Dfile.encoding=#fileEncoding# -Xmx#maxhe
#
# placeholders ("#bla#" in command gets replaced with conter
# Note: "#wekajar#" gets replaced by the launcher class, sir
#   provided as parameter
maxheap=512m
# The MDI GUI
#mainclass=weka.gui.Main
# The GUIChooser
mainclass=weka.gui.GUIChooser
# The file encoding: use "utf-8" instead of "Cn1252" to disc
```

GUI I: WEKA Explorer and CLI

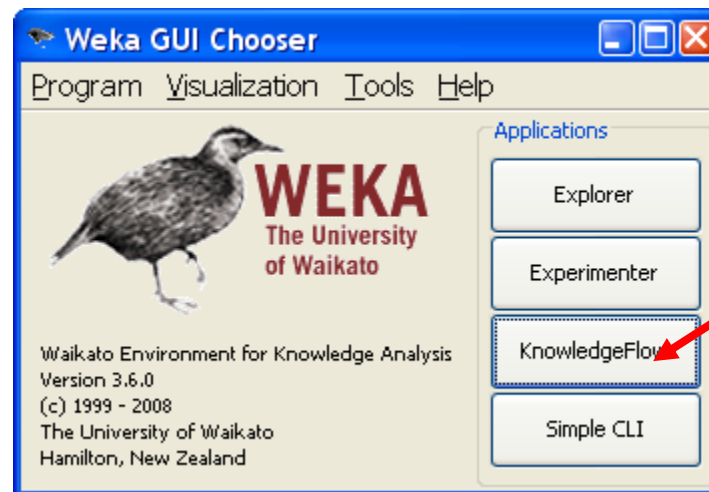
- Everything is in main memory: dataset, filter, model
- No large-scale data mining

GUI II. WEKA Knowledge Flow

- Design configuration for streamed data processing
- Specify data stream and run algorithms which stream data from one component to another
- If the algorithm allows incremental filtering and learning, data will be loaded sequentially from disk

Comparing classifiers.

Knowledge flow



Knowledge flow tabs

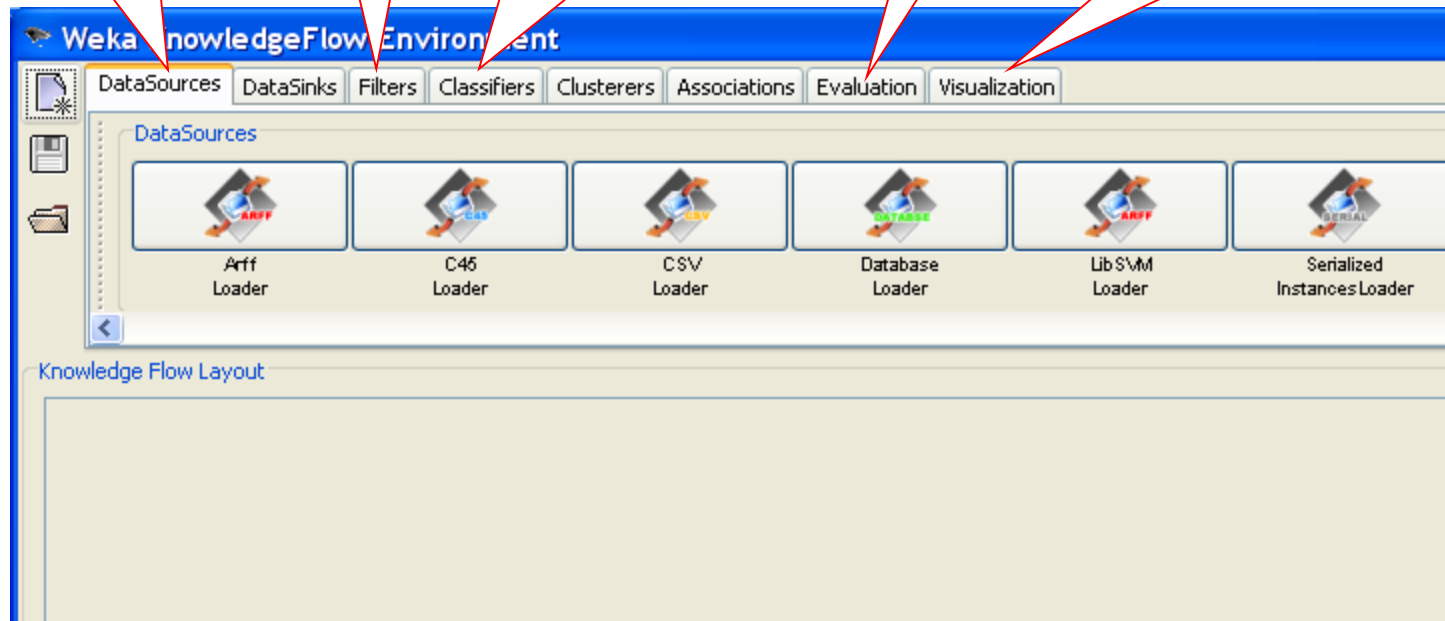
DATA
SOURCES

FILTERS

CLASSIFIERS

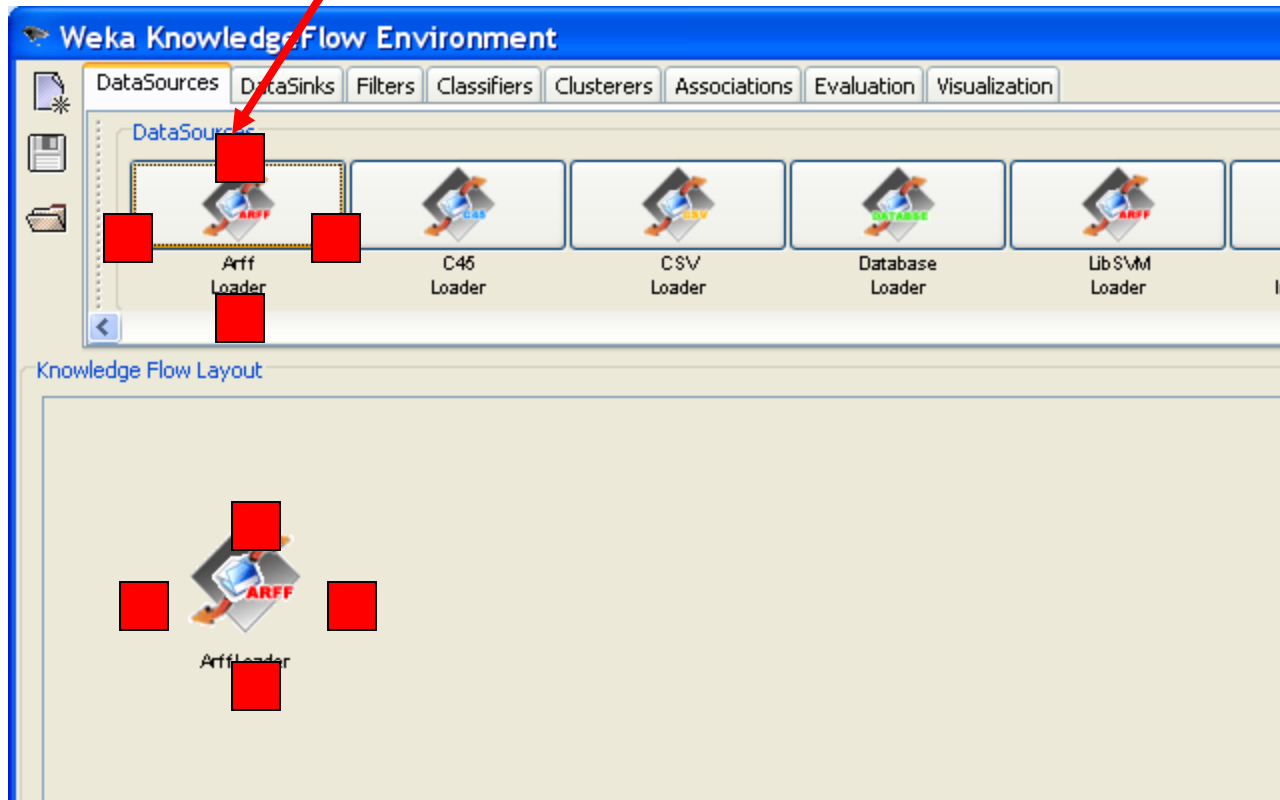
EVALUATION

VISUALIZATION

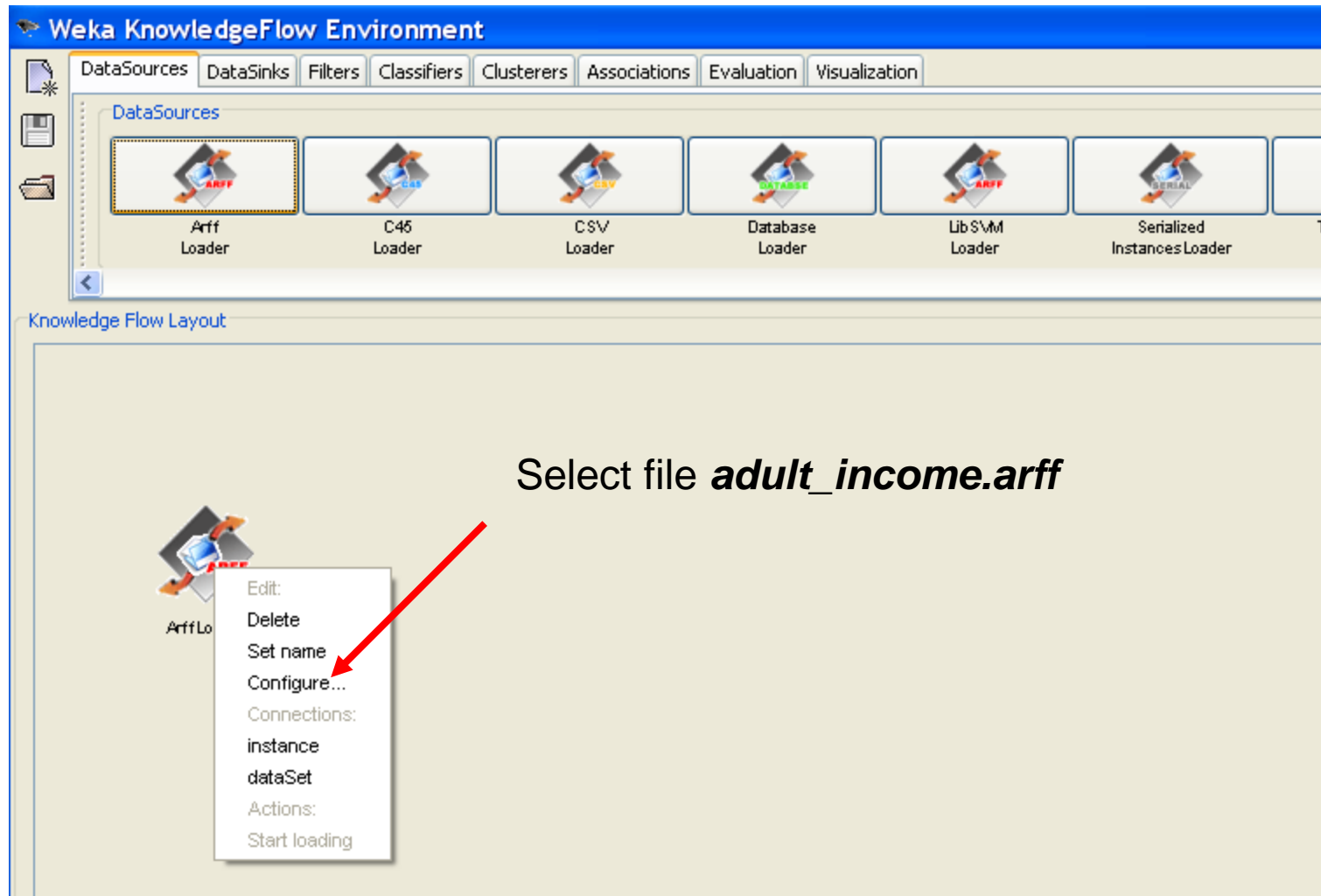


Loading the data

Click



Loading the data



Attributes of interest:

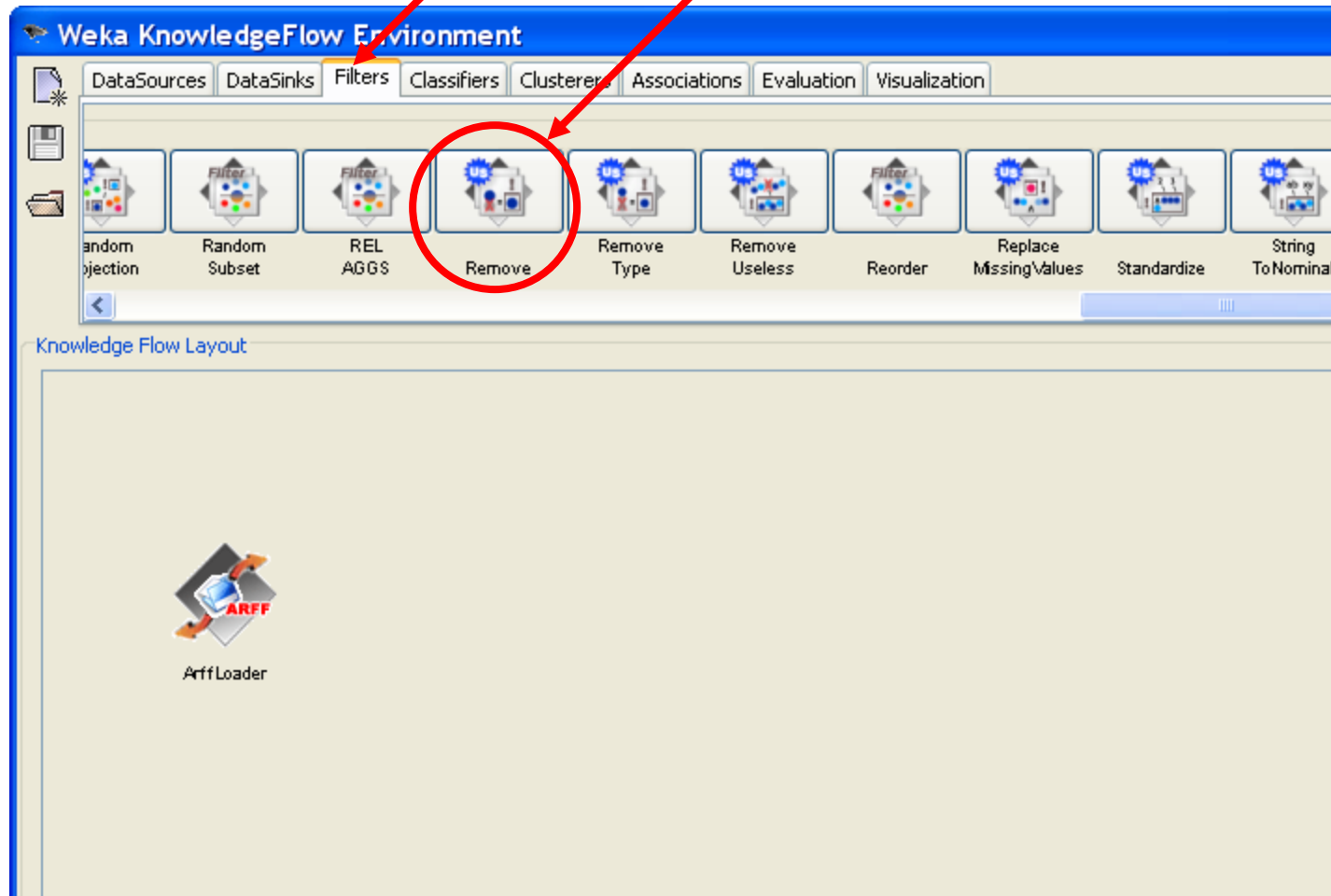
age, education,
class (income >50 K: YES,NO)

1. @attribute Age numeric
 3. @attribute Education {Preschool,1st-4th,5th-6th,7th-8th,9th,10th,11th,12th,Prof-school,HS-grad,Some-college,Assoc-voc,Assoc-acdm,Bachelors,Masters,Doctorate}
- last @attribute class {>50K, <=50K}

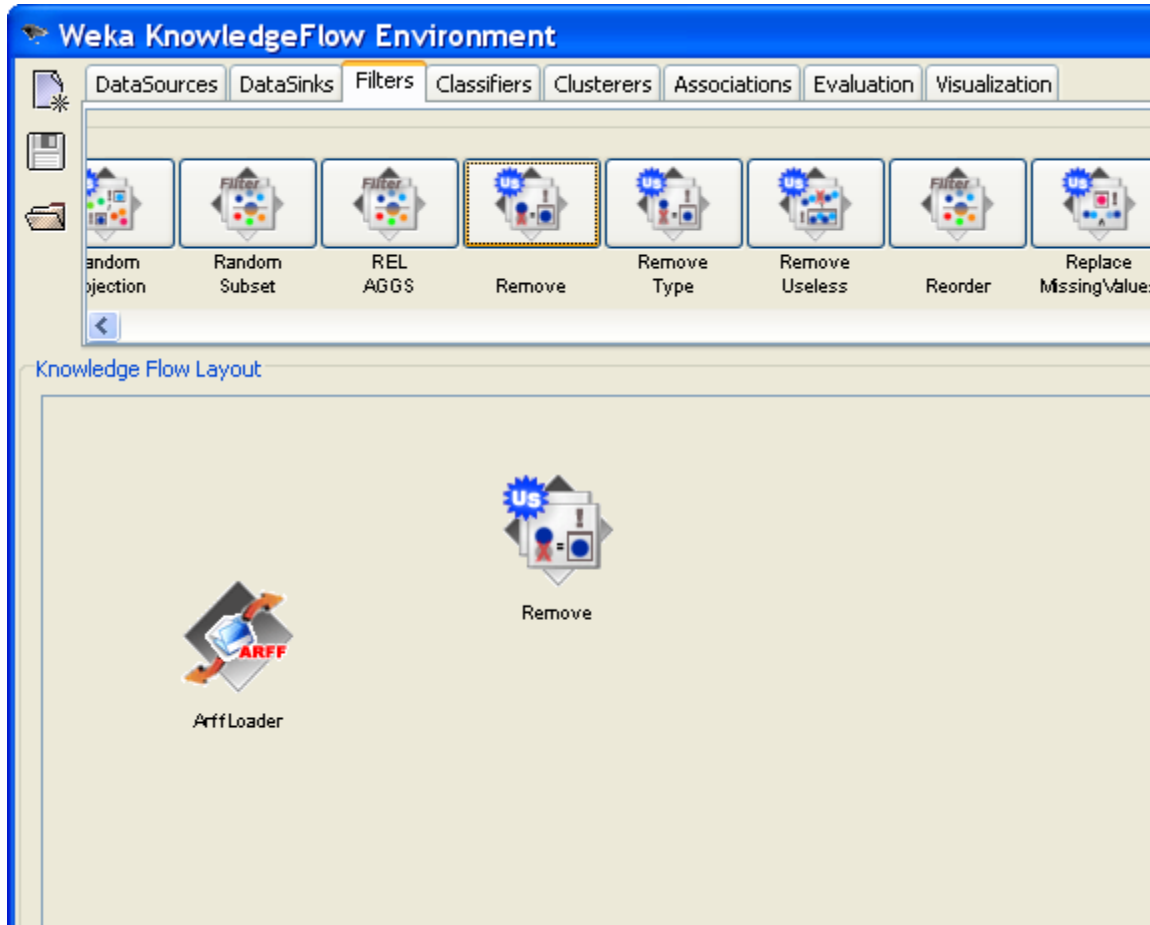
We remove all other attributes and leave only attributes 1,3, last – for simplicity

We build a classifier, which predicts income based on age and education

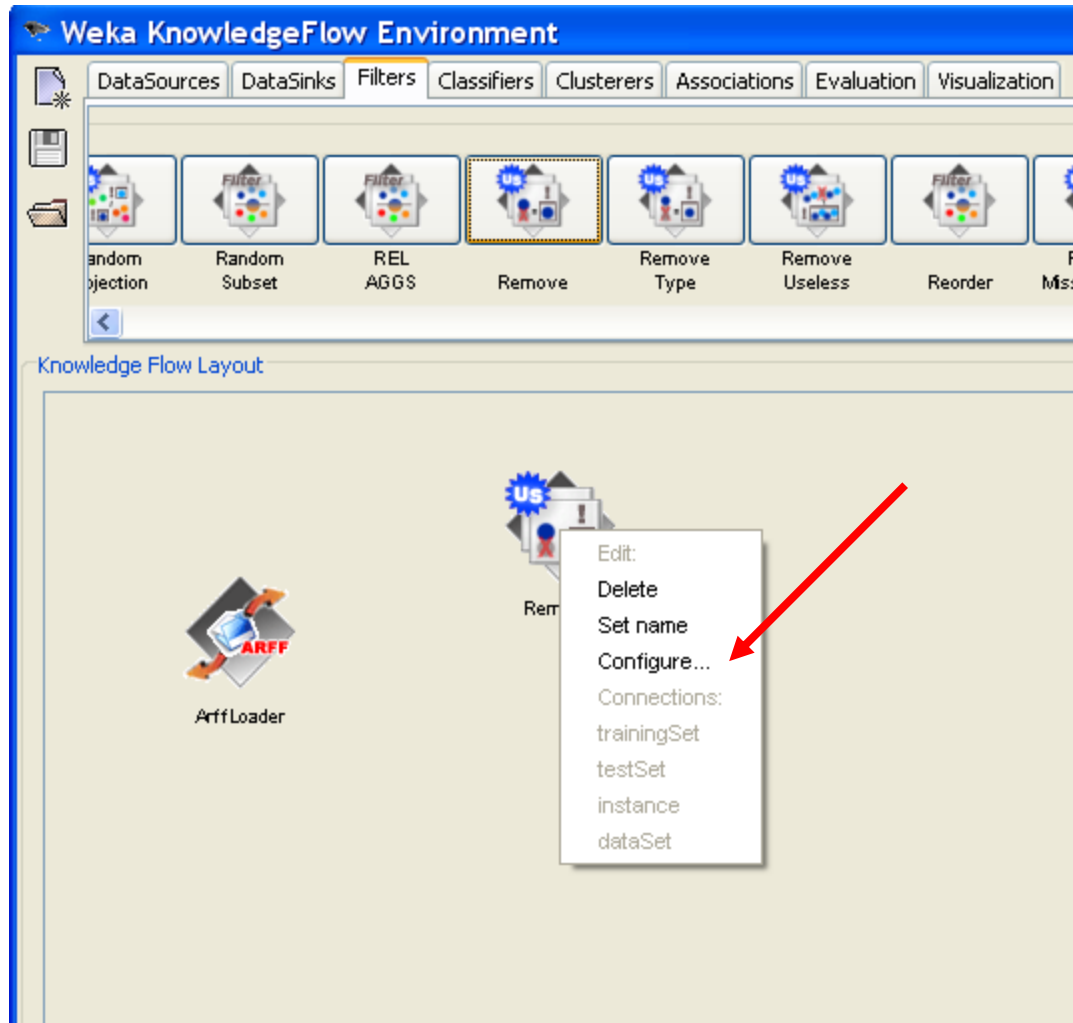
Removing attributes



Removing attributes

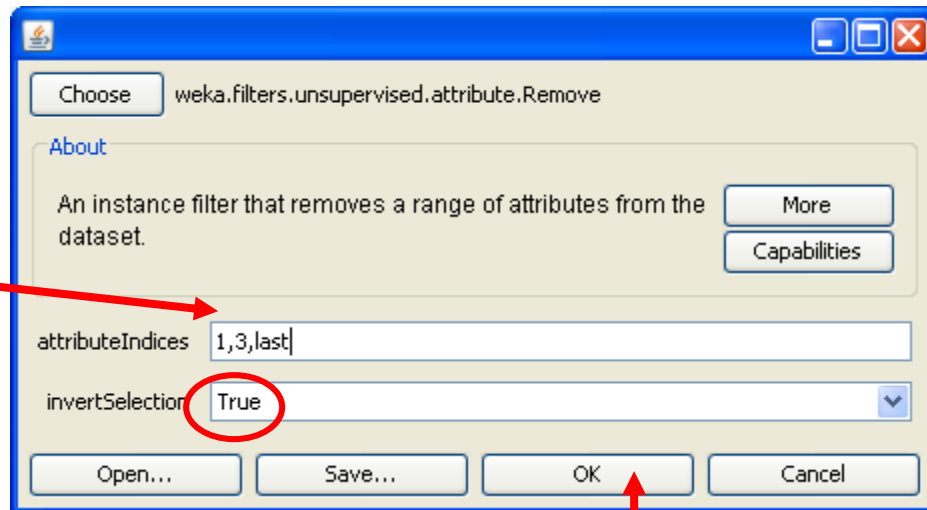


Removing attributes



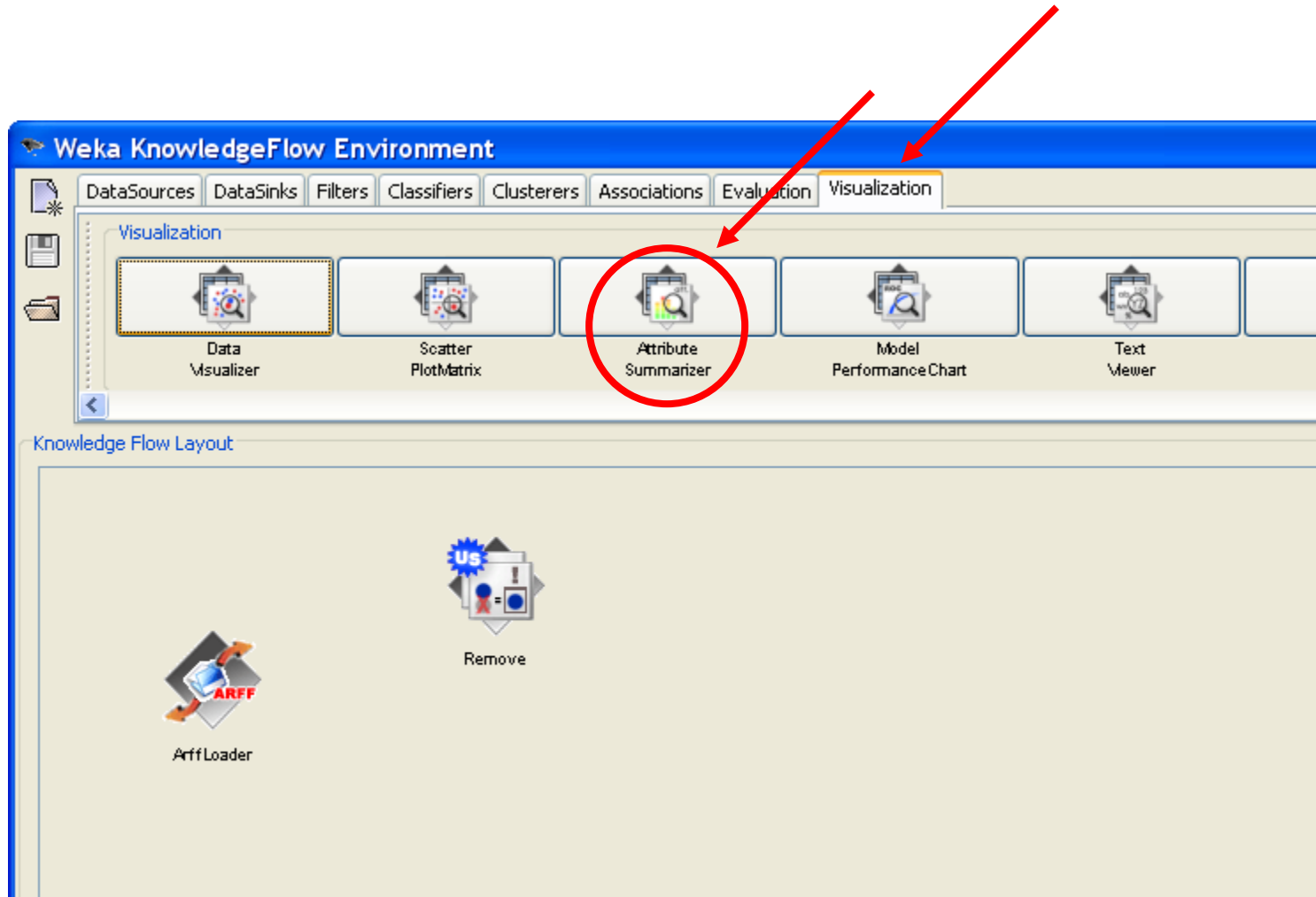
Removing attributes

What not to remove

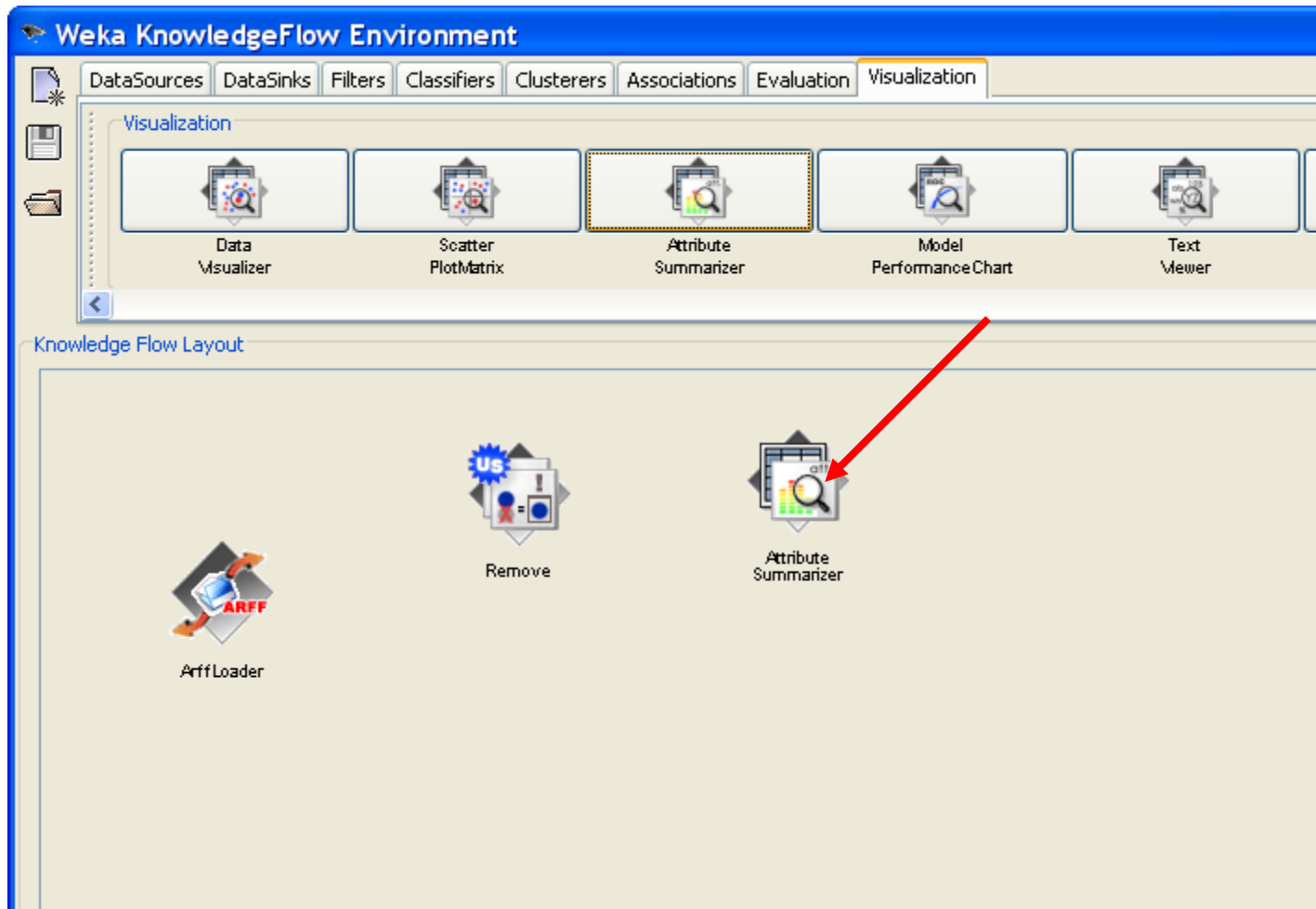


It means: remove all except attributes 1,3,last

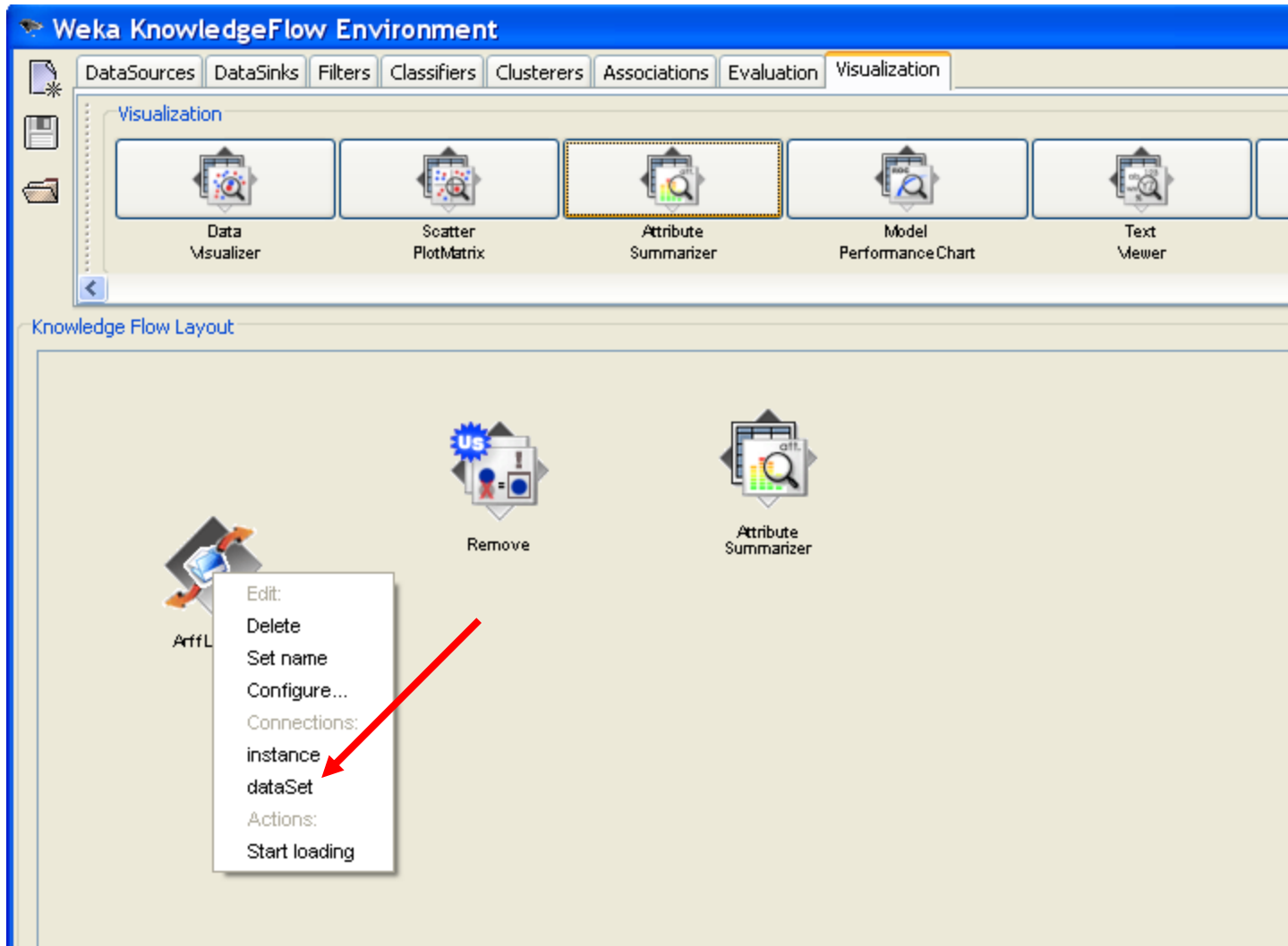
Visualize data



Visualize data



Connect the flow



Connect the flow:

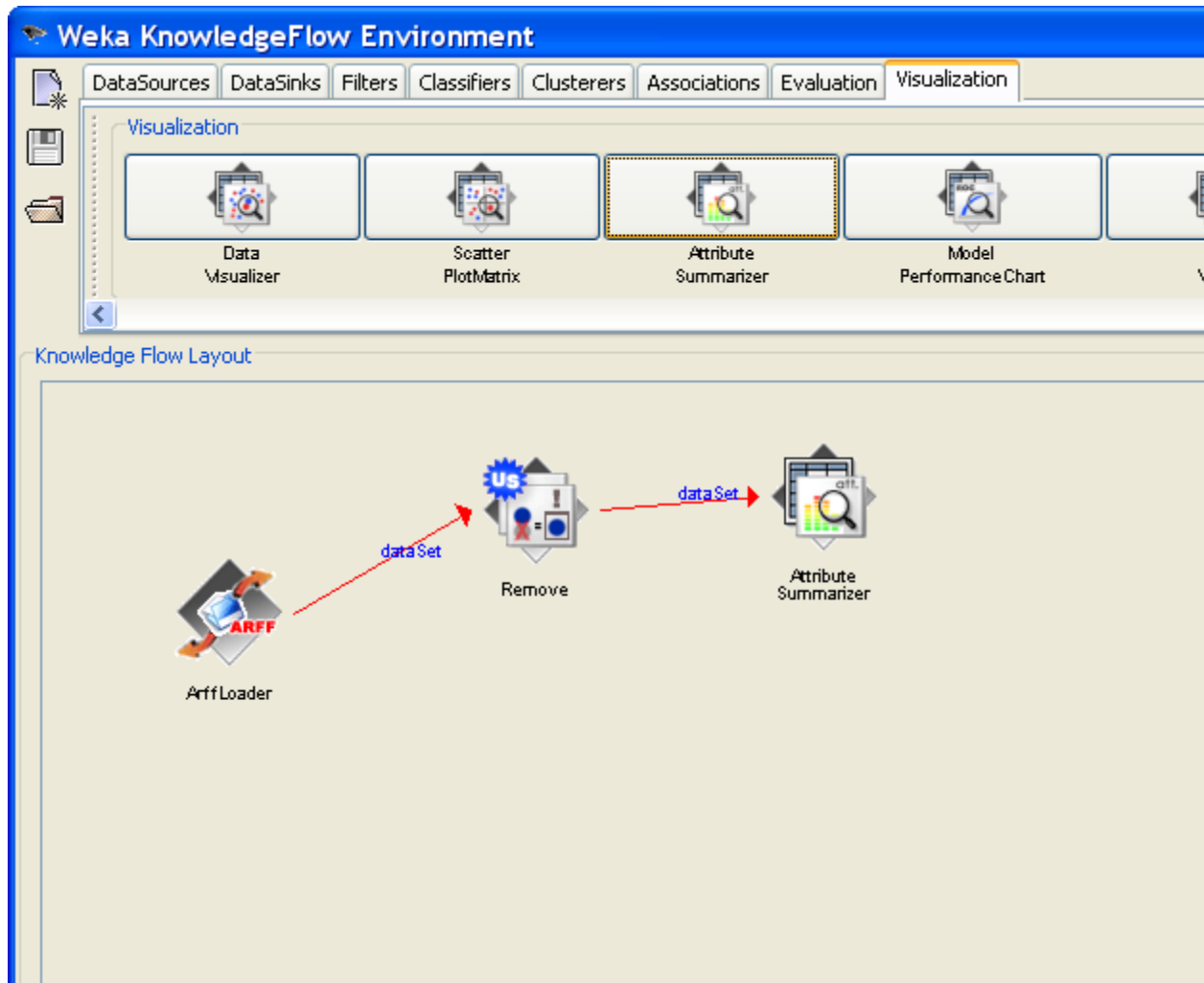
from data loader to attribute remover

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a horizontal menu bar includes tabs for "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is currently selected. Below the menu bar, a "Visualization" panel shows several tool icons: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer" (which is highlighted with a dashed border), "Model Performance Chart", and "Text Viewer".

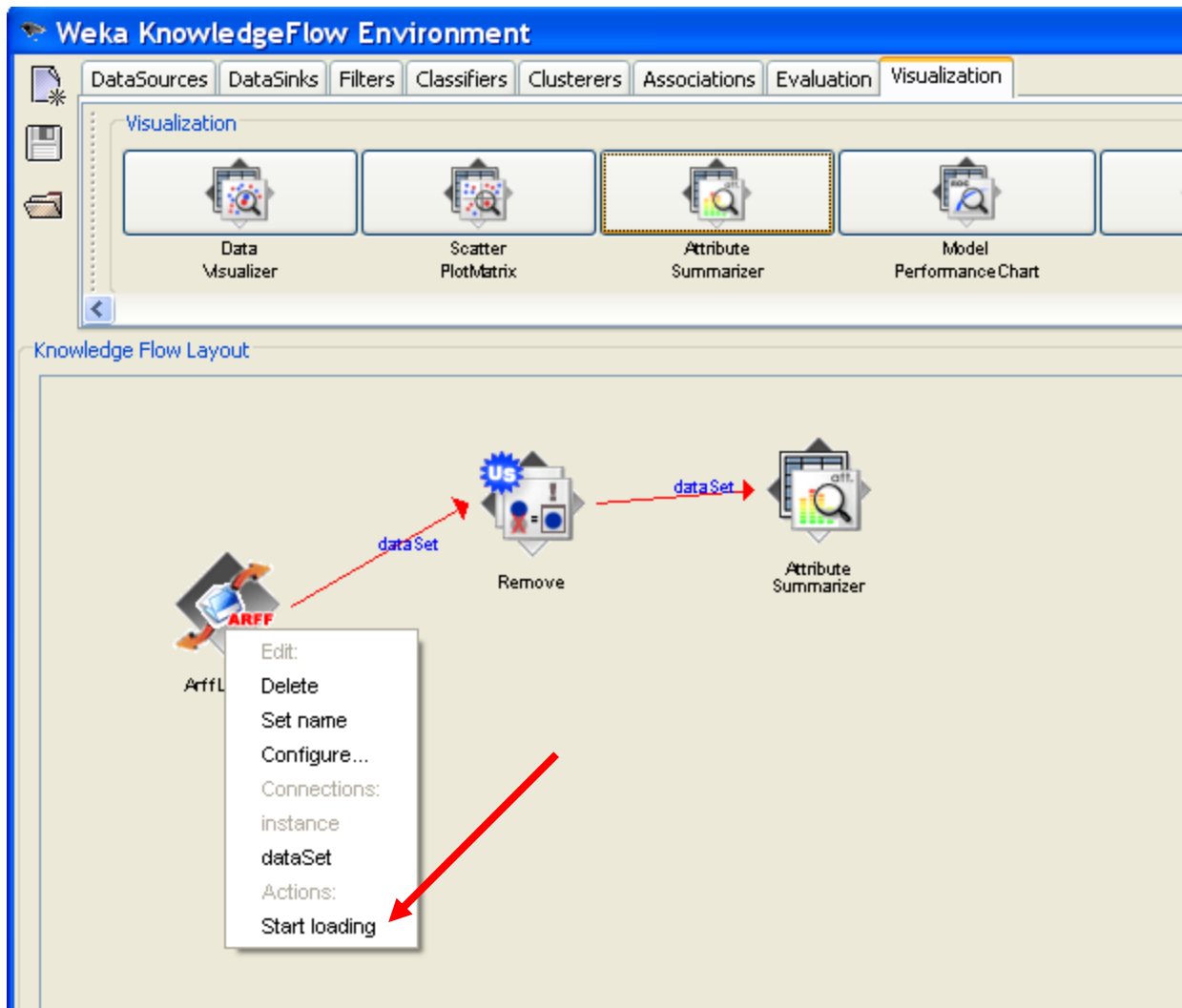
The main workspace, titled "Knowledge Flow Layout", contains a workflow diagram. On the left, an "ArffLoader" icon is connected to a "Remove" icon by a red arrow labeled "dataSet". The "Remove" icon is a blue gear with a white "U" and a red "X" on it. To the right of the "Remove" icon is an "Attribute Summarizer" icon, which is a blue gear with a white "att." and a magnifying glass on it.

Connect the flow:

from attribute remover to summarizer



Start data flow

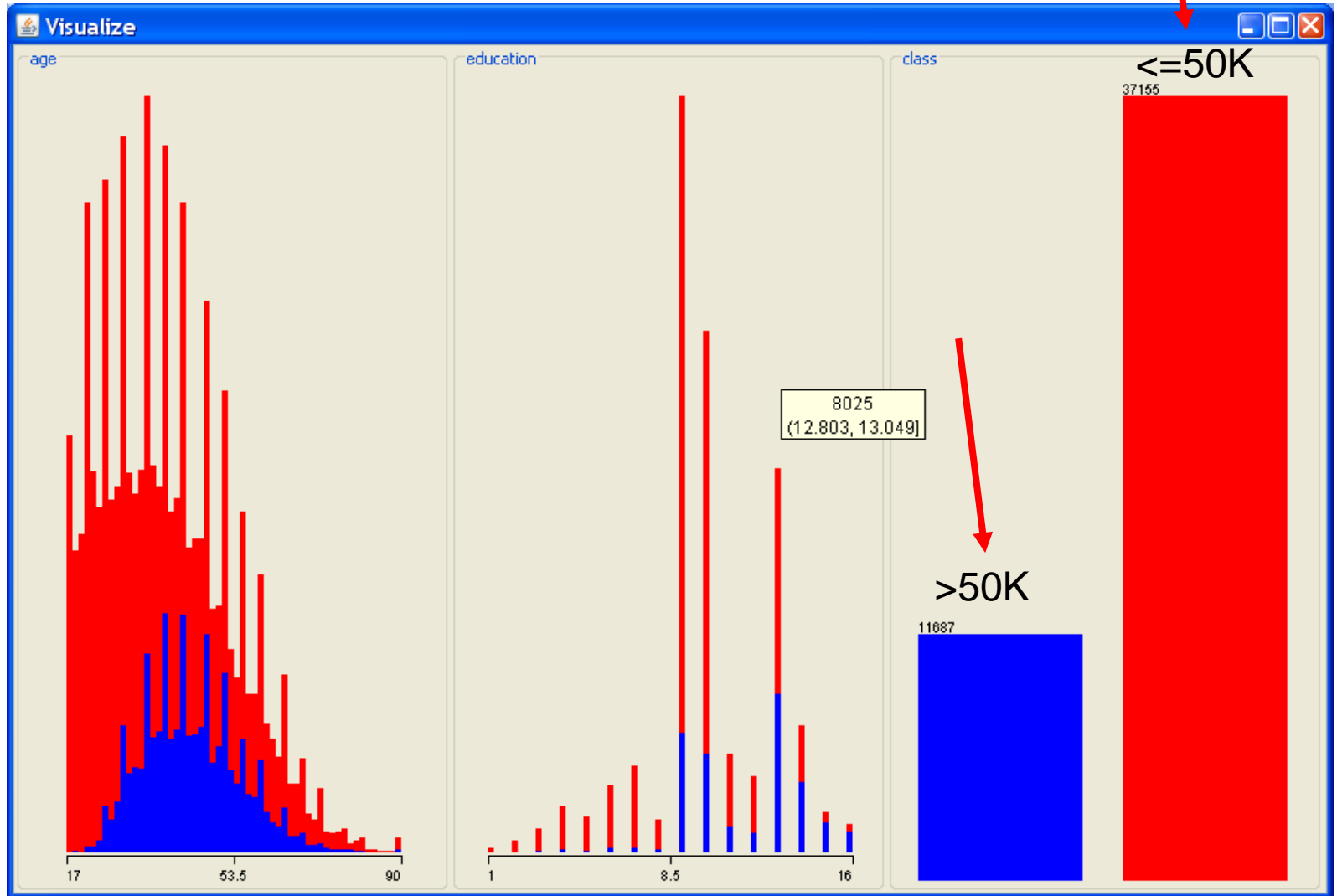


Visualize the data

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, the title bar reads "Weka KnowledgeFlow Environment". Below it, a menu bar includes "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is selected, showing a palette of visualization tools: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer", "Model Performance Chart", and "Text Viewer".

The main workspace, titled "Knowledge Flow Layout", contains a workflow diagram. It starts with an "ArffLoader" node (labeled "ARFF") connected to a "Remove" node (labeled "Us") via a red arrow labeled "data Set". The "Remove" node is then connected to an "Attribute Summarizer" node (labeled "Att Sum") via another red arrow labeled "data Set". A context menu is open over the "Attribute Summarizer" node, listing "Edit:", "Delete", "Actions:", and "Show summaries". A red arrow points to the "Show summaries" option.

Visualize the data



Assigning the class

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, the title bar reads "Weka KnowledgeFlow Environment". Below it, a series of tabs are visible: "DataSources", "DataSinks", "Filters", "Classifiers", "Clusters", "Associations", "Evaluation", and "Visualization". The "Evaluation" tab is currently selected and highlighted with a red arrow pointing from the main title above. Underneath the tabs, a toolbar contains several icons for evaluation tasks. The "Class Assigner" icon is circled in red. Below the toolbar, the "Knowledge Flow Layout" area shows a workflow diagram. The workflow starts with an "ArffLoader" icon, which outputs "data" to a "Remove" icon. The "Remove" icon then outputs "data Set" to two other icons: "Attribute Summarizer" and "ClassAssigner". The "ClassAssigner" icon is also circled in red.

Weka KnowledgeFlow Environment

DataSources DataSinks Filters Classifiers Clusters Associations Evaluation Visualization

Evaluation

Training SetMaker Test Set Maker Cross Validation FoldMaker Train Test SplitMaker Instance Stream To BatchMaker Class Assigner Class Value Picker Classifier Performance Eval

Knowledge Flow Layout

ArffLoader Remove Attribute Summarizer ClassAssigner

data data Set data Set

Configuring class assigner

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusters, Associations, Evaluation, and Visualization. The Evaluation tab is active, showing a workflow of nodes: ArffLoader, Remove, Attribute Summarizer, and ClassAssigner. A dialog box titled "ClassAssignerCustomizer" is open, showing the "About" section with the text "Designate which column is to be considered the class column in incoming data" and a text input field labeled "classColumn" containing the value "last". A red circle highlights the "classColumn" field. The workflow diagram shows data flowing from ArffLoader to Remove, and then from Remove to both Attribute Summarizer and ClassAssigner, with labels "data" and "data Set" indicating the data flow.

Weka KnowledgeFlow Environment

DataSources DataSinks Filters Classifiers Clusters Associations Evaluation Visualization

Evaluation

ClassAssignerCustomizer

About

Designate which column is to be considered the class column in incoming data

classColumn last

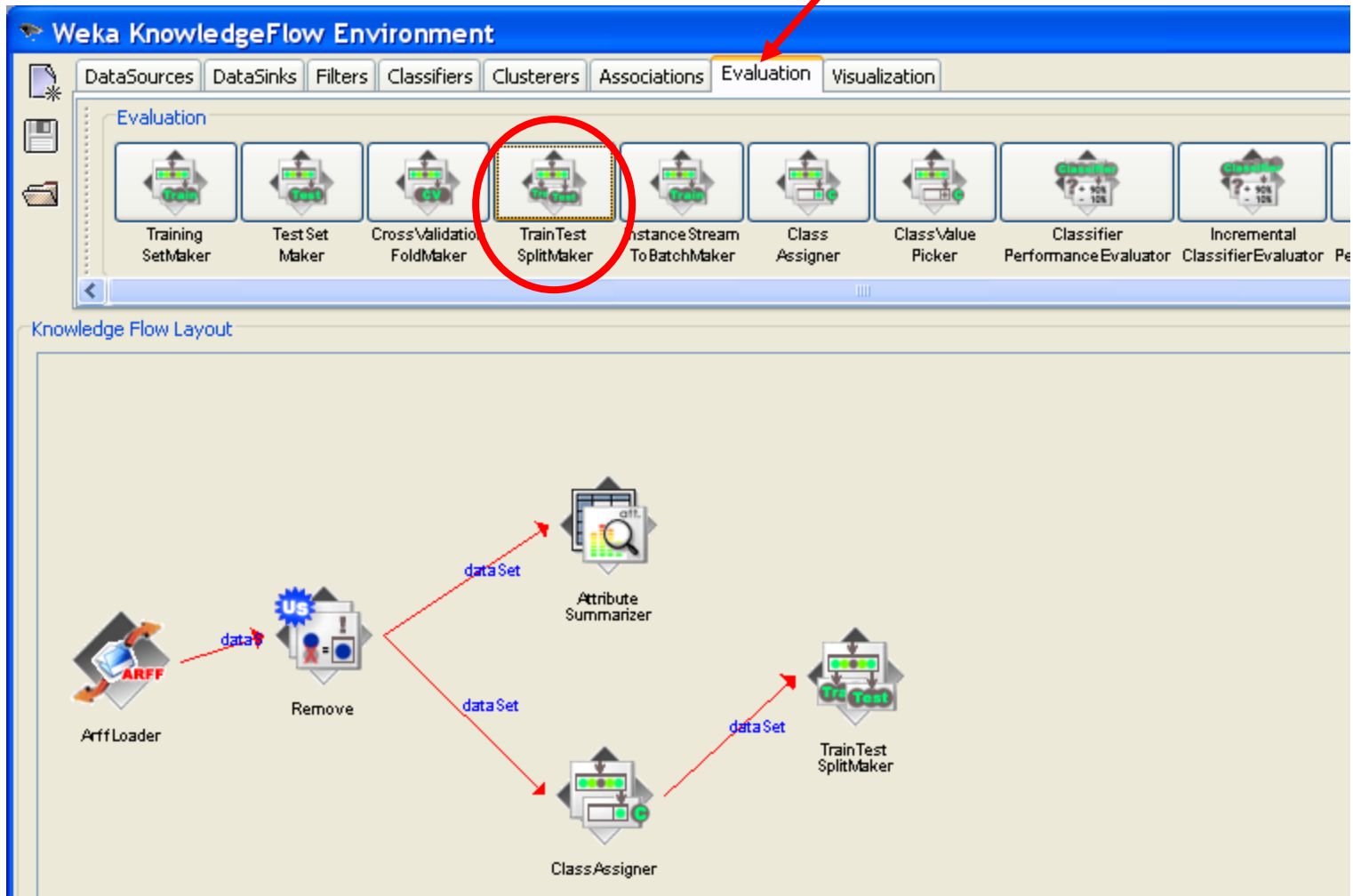
ArffLoader

Remove

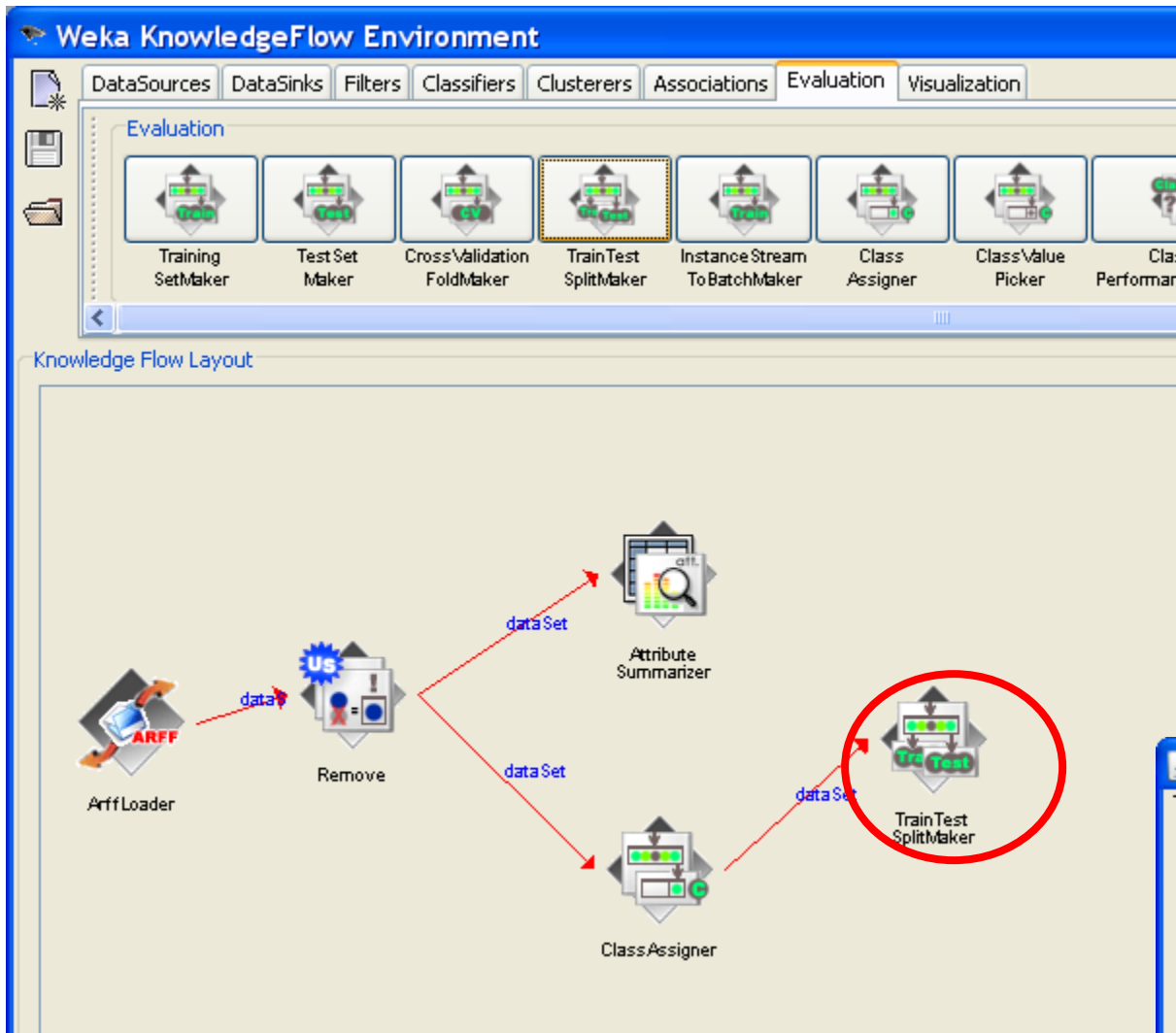
Attribute Summarizer

ClassAssigner

Subdivision of the dataset into “training” and “test” set

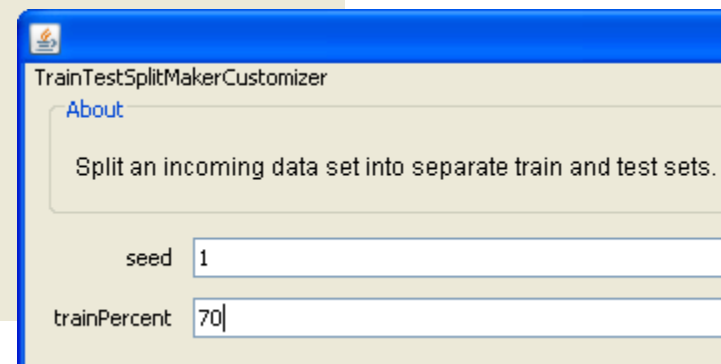


Subdivision of the dataset into “learning” and “test” set



We want to build our prediction model on the 70% of the whole dataset, and test on the remaining

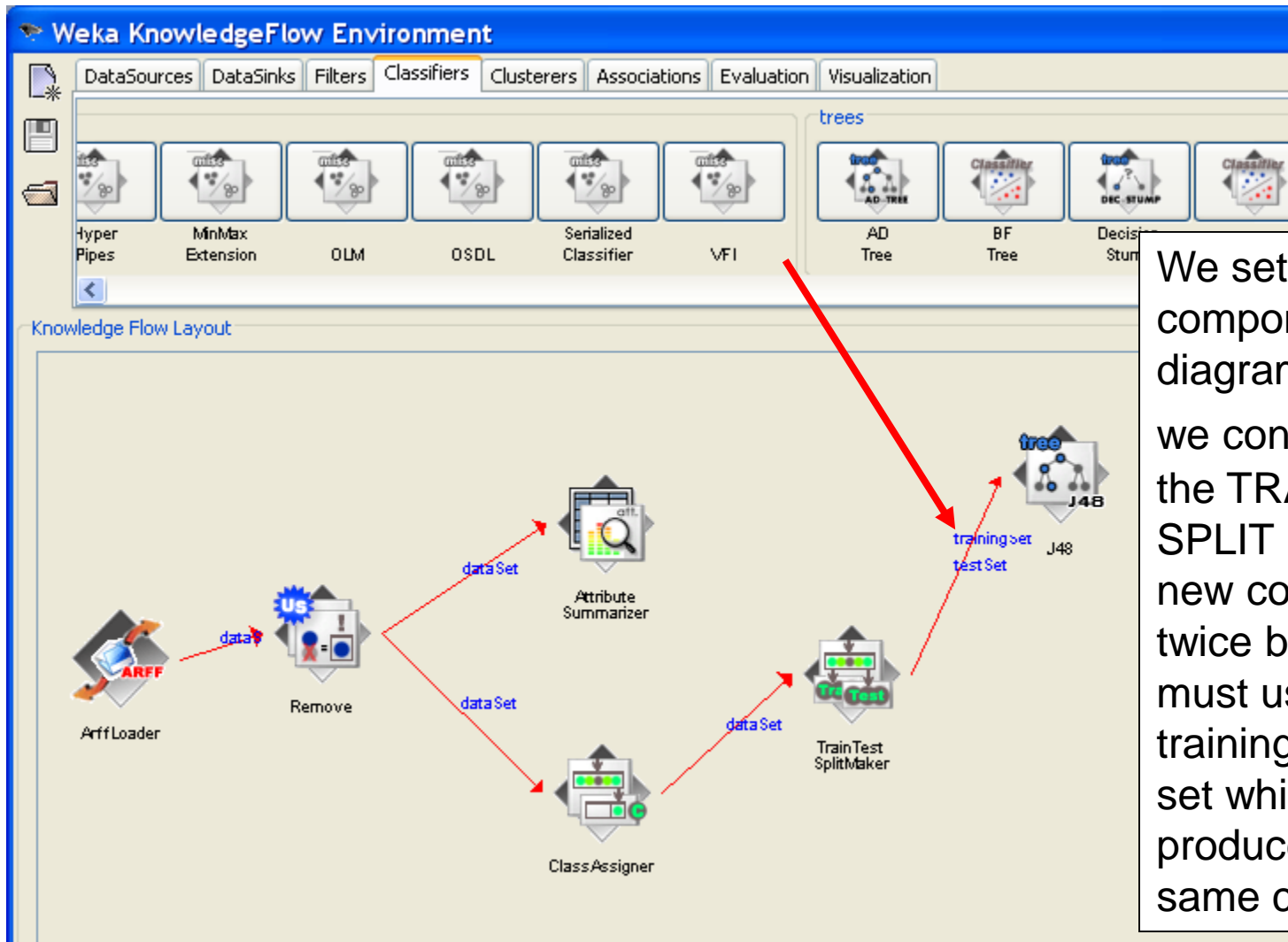
So, we set the TRAINTEST SPLIT MAKER (EVALUATION) in the diagram and configure its parameters.



Choosing discrete classifier – decision tree

The screenshot displays the Weka KnowledgeFlow Environment interface. The top menu bar includes 'DataSources', 'DataSinks', 'Filters', 'Classifiers', 'Clusters', 'Associations', 'Evaluation', and 'Visualization'. The 'Classifiers' tab is active, showing a list of classifiers. A red arrow points from the title 'Choosing discrete classifier – decision tree' to the 'Classifiers' tab. Another red arrow points to the 'trees' sub-tab, which is highlighted. Under the 'trees' sub-tab, several classifier icons are visible: 'AD Tree', 'BF Tree', 'Decision Stump', 'FT', 'Id3', and 'J48'. The 'J48' icon is circled in red. Below the classifier list, the 'Knowledge Flow Layout' area shows a workflow diagram. The workflow starts with an 'ArffLoader' node, followed by a 'Remove' node. From 'Remove', two 'data Set' arrows lead to 'Attribute Summarizer' and 'ClassAssigner' nodes. From 'ClassAssigner', a 'data Set' arrow leads to a 'Train Test SplitMaker' node. A 'J48' classifier icon is also present in the layout area.

Connecting classifier to the data



We set J48 component in the diagram, we connect **twice** the TRAIN TEST SPLIT MAKER to this new component: twice because we must use together the training and the test set which are produced by the same component.

Adding visualizer to see the classification results

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a horizontal menu bar includes tabs for "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is highlighted with a red circle. Below the menu bar, a "Visualization" panel contains several tool icons: "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer", "Model Performance Chart", "Text Viewer", and "Graph Viewer". The "Graph Viewer" icon is also circled in red. A red arrow points from this icon to the "Graph" node in the workflow below.

The main area, titled "Knowledge Flow Layout", shows a workflow diagram with the following components and connections:

- ArffLoader** (input) connects to **Remove** via a red arrow labeled "data".
- Remove** (with a "Us" icon) branches into two paths:
 - One path goes to **Attribute Summarizer** via a red arrow labeled "data Set".
 - Another path goes to **ClassAssigner** via a red arrow labeled "data Set".
- ClassAssigner** connects to **Train Test SplitMaker** via a red arrow labeled "data Set".
- Train Test SplitMaker** connects to **J48** (with a "tree" icon) via a red arrow labeled "training set" and "test Set".
- J48** connects to **GraphViewer** via a red arrow labeled "graph".

Perform classification

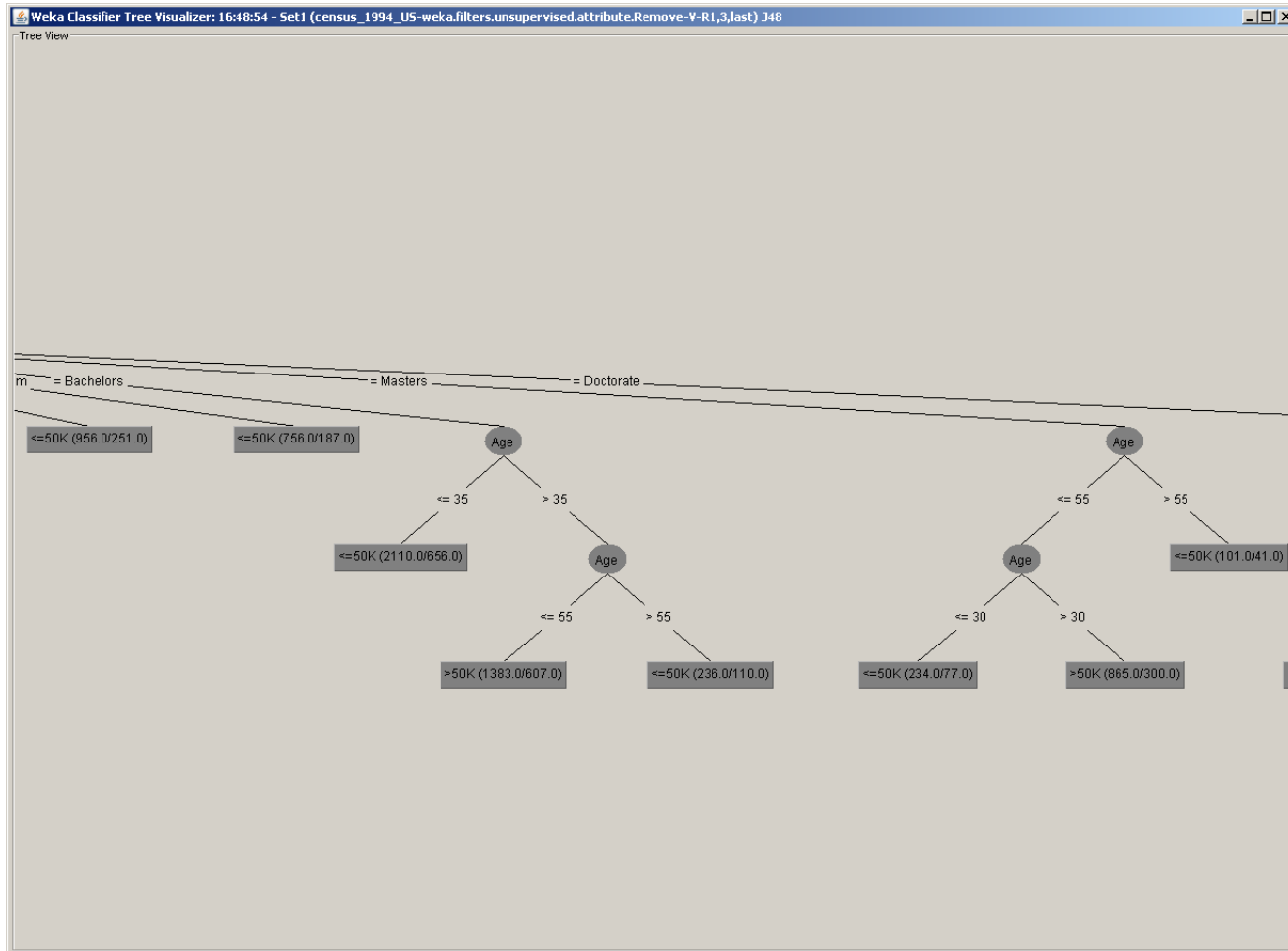
The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a horizontal menu lists several categories: DataSources, DataSinks, Filters, Classifiers, Clusterers, Associations, Evaluation, and Visualization. The "Visualization" category is currently selected and highlighted.

Under the "Visualization" menu, a row of six icons is shown, each with a label below it: Data Visualizer, Scatter PlotMatrix, Attribute Summarizer, Model Performance Chart, Text Viewer, and Graph Viewer. The Graph Viewer icon is highlighted with a dashed yellow border.

The main area of the interface is titled "Knowledge Flow Layout" and contains a workflow diagram. The workflow starts with an "Artificial Data" node on the left. A red arrow labeled "data" points from this node to a "US" node. From the "US" node, two red arrows labeled "data Set" branch out: one to an "Attribute Summarizer" node and another to a "Class Assigner" node. From the "Class Assigner" node, a red arrow labeled "data Set" points to a "Train Test SplitMaker" node. From the "Train Test SplitMaker" node, two red arrows labeled "training set" and "test Set" point to a "tree J48" node. Finally, a red arrow labeled "graph" points from the "tree J48" node to a "Graph Viewer" node on the far right.

A context menu is open over the "Artificial Data" node, listing the following options: Edit, Delete, Set name, Configure..., Connections: instance, dataSet, Actions: Start loading. A red arrow points from the "Start loading" option to the "US" node.

Show classification results (decision tree)

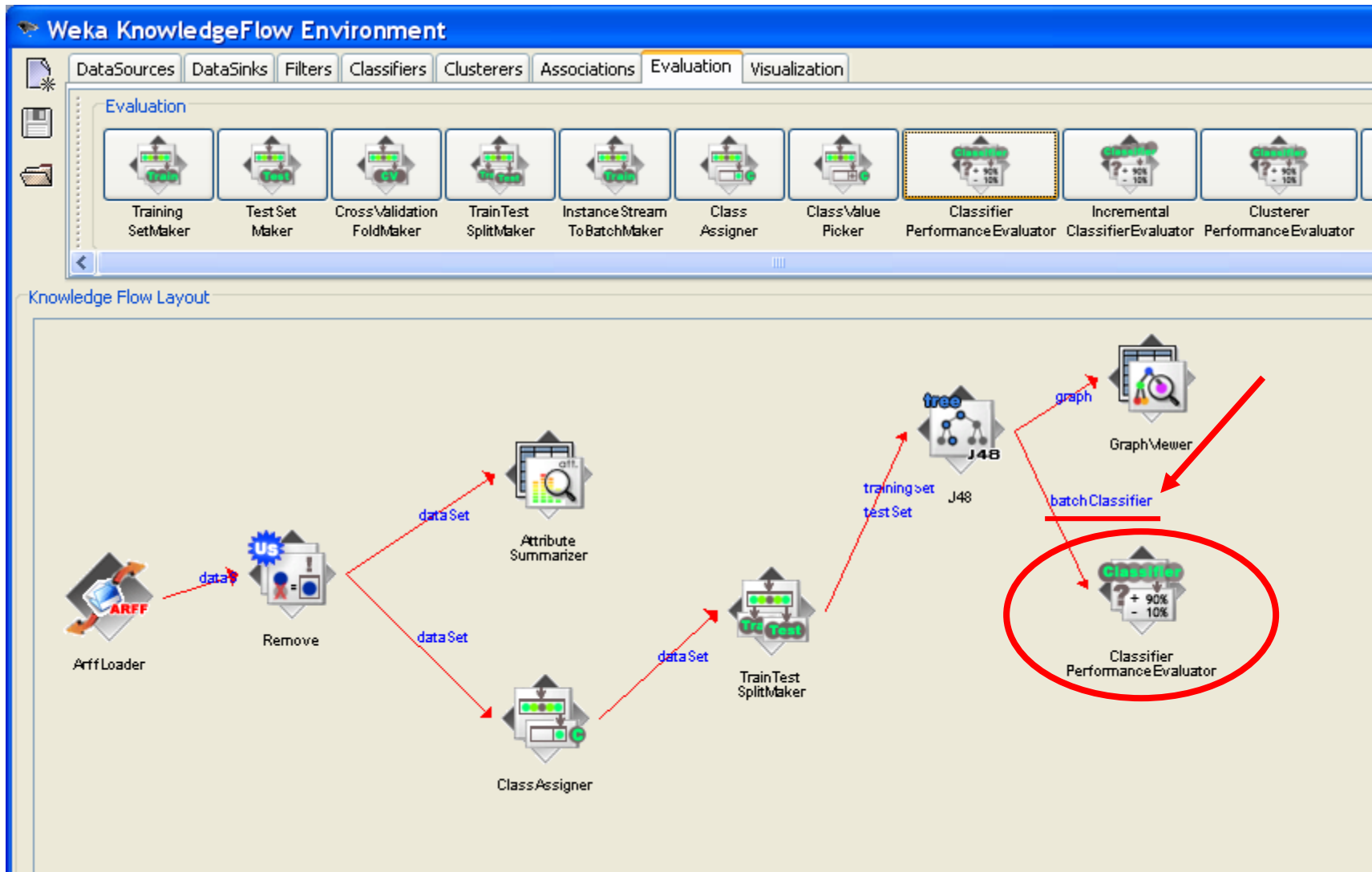


Classifier evaluation

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, the title bar reads "Weka KnowledgeFlow Environment". Below it, a menu bar includes "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Evaluation" menu is circled in red. Below the menu bar, a toolbar shows various workflow components. The "Classifier Performance Evaluator" component is also circled in red. The main workspace, titled "Knowledge Flow Layout", contains a workflow diagram with the following components and connections:

- ArffLoader** (input) connects to **Remove** via a "data" edge.
- Remove** (with "Us" icon) branches into two "data Set" edges:
 - One to **Attribute Summarizer**.
 - One to **ClassAssigner**.
- ClassAssigner** connects to **Train Test SplitMaker** via a "data Set" edge.
- Train Test SplitMaker** connects to **J48** (Classifier) via a "training set" and "test Set" edge.
- J48** connects to **GraphViewer** via a "graph" edge.

Connecting classifier to the evaluator



Selecting performance model: chart

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, a blue header bar contains the text "Weka KnowledgeFlow Environment". Below this, a series of tabs are visible: "DataSources", "DataSinks", "Filters", "Classifiers", "Clusterers", "Associations", "Evaluation", and "Visualization". The "Visualization" tab is currently selected and highlighted with a red arrow pointing to it from above. Underneath the tabs, a row of visualization tool icons is shown. The "Model Performance Chart" icon is circled in red. Other icons include "Data Visualizer", "Scatter PlotMatrix", "Attribute Summarizer", "Text Viewer", "Graph Viewer", and "Strip Chart".

Below the visualization tools, the "Knowledge Flow Layout" section shows a workflow diagram. The flow starts with an "ArffLoader" node connected to a "Remove" node (marked with a "Us" icon). From "Remove", two "data Set" arrows branch out to "Attribute Summarizer" and "ClassAssigner". From "ClassAssigner", a "data Set" arrow leads to "Train Test SplitMaker". From "Train Test SplitMaker", two "data Set" arrows lead to "J48" (a "tree" classifier) and "Classifier Performance Evaluator". From "J48", a "graph" arrow leads to "Graph Viewer" and a "batch Classifier" arrow leads to "Classifier Performance Evaluator". From "Classifier Performance Evaluator", a "threshold Data" arrow leads to "Model Performance Chart". A red arrow points from the "Model Performance Chart" icon in the visualization toolbar to this "Model Performance Chart" node in the workflow.

Running the model

Weka KnowledgeFlow Environment

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

Visualization

Data Visualizer | Scatter PlotMatrix | Attribute Summarizer | **Model Performance Chart** | Text Mewer | Graph Mewer | Strip Chart

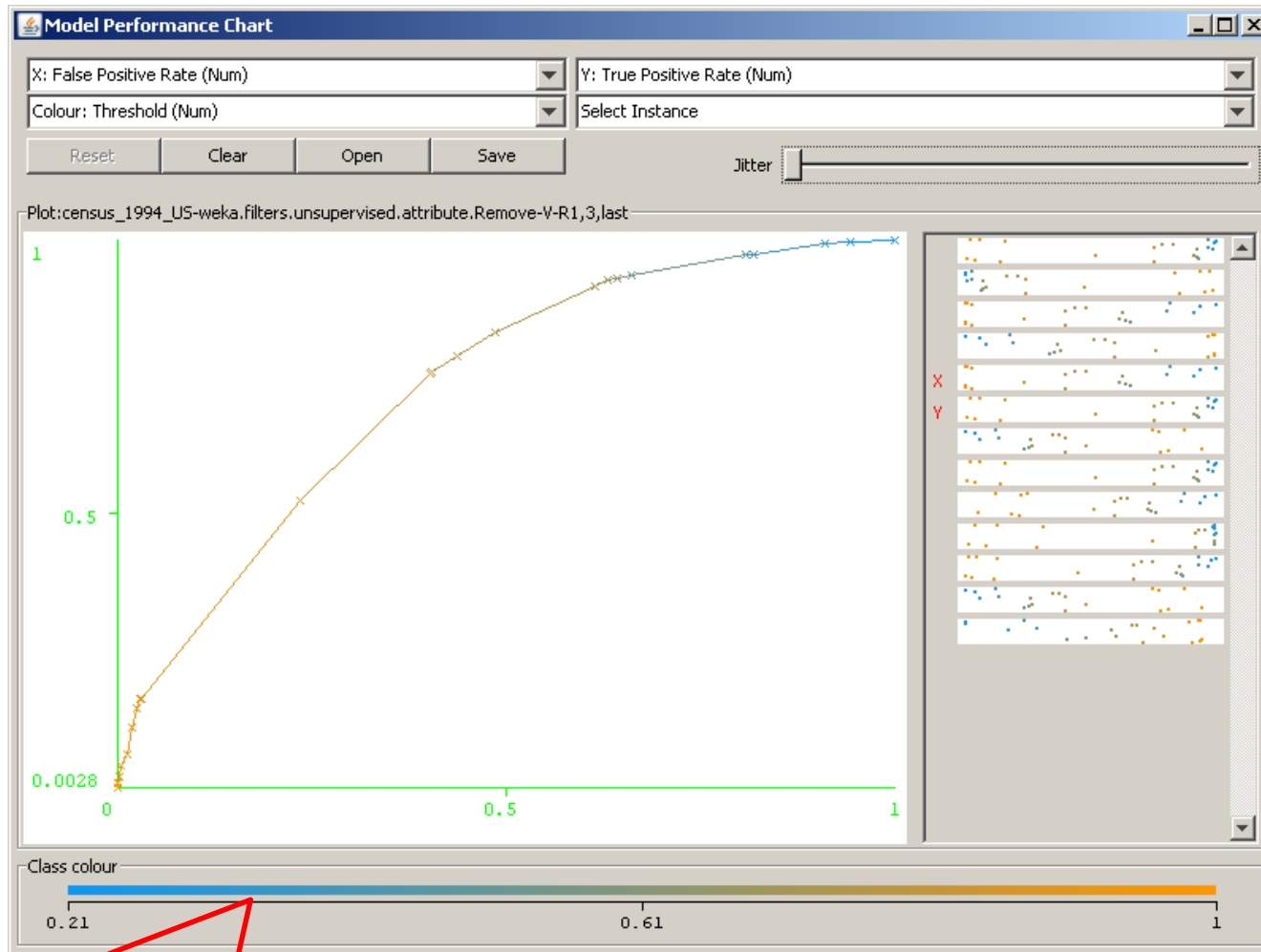
Knowledge Flow Layout

```
graph LR; AP[AP] -- data --> CA[Class Assigner]; AP -- data Set --> AS[Attribute Summarizer]; CA -- data Set --> TTS[Train Test SplitMaker]; TTS -- training set --> J48[J48]; TTS -- test Set --> J48; J48 -- batch Classifier --> CPE[Classifier Performance Evaluator]; CPE -- threshold Data --> MPC[Model Performance Chart]; J48 -- graph --> GM[Graph Mewer];
```

Context Menu:

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

Show chart: View ROC curve

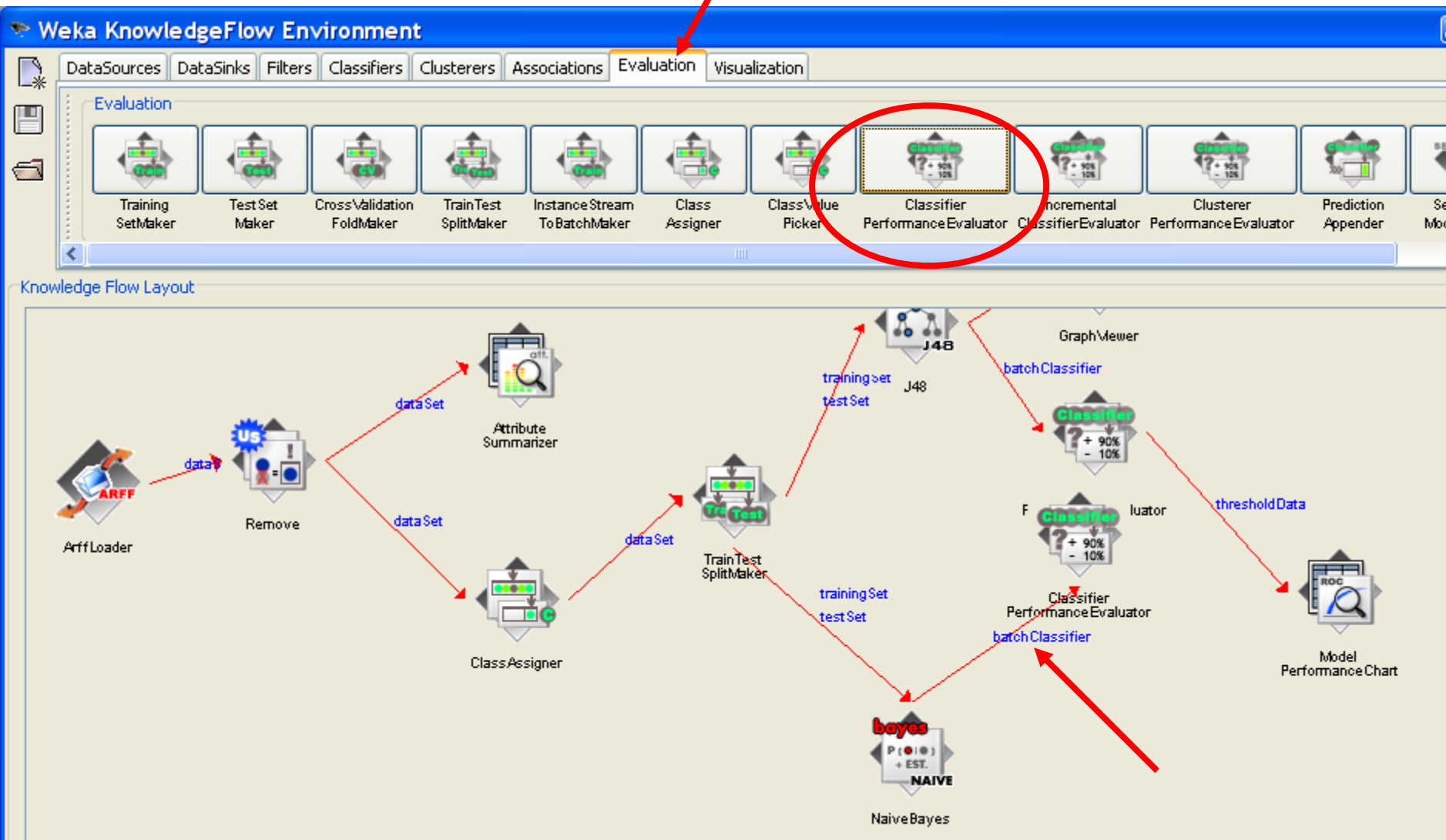


Threshold value for dividing positives from negatives

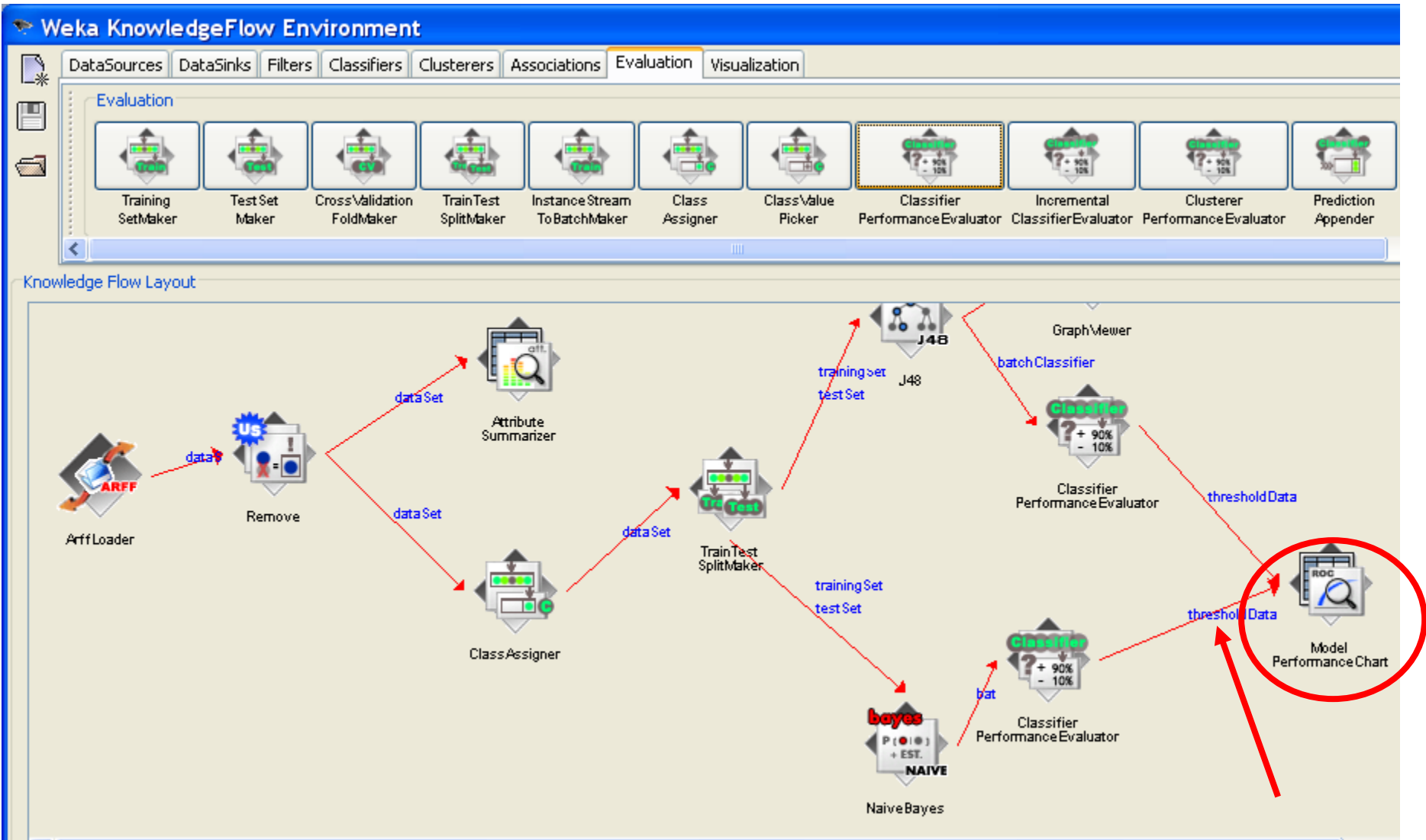
Adding Naïve Bayes classifier

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, the 'Classifiers' tab is selected in the toolbar. Below the toolbar, a row of classifier icons is shown, with the 'Naive Bayes' icon circled in red. The main workspace, titled 'Knowledge Flow Layout', contains a workflow diagram. The workflow starts with an 'ArffLoader' node connected to a 'Remove' node. The 'Remove' node outputs two 'data Set' streams: one to an 'Attribute Summarizer' node and another to a 'Class Assigner' node. The 'Class Assigner' node outputs a 'data Set' stream to a 'Train Test SplitMaker' node. The 'Train Test SplitMaker' node outputs two streams: 'training set' and 'test Set', both of which are circled in red. The 'training set' stream is connected to a 'Naive Bayes' classifier node (also circled in red). The 'test Set' stream is connected to a 'tree J48' node. The 'tree J48' node outputs a 'graph' stream to a 'Graph Viewer' node and a 'batch Classifier' stream to a 'Classifier Performance Evaluator' node. The 'Classifier Performance Evaluator' node outputs a 'threshold Data' stream to a 'Model Performance Chart' node.

Adding separate performance evaluator for Naïve Bayes classifier



Connecting second performance evaluator to the same Model Performance Chart



Run both classifiers

Weka KnowledgeFlow Environment

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

Evaluation

Training SetMaker | Test Set Maker | CrossValidation FoldMaker | TrainTest SplitMaker | InstanceStream ToBatchMaker | Class Assigner | ClassValue Picker | Classifier Performance Evaluator | Incremental Classifier Evaluator | Clusterer Performance Evaluator | Prediction Appender | Serial ModelS

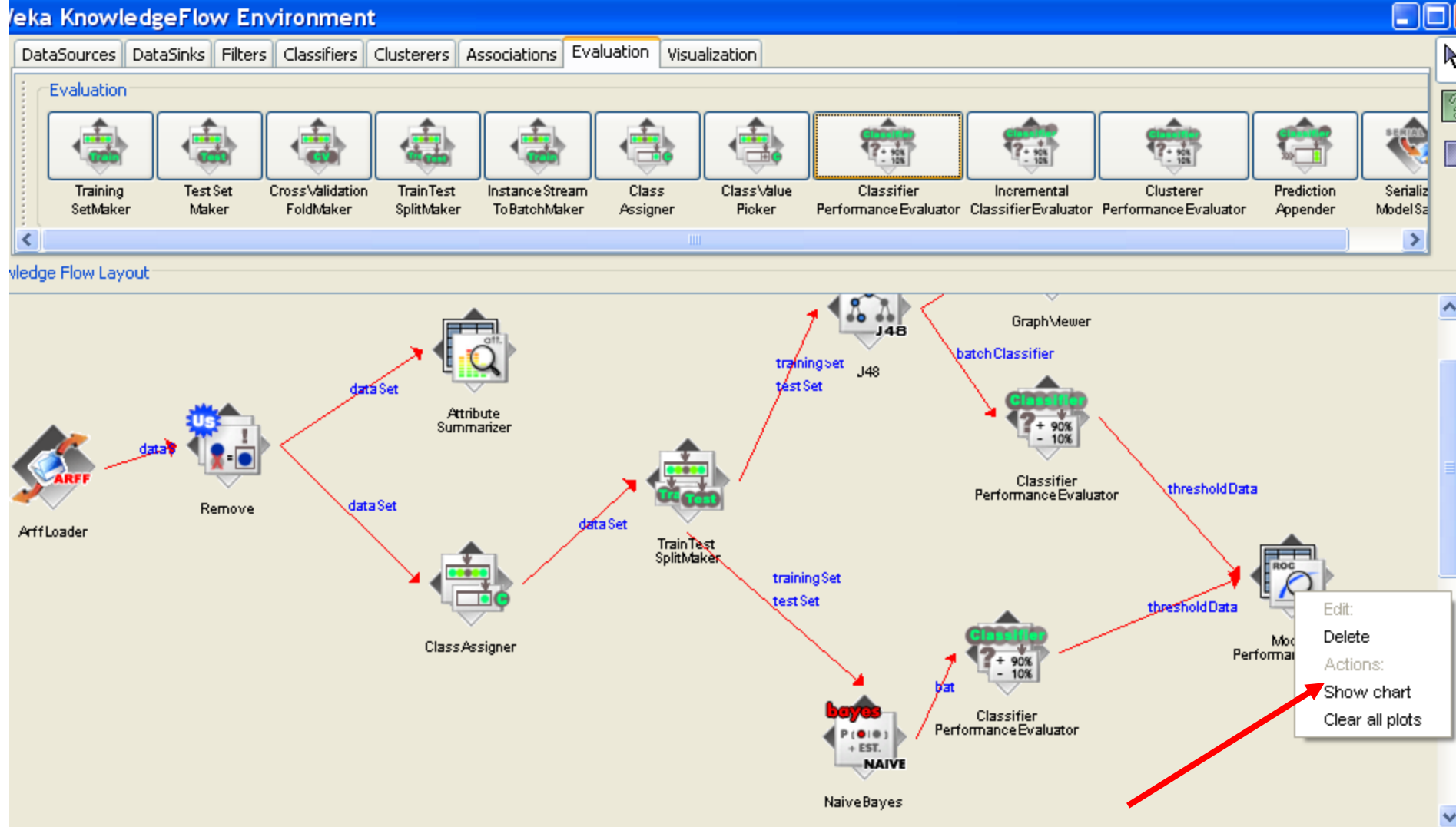
Knowledge Flow Layout

```
graph TD; ArffLoader[Arff Loader] -- data --> US[US]; US -- data set --> AttributeSummarizer[Attribute Summarizer]; US -- data set --> TrainTestSplitMaker[TrainTest SplitMaker]; TrainTestSplitMaker -- training set --> NaiveBayes[boyes NAIVE]; TrainTestSplitMaker -- test set --> NaiveBayes; TrainTestSplitMaker -- training set --> J48[J48]; TrainTestSplitMaker -- test set --> J48; NaiveBayes -- bat --> NaiveBayesEvaluator[Classifier Performance Evaluator]; J48 -- batchClassifier --> J48Evaluator[Classifier Performance Evaluator]; NaiveBayesEvaluator -- thresholdData --> ModelPerformanceChart[Model Performance Chart]; J48Evaluator -- thresholdData --> ModelPerformanceChart;
```

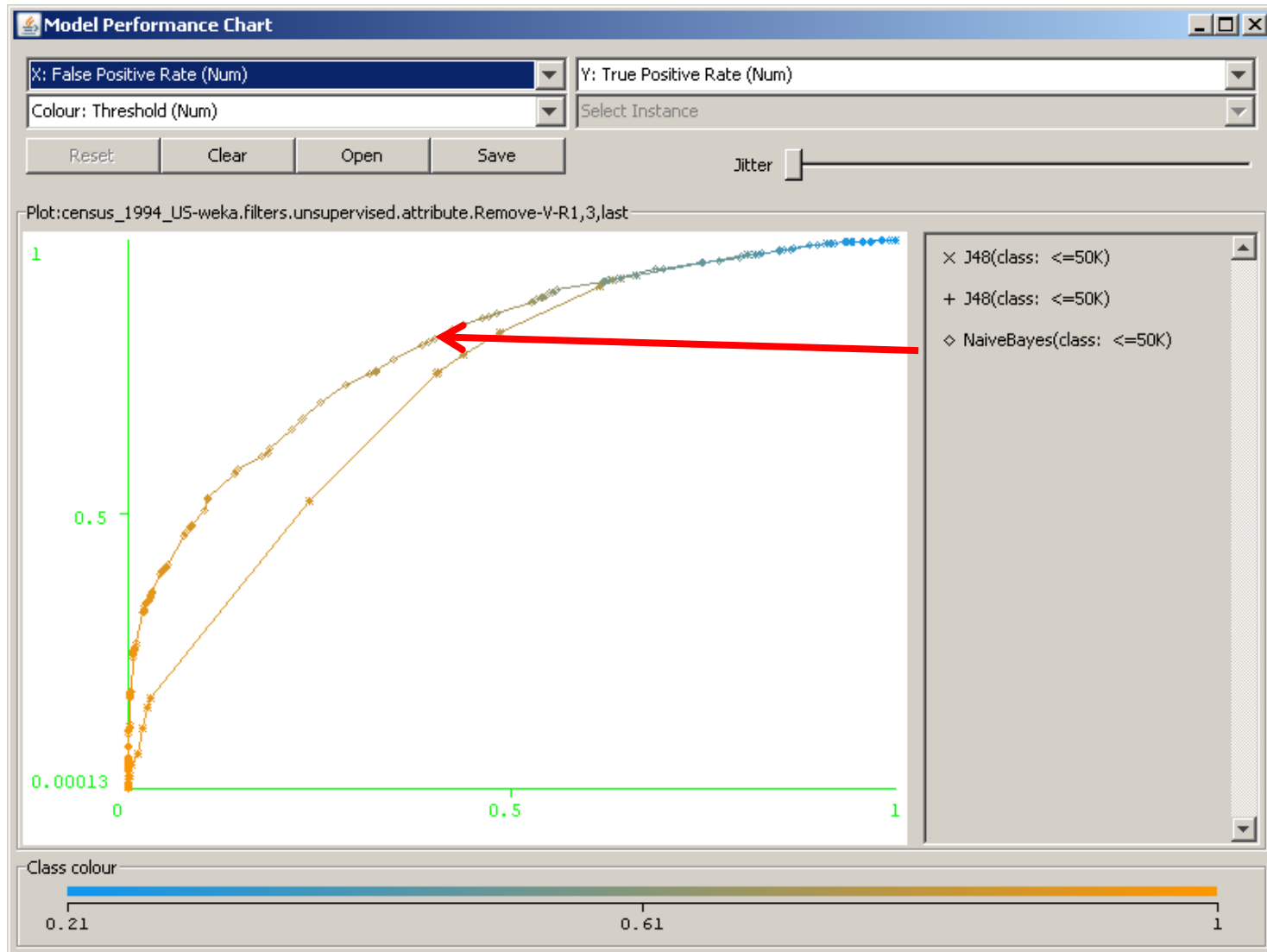
Arff Loader context menu:

- Edit:
- Delete
- Set name
- Configure...
- Connections:
- instance
- dataSet
- Actions:
- Start loading

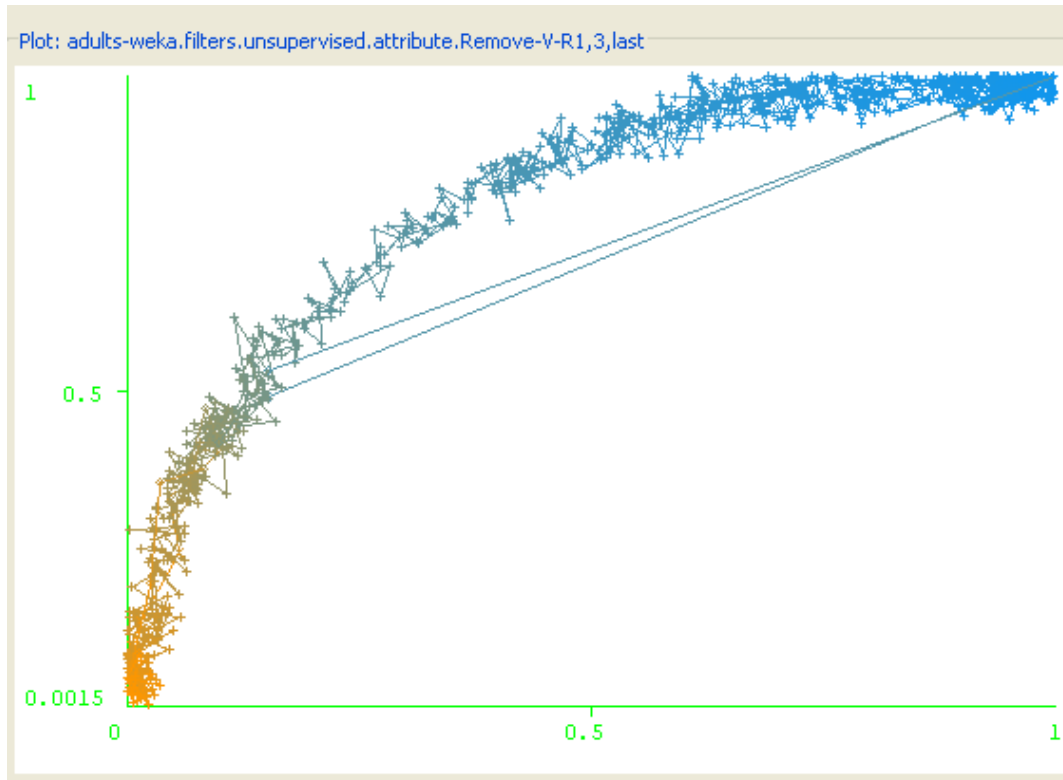
View ROC curves for both classifiers



Compare classifiers using their ROC curves



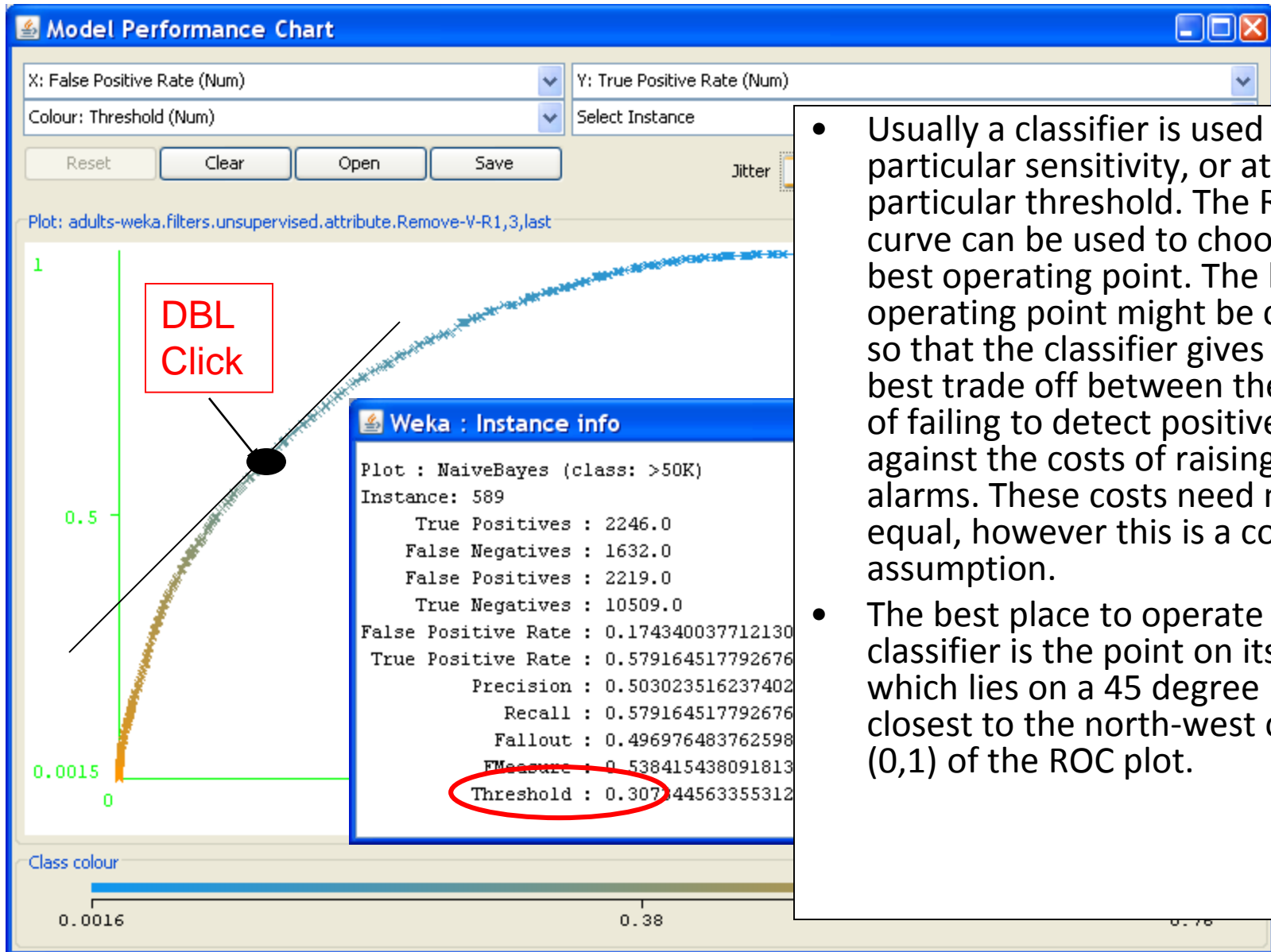
How good is the classifier



The area under the ROC curve shows the quality of a classifier – not accuracy, but the ability to separate between positive and negative instances.

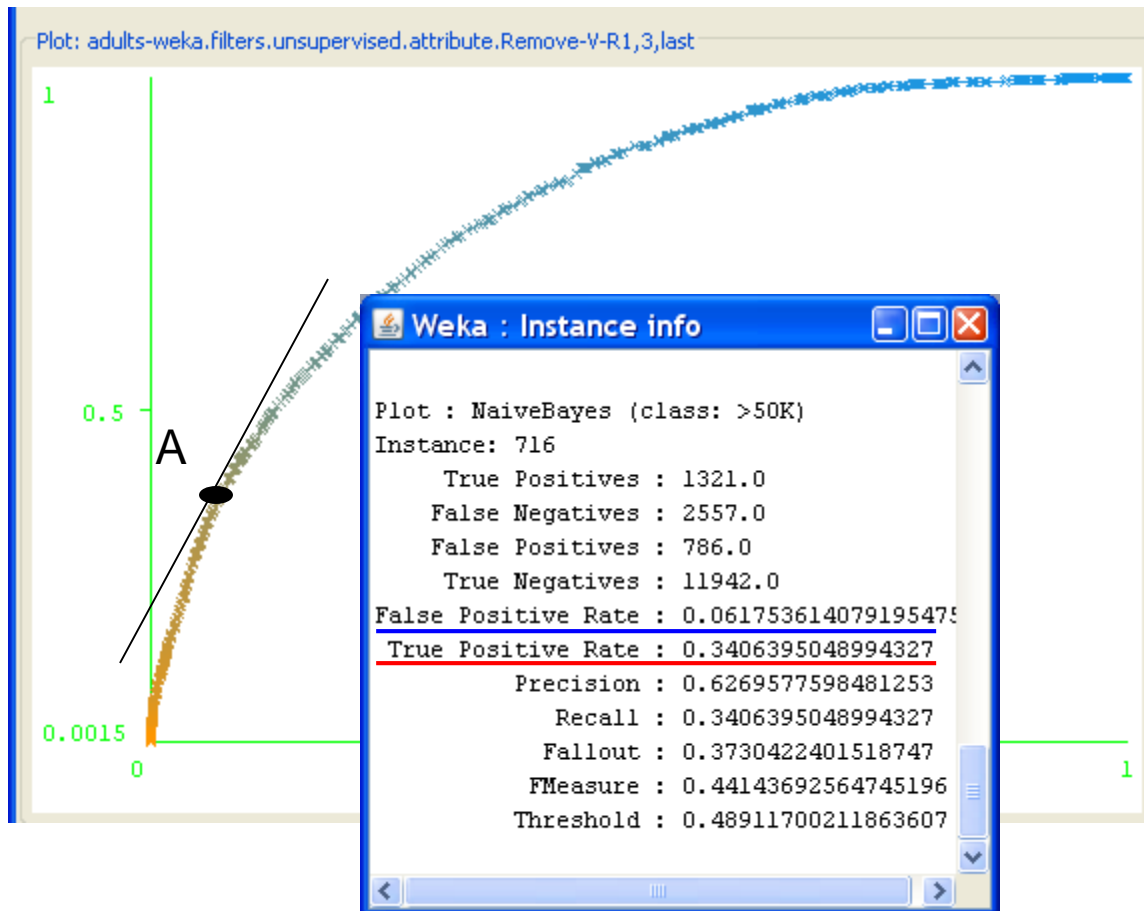
What classifier is better?

Choosing the Operating Point



- Usually a classifier is used at a particular sensitivity, or at a particular threshold. The ROC curve can be used to choose the best operating point. The best operating point might be chosen so that the classifier gives the best trade off between the costs of failing to detect positives against the costs of raising false alarms. These costs need not be equal, however this is a common assumption.
- The best place to operate the classifier is the point on its ROC which lies on a 45 degree line closest to the north-west corner (0,1) of the ROC plot.

Cost sensitive operating points

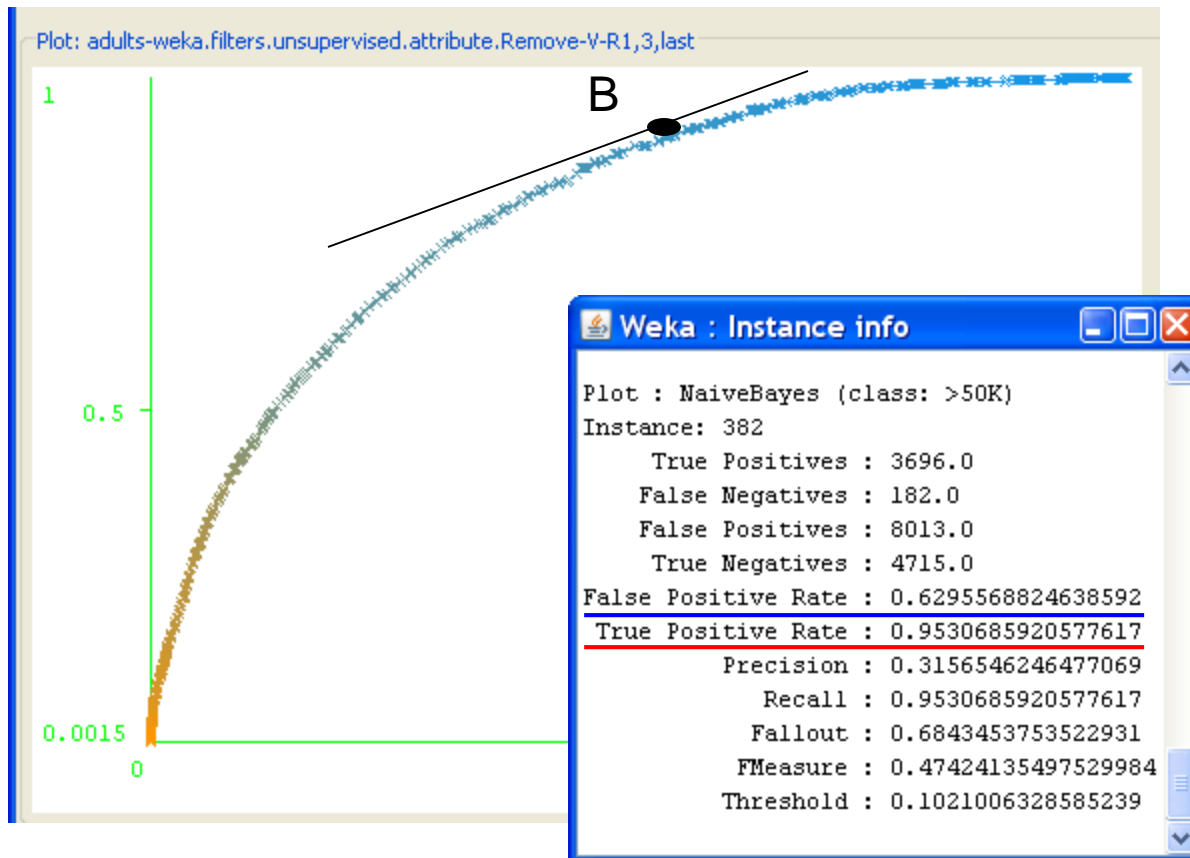


Is this threshold good :

for cancer detection?

for targeting potential customers?

Cost sensitive operating points



Is this threshold good :

for cancer detection?

for targeting potential customers?

Conclusions

- WEKA is a powerful datamining tool with the state-of-the art GUI, but is not very easy to use



- There are other open source data mining tools:
 - Orange:
 - <http://www.ailab.si/orange>
 - Tanagra:
 - <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>