ROC curves

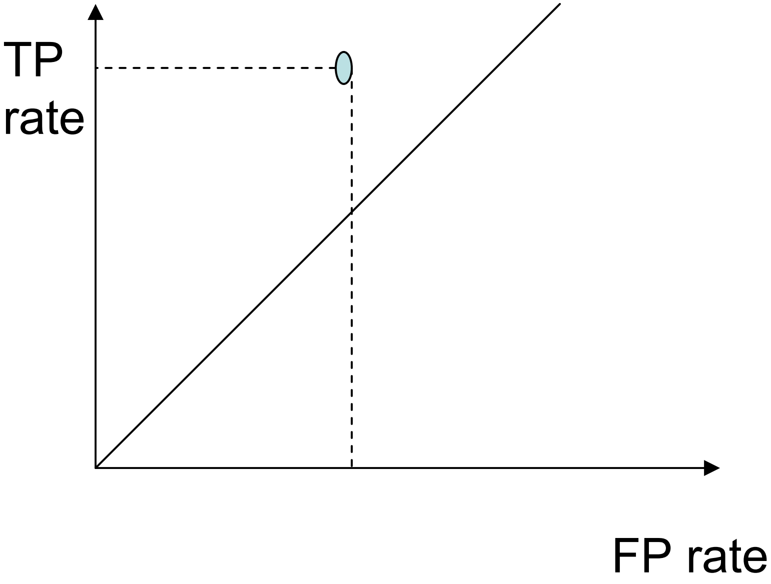
Data Mining Lab 5

Lab outline

- Remind what ROC curve is
- Generate ROC curves using WEKA
- Some usage of ROC curves

Point in ROC space

		TRUE CLASS	
		YES	NO
PREDICTED CLASS	YES	TP	FP
	NO	FN	TN
Total:		P	N



FP rate: FP/N TP rate: TP/P (recall)

FN rate: FN/N TN rate: TN/P

Classifier accuracy: (TP+TN)/(P+N)

Shows how good is classifier in discriminating positive instances from the negative ones

ROC curve of a probabilistic classifier

Naïve Bayes, for example, outputs the probability of an instance in a testing set to be classified as YES

Outlook	Temp	Windy	**P(Y|E)**	Real class
overcast	mild	yes	**0.95**	YES
rainy	mild	no	**0.80**	YES
rainy	cool	yes	**0.60**	NO
sunny	mild	no	**0.45**	YES
sunny	cool	no	**0.40**	NO
sunny	hot	no	**0.35**	NO
sunny	hot	yes	**0.25**	NO

ROC curve of a probabilistic classifier

In a general case, we classify an instance as YES if the probability is more than 50%

Classified as YES

Classified as NO

Outlook	Temp	Windy	**P(Y|E)**	Real class
overcast	mild	yes	**0.95**	YES
rainy	mild	no	**0.80**	YES
rainy	cool	yes	**0.60**	NO
sunny	mild	no	**0.45**	YES
sunny	cool	no	**0.40**	NO
sunny	hot	no	**0.35**	NO
sunny	hot	yes	**0.25**	NO

Operating threshold

ROC curve of a probabilistic classifier

We compute the confusion matrix

		TRUE CLASS	
		YES	NO
PREDICTED CLASS	YES	2 (TP)	1 (FP)
	NO	1 (FN)	3 (TN)
Total:		3 (P)	4 (N)

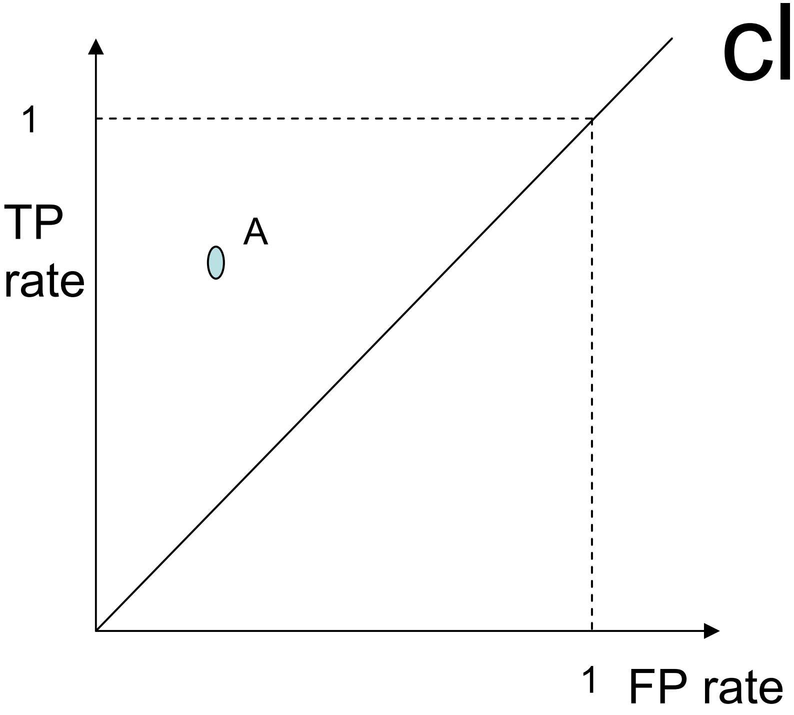
Outlook	Temp	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

And the TP and FP rates:

TP rate: TP/P=2/3≈0.7

FP rate: FP/N=1/4=0.25

ROC curve of a probabilistic classifier



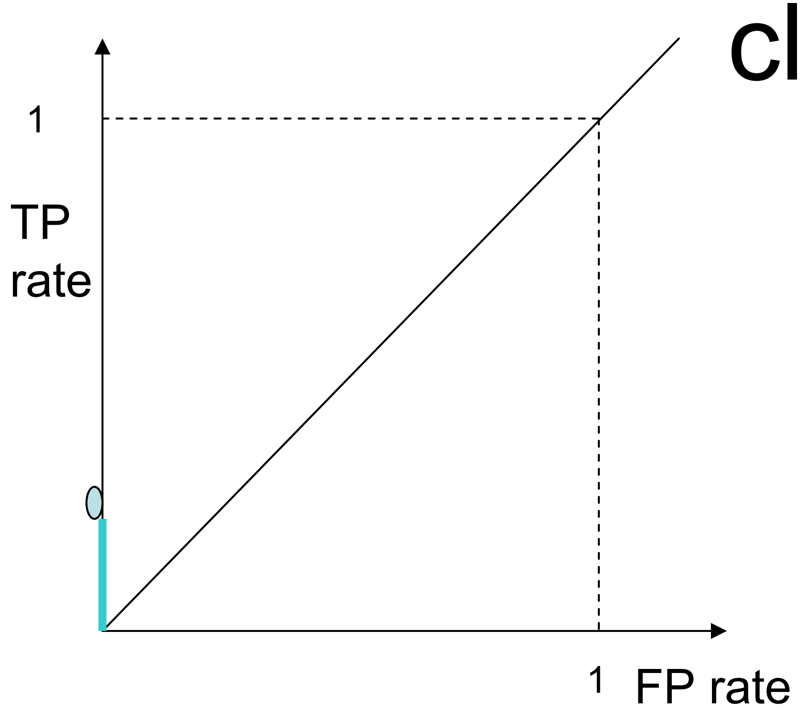
Outlook	Temp	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

This corresponds to point A in a ROC space

FP rate: FP/N=1/4=0.25

TP rate: TP/P=2/3≈0.7

ROC curve of a probabilistic classifier



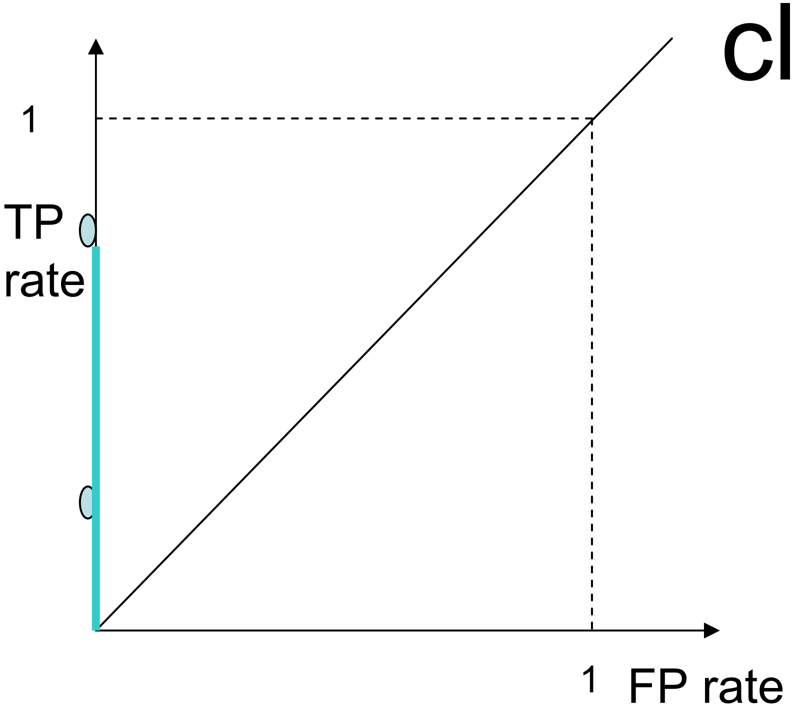
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: FP/N=0/4=0

TP rate: TP/P=1/3≈0.3

ROC curve of a probabilistic classifier



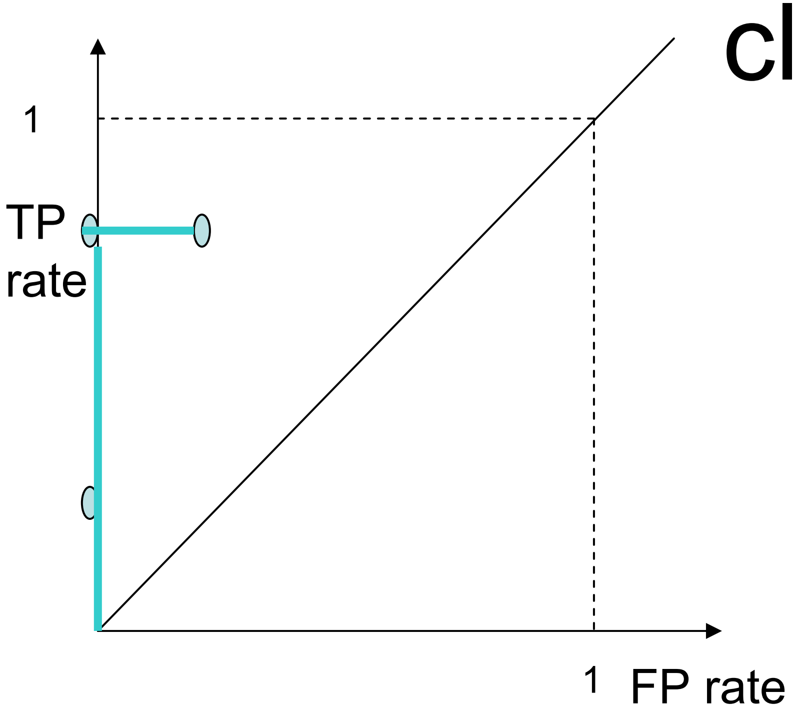
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: FP/N=0/4=0

TP rate: TP/P=2/3≈0.7

ROC curve of a probabilistic classifier



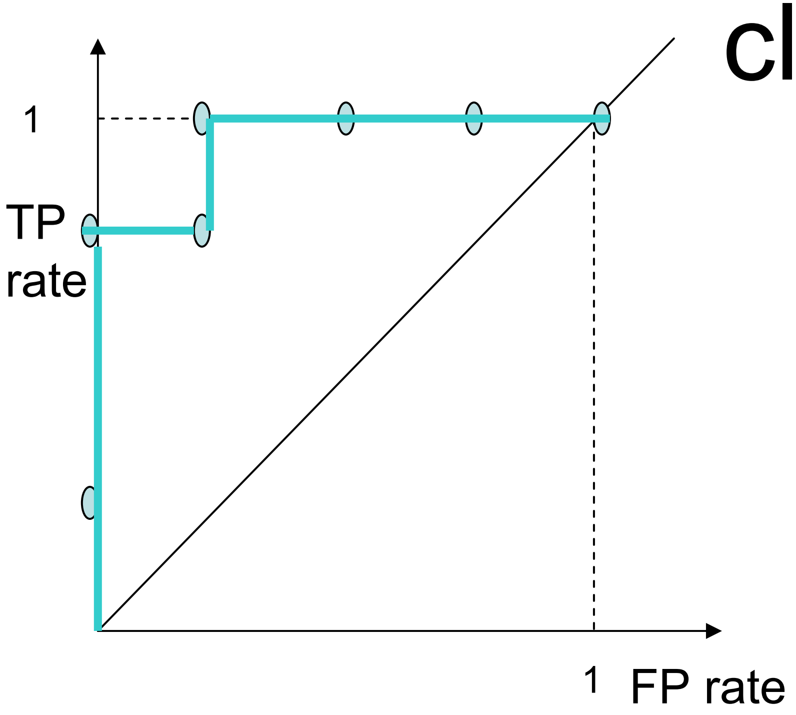
For different threshold values we get different points in the ROC space

Outlook	Temp	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	NO	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: FP/N=1/4=0.25

TP rate: TP/P=2/3≈0.7

ROC curve of a probabilistic classifier



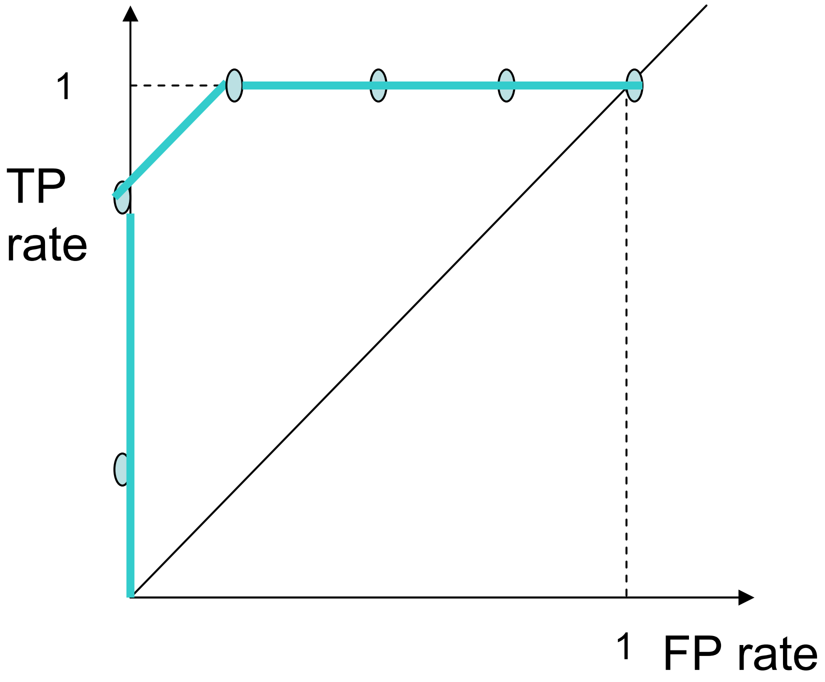
For different threshold values we get different points in the ROC space

Outlook	Tem p	Windy	P(Y|E)	Predicted class	Real class
overcast	mild	yes	0.95	YES	YES
rainy	mild	no	0.80	YES	YES
rainy	cool	yes	0.60	YES	NO
sunny	mild	no	0.45	YES	YES
sunny	cool	no	0.40	NO	NO
sunny	hot	no	0.35	NO	NO
sunny	hot	yes	0.25	NO	NO

FP rate: FP/N=1/4=0.25

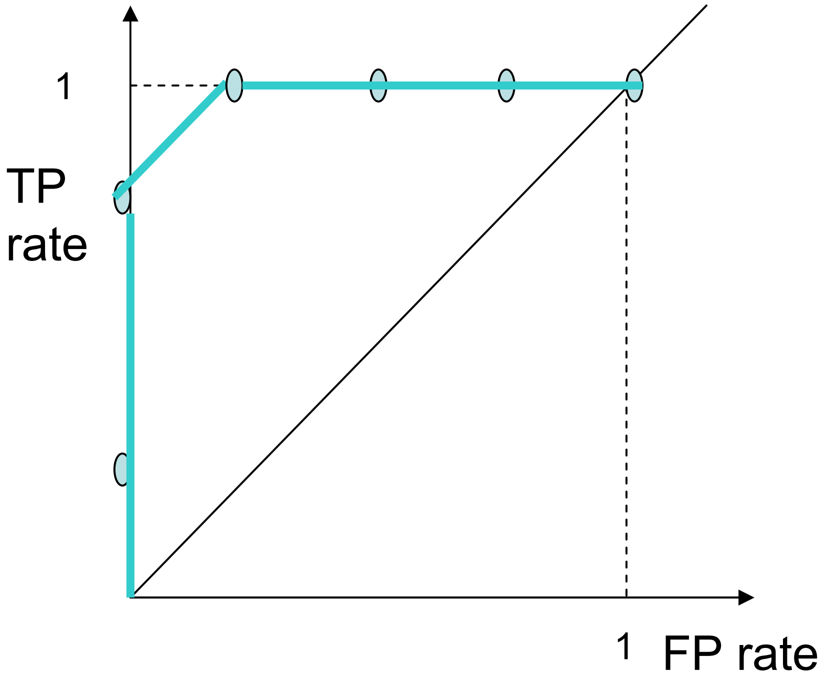
TP rate: TP/P=3/3=1.0, etc…

ROC curve of a probabilistic classifier

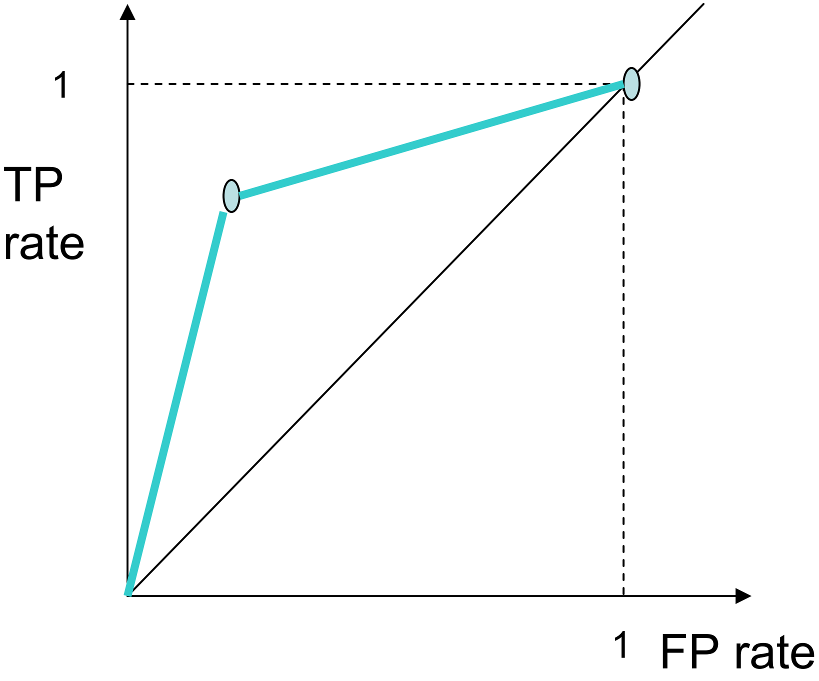


At the end we get the ROC curve
for Naïve Bayes classifier

ROC curve of a probabilistic classifier vs discrete classifier



ROC curve for Naïve Bayes classifier (probabilistic)



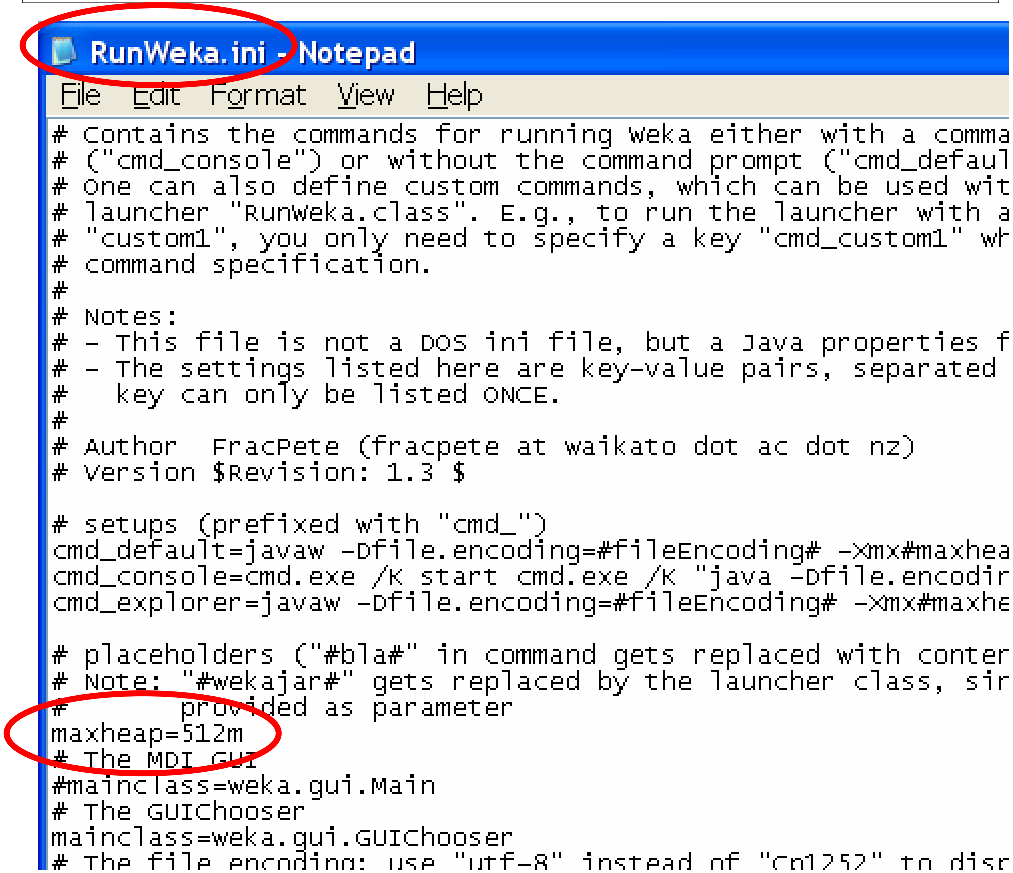
ROC curve for Decision Tree classifier (discrete)

Lab outline

- Remind what ROC curve is
- Generate ROC curves using WEKA
- Some usage of ROC curves

Preparation

Step 1. Increase Java heap size



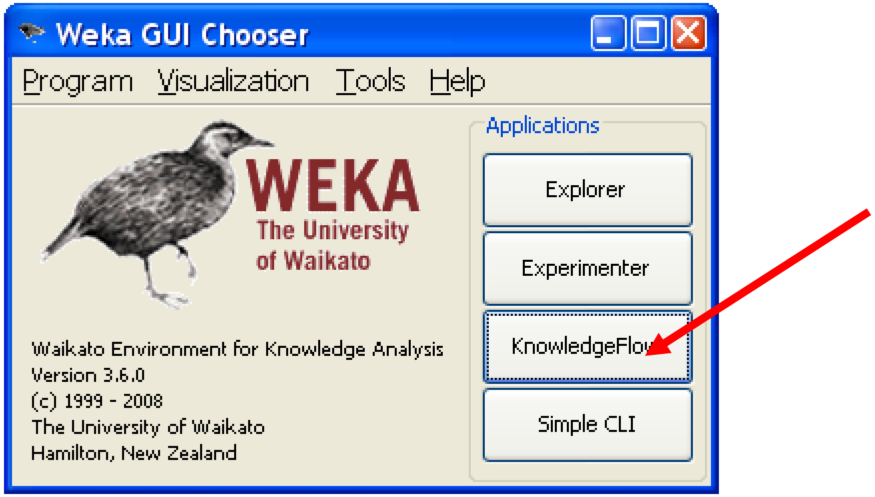
```
# Contains the commands for running Weka either with a comma
# ("cmd_console") or without the command prompt ("cmd_defaul
# One can also define custom commands, which can be used wit
# launcher "RunWeka.class". E.g., to run the launcher with a
# "custom1", you only need to specify a key "cmd_custom1" wh
# 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 "Cp1252" to disr
```

Step 2. Download input data file

adult_income.arff

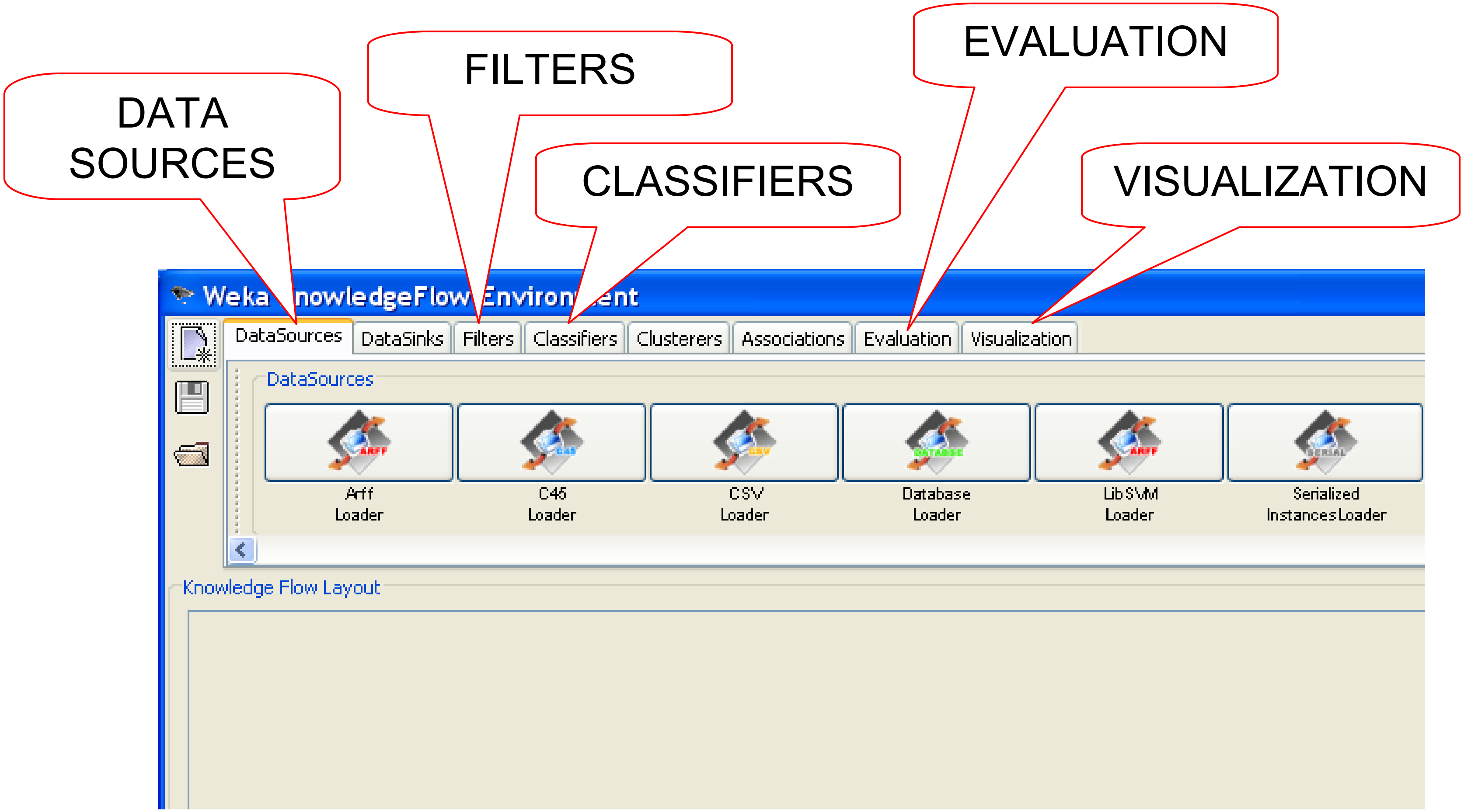
into your home directory

Comparing classifiers. Knowledge flow



Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand

Knowledge flow tabs



DATA SOURCES

FILTERS

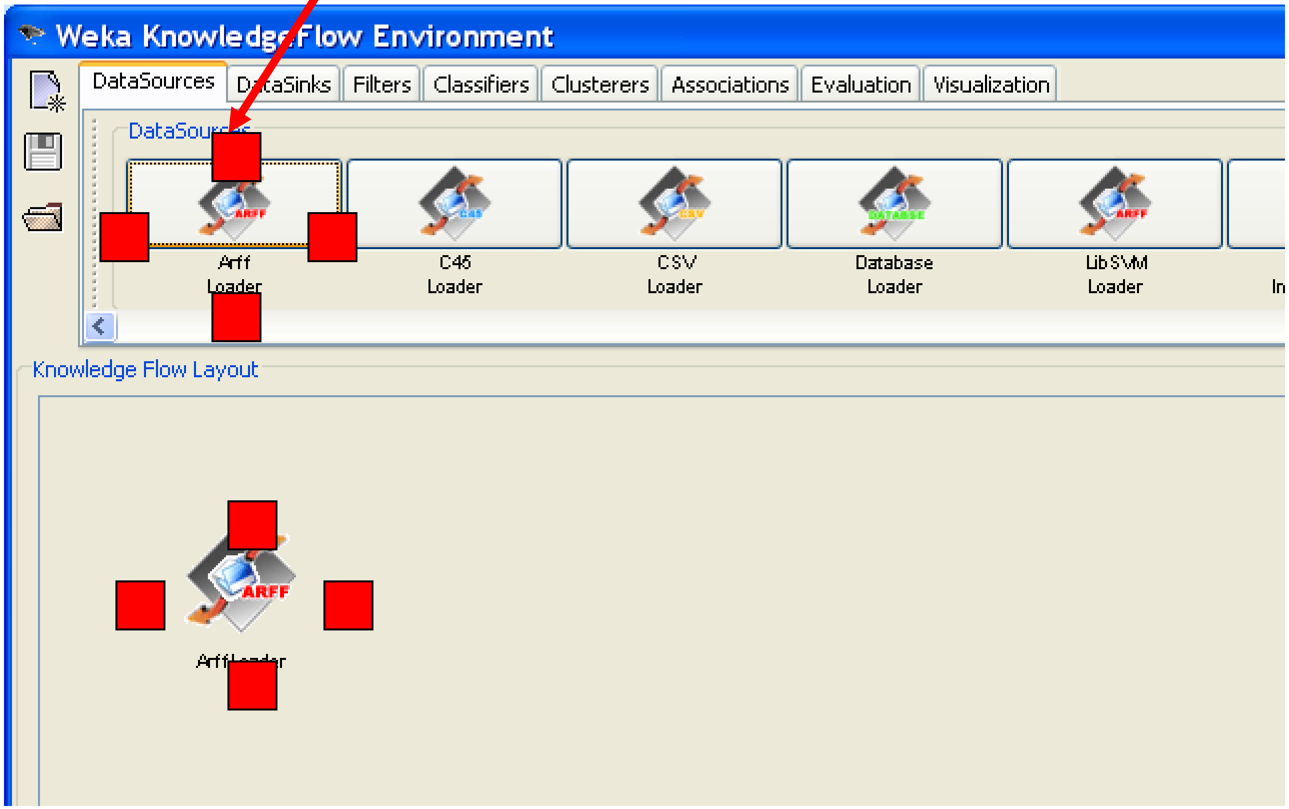
CLASSIFIERS

EVALUATION

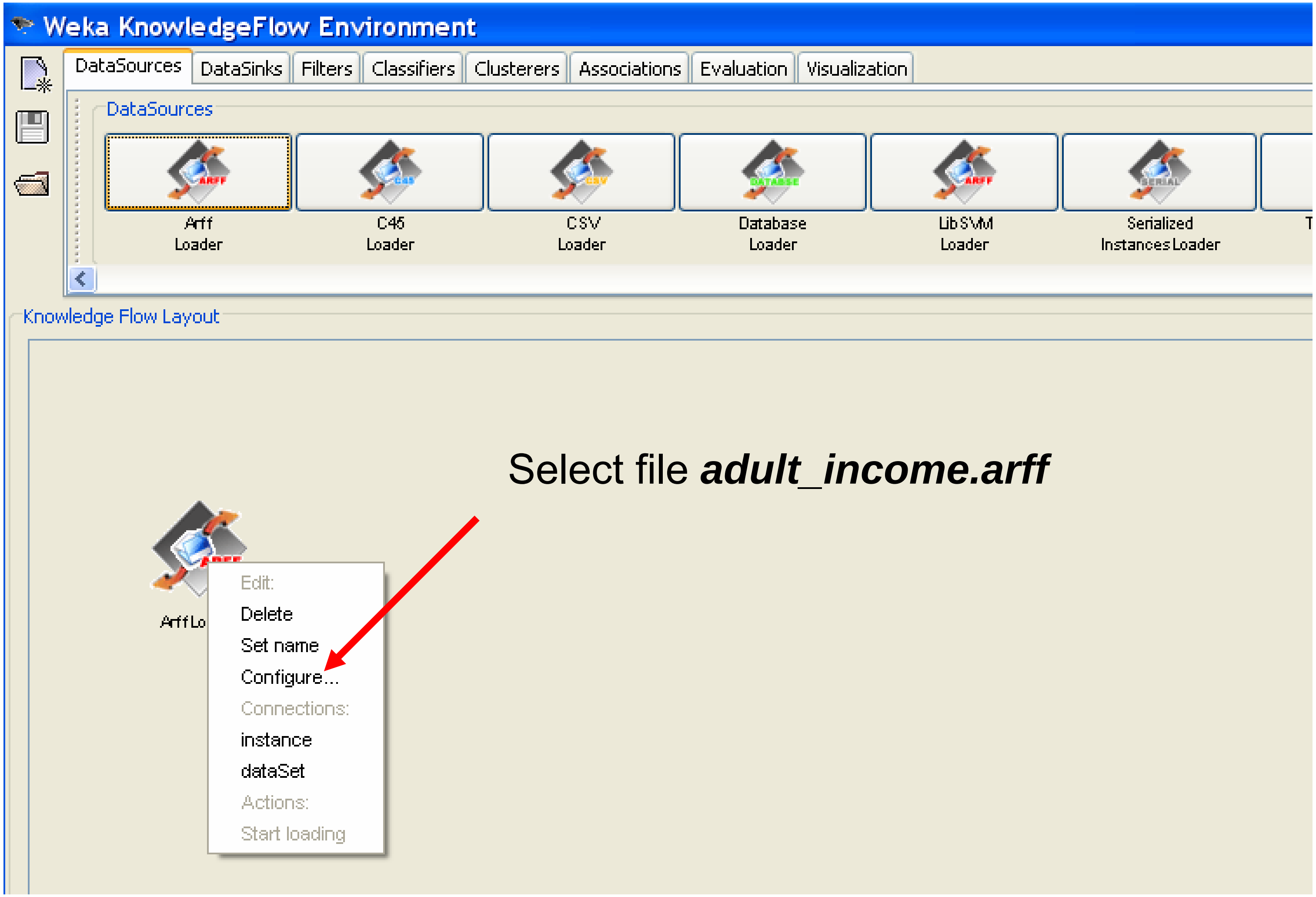
VISUALIZATION

Loading the data

Click



Loading the data



Select file ***adult_income.arff***

Data file ***adult_income.arff***

@relation adults

1. @attribute age numeric
2. @attribute workclass {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
3. @attribute education real
4. @attribute marital_status {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}
5. @attribute occupation {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}
6. @attribute sex {Male, Female}
7. @attribute native_country {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands}
8. @attribute class {>50K, <=50K}

Data from US census

Attributes of interest:

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

```
1.    @attribute age numeric
3.    @attribute education real
last   @attribute class {>50K, <=50K}
```

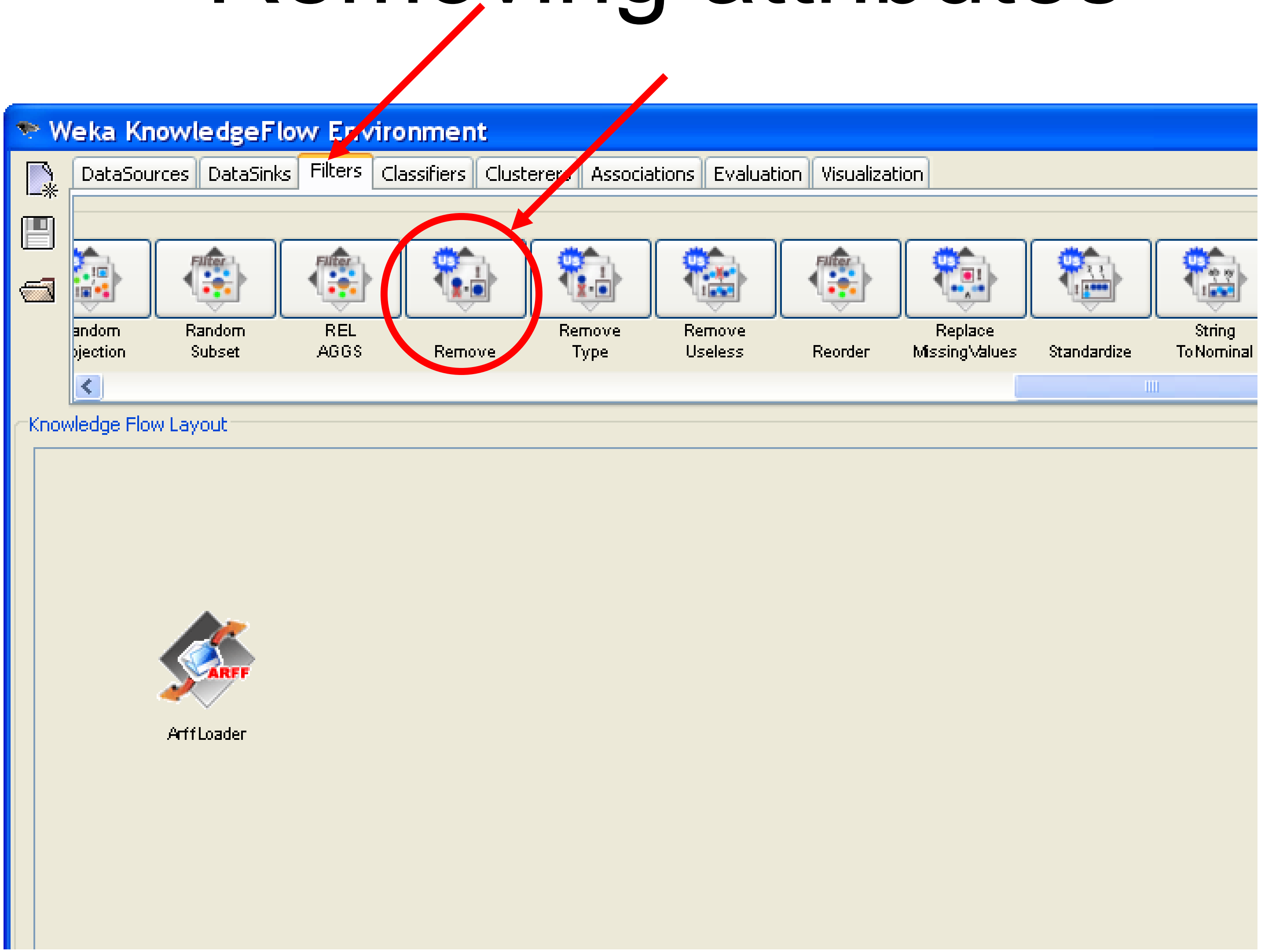
Numeric codes for education levels

Doctorate	16
Prof-school	15
Masters	14
Bachelors	13
Assoc-acdm	12
Assoc-voc	11
Some-college	10
HS-grad	9
11th	-7
12th	8
10th	6
9th	5
7th-8th	4
5th-6th	3
1st-4th	2
Preschool	1

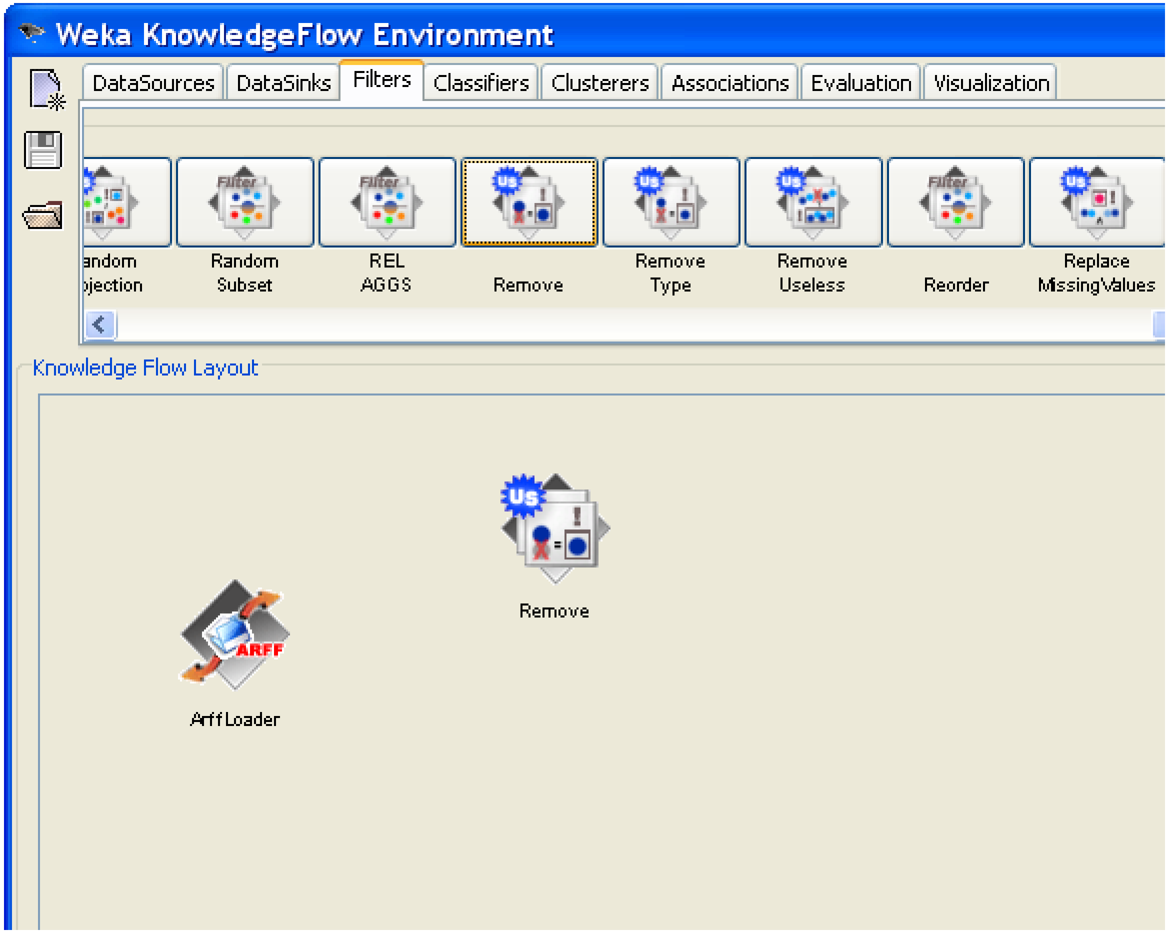
We remove all other attributes and leave only attributes 1,3, last – for simplicity

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

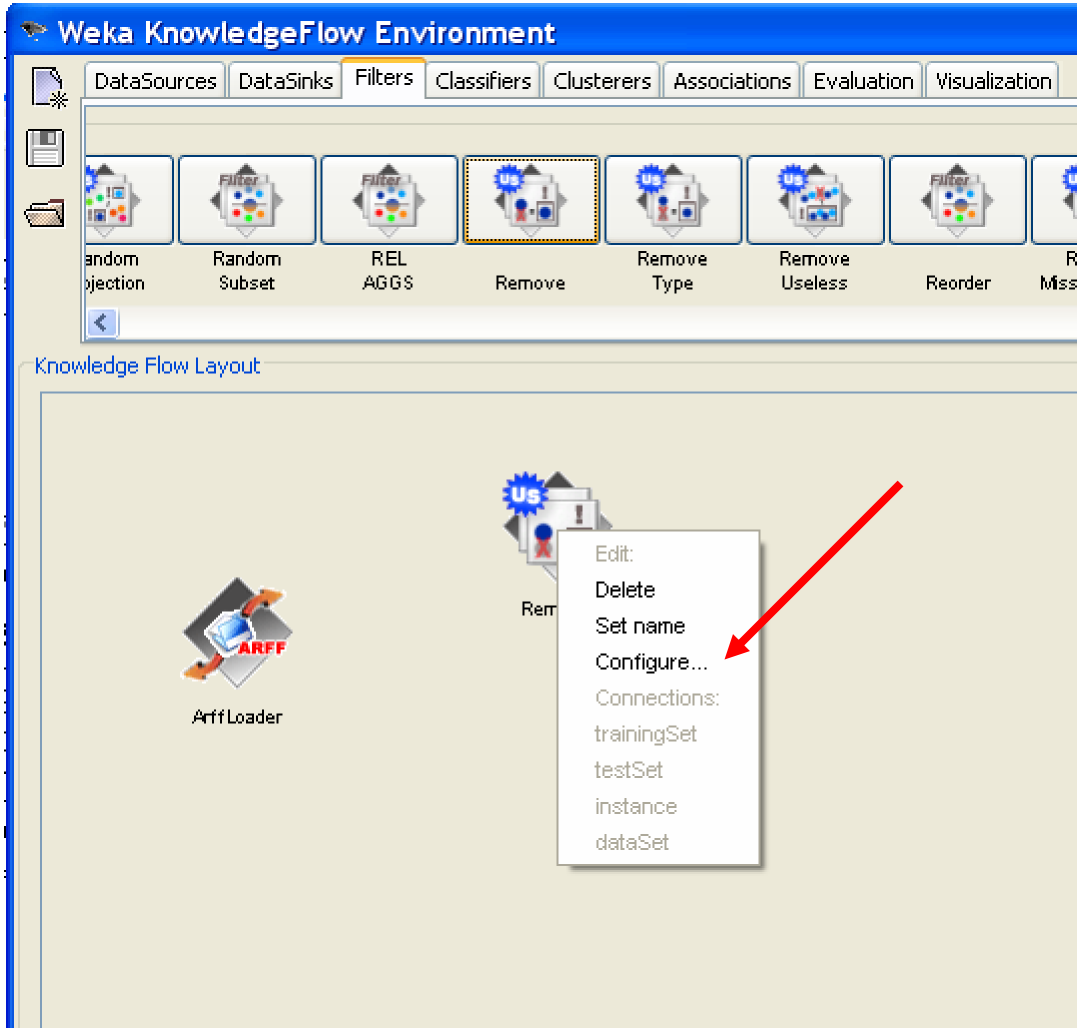
Removing attributes



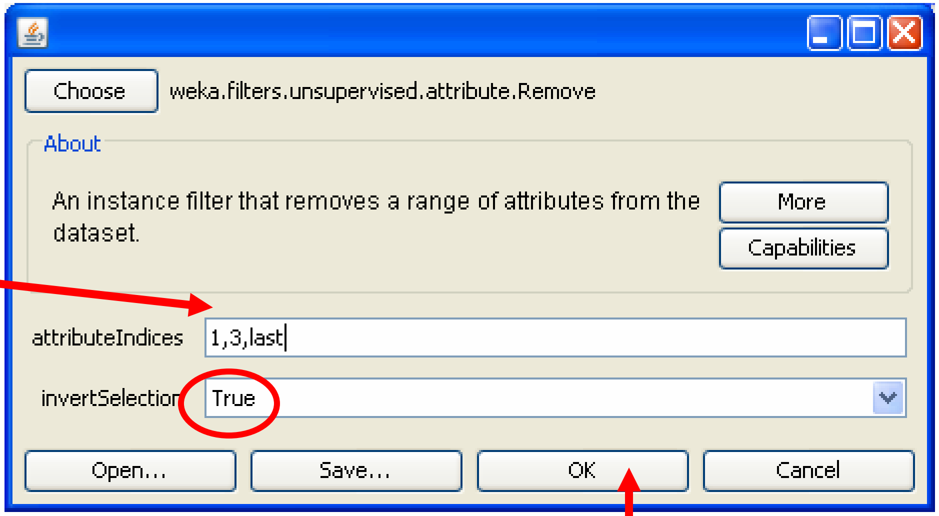
Removing attributes



Removing attributes



Removing attributes



Choose weka.filters.unsupervised.attribute.Remove

About

An instance filter that removes a range of attributes from the dataset.

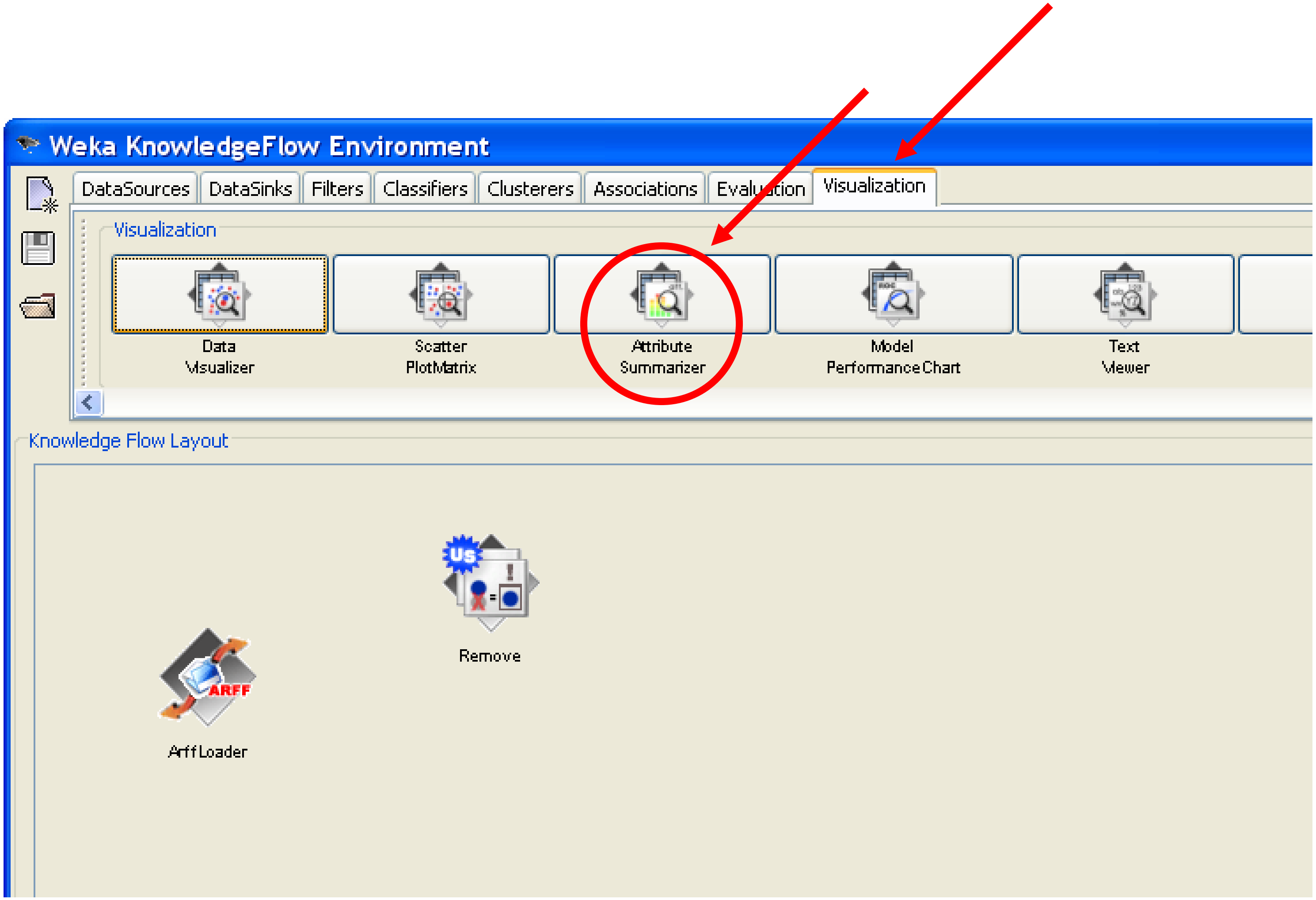
attributeIndices 1,3,last

invertSelection True

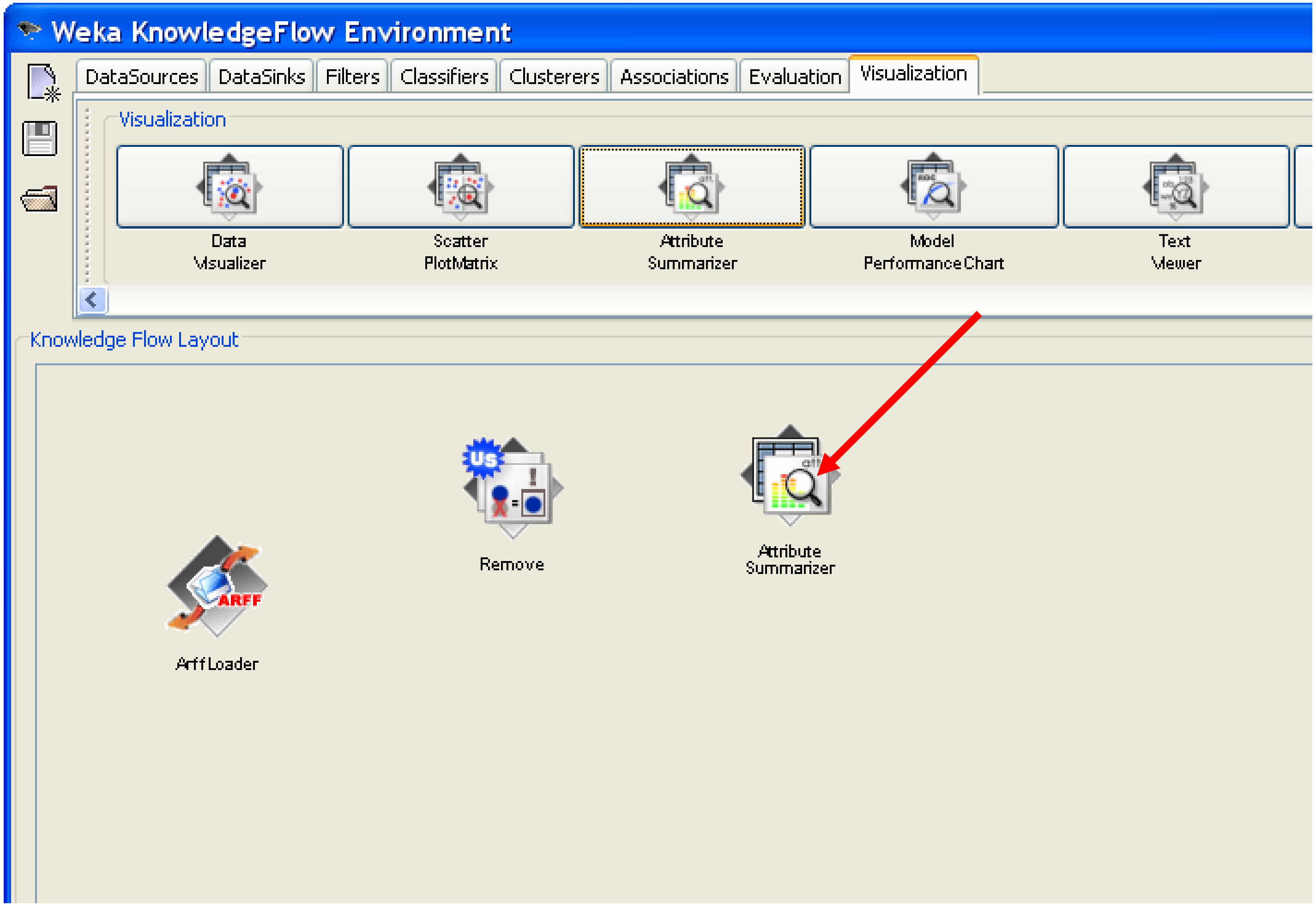
Type

Means remove all except attributes 1,3,last

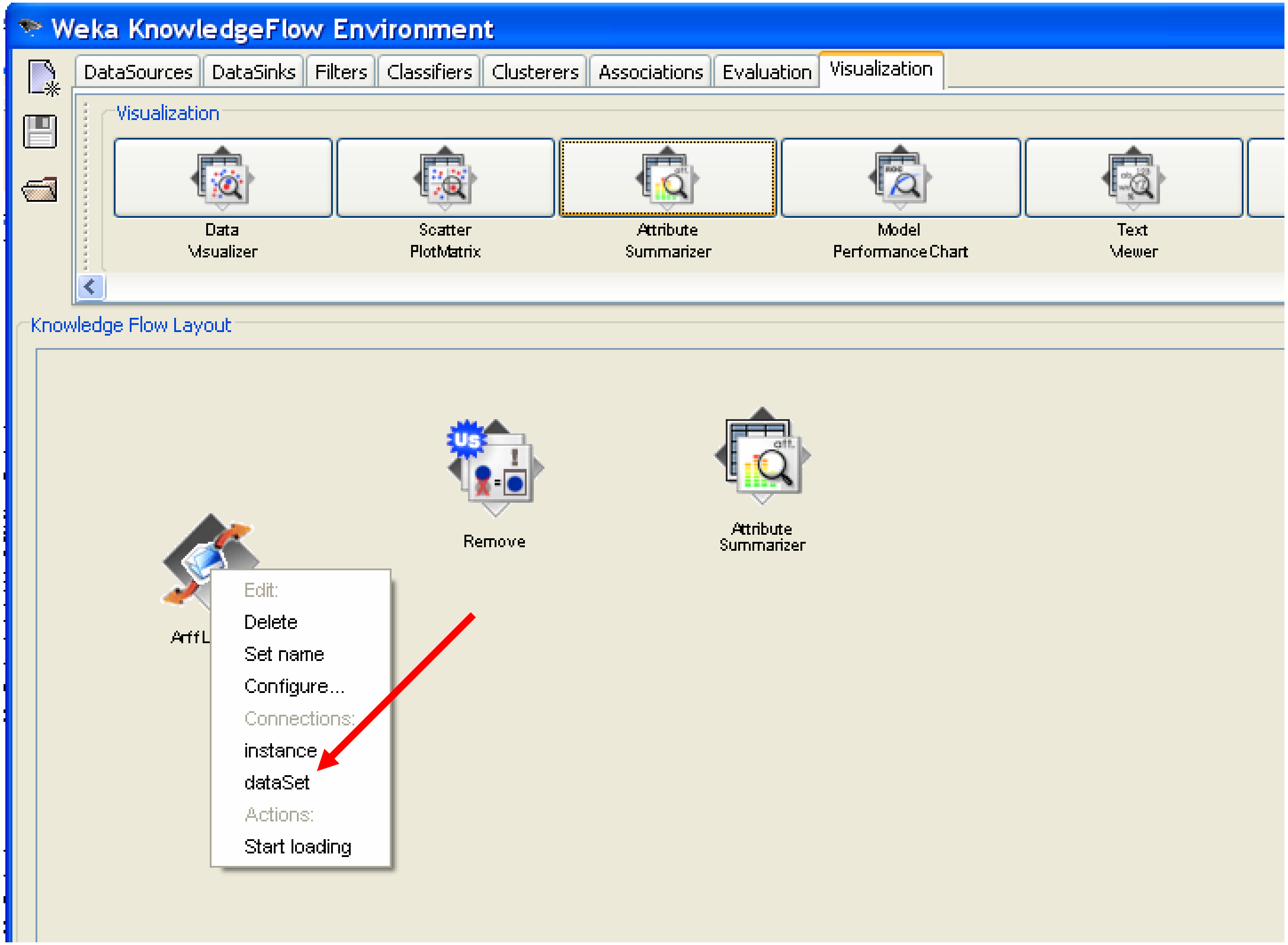
Visualize data



Visualize data

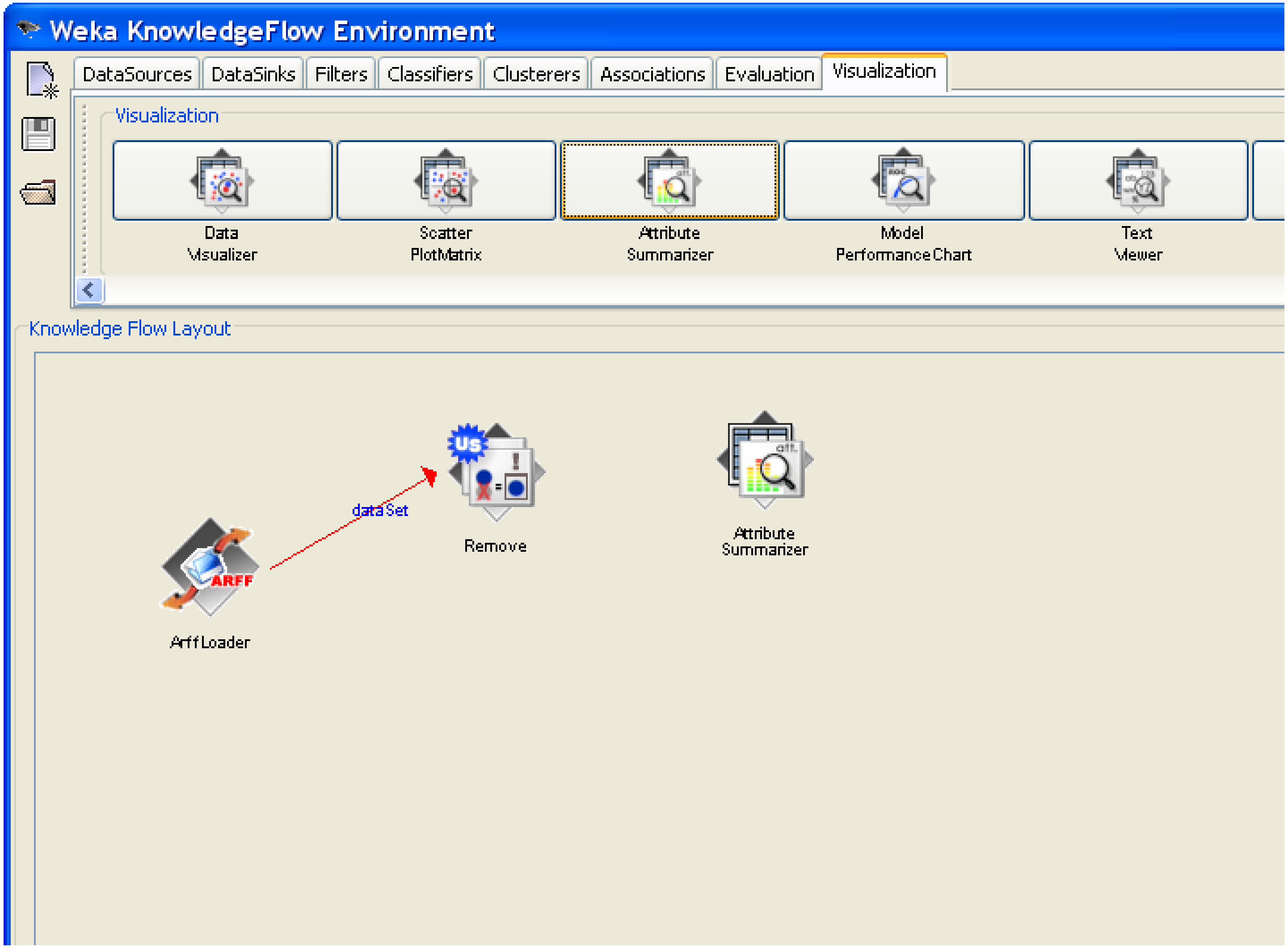


Connect the flow



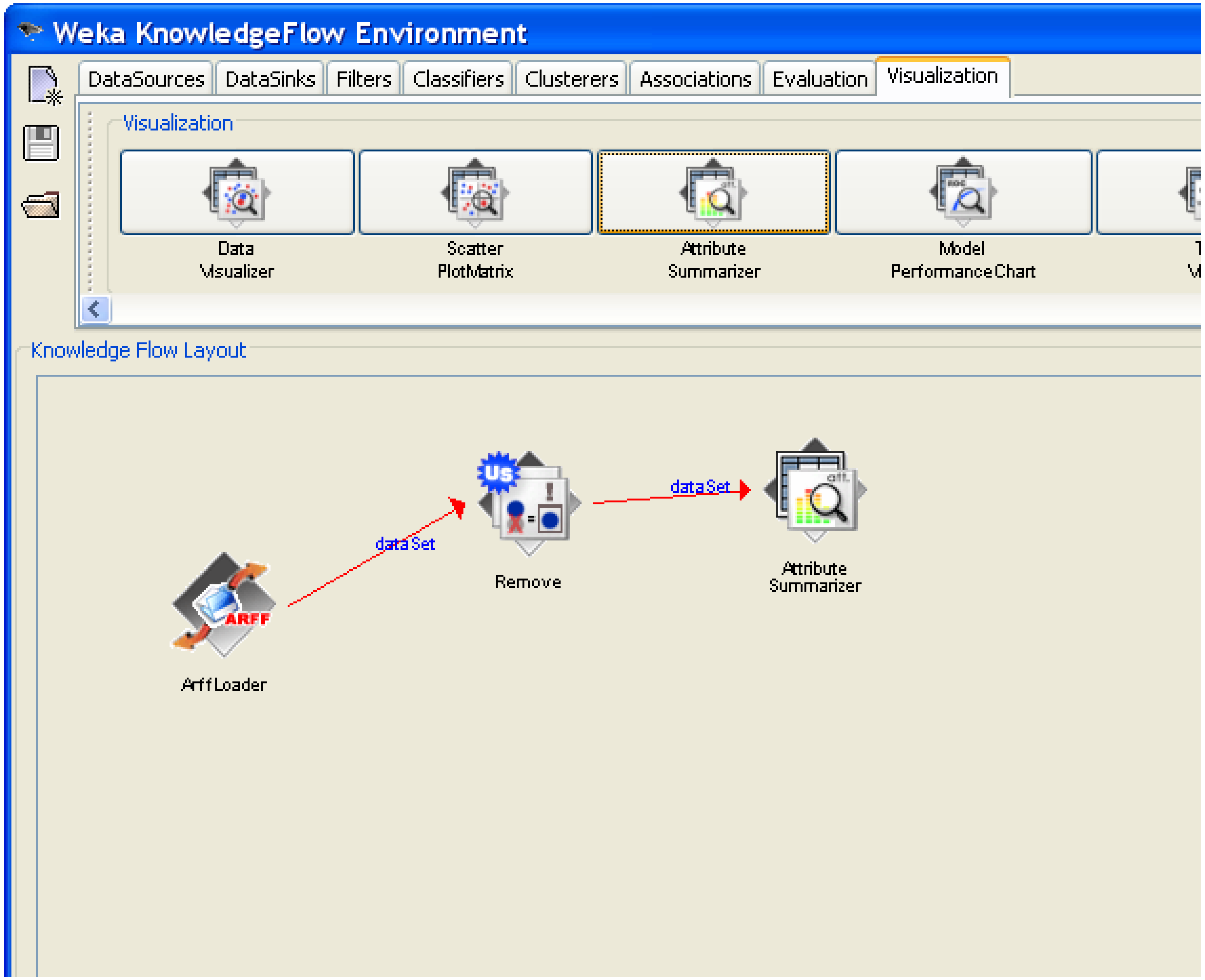
Connect the flow:

from data loader to attribute remover

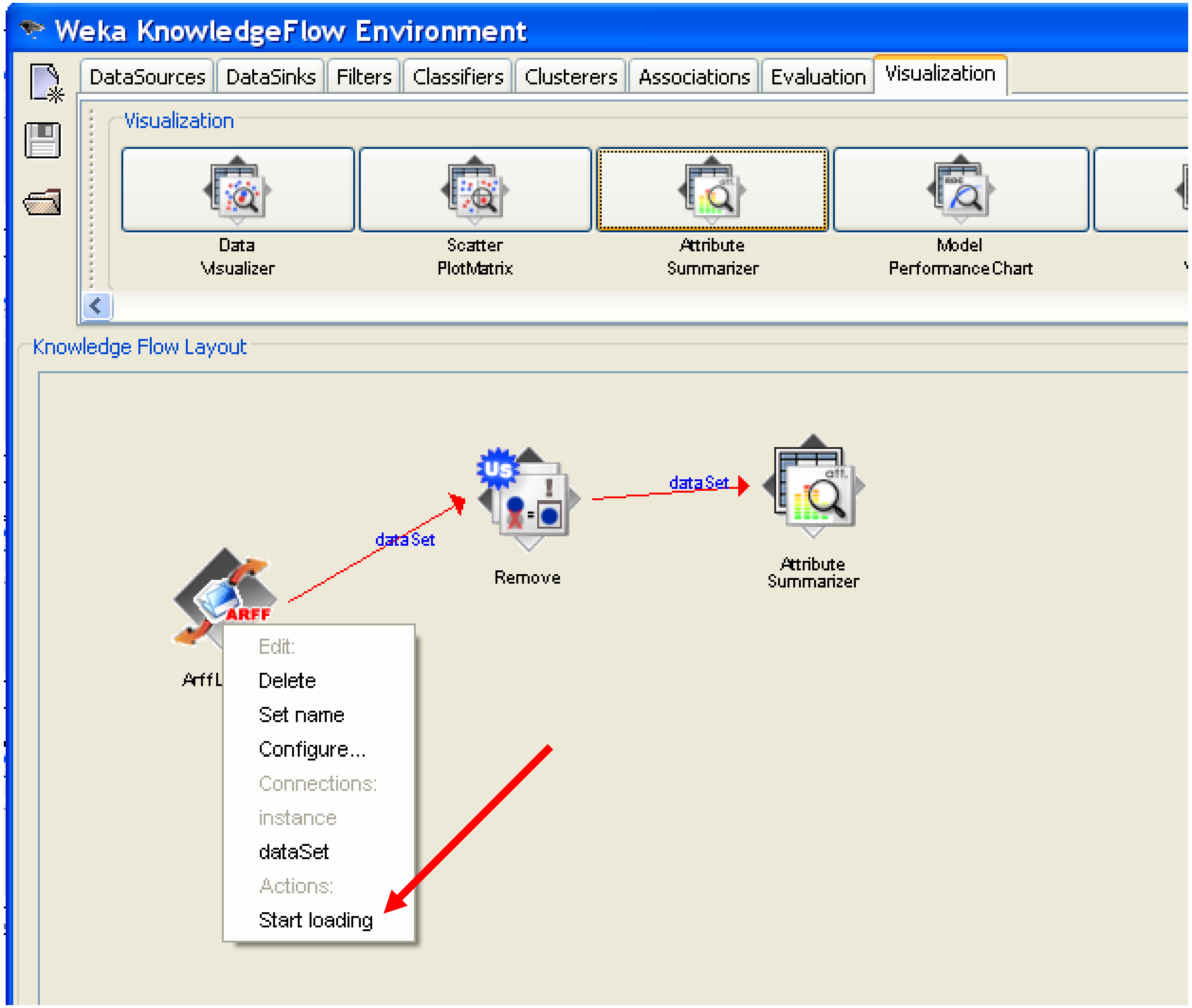


Connect the flow:

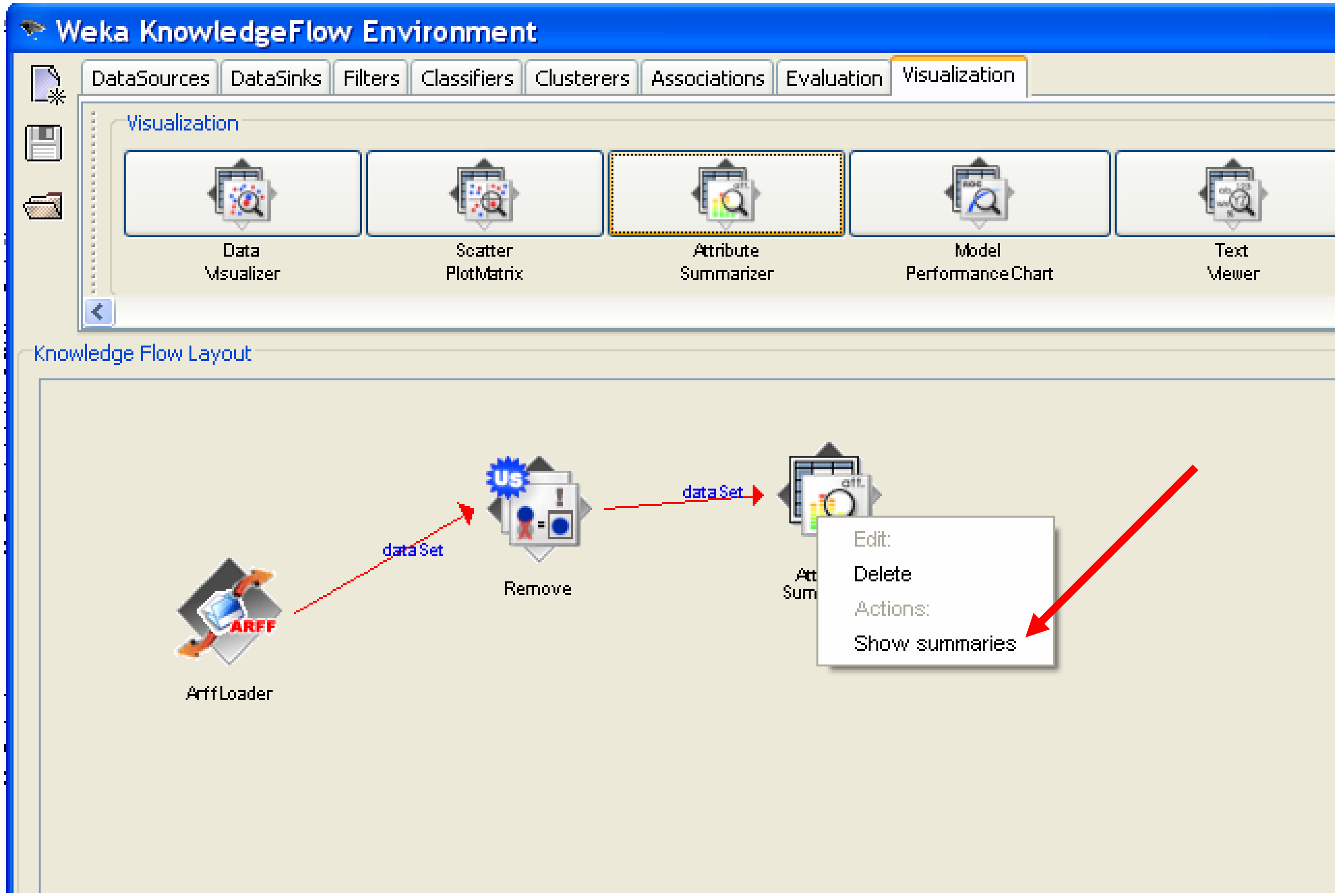
from attribute remover to summarizer



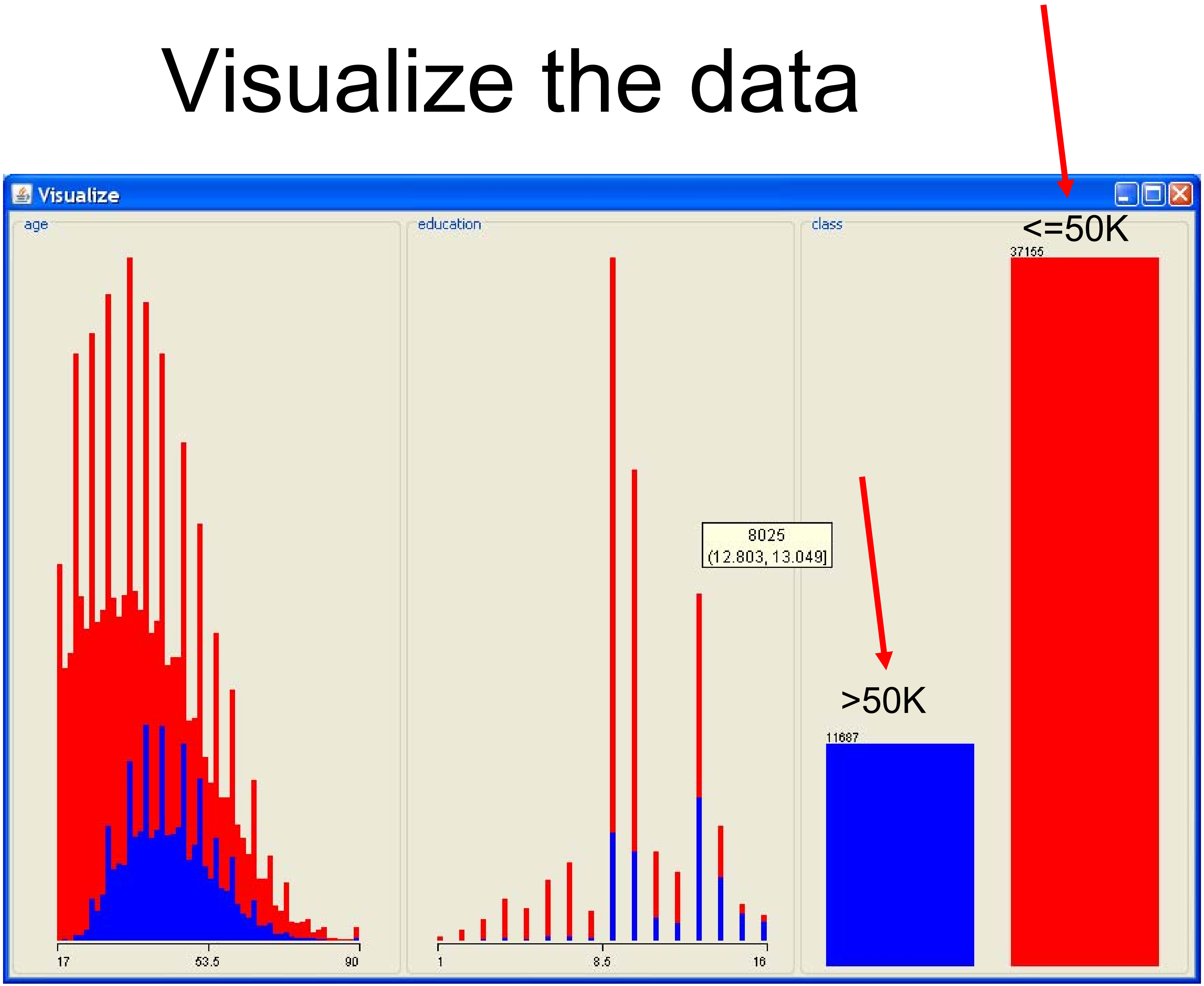
Start data flow



Visualize the data



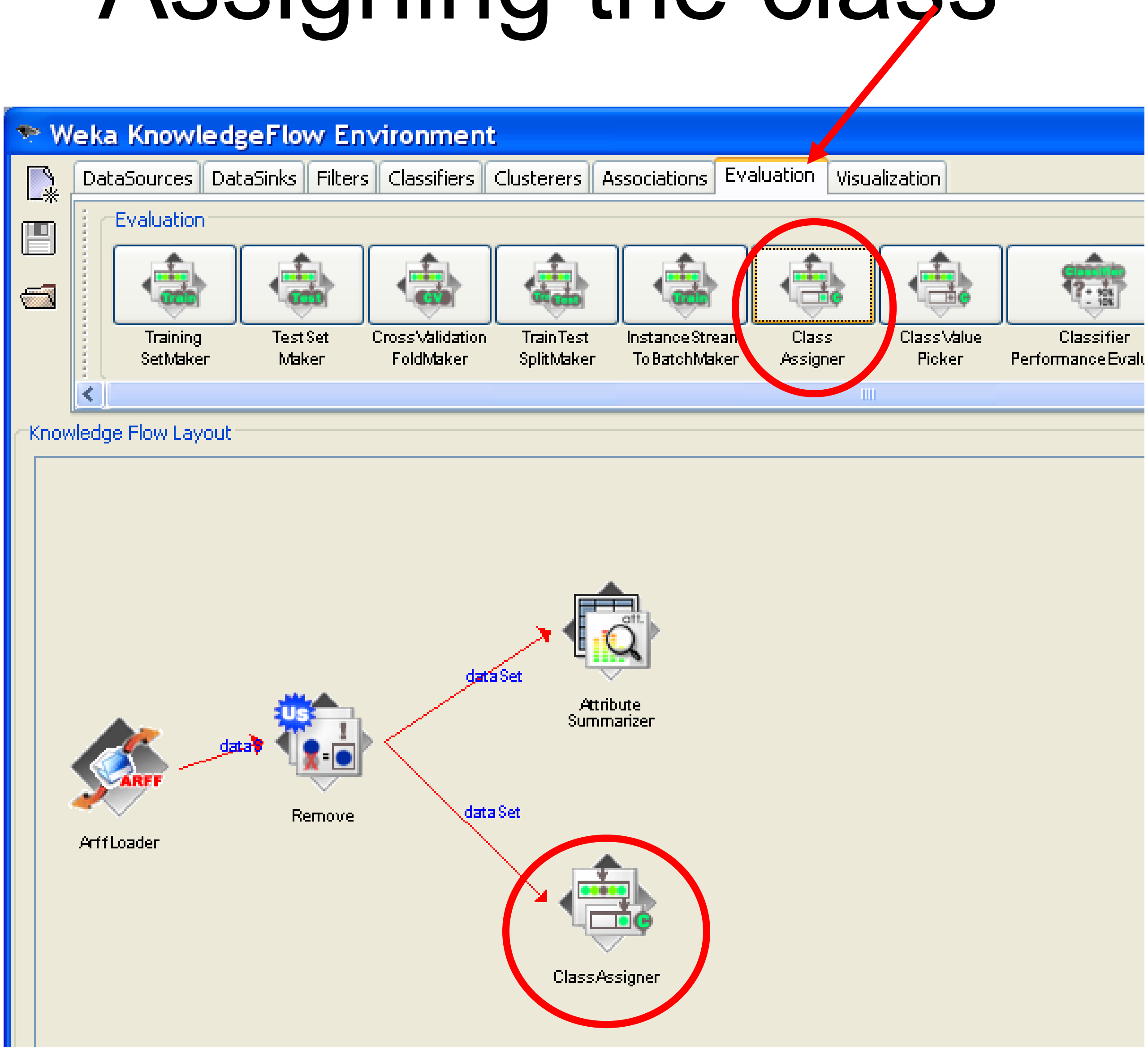
Visualize the data



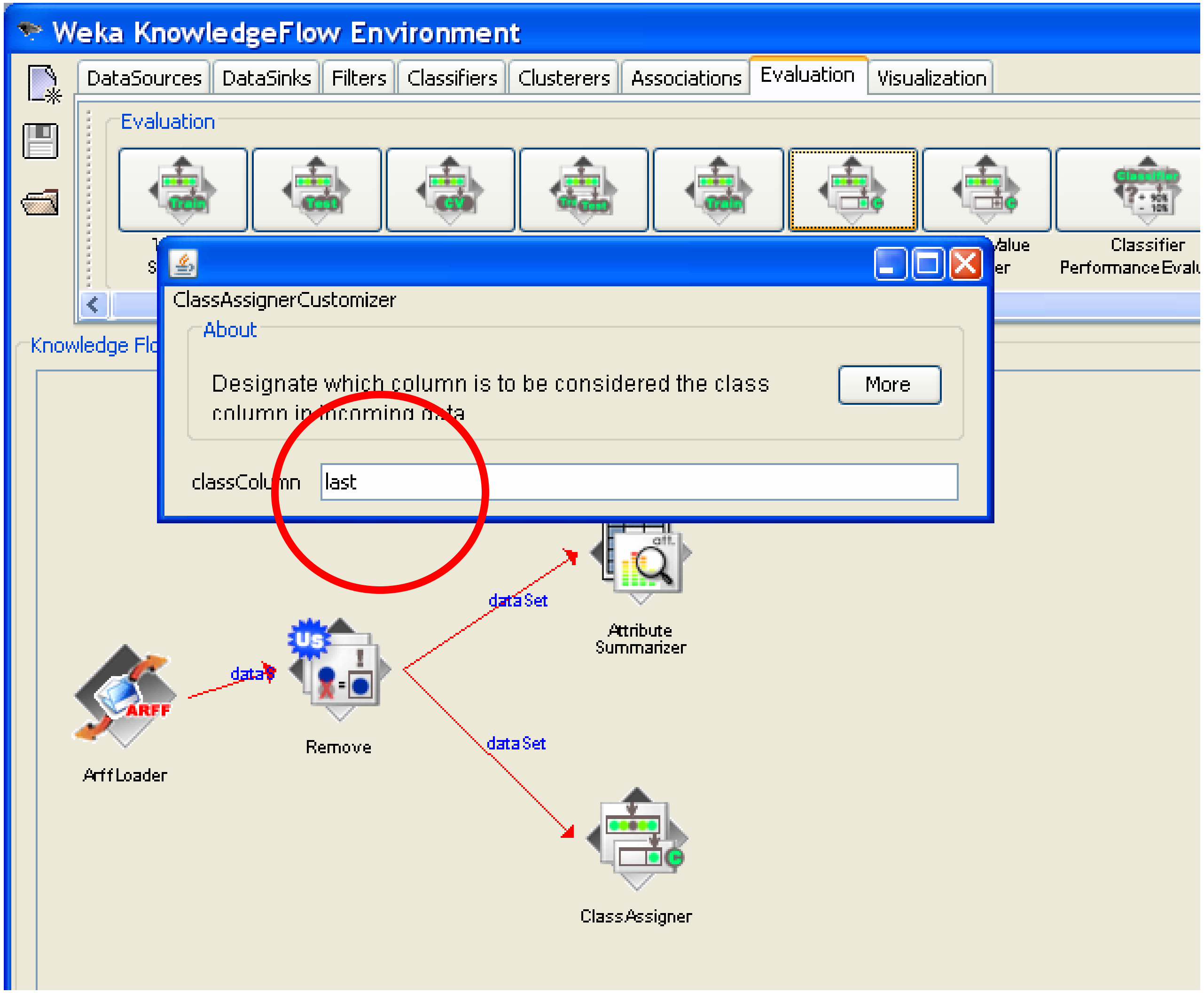
<=50K

>50K

Assigning the class



Configuring class assigner



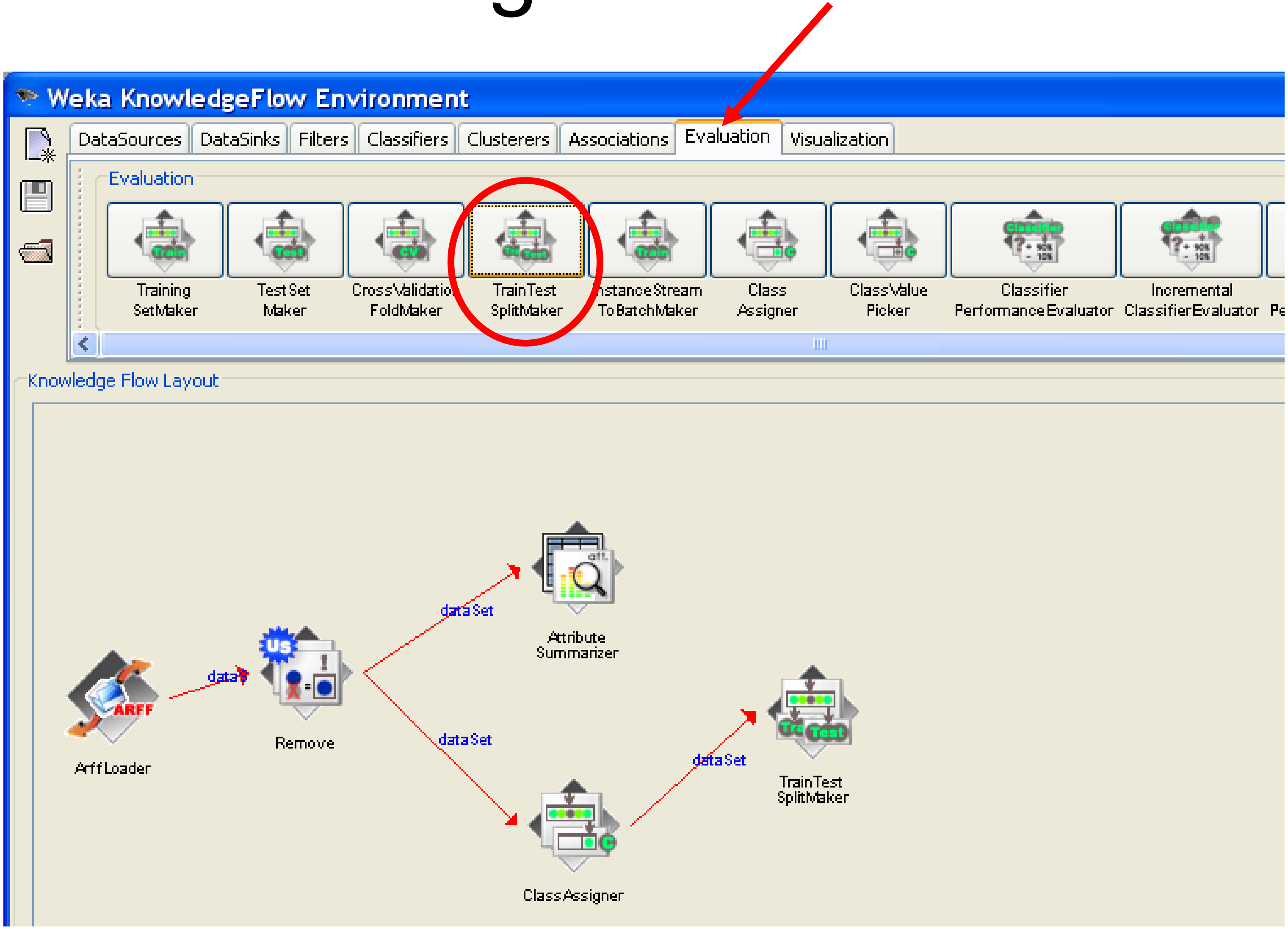
ClassAssignerCustomizer

About

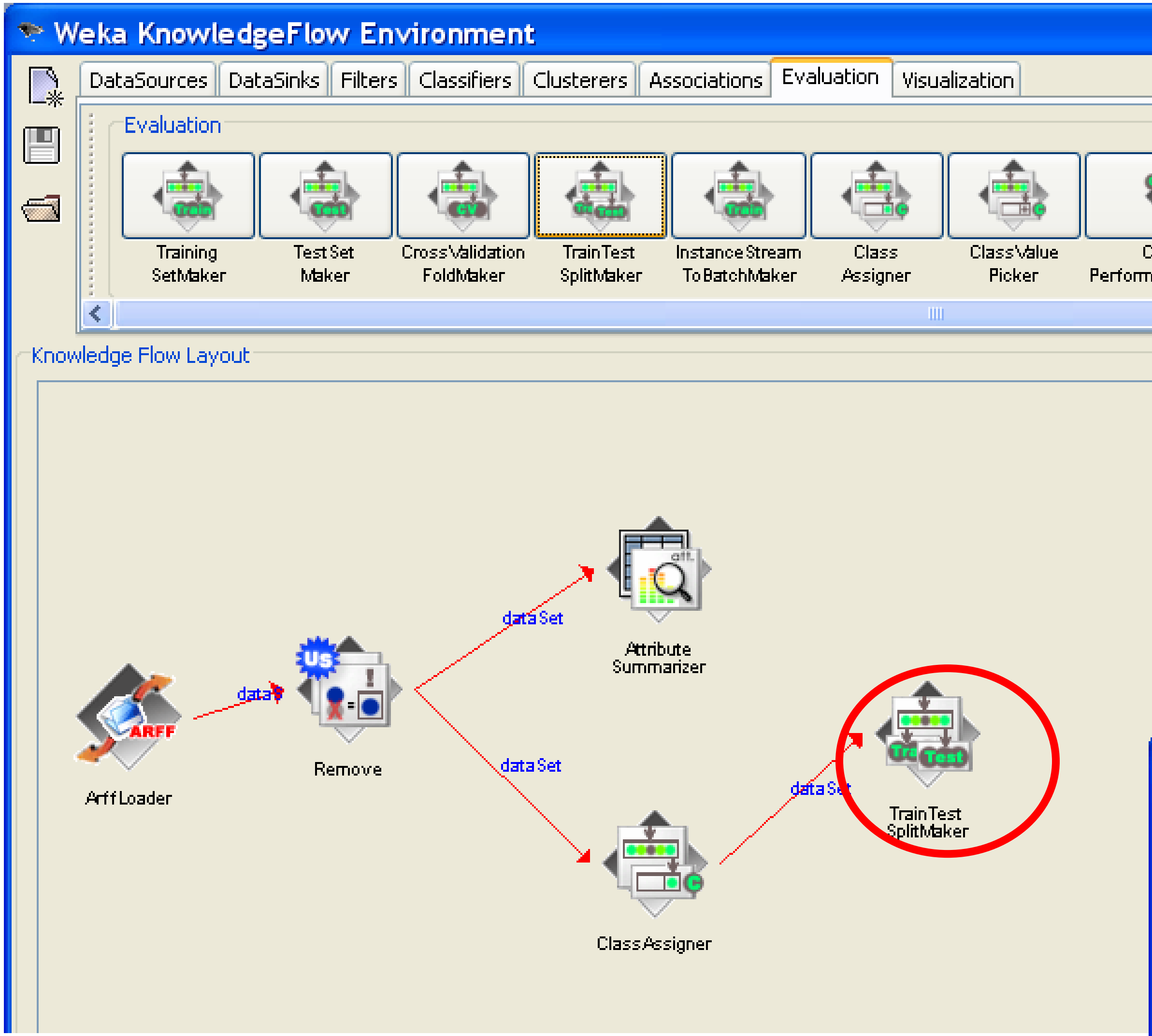
Designate which column is to be considered the class column in incoming data

classColumn: last

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



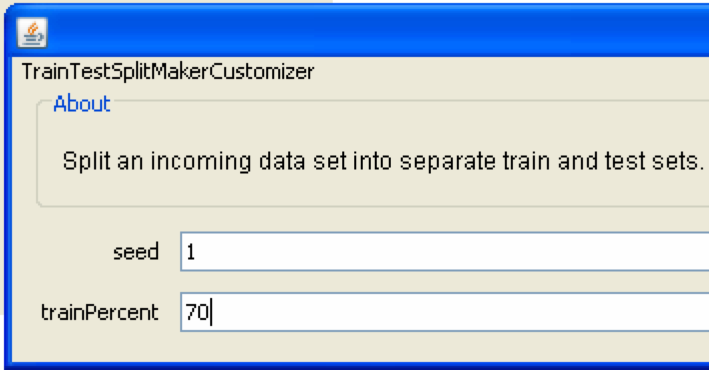
Subdivision of the dataset into “learning” and “test” set



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

and compute the ROC curve on the remaining.

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

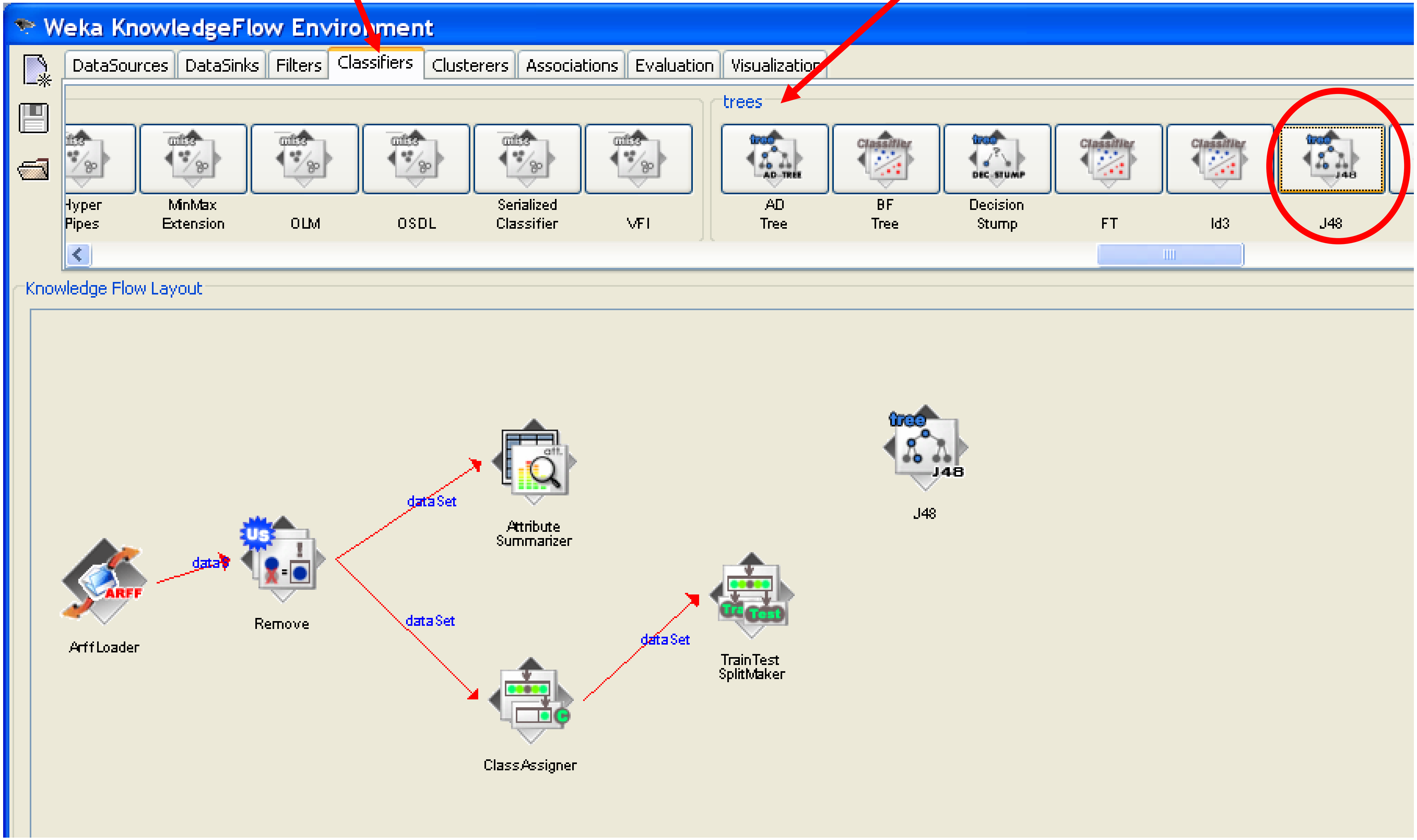


Split an incoming data set into separate train and test sets.

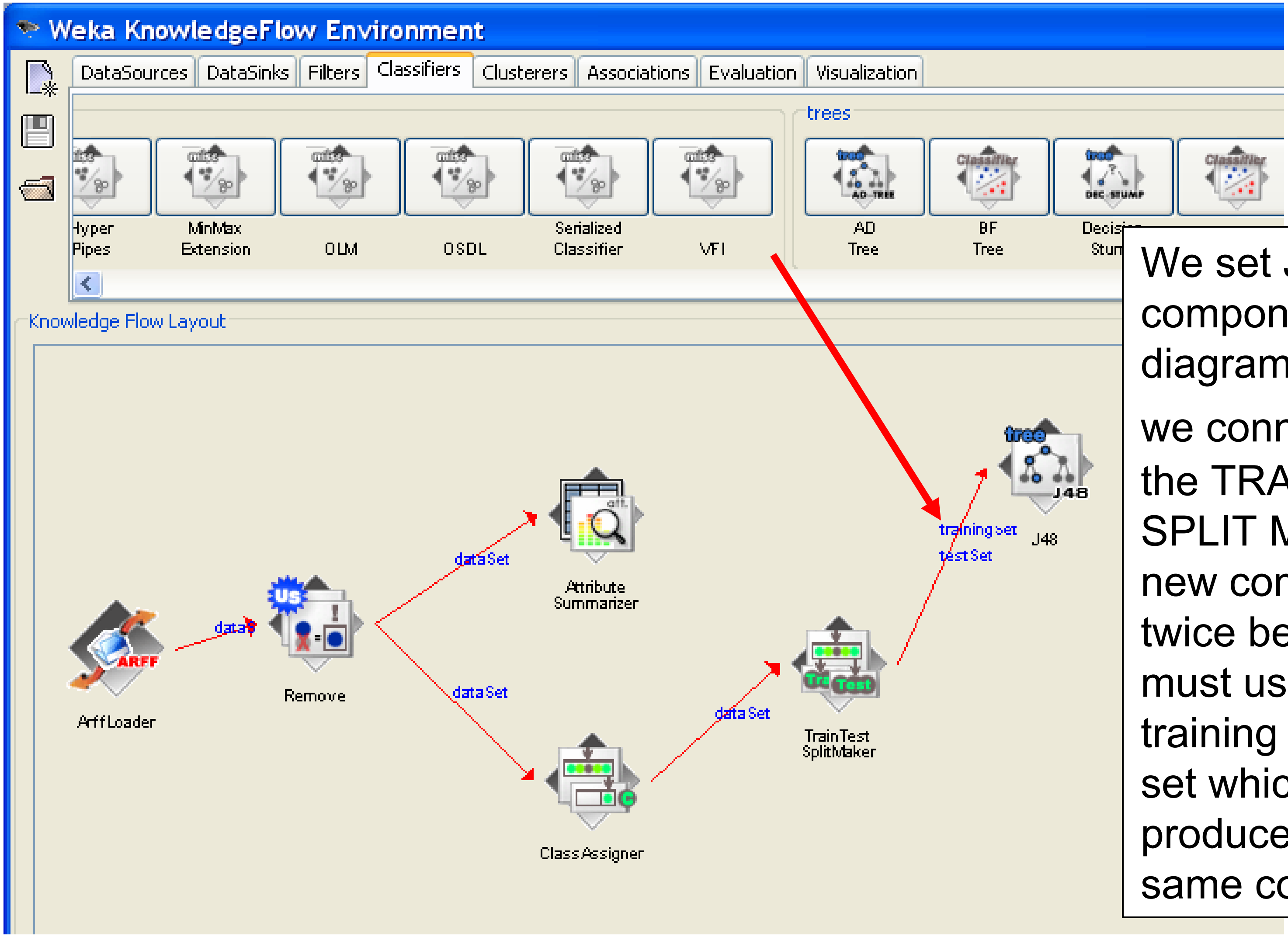
seed 1

trainPercent 70

Choosing discrete classifier – decision tree



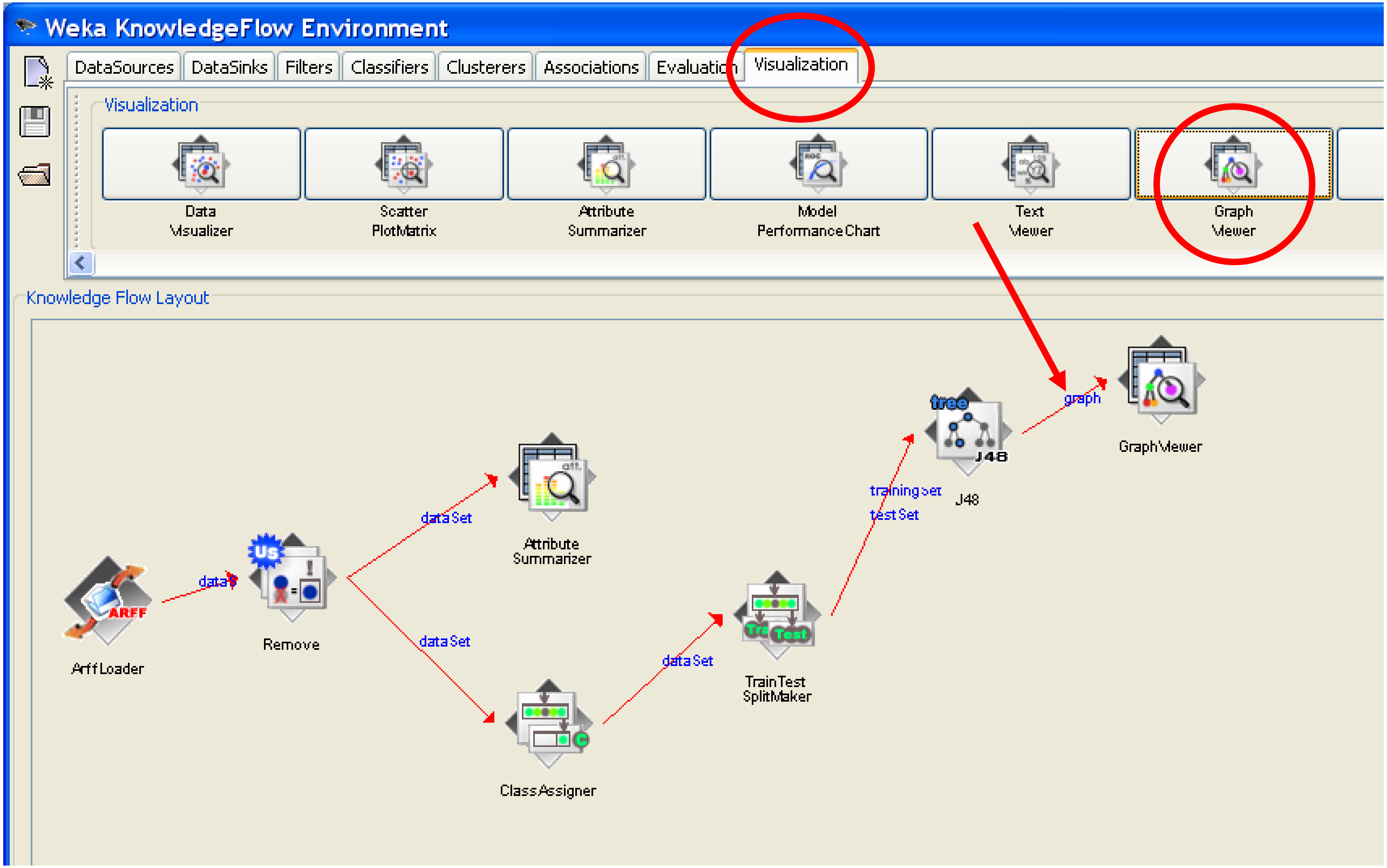
Connecting classifier to the data



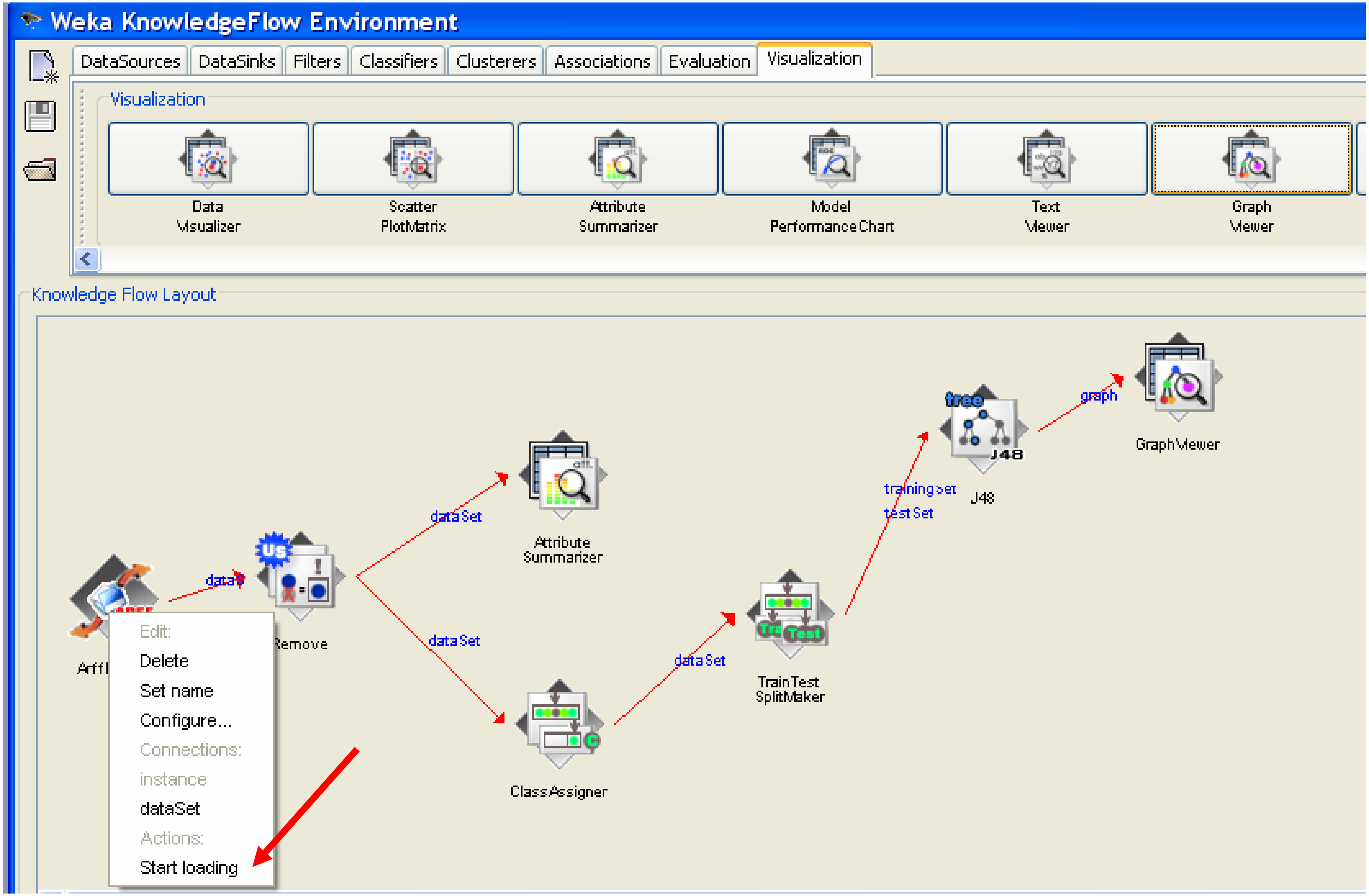
We set J48 component in the diagram,

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.

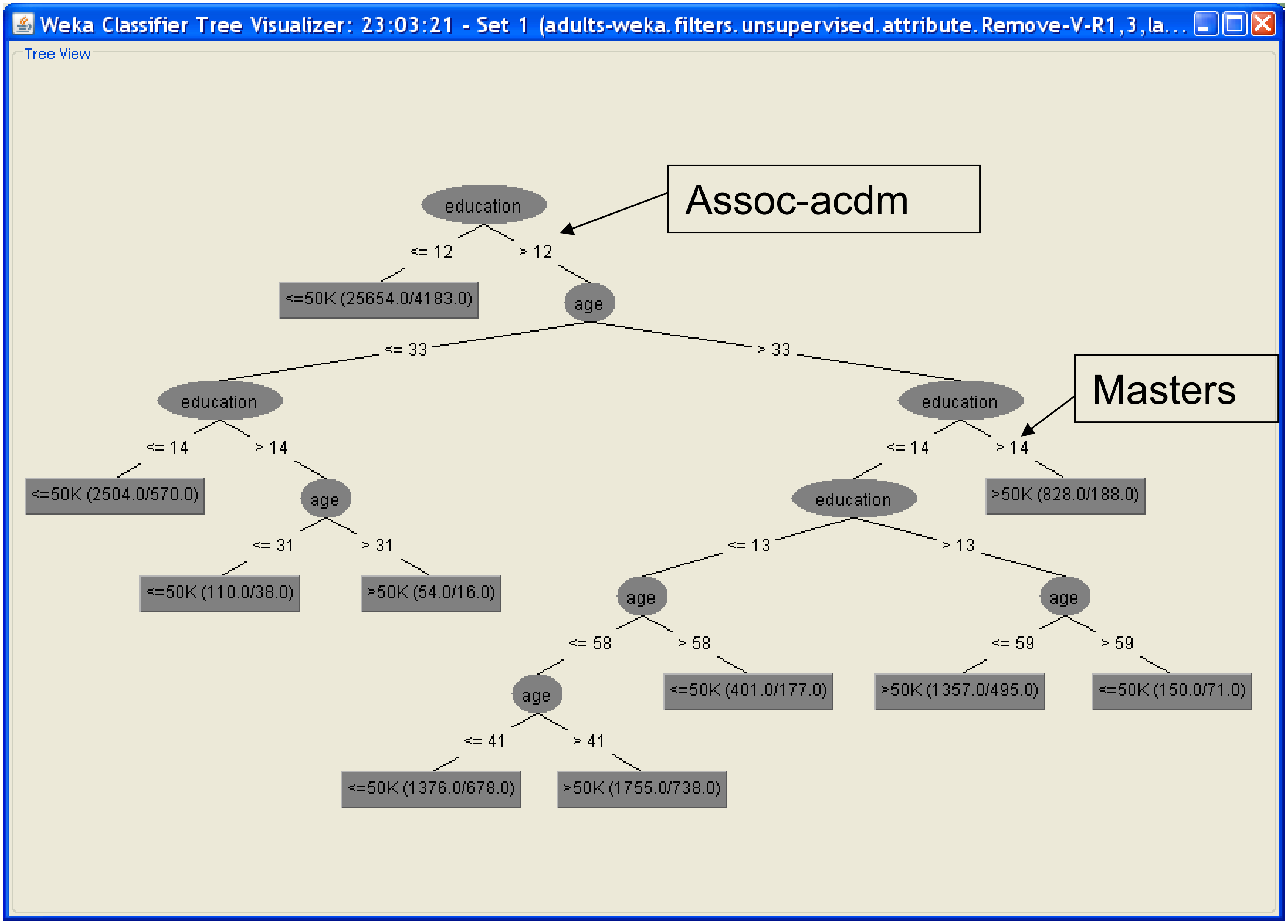
Adding visualizer to see the classification results



Perform classification



Show classification results (decision tree)



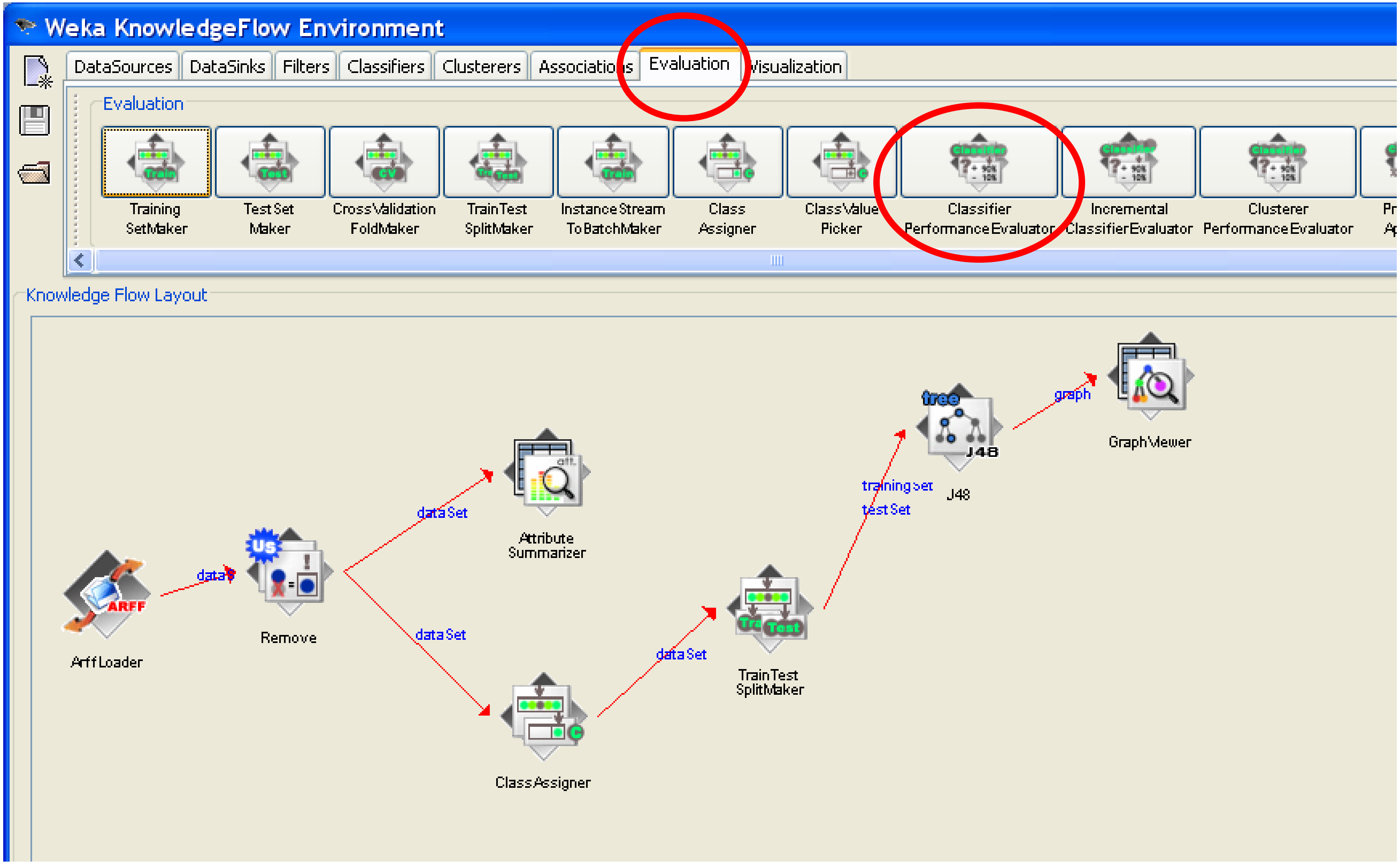
Weka Classifier Tree Visualizer: 23:03:21 - Set 1 (adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,la...

Tree View

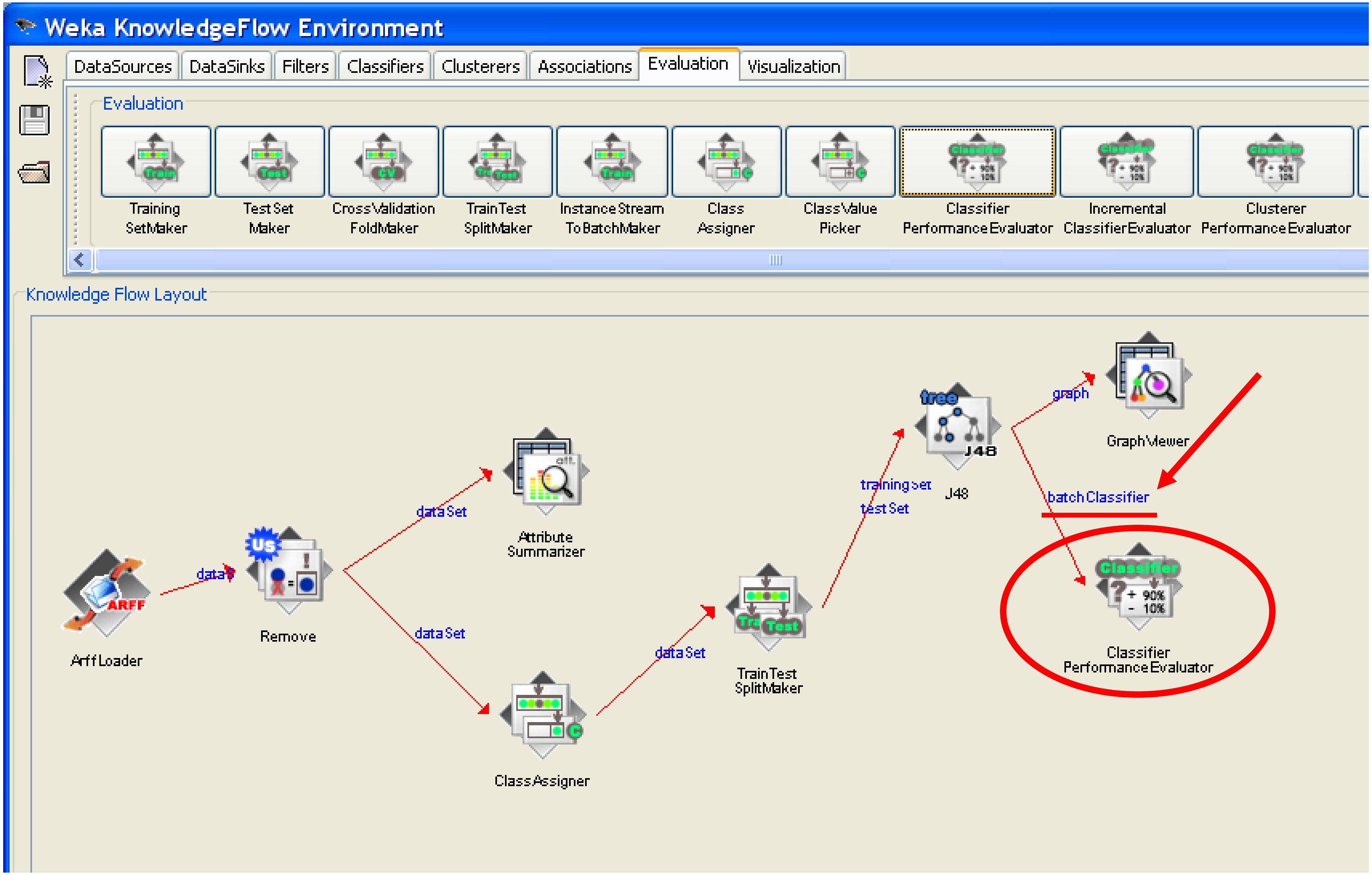
Assoc-acdm

Masters

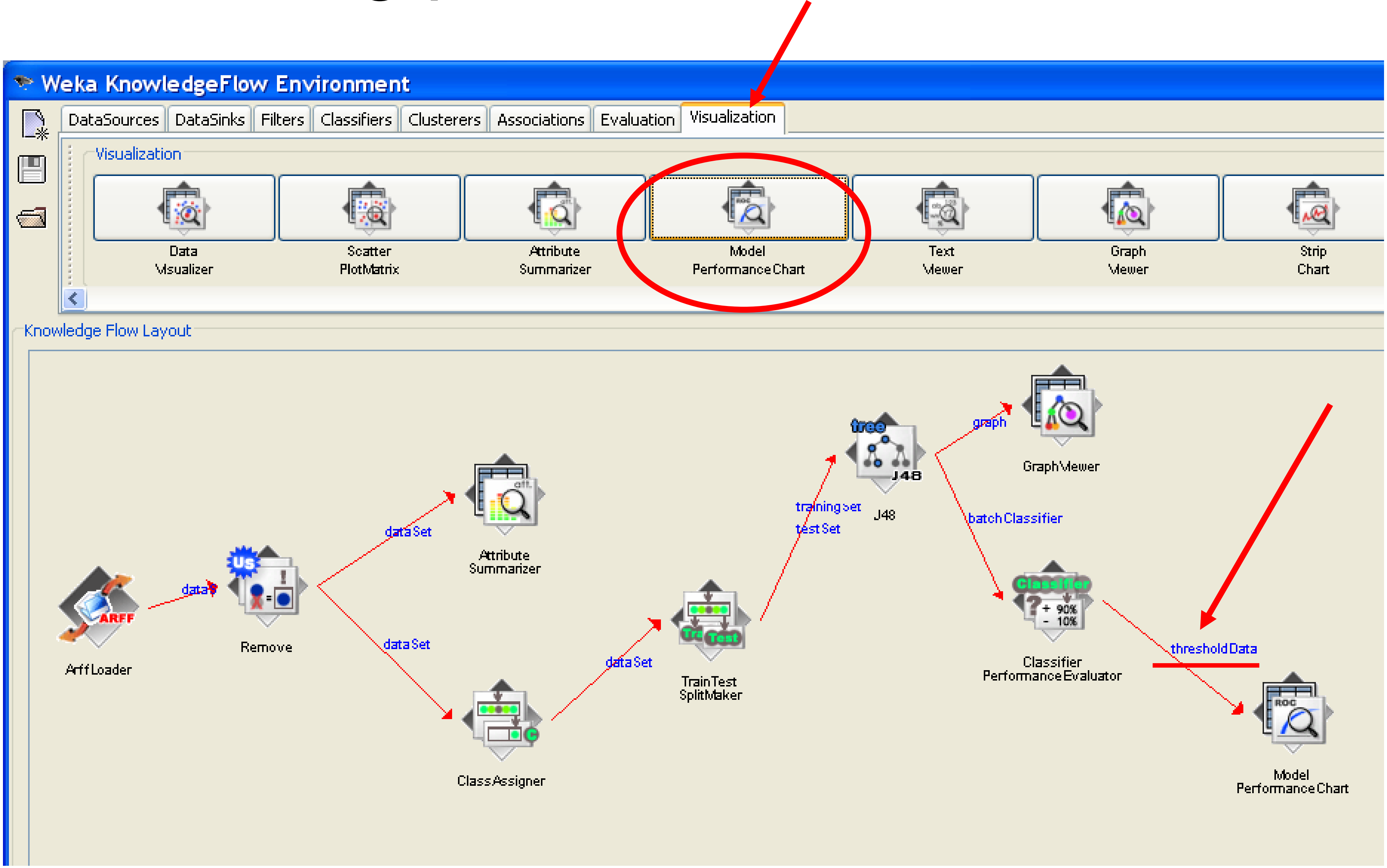
Classifier evaluation



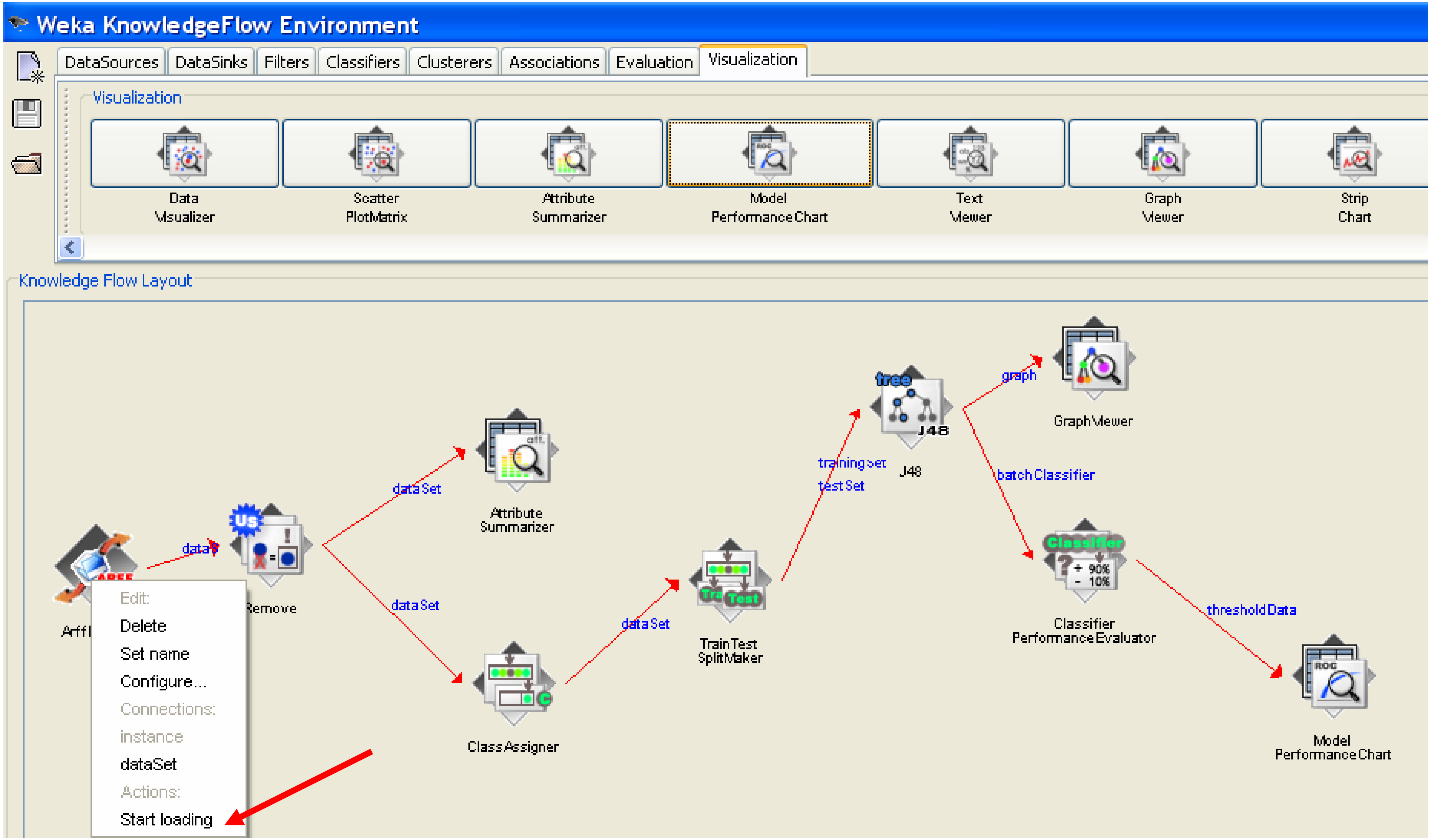
Connecting classifier to the evaluator



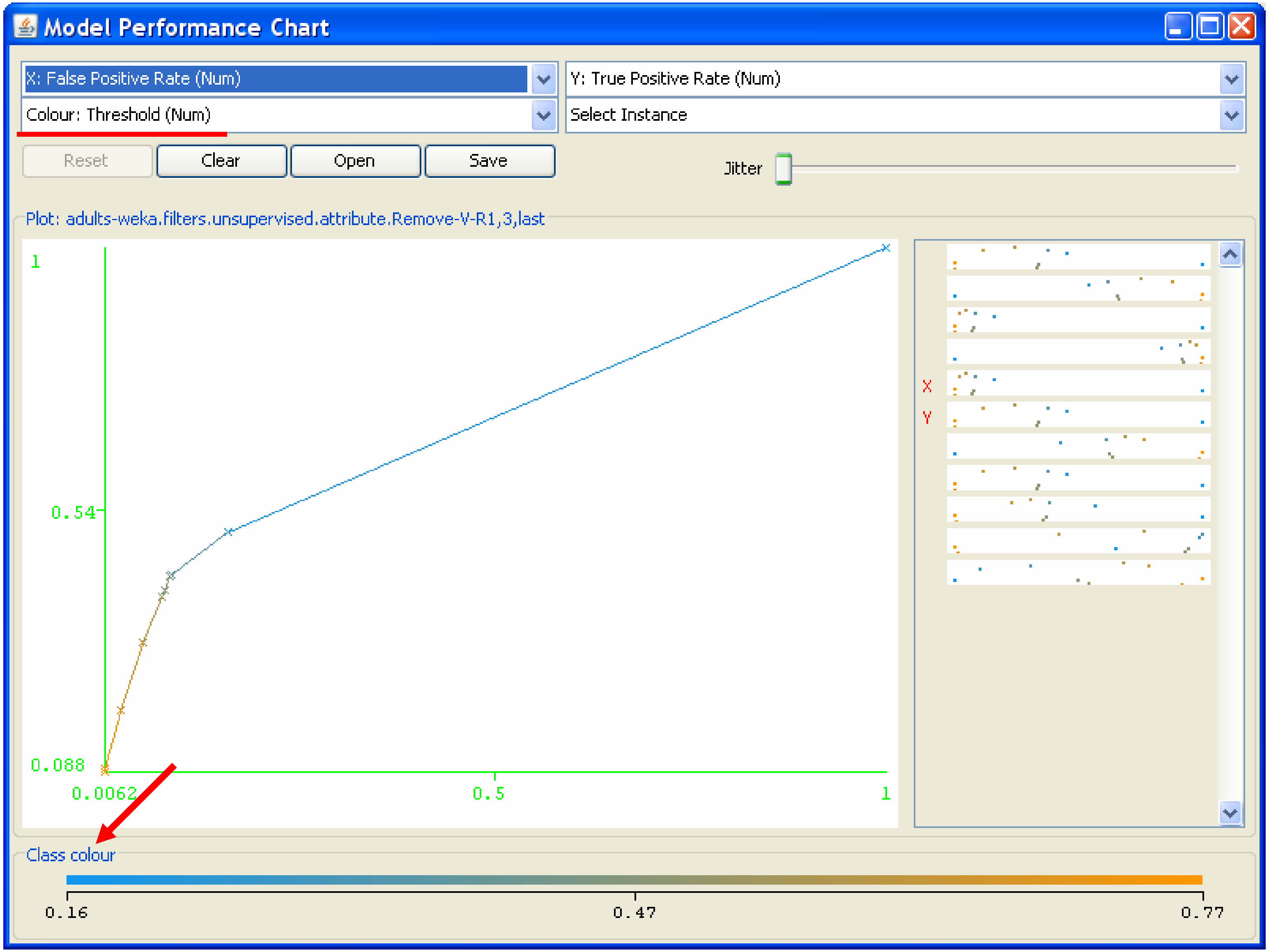
Selecting performance model: chart



Running the model



View ROC curve



Model Performance Chart

X: False Positive Rate (Num)

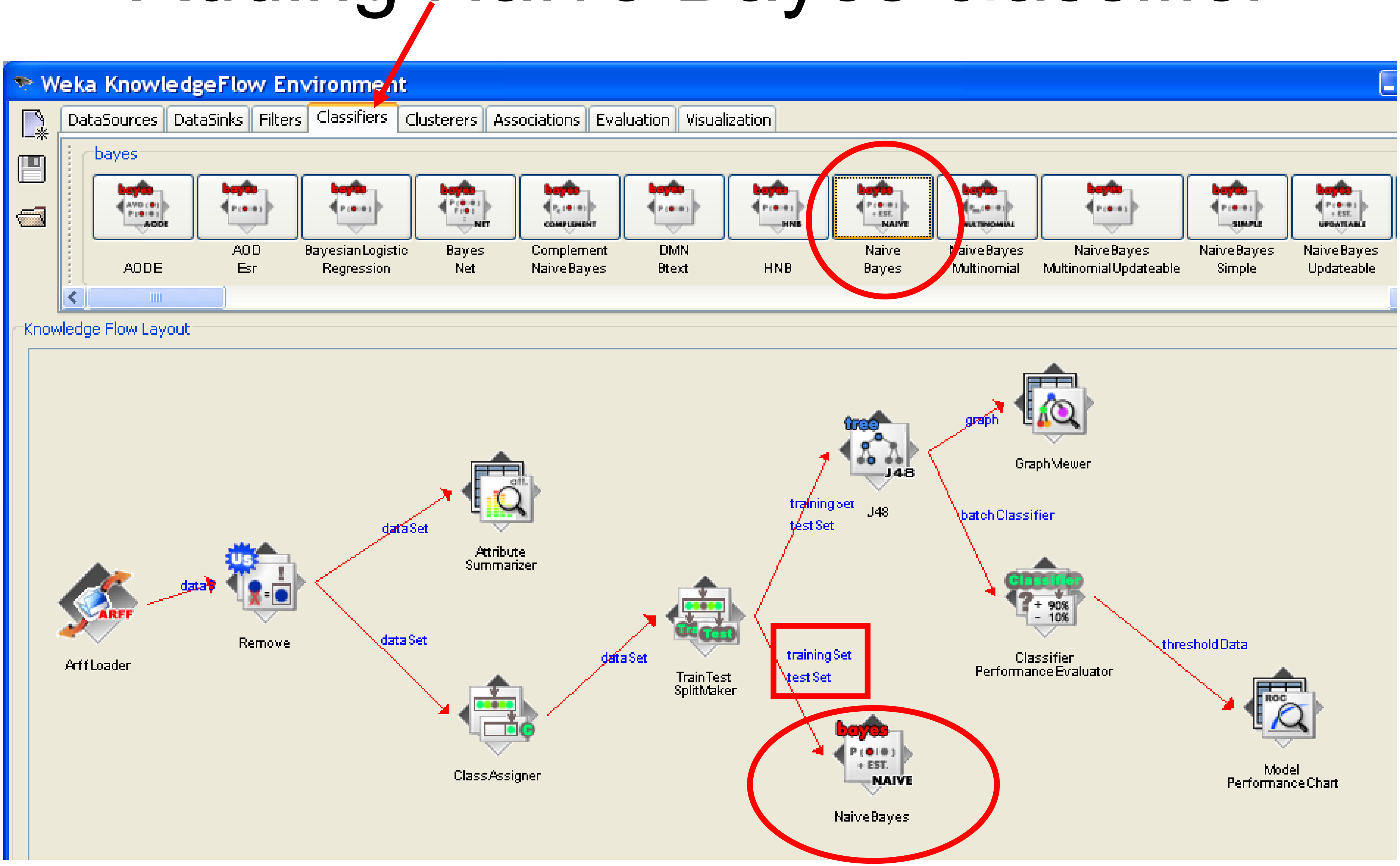
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last

Adding Naïve Bayes classifier



Weka KnowledgeFlow Environment

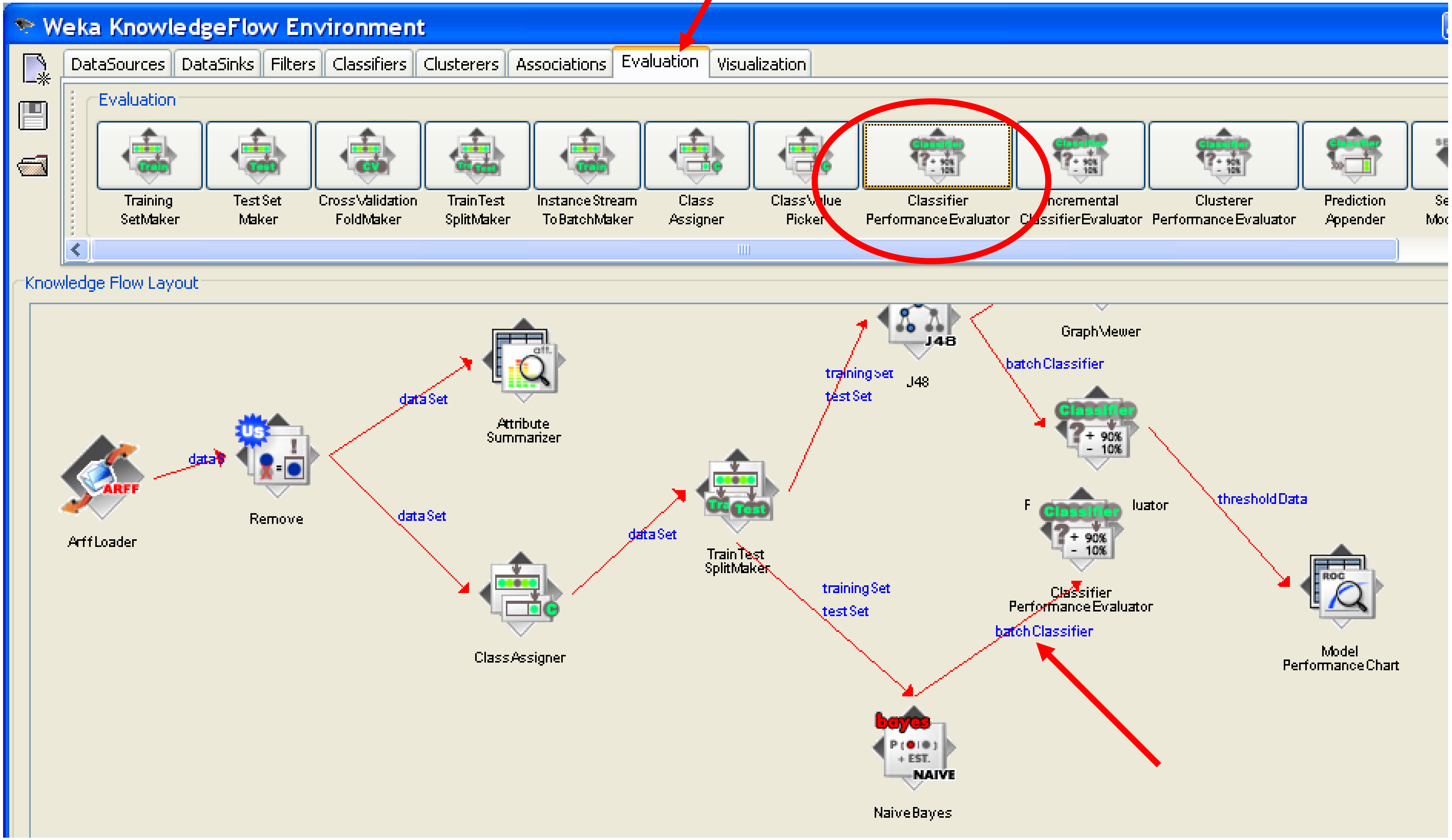
DataSources DataSinks Filters Classifiers Clusterers Associations Evaluation Visualization

bayes

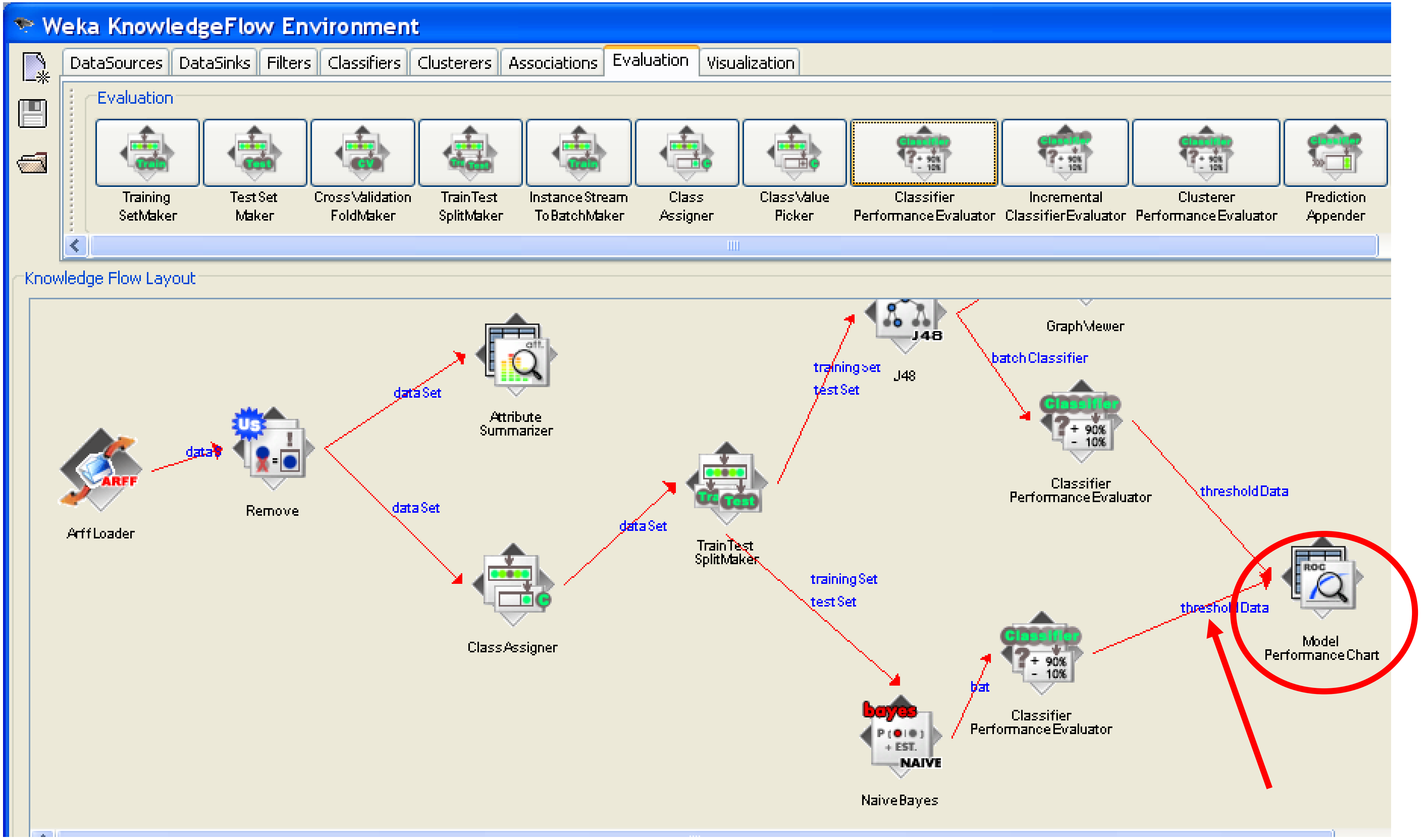
AODE | AOD Esr | Bayesian Logistic Regression | Bayes Net | Complement Naive Bayes | DMN Btext | HNB | Naive Bayes | Naive Bayes Multinomial | Naive Bayes Multinomial Updateable | Naive Bayes Simple | Naive Bayes Updateable

Knowledge Flow Layout

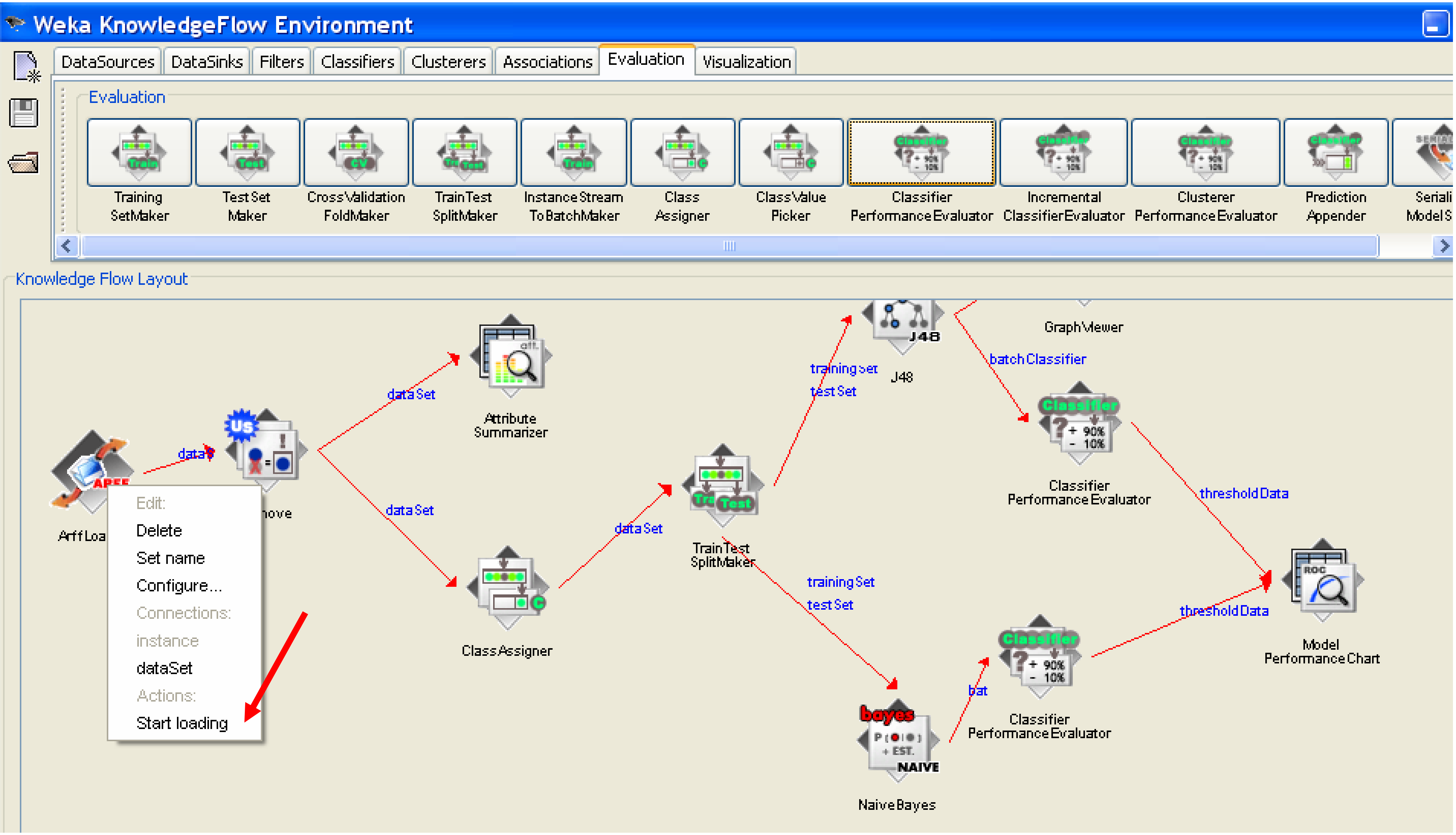
Adding separate performance evaluator for Naïve Bayes classifier



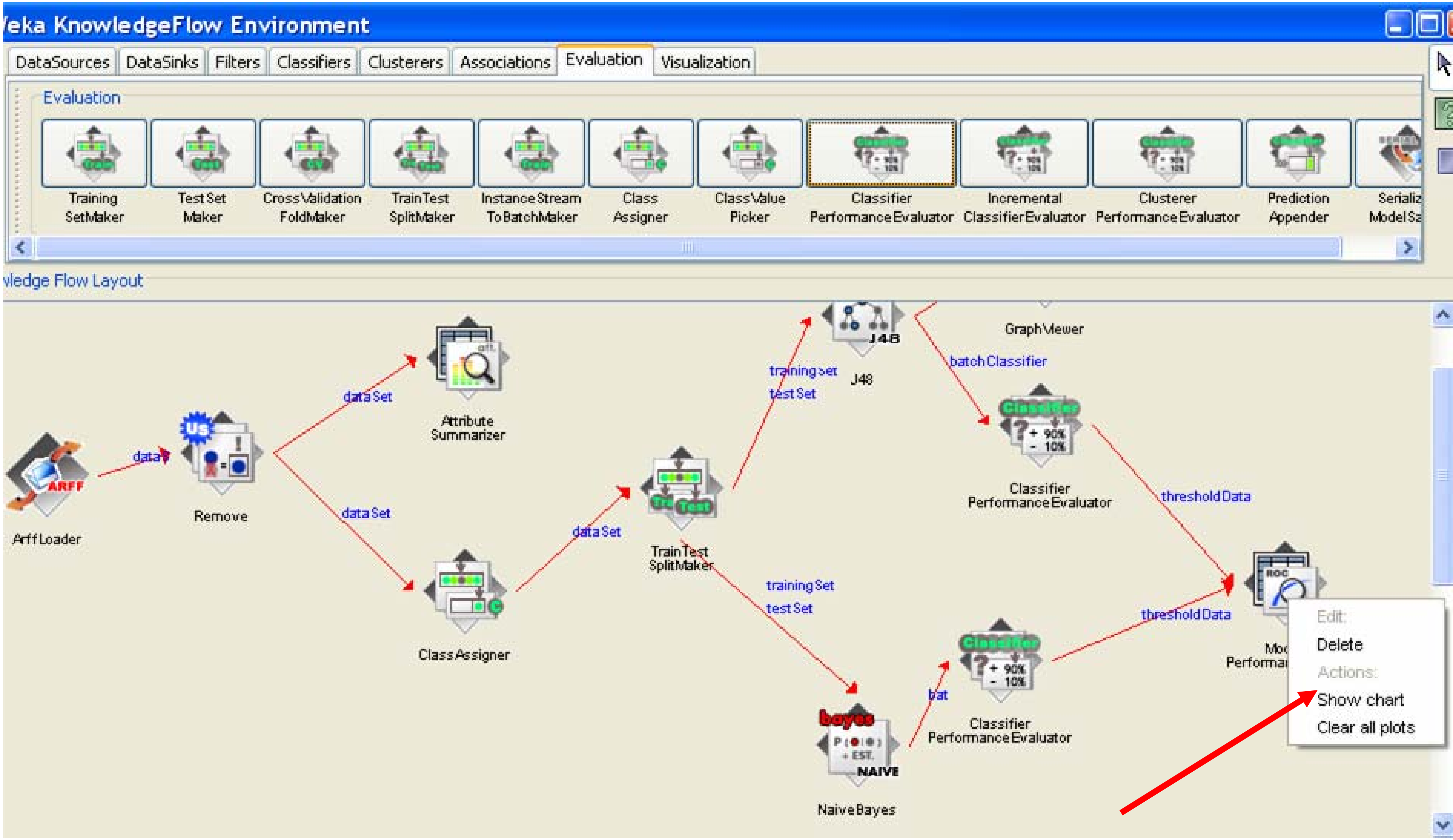
Connecting second performance evaluator to the same Model Performance Chart



Run both classifiers



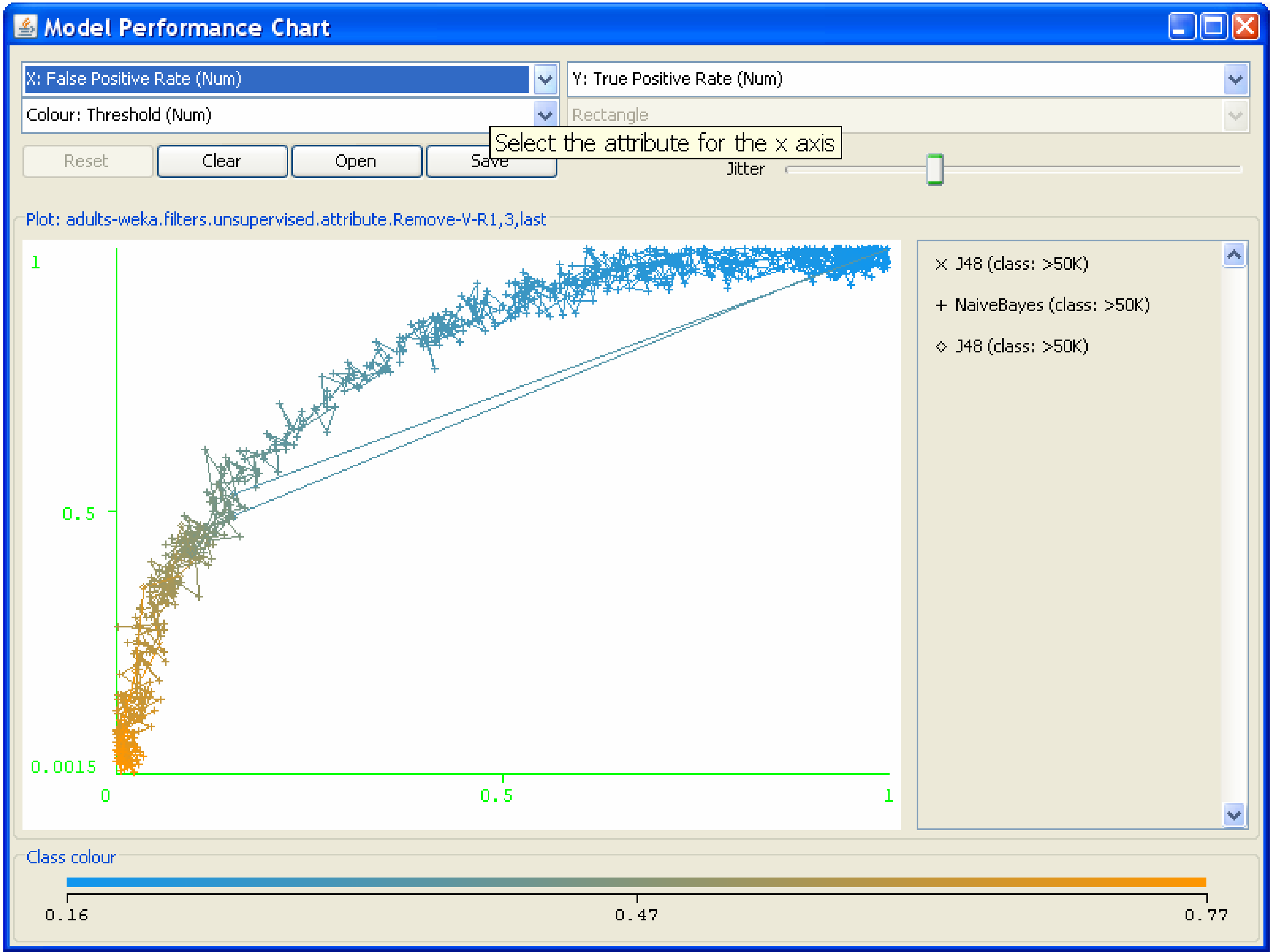
View ROC curves for both classifiers



Lab outline

- Remind what ROC curve is
- Generate ROC curves using WEKA
- Some usage of ROC curves

Compare classifiers using their ROC curves



Model Performance Chart

X: False Positive Rate (Num)

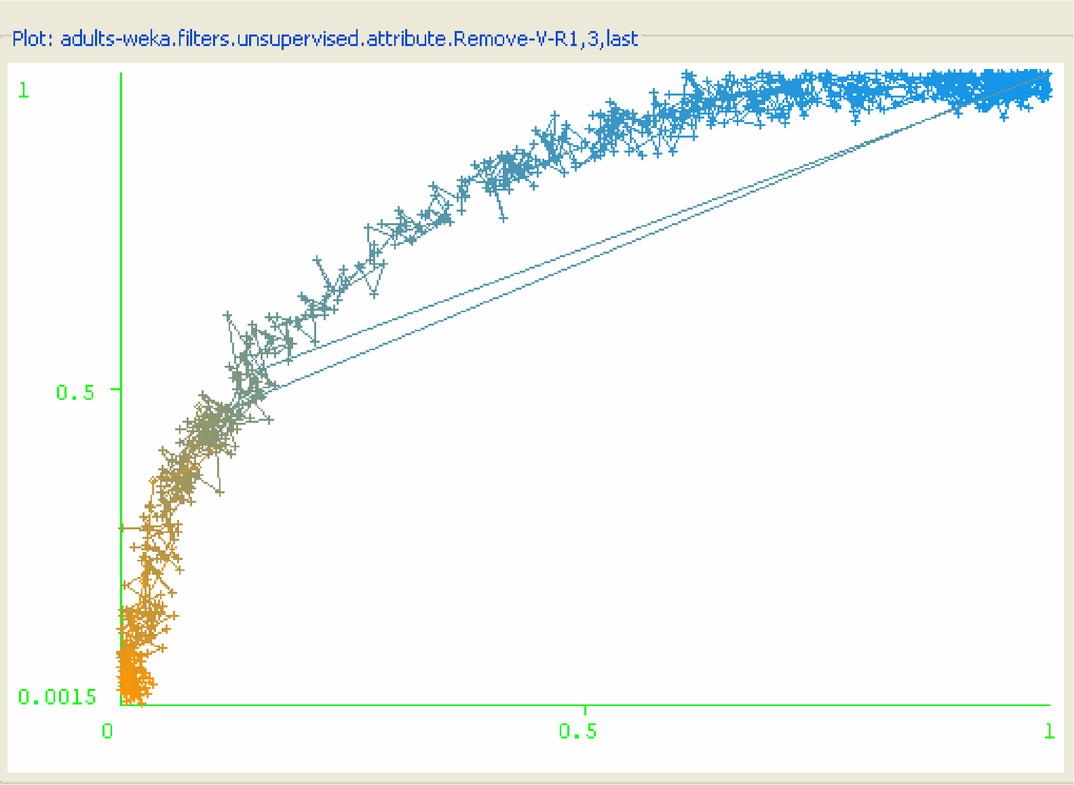
Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select the attribute for the x axis

Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last

How good is the classifier

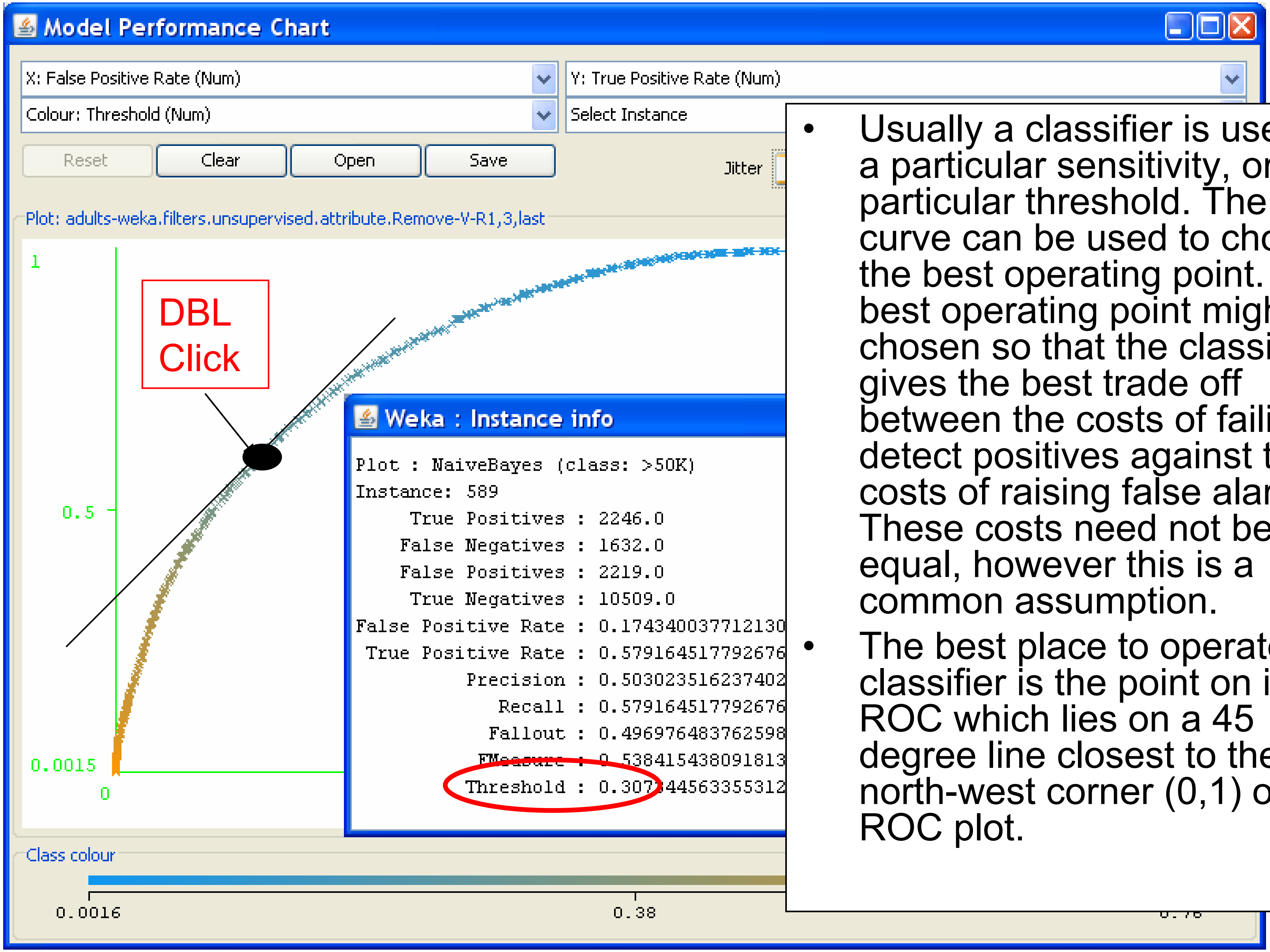


Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last

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?

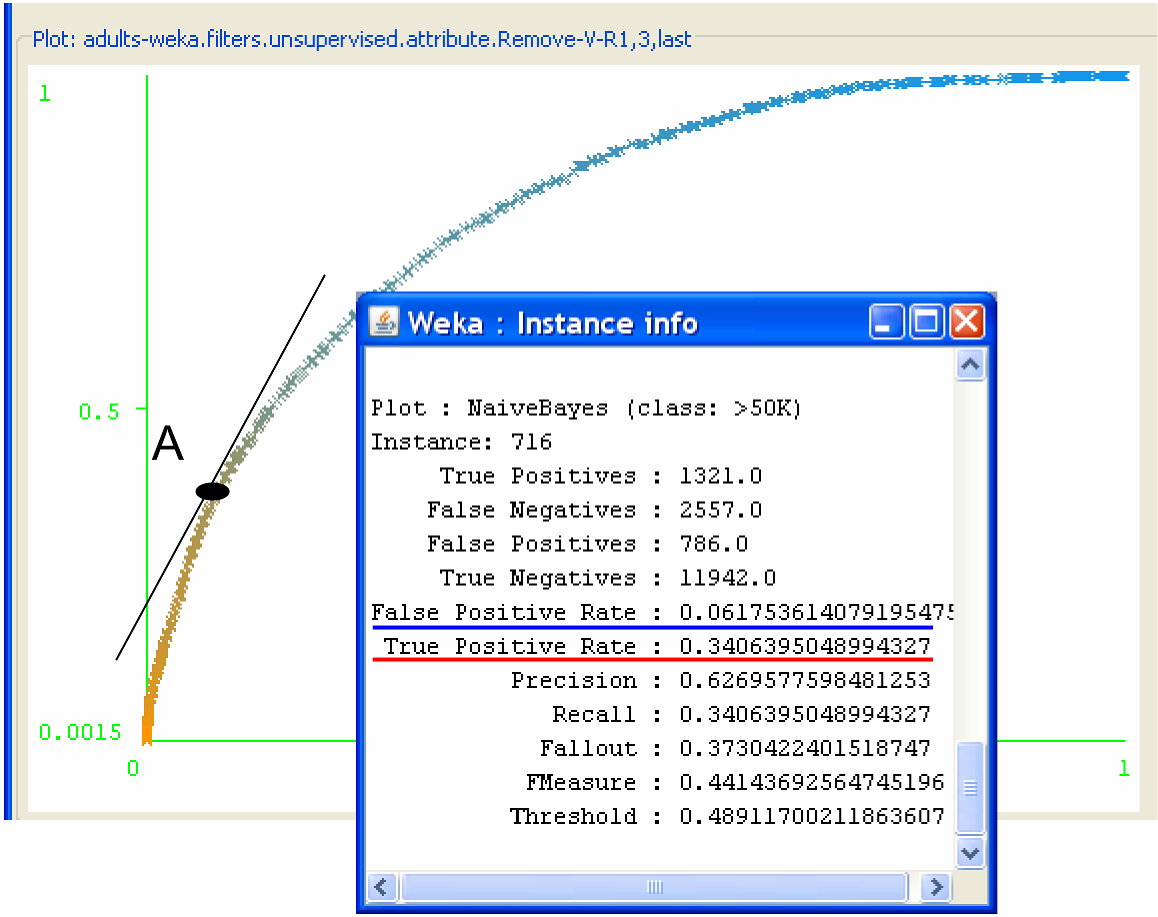
Choosing the Operating Point



```
Plot : NaiveBayes (class: >50K)
Instance:  589
     True Positives : 2246.0
    False Negatives : 1632.0
    False Positives : 2219.0
     True Negatives : 10509.0
False Positive Rate : 0.17434003771213O
 True Positive Rate : 0.579164517792676
          Precision : 0.503023516237402
             Recall : 0.579164517792676
            Fallout : 0.496976483762598
           FMeasure : 0.538415438091813
          Threshold : 0.30734456335312
```

- 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


Plot: adults-weka.filters.unsupervised.attribute.Remove-V-R1,3,last
1
0.5
0.0015
0
1
A
Weka : Instance info

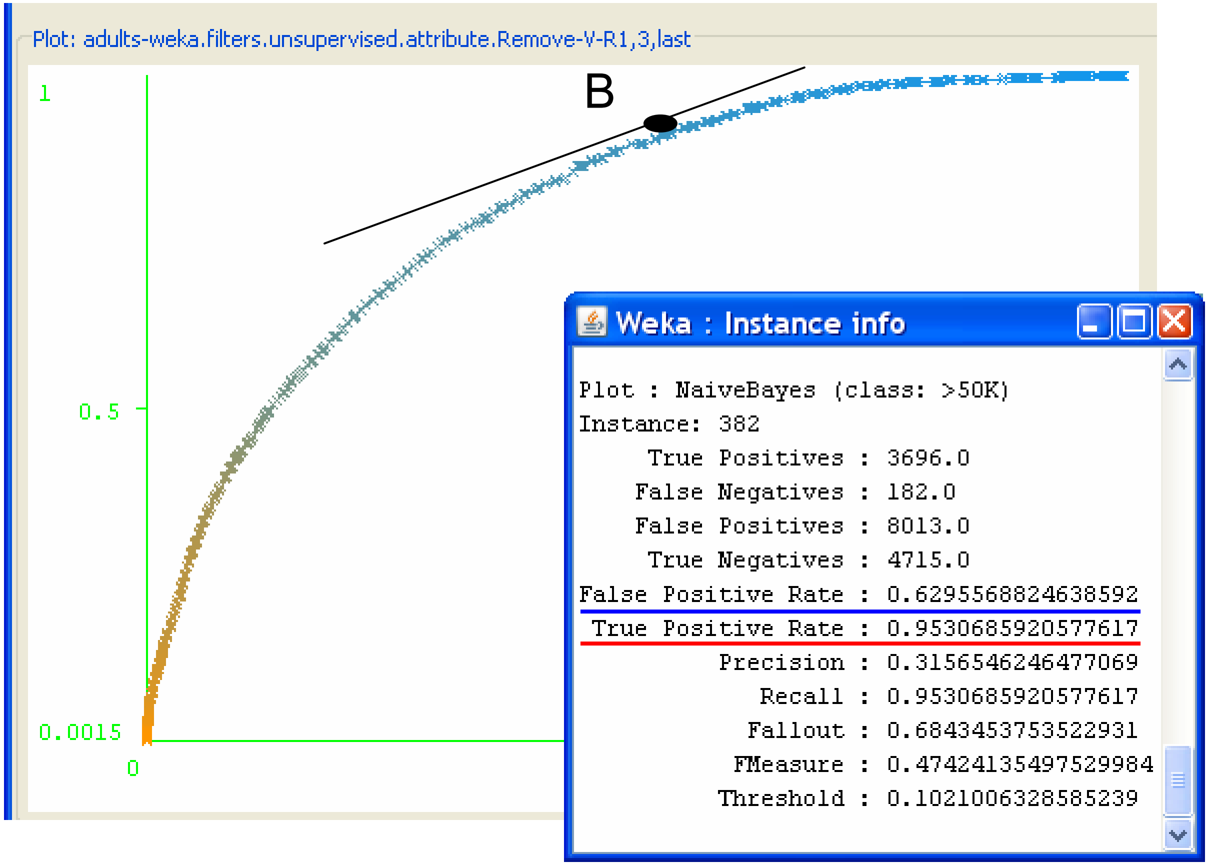
```
Plot : NaiveBayes (class: >50K)
Instance: 716
     True Positives : 1321.0
    False Negatives : 2557.0
    False Positives : 786.0
     True Negatives : 11942.0
False Positive Rate : 0.061753614079195475
 True Positive Rate : 0.3406395048994327
          Precision : 0.6269577598481253
             Recall : 0.3406395048994327
            Fallout : 0.3730422401518747
           FMeasure : 0.4414369256474519б
          Threshold : 0.48911700211863607
```

Is this threshold good :

for cancer detection?

for targeting potential customers?

Cost sensitive operating points



```
Plot : NaiveBayes (class: >50K)
Instance: 382
     True Positives : 3696.0
    False Negatives : 182.0
    False Positives : 8013.0
     True Negatives : 4715.0
False Positive Rate : 0.6295568824638592
 True Positive Rate : 0.9530685920577617
          Precision : 0.3156546246477069
             Recall : 0.9530685920577617
            Fallout : 0.6843453753522931
           FMeasure : 0.47424135497529984
          Threshold : 0.1021006328585239
```

Is this threshold good :

for cancer detection?

for targeting potential customers?

Conclusions

- WEKA is a powerful datamining tool, but is not very easy to use

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