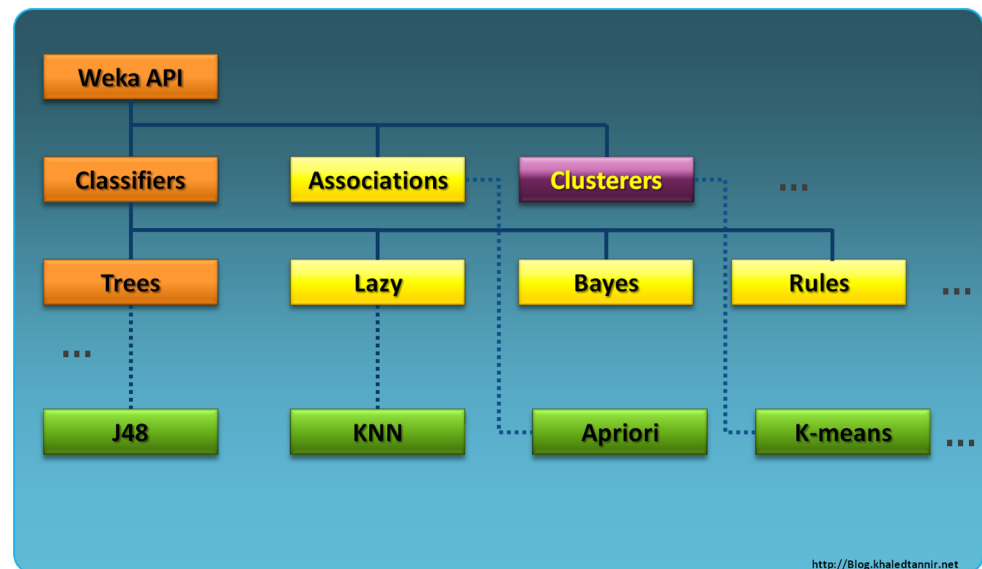


# Data Mining with Weka

Jia Wang

# Weka

- What's Weka?
  - a **collection** of algorithms
  - based on **Java**
- What can it do?
  - Data pre-processing
  - Feature selection
  - Classification
  - Clustering
  - Association rule/ frequent itemset mining
  - Visualization of data and models



# Workflow

1. Data preparation & loading
2. Data visualization & pre-processing
3. Configure/run data mining algorithm
4. Save/visualize trained models

# Data preparation & loading

- Data format: .arff file (.csv also supported)

```
% 1. Title: Iris Plants Database %
```

```
@RELATION iris
```

```
@ATTRIBUTE sepallength NUMERIC
```

```
@ATTRIBUTE sepalwidth NUMERIC
```

```
@ATTRIBUTE petallength NUMERIC
```

```
@ATTRIBUTE petalwidth NUMERIC
```

```
@ATTRIBUTE class {Iris-setosa,Iris-versicolor, Iris-virginica}
```

```
@DATA
```

```
5.1,3.5,1.4,0.2,Iris-setosa
```

```
4.9,3.0,1.4,0.2,Iris-setosa
```

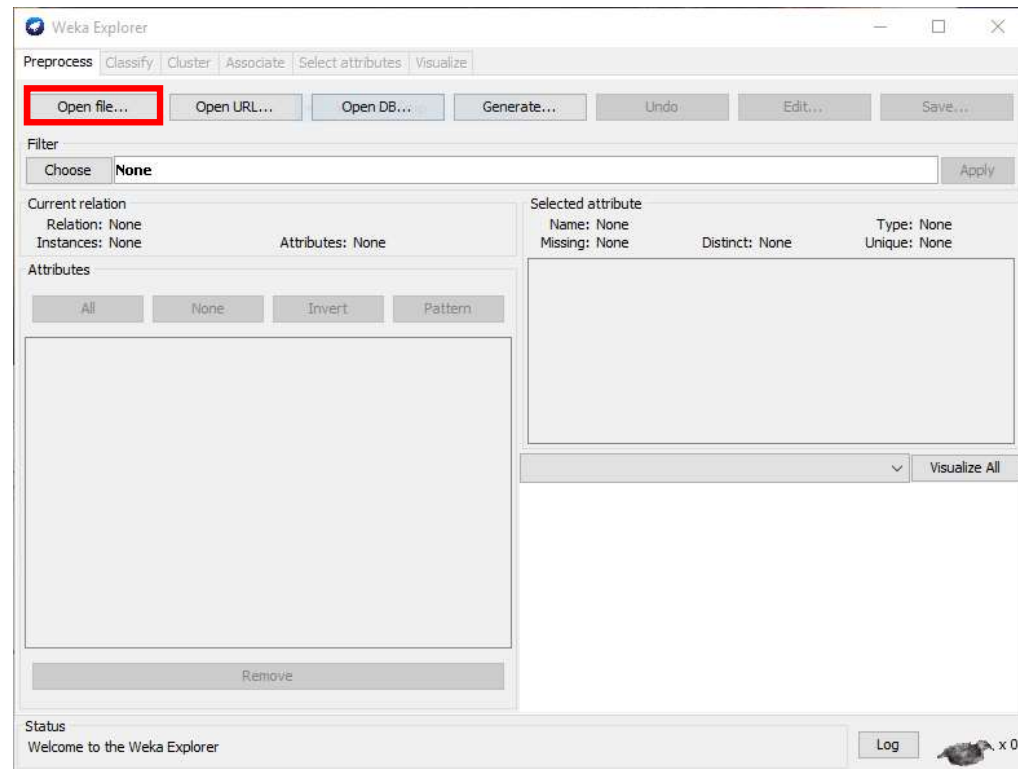
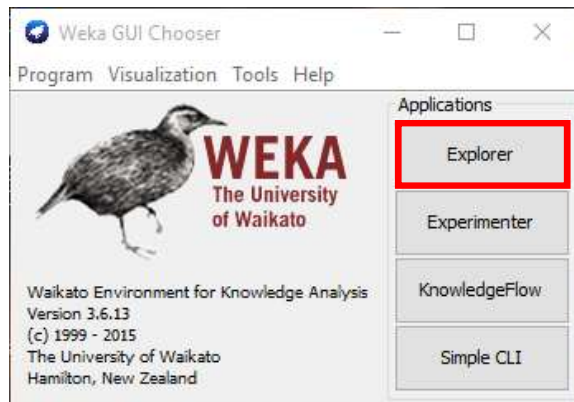
```
4.7,3.2,1.3,0.2,Iris-setosa
```

```
...
```

# Data preparation & loading

- Steps

1. Run weka GUI
2. Click 'Explorer'
3. 'Open file...'

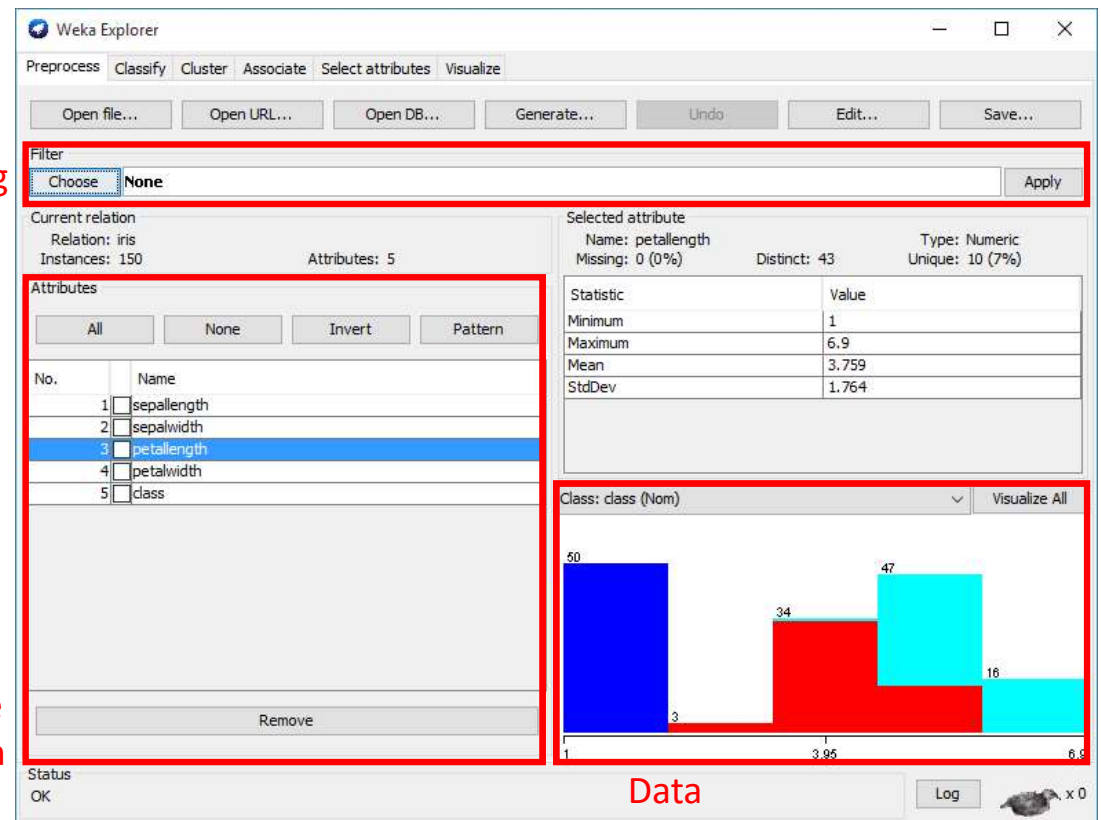


# Data visualization & pre-processing

- Attribute selection
  - i.e., attributes used in training
- Data filtering & formatting
- Visualize data

filtering

attribute selection



Data visualization

# Specify tasks and evaluation methods

1. Click 'classify' tab
2. Choose algorithm and specify its parameters
3. Specify train/test data
4. Select target attribute
5. Click 'start'

**Classifiers you have trained so far**

**evaluation results**

**Classifier output**

Metric	Value
Kappa statistic	0.9408
Mean absolute error	0.0396
Root mean squared error	0.1579
Relative absolute error	8.8979 %
Root relative squared error	33.4091 %
Total Number of Instances	51

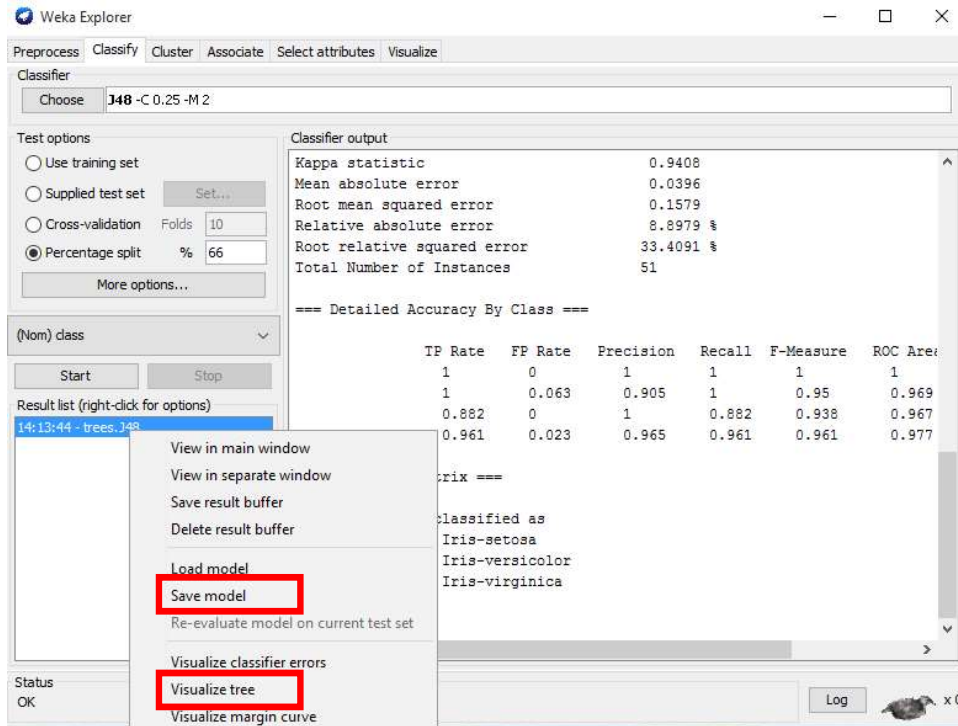
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	1	0	1	1	1	1
1	0.882	0.063	0.905	1	0.95	0.969
Weighted Avg.	0.961	0.023	0.965	0.961	0.961	0.977

=== Confusion Matrix ===

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	15	c = Iris-virginica

# Save/visualize trained models



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 Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

Classifier output

Kappa statistic 0.9408

Mean absolute error 0.0396

Root mean squared error 0.1579

Relative absolute error 8.8979 %

Root relative squared error 33.4091 %

Total Number of Instances 51

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	1	0	1	1	1	1
2	1	0.063	0.905	1	0.95	0.969
3	0.882	0	1	0.882	0.938	0.967
4	0.961	0.023	0.965	0.961	0.961	0.977

====

Classified as

Iris-setosa

Iris-versicolor

Iris-virginica

Result list (right-click for options)

14:13:44 - trees.J48

View in main window

View in separate window

Save result buffer

Delete result buffer

Load model

**Save model**

Re-evaluate model on current test set

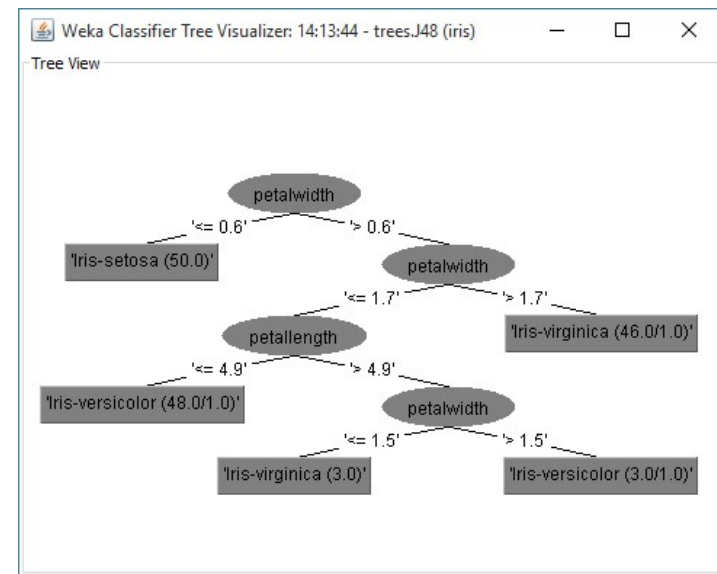
Visualize classifier errors

**Visualize tree**

Visualize margin curve

Status OK

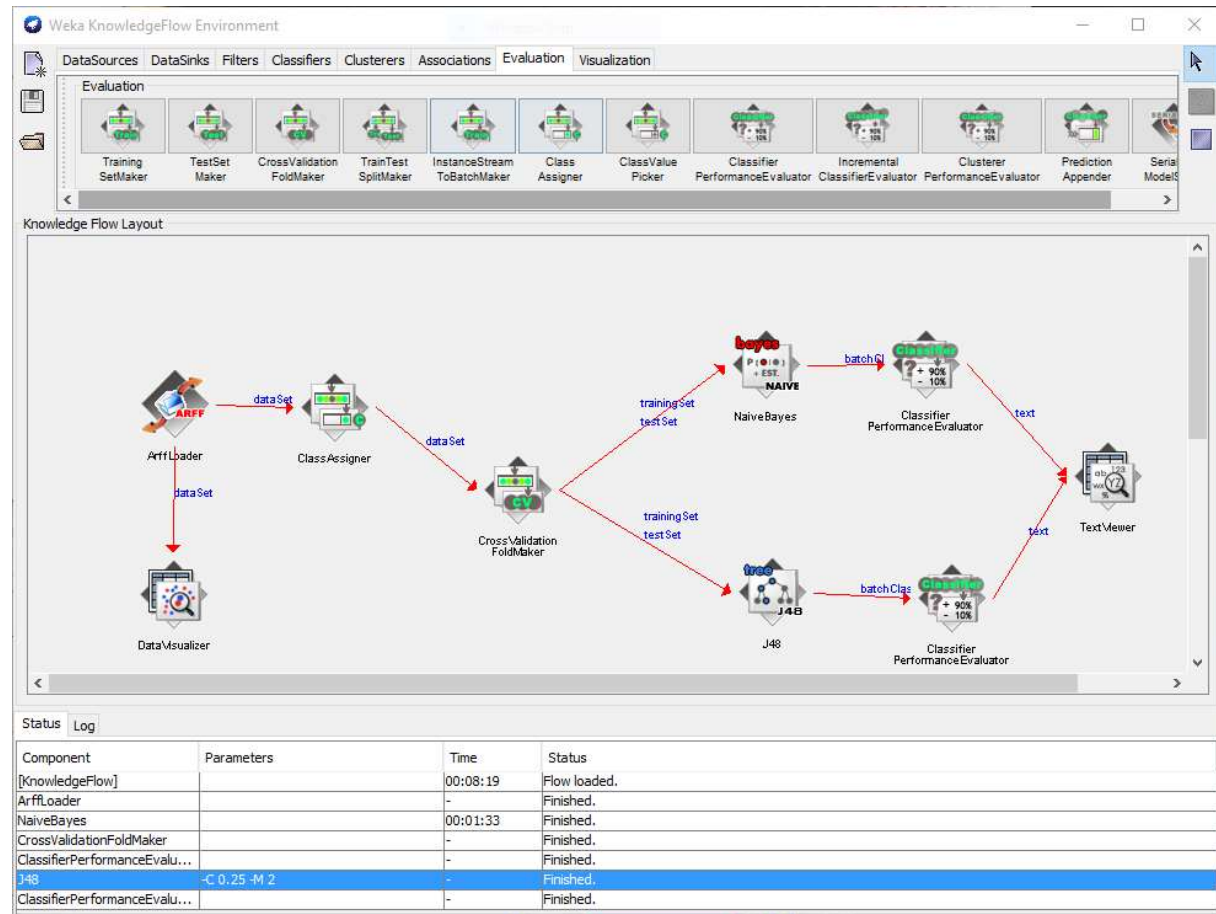
Log





# Knowledge Flow: a GUI for Workflow

- Specify workflow by a graphical interface



# Emsemble classifiers

- Adaboost

