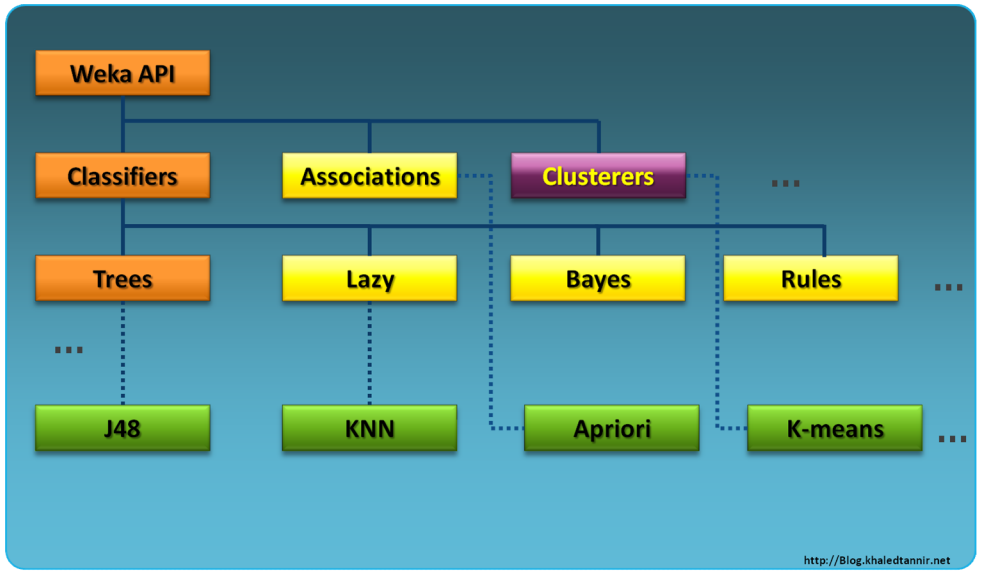
Data Mining with Weka

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

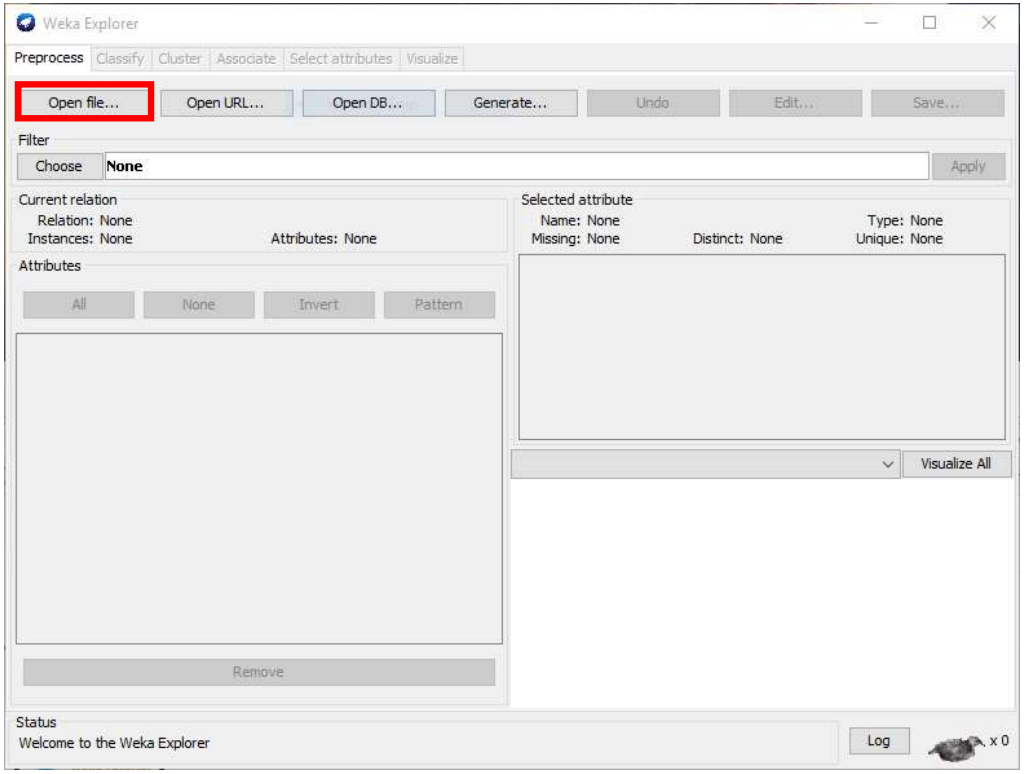
Data preparation & loading

- 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
…
```

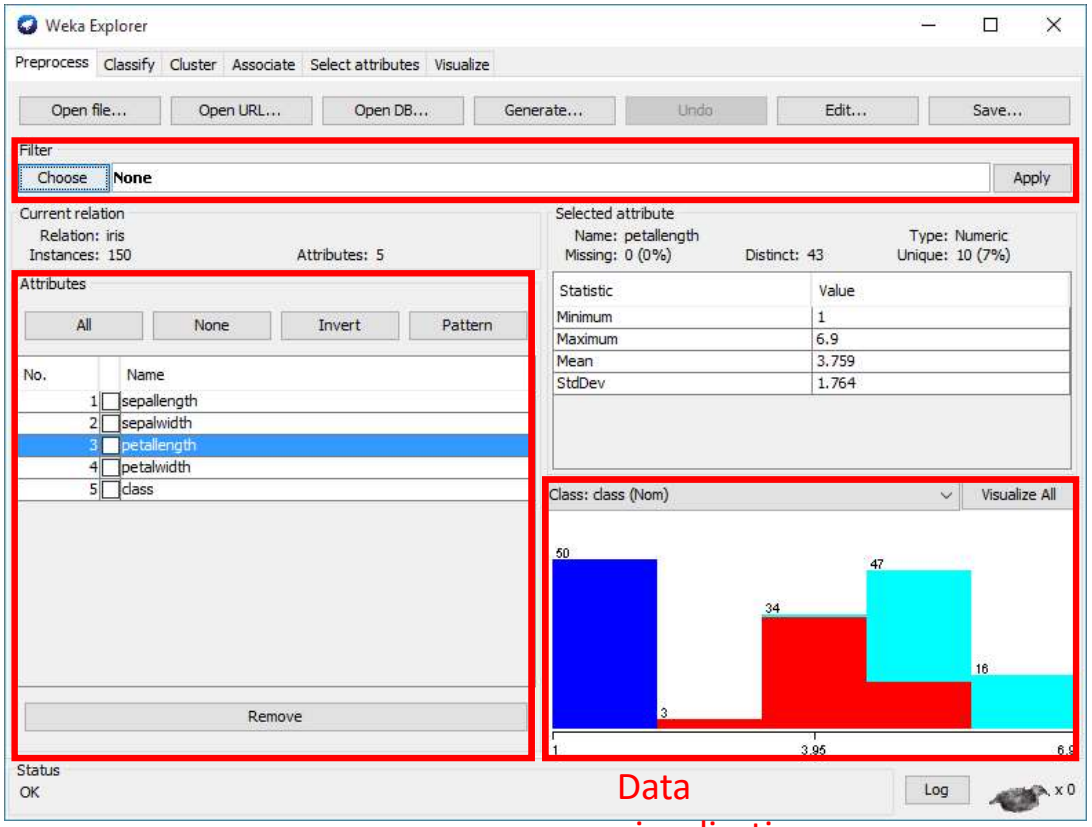
Data preparation & loading

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



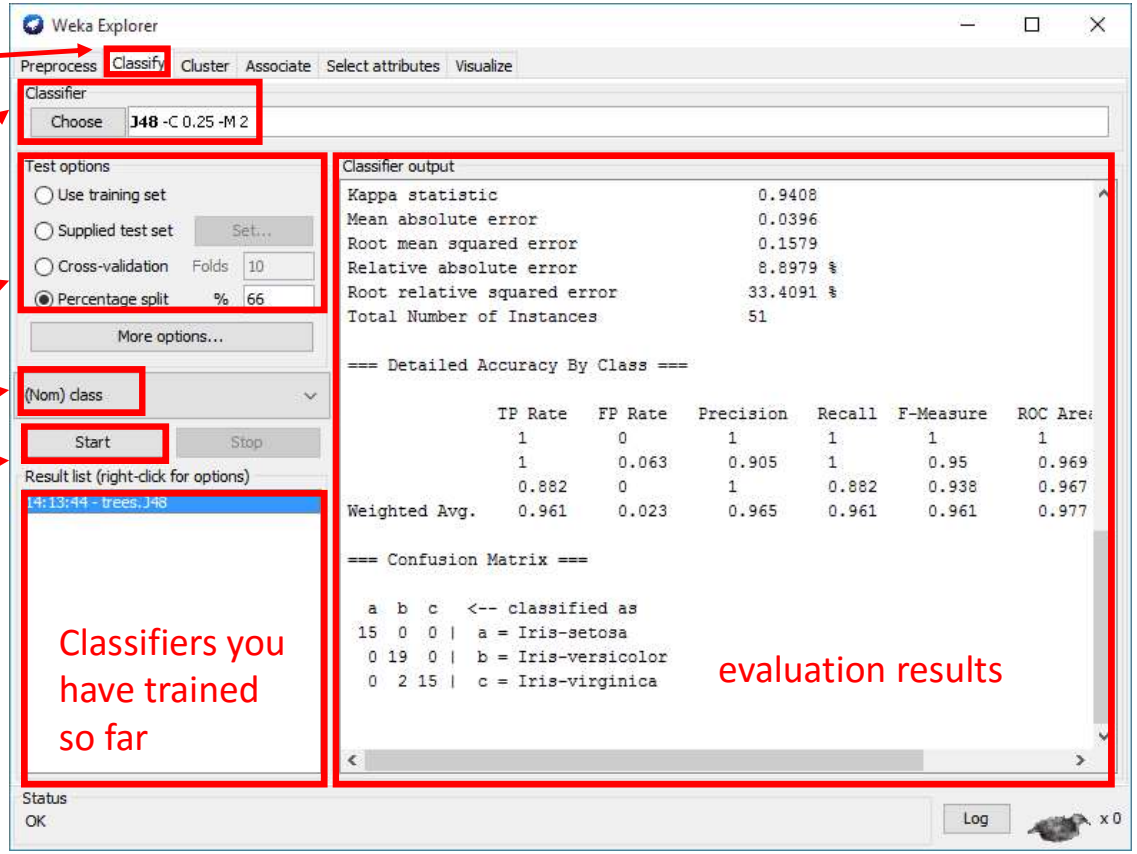
Data visualization & pre-processing

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

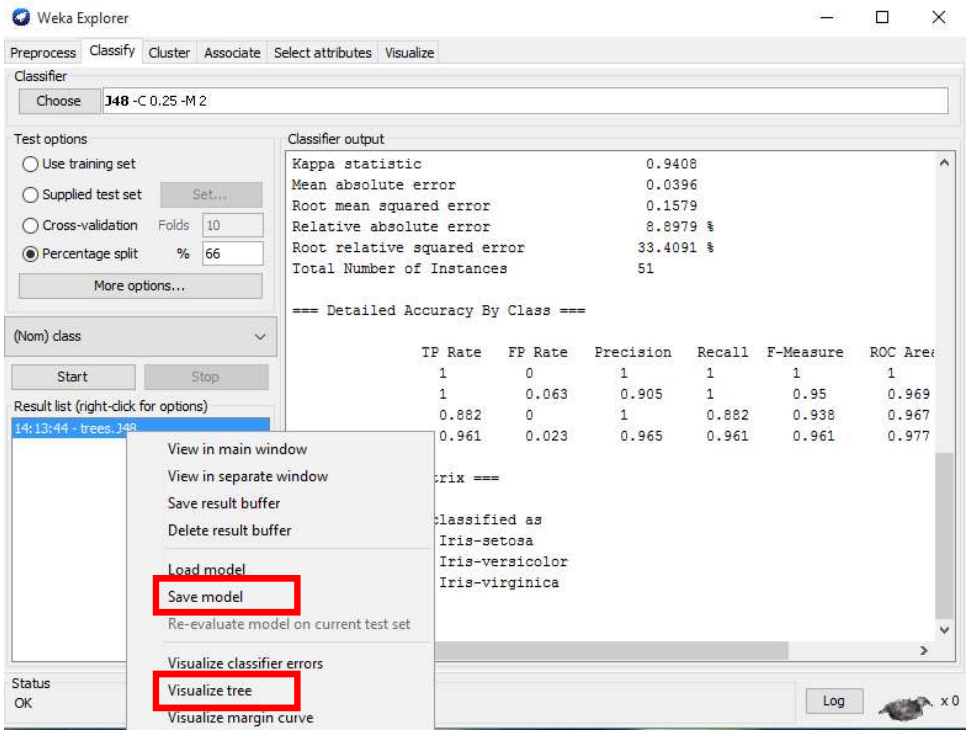


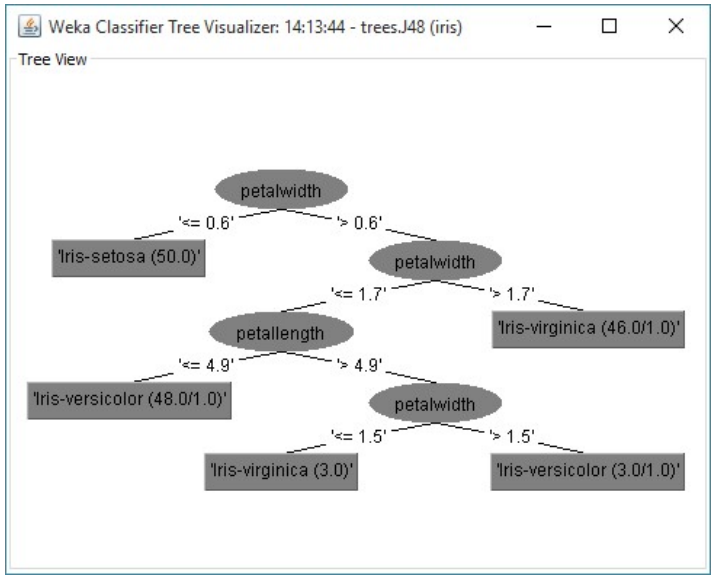
Specify tasks and evaluation methods

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



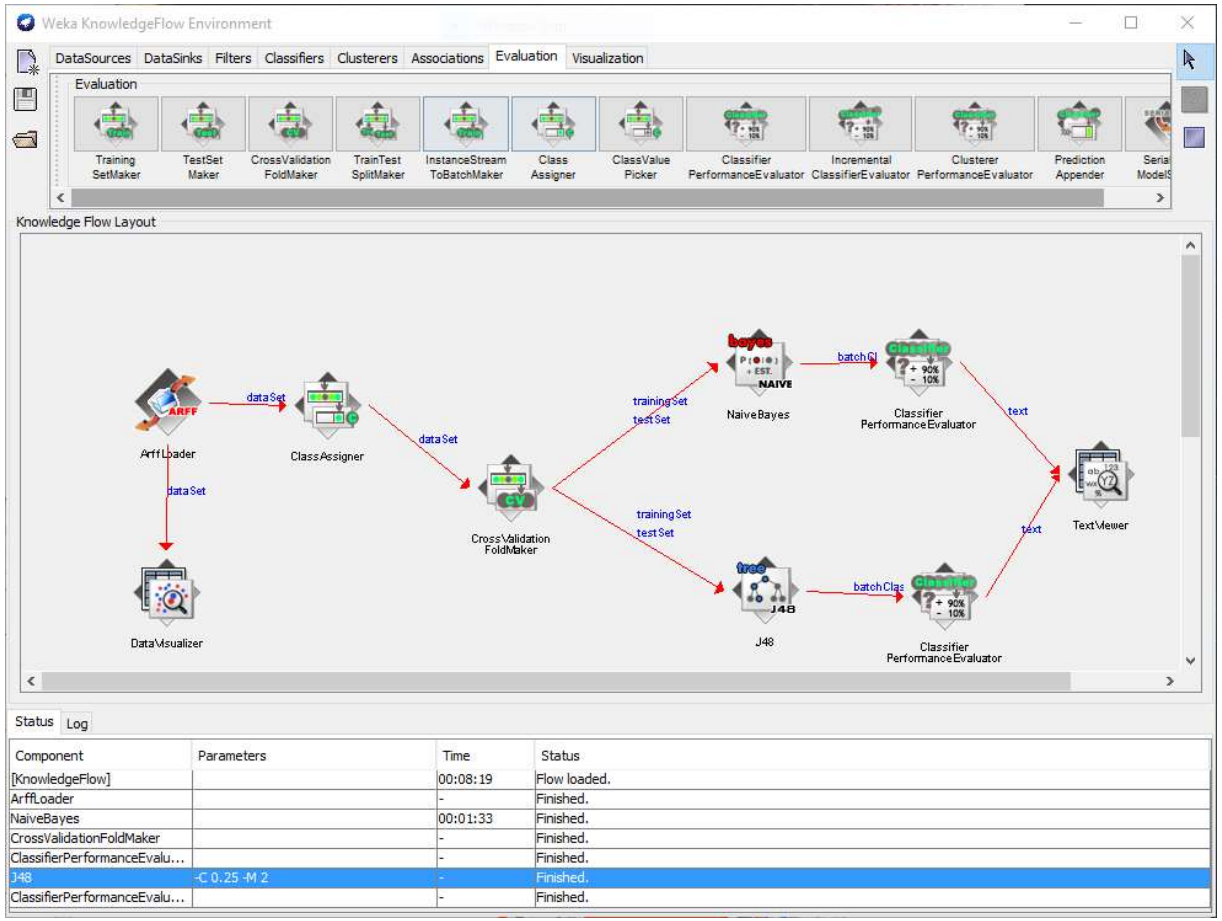
Save/visualize trained models





Knowledge Flow: a GUI for Workflow

- Specify workflow by a graphical interface



Emsemble classifiers

- Adaboost