Big Data Technical Platform I

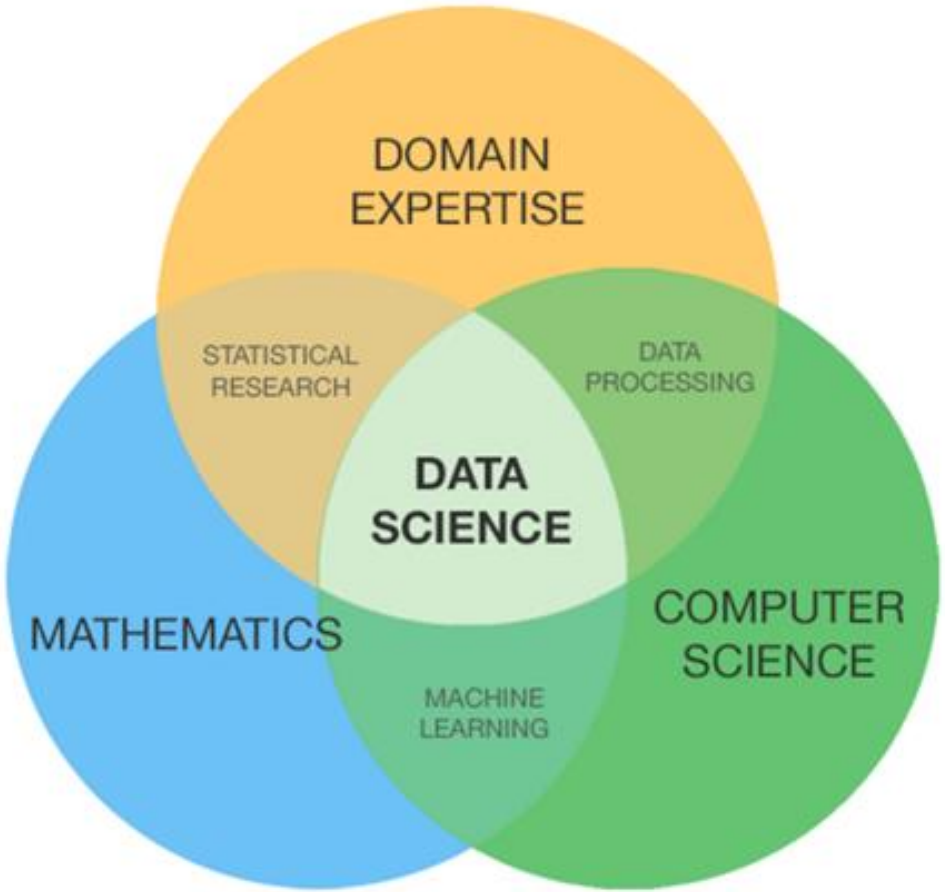
Outline

- Overview
- Weka
- R
- Python - scikit-learn

Overview

Data Science

- Big data
  - 3V: Velocity, Volume, Variety
- Data Science
  - Statistical Research
  - Machine Learning
  - Data Mining
  - Cross/multi-Domain Knowledge
  - …



Platform/Service for Data analysis

amazon web services

WEKA The University of Waikato

SKYTREE BIG DATA ANALYTICS

MATLAB

Microsoft Azure

rapidminer

APACHE Spark MLlib

Google Cloud Platform

sas

Open for Innovation KNIME

R

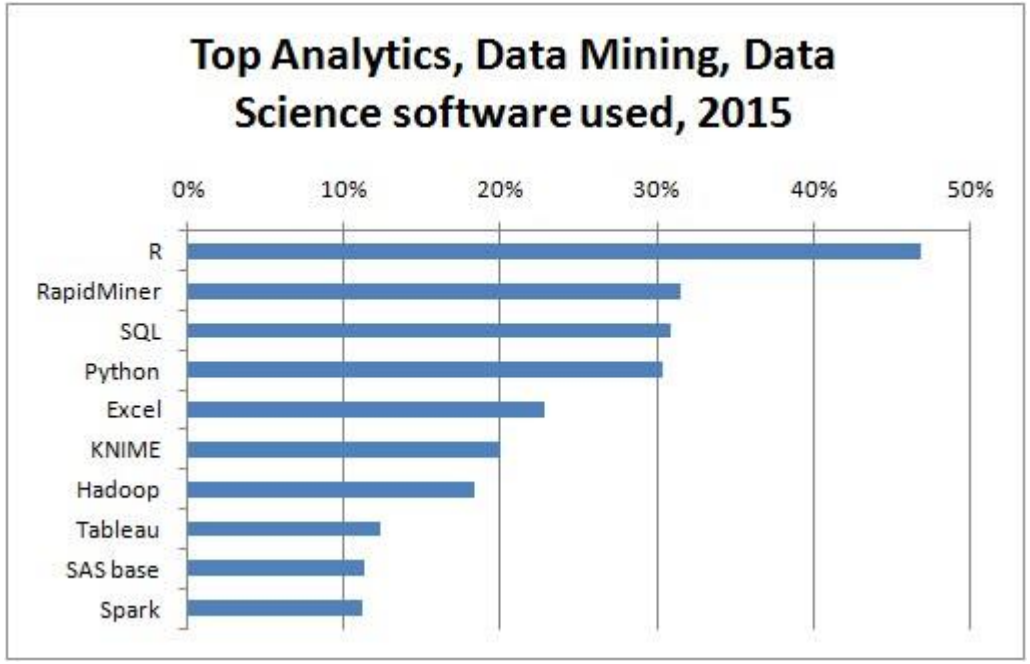
SPSS

scikit learn

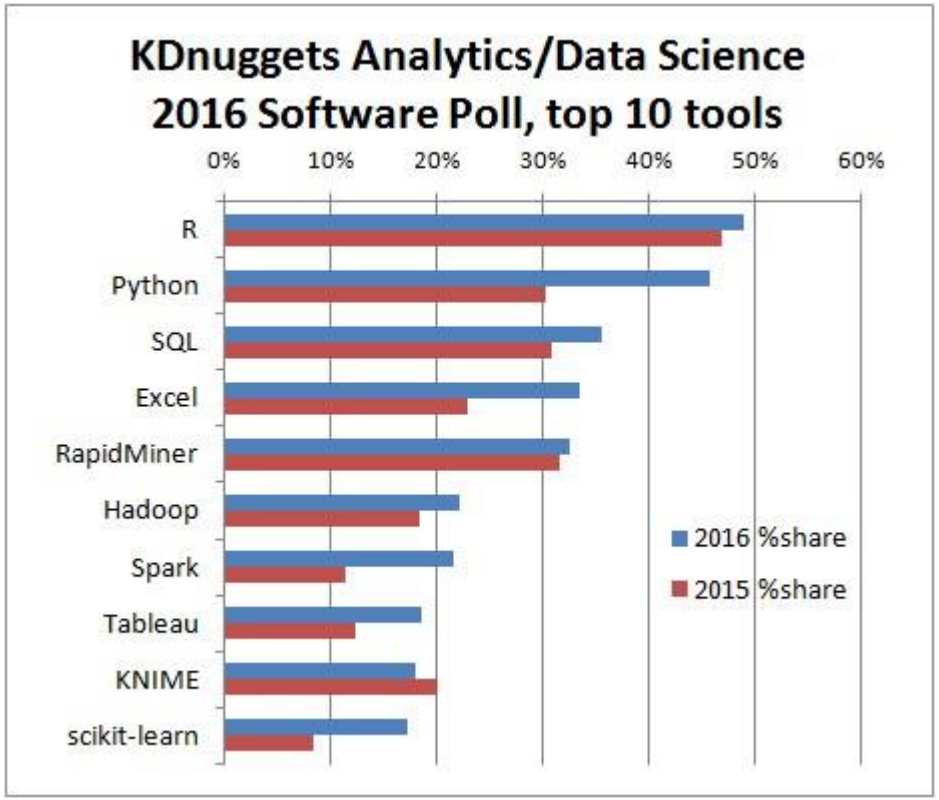
Top 10 Analytics Tools

What programming/statistics languages you used for an analytics / data mining / data science work in 2014?	
Language used	% voters in 2014 (719 total) % voters in 2013 (713 total) % voters in 2012 (579 total)
R (352 voters in 2014)	49.0% 60.9% 52.5%
SAS (262)	36.4% 20.8% 19.7%
Python (252)	35.0% 38.8% 36.1%
SQL (220)	30.6% 36.6% 32.1%
Java (89)	12.4% 16.5% 21.2%
Unix shell/awk/sed (63)	8.8% 11.1% 14.7%
Pig Latin/ Hive/ other Hadoop-based languages (61)	8.5% 8.0% 6.7%
SPSS (58)	8.1% not asked not asked
MATLAB (45)	6.3% 12.5% 13.1%

Top 10 Analytics Tools



Top 10 Analytics Tools



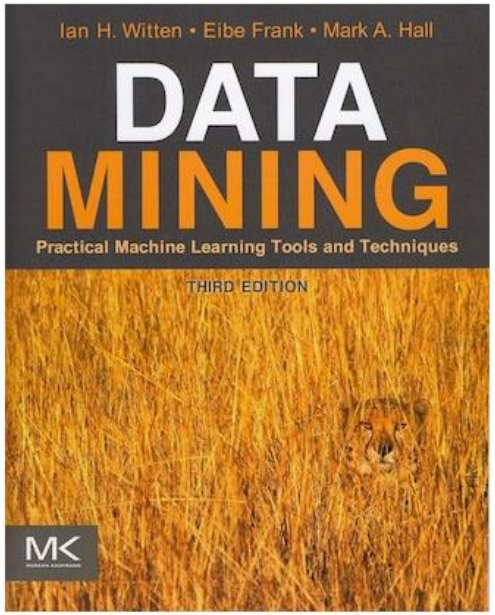
Weka

Introduction

- 懷卡托智能分析環境(**W**aikato **E**nvironment for **K**nowledge **A**nalysis)
  - Machine Learning Group at the University of Waikato
- Open source software in java issued under the GNU General Public License
- Use machine learning for data mining tasks
  - Including **Data Pre-processing**、**Classification**、**Regression**、**Clustering**、**Association** and **Visualization**

Introduction

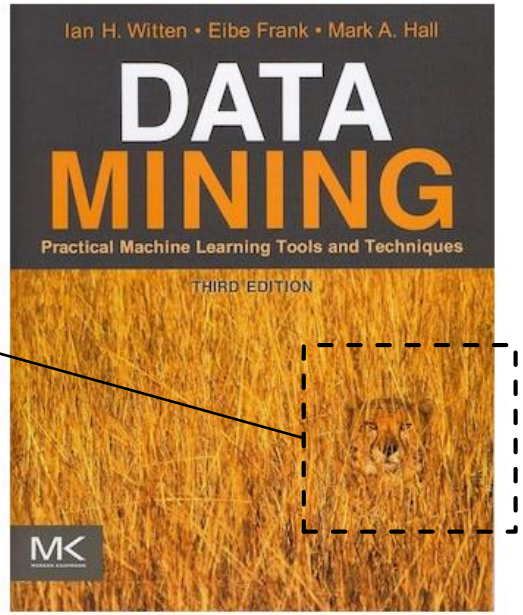
- Data Mining: Practical Machine Learning Tools and Techniques （3rd Edition）
  - It was written a companion book for the WEKA software
  - Shows you how to use the WEKA machine learning workbench



Introduction

- Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition)
  - It was written a companion book for the WEKA software
  - Shows you how to use the WEKA machine learning workbench

Is the computer able to find an animal on this image?



Advantages of WEKA

- Free availability
- Portability
- Fully implemented in the java programming language
- Combine data preprocessing and modeling techniques
- Easy to use GUI
- Provides access to SQL databases

Download

- http://www.cs.waikato.ac.nz/ml/weka/index.html

WEKA
The University of Waikato

Machine Learning Group at the University of Waikato

Project | Software | Book | Publications | People | Related

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

We have put together several free online courses that teach machine learning and data mining using Weka. Check out the **website for the courses** for details on when and how to enrol. The videos for the courses are available **on Youtube.**

Yes, it is possible to apply Weka to **big data**!

Getting started

- **Requirements**
- **Download**
- **Documentation**
- **FAQ**
- **Getting Help**

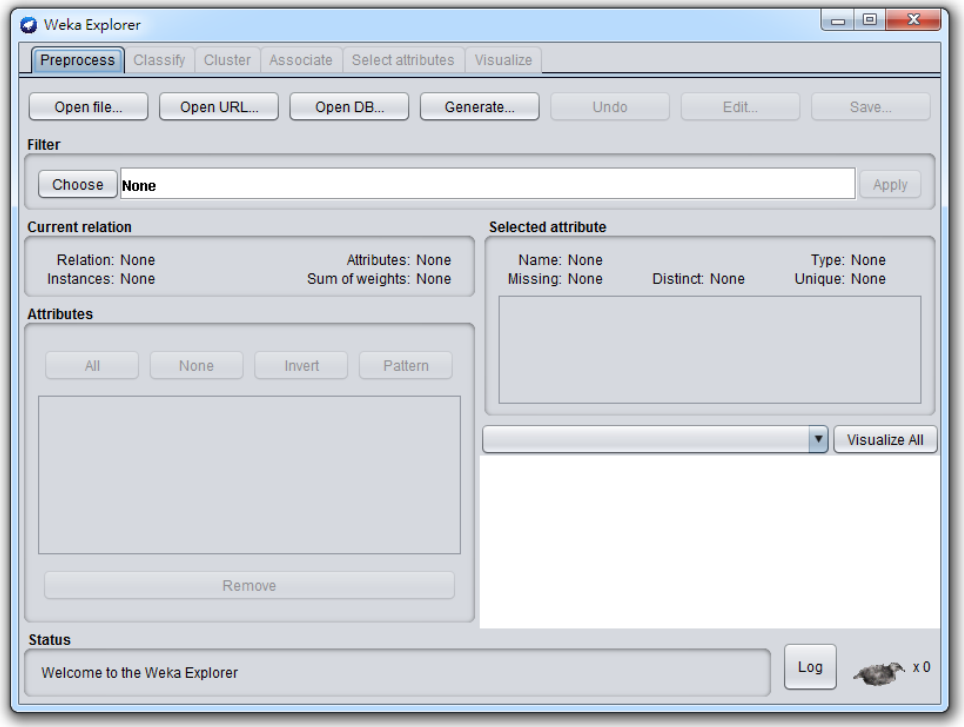
Further information

- **Citing Weka**
- **Datasets**
- **Related Projects**
- **Miscellaneous Code**
- **Other Literature**

Developers

- **Development**
- **History**
- **Subversion**
- **Contributors**

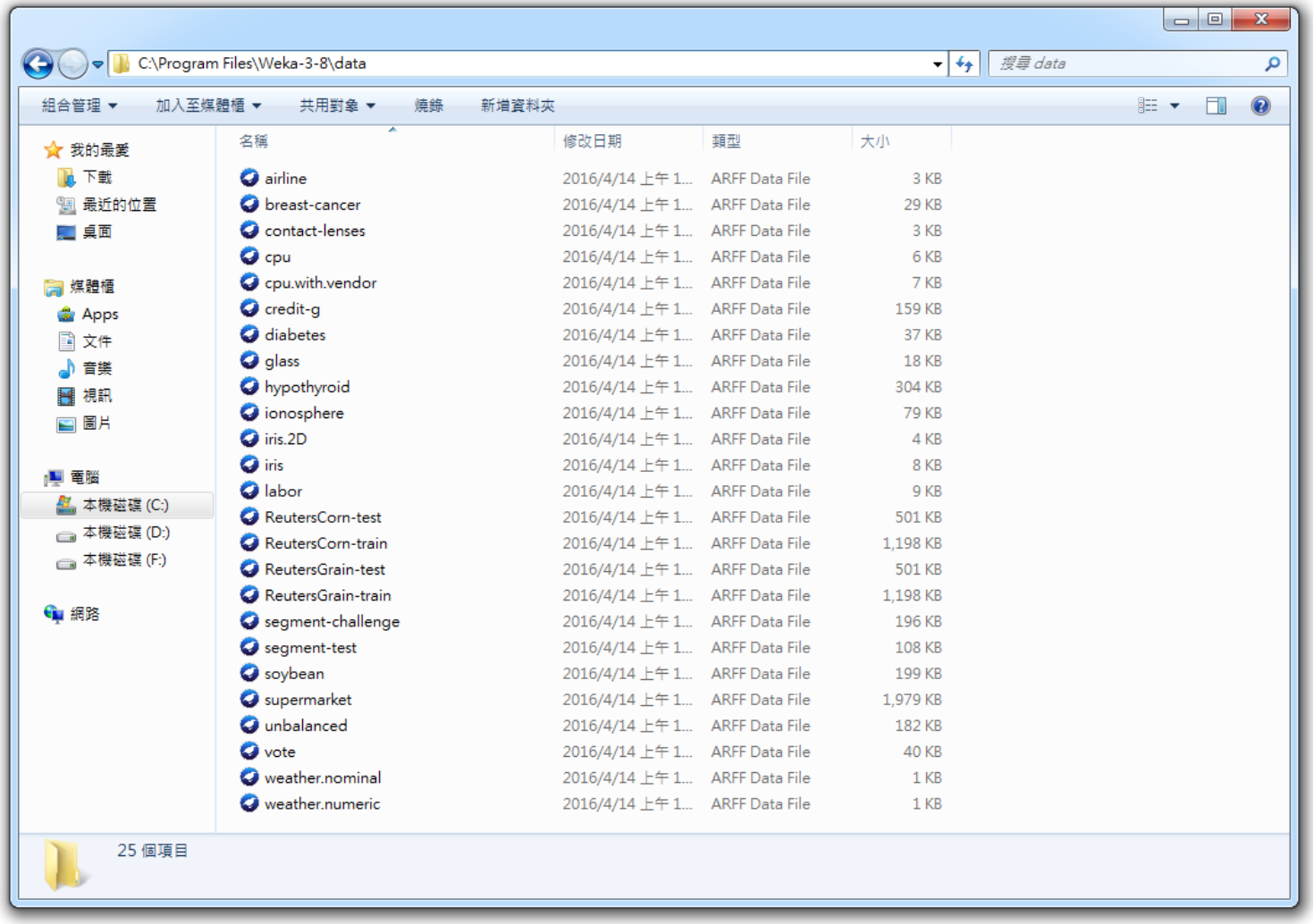
Weka GUI Chooser



WEKA Interface

- kinds of Weka modes
  - Explorer
    - An environment for exploring data with WEKA
  - Experimenter
    - An environment for performing experiments and conducting statistical tests between learning schemes
  - KnowledgeFlow
    - This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning
  - Workbench
    - This environment supports all modes in Weka
- Command line interface

Data in Weka


C:\Program Files\Weka-3-8\data
搜尋 data
組合管理
加入至媒體櫃
共用對象
燒錄
新增資料夾
我的最愛
下載
最近的位置
桌面
媒體櫃
Apps
文件
音樂
視訊
圖片
電腦
本機磁碟 (C:)
本機磁碟 (D:)
本機磁碟 (F:)
網路
名稱
修改日期
類型
大小
airline 2016/4/14 上午 1... ARFF Data File 3 KB
breast-cancer 2016/4/14 上午 1... ARFF Data File 29 KB
contact-lenses 2016/4/14 上午 1... ARFF Data File 3 KB
cpu 2016/4/14 上午 1... ARFF Data File 6 KB
cpu.with.vendor 2016/4/14 上午 1... ARFF Data File 7 KB
credit-g 2016/4/14 上午 1... ARFF Data File 159 KB
diabetes 2016/4/14 上午 1... ARFF Data File 37 KB
glass 2016/4/14 上午 1... ARFF Data File 18 KB
hypothyroid 2016/4/14 上午 1... ARFF Data File 304 KB
ionosphere 2016/4/14 上午 1... ARFF Data File 79 KB
iris.2D 2016/4/14 上午 1... ARFF Data File 4 KB
iris 2016/4/14 上午 1... ARFF Data File 8 KB
labor 2016/4/14 上午 1... ARFF Data File 9 KB
ReutersCorn-test 2016/4/14 上午 1... ARFF Data File 501 KB
ReutersCorn-train 2016/4/14 上午 1... ARFF Data File 1,198 KB
ReutersGrain-test 2016/4/14 上午 1... ARFF Data File 501 KB
ReutersGrain-train 2016/4/14 上午 1... ARFF Data File 1,198 KB
segment-challenge 2016/4/14 上午 1... ARFF Data File 196 KB
segment-test 2016/4/14 上午 1... ARFF Data File 108 KB
soybean 2016/4/14 上午 1... ARFF Data File 199 KB
supermarket 2016/4/14 上午 1... ARFF Data File 1,979 KB
unbalanced 2016/4/14 上午 1... ARFF Data File 182 KB
vote 2016/4/14 上午 1... ARFF Data File 40 KB
weather.nominal 2016/4/14 上午 1... ARFF Data File 1 KB
weather.numeric 2016/4/14 上午 1... ARFF Data File 1 KB
25 個項目

Data Format

- Use flat text files to describe the data
- Data can be imported from a file in various formats and read from a URL or from SQL database (using JDBC)
  - **ARFF**、**CSV**、C4.5、binary.

Data Format

```
@RELATION iris
```

```
@ATTRIBUTE sepallength  REAL
@ATTRIBUTE sepalwidth   REAL
@ATTRIBUTE petallength  REAL
@ATTRIBUTE petalwidth   REAL
@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
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
```

sparse data

```
@data
0, X, 0, Y, "class A"
0, 0, W, 0, "class B"
```

```
@data
{1 X, 3 Y, 4 "class A"}
{2 W, 4 "class B"}
```

Analysis case: classification <Iris data>

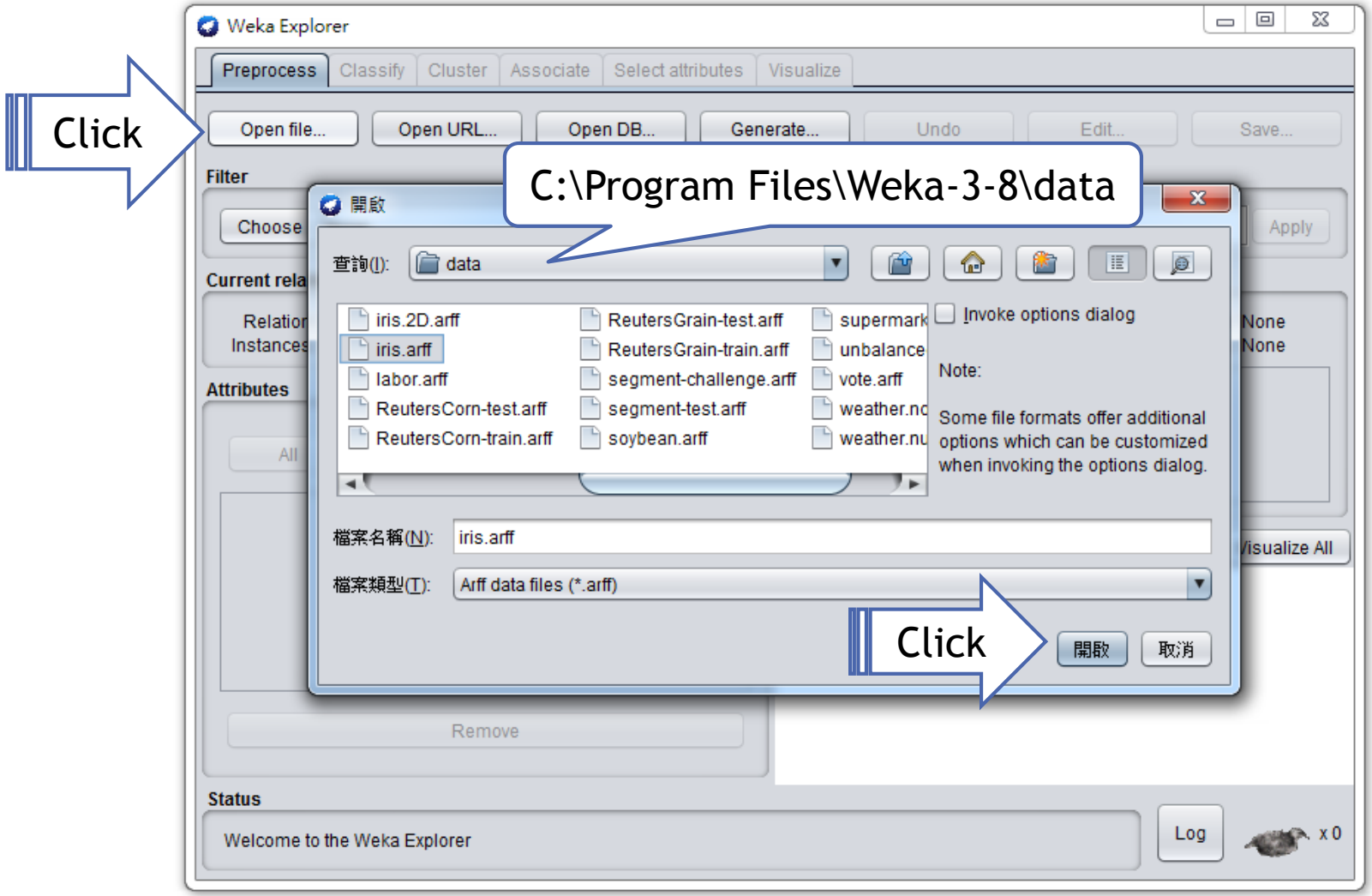
- The best known database to be found in the pattern recognition
- The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant
- One class is linearly separable from the other 2
- Predicted attribute: class of iris plant.

Iris Setosa

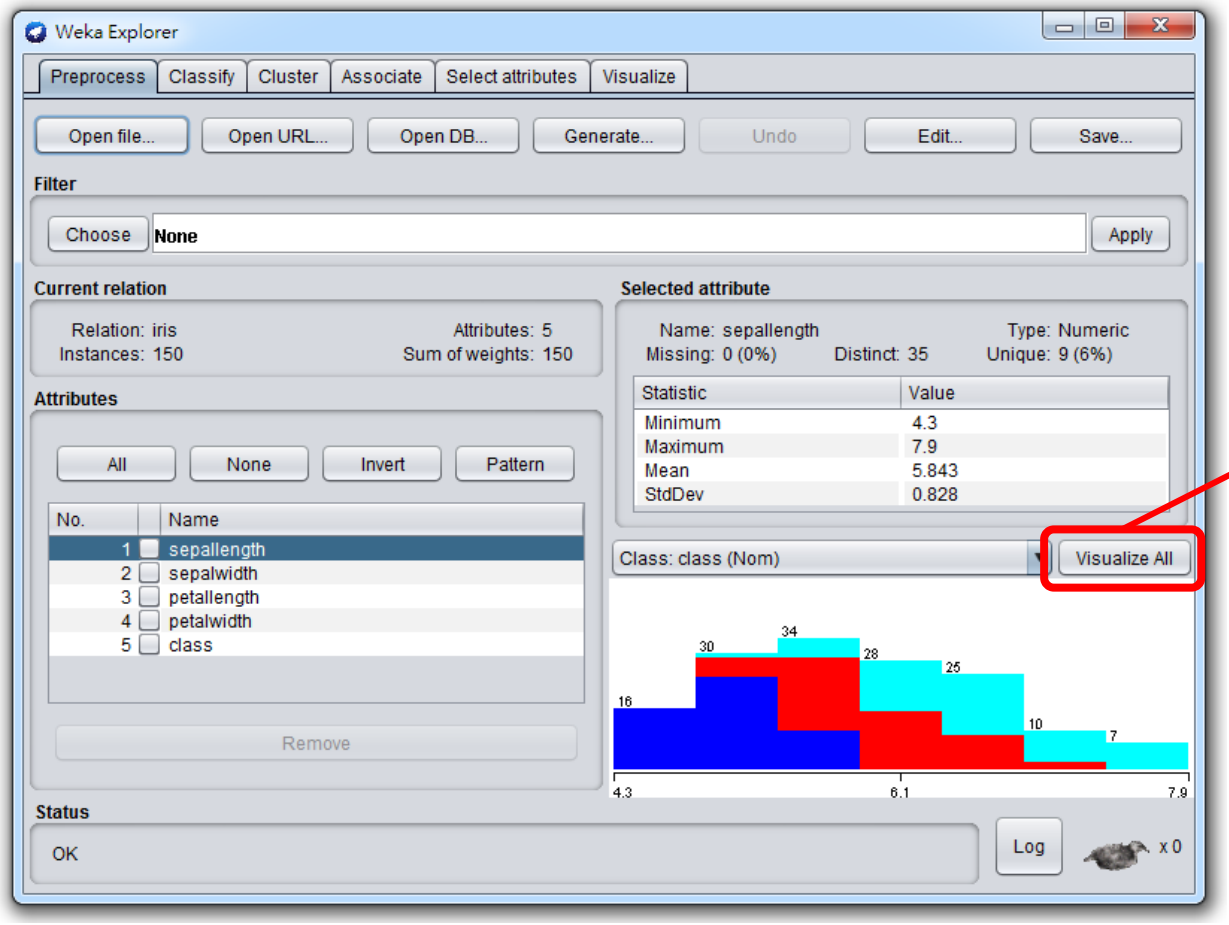
Iris Versicolour

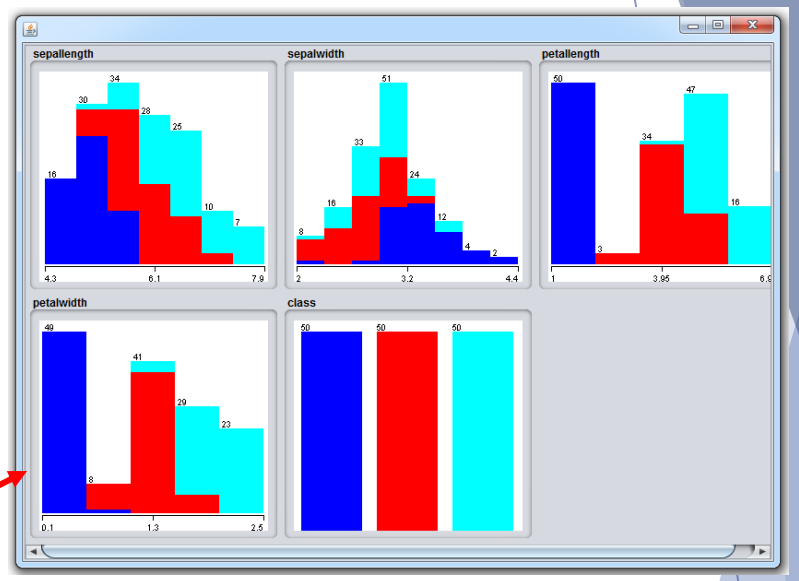
Iris Virginica

Analysis case: classification <Iris data>



Analysis case: classification <Iris data>

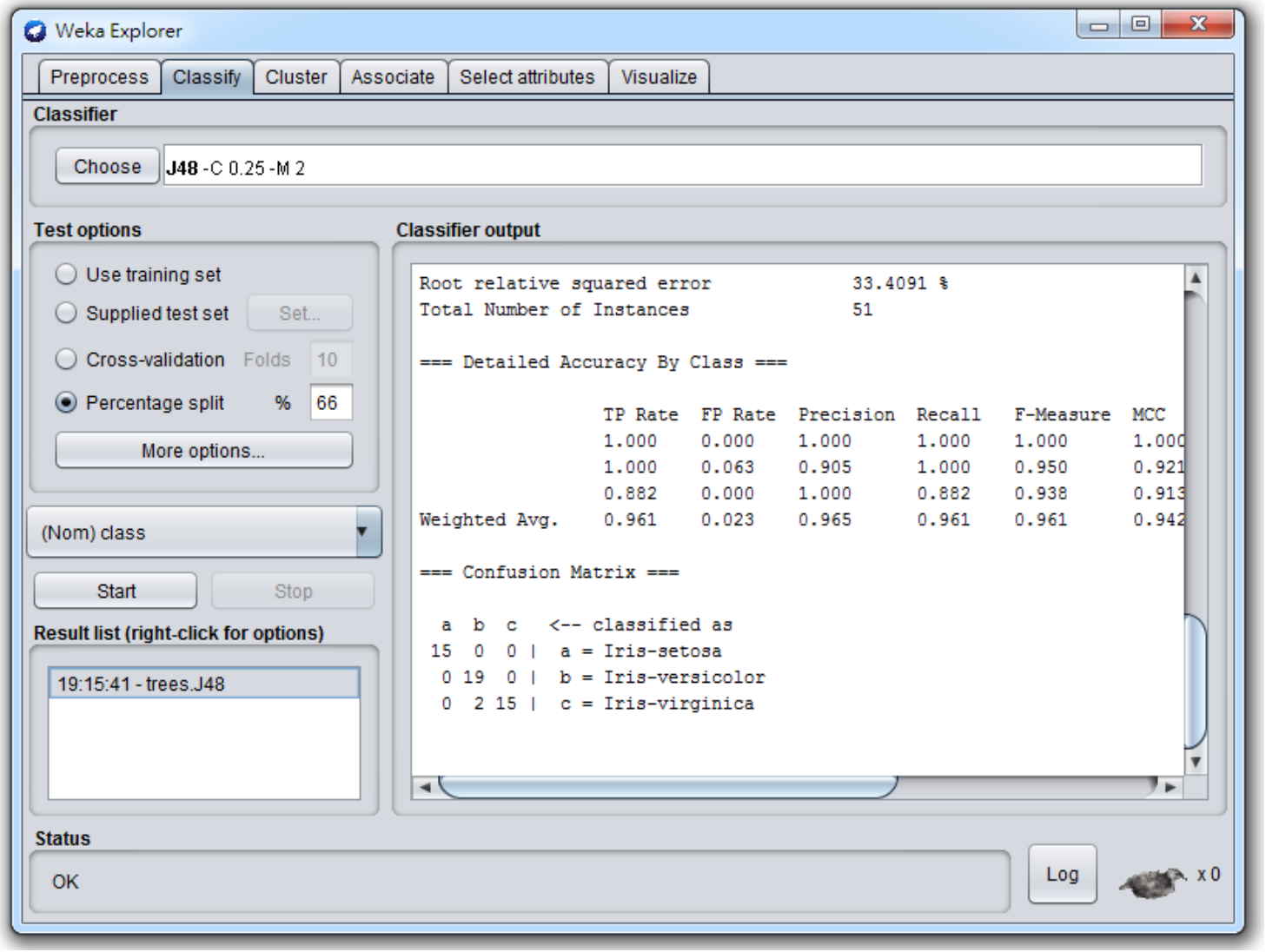




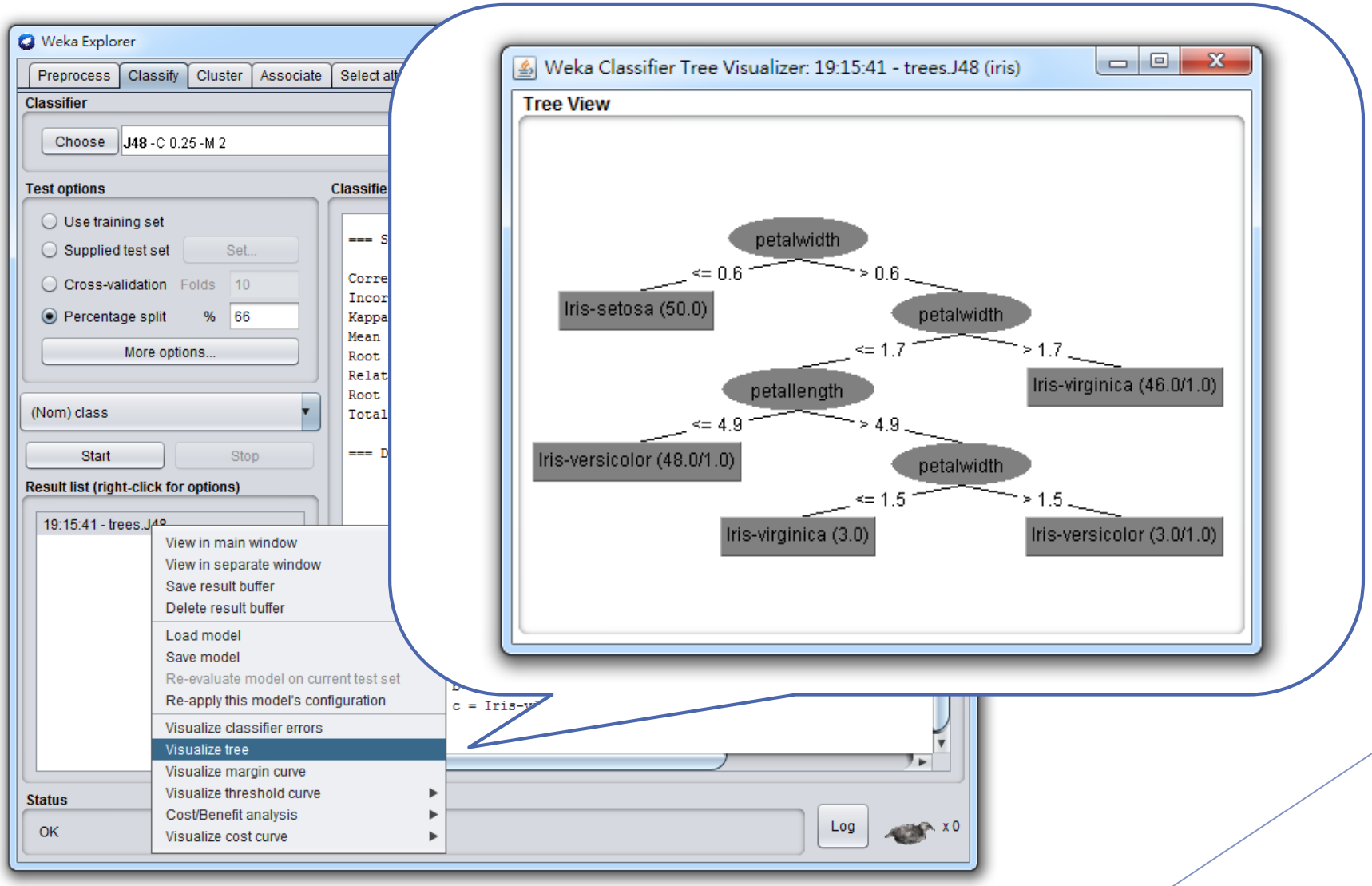
Classifiers

- Classifiers in WEKA are the models for predicting nominal or numeric quantities.
- The learning schemes available in WEKA include
  - Decision trees
  - Instance-based classifiers
  - Support vector machines
  - Logistic regression
  - Bayes' net
  - Meta classifiers

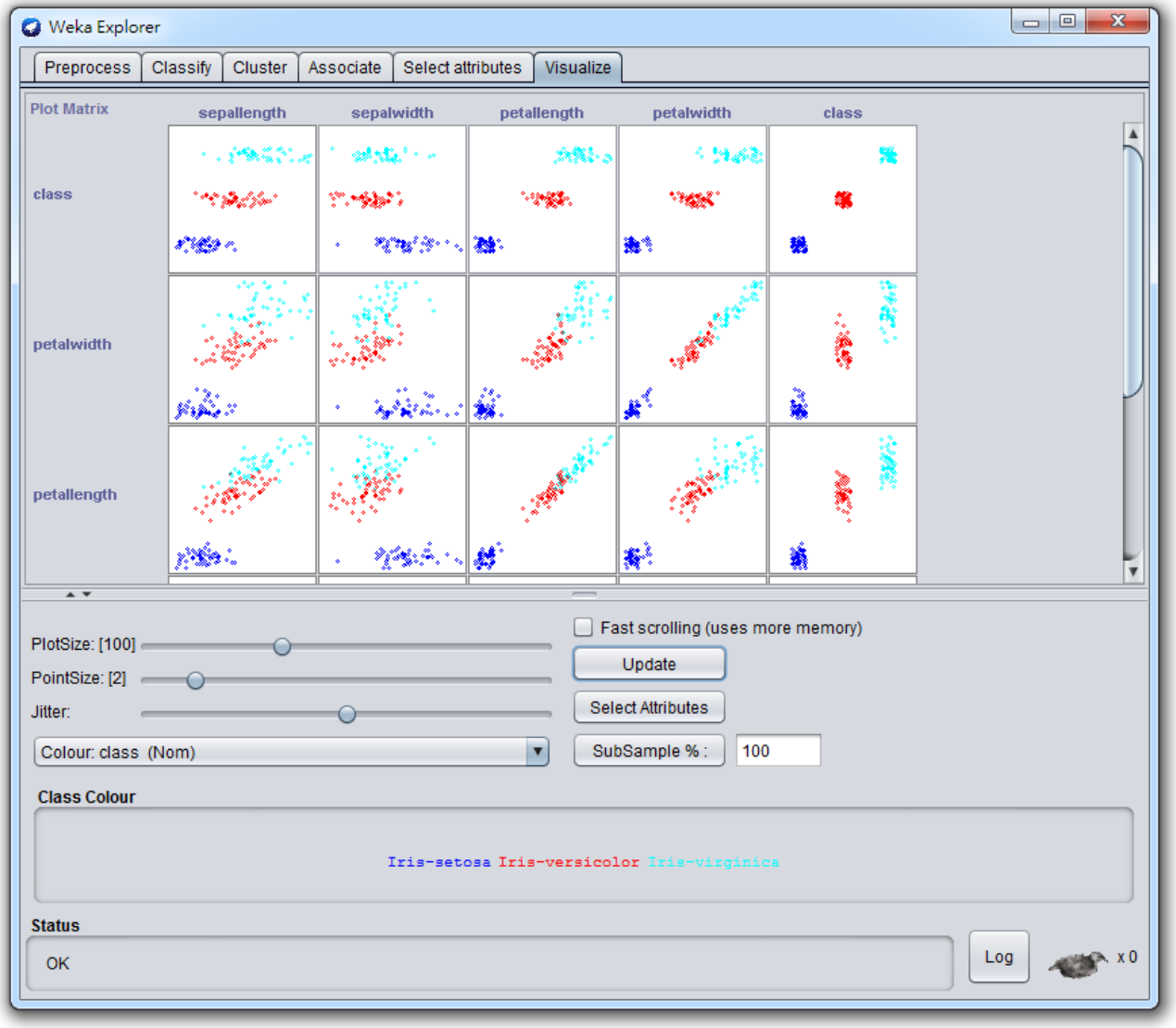
Analysis case: classification <Iris data>



Analysis case: classification <Iris data>

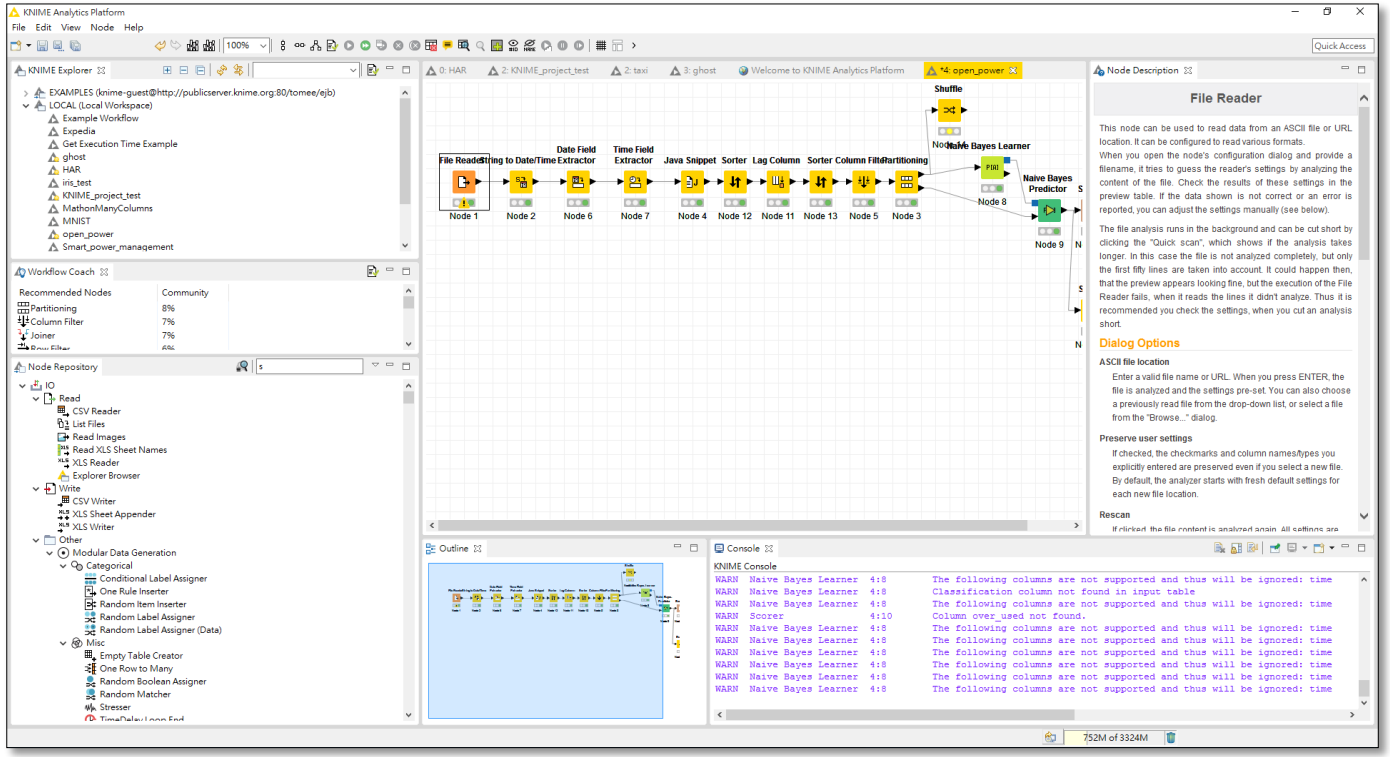


Analysis case: classification <Iris data>



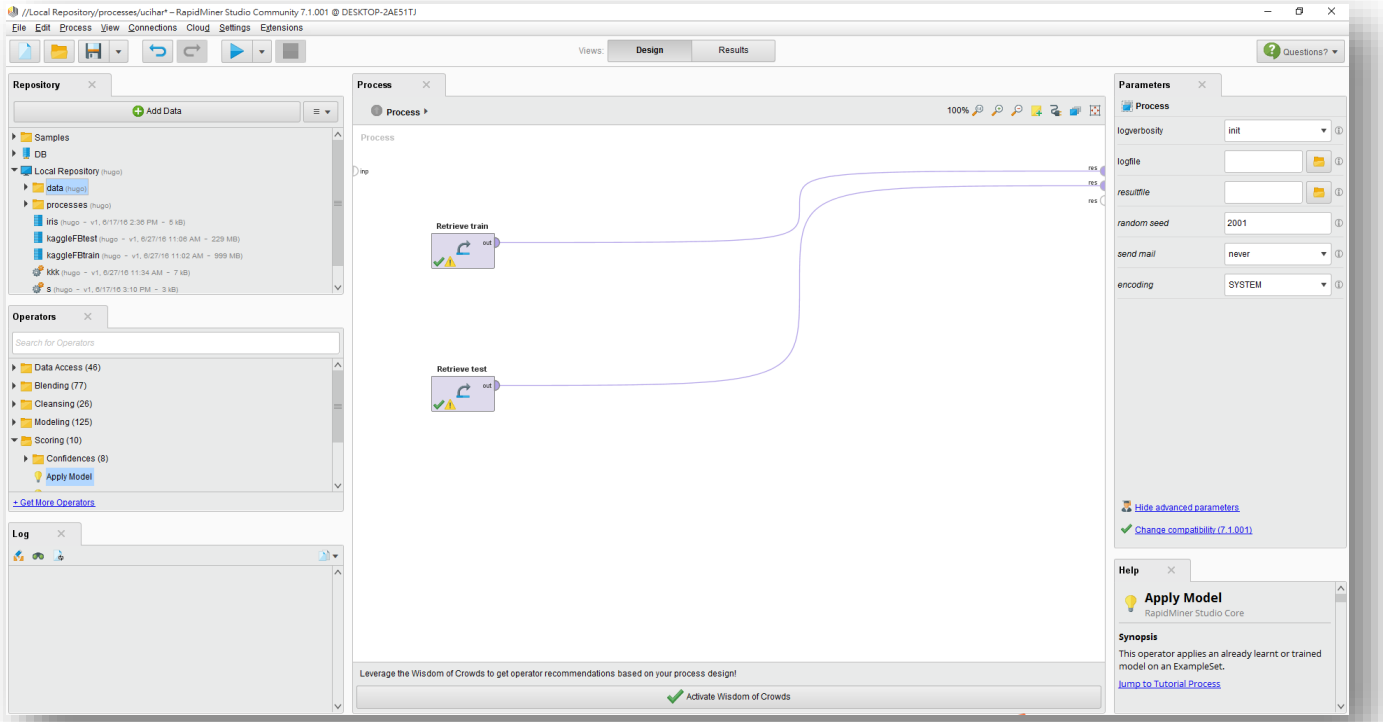
Other Java based Platforms

- KNIME (Konstanz Information Miner)
  - http://www.knime.org/



Other Java based Platforms

- RapidMiner
  - http://rapidminer.com/



R

Introduction

- Statistical computing and graphics
  - A programming language
  - Software environment
- Widely use on
  - Statisticians and data miners
  - Developing statistical software
  - Data analysis
- Interpreted language
  - Users typically access it through a command-line interpreter
- Open source and a powerful tool for data mining and data analysis

Comprehensive R Archive Network (CRAN)

- A network of ftp and web servers around the world
  - Store identical, up-to-date, versions of code and documentation for R
- Currently, the **CRAN** package repository features **7821** available packages
  - Including association rule、sequential patterns、classification、regression and clustering
- http://cran.r-project.org/

Association Rules and Sequential Patterns

- arules
  - Mining Association Rules and Frequent Itemsets.
  - Provides interfaces to C implementations of the association mining algorithms Apriori and Eclat by C. Borgelt.
- arulesSequences
  - Mining Frequent Sequences
  - Provides interfaces to the C++ implementation of cSPADE by Mohammed J. Zaki
- arulesViz
  - Visualizing Association Rules and Frequent Itemsets

Association Rules and Sequential Patterns

```
install.packages('arules')
library('arules')
data("Adult")
rules <- apriori(Adult,parameter = list(supp = 0.5, conf = 0.9, target = "rules"))
labels(rules)
```

```
> labels(rules)
 [1] "{} => {capital-gain=None}"
 [2] "{} => {capital-loss=None}"
 [3] "{hours-per-week=Full-time} => {capital-gain=None}"
 [4] "{hours-per-week=Full-time} => {capital-loss=None}"
 [5] "{sex=Male} => {capital-gain=None}"
 [6] "{sex=Male} => {capital-loss=None}"
```

```
[47] "{workclass=Private,race=White,capital-loss=None} => {capital-gain=None}"
[48] "{workclass=Private,capital-gain=None,native-country=United-States} => {capital-loss=None}"
[49] "{workclass=Private,capital-loss=None,native-country=United-States} => {capital-gain=None}"
[50] "{race=White,capital-gain=None,native-country=United-States} => {capital-loss=None}"
[51] "{race=White,capital-loss=None,native-country=United-States} => {capital-gain=None}"
[52] "{race=White,capital-gain=None,capital-loss=None} => {native-country=United-States}"
```

Classification

- C50
  - C5.0 Decision Trees and Rule-Based Models
- Party
  - A computational toolbox for recursive partitioning.
  - ctree(), provides an implementation of conditional inference trees
  - cforest() provides an implementation of Breiman's random forests
- e1071
  - Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier

Classification

- xgboost
  - Extreme Gradient Boosting
- class
  - Various functions for classification, including k-nearest neighbor, Learning Vector Quantization and Self-Organizing Maps.
- kknn
  - weighted k-Nearest Neighbors for Classification, Regression and Clustering
- nnet
  - Feed-Forward Neural Networks and Multinomial Log-Linear Models

Classification

Package: class

Package: kknn

```
> pred_knn = knn(train_data, test_data, train_lable, k = 1, prob=TRUE)
> pred_kknn = kknn(i ~ . ,pima_train_data,test_data,k=1,distance = 2,kernel="rectangular")
> accuracy_knn = sum(pred == test_lable)/length(test_lable)
> accuracy_kknn = sum(pred2$fitted.values == test_lable)/length(test_lable)
> accuracy_knn
[1] 0.7254902
> accuracy_kknn
[1] 0.7385621
```

Clustering

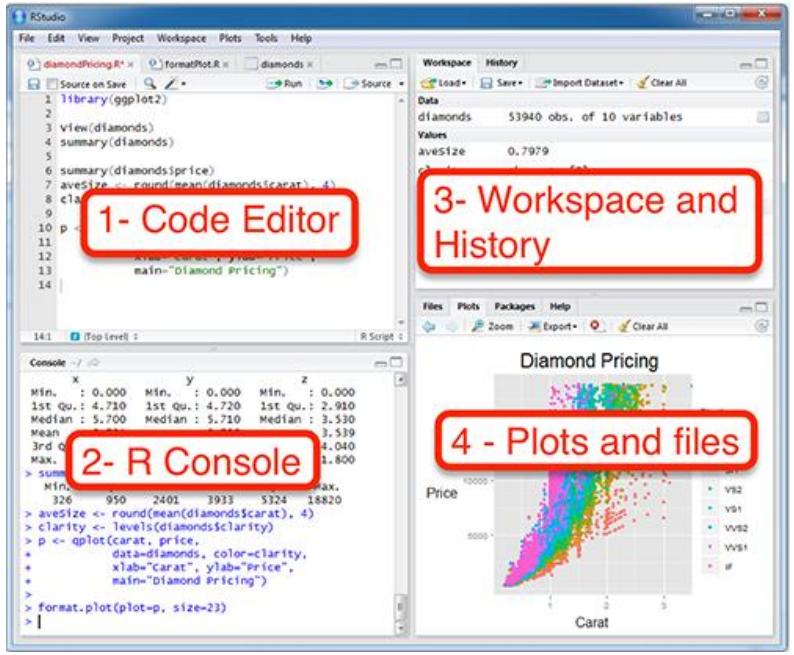
- cluster
  - Provides Hierarchical Clustering methods
- fpc
  - Various methods for clustering and cluster validation. Fixed point clustering. Linear regression clustering. DBSCAN clustering.
- cclust
  - Convex Clustering methods, including K-means algorithm, On-line Update algorithm and Neural Gas algorithm

Clustering

```
> library(cclust)
> data = read.table('C://Users/GYLi/Downloads/taiwan.dat',sep=',')
> data = as.matrix(data)
> op <- options(digits.secs = 6)
> Sys.time()
[1] "2016-01-21 15:50:35.105982 CST"
> km = kmeans(data,centers = 5,iter.max = 1000)
> Sys.time()
[1] "2016-01-21 15:50:39.44922 CST"
> km_cclust = cclust(data,centers = 5,iter.max = 1000,dist = "euclidean",method = "kmeans")
> Sys.time()
[1] "2016-01-21 15:50:53.413028 CST"
> sum(km$withinss)
[1] 680244.7
> sum(km_cclust$withinss)
[1] 680244.7
> Sys.time()
[1] "2016-01-21 15:50:53.41502 CST"
> km = kmeans(data,centers = 10,iter.max = 1000)
> Sys.time()
[1] "2016-01-21 15:51:00.728446 CST"
> km_cclust = cclust(data,centers = 10,iter.max = 1000,dist = "euclidean",method = "kmeans")
> Sys.time()
[1] "2016-01-21 15:51:34.224363 CST"
> sum(km$withinss)
[1] 268327.8
> sum(km_cclust$withinss)
[1] 268327.8
> km$iter
[1] 7
> km_cclust$iter
[1] 76
```

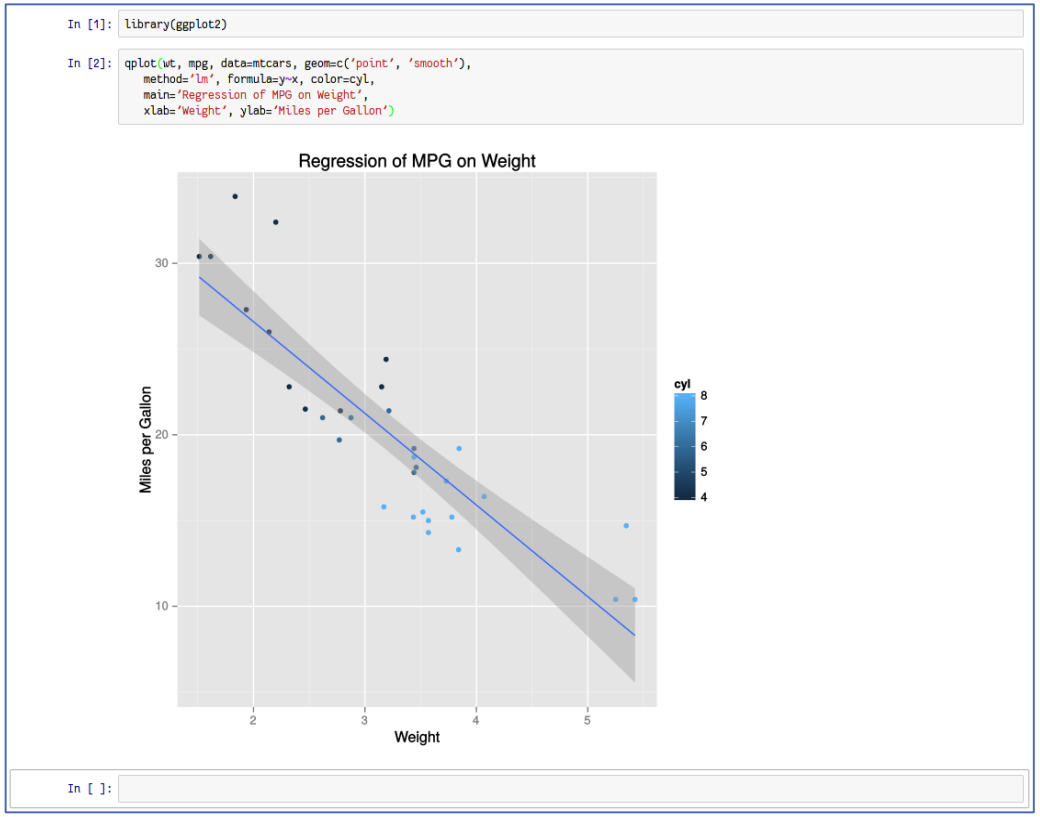
RStudio

- http://www.rstudio.com/
- An integrated development environment (IDE) for R
  - Including console, syntax-highlighting editor, plotting, history, debugging and workspace management



IRkernel: R kernel for Jupyter

- http://irkernel.github.io/



Python - scikit-learn

Introduction

- Scikit-learn is a python module
  - State-of-art machine learning algorithms for medium-scale problems
- Built on NumPy、SciPy and matplotlib
- Open source、commercially usable

Supervised Learning

- sklearn.linear_model: Generalized Linear Models (38)
  - Logistic regression, Ridge Regression, Lasso, Stochastic Gradient Descent...
- sklearn.svm: Support Vector Machines (9)
- sklearn.neighbors: Nearest Neighbors (13)
- sklearn.naive_bayes: Naive Bayes (3)
- sklearn.tree: Decision Trees (5)
- sklearn.ensemble: Ensemble Methods (14)
  - Random Forests, AdaBoost, Gradient Tree Boosting...

Supervised Learning : randomforest

```
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
```

Import modules

```
X = [[0, 0, 1], [1, 1, 1]]
y = [0, 1]
```

Data

```
clf = RandomForestClassifier(n_estimators=10, n_jobs=-1)
```

Call the classifier

```
clf.fit(X, y)
y_pred = clf.predict([0.5, 1, 0.5])
y_pred_proba = clf.predict_proba([0.5, 1, 0.5])
```

Build model and predict

```
print clf.feature_importances_
print metrics.accuracy_score(y, y_pred)
```

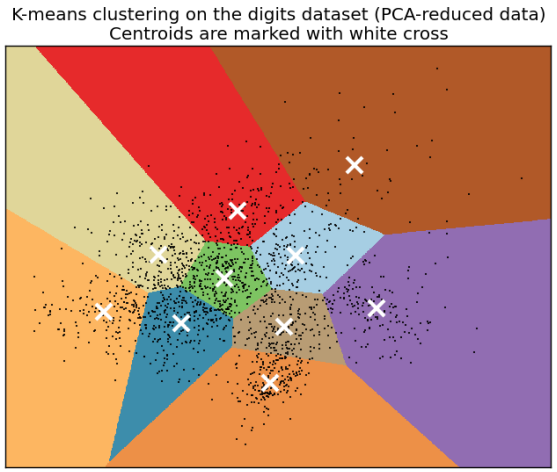
Evaluation and see feature importances

Unsupervised Learning

- sklearn.cluster: Clustering (9)
  - K-means, DBSCAN, Mean Shift, Spectral clustering…
- sklearn.cluster.bicluster: Biclustering (2)

Unsupervised Learning : k-means

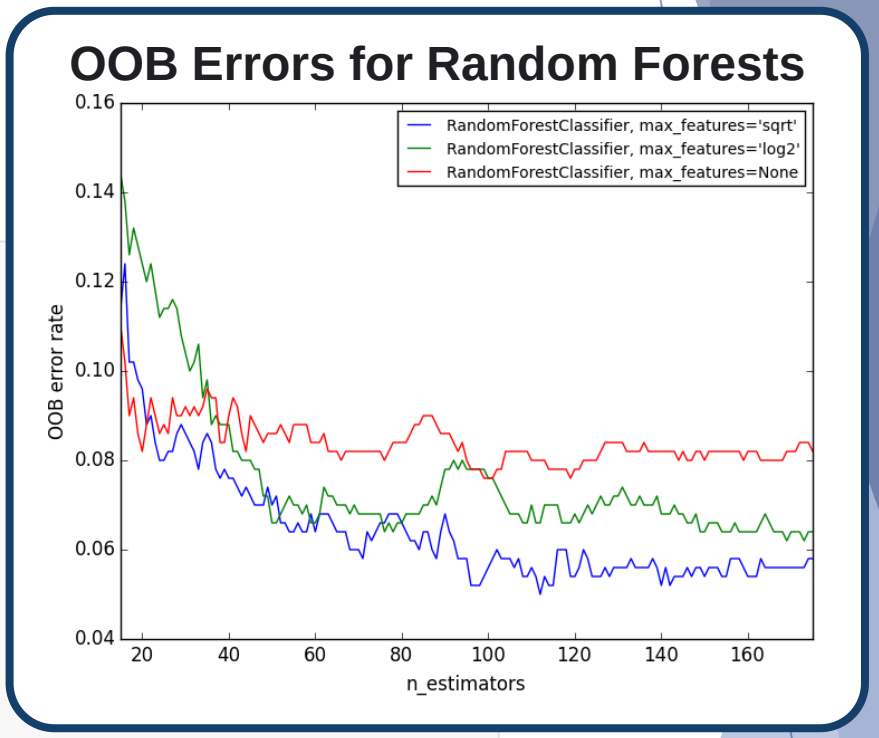
```
from sklearn.cluster import Kmeans

clf = Kmeans(n_clusters=10)
clf.fit(X)
pred = clf.predict(pred_X)

print clf.cluster_centers_          # Coordinates of cluster centers
print clf.labels_                   # Labels of each point
```



Online Document

Python source code: `plot_ensemble_oob.py`

```
import matplotlib.pyplot as plt

from collections import OrderedDict
from sklearn.datasets import make_classification
from sklearn.ensemble import RandomForestClassifier, ExtraTreesClassifier

# Author: Kian Ho <hui.kian.ho@gmail.com>
#         Gilles Louppe <g.louppe@gmail.com>
#         Andreas Mueller <amueller@ais.uni-bonn.de>
#
# License: BSD 3 Clause

print(__doc__)

RANDOM_STATE = 123

# Generate a binary classification dataset.
X, y = make_classification(n_samples=500, n_features=25,
                           n_clusters_per_class=1, n_informative=15,
                           random_state=RANDOM_STATE)
```



IPython

IP[y]: IPython
Interactive Computing

jupyter

- http://ipython.org/
- A powerful interactive shell
- A kernel for Jupyter
  - http://jupyter.org/index.html
- Support for interactive data visualization and use of GUI toolkits
- Flexible, embeddable interpreters to load into your own projects
- Easy to use, high performance tools for parallel computing

IPython

IP[y]: IPython
Interactive Computing

IP[y]: Notebook spectrogram Last Checkpoint: a few seconds ago (autosaved)

IPython (Python 3)

File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

Simple spectral analysis

An illustration of the Discrete Fourier Transform using windowing, to reveal the frequency content of a sound signal.

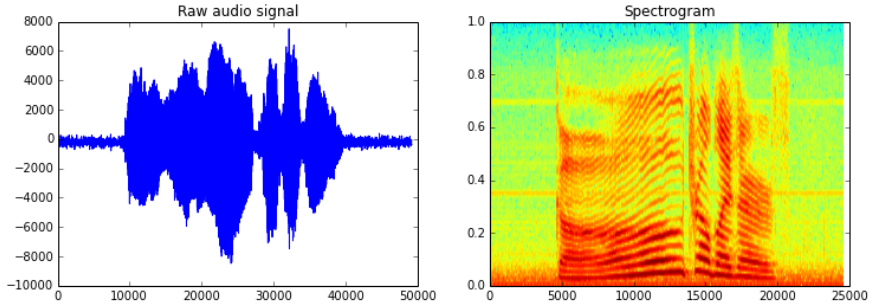
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

We begin by loading a datafile using SciPy's audio file support:

In [1]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin `specgram` routine:

In [2]:
```
%matplotlib inline
from matplotlib import pyplot as plt
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```



Python packages

- PIP
  - http://pypi.python.org/pypi/pip
  - A tool for installing and managing Python packages
- Anaconda
  - http://www.continuum.io/
  - Open data science platform powered by Python
  - The open source version of Anaconda
    - Including over 100 of the most popular Python, R and Scala packages for data science