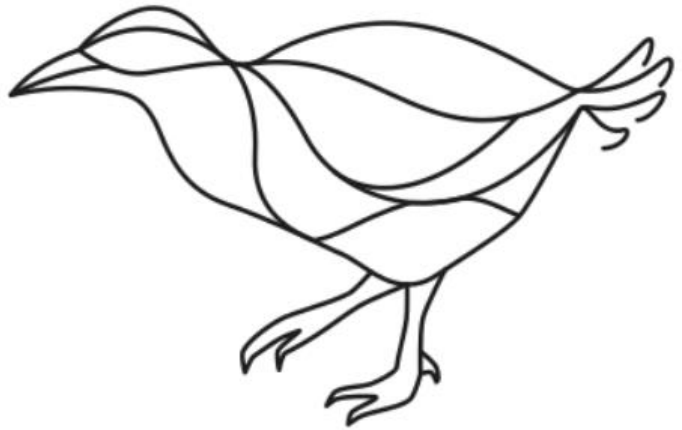
Introduction to Weka

Guilherme Alves

guilherme.alves-da-silva@univ-lorraine.fr

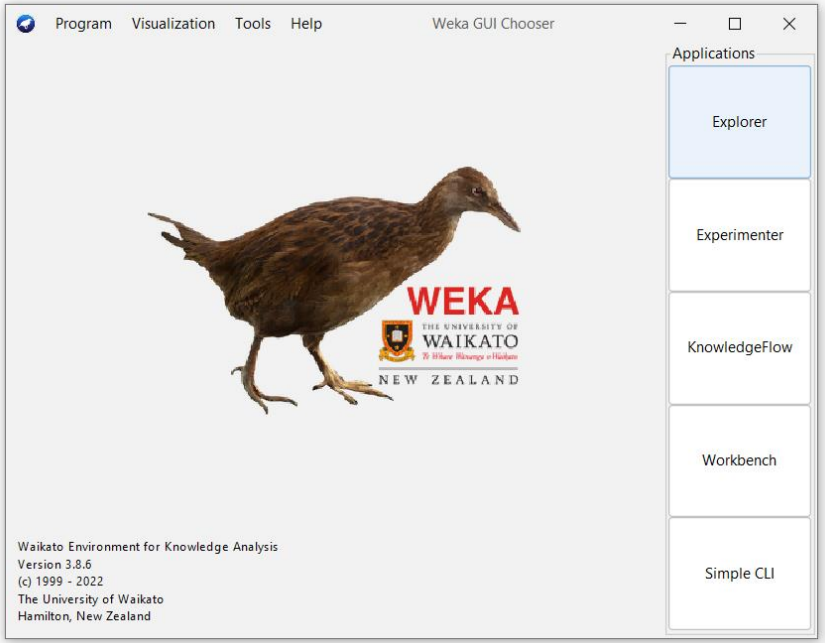
WEKA: The Tool

- Open source & free
- Graphical interface
  - no code ;)
- Developed and maintained by the University of Waikato, NZ



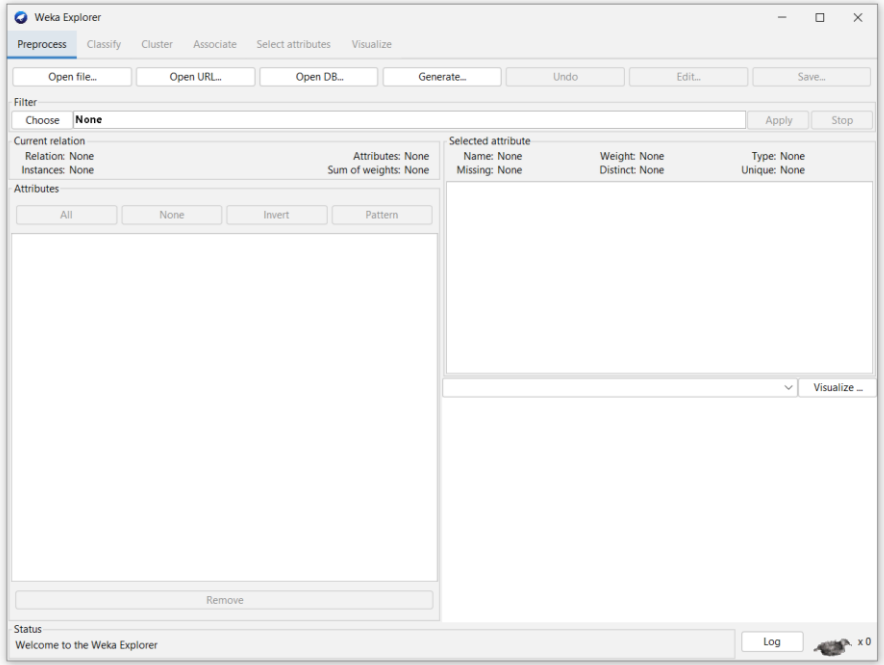
Weka: Main Interface

- Check if you have Java (JRE) installed in your machine
  - In a terminal:
    - command `java --version`
- Launch Weka
- Go to “Explorer”



Weka: Explorer Interface

- **Preprocess**: select and process attributes (filters, deletion, etc)
- **Classify**: train and test models for classification and regression
- **Cluster**: perform cluster analysis
- **Associate**: learn association rules
- **Select attributes**: methods to select the most relevant attributes
- **Visualize**: tools to visualize the data (2D)



Hands on:
Training a classifier

Iris dataset

Weka: Explorer Interface

- Supported file formats:
  - CSV, ARFF (Attribute-Relation File Format), etc.
- CSV: Comma-Separated Values
  - Text file separated by , (comma)
  - You can open it in a plain text editor
  - Each line: one object (data instance)

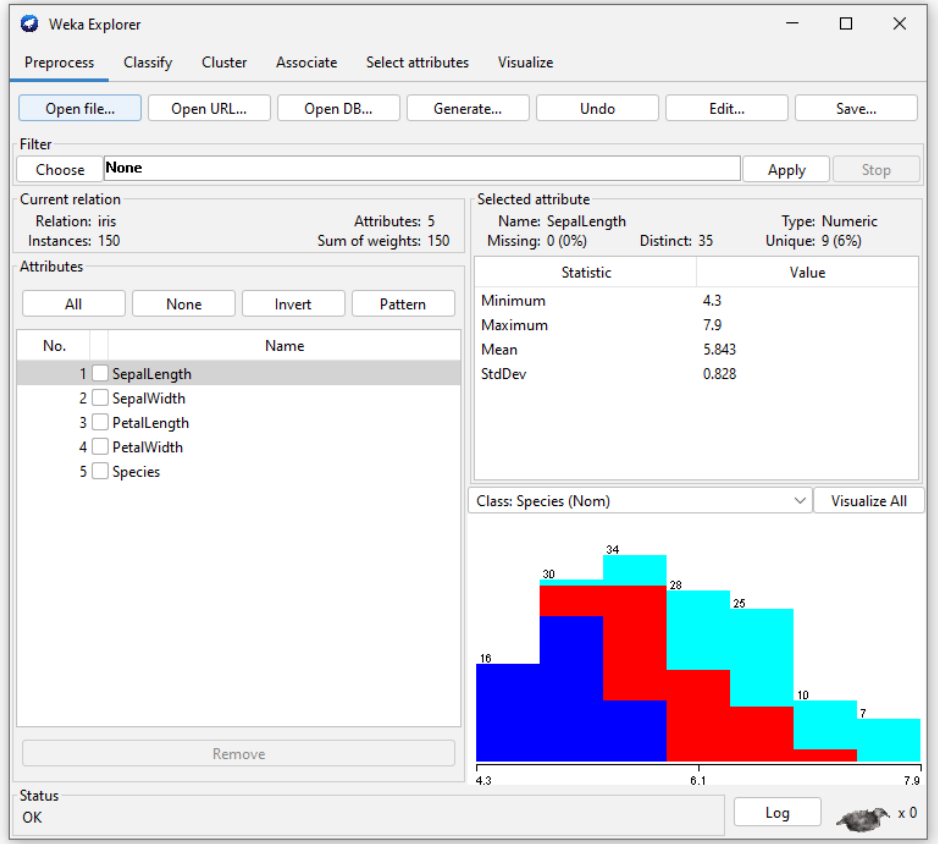
SepalLength	SepalWidth	PetalLength	PetalWidth	Species (CLASS)
5.1	3.8	1.9	0.4	Iris-setosa
4.8	3.0	1.4	0.3	Iris-setosa
5.1	3.8	1.6	0.2	Iris-setosa
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor

iris.csv

```
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
```

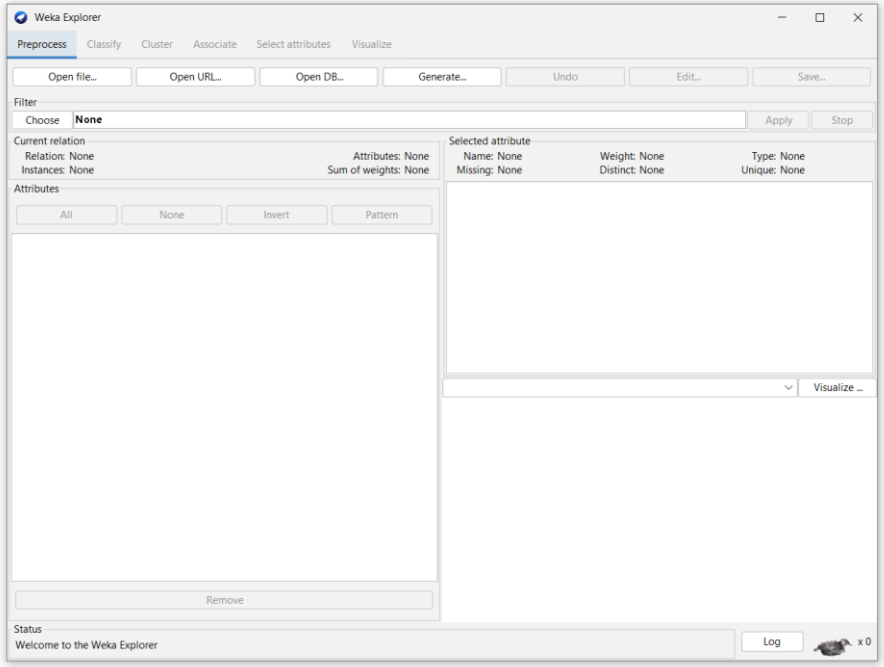
Weka: Preprocess

- **Loaders**: load data into Weka…
  - …from a file, a URL, a database (DB) or
  - …from a generator (synthetic data)
- **Filters**: manipulate the data
  - e.g. convert the type of attributes
    - numeric → categorical
- **Attributes**: list all attributes loaded
  - You can remove attributes
- **Selected attribute**: shows statistics about an attribute + histogram



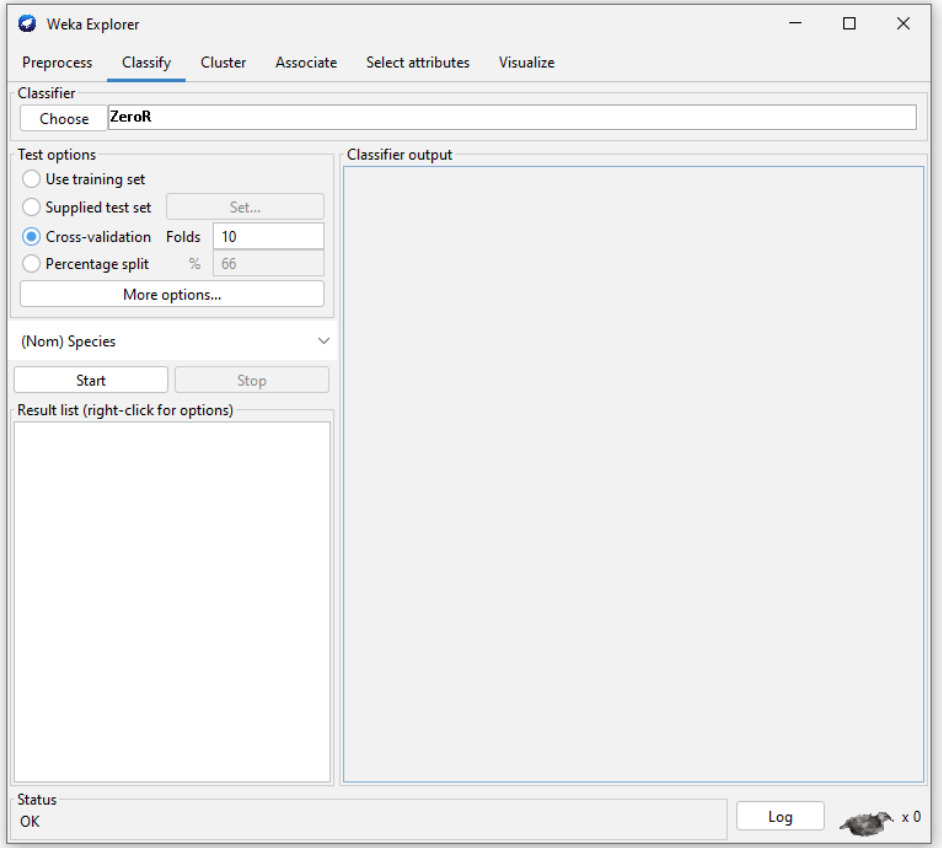
Weka: Preprocess

1. Load the dataset
   - Button “Open file…”
2. Check
   - The distribution of each feature (attribute)
   - Should an attribute (or more) be removed?
   - Should you pre-process any attribute?
3. Then go to
   - “Classify” tab



Weka: Classify

- **Classifier**: lists all installed algorithms that can train a classifier
- **Test options**: indicates how the classifier is evaluated
  - Is the training set used as test set as well?
  - Should Weka split the dataset into training and test sets? How?
- **Classifier output**: details the obtained results
- **Result list**: lists the latest results



Weka: learning a decision tree

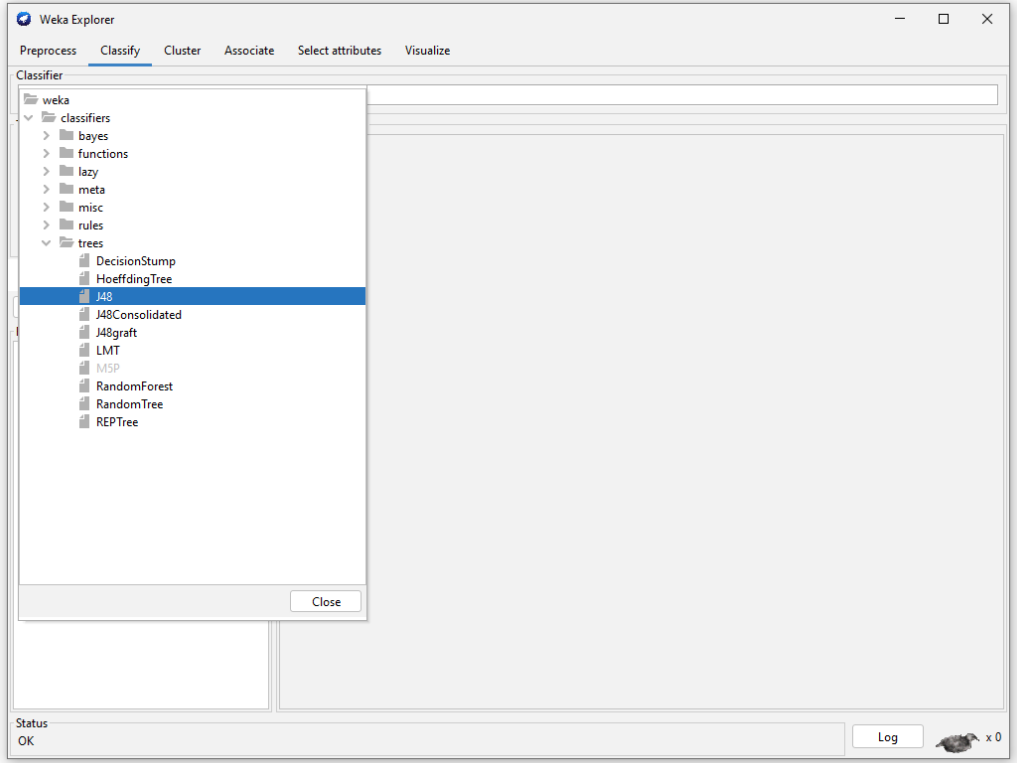
1. **Experimental setup**
   - **Classifier**: select the algorithm J48: *weka / classifiers / tree / J48*
   - **Test options**: Percentage-split = 66% (training)
   - Be sure that "Species" is selected as attribute class

2. Launch the algorithm by clicking at "Start"

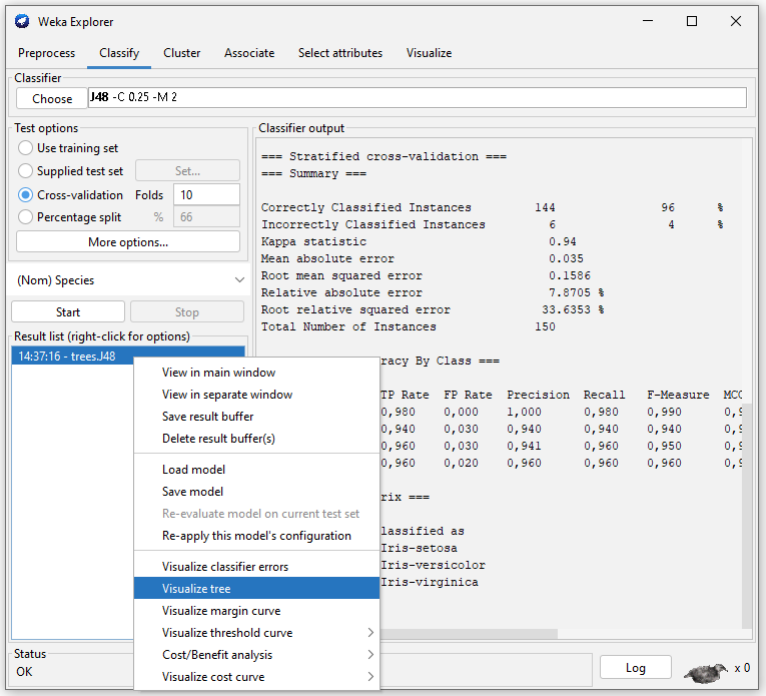
3. **Once the training process is finished**: Status bar = "OK"
   - Look at the **classifier output** box!

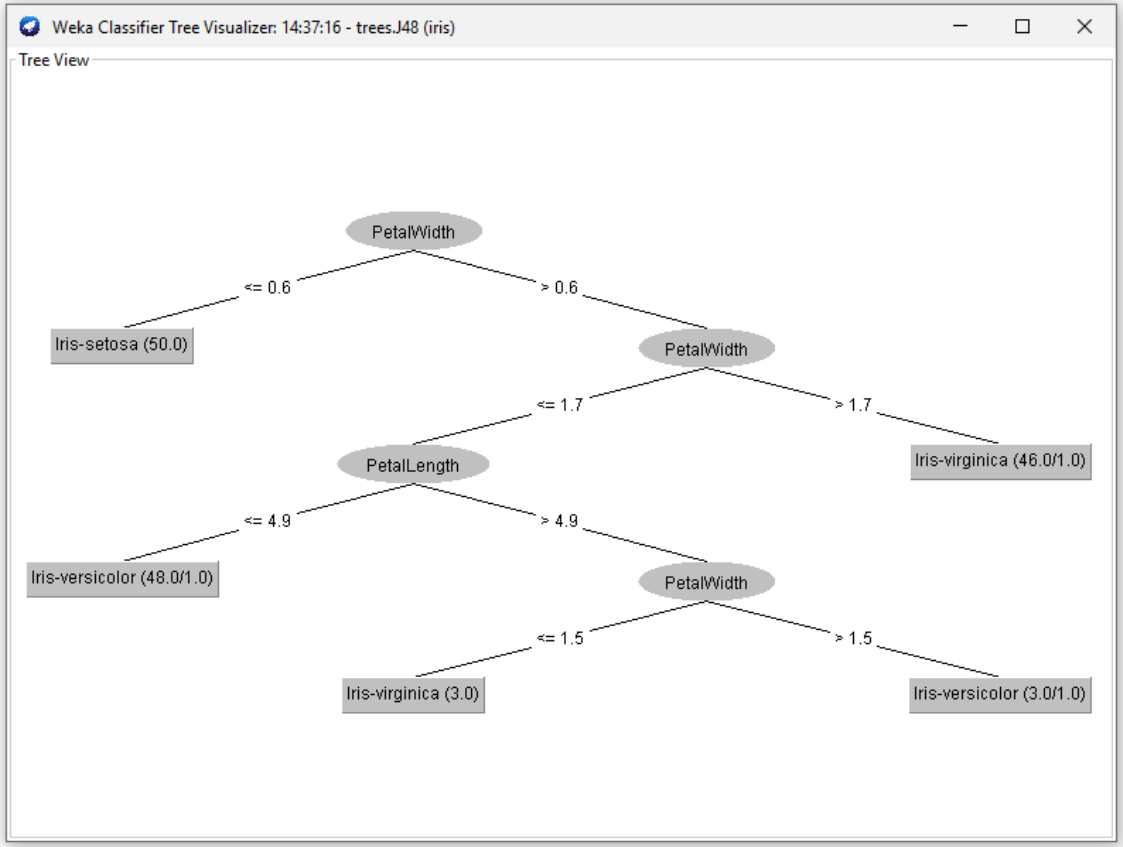
4. **Questions**:
   a) What do you see?
   b) How well is the performance of this decision tree?
   c) What else?

5. **Can you build a different decision tree/classifier?**



Visualizing a decision tree





Weka: Visualize the data