ID Tree generation and visualization with Weka (run example in Weka)

Identification Trees

Last, but not least, an ID tree is a decision tree in which all possible divisions is created by training the tree against a list of known data. The purpose of an ID tree is to take a set of sample data, classify the data and construct a series of test to classify an unknown object based on like properties.

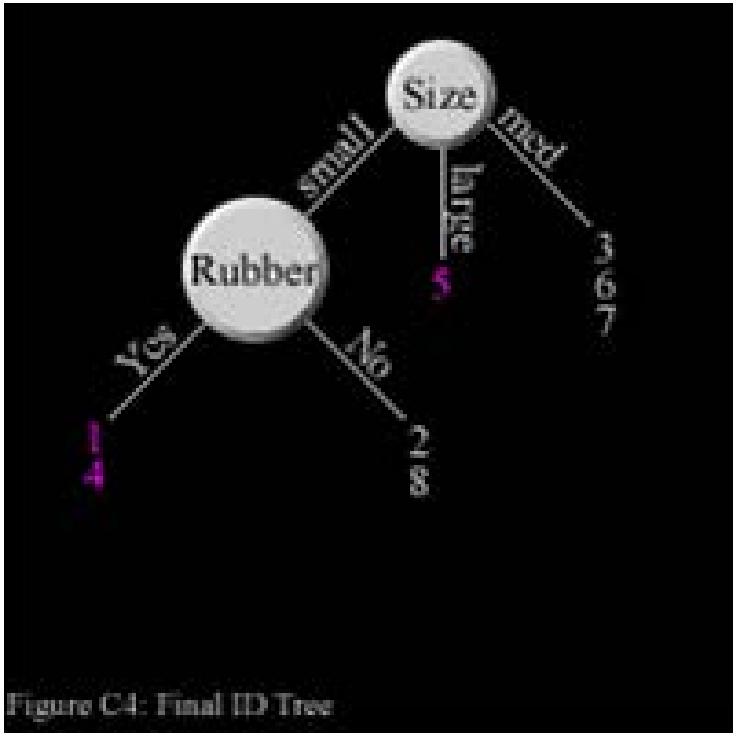
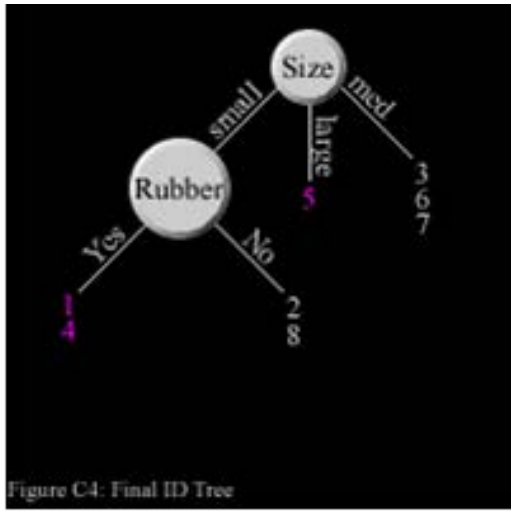


Figure C4: Final ID Tree

Refined Rules



R1: if (size = large)
then (ball does bounce)
R2: if (size = medium)
then (ball does not bounce)
R3: if (size = small)
(rubber = no)
then (ball does not bounce)
R4: if (size = small)
(rubber = yes)
then (ball does bounce)

R1: if (size = large)
then (ball does bounce)
R2: if (size = medium)
then (ball does not bounce)
R3: if (rubber = no)
then (ball does not bounce)
R4: if (size = small)
(rubber = yes)
then (ball does bounce)

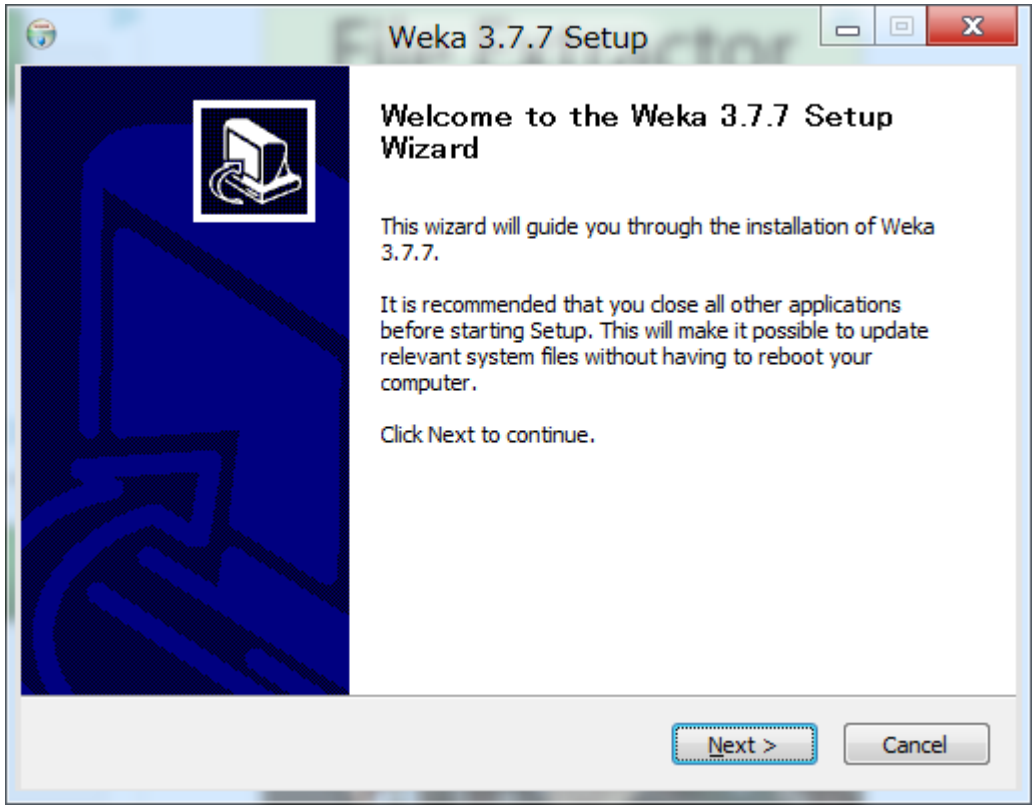
Rules are used in rule-based (forward chaining or backward chaining) systems.

How to get and start Weka

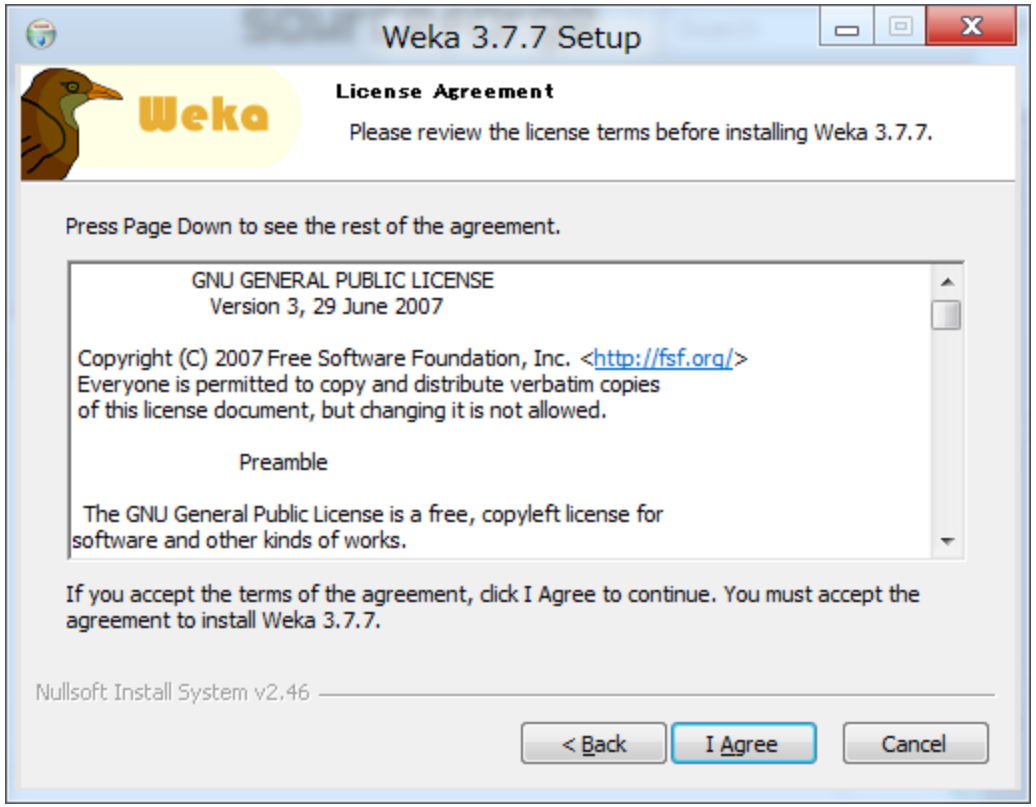
- Go to this page
  http://www.cs.waikato.ac.nz/ml/weka/
- Download Weka from
  Weka-3.7 Weka-download .exe file
- Start Weka
- Press Explorer

Weka Setup

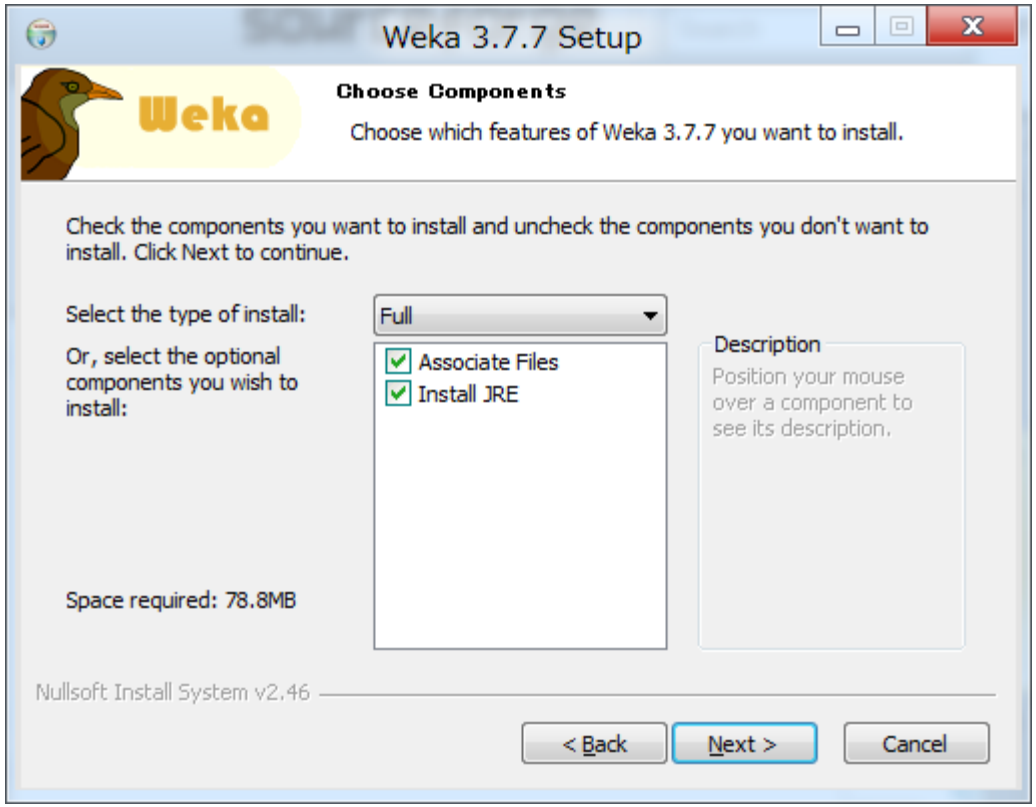
- After download, please run exe file.



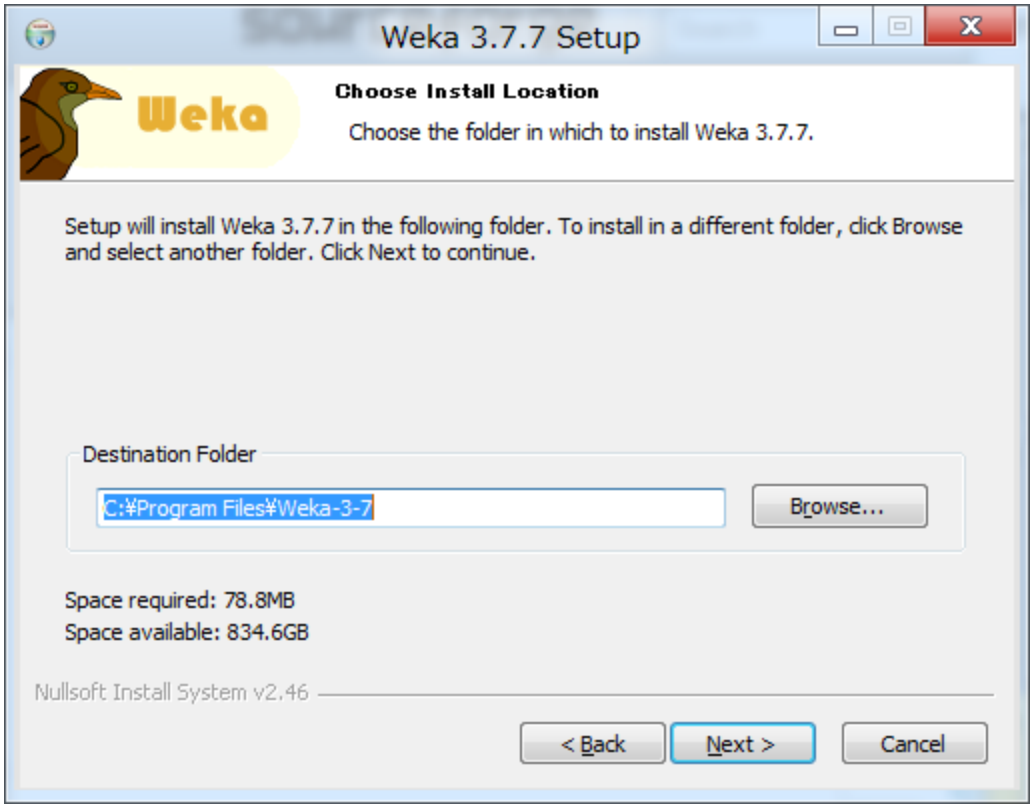
- Agree the license



- Select the type of install “Full” and go next.

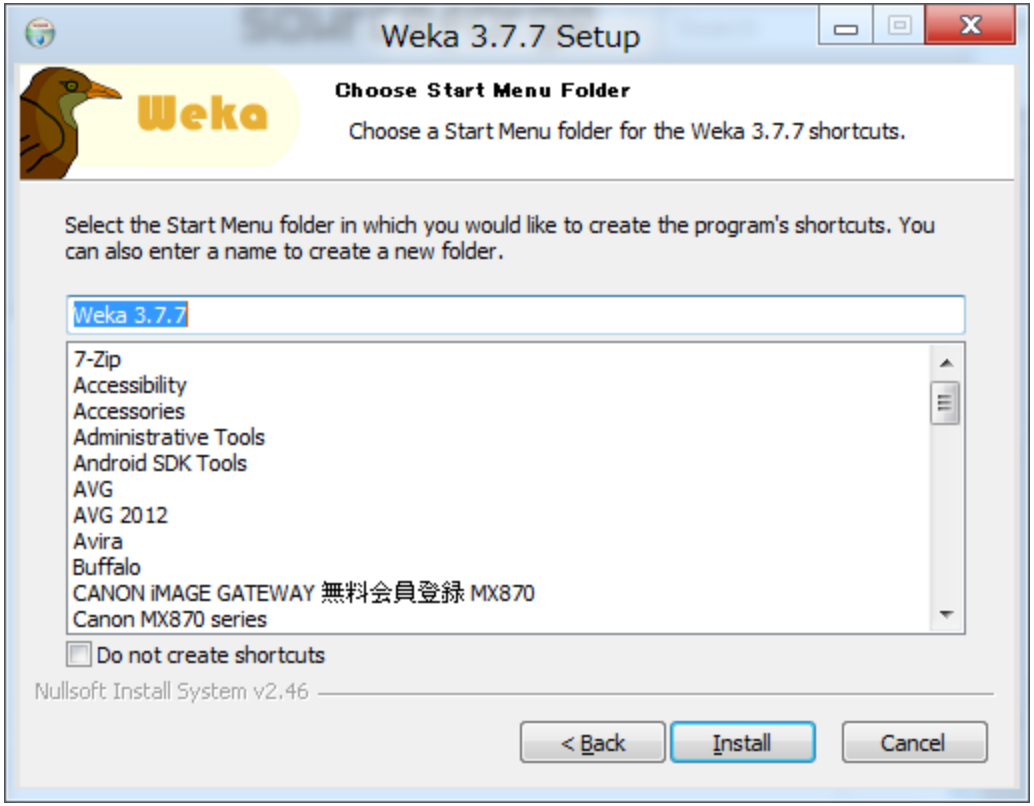


- Set install directory.



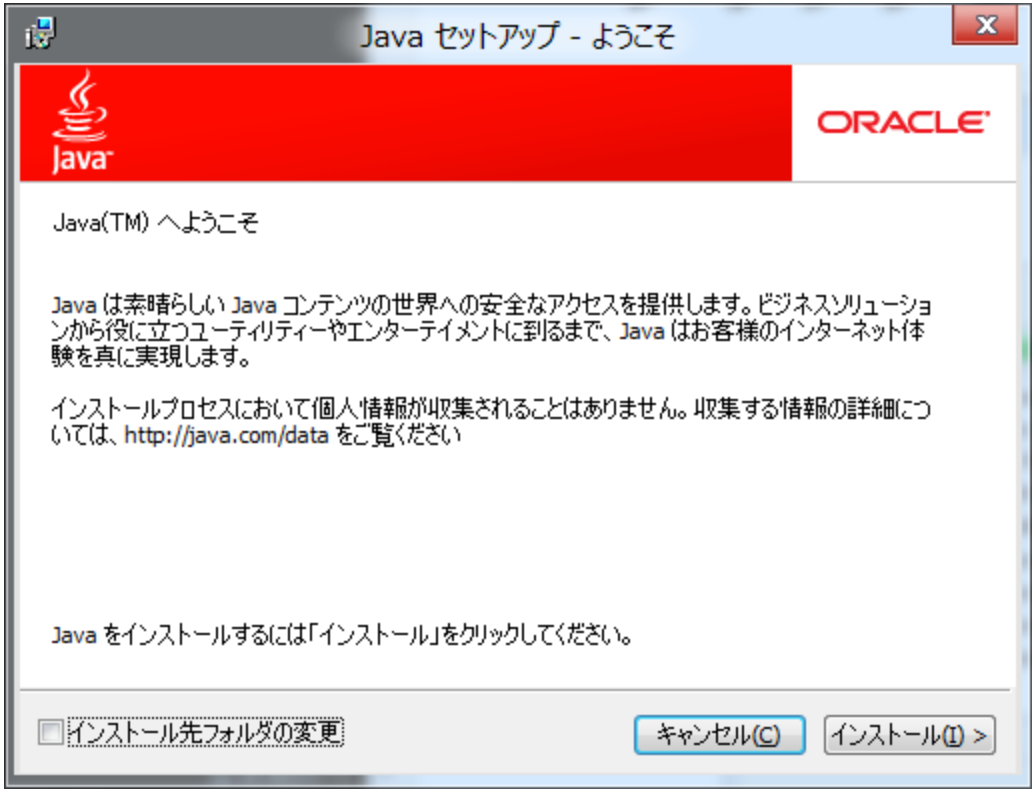
- Set to add start menu or not,

  to create shortcut or not.

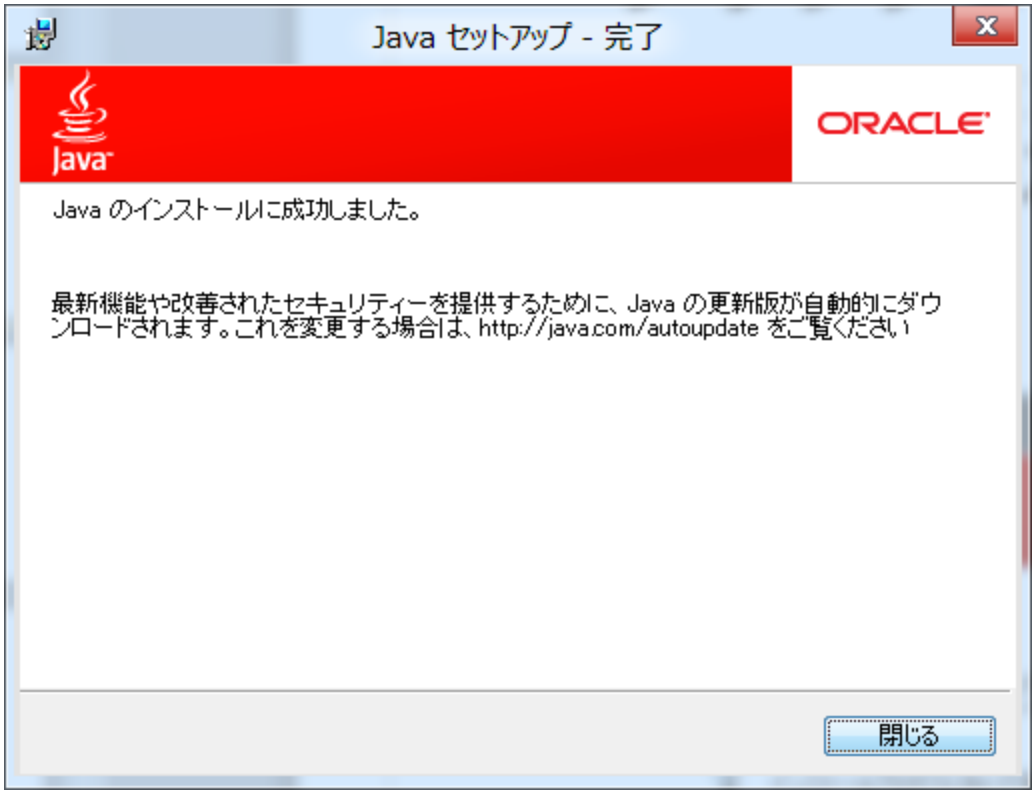


JRE

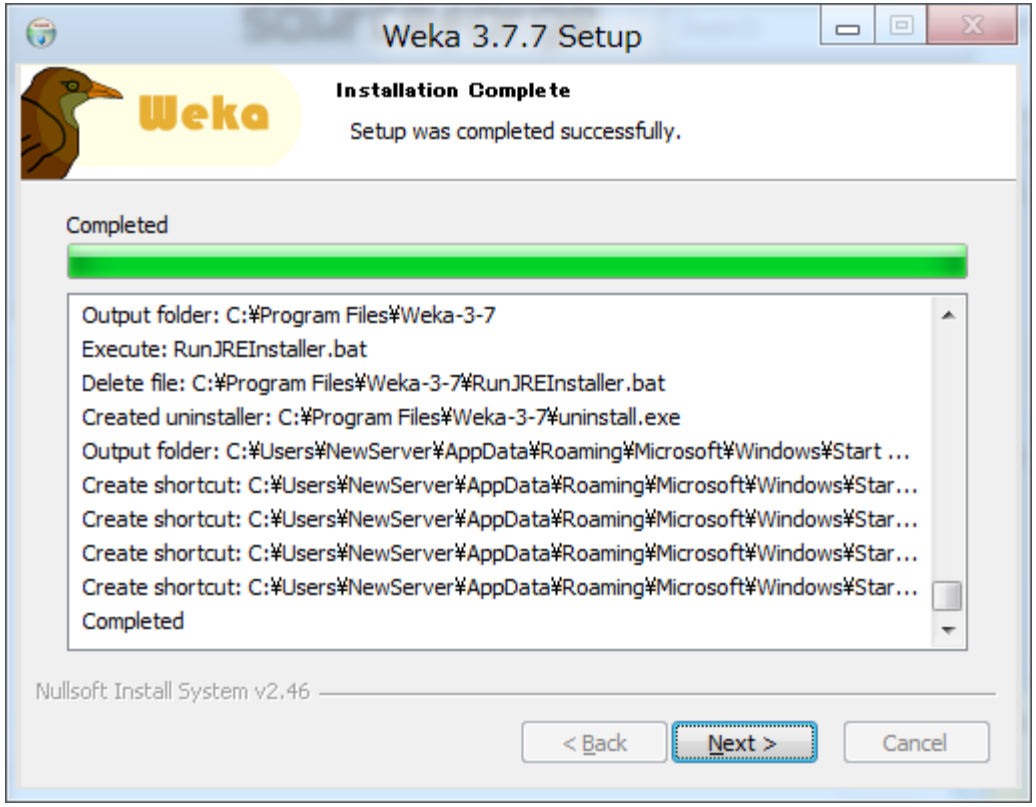
- Before installing Weka, JRE's installation starts.



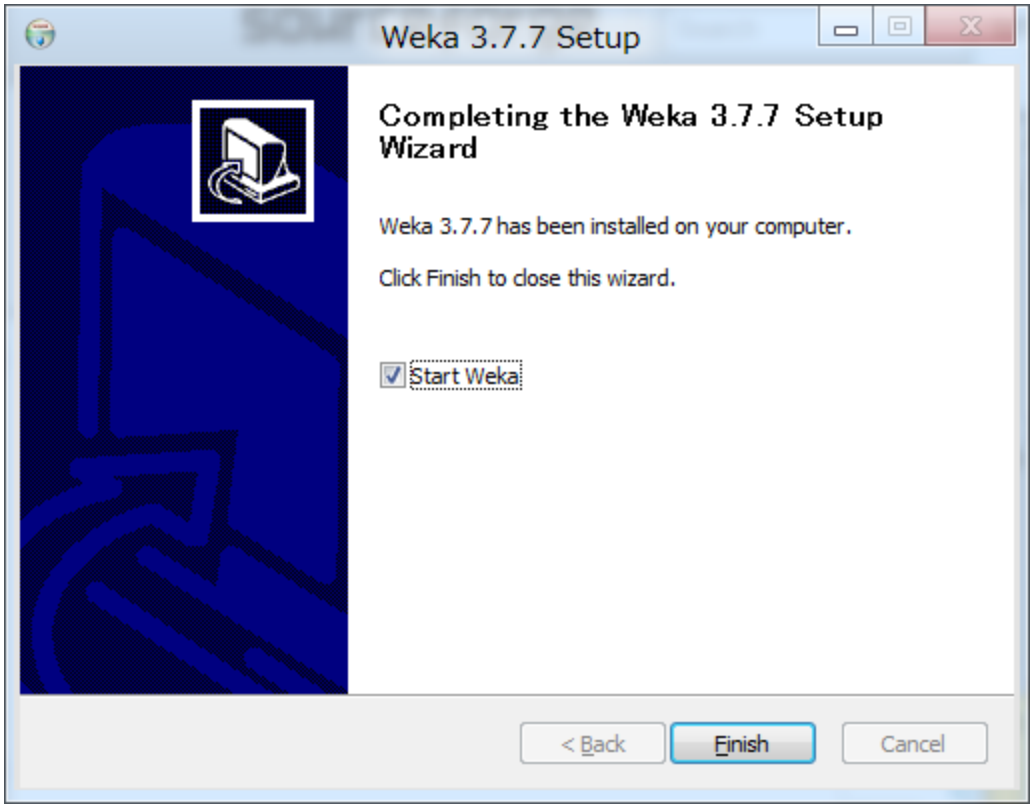
- After JRE’s install, close the window.



- After installing JRE, Weka installation starts.

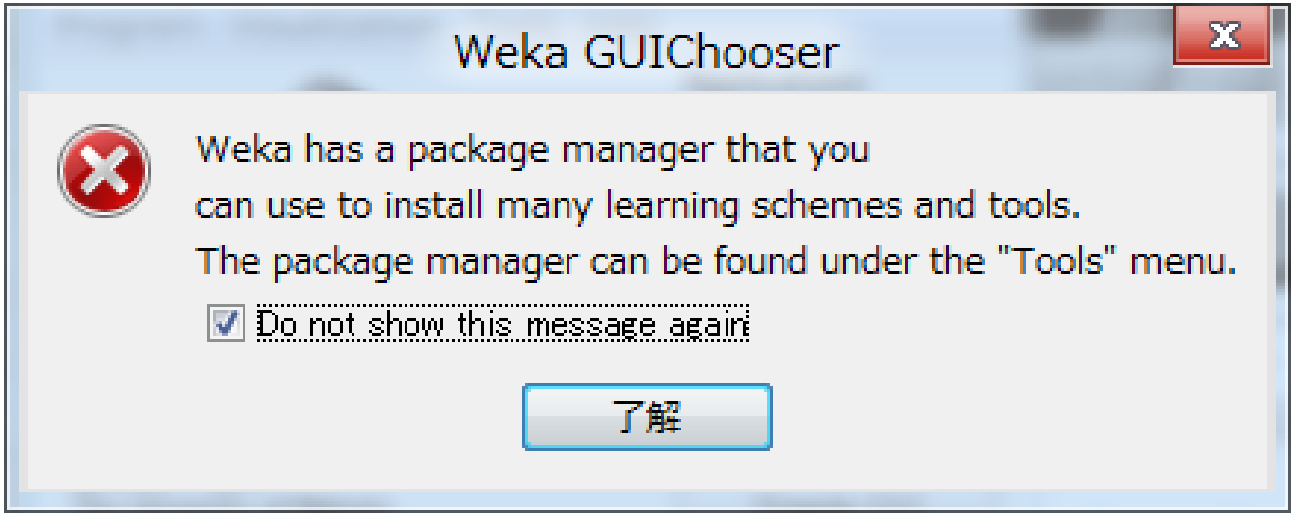


- Click Finish button after checking Start Weka.



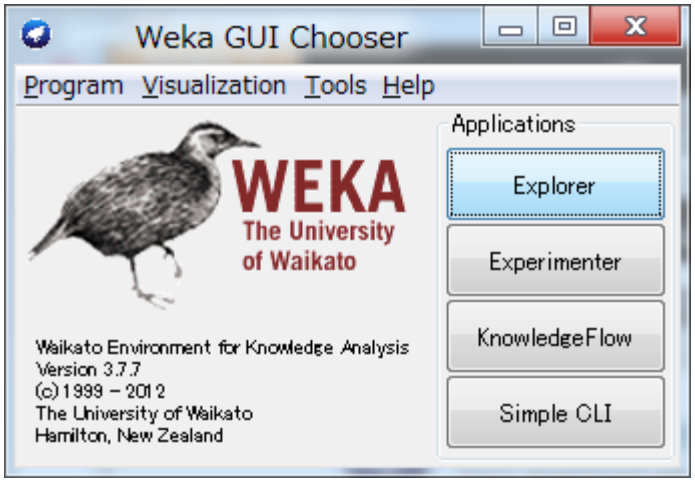
Start Weka

- At the first time, the warning like below is appeared. Please check and click ok.



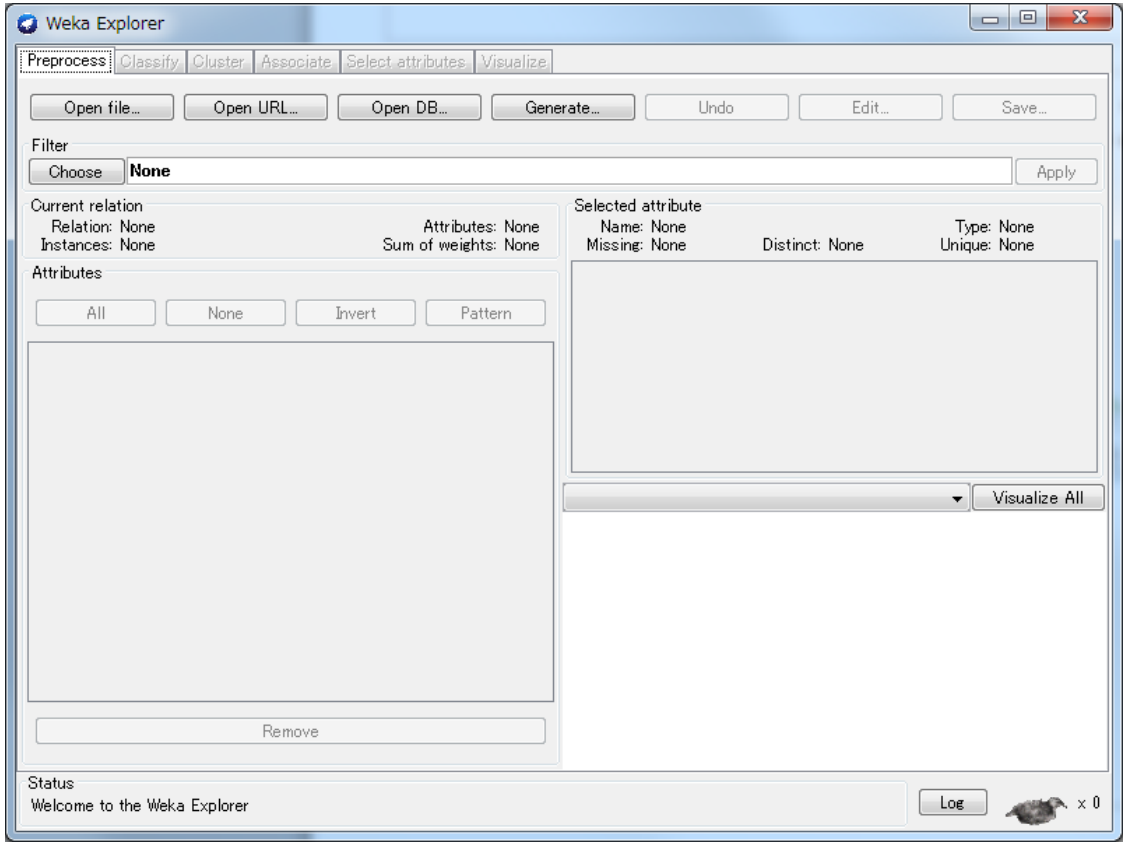
Use Explorer

- This is the Weka GUI.
- To start data mining, click Explorer.



Wekaの実行

- この画面からデータを開いたりデータマイニングをしたりします



データセットの作成

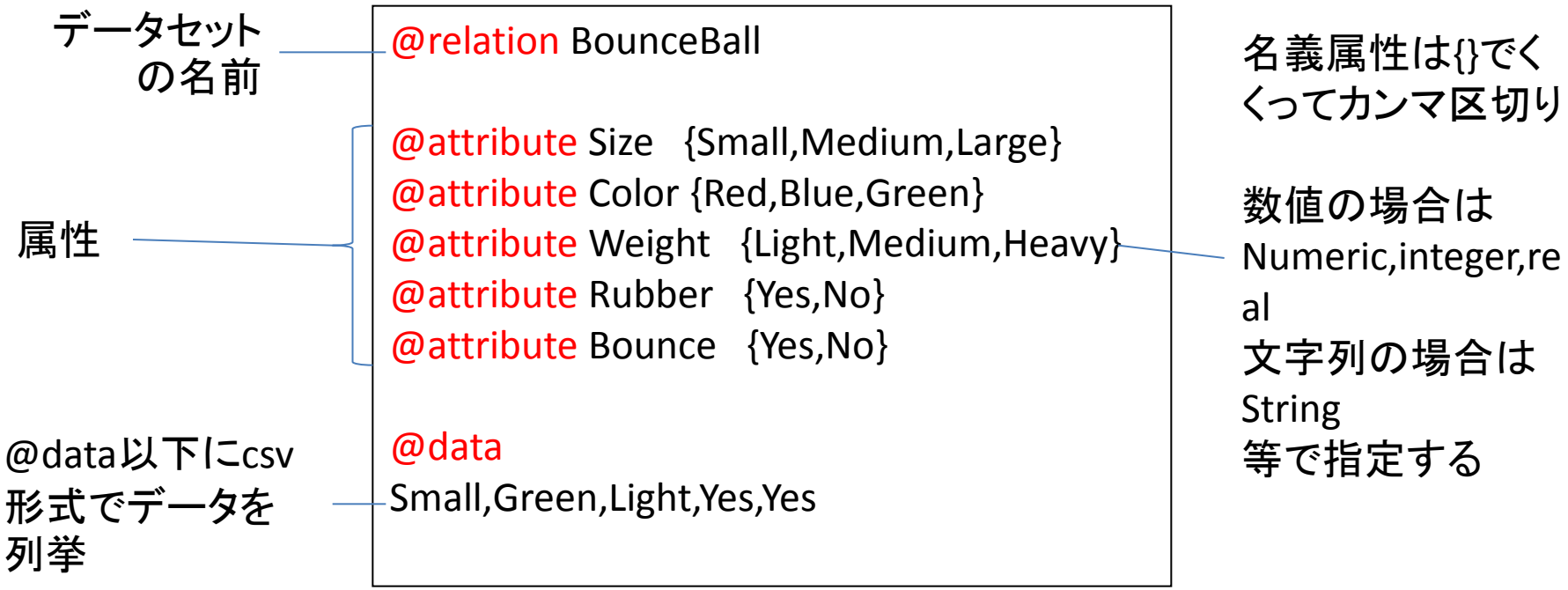
- 例としてbounce ballのデータセットを作成し、データマイニングをしてみましょう

Ball	Size	Color	Weight	Rubber?	Result(Bounce?)
1	Small	Green	Light	Yes	Yes
2	Small	Blue	Medium	No	No
3	Medium	Red	Medium	No	No
4	Small	Red	Medium	Yes	Yes
5	Large	Green	Heavy	Yes	Yes
6	Medium	Blue	Heavy	Yes	No
7	Medium	Green	Heavy	Yes	No
8	Small	Red	Light	No	No

ball.csv → ball.arff　open ball.csv in openFile and save .arff format

データセットの書式

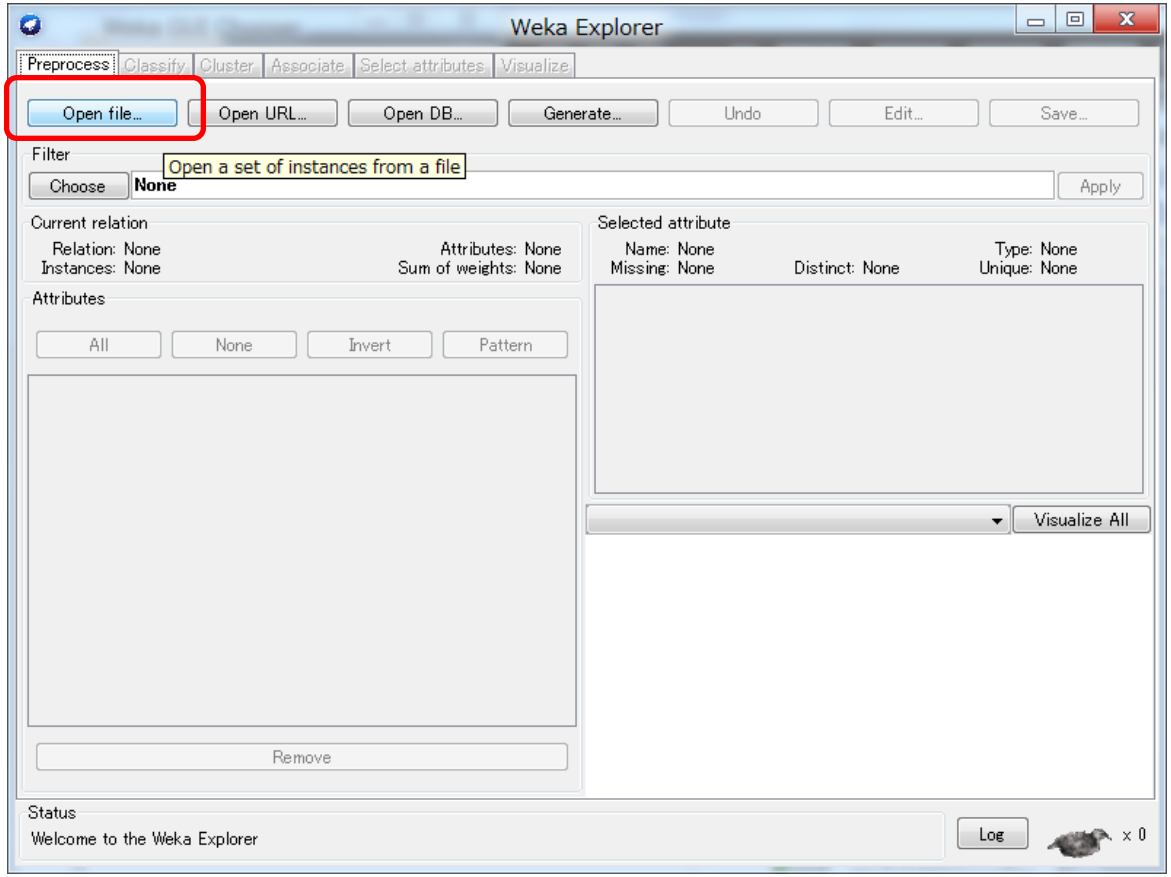
- Wekaではarff形式のファイルが推奨されている



※dataフォルダ内のサンプルも参考にしてもよい
※Wekaでは大文字小文字を区別します
※csv形式のファイルも読み込めますが、推奨はされません

データセットの読み込み

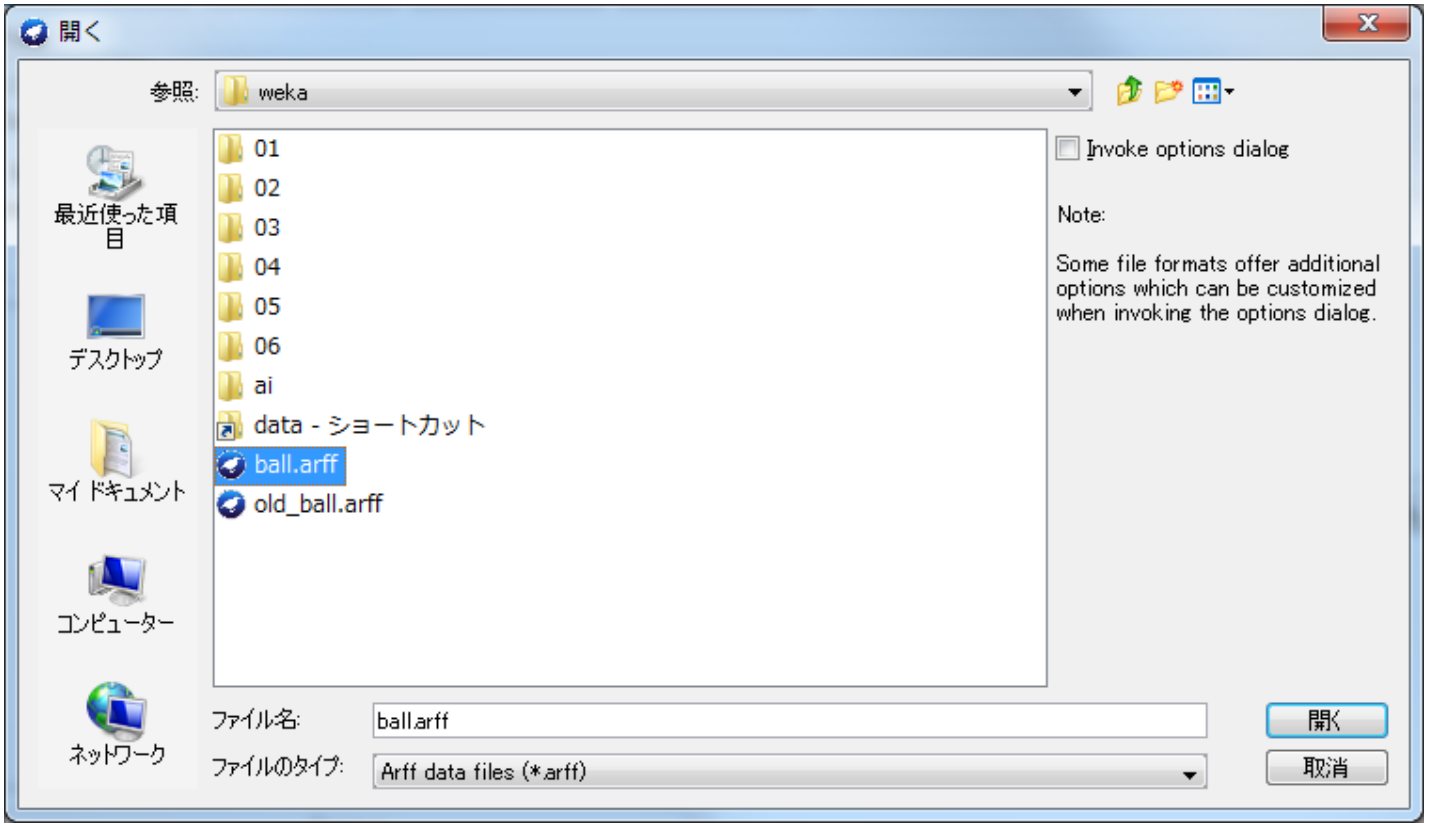
- Preprocessタブを開き、Open fileを押します



データセットの読み込み

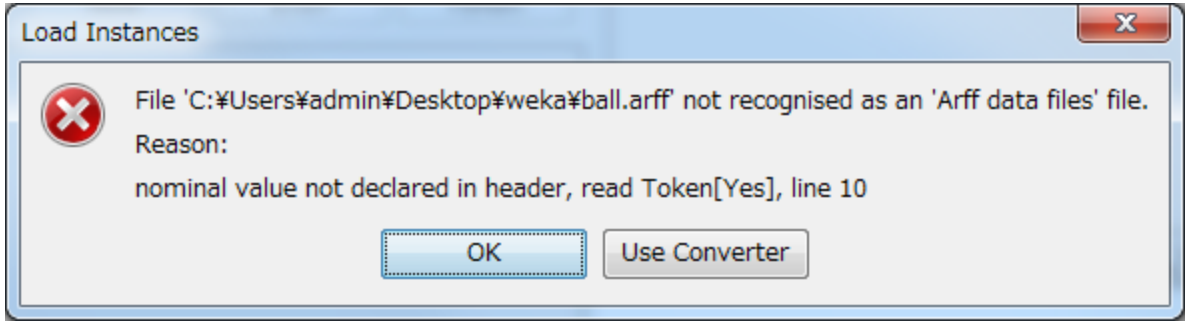
- いま、作成したBounceBallのデータセットを読み込みます

ball.arff



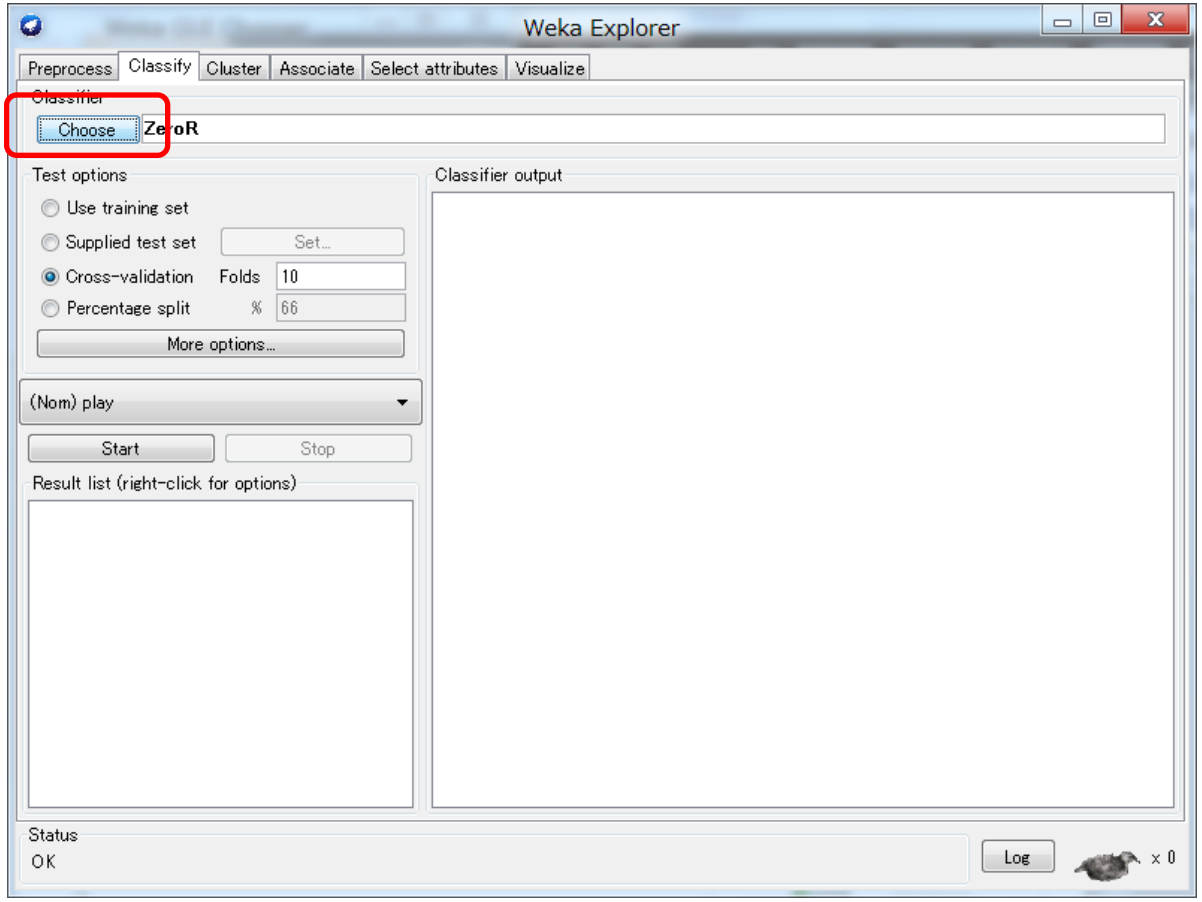
データセットの読み込み

- 書式が間違っている場合は、以下の様なエラーが出るので、間違っている部分を修正しましょう



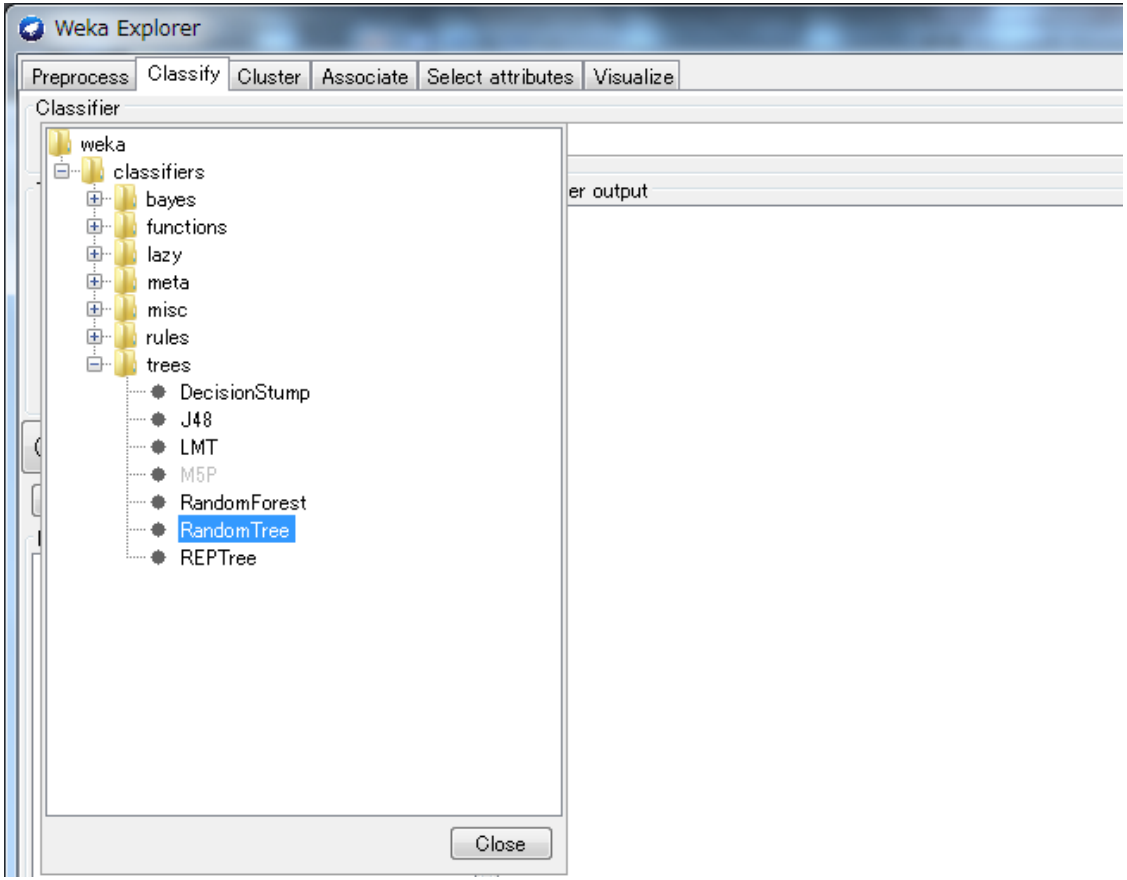
Wekaの実行

- Classifyタブを開き、Chooseを押します



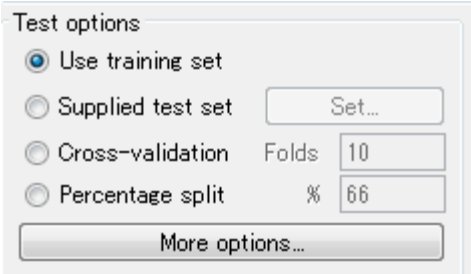
アルゴリズムの指定

- アルゴリズムの一覧が表示されるので、treesフォルダの中にあるRandom Treeを選択します



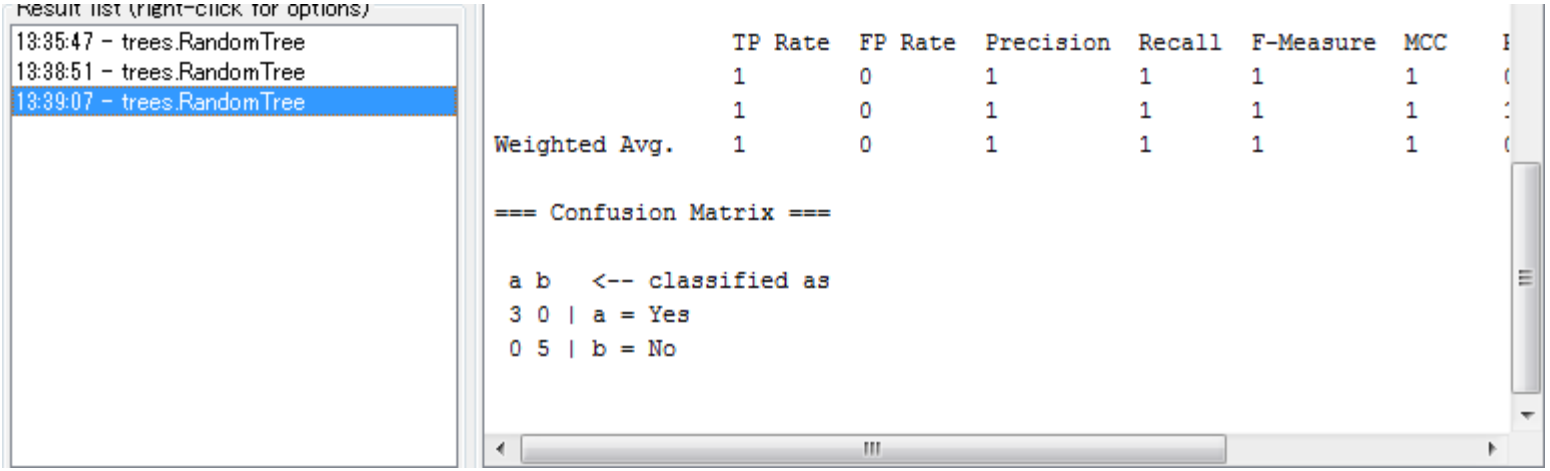
評価方法の指定

- 初期設定ではCross-validation(交差検定)になっていますが、Use training setに変更します(全てを学習データとする)



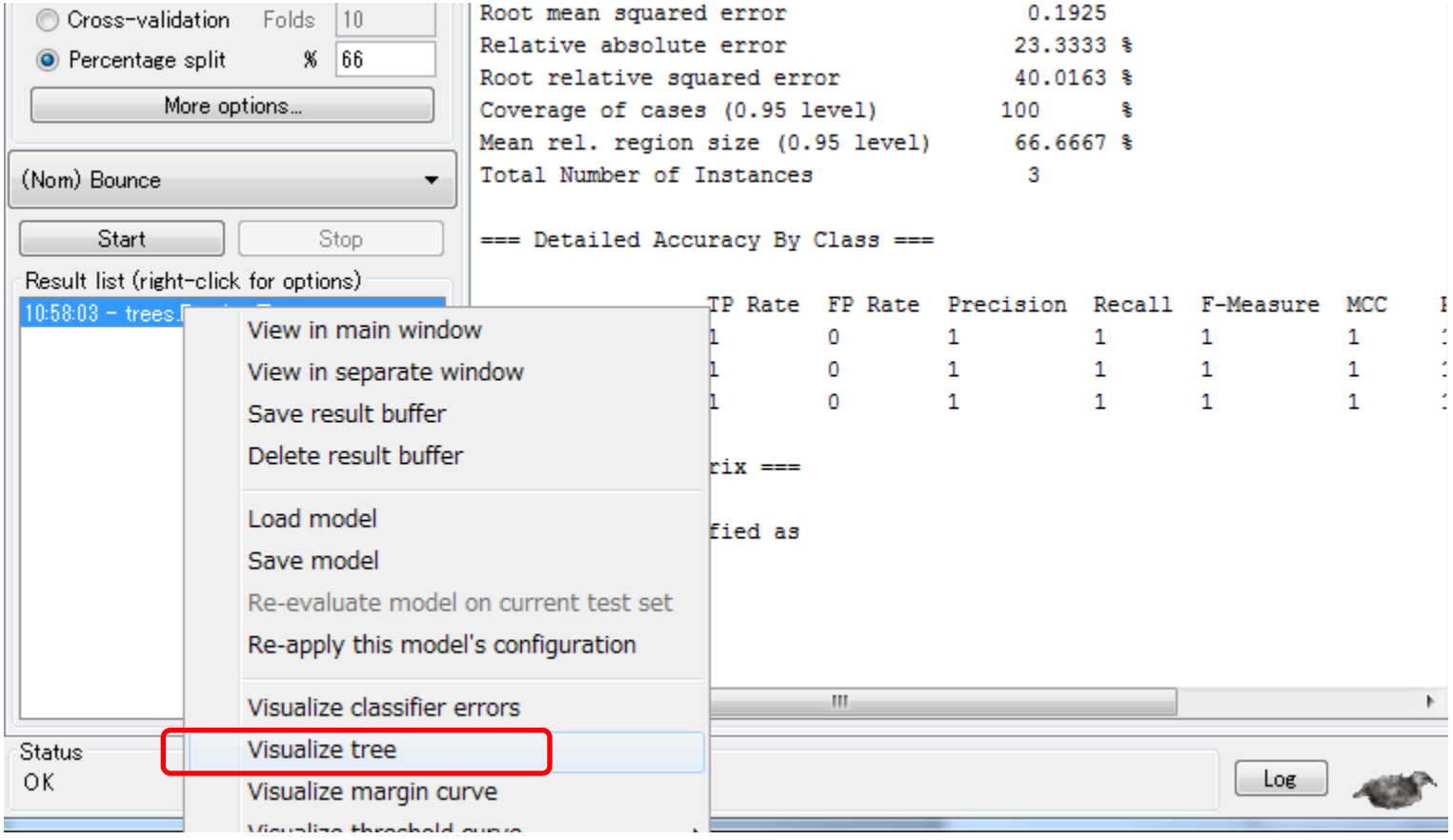
Wekaの実行

- Startを押すと解析を実行します



Wekaの実行

- 結果が右側に表示されたら、結果のリストを右クリックし、Visualize treeを選択します



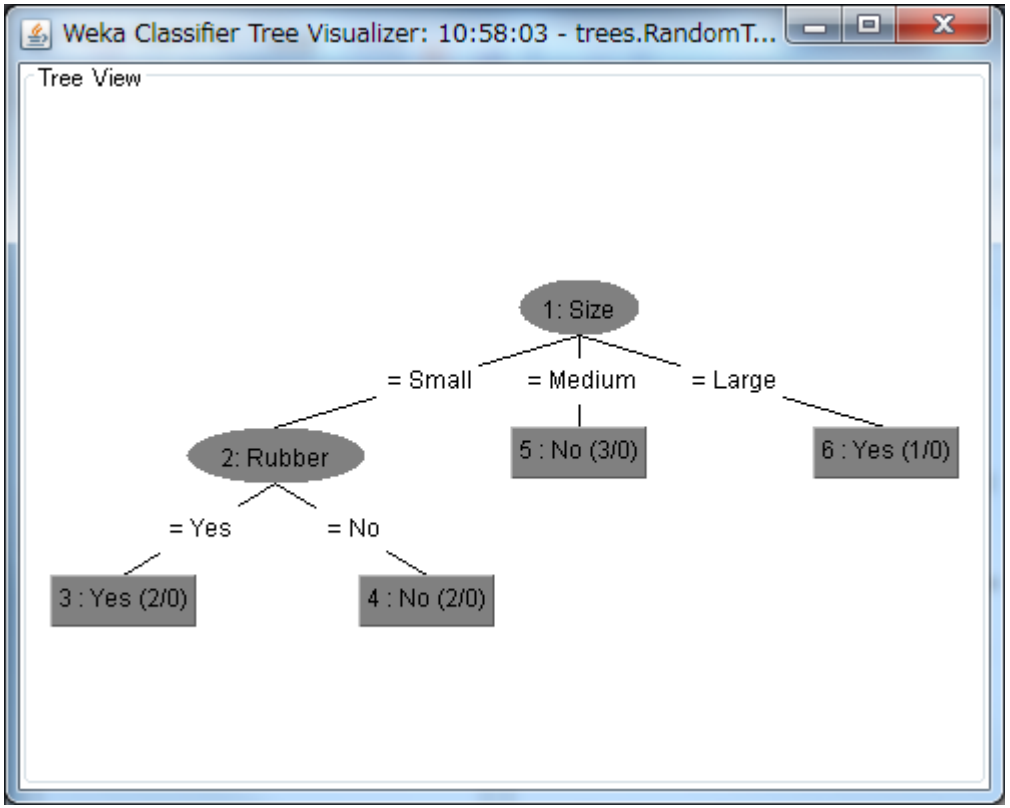
Wekaの実行

- データのbounceにおける決定木が生成されました

```
RandomTree
==========

Size = Small
|   Rubber = Yes : Yes (2/0)
|   Rubber = No : No (2/0)
Size = Medium : No (3/0)
Size = Large : Yes (1/0)

Size of the tree : 6
```



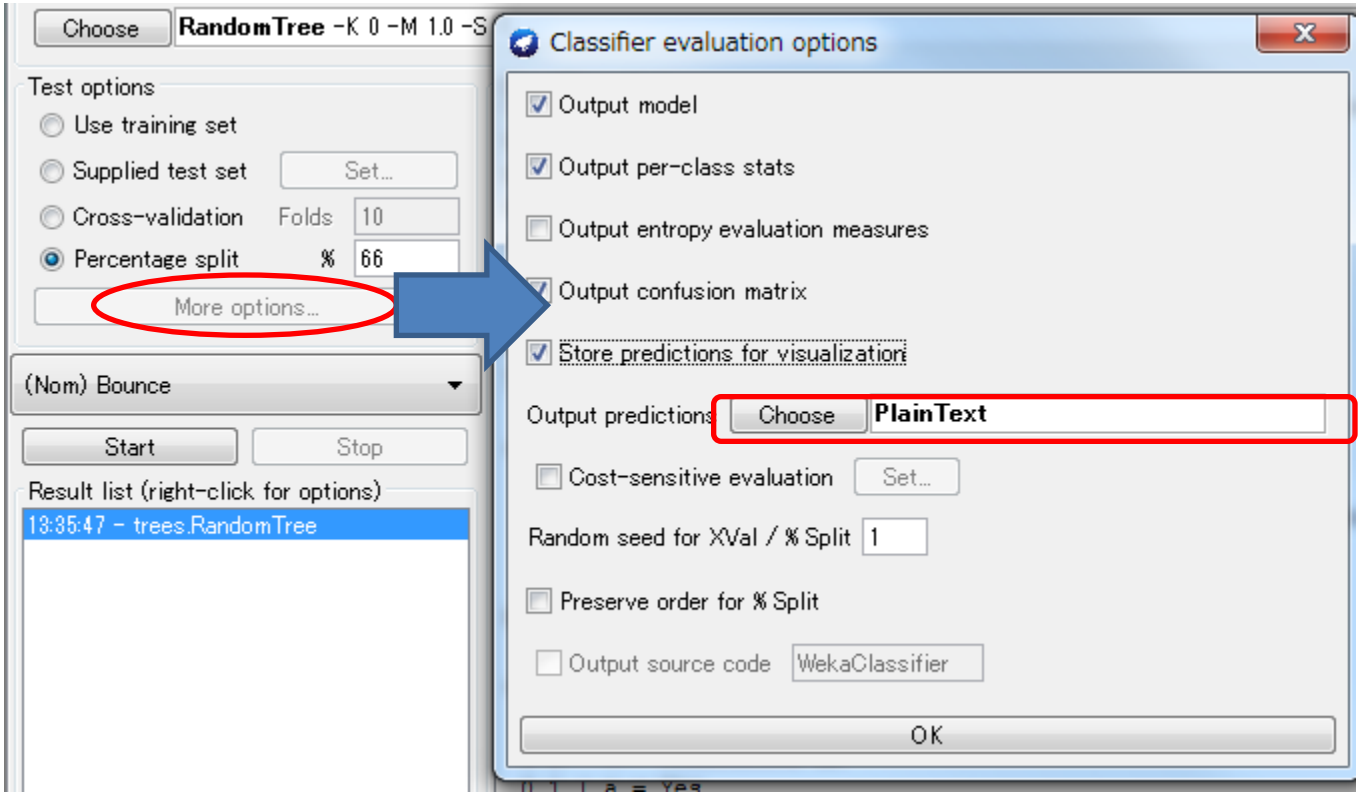
Predict new data

- Add unknown data in the dataset. Unknown data is presented in ?.

```
@data↓
Small,Green,Light,Yes,Yes↓
Small,Blue,Medium,No,No↓
Medium,Red,Medium,No,No↓
Small,Red,Medium,Yes,Yes↓
Large,Green,Heavy,Yes,Yes↓
Medium,Blue,Heavy,Yes,No↓
Medium,Green,Heavy,Yes,No↓
Small,Red,Light,No,No↓
Small,Blue,Light,Yes,?←
```

Add setting

- Click More Options and change Output predictions to PlainText from null.



Result of prediction

- Start mining, the unknown data is predicted “Yes”.

```
=== Predictions on training set ===

 inst#     actual  predicted error prediction
     1      1:Yes      1:Yes       1
     2       2:No       2:No       1
     3       2:No       2:No       1
     4      1:Yes      1:Yes       1
     5      1:Yes      1:Yes       1
     6       2:No       2:No       1
     7       2:No       2:No       1
     8       2:No       2:No       1
     9        1:?      1:Yes       1
```

Ball	Size	Color	Weight	Rubber?	Result (Bounces?)
1	Small	green	Light	yes	yes
2	Small	blue	Medium	no	no
3	Medium	red	Medium	no	no
4	Small	red	Medium	yes	yes
5	Large	green	Heavy	yes	yes
6	Medium	blue	Heavy	yes	no
7	Medium	green	Heavy	yes	no
8	Small	red	Light	no	no

Figure C1: Identification Tree Training Data

Work in class:

Implement and generating ID tree on the
bouncing ball dataset, ball.arff
using Weka.
(correctly generate ball.arff file and then use Weka)

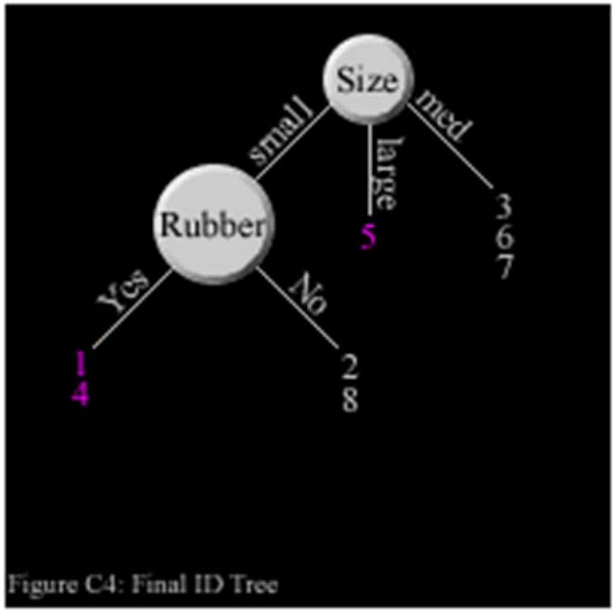
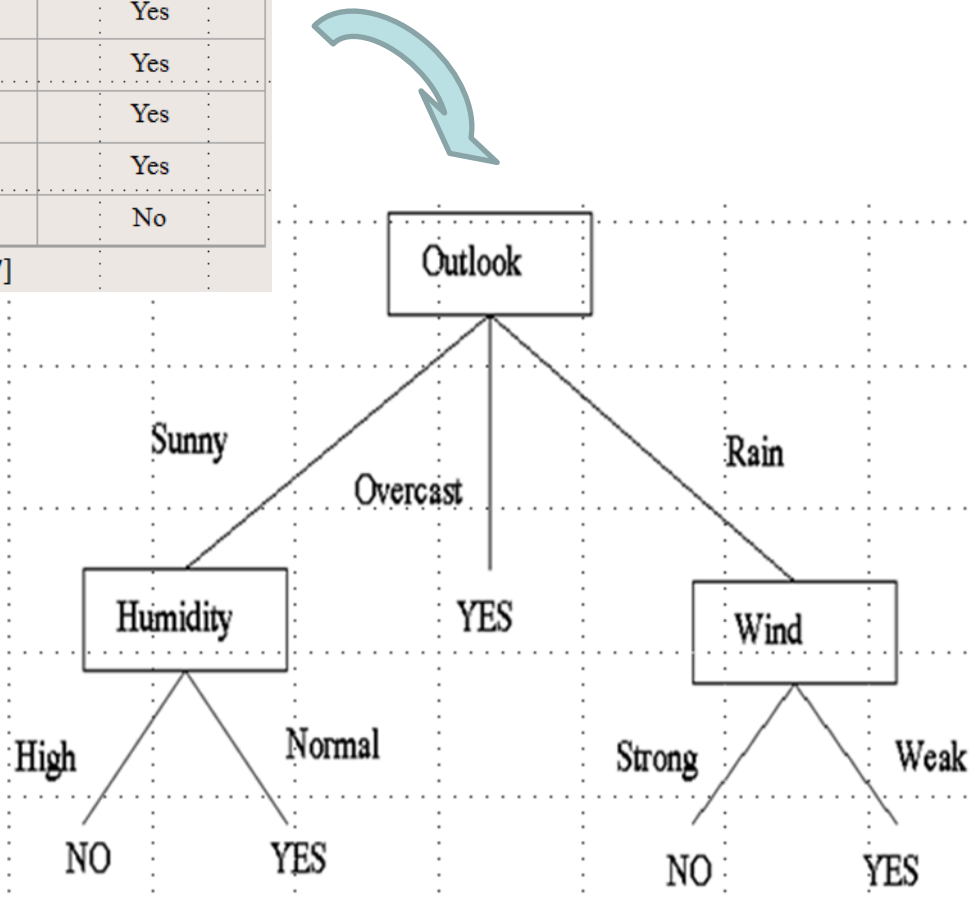


Figure C4: Final ID Tree

An example:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

[See: Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997]



Home Work

1. Please read the article decision-tree-article
   (no submission)