Introduction to WEKA

Preprocess and Classify

PL03

[Hugo Peixoto] – 2018/2019, Universidade do Minho

Material

http://hpeixoto.github.io/dc

WEKA

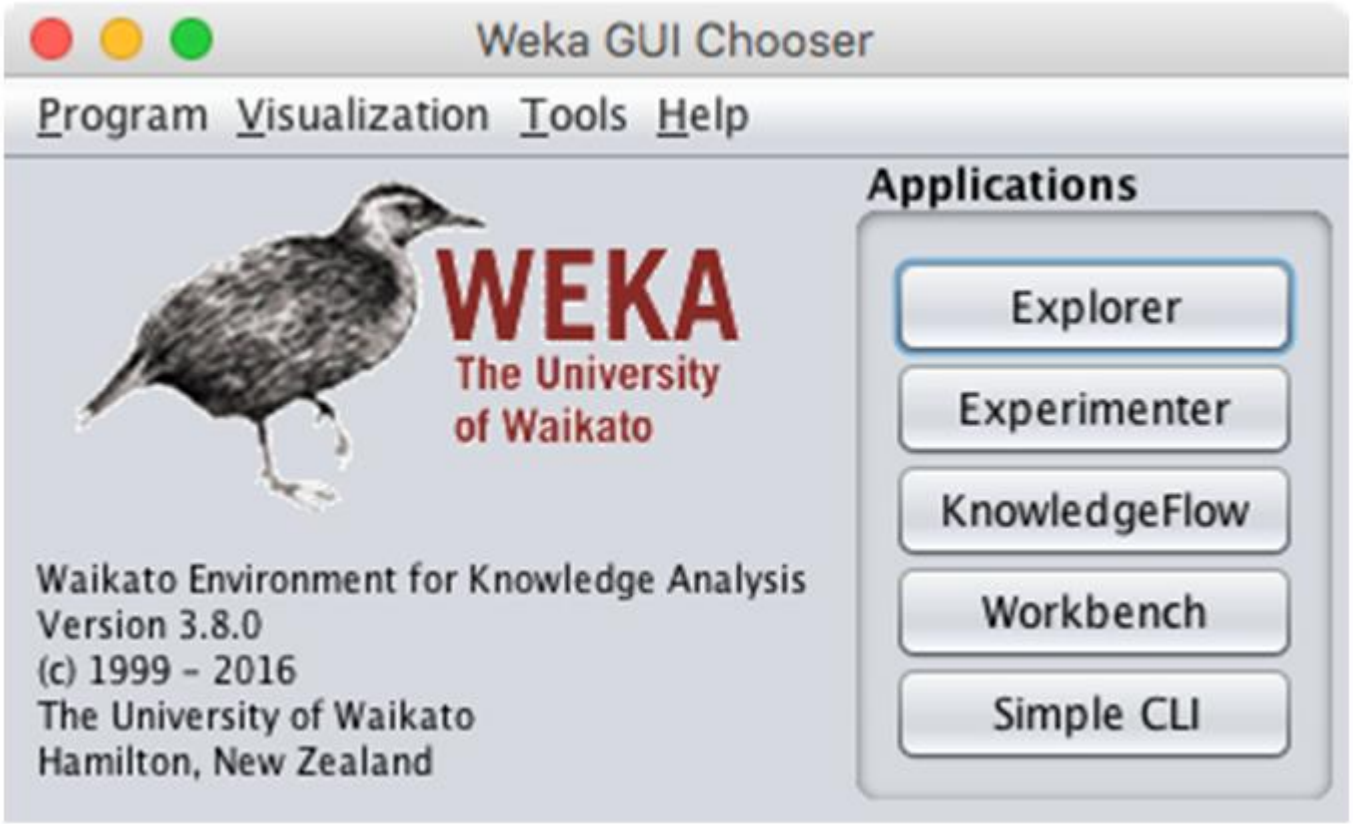
Weka is a collection of machine learning algorithms for data mining tasks.

WEKA

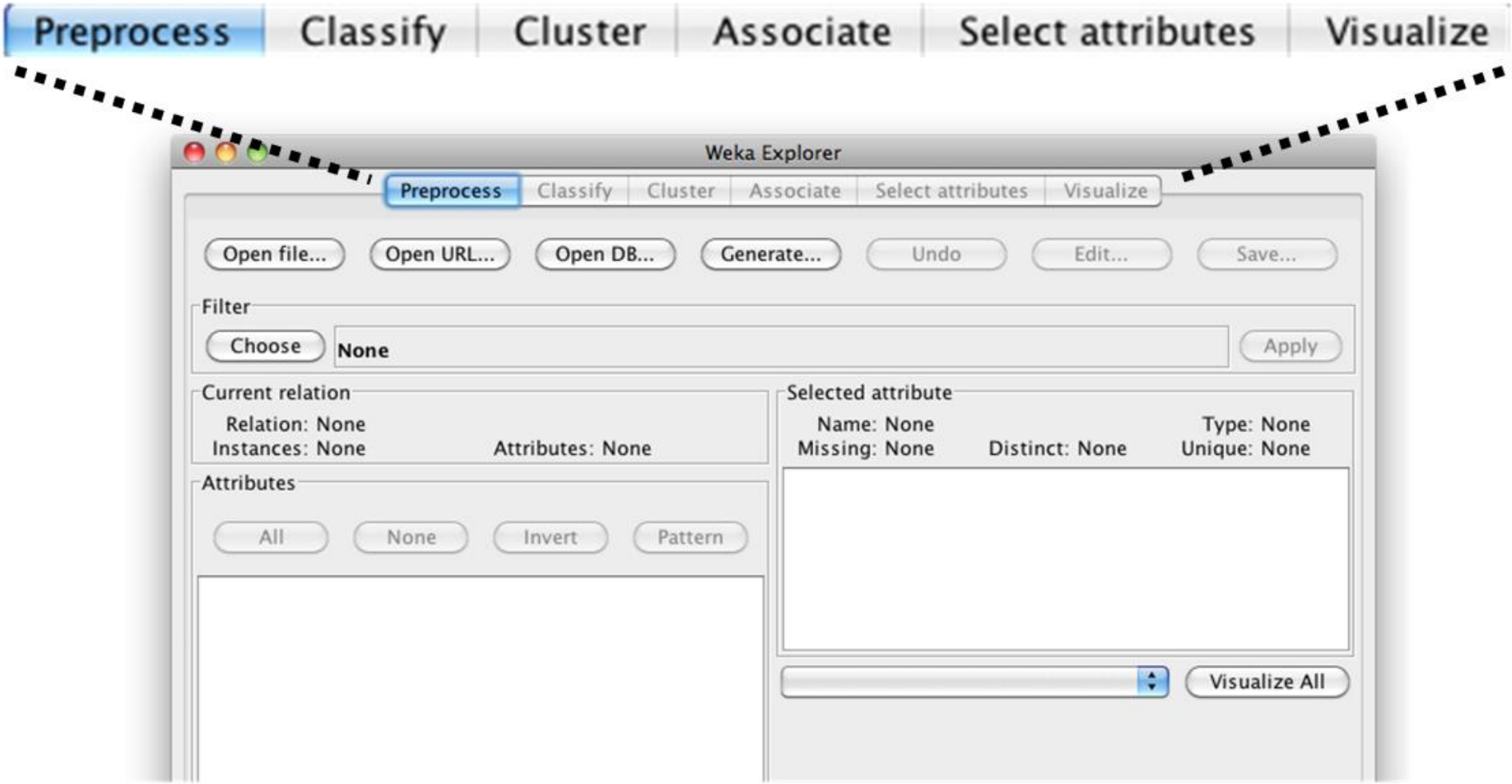
WEKA download

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

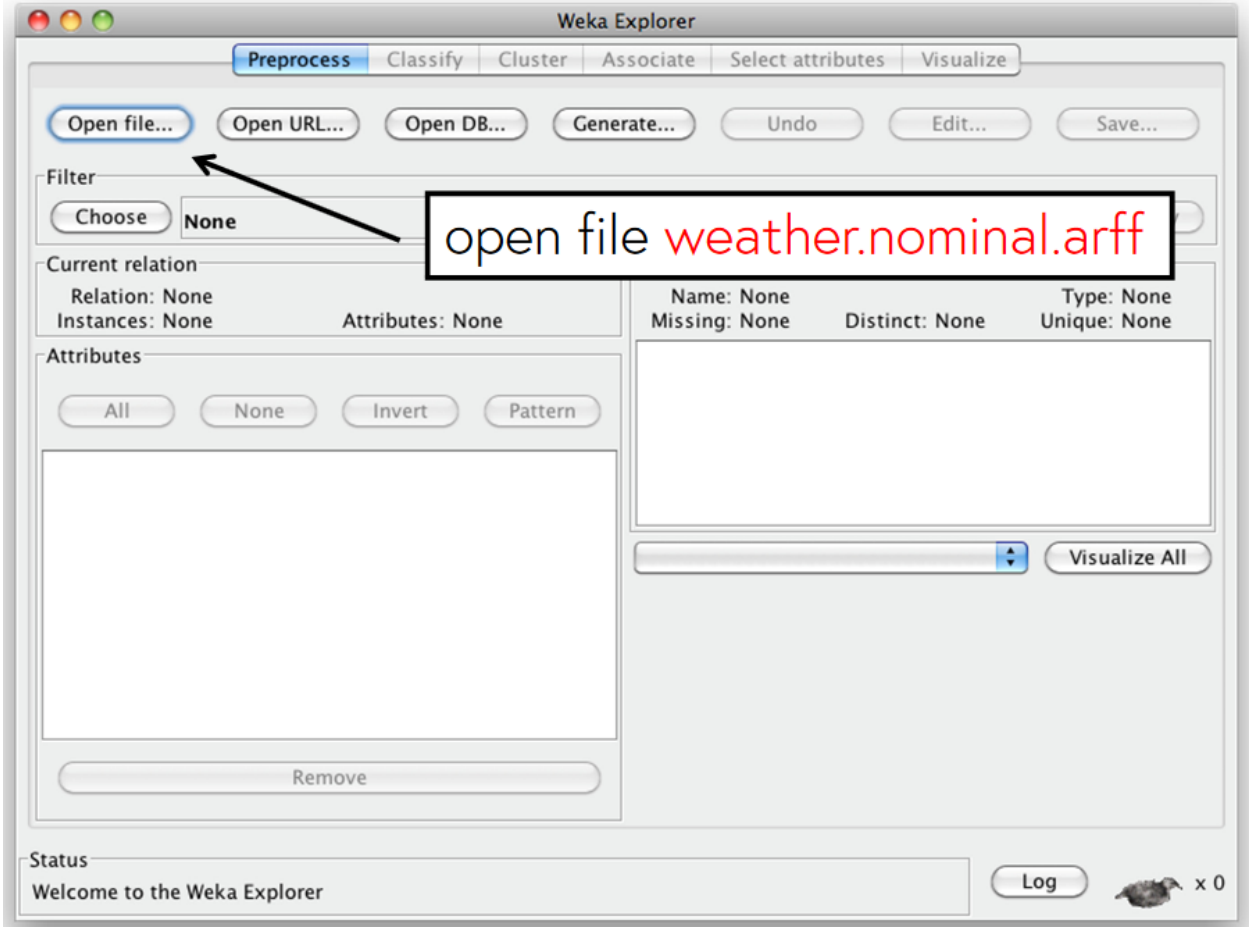
WEKA


Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.8.0
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI

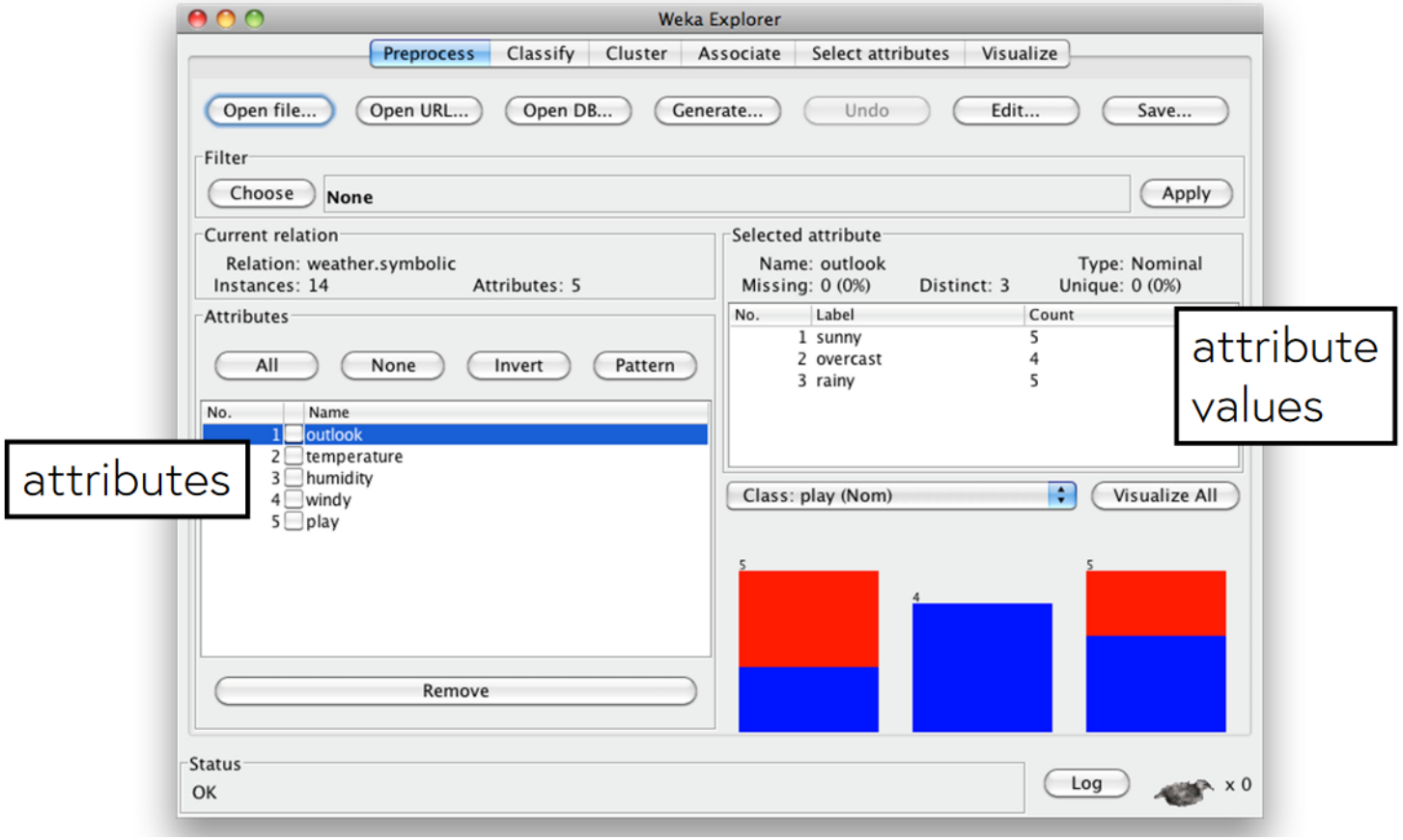
WEKA


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All

WEKA

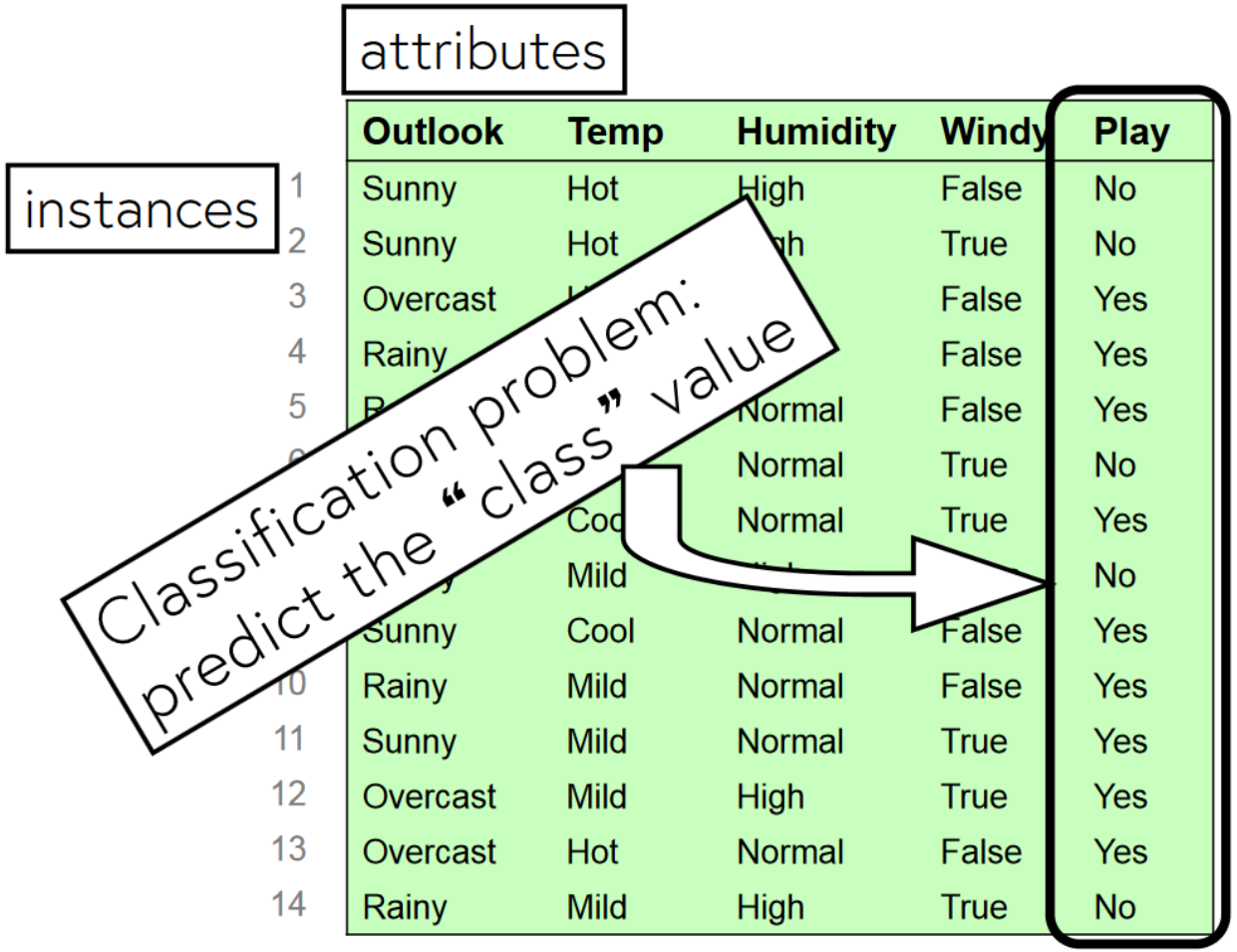

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
open file weather.nominal.arff
Current relation
Relation: None
Instances: None
Attributes: None
Name: None
Missing: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Visualize All
Remove
Status
Welcome to the Weka Explorer
Log
x 0

WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: outlook
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No.
Label
Count
1 sunny 5
2 overcast 4
3 rainy 5
Class: play (Nom)
Visualize All
5
4
5
Status
OK
Log
x 0
attributes
attribute values

WEKA

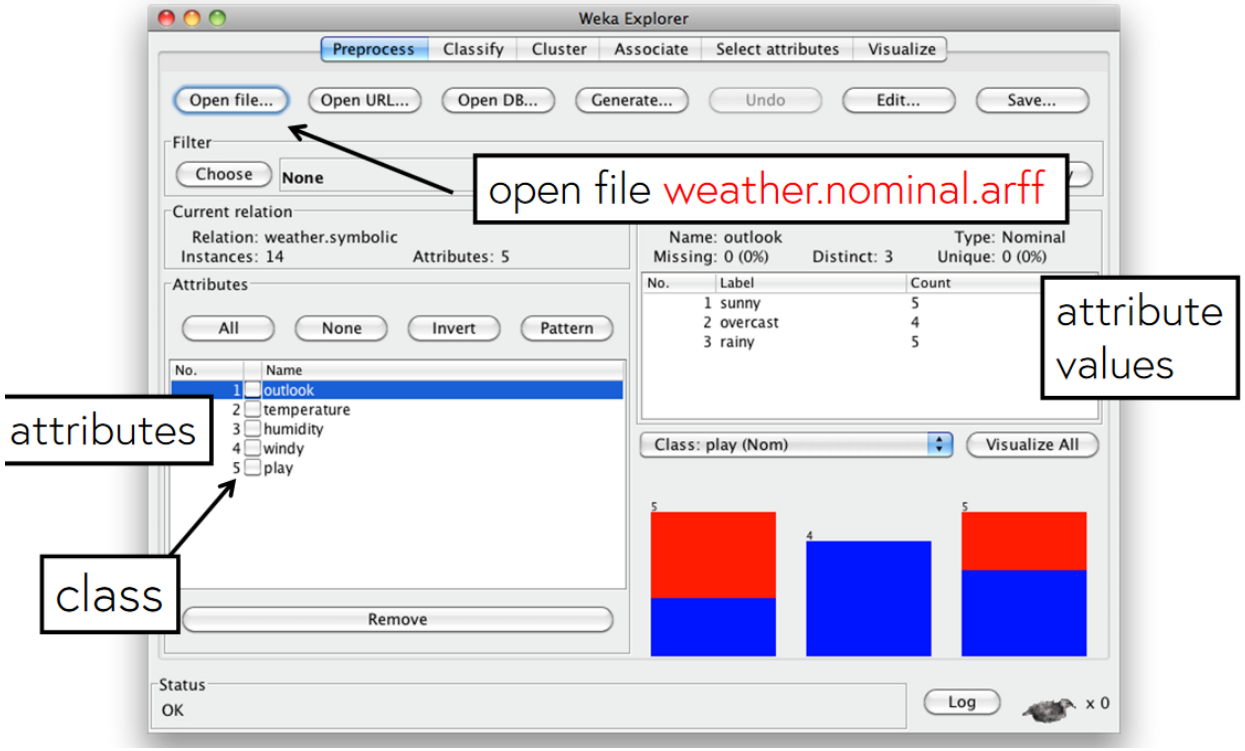
Weather.arff



	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	[illegible]	True	No
3	Overcast	[illegible]	[illegible]	False	Yes
4	Rainy	[illegible]	[illegible]	False	Yes
5	[illegible]	[illegible]	Normal	False	Yes
[illegible]	[illegible]	[illegible]	Normal	True	No
[illegible]	[illegible]	Cool	Normal	True	Yes
[illegible]	[illegible]	Mild	[illegible]	[illegible]	No
[illegible]	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

WEKA

Weather.arff

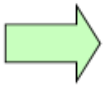


WEKA

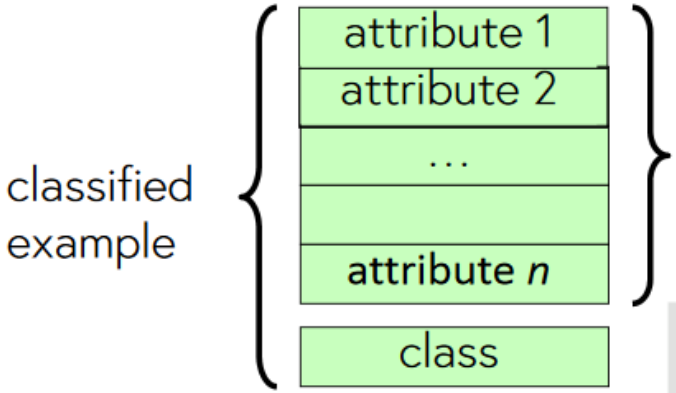
Classification

sometimes called "supervised learning"

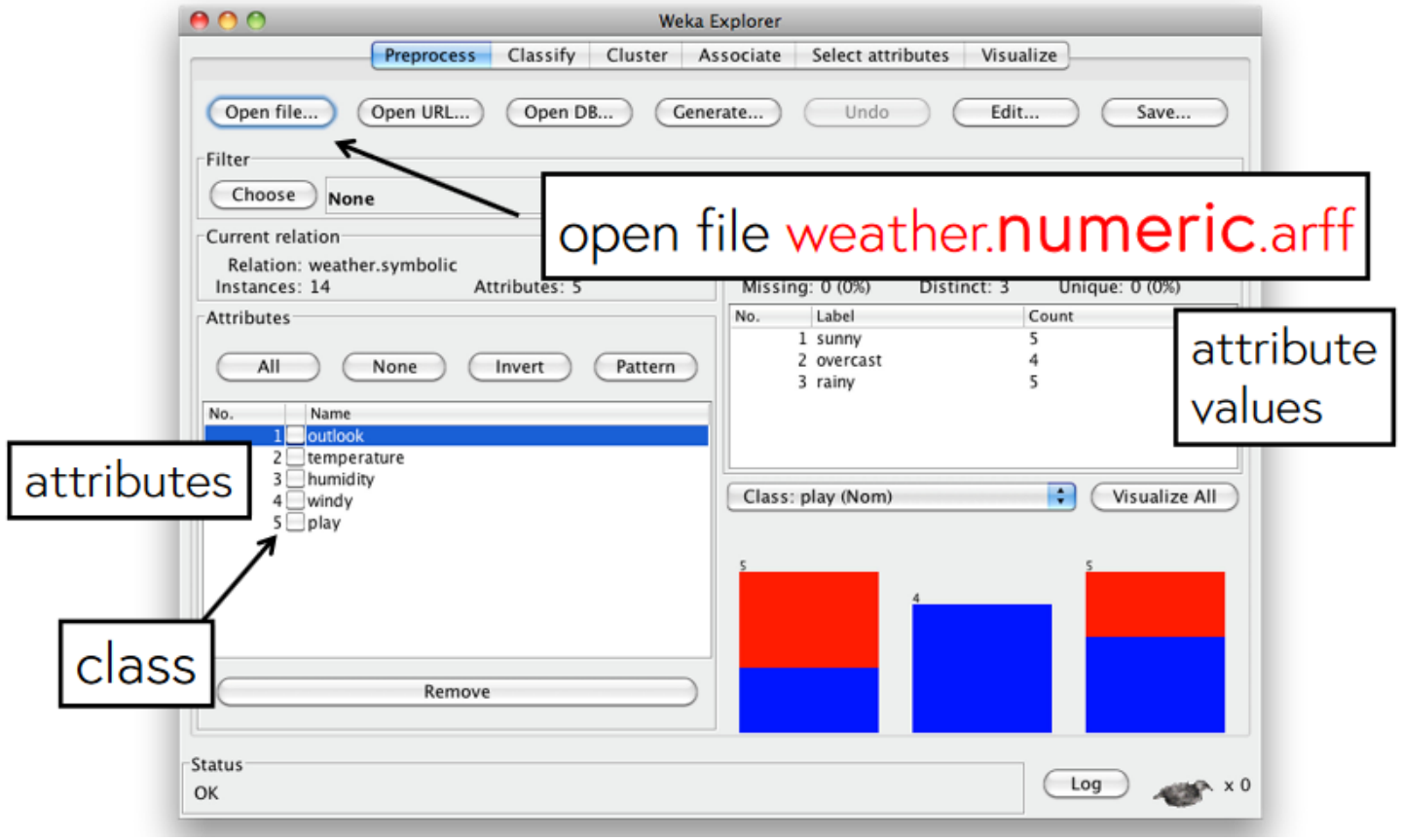
Dataset: classified examples



"Model" that classifies new examples



WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No.
Label
Count
1 sunny 5
2 overcast 4
3 rainy 5
Class: play (Nom)
Visualize All
Status
OK
Log
x 0
open file weather.numeric.arff
attribute values
attributes
class

BUILD CLASSIFIER

USE J48 TO ANALYZE THE GLASS DATASET

- Open file glass.arff
- Check the available classifiers
- Choose the J48 decision tree learner (trees>J48)
- Run it
- Examine the output
- Look at the correctly classified instances
  - ... and the confusion matrix

BUILD CLASSIFIER

INVESTIGATE J48

- Open the configuration panel
- Check the More information
- Examine the options
- Use an unpruned tree
- Look at leaf sizes
- Set minNumObj to 15 to avoid small leaves
- Visualize tree using right-click menu

BUILD CLASSIFIER

Pruning (decision trees)

is a technique in machine learning that reduces the size of **decision trees** by removing sections of the **tree** that provide little power to classify instances. **Pruning** reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

EXERCÍCIO GRUPO

DETERMINAR NECESSIDADE DE COMPONENTES SANGUÍNEOS

No bloco operatório de uma unidade de saúde é de vital importância determinar com a devida antecedência a potencial necessidade de um paciente, que vai ser intervencionado, vir a receber uma ou mais transfusões de componentes sanguíneos (sangue, plasma, etc).

Esta necessidade advém de diversos fatores, tais como:

- Custo unitário de cada fornecimento de componentes sanguíneos;
- Escassez de oferta de componentes sanguíneos;
- Correto tratamento do paciente em caso de necessidade de transfusão;
- Diminuir os desperdícios com o deteriorar de componentes sanguíneos afetos a cirugias que não são usados;
- entre outros…

EXERCÍCIO

DETERMINAR NECESSIDADE DE COMPONENTES SANGUÍNEOS

Num exercício de brainstorm refletir sobre os principais dados que deveriam estar acessíveis.

nota: Não existe nenhuma limitação técnica nem legal para a recolha de qualquer input. Pelo que todos os dados estarão acessíveis caso sejam considerados.

EXERCISES

FE02

Introduction to WEKA

Preprocess and Classify

PL03

[Hugo Peixoto] – 2018/2019, Universidade do Minho