Data Mining with Weka

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Agenda

- **What is Weka**
- **Getting started with Weka**
  - Install Weka
  - Explore the "Explorer" interface and datasets
  - Use filters
  - Visualize the data set
- **Data Mining Task—Classification**
  - Algorithms/Functions
  - Evaluation
- **The Data Mining Process**

What is Weka?

- **What is Weka**
  - --A bird found only in New Zealand

- **Data mining workbench**
  - Waikato Environment for Knowledge Analysis

- **Machine learning algorithms for data mining tasks**
  - 100+ algorithms for classification
  - 75 for data preprocessing
  - 25 to assist with feature selection
  - 20 for clustering, finding association rules, etc

Getting Started with Weka

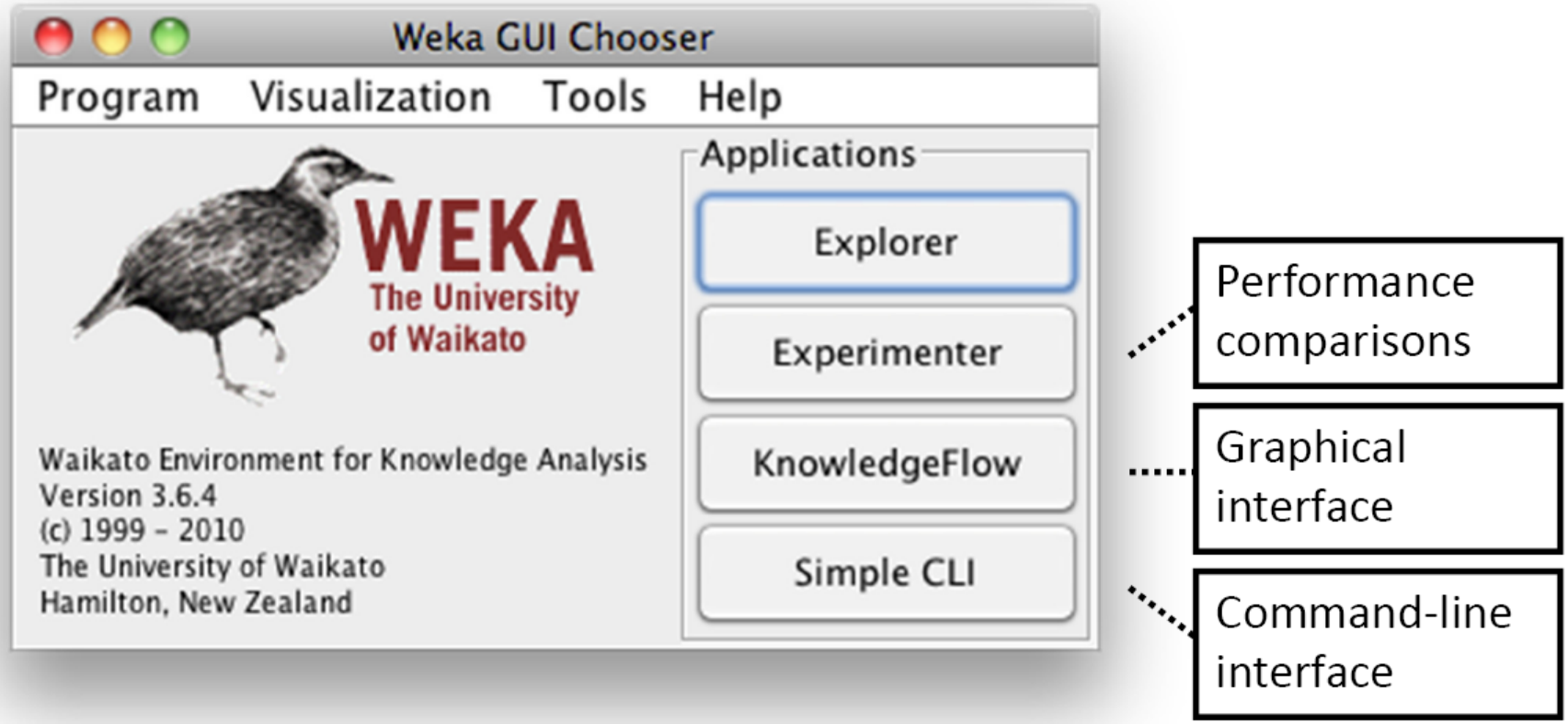
Getting Started with Weka

- **Download from**
  - http://www.cs.waikato.ac.nz/ml/weka
  - (for Windows, Mac, Linux)

- **Weka 3.8**
  - (the latest stable version of Weka)
  - (includes datasets for the course)
  - (it's important to get the right version)

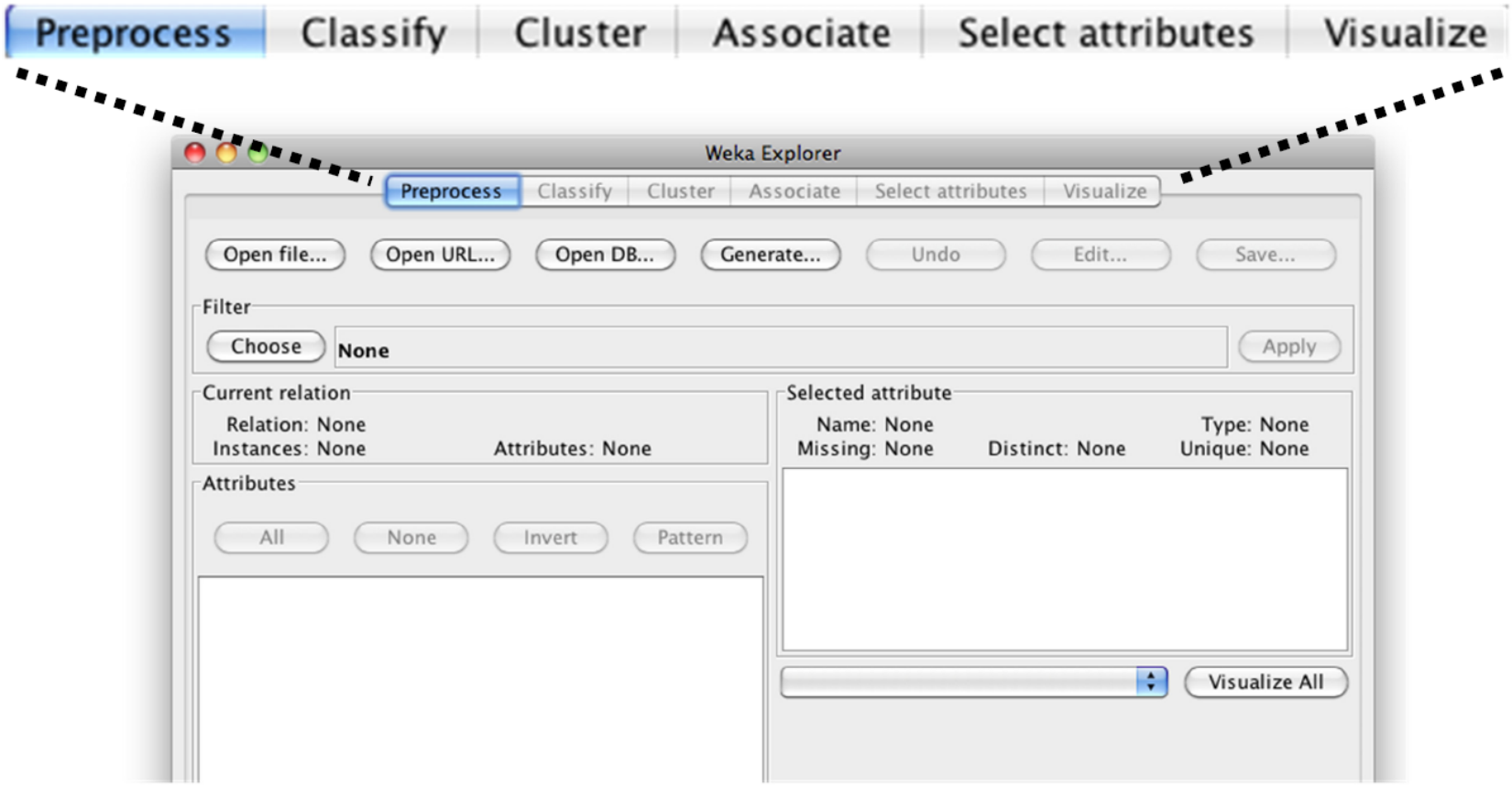
Getting Started with Weka

- **The Interface**



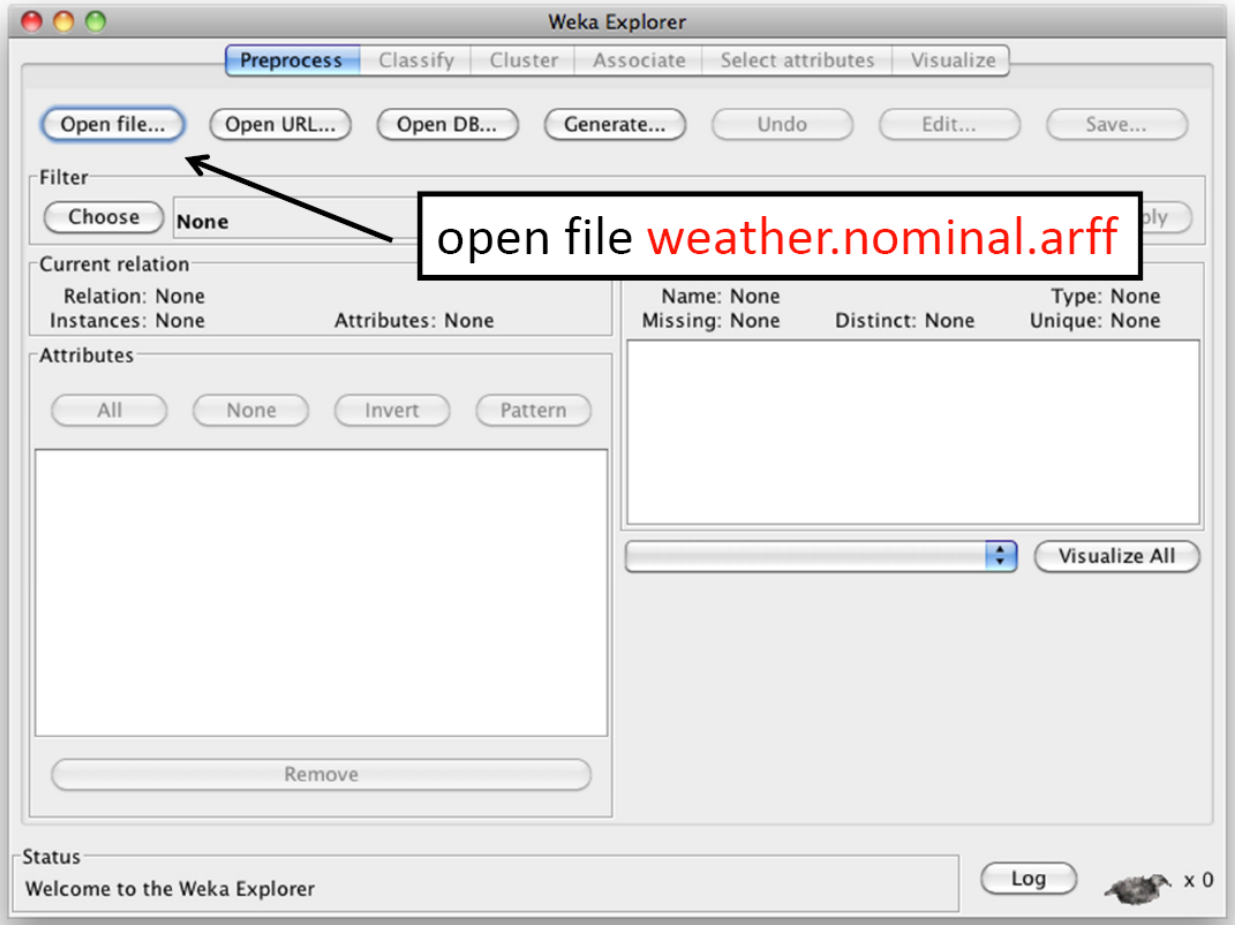
Getting Started with Weka

- **Exploring the Explorer**



Getting Started with Weka

- **Explore the Data Set**



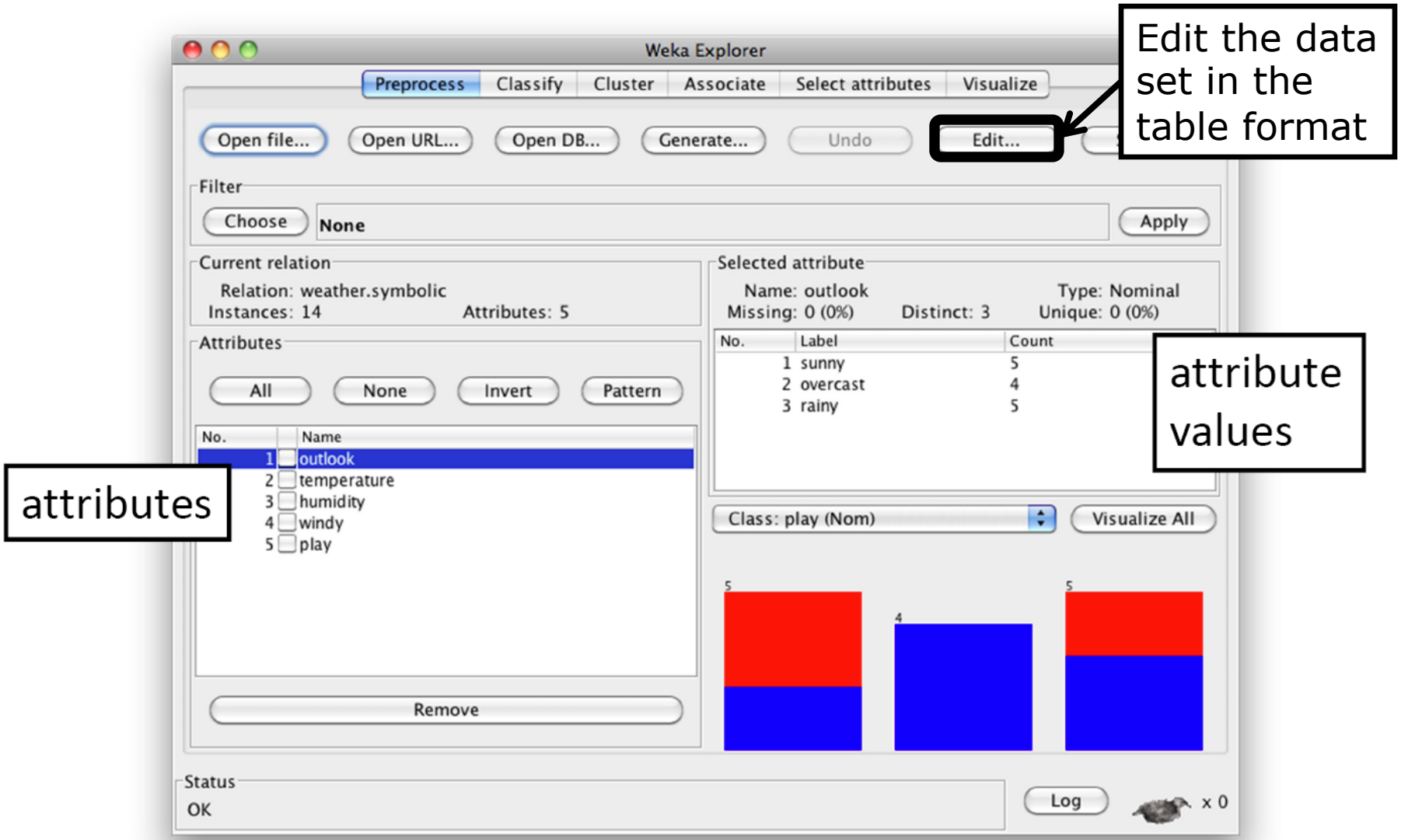
Getting Started with Weka

attributes

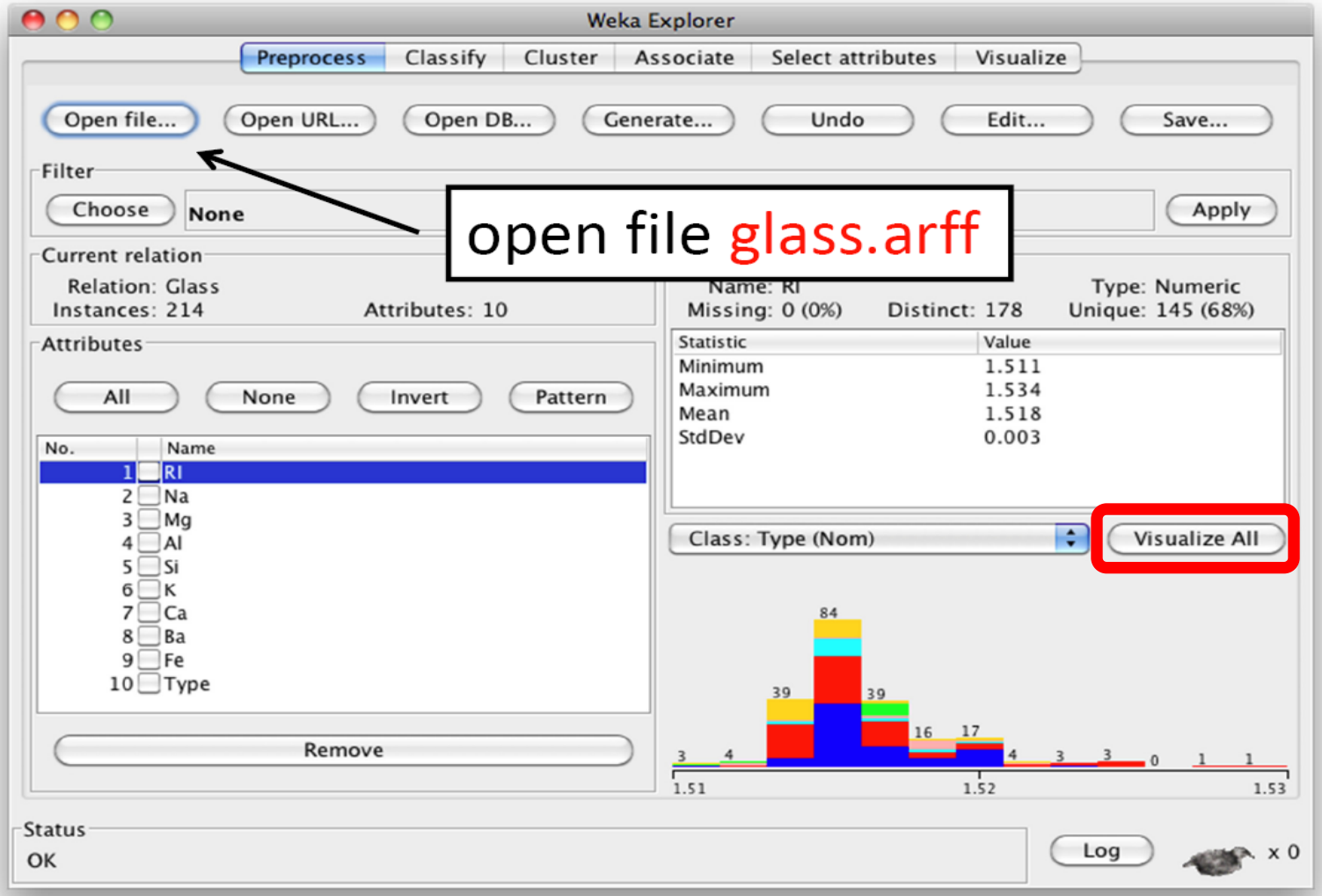
instances

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	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

Getting Started with Weka



Getting Started with Weka



Getting Started with Weka

- **WEKA only deals with "flat" files**
  - ARFF file (Attribute Relation File Format) is the default file type in Weka but data can also be imported from various formats.
    - ARFF has two sections:
      - the Header information defines attribute name, type and relations.
      - The Data section lists the data records
    - CSV: Comma Separated Values (text file)

Getting Started with Weka

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
……
```

nominal attribute

numeric attribute

Flat file in ARFF format

Getting Started with Weka

- **Use a filter to remove an attribute**
  - Open weather.nominal.arff (again!)
  - Check the filters
    - – supervised vs unsupervised
    - – attribute vs instance
  - Choose the *unsupervised attribute* filter *Remove*
  - Check the *More* information; look at the options
  - Set *attributeIndices* to *3* and click OK
  - Apply the filter
  - Recall that you can *Save* the result
  - Press *Undo*

Getting Started with Weka

- **Remove instances where *humidity* is *high***
  - Supervised or unsupervised?
  - Attribute or instance?
  - Select *RemoveWithValues*
  - Set *attributeIndex*
  - Set *nominalIndices*
  - Apply
  - Undo

Getting Started with Weka

- **Using the Visualize panel**
  - Open *iris.arff*
  - Go to Visualize panel
  - Click one of the plots; examine some instances
  - Set x axis to *petalwidth* and y axis to *petallength*
  - Click on Class color to change the color
  - Bars on the right change correspond to attributes:
    - click for x axis; right-click for y axis
  - Show Select Instance: Rectangle option
  - Submit, Reset, Clear and Save

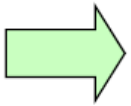
Data Mining Task—Classification

Data Mining Task—Classification

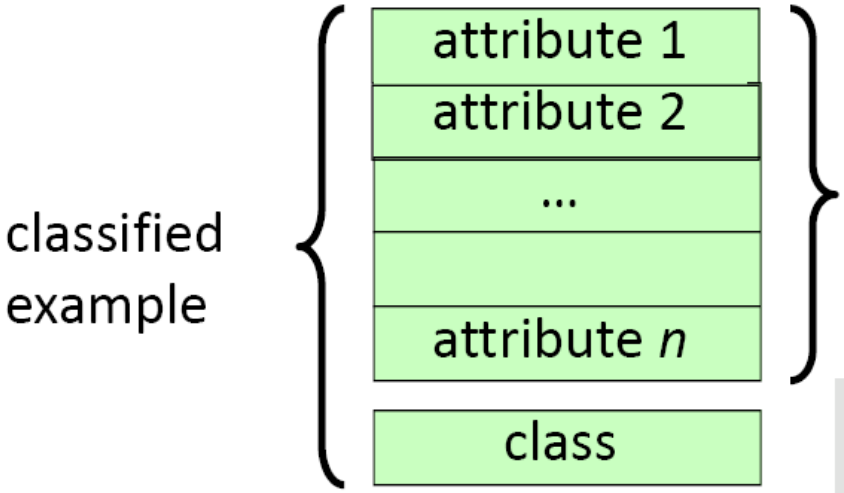
Classification

sometimes called "supervised learning"

Dataset: classified examples



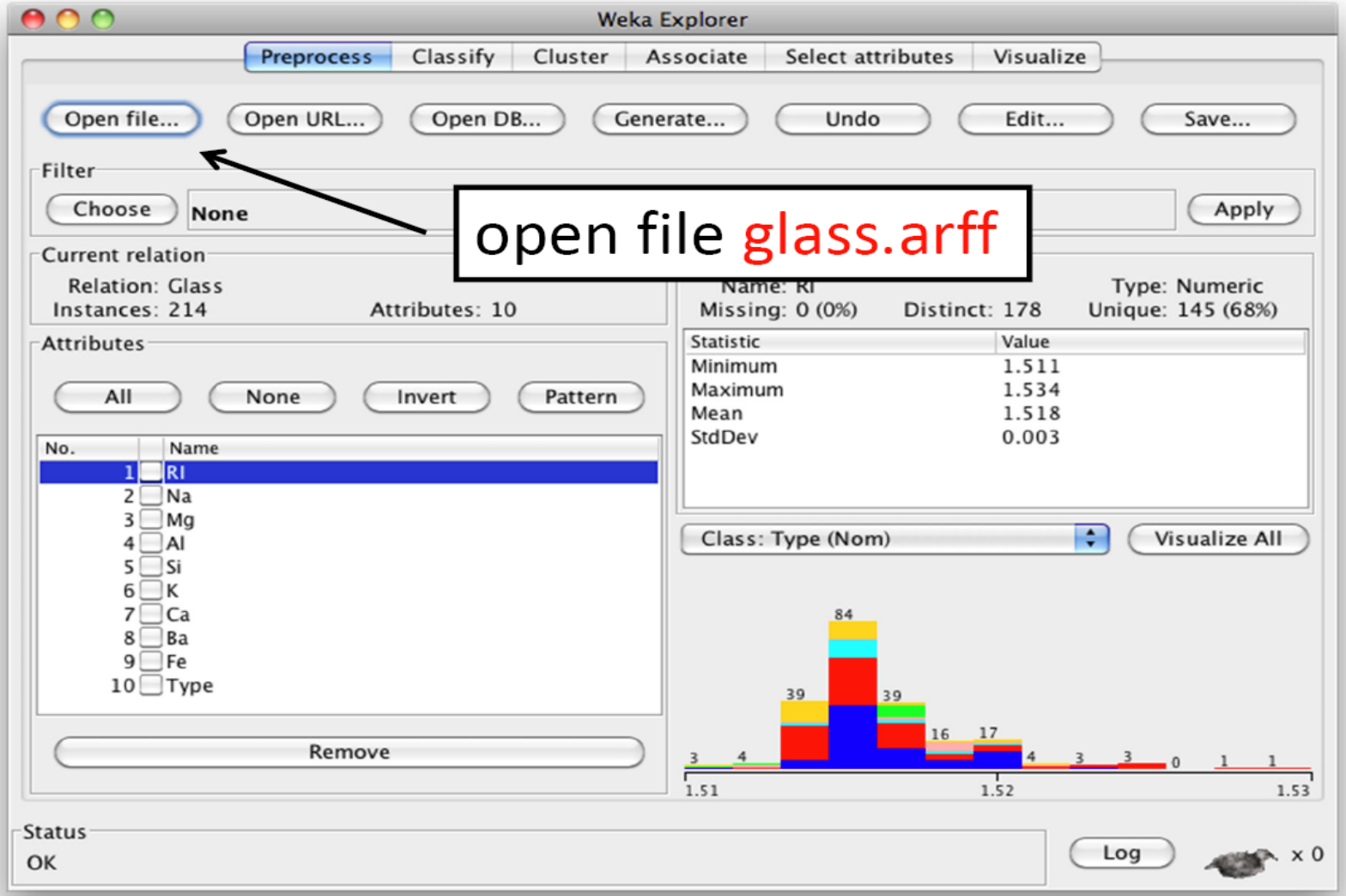
"Model" that classifies new examples



Data Mining Task—Classification

- **100+ Available Algorithms/Classifiers**
  - Bayes Classifiers: bayes>*NaiveBayes* …
  - Tree Classifiers: trees>*J48* …
  - Regression: functions>*LinearRegression*; *Logistic* …
  - Rule Based: rules>*ZeroR*; *OneR* …
  - Lazy Learner Classifiers: lazy>*IBk* …
  - SVM (Support Vector Machines): functions>*LibSVM* …
  - Ensemble Classifiers: meta>*Bagging*; *Boosting* …

Data Mining Task—Classification



Data Mining Task—Classification

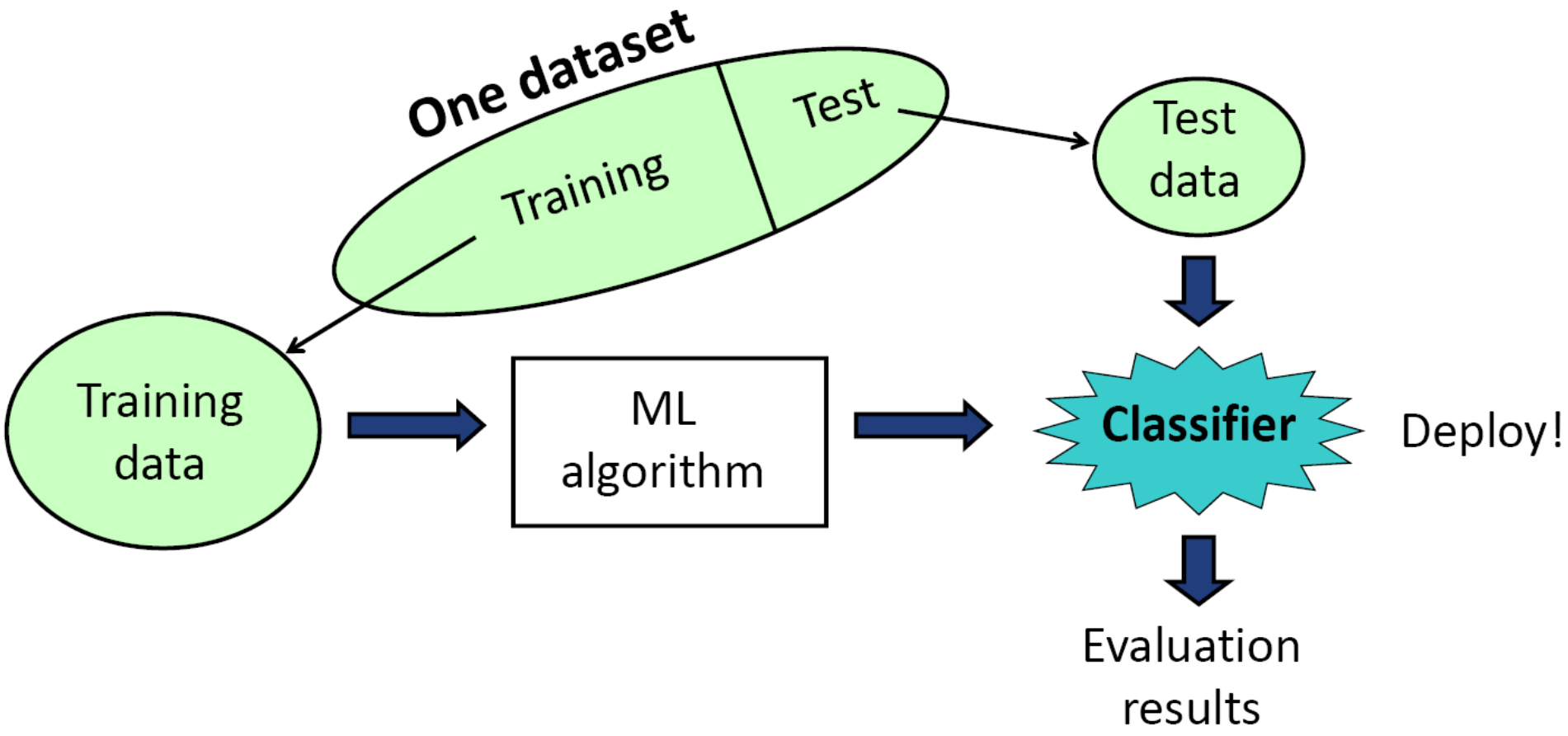
- **Use J48 to analyze the glass dataset**
  - Choose the J48 decision tree learner (trees>J48)
    - Open the configuration panel
    - Check the More information
    - Examine the options
  - Run it; Examine the output
    - Look at the correctly classified instances
    - and the confusion matrix
  - Visualize tree using right-click menu
  - Look at leaf sizes
  - Set *minNumObj* to 15 to avoid small leaves

Data Mining Task—Classification

- **Classification Boundaries**
  - Weka's Boundary Visualizer
    - WEKA GUI Chooser: *Visualization>BoundaryVisualizer*
    - Open *iris.arff*
    - Note: *petallength* on X, *petalwidth* on Y
    - Classifier: Choose *rules>OneR*
    - Check *Plot training data*
    - Click *Start*
    - In the Explorer, examine OneR's rule

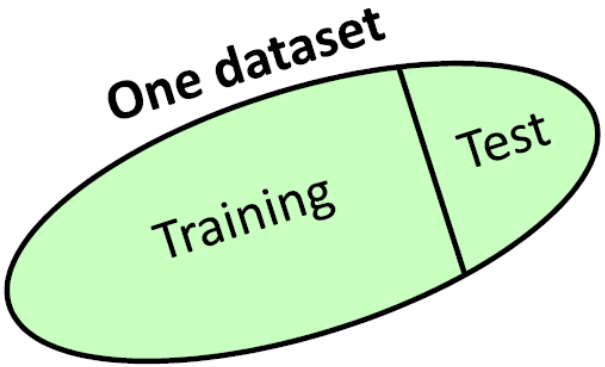
Data Mining Task—Classification

- **Evaluation: Training and Testing**

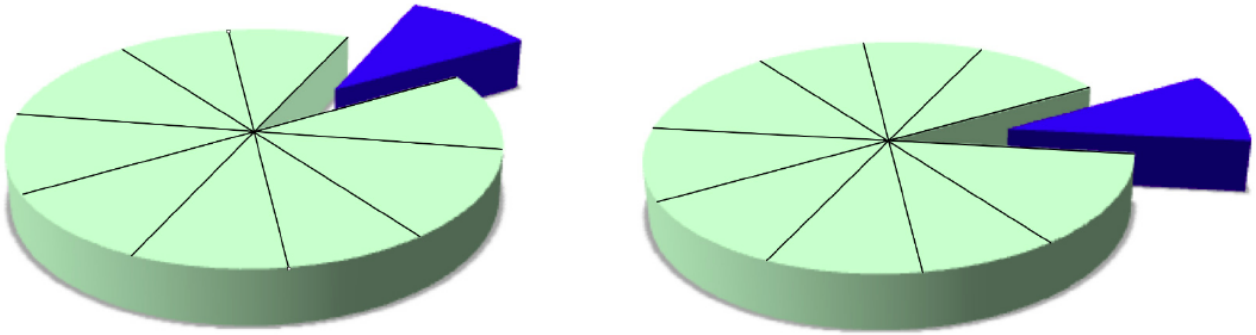


Data Mining Task—Classification

- **Evaluation: Cross-Validation**
  - **Repeated holdout**
    - **For example: hold out 10% for testing, repeat 10 times)**



- Cross-validation really is better than repeated holdout.
- It reduces the variance of the estimate.

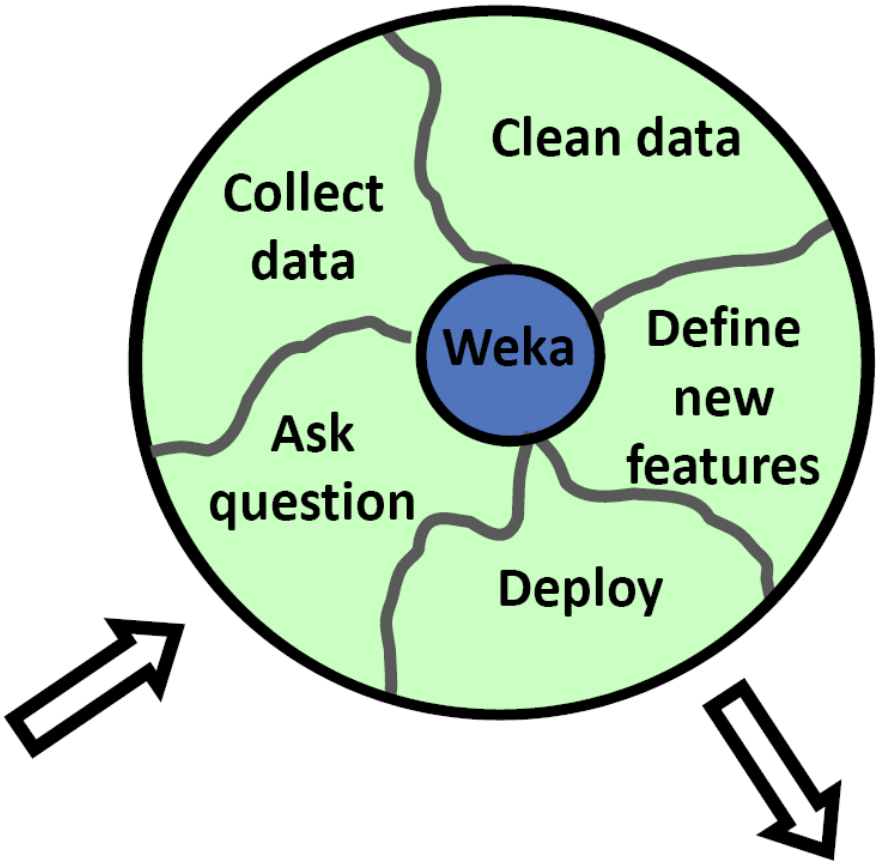


(repeat 10 times)

The Data Mining Process

The Data Mining Process

- **The Data Mining Process**



The Data Mining Process

- **Ask a question**
  - – what do you want to know
- **Gather data**
  - – more data beats a clever algorithm
- **Clean the data**
  - – real data is very messy
- **Define new features**
  - – feature engineering—the key to data mining
- **Deploy the result**
  - – technical implementation
  - – convince your boss!

The Data Mining Process

- **(Selected) filters for feature engineering**
  - **AddExpression (MathExpression)**
    - Apply a math expression to existing attributes to create new one (or modify existing one)
  - **Center (Normalize) (Standardize)**
    - – Transform numeric attributes to have zero mean (or into a given numeric range) (or to have zero mean and unit variance)
  - **Discretize (also supervised discretization)**
    - – Discretize numeric attributes to have nominal values
  - **PrincipalComponents**
    - – Perform a principal components analysis/transformation of the data

  ……

More Weka Tutorial Resources

- Weka MOOC Youtube
  - https://www.youtube.com/user/WekaMOOC
- Weka—Regression Task Example
  - http://www.ibm.com/developerworks/library/os-weka1/

Public Data Repositories

- **Links to sites with publicly available datasets**
  - There is overlap among the datasets provided at the different sites:
    - University of California Irvine Machine Learning Repository
      - a large repository of datasets supplied by individuals
    - ACM Data Mining and Knowledge Discovery Cup Center
      - contains links to instructions and datasets for the annual KDD contest

Questions?

Thanks!