THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 1 – Lesson 1

Introduction

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Data Mining with Weka

… **a practical course on how to use Weka for data mining**

… **explains the basic principles of several popular algorithms**

Ian H. Witten

University of Waikato, New Zealand

Data Mining with Weka

- **What's data mining?**
  - *We are overwhelmed with data*
  - *Data mining is about going from data to information, information that can give you useful predictions*
- **Examples??**
  - *You're at the supermarket checkout.*
    You're happy with your bargains …
    … and the supermarket is happy you've bought some more stuff
  - *Say you want a child, but you and your partner can't have one.*
    Can data mining help?
- **Data mining vs. machine learning**

Data Mining with Weka

- **What's 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

Data Mining with Weka

What will you learn?

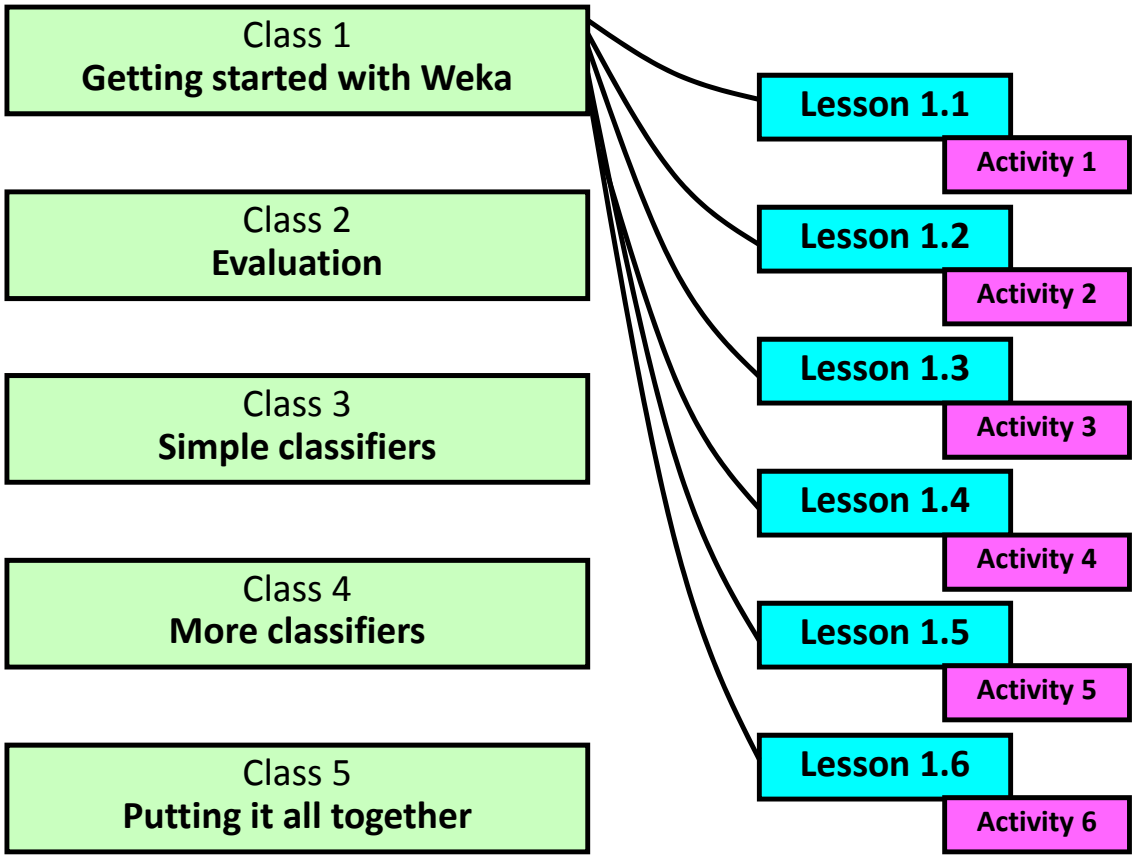
- Load data into Weka and look at it
- Use filters to preprocess it
- Explore it using interactive visualization
- Apply classification algorithms
- Interpret the output
- Understand evaluation methods and their implications
- Understand various representations for models
- Explain how popular machine learning algorithms work
- Be aware of common pitfalls with data mining

Use Weka on your own data
… and understand what you are doing!

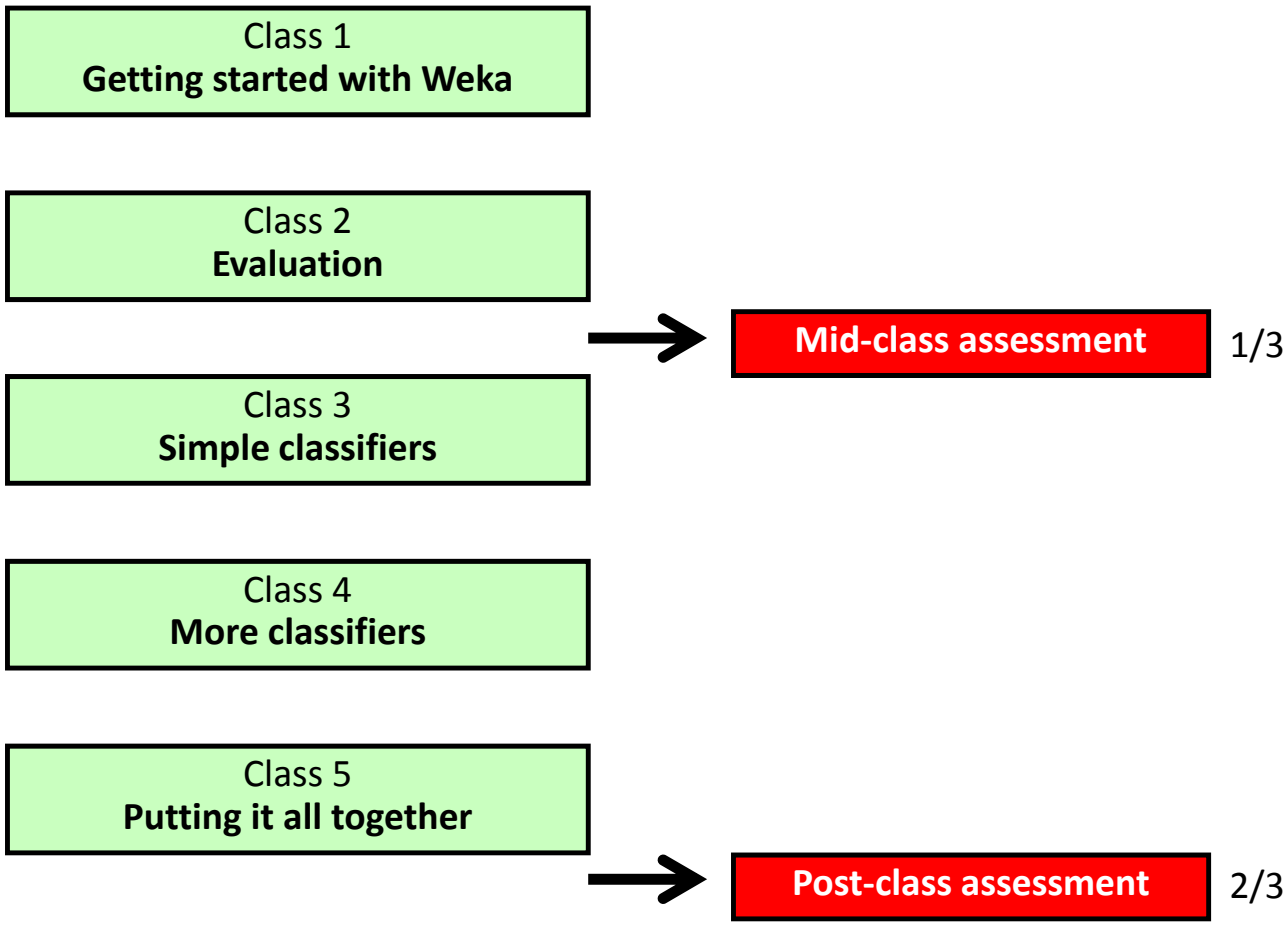
Class 1: Getting started with Weka

- Install Weka
- Explore the “Explorer” interface
- Explore some datasets
- Build a classifier
- Interpret the output
- Use filters
- Visualize your data set

Course organization



Course organization

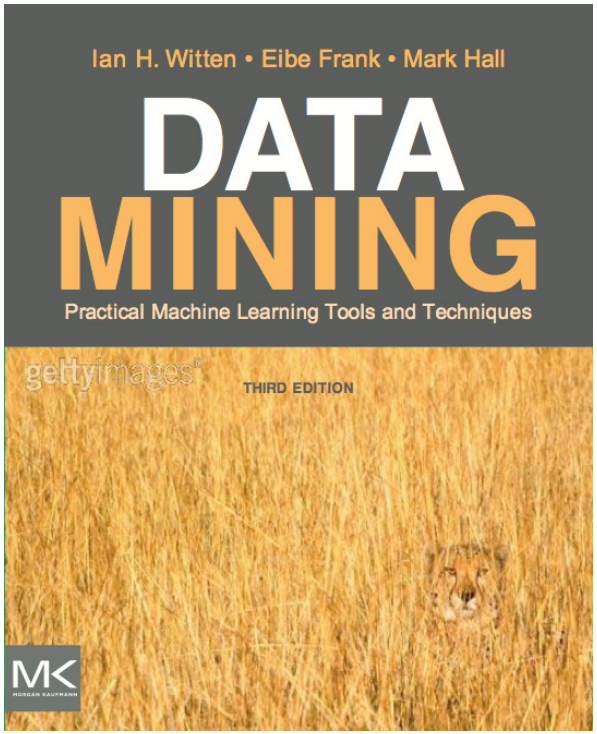


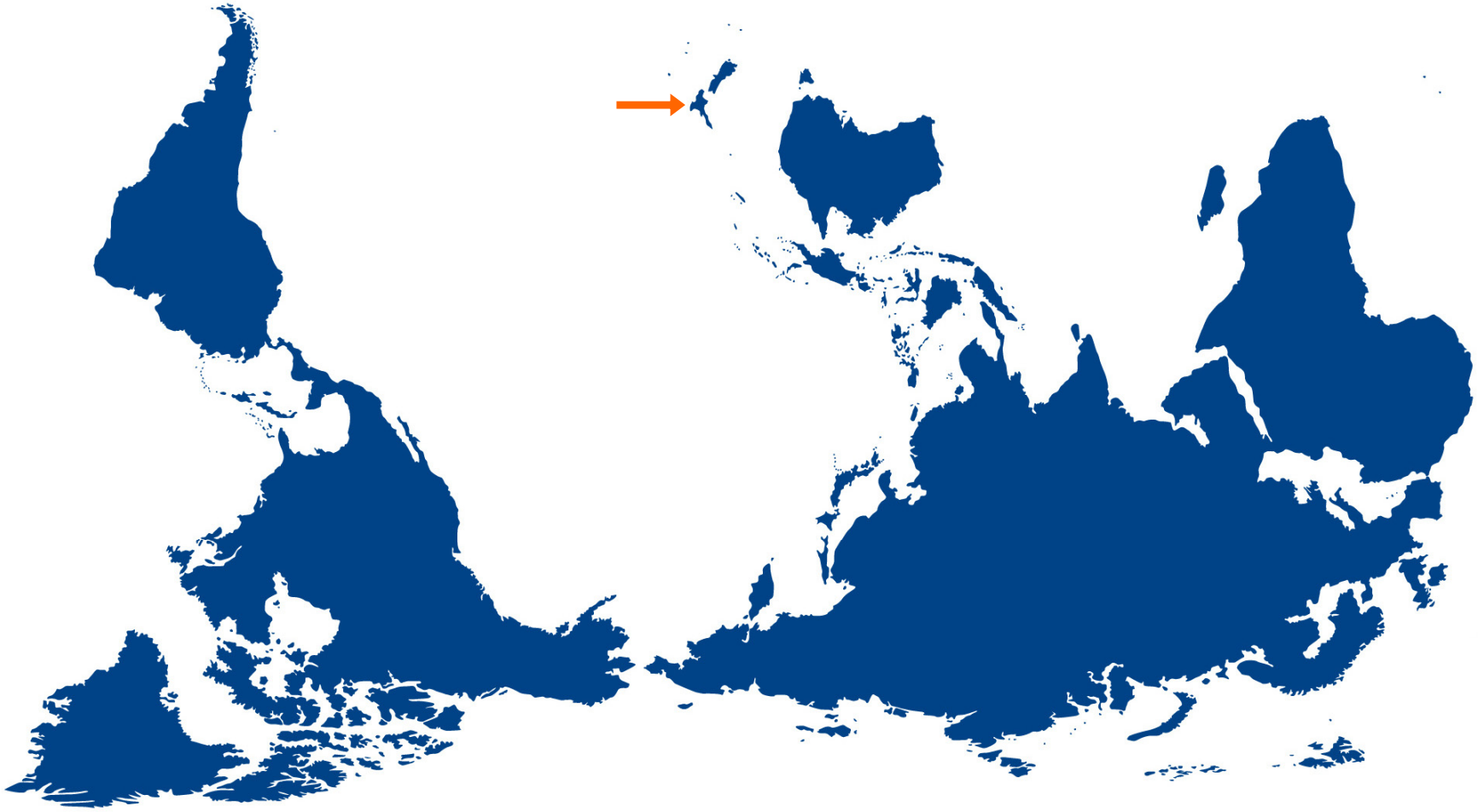
Textbook

This textbook discusses data mining, and Weka, in depth:

> *Data Mining: Practical machine learning tools and techniques*,
> by Ian H. Witten, Eibe Frank and Mark A. Hall. Morgan Kaufmann, 2011

The publisher has made available parts relevant to this course in ebook format.





THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 1 – Lesson 2

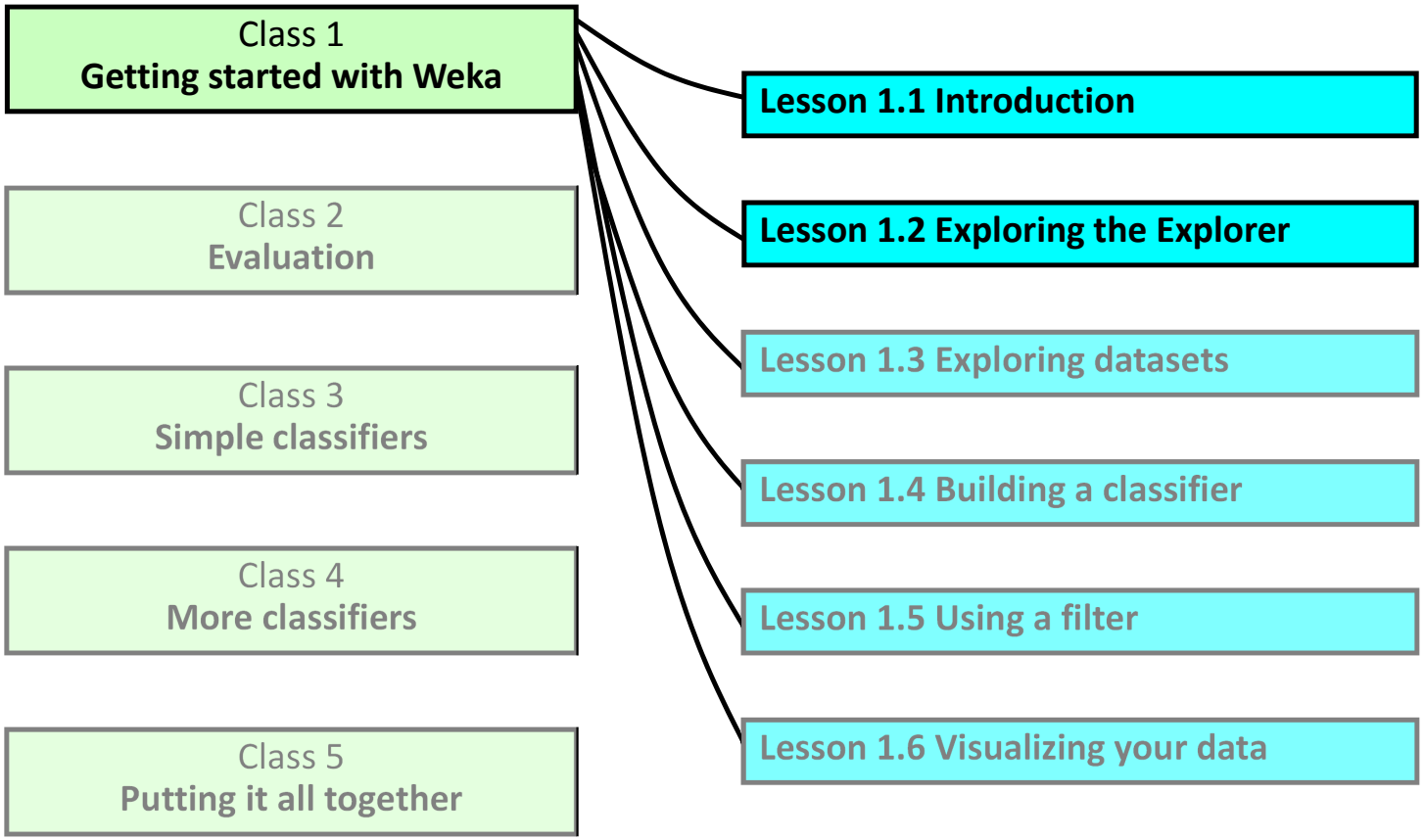
Exploring the Explorer

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 1.2: Exploring the Explorer



Lesson 1.2: Exploring the Explorer

Download from

http://www.cs.waikato.ac.nz/ml/weka

(for Windows, Mac, Linux)

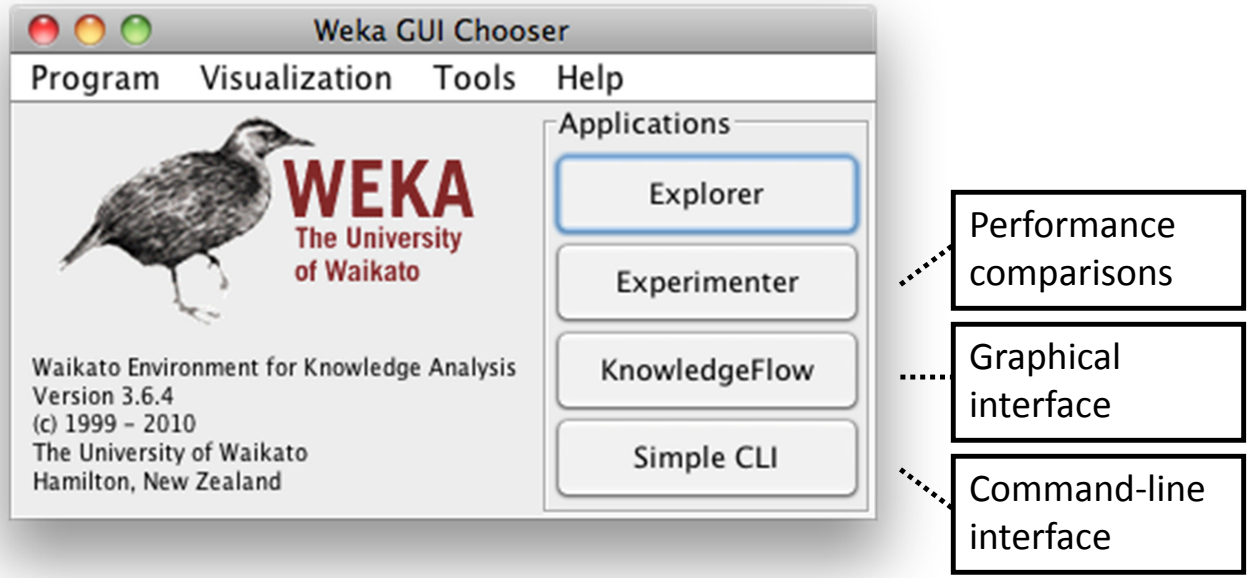
Weka 3.6.10

(the latest stable version of Weka)

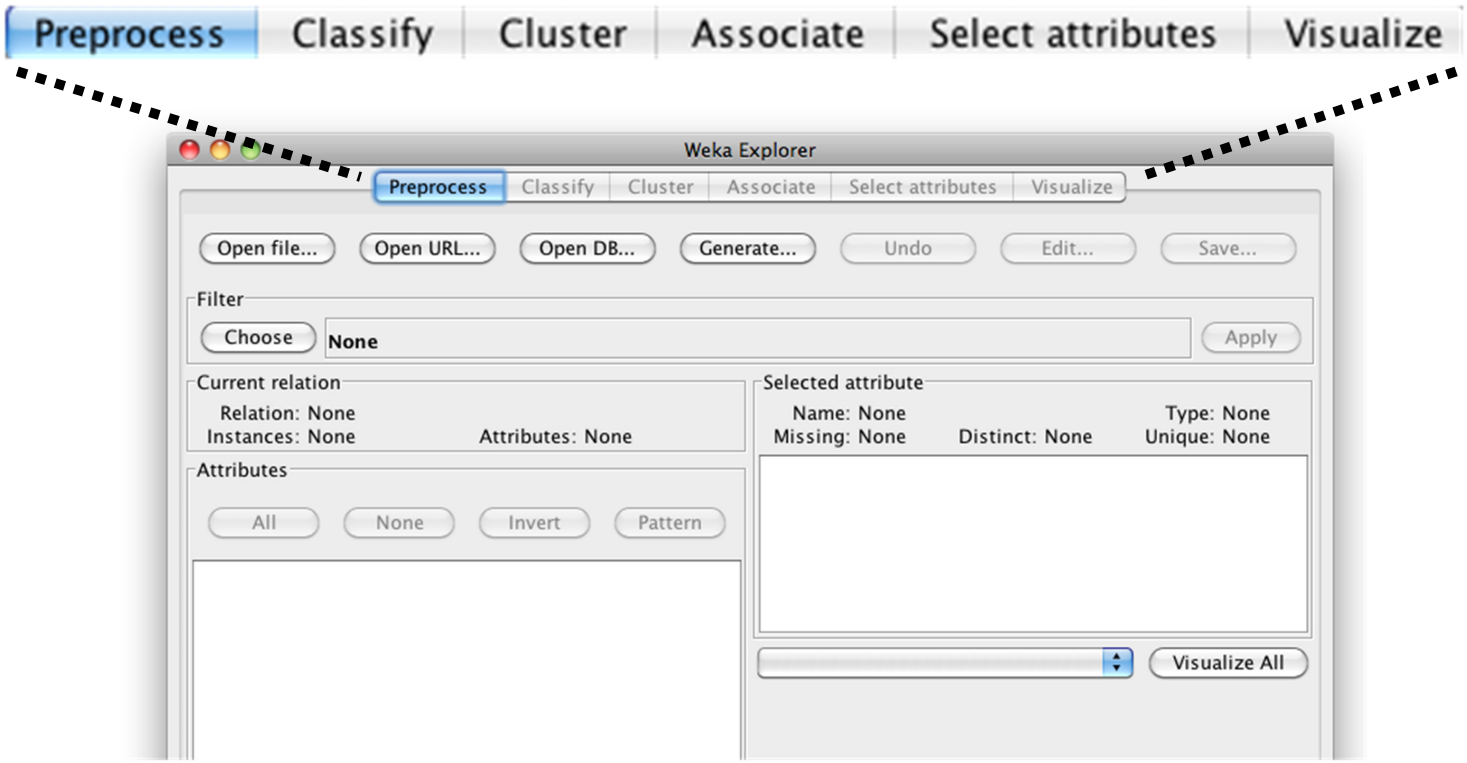
(includes datasets for the course)

(it's important to get the right version, 3.6.10)

Lesson 1.2: Exploring the Explorer



Lesson 1.2: Exploring the Explorer



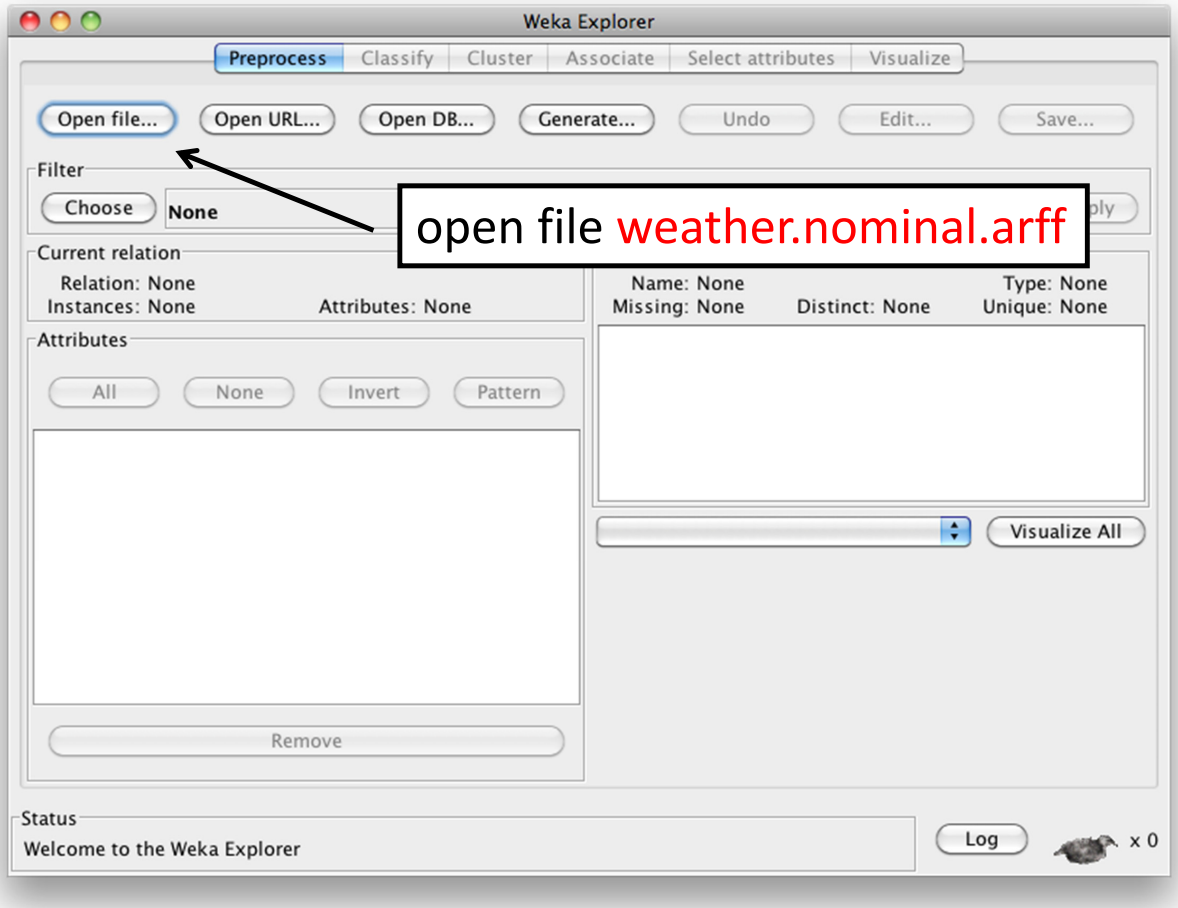
Lesson 1.2: Exploring the Explorer

attributes

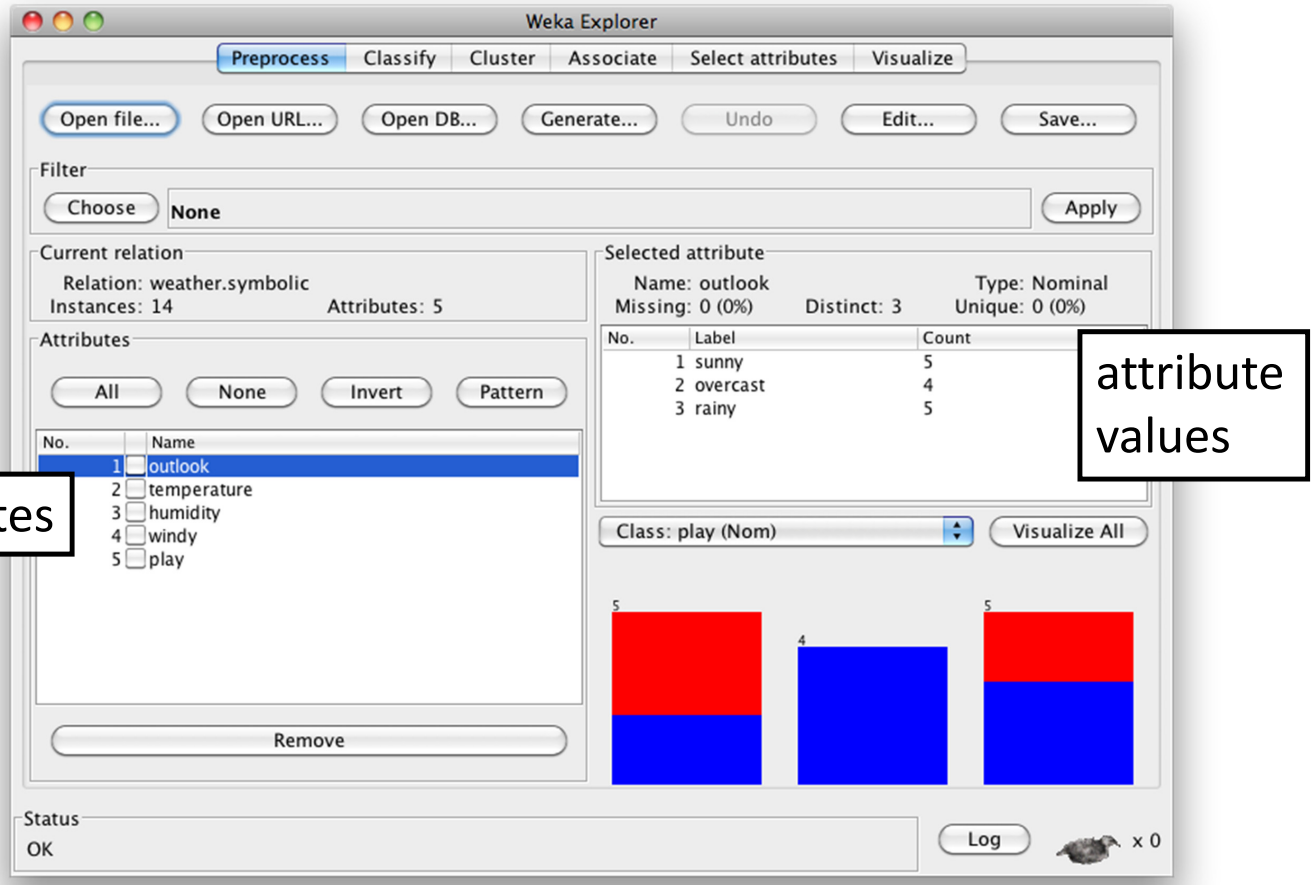
instances

	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

Lesson 1.2: Exploring the Explorer



Lesson 1.2: Exploring the Explorer

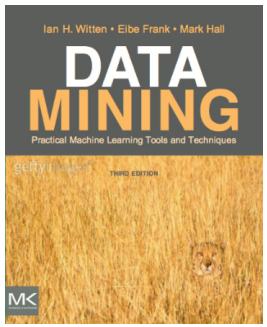


attributes

attribute values

Lesson 1.2: Exploring the Explorer

- Install Weka
- Get datasets
- Open Explorer
- Open a dataset (*weather.nominal.arff*)
- Look at attributes and their values
- Edit the dataset
- Save it?



Course text

- Section 1.2 *The weather problem*
- Chapter 10 *Introduction to Weka*

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 1 – Lesson 3

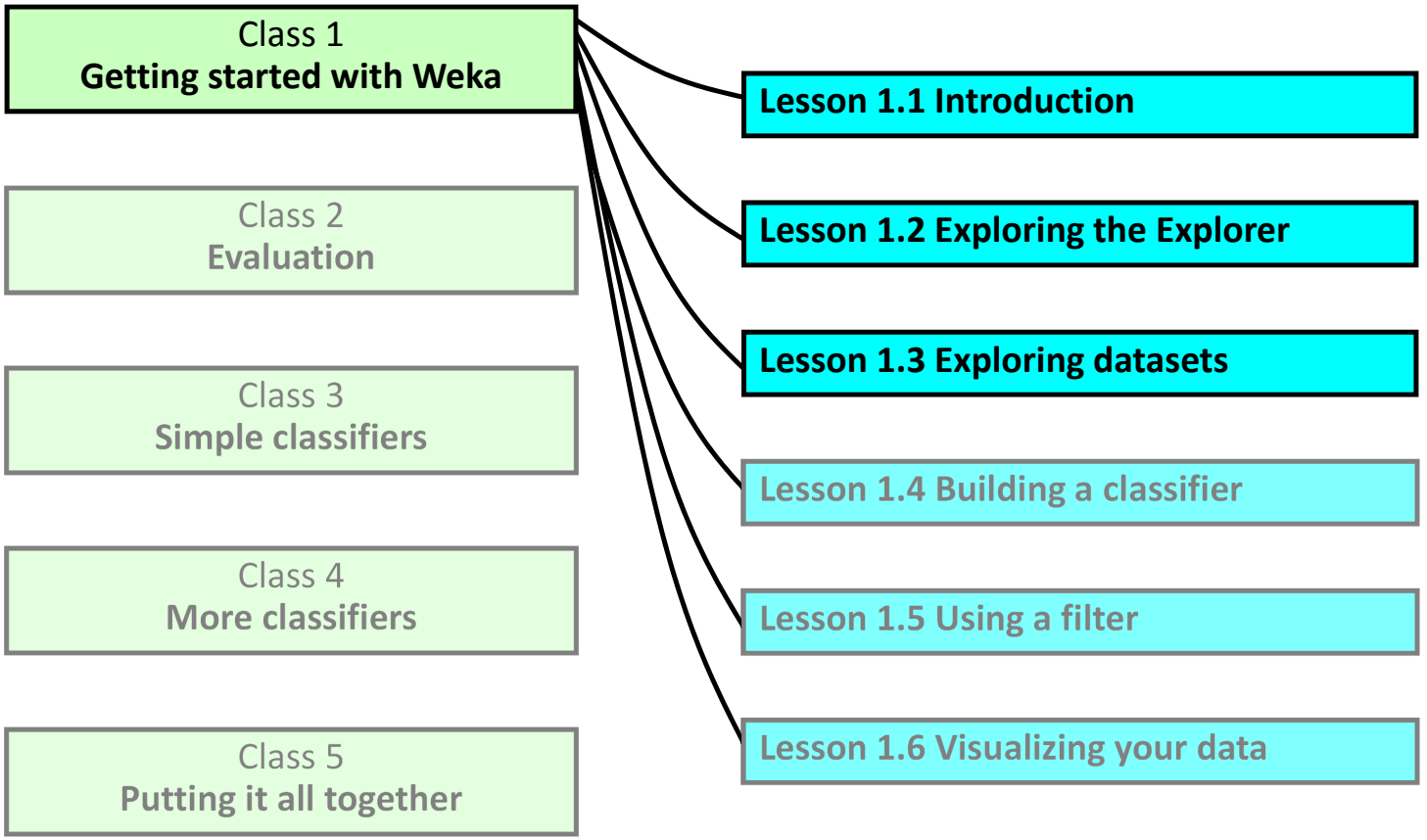
Exploring datasets

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 1.3: Exploring datasets



Lesson 1.3: Exploring datasets

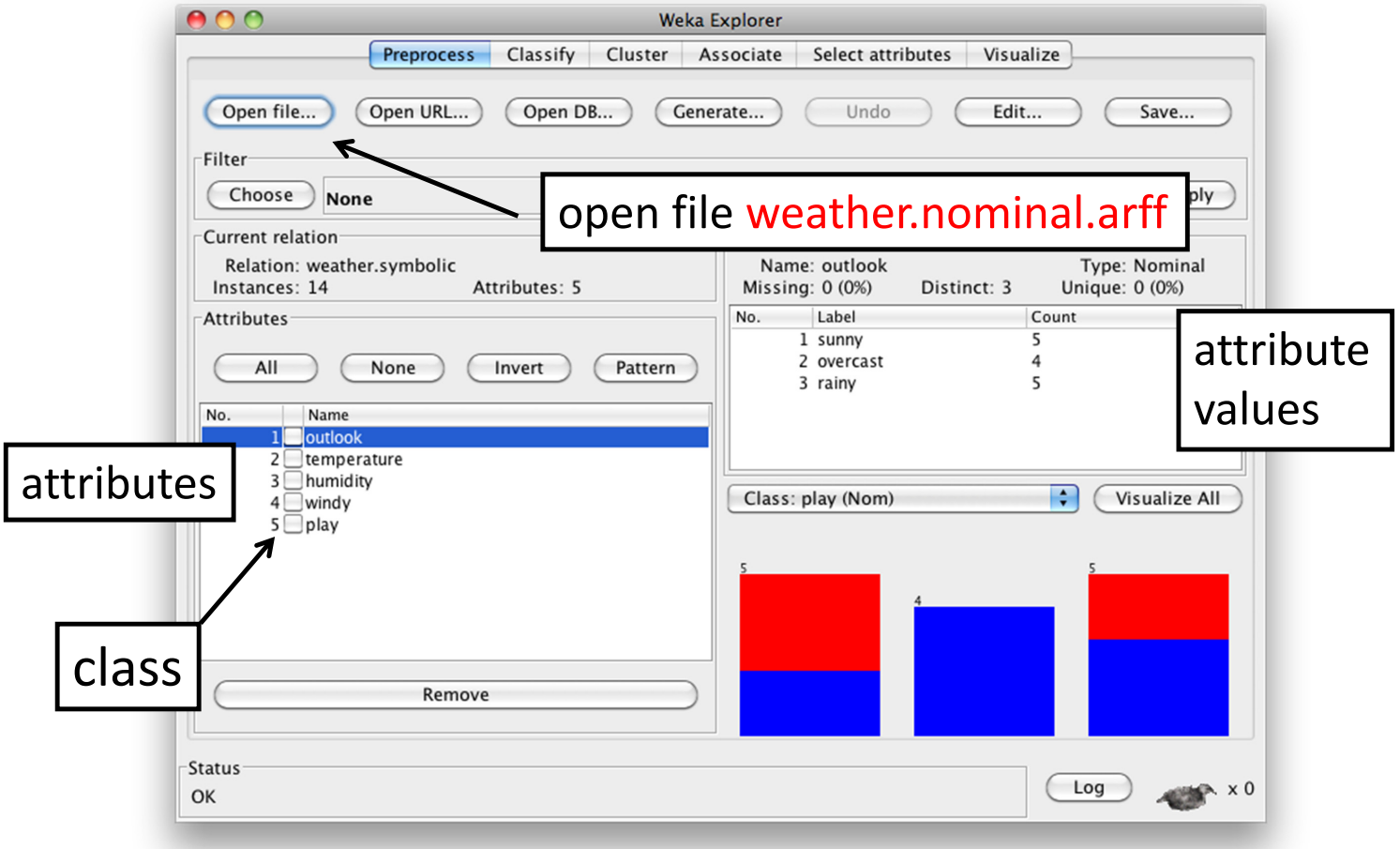
attributes

instances

	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	[illegible]gh	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]	Coo[illegible]	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

Classification problem: predict the "class" value

Lesson 1.3: Exploring datasets

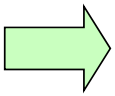


Lesson 1.3: Exploring datasets

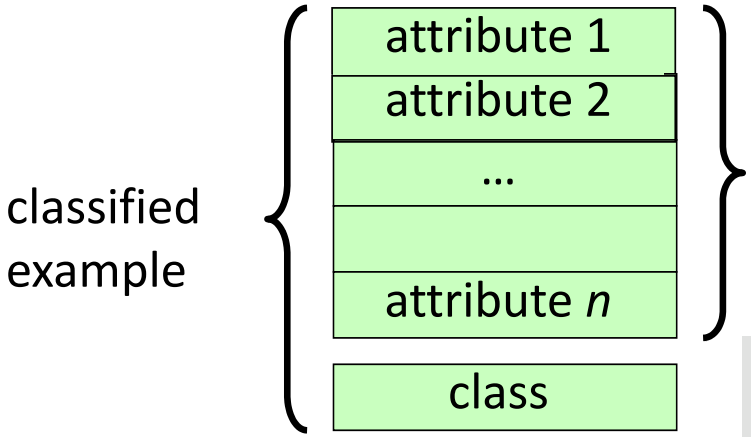
Classification

sometimes called "supervised learning"

Dataset: classified examples



"Model" that classifies new examples

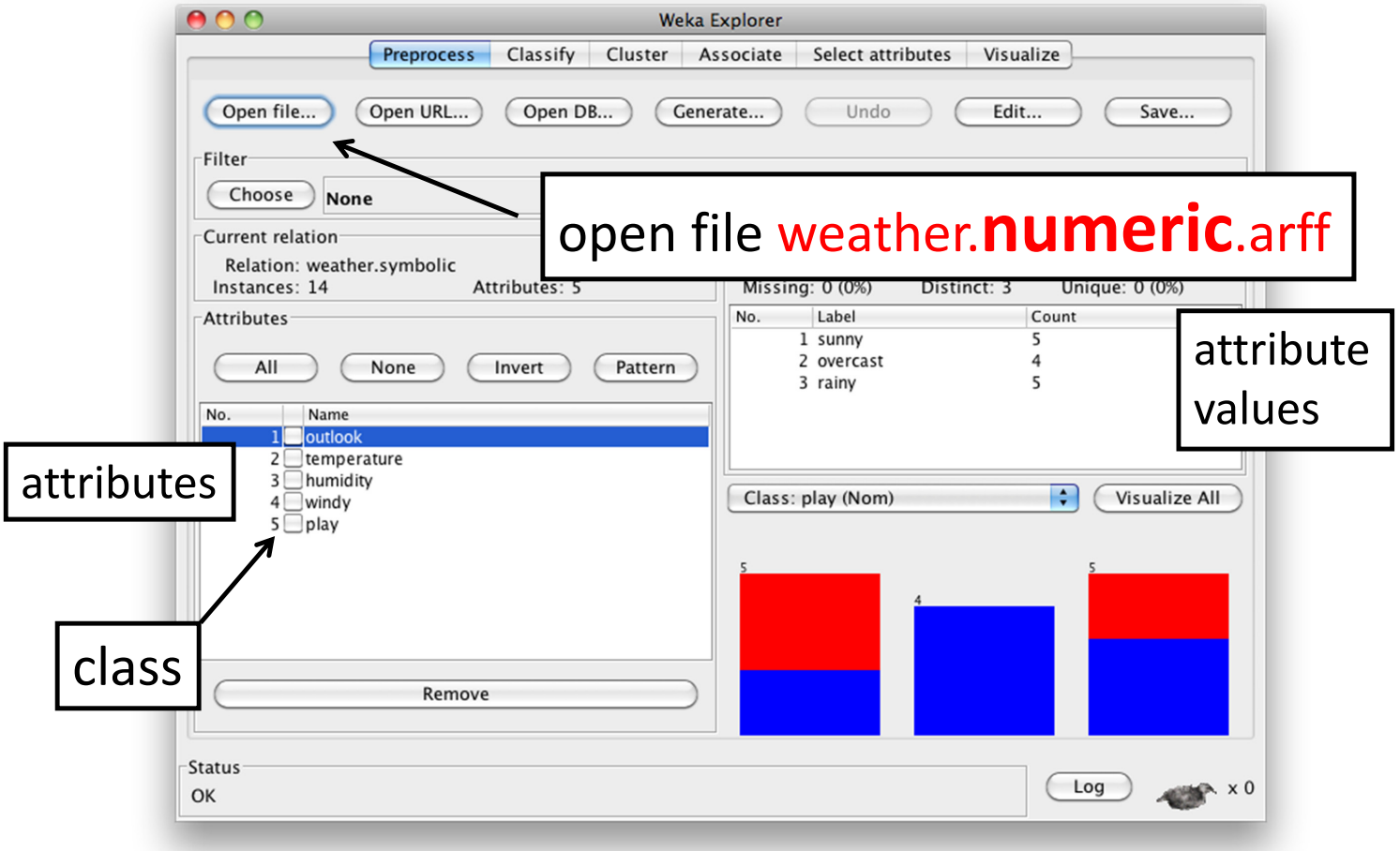


instance:
fixed set of features

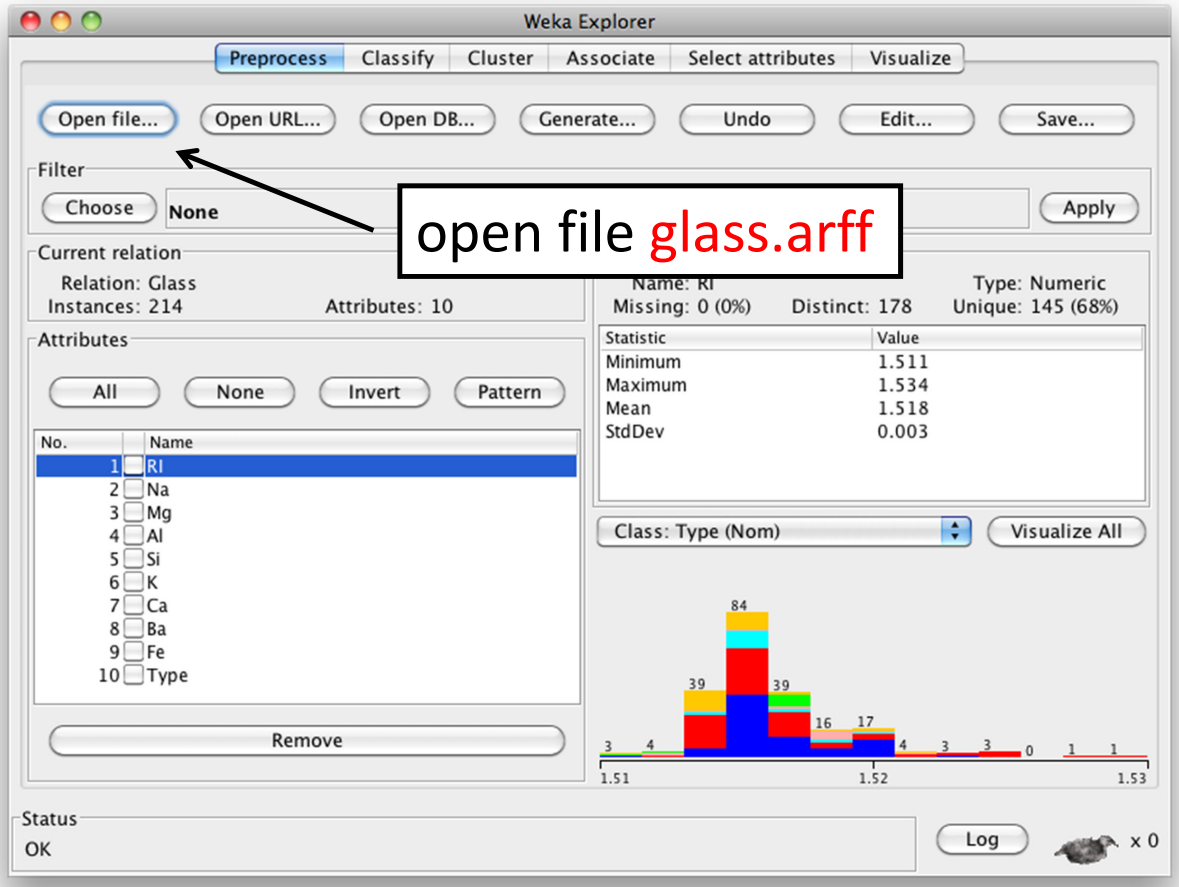
discrete ("nominal")
continuous ("numeric")

discrete: "classification" problem
continuous: "regression" problem

Lesson 1.3: Exploring datasets

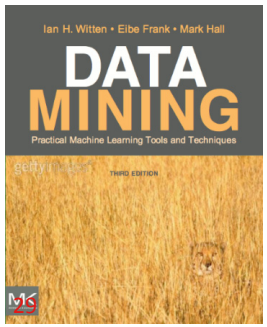


Lesson 1.3: Exploring datasets



Lesson 1.3: Exploring datasets

- The classification problem
- *weather.nominal*, *weather.numeric*
- Nominal vs numeric attributes
- ARFF file format
- *glass.arff* dataset
- Sanity checking attributes



Course text

- Section 11.1 *Preparing the data*
  Loading the data into the Explorer

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 1 – Lesson 4

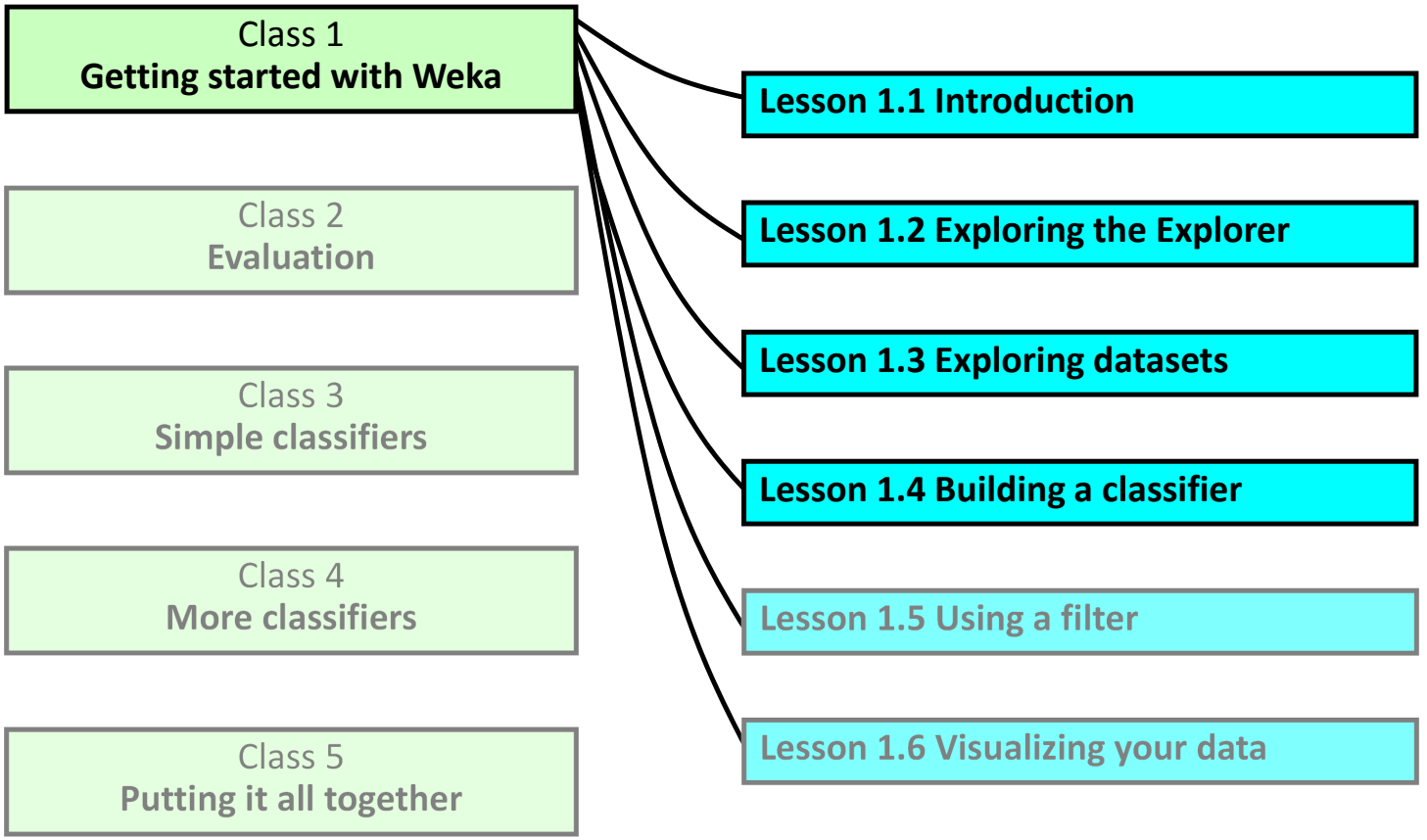
Building a classifier

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 1.4: Building a classifier



Lesson 1.4: Building a classifier

Use J48 to analyze the glass dataset

- Open file glass.arff

  (or leave it open from the last lesson)
- 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

Lesson 1.4: Building a classifier

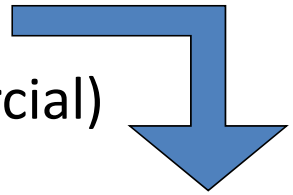
Investigate J48

- 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

Lesson 1.4: Building a classifier

From C4.5 to J48

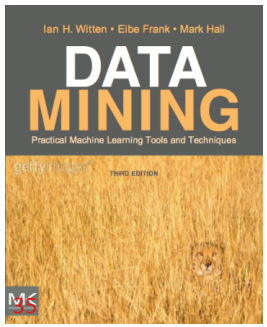
- ID3 (1979)
- **C4.5** (1993)
- C4.8 (1996?)
- C5.0 (commercial)



J48

Lesson 1.4: Building a classifier

- Classifiers in Weka
- Classifying the *glass* dataset
- Interpreting J48 output
- J48 configuration panel
- … option: pruned vs unpruned trees
- … option: avoid small leaves
- J48 ~ C4.5



Course text

- Section 11.1 *Building a decision tree*
  Examining the output

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 1 – Lesson 5

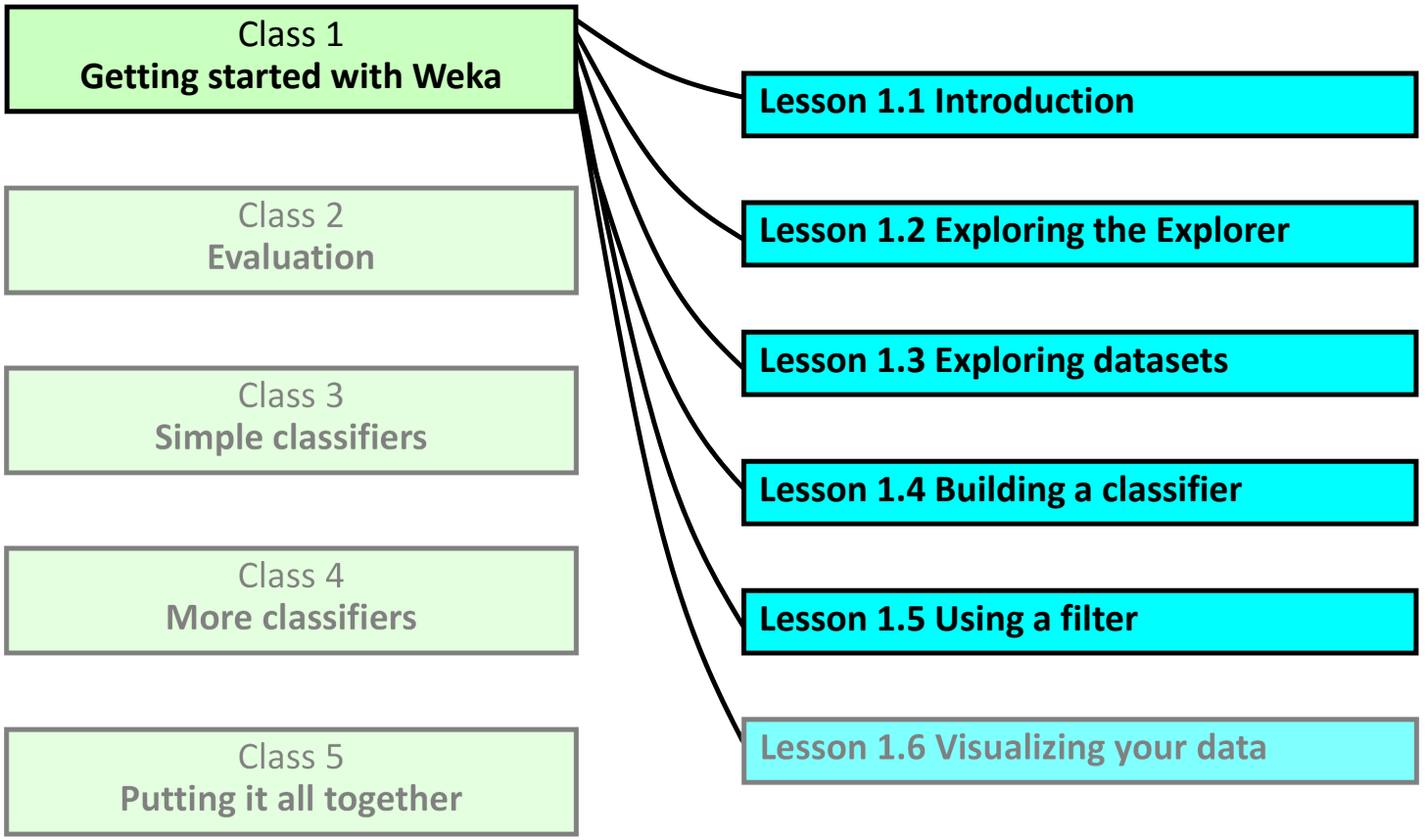
Using a filter

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 1.5: Using a filter



Lesson 1.5: Using a filter

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*

Lesson 1.5: Using a filter

Remove instances where *humidity* is *high*

- Supervised or unsupervised?
- Attribute or instance?
- Look at them
- Select *RemoveWithValues*
- Set *attributeIndex*
- Set *nominalIndices*
- Apply
- *Undo*

Lesson 1.5: Using a filter

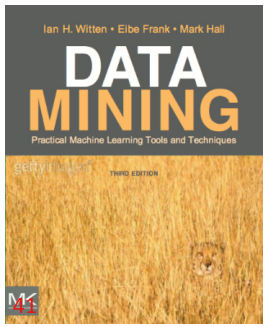
Fewer attributes, better classification!

- Open glass.arff
- Run J48 (trees>J48)
- Remove Fe
- Remove all attributes except RI and MG
- Look at the decision trees

- Use right-click menu to visualize decision trees

Lesson 1.5: Using a filter

- Filters in Weka
- Supervised vs unsupervised,
  attribute vs instance
- To find the right one, you need to look!
- Filters can be very powerful
- Judiciously removing attributes can
  - improve performance
  - increase comprehensibility



Course text

- Section 11.2 *Loading and filtering files*

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 1 – Lesson 6

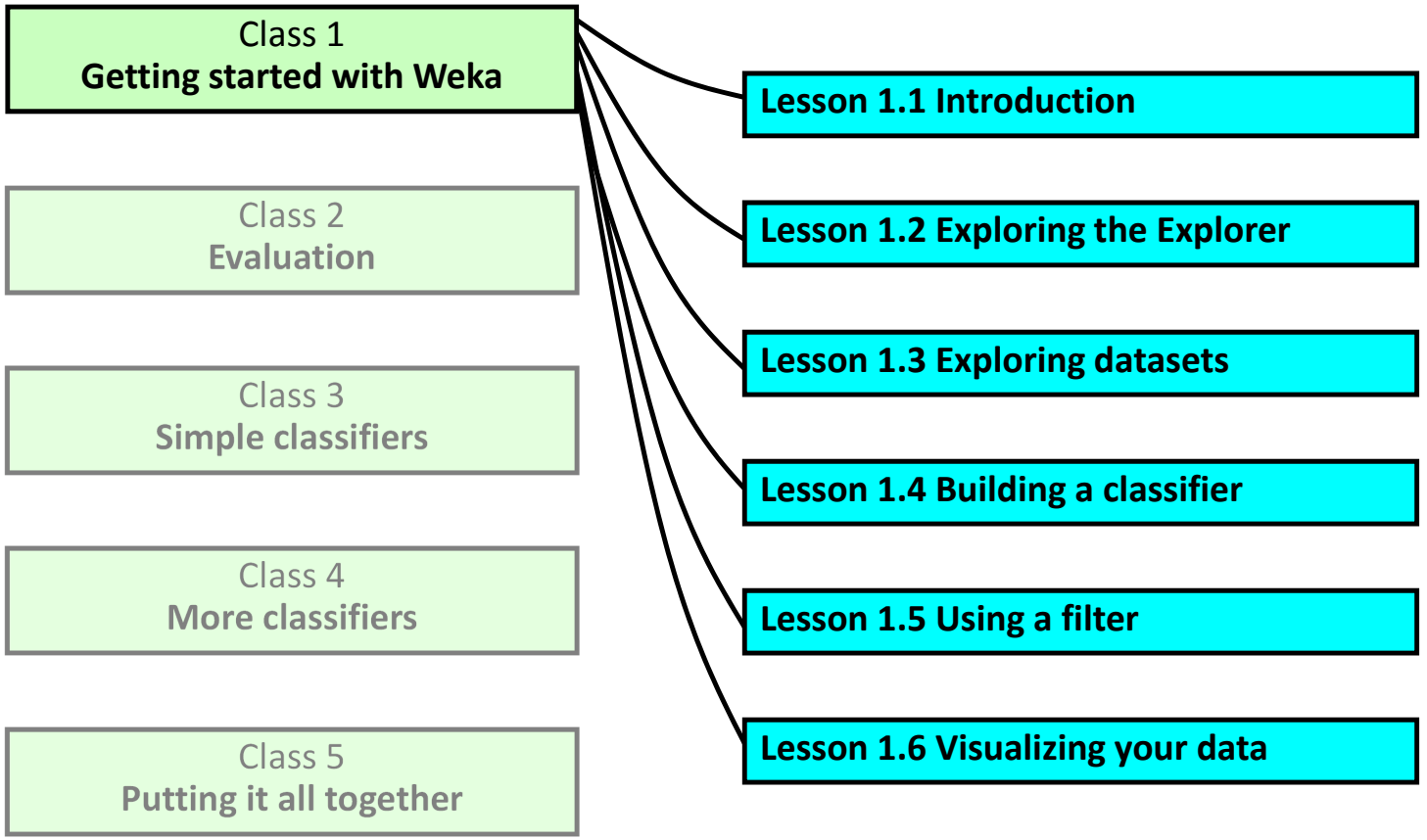
Visualizing your data

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 1.6: Visualizing your data



Lesson 1.6: Visualizing your data

Using the Visualize panel

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

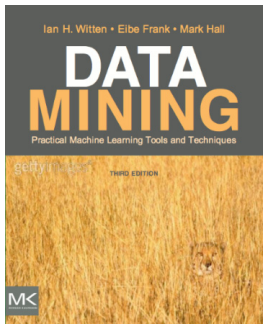
Lesson 1.6: Visualizing your data

Visualizing classification errors

- Run J48 (trees>J48)
- Visualize classifier errors (from Results list)
- Plot predictedclass against class
- Identify errors shown by confusion matrix

Lesson 1.6: Visualizing your data

- Get down and dirty with your data
- Visualize it
- Clean it up by deleting outliers
- Look at classification errors
  - (there's a filter that allows you to add classifications as a new attribute)



Course text

Section 11.2 Visualization

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz