

# WEKA簡介與實作

## Chapter 1. 認識Weka

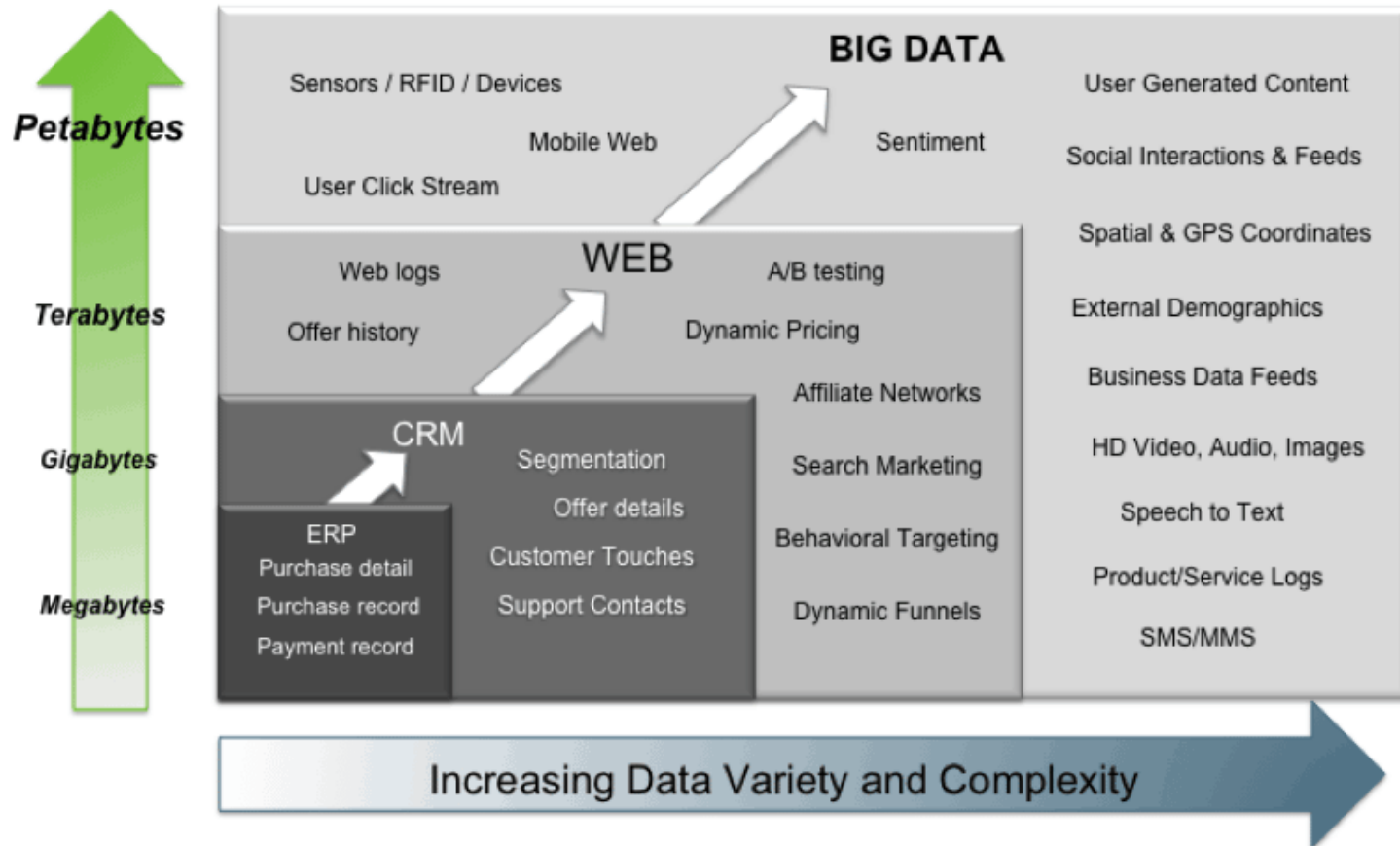
布丁布丁吃布丁

2019年8月9日



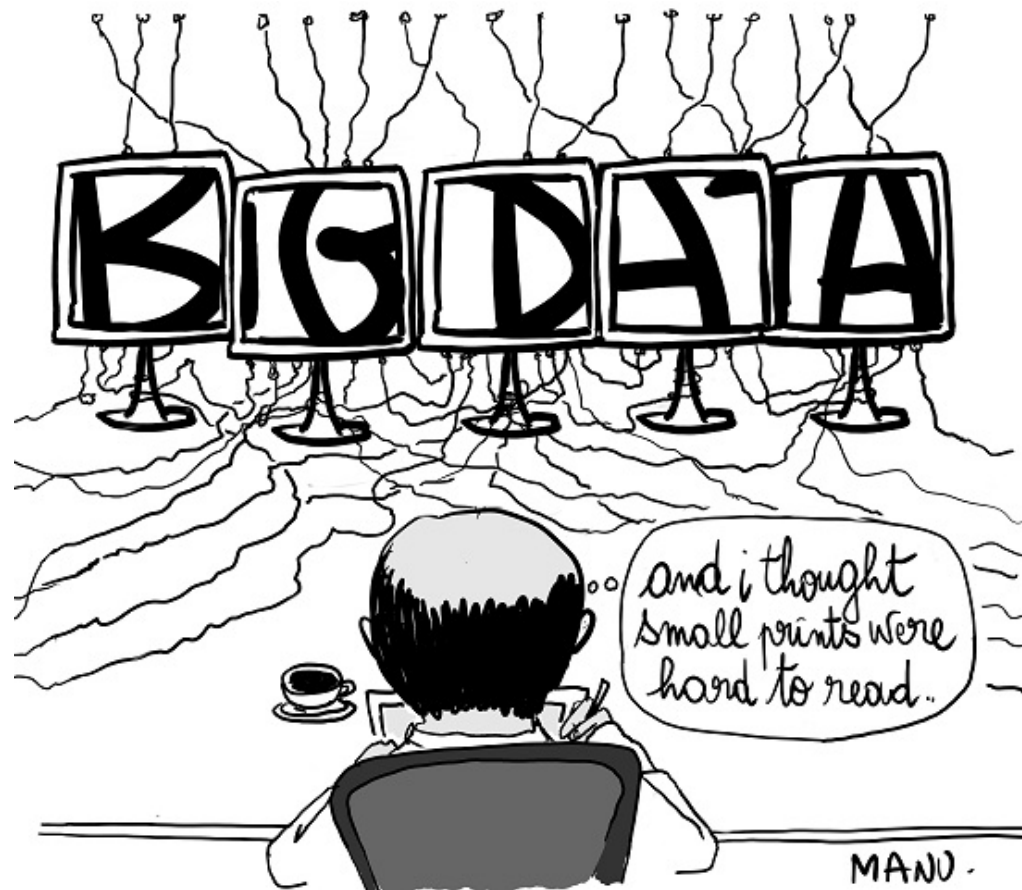
# 什麼是大數據？

Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

# BIG DATA



就是你看不完的資料







# 如何找出資料裡的異常個案？

## 如何探索資料的共同模式？

如何找出跟特定類別  
最相關的屬性規則？

# 如何預測資料的可能類別？

## 如何預測資料的接近數值？





# Weka懂

如何找出跟特定類別最相關的屬性規則？

# 如何預測資料的可能類別？

# 如何預測資料的接近數值？



# 課程大綱 (1/2)



## Chapter 1. 認識Weka

1. 認識Weka
2. Weka的資料來源
3. 準備Weka :  
下載、安裝與設定
4. 認識Weka架構

## Chapter 2. 探索性與比較性分析

5. 探索性分析：分群
6. 探索性分析：異常偵測
7. 比較性分析：  
關聯規則探勘

# 課程大綱 (2/2)

## Chapter 3. 預測性分析

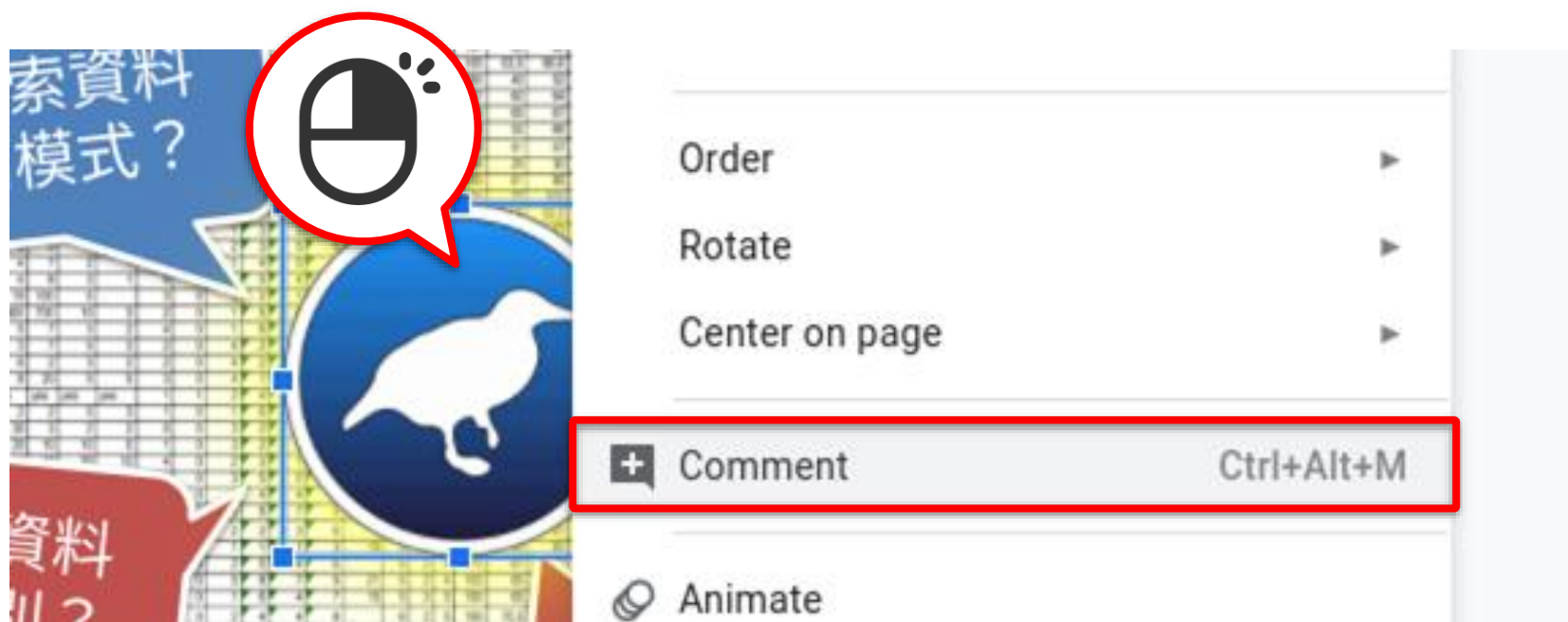
- 8. 預測性分析：分類
- 9. 預測性分析：迴歸

## Chapter 4. 進階應用與結語

- 10. Weka的進階應用
- 11. 結語



# 請多多善用「註解」





Part 1.

# 認識Weka





# Weka的出生地 紐西蘭懷卡託大學





# 開放原始碼工具 Weka

- 紐西蘭懷卡托大學機器學習實驗室專為學習資料探勘所開發的Java軟體，可用於**研究**、**教學**、**應用**等各種用途
- 包含完整的資料探勘處理流程，含括**資料前處理**工具、機器學習**演算法**、成效**評估**方法、資訊**視覺化**報表摘要
- 兼具**圖形化使用者介面**與**指令列**應用工具
  - 易於比較不同演算法的分析結果
  - 模組化設計，能夠擴充不同的演算法
- **跨平臺**：Windows、Mac OS、Linux
- 1993年開發初版，至今最新版本是2018年發佈的3.9.3



# Weka命名的由來



- Weka是指紐西蘭秧雞 (Gallirallus australis)
- 紐西蘭地區的一種不會飛的特有種鳥類

W  
a  
i  
k  
a  
t  
o  
E  
n  
v  
i  
r  
o  
n  
m  
e  
n  
t  
f  
o  
r  
K  
n  
o  
w  
l  
e  
d  
g  
e  
A  
n  
a  
l  
y  
s  
i  
s



W  
e  
k  
a

# Weka對於資料探勘的支援

Cluster  
分群

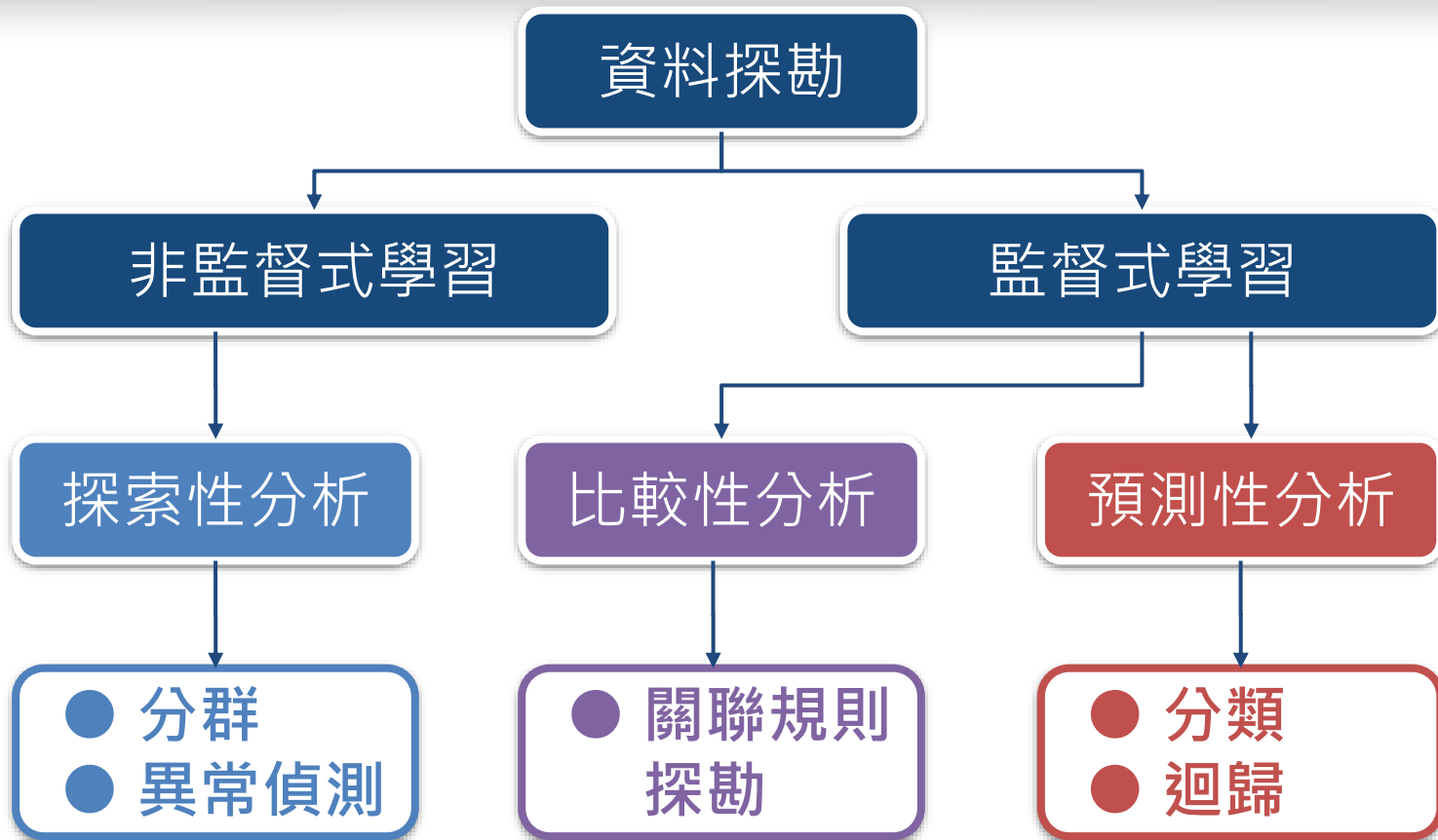
Association Rule  
關聯式規則

Classification  
分類

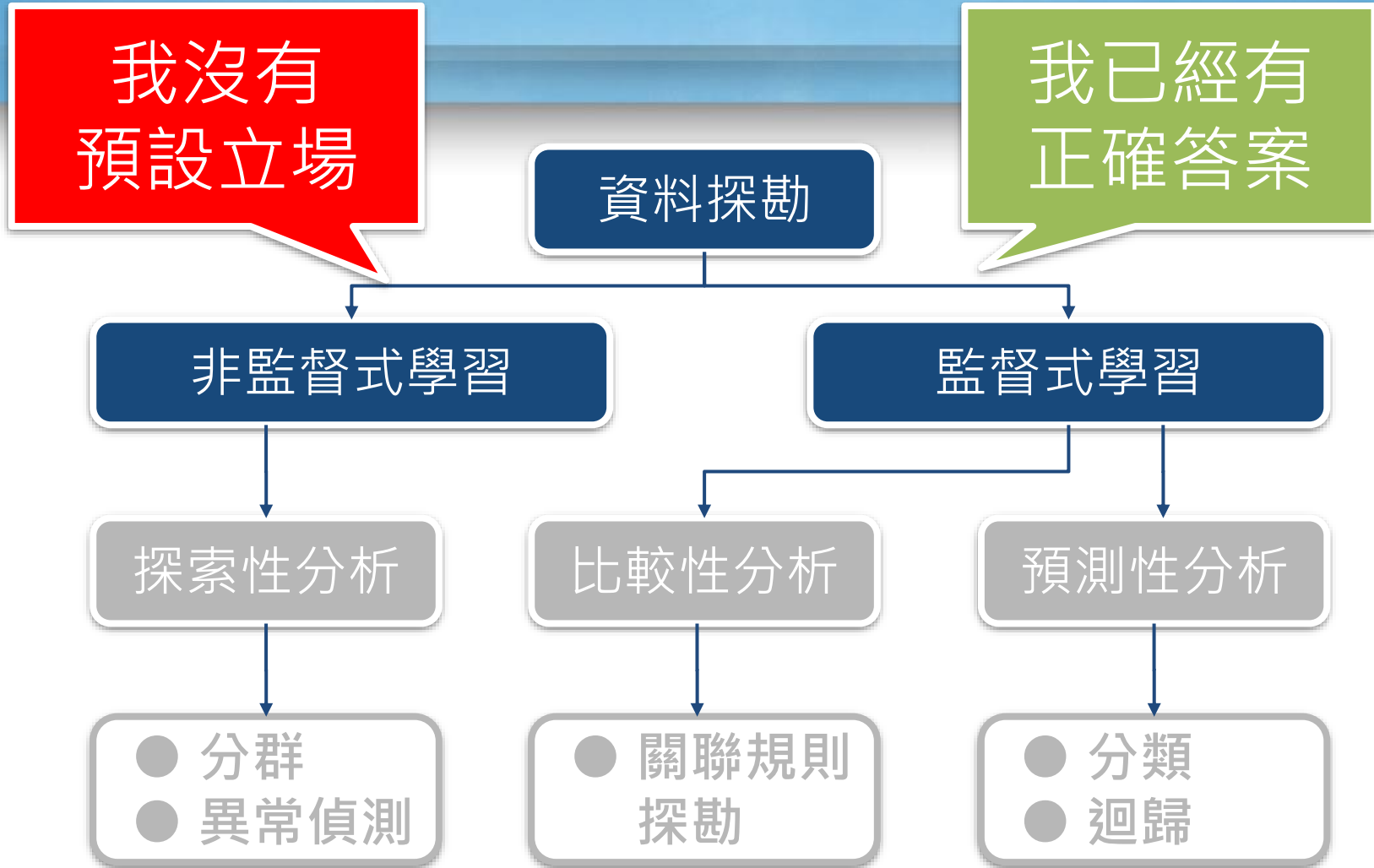




# 本課程對於資料探勘的分類



# 資料探勘的目標



# 探索性分析的目標

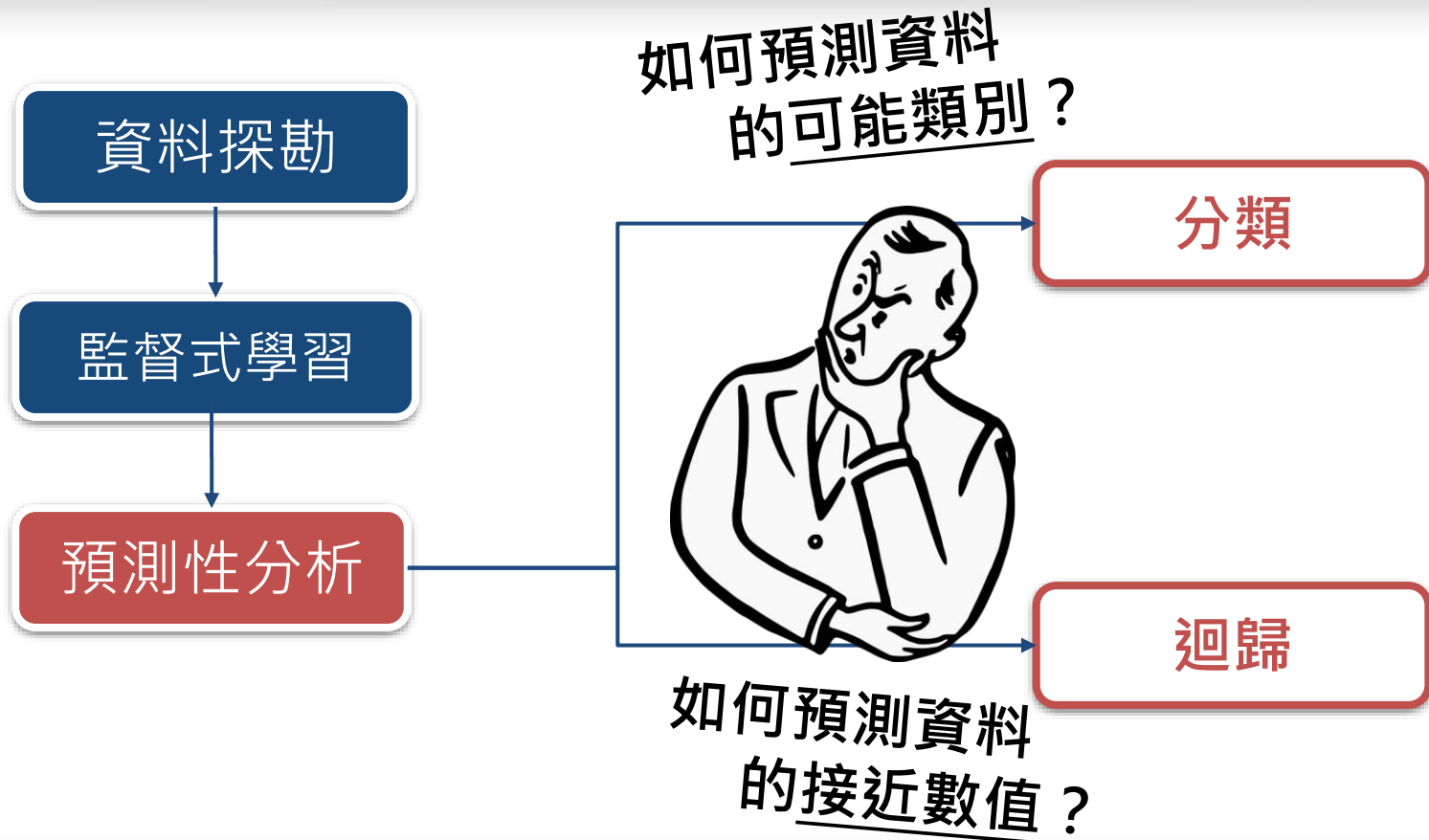




# 比較性分析的目標



# 預測性分析的目標



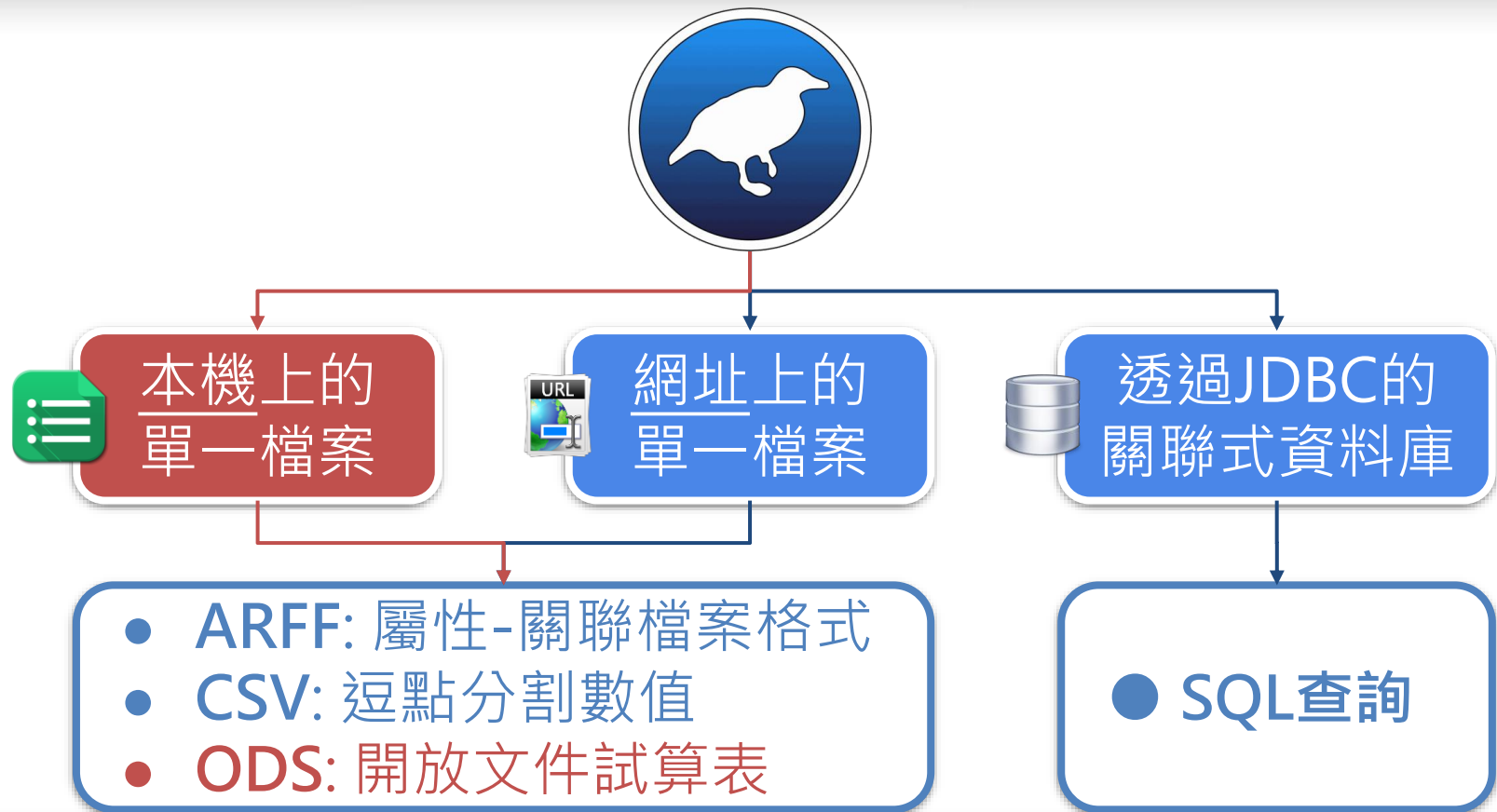


Part 2.

# Weka的資料來源



# Weka可接受的資料來源

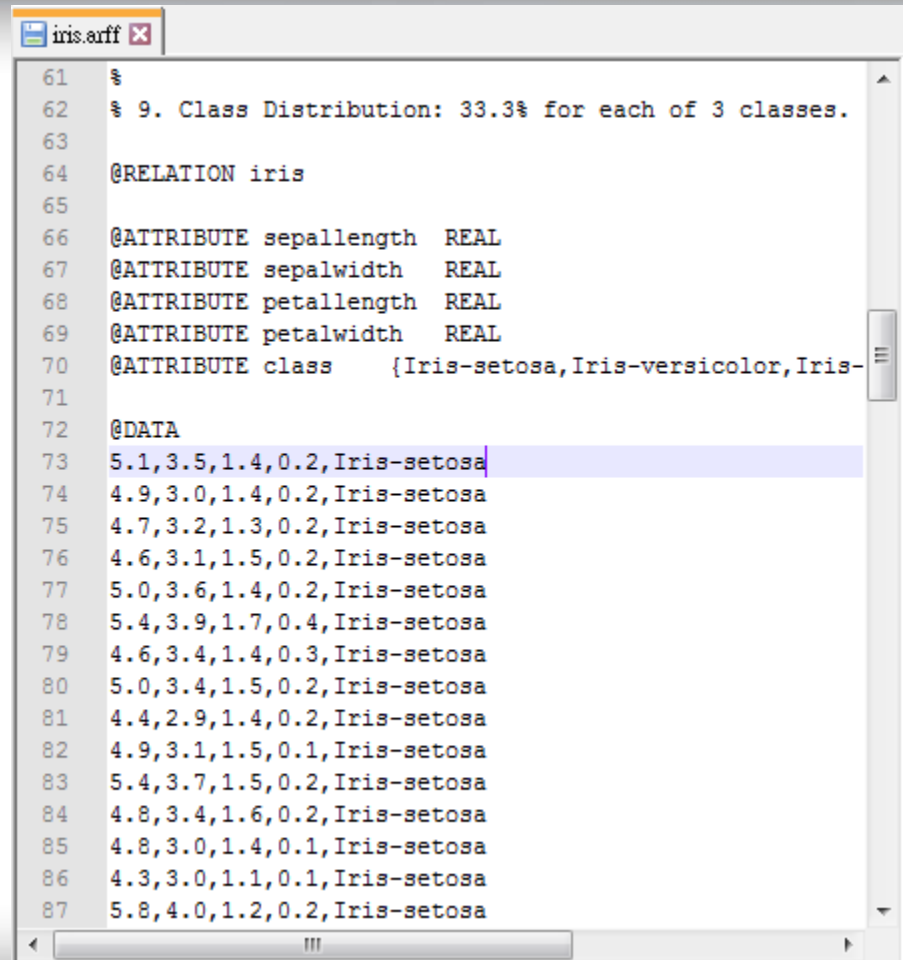


# 單一檔案格式

## ARFF

### Attribute-Relation File Format 屬性-關聯檔案格式

1. 開頭註解：說明檔案內容，以%開頭
2. 檔案標題：以@RELATION開頭
3. 屬性定義：以@ATTRIBUTE開頭，定義屬性的資料類型
4. 資料案例：位於@DATA之後，一行一個案例




```
61 %  
62 % 9. Class Distribution: 33.3% for each of 3 classes.  
63  
64 @RELATION iris  
65  
66 @ATTRIBUTE sepallength REAL  
67 @ATTRIBUTE sepalwidth REAL  
68 @ATTRIBUTE petallength REAL  
69 @ATTRIBUTE petalwidth REAL  
70 @ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-  
71  
72 @DATA  
73 5.1,3.5,1.4,0.2,Iris-setosa  
74 4.9,3.0,1.4,0.2,Iris-setosa  
75 4.7,3.2,1.3,0.2,Iris-setosa  
76 4.6,3.1,1.5,0.2,Iris-setosa  
77 5.0,3.6,1.4,0.2,Iris-setosa  
78 5.4,3.9,1.7,0.4,Iris-setosa  
79 4.6,3.4,1.4,0.3,Iris-setosa  
80 5.0,3.4,1.5,0.2,Iris-setosa  
81 4.4,2.9,1.4,0.2,Iris-setosa  
82 4.9,3.1,1.5,0.1,Iris-setosa  
83 5.4,3.7,1.5,0.2,Iris-setosa  
84 4.8,3.4,1.6,0.2,Iris-setosa  
85 4.8,3.0,1.4,0.1,Iris-setosa  
86 4.3,3.0,1.1,0.1,Iris-setosa  
87 5.8,4.0,1.2,0.2,Iris-setosa
```

# 單一檔案格式 CSV

## Comma-Separated Values 逗點分割數值

1. 屬性：以**逗點**區隔的每個欄位
2. 屬性標題：第一行為屬性標題，以逗點區隔每個欄位
3. 資料案例：第二列之後的**每一行**為一案例
4. 資料類型：需由程式自動判斷



```
iris.csv - Notepad++
檔案(F) 編輯(E) 搜尋(S) 檢視(V) 編碼(N) 語言(L) 設定(I) 巨
iris.csv x
1 sepallength,sepalwidth,
2 5.1,3.5,1.4,0.2,Iris-se
3 4.9,3,1.4,0.2,Iris-setosa
4 4.7,3.2,1.3,0.2,Iris-setosa
5 4.6,3.1,1.5,0.2,Iris-setosa
6 5,3.6,1.4,0.2,Iris-setosa
7 5.4,3.9,1.7,0.4,Iris-setosa
8 4.6,3.4,1.4,0.3,Iris-setosa
9 5,3.4,1.5,0.2,Iris-setosa
10 4.4,2.9,1.4,0.2,Iris-setosa
11 4.9,3.1,1.5,0.1,Iris-setosa
12 5.4,3.7,1.5,0.2,Iris-setosa
13 4.8,3.4,1.6,0.2,Iris-setosa
14 4.8,3,1.4,0.1,Iris-setosa
15 4.3,3,1.1,0.1,Iris-setosa
16 5.8,4,1.2,0.2,Iris-setosa
17 5.7,4.4,1.5,0.4,Iris-setosa
Normal text file length: 4,611 lines: 152 Ln: 1 Col: 52 Sel: 0|0
```



# 單一檔案格式 ODS



OpenDocument Spreadsheet

## 開放文件試算表

1. 屬性：每一直欄為一屬性
2. 屬性標題：第一橫列為屬性標題
3. 資料案例：第二列之後的每一列為一案例
4. 資料類型：依細格資料類型設定
5. 主要使用LibreOffice編輯

iris.ods (read-only) - LibreOffice Calc

	A	B	C	D	E	F
1	sepalwidth	petalwidth	class			
2	5.1	3.5	1.4	0.2	Iris-setosa	
3	4.9	3	1.4	0.2	Iris-setosa	
4	4.7	3.2	1.3	0.2	Iris-setosa	
5	4.6	3.1	1.5	0.2	Iris-setosa	
6	5	3.6	1.4	0.2	Iris-setosa	
7	5.4	3.9	1.7	0.4	Iris-setosa	
8	4.6	3.4	1.4	0.3	Iris-setosa	
9	5	3.4	1.5	0.2	Iris-setosa	
10	4.4	2.9	1.4	0.2	Iris-setosa	
11	4.9	3.1	1.5	0.1	Iris-setosa	
12	5.4	3.7	1.5	0.2	Iris-setosa	
13	4.8	3.4	1.6	0.2	Iris-setosa	
14	4.8	3	1.4	0.1	Iris-setosa	
15	4.3	3	1.1	0.1	Iris-setosa	
16	5.8	4	1.2	0.2	Iris-setosa	
17	5.7	4.4	1.5	0.4	Iris-setosa	
18	5.4	3.9	1.3	0.4	Iris-setosa	
19	5.1	3.5	1.4	0.3	Iris-setosa	
20	5.2	2.8	1.7	0.2	Iris-setosa	

# 相關詞彙定義

## 案例、屬性 (1/2)

### 案例 (Instance)

抽樣對象、個案、  
觀察值個體

### 屬性 (Attribute)

特徵、變項、觀察值

案例1



案例2



屬性

- 名字：豪快綠
- 攻擊力：9
- 防禦力：5

屬性

- 名字：猛牛紫
- 攻擊力：3
- 防禦力：12

# 相關詞彙定義

## 案例、屬性 (2/2)

屬性

屬性標題

名字	攻擊力	防禦力
豪快綠	9	5
猛牛紫	3	12

案例





# Attribute Type

## 屬性的資料類型

資料類型分類		舉例
主要資料類型	Nominal 類別型	● male ● 臺南
	Numeric 數值型	● 1 ● 0.75
特殊資料類型	String 字串型 (文字型)	● This is a pen ● 這是一隻筆
	Boolean 布林值 (是或否)	● t ● f (建議以類別型取代)
缺失資料	Missing Value 缺失值/未知值	?



Part 3.

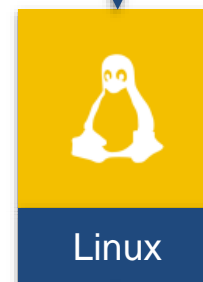
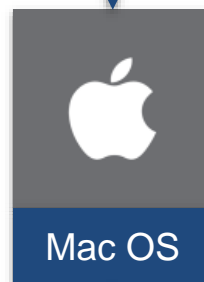
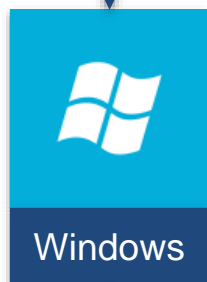
# 準備Weka

## 下載、安裝與設定

# Weka的下載



請下載  
includes Java VM  
版本  
(有分64 bit  
或32 bit)



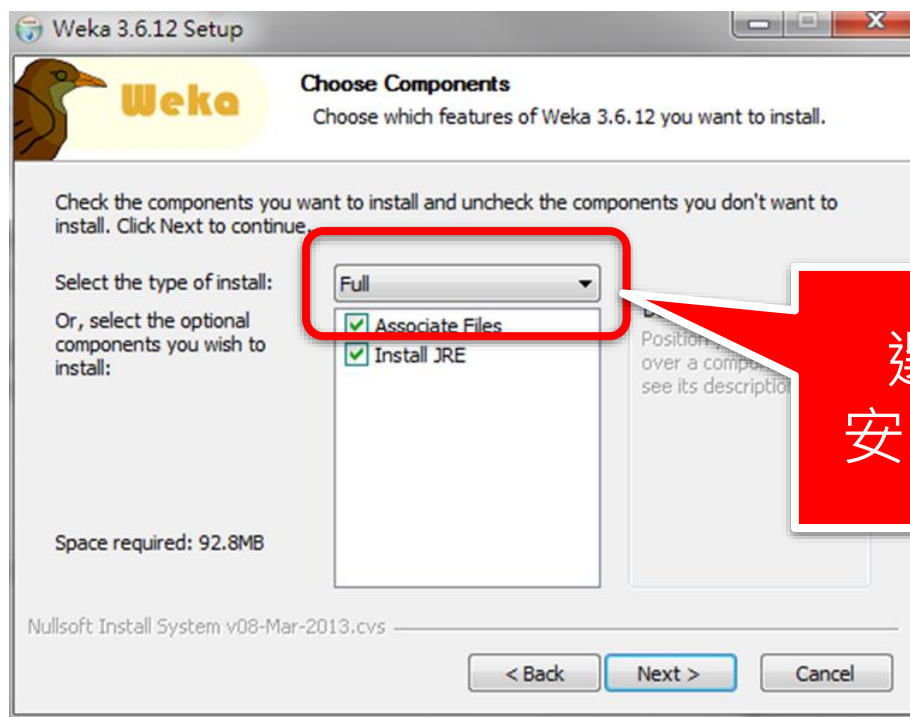
所需環境：



Java VM

# Weka的安裝

安裝精靈，容易上手

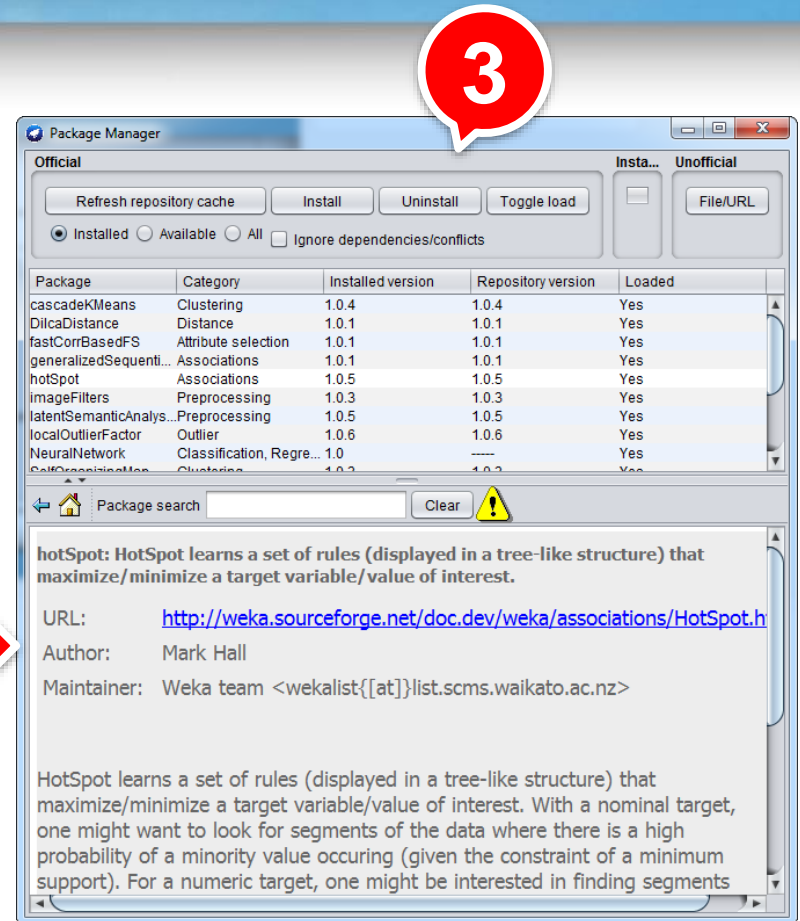
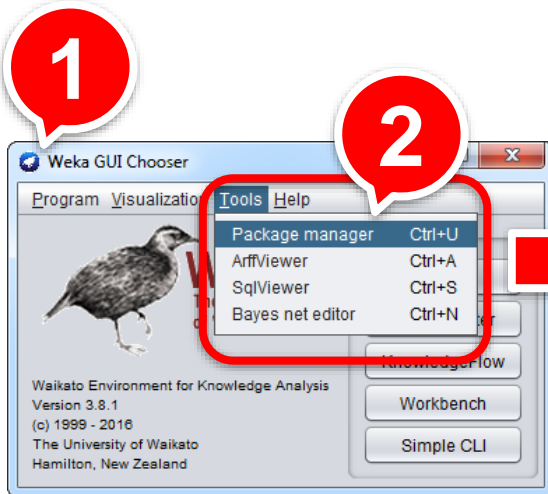


選「Full」  
安裝所有資料



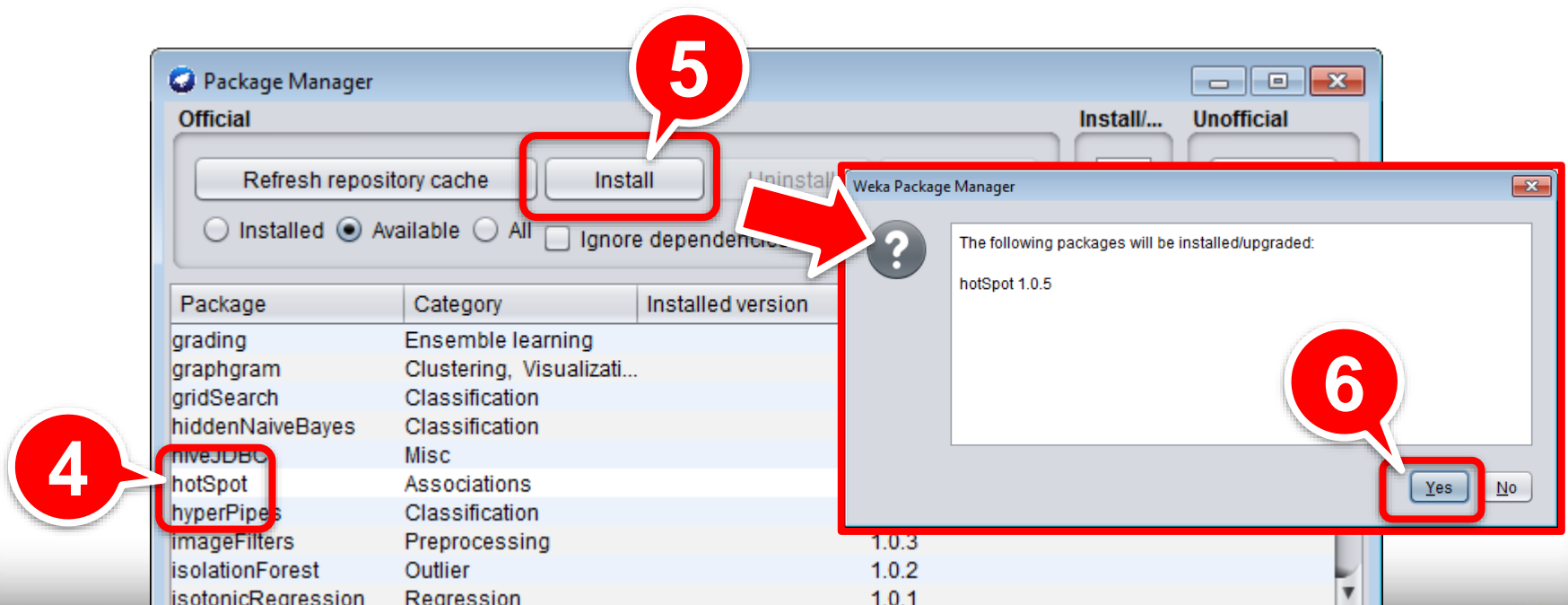
# STEP 1. 安裝官方套件 (1/3)

1. 開啟Weka
2. Tools ⇒ Package Manager
3. Package Manager主視窗



# STEP 1. 安裝官方套件 (2/3)

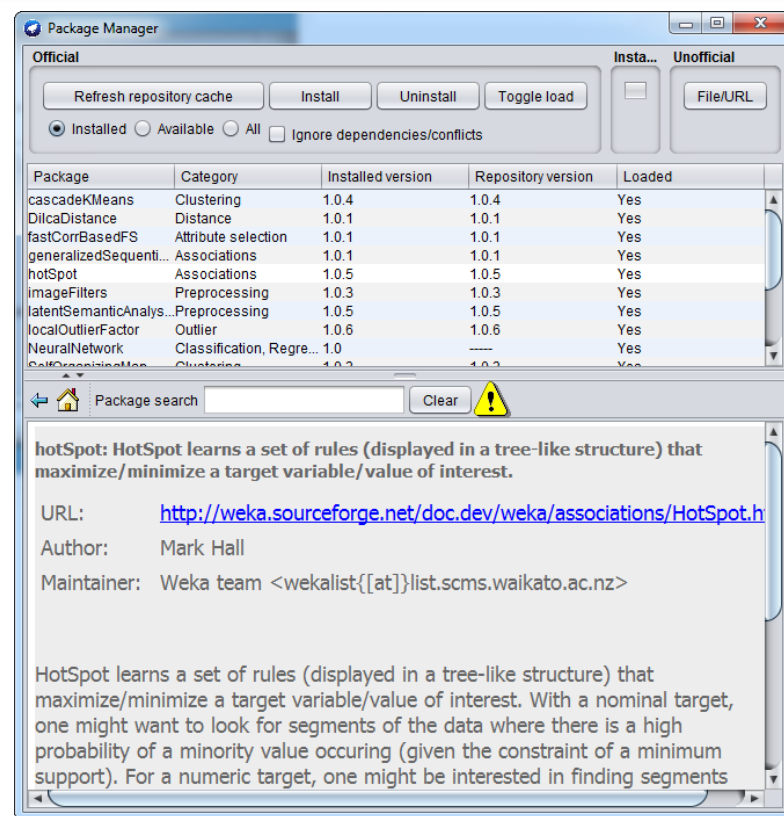
4. 找到要安裝的套件，例如 **cascadeKMeans**
5. **Install**
6. 確認安裝，**Yes**



# STEP 1. 安裝官方套件 (3/3)

請按照以上步驟，安裝以下套件吧：

- **cascadeKMeans**  
分群演算法
- **hotSpot**  
關聯規則探勘演算法
- **localOutlierFactor**  
異常偵測演算法



# STEP 2. 安裝自製套件 (1/2)

## WekaODF

1. 在Blog上，下載WekaODF1.0.5.zip檔案

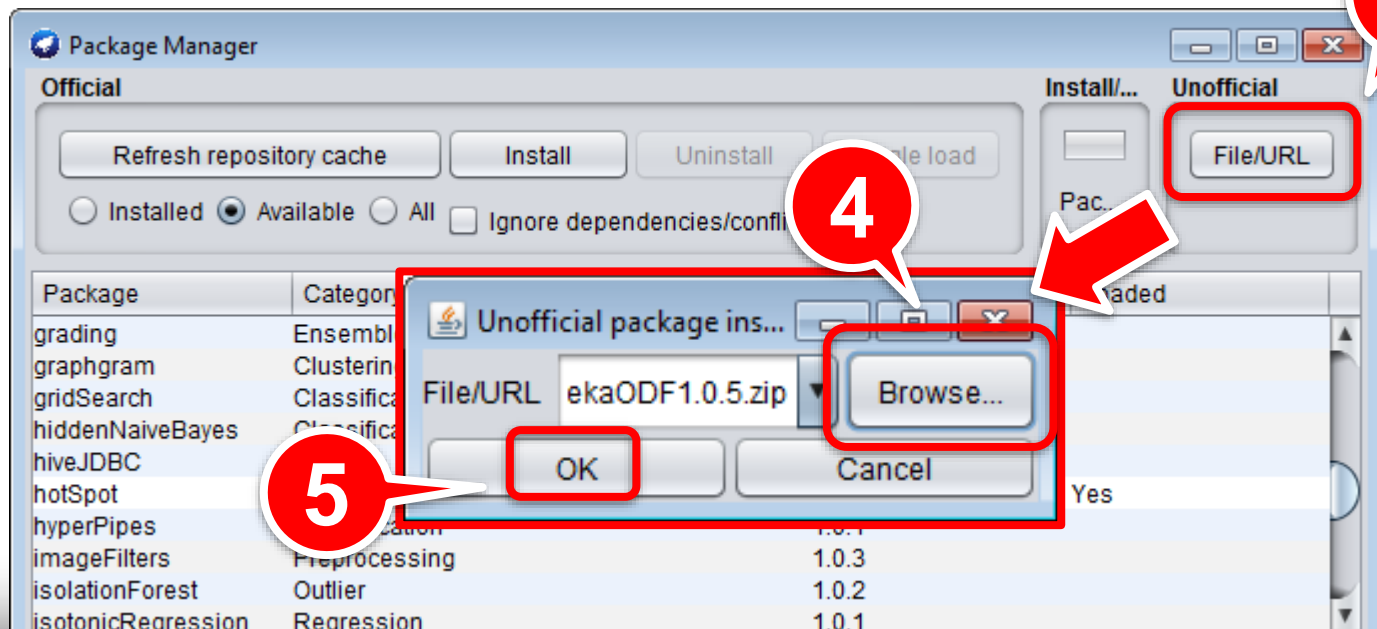




# STEP 2. 安裝自製套件 (2/2)

## WekaODF

2. 開啟Package Manager主視窗
3. 在右上角Unofficial按下File/URL
4. 按Browse，選擇剛剛下載的WekaODF1.0.5.zip檔案
5. 確認安裝，Yes



# STEP 3. 關閉Weka，再重新啟動



因為安裝了套件  
請重新啟動Weka吧

# 安裝套件



上機啦！

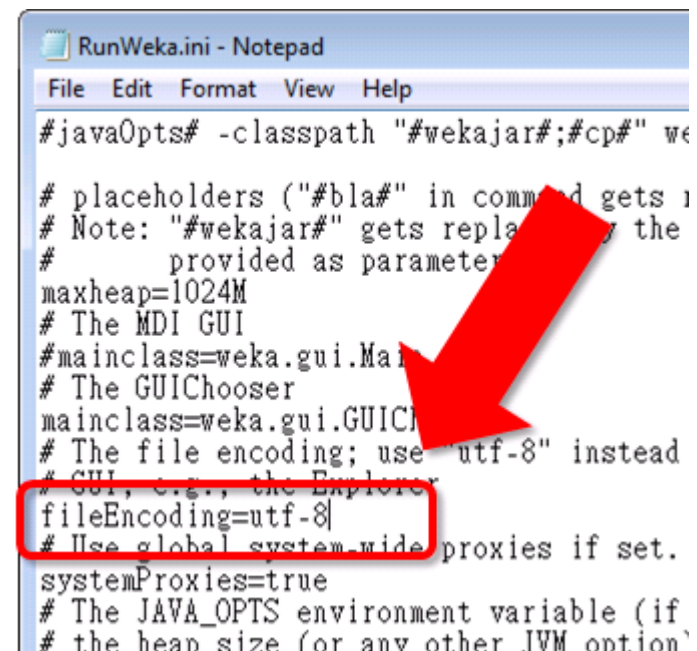
1. 安裝官方套件
  - cascadeKMeans  
分群演算法
  - localOutlierFactor  
異常偵測演算法
2. 安裝自製套件
  - WekaODF
3. 關閉Weka，再重新啟動



# (只有Windows作業系統需要設定) 讓Weka能夠讀取中文

1. 開啟Weka安裝目錄，預設為  
C:\Program Files\Weka-[版本號]
2. 用文字編輯器開啟RunWeka.ini
3. 將以下設定  
fileEncoding=**Cp1252**  
改成  
fileEncoding=**utf-8**
4. 儲存，重新啟動Weka

詳細操作請看[Blog: 如何在Weka中顯示中文](http://blog.pulipuli.info/2017/06/wekautf8-how-to-process-chinese-data-in.html)



```
RunWeka.ini - Notepad
File Edit Format View Help
#javaOpts# -classpath "#wekajar#;#cp#" we
# placeholders ("#bla#" in command gets r
# Note: "#wekajar#" gets replaced by the
# provided as parameter
maxheap=1024M
# The MDI GUI
#mainclass=weka.gui.Ma
# The GUIChooser
mainclass=weka.gui.GUICh
# The file encoding; use "utf-8" instead
# GUI, e.g., the Explorer
fileEncoding=utf-8
# Use global system-wide proxies if set.
systemProxies=true
# The JAVA_OPTS environment variable (if
# the heap size (or any other JVM option)
```

# LibreOffice Calc下載

## LibreOffice

- LibreOffice辦公室套裝軟體的試算表工具
- LibreOffice是跨平臺的開放自由軟體，是編輯開放文件格式(ODF)的最佳選擇
- 開放文件格式包含文件(ODT)、**試算表(ODS)**、投影片(ODP)等多種類型格式
- 開放文件格式是我國政府的主要通用格式



<https://zh-tw.libreoffice.org/download/libreoffice-fresh/>





Part 4.

# 認識Weka架構

# Weka的功能架構



探索器

實驗器

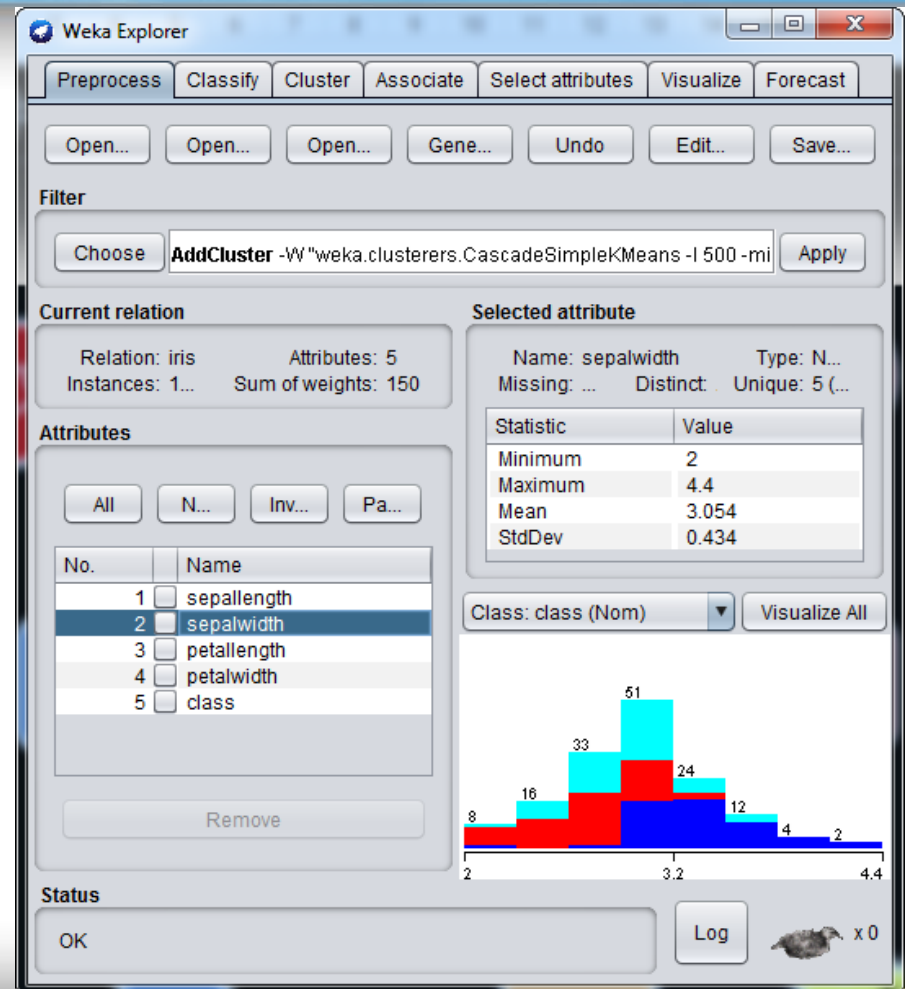
知識流

命令列

# Explorer 探索器

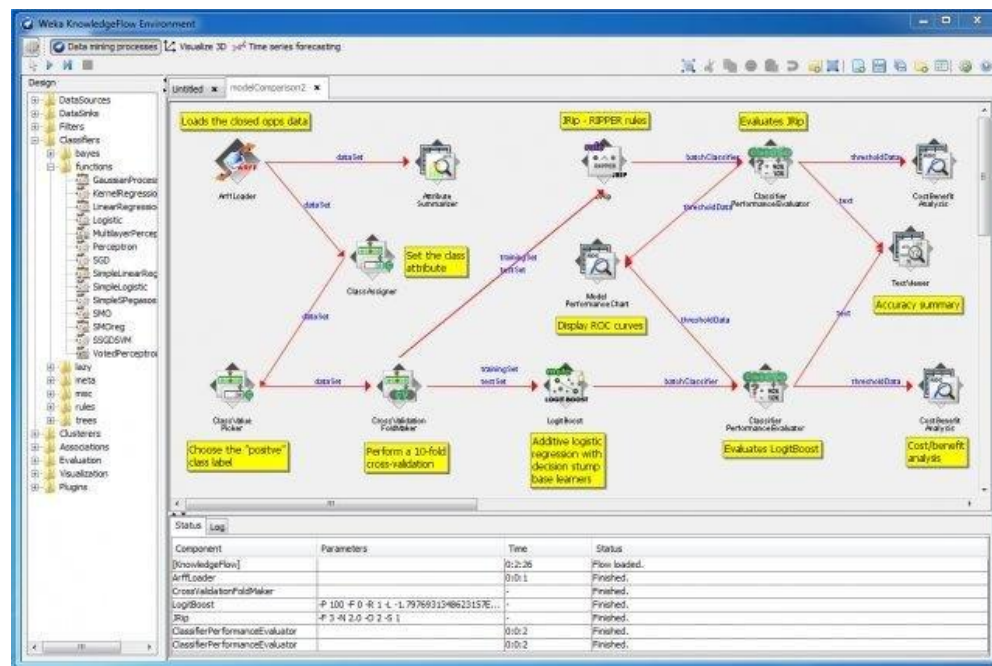


- Weka主要的圖形化使用者介面
- 以頁籤、下拉式選單、欄位設定等表單元件，讓使用者輕易進行資料分析與探勘
- 直接提供各種視覺化圖表，展現分析結果
- 一次只能分析一份資料集



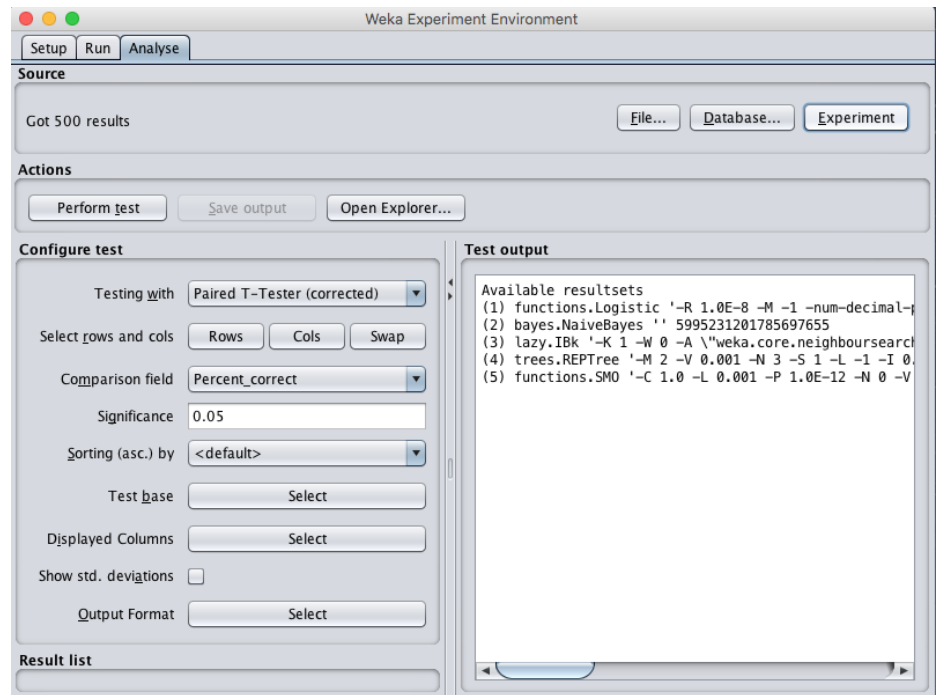
# Knowledge Flow 知識流

- 以資料流的形式，定義資料探勘中的所有步驟
- 不僅支援單一檔案，還能支援分析批次、增量的串流資料
- 類似商業智慧 (Business Intelligence, BI)



# Experimenter 實驗器

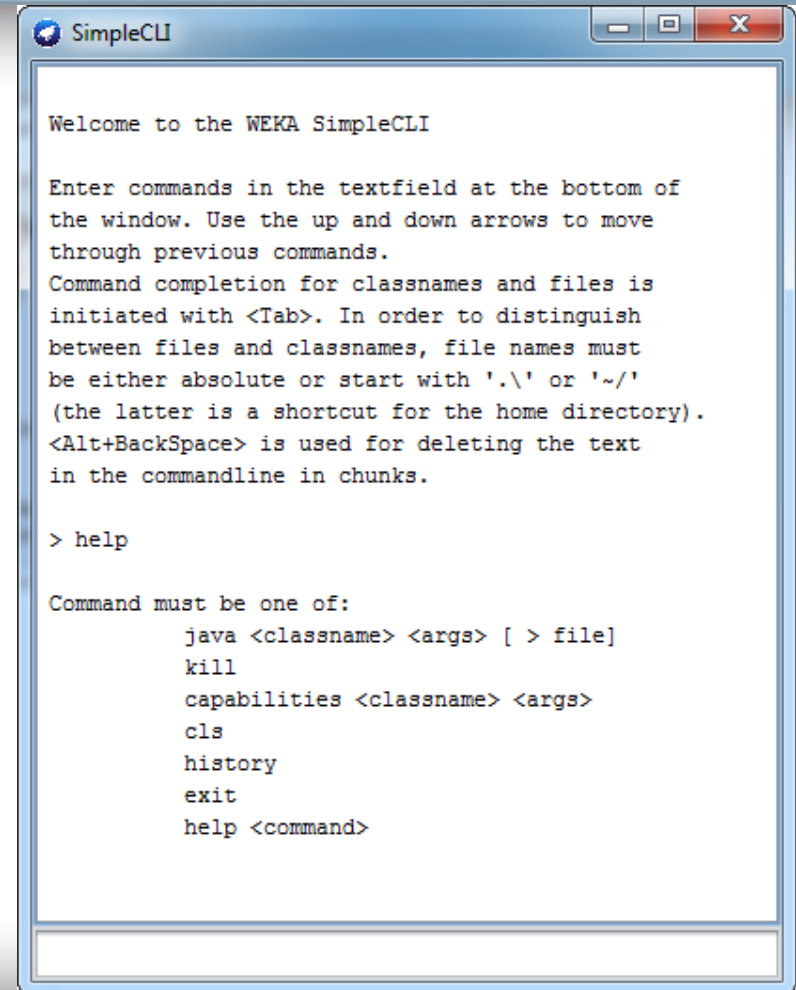
- 適合處理多份資料、多種不同的資料探勘演算法、多種不同的參數設定的情況使用
- 特別適合資料庫存取、多電腦的分散式運算
- 設定後可批次且自動執行，使用者僅需等待分析結果即可





# Simple CLI 命令列

- 能夠使用完整進階指令，突破圖形化使用者界面的限制
- 記憶體消耗較少
- 可先從探索器擬定命令參數，再以命令列批次執行



```
SimpleCLI

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or '~/ '
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
    java <classname> <args> [ > file]
    kill
    capabilities <classname> <args>
    cls
    history
    exit
    help <command>
```

# 本課程的Weka流程

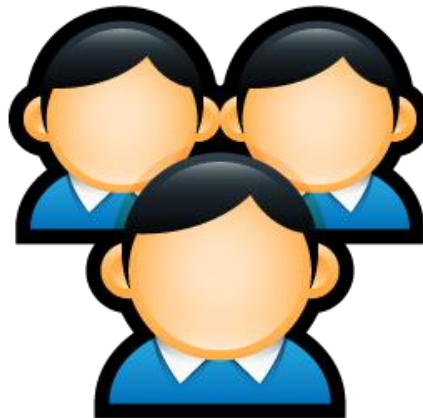
1. 案例說明
2. 演算法簡介
3. 實作
  - a. 取得資料集
  - b. 使用Weka的探索器進行分析
4. 檢視分析結果



# 實作資料集

## 學生成績資料集 (1/3)

- 學生成績資料集是Cortez等人(2008)年從兩所葡萄牙學校蒐集649位學生、33種屬性的開放資料集
- 屬性包括學生個人資料、家庭狀況、就學狀況、學校生活、課堂表現



※ 本教學取其資料集內容，因應教學內容而作調整

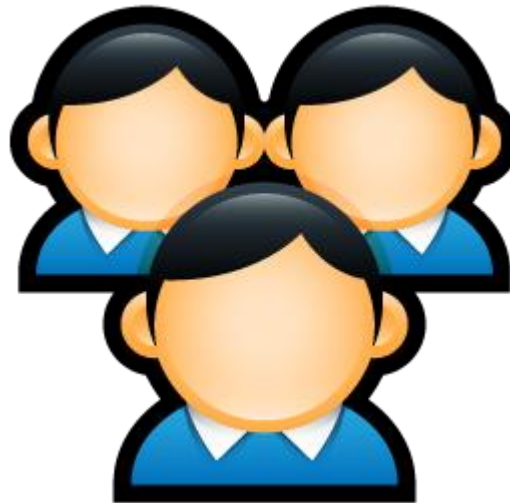
# 實作資料集

## 學生成績資料集 (2/3)

### Nominal Type 類別型屬性

共15種

- 性別
- 就學理由
- 是否補習
- 學校



### Numeric Type 數值型屬性

共18種

- 年齡
- 雙親教育程度
- 缺席次數
- 課堂成績

※ 完整的屬性說明請看論文

# 實作資料集

## 學生成績資料集 (3/3)

Untitled 1 - LibreOffice Calc

File Home Insert Layout Data Review View Tools

General

Home

A1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Gender	Age	Address	FamiSize	ParentsStat	MonEdu	FatEdu	MonJob	FatJob	ChoSchRea	Guardian	TravelTime	StudyTime	ClassFailure	schoolsuc
2	female	15	urban	>3	together		4	4 teacher	services	course	mother	1	3	0	no
3	female	16	urban	>3	together		4	2 services	other	course	mother	1	2	0	no
4	male	16	urban	>3	together		3	3 services	other	home	mother	1	2	0	no
5	female	17	urban	>3	together		3	4 services	other	course	mother	1	3	0	no
6	female	16	urban	>3	together		2	1 other	other	course	mother	1	2	0	no
7	male	16	urban	>3	together		2	1 other	other	course	mother	3	1	0	no
8	male	15	urban	<=3	together		4	3 teacher	services	home	mother	1	3	0	no
9	male	15	urban	>3	together		4	2 other	other	course	mother	1	4	0	no
10	male	15	urban	>3	together		4	3 teacher	other	home	mother	1	2	0	no
11	female	16	urban	>3	together		4	3 health	other	home	mother	1	2	0	no
12	male	16	urban	>3	together		2	3 other	other	home	father	2	1	0	no
13	male	16	urban	<=3	together		1	1 other	other	home	mother	2	2	0	no
14	female	17	urban	>3	together		2	1 services	other	course	mother	2	2	0	no
15	male	15	urban	>3	together		4	4 services	services	reputation	mother	2	2	0	no
16	male	16	urban	>3	together		4	4 health	other	course	mother	1	1	0	no
17	male	18	urban	>3	together		4	2 teacher	other	home	mother	1	2	0	no
18	female	18	urban	>3	together		3	4 other	other	course	mother	1	1	0	no
19	male	15	urban	>3	together		4	2 teacher	other	home	mother	1	2	0	no
20	female	17	urban	<=3	together		4	2 health	other	reputation	mother	1	2	0	no

Sheet1 Pivot Table\_Sheet1\_3 Pivot Table\_Sheet1\_2 Pivot Table\_Sheet1\_1

Find Find All Formatted Display Match Case

Sheet 1 of 4 Default English (USA) Average: ; Sum: 0 100%



# 準備好了嗎？



## Chapter 2.

## 探索性與比較性分析



GO

