R 軟體統計分析應用(一)
- 大數據分析與R軟體基本操作

日期 2016 06/19

地點 校務研究辦公室

資訊科技的運用-決策的依據

數據

圖表

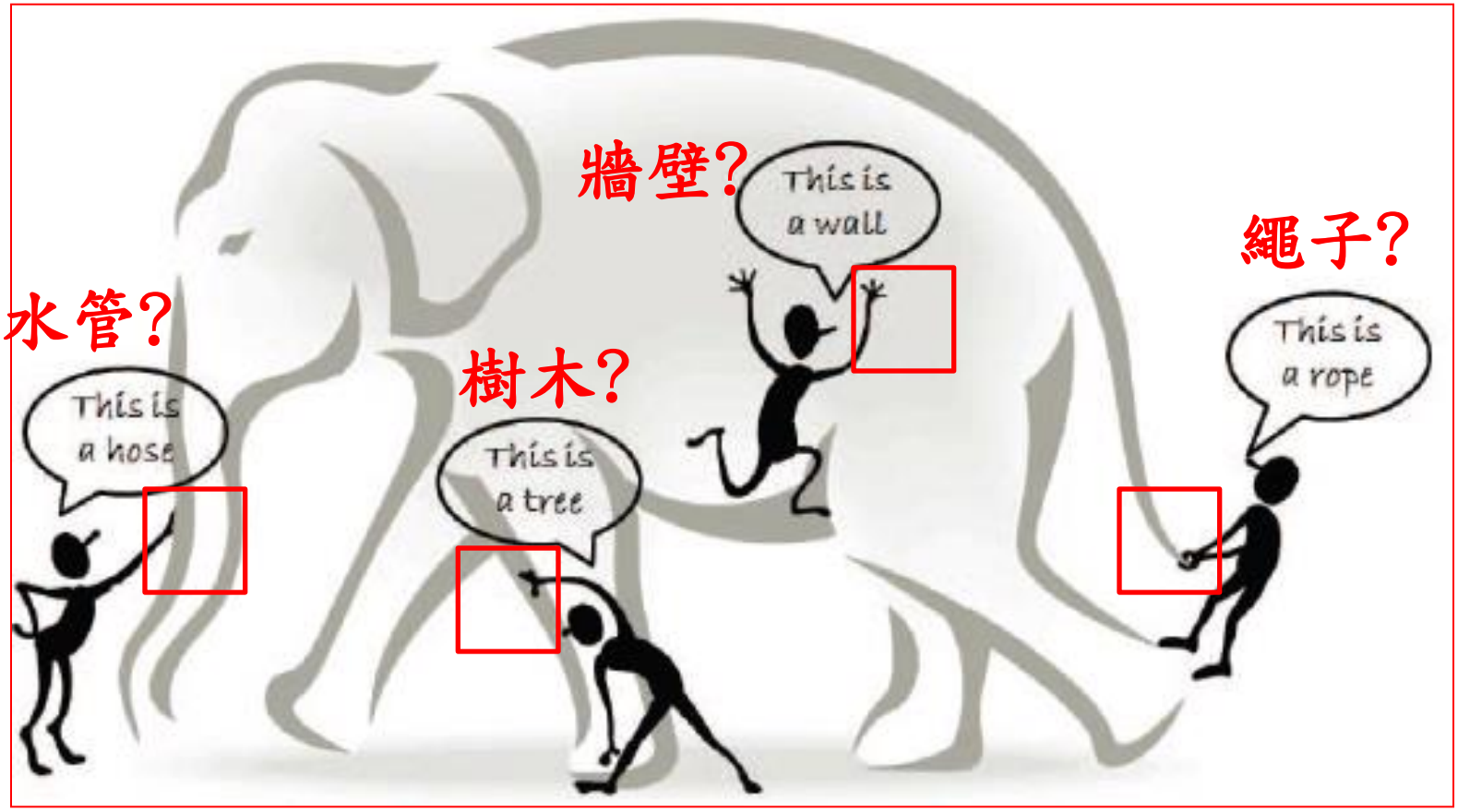
資料量大 ←

資料 資訊 知識 智慧 預測 決策

→ 重要性高

資訊科技的運用-決策與資料的關係

當資料量很大時 是否會容易做出錯誤決策呢



資訊科技的運用-資料的多寡

資訊科技的運用-資料的來源

傳統商業資料-CRM, ERP, 網站交易, 總分類帳, 醫療資料

- Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.

機器產生資料-通聯記錄, 網站記錄, 智慧電錶, 感測紀錄
半導體製程參數, 氣象變化資料

- Machine-generated /sensor data – includes Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data.

社交網站資料-Facebook, Twitter, Blog, Forum, Google

- Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook

資訊科技的運用-政府部門的開放資料 01

開放資料（Open Data）

議題1、「健康照護」應用議題（主辦機關：衛福部）

一、解決課題：

(一) 加值運用傳染病巨量數據，提供疫情警示資訊，以及早因應。

(二) 掌握我國癌症趨勢和預測，以利及早因應與管控。

(三) 加值運用健保巨量數據，評估主要慢性病(心臟病、中風、三高及慢性阻塞性肺病(COPD)等)發生率及未來發生趨勢，以利相關單位及早因應與管控。

資訊科技的運用-政府部門的開放資料 02

開放資料（Open Data）

議題 2、「毒藥品防制」應用議題（主辦機關：衛福部）

一、 **解決課題：**

進行當前毒品情勢分析及藥物濫用情勢分析，作為毒品/藥物濫用防制政策擬定參考。

資訊科技的運用-政府部門的開放資料 03

開放資料（Open Data）

議題 3、「穩健財政收支」應用議題（主辦機關：財政部）

一、 **解決課題：**

推估我國財政收支發展情形，強化預算收支控管，檢討收支結構，縮減財政赤字差短，以減少債務之舉借數額，建立財政收入與支出間、及財政收支與經濟成長間之連動關係，俾達成財政穩健目標，為各項政務推動及經濟發展注入活水。

資訊科技的運用-政府部門的開放資料 04

開放資料（Open Data）

議題 4、「自然環境保護」應用議題（主辦機關：環保署）

一、 **解決課題：**

掌握自然生活環境品質變化(氣候、水質、地質、空氣)及建立預測模型，以利及早因應降低其對民眾、生態之影響或作為環評之參據。

資訊科技的運用-政府部門的開放資料 05

開放資料（Open Data）

議題 5、「災害預警」應用議題（主辦機關：科技部）

一、 解決課題：

整合社群網路情資分析，強化現有複合性災害預警監測及疏散撤離之能量，以利進行「超前部署、預置兵力、隨時防救」，保障人民生命財產安全。

資訊科技的運用-Facebook 巨量資料

資訊科技的運用-巨量資料的特點

- Volume. Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. With more than 25,000 airline flights per day, the daily volume of just this single data source runs into the Petabytes. Smart meters and heavy industrial equipment like oil refineries and drilling rigs generate similar data volumes, compounding the problem.

Volume(巨量)

- Velocity. Social media data streams – while not as massive as machine-generated data – produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day). **Velocity(快速)**
- Variety. Traditional data formats tend to be relatively well defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. As new services are added, new sensors deployed, or new marketing campaigns executed, new data types are needed to capture the resultant information.

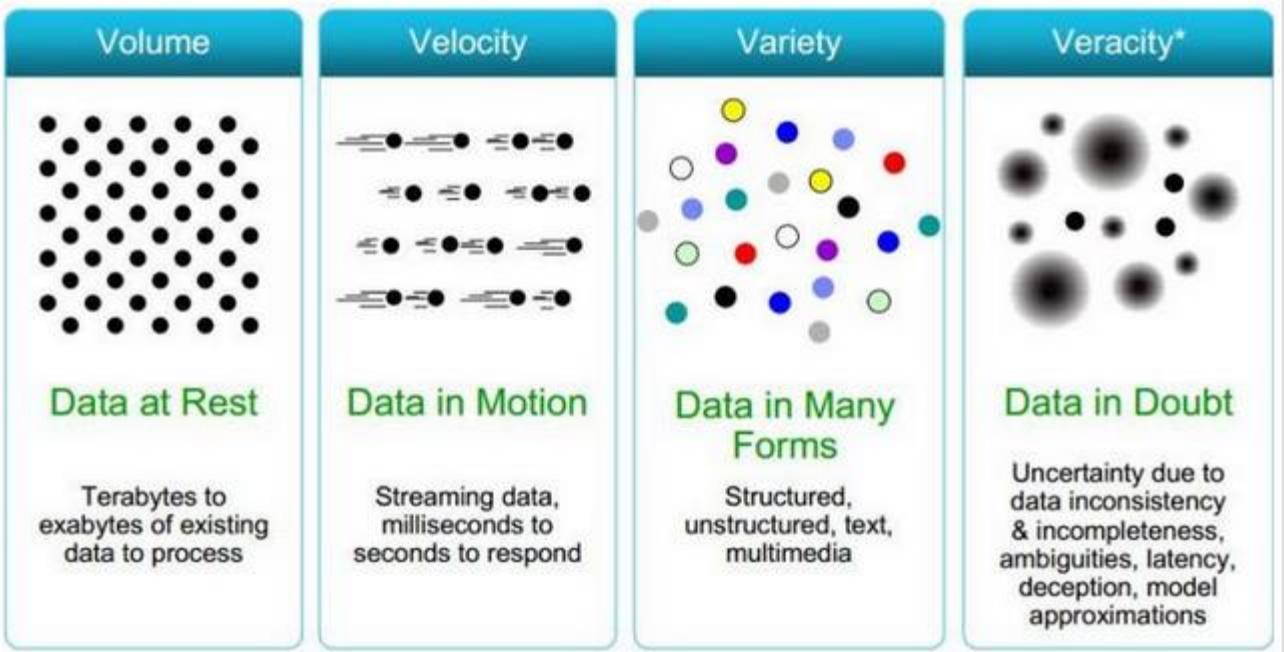
Variety(多樣)

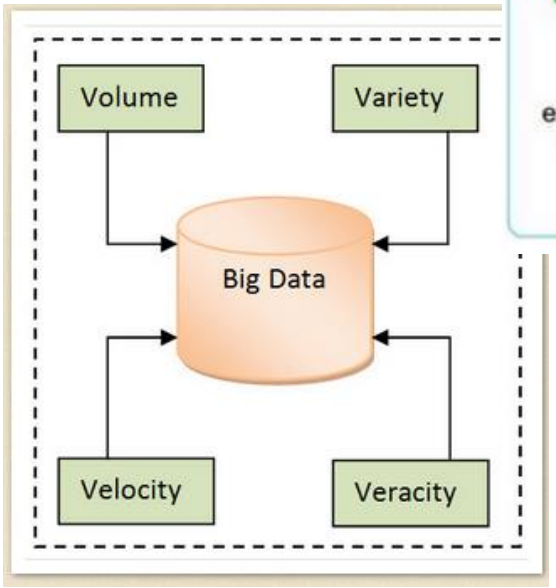
資訊科技的運用-巨量資料的價值

Value(價值)

- Value. The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data; the challenge is identifying what is valuable and then transforming and extracting that data for analysis.

資訊科技的運用-巨量資料的真實性

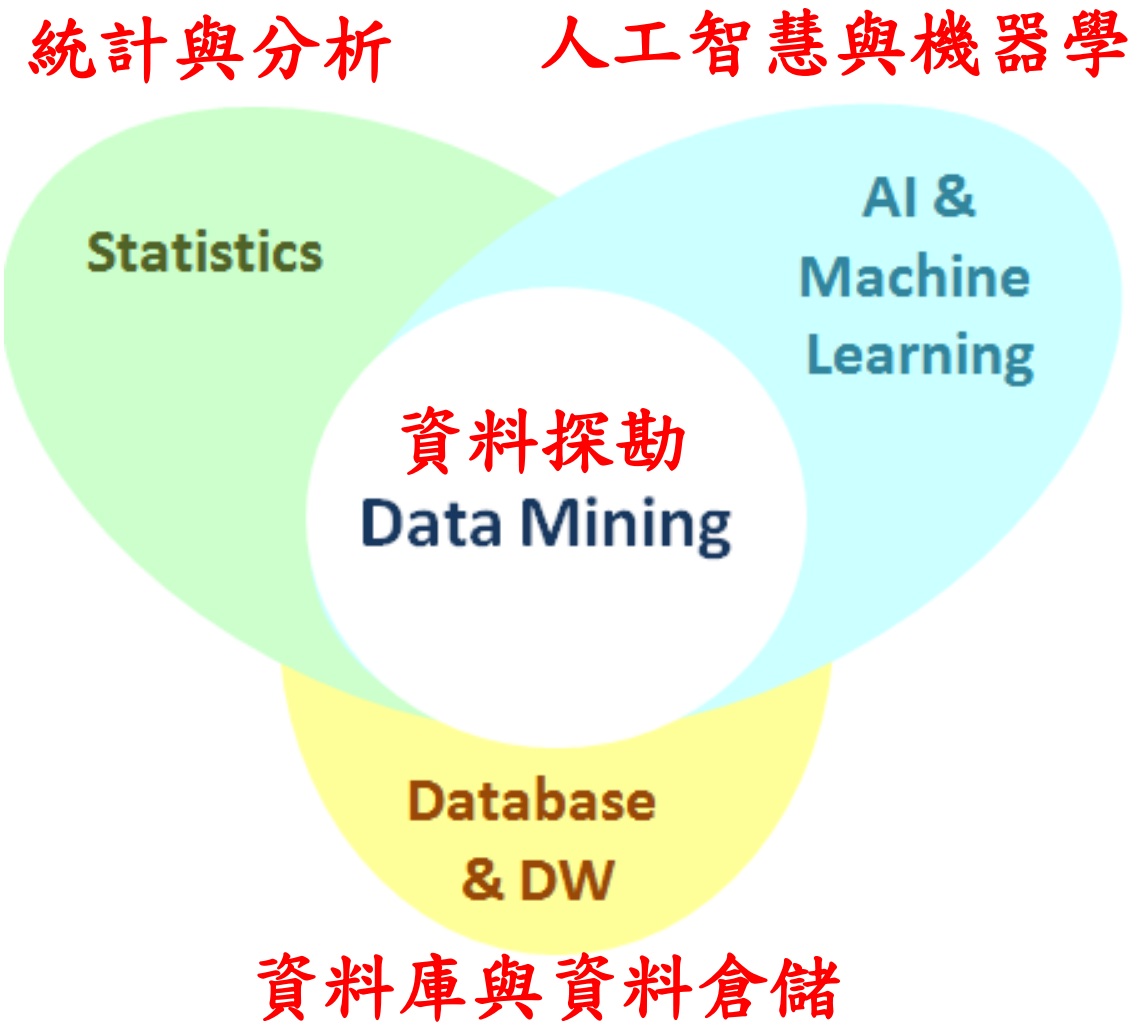




『volume（大量性）』、『variety（多樣性）』、

『velocity（速度性）』、『veracity（真實性）』

資料處理領域涵蓋的技術-Data Mining



資料處理領域涵蓋的技術-Data Science

統計與分析

機器學習(依賴電腦)
最佳化技術

溝通理解與
說故事能力

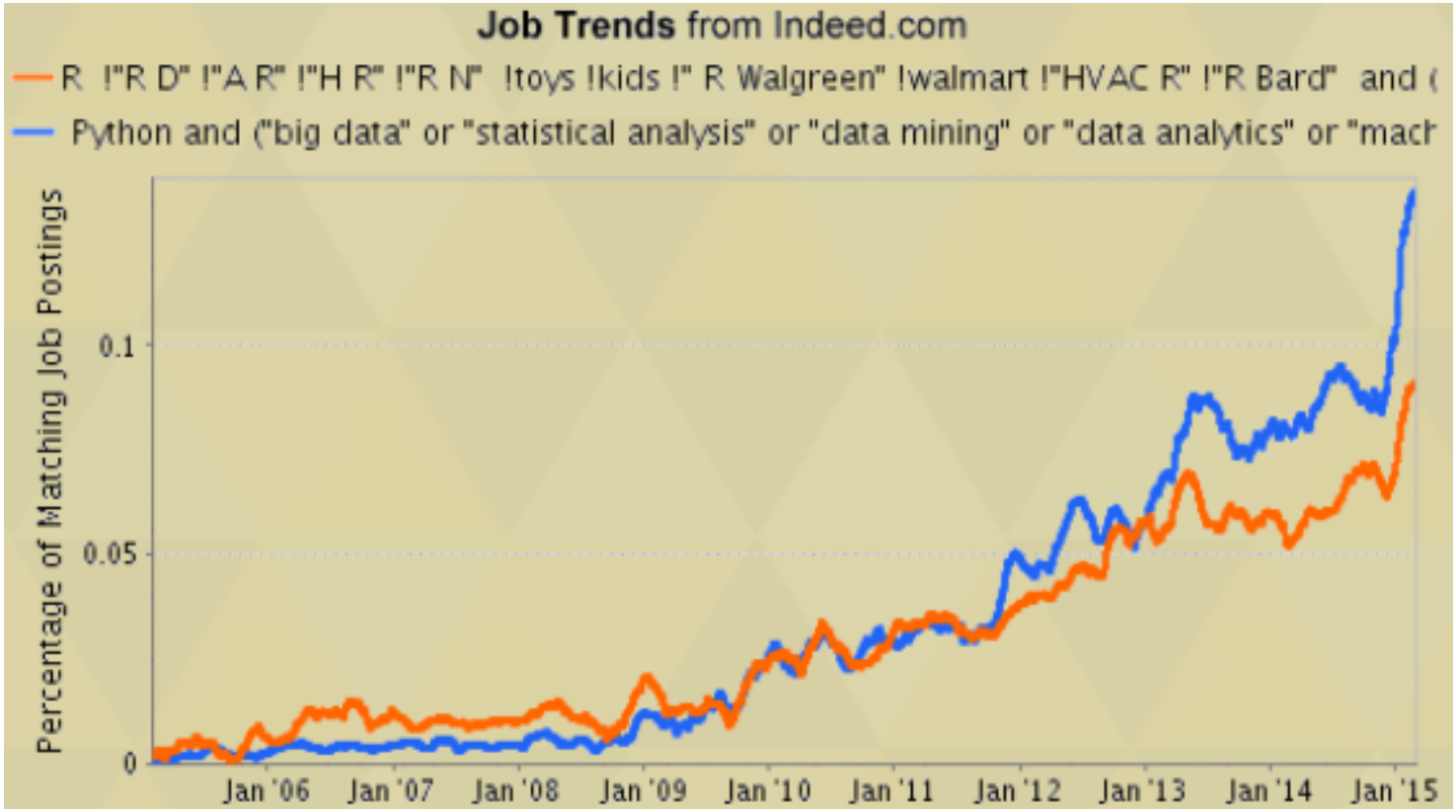
程式設計
演算法(依賴人腦)
資料結構

巨量資料
雲端運算

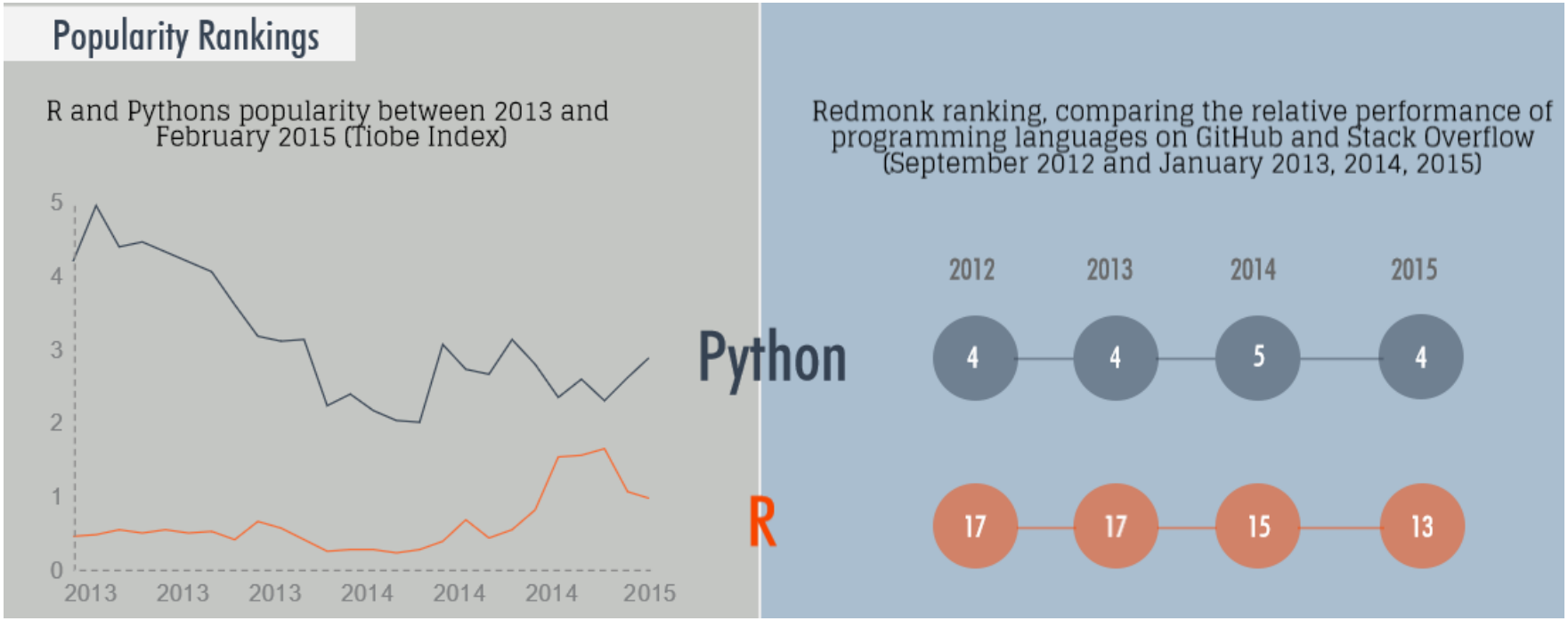
視覺化
圖表工具

專業領域
背景知識

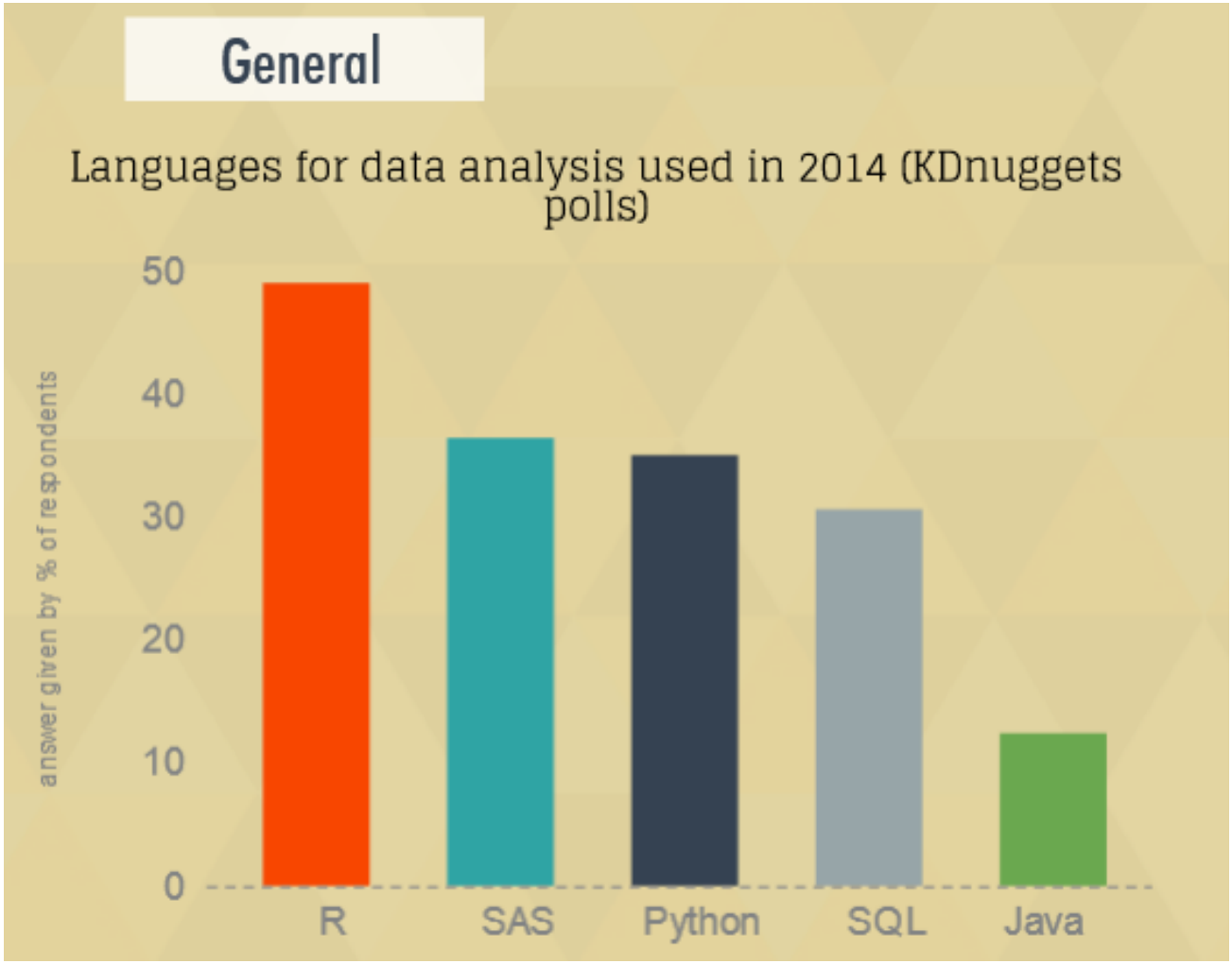
資料處理領域涵蓋的技術- R & Python 01



資料處理領域涵蓋的技術- R & Python 02

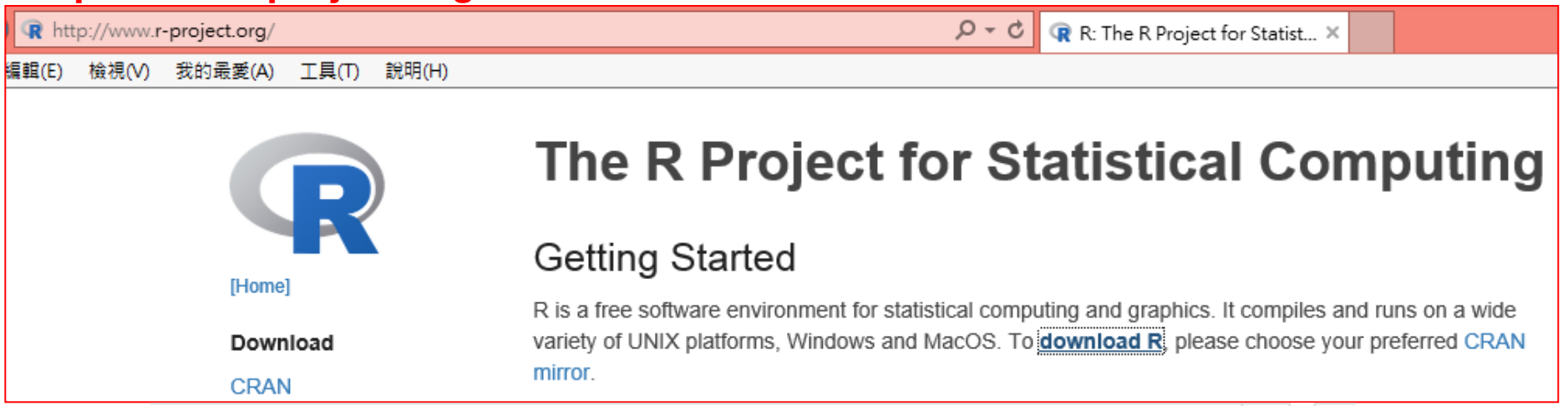


資料處理領域涵蓋的技術- R & Python 03



R 軟體資料分析工具

http://www.r-project.org



R 軟體分析工具-鳶尾花 01

RGui (64-bit)

檔案 編輯 視窗

R Console

```
> data()
```

R 軟體內建範例資料集

R data sets

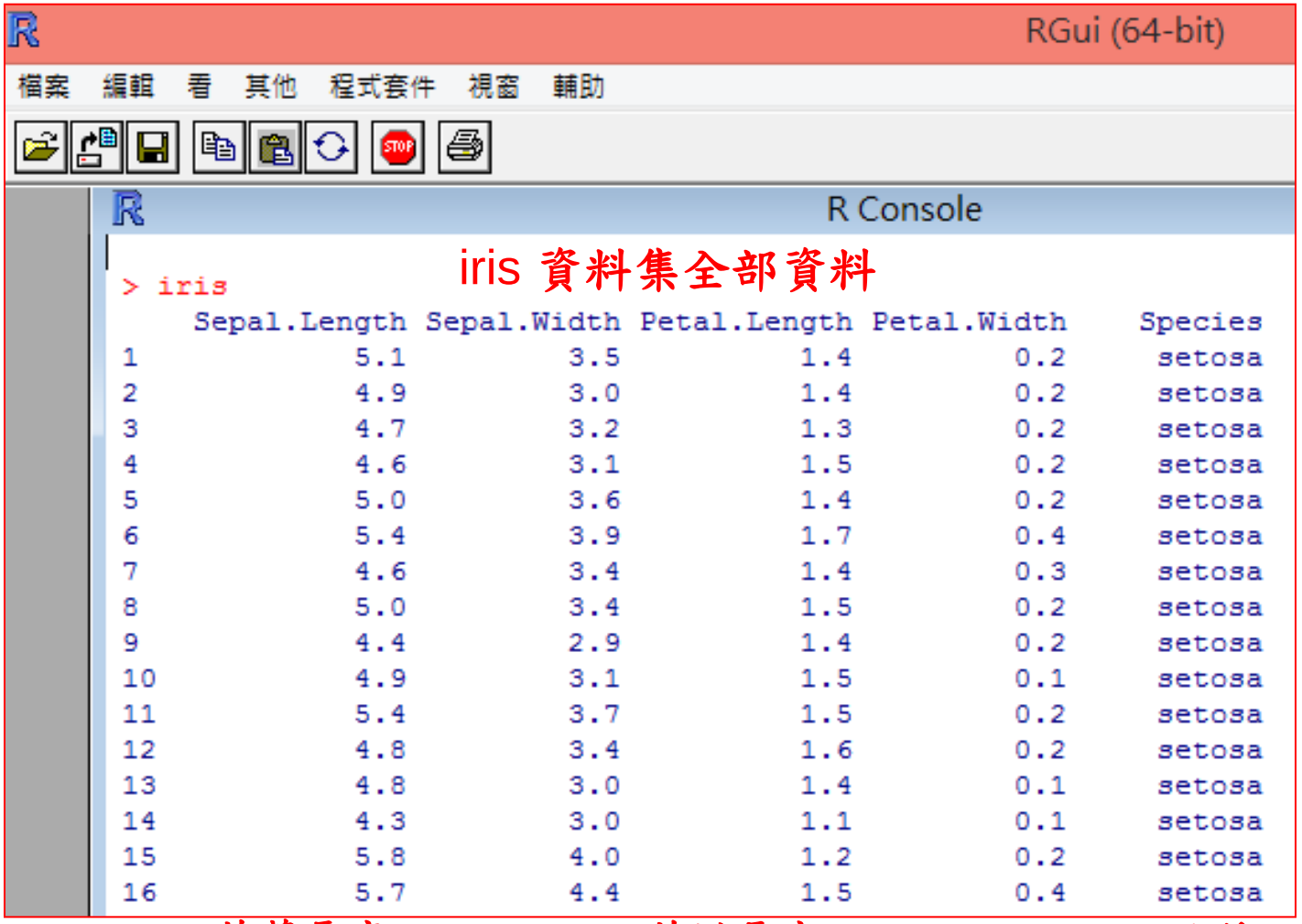
```
beaver1 (beavers)      Body Temperature S
beaver2 (beavers)      Body Temperature S
cars                   Speed and Stopping
chickwts               Chicken Weights by
co2                    Mauna Loa Atmosphe
crimtab
discover
esoph
euro                   Conversion Rates o
euro.cross (euro)      Conversion Rates o
eurodist               Distances Between European Cities and
                       Between US Cities
faithful               Old Faithful Geyser Data
fdeaths (UKLungDeaths)
                       Monthly Deaths from Lung Diseases in the
freeny.y (freeny)
infert
iris                   Edgar Anderson's Iris Data
iris3                  Edgar Anderson's Iris Data
islands                Areas of the World's Major Landmasses
ldeaths (UKLungDeaths)
lh                     Luteinizing Hormone in Blood Samples
longley                Longley's Economic Regression Data
lynx                   Annual Canadian Lynx trappings 1821-1934
mdeaths (UKLungDeaths) Monthly Deaths from Lung Diseases in the UK
morley                 Michelson Speed of Light Data
```

co2 二氧化碳濃度

冒納羅亞火山

iris 鳶尾花

R 軟體分析工具-鳶尾花 02



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa

花萼長度　　花瓣長度　　品種

花萼寬度　　花瓣寬度

R 軟體分析工具-鳶尾花 03

```
> head(iris,5)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
> tail(iris,5)
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
> mean(Sepal.Length)
[1] 5.843333
> var(Sepal.Length)
[1] 0.6856935
> sd(Sepal.Length)
[1] 0.8280661
>
```

iris 資料集前面 5 筆資料

iris 資料集最後 5 筆資料

花萼長度平均值

花萼長度變異數

花萼長度標準差

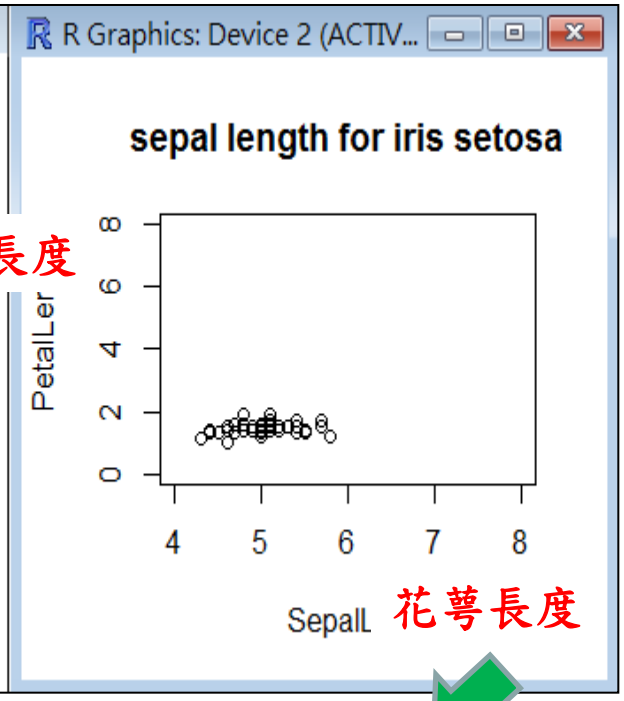
attach(iris)

R 軟體分析工具-鳶尾花 04

R 軟體分析工具-鳶尾花 05

R Console

```
> attach(iris)
The following objects are masked from iris (pos = 3):

    Petal.Length, Petal.Width, Sepal.Length, Sepal.Width, Species

The following objects are masked from iris (pos = 4):

    Petal.Length, Petal.Width, Sepal.Length, Sepal.Width, Species

The following objects are masked from iris (pos = 5):

    Petal.Length, Petal.Width, Sepal.Length, Sepal.Width, Species

> plot(Sepal.Length[Species=="setosa"], Petal.Length[Species=="setosa"],
+ pch=1, col="black", xlim=c(4,8), ylim=c(0,8), main="sepal length for iris setosa",
+ xlab="SepalLen", ylab="PetalLen")
>
>
>
```



➢ plot(Sepal.Length[Species=="setosa"], Petal.Length[Species=="setosa"],

+ pch=1, col="black", xlim=c(4,8), ylim=c(0,8), main="sepal length for iris setosa",

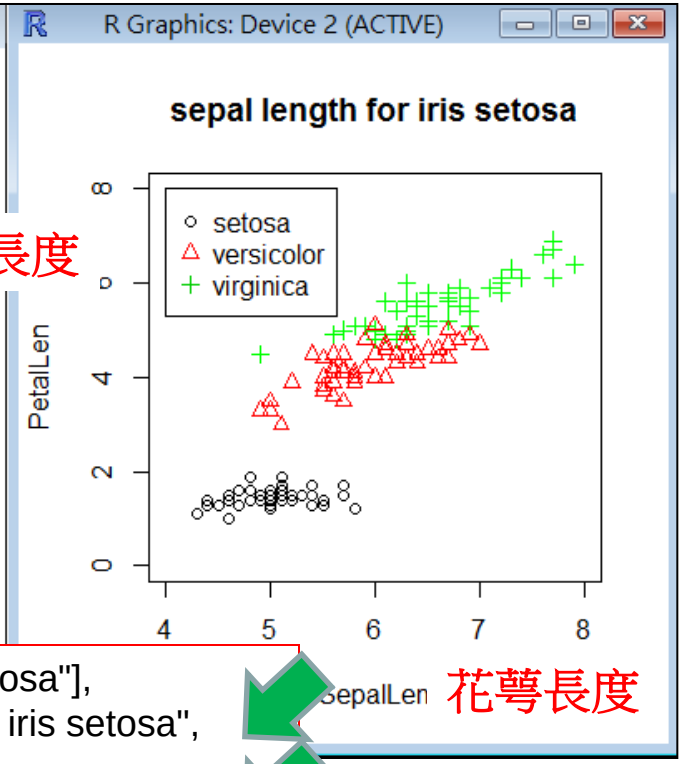
+ xlab="SepalLen", ylab="PetalLen")

R 軟體分析工具-鳶尾花 06

R Console

```
> plot(Sepal.Length[Species=="setosa"], Petal.Length[Species=="setosa"],
+ pch=1, col="black", xlim=c(4,8), ylim=c(0,8), main="sepal length for iris setosa",
+ xlab="SepalLen", ylab="PetalLen")
>
> points(Sepal.Length[Species=="virginica"],Petal.Length[Species== "virginica"]
+ ,pch=3,col="green")
>
> points(Sepal.Length[Species=="versicolor"],Petal.Length[Species== "versicolor"],
+ ,pch=2,col="red")
>
> legend(4,8,legend=c("setosa","versicolor","virginica"),col=c(1,2,3), pch=c(1,2,3))
```

花瓣長度

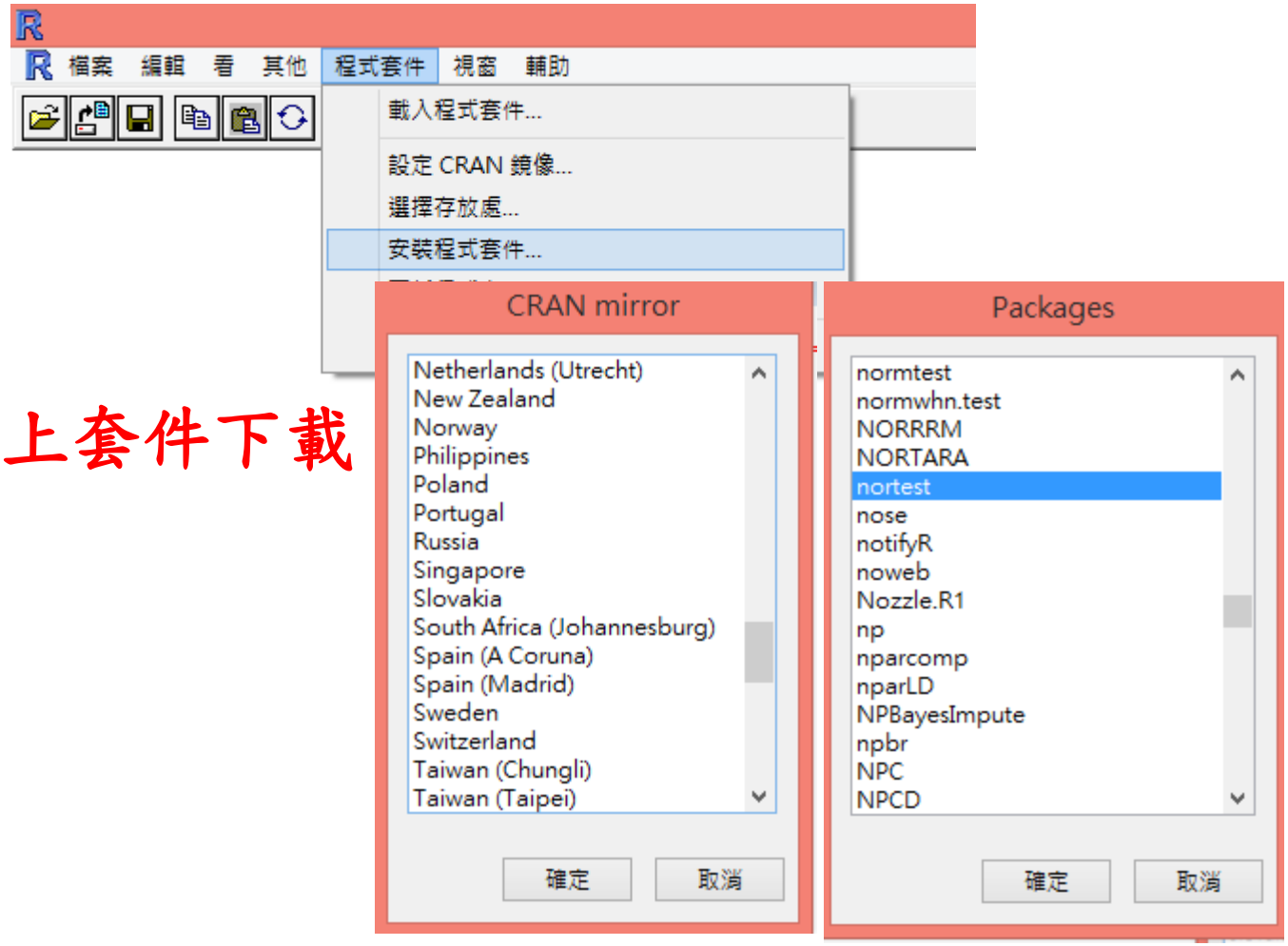


花萼長度

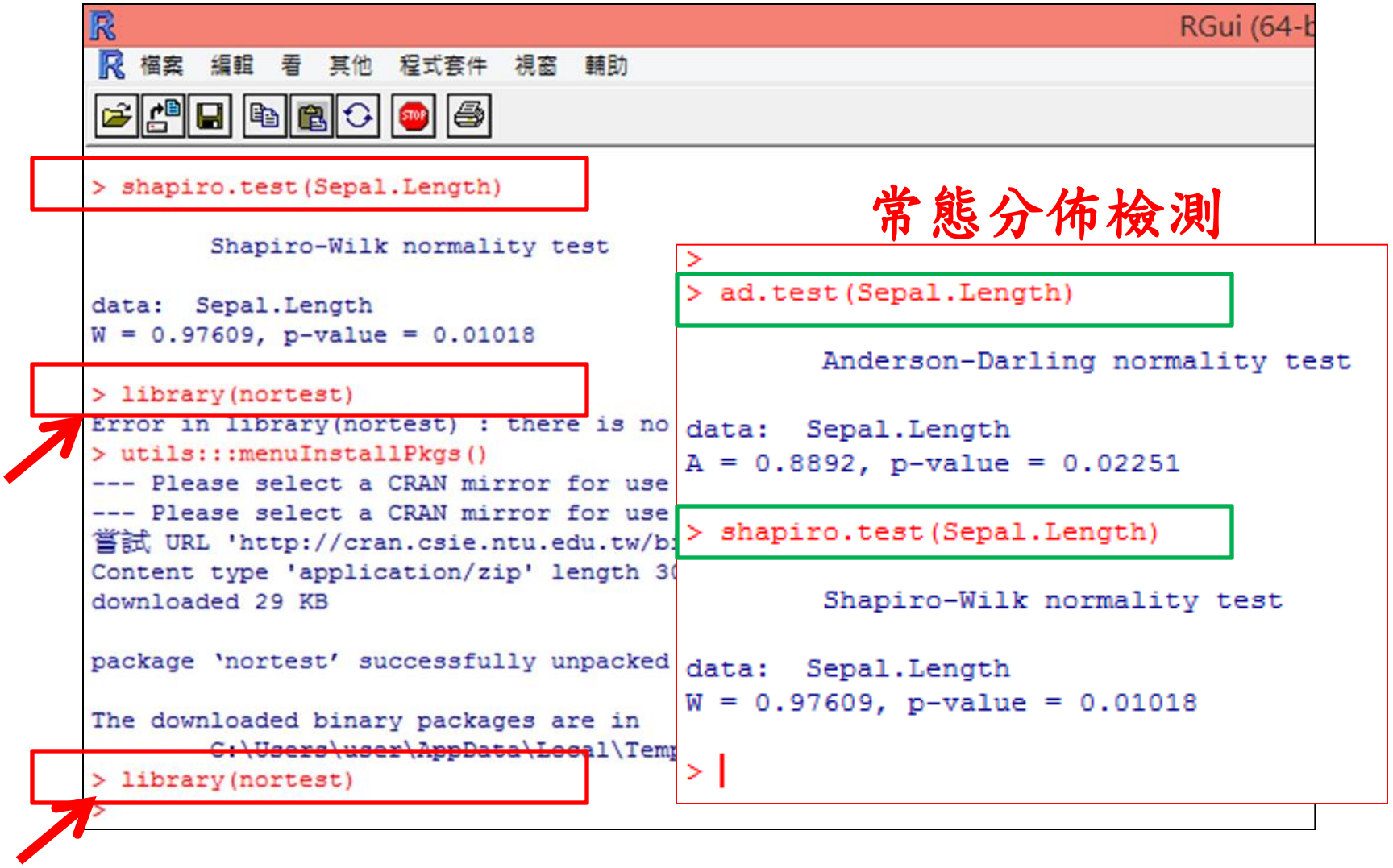
> plot(Sepal.Length[Species=="setosa"], Petal.Length[Species=="setosa"],
+ pch=1, col="black", xlim=c(4,8), ylim=c(0,8), main="sepal length for iris setosa",
+ xlab="SepalLen", ylab="PetalLen")
> points(Sepal.Length[Species=="virginica"],Petal.Length[Species== "virginica"],
+ ,pch=3,col="green")
> points(Sepal.Length[Species=="versicolor"],Petal.Length[Species== "versicolor"],
+ ,pch=2,col="red")
> legend(4,8,legend=c("setosa","versicolor","virginica"),col=c(1,2,3), pch=c(1,2,3))

R 軟體分析工具-鳶尾花 07

線上套件下載



R 軟體分析工具-鳶尾花 08



R 軟體分析工具-鐵達尼號 01

關聯規則分析(Association Rules)

```
> library(arules)
Error in library(arules) : there is no package called 'arules'
> utils:::menuInstallPkgs()
Warning in install.packages(NULL, .libPaths()[1L], dependencies = NA,$
  'lib = "C:/Program Files/R/R-3.2.0/library"' is not writable
嘗試 URL 'http://cran.csie.ntu.edu.tw/bin/windows/contrib/3.2/arules_$
Content type 'application/zip' length 1782667 bytes (1.7 MB)
downloaded 1.7 MB

package 'arules' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\user\AppData\Local\Temp\RtmpEVFT2Q\downloaded_packag$
>
> library(arules)
Loading required package: Matrix

Attaching package: 'arules'

The following objects are masked from 'package:base':

    %in%, write
```

載入關聯分析套件 arules

R 軟體分析工具-鐵達尼號 02

```
> load("c:/users/user/downloads/titanic.raw.rdata")
> dim(titanic.raw)
[1] 2201    4
> aa <- sample(1:nrow(titanic.raw), 16)
> titanic.raw[aa,]
     Class    Sex   Age Survived
709    3rd   Male Adult       No
2105   2nd Female Adult      Yes
433    3rd   Male Adult       No
729   Crew   Male Adult       No
802   Crew   Male Adult       No
723   Crew   Male Adult       No
63     1st   Male Adult       No
1569   1st   Male Adult      Yes
1661   3rd   Male Adult      Yes
1699  Crew   Male Adult      Yes
672    3rd   Male Adult       No
484    3rd   Male Adult       No
2013   1st Female Adult      Yes
1702  Crew   Male Adult      Yes
1051  Crew   Male Adult       No
1417   3rd Female Adult       No
```

匯入外部資料集
4欄位 共2201筆
顯示前面 16 筆

R 軟體分析工具-鐵達尼號 03

顯示titanic.raw資料集結構

```
> str(titanic.raw)
'data.frame':   2201 obs. of  4 variables:
 $ Class   : Factor w/ 4 levels "1st","2nd","3rd",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Sex     : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Age     : Factor w/ 2 levels "Adult","Child": 2 2 2 2 2 2 2 2 2 2 ...
 $ Survived: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

執行 apriori 演算法

```
> rulesa <- apriori(titanic.raw)

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE     0.1      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10 item(s), 2201 transaction(s)] done [0.00s].
sorting and recoding items ... [9 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.01s].
writing ... [27 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

R 軟體分析工具-鐵達尼號 04

顯示 apriori 演算法的執行結果

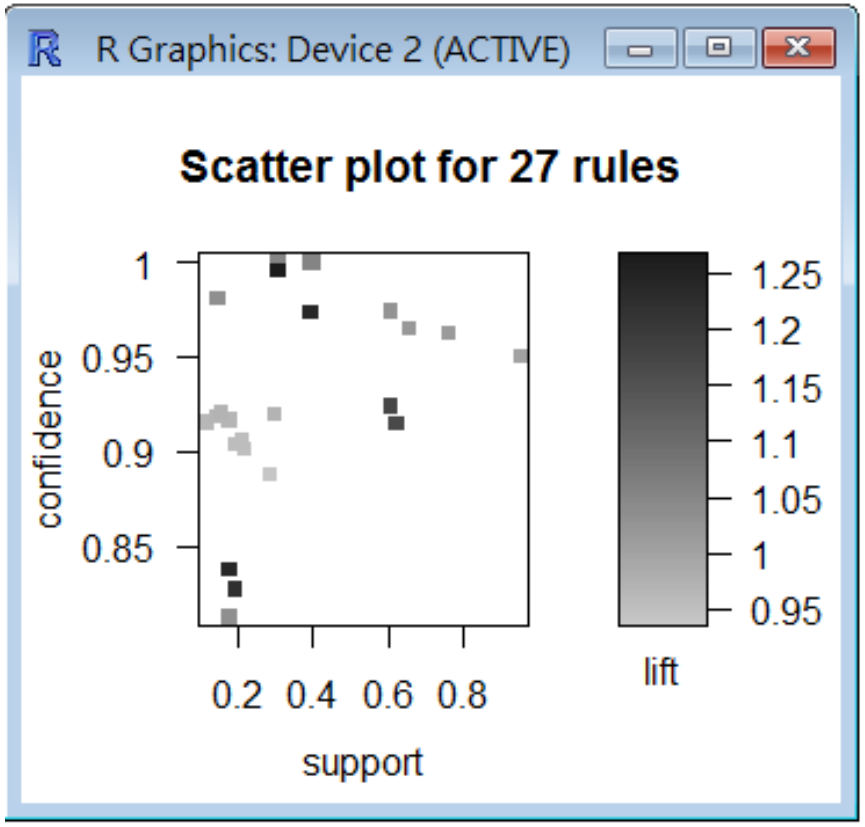
```
> inspect(rulesa)
    lhs                    rhs                support    confidence      lift
1   {}                  => {Age=Adult}    0.9504771  0.9504771 1.0000000
2   {Class=2nd}         => {Age=Adult}    0.1185825  0.9157895 0.9635051
3   {Class=1st}         => {Age=Adult}    0.1449341  0.9815385 1.0326798
4   {Sex=Female}        => {Age=Adult}    0.1930940  0.9042553 0.9513700
5   {Class=3rd}         => {Age=Adult}    0.2848705  0.8881020 0.9343750
6   {Survived=Yes}      => {Age=Adult}    0.2971377  0.9198312 0.9677574
7   {Class=Crew}        => {Sex=Male}     0.3916402  0.9740113 1.2384742
8   {Class=Crew}        => {Age=Adult}    0.4020900  1.0000000 1.0521033
9   {Survived=No}       => {Sex=Male}     0.6197183  0.9154362 1.1639949
10  {Survived=No}       => {Age=Adult}    0.6533394  0.9651007 1.0153856
11  {Sex=Male}          => {Age=Adult}    0.7573830  0.9630272 1.0132040
12  {Sex=Female,
     Survived=Yes}      => {Age=Adult}    0.1435711  0.9186047 0.9664669
13  {Class=3rd,
     Sex=Male}          => {Survived=No}  0.1917310  0.8274510 1.2222950
14  {Class=3rd,
     Survived=No}       => {Age=Adult}    0.2162653  0.9015152 0.9484870
15  {Class=3rd,
     Sex=Male}          => {Age=Adult}    0.2099046  0.9058824 0.9530818
16  {Sex=Male,
     Survived=Yes}      => {Age=Adult}    0.1535666  0.9209809 0.9689670
```

R 軟體分析工具-鐵達尼號 05

繪圖 apriori 演算法的執行結果

```
> plot(rulesa)
> summary(titanic.raw)
  Class          Sex          Age         Survived
 1st :325   Female: 470   Adult:2092   No :1490
 2nd :285   Male  :1731   Child: 109   Yes: 711
 3rd :706
 Crew:885
```

titanic.raw 資料集摘要資訊



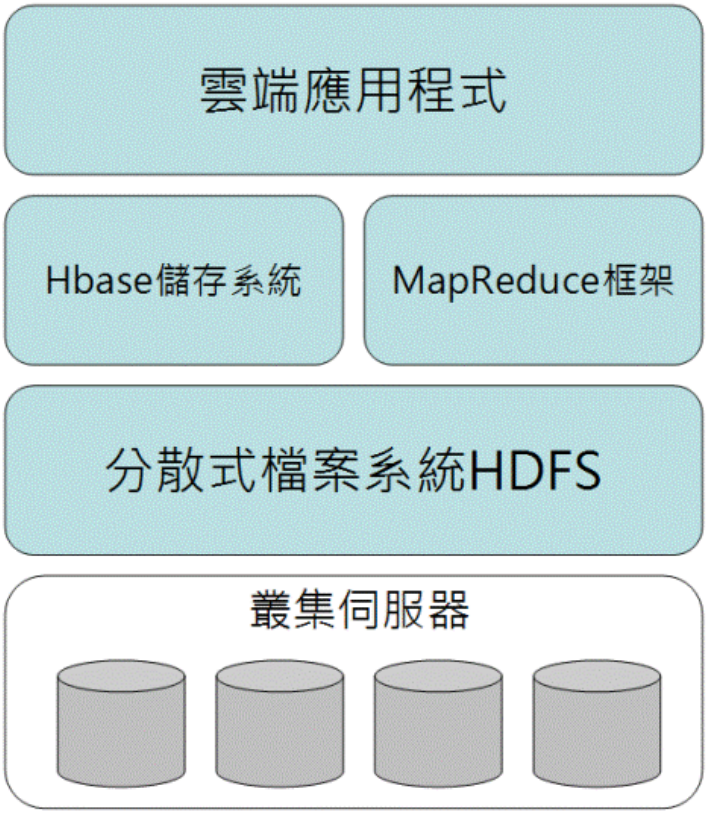
Support: 支持程度

Confidence: 信心程度

Lift: 增益

須線上下載 arulesViz 套件
並執行 library(arulesViz)

雲端分散式處理平台

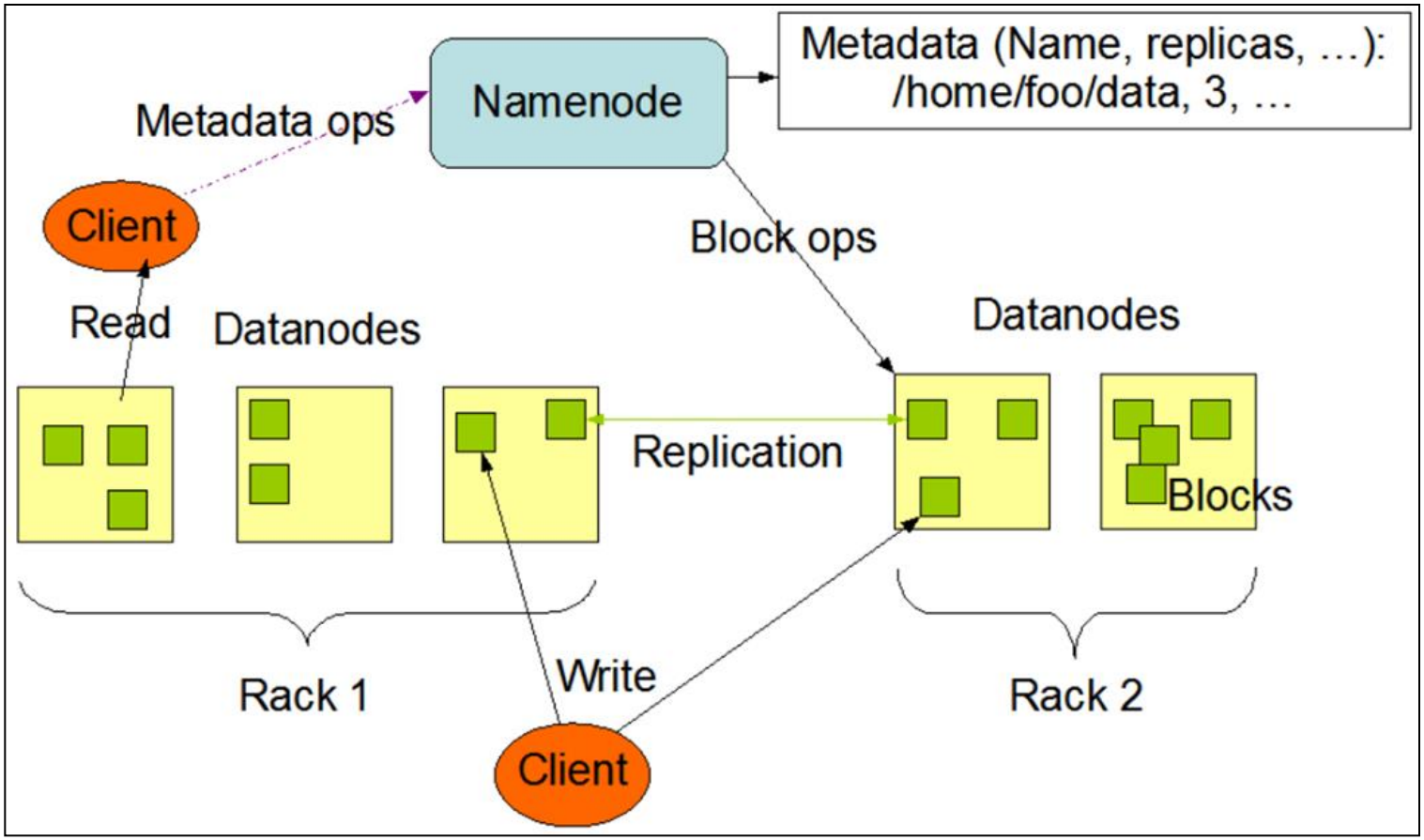


雲端儲存系統(Hbase)(MongoDB)

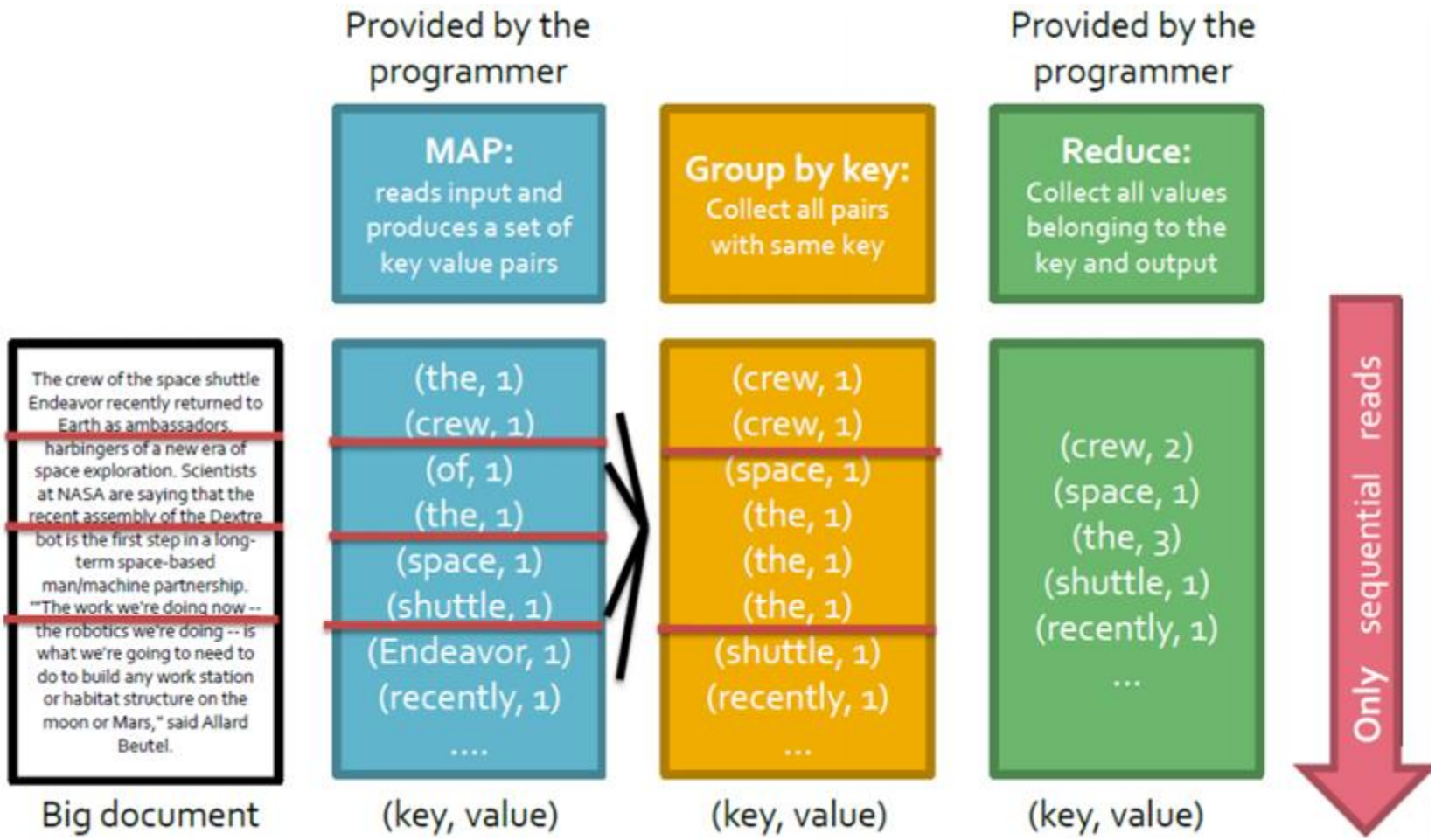
雲端運算框架(Hadoop MapReduce)

雲端檔案系統(Hadoop HDFS)

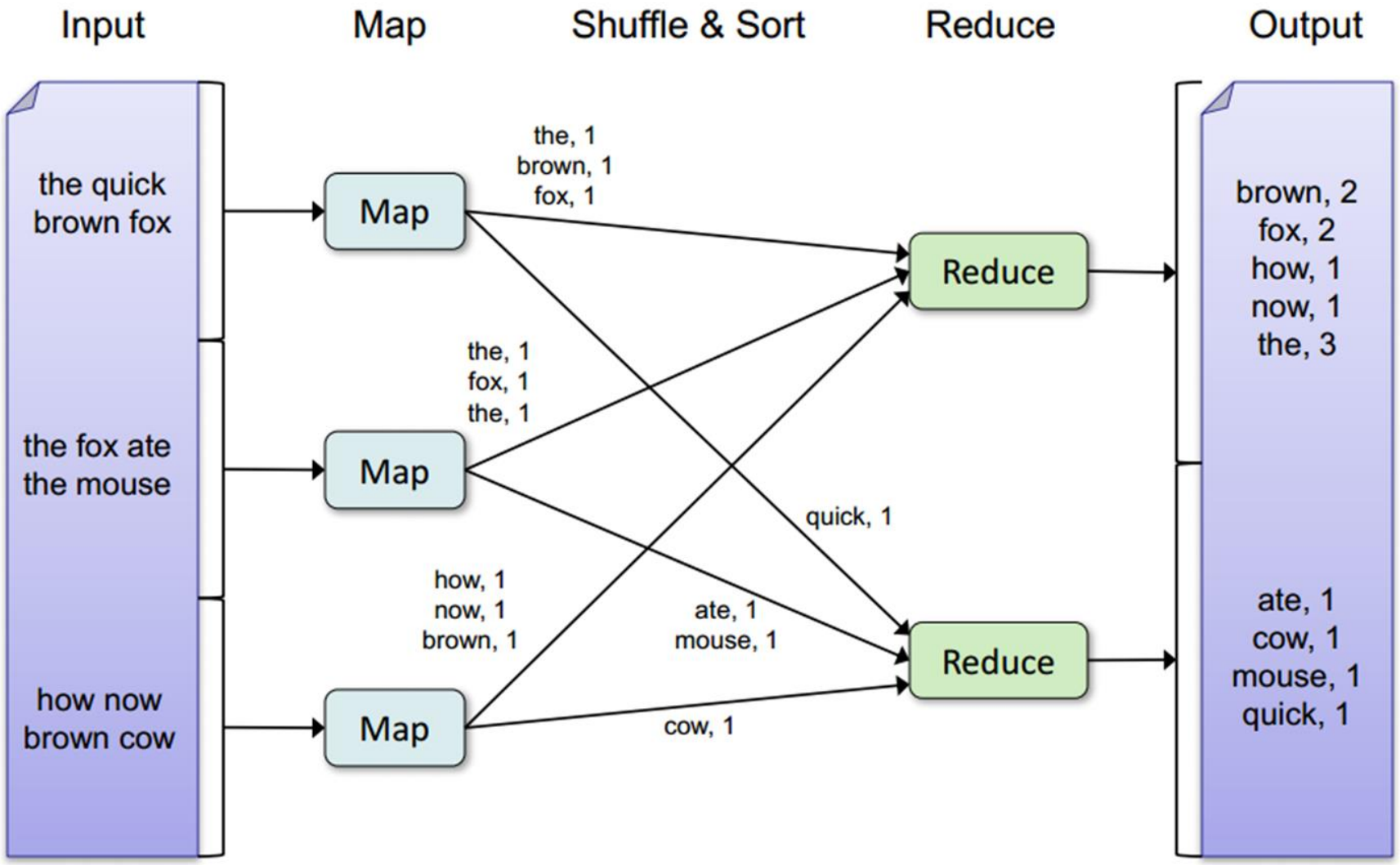
雲端分散式處理平台-HDFS



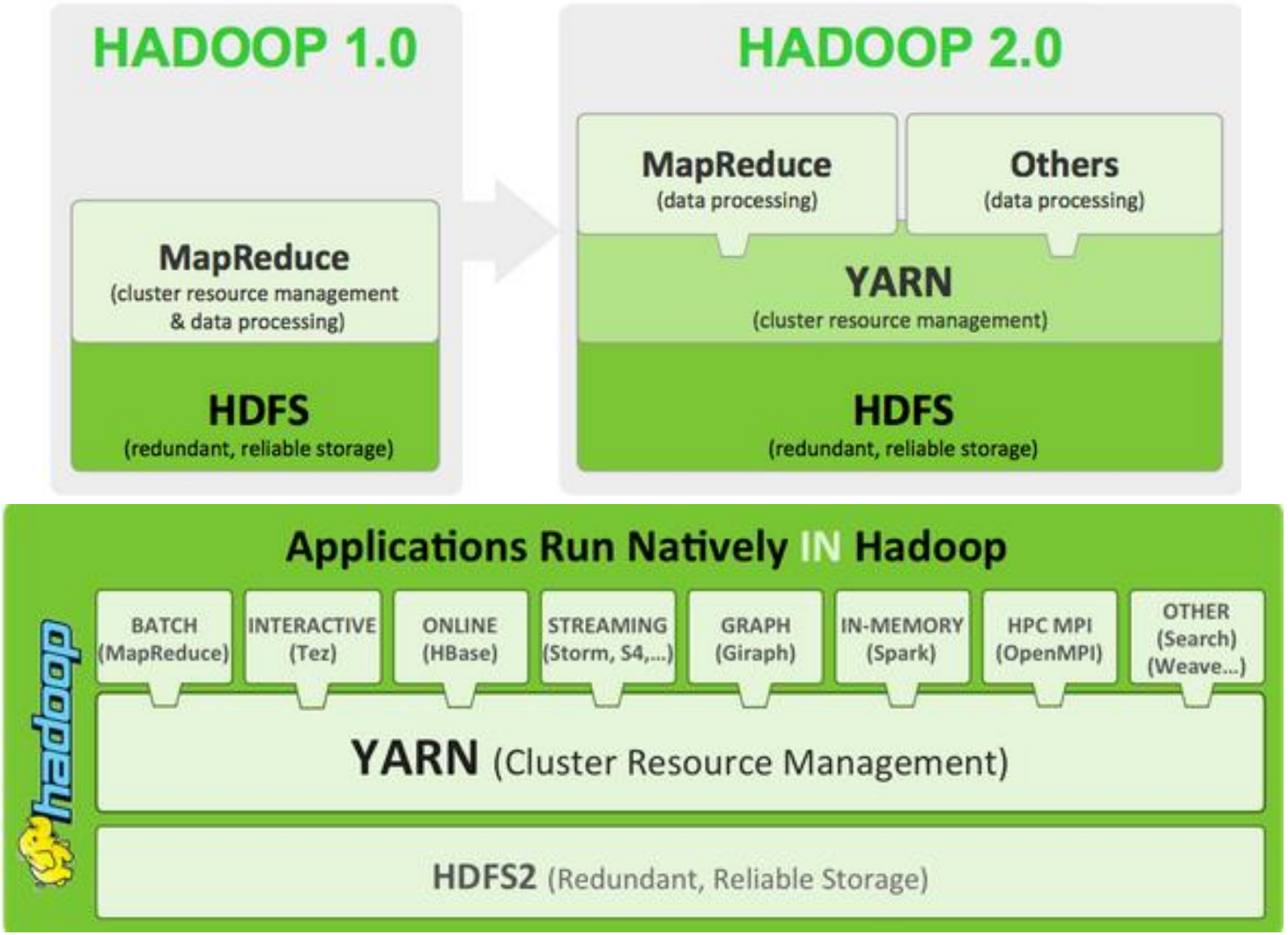
雲端分散式處理平台-MapReduce



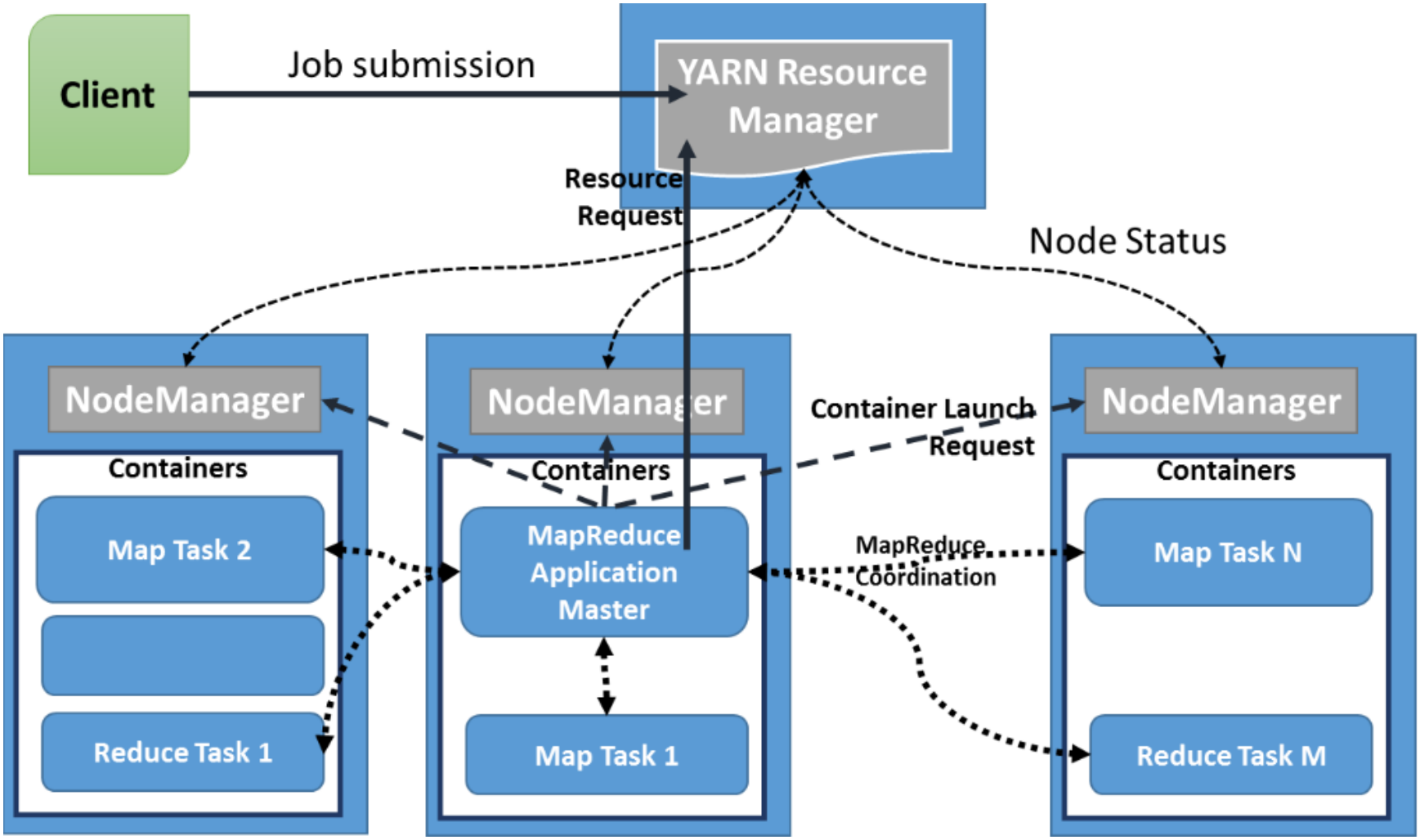
雲端分散式處理平台-MapReduce



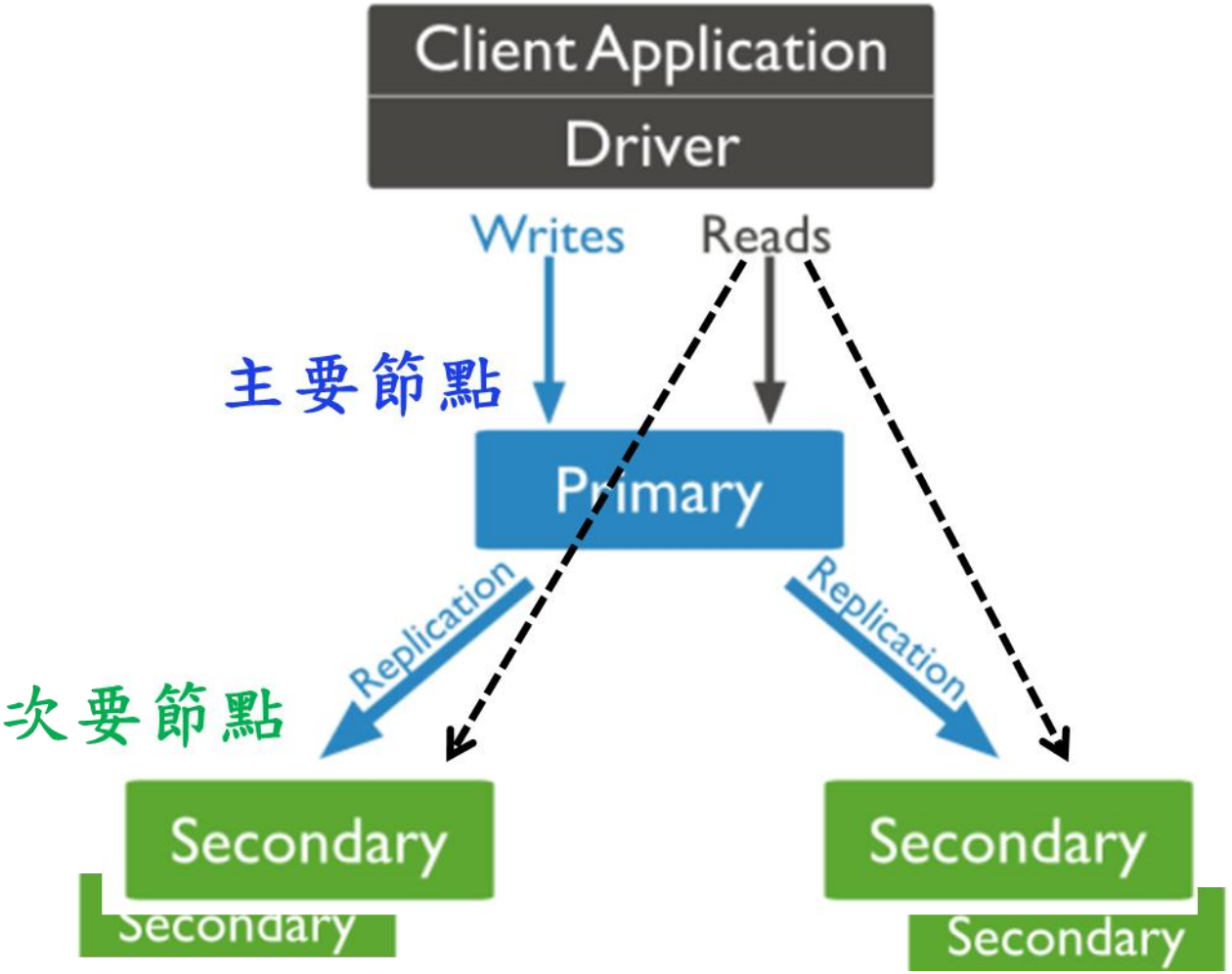
雲端分散式處理平台-Hadoop



雲端分散式處理平台-YARN



雲端分散式處理平台-MongoDB



雲端分散式處理平台-結合R軟體

Hadoop

- Hadoop (or YARN) - a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models
- RHadoop - a collection of five R packages that allow users to manage and analyze data with Hadoop, developed by Revolution Analytics
- RHIPE - an R and Hadoop Integrated Programming Environment

Spark

- Spark - a fast and general engine for large-scale data processing, which can be 100 times faster than Hadoop
- SparkR - R frontend for Spark

MongoDB

- MongoDB - an open-source document database
- R packages: rmongodb, RMongo
- A nice example of rmongodb

謝 謝 指 教