育達科技大學
Yu Da University of Science and Technology

BIG Data Application

(Business Management Seminar)

Ming-Chang Lee, Ph.D.

Department and Graduate Institute of Information Management

Yu Da University of Science and Technology

November 19, 2018


BIG
DATA
USE
ONE
SINCE
SOFTWARE
UBIQUITOUS
REQUIRING
BUSINESS
MOVING
ANALYTICS
SAN
RADIO-FREQUENCY
COMPLEXITY
SOLID
WIRELESS
TOLERABLE
PARALLEL
SIZE
NEEDED
QUALITIES
PETABYTES
INTERNET
PROCESSING
LOGS
TECHNOLOGIES
USED
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
REFINING
EVERY
CASE
LARGER
STORAGE
DATABASES
RESEARCH
MPP
TERABYTES
COST
CONTINUES
CITATION
COMPLEX
INDEXING
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
AMOUNT
DESCRIBING
ELAPSED
CURRENT
THOUGHT
SENSOR
ARCHIVES
WORKING
USING
TYPES
GARTNER
MASSIVELY
BIOLOGICAL
PERFORMANCE
RELATED
ALSO
HUNDREDS
CREATED
DISK
RELATIONAL
DESKTOP
CURRENTLY
TENS
CAPACITY
FORMS
PRACTITIONERS
ORGANIZATIONS
NOW
WORLD'S
PRESENTATIONS
FC
ZETTABYTES
GENOMICS
COMPUTING
TOOLS
SET
SHARED
COMBAT
SIGNIFICANT
TIME
WITHIN
PROCESS
INCLUDE
SYSTEMS
DEFINITION
BIOGEOCHEMICAL
NETWORKS
RECONSIDER
SEARCH
OPPORTUNITIES
INFORMATION
CONNECTOMICS
RECORDS

http://rwepa.blogspot.com/

About Me

- Ph. D., Department of Industrial and Systems Engineering, Chung Yuan Christian University.

- Shipping analyst
- Marketing analyst
- Invited Talks: +1800 hours

- Data mining
- Deep learning
- Visualization
- Optimization

- R/Python
- Office / VBA
- Database/SQL

Content

1. Big Data
2. R / RStudio
3. Big Data Application

1. Big Data

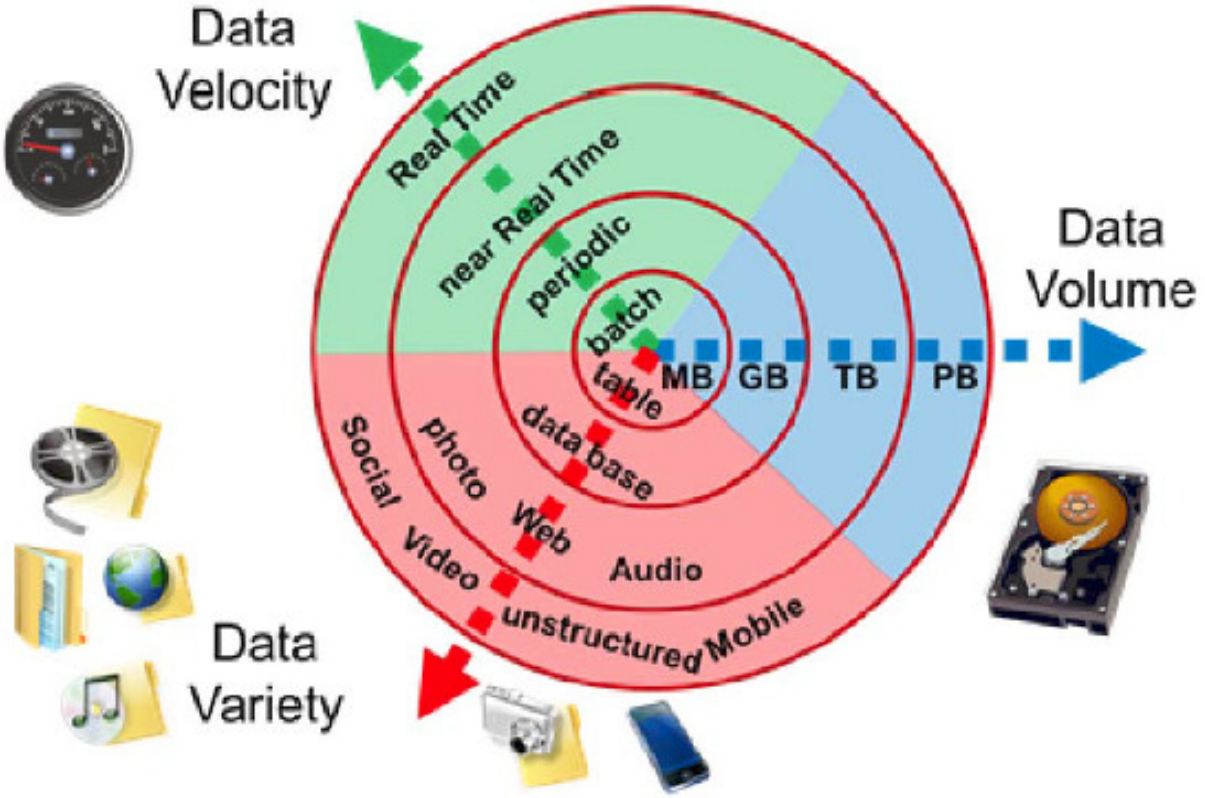
Big Data?

- Data sets that are too large or complex for traditional data-processing application software to adequately deal with.

Reference: https://en.wikipedia.org/wiki/Big_data

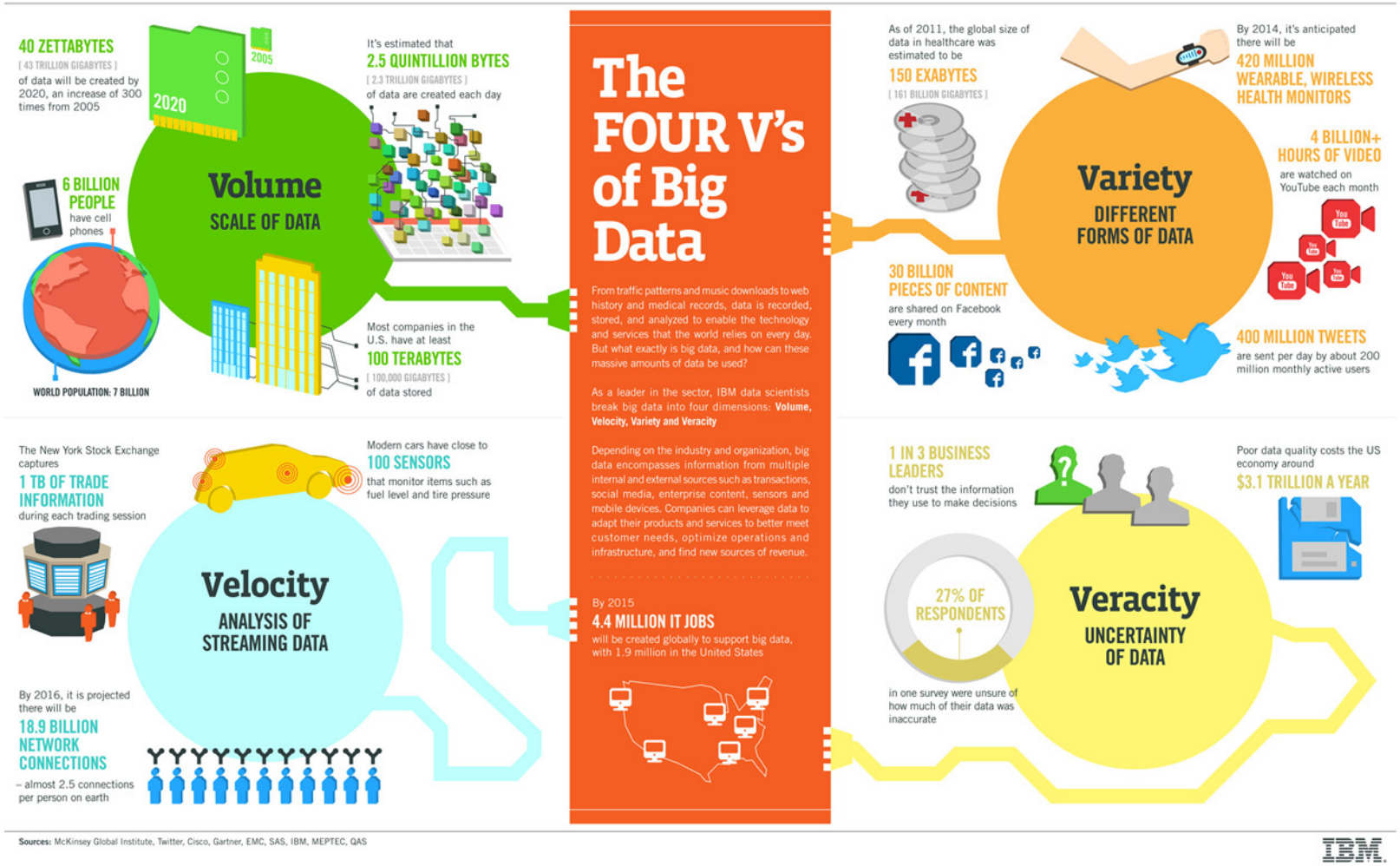
Big Data - 3V

- Doug Laney, 2001:



Reference: Lancy, D., 3D Data Management: Controlling Data Volume, Velocity, and Variety, Gartner, 2001.

IBM Big Data - 4V



Reference: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Veracity

- Veracity refers to the quality or trust worthiness of the data.
- A common complication is that the data is saturated with both useful signals and lots of noise (data that can't be trusted).

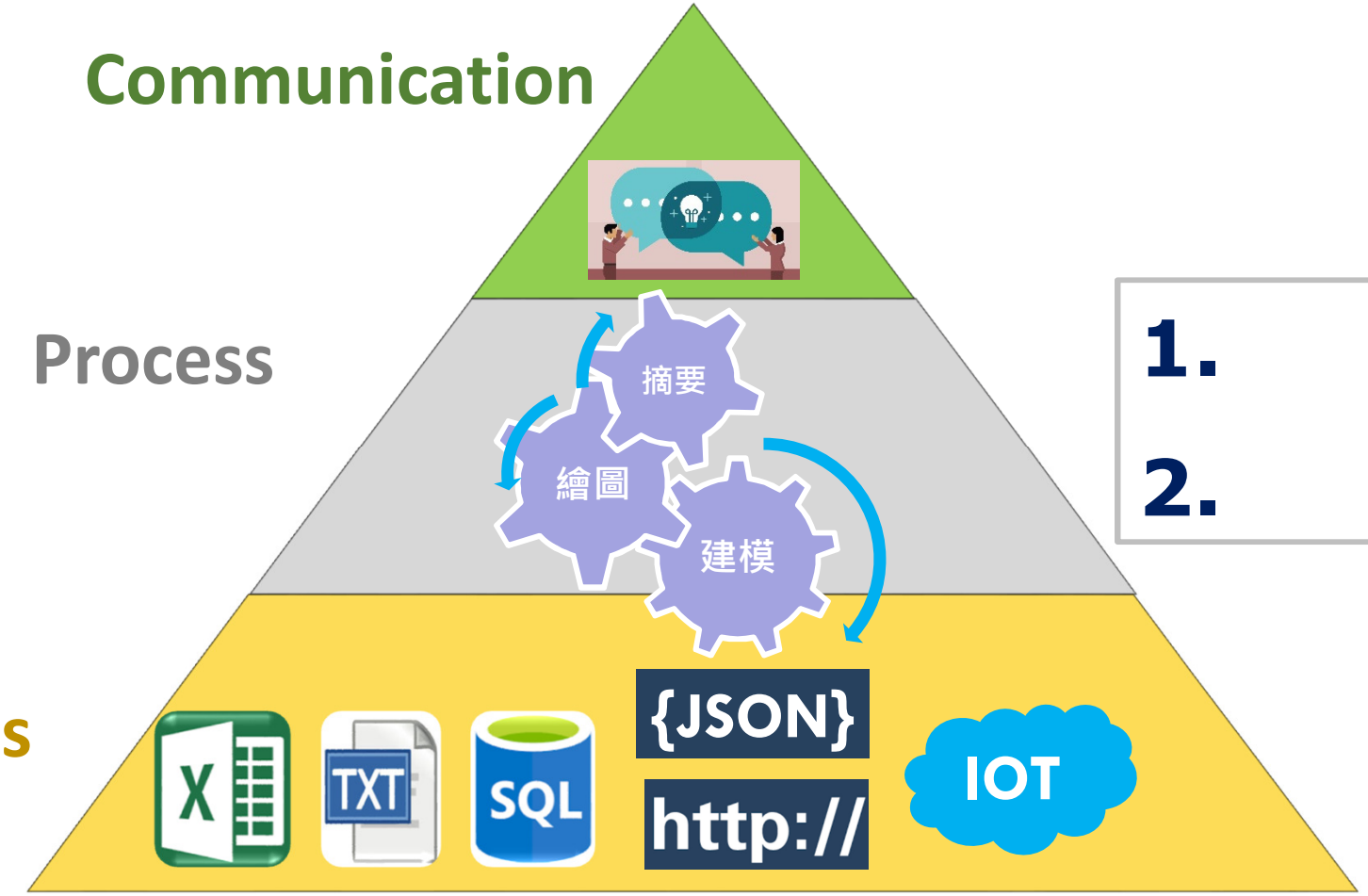
Why Big Data now?

- More data are being collected and stored
- Open sources
- Commodity hardware
- Cloud
- IOT, …

Big Data Challenges

- Capturing data
- Data storage
- Data analysis, querying
- Search
- Sharing
- Transfer
- Visualization
- Communication, …

APC approach



1.

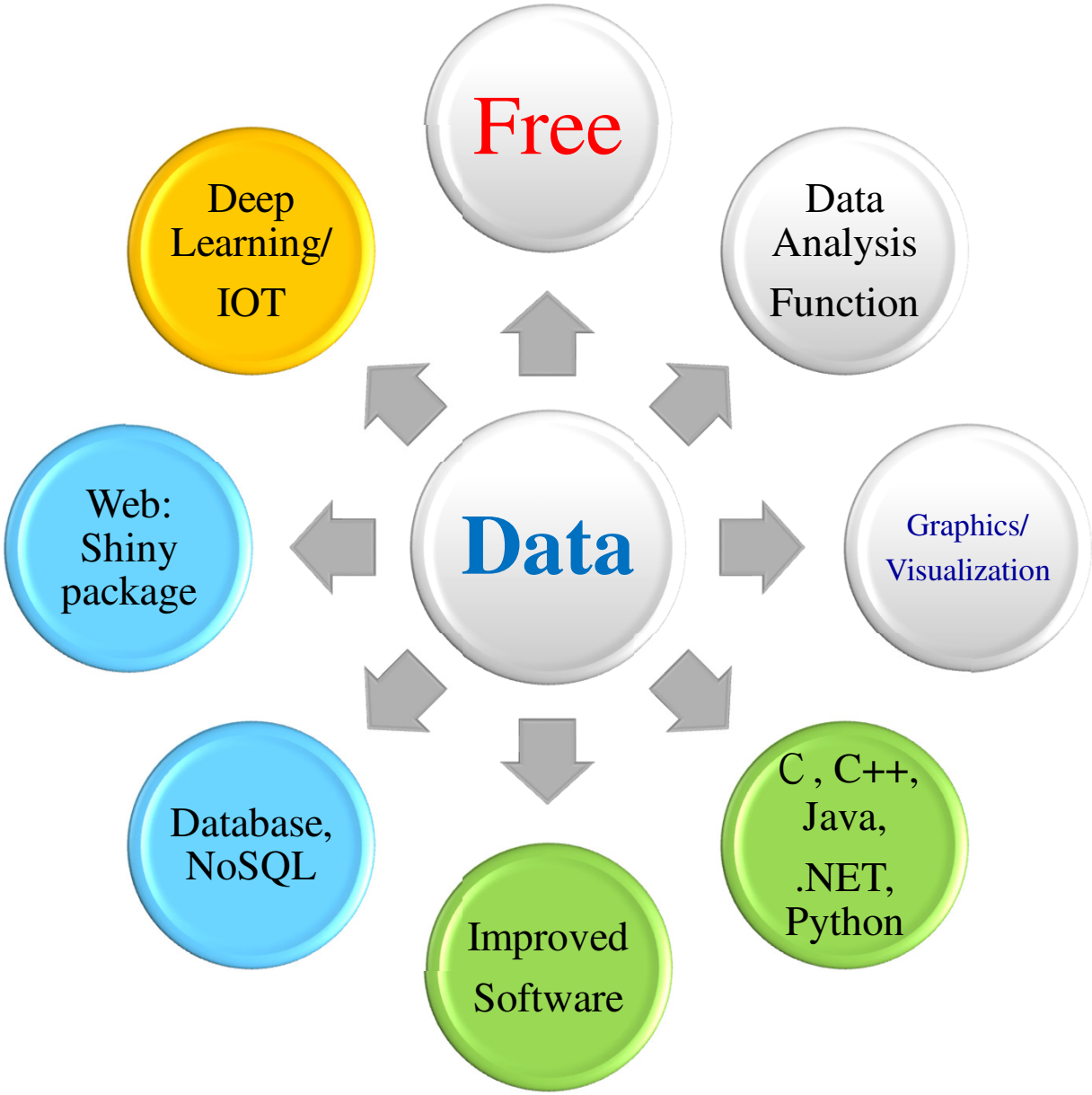
2.

2. R / RStudio

R!

- 1976 - John Chambers, Rick Becker, and Allan Wilks developed S language at Bell Labs.
- 1993 - Ross Ihaka and Robert Gentlemen developed R at University of Auckland in New Zealand.
  - **R is a programming language with the features:**
    - **Statistical analysis**
    - **Graph**
    - **Visualization**
- 1997年－R development core team
  - February 2000 － R 1.0.0
  - March 2013 － R 2.15.3
  - July 2018 － R 3.5.1

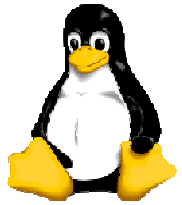
R - Features


Free
Data Analysis Function
Graphics/ Visualization
C , C++, Java, .NET, Python
Improved Software
Database, NoSQL
Web: Shiny package
Deep Learning/ IOT
Data

R- Download

- Web : http://www.r-project.org/
- Download \ CRAN
- Taiwan CRAN

  https://ftp.yzu.edu.tw/CRAN/
  http://ftp.yzu.edu.tw/CRAN/
  http://cran.csie.ntu.edu.tw/



- Download R for Windows

  - Download R for Linux
  - Download R for (Mac) OS X
  - Download R for Windows

R- Download

- base:

Download R 3.5.1 for Windows (62 megabytes, 32/64 bit)

- R install guide, 2006
  - http://web.ydu.edu.tw/~alan9956/docu/refer/R01_install.pdf
- Installed path
  - C:\R\R-3.5.1

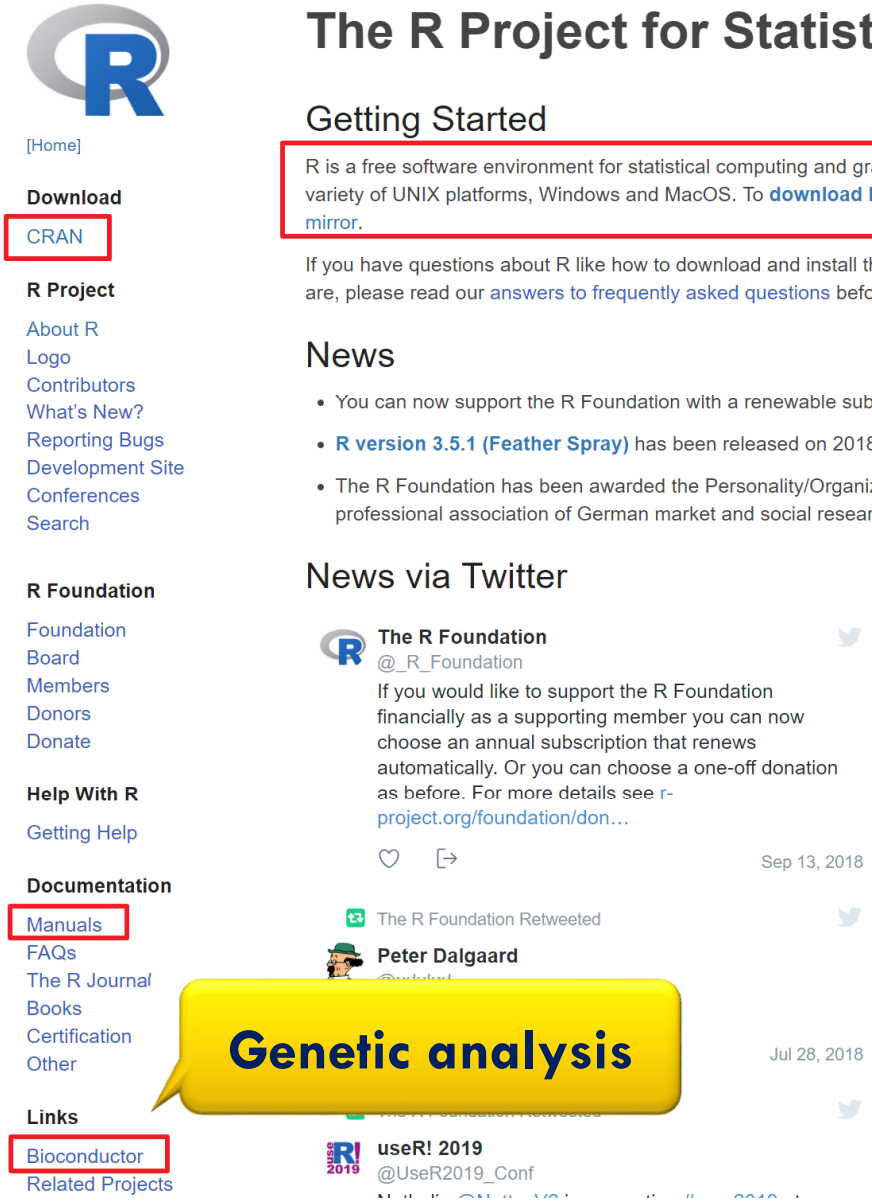
http://www.r-project.org/

Download

Free

Manuals

Genetic analysis



R - Manuals

The R Manuals

edited by the R Development Core Team.

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platform-specific pages, but most parts version of the manuals for each platform are part of the respective R installations. The manuals change with R, hence we provide versions for the most recent version for the patched release version (R-patched) and finally a version for the forthcoming R version that is still in development (R-devel).

Here they can be downloaded as PDF files, EPUB files, or directly browsed as HTML:

Manual	R-release	R-patched
An Introduction to R is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.	HTML | PDF | EPUB	HTML | PDF | EPUB
R Data Import/Export describes the import and export facilities available either in R itself or via packages which are available from CRAN.	HTML | PDF | EPUB	HTML | PDF | EPUB
R Installation and Administration	HTML | PDF | EPUB	HTML | PDF | EPUB
Writing R Extensions covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.	HTML | PDF | EPUB	HTML | PDF | EPUB
A draft of **The R language definition** documents the language *per se*. That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions.		
R Internals: a guide to the internal structures of R and coding standards for the core team working on R itself.		
The R Reference Index: contains all help files of the R standard and recommended packages in printable form. (9MB, approx. 3500 pages)		

contributed documentation (Free!)

Translations of manuals into other languages than English are available from the contributed documentation section (only a few translations are available).

R - Manuals

Contributed Documentation

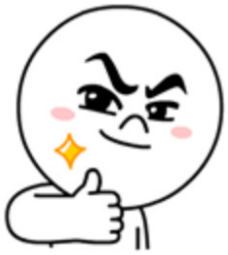
Note: The CRAN area for contributed documentation is frozen and no longer actively maintained.

English --- Other Languages

Manuals, tutorials, etc. provided by users of R. The R core team does not take any responsibility for contents, but we appreciate the effort very much and encourage everybody to contribute to this list! To submit, follow the submission instructions on the CRAN main page. All material below is available directly from CRAN, you may also want to look at the list of other R documentation available on the Internet.

Note: Please use the directory listing to sort by name, size or date (e.g., to see which documents have been updated latetly).

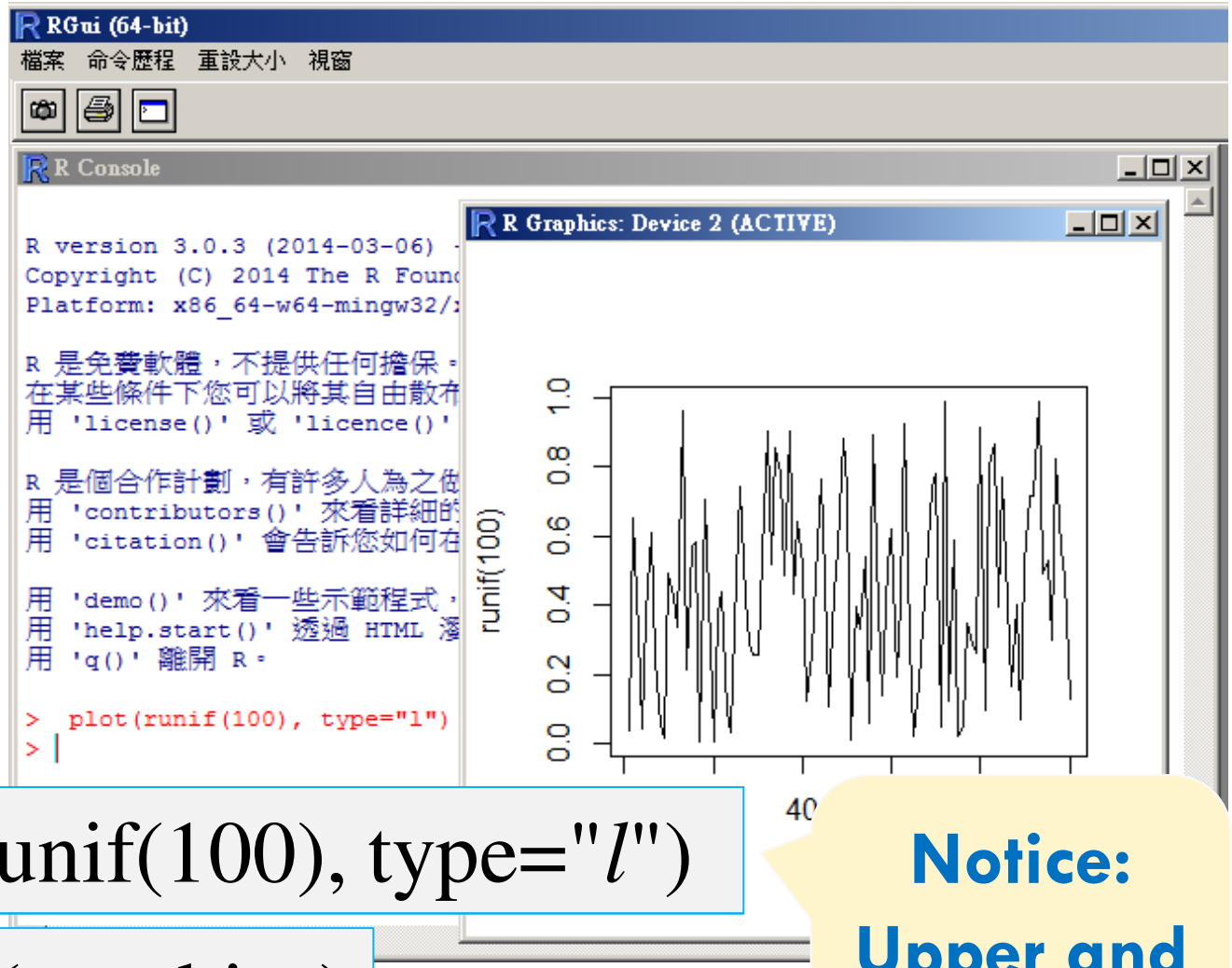
English Documents



Documents with more than 100 pages:

- **"Visual Statistics. Use R!"** by Alexey Shipunov (PDF, 2017-05-15, 388 pages) erials are accessible from Alexey Shipunov's English R page .
- **"Using R for Data Analysis and Graphics - Introduction, Examples and Commentary"** by John Maindonald (PDF, data sets and scripts are available at JM's homepage).
- **"Practical Regression and Anova using R"** by Julian Faraway (PDF, data sets and scripts are available at the book homepage).

R - Practice




```
plot(runif(100), type="l")
```

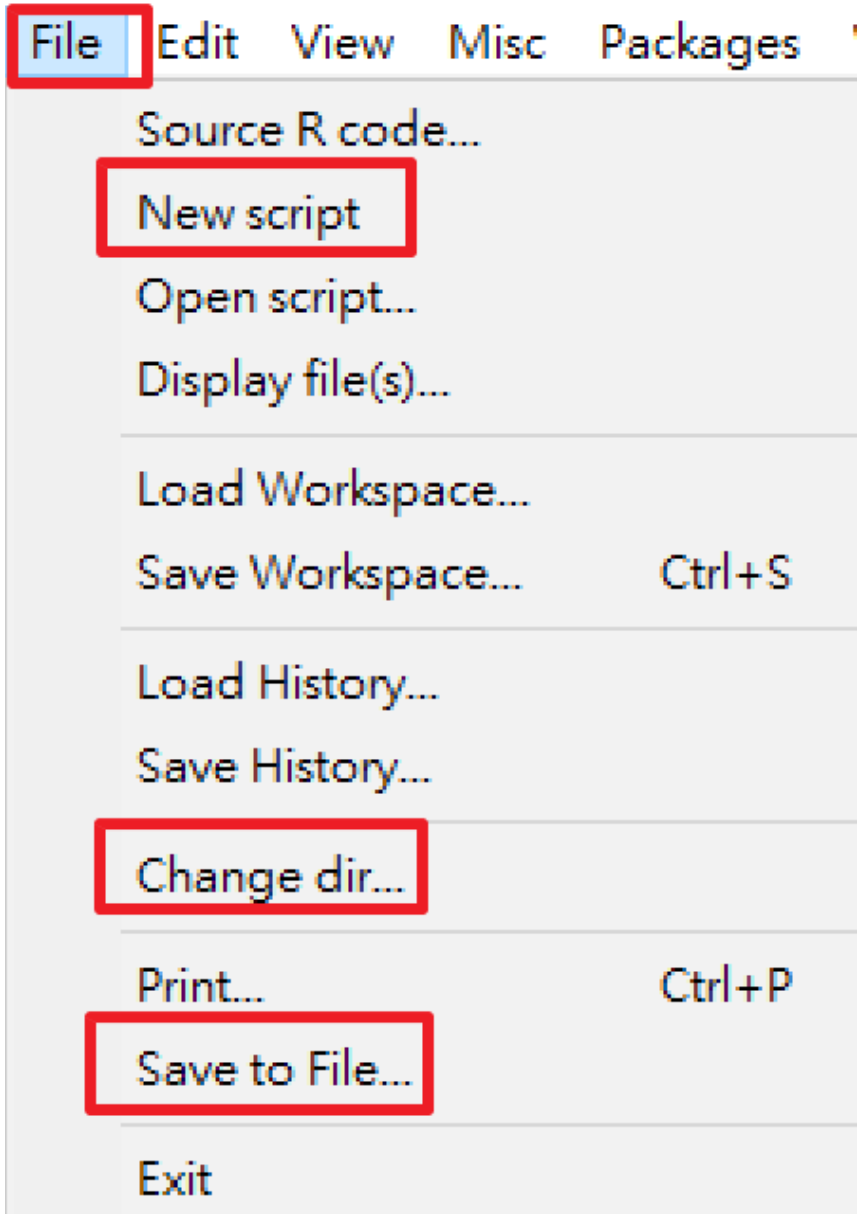
```
demo(graphics)
```

```
demo(persp)
```

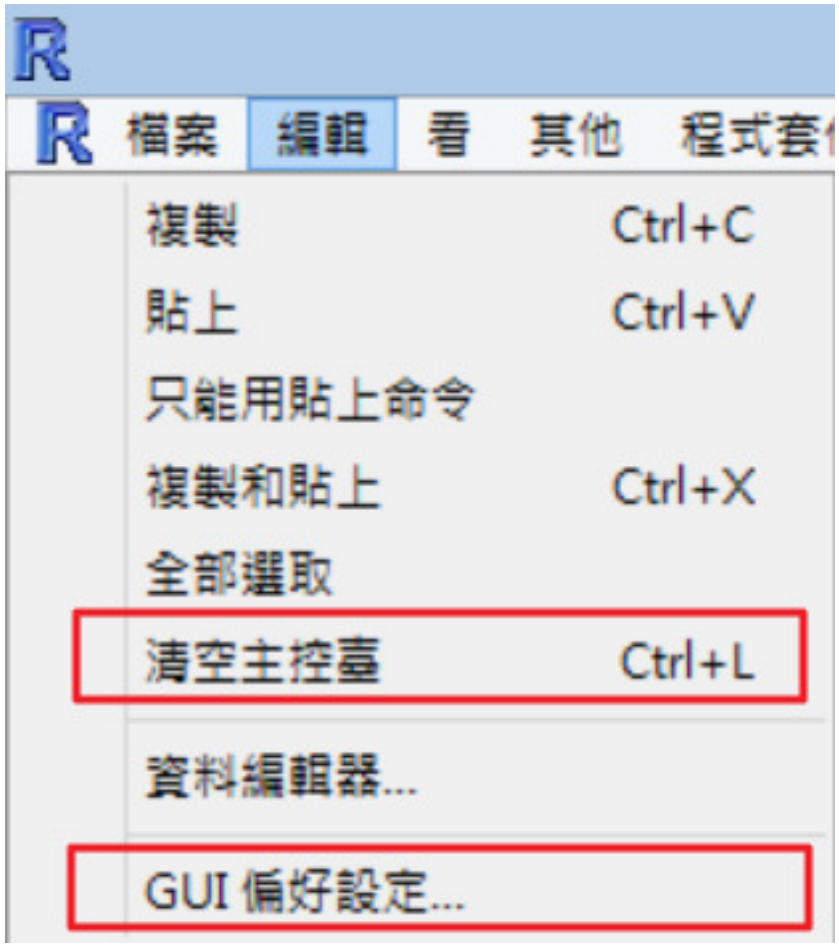
Notice: Upper and lower case

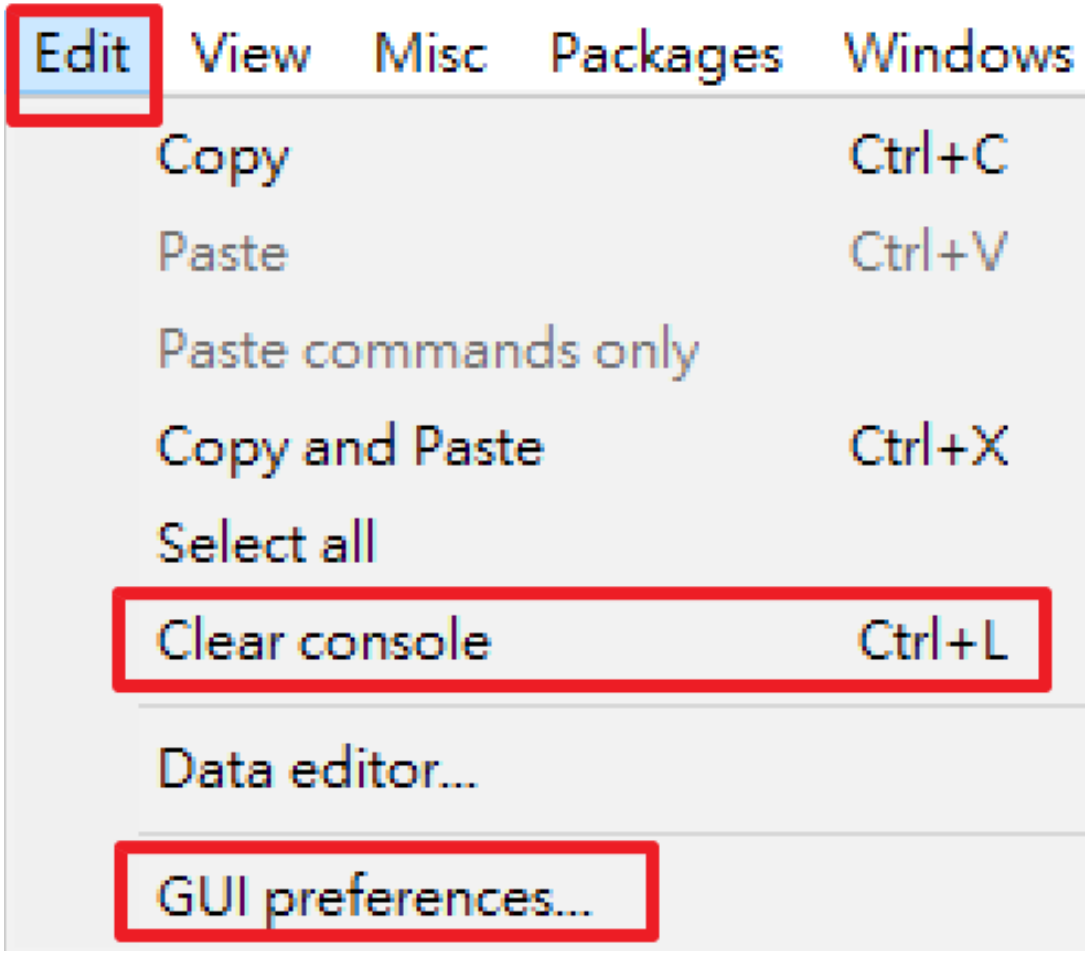
R – Menu [File]


檔案
編輯
看
其他
輸入 R 程式碼...
建立新的命令稿
開啟命令稿...
顯示檔案...
載入工作空間...
儲存工作空間...
載入命令歷程...
儲存命令歷程...
變更現行目錄...
列印...
儲存到檔案...
離開


File
Edit
View
Misc
Packages
Source R code...
New script
Open script...
Display file(s)...
Load Workspace...
Save Workspace... Ctrl+S
Load History...
Save History...
Change dir...
Print... Ctrl+P
Save to File...
Exit

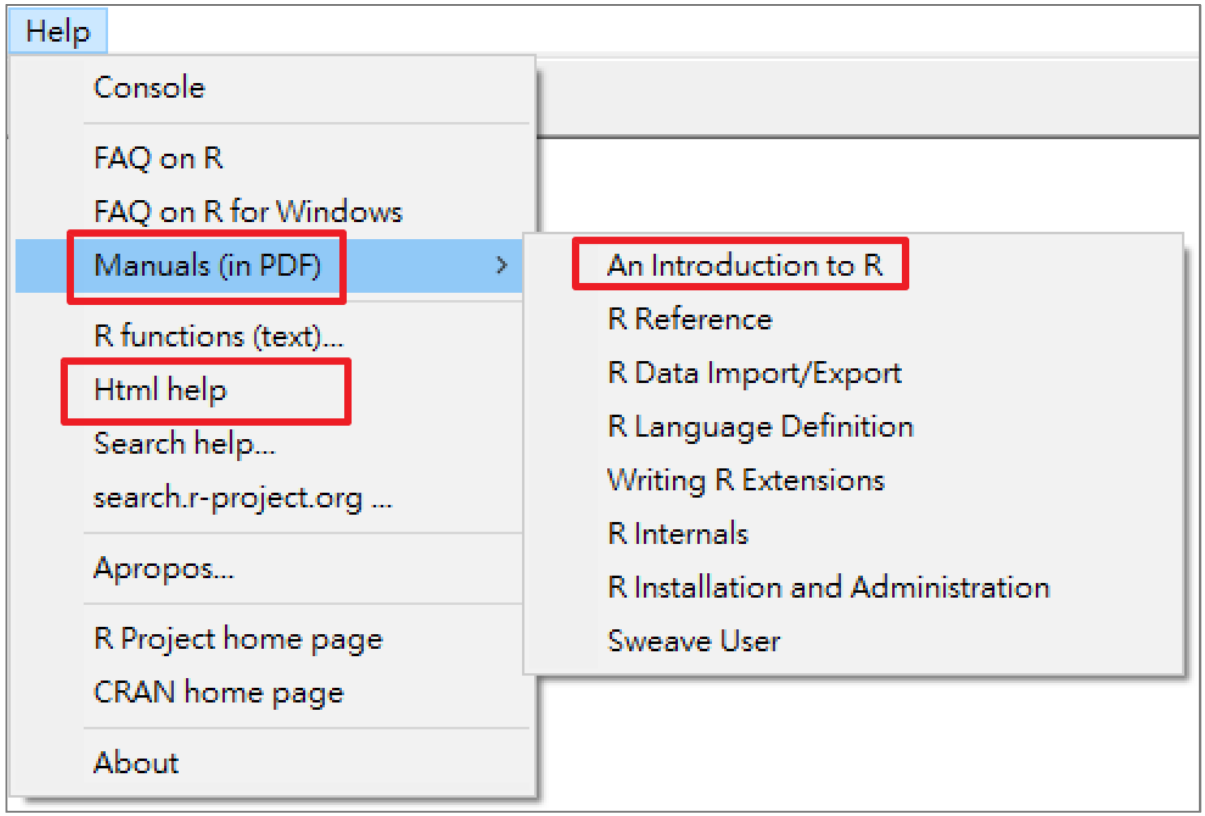
R – Menu [Edit]


檔案
編輯
看
其他
複製 Ctrl+C
貼上 Ctrl+V
只能用貼上命令
複製和貼上 Ctrl+X
全部選取
清空主控臺 Ctrl+L
資料編輯器...
GUI 偏好設定...

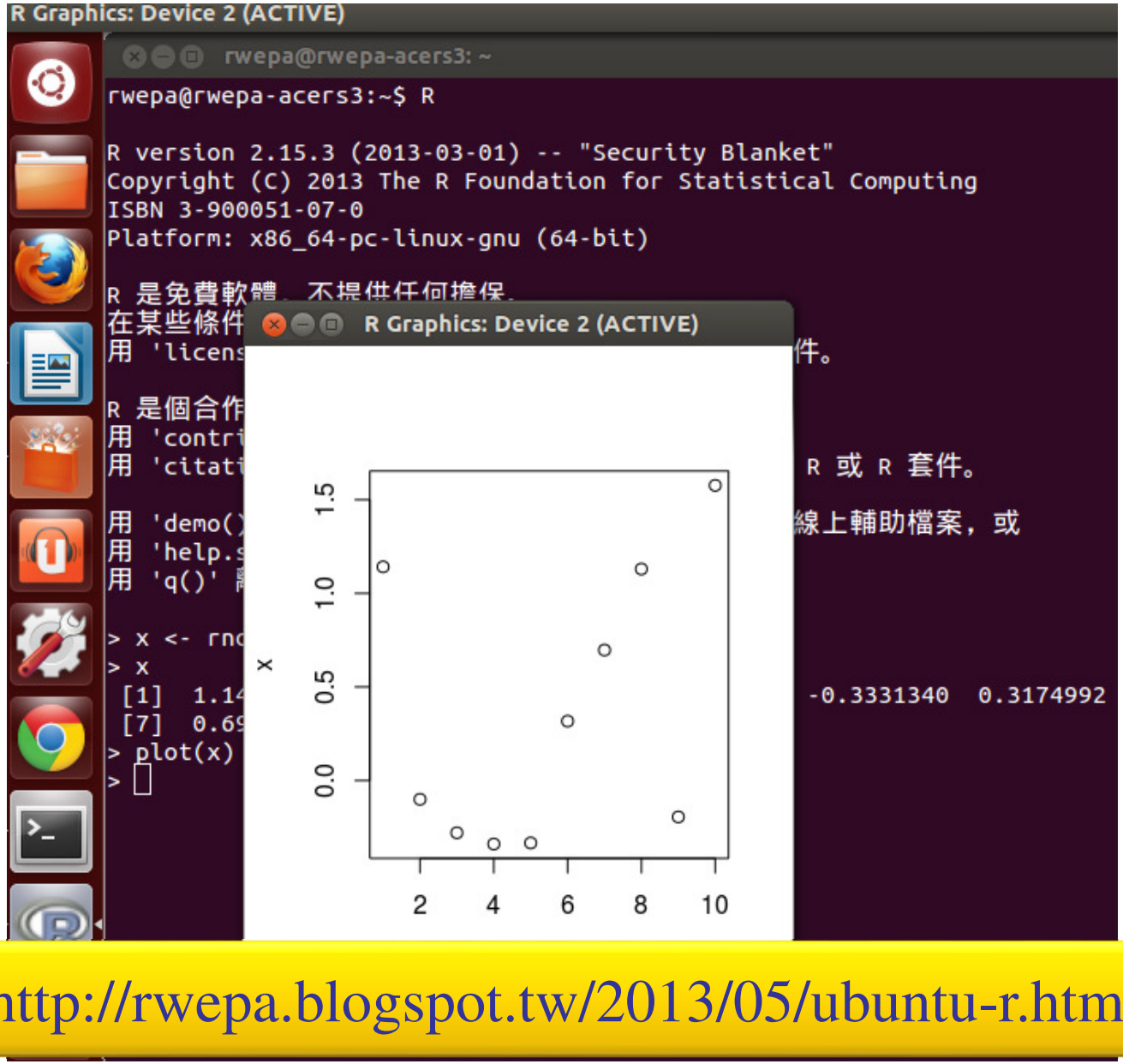

Edit
View
Misc
Packages
Windows
Copy Ctrl+C
Paste Ctrl+V
Paste commands only
Copy and Paste Ctrl+X
Select all
Clear console Ctrl+L
Data editor...
GUI preferences...

R – Menu [Help]


編輯
看
其他
程式套件
視窗
輔助
主控臺
關於 R 的 FAQ
關於 Windows 上 R 的 FAQ(W)
手冊 (PDF檔案)
R 函式 (文字檔案)...
Html 輔助
搜尋輔助...
search.r-project.org ...
關於...
R 首頁
CRAN 首頁
關於


Help
Console
FAQ on R
FAQ on R for Windows
Manuals (in PDF)
R functions (text)...
Html help
Search help...
search.r-project.org ...
Apropos...
R Project home page
CRAN home page
About
An Introduction to R
R Reference
R Data Import/Export
R Language Definition
Writing R Extensions
R Internals
R Installation and Administration
Sweave User

Ubuntu - R

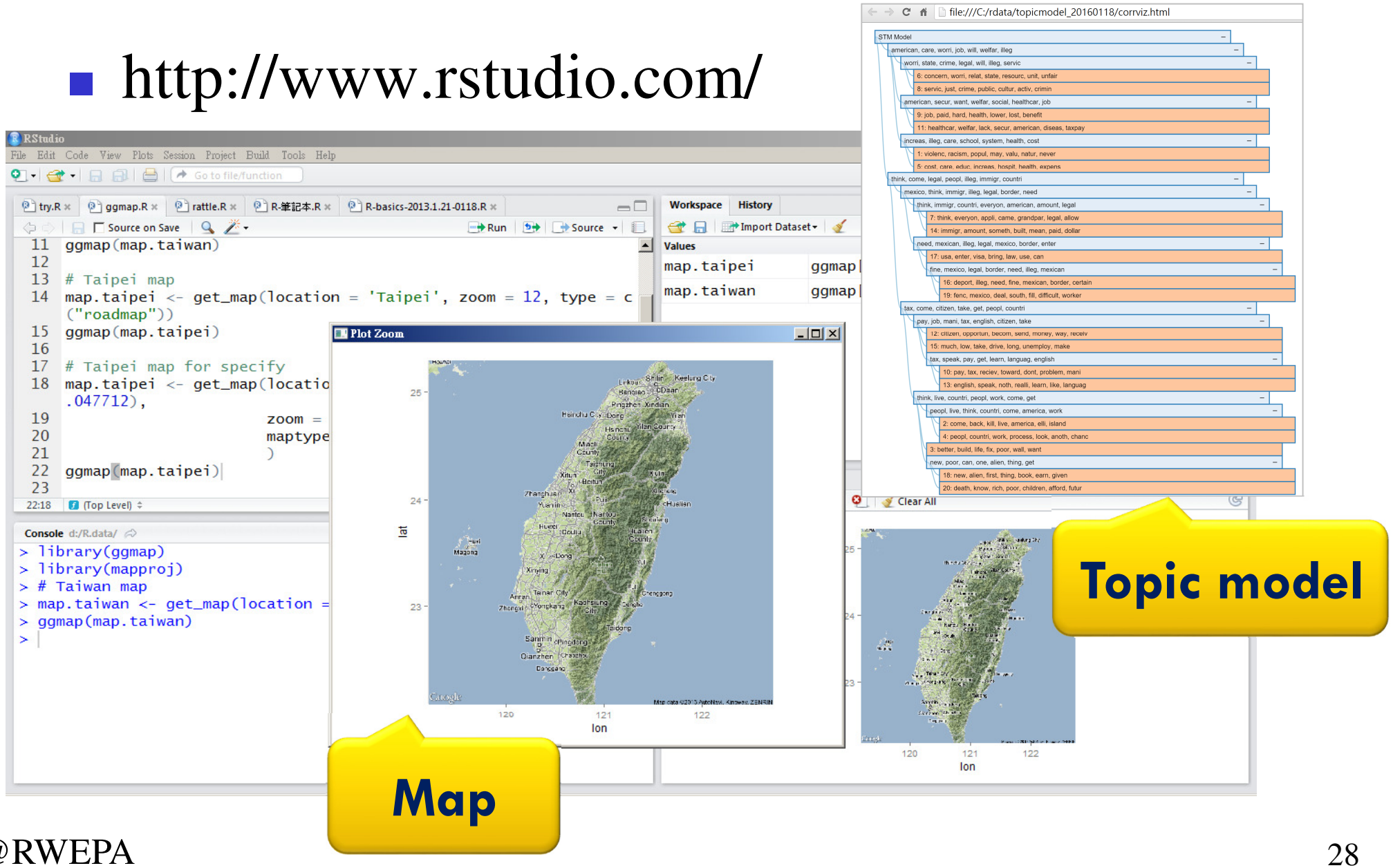


http://rwepa.blogspot.tw/2013/05/ubuntu-r.html

RStudio

Integrated Development Environment- RStudio

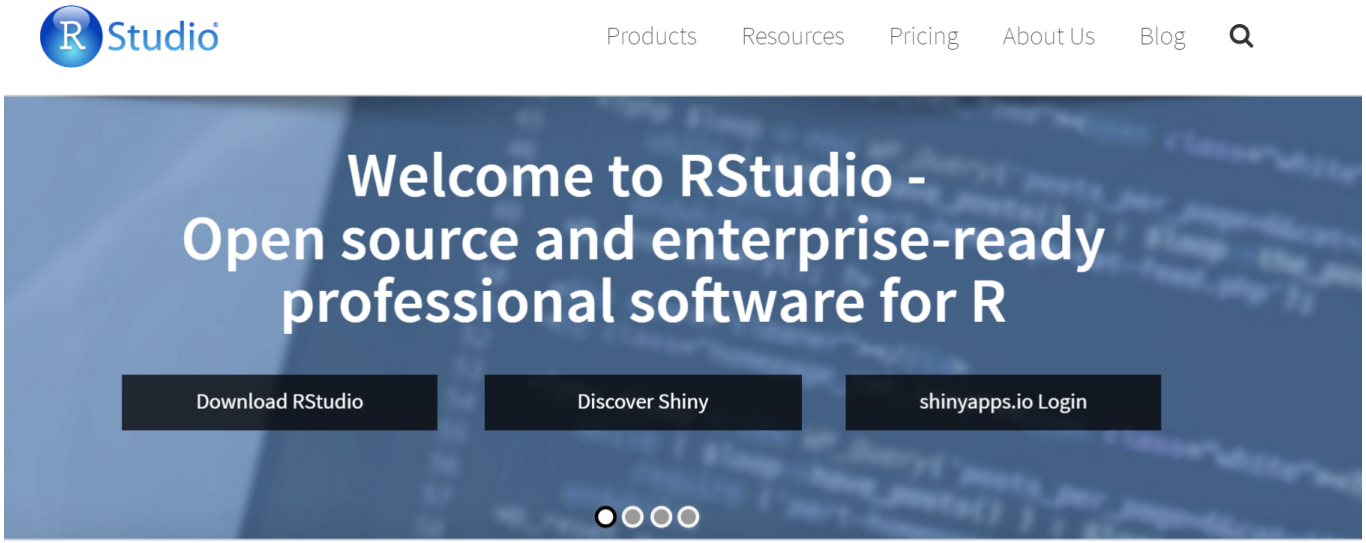
- http://www.rstudio.com/

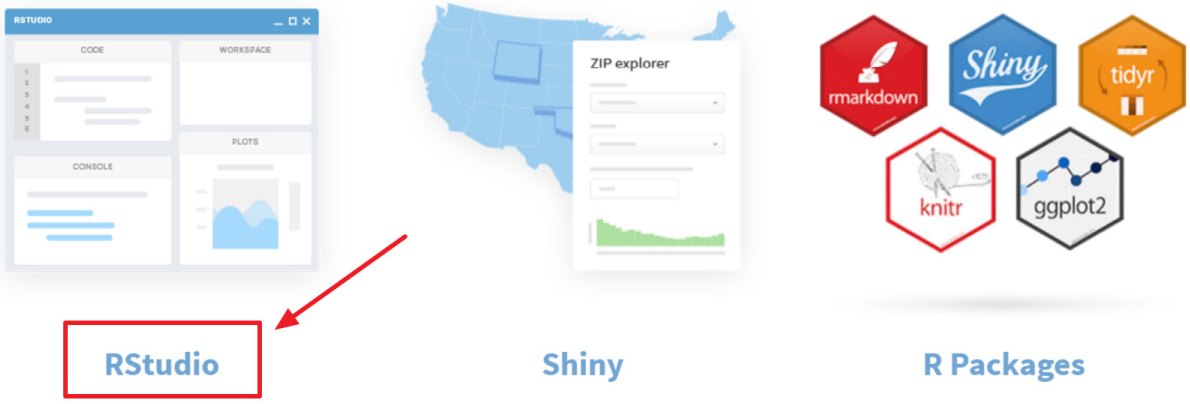


RStudio - Download

- http://www.rstudio.com/

Try: RStudio daily





RStudio - Download

RStudio Products Resources Pricing About Us Blog

Choose Your Version of RStudio

Desktop Server

	RStudio Desktop (Free License)	RStudio Desktop (Commercial License)	RStudio Server (Free License)	RStudio Server Pro (Commercial License)
Integrated Development Environment for R	✓	✓	✓	✓
Priority support		✓		✓
Access via Web Browser			✓	✓
Enterprise Security and Access Controls				✓
Project Sharing				✓
Access to Multiple Versions of R				✓
Multiple Concurrent Sessions				✓
Administrative Dashboard				✓
Load Balancing and Resource Management				✓
License	AGPL	Commercial	AGPL	Commercial
	DOWNLOAD	DOWNLOAD	DOWNLOAD	DOWNLOAD

Free

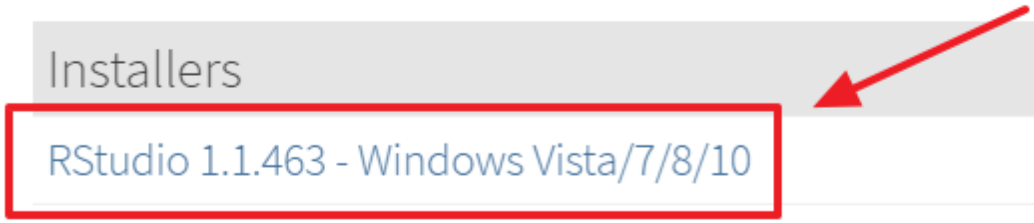
RStudio - Download

RStudio Desktop 1.1.463 — Release Notes

RStudio requires R 3.0.1+. If you don't already have R, download it here.

Linux users may need to import RStudio's public code-signing key prior to installation, policy.

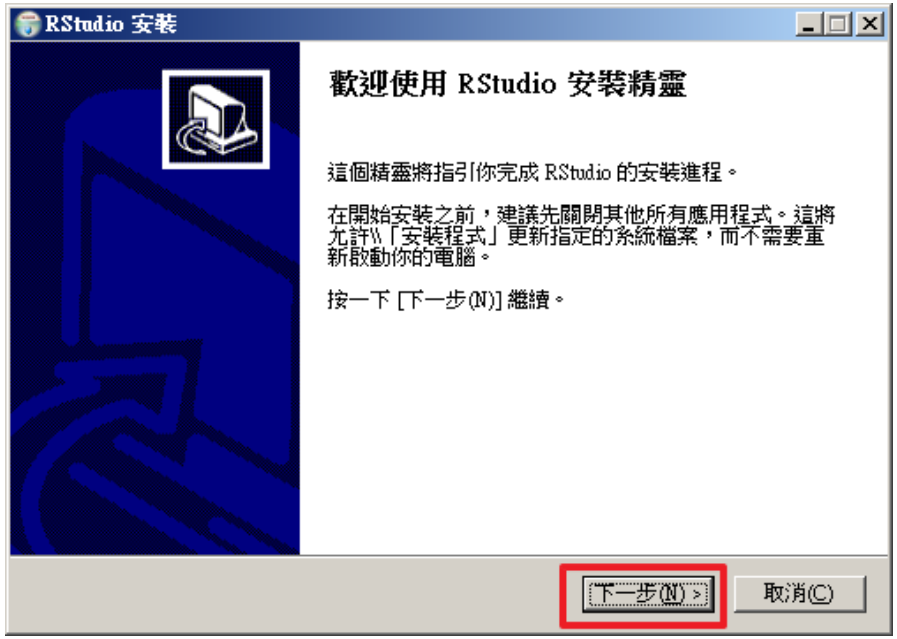
Installers for Supported Platforms



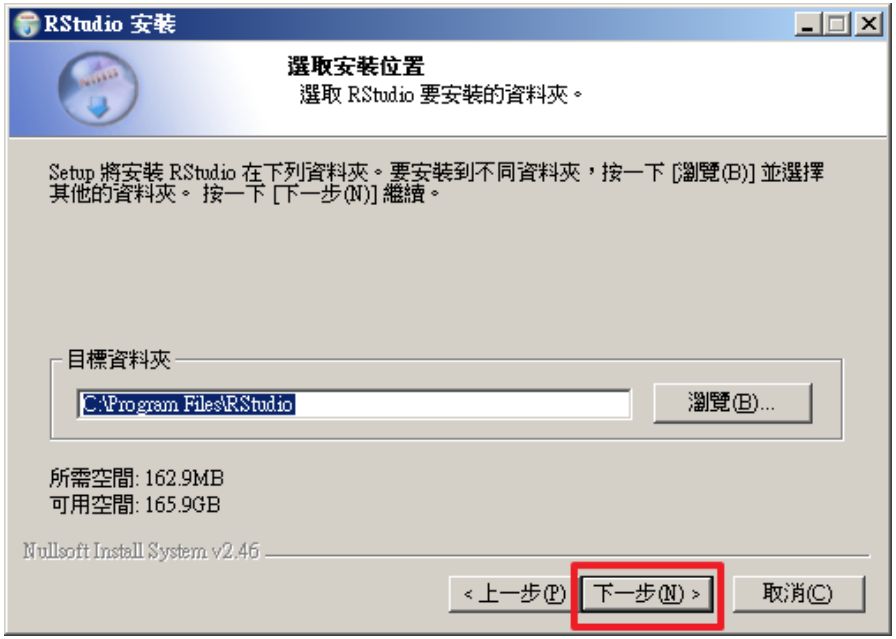
Installers	Size	Date
RStudio 1.1.463 - Windows Vista/7/8/10	85.8 MB	2018-10-29
RStudio 1.1.463 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-10-29
RStudio 1.1.463 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-10-29

RStudio - Installation

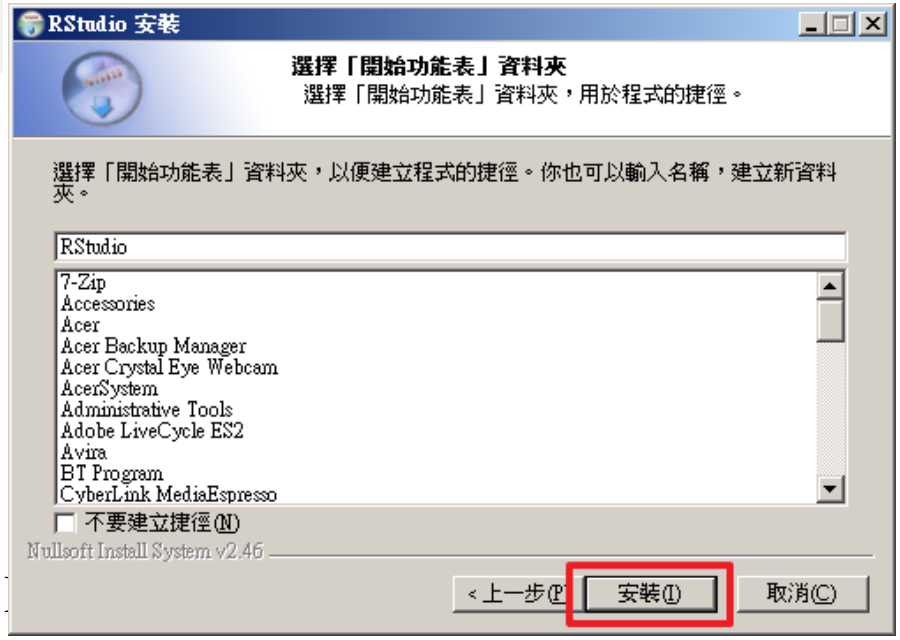
1



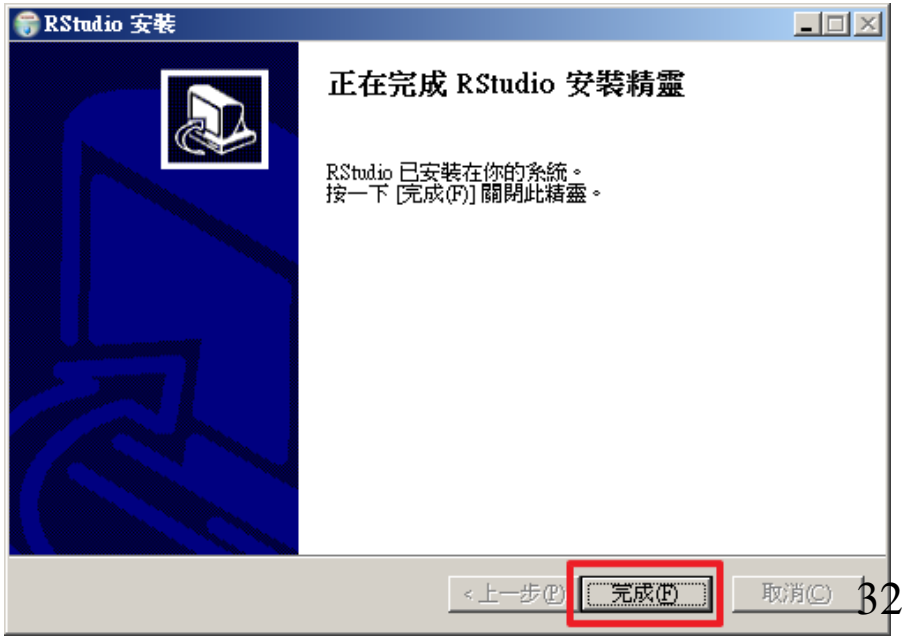
2



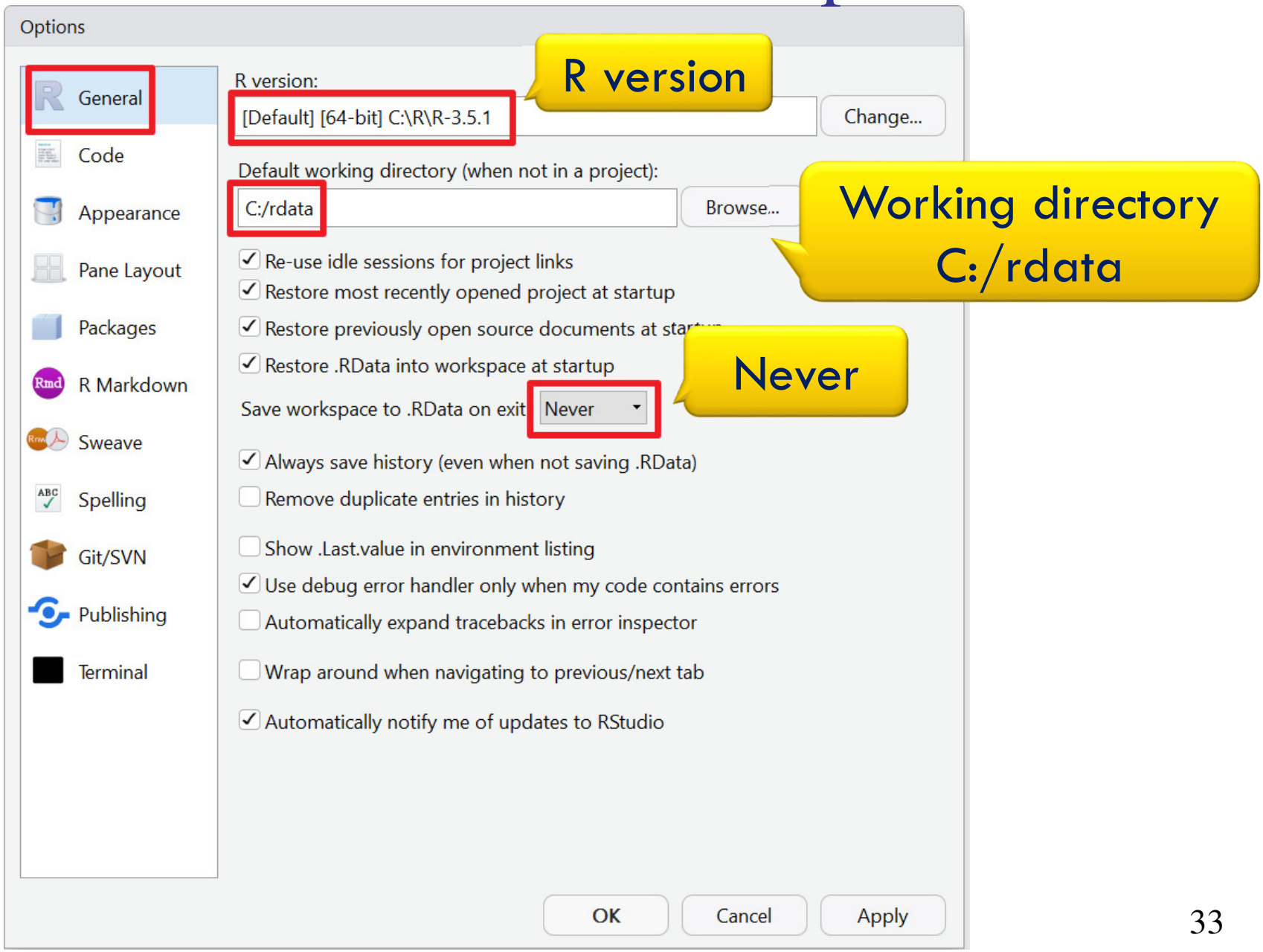
3



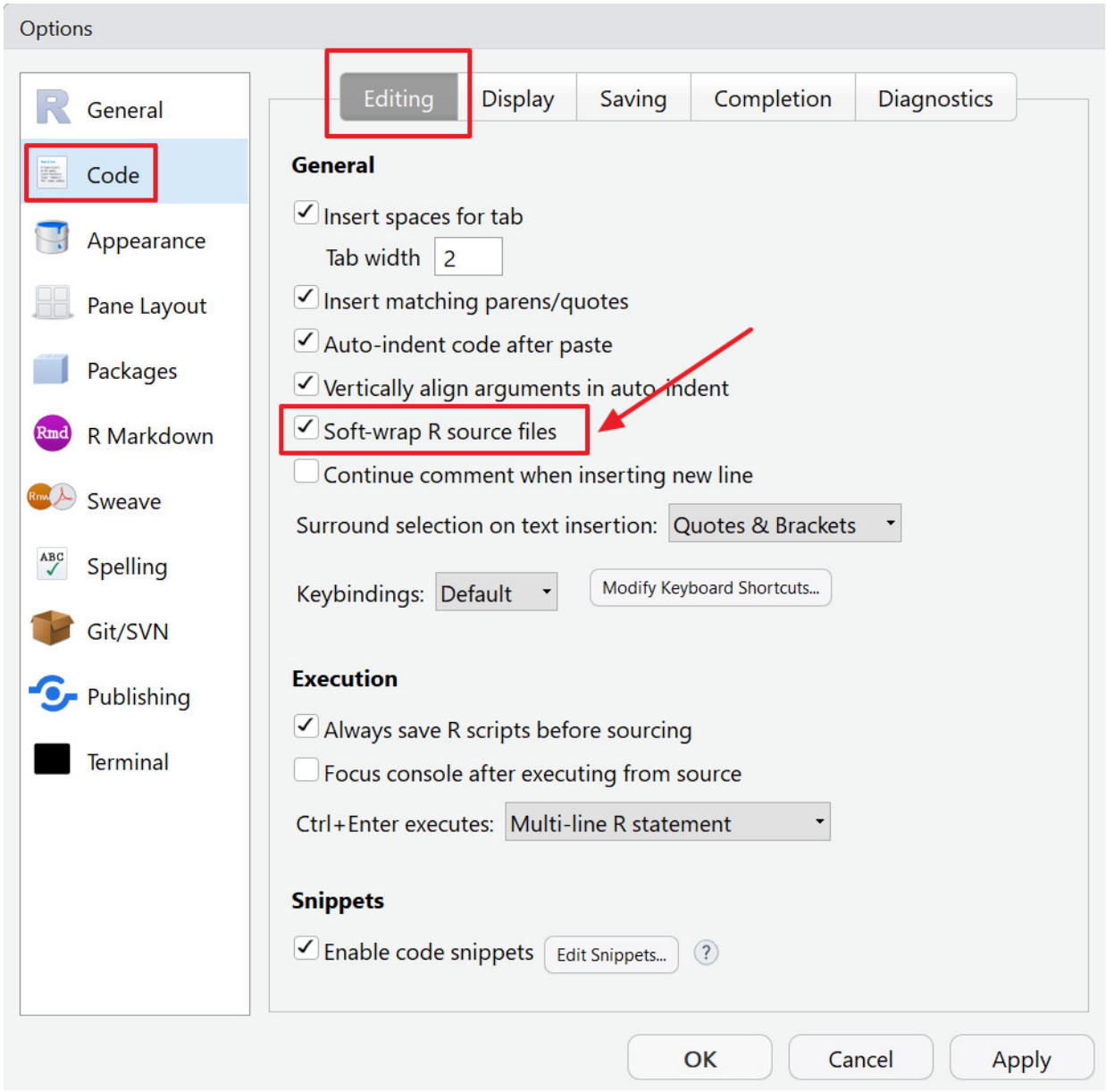
4



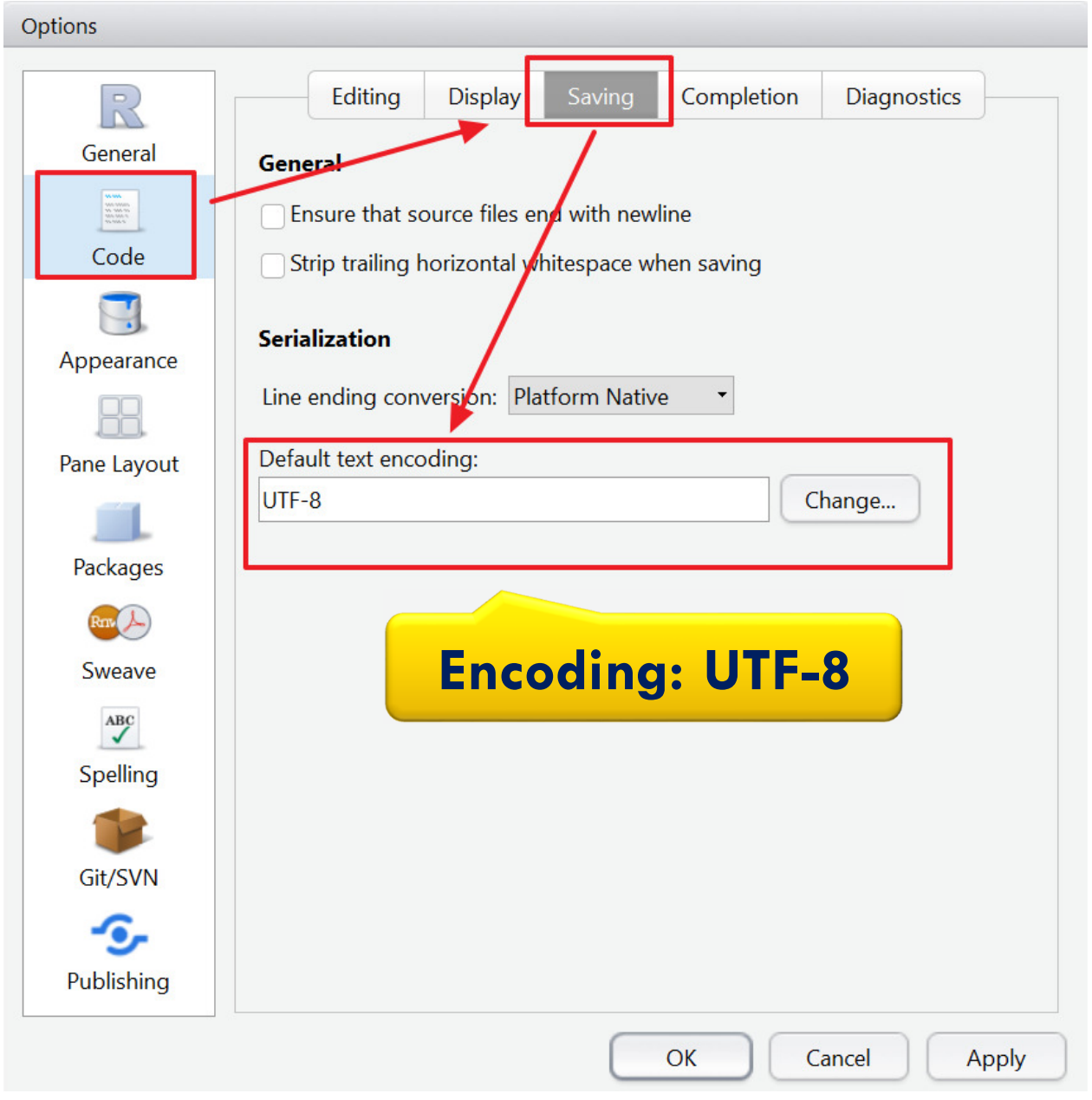
RStudio - Tools \ Global Options



RStudio - Options

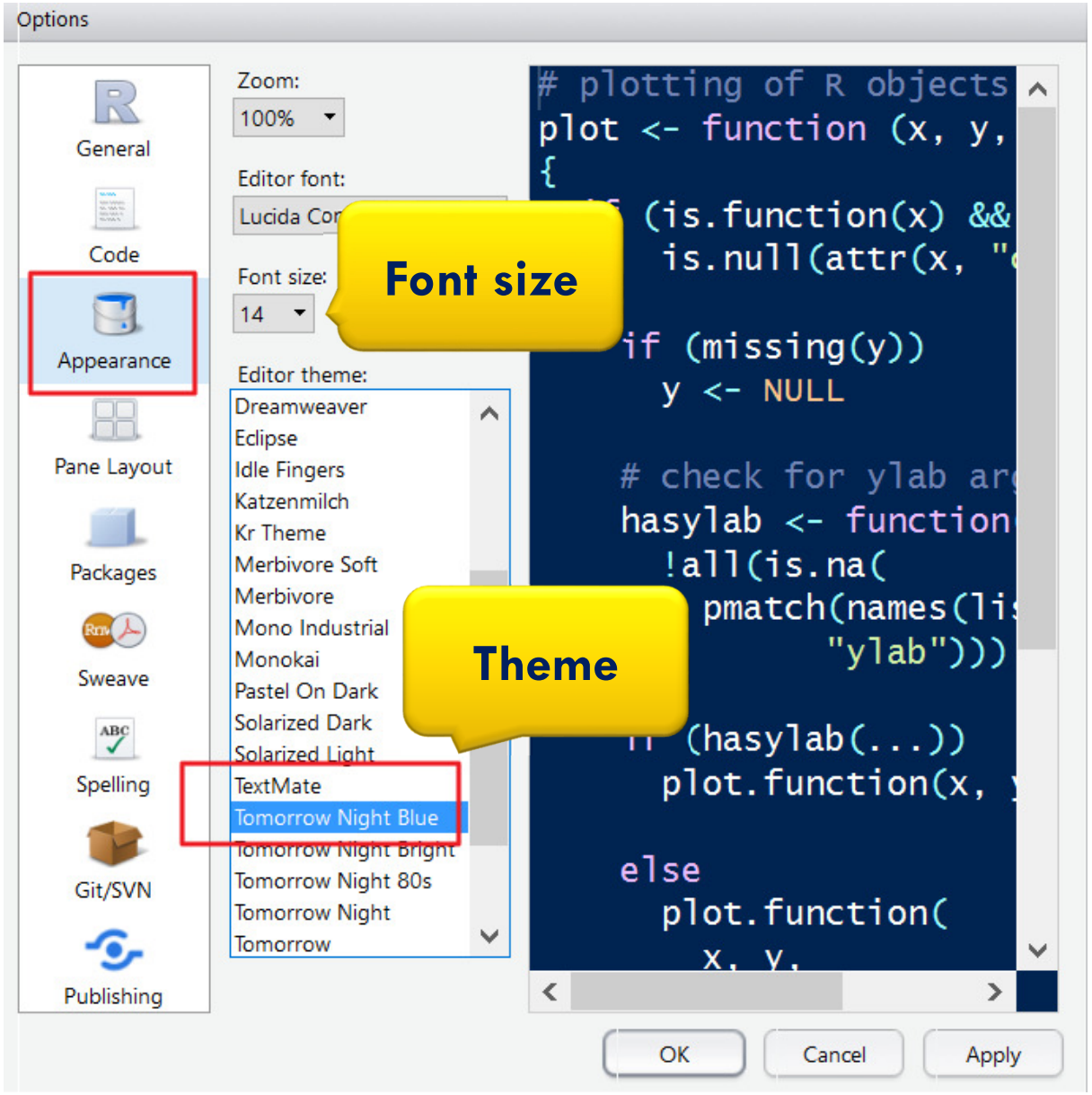


RStudio - Options

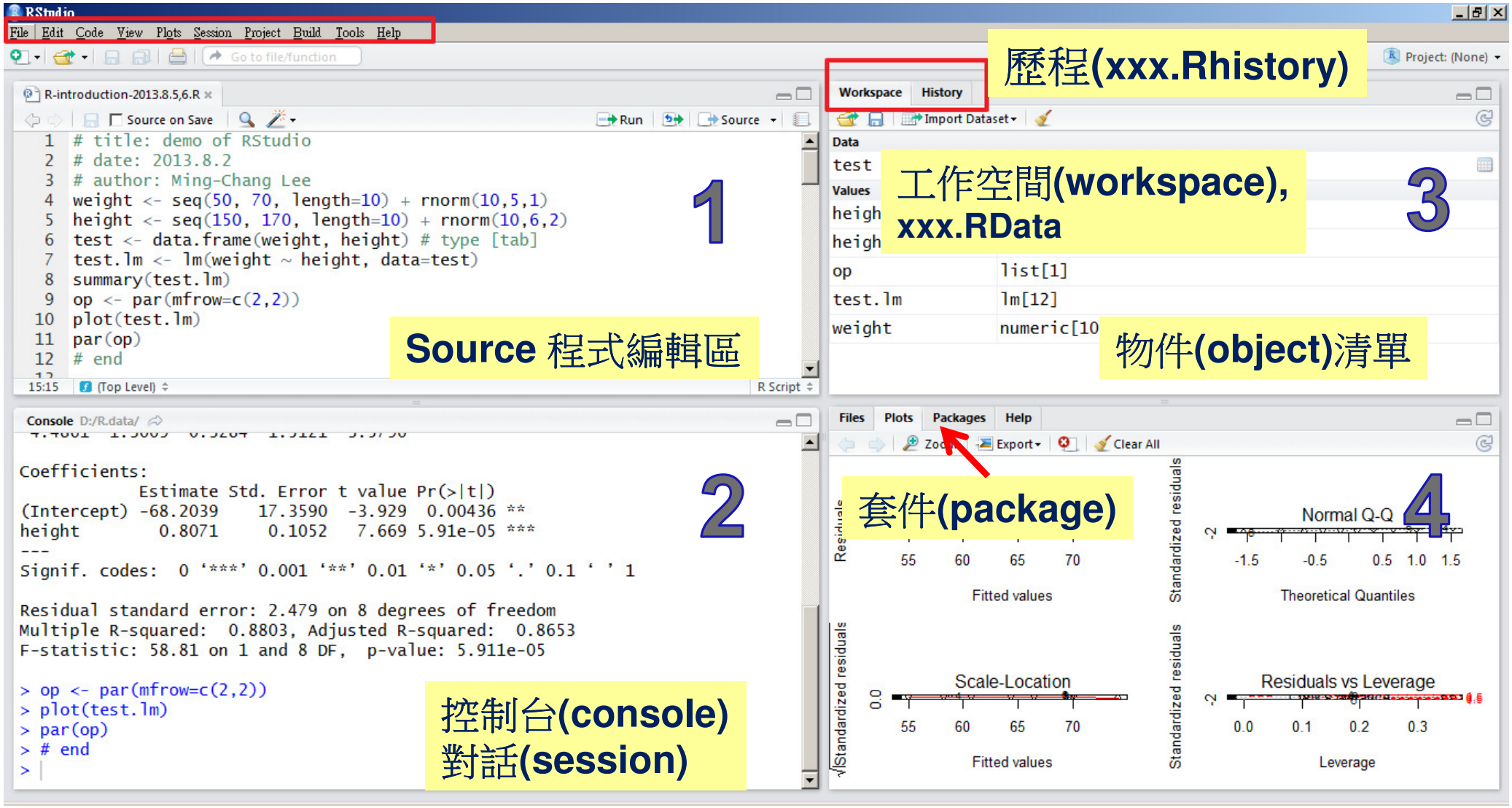


RStudio - Options

- Appearance \ Editor theme
- Default: TextMate



RStudio - Environment



CTRL + SHIFT + F10: 重新啟動R

RStudio - Keyboard Shortcuts

Keyboard Shortcuts	Feature
Ctrl + Shift + N	Create new script
Ctrl + S	Save
Ctrl + Shift + R	Insert section
Alt + -	Insert assignment operator
Ctrl + Shift + C	Comments
Ctrl + Enter	Run selected lines
Ctrl + Shift + F10	Restart R
Alt + Shift + K	Shortcuts reference

R + Editor

- R – Naïve environment
- RStudio – IDE
- Eclipse
  - StatET 3.4.1:
    An Eclipse based IDE (integrated development environment) plug-in for R.
  - http://www.walware.de/goto/statet

- R Tools for Visual Studio
  - https://docs.microsoft.com/zh-tw/visualstudio/rtvs/installing-r-tools-for-visual-studio

Packages 套件

Packages

- Shared R codes
- Components:
  - R codes
  - R functions
  - Datasets
  - Help document
- Basic packages: 30
- Installed directory: \R-3.5.1\library

Packages

- **Step 1: install.packages**("PackageName")
- **Step 2: library**(PackageName)
- Install and load "e1071" (machine learning)

```
> install.packages("e1071")
trying URL 'http://cran.cs.pu.edu.tw/bin/windows/contrib/3.0/e1071_1.6-1.zip'
Content type 'application/zip' length 514468 bytes (502 Kb)
opened URL
downloaded 502 Kb

package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\Administrator\AppData\Local\Temp\RtmpoHSOAk\downloaded_packages
> library(e1071)
Loading required package: class
>
```

# loaded packages
search()

Packages - 38 views

(2018.11.18)

Contributed Packages

Available Packages

Currently, the CRAN package repository features 13398 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

sorted by name

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.

CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 38 views are available.

38 views

Packages - 38 views

RWEPA → task

2013年10月8日 星期二

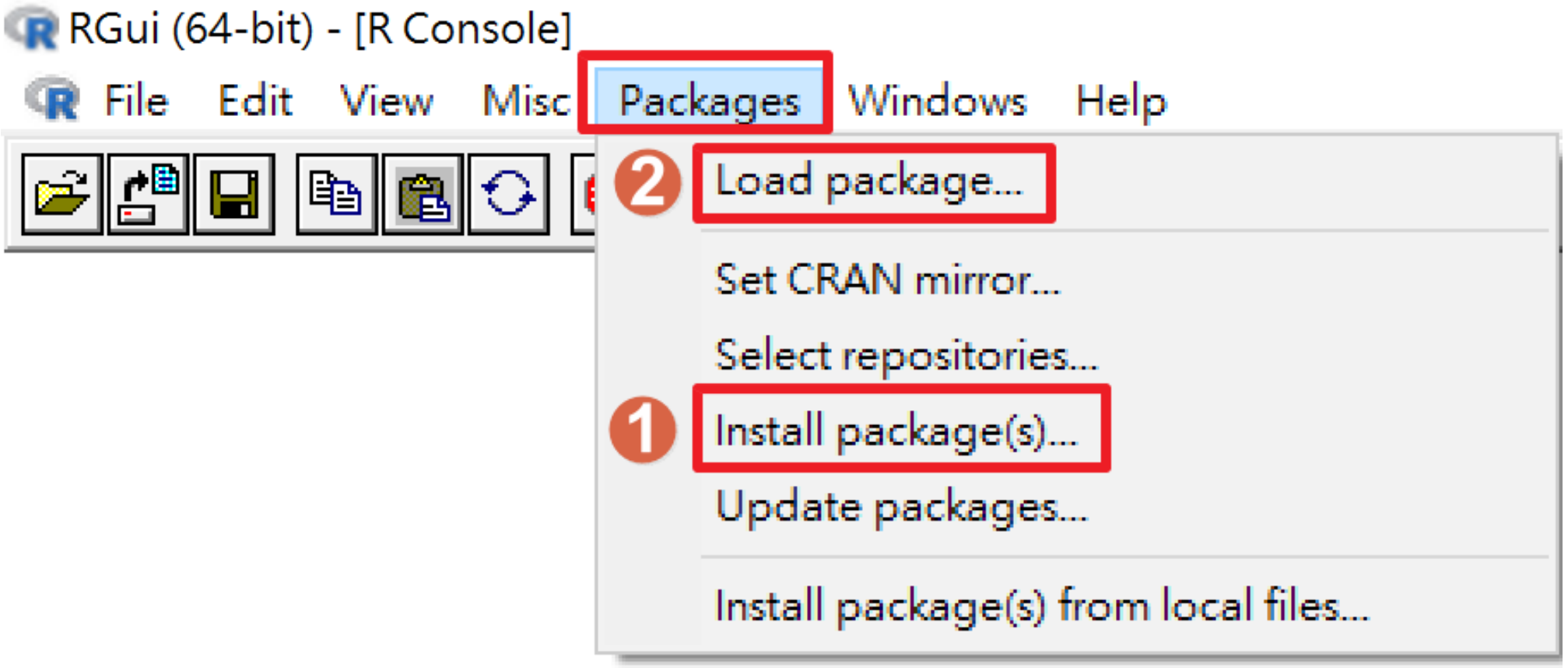
Task Views - R套件區分成38個類別

更新日期: 2018.11.18
CRAN Task View : https://cran.r-project.org/web/views/
套件區分成 38 個類別，中文說明如下:

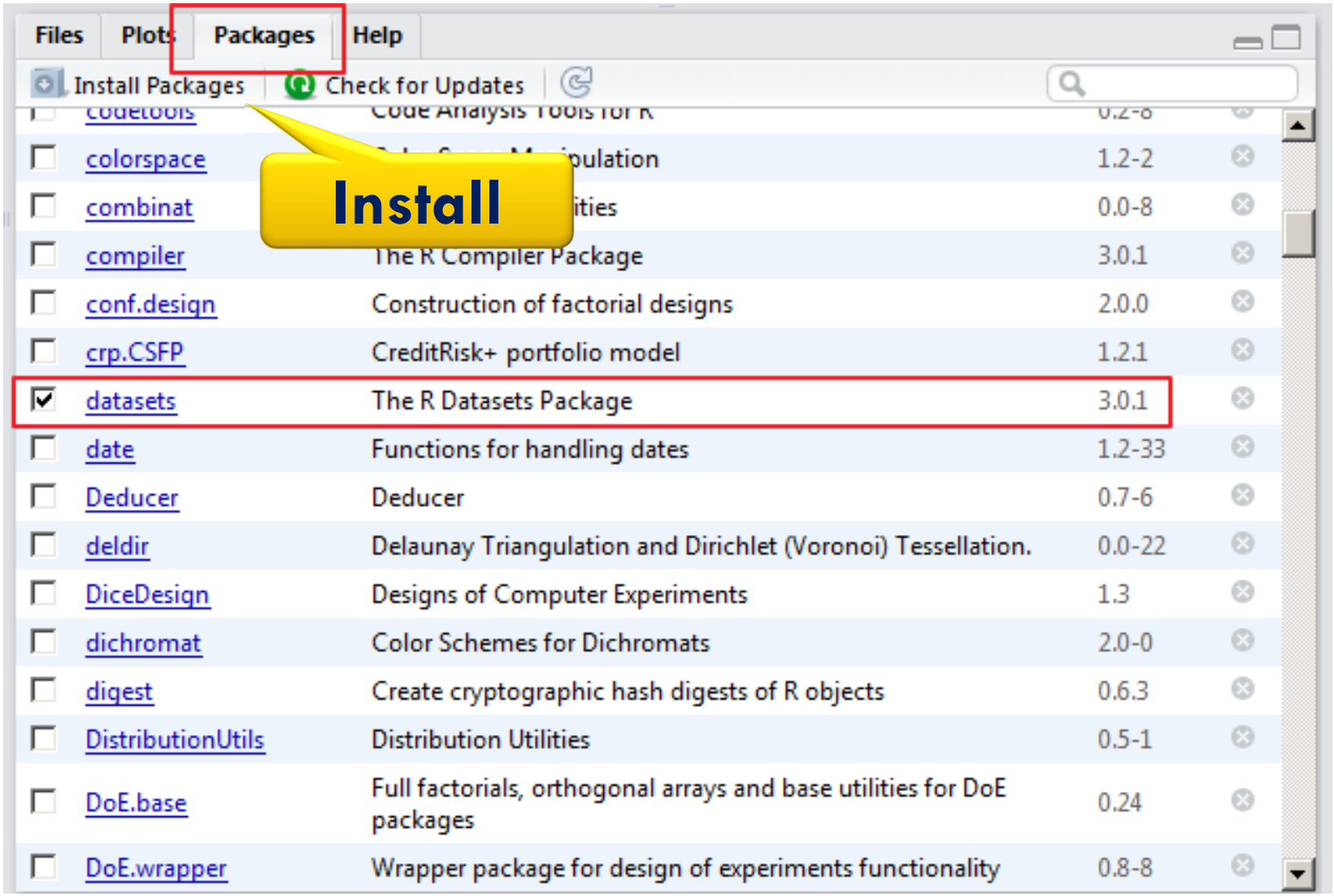
編號	主題	英文說明	中文說明
01	**Bayesian**	Bayesian Inference	貝氏統計
02	**ChemPhys**	Chemometrics and Computational Physics	計量化學學, 計算物理
03	**ClinicalTrials**	Clinical Trial Design, Monitoring, and Analysis	臨床試驗設計, 監測和分析
04	**Cluster**	Cluster Analysis & Finite Mixture Models	群集分析, 有限混合模型
05	**Databases**	Databases with R	R與資料庫連接
06	**DifferentialEquations**	Differential Equations	微分方程
07	**Distributions**	Probability Distributions	機率分配
08	**Econometrics**	Computational Econometrics	計量經濟
09	**Environmetrics**	Analysis of Ecological and Environmental Data	生態, 環境資料分析

```
# Installed list
x <- installed.packages()
x
```

R - Packages


RGui (64-bit) - [R Console]
File
Edit
View
Misc
Packages
Windows
Help
2
Load package...
Set CRAN mirror...
Select repositories...
1
Install package(s)...
Update packages...
Install package(s) from local files...

RStudio - Packages

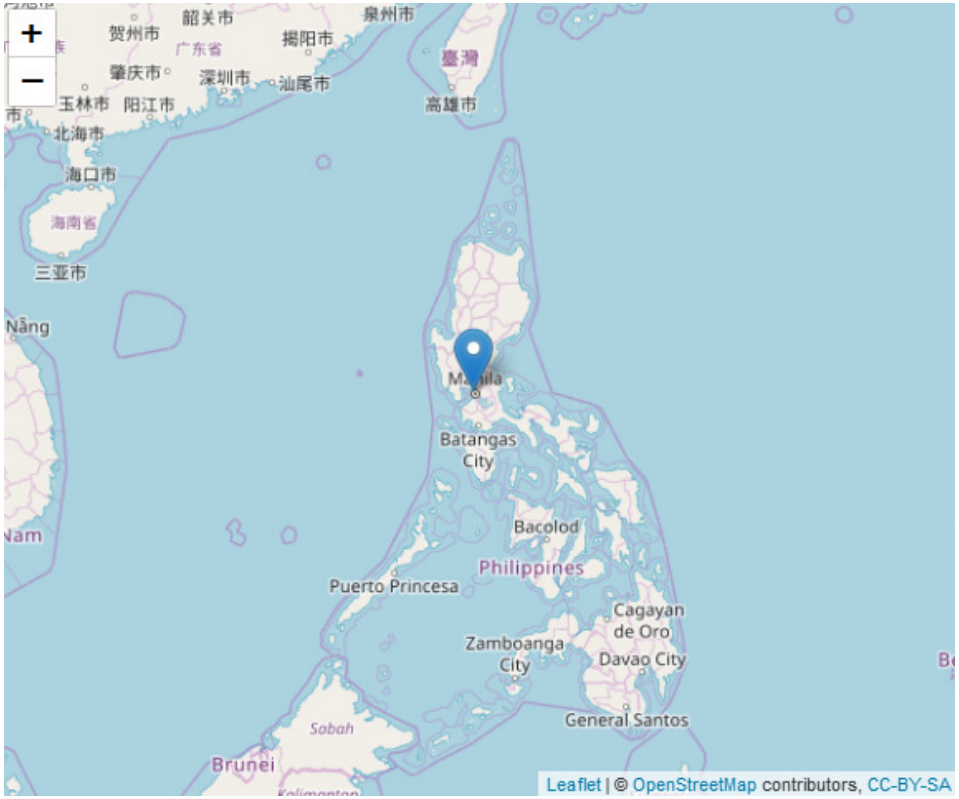


☐	codetools	Code Analysis Tools for R	0.2-8
☐	colorspace	...pulation	1.2-2
☐	combinat	...ities	0.0-8
☐	compiler	The R Compiler Package	3.0.1
☐	conf.design	Construction of factorial designs	2.0.0
☐	crp.CSFP	CreditRisk+ portfolio model	1.2.1
☑	datasets	The R Datasets Package	3.0.1
☐	date	Functions for handling dates	1.2-33
☐	Deducer	Deducer	0.7-6
☐	deldir	Delaunay Triangulation and Dirichlet (Voronoi) Tessellation.	0.0-22
☐	DiceDesign	Designs of Computer Experiments	1.3
☐	dichromat	Color Schemes for Dichromats	2.0-0
☐	digest	Create cryptographic hash digests of R objects	0.6.3
☐	DistributionUtils	Distribution Utilities	0.5-1
☐	DoE.base	Full factorials, orthogonal arrays and base utilities for DoE packages	0.24
☐	DoE.wrapper	Wrapper package for design of experiments functionality	0.8-8

leaflet package

https://rstudio.github.io/leaflet/

```
library(leaflet)
m <- leaflet() %>%
  addTiles() %>%
  addMarkers(lng=120.973551, lat=14.593105, popup="Home") %>%
  setView(lng = 120.973551, lat = 14.593105, zoom = 5)
m  # Print the map
```



Data type

- Integer
- Numeric
- Character
- Logical

Date type - integer

```
> # 整數
> x1 <- c(1:10)
> x1
 [1]  1  2  3  4  5  6  7  8  9 10
> typeof(x1)
[1] "integer"
> is.integer(x1)
[1] TRUE
> is.numeric(x1)
[1] TRUE
> class(x1)
[1] "integer"
```

Data type - numeric

```
> # 數值
> x2 <- c(1.1, 1.3, 1.5, 1.7, 2)
> x2
[1] 1.1 1.3 1.5 1.7 2.0
> typeof(x2)
[1] "double"
> is.integer(x2)
[1] FALSE
> is.numeric(x2)
[1] TRUE
> class(x2)
[1] "numeric"
```

Data type - character

```
> # 字串資料
> x3 <- c("台北市", "新北市", "台中市", "台南市", "高雄市")
> x3
[1] "台北市" "新北市" "台中市" "台南市" "高雄市"
> typeof(x3)
[1] "character"
> is.character(x3)
[1] TRUE
> class(x3)
[1] "character"
> x4 <- c(1, 2.3, "巨量資料")
> x4
[1] "1"        "2.3"      "巨量資料"
> class(x4)
[1] "character"
```

force transfer to character

Data type - logical

- 邏輯值包括: TRUE 或 FALSE
- TRUE→ 1
- FALSE→ 0

```
> # 邏輯值
> TRUE*2
[1] 2
> FALSE*3
[1] 0
```

Data type - logical

```
> x5 <- x2 >= 1.4
> x5
[1] FALSE FALSE  TRUE  TRUE  TRUE
> typeof(x5)
[1] "logical"
> is.integer(x5)
[1] FALSE
> is.numeric(x5)
[1] FALSE
> is.character(x5)
[1] FALSE
> is.logical(x5)
[1] TRUE
> class(x5)
[1] "logical"
```

Data object

Data object

向量 vector

北部	中部	南部

矩陣 matrix

1	3	5
2	4	6

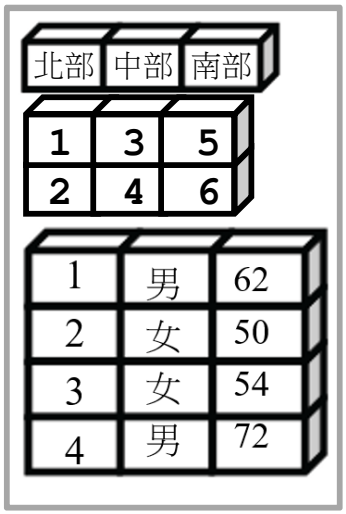
陣列 array

1.1	4.4	7.7
2.2	5.5	8.8
3.3	6.6	9.9

資料框 data.frame

1	男	62
2	女	50
3	女	54
4	男	72

串列 list

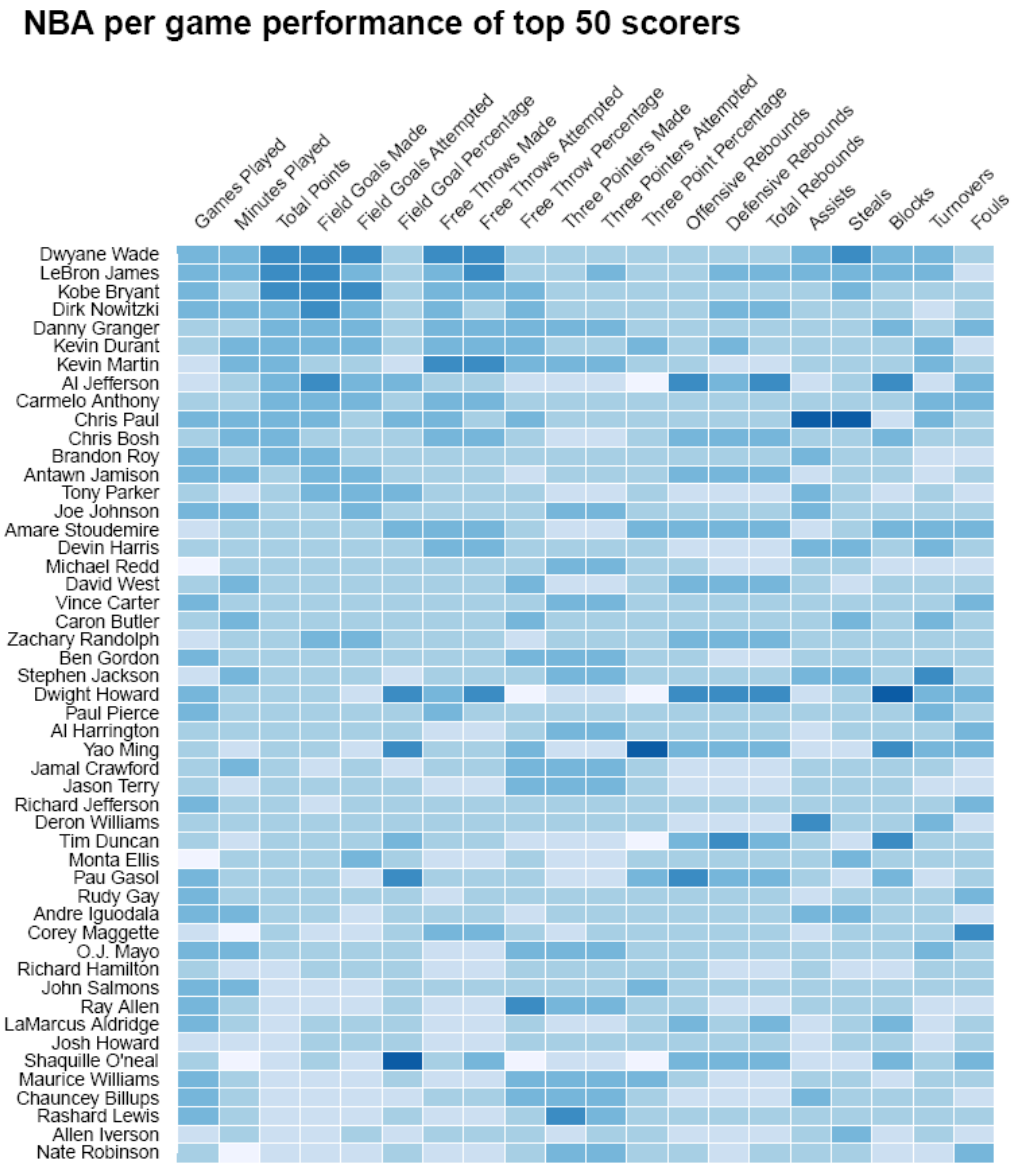


vector

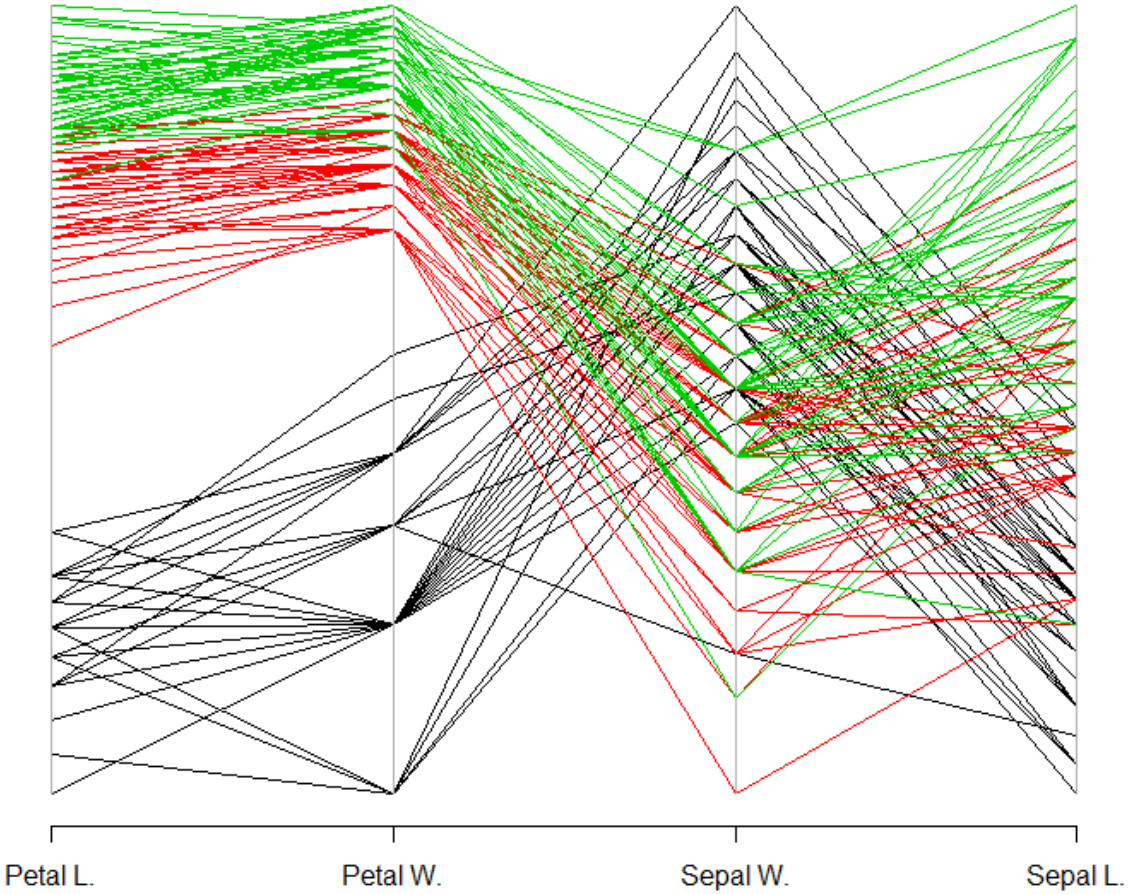
matrix

data.frame

image {graphics}

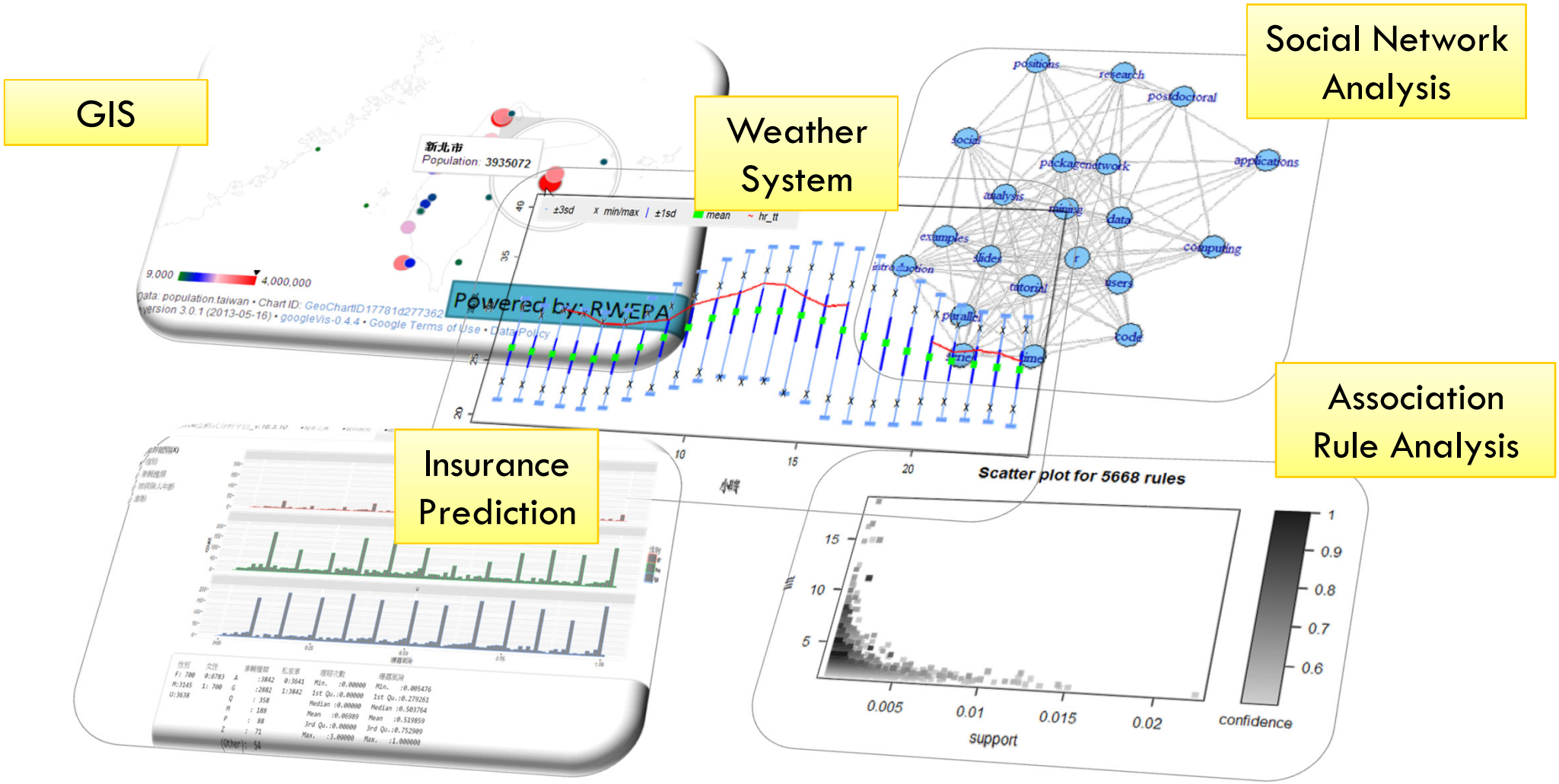


parcoord {MASS}
Parallel coordinates



3. Big Data Application

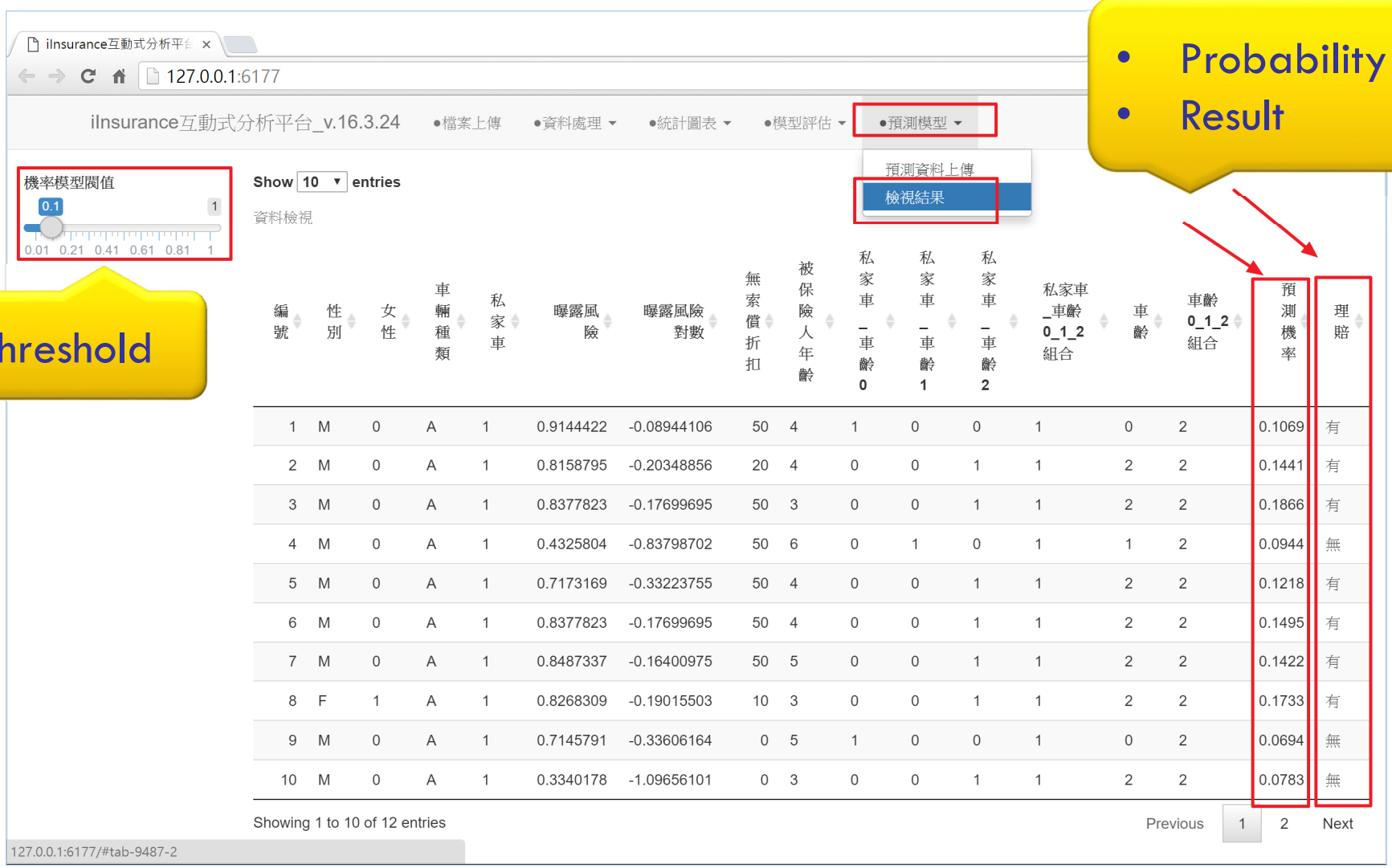
Big Data Application



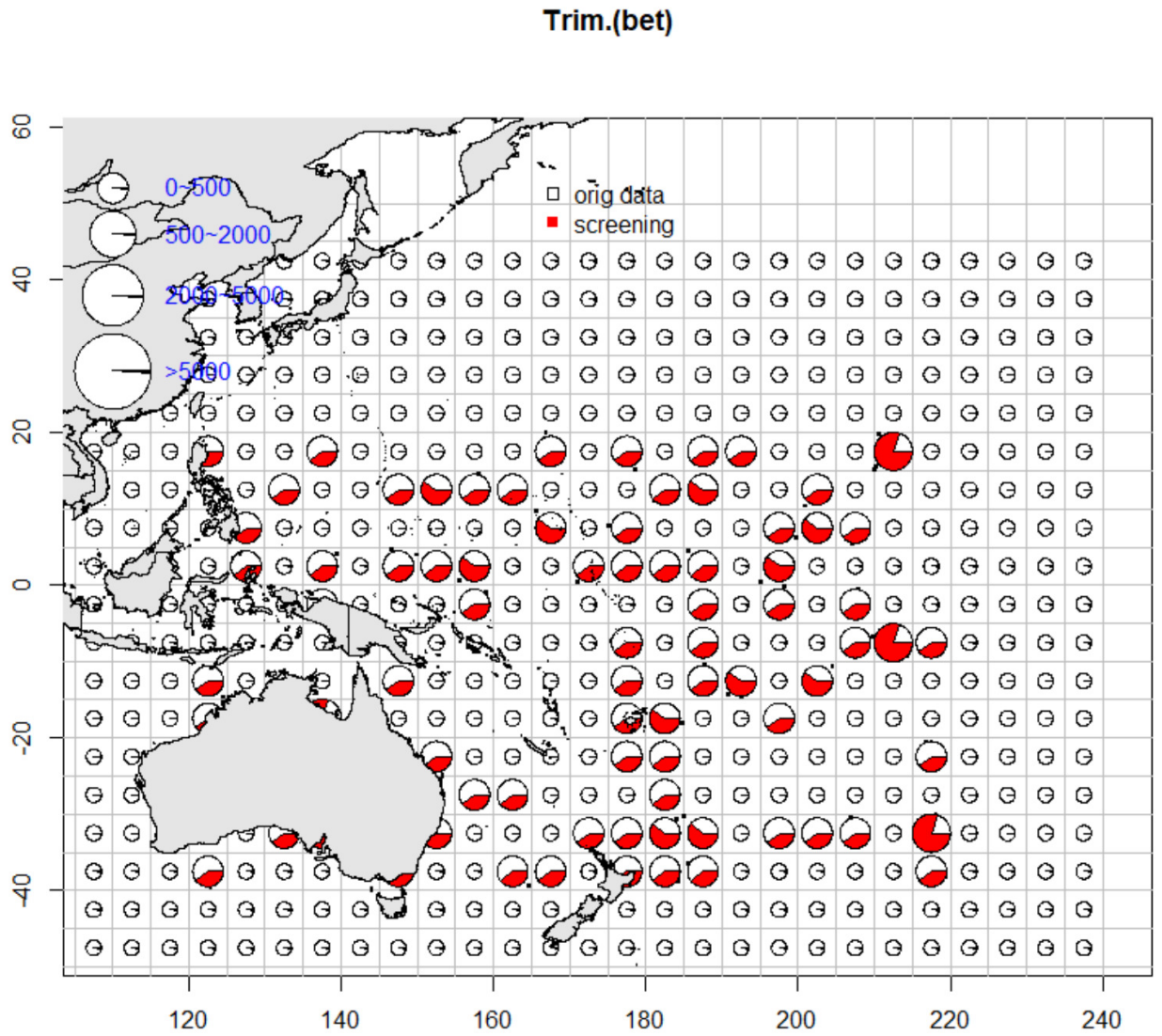
Weather System (data size=16 Million)

http://rwepa.ddns.net:3838/sample-apps/cwb_vis_qc/

Insurance Forecasting



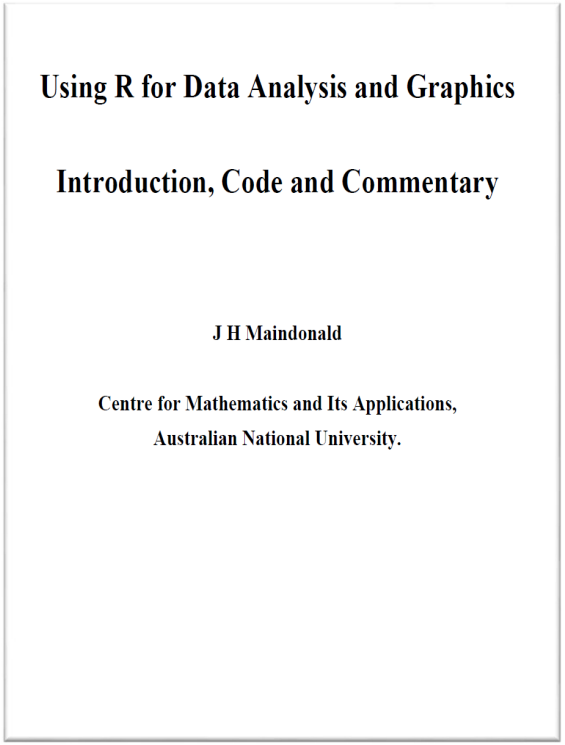
Space Pie Chart Detector



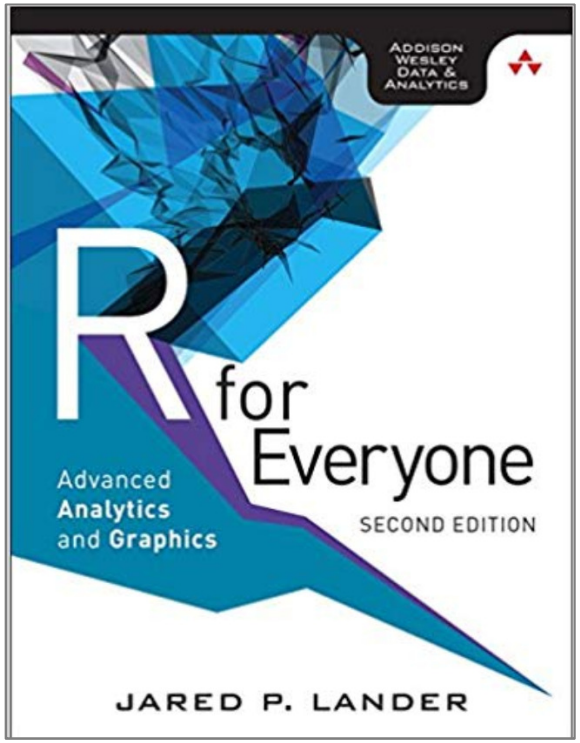
Open Data

- OPEN DATA TAIWAN
  - https://data.gov.tw/
- OPEN DATA PHILIPPINES
  - https://data.gov.ph/
- UCI Machine Learning Repository
  - https://archive.ics.uci.edu/ml/datasets.html

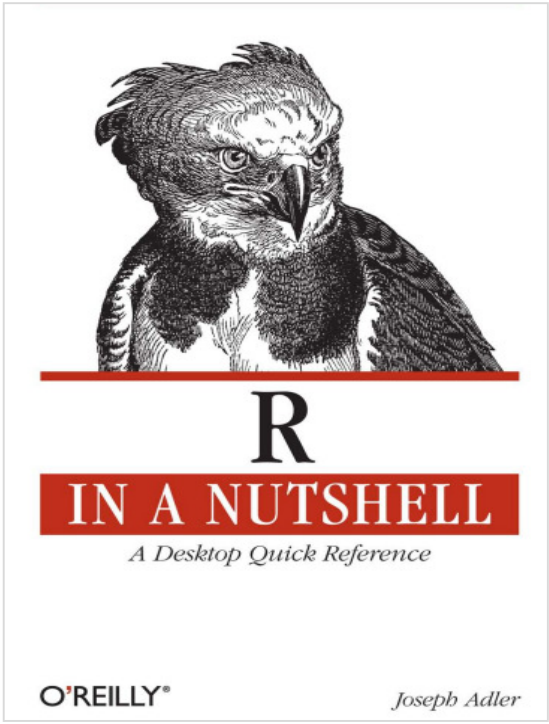
References



(Basic)



(Intermediate)



(Advanced)

Thank you

alan9956@gmail.com

http://rwepa.blogspot.tw/