

R for Bioinformatics

Introduction, Programming, Data Analysis and
Visualization

Introduction to Data Analysis and R

Gang Chen
chengang@bgitechsolutions.com

November 1, 2013

Outline

- 1 Data Analysis
- 2 Data Aanlysis and R
- 3 Hello R!
- 4 Development Environment
- 5 References

Next

- 1 Data Analysis
- 2 Data Aanlysis and R
- 3 Hello R!
- 4 Development Environment
- 5 References

Data Analysis

Wikipedia

Analysis of data is a process of **inspecting**, **cleaning**, **transforming**, and **modeling** data with the goal of discovering useful information, suggesting conclusions, and supporting decision making.

Data Analysis

- 1 Collecting →
- 2 Cleaning →
- 3 Transforming →
- 4 Modeling →
- 5 Visualizing →
- 6 Knowledge

Biological Data Analysis

- 1 Sequencing →
- 2 QC →
- 3 Alignment ...→
- 4 GWAS, EWAS ...→
- 5 Manhattan Plot, Q-Q plot ...→
- 6 Paper?

Next

- 1 Data Analysis
- 2 Data Analysis and R**
- 3 Hello R!
- 4 Development Environment
- 5 References

What is R?

R

R is a **free** software environment for statistical computing and graphics.

----R-project.org

Data Analysis

- ① Collecting →
- ② Cleaning →
- ③ Transforming →
- ④ Modeling →
- ⑤ Visualizing →
- ⑥ Knowledge

R

- **Rcurl...**
- **gsub, unique...**
- **reshape...**
- **e1071...**
- **ggplot2...**
- **knitr...**

Biological Data Analysis and R

Biological Data Analysis

- 1 Sequencing →
- 2 QC →
- 3 Alignment ...→
- 4 GWAS, EWAS ...→
- 5 Manhattan Plot, Q-Q plot ...→
- 6 Paper?

R

- **Rsamtools, Affy...**
- **genomicRanges...**
- **reshape...**
- **e1071...**
- **ggbio...**
- **knitr, shiny...**

History

- Version 0.16, This is the last alpha version developed primarily by Ihaka and Gentleman. Much of the basic functionality from the "White Book" (see S history) was implemented. The mailing lists commenced on April 1, 1997.
- Version 0.49, April 23, 1997, This is the oldest available source release, and compiles on a limited number of Unix-like platforms. CRAN is started on this date, with 3 mirrors that initially hosted 12 packages. Alpha versions of R for Microsoft Windows and Mac OS are made available shortly after this version.
- Version 0.60, December 5, 1997, R becomes an official part of the GNU Project. The code is hosted and maintained on CVS.

History

- Version 1.0.0, February 29, 2000, Considered by its developers stable enough for production use.[28]
- Version 1.4.0, S4 methods are introduced and the first version for Mac OS X is made available soon after.
- Version 2.0.0, October 4, 2004, Introduced lazy loading, which enables fast loading of data with minimal expense of system memory.
- Version 2.1.0, Support for UTF-8 encoding, and the beginnings of internationalization and localization for different languages.
- Version 2.11.0, April 22, 2010, Support for Windows 64 bit systems.
- Version 2.13.0, April 14, 2011, Adding a new compiler function that allows speeding up functions by converting them to byte-code.

History

- Version 2.14.0, October 31, 2011, Added mandatory namespaces for packages. Added a new parallel package.
- Version 2.15.0, March 30, 2012, New load balancing functions. Improved serialization speed for long vectors.
- Version 3.0.0, April 3, 2013, Support for numeric index values 231 and larger on 64 bit systems.

R in China

- 2004, official documents are translated into Chinese
- 2006, some books on R in Bioinformatics
- 2008, the first R conference was hold at Renming University, Beijing.
- 2009 to 2013, China R Conference is hold at Beijing and Shanghai each year.
- 2012, popular R books are translated into Chinese.
- 2013, R in Action ggplot2 R in a nutshell ...are published in China.
- 2013, CUHK-R course is launched.

Applications of R

Applications

- Statistical analysis
- Data Mining
- Life Science
- Business Intelligence
- Data Visualization
- Social Network
- eCommerce
- Integrated Circuit
- Financial
- Media
- Consulting
- ...

Attendee of R community



Facultad de Ciencias Económicas y Empresariales
www.uclm.es/ab/fcee



Pros and Cons

Bo Cowgill, Google

"The best thing about R is that it was developed by statisticians. The worst thing about R is that ... it was developed by statisticians."

Next

- 1 Data Analysis
- 2 Data Aanlysis and R
- 3 Hello R!**
 - Hello R!
 - Hello Statistical Analysis!
 - Hello Plot!
- 4 Development Environment
- 5 References

Hello R!

Hello R!

Hello R!

```
print("Hello R!")
```

```
## [1] "Hello R!"
```

Hello R!

Hello Statistical Analysis!

Hello Statistical Analysis

```
data(mtcars)
cor(mtcars$mpg, mtcars$wt)

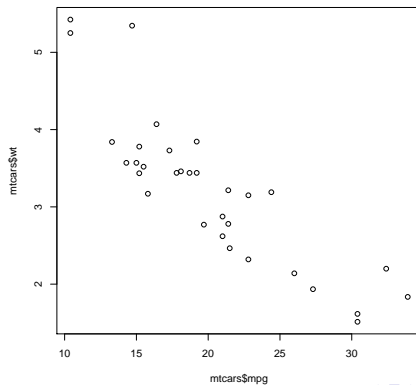
## [1] -0.8677
```

Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)  
plot(mtcars$mpg, mtcars$wt)
```

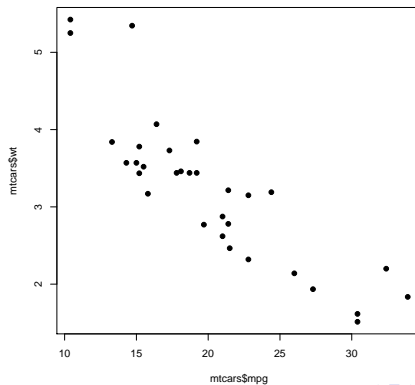


Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)  
plot(mtcars$mpg, mtcars$wt, pch = 19)
```

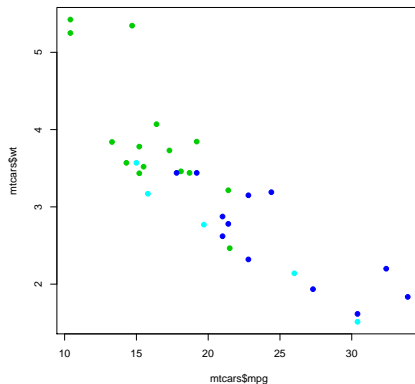


Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)  
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear)
```

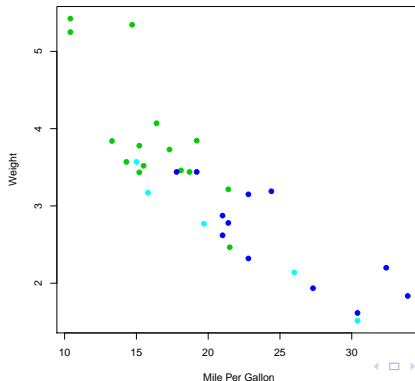


Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mile Per Gallon",
     ylab = "Weight")
```

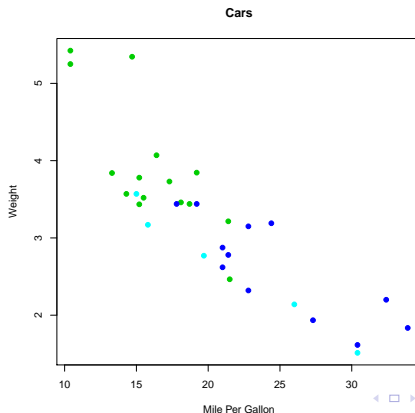


Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mile Per Gallon",
     ylab = "Weight", main = "Cars")
```

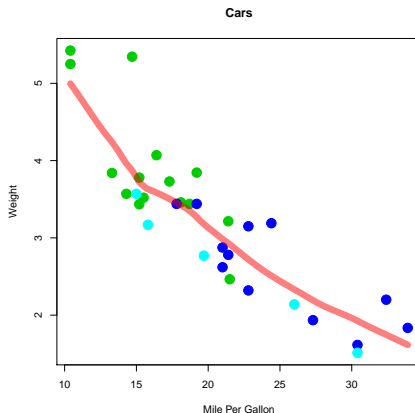


Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mile Per Gallon",
     ylab = "Weight", main = "Cars", cex = 2)
lines(loess.smooth(mtcars$mpg, mtcars$wt), col = rgb(1, 0, 0, 0.5), lwd = 10)
```

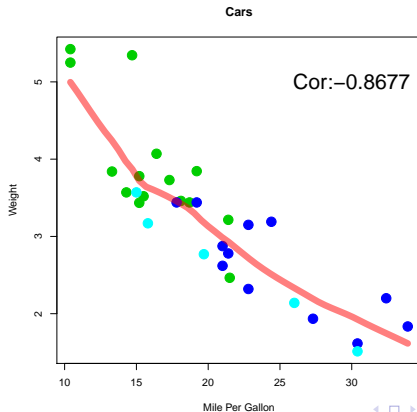


Hello R!

Hello Plot!

Hello Plot

```
data(mtcars)
plot(mtcars$mpg, mtcars$wt, pch = 19, col = mtcars$gear, xlab = "Mile Per Gallon",
     ylab = "Weight", main = "Cars", cex = 2)
lines(loess.smooth(mtcars$mpg, mtcars$wt), col = rgb(1, 0, 0, 0.5), lwd = 10)
text(30, 5, paste("Cor:", round(cor(mtcars$mpg, mtcars$wt), 4), sep = ":"), cex = 2)
```



Next

- 1 Data Analysis
- 2 Data Aanlysis and R
- 3 Hello R!
- 4 Development Environment**
 - Obtaining and installing R
 - R in Command Line
 - Editors and IDEs
- 5 References

Download and Installation

Download

CRAN

Installation

- R: Linux(apt, yum), Mac OS, Windows
- Rtools: Windows
- packages: CRAN, devtools, github, local file

R Commands

R Commands

```
R CMD command args
```

```command``:`

**INSTALL** Install add-on packages

**REMOVE** Remove add-on packages

**BATCH** Run R in batch mode

# R Command Options

## R Command Options

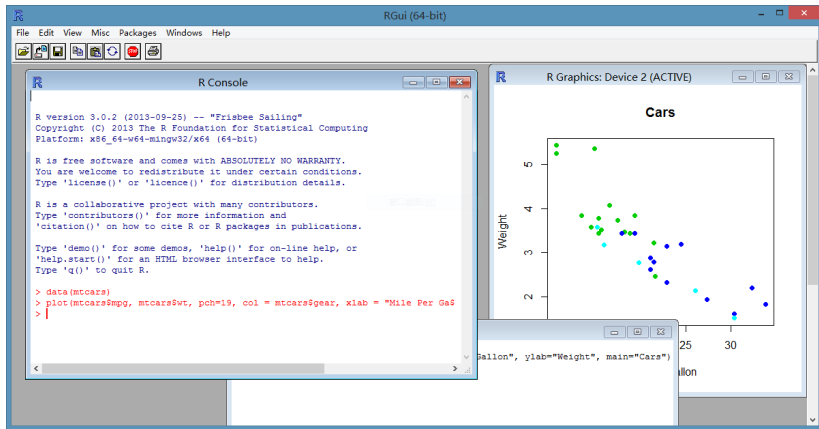
- h, --help Print usage message and exit
- version Print version information and exit
- save Do save workspace at the end of the session
- no-save Don't save it
- restore Do restore previously saved objects
- no-restore Don't restore anything
- vanilla Combine --no-save, --no-restore, --no-site-file, --no-init-file and --no-environ
- f file, --file=file Take input from `file`
- e expression Use `expression` as input

# Editors and IDEs

## Editors

- R terminal
- Rgui
- VIM + Vim-R-plugin
- Emacs + ESS
- Notepad++ + NppToR
- ...

# R Terminal and Rgui





# R Terminal and Rgui

## R

- Ctrl + R : run
- Tab: auto complete
- arrow up and down: history

## R and Texteditor

- copy and paste
- `source("source.R")`

## source

```
sourceDir <- function(path, trace = TRUE, ...) {
 for (nm in list.files(path, pattern = "[.] [RrSsQq]$", ...)) {
 if (trace)
 cat(nm, " :")
 source(file.path(path, nm), ...)
 if (trace)
 cat("\n")
 }
}
```

# VIM + Vim-R-plugin

```

script2.R zzz.R RNA-Seq.R
fmcS-manual x, Row = NULL, Col = if (symm) "Row" else NULL, distfun, hclust
fmcS-manual
fmcS-manual
Rdlatex.log [Scratch] [Preview] 1,1 All
rpackage.Rche y <- matrix(rnorm(50), 10, 5, dimnames=list(paste("g", 1:10, sep="
00_pkg_src/ ## Row clustering
mypackage/ hr <- hclust(as.dist(1-cor(t(y), method="pearson")), method="compl
00check.log ## Column clustering
00install.o hc <- hclust(as.dist(1-cor(y, method="spearman")), method="comple
mypackage-Ex ## Plot heatmap
mypackage-Ex heatmap.2(y, Row=as.dendrogram(hr), Col=as.dendrogram(hc), scale
mypackage-Ex ## Return matrix with row/column sorting as in heatmap
rpackage/ y[rev(hr$labels[hr$order]), hc$labels[hc$order]]
man/ heatmap.colors
l-colAg.Rd heatmap function grDevices
mypackage- heatmap function stats ample
R/ fmcS(sdfset[[1]], sdfset[[2]], fast=T)
myfct.R result <- fmcS(sdfset[[1]], sdfset[[2]])
DESCRIPTION mcs <- fmcS(sdfset[[1]], sdfset[[2]], au=2, bu=1, matching.mode="a
NAMESPACE mcs
Read-and-del script2.R [+] 12,1 33%
fmcS_1.0.tar
litter.png
matrix.xls
myfct.R
rpackage_1.0.
otes.R
verLapper.R
angeoverlap
IA-Seq.R
cript1.R
cript2.R
FStreamer.R
est.sdf
est.svg
ips_and_trick
zzz.R

```

```

X g9 0.1416941
g10 0.5772262 0.3061073
> as.matrix(c)[1:4,1:4]
 g1 g2 g3 g4
g1 1.0000000 -0.7240061 0.8050921 0.2327069
g2 -0.7240061 1.0000000 -0.5586679 -0.7823333
g3 0.8050921 -0.5586679 1.0000000 0.2567203
g4 0.2327069 -0.7823333 0.2567203 1.0000000
> y
 t1 t2 t3 t4 t5
g1 -0.2608109 -2.1287458 0.5436205 -0.1962956 0.5136432
g2 -2.0478162 -0.2318061 -2.1907113 -0.9185012 -1.1459074
g3 -0.1814785 -0.5137189 1.2004188 -0.2185163 0.9562711
g4 0.2493454 -0.5782053 0.7562372 -0.6441311 -1.0792957
g5 0.1082261 -1.8310231 -0.3319702 0.5535095 0.0165956
g6 0.2596634 -0.8048402 -0.3751721 -0.6061271 -1.4533725
g7 0.4497986 -0.6475571 1.1905096 1.2794214 0.1432148
g8 -1.0501454 -0.3717143 0.2831488 -1.6238084 0.3429913
g9 -0.7831244 0.8490208 1.1253892 -0.4341535 0.6912465
g10 -1.7273262 0.3621398 2.2920425 -0.9175735 -1.6735589

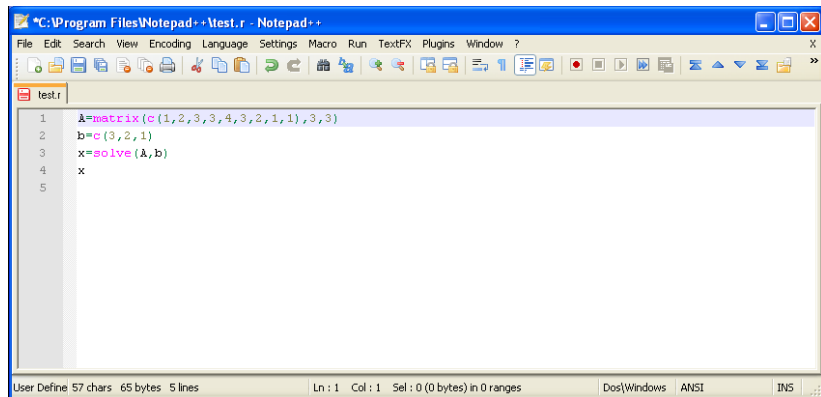
```

```

heatmap.2 package:gplots R Documente
n
Enhanced Heat Map
Description:
A heat map is a false color image (basically 'image(t(x))') w
a
dendrogram added to the left side and/or to the top. Typica
reordering of the rows and columns according to some set of v
es
(row or column means) within the restrictions imposed by the
dendrogram is carried out.

```

# Notepad++ + NppToR

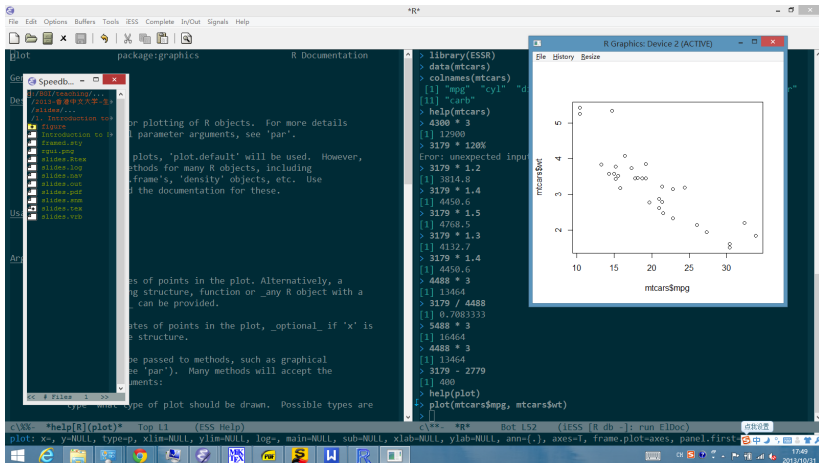


The screenshot shows the Notepad++ application window. The title bar reads "\*C:\Program Files\Notepad++\test.r - Notepad++". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Macro, Run, TextFX, Plugins, Window, and ?. The toolbar contains various icons for file operations, editing, and development. The editor area shows a file named "test.r" with the following R code:

```
1 A=matrix(c(1,2,3,3,4,3,2,1,1),3,3)
2 b=c(3,2,1)
3 x=solve(A,b)
4 x
5
```

The status bar at the bottom displays: "User Define 57 chars 65 bytes 5 lines", "Ln : 1 Col : 1 Sel : 0 (0 bytes) in 0 ranges", "Dos\Windows", "ANSI", and "INS".

# Emacs + ESS



# Emacs + ESS

## What is ESS?

Emacs Speaks Statistics (ESS) is an add-on package for emacs text editors such as **GNU Emacs** and XEmacs. It is designed to support editing of scripts and interaction with various statistical analysis programs such as **R**, S-Plus, SAS, Stata and JAGS.

## ESS Website

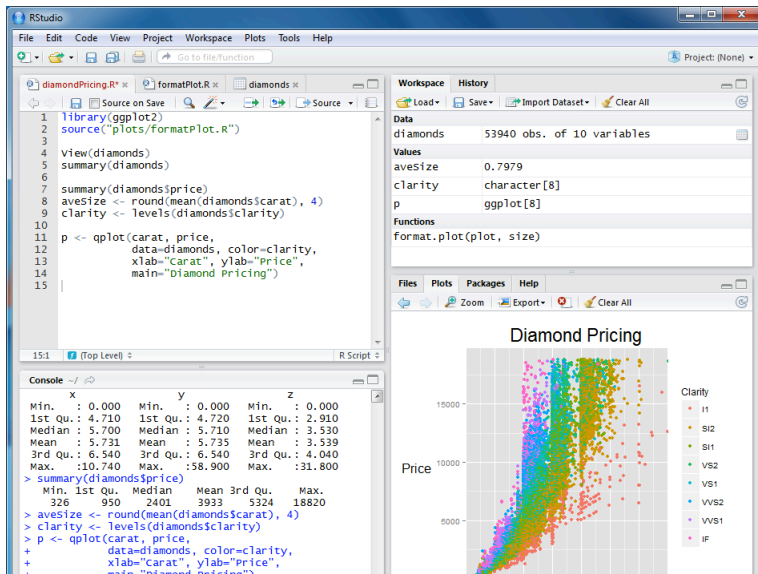
<http://ess.r-project.org/>

# IDEs

## IDEs

- RStudio: local and cloud-based
- TinnR
- StatET: eclipse for R
- ...

# RStudio



# Next

- 1 Data Analysis
- 2 Data Aanlysis and R
- 3 Hello R!
- 4 Development Environment
- 5 References**



# Books

- R in action (also in Chinese)
- Introduction to R (also in Chinese)
- R for beginner (also in Chinese)
- R in a Nutshell (Chinese version is in press)
- The art of R programming (also in Chinese)
- ggplot2. Elegant Graphics for Data Analysis (also in Chinese)

# Websites

- R-project and CRAN
- COS.name (Chinese)
- Quick-R
- <http://had.co.nz/>, Hadley Wickham
- Twitter, github, RForge
- Google

# Websites

- R-project and CRAN
- COS.name (Chinese)
- Quick-R
- <http://had.co.nz/>, Hadley Wickham
- Twitter, github, RForge
- Google Baidu?

# Journals

- The R Journal
- Journal of Statistical Software
- BMC Bioinformatics: Software
- Bioinformatics: Application Note