



JIGSAW ACADEMY

Analytics for Professionals

INTRODUCTION TO R



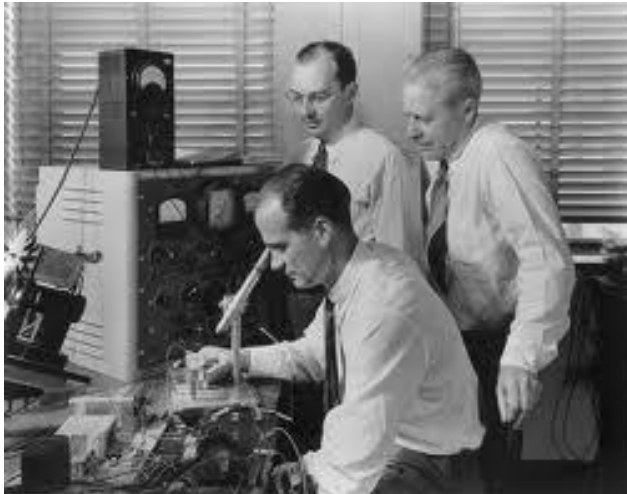
AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why use R?
- References for learning R



HISTORY AND EVOLUTION OF R

Origin in the Bell Labs in the 1970's



Bell Laboratories



HISTORY AND EVOLUTION OF R

R has developed from the S language

S Version 1

S Version 2

S Version 3

S Version 4

Developed 30 years ago for research
applied to the high-tech industry





HISTORY AND EVOLUTION OF R

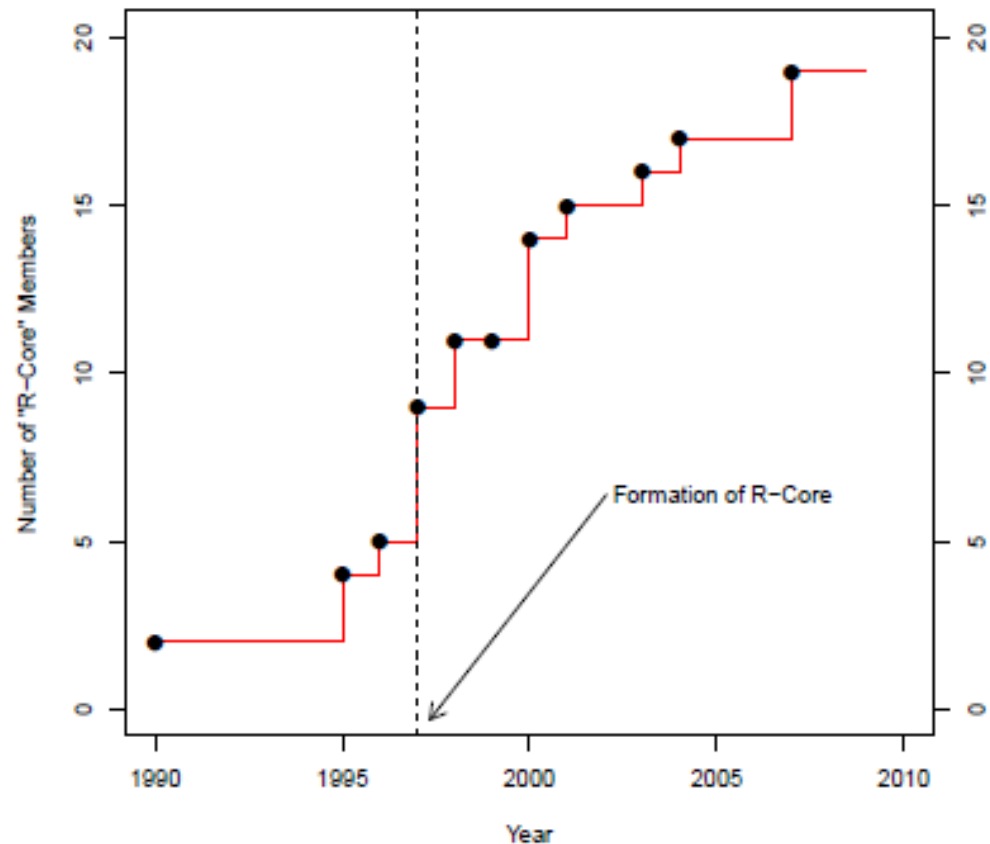
The regular development of R

1990's: R developed concurrently with S

1993: R made public

Acceleration of R development

- R-Help and R-Devl mailing-lists
- Creation of the R Core Group



Source: R Journal Vol 1/2



HISTORY AND EVOLUTION OF R

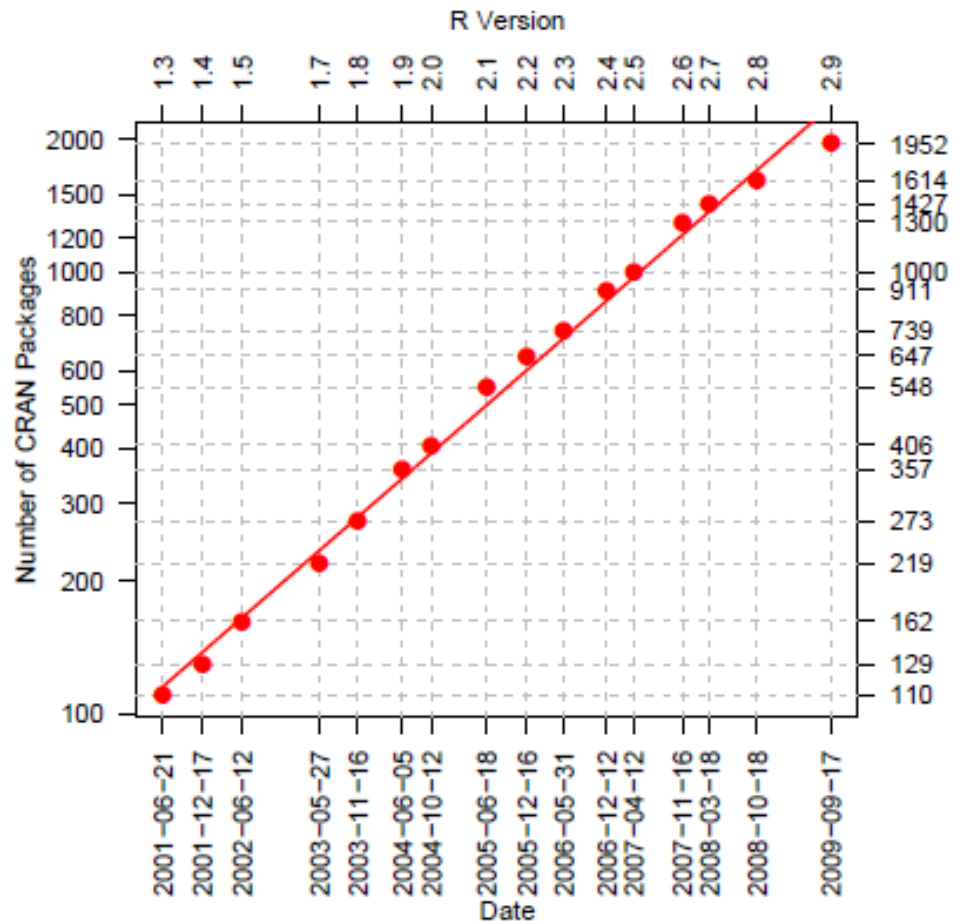
Growing number of packages

2001: ~100 packages

2009: Over 2000 packages

2000: R version 1.0.1

Today: R version 2.14





HISTORY AND EVOLUTION OF R

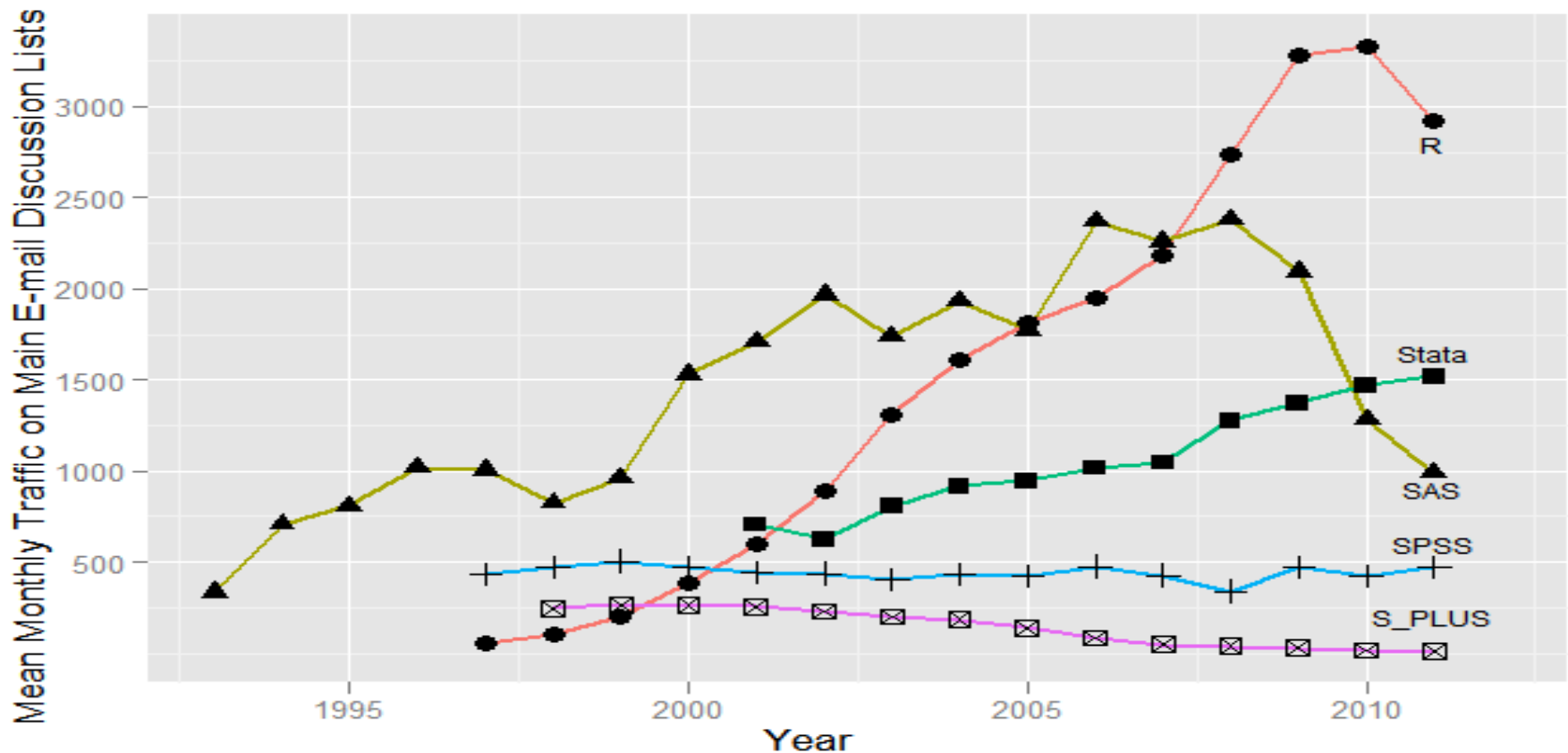
Explosion of R popularity in the last decade

- Object-oriented, growing user base, scripting features
- Free and open-source
- Irrational reasons: R seen as « cool »



HISTORY AND EVOLUTION OF R

Comparison of Mailing Lists
















Evolution of the traffic on software main mailing-lists. Source: R.A. Muenchen, r4stats.com



HISTORY AND EVOLUTION OF R

Popularity amongst programming languages

What programming languages you used for data mining / data analysis in the past 12 months? [570 voters]

R (257)	 45%
SQL (184)	 32%
Python (140)	 25%
Java (139)	 24%
SAS (121)	 21%
MATLAB (83)	 15%
C/C++ (73)	 13%
Unix shell/awk/gawk/sed (59)	 10%
Perl (45)	 7.9%
Hadoop/Pig/Hive (35)	 6.1%
Lisp (4)	 0.7%
Other (70)	 12.0%
None (7)	 1.2%



HISTORY AND EVOLUTION OF R

Number of Blogs

Software	Number of Blogs
R	365
SAS	40
Stata	8
Others	0-3

Data as on Mar 2012



AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why using R?
- References for learning R

PRINCIPLE AND SOFTWARE PARADIGM



R is not really a (statistical) software

- R is rather a programming language
- Limited user-friendly interfaces for data analysis
- Is object oriented and almost non declarative
- Similar to programming languages like Fortran, C, Java, Python



PRINCIPLE AND SOFTWARE PARADIGM

R has limited Graphical User Interface (GUI) options

Recent endeavours to enhance R user-friendliness

Several GUIs in development

R-commander

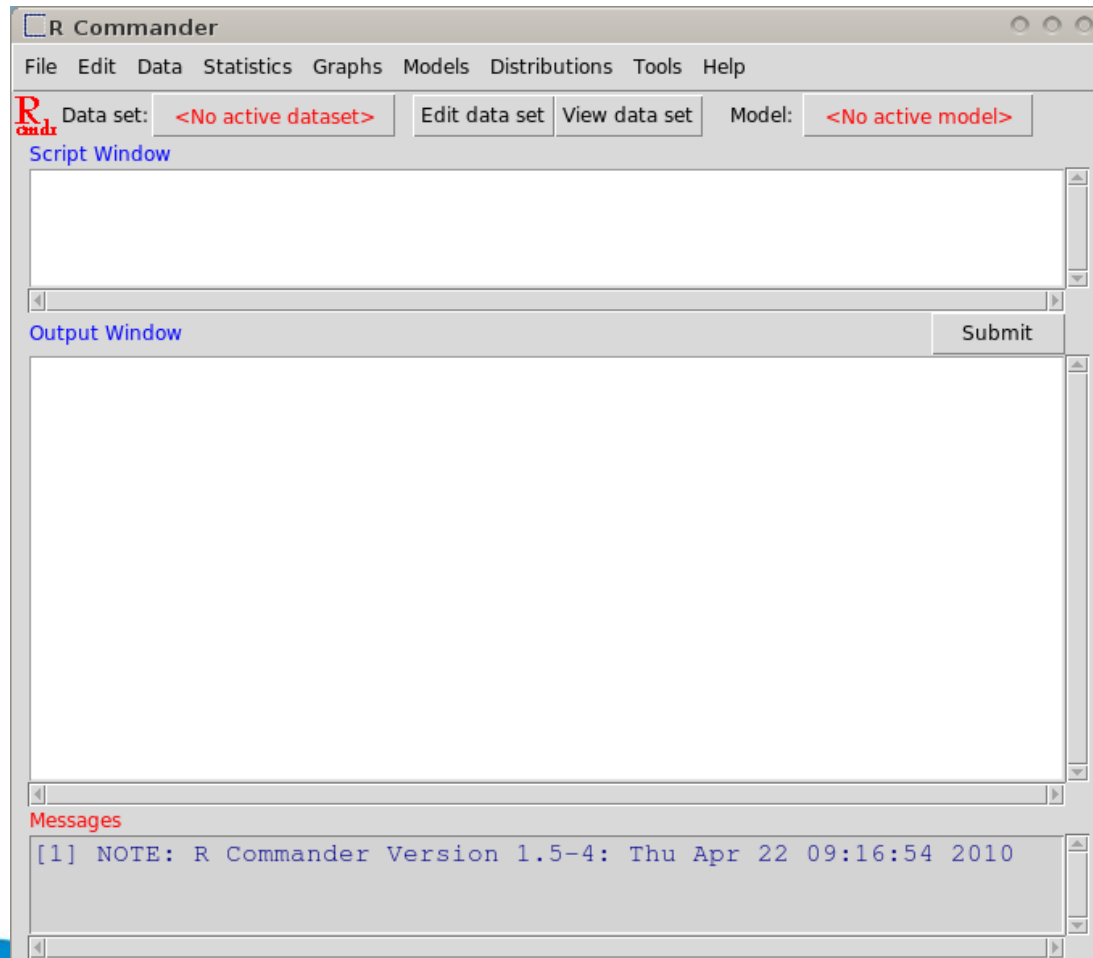
RKWard

Rattle

PRINCIPLE AND SOFTWARE PARADIGM



R Commander (RCmdr)





PRINCIPLE AND SOFTWARE PARADIGM

RKward

traffix.Rdata - ppscales - RKward

File Edit View Workspace Run Analysis Plots Distributions Windows Settings Help

Dataset Script File Open R Script File Open Recent Previous Window Next Window Cut Copy Paste Paste inside selection

Workspace

- All Non-Functions Functions
- scaleweights ppscales print.html
- Show All Environments
- Show Hidden Objects

Name Label

- package:graphics
- package:stats
- package:rkward
- .GlobalEnv
- x
- test
- shutDevice
- scores
- scaleweights
- sartpPlot
- sartp
- ppscales
 - vp
 - timestamp
 - sn9
 - sn8
 - sn7
 - sn6
 - sn5
 - sn4
 - sn3

Update

	pn6	an1	pn4	ac13	ac8	ac11	
Name	pn6	an1	pn4	ac13	ac8	ac11	
Label	Personal ...	Awarenes...	Personal ...	Awarenes...	Awarenes...	Awarenes...	Pe
Type	Number	Number	Number	Number	Number	Number	N
Format							
Levels							
1	3	5	2	4	3	1	
2	3	6	2	1	6	6	
3	7	7	1	2	7	1	
4	5		5	2	7	5	
5	3	4	4	4	1	6	
6	4	7	3	invalid	4	2	
7	2	6	2	2	5	2	
8	3	5	2	6	3	6	
9	5	5	4	6	5	7	
10	1	3	7	6	1	3	
11	6	7	7	2	6	5	
12	2	5	2	3	5	1	

Command log Pending Jobs R Console Help search

Ready. /home/thomas R engine idle



PRINCIPLE AND SOFTWARE PARADIGM

Rattle

R Data Miner - [Rattle (audit.csv)]

Project Tools Settings Help Rattle Version 2.6.6 logaware.com

Execute New Open Save Report Export Stop Quit

Data Explore Test Transform Cluster Associate Model Evaluate Log

Source: ☒ Spreadsheet ☐ ARFF ☐ ODBC ☐ RDataset ☐ RData File ☐ Library ☐ Corpus ☐ Script

Filename: Separator: Decimal: ☒ Header

☒ Partition Seed:

☒ Input ☒ Ignore Weight Calculator:

Target Data Type
☒ Auto ☐ Categorical ☐ Numeric ☐ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	ID	Numeric	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2000
2	Age	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 67
3	Employment	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 8 Missing: 100
4	Education	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 16
5	Marital	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 6
6	Occupation	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 14 Missing: 101
7	Income	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2000
8	Gender	Categorical	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
9	Deductions	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 41
10	Hours	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 68
11	IGNORE_Accounts	Categorical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	Unique: 33 Missing: 43
12	RISK_Adjustment	Numeric	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 310
13	TARGET_Adjusted	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2

Roles noted. 2000 observations and 9 input variables. The target is TARGET_Adjusted. Categorical 2. Classification models enabled.

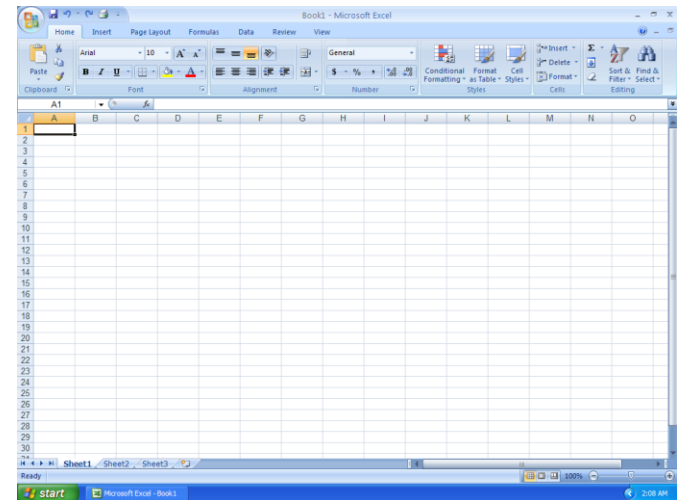


PRINCIPLE AND SOFTWARE PARADIGM

Inherent limitations of pervasive Excel-like spreadsheets



VS.



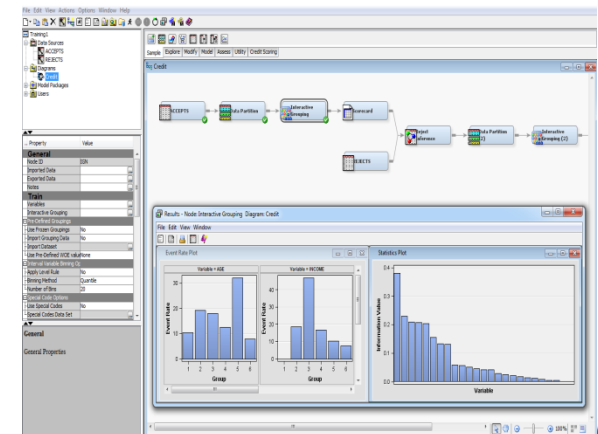
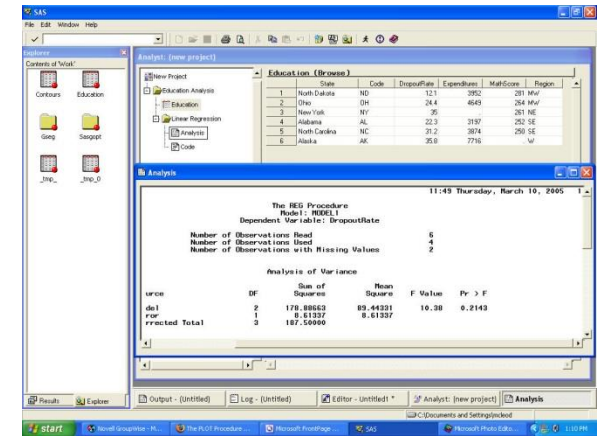


PRINCIPLE AND SOFTWARE PARADIGM

Sophisticated but costly SAS



VS.



Screenshot of SAS enterprise Miner 7.1. Source: sas.com



AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why using R?
- References for learning R



DESCRIPTION OF R INTERFACE

R console

The screenshot shows a desktop environment with a dark blue background. On the left side, there are three desktop shortcuts: Mozilla Firefox, R 2.14.0, and R x64 2.14.0. The RGui window is open, displaying the R console. The console window has a menu bar (File, Edit, View, Misc, Packages, Windows, Help) and a toolbar with icons for file operations. The console text includes the R version (2.14.0), platform (i386-pc-mingw32/i386 (32-bit)), and various help messages. A warning message is displayed at the bottom of the console, indicating that the library path is not writable and that some packages are not available. The RGui window is titled 'RGui' and has standard window controls (minimize, maximize, close). The R console window is titled 'R Console' and has its own window controls. The R console text is as follows:

```
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Warning in install.packages(necessary[!installed]) :
  'lib = "C:/Program Files/R/R-2.14.0/library"' is not writable
Warning: unable to access index for repository http://software.rc.fas.harvard.edu
Warning message:
In getDependencies(pkgs, dependencies, available, lib) :
  packages 'TinnR', 'svSocket' are not available (for R version 2.14.0)
Error in library(TinnR) : there is no package called 'TinnR'
> |
```

R desktop shortcut

RGui: R basic interface

R command line (space to write instructions)



DESCRIPTION OF R INTERFACE

Using the command line in R console

First false sentence
followed by R's
error message

Second correct
sentence

Declaration and
printing of the
sentence as a R
object

Simple math
computations

Basic information
about the R object
containing the
sentence

```
RGui
File Edit View Misc Packages Windows Help

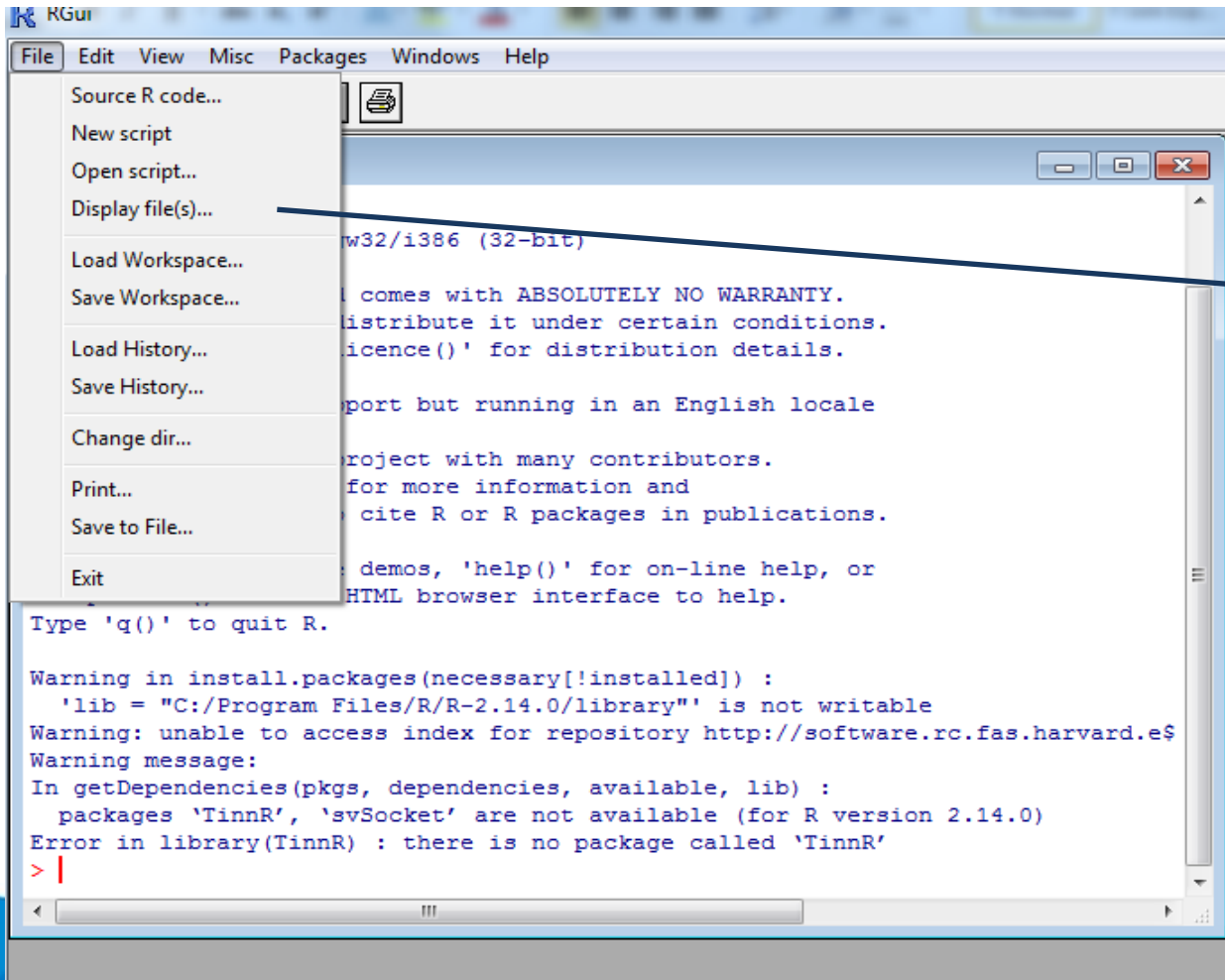
R Console

lib = "C:/Program Files/R/R-2.14.0/library" is not writable
Warning: unable to access index for repository http://software.rc.fas.harvard.edu
Warning message:
In getDependencies(pkgs, dependencies, available, lib) :
  packages 'TinnR', 'svSocket' are not available (for R version 2.14.0)
Error in library(TinnR) : there is no package called 'TinnR'
> Welcom to the Jigsaw R intriduction course
Error: unexpected symbol in "Welcom to"
> "Welcom to the Jigsaw R intriduction course"
[1] "Welcom to the Jigsaw R intriduction course"
> Welcome="Welcom to the Jigsaw R intriduction course"
> Welcome
[1] "Welcom to the Jigsaw R intriduction course"
> 3+4
[1] 7
> ln(1)
Error: could not find function "ln"
> log(1)
[1] 0
> is.vector(Welcome)
[1] TRUE
> summary(Welcome)
      Length      Class      Mode 
      1 character character
```



DESCRIPTION OF R INTERFACE

RGui menu: File tab

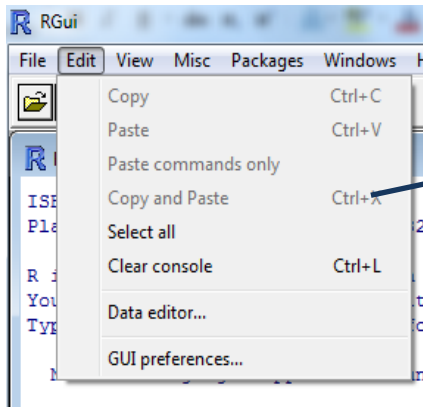


**File tab: Usual basic
and general
operations**

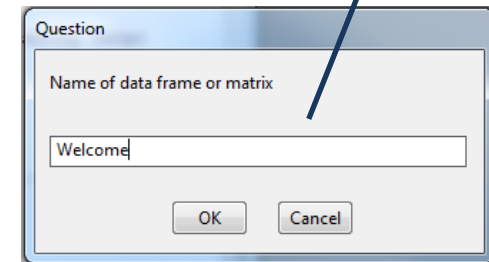


DESCRIPTION OF R INTERFACE

RGui menu: Edit tab

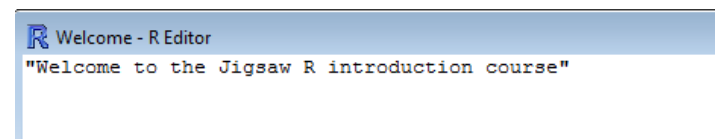


Edit tab: basic and general editing



Data editor: entering the object's name

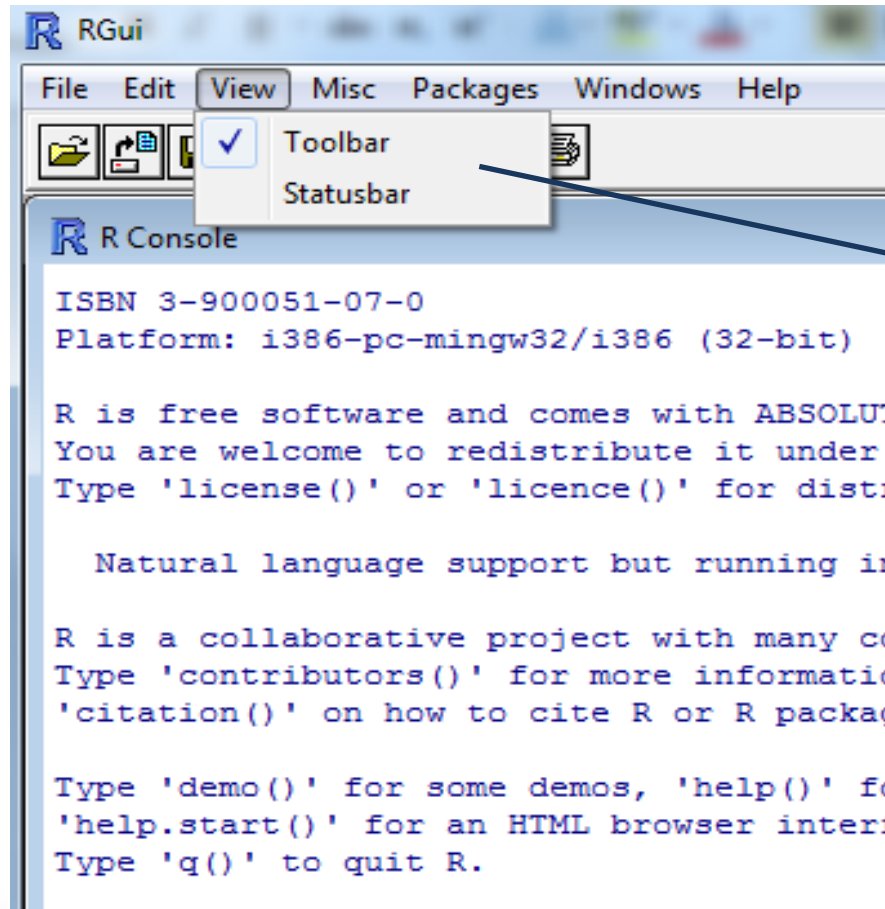
Results of the data editor





DESCRIPTION OF R INTERFACE

RGui menu: View tab

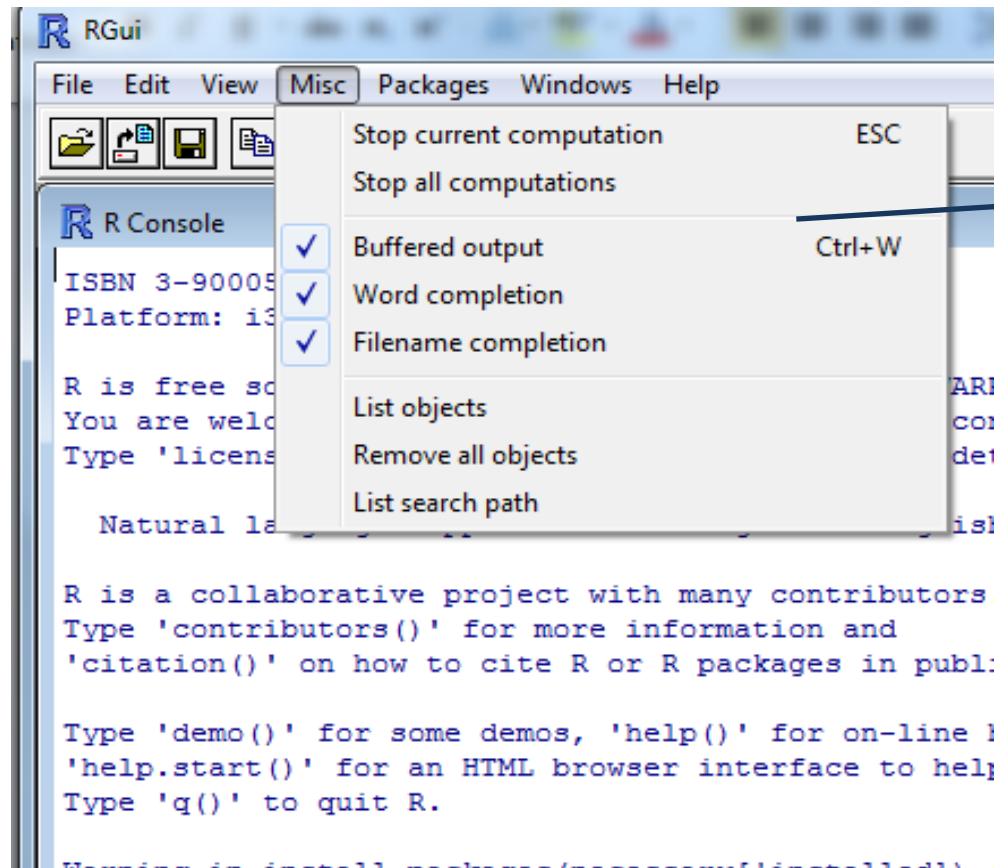


View tab: viewing
Toolbar and/or
Status bar



DESCRIPTION OF R INTERFACE

RGui menu: Misc tab

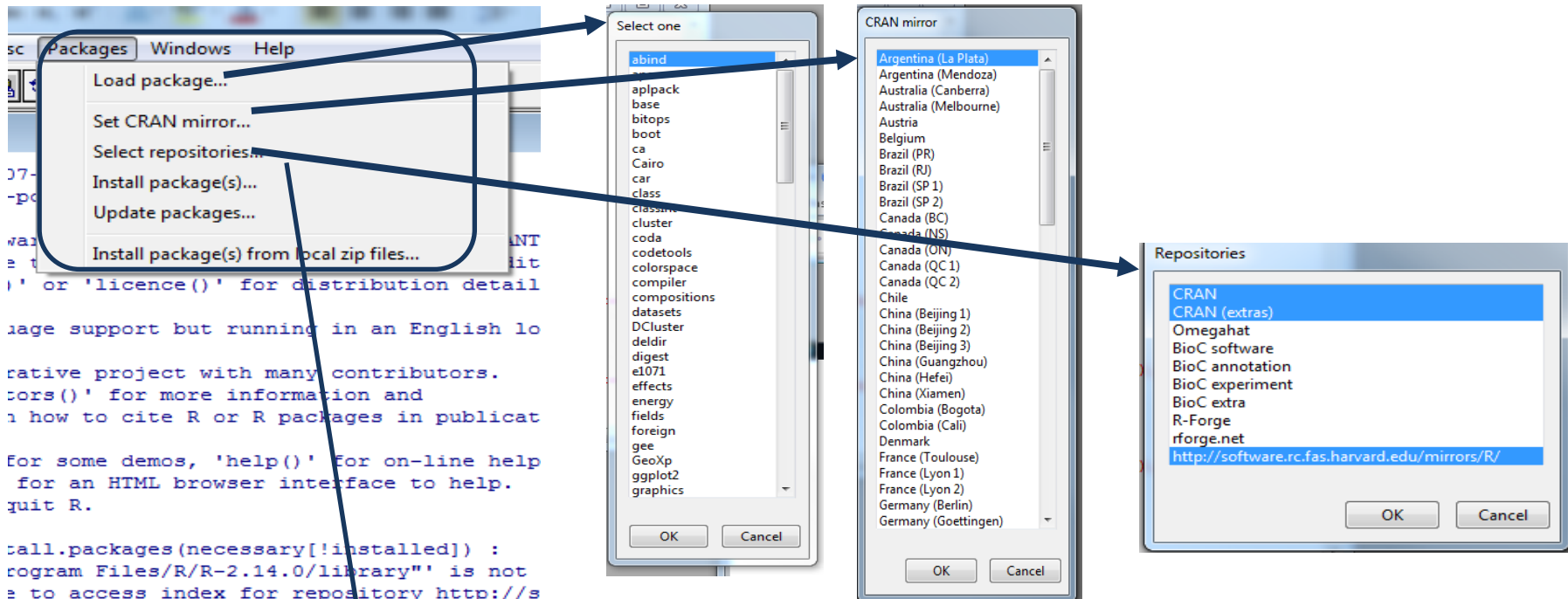


Misc tab:
diverse
operations



DESCRIPTION OF R INTERFACE

RGui menu: Packages tabs

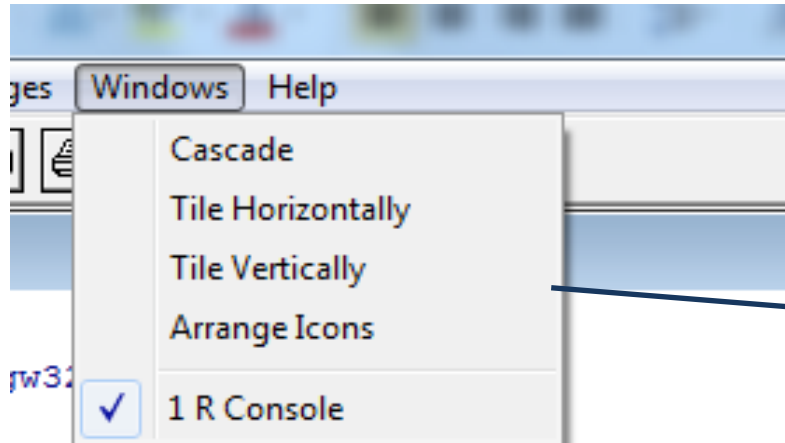


Packages tab:
adding functions
to R foundation



DESCRIPTION OF R INTERFACE

RGui menu: Windows tab



Windows tab:
usual options
to arrange the
tiles

```
It comes with ABSOLUTELY NO WARRANTY.  
It is distributed under certain conditions.  
See 'licence()' for distribution details.
```

```
It is not supported but running in an English locale.
```

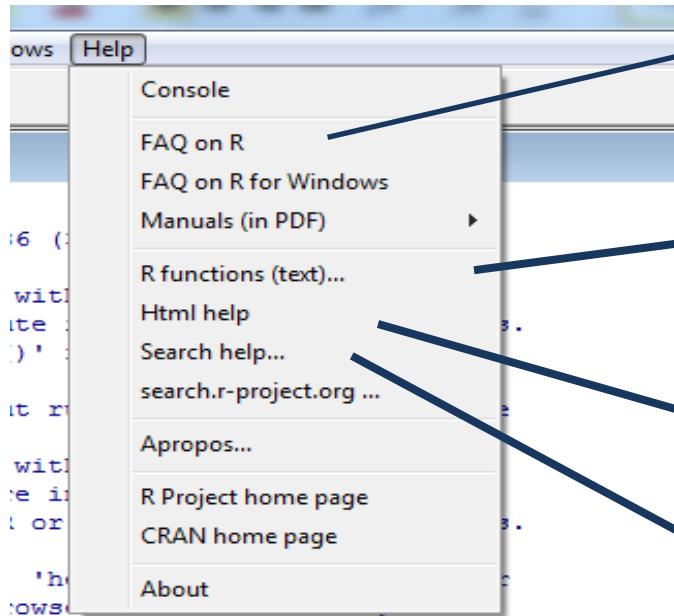
```
It is a project with many contributors.  
See 'help()' for more information and  
to cite R or R packages in publications.
```

```
See 'demo()' for on-line help,
```

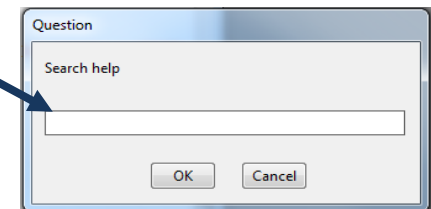
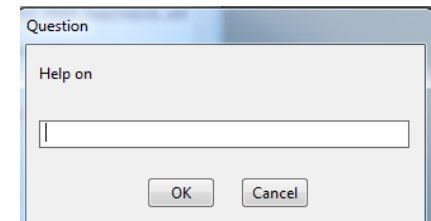
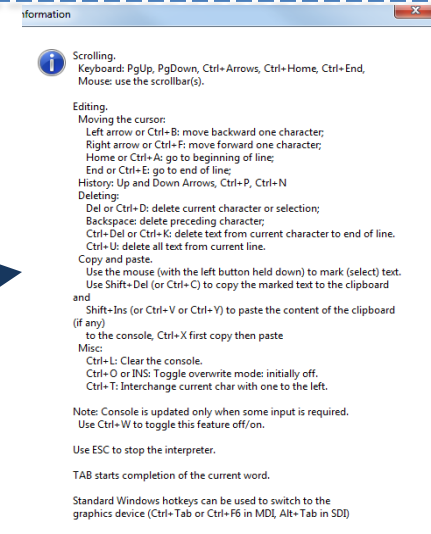


DESCRIPTION OF R INTERFACE

RGui menu: Help tab



Help tab: very important links to help



```
necessary[!installed]) :  
R-2.14.0/library" is not writable  
lex for repository http://software.rc.fas
```



AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why using R?
- References for learning R



ADVANTAGES OF R

R “philosophy”

- Open source code
- You can access the code of the software
- In-depth understanding of what R does
- Modify the code

Example “mgcv” package webpage

Address of the
« mgcv » package

Link with Package
sources (.tar.gz
file)

mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL

Routines for GAMs and other generalized ridge regression with multiple smoothing parameter selection by GCV, REML or UBRE/AIC. Also GAMMs by REML or PQL. Includes a gam() function.

Version: 1.7-13
Priority: recommended
Depends: R (≥ 2.14.0), stats, graphics
Imports: nlme, methods, Matrix
Suggests: nlme (≥ 3.1-64), splines, Matrix, parallel
Published: 2012-01-22
Author: Simon Wood
Maintainer: Simon Wood <simon.wood at r-project.org>
License: GPL (≥ 2)
Citation: mgcv citation info
In views: Bayesian, Econometrics, Environmetrics, SocialSciences
CRAN checks: mgcv results

Downloads:

Package source: [mgcv 1.7-13.tar.gz](#)
MacOS X binary: [mgcv 1.7-13.tgz](#)
Windows binary: [mgcv 1.7-13.zip](#)
Reference manual: [mgcv.pdf](#)
News/ChangeLog: [ChangeLog](#)
Old sources: [mgcv archive](#)

Reverse dependencies:

Reverse depends: [AdaptFitOS](#), [coenoflex](#), [COZIGAM](#), [diseasemapping](#), [dlmap](#), [DSpat](#), [ez](#), [fRegression](#), [gamlss.add](#), [gamm4](#), [Haplin](#), [JointModeling](#), [labdsv](#), [MAPLES](#), [modTempEff](#), [mombf](#), [paltran](#), [pcurve](#), [PL.popN](#), [reams](#), [refund](#), [RLRsim](#), [season](#), [SemiParBIVprobit](#), [spatstat](#), [STAR](#), [StatDA](#), [TSA](#), [tsDyn](#)

Reverse imports: [landsat](#), [openair](#), [scam](#), [tripEstimation](#)

Reverse suggests: [amer](#), [BiodiversityR](#), [car](#), [caret](#), [catdata](#), [demography](#), [flexmix](#), [granova](#), [HSAUR2](#), [latticeExtra](#), [LMERConvenienceFunctions](#), [MatchIt](#), [mediation](#), [MuMin](#), [PL.popN](#), [rasterVis](#), [Rcmdr](#), [RcmdrPlugin.HH](#), [RcmdrPlugin.IPSUR](#), [reldist](#), [rioja](#), [scam](#), [simex](#), [taRifx](#), [vegan](#), [Zelig](#)

Reverse enhances: [sfsmisc](#)

Screenshot of the CRAN webpage of the « mgcv » package. Source: [CRAN](#)



ADVANTAGES OF R

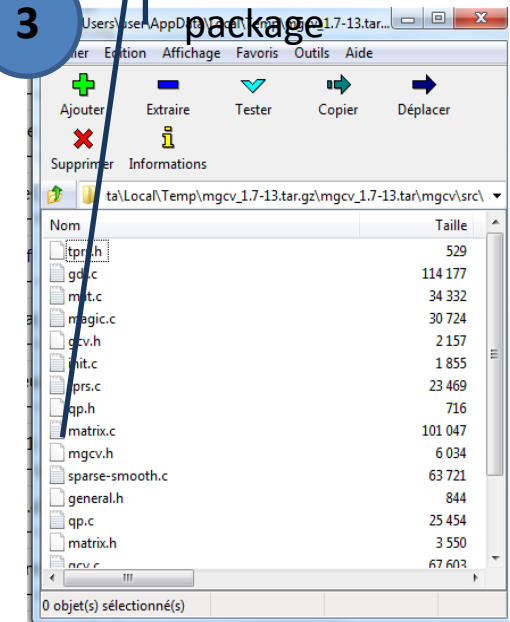
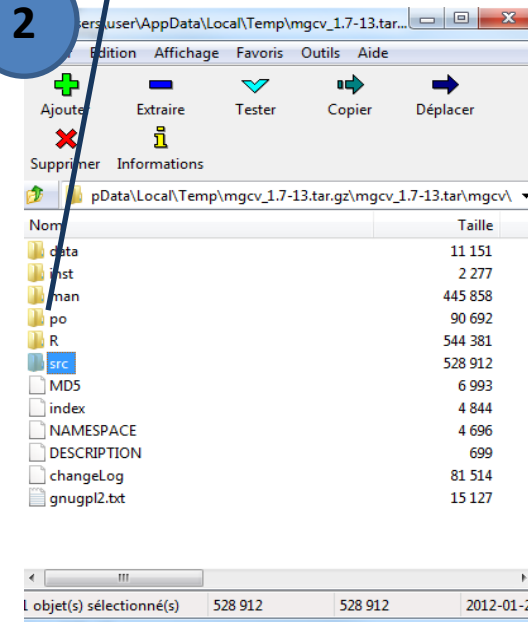
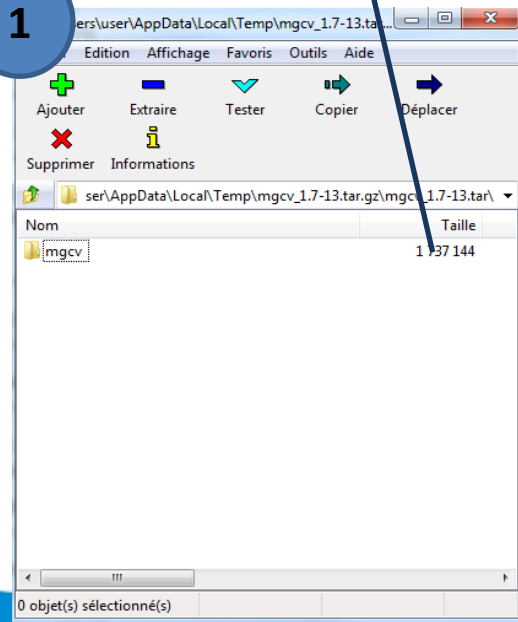
R access to source code

Example of source code of the “mgcv” package

Unzipping
[mgcv_1.7-13.tar.gz](#)
file (with 7zip)

List of directories
in the « mgcv »
package

List of functions (i.e
open code) in the « src »
(i.e code sources)
directory the « mgcv »
package



Screenshot of unzipping the « mgcv » package and browsing through the package's files.



ADVANTAGES OF R

R is free

Software	Academics	Demo	Commercial (basic)	Commercial (full)
R	Free	Free	Free	Free
SAS	Free to \$100s	Not available	\$1 000s	\$10 000s
Statistica	\$100s	30 days limit	~\$1 000	\$10 000
Excel (Microsoft)	Free to \$10s	Limited	~\$100	\$100s
SPSS (IBM)	\$100s	14 days limit	~\$2 000	\$1 000s



ADVANTAGES OF R

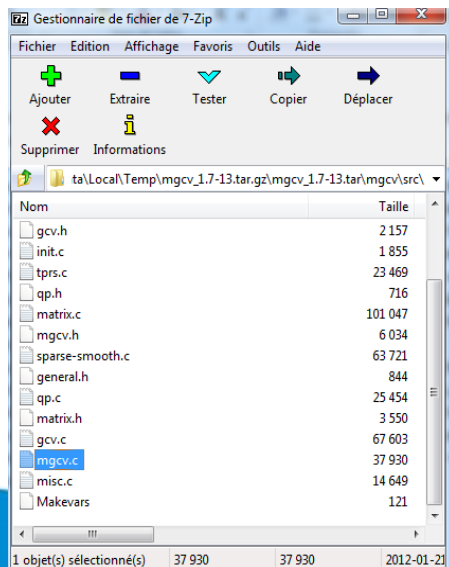
Interface with other languages and scripting capabilities

Interfaces with virtually any other programming language

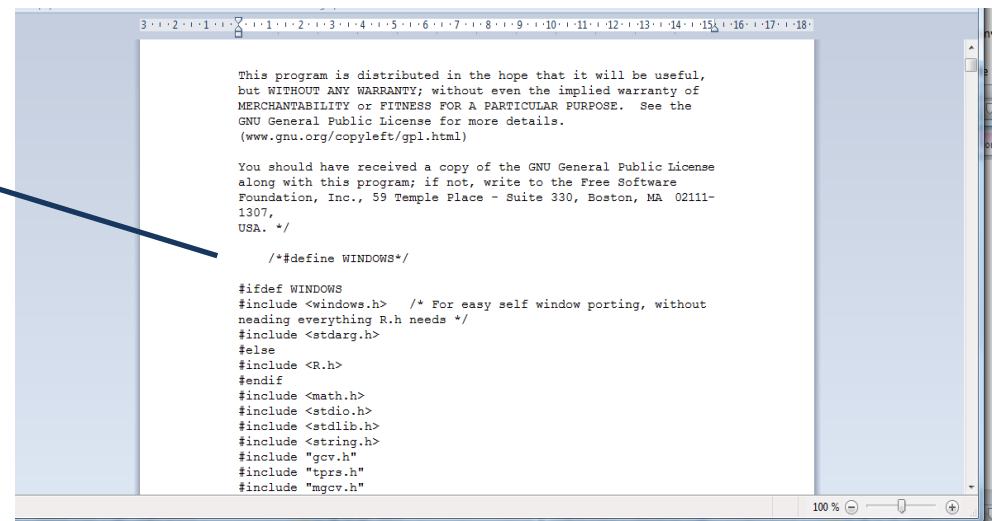
- Fortran, C, C++, Python...
- Tailor or rewrite your old codes in R

R as a scripting language

- R scripts can launch or be launched by other languages



« mgcv.c » file
in the
« mgcv »
package
coded in
typical C
programming
language

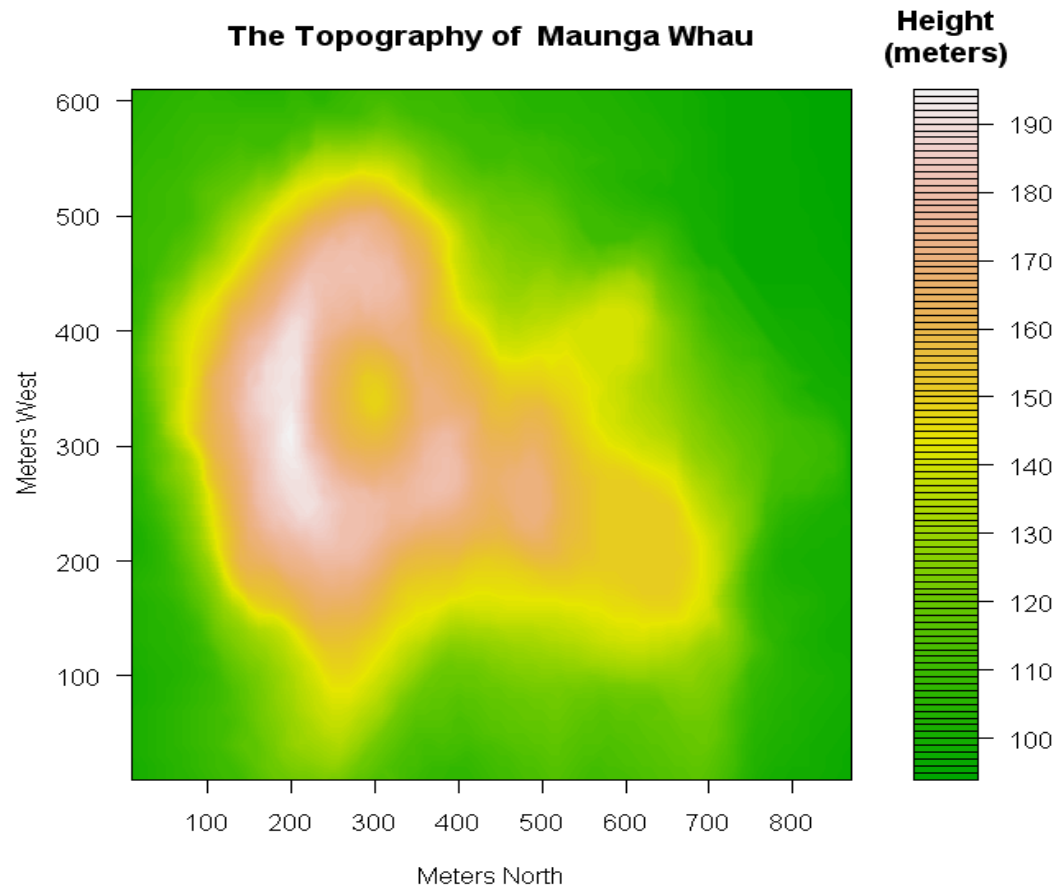


Screenshot of the file « mgcv.c » of the « mgcv » package open in WordPad



ADVANTAGES OF R

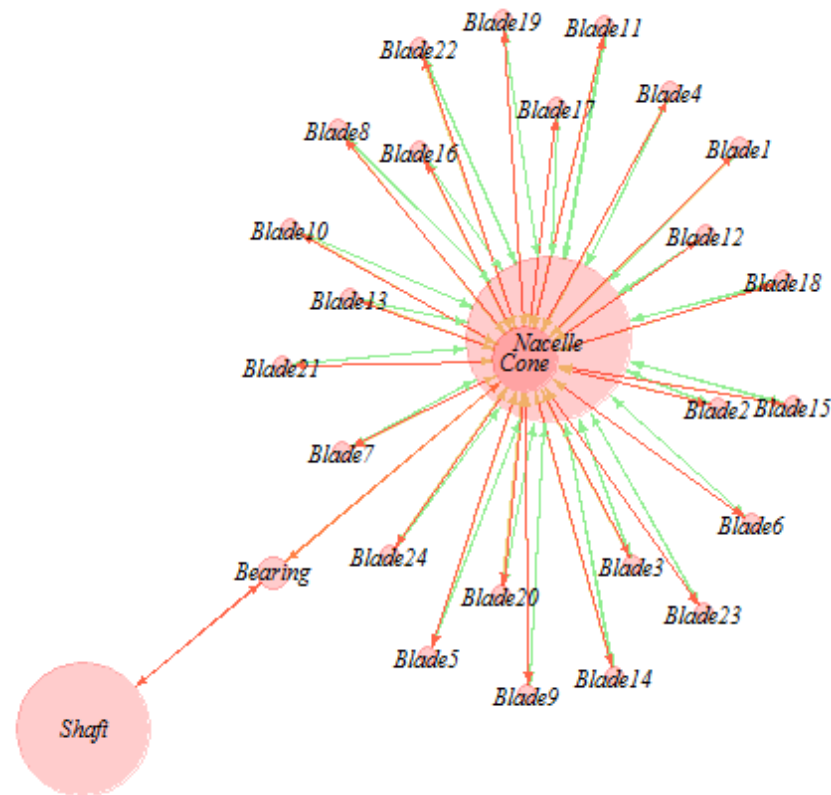
R visualization capabilities





ADVANTAGES OF R

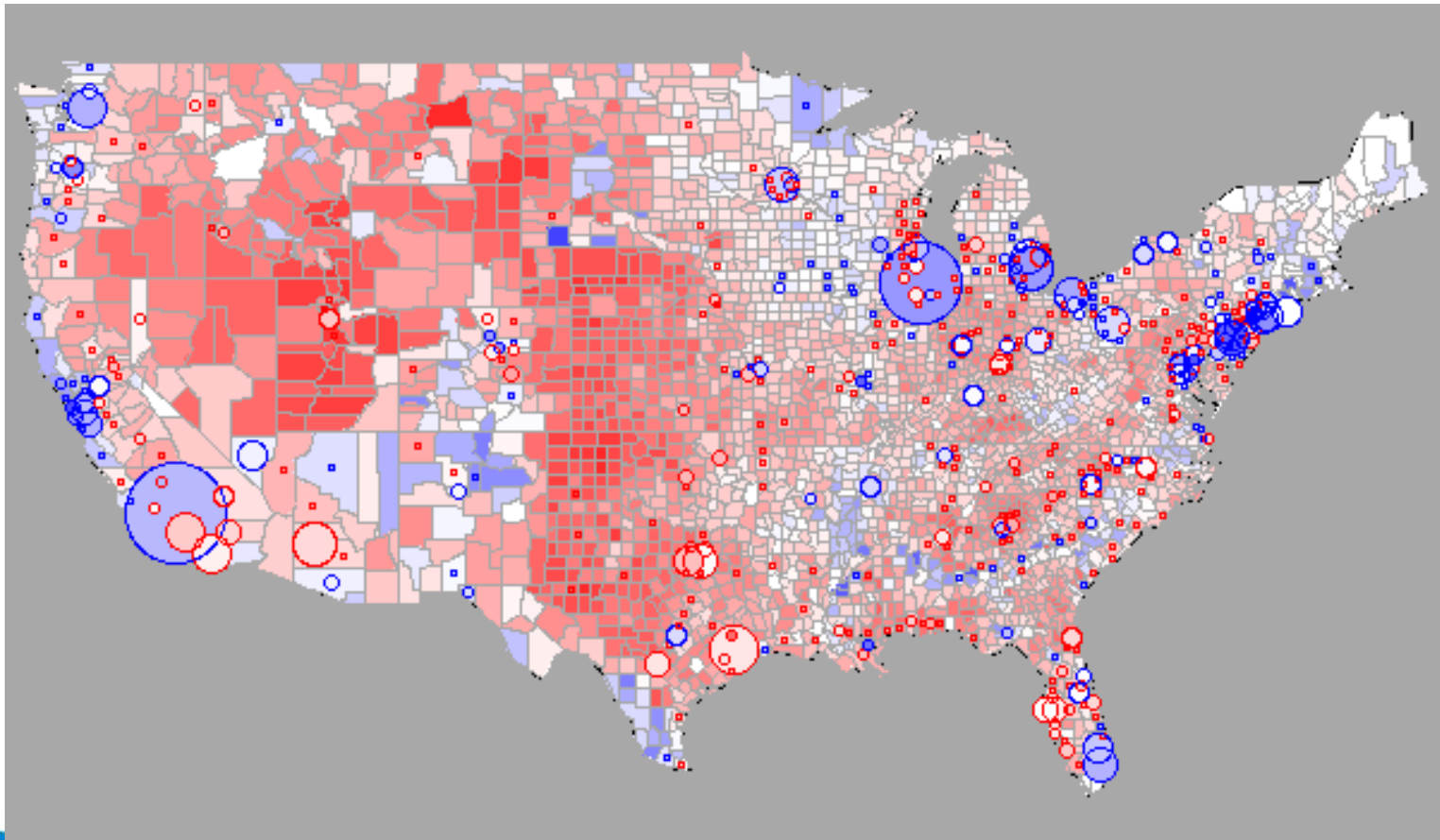
R visualization capabilities





ADVANTAGES OF R

R visualization capabilities





ADVANTAGES OF R

R role in academia

- R ~ tool used by the finest researchers
- **Top-notch analytics capabilities**



Screenshot of a user's Facebook map . Source: [Paul Butler/Facebook](#), DG Rossiter, [spatialanalysis.co.uk](#)



ADVANTAGES OF R

To summarize

Free open source philosophy

- R websites with many examples
- Free books
- Free online open courses
- Twitter accounts

Online help and discussion

- Mailing-lists
- Very active and diverse forums
- Communities of developers and helpers



AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why using R?
- References for learning R



DRAWBACKS OF R

Average memory performance

Poor management of large datasets

- Avoid imbricated loops
- Prefer R advanced language for data structure

Complicated structure of packages in R

- Dozen of packages
- To be loaded every time in memory

R packages to better manage memory

- Rhadoop (inspiration from Google)
- Ff
- bigmemory



DRAWBACKS OF R

Average computing performance

No default parallel execution

- R packages to use several cores
- Top skills needed for high performance computing

A high-level programming language

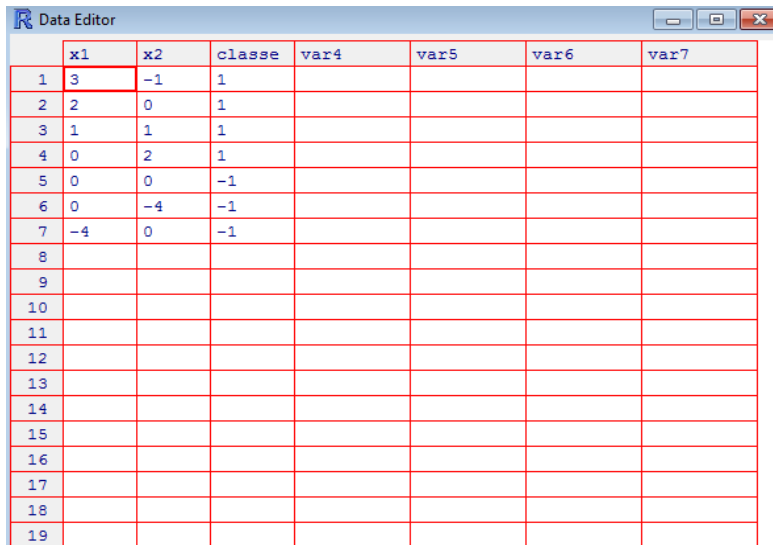
- Abstract and modern (Python...)
- More productive coding
- But further from « machine language »...
- ... meaning 100 times slower than C



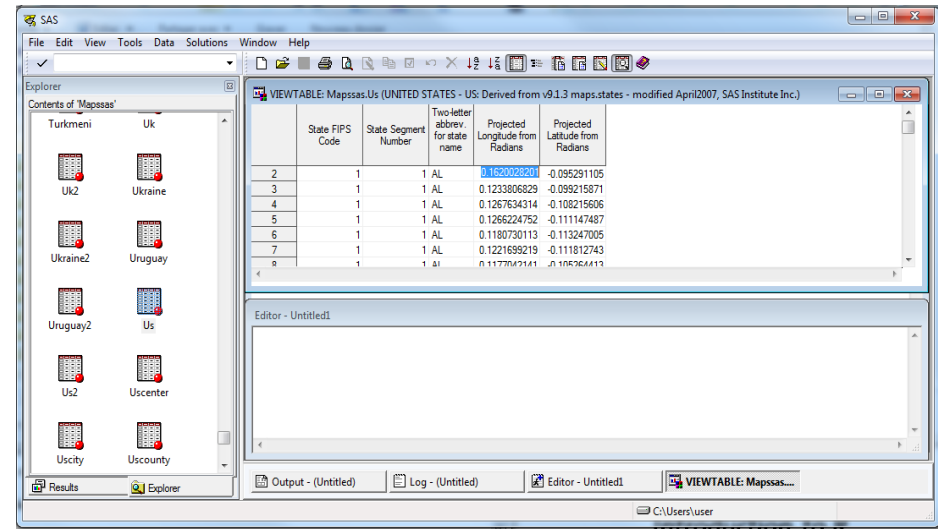
DRAWBACKS OF R

Difficult data visualization and management

Difficult to inspect data sets



	x1	x2	classe	var4	var5	var6	var7
1	3	-1	1				
2	2	0	1				
3	1	1	1				
4	0	2	1				
5	0	0	-1				
6	0	-4	-1				
7	-4	0	-1				
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							



VIEWTABLE: Mapssas.Us (UNITED STATES - US: Derived from v9.1.3 maps.states - modified April2007, SAS Institute Inc.)

	State FIPS Code	State Segment Number	Two-letter abbrev. for state name	Projected Longitude from Radians	Projected Latitude from Radians
2	1	1	AL	0.1620028201	-0.095291105
3	1	1	AL	0.1233806829	-0.099215871
4	1	1	AL	0.1267634314	-0.108215606
5	1	1	AL	0.1266224752	-0.111147487
6	1	1	AL	0.1180730113	-0.113247005
7	1	1	AL	0.1221699219	-0.111812743
8	1	1	AL	0.1177047141	-0.106364413

Screenshot of the R data editor and « Viewtable » tab in SAS 9.3



DRAWBACKS OF R

Difficult architecture management

Problems for large organizations

- R made of several thousands independent packages
- No deployment plan for complex organizations
- No installation support

Lack of code accountability

- Thousands of individual independent R developers
- Nobody responsible for the quality of the code

Potentially high hidden costs with R

- Total cost may favour commercial solutions for complex computations made in large corporations





DRAWBACKS OF R

Relatively difficult to learn

Steep learning curve

- R code far from undergrad computer science courses
- Very complex data structures (useful if mastered)
- Is R's syntax not logical?



Still, not more difficult to learn than SAS

- Both SAS and R more abstract than basic programming languages (Fortran, C...)
- Difficult to learn = more rewarding professionally!!





AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why use R?
- References for learning R



SO WHY LEARN R?

More positive than negative points

No language is perfect!!

- Contradictory objectives to meet
- Strengths and weaknesses of each language

Effect of legacy and the culture of the organization

- Use existing solutions (system architecture, BA tools...)
- Habits in business analytics

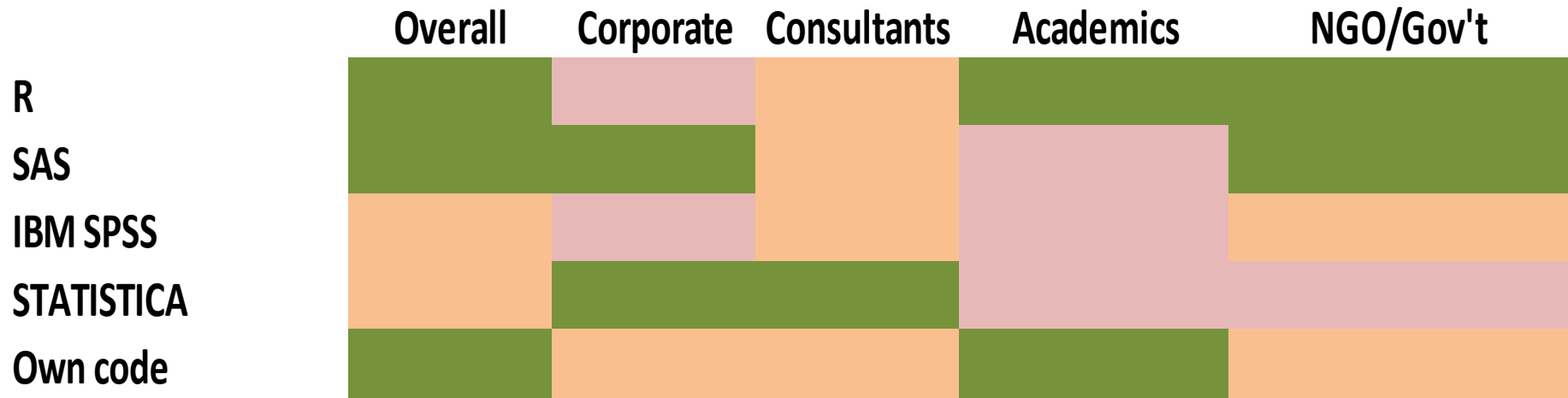
Different needs imply different tools

- Large corporations + defined procedures → SAS-like
- Less financial resources + quick proof of concept → R



SO WHY LEARN R?

Very appealing solution



Popularity of business analytics software (green = very popular, red = unpopular). Source: [Rexer Analytics](#)



AGENDA

- History and evolution of R
- Principle and software paradigm
- Description of R interface
- Advantages of R
- Drawbacks of R
- So why using R?
- References for learning R



REFERENCES FOR LEARNING R

Books

Many books available: choose the one that fits you!

- Style, pedagogy, theory vs practice
- Browse several books at local library or store

Springer's UseR! Series (<http://www.springer.com/series/6991>)

- Recent, concise, good quality, affordable, diverse

Pure rookies: « A beginners' guide to R », « R by example »

One step forward: « Business analytics for managers »

Intensive Excel users: « R through Excel »

O'Reilly R series (for programmers)

« R cookbook », « R in a nutshell »



REFERENCES FOR LEARNING R

Websites

R official websites

- The R project for statistical computing (www.r-project.org)
- Mailing lists (« [R-help](#) », Special Interest Groups) and R journal
- Official (austere) manuals (« An introduction to R »)

Other websites

- UCLA online R resources <http://www.ats.ucla.edu/stat/r/>)
- R blogs aggregator (www.r-bloggers.com)
- Social networks: LinkedIn groups (The R project for statistical computing), Twitter accounts (@RevolutionR, @inside_R), jobboards (Analytical Bridge...)



REFERENCES FOR LEARNING R

Conferences

Growing number of conferences about R

Official International R UseR! conference

- Annual during a few days in new venue (Google it!)
- Lots of materials about many topics

Other conferences or venues

- Conferences about business analytics (data mining, specialized topics...) with sessions involving R
- Find (or even start!) a R user group close to your location ([R Wiki](#) geographical list, map of groups on « meetup.com »)
- Events and news from R-bloggers blog