



a free software platform for statistical computing  
and myriads of other stuff

Jérôme Sueur & Stéphane Dray

MUSÉUM  
NATIONAL  
D'HISTOIRE  
NATURELLE  
● ● ● ●



28/11/2013

Journées du Centre Blaise Pascal  
Data analysis and modelling in life sciences

# What is it?

Born in 1993

Developed by **Robert Gentleman** and **Ross Ihaka**  
University of Auckland, New-Zeland



©New-York Times

## What is it?



- ▶ A language and environment for statistical computing and graphics

## What is it?



- ▶ A language and environment for statistical computing and graphics
- ▶ Derived from S developed at Bell Laboratories

## What is it?



- ▶ A language and environment for statistical computing and graphics
- ▶ Derived from S developed at Bell Laboratories
- ▶ Licence: Free Software Foundation's GNU General Public License

## What is it?



- ▶ A language and environment for statistical computing and graphics
- ▶ Derived from S developed at Bell Laboratories
- ▶ Licence: Free Software Foundation's GNU General Public License
- ▶ All OS (Windows, MacOS, Linux, FreeBSD)

## What is it?



- ▶ A language and environment for statistical computing and graphics
- ▶ Derived from S developed at Bell Laboratories
- ▶ Licence: Free Software Foundation's GNU General Public License
- ▶ All OS (Windows, MacOS, Linux, FreeBSD)
- ▶ Collaborative project

## What is it?



- ▶ A language and environment for statistical computing and graphics
- ▶ Derived from S developed at Bell Laboratories
- ▶ Licence: Free Software Foundation's GNU General Public License
- ▶ All OS (Windows, MacOS, Linux, FreeBSD)
- ▶ Collaborative project
- ▶ Customizable to your own needs (open source)



## Who contribute?

- ▶ Core team: an international team 20 people mainly coming from statistics

## Who contribute?

- ▶ Core team: an international team 20 people mainly coming from statistics
- ▶ Users: about  $2.10^6$

## Who contribute?

- ▶ Core team: an international team 20 people mainly coming from statistics
- ▶ Users: about  $2.10^6$
- ▶ You!

## What is it done for?

He can do everything or almost...



- ▶ mathematics
- ▶ physics
- ▶ chemistry
- ▶ astronomy
- ▶ applied statistics
- ▶ spatial data analysis
- ▶ financial sciences
- ▶ social sciences
- ▶ text mining

## What is it done for?

He can do everything or almost...



- ▶ mathematics
- ▶ physics
- ▶ chemistry
- ▶ astronomy
- ▶ applied statistics
- ▶ spatial data analysis
- ▶ financial sciences
- ▶ social sciences
- ▶ text mining
- ▶ life sciences (ecology, genomics, phylogeny, geology, archeology,...)

# What is it done for?

He can do everything or almost...



- ▶ mathematics
- ▶ physics
- ▶ chemistry
- ▶ astronomy
- ▶ applied statistics
- ▶ spatial data analysis
- ▶ financial sciences
- ▶ social sciences
- ▶ text mining
- ▶ life sciences (ecology, genomics, phylogeny, geology, archeology,...)
- ▶ **your research domain**

# What is it done for?

## Task views

Bayesian Inference, Chemometrics and Computational Physics, Clinical Trial Design, Monitoring, and Analysis, Cluster Analysis & Finite Mixture Models, Differential Equations, Probability Distributions, Computational Econometrics, Analysis of Ecological and Environmental Data, Design of Experiments (DoE) & Analysis of Experimental Data, Empirical Finance, Statistical Genetics, Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization, High-Performance and Parallel Computing with R, Machine Learning & Statistical Learning, Medical Image Analysis, Meta-Analysis, Multivariate Statistics, Natural Language Processing, Numerical Mathematics, Official Statistics & Survey Methodology, Optimization and Mathematical Programming, Analysis of Pharmacokinetic Data, Phylogenetics, Especially Comparative Methods, Psychometric Models and Methods, Reproducible Research, Robust Statistical Methods, Statistics for the Social Sciences, Analysis of Spatial Data, Handling and Analyzing Spatio-Temporal Data, Survival Analysis, Time Series Analysis, Web Technologies and Services, gRaphical Models in R,

# What is it done for?

## Task views

Bayesian Inference, Chemometrics and Computational Physics, Clinical Trial Design, Monitoring, and Analysis, Cluster Analysis & Finite Mixture Models, Differential Equations, Probability Distributions, Computational Econometrics, Analysis of Ecological and Environmental Data, Design of Experiments (DoE) & Analysis of Experimental Data, Empirical Finance, Statistical Genetics, Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization, High-Performance and Parallel Computing with R, Machine Learning & Statistical Learning, Medical Image Analysis, Meta-Analysis, Multivariate Statistics, Natural Language Processing, Numerical Mathematics, Official Statistics & Survey Methodology, Optimization and Mathematical Programming, Analysis of Pharmacokinetic Data, Phylogenetics, Especially Comparative Methods, Psychometric Models and Methods, Reproducible Research, Robust Statistical Methods, Statistics for the Social Sciences, Analysis of Spatial Data, Handling and Analyzing Spatio-Temporal Data, Survival Analysis, Time Series Analysis, Web Technologies and Services, gRaphical Models in R, **your task view**



## How does it look like? – SHELL

```
$ R CMD BATCH scriptfile.R outputfile.out
```

or

```
$ Rscript --slave scriptfile.R arg1 arg2 arg3 > results.out
```

or to make an R script file executable, add a header line in the R file:

```
> #!/usr/bin/Rscript --slave  
> # Rscript here
```

then make the .R scriptfile executable and invoke it directly:

```
$ chmod +x scriptfile.R  
$ scriptfile.R input-value
```

## How does it look like? – TERMINAL

```
bioac@bioac-Latitude-E6430: ~  
bioac@bioac-Latitude-E6430:~$ R  
  
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"  
Copyright (C) 2013 The R Foundation for Statistical Computing  
Platform: i686-pc-linux-gnu (32-bit)  
  
R est un logiciel libre livré sans AUCUNE GARANTIE.  
Vous pouvez le redistribuer sous certaines conditions.  
Tapez 'license()' ou 'licence()' pour plus de détails.  
  
R est un projet collaboratif avec de nombreux contributeurs.  
Tapez 'contributors()' pour plus d'information et  
'citation()' pour la façon de le citer dans les publications.  
  
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide  
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.  
Tapez 'q()' pour quitter R.  
  
> █
```

# How does it look like? – GUI

The screenshot displays the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Project, Build, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar. The main workspace is divided into two panes: the Console on the left and the Packages pane on the right.

**Console:**

```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: i686-pc-linux-gnu (32-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et 'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

> 1+1
[1] 2
> |
```

**Packages Pane:**

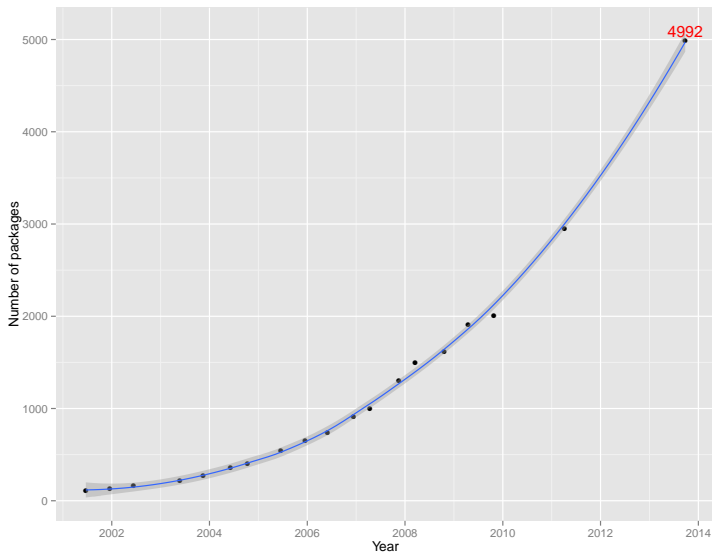
Package Name	Description	Version	Update Icon
<input type="checkbox"/> <a href="#">audio</a>	Audio Interface for R	0.1-4	🔄
<input type="checkbox"/> <a href="#">audio</a>	Audio Interface for R	0.1-4	🔄
<input type="checkbox"/> <a href="#">audiolyzR</a>	audiolyzR: Give your data a listen	0.4-9	🔄
<input type="checkbox"/> <a href="#">boot</a>	Bootstrap Functions (originally by Angelo Canty for S)	1.3-9	🔄
<input type="checkbox"/> <a href="#">cacheSweave</a>	Tools for caching Sweave computations	0.6-1	🔄
<input type="checkbox"/> <a href="#">class</a>	Functions for Classification	7.3-9	🔄
<input type="checkbox"/> <a href="#">cluster</a>	Cluster Analysis Extended Rousseeuw et al.	1.14.4	🔄
<input type="checkbox"/> <a href="#">codetools</a>	Code Analysis Tools for R	0.2-8	🔄
<input type="checkbox"/> <a href="#">colorspace</a>	Color Space Manipulation	1.2-2	🔄
<input type="checkbox"/> <a href="#">colorspace</a>	Color Space Manipulation	1.2-2	🔄
<input type="checkbox"/> <a href="#">compiler</a>	The R Compiler Package	3.0.2	🔄
<input type="checkbox"/> <a href="#">csound</a>	Accessing Csound functionality through R	0.1-1	🔄
<input checked="" type="checkbox"/> <a href="#">datasets</a>	The R Datasets Package	3.0.2	🔄
<input type="checkbox"/> <a href="#">dichromat</a>	Color Schemes for Dichromats	2.0-0	🔄

## How does it work? – main structure

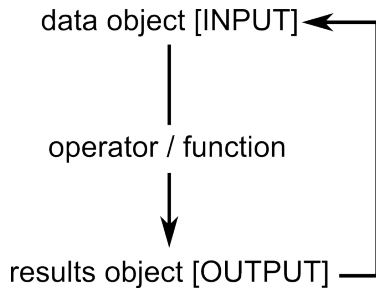
base (Core team) + packages (users)

## How does it work? – main structure

base (Core team) + packages (users)



## How does it work? – objects



## How does it work? – objects

R is an object oriented program language

```
> v <- 1:5
```

```
> v
```

```
[1] 1 2 3 4 5
```

```
> c <- c("hello world", "goodbye moon")
```

```
> c
```

```
[1] "hello world" "goodbye moon"
```

```
> m <- matrix(6:30, nc=5)
```

```
> m
```

```
      [,1] [,2] [,3] [,4] [,5]  
[1,]    6   11   16   21   26  
[2,]    7   12   17   22   27  
[3,]    8   13   18   23   28  
[4,]    9   14   19   24   29  
[5,]   10   15   20   25   30
```

```
> df <- crabs
```

```
> head(df)
```

```
   sp sex index  FL RW  CL  CW  BD  
1  B  M     1  8.1 6.7 16.1 19.0 7.0  
2  B  M     2  8.8 7.7 18.1 20.8 7.4  
3  B  M     3  9.2 7.8 19.0 22.4 7.7  
4  B  M     4  9.6 7.9 20.1 23.1 8.2  
5  B  M     5  9.8 8.0 20.3 23.0 8.2  
6  B  M     6 10.8 9.0 23.0 26.5 9.8
```

## How does it work? – objects

```
> v[2:3]
[1] 2 3

> m[1:2, 1:2]
      [,1] [,2]
[1,]    6   11
[2,]    7   12

> head(df$CL)
[1] 16.1 18.1 19.0 20.1 20.3 23.0

> vm <- cbind(v,m)
> vm
      v
[1,] 1  6 11 16 21 26
[2,] 2  7 12 17 22 27
[3,] 3  8 13 18 23 28
[4,] 4  9 14 19 24 29
[5,] 5 10 15 20 25 30
```



## How does it work? – pre-programmed loops and conditions

```
> v^2
[1] 1 4 9 16 25
> apply(m, MARGIN=2, FUN=sum)
[1] 40 65 90 115 140
> m[m < mean(m)] <- 0
> m
      [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0   21   26
[2,]    0    0    0   22   27
[3,]    0    0   18   23   28
[4,]    0    0   19   24   29
[5,]    0    0   20   25   30
```

## How does it work? – generic functions

Generic function: a single function that works differently depending on the class of the input object

- ▶ `print()`: print an object
- ▶ `plot()`: graphical display of an object
- ▶ `summary()`: returns a summary of results

## How does it work? – generic functions

```
> summary(weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.590  4.388   4.750   4.846   5.217   6.110
```

```
> summary(lm(weight ~ group))
```

```
Call:
```

```
lm(formula = weight ~ group)
```

```
Residuals:
```

```
   Min       1Q   Median       3Q      Max
-1.0710 -0.4938  0.0685  0.2462  1.3690
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.2202  22.850 9.55e-15 ***
groupTrt     -0.3710     0.3114  -1.191  0.249
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6964 on 18 degrees of freedom
```

```
Multiple R-squared:  0.07308,    Adjusted R-squared:  0.02158
```

```
F-statistic: 1.419 on 1 and 18 DF,  p-value: 0.249
```

## How does it work? – functions

Very easy to write new functions that take expressions as input.

A new function named `f2c` that takes a temperature `x` in Fahrenheit as input and returns a temperature in Celsius:

```
> f2c <- function(x)
  {
    res <- (x-32)*5/9
    return(res)
  }
> f2c(70)

[1] 21.11111
```

## How does it work? – high quality graphics

- ▶ 2D and 3D

## How does it work? – high quality graphics

- ▶ 2D and 3D
- ▶ statics and dynamics

## How does it work? – high quality graphics

- ▶ 2D and 3D
- ▶ statics and dynamics
- ▶ interactive

## How does it work? – high quality graphics

- ▶ 2D and 3D
- ▶ statics and dynamics
- ▶ interactive
- ▶ single or trellis plot



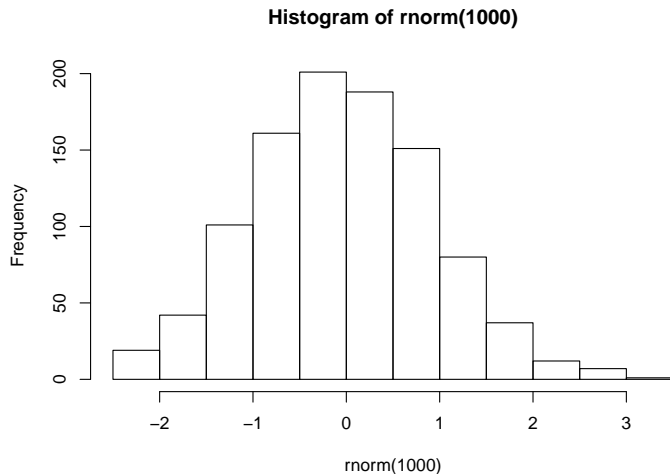
## How does it work? – high quality graphics

- ▶ 2D and 3D
- ▶ statics and dynamics
- ▶ interactive
- ▶ single or trellis plot
- ▶ high resolution printing

# How does it work? – high quality graphics

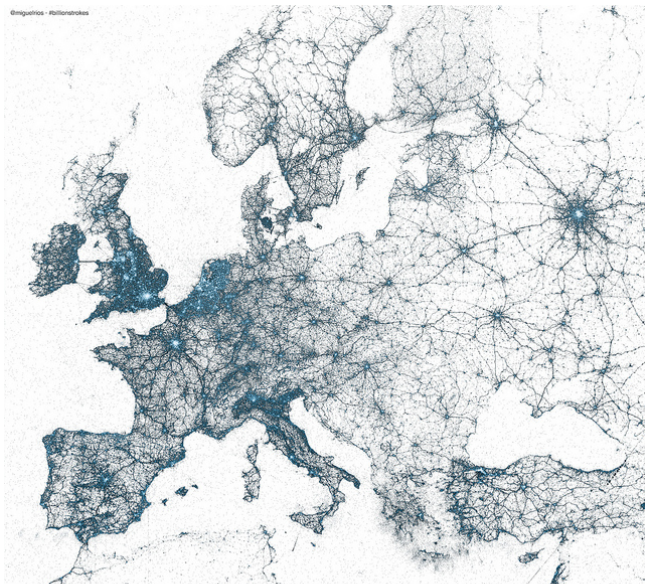
## Basic graphic

```
> hist(rnorm(1000))
```



## How does it work? – high quality graphics

Tweet traffic: every dot is a Tweet, and the color is the Tweet count from 2009 to May 2013. Twenty lines of R code using ggmap package



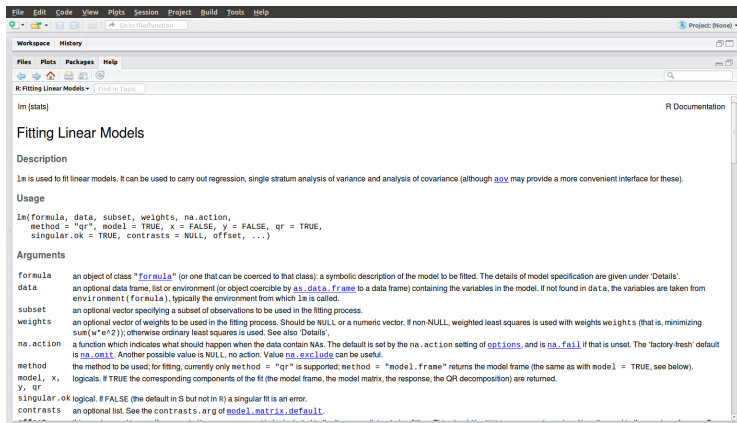
## How does it work?: other languages

R can interface with other languages:

- ▶ C / C++
- ▶ Fortran
- ▶ Java
- ▶  $\text{\LaTeX}$
- ▶ any program that can be command driven

# Any help? – inside R

> ?lm



The screenshot shows the R help window for the `lm` function. The title bar indicates the current project is '(None)'. The main content area is titled 'Fitting Linear Models' and includes a description, usage, and arguments.

**Description**

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although [aov](#) may provide a more convenient interface for these).

**Usage**

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

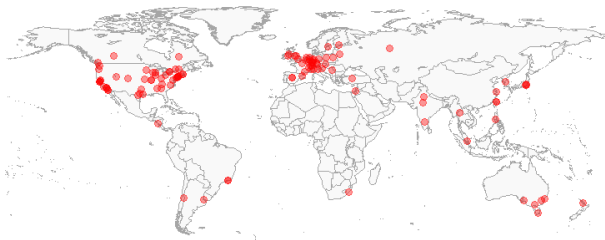
**Arguments**

<code>formula</code>	an object of class " <a href="#">formula</a> " (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.
<code>data</code>	an optional data frame, list or environment (or object coercible by <a href="#">as.data.frame</a> to a data frame) containing the variables in the model. If not found in <code>data</code> , the variables are taken from <code>environment(formula)</code> , typically the environment from which <code>lm</code> is called.
<code>subset</code>	an optional vector specifying a subset of observations to be used in the fitting process.
<code>weights</code>	an optional vector of weights to be used in the fitting process. Should be <code>NULL</code> or a numeric vector. If non- <code>NULL</code> , weighted least squares is used with weights <code>weights</code> (that is, minimizing $\sum(w \cdot e^2)$ ); otherwise ordinary least squares is used. See also 'Details'.
<code>na.action</code>	a function which indicates what should happen when the data contain NAs. The default is set by the <code>na.action</code> setting of <a href="#">options</a> , and is <code>na.fail</code> if that is unset. The 'factory-fresh' default is <code>na.omit</code> . Another possible value is <code>NULL</code> , no action. Value <code>na.exclude</code> can be useful.
<code>method</code>	the method to be used; for fitting, currently only <code>method = "qr"</code> is supported; <code>method = "model.frame"</code> returns the model frame (the same as with <code>model = TRUE</code> , see below).
<code>model, x, y, qr</code>	logicals. If <code>TRUE</code> the corresponding components of the fit (the model frame, the model matrix, the model matrix, the response, the QR decomposition) are returned.
<code>singular.ok</code>	logical. If <code>FALSE</code> (the default in S but not in R) a singular fit is an error.
<code>contrasts</code>	an optional list. See the <code>contrasts.arg</code> of <a href="#">model.matrix.default</a> .

## Any help? – RUG (R User Group)

- ▶ 56 RUGS in 25 countries
- ▶ 4 RUGS in France
  - ▶ semin-R – MNHN, INED, Univ. Paris-Descartes, Paris
  - ▶ R-Lyon, Univ. Lyon, Lyon
  - ▶ GUR – Cirad, Montpellier
  - ▶ fl\tauR – INSEE, Paris

R User Groups Worldwide



© Revolution Analytics

# Any help? – Blogs

452 blogs aggregated by R-bloggers.

## R-bloggers

R news and tutorials contributed by (452) R bloggers

[Home](#)[About](#)[add your blog!](#)[Contact us](#)[RSS](#)[R jobs](#)

### WELCOME!

Here you will find daily **news and tutorials about R**, contributed by over 450 bloggers. You can subscribe for e-mail updates:

Your e-mail here

Subscribe

10042 readers

BY FREQUENTER

And get updates to your Facebook:



R bloggers

J'aime

10 603 personnes aiment R bloggers.



If you are an **R blogger yourself** you are invited to add your own R content feed to this site

(Non-English R

## The R Backpages

November 7, 2013

By Joseph Rickert



by Joseph Rickert As an avid newspaper reader (I still get the print edition of the New York Times delivered every Sunday morning) I have always thought that some of the most interesting news is to be found in the back pages. So, in that spirit here are some things that I thought might be fit to print. Plotty...

Read

more +

## Webinar replay: What's new in Revolution R Enterprise 7

November 6, 2013

By David Smith

In case you missed yesterday's webinar, the slides and replay are now available for Introducing Revolution R Enterprise 7: The Big Data Big Analytics

## EXIF with R | rCharts + catcorrjs + exiftool

November 6, 2013

By kkr



### TOP 3 POSTS FROM THE PAST 2 DAYS

A Mitochondrial Manhattan Plot  
Ace Code Editor in Shiny (shinyAce)  
Installing R packages

Search & Hit Enter

### TOP 9 ARTICLES OF THE WEEK

1. Installing R packages
2. R and my divorce from Word
3. Using apply, sapply, lapply in R
4. Display googleVis charts within RStudio
5. What Hadley Wickham uses
6. Dream Team - combining Tableau and R
7. Data Preparation - Part I
8. Basics of Histograms
9. Select operations on R data frames

### SPONSORS



## Any help? – Conferences

- ▶ International: useR! [every year, last: University of Castilla-La Mancha, Spain]
- ▶ French: Rencontres R [Bordeaux 2012, Lyon 2013, Montpellier 2014]



Deuxièmes rencontres R

Lyon, 27 et 28 juin 2013



Accueil

Actualités

Conférenciers Invités

Programme

Tutoriels

Soumission

[Informations pratiques](#)

Inscription

Comités

Contact

Partenaires ▾

Recherche

### Informations pratiques

Les journées auront lieu à Lyon, les jeudi 27 et vendredi 28 juin 2013 sur le campus de la Doua. Une [série de tutoriels](#) sur des aspects de R spécifiques ou avancés seront proposés le mercredi 26 juin après-midi.

### Inscriptions et tarifs

Les inscriptions se font en ligne sur ce [site](#).

Le tarif des inscriptions est le suivant :

- **Tarif normal** (jusqu'au 27 mai) : 120 euros
- **Inscription tardive** (du 28 mai au 10 juin) : 150 euros
- **Tarif étudiant** (jusqu'au 10 juin) : 90 euros

### Hébergement et restauration

L'inscription inclut les deux pauses déjeuner (27 et 28 juin) et un apéritif dînatoire le 27 au soir.

Nous proposons, au moment de l'inscription, une solution d'hébergement sur le campus de la Doua (50 appartements étudiants à 60 euros la nuit).



## Any help? – and plenty of other things...



- ▶ discussion lists
- ▶ manuals
- ▶ tutorials
- ▶ courses
- ▶ books
- ▶ journals
- ▶ ...

## Any issue?

### Some known issues:

- ▶ Some code inconsistency
- ▶ Function redundancy
- ▶ Memory allocation not optimized (can be quite slow...)
- ▶ No reviewing process on packages
- ▶ Not optimized for parallel computing
- ▶ ...

## Any issue?

### Some known issues:

- ▶ Some code inconsistency
- ▶ Function redundancy
- ▶ Memory allocation not optimized (can be quite slow...)
- ▶ No reviewing process on packages
- ▶ Not optimized for parallel computing
- ▶ ...

### A solution?:

- ▶ R++, the next step, a project headed by Christophe Genolini (Université Paris Nanterre)

**THANK YOU!**