Getting starting with **R**

Alexandre Courtiol

Leibniz Institute of Zoo and Wildlife Research

June 2018

Getting started with **R**

About this course

I will:

- give you all the slides, so write only what is not being displayed
- show you what are the main principles of the **R** language
- provide you with short examples on how to use some of the most useful functions in **R**
- provide you with suggestions on how to learn more about **R** on your own

About this course

I will:

- give you all the slides, so write only what is not being displayed
- show you what are the main principles of the **R** language
- provide you with short examples on how to use some of the most useful functions in **R**
- provide you with suggestions on how to learn more about **R** on your own

You will not:

- be fluent in the **R** language after one week
- be able to remember how to implement many of the things we will see

About this course

I will:

- give you all the slides, so write only what is not being displayed
- show you what are the main principles of the **R** language
- provide you with short examples on how to use some of the most useful functions in **R**
- provide you with suggestions on how to learn more about **R** on your own

You will not:

- be fluent in the **R** language after one week
- be able to remember how to implement many of the things we will see

You should:

- accept that at the beginning it will be a little too abstract
- focus on the big picture (understand the logic, consider the long term gains)
- ask any silly question that pops up in your creative mind

Who am I?

- evolutionary biologist / statistician
- studies in France (Montpellier), postdoc in the UK (Sheffield)
- senior researcher at Leibniz IZW / lecturer at Freie University
- experience with **R**:
  - 2003 –: studying **R** (still ongoing)
  - 2008 –: using **R** most days
  - 2010 –: teaching **R**
  - 2016 –: developing **R** packages

Getting started with **R**

Getting started with R

R in brief

R is a programming language and software environment for statistical computing & graphics.

Key points about **R**:

- free for all
- open source (explore: `https://github.com/wch/r-source`)

R in brief

R is a programming language and software environment for statistical computing & graphics.

Key points about **R**:

- free for all
- open source (explore: `https://github.com/wch/r-source`)
- polyvalent:
  - from laptop to most advanced supercomputers
  - local or remote
  - Windows, MacOS, linux or many other Unix-based systems

R in brief

R is a programming language and software environment for statistical computing & graphics.

Key points about **R**:

- free for all
- open source (explore: `https://github.com/wch/r-source`)
- polyvalent:
  - from laptop to most advanced supercomputers
  - local or remote
  - Windows, MacOS, linux or many other Unix-based systems
- rich (tons of **R** packages out there)
- cutting edge (check updates for today: `http://dirk.eddelbuettel.com/cranberries/cran/updated/`)

R in brief

R is a programming language and software environment for statistical computing & graphics.

Key points about **R**:

- free for all
- open source (explore: `https://github.com/wch/r-source`)
- polyvalent:
  - from laptop to most advanced supercomputers
  - local or remote
  - Windows, MacOS, linux or many other Unix-based systems
- rich (tons of **R** packages out there)
- cutting edge (check updates for today: `http://dirk.eddelbuettel.com/cranberries/cran/updated/`)
- used by millions
- **R** is the best software environment for statistical computing, but it is far from perfect!

Getting started with R

A short history of **S**/**R**

S (`http://ect.bell-labs.com/sl/S/`)

- 1976-1980: version 1: interactive statistical system, Fortran based (Becker, Chambers, & al. at Bell Labs)
- 1980-1988: version 2: portable version (thanks to Unix)
- 1988: version 3 (**S**3): "everything is an object" paradigm, C-based (very much like R)
- 1991: a large statistical modeling toolbox is added to **S**3
- 1993: **S**+ exclusive license (to StatSci, later MathSoft, later SolutionMetrics)
- 1998: version 4 (**S**4): advanced object-oriented features
- 2012: **S**+ becomes TIBCO Enterprise Runtime for R (TERR)

A short history of **S**/**R**

S (`http://ect.bell-labs.com/sl/S/`)

- 1976-1980: version 1: interactive statistical system, Fortran based (Becker, Chambers, & al. at Bell Labs)
- 1980-1988: version 2: portable version (thanks to Unix)
- 1988: version 3 (**S**3): "everything is an object" paradigm, C-based (very much like R)
- 1991: a large statistical modeling toolbox is added to **S**3
- 1993: **S**+ exclusive license (to StatSci, later MathSoft, later SolutionMetrics)
- 1998: version 4 (**S**4): advanced object-oriented features
- 2012: **S**+ becomes TIBCO Enterprise Runtime for R (TERR)

R (`https://www.r-project.org/about.html`)

- 1993: the replication of **S** as the **R** project starts (Ihaka & Gentleman at University of Auckland)
- 23/04/1997: first version of **R** archieved on The Comprehensive R Archive Network (CRAN)
- 05/12/1997: **R** version 0.6 is part of GNU project ("freedom to share, freedom to chang")
- 29/02/2000: **R** version 1.0 (judged stable enough for production use by the R Development Core Team)

Getting started with R

Is **R** good for you?

Good for:

- data manipulation
- plots, including GIS
- analysing small, medium and big data
- programming around data

Is **R** good for you?

Good for:

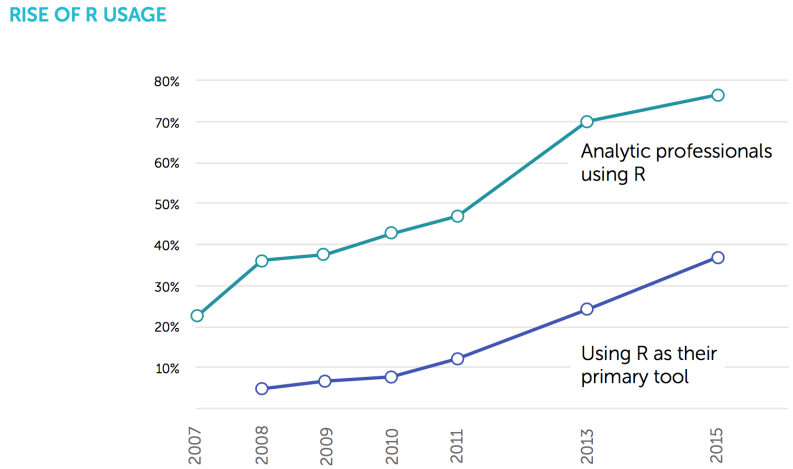
- data manipulation
- plots, including GIS
- analysing small, medium and big data
- programming around data

Not optimal for:

- beginners
- data entry
- formal algebra

Getting started with R

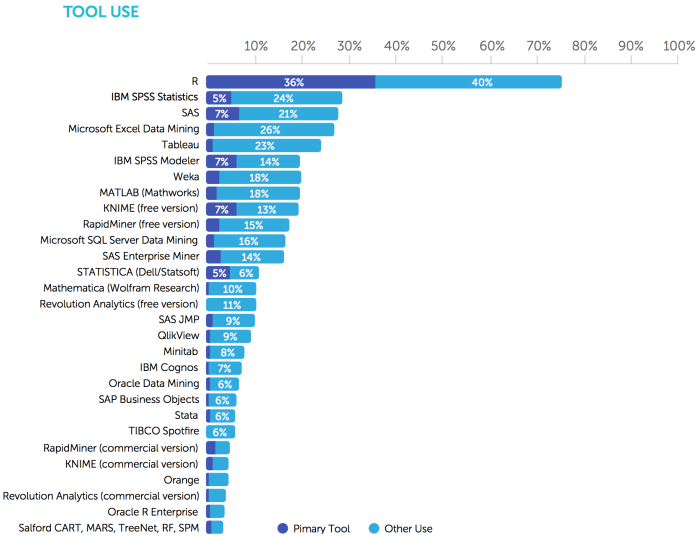
Who uses **R**?



Rexer Analytics

[1220 analytic professionals from 72 countries participated in this survey]

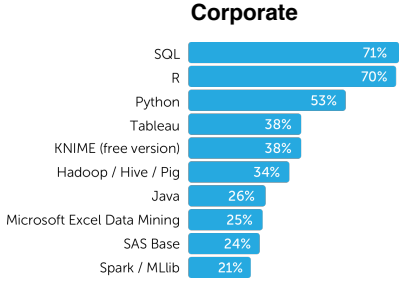
What else?

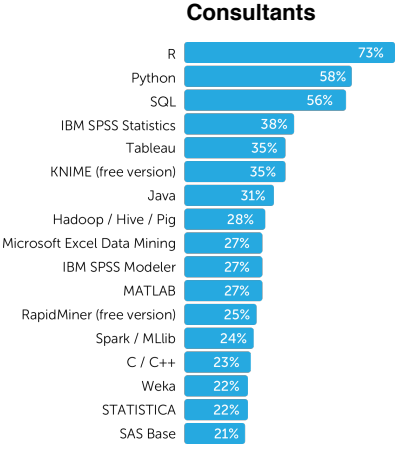


Rexer Analytics

Most Data Scientists use Multiple Tools

What data science / analytic tools, technologies, and languages did you use in the past year?

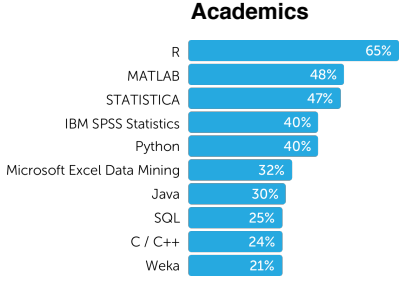


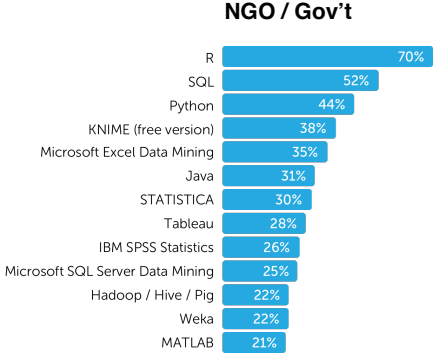


All tools used by more than 20% of a group are shown

Most Data Scientists use Multiple Tools

What data science / analytic tools, technologies, and languages did you use in the past year?





All tools used by more than 20% of a group are shown

Rich companies rely on **R** too!

Some examples:
(`http://blog.revolutionanalytics.com/2014/05/companies-using-r-in-2014.html`)

- Facebook (data analysis, big-data visualization, user behaviour analysis)
- Google (advertising effectiveness, economic forecasting, and big-data statistical modeling)
- Twitter (data visualization and semantic clustering)
- The City of Chicago (food poisoning monitoring)
- The New York Times (interactive features such as the Dialect Quiz and the Election Forecast)
- Microsoft (Xbox matchmaking)
- The Human Rights Data Analysis Group (counts of casualties in war zones)
- ANZ Bank (credit risk analysis)
- The FDA (regulatory drug approvals process)
- Monsanto (statistical analysis in plant breeding, fertility mapping and yield forecasting)
- Lloyds of London (risk analysis and catastrophe modeling)
- RealClimate.org (climate change analysis)
- NOAA (flood warnings)

Getting started with **R**

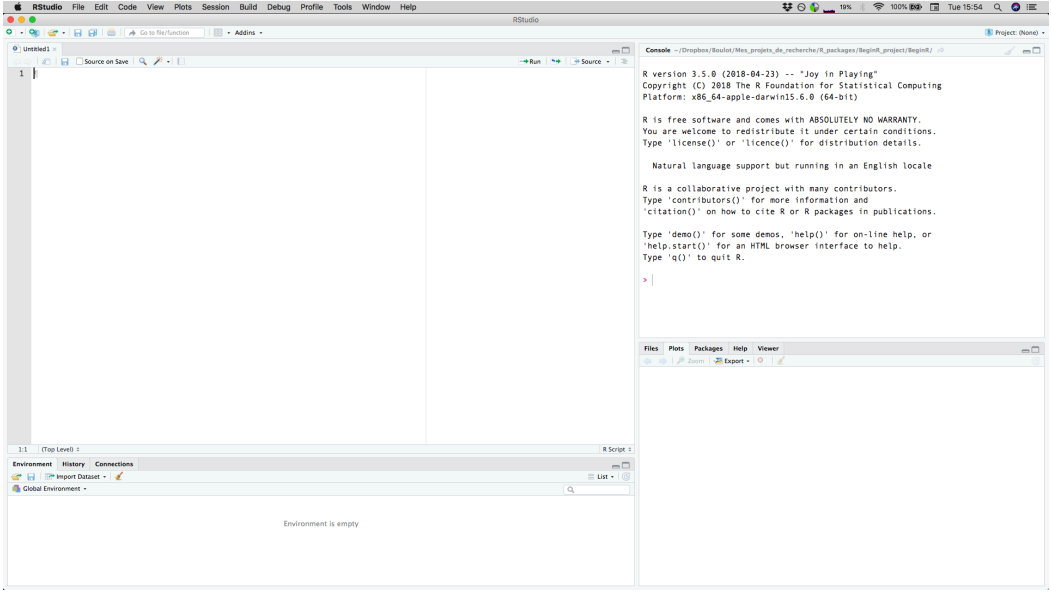
Getting started with R

Installation steps

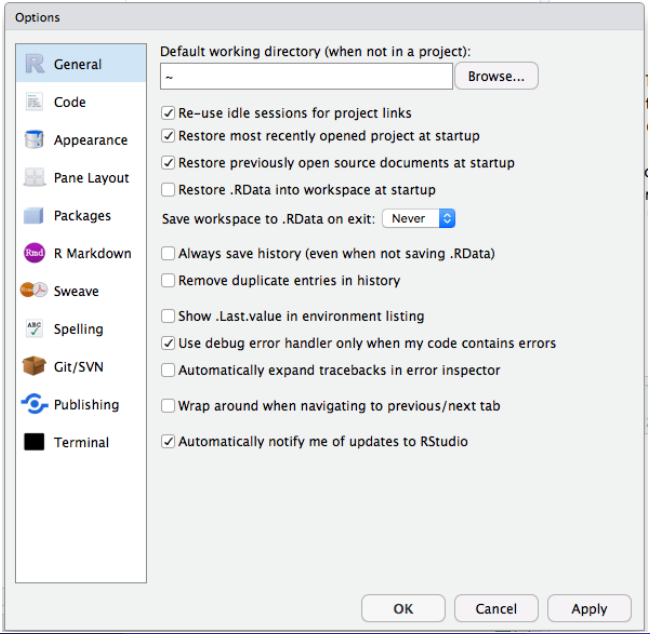
1. check that you do get internet access
2. install **R**: `https://cran.r-project.org/`
3. install RStudio: `https://www.rstudio.com/products/rstudio/download/`
4. open RStudio

Note: we will use RStudio but you don't have to (RStudio is free but not for all).

RStudio


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
RStudio
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
Console ~/Dropbox/Boulot/Mes_projets_de_recherche/R_packages/BeginR_project/BeginR/
R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-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.
>
Files Plots Packages Help Viewer
Zoom
Export
(Top Level)
R Script
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty

Better default setting for RStudio



Getting started with R

Basic arithmetic

Try in the following in the "Console" pannel:

```
1 + 1
## [1] 2
1 - 1
## [1] 0
2 * pi
## [1] 6.283185
3 / 2
## [1] 1.5
10 %% 3
## [1] 1
5^2
## [1] 25
5^2 + 1
## [1] 26
5^(2 + 1)
## [1] 125
```

Conclusion: you may never need a hand calculator anymore!

Getting started with R

The concept of an **R** script

All instructions must be written as a computer script!

- it is just a text file (no need for R to read it, it never gets corrupted)
- the script must be saved at a known location
- all non-**R** instruction must be preceeded by #

Why bother?

- transparent & reproducible
- easy to share & modify

Good practice

1. only use the "Console" pannel to mess around
2. write a script and comment it properly
3. make sure your script always work
4. store each result as an object with a useful name

Getting started with R

Creating objects

Objects are being assigned using the "arrow" operator:

```
one.plus.one <- 1 + 1 # storing the result
```

Objects are being used through their name (that is the whole point):

```
one.plus.one # displaying the result
## [1] 2
one.plus.one.plus.one <- one.plus.one + 1
one.plus.one.plus.one
## [1] 3
```

Tip:

```
(one.times.two <- 1 * 2) # storing and displaying the result
## [1] 2
```

Note 1: avoid spaces & weird characters in object names to avoid troubles.
Note 2: names are case sensitive.

Common mistakes

The huge majority of beginners problems are typos:

```
one.plus.one
## [1] 2
one.plus.two
## Error in eval(expr, envir, enclos): object 'one.plus.two' not found
one.plusone
## Error in eval(expr, envir, enclos): object 'one.plusone' not found
1 +
one.plus.one <- 1 + 1
## Error in 1 + one.plus.one <- 1 + 1: target of assignment expands to non-language object
```

The concept of an **R** object

What is an object?

- everything in **R** is an object
- objects have names
- objects allow abstraction
- objects belongs to classes for which specific methods exist (and can be created)

Note: we will come back on that later (for the programming session).

Note for geeks who know other computer languages

R objects are (by default) not mutable (there is copy on demand):

```
a <- 1
b <- a
b <- b + 1
b
```

```
## [1] 2
```

```
a
```

```
## [1] 1
```

Getting started with R

Functions

```
citation()  # function showing how to cite R
```

```
##
## To cite R in publications use:
##
##   R Core Team (2018). R: A language and environment
##   for statistical computing. R Foundation for
##   Statistical Computing, Vienna, Austria. URL
##   https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2018},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating
## R, please cite it when using it for data analysis.
## See also 'citation("pkgname")' for citing R packages.
```

```
help(citation) # getting help for this function
?citation() # same but shorter (syntactic sugar)
```

Functions

```
mean()
```

```
?mean()
```

```
Usage:
     mean(x, ...)
     ## Default S3 method:
     mean(x, trim = 0, na.rm = FALSE, ...)

Arguments:
       x: An R object.  Currently there are methods for numeric/logical
          vectors and date, date-time and time interval objects, and
          for data frames all of whose columns have a method.  Complex
          vectors are allowed for 'trim = 0', only.
    trim: the fraction (0 to 0.5) of observations to be trimmed from
          each end of 'x' before the mean is computed.  Values of trim
          outside that range are taken as the nearest endpoint.
   na.rm: a logical value indicating whether 'NA' values should be
          stripped before the computation proceeds.
[...]
```

Syntax for functions

Basic syntax:

```
sign(x = -5)
## [1] -1
sign(-5) # dangerous: avoid!
## [1] -1
sign(y = -5)
## Error in sign(y = -5):  supplied argument name 'y' does not match 'x'
```

Equal signs and arrows are not the same:

```
sign(y <- -5) # dangerous: avoid!
## [1] -1
y
## [1] -5
sign(x = y <- -5) # same as above
## [1] -1
```

Syntax for functions

Not puting the parentheses shows the definition of the function:

```
sign
## function (x)  .Primitive("sign")
```

All functions need parentheses and exceptions correspond to syntactic sugar:

```
1 + 1
## [1] 2
`+`(1, 1)
## [1] 2
a <- 1
a
## [1] 1
`<-`(a, 1)
a
## [1] 1
```

Key principles of the **R** language

- Everything that exists in **R** is an object
- Everything that happens in **R** is a function call

John M. Chambers

Finding functions

To find the name of the function you are look for, you may try:

```
??"linear model"
```

or

```
help.search(pattern = "linear model", package = "stats")  # if you know where to look for
```

Getting started with R

The concept of an **R** package

Packages extend **R** functionalities:

- for most users; e.g. `ggplot2`
- for specific users; e.g. `IsoriX`
- for developpers; eg. `Rcpp`

Key facts about packages:

- a package is a folder (often compressed) containing **R** functions, data & documentation
- a library is an installed package
- there are tons of packages out there:
  - 12606 packages are available on `cran.r-project.org`
  - $\sim$ 1500 packages aimed at bioinformatic on `bioconductor.org`
  - many more on `github.com`
  - many more shared between users in other ways

Note: packages can be used to create research compendia!

Installing a package

Simple situation: the package is available as binary for your system on CRAN

```
install.packages("dpylr")
```

In general, the installation procedure depends on:

- where the package is being hosted (local, CRAN, bioconductor, GitHub, other)
- if the package contains sources in another language that have been compiled or not (yes, no)

In order to be able to install packages that require compilation (and thus have access to more or newer version of packages), you need to install:

- Rtools if you use Windows (`https://cran.r-project.org/bin/windows/Rtools/`)
- Xcode if you use macOS (`https://developer.apple.com/xcode/`)
- nothing if you use Linux or other Unix-based system

Installing the package for this course

The package is not on CRAN as I want to be able to update it instantaneously and have potentially large files.

I host the package here: `https://github.com/courtiol/BeginR`

You should install it as follows:

```
install.packages("drat")
drat::addRepo("courtiol")
install.packages("BeginR")
```

Loading a package

Loading a package makes the exported functions of the package and its data (if lazy-loaded) available to the user.

Example:

```
library(BeginR)
##
##  ########################################
##  #                                      #
##  #  This is the package for the course  #
##  #                                      #
##  #        'Getting Started with R'      #
##  #                                      #
##  #      Version 0.0.0.9000 installed    #
##  #                                      #
##  #    To access the slides, just type   #
##  #                                      #
##  #  browseVignettes(package = 'BeginR')  #
##  #                                      #
##  ########################################
```

Getting started with R

Updating R packages

Some things to know:

- **R** packages evovle quickly
- young **R** packages can be very buggy
- packages are not reviewed

Good practice:

- update your **R** packages daily

```
update.packages(ask = FALSE)
```

- check what is being changed if you heavily rely on a recent package
- contact the maintainer when you spot bugs (but write minimal reproductive examples!)

Updating R

Some things to know:

- **R** has many bugs (like all other software)
- **R** bugs are reported, discussed and solved in the open (unlike most other software): `https://bugs.r-project.org/bugzilla3/`
- each new version of **R** is in general more efficient, richer, and less buggy

What to do?

- check for **R** new versions on CRAN
- check for what has changed if you fancy (`https://cran.r-project.org/index.html`)
- install the new version of **R** (unless it is not a minor update that you don't need)
- re-install all your packages

Note 1: some packages can help to do this: `InstallR` on Windows and `UpdateR` on macOS.
Note 2: also update RStudio for full compatibility with **R**.

Getting started with R

Useful resources

- the help files and official documentation (boring but thorough `https://cran.r-project.org/manuals.html`)
- a nice blog: `http://www.r-bloggers.com/`
- 2 **R** local meetup groups: `https://www.meetup.com/Berlin-R-Users-Group/` & `https://www.meetup.com/rladies-berlin/`

Books:

Other way to learn

Afer having learned some basics, just open the console and test your understanding by performing experiments!