

# Data Analysis and Unsupervised Learning

## Introduction to R

MAP573 – Julien Chiquet

École Polytechnique, Autumn semester, 2020-2021

<https://jchiquet.github.io/MAP573>



# Outline

- ① What is R?
- ② Why R?
- ③ A dummy Rsession
- ④ R Markdown

# Outline

- 1 What is R?
- 2 Why R?
- 3 A dummy Rsession
- 4 R Markdown

# R ?

## In a nutshell

R is a scientific software specialized in calculation and statistical analysis.

It is also

- a programming language,
- a environment/interpreter,
- an open-source project (GNU-R)
- a multi-plateform software (Linux, Mac, Windows)

## A bit of history

- 1970s: S-language developed at Bell labs (Chambers, Beckers)
- 1980s: S-PLUS developed at AT&T. Lab
- 1990s: R is developed as a GNU/GPL open-source counterpart to S by Gentleman and Ihaka (Auckland university)
- 1997: The R-core team now leads the development
- 2002: The R fondation is created and chaired by Gentleman and Ihaka
- 2011: first public of R-studio (JJ Allaire)
- 2019: Rstudio lead scientist H. Wickham receives COPSS Award (statistician Nobel price)

# Remarkable basics features

## Scientific Computing

- linear algebra
- statistical models and data analysis

## Data manipulation and visualization

- import, export, transformation
- great, versatile plotting system

## Interfacing is easy

- to most programming languages (C/C++, Python)
- to most database systems (SQL, postgres)
- for distributed computing (Hadoop, H2O, spark)

## Package manager

- Extremely versatile

# Why R ?

## Community

- SatRDay, R user groups, meet-up, conference
- CRAN community <https://cran.r-project.org/>
- Rstudio community <https://community.rstudio.com/>
- R dev/package well integrated on [github](#)

## Packages manager

- more than 13,000 community-based libraries
- cutting-edges statistical methods
- easy to learn even for non-statistician/data scientist

## Reproducibility

- Rmarkdown is not just notebook
- Great for interfacing, plotting, scientific reports

# Why not R?

- Easy to make dirty code (less and less true)
- Not typed language, not compilation by default: may be slow
- Many ways to do the same things
- Less well interfaced to ML/Deep-learning library than Python



# Why R again?

## The Rstudio group

Even if it is a company...

- Rstudio API is a great all-in-one tool for data analysis and development
- Cleaner implementation (tidyverse and co)
- New functionalities (unitary test, github integration)
- Interface to deep learning tools (Tensor Flow, Keras, Torch, etc.)
- Interface with Python (reticulate)
- Nice surrogate Oriented-Object programming with R6

~> Rstudio basically saved R from Python

# Outline

- ① What is R?
- ② Why R?
- ③ A dummy Rsession
- ④ R Markdown

# The Rstudio API

- A full API with code, interpreter, workspace and plots
- Package developement and external code integration are easier
- Notebooks integration with Rmarkdown
- Interface with github

⇒ Rstudio is a state-of-the-art tool for efficient development in R

# My favorites shortcuts

- `ctrl + return`: execute current selection in console
- `ctrl + 1/2/3/4`: navigate between panels
- `ctrl + down/up`: navigate between tabs
- `ctrl + shift + k`: knit current document

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The main editor window shows a script with the following code:

```
1 k <- 2+2  
2 y <- 3+5  
3
```

The Environment panel on the right shows the Global Environment with the following values:

Variable	Value
x	4
y	8

The Files panel on the right shows the User Library with the following packages and versions:

Package Name	Description	Version
acepack	ace() and avar() for selecting regression transformations	1.3-3.3
aricode	Compute rand index	2015.06.12
BiocInstaller	Install/Update Bioconductor and CRAN Packages	1.18.2
biotools	Tools for Biometry and Applied Statistics in Agricultural Science	2.1
bitops	Bitwise Operations	1.0-6
blockseg	Two dimensional change-points detection	1.0
blocseg	Two dimensional change-points detection	1.0
car	Companion to Applied Regression	2.0-25
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
cgdsr	R-Based API for accessing the MSKCC Cancer Genomics Data Server (CGDS)	1.1.33
clusterpath	Fast agglomerative convex clustering, non-Rcpp implementation	1.2
colorspace	Color Space Manipulation	1.2-6
crayon	Colored Terminal Output	1.2.1
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Cryptographic Hash Digests of R Objects	0.6.8

The Console panel at the bottom shows the following output:

```
~/Documents/Teachings/2015-2016/L3_GBI/ISV51/td1-issv51/ >  
en ligne ou "help.start()" pour obtenir l'aide au format HTML.  
Tapez 'q()' pour quitter R.  
  
WARNING: Your CRAN mirror is set to "http://cran.rstudio.com/" which has an insecure (non-HTTPS) URL. The repository was  
likely specified in .Rprofile or Rprofile.site so if you wish to change it you may need to edit one of those files. You  
should either switch to a repository that supports HTTPS or change your RStudio options to not require HTTPS download  
ds.  
  
To learn more and/or disable this warning message see the "Use secure download method for HTTP" option in Tools -> Glob  
al Options -> Packages.  
> 2  
[1] 2  
> 2+2  
[1] 4  
k <- 2+2
```

# Outline

- ① What is R?
- ② Why R?
- ③ A dummy Rsession
- ④ R Markdown

# A dummy Rsession

# Outline

- ① What is R?
- ② Why R?
- ③ A dummy Rsession
- ④ R Markdown

# R Markdown



Figure 2: an authoring framework for data science



# R Markdown?

- Markdown is a *lightweight markup language* with plain text formatting syntax that can be converted to HTML. It is completely independent from R. The extension is typically `.md`.
- R Markdown is an *extension of the markdown syntax* that enables R code to be executed. The extension is typically `.Rmd`.
- `rmarkdown` is a library/package which processes and converts `.Rmd` files into a number of different formats, including HTML or `.pdf`. The core function is `rmarkdown::render()`.
- `knitr` is a library/package which processes plain text document with embedded code, executes the code and 'knits' the results back into the document. The core function is `knitr::knit()`.

```
install.packages("rmarkdown")  
install.packages("knitr")
```

# How does it work?

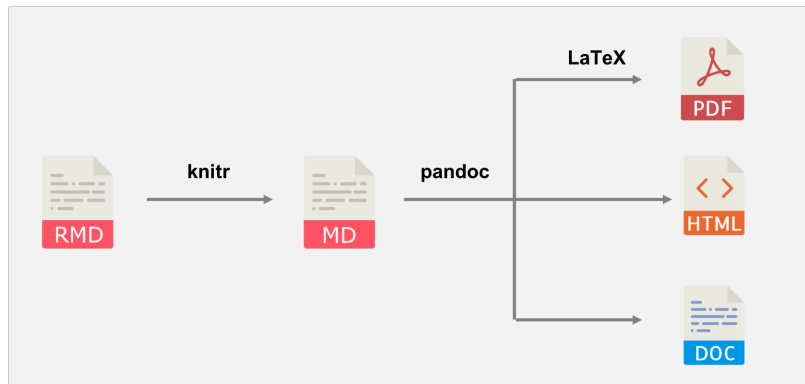
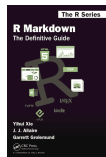


Figure 3: R Markdow workflow

# References

Rmarkdown: Dynamic Documents for R (Allaire et al., 2020),  
<https://bookdown.org/yihui/rmarkdown/>



Knitr: A General-Purpose Package for Dynamic Report Generation in R (Xie, 2020), <https://yihui.name/knitr/>



Rstudio doc

See <https://rmarkdown.rstudio.com/>

# R Markdown possibilities

See <https://rmarkdown.rstudio.com/>

## Handle various inputs

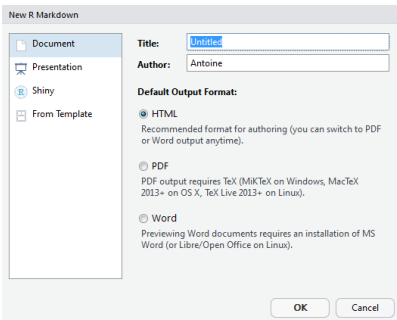
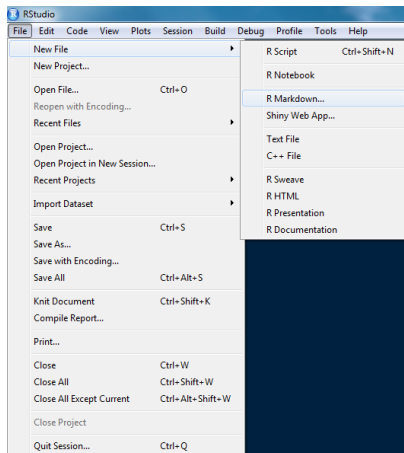
- Markdown Syntax ([Markdown reference cheat sheet](#))
- $\text{\LaTeX}$ (Advanced mathematical expressions)
- HTML/javascript
- Code chunks (R, Python, Julia and more)

## Handle various output

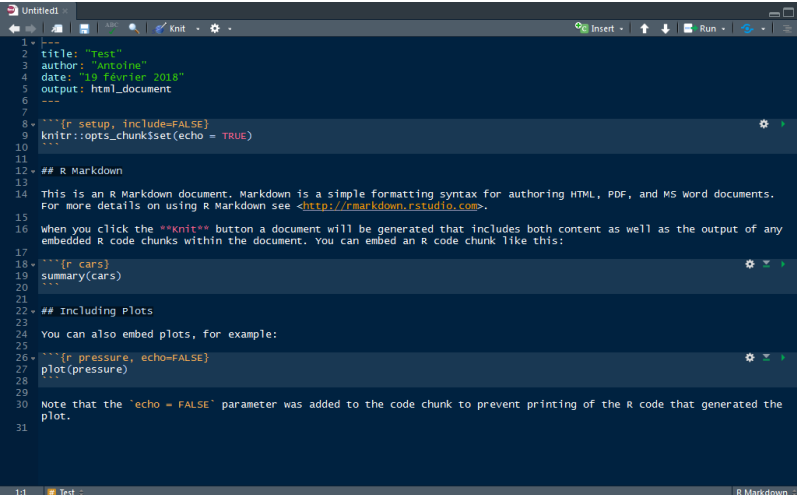
- Rstudio Notebook
- HTML report (static, dynamic)
- HTML website (static, dynamic)
- PDF document
- Doc documents

~> More than a Jupyter notebook

# Create a new .Rmd



# New .Rmd



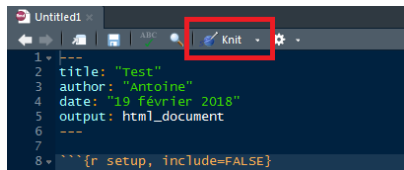
The screenshot shows the RStudio interface with a new R Markdown document titled 'Untitled1'. The editor has a dark theme. The code is as follows:

```
1 ---
2 title: "Test"
3 author: "Antoine"
4 date: "19 février 2018"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS word documents.
15 For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
16
17 when you click the **knit** button a document will be generated that includes both content as well as the output of any
18 embedded R code chunks within the document. You can embed an R code chunk like this:
19
20 ```{r cars}
21 summary(cars)
22 ```
23
24 ## Including Plots
25
26 you can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30 ```
31
32 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the
33 plot.
```

At the bottom of the window, the status bar shows '1:1 Test' on the left and 'R Markdown' on the right.

# Compile .Rmd

Use the Knit button to produce a HTML file



Shortcut: Ctrl + Maj + K

# References

Many ideas/examples inspired/stolen from the following books:

Advanced R (Wickham, 2014), <http://adv-r.had.co.nz/>



A Language and Environment for Statistical Computing (R Core Team, 2017), <https://www.R-project.org/>



Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... Chang, W. (2020). *Rmarkdown: Dynamic documents for R*. Retrieved from <https://bookdown.org/yihui/rmarkdown>

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>