Data Analysis and Unsupervised Learning

Introduction to R

MAP573 – Julien Chiquet

École Polytechnique, Autumn semester, 2020-2021

https://jchiquet.github.io/MAP573

Outline

Outline

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: `R`studio lead scientist H. Wickham receives COPSS Award (statistician Nobel price)

Remarkable basics features

Scientific Computing

- linear algebra
- statistical models and data analysis

Data manipulation an visualization

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

Interfacing is easy

- to most programming languages (C/C++, Python)
- to most database systems (SQL, postgrey)
- for distributed computing (Hadoop, H20, 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

$\rightsquigarrow$ Rstudio basically saved `R` from `Python`

Outline

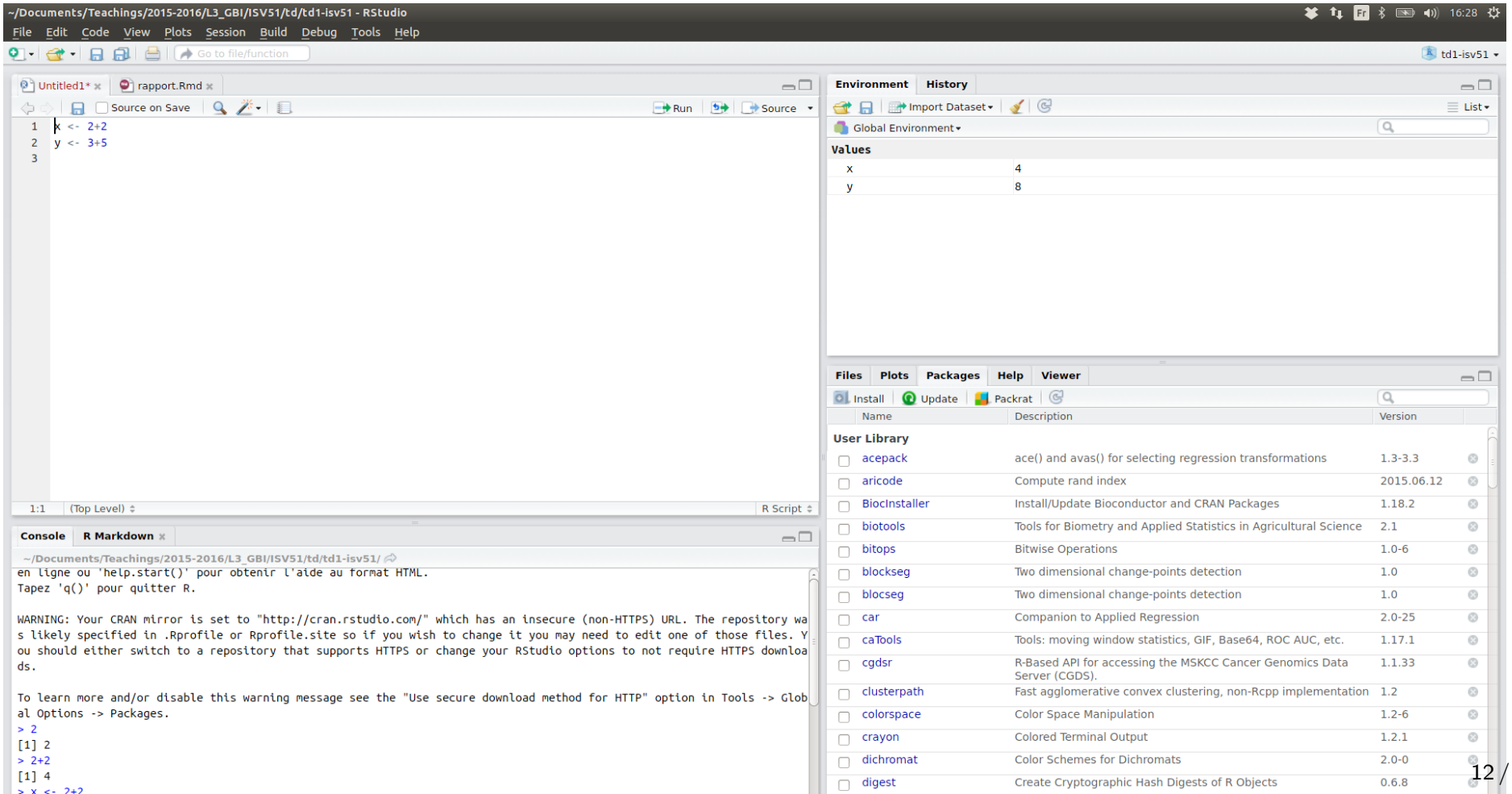
The Rstudio API

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

$\rightsquigarrow$ Rstudio is a state-of-the-art tool for efficent 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 doccument



Outline

A dummy Rsession

Outline

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 extention 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()`.

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

How does it work?

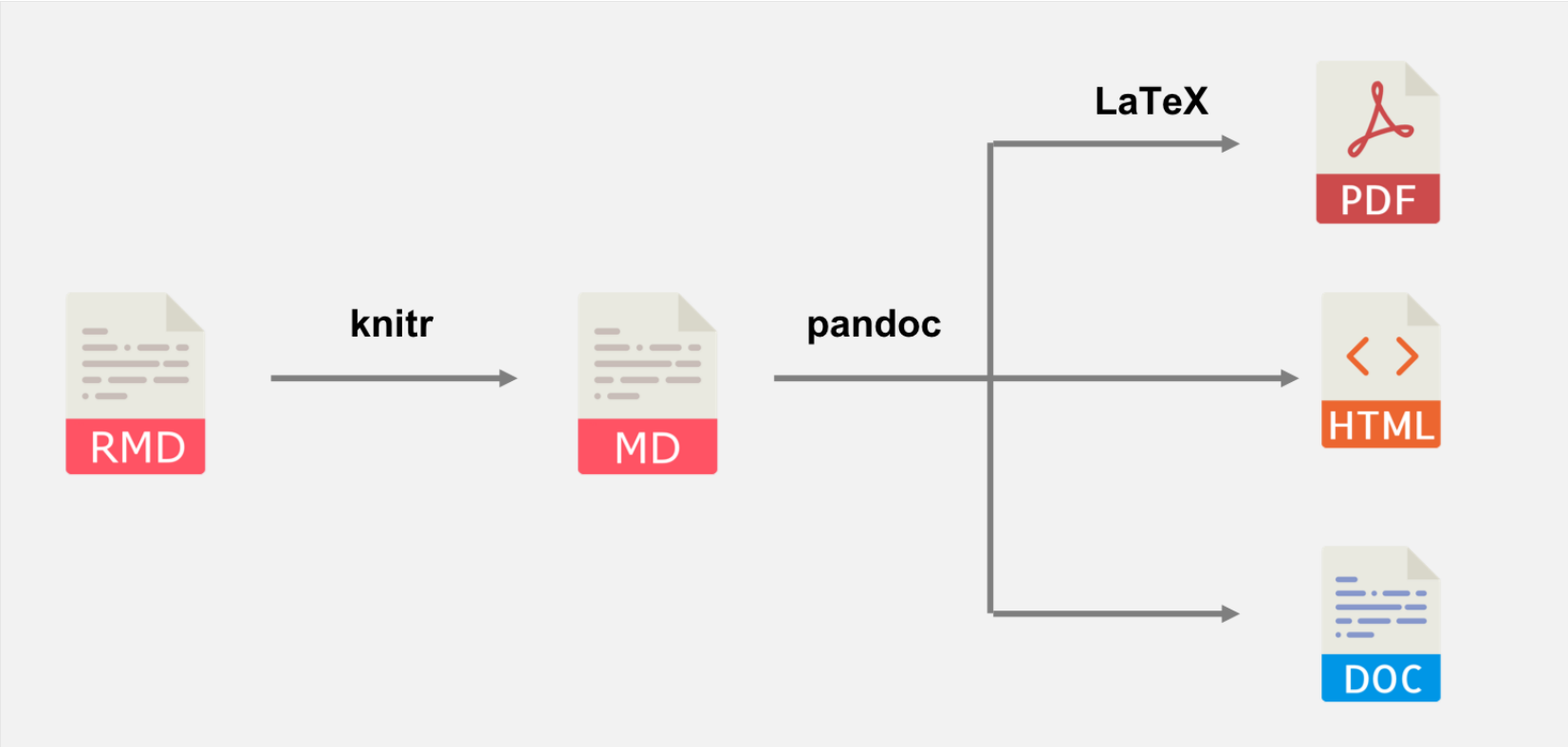
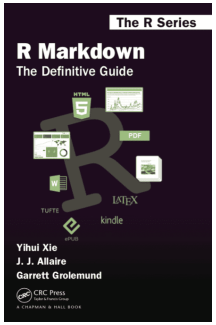


Figure 3: R Markdow workflow

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

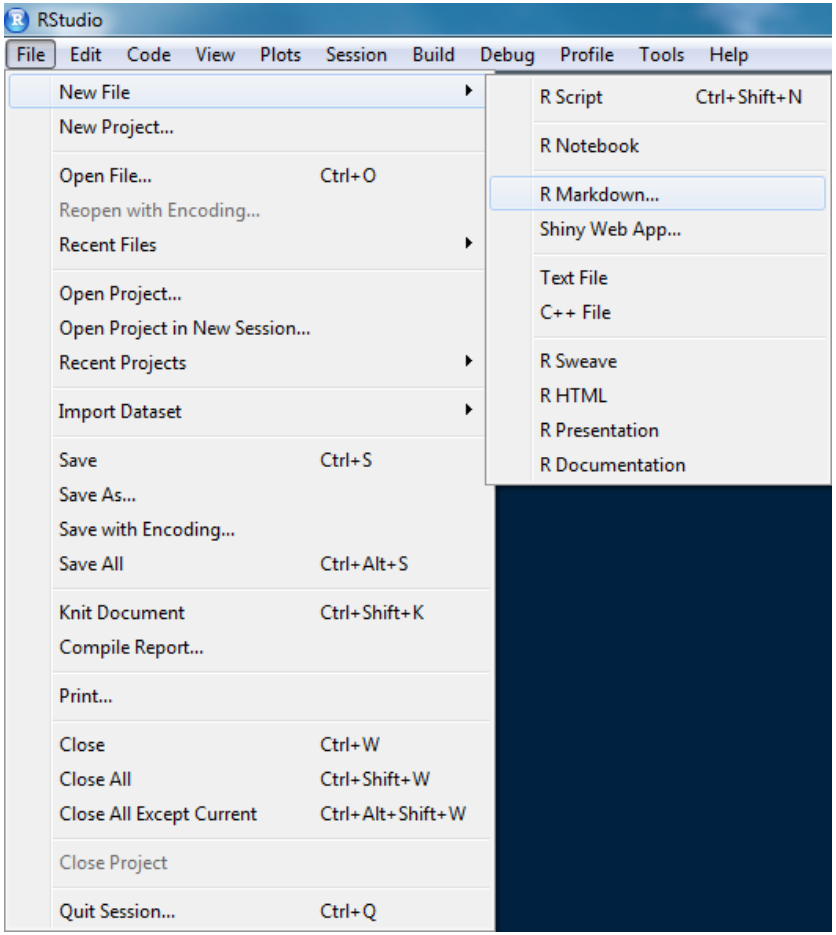
- Markdown Syntax (Markdown reference cheat sheet)
- LaTeX(Advanced mathematical expressions)
- HTML/javascript
- Code chunks (R, Python, Julia and more)

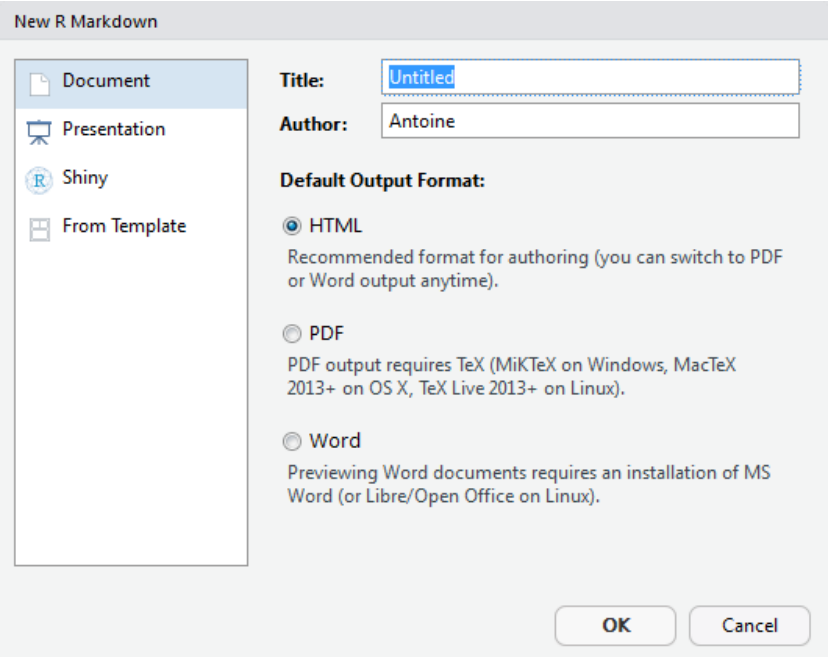
Handle various output

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

$\rightsquigarrow$ More than a Jupyter notebook

Create a new `.Rmd`





New .Rmd

Untitled1

Knit Insert Run

```
---
title: "Test"
author: "Antoine"
date: "19 février 2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

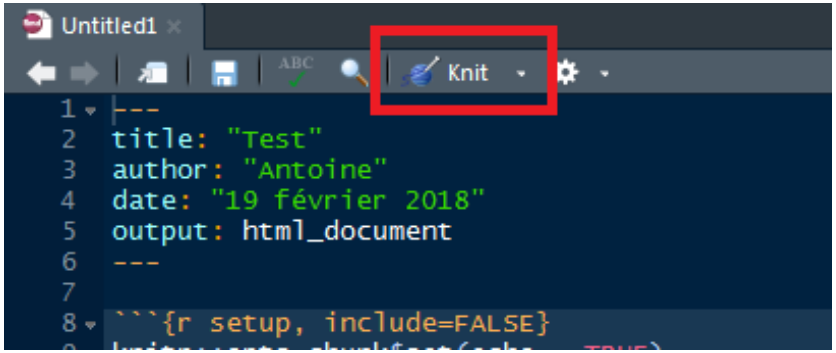
```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```

1:1 Test R Markdown

Compile `.Rmd`

Use the `Knit` button to produce a HTML file

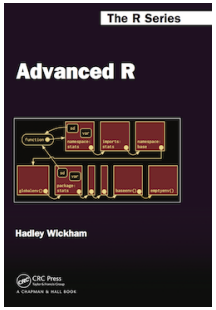


Shortcut: Ctrl + Maj + K

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