

Collaborative Reproducible Research with the R Ecosystem

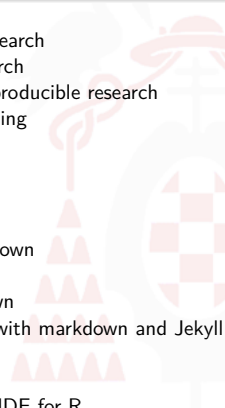
Daniel Rodriguez, David F. Barrero

Departamento de Automática
Universidad de Alcalá



/gso>

Table of Contents

- 
- 1 Introduction
 - The practice of research
 - Reproducible research
 - Components of reproducible research
 - Literate Programming
 - T_EX and L_AT_EX
 - 2 Markdown
 - First contact
 - Using markdown
 - “Advanced” markdown
 - Metadata
 - Academic markdown
 - Creating websites with markdown and Jekyll
 - 3 R Ecosystem
 - The R language
 - R Studio as Open IDE for R
 - Packages for Reproducible Research
 - 4 Collaborating and sharing your research
 - Git
 - 5 Conclusions

Introduction

The practice of research (I)

What is research?

- In short, writing papers

Many processes involved

- Code development
- Dataset collection
- Scripts for everything
- Figures creation
- Statistics and analysis

The review comes, after 6 months

- ... you must repeat an experiment
- ... or change a figure or ...

Introduction

The practice of research (I)

What is research?

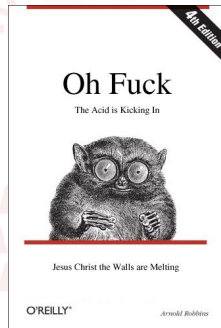
- In short, writing papers

Many processes involved

- Code development
- Dataset collection
- Scripts for everything
- Figures creation
- Statistics and analysis

The review comes, after 6 months

- ... you must repeat an experiment
- ... or change a figure or ...



Introduction

The practice of research (II)

Even worse: Imagine you need to reproduce other's experiments

- Or any one else reproduce your experiments ... **in teams**

Need of many assets

- Experiment run
- Data analysis
- Experiment documentation
- Paper authoring
- **On-line publication**

Need of procedures and tools to handle this

Our objective

Formalize the whole publishing *collaborative* process

Introduction

Reproducible research

“Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.”

<http://reproducibleresearch.net/>

Reproducible research is a cornerstone of research

- Science **must** be reproducible
- Repetible \neq reproducible

Reproducible research is good

- For science
- For you

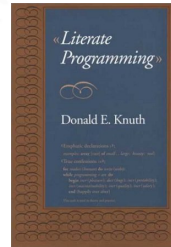
Introduction

Components of reproducible research

Task	Tool
Code development	SVN, Git
Collaboration	Redmine, GitHub
Data format	CSV, YAML, JASON
Experiment run	Python, Lua
Statistical analysis	R, RStudio
Documentation	Markdown
Results writting	Bookdown, Knitr, Swave
Assets publishing	GitHub

Literate Programming

Literate programming is an approach to programming introduced by D. Knuth in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated.



- https://en.wikipedia.org/wiki/Literate_programming

T_EX

T_EX(= tau epsilon chi, and pronounced similar to "blecch", not to the state known for 'Tex-Mex' chili) is a computer language designed for use in typesetting; in particular, for typesetting math and other technical (from Greek "techne" = art/craft, the stem of 'technology') material.
(More info)

- It has been around for many years...
-

T_EX and L^AT_EX

L^AT_EX, which is pronounced «Lah-tech» or «Lay-tech» (to rhyme with «blech» or «Bertolt Brecht»), is a document preparation system for high-quality typesetting. It is most often used for medium-to-large technical or scientific documents but it can be used for almost any form of publishing.

<http://www.latex-project.org/about/>

- It also has been around for many years...
-

Markdown

First contact (I)

Markdown: Trivial markup

- Simple
- Very simple
- Extremely simple
- Did I say it's simple?

VERY powerful

- Several outputs
- Professional quality
- ... and simple!

Markdown example

I am a header

I am a subheader

Regular , *italic* and **bold**

– List item 1

– List item 2

[I am a link](<http://foo.com>)

![I am a pic](markdown.png)

~~~C

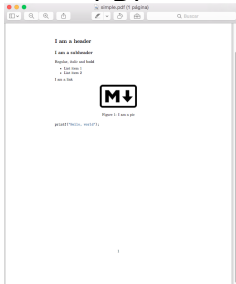
```
printf("Hello , world");
```

~~~

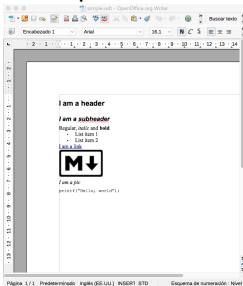
Markdown

First contact (II)

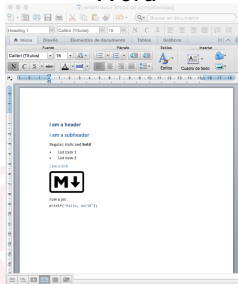
PDF



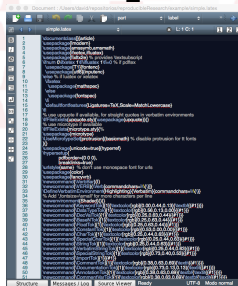
OpenOffice



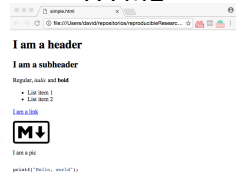
Word



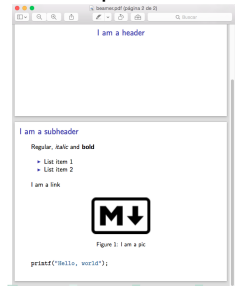
L^AT_EX



HTML



Beamer presentation



Markdown

Using markdown

The easy style: Markdown editors

- Desktop: Remarkable, MacDown, ...
- Online editors: (Dillinger), (StackEdit)

The hacker style: Command-line markdown parser, **pandoc**

- 1 Edit file with any text editor
- 2 “Compile” the document
 - `pandoc simple.md -o simple.pdf`
 - `pandoc simple.md -o simple.docx`

The professional style: Integrated in a ecosystem

- GitHub
- R markdown, Knitr, ...
- iPython, Jupiter, ...

Markdown

“Advanced” markdown (I)

Markdown is loosely defined

- Several elements not defined in original markdown
- Variations among implementations
- We follow GitHub markdown style

Lists

```
* Bullet 1
+ Cross 1
  - Nested item 1
- Nested item 2
1. Ordered item 1
2. Ordered item 2
```

Tables

Header 1	Header 2
Cell (1,1)	Cell (1,2)
Cell (2,1)	Cell (2,2)

Markdown

“Advanced” markdown (II)

Blockquotes

Don Quixote said:

> Until death it is all life

Mentions

@username

Used also to reference

Maths

Same syntax than \LaTeX

$\text{\$}\sqrt{\text{\frac{a}{b}}}\text{\$} \rightarrow \sqrt{\frac{a}{b}}$

$\text{\$}\lim_{x \rightarrow \infty} e^{-x}\text{\$} \rightarrow \lim_{x \rightarrow \infty} e^{-x}$

(Mastering Markdown) (Markdown Cheatsheet)

Markdown

Metadata (I)

Metadata: Data about data

- In our case title, authors, affiliation, etc
- Header (Valid in pandoc)

```
title:  'Constructing and reconstructing childhood:  
        Pursuit of the perfect donut'
```

```
author:
```

- Homer J. Simpson
- Lisa Simpson

```
institute:
```

- University of Springfield

```
date: "January, 1, 2016"
```

```
tags: [nothing, nothingness]
```

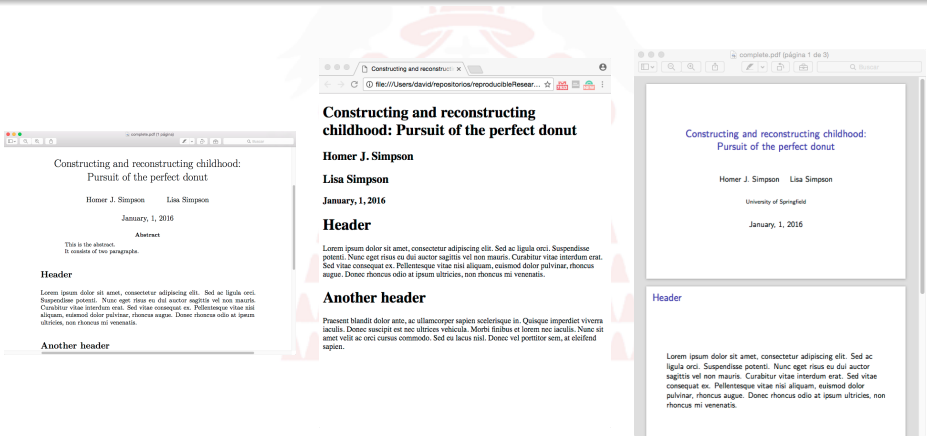
```
abstract: |
```

```
    This is the abstract.
```

```
    It consists of two paragraphs.
```

```
...
```


Markdown Metadata (II)



Markdown

Metadata (III)

Template

```
---  
title: 'My title'  
author:  
- Author 1  
- Author 2  
institute:  
- Institution  
...  
  
My text goes here
```

Markdown

Academic markdown

TODO



Markdown

Creating websites with markdown and Jekyll

TODO
(Better located after Github?)



What is R?

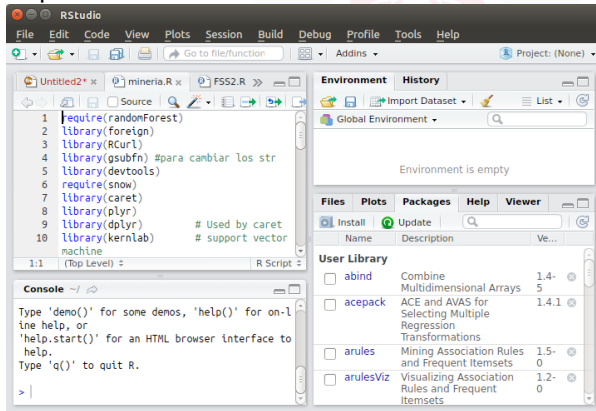
R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

Wikipedia:

[https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

What is RStudio?

Popular IDE for R.



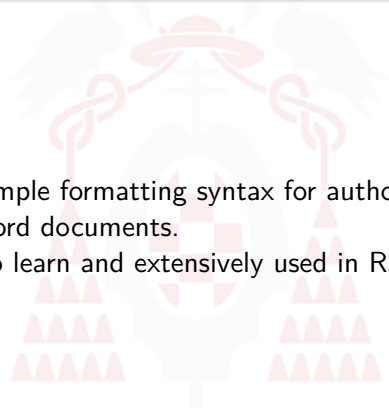
<https://www.rstudio.com/>

RStudio?

Well known package authors work for RStudio.



Markdown



Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents.
Extremely easy to learn and extensively used in RStudio.

knitr

knitr is an engine for dynamic report generation with R. It is a package in the statistical programming language R that enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents.

<https://en.wikipedia.org/wiki/Knitr>

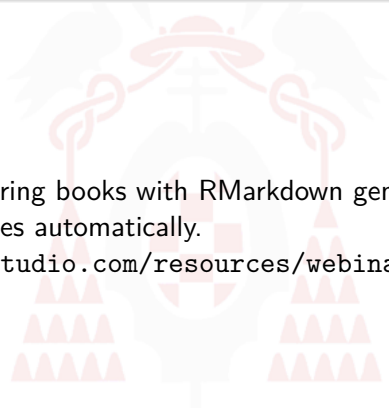
knitr extracts R code in the input document, evaluates it and writes the results to the output document

Rnw, Markdown, HTML and LaTeX

<http://yihui.name/knitr/>

<http://yihui.name/knitr/demo/minimal/>

Bookdown



It is way of authoring books with RMarkdown generating PDF, handouts and slides automatically.

<https://www.rstudio.com/resources/webinars/introducing-bookdown/>

Rnw

Inserts R into Latex documents.



Git and GitHub



Git

Git

[https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

GitHub

GitHub

www.github.com

Publish Websites, etc.



Conclusions

- Learn Bash and other tools for it: awk, make, sort,
- Learn Git, R or Python.
- If you use R/Python for your research automate as much as possible. It is much easier to modify.
- Try to stick to Open/libre software GNU/Linux, R, Python, etc.
- Etc.