Reproducible research

Module 12

Leonardo Collado-Torres

June 19, 2015

Reproducible research

- Have you heard about it?
- Maybe you've heard about the Cousera Reproducible Research course
- How would you define it?
- Is it the same as replicability?
- Is it important?

```
Science moves forward then discoveries are replicated
and reproduced
```

Implementing Reproducible Research

Replication

```
Replication, the practice of independently implementing
scientific experiments to validate specific findings,
 is the cornerstone of discovering scientific truth.
```

Implementing Reproducible Research

Reproducibility

```
Reproducibility can be thought of as a different
standard of validity from replication because it
forgoes independent data collection and uses the
methods and data collected by the original investigator.
```

Implementing Reproducible Research

A bit more practical

```
The sharing of analytic data and computer codes
uses to map those data into computational results
is central to any comprehensive definition of
reproducibility.
```

Implementing Reproducible Research

Why its important?

```
Except for the simplest of analyses, the computer
code used to analyze a dataset is the only record
that permits others to fully understand what a
researcher has done.
```

Implementing Reproducible Research

Scrapping example

- Open Scrapping Ravens data from ESPN
- Is this document reproducible?
- What information is missing?
- How could we improve it?
- What about rMaps Mexico map?

R session information

```
sessionInfo()
```

```
## R version 3.2.0 Patched (2015-05-18 r68382)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
## Running under: OS X 10.8.5 (Mountain Lion)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] devtools_1.8.0 knitr_1.10.5   rmarkdown_0.7
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.11.6      digest_0.6.8     git2r_0.10.1     formatR_1.2
##  [5] magrittr_1.5     evaluate_0.7     stringi_0.4-1    curl_0.8
##  [9] rstudioapi_0.3.1 xml2_0.1.1       tools_3.2.0      stringr_1.0.0
## [13] yaml_2.1.13      rversions_1.0.1  memoise_0.2.1    htmltools_0.2.6
```

```
library('devtools'); session_info()
```

```
## Session info -------------------------------------------------------------
```

```
##  setting  value
##  version  R version 3.2.0 Patched (2015-05-18 r68382)
##  system   x86_64, darwin10.8.0
##  ui       AQUA
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York

## Packages ------------------------------------------------------------------

##  package    * version date       source
##  curl         0.8     2015-06-06 CRAN (R 3.2.0)
##  devtools   * 1.8.0   2015-05-09 CRAN (R 3.2.0)
##  digest       0.6.8   2014-12-31 CRAN (R 3.2.0)
##  evaluate     0.7     2015-04-21 CRAN (R 3.2.0)
##  formatR      1.2     2015-04-21 CRAN (R 3.2.0)
##  git2r        0.10.1  2015-05-07 CRAN (R 3.2.0)
##  htmltools    0.2.6   2014-09-08 CRAN (R 3.2.0)
##  knitr      * 1.10.5  2015-05-06 CRAN (R 3.2.0)
##  magrittr     1.5     2014-11-22 CRAN (R 3.2.0)
##  memoise      0.2.1   2014-04-22 CRAN (R 3.2.0)
##  Rcpp         0.11.6  2015-05-01 CRAN (R 3.2.0)
##  rmarkdown  * 0.7     2015-06-13 CRAN (R 3.2.0)
##  rstudioapi   0.3.1   2015-04-07 CRAN (R 3.2.0)
##  rversions    1.0.1   2015-06-06 CRAN (R 3.2.0)
##  stringi      0.4-1   2014-12-14 CRAN (R 3.2.0)
##  stringr      1.0.0   2015-04-30 CRAN (R 3.2.0)
##  xml2         0.1.1   2015-06-02 CRAN (R 3.2.0)
##  yaml         2.1.13  2014-06-12 CRAN (R 3.2.0)
```

Reproducible documents

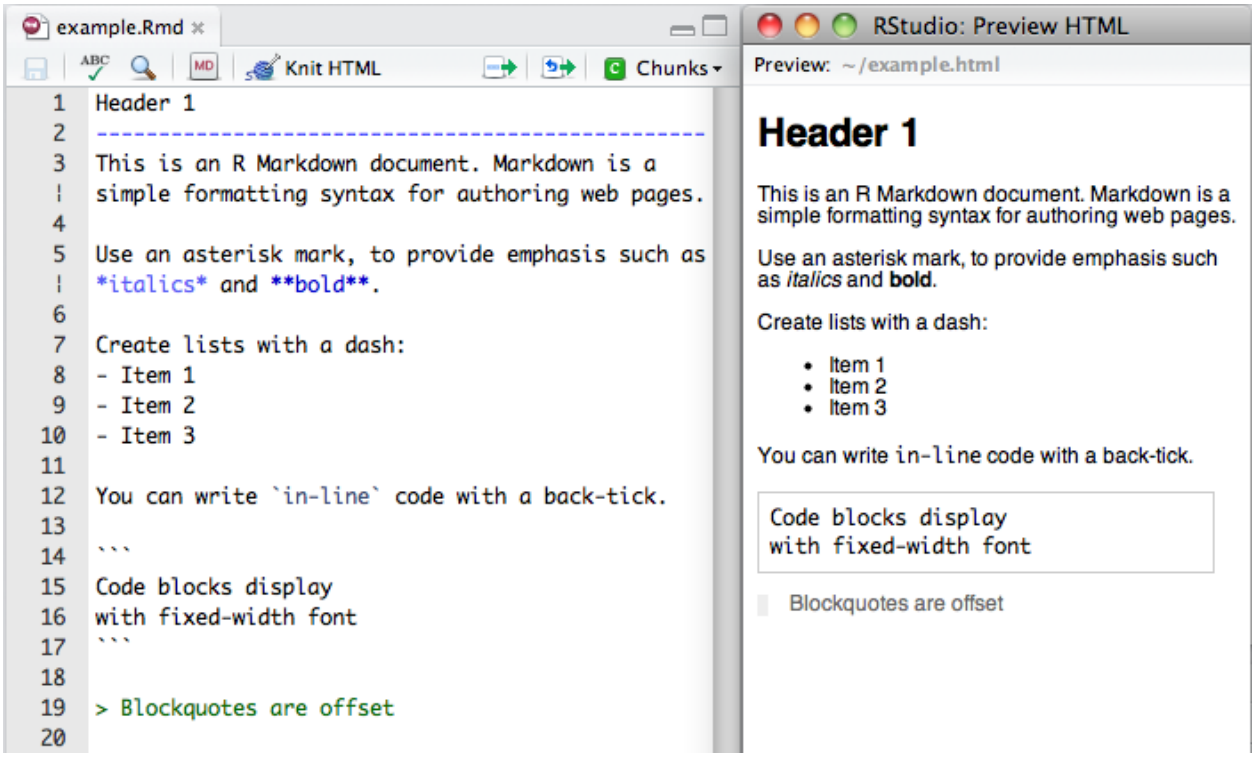
- Have you ever had your code in one file, your description of the results in another file?
- Ever made copy-paste mistakes?
- What if you were asked to change some models or revise the document?
- Was it easy to maintain?

- What would be a reproducible document for you?

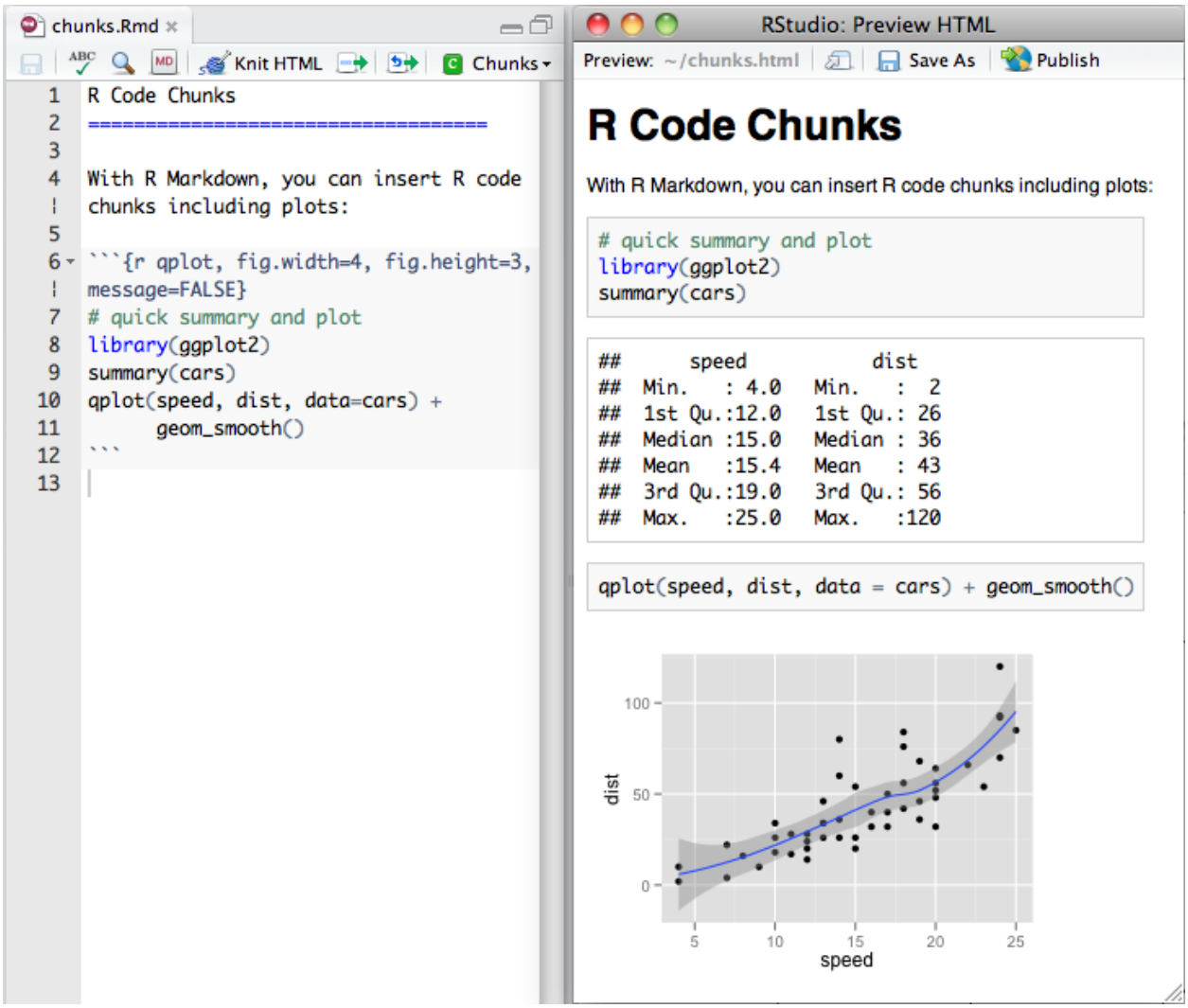
Reproducible docs in R

- R Markdown is the easiest
- It's based on Markdown: simple human readable syntax that generates HTML docs
- Also uses Pandoc to create files in many formats
- You maintain a **single** file! It has the code, figures and description of results. It then creates a file in the format you want to share with others.

Markdown


example.Rmd
Knit HTML
Chunks
1 Header 1
2 ----------
3 This is an R Markdown document. Markdown is a
4 simple formatting syntax for authoring web pages.
5
6 Use an asterisk mark, to provide emphasis such as
7 *italics* and **bold**.
8
9 Create lists with a dash:
10 - Item 1
11 - Item 2
12 - Item 3
13
14 You can write `in-line` code with a back-tick.
15
16 ```
17 Code blocks display
18 with fixed-width font
19 ```
20
21 > Blockquotes are offset
RStudio: Preview HTML
Preview: ~/example.html
Header 1
This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.
Use an asterisk mark, to provide emphasis such as italics and bold.
Create lists with a dash:
Item 1
Item 2
Item 3
You can write in-line code with a back-tick.
Code blocks display
with fixed-width font
Blockquotes are offset

Markdown + R code



Examples

- R Programming for Data Science (can buy for $0.00)
- Advanced R, for example Data structures
- Polyester: simulating RNA-seq datasets with differential transcript expression paper which creates this html.
- Timing information report which renders into this html. Note that it has an interactive table.

Misc

- R Markdown: where to go to learn more
- Example R Markdown doc
- An example interactive document: beyond the scope of this course, but cool to know it exists.
- Another intro to R Markdown presentation
- Where do I start using Bioconductor?