An introduction to R markdown for reproducible research

Stephen J Eglen
https://sje30.github.io
sje30@cam.ac.uk

Cambridge Computational Biology Institute
University of Cambridge
@StephenEglen

Slides: http://bit.ly/eglen2018-rmd (CC-BY license)

Acknowledgements

Laurent Gatto, Mike Smith.

The reproducibility crisis

Many key findings in publications are either not independently verified, or fail verification when it is attempted (Baker, 2016).

Duke oncogenomics scandal. Awesome detective work by Keith Baggerley and Kevin Coombes. https://www.youtube.com/watch?v=7gYIs7uYbMo

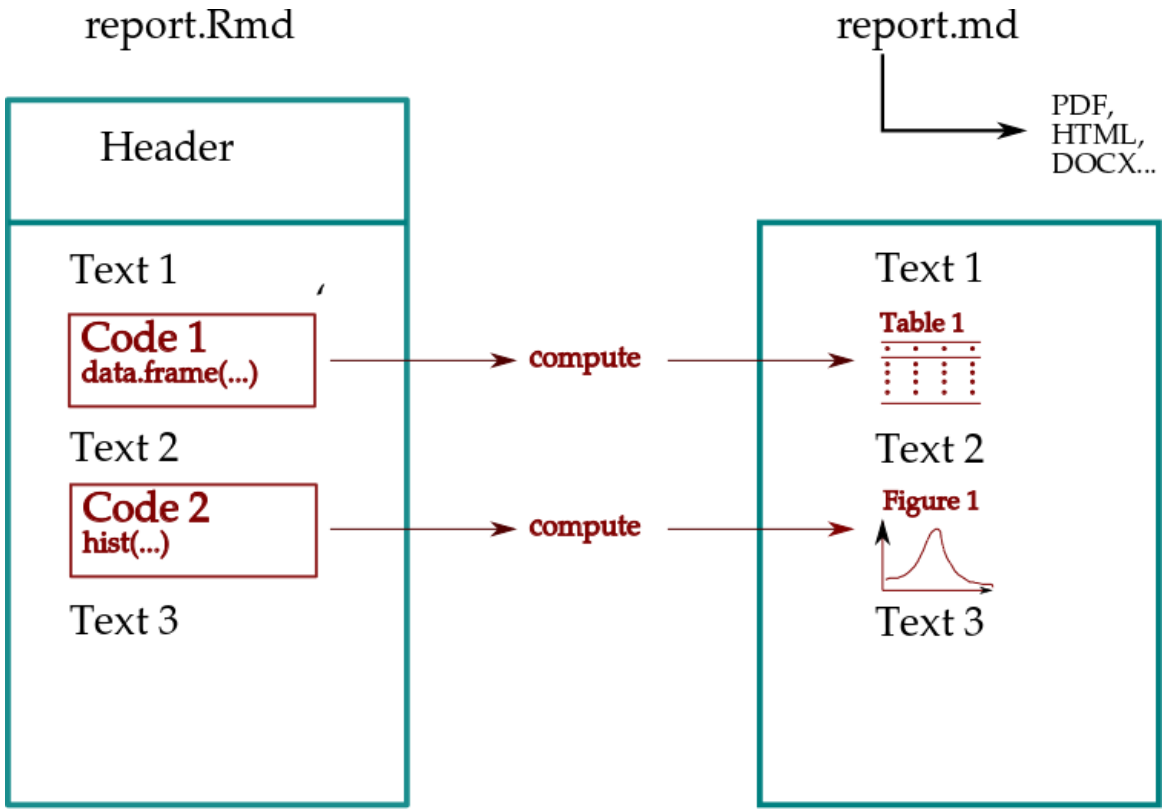
Disclaimer: do I mean "reproducibility" or "replicability"? (Barba 2018) https://arxiv.org/pdf/1802.03311.pdf

10.1038/533452a

Literate programming

Donald Knuth invented the literate programming environment to combine code and documentation into one file.

In our context, we interweave R code and markdown. R code is executed and results (text, figures, tables) fed back into markdown. Markdown then translated into html (or pdf or Word or slides...).



Moral or selfish approach?

Markowetz *Genome Biology* (2015) 16:274
DOI 10.1186/s13059-015-0850-7

Genome Biology

COMMENT Open Access

Five selfish reasons to work reproducibly

Florian Markowetz

Abstract

And so, my fellow scientists: ask not what you can do for reproducibility; ask what reproducibility can do for you! Here, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

Keywords: Reproducibility, Scientific career

how science actually is. And, whether you like it or not, science is all about more publications, more impact factor, more money and more career. More, more, more… so how does working reproducibly help me achieve more as a scientist.

Reproducibility: what's in it for me?

In this article, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

Selfish reasons to share

Why not align what is good for science with what is good for scientists?

1. Funding mandates (REF + enforcement from Wellcome Trust)
2. Credit through data papers
3. Fixes data bugs / errors in analysis
4. Prevent data loss (Vines et al 2014). e.g. students have a habit of leaving…
5. Your future self is probably one of the main beneficiaries of sharing.
6. *Now* is a very good time to be an open scientist.
7. Leads to further collaborations
8. Reviewers can do more work…

Simple example of reproducible research

Eglen SJ (2016) Bivariate spatial point patterns in the retina: a reproducible review. Journal de la Société Française de Statistique 157:33–48.

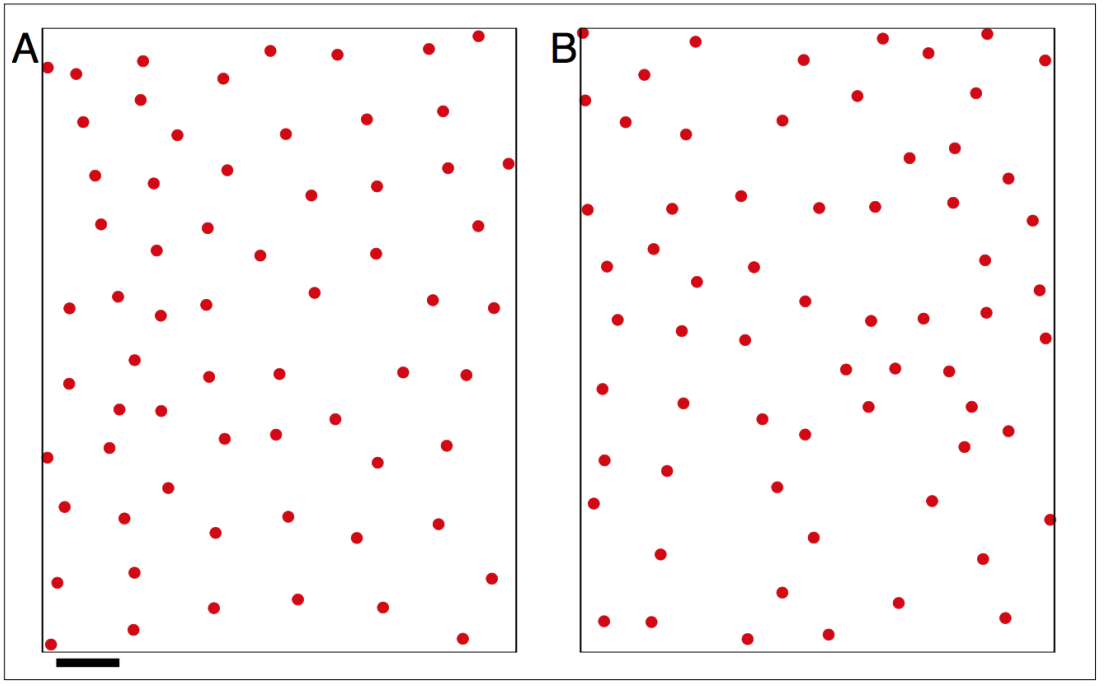


FIGURE 2. *An example retinal mosaic : beta on-centre retinal ganglion cells (Wässle et al., 1981). On the left is the observed map, and the right is an example univariate simulation with matching field and density of points. Scale bar is 100 μm ; soma are drawn to scale with a radius of 9 μm.*

See paper or code. Docker image.

See other examples at: https://rmarkdown.rstudio.com/gallery.html

What is markdown?

A light-weight markup language for generating HTML.

```
### Example

Here is some *markup text* with **bold** and a
[link](http://www.rstudio.org).
Maths can be included $x^2 + y^2 = z^2$.
```

Example

Here is some *markup text* with **bold** and a link. Maths can be included $x^2 + y^2 = z^2$.

Embedding code: inline

Take a chunk like:

Simple snippets of code can be embedded like ‘r sqrt(144)‘ is the square root of 144.

(NB: the backtick character above.)

which makes:

Simple snippets of code can be embedded like 12 is the square root of 144.

Embedding code: chunks

Or a chunk of code

````
```{r eval=TRUE}
n = 10
rnorm(n)
```
````

which makes this output:

```r
n = 10
rnorm(n)
```

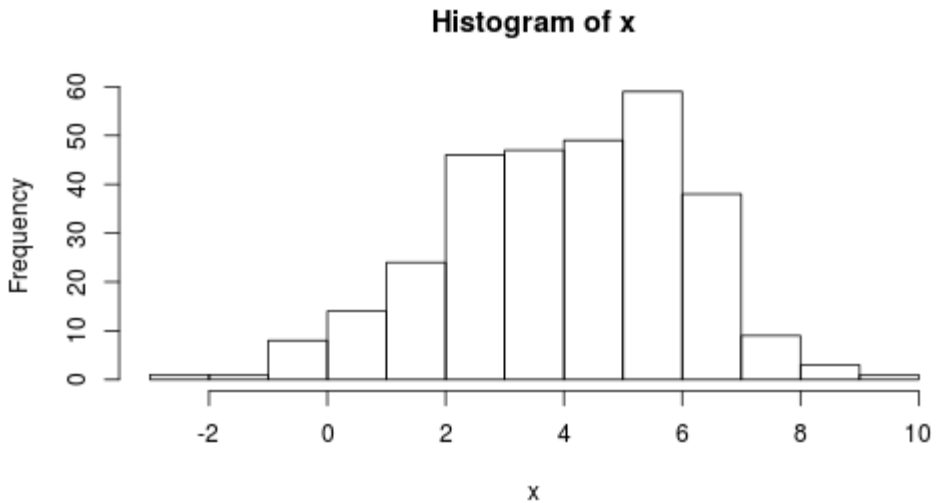
```
##  [1]  1.59394733 -1.31471275  0.19829420  0.78139398 -1.29532884
##  [6] -1.65744397  0.05817007 -0.03310738 -0.18284250  0.52807169
```

Figure outputs

If your code generates a figure, it is saved into `figures/` folder and included in output. (See bookdown for captions and references).

````
```{r eval=TRUE, echo=TRUE, fig.height=4}
x = rnorm(300, mean=4, sd=2)
hist(x)
```
````

```r
x = rnorm(300, mean=4, sd=2)
hist(x)
```



Next steps

- Material from today at: http://github.com/sje30/2018-12-07-rmd
- Jupyter as alternative format for notebook computation.
- Two-way Rmd <-> docx conversion: https://noamross.github.io/redoc/