Simple steps to improve reproducibility of your computational 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-rse (CC-BY license)

Acknowledgements

Paul Charlseworth, Ellese Cotterill, Catherine Cutts, Tom Edinburgh, BBSRC, EPSRC, Wellcome Trust, Software Sustainability Institute.

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

Inverse problems are hard

Mark	grade
70-100	A
60-69	B
50-59	C
40-49	D
0-39	F

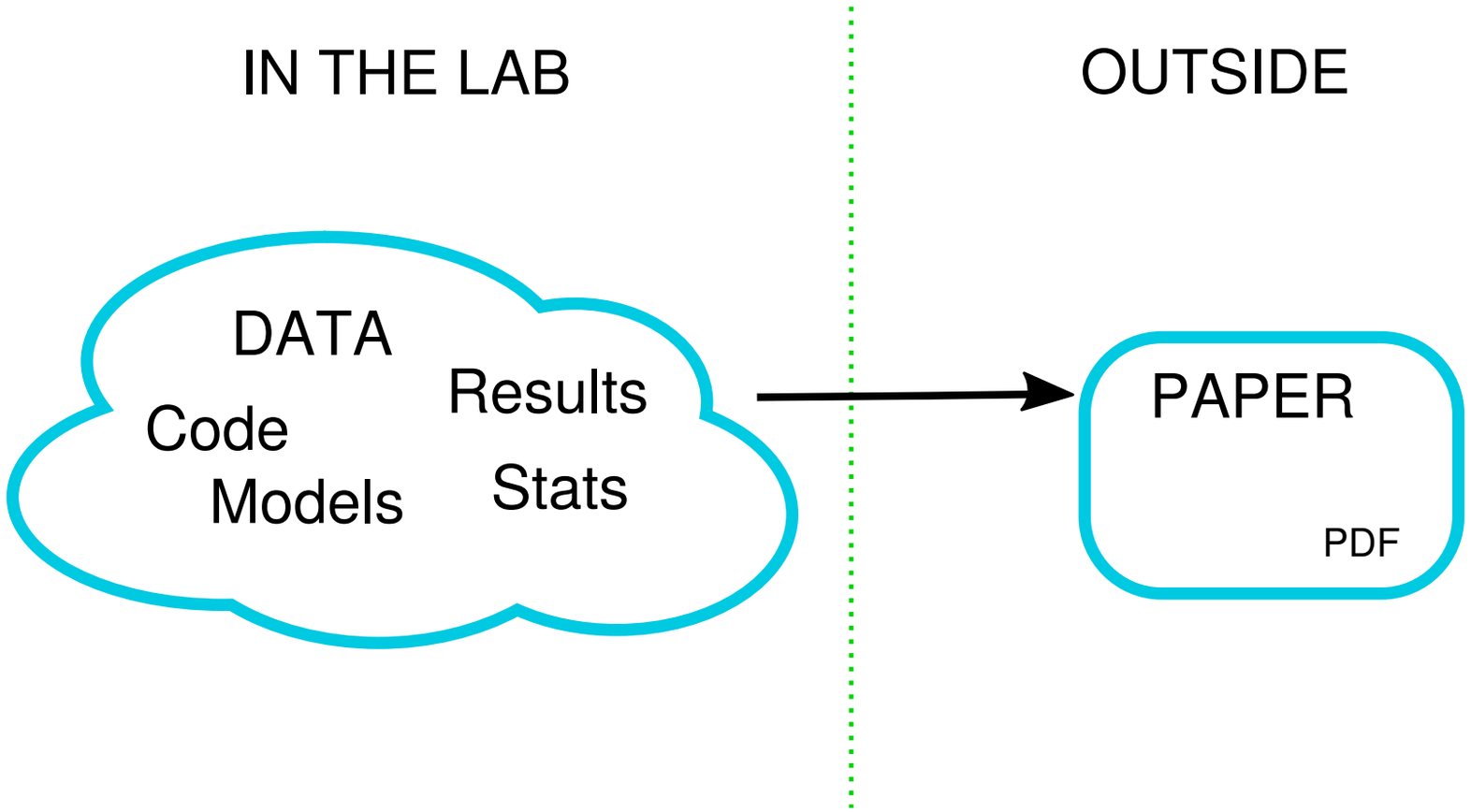
Forward problem

I scored 68, what was my grade?

Inverse problem

I got a B, what was my score?

Research sharing: the inverse problem



Where is the scholarship?

> An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures.

[Buckheit and Donoho 1995, after Claerbout]

10.1007/978-1-4612-2544-7_5

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?

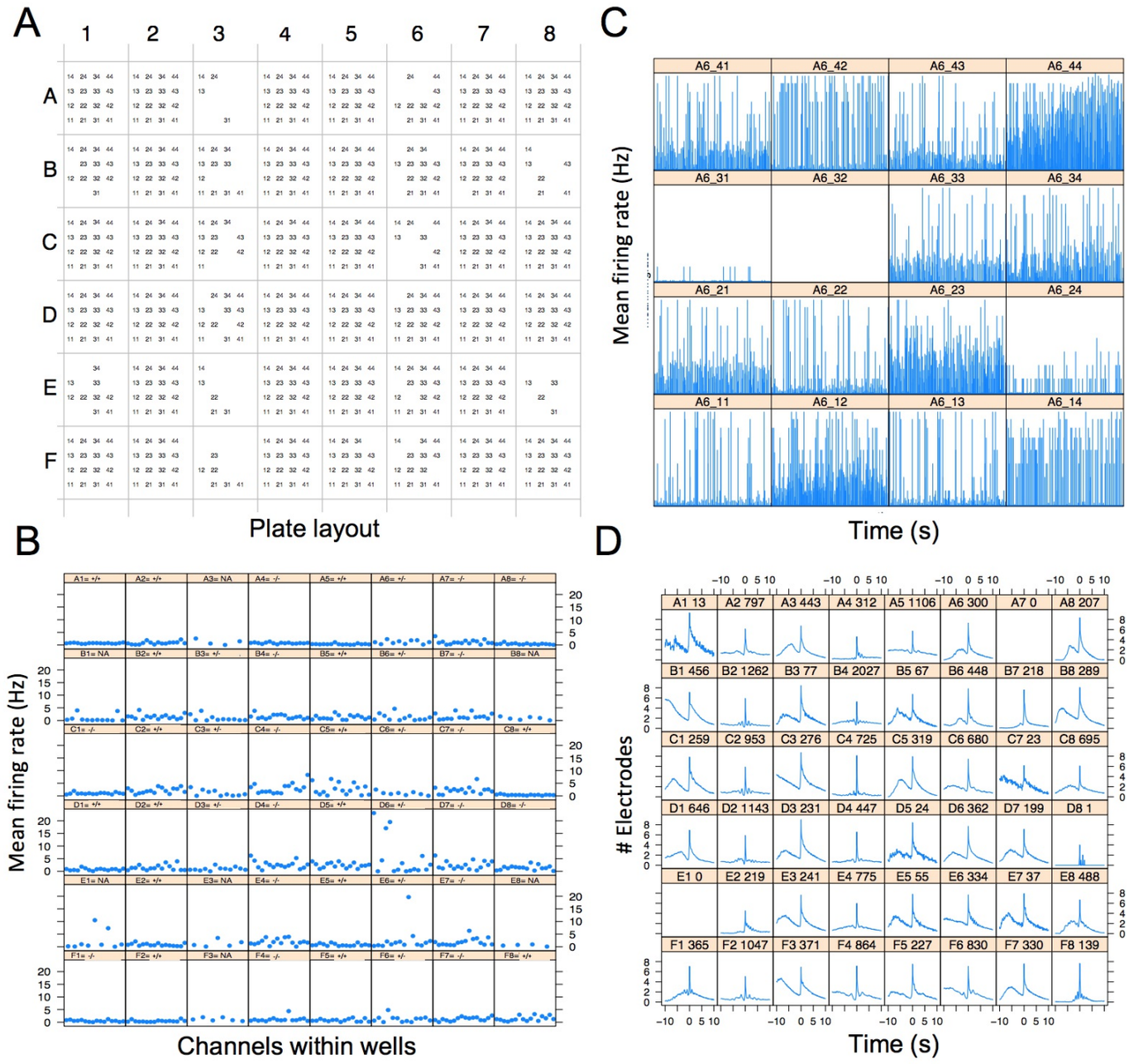
In this article, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

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. is a very good time to be an open scientist.
7. Leads to further collaborations
8. Reviewers can do more work…

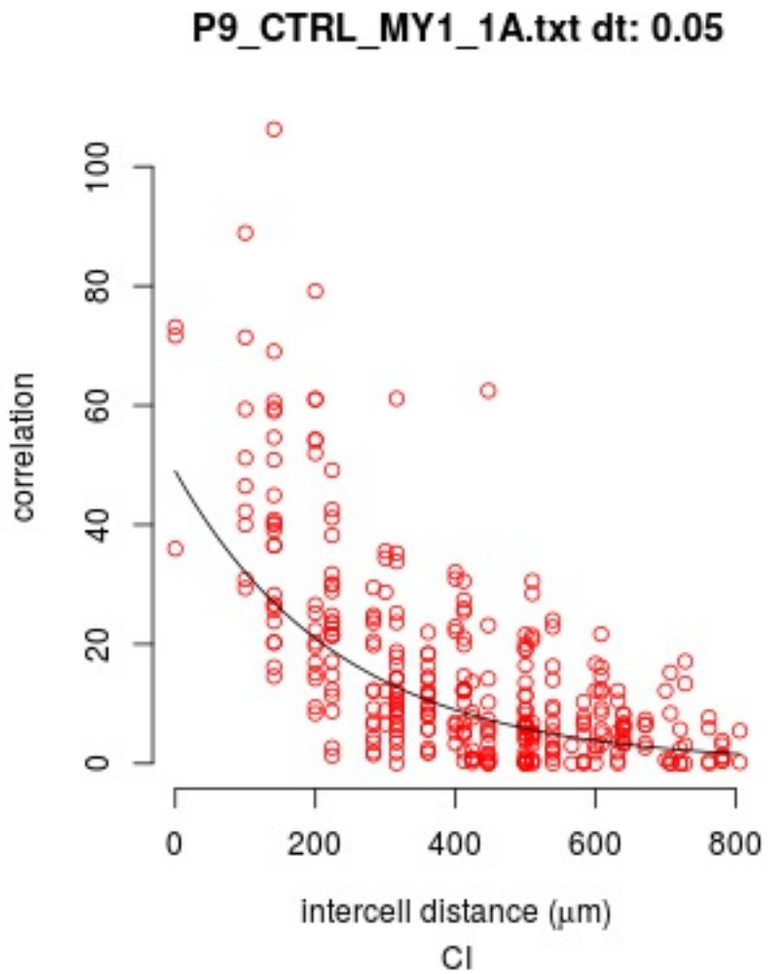
meaRtools: Tools for MEA analysis

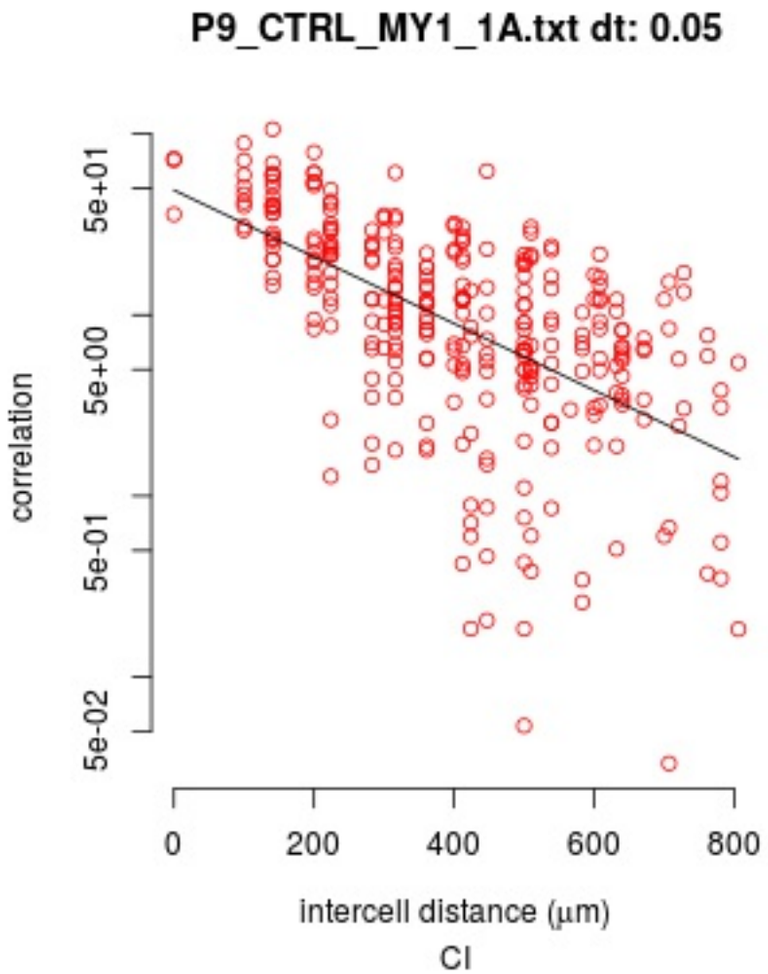


10.1101/242065

Reviewers doing your work

I would use an ordinate log scale for this bottom right panel (as done in Fig. 3). But since the authors gave me everything, I can do it! by redefining fourplot as follows:





Code review pilot

EDITORIAL

nature neuroscience

Extending transparency to code

Reproducibility initiatives seek to promote greater transparency and sharing of scientific reagents, procedures and data. Less recognized is the need to share data analysis routines. *Nature Neuroscience* is launching a pilot project to evaluate the efficacy of sharing code.

COMMENTARY

Toward standard practices for sharing computer code and programs in neuroscience

Stephen J Eglen[1], Ben Marwick[2], Yaroslav O Halchenko[3], Michael Hanke[4,5], Shoaib Sufi[6], Padraig Gleeson[7], R Angus Silver[7], Andrew P Davison[8], Linda Lanyon[9], Mathew Abrams[9], Thomas Wachtler[10], David J Willshaw[11], Christophe Pouzat[12] & Jean-Baptiste Poline[13]

Computational techniques are central in many areas of neuroscience and are relatively easy to share. This paper describes why computer programs underlying scientific publications should be shared and lists simple steps for sharing. Together with ongoing efforts in data sharing, this should aid reproducibility of research.

10.1038/nn.4550

Specific recommendations

1. Include enough code to reproduce key figure/result from your paper ("modeldb").
2. Provide toy examples if your project is too intensive to expect others to run in a few hours.
3. Version control (github)
4. Licence (MIT)
5. Provide data
6. Provide tests
7. Use standards
8. Use permanent URLs (Zenodo/figshare)

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.

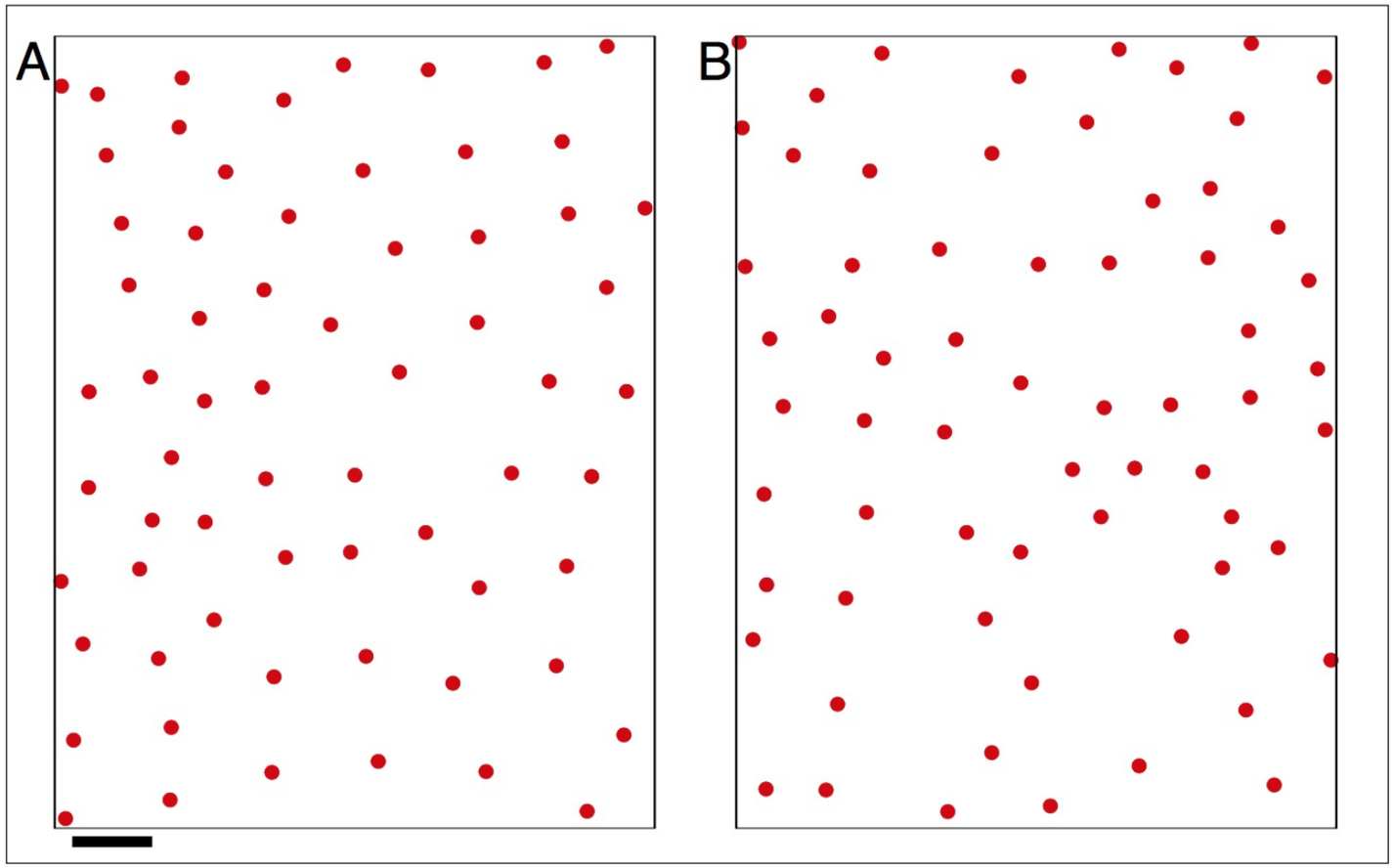


FIGURE 2. *An example retinal mosaic : beta on-centre retinal ganglion cells (Wässle et al., 1981). On the left is the observed map, and the right is an example univariate simulation with matching field and density of points. Scale bar is 100 μm ; soma are drawn to scale with a radius of 9 μm.*

See paper or code. Docker image.

New tools

1. **Docker** Can bundle entire open-source evironment for others to share.
2. Jupyter notebooks
3. binder = Docker + jupyter + cloud compute
4. Code ocean, and alternatives, being supported by some journal publishers, e.g. CUP.

Mesoscale two-photon imaging with the 2p-RAM

launch binder

Notebooks and data acquired with the two-photon random access mesoscope (2p-RAM), accompanying

A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging

Sofroniew, N. J. 1, *, Flickinger, D. 1, *, King, J. 2, Svoboda, K. 1

1 Janelia Research Campus, Ashburn VA 20147, USA 2 Vidrio Technologies, Ashburn VA 20147, USA

*These authors contributed equally to this work

https://github.com/sofroniewn/2pRAM-paper

Binder example for teaching

Two hour introduction to computational neuroscience (this Thursday if anyone is interested…)

Binder Github

Find a code buddy

- We ask our students to submit a .Rnw file rather than a pdf. You get a zero if I can't compile the pdf.
- So, ask someone else if they can run your code.
- Bioconductor team performs code review
- Journals gradually moving in this direction

Third most important file in github repo?

Credit: Arfon Smith

Third most important file in github repo?

- First: LICENSE

Credit: Arfon Smith

Third most important file in github repo?

- First: LICENSE
- Second: README.md

Credit: Arfon Smith

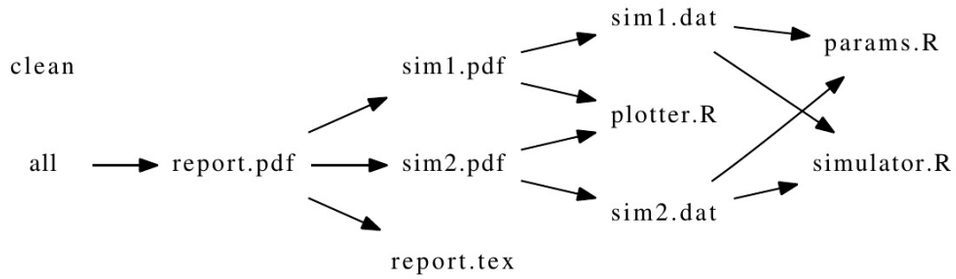
Third most important file in github repo?

- First: LICENSE
- Second: README.md
- Third: Makefile

Credit: Arfon Smith

Make

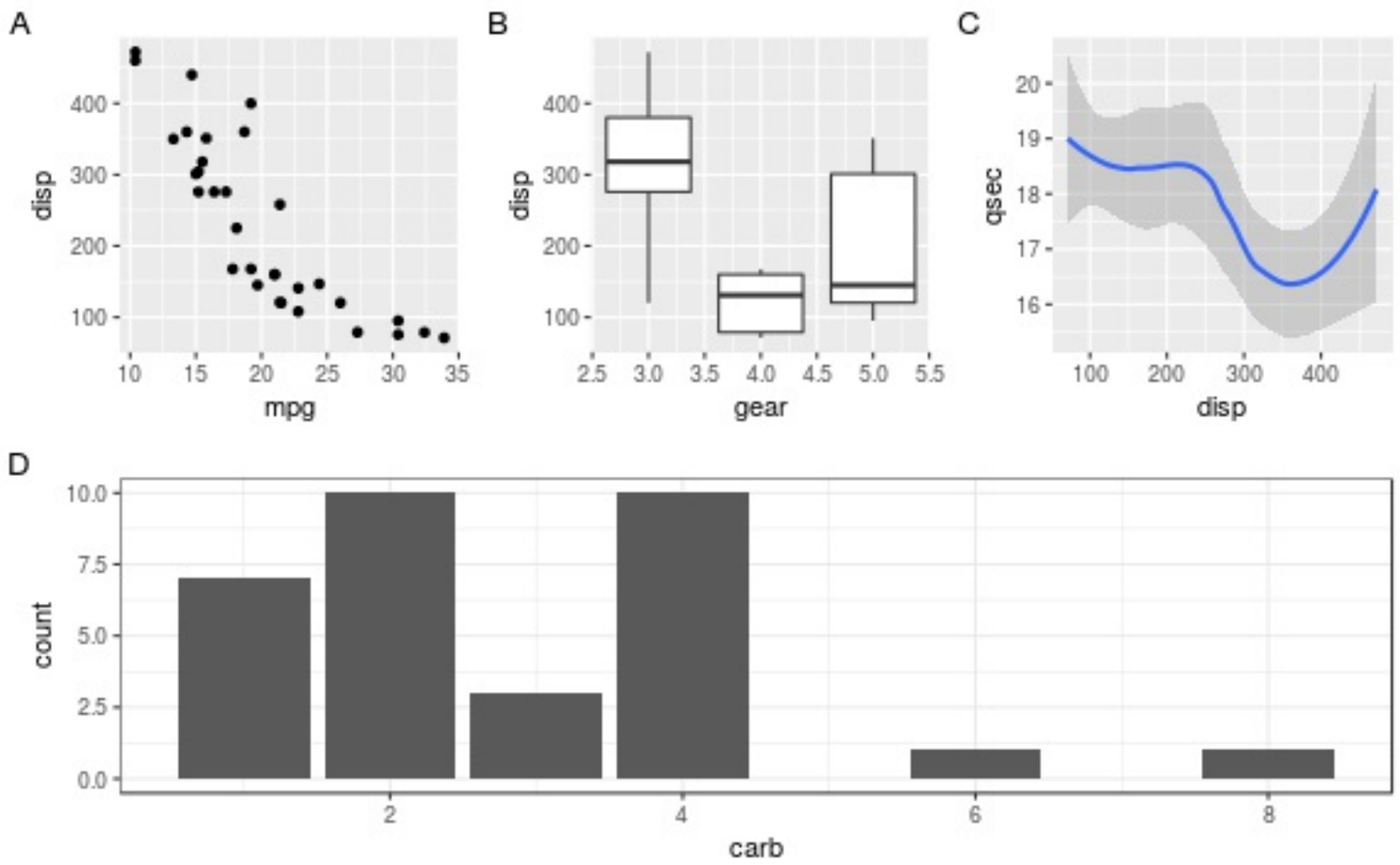
Make (or SnakeMake) are great at reducing cognitive load.




```
report.pdf: report.tex sim1.pdf sim2.pdf
    texi2pdf report.tex

sim1.dat: params.R simulator.R
    Rscript simulator.R rnorm > sim1.dat

sim2.dat: params.R simulator.R
    Rscript simulator.R runif > sim2.dat

sim1.pdf: sim1.dat plotter.R
    Rscript plotter.R sim1.dat

sim2.pdf: sim2.dat plotter.R
    Rscript plotter.R sim2.dat


.PHONY: all clean

all: report.pdf

clean:
    rm -f report.pdf report.log report.aux
    rm -f sim1.* sim2*
```

Reproducible figures

```
library(ggplot2); library(patchwork) # github.com/thomasp85/patchwork
p1 = ggplot(mtcars) + geom_point(aes(mpg, disp)) + labs(tag="A")
p2 = ggplot(mtcars) +
  geom_boxplot(aes(gear, disp, group = gear)) + labs(tag="B")
p3 = ggplot(mtcars) + geom_smooth(aes(disp, qsec)) + labs(tag="C")
p4 = ggplot(mtcars) + geom_bar(aes(carb)) + labs(tag="D")
((p1 | p2 | p3) /  p4 ) + theme_bw()
```



Summary

- Find the selfish reasons to make your research reproducible.
- Adopt good practices to help you on your way.
- Writing code in groups can be very motivating.
- Use new tech if you want, but old tech works too.

Challenges

- Long computation times (CODE CHECK).
- licensed software complicates everything.
- Can journals handle reproducible documents?
- When is the best time to think reproducibly?
  - Too early (explore first)
  - Too late (paper now out)?
- Technical challenges << Societal challenges