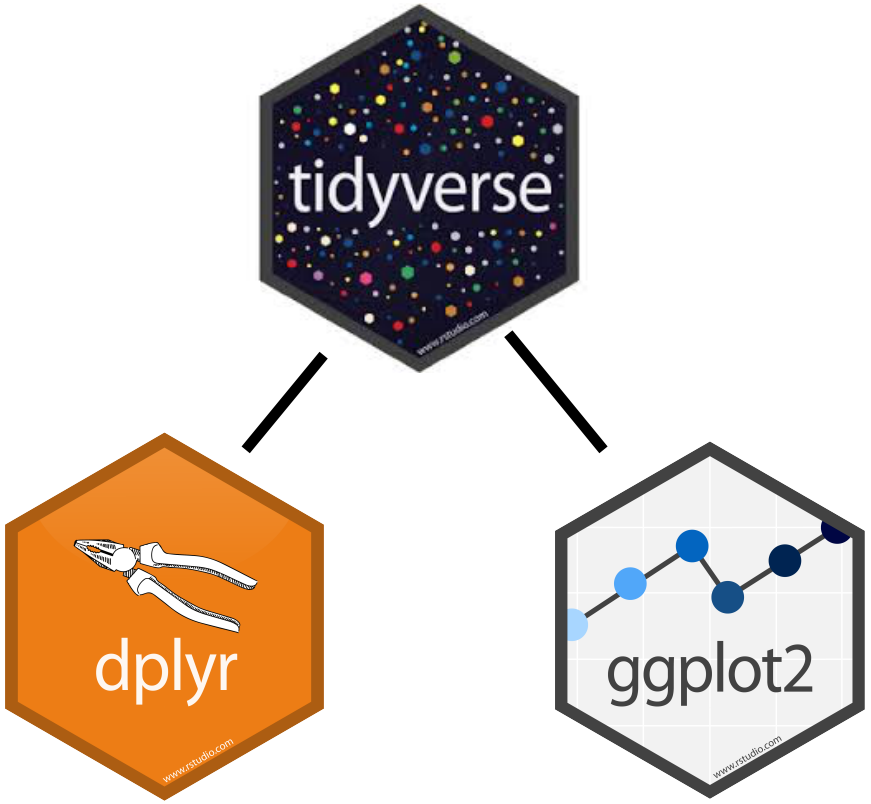
Reproducibility (and the failures)

10 February 2020

Modern Research Methods

Last couple weeks: Working with data in the tidyverse



filter(), select(), arrange(), mutate(), group_by(), summarize(), slice(), distinct()

ggplot()

- Load packages – library()
- Read in data – read_csv()
- Save plots – pdf()…dev.off()
- Knit file to html for sharing
- Use Rmarkdown to make reproducible reports

Literate Programming

Artwork by @allison_horst

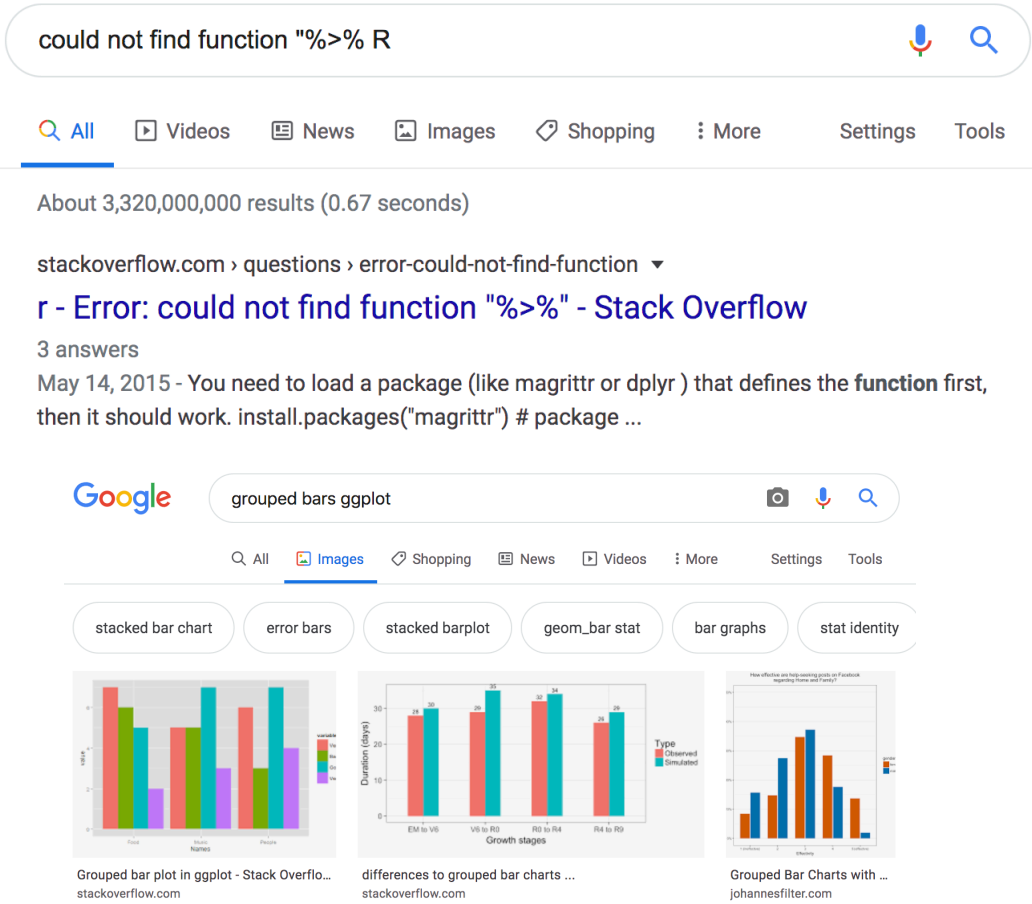
Debugging

Finding the root cause of a problem is challenging, and it's a **skill**

Strategies:

1. Google – whenever you see an error message you don't understand, start by googling it.
2. Make it repeatable. Make a reproducible example.
3. Figure out where the bug is, and then figure out how to fix it.

Experienced programmers do this allll the time.



(Wickham, 2019)

The R community is really wonderful!

Twitter: #rstats
Rladies
Rstudio

Artwork by @allison_horst

You now know the core of the tidyverse

- As the course progresses, we'll continue to learn new tools in the tidyverse
- We've just been using a few packages but there are over 15,000 packages available that do everything you possibly can imagine (Explore: https://www.r-pkg.org/)
- Resource for learning more on your own:
  - Tidyverse website: https://www.tidyverse.org/
  - R Studio Education: https://education.rstudio.com/learn/
  - R for Data Science: https://r4ds.had.co.nz/

Why are these tools useful for psychologists?

The Single Experiment

Population

Question

Hypothesis

Exp. Design

Experimenter

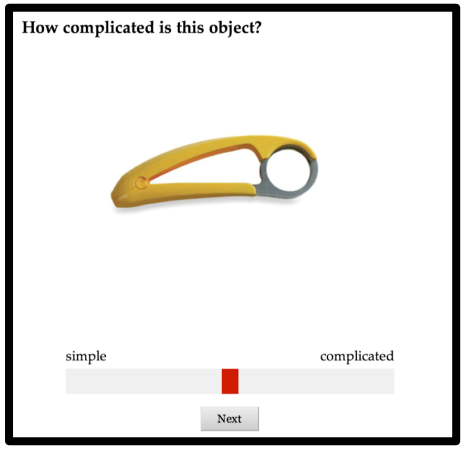
Data

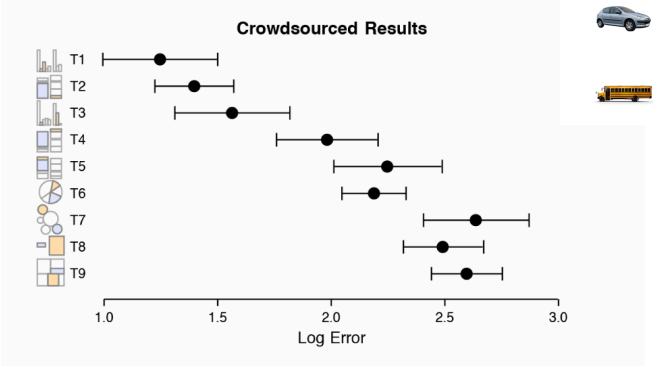
Analyst

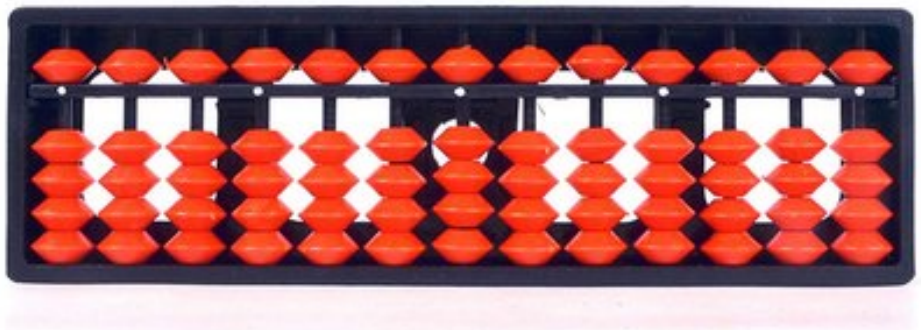
Code

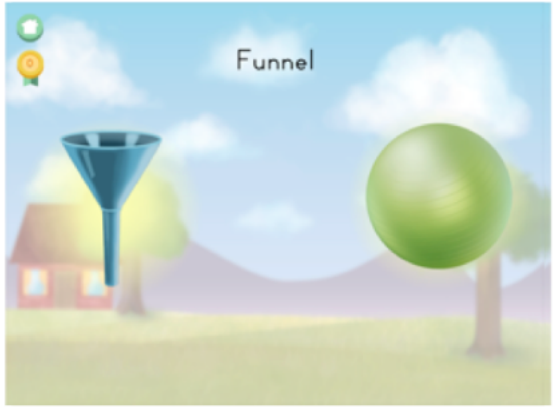
Estimate

Claim









Population

Question

Hypothesis

Exp. Design

Experimenter

Data

Analyst

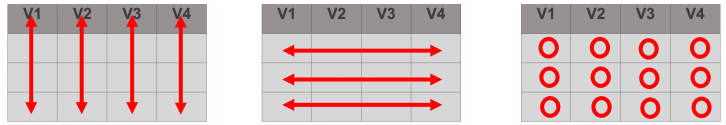
Code

Estimate

Claim

The tidyverse is a powerful and consistent ecosystem for analyzing, exploring, and presenting data.

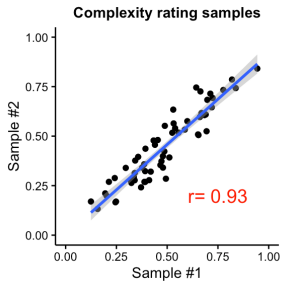
Tidy data



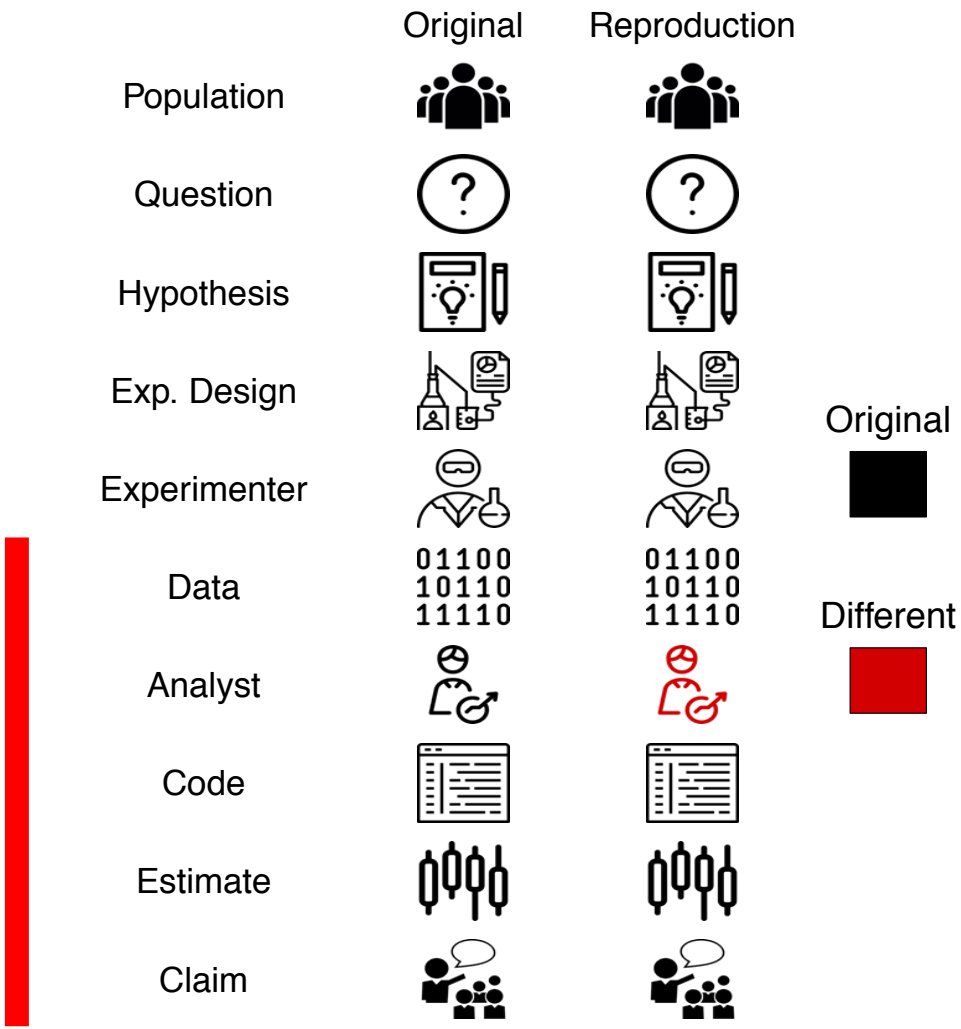
You, an R coder

R markdown, using dplyr verbs

ggplot



These tools also help make the analyses of your experiment **reproducible**



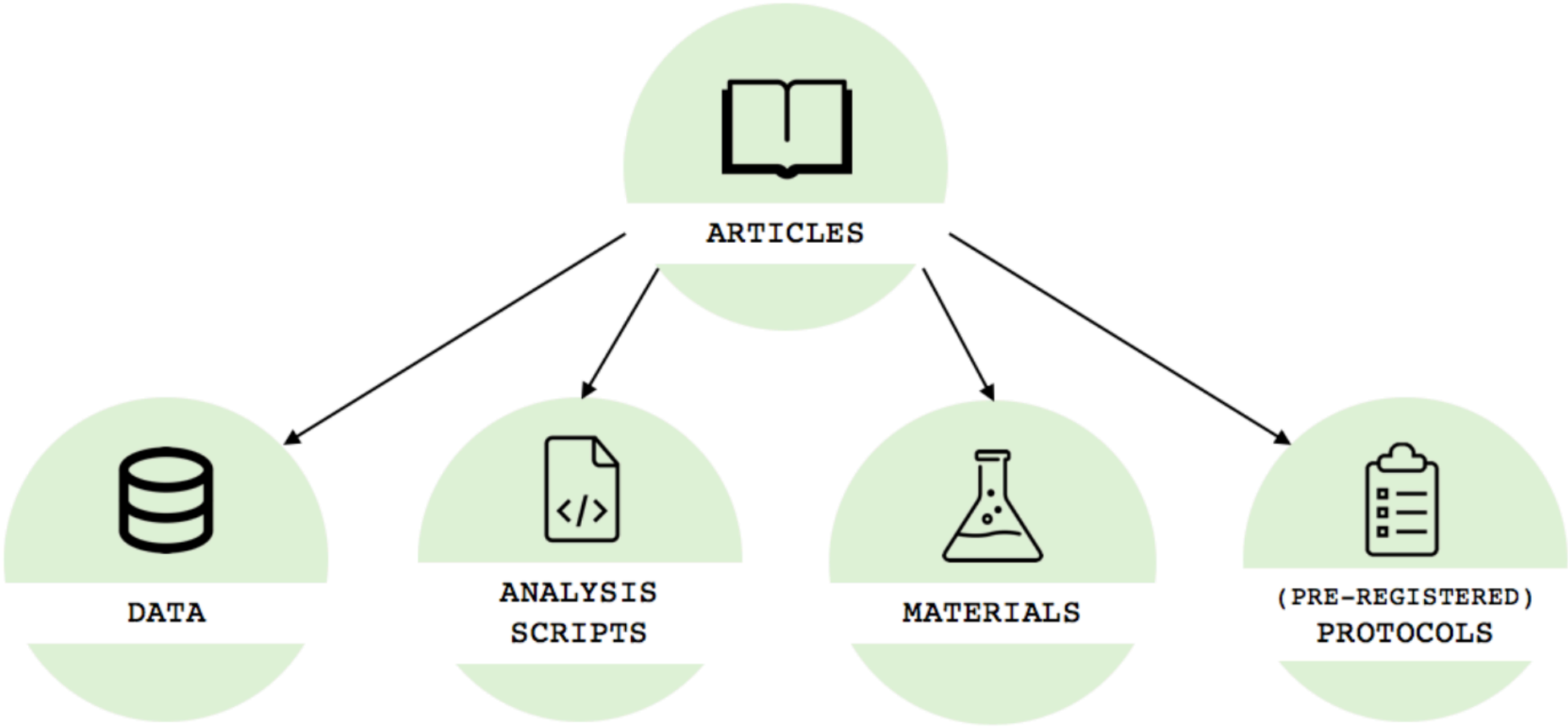
(Patil, Peng, & Leek, 2019)

REPRODUCE = Get same result from same dataset.

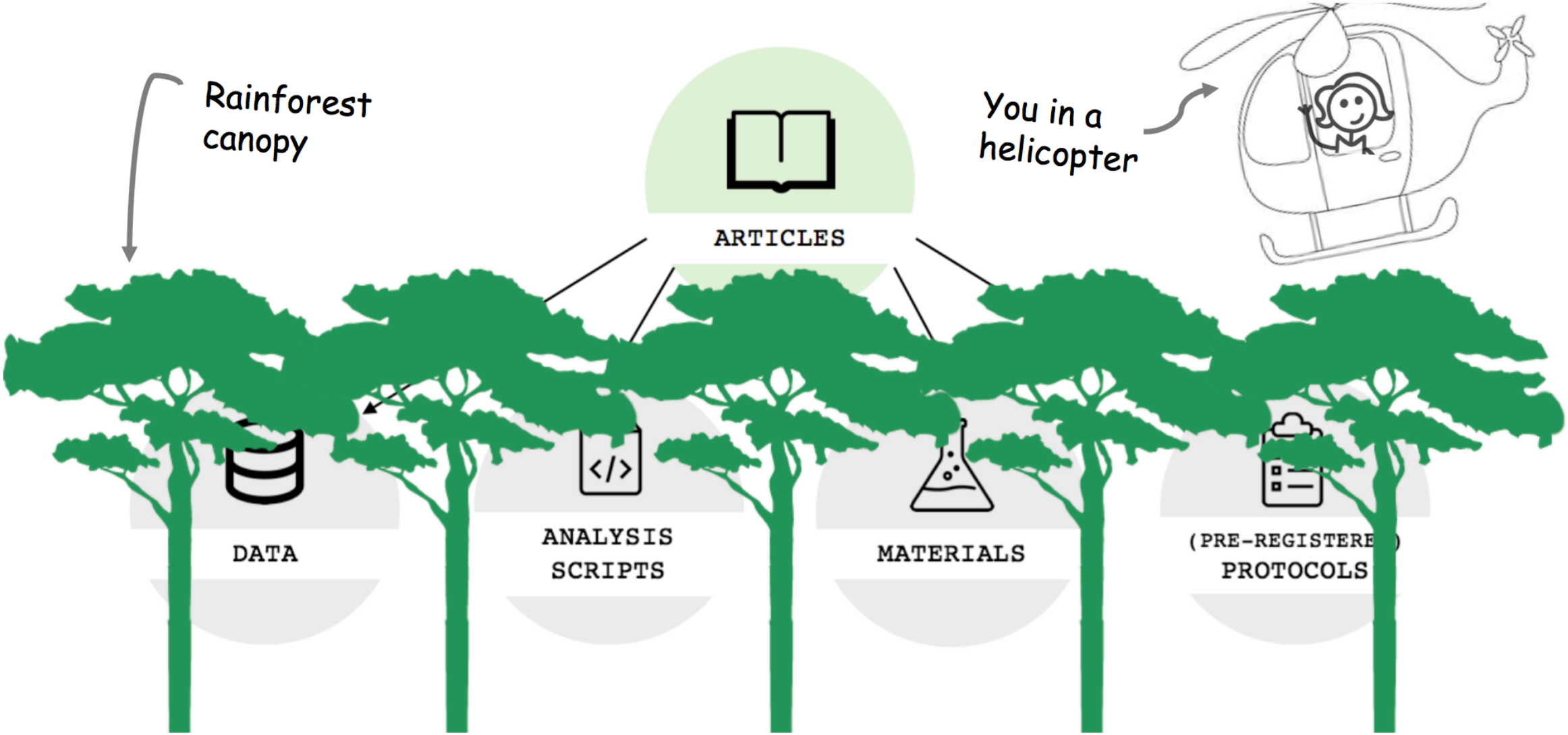
REPRODUCE = "...a second researcher might use the same raw data [and] implement the same statistical analysis in an attempt to yield the same results.... Reproducibility is a minimum necessary condition for a finding to be believable and informative." – NSF Report.

Why is reproducibility important?

The Modern Scholarly Record



The Modern Scholarly Record

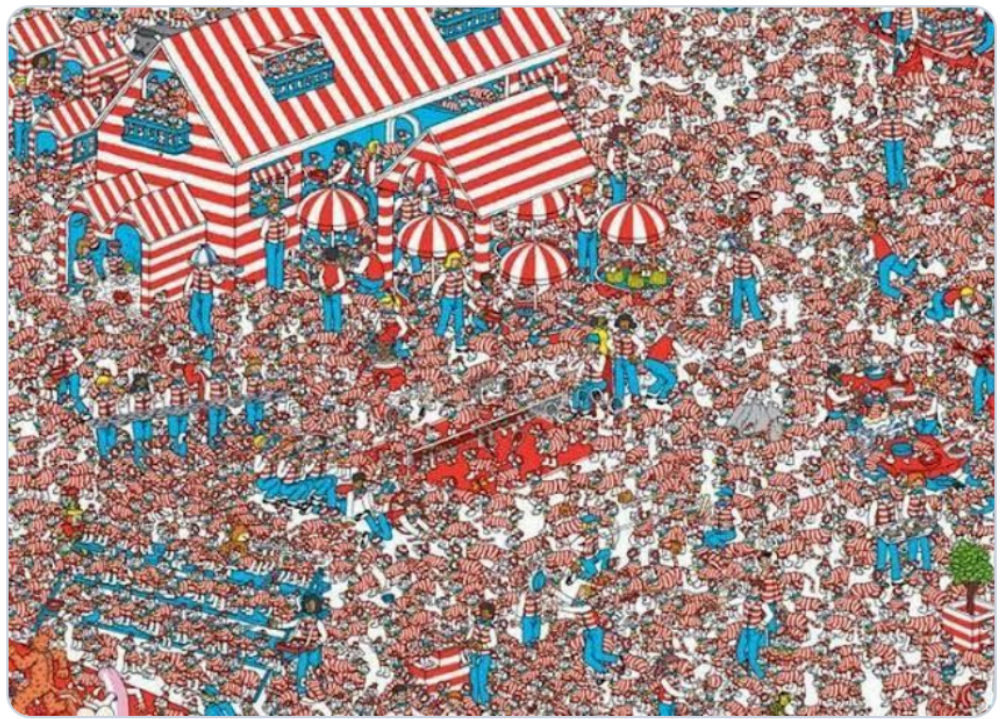


Guy Prochilo 🏳️‍🌈
@GuyProchilo

"Thank you for your interest in our paper. Please find attached the code for reproducing our findings"

The code

#phdchat #rstats



2:26 PM · Feb 5, 2020 · Buffer

5 Retweets **40** Likes

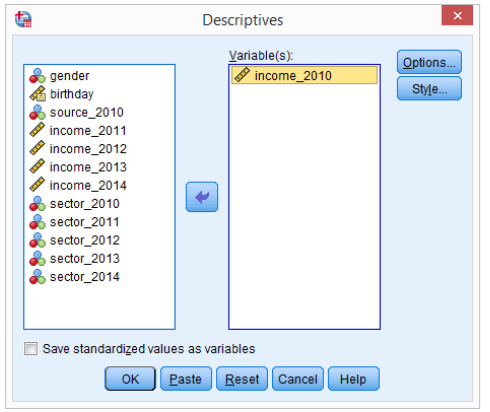
What makes an analysis irreproducible?

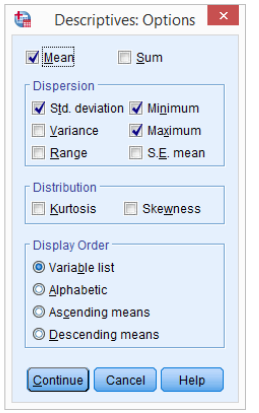
Either by later-you or another analyst.

- Original data is lost/not accessible
- Outdated/unavailable software
- Point and click software – hard to save steps
- Have data but don't know what variables correspond to
- Ambiguous verbal description of the analysis.

SUM =C5-

	A	B	C	D
24	**Credit**			
25	**Visa**	8/5/2008	$75.00	$0.00
26	**Mastercard**	8/5/2008	$37.42	$23.51
27	**Discover**	8/5/2008	$30.52	$30.00
28	**Store Credit Card**	8/5/2008	$87.56	$66.79
29	**Total**		$1,397.	
30	**Remaining**		=C5-	
31				





Results and Discussion

There was not a significant effect of sampling on generalization ($\chi^2(1) = 0.89$, $p = .34$; $d = 0.33$ [−0.22, 0.88]). Proportions and effect sizes are shown in Figures 3 and 4, respectively.

Open Science

Movement to make experimental materials available to others

- Stimuli
- Experimental data
- Experimental code

https://www.youtube.com/watch?v=1rFWeTryiW4&feature=youtu.be

How reproducible is psychological research?

Data not typically available from published papers in psychology

Study	Field	Papers checked	% data available*
Wicherts et al. (2006)	Psychology	141	27%
Vanpaemel et al. (2015)	Psychology	394	38%
Vines et al. (2014)	Ecology	516	19%
Hardwicke & Ioannidis (2018)	Psychology & Psychiatry	111	14%

Journals are creating policies that mandate data sharing

A mandatory open data policy was introduced at the journal Cognition on 1st March, 2015:

"All empirical papers must archive their data upon acceptance in order to be published unless the authors provide a compelling reason why they cannot."

"The data must be in a form that allows all reported statistical analyses to be reproduced while retaining the confidentiality of individual participants. This entails that the data are formatted and documented in a way that makes the structure of the data set readily apparent."

Data availability statement

Author note

This work was supported by an Economic and Social Research Council grant (ES/K004948/1) and a European Research Council consolidator grant (817492-SAMPLING) to ANS and an Economic and Social Research Council grants (ES/K002201/1 and ES/N018192/1) and a Leverhulme Trust grant (RP2012-V-022) grant to NS. The authors thank Jerome Busemeyer and Richard Shiffrin for helpful discussions. The data as well as analysis code from all experiments is available on the Open Science Framework: http://doi.org/10.17605/OSF.IO/8QS6J.

(Liao & Anderson, 2020)

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

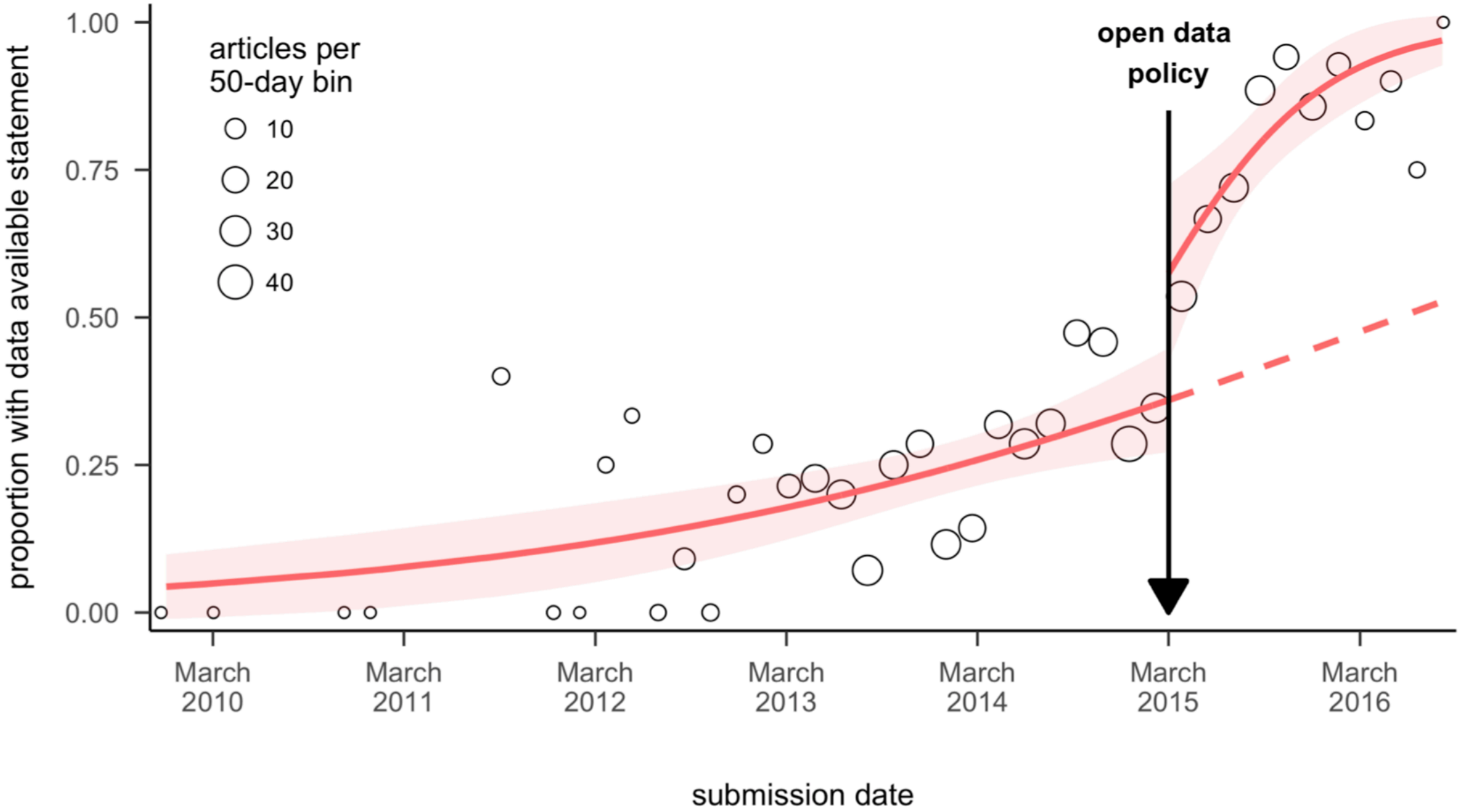
Cite this article: Hardwicke TE *et al*. 2018 Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R. Soc. open sci.* **5**: 180448.
http://dx.doi.org/10.1098/rsos.180448

Received: 19 March 2018
Accepted: 25 June 2018

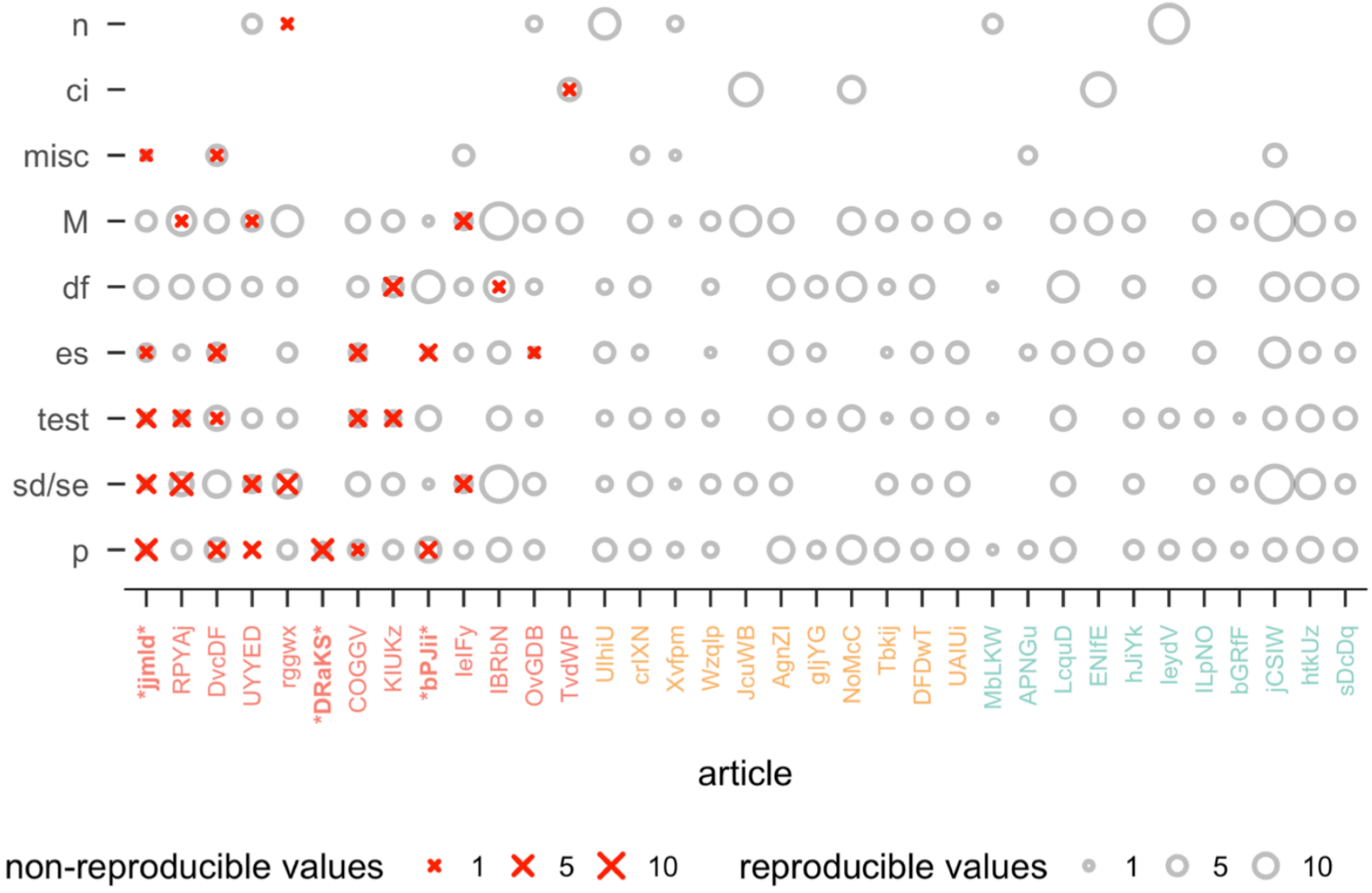
Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*

Tom E. Hardwicke[1], Maya B. Mathur[2,4], Kyle MacDonald[3], Gustav Nilsonne[3,5,6], George C. Banks[7], Mallory C. Kidwell[8], Alicia Hofelich Mohr[9], Elizabeth Clayton[10], Erica J. Yoon[3], Michael Henry Tessler[3], Richie L. Lenne[11], Sara Altman[3], Bria Long[3] and Michael C. Frank[3]

Is data available?



Are target values reproducible?

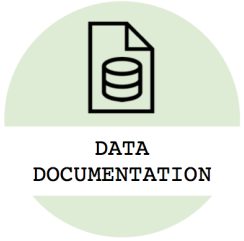


Assessment of reproducibility in psychology from Hardwicke et al. (2018)

- *Cognition*'s open data policy was highly effective at increasing data availability, but fell short of ideal
- Open data alone is clearly not enough to achieve the benefits envisioned by proponents of data sharing
- Reproducibility issues may be major barriers to data re-use but do not necessarily undermine substantive conclusions.
- Scientists are only human and inherit all the fallibilities that come with that. From this perspective, it is not surprising that analysis pipelines are peppered with errors.
- We should adopt strategies to reduce the likelihood of the errors that inevitably arise in computational work.

Open science and your future self

Reproducibility Solutions

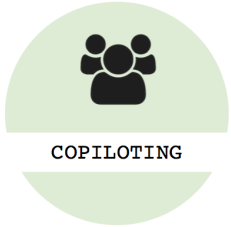


Data documentation

Create a 'codebook' which describes the structure and content of your data files. Consider organizing the data in 'Tidy' format.

Literate programming

Use R Markdown to combine your analysis code with regular prose. Using comments to explain your analysis helps others (and your future self) to understand what you did.



Co-piloting

Team up with the person sat next to you. Checking each other's work may help to reduce the chance of human error.

Dynamic report generation

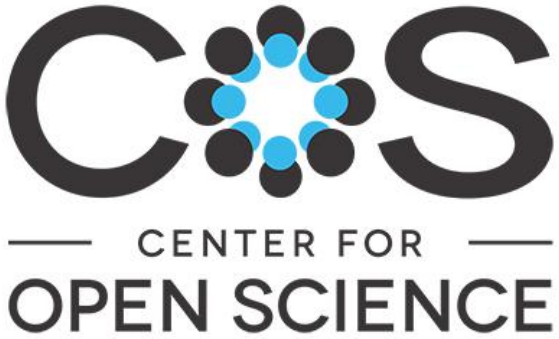
Use knitR to generate research reports directly from core research artifact (data, analysis scripts). A reader can now trace the provenance of reporte values to their source. Voilà! Reproducibility.

Version control

Use Github to keep track of changes to your project. This facilitates collaboration and helps with error detection.

Next Time: Tools for Reproducible Workflows

Version Control



Excuse Me, Do You Have a Moment to Talk About Version Control?

Jennifer Bryan

RStudio and the Department of Statistics, University of British Columbia, Vancouver, Canada

ABSTRACT

Data analysis, statistical research, and teaching statistics have at least one thing in common: these activities all produce many files! There are data files, source code, figures, tables, prepared reports, and much more. Most of these files evolve over the course of a project and often need to be shared with others, for reading or edits, as a project unfolds. Without explicit and structured management, project organization can easily descend into chaos, taking time away from the primary work and reducing the quality of the final product. This unhappy result can be avoided by repurposing tools and workflows from the software development world, namely, distributed version control. This article describes the use of the version control system Git and the hosting site GitHub for statistical and data scientific workflows. Special attention is given to projects that use the statistical language R and, optionally, R Markdown documents. Supplementary materials include an annotated set of links to step-by-step tutorials, real world examples, and other useful learning resources. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2017
Revised October 2017

KEYWORDS
Data science; Git; GitHub; R language, R Markdown; Reproducibility; Workflow

1. Why Git?

Why would a statistician use a version control system, such as *https://git-scm.com* Git (*Git* n.d.)? And what is the point of hosting your work online, for example, on *https://github.com* GitHub (*GitHub* n.d.)? Could the gains possibly justify the inevitable pain?

I say yes, with the zeal of the converted.

There are many benefits of using hosted version control in your statistical practice:

- Doing your work becomes tightly integrated with organizing, recording, and disseminating it. It is not a separate, burdensome task you are tempted to neglect.
- Collaboration is much more structured, with powerful tools for asynchronous work and managing versions.
- The marginal effort required to create a web presence for a project is negligible.

out tools that soften Git's sharpest edges, I recommend specific habits and attitudes that reduce frustration.

2. What is Git?

Git is a *version control system*. Its original purpose was to help groups of developers work collaboratively on big software projects. Git manages the evolution of a set of files—called a *repository* or *repo*—in a sane, highly structured way. It is like the "Track Changes" feature from Microsoft Word, but more rigorous, powerful, and scaled up to multiple files.

Git has been repurposed by the data science community (Ram 2013; Bartlett 2016; Perez-Riverol et al. 2016). We use it to manage the motley collection of files that make up typical data analytical projects, which consist of data, figures, reports, and source code. Even those who identify more as statisticians than data scientists generally have a similar mix of files that are the

Acknowledgements

Slides 9-10, 15-16, 19-23 adopted from Tom Hardwicke by CC