

Reproducible Workflows for Better Science and Efficient Collaboration

Francisco Rodriguez-Sanchez
@frod_san

<https://frodriguezsanchez.net>

The Reproducibility Crisis Revolution

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

- Goal: Replicate 193 experiments from 53 papers

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

- Goal: Replicate 193 experiments from 53 papers
- Finally: 50 experiments from 23 papers

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

- Goal: Replicate 193 experiments from 53 papers
- Finally: 50 experiments from 23 papers
- ~Half didn't replicate (much smaller effect sizes)

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

- Goal: Replicate 193 experiments from 53 papers
- Finally: 50 experiments from 23 papers
- ~Half didn't replicate (much smaller effect sizes)
- No paper reported all required data

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

- Goal: Replicate 193 experiments from 53 papers
- Finally: 50 experiments from 23 papers
- ~Half didn't replicate (much smaller effect sizes)
- No paper reported all required data
- Impossible to repeat experiments w/o contacting authors

NEWS | 09 December 2021

Half of top cancer studies fail high-profile reproducibility effort

- Goal: Replicate 193 experiments from 53 papers
- Finally: 50 experiments from 23 papers
- ~Half didn't replicate (much smaller effect sizes)
- No paper reported all required data
- Impossible to repeat experiments w/o contacting authors
- 1/3 authors didn't respond or help



Sylvain Deville
@DevilleSy

...

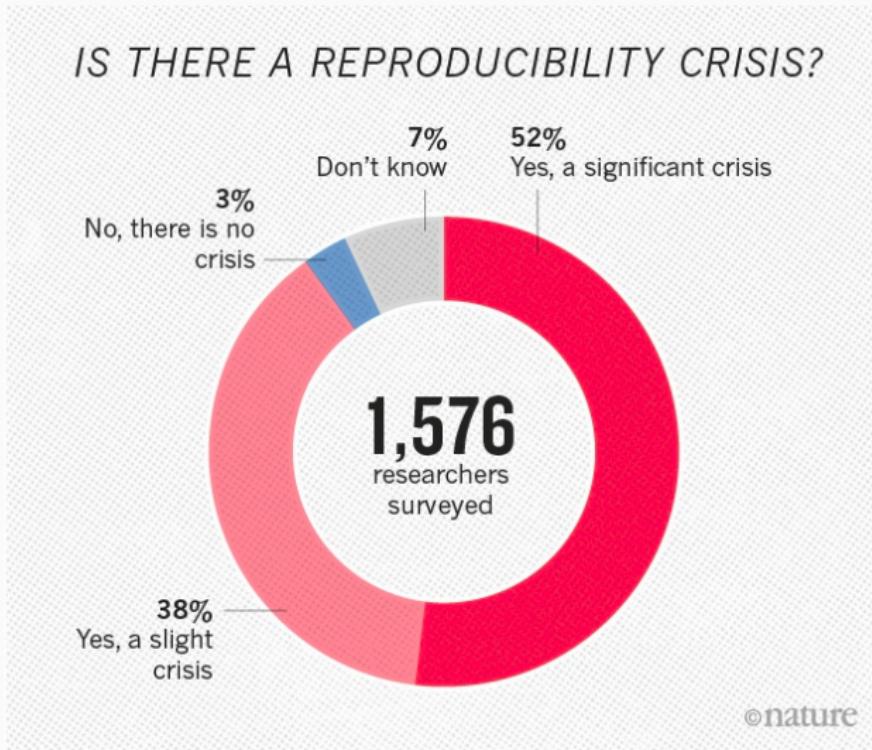
Trying to reproduce the results of a paper using only what's in the Methods section



Most scientific articles

are NOT reproducible

The reproducibility crisis



Reproducibility

CRISIS

REVOLUTION

Reproducibility vs Replicability

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

The Turing Way

We can't guarantee that
our results are **replicable**.

But at least
they should be **reproducible**.

Most scientific articles

are NOT reproducible

The prevalence of statistical reporting errors in psychology (1985–2013)

Michèle B. Nuijten¹ · Chris H. J. Hartgerink¹ · Marcel A. L. M. van Assen¹ ·
Sacha Epskamp² · Jelte M. Wicherts¹

WHAT STATCHCHECK LOOKS FOR

This computer algorithm scans papers for statistical tests, uses reported results to recompute the P value and flags up inconsistencies.

Type of test

The t -test assesses differences between two groups.

Test statistic

Compares observed values with those expected under the null hypothesis.

$$t(37) = 4.93, P < 0.01$$

Degrees of freedom

Accounts for size of sample.

P value

The likelihood of observing differences as extreme, or more so, if the null hypothesis is true.

The prevalence of statistical reporting errors in psychology (1985–2013)

Michèle B. Nijsten¹ · Chris H. J. Hartgerink¹ · Marcel A. L. M. van Assen¹ ·
Sacha Epskamp² · Jelte M. Wicherts¹

1/2 articles: inconsistencies in p-values

1/8 articles: grossly inconsistent p-values

(affecting conclusions -> significance)

In ecology

< 20% articles are reproducible

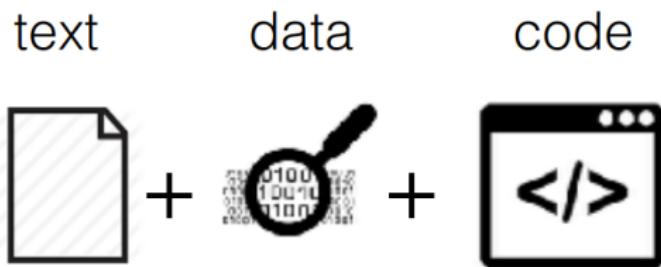
[Culina et al 2020](#)

We can't even reproduce our own work

Data/Code lost or unusable

qualitative_data.csv	04/07/2016 15:50
cleandata.xlsx	25/06/2015 01:14
cleandata_YC.xlsx	30/06/2015 16:22
COORDENADAS PACO_20-05-2016 CON REVIEWS.xlsx	20/05/2016 16:23
COORDENADAS PACO_20-05-2016 CON REVIEWS_FRS.xlsx	27/05/2016 19:41
COORDENADAS_paper195(Girella_elevata).xlsx	08/06/2016 13:09
coordenadas_raw_2016-06-08.xlsx	09/06/2016 15:53
coordenadas_raw_2016-06-08_old.xlsx	08/06/2016 16:00
coordenadas_raw_2016-06-21.xlsx	21/06/2016 16:12
coords_2015-09-09_modif.xlsx	05/11/2015 15:23
coords_2015-10-11_modif_YC.xlsx	17/11/2015 13:37
coords_2015-10-11_modif_YC_PACO.xlsx	17/11/2015 17:06
coords_2015-10-18_modif_YC.xlsx	18/11/2015 17:24
coords_2015-12-26_modif_YC.xlsx	30/03/2016 19:38
coords_2016-04-02.xlsx	06/04/2016 17:46
coords_2016-04-02_YC.xlsx	06/04/2016 18:03
coords_2016-04-08_YC.xlsx	11/04/2016 13:51
dataset_Y_coords_09_09_15.xlsx	23/09/2015 17:18
Datos metaanalisis_18-04-2016.xlsx	19/04/2016 16:24
FINAL METAANALYSIS_14-6-2016_WITH REVIEWS.xlsx	21/06/2016 16:15
FINAL METAANALYSIS_16-6-2016_WITH REVIEWS.xlsx	21/06/2016 16:13
FINAL METAANALYSIS_2016-04-27_WITH REVIEWS.xlsx	25/05/2016 18:05
FINAL METAANALYSIS_2016-04-27_WITH REVIEWS_FRS.xlsx	27/05/2016 18:44
FINAL METAANALYSIS_2016-04-29_EXCLUDING REVIEWS.xlsx	08/06/2016 13:06
FINAL VOTECOUNTING_1-7-2016.xlsx	04/07/2016 15:46
fitnessdata_2016-06-22.xlsx	22/06/2016 21:00
ifs for Bastien_19-3-2016_YC.xlsx	28/03/2016 19:26
Metaanalysis final_01-05-2015 with coordinates.xlsx	18/05/2015 19:20
Metaanalysis final_22-05-2015 coords.xlsx	24/06/2015 15:50
Metaanalysis final_25-06-2015.xlsx	30/06/2015 16:55
Metaanalysis y coords revisadas_06-08-2015_AH_E.xlsx	23/09/2015 12:57

What's a reproducible article?



A scientific article is reproducible if there is **computer code** that can regenerate all results and figures from the **original data**

A scientific article is **advertising**, not scholarship.

The actual scholarship is the **full software environment, code and data**, that produced the result.

Claerbout & Karrenbach 1992

Are we sharing the data?

PERSPECTIVE

Public Data Archiving in Ecology and Evolution: How Well Are We Doing?

Dominique G. Roche^{1,2*}, Loeske E. B. Kruuk^{1,3}, Robert Lanfear^{1,4}, Sandra A. Binning^{1,2}

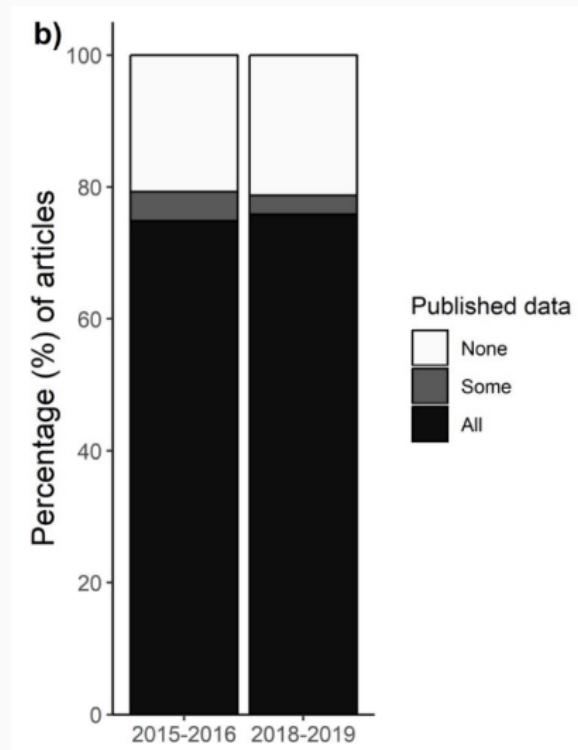
1 Division of Evolution, Ecology and Genetics, Research School of Biology, The Australian National University, Canberra, Australian Capital Territory, Australia, **2** Éco-Éthologie, Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland, **3** Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom, **4** Department of Biological Sciences, Macquarie University, Sydney, Australia

* dominique.roche@mail.mcgill.ca

Abstract

Policies that mandate public data archiving (PDA) successfully increase accessibility to data underlying scientific publications. However, is the data quality sufficient to allow reuse and reanalysis? We surveyed 100 datasets associated with nonmolecular studies in journals that commonly publish ecological and evolutionary research and have a strong PDA policy. Out of these datasets, 56% were incomplete, and 64% were archived in a way that partially or entirely prevented reuse. We suggest that cultural shifts facilitating clearer benefits to authors are necessary to achieve high-quality PDA and highlight key guidelines to help authors increase their data's reuse potential and compliance with journal data policies.

Are we sharing data?



Are we sharing data?

Quickly getting better

Scientific Life

Early Career
Researchers Embrace
Data Sharing

Hamish A. Campbell,^{1,*}
Mariana A. Micheli-Campbell,¹
and Vinay Udyawer²

[Campbell et al. 2019](#)

Are we sharing the code?

Code exists but rarely shared

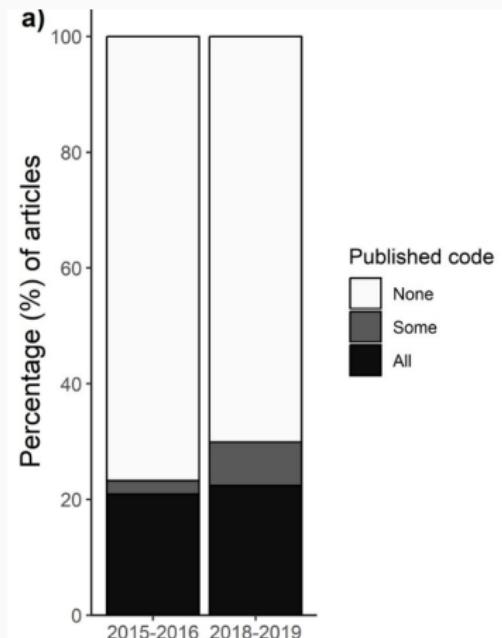
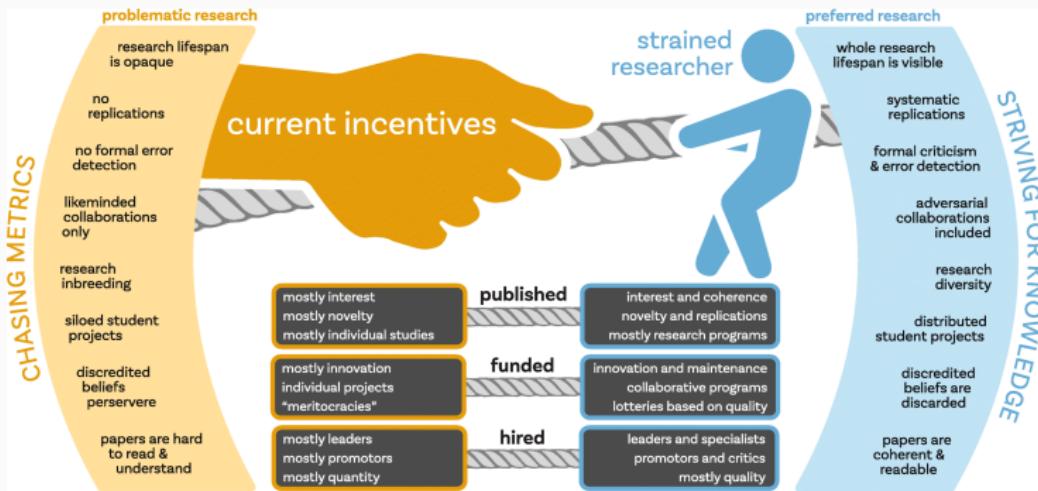


Fig 1. Code-sharing is at its infancy in ecology, where

WHY?

Poor incentives



O'Dea et al 2021

Doing reproducible research can be costly

The Costs of Reproducibility

Russell A. Poldrack^{1,*}

¹Department of Psychology, Stanford University, Stanford, CA, USA

*Correspondence: poldrack@stanford.edu

<https://doi.org/10.1016/j.neuron.2018.11.030>

PERSPECTIVE

Open science challenges, benefits and tips in early career and beyond

Christopher Allen^{1,*}, David M. A. Mehler^{1,2,*}

Credit data generators for data reuse

To promote effective sharing, we must create an enduring link between the people who generate data and its future uses, urge **Heather H. Pierce** and colleagues.

Pierce et al 2019

Publish your computer code: it is good enough

*Freely provided working code – whatever its quality – improves programming and enables others to engage with your research, says **Nick Barnes**.*

Barnes 2010

- Improve training
- Avoid shaming -> constructive critique
- Ugly code better than no code

Why doing reproducible research?

Reproducibility: good for you,
good for everyone

Automation (good code) saves time

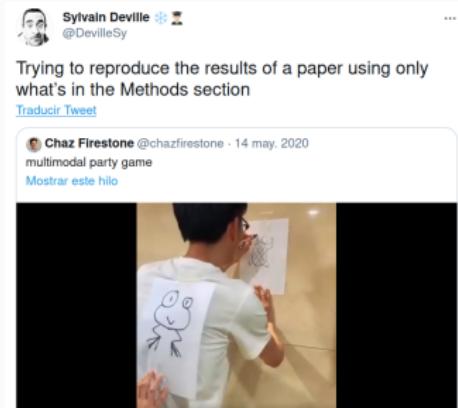


Trevor Branch
@TrevorABranch

...

My rule of thumb: every analysis you do on a dataset will have to be redone 10–15 times before publication. Plan accordingly. [#Rstats](#)

Code = fully traceable, reproducible analysis



Code advantages:

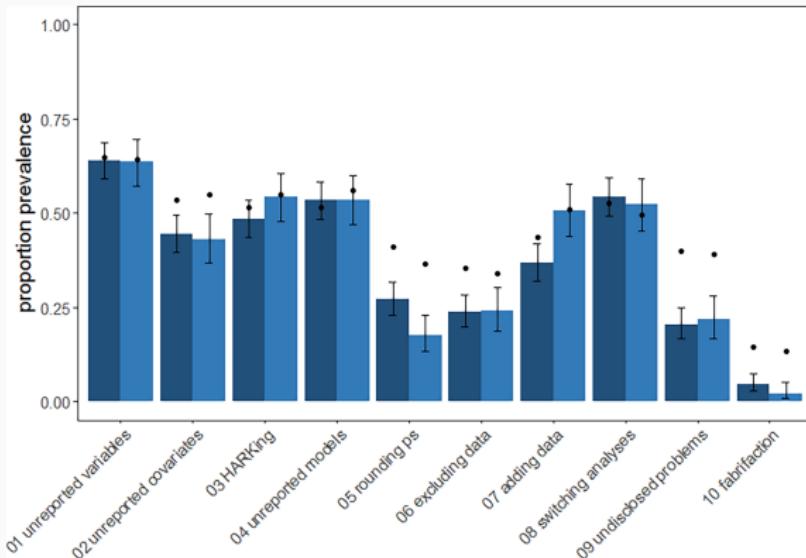
- Easier writing
- Easier, deeper review
- Reusable

Transparency prevents bad practices

RESEARCH ARTICLE

Questionable research practices in ecology and evolution

Hannah Fraser^{1*}, Tim Parker², Shinichi Nakagawa³, Ashley Barnett¹, Fiona Fidler^{1,4}



p-hacking, HARKing, data fabrication...

Transparency avoids unnecessary disputes

DOI:10.1063/PT.6.1.20180822a

22 Aug 2018 in **Research & Technology**

The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top theorists.

Ashley G. Smart

Over the next seven years, the perplexing discrepancy would ignite a bitter conflict, with junior scientists caught in the crossfire. At stake were not only the reputations of the two groups but also a peculiar theory that sought to explain some of water's deepest and most enduring mysteries. Earlier this year, the dispute was finally settled. And as it turns out, the entire ordeal was the result of botched code.

Transparency brings better science



Alexey Shiklomanov
@ashiklom711

...

I'm co-author on a study currently published only as a publicly available discussion paper. My code was on GitHub.

A colleague read the paper, thought the results looked weird, checked my code, found a bug and emailed me about it.

This is how science should work. [#openscience](#)

Many journals (and funders) value/require reproducibility

As a condition for publication in ESA journals, all underlying data and statistical code pertinent to the results presented in the publication must be made available in a permanent, publicly accessible data archive or repository, with rare exceptions (see



Many journals value reproducibility

'Papers with exemplary data and code archiving
are **more valuable** for future research and [...]
will be given **higher priority** for publication'
(*Molecular Ecology*)

RESEARCH ARTICLE

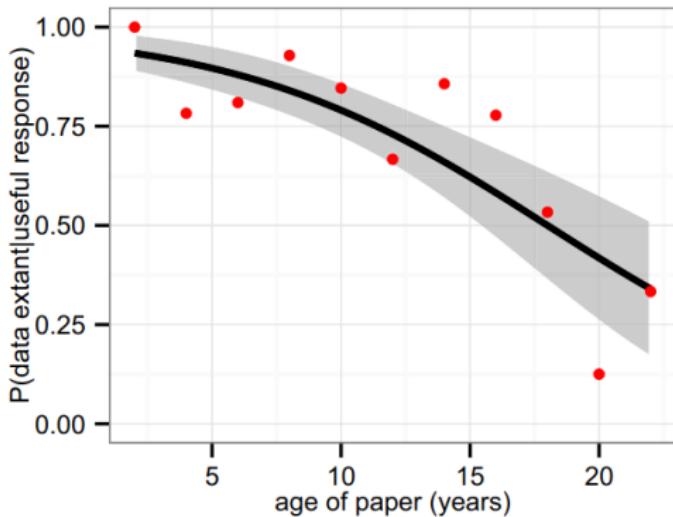
The citation advantage of linking publications to research data

Giovanni Colavizza^{1,2}, Iain Hrynaszkiewicz^{3,4}, Isla Staden^{1,5}, Kirstie Whitaker^{1,6},
Barbara McGillivray^{1,6*}

Colavizza et al 2020

Let's stop losing data & code

The Availability of Research Data Declines Rapidly with Article Age



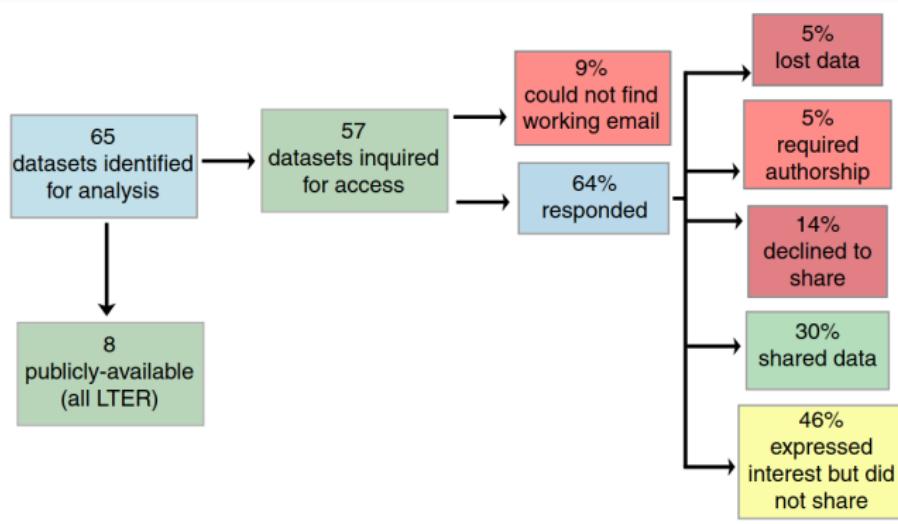
Vines et al 2014

Open data & code enable synthesis

REVIEW

Advances in global change research require open science by individual researchers

ELIZABETH M. WOLKOVICH^{*†}, JAMES REGETZ[‡] and MARY I. O'CONNOR[†]



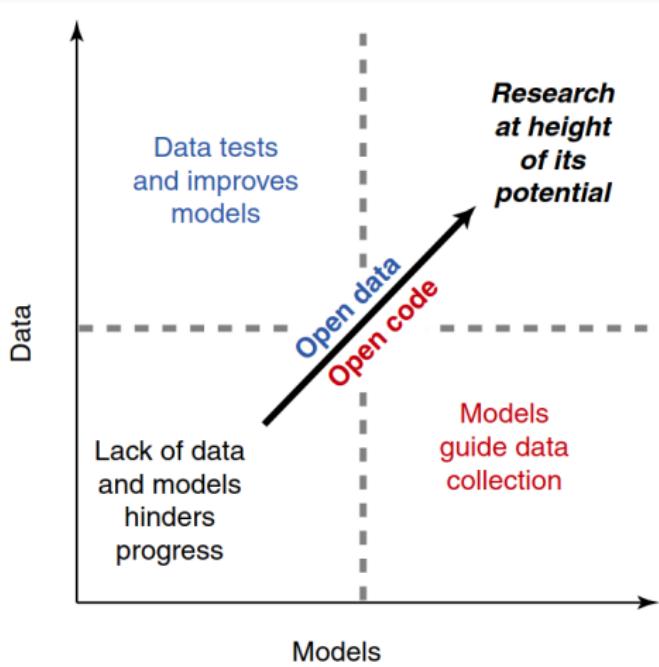
Wolkovich et al 2012

Open data & code enable synthesis

REVIEW

Advances in global change research require open science by individual researchers

ELIZABETH M. WOLKOVICH^{*†}, JAMES REGETZ[‡] and MARY I. O'CONNOR[†]



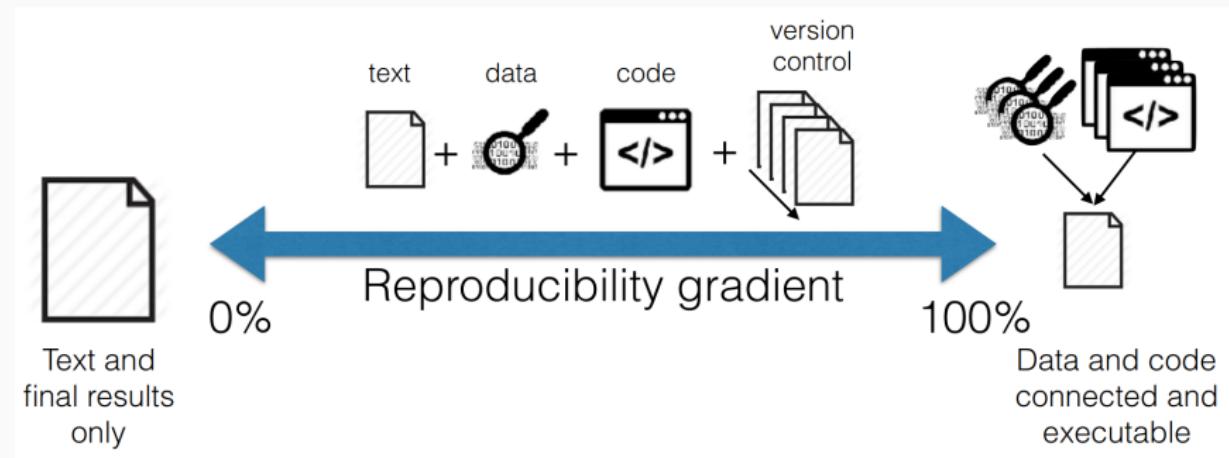
Reproducible workflows facilitate collaboration

and make everyone happier



How to do reproducible research

Reproducibility is a gradient



Rodríguez-Sánchez et al. 2016 (modif. Peng 2011)

Basic reproducibility

Basic reproducibility

- MANUSCRIPT (Text + Tables + Figures)
- DATA in permanent archive (see [Tierney & Ram 2020](#))
- CODE in permanent archive (see [Eglen et al 2016](#))

Permanent archive:

- Zenodo, Dryad, OSF, Figshare, Data Paper...
- NOT GitHub, website...

How to share data

- **Open** format (csv, txt)
- **README** (who, what, when, where, why, how)
- **Describe variables**
- **Licence** (CC0, CC-BY, ODbL)
- **Citation** (DOI)
- **Metadata** standardised (JSON, XML)

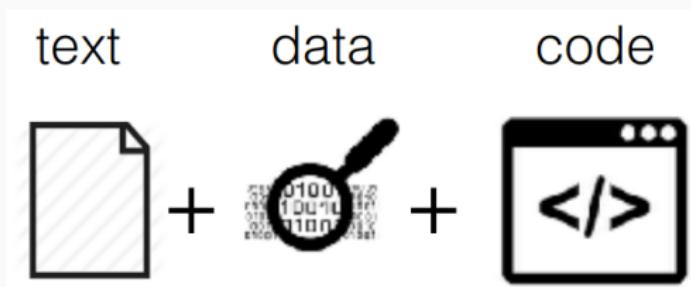
Tierney & Ram 2020

How to share code

- Scripts: plain text (.R)
- Permanent archive (eg. Zenodo) with DOI (citable)
- Licence
- README

Eglen et al 2016

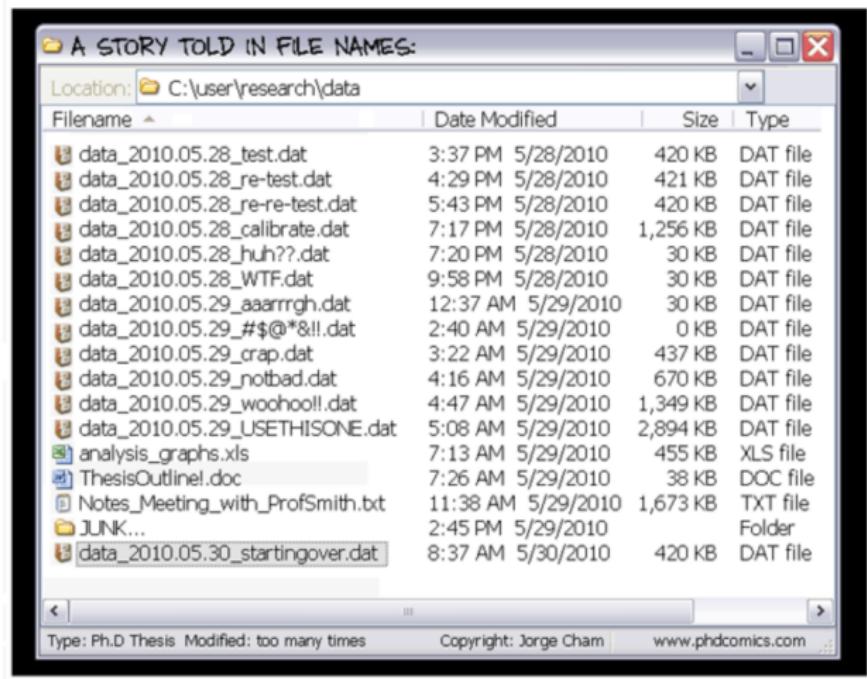
Basic reproducibility



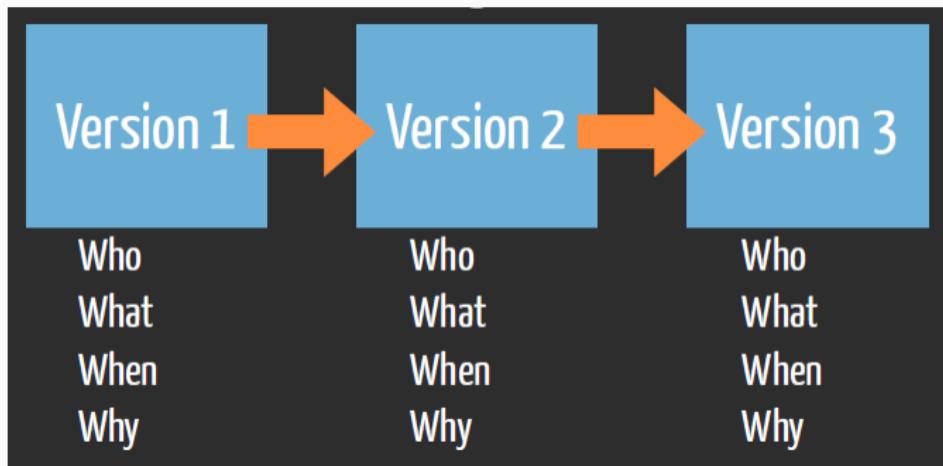
DATA + CODE

- analysis fully **traceable**
- results can be **regenerated**

Version control



Version control with git

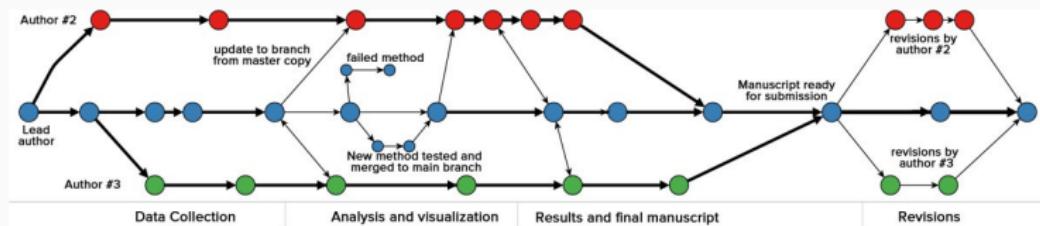


R. Fitzjohn

Much to learn from software engineering

Git can facilitate greater reproducibility and increased transparency in science

Karthik Ram



Ram 2013

Automatic checks with Continuous Integration

Reproducibility of computational workflows is automated using continuous analysis

Brett K Beaulieu-Jones¹ & Casey S Greene²

Pakillo / Carex.bipolar  

Current	Branches	Build History	Pull Requests	More options	☰
✓ master	Pakillo	add two more articles to pkgdown	→ #7 passed → 1c006ff ↗ ⌚ 3 min 22 sec 🕒 a day ago		
✓ master	Pakillo	added leaflet occurrence maps to appear as a	→ #6 passed → 57f5374 ↗ ⌚ 5 min 23 sec 🕒 a day ago		
✓ master	Pakillo	build site with pkgdown	→ #5 passed → 6108a7a ↗ ⌚ 17 min 35 sec 🕒 a day ago		
✗ master	Pakillo	still trying to fix error with sf in travis (via rmat)	→ #4 failed → 2c922d4 ↗ ⌚ 16 min 58 sec 🕒 2 days ago		
✗ master	Pakillo	adding more sf dependencies to travis	→ #3 errored → 5a60b49 ↗ ⌚ 13 min 59 sec 🕒 2 days ago		
✗ master	Pakillo	trying to fix error with rgdal on travis	→ #2 errored → 076af29 ↗ ⌚ 14 min 15 sec 🕒 2 days ago		
✗ master	Pakillo	add travis	→ #1 errored → 4bc6e8 ↗ ⌚ 18 min 54 sec 🕒 3 days ago		

Structuring projects

One Project = One Folder

```
myproject
|
|- data
|
|- code
|
|- output (figures etc)
|
|- manuscript
```

Project-Oriented Workflow: advantages

- Self-contained
- Easy to navigate (file paths)
- Easy to share

Rstudio projects

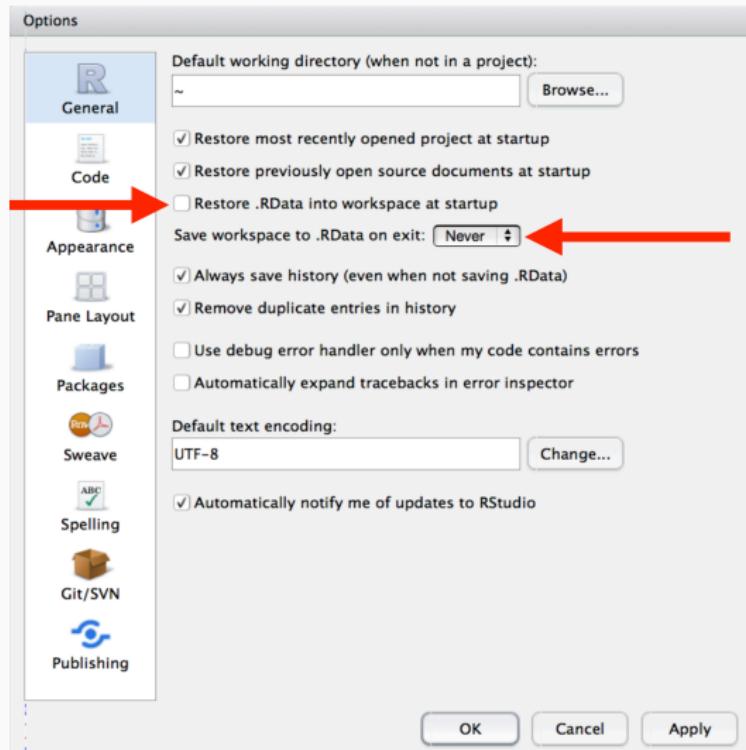
New Project

Create project from:

-  **New Directory**
Start a project in a brand new working directory >
-  **Existing Directory**
Associate a project with an existing working directory >
-  **Version Control**
Checkout a project from a version control repository >

Cancel

Avoid saving workspace



<https://rstats.wtf>

Use `here` for file paths



```
setwd('C:/Users/PACO/myproject')
```



```
library('here')
```

```
mydata <- here('data', 'mydata.csv')
```

fertile package: real-time feedback on reproducibility

```
library('fertile')

setwd("C:/Users/FRS")
```

Error: setwd() is likely to break reproducibility. Use here::here() instead.

<https://github.com/baumer-lab/fertile>

Structuring projects: guidelines

Guidelines for structuring projects

- All files in **same directory**

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**
- **Modular code** (functions)

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**
- **Modular code** (functions)
- Output **disposable & separate** from code

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**
- **Modular code** (functions)
- **Output disposable & separate** from code
- **makefile** runs analyses in **appropriate order**

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**
- **Modular code** (functions)
- **Output disposable & separate** from code
- **makefile** runs analyses in **appropriate order**
- **Software dependencies** under control

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**
- **Modular code** (functions)
- **Output disposable & separate** from code
- **makefile** runs analyses in **appropriate order**
- **Software dependencies** under control
- **README**

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Guidelines for structuring projects

- All files in **same directory**
- Raw data separate from **clean data**
- **Modular code** (functions)
- **Output disposable & separate** from code
- **makefile** runs analyses in **appropriate order**
- **Software dependencies** under control
- README
- License

Noble 2009, Rodriguez-Sanchez et al 2016, Wilson et al 2017

Project organisation example

myproject

- data
 - data-raw
 - data-clean
- code
- output (figures etc)
- manuscript
- README
- License
- Makefile

- What
- Who
- How
- Licence
- Citation
- etc

README.md

pandanusicotopes

 [Launch binder](#)

This repository contains the data and code for our paper:

Florin, A. et al. (2020). *Palaeoprecipitation data from Madjedbebe, northern Australia: A novel proxy from ancient pandanus.*

How to cite

Please cite this compendium as:

Marwick, B., A. Florin et al., (2020). *Compendium of R code and data for Palaeoprecipitation data from Madjedbebe, northern Australia: A novel proxy from ancient pandanus.* Accessed 16 Oct 2020. Online at <https://doi.org/xxx/xxx>

How to download

You can download the compendium as a zip from this URL: <https://github.com/benmarwick/pandanusicotopes/archive/master.zip>

Licenses

Text and figures : CC-BY-4.0

Code : See the [DESCRIPTION](#) file

Data : CC-0 attribution requested in reuse

Document your data

<https://docs.ropensci.org/dataspice/>

```
library("dataspice")
create_spice()    # create CSV templates for metadata

edit_creators()  # open Shiny apps to edit the CSVs
prep_access()
edit_access()
prep_attributes()
edit_attributes()
edit_biblio()

write_spice()    # write machine-readable metadata

build_site()    # build human-readable metadata report
```

Write modular code

Break up scripts

```
prepare_data.R
```

```
run_analysis.R
```

```
make_figures.R
```

(and `makefile` will run them in the right order)

makefile runs code in appropriate order

makefile.R

```
source("prepare_data.R")  
  
source("run_analysis.R")  
  
source("make_figures.R")
```

Don't Repeat Yourself (DRY)

```
dataset %>%  
  filter(species == "Laurus nobilis") %>%  
  ggplot() +  
  geom_point(aes(x, y))  
  
dataset %>%  
  filter(species == "Laurus azorica") %>%  
  ggplot() +  
  geom_point(aes(x, y))
```

Don't Repeat Yourself

Write functions (documented + tested)

```
plot_species <- function(sp, data) {  
  data %>%  
    filter(species == sp) %>%  
    ggplot() +  
    geom_point(aes(x, y))  
}
```

Don't Repeat Yourself

Use functions

```
plot_species(sp = "Laurus nobilis", dataset)
```

```
plot_species(sp = "Laurus azorica", dataset)
```

Don't Repeat Yourself

Use for loops

```
for (i in species) {  
  plot_species(sp = i, dataset)  
}
```

Don't Repeat Yourself

Good ol' `lapply`

```
lapply(species, plot_species, data = dataset)
```

Don't Repeat Yourself

```
library("purrr")  
  
map(species, plot_species, data = dataset)
```

Comment your code

Why rather than What

```
## Response is not linear, so fit gam rather than lm  
  
model.height <- gam(height ~ s(diameter), data = trees)
```

Use meaningful names for objects

```
m1 <- lm(height ~ diameter, data = trees)
m2 <- gam(height ~ s(diameter), data = trees)
```

Use meaningful names for objects

```
m1 <- lm(height ~ diameter, data = trees)  
m2 <- gam(height ~ s(diameter), data = trees)
```

```
model.linear <- lm(height ~ diameter, data = trees)  
model.gam <- gam(height ~ s(diameter), data = trees)
```

Project templates

Automatic project creation with template

```
library('template')
```

```
new_project("mynewproj",
            package = FALSE)
```

	analyses
	data
	data-raw
	manuscript
	R
	.Rproj.user
	makefile.R
 mynewproj.Rproj	
	README.Rmd
	.gitignore

template: New projects also on GitHub

```
new_project("mynewproj",
            package = FALSE,
            github = TRUE)
```

The screenshot shows a GitHub repository page for 'Pakillo / mynewproj'. The repository is private. The 'Code' tab is selected. The master branch has 1 branch and 0 tags. The commit history shows an 'Initial commit' by 'Pakillo' 2 minutes ago, which includes files: .gitignore, README.Rmd, makefile.R, and mynewproj.Rproj.

File	Commit	Time
.gitignore	Initial commit	2 minutes ago
README.Rmd	Initial commit	2 minutes ago
makefile.R	Initial commit	2 minutes ago
mynewproj.Rproj	Initial commit	2 minutes ago



workflowr:
reproducible projects with
website

wflow_start creates project scaffolding

```
library('workflowr')  
  
wflow_start("newproject")
```

	analysis
	code
	data
	docs
	output
	.git
	newproject.Rproj
	README.md
	_workflowr.yml
	.gitattributes
	.gitignore
	.Rprofile

wflow_open starts new analysis

```
wflow_open("analysis/first-analysis.Rmd")
```

```
---
title: "first-analysis"
author: "Pakillo"
date: "2021-06-15"
output: workflowr::wflow_html
editor_options:
  chunk_output_type: console
---
|
## Introduction

```{r}
data(iris)
plot(iris)
```

```

wflow_build runs analyses and generates website

wflow_build()

newproject Home About License

Introduction first-analysis

Pakillo
2021-06-15

workflow ✓

Introduction

```
data(iris)
plot(iris[, 1:2])
```

A scatter plot showing the relationship between Sepal.Length (x-axis) and Sepal.Width (y-axis). The x-axis ranges from 4.5 to 8.0, and the y-axis ranges from 2.0 to 4.0. The data points are open circles, forming three distinct clusters that correspond to the three classes in the Iris dataset: Setosa, Versicolor, and Virginica.

Past versions of unnamed-chunk-1-1.png

Session information

wflow_publish commits changes & updates everything

```
wflow_publish(c("analysis/first-analysis.Rmd",
               "analysis/index.Rmd",
               "analysis/about.Rmd",
               "analysis/license.Rmd"),
               message = "Publish initial analyses")
```

Connect with GitHub/GitLab and deploy website

```
wflow_use_github("Pakillo")  
wflow_git_push()
```

Research compendia: projects as packages

Projects as packages

- Standard structure

Rodríguez-Sánchez et al. 2016, Marwick et al 2018, but see [McBain 2020](#)

Projects as packages

- Standard structure
- Promotes modular code, documented and tested

Rodríguez-Sánchez et al. 2016, Marwick et al 2018, but see [McBain 2020](#)

Projects as packages

- Standard structure
- Promotes modular code, documented and tested
- Easy to share and run

Rodríguez-Sánchez et al. 2016, Marwick et al 2018, but see [McBain 2020](#)

Projects as packages

- Standard structure
- Promotes modular code, documented and tested
- Easy to share and run
- Automatic checks (Continuous Integration)

Rodríguez-Sánchez et al. 2016, Marwick et al 2018, but see [McBain 2020](#)

Projects as packages

- Standard structure
- Promotes modular code, documented and tested
- Easy to share and run
- Automatic checks (Continuous Integration)
- Automatic code review (**goodpractice**)

Rodríguez-Sánchez et al. 2016, Marwick et al 2018, but see [McBain 2020](#)

Projects as packages

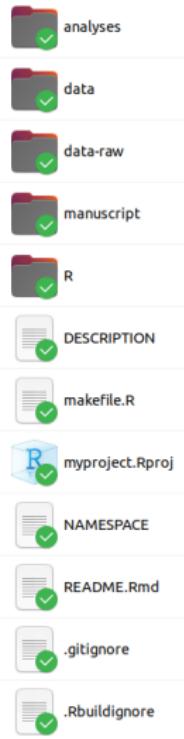
- Standard structure
- Promotes modular code, documented and tested
- Easy to share and run
- Automatic checks (Continuous Integration)
- Automatic code review (**goodpractice**)
- Easily create website with `pkgdown`

Rodríguez-Sánchez et al. 2016, Marwick et al 2018, but see [McBain 2020](#)

Creating package structure with template

```
library('template')
```

```
new_project("myproject",
            package = TRUE)
```

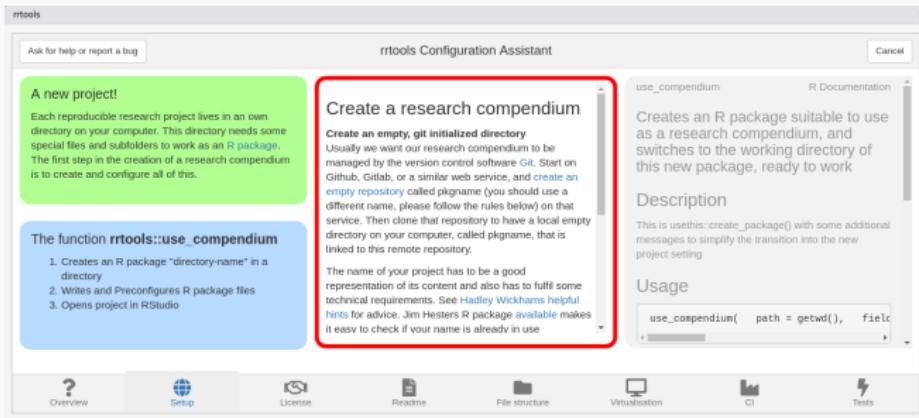


rrtools

rrtools creates research compendia

```
library("rrtools")  
  
use_compendium("~/myproject/")
```

Rstudio addin: <https://github.com/nevrome/rrtools.addin>



rrtools: project structure

```
- README
- LICENSE
- DESCRIPTION
- travis.yml
- Dockerfile
- analysis/
|
|- paper/
    |- paper.Rmd
    |- references.bib
|
|- figures/
|
|- data/
    |- raw_data/
    |- derived_data/
```



rcompendium

rcompendium creates new project with all scaffolding

```
library('rcompendium')  
  
new_compendium()
```

- R package structure
- GitHub repository
- Automatic testing & website update

Continuous Integration (GitHub Actions, GitLab CI...)

Automatic testing with every commit!

| Pakillo / Carex.bipolar | | | | |
|-------------------------|--|-----------------------------|---------------------------------|------------------------------|
| Current | Branches | Build History | Pull Requests | |
| ✓ master
④ Pakillo | add two more articles to pkgdown | → #7 passed
④ 1c006ff ↗ | ⌚ 3 min 22 sec
🕒 a day ago | More options |
| ✓ master
④ Pakillo | added leaflet occurrence maps to appear as a | → #6 passed
④ 57f5374 ↗ | ⌚ 5 min 23 sec
🕒 a day ago | |
| ✓ master
④ Pakillo | build site with pkgdown | → #5 passed
④ 6108a7a ↗ | ⌚ 17 min 35 sec
🕒 a day ago | |
| ✗ master
④ Pakillo | still trying to fix error with sf in travis (via mat | → #4 failed
④ 2c922d4 ↗ | ⌚ 16 min 58 sec
🕒 2 days ago | |
| ! master
④ Pakillo | adding more sf dependencies to travis | → #3 errored
④ 5a60b49 ↗ | ⌚ 13 min 59 sec
🕒 2 days ago | |
| ! master
④ Pakillo | trying to fix error with rgdal on travis | → #2 errored
④ 076af29 ↗ | ⌚ 14 min 15 sec
🕒 2 days ago | |
| ! master
④ Pakillo | add travis | → #1 errored
④ 4bce6e8 ↗ | ⌚ 18 min 54 sec
🕒 3 days ago | |

Minimalistic compendium

<https://github.com/cboettig/compendium>

- DESCRIPTION (dependencies)
- Manuscript (Rmd)
- GitHub Actions

Data management

Data management

See <https://dataoneorg.github.io/Education/bestpractices/>

- 1. Planification (e.g. DMPTool)
- 2. Collection
- 3. Metadata description (dataspice, EML, Data Packages, DataPackageR)
- 4. Quality control (e.g. assertr, validate, pointblank)
- 5. Storage

Document your data

<https://docs.ropensci.org/dataspice/>

```
library("dataspice")
create_spice()    # create CSV templates for metadata

edit_creators()  # open Shiny apps to edit the CSVs
prep_access()
edit_access()
prep_attributes()
edit_attributes()
edit_biblio()

write_spice()    # write machine-readable metadata

build_site()    # build human-readable metadata report
```

Check data before analysis

```
library("assertr")  
  
dataset %>%  
  assert(within_bounds(0, 0.20), fruit.weight) %>%  
  assert(in_set("black", "red"), colour)
```

Check out also [pointblank](#)

Editorial expression of concern

IN THE 3 June issue, *Science* published the Report “Environmentally relevant concentrations of microplastic particles influence larval fish ecology” by Oona M. Lönnstedt and Peter Eklöv (1). The authors have notified *Science* of the theft of the computer on which the raw data for the paper were stored. These data were not backed up on any other device nor deposited in an appropriate repository. *Science* is publishing this Editorial Expression of Concern to alert our readers to the fact that no further data can be made available, beyond those already presented in the paper and its supplement, to enable readers to understand, assess, reproduce, or extend the conclusions of the paper.

Jeremy Berg

Editor in Chief

Use the **cloud**: safe, persistent, easy to share

- [Open Science Framework](#)
- GitHub
- Dropbox
- Figshare, Zenodo, etc
- See all data repositories in www.re3data.org

Tidy data

Tidy data

| country | year | cases | population |
|-------------|------|--------|------------|
| Afghanistan | 1999 | 745 | 1587071 |
| Afghanistan | 2000 | 2666 | 20395360 |
| Brazil | 1999 | 37737 | 172006362 |
| Brazil | 2000 | 80488 | 172004898 |
| China | 1999 | 212258 | 1272015272 |
| China | 2000 | 213766 | 128012583 |

variables

| country | year | cases | population |
|-------------|------|--------|------------|
| Afghanistan | 1999 | 745 | 1587071 |
| Afghanistan | 2000 | 2666 | 20395360 |
| Brazil | 1999 | 37737 | 172006362 |
| Brazil | 2000 | 80488 | 172004898 |
| China | 1999 | 212258 | 1272015272 |
| China | 2000 | 213766 | 128012583 |

observations

| country | year | cases | population |
|-------------|------|--------|------------|
| Afghanistan | 1999 | 745 | 1587071 |
| Afghanistan | 2000 | 2666 | 20395360 |
| Brazil | 1999 | 37737 | 172006362 |
| Brazil | 2000 | 80488 | 172004898 |
| China | 1999 | 212258 | 1272015272 |
| China | 2000 | 213766 | 128012583 |

values

| country | year | cases | country | 1999 | 2000 |
|-------------|------|--------|-------------|--------|--------|
| Afghanistan | 1999 | 745 | Afghanistan | 745 | 2666 |
| Afghanistan | 2000 | 2666 | Brazil | 37737 | 80488 |
| Brazil | 1999 | 37737 | China | 212258 | 213766 |
| Brazil | 2000 | 80488 | | | |
| China | 1999 | 212258 | | | |
| China | 2000 | 213766 | | | |

table4

COMMENT**Open Access**

CrossMark

Gene name errors are widespread in the scientific literature

Mark Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene sym-

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and **special characters** in column names.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data.** Do all data manipulation through code.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data.** Do all data manipulation through code.
- Export data as plain text (txt, csv).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data.** Do all data manipulation through code.
- Export data as plain text (txt, csv).
- <http://www.datacarpentry.org/spreadsheet-ecology-lesson/>

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data.** Do all data manipulation through code.
- Export data as plain text (txt, csv).
- <http://www.datacarpentry.org/spreadsheet-ecology-lesson/>
- <http://kbroman.org/dataorg/>

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or YYYY-MM-DD as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data.** Do all data manipulation through code.
- Export data as plain text (txt, csv).
- <http://www.datacarpentry.org/spreadsheet-ecology-lesson/>
- <http://kbroman.org/dataorg/>
- Broman & Woo: Data organization in spreadsheets

Common spreadsheet errors

More than one variable per column

| Date collected | Plot | Species-Sex | Weight |
|----------------|------|-------------|--------|
| 1/9/78 | 1 | DM-M | 40 |
| 1/9/78 | 1 | DM-F | 36 |
| 1/9/78 | 1 | DS-F | 135 |
| 1/20/78 | 1 | DM-F | 39 |
| 1/20/78 | 2 | DM-M | 43 |
| 1/20/78 | 2 | DS-F | 144 |
| 3/13/78 | 2 | DM-F | 51 |
| 3/13/78 | 2 | DM-F | 44 |
| 3/13/78 | 2 | DS-F | 146 |

| Date collected | Plot | Species | Sex | Weight |
|----------------|------|---------|-----|--------|
| 1/9/78 | 1 | DM | M | 40 |
| 1/9/78 | 1 | DM | F | 36 |
| 1/9/78 | 1 | DS | F | 135 |
| 1/20/78 | 1 | DM | F | 39 |
| 1/20/78 | 2 | DM | M | 43 |
| 1/20/78 | 2 | DS | F | 144 |
| 3/13/78 | 2 | DM | F | 51 |
| 3/13/78 | 2 | DM | F | 44 |
| 3/13/78 | 2 | DS | F | 146 |

Source: Data Carpentry

Multiple tables

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AF | AI | | | | | | | | | | |
|-------------------------|------------------------|---------|---|---|---|---------|-----|-------|----|---------|------|------|------|-------------|------|----------|----|---------|-----|------|------|---------|-------------|------------|---------|----|-----|------|------|---------|-------------|----------|---------|-----|-----|------|------|------|-------------|----|---|---|---|
| 1 lake site May 29 2012 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 1 | T1 | 1 | 1 | 2 | T1 | 2.6 | 0.51 | 1 | T1 | 6 | 85 | 91 | T1 | 30.4 | 15.47126 | 1 | T1 | 17 | 80 | 97 | avr | SEM | 1 | T1 | 52 | 191 | 243 | avr | SEM | 1 | T1 | 52 | 191 | 243 | | | | | | | | |
| 5 | 2 | T1 | 1 | 2 | 3 | T2 | 0.2 | 0.2 | 2 | T1 | 8 | 13 | 21 | T2 | 0.2 | 0.2 | 2 | T1 | 44 | 136 | 180 | T1 | 77.8 | 30.384865 | 2 | T1 | 50 | 270 | 320 | T1 | 541.6 | 50.313 | 2 | T1 | 50 | 270 | 320 | | | | | | |
| 6 | 3 | T1 | 1 | 3 | 4 | control | 0.2 | 0.2 | 3 | T1 | 11 | 0 | 11 | control | 0.6 | 0.6 | 3 | T1 | 18 | 0 | 18 | T2 | 1.8 | 1.5620499 | 3 | T1 | 6 | 0 | 6 | T2 | 0.2 | 0.2 | 3 | T1 | 6 | 0 | 6 | | | | | | |
| 7 | 4 | T1 | 1 | 0 | 1 | | | | 4 | T1 | 0 | 6 | 6 | | | | 4 | T1 | 0 | 14 | 14 | control | 0.4 | 0.244949 | 4 | T1 | 0 | 39 | 39 | control | 0 | 0 | 4 | T1 | 0 | 39 | 39 | | | | | | |
| 8 | 5 | T1 | 0 | 3 | 3 | | | | 5 | T1 | 3 | 20 | 23 | | | | 5 | T1 | 10 | 70 | 80 | | | 5 | T1 | 4 | 96 | 100 | | | | 5 | T1 | 4 | 96 | 100 | | | | | | | |
| 9 | 6 | T2 | 1 | 0 | 1 | | | | 6 | T2 | 0 | 0 | 0 | | | | 6 | T2 | 1 | 7 | 8 | | | 6 | T2 | 0 | 1 | 1 | | | | 6 | T2 | 0 | 1 | 1 | | | | | | | |
| 10 | 7 | T2 | 0 | 0 | 0 | | | | 7 | T2 | 0 | 0 | 0 | | | | 7 | T2 | 0 | 1 | 1 | | | 7 | T2 | 0 | 0 | 0 | | | | 7 | T2 | 0 | 0 | 0 | | | | | | | |
| 11 | 8 | T2 | 0 | 0 | 0 | | | | 8 | T2 | 1 | 0 | 1 | | | | 8 | T2 | 0 | 0 | 0 | | | 8 | T2 | 0 | 0 | 0 | | | | 8 | T2 | 0 | 0 | 0 | | | | | | | |
| 12 | 9 | T2 | 0 | 0 | 0 | | | | 9 | T2 | 0 | 0 | 0 | | | | 9 | T2 | 0 | 0 | 0 | | | 9 | T2 | 0 | 0 | 0 | | | | 9 | T2 | 0 | 0 | 0 | | | | | | | |
| 13 | 10 | T2 | 0 | 0 | 0 | | | | 10 | T2 | 0 | 0 | 0 | | | | 10 | T2 | 0 | 0 | 0 | | | 10 | T2 | 0 | 0 | 0 | | | | 10 | T2 | 0 | 0 | 0 | | | | | | | |
| 14 | 11 | control | 0 | 0 | 0 | | | | 11 | control | 0 | 0 | 0 | | | | 11 | control | 0 | 0 | 0 | | | 11 | control | 0 | 0 | 0 | | | | 11 | control | 0 | 0 | 0 | | | | | | | |
| 15 | 12 | control | 0 | 0 | 0 | | | | 12 | control | 0 | 0 | 0 | | | | 12 | control | 0 | 0 | 0 | | | 12 | control | 0 | 0 | 0 | | | | 12 | control | 0 | 0 | 0 | | | | | | | |
| 16 | 13 | control | 0 | 0 | 0 | | | | 13 | control | 0 | 0 | 0 | | | | 13 | control | 0 | 0 | 0 | | | 13 | control | 0 | 0 | 0 | | | | 13 | control | 0 | 0 | 0 | | | | | | | |
| 17 | 14 | control | 0 | 0 | 0 | | | | 14 | control | 0 | 0 | 0 | | | | 14 | control | 0 | 1 | 1 | | | 14 | control | 0 | 0 | 0 | | | | 14 | control | 0 | 0 | 0 | | | | | | | |
| 18 | 15 | control | 1 | 0 | 1 | | | | 15 | control | 0 | 3 | | | | | 15 | control | 0 | 1 | 1 | | | 15 | control | 0 | 0 | 0 | | | | 15 | control | 0 | 0 | 0 | | | | | | | |
| 19 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 20 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 21 | Barn site May 29, 2012 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 23 | 1 | T1 | 3 | 3 | 6 | | | | 29 | May | plot | bug1 | bug2 | gen
eral | | | | 29 | Jun | plot | bug1 | bug2 | gen
eral | | | 19 | Jun | plot | bug1 | bug2 | gen
eral | | | 26 | Jun | plot | bug1 | bug2 | gen
eral | | | | |
| 24 | 2 | T1 | 1 | 4 | 5 | | | | 2 | T1 | 26 | 36 | 74 | 110 | | | | 2 | T1 | 65 | 502 | 567 | | | 2 | T1 | 0 | 0 | 0 | | | 2 | T1 | 0 | 0 | 0 | | | 2 | T1 | 0 | 0 | 0 |
| 25 | 3 | T1 | 0 | 0 | 0 | T1 | 2.4 | 1.288 | 3 | T1 | 13 | 0 | 13 | T1 | 30.6 | 20.10124 | 3 | T1 | 10 | 7 | 27 | T1 | 119.4 | 111.912882 | 3 | T1 | 12 | 20 | 32 | T2 | 0.4 | 0.4 | 3 | T1 | 12 | 20 | 32 | | | | | | |
| 26 | 4 | T1 | 0 | 0 | 0 | T2 | 0.4 | 0.245 | 4 | T1 | 7 | 0 | 7 | T2 | 5 | 0.774597 | 4 | T1 | 0 | 6 | 6 | T2 | 5 | 2.1908902 | 4 | T1 | 0 | 16 | 16 | control | 1.2 | 0.5831 | 4 | T1 | 0 | 16 | 16 | | | | | | |
| 27 | 5 | T1 | 0 | 1 | 1 | control | 1 | 0.316 | 5 | T1 | 2 | 0 | 2 | control | 2.2 | 1.714643 | 5 | T1 | 0 | 2 | 2 | control | 2.8 | 0.969556 | 6 | T2 | 0 | 0 | 0 | control | 2.8 | 0.969556 | 6 | T2 | 0 | 0 | 0 | | | | | | |
| 28 | 6 | T2 | 0 | 0 | 0 | | | | 6 | T2 | 1 | 0 | 1 | | | | 6 | T2 | 0 | # | 8 | | | 6 | T2 | 0 | 0 | 0 | | | 6 | T2 | 0 | 0 | 0 | | | | | | | | |
| 29 | 7 | T2 | 0 | 0 | 0 | | | | 7 | T2 | 0 | 4 | 4 | | | | 7 | T2 | 0 | 12 | 12 | | | 7 | T2 | 0 | 0 | 0 | | | 7 | T2 | 0 | 0 | 0 | | | | | | | | |
| 30 | 8 | T2 | 0 | 1 | 1 | | | | 8 | T2 | 0 | 0 | 0 | | | | 8 | T2 | 0 | 0 | 0 | | | 8 | T2 | 0 | 0 | 0 | | | 8 | T2 | 0 | 0 | 0 | | | | | | | | |
| 31 | 9 | T2 | 0 | 1 | 1 | | | | 9 | T2 | 0 | 0 | 0 | | | | 9 | T2 | 3 | 0 | 3 | | | 9 | T2 | 0 | 0 | 0 | | | 9 | T2 | 0 | 0 | 0 | | | | | | | | |
| 32 | 10 | T2 | 0 | 0 | 0 | | | | 10 | T2 | 0 | 0 | 0 | | | | 10 | T2 | 2 | 0 | 2 | | | 10 | T2 | 0 | 2 | 2 | | | 10 | T2 | 0 | 2 | 2 | | | | | | | | |
| 33 | 11 | control | 0 | 0 | 0 | | | | 11 | control | 0 | 1 | 1 | | | | 11 | control | 0 | 5 | 5 | | | 11 | control | 0 | 2 | 2 | | | 11 | control | 0 | 2 | 2 | | | | | | | | |
| 34 | 12 | control | 0 | 1 | 1 | | | | 12 | control | 0 | 0 | 0 | | | | 12 | control | 1 | 1 | 1 | | | 12 | control | 1 | 0 | 1 | | | 12 | control | 1 | 0 | 1 | | | | | | | | |
| 35 | 13 | control | 0 | 1 | 1 | | | | 13 | control | 0 | 0 | 0 | | | | 13 | control | 0 | 0 | 0 | | | 13 | control | 0 | 0 | 0 | | | 13 | control | 0 | 0 | 0 | | | | | | | | |
| 36 | 14 | control | 1 | 1 | 1 | | | | 14 | control | 0 | 1 | 9 | | | | 14 | control | 0 | 5 | 5 | | | 14 | control | 0 | 3 | 3 | | | 14 | control | 0 | 3 | 3 | | | | | | | | |
| 37 | 15 | control | 0 | 2 | 2 | | | | 15 | control | 0 | 1 | 1 | | | | 15 | control | 0 | 2 | 2 | | | 15 | control | 0 | 1 | 0 | | | 15 | control | 0 | 1 | 0 | | | | | | | | |
| 38 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 39 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Multiple tabs

Could you avoid new tab by adding a column to original spreadsheet?

Using formatting, comments, etc to convey information

| Plot: 2 | | | |
|----------------|---------|-----|--------|
| Date collected | Species | Sex | Weight |
| 1/8/14 | NA | | |
| 1/8/14 | DM | M | 44 |
| 1/8/14 | DM | M | 38 |
| 1/8/14 | OL | | |
| 1/8/14 | PE | M | 22 |
| 1/8/14 | DM | M | 38 |
| 1/8/14 | DM | M | 48 |
| 1/8/14 | DM | M | 43 |
| 1/8/14 | DM | F | 35 |
| 1/8/14 | DM | M | 43 |
| 1/8/14 | DM | F | 37 |
| 1/8/14 | PF | F | 7 |
| 1/8/14 | DM | M | 45 |
| 1/8/14 | OT | | |
| 1/8/14 | DS | M | 157 |
| 1/8/14 | OX | | |
| 2/18/14 | NA | M | 218 |
| 2/18/14 | PF | F | 7 |
| 2/18/14 | DM | M | 52 |

measurement device not calibrated

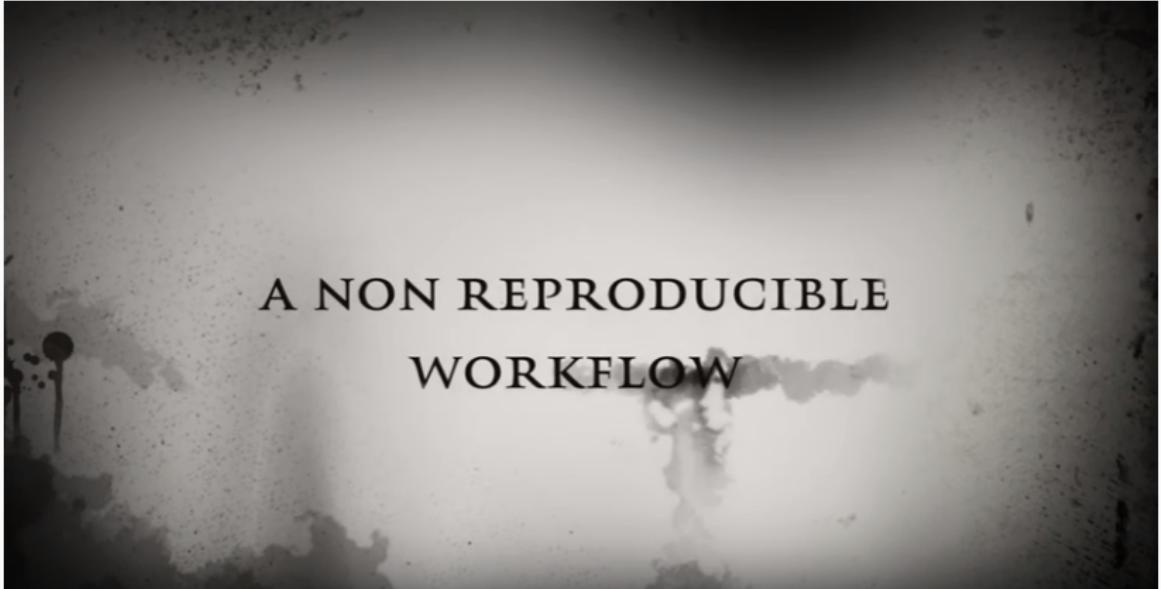
| Date collected | Species | Sex | Weight | Calibrated |
|----------------|---------|-----|--------|------------|
| 1/8/14 | NA | | | |
| 1/8/14 | DM | M | 44 | Y |
| 1/8/14 | DM | M | 38 | Y |
| 1/8/14 | OL | | | |
| 1/8/14 | PE | M | 22 | Y |
| 1/8/14 | DM | M | 38 | Y |
| 1/8/14 | DM | M | 48 | Y |
| 1/8/14 | DM | M | 43 | Y |
| 1/8/14 | DM | F | 35 | Y |
| 1/8/14 | DM | M | 43 | Y |
| 1/8/14 | DM | F | 37 | Y |
| 1/8/14 | PF | F | 7 | Y |
| 1/8/14 | DM | M | 45 | Y |
| 1/8/14 | OT | | | |
| 1/8/14 | DS | M | 157 | N |
| 1/8/14 | OX | | | |
| 2/18/14 | NA | M | 218 | N |
| 2/18/14 | PF | F | 7 | Y |
| 2/18/14 | DM | M | 52 | Y |

Your turn: tidy up this messy dataset

<https://ndownloader.figshare.com/files/2252083>

Reproducible dynamic documents with Rmarkdown

A scary movie... with happy ending



A NON REPRODUCIBLE
WORKFLOW

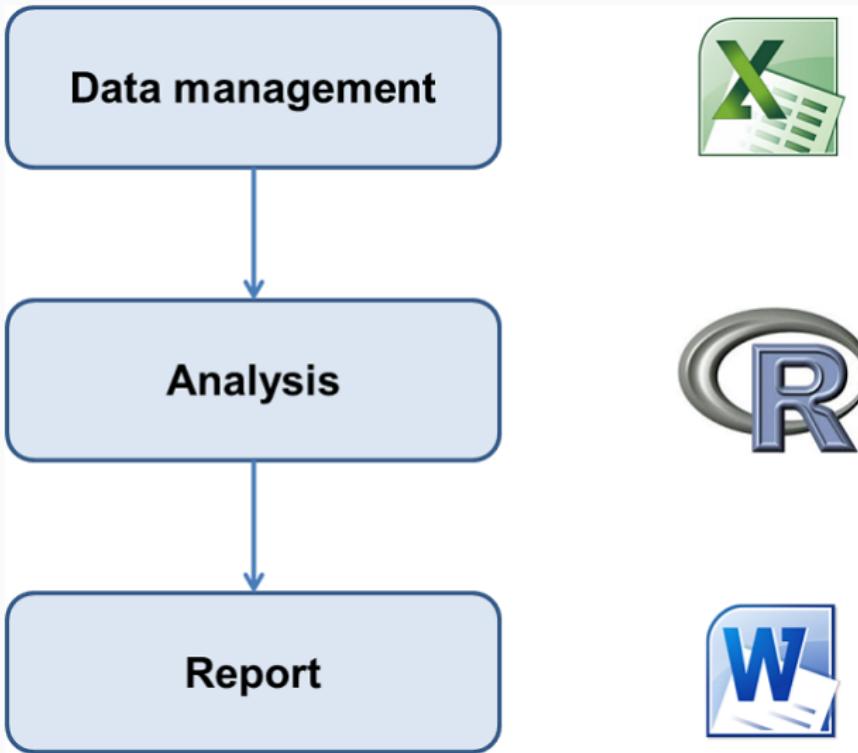
<https://youtu.be/s3JldKoA0zw>

A typical research workflow

1. Prepare data (spreadsheet)
2. Analyse data (R)
3. Write report/paper (Word)
4. Start the email attachments
nightmare...



This workflow is broken



Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for ...?

Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for ...?
- What dataset was used? Which individuals were left out? Where is the clean dataset?

Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for ...?
- What dataset was used? Which individuals were left out? Where is the clean dataset?
- Oops, there is an error in the data. Can you repeat the analysis? And update figures/tables in Word!

Manual copy-paste is tedious & problematic

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|----------|
| (Intercept) | -0.0651657 | 0.4264970 | -0.153 | 0.879 |
| sunshine | 0.0100228 | 0.0004232 | 23.683 | <2e-16 |

'Transcribing numbers from stats software by hand was the largest source of errors'

(Eubank 2016)



Trevor A. Branch

@TrevorABranch

 Follow

My rule of thumb: every analysis you do on a dataset will have to be redone 10–15 times before publication. Plan accordingly. #Rstats

Your **closest collaborator** is you 6 months ago,
and you don't respond to emails.

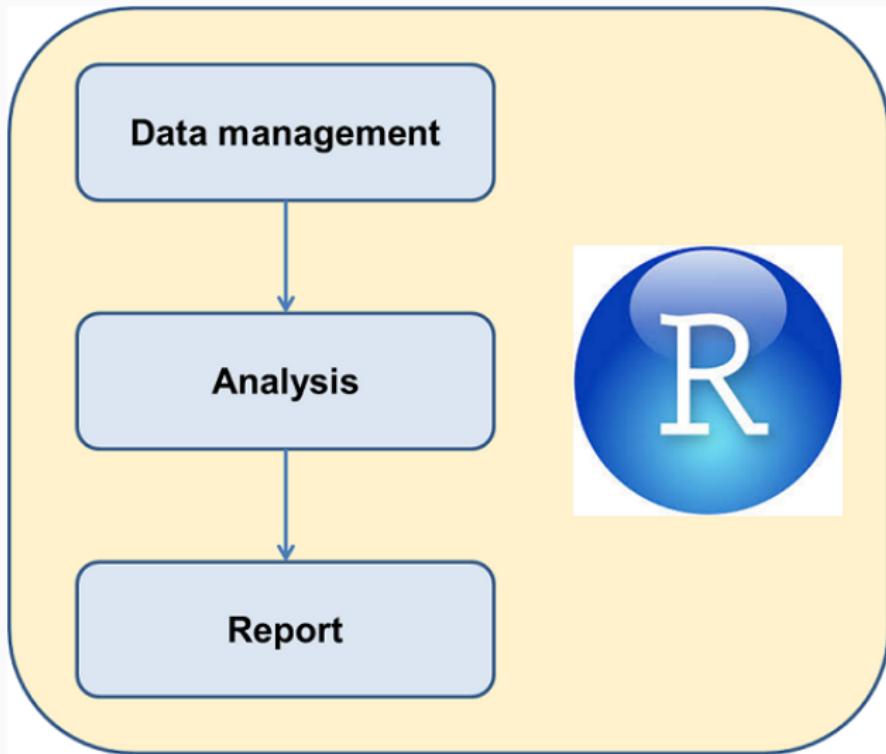
(P. Wilson)

Even **you** will struggle to reproduce
your own results from a few weeks/months ago.

Writing reproducible manuscripts is hard

Revising non-reproducible manuscripts is even harder

Also, please note that because rev#1
asked to re-calculate effect sizes (...)
we need to change every single
number in the main text.



Rmarkdown documents

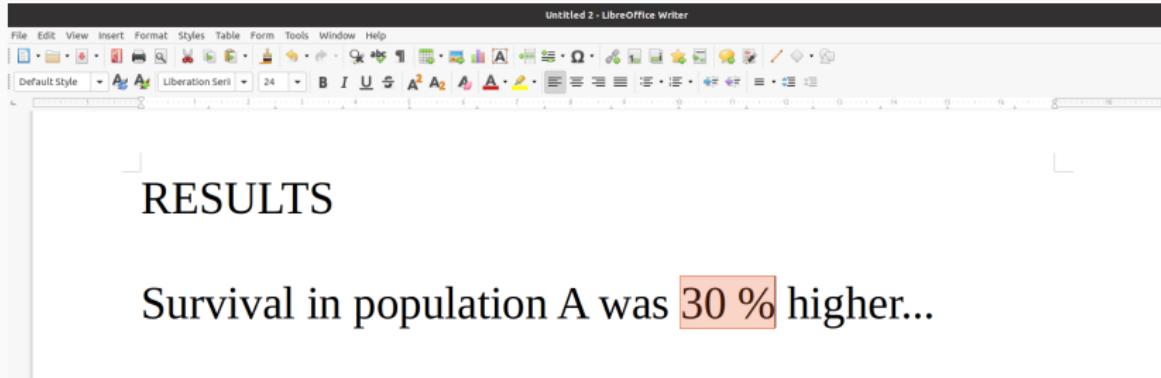
- Fully reproducible (trace all results inc. tables and plots)
- Dynamic (regenerate with 1 click)
- Multiple outputs:
 - documents (HTML, Word, PDF)
 - presentations (HTML, PDF, PowerPoint)
 - books
 - websites...



Where does this value come from?

RESULTS

Survival in population A was 30 % higher...

A screenshot of a LibreOffice Writer document. The title bar says "Untitled 2 - LibreOffice Writer". The menu bar includes File, Edit, View, Insert, Format, Styles, Table, Form, Tools, Window, Help. The toolbar below has icons for various functions like bold, italic, underline, and tables. The main text area contains the word "RESULTS" in large, bold, black font, and the sentence "Survival in population A was 30 % higher...". The "30 %" is highlighted with a pink rectangular box.

Dynamic documents with Rmarkdown

Rmarkdown:

Survival in population A was `r surv.diff` % higher

Output:

Survival in population A was 30 % higher

Dynamic documents with Rmarkdown

```
mydata <- read.csv("data.txt")
```

Rmarkdown:

We measured `r nrow(mydata)` individuals

Output:

We measured **100** individuals

Much better than copy-paste!

Rmarkdown: code (R, Python, etc) + text (Markdown)

```
---
```

```
title: "Does sunshine make people happy?"  
author: "FRS"  
output: word_document  
---  
  
## Introduction  
  
It is well known that individual well-being can be influenced by climatic conditions.  
  
## Methods  
  
```{r echo=FALSE}  
Read data
data <- read.table("data.txt", header = TRUE)

Fit linear model
model <- lm(happiness ~ sunshine, data = data)
```
```

Metadata
(YAML)

Text
(Markdown)

Code
(R, Python...)

We collected data on `r nrow(data)` individuals and fitted a linear model.

Code chunk options

```
```{r echo=FALSE, eval=TRUE, cache=TRUE, fig.height=3}
plot(iris)
```
```

<https://yihui.org/knitr/options/>

Code chunk options

```
```{r}
#| echo = FALSE
#| eval = TRUE
#| fig.cap = "My figure caption"
plot(iris)
```
```

Naming chunks helps debugging

```
processing file: test.Rmd
|.....
ordinary text without R code | 14%

|.....
label: setup (with options) | 29%
List of 1
$ include: logi FALSE

|.....
ordinary text without R code | 43%

|.....
label: read.data | 57%
|.....
ordinary text without R code | 71%

|.....
label: plot (with options) | 86%
List of 1
$ echo: logi FALSE

Quitting from lines 28-29 (test.Rmd)
Error in eval(predvars, data, env) : object 'specie' not found
Calls: <Anonymous> ... plot.formula -> eval -> eval -> <Anonymous> -> eval -> eval
Execution halted
```

Naming chunks helps navigating long docs

```
1 ---  
2 title: "My Analysis"  
3 author: "FRS"  
4 output: html_document  
5 ---  
6  
7 ```{r setup, include=FALSE}  
8 knitr::opts_chunk$set(echo = TRUE)  
9 ```  
10  
11 This is an R Markdown document. Markdown is a simple  
12 My Analysis for authoring HTML, PDF, and MS Word  
Chunk 1: setup re details on using R Markdown see  
Chunk 2: read.data .rstudio.com>.  
Chunk 3: plot
```

11:60

(Top Level) ▾

R Markdown ▾

Naming chunks: figure files take chunk name



unnamed-chunk-1-1.png



unnamed-chunk-1-2.png



unnamed-chunk-1-3.png



unnamed-chunk-1-4.png

Not only R! Python, Julia, C++, SQL, Stan, etc

knitr engines:

```
[1] "asis"      "asy"       "awk"        "bash"       "block"      "block2"
[7] "bslib"     "c"         "cat"        "cc"        "coffee"    "comment"
[13] "css"       "dita"      "dot"        "embed"     "eviews"    "exec"
[19] "fortran"   "fortran95" "gawk"      "go"        "groovy"    "haskell"
[25] "highlight" "js"        "julia"      "lein"      "mysql"     "node"
[31] "octave"    "perl"      "php"        "psql"      "python"    "R"
[37] "Rcpp"      "Rscript"   "ruby"       "sas"       "sass"      "scala"
[43] "scss"      "sed"       "sh"        "sql"       "stan"      "stata"
[49] "targets"   "tikz"      "verbatim"  "zsh"
```

Markdown: easy text formatting

```
# Header  
## Subheader  
*italic*  
**bold**  
[a link](https://example.com)
```

.

Handy: <https://thinkr-open.github.io/remedy/>

Or use [Visual Markdown Editor](#)

Regenerate Word/PDF/HTML with one click

```
---
```

```
title: "Does sunshine make people happy?"  
output: pdf_document  
bibliography: refs.bib  
---
```

Introduction

Climate influences individual well-being [Rehdanz_2005]. However, ...

Methods

```
```{r echo=FALSE}  
read data
data <- read.table("data.txt", header=T)
data[10,1] <- 11 # correct error
```

```
fit linear model
model <- lm(happiness ~ sunshine, data=data)
```

We collected data on `r nrow(data)` individuals and fitted a linear model.

### # Results

We found that...

```
```{r echo=FALSE, results='asis'}  
# make table with model output  
print(xtable::xtable(model), comment = FALSE)
```

```
```{r echo=FALSE, fig.height=3, fig.width=3, fig.align='center'}  
visreg::visreg(model) # plot
```

### # Discussion

Our results confirm that happiness is related to sunshine (slope = `r coef(model)[2]`).

### # References

**a**

## Does sunshine make people happy?

### Introduction

Climate influences individual well-being (Rehdanz and Maddison 2005). However, ...

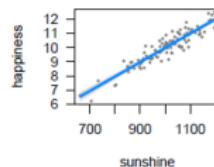
### Methods

We collected data on 100 individuals and fitted a linear model.

### Results

We found that...

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.0986	0.4271	-0.23	0.8180
sunshine	0.0101	0.0004	23.75	0.0000



### Discussion

Our results confirm that happiness is related to sunshine (slope = 0.0100652).

### References

Rehdanz, Katrin, and David Maddison. 2005. "Climate and Happiness." *Ecological Economics* 52 (1). Elsevier BV: 111–25. doi:10.1016/j.ecolecon.2004.06.015.

**b**

## Spotted error in the data? No problem!

- Make changes in Rmarkdown document
- Click **Knit** in Rstudio
- Report will **update automatically!**

# Why Rmarkdown?

The screenshot shows a blog post titled "Convince me to start using R Markdown" by Darren Dahly. The post features a list of five steps for convincing others to use R Markdown.

**Convince me to start using R Markdown**

R Markdown rmarkdown

Darren\_Dahly

1. Start using R Markdown to generate reports of your data analyses.
2. If the data changes, rerun the report with a click of the mouse.
3. Take 3 days off of work.
4. On the 4th day, tell your collaborators that the re-analysis is complete.
5. Be hailed as a hero.

<https://community.rstudio.com/t/convince-me-to-start-using-r-markdown/1636/12>

Your turn

---

# Create, edit and share Rmarkdown document

File > New File > Rmarkdown

Write text

Insert code chunks

Change chunk options (echo, eval, etc)

HTML/Word/PDF output

PDF generation requires LaTeX

```
library('tinytex')

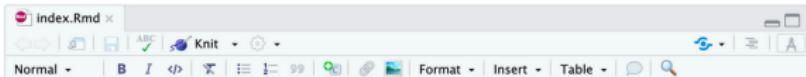
install_tinytex()
```

## Rmarkdown bells and whistles

---

# 'Visual Rmarkdown': Rmd as in word processor

The editor toolbar includes buttons for the most commonly used formatting commands:



Additional commands are available on the **Format**, **Insert**, and **Table** menus:

A screenshot of the RStudio Visual Markdown menu bar with three menus open: Format, Insert, and Table. The Format menu contains options for text styling (Bold, Italic, etc.), block types (Text, Bullets &amp; Numbering, Blockquote, Line Block, Div Block, Code Block), and document cleanup (Clear Formatting, Edit Attributes). The Insert menu contains options for Rmd Chunk, Image, Link, Horizontal Rule, Definition, Inline Math, Display Math, Footnote, Citation, Div Block, Code Block, YAML Block, and Comment. The Table menu contains options for inserting tables, managing rows and columns, and deleting rows or tables.

<https://rstudio.github.io/visual-markdown-editing>

## Automatic table generation

```
model <- lm(happiness ~ sunshine, data = mydata)
xtable(model)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.0652	0.4265	-0.15	0.8789
sunshine	0.0100	0.0004	23.68	0.0000

Many alternatives: `gtsummary`, `modelsummary`, `huxtable`, etc

## equatiomatic describes model structure

We fitted a linear model:

```
library('equatiomatic')
model <- lm(happiness ~ sunshine, data = mydata)
extract_eq(model)
```

$$\text{happiness} = \alpha + \beta_1(\text{sunshine}) + \epsilon \quad (1)$$

# Models that describe themselves!

```
library("report")
model <- lm(happiness ~ sunshine, data = mydata)
report(model)
```

We fitted a linear model (estimated using OLS) to predict happiness with sunshine (formula:  $\text{happiness} \sim \text{sunshine}$ ). The model explains a statistically significant and substantial proportion of variance ( $R^2 = 0.85$ ,  $F(1, 98) = 560.90$ ,  $p < .001$ , adj.  $R^2 = 0.85$ ). The model's intercept, corresponding to  $\text{sunshine} = 0$ , is at -0.07 (95% CI [-0.91, 0.78],  $t(98) = -0.15$ ,  $p = 0.879$ ). Within this model:

- The effect of sunshine is statistically significant and positive ( $\beta = 0.01$ , 95% CI [9.18e-03, 0.01],  $t(98) = 23.68$ ,  $p < .001$ ; Std.  $\beta = 0.92$ , 95% CI [0.85, 1.00])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

# Insert equations with LaTeX

Using LaTeX:

```
$$
y \sim N(\mu, \sigma^2)
$$
```

$$y \sim N(\mu, \sigma^2)$$

- Mathpix: <https://github.com/jonocarroll/mathpix>

# Citing bibliography

Insert Citation

My Sources

- Bibliography
- Zotero
- My Library
- From DOI
- Crossref
- DataCite
- PubMed

Search for citation

	@boghizadehfini2020 What dentists need to know about COVID-19	Baghizadeh Fini, M 2020	
	@bostanciklioglu2020 Severe Acute Respiratory Syndrome Coronavirus 2 is Penetrating to Dementia Re...	Bostanciklioglu, M 2020	
	@fran Functional reactive animation	Elliott, C, and Hudak, P 1997	
	@guo2020 The origin, transmission and clinical therapies on coronavirus disease 2019 (CO...	Guo, Y, Cao, Q, Hong, Z, Tan, Y, Chen, et al. 2020	
	@hu2020 The cytokine storm and COVID-19	Hu, B, Huang, S, and Yin, L 2020	
	@malik2020 Coronavirus Disease Pandemic (COVID-19): Challenges and a Global Perspective	Malik, Y, Kumar, N, Sircar, S et al. 2020	
	@R-base R: A language and environment for statistical computing	R Core Team 2017	

Selected Citation Keys

Add to bibliography: book.bib

<https://rstudio.github.io/visual-markdown-editing/#/citations>

## Using BibTeX file with references

```

```

```
title: "My awesome Rmd"
```

```
output: html_document
```

```
bibliography: references.bib
```

```
--
```

## Format bibliography for any journal

```

```

```
title: "Does sunshine make people happy?"
author: "FRS"
output: word_document
bibliography: myrefs.bib
csl: ecology-letters.csl
```

```

```

Thousands of Citation Styles:

<https://www.zotero.org/styles>

<https://github.com/citation-style-language/styles>

# Rmarkdown templates

- rticles
- papaja
- rrtools
- pinc
- rmdTemplates
- pagedreport
- GitHub!

## My cool paper written in Rmarkdown

F. Rodriguez-Sánchez<sup>a,1,2</sup> and And François<sup>a,b</sup>

<sup>a</sup>Some Institute of Technology, Department, Street, City, State, Zip; <sup>b</sup>Another University Department, Street, City, State, Zip

This manuscript was completed on September 10, 2010.

Please provide an abstract of no more than 250 words in a single paragraph. Abstracts should explain to the general reader the major contributions of the article. References in the abstract must be cited in full when the abstract reads and cited in the text.

one | two | option1 | option2 | option3

This PNAS journal template is provided to help you write your work in the current journal format. Instructions for use are given below.

Note: please start your introduction without including the word "Introduction" as a section heading (except for math articles in the Physical Sciences section); this heading is implied in the first paragraph.

### Guide to using this template

Please note that while this template provides a preview of the typesetting requirements for a manuscript to help in its preparation, it will not necessarily be the final output of the journal. For more detailed information please see the [PNAS Information for Authors](#).

**Author Affiliations.** Include departments, institutions, and complete address, with the ZIP/postal code, for each author. Use lowercase letters to denote the institution for each author as shown in the example. Authors with an ORCID ID may apply this to a published article.

**Submitting Manuscripts.** All authors must submit their article to [PNAScentral](#). All authors must submit their article to [PNAScentral](#). If you are using [Word](#) to write your paper, please use the [Word2PNAS](#) option in the top bar of the editor window.

**Format.** Many authors find it useful to organize their manuscripts with the following order of sections: Title, Author Affiliations, Keywords, Abstract, Significance Statement, Results, Discussion, Materials and Methods, Acknowledgments, and References. Other sections, such as *Supplementary Information*, can be included at the end of the manuscript.

**Manuscript Length.** PNAS generally uses a two-column format and accepts 67 characters, including end-of-line line feeds. The maximum length of a Division/Subdivision research article is six pages and a PNAS PLUS research article is ten pages including all figures, tables, and equations. When submitting tables, figures, and/or equations in addition to text, keep the total for your manuscript under 30,000 characters (including spaces) for Division/Subdivision and 70,000 characters (including spaces) for PNAS PLUS.

**References.** References should be cited in numerical order as they appear in text; this will be done automatically via bibtex, e.g., (1) and (2, 3). All references, including those for the SI, should be included in the main document file. References appearing in both sections should not be duplicated. SI references



Fig. 1. Placeholder image of a bright green tree frog to show publication styling.

included in tables should be included with the main reference section.

**Data & Additions.** PNAS must be able to archive the data essential to a published article. Where such archiving is not possible, deposition of data in public databases, such as GenBank, Aragénogenome, Protein Data Bank, UniProt, and others outlined in the Information for Authors, is required. Authors should indicate in the manuscript where data are available online and their use has or is having an acceptance of a manuscript for publication.

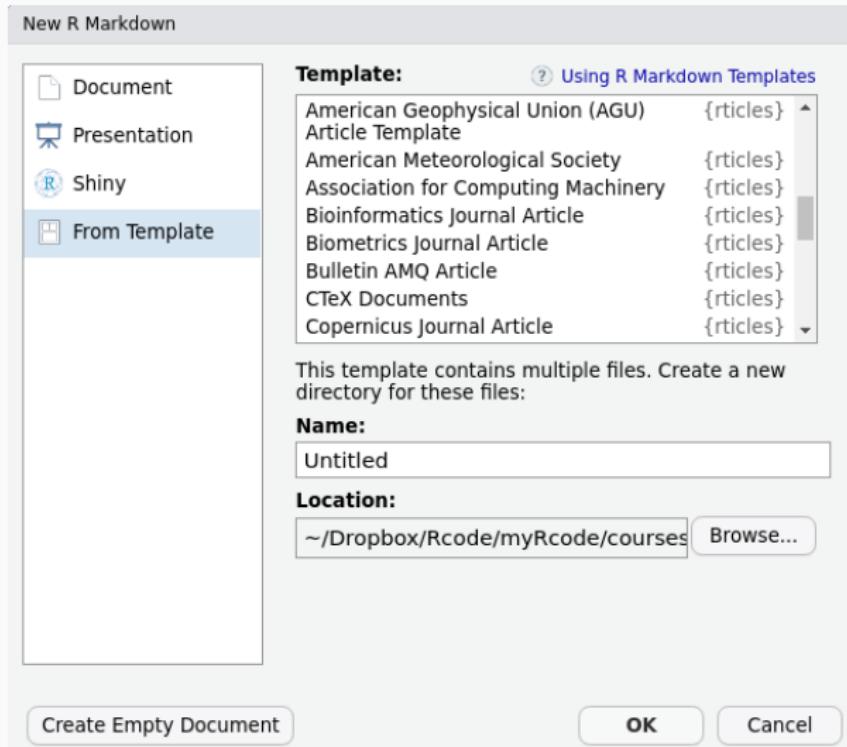
### Significance Statement

Authors must submit a 150-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of specialty. This summary of the work is intended to aid a broad readership. The Significance Statement appears in the paper itself and is required for all submission projects.

Please provide details of author contributions.

Please declare conflict of interests.

# Accessing Rmd templates



## Revise writing style: gramr

The screenshot shows a user interface for a grammar checker named "gramr". On the left, there is a sidebar titled "Ignore" containing a list of writing style rules, each with a checkbox. The rule "So' at start of sentence" has a checked checkbox. At the bottom of this sidebar are two buttons: "Next" and "Finish". The main area is titled "Text to Check" and contains a text box with the following content:

So the cat was stolen. This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <<http://rmarkdown.rstudio.com>>.

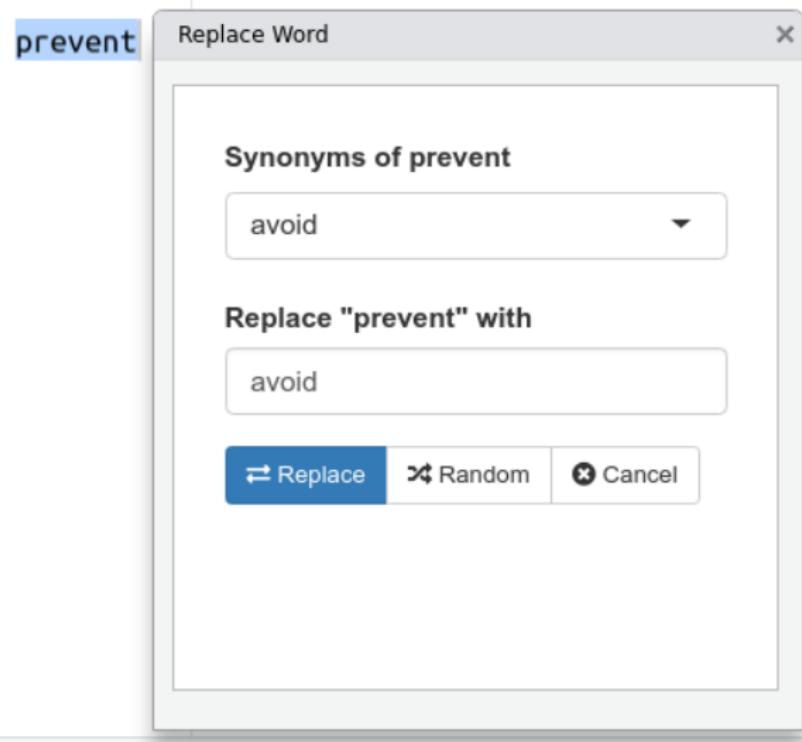
Below the text box, a message box displays the feedback: "was stolen" may be passive voice.

<https://github.com/ropenscilabs/gramr>

## Spell and grammar checking

<https://github.com/nevrome/wellspell.addin>

## Find synonyms



<https://github.com/gadenbuie/synamyn>

## Word count and readability

Method	koRpus	stringi
Word count	107	104
Character count	604	603
Sentence count	10	Not available
Reading time	0.5 minutes	0.5 minutes

<https://github.com/benmarwick/wordcountaddin>

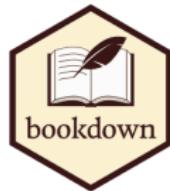
# Write books, theses, with bookdown

## BOOKDOWN

### Write HTML, PDF, eBook, and Kindle books with R Markdown

The `bookdown` package is an [open-source R package](#) that facilitates writing books and long-form articles/reports with R Markdown. Features include:

- Generate printer-ready books and eBooks from R Markdown documents.
- A markup language easier to learn than LaTeX, and to write elements such as section headers, lists, quotes, figures, tables, and citations.
- Multiple choices of output formats: PDF, LaTeX, HTML, EPUB, and Word.
- Possibility of including dynamic graphics and interactive applications (HTML widgets and Shiny apps).
- Support a wide range of languages: R, C/C++, Python, Fortran, Julia, Shell scripts, and SQL, etc.
- LaTeX equations, theorems, and proofs work for all output formats.
- Can be published to GitHub, [bookdown.org](#), and any web servers.
- Integrated with the RStudio IDE.
- One-click publishing to <https://bookdown.org>.



[https://bookdown.org/](https://bookdown.org)

# Presentation Ninja

with xaringan

Yihui Xie

RStudio, PBC

<https://slides.yihui.org/xaringan/>

# Parameterised reports

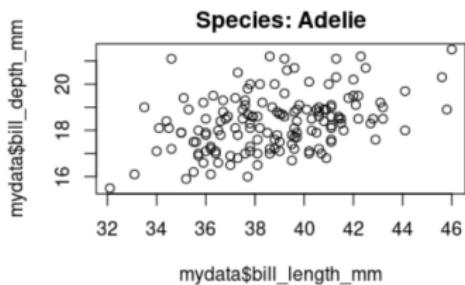
```

```

```
title: "My template report"
output: html_document
params:
 sp: Adelie

```

```
[{"r": true, "text": "library(palmerpenguins)\nlibrary(ggplot2)\n\nmydata <- subset(penguins, species == params$sp)\n\nplot(mydata$bill_length_mm, mydata$bill_depth_mm,\n main = paste0(\"Species: \" , params$sp))\n"}]
```



## Render thousands of individual reports from Rmd template

```
library('rmarkdown')

for (i in unique(penguins$species)) {

 render("template_report.Rmd",
 params = list(sp = i))

}
```

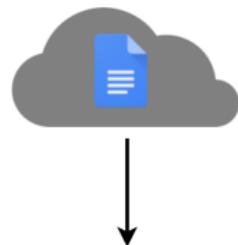
# Collaborative writing

- GitHub, GitLab, etc
- Google Docs ([trackdown](#))
- [redoc](#)

Locally



Google Docs



Share link with collaborators



Integrate changes

## Rmarkdown resources

---

# Rmarkdown website

<http://rmarkdown.rstudio.com/>

## R Markdown

from RStudio



The image shows a collage of screenshots from RStudio, illustrating the capabilities of R Markdown. It includes a scatter plot with a color gradient, a histogram, a line graph, and several code snippets in R. One snippet shows how to generate a density plot:

```
a = approx(x, y, n=1000, add=T)
ggplot(a, aes(x=x, y=y)) + geom_point() + theme_minimal()
```

**More Examples:**

The rest of this document consists of a few test cases to show you everything that works well in (slightly more complicated scenarios). First we generate two plots in one figure environment with the check option `fig.show = "check"`:

```
a = approx(x, y, n=1000, add=T)
ggplot(a, aes(x=x, y=y)) + geom_point() + theme_minimal()
b = ggplot(mtcars, aes(wt, mpg, color = factor(cyl))) +
 geom_point() + theme_minimal()
```

Get Started    Gallery    Formats    Articles    

Analyze. Share. Reproduce.

Your data tells a story. Tell it with R Markdown.  
Turn your analyses into high quality documents,  
reports, presentations and dashboards.

# Rmarkdown cheat sheet

**R Markdown Cheat Sheet**  
learn more at [rmarkdown.rstudio.com](http://rmarkdown.rstudio.com)



## Workflow

1 Open a new Rmd file by File > New File > R Markdown template  
2 Write document by editing template  
3 Knit document to create report Use knit button or render() to knit  
4 Preview Output in IDE window  
5 Publish (optional) Go web or server. Sync publish button to accounts of GitHub, RHub, Connect, RStudio Cloud, Find in document, File path to output document  
6 Examine building log in R Markdown console  
7 Use output file that is saved alongside Rmd file  
8 Render code with `render()`  
9 Run code chunks with `knit()` or `render()`  
10 Run all previous chunks with `knit_all()`  
11 Run current chunk with `knit_current()`

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents.

**Summary (RMD)**

```

title: "R Markdown"
author: "RStudio Team"
date: "2015-07-09"
output: html_document

This is an R Markdown document.
Markdown is a simple formatting syntax for authoring
HTML, PDF, and MS Word documents.

It uses conventional text formating
conventions and provides the ability to
embed R code and results in the document.

You can use the code editor to edit this
document directly, or you can edit it
in RStudio and then click the Knit button
in the toolbar above to render the results
of the code down below.

You will see the location of the
rendered document in the working directory

Text
#-----
Narration framework with
mathematical notation.

Code chunks
#-----
Chunks of embedded
code. Each chunk
• Begins with `##` or
ends with `###`.
If R Markdown will run
the code in a chunk,
it will use the location of
the chunk to run the
working directory

#-----
Inline code
#-----
Insert with r code
Built with 3.2.3
#-----
Important chunk options
#-----
cache - cache results for future
knits (default = FALSE)
cache.path - directory to save cached
results (default = "cache")
child - file to knit and then include
(default = NULL)
collapse - collapse all output into single
block (default = FALSE)
comment - prefix for each line of results
(default = "#")
#-----
Options not listed above: R options, progress, quiet, background, cache commmons, cache.lazy, cache.reuse, doc.opts, diag, engine.opts, engine.path, fig.wid, fig.height, fig.width, fig.cap, fig.pch, fig.props, fig.title, fig.scale, fig.show, fig.size, fig.censor, fig.interv, out.extra, out.height, out.width, pretty, plot, rbind, rcontext, rsize, rlog, rpath
```

For more details on using R Markdown see <http://rmarkdown.rstudio.com>

## Embed code with knitr syntax

Inline code  
Insert with `r code`  
Built with 3.2.3  
Cache results with `r getresults()`  
-----  
Important chunk options  
cache - cache results for future knits (default = FALSE)  
cache.path - directory to save cached results (default = "cache")  
child - file to knit and then include (default = NULL)  
collapse - collapse all output into single block (default = FALSE)  
comment - prefix for each line of results (default = "#")  
Options not listed above: R options, progress, quiet, background, cache commmons, cache.lazy, cache.reuse, doc.opts, diag, engine.opts, engine.path, fig.wid, fig.height, fig.width, fig.cap, fig.pch, fig.props, fig.title, fig.scale, fig.show, fig.size, fig.censor, fig.interv, out.extra, out.height, out.width, pretty, plot, rbind, rcontext, rsize, rlog, rpath

## Code chunks

One or more lines starting with `##` and followed by one or more blank lines (or `##` and `##` followed by one or more blank lines). Place chunk options within the brace, after `##` and before `##`.  
-----  
depends - chunk dependencies for caching (default = NULL)  
echo - print code in output document (default = TRUE)  
engine - code language used in chunk (default = "R")  
error - display error messages in document (default = TRUE)  
eval - run code in chunk (default = TRUE)  
fig.height - height of figure in inches (default = 5)  
highlight - highlight source code (default = TRUE)  
include - include chunk in document after rendering (default = TRUE)  
message - display code messages in document (default = TRUE)  
results - default = "markup"  
results - "asis" - print raw code  
"hide" - do not display results  
"hold" - put all results below all code  
tidy - tidy code for display (default = TRUE)  
warning - display code warnings in document (default = TRUE)

## Global options

Set with `knitr::opts_chunk$set()`, e.g.  
`knitr::opts_chunk$set(cache = TRUE)`

## Parameters

Parameterize your documents to reuse with different inputs (e.g., data sets, values, etc.)

### Add parameters

Create and set the parameters in the `params` argument of `param`.

### Call parameters

Call parameter `param` in code as `param`.

### Set parameters

Set values with `Knit with parameters` or the `params` argument of `render`.

```
param: {
 n: 100
 d: "sys.argv[1]"
}

Call parameter:
param
n: 100
d: "sys.argv[1]"

Set parameters:
Knit with parameters:
param: {
 n: 100
 d: "sys.argv[1]"
}
params: {
 n: 100
 d: "sys.argv[1]"
}

render("doc.Rmd",
 params: {
 n: 1,
 d: as.Date("2015-01-01")
})
```

[https://www.rstudio.org/links/r\\_markdown\\_cheat\\_sheet](https://www.rstudio.org/links/r_markdown_cheat_sheet)

## Rmarkdown reference guide

<h1>R Markdown</h1>	<h2>Reference Guide</h2>	
Learn more about R Markdown at <a href="#">markdown.rstudio.com</a>		
Learn more about Interactive Docs at <a href="#">shiny.rstudio.com/articles</a>		
<b>Contents:</b>		
1. <a href="#">Markdown Syntax</a>		
<b>2. Knitr chunk options</b>		
3. <a href="#">Pandoc options</a>		
<h3>Syntax</h3>	<h3>Becomes</h3>	
<p>Make a code chunk with three back ticks followed by an <code>r</code> in braces. End the chunk with three back ticks:</p> <pre>```{r} paste("Hello", "World!") ```</pre>	<p>Make a code chunk with three back ticks followed by an <code>r</code> in braces. End the chunk with three back ticks:</p> <pre>parse("Hello", "World!")  # [1] "Hello World"</pre>	
<p>Place code inline with a single back ticks. The first back tick must be followed by an <code>R</code>, like this <code>r` paste("Hello", "World!")`</code>:</p>	<p>Place code inline with a single back ticks. The first back tick must be followed by an <code>R</code>, like this <code>r` Hello World`</code>:</p>	
<p>Add chunk options within braces. For example, <code>echo=FALSE</code> will prevent source code from being displayed:</p> <pre>## {r eval=TRUE, echo=FALSE} paste("Hello", "World")</pre>	<p>Add chunk options within braces. For example, <code>echo=FALSE</code> will prevent source code from being displayed:</p> <pre>## [1] "Hello World"</pre>	
<p>Learn more about chunk options at <a href="#">http://yihui.name/knitr/options</a></p>		
<h3>Chunk options</h3>		
option	default value	description
<b>Code execution</b>		
<code>child</code>	<code>NULL</code>	A character vector of filenames. Knitr will kill the files and place them into the main document.
<code>code</code>	<code>NULL</code>	Set to a file. Knitr will replace the code in the chunk with the code in the code option.
<code>engine</code>	<code>"R"</code>	Knitr will evaluate the chunk in the named language, e.g. <code>engine = "python"</code> . <code>run.names(knit\$left_engine\$get())</code> to see supported languages.
<code>eval</code>	<code>TRUE</code>	If <code>FALSE</code> , Knitr will not run the code in the code chunk.
<code>include</code>	<code>TRUE</code>	If <code>FALSE</code> , Knitr will not include the chunk but will include the chunk in the final document.
<code>part</code>	<code>TRUE</code>	If <code>FALSE</code> , Knitr will not include the chunk when running <code>knit2html</code> to extract the source code.
<b>Results</b>		
<code>collapse</code>	<code>FALSE</code>	If <code>TRUE</code> , knitr will collapse all the source and output blocks created by the chunk into a single block.
<code>echo</code>	<code>TRUE</code>	If <code>FALSE</code> , knitr will not display the code in the code block above it's results in the final document.
<code>results</code>	<code>"markup"</code>	If <code>"asis"</code> , knitr will not display the code's results in the final document. If <code>"asis"</code> , knitr will wait through results without reformatting them (useful if results return raw HTML, etc.)
<code>error</code>	<code>TRUE</code>	If <code>FALSE</code> , knitr will not display any error messages generated by the code.
<code>message</code>	<code>TRUE</code>	If <code>FALSE</code> , knitr will not display any messages generated by the code.
<code>warning</code>	<code>TRUE</code>	If <code>FALSE</code> , knitr will not display any warning messages generated by the code.
<b>Code styling</b>		
<code>background</code>	<code>"#FFFFFF"</code>	A background color for chunks in LaTeX output.
<code>comment</code>	<code>"%"</code>	A character string. Knitr will append the string to the start of each line of results in the final document.
<code>highlight</code>	<code>TRUE</code>	If <code>TRUE</code> , knitr will highlight the source code in the final output.
<code>prompt</code>	<code>FALSE</code>	If <code>TRUE</code> , knitr will add it to the start of each line of code displayed in the final document.
<code>size</code>	<code>"normalize"</code>	Fontsize for LaTeX output.
<code>strip.white</code>	<code>TRUE</code>	If <code>TRUE</code> , knitr will remove white spaces that appear at the beginning or end of a code chunk.
<code>tidy</code>	<code>FALSE</code>	If <code>TRUE</code> , knitr will tidy code chunks for display with the <code>tidy_source</code> function in the <code>knitr</code> package.
 Studio	Updated 10/30/2014	© 2014 RStudio, Inc. All rights reserved.

[https://www.rstudio.org/links/r\\_markdown\\_reference\\_guide](https://www.rstudio.org/links/r_markdown_reference_guide)

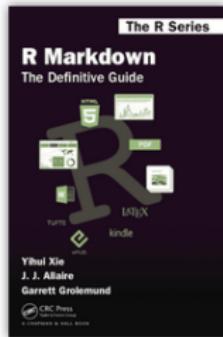
## R Markdown: The Definitive Guide

by Yihui Xie, J. J. Allaire, Garrett Grolemund

2018-09-11

Star

239



The first official book authored by the core R Markdown developers that provides a comprehensive and accurate reference to the R Markdown ecosystem. With R Markdown, you can easily create reproducible data analysis reports, presentations, dashboards, interactive applications, books, dissertations, websites, and journal articles, while enjoying the simplicity of Markdown and the great power of R and other languages. *Read more →*

<https://bookdown.org/yihui/rmarkdown/>

<https://bookdown.org/yihui/rmarkdown-cookbook/>

# Welcome to Quarto

Quarto® is an open-source scientific and technical publishing system built on [Pandoc](#)

- Create dynamic content with [Python](#), [R](#), [Julia](#), and [Observable](#).
- Author documents as plain text markdown or [Jupyter](#) notebooks.
- Publish high-quality articles, reports, presentations, websites, blogs, and books in [HTML](#), [PDF](#), [MS Word](#), [ePub](#), and more.
- Author with scientific markdown, including equations, citations, crossrefs, figure panels, callouts, advanced layout, and more.

<https://quarto.org/>

Your turn

---

## Your turn

- Try visual markdown editor
- Add bibliography
- Try templates (rticles, rmdTemplates)
- Parameterised reports (e.g. different iris or penguin species)

## Workflow management

---

In complex projects we must **keep pieces organised**

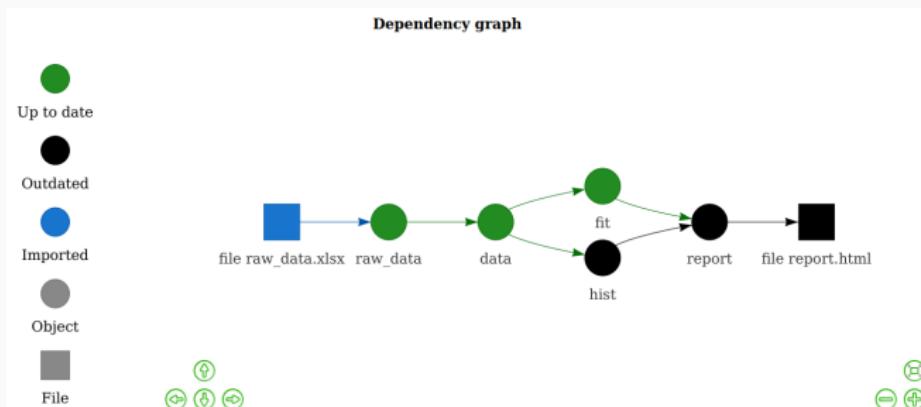


`makefile` runs all code in right order

`makefile.R`

```
source("clean_data.R")
source("fit_model.R")
render("report.Rmd")
```

# targets: advanced workflow management



<https://docs.ropensci.org/targets/>

Your turn

---

Write makefile.R for your project

Try `targets` minimal example

<https://github.com/wlandau/targets-minimal>

## Controlling software dependencies

---



Package changes can break your analysis

How to reproduce your analysis in a year,  
or different computer?

# sessionInfo records OS & used packages

```
sessionInfo()
```

```
R version 4.2.0 (2022-04-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.4 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
```

```
locale:
[1] LC_CTYPE=en_GB.UTF-8 LC_NUMERIC=C
[3] LC_TIME=es_ES.UTF-8 LC_COLLATE=en_GB.UTF-8
[5] LC_MONETARY=es_ES.UTF-8 LC_MESSAGES=en_GB.UTF-8
[7] LC_PAPER=es_ES.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] stats graphics grDevices utils datasets methods base
```

```
other attached packages:
[1] report_0.5.5 equatiomatic_0.3.1 xtable_1.8-4 knitr_1.40
```

```
loaded via a namespace (and not attached):
[1] Rcpp_1.0.9 mvtnorm_1.1-3 lattice_0.20-45 tidyR_1.2.0
[5] zoo_1.8-10 assertthat_0.2.1 digest_0.6.29 utf8_1.2.2
[9] mime_0.12 R6_2.5.1 backports_1.4.1 evaluate_0.16
[13] coda_0.19-4 pillar_1.8.1 rlang_1.0.5 multcomp_1.4-20
[17] performance_0.9.2 rstudioapi_0.14 Matrix_1.4-1 effectsSize_0.7.0.5
[21] rmarkdown_2.16 splines_4.2.0 stringr_1.4.1 shiny_1.7.2
[25] broom_1.0.1 compiler_4.2.0 httpuv_1.6.5 xfun_0.32
[29] pkgconfig_2.0.3 parameters_0.18.2 htmltools_0.5.3 insight_0.18.2
[33] tidyselect_1.1.2 tibble_3.1.8 codetools_0.2-18 fansi_1.0.3
[37] lubridate_1.8.0 dplyr_1.0.8 MASS_7.3-56 grid_4.2.3
[41] R6_2.5.1 purrr_0.3.4 Rcpp_1.0.9 magrit_2.0.1
[45] assertthat_0.2.1 digest_0.6.29 backports_1.4.1 evaluate_0.16
[49] R6_2.5.1 pillar_1.8.1 rlang_1.0.5 multcomp_1.4-20
[53] purrr_0.3.4 dplyr_1.0.8 MASS_7.3-56 grid_4.2.3
[57] R6_2.5.1 purrr_0.3.4 assertthat_0.2.1 digest_0.6.29
[61] pillar_1.8.1 rlang_1.0.5 backports_1.4.1 evaluate_0.16
[65] dplyr_1.0.8 R6_2.5.1 pillar_1.8.1 rlang_1.0.5
[69] assertthat_0.2.1 digest_0.6.29 backports_1.4.1 evaluate_0.16
[73] R6_2.5.1 pillar_1.8.1 rlang_1.0.5 multcomp_1.4-20
[77] purrr_0.3.4 dplyr_1.0.8 MASS_7.3-56 grid_4.2.3
[81] R6_2.5.1 purrr_0.3.4 assertthat_0.2.1 digest_0.6.29
[85] pillar_1.8.1 rlang_1.0.5 backports_1.4.1 evaluate_0.16
[89] dplyr_1.0.8 R6_2.5.1 pillar_1.8.1 rlang_1.0.5
[93] assertthat_0.2.1 digest_0.6.29 backports_1.4.1 evaluate_0.16
```

## checkpoint reconstructs packages in given date

```
library('checkpoint')

checkpoint("2019-10-08")

source("analysis.R")
```

1. Detects packages used
2. Installs version from given date (only CRAN)
3. Independent install (not messing w/ main library)

## automagic records & install packages (CRAN + GitHub)

```
automagic::make_deps_file()
```

File `deps.yaml` records dependencies:

```
- Package: equatiomatic
 Repository: CRAN
 Version: 0.1.0

- Package: report
 GithubUsername: easystats
 GithubRepo: report
 GithubRef: HEAD
 GithubSHA1: c48a4bb0a40df7116bc502aa3ce2cbbc9d70b7e2
```

To install all those dependencies:

```
automagic()
```

## renv also controls dependencies

```
renv::init()
Create private package library for project

renv::snapshot()
Capture dependencies in lockfile

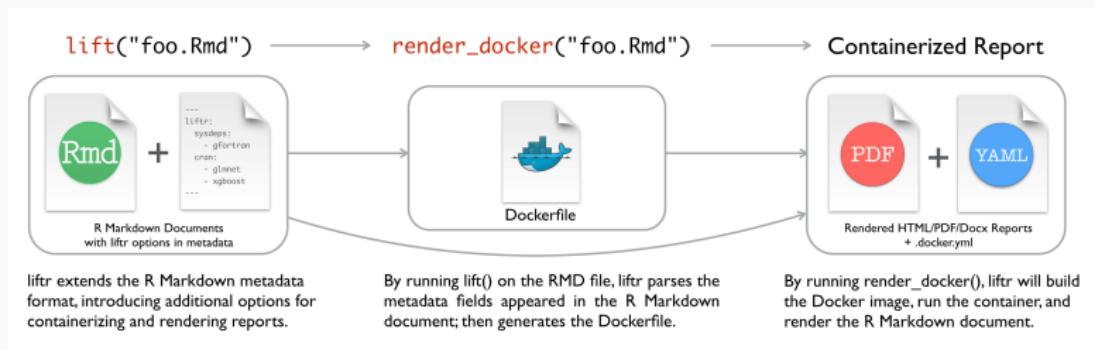
renv::restore()
Regenerate dependencies from lockfile
```

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

To ensure reproducibility,  
besides R packages  
we also need to control  
**computational environment**

Docker recreates virtual systems  
from a Dockerfile

# liftr: process Rmd in Docker container



<https://liftr.me/>

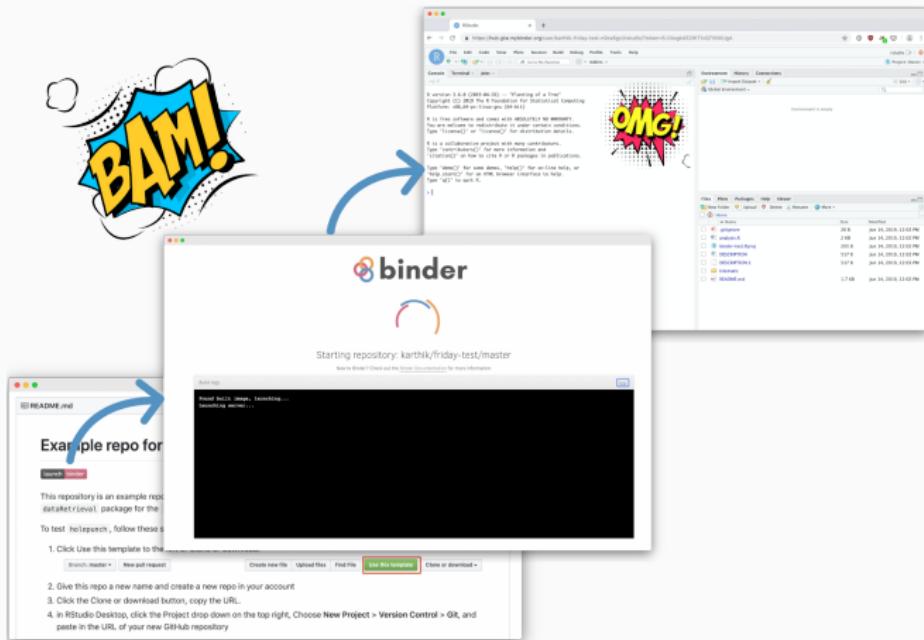
## containerit creates Dockerfile

```
library("containerit")

dockfile <- dockerfile(from = "mypaper.Rmd")
```

<https://o2r.info/containerit>

# holepunch: reproduce analysis in the cloud (Binder)



<https://karthik.github.io/holepunch/>

Your turn

---

## Checkpoint

- Create script/Rmd using different packages
- Call **checkpoint** on former date

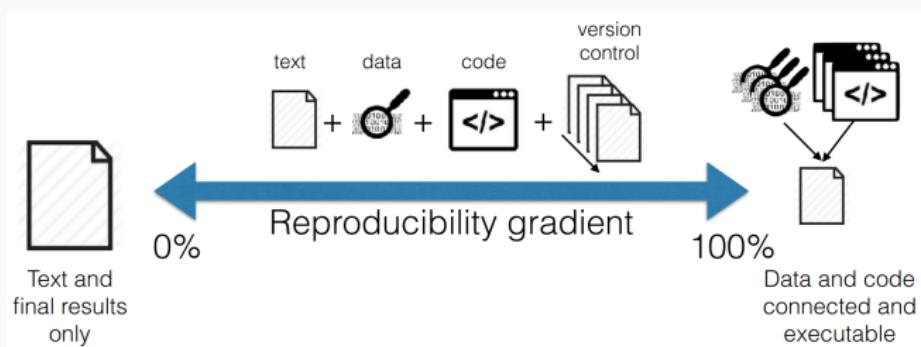
- Record dependencies:
  - `automagic::make_deps_file()`
  - `renv::snapshot`
- Recreate packages
  - `automagic()`
  - `restore()`

# How to write more reproducible code

- BES guide to reproducible code
- Turing Way
- Good enough practices in scientific computing
- Ciencia reproducible: qué, por qué, cómo
- <https://rstats.wtf>
- `fertile` package
- CodeCheck

# Reproducibility

- Good for you, good for science
- Requires systemic changes
- Reproducibility gradient: step by step



Happy collaboration!

