

rOpenSci Community Manager Joins Her Community

Learning R by Listening

Stefanie Butland

rOpenSci Community Manager

@stefaniebutland

@rOpenSci

May 22, 2019 - R-Ladies Seattle



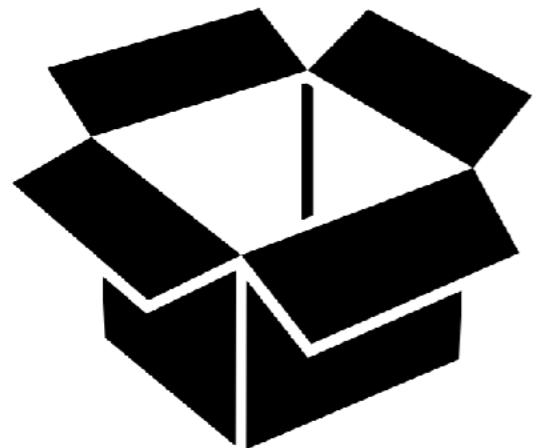
I contain multitudes

- Walt Whitman h/t Ed Yong



rOpenSci

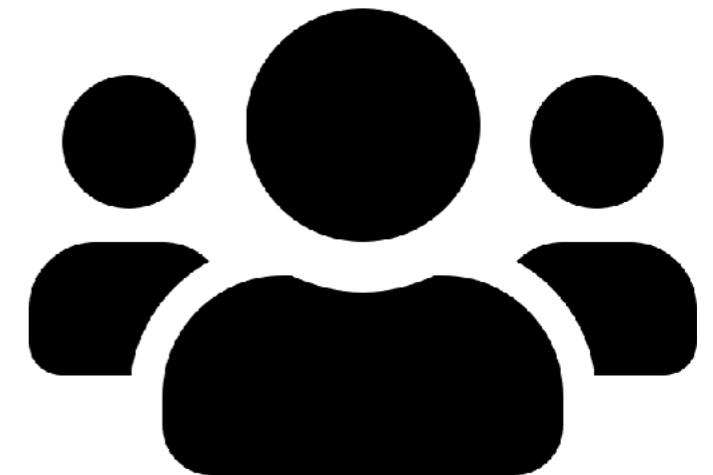
technical & social infrastructure



Software



Peer review



Community

rOpenSci by the numbers

5 staff

1 postdoc

2 Bioconductor pkgs

192 CRAN pkgs

287 total pkgs

~ 500 code contributors

> 500 citations

LOTS! of community members

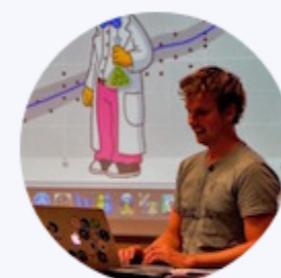
1 awesome open system for peer review of software



Karthik Ram



Scott Chamberlain



Jeroen Ooms



Stefanie Butland



Maëlle Salmon



Dan Sholler

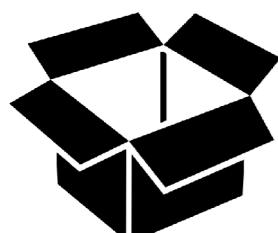
ropensci.org/packages/

FILTERS

- All
 - Altmetrics
 - Data Publication
 - Tools
 - Visualization
 - Databases
 - Geospatial
 -
- Web
 - Images Processing
 - Literature
 - Computing Infrastructure
 - Security
 - Taxonomy
 -
- CRAN / BIOC available
 - Active repos

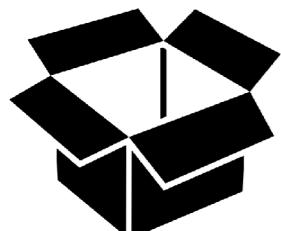
x

	NAME	MAINTAINER	DESCRIPTION	DETAILS	
	monkeylearn	Maëlle Salmon	Accesses the Monkeylearn API for Text Classifiers and Extractors	CRAN	
	opencage	Maëlle Salmon	Interface to the OpenCage API	CRAN	
	riem	Maëlle Salmon	Accesses Weather Data from the Iowa Environment Mesonet	CRAN	
	ropenaq	Maëlle Salmon	Accesses Air Quality Data from the Open Data Platform OpenAQ	CRAN	
	rtimecropem	Maëlle Salmon	Supports the Analysis of RTI MicroPEM Output Files	CRAN	



4 of 6 new maintainers are women!

- [RSelenium](#): now maintained by Ju Kim
- [chromer](#): now maintained by Paula Andrea
- [qualtRics](#): now maintained by Julia Silge
- [rsnps](#): now maintained by Julia Gustavsen and Sina Rüeger
- [rdpla](#): now maintained by Alyssa Columbus
- [webchem](#): now maintained by Erik Sapper



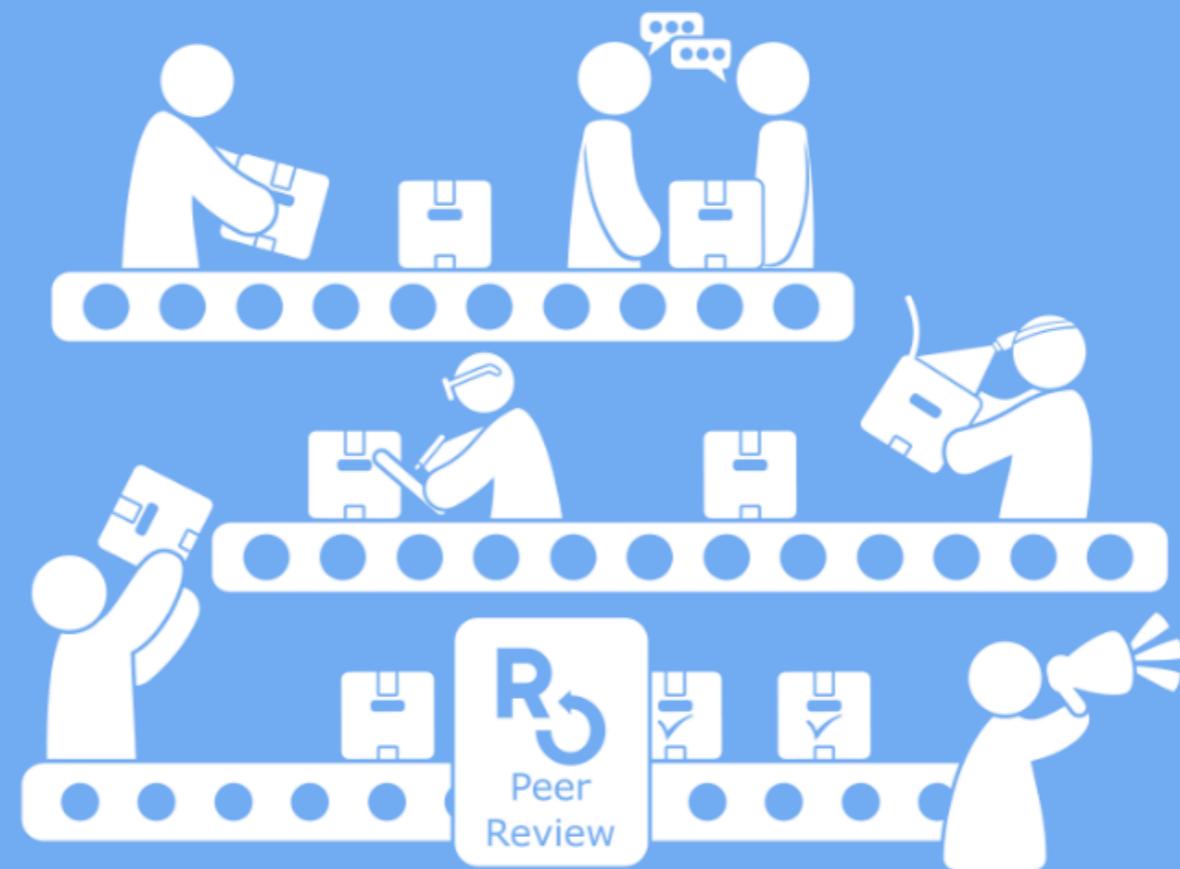


academic peer review of
research publications?

ropensci.github.io/dev_guide/

rOpenSci Packages

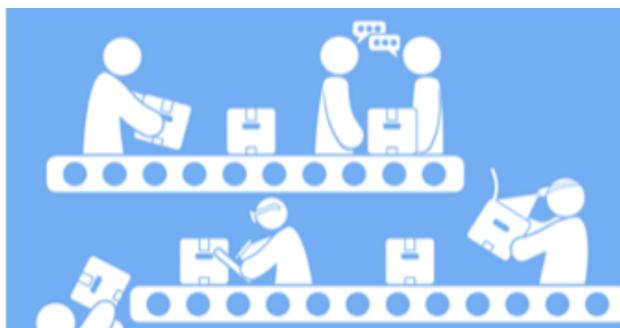
*Development, Maintenance,
and Peer Review*



Software peer review



July 2018



 **treestartR v0.1.0**  Reference

 build passing  codecov 65%  Peer Reviewed

treeStartR

treeStartR is an R package for automated generation of starting trees for total-evidence phylogenetic analyses. Phylogenetic trees, and particularly time-scaled phylogenetic trees, are increasingly estimated using parameter-rich models of evolution (Lartillot, Lepage, and Blanquart 2009; Wright, Lloyd, and Hillis 2016) and models incorporating macroevolutionary processes (Heath, Huelsenbeck, and Stadler 2014). Finding a starting tree with a computable likelihood to perform a Bayesian MCMC under these complex models can be a challenge, particularly when estimation involves many taxa, large datasets, and missing data. Different phylogenetic estimation packages allow users to find starting trees in different ways, such as estimating a tree under parsimony (Stamatakis (2014)) or neighbor-joining (Bouckaert et al. 2014), randomly adding taxa to the tree (Bouckaert et al. 2014; Stamatakis 2014), or allowing the user to specify the tree. Addition of taxa is usually performed based on data - that is, using an algorithm (such as parsimony or neighbor-joining) to add the tips to the tree. However, in analyses of the fossil record, specimens may be included for which there are no molecular or morphological characters available, but the taxonomy of the

Links

Download from CRAN at <https://cloud.r-project.org/package=treestartR>

Report a bug at <https://github.com/ropensci/treeStartR/issues>

License

MIT + file [LICENSE](#)

Developers

April Wright
Author, maintainer

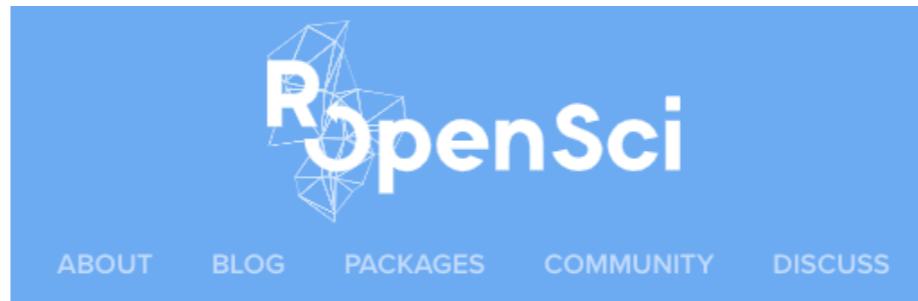


v 2018

Software peer review



July 2018



Generating reasonable starting trees for complex phylogenetic analyses

April Wright

December 11, 2018

I never really thought I would write an R package. I use R pretty casually. Then, this year, I was invited to participate during the last week of the [Analytical Paleobiology short course](#), an intensive month-long experience in quantitative paleontology. I was thrilled to be invited. But I got a slight sinking feeling in my stomach when I realized all the materials were in R.

And so I, a Pythonista, decided I would spend some of my maternity leave writing R packages to try to blend in with students who had spent the month living and breathing R.



Nov 2018

Review for us:

ropensci.org/onboarding

“As a user, you are qualified as a potential package reviewer if you have some appreciation for what makes your favourite packages useful”

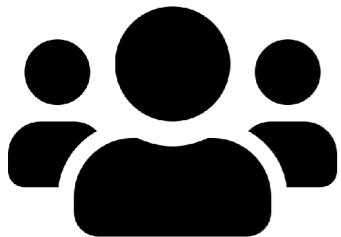
- Miles McBain



Experience in your midst!

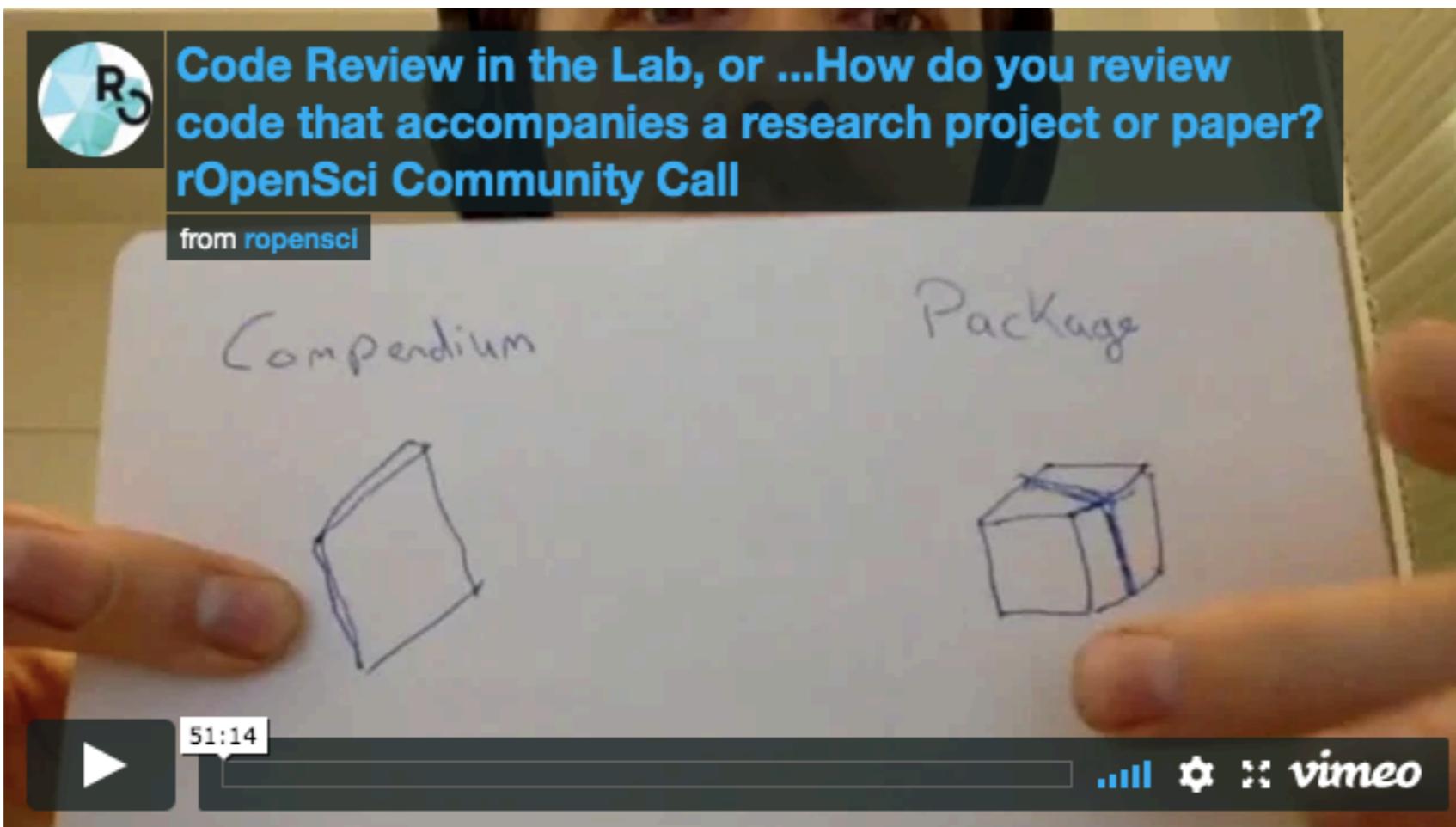


- attended an rOpenSci unconf
- contributed a blog post
- maintain an rOpenSci package
- reviewed
- used
- commented in a forum thread
- attended a Community Call
- read a blog post
- ❤️'d or RT'd a tweet



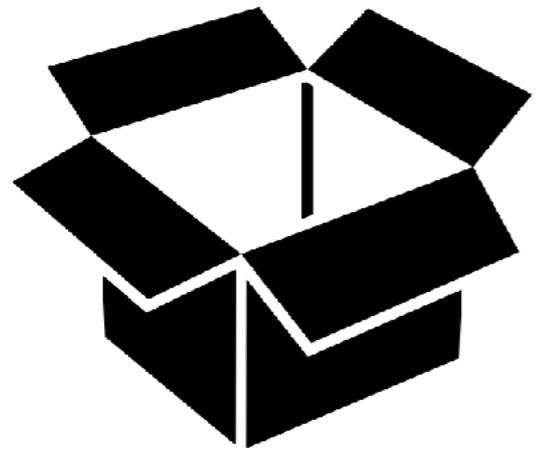
Code review in the lab?

Community Call video, discussion, summary



<https://ropensci.org/commcalls/2018-10-16/>

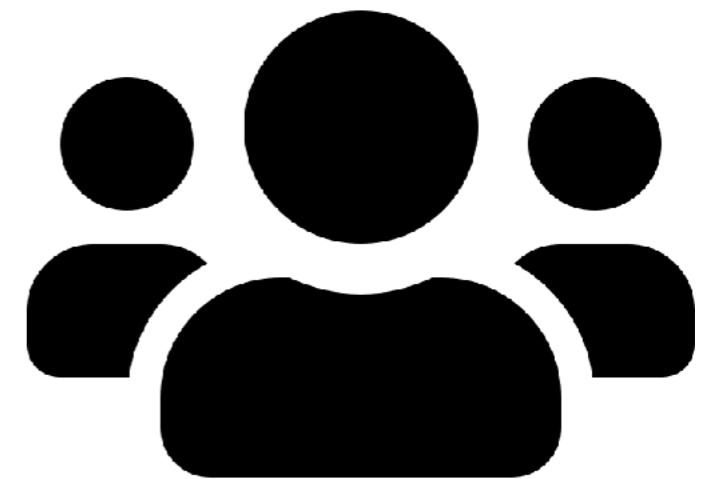
<https://ropensci.org/>



/packages



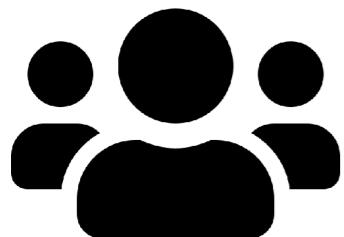
/software-review



/community

I taught myself some R

This is about me ...
but really it's about you



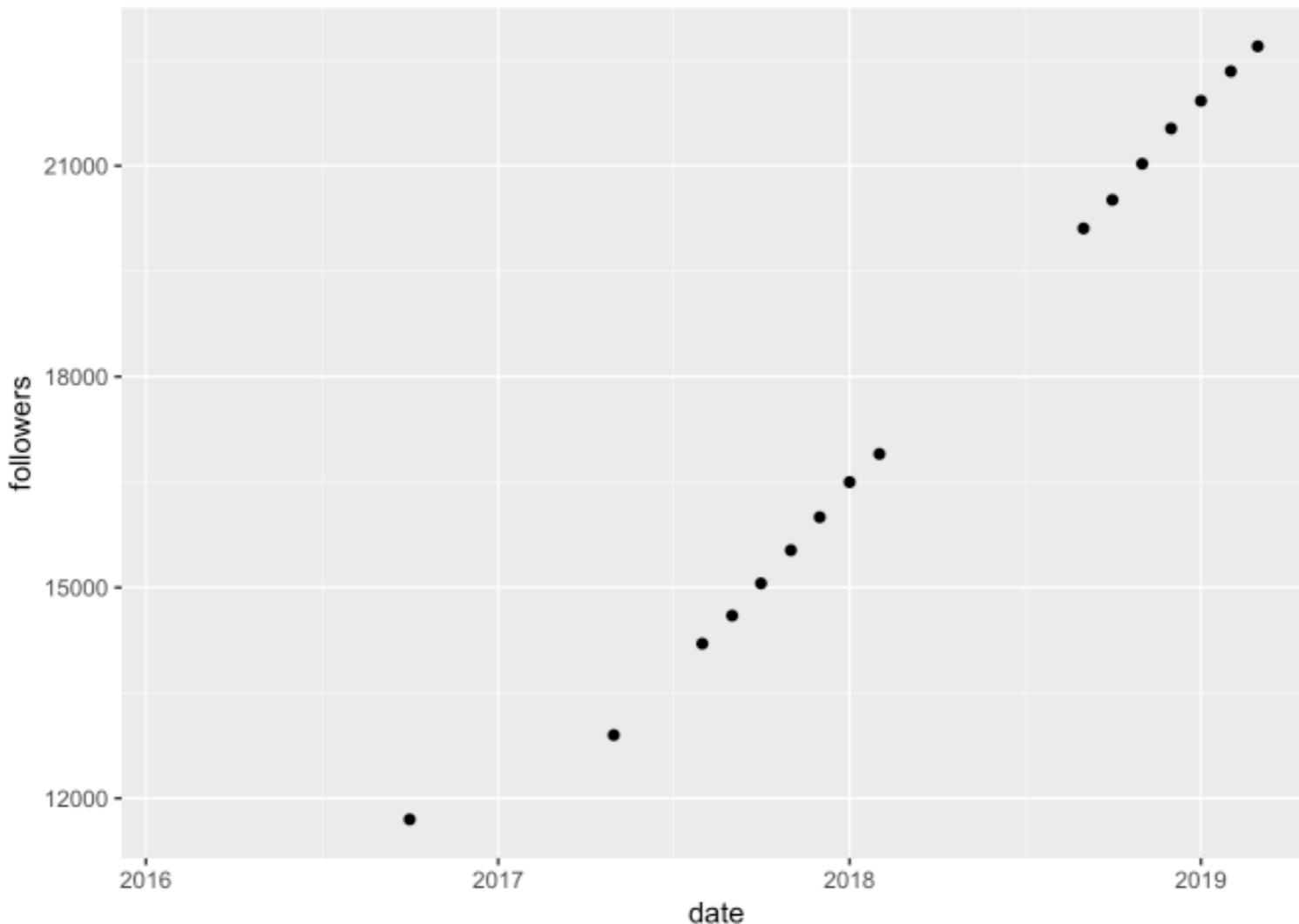


Between January 10th and March 7th, I made an RMarkdown report with the **plots** I wanted to make...

This is what I came for

```
# use ggplot2 to make point or line plots of date vs tweets, mentions, new followers etc  
# followers column is the only one that is not 100% objective, since it is not archived by twitter analytics the way mentions etc are  
ggplot(data = metrics, aes(x=date, y=followers)) +  
  geom_point()
```

```
## Warning: Removed 22 rows containing missing values (geom_point).
```



Number of rOpenSci followers on Twitter has increased steadily since September 2016.

...from a csv file with some data I collected and whose **date column** is formatted correctly, where I understand the **provenance** of the data...

Read csv of Twitter data

```
# load data from csv -----
# careful: read_csv is tidyverse; read.csv is not
# metrics is an "object"

# > here::here()
# [1] "/Users/stefaniebutland/GitHub/stefaniebutland/twitter_time"

# I learned the hard way that lubridate::here() masked here::here() so need
# package to avoid errors

metrics <- read_csv(here::here("data", "2019-03-02_twitter_metrics_monthly_rOpenSci.csv"))
```

JAN 2019 SUMMARY

Tweets	48	Tweet impressions	23.3K
--------	----	-------------------	-------

Profile visits	549	Mentions	64
----------------	-----	----------	----

New followers	29
---------------	----

Display table of Twitter data

```
knitr::kable(metrics, caption = "Twitter metrics table")
```

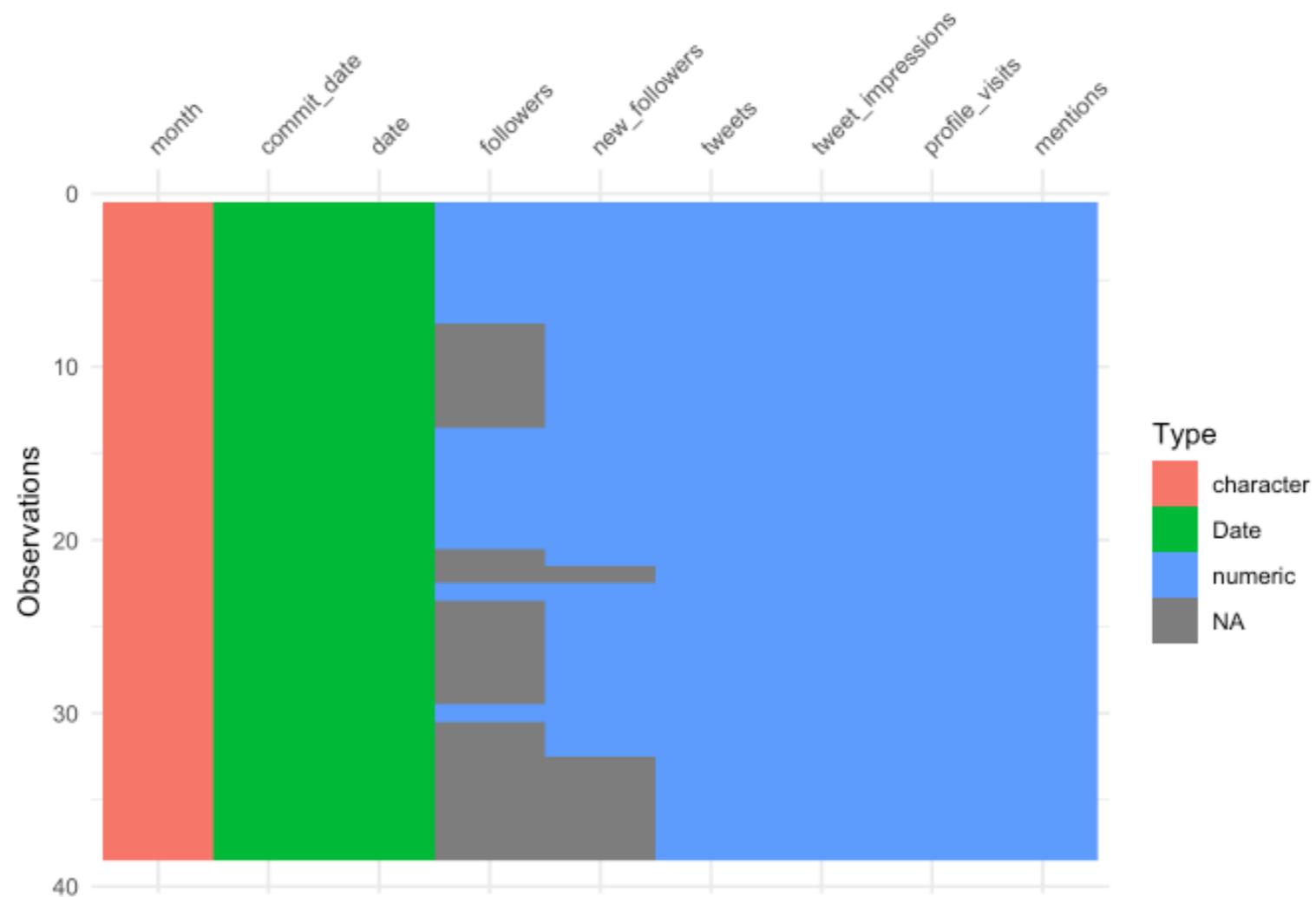
Twitter metrics table

commit_date	date	month	followers	new_followers	tweets	tweet_impressions	profile_visits	mentions
2019-03-01	2019-03-01	Feb	22701	351	20	205000	2327	348
2019-02-01	2019-02-01	Jan	22345	405	26	217000	3603	292
2019-01-02	2019-01-01	Dec	21925	333	28	211000	2369	212
2018-12-02	2018-12-01	Nov	21531	476	38	244000	4104	497

...I tasted the magic of **the visdat package** and knew where my **missing data** was...

Try `visdat` magic

```
# use visdat package to visualize missing data
# this was the first package I tried after reading in csv, initially just for fun
vis_dat(metrics)
```



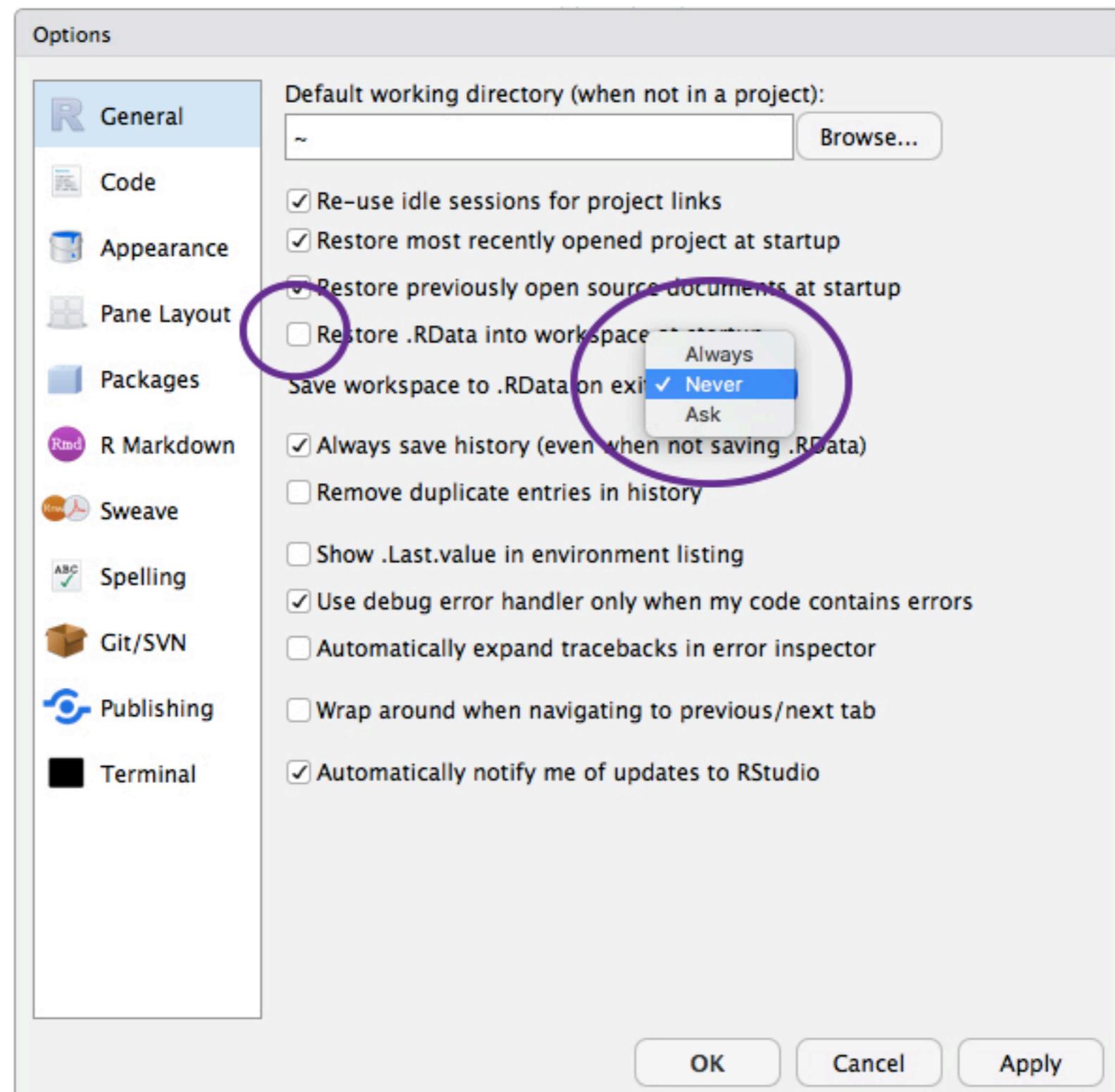
...and I'm working in RStudio connected to my own public GitHub repository with data and scripts in the right folders...

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the path `~/GitHub/stefaniebutland/twitter_time - master - RStudio`.
- File Explorer:** Displays the project structure:

Name	Type	Size	Last Modified
<code>.gitignore</code>	File	61 B	Apr 8, 2019
<code>.Rhistory</code>	File	10.7 KB	Apr 8, 2019
<code>data</code>	Folder		
<code>docs</code>	Folder		
<code>plots</code>	Folder		
<code>README.md</code>	File	61 B	Feb 7, 2019
<code>scripts</code>	Folder		
<code>twitter_time.Rproj</code>	File	205 B	May 13, 2019
- Code Editor:** Shows the `tweets_ropensci.R` script with the following code:32 ## Data Analysis
33 ### Load Packages
34 ```{r load pkgs, message=FALSE, collapse=TRUE}
35 # load packages -----
36 library(tidyverse)
37 library(knitr)
38 library(here)
39 library(useThis)
40 library(visdat)
41 library(lubridate)
42
43 useThis::use_github()
44 ...
45
46 ### Read csv of Twitter data
47 ```{r read_csv, echo=TRUE, message=FALSE, results='hide'}
48 # load data from csv -----
49 # careful: read_csv is tidyverse; read.csv is not
50 # metrics is an "object"
51
52 # > here::here()
53 # [1] "/Users/stefaniebutland/GitHub/stefaniebutland/twitter_time"
54
55 # I learned the hard way that lubridate::here() masked here::here() so need
to call specific function within the package to avoid errors
56
57 metrics <- read_csv(here::here("data",
"2019-03-02_twitter_metrics_monthly_rOpenSci.csv"))
58 ...
30:125 # Data Collection
R Markdown
- Environment:** Shows the `metrics` object in the Global Environment.
- Console:** Shows the command `# Data Collection`.

...never saving my RStudio workspace when I exit...



...using `here::here()` so I don't hardcode my local file paths so Jenny Bryan doesn't come set my laptop on fire. And once I used “masking the function” and “calling the function from the package” in a sentence!

The screenshot shows a Twitter conversation. On the left, Hadley Wickham (@hadleywickham) posts a warning about using absolute file paths in R scripts, which can lead to security issues. On the right, mara (@mara) responds to a thread about conflicts between different 'here' functions in the tidyverse and lubridate packages.

Hadley Wickham

@hadleywickham

Following

The only two things that make @JennyBryan . Instead use projects + here::here()
#rstats

If the first line of your R script is
`setwd("C:\Users\jenny\path\that\only\I\have")`
I* will come into your office and
SET YOUR COMPUTER ON FIRE .

* or maybe Timothée Poisot w

4:50 PM - 10 Dec 2017

mara Sustainer

Conflict between here in, uh, here, and here in lubridate?

tidyverse lubridate

Personally, I almost always use `here::here()`, even when I *don't* have lubridate attached. But, as jenny mentioned, you've described the situation as it is.

1 2018-03-02

305 Retweets 992 Likes



3 to my (beginner) R success

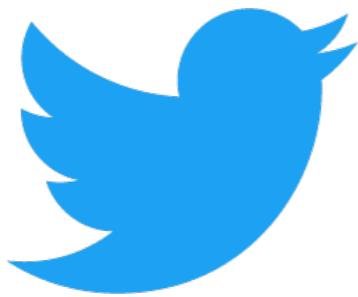
- in-person and online trust networks
(aka community)
- a specific modest project
- the “right” tutorials and resources
for me



In-person and online trust networks

Omigosh but this
community is everything

- Me

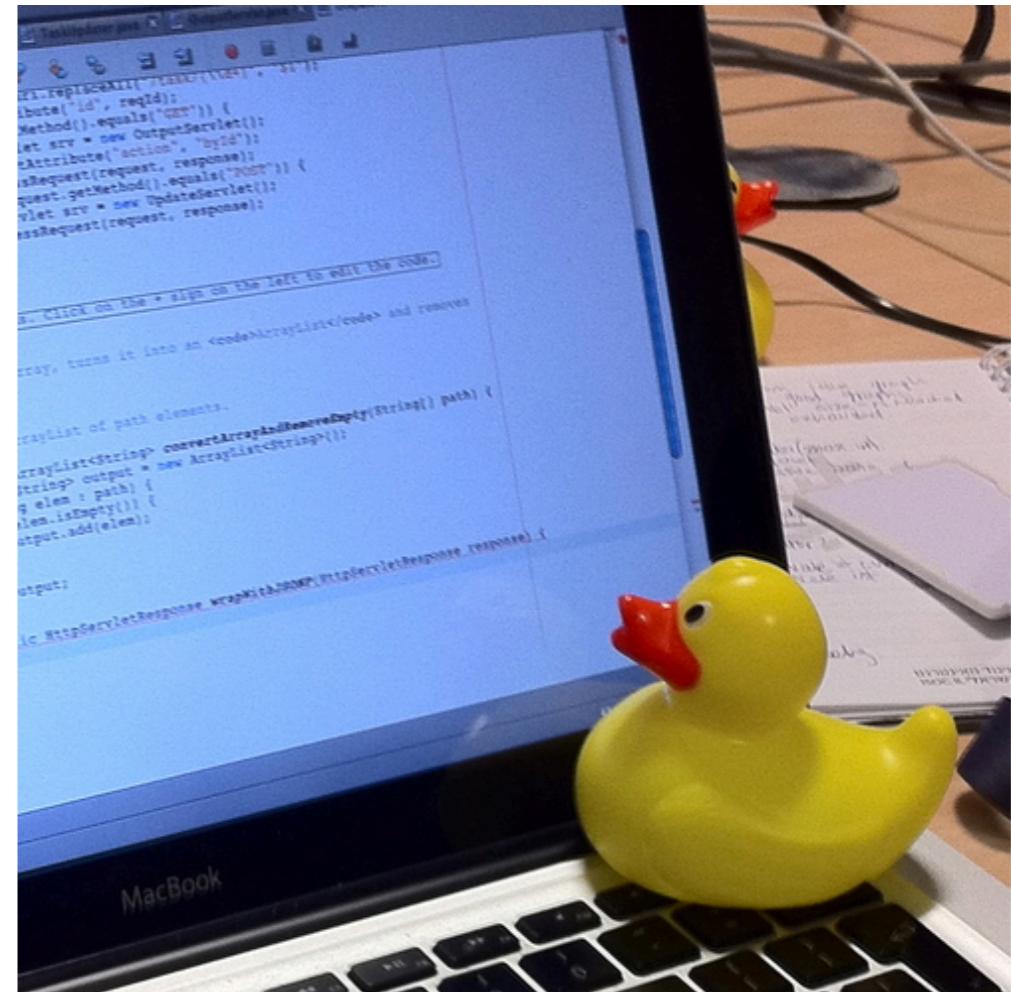


#rstats



Human rubber duck debuggers

"wait, my x-axis of dates sorted itself alphabetically?!"



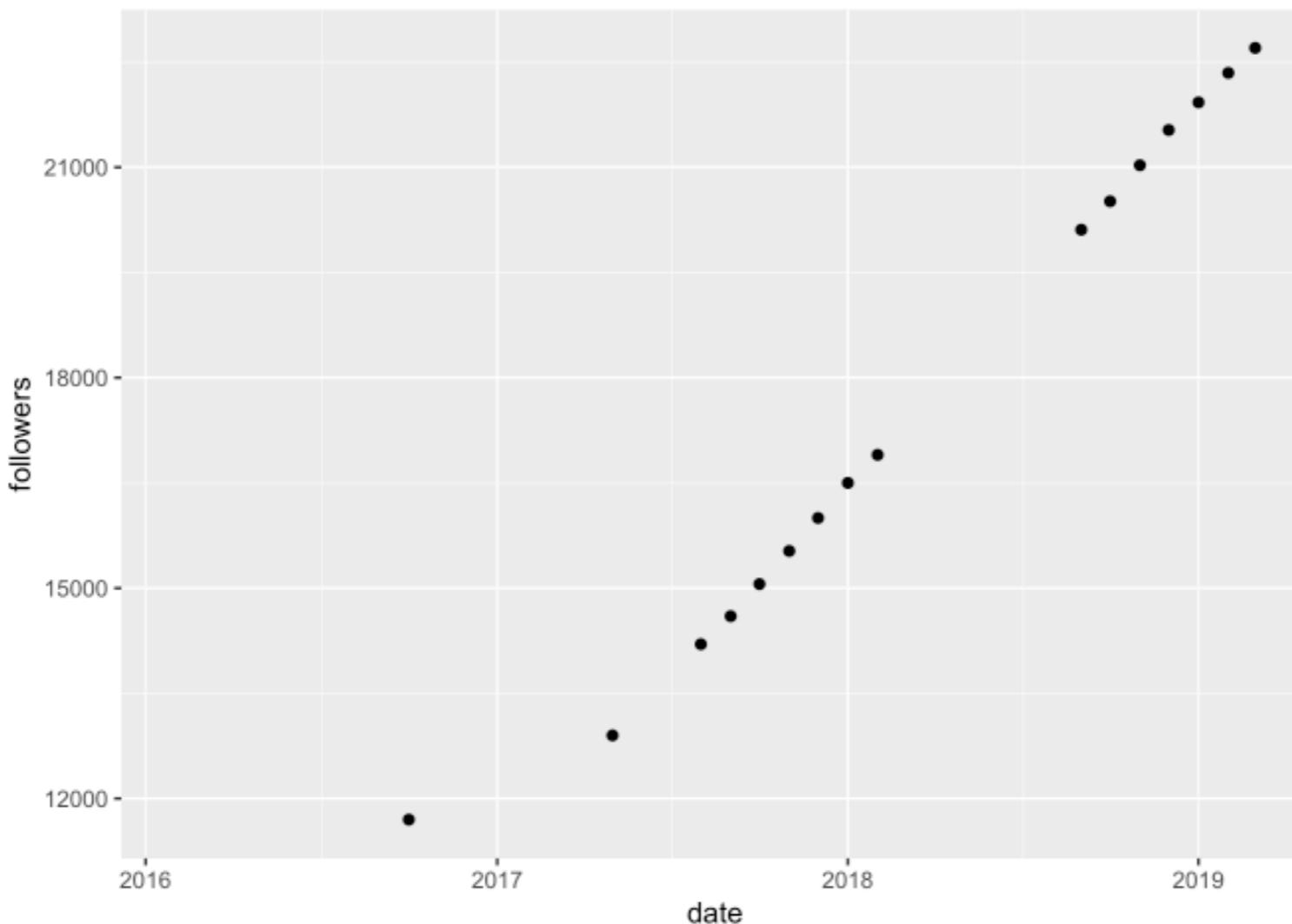


A specific modest project

This is what I came for

```
# use ggplot2 to make point or line plots of date vs tweets, mentions, new followers etc  
# followers column is the only one that is not 100% objective, since it is not archived by  
# twitter analytics the way mentions etc are  
ggplot(data = metrics, aes(x=date, y=followers)) +  
  geom_point()
```

```
## Warning: Removed 22 rows containing missing values (geom_point).
```



Number of rOpenSci followers on Twitter has increased steadily since September 2016.



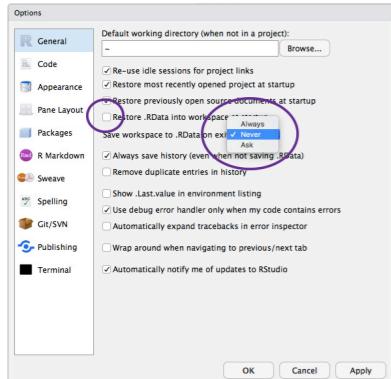
The “right” resources for me

designed by us, for us.
~~100~~ oriented towards R beginners
aims to provide a solid foundation of R skills

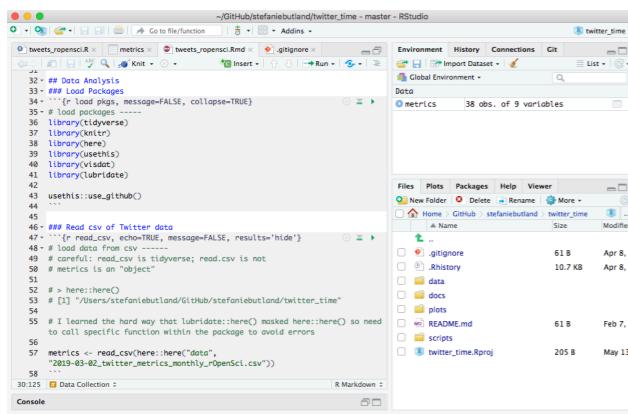


Jen Richman, Lisa Williams, Danielle Navarro
<https://rladiessydney.org/ryouwithme>

Where did you learn that?



never saving my
RStudio
workspace when I
exit



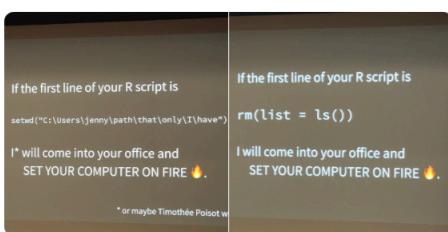
working in
RStudio with
data and scripts
in the right
folders

R-Ladies Sydney



Hadley Wickham @hadleywickham Following

The only two things that make @JennyBryan Instead use projects + here::here()
#rstats



4:50 PM - 10 Dec 2017

305 Retweets 992 Likes

here::here()

Where did you learn that?

working in **RStudio** connected
to my own public **GitHub**
repository

Happy Git and GitHub for the useR

Jenny Bryan, the STAT 545 TAs, Jim Hester

Let's Git started

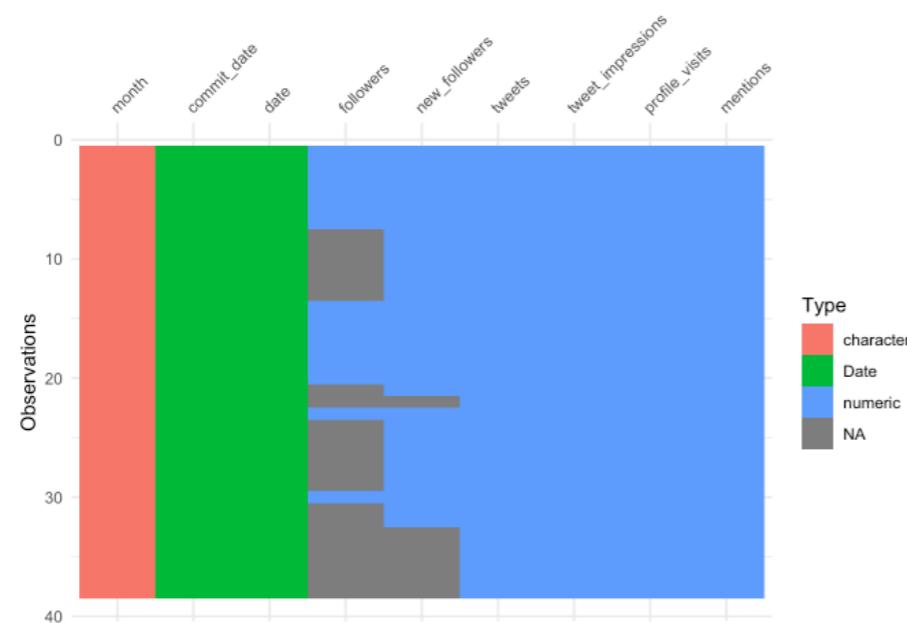


Where did you learn that?

I knew where my **missing data** was

Try `visdat` magic

```
# use visdat package to visualize missing data
# this was the first package I tried after reading in csv, initially just for fun
vis_dat(metrics)
```



The rOpenSci blog!

A screenshot of a blog post on the rOpenSci website. The title is "Onboarding visdat, a tool for preliminary visualisation of whole dataframes". The author is Nicholas Tierney, and the date is August 22, 2017. A call-to-action button at the bottom right says "Take a look at the data".

Onboarding visdat, a tool for preliminary visualisation of whole dataframes

Nicholas Tierney

August 22, 2017

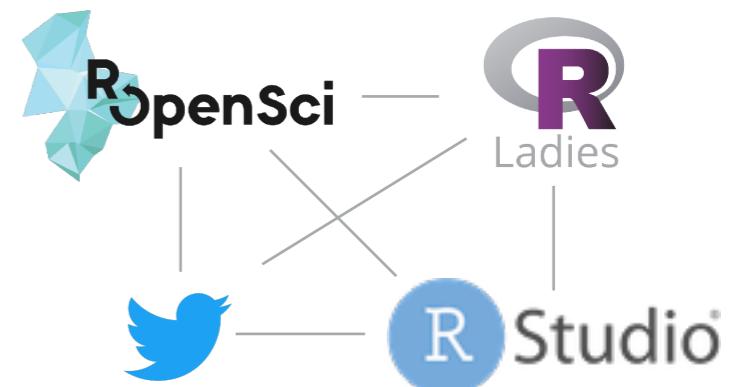
Take a look at the data

Where did you learn that?

“wait, my x-axis of dates sorted itself alphabetically?!”

lubridate!

I used “masking the function” and
“calling the function from the package”
in a sentence

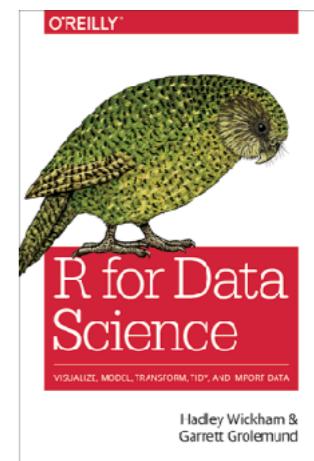


[Conflict between here in, uh, here, and here in lubridate?](#)

tidyverse lubridate

mara Sustainer 1 2018-03-02

Personally, I almost always use `here::here()`, even when I *don't* have lubridate attached. But, as jenny mentioned, you've described the situation as it is.



3 to my success

- in-person and online trust network
- a specific modest project
- the “right” tutorials and resources for me

“Suck until you don't”

–Pat Schloss

“If it hurts, do it more often”

- Martin Fowler h/t Jenny Bryan

chadfawcett 2:03 PM
@stefanie wreaking havoc on the R repo again... 😬

Screen Shot 2019-02-05 at 2.01.27 PM.png ▾

twoStoryRobot/edynamic-portal-app · You commented a day ago

Show more

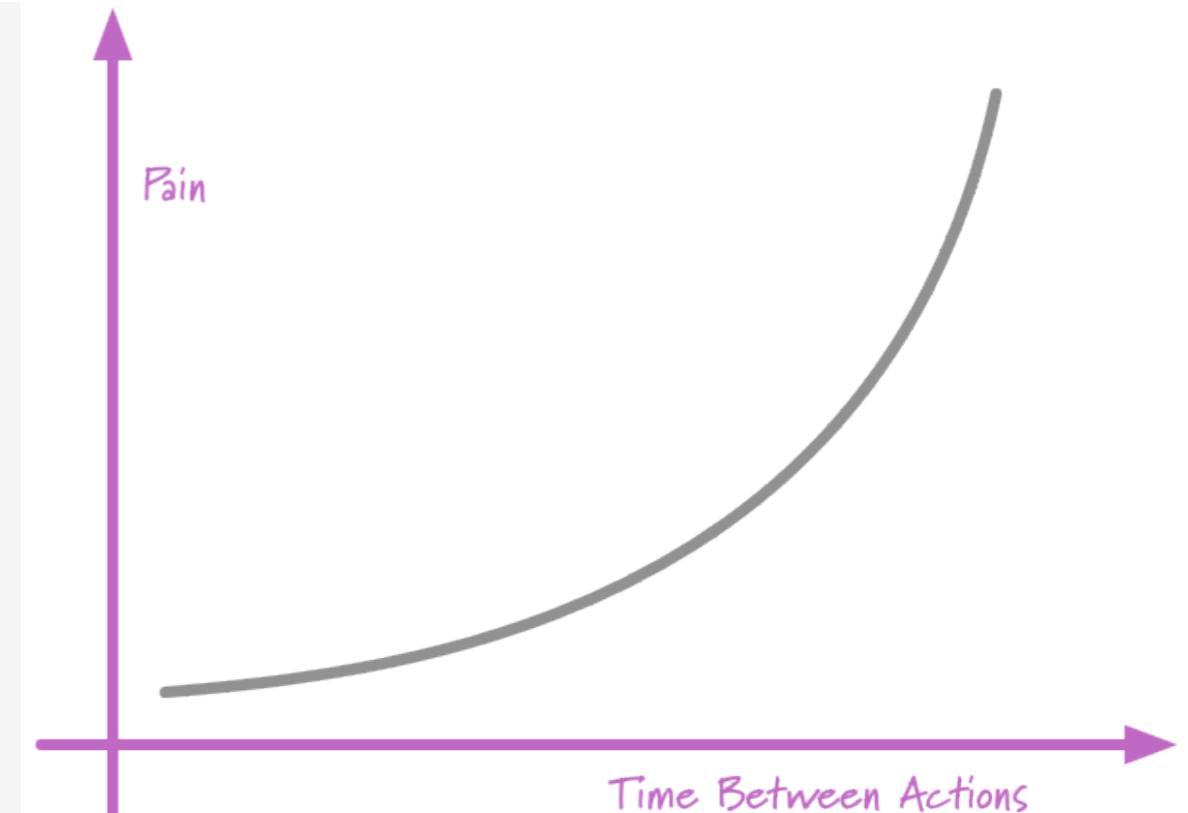
All activity

stefaniebutland pushed to KamloopsR/Meetup 2 hours ago

1 commit to master

9376190 haha edited master directly itsonlytext

1 😬 1 😂



Appreciation

Kara Woo and Monica Gerber for
saying this is a talk they'd like to hear

Maëlle Salmon for talk feedback

rOpenSci team for sustaining a
values-driven organization

Get Involved!

 @rOpenSci - <https://ropensci.org/community/>

https://ropensci.github.io/dev_guide/

 @TheCSCCE - Community Management

 @StefanieButland