PILGRIM'S PROGRESS

a journey from confusion to contribution

Mara Averick

TIDYVERSE DEV ADVOCATE, RSTUDIO

Not a Real Data Scientist™

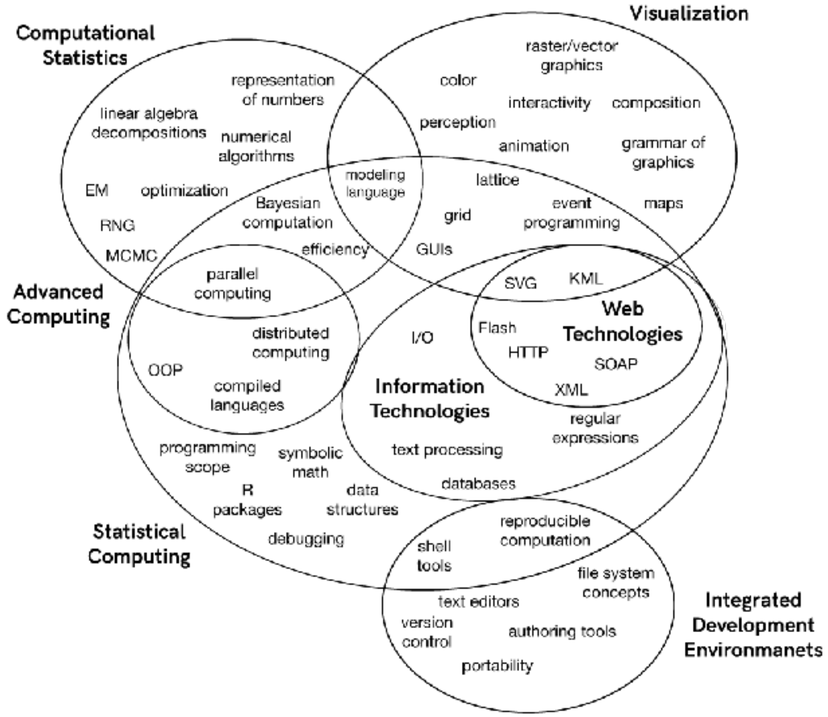
Mara Averick

TIDYVERSE DEV ADVOCATE, RSTUDIO

Not a Real Data Scientist™

Mara Averick

TIDYVERSE DEV ADVOCATE, RSTUDIO



Mara Averick

TIDYVERSE DEV ADVOCATE, RSTUDIO

Less true these days!

An aside on the title

“Like many social groups that do not reproduce themselves biologically, the experimental particle physics community renews itself by training novices.”

— Sharon Traweek, *Pilgrim's Progress: Male Tales Told During a Life in Physics*

In Beamtimes and Lifetimes: The World of High Energy Physics. (1988). Cambridge, MA: Harvard University Press.

SCIENCE & SOCIETY

DDTX 2018

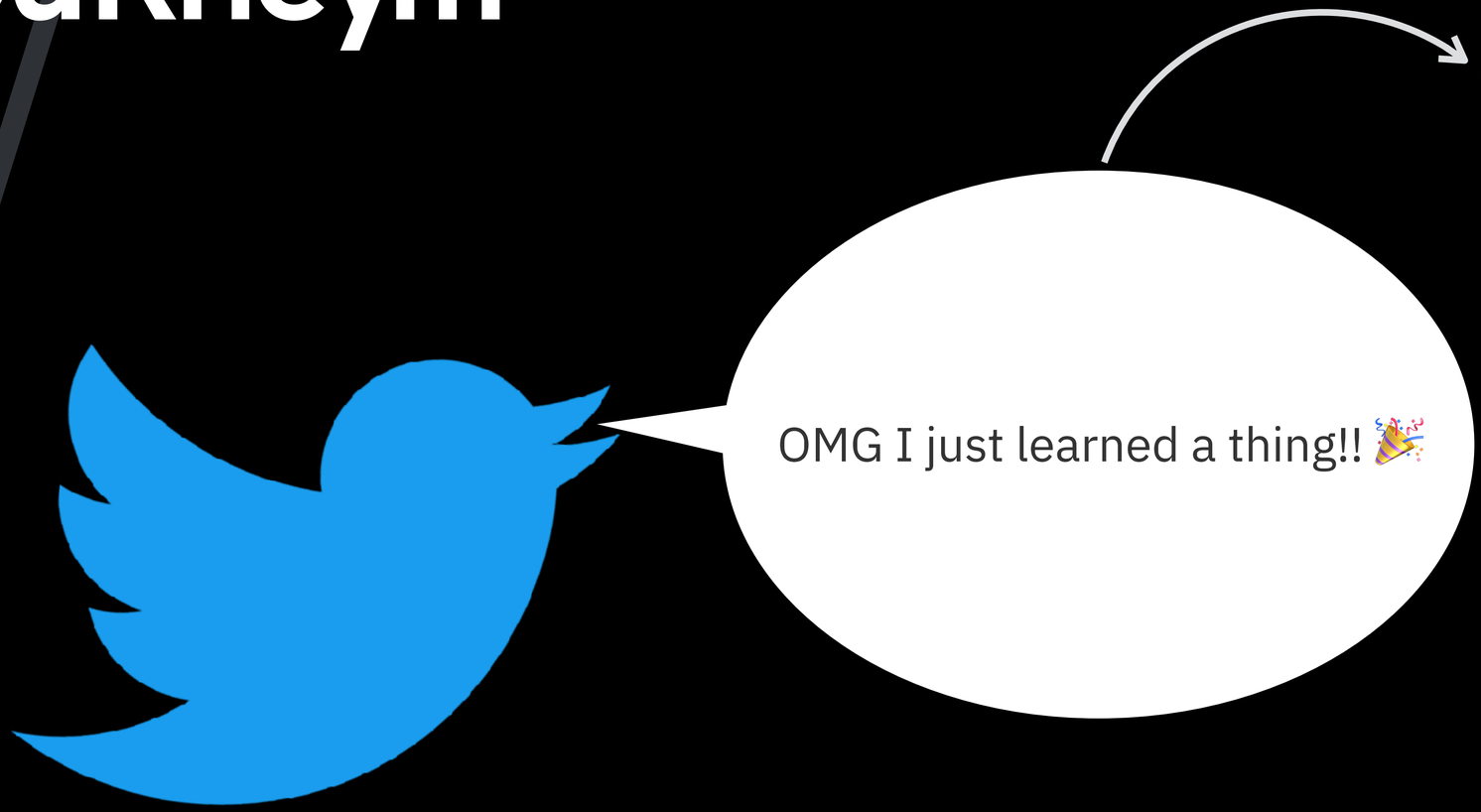
An aside on the title

An aside on the title

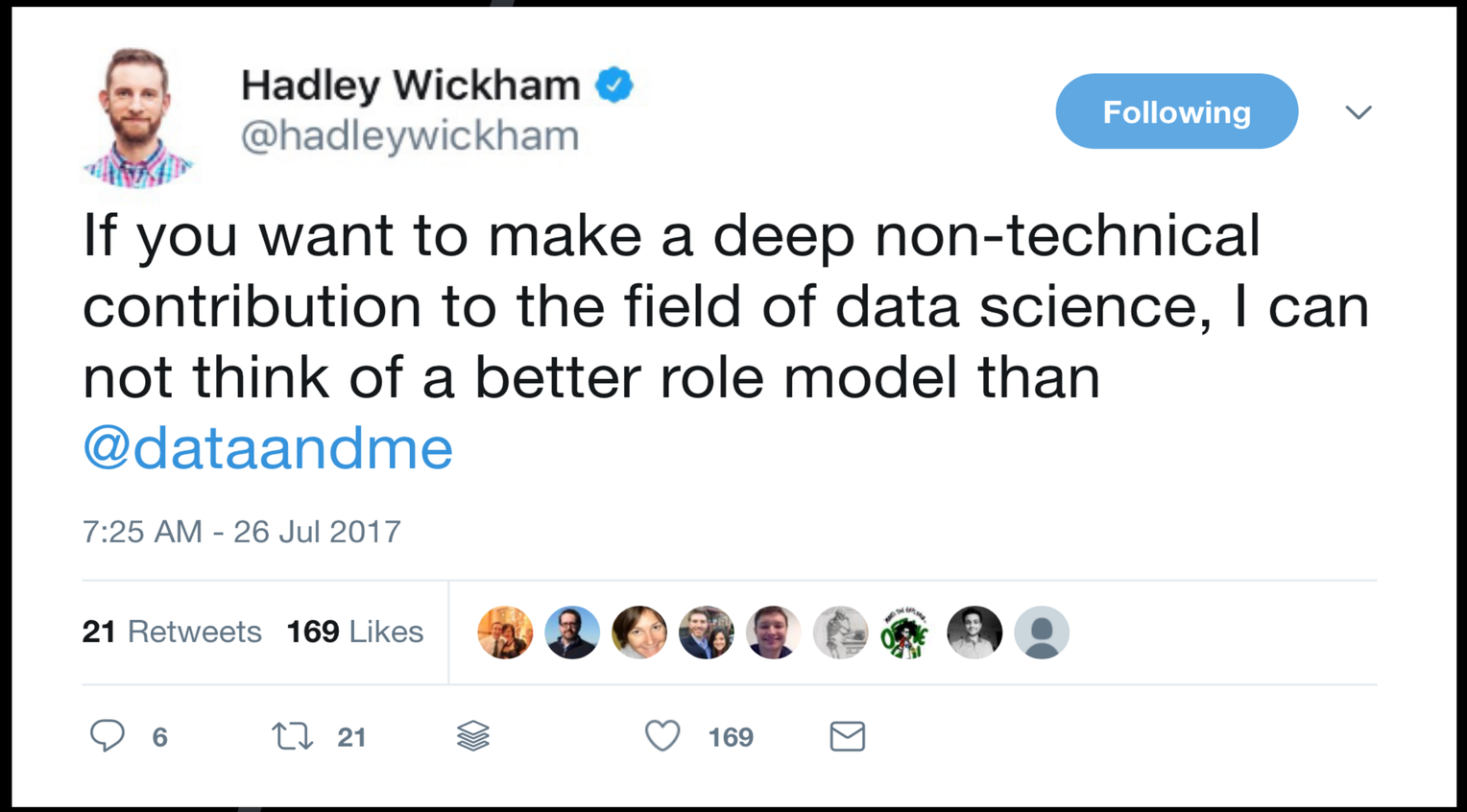
my jouRney...

my jouRney…

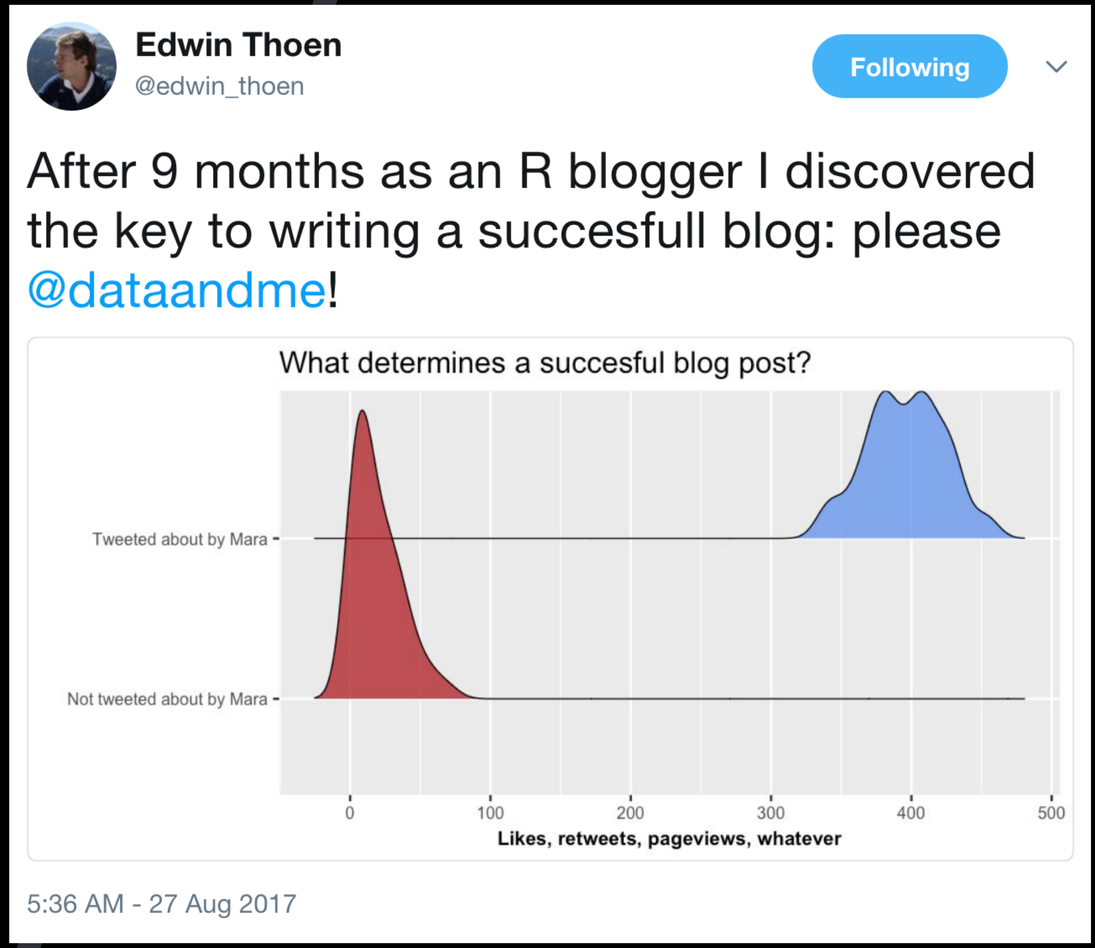
my jouRney...



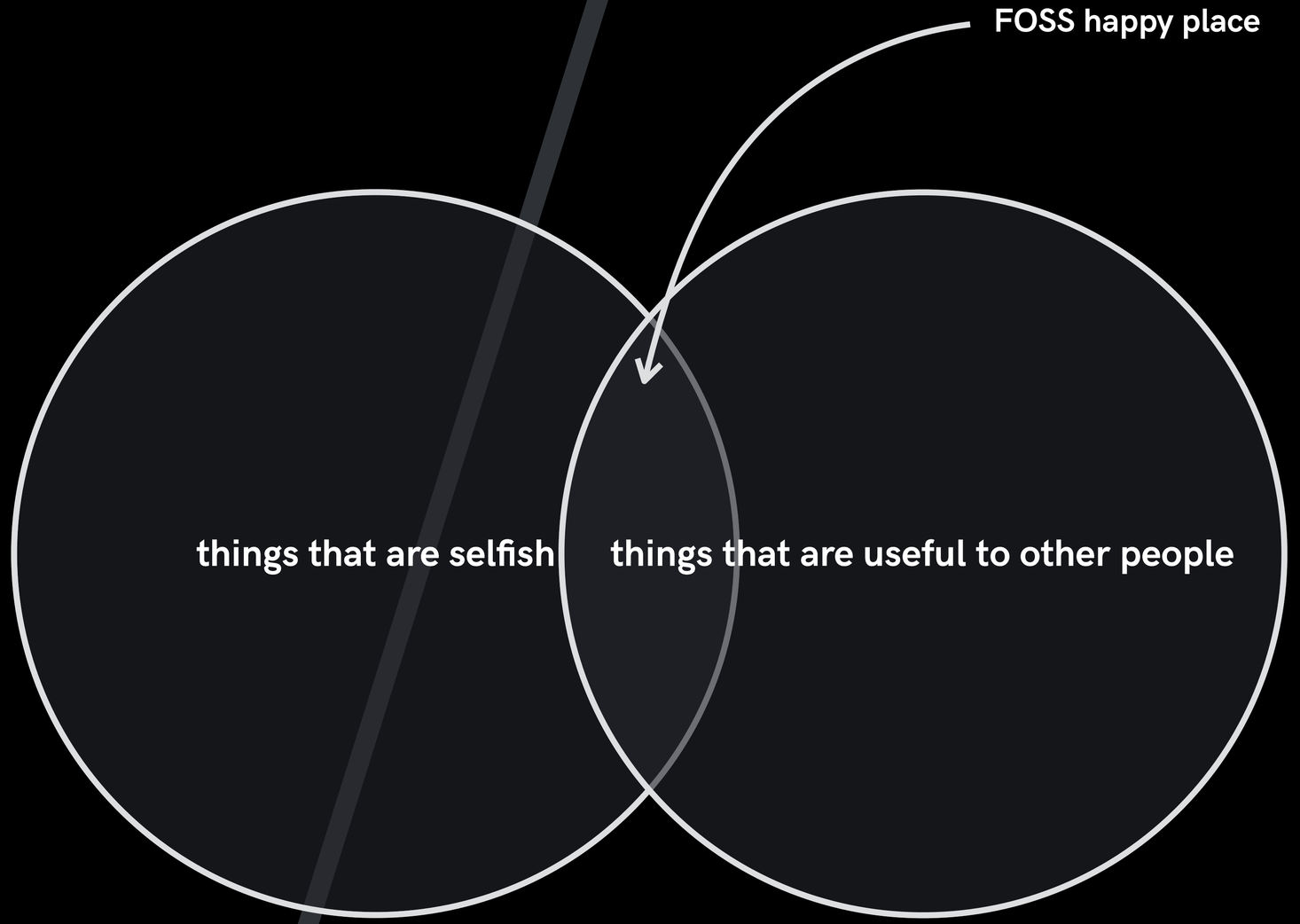
but…


Hadley Wickham
@hadleywickham
Following
If you want to make a deep non-technical contribution to the field of data science, I can not think of a better role model than @dataandme
7:25 AM - 26 Jul 2017
21 Retweets 169 Likes
6
21
169

but…


Edwin Thoen
@edwin_thoen
Following
After 9 months as an R blogger I discovered the key to writing a succesfull blog: please @dataandme!
What determines a succesful blog post?
Tweeted about by Mara
Not tweeted about by Mara
0
100
200
300
400
500
Likes, retweets, pageviews, whatever
5:36 AM - 27 Aug 2017

but...


FOSS happy place
things that are selfish
things that are useful to other people


Hadley Wickham
@hadleywickham
Follow
Replying to @juliasilge @arnicas @dataandme
I really didn't want to imply that Mara doesn't have awesome technical skills ☹ She does!
8:32 AM - 26 Jul 2017

Hadley Wickham ✓
@hadleywickham

Follow

Replying to @juliasilge @arnicas @dataandme

I really didn't want to imply that Mara doesn't have awesome technical skills 😕 She does!

8:32 AM - 26 Jul 2017

Lucy 🌻
@LucyStats

Follow

Replying to @hadleywickham @juliasilge and 2 others

I read as @dataandme has the ability to translate quite technical material to a non-tech audience, so if you're non-tech she's a 🔥 go to! 👯🏆

8:37 AM - 26 Jul 2017

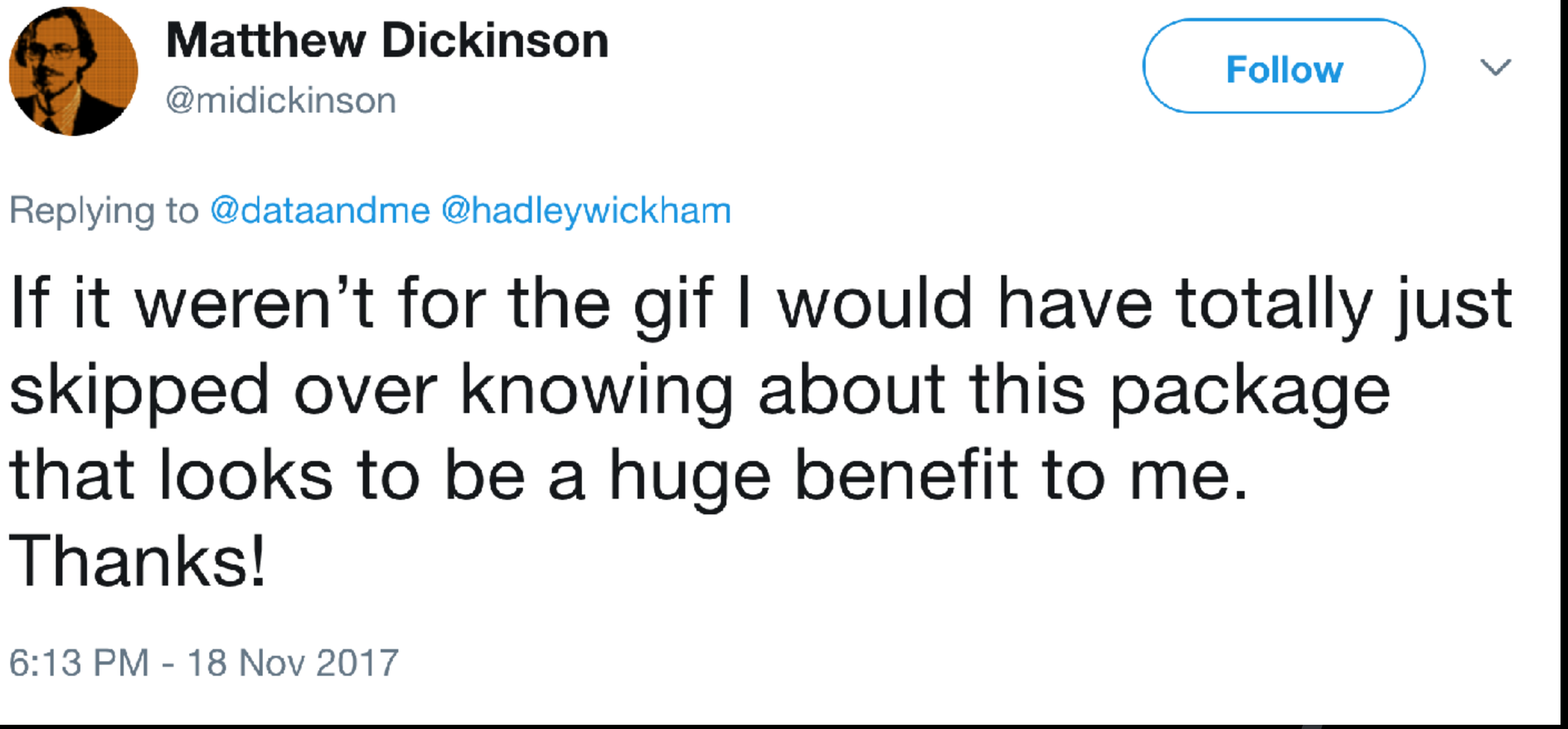
ex•o•ter•ic

adj. *understandable by outsiders*

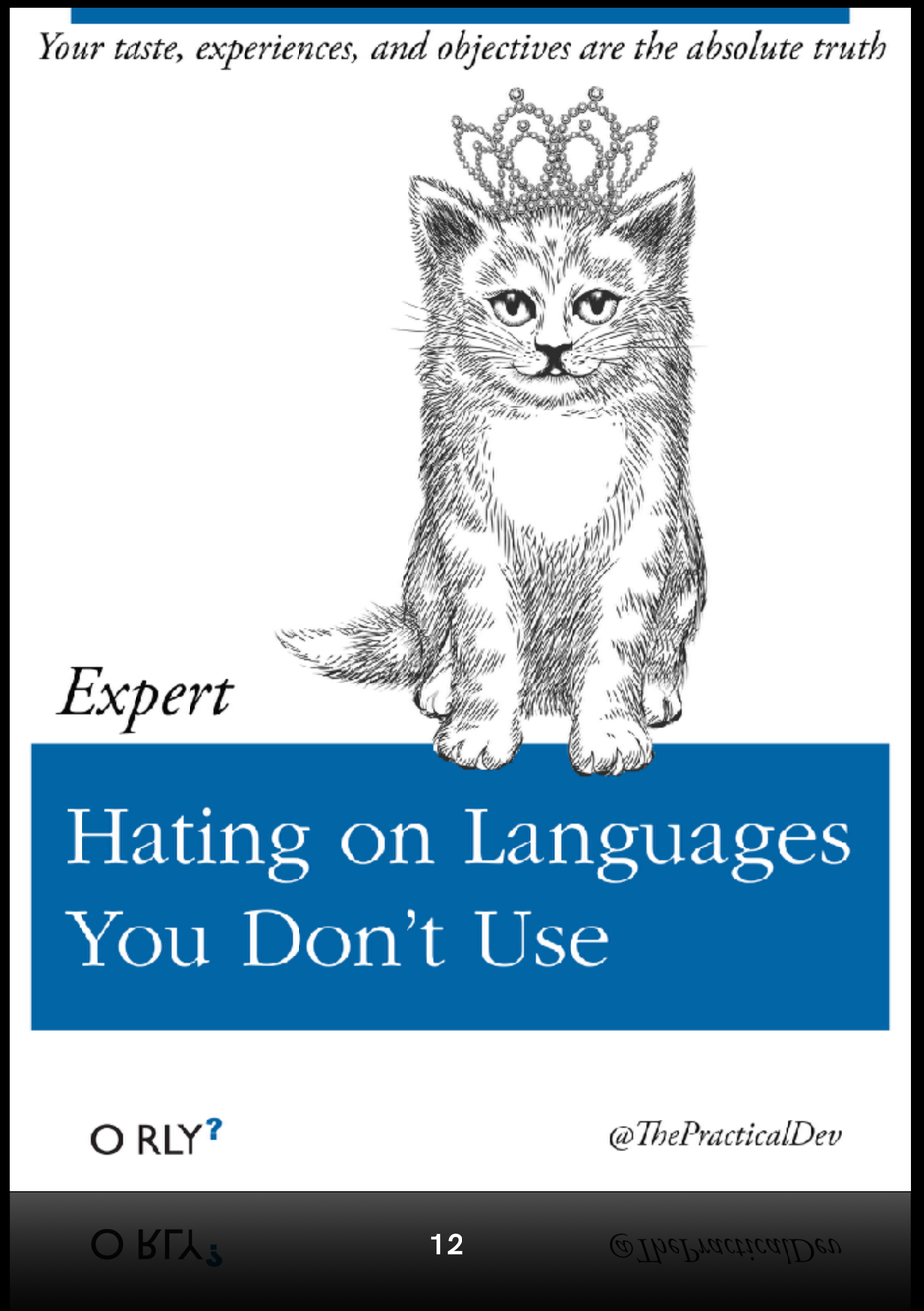
or the general public

you never know...

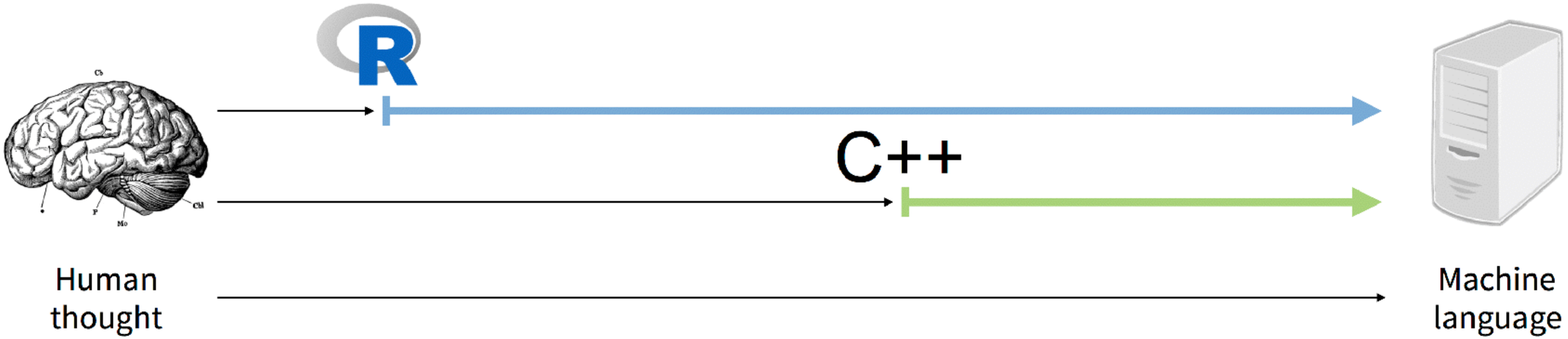
you never know...



This Talk Will Not Cover



R - a computer language for scientists



via Garrett Grolemund


tidyverse
www.rstudio.com

What about this so-called tidyverse?

The tidyverse is an opinionated collection of R packages designed for data science.

All packages share an underlying design philosophy, grammar, and data structures.

source: https://www.tidyverse.org/

TIDY TOOLS

Functions should be...

SIMPLE

Do one thing and do it well.

COMPOSABLE

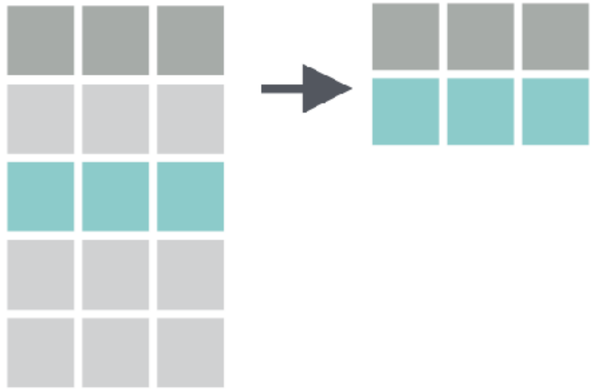
Combine with other functions for multi-step operations.

DESIGNED FOR HUMANS

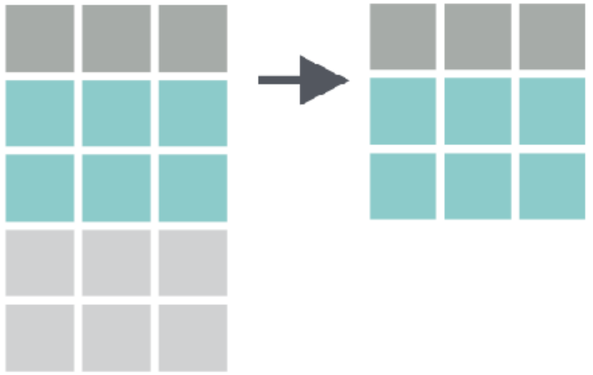
Use evocative verb names, making them easy to remember.

Source: Wickham, Hadley. 2017-11-13. “The tidy tools manifesto.” https://cran.r-project.org/web/packages/tidyverse/vignettes/manifesto.html

FUNCTION EXAMPLES



filter(.data, …**)** Extract rows that meet logical criteria. Also **filter_()**. *filter(iris, Sepal.Length > 7)*



top_n(x, n, wt**)** Select and order top n entries (by group if grouped data). *top_n(iris, 5, Sepal.Width)*

COMPOSE WITH THE PIPE

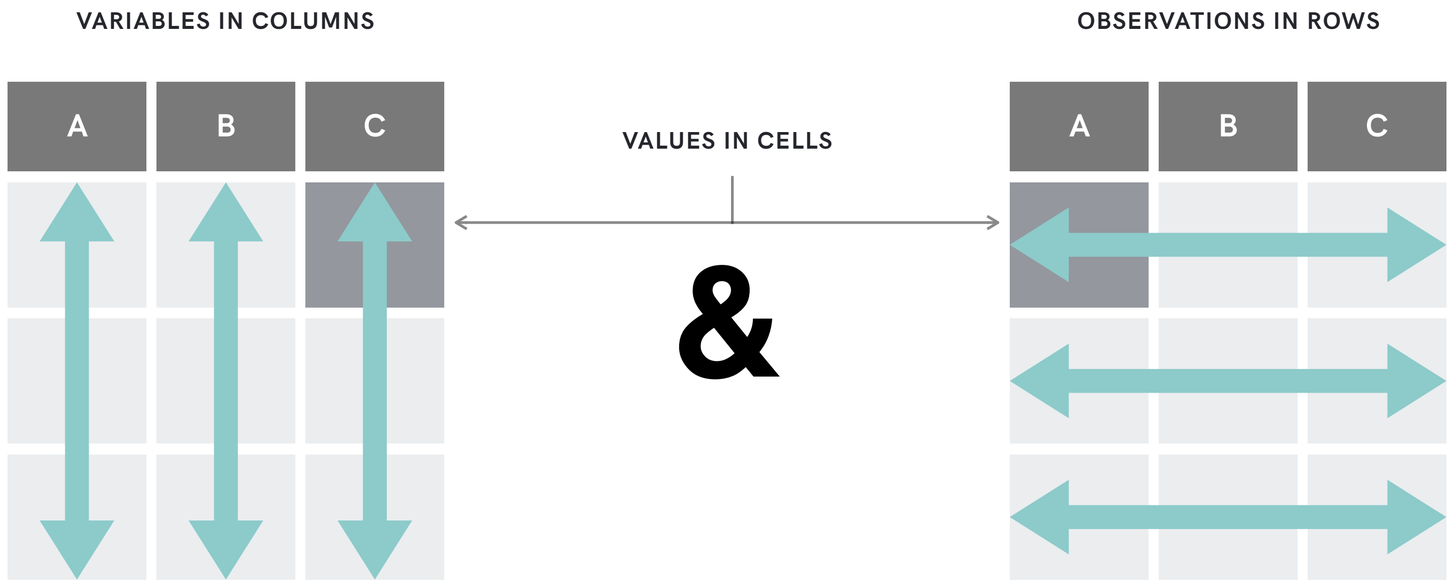
```
iris %>%
    filter(Sepal.Length > 7) %>%
    top_n(5, Sepal.Width)
```

COMPOSE WITH THE PIPE

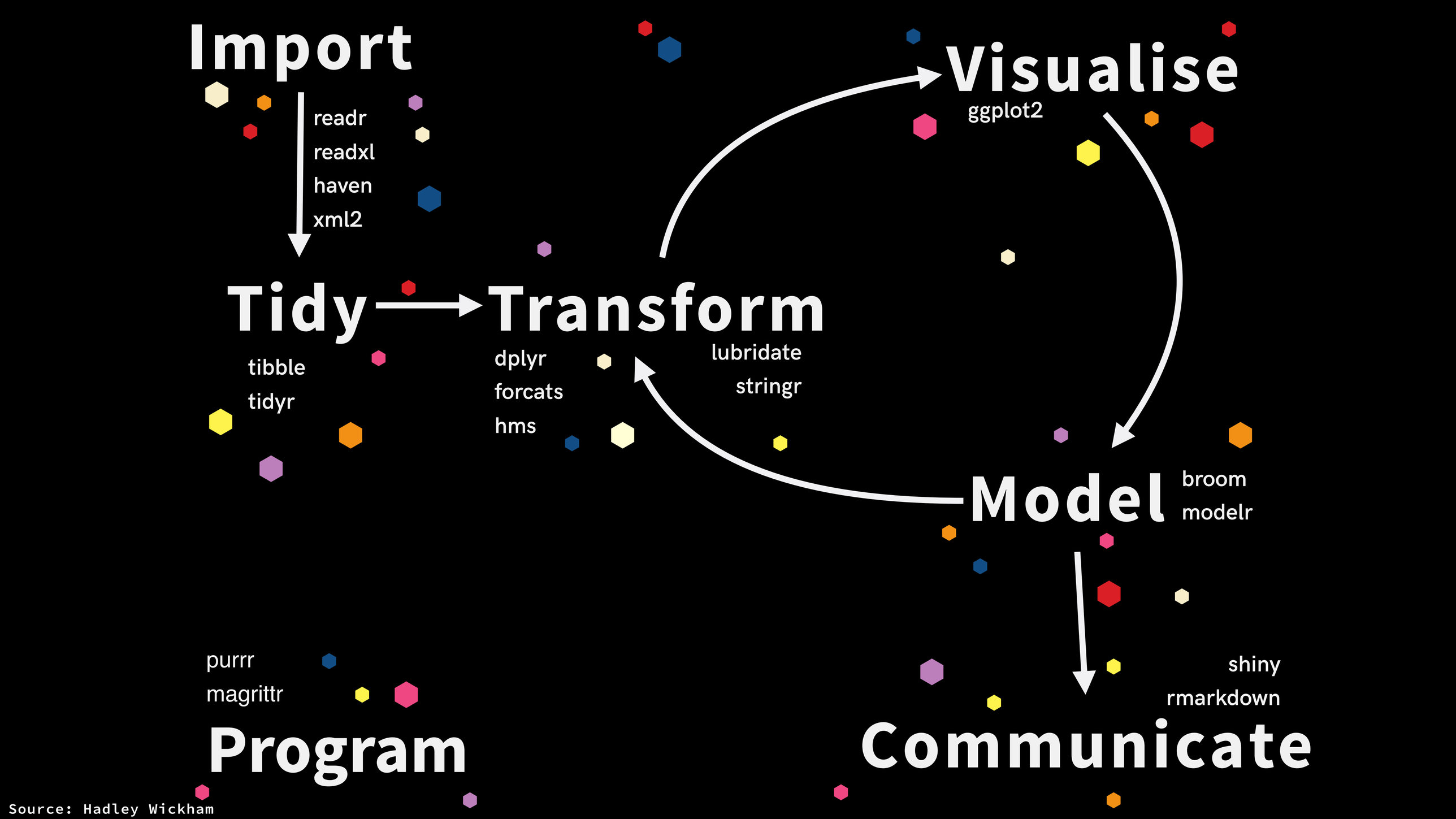
COMPOSE WITH THE PIPE

```
iris %>%
    filter(Sepal.Length > 7) %>%
    top_n(5, Sepal.Width)
```

TIDY DATA



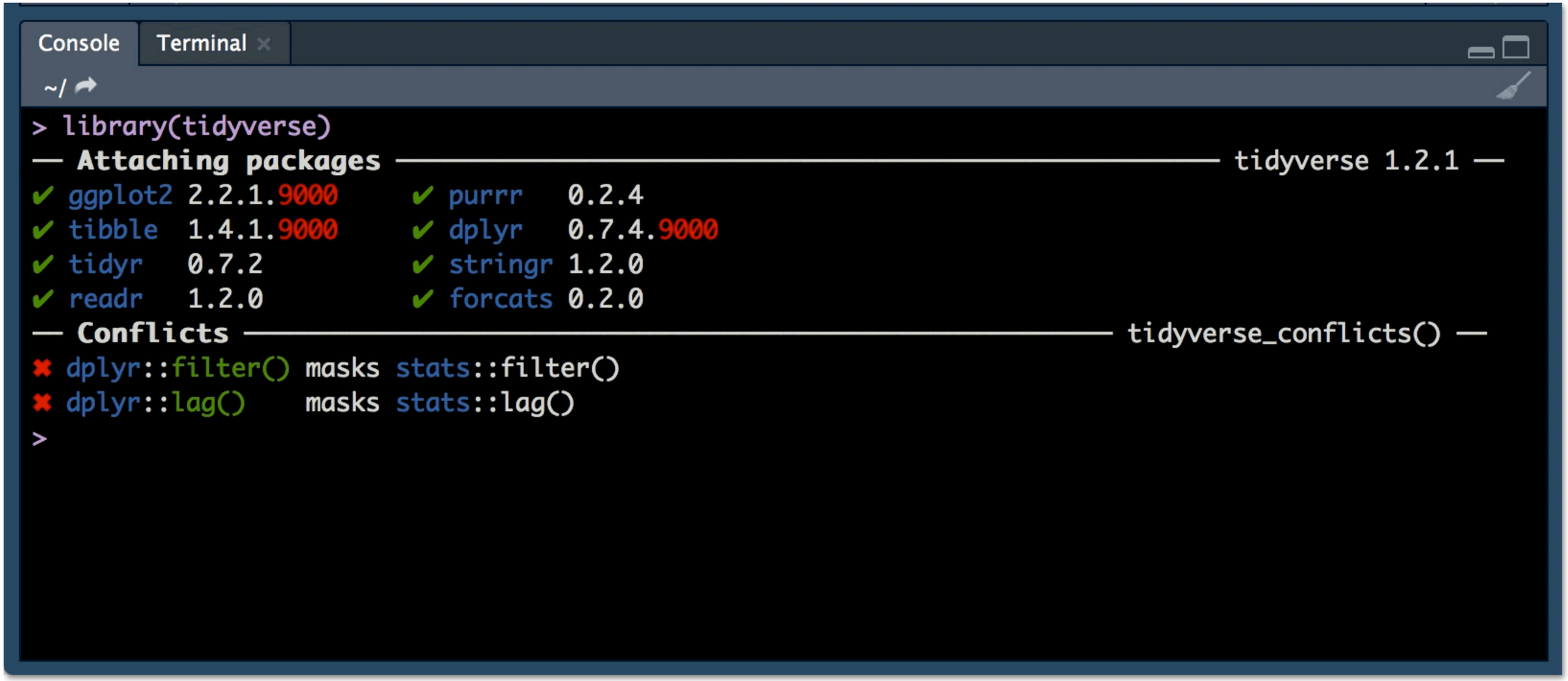
Source: Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1-23. doi:http://dx.doi.org/10.18637/jss.v059.i10.



Source: Hadley Wickham

install.packages("tidyverse")

library(tidyverse)



TIDYVERSE PACKAGES

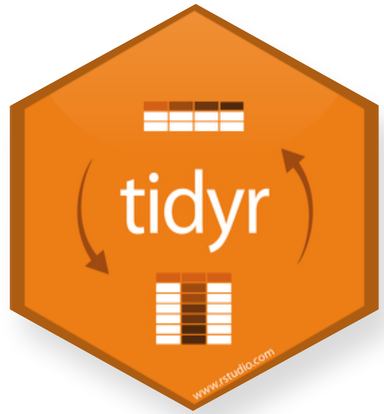
THE CORE

TIDYVERSE

a wrapper package that makes it easy to install and load core packages from the tidyverse in a single command

READR

a fast and friendly way to read in and parse rectangular data (like csv, tsv, and fwf)



TIDYR

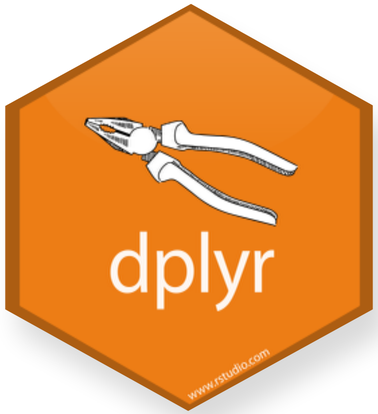
a set of verbs that help you get to tidy data, allowing you to work with other tidyverse packages and store results

TIDYVERSE PACKAGES

THE CORE

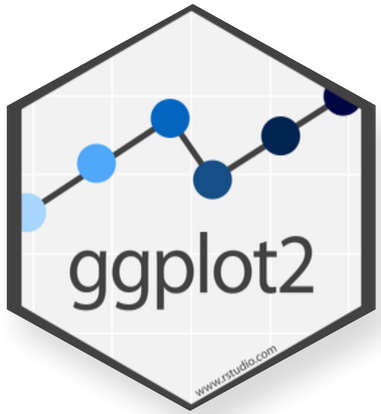
TIBBLE

a modern reimagining of data.frames that do less and complain more forcing you to confront problems earlier



DPLYR

a grammar of data manipulation with a set of verbs to solve common data wrangling problems



GGPLOT2

a system for declaratively creating graphics, based on The Grammar of Graphics

TIDYVERSE PACKAGES

THE CORE

STRINGR

a cohesive set of functions designed to make working with strings as easy as possible

FORCATS

a suite of useful tools that solve common problems with factors, which R uses to handle categorical variables

PURRR

a consistent toolkit for enhancing R's functional programming, and working with functions and vectors

TIDYVERSE PACKAGES

SOME NON-CORE

MAGRITTR

offers a set of operators (e.g. %>%) which make code more readable by structuring sequences of operations

READXL

makes it easy to get data out of Excel and into R, and work with tabular data in R

LUBRIDATE

provides robust methods for working with date-times in R, and functionality not offered in base R

TIDYVERSE PACKAGES

SOME MORE NON-CORE

HMS

provides a simple class for storing durations or time-of-day values

BROOM

takes untidy model outputs of predictions and estimations to the tidy data we want to work with

HAVEN

enables R to read and write various data formats used by other statistical packages

TIDYVERSE PACKAGES

SOME MORE NON-CORE

GOOGLEDRIVE

allows you to interact with files on Google Drive from R

RMARKDOWN

an authoring framework for data science that allows you to combine prose, code, and output

SHINY

makes it easy to build interactive web apps straight from R

CONTRIBUTING TO THE

tidyverse

Contributing to FOSS

Pintscher, Lydia, Ed. 2012. *Open Advice: Foss: What We Wish We Had Known When We Started.*

photo cred: Sail Fish Scuba https://sailfishscuba.com/manowar/

Contributing to FOSS

WHAT HOLDS PEOPLE BACK?

Pintscher, Lydia, Ed. 2012. *Open Advice: Foss: What We Wish We Had Known When We Started.*

photo cred: Sail Fish Scuba https://sailfishscuba.com/manowar/

Contributing to FOSS

WHAT HOLDS PEOPLE BACK?

- *“I can't write code.”*

Pintscher, Lydia, Ed. 2012. *Open Advice: Foss: What We Wish We Had Known When We Started.*

photo cred: Sail Fish Scuba https://sailfishscuba.com/manowar/

Contributing to FOSS

WHAT HOLDS PEOPLE BACK?

- *“I can't write code.”*
- *“I'm not really good at this.”*

Pintscher, Lydia, Ed. 2012. *Open Advice: Foss: What We Wish We Had Known When We Started.*

photo cred: Sail Fish Scuba https://sailfishscuba.com/manowar/

Contributing to FOSS

WHAT HOLDS PEOPLE BACK?

- *“I can't write code.”*
- *“I'm not really good at this.”*
- *“I'd just be a burden.”*

Pintscher, Lydia, Ed. 2012. *Open Advice: Foss: What We Wish We Had Known When We Started.*

photo cred: Sail Fish Scuba https://sailfishscuba.com/manowar/

Contributing to FOSS

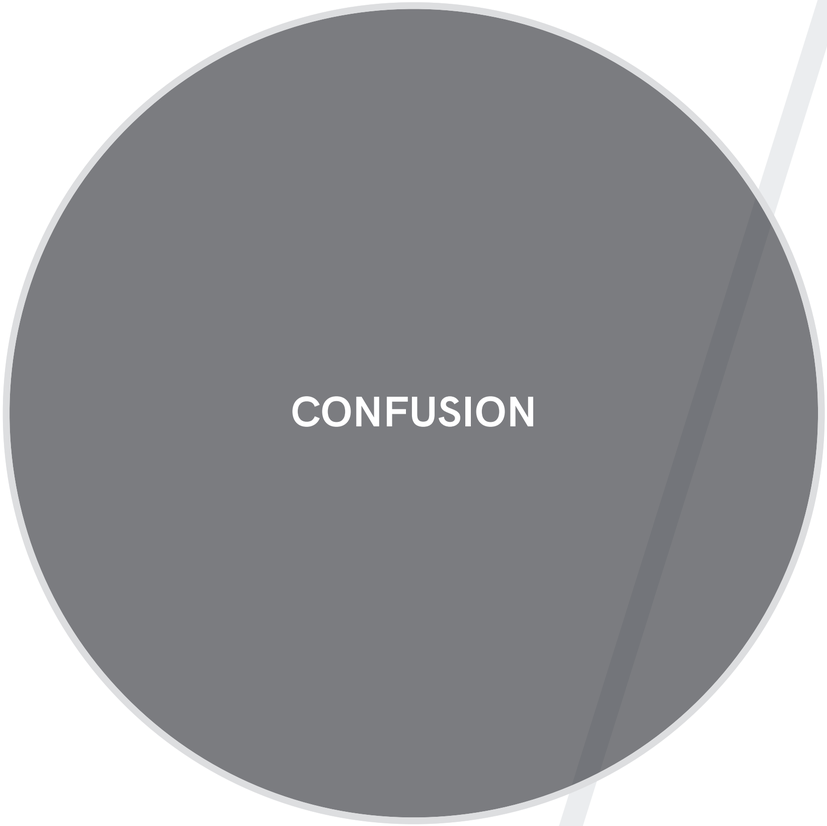
WHAT HOLDS PEOPLE BACK?

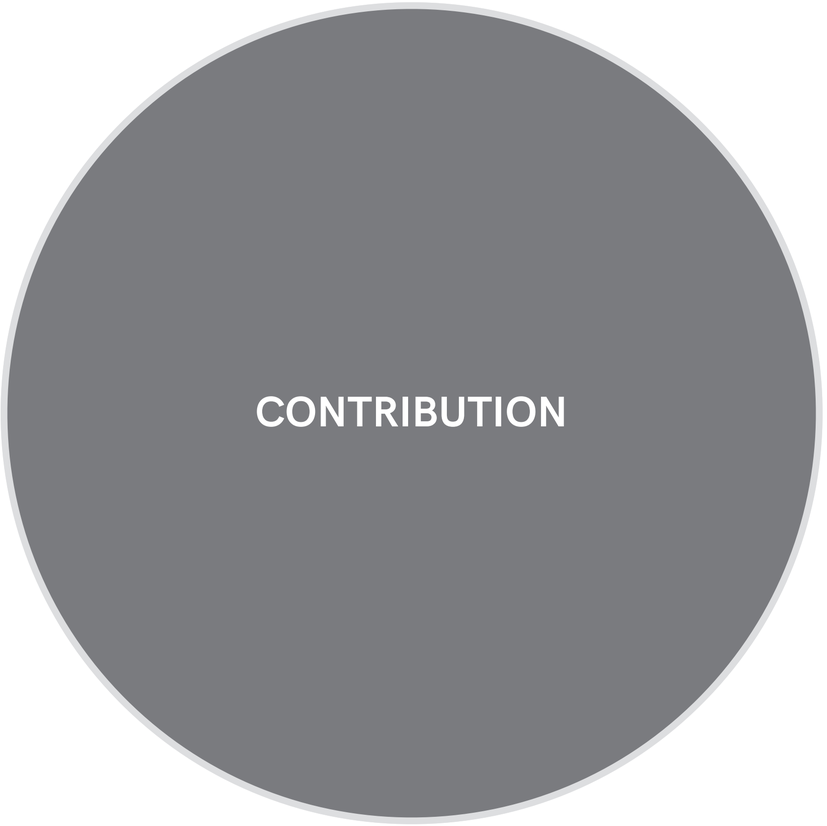
- *“I can't write code.”*
- *“I'm not really good at this.”*
- *“I'd just be a burden.”*
- *“They already have enough people smarter than me.”*

Pintscher, Lydia, Ed. 2012. *Open Advice: Foss: What We Wish We Had Known When We Started.*

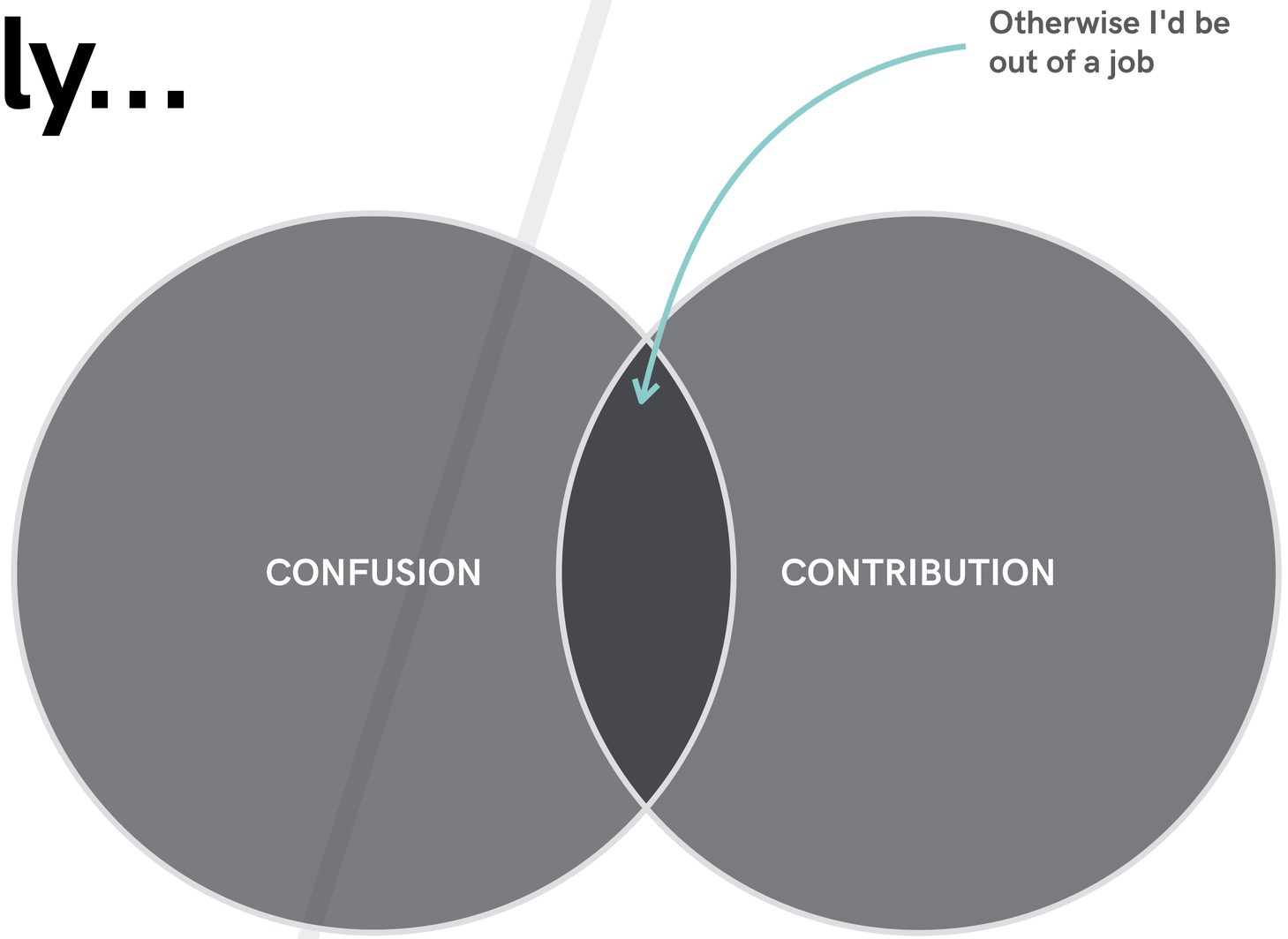
photo cred: Sail Fish Scuba https://sailfishscuba.com/manowar/

Luckily…





Luckily...



Ask questions

The most useless problem statement that one can face is "it doesn't work", yet we seem to get it far too often.

– Thiago Maciera

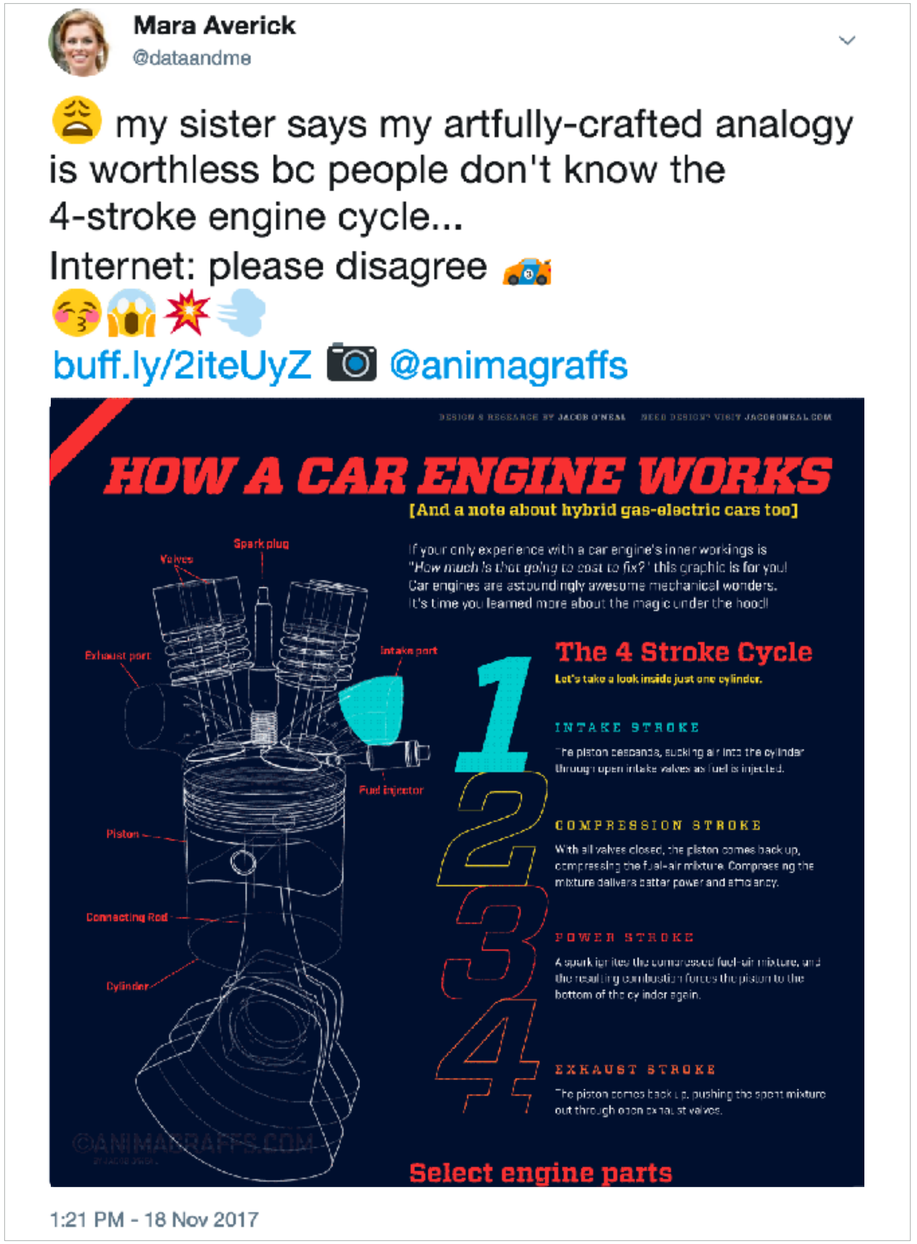
Maciera, Thiago. 2012. "The Art of Problem Solving." In Open Advice: FOSS: What We Wish We Had Known When We Started, edited by Lydia Pintscher, 55–61.

The newcomer's paradox…

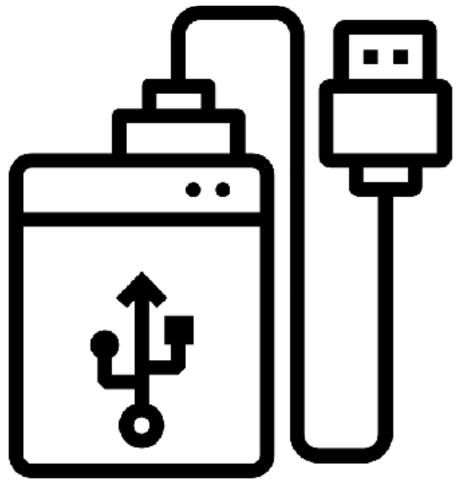
When you ask for help, some friendly soul will no doubt tell you that "it's easy, just do foo, bar and baz." *Except for you, it is not easy, there may be no documentation for foo, bar is not doing what it is supposed to be doing and what is this baz thing anyway with its eight disambiguation entries on Wikipedia?*

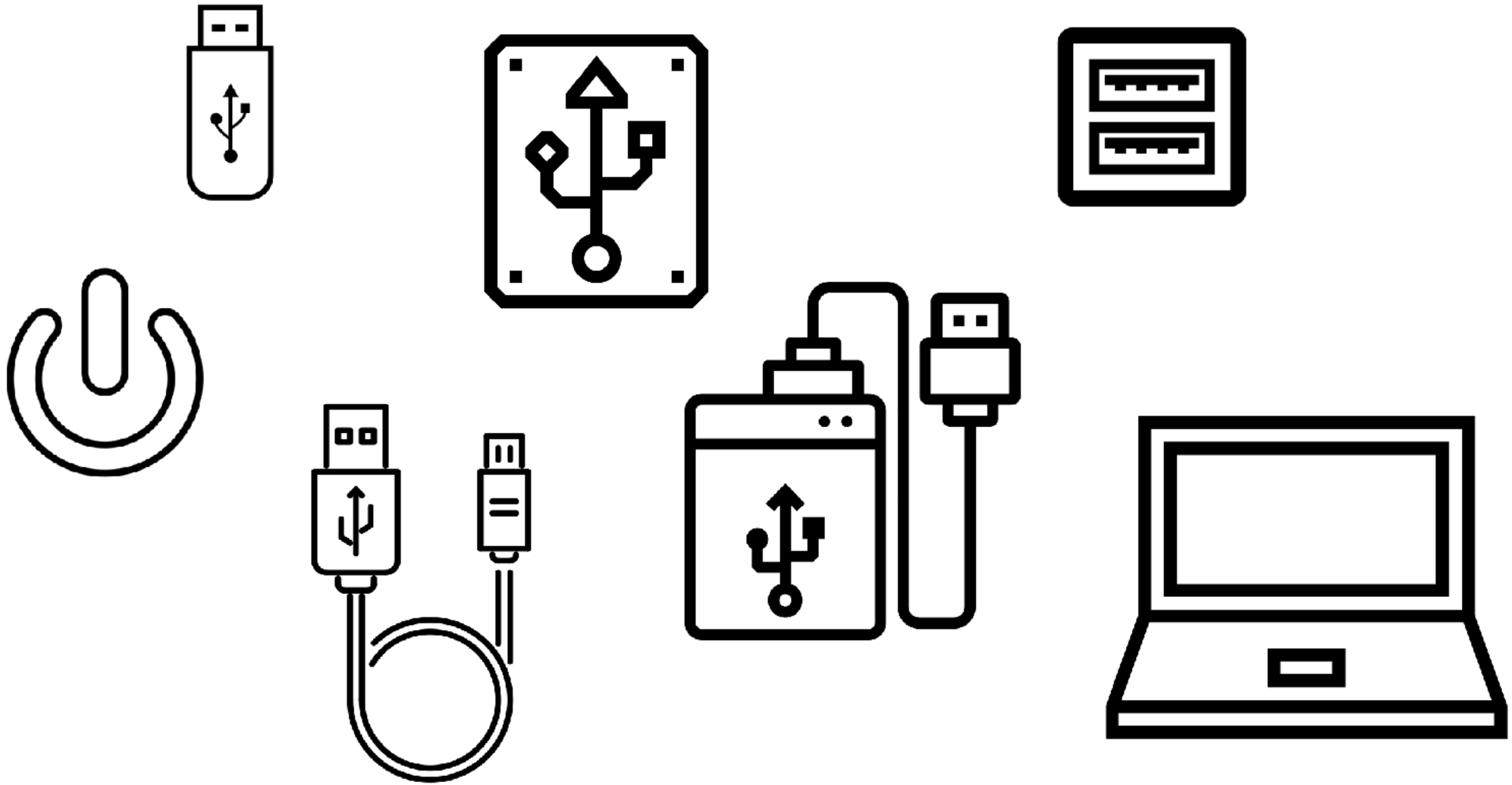
— Leslie Hawthorne

"You'll Eventually Know Everything They've Forgotten." In *Open Advice: FOSS: What We Wish We Had Known When We Started*, edited by Lydia Pintscher, 29–32.



https://animagraffs.com/how-a-car-engine-works/





where to ask

Twitter

StackOverflow

RStudio Community

the magic of reprex

reproducible
example

reprex raison d'être

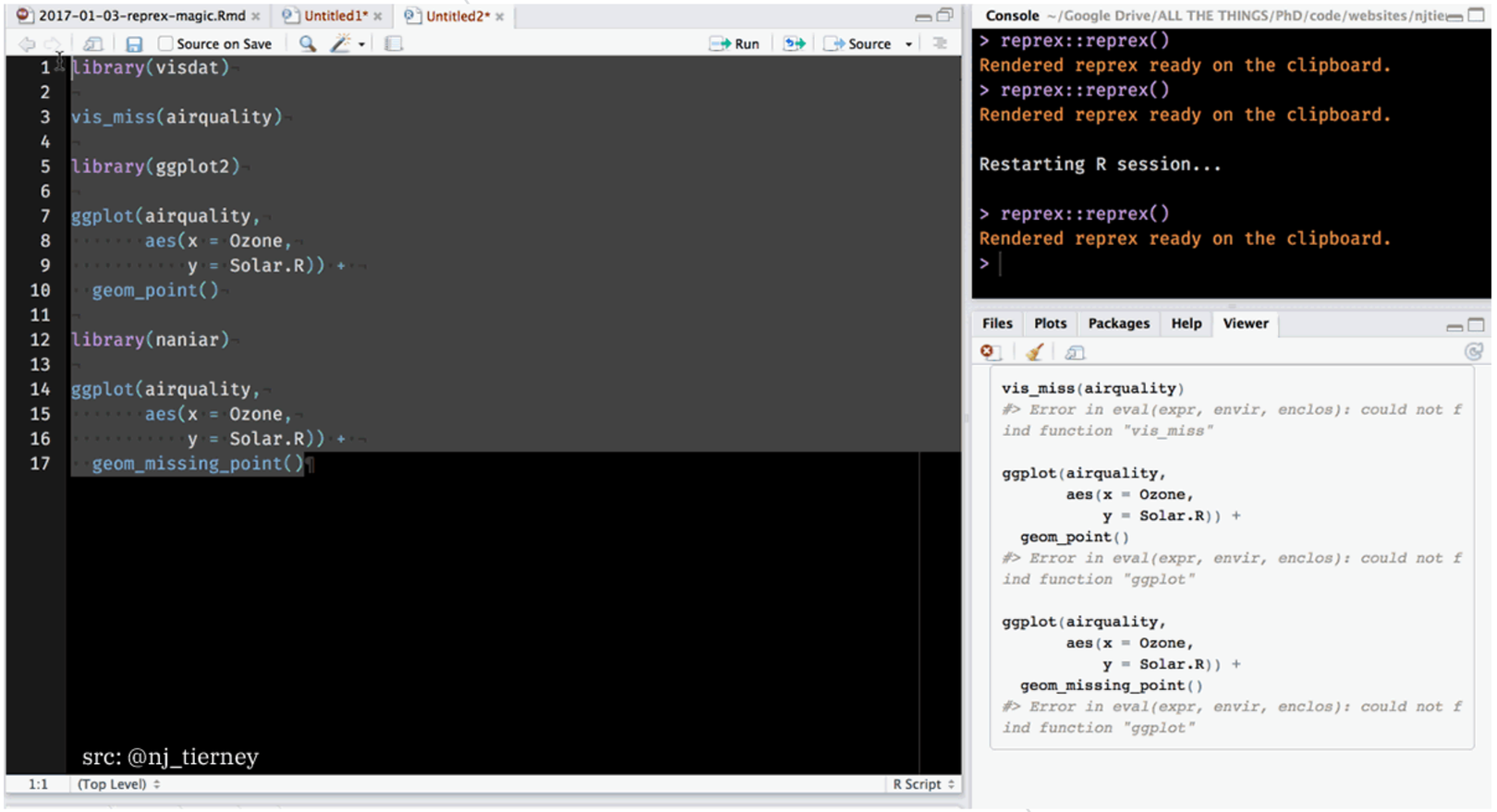
Keys to reprex-cellence

✓ Code that **actually runs**

✓ Code that **doesn't have to be run**

✓ Code that **can be easily run**

Source: Jenny Bryan, 2017. "reprex: the package, the point." https://speakerdeck.com/jennybc/reprex-help-me-help-you



Source: Nick Tierney. "Magic reprex." 2017-01-11 <http://www.njtierney.com/post/2017/01/11/magic-reprex/>

Nailing those reprexes?

- Help others ask questions.
- Answer questions.
- Write about it.

where to answer

StackOverflow

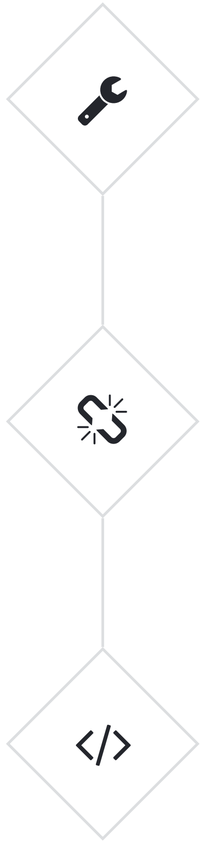
RStudio Community

GitHub

File issues

Remember, behind every octocat there is an actual human...

the anatomy of an issue



PROBLEM DESCRIPTION

EXPECTED BEHAVIOUR

REPREX

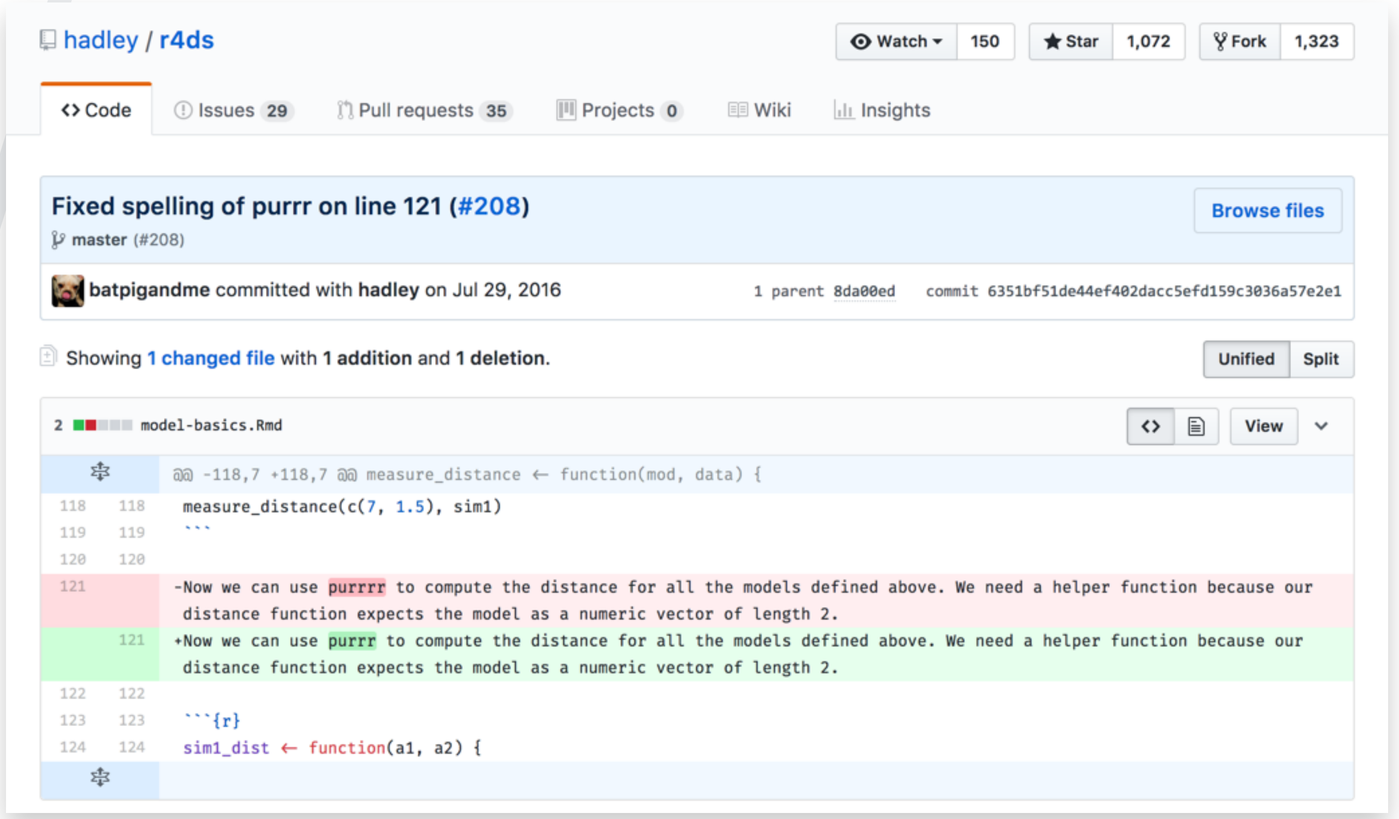
Contribute documentation

“Innocence lost is not easily regained. The designer simply cannot predict the problems people will have, the misinterpretations that will arise, and the errors that will get made.”

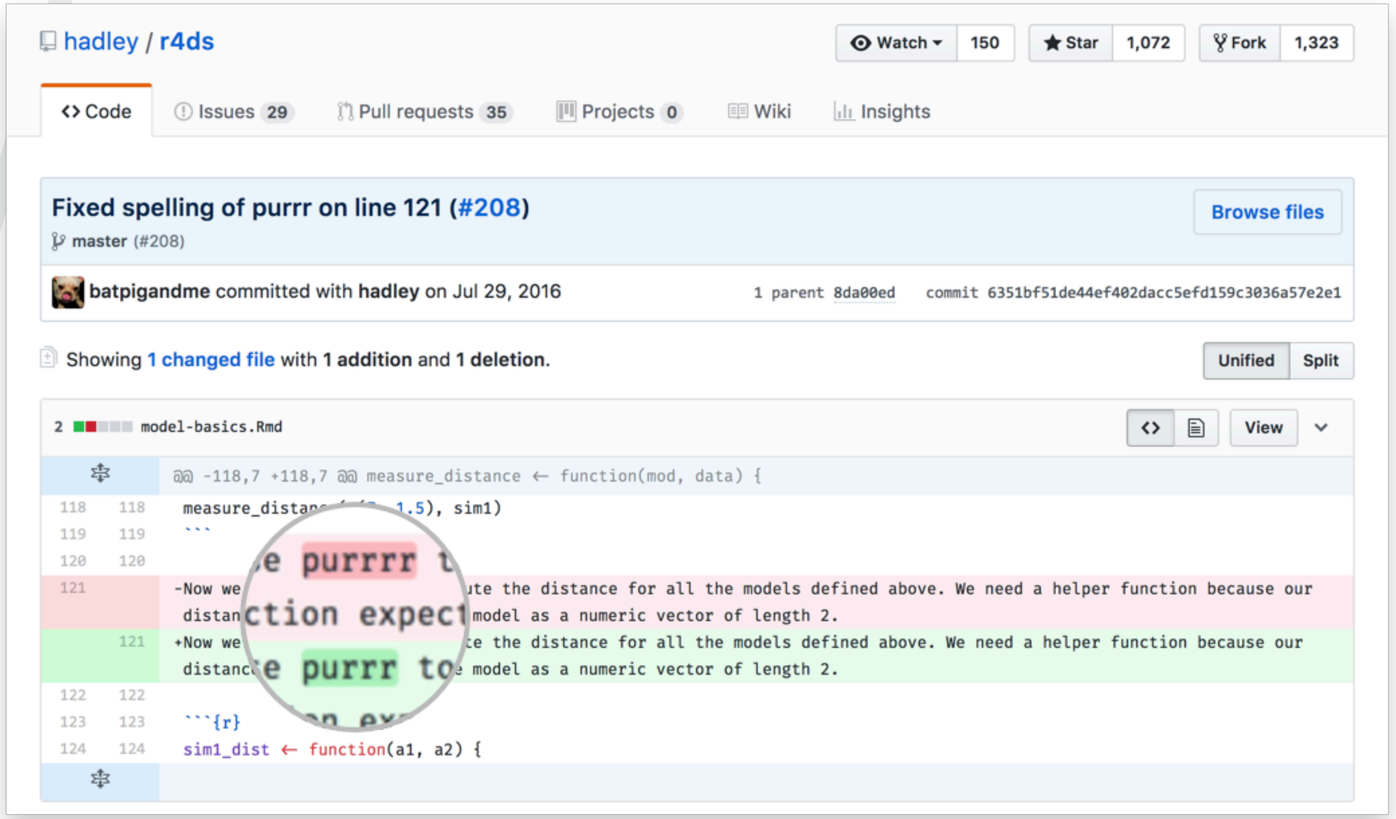
— Donald Norman, *The Design of Everyday Things*

Hawthorn, Leslie. 2012. “You’ll Eventually Know Everything They’ve Forgotten.” In *Open Advice: FOSS: What We Wish We Had Known When We Started2*, edited by Lydia Pintscher, 29–32.

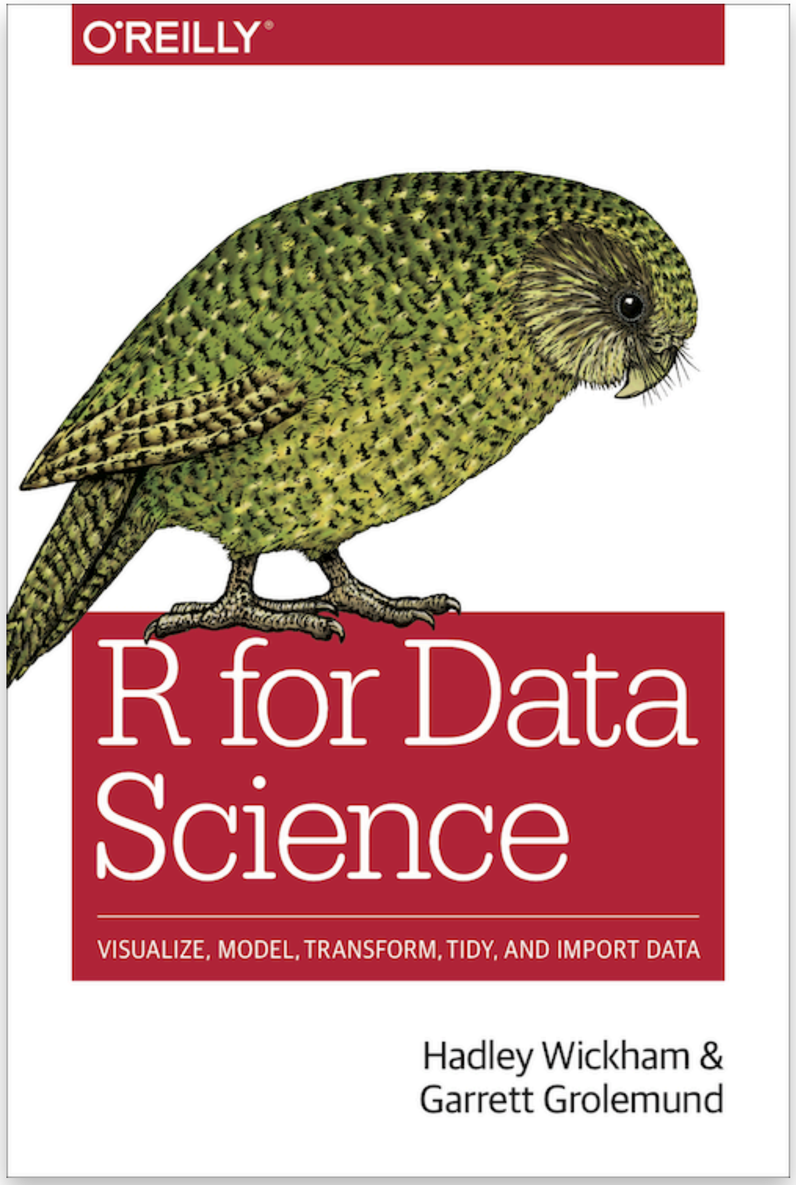
My first "contribution"



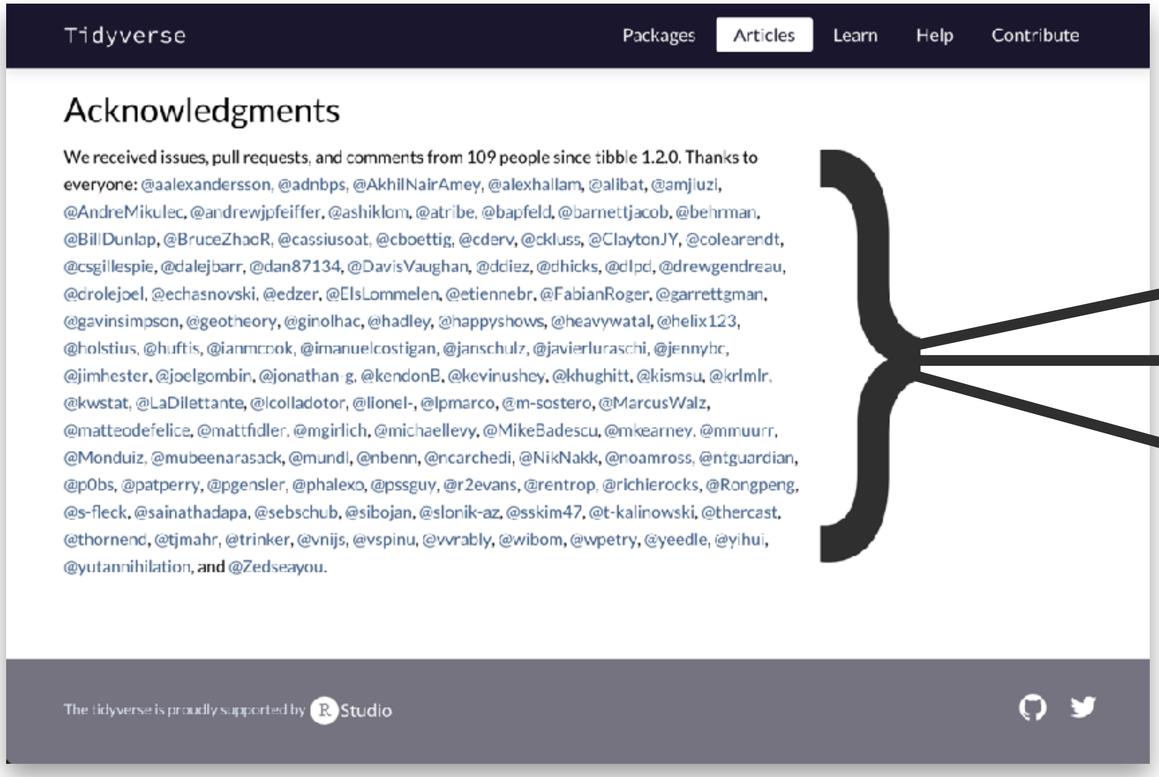
My first "contribution"



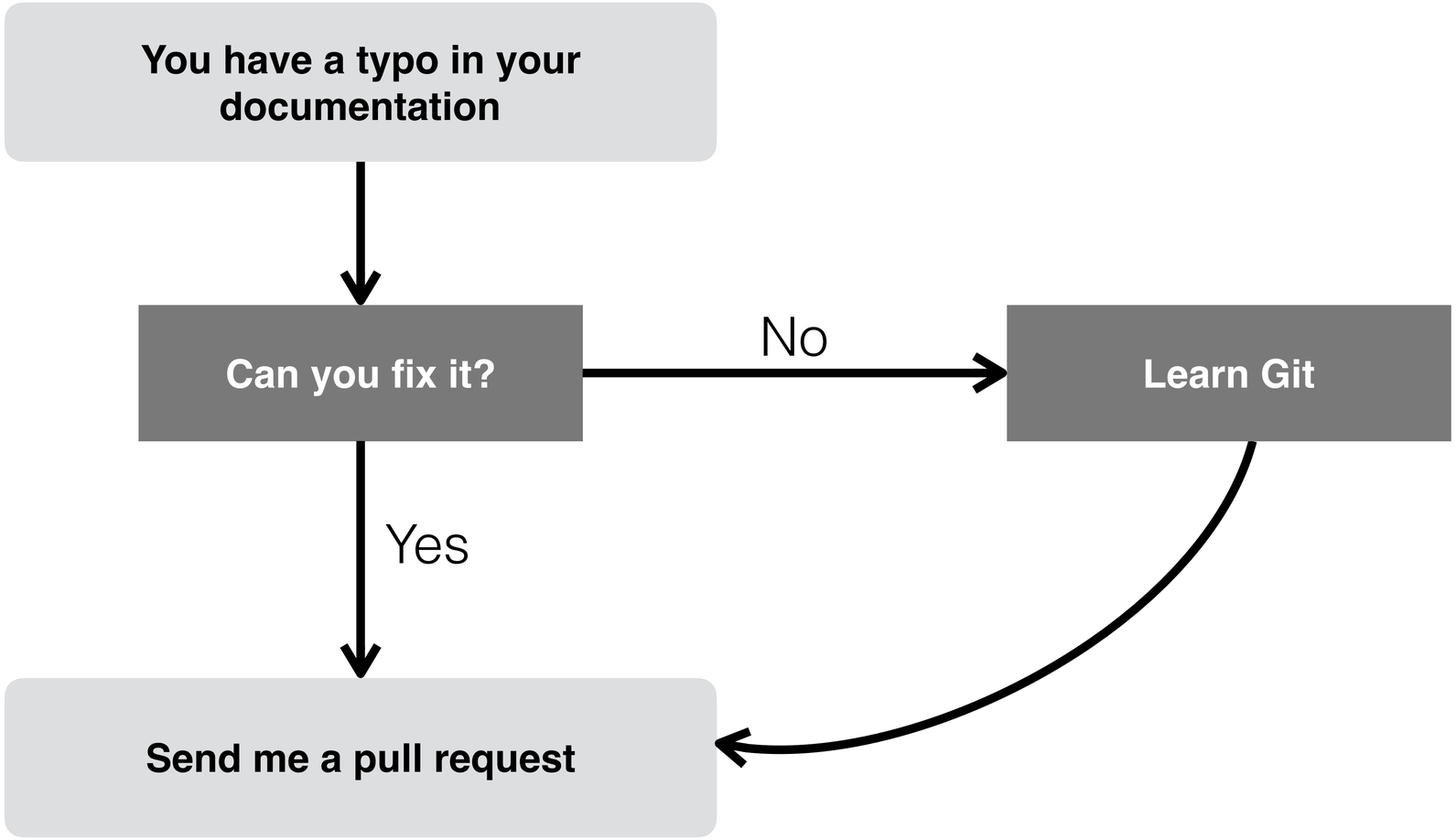
My first “contribution”



every contribution counts…

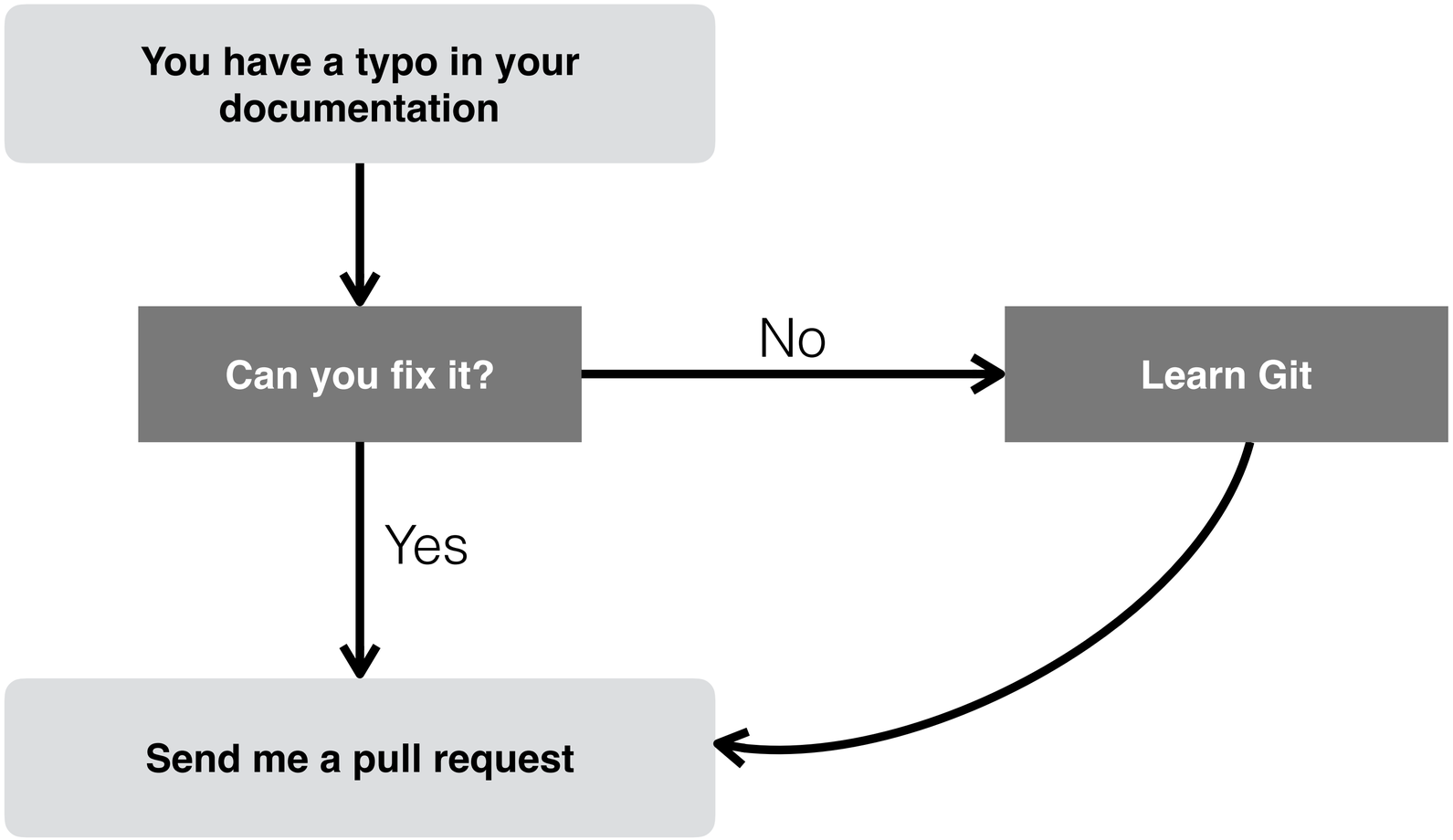


tpyos

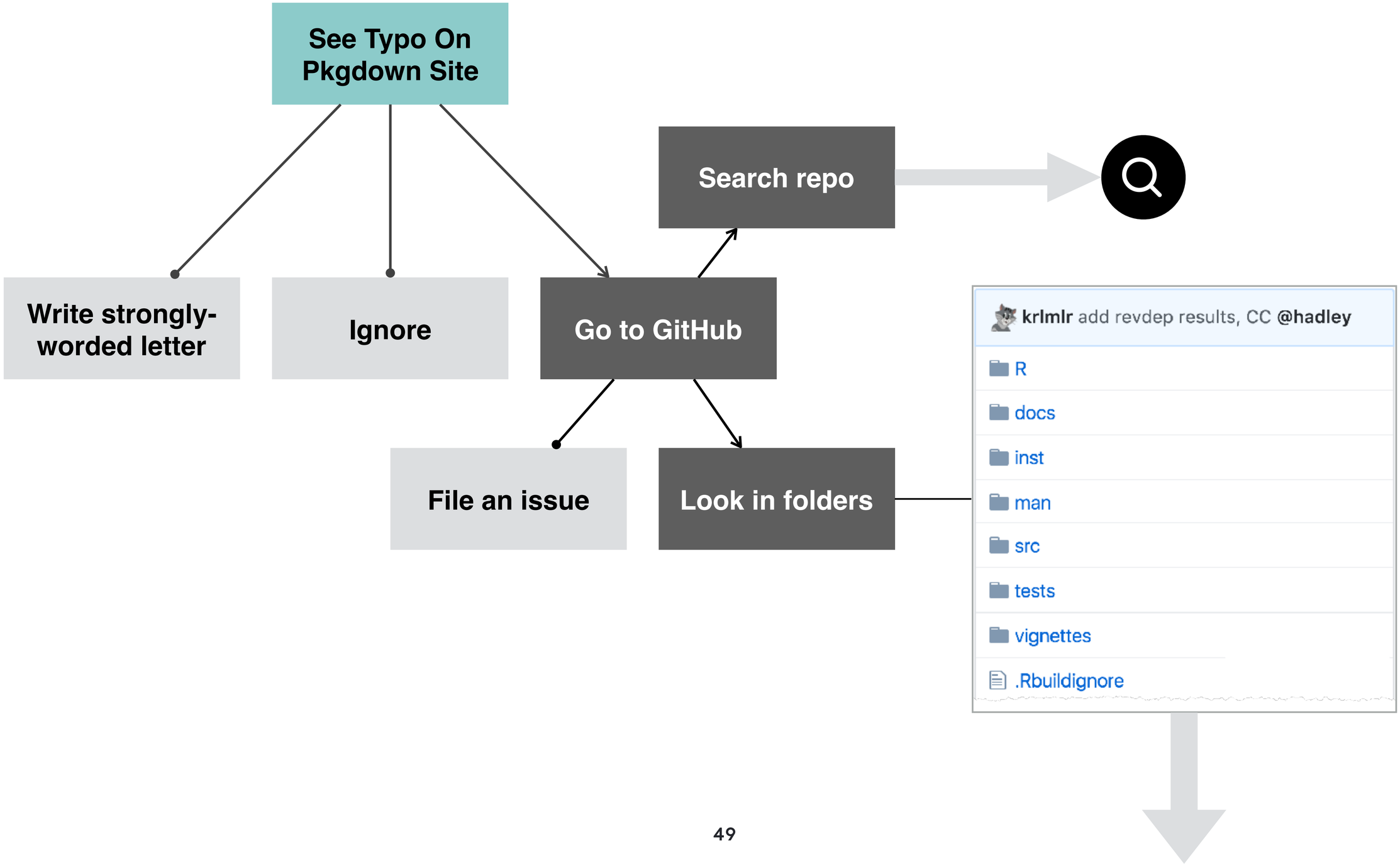


Adapted from: You Do Not Need to Tell Me I Have A Typo in My Documentation by Yihui Xie

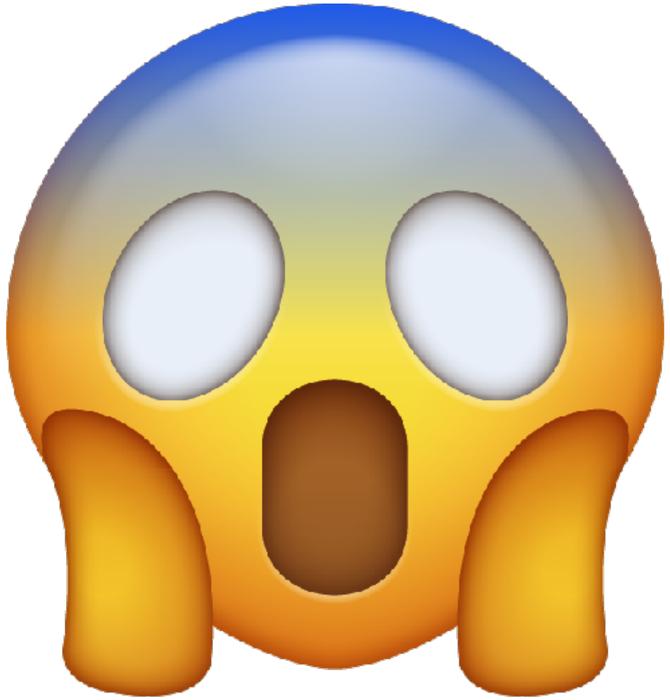
typos

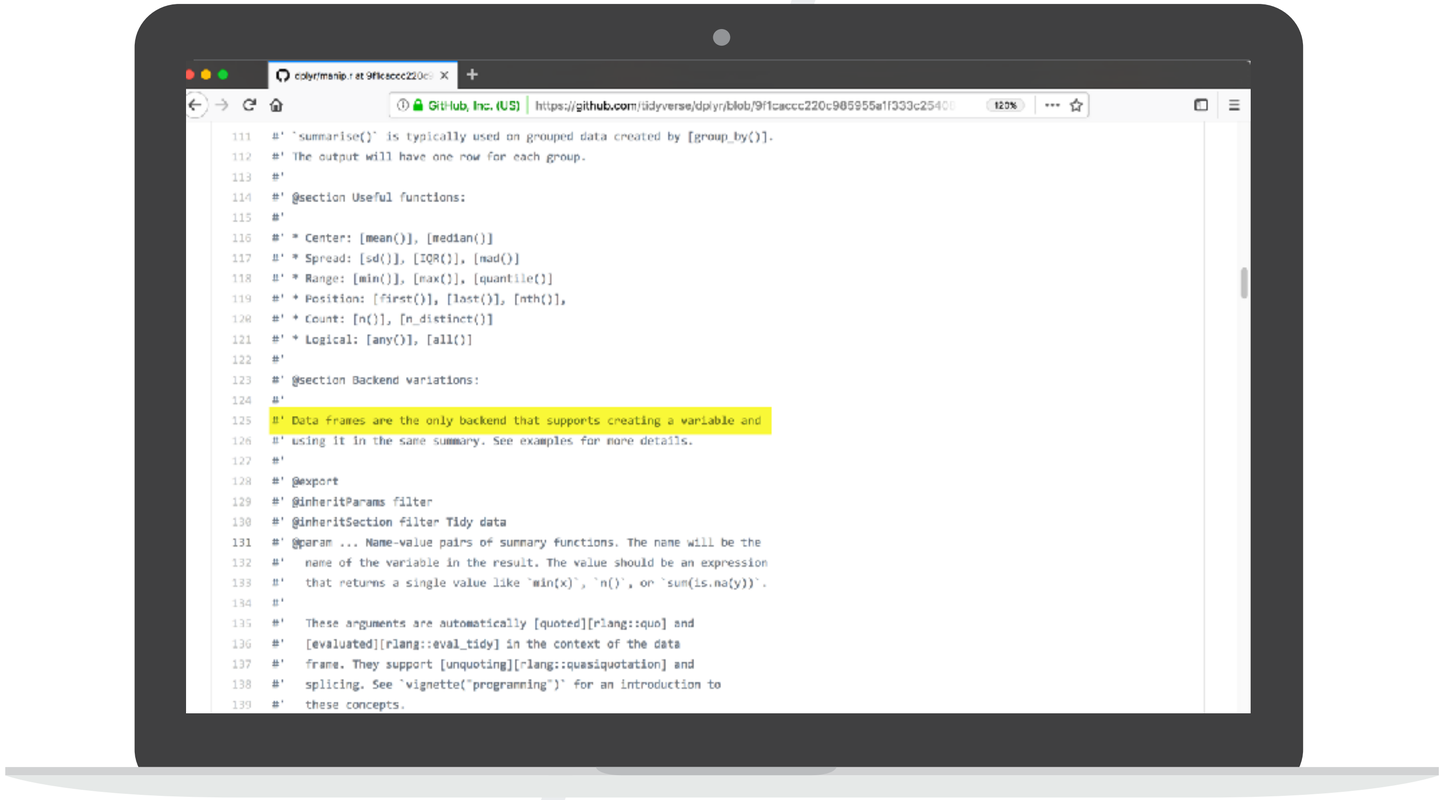


Adapted from: You Do Not Need to Tell Me I Have A Typo in My Documentation by Yihui Xie

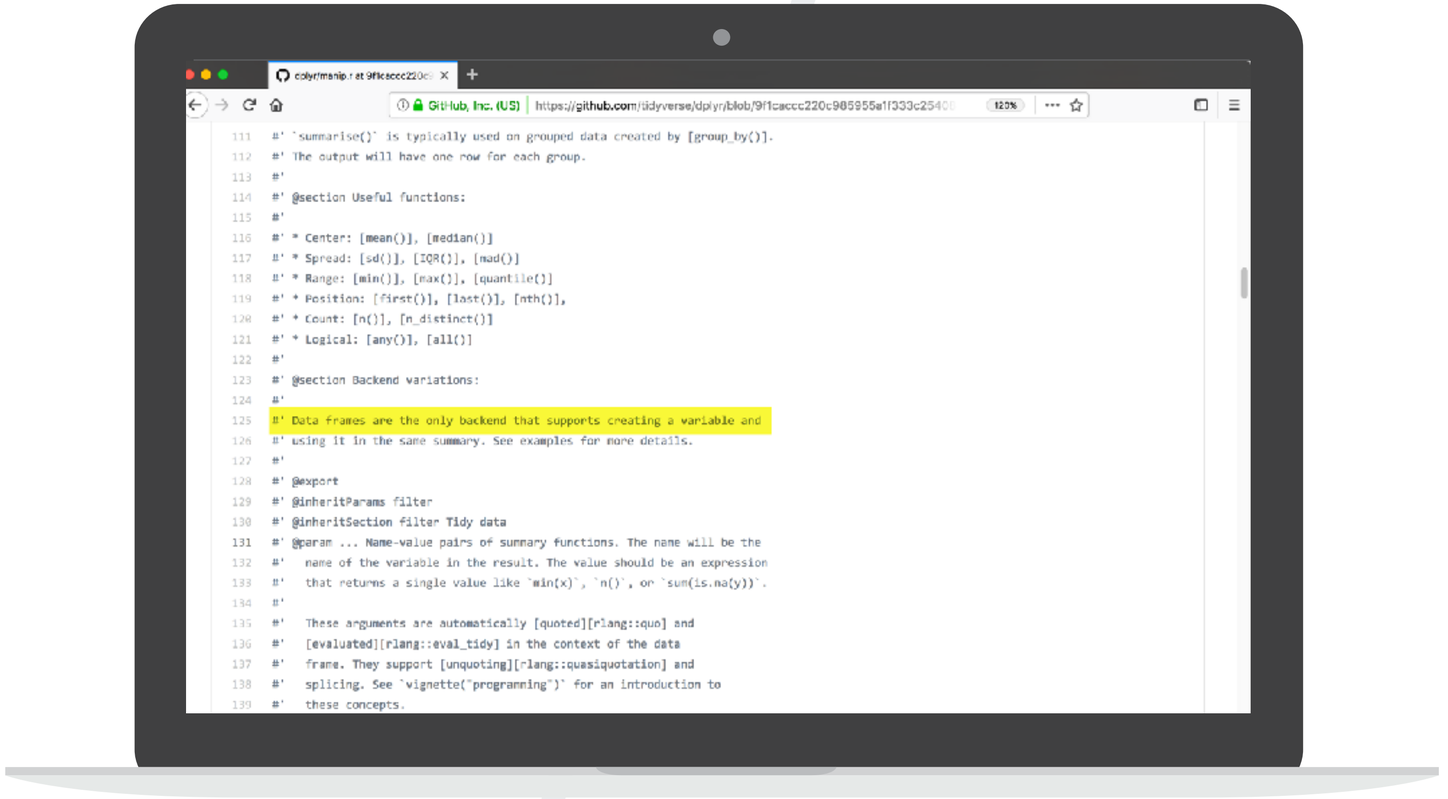

See Typo On Pkgdown Site
Write strongly-worded letter
Ignore
Go to GitHub
Search repo
File an issue
Look in folders
krlmlr add revdep results, CC @hadley
R
docs
inst
man
src
tests
vignettes
.Rbuildignore

Source code?!



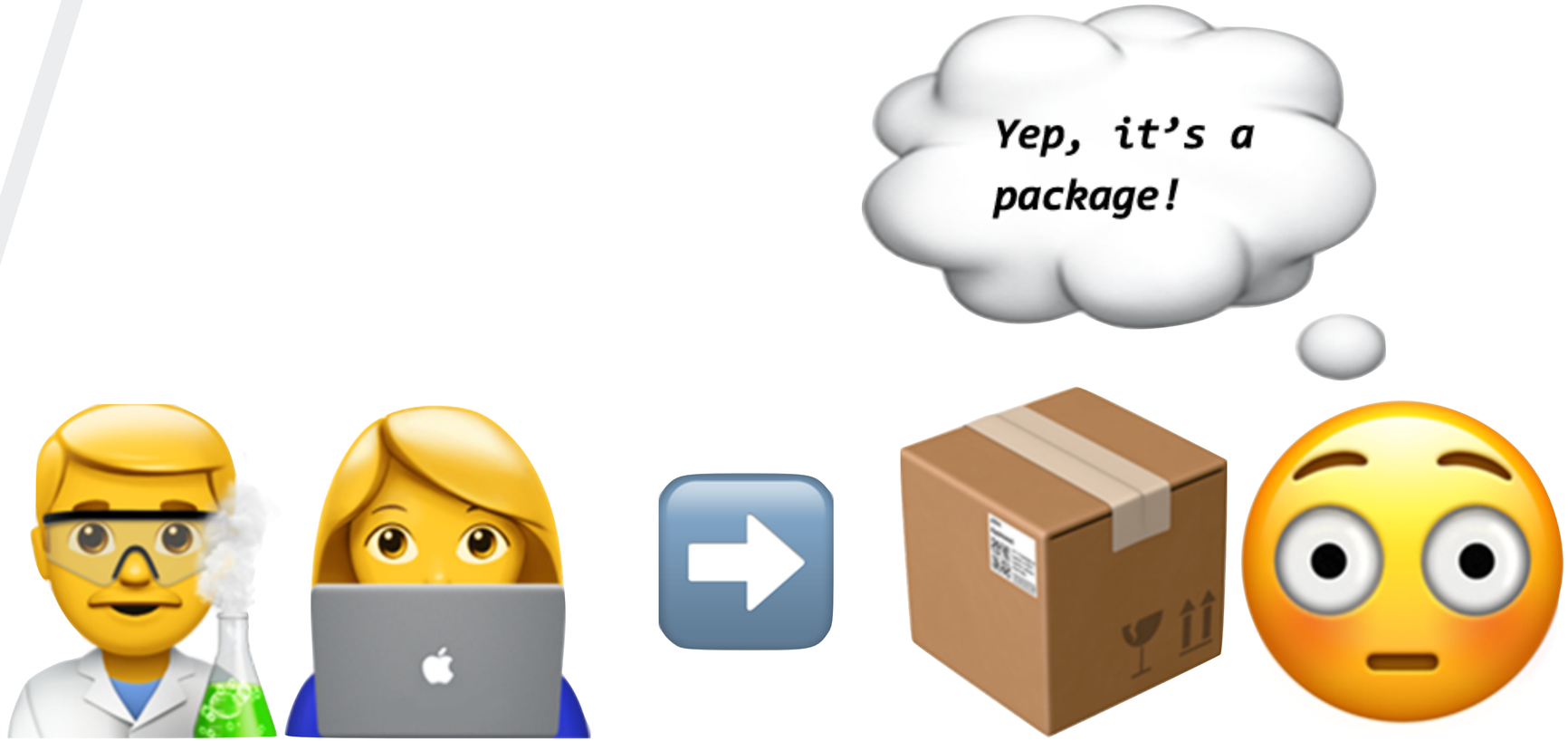

GitHub, Inc. (US)
https://github.com/tidyverse/dplyr/blob/9f1caccc220c965955af333c25408
#' `summarise()` is typically used on grouped data created by [group_by()].
#' The output will have one row for each group.
#'
#' @section Useful functions:
#'
#' * Center: [mean()], [median()]
#' * Spread: [sd()], [IQR()], [mad()]
#' * Range: [min()], [max()], [quantile()]
#' * Position: [first()], [last()], [nth()],
#' * Count: [n()], [n_distinct()]
#' * Logical: [any()], [all()]
#'
#' @section Backend variations:
#'
#' Data frames are the only backend that supports creating a variable and
#' using it in the same summary. See examples for more details.
#'
#' @export
#' @inheritParams filter
#' @inheritSection filter Tidy data
#' @param ... Name-value pairs of summary functions. The name will be the
#'   name of the variable in the result. The value should be an expression
#'   that returns a single value like `min(x)`, `n()`, or `sum(is.na(y))`.
#'
#'   These arguments are automatically [quoted][rlang::quo] and
#'   [evaluated][rlang::eval_tidy] in the context of the data
#'   frame. They support [unquoting][rlang::quasiquotation] and
#'   splicing. See `vignette("programming")` for an introduction to
#'   these concepts.

Source code?!


GitHub, Inc. (US)
https://github.com/tidyverse/dplyr/blob/9f1caccc220c985955af333c25408
#' `summarise()` is typically used on grouped data created by [group_by()].
#' The output will have one row for each group.
#'
#' @section Useful functions:
#'
#' * Center: [mean()], [median()]
#' * Spread: [sd()], [IQR()], [mad()]
#' * Range: [min()], [max()], [quantile()]
#' * Position: [first()], [last()], [nth()],
#' * Count: [n()], [n_distinct()]
#' * Logical: [any()], [all()]
#'
#' @section Backend variations:
#'
#' Data frames are the only backend that supports creating a variable and
#' using it in the same summary. See examples for more details.
#'
#' @export
#' @inheritParams filter
#' @inheritSection filter Tidy data
#' @param ... Name-value pairs of summary functions. The name will be the
#'   name of the variable in the result. The value should be an expression
#'   that returns a single value like `min(x)`, `n()`, or `sum(is.na(y))`.
#'
#'   These arguments are automatically [quoted][rlang::quo] and
#'   [evaluated][rlang::eval_tidy] in the context of the data
#'   frame. They support [unquoting][rlang::quasiquotation] and
#'   splicing. See `vignette("programming")` for an introduction to
#'   these concepts.

Go to the source…

require(n00bs)



require(n00bs)

When assigning reviewers to a submission, we aim to pair experienced reviewers with new ones, or reviewers with expertise on a package's programming methods with those experienced in its field of application.

How rOpenSci uses Code Review to Promote Reproducible Science

ropensci.org highlighted with Highly

BABY STEPS WITH MORE PACKAGES

ROXYGEN2

generates documentation from specially-formatted comments, used by all tidyverse packages

DEVTOOLS

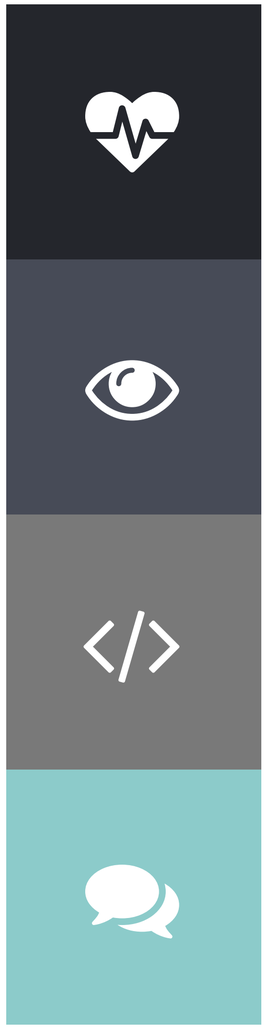
makes package development easier by providing R functions that simplify common tasks

TESTTHAT

provides functions that make it easy to create unit tests for R packages, used throughout tidyverse

roxygen2: In-Line Documentation for R. R package v 6.0.1. by Hadley Wickham, Peter Danenberg and Manuel Eugster (2017). https://CRAN.R-project.org/package=roxygen2

Hints for happy contributing in the tidyverse



GET THE PULSE OF A PROJECT

WATCH THE REPO

READ THE CODE

DISCUSS YOUR IDEAS

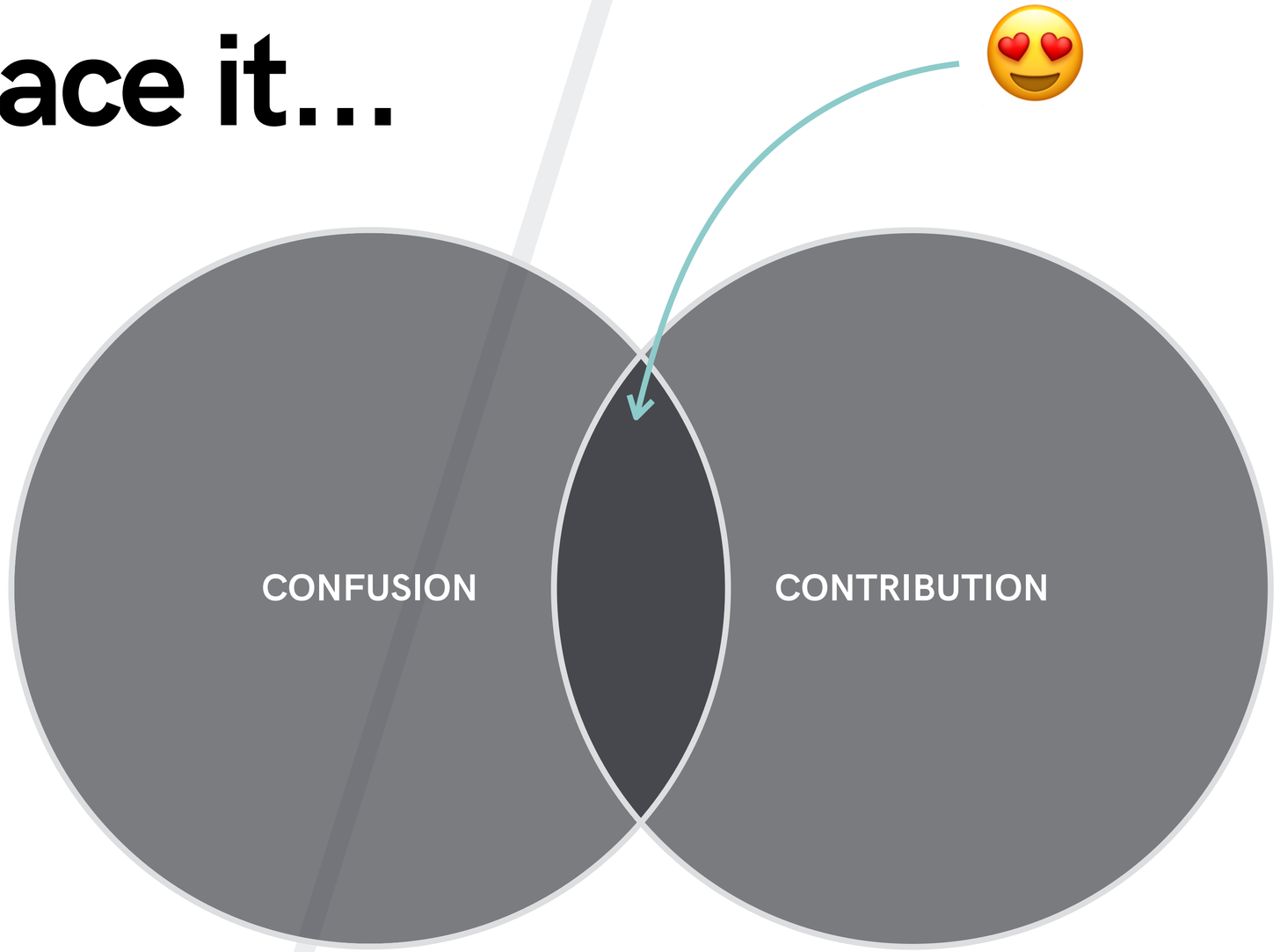
leaRn out loud

OMG, I just learned a thing!

give back


Acting out of the goodness of your heart, or something
Thanklessly Maintaining Open Source Software
O RLY?
@ThePracticalDev

embrace it...



Thank You

http://bit.ly/mara-ddtx






Thank You

http://bit.ly/mara-ddtx