

Creating Reproducible Data Science

Alex K. Gold
Solutions Engineer

 @alexkgold



github.com/akgold/2019-09-17_john_deere_webinar

Plan and Learning Objectives

Why care about reproducibility?

Or portability?



Why care about reproducibility?

Or portability?

Sharing!



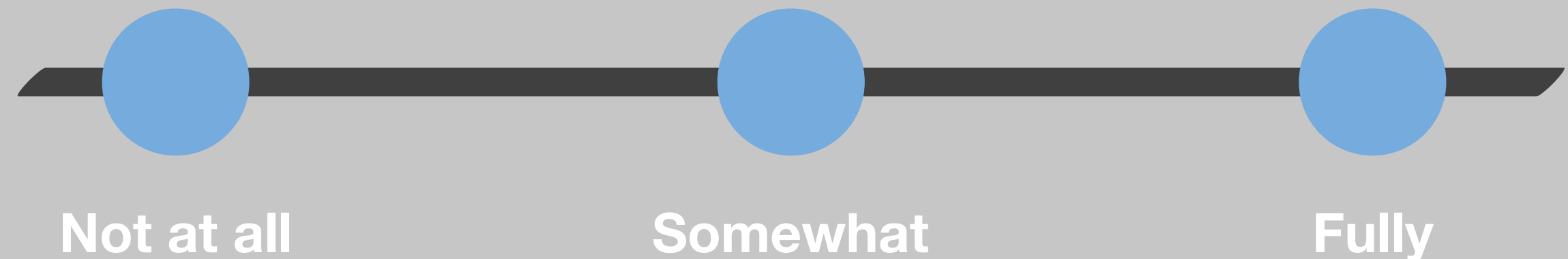
- With Colleagues
- With Future You
- They don't answer 📞

**How much
reproducibility
is enough?**

More Reproducible



More Work



A Taxonomy Of Irreproducibility

**Code that won't
run on someone
else's machine**



Difficulty Finding Latest Version



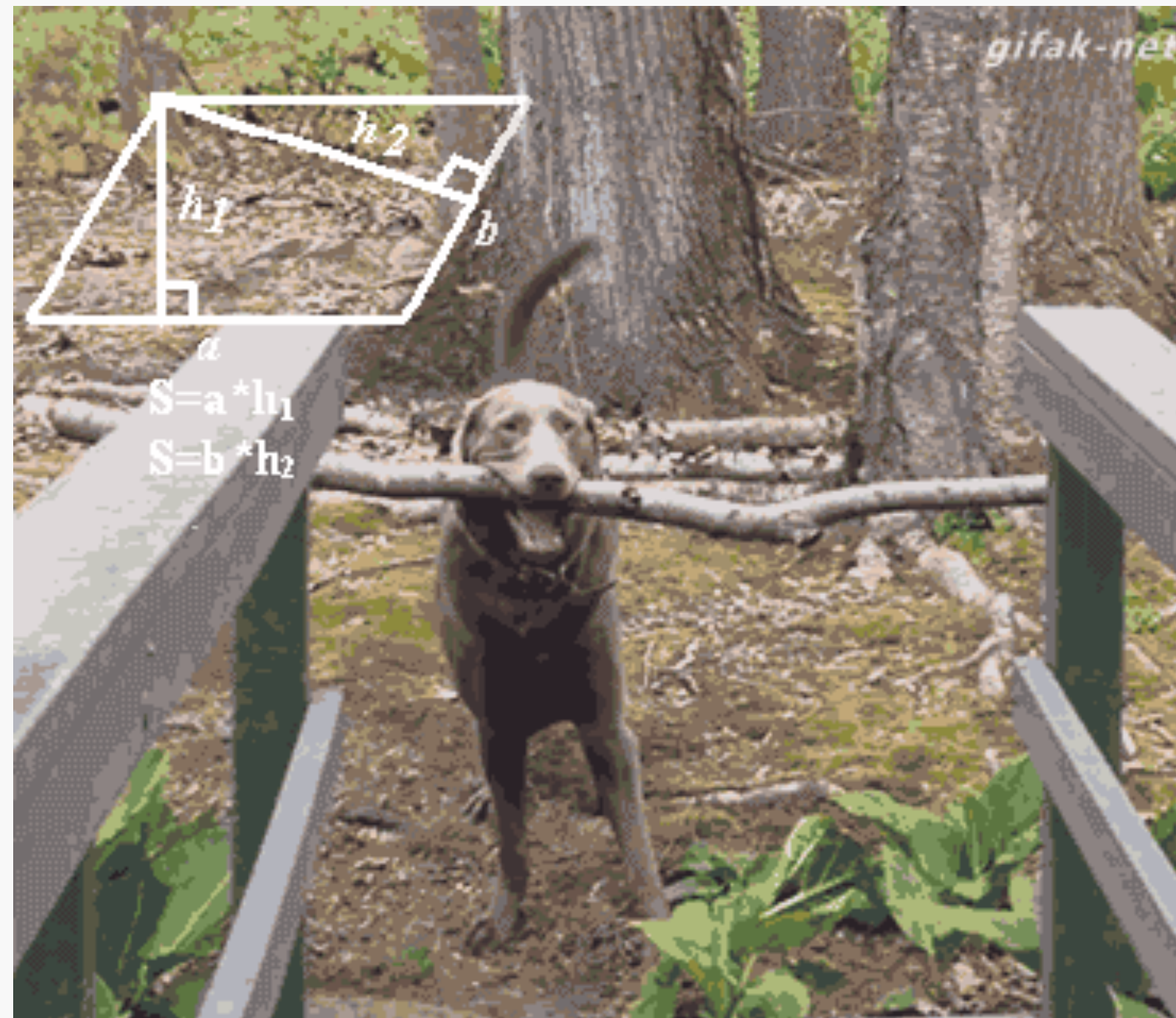
**Things that are
tedious and
you'll need again**



Environmental factors that will break



Solutions



**Code that breaks on someone
else's machine**

Code that breaks on someone else's machine



-based Workflow



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 🔥**



If the first line of your R script is

```
rm(list = ls())
```

**I will come into your office and
SET YOUR COMPUTER ON FIRE 🔥**

**Avoiding
Computer fires**

Project-based Workflow
and here :: here

**Avoiding
Computer fires**

Demo!

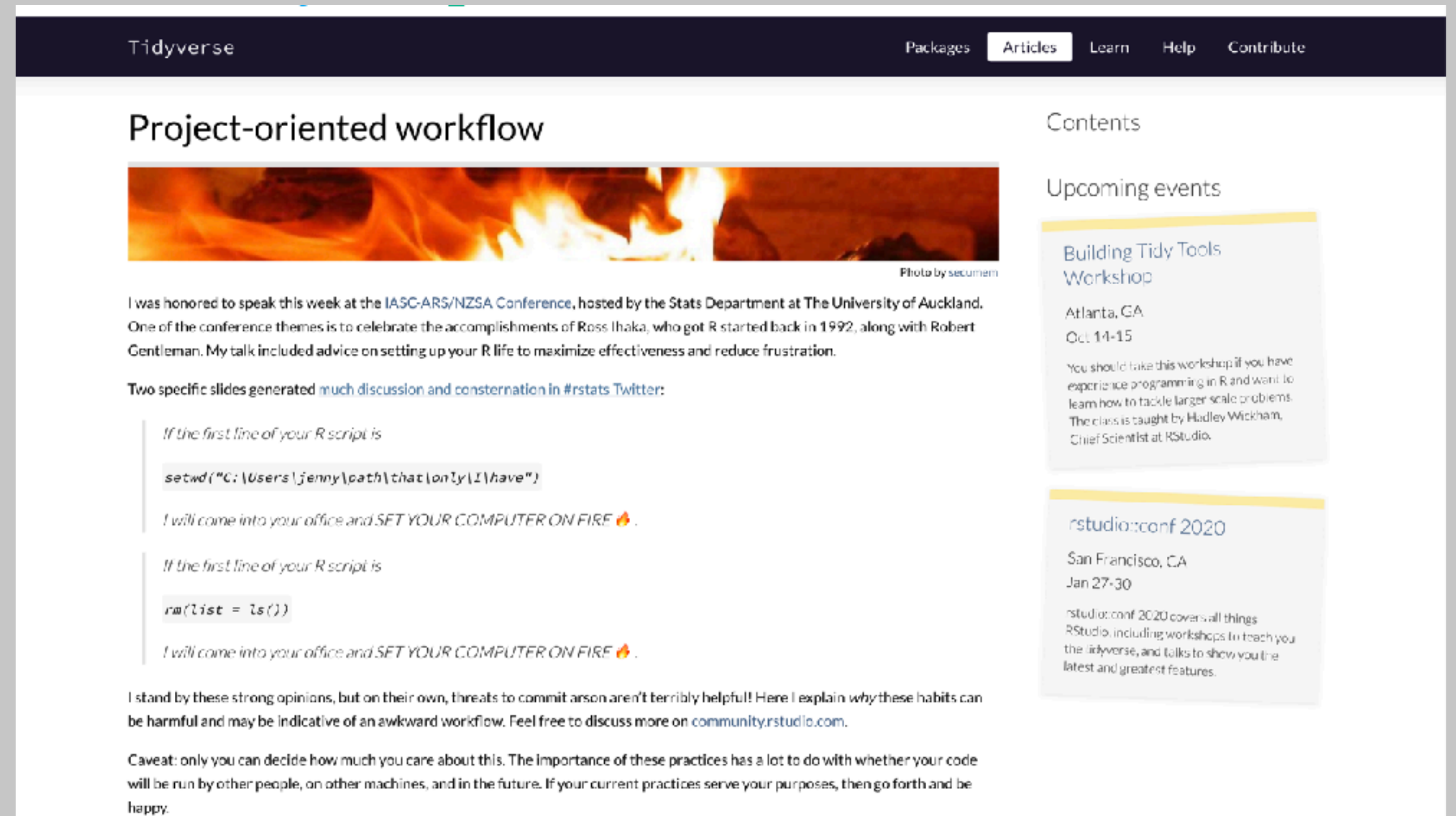
Code Structure

```
~/rstudio/2019-09-17_john_deere_webinar master tree dir_demo
dir_demo
├── README.md
├── code
│   ├── 00_data_cleaning.Rmd
│   ├── 00_data_cleaning.html
│   ├── 00_data_cleaning_files
│   ├── 01_model.Rmd
│   ├── 01_model.html
│   ├── 02_report.Rmd
│   ├── 02_report.html
│   └── 02_report_files
├── data
│   └── clean_flights.csv
├── models
│   └── log_reg.rds
└── plots
    └── delay_plot.png
```

Project-Based Workflow

Learn More

tidyverse.org/articles/2017/12/workflow-vs-script



The screenshot shows the Tidyverse website with a dark blue header containing navigation links: Packages, Articles (selected), Learn, Help, and Contribute. The main article is titled "Project-oriented workflow" and features a large image of fire. The text discusses the author's experience at the IASC-ARS/NZSA Conference and their stance on project-oriented workflows. It includes two code snippets with comments about setting the working directory and the potential for mischief. The article also mentions a talk at the conference and a caveat about the importance of these practices. On the right side, there are two event listings: "Building Tidy Tools Workshop" in Atlanta, GA, and "rstudio::conf 2020" in San Francisco, CA.

Tidyverse Packages Articles Learn Help Contribute

Project-oriented workflow

Photo by secumem

I was honored to speak this week at the [IASC-ARS/NZSA Conference](#), hosted by the Stats Department at The University of Auckland. One of the conference themes is to celebrate the accomplishments of Ross Ihaka, who got R started back in 1992, along with Robert Gentleman. My talk included advice on setting up your R life to maximize effectiveness and reduce frustration.

Two specific slides generated [much discussion and consternation in #rstats Twitter](#):

```
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 🔥.  
  
If the first line of your R script is  
rm(list = ls())  
  
I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.
```

I stand by these strong opinions, but on their own, threats to commit arson aren't terribly helpful! Here I explain *why* these habits can be harmful and may be indicative of an awkward workflow. Feel free to discuss more on [community.rstudio.com](#).

Caveat: only you can decide how much you care about this. The importance of these practices has a lot to do with whether your code will be run by other people, on other machines, and in the future. If your current practices serve your purposes, then go forth and be happy.

Contents

Upcoming events

Building Tidy Tools Workshop

Atlanta, GA
Oct 14-15

You should take this workshop if you have experience programming in R and want to learn how to tackle larger scale problems. The class is taught by Hadley Wickham, Chief Scientist at RStudio.

rstudio::conf 2020

San Francisco, CA
Jan 27-30

rstudio::conf 2020 covers all things RStudio, including workshops to teach you the tidyverse, and talks to show you the latest and greatest features.

Difficulty Finding Latest Version

Difficulty Finding Latest Version



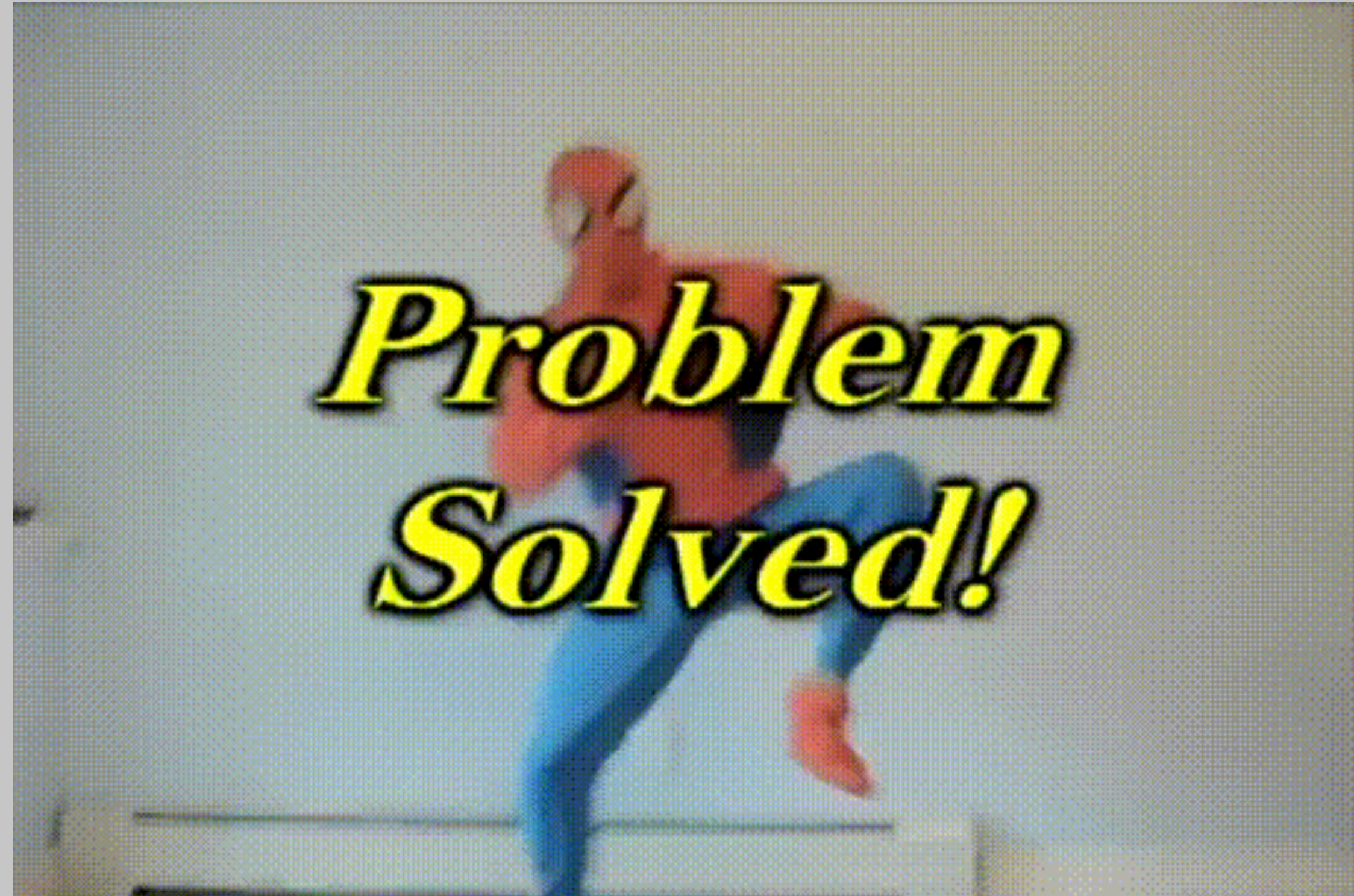
Version Control

Version Control



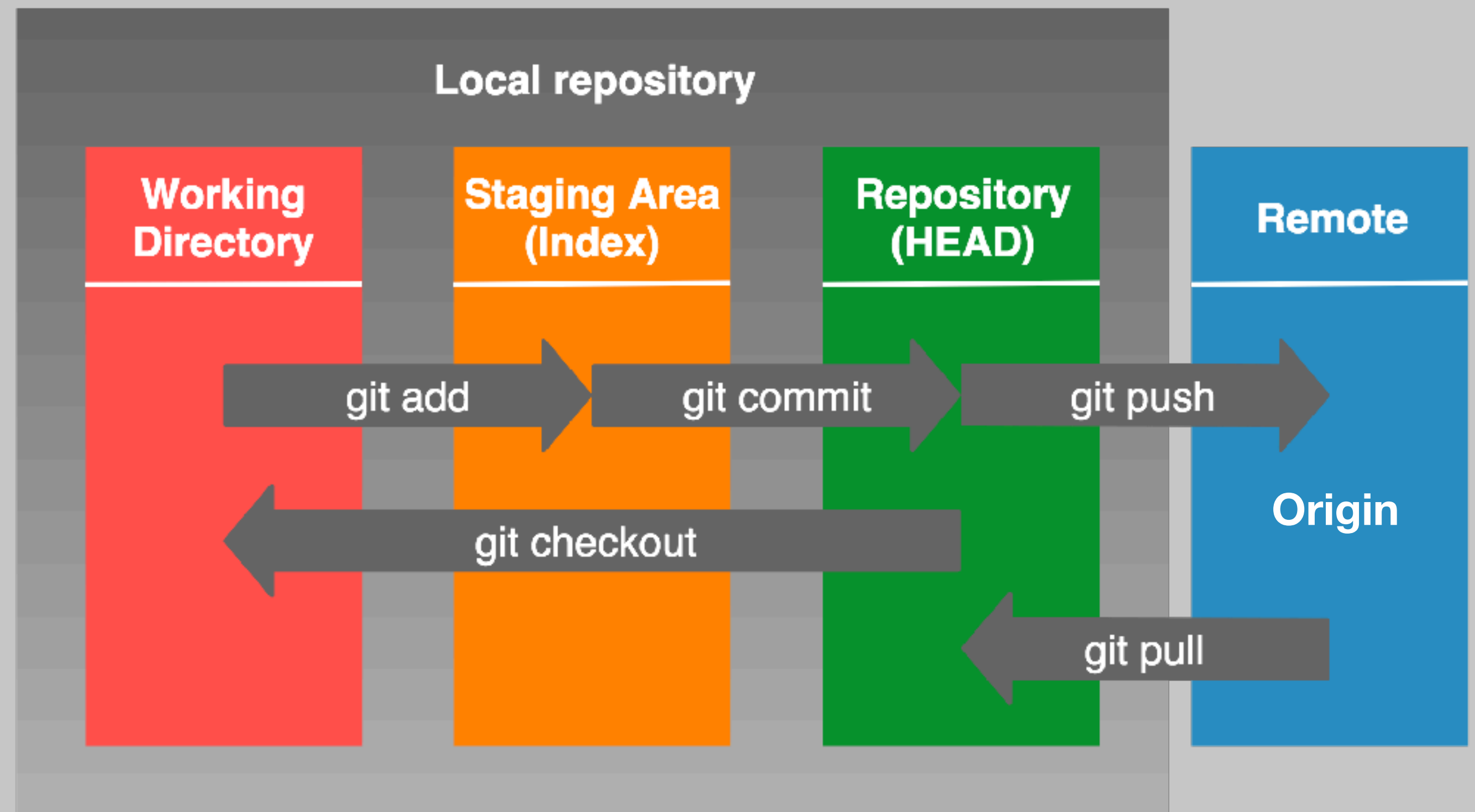
Version Control

Why would I use it?



Version Control

A few terms.
(Sorry)



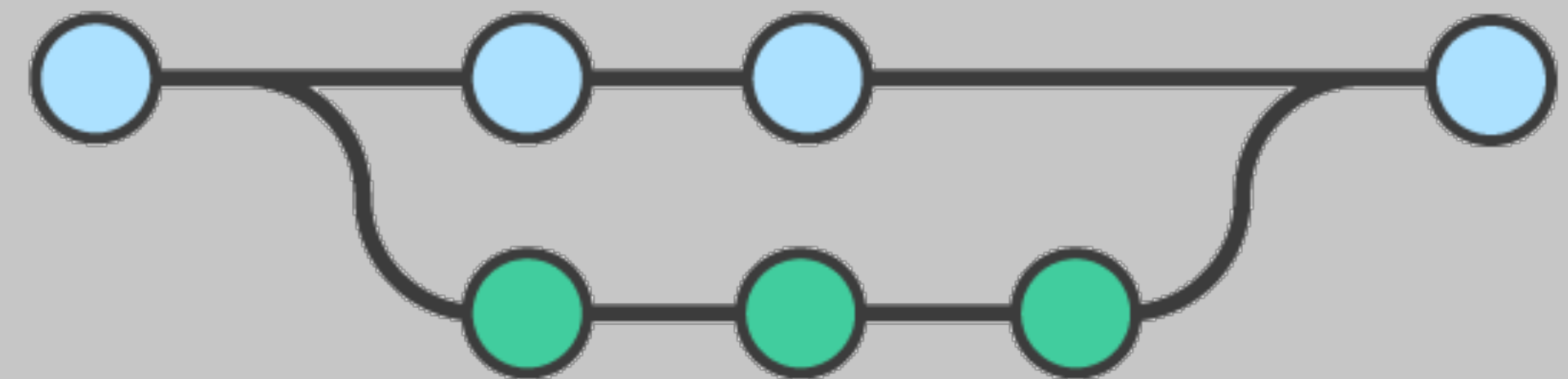
Cute dog
break.



Version Control

Branching

Master

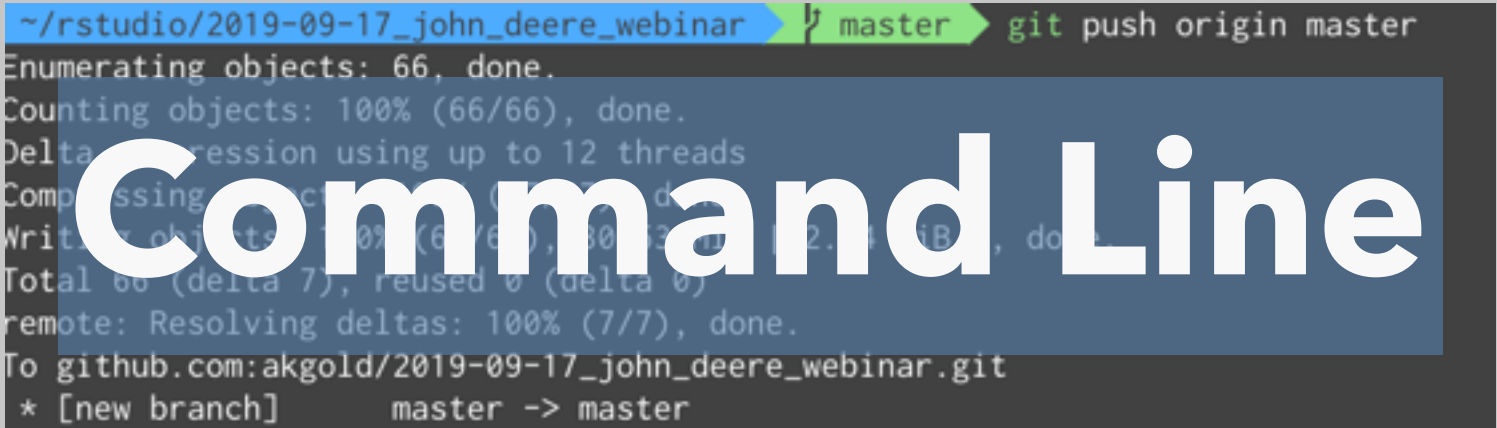
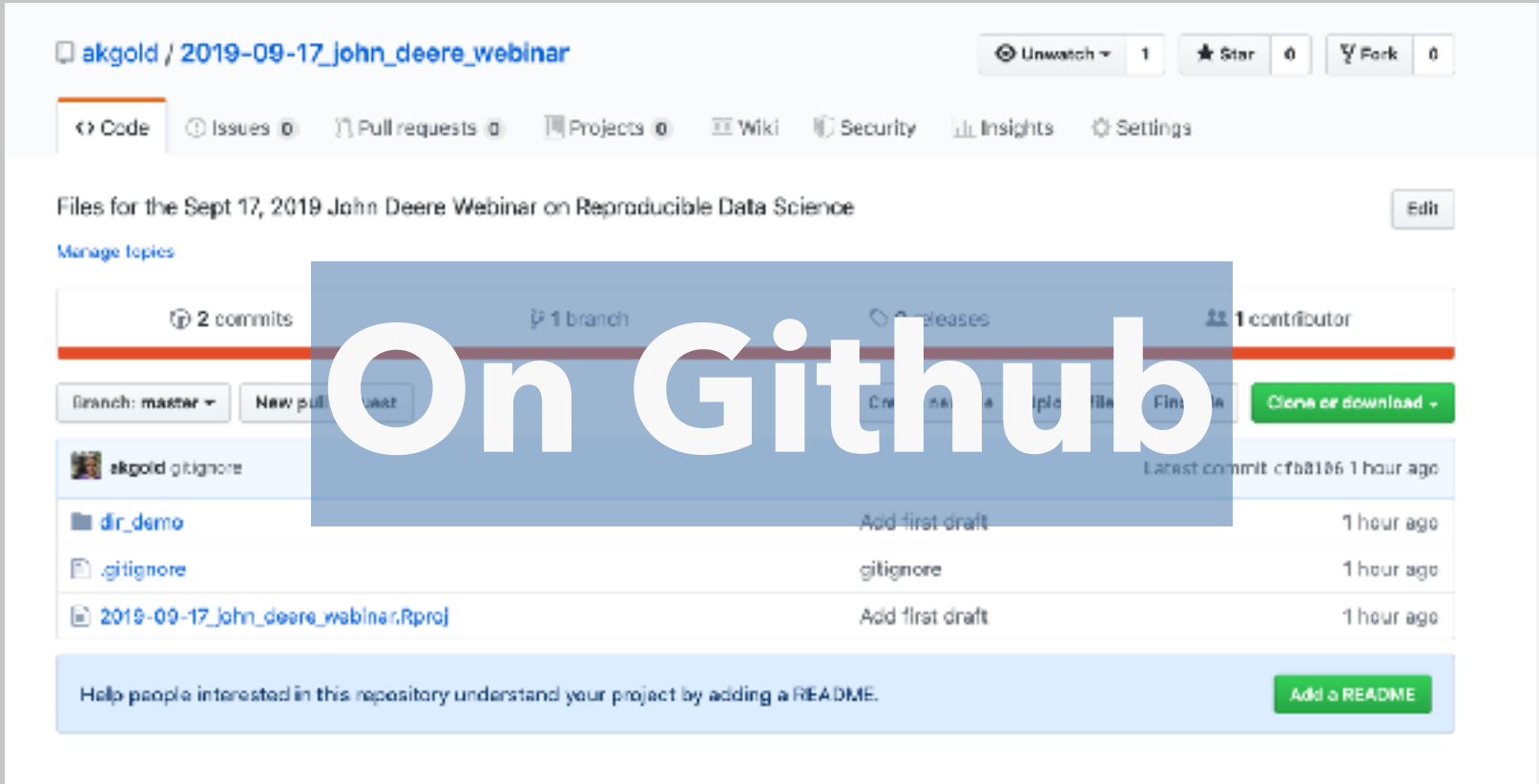
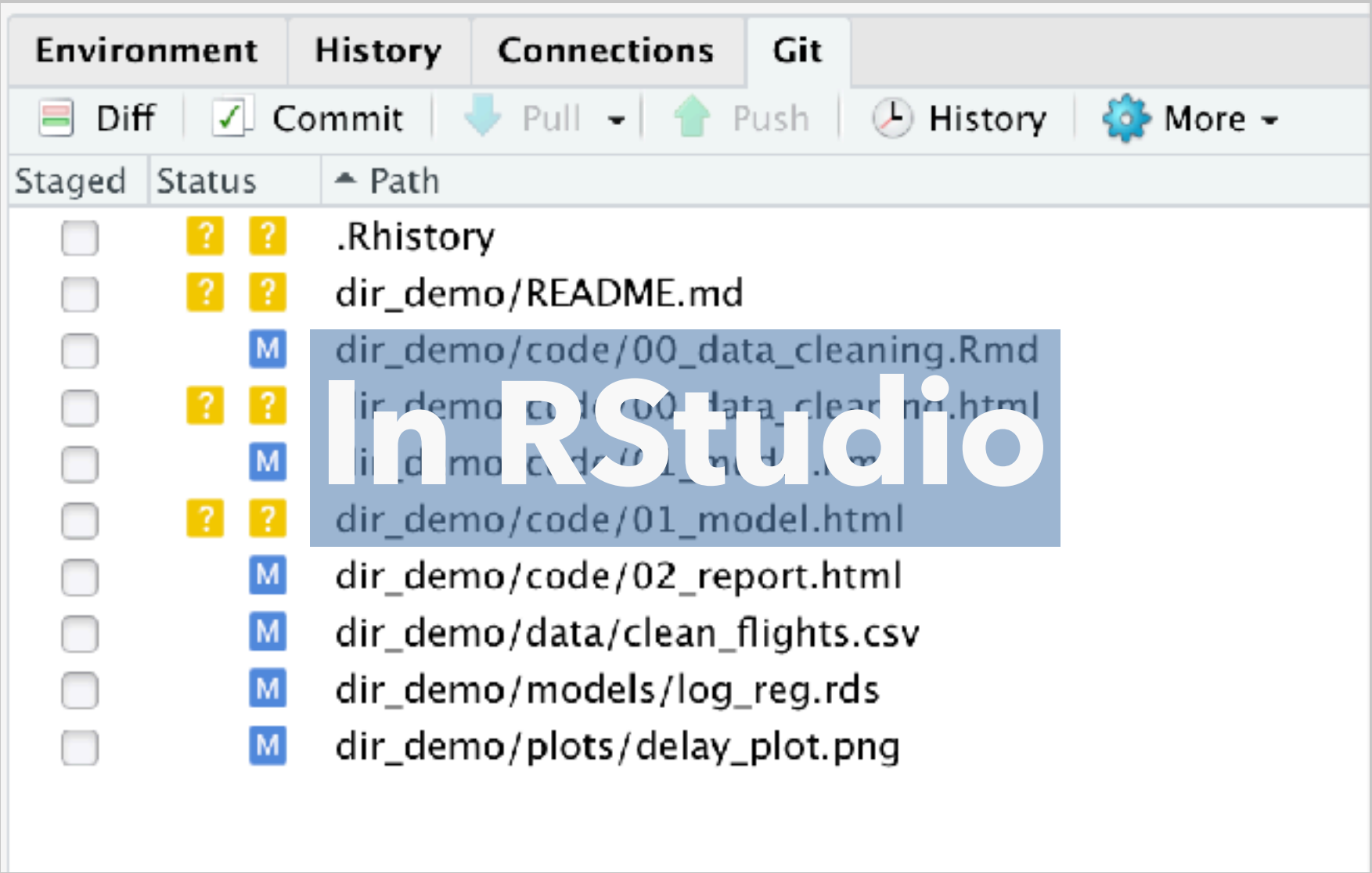


“Feature”

Git vs Github



Where can I git?

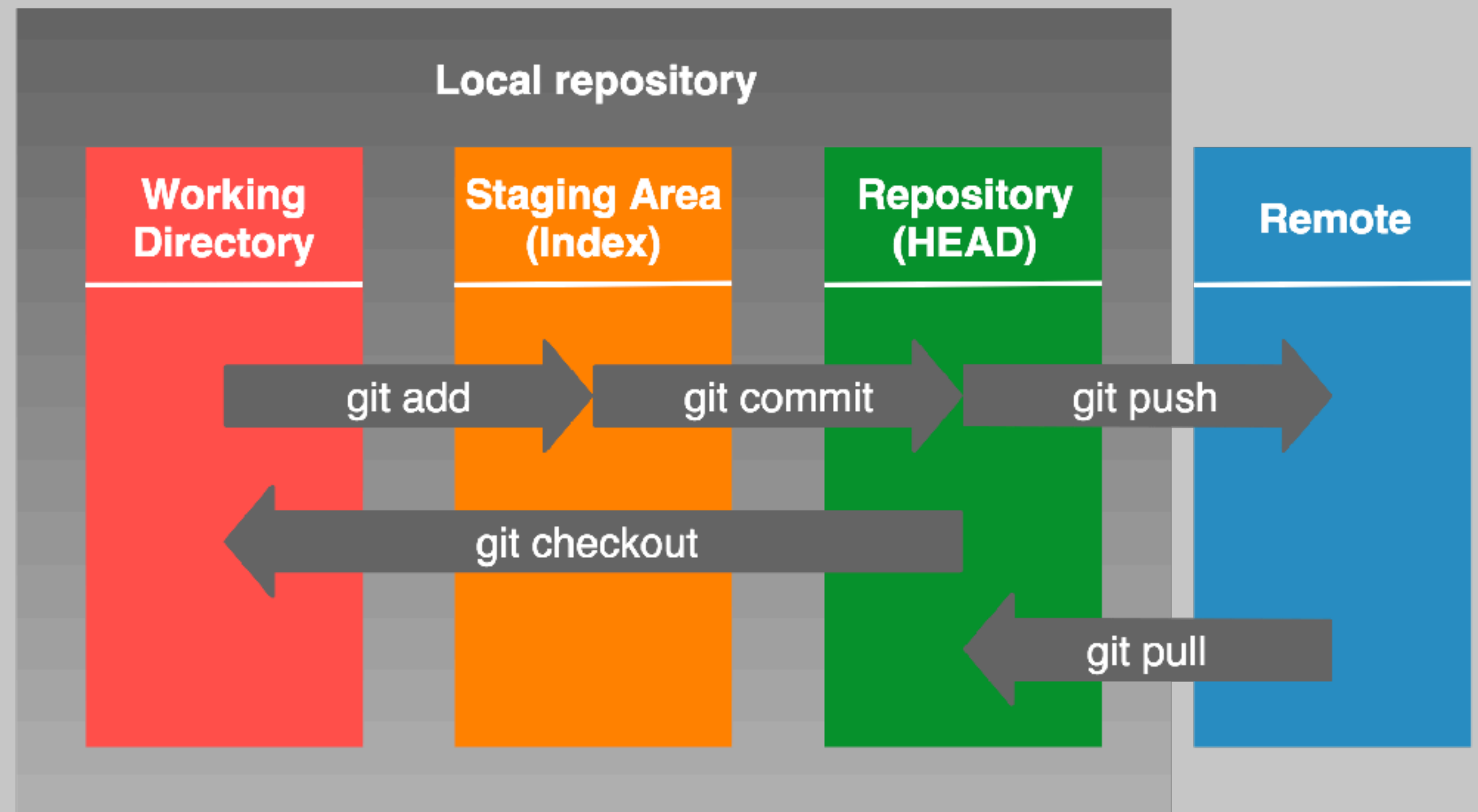


Version Control

Demo!

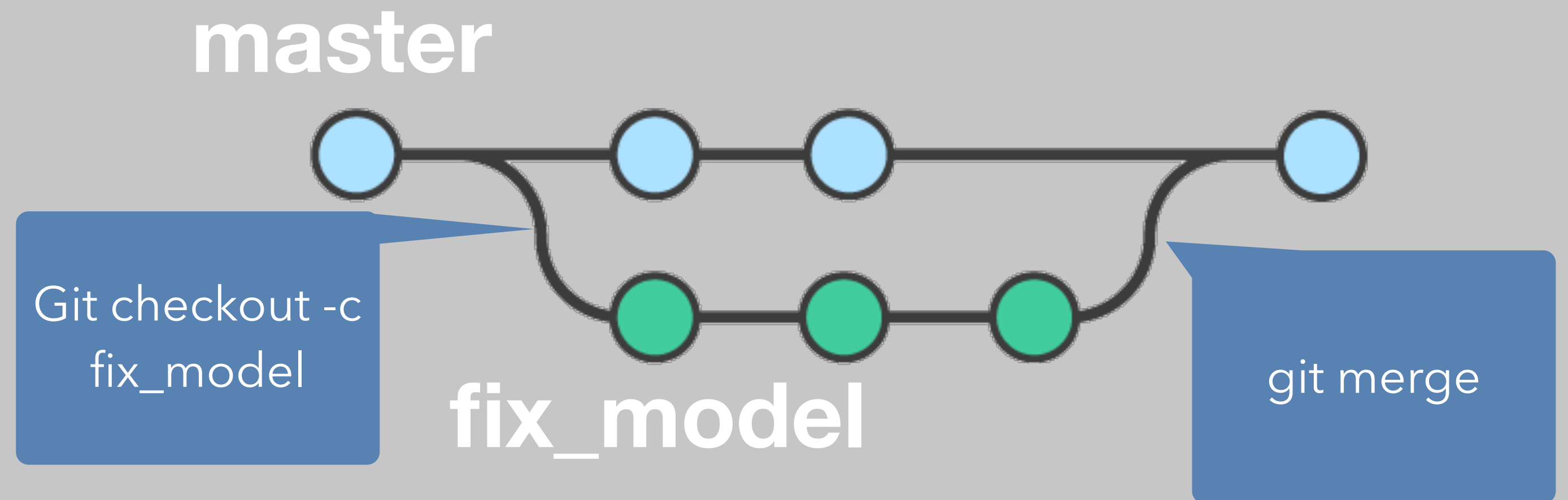
Version Control

Command Review



Version Control

Branching



Version Control

Learn More

www.happygitwithr.com



**Things that are tedious and
you'll need again**

Things that are tedious and you'll need again

R Tools





David Robinson

@drob

When you've written the same code 3 times, write a function

When you've given the same in-person advice 3 times, write a blog post

9:22 PM · Nov 8, 2017 · [Twitter for iPhone](#)

When you've used the same

- function
- RMarkdown document
- boilerplate Shiny code

3x

write a package



**Code Snippets,
Functions,
And Templates**

Demo!



R Packages

r-pkgs.had.co.nz



**Environmental factors that
will break**

Environmental factors that will break

Controlling
Environments

**Why would my
environment
break?**



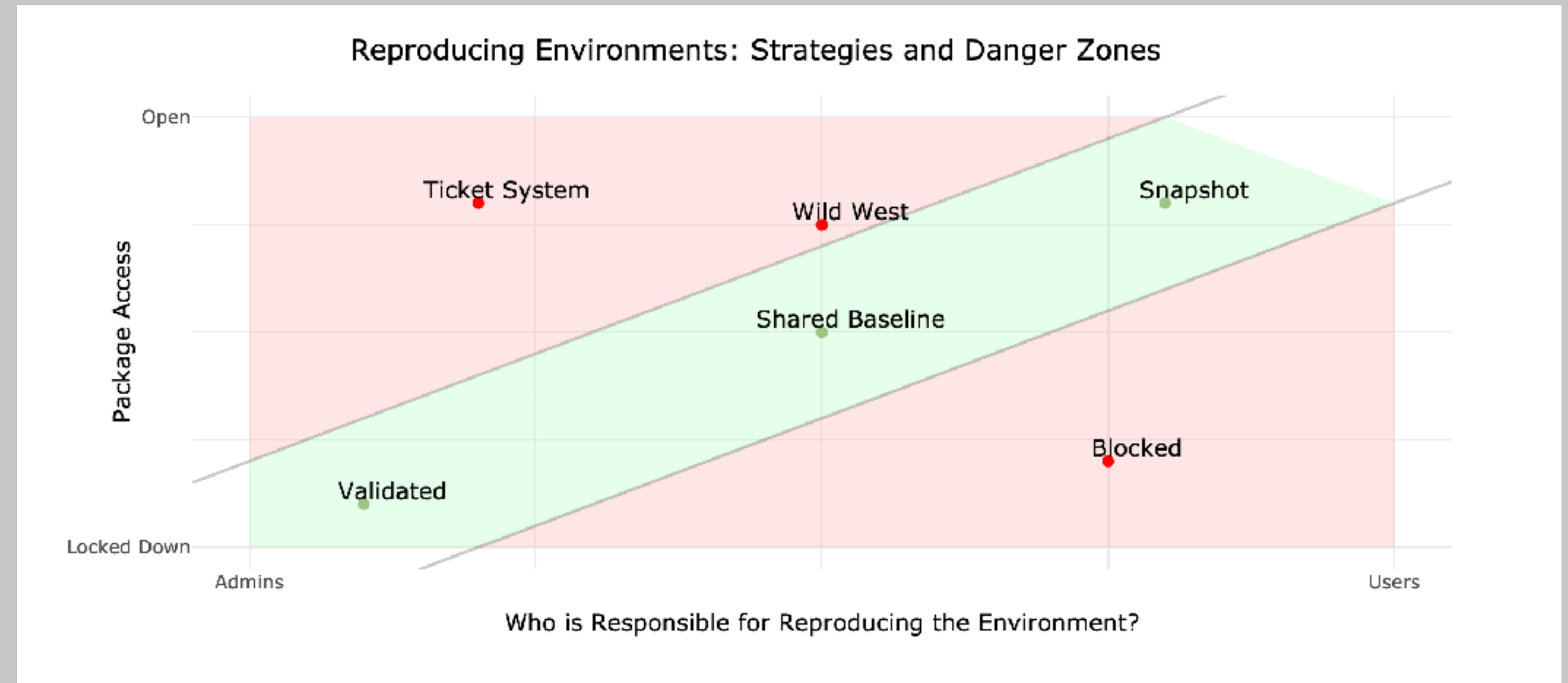
Reproducing Environments

(Advanced Reproducibility)

**Save package state
using packrat/renv**

Reproducing Environments

(Advanced Reproducibility)



Reproducing Environments

Learn More

environments.rstudio.com



The End

Code only for your machine

- ★ -based workflow

- ★ tidyverse.org/articles/2017/12/workflow-vs-script

Trying to find and coordinate versions

- ★ Use 

- ★ happygitwithr.com

Reusing tedious work

- ★ Write a 

- ★ r-pkgs.had.co.nz

Reproducing environments

- ★ [packrat/renv](https://packrat.rstudio.com/)

- ★ environments.rstudio.com

