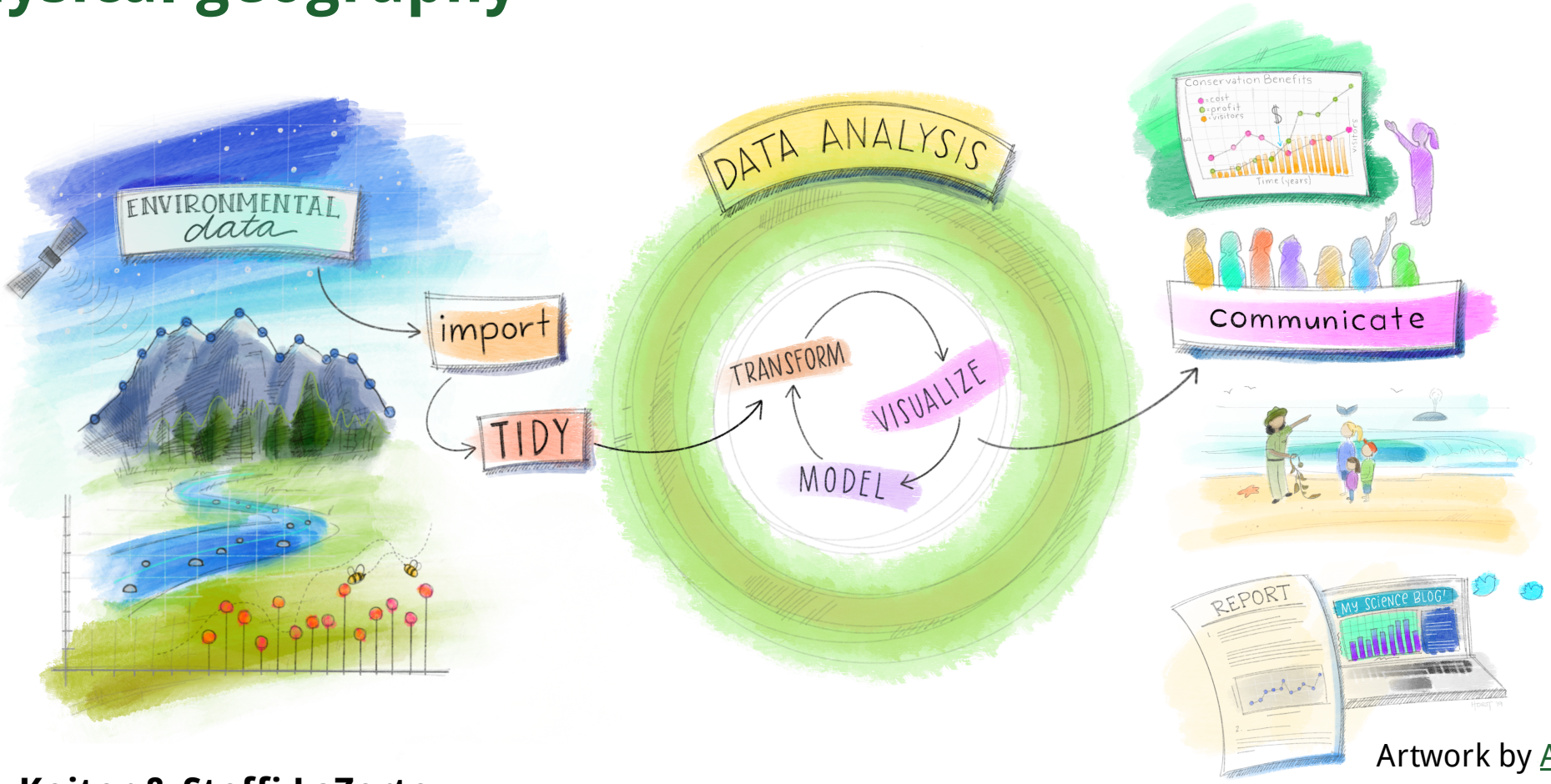
38:279 Introduction to Geographic Research Methods

Importance of reproducible data analysis in physical geography



Artwork by Allison Horst

Updated from Grolemund & Wickham's classic R4DS schematic, envisioned by Dr. Julia Lowndes for her 2019 useR! keynote talk

Alex Koiter & Steffi LaZerte

Introductions

Who am I?

Who are you?

- Preferred name
- Experience with research and/or data
- Interesting fact about yourself and/or something you are proud of having done or become

Introduction

Reproducibility and Replicability

What's the difference?

Introduction

Reproducibility

- The data analysis can be successfully reproduced

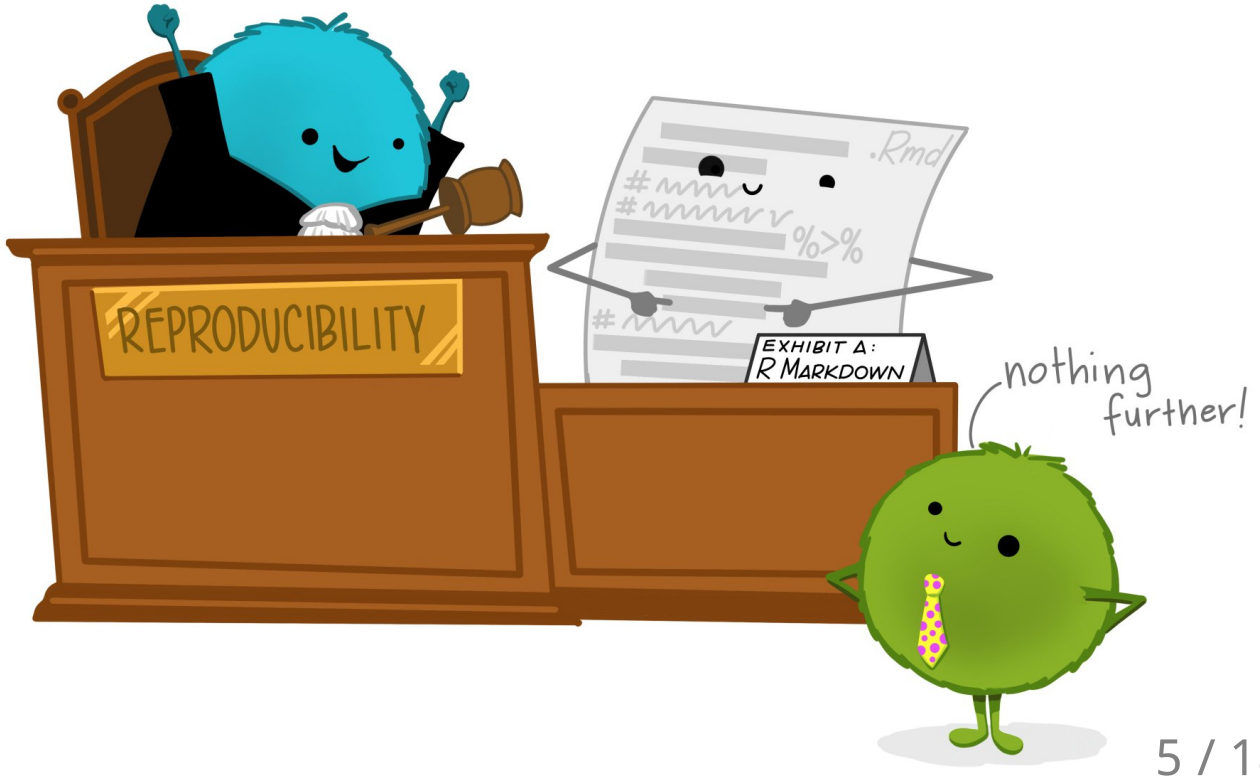
Replicability

- A separate investigator conducted an independent study and came to the same conclusion as the original study

Reproducibility

- Less about ensuring the correctness of the results
- More about being transparent and understanding exactly what was done
  - This is especially important in large and complex datasets

A study can be reproducible and still be wrong



Artwork by Allison Horst

Computational reproducibility

Important because:

- Allows us to evaluate the data, analyses, and models on which conclusions are drawn
- Allows you to revisit your own work (e.g., incorporate a suggestion)

It is difficult to reproduce because:

- Data is not made available
- Method sections of papers often do not provide enough detail
- Use of graphical programs (clicks and drop down menus)
- Not making code available (R, Python, MATLAB)

Working with large data sets

Data aquisition

- Documenting getting/downloading/importing data sets
  - Always maintain the original data (unmodified)

Working with large data sets

Data munging/wrangling

1. Formatting
2. Merging
3. Quality assurance
    - NA's
    - 0's
    - Detection limits
    - Outliers
    - Typos/errors

Need to document every change you make

Easy with small data sets with and simple structure

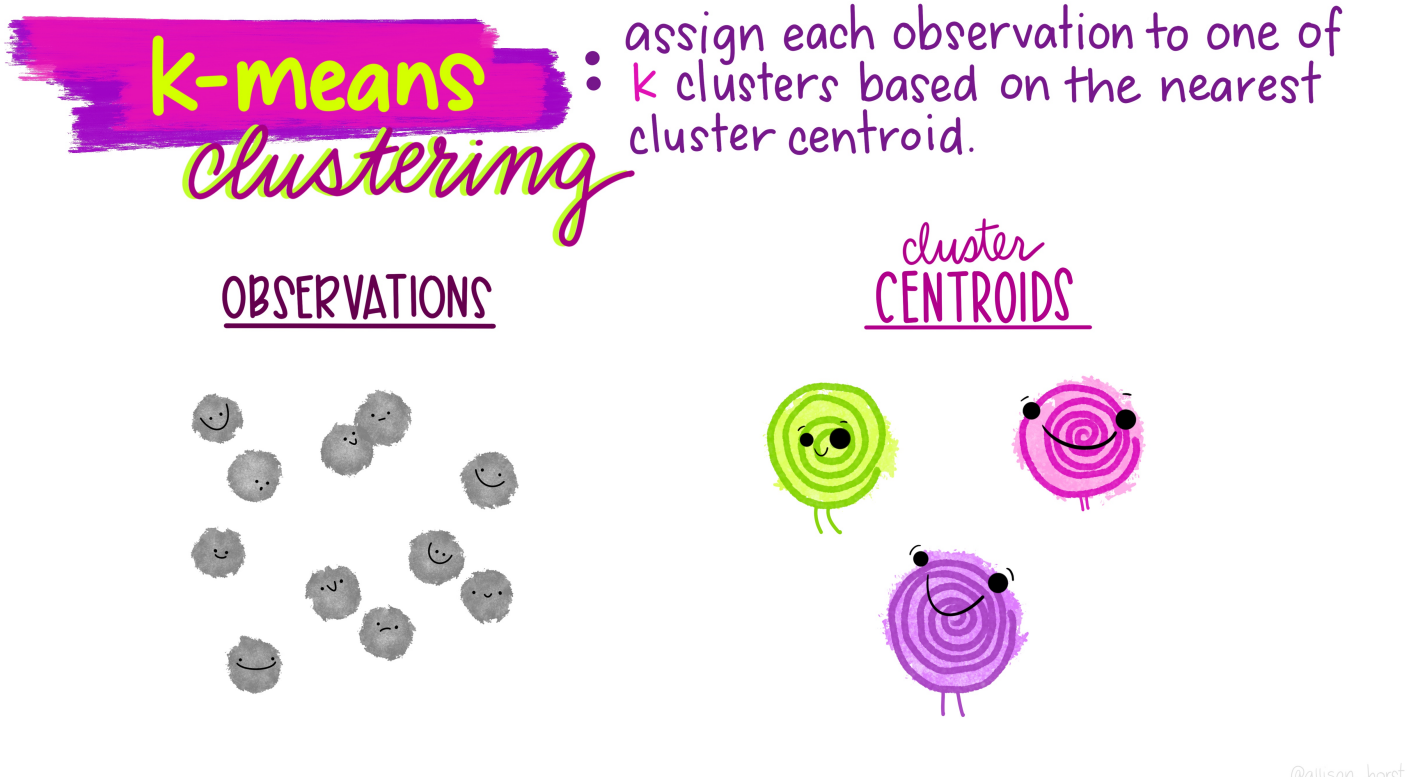
Artwork by Allison Horst

Working with large data sets

Data analysis

- Analysis/figures
  - Data used
  - Analysis used (trial and error)
  - Parameters
  - Diagnostics
  - Figure creation process
  - Software versions

Hard to write papers if you don't keep track of this!



Artwork by Allison Horst

Why is this not practiced?

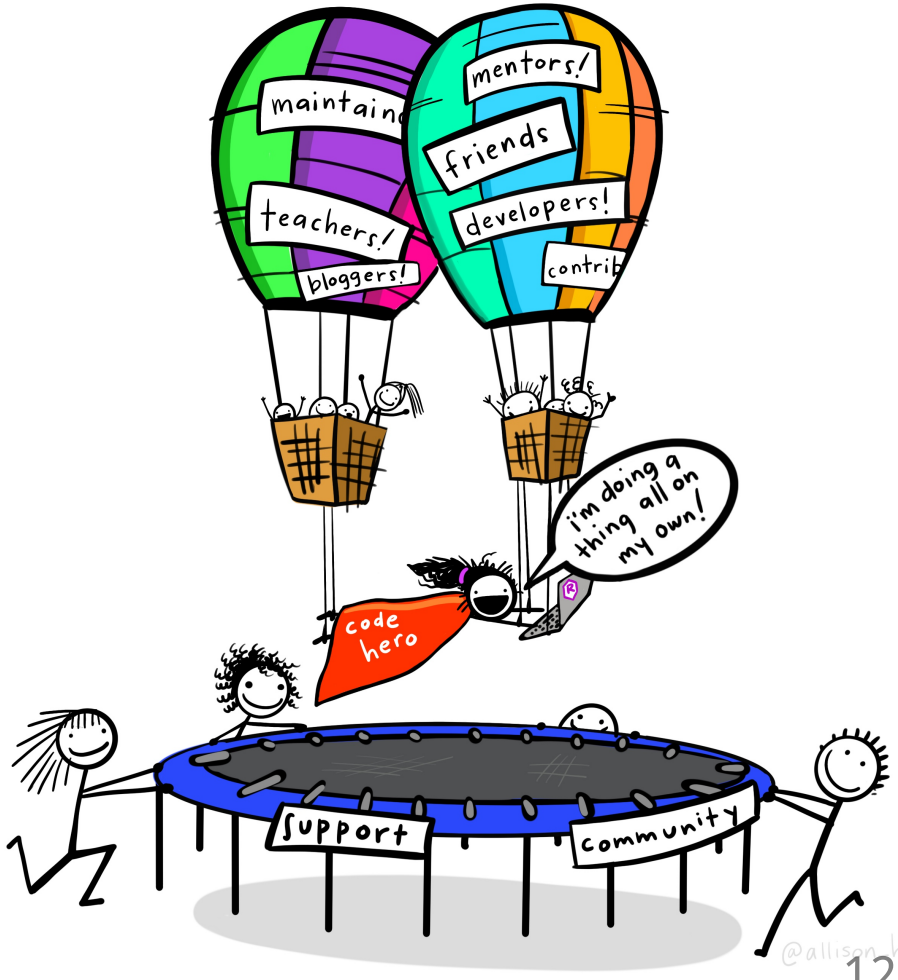
Take a few minutes to come up with a few reasons this is not always done

Why is this not practiced?

- Don't know how
- Too busy
- It's internal work
- Worried about being copied
- Rigged the data

Why is this not practiced?

- Don't know how - **learn! lots of support and tools**
- Too busy - **often faster in the long run**
- It's internal work - **often a need to share**
- Worried about being copied - **in practice low risk**
- Rigged the data - **you have bigger problems**



Artwork by Allison Horst

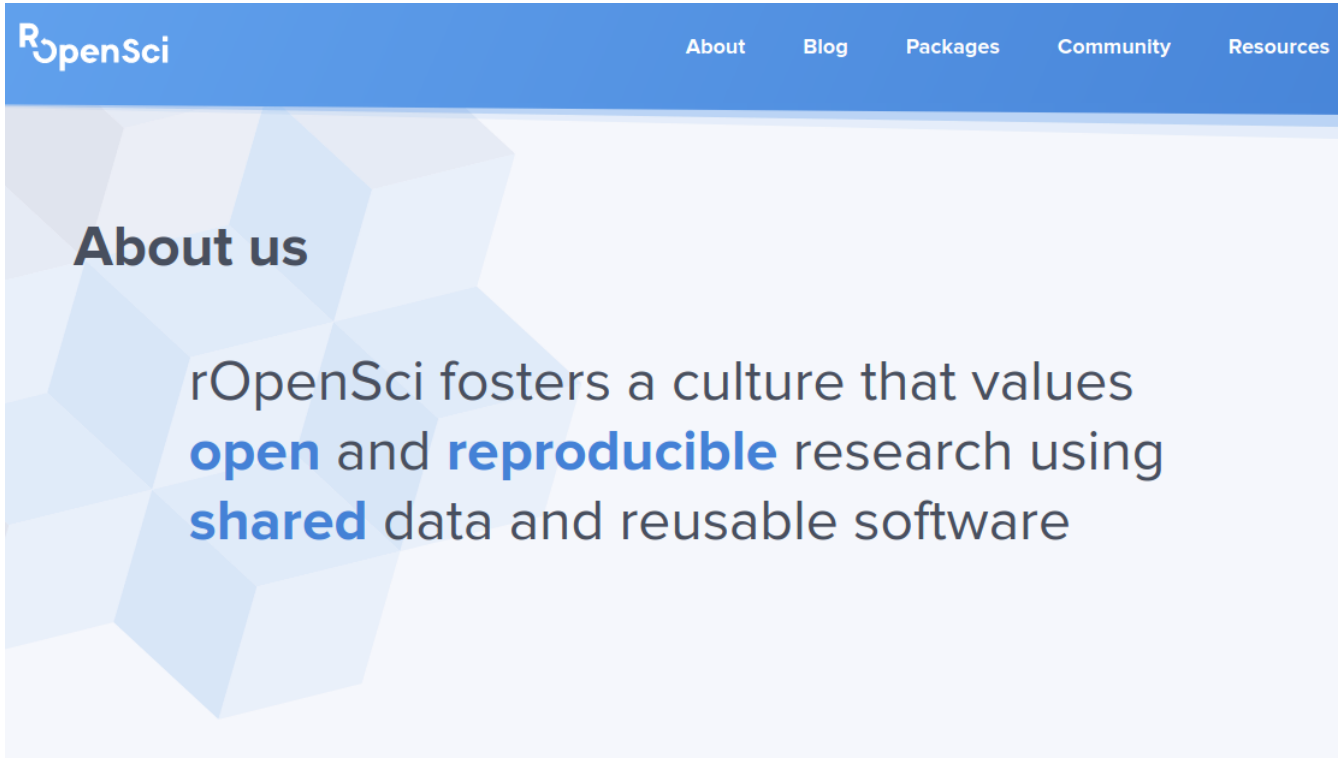
How can you achieve this?

Keeping track

- Extensive notes
  - What, when, with what
- Programmatically
  - Scripts, R, Python, MATLAB, etc.
- Version control
  - git, GitHub, GitLab

Sharing

- Open Science!
  - GitHub, GitLab, etc.
  - Open Science Framework (OSF)
- Journal Supplementary materials



https://ropensci.org

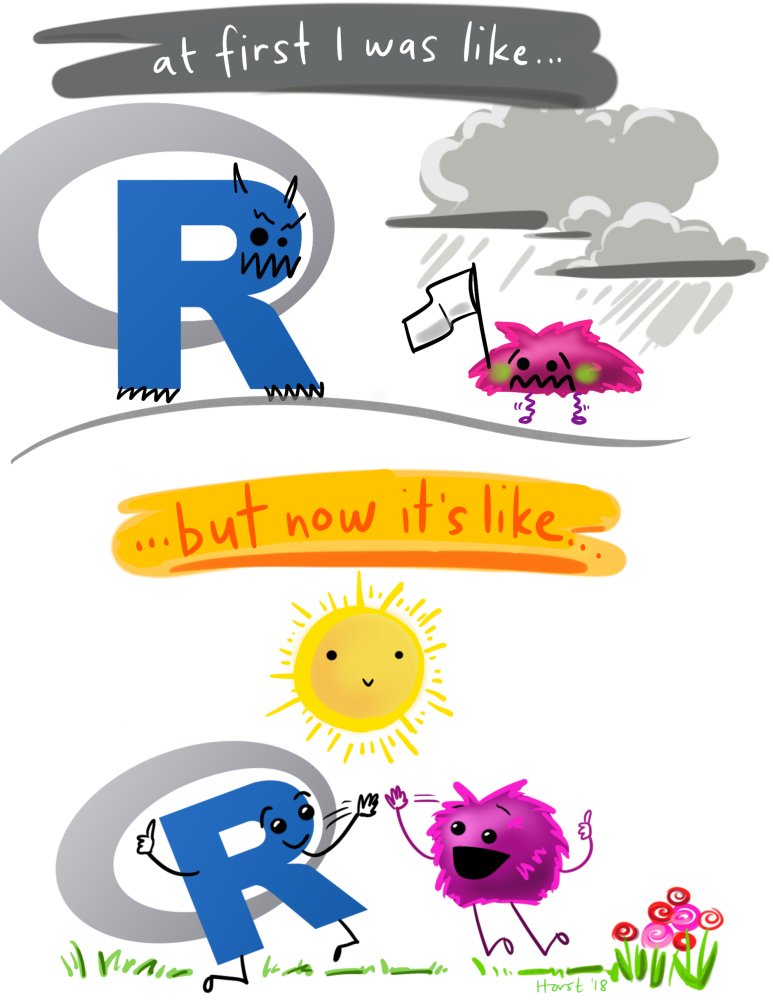
What do we use?

R

- Statistical programming language
- Free and open source

Want to learn?

- rOpenSci organization
- R for Data Science online book
- RStudio Primers interactive online exercises
- Attend classes or workshops
  - Like one of Steffi's Introduction to R workshops (shameless self-promotion)



Artwork by Allison Horst

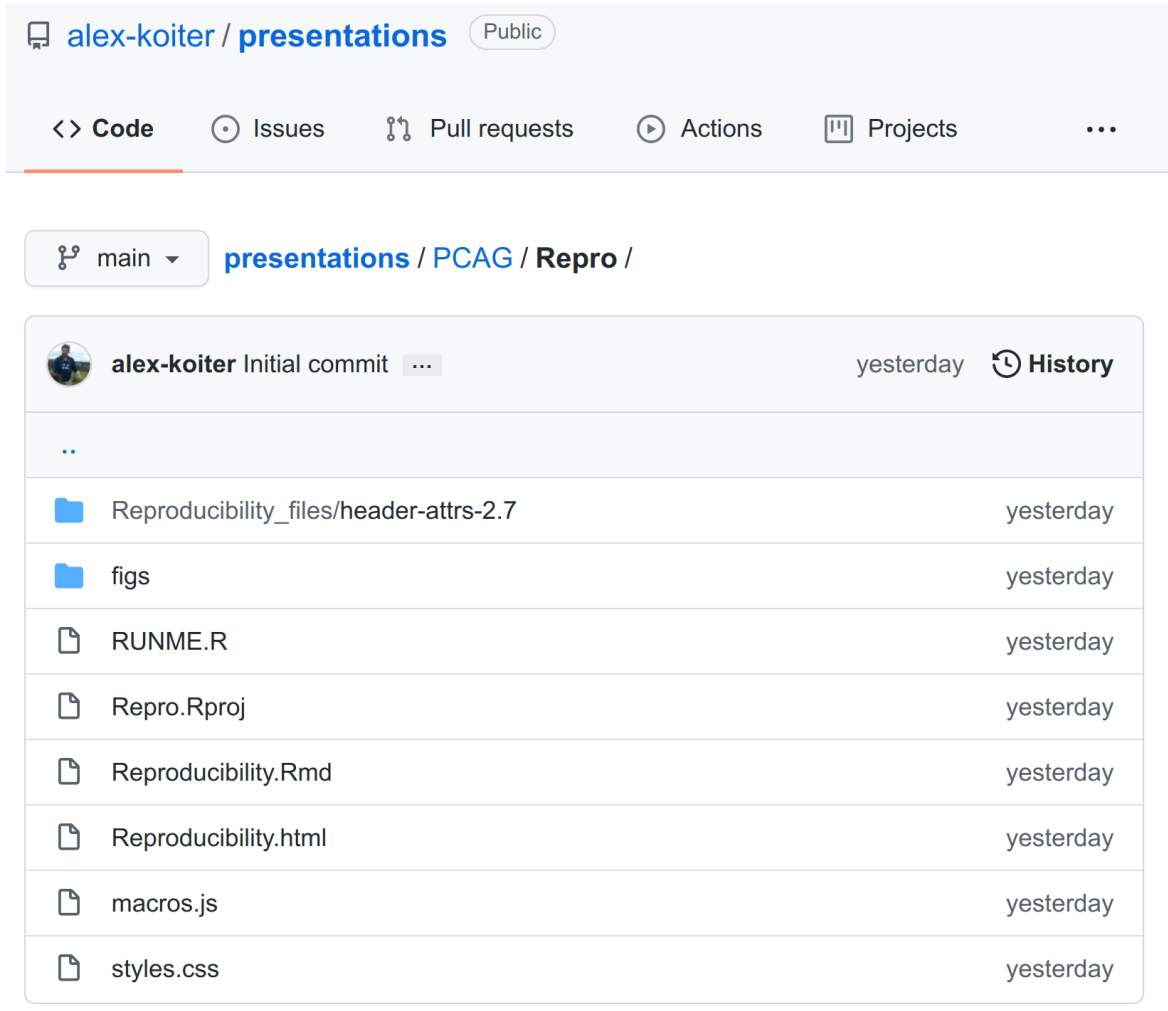
What do we use?

git & GitHub

- **git** is a version control system
  - keeps tracks of changes
- **GitHub** is an online home for git projects
  - allows collaboration and sharing
- like R, tricky to learn, but oh so powerful!

Want to learn?

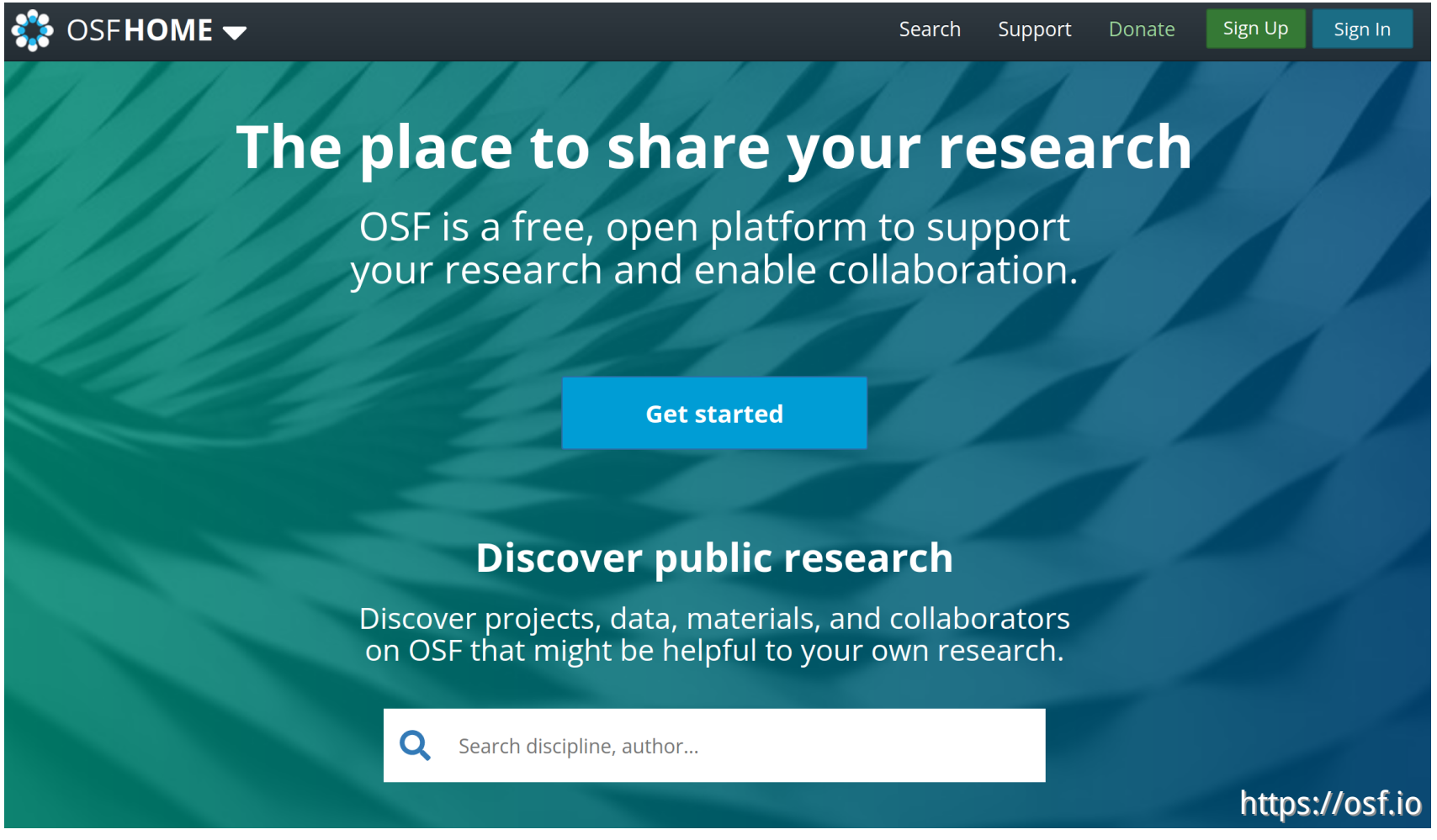
- Happy Git with R online book



This presentation is on GitHub!
(and is reproducible and open)

What do we use?

Open Science Framework

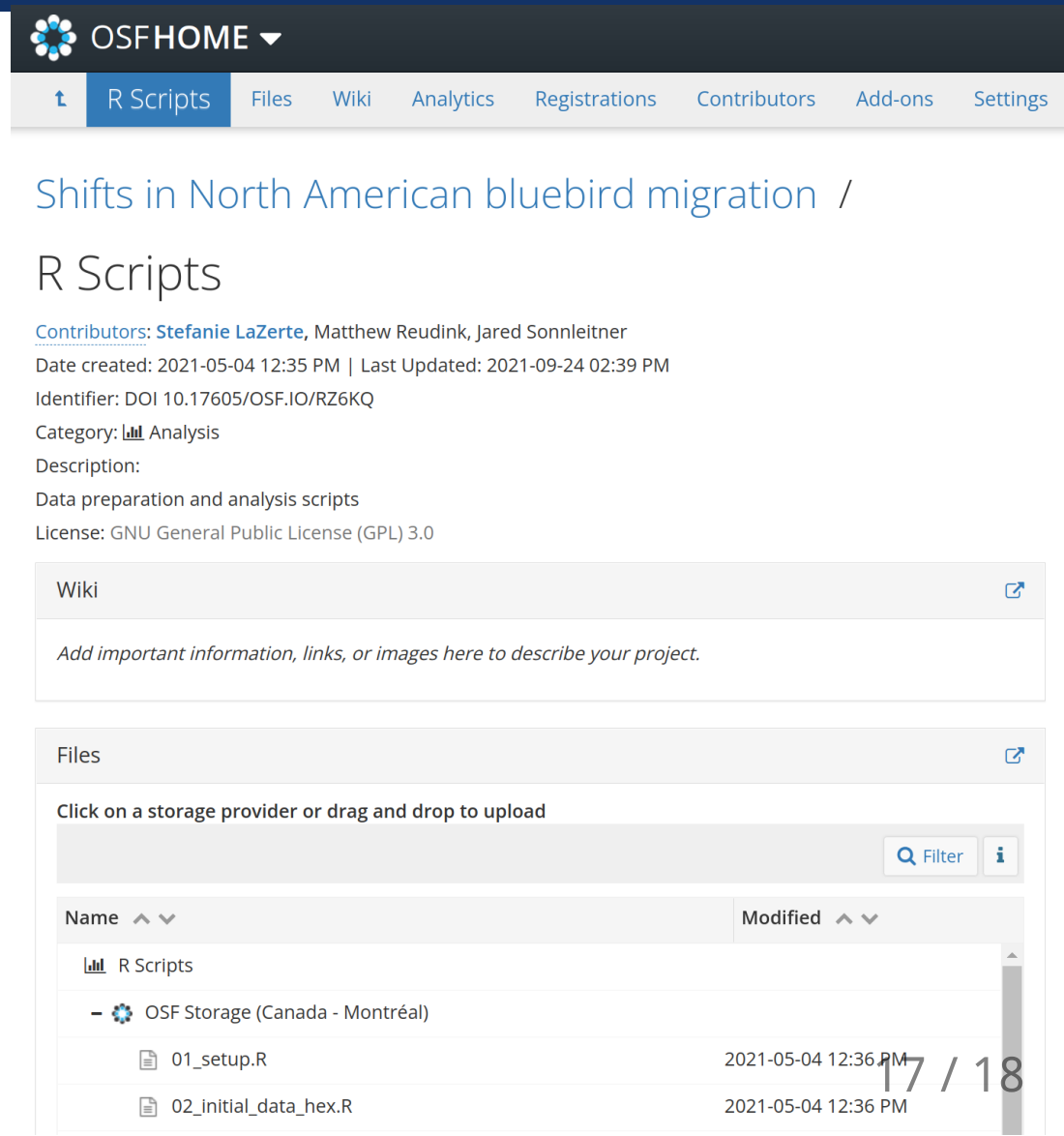


What do we use?

Open Science Framework

- Can integrate with GitHub, Dropbox, Zotero, etc.
- Use as Dropbox-like storage and sharing
  - Drag and drop!
- Make parts private or public
- Create DOIs for referencing in publications

Great way to get your feet wet!



Making science stronger

- Peer review is difficult if we don't know how things were done
- Mistakes happen
  - Can only be fixed if found
  - This is not a sign of weakness - hiding or not learning from them is
- Reducing the need to reinvent the wheel for similar projects/analysis
  - Easier to build upon previous work
- Accessibility

Making science stronger

- Peer review is difficult if we don't know how things were done
- Mistakes happen
  - Can only be fixed if found
  - This is not a sign of weakness - hiding or not learning from them is
- Reducing the need to reinvent the wheel for similar projects/analysis
  - Easier to build upon previous work
- Accessibility

Thank you!

BRANDON UNIVERSITY

@Alex_Koiter | alex-koiter | alexkoiter.ca
@steffilazerte | steffilazerte | steffilazerte.ca
Slides: https://github.com/alex-koiter/presentations (PDF)
Created with the R package xaringan, using remark.js, knitr, and R Markdown
Icons from Ionicons; Compiled on 2021-11-25

Dr. Steffi LaZerte
Analysis and Data Tools for Science