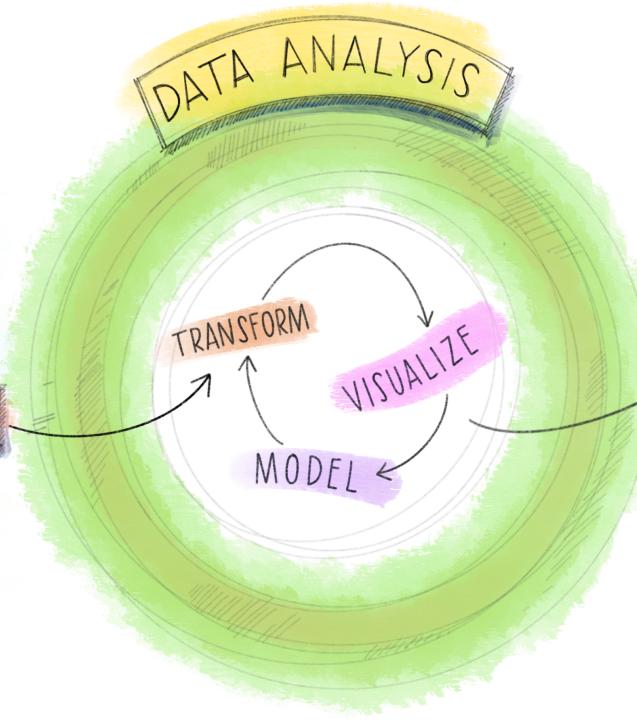


38:279 Introduction to Geographic Research Methods

# Importance of reproducible data analysis in physical geography



Alex Koiter & Steffi LaZerte

Artwork by [Allison Horst](#)

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

# 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

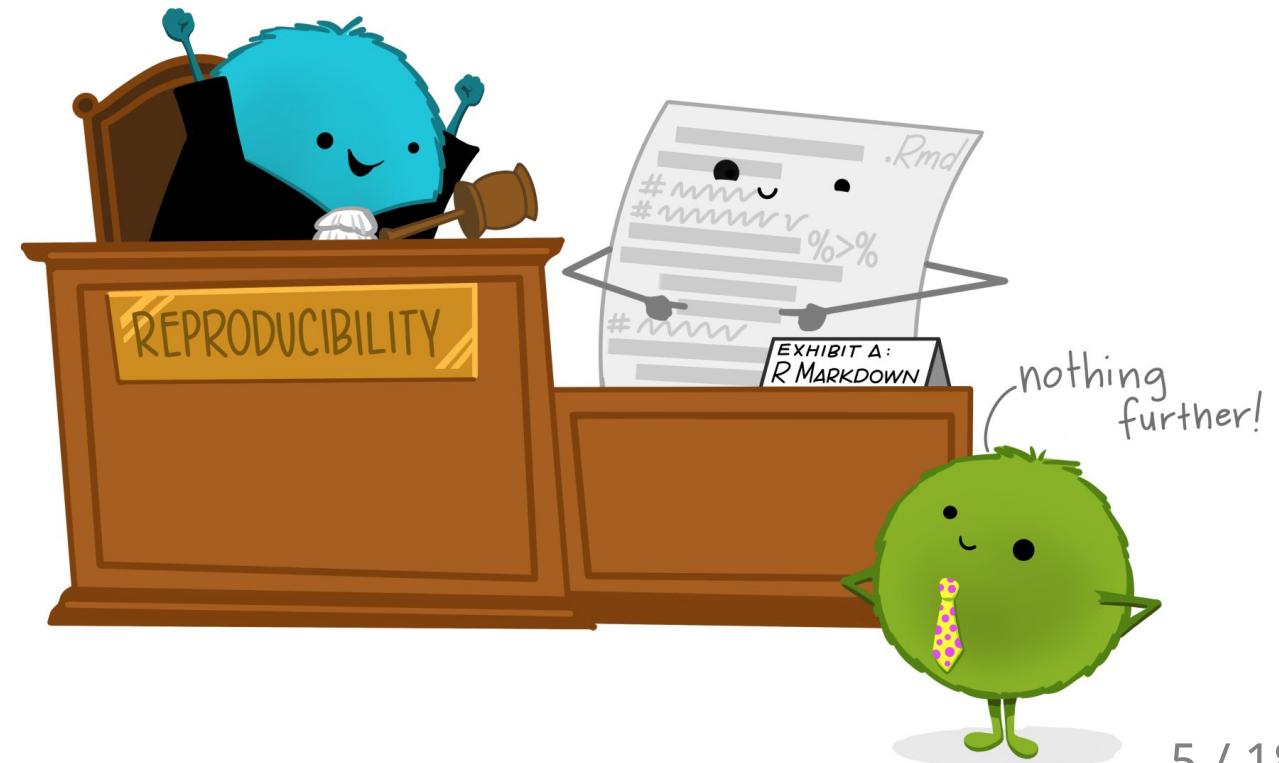
## Replicability

- 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



# 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 acquisition

- 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**

# Working with large data sets

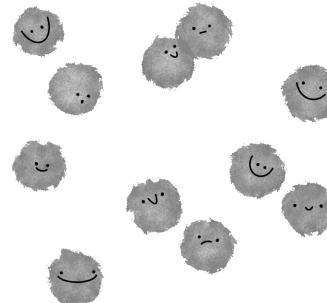
## Data analysis

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

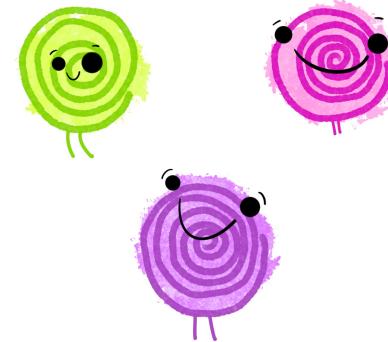
Hard to write papers if you  
don't keep track of this!

## k-means clustering

### OBSERVATIONS



### cluster CENTROIDS



- assign each observation to one of k clusters based on the nearest cluster centroid.

@allison\_horst

# Why is this not practiced?

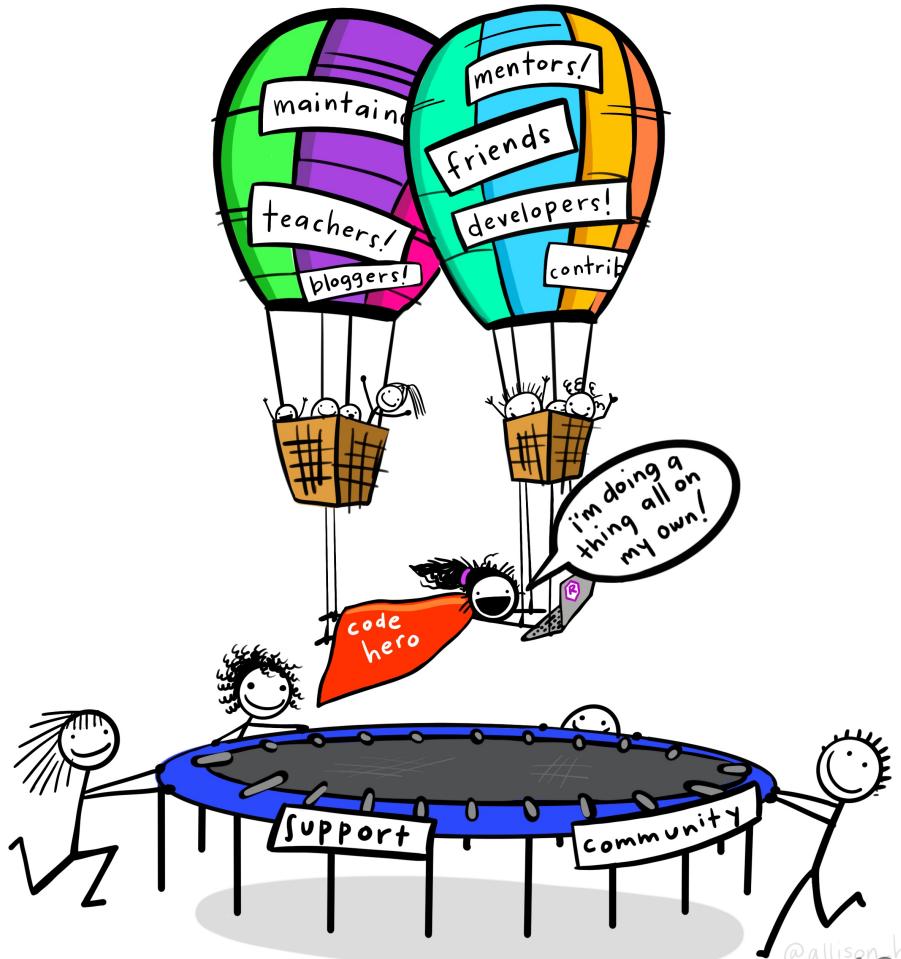
Take a few minutes to come up with a few reasons this is not always done

# 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?

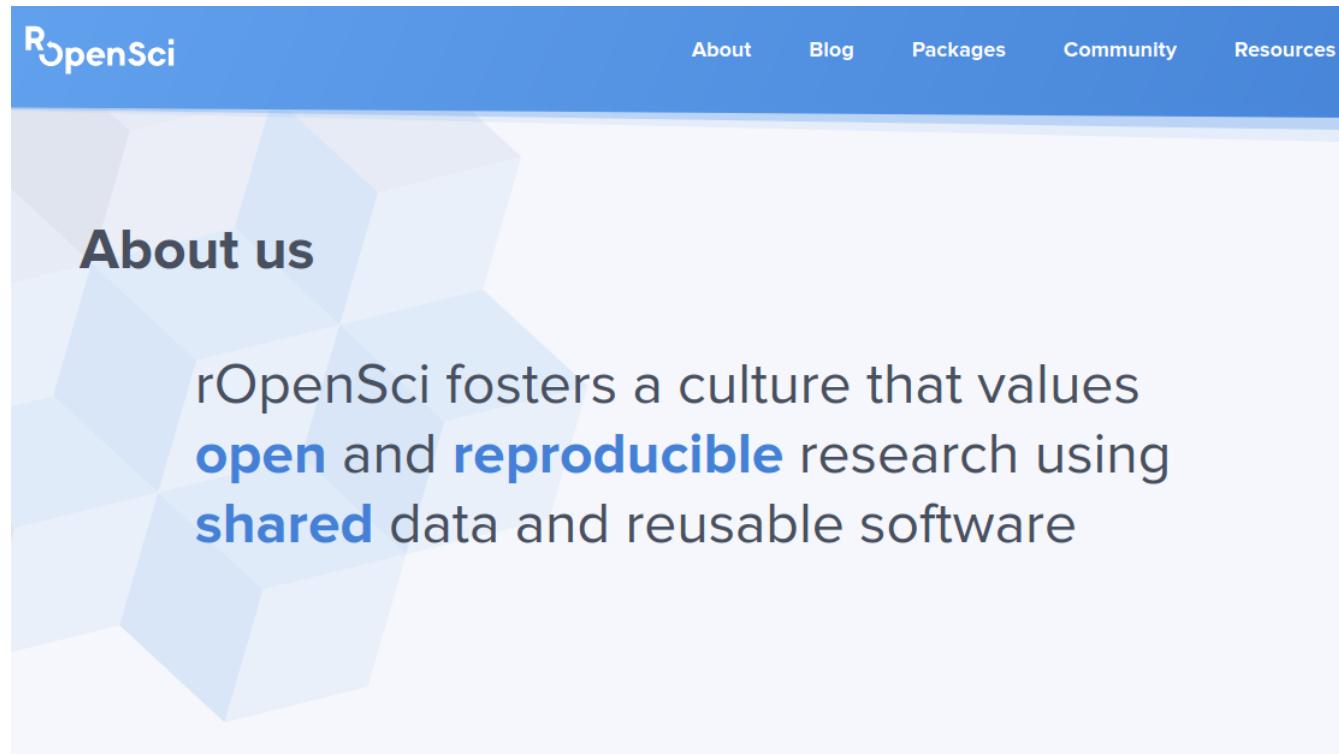
- Don't know how - **learn! lots of support and tools**
- Too busy - **often faster in the long run**
- It's internal work - **often a need to share**
- Worried about being copied - **in practice low risk**
- Rigged the data - **you have bigger problems**



# How can you achieve this?

## Keeping track

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



<https://ropensci.org>

## Sharing

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

# 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)



# 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

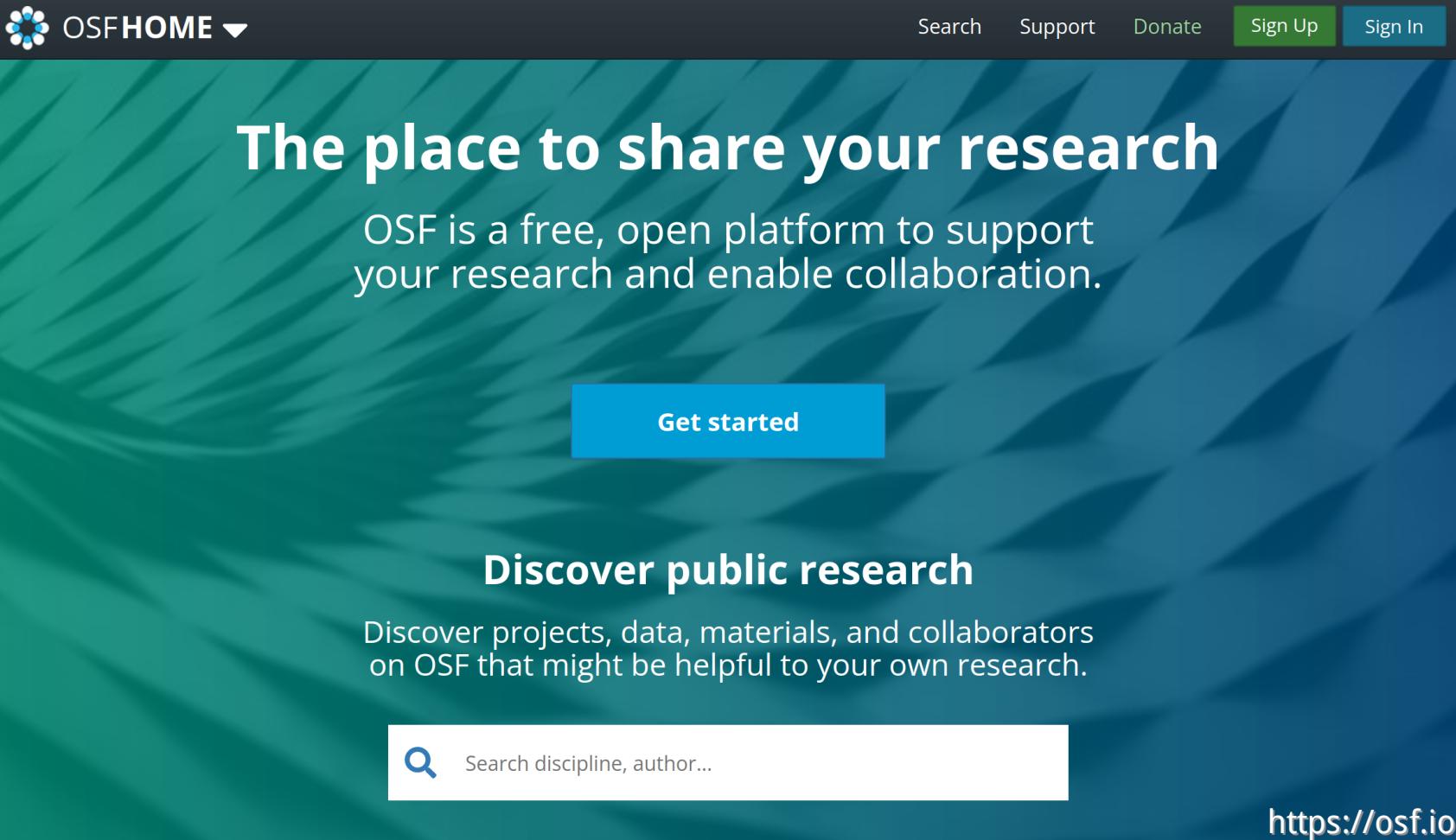
The screenshot shows a GitHub repository page for 'alex-koiter / presentations'. The repository is public. The 'Code' tab is selected. The commit history shows an initial commit by 'alex-koiter' labeled 'Initial commit' from 'yesterday'. Below it, there are ten commits corresponding to files in the repository: 'Reproducibility\_files/header-attrs-2.7', 'figs', 'RUNME.R', 'Repro.Rproj', 'Reproducibility.Rmd', 'Reproducibility.html', 'macros.js', and 'styles.css'. Each of these commits was made 'yesterday'.

File	Commit Date
Reproducibility_files/header-attrs-2.7	yesterday
figs	yesterday
RUNME.R	yesterday
Repro.Rproj	yesterday
Reproducibility.Rmd	yesterday
Reproducibility.html	yesterday
macros.js	yesterday
styles.css	yesterday

This presentation is on [GitHub!](#)  
(and is reproducible and open)

# What do we use?

## Open Science Framework



The screenshot shows the homepage of the Open Science Framework (OSF). At the top, there is a dark navigation bar with the OSF logo, a "HOME" dropdown menu, and links for "Search", "Support", "Donate", "Sign Up", and "Sign In". The main background features a teal-to-blue gradient with diagonal stripes. The central text reads "The place to share your research" in large white font, followed by a description: "OSF is a free, open platform to support your research and enable collaboration." Below this is a blue "Get started" button. Further down, there is another section titled "Discover public research" with the subtext: "Discover projects, data, materials, and collaborators on OSF that might be helpful to your own research." At the bottom left is a search bar with a magnifying glass icon and the placeholder text "Search discipline, author...". The URL "https://osf.io" is visible at the bottom right.

OSFHOME ▾

Search Support Donate Sign Up Sign In

# The place to share your research

OSF is a free, open platform to support your research and enable collaboration.

Get started

## Discover public research

Discover projects, data, materials, and collaborators on OSF that might be helpful to your own research.

Search discipline, author...

<https://osf.io>

# 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!

The screenshot shows the OSFHOME interface for an R Scripts project. At the top, there's a navigation bar with links for R Scripts, Files, Wiki, Analytics, Registrations, Contributors, Add-ons, and Settings. The R Scripts tab is active. Below the navigation, the project title is "Shifts in North American bluebird migration / R Scripts". It lists contributors (Stefanie LaZerte, Matthew Reudink, Jared Sonnleitner), date created (2021-05-04 12:35 PM), last updated (2021-09-24 02:39 PM), identifier (DOI 10.17605/OSF.IO/RZ6KQ), category (Analysis), and a description: "Data preparation and analysis scripts". The license is GNU General Public License (GPL) 3.0. There are sections for "Wiki" and "Files". The "Wiki" section has a placeholder: "Add important information, links, or images here to describe your project." The "Files" section has a message: "Click on a storage provider or drag and drop to upload". A table lists files: Name (R Scripts, OSF Storage (Canada - Montréal)), Modified (2021-05-04 12:36 PM, 2021-05-04 12:36 PM), and two file entries: 01\_setup.R and 02\_initial\_data\_hex.R. A "Filter" button is also visible.

# 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!**

 @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