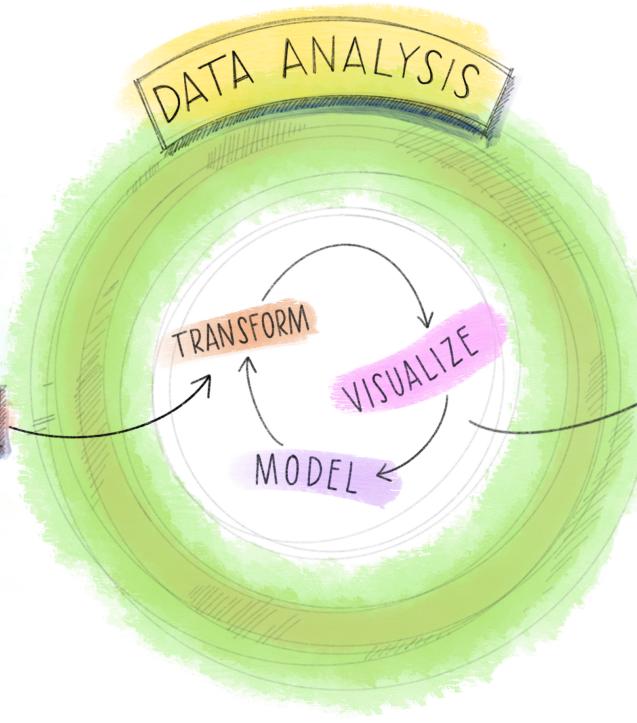


38:279 Introduction to Geographic Research Methods

Importance of reproducible data analysis in physical geography



Alex Koiter & Steffi LaZerte

Artwork by [Allison Horst](#)

Updated from Grolmund & 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

Whats the difference?

Introduction

Reproducibility

- The data analysis can be successfully reproduced

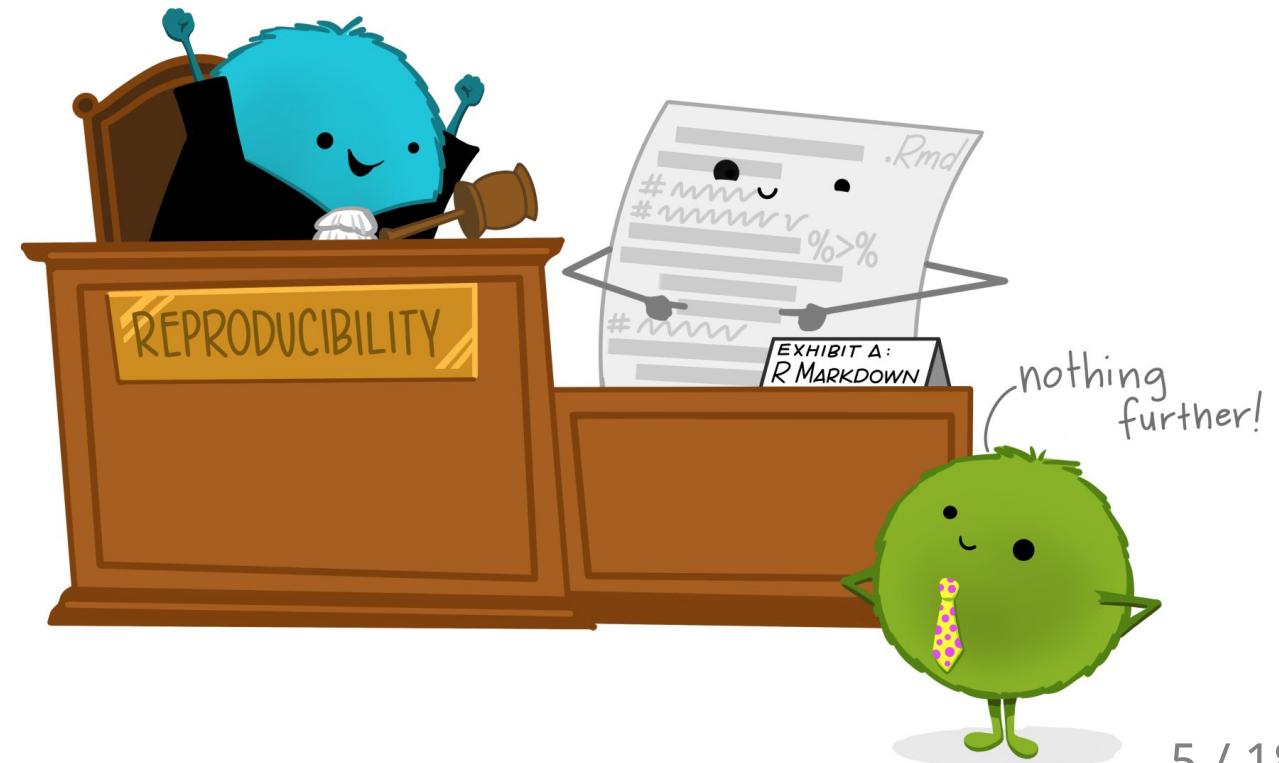
Replicability

- 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

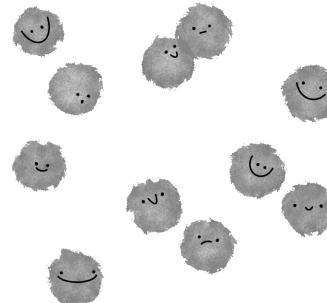
Data analysis

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

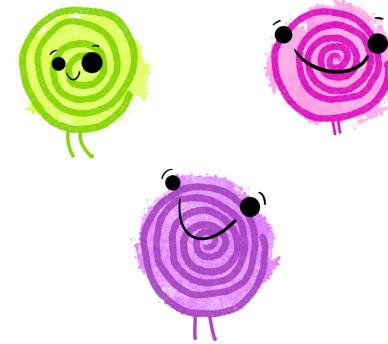
Hard to write papers if you
don't keep track of this!

k-means clustering

OBSERVATIONS



cluster CENTROIDS



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

Why is this not practiced?

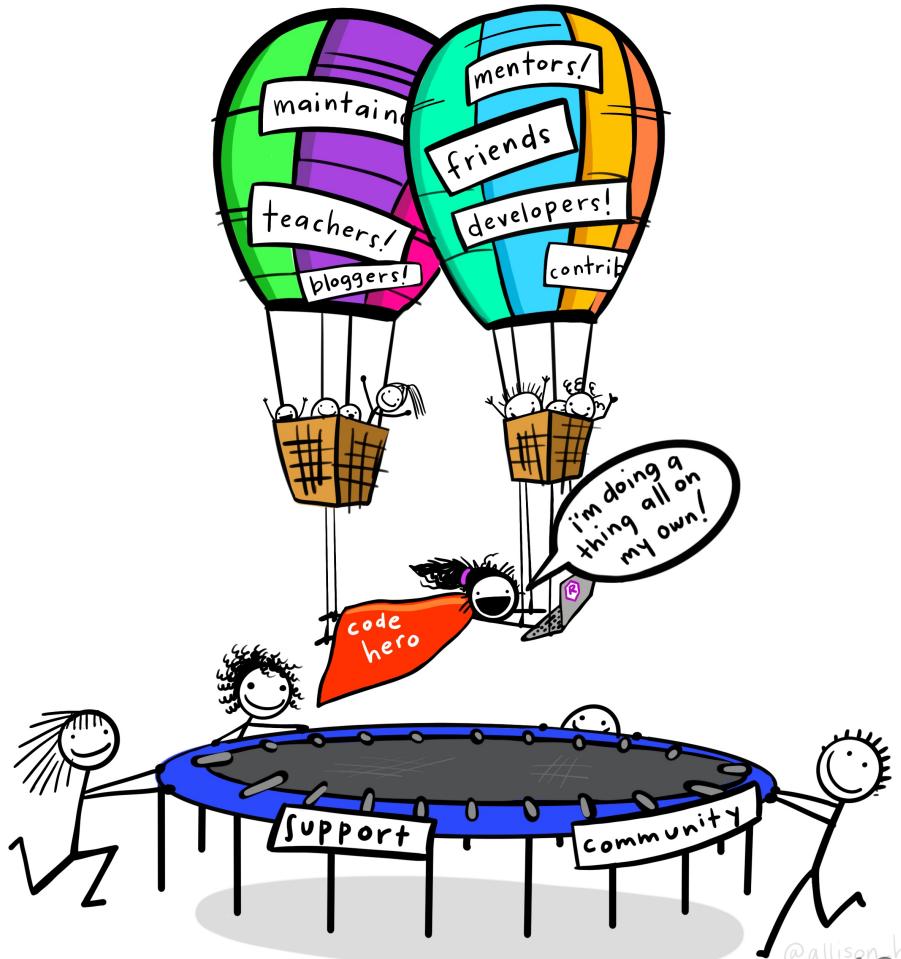
Take a few minutes to come up with a few reasons this is not always done

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](#)

The screenshot shows the rOpenSci website's "About us" page. The header features the rOpenSci logo and navigation links for About, Blog, Packages, Community, and Resources. The main content area has a light blue background with a geometric polygon pattern. The title "About us" is centered in bold black font. Below it, a paragraph explains the organization's mission: "rOpenSci fosters a culture that values **open** and **reproducible** research using **shared** data and reusable software".

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)



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

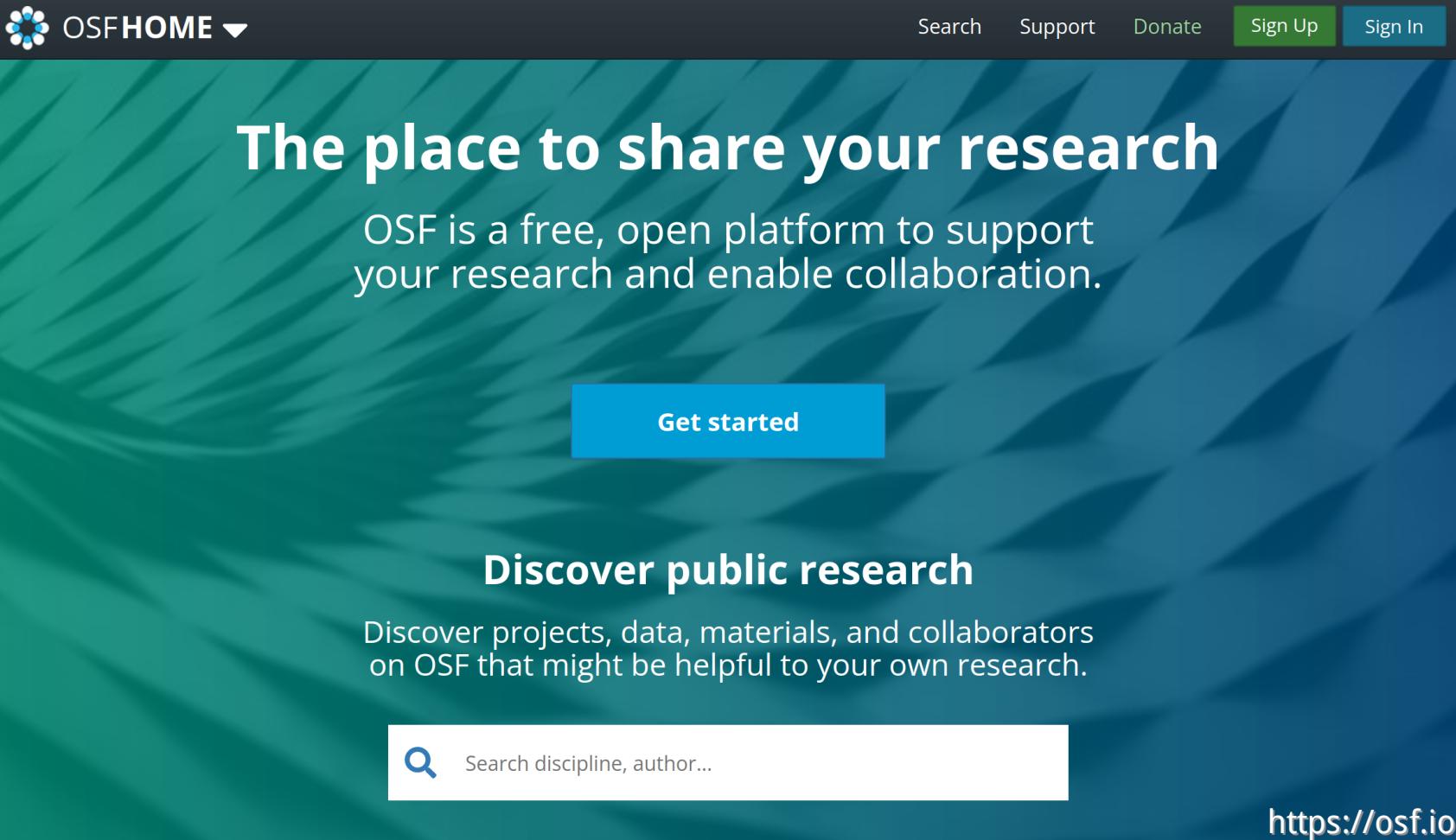
The screenshot shows a GitHub repository page for 'alex-koiter / presentations'. The repository is public. The 'Code' tab is selected. The commit history shows an initial commit by 'alex-koiter' labeled 'Initial commit' from 'yesterday'. Below it, there are ten commits corresponding to files in the repository: '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 a project titled "Shifts in North American bluebird migration / R Scripts". The top navigation bar includes links for R Scripts, Files, Wiki, Analytics, Registrations, Contributors, Add-ons, and Settings. Below the title, it displays contributor information (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), description (Data preparation and analysis scripts), and license (GNU General Public License (GPL) 3.0). The "Wiki" section contains a placeholder for adding project details. The "Files" section has a drag-and-drop area and a table listing files: "R Scripts" (modified 2021-05-04 12:36 PM), "OSF Storage (Canada - Montréal)" (modified 2021-05-04 12:36 PM), "01_setup.R" (modified 2021-05-04 12:36 PM), and "02_initial_data_hex.R" (modified 2021-05-04 12:36 PM). A search bar labeled "Filter" and a help icon are 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

