Outline

* Lots of researchers use code for their data wrangling and analyses, because code easier to share, and to use to reproduce results than point and click software
* BUT few researchers share their code. This paper from last year found that of 350 studies in top journals like Ecology, Nature and Science, despite 80% of studies sharing their data (shown on the left), only 20% shared their code (shown on the right)
* and only some code works correctly. Here’s a list of 4 studiesHere’s a list of 4 studies, they range in fields like Ecology, Psych and Biomed. Each tried to run other study’s code to reproduce results, and the proportion of results they *could* reproduce is on the left. What I want you to notice is the highest number is 50%. That means whether a code script reproduces results usually has a lower chance than a coin toss. That’s not great
* Code reproducibility IS HARD! And here are a few reasons why:
  + Lots of us are self-trained coders
  + Packages constantly update
  + The things we study have lots of variation, which can break models for example
  + Academic institutions and funding schemes don’t necessarily incentivise reproducibility
  + And researchers are busy! They have lots to do.
* But if I’m being honest, maybe it’s because it’s easier to do the same thing you always do. New stuff is scary.
* But code reproducibility does matter. Journals are beginning to require code availability statements. Some journals allow peer reviewers to require code to review a paper properly. So sharing code is important for publication.
* But even more importantly, your should care because reproducibility is fundamental to science and knowledge building. If you find a significant result, other’s can’t build on it if they can’t figure out whether it was true.
* At the ALA, myself and my colleagues work on many projects that try to support reproducibility in R, namely the galah package, ALA Labs, and more recently the Cleaning Biodiversity Data in R book. Each of these aim to make it easier to download data, use that data, and wrangle that data transparently and repeatably
* Now I want to share some tips of what I’ve learned through those projects about reproducible workflows in R. To show you, let’s make a code snippet reproducible.

Main points

* The setup
  + Create an R Project (and the here package)
  + GitHub setup:
    - Sign up, link to R Studio
    - usethis
    - This is the best way to progress work while still saving old things that you can copy paste back in if you need
  + renv/groundhog/sessionInfo
  + Folder structure and file paths (good naming practices)
  + Backed up data locally/repository/DOIs
* Human readable
  + Clear notes & object names (avoid abbreviations)
  + Line breaks
  + Separate into chunks that reflect smaller but meaningful steps
* Render

Things I want to emphasize that are related to ALA

* Data DOIs – atlas\_occurrences(), or an ALA DOI (galah)
* We’ve made many small reproducible code write-ups in ALA Labs
* Data cleaning book can help with data wrangling and validation

Conclusion

* People should care about! Reproducibility is a fundamental tenet of scientific theory building. It also prevents the frequent need to start again from scratch when a project cannot be replicated
* People (particularly ECRs) are getting noticeably better at this over time.
* The best way to improve reproducibility individually is to talk about it with others, and as you become more senior, teach up-and-comers and push for change at a structural level

Additional points

* Worried about getting scooped? Use embargo periods on OSF