Version Control and Reproducibility with GitHub and Zenodo

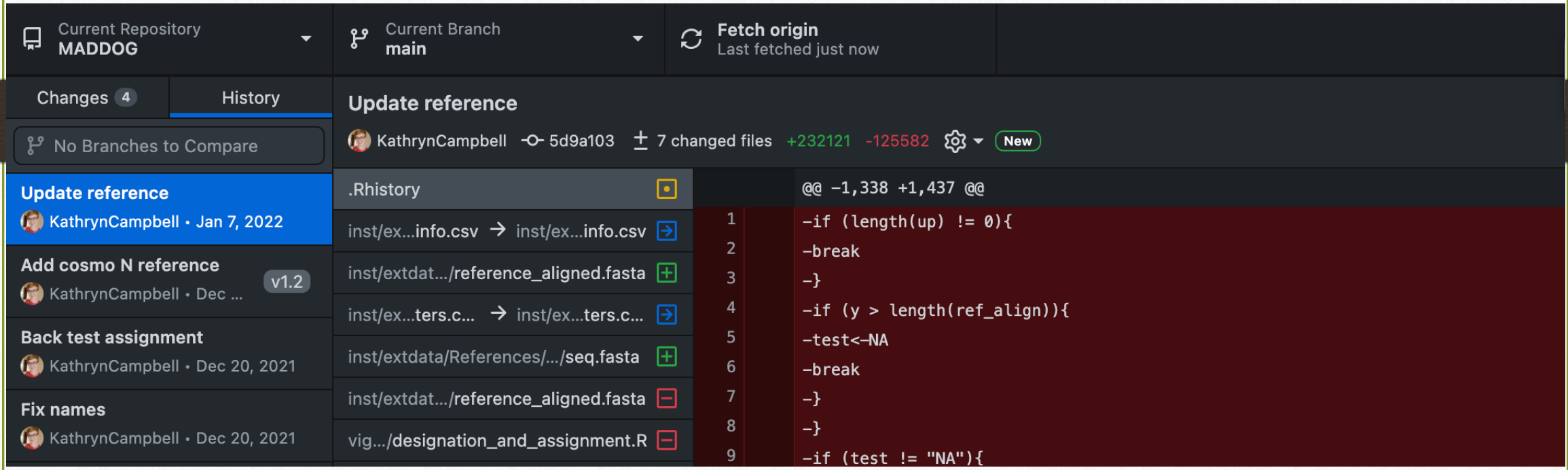
Kathryn Campbell

Rabies Lab Meeting 24th March

Introduction to Version Control

- The practice of tracking and managing changes to software code.
- Keeps track of every modification to the code in a special kind of database.
- Help software teams manage changes to source code over time.
- If a mistake is made, developers can turn back the clock and compare earlier versions of the code to help fix the mistake while minimizing disruption to all team members.

Version Control in GitHub



Introduction to Reproducibility

- The ability to replicate the analyses undertaken and produce the same findings
- Ensures the study, methods and findings are robust
- Other researchers need to be able to use any data made publicly available, tools/pipelines developed or study designs generated by your study in their own research
  - **Can other people understand and replicate the study?**
- The raw data and code used for analyses need to be clear and understandable;
  - **How do we get from the raw data to the figures/outputs from the code?**

Data Formatting

- Raw data needs to be formatted in a way that others (and you!) can easily understand and use
- **Don't include spaces or special characters in column or row names**; this makes it difficult to call the column when analyzing the data in R
- Crucial that all **columns have names**, and there is only **one data entry per cell**

ID	country	year	sequence_host
JQ685894	United States	2009	Vulpes vulpes
JQ685944	United States	1984	Mephitis mephitis
JQ685967	United States	2012	Mephitis mephitis
JQ685970	United States	1974	Mephitis mephitis
JQ944704	Russia	2009	Canis familiaris
JQ944705	Russia	2008	Canis familiaris
JQ944706	Russia	2008	Canis familiaris
JQ944708	Russia	2008	Vulpes vulpes

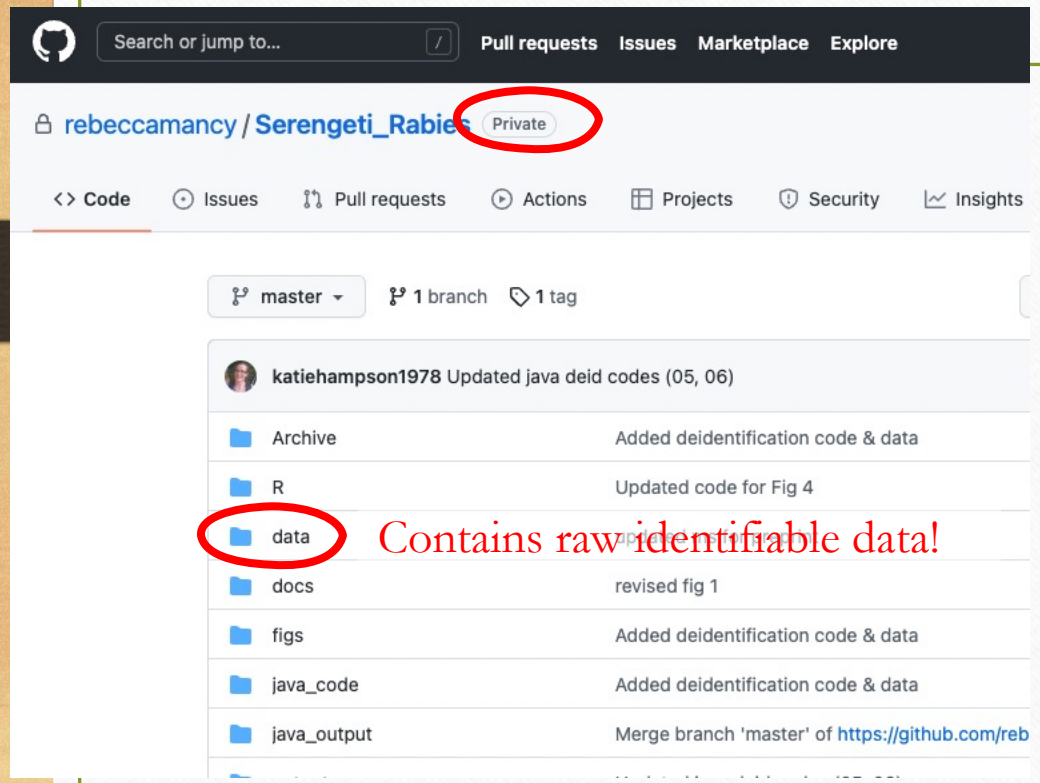
country	sequence IDs	years of collection	sequence hosts
United States	JQ685894, JQ685944, JQ685967, JQ685970	2009, 1984, 2012, 1974	Vulpes vulpes, Mephitis mephitis
Russia	JQ944704, JQ944705, JQ944706, JQ944708	2009, 2008	Canis familiaris, Vulpes vulpes

De-identifying data (anonymizing)

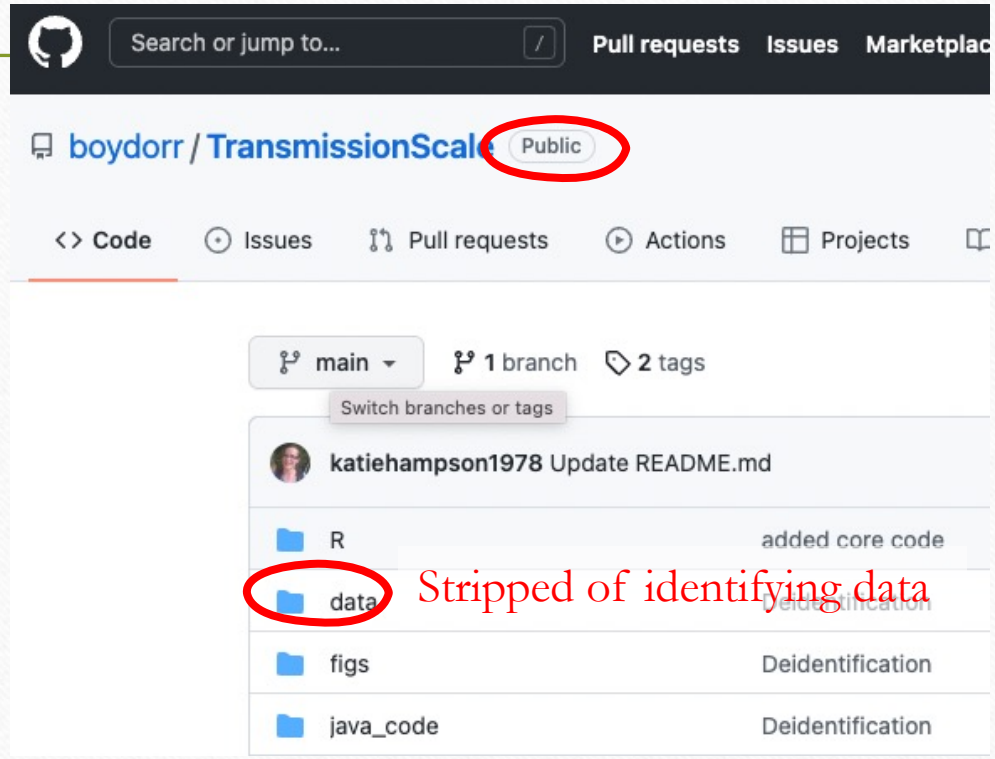
- What is sensitive data?
  - Names, telephone numbers, GPS coordinates, age, address, or other information that could allow someone to identify a person and their medical history
  - Our work has a LOT of sensitive data:
    - Contact tracing, IBCM, surveillance, household surveys, even dog vaccination registers!!!

Github repositories

- Sensitive - PRIVATE



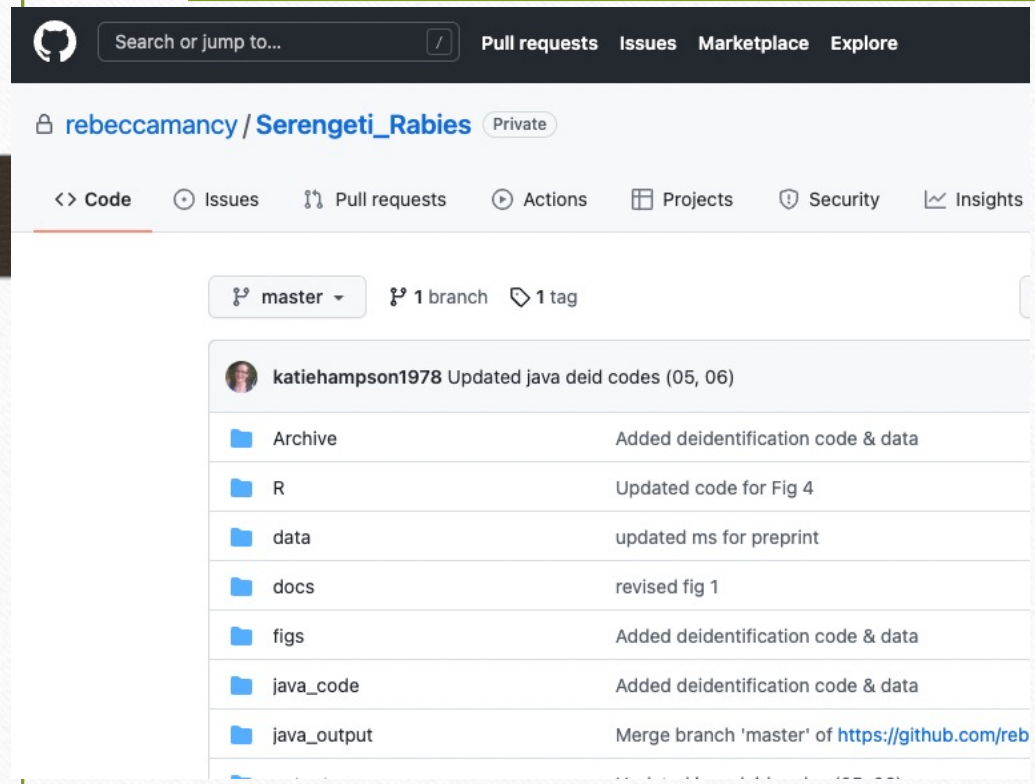
- Deidentified - PUBLIC



- Can I just make my Github repo public? NO – because its history is tracked!

In your private Github repository:

- Write code to strip data to bare minimum needed for analysis and removing identification/ sensitive information



Data cleaning now outputs deidentified data

Code duplicated to analyse deidentified data

Ideally do this at the earliest stage of analysis.
But not always possible if you need raw data to understand connections in the data!

Examples

- Import into R raw data with all its sensitive information

```
39
40 # Read in data from WiseMonkey and subset to Serengeti ----
41 animalCTts <- "20210603223010" # WM time signature
42 humanCTts <- "20210529185812"
43 biting_animals <- read.csv(paste("data/Tanzania_Animal_Contact_Tracing_",
44                                   animalCTts, ".csv", sep = ""),
45                             stringsAsFactors = FALSE)
```

R	S	T	U	V	W	X	Y	Z	AA	AB	
UTM Easting	UTM Northir	District	Village	Biter ID	ID	Uncertainty	Species	Other specie	Owner	Owner name	Anim
		Serengeti	Mbirikiri	0	1961	incorrect gps	Wildlife: White tailed mo		Not applicable		FA
670939.103	9798185.07	Serengeti	Nyirongo	0	2569		Domestic dog		Unknown		FA
651592.14	9819956.06	Serengeti	Nyamakobiti	0	3493		Domestic dog		Unknown		FA
673769.152	9815641.03	Serengeti	Sogoti	0	3542		Domestic dog		Unknown		FA
678956.113	9807974.01	Serengeti	Rung'abure	0	3638		Domestic dog		Unknown		FA
697365.118	9788764.04	Serengeti	Bonchugu	3756	3775		Domestic dog		Known	Wegesa Mar	TI
687375.082	9809752.09	Serengeti	Manyata	3778	3777	what happer	Domestic dog		Known	Mwita Ngocl	TI
648622.057	9822383.98	Serengeti	Nyamakobiti	0	4716		Domestic dog		Known	Chacha mwit	TI

GPS

Owner name

Examples

- Save information with only useful variables (makes files nice and small too!)

De-identified data object

Still has ~20 variables but none sensitive

```
147  biting_animals_deid <- biting_animals %>%
148    dplyr::select(UTM.Easting.jitter, UTM.Northing.jitter, Biter.ID, ID, Chain.ID,
149                  Species, Other.species, Owner, Suspect, Rabid,
150                  Date.bitten.known, Date.bitten, Date.bitten.uncertainty,
151                  Symptoms.started.known, Symptoms.started, Symptoms.started.accuracy,
152                  Incubation.period, Incubation.period.units, Infectious.period, Infectious.period.units,
153                  Outcome, Action, Dogs.bitten, Animals.bitten, Carnivores.bitten, Locations)
154  saveRDS(object = biting_animals_deid, paste0("output/clean_bite_data_no_densities_deid.rda"))
155
```

Saved with a useful (explanatory name) in outputs so not confused with raw (sensitive) data

Examples

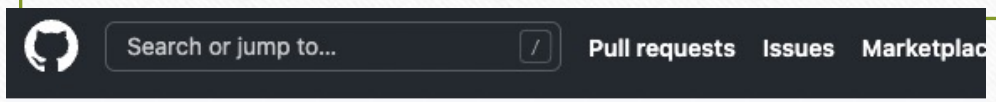
- Jitter GPS locations
  - Add random numbers 0-1 km to all X points; repeat for Y points

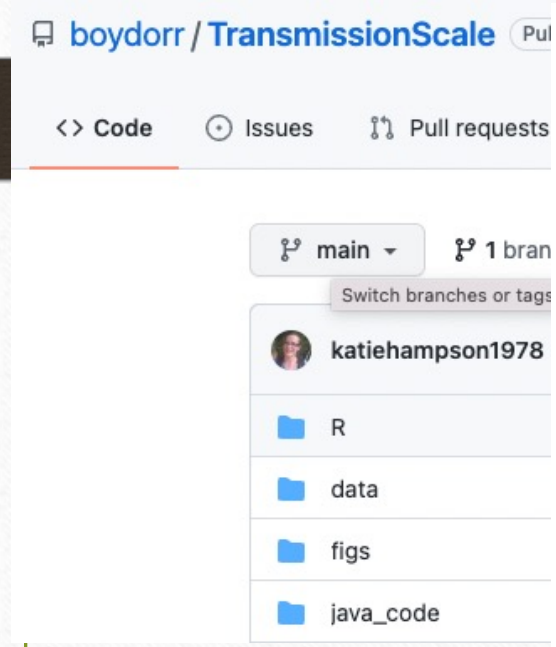
```
142
143  # DE-IDENTIFY!
144  biting_animals$UTM.Easting.jitter <- jitter(biting_animals$UTM.Easting, amount =1000)
145  biting_animals$UTM.Northing.jitter <- jitter(biting_animals$UTM.Northing, amount =1000)
146
```

Save to a new variable name – do not to overwrite the raw spatial data which may be essential to fine scale spatial analyses!

Finally create a new PUBLIC repository and only move over deidentified data and scripts that run with these data

- Deidentified - PUBLIC
- AND warn future users that spatial data is jittered! (they need to set up a collaboration and IRB for using raw sensitive data!)





README.md

Rabies shows how scale of transmission can enable acute infections to persist at low prevalence

Authors: Rebecca Mancy, Malavika Rajeev, Ahmed Lugelo, Kirstyn Brunker, Sarah Cleaveland, Elaine A. Ferguson, Karen Hotopp, Rudovick Kazwala, Matthias Magoto, Kristyna Rysava, Daniel T. Haydon, Katie Hampson

This repository contains all the code and de-identified data in the paper by Mancy et al. 2022.

Geographic masking was used to de-identify the spatial data on rabies transmission (which is mostly localized to households of either dog owners or persons bitten by rabid animals), by jittering XY locations within a 1km radius. Running the transmission tree code on these jittered data will therefore not generate precisely the same inferred progenitors as from the raw data because the very localized spatial structure is masked. The resulting consensus trees are provided from the original raw data as are the distances between cases and to contacts and the inferred step lengths of rabid animals.

For more information: katie.hampson@glasgow.ac.uk

Work through your plans and your repo with a member of the team!

- **We are here to check and troubleshoot**
- And we (me!) have responsibility to make sure nothing sensitive is accidentally released!
- We should also independently run the code and check everything works, which is also a useful to minimize errors before publication!

A CODING PARTNER IS GREAT PRACTICE & BUILDS SKILLS AS THEY WILL OFTEN KNOW SHORT CUTS & TRICKS

Don't worry about sharing messy code:

It is the best way to learn & get better!

Introduction to GitHub

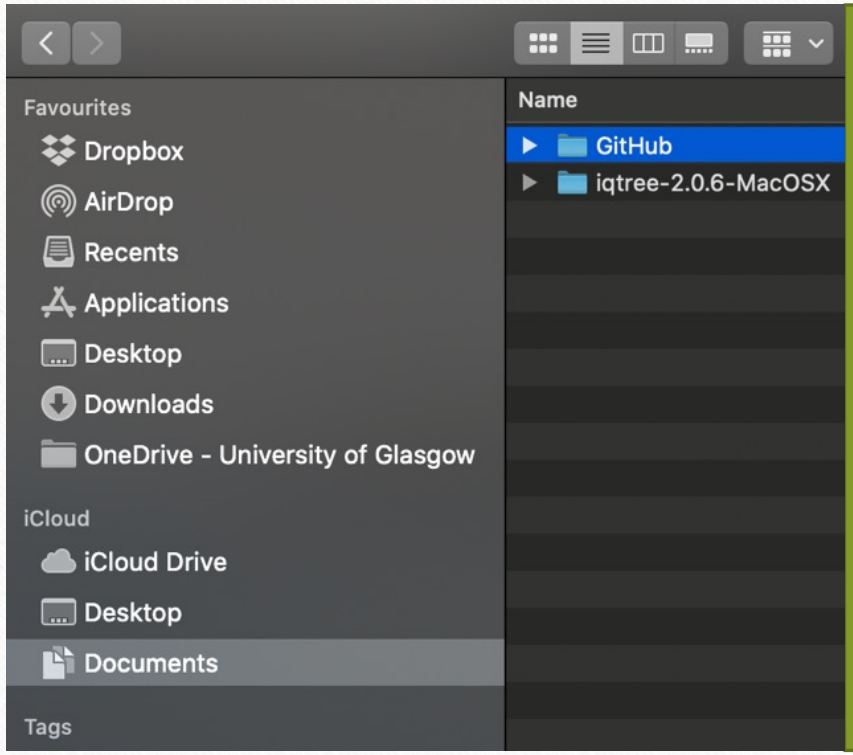
- GitHub is a free, user-friendly website and cloud-based service that helps developers store, manage and distribute their code.
- Consists of Repositories (or "repos") that contain all the code, data, figures, outputs etc for a particular project.
- Multiple people can have access to and work on a repository, allowing easy collaboration

INTERACTIVE ELEMENT:

If you haven't already, sign up at **https://github.com/**

And download GitHub desktop: **https://desktop.github.com/**

Setting up GitHub



INTERACTIVE ELEMENT:

Make a folder called “Github” or “Git” somewhere easy to find (e.g. your desktop or documents)

Open up GitHub desktop. You may be prompted to login if this is the first time using it.

Cloning the Example Repository

INTERACTIVE ELEMENT:

I'll show you how to clone a repository.

This is the repository name you'll need:
KathrynCampbell/Example_Repository

GitHub Interactive Element Checklist

- Finding the Repository
- File Structure and names
- Creating an R project
- Saving R outputs within the repository
- Creating your own repository from a directory
- Seeing the repository on GitHub.com
- Pushing and pulling
- Examining changes (version control) – the 4 icons (add, rename, modify, delete)

README file

INTERACTIVE ELEMENT

I'll show you how to edit the README file and why it's important.

Here's an example: https://github.com/KathrynCampbell/MADDOG

Introduction to Zenodo

- Zenodo makes the sharing, curation and publication of data and software a reality for all researchers
- Allows you to publish your code, repository or package and receive a DOI you can use to share the work
- Creation of 'releases' that allow you to work on code, push it to GitHub so others can collaborate, but not make it 'live' to the public until you're ready

INTERACTIVE ELEMENT:

If you haven't already, sign up at https://zenodo.org/ using your GitHub account

Zenodo Interactive Element Checklist

- Navigating to the GitHub section of Zenodo
- Selecting a repository
- Creating a release
- Getting a DOI
- Adding the DOI to your README
- Using the DOI
- Consistent DOI across releases