Reproducible Research and the Data Lifecycle

Week 1
Data Science Workshop for NGA LTER REU Students

Goals for Today

- Understand the motivation behind the upcoming lessons
- Introduce terminology you (hopefully) will see throughout your careers
- Introduce some best practices regarding data at the **beginning** of your summer research (instead of at the end)

Code of Conduct

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Focus on what is best for the community
- Show courtesy and respect towards other community members

Code of Conduct from The Carpentries: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

Getting Started

Let’s do some polls.

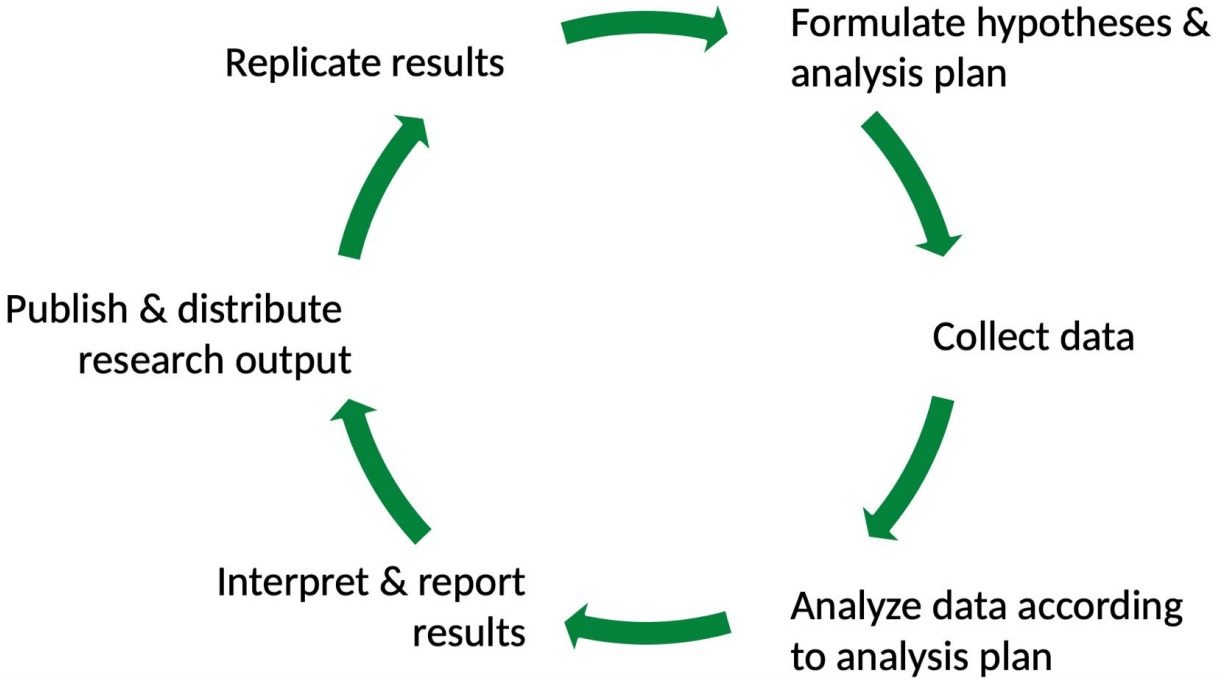
Principles

The Scientific Method

1. Formulating a hypothesis
2. Designing the study
3. Running the study and collecting the data
4. Analyzing the data
5. Reporting the study

OSC
LMU Open Science Center

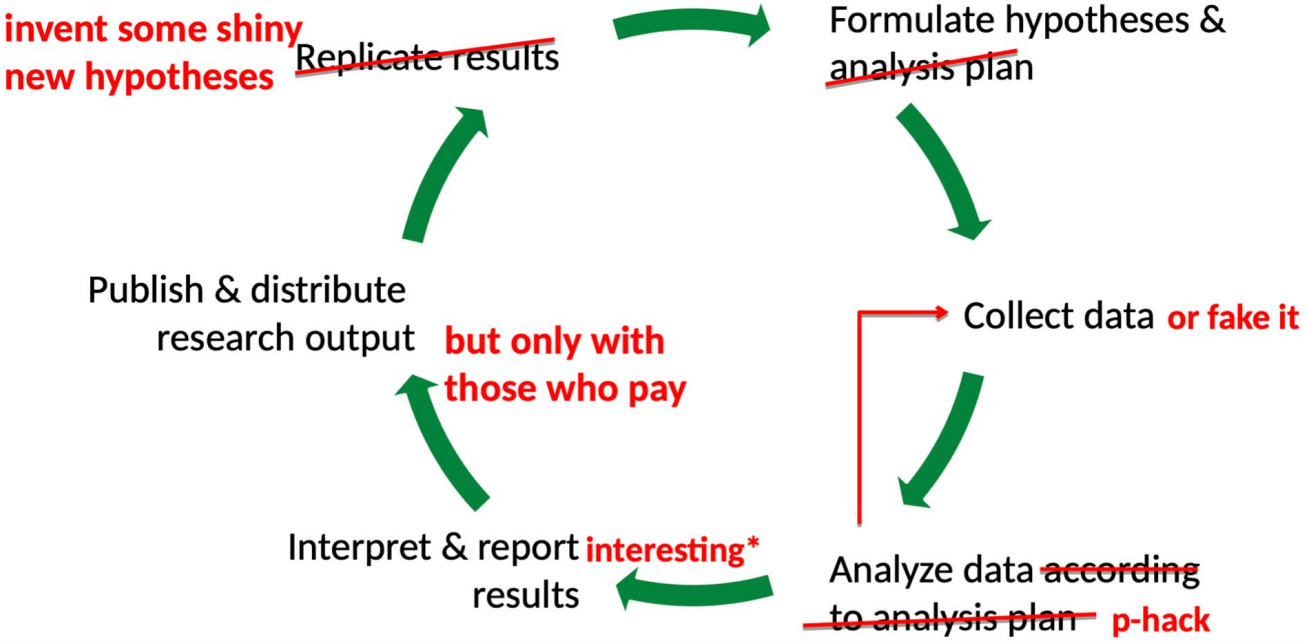
The Confirmatory Research Process



f.io/z7954/

Wagenmakers et al. (2012)

OSC
LMU Open Science Center



* p < .05; that fit a theory; that are surprising / publishable...

https://osf.io/z7954/

P-hacking is fun! (But don't do it)

" "p-hacking" occurs when researchers collect or select data or statistical analyses until nonsignificant results become significant."

- "The Extent and Consequences of P-Hacking in Science", Head et al., 2015

Let's play with it:

538's Hack your way to scientific glory

Frameworks that drive best practices

- Open Science
- Reproducibility / Replicability
- Data Lifecycle

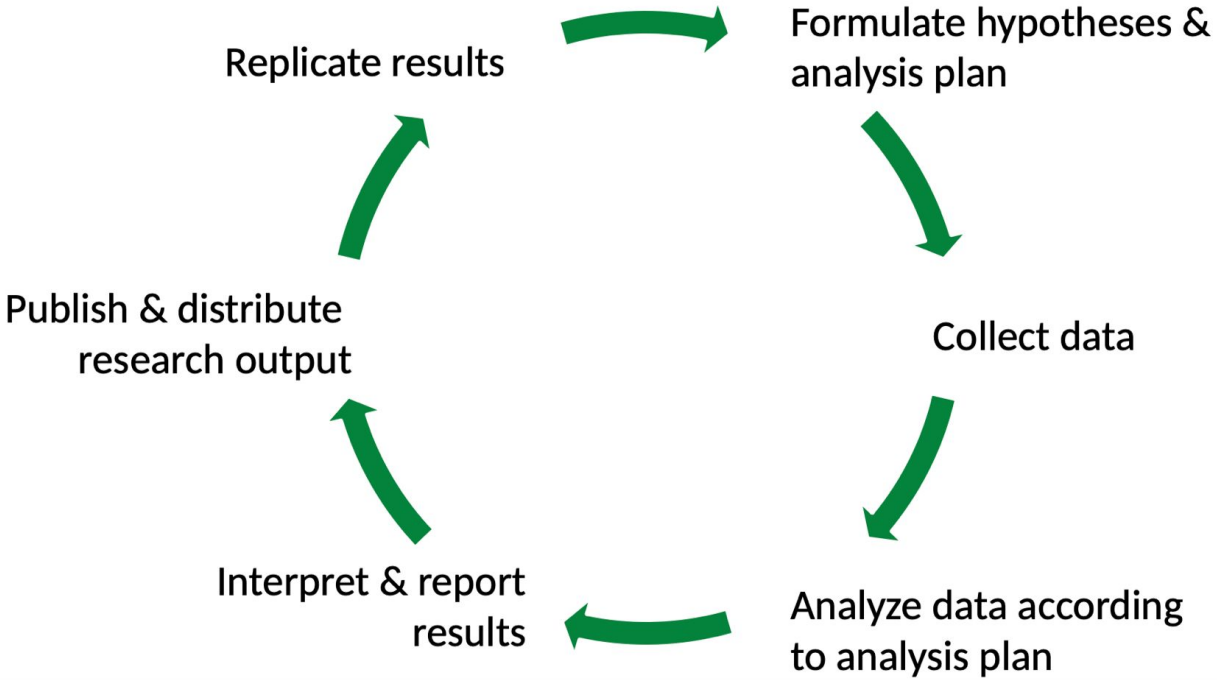
Open Science

“Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks”

- “Open Science now: A systematic literature review for an integrated definition”, Vicente-Sáez & Martínez-Fuentes, 2018

OSC
LMU Open Science Center

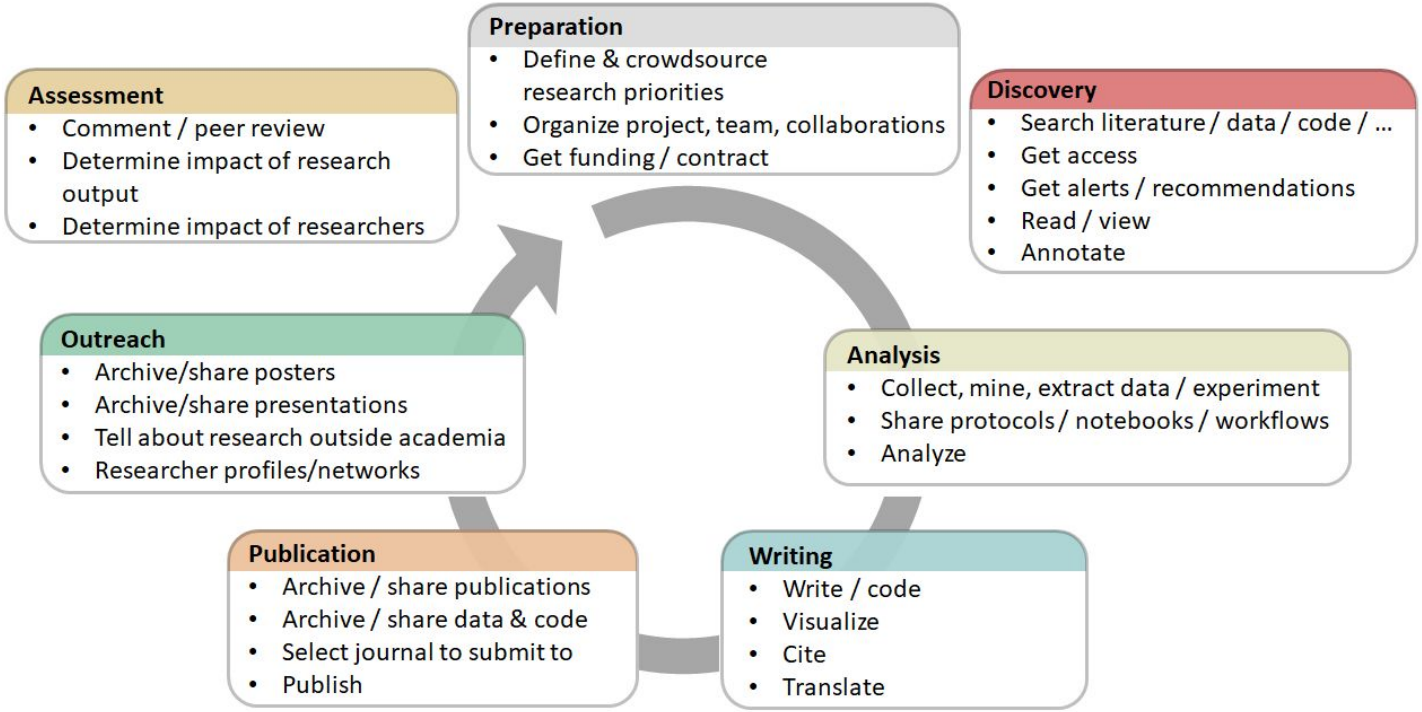
The Confirmatory Research Process



https://osf.io/z7954/

Wagenmakers et al. (2012)

Open Science



Open Science Training Handbook

Reproducible (vs Replicable)

Reproducibility is often defined as:

the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

This is distinct from **Replicability**:

which refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected.

https://open-science-training-handbook.gitbook.io/book/open-science-basics

Reproducible (vs Replicable)

Reproducibility is often defined as:

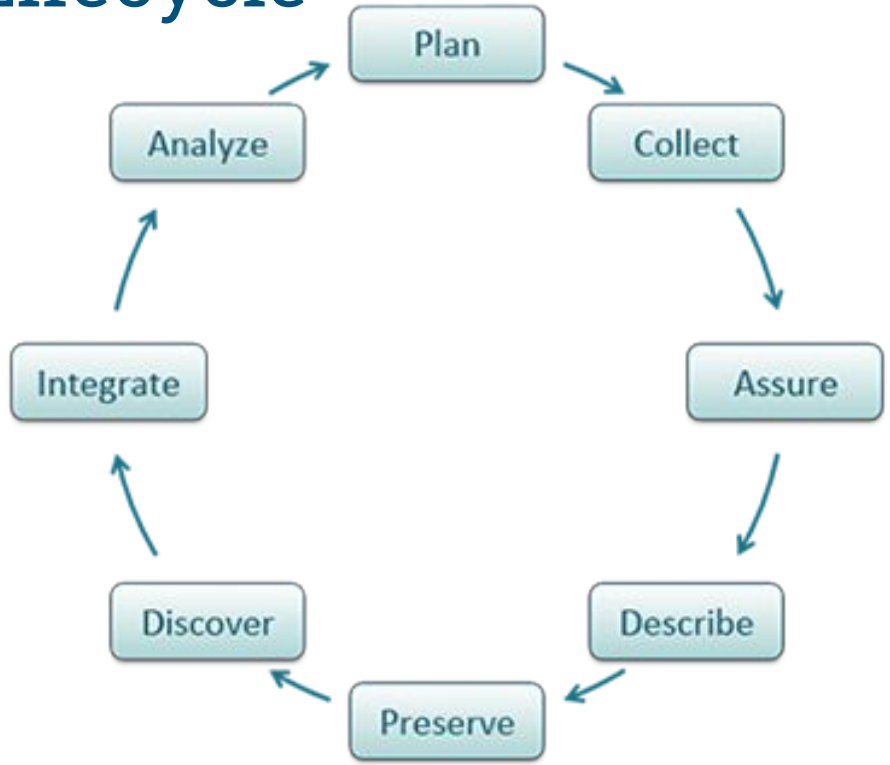
the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results.

This is distinct from **Replicability**:

which refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected.

https://open-science-training-handbook.gitbook.io/book/open-science-basics

The Data Lifecycle



https://www.dataone.org/data-life-cycle/

How Does This All Relate to “Data”?

- Data can be fudged
  - Intentionally
  - Unintentionally (motivated reasoning)
- Data can be lost
- Data will sometimes need updating
- Data will be shared
  - At very least: with advisors and collaborators
  - Required: with scientific community
  - Maximally: with Everybody in the Whole World !!

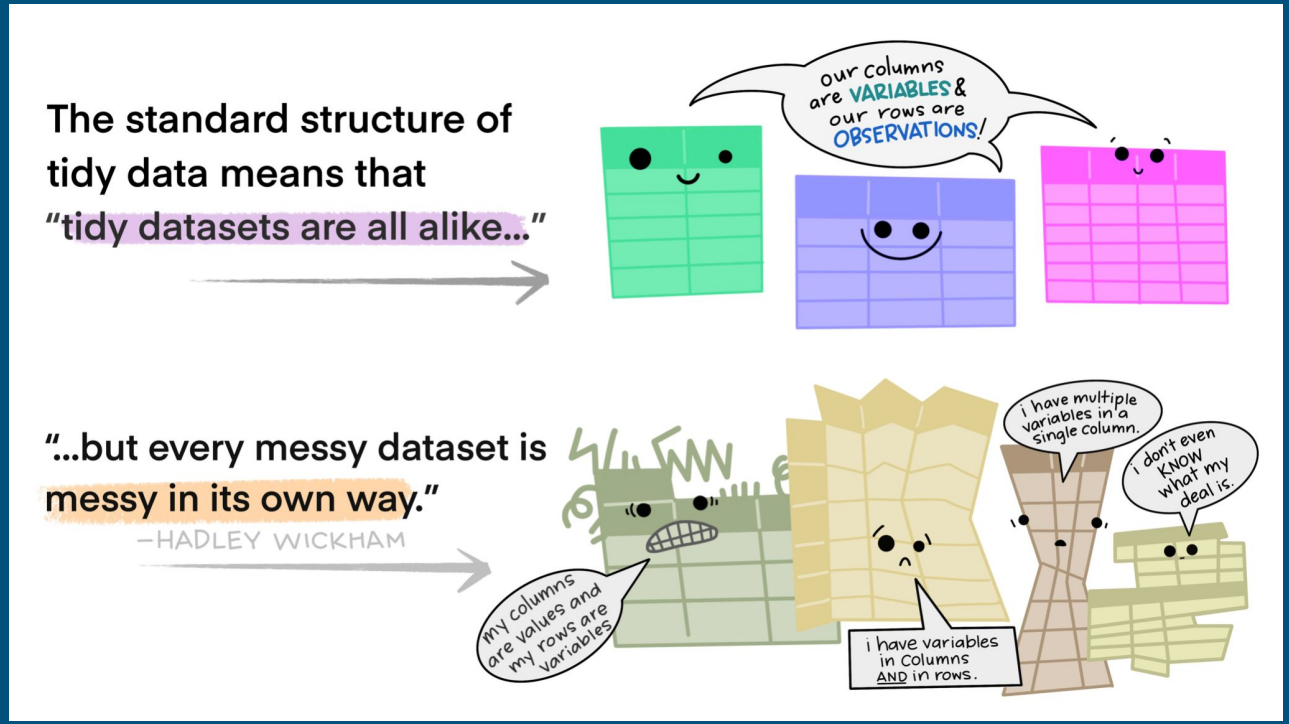
Practices

Preview of Topics

- Organization
  - Naming
  - Spreadsheet Design
- Automation
  - Code
- Record Keeping
  - Metadata
  - Version Control
- Publication
  - Data Archives and Repositories

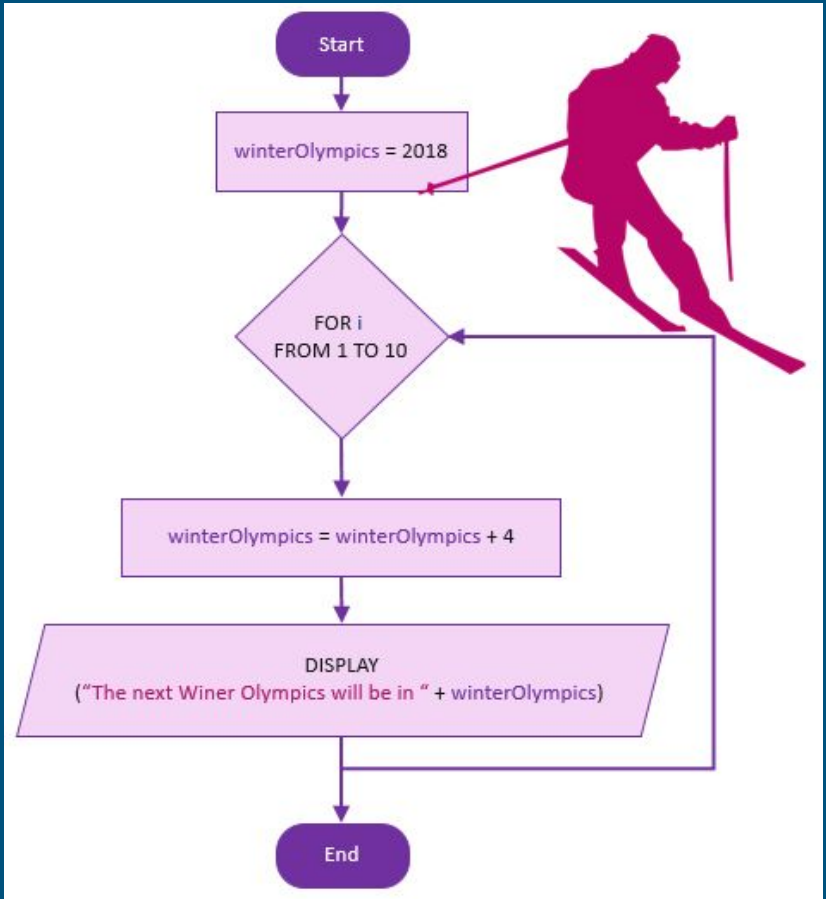
Organization (Tidy Data)

https://www.openscapes.org/blog/2020/10/12/tidy-data/



Automation

- Code reduces errors
- Documented code creates a record of processing steps
- Other people's code = less work!
  - Use standard libraries whenever possible - don't reinvent



Metadata

Data about data ...

Time (PST)	Bottom depth (m)	Sample depth (m)	Temp. (°C)	Salin
6:14:56	777	2	18.64	33.7
2:31:33	3954	2	14.71	33.5
5:39:03	4212	2	17.25	33.3
2:20:07	4186	2	14.68	33.1
1:00:17	4234	2	17.4	33.0
2:23:24	922	2	15.59	33.7
9:13:50	71	2	13.71	33.8

NGA LTER

NGA-LTER

Northern Gulf of Alaska Long-Term Ecological Research

Cruise Report July 2018

Cruise ID: WoIJ2018

Funding Sources: NSF, NPRB, AOOS, EVOS/GWA

Cruise: EDDIES | Leg: 1 | Cast: 8 | Ty

Date: 16 Jun 04 | Time: 0253 | Lat: 33 41.809 | Long: 63 9.912

Date: 16 Jun 04 | Time: 0331 | Lat: 33 41.84 | Long: 63 9.74

N #	Depth	Niskin Temp	Helium	Oxygens $O2_1$	Oxygens $O2_2$ $O2_3$	DIC/Alk	TOC/TON	Salts	Nuts	Bact	
1	0										
2	0	23.8		1		91	91		91		8
3	0										
4	20	23.4		2		92 Cont. inside bottle	92		92		8
5	20										
6	40	21.2		3		93	93		93		8
7	40										
[illegible]	50	20.9		4		94	94		94		[illegible]

Data Archives and Repositories

There's a million of them:

MetaZooGene

EcoTaxa 2.0

DataONE

NCEI
NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

EDI

R2R
ROLLING DECK TO REPOSITORY

BCO-DMO
Biological & Chemical Oceanography Data Management Office

DRYAD

iDigBio
Integrated Digitized Biocollections

Version Control (GitHub)

System that highlights changes to text files …

```
2 ctd_utilities/__init__.py

@@ -21,5 +21,5 @@
21  21  from .ctd_objects.dataset import CtdDataset
22  22
23  23  # functions to manipulate and plot single casts
24     -from .plot_cast import plot_profile
    24 +from .plot_cast import plot_profile, profiles_as_subplots
25  25
```

- Takes time to learn
- Enables collaboration
- Open Source

Conclusions

- Science benefits from Reproducible Research
- Reproducible Research depends on effective data management and analysis
- Effective data management and analysis depends on tools
- We are going to practice using the tools!

Finish-up

Please take a moment to write in the chat …

- Something that surprised you, or
- Something you want to learn more about

- Assignment #1 on GitHub
- Next week we will be at a different Zoom Link
- Go over poll results

Resources

- University of Washington Library Reproducibility Resources
- Open Science Training Handbook
- Open Science Slides, Open Science Foundation (OSF)
- Consolidating Teaching Resources, Tufts University
- Open Science Workshop Materials, LMU Open Science Center
- Initial Steps Toward Reproducible Research
- Data Organization - organizing data in spreadsheets