# LAS 6292: Reproducible Data (Re)Organization

updated: 2021-03-18

## Pre-Class Preparation (Instructor):

**Send in an email to students:**

- content of any pre-class emails.

**Bring to Class:**

- Snacks
- Flip charts and markers
- Dry write markers
- Tent cards for student names
- website R resources
- RStudio shortcuts PC or mac
- Cheat Sheets Rstudio base r,tidyverse, etc.

## Objectives and Competencies:

- Understand and be able to explain the importance of a well-documented, reproducible workflow
- Be able to outline a logical workflow for data importing and correction
- Be able to load a file as a dataframe in RStudio, edit the data frame, and save the corrected version as a csv file

## Pre-class Preparation (Students):

**Readings**

1. Laskowski, 2020. What to do when you don't trust your data anymore.
   [read online] [download pdf]
2. Pennisi, E. 2020. Spider biologist denies suspicions of widespread data fraud in his animal personality research. Science.
   [read online] [download pdf]
3. Alston, J. M., and Rick, J. A.. 2020. A Beginner's Guide to Conducting Reproducible Research. Bull Ecol Soc Am 00( 00):e01801. [read online] [download pdf]

4. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510.
[read online] [download pdf]

**Online Lectures:**

These introductions to R and R Studio were made by Professor Ethan White (UF-WEC). They are a good overview of some R basics.

- Intro to R and RStudio.
- Navigate R and RStudio web page
- Intro to R Packages
- Expressions and Variables in R

**Reminder:** Don't forget to install the `Tidyverse` library; you can read how on this page. And if you want to get ahead to what we will be doing in class, you can read about how to Set up a Project in R

Computer programming is challenging to learn and teach, especially remotely. The next few sessions we will be using modified version of a style known as:

```
(1) I do it
(2) We do it
(3) You do it
```

This means you first observe me doing something in Rstudio by watching a video (or in this case you'll watch a video of Prof. Ethan White). The in class we will do some tasks together at the same time (aka "live-coding"). This is useful, because you can ask questions as you go. Then, you will work on an assignment based on the tasks we did together, which you will submit as your assignment.

## Class Outline

**Topic 1: Intro to Reproducibility: (15 min)**

- Why should we practice 'Reproducible research'?

**Breakout: (10 min)**

- Write the documentation to make a peanut-butter and jelly sandwich

**Returning results: (20 min)**

- Sandwich or No Sandwich?

**Break (10 min)**

**Coding Corrections: Intro (15 min)**

- We're going to practice the routine for reproducible data correction. Applies to any coding problem.
- Step 1: step away from the computer. My first step in the process is to look over the file and map out what needs to be done.
- Remember last time when we looked at the data and found mistakes? Take 5 minutes and do that again.

**Coding Corrections: Implement the Changes(60 min)**

- **Key Points to teach emphasize:**

  1. to keep track of progress, we will try the following: raise hand = I need help, thumbs up = I'm good.
  2. Create an RStudio Project: `data_dirty`, `data_clean`, `scripts`
  3. Move the file into the data_raw folder
  4. New .r file
  5. add key info: session info(), name and what for, etc.
  6. Load tidyverse
  7. import csv
  8. make changes

  - add column
  - change tolower
  - save as csv
  - sweep the environment, restart r, rerun to make sure it works
  - read colnmaes
  - change column names
  - get summary of data types str, glimpse
  - change Data Types
  - change to a column to lower
  - change a column to upper
  - order factor
  - add a column
  - add a calculation
  - Export clean data

**KEY MESSAGES:** (1) Keep raw data raw (2) annotate. lots.

**Coding Corrections: Now do it on your own (50 min)**

https://dataoneorg.github.io/Education/lessons/09_analysis/09_analysis.pdf

Now do it on your own, with a new dataset.

Here is the dataset, and here are the instructions. Can you reproduce the results? don't forget to sweep the environment and restart r

https://www.dataquest.io/blog/load-clean-data-r-tidyverse/

## Tools & Resources

## Sources