Reproducibility

Rosana Zenil-Ferguson

4/5/2021

Useful advice

- Your closest collaborator is you six months ago, and you do not reply to emails -*Karl Broman*
- The most important tool is the mindset, when starting, that the end product will be reproducible -*Keith Baggerly*
- Assume that everything that you are doing right now will need to be redone at some point in the future: be prepared

Five stages of reproducibility

(From Claudia Solis-Lemus PhD)

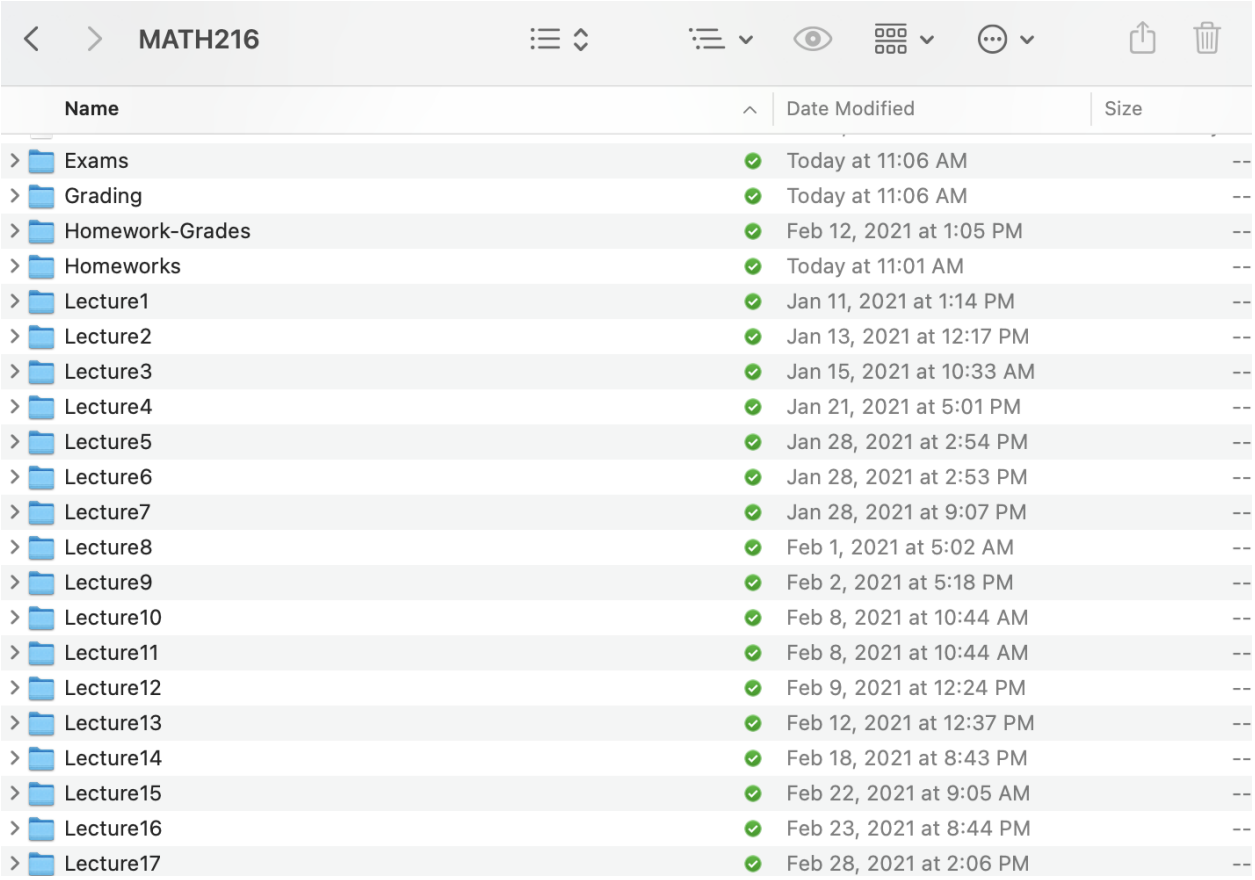
1. Denial: I do not need to be reproducible. I have not kept track of code/scripts in years and I have been just fine. People exagerate. We do not have to be that paranoid
2. Anger: Why do I have to write these stupid notes!? It takes twice the time to write notes and do the work. I could simply do the work! This is stupid and ridiculous! I am just wasting my time with notes and comments that nobody cares about!
3. Bargaining: Well, perhaps it is ok if I only keep notes in the very final script or the very final function. That makes sense. No one needs to know or would even care to read my other code. Yes, maybe it is ok if I only comment at the end on the project

Five stages of reproducibility

(From Claudia Solis-Lemus PhD)

4. Depression: I do not understand my notes. The comments that I made a year ago do not mean anything to me anymore. This has totally failed. I am a reproducibility failure. If I am not able to understand my own notes, no one will
5. Acceptance: I understand that being reproducible is a process. No one does this right the first time. No one does it right period. We are all learning, and all I can do is try my best to make notes/comments and be honest and open about my research process

Basic organizational skills- Directories



MATH216

Name	Date Modified	Size
Exams	Today at 11:06 AM	--
Grading	Today at 11:06 AM	--
Homework-Grades	Feb 12, 2021 at 1:05 PM	--
Homeworks	Today at 11:01 AM	--
Lecture1	Jan 11, 2021 at 1:14 PM	--
Lecture2	Jan 13, 2021 at 12:17 PM	--
Lecture3	Jan 15, 2021 at 10:33 AM	--
Lecture4	Jan 21, 2021 at 5:01 PM	--
Lecture5	Jan 28, 2021 at 2:54 PM	--
Lecture6	Jan 28, 2021 at 2:53 PM	--
Lecture7	Jan 28, 2021 at 9:07 PM	--
Lecture8	Feb 1, 2021 at 5:02 AM	--
Lecture9	Feb 2, 2021 at 5:18 PM	--
Lecture10	Feb 8, 2021 at 10:44 AM	--
Lecture11	Feb 8, 2021 at 10:44 AM	--
Lecture12	Feb 9, 2021 at 12:24 PM	--
Lecture13	Feb 12, 2021 at 12:37 PM	--
Lecture14	Feb 18, 2021 at 8:43 PM	--
Lecture15	Feb 22, 2021 at 9:05 AM	--
Lecture16	Feb 23, 2021 at 8:44 PM	--
Lecture17	Feb 28, 2021 at 2:06 PM	--

Figure 1: My math class folder

Basic organizational skills- Directories

Figure 2: Lecture folder

Tracking everything you do with Rmarkdown

For example, I like it for my class and tutorials because it can write math

```
---
title: "Lecture 4"
author: "MATH 216"
date: "1/20/2021"
output: beamer_presentation
---

## Derivatives
In MATH 215 you learned that a derivative is no other thing that the slope of a tangent line. In biology, we interpret derivatives as **the rate of change** at a given point.


The formal definition of the derivative is

$$f'(a)=\lim_{h \to 0} \frac{f(a+h)-f(a)}{h}$$
```

$$f'(a)=\lim_{h \to 0} \frac{f(a+h)-f(a)}{h}$$

```
However, for the purpuse of this class we will just revise the essential derivatives and rules.

## Tangent of a function-Rate of change
```

Derivatives
Tangent of a func…
Interpreting deriv…
Example 1. Deriv…
Example 1.
Basic differentiati…
Basic differentiati…
Basic differentiati…
New derivatives f…
New derivatives f…
Example 4: The r…
Example 4: How …
New derivatives f…

Figure 3: Writing math in Rmarkdown

Tracking everything you do with Rmarkdown

It can also accumulate all your code

```
## Example 1. Covid-19 pandemic: Exponential Growth
When the speed of growth is *proportional to the size of the population*, that's exponential growth
```{r, message=FALSE}
Date<-seq(as.Date("2020-02-23"), as.Date("2020-03-23"), by="days")
Cases<- c(10,11,13,18,22,30,42,47,69,109,164,220,271,352,412,469,617,876,1292,1766,2244,2605,3047,3657,4427,5426,6479,7738,8934,10312)
covid_UK<-data.frame(Date,Cases)
```

```{r, echo=FALSE, results='asis'}
library(knitr)
kable(covid_UK[1:10,],caption="Covid-19 cases in the UK")
```

```{r, message=FALSE}
log.covid_UK<-data.frame(Date,log(Cases))
```
Example 1. Covid-19 pandemic: Exponential Growth
```

Models of Popula…
Exponential growth
Example 1. Covid…
Example 1. Covid…
Example 2. Diver…
Logistic Growth d…
Example: Logistic…
Example: Logistic…

Figure 4: R code in Rmarkdown

To git or not to git

- Personally: I don't have a preference. I think you can be as organized as possible in folders as long as you comment your code
- If you don't use git as often, use dropbox- $10 a month (I don't think about backups)

The basic data vs. the data transform

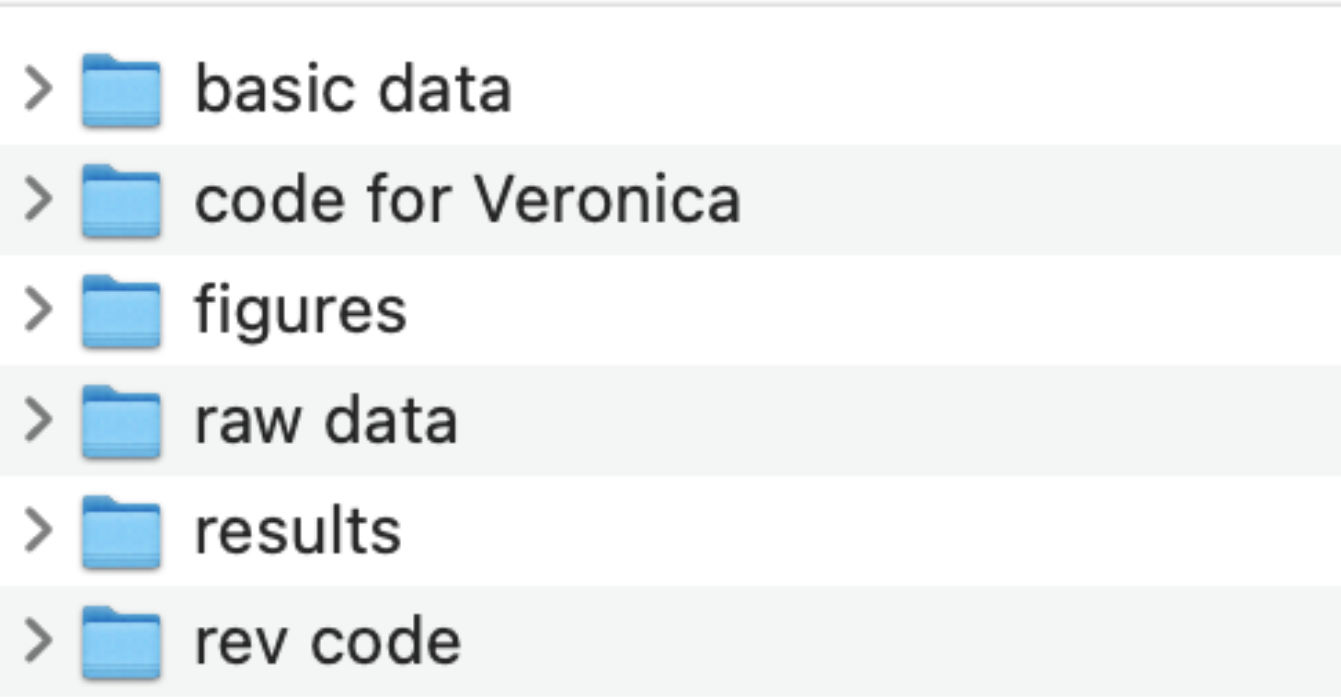


Figure 5: Raw data folder: Do not touch!

The data transform file

datatransform.R

datatransform.R 〉 No Selection

```r
library("phytools")
library("treeplyr")
# Last Updated: 10/22/2020
# Read original tree 106 tips
thal.tree<-read.tree("~/Dropbox/thaldiv/raw data/Thal106.divtime.tre")
plotTree(thal.tree, fsize=0.4)
#write.csv(thal.tree$tip.label,"~/Dropbox/thaldiv/raw data/tiplabel.csv")
# Created a dataset with labels that can be matched
## ? there are some repeated taxa with different accessions I assume. Which one to match??
labelstomatch<-read.csv("~/Dropbox/thaldiv/raw data/originaltiplabel.csv",header=TRUE)
thal.tree$tip.label<-labelstomatch[,3]
# Plot tree with matching labels
plotTree(thal.tree, fsize=0.4)


#Created a simpler CSV file containing the basic data
#### pollination
#    Insect=0
#    Wind=1
#### sexual_system
#    Hermaphroditic= 0
#    Dioecy= 1
#    Andromonoecy= 2
#    Gynomonoecy=3
#### ploidy
#    Diploid=0
#    Polyploid=1
##########
### all unknowns marked as ? and all with both traits marked as (0 1)
```

Figure 6: datatransform.R