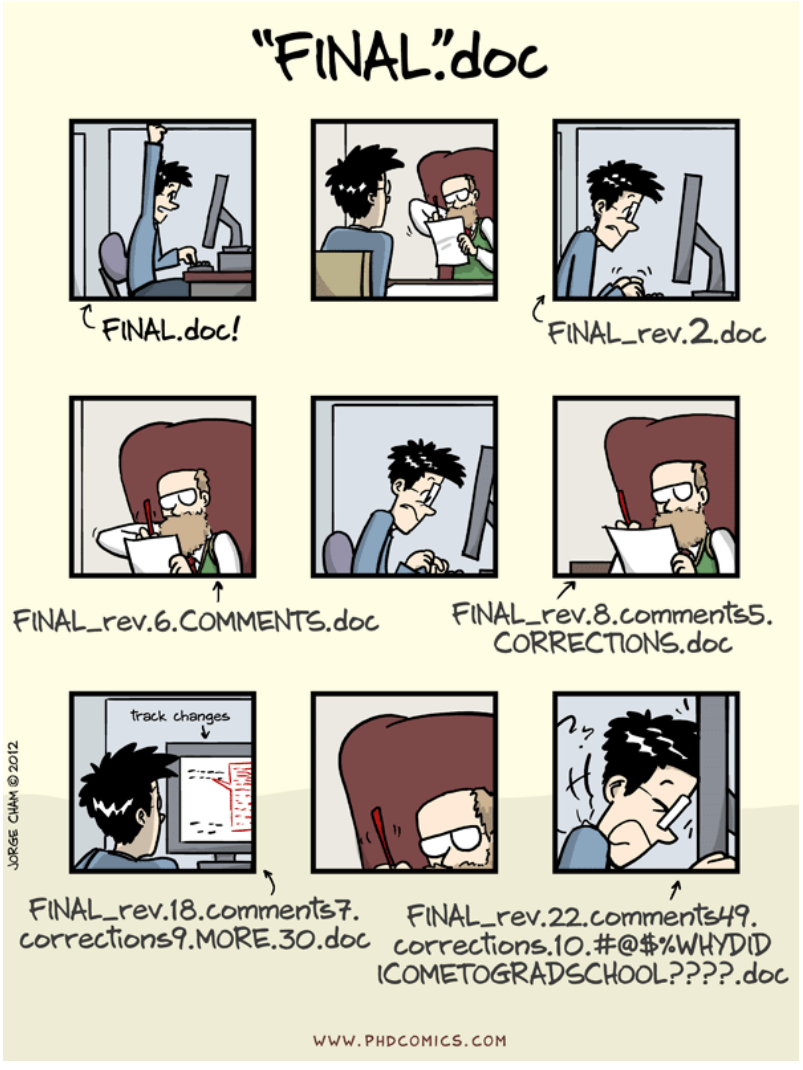
Developing a Workflow to Maximize Reproducibility and Research Impact: Managing Data, Computer Code, and Projects for Success

John R. Fieberg & Althea A. ArchMiller

4/11/2017

Why worry about reproducibility?

Working towards future reproducibility makes my code easier for my collaborators (and me) to read, run, and debug today,
and that's why I think reproducibility is a **win-win for all researchers**."
-Althea



Why worry about reproducibility?

"[Reproducibility] provides security, saves time, and forces me to be more thoughtful about my workflow." - Ethan Young

- make your life easier! Now, and in the future
- collaborations
- broader research impact
- increased citations
- transparency
- grant and journal requirements

Is my research reproducible?

- Are your research documents stored in these formats?
  - .csv
  - .txt
  - .pdf
  - .html
  - .R/.Rdata
    - YES!
  - .doc/.docx
  - .sas
  - .xls/.xlsx
  - any other proprietary file format
    - NO!

Is my research reproducible?

- Is your code linear?
  - Clear environment often and at beginning of script
  - Don't save .Rdata or history
  - Each program should focus on one main task or analysis
  - Don't rely on manual commenting/uncommenting

Is my research reproducible?

- Is your code linear?
  - Clear environment often and at beginning of script
  - Don't save .Rdata or history
  - Each program should focus on one main task or analysis
  - Don't rely on manual commenting/uncommenting

So, what's wrong here?

```
# What variables are significant?
lm.out <- lm(weight ~ height, data = trial.data)
remove(lm.out) # clear previous lm.out for each
               #   new lm() definition above

# Is the relationship significant?
#     (If not, clear and try a new regressor)
summary(lm.out)
```

Is my research reproducible?

- Are your files easily shared with others?
  - Organized directory structure
  - Files relatively linked
  - Well-documented & commented
  - Consistency in coding practices

"The point of having style guidelines is to have a common vocabulary of coding so people can concentrate on *what* you are saying, rather than on *how* you are saying it." - Google's R Style Guide

Workshop Outline

The goal for this workshop is to help you develop the tools to develop a workflow to maximize reproducibility, collaborations, and research impact.

1. RStudio Projects for organizing data, code, and output

Workshop Outline

The goal for this workshop is to help you develop the tools to develop a workflow to maximize reproducibility, collaborations, and research impact.

1. RStudio Projects for organizing data, code, and output
2. R-Markdown and R-Oxygen for documenting your code and creating reproducible reports

Workshop Outline

The goal for this workshop is to help you develop the tools to develop a workflow to maximize reproducibility, collaborations, and research impact.

1. RStudio Projects for organizing data, code, and output
2. R-Markdown and R-Oxygen for documenting your code and creating reproducible reports
3. GitHub for version-control, collaborating and archiving

1. RStudio Projects

Think about a typical research project, maybe a dissertation chapter or an experiment that you've managed from data collection through publication. What are typical **folders** that you've used?

1. RStudio Projects

Think about a typical research project, maybe a dissertation chapter or an experiment that you've managed from data collection through publication. What are typical **folders** that you've used?

- Raw data
- Processed data
- Analysis scripts
- Paper/Manuscript-related documents
- Sharing documents ("transmittals")
- Metadata
- Maps or other deliverables

RStudio Projects provide an opportunity for you to organize and manage all of these types of folders in **one place** in a way that **relatively links** everything together and **eases sharing**.

1. RStudio Projects

Think about a typical research project, maybe a dissertation chapter or an experiment that you've managed from data collection through publication. What are typical **folders** that you've used?

- Raw data
- Processed data
- Analysis scripts
- Paper/Manuscript-related documents
- Sharing documents ("transmittals")
- Metadata
- Maps or other deliverables

RStudio Projects provide an opportunity for you to organize and manage all of these types of folders in **one place** in a way that **relatively links** everything together and **eases sharing**.

Up next, Activity 1!

Activity 1: Data management and updating

Here, we will read in and process three weeks of experimental data and do some preliminary analysis. Then, we will get a final (4th) week of data, which we will merge with the original data.

The goals are to:

1. Be introduced to RStudio
2. Create a framework for keeping data organized and up-to-date
3. Automatically update our analyses based on the master dataset

Context: Abundance data from ~75 invertebrate species sampled on various beaches along the Dutch coast.

Zuur, A.F., E.N. Ieno, and G.M. Smith (2007) Analysing Ecological Data. Springer, New York.

Activity 1: Data management and updating

Before we begin today, we need sync your individual versions of the workshop documents with Althea's master branch:

Activity 1: Data management and updating

Before we begin today, we need sync your individual versions of the workshop documents with Althea's master branch:

1. Open RStudio and your reproducibility_workshop.rproj. (File > Open Project...)
2. Open shell (Tools > Shell...)
3. Type in exactly, then press enter:

```
$ git fetch upstream
```

4. Type in exactly, then press enter:

```
$ git checkout master
```

5. Type in exactly, then press enter:

```
$ git merge upstream/master
```

Activity 1: Data management and updating

Now create a new folder in student_folders/ for all of today's activities. Name the folder after yourself (or an alias).

Open a new R Script file and save it to that new folder as **"activity1a_data_processing.R"**

First, we will read in first three weeks of data and combine them, process the data a little bit, and save the merged/processed data for analysis.

Secondly, we will save another new R Script file as **"activity1b_data_analysis.R"** and do (preliminary) regression analysis.

Finally, we will pretend to have just gotten the final week's data in and update everything in a "reproducible" way.

1. RStudio Projects

Other links

`https://swcarpentry.github.io/r-novice-gapminder/02-project-intro/`

Data Mangement Tips

- Treat data as read-only
  - Don't use Excel, etc, to manipulate raw data
  - Use a single R program for all manipulation
  - Save "cleaned" or "processed" data in easily loadable formats
- Differentiate data types with folders *raw* versus *processed* versus *output* (e.g., linear regression objects, etc)
- Write dates in YYYYMMDD or equivalent format

Tips

- Don't use github with large files :-(
- Create new projects in GitHub first, then sync them with RStudio

Why R-Markdown for manuscripts?

"I can do reproducible work in R (making me happy) and format the output report in Word (making my collaborators happy)" - Richard Layton `http://rmarkdown.rstudio.com/articles_docx.html`