Notes for Participants

Keith Baggerly

10/19/2018

Table of Contents

# Overview

Greetings!

You’re getting this note because you’re a potential participant in the 2018 FDA Short Course on Reproducible Research (RR), and I wanted to let you know

* [what it’ll cover](#what-well-cover),
* [what you’ll need](#what-youll-need) to get the most out of the course, and
* [where to go](#where-to-get-more-details) on GitHub to get more info, including links to slides and code as we get closer to the course date.

One quick note about terminology. We distinguish between reproducibility and replicability. We say a finding is

* **reproducible** if someone else, given the same data you started with, can readily obtain the same results
* **replicable** if someone else, when running a qualitatively similar study with new data, is very likely to get results similar to those you reported

Most of the course will be about the former, but we will touch on the latter.

# What We’ll Cover

The course is split into 4 parts, which vary in length and the level of background required.

* **Part 1** (1.5hr, general audience): *Why it’s important*. We use in-depth analyses of a few problematic case studies to show how basic mistakes can and have negatively impacted patient care. We’ll apply forensic bioinformatics to reverse engineer what must have happened in these cases, and discuss the breadth of the problem.
* **Part 2** (4 hr, aimed at folks analyzing data): *How to Do It Right - Workflows*. We’ll introduce ways of organizing data, code, and text in ways that will make it easier for others to get the same results you got, and tools available today which make implementing these practices easier. We’ll do some of this live with toy examples, and examine more realistic examples with prebuilt analyses. My own examples will use R and RStudio, but the general concepts presented here can be adapted to other programming environments. A short topic list includes
  + literate programming with markdown/markdown
  + working on the web with git/github
  + README files
  + folder structures
  + projects, encapsulation, and here
  + gathering raw data
  + make, appoximations, and file naming
  + report structure
  + sanity checks
* **Part 3** (2 hr, data analysts using R): *How to Do It Right - R Packages*. For folks who are doing analyses in R, R packages are everywhere. Recent packages, specifically devtools and roxygen2, have made it much easier to write even more new packages, so you can now write a package in 15 minutes or so (with some practice). We’re going to show you how, and then use what we’ve learned to share scripts that make organization easier. We’ll discuss
  + the structure of R packages
  + DESCRIPTIONs, LICENSEs, and documentation
  + vignettes
  + storing data in packages
  + storing templates in packages
  + storing analyses in packages
* Part 4 (0.5hr, general audience): *Looking at Replicability*. Even when findings are reproducible, they may not be replicable. We review some common problems leading to poor replicability, and discuss various sanity checks we can try putting in place to let us catch some of the bigger ones.

# What You’ll Need

Parts 1 and 4 will involve me presenting, so you’ll need an open mind and a willingness to ask questions.

Parts 2 and 3 will involve a good deal of live demo and opportunities to try it yourself, so you’ll want to have a laptop you can create and edit files on with recent versions of the relevant software (all free) preinstalled. Specifically, you’ll want R, RStudio, and a bunch of R packages.

## R and RStudio

On my MacBook Pro laptop (OS X 10.13.6) I’m currently running

* R version 3.5.1 (2018-07-02) and
* RStudio version 1.2.1047

R is available from CRAN, <https://cran.r-project.org/>; a list of CRAN mirrors is available from the R Project, <https://www.r-project.org/>.

RStudio 1.2 is currently in “preview release”, and is available here <https://www.rstudio.com/products/rstudio/download/preview/>

RStudio 1.1, the “official version”, is available here <https://www.rstudio.com/products/rstudio/download/>

I’d encourage you to get 1.2 for two reasons. First, it’s likely to be officially released quite soon, <https://community.rstudio.com/t/rstudio-1-2-release-date-interval/9522>. Second, it includes better support for Python via the reticulate package, and the capability to produce more types of output files, including PowerPoint.

## R Packages

The main R packages I’ll use (in alphabetic order, with my version numbers shown) are:

* devtools 1.13.6
* downloader 0.4
* here 0.1
* knitr 1.20
* readr 1.1.1
* rmarkdown 1.10
* roxygen2 6.1.0

I may use other packages from the tidyverse without extensive discussion:

* tidyverse 1.2.1

### Windows and RTools

For those of you running Windows machines, you’ll also need to install

* [Rtools](https://cran.r-project.org/bin/windows/Rtools/)

for Part 3 of the course, in order to get new packages you create to compile properly. Jeff Leek gives a slightly more expansive description of installation [here](http://jtleek.com/modules/01_DataScientistToolbox/02_10_rtools/).

### tinytex and PDF Output

The rmarkdown package uses LaTeX formatting to generate pdf versions of reports (as opposed to figures), so if you want to generate pdf reports, you’ll need to have some version of LaTeX installed on your computer.

If you don’t have LaTeX installed, probably the easiest way for R users to address this is by using the tinytex package. Installation and use of tinytex is described here:

<https://yihui.name/tinytex/>

### Bonus: Git and GitHub

As we go along, I’ll be saving my work using the [git](https://git-scm.com/) version control system, and posting materials to the course’s repository on [github](https://github.com/) both so you’ll all be able to access everything after the course and so you’ll be able to see web interactivity and sharing “live”. We won’t discuss the mechanics of these tools much in the course itself due to time constraints, but you may want to install git on your machine and set up an account on github if you find the illustrated usages interesting.

# Where to Get More Details

I’ve set up a web page for the course at GitHub:

<https://github.com/kabagg/fda_2018_rr>

These notes are available from the website now.

As they become available, this page will have links and/or text added for

* a more detailed course syllabus
* places for further reading
* datasets for download
* R packages I might’ve missed above
* presentation slides

Other things I think of or have suggested to me.