SISBID 2017 Module 3: Reproducible Research

Keith Baggerly and Karl Broman

July 17-19, 2017

This module is part of the Summer Institute in Statistics for Big Data!

Taught by

Keith A. Baggerly kabagger@mdanderson.org

and

Karl Broman kbroman@biostat.wisc.edu

Course Goals

- pdf
- html
- MS Word
- Rmd Source

Homework

- pdf
- html
- MS Word
- Rmd Source

Cheat Sheets

Karl's Software Carpentry Course

These are from RStudio's list

- Rmarkdown; there's also a reference guide
- Package Development/Devtools

There are many other sheets there (including some for user contributions and translations), so check it out!

These are from GitHub

- Git/GitHub
- Git
- Links to Translations
- More Resources

Course Syllabus and Lecture Materials

Day 1, Jul 17, 2017

Session 1, 8:30-10

Lecture 0, Basic Intro, Keith, 5-10 min pdf, printable pdf Introduction to the course, administration, course goals Definitions - reproduction vs replication

Lecture 1, Intro and Common Problems, Karl, 40 min pdf, printable pdf An introduction to reproducible research by way of commonly encountered problems, plus discussion of the organization of project files

Lecture 2, A Train Wreck, Keith, 40 min pdf, printable pdf A case study describing just how bad things can get, with clinical implications

Session 2, 10:30-12

Lecture 3, R Markdown and Literate Programming, Karl, 45 min RMarkdown notes, Rmd example An introduction to R markdown, RStudio, and Literate Programming, with examples illustrating how to produce reproducible reports

Homework part 1, participants, 45 min Set up the analysis folder, write the preprocessing script in R markdown, compile to html / pdf / word

Session 3, 1:30-3

Lecture 4, R Packages, Keith, 45-60 min (much live demo) pdf, printable pdf How to write R packages quickly and easily with devtools, roxygen2, rmarkdown, and knitr - overhead, code, data, vignettes, clean code, and templates

Homework part 2, participants, 30 min writing a basic package

Session 4, 3:30-5

Lecture 5, EDA, Big Jobs, and Automation, Karl, 75 min (includes some short activities) pdf, printable pdf, activity 1 spin example, activity 2 caching example Capturing exploratory data analysis, handling the challenges arising when data or jobs are big enough to make rerunning unpleasant or infeasible, and a brief introduction to automation with GNU Make

Lecture 6, Vitamin D, Keith, 10-15 min pdf, printable pdf Discussion of how recommendations are set, and reconstruction of analyses obscured by lack of code and misapplied terminology. Linked to course homeworks.

R package sisbid3, with a vignette on adding data to R packages just the vignette report fitting logistic regression to Priemel et al

Day 2, Jul 18, 2017

Session 5, 8:30-10

Lecture 7, Problems with Replication, Keith, 40 min pdf, printable pdf A review of several factors which can make results harder to replicate (be seen again with new samples) vs hard to reproduce (starting from the same raw data)

Lecture 8, Git on your Computer, Keith, 50 min, mostly live pdf, printable pdf Using git to track files and versions; init, status, add, commit, branch, checkout, merge

Session 6, 10:30-12

Lecture 9, Git with GitHub, Keith, 45 min pdf, printable pdf Introducing GitHub, perhaps the "killer app" for git; working with remote repositories, bare repos, cloning, pull, push

Homework, participants, 45 min Establishing a repo at GitHub Post your package to GitHub

This session will be a mixture of lecture and live demo.

Session 7, 1:30-3

Lecture 10, Collaborating with Git, Keith, 45 min pdf, printable pdf Working with others, making comments, providing feedback, fixing errors

Homework, participants, 45 min Working with your neighbor's repos

Session 8, 3:30-5

Homework, participants, 45 min Add comments and vignettes to your package on GitHub

Lecture 11, Implementing RR at MDACC, Keith, 45 min pdf, printable pdf A review of ongoing efforts within the biostat department at MD Anderson to produce reproducible reports, and how we took a report written a few years ago using a mix of R and Stata and revamped it in R/rmarkdown to emulate not just the results but also the "look and feel" of the initial MS word output. Hits on tables and figures in rmarkdown, references, reformatting headers.

Day 3, Jul 19, 2017

Session 9, 8:30-10

Lecture 12, Writing Good Reports, Keith, 45 min pdf, printable pdf The "non-codeable" parts of reproducibility - trying to increase the odds your collaborators will understand what it is you're trying to do.

Homework, participants, 45 min Automating common tasks with templates - report structures, directory structures, and look and feel

Session 10, 10:30-12

Lecture 13, Summary and Wrapup, Karl, 45 min pdf, printable pdf Maintaining the Mindset, summary of the course, sharing your work, some links to resources, and brief discussions of topics that we didn't cover

Final Class Discussion

Evals, participants, 5 min

Recommended Reading/Browsing

General

- Christopher Gandrud, Reproducible Research with R and Rstudio, 2e (2015)
- Hadley Wickham, R Packages (2015)
- Yihui Xie, Dynamic Documents with R and knitr, 2e (2015)
- Karl Broman's Tools for RR Course
- Other links: `resources.md`