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](https://www.biostat.washington.edu/suminst/sisbid)!

Taught by

Keith A. Baggerly  
[kabagger@mdanderson.org](mailto:kabagger@mdanderson.org)

and

Karl Broman  
[kbroman@biostat.wisc.edu](mailto:kbroman@biostat.wisc.edu)

# Course Goals

* Course Goals  
  [pdf](RRcourse_goals.pdf), [html](RRcourse_goals.html), [MS Word](RRcourse_goals.docx), [Rmd Source](RRcourse_goals.Rmd)

# Homework

* Homework  
  [pdf](homework.pdf), [html](homework.html), [MS Word](homework.docx), [Rmd Source](homework.Rmd)

# Cheat Sheets

Karl's [Software Carpentry Course](https://kbroman.wordpress.com/2015/04/29/cheat-sheets-for-r-based-software-carpentry-course/)

These are from RStudio's [list](https://www.rstudio.com/resources/cheatsheets/)

* [Rmarkdown](http://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf); there's also a [reference guide](http://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf)
* [Package Development/Devtools](http://www.rstudio.com/wp-content/uploads/2015/06/devtools-cheatsheet.pdf)

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

These are from GitHub

* [Git/GitHub](https://services.github.com/on-demand/downloads/github-git-cheat-sheet.pdf)
* [Git](https://education.github.com/git-cheat-sheet-education.pdf)
* [Links to Translations](https://services.github.com/resources/cheatsheets/)
* [More Resources](https://help.github.com/articles/git-and-github-learning-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](2017_03_00_SISBID_basic_intro.pdf), [printable pdf](2017_03_00_SISBID_basic_intro_printable.pdf)  
Introduction to the course, administration, course goals  
Definitions - reproduction vs replication

**Lecture 1, Intro and Common Problems, Karl, 40 min** [pdf](2017_03_01_SISBID_introduction_slides.pdf), [printable pdf](2017_03_01_SISBID_introduction_printable.pdf)  
An introduction to reproducible research by way of commonly encountered problems

**Lecture 2, A Train Wreck, Keith, 40 min** [pdf](2017_03_02_SISBID_train_wreck.pdf), [printable pdf](2017_03_02_SISBID_train_wreck_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** [Rmd example](2017_03_03_SISBID_example.Rmd) [md source](2017_03_03_SISBID_Rmarkdown.md)  
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)**  
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, Big Jobs, Karl, 75 min (includes some workalong activities)** [pdf](2017_SISBID_03_05_bigjobs_slides.pdf), [printable pdf](2017_SISBID_03_05_bigjobs_printable.pdf), [activity 1 spin code](2017_SISBID_03_05_bigjobs_activity1_spin.R), [activity 2 caching Rmd](2017_SISBID_03_05_bigjobs_activity2_cache.Rmd)  
A discussion of challenges arising when data or jobs are big enough to make rerunning unpleasant or infeasible

**Lecture 6, Vitamin D, Keith, 10-15 min** [pdf](2017_SISBID_03_06_vitamin_d.pdf), [printable pdf](2017_SISBID_03_06_vitamin_d.pdf)  
Discussion of how recommendations are set, and reconstruction of analyses obscured by lack of code and misapplied terminology. Linked to course homeworks.

## Day 2, Jul 18, 2017

### Session 5, 8:30-10

**Lecture 7, Problems with Replication, Keith, 40 min**  
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**  
Using git to track files and versions; init, status, add, commit, branch, checkout, merge

### Session 6, 10:30-12

**Lecture 8, Git with GitHub, Keith, 45 min**  
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 9, Collaborating with Git, Keith, 45 min**  
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 10, Implementing RR at MDACC, Keith, 45 min**  
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 11, Writing Good Reports, Keith, 45 min**  
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 12, Summary and Wrapup, Karl, 45 min** [pdf](2017_SISBID_03_05_summary_slides.pdf), [printable pdf](2017_SISBID_03_05_summary_printable.pdf)  
Maintaining the Mindset

**Final Class Discussion**

**Evals, participants, 5 min**

# Recommended Reading/Browsing

## General

* [Christopher Gandrud, Reproducible Research with R and Rstudio, 2e (2015)](http://www.amazon.com/Reproducible-Research-Studio-Second-Chapman-ebook/dp/B010ACWGBI/ref=tmm_kin_title_0?_encoding=UTF8&sr=&qid=)
* [Hadley Wickham, R Packages (2015)](http://www.amazon.com/R-Packages-Hadley-Wickham-ebook/dp/B00VAYCHL0/ref=pd_sim_351_6?ie=UTF8&refRID=1E8HS30WBHRCW45SEWXM)
* [Yihui Xie, Dynamic Documents with R and knitr, 2e (2015)](http://www.amazon.com/Dynamic-Documents-knitr-Second-Chapman-ebook/dp/B00ZBYPJEW/ref=tmm_kin_title_0?_encoding=UTF8&sr=&qid=)

[Karl Broman's Tools for RR Course](http://kbroman.org/Tools4RR/)