R for Scientific Visualization, Collaboration and Thinking

Daniel Hammarström

Invalid Date

Table of contents

Preface

The purpose of this collection of workshops is to provide basic skills and understanding of scientific visualization, collaboration, and thinking. These concepts are often not taught explicitly in postgraduate courses. Instead, they are assumed to be known and understood or somehow magically infused in students through their respective scientific environments. Instead of relying on magic, professional skills should be taught and developed continuously, explicitly, and systematically.

The first workshop deals with data visualization which is a fundamental skill for any scientist. Data visualization, like writing, serves multiple purposes in day-to-day scientific practice. Making graphs before data collection can help the researcher better formulate the scientific problem or explain core concepts for oneself, committees or funding agencies. Visualizations will also be important for exploratory data analysis, for discovering problematic data formatting or important patterns related to the scientific problem. Lastly, visualizations will aid in communicating scientific results to other researchers and the public.

Two workshops are devoted to collaboration in data-intensive science. The first workshop deals with the basics of version control and the second workshop deals with structuring collected data into data packages for analysis and sharing. Version control is a way to keep track of changes in a set of files. It also provides opportunities for effective collaboration in complex projects. Using the R package system for data management and sharing is a way to ensure that data is well documented and downstream analyses are reproducible.

The last workshop deals with how we can develop scientific thinking using scientific models and data simulations. We will use graphical models to draw out assumptions and explore simulations as a way to test the robustness of our models and plan experiments.

1 Data visualization

1.1 A grammar of graphics - `ggplot2`

1.1.1 Building blocks of a plot

Wickham (2010)

2 Version control and collaborative scientific coding

2.1 Introduction to git and GitHub

2.2 A workflow for a scientific paper

2.2.1 The final result

2.2.2 Initial setup

2.2.3 Repositories and branches

2.2.4 Milestones, issues, and pull requests

2.2.5 A style guide for scientific coding

We adhere to the tidyverse style guide. The following is a summary of the most important points with some modifications.

2.2.6 Publishing

3 R packages for data management and sharing

3.1 Increased reproducibility

A basic principle in reproducible research is that the data and code used to generate results should be made available to others.

3.2 R package basics

3.2.1 Incorporating data into R packages

3.2.2 Documenting data

3.3 Sharing packages

4 Scientific models and data simulations

References

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19 (1): 3–28. https://doi.org/10.1198/jcgs.2009.07098.