

R, Interactive Graphics, and Data Visualization

Jason Heppler and Lincoln Mullen

DHSI 2016

Course description

The visualization of historical and literary data has become a common practice in digital humanities, drawing on older traditions of visualizing in these disciplines. A variety of out-of-the-box tools exist for easily jumping in to data and information visualization, but when we use these tools we run the risk of research questions being wedged into a tool rather than the tool fitting the research. This course introduces students to humanities visualizations, using a programming language that let researchers prioritize their questions over the requirements of ready-made tools. Students will learn how to iteratively create plots and maps using the R statistical programming language, as well as how to manipulate data so that it can be visualized. Students will become familiar with the entire pipeline of visualization—from data manipulation to exploratory graphics to online interactive visualizations. In addition, the course will offer an introduction to Shiny, a framework for writing interactive websites with analysis in R.

Syllabus

Day 1 | Static visualization with ggplot2

Discussions will include an overview of how humanities scholars are using R (and other languages like Python), the grammar of graphics and the ggplot2 package, and the selection of a visualization project for attendees to work on during the course.

Morning

- What is R? Why R?
- The grammar of graphics
- Basic plots and data structures

Afternoon

- Intermediate plotting with ggplot2
- Reproducible research with rmarkdown documents
- Lab: Installing R, RStudio, and packages (as necessary)

Readings

Hadley Wickham, “Introduction,” *R for Data Science* (O’Reilly, 2016).

Paul Torfs & Claudia Brauer, “A (very) short introduction to R” (2014).

Taylor Arnold and Lauren Tilton, “Set-Up” and “A Short Introduction to R,” *Humanities Data in R* (Springer, 2015)

Day 2 | Data manipulation with dplyr and tidyr

Discussions will include best practices for data formats. Attendees will gather and manipulate the data for their projects.

Morning

- Data manipulation verbs from dplyr and tidyr

Afternoon

- More data manipulation
- Lab: Gathering, cleaning, and tidying data for the course

Readings

Hadley Wickham, “Tidy Data and Tidy Tools,” NYC Open Statistical Computing Meetup, Dec. 2011 (video).

Hadley Wickham, “Tidy data,” *R for Data Science* (O’Reilly, 2016).

Day 3 | Interactive graphics with htmlwidgets

Discussions will include an overview of how humanities scholars are creating online visualizations, and the use of JavaScript visualization libraries in R via the htmlwidgets package. Attendees will begin the online portion of their visualization.

Morning

- Time series with dygraphs
- Interactive ggplots with ggiraph
- Network graphics

Afternoon

- Maps with leaflet

Relevant Readings

Taylor Arnold and Lauren Tilton, “Geospatial Data,” *Humanities Data in R* (Springer, 2015)

Robin Lovelace and James Cheshire, “Introduction to Spatial Data and ggplot2,” Spatial.ly, Dec. 9, 2013.

Steven Brey, “Working with Geospatial Data” and Bethany Yollin, “Working with Geospatial Data (and ggplot2)”

Shawn Graham, Ian Milligan, and Scott Weingart, “Network Analysis” and “Network Analysis Fundamentals,” *The Historian’s Macroscope* (2015).

Scott Weingart, “Demystifying Networks, Parts I & II,” *Journal of Digital Humanities* 1, no. 1 (2011).

Day 4 | Interactive analysis with Shiny

An introduction to the Shiny web framework for R, which allows users to create a website which connect arbitrary inputs to output from plots and htmlwidgets.

Morning

- Introduction to Shiny

Afternoon

- Lab: Hands-on with networks

Readings

RStudio, “Teach yourself Shiny” (2014).

Day 5 | Final

Morning

- Work on final projects

Afternoon

- Show and tell
- Discussion of projects

Preparing for this class

There are several things that you can do to prepare for this course.

First, come prepared by installing R and RStudio. You can find instructions for installing R for your operating system at CRAN, and for installing RStudio Desktop at RStudio’s website. While a version of RStudio will be available for participants via a server, it will be far better to have R installed on your own platform. Help will be provided on the first day.

Second, pick a dataset in your field that you would like to visit, or better yet, several. We will go over how to prepare the data for visualization, but you should acquire the data before coming to the course.

Third, try to gain a basic familiarity with R as explained in the suggested readings. We will provide an introduction in the class, but you will get much more out of the course if you work through the readings in advance.

Bibliography

Jim Albert and Maria Rizzo, *R by Example* (Springer, 2012)

An example-based introduction to R useful for an introduction into the range of statistical and probability computation.

Taylor Arnold and Lauren Tilton, *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text* (Springer, 2015)

A guide on the key concepts of R and its application to digital humanities, including network theory, geospatial analysis, textual analysis, and images.

Christopher Gandrud, *Reproducible Research with R & RStudio* (CRC Press, 2015)

An overview of tools for data gathering, preparation, and presentation, aimed at a reproducible research workflow. The book addresses several key R libraries, including rmarkdown, knitr, and RStudio.

Matthew L. Jockers, *Text Analysis with R for Students of Literature* (Springer, 2014)

Jockers guides users through the use of R for students and scholars of literature and text, or those interested in computational textual analysis. Approaches include word frequency, token distribution, clustering, classification, and topic modeling.

Daniel Kaplan, *Data Computing: An Introduction to Wrangling and Visualization with R* (Project Mosaic, 2015).

An excellent introduction for beginners interested in data wrangling and visualization with R, relying largely on the ever-useful Hadleyverse collection of packages.

Lincoln Mullen, *Digital History Methods in R*

An in-progress book focused on teaching historians to apply R to digital history. Topics include working with and manipulating different types of data as well as the application of various techniques including plotting, statistics, geospatial analysis, natural language processing, topic modeling, and networks.

Google, Intro to R

An introduction to R from Google Developers.

Hadley Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).

Wickham's introduction to the ggplot2 package for data visualization, a useful reference work for creating and designing graphics in R.

Hadley Wickham, *Advanced R* (Chapman & Hall, 2014).

Aimed at intermediate and advanced R users, Wickham's book is less a work of reference and more a guide about how the R language works and some the more advanced features of the language.

Use R! series

The book series by Springer focused on shorter books that focus on specific areas or statistical topics.