SigSci BioDataSci:
Signature Science Biological Data Science Workshop Series

*Stephen D. Turner**

Contents

1 About the Course

This course will introduce methods, tools, and software for reproducibly managing, manipulating, analyzing, and visualizing large-scale biological data. Specifically, the course introduces the R statistical computing environment and packages for manipulating and visualizing high-dimensional data, covers strategies for reproducible research, and culminates with analyses of real experimental high-dimensional biological data. The course is all hands-on and interactive, with live demonstration and problem-solving in place of lecture-style instruction.

This is an *applied data science* course. This is not a statistics class or a machine learning course. There is a session devoted to essential statistical analysis, but this single session offers neither a comprehensive background on underlying theory nor in-depth coverage of implementation strategies using R. There is also a session on predictive modeling using some machine learning strategies, and another session on natural language processing, but this is by no means a deep dive into machine learning, nor is it a comprehsive course in computational linguistics. Some general knowledge of statistics and study design is helpful, but isn't required.

This is not a *"Tool X"* or *"Software Y"* class. Participants should take away from this series the ability to use an extremely powerful scientific computing environment (R) to do many of the things that they will do *across study designs and disciplines* – managing, manipulating, visualizing, and analyzing large, sometimes high-dimensional data. This data might be microarray genotype data, gene expression data, microbial genomics data, public health/demographic data, or something not biologically related at all such as movie preference trends from Netflix, truck routing data from FedEx, or sentiment analyses in several great works of fiction. Regardless, the same computational know-how and data literacy is required to do the same kinds of basic tasks in each. This workshop series may feature specific tools here and there but these are not important – the same specific software or methods will likely not be used 10 years from now, but underlying data and computational foundation will be. *That* is the goal of this course – to arm participants with a basic foundation, and more importantly, to enable participants to figure out how to use *this tool* or *that tool* on their own, when they need to.

The core curriculum of this workshop series is *not* modular. The course series builds on itself. Later elective workshops in the series will only be useful if the participant has attended all prior core curriculum sessions.

*sturner@signaturescience.com

2 *SigSci BioDataSci* Proposed Schedule

Each workshop session is 3 hours long. As noted previously, the core curriculum of this workshop series is *not* modular. The course series builds on itself. Later "elective" workshops in the series will only be useful if the participant has attended all prior workshops, and attendance is required at *all* core curriculum sessions. Elective sessions *may* be missed if the participant reviews the course material for that session, but there is no out-of-class substitute for the in-class hands-on instruction in the core curriculum sessions. The course could be run as a series of 3-hour weekly sessions across 5-7 weeks, or as a 2-3 day intensive bootcamp.

2.1 Core Curriculum

2.1.1 Week 1: Intro to R

This novice-level introduction is directed toward scientists with little to no experience with statistical computing or bioinformatics. This interactive introduction will introduce the R statistical computing environment. The first part of this workshop will demonstrate very basic functionality in R, including functions, functions, vectors, creating variables, getting help, filtering, data frames, plotting, and reading/writing files.

Learning Objectives

- Become familiar with the RStudio interface and project management using RStudio
- Using R scripts to make analyses reproducible
- Perform basic arithmetic operations in R
- Using functions, creating variables, getting help
- Installing and loading R packages
- Importing and inspecting data

2.1.2 Week 2: Advanced Data Manipulation with R

Data analysis involves a large amount of janitor work – munging and cleaning data to facilitate downstream data analysis. This session assumes a basic familiarity with R and covers tools and techniques for advanced data manipulation. It will cover data cleaning and "tidy data," and will introduce R packages that enable data manipulation, analysis, and visualization using split-apply-combine strategies. Upon completing this lesson, participants will be able to use the *dplyr* package in R to effectively manipulate and conditionally compute summary statistics over subsets of a "big" dataset containing many observations. We will also cover use of the *tidyr* package to effectively transform data from wide to long formats, and back again.

Learning Objectives

- Employ the `filter` operation to return only rows of data meeting a condition
- Employ the `select` function to subset data including only columns of interest
- Employ the `mutate` function to modify existing data or create new data
- Employ the `arrange` function to sort data by columns of interest
- Use the `group_by` and `summarize` functions in combination to perform summary and statistical analyses over subgroupings of data
- Employ the 'pipe' operator, `%>%`, to link together a sequence of functions
- Reformat and reshape "messy" wide data to a tidy format using functions from the tidyr package

2.1.3 Week 3: Advanced Data Visualization with R and ggplot2

This session will cover fundamental concepts for creating effective data visualization and will introduce tools and techniques for visualizing large, high-dimensional data using R. We will review fundamental concepts for visually displaying quantitative information, such as using series of small multiples, avoiding "chart-junk," and maximizing the data-ink ratio. After briefly covering data visualization using base R graphics, we will introduce the *ggplot2* package for advanced high-dimensional visualization. We will cover the grammar of graphics (geoms, aesthetics, stats, and faceting), and using ggplot2 to create plots layer-by-layer. Upon completing this lesson, participants will be able to use R to explore a high-dimensional dataset by faceting and scaling arbitrarily complex plots in small multiples.

Learning Objectives

- Understand the grammar of graphics, and about building a plot layer by layer
- Map features of the data to aesthetics of a plot
- Rescale data for more effective visualization
- Create typical visualizations, such as scatter plots, histograms, density plots, boxplots, and their alternatives.
- Faceting plots to show visualizations in small multiples
- Creating publication-ready plots and using themes

2.1.4 Week 4: Reproducible Research & Automated Reporting with Dynamic Documentation

Scientific research is plagued by reproducibility issues. This session covers some of the barriers to reproducible research and how to start to address some of those problems during the data management and analysis phases of the research life cycle. In this session we will cover using R and dynamic document generation with RMarkdown and RStudio to weave together reporting text with executable R code to automatically generate reports in the form of PDF, Word, or HTML documents.

Learning Objectives

- Understand the benefits of using dynamic documentation for reproducible research
- Using markdown as a markup / formatting language
- Embedding R code in an RMarkdown document
- Compiling Rmarkdown to an HTML or PDF report

2.1.5 Week 5: Refresher session / BYOD

This is a refresher session where we will integrate all of the skills we have practiced in the course so far. I will also provide this as a "Bring Your Own Data" (BYOD) session, where we will work through specific problems live in class, allowing all participants to see how we can approach a problem using the principles and techniques already covered.

2.2 Electives

2.2.1 Week 6: Essential Statistics

This session will provide hands-on instruction and exercises covering basic statistical analysis in R. This will cover descriptive statistics, t-tests, linear models, chi-square, clustering, dimensionality reduction, and resampling strategies. We will also cover methods for "tidying" model results for downstream visualization and summarization.

Learning Objectives

- Using exploratory data analysis and descriptive statistics to get a "feel" for the data you are working with
- Implementing statistical tests for continuous outcomes in R: t-tests, ANOVA, simple linear regression, and multiple linear regression
- Implementing statistical tests for categorical outcomes in R: chi-square tests, fisher exact tests, logistic regression
- Perform power and sample size analysis using R
- "Tidying" the results of statistical analysis

2.2.2 Week 7: Predictive Modeling & Forecasting

This session will provide hands-on instruction for using machine learning algorithms to predict a disease outcome. We will cover data cleaning, feature extraction, imputation, and using a variety of models to try to predict disease outcome. We will use resampling strategies to assess the performance of predictive modeling procedures such as Random Forest, stochastic gradient boosting, elastic net regularized regression (LASSO), and k-nearest neighbors. We will also demonstrate demonstrate how to *forecast* future trends given historical infectious disease surveillance data using methodology that accounts for seasonality and nonlinearity.

Learning Objectives

- Using exploratory data analysis & reviewing data visualization techniques to get a "feel" for the data you are working with
- Feature extraction and variable re-coding for machine learning analysis
- Imputing missing data
- Using the caret package for automated model training and testing
- Understand how resampling techniques can be used to develop a predictive model
- Assess the performance of a variety of predictive models on a particular data set: random forest, support vector machines, k-nearest neighbor, and elastic net regularized > regression
- Introduce forecasting and time series analysis

2.2.3 Week 8: Text Mining

This session will provide an overview of fundamental principles in text mining, and introduces the *tidytext* package that allows one to apply to text data the same "tidy" methods that have been used throughout this course with for wrangling and vizualizing text data.

Learning Objectives / Topics Covered

- Tokenizing text
- Stop words
- Sentiment analysis
  - Trajectory of sentiment over time
  - Contribution of terms to sentiment in a corpus
- Word/document frequency: TF-IDF statistics.
- Topic modeling
  - Latent Dirichlet allocation
  - Clustering, document-topic, topic-term modeling

3 Pre-requisites / Computing requirements

Pre-requisites: **There are none!** Some knowledge of statistics might be useful but isn't strictly required.

Computing requirements: Each class involves lots of hands-on practice coding, in-class exercises, and short homework assignments. Participants will be required to have all necessary software installed and tested prior to the first class. All software is freely available and open source.

- **R**: https://www.r-project.org/
- **RStudio**: https://www.rstudio.com/
- **R Packages**. These are installed from within R. An incomplete list is given below. Students will need to be able to install packages as needed, on the fly, during class. This can *usually* be done without privilege escalation / administrative access.
  - dplyr (https://cran.r-project.org/package=dplyr)
  - readr (https://cran.r-project.org/package=readr)
  - tidyr (https://cran.r-project.org/package=tidyr)
  - ggplot2 (https://cran.r-project.org/package=ggplot2)
  - tinytex (https://cran.r-project.org/package=tinytex)
  - knitr (https://cran.r-project.org/package=knitr)
  - rmarkdown (https://cran.r-project.org/package=rmarkdown)
  - purrr (https://cran.r-project.org/package=purrr)
  - skimr (https://cran.r-project.org/package=skimr)
  - tidytext (https://cran.r-project.org/package=tidytext)
  - tm (https://cran.r-project.org/package=tm)
  - topicmodels (https://cran.r-project.org/package=topicmodels)
  - caret (https://cran.r-project.org/package=caret)
  - ModelMetrics (https://cran.r-project.org/package=ModelMetrics)
  - generics (https://cran.r-project.org/package=generics)
  - gower (https://cran.r-project.org/package=gower)
  - randomForest (https://cran.r-project.org/package=randomForest)
  - gbm (https://cran.r-project.org/package=gbm)
  - glmnet (https://cran.r-project.org/package=glmnet)
  - kknn (https://cran.r-project.org/package=kknn)
  - mice (https://cran.r-project.org/package=mice)
  - prophet (https://cran.r-project.org/package=prophet)

More extensive setup and installation instructions will be provided prior to the first workshop session.