Data Science for Biostatisticians

- How to use the command line
    - Basic Unix commands
    - Chaining commands
    - I/O
    - Piping and redirection
    - Using a text editor
    - Installing software via a package manager
    - Installing software from source
- How to obtain data
    - From text files (CSV, TSV)
    - From unstructured text
    - From JSON files
    - From XML files
    - From Excel files
    - From a web (REST) API
    - From scraping the web
    - From a relational database
    - From a NoSQL database
    - From a SPARQL database
- How to clean data
    - Cleaning text data with string functions
    - Cleaning text with regular expressions
    - Setting validation rules
    - Mapping to standard vocabularies
- How to work with DataFrames
    - Anatomy of a DataFrame
    - Indexing and extracting subsets
    - Sorting
    - Tall to Wide
    - Wide to Tall
    - The split-apply-combine pattern
    - Grouping
    - Function application
    - Concatenating DataFrames
    - Merging and joining DataFrames
    - Saving and exporting DataFrames

- How to explore data
    - Summary statistics
    - Constructing meaningful tables
    - Introduction to statistical visualization
    - Working with domain experts
- How to work with missing data and outliers
    - Identifying outliers
    - Types of missing data
    - Simple methods
    - Single imputation
    - Multiple imputation
- How to model data
    - Review of probability models
    - Review of data transformations
    - Review of linear models
    - Review of Generalized linear models
    - Multi-level models
    - Fitting models to data
    - Fitting nonlinear models to data
    - Interpreting model fits
    - Checking model assumptions
    - Using simulations
    - Combining information from multiple sources
- How to build a pipeline
    - Why build a pipeline
    - Chaining steps
    - Cross-validation and pipelines
- How to display results
    - Native plotting libraries
    - Graphics for the web
    - Graphics for high-dimensional data
    - Making interactive plots
    - Plotting geographical data
    - Plotting graphical and network data
- How to improve performance
    - Profiling and benchmarking code
    - Understanding algorithmic complexity
    - Choice of data structures and algorithms
    - Wrapping native code
    - Running jobs in parallel
- How to make analysis reproducible
    - Literate programming

    - Version control
    - Automating tasks
    - Using VMs and Docker containers for reproducible environments
    - Testing code and test generators
    - Code coverage
    - Use of `make` to build projects
- How to work with data too big for RAM
    - Buy more RAM
    - Change of algorithm/data structure
    - Lazy evaluation
    - Memory Mapping
    - Sub-sampling
    - Distributed computing
- Case studies (interspersed throughout course)
    - Working with EHR data
    - Working with image data
    - Working with genomics data
    - Working with survey data
    - Working with Quantified Self data