DataSci Cleaning Data Lecture Notes: Obtaining Data Motivation

About Course

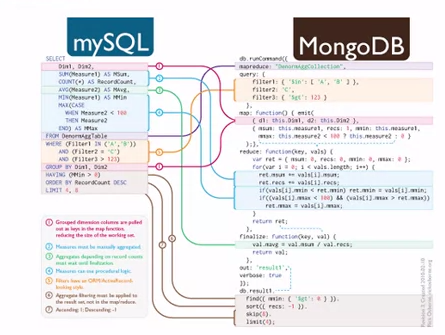
1. Covers Basic ideas behind getting data ready for analysis
   1. Finding and extracting raw data
   2. Tidy data principals and how to make data tiny
      1. Tiny = Easy to use
   3. Practical implementation through a range of R packages
2. Course Requires
   1. The Data Scientist’s Toolbox class
   2. R Programming class
3. Useful Knowledge Base
   1. Exploratory analysis (like stats classes)
      1. For plotting data while cleaning
   2. Reporting Data and Reproducible Research
      1. Creating scripts that can reproduce your analysis.

Initial Data Presentation: Desire vs Reality

1. Ideally data would be listed in a “matrix”-style excel sheet
2. In reality, it’s usually a horrible very bad mess in a Txt file
3. Or, it could be json format or source code from a website

Where is Data Contained

1. May need to extract it, from a database
2. Various databases can have different formatting, such as



1. Can use “get requests” from websites

Goal of this Course

1. Take Raw Data, Process it via a script, Tidy the data up, analyst the data, then communicate your findings