DataSci Cleaning Data Lecture Notes: Raw and Processed Data

Raw Vs Process Data

1. Raw Data
   1. Original Source of data
   2. Difficult to use for data analyses
   3. Data analysis includes processing the raw data
   4. Raw data may only need to be processed once
2. Processed data
   1. Data that is ready for analysis
   2. Processing can include merging, subsetting, transforming, etc
   3. There may be standards for processing
   4. All steps to process data should be recorded – very important part of data analysis

Data Processing Pipeline

1. Example 1 - Illumina hiSeq machine – sequences DNA
   1. Give machine fragments of DNA on a slide, may only have 500 letters of the sequence
   2. Chemicals are added to create multiple copies of that sequence on the slide – called sequencing by synthesis
   3. Each letter in synthetic sequencing is assigned a color
   4. This makes images on the slide. The machine then zooms in on the slide. Whichever color is dominant on the image is given to represent that DNA letter
   5. Then the machine gives this back to the researchers in a file formatted so that each fragment on the slides is assigned a batch of letters
   6. In this example, the raw data is each image that needs to be processed, or each fragment that needs to be synthesized, or each letter on the output file. Each step is an example of raw data. If any step along the line contains an error, the output file will be wrong. Each step must be recorded and reproducible to verify the processed data.