**Option 1: National Healthcare Performance Data Sets**

The [medicare.gov](https://data.medicare.gov/) sets provides multiple data sets about hospital performance, insurance providers, readmissions mortality rates, complications and infections etc. Identifying trends among these variables to predict a particular performance outcome such as readmission rate is valuable to the integrity of our healthcare system. Some advantages of these data sets are that they collect information from the same participating hospitals allowing me to more cleanly merge various data sets during analysis (i.e. I can merge along hospital name for instance). Also, it could be interesting to look at geographic information and do some visualizations of national regional trends. I am agnostic as far as hypotheses are concerned but will suss that out during exploration. **Pros**: multiple big data sets and relevant to my career interests in healthcare. **Cons**: Are these data sets already well curated? Also, do we do a similar mini-project down the line? I saw a comment in the Springboard forum asking about hospital readmissions rates. I’d rather not double dip.

**Option 2: Global Pharmaceutical Market Data Sets**

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) appears to be a good data source for healthcare information. I moonlight in tech transfer for my university in the life sciences and think that it would be interesting to evaluate the [2016 global data on the pharmaceutical market](http://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-health-statistics/pharmaceutical-market-edition-2016_d9d50c86-en) data set. I would explore R&D efforts/spending and drug trends by country. This is a single data set and includes nice data dictionaries. I’m not sure if a capstone project based on a single data set is unambitious; if so, I could also look epidemiological data by select countries to evaluate deeper trends or/and test hypotheses based on my initial data exploration. **Pros**: big data set and relevant to my professional trajectory. **Cons**: will this be too clean?

**FYI**: The [OECD](https://data.oecd.org/) is a great resource for this type of data from my brief review.

**Option 3: Brain Protein Expression in Mouse Models of Downs Syndrome**

This [data set via the UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression) evaluates cortical expression of protein levels in trisomy 21 mice under various learning paradigms. This data set was analyzed in a [PLosONE publication](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129126) for which the authors used a combination of supervised and unsupervised learning to look at protein expression patterns across genotype and condition. **Pros**: great opportunity to test machine learning. **Cons**: the data is published and I’m cautious my awareness of the results could overly scaffold my learning experience. I know we are not discouraged from using data that has been evaluated; however, I don’t want to completely remove the discovery aspect of my project.