**WEEK 3 - 2015-11-04**

**1 - Types of Data Science Questions**

**Order of difficulty**

* Descriptive
  + **Goal: Describe a set of Data**
  + The first kind of data analysis performed
  + Commonly applied to census data
  + Description and interpretation are different steps
  + Descriptions can usually not be generalized without additional statistical modeling. Describe what you see but not what it means for next person.
  + Example <http://www.census.gov/2010census/> (just decribing analys)
  + Example <https://books.google.com/ngrams> (One-way-info)
* Exploratory
  + **Goal: Find relationships you didnt know about**
  + Good for discovering new connections
  + Good for defining future studies
  + Exploratory Analysis are usually not the final say
  + Exploratory Analysis ALONE should NOT be used for predicting.
  + Correlation does not imply causasion. (emphasize that a [correlation](https://en.wikipedia.org/wiki/Correlation_and_dependence) between two variables does not necessarily imply that one [causes](https://en.wikipedia.org/wiki/Causality) the other)
* Inferential
  + **Goal: Use a relatively small sample of data to say something about a bigger population. Generalized on larger population**
  + Inference is commonly the goal of statistical models
  + Inference involves estimating both the quantity you care about and your uncertainty about your estimate
  + Inference depends heavily on both the population and the sampling scheme
* **Predictive**
  + **Goal: To use data on some Objects to predict values for another object**
  + **If X predicts Y - It does not mean X causes Y**
  + Accurate prediction depends on measuring the right variables
  + Although there are better and worse prediction models, more data and a simple model works really well
  + Prediction is really hard, specially on future referenses
* **Causal**
  + **Goal: To find out what happens to one variable when you make another variable change**
  + Usually randomized studies are required to identify causation
  + Casual relationships are usually identified as average effects, but may not apply to every individual. Example. If you give population this drug, the average people will live longer.
  + Casual models are usually the “gold snadard” for data analysis. **Diagnostic test or benchmark that is the best available under reasonable conditions.**
* Mechanistic
  + **Goal: Understand the exact changes in variables that lead to changes in other variables for individual object**
  + Hard to infer, except in simple situations