**Big Data Final Project Ideas**

1. Trump Tweets
   1. Data download link: <https://www.kaggle.com/kingburrito666/better-donald-trump-tweets>
   2. Data manipulation
      1. Divide text to words
      2. Do the manipulation in excel
   3. Models
      1. Y1 variable: is it Android or Iphone
      2. Y2 variable: likes or retweets
      3. X variables: words, time, photo, hashtag, iphone/android, media type
   4. Key questions
      1. Can we predict who wrote the tweet based on the words?
      2. Can we predict number of likes/retweets based on the words?
   5. Topics from class
      1. FDR
      2. LASSO
      3. Bootstrapping
      4. Causal inference (maybe..)
      5. Clustering (cluster of words indicative of topic?)
   6. Issues
      1. The research question won’t be deep
      2. Too similar to midterm?
2. Divvy Data
   1. Data manipulation
      1. Divide time into dummy (hours or by grouping them meaningfully)
      2. Do the manipulation in R
   2. Models
      1. Y1 variable: count
      2. Y2 variable: casual vs. registered
      3. X variables: season, windspeed, humidity, time (24 hr dummy), temp, working day, holiday
   3. Key questions
      1. Can we predict the number of bike usage based on time/or other variables?
   4. Topics from class
      1. FDR
      2. LASSO (maybe not…)
      3. Bootstrapping
      4. Causal inference
      5. Clustering (cluster of time periods)
   5. Issues
      1. The research question won’t be deep
      2. X variables still too little? (might need to check with Prof)