1. USA Today keeps a record of all NFL arrests/crimes in this web page: <http://www.usatoday.com/sports/nfl/arrests/>   
   Are these arrests seasonally correlated? Are they correlated whether the teams wins or loses? Are the different types of crimes related?
2. Twitter Projects:
   1. Sentiment analysis comparison between different search terms.
   2. Finding related words.
   3. Topic modeling on tweets.
   4. Comparing tweet subject frequencies.
3. Spam/Fraud modeling of emails (<https://archive.ics.uci.edu/ml/datasets/Spambase> ), texts (<https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection> ), or other text sources.
4. Kaggle’s public data sets: <https://www.kaggle.com> .
5. Also think about taking any of the above data sets and merging/correlating them with:
   1. Census Data: [http://www.census.gov/data/data-tools.html#](http://www.census.gov/data/data-tools.html)
   2. Weather Data: See weather\_retrieval.R
6. Many major websites have API’s that you can access for free.
   1. Twitter
   2. Linkedin
   3. Foursquare
   4. Open Street Map
7. Amazon released an anonymized purchase dataset here: <http://snap.stanford.edu/data/amazon-meta.html> , which is part of a larger collection of network type data here: <http://snap.stanford.edu/data/index.html>
8. Game Databases:
   1. Chess game database: <http://www.ficgs.com/directory_databases.html>
   2. Poker hand database: <https://archive.ics.uci.edu/ml/datasets/Poker+Hand>
9. Wikipedia traffic statistics:
   1. <https://aws.amazon.com/datasets/6025882142118545>
10. Yelp Dataset Challenge: <https://www.yelp.com/dataset/challenge>
11. UCI Machine Learning Database <https://archive.ics.uci.edu/ml/datasets.html>

# Previously tackled problems:

* Is there a correlation between airline delays and weather? Can we predict the delay time?
* Are restaurant inspections done at random? Or do they depend on popularity, food type, lat/long…?
* Are profane tweet locations and frequency related to crime locations?
* Does the home football team really have a 3pt advantage on average? Do the Vegas betting lines take this into account?
* Are any major stocks in the NYSE highly correlated with each other? With a one, two, or three day lag?
* Is global warming responsible for the higher frequency of Atlantic Ocean hurricanes?
* Is the NIKE SPARQ (measurement of football player’s athleticism) related to season performance or pick order?
* Can we predict if an online news article will be popular? Does this depend on subject, time of post, number of pictures/videos, etc…?
* Is it possible to predict web server response times based on site content and the url?