**Final Project, Part 1: *Proposal***

**Problem Statement**

1. Hypothesis/assumptions:
   1. Before most NFL football games, there is a team that is favored to win by a specific number of points given by the “spread”.
      1. There are a small number of games considered perfectly even where the spread is equal to zero.
   2. Theoretically, there are even odds of the favored team “beating” the spread (i.e. winning by more than the spread).
   3. Whether the favored team will “beat” the spread in a Super Bowl can be predicted from past data from that season.
      1. The relevant data comes from the games in which either of the two teams in the Super Bowl participated.
      2. The prediction may depend on:
         1. The size of the spread
         2. The amount of points by which the favored team wins (in games that they won) on average during the season
         3. The amount of points by which the unfavored team loses (in games that they lost) on average during the season
         4. Consecutive games won by the favored team leading up to the Super Bowl (i.e. momentum)
         5. Consecutive games won by the unfavored team leading up to the Super Bowl (i.e. momentum)
         6. Average amount of points by which the favored team has won in past matchups between the two teams over the last 5 seasons
         7. Other features to be defined
2. Goals and success metrics
   1. Goal: Determine which of the features should be included in the model and which type of model to use to get the most accurate predictions.
   2. Success Metric: Use data from 2000 to 2016 to train a model, and correctly predict whether the Patriots beat the spread in the 2017 Super Bowl.
      1. We already know the outcome but the model won’t be trained on that data.
3. Risks or limitations
   1. Presumably both teams in the Super Bowl did well during the season so there may not be many losses during the season (i.e. small sample size) to assess the unfavored team’s average loss margin.
   2. There may not be many recent matchups between the two teams (i.e. small sample size) to assess the average amount of points by which the favored team has won. This will also be the only input that takes into account data outside of the current season, which may be less relevant.
   3. Injuries are common in the NFL. The performance of a team in the Super Bowl may differ greatly from prior games that season if key players are injured or return from injury. That data is not given in the relevant data set and if available, may be complex to incorporate.
   4. League rules change over time, so it’s possible that only data from the “modern era” (e.g. 2000 and later) is useful in predicting Super Bowl outcomes today.

**Relevant Data Set**

1. [Link to data](https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data/version/1)
2. Description: NFL football games since the 1966 season with game results and descriptive info including if a playoff game, played at a neutral site, and weather information if available. Data set was built from publicly available NFL data, weather provided by the NOAA, and betting data from a variety of sources but cross-referenced with Pro Football Reference.
3. Data fields:
   1. Date of game
   2. NFL season
   3. Week of season
   4. Home team name
   5. Away team name
   6. Stadium name
   7. ID of favored team
   8. Spread in points for favorite
   9. Over under points
   10. Weather detail
   11. Temperature in degrees F
   12. Wind speed in mph
   13. Humidity
   14. Home team points scored
   15. Away team points scored
   16. If game played at neutral location
   17. If game is a playoff game