## NFL Plays Analytics

Predicting offensive football plays

# Data Wrangling

My ultimate goal in this project is to build a multiclass model that predicts opponent’s next offensive play. The main open-source dataset can be found here, nfldb; it is a Postgresql relational database containing 8 tables, covering play-by-play, game scores, player stats, schedules, and much more from the 2009-2016 seasons and playoffs.

I explored the various tables with SQL queries to better understand the scope of the data available, and determined that I would begin wrangling data from 3 of the tables - plays, games, and drives table which shared game and drive keys useful for merging later. For initial programming and processing ease, I focused on the 2016 Regular season data for every team. To obtain labels for play type and direction, I searched each plays description for pass vs rush and left vs middle vs right, using string manipulation via regular expressions. This wrangling was used to produce my target variable. In the process, several plays and play types with insufficient data or outliers were excluded, including kicks, punts, two-point conversions, and quarterback scrambles.

Score tuples, game clock quarter and time, and variety of other categorical variables with string representations were split converted to integers, category codes, and/or time floats as necessary. I shifted statistics forward for each play to prevent leaking future information into the model for a given play. Then, I accumulated these stats separately across drive possessions (collections of plays) and games for each team and game. I zero-filled the resulting missing data for the beginning of drives and forward-filled stats for accumulated game stats.

Following processing plays, I merged this dataframe with general game and drive data that was queried from the database. These merges provided additional data about home vs away teams, day of the week, week of the season, and summary stats from previous play drives. With the current wrangling setup, ~32,000 plays and ~200 features emerged which extrapolates to ~320,000 usable plays across the entire database. The features are likely to grow too, as I incorporate player-specific stats and other potentially important indicator variables.