The problem I intend to solve is to predict the outcome of a given soccer match in the top 5 European Leagues: England, Spain, Germany, Italy, and France. I intend to use a data set containing all the games played in these leagues from the 2008-2016 seasons. This dataset includes the following: Players and Teams' attributes sourced from EA Sports' FIFA video game series, team line-ups with squad formations, betting odds from over 10 providers, detailed match events (goal types, possession, corner, cross, fouls, cards etc). The first challenge is determining which variables most directly affect the result.

I will mostly look at the following variables: the attacking strength of the home team (tendency to score goals), the defensive weakness of the away team (tendency to concede goals) and the home advantage effect. These three parameters determine the goal-scoring rate of the home team. The parameters that determine the goal-scoring rate of the away team are: the attack strength of the away team and, the defensive weakness of the home team.

Notice that the difference between these two quantities: attacking strength of the home team, and attacking strength of the visiting team, will determine the winner. The distribution this difference has the same form as the one for the difference of two independent Poisson variates. I will use this model to predict the winner of a matchup between a home team and an away team. Apart from the finding the winner of a game, my goal is to display the expected winners for all the matches in a weekend in the leagues mentioned above and display them to the user using Tkinter. The program should also display betting odds for a home win, draw, and away win and suggest which bet to make. Additionally it should present a preliminary score based on the above metrics.

This problem consists of breaking down game data into manageable chunks, filtering out the relevant variables and analyzing it. I will use pandas to interact and manipulate the databases. This will allow me to represent the data as a two-dimensional tabular, data structure with both row and column labels. To work with the data I will use numpy for operations on ndarrays and scipy for generating the poisson distribution. Finally, I will display the results on Tkinter.

Update 1:
No major design modification. Displaying results on Tkinter is still the plan
Update 2:
Results are successfully displayed on Tkinter, as well as an additional Head-to-Head feature that was unforeseen before TP3 (Thanks Alex). I also included an interesting graphic for each league that gives the offensive characteristics of the league as a whole by looking at the distribution of goals scored by any team from that league