DATA SOURCES

---

1. Fox Sports World Cup Records:

   https://www.foxsports.com/soccer/fifa-world-cup-men/stats

2. Wikipedia Stats on FIFA:

   https://en.wikipedia.org/wiki/FIFA_World_Cup_records_and_statistics

PROPOSAL (Due 4/12 @ 11:59 PM)

---

1. **What is the topic of your project? What question would you like to answer?**
   a. Topic:

      ELO score and Final Winner Prediction of FIFA World Cup

   b. Question:

      i. Step 1: Starting from the quarter-final, predict the winning/ losing probability of each possible rivalry pair. Construct a 4*4 matrix with $P_{i,j}$ being the probability of Team i winning against Team j.

      ii. Step 2: Given the draws for that particular year, get the probability of each team being the champion using: theoretical calculations and Monte Carlo simulation.

      iii. Compare with empirical results:

      How does the regression-predicted $P_{i,j}$ compare with historical results?

How does the final championship prediction compare with historical results?

2. **What methods or ideas will you use from the second part of the class?**
    a. Logistic Regression
    b. Tournament simulation (Monte-Carlo Simulation)
3. **What datasets are you going to use?**
    a. The possible data sources are listed above. I think it should be in a reasonable workload to go through the all the historical data, and extract team statistics (number of goals/ number of interceptions, etc.) from the website.
    b. There may also be some tidier dataset on Kaggle, which I could search for and inspect later.
4. **Do you have questions about how to start that we can help with?**
    a. Double check on model training and evaluation procedure:

    Split the collected data into two parts: earlier years for training, and more recent years for evaluation.

        i. Training: using rivalry team statistics from that *particular* year.
        ii. Evaluation: For Year *X*, use rivalry team statistics from the start to Yeah *X-1*. In other words, during evaluation we will use averaged "historical data" (relative to the year we are evaluating on) as input features.
        iii. Rationale: The goal of the evaluation phase is to see our model's ability to predict. In reality, when predicting an upcoming match result we can only rely on previous records. However in training, in order to better capture feature dependence and importance to the final outcome, it would be most

beneficial for the model to see up-to-date and aligned input statistics and output probabilities/ results.

b. It would be great if I could get some hint/ guidance on finding more handy datasets!