# Assignment #1: Sports Data

A Survey of Sports Data in Soccer

## Box Score Data

**High:** The Box Score in nearly all professional leagues is readily available on the internet, most notably through the statistics tab for more recent games on Google's knowledge panels whenever a game is searched. However, more comprehensive box scores dating back to 1993 can be found on `http://www.fbref.com` (although there are game scores dating back to 1889). Currently there is box score data for every game in 145 competitions over 47 countries, with box scores being generated the day after the game is completed. A comprehensive of all leagues can be found here.

For example, fbref contains box score data from a match from January 16th, 1993 all the way to a match from January 9th, 2024(yesterday as of day of writing). Unfortunately, it looks like there is no way to cleanly get this data into a clean csv form, meaning that an R package(worldfootballR) or web scraping would be needed to turn the data into a clean dataframe.

## Event Data

**Medium:** A small subset of datasets of event data are available to the public. For example, on kaggle, there is a dataset containing event data from the top 5 leagues from the 2011-2012 season to the 2016-2017 season. However, more recent and up to date event data(i.e. event data from present games) are not available to the public for free and must be purchased through a data company.

However, data companies like statsbomb release a small subset of event data which can be accessed through their R and python packages(statsbombr and statsbombpy, respectively). These packages contain event data for a small subset of competitions(like the Bundesliga), but one must pay for API access to get all of statsbomb's leagues.

## Tracking Data

**Medium:** A small subset of tracking data is available to the public through free data releases from soccer data companies. For example, Statsbomb 360 Tracking Data for the Women's Euro 2022 was made publicly avaliable through Statsbomb's R and Python packages. There also exists tracking data for games between college-aged athletes at a Japanese university, although the season and league are unknown.

However, like the event data, more recent and up to date tracking data, as well as data for other leagues, are not available to the public for free. For example, to get Statsbomb 360 tracking data through either of the Python or R packages, one must purchase a data subscription.