Assignment #1: Sports Data

Your task is to go out and survey the landscape of data in the sport of your choice.

Why are you being asked to do this?

We will be needing publicly available sports data throughout this semester for assignments and the project. We will share our finadings with each other so that we are all aware of data avaialble across many sports.

What (exactly) are you being asked to do?

Choose a sport. If you are undecided between two sports, choose the one that you think is less popular. Scour the internet to find publicly available datasets for your chosen sport, and summarize the state of data as high, medium, low or none, in each of three categories: box score data, event data and tracking data.

None: Not even professional teams have access to these data.
Low: Teams have access, but these data aren't available publicly for any meaningful sample of games.
Medium: There is at least one meaningful (e.g. a historical season of data) dataset publicly available.
High: Present-day dadta are publicly available as they are generated.

For the medium and high categories (if you have any), provide a short description of the publicly available data (year(s), league(s), sample size) and a URL link to the data.

Submission Requirements

- A one-page PDF summarizing your findings (see sample below)

Reminder

- Please anonymize your submission by removing any personally identifiable information.

How will your grade be determined?

This assignment will be graded for completion, meaning that as long as you complete the assignment, your grade will be 100%.

SMGT 430 Assignment #1

Sport

Volleyball

Box Score Data

Availability: High. Box scores and play-by-play are available for NCAA Division I women's volleyball, directly from the NCAA (example: https://www.ncaa.com/game/6191666). Unfortunately, you would have to scrape the data from the website because there is no way to download the data in a tidy format. Because the play-by-play data only describes the outcome of each point, this does not quite qualify as event data.

Event Data

Availability: Low. NCAA teams have access to high quality touch-by-touch data through a company called Volleymetrics. For each touch (pass, set, attack, etc.), we know the player, location, time and quality of the touch. Unfortunately, none of these data are publicly available. There is a website (volleydork.com) where the author Chad Gordon has published some analytics based on these data, but it is not clear how he got access to the data.

Tracking Data

Availability: None. Many people have started experimenting with computer vision for player and ball tracking in volleyball as side hobbies, but for now, not even teams have acccess to these data. There have been some academic papers published on the topic.

Note: This is just a sample meant to demonstrate the format of the assignment. I hope you will go into slightly more detail than this in your submission. The expected time spend on this assignment is 2 hours.