CMSC 12200 Final Project Proposal
Team 3+1

Members:
- Joey Cipriano (ciprianoj)
- Peter Tang (peterttang)
- Xingyu Wang (xingyuwang)
- Kevin Yan (kyan1)

The Effect of Nike Vaporfly Shoes on Running Performance

Our project aims to do an analysis on running data before and after the release of Nike's now controversial Vaporfly 4% sneakers. Marketed as increasing running economy by 4% for the wearer, there is now ongoing debate over whether these shoes should be allowed in the professional running sphere, largely in response to the number of records being broken since Vaporfly's release in 2017.

Some studies of the Vaporfly have already been done, and we want to contribute our own analysis, inspired by this New York Times study from 2018. The article uses data from an app called Strava, where runners can share information on their workouts and runs, including the time and distance they ran, and most importantly *what shoes they were wearing*.

Since Strava's API offers rather limited functionality, we will scrape Strava's site for data from major marathons (e.g. Boston, New York, Chicago) and link that data with information scraped from official marathon databases. Most large marathons have very well-kept records and often contain additional information not always available on Strava (such as age, sex, hometown).

Finally, we plan to create a website which displays our findings on how Vaporfly shoes affected different runners. In particular, the user would be able to see general descriptives on the effects, and could give certain inputs, such as their age or current marathon pace. We will then offer a prediction of how much the Vaporfly would improve their running, based on runners similar to them in our data.

Possible difficulties include data cleaning, as Strava data is self-reported and may contain various misspellings or inconsistencies. A difficult question is how to measure increases in running ability that are *caused* by the Vaporfly. There may be selection bias in who chooses to buy Vaporfly shoes, and nonlinearities in how runners improve year to year, which we will have to take into consideration.

Data Sources:
- Strava, via HTML scraping
- Boston, New York, Chicago, and other marathon databases, via HTML scraping

Timeline:
1. **Clean Strava data.** Complete by: **Feb 14**, Week 6.
   - Scrape Strava data on marathons before/after 2017, filtering for runners with recorded shoe data and finding all those with Vaporfly shoes.
2. **Link with marathon data**. Complete by: **Feb 28**, Week 8.
   - Scrape marathon data from corresponding websites. Do record linkage with the Strava database using name and marathon finish time as identifiers.
3. **Analysis and website creation**. Complete by: **March 13**, Week 10.
   - Slice the data in appropriate ways to extrapolate causal effect of Vaporfly shoes. Configure website to take user input and display predicted results as a function of that.