**Quantifying Uncertainty in Live Marathon Finish Time Predictions**

The goal of this project is to develop a model that can predict a marathon runner’s finish time using their in-race splits. Finish times can be estimated mid-race using naïve extrapolation, using the average pace so far in the race. There are many issues from making predictions this way: (1) it does not consider any context occurring within the race that can determine if a runner is likely to finish much slower or faster than expected, and (2) the prediction is just a single point estimate that has no information about how certain that estimate is. I propose using a Bayesian model because it solves each of the 2 previous issues. Bayesian models use condition probability to account for in-race context important in predicting the outcome, and Bayesian models output a probability distribution that allows us to quantify how certain we are of the predicted finish time.

Data:

The data from this project was scraped from the Boston Athletic Association website.

Notes:

**POSTERIOR ∝ LIKELIHOOD \* PRIOR**

* P(finish | 5K) ∝ P(5K | finish) \* P(finish)
* P(finish | 5K, 10K) ∝ P(10K | finish, 5k) \* P(finish|5K)
  + Assumption: P(10K | finish, 5K) simplifies to P(10K|finish) - BDA pg. 11
* P(finish | 5K, 10K, 15K) ∝ P(15K | finish, 5K, 10K) \* P(finish | 5K, 10K)
  + Assumption: P(10K | finish, 5K, 10K) simplifies to P(15K|finish) - BDA pg. 11
* …

Initial prior distribution: density of all finish times in dataset (probably should filter using information about runner like age, sex, experience, etc.)