# Introduction

In this Data Era, there are potentials in analyzing data and bringing insights to a problem we have interests in. For instance, to find out whether there is a relationship between physical performance and aging and what the relationship is, road races data can be the data we could do analysis on. This kind of data is collected by the race organizers and it is at individual level. Data is often published on the Web. We can bring some insights to the question we have interests in.

Cherry Blossom Ten Mile Run is one of the annual races held in Washington D.C. early April which the cherry blossom is blooming around this time. This race run started in 1973 as a training run for runners planning to compete in Boston Marathon. Its participants range from age 9 to 89. The organizer publishes the results at http://www.cherryblossom.org/. The data published provides a very helpful resource including name, sex, hometown and final time of participants for us to learn the relationship between physical performance and aging.

The original analysis was provided by the book of Daniel Kaplan and Deborah Nolan in the Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. The data has been already web-scraped and it has results from all years from 1999 to 2012. The data is divided by genders meaning MenTxt for men and WomenTxt for women.

To answer the above questions, we are using the data for each gender and a Loess model to predict the smoothed run times after two rounds of normalization. Then we have done the distribution comparison for men runners and also for women runners in the year of 1999 and in the year of 2012.

Resources: Daniel Kaplan and Deborah Nolan in the Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving

## Results

1. For men data only, I think we need a t test to compare them

density plot: In this visualization, we are comparing the density distribution from the year of 1999 and the year of 2012. We find that they have different distributions. 2012 density curve has a more slim shape on the top and it is slightly right skewed while 1999 curve has a broader shape and almost normally distribute. The density distribution shows evidence of age population shifting when comparing men runners from these two years. When you look at ages where the peaks of two curves appear, you can see that men runners are actually younger of 2012 that the peak appears around 30 years old while the peak appears around 40 of 1999. (can use a test to verify this) figure 11

qq-plot: now let us look into qq-plot where we can compare two sets of quantiles against each other. we can see these two sets of quantiles are roughly forming a straight line at the beginning then starts to curve off between age of 30 and age of 58. In this age range between 30 to 58, you can see that 2012 has lower ages. After age of 60, it forms to a straight line again then finishes with a small portion of curve off the extremity. figure 12

Loess fit line: As you can see the two Loess fit curves almost have the same shape but again 2012 curve is higher than 2012. This is consistent with what we have found in the density plot. This visualization helps with telling the differences of the predicted run time between these two years. At the very beginning, the run time difference is quite big where the biggest difference appears around people in their 40s. Then it starts to narrow down at age of 50 where the smallest difference appears between people who are at their 50s and 60s. After age of 60, it gets wider but still smaller than the difference before age of 40. Figure 13 is the fit curves and Figure 14 is the difference plot.

statistics description.( I think we need a statistic test.)