**Formula 1 Results Analysis & Prediction**

**Team GottaGoFast:**   
Zachary Lawell, Lezeh Foy, Deepak Verma

MSCS 5610 – 701 Data Mining (Spring 2019)

Marquette University

# Background and Project Description

Formula One (also Formula 1 or F1) is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA). The FIA F1 World Championship has been a premier event around the world since its inaugural season in 1950. The word "formula" in the name refers to the set of very strict rules and regulations to which all participants' cars must conform.

Tucked inside each car is around 1.25 kilometers of wiring and more than 200 sensors, some of which give readings up to 1,000 times per second. Sensors around the track and highly precise race statistics from each event have also been gathered since inception. This makes F1 is a treasure trove of big data. Data is a gathered and analyzed in real-time using high performance compute clusters. The results of which are utilized by the teams to provide real-time advice to drivers during a race to gain an edge. The teams competing in F-1 are at the forefront of technological advancement and therefore hige investment can drive developments which make a difference on track. Average mid-field teams spend around $125m to $150m per annum, whereas top teams spend $350m to $410m per annuum[3]. The time difference between these teams are less 0.06 seconds per lap. At the end of each season, the top 10 teams share a winning of approximately $810m (as of 2018)[4].

Gathering data and quickly analyzing it to predict outcomes via the application of machine learning techniques can not only help the teams gain an edge where a single position advancement results in millions of dollars of winnings. Analysis and predictions of data from prior data could allows race organizers, media, and fans to quickly narrow down the likely outcomes of races. However too many factors affect the outcomes just as with any sport.

# Dataset

For this project we are exploring the dataset from Formula 1 racing available from http://ergast.com/mrd/. The Ergast Developer API is an experimental web service which provides a historical record of motor racing data for non-commercial purposes. The API provides data for the Formula One series, from the beginning of the world championships in 1950. This dataset contains data from 1950 all the way through the 2018 season, and consists of tables describing race results, constructors results, constructors, race drivers, lap times, pit stops, qualification results, and many more as per the scheme provided at <http://ergast.com/schemas/f1db_schema.txt> .

# Project Goals

In this project we’ll utilize the techniques studied in the Data Mining class against the dataset chosen. This will include pre-processing the data, performing initial visualization and observatory analysis. We then intend to apply some unsupervised and supervised techniques to model the data and predict outcomes. Some of the initial goals include:

* Determining the fewest number of the most important attributes for modeling the data
* Utilizing unsupervised machine learning to discover any interesting patterns
* Utilizing supervised machine learning to predict the likelihood of a race

These goals will evolve as we become more familiar with the dataset and perform the initial exploratory analysis. From initial visual analysis of the database scheme the objectives include:

* Normalizing the data by combining tables from the relational database schema
* Charting some statistical analysis such as the team and drivers with most historic wins, countries with the most wins
* Determing if speeds have gone up or down over the years
* Determining if reliability has increased or reduced
* Finding correlations between data elements such as drivers and track, qualification position and race win by track
* Applying clustering techniques against various data elements
* Finding correlations between data attributes
* Selecting features that can predict the winning race outcome and its probability

# Tools

The tools chosen to perform the data pre-processing, visualization, analysis and modeling is python and its various data analysis libraries using Jupyter notebook as the interface for visualization. The github webpage for the project is at <https://github.com/dvermagithub/GottaGoFast>

# Literature Review

X research are cited below.

Sensorless but not Senseless: Prediction in Evolutionary Car Racing   
<https://ieeexplore.ieee.org/abstract/document/4218909>

A machine learning framework for sport result prediction   
<https://doi.org/10.1016/j.aci.2017.09.005>

Hierarchical Models for Permutations, Analysis of Auto Racing Results, Todd Graves, C. Shane Reese & Mark Fitzgerald <https://amstat.tandfonline.com/doi/abs/10.1198/016214503000053#.XKtTOphKiUk>

Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing Theja Tulabandhula and Cynthia Rudin  
<https://www.liebertpub.com/doi/10.1089/big.2014.0018>

Predicting Return on Investment In Sport Sponsorship - Modeling Brand Exposure, Price, and ROI In Formula One Automotive Competition <http://www.journalofadvertisingresearch.com/content/54/4/435.full.pdf+html>

# References

[3] <https://www.racefans.net/2018/12/26/the-cost-of-f1-revealed-how-much-teams-spent-in-2018-part-two/>

[4] <https://www.racefans.net/2018/08/01/formula-1-teams-prize-money-payments-2018-revealed/>