**Problem:**

A foundational problem for Redback Operations is the ability to provide real-time feedback to its users on several key performance indicators. Using IoT sensor data as well as information from in-game processes, our goal should be able to convert that data into performance indicators – and performance improvements for our users. A sophisticated data-driven response provides the opportunity to develop tools such as AI for enhanced practice, evaluations of performance, coaching, and personalized exercise regimes for our users. This would lead hopefully to real-time feedback on wearable devices, quantification of data when synched with personalized devices such as smart-watch, smartphones, etc., and social engagement through online profiles and fitness gamification.

**So, the main question we are working with here – as a starting point – is whether:**

Given a future race event and a database of historical races of a rider, can we predict the loss or gain of time for this rider, relative to his competitors? (This is based on analysis by Leeuw et al 2020 – can we turn this into something fruitful for our company?)

This question uses *Time-Series Regression*, “in time series regression we are primarily interested in learning how to aggregate a time series into a single number (in our case win or loss of time) (Leeuw et al 2020).

In the particular training instance used to test our time-series regression algorithm, we will be taking information from procyclingstats.com to test a time-series regression algorithm.

**Problems involved**

1. **Web Scraping**

Procyclingstats.com has no easy way to access data into a plain-text file that lists the data. As such, I will need to build a web-scraper using BeautifulSoup or Selenium to get the required data. My focus for training data will be Cadel Evans’ Tour de France runs from 2005-to 2013).

1. **Data Cleaning and Pre-processing**

I’m not sure at this stage how complete the data is – or what type of pre-processing will be required. For instance, The Tour de France has two types of race start – Mass starts and time trials. Mass Starts are defined are race starts where every single rider begins at the same time. Time Trials occur when each cyclist rides alone, or in some cases only with teammates. Following Leeuw et al 2020 I will focus only on those stages in the Tour de France which began with mass starts.

1. **What is the Target variable?**

Given the complexities of professional cycling competitions, the goal often is not to be first – but rather to limit time losses compared to other competitors. A successful performance might not be one in which the rider achieves high standing in the race. A potential fix for this is to compare the finish time of our chosen rider with the average finish time of the ten best riders in the general classification.

1. **Feature Engineering**

Cycling features can be complex – especially when describing the terrain of the stage being cycled. How do we accommodate this? What type of engineering is needed to make this successful (Data Engineering team might have some ideas)

**Potential Timeline (Possibly the end of Trimester)**

1. Build a web-scraper that can scrape from Procyclingstats.com for Cadel Evans’ Tour de France (2005-2020). Though it may be easier to scrape all riders in those races and clean data to feature Cadel Evans later.
2. Data Cleaning and Pre-processing
3. Develop, Test and Evaluation a Time-series regression algorithm
4. IF this works, how can it be transferred to situations not involving professional athletes. What type of features should we be focusing on?
5. How do we get this algorithm to work in relation to IoT and gaming?
6. Can the time-series regression be improved upon using Deep Learning time-series algorithms?

**Bibliography and Resources**

Leeuw et al (2020). *Time Series Regression in Professional Road Cycling*. In Appice, A., Tsoumakas, G., Manolopoulos, Y., Matwin, S (Eds.) *Discovery Science 23rd International Conference DS 2020 Thessaloniki, Greece, October 19-21, 2020, Proceedings*

Procyclingstats.com

Lazzeri, F (2021) Machine Learning for Time-Series Forecasting with Python

Nielsen, A (2020) Practical Time Series Analysis: Prediction with Statistics & Machine Learning