**Capstone Project Proposal**
**Harrison Carter**
**Joseph Mata**
**NYC-DS-051622**

# Henry’s Fork Sonde Data Analysis

## Business Understanding

My family has been fishing at the Henry’s Fork in Idaho for three generations, and stewardship of the river is a cause close to my heart. The local Henry’s Fork Foundation is a nonprofit whose goal is “conserve, protect, and restore the unique fisheries, wildlife, and aesthetic qualities of the Henry's Fork and its watershed.” I would like to help them determine the factors that go into the turbidity at a site they have identified as being of particular interest. Turbidity is a measure of the suspended solids in a water sample, which in turn affects the availability of light for macrophytes growing in the river, the architecture of the riverbed, which consequently affect fish habitat and food availability. In this way, not only the low visibility affects sport fishing, but the availability of healthy fish itself. This is of interest to the HFF board for numerous reasons, not the least of which is the healthy balance of tourism and industry, where industry is represented by the dam at Island Park Reservoir.

## Data Understanding

The data I am collecting is available on the Henry’s Fork Foundation website. In addition, I have received climate and hydrology data from the Senior Scientist at HFF that is pertinent to the turbidity at the site in question. This includes the temperature, dissolved oxygen, conductivity, phytoplankton, and cyanobacteria content, as well as our target, turbidity, ranging back to when the data collection devices were first installed. There is also reservoir data containing inflow/outflow rates, as well as the volume and elevation of the water in it, which is likely a set of determining factors on its own. Finally, I am also in possession of the snowpack and water runoff data for the local tributary rivers, which may or may not play a role on their own.

## Data Preparation

A good portion of the data is in time series format, and sampled in 15 minute increments, which will need to be downsampled into daily mean data to be compatible with the reservoir data. Some of the data is also in water year and irrigation year format, which I still need to consider how to wrangle. Then I will concatenate this data containing the target variable with the two other data sets for water runoff and climate data once the time signatures match. Then the data will require interpolation or dropping of values for many small stretches of time when maintenance was performed on the sondes. Then removing seasonality and other trends will be performed to identify and isolate the determining factors. The dataset with the most rows is 285,000, and the least is 12,365, which will likely be closer to the size of the actual dataset before the train/test split.

## Evaluation

Tentatively, I would like to use time series regression to determine the factors that increase turbidity, and potentially predict future values, ideally across entire seasons. I would like to have a model where the water conditions can be plugged into the model and the relative turbidity would come out. This also may warrant further investigation of how to mitigate the river turbidity by providing recommendations to the Foundation, whether it be public outreach or industrial negotiation.

## Deployment

Ultimately, I would like to make an app (possibly using Flask?) that is an interactive graph of turbidity and contributing factors, as well as a calculator. This would have particular use at the HFF for both scientists and members as a way to consider the human factors that impact turbidity and how we can mitigate them.