# Motivation

One of the most consistent challenges to engineers since the dawn of the industrial revolution is the prediction of fluid flow through pipes and machinery. There have been two general approaches to this problem:

## First Approach

1. Measure the flow of fluids through the machinery of interest
2. Measure the properties of that fluid
3. Measure the environmental conditions (e.g. temperature & pressure)
4. Use the laws of physics to explain the relationship between the fluid flow, the fluid properties and environmental conditions
5. Use those laws to predict fluid flow in the future.

## Second Approach

1. Measure the flow of fluids through the machinery of interest
2. Measure the properties of that fluid
3. Measure the environmental conditions (e.g. temperature & pressure)
4. Search for a relationship between the fluid flow, the fluid properties and the environmental conditions
5. Use that relationship to predict fluid flow in the future.

The first approach describes the typical engineering workflow that failed very early. We are only now, 200 years and billions of research dollars later, getting this approach to work consistently using high powered computers and computational fluid dynamics (CFD) models.

The second approach sounds a lot like data science, especially if we choose to define the “relationship” as a regression model or a neural network. However, even if the engineers were to drop their empirical fluid flow models tomorrow and start using CFD exclusively, the need for a data-science-like approach will not go away.

This is because the CFD models still require fluid properties. Some of these properties relevant to flow are cheap and easy to measure, such as density. Others are harder and therefore more expensive to obtain such as apparent molecular mass. Engineers will routinely use empirical models to estimate the values of harder-to-measure properties using easier-to-measure properties, and therefore the better the fluid property models the better the fluid flow models, CFD or otherwise.

Goossens (1995) presents a fluid property model that predicts the apparent molecular mass of a liquid composed of a complex mixture of organic molecules with the density and boiling point of that mixture.

A quick search of the chemical engineering literature reveals that, starting in 1969, at least 9 empirical equations that were published before the Goossens paper were designed to predict exactly the same relationship, and 6 empirical equations were published after, the latest bearing a date of 31 January 2023.

This would suggest that over the last 60 years or so, the performance of empirical equations used by engineers to substitute for unmeasured molecular mass data has been less than satisfactory.

What is different about the Goossens paper, however, is that the dataset used to generate the correlation model was included. This means that if we assume no malice in the construction of the dataset, then we can test some or all the models in the literature against this dataset.

However, with a data science mindset, we can do much, much more. For starters, we could unleash a plethora of different machine learning models against the data to see which one performs the best. However, with only 70 observations, we will have a data scarcity challenge for many models.

This has not stopped researchers from publishing at least 2 ANN models in the chemical engineering literature. Synthetic data generation is an option here, but we are unsure at this time whether this approach was used.

Alternatively, we could retreat to regression modelling, where the (relatively) small number of observations will not be an issue. We could employ non-linear regression to re-fit the constants used in the empirical equations mentioned above to the Goossens data then measure the performance of the revised models.

Finally, we could explore symbolic regression, in which we not only search for the best constants, but we also search for the best structure for the correlating equation. This is a fascinating option because a brute force implementation could run for infinite time and never find a reasonable solution.

## The Good

If it works, it is possible that we may find a completely new empirical model for this old problem. Or we may find that our new model is very similar to an older model, therefore validating one older model over other older models.

In either case, we are building on the regression courses taken as part of this degree program and taking it a step further.

There appear to be several python libraries under active development or maintenance that build these models on top of genetic algorithms.

## The Bad

There is no guarantee that this thing will work. Some of the libraries out there are secretive about what they are actually doing under the hood while they wait for their patents to come through.

There appear to be two R libraries that build these models, but only one is under active development or maintenance, so we may have to spend more time writing code if we choose R.

Finally, although the workflow proposed for this project is 100% data science, the application is deep in the bowels of chemical engineering. Can one make the Navy Federal Credit Union look at a report on “Apparent Molecular Mass”?

## The Ugly

One way out of the “irrelevance” trap is to choose a different dataset. There exists a series of datasets based on the Feynman lectures. One could argue perhaps that theoretical physics is less obscure than chemical engineering. However, our contribution to data science using these datasets is likely to be zero, as they have been used to symbolic regression already. Is this important?

Or we can look for a dataset on a more accessible topic (Dr. Seals “Scooby Doo” dataset comes to mind) and attempt to make the case that a mathematical description of the relationship between predictor and response variables may be as valuable as the raw prediction itself, as it is more explainable.

There are two closely related technologies to this one. The first is the box-cox transform (Dr. Amin’s class), which even though designed to address normality issues could be used to formulate non-linear regression.

The second is alternating conditional expectations, in which we transform both the predictor and the regressor variables to linearize the relationship between the two. Conceptually, it is quite elegant, and there is at least one R library (‘acepack’) that facilitates this kind of regression.