**Machine Learning Selection**

**Understanding the Problem**

Before we go into model selection we first need to identify the problem at hand. For my project the problem involves using several predictor variables and asses their impact on a response.

With that said this problem can be limited to a set of algorithms that deal with prediction and not classification. My professor and I discussed a few possible supervised algorithms that I think will work well with this data set. I decided to first use a multi-variable linear regression as my baseline model. I would then extend this to a polynomial or step wise regression if there is room for improvement. Lastly, I will try a random forest and compare the summary statistics amongst all and decide which model performs best with the test data.

My data is aggregated by advertiser spend by day and new customer counts. I won’t be needing time in these exercises so my dependent variable will be customer count while my independent variables will be advertiser spend. I will divide my data set into a training and test set using a 70-30 split.

Once my analysis is done will compare R squared and adjusted R squared values amongst the models and rank them accordingly. Then I will see which model performs best at predicting customer counts given spend on my test set.