**Description of the problem and a discussion of the background. Explicitly describe your audience and why they would care about your problem.**

Traffic accidents are one of the leading causes of serious injury and fatalities in the United States. In 2018 in Texas, where I live, one crash occurred every 58 seconds and one person died due to a crash every 2 hours and 24 minutes (Source: Texas Department of Transportation). According to the Association for Safe International Road Travel, 1.3 million people die in car accidents every year. Many traffic accidents could be prevented by awareness of the danger present in situations, changes in driving laws, and improved driver’s education.

Creating a machine learning algorithm to predict the severity of crashes would be useful to drivers, as well as law-making bodies. Both drivers and law-making bodies are invested in reducing crashes, in order to save lives and money. Drivers could use the knowledge to decide the safest times and conditions to drive under, as well as to be aware of situations that require extra caution. Law-making bodies could use the model to make decisions about changing speed-limits, fixing road conditions, and implementing safe driving laws to address major risk factors.

**Description of the data and how it will be used to solve problem.**

The data I will be using to build the model is a record of over 100,000 accidents in Seattle, labelled by the level of severity of the accident. The data contains 37 different attributes about crashes which could be used to build the model. The main attributes I expect to use are the road and weather conditions, speeding information, lighting conditions, and whether the driver was under the influence. I do not expect to use specific location much as a predictor, in order to make the model generalizable to other cities besides Seattle; I will use the location *type* to predict which type of location is related to accident severity. (A complete list of attributes can be found here: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>)

I will conduct exploratory data analysis to determine which attributes might be most useful in constructing the model and filter the data down to an appropriate number of predictive attributes. I will prepare the data set by removing incomplete records and formatting data to make it usable by the classification model. I will balance the data, to avoid creating a biased model, as well as separate the data into training and testing sets, to verify the accuracy of the model. I will use the data to build a logistic regression model that predicts the probability that an accident will be severe. Different combinations of features will be tested to determine which provides the most accurate modelling of accident severity. This type of model will be useful because it will provide the user with probabilities, as well as the ability to understand the impact of a particular feature on the crash outcome.