Predicting Vehicle Collision Severity

**Carlos Fortin**

**September 2020**

# Introduction

## Background

Today, most people in the developed world drive some type of motor vehicle daily. They drive to and from work, take their loved ones on vacations trips, their children to school and often do daily chores such as going to the supermarket utilizing their vehicles. Unfortunately, the more vehicles the more accidents take place. Some of these accidents are severe and can cost lives.

For an accident to occur, there are many factors involved such as traffic violations (speeding, running a red light, etc.), mechanical failures such as a flat tire causing the driver to lose control and sometimes weather or road conditions might be poor causing the driver to struggle to keep control.

## Problem

The purpose of this report is to identify key features that increase the likelihood of an accident to occur and use them to create a model that can predict the severity of that accident so that future drivers might be able to understand the risk associated with their driving under specific conditions.

# Data Acquisition and Cleaning

## Data Source

For this analysis and the model, the data from <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv> was used. This data originally contained +190,000 rows and 38 columns (features).

## Data Cleaning

Before any analysis and modeling, the data was first studied and cleansed. At first glance, there were several features that seemed to be redundant and missing so those were removed. Additionally, extreme outliers or data that seemed to be entered in mistake were also removed.

The features that were dropped from this data frame were: 'ADDRTYPE','INTKEY','X','Y','INCDATE','PEDROWNOTGRNT','INATTENTIONIND','ST\_COLCODE','ST\_COLDESC','HITPARKEDCAR','JUNCTIONTYPE','COLDETKEY','EXCEPTRSNCODE','EXCEPTRSNDESC','SEVERITYCODE.1','SDOT\_COLCODE','SDOT\_COLDESC','SDOTCOLNUM','SEGLANEKEY', 'CROSSWALKKEY','STATUS','REPORTNO'.

The features above were removed from the analysis since they describe more the conditions after the accident occurred and describe consequences of the accident after it occurred. For this analysis, the objective is to predict the severity of an accident given certain conditions prior to an accident so the features above were not needed.

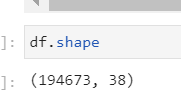
The following features were kept for further analysis:

'SEVERITYCODE', 'OBJECTID','INCKEY', 'LOCATION', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT',’'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDTTM','UNDERINFL', 'WEATHER', 'ROADCOND',

'LIGHTCOND', 'SPEEDING'.

Additionally, the data was further cleaned by dropping blanks from Weather, Road conditions, lighting, and speeding and driving under the influence. Some features contents were also modified to simplify analysis. For example, certain entried contained 0,1,N,Y. The N were all converted to 0 and the Y were all converted to 1.

Starting data frame:



Cleaned data frame:

