**Predicting the severity of a car accident**

Rouslan Gabissov

Augustus 21, 2020

# Introduction

## Background

Each year many car accidents take place. The consequences of these accidents have different dimensions:

* Human cost: injuries and fatalities of involved persons
* Material cost: damages to vehicles, road infrastructure and other property
* Business cost: cost associated with traffic delays

Everybody recognizes the importance of reducing car accidents. Governments are concerned with public health and want to make sure that the number is as low as possible. Business want to avoid costs associated with them. And, ordinary people of course want to get to their destination well and on time.   
Therefore, it is advantageous to everybody to have ability to predict probability and severity of an accident.

## Problem

The features that could predict severity of possible accident might include location, day of week, hour, weather condition, vehicle type, roadway configuration, road surface and traffic control.

The aim of this project is to predict severity of possible car accident based on the available data.

## Interest

Car (software in built-in gps systems) and gps navigation software companies (like Google Maps and Waze) could be interested in incorporating predication of severity of possible incidents into their software. In addition, the model could be interesting to public officials interested in improving road safety by for example selecting locations for extra traffic controls or adjusted speed regimes.

# Data acquisition

## 2.1 Data source

The project’s data is related to Canada and provided by Transport Canada and Statistics Canada. The data set can be found on Kaggle [here](https://www.kaggle.com/tbsteal/canadian-car-accidents-19942014). The set contains 5.86m records observed in the period 1999 to 2014 and 22 columns. All of the features required are present with expectation of location.  
Pre-processing will be required. For instance, there could be duplicated categories in the form of '01' and '1'. The data formatting does not always appear to be consistent. Additionally, some data is unknown or not provided by the jurisdiction.