1. Introduction: The project

1.1. Scenario

“Tomorrow evening I have to go visit my brother's family. He lives in another neighborhood, on the other side of town. Shall I drive there or take the metro? Driving is certainly more convenient; will I run into an accident if I take the car? Shall I move the visit to another day?

This is a typcal everyday situation. To drive or not to drive? Take the car or find another way? If only someone could know. Answering this question is the scope of this project: provide people with a predictive tool, able to warn them about the likelihood of car accidents and their severity, allowing them to make a more informed decision regarding their future travel plans, and choice of transportation.

1.2 Scope of the project

In this project we will try to predict the likelihood and the severity of road accidents in the city of Seattle based on historical data collected by the Seattle Police Department (SPD), and the Seattle Department of Transportation (SDOT), from 2004.

The **target market** of the project is:

- anyone with a driving license, and a mean of transport (the drivers), interested in knowing the probability of getting into a car accident and how severe it would be, given factors such as weather, time of the year, location, road conditions..., so that they would drive more carefully or even change their travel plans, if possible;

- public authorities such as police, departments of transportation, road authorities, healthcare providers... in order to be better prepared in dealing with such events knowing the probability and the scale of the problem.

Probability and severity of accidents will be assessed using the available historical data in terms of severity classes, and other available information such as collision locations, date and time of the event, weather, road and light conditions...

In order to deliver a product able to satisfy all relevant stakeholders' needs, we'll use an analytic, machine-learning driven approach to build 2 predictive classification models, as follows:

- **1st model - Classification of Risk**. Is the risk of getting into car accident higher than usual? The model will return a binary-class prediction, as follows:
    1. Low risk: input feature data match conditions whose “number of collisions” (or frequency distribution) is less than the average number of collisions.
    2. High risk: input feature data match conditions whose “number of collisions” (or frequency distribution) is greater or equal to the average number of collisions.

- **2nd model - Classification of Collision Severity**: how severe the potential collision is. The model will return a binary-class prediction, as follows:
    1. Not severe (property damaged only);
    2. Severe (injuries and/or fatalities).