2. Data

2.1 Data Requirements

In order to build a predictive model able to forecast the risk and severity of vehicles' collisions, we need historical data, as more comprehensive as possible, on road accidents in the region/city of interest, in our case Seattle. For better predictive results, the dataset should include attributes such as:

- Time of the year
- Day of the week
- Time of the day
- Incident's location
- Weather condition
- Light condition
- Road condition
- The collision severity classification

It is important to keep in mind that the selected analytic approach, in this project, is predictive, and not explanatory or descriptive.

> *In predictive modeling [...] criteria for choosing predictors are quality of the association between the predictors and the response, data quality, and* ***availability of the predictors at the time of prediction****, known as ex-ante availability. In terms of ex-ante availability, whereas chronological precedence of X to Y is necessary in causal models, in predictive models not only must X precede Y , but* ***X must be available at the time of prediction****.*
>
> — Prof. Galit Shmueli, Institute of Service Science, College of Technology Management, National Tsing Hua University. Reference: To Explain or to Predict?

Information such as the type of collision, the number of people involved, the kind of reported damage... can be very insightful, and have great causal meaning when it comes to explain the severity of an accident, but can't really be used to build a predictive model because these pieces of information are available only once an incident has already actually occurred (ex: how can someone know the kind of accident, or the number of people involved, when trying to predict in advance the likelihood and the severity of the accident itself?).

Therefore, in the attribute selection process, we're going to select only the attribute that can be actually used as input for our predictive model.

2.2. Data Source

For the scope of this project I am going to use a dataset containing information about all types of collisions that have happened in the city of Seattle, collected by Seattle Police Department (SPD) and recorded by Traffic Records.

2.3. Data Understanding

The dataset has been built with data on all types of collisions from January 1st, 2004, to May 20th, 2020, containing **194,673 records of collisions, with 37 attributes** (severity label column not included). The database contains a mix of data types, including integers, floats, and text type. 136,485 collisions (70% of total) belong to the Severity Class 1 (Property Damage Only); the dataset is quite imbalanced; this is something we should keep in mind in the stage of model testing, and, in case, take some corrective actions.