# Introduction

To complete the capstone project of the IBM Professional Data Science certificate, we had to create a model to predict the severity of an accident.

We used the Seattle collision Data supplied by: SDOT Traffic Management Division, Traffic Records Group.

Except the purpose and the Dataset, we were free.

After completing this exam, I wanted to go further and see if I could have better results with other algorithms.

Then, everything started with one question “What is the optimal Data split?”.

I found no definite answers on internet, as I wanted to put some answers on those question, I decided to explore every question I could have, and this is precisely the purpose of the research project.

Every time a new question arises, I will write it in this research project, and write a script to automate the finding that can be applied on every dataset.

# The questions we explore

1. Optimal Data split between training, validation and tunning
2. Compare the cross validation and manual split in terms of accuracy and speed
3. Explore the parameters of the K-Nearest Neighbours and the Random Forest algorithm
4. Analyse the impact on accuracy and speed to use only a small part of the data to choose the parameters. Do the parameters differ between small and full set? How does the size of the dataset impact the results?
5. Compare the difference in speed and accuracy between the K-Nearest Neighbours and the Random Forest algorithm. What if we have no clue about the optimal parameters?