## Forward

I did the exercise for the previous session of the Capstone project. But I was delayed and had to postpone my submission. However the problem changed between the last session and this one. So the problem handled here is different from the one asked, but I’d rather avoid having to start again from scratch if possible. Sorry for the inconvenience and thank you for your understanding!

## Introduction to the business problem

Car accidents are one of the major causes of death nowadays. Being able to better understand what causes them could greatly help reduce the number of accident and save lives. In this project we try to build predictors of what are the main causes of road accidents and more specifically predict the severity of said accident in term injuries or death of people involved. In this project we'll mainly focused of accidents occurring in Belgium between 2005 and 2019. The results could help various stakeholders:

* Help roads users (drivers, cyclist, pedestrians,) better understand factors that increase the risk of accident and adapt their behaviour accordingly.
* Help road safety agencies design awareness campaigns focusing on the most appropriate message
* Help emergency services better anticipate when sever accident are likely to occur to size response teams accordingly

Other more long term potentials could be considered like guiding the development of new or improved safety features in vehicles or on roads. Such algorithms could also help autonomous vehicles identify risky conditions and adapt the driving behaviour accordingly.

## Presentation of the data

To build our model we'll use historical car accident casualties’ data available on the Belgian statistical website [www.data.gov.be](http://www.data.gov.be). This data are available on an annual basis for every year between 2005 and 2019. and includes only the accidents with casualties expressed as "slightly injured", "severely injured" and/or "death" as well as the number of victims of each type for each accident. The dataset also contains various information about the accidents like:

* Date, day of the week and time
* Lighting conditions (day, sunrise/sunset, night with public lighting and night without public lighting)
* Location of the accident (address), type of neighbourhood (inside or outside city) and type of road (local, regional or highway)
* Identity of the victim (age category and sex) and type (driver, passenger, pedestrian, cyclist)
* Type of vehicle involved

Unfortunately some potentially intersecting features are missing. More details about weather conditions could be useful (rain, frost, snow,) but are not directly available. This information could however be reconstructed using historical weather data publicly available from weather data provider using date, time and location of each accident (e.g. from <https://dev.meteostat.net/>). Other missing information like was speeding, alcohol or drugs involved will probably be difficult to reconstruct. Age, time and day of the week of the accident could give us proxies of this but we probably need new dataset including these features to properly account for them.

The work was divided in 2 main parts.

* The analysis of the data were the data are imported and cleaned. And some statistical analysis are done to g better understand the data at hand
* The modelling of the data were we tried to design algorithm to predict the severity of an accident.