Background

According to WHO, the number of road traffic deaths rising steadily up to 1.35 million in 2016. It is the 8th leading cause of death, less likely to survive than AIDs.

Problem

Prevention is always better than cure. This project is to predict the injury in car accidents. This is concerned by drivers for the purchase of insurance as well as the insurance company to adjust the insurance premium and the claim. On top of those, minimize the injury is most important purpose.

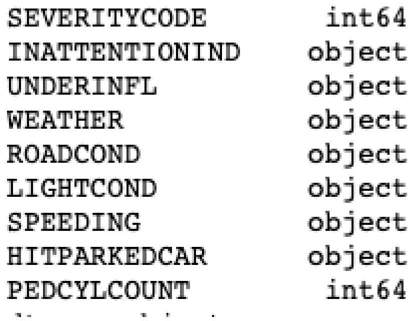
Data sources

The data I will us is obtained from this class's, provided by SDOT Traffic Management Division, Traffic Records Group, from 2004 to Present. The dataset comes with many factors that may affect the probability of accidents: Geographic data, accident details, timestamp, road types at which accident happened like intersection, driver behavior, weather, road condition, light condition.

Data cleaning and Feature selection

There were 194673 samples and 38 features in the data. Looking into the features, some id-type features don't help the analysis and thus dropped out. Duplicated and details of the crashes will also be removed because we are focusing on the injury only. As such I picked road types at which accident happened, driver behavior, weather, road condition, light condition.

In other words,



```
SEVERITYCODE       int64
INATTENTIONIND    object
UNDERINFL         object
WEATHER           object
ROADCOND          object
LIGHTCOND         object
SPEEDING          object
HITPARKEDCAR      object
PEDCYLCOUNT        int64
```

Where

1)Severity code is the prediction group,

2)inattentionind is the driver attention status,

3) underinfl is the status that driver is involved in drugs and alcohol

4) Weather, road and Light condition

5) whether the driver was speeding during the accident

6) hitparkedcar is whether the accident happened in car park

7) pedcylcount is whether bicycle involved in the accident