**Behind The Wheel**

**Mustafa Barış Emektar** **- 2304533**

**Berk Güler - 2310092**

# Short Project Description

In this project, we’ll dive into a comprehensive dataset that captures accidents across the United States over a span of several years. The features like location, weather conditions, time of day, and road characteristics of the US Accidents dataset will be investigated. The goal is to explore the meaningful relationships between these features, the underlying dynamics of accidents, and factors that influence accident severity. Additionally, actionable insights to prevent accidents will be discovered.

The problem of analyzing and understanding accidents is of extreme importance due to several reasons:

* **Human Lives and Public Safety:** Accidents result in loss of life, injuries, and trauma. By studying accident patterns, we can identify preventive measures to reduce fatalities and injuries. Also, we can develop targeted policies to improve road safety.
* **Economic Impact:** Accidents lead to significant economic losses through medical expenses, property damage, and insurance claims. Understanding the causes and severity of accidents can help allocate resources more effectively.
* **Infrastructure and Traffic Management:** Accidents impact traffic flow, road infrastructure, and emergency response systems. Analyzing accident data helps to make decisions on road design, traffic management, and emergency services.

In summary, solving this problem contributes to safer roads, reduced economic burden, and improved public well-being.

# Methodology

## What is the problem/question you are solving?

The problem being addressed in this project is to analyze and understand accident patterns and severity using the US Accidents dataset. Specifically, the exploration of relationships between various features (such as weather conditions, time of day, and location) and their impact on accident outcomes is aimed at. By doing so, we can identify actionable insights for improving road safety and minimizing the consequences of accidents.

## Dataset

For this project, we will be using the US Accidents dataset from Kaggle. It covers accidents that occurred in the United States from 2016 to 2023. Here is why this dataset is appropriate to use:

* **Time Scope and Size**: The dataset covers 7 years of accidents that occurred in the 49 states of the USA. It consists of 7.7 million accident records with 3.06GB of data.
* **Rich Features**: The dataset provides a rich source of data for our analysis. It contains detailed information about each accident, including features such as location, severity, weather conditions, time of day, and road characteristics. It contains 46 columns (different features). These (possibly) relevant features can help us understand accident patterns and severity.
* **Severity Labels**: The dataset includes labels for accident severity (e.g., fatal, minor), which is crucial for our analysis. Without these labels, we would treat all accidents equally, missing critical insights into the impact of accidents.

## How will you solve the problem? Plan of action.

Here’s how we will approach this problem:

* **Exploratory Data Analysis (EDA):** We will start by exploring the dataset: visualize distributions, correlations, and summary statistics. Key features like location, time, and weather that might influence accident severity will be identified.
* **Feature Engineering:** New features can be created if needed (e.g., day of the week, time of day). Missing values and outliers will be handled if they exist.
* **Relationship Discovery:** The relationships between features will be investigated. For instance, are certain weather conditions associated with more severe accidents? Does accident severity vary by location? Are there temporal patterns? For instance, are there more accidents during rush hours?
* **Machine Learning Models:** Regression models like linear regression can be considered to predict accident severity based on relevant features. Also, classification models like decision trees and random forests can be utilized to predict if the accident is fatal or not.
* **Scaling Up:** To scale up, Apache Spark can be used for handling large datasets efficiently since it parallelizes computations to speed up processing and optimizes memory/storage usage.
* **Visualization and Interpretation:** Visualizations like heatmaps, graphs, or scatter plots will be created to highlight relationships. Also, temporal trends will be shown, and geographical maps will be used to show accident hotspots. The model results will be interpreted to understand which features contribute most to accident severity.

## How will you evaluate your method?

When assessing the regression models; mean absolute error (MAE) and root mean squared error (RMSE) can be used. Also; accuracy, precision, recall, and F1-score are some of the metrics that can be used to evaluate classification models.

Our model’s performance will be compared against simple baselines:

* **Predicting Mean Severity:** A straightforward baseline is to predict the mean severity of all accidents. Our model should outperform this basic approach.
* **Using Basic Rules**: The model will also be compared against rules based on common sense or domain knowledge. For instance, it is considered that accidents during rush hour are more serious. The performance of our model should be better than these rule-based baselines.