### Introduction to the Project

* **Business Context:**
  + This project is embedded in ongoing efforts to enhance road safety. By analyzing historical accident data provided by the Observatoire National Interministériel de la Sécurité Routière (ONISR), we aim to identify high-risk areas, understand the underlying causes of accidents.
* **Technical Point of View:**
  + Technically, the project involves handling and analyzing a large volume of data extracted from the BAAC (Bulletin et al.) files. The process includes data cleaning, pre-processing, visualization, and statistical analysis. Advanced data analytics techniques, including machine learning, will be applied to build predictive models and derive actionable insights.
* **Economic Point of View:**
  + Economically, reducing road accidents translates to significant cost savings for the government and citizens. By preventing accidents, we can reduce healthcare costs, lower the financial burden on victims and their families, and minimize the economic impact of traffic disruptions. Enhanced road safety also leads to increased productivity and economic efficiency.
* **Scientific Point of View:**
  + Scientifically, this project contributes to road safety research by providing a data-driven approach to understanding accident patterns and their causes. It supports the development of predictive models and safety interventions based on empirical evidence, enhancing our ability to prevent accidents and save lives.

### Objectives

* **Main Objectives:**
  + The primary objective is to analyze and visualize accident data to identify patterns, risk factors, and high-risk locations. Specifically, we aim to:
    - Identify temporal and spatial trends in road accidents.
    - Determine the demographic profiles of accident victims.
    - Understand the impact of weather, road conditions, and other environmental factors on accident rates.
    - Develop predictive models to forecast accident hotspots and severity.

### Understanding and Manipulation of Data

#### Framework

* **Data Sets Used:**
  + The primary dataset used is the annual accident data from ONISR, covering 2019 to 2022. This dataset is structured into four primary files: Caractéristiques (accident characteristics), Lieux (locations), Véhicules (vehicles involved), and Usagers (users involved).
* **Availability:**
  + The data is publicly available on the ONISR website, but certain privacy-sensitive information is omitted. This ensures compliance with data protection regulations while providing comprehensive accident data for analysis.
* **Volume of Dataset:**
  + The dataset comprises millions of records, with each year contributing a substantial number of entries. Each record includes detailed information about the accident, including date, time, location, vehicle types, and user demographics (see Appendix 1)

#### Relevance

* **Relevant Variables:**
  + Key variables identified include:
    - **Temporal Variables:** Date, time of accident.
    - **Spatial Variables:** Location coordinates, road type, intersection type.
    - **Environmental Variables:** Weather conditions, lighting conditions.
    - **Vehicle Variables:** Vehicle type, vehicle maneuver before the accident.
    - **User Variables:** Age, gender, injury severity.
* **Target Variable:**
  + The primary target variable is the severity of injuries sustained, categorized into:
    - Indemne (uninjured)
    - Tué (fatal)
    - Blessé hospitalisé (hospitalized injury)
    - Blessé léger (minor injury)
* **Dataset Features:**
  + The dataset provides a comprehensive view of each accident, including:
    - Detailed accident descriptions (time, location, conditions).
    - Demographic information of involved parties.
    - Vehicle details and maneuvers.
* **Data Limitations:**
  + Limitations include missing values, inconsistencies in data recording over different years, and the exclusion of certain personal information for privacy reasons.

#### Pre-processing and Feature Engineering

* **Data Cleaning:**
  + The data cleaning process involved:
    - Handling missing values through imputation or exclusion.
    - Standardizing variable formats (e.g., date and time formats).
    - Correcting inconsistencies and errors in the dataset.
* **Normalization/Standardization:**
  + Normalization was applied to ensure consistent scales across variables. For instance, numerical variables like age and vehicle speed were standardized to have a mean of zero and a standard deviation of one.
* **Dimension Reduction Techniques:**
  + Techniques like Principal Component Analysis (PCA) were considered to reduce the dataset's dimensionality. This helps focus on the most informative variables and improves model performance by reducing noise and redundancy.

#### Visualizations and Statistics

* **Variable Relationships:**
  + Relationships between variables were explored using:
    - **Correlation Analysis:** Identifying correlations between different variables.
    - **Scatter Plots and Heatmaps:** Visualizing relationships between pairs of variables.
    - **Pair Plots:** Showing pairwise relationships in the dataset.
* **Data Distribution:**
  + Data distribution was analyzed to understand the central tendencies and variability. Tools used include:
    - **Histograms:** Showing the distribution of continuous variables.
    - **Box Plots:** Identifying outliers and understanding the spread of data.
* **Statistical Analyses:**
  + To validate the observed relationships and differences, statistical tests, such as chi-square tests for categorical variables and t-tests for continuous variables, were conducted.
* **Conclusions for Modeling:**
  + The exploratory analysis provided crucial insights for the modeling phase. Key findings include:
    - Temporal trends indicate peak accident times.
    - High-risk locations are identified through spatial analysis.
    - Demographic patterns highlighting vulnerable groups.
    - Environmental factors affecting accident severity.
  + These insights will guide feature selection and model development, ensuring the predictive models are robust and relevant to the identified risk factors.