# Data mining project

## Chicago Vehicles Crashes

A.Y. 2024/2025

The project is about traffic incidents in Chicago. It aims to simulate a decision support system for an assurance company. Attached to this document, you can find 3 distinct files: crashes.csv, people.csv, and vehicles.csv.

- crashes.csv contains the main body of data: a table with data about road incidents between January 2016 and December 2024 in Chicago, USA. The same table also includes information about the causes of the incident, some road properties, and some information about the reported injuries.
- people.csv contains information about the people involved in the incidents, including their sex, age, city of residency, etc.
- vehicles.csv contains information about the vehicle(s) involved in the incidents. This file includes information about the vehicle and the information collected by the police after the incident.

The **project** consists of data analysis based on data mining tools. It has to be performed by using Python. The guidelines require addressing specific tasks, and results must be reported in a unique paper. This paper's total length must be **20 pages** of text including figures. The students must deliver both: paper and well-commented Python notebooks. The paper must contain important aspects for the evaluation.

## Dataset description

The data are divided in 3 csv files, which can be downloaded by the following URLs. Moreover, in the following web pages, you will also find a **detailed description** of the columns of each of the 3 aforementioned files.

- **crashes.csv**
  **https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data**
- **people.csv**
  **https://data.cityofchicago.org/Transportation/Traffic-Crashes-People/u6pd-qa9d/about_data**
- **vehicles.csv**
  **https://data.cityofchicago.org/Transportation/Traffic-Crashes-Vehicles/68nd-jvt3/about_data**

Alternatively, you can download the 3 csv files, in zip format from the following URL:
GOOGLE DRIVE DOWNLOAD.

# Task1: Data Understanding and Preparation (30 points)

### Task 1.1: Data Understanding

Crashes.csv file must be **mandatory** analyzed, while people.csv and vehicle.csv are optional. Nevertheless, they contain crucial information for the following tasks. Hence, we suggest to exploit the information of at least one of the two other csv files.

Explore the incidents dataset with the analytical tools studied and write a concise "data understanding" report assessing data quality, the distribution of the variables and the pairwise correlations.

### Task 1.2: Data Preparation

Analyze the features assessing their quality and analyzing their statistics and distributions; improve the quality of your data and prepare it by extracting new features.

You must create an incident **profile for each month and year of the police department** (please, consider a different profile of the month in different years, so Jan 2021 is different from Jan 2022). The police department is identified by the column *BEAT_OF_OCCURRENCE*. Examples of indicators to be computed are:

1. The average age of involved people in a department in a certain month
2. The average speed limit in a department in a certain month with respect to the whole city
3. Estimating level of injuries in each department in each month.

Note that these examples are not mandatory. You can derive indicators that you prefer and that you consider interesting for describing the incidents.

It is MANDATORY that each team defines some indicators. Each of them has to be correlated with a description and when it is necessary also its mathematical formulation. The extracted variables will be useful for the clustering analysis (i.e., the second project's task). Once the set of indicators is computed, the team has to explore the new features for a statistical analysis (distributions, outliers, visualizations, correlations).

**Subtasks of DU**:

- Data semantics for each feature that is not described above and the new one defined by the team
- Distribution of the variables and statistics
- Assessing data quality (missing values, outliers, duplicated records, errors)
- Variables transformations
- Pairwise correlations and eventual elimination of redundant variables.

Nice visualization and insights can be obtained, exploiting the latitude and longitude features (e.g. https://plotly.com/python/getting-started/).

## Task 2: Clustering analysis (30 POINTS - 32 with optional subtask)

Based on the features extracted in the previous task, explore the **monthly profile of the department (**please, consider a different profile of the month in different years, so Jan 2021 is different from Jan 2022**)** using various clustering techniques. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

**Subtasks**

- Clustering Analysis by K-means on the entire dataset:
    1. Identification of the best value of k
    2. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset
    3. Evaluation of the clustering results
- Analysis by density-based clustering**.** In this task, choose **one state** in the dataset:
    1. Study of the clustering parameters
    2. Characterization and interpretation of the obtained clusters
- Final evaluation of the best clustering approach and comparison of the clustering obtained
- **Optional (2 points):** Explore the opportunity to use alternative clustering techniques in the library: https://github.com/annoviko/pyclustering/

## Task 3: Predictive Analysis (32 POINTS)

Consider the problem of predicting for each **monthly profile of the department (**please, consider a different profile of the month in different years, so Jan 2021 is different from Jan 2022**),** the DAMAGE of the incidents. Therefore you must **create a target variable**, by aggregating the values of DAMAGE of the incidents in each department, year, month. The students need to:

1) define new features that enable the classification. Please, reason on the suitability of the features defined for the clustering analysis. In case these features are not suitable for the above prediction problem you can also change the indicators.
2) perform the predictive analysis comparing the performance of different models discussing the results and discussing the possible preprocessing applied to the data for managing possible identified problems that can make the prediction hard. Note that the evaluation should be performed on both training and test sets.

**Mandatory models**: The following models must be analyzed: Decision Tree, KNN, Ensemble Methods.

**Optional (2 points):** Explore the opportunity to use alternative classification models.

# Rules for final delivery and Exam

**Project Delivery.**The final deadline indicated by the teacher by email is **STRICT**. Each group must deliver by email to anna.monreale@unipi.it, mattia.setzu@unipi.it, and lorenzo.mannocci@phd.unipi.it a zipped folder named **DM_GroupID.zip** and containing 4 folders and 1 pdf file:

1. a folder named **DM_GroupID_TASK1**, containing source code of data understanding
2. a folder named **DM_GroupID_TASK2**, containing source code of data clustering
3. a folder named **DM_GroupID_TASK3**, containing source code of classification
4. a folder named **DM_GroupID_TASK4**, containing source code of time series analysis/explanation analysis
5. a pdf file with maximum 20 pages including figures discussing the results of the 4 tasks. The name of this file must be: **DM_Report_GroupID.pdf.** The file must contain the list of authors (i.e., members of the group).

We prefer to have group presentations of the project. If this is impossible we can find a solution together.