# CMP4011 Big Data and Cloud Computing

# Project Report

## Team 10

| Name | Sec | B.N | Code |
|---|---|---|---|
| Ahmed Osama Helmy | 1 | 5 | 9213061 |
| Omar Mahmoud | 1 | 29 | 9210758 |
| Abdallah Ahmed | 1 | 25 | 9210652 |
| Aliaa Gheis | 1 | 27 | 9210694 |

## Problem Statement:

Road safety is a critical concern, and understanding accident patterns can help cities improve traffic management and reduce accident rates. This project aims to analyze accident data to identify high-risk locations, contributing factors, and potential mitigation strategies. By leveraging big data processing, we will extract valuable insights for transportation authorities and urban planners.

## Dataset

**Dataset Name:** US Accidents (2016 - 2023)

**Link:** https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

**Description:**

- Contains 7.7 million accident records with 46 columns
- Main features of accident are (Severity of Accident, Time, Location, Weather, Road Characteristics)

## Planned Approach or Proposed Solution:

1. **Exploratory Data Analysis (EDA):**
    - Analyze accident severity distribution and trends over time.
    - Identify correlations between weather conditions, traffic signals, and accident severity.
2. **Descriptive Analysis:**
    - Implement K-Means clustering with MapReduce to identify accident hotspots
    - Apriori algorithm for association rule mining to discover patterns in accident contributing factors
3. **Predictive Analysis:**
    - Apply Classification models (Random Forest, SVM) to predict accident severity based on environmental and traffic factors.
    - Implement Regression models (Linear Regression, Gradient Boosting) to estimate accident frequency per location.
4. **Big Data Implementation:**
    - Use Python & Apache Spark.
    - Implement K-Nearest Neighbors (KNN) using MapReduce to classify accident severity based on past accident characteristics.
    - Deploy on Azure (Maybe Azure HDInsight) for real-time accident risk analysis.