# A Scenario-based Database for Connected and Autonomous Driving in A Smart City

### Introduction

In this project, we aim to build a scenario-based driving database that is dedicated to connected and autonomous vehicles. We plan to record and model the dynamic traffic information in Pittsburgh from heterogeneous driving data such as lidar point cloud, vision information, GPS, etc. After that, an unsupervised learning approach based on nonparametric Bayesian will be applied to learn and recognize driving scenarios by segmentation. A user-friendly web application will be developed to provide the dataset to public from a scenario perspective.

### Types of data produced

The traffic information will be recorded by a data collection platform which is equipped with multiple advanced sensors including Lidar, high-resolution camera, radar, GPS, IMU units, and vehicle information such as steering wheels and braking pedals. The platform is able to capture the complex and informative real-world driving scenarios and categorize them as high-dimensional and heterogeneous time series data. The on-vehicle sensors are organized by ROS (Robot Operating System). The raw data will be recorded as rosbag files and image sequences.

### Data and metadata standards

The on-vehicle ECU installed with ROS will serve as the hardware terminal. All the collected raw data including GPS trajectory, images, Lidar point clouds, and steering input etc will be packaged as rosbag files. After data cleansing, and processing, the raw data will be converted into multidimensional time series and will be uploaded to MySQL database. The PI and research assistants will code algorithms using SQL and python that is compatible for any platform. The unsupervised learning method will be applied to extract and segment the traffic scenarios. The index of traffic scenarios will be updated with the time series traffic data in the MySQL database. A web application will be provided for users to query traffic data.

### Policies for access and sharing

Throughout the duration of the work, the PI will in a timely manner communicate any significant findings with the scientific community through journal publications, national and international conference presentations, and seminars. The reported results will be made available to the research community, where possible and permitted and upon request.

### Policies for re-use, redistribution

Data derived from this project shall be retained for at least one year. The selected research results will be open source and shared to the research community through technical reports or publications. The data in this project does not contain private or confidential information. The research results are belonged to PI’s university, selected result data and visualization can be obtained upon request by asking PI and research assistants.

### Plans for archiving & preservation

All collected data will reside on PCs and workstations belonging to the PIs’ university. A data server which is expected to have MySQL installed will be set up to store all the data and hold the web application. All data will be regularly backed up either onto multiple external hard drives, or a centralized backup cloud storage, to ensure full data recovery in the event of equipment failure. In the case of catastrophic failures, we will maintain both the data server and the web application indefinitely.

### Software Sharing Plan

Our existing software code base for data processing and integration, as well as the unsupervised learning algorithm, has been developed using open source software and libraries including MySQL and python. The data integration tool is belonged to PI’s university. The selected developed tools and web application access will be made available to the research community, where possible and permitted and upon request by asking PI and research assistants.