1. Data preparation: Xie/Bugatti – 2018/2019

* **Data formatting**

1. The original two files contain data for both 2018 and 2019, encompassing records from all sensors. Our task is to separate and extract data for each individual sensor for the two respective years.
2. We extract sensor files for each unique detector ID for both the years 2018 and 2019. This involves connecting to a MySQL database, fetching data from the 2018 and 2019 tables in chunks, creating separate CSV files for each detector ID found in the fetched data, and storing them in respective folders on the desktop for both years. The data is still not aggregated.
3. Add a header to the files if not already there.
4. To aggregate lane data and convert it into our desired format, we consolidate unique timestamps into single rows, merging volume, speed, and occupancy values across lanes. Then, we transform individual lane data into aggregated values by calculating the mean speed and sum of occupancy and volume. Check if it has been transformed to our desired format – has 4 columns.
5. Get a common group of sensors for 2018 and 2019. There are initial 770 common sensors.

* **Data quality check**

## Sensor data aggregation

* 1. Uploaded the data files (2) received from XIE to the shredder.
  2. Utilized column 22, '**DetectorID**', to extract data from the database and store it as separate CSV files for each detector. In each detector file, every row represents a lane value and the corresponding measurement timestamp.

Observation:

* There are 825 sensor files in 2018 and 868 in 2019.
* The timestamp was supposed to be in 15 minutes interval which it is not.

Notably, if a unique timestamp is recorded, it signifies that all lanes were measured simultaneously and will be treated as a single record when aggregating lane values.

## Data Analysis and Visualization

* 1. Get a sense of the missing values. There is supposed to have 35040 rows in a year with a 15-minute time interval. The time since it’s not consistently sampled needs to be fixed at 15-minute intervals.
  2. Separate sensor files according to it regions:

## Dealing with missing values