1. Data Cleaning, Data Source: Xie/Bugatti – 2018/2019

# Process Outline

## 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.

* 1. Aggregate lane data by consolidating unique timestamps into single rows, merging volume, speed, and occupancy values. This transformation involves converting the format from individual lanes to aggregated values.

Key considerations include:

* Merging lanes into aggregated values.
* Calculating mean for speed and sum for occupancy and volume for aggregation.
* Get a common group of sensors for 2018 and 2019. There are 770 common sensors.

## 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