PROJECT GOALS
1. Identify Aggregated Cruising
2. Differentiate Parking and For-Hire Vehicles
3. Protect Privacy

DATA PROCESSING AND CLASSIFICATION

1. ESTIMATE PATHS: The first step in our analysis is to derive probable routes from the scattered sensor readings. Identifying precise trips or paths would be very difficult to accomplish due to the data consistency and reliability issues. We estimate path information through a series of algorithms that group readings, remove paths that are too short to process, and clean false detections, resulting in a derived route that is mapped the actual street grid in Seattle.

2. EXTRACT METADATA: Metadata features are extracted from the estimated paths. These represent travel characteristics that can be used for classification and machine learning.

3. LABEL VEHICLES-FOR-HIRE: We differentiate vehicles-for-hire from other traffic using a simple algorithm, prior to machine learning. If a vehicle leaves the sensor grid (is not detected) for more than 15 minutes, and if it does so more than 5 times throughout a day, it is likely either a for-hire vehicle or a bus. Bus drivers can then be filtered out by their “dispersion ratio” (total number of times a vehicle is detected divided by the number of unique sensors that detected it). If this ratio is around 3, the traveler did not cover the same ground many times, and so cannot be a bus driver.

4. SEMI-SUPERVISED MACHINE LEARNING: Cruising classification was completed in multiple steps. First, the distance ratio (shortest distance between start and end sensor hits / routed distance) was used to classify the cases as either probably cruising or not probably cruising. An illustrated below, a trip that follows the most direct path is probably not cruising, and a path that meanders significantly is probably cruising. A distance ratio of 7.0 suggests that the path traveled 6 times more than necessary, contributing to traffic and congestion along the way.

5. AGGREGATE FOR HEATMAP: We developed aggregation scripts to summarize the results. To protect privacy, only the aggregate data stream and heatmap were available for analysis. We used a web app to demonstrate potential uses. The algorithms and data would be a potential candidate for use in a linked data repository with strong governance, such as the University of Washington Transportation Data Collaborative.

RESULTS
Approximately 35% of the sample data was labeled as cruising and visualized. Of that amount, activity attributed to vehicles-for-hire is in the range of 10% or fewer. We found that the number of time crossed, average speed, and percentage of time driving were the most important features considered by the gradient boosting classifier.