# Project Proposal: Short-Term Traffic Congestion Forecasting for Seattle’s I-5 Corridor

## Team Members

Hannah Belville, Cong Ho, Paul Skentzos

## Problem Statement

Seattle is consistently ranked as one of the most congested cities in the United States, with drivers losing significant time and productivity each year. Traffic congestion not only costs the regional economy but also leads to increased vehicle emissions and delayed emergency response times. Addressing this problem requires predictive modeling that allows transportation agencies to intervene proactively.

The goal of this project is to develop a robust, data-driven model for short-term traffic congestion forecasting on a critical segment of the Puget Sound region’s infrastructure: Interstate 5 (I-5) within the Seattle metropolitan area.

### The primary question we aim to answer

Can we accurately predict the average vehicle speed (or travel time) on a specific congested segment of the I-5 freeway 15 and 30 minutes into the future, given current and recent historical traffic data, temporal features, and relevant weather conditions?

### The Importance of this Question

Accurate short-term forecasting is crucial for enabling proactive management of Intelligent Transportation Systems (ITS), such as dynamic ramp metering and traveler services, and improving resource allocation for WSDOT and emergency services to minimize secondary incidents and reduce overall delay.

## Data Sources

We plan to use a combination of publicly available, high-quality datasets, which are sufficient and standard for this type of time-series analysis:

* **WSDOT Traffic Data**: We will use archived loop detector data from the Traffic Data Acquisition System (TDAS) for I-5 in the Seattle area (e.g., Federal Way to Lynnwood). This data, which includes volume, occupancy, and speed metrics at 5-minute intervals, will serve as both our core feature set (lagged values) and our prediction target. The data is available via the WSDOT public data portal, and data acquisition and initial parsing is the immediate next step (Week 4).
* **NOAA / Local Weather API**: We will integrate historical daily or hourly weather data (temperature, precipitation, visibility) for the King County area. This serves as an Exogenous Feature, as weather is a known, non-linear contributor to traffic speed and congestion, improving overall predictive accuracy. NOAA data is publicly available, requiring only temporal matching with the traffic data.
* **Seattle Open Data Portal (SDOT Collisions)**: Collision records for the relevant period will be utilized as a Feature Augmentation. This allows us to identify and model the effect of non-recurrent congestion (unexpected accidents) on typical flow patterns. The data is readily available and will be mapped temporally to the I-5 sensor data.

**Data Preprocessing and Cleaning Required:** Yes, a significant amount of preprocessing will be necessary to prepare the data. This effort includes handling detector noise and imputing missing time intervals in the traffic data, aligning the different data sources, and extracting essential temporal features (e.g., day of the week, holiday status, time-of-day transformations).

## Analytical Approach

Our general problem (predicting future I-5 speed) will be expressed as a supervised time-series regression problem.

1. **Model Formulation:**
2. **Baseline Modeling:** We will first establish a benchmark using a classic time-series model.
3. **Advanced Modeling:** We will implement some advanced machine learning techniques, that could focus on:
   * **Gradient Boosting Machines (e.g., XGBoost):**
4. **Evaluation:** Model performance will be evaluated primarily using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to quantify the magnitude of prediction errors in miles per hour.

## Solution Technologies

* **Core Programming:** Python (Jupyter Notebooks) for all processing and modeling.
* **Data Handling:** Pandas and NumPy for data cleaning and manipulation.
* **Modeling Frameworks:** Scikit-learn for baseline and ensemble models; Possibly PyTorch .
* **Data Visualization:** Matplotliband Seaborn, initially, for display of data.
* **Data Size:** It is believed that the data sets will be able to be processed on the team’s individual computers. As we get deeper into the data, this assumption could be challenged.

## Challenges

The main challenges will be in modeling surprise congestion and achieving high accuracy with any models we might make use of.

Accidents and some weather events are “patterns of life” that are abnormal to traffic flows, and this could present issues when performing analysis.

## Citations

This investigation is informed by previous work in the field:

* WSDOT’s past research on freeway congestion prediction using pattern recognition and time-series modeling (WSDOT Research Report No. 381.1, *Freeway Congestion Prediction*).
* Academic and industry work, such as the JamBayes and Clearflow projects by Microsoft Research, which focused on learning and forecasting traffic flow in the Seattle area.
* The general literature on traffic flow prediction using machine learning, particularly the comparison of statistical, machine learning, and deep learning methods (LSTM/GRU) on time-series traffic data.

## Group Dynamics

* **Communication:** Our primary method of communication will be email for formal updates and scheduling, and a group channel on Microsoft Teams or Discord or something similar, informal discussions, and status updates.
* **Data and Code Sharing:** We will use a GitHub repository to share all code (Jupyter Notebooks, Python scripts), documentation, and data samples. Depending on the size and format of raw data files, alternative locations might be used for storage.
* **Periodic Meetings:** **We expect to have regular check-ins (formal time to be decided and scheduled in the next week), reviews of the past week’s progress, and assignment of next steps, which will be dependent on work schedules, etc.**
* Next Step (Week 4):
  + **Cong Ho (Data Engineer):** Lead data acquisition: Download, parse, and store the WSDOT Traffic Data and initial SDOT Collision data.
  + **Hannah Belville (Data Analyst):** Lead exploratory data analysis (EDA): Visualize traffic speed patterns and identify missing value percentages.
  + **Paul Skentzos (ML Specialist):** Lead literature review and analytical setup: Research best practices for time-series feature engineering and select initial hyperparameter ranges for the baseline model.