An Interactive System for Spatiotemporal Prediction and Visualization of US Traffic Delays

Bryan Cheung Daniel Do Jason Harris Mark Nassar Keegan Valerio Rui-Jia Zhang

Summary

Traffic delays caused by accidents lead to **socio-economic inefficiencies** including (but not limited to) unreliable travel times, spillover to secondary routes, and increased risk of additional accidents. Building an **understanding of factors that influence accident traffic delay severity** can inform the public, policy formation, and design. The objective of this application is to use historic data collected from US traffic accidents to **predict the impact on traffic flow.** Additionally, we aim to provide an **interactive user interface** to facilitate making these predictions, while simultaneously allowing a user to further **investigate past traffic incidents** in both space and time.

Data

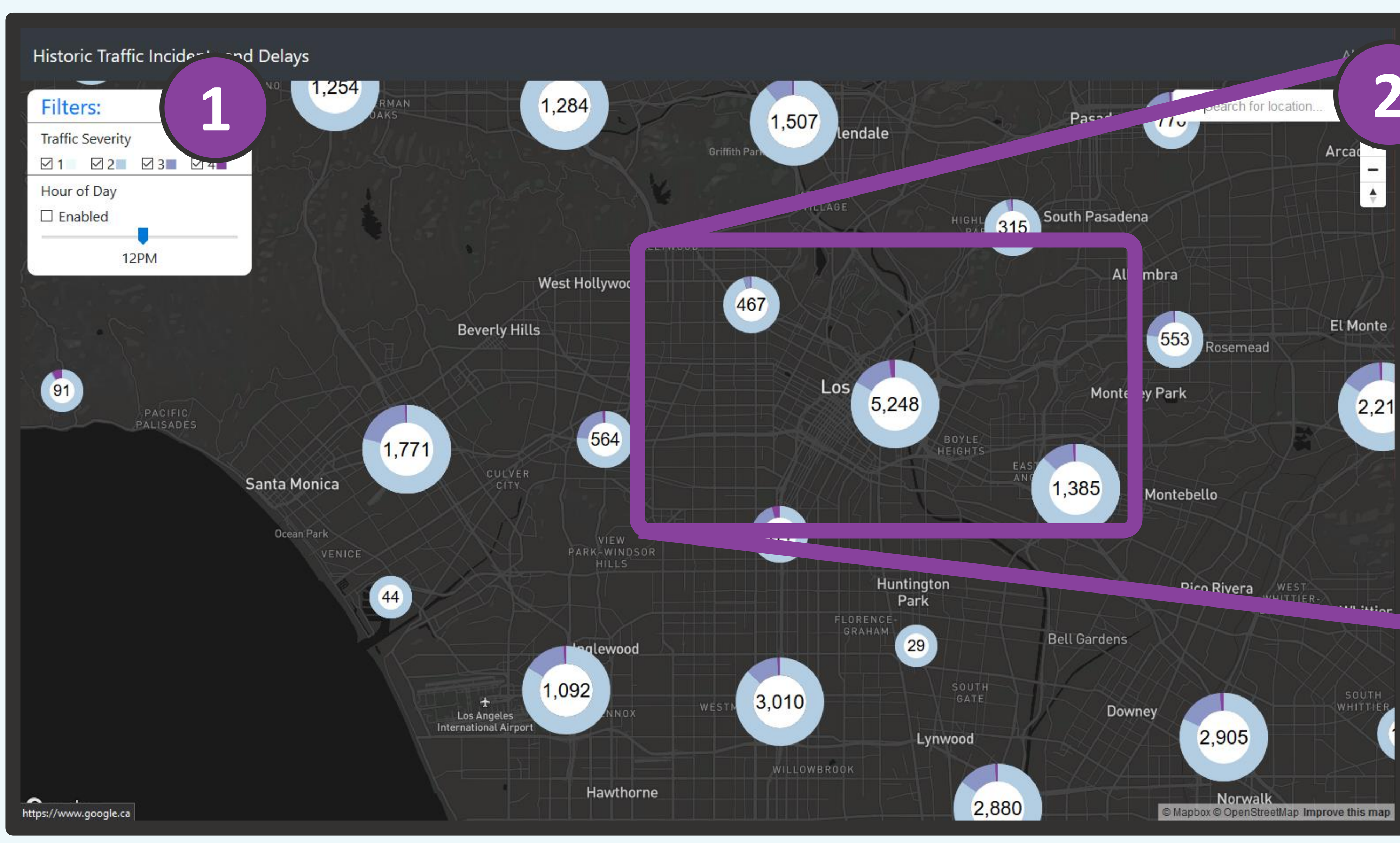
Our analysis is using the "US Accidents" dataset, created by Sobhan Moosavi, **downloaded** from Kaggle. This is a 1.24 GB csv file which contains 49 attributes for **3.5 million traffic accidents** from February 2016 to June 2020 in the United States.

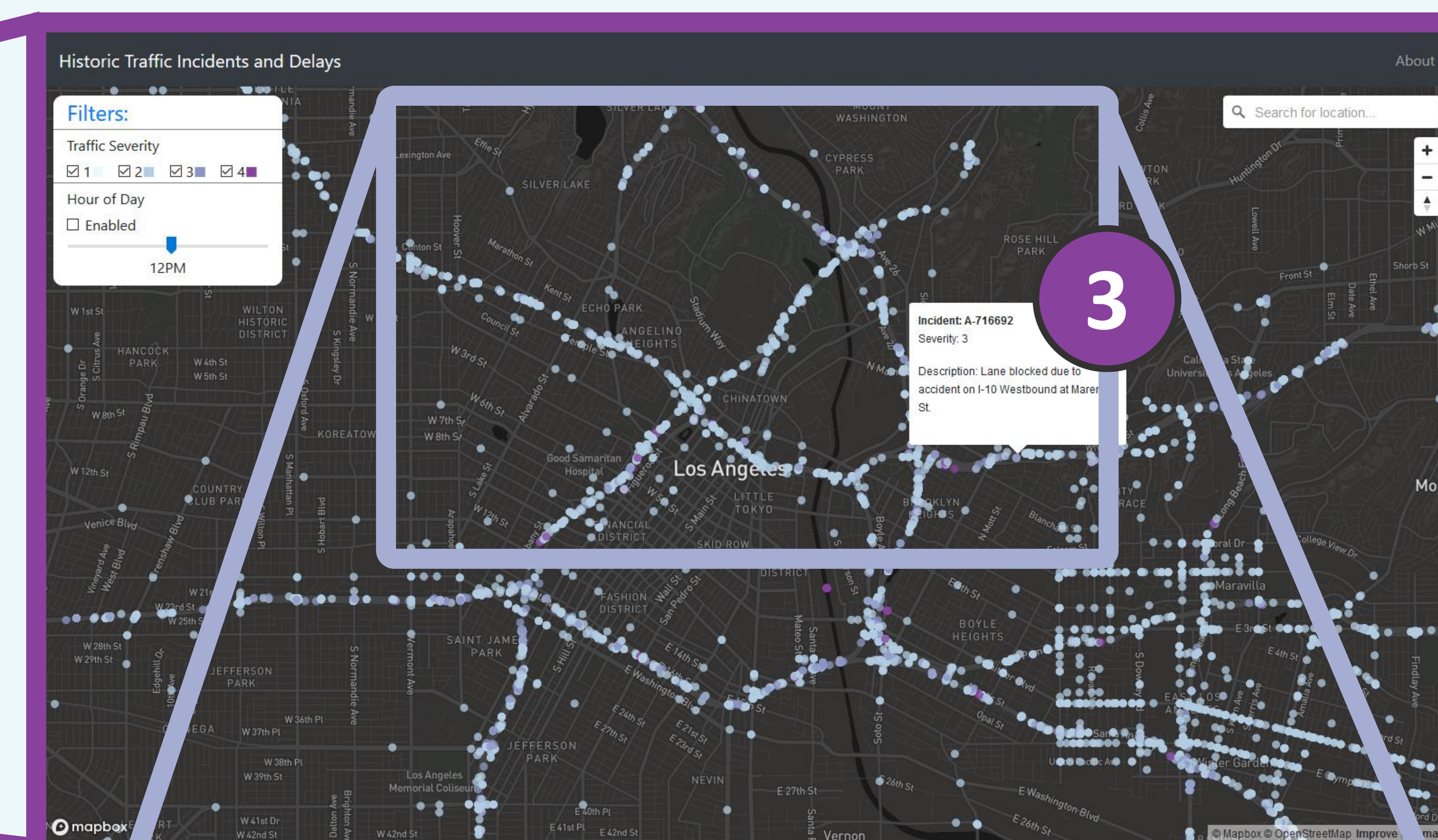
Current Methods

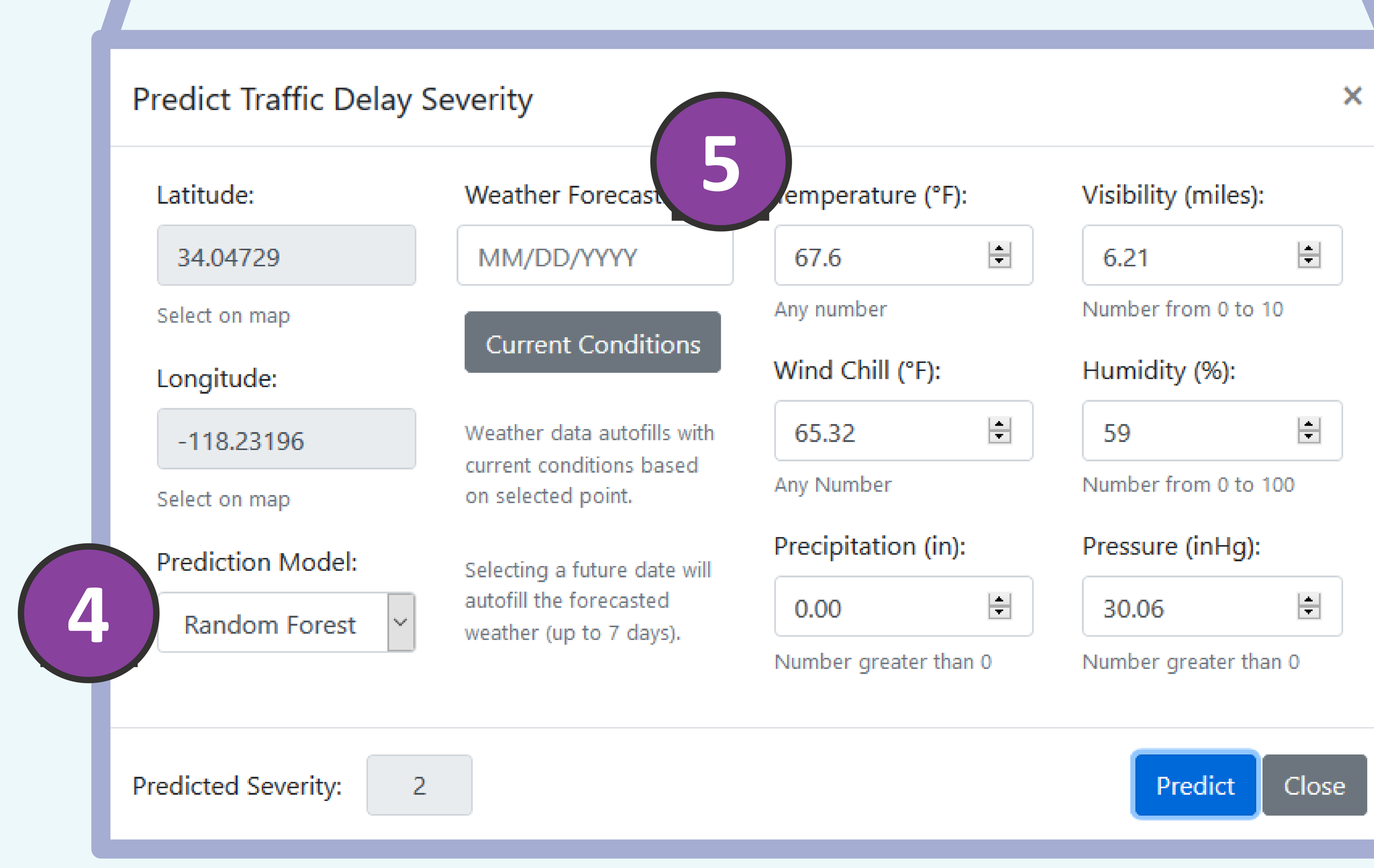
The NHTSA and Ford have applications to visualize traffic accidents; however, these are limited in detail and they **do not feature any prediction.** Our methods allow for an **interactive visualization** as well as the ability to predict traffic accident delay severity.

Data Balancing

Data imbalances were addressed before modeling. Severity class 2/3 comprise roughly 99.99% of the dataset, but severity 1/4 only occurs in roughly **1 in 10,000** records. We tested cost-sensitive and classifier-specific solutions that tweak the algorithm itself. We finalized using the over and und under-sampling technique **SMOTE-ENN** algorithm which essentially creates synthetic examples of the minority classes to increase the chances of predictions of the minority classes. After balancing the original dataset, the dataset increased to roughly **7 million records**.







This visualization displays traffic accident delay severity on an interactive map made using HTML, JavaScript, and the mapping library Mapbox. **Left figure:** At wide zoom levels, donut charts are displayed to visualize grouped historic accidents and their severity**. Top right figure:** Narrow zoom levels display individual accidents, color-coded to severity. **Bottom right figure:** Clicking anywhere on the map opens a popup which allows a user to predict severity at the selected location. **Additional functionalities include:**

- **[1]** Use the filter box in the top-left corner to adjust the visibility of historic data points based on severity or hour-of-occurrence
- **[2]** Use the search bar in the top-right corner to fly to any location in the continental US
- **[3]** Hovering over individual data points pops up additional details about the selected incident
- **[4]** Use a drop-down to select the modeling algorithm used to make predictions
- **[5]** Ability to prefill weather conditions with current and forecasted data

Modeling Traffic Delay Severity

The prediction of severity distills down to a classification problem. The team **experimented with numerous algorithms** utilizing a single-machine. Given the dataset's size, some algorithms could not complete in a timely fashion without moving to a parallelized platform such as Spark. The final models that are utilized were:

- **K-Nearest Neighbors -** This algorithm was chosen as a starting point since it is simple and easy-to-implement.
- **Simple Decision Tree** – Intuitive, easy-to-explain, requires less data preparation compared to other algorithms.
- **Random Forest** - Ensemble method consisting of weak learners (trees). Because of this, as more trees are added, it becomes more accurate, more stable, limits overfitting, and bias error present in individual trees.

Experiments and Results

Model:

The model experiments included manual tuning and cross-validated grid search tuning. Model performance metrics included ROC, AUC, and accuracy. Performance on balanced data yielded **accuracies of 59%, 71%, and 74%** for KNN, Decision Tree, and Random Forest models, respectively. Rather than a model competition, the visualization allows the user to **select the prediction model.**

Visualization:

The visualization was evaluated based on user testing. **A script was created** which utilized all of the visualization functionalities and the **time to complete** the script was recorded. The **intuitiveness of the UI** was also recorded for each user (e.g. clicked on something that is not a button). Adjustments were made between each user test. On average, the **time spent on the script decreased with each test**, as shown below, as intuitiveness was improved.

There are methods that can visualize traffic accidents; however, **none exist that can predict the severity** of traffic accident delays.

Team 088 - The Multiarmed Bandits