# Analyzation of Traffic Patterns in Relation to Delays in Boston Public Transit

Cristian Mendivil
crme7282@colorado.edu

Lucas Lyon
luly2738@colorado.edu

Vamshi Arugonda
vaar2387@colorado.edu

Theodore Freeman
thfr5570@colorado.edu

## Abstract

Do delays in public transit times correlate with increased travel times for uber? What are the most influential weather patterns affecting bus travel time?

{Brief summary of result}

## 1. Introduction

The authors of this paper will layout the methods and sources of data for an analysis of traffic, ride sharing, and public transportation in the Boston metro area. Data will be sources from 2016 to present, and most of the data will serve to provide a baseline for normal travel times between various sectors of the city. This paper will go into the application of knowledge gained, previous work that has been done on the data, details about the data sets and where to download them, methods of evaluation, data mining tools, temporal-based milestones, and a summary of peer review that was received in class.

## 2. Related Work

There has been quite a bit of prior work about traffic prediction and analysis. One such study was done by a group of researchers at the University of Southern California with a goal of accurately predicting and quantifying impact of traffic incidents. This is a pretty good study as in their conclusion they claim that their model can increase "prediction accuracy of baseline approaches by up to 45% [1]" for the impact of traffic incidents on road networks. We have still yet to read through the whole paper, but get the feeling that it will be a valuable source for inspiration for where we can take our project, and has it is own references and prior work which we can also look through and possibly utilize. Another group did some work on developing a support system for using real time bus location data to accurately estimate arrival times. This study may be useful considering that all the data we intend on using is public transit data or Uber. Perhaps it can give us some ideas of how to use our public transit data in a cleverer way. A way in which our project will be different from the described research above is in a couple of ways. First, the most recent of these projects was done in 2016 so there is potential at least to have more currently relevant results. Second our work is going to try to learn how individual traffic delays affect city-wide transit rather than just providing time estimates for when the next bus will arrive or route prediction for obstruction avoidance. Both projects may be useful to us though by providing different ideas for how to use and view our data as well as what we might avoid. If we find that we are getting stuck in a corner though and neither of these are able to help get us out it seems there is plenty of

[1]B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, “Forecasting Spatiotemporal Impact of Traffic Incidents on Road Networks,” in 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 587–596.

other research out there which if we searched for we may be able to find our answers.

## 3. Data Set

Our team will utilize two data sets: one from Uber, and one from MBTA. The Uber dataset must be downloaded in quarter-year increments from *https://movement.uber.com/*. The MBTA dataset can be downloaded in one file from their dashboard at *http://www.mbtabackOntrack.com/performance/index.html\#/download*. You must select the radio box "Reliability". Our team had difficulty downloading the entire dataset at once, and had to split the download into three time frames: January 1st 2016 - January 1st 2017, January 2nd 2017 - January 1st 2018, and January 2nd 2018 - March 5th 2018.

The MBTA Reliability dataset has 395,130 rows and 9 attributes. The attributes include service date and time, whether the row is for Off-Peak service or Peak service times, the type of transport (including rail, commuter, and bus), the route line, stop station, metric measured (including Passenger Wait Time and Schedule Adherence), and varying numerators and denominators for those metrics.

We will primarily use the service date and time, type of transport, stop station, metric type, and numerator and denominator attributes. The Peak vs Off-Peak hours attribute is not very helpful for establishing a link between delays in public transit and traffic, but will be helpful to establish two baselines for normal wait times, one during peak hours and one during off-peak hours.

The Uber Movement dataset is split into 7 distinct .csv files, each containing 3 months worth of travel times between every Uber-defined source and destination in Boston. Each file has 7 attributes: sourceid, dstid, hod, meanTravelTime, standardDeviationTravelTime, geometricMeanTravelTime, and geometricStandardDeviationTravelTime. Every attribute in the file will be useful for building a model of baseline travel times between different sectors of the city.

## 4. Main Techniques Applied

[TODO]

## 5. Key Results

[TODO]

## 6. Applications

[TODO]