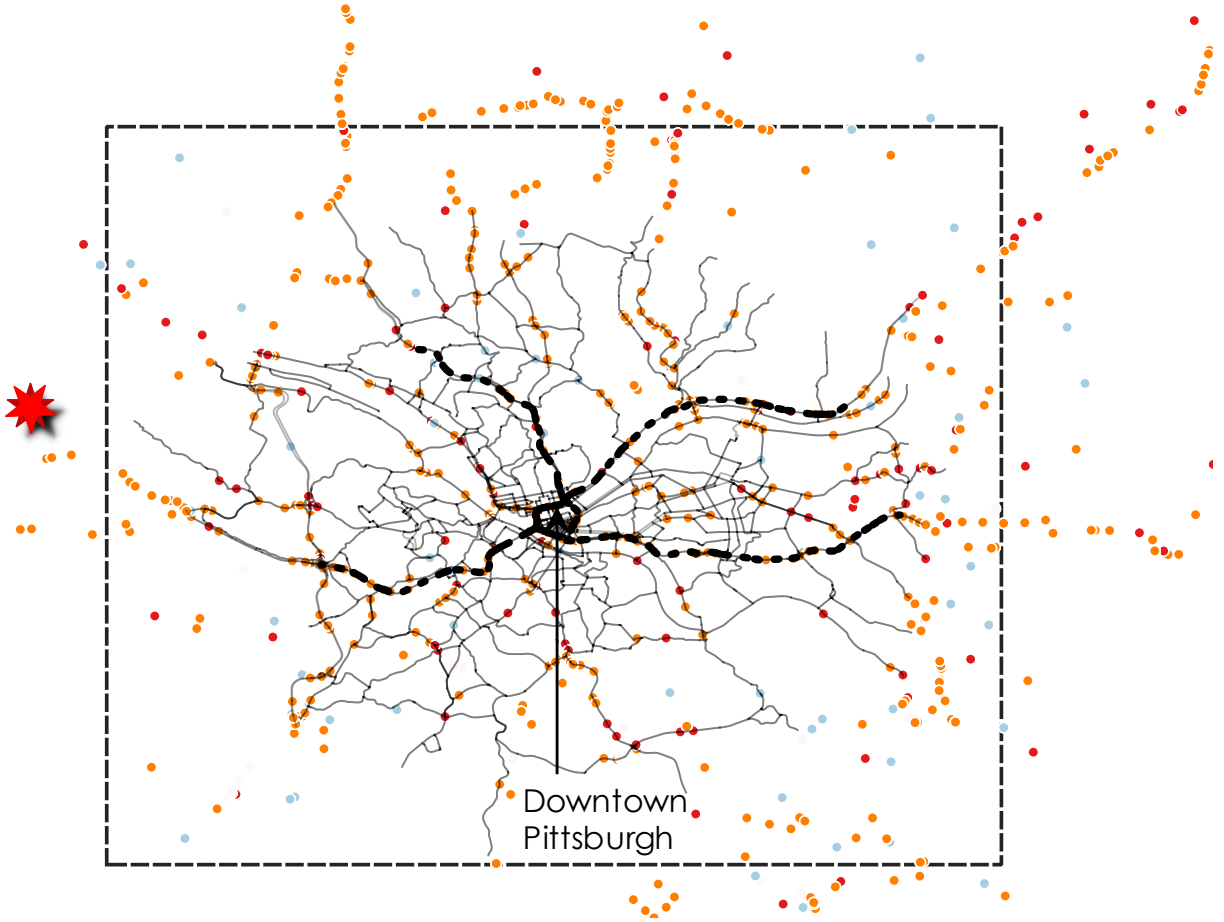
From Twitter to Traffic Predictor: Next-Day Morning Traffic Prediction Using Social Media Data

Weiran Yao
PhD student
Carnegie Mellon University

Sean Qian
Associate Professor
Carnegie Mellon University



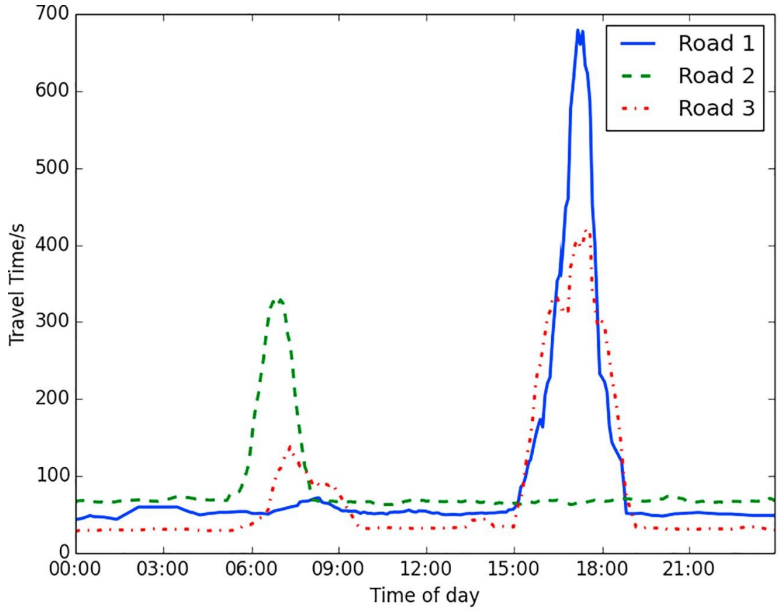
2020 TRB Annual Meeting @ Washington D.C.
New Solutions to Traffic Monitoring Challenges

Funding to attend this conference was provided by the CMU GSA/Provost Conference Funding

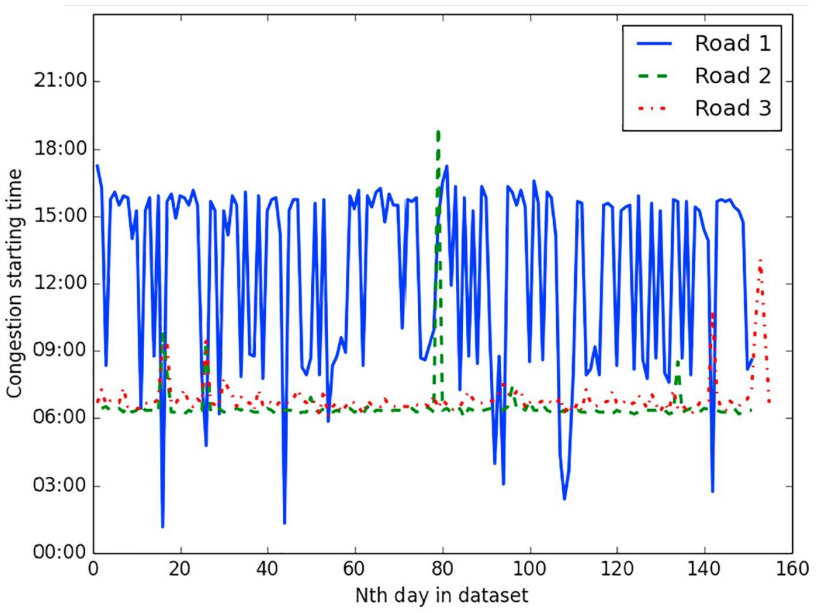
Challenges of predicting next-day morning traffic

- **Definition:** Predict morning traffic before early morning or even earlier
- **Background:** 13% of the population commute before 6 am; 4.4 % by 5 am (American Community Survey, 2015)
- **Motivation:** To provide **travel information**, traffic prediction of morning rush hour traffic before early morning (e.g. 5 am) is needed.

- **Challenge:** However, **real-time** and **historical traffic** are **not helpful:**



(a)



(b)

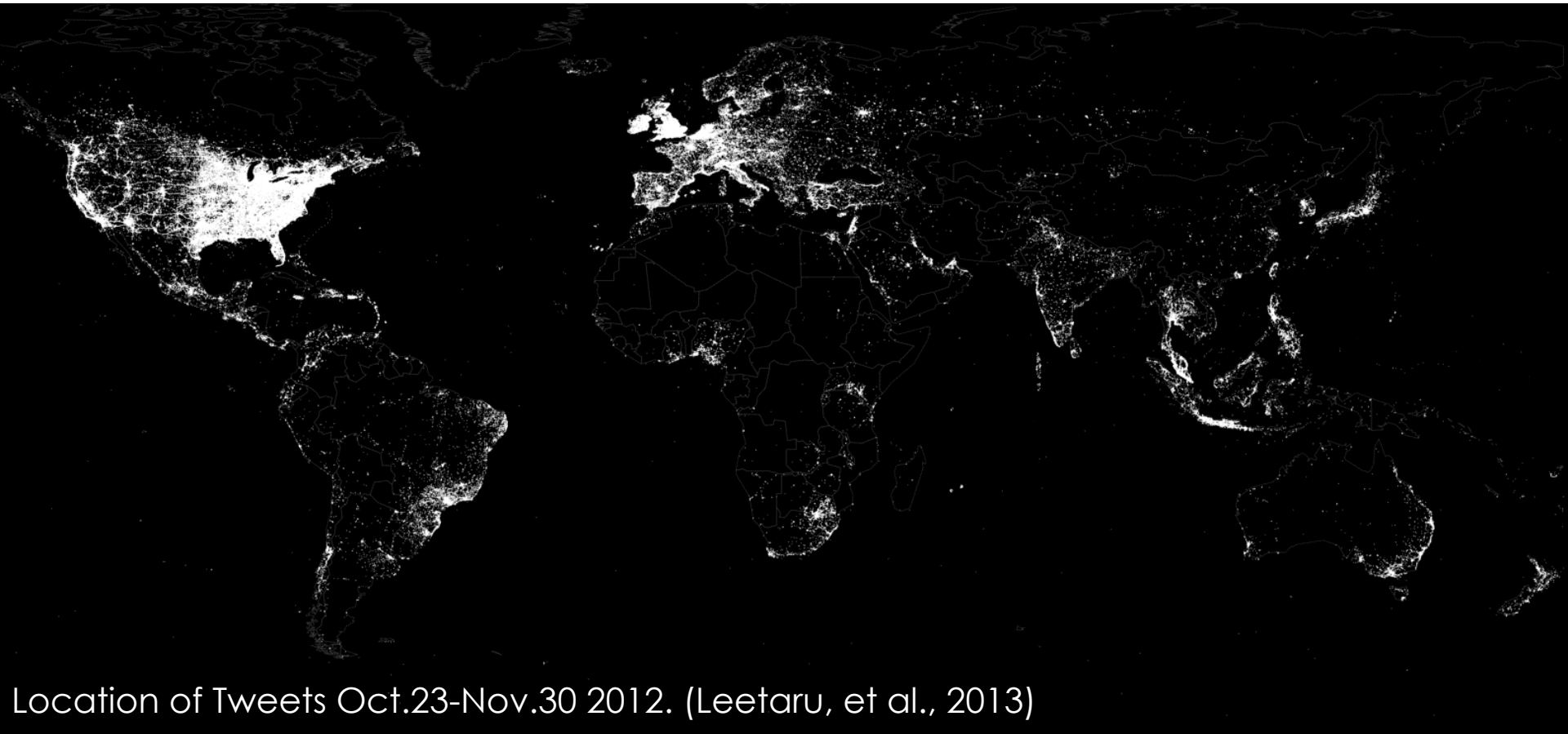
(a) Traffic **breakdown suddenly**

(b) Day-to-day **variance** is high

(Zhang and Qian, 2019)

Opportunities

- **Travel demand** on each day (departure time, mode, etc.) may be explained by commuters' activities at midnight or early in the morning;
- The rise of social media and analytics offer new tools to **sensing crowd activities** during late night and early morning;
- Rarely can other public or open-source crowdsourcing data collect such "private" data in large scale and on daily basis.



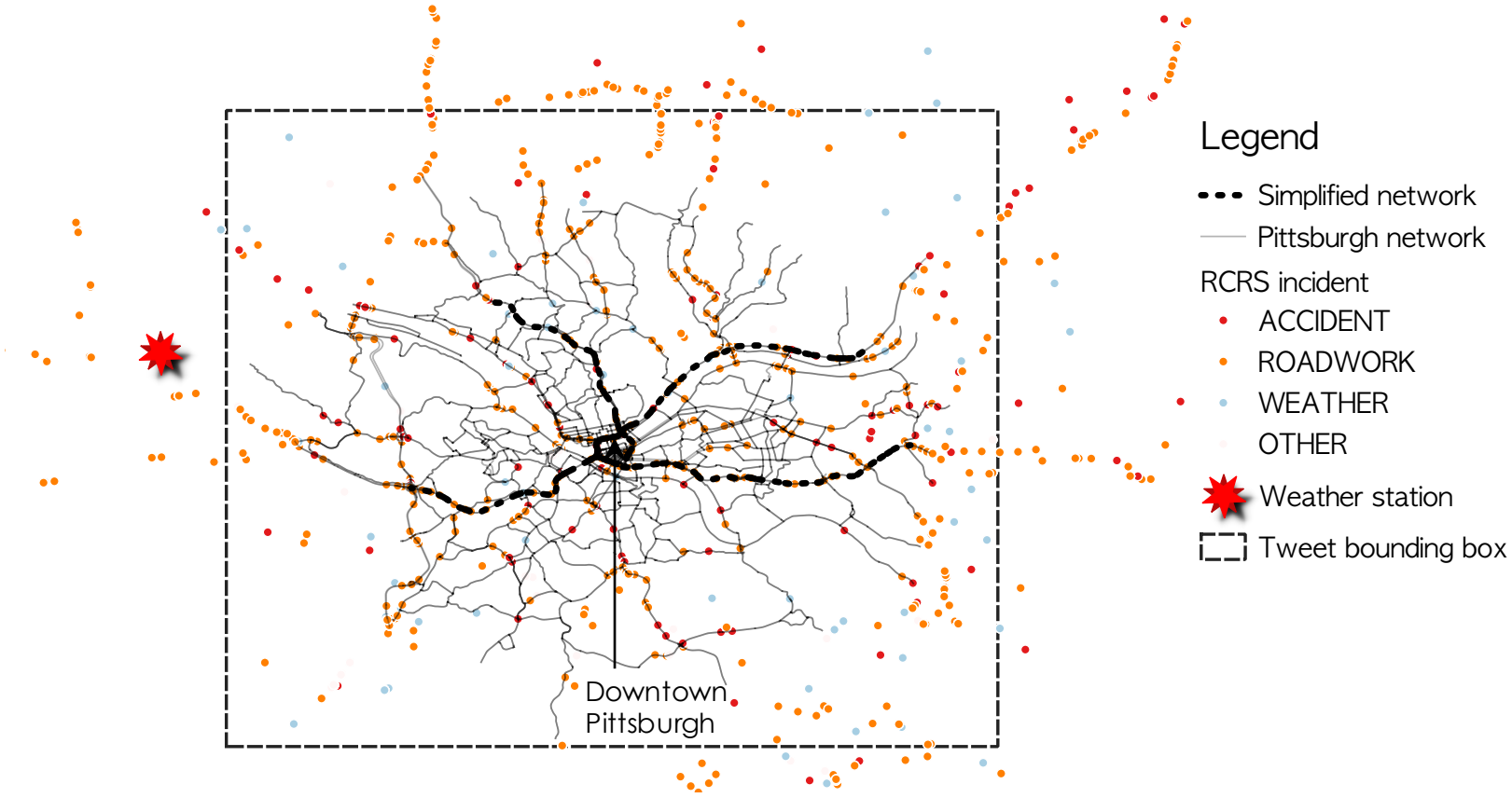
Location of Tweets Oct.23-Nov.30 2012. (Leetaru, et al., 2013)

Research objectives

- Predict morning peak hour traffic with demand (tweet, temporal, etc.) and supply (traffic incidents, weather, etc.) side features available before early morning (i.e., before 5 am, 3 am or 0 am).
  - ❑ Characterize spatiotemporal morning road traffic patterns;
  - ❑ Extract meaningful features which explain morning traffic;
  - ❑ Develop interpretable predictive frameworks which utilize spatiotemporal road traffic patterns

- Explain how user's tweeting activities in the late night and early morning impact next-day morning traffic, while controlling for the effects of supply-side features.

Datasets collected in Pittsburgh metropolitan regions



- INRIX **traffic speed** (87 TMC segments) sampled every 5 min
  - Covers 4 main highways in Pittsburgh (I-376 W; PA-28S; I-376N, I-279S)
- PennDOT RCRS **incident** datasets (accident, roadwork, severe weather, etc.)
- **Weather** underground (hourly temp, humid, visibility, etc.)
- **Twitter** geocoded streaming data from Jan-Dec 2014

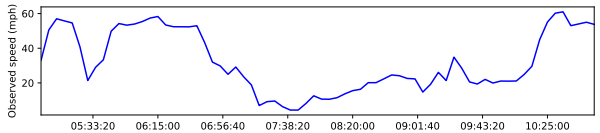
Traffic output: characterizing morning road traffic

- Traffic speed on segment is first normalized into Travel Time Index (TTI) using reference (free-flow) speed (85 percentile speed among all periods).
- PCA and K-Means clustering are performed on road congestion profiles to identify morning road traffic clusters.

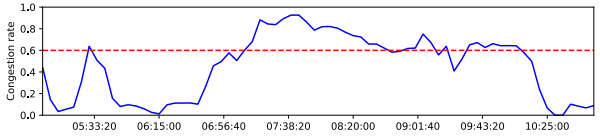
1. Speed processing

Traffic speed for each TMC segment during morning periods is recorded as one-dimensional time-series.

Speed curves



Travel time index (TTI) curves



2. Clustering analysis

Spatiotemporal congestion patterns are characterized road by road. Congestion rate data on all segments on a recurrent congested road are used to build the road's daily congestion profiles and to identify typical patterns.

Segment TTI

TMC (order 1)

TMC (order 2)

TMC (order n)

$$\begin{matrix} TTI_{n1}^{1} & TTI_{n2}^{1} & \dots & TTI_{nt}^{1} \\ TTI_{n1}^{2} & TTI & \dots & TTI_{nt}^{2} \\ \vdots & \vdots & \ddots & \vdots \\ TTI_{n1}^{d} & TTI_{n2}^{d} & \dots & TTI_{n1}^{d} \end{matrix}$$

Road congestion profiles

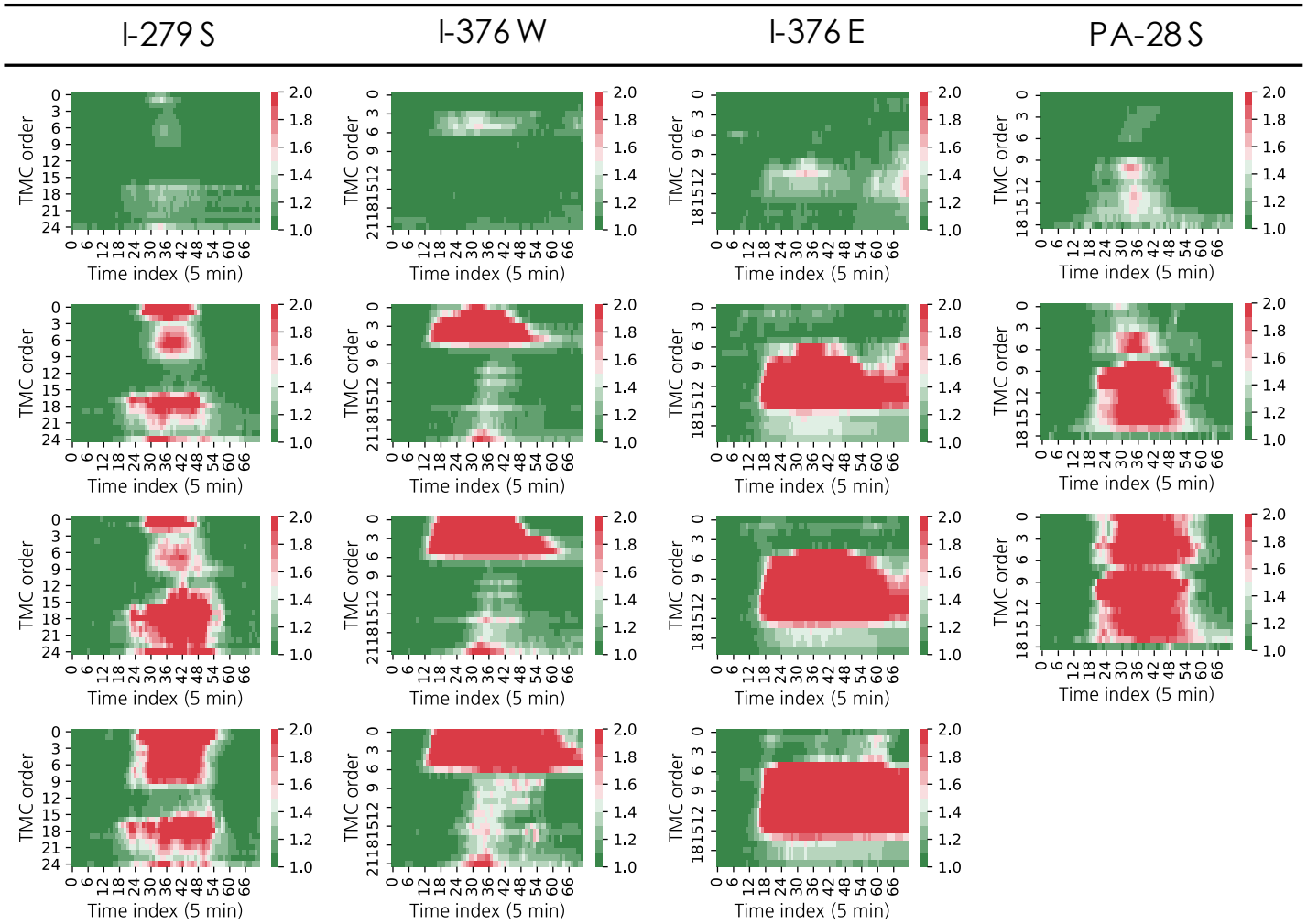
Segment TTI curves are concatenated to construct road congestion profiles.

$$\begin{matrix} TTI_{11}^{1} & TTI_{12}^{1} & \cdots & TTI_{1t}^{1} & TTI_{21}^{1} & \cdots & TTI_{31}^{1} & \cdots & TTI_{nt}^{1} \\ TTI_{11}^{2} & TTI_{12}^{2} & \dots & TTI_{1t}^{2} & TTI_{21}^{2} & \cdots & TTI_{31}^{2} & \dots & TTI_{nt}^{2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ TTI_{11}^{d} & TTI_{12}^{d} & \dots & TTI_{1t}^{d} & TTI_{21}^{d} & \dots & TTI_{31}^{d} & \dots & TTI_{nt}^{d} \end{matrix}$$

K-means clusters

PCA is first used for dimension reduction. K-means clustering is performed on reduced congestion profiles to extract spatiotemporal congestion patterns.

Traffic output: characterizing morning road traffic



- Because of spillback effects, morning road traffic shows **ordered** spatiotemporal clustering patterns

Traffic output: characterizing morning road traffic

The morning traffic consists of 72 data points of TMC speed measured every 5 minutes. We propose to describe morning traffic with **3 variables**, including:

- congestion status (S): a road segment is defined as congested if the observed travel time index is greater than 2 for at least 15 minutes.

- congestion starting time (CST): denotes the starting point of the congested period, and if multiple congested periods exist, CST is the starting point of the earliest period.
- congestion duration (CD): measures the length of congestion and is thus defined as the interval between the first congestion starting point and the last congestion ending point.

Traffic output representation is **easier to predict**, while **sufficient as** ahead-of-time **travel information** for morning commuters.

Tweet processing pipeline

We claim that tweets capture **three types of information** which can explain next-day morning traffic variances.

- **Sleep-wake pulses:** tweeting activities in late night and early morning capture the sleep-wake patterns in urban districts, and thus explain next-day morning traffic;

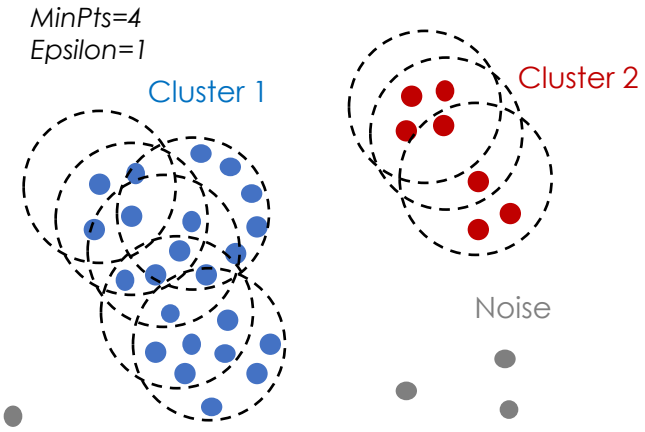
- **Event indicators:** geocoded tweet counts and sentiment, aggregated by space and time, have been proved in many prior works as an effective indicator for events and holidays, which impact next-day traffic;

- **(Planned) traffic incidents:** Twitter accounts owned by public traffic agencies and media (e.g., @511PAPittsburgh) automatically report real-time traffic incidents (e.g., crashes and planned roadworks, etc.), which have impacts on next-day morning traffic.

Tweet processing pipeline: timeline activity augmenter

Geocoded tweets are sparse because the sampled users change every day, adding noises to the estimates of daily sleep-wake variances. User timeline tweets (non-geocoded tweets, favorites, retweets) are retrieved to augment geocoded tweet activities.

1. **Resident filtering:** user profiles are used to identify local residents of Pittsburgh, with carefully-implemented Regular Expressions (REs).
2. **Home location inference:** (1) DBSCAN to identify user's frequently visited places; (2) a rule-based classifier is tuned to locate user's homes from coordinate clusters.
3. **Geotagging:** The missing tweeting coordinates between late night and early morning (9 pm - 5 am), with their inferred home locations by assuming residents sleep at home.



Tweet processing pipeline: cleaner+geocoder (user tweets)

Cleaner

- Spam remover: *Botometer API* (Davis et al, 2016) is applied to find spam users with features extracted from *available account meta-data*, *following patterns*, and *content scraped from 200 most recent tweets*

- Tweet content cleaner:

1. Lower the text content;
2. Remove special tweet entities (e.g. urls, emojis, email addresses, phone numbers, user names, etc.);
3. Segment hashtags to words and remove \#, e.g., \#LetsGoPens -- lets go pens;
4. Concatenate consecutive (>3) single-character tokens, e.g. Ain't H A P P Y -- ain't happy;
5. Remove repeated suffix, e.g., Soooo goood lololol...-- so good lol;
6. Translate slangs to formal words with a slang dictionary, e.g. lol -- laughing out loud;
7. Remove special characters and brackets, e.g., *([)]\&=;
8. Strip and remove extra whitespaces;
9. Add ending mark ``.'' to unfinished sentences and fill empty tweets with an ending mark.

Geocoder: Individual tweets are spatially joined with census tracts on their posting coordinates.

Tweet processing pipeline: cleaner+geocoder (incident)

Cleaner

- @511PAPittsburgh reports real-time traffic incident status in Southwest PA. The account with a series of computer-generated tweets:

  2019-12-27 06:42 -- Multi vehicle crash on I-376 eastbound at Mile Post: 74.0. There is a lane restriction
  2019-12-27 07:18 -- UPDATE: Multi vehicle crash on I-376 eastbound at Mile Post: 74.0. All lanes closed
  2019-12-27 08:02 -- CLEARED: Multi vehicle crash on I-376 eastbound at Mile Post: 74.0.

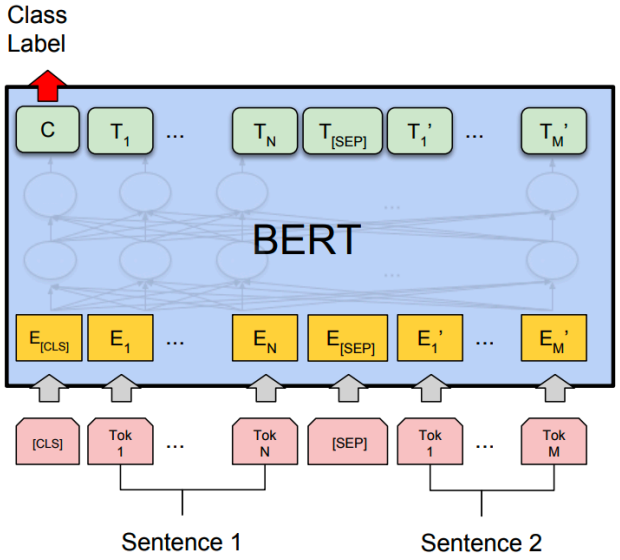
- Considering the fixed content format, we carefully implement Regular Expressions (REs) to extract:
  - ❑ Highway or road names (e.g., I-376)
  - ❑ Direction (e.g., eastbound)
  - ❑ Exit or mile post (e.g., 74.0)
  - ❑ Incident type (e.g., crash)
  - ❑ Lane status (e.g., lane restriction -- full closure -- open)
  - ❑ Tweet flag (e.g., occur -- update -- clear) from each incident tweet.

Geocoder: we applied the GIS developed in (Gu et al, 2016) to translate the parsed incident highway/road name and exit/mile post into incident latitude/longitude coordinates with a predefined dictionary.

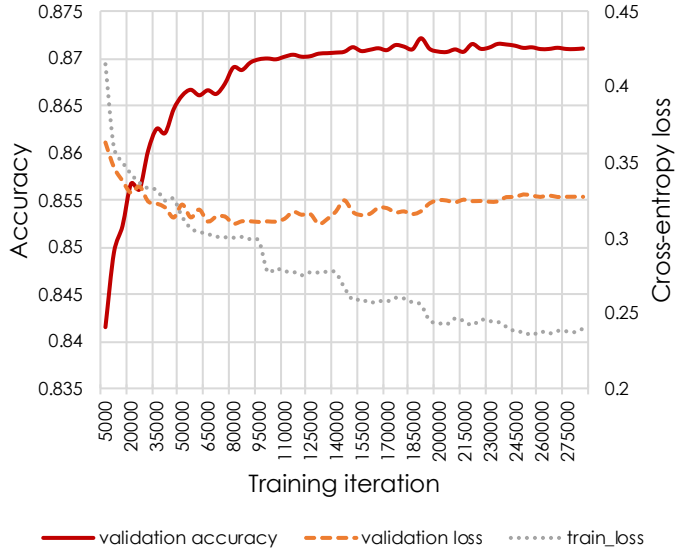
Tweet processing pipeline: encoder (user tweets)

- **Spatiotemporal sleep-wake pulses:** two histogram-vectors describe the distributions of sleeping and waking up times of influential tweet users with the last augmented tweets residents post between 9 pm-3 am and the first tweets they post between 3 am-5 am.

- **Event indicators:** The number of geocoded tweets and the percentage of neutral tweets (trained with BERT) in the six pre-defined periods, among all sentiment categories, are also encoded as event indicators.



(Delvin, et al, 2019)



Tweet processing pipeline: encoder (incident tweets)

2019-12-27 06:42 -- Multi vehicle crash on I-376 eastbound at Mile Post: 74.0. There is a lane restriction
2019-12-27 07:18 -- UPDATE: Multi vehicle crash on I-376 eastbound at Mile Post: 74.0. All lanes closed
2019-12-27 08:02 -- CLEARED: Multi vehicle crash on I-376 eastbound at Mile Post: 74.0

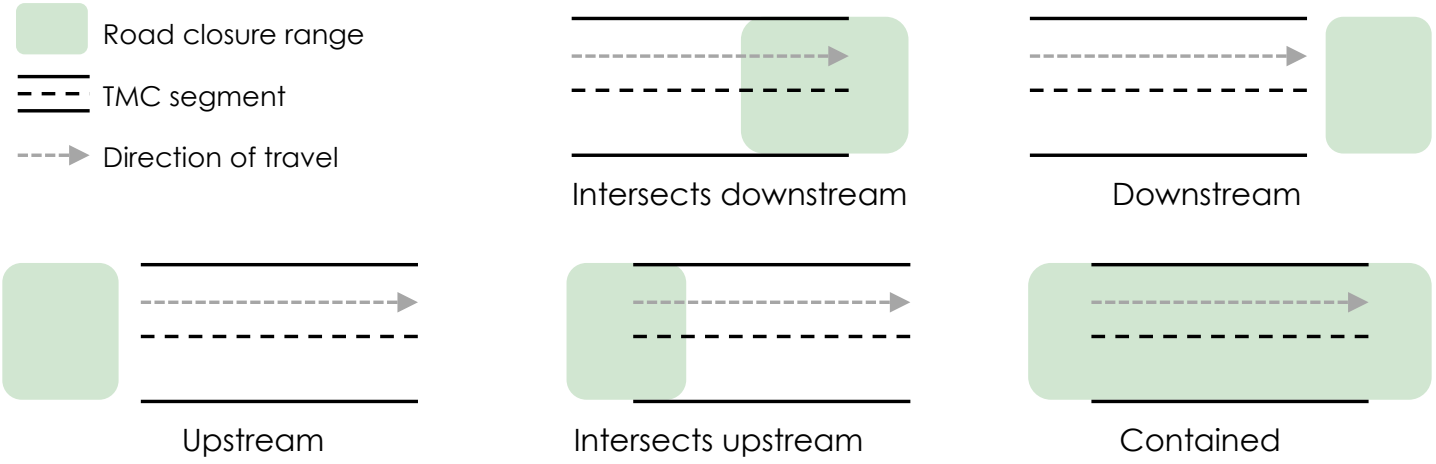
To integrate with PennDOT RCRS incident dataset, the same set of features. are extracted, including:

- **Closure/open timestamps:** defined as the first and last tweeting timestamps of the series of tweets describing an incident
- **Highway/road names:** same format as RCRS
- **Incident start/end coordinates:** If two or more exits/mile posts appear in a tweet, incident start location is defined as the coordinates of the smallest exit/mile post and incident end location uses the coordinates of the largest exit/mile post. Otherwise, the incident start and end locations both use the coordinates of the incident.
- **lane closure type:** two types of lane closure, i.e., partial and full closure, are used to describe the severity of incident.

Traffic incident processing

Three aspects of incident impacts are considered: (1) lane closure types, (2) incident location and (3) incident time window:

- **Lane closure types:** 2 types of lane closure are encoded separately.
- **Incident location:** distance between incident start/end location and segment start/end location is mapped into vector [DS, C, US] to represent the five possible incident-segment location.
- **Incident time window:** the effective time-window of an incident



Weather and time features

Weather features

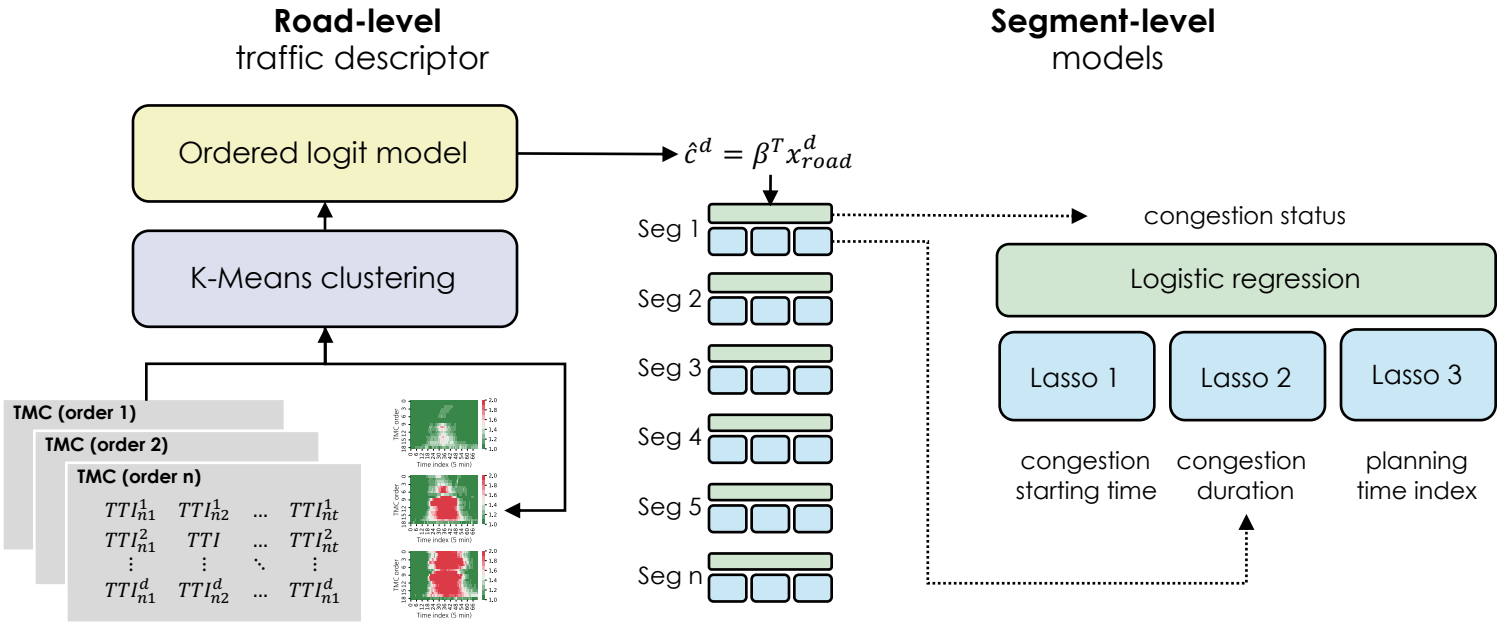
six continuous variables -- temperature, humidity, wind speed, pressure, visibility and hourly precipitation, and a binary variable -- pavement condition.

Time features

Time features include four categorical variables: week-of-year, month-of-year, day-of-week, and holiday.

- For the cyclic month and week of year variables, we use sine and cosine functions to transform them into clockwise encoded features.
- For day-of-week and holiday variables, we apply one-hot encoding after combining similar time features.

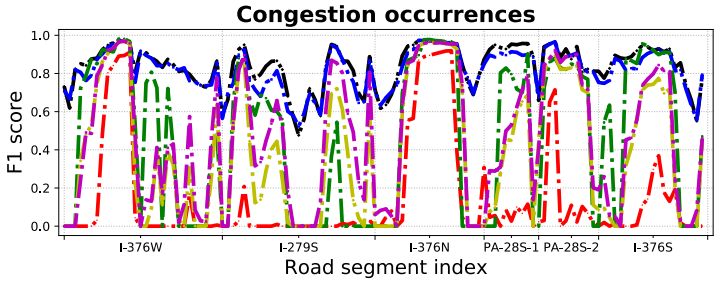
Multi-level predictive model architecture

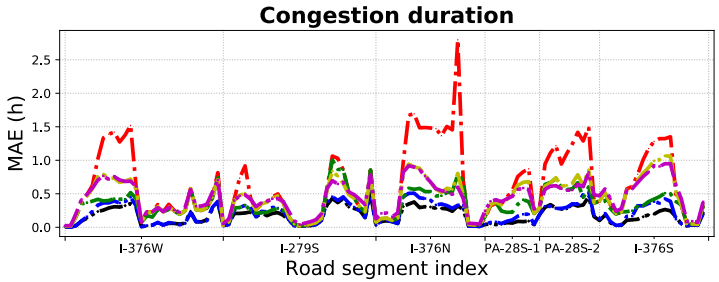


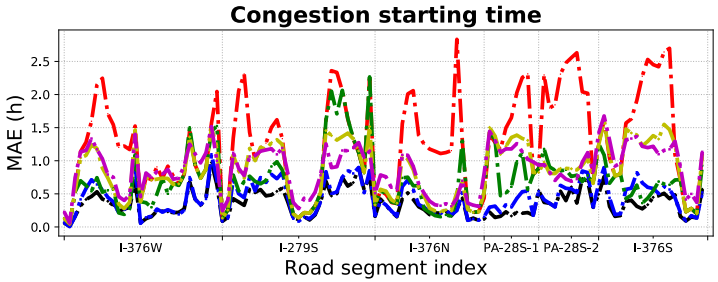
- **Road-level descriptor:** L2-regularized ordered logit regression trained to predict ordered road traffic cluster index. RFE is applied for feature selection to remove unrelated spatiotemporal features.

- **Segment-level classifier:** L1 regularized logistic regression is trained for each segment with **descriptor feature** included.
- **Segment-level regressor:** LASSO trained for each segment for predicting congestion starting time and duration with **descriptor feature** included.

Experiments: model prediction performances







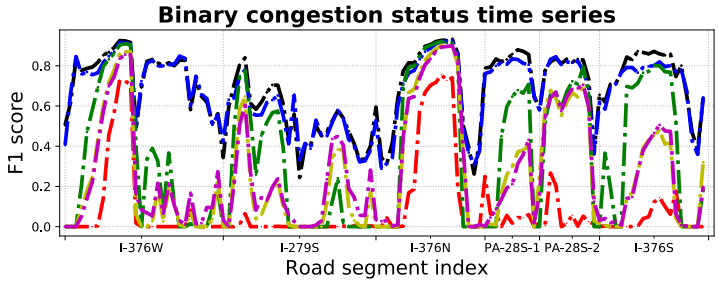
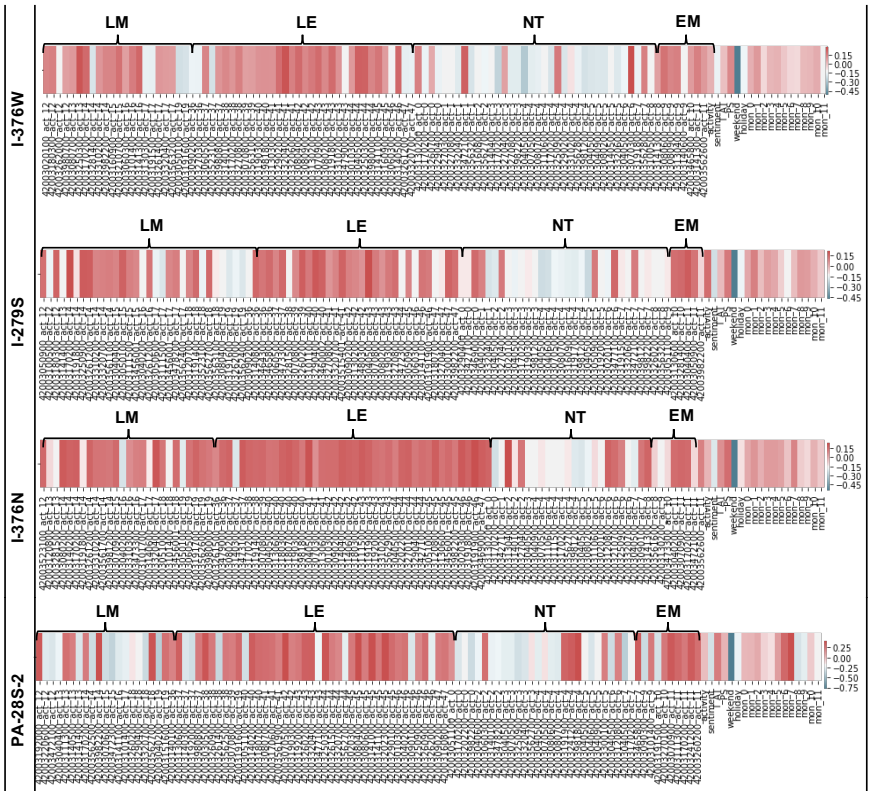


Table 5
Model comparison results by method.

Method	Congestion measurements							
	Occurrence (%)		Starting time (h)		Duration (h)		Binary time-series (%)	
	Accuracy	F1	RMSE	MAE	RMSE	MAE	Accuracy	F1
Logit+Lasso+Olm	0.909	0.838	0.989	0.353	0.399	0.172	0.954	0.691
Logit+Lasso	0.888	0.817	1.103	0.431	0.460	0.212	0.948	0.672
AR-MIMO	0.653	0.148	2.003	1.282	1.001	0.636	0.901	0.093
Olm+Kernel	0.809	0.459	1.398	0.694	0.653	0.343	0.944	0.359
HM	0.765	0.396	1.601	0.854	0.803	0.446	0.927	0.263
HM-M	0.782	0.461	1.359	0.881	0.725	0.451	0.902	0.275

Experiments – how tweets explain next-day traffic

- Visualizing the weights of road-level traffic descriptor



- The earlier people rest (+ in LE and – in NT), the more congested roads will be in the next morning.
- People's tweeting activities in the early morning (EM) are positively associated with congestion in morning peak hours.
- %Neu in LE is positively correlated with tomorrow's road congestion

- Most features are extracted from LE and NT, which can be extracted earlier than 5 am of the next-day.

Experiments – how advance can next-day traffic be predicted

- Model is insensitive to the increase of forecasting horizons till 0 am;
- From a practical perspective, the optimal implementation of our framework is to predict morning congestion at 0 am, which leaves enough time for decision making while keeping adequate prediction performances.

Table 6
Prediction errors with different forecasting horizons.

Congestion measurements		Forecasting horizons		
		6 hours (5 AM)	8 hours (3 AM)	11 hours (0 AM)
Occurrence (%)	Accuracy	0.909	0.912	0.909
	F1	0.838	0.843	0.839
Starting time (h)	RMSE	0.989	0.979	0.987
	MAE	0.353	0.348	0.356
Duration (h)	RMSE	0.399	0.400	0.401
	MAE	0.172	0.171	0.173
Binary time-series (%)	Accuracy	0.954	0.953	0.953
	F1	0.691	0.691	0.690

Takeaways

- We find that the earlier people rest, the more congested roads will be in the next morning. In addition, people's tweeting activities in the early morning are positively associated with congestion in morning peak hours
- People's tweeting activity features extracted before the start of morning periods at 5 am, or as late as the midnight before can explain morning traffic.
- Multi-level predictive architecture helps capture the spatiotemporal road traffic pattern and improves segment-level traffic prediction
- Evaluation studies support that our framework can accurately predict next-day morning **congestion starting time with an average error of 21 minutes** and **congestion duration with an average error of 10 minutes** using data provided before 0 am.

Thank you!
Questions or comments?