Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights

Sobhan Moosavi, Mohammad Hossein Samavatian,
Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath

Department of Computer Science and Engineering
The Ohio State University

ACM SIGSPATIAL 2019
Chicago, IL (November 5 – 8, 2019)

NSF Grant: EAR-1520870

OSC Grant: PAS0536

Traffic Accidents: Important Public Safety Challenges

- About 1.25 million traffic death in 2013
  - Based on the Global Status Report [28]
- In the United States*:
  - ~ 6 million accidents per year (officially reported)
  - ~ 2.3 million car accidents injuries or disabilities per year
  - ~ 37,000 traffic death per year
  - ~ $230 billion cost per year
- This is not even the whole picture!
  - About 10 million or more crashes go unreported
  - Source: National Highway Traffic Safety Administration

* Source: Association For Safe International Road Travel

Categories of Prior Research

- Analysis of Environmental Stimuli
  - Studying the impact of stimuli on the occurrence of accidents
  - Examples of stimuli: weather, traffic, and properties of road
  - Existing studies: *Eisenberg 2004; Jaroszweski and McNamara 2014; Mannering et al. 2016; Tamerius et al. 2016; Theofilatos 2017*
- Accident Frequency Prediction
  - Predicting the frequency of accidents within a place during a time interval
  - Existing studies: *Chang 2005, Caliendo et al. 2007, Najjar 2017, Yuan 2018*
- Accident Risk Prediction
  - Predicting the possibility of an accident within a place during a time interval
  - Existing studies: *Lin et al. 2015, Chen et al. 2016, Wenqi et al. 2016, Yuan et al. 2017*

Shortcomings of Literature

- Small-scaled datasets
- Expensive data sources
- Inapplicable for real-time purposes

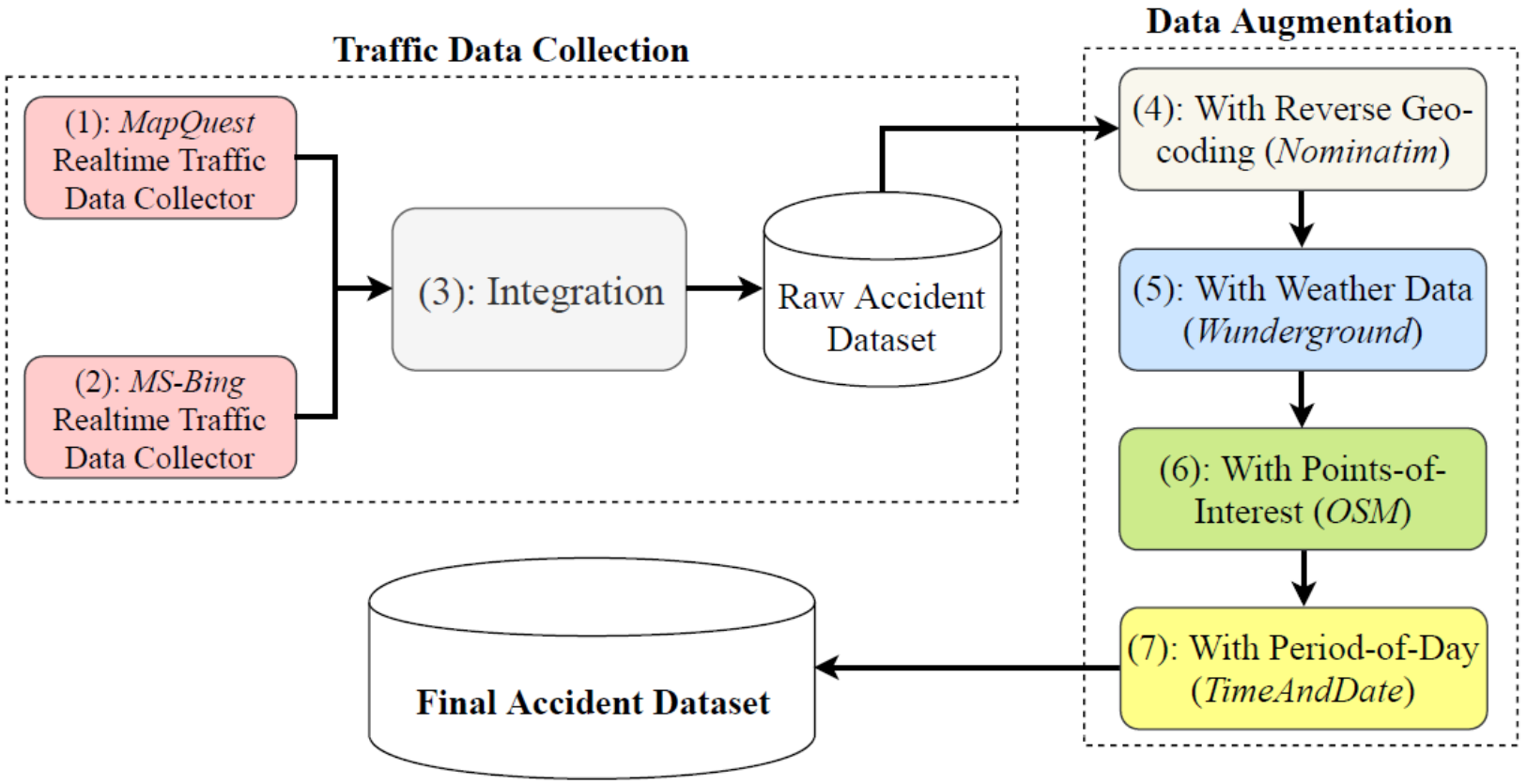
Important Contributions

- We proposed a new process to build a large-scale traffic accident dataset
  - Dataset is available at https://smoosavi.org/datasets/us_accidents
- We gleaned a variety of insights by analyzing the resulting dataset
- We proposed a new solution for real-time traffic accident prediction

Large-scale Traffic Accident Dataset

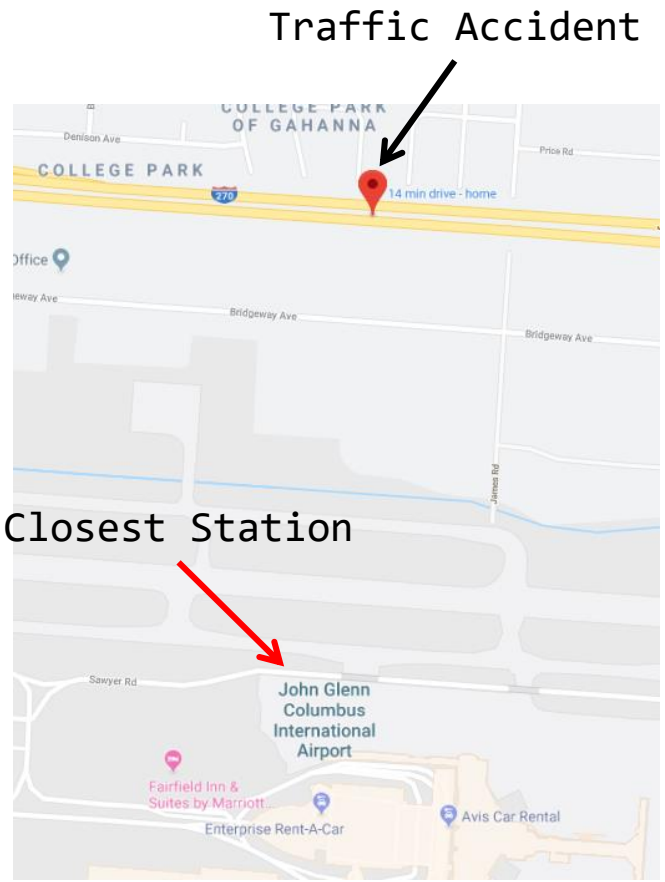
Process of Building the Dataset

- We propose a process to collect, augment, and publish a large-scale accident data



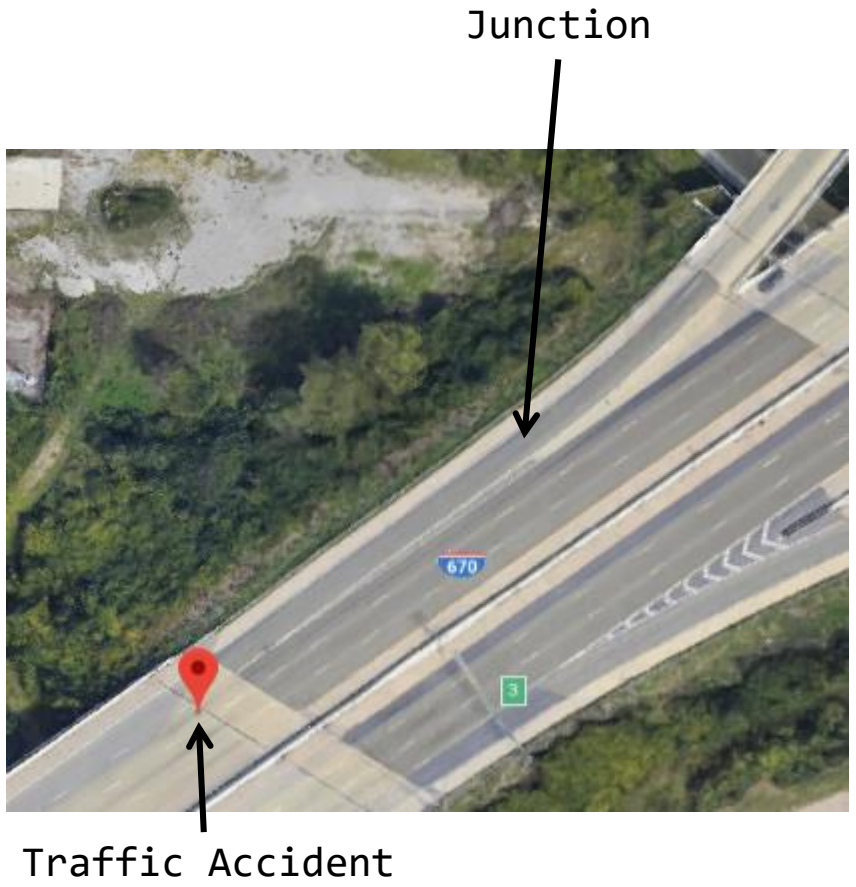
Data Augmentation By Weather

- Augmenting with *Weather* Data
  - Find the closest weather station
  - Find the closest sensor reading
  - Historical weather data was collected from *Weather Underground*



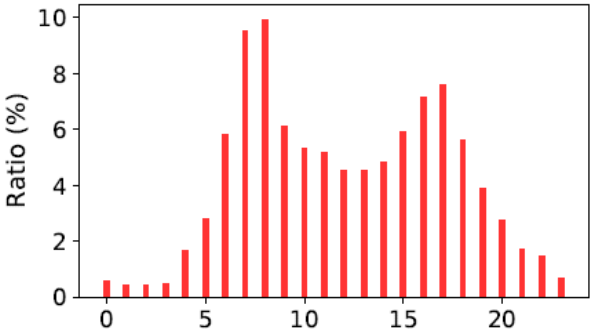
Data Augmentation By POI

- Augmentation with *Points-of-Interest*
  - POI: A location on the map with a *special* property (e.g., intersection, junction, and amenity)
  - Collected from Open Street Map (OSM)
  - How to be assigned? Using distance threshold
  - Employed a *data-driven process* to find the best distance threshold

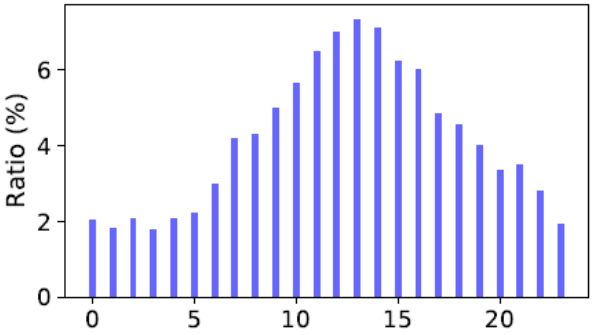


Final Dataset: US-Accidents (2016 – 2019)

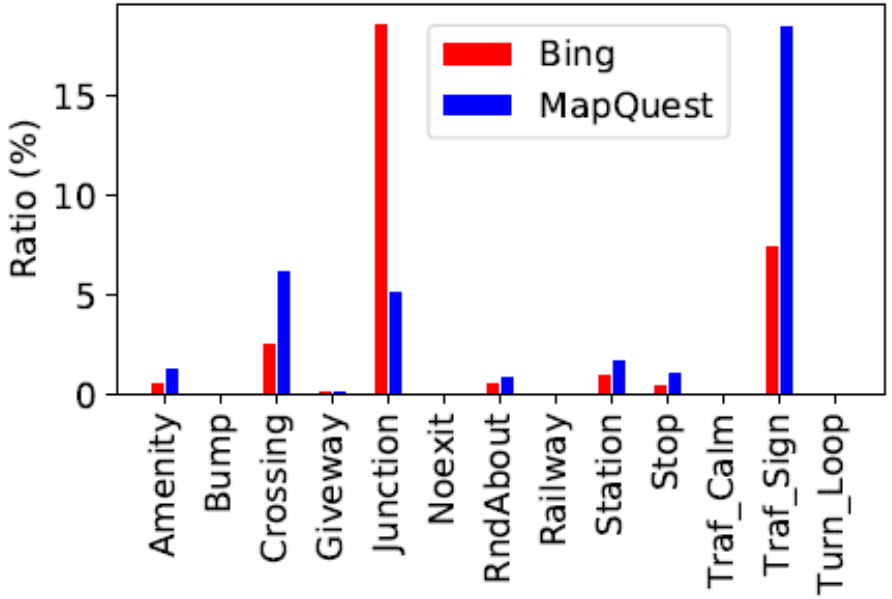
- Total accidents: 2.25 million
- Total attributes: 45
- Coverage: 49 states of the US

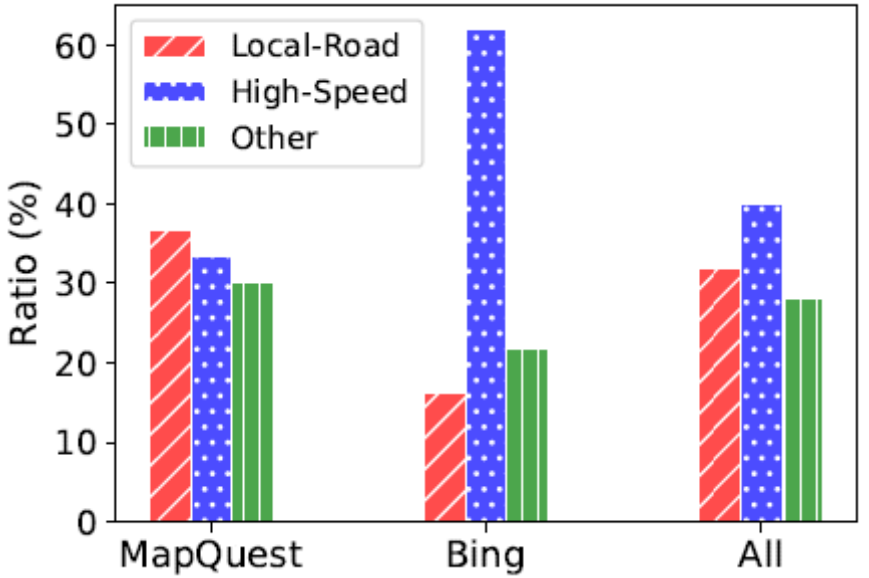


(b) Hour of Day (weekdays)



(c) Hour of Day (weekends)

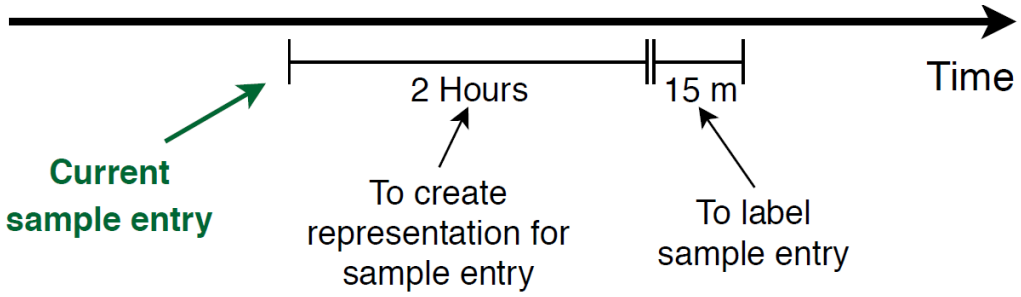




Accident Prediction Framework

Problem Statement

- Given
  - A spatial region R (size: 5km x 5km)
  - A database of traffic events E
  - A database of weather information W
  - A database of points-of-interest P



- Create
  - A representation F_{RT} for R during a time interval $T = 15$ minutes
  - Label F_{RT} by L (0 or 1)

- Find
  - A model M to predict L when using information from the past two hours

F_{RT}: A Heterogeneous Representation

Time Sensitive

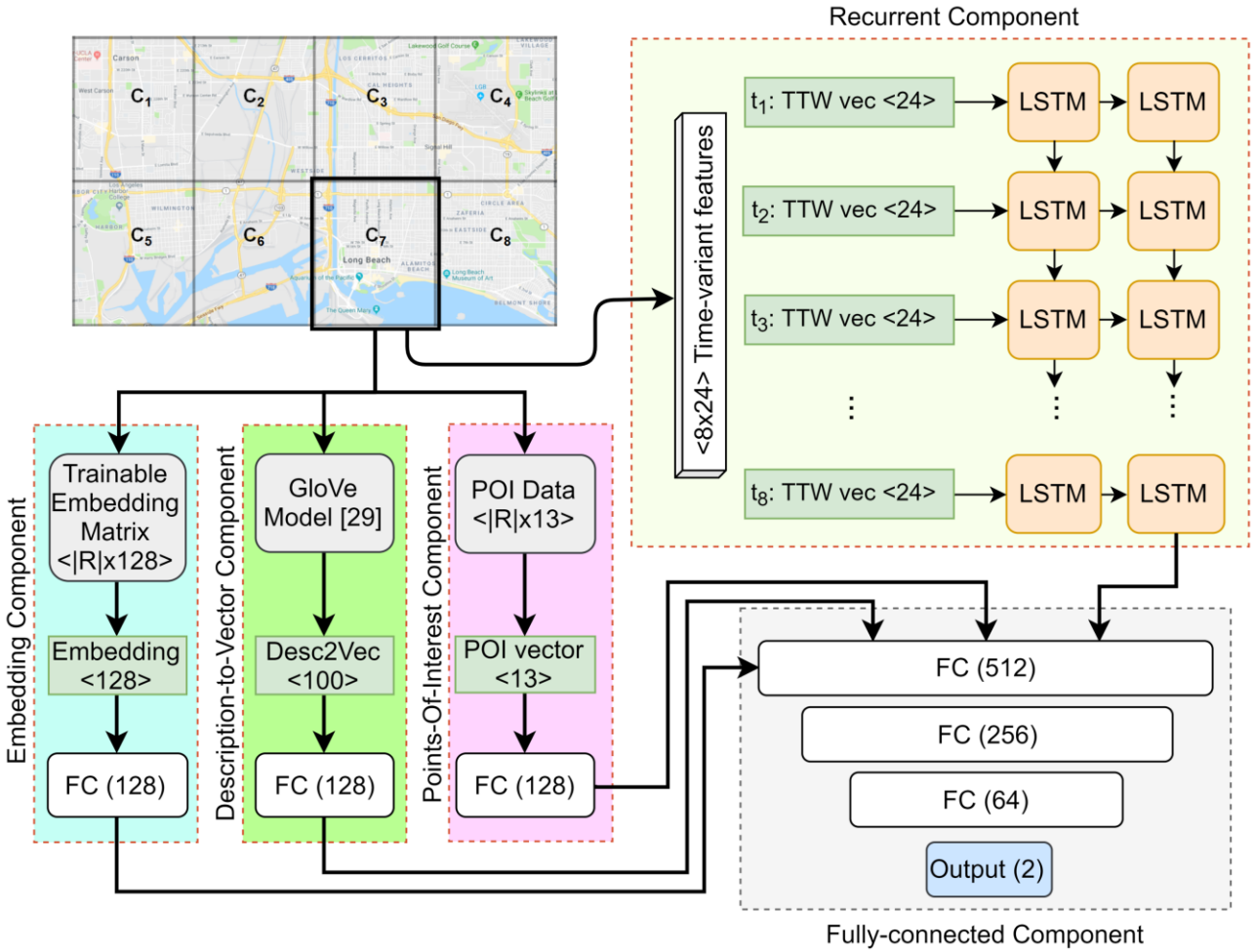
- **Traffic**: a quantitative vector of size 7 to account for various traffic events for R during T
- **Time**: TOD (weekday or weekend), HOD (5 time-intervals), and Daylight (day or night)
- **Weather**: a vector representing 10 weather attributes for R during T

Time Insensitive

- **POI**: a quantitative vector for the number of POIs inside R
- **Desc2Vec**: an embedding representation for the description of past traffic events inside R

Deep Accident Prediction (DAP) Model

- Includes five components
  - Recurrent
  - Description-to-Vector
  - Points-Of-Intertest
  - Embedding
  - Fully-connected



Experimental Setup

- We chose six cities
  - Atlanta, Austin, Charlotte, Dallas, Houston, and Los Angeles
- Data
  - From June 2018 to August 2018 (12 weeks)
  - The first 10 weeks as the train and the last two weeks as the test set
- Employed negative-sampling to account for data sparseness

Experimental Setup (Cont'd)

Baseline Models

- Logistic Regression (LR)
- Gradient Boosting Classifier (GBC)
- Deep Neural Network (DNN)
  - A four-layer neural network with feed-forward layers of size 512, 256, 64, and 2

Evaluation Metric

- *F1-score*
- Reported for each class separately, and the *weighted F1-score*

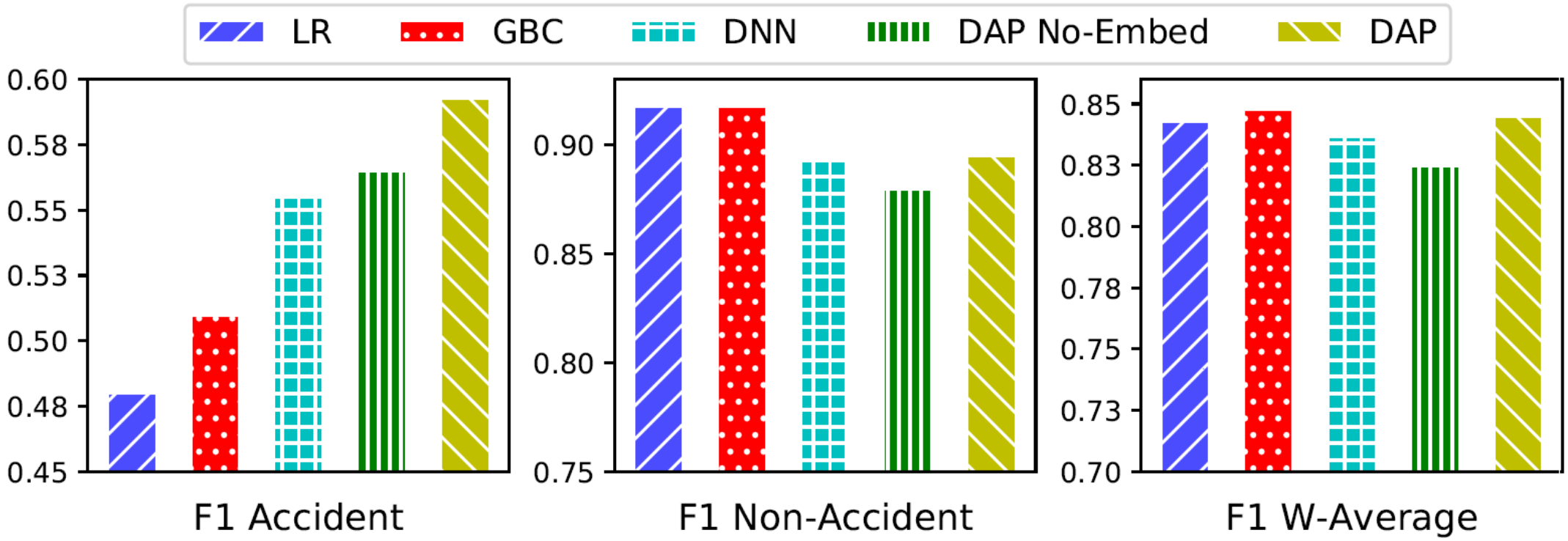
Results (Model Comparison)

City \ Model	LR			GBC			DNN			DAP-NoEmbed			DAP		
	Acc	Non-Acc	W-Avg	Acc	Non-Acc	W-Avg	Acc	Non-Acc	W-Avg	Acc	Non-Acc	W-Avg	Acc	Non-Acc	W-Avg
Atlanta	0.54	0.91	0.83	0.57	0.91	0.84	0.62	0.89	0.83	0.62	0.91	0.84	**0.65**	0.89	0.84
Austin	0.58	0.93	0.87	0.61	0.93	0.87	0.62	0.92	0.87	0.62	0.93	0.87	**0.64**	0.91	0.87
Charlotte	0.56	0.91	0.83	0.60	0.91	0.84	0.61	0.87	0.82	0.61	0.87	0.81	**0.63**	0.87	0.82
Dallas	0.30	0.94	0.87	0.32	0.94	0.87	0.36	0.94	0.87	0.43	0.88	0.83	**0.50**	0.93	0.88
Houston	0.49	0.94	0.88	0.51	0.94	0.88	**0.59**	0.93	0.88	0.58	0.92	0.88	0.58	0.93	0.88
Los Angeles	0.41	0.88	0.78	0.45	0.88	0.79	0.53	0.81	0.75	0.53	0.77	0.72	**0.56**	0.84	0.78

- **Acc**: F1-score reported for Accident cases
- **Non-acc**: F1-score reported for Non-accident cases
- **W-Avg**: F1-score reported as a weighted average for all cases

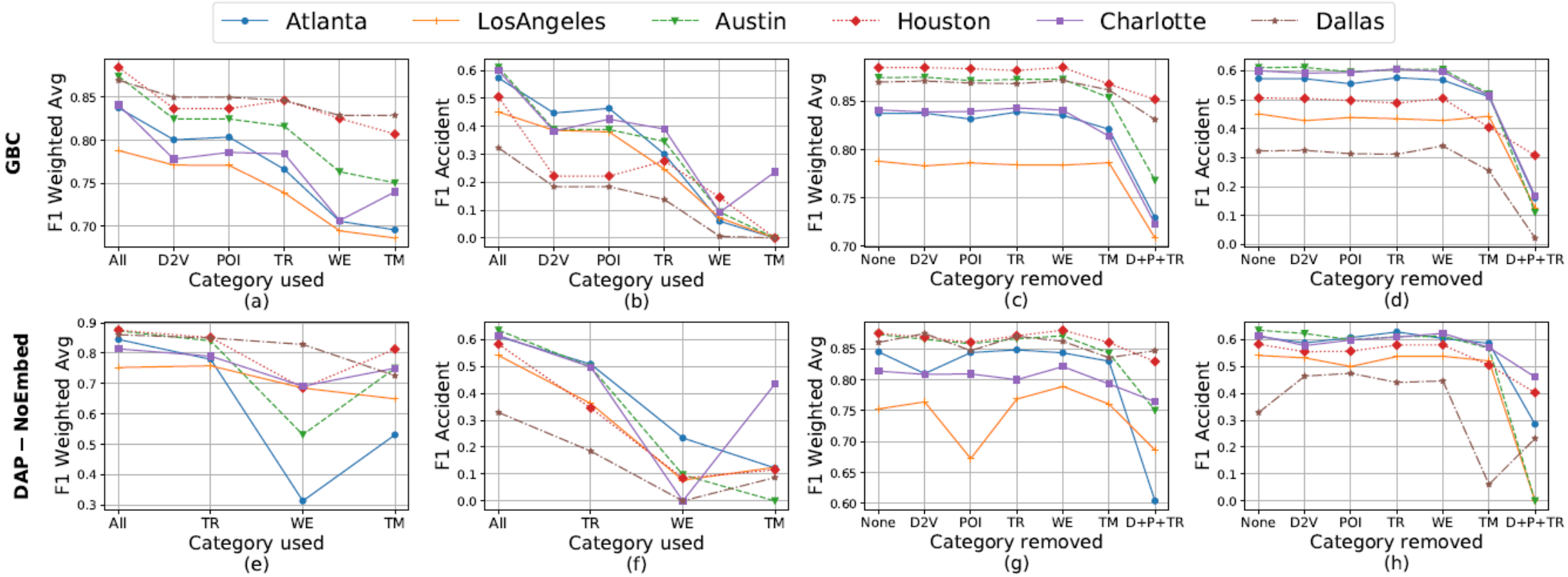
- Our proposed model performs the best for 5 out of 6 cities

Results (Model Comparison)



- Our proposed model performs the best for 5 out of 6 cities

Results (Attribute Analysis)



Summary and Future Work

- We presented a process to build a countrywide traffic accident dataset
- We presented a new approach for real-time traffic accident prediction
- Our results are comparable with the state-of-the-art solutions which rely on extensive data sources
- Future work: employing other publicly available sources of environmental data (e.g., demographic information, average daily traffic reports, etc.)

Questions

Augmentation with POI

Problem: How to find the best distance threshold to relate the location of an accident to an existing POI?

Solution: We use natural language description of traffic accidents to find the best distance threshold

Augmentation with POI (Cont'd)

- Employed *regular expression patterns* to specify the type of location
  - Identified 27 regular expression patterns
  - Four of those patterns can specify location properties

Source	Description	Type
MapQuest	Serious accident **on** 4th Ave **at** McCullaugh Rd.	Intersection
MapQuest	Accident **on** NE-370 Gruenther Rd **at** 216th St.	Intersection
MapQuest	Accident **on** I-80 **at Exit** 4A Treasure Is.	Junction
MapQuest	Accident **on** I-87 I-287 Southbound **at Exit** 9 I-287.	Junction
Bing	**At** Porter Ave/**Exit** 9 - Accident. Left lane blocked.	Junction
Bing	**At** IL-43/Harlem Ave/**Exit** 21B - Accident.	Junction
Bing	**Ramp to** I-15/Ontario Fwy/Cherry Ave - Accident.	Junction
Bing	**Ramp to** Q St - Accident. Right lane blocked.	Junction

Augmentation with POI (Cont'd)

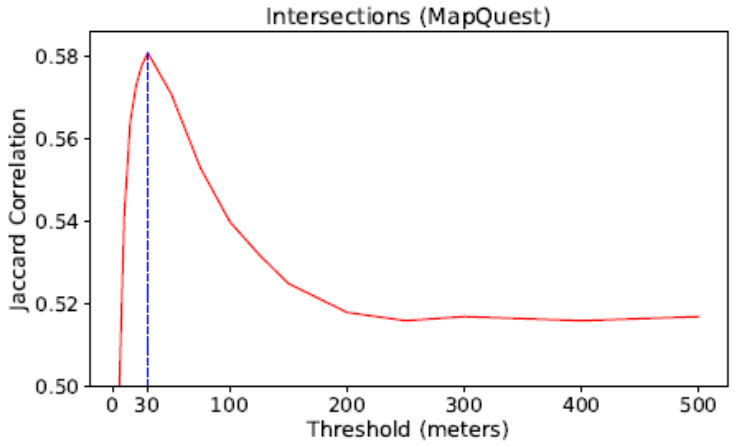
- We can potentially annotate some accident records by regular expression
- We can potentially annotate all accident records by POI tags located up to a distance threshold from them

Algorithm 1: Find Annotation Correlation

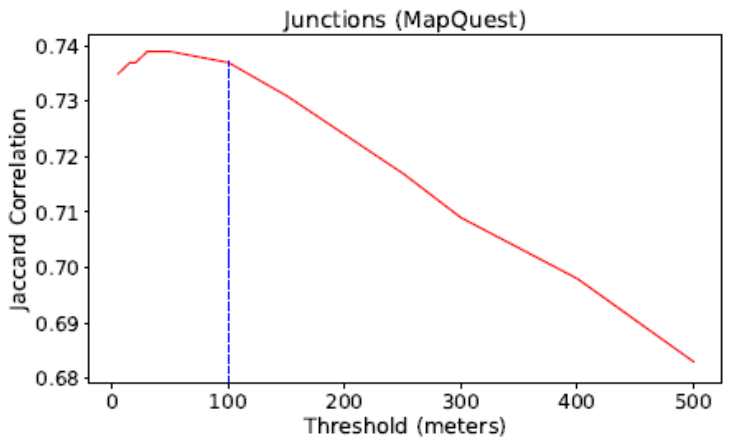
1: Input: a dataset of traffic accidents $\mathcal{A}$, a database of points-of-interest $\mathcal{P}$, and a distance threshold τ.
2: Extract and create a set of regular expression patterns RE to identify a specific POI v.
3: Create set S_1: for each traffic accident $a \in \mathcal{A}$, we add it to S_1 if its natural language description $a.desc$ can be matched with at least one regular expression in set RE.
4: Create set S_2: for each traffic accident $a \in \mathcal{A}$, we add it to S_2 if there is at least one POI $p \in \mathbf{P}$ of type v, where $haversine\_distance\,(a, p) \leq \tau$.
5: Output: Return $Jaccard\,(S_1, S_2)$.

Augmentation with POI (Cont'd)

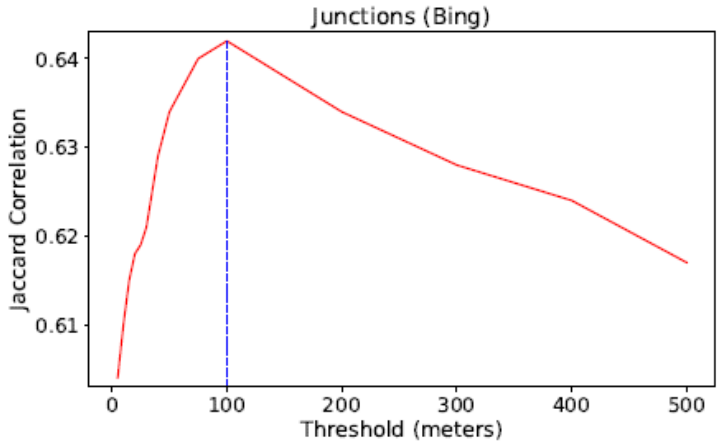
- Correlation analysis was performed on a sample of 100,000 accidents
  - Only for *Intersection* (crossing, stop-sign, and traffic signal) and *Junction*



(a) Using MapQuest for Intersection



(b) Using MapQuest for Junction



(c) Using Bing for Junction