Levering Telematics & Contextual Data Analysis for Driving Risk Prediction

Srinivasan Parthasarathy

Department of Computer Science and Engineering

The Ohio State University

Joint work with S. Moosavi (OSU → Lyft), M. Sammavatian and R. Ramnath (OSU)

December 15 2019

GRT00053368

PAS0536

EAR-1520870

Outline

Motivation and Research Problem

Summary of Contributions

I: Characterizing Driving Context

II: Characterizing Driving Style

III: Context-aware Driving Risk Prediction

Conclusion and Future Work

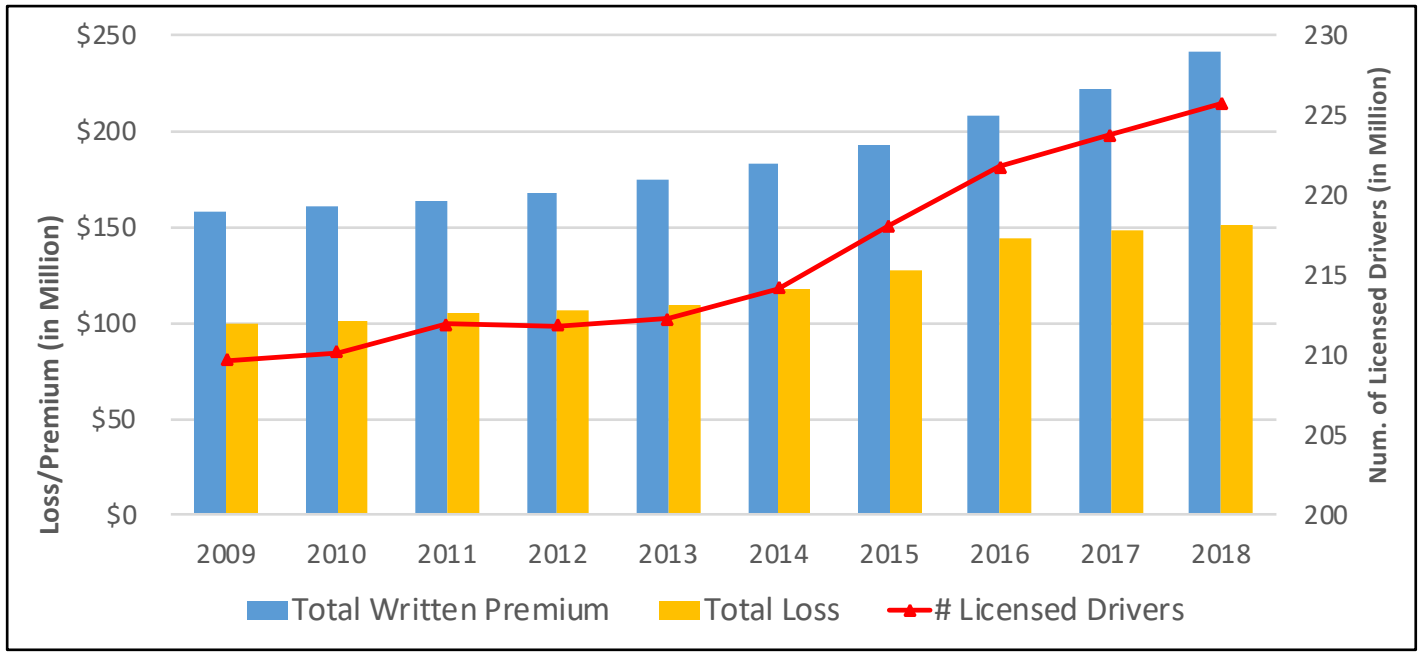
Motivation (Human cost)

- Traffic Accidents 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

Source: Association For Safe International Road Travel

Motivation (Business Cost)

- The automobile insurance in the US: loss to written premium
  - Sources: *Insurance Information Institute* and the *Federal Highway Administration*



$$\frac{loss}{premium} \approx 64.5\%$$

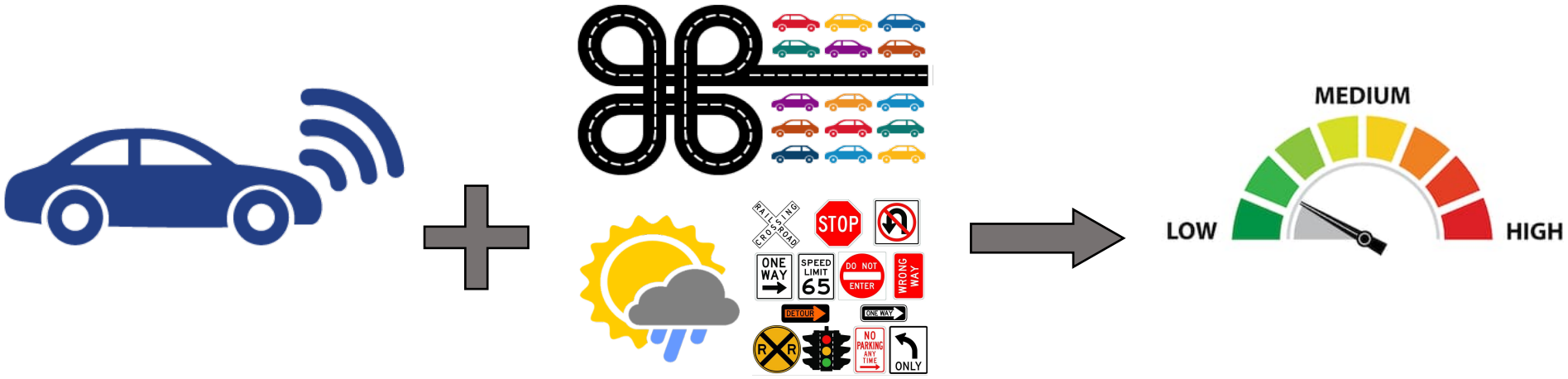
How to reduce the loss?

- By helping Drivers
  - To improve their skills and adjust their behavior and reduce premiums
- By helping Insurance Companies
  - To better predict risk and reduce insurance loss
- By helping Cities
  - To prevent disastrous events, better manage the traffic, and redesign transportation infrastructures if necessary

Research Problem: How to Determine Driving Risk?

Constraints

- Driving risk depends on the personality of drivers
- Driving risk depends on driving context

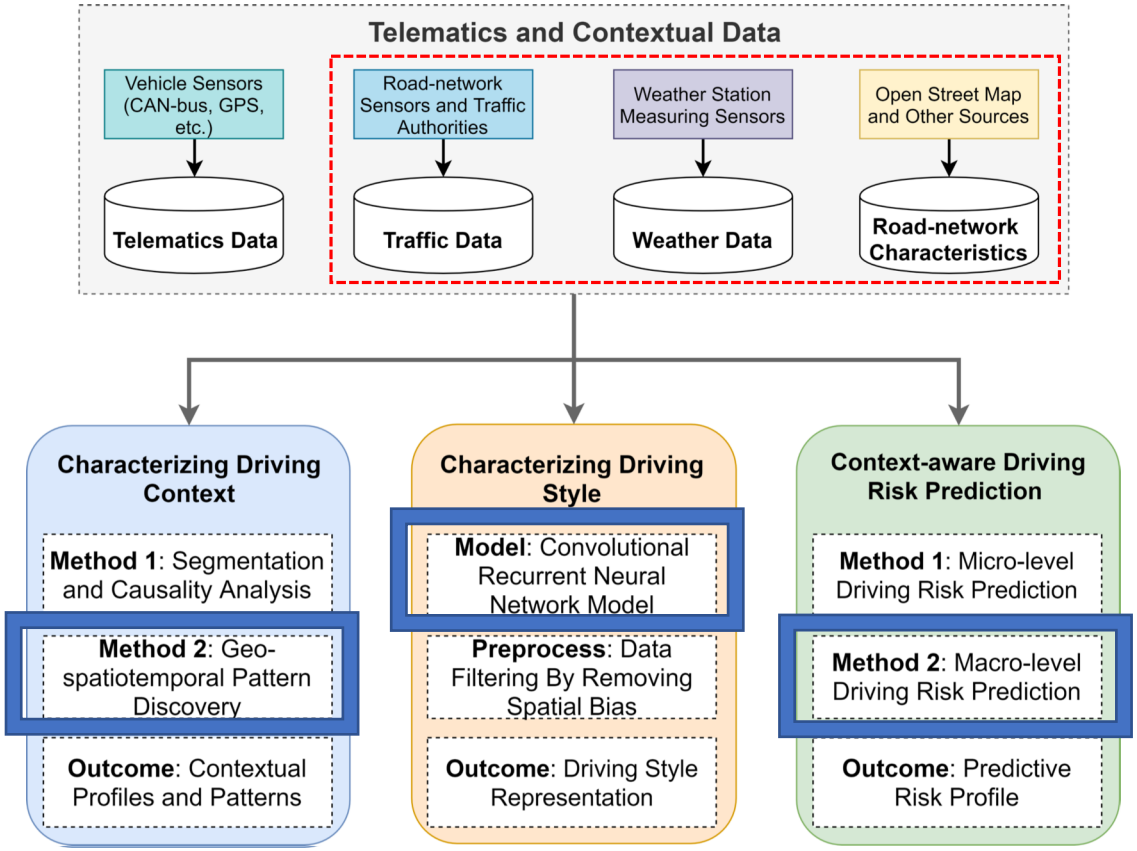


Telematics Data

Contextual Data

Driving Risk

Summary of Contributions

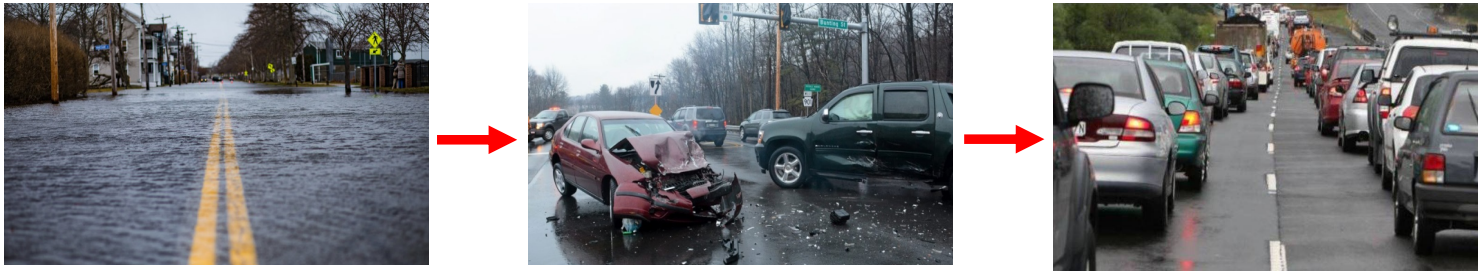


I: Characterizing Driving Context

Geo-spatiotemporal Pattern Discovery (ACM SIGKDD 2019)

- Geo-spatiotemporal data: associated with geo-location and time
  - Examples: traffic events, weather events, etc.

- Pattern discovery on geo-spatiotemporal data: co-location, co-occurrence, cause and effect, etc.



Rain → **Accident** → **Congestion**

Background and Motivation

- Importance of these patterns
  - Beneficial for urban planning, traffic management, and disaster prediction

- We propose a new framework to:
  - Discover *Propagation* Patterns (or *short-term* impacts)
  - Discover *Influential* Patterns (or *long-term* impacts)

Dataset: Large-scale Traffic and Weather Events

Traffic events: from streaming traffic reports

Entity Type	Raw Count	Relative Frequency
Accident	1,169,507	8.9%
Broken-Vehicle	308,112	2.34%
Congestion	10,542,020	80.18%
Construction	209,933	1.60%
Event	32,817	0.25%
Lane-Blocked	246,832	1.88%
Flow-Incident	637,489	4.85%
Total	13,146,710	100%

Weather events: from historical observations

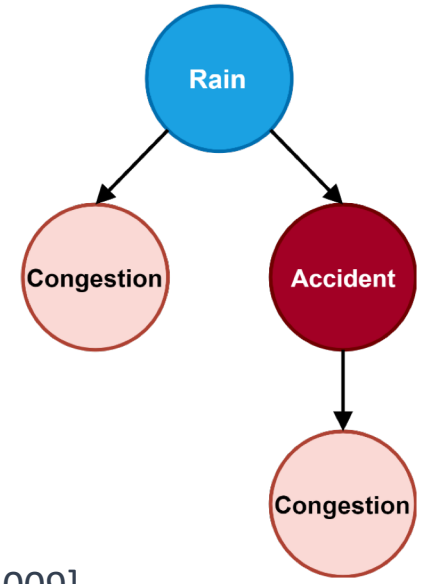
Entity Type	Raw Count	Relative Frequency
Severe-Cold	67,285	3.09%
Fog	454,704	20.87%
Hail	1,252	0.06%
Rain	1,384,588	63.54%
Snow	236,546	10.86%
Storm	14,863	0.68%
Precipitation	19,711	0.9%
Total	2,178,949	100%

Data was collected from Aug 2016 to Aug 2018 for the contiguous United States (49 states)

Short-term Pattern Discovery

- Propagation of events on a short-term basis

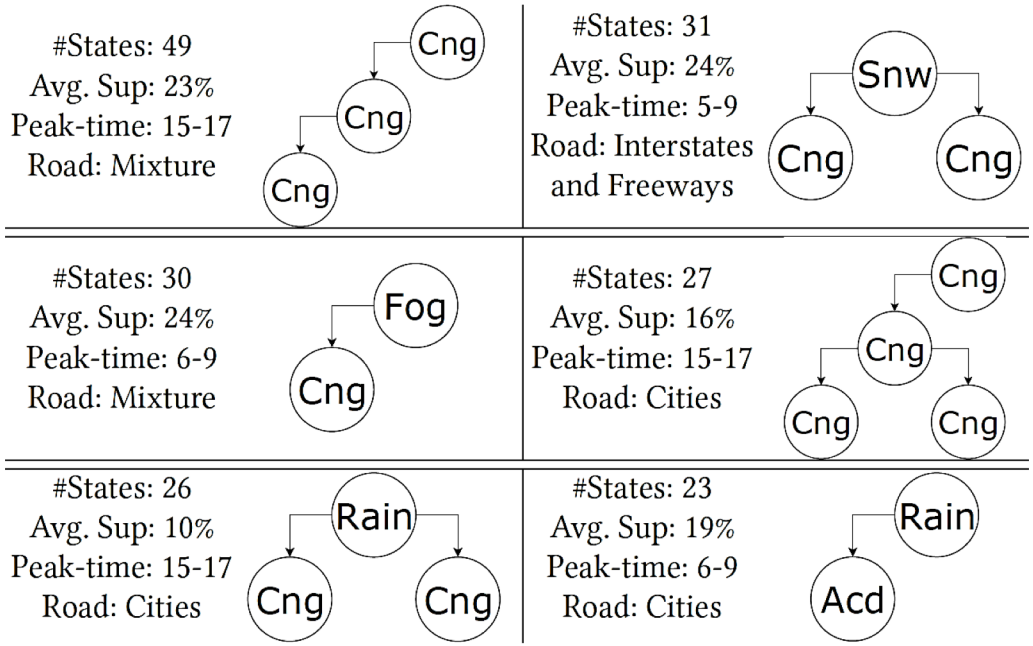
Example



Frequent tree pattern mining approach: [Zaki 2005, Tatikonda&Parthasarathy 2009]

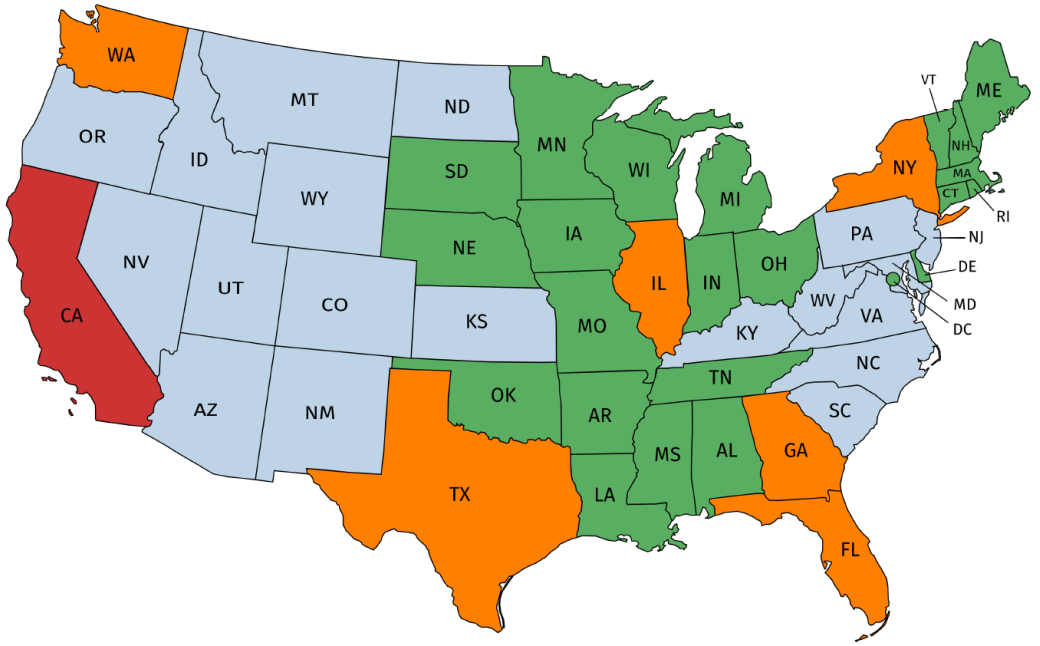
Short-term Pattern Discovery (Cont'd)

- Extracted Relations: ~ 6 million
- Extracted Trees: ~ 1.7 million
- Extracted 90 unique patterns across 49 States
  - *Embedded*, *un-ordered*, *frequent* tree patterns



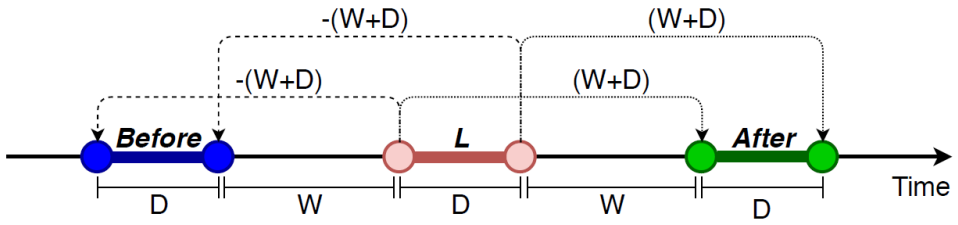
Short-term Pattern Discovery (Cont'd)

- Represented each state with a one-hot vector of size 90
- Identified 4 clusters of states using K-means
- These clusters represent similarities based on short-term impacts



Long-Term Pattern Discovery

- Impact of long-term events on their neighborhood
  - **Example**: *major construction → more traffic jams*
- Long-term event: longer than 5 hours
- Main task: Comparing current with before and after time-intervals



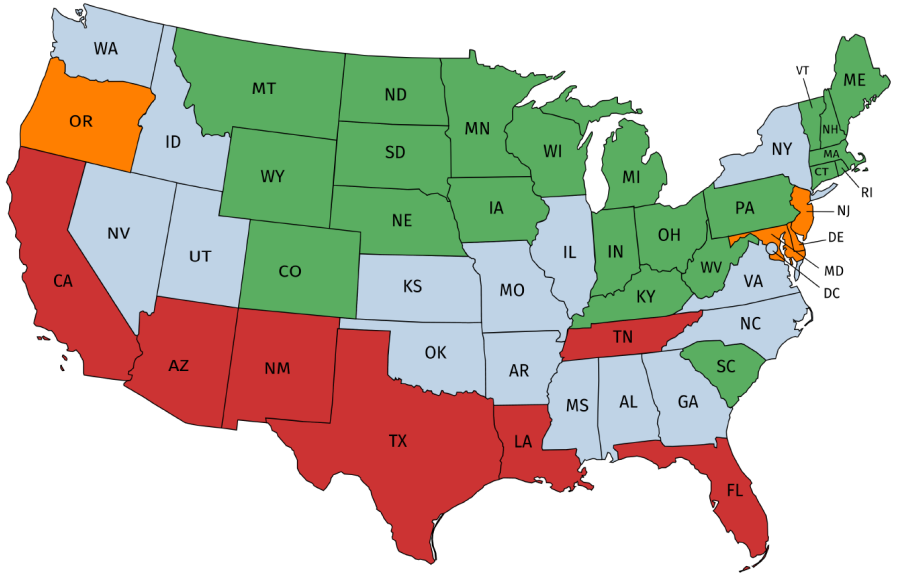
Long-Term Pattern Discovery (Cont'd)

- Statistical significance testing to determine impacts
  - Positive: the presence of a long-term event → increase in the number of nearby events
  - Negative: the presence of a long-term event → decrease in the number of nearby events

- **Results**: Impact by Location
- **Observation**: CA, FL, and TX are top states with the most traffic issues

Color Code

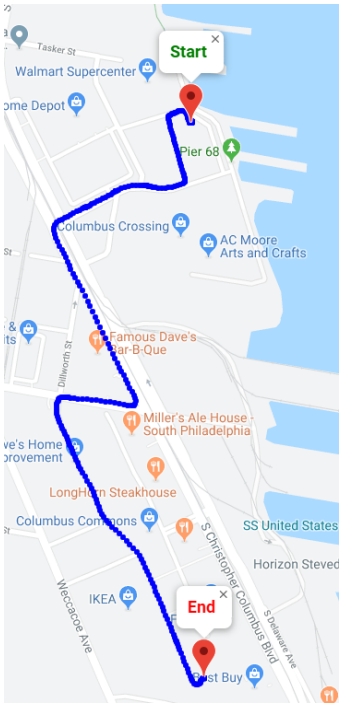
- Positive impact by 99% confidence
- Positive impact by 95% confidence
- Positive impact by 90% confidence
- No significant impact



II: Characterizing Driving Style

Characterizing Driving Style (2018 – 2019)

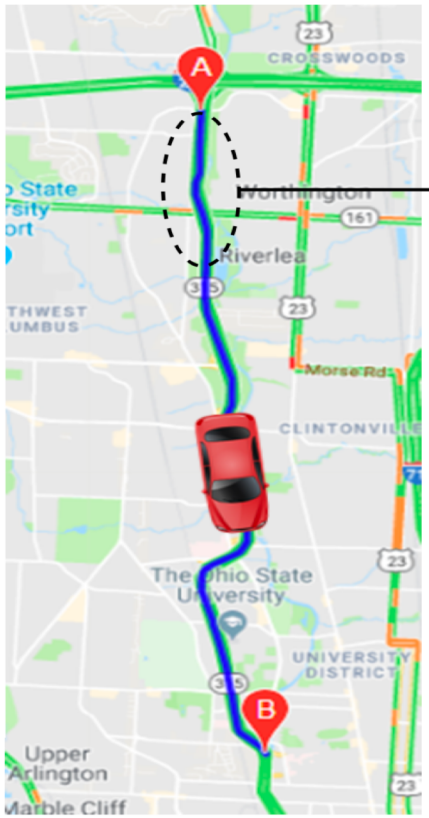
Trajectory



Who is the driver?

Constraint: exploit information on how people drive, instead of where they drive!

Feature extraction



(a)

(b)

Features	T_1	T_2	T_3	T_4	...	T_{256}
Speed	4.6	4.9	6.4	6.0	...	6.4
Acceleration	-1.0	0.3	1.5	-0.4	...	-1
GPS_Speed	4.2	4.7	6.2	6.1	...	6.5
GPS_Accel	-0.8	0.5	1.5	-0.1	...	-0.8
Angular_Speed	0.2	0.25	0.32	0.21	...	0.21
RPM	1250	1400	2400	2100	...	2360
Heading	156	158	168	168	...	273
Acceleration_X	0.85	0.89	1.2	-0.2	...	-0.6
Acceleration_Y	0.4	0.63	0.92	-0.31	...	-0.5
Acceleration_Z	0.01	0.00	0.01	0.00	...	0.00

(c)

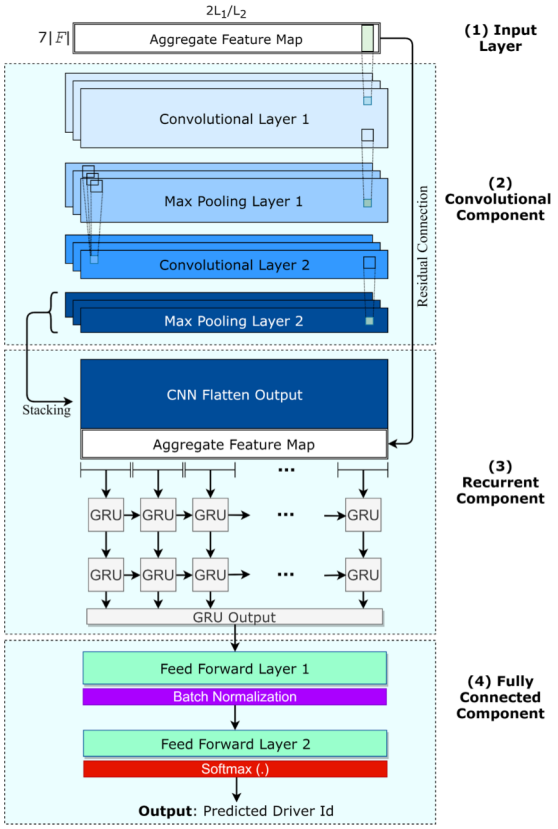
Mean	5.48	5.21	...	6.21
Min	4.6	4.2	...	5.5
Max	6.4	6.7	...	6.8
P-25%	4.83	4.45	...	5.73
P-50%	5.45	5.32	...	6.2
P-75%	6.1	6.4	...	6.5
Std	0.86	1.2	...	0.21

70 (features)

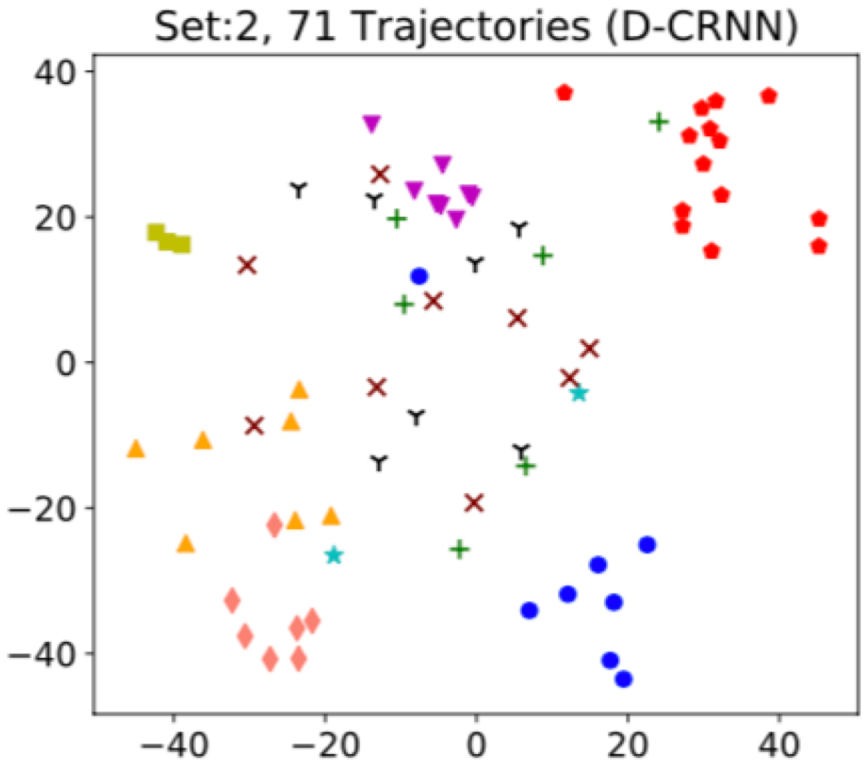
128 (time)

Characterizing Driving Style (Cont'd)

- We proposed several data sampling strategies to avoid spatial bias
- We developed a neural network architecture to encode driving style
- We thoroughly tested our proposal based on real-world data



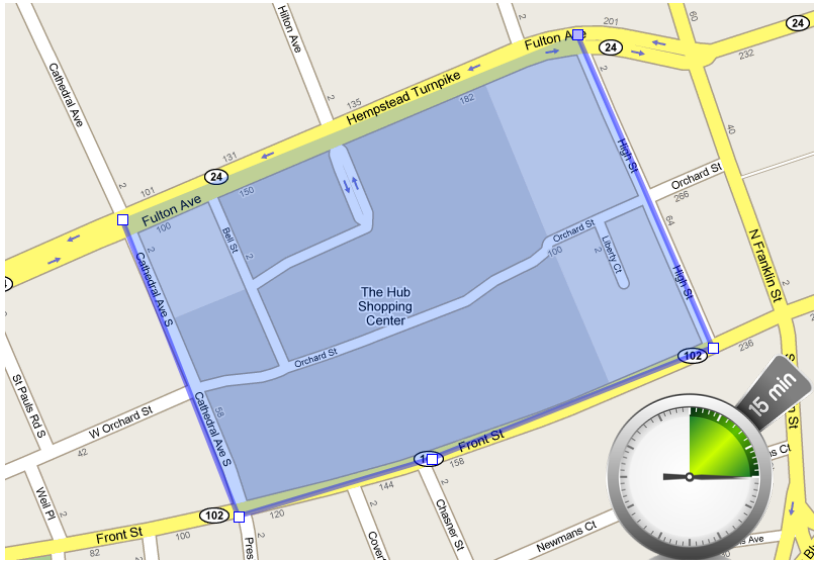
Representation of Driving Style



III: Context-aware Driving Risk Prediction

Micro and Macro-level Driving Risk Prediction

Micro-level: Prediction for a Driver



Macro-level: Prediction for a Region

Macro-level Driving Risk Prediction: Traffic Accident Prediction (ACM SIGSPATIAL 2019)

- Traffic Accidents: explicit indicators of driving risk
  - A global status report: 1.25 million traffic death in 2013

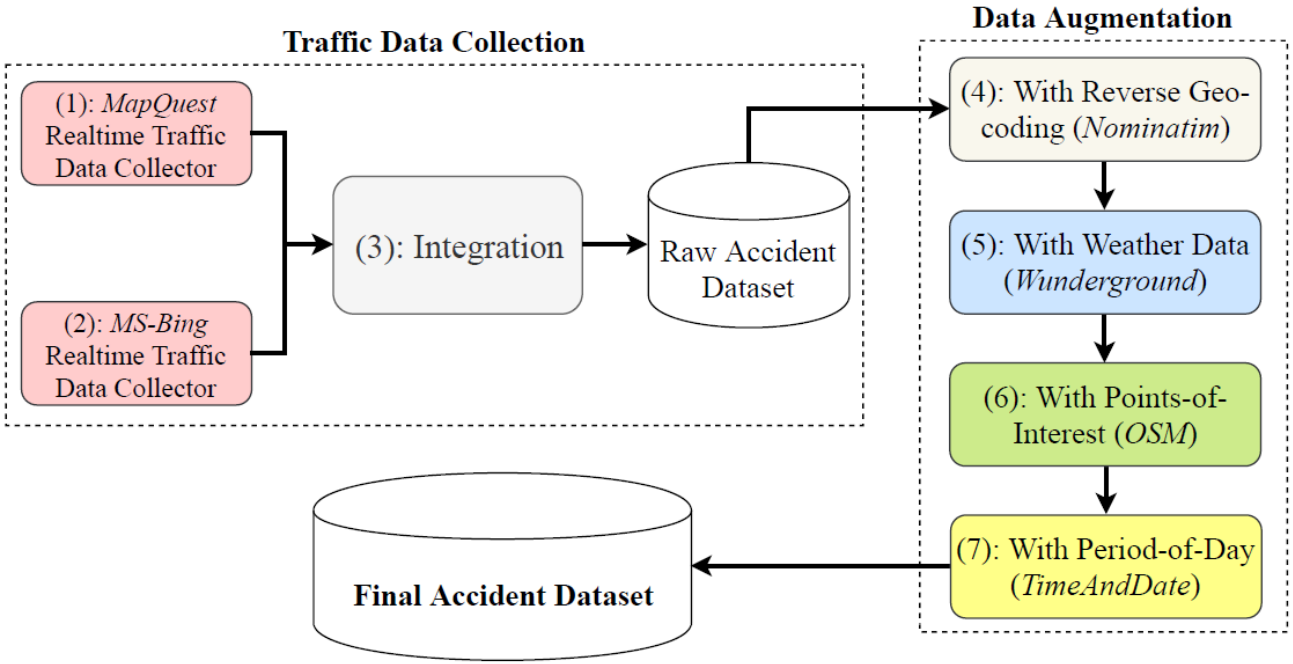
- Related studies over the past few decades
  - Analyzing the impact of environmental stimuli
  - Predicting the frequency of accidents
  - Predicting the risk of accidents

Existing Studies Suffer From …

- Using small-scale datasets
- Utilizing expensive data sources
- Being inapplicable for real-time purposes

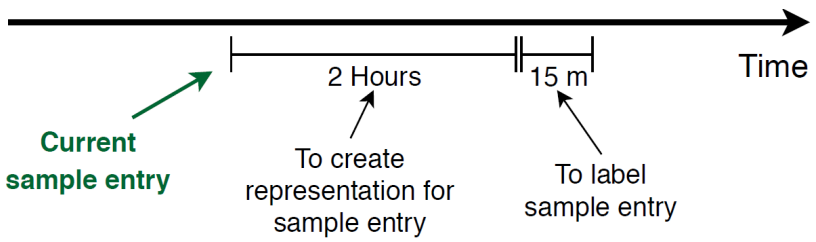
Large-Scale Traffic Accident Dataset

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



Traffic Accident Prediction: 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 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

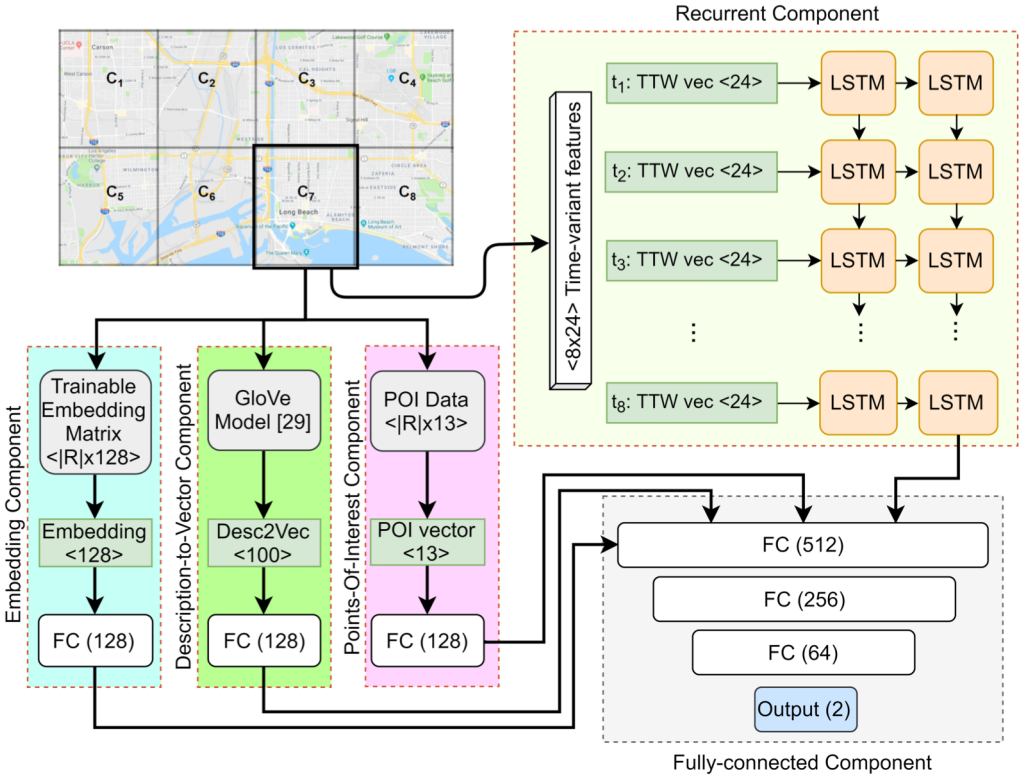
- **Traffic**: a quantitative vector of size 7 to account for various traffic events for R during T
- **Time**: TOD (weekday or weekend), HOD (5 time-intervals), and Daylight (day or night)
- **Weather**: a vector representing 10 weather attributes for R during T

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
    - Captures time-varying info
  - Description-to-Vector
    - Captures prior accident history
  - Points-Of-Interest
    - Captures roadway context
  - Embedding
    - Captures spatial heterogeneity
  - Fully-connected
    - Integrates



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 imbalance issue

Results (Model Comparison, F1-score)

City \ Model	LR	GBC	DNN	DAP-NoEmbed	DAP
Atlanta	0.54	0.57	0.62	0.62	**0.65**
Austin	0.58	0.61	0.62	0.62	**0.64**
Charlotte	0.56	0.60	0.61	0.61	**0.63**
Dallas	0.30	0.32	0.36	0.43	**0.50**
Houston	0.49	0.51	**0.59**	0.58	0.58
Los Angeles	0.41	0.45	0.53	0.53	**0.56**

Baselines

- Logistic Regression (LR)
- Gradient Boosting Classifier (GBC)
- Deep Neural Network (DNN)
- DAP without Embedding Component (DAP-NoEmbed)

Ongoing Work: Micro-level Driving Risk Prediction

How to determine driving risk?

- Traditionally: by using demographic data
- Today: by using demographic + telematics data

I: Contextualizing Telematics Data

- Maximizing the usage of telematics and contextual data
- Providing various views on driving behavior

II: Building Risk Cohort Classifier

- Refining risk labels by a data-driven process
- Building a classifier based on contextualized telematics data

III: Online Risk Cohort Prediction

- Utilizing contextualized telematics data
- Predicting risk cohort for drivers in real-time

Summary and Future Work

Published Datasets

- **DACT**: A Dataset of Annotated Car Trajectories for driving behavior analysis and transportation research (2017)
- **Large-scale Traffic and Weather Events (LSTW)**: A large-scale dataset of traffic and weather events, containing 25 million events (2019)
- **US-Accidents**: A data of 2.25 million traffic accidents (2019)

Summary

- Several models were proposed to leverage telematics and contextual data to:
  - Characterize driving context
  - Characterize driving style
  - Predict driving risk

- Our solutions can help …
  - Drivers to improve their skills
  - Insurance companies to better predict risk and reduce loss
  - Cities to better manage traffic and reduce disastrous events

Future Work

- Extend the usage of Telematics Data
  - Using more sensors and finer-grained data collection rates
- Extend the usage of Contextual Data
  - Utilizing detailed road data, weather data, etc.
- Augment the Risk labels
  - Employing risk labels for sub-trajectories or a series of actions
- Evaluate models in real-world
  - Employing A/B testing

Q&A