

# CSC 369 FINAL PROJECT

Winter Quarter 2020


Derek Kelley, Kevin Yoo,  
Joulien Ivanov, Tyler Davis

# OUR DATA


US Traffic Accident Dataset (2016-2019)  
*(sampled 750,000 records out of 3.0 million records)*

- Source: Kaggle

# QUESTIONS TO ANSWER


1. Severity & Roadway Type in top 5 states with the highest rate of accidents
  2. Severity & Description
  3. Severity & Time Year / Day
  4. Reporting Source & Roadway Type / Severity
- 
- Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

# METHODOLOGY

- Use the Hadoop technology in order to leverage the distributed computing logic of Spark and efficiently analyze large amounts of data
  - Find interesting correlations/patterns in the data in order to draw important conclusions about US automobile accidents
- 
- A series of three parallel white diagonal lines extending from the bottom right towards the top right of the slide.

# Q1

Which 5 states have the highest rate of car accidents?  
For each of the 5 states, on which road types did most accidents occur and is there a correlation between the road type and severity of the accident?

Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

# Q1 (CONT.)

Method:

- 2 RDDs:

- Map(state, # of accidents)

- Map(state, SortedList(roadtype, # of accidents, avg severity))

- Road types were extracted using regular expressions

- “I-5 W” contains “I-`{number}` `{bound}`”
  - “CA-1 S” contains “`{2 char abbv for state}`-`{number}` `{bound}`”

# Q1 (CONT.)

## Results:

- CA 9768
  - (Local Street, 4039, 2.23), (Interstate, 2352, 2.92), (Freeway, 1748, 2.58), (State Highway, 1388, 2.09), (Highway, 241, 2.08)
- TX 4877
  - (Local Street, 3837, 2.13), (Interstate, 417, 2.87), (State Highway, 367, 2.43), (Freeway, 213, 2.56), (Highway, 43, 2.28)
- FL 3034
  - (Local Street, 2329, 2.18), (Interstate, 619, 2.92), (Highway, 53, 2.15), (State Highway, 33, 2.39)
- PA 1835
  - (Local Street, 1573, 2.08), (Interstate, 136, 2.86), (State Highway, 80, 2.31), (Highway, 46, 2.37)
- NY 1732
  - (Local Street, 1448, 2.37), (Interstate, 253, 2.82), (State Highway, 16, 2.31), (Highway, 15, 2.13)

## Q1 (CONT.)

### Observations From Findings:

- California had the highest number of accidents in this sample
- For all 5 top states, the most accidents occurred on local streets (1<sup>st</sup>) then on interstates (2<sup>nd</sup>)
- For all 5 top states, the road with the highest average severity was the Interstates



## Q2

Is there a correlation between the natural language description format and the severity of the accident?  
Does the API source matter?

Map Quest's 2 Main Types of Descriptions:

Accident on Reveille St at I-45.

Left lane blocked due to accident on US-6 Hartford Ave Westbound

Bing's 2 Main Types of Descriptions:

At Parkview Blvd - Accident.

Closed at MD-224/Chicamuxen Rd - Road closed due to accident.

## Q2 (CONT.)

Method: Filter by description and API, then average severity

Performance: 15.8 sec

Results: (Average severity per description type & API)

Description / API > v	MapQuest	Bing
"Accident on..."	2.19	
"Lane blocked due to..."	2.59	
Other	2.31	
"- Accident"		2.18
"- Road closed due to accident"		4.0
Other		2.37

## Q2 (CONT.)

### Observations from Findings:

- For MapQuest, accidents that have the description type of “Lane blocked due to...” are 13% more severe than accidents with the description type “Accident on...”
- For Bing, accidents that have the description type of “- Road closed due to accident” are 60% more severe than accidents with the description type “- Accident”
- “- Road closed due to accident” description type from the Bing API had an average severity of 4.0 (max severity)

## Q3

Part 1) What days of the year have the most accidents?

Part 2) What hours are most accident prone?

Part 3) What hours have most accidents with severity 4?



# Q3 – PART 1 What days of the year have the most accidents?

## Method:

- ▶ 1 RDD:
  - ▶ Map accidents to (month-day, count) pairs
  - ▶ MapValues to (# of accidents) / (# of accidents per day on average)

## Results:

(month-day   number of times over daily average*)	
▶ 12-12	3.29
▶ 12-13	3.29
▶ 11-06	3.27
▶ 12-11	3.26
▶ 12-19	3.25

\*daily accidents average: 877

## Q3 – PART 1 What days of the year have the most accidents?

### Observations from Findings:

These top 5 dates (and in fact *top 10* dates) all have something in common: they occur in **late fall** or **early winter**.

For many parts of the US, it is characteristic for the first snowfalls or storms to take place during that time of the year. Often, the first storms are the most brutal as a majority of people are under-prepared. Not only is the weather more unreliable, but holidays add to the problem as many people embark on journeys to visit friends and family or simply go out shopping for presents.

## Q3 – PART 2 What hours are most accident prone?

### Method:

- ▶ 1 RDD:
  - ▶ Map accidents to (hour, count) pairs
  - ▶ ReduceByKey and Sort

### Results:

(hour am/pm | number of accidents)

- ▶ 8 am | 63562
- ▶ 7 am | 61531
- ▶ 17 (5 pm) | 48144
- ▶ 16 (4 pm) | 46538
- ▶ 15 (3 pm) | 38629

## Q3 – PART 2 What hours are most accident prone?

### **Observations from Findings:**

Perhaps unsurprisingly, the top hours with most crashes happen to be 8 am and 7am followed by 5 pm and 4 pm.

It is no coincidence that most people have work either from 8-4pm or 9-5pm. As one can imagine, as more people leave to go to work and come back, more drivers enter the roads. With more drivers, peak traffic is reached, and as a result, the chances of an accident are magnified.



## Q3 – PART 3 What hours have most accidents with severity 4?

### Method:

- ▶ 1 RDD (same as Part 2 but filtered with severity == 4)

### Results:

(hour am/pm | number of accidents)

- ▶ 16 (4 pm) | 1075
- ▶ 17 (5 pm) | 1056
- ▶ 8 am | 1039
- ▶ 15 (3 pm) | 1037
- ▶ 13 (1 pm) | 1013

## Q3 – PART 3 What hours have most accidents with severity 4?

### **Observations from Findings:**

The peak hour with most dangerous accidents is 4 pm. As we learned earlier, traffic is entering one of its peak hours at that time.

It is possible to correlate that because there were fewer cars before 4 pm on the road, it is more likely that people who were anxious to leave work will speed their way home. The problem with driving at faster speeds is that more catastrophic accidents are likely; this ultimately leads to a larger quantity of accidents with a severity 4 categorization.

Total time to run all of Q3: 18.551s

# Q4

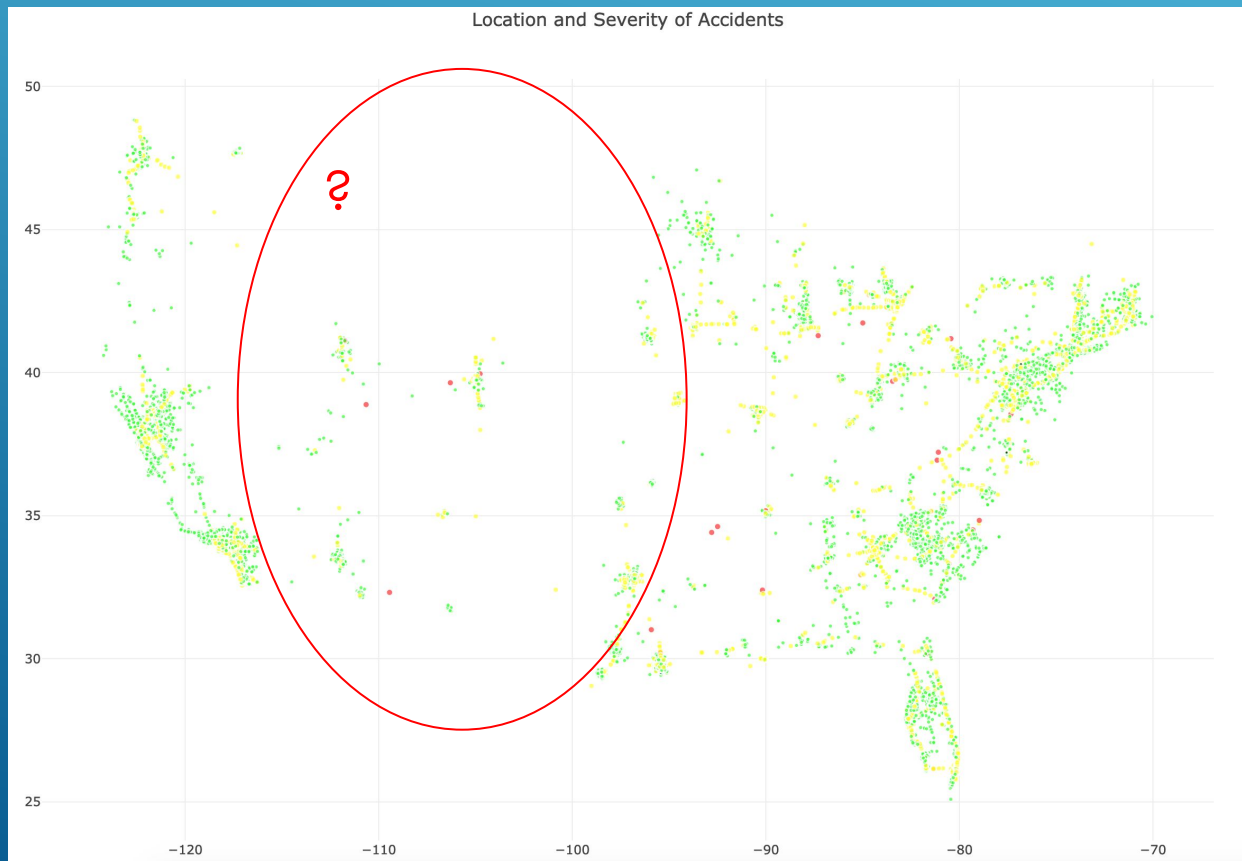
Part 1) Do different reporting sources report accidents differently based on roadway?

Part 2) Do different reporting sources report severity differently?

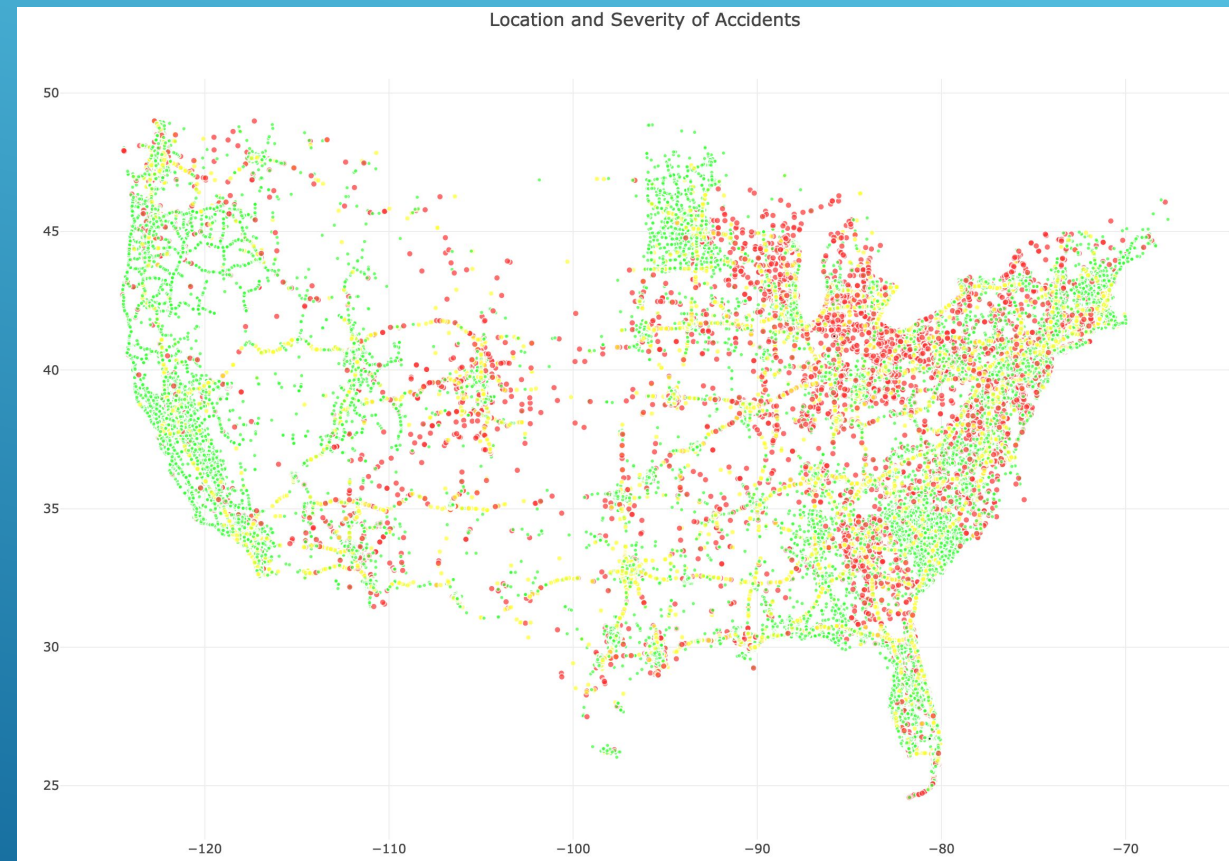
Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

# Q4 – Inspiration

## MapQuest only



## Bing, MapQuest, and Bing-Mapquest



## Q4 – PART 1 Reporting source and roadway type

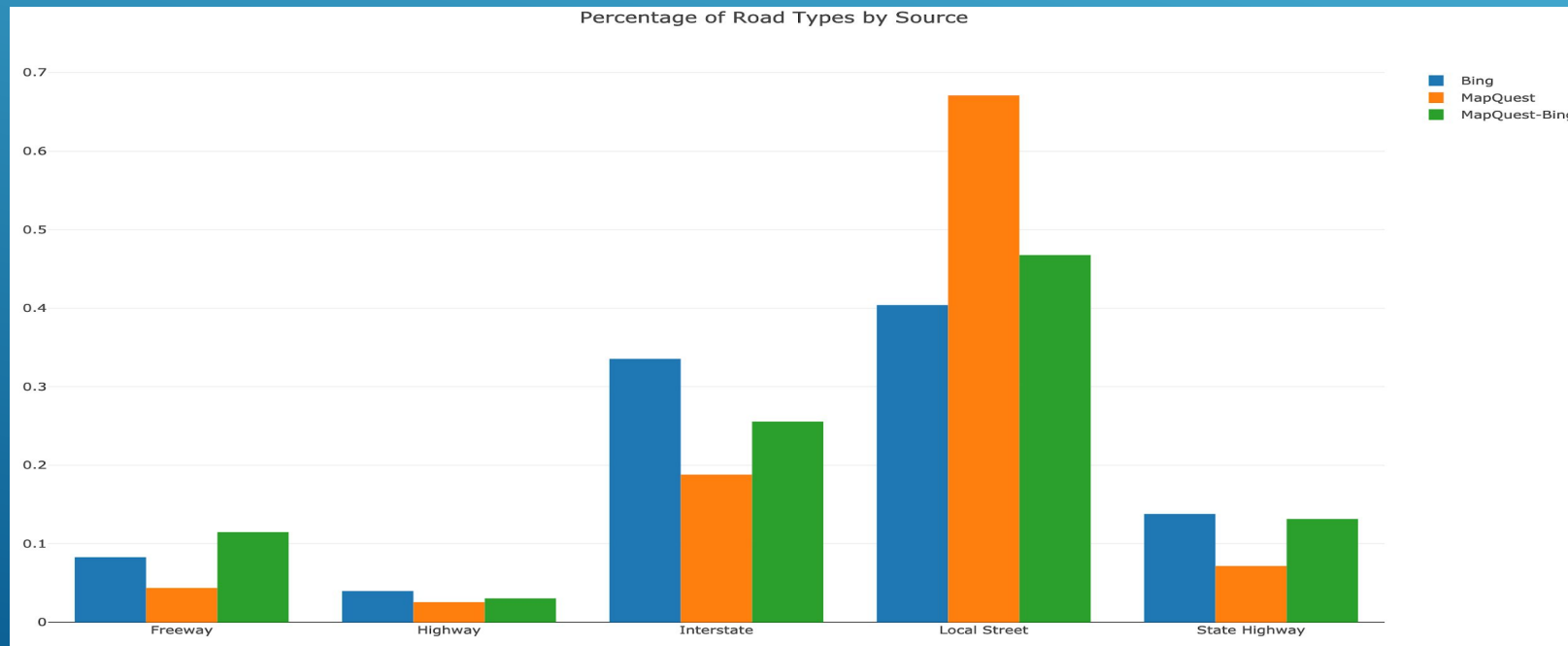
### **Method:**

- ▶ Use the roadway mapping from Q1 to get the roadway type for each accident
- ▶ Partition the data set based on reporting source and roadway type
- ▶ Count the number of accidents for every source and roadway type
- ▶ Divide the count of accidents by the total number of accidents reported by that source
- ▶ End result is the percentage of accidents reported by each source that occurred on a specific type of road

# Q4 – PART 1 Reporting source and roadway type

## Findings and Observations:

A much higher proportion of Bing's reported accidents occur on interstates and state highways as compared to MapQuest. This would explain the vast swaths of the US that didn't report accidents on MapQuest. This may indicate a difference in the way Bing and MapQuest source traffic data.



# Q4 – PART 2 Reporting source and severity

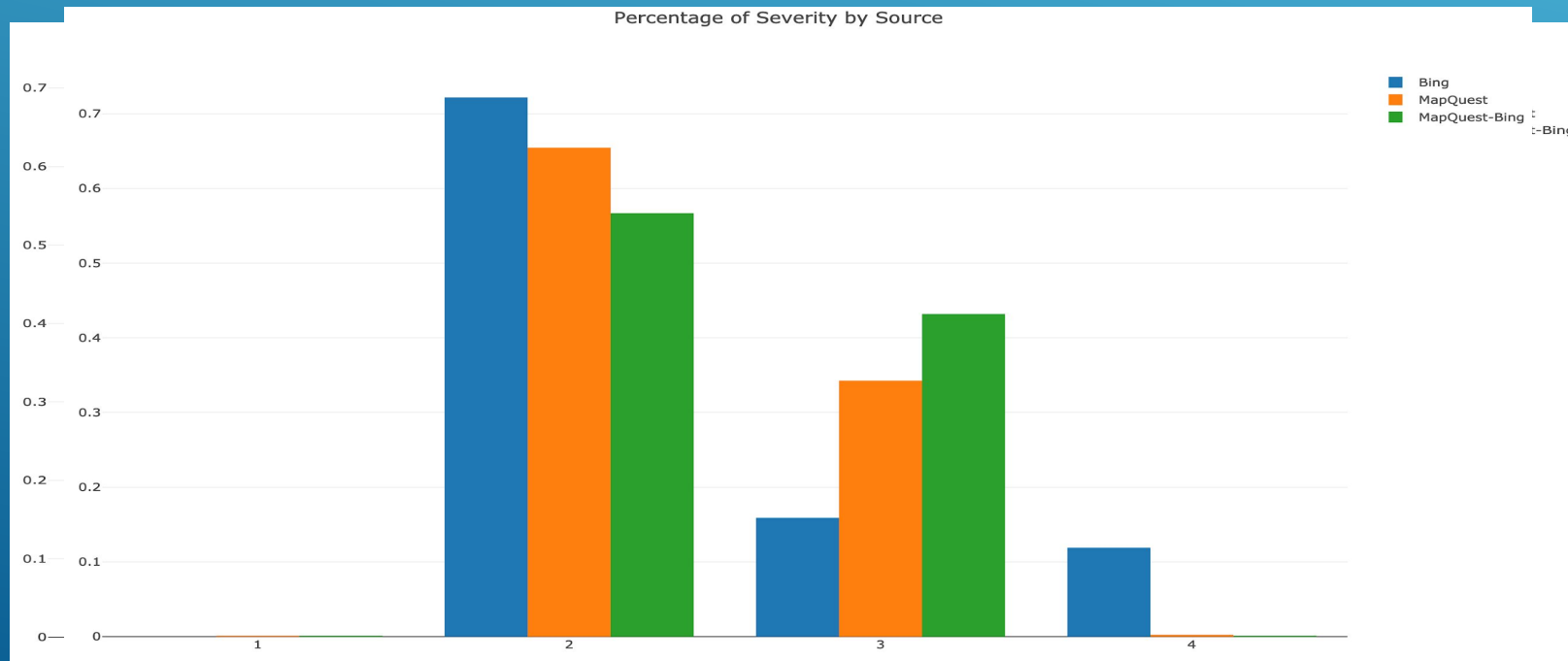
## **Method:**

- ▶ Use a similar method as part 1
- ▶ Count all accidents by source and severity
- ▶ Divide the number of accidents by the total number of accidents from the reporting source

## Q4 – PART 2 Reporting source and severity


### Findings and Observations:

A much higher proportion of Bing's reported accidents were the maximum severity level. While a far higher proportion of MapQuest's reports were level 3. This could suggest a further difference in classification; however, Bing reports more interstate accidents (which are more severe from Q1).





# WHAT WE LEARNED

1. There are a lot of insights that can be made from analyzing car accident data, which can be used to better improve road construction in the future
  2. Scala is your friend when analyzing large amounts of data
  3. Do the simple thing and don't try to complicate your code (using `groupByKey`, `aggregate`, etc. when appropriate)
- 
- A series of three parallel white diagonal lines in the bottom right corner of the slide, extending from the bottom edge towards the right edge.

# OBSTACLES

1. Filling NULL values in the dataset
2. Using Regular Expressions to recognize string patterns

QUESTIONS?

