# COMS E6898: FINAL PRESENTATION

SUBMITTED BY:
NEHA GUPTA
(NG2565)

# TOPIC: REAL-TIME TRAFFIC MONITORING

- Current Real-time Traffic Monitoring and Guidance Systems:
  - Use Cellular Network
  - Radio Signals
- Commercial applications use Satellite Data or a combination of above

**Novelty of approach**

- This is a novel approach in the way it uses the real time street-cam data (collected from various locations in the New York City) and the tweets from twitter to generate real-time navigation and supervision features.
- In [7], the cameras are sending tweets on twitter which is different from this approach.

**Motivation**

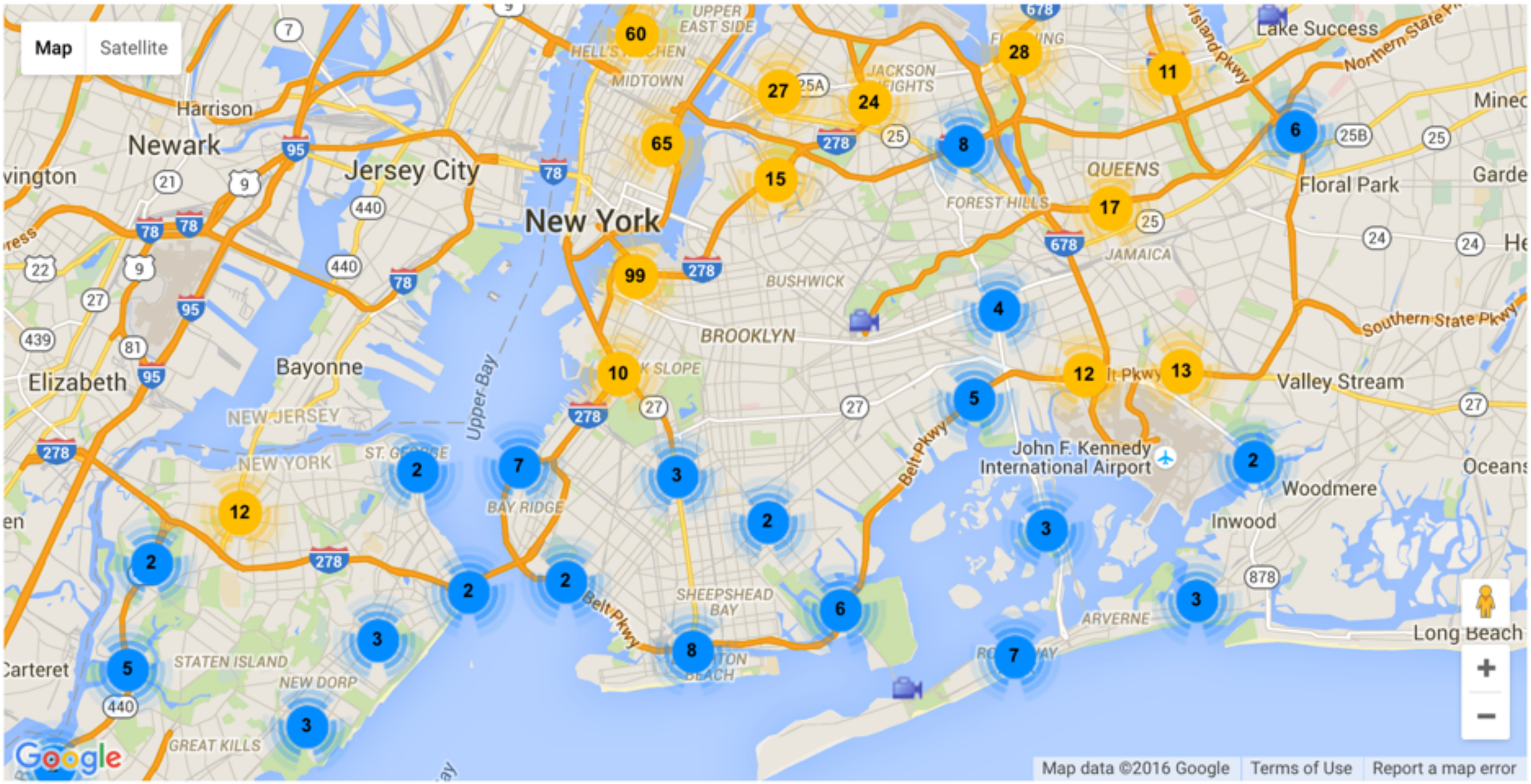
The paper "Tweeting Cameras for Event Detection" [7] inspired me to take up this project.

**Besides, the thrill to work on a real life dataset!**

# Real Time Traffic Information

Main Map | Mobile | Cameras List | Midtown Map | Traffic Speed | Glossary | Help

**Search Camera By Name:** GO



# PROPOSAL

We will be undertaking the following tasks for this project:

1. Start with a waze report, find relevant waze reports that are talking about the same event,  then correlate this filtered data with tweets from twitter and find common occurrences of the same event in both waze and twitter.

2. Start with Twitter data, process the tweets on Twitter to find relevant new events. These may not be present in waze.

3. Design an interactive annotation tool that will serve as a user-interface to link waze events with the images. This task has been de-scoped after submission of the project proposal so it will be not be analyzed here.

# PROPOSED DATA

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 Next > Last >>

| Actions | | tweetid | time | long | lat | addr | content | userid |
|---|---|---|---|---|---|---|---|---|
| Edit | Delete | 1 | 2015-12-10 00:00:04 | -76.33805556 | 42.12361111 | | Wind 3.7 mph S. Barometer 29.972 in  Steady. Temp... | 19719569 |
| Edit | Delete | 2 | 2015-12-10 00:00:06 | -74.0007613 | 40.7207559 | Peterborough | NEW BLOG POST: A Morning in Cobourg  https://t.co... | 362587078 |
| Edit | Delete | 3 | 2015-12-10 00:00:06 | -74.0007613 | 40.7207559 | East Sussex | Lily Pads  https://t.co/cey4lQ2e5f | 19446802 |
| Edit | Delete | 4 | 2015-12-10 00:00:11 | -73.9803314 | 40.7744598 | | Joanne Trattoria Fam #joannetrattoria #family #ho... | 14955733 |
| Edit | Delete | 5 | 2015-12-10 00:00:13 | -74.0059731 | 40.7143528 | Charlotte | #Hospitality #Job alert: CAFE MANAGER SENIOR \| Co... | 126371773 |
| Edit | Delete | 6 | 2015-12-10 00:00:14 | -73.98757295 | 40.74559825 | New York | Pin it together for the holidays. Shop a piece of... | 131743957 |
| Edit | Delete | 7 | 2015-12-10 00:00:18 | -73.98342497 | 40.72632363 | | Had the distinct pleasure of seeing shaunbarker12... | 465963214 |
| Edit | Delete | 8 | 2015-12-10 00:00:28 | -73.9746896 | 40.78376511 | | Celebration! (at @TheMillingRoom in New York  NY)... | 15997164 |
| Edit | Delete | 9 | 2015-12-10 00:00:28 | -73.93162786 | 40.66841131 | | I'm at Eastern Parkway & Utica Avenue in Broo... | 16532821 |
| Edit | Delete | 10 | 2015-12-10 00:00:37 | -74.0007613 | 40.7207559 | New York City | Comic Book Emotions.  https://t.co/z4WHuo6yfc | 35484447 |

Tweets Table in Database

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 Next > Last >>

| Actions | | name | location_lat | location_long |
|---|---|---|---|---|
| Edit | Delete | 1_Ave_@_110_St | 40.79142678 | -73.93807411 |
| Edit | Delete | 1_Ave_@_124_St | 40.80042614 | -73.93155098 |
| Edit | Delete | 1_Ave_@_23_St | 40.73597417 | -73.97828579 |
| Edit | Delete | 1_Ave_@_42_St | 40.74803726 | -73.96948814 |
| Edit | Delete | 1_Ave_@_79_St | 40.77144187 | -73.95249367 |
| Edit | Delete | 1_Ave_@_86_St | 40.7760243 | -73.94931793 |
| Edit | Delete | 1_Ave_@_96_St | 40.783304 | -73.944662 |
| Edit | Delete | 1_Ave_@_E_14_St | 40.731331 | -73.982561 |
| Edit | Delete | 1_Ave_and_34_St | 40.74296517 | -73.97330761 |
| Edit | Delete | 11_Ave_@_34_ST | 40.75492947 | -74.0018034 |
| Edit | Delete | 11_Ave_@_42_St | 40.75990312 | -73.99815559 |
| Edit | Delete | 12_Ave_@_14_St | 40.74182715 | -74.0087986 |
| Edit | Delete | 12_Ave_@_22_St | 40.74771214 | -74.00789738 |
| Edit | Delete | 12_Ave_@_34_St | 40.75629482 | -74.00450706 |
| Edit | Delete | 12_Ave_@_42_St | 40.76126838 | -74.00090218 |
| Edit | Delete | 12_Ave_@_57_St | 40.77072685 | -73.99420738 |
| Edit | Delete | 2_Ave_@_110_St | 40.792531 | -73.9402627 |
| Edit | Delete | 2_Ave_@_125_St | 40.80195299 | -73.93339634 |

Camera Table in Database

# PROPOSED DATA

- The dataset has been collected from various sources (waze, new york city data from http://dotsignals.org, twitter data all collected over a period of one month) and organized in a database that can be queried for different analyses.

- Further, the waze table has 417990 records (number of reports), the images table has 27927 records (images) and the twitter data has 763890 records (tweets).

# APPROACH FOR TWEETS DATA

- For each ‘user id’, we can run one of the following steps and store the tweets category for each user. This will be any one of the following {ACCIDENT, CHIT_CHAT, HAZARD, JAM, MISC, POLICE, ROAD_CLOSED}.

- **Topic modeling using LDA**: As pointed out in [3], Topic Modeling with Latent Dirichlet Allocation (LDA) [2] is a popular unsupervised method for discovering latent semantic properties of a document collection.
    - document classification
    - clustering
    - information extraction.
- However, LDA is sensitive to noise
    - NLTK:
        - removing the stop words
    - language detection especially when there are short words in the message.
- From [3], topic modeling with LDA works best when there is little or no redundancy in the training data. So, as the tweets are limited to 140 characters and the number of topics to be modeled are K=7, this approach should give good results.

# APPROACH FOR TWEETS DATA

- **Backup-plan:** method suggested in [5], i.e. 'Short text language detection using infinity-gram'. Can detect 19 languages with 99% accuracy.

- **Third party APIs** e.g. MonkeyLearn API [6], that can categorize the tweets based on the events of interest.
  - higher accuracy among its peers (AlchemyAPI, Datumbox, Metamind etc).
  - Sample output:

# APPROACH FOR CAMERA DATA

- Common parameter in the ‘camera’ and ‘tweets’ data is latitude and longitude (location) of the camera and the user, so:
  - select all entries that pertain to a certain area i.e. (lat. + r, long. + r). Radius ‘r’ is tunable by the program.
    - Python geocoder library geopy can be used to manipulate the [longitude, latitude] data. No Twitter APIs needed here.
- Select from the tweets database all tweets that are pertaining to above location diameter.

- Now, in this set of tweets check if the tweets are talking about any of these events of interest {ACCIDENT, CHIT_CHAT, HAZARD, JAM, MISC, POLICE, ROAD_CLOSED}. If so, we form clusters of tweets belonging to each of these categories in the selected longitude and latitude zone.

# FINAL DESIGN

- So, From step 1:
  - tweet and tweet-category of all the users.

- From step 2,
  - tweets pertaining to all the geographical areas of interest to us (this is tunable by the program).

**This information can be further used by the program to create a visualization of the tweets and camera data in real-time as per the goal of the project.**

# DATA/EQUIPMENT NEEDED

A workstation or laptop installed with all the required python libraries and connection to the database that stores the dataset.

Depending on the actual implementation, the python libraries that need to be installed will differ.

# RESULT EVALUATION

- As this is a non-classic, non-trivial problem, it is difficult to measure or find the ground truth for  such a large dataset. However, the predictions of any machine learning algorithm can have errors:
  - Miss-detection (for e.g. there were some accidents which the algorithm could not detect)
  - False alarms (for e.g. the algorithm predicted there was an accident when there was none). Miss-detections are hard to find manually for this large dataset. One alternative is to consider some sub-samples for some of the predictions and check the number of false alarms.

- Performance measures can be the detection latency of events by the program as compared to those actually reported (DOT can serve as the ground-truth provider in this case).

# FUTURE WORK

- Multiple Tweets by the same user in different geo locations.
- Tweet representation on the map - APIs, what if too many tweets.
- Load testing the program.
- Not real-time, what does it take to go live?

# REFERENCES


1. **https://geopy.readthedocs.io/en/1.10.0/**
2. **Blei D, Ng A, Jordan M, n.d., Latent Dirichlet allocation. Journal of Machine Learning Research 3: 993–1022**
3. **Cohen et al., "Redundancy-Aware Topic Modeling for Patient Record Notes."**
4. **Blog post: http://alexperrier.github.io/jekyll/update/2015/09/04/topic-modeling-of-twitter-followers.html**
5. **Blog post: https://shuyo.wordpress.com/2012/02/21/language-detection-for-twitter-with-99-1-accuracy/**
6. **http://docs.monkeylearn.com/article/api-reference/**
7. **Wang and Kankanhalli, "Tweeting Cameras for Event Detection."**
8. **Kam-Yiu Lam et al., "RETINA: A REal-time TraffIc NAvigation System."**