On the importance of keywords for the application of Twitter posts for traffic incident detection

Camille Kamga

Anil Yazici

Sandeep Mudigonda

Wei Hao

Nathalie Martinez

Traffic Incidents

- Roadway incidents → 57.9% of the total delay on road networks.
- Improve roadway geometric design for safer driving
- Mitigate incident impacts:

→ 1 min less incident duration → 4-6 min/vehicle delay saving & 9 gal fuel, 0.7 kg HC, 9 kg CO, 1.3 kg NO)

→ Reduce detection and clearance times

  - **Gather** and disseminate the **incident information** fastest way possible efficient response

Crowdsourced social media (Twitter) data can help

  - Harvest the information content of crowd-sourced online Twitter feeds
  - Use as an incident management (IM) support tool

Oak Ridge National Lab Report by Shih-Miao et al., 2004 Texas Transportation Institute, 2012

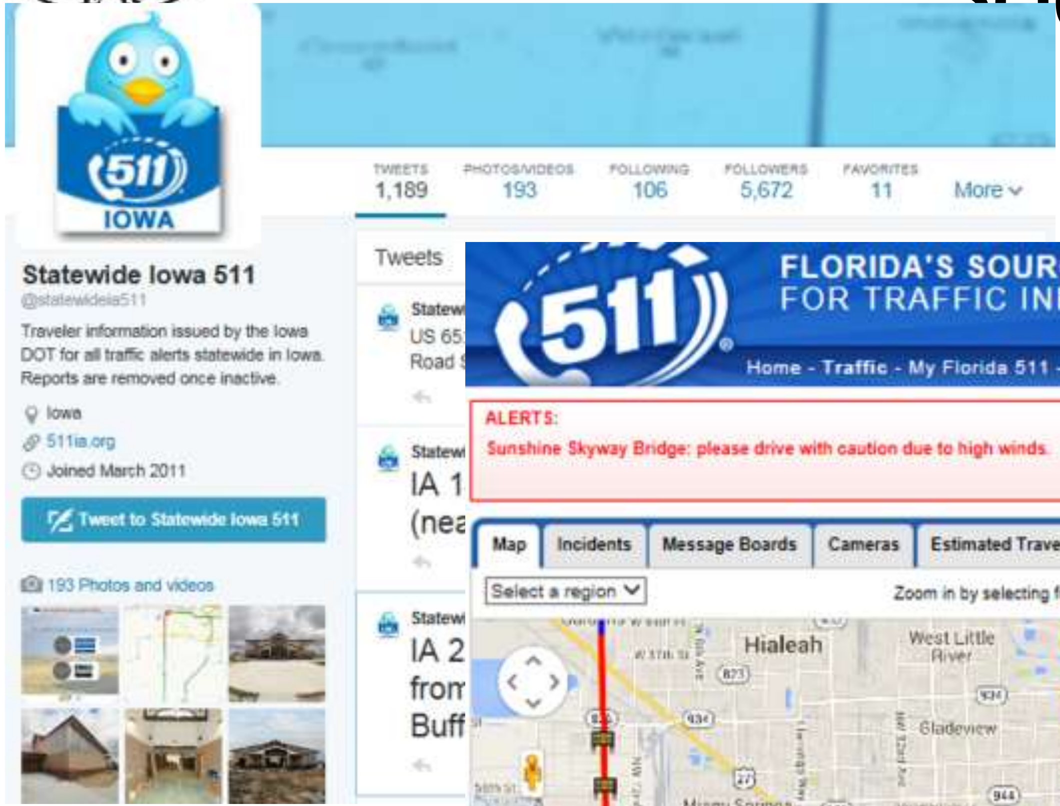
Use of Social Media

- Web 2.0 → user generated content → everybody is a "reporter"

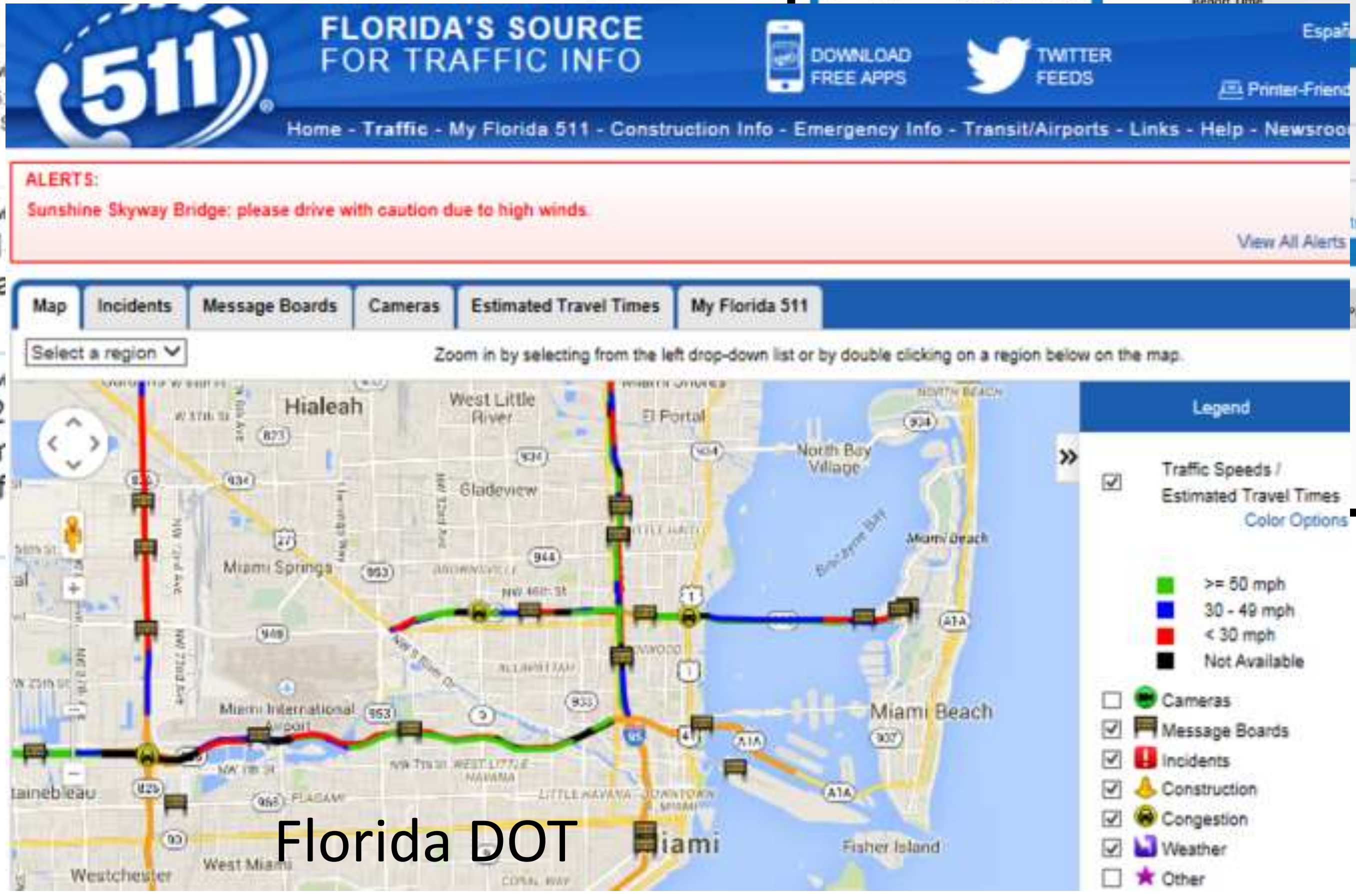
Social media feeds as information source

- Brand adoption; Political public opinion; "meet up";
- Monitor disease outbreaks; Disaster information
- Transportation
  - Surveys: policy, demand, etc.
  - Transit service disruptions real-time interaction
  - Potential for extracting real-time information

Transportation Agency Adoptions of Social Media



Iowa DOT



Florida DOT

Utah DOT

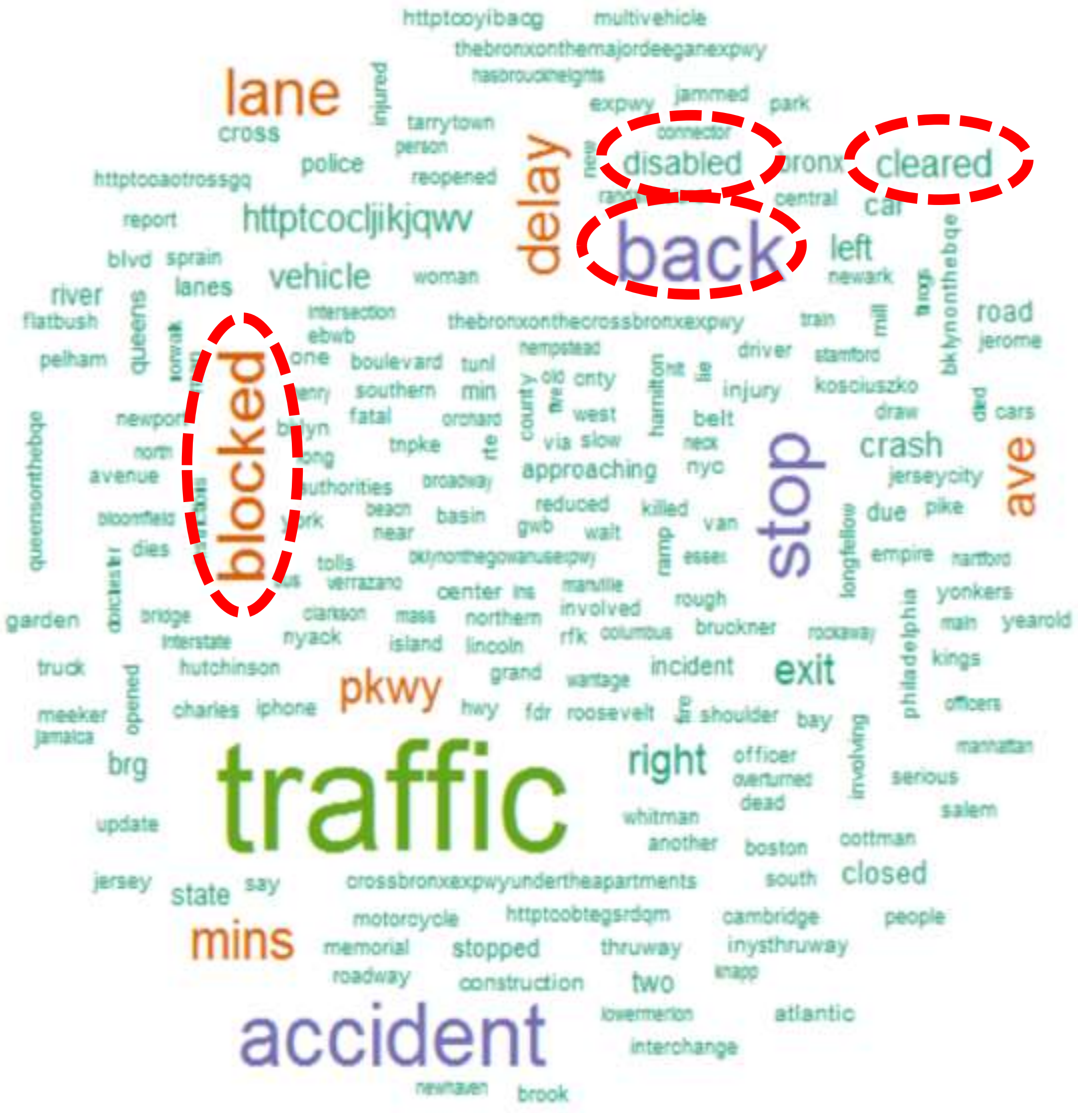
Information Extraction from Social Media

- ”needle in a haystack” problem (Grant-Muller et al., 2014).
- Natural language form → 80% unstructured (Liu et al., 2011),
    - Ungrammatical, abbreviated
- Approach:
    1. **Information retrieval**: query-based
    2. **Information extraction**: text → relevant information
    - “Dictionary” → List of common words → best “candidate” tweets
    - Context dependent, different set for different purposes
    - Lack/ambiguity of context → challenge! (Pereira et al., 2014)
    3. **Prediction**: extracted information → predict future transportation states

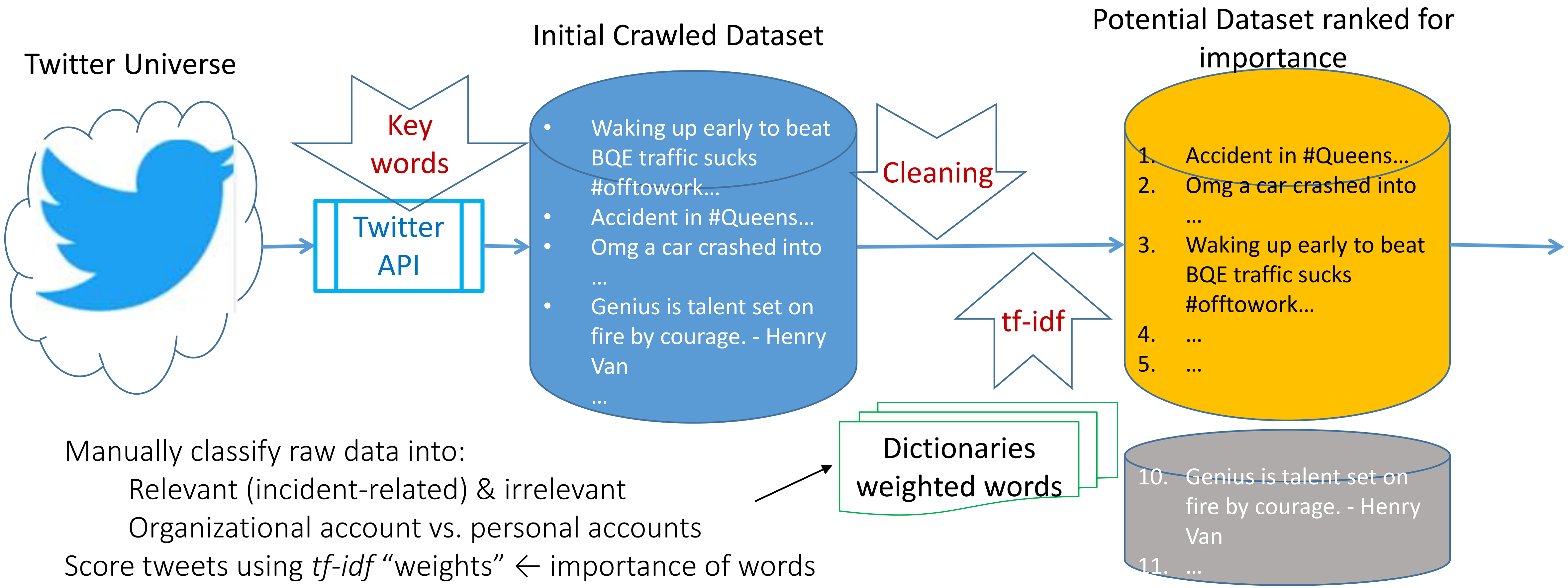
- Most “ … 511, DOT
- Early d …
  - Usu …
  - Imp … eets

  →Di …

  →



Proposed Methodology



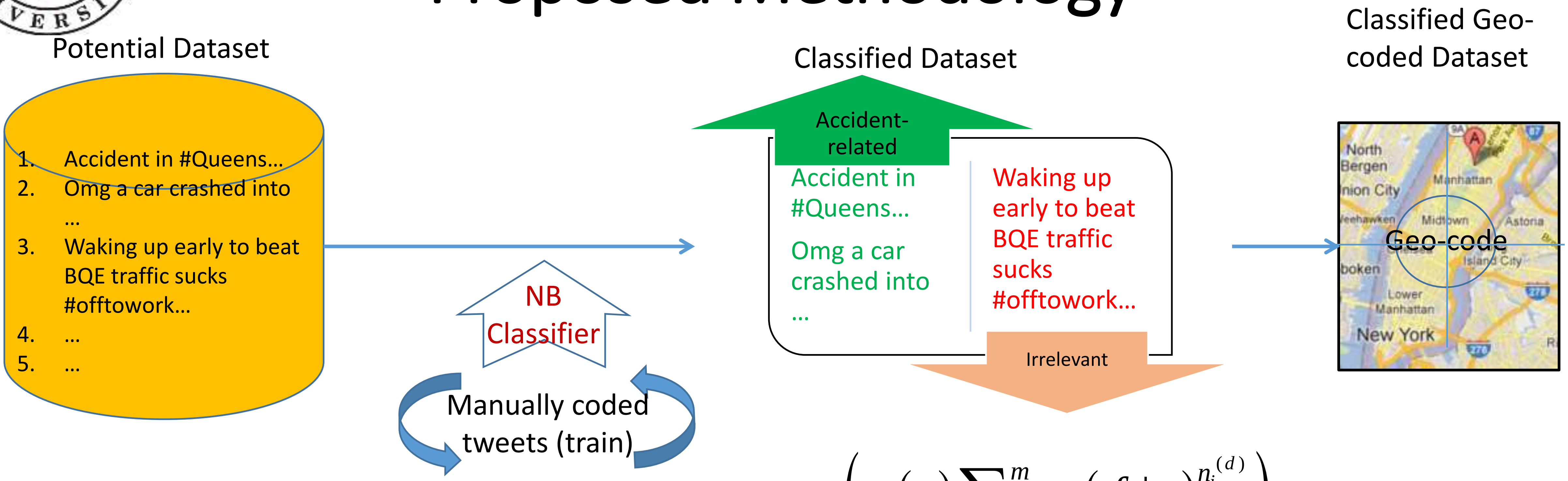
Manually classify raw data into:

- Relevant (incident-related) & irrelevant
- Organizational account vs. personal accounts

Score tweets using *tf-idf* “weights” ← importance of words

$$tf(t,d)=\frac{f(t,d)}{\max\{f(w,d):w\in d\}} \quad idf(t,D)=\log\left(\frac{N}{|\{d\in D:t\in d\}|}\right)$$

Proposed Methodology



- Naïve-Bayesian (NB) Classifier

$$P_{NB}(c\,|\,d) := \frac{\left(p(c)\sum_{i=1}^{m} p(f\,|\,c)^{n_i(d)}\right)}{P(d)}$$

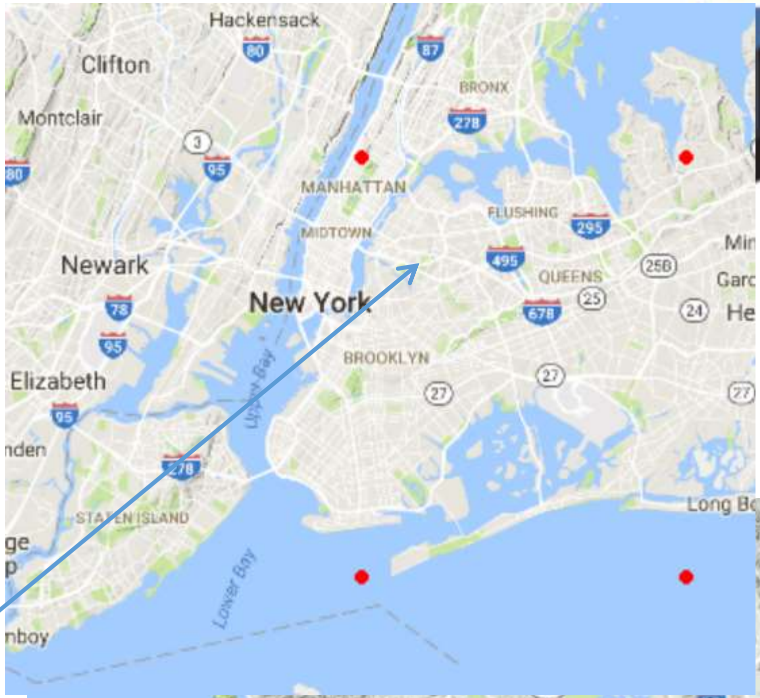
→What is the probability that a tweet is relevant given that it includes “car” and “crash”?

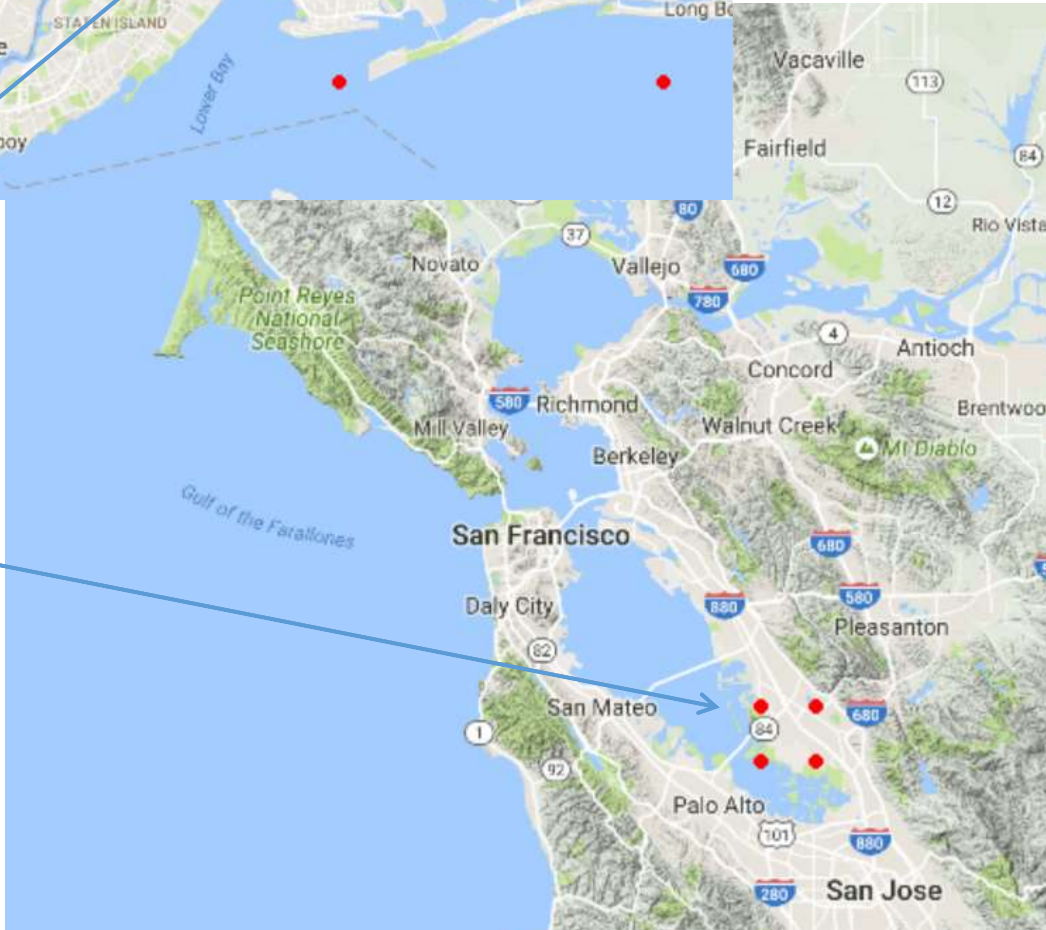
- NB for each account type (Organizational vs. personal)

Geocoding

- < 3% tweets have accurate geo-location

Account	Tweet text	Geocode Reported
@TotalTraffic NYC	Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of 6 mins #traffic	-73.9626, -73.9626, -73.6998, -73.6998, 40.5417, …,
@sfgiantsfan1	@KTVU there was a high speed crash on Thornton ave in Newark car flipped several times before bursting into flames	-122.0731, -122.0731, -121.9876, -121.9876, 37…
@511NY	Accident with property damage on #US9 NB at Montrose station rd	-73.9535, -73.9535, -73.9166, -73.9166, 41.2298, …,





Geocoding

- Regular expressions (ave, pkwy, hwy, st, rd, at, near, between…)
- Hastags (#Queens)
- Location

	Tweet text	Geocode Reported	Location
@TotalTrafficNYC	Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of 6 mins #traffic	-73.9626, -73.9626, -73.6998, -73.6998, 40.5417, …,	Queens, NY
@	@KTVU there was a high speed crash on Thornton ave in Newark car flipped several times before bursting into flames	-122.0731, -122.0731, -121.9876, -121.9876, 37…	Newark, CA

Impact of dictionaries

6900 randomly selected public tweets collected using Twitter API.

Manually coded raw data:

- incident-related & irrelevant
- Organizational vs. personal

$$Normalized\ tfidf(S) = \frac{\sum_{for\ all\ t\ in\ d} tfidf(t,d)}{\sum_{t \in S} t}$$

Organization accounts	Personal accounts
"exit "ave" "accident" "lane" "block" "delay" "min" "pkwy" "traffic" "right" "back" "stop" "crash" "clear" "close" "left" "vehicle" "road" "disable"	"accident" "just" "car" "traffic" "got" "bridge" "block" "crash" "highway" "thank" "get" "road" "today"

Filtered based on a 20th percentile of normalized *tf-idf*	Organizational tweets	Personal tweets	Total
Organizational dictionary	435	4	439
Personal dictionary	409	49	458
Organizational + personal dictionary	469	18	487

Impact of dictionaries

Relevant tweet		Account type	Using organizational + personal keywords	Using only organizational keywords	Using only personal keywords
#1	State troopers just blocked the ramps leading from route 138 in Canton onto 93 due to serious crash #WCVB	Agency	0.27	0.27	0.8
#2	Omg a car crashed into the paramus Wendy's @amandabootsy http://t.co/C4DwTEIyHN	Personal	0.2	0.16	0.4
#3	@crosattto it was a bad wreck that a car went straight into the wall and went up in flames. http://t.co/XCvA7QkAF8	Personal	0.04	0	0.1
#4	car on fire on Lower level of Verrazano Bridge. 🚙🔥🚒🚨💦 @ Verrazano Bridge Tolls https://t.co/lpEPEGGXWn	Personal	0.34	0	1.5

Classification using different dictionaries

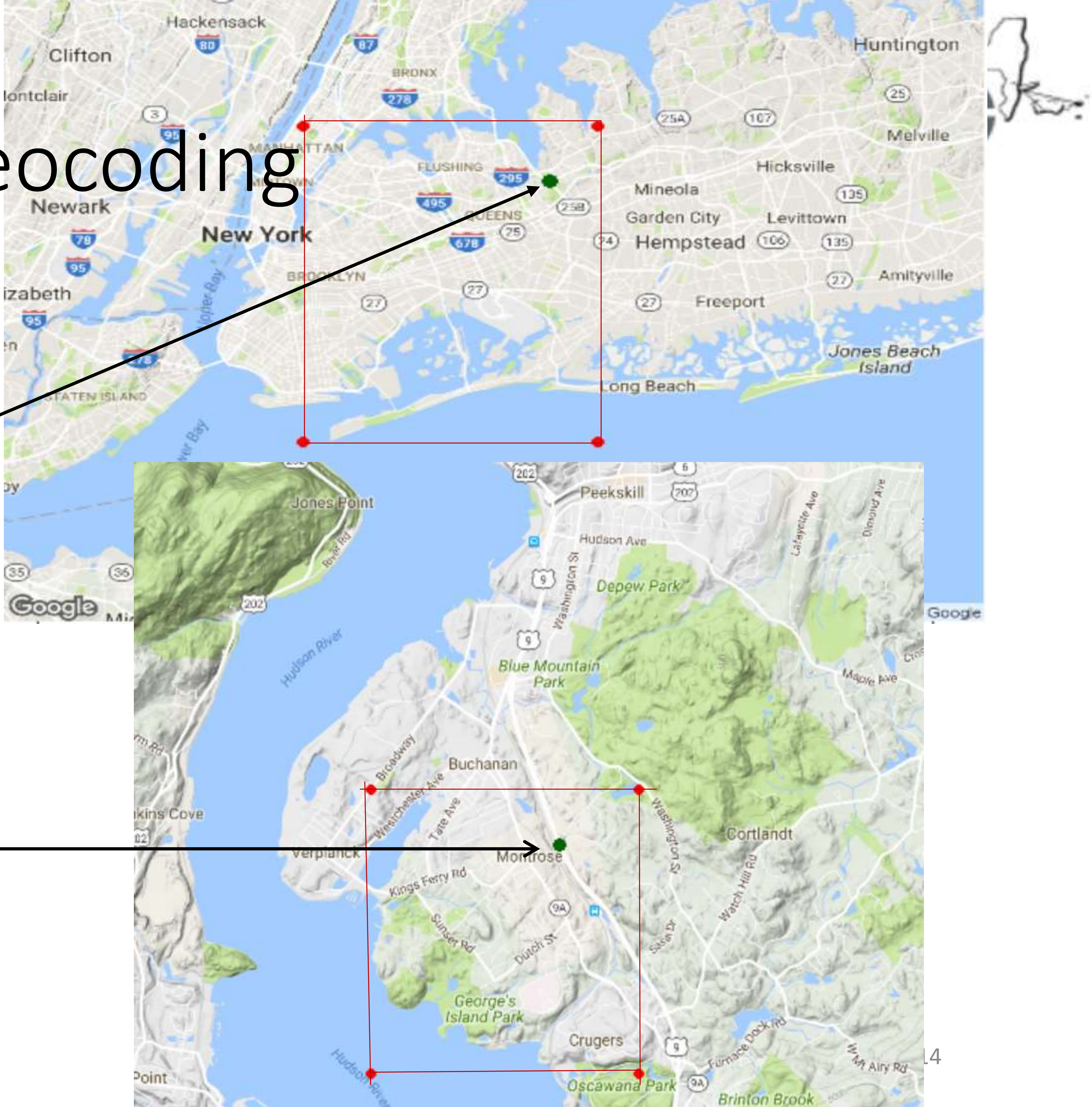
- Raw data → 80% training, 20% test
- NB_{org} using only organizational dictionary.
- NB_{all} using organizational and personal dictionary.
- NB_{pers} using only personal dictionary.

Classifier	Accuracy in predicting relevant tweets
NB_{org}	75.6%
NB_{all}	85.5%

Classifier	Accuracy in predicting relevant *personal* tweets
NB_{org}	50.5%
NB_{all}	54%
NB_{per}	74.4%

Geocoding

Account	Tweet text	Geocode Reported
@TotalTrafficNYC	Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of 6 mins #traffic	-73.9626, -73.9626, -73.6998, -73.6998, 40.5417, …,
@511NY	Accident with property damage on #US9 NB at Montrose station rd	-73.9535, -73.9535, -73.9166, -73.9166, 41.2298, …,



Summary

- All incident information is useful for early detection
- Dictionaries derived from prominent accounts give lesser importance to personal accounts
- Personal dictionaries are more effective in
  - Filtering potentially useful tweets
  - Classification of relevant tweets
- Geocoding requires analysis of regular expressions, hashtags, location of account, neighborhood information

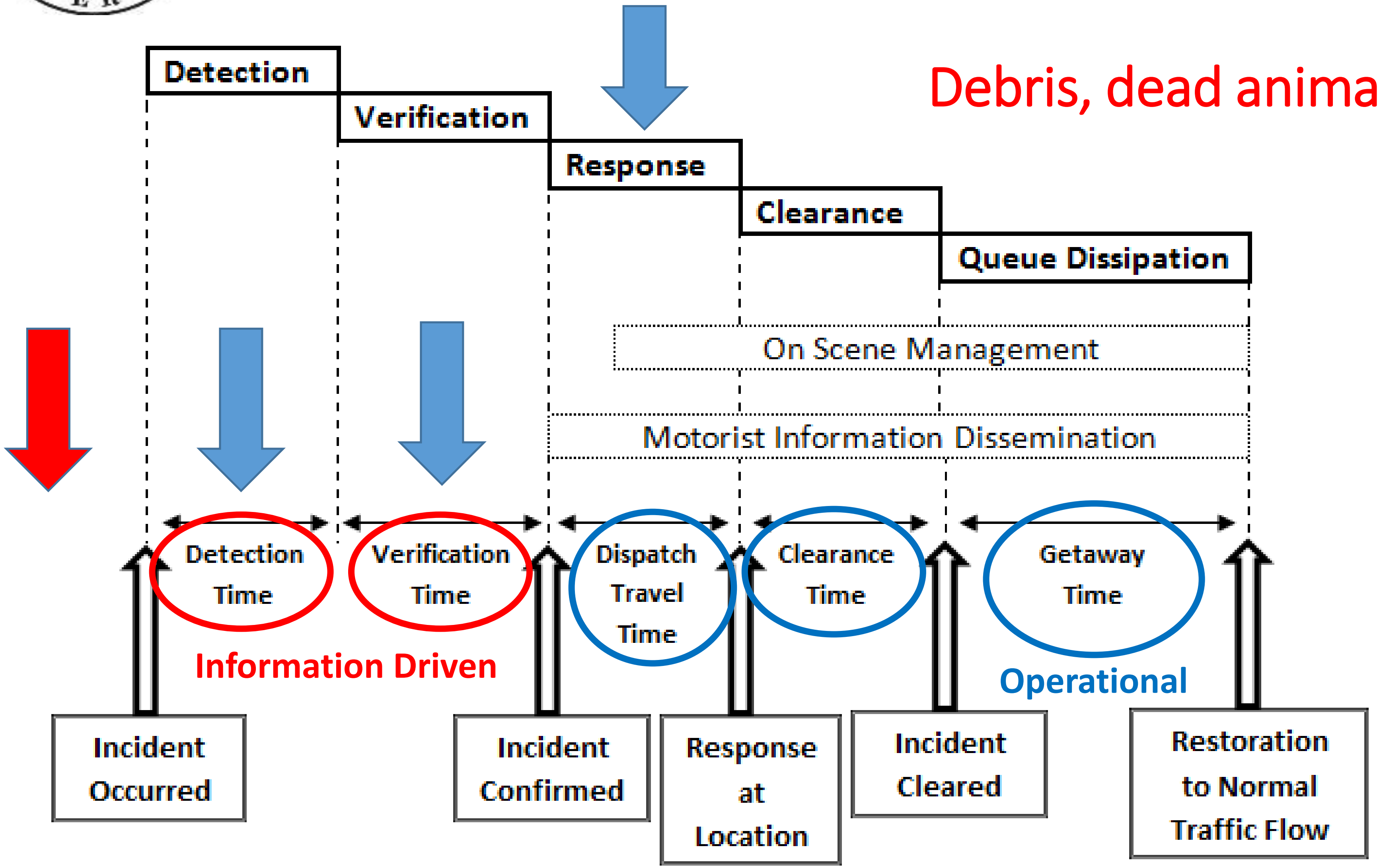
Remarks

- More raw data for personal tweets
- Extra effort for identifying personal & organization (automated)

- IM → incidence, location and time
  - Geo-coding : 3% on all tweets
  - Further text analysis
  - Time of tweet not always incident time

Future Potential

Debris, dead animal →Accident prevention!



New York State
Department of
Transportation

Thank you! @nyserda @nysdot #Questions?

Energy. Innovation. Solutions.