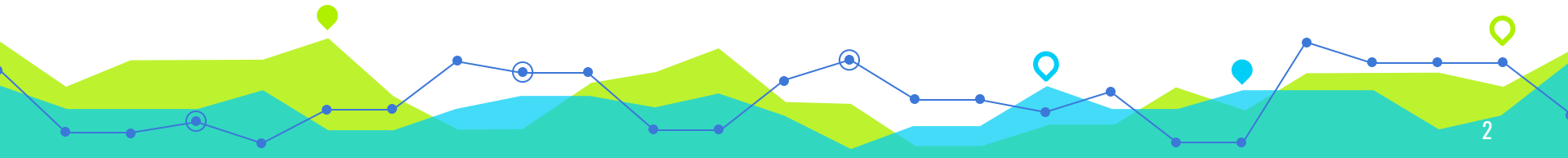
Road Closure Location Extraction From Twitter

Temple Moore, Nathan Jacques, and David Trichter
Data Scientists

OUR PROBLEM:

Leveraging social media sources such as Twitter, we want to identify real time road closures, damaged roads, traffic congestion, flooding, and other blocked routes that may affect travel time, travel safety, and accessibility to emergency response crews.

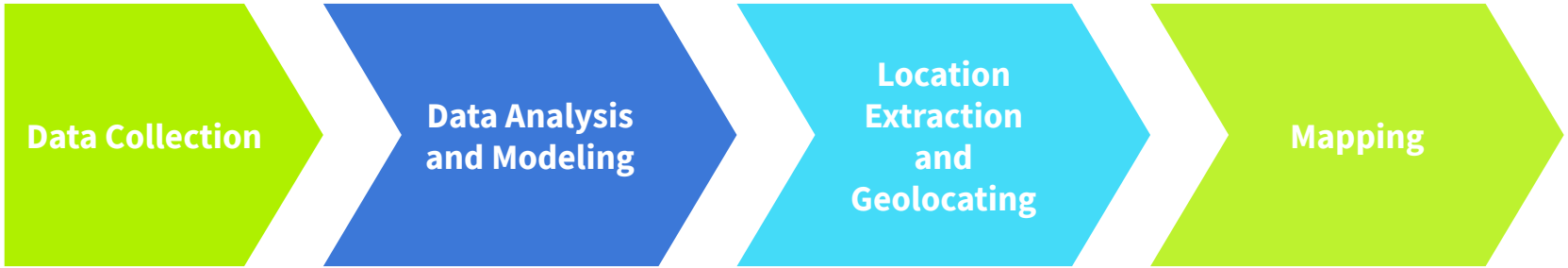


Case study:
HURRICANE MATTHEW

- Hurricane Matthew claimed eleven lives in the state of Florida.
- Fatalities were noted in Volusia, Miami-Dade, Duval, Orange, and Putnam Counties.
- The most substantial damage occurred in Volusia, Flagler, St. Johns, Duval, and Nassau counties. More than 12,000 homes in Volusia sustained damage.
- Total economic damage was estimated to exceed $2.0 billion USD in the state.

Matthew's best track positions and intensity (Source: NHC)

OUR WORKFLOW



METHODOLOGY

Data Acquisition

- Twitter API
- HERE API
- 511 Twitter Accounts
- Proprietary Interstate Exit Data

EDA and Modeling

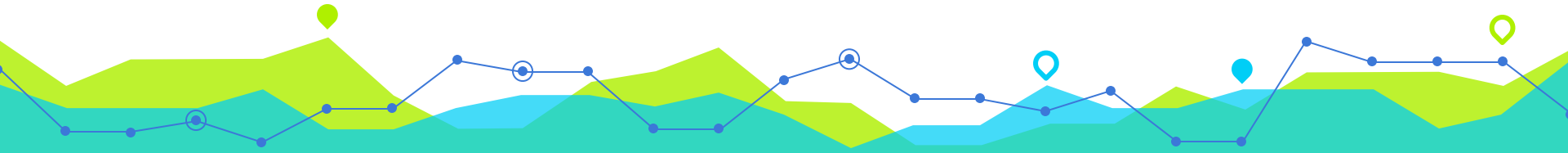
- TF-IDF, Count Vectorizer
- Keyword Filtering
- Logistic Regression, Gradient Boosting

NER and Geolocation

- SpaCy Named Entity Recognition
- RegEx
- Geolocating Twitter Text

Mapping

- HERE API
- Google Maps API



Data Collection

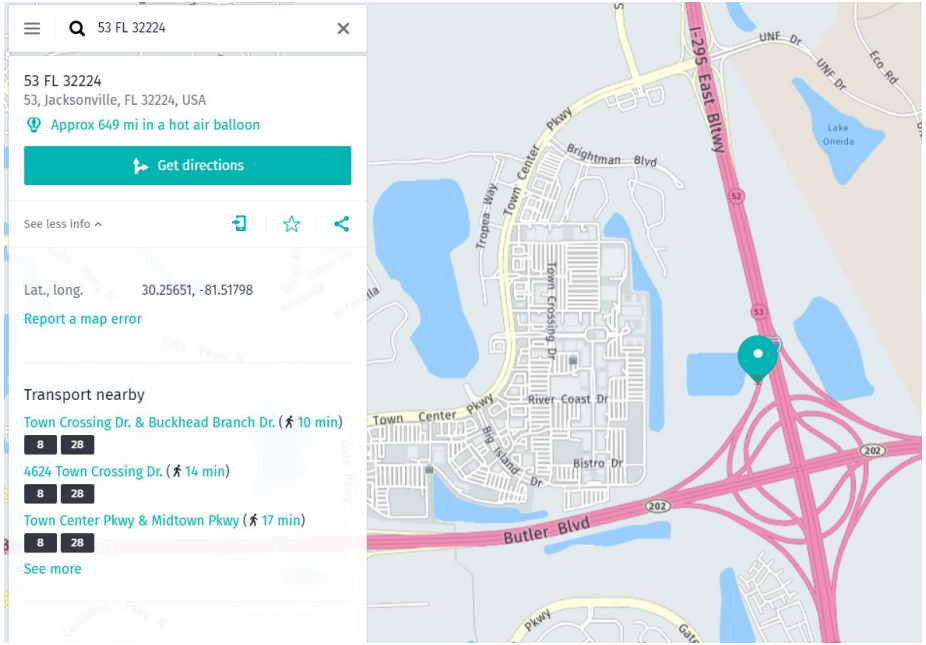
1

Twitter API and Road Exits

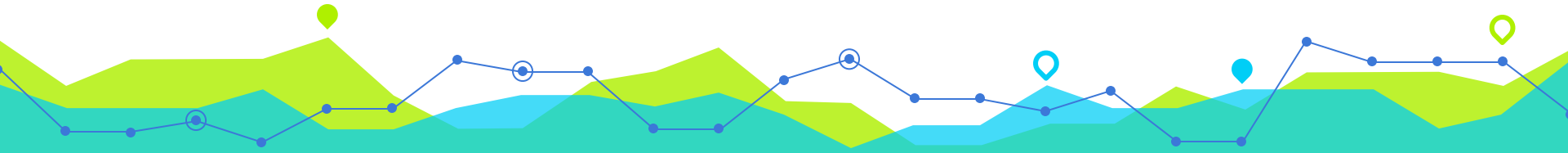
TWITTER DATA

- Twitter API
- Extracted over 24,000 historical tweets from state-run 511 Twitter accounts using Tweepy.
- Historical Tweets dated from of October 4, 2016 - October 14, 2016.
- Curated lists on Twitter provide accurate and reliable data

RETRIEVING INTERSTATE EXITS



- 511 Tweets are formatted reliably; containing road names, intersections, and exits
- No efficient or cheap way to collect geolocated interstate exits or cross streets
- Manual geolocation limited to region



Data Analysis and Modeling

Keyword Classification and Supervised Models

2

MODELING AND PREPROCESSING

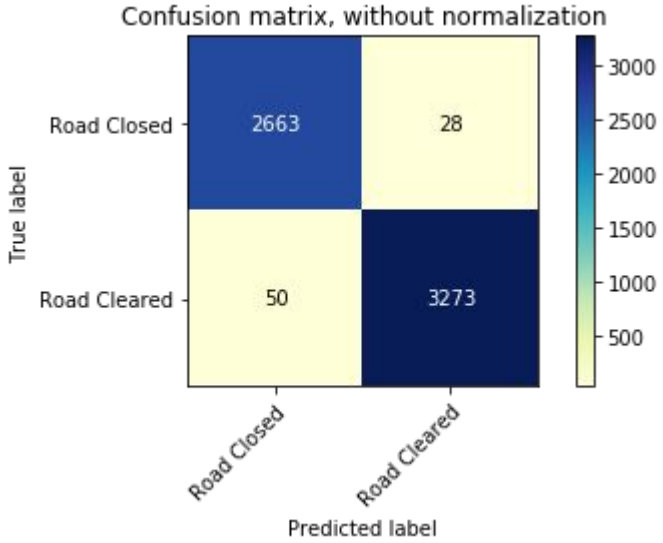
- Over 24,000 historical Tweets were cleaned with RegEx
- Classified each tweet as “closed” based on keyword classification
- Logistic Regression for interpretability, Gradient Boosting to curb overfitting
- Trained models on historic tweets, evaluated performance on real-time tweets

KEYWORD CLASSIFICATION

- **Tweets from reliable and verified sources contained similar language**
- Words such as “closed”, “flood”, “closure”, “disabled” appeared in tweets frequently
- Filtered out Tweets with words that would cause false positives: “cleared”, “opening”, “lifted”

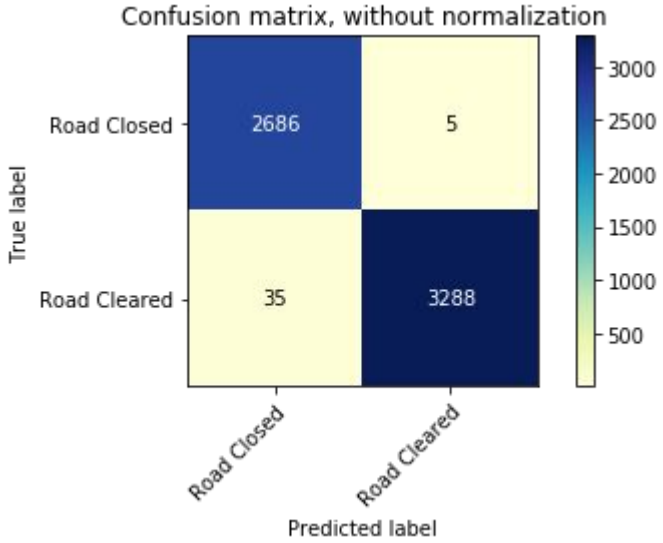
LOGISTIC REGRESSION

- Count Vectorized words
- **Accuracy score** of **99.92%** on the training set, and **99.90%** on the testing set
- The **ROC AUC** of this model was **0.9872**
- **Sensitivity** of **98.95%** and a **Specificity** of **98.49%** on the testing set.



GRADIENT BOOSTING

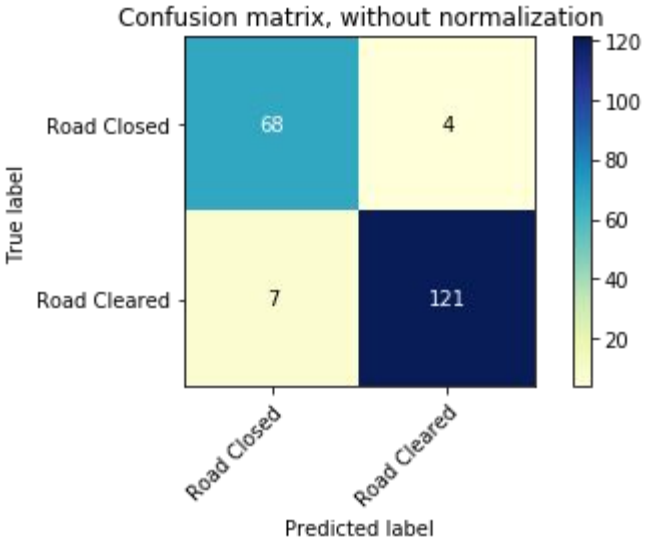
- TF-IDF Vectorized words
- **Accuracy score** of **99.99%** on the training set, and **99.96%** on the testing set
- The ROC AUC of this model was 0.9938
- **Sensitivity** of **99.81%** and a **Specificity** of **98.94%** on the testing set.



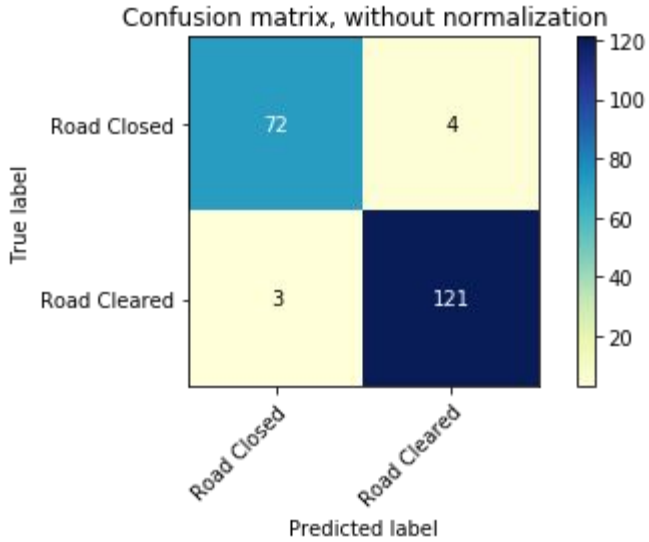
REAL-TIME DATA

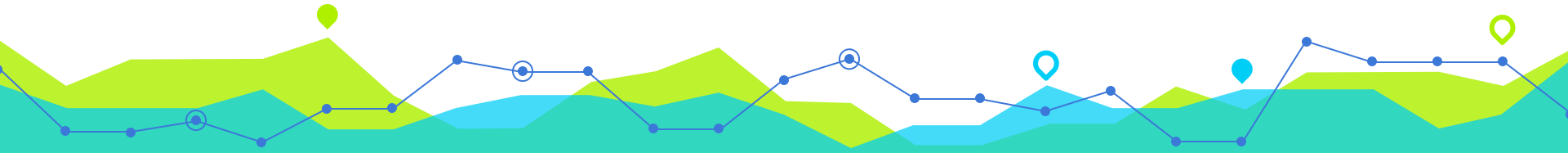
- Using models trained on historical tweets, we evaluated their performance on tweets taken real time.
- The Logistic Regression Model had an accuracy of 93.33% on testing data, a Sensitivity of 95.65%, and a Specificity of 91.89%.
- The Gradient Boosted Model had an accuracy of 96.67% on testing data, a Sensitivity of 100%, and a Specificity of 94.59%.
- These models can reliably be run on Tweets taken real time from curated lists

LOGISTIC REGRESSION



GRADIENT BOOSTING





NER and Geolocation

3

SpaCy Entity Recognition and Location Matching

spaCy

NAMED ENTITY RECOGNITION

Locates and classify named entity mentions in unstructured text into pre-defined categories such as the place names, organizations, locations, etc.

NAMED ENTITY RECOGNITION

Atlantic Blvd SR 10 LOC closure detour in Jacksonville GPE # Matthew PERSON

IN DUVAL COUNTY GPE SR GPE -A1A IS CLOSED SOUTH OF BUTLER BLVD SR-202 DUE ORG TO SEVERE WEATHER AND FLOODING PLEASE SEEK AN ALTERNATE ROUTE

HOW TO GET LOCATIONS FROM TWEETS

SpaCy NER

SpaCy locates named entities within the Tweet body, extracts them for use in a mapping search query

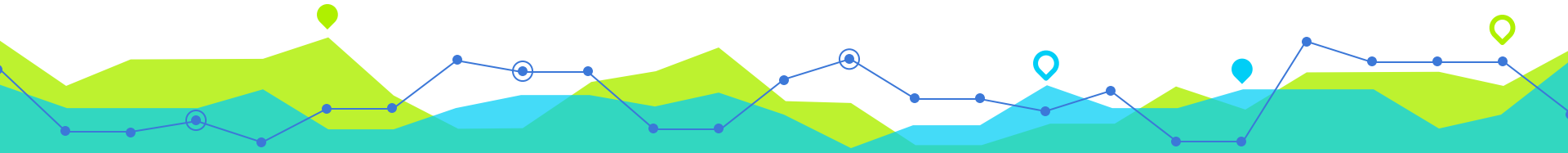
Exit Data

Using our proprietary map of exit coordinates in Florida, map tweets based on matching Interstate and Exit numbers

Intersections and Full Text

If two roads are extracted, find roads in exit dataset or NER road names, or simply use full text of tweet

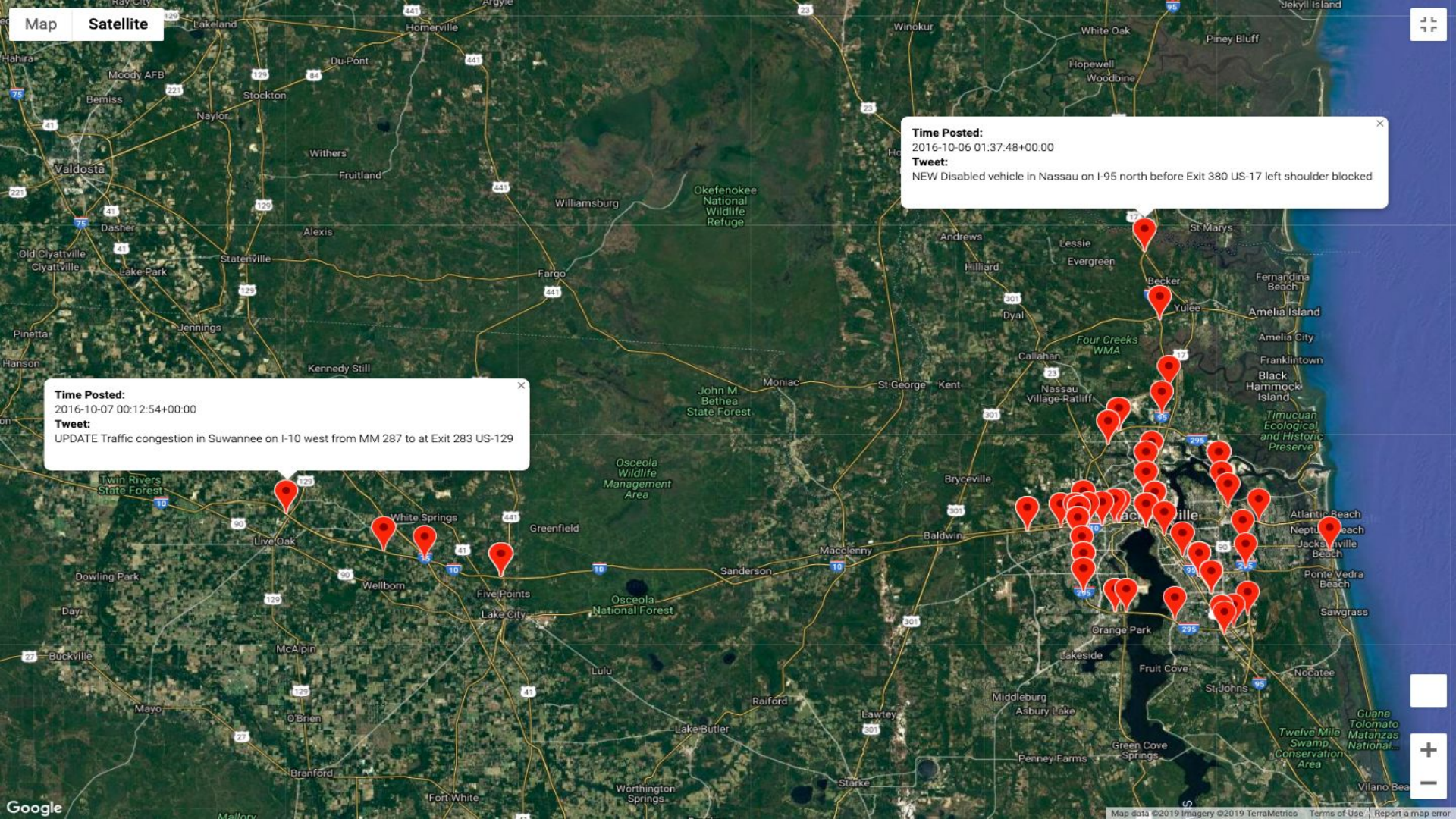

WHICH TWEETS WERE PLOTTED?



Mapping

4

Using Extracted Locations, Coordinates, and Google Maps API


Map
Satellite
Time Posted:
2016-10-06 01:37:48+00:00
Tweet:
NEW Disabled vehicle in Nassau on I-95 north before Exit 380 US-17 left shoulder blocked
Time Posted:
2016-10-07 00:12:54+00:00
Tweet:
UPDATE Traffic congestion in Suwannee on I-10 west from MM 287 to at Exit 283 US-129
Google
Map data ©2019 Imagery ©2019 TerraMetrics
Terms of Use
Report a map error

LEVERAGING API TECHNOLOGIES

- Combined Herepy and the Here API with the gmaps for Google Maps.
- Here.com had more accurate search results for mapping preliminary test searches.
- A Google Map output is a more comfortable platform for end user

USER FEATURES

- Search by user-defined region, zoom level, map type.
- Returns an output DataFrame for future analysis of which tweets were actually mapped
- Returns an interactive google map allowing for deeper understanding of patterns and importance of tweets

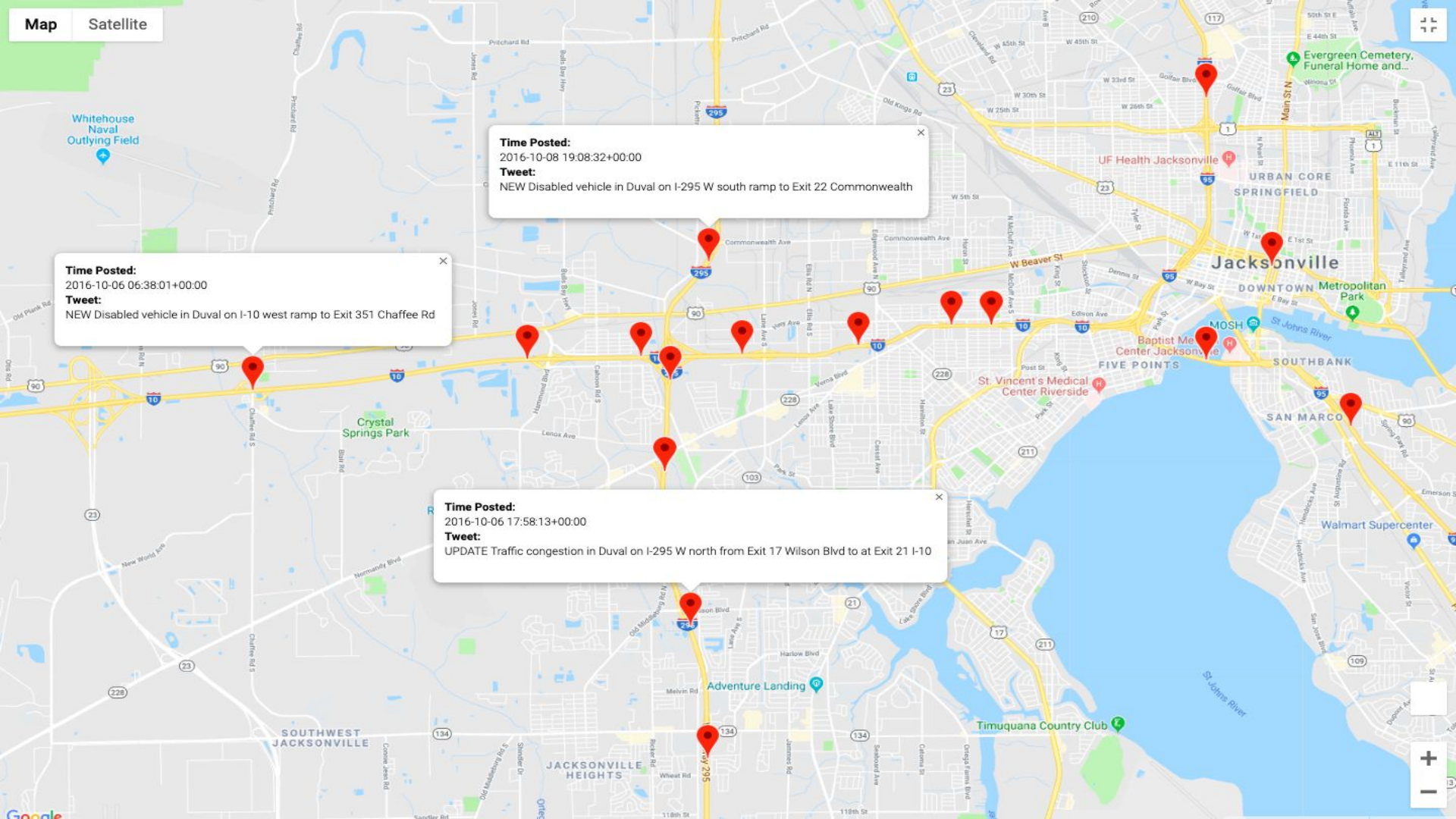
IMPORTANCE OF PRECISION

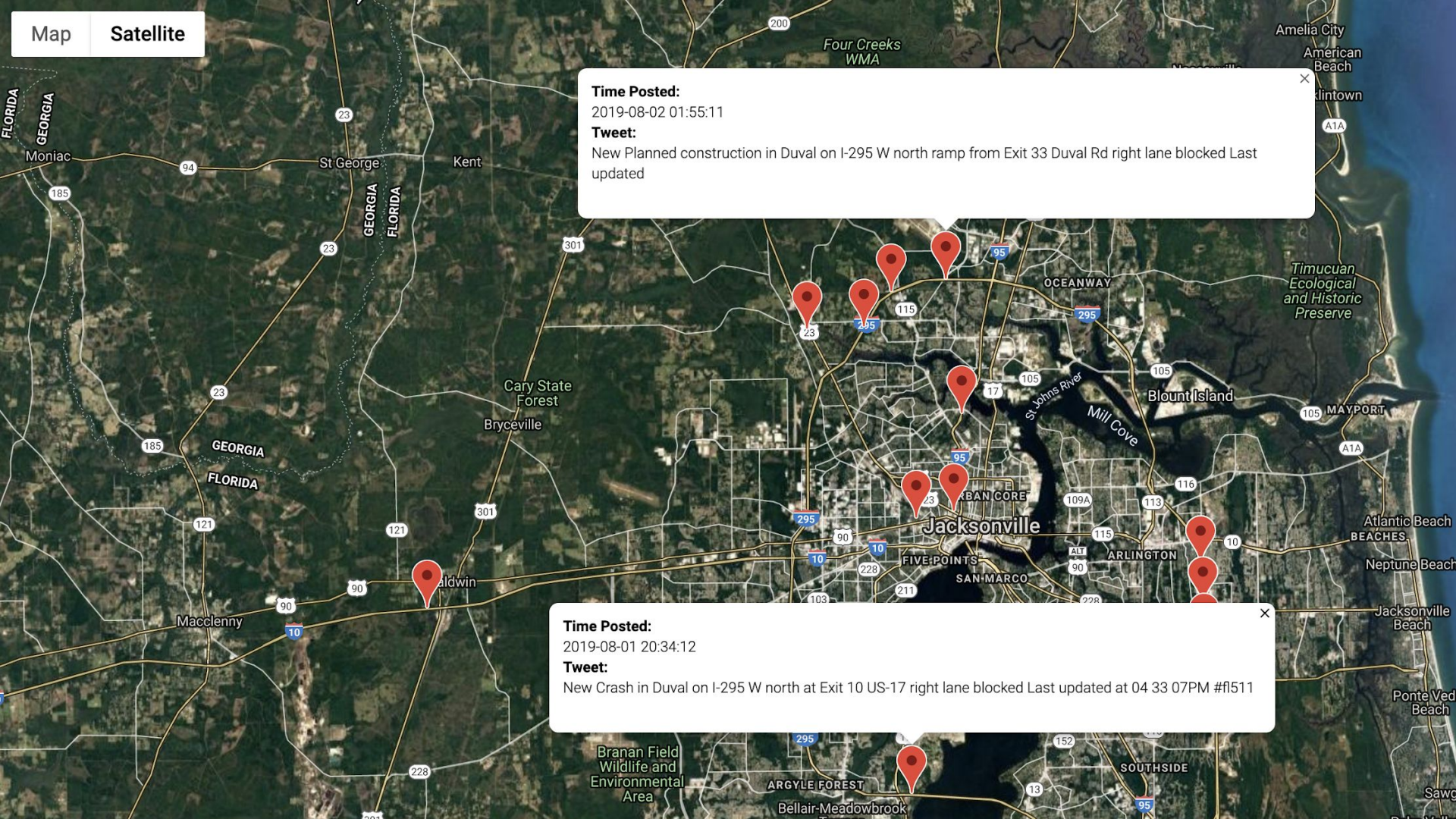
- 28 False Positives along with 143 True Positives resulted in a **precision of 83.6%** for our test set.

FALSE POSITIVES MAP

IMPORTANCE OF PRECISION 2

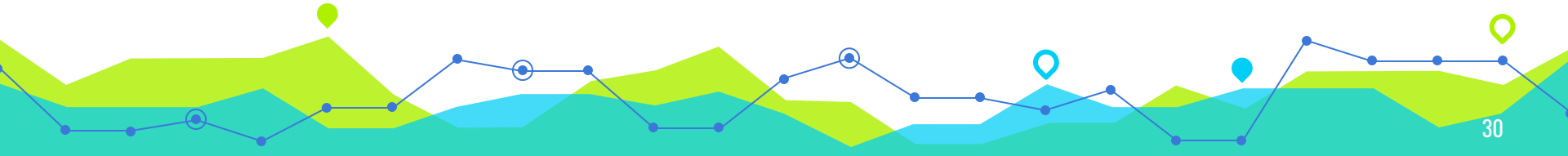
- Of the 143 True Positives, the vast majority were matched within a few feet of the location tweeted.
- Allows for this data to be practically used to avoid specific exits on the highway, or respond to an emergency


Map
Satellite
Time Posted:
2016-10-08 19:08:32+00:00
Tweet:
NEW Disabled vehicle in Duval on I-295 W south ramp to Exit 22 Commonwealth
Time Posted:
2016-10-06 06:38:01+00:00
Tweet:
NEW Disabled vehicle in Duval on I-10 west ramp to Exit 351 Chaffee Rd
Time Posted:
2016-10-06 17:58:13+00:00
Tweet:
UPDATE Traffic congestion in Duval on I-295 W north from Exit 17 Wilson Blvd to at Exit 21 I-10
Whitehouse Naval Outlying Field
Crystal Springs Park
Jacksonville
DOWNTOWN
URBAN CORE
SPRINGFIELD
FIVE POINTS
SOUTHBANK
SAN MARCO
Metropolitan Park
UF Health Jacksonville
St. Vincent's Medical Center Riverside
Evergreen Cemetery, Funeral Home and...
Walmart Supercenter
Adventure Landing
Timuquana Country Club
SOUTHWEST JACKSONVILLE
JACKSONVILLE HEIGHTS
W Beaver St
St Johns River
Commonwealth Ave
Lenox Ave
Harlow Blvd
Melvin Rd
Wheat Rd
Google


Map
Satellite
Time Posted:
2019-08-02 01:55:11
Tweet:
New Planned construction in Duval on I-295 W north ramp from Exit 33 Duval Rd right lane blocked Last updated
Time Posted:
2019-08-01 20:34:12
Tweet:
New Crash in Duval on I-295 W north at Exit 10 US-17 right lane blocked Last updated at 04 33 07PM #fl511
Four Creeks WMA
Amelia City
American Beach
Moniac
St George
Kent
FLORIDA
GEORGIA
Timucuan Ecological and Historic Preserve
OCEANWAY
Cary State Forest
Bryceville
Blount Island
St Johns River
Mill Cove
MAYPORT
Jacksonville
FIVE POINTS
SAN MARCO
ARLINGTON
Atlantic Beach
BEACHES
Neptune Beach
Jacksonville Beach
Macclenny
Baldwin
Branan Field Wildlife and Environmental Area
ARGYLE FOREST
Bellair-Meadowbrook
SOUTHSIDE
Ponte Vedra Beach

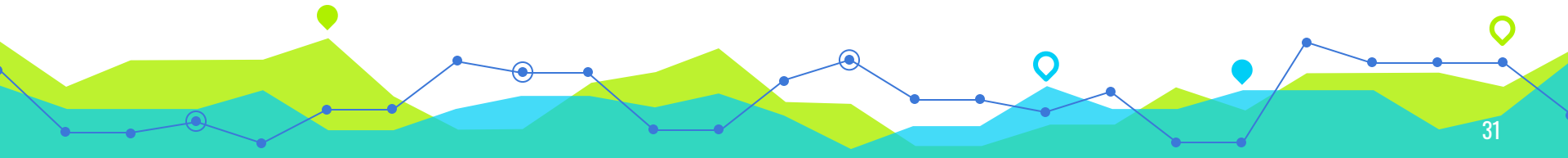
CAVEATS

- Only 60% of our road closure classified data made it onto the map.
- Geolocating Interstate and Highway exits or mile markers is inefficient or expensive
- Despite accurate classification, mapping Tweets to 100% accuracy is difficult due to variance in language, search algorithms
- SpaCy is powerful but computationally expensive
- We would like to deploy all pieces of this platform together, but some modules have incompatibility issues



CONCLUSIONS AND RECOMMENDATIONS

- Reliable and parsable data is difficult to find on Twitter, but official accounts can be trained to provide machine-interpretable data
- Despite using a limited geolocation dataset, we were able to still use the text of Tweets to find road closures, even in real time
- Classification methods used were very accurate
- Further Iterations would implement cross street and intersection functionality (one debugging session away), and to increase number of regions/states/languages



THANK YOU!

Any questions?