THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

A Decision Support System for Effective Police Patrolling

August 2019

Varsha Waingankar

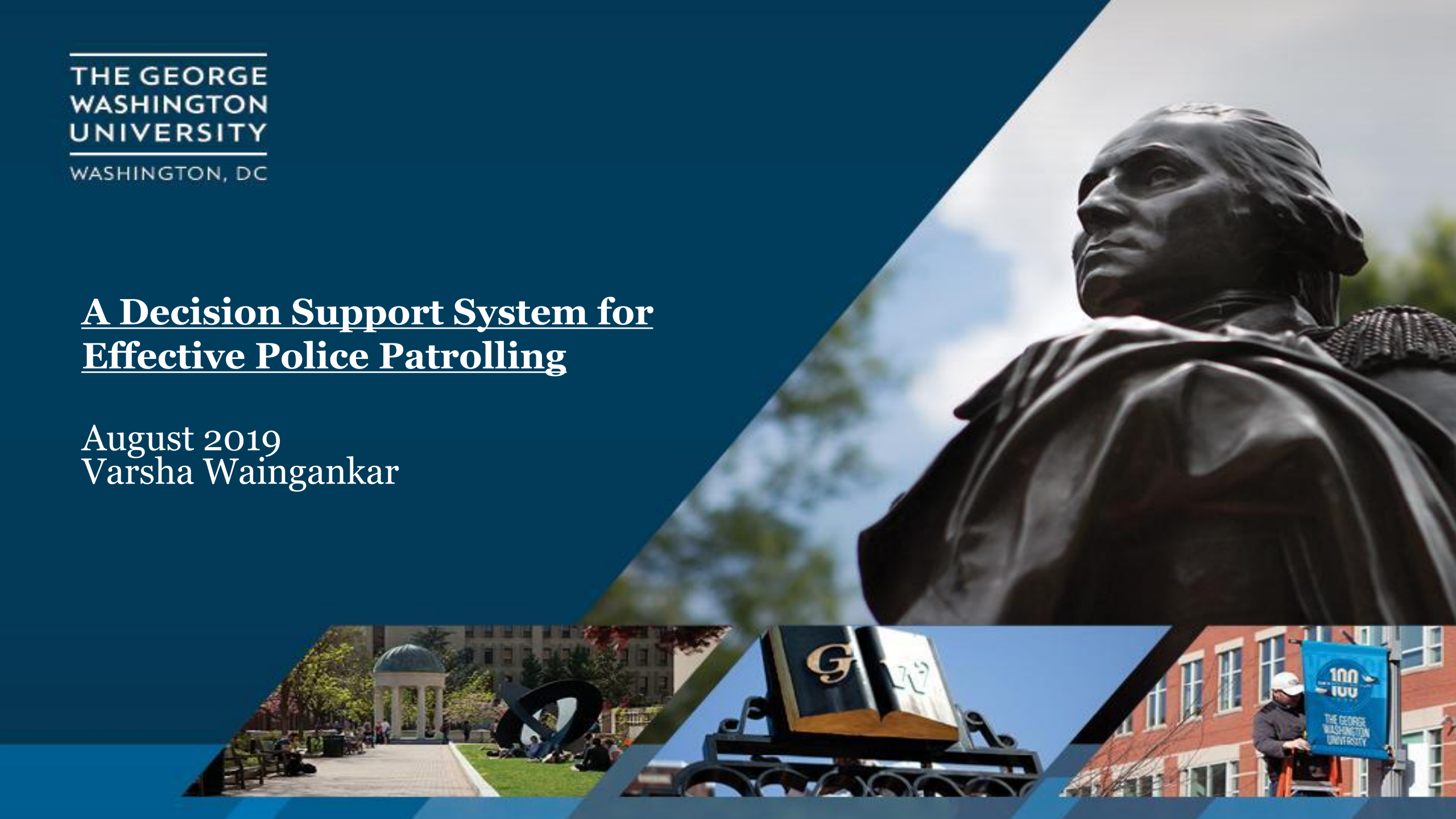


Table of Contents

Problem Statement

1. Effective use of police or patrol cars can be critical for overstretched municipalities across the country.

2. The goal is to help decision makers to more efficiently allocate police patrols and officer time, which are very expensive and scarce for many jurisdictions

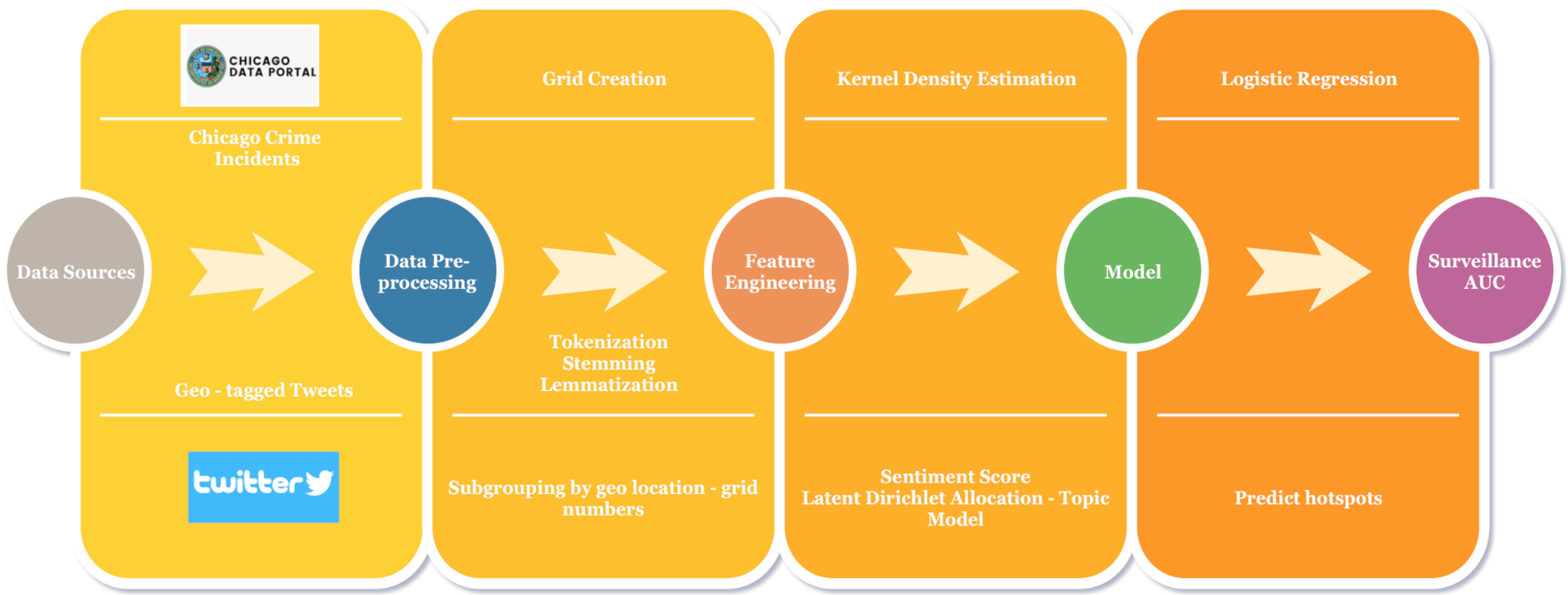
Context

- Report found that from 1997 to 2016 in the United States, the rate of full-time sworn officers per 1,000 residents decreased by 11 percent.
- During the same period, the general population in the U.S. increased by 21 percent.
- The population of Chicago is close to 2.9 million and total number of police officers is 12k
- In Chicago for every 10k population, there are just 44 officers

- Reference

https://www.sciencedirect.com/science/article/pii/S0167923614000268

Roadmap



Planning Phase Preparation Phase Modeling Phase Conclusion

Data

Crime Data

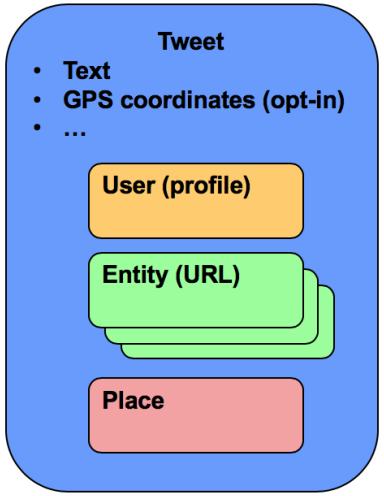
- Using the Chicago Crime Data Portal, information was collected on all crimes documented between June 22, 2019 and August 1, 2019.
- Each crime record in our subset contained a timestamp of occurrence, latitude/longitude coordinates of the crime at the city- block level, and one of 25 types (e.g., ASSAULT and THEFT).

Twitter Data

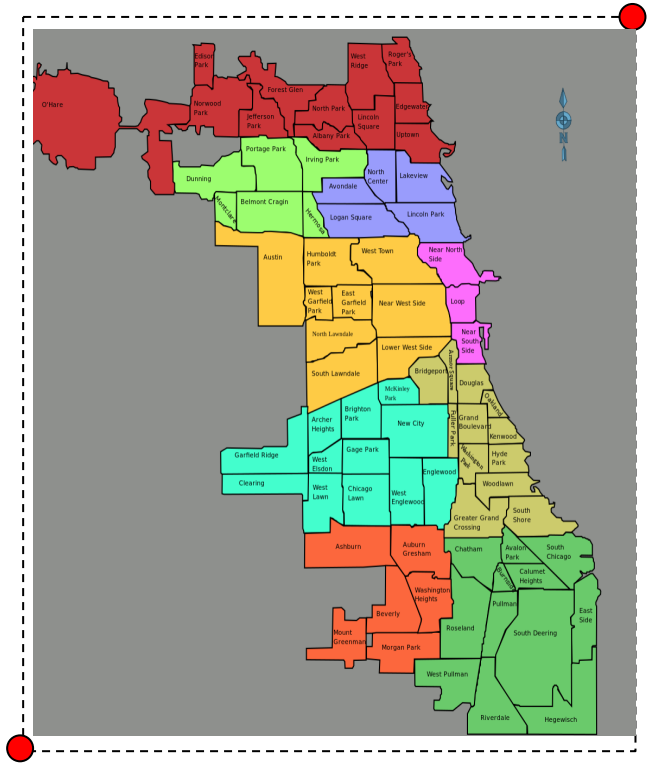
- During the same time period, tweets tagged with GPS coordinates falling within the city limits of Chicago, Illinois were also collected
- This was done by using the official Twitter Streaming API, defining a collection bounding box with coordinates [– 87.94011,41.64454] (lower-left corner) and [– 87.52413,42.02303] (upper-right corner).

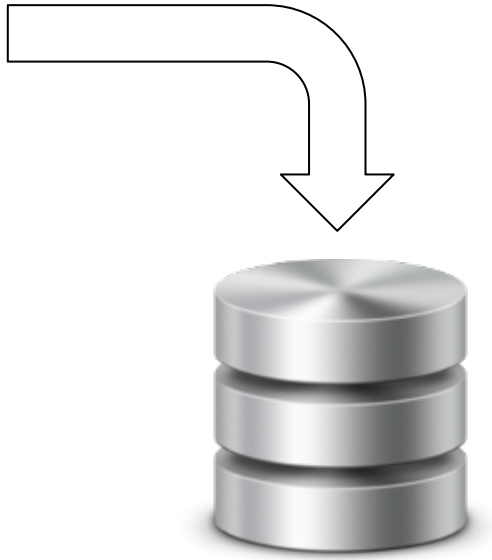
Twitter Streaming API

- Example stream: Filter



Long: -87.5241371038858
Lat: 42.0230385869894

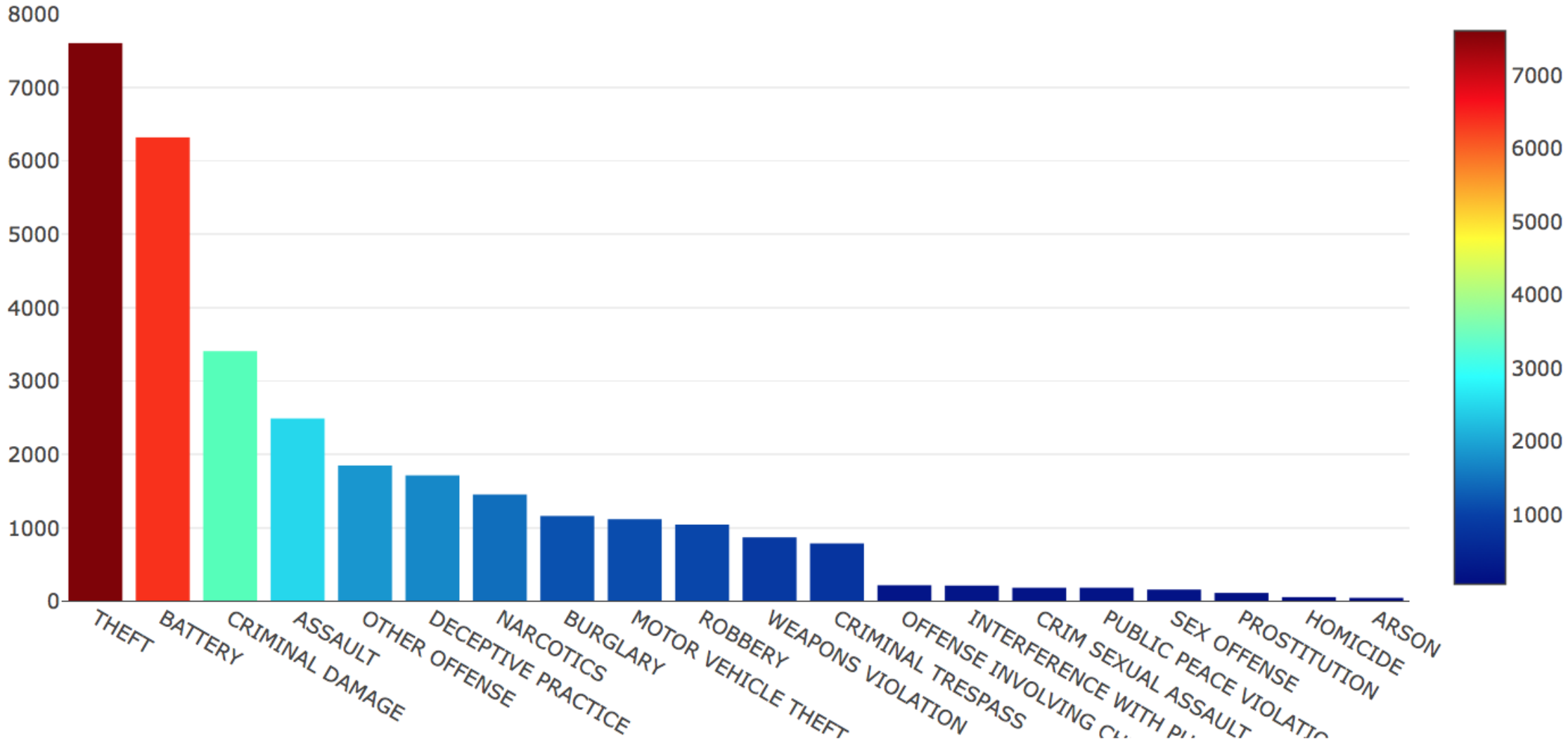




Dropbox

Long: -87.9401140825184
Lat: 41.6445431225492

Crime Categories



Grid Creation

- From the locations of known crimes of type T within the training window (**these points received a label True)**
- From a grid of evenly spaced points at 200- meter intervals, not coinciding with points from the first set (**these points received a label False**).

Grid of Locality

	latitude	latitude_index	longitude	longitude_index
0	41.644540	0	-87.940110	0
1	41.646341	0	-87.940136	1
2	41.648142	0	-87.940162	2
3	41.649944	0	-87.940189	3

Data Crime Type Theft

'latitude', 'longitude', 'latitude index', 'longitude index', 'crime'

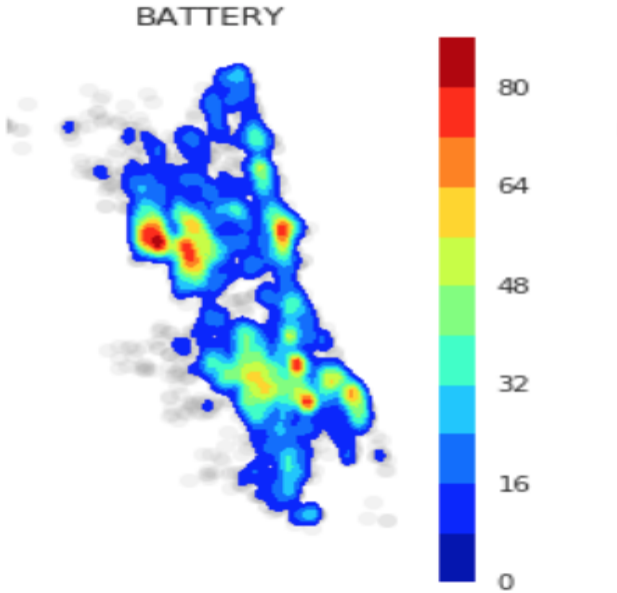
Long: -87.9401140825184
Lat: 41.6445431225492

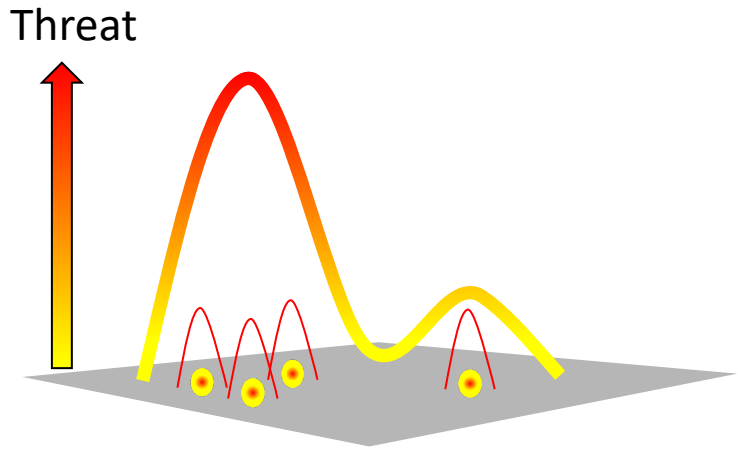
City grid divided into 200 meter squares

Long: -87.9401140825184
Lat: 41.6445431225492

Kernel Density Estimation

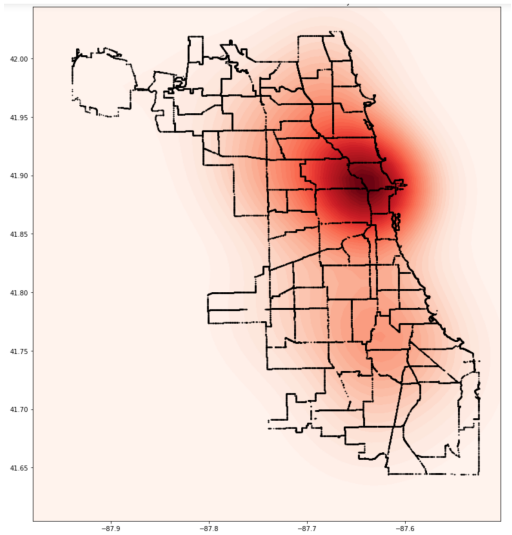
- Kernel Density Estimation (KDE) is a non-parametric technique for density estimation.
- A known density function (the kernel) is averaged across the observed data points to create a smooth approximation.
- Estimation data: historical crime record

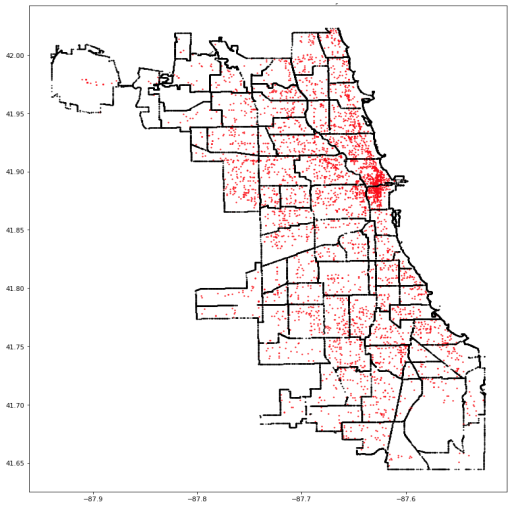




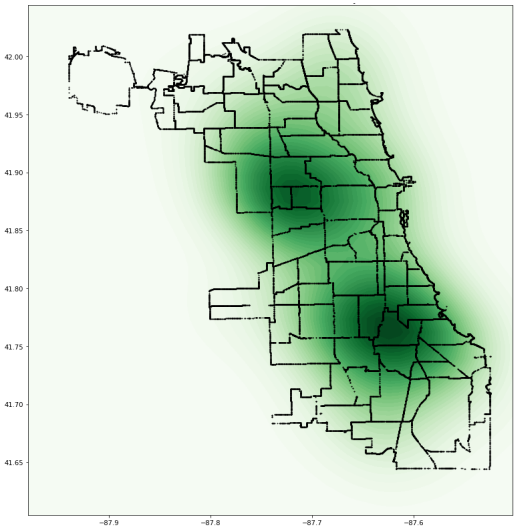
Hot map for crime types using KDE

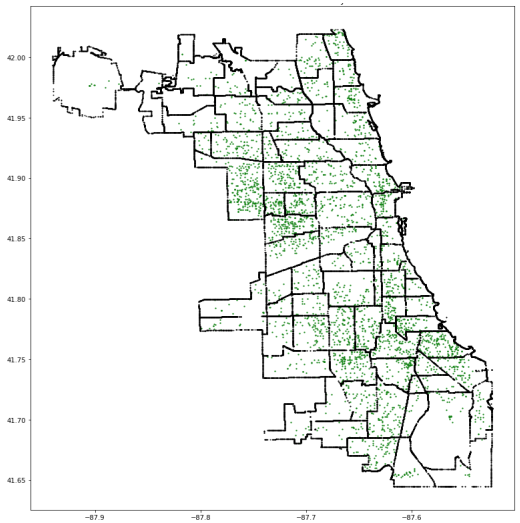
Theft



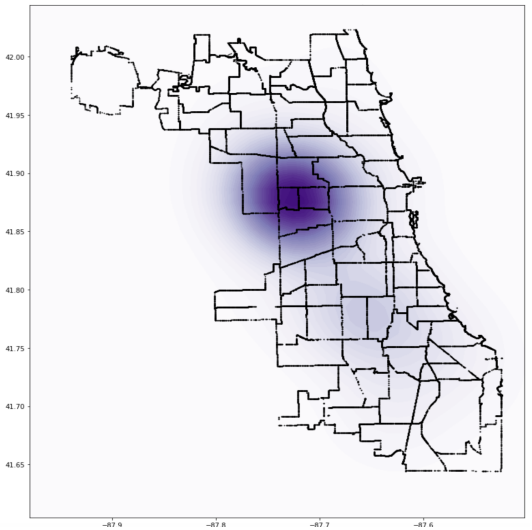


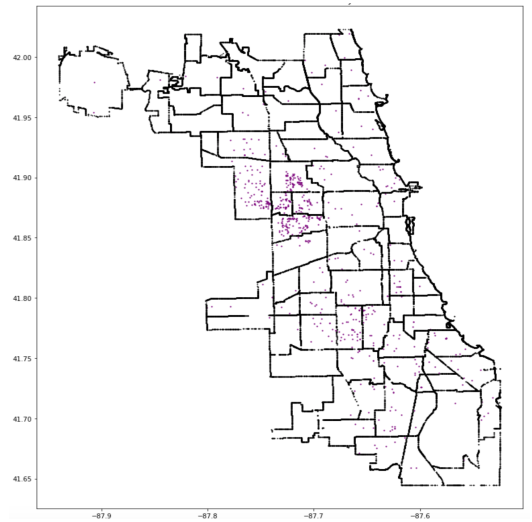
Battery



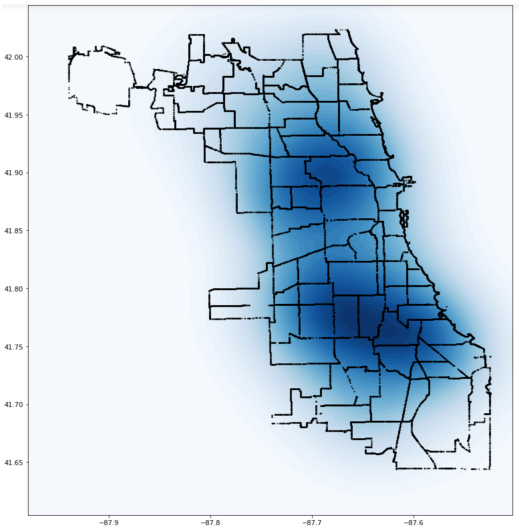


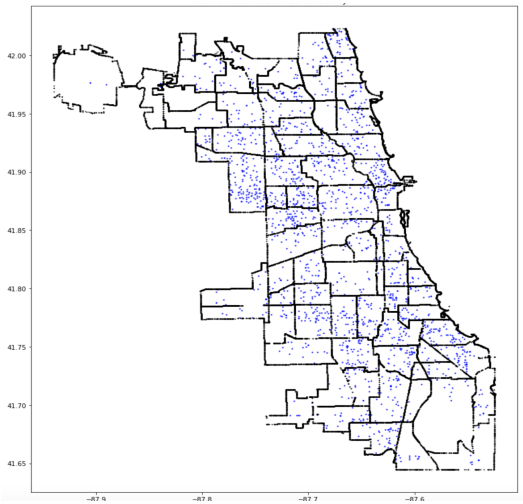
Narcotics





Burglary





Planning Phase

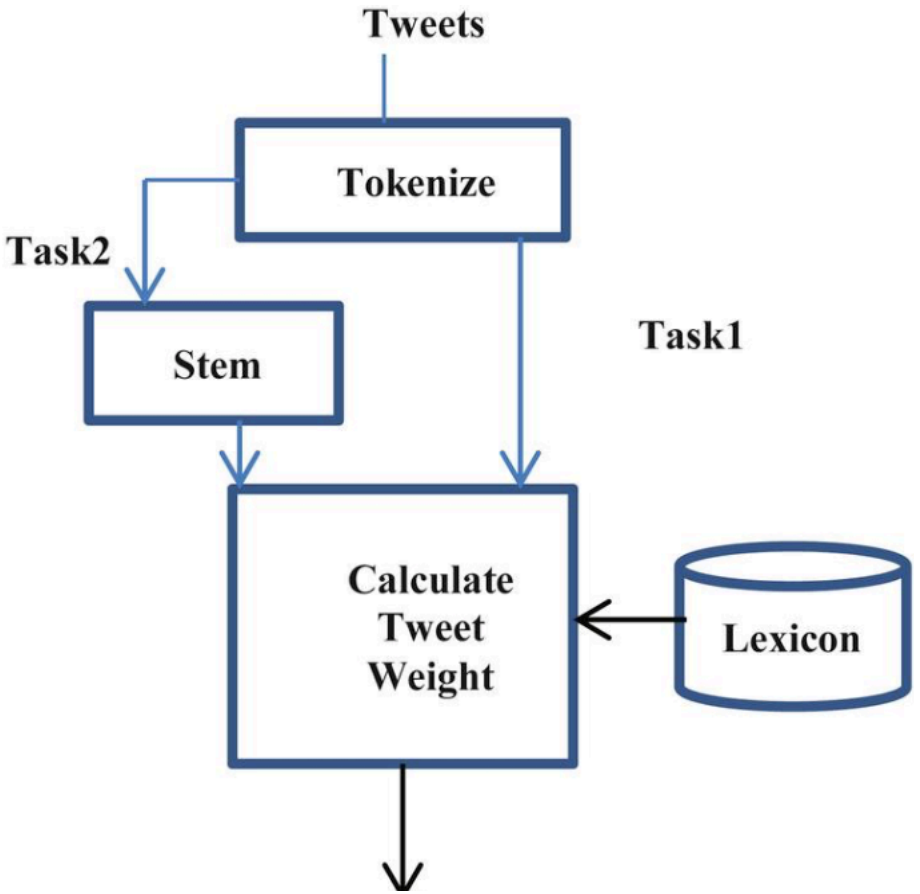
Preparation Phase

Modeling Phase

Conclusion

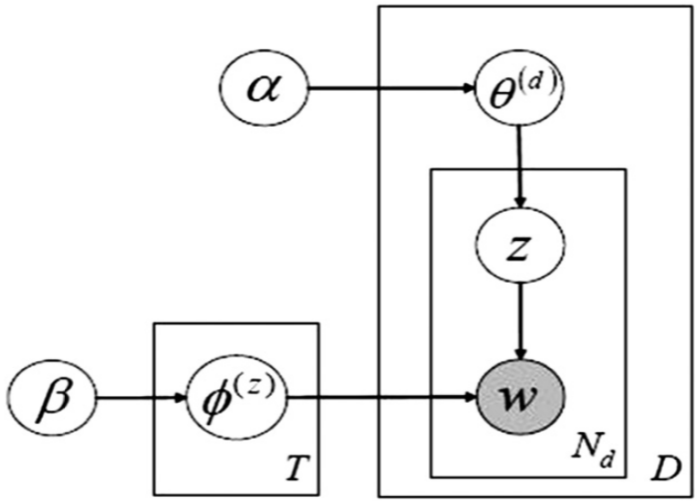
Sentiment Analysis on Tweets

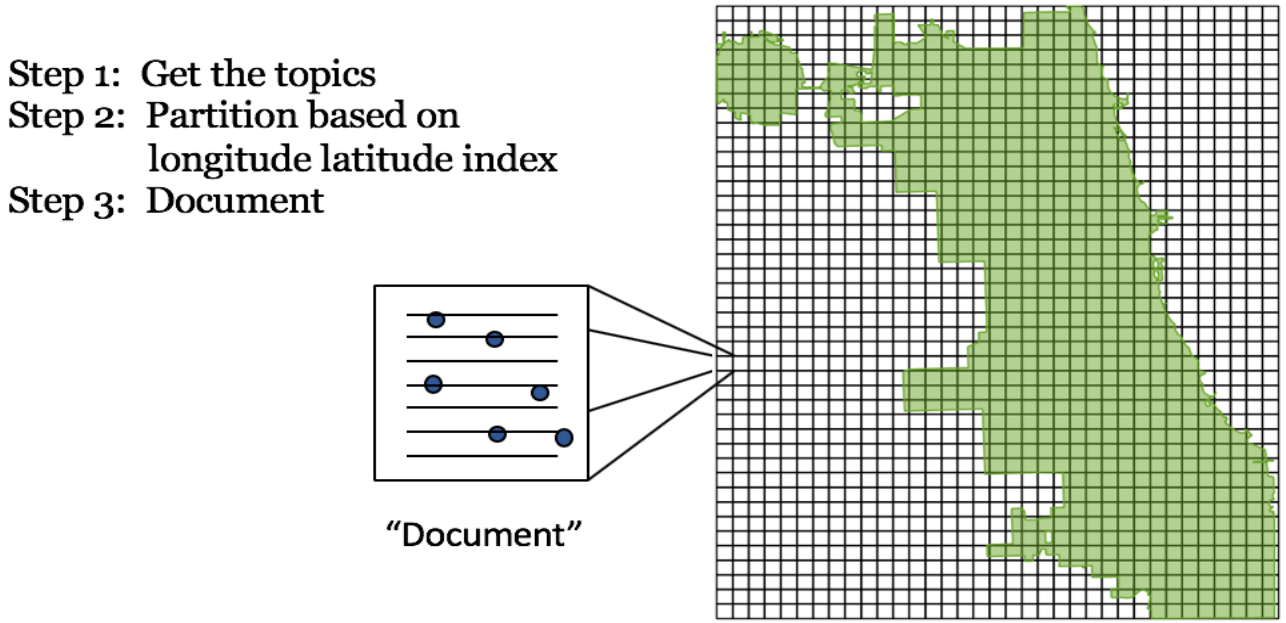
- There are two major approaches:
- Supervised machine learning approaches
- Unsupervised lexicon-based approaches
- Unsupervised method for predicting the sentiment by using lexicons that have detailed information, specially curated and prepared just for sentiment analysis was used.
- AFINN lexicon for sentiment extraction and analysis was implemented



Topic Modeling – Latent Dirichlet Allocation

- Preprocess text, apply tokenization, stemming, lemmatization
- Remove the stop words
- Apply topic models

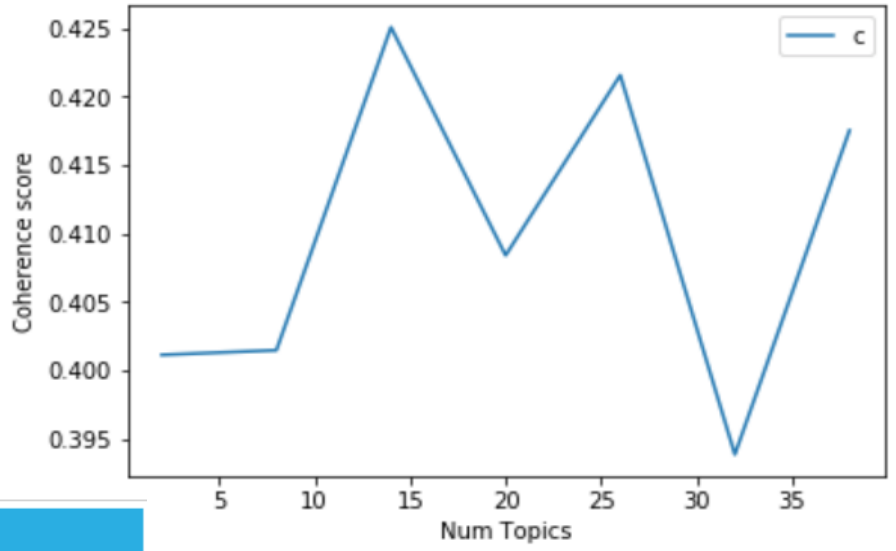




Neighborhood boundaries for computing tweet-based topics. Used only the green neighborhoods (i.e., those within the city boundary) for analysis.

Coherence Score

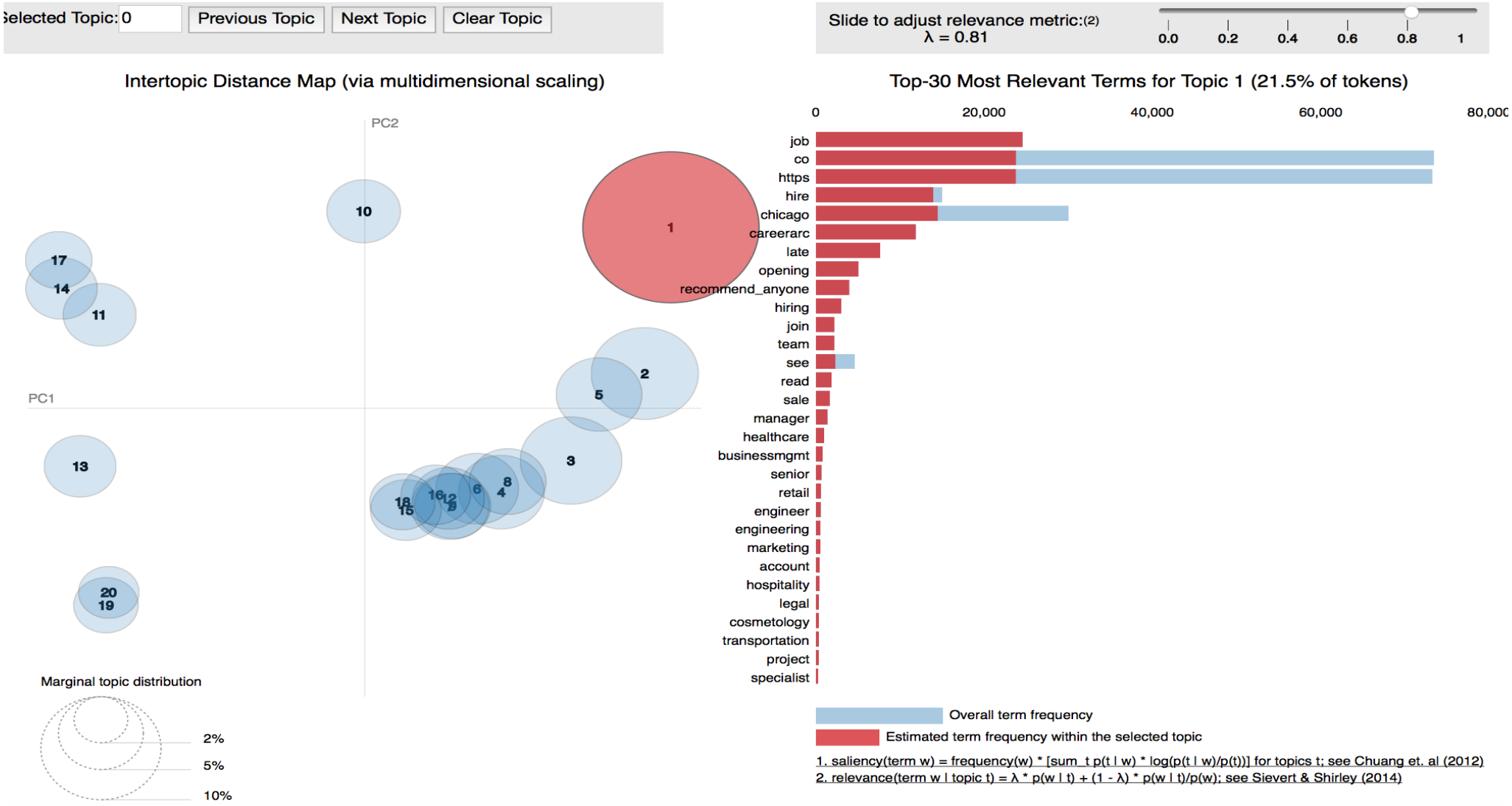
- This score is trying to quantify the semantic similarities of the high scoring words within each topic.
- A high score means the result is more human-interpretable that indicates a better model.
- Document collections start as matrices of terms and documents



	Topic 1	Topic 2	Topic 3
Document 1	1	0	1
Document 2	0	0	0
Document 3	0	0	1
Document 3	0	1	0

	Word 1	Word 2	Word 3
Topic 1	1	1	0
Topic 2	0	1	1
Topic 3	1	0	1

LDA Visualization



Model Execution and Performance Evaluation

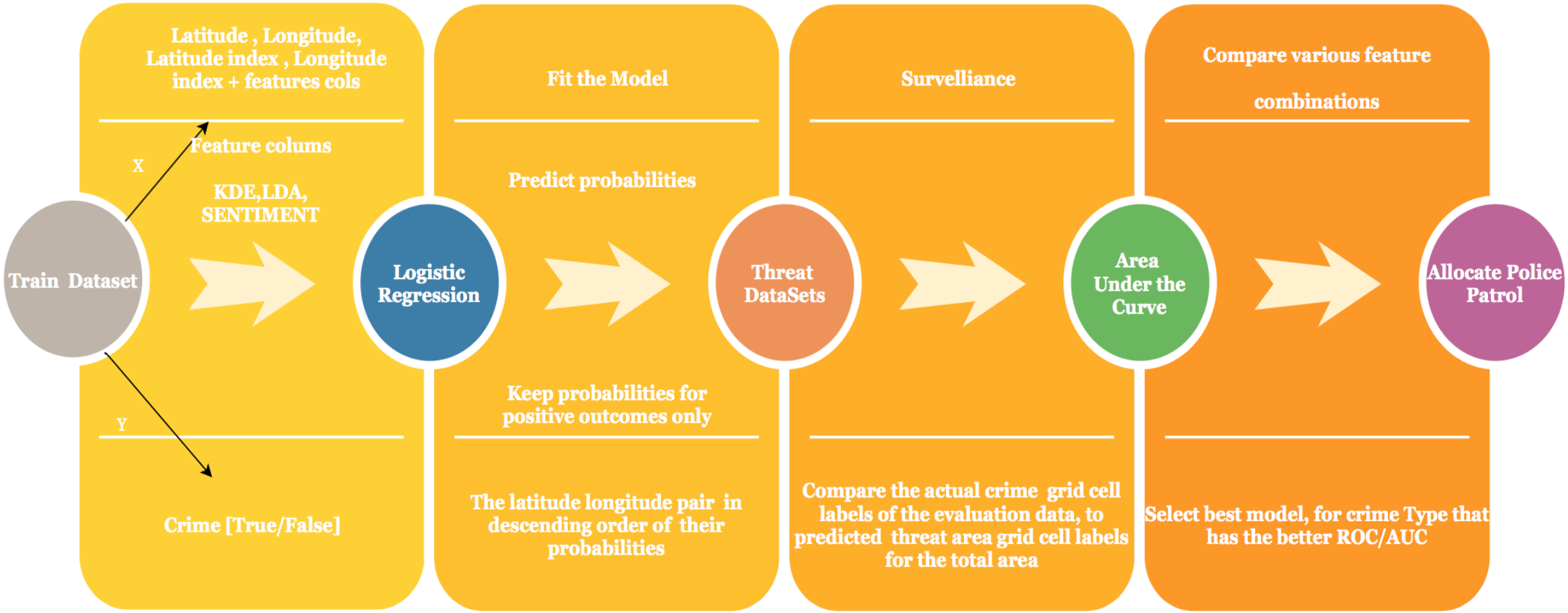
Model

- Training the model on 31 day window for crime type T.
- Making T predictions for the first day following the training window, sliding one day into the future and repeating.

Evaluation

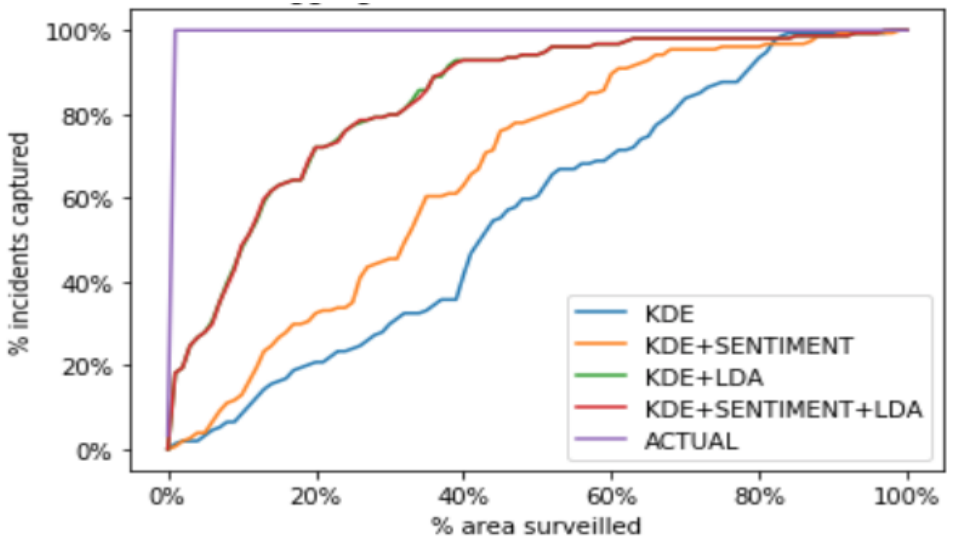
- Project the surveillance plot that measures the percentage of true T crimes during prediction window , that occur within the most threatened area, according to models prediction for T
- Lastly, because each model execution produced a series of surveillance plots for crime type T, one for each prediction day, aggregate the plots to measure overall performance

Model Execution



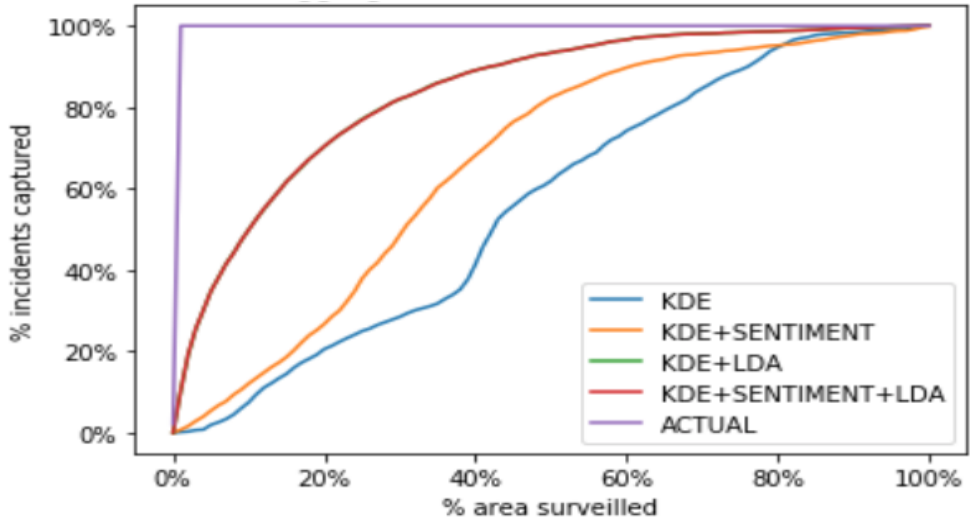
Crime Type - Theft

AUC – One day Prediction




```
56.538961038961034 KDE
67.18831168831167 KDE+SENTIMENT
84.07142857142856 KDE+LDA
84.04545454545455 KDE+SENTIMENT+LDA
100.03246753246754 ACTUAL
```

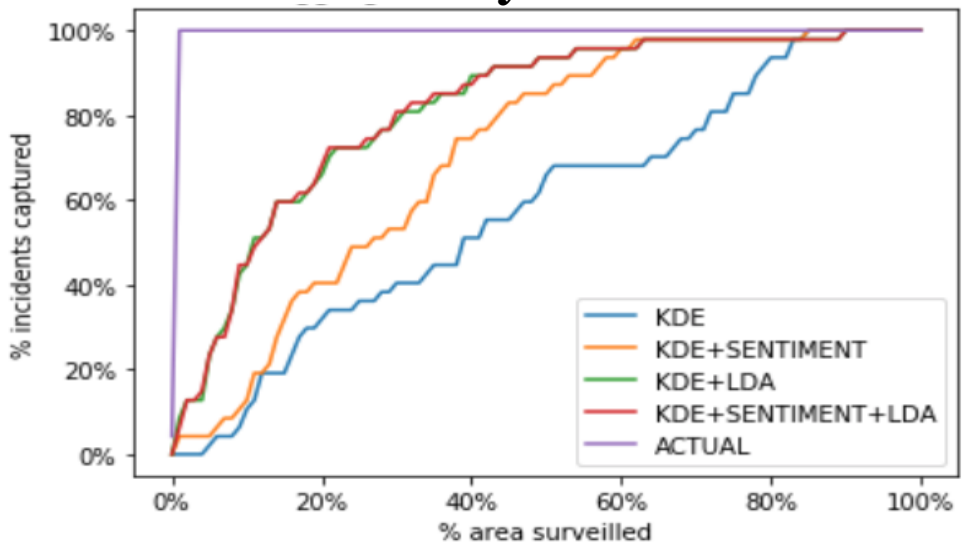
AUC – Aggregated Plot for 10 days prediction




```
56.8164610444482 KDE
66.96569207368599 KDE+SENTIMENT
84.40949805644753 KDE+LDA
84.38837248605712 KDE+SENTIMENT+LDA
```

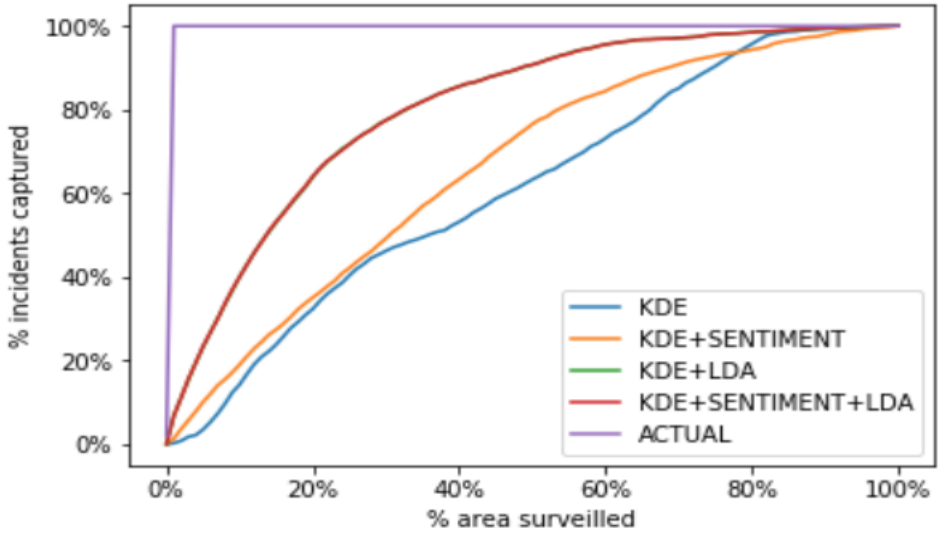
Crime Type - Battery

AUC – One day Prediction




```
58.72340425531915 KDE
71.8936170212766 KDE+SENTIMENT
82.1063829787234 KDE+LDA
82.25531914893617 KDE+SENTIMENT+LDA
100.04255319148936 ACTUAL
```

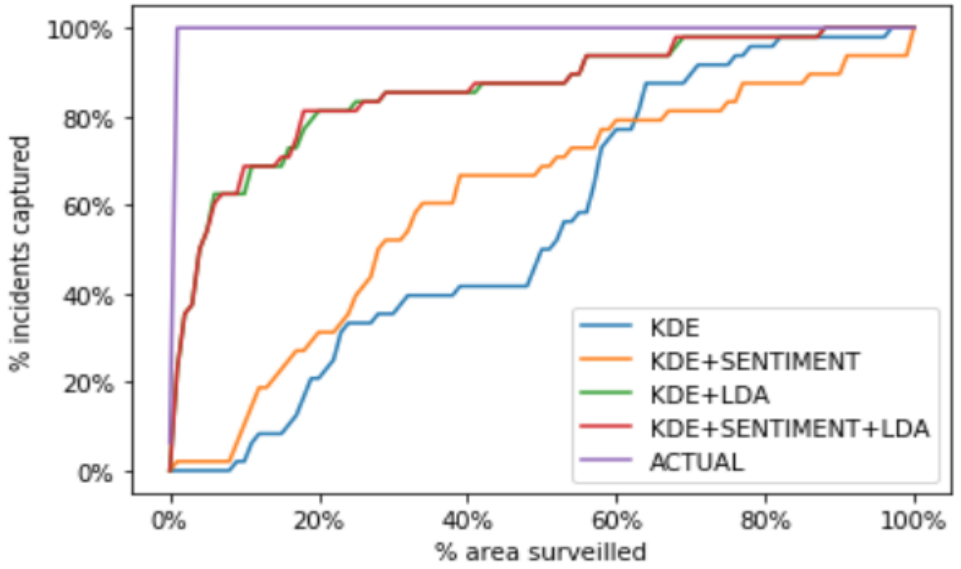
AUC – Aggregated Plot for 10 days prediction




```
61.76088314934388 KDE
66.78816913143095 KDE+SENTIMENT
80.8117058946053 KDE+LDA
80.836075817538 KDE+SENTIMENT+LDA
100.0202041241408 ACTUAL
```

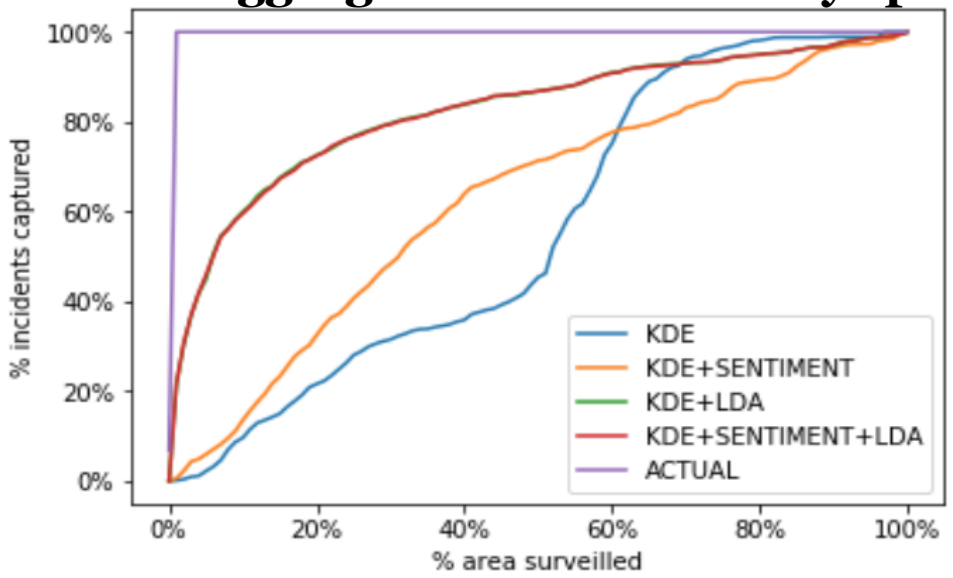
Crime Type - Narcotics

AUC – One day Prediction




```
56.395833333333336 KDE
61.541666666666664 KDE+SENTIMENT
86.1875 KDE+LDA
86.33333333333331 KDE+SENTIMENT+LDA
100.0625 ACTUAL
```

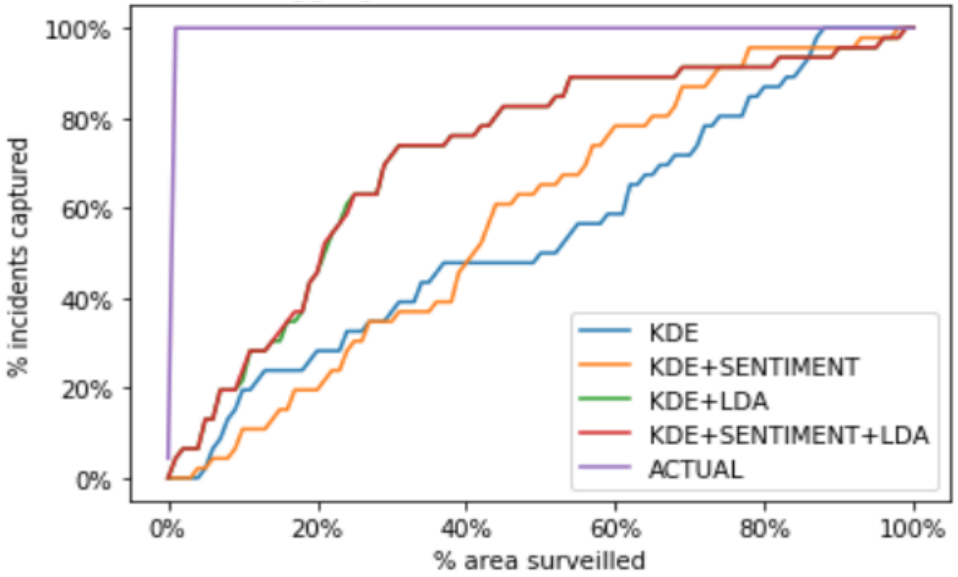
AUC – Aggregated Plot for 10 days prediction




```
56.63587684069611 KDE
62.99732262382865 KDE+SENTIMENT
82.75033467202141 KDE+LDA
82.71820615796518 KDE+SENTIMENT+LDA
100.06827309236948 ACTUAL
```

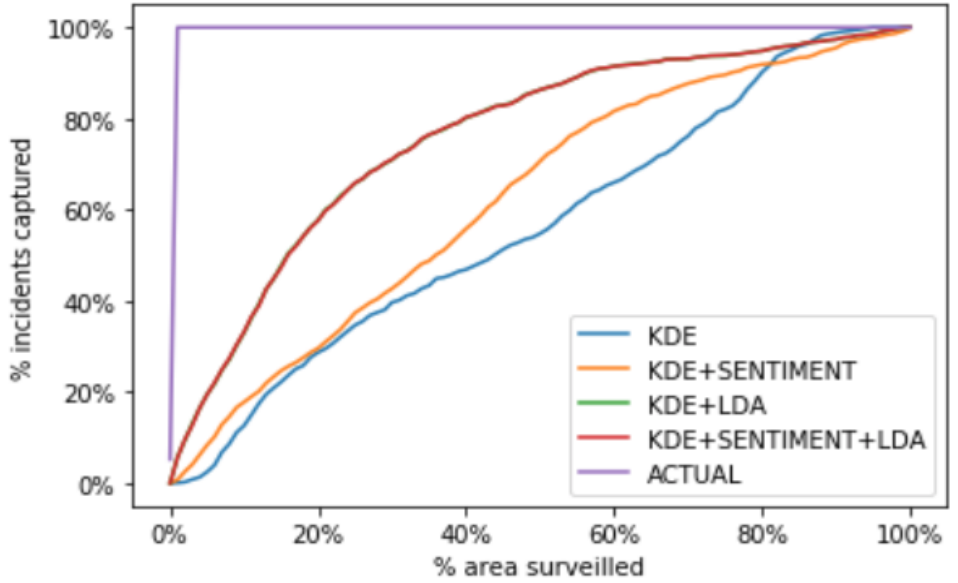
Crime Type - Burglary

AUC – One day Prediction




```
55.21739130434783 KDE
58.45652173913044 KDE+SENTIMENT
73.95652173913044 KDE+LDA
73.8913043478261 KDE+SENTIMENT+LDA
100.04347826086956 ACTUAL
```

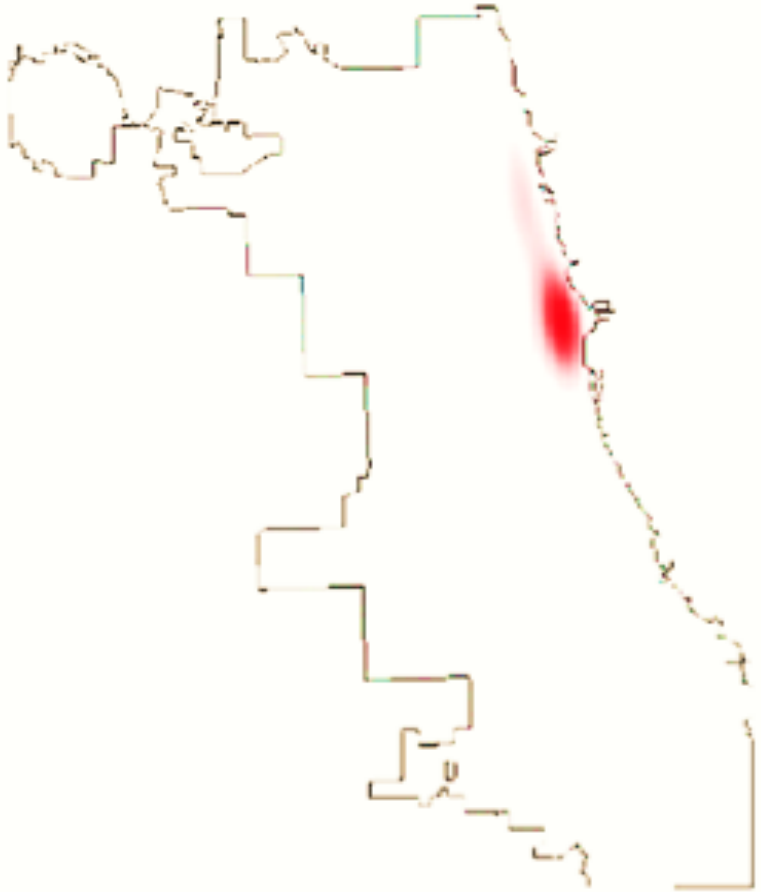
AUC – Aggregated Plot for 10 days prediction



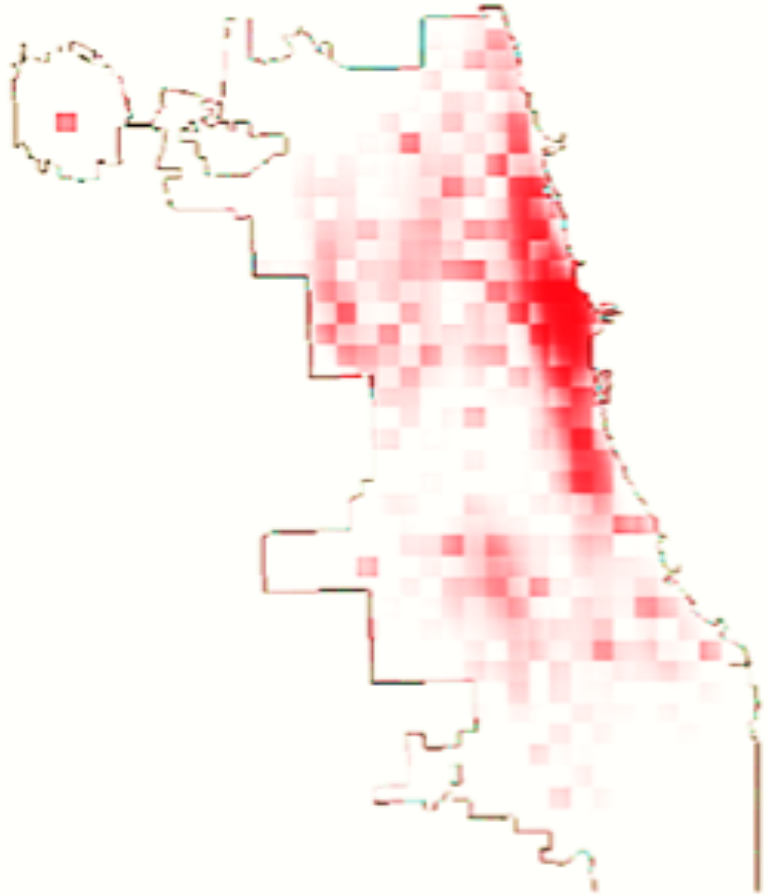

```
56.898360655737704 KDE
62.90409836065574 KDE+SENTIMENT
76.51803278688526 KDE+LDA
76.50819672131149 KDE+SENTIMENT+LDA
100.05245901639344 ACTUAL
```

Crime prediction (Theft)

Predicted hotspots using KDE



Predicted hotspots using KDE+LDA



Planning Phase

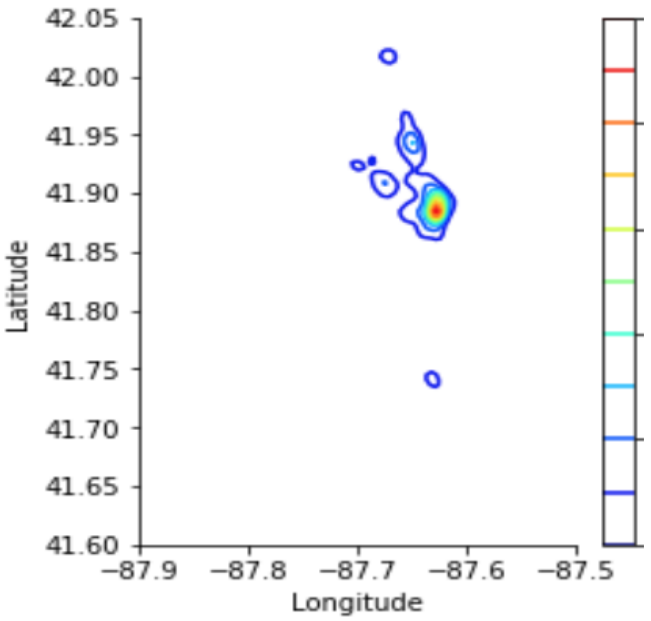
Preparation Phase

Modeling Phase

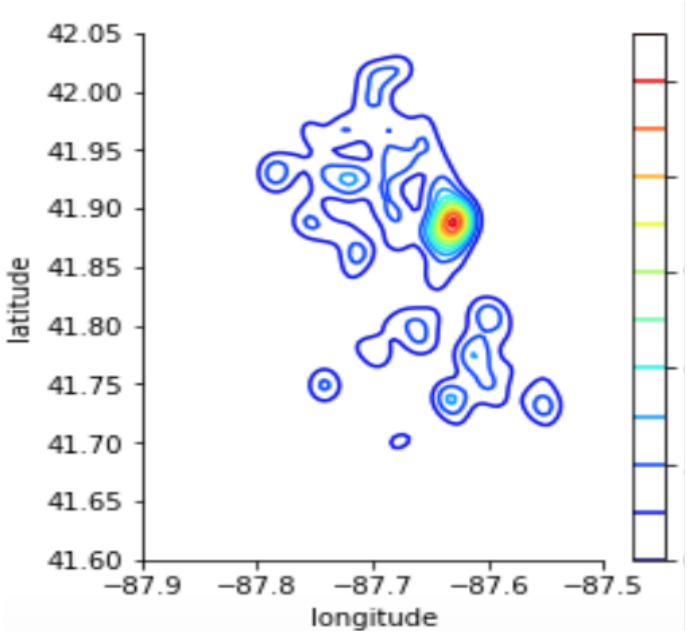
Conclusion

Model Evaluation

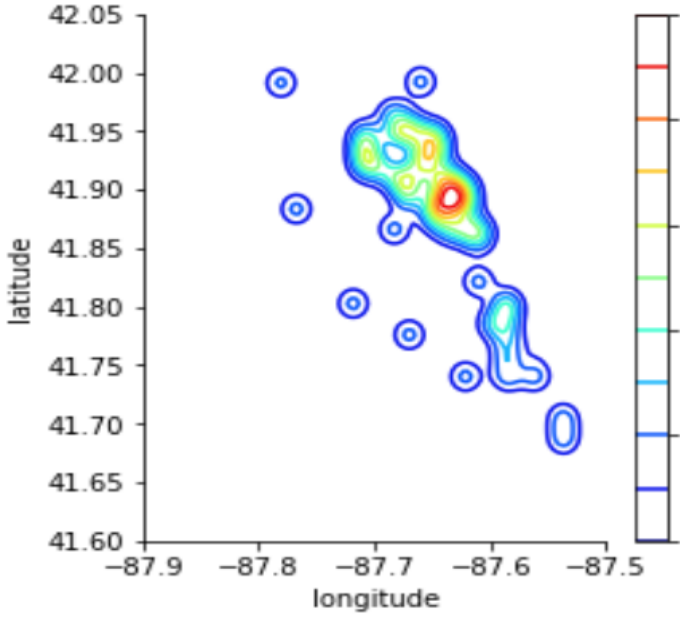
KDE generated Hotspots



Actual Hotspots



KDE+LDA generated Hotspots



Planning Phase

Preparation Phase

Modeling Phase

Conclusion

Conclusion

- The best model for crime hotspot prediction is using **Kernel density estimation and LDA** topic models, as it yields the best AUC (Area Under the Curve).

- Using the best model, decision makers can allocate scarce resources (e.g., police patrols) across the geographic space, more efficiently, leading to reduction in wasted effort and decrease in crime response time.

- Only for 19 of the 25 crime types studied, the addition of Twitter data improves crime prediction performance versus a standard approach based on kernel density estimation.


danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
dank je
thank you
gracias
tapadh leat
bedankt
hvala
mauruuru
mochchakkeram
dziękuję
go raibh maith agat
sagolun
sukriya
kop khun krap
obrigado
arigatō
takk
dakujem
grazie
мерси
terima kasih
감사합니다
merci
ευχαριστώ