Toward Predictive Crime Analysis via Social Media, Big Data, and GIS

Anthony J. Corso

Agenda

- Introduction
  - Background
  - Objectives
  - Problem
- Data
- Artifact Outcomes
- Current and Future Research
- Conclusion

Introduction

- **Objectives**
  - **Realize links between crime and domain specific data**
  - **Investigate Spatial and Temporal Analysis of Crime, Nearest Neighbor Clustering, Kernel Density Estimation, and Risk Terrain Modeling techniques**
  - **Investigate linguistic analysis of social media combined with crime and domain specific data for real-time predictive capabilities**
- **Problem**
  - **Identify predictive real-time ??? via social media and historical record**

Theory Base for Research

- **Abstract**
- **Concrete**

Nominal
Ordinal
Interval
Ratio

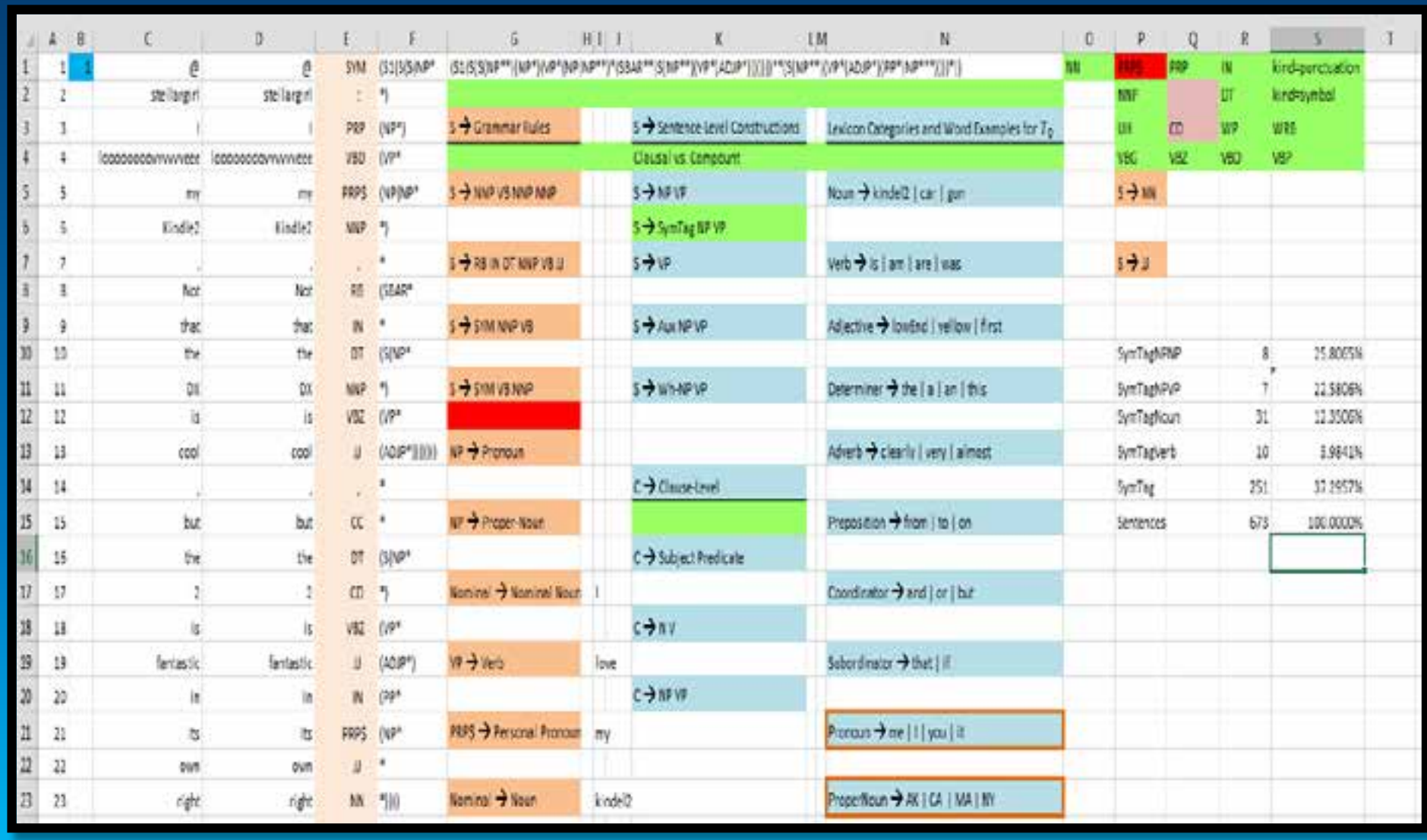
Reynolds, P. D. 2007. *A primer in theory construction*, Boston.

Tweet Anatomy

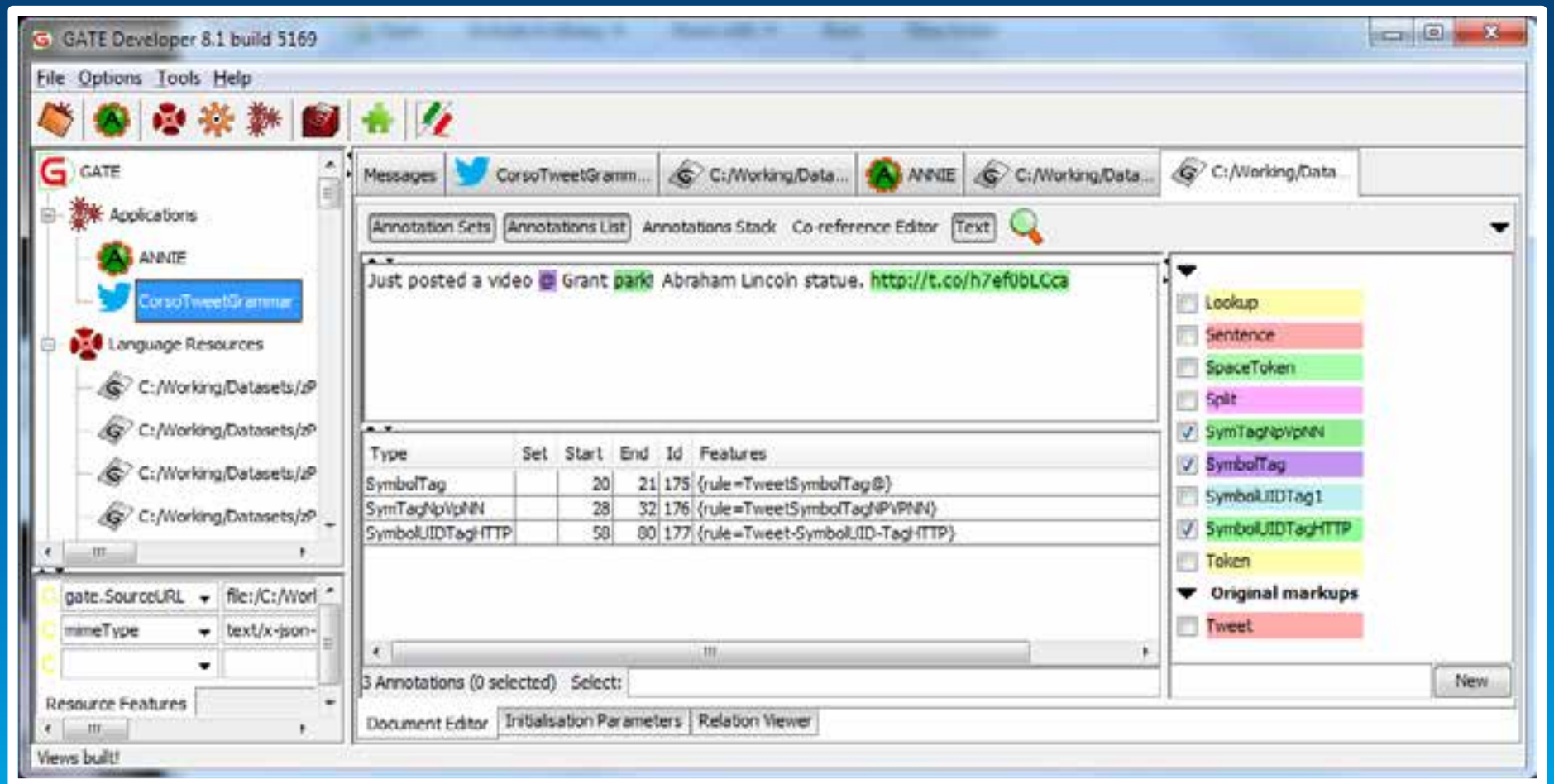

```
"Wed Oct 01 08:06:57 +0000 2014","id":517224133898661296,"id_str":"517224133898661296","text":"@MorganEastwood Evidently, your dad was going to be on the original B
"Wed Oct 01 08:07:12 +0000 2014","id":517224197249306624,"id_str":"517224197249306624","text":"I done been shot and had my head caved in so don't expect me to break
"Wed Oct 01 08:07:16 +0000 2014","id":517224213443543040,"id_str":"517224213443543040","text":"#PARANORMAL #FANTASY\n@FrostFyre\nWYTCHFAE\nFun, Action-Packed Romanc
"Wed Oct 01 08:07:26 +0000 2014","id":517224253188759553,"id_str":"517224253188759553","text":"@maimu92 omg thanks so much!!!!","source":"\u003ca href=\"http:\/\/tw
"Wed Oct 01 08:07:32 +0000 2014","id":517224280585945088,"id_str":"517224280585945088","text":"bout to catch des\ud83d\ude34 on u mfs been a long day","source":"\u0
"Wed Oct 01 08:07:41 +0000 2014","id":517224317420306432,"id_str":"517224317420306432","text":"I believe in the 3 strike rule n I got 2 strikes already","source":"\
"Wed Oct 01 08:07:46 +0000 2014","id":517224336584105984,"id_str":"517224336584105984","text":"@xo_leena you best believe ill be there with you baby.","source":"\u0
"Wed Oct 01 08:07:47 +0000 2014","id":517224343345311744,"id_str":"517224343345311744","text":"Somebody Tml!!","source":"\u003ca href=\"http:\/\/twitter.com\/downlo
"Wed Oct 01 08:07:49 +0000 2014","id":517224348990832640,"id_str":"517224348990832640","text":"I blame that vine I watched like 50 times","source":"\u003ca href=\"h
"Wed Oct 01 08:07:55 +0000 2014","id":517224376442585088,"id_str":"517224376442585088","text":"Then he have the nerve to wake up &amp;amp; tell me to get off the phone.
"Wed Oct 01 08:08:18 +0000 2014","id":517224471527055361,"id_str":"517224471527055361","text":"Oj","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\"
```


```
"@MorganEastwood Evidently, your dad was going to be on the original Batman. What happened
"I done been shot and had my head caved in so don't expect me to break up no fights ....","
"#PARANORMAL #FANTASY\n@FrostFyre\nWYTCHFAE\nFun, Action-Packed Romance!\nhttp:\/\/t.co\/9c
```

Tweet Grammar



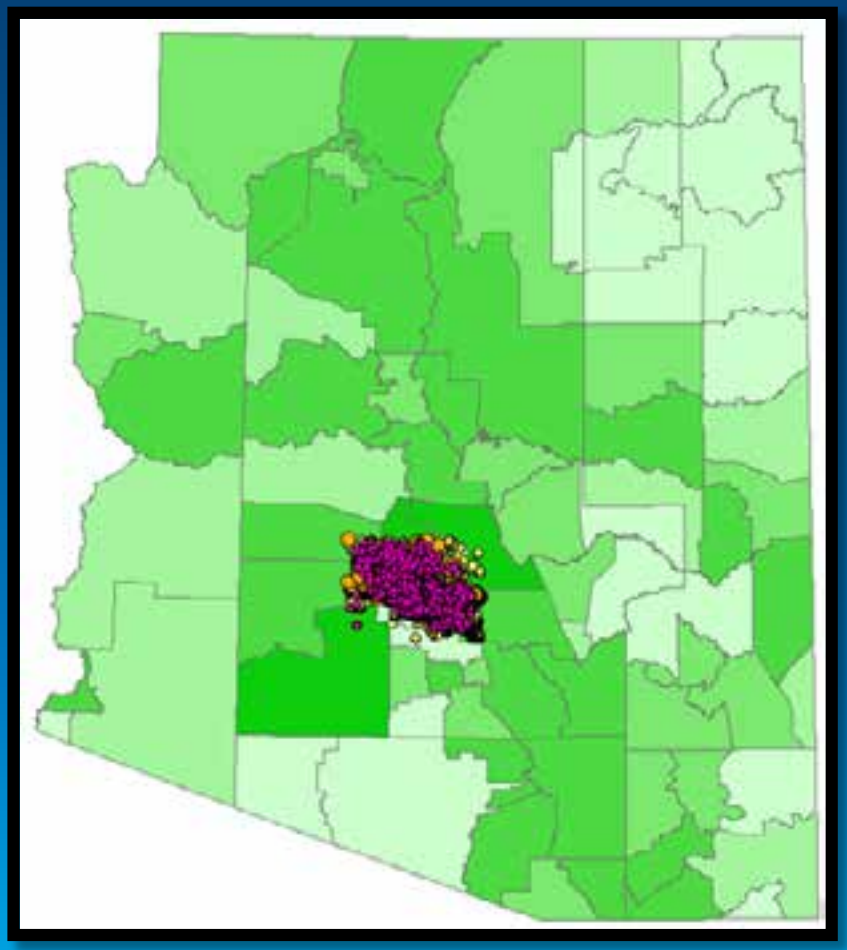
Natural Language Processing



Big Data

- **What does it mean???**
  - **Millions, billions, trillions of records**
  - **A lot of data, A lot of data, A lot of data, A lot of data**
  - **Not, A lot of data is still small data**
- **Problem**
  - **Data pipe**
    - **JavaScript Object Notation (JSON)**
    - **General Architecture for Text Engineering (GATE)**
    - **Java**
    - **ArcGIS**

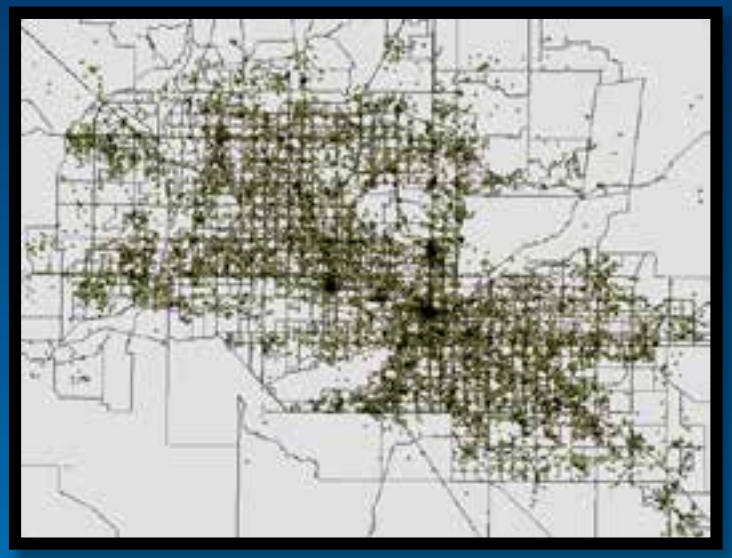
Data, Maps, and GIS Overview

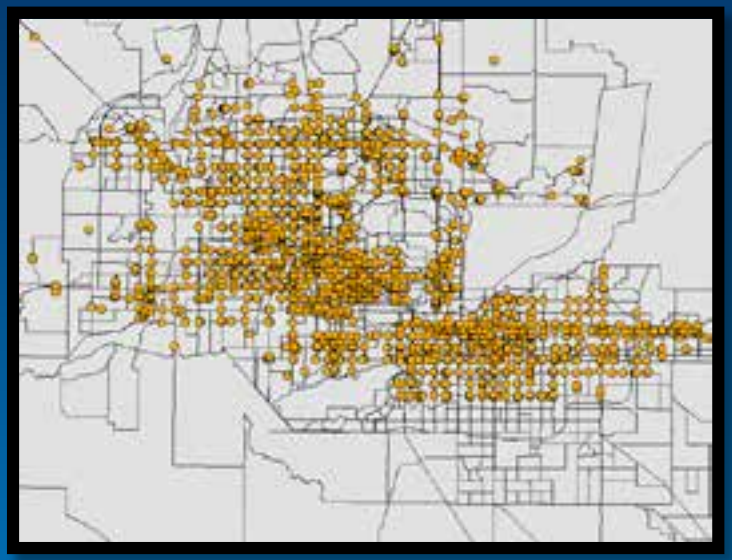


Small Scale

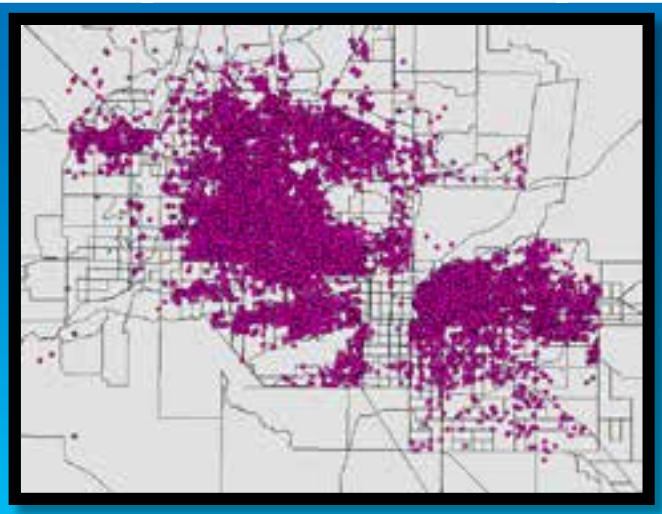
Large Scale

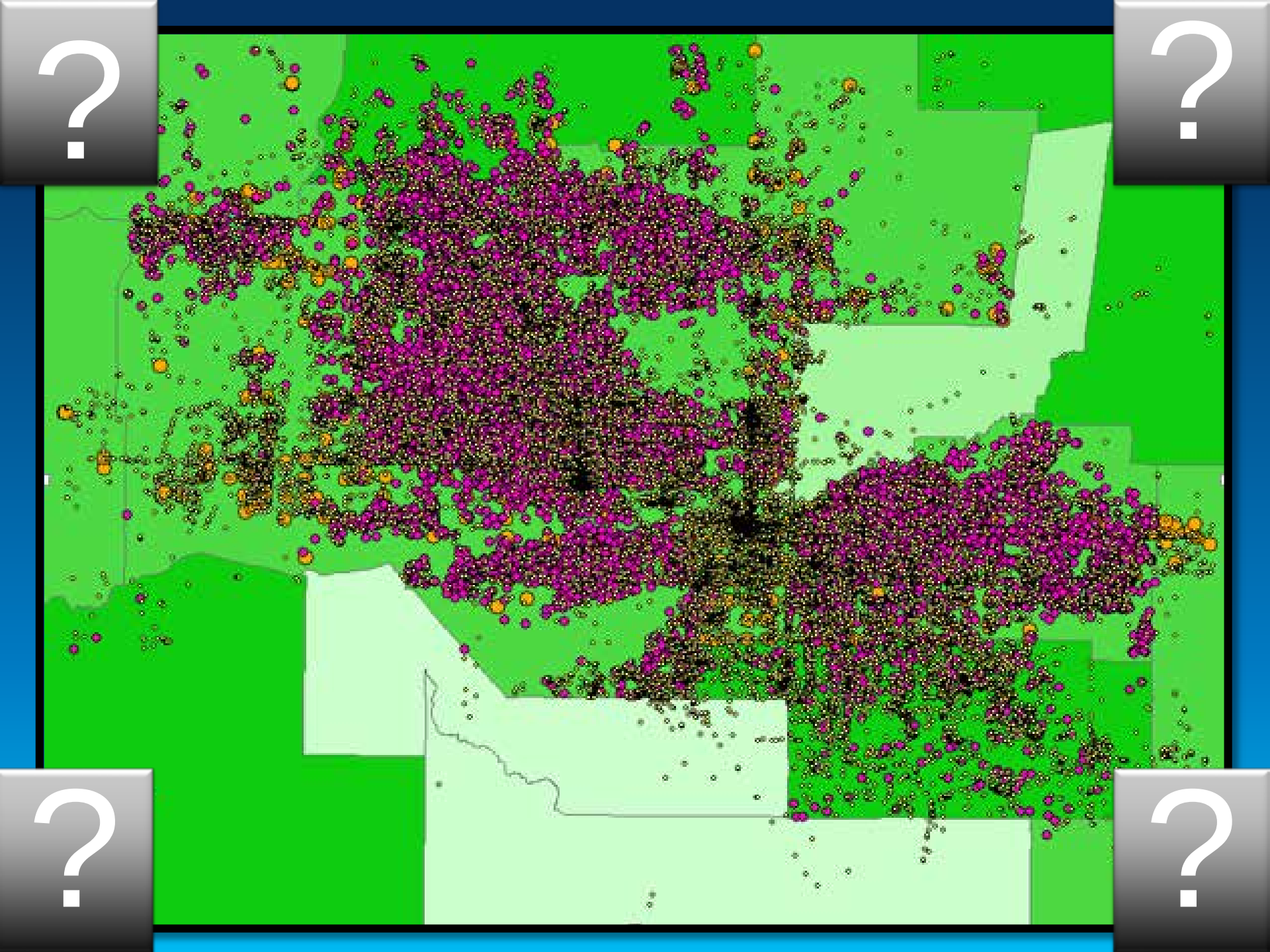
Data, Maps, and GIS Integration



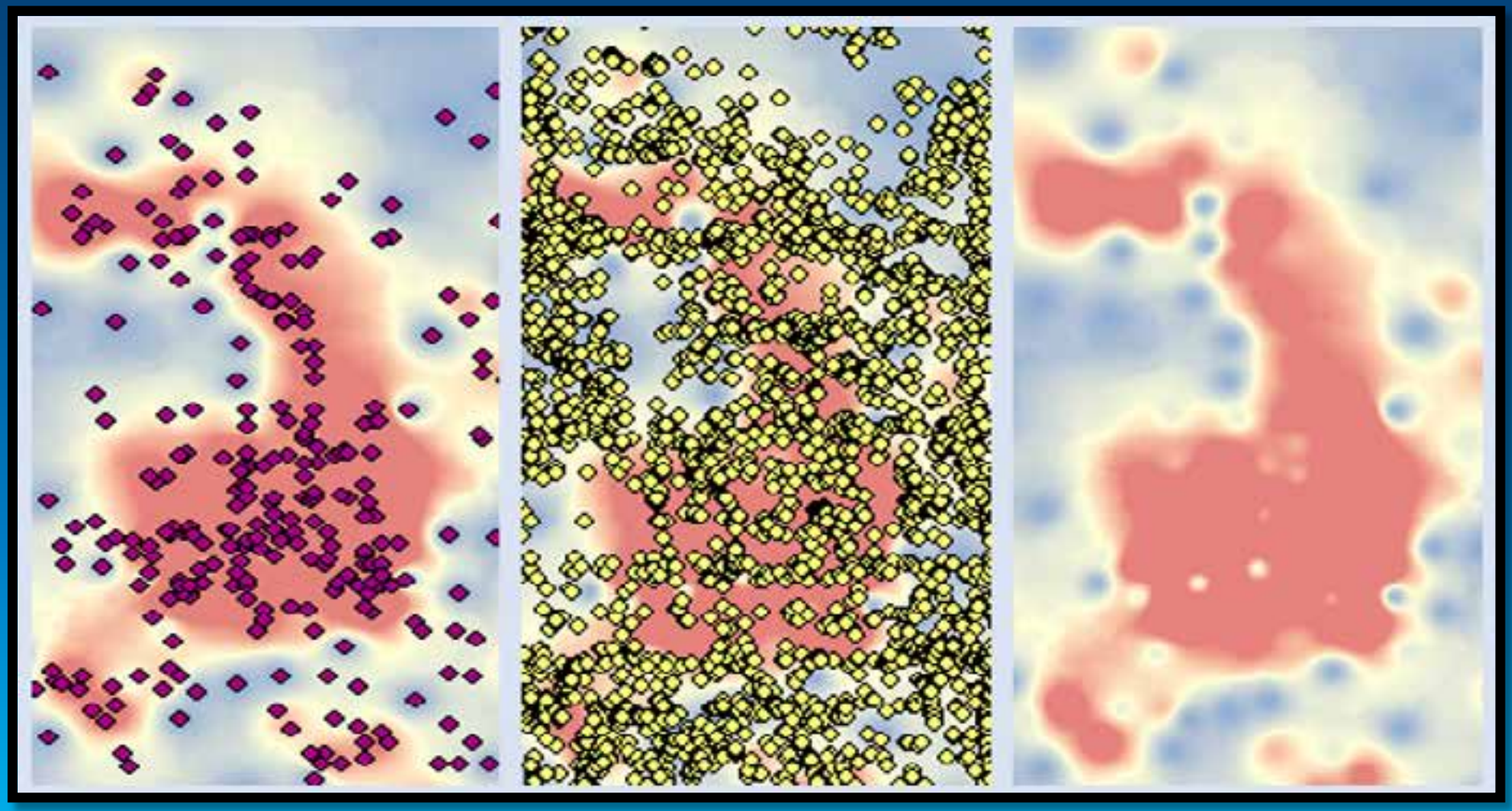


Tweet Corpus	Crime Data	SNAP Locations

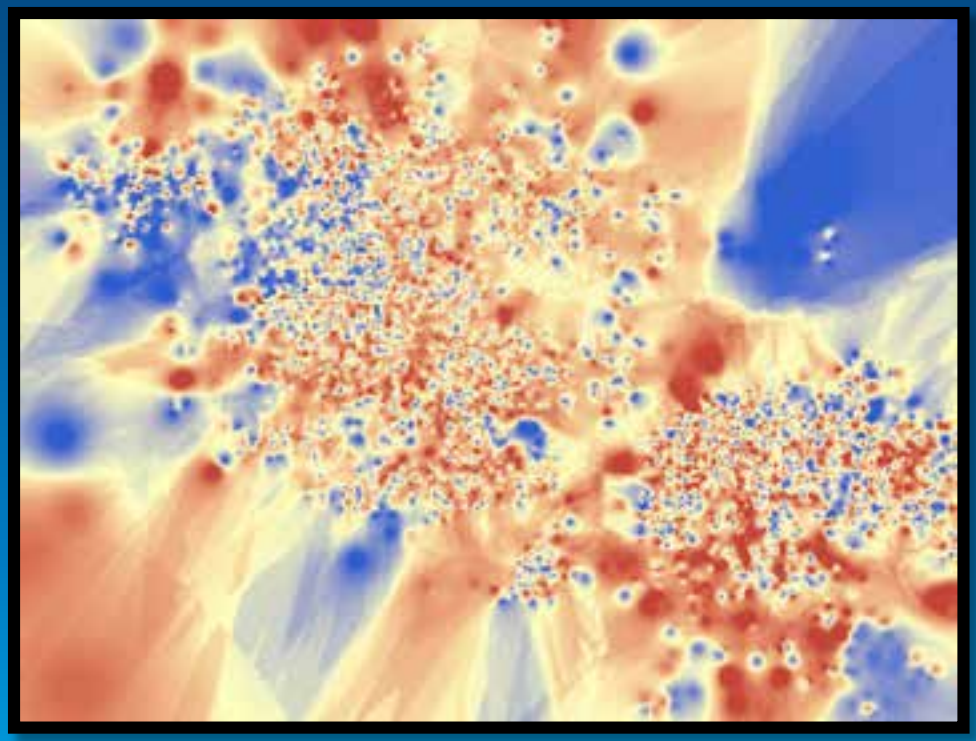




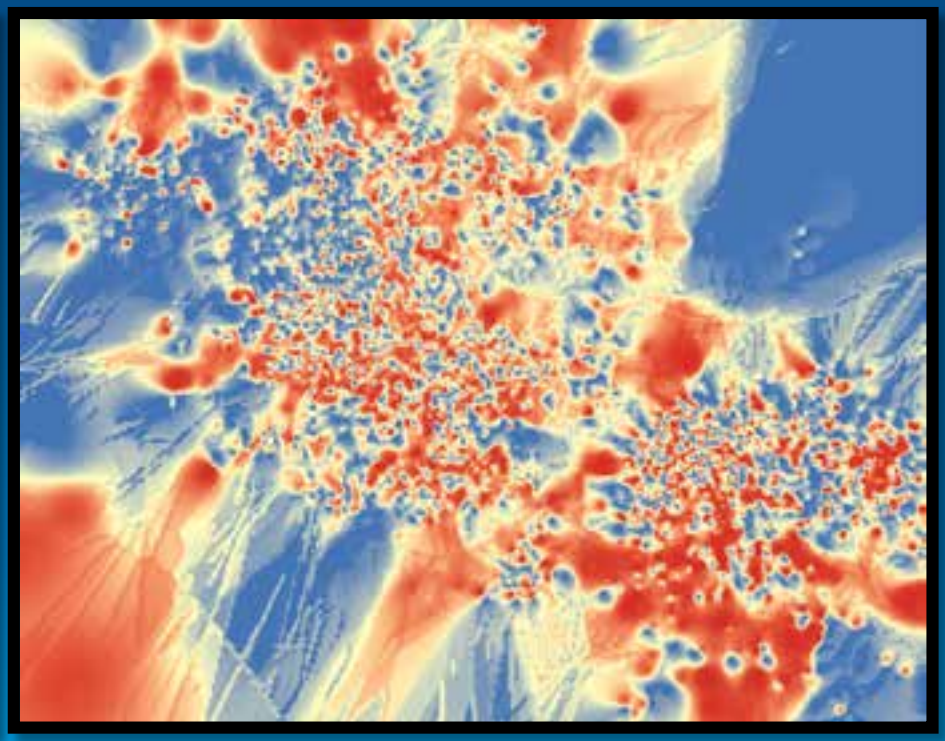
Hot Spot Maps Preliminary



Retrospective Hot Spot Maps



Hot Spot No Social Media



Hot Spot With Social Media

Regression Analysis

Summary of OLS Results

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]
Intercept	4.703064	0.019981	235.374965	0.000000*	0.020054	234.515788	0.000000*
WORDNET	0.104344	0.054382	1.918726	0.055029	0.053274	1.958633	0.050163

OLS Diagnostics

Input Features:	crimeTweetJoin	Dependent Variable:	CRIMETYPE
Number of Observations:	20000	Akaike's Information Criterion (AICc) [d]:	95412.149378
Multiple R-Squared [d]:	0.000184	Adjusted R-Squared [d]:	0.000134
Joint F-Statistic [e]:	3.681511	Prob(>F), (1,19998) degrees of freedom:	0.055033
Joint Wald Statistic [e]:	3.836245	Prob(>chi-squared), (1) degrees of freedom:	0.050156
Koenker (BP) Statistic [f]:	9.554071	Prob(>chi-squared), (1) degrees of freedom:	0.001995*
Jarque-Bera Statistic [g]:	1439.378234	Prob(>chi-squared), (2) degrees of freedom:	0.000000*

R-Squared High

Contributions and Outcomes

- **Predictive Crime Analysis via Social Media**
  - **Hypothesis: A GIS hot spot map or risk terrain model increases in precision and accuracy as a social media corpus is operationalization into GIS risk layers.**
  - **Independent Variable: Each different mapping treatment to produce a visualization**

  - **STAC: Spatial and Temporal Analysis of Crime: geographically locates the densest clusters of incidents on the map**

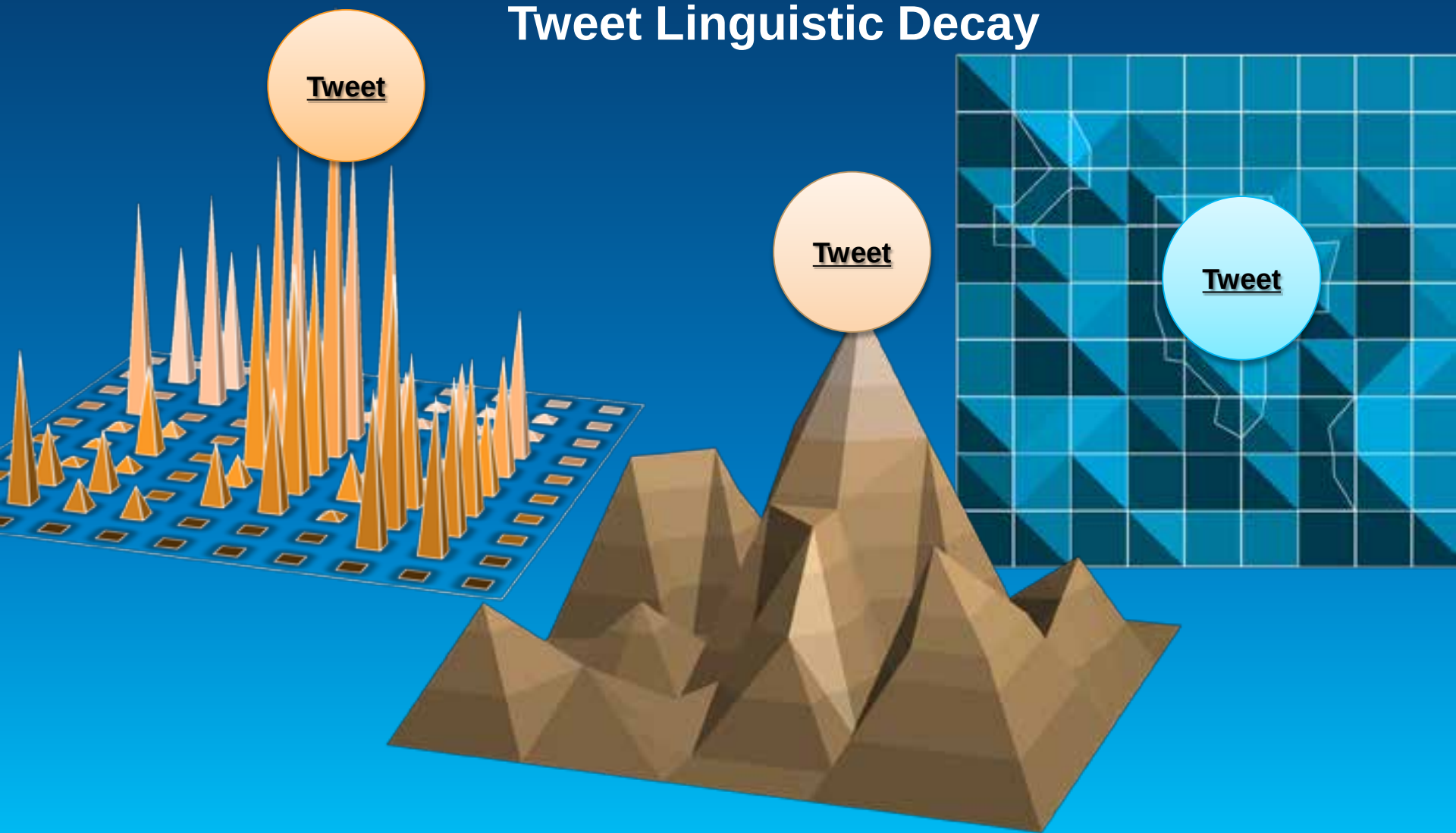
  - **Nnh: Nearest Neighbor Clustering: technique is based on a threshold distance to which the crime incidents are compared to identify clusters, e.g., city block**

  - **KDE: Kernel Density Estimation: continuous smoothed surface with variation of the density of crime over a study area**

  - **RTM: Two factors that have different operational spatial influences but can be spatially joined. The spatial overlap of risk factors creates a more risky environment where crime would be expected to occur in the future.**

Current Research College Crime

Future Research Prediction



Conclusion and Questions

- **Integrating domain specific and crime data**
- **A predictive social media artifact is possible**
- **GIS RTM artifact construction is encouraging**