Integrating Data from Social Media for Anticipatory Policing and Intelligence

Donald E. Brown and Matthew S. Gerber

Predictive Technology Laboratory
University of Virginia
Charlottesville, VA 22904

Acknowledgement: This work was partially supported by grants from the Army Research Laboratory and TSC.
Agenda

1. Criminal Site Selection
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - Temporal Predictions

2. Social Media Mining for Criminal Event Prediction
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
   - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
Predicting Crime Locations & Time

- **Point Process Models**

  - Predicting Crime Locations & Time
  - **Point Process Models**
    - Kernel Density Estimation (Eck et al. 2005, Gorr 2012)
    - Self-Exciting Point Process Models (Mohler 2011)
    - Criminal Site Selection (CSS)
      - Logistic Regression (Brown et al. 2004)
      - Spatial-Temporal GAM (Wang and Brown 2012)
    - Multilevel Group and Spatial-Temporal Models (Huddleston and Brown 2009, Fox and Brown, 2012)
    - Data Mining Methods (CART, MARS, Random Forest) (McCue 2007)
    - Kriging (Kerry et al. 2010)
Point Process Models
- Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
Predicting Crime Locations & Time

- Point Process Models
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)
Predicting Crime Locations & Time

- Point Process Models
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)
- Criminal Site Selection (CSS)

Integration of Social Data

D.E. Brown

CSS
Criminal Choice
Space
Anchor Points
Time

Media + CSS
Text Mining
Twitter + CSS
Small Set Examples
Chicago Examples

Asymmetric Threat Tracker (ATT)

Conclusions

Predictive Technology Laboratory, UVA
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)

- **Criminal Site Selection (CSS)**
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)

- **Criminal Site Selection (CSS)**
  - Logistic Regression (Brown et al. 2004)
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)

- **Criminal Site Selection (CSS)**
  - Logistic Regression (Brown et al. 2004)
  - Spatial-Temporal GAM (Wang and Brown 2012)
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)
- **Criminal Site Selection (CSS)**
  - Logistic Regression (Brown et al. 2004)
  - Spatial-Temporal GAM (Wang and Brown 2012)
  - Multilevel Group and Spatial-Temporal Models (Huddleston and Brown 2009, Fox and Brown, 2012)
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)

- **Criminal Site Selection (CSS)**
  - Logistic Regression (Brown et al. 2004)
  - Spatial-Temporal GAM (Wang and Brown 2012)
  - Multilevel Group and Spatial-Temporal Models (Huddleston and Brown 2009, Fox and Brown, 2012)

- **Data Mining Methods** (CART, MARS, Random Forest) (McCue 2007)
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)
- **Criminal Site Selection (CSS)**
  - Logistic Regression (Brown et al. 2004)
  - Spatial-Temporal GAM (Wang and Brown 2012)
  - Multilevel Group and Spatial-Temporal Models (Huddleston and Brown 2009, Fox and Brown, 2012)
- **Data Mining Methods (CART, MARS, Random Forest)** (McCue 2007)
- **Kriging** (Kerry et. al. 2010)
Predicting Crime Locations & Time

- **Point Process Models**
  - Kernel Density Estimation (Eck et. al. 2005, Gorr 2012)
  - Self-Exciting Point Process Models (Mohler 2011)

- **Criminal Site Selection (CSS)**
  - Logistic Regression (Brown et al. 2004)
  - Spatial-Temporal GAM (Wang and Brown 2012)
  - Multilevel Group and Spatial-Temporal Models (Huddleston and Brown 2009, Fox and Brown, 2012)

- **Data Mining Methods** (CART, MARS, Random Forest) (McCue 2007)
- Kriging (Kerry et. al. 2010)
Criminal Choice

- Cash Present
- Easy Entry
- Not Occupied
- Movements Hidden
- Guardian
- Distance
- Stereo

Predictive Technology Laboratory, UVA
Models of Criminal Choice

Crime based on decisions by criminals.
Models of Criminal Choice

- Crime based on decisions by criminals.
- Model the criminal choice process and use the models to predict.
Models of Criminal Choice

- Crime based on decisions by criminals.
- Model the criminal choice process and use the models to predict.
- Random Utility Maximization (McFadden, 1973)
Crime based on decisions by criminals.

Model the criminal choice process and use the models to predict.

Random Utility Maximization (McFadden, 1973)

\[ U(d, s_i) = V(d, s_i) + \varepsilon(d, s_i) \]
Models of Criminal Choice

- Crime based on decisions by criminals.
- Model the criminal choice process and use the models to predict.
- Random Utility Maximization (McFadden, 1973)

\[ U(d, s_i) = V(d, s_i) + \varepsilon(d, s_i) \]
- Individual \( d \), location, \( s_i \)
Models of Criminal Choice

- Crime based on decisions by criminals.
- Model the criminal choice process and use the models to predict.
- Random Utility Maximization (McFadden, 1973)
  \[ U(d, s_i) = V(d, s_i) + \varepsilon(d, s_i) \]
  - Individual \( d \), location, \( s_i \)
  - \( V(d, s_i) \) is deterministic
Models of Criminal Choice

- Crime based on decisions by criminals.
- Model the criminal choice process and use the models to predict.
- Random Utility Maximization (McFadden, 1973)
  \[ U(d, s_i) = V(d, s_i) + \varepsilon(d, s_i) \]
  - Individual \( d \), location, \( s_i \)
  - \( V(d, s_i) \) is deterministic
  - \( \varepsilon(d, s_i) \) is stochastic
Models of Criminal Choice

- Crime based on decisions by criminals.
- Model the criminal choice process and use the models to predict.
- Random Utility Maximization (McFadden, 1973)
  \[ U(d, s_i) = V(d, s_i) + \varepsilon(d, s_i) \]
  - Individual \( d \), location, \( s_i \)
  - \( V(d, s_i) \) is deterministic
  - \( \varepsilon(d, s_i) \) is stochastic
- With choice set \( A \), common assumptions lead to
  \[ P(a_i|A, d) = \frac{\exp(V(d, s_i))}{\sum_{a_j \in A} \exp(V(d, s_j))} \]
Agenda

1. **Criminal Site Selection**
   - Criminal Choice
   - **Spatial Predictions**
     - Anchor Point Discovery or Geographic Profiling
     - Temporal Predictions

2. **Social Media Mining for Criminal Event Prediction**
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
   - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
Criminal Site Selection

Time Axis

Feature Space

Geographic Space

Predictive Technology Laboratory, UVA
Spatial-Temporal Generalized Additive Models (STGAM)

- Extends previous work to use spatial-temporal data to predict criminal incidents.
Spatial-Temporal Generalized Additive Models (STGAM)

- Extends previous work to use spatial-temporal data to predict criminal incidents.
- Generalized additive models provide an effective approach that captures the nonlinearities in the prediction problem.
Spatial-Temporal Generalized Additive Models (STGAM)

- Extends previous work to use spatial-temporal data to predict criminal incidents.
- Generalized additive models provide an effective approach that captures the nonlinearities in the prediction problem.
- The STGAM provides a framework to incorporate unstructured data, such as, text.
Process: Data

Predictive Technology Laboratory, UVA
Process: Feature Computation

- Area to Model
- GIS Layers
- Demographic Data
- Incident Data

Grids
Distance Features
Demographic Features
Incident Feature

for each time interval t
Process: Modeling

Area to Model  
GIS Layers  
Demographic Data  
Incident Data

- Grids  
- Distance Features  
- Demographic Features  
- Incident Feature

for each time interval t

The Spatio-Temporal GAM
Process: Prediction

The Spatio-Temporal GAM

Input

Output

Predictive Technology Laboratory, UVA
Spatio-Temporal Generalized Additive Models (STGAM)

\[ \mathcal{g} (E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( Y_{NN(g),t-1:t-K} \right) \]

\( \mathcal{g} \): a link function
Spatio-Temporal Generalized Additive Models (STGAM)

STGAM:

$$g \left( E[y_{g,t}] \right) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right)$$

- $g$: a link function
- $y_{g,t}$: the response variable at location $s_g$ and time $t$
Spatio-Temporal Generalized Additive Models (STGAM)

STGAM:

\[ g \left( E[y_{g,t}] \right) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{g,t,n}) + f_T \left( f_{\text{feature}}(Y_{NN(g), t-1:t-K}) \right) \]

- \( g \): a link function
- \( y_{g,t} \): the response variable at location \( s_g \) and time \( t \)
- \( E[y_{g,t}] \): the expectation of the response \( y_{g,t} \)
Spatio-Temporal Generalized Additive Models (STGAM)

STGAM:

\[ g( E[y_{g,t}] ) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}}(Y_{\text{NN}}(g), t-1:t-K) \right) \]

- \( g \): a link function
- \( y_{g,t} \): the response variable at location \( s_g \) and time \( t \)
- \( E[y_{g,t}] \): the expectation of the response \( y_{g,t} \)
- \( f_0 \): a smooth function over the entire space \( S \)
Spatio-Temporal Generalized Additive Models (STGAM)

STGAM:

\[ g(E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right) \]

- \( g \): a link function
- \( y_{g,t} \): the response variable at location \( s_g \) and time \( t \)
- \( E[y_{g,t}] \): the expectation of the response \( y_{g,t} \)
- \( f_0 \): a smooth function over the entire space \( \mathcal{S} \)
- \( N \): the total number of structured features
Spatio-Temporal Generalized Additive Models (STGAM)

**STGAM:**

\[ \mathcal{g}(E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right) \]

- \( \mathcal{g} \): a link function
- \( y_{g,t} \): the response variable at location \( s_g \) and time \( t \)
- \( E[y_{g,t}] \): the expectation of the response \( y_{g,t} \)
- \( f_0 \): a smooth function over the entire space \( \mathcal{S} \)
- \( N \): the total number of structured features
- \( f_n \): the smooth function of the \( n^{th} \) structured feature
Spatio-Temporal Generalized Additive Models (STGAM)

STGAM:

\[ g(E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T\left(f^{\text{feature}}(Y_{NN(g),t-1:t-K})\right) \]

- \( g \): a link function
- \( y_{g,t} \): the response variable at location \( s_g \) and time \( t \)
- \( E[y_{g,t}] \): the expectation of the response \( y_{g,t} \)
- \( f_0 \): a smooth function over the entire space \( \mathbb{S} \)
- \( N \): the total number of structured features
- \( f_n \): the smooth function of the \( n^{th} \) structured feature
- \( x_{[g,t,n]} \): the \( n^{th} \) structured feature associated with location \( s_g \) and time \( t \)

Predictive Technology Laboratory, UVA
Spatio-Temporal Generalized Additive Models (STGAM) for Threat Surfaces

STGAM:

\[ \mathbb{E}(y_{g,t}) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right) \]

- \( f_T \): a smooth function to be estimated from data about the temporal effect
Spatio-Temporal Generalized Additive Models (STGAM) for Threat Surfaces

**STGAM:**

\[
\mathbb{g}(E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T\left(f_{\text{feature}}(Y_{NN(g),t-1:t-K})\right)
\]

- \(f_T\): a smooth function to be estimated from data about the temporal effect
- \(f_{\text{feature}}\): a predefined function to calculate temporal features from the past response values:
Spatio-Temporal Generalized Additive Models (STGAM) for Threat Surfaces

STGAM:

$$\mathbb{E}(y_{g,t}) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right)$$

- $f_T$: a smooth function to be estimated from data about the temporal effect
- $f_{\text{feature}}$: a predefined function to calculate temporal features from the past response values:
  - For classification problems, a good choice of $f_{\text{feature}}$ is $f_{\text{feature}} = \kappa_{g,t}$, where $\kappa$ is a temporal dummy variable.
Spatio-Temporal Generalized Additive Models (STGAM) for Threat Surfaces

STGAM:

\[ g(E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T(f_{\text{feature}}(Y_{NN(g)},t-1:t-K)) \]

- \( f_T \): a smooth function to be estimated from data about the temporal effect
- \( f_{\text{feature}} \): a predefined function to calculate temporal features from the past response values:
  - For classification problems, a good choice of \( f_{\text{feature}} \) is \( f_{\text{feature}} = \kappa_{g,t} \), where \( \kappa \) is a temporal dummy variable.
- \( NN(g) \): spatial grids close to \( s_g \)
Spatio-Temporal Generalized Additive Models (STGAM) for Threat Surfaces

**STGAM:**

\[
\mathbb{E}(E[y_{g,t}]) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f^{\text{feature}}(Y_{NN(g),t-1:t-K}) \right)
\]

- **\(f_T\):** a smooth function to be estimated from data about the temporal effect
- **\(f^{\text{feature}}\):** a predefined function to calculate temporal features from the past response values:
  - For classification problems, a good choice of \(f^{\text{feature}}\) is \(f^{\text{feature}} = \kappa_{g,t}\), where \(\kappa\) is a temporal dummy variable.
- **\(NN(g)\):** spatial grids close to \(s_g\)
- **\(Y_{NN(g),t-1:t-K}\):** the previous values of \(Y\) in the neighborhood \(NN(g)\) between time \(t - 1\) and \(t - K\)
Properties of STGAM

- STGAM has properties of GAM:
Properties of STGAM

- STGAM has properties of GAM:
  - Can include many predictors.
Properties of STGAM

- STGAM has properties of GAM:
  - Can include many predictors.
  - Captures nonlinearities.
Properties of STGAM

- STGAM has properties of GAM:
  - Can include many predictors.
  - Captures nonlinearities.
  - Reasonable interpretability.
Properties of STGAM

- STGAM has properties of GAM:
  - Can include many predictors.
  - Captures nonlinearities.
  - Reasonable interpretability.
- We can efficiently estimate STGAM parameters with penalized iteratively reweighted least squares.
Evaluation of Threat Surfaces with Surveillance Plots

- **HRP**: percentage of the high risk area predicted by the models

\[ \text{HRP}_\delta = \frac{\| \{(s_g, t) | \Pr(y_g, t = 1) > \delta \} \|}{\| S^* \times T^* \|} \]

- **TIP**: percentage of incidents (from test set) that happened within the high risk area

\[ \text{TIP}_\delta = \frac{\| \{ y_g, t = 1 | (s_g, t) \subset \{(s_g, t) | \Pr(y_g, t = 1) > \delta \} \} \|}{\| \{ y_g, t = 1 \} \|} \]
Evaluation of Threat Surfaces with Surveillance Plots

- **HRP**: percentage of the high risk area predicted by the models
  \[
  \text{HRP}_\delta = \frac{\| \{(s_g, t) | Pr(y_g, t = 1) > \delta \} \|}{\| S^* \times T^* \|}
  \]

- **TIP**: percentage of incidents (from test set) that happened within the high risk area
  \[
  \text{TIP}_\delta = \frac{\| \{ y_g, t = 1 | (s_g, t) \subset \{(s_g, t) | Pr(y_g, t = 1) > \delta \} \} \|}{\| \{ y_g, t = 1 \} \|}
  \]

- Based on the above plot, about 70% of real incidents happened within the top 20% area with the highest predicted risk.
Example: Charlottesville, VA

- Criminal incident data for Charlottesville, Virginia from April 2001 to February 2005;
Example: Charlottesville, VA

- Criminal incident data for Charlottesville, Virginia from April 2001 to February 2005;
- Geographic information for Charlottesville from multiple sources;

Legend:
- Incident(B&E)
- Small businesses
- Hospitals
- K 12
- Interstates Roads
- Local Roads
- Charlottesville
Example: Charlottesville, VA

- Criminal incident data for Charlottesville, Virginia from April 2001 to February 2005;
- Geographic information for Charlottesville from multiple sources;
- Demographic data for Charlottesville measured in census block groups;

Legend
- Incident (B&E)
- Small businesses
- Hospitals
- K 12
- Interstates Roads
- Local Roads
- Charlottesville

Example: Charlottesville, VA

Criminal incident data for Charlottesville, Virginia from April 2001 to February 2005;

Geographic information for Charlottesville from multiple sources;

Demographic data for Charlottesville measured in census block groups;
Example: Charlottesville, VA

- Criminal incident data for Charlottesville, Virginia from April 2001 to February 2005;
- Geographic information for Charlottesville from multiple sources;
- Demographic data for Charlottesville measured in census block groups;
- Grid size: 0.02mile × 0.02mile;
Example: Charlottesville, VA

- Criminal incident data for Charlottesville, Virginia from April 2001 to February 2005;
- Geographic information for Charlottesville from multiple sources;
- Demographic data for Charlottesville measured in census block groups;
- Grid size: 0.02mile × 0.02mile;
- Time interval: 1 month.
Threat Surface Evaluations: S-T GAM vs. Spatial GLM and Hot Spot Model
Criminal Site Selection

1. Criminal Choice
2. Spatial Predictions
3. Anchor Point Discovery or Geographic Profiling
4. Temporal Predictions

Social Media Mining for Criminal Event Prediction

1. Text Mining
2. Social Media for CSS
3. Crime Prediction using Small Set (Media) Twitter Posts
4. Crime Prediction using Chicago Twitter Posts

Asymmetric Threat Tracker (ATT)

Conclusions
Geographic profiling is an investigative technique used by police that uses the known locations of a crime series to determine a serial offender’s anchor point (residence or workplace).
Geographic profiling is an investigative technique used by police that uses the known locations of a crime series to determine a serial offender’s anchor point (residence or workplace).

APD models a criminal’s journey to crime, the general tendency for criminal activity to decrease as distance from a criminal’s anchor point increases.
Geographic profiling is an investigative technique used by police that uses the known locations of a crime series to determine a serial offender’s anchor point (residence or workplace).

APD models a criminal’s journey to crime, the general tendency for criminal activity to decrease as distance from a criminal’s anchor point increases.

Most APD methods produce a jeopardy surface to identify the most likely area for the offender’s anchor point.
Example: Vancouver (Rossmo, 2000)

Figure 6.1


Problems with APD Methods

- Most accurate approach is simple centro-graphic modeling such as calculation of the Center of Minimum Distance (CMD) (Paulson, 2006; Levine and Block, 2011).
- Distance-decay mathematical models do not consider the impact of the environment on the criminal’s choice set - i.e. the target backcloth is ignored.
CSS for APD

- Link crime series in the training data set
CSS for APD

- Link crime series in the training data set
- Develop a CSS Model for a given geographic area
CSS for APD

- Link crime series in the training data set
- Develop a CSS Model for a given geographic area
- Link a crime series for the “unknown” group
CSS for APD

- Link crime series in the training data set
- Develop a CSS Model for a given geographic area
- Link a crime series for the “unknown” group
- Using Bayes Rule, calculate the joint posterior probability density for the group anchor point from the group’s linked crime series for every grid cell
CSS for APD

- Link crime series in the training data set
- Develop a CSS Model for a given geographic area
- Link a crime series for the “unknown” group
- Using Bayes Rule, calculate the joint posterior probability density for the group anchor point from the group’s linked crime series for every grid cell
- Map the joint posterior probability density for every grid cell to produce a jeopardy surface for the group anchor point
Santa Anna Gang Data

Gang Incident Tracking System (GITS) Database for Santa Ana, California.
Santa Anna Gang Data

- Gang Incident Tracking System (GITS) Database for Santa Ana, California.
Example Anchor Point Prediction
## Cross-Validation Results

<table>
<thead>
<tr>
<th>Gang</th>
<th>Crime Count</th>
<th>Error CMD</th>
<th>Distance CSS</th>
<th>Search CMD</th>
<th>Cost CSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>953</td>
<td>710</td>
<td>1200</td>
<td>444</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>363</td>
<td>535</td>
<td>176</td>
<td>334</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>40</td>
<td>100</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1563</td>
<td>1535</td>
<td>3132</td>
<td>2761</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>410</td>
<td>69</td>
<td>196</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>637</td>
<td>273</td>
<td>552</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>1277</td>
<td>1112</td>
<td>2144</td>
<td>1653</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>2276</td>
<td>2334</td>
<td>6908</td>
<td>6920</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>34</td>
<td>20</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>574</td>
<td>534</td>
<td>420</td>
<td>339</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>1982</td>
<td>1957</td>
<td>5272</td>
<td>4421</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>2328</td>
<td>1807</td>
<td>7136</td>
<td>5564</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>1844</td>
<td>1632</td>
<td>4644</td>
<td>4219</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>601</td>
<td>462</td>
<td>480</td>
<td>327</td>
</tr>
<tr>
<td>15</td>
<td>32</td>
<td>470</td>
<td>73</td>
<td>292</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>780</td>
<td>519</td>
<td>760</td>
<td>330</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>451</td>
<td>174</td>
<td>248</td>
<td>44</td>
</tr>
</tbody>
</table>

**Average**

|             | 13 | 976 | 814 | 1974 | 1616 |
1. **Criminal Site Selection**
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - **Temporal Predictions**

2. **Social Media Mining for Criminal Event Prediction**
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
   - Crime Prediction using Chicago Twitter Posts

3. **Asymmetric Threat Tracker (ATT)**

4. **Conclusions**
Temporal predictions can be leveraged to provide:
Criminal Event Forecasting

- Temporal predictions can be leveraged to provide:
  - More informative, well-calibrated forecasts to support resource allocation decisions at the operational level of security planning
Temporal predictions can be leveraged to provide:

- More informative, well-calibrated forecasts to support resource allocation decisions at the operational level of security planning
- Improved (geographic) mission assignment at the operational level by linking decisions about spatial areas of responsibility to forecasts of times of future criminal activity.
Temporal predictions can be leveraged to provide:

- More informative, well-calibrated forecasts to support resource allocation decisions at the operational level of security planning
- Improved (geographic) mission assignment at the operational level by linking decisions about spatial areas of responsibility to forecasts of times of future criminal activity.

Our methodology formally links the analytic products used for spatial event prediction with time series models for forecasting incident intensities.
Example: Pittsburgh Burglaries 2008

- 3093 burglaries in 2008
- Source: (Gorr 2012)
Goal: Accurately forecast the burglary count in each of the city’s 6 precincts.
Goal: Accurately forecast the burglary count in each of the city’s 6 precincts.

Let $N_t$ be the city count at time $t$ and $N_{kt}$ be the precinct count, $k = 1, \ldots, 6$. 
Use STGAM at the tactical level to map the spatial process.

\[ \pi_{g,t} = g \left( E[y_{g,t}] \right) = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T \left( f_{\text{feature}} \left( Y_{NN(g),t-1:t-K} \right) \right) \]
Use STGAM at the tactical level to map the spatial process.

\[ \pi_{g,t} = \mathcal{G}(E[y_{g,t}]) \]

\[ = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) \]

\[ + f_T \left( f_{\text{feature}}(Y_{NN}(g), t-1:t-K) \right) \]
**Approach - Step 1**

- Use STGAM at the tactical level to map the spatial process.

\[
\pi_{g,t} = \mathcal{G} \left( E[y_{g,t}] \right) \\
= f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) \\
+ f_T \left( f_{\text{feature}} \left( Y_{NN(g),t-1:t-K} \right) \right)
\]
Use STGAM at the tactical level to map the spatial process.

\[ \pi_{g,t} = \mathcal{G} \left( E[y_{g,t}] \right) \]

\[ = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{g,t,n}) \]

\[ + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right) \]
Approach - Step 1

- Use STGAM at the tactical level to map the spatial process.

\[ \pi_{g,t} = \mathcal{G}(E[y_{g,t}]) \]

\[ = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) \]

\[ + f_T \left( f_{\text{feature}}(Y_{NN(g),t-1:t-K}) \right) \]
Approach - Step 2

Risk-weight the threat scores for the six precincts \((k = 1, \ldots, 6)\) by integrating over the spatial density function.

\[
W_k = \frac{\pi_k}{\sum_{i=1}^{I} \pi_i} = \frac{\sum_{j \in D_j} \pi_j}{\sum_{i=1}^{I} \pi_i}
\]
Use exponential smoothing at the city level to forecast the city-wide, weekly total

\[ F_t = \alpha N_{t-1} + (1 - \alpha) F_{t-1} \]

Forecast the weekly precinct amount using a spatial threat-weighting of the city-level forecast.

\[ F_{kt} = w_k F_t \]
1. Point Layer: Burglaries Observed
Modeling the Spatial Process

2. STGAM

\[ \pi_{g,t} = f_0(s_g) + \sum_{n=1}^{N} f_n(x_{[g,t,n]}) + f_T\left(f_{\text{feature}}(Y_{\text{NN}(g),t-1:t-K})\right) \]
Modeling the Spatial Process

3. Threat Surface
Modeling the Spatial Process

4. Precinct Layer
Modeling the Spatial Process

5. Precinct Weights

\[ w_k = \frac{\sum_{j \in D_j} \pi_j}{\sum_{i=1}^{I} \pi_i} \]

6. Precinct Weight Table

<table>
<thead>
<tr>
<th>Precinct</th>
<th>( w_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.13%</td>
</tr>
<tr>
<td>2</td>
<td>11.20%</td>
</tr>
<tr>
<td>3</td>
<td>24.18%</td>
</tr>
<tr>
<td>4</td>
<td>19.14%</td>
</tr>
<tr>
<td>5</td>
<td>24.39%</td>
</tr>
<tr>
<td>6</td>
<td>5.96%</td>
</tr>
</tbody>
</table>
Pittsburgh Burglary Results in RMSE

<table>
<thead>
<tr>
<th>Precinct</th>
<th>Exp.</th>
<th>ARIMA</th>
<th>TSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.96</td>
<td>3.96</td>
<td>3.88</td>
</tr>
<tr>
<td>2</td>
<td>2.99</td>
<td>3.23</td>
<td>2.91</td>
</tr>
<tr>
<td>3</td>
<td>4.24</td>
<td>4.35</td>
<td>3.84</td>
</tr>
<tr>
<td>4</td>
<td>3.96</td>
<td>4.15</td>
<td>3.72</td>
</tr>
<tr>
<td>5</td>
<td>3.97</td>
<td>4.36</td>
<td>3.82</td>
</tr>
<tr>
<td>6</td>
<td>1.62</td>
<td>1.72</td>
<td>1.75</td>
</tr>
</tbody>
</table>

Aggregates

Root Mean Square Error (RMSE)

Comparison of Aggregate RMSE Over Time

Predictive Technology Laboratory, UVA
TSF reduces RMSE but not significant at $p < 0.05$. 
Pittsburgh Results Summary

- TSF reduces RMSE but not significant at $p < 0.05$.
- Takes about 10 weeks for the TSF to dominate the exponential smoothing approach.
Pittsburgh Results Summary

- TSF reduces RMSE but not significant at $p < 0.05$.
- Takes about 10 weeks for the TSF to dominate the exponential smoothing approach.
- Models can be calibrated.
Agenda

1. Criminal Site Selection
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - Temporal Predictions

2. Social Media Mining for Criminal Event Prediction
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
   - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
Text mining:

- General steps of text mining:
  1. Transform text into structured data
  2. Build mathematical models on structured data

Text mining models:
- Term-Frequency/Inverse-Document-Frequency (TF-IDF) (Buckley & Salton, 1988)
- Latent Semantic Indexing (LSI) (Deerwester, et al., 1990)
- Probabilistic LSI (pLSI) (Hoffman, 1999)
- Multinomial Inverse Regression, (MIR) (Taddy, 2012)

Predictive Technology Laboratory, UVA
Text mining:
- General steps of text mining:
  1. Transform text into structured data

Text mining: General steps of text mining:

1. Transform text into structured data
Text mining:

- General steps of text mining:
  1. Transform text into structured data
  2. Build mathematical models on structured data

Text mining:

General steps of text mining:

1. Transform text into structured data
2. Build mathematical models on structured data
Text Mining

- Text mining:
  - General steps of text mining:
    1. Transform text into structured data
    2. Build mathematical models on structured data
  - Text mining models:
Text Mining

Text mining:

General steps of text mining:
1. Transform text into structured data
2. Build mathematical models on structured data

Text mining models:
- Term-Frequency/Inverse-Document-Frequency (TF-IDF) (Buckley & Salton, 1988)
Text Mining

- Text mining:
  - General steps of text mining:
    1. Transform text into structured data
    2. Build mathematical models on structured data
  - Text mining models:
    - Term-Frequency/Inverse-Document-Frequency (TF-IDF) (Buckley & Salton, 1988)
    - Latent Semantic Indexing (LSI) (Deerwester, et al., 1990)
Text Mining

Text mining:
- General steps of text mining:
  1. Transform text into structured data
  2. Build mathematical models on structured data

Text mining models:
- Term-Frequency/Inverse-Document-Frequency (TF-IDF) (Buckley & Salton, 1988)
- Latent Semantic Indexing (LSI) (Deerwester, et al., 1990)
- Probabilistic LSI (pLSI) (Hoffman, 1999)
Text Mining

- **Text mining:**
  - General steps of text mining:
    1. Transform text into structured data
    2. Build mathematical models on structured data
  - Text mining models:
    - Term-Frequency/Inverse-Document-Frequency (TF-IDF) (Buckley & Salton, 1988)
    - Latent Semantic Indexing (LSI) (Deerwester, et al., 1990)
    - Probabilistic LSI (pLSI) (Hoffman, 1999)
Text mining:

General steps of text mining:
1. Transform text into structured data
2. Build mathematical models on structured data

Text mining models:
- Term-Frequency/Inverse-Document-Frequency (TF-IDF) (Buckley & Salton, 1988)
- Latent Semantic Indexing (LSI) (Deerwester, et al., 1990)
- Probabilistic LSI (pLSI) (Hoffman, 1999)
- Multinomial Inverse Regression, (MIR) (Taddy, 2012)
Markov models: N-gram language model and hidden Markov model (HMM)
Natural Language Processing

- Markov models: N-gram language model and hidden Markov model (HMM)
- Semantic role labeling (SRL)
Agenda

1. Criminal Site Selection
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - Temporal Predictions

2. Social Media Mining for Criminal Event Prediction
   - Text Mining
   - **Social Media for CSS**
     - Crime Prediction using Small Set (Media) Twitter Posts
     - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
Extracting Structured Data from Text

Documents -> Unstructured Text -> Structured Vectors

- (0.1, 0, 2, 0.1, 0, 21, 0.35)
- (0.25, 0, 0, 0, 0.75)
- (0.1, 0, 0, 0.6, 0.3)
- ...
- ...
- (0.32, 0.18, 0, 0.5)
Extracting Structured Data from Text

Structure text by vectors

\[ x_{[g, t, txt]} = \langle \text{word}_1, \text{word}_2, \ldots, \text{word}_{ng} \rangle \]

↓ (step 1)

\[ x_{[g, t, txt]} = (\text{word}'_1, \text{word}'_2, \ldots, \text{word}'_{ng}) \]

↓ (step 2)

\[ x_{[g, t, txt]} = (\text{topic}_1, \ldots, \text{topic}_m) \]

The critical parts are:
1. how to extract important words from each document?
2. how to extract topics from words?

Unstructured Text

Structured Vectors
Extracting Structured Data from Text

Structure text by vectors

\[ x_{[g,t,txt]} = \langle \text{word}_1, \text{word}_2, \cdots, \text{word}_{ng} \rangle \]
\[ \Downarrow \text{ (step 1)} \]
\[ x_{[g,t,txt]} = (\text{word}'_1, \text{word}'_2, \cdots, \text{word}'_{ng}) \]
\[ \Downarrow \text{ (step 2)} \]
\[ x_{[g,t,txt]} = (\text{topic}_1, \cdots, \text{topic}_m) \]

The critical parts are:

1. how to *extract important words* from each document?
2. how to *extract topics* from words?
Extracting Structured Data from Text

Structure text by vectors

\[ x_{[g,t,txt]} = \langle \text{word}_1, \text{word}_2, \ldots, \text{word}_{ng} \rangle \]
\[ \downarrow \text{ (step 1)} \]
\[ x_{[g,t,txt]} = (\text{word}_1', \text{word}_2', \ldots, \text{word}_{ng}') \]
\[ \downarrow \text{ (step 2)} \]
\[ x_{[g,t,txt]} = (\text{topic}_1, \ldots, \text{topic}_m) \]

The critical parts are:

1. how to extract important words from each document? Semantic Role Labeling (SRL)
2. how to extract topics from words? Latent Dirichlet Allocation Models (LDA)
The Latent Dirichlet Allocation Model (LDA)

- LDA is a popular generative probabilistic language model.
The Latent Dirichlet Allocation Model (LDA)

- LDA is a popular generative probabilistic language model.
- It assumes a document is a bag of words.

LDA is a popular generative probabilistic language model.
It assumes a document is a bag of words.
The Latent Dirichlet Allocation Model (LDA)

- LDA is a popular generative probabilistic language model.
- It assumes a document is a bag of words.
- It doesn’t consider syntax or semantics.

LDA is a popular generative probabilistic language model. It assumes a document is a bag of words. It doesn’t consider syntax or semantics.
The Latent Dirichlet Allocation Model (LDA)

- LDA is a popular generative probabilistic language model.
- It assumes a document is a bag of words.
- It doesn’t consider syntax or semantics.
- Incorporating semantic analysis with LDA can improve the performance.

LDA is a popular generative probabilistic language model. It assumes a document is a bag of words. It doesn’t consider syntax or semantics. Incorporating semantic analysis with LDA can improve the performance.
LDA is a popular generative probabilistic language model. It assumes a document is a bag of words. It doesn’t consider syntax or semantics. Incorporating semantic analysis with LDA can improve the performance. Semantic analysis can help decide which words are important in the topics.
Semantic Role Labeling (SRL)

- Semantic role labeling (SRL) is a natural language processing technique.
Semantic Role Labeling (SRL)

- Semantic role labeling (SRL) is a natural language processing technique.
- SRL automatically extracts:
Semantic Role Labeling (SRL)

- Semantic role labeling (SRL) is a natural language processing technique.
- SRL automatically extracts:
  - the verbal and nominal predicates (usually called events) mentioned a sentence,
Semantic Role Labeling (SRL) is a natural language processing technique. SRL automatically extracts:

- the verbal and nominal predicates (usually called events) mentioned in a sentence,
- the entities associated with the events,
Semantic Role Labeling (SRL)

- Semantic role labeling (SRL) is a natural language processing technique.
- SRL automatically extracts:
  - the verbal and nominal predicates (usually called events) mentioned a sentence,
  - the entities associated with the events,
  - and the roles of the entities with respect to the events.
Semantic Role Labeling (SRL)

- Semantic role labeling (SRL) is a natural language processing technique.
- SRL automatically extracts:
  - the verbal and nominal predicates (usually called events) mentioned a sentence,
  - the entities associated with the events,
  - and the roles of the entities with respect to the events.

Example:
Semantic role labeling (SRL) is a natural language processing technique. SRL automatically extracts:

- the verbal and nominal predicates (usually called events) mentioned in a sentence,
- the entities associated with the events,
- and the roles of the entities with respect to the events.

Example:

An Example of SRL Analysis

\[ e_1: \text{warning \ TRAFFIC} \] \[ e_1: \text{ALERT} \]: \[ e_2: \text{entity \ Rt. \ 20} \] \[ e_2: \text{closed} \]
\[ e_2: \text{cause \ due \ to \ a \ wreck} \].
Semantic Role Labeling-Based Latent Dirichlet Allocation (SRL-LDA)

- SRL extracts important words mentioned in a document.
Semantic Role Labeling-Based Latent Dirichlet Allocation (SRL-LDA)

- SRL extracts important words mentioned in a document.
- LDA extracts latent variables about the important words.
Semantic Role Labeling-Based Latent Dirichlet Allocation (SRL-LDA)

- SRL extracts important words mentioned in a document.
- LDA extracts latent variables about the important words.
- Mathematically, SRL-LDA can be formulated as follows:

\[
Pr(D) = Pr(w) = \prod_{w_i \in w^*} \sum_{t=1}^{T} Pr(w_i|z_i = t) \cdot Pr(z_i = t) \cdot \prod_{w_i \in w - w^*} Pr(w_i|SRL)
\]
Semantic Role Labeling-Based Latent Dirichlet Allocation (SRL-LDA)

- SRL extracts important words mentioned in a document.
- LDA extracts latent variables about the important words.
- Mathematically, SRL-LDA can be formulated as follows:

\[
Pr(D) = Pr(w) = \prod_{w_i \in w^*} \sum_{t=1}^{T} Pr(w_i | z_i = t) \cdot Pr(z_i = t) \cdot \prod_{w_i \in w \setminus w^*} Pr(w_i | SRL)
\]

- A document \(d\) can be represented by the document-topic distribution \(\theta_d\).
SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
Properties of SRL-LDA

- SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
- SRL-LDA can partly solve the synonymy and polysemy problems.
Properties of SRL-LDA

- SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
- SRL-LDA can partly solve the synonymy and polysemy problems.
- SRL-LDA can be estimated easily and efficiently with current algorithms.
Properties of SRL-LDA

- SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
- SRL-LDA can partly solve the synonymy and polysemy problems.
- SRL-LDA can be estimated easily and efficiently with current algorithms.
- Compared to LDA,
Properties of SRL-LDA

- SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
- SRL-LDA can partly solve the synonymy and polysemy problems.
- SRL-LDA can be estimated easily and efficiently with current algorithms.
- Compared to LDA,
  - SRL-LDA can utilize semantic information of words in documents.
Properties of SRL-LDA

- SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
- SRL-LDA can partly solve the synonymy and polysemy problems.
- SRL-LDA can be estimated easily and efficiently with current algorithms.
- Compared to LDA,
  - SRL-LDA can utilize semantic information of words in documents.
  - SRL-LDA provides more meaningful topics.
Properties of SRL-LDA

- SRL-LDA can represent the unstructured textual data with meaningful vectors in low dimensions.
- SRL-LDA can partly solve the synonymy and polysemy problems.
- SRL-LDA can be estimated easily and efficiently with current algorithms.
- Compared to LDA,
  - SRL-LDA can utilize semantic information of words in documents.
  - SRL-LDA provides more meaningful topics.
  - SRL-LDA is more suitable for the problems with a limited number of documents.
Agenda

1. Criminal Site Selection
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - Temporal Predictions

2. Social Media Mining for Criminal Event Prediction
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
   - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
We applied SRL-LDA models to predict daily hit and run incidents in Charlottesville, VA, using Twitter posts.
We applied SRL-LDA models to predict daily hit and run incidents in Charlottesville, VA, using Twitter posts.

Data:
We applied SRL-LDA models to predict daily hit and run incidents in Charlottesville, VA, using Twitter posts.

Data:
- Twitter posts from CBS19 and actual hit-and-run incidents in Charlottesville, Virginia from February 22, 2011 to October 21, 2011
We applied SRL-LDA models to predict daily hit and run incidents in Charlottesville, VA, using Twitter posts.

Data:
- Twitter posts from CBS19 and actual hit-and-run incidents in Charlottesville, Virginia from February 22, 2011 to October 21, 2011
- Training data: the data before September 17, 2011
- Test data: the data after September 18, 2011
Semantic role labeling:
- The verb-based SRL by Punyakanok et al.\(^1\)
- The noun-based SRL by Gerber and Chai\(^2\)

---


\(^2\) M. Gerber, J. Chai, and A. Meyers. The role of implicit argumentation in nominal SRL. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL,
Semantic role labeling:
- The verb-based SRL by Punyakanok et al.\(^1\)
- The noun-based SRL by Gerber and Chai\(^2\)

Topic Extraction: 10-topic LDA model

---


\(^2\) M. Gerber, J. Chai, and A. Meyers. The role of implicit argumentation in nominal SRL. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL,
Semantic role labeling:
- The verb-based SRL by Punyakanok et al.\(^1\)
- The noun-based SRL by Gerber and Chai\(^2\)

Topic Extraction: 10-topic LDA model

Predictive Model:

\[
\text{logit} \left( Pr[y_{d+1} = 1] \right) = \beta_0 + \beta_1 T_{d,1} + \cdots + \beta_k T_{d,k}
\]

---


\(^2\) M. Gerber, J. Chai, and A. Meyers. The role of implicit argumentation in nominal SRL. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL,
Results: Topics from SRL-LDA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>close</td>
<td>announce</td>
<td>arrest</td>
<td>plan</td>
<td>say</td>
<td>expect</td>
<td>report</td>
<td>say</td>
<td>accuse</td>
<td>win</td>
</tr>
<tr>
<td>fire</td>
<td>announce</td>
<td>suspect</td>
<td>sentence</td>
<td>report</td>
<td>remain</td>
<td>close</td>
<td>found</td>
<td>make</td>
<td>cbs</td>
</tr>
<tr>
<td>crash</td>
<td>say</td>
<td>charge</td>
<td>kill</td>
<td>student</td>
<td>include</td>
<td>claim</td>
<td>die</td>
<td>traffic</td>
<td>sports</td>
</tr>
<tr>
<td>charge</td>
<td>make</td>
<td>conference</td>
<td>use</td>
<td>vote</td>
<td>rbs</td>
<td>warning</td>
<td>crash</td>
<td>break</td>
<td>come</td>
</tr>
<tr>
<td>look</td>
<td>search</td>
<td>death</td>
<td>plead</td>
<td>update</td>
<td>close</td>
<td>confirm</td>
<td>investigate</td>
<td>trial</td>
<td>hold</td>
</tr>
<tr>
<td>delay</td>
<td>hanchettjim</td>
<td>murder</td>
<td>life</td>
<td>tell</td>
<td>close</td>
<td>block</td>
<td>fall</td>
<td>set</td>
<td>approve</td>
</tr>
<tr>
<td>cause</td>
<td>confirm</td>
<td>shoot</td>
<td>ask</td>
<td>hear</td>
<td>cancel</td>
<td>robbery</td>
<td>raise</td>
<td>service</td>
<td>help</td>
</tr>
<tr>
<td>come</td>
<td>hold</td>
<td>rsb</td>
<td>sell</td>
<td>work</td>
<td>release</td>
<td>issue</td>
<td>identify</td>
<td>hope</td>
<td>hold</td>
</tr>
<tr>
<td>injure</td>
<td>run</td>
<td>deny</td>
<td>statement</td>
<td>speak</td>
<td>follow</td>
<td>declare</td>
<td>seek</td>
<td>find</td>
<td>start</td>
</tr>
<tr>
<td>watch</td>
<td>move</td>
<td>receive</td>
<td>connection</td>
<td>head</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table**: Top 10 most likely words for each topic
Results: Topics from SRL-LDA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>close</td>
<td>announce</td>
<td>arrest</td>
<td>plan</td>
<td>say</td>
<td>expect</td>
<td>report</td>
<td>say</td>
<td>say</td>
<td>accuse</td>
</tr>
<tr>
<td>fire</td>
<td>make</td>
<td>suspect</td>
<td>sentence</td>
<td>report</td>
<td>remain</td>
<td>close</td>
<td>found</td>
<td>make</td>
<td>win</td>
</tr>
<tr>
<td>crash</td>
<td>conference</td>
<td>charge</td>
<td>kill</td>
<td>student</td>
<td>include</td>
<td>damage</td>
<td>die</td>
<td>traffic</td>
<td>cbs</td>
</tr>
<tr>
<td>charge</td>
<td>search</td>
<td>death</td>
<td>use</td>
<td>vote</td>
<td>rsb</td>
<td>open</td>
<td>crash</td>
<td>come</td>
<td>come</td>
</tr>
<tr>
<td>look</td>
<td>hanchettjim</td>
<td>murder</td>
<td>plead</td>
<td>update</td>
<td>warning</td>
<td>confirm</td>
<td>report</td>
<td>found</td>
<td>hold</td>
</tr>
<tr>
<td>delay</td>
<td>confirm</td>
<td>shoot</td>
<td>life</td>
<td>tell</td>
<td>close</td>
<td>block</td>
<td>robbery</td>
<td>raise</td>
<td>approve</td>
</tr>
<tr>
<td>cause</td>
<td>hold</td>
<td>rsb</td>
<td>ask</td>
<td>hear</td>
<td>cancel</td>
<td>robbery</td>
<td>issue</td>
<td>identify</td>
<td>help</td>
</tr>
<tr>
<td>come</td>
<td>deny</td>
<td>deny</td>
<td>sell</td>
<td>work</td>
<td>release</td>
<td>declare</td>
<td>follow</td>
<td>seek</td>
<td>start</td>
</tr>
<tr>
<td>injure</td>
<td>receive</td>
<td>connection</td>
<td>statement</td>
<td>speak</td>
<td>rain</td>
<td>follow</td>
<td>identify</td>
<td>seek</td>
<td>look</td>
</tr>
<tr>
<td>watch</td>
<td>move</td>
<td>connection</td>
<td>missing</td>
<td>head</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table: Top 10 most likely words for each topic

- Topic 1, 4, 6, and 8 were significantly related to the incident probability.
Results: Topics from SRL-LDA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
<th>Topic 9</th>
<th>Topic 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>close</td>
<td>announce</td>
<td>arrest</td>
<td>plan</td>
<td>say</td>
<td>expect</td>
<td>report</td>
<td>say</td>
<td>accuse</td>
<td>win</td>
</tr>
<tr>
<td>fire</td>
<td>say</td>
<td>suspect</td>
<td>sentence</td>
<td>report</td>
<td>remain</td>
<td>close</td>
<td>found</td>
<td>make</td>
<td>cbs</td>
</tr>
<tr>
<td>crash</td>
<td>make</td>
<td>charge</td>
<td>kill</td>
<td>student</td>
<td>include</td>
<td>damage</td>
<td>die</td>
<td>traffic</td>
<td>sports</td>
</tr>
<tr>
<td>charge</td>
<td>conference</td>
<td>death</td>
<td>use</td>
<td>vote</td>
<td>rsb</td>
<td>open</td>
<td>crash</td>
<td>break</td>
<td>come</td>
</tr>
<tr>
<td>look</td>
<td>search</td>
<td>murder</td>
<td>plead</td>
<td>update</td>
<td>warning</td>
<td>confirm</td>
<td>crash</td>
<td>set</td>
<td>hold</td>
</tr>
<tr>
<td>delay</td>
<td>hanchettjim</td>
<td>shoot</td>
<td>life</td>
<td>tell</td>
<td>close</td>
<td>block</td>
<td>raise</td>
<td>service</td>
<td>help</td>
</tr>
<tr>
<td>cause</td>
<td>confirm</td>
<td>shoot</td>
<td>ask</td>
<td>hear</td>
<td>cancel</td>
<td>robbery</td>
<td>identify</td>
<td>approve</td>
<td>start</td>
</tr>
<tr>
<td>come</td>
<td>hold</td>
<td>rsb</td>
<td>sell</td>
<td>work</td>
<td>release</td>
<td>issue</td>
<td>seek</td>
<td>help</td>
<td>start</td>
</tr>
<tr>
<td>injure</td>
<td>run</td>
<td>deny</td>
<td>statement</td>
<td>speak</td>
<td>rain</td>
<td>declare</td>
<td>seek</td>
<td>find</td>
<td>stop</td>
</tr>
<tr>
<td>watch</td>
<td>move</td>
<td>receive</td>
<td>missing</td>
<td>head</td>
<td>follow</td>
<td>follow</td>
<td>step</td>
<td>stop</td>
<td>lawsuit</td>
</tr>
</tbody>
</table>

Table: Top 10 most likely words for each topic

- Topic 1,4,6, and 8 were significantly related to the incident probability.
- The estimated model:

\[
\text{logit} \left( Pr[y_{t+1} = 1] \right) = 0.4 + 0.71 T_{t,1} + 0.88 T_{t,4} + 0.72 T_{t,6} + 0.61 T_{t,8}
\]

D.E. Brown

Predictive Technology Laboratory, UVA
Prediction Performance of SRL-LDA

![Graph showing the prediction performance of SRL-LDA with false positive rate on the x-axis and average true positive rate on the y-axis. The graph includes error bars for each data point.]

- Average true positive rate
- False positive rate

**Legend:**
- Predictive Technology Laboratory, UVA
Model Comparison: Prediction Performance of LDA only

![Graph showing the relationship between false positive rate and average true positive rate for LDA only prediction performance. The graph includes error bars and a trend line.]
Model Comparison: Prediction Performance of sLDA

---

Breaking and Entering Crimes

Breaking and entering crimes in Charlottesville, VA.
Breaking and Entering Crimes

- Breaking and entering crimes in Charlottesville, VA.
- Data:
  - Breaking and entering crimes that occurred in Charlottesville during the period of March 1st - October 31st, 2011
  - Twitter posts from CBS 19 during the same time period
  - Geographic and demographic data used in the H & R.
  - Grid size: 0.02 miles \( \times \) 0.02 miles
  - Time interval: 24 hours
Results

- Models in comparison:
  1. STGAM: STGAM using 18 distance and demographic features
  2. STGAM+SRL-LDA: STGAM using additional 10 textual features extracted by SRL-LDA
  3. STGAM+LDA: STGAM using additional 10 textual features extracted by LDA
Results

Models in comparison:

1. **STGAM**: STGAM using 18 distance and demographic features
2. **STGAM+SRL-LDA**: STGAM using additional 10 textual features extracted by SRL-LDA
3. **STGAM+LDA**: STGAM using additional 10 textual features extracted by LDA

![Graph showing comparative performance of different models](chart.png)
Agenda

1. Criminal Site Selection
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - Temporal Predictions

2. Social Media Mining for Criminal Event Prediction
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
     - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
Chicago Analysis

- Chicago crime data for 2012
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
- 30-day moving window for training set and next day test
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
- 30-day moving window for training set and next day test
- \( \sim 800k \) GPS-tagged tweets per 30 days
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
- 30-day moving window for training set and next day test
- $\sim 800k$ GPS-tagged tweets per 30 days
- Tokenization and word filtering using a Twitter-specific tagger from CMU
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
- 30-day moving window for training set and next day test
- \( \sim 800k \) GPS-tagged tweets per 30 days
- Tokenization and word filtering using a Twitter-specific tagger from CMU
- No word stemming
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
- 30-day moving window for training set and next day test
- \( \sim 800k \) GPS-tagged tweets per 30 days
- Tokenization and word filtering using a Twitter-specific tagger from CMU
- No word stemming
- 100-500 topic models, with 500 doing best
Chicago Analysis

- Chicago crime data for 2012
- Geo-tagged tweets
- 30-day moving window for training set and next day test
- \( \sim 800k \) GPS-tagged tweets per 30 days
- Tokenization and word filtering using a Twitter-specific tagger from CMU
- No word stemming
- 100-500 topic models, with 500 doing best
- KDE used instead of full CSS (i.e., ST-GAM)
Assault

Results for "ASSAULT" (464)

% incidents captured

% area surveilled

○ 0_500_PTLATT.Models (AUC=0.71)
△ 500_1500_PTLATT.Mod (AUC=0.74)
## Example Topics: Assault

<table>
<thead>
<tr>
<th>Topic 97</th>
<th>Topic 340</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.2</td>
<td>-3.2</td>
</tr>
<tr>
<td>monitor</td>
<td>cybercoders</td>
</tr>
<tr>
<td>tim</td>
<td>businessmgmt</td>
</tr>
<tr>
<td>himym</td>
<td>nettempsjobs</td>
</tr>
<tr>
<td>doo</td>
<td>virtual</td>
</tr>
<tr>
<td>sloot</td>
<td>cultural</td>
</tr>
<tr>
<td>jwow</td>
<td>foundation</td>
</tr>
<tr>
<td>gatorade</td>
<td>guidance</td>
</tr>
</tbody>
</table>
Liquor Law Violation

Results for "LIQUOR-LAW-VIOLATION" (18)

% incidents captured

% area surveilled

0.0 0.2 0.4 0.6 0.8 1.0

0.0 0.2 0.4 0.6 0.8 1.0

φ 500_PTL_ATT.Mod (AUC=0.68)
△ 500_2000_PTL_ATT.Mod (AUC=0.69)
Example Topics: Liquor Law Violation

<table>
<thead>
<tr>
<th>Topic 274</th>
<th>Topic 435</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>4.5</td>
</tr>
<tr>
<td>nah</td>
<td>walhi</td>
</tr>
<tr>
<td>oh</td>
<td>coffee</td>
</tr>
<tr>
<td>hoe</td>
<td>company</td>
</tr>
<tr>
<td>tf</td>
<td>gym</td>
</tr>
<tr>
<td>chill</td>
<td>metropolis</td>
</tr>
<tr>
<td>lmfaoooo</td>
<td>somalia</td>
</tr>
<tr>
<td>uh</td>
<td>morning</td>
</tr>
</tbody>
</table>
Integrating Social Data

CSS
Criminal Choice
Space
Anchor Points
Time

Media + CSS
Text Mining
Twitter + CSS
Small Set Examples
Chicago Examples

Asymmetric Threat Tracker (ATT)

Conclusions

Offenses Involving Children

Results for "OFFENSE-INVOLVING-CHILDREN" (70)

Predictive Technology Laboratory, UVA
Example Topics: Offenses Involving Children

<table>
<thead>
<tr>
<th>Topic 176</th>
<th>Topic 134</th>
</tr>
</thead>
<tbody>
<tr>
<td>-7.0</td>
<td>5.7</td>
</tr>
<tr>
<td>bank</td>
<td>follow</td>
</tr>
<tr>
<td>tax</td>
<td>angelo</td>
</tr>
<tr>
<td>court</td>
<td>share</td>
</tr>
<tr>
<td>branch</td>
<td>listen</td>
</tr>
<tr>
<td>questions</td>
<td>deniro</td>
</tr>
<tr>
<td>broke</td>
<td>capone</td>
</tr>
<tr>
<td>loan</td>
<td>prod</td>
</tr>
</tbody>
</table>
Results for "OTHER-OFFENSE" (508)

- 0_500_PTL.ATT.Models (AUC=0.68)
- 500_500_PTL.ATT.Mode (AUC=0.72)
Example Topics: Other Offenses

Topic 397 ($w=3.54862548797438$): chicago ohare ord airport international im il united time club morning pic hilton people city coffee good snow enjoying daily

Topic 33 ($w=-3.36817467301547$): nf yuh sweetheart vip turnup kenwood stoney 87th 4sho thestruggle stony gby tiff hp buddy walter pu lawno bringthachiback kp
Results for "BATTERY" (1395)

- 0_500_PTL_ATT.Models (AUC=0.71)
- 500_500_PTL_ATT.Mode (AUC=0.74)
Theft
Example Topics Theft

Topic 323 ($w = -2.2647679060852$): rabbit soxfest signal mega property verdict reddit alias photoshop upset dibs bobble hamm roberts educating mccarthy kombucha leavin grains scholars

Topic 246 ($w = -2.19336139576587$): jewel portillos longhorn steakhouse oscoc barracos ball sum1 deodorant dist region kmart charles chuck caravan centers baccis cheeses recruiting soiled
Integrating Social Data

D.E. Brown

CSS
Criminal Choice
Space
Anchor Points
Time

Media + CSS
Text Mining
Twitter + CSS
Small Set Examples
Chicago Examples

Asymmetric Threat Tracker (ATT)

Conclusions

Prostitution

Results for "PROSTITUTION" (47)

% incidents captured

% area surveilled

0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

0_500_PTL_ATT_Models (AUC=0.87)
500_1000_PTL_ATT_Mod (AUC=0.9)

Predictive Technology Laboratory, UVA
Integrating Social Data

D.E. Brown

CSS
Criminal Choice
Space
Anchor Points
Time

Media + CSS
Text Mining
Twitter + CSS
Small Set Examples
Chicago Examples

Asymmetric Threat Tracker (ATT)

Conclusions

---

Agenda

1. Criminal Site Selection
   - Criminal Choice
   - Spatial Predictions
   - Anchor Point Discovery or Geographic Profiling
   - Temporal Predictions

2. Social Media Mining for Criminal Event Prediction
   - Text Mining
   - Social Media for CSS
   - Crime Prediction using Small Set (Media) Twitter Posts
   - Crime Prediction using Chicago Twitter Posts

3. Asymmetric Threat Tracker (ATT)

4. Conclusions
Asymmetric Threat Tracker

ATT Engine

GIS + Models

Statistical Engine: R

Preprocessing

Prediction

ATT Library

Data ingest
AOI definition
Model definition

Model training
Model execution
Result visualization

Interfaces
- GUI application
- Web application
- Web service
Displays in Spinoza
Conclusions

- CSS has proven valuable for law enforcement and security.
Conclusions

- CSS has proven valuable for law enforcement and security.
- Incorporating text in CSS model has shown operational lift.
Conclusions

- CSS has proven valuable for law enforcement and security.
- Incorporating text in CSS model has shown operational lift.
- Next Steps
Conclusions

- CSS has proven valuable for law enforcement and security.
- Incorporating text in CSS model has shown operational lift.
- Next Steps
  - CSS models with Chicago tweets
Conclusions

- CSS has proven valuable for law enforcement and security.
- Incorporating text in CSS model has shown operational lift.

Next Steps
- CSS models with Chicago tweets
- Tweets and APD models

CSS has proven valuable for law enforcement and security. Incorporating text in CSS model has shown operational lift. Next Steps
- CSS models with Chicago tweets
- Tweets and APD models
Conclusions

- CSS has proven valuable for law enforcement and security.
- Incorporating text in CSS model has shown operational lift.

Next Steps
- CSS models with Chicago tweets
- Tweets and APD models
- Tweets and temporal models
Conclusions

- CSS has proven valuable for law enforcement and security.
- Incorporating text in CSS model has shown operational lift.
- Next Steps
  - CSS models with Chicago tweets
  - Tweets and APD models
  - Tweets and temporal models
- ATT will be available as open source, soon.