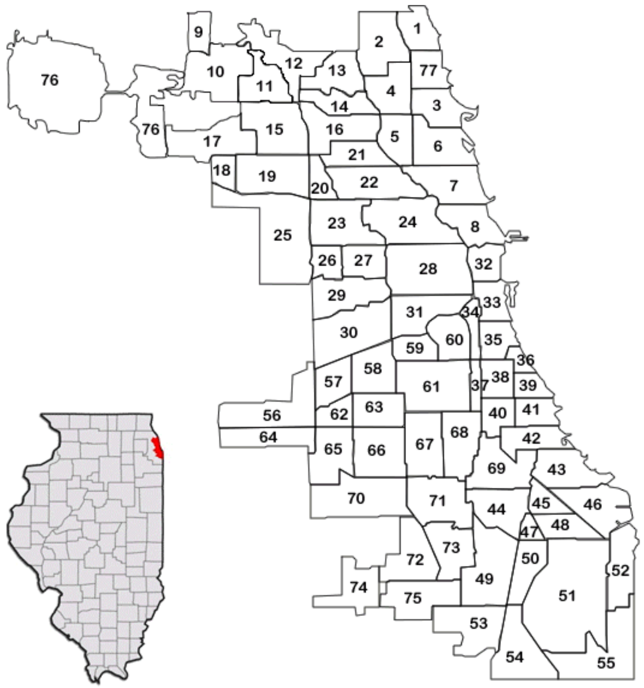
Detecting Spatially Localized Subsets of Leading Indicators for Event Prediction

Daniel B. Neill
Event and Pattern Detection Laboratory
H.J. Heinz III College, Carnegie Mellon University
neill@cs.cmu.edu

Joint work with Seth Flaxman, Amrut Nagasunder, Wil Gorr (CMU); Brett Goldstein (City of Chicago).

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330.

Background: Crime Prediction in Chicago



Since 2009, we have been working with the Chicago Police Department (CPD) to predict and prevent emerging clusters of violent crime.

Our new crime prediction methods have been incorporated into our **CrimeScan** software, run twice a day by CPD and used operationally for deployment of patrols.

From the Chicago Sun-Times, February 22, 2011:

"It was a bit like "Minority Report," the 2002 movie that featured genetically altered humans with special powers to predict crime. The CPD's new crime-forecasting unit was analyzing 911 calls and produced an intelligence report predicting a shooting would happen soon on a particular block on the South Side. Three minutes later, it did…"

CrimeScan

The key insight of our method is to **use detection for prediction**:

We can **detect emerging clusters** of various leading indicators (minor crimes, 911 calls, etc.) and use these to **predict** that a cluster of violent crime is likely to occur nearby.

Some advantages of the CrimeScan approach:

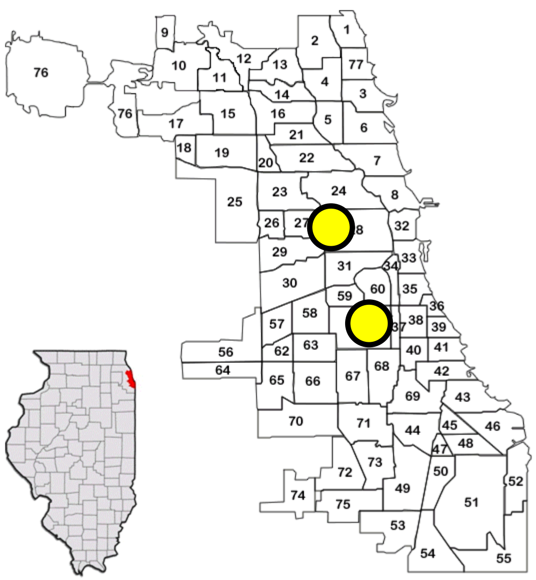
- Advance prediction (up to 1 week) with high accuracy.
- High spatial and temporal resolution (block x day).
- Predicting **emerging hot spots** of violence, as opposed to just identifying bad neighborhoods.

How to detect leading indicator clusters?
How to use these for prediction?
***** Which leading indicators to use? *****

CrimeScan: Cluster Detection

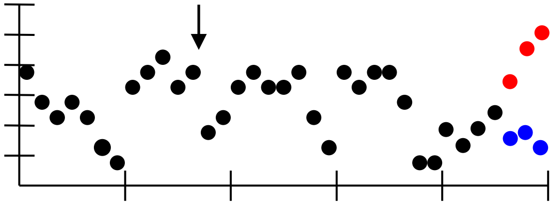
We aggregate daily counts for each leading indicator at the block level, and search for **clusters** of nearby blocks with recent counts that are significantly higher than expected.

Imagine moving a circular window around the city, allowing the center, radius, and temporal duration to vary.



Is there any spatial window and duration T such that counts have been higher than expected for the last T days?

Time series of past counts



Actual counts of last 3 days

Expected counts of last 3 days

CrimeScan: Cluster Detection

We aggregate daily counts for each leading indicator at the block level, and search for **clusters** of nearby blocks with recent counts that are significantly higher than expected.

Imagine moving a circular window around the city, allowing the center, radius, and temporal duration to vary.

We find the highest-scoring space-time regions, where the score of a region is computed by the likelihood ratio statistic.

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)}$$

Alternative hypothesis: cluster in region S

Null hypothesis: no clusters

These are the most likely clusters; we compute the p-value of each cluster by randomization, and report clusters with p-values < α.

CrimeScan: Prediction

The current deployed version of CrimeScan uses a very simple prediction rule:

"Areas which are closer to a significant cluster of any of the monitored LI are assumed more likely to have a spike in VC within the next 1 week."

Total proximity to leading indicator clusters is computed using a Gaussian kernel:

$$\text{score} = \sum \exp(-d_i^2/2)$$

(d_i is distance to the i^{th} leading indicator cluster)

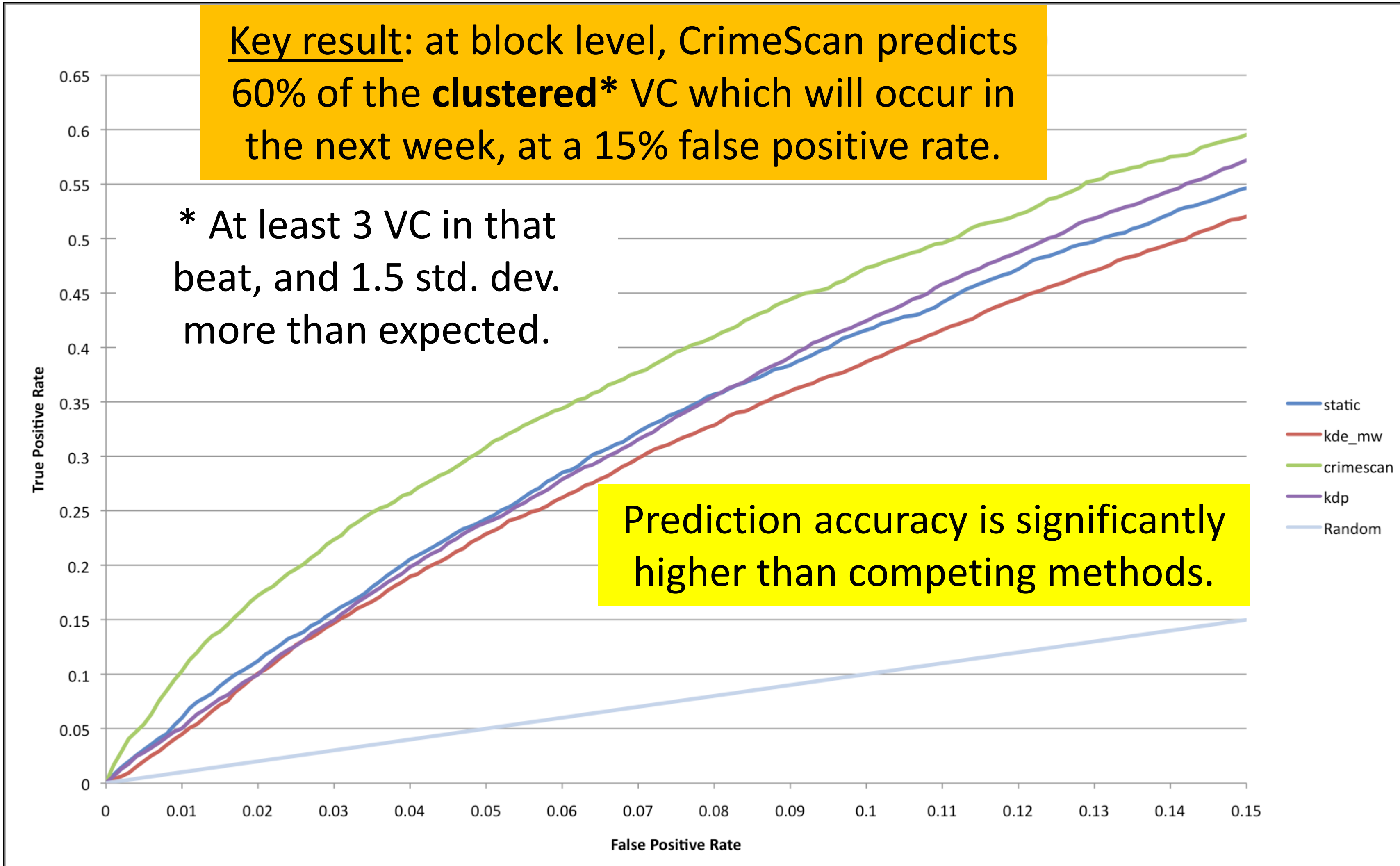
We are also investigating the use of logistic regression for prediction (results not shown).

CrimeScan: Preliminary Results

<u>Key result</u>: at block level, CrimeScan predicts 60% of the **clustered*** VC which will occur in the next week, at a 15% false positive rate.

* At least 3 VC in that beat, and 1.5 std. dev. more than expected.

Prediction accuracy is significantly higher than competing methods.



Which Predictors to Use?

Challenge #1: hundreds of possible predictors, including minor crimes, 911 emergency calls, 311 calls for service, etc.

Challenge #2: different data sources, or combinations of sources, may be predictive in different areas of the city.

We wish to learn which combinations of sources are predictive, and where, using **cross-correlation analysis** of historical data.

Typical formulation: given an independent variable time series X and a dependent variable time series Y, maximize correlation between X and lagged Y, over a range of lags $L = L_{min} \ldots L_{max}$.

For which subset of leading indicators, and which subset of locations, is cross-correlation maximized?

Maximizing cross-correlation

Given monitored locations s_i (i = 1..N), we observe the multiple independent variable time series $x_{i,m}^t$ (m = 1..M) and the dependent variable time series y_i^t at each location.

Our goal is to maximize the correlation r(X, Y) over **all subsets** of leading indicators, **all proximity-constrained subsets** of locations, and **all lags** L = $L_{min}..L_{max}$:

$$\max_{S \subseteq \{s_1..s_N\},\, D \subseteq \{d_1..d_M\},\, L \in \{L_{\min}..L_{\max}\}} r(X, Y)$$

$$\text{where } X = \sum_{d_m \in D} \sum_{s_i \in S} x_{i,m}^t \text{ and } Y = \sum_{s_i \in S} y_i^{t+L}$$

aggregated independent var. time series

aggregated, lagged dependent var. time series

Maximizing cross-correlation

How to **efficiently** maximize correlation r(D, S, L) over $2^N \times 2^M$ subsets of locations and predictors?

Iterative framework (outer loop):

1) Randomly initialize subset of streams D.
2) Optimize over locations: $S = \arg\max_S r(D, S, L)$
3) Optimize over streams: $D = \arg\max_D r(D, S, L)$
4) Repeat steps 2-3 until convergence.
5) Repeat steps 1-4 for R random restarts.
6) Repeat steps 1-5 for each lag L.

Optimizing over subsets of streams

Given fixed S and L, we want to find a set D to maximize r(D, S, L).
We write: $X = \sum_{d_m \in D} X_m$; $X_m = \sum_{s_i \in S} x_{i,m}$; and $Y = \sum_{s_i \in S} y_i$.

Then we maximize $r(D \mid S, L) = r(X, Y) = \dfrac{X \cdot Y}{\|X\| \|Y\|} = \dfrac{\sum_{d_m \in D} (X_m \cdot Y)}{\left\|\sum_{d_m \in D} X_m\right\| \|Y\|}$

Now we would like to write this expression as a **convex function** of two **additive sufficient statistics**, $r(D \mid S, L) = F(C, B)$ where $C = \sum_{d_m \in D} C_m$ and $B = \sum_{d_m \in D} B_m$.

If we can do this, we can show that the optimal D consists of the k streams with highest ratio C_m / B_m, for some $k \in \{1..N\}$.

This **linear-time subset scanning (LTSS)** property allows us to find the exact maximum over the 2^M subsets in $O(M \log M)$.

Optimizing over subsets of streams

Given fixed S and L, we want to find a set D to maximize r(D, S, L).
We write: $X = \sum_{d_m \in D} X_m$; $X_m = \sum_{s_i \in S} x_{i,m}$; and $Y = \sum_{s_i \in S} y_i$.

Then we maximize $r(D \mid S, L) = r(X, Y) = \dfrac{X \cdot Y}{\|X\|\,\|Y\|} = \dfrac{\sum_{d_m \in D} (X_m \cdot Y)}{\left\|\sum_{d_m \in D} X_m\right\| \|Y\|}$

Now we would like to write this expression as a **convex function** of two **additive sufficient statistics**, $r(D \mid S, L) = F(C, B)$ where $C = \sum_{d_m \in D} C_m$ and $B = \sum_{d_m \in D} B_m$.

We can write $r(D \mid S, L) = \dfrac{C}{\|Y\|\sqrt{B}}$

additive sufficient statistic: $C = \sum C_m = \sum (X_m \cdot Y)$

not an additive sufficient statistic!
$B = \sum_{d_m \in D} (X_m \cdot X_m)$ **+** $\boldsymbol{\sum_{d_i, d_j \in D, i \neq j} (X_i \cdot X_j)}$

<u>Solution</u>: we can approximate the all-pairs computation using the **average** dot product of stream d_m with an arbitrary set of streams.

Iterative average dot product (IADP)

We can w

Since the optimal subset D is unknown, we compute the average dot product of each stream D_m with an arbitrary subset of streams D' ($D_m \notin D'$): $Q_m = \frac{1}{|D'|} \sum_{d_i \in D'} (X_m \cdot X_i)$

Then $B \approx \sum_{d_m \in D} B_m$, where $B_m = X_m \cdot X_m + (|D|-1)\, Q_m$. We have approximated r(D | S, L) with a function which can be exactly and efficiently optimized using the LTSS property!

$B = \sum_{d_m \in D} (X_m \cdot X_m) + \sum_{d_i, d_j \in D,\, i \neq j} (\mathbf{X_i} \cdot \mathbf{X_j})$

However, the approximation may be poor when D' is far from D.

Our solution is to **iterate**: at each step, we set D' equal to the best subset D found on the previous step, and repeat until convergence.

Optimizing over subsets of locations

Given fixed D and L, we want to find a set S to maximize r(D, S, L).
We write: $X = \sum_{s_i \in S} X_i$; $X_i = \sum_{d_m \in D} x_{i,m}$; and $Y = \sum_{s_i \in S} y_i$.

Then we maximize r(S | D, L) = r(X, Y) =
$$\frac{\left(\sum_{s_i \in S} X_i\right) \cdot \left(\sum_{s_i \in S} Y_i\right)}{\left\|\sum_{s_i \in S} X_i\right\| \left\|\sum_{s_i \in S} Y_i\right\|}$$

This expression is more difficult to approximate by a function that satisfies LTSS because we have summations both over X_i and Y_i, resulting in "all-pairs" computations both in the numerator and in the denominator.

The iterative average dot product method can also be applied in this setting, but now we must make five approximations instead of one. Details are provided in the full paper (Flaxman and Neill, 2012, submitted).

Results: Comparison of Methods

For IADP and several competing methods, we maximized cross-correlation over subsets of predictors (and locations) for each of the 77 Chicago neighborhoods.

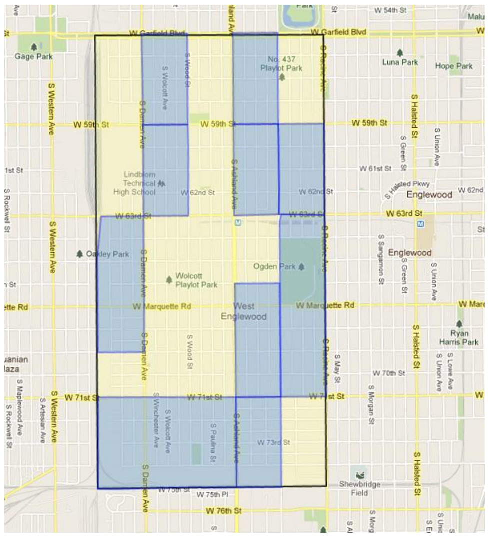
We then computed the average cross-correlation found by each method.

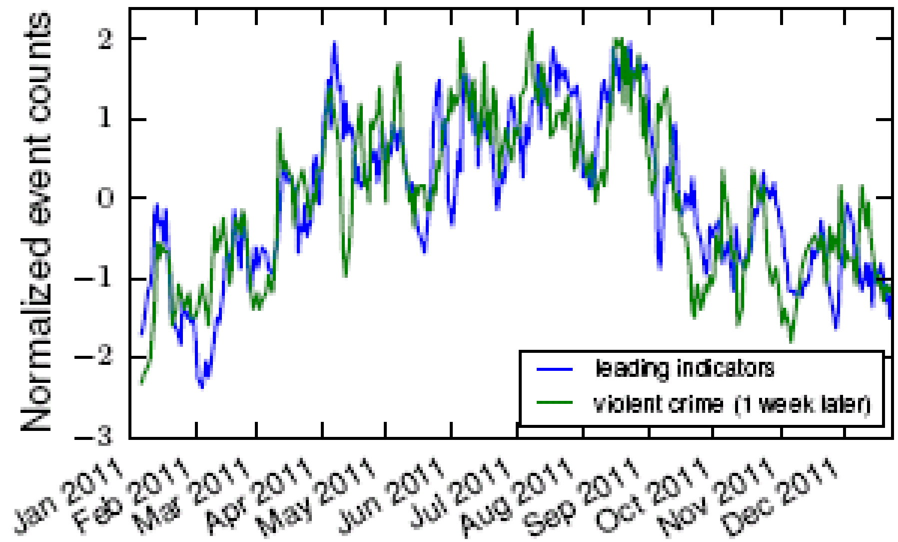
Method	Average cross-correlation
IADP, searching over subsets of census tracts within each neighborhood.	.546
IADP, treating each neighborhood as a single location.	.423
Google Correlate	.404
LASSO	.325

← By jointly optimizing over subsets of locations and streams, we find areas with much stronger cross-correlations between independent and dependent variables.

← Improved feature selection: Searching over subsets of streams for each neighborhood, we find significantly higher correlations than previous methods.

Results: Exploratory Analysis





Considering all subsets of census tracts within each of the 77 neighborhoods of Chicago, 28 different potential predictors, and a 1-week lag, we found a correlation of **r = .786** between violent crime and a subset of 12 leading indicators, for 10 census tracts in the West Englewood neighborhood.

Total run time for all 77 neighborhoods was **2.1 hours**.

Conclusions and Ongoing Work

CrimeScan is a new and powerful methodology for crime prediction which has been very successful in practice.

We are in the process of extending CrimeScan by developing novel methods to choose an optimal set of spatially varying leading indicators for prediction.

Our results suggest that different subsets of leading indicators have high predictive accuracy in different areas, and that our new methods can efficiently optimize cross-correlation over **subsets** of locations and streams.

Our next step is to determine whether the optimized, spatially varying subset of leading indicators can be used to improve the overall predictive accuracy of CrimeScan.