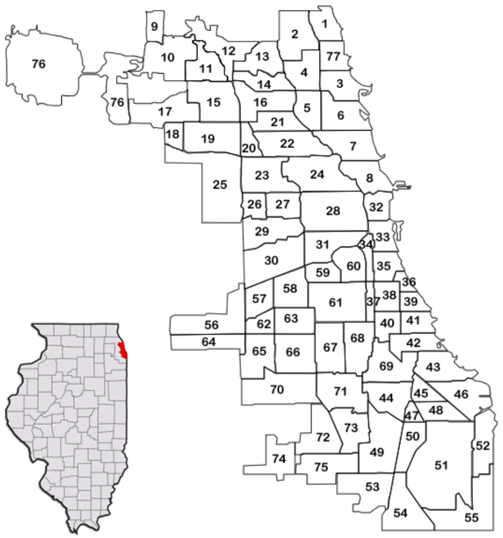
Predicting and Preventing Emerging Outbreaks of Crime

Mallory Nobles
Daniel B. Neill
Seth Flaxman

Event and Pattern Detection Laboratory
Carnegie Mellon University
Heinz College

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

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, which has been used operationally by CPD 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.

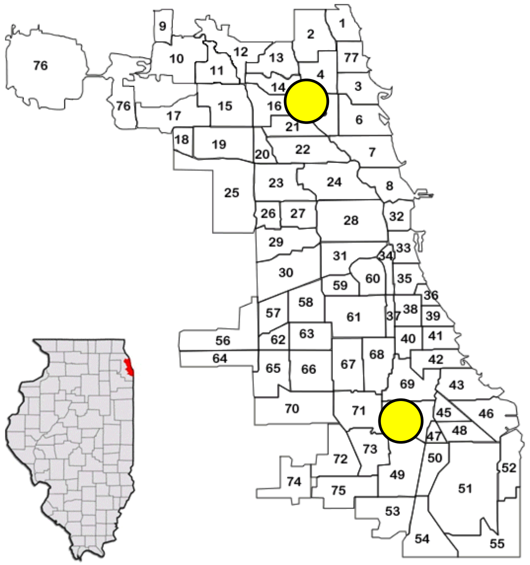
How to detect leading indicator clusters?
How to use these for prediction?
Which leading indicators to use?

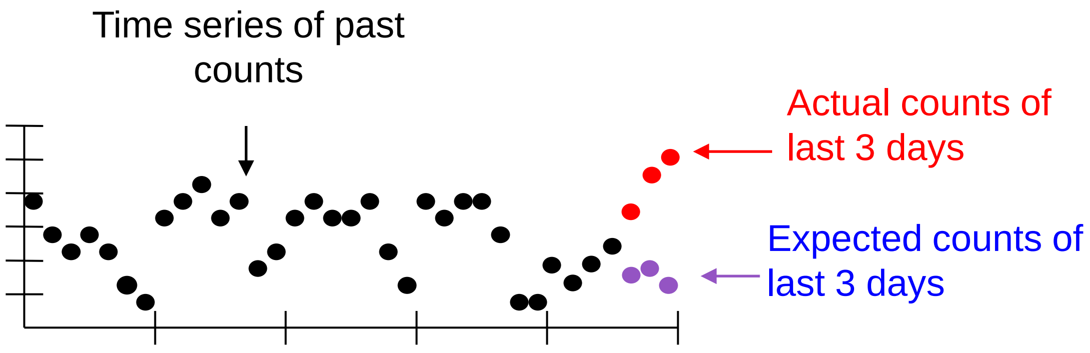
Cluster Detection

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 significantly higher than expected for the last T days?





Expectation Based Scan Statistic

We find the highest-scoring space-time regions, where the score of a region is computed by the likelihood ratio statistic, $F(S) = \frac{\Pr(Data \mid H_1(S))}{\Pr(Data \mid H_0(S))}$.

The expectation based scan statistic assumes counts are Poisson distributed: $\mathbf{c_i^t \sim \text{Poisson}(q_i^t b_i^t)}$

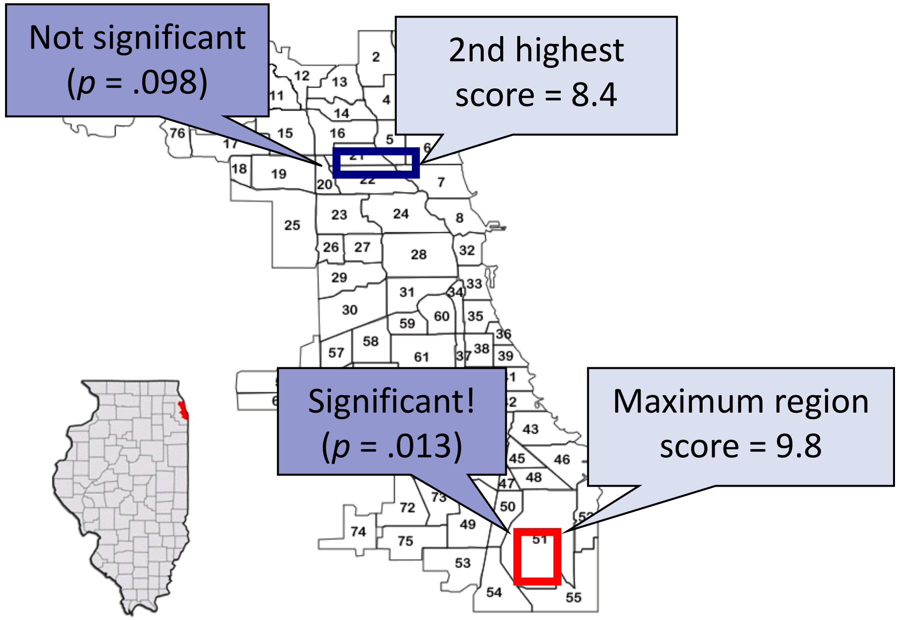
- q_i^t is relative risk.
- b_i^t is **expected count** under H_0, estimated by time series analysis of historical data.

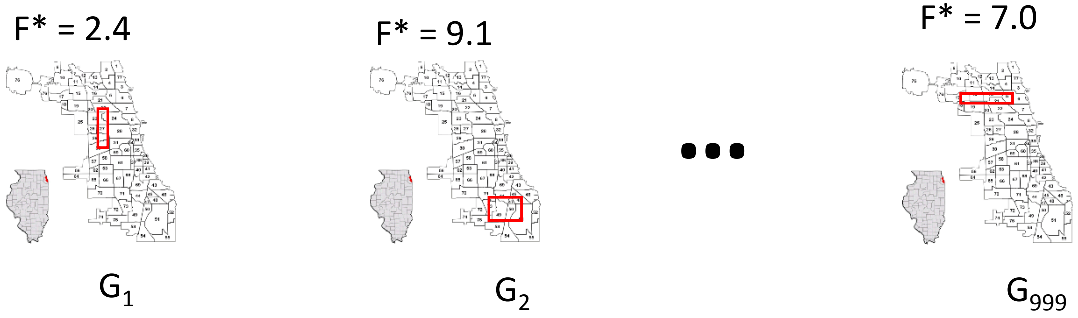
Null Hypothesis $H_0(S)$	Alternative Hypothesis $H_1(S)$
Assumes no clusters.	Assumes cluster in region S
We expect counts to be equal to baselines: $q_i^t = 1$ everywhere.	We expect increased risk in space-time region S: $q_i^t = q_{in}$ in S, for $q_{in} > 1$, and $q_i^t = 1$ outside.

This formulation gives a simple and computationally efficient likelihood ratio statistic:

$$\mathrm{F}(S) = \left(\frac{C}{B}\right)^C e^{B-C}, \text{ where } C = \sum_S c_i^t \text{ and } B = \sum_S b_i^t.$$

Randomization Testing





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

- Randomly generate counts for R = 999 replica datasets under H_0 (assuming no events).
- Find maximum region score $F^* = \max_S F(S)$ of each replica.
- p-value of region S = $(R_B+1) / (R+1)$, where R_B = # of replicas with $F^* \geq F(S)$.
- All regions with p-values < α are significant at level α.

Prediction

We are currently investigating several different methods for prediction of clusters of violent crime, using the detected leading indicator clusters as features of a predictive model.

Model-based prediction:

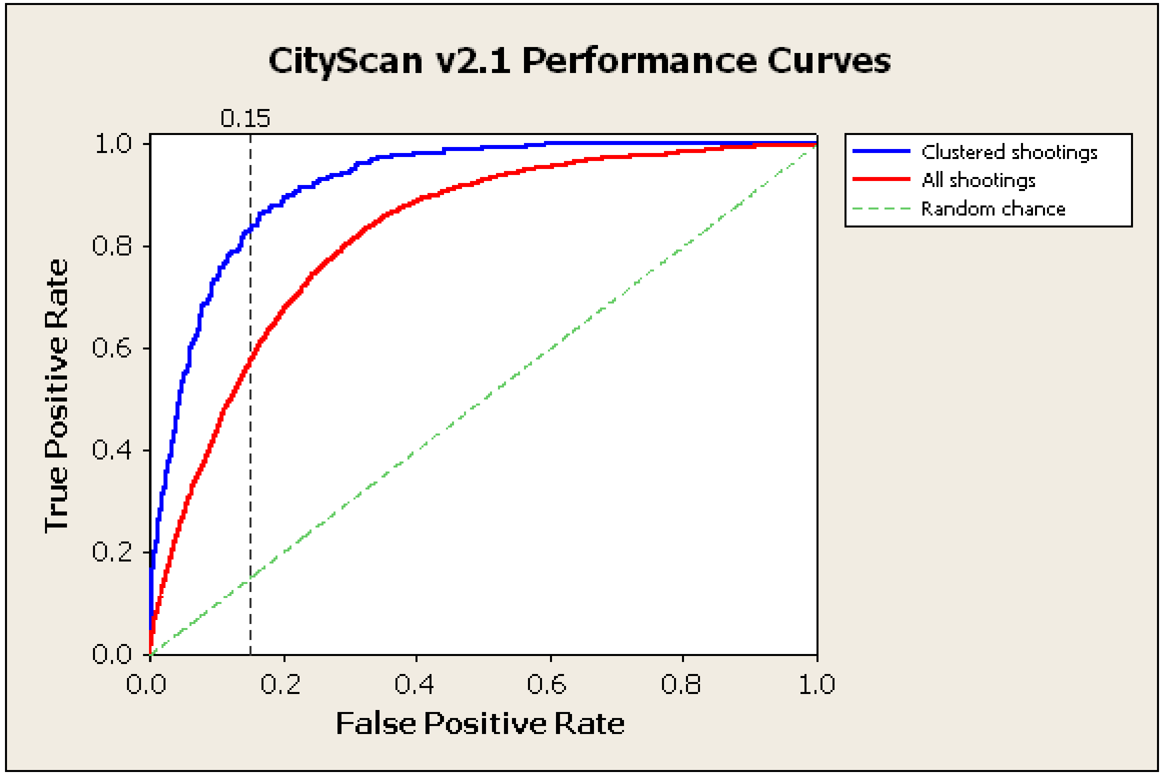
We learn a sparse (penalized) logistic regression model, with binary features including the presence of each type of leading indicator cluster within some radius.

$$\log(p / (1-p)) = \beta_0 + \beta_1 x_1 + \ldots$$

Advantages:

- Can learn which leading indicator types are most relevant for prediction
- Can include various additional features (month, day of week, weather...)

Model-based prediction results



In our preliminary evaluation on 2011-2013 data, the latest version of CityScan predicts
83% of clustered shootings/homicides and
57% of all shootings/homicides
at a **15%** false positive rate.
(Keep in mind that only 15% would be predicted by chance at this false positive rate!)

Leading Indicator Selection

Which types of calls to 911 predict clusters of homicides and aggravated battery with a handgun (shootings)?

<u>Data:</u>

- Dispatcher calls from January 2007 – May 2010, coded by 1 of 271 types (~9 million):
  - “01-01-2010”, “12:25:00”, “ARSON”, 1172456, 1834562
  - “01-02-2010”, “19:55:00”, “THEFT”, 1173123, 1831123
- All shootings/homicides from January 2007 – May 2010 (9,087 total)
  - “01-01-2010”, “19:00:37”, “HOMICIDE”, 1172001, 1834023
  - “01-07-2010”, “19:55:00”, “HOMICIDE, 1173934, 1831384

A data driven approach to this leading indicator selection problem considers the space-time interaction between pairs of point process.

Statistical Test for Space Time Interaction

If two events are close in time, they are likely to be close in space [Diggle, 1995]. Space-time clustering exists if, among those events that are close in time, there are events that are closer in space that would be expected due to chance alone.

Space time interaction tests evaluate if there is space-time clustering of events after adjusting for purely spatial and purely temporal clustering.

A widely used statistical technique for testing space time interaction is the Knox test.

Knox Test (1964)

For two point processes, $P_1 = \{(s_i^1, t_i^1): i = 1, \ldots, n_1\}$, and $P_2 = \{(s_i^2, t_i^2): i = 1, \ldots, n_2\}$, put the $N = n_1 * n_2$ pairs of points into a contingency table.

	Close in space	**Not close in space**	
Close in time	X	a	$= N_t$
Not close in time	b	c	
	$= N_s$		

Test statistic: $\frac{X}{N} - \frac{N_t}{N}\frac{N_s}{N}$

- This is essentially a χ^2 test for significance.
- The Mantel test, another commonly used test for space-time interaction, is similar to the Knox test, but requires no threshold values.

Shortcomings

Classical space-time interaction tests only consider discretized and/or linear dependence (correlation).

They focus exclusively on interpoint (Euclidean) distances.

There is no way to incorporate additional covariates and spatial or temporal structure into these tests.

Kernel Based Space-Time Interaction Test

Given points $P = \{p_i = (s_i\ ;\ t_i)\} \sim S\ X\ T$ we have two ways of measuring similarity:

- Similarity in space: $k\left(p_i\ ;\ p_j\right) = k(s_i\ ;\ s_j)$
- Similarity in time: $l\left(p_i\ ;\ p_j\right) = l(s_i\ ;\ s_j)$
- k and l are kernels: real-valued paired similarity functions where larger values indicate more similarity

Are these two notions of similarity independent?

Under widely met conditions, P(S,T) = P(S)P(T) if and only if the Hilbert Schmidt Independence Criterion is zero.

$$\mathrm{HSIC} = \frac{1}{n^2}\sum_{i,j} k(s_i, s_j)\ell(t_i, t_j) - \frac{2}{n^3}\sum_{i,j,r} k(s_i, s_j)\ell(t_i, t_r) + \frac{1}{n^4}\sum_{i,j,q,r} k(s_i, s_j)\ell(t_q, t_r)$$

Results

HSIC < .01	HSIC and Knox < .01	Knox < .01
auto accident pd	10-1	arson report
battery jo	death removal	auto theft ip
battery victim inj	evidence technician (pri. 1)	criminal tres. (ov)
beat team meeting (ov)	evidence technician (pri. 2)	evidence technician (pri. 3)
crim dam to prop rpt	gambling	found property
mental unauth absence	gang disturbance	k9 request
mission	outdoor roll call	kidnapping report
person with a gun	person shot	notify
robbery victim injured	plan 1-5	person stabbed
theft ip	shots fired	pick up car
	shots fired (ov)	polling place check
		suspicious person (ov)
		transport

Results

HSIC < .01	HSIC and Knox < .01	Knox < .01
auto accident pd	10-1	arson report
battery jo	death removal	auto theft ip
battery victim inj	evidence technician (pri. 1)	criminal tres. (ov)
beat team meeting (ov)	evidence technician (pri. 2)	evidence technician (pri. 3)
crim dam to prop rpt	gambling	found property
mental unauth absence	gang disturbance	k9 request
mission	outdoor roll call	kidnapping report
person with a gun	person shot	notify
robbery victim injured	plan 1-5	person stabbed
theft ip	shots fired	pick up car
	shots fired (ov)	polling place check
		suspicious person (ov)
		transport

Expert rating: 5 (most reasonable)

Results

HSIC < .01	HSIC and Knox < .01	Knox < .01
auto accident pd	10-1	arson report
battery jo	death removal	auto theft ip
battery victim inj	evidence technician (pri. 1)	criminal tres. (ov)
beat team meeting (ov)	evidence technician (pri. 2)	evidence technician (pri. 3)
crim dam to prop rpt	gambling	found property
mental unauth absence	gang disturbance	k9 request
mission	outdoor roll call	kidnapping report
person with a gun	person shot	notify
robbery victim injured	plan 1-5	person stabbed
theft ip	shots fired	pick up car
	shots fired (ov)	polling place check
		suspicious person (ov)
		transport

Expert rating: 5 (most reasonable), 4

Results

HSIC < .01	HSIC and Knox < .01	Knox < .01
auto accident pd	10-1	arson report
battery jo	death removal	auto theft ip
battery victim inj	evidence technician (pri. 1)	criminal tres. (ov)
beat team meeting (ov)	evidence technician (pri. 2)	evidence technician (pri. 3)
crim dam to prop rpt	gambling	found property
mental unauth absence	gang disturbance	k9 request
mission	outdoor roll call	kidnapping report
person with a gun	person shot	notify
robbery victim injured	plan 1-5	person stabbed
theft ip	shots fired	pick up car
	shots fired (ov)	polling place check
		suspicious person (ov)
		transport

Expert rating: 5 (most reasonable), 4, 3

Results

HSIC < .01	HSIC and Knox < .01	Knox < .01
auto accident pd	10-1	arson report
battery jo	death removal	auto theft ip
battery victim inj	evidence technician (pri. 1)	criminal tres. (ov)
beat team meeting (ov)	evidence technician (pri. 2)	evidence technician (pri. 3)
crim dam to prop rpt	gambling	found property
mental unauth absence	gang disturbance	k9 request
mission	outdoor roll call	kidnapping report
person with a gun	person shot	notify
robbery victim injured	plan 1-5	person stabbed
theft ip	shots fired	pick up car
	shots fired (ov)	polling place check
		suspicious person (ov)
		transport

Expert rating: 5 (most reasonable), 4, 3, 2

Results

HSIC < .01	HSIC and Knox < .01	Knox < .01
auto accident pd	10-1	arson report
battery jo	death removal	auto theft ip
battery victim inj	evidence technician (pri. 1)	criminal tres. (ov)
beat team meeting (ov)	evidence technician (pri. 2)	evidence technician (pri. 3)
crim dam to prop rpt	gambling	found property
mental unauth absence	gang disturbance	k9 request
mission	outdoor roll call	kidnapping report
person with a gun	person shot	notify
robbery victim injured	plan 1-5	person stabbed
theft ip	shots fired	pick up car
	shots fired (ov)	polling place check
		suspicious person (ov)
		transport

Expert rating: 5 (most reasonable), 4, 3, 2, 1 (least reasonable)

Conclusions & Contributions

We showed that detection of clusters of leading indicators can be used to effectively predict clusters of violent crime.

We formulated the leading indicators question as space-time interaction between pairs of point processes.

We introduced a new kernel-based space-time interaction test. This is a more flexible test than classical space-time interaction tests, since kernels can encode more than just distance between points and the HSIC can test for non-linear dependence. HSIC had comparable performance to classical tests, and the parameter choices were less critical.

From CrimeScan to CityScan…

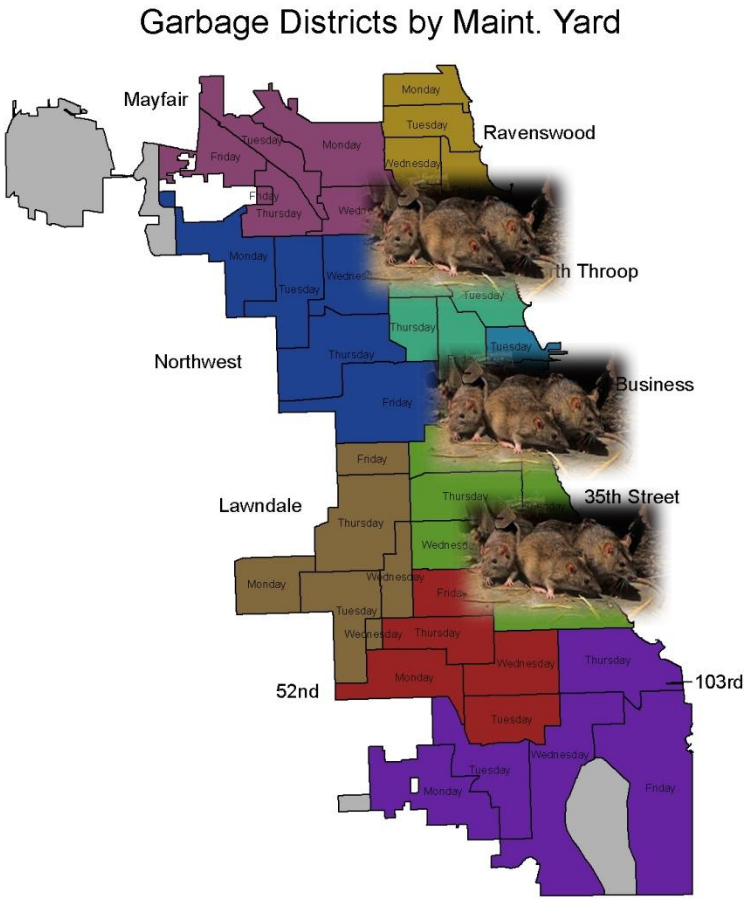
Working with the City of Chicago's Department of Innovation and Technology, we are currently using our new methods to analyze other data relevant to the city.

Most interestingly, we have very promising initial results for prediction of emerging patterns of 311 calls.

Examples: abandoned buildings, graffiti, sanitation complaints, rodent removal, garbage carts…

Our CrimeScan software has been renamed "CityScan" and will be an essential component of the city's new **Chicago SmartData platform** for real-time predictive analytics and decision making, with applications including rodent control, preventing STIs, and emergency response.

Preventing Rat Infestations with CityScan



We are currently performing a controlled experiment with Chicago's Dept. of Streets and Sanitation, with the goal of predicting and preventing rodent infestations.

- Infestations are measured by "rodent complaint" 311 calls.
- Leading indicators are other 311 call types.

<u>"Treatment" garbage districts</u>:
We predict rodent complaints using CityScan and use our predictions to direct the city's preventative rat baiting crews.

<u>"Control" garbage districts</u>:
Preventative baiting performed as usual.