Using Big Data to Identify Hotspots of Pedestrian Crashes in Manhattan

Presented by Prof. Kaan Ozbay
November 19th, 2014

Kaan Ozbay, Kun Xie and Hong Yang
Center for Urban Science + Progress (CUSP)
Department of Civil & Urban Engineering (CUE)
New York University

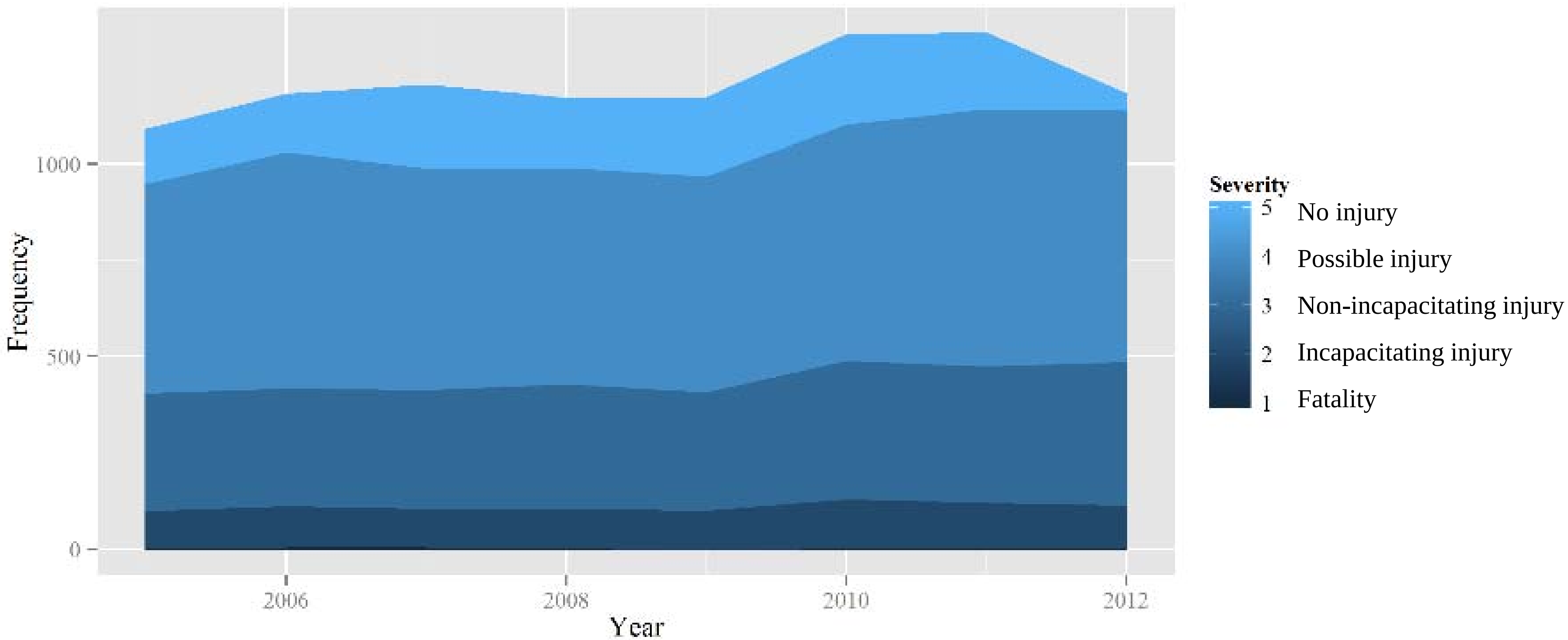
Introduction

- Pedestrian Safety Situation in Manhattan (2005~2012)
  - A total of 9664 pedestrian crashes occurred
  - About 9.4 % of them (906) involved serious injuries and fatalities
- Importance of Identifying Hotspots of Pedestrian Crashes
  - Vision Zero Action Plan was launched in 2014, aimed at reducing the crash rate and relieving crash severity
  - Accurate identification of these hotspots can result in efficient allocation of government resources
- Two Important Factors in Hotspot Identification:
  - a) Different costs of crashes by severity
  - b) Effects of crash exposures such as traffic volume, road length, etc.

NYU

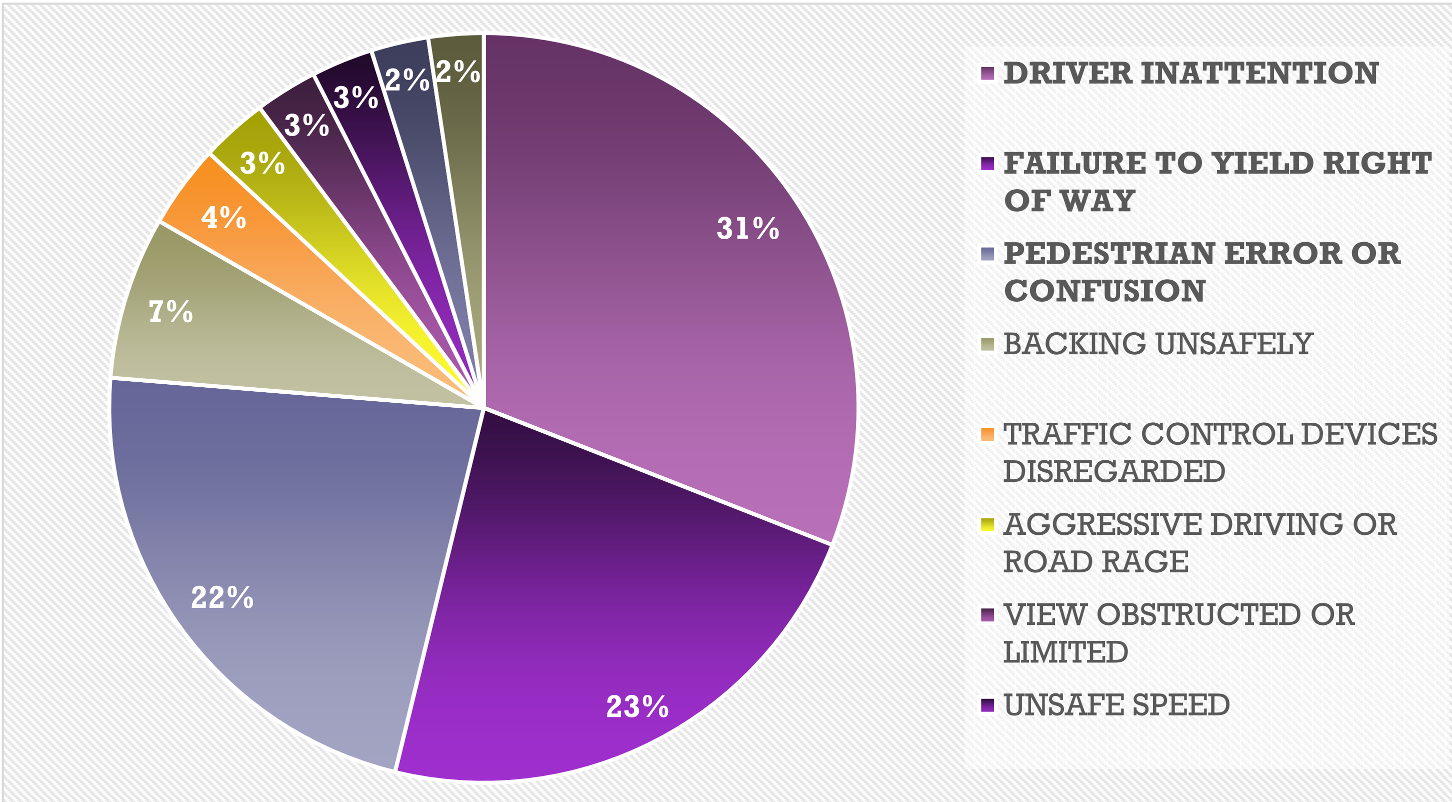
Descriptive Analysis

- Pedestrian Crash Frequency by Severity (2005~2012)



Descriptive Analysis

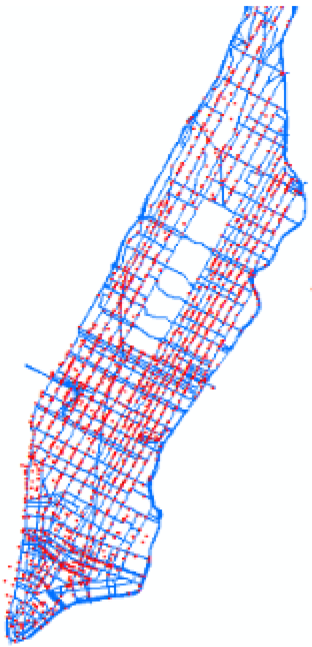
- Pedestrian Crash Causes



NYU

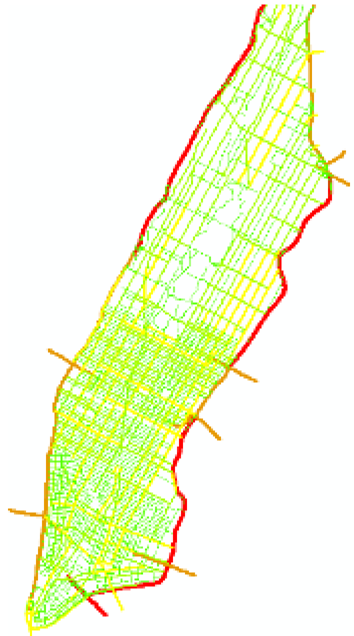
“Big Data” Used

- A massive amount of data from a variety of sources were collected. The total size of datasets is over 100 GB.



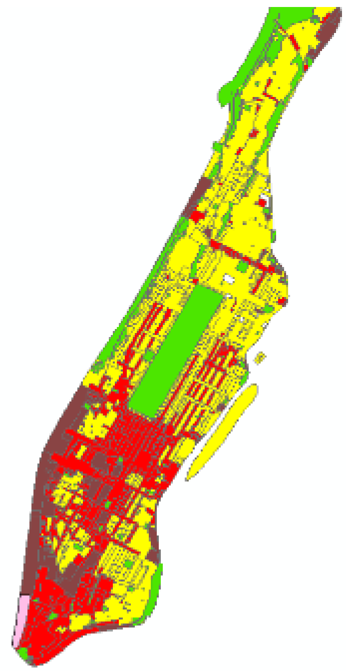
Crash
- No injury
- Possible injury
- Fatality
- …

(Source: NYSDOT)



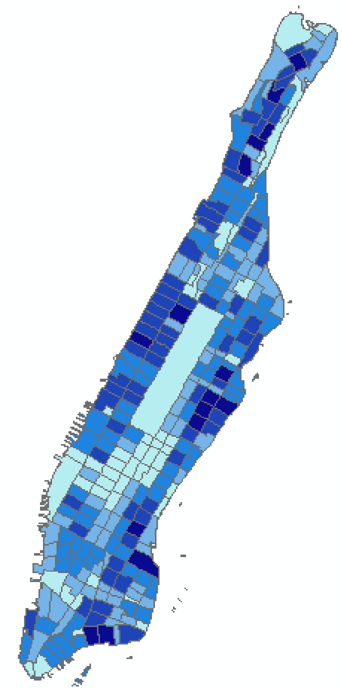
Traffic
- Traffic volume
- Taxi trips
- MTA turnstile
- …

(Source: NYSDOT, TCL, MTA)



Land Use
- Source:
- Residential
- Commercial
- …

(Source: NYCDCP)



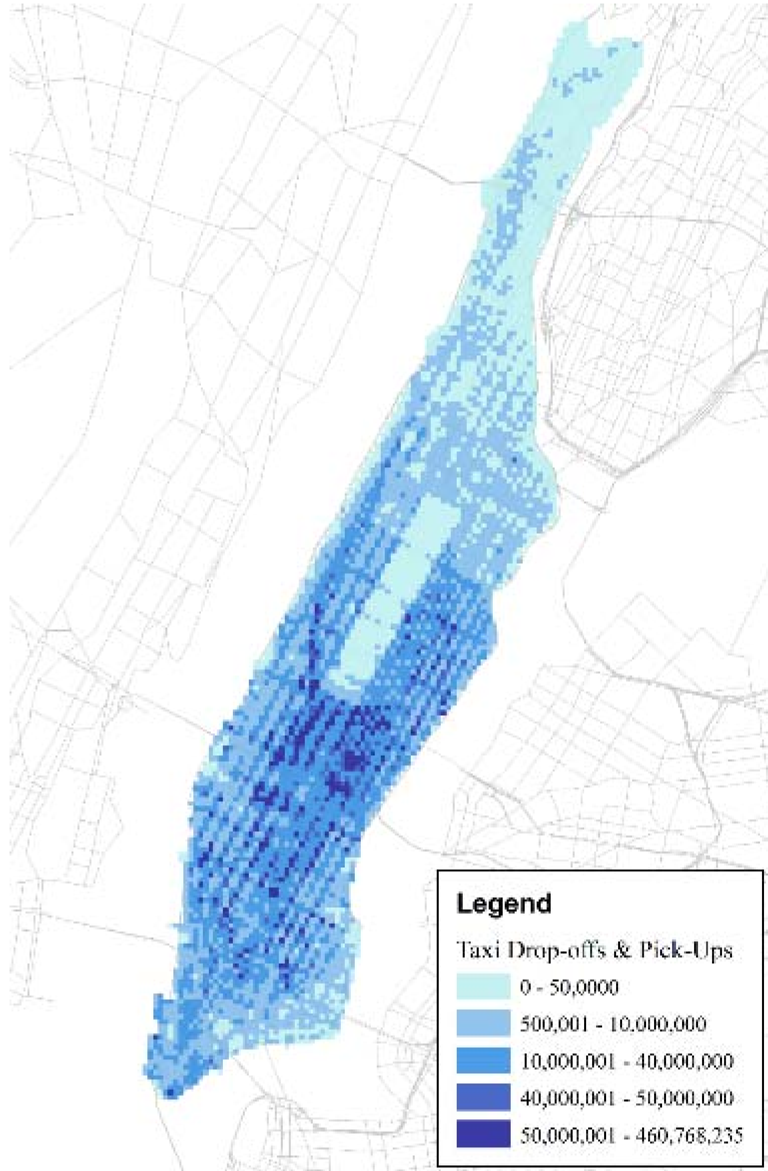
Socioeconomic
- Population
- Employment
- GDP
- …

(Source: US Census Bureau)

+ "Big Data" Used: Taxi Trip Data

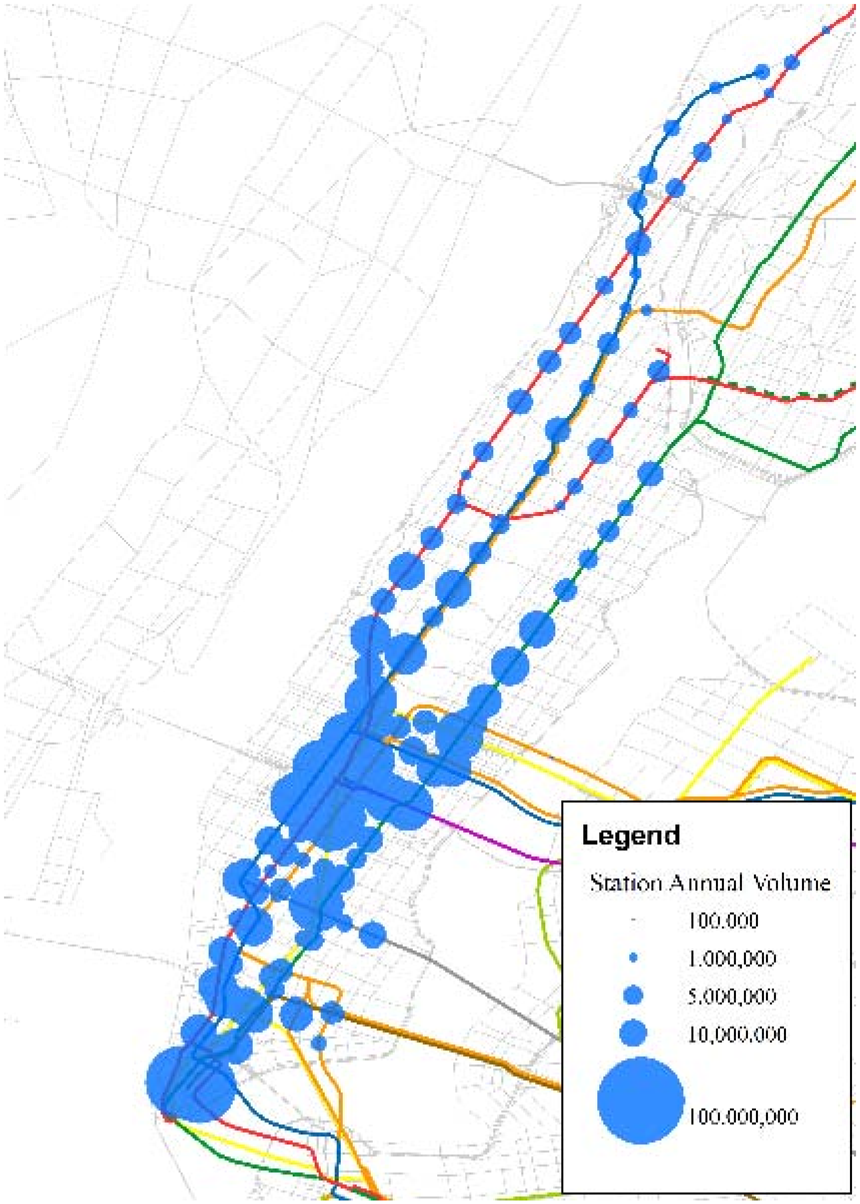
- Taxi pick-up and drop-off data from 2008 to 2012. Size of dataset is over 100 GB

- Taxi trips concentrate on main corridors such as 5 Ave and 6 Ave.



NYU

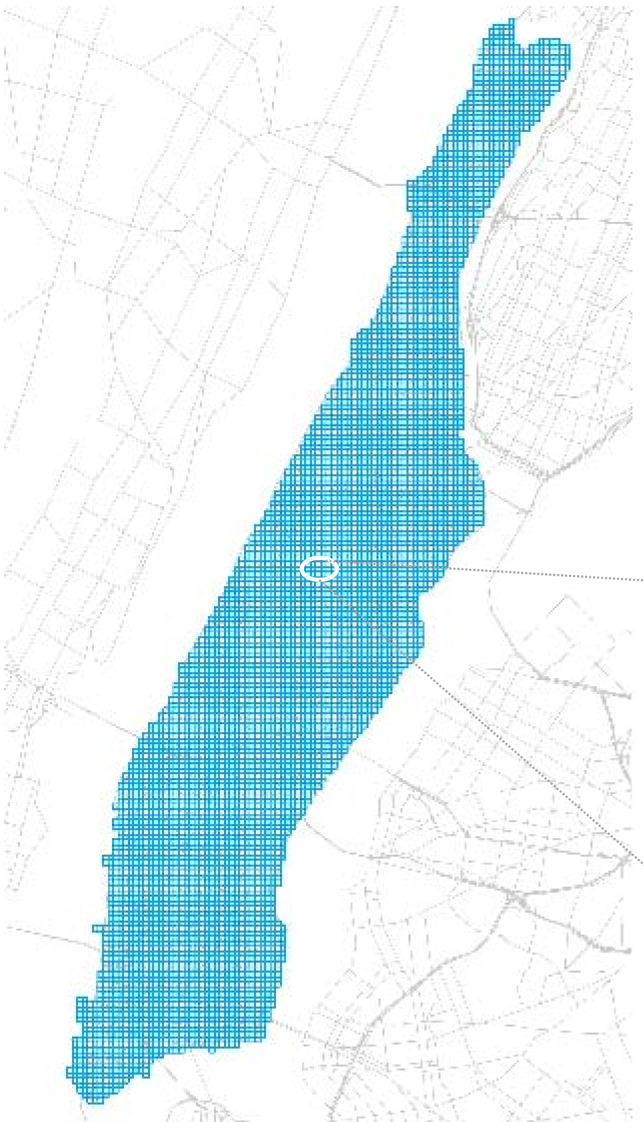
"Big Data" Used: MTA Turnstile Data

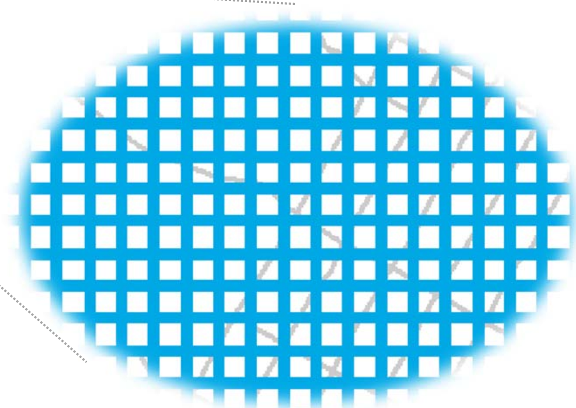


- Refreshed weekly, available up until May 05, 2010
- Midtown and downtown have large passenger volumes

Grid Cells

- Basic geographical unites of analysis: grid cells (300×300 feet^2)
- Traffic, land use, demographic and socioeconomic features were captured for each cell





Grid Cell (300×300 feet^2)

Spread of Crash Cost

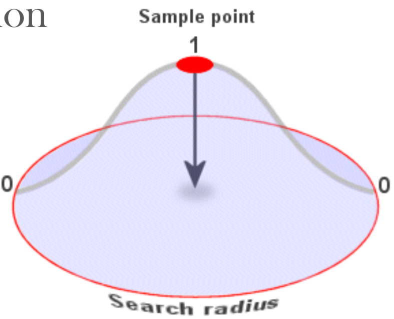
- Crash Cost by Severity

Crash Type	Comprehensive Cost per Crash ($)
Fatality	4,538,000
Incapacitating injury	230,000
Non-incapacitating injury	58,700
Possible injury	28,000
Property damage only	2,500

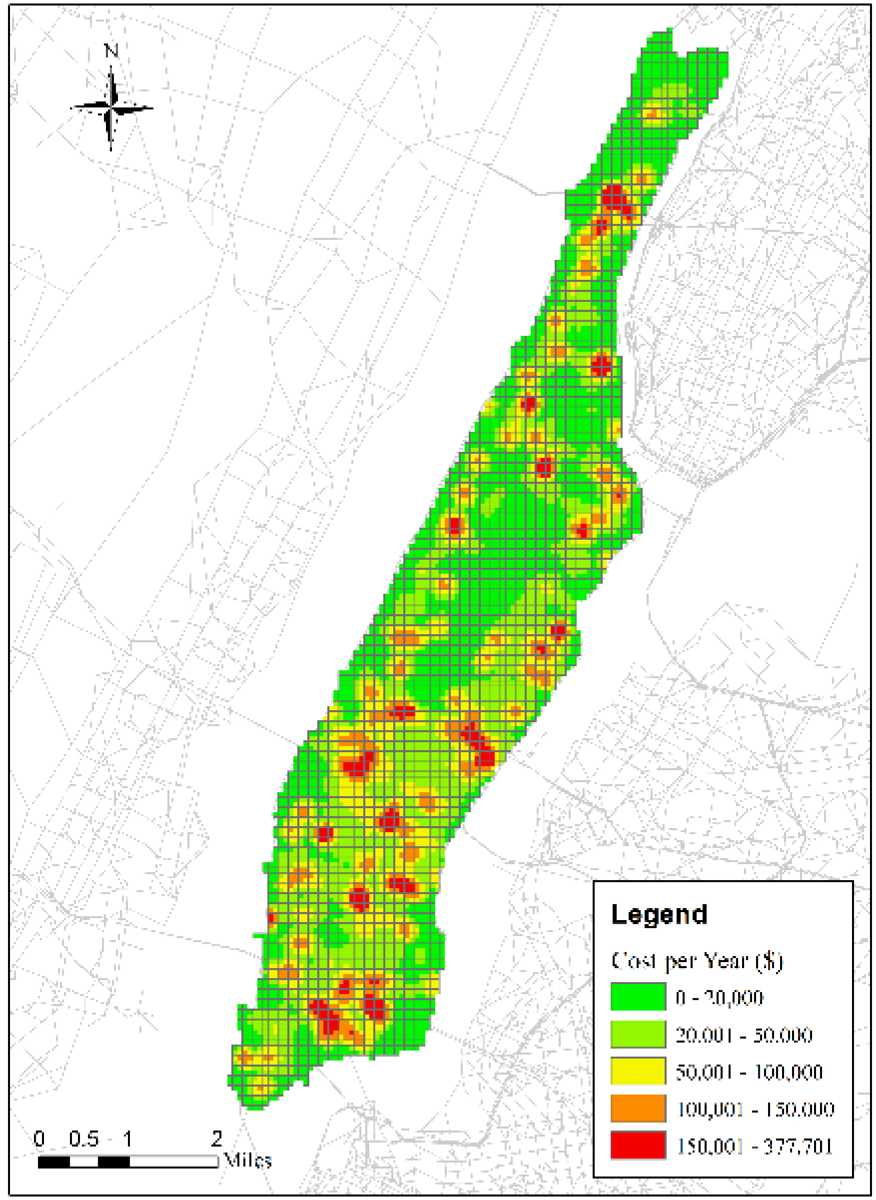
(Source: National Safety Council. All values were converted to 2012 dollars)

- 2-D Kernel Density Function

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right)$$

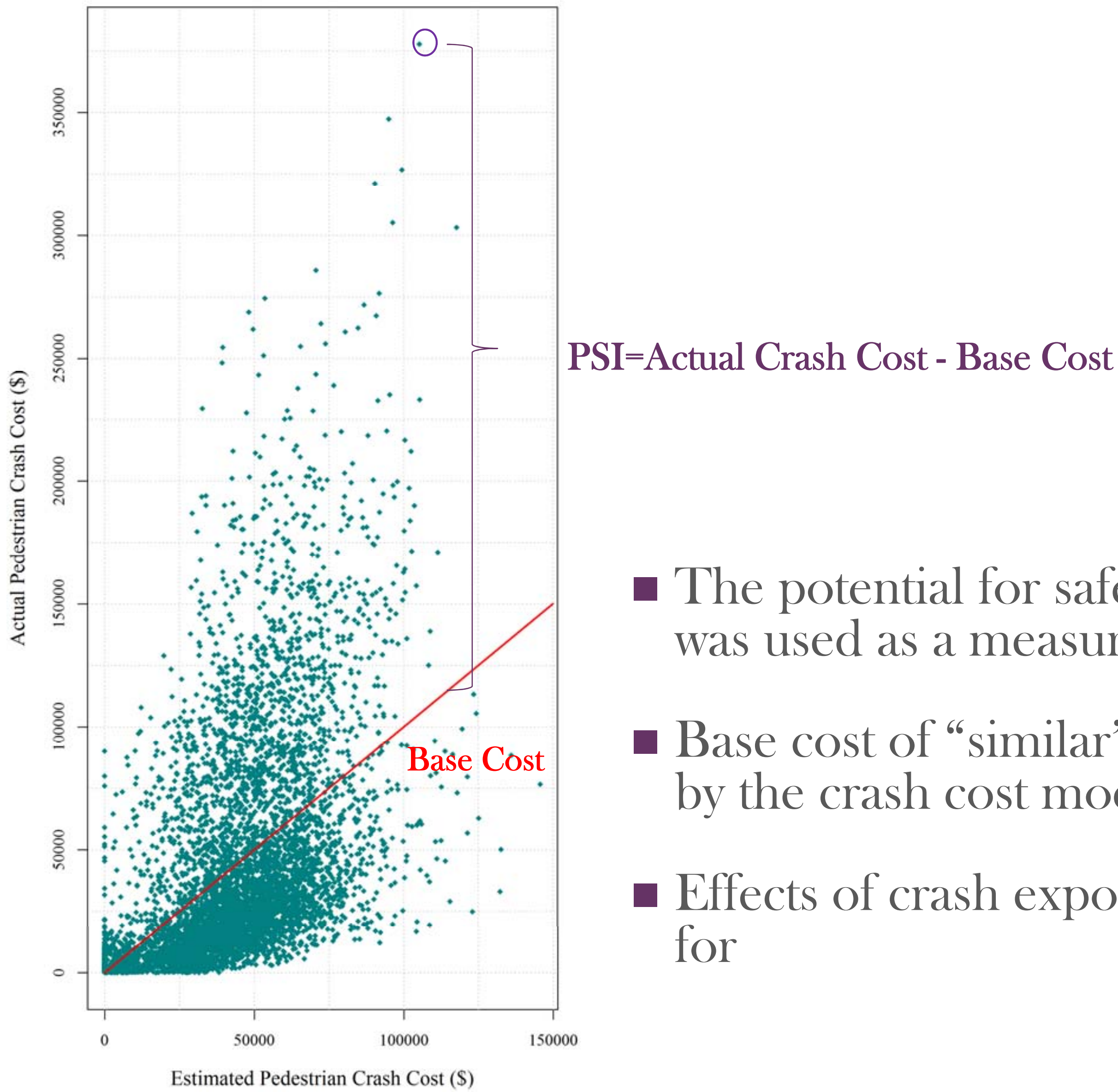


$\lambda(s)$: Density at location s
r: Bandwidth (1000 feet is used here)
d_{is} : Distance from location s to crash i
k(.): kernel function (Gaussian function is used here)



NYU

Potential for Safety Improvement (PSI)



- The potential for safety improvement (PSI) was used as a measure to rank crash hotspots
- Base cost of "similar" sites can be estimated by the crash cost model
- Effects of crash exposures can be accounted for

Crash Cost Model

- Linear Model
  - Develop a linear relationship between dependent variable crash cost and independent variables such as *taxi trips*, *truck ratio*, *population*, etc.

$$y_i = \beta x_i + \mu_i,\ \mu_i \sim N(0,\sigma^2)$$

y_i : Pedestrian crash cost per year ($)
x_i : Independent variables
β : Coefficients of x_i
μ_i : Error term

  - Weakness of linear model
    - Ignore the fact that crash cost is left-censored at zero.
    - Have the chance to give a negative prediction of the crash cost

NYU

Crash Cost Model

- Tobit Model
  - Appropriate for describing relationship between a non-negative dependent variables (crash cost) and independent variables.

$$y_i = \begin{cases} y_i^* & if\ y_i^* > 0 \\ 0 & if\ y_i^* \le 0 \end{cases}$$

$$y_i^* = \beta x_i + \mu_i,\ \mu_i \sim N(0, \sigma^2)$$

y_i : Pedestrian crash cost per year ($)
y_i^* :Latent variables ($)
x_i : Independent variables
β : Coefficients of x_i
μ_i : Error term

NYU

Modeling Results

- Model Comparison: Tobit model vs Linear Model

	Log-likelihood	AIC	BIC
Linear model	-74373.18	148774.4	148868.60
Tobit model	-72883.64	145795.3	145889.50

  - The Tobit model outperforms the linear model by presenting higher log-likelihood and lower AIC and BIC.

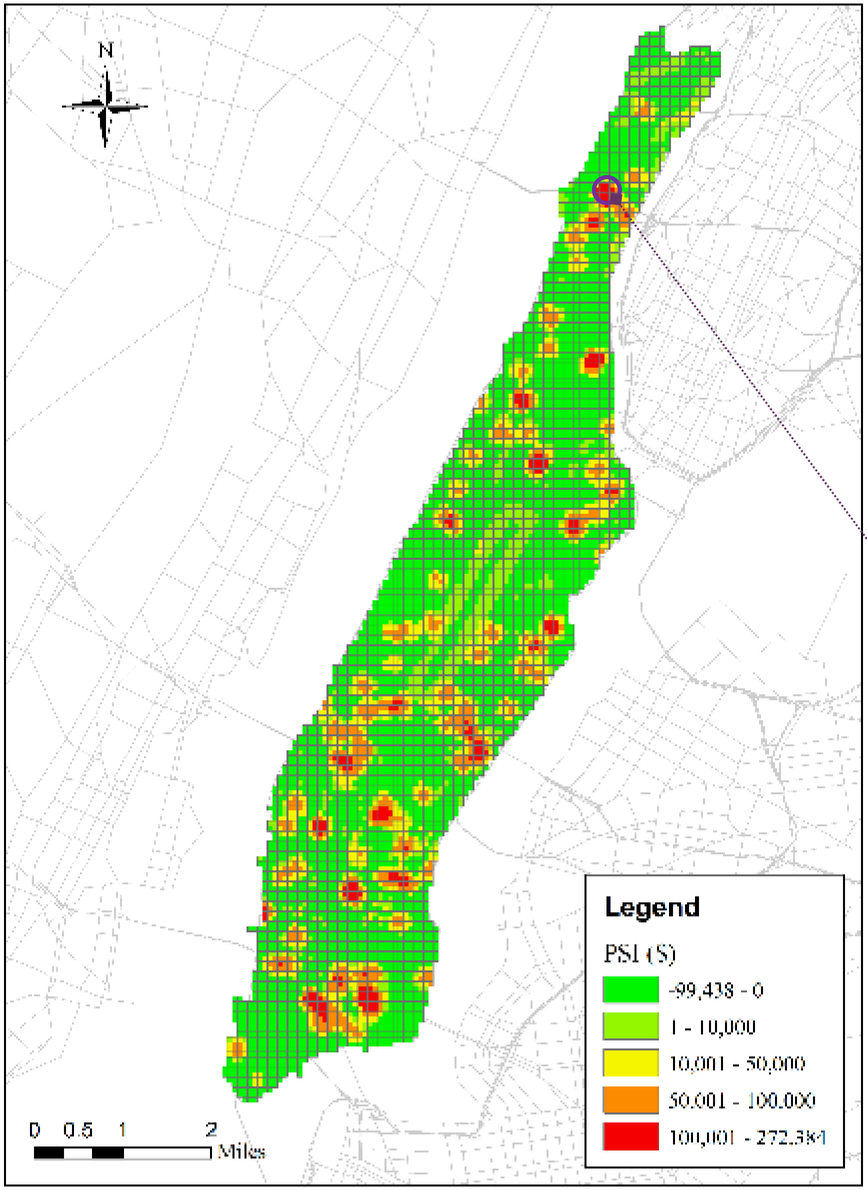
- Results of the Tobit Model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-1.34E+04	1.73E+03	-7.745	9.55e-15 *
Vehicle mile traveled	7.99E-04	3.65E-04	2.189	0.028603 *
Taxi trips (10^3)	3.15E+01	3.45E-00	3.027	0.002471 *
Subway passengers (10^3)	1.77E+01	1.62E-01	10.942	< 2e-16 *
Truck ratio	9.64E+04	[illegible]	[illegible]	< 2e-16 *
Bus stop density	[illegible]	[illegible]	17.052	< 2e-16 *
Length of sidewalks	[illegible]	[illegible]	[illegible]	4.48e-05 *
Total population	[illegible]	[illegible]	[illegible]	2.66e-14 *
Ratio of population over 6[illegible]	[illegible]	[illegible]	2.634	0.008432 *
Unemployment	[illegible]	[illegible]	5.094	3.51e-07 *
Ratio of commercial areas	1.44E+04	2.84E+03	5.051	4.39e-07 *
Ratio of residential areas	7.97E+03	2.34E+03	3.403	0.000667 *
Ratio of manufactural areas	8.37E+03	3.09E+03	2.708	0.006764 *

One unit increase is
One unit increase is
expected to increase the annual crash cost by 17.7 $

*Indicate variables which are statistically significant

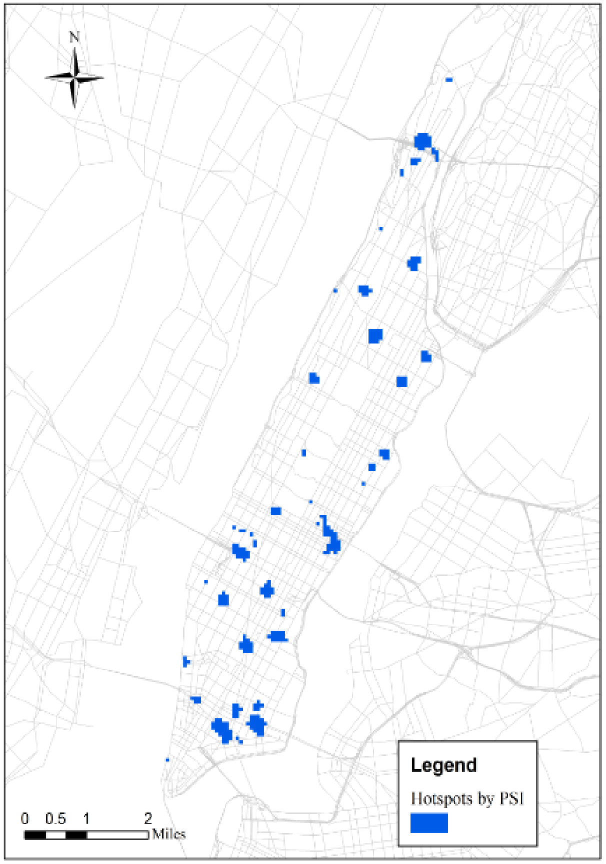
Hotspot Identification



Spot with the greatest improvement potential: Broadway (from 180th to 181st ST)

272,384 $ can be saved annually from this spot!

Comparisons of Hotspots Identified



- Identify top 300 hotspots: by crash frequency vs by PSI
- Only 40 hotspots (about 13.3%) are overlapped
- Hotspots identified by PSI tend to be on continuous regions

NYU

Thank You!

Kaan Ozbay, Kun Xie and Hong Yang
Center for Urban Science + Progress (CUSP)
Department of Civil & Urban Engineering (CUE)
New York University