Benjamin Rothschild
MACSS Thesis Proposal Workshop
May 6, 2017

**Research Question**
How accurately can I create a model to predict the locations of homicides in Chicago?

**Introduction**
Many US cities have learned that to operate more efficiently they can use the data they collect to learn how to allocate their resources better. These predictive models have already helped in a wide array of city applications from distributing health inspectors to food service establishments, planning transportation routes to new areas, or to allocating police to crime hotspots. In my research, I aim to show that by adding additional data from third parties such as Zillow, Twitter, or Yelp, I can increase the predictive accuracy of crime hotspot detection in Chicago.

**Literature Review**
Research on crime and policing spans academia, industry, and government. In this literature review I will briefly cover the academic work as well as give an overview of the progress some private companies have made solving similar research questions. Lastly, I will cover some methods that states and cities currently use to optimized resource allocation in different departments.

Research on crime spans several academic departments such as economics, sociology, and psychology and has been a topic of research for over 150 years. In 1835 Adolphe Quetelet, a Belgian sociologist and mathematician, analyzed crime statistics in France which were first tracked and published in 1827.[1] In the economics literature, the study of crime started with Becker's model of rationality and criminal activity. Becker's model states that crime comes from "the choices that reasoning individuals make between criminal and non-criminal courses of action, choices informed by predictions of the likely merits and demerits of available alternatives". Further study using this model has been done around deterrence methods such as how capital punishment has decreased homicides in the United States (Ehlrich and Bower 1987) and how social interactions are inversely related to the degree of crime (Glaeser, Sacerdote, Scheinkman).[2]

[1] Using mathematical models, Quetelet found that some people were more likely commit crimes than others such as younger, poorer, and unemployed.
[2] Glaeser, Sacerdote, Scheinkman refuted Becker's individual agent-based model and instead argued that social interactions create enough variability between individuals to explain the high variance of crime rates. In this model, the amount of social interaction is inversely related to the degree of the crime which means that social interactions are higher in minor crimes and lower in more serious crimes like homicide.

In the psychology literature, the seminal work by Gottfredson and Hirschi (1990) argues that criminals are not rational agents and instead believe the essential element in determining if someone is a criminal is the absence of self-control. People with self-control will consider the consequences of their behavior while those without do not. They also state that self-control is a learned behavior at a young age and is very resistant to change. In their book, they use this framework to explain why certain criminals decide to commit a crime, specifically looking at white-collar crime, homicide, and gang violence.

The rise of Big Data has produced a new wave of research on the study of urban life. Whereas previously research relied on highly aggregated statistics and surveys, new data sources such as sensors and digitized records have produced rich datasets that can be used to analyze classic questions in new ways. For example, some researchers were successfully able to predict the median income of residents in New York City by training a computer vision model on Google Street View images (Glaeser et all 2016). Some other data sources that have successfully been used in urban science research include data from Zillow and Craigslist for home prices, Yelp and Foursquare to predict foot traffic in restaurants and stores. In addition to datasets produced by private companies, the Open Data movement has also been invaluable to the study of cities. Many cities and have started to publically release datasets online to be used for analysis by businesses and researchers. By combining these data sources with city records new areas of research is possible.

One example of this phenomena is that researchers used text analysis on Yelp reviews to predict if a restaurant had a health violation. They matched 1,756 restaurant inspections with reviews on Yelp from 2006 to 2013 and used a SVM model with 10-fold cross validation to predict what restaurants were more likely to have food violations. They predicted 82% of severe offenders from places with no violation (Kang, et al. 20130). In Chicago, data scientists in the city government created a prediction model to more efficiently distribute food inspection officials to food establishments across the city. There are over 15,000 food establishments that are subject to sanitation inspections and only three dozen food inspectors. To better allocate these inspectors to restaurants, data scientists collected data on complaints and matched them to the food establishments business characteristics. This model found that food inspection outcomes were correlated with many factors such as sanitation complaints, nearby burglaries, length of time since the last inspection, and if the establishment had previous critical or serious violations. By using this model to allocate their food inspectors, the city improved the time to discovery of critical violations by 7 days.[3] Similar to these approaches, my research aims to help cities better allocate their resources by using novel "Big Data" data sources.

In the research area of crime detection and predictive policing, the biggest study to date looked at how effective predictive models were at lowering crimes in Los Angeles and Kent, UK (Courneya et al 2003). Their hypothesis was that since the concentration of police resources in stable crime hotspots has been effective in reducing crime, if it was possible more accurately detect these hotspots to allocate policing resources crime would also decrease. They first

[3] The researchers published their methods and code here: https://chicago.github.io/food-inspections-evaluation/

created a model to predict crime hotspots. In Los Angeles they focused on burglary, car theft, and criminal damage which comprise about 55% of crime volume. They used an epidemic-type aftershock sequence (ETAS) model that estimates the risk associated with both long-term hotspots and short-term models of near-repeat risk. From the results of this model they then conducted two controlled experiments with a division of the Los Angeles Police department and Kent Police Department. In the experiments the ETAS algorithm was put head-to-head with hotspot maps procures by a dedicated crime analyst. Compared to a dedicated crime analyst the predictive model could predict crime successfully at a rate of 9.8% and 6.8% compared to 6.8% and 4.0%. The next step in their research was to use this predictive model to control where police go to measure if the improved policing coverage could lower crime rates. They conducted a controlled experiment and found that on average when using the ETAS forecasts there was a 7.4% reduction in crime volume as a function of patrol time, whereas patrols based upon analyst predictions showed no significant effect. The researchers conclude that dynamic police patrol in response to ETAS crime forecasts can disrupt opportunities and lead to real crime reductions.

There is also a small body of research on improving the accuracy of hotspot maps. Since crime hotspot maps are a widely-used method of displaying spatial crime patterns and allocating police resources, researchers have developed a methodology to better predict crime hotspots. They argue that current hotspot maps often utilize too little of available data and fail to capture short-term changes in risk. They improve on these methods by using an Expectation Maximization algorithm which can easily be deployed on a desktop computer connected to a police agency Risk Management System (Mohler 2014). The researchers applied this methodology to homicide and gun crime in Chicago and make their results and methodology available online.

In addition to academic research there have been several private companies that offer software solutions to police departments and thus contributed to this research topic. Often these companies will describe their methods online or publish to academic journals. One example is CivicScape, a company founded by University of Chicago professors and alumni. They leverage information from historical crime-trends, on-the-ground intelligence from police officers, and input from the community and provide a deployable solution to police departments. The input datasets they use include recent crime activity, 311 calls (community input), census tract data, weather forecasts. They use an ensemble of feed-forward neural networks tuned to the specific crime type and location in the city and produce predictions on a three-block radius for every hour. They have open sourced their methodology and data sources on Github.[4] Another company PredPol provides similar predictive models to departments in the United States. They use three data points, crime type, crime location, and crime date to provide agencies with custom crime predictions, usually pinpointing areas within a 500 foot by 500 foot box. They update their models every 6 months and are aimed at supporting dedicated crime analysts in making resource allocation and patrol decisions. They have published their results in several academic papers (Mohler 2013).

[4] *https://github.com/CivicScape/CivicScape*

**Data**
I have identified a few data sources to use my paper:

Plenar.io (http://plenar.io). This site compiles many of the datasets published on the City of Chicago Data Portal such as crime statistics, 311 Calls, Food inspections, etc and make them available online. They compile all the datasets onto the same "timeline" and "map". This will hopefully save me some time in cleaning data.
Datasets I have identified on this website to use are:

- 311 Service Requests
- Business Licenses
- Environmental Inspections
- 311 Vacant Buildings
- Crimes
- Food Inspections
- Red Light Tickets
- Divvy Trips
- Building Violations

Craigslist. I have thought about trying to use craigslist to determine what people sell at certain locations, maybe this will influence crime rates. Not sure if data is public/available as I would need historical data.

Twitter. For geo-tagged tweets in Chicago, can I figure out a sentiment score and add that to my model? Need to look into this more.

**Methodology**
In my analysis I plan to use a Special Regression with endogenous regressors. To measure the predictive power of my model I will consider the Error Rate. I will try to tune my model to have the highest predictive accuracy (possibly at the expense of clear explanation). I will do the analysis in python using the PySAL and for mapping I will use MapBox. I will consider different prediction timespans (hourly, daily, etc) and consider tradeoffs between predictive power and usefulness.

**References**
Becker, Gary S. "Crime and punishment: An economic approach." *The Economic Dimensions of Crime*. Palgrave Macmillan UK, 1968. 13-68.

Courneya, Kerry S., et al. "*Randomized controlled trial of exercise training in postmenopausal breast cancer survivors: cardiopulmonary and quality of life outcomes." Journal of clinical oncology 21.9 (2003): 1660-1668.*

Ehrlich, Isaac, and George D. Brower. "On the issue of causality in the economic model of crime and law enforcement: Some theoretical considerations and experimental evidence." *The American Economic Review*77.2 (1987): 99-106.

Glaeser, Edward L., et al. "Big data and big cities: The promises and limitations of improved measures of urban life." *Economic Inquiry* (2016).

Glaeser, Edward L., Bruce Sacerdote, and Jose A. Scheinkman. "Crime and social interactions." *The Quarterly Journal of Economics* 111.2 (1996): 507-548.

Gottfredson, Michael R., and Travis Hirschi. *A general theory of crime*. Stanford University Press, 1990.

Kang, Jun Seok, et al. "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." *EMNLP*. 2013.
*Mohler, George. "Marked point process hotspot maps for homicide and gun crime prediction in Chicago." International Journal of Forecasting 30.3 (2014): 491-497.*

Mohler, George O., et al. "Self-exciting point process modeling of crime." *Journal of the American Statistical Association* 106.493 (2011): 100-108.

Quetelet, Adolphe. *Sur l'homme et le développement de ses facultés ou essai de physique sociale*. Vol. 1. Bachelier, 1835.