

# CSI Chicago Project Proposal

Let's make less crimes happen

Friedrich Amouzou

University of Colorado Boulder  
Boulder, Colorado  
Friedrich.Amouzou@colorado.edu

Maria Pazos\*

University of Colorado Boulder  
Boulder, Colorado  
Maria.Pazos@colorado.edu

Maria Knigge

University of Colorado Boulder  
Boulder, Colorado  
Maria.Knigge@colorado.edu

Brandon Stone

University of Colorado Boulder  
Boulder, Colorado  
Brandon.n.Stone@colorado.edu

## ABSTRACT

The city of Chicago, Illinois has a significantly higher crime rate than the rest of the United States [3]. Researchers have investigated crime in Chicago for decades but the majority of studies exclusively focus on violent crime [4]. The purpose of this project is to find interesting relationships between the various types of crime that occur in Chicago, the date they occur, and their location. By data mining the Chicago Police Department's CLEAR database of reported instances of crime we could potentially find unintuitive information that can aid in the process of crime prevention and reduction. This project intends to provide a more holistic view of the geo-temporal relationships of crime in Chicago.

## CCS CONCEPTS

• **Information systems** → **Data analytics**; Data mining; • **Mathematics of computing** → *Exploratory data analysis*;

---

\*goes by Sol

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).  
*FakeConference, March 2018, Boulder, Colorado USA*  
© 2018 Copyright held by the owner/author(s).  
ACM ISBN 123-4567-89-012/34/56.  
[https://doi.org/12.345/123\\_4](https://doi.org/12.345/123_4)

## KEYWORDS

Chicago, data mining, crime, research, CLEAR

### ACM Reference Format:

Friedrich Amouzou, Maria Knigge, Maria Pazos, and Brandon Stone. 2018. CSI Chicago Project Proposal: Let's make less crimes happen. In *Proceedings of ACM Never Happened (FakeConference)*. ACM, New York, NY, USA, 4 pages. [https://doi.org/12.345/123\\_4](https://doi.org/12.345/123_4)

## 1 PROBLEM STATEMENT/MOTIVATION

Prior research on crime in Chicago has been done but the studies tend to be longterm or specific. Previous studies focused exclusively on gun violence [4]; or were completely comprehensive, but took 48 years for data to be gathered and analyzed [3]. We want to analyze available data on crime cross referenced with its geographical position to get a robust source to anticipate crime. We want to cross reference all crimes with their coordinate location to get a comprehensive view of all crime throughout the city; and receive results sooner than 48 years.

## 2 LITERATURE SURVEY

Multiple universities have conducted studies on crime in Chicago in the past. These studies tend to take a long time to gather data or are specifically focused on a certain aspect of crime. Three major studies stood out when looking for previous work done in this field.

## 2.1 University of Chicago

The University of Chicago does ongoing research with their Crime Lab which studies different social programs and policies to reduce crime and violence [1]. The Crime Lab is part of the University's Urban labs and was founded in 2008. The lab focuses on the cause of crime and developing and evaluating new ways to combat crime.

## 2.2 University of Wisconsin - Madison

Assistant professor Robert Vargas is focusing on the relations between gangs and student achievement [2]. His work is centered around the fact that, "You can build great schools and great community programs, but if these kids don't feel safe walking around in their own neighborhood, it's kind of a moot point." [2]

## 2.3 Yale University

The study conducted by Andrew Papachristos, Ph.D aims to be a comprehensive look at the trend of crime in Chicago. The study encompasses many attributes of crime around the city and took 48 years to gather and analyze data. Trends in crime over time were heavily analyzed in this study however not a lot of attention was given to the relation between geographical location and crime [3].

## 3 DATA SET

The principal dataset, found here:

<https://catalog.data.gov/dataset/crimes-2001-to-present-398a4>

is the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) database. It's a relational table containing over six million reported instances of crime in Chicago starting from January 1st, 2001 to a week before the present. Each instance of crime has twenty-two attributes associated with it:

### 3.1 Crime Identification

Two of the attributes, *ID* and *Case Number* are exclusively used to identify the particular instance of crime. There's also the *Updated On* attribute which indicates when the crime was last modified in the database. These

attributes, along with the geo-temporal attributes identify which individual incidents are part of the same overall crime.

## 3.2 Types of Crime

Nearly a third of the data's attributes describe the nature of the crime. This includes nominal data like the literal descriptions of the crime in the *Primary Type* and *Description* attributes which categorize the crime and its severity in written word. Additionally, there are the *IUCR* and *FBI Code* attributes which provide the Illinois Uniform Crime Reporting and Federal Bureau of Investigation codes respectively: standardized government categories for the crimes. There are also two binary attributes *Arrest* and *Domestic* which indicate whether an arrest was made and whether the crime is classified as domestic.

## 3.3 Geo-temporal

The majority of the attributes indicate when and where the crime occurred. The numeric interval attributes *Date* and *Year* describe the time the crime occurred down to the nearest second<sup>1</sup>. The *X Coordinate*, *Y Coordinate*, *Latitude*, *Longitude*, and *Location* (which is a combination of the two former) describe the precise location of where the crime occurred. These attributes will be especially useful for mapping out our results. The rest of these attributes are nominal descriptions and categorizations of the location: *Block*, *Beat*, *District*, *Ward*, *Community Area*, *Location Description*. All together the geo-temporal attributes of this dataset provide a in-depth account of where these crimes transpired.

## 4 PROPOSED WORK

### 4.1 Preprocessing

We will be removing any incomplete police reports (i.e., we will remove data containing NULL values). Additionally, because the dataset contains over six million entries, we will need to reduce the dataset by eliminating uninformative and unnecessary features. Rows that include missing data and/or did not end in any reprimand will be ignored. This will give us a rich view at the many aspects of crime while also eliminating false positives; false reports, incomplete reports, etc. This

<sup>1</sup>Although this is seemingly only accurate to the nearest hour

will allow us to process the data with greater speed and get clearer results.

## 4.2 Data Analysis

Depending on the type of feature (e.g., latitude and longitude of the crime versus the block on which it occurred), we will use different methods of prediction. Multilinear regression techniques will be implemented to predict numeric and non-categorical data, like the latitude or longitude of a crime. We will use Logit and Probit Regression in order to classify data non-numeric data, where Logit Regression will be used to predict binary-valued features and Probit Regression will be applied to features that can be well-modeled by a Normal distribution. K Nearest Neighbor (KNN) methods of classification may also be applied, depending on the success of aforementioned techniques. Clustering<sup>2</sup> will allow us to perform more open-ended classification, which will potentially allow us to reintroduce police reports with missing locational or criminal information. Our work differs from previous work in the field because we are aiming for a comprehensive view of crime based on geographical location and time of day. We are gathering data to analyze chance of crime on any given day in an area rather than looking at the trend of crime over time.

## 5 EVALUATION METHODS

We will evaluate the efficacy of our method to predict where crimes occur by verifying that test data will generate an accurate prediction of a crime's geographic location 90-95% of the time.

## 6 TOOLS

We will be using Python, Jupyter Notebooks and R for this project, relying heavily on Python as our main tool. We will utilize Python's libraries, including numpy, pandas, scipy.stats, etcetera.

## 7 MILESTONES

- We expect the data processing to be the most difficult and time consuming part of this project. This portion of the project is to be completed by March 20th.

<sup>2</sup>Time permitting

- By April 3rd we should have completed and debugged our code, and we should be producing results and graphs, so that the only work left will be completing the final paper and the presentation.
- The rough draft of the final paper should be completed by April 19th.
- We will also contact other groups with similar projects by April 24th to see if we can share any relevant data. If there is any relevant data between our groups, we will include it in a "Similar Projects" section of our final paper and our final presentation.
- The final draft of our paper and presentation will be completed by April 26th, in order to have enough time for last minute corrections and such before the due date, May 1st.

## 8 SUMMARY OF PEER REVIEW SESSION

There are two other groups in our data mining class doing a project on crime in Chicago. Group 3, Chicago Crimes: Substance Abuse and Violence, is cross-referencing their dataset with Chicago laws and policies to see how they affect the geo-temporal distribution of crime. We differ from Group 6 in that we are looking for general predictive trends rather correlations to laws. Group 6, Make Chicago Safe Again, is focused on finding interesting trends and outliers and not necessarily mapping crime to regions. We differ from Group 3 in that we are visually mapping our results and ignoring outliers. Another group in our data mining class is focusing on crime statistics as well, albeit in Los Angeles rather than Chicago. We will corroborate our final results with these four groups in order to discover and confirm interesting conclusions we derive.

## ACKNOWLEDGMENTS

The authors would like to thank their mothers and our wonderful professor, Elle Boese.

## REFERENCES

- [1] [n. d.]. Urban Labs Crime Lab. ([n. d.]). <https://urbanlabs.uchicago.edu/labs/crime>
- [2] Jim Dayton. 2015. Professor studies impact of Chicago gang violence. (Feb 2015). <https://news.wisc.edu/professor-studies-impact-of-chicago-gang-violence/>

- [3] Andrew V Papachristos. 1970. 48 Years of Crime in Chicago: A Descriptive Analysis of Serious Crime Trends from 1965 to 2013. (Jan 1970). <https://isps.yale.edu/research/publications/isps13-023>
- [4] Mark Peters. 2017. UChicago Crime Lab releases 2016 gun violence report. (Jan 2017). <https://news.uchicago.edu/article/2017/01/19/uchicago-crime-lab-releases-2016-gun-violence-report>