Nathan Cooper Jones
James Guerrera-Sapone
Bhoj Rani Soopal

Project Proposal

Research goal: The goal of this project is to use both historical and recent crime data in the Illinois Institute of Technology campus for **understanding relationships present in the data** (i.e. how block, time, date, weather, and crime type tend to interact with one another) and develop a model that allows us to **predict the likelihood of crimes** based on chosen predictors present in the dataset.

Research questions:

1) How do different crime characteristics such as block, time, date, weather, crime type, crime severity (i.e. arrest or not), etc. tend to interact with one another in the Illinois Institute of Technology area?
2) Of said characteristics, which prove to have a stronger relationship with predicting when and what kind of crime will occur in the Illinois Institute of Technology?
3) How does the model for Illinois Institute of Technology compare to surrounding schools in the Chicago area (UIC, University of Chicago, etc.) in terms of crime?
4) Bring transparency to crime reporting to all Illinois Tech affiliates in a more accurate, accessible, and informative manner compared to the current approach via Public Safety's blog or IITAlert.

Proposed methodology:

- For descriptive analytics:
    - Heatmaps, line graphs, bar charts to show crime frequency, locations, etc.
- For predictive analytics:
    - Linear modelling, K-means clustering, neural networks (perhaps) to develop a predictive model based on a finely-tuned set of features.
        - Baseline model can be the entire city of Chicago *or* can be surrounding universities so we can get a relative normal amount of crime for a college campus in Chicago.
    - We can measure analysis results by testing this model with incoming data and seeing its accuracy and precision, finely-tuning it with this data until the model is reasonably accurate.
    - We need to research how best to work with geospatial and temporal data for both predictive models and visualizing for the user.

Project Outline

Data sources and reference data:

- Public police records ranging from 2001 to present for all criminal activity to occur in the city of Chicago:
    - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data
- Public police records ranging from 2001 to present for all criminal activity to occur within the confines of the Illinois Institute of Technology main campus (with an extra block of range added to all sides):
    - https://data.cityofchicago.org/Public-Safety/Illinois-Institute-of-Technology-Crime-Data/a56f-m4am
- Visual map representation of the dataset above:
    - https://data.cityofchicago.org/Public-Safety/Illinois-Institute-of-Technology/j79s-73yw
- Illinois Tech Public Safety Crime Log (potential to scrape this to be compared to public police record dataset):
    - https://blogs.iit.edu/public_safety/
- Chicago Weather Dataset
- Chicago Population Data (for crime normalization)

Dataset Description:

- Current size of the dataset taking Illinois Tech's campus with an error margin of 2 blocks is about 9,996, but this is subject to updates daily.
- The dataset might have missing data depending on the confidentiality of the crime. For these, the value could be imputed or left missing - but we should refrain from simply removing the row entirely unless the amount missing is truly significant.
    - We will have to produce a system that allows us to impute missing values to gain as much information from each row through thorough inspection of the dataset and consideration of our ultimate goals (ex: if we have a missing block but location coordinates, we will be able to use the row, but missing location coordinates and block fields entirely might mean we have to throw the entire crime out from consideration since there is no way to find the true origin of the crime).
- In addition, location data is shifted a variable amount from its true location to protect some anonymity. To address this, we may have to expand the location boundary around campus to capture crimes that may have been shifted off of campus from this transformation.
    - Using Public Safety's blog (who does *not* shift crime locations) to match actual crimes that have occurred in the dataset that might have shifted locations, we

should be able to develop a model for addressing this location shift in crimes to still generate a full range of crimes that occurred on or very close off of campus.

- The *Block* field does not say the full block with some of it censored - this should not prohibit much analysis as Illinois Tech only has a handful of valid blocks that are immediately obvious to identify, especially paired with both the *X*- and *Y-Coordinate* fields (ex: *033XX S STATE ST* is simple to extract the true block from).

Field (20 fields)	**Description**
Case Number	Unique identifier for each case
Date	Date and time for each case
Block	The block that the crime occurred on
Primary Type	the kind of crime that was committed
Description	String data simple descriptors for each crime
Location Description	The type of area where the crime as committed (ex. Appartment, Sidewalk)
Arrest	Binary attribute indicating whether an arrest was made or not
Domestic	Binary attribute indicating whether a crime was domestic or not
Beat	Beat number at time of crime
District	District that crime occurred in
Ward	Ward that crime occurred in
Community Area	Area that crime occurred in
FBI Code	FBI code for crime
X coordinate	X coordinate in city of crime
Y coordinate	Y coordinate in city of crime
Year	Year that crime occurred
Updated on	Last time record was updated
Latitude	Latitude of crime location
Longitude	Longitude of crime location
Location	Latitude and longitude of crime location as an ordered pair

Many of the attributes such as *Year*, *Latitude*, *Longitude*, *Location*, *Beat*, *Community Area*, *District*, *Ward*, *FBI Code*, and *Updated on* are unnecessary and can be pruned from the dataset. This will reduce the dimensionality and increase the simplicity of the dataset, especially when used for predictive modelling.

Literature review and related work:

- Kang, H., & Kang, H. (2017). Prediction of crime occurrence from multi-modal data using deep learning. Plos ONE, 12(4), 1-19. doi:10.1371/journal.pone.0176244
    - This article approaches an analysis of crimes in the Chicago area by considering environmental factors and context by analyzing images obtained via Google Street view. When paired with information on "demographic, housing, education… economic information… weather data" (2) and even social media activity around the time of the crime, they were able to build a model much more advanced than any current crime predictor. This is done through a multi-tiered process that begins with data collection from multiple sources, an advanced preprocessing stage where data goes through two different tests to determine if it is 'good data' or not, and then multiple different models all fed into a deep learning neural network system finely tuned to make accurate predictions based on recent crime data.
- Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. Statistical Analysis & Data Mining, 9(3), 139-154. doi:10.1002/sam.11312
    - This paper discusses the most important data mining techniques used in analyzing crime, which "include entity extraction, clustering, association rule mining, decision trees, support vector machines, naive Bayes rule, neural networks and social network analysis" (2). The paper is built around the applications of these different techniques working with crime data; in particular, the paper details how clustering is used when dealing with crime analysis, using the techniques to locate 'hotspots' where a surge of criminal activity occurs. From this, it seems k-means clustering algorithms will aid in the predictive phase the project by searching for popular events or details not immediately visible just from the raw data alone. Surprisingly, association analysis does not prove to be too useful in this field for anything other than outlier analysis, which conflicted with my initial thoughts on the matter to use it as a relationship linker between two distinct events. As for decision trees, previous research has suggested their use comes through a trial and error process for predicting crimes, as this tool excels in determining the probability of future events based on past ones.

- EDA analyses of Chicago crime to base our first steps of EDA off of: https://www.kaggle.com/fahd09/eda-of-crime-in-chicago-2005-2016/notebookand https://www.kaggle.com/amunnelly/crime-in-chicago

Tools we plan to use:

- R and RStudio (of course).
- Tableau for high quality visualizations.
- PostgreSQL for database management.
- The SODA API allows us to use the most recent crime data from the data portal without having to download the data to a local machine or server.
    - https://dev.socrata.com/