# Crime Data Analysis and Prediction

Wing Fung, Hui
CUHK
1155152572@link.cuhk.edu.hk

Ka Ho, Poon
CUHK
1155145546@link.cuhk.edu.hk

Wang Lung, Yen
CUHK
1155152698@link.cuhk.edu.hk

Chi Chung, Wu
CUHK
1155145519@link.cuhk.edu.hk

## 1. Introduction

Living in a safe place with a low crime rate has always been so important for everyone. However, no matter how good the public security is, crime still exists in many different forms. To raise the awareness of the general public, the New York City Police Department (NYPD) has uploaded the police complaints data from the year 2006 to 2019 in their website and the dataset is open to the general public. Although it is out of our control to prevent crime from happening, at least we would make good use of the data available on hand to identify the trend in crime in the past and in an attempt to predict and prevent the future crime.

In this project, we would like to apply the knowledge acquired in the course to perform crime analysis and predictions. We will examine the crime data in New York City, due to the volume and variety of the dataset. The data computation will take place on Amazon Cloud Platform and we would also like to build a crime prediction model using time-series forecasting technique. We hope our study will contribute to crime prevention in our society.

## Topics Related

• MapReduce • Apriori algorithm • K-Mean Clustering

• Time series forecasting

## 2 Related Work

The topic of crime prevention is the top priority for governments and academia. Many academic studies, reports, and applications are developed to predict crimes. The Los Angeles Police Department can be a valid example. This organization introduces a system from PredPol, that can provide location-based proactive policing by using machine learning algorithms. Crime prediction is one of our targets too. This predictive result is valuable to optimize the police resource allocation. Moreover, some researches focus on the crime forecast with various algorithms. However, some of them have ignored the basic information of crime. The research "Exploratory Data Analysis And Crime Prediction In San Francisco" compares algorithms' accuracy rate in crime prediction with only the type, time, and location. The age, race, sex of suspects and victims, which can provide valuable imformation for cirme investigation and prevention, are ignored. Thus, in this project, we not only provide a method to extract the crime hotspots and period, but also try to “draw” portraits of suspects and victims for various types of crime. In the end, we will develop a system to predict the crime occurrences in the future.

## 3 Deliverable Plan

Most computation and MapReduce tasks are performed on the Amazon cloud platform, and at the end or project, we will bundle all the AWS settings and deliver it as a crime data analysis engine. It has the capability to fit-in a crime data set to generate crime data analysis results. Also, we will build a crime predictive model, based on time-series forecasting methods. Finally, all the crime data analysis and prediction results will be delivered as graphical presentations, it will include various figures and charts for comparing the crime and the experimental results of different algorithms. We will use Tableau to generate the graphs.

## 4 Dataset

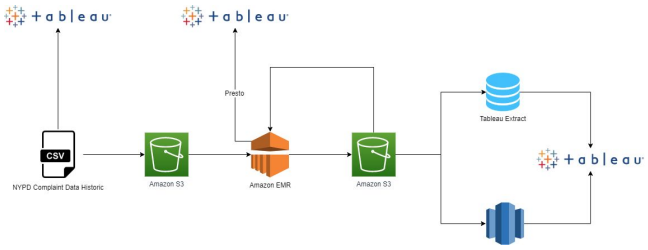
With the perfection and dissemination of technology, it is a trend that the governments develop open data platforms. Open data access policies are giving rise to unforeseen improvements in public and private sector transparency, leading the collective wisdom of the organizations. Many countries, including Canada, the United Kingdom, and the United States, publish the crime datasets.

The NYPD Complaint Data Historic, published by the New York City Police Department, is one of them. This dataset contains the New York City Police Department’s data from 2006 to the end of last year(2019). 6.98 million rows and 35 columns can be reviewed. Compared with datasets from other cities, this dataset equips with a massive amount of data and maximum data dimensions. Based on the volume and the variety of the NYPD dataset, it is chosen for this project.

Related columns in NYPD datasets:

| Column | Description |
|---|---|
| CMPLNT_FR_DT | Date of offence occurrence |
| CMPLNT_FR_TM | Time of offence occurrence |
| PD_DESC | Offense Description |
| Lat_Lon | Geospatial Location Point of offence |
| SUSP_AGE_GROUP | Suspect's Age |
| SUSP_RACE | Suspect's Race |
| SUSP_SEX | Suspect's Sex |
| VIC_AGE_GROUP | Victim's Age |
| VIC_RACE | Victim's Race |
| VIC_SEX | Victim's Sex |

# 5 Technologies




## 5.1 Amazon EMR

Amazon EMR is a platform to process data using open source tools, including Apache Hadoop, Apache HBase, and Apache Flink. The platform is used to run the transformation task for a large amount of data with a strength of HDFS.

## 5.2 Tableau Desktop

Tableau is an analysis tool providing live visual analytics fuel unlimited data exploration. This tool is primarily responsible for analyzing transformed data in S3 and visualizing specific data in various charts and maps.

# 6 Algorithms

## 6.1 K-mean Clustering

K-mean clustering is a method to partition data into k clusters which each datapoint belongs to the cluster with the nearest mean. We use K-means clustering to cluster the crime data in geographical and time domain. As the geospatial location (Lat_Lon) of each offence is provided, we take the geographic distance as the objective function. This would help us find out the crime hotspots. Also, we adopt the occurrence time (CMPLNT_FR_TM) of offences and use the time distance as our objective function. This would help us obtain the frequent time periods of each crime.

## 6.2 Apriori Algorithm

Apriori algorithm is one of the basic algorithms for mining frequent patterns. It scans the dataset to collect all itemsets that satisfy a predefined minimum support. Our goal of using this method is to find the portrayals of suspects and victims of each type of crime. Hence, we implemented the algorithm on sex, race, age features of suspects and victims.

## 6.3 Time Series Forecasting

Time series is a series of data points indexed in time order. We treat the number of crimes that happen each day (CMPLNT_FR_DT) as our time serialized datapoint. We try to forecast the number of crimes that will happen in the future. We will compare the classic autoregressive and moving average (ARMA) models and prophet algorithm released by facebook in 2018. We allocate 90% of our samples for training and 10% for testing. We compare the testing results of these two algorithms in order to find the more robust method for this problem.

# 7 Demonstraction

First, we will demonstrate using MapReduce to process the NYPD crime dataset through Amazon EMR big data platform. Second, we will use Tableau with a graphical user interface to start the analysis and prediction process based on the captured values. Last but not least, we will demonstrate transforming the result by using Tableau Prep Builder to present our observed patterns and predictions.

# 8 Timeline