***Chicago Police Department Project Proposal***

Introduction

Throughout the duration of the CPDB project, our team would like to explore two high level questions that are highly related to each other. First, we would like to understand the characteristics of officers who we believe exhibit signs of consistently overstepping their statute of limitations, and abuse citizens repeatedly. For simplicity, we will refer to such officers as repeat offenders. Additionally, assuming that this cohort can effectively be roughly defined by our analysis, we would then like to examine if there is evidence of systematic protection for the behavior of these officers.

Throughout the proposal we will use the following terms:

* Complaint: A report indicating an incident, but not an explicit legal action
* Allegation: A legal filing derived from a complaint (or possibly not) accusing a party of wrongdoing.
* Acquittal: An instance where someone has be accused of wrongdoing via an allegation, but the justice system determined that the claim was wrong. A case of innocent.
* Conviction: An instance where someone has been accused of wrongdoing via an allegation, and the justice system has determined the claim was truthful. A guilty case.
* TRR: A police report that is filed by police officers following an incident involving police force.

Description and summarization:

The first step in this process will be to determine how to effectively identify repeat offenders. To do this, some quickly computed statistics can be very helpful.

Q #1- *First, what are the average number of complaints and allegations filed against each officer over a specific period of time, and what does the holistic distribution most resemble?*

Q #2- *What is the range of complaints to allegations as a ratio for all officers? Are officers more likely to hold a specific position when they have allegations filed against them - i.e. what are the average number of complaints and acquittals filed for officers at different stages of tenure?*

Q #3- *How many police officers have a significant number of complaints against them that don’t have corresponding TTR police reports?*

Q #4- *We would also like to see the average number of complaints that are filed for off-duty officers, and how this relates to the other metrics.*

Computing these statistics will allow us to get a grasp on police officer behavior at a high level.

Data Integration:

A next fundamental step is understanding the variables computed above from a normalized sense. Explicitly - we want to see how the statistics above change when we incorporate environmental factors, such as an officer operating in a higher crime area and officers interacting with more severe crimes. In order to accomplish this we would like to incorporate the list of All Sworn Employees, Chicago Arrest Data and the Chicago Crime Data. Details of schema retation are shown at the end of this document.

Q #5- *How does normalizing the number of claims for patrol area affect the statistics computed above in the context of all officers (including officers without claims against them)?*

We would like to use the average number of crimes that are committed in the area where each officer operates to do this. Arrests can also be used to see if there is a substantial difference from environmental factors such as crimes being unreported. Ideally, this will also take into account the element of time, as neighborhoods change over the long term. Some specifics are as follows:

Q #6- *Are officers in more crime ridden areas more likely to have more complaints filed against them?*

Q #7- *For each officer, what is the relationship between claims and actually arresting civilians? Is there a pattern here for repeat offenders targeting specific populations? More explicitly, what are the demographics of the victims of officers who we see are exhibiting traits of being repeat offenders vs. their victims demographic information?*

Arrest and crime data will be critical to understanding and computing the above questions.

Workflow analytics:

An excellent spark visualization would be to explore who repeat officers are working with when complaints are filed against them. Once we have an idea of the population of repeat offenders, we can examine if there are clusters of co-accusers.

Q #8- *Do complaints tend to be filed against a cluster of police officers? How do these potential clusters change over time.*

This may indicate that an officer has started patrolling a different area because they are seeking to find victims who are less likely to report abuse based on one incident.

Additionally, we would like to explore the relationship between complaint history and the rate at which new complaints are filed.

Q #9- *Are we able to see if once police officers potentially “get away” with certain behaviors, they are much more likely to keep committing the same acts repeatedly?*

A spark workflow would be great for assisting with a visualization of this potential behavior.

Machine learning:

In terms of applying machine learning to our dataset we like to utilize both supervised and unsupervised machine learning methods. Our supervised method will be used to see if we can build a model to predict officers that will become repeat offenders using the following features:

* Demographics
* Tenure
* Rank
* # of awards
* District/Beat
* Other significant attributes we discover from previous analyses

Additionally, it would be great to use the unsupervised method of k-means clustering to try and identify boundaries between distinct groups of officers based on behavioral statistics as described above.

Modeling with Neural Networks:

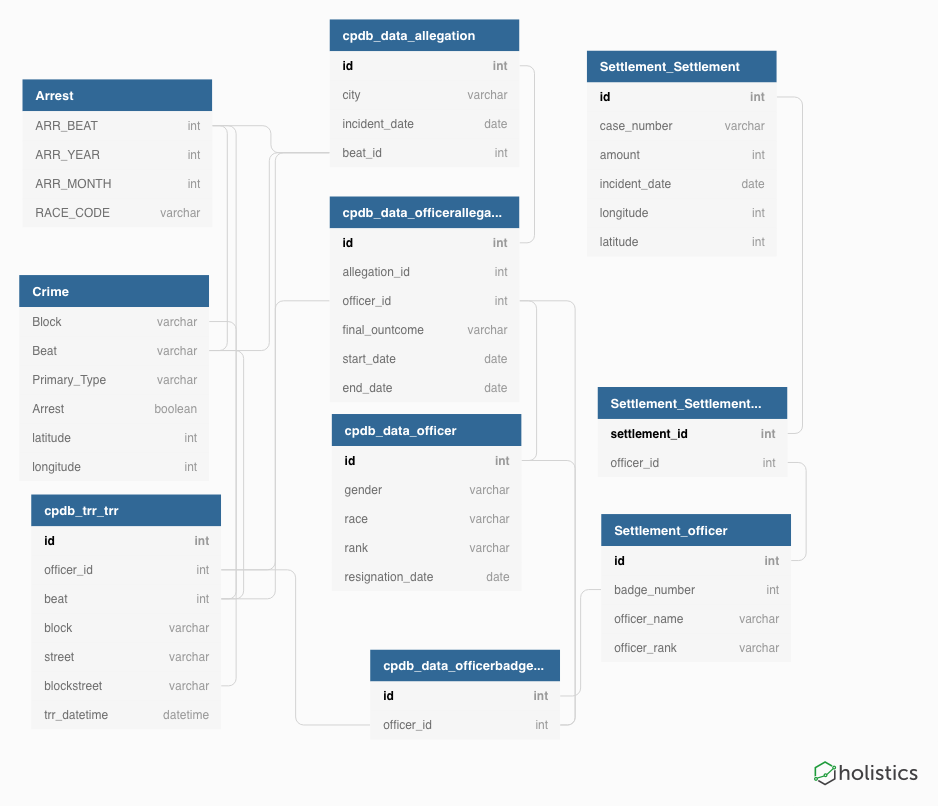
Using neural networks presents a great opportunity to dissect the large amount of text that we will have access to through our reports. We would like to explore the relationship of trying to predict or classify if the officer involved in either the claim report, allegation report, or TTR report is within our repeat offending officers cluster. This would ideally be done using decision trees, and we can take advantage of the metrics above that we have computed. The reason why we want to explore all of these types of reports are we want to see if there are any discernible patterns in the way that these officers tend to write or be described within them - whether from their own personal perspective or one from a potential victim. We would also like to perform a sentiment analysis on the text of these reports to provide additional data on the style of the author - which would be very hard to discern otherwise.

Furthermore, we would like to use a neural network to see if we can successfully predict the location of an incident purely using the text in either the complaints or the TTR reports. The purpose of this would be to try and uncover relationships between specific crime areas that may be hard to pull out with more macro based measures.

Visualization:

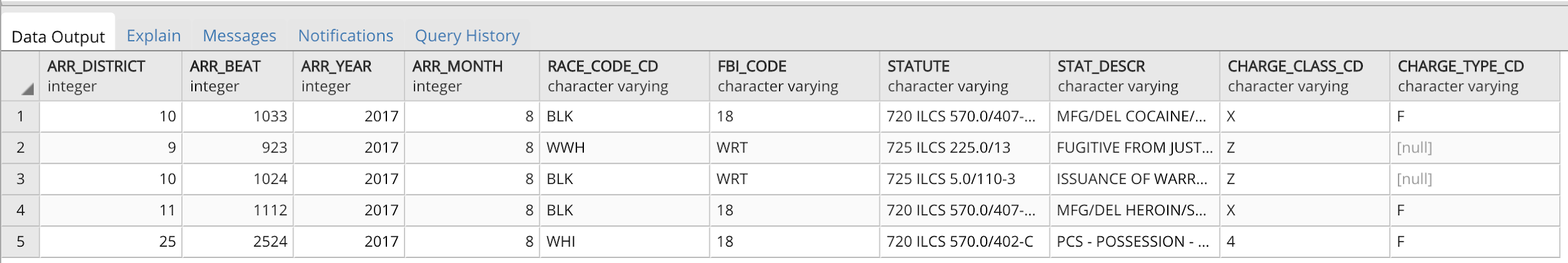
Visualization is an extremely powerful tool, and it will be important to crafting an impactful representation of our results. First and foremost, we would like to visualize the statistics computed in the description of average officers and repeat offenders side by side in histogram form, very similar to the example provided. Ideally, we would like to visualize this before and after normalization from crime and arrest data. Additionally, another cohort that we are very interested in exploring is officers who have received an award. We are interested in this group of officers because we want to explore the premise of the white knight - are complaints less likely to become allegations once officers are deemed “good” from an award. Ideally, we would like to explore this in a similar histogram style, comparing metric values from the awarded officers and the average officer, but now potentially excluding repeat offenders.

Supplemental Schema Details

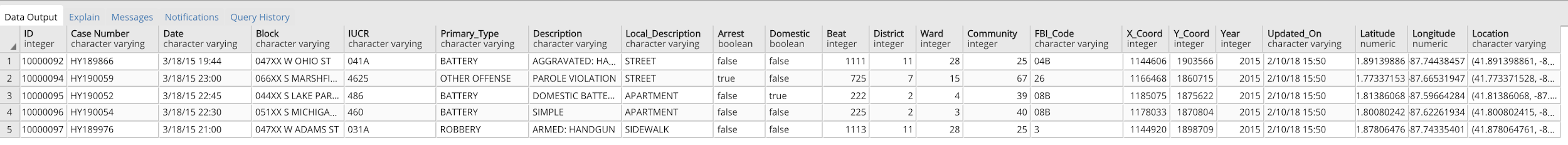
The following figure details how we will integrate the Arrest, Crime, and Settlement sources into the main CPDB database. 

External databases Additional Detail

**The Arrest database** is sourced from the following link: <https://home.chicagopolice.org/online-services/public-arrest-data/>



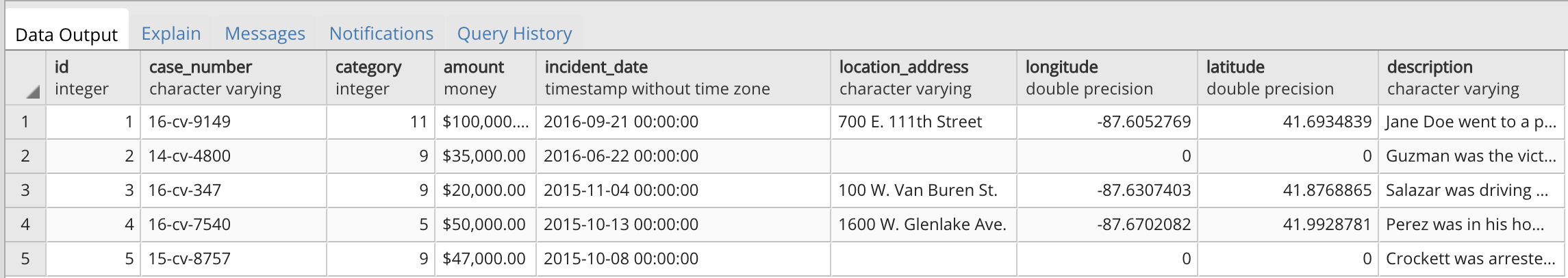
**The Crime database** is sourced from the following link: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>



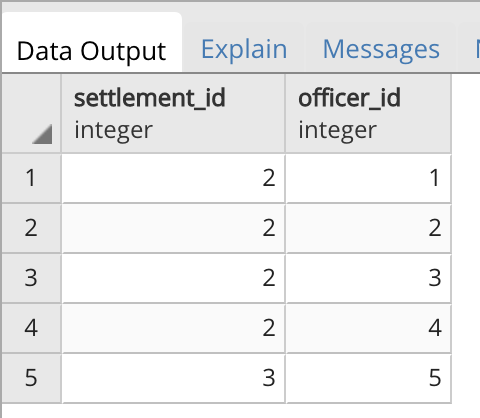
**The Settlement database** is sourced from the following link:

<http://projects.chicagoreporter.com/settlements/search/cases>

Settlement Table:



Settlement\_Officer Table:



Officer Table:

