Data Science Seminar - Project Proposal

The Silver Imps
Song Luo, Jing Jiang, Yuxin Chen

Background and Theme

In an assigned reading article *How to predict bad cops in Chicago*, a term "repeaters" is mentioned. **Our Definition for "Repeaters"**: the top 10% of officers with most complaints. That is to say, if we sort each officer's number of complaints in descending order, we take the top 10% of officers as repeaters. Their misconduct is also of repetitive pattern, namely their most-recent misconduct can be traced back to previous allegations.

We want to do research on "repeaters" and see if there is any trend and trait of the "repeaters" that is worth noting. Also, we want to expand on the time scope in the article and see if we can still predict their future misconduct with previous data.

Research Questions

To evaluate the trend of change for repeaters we decide to use fixed-length time span. Since the most recent incident date is somewhere around 2018, we will have three timespans 2014-2018, 2010-2014, 2006-2010. We will answer the following questions with respect to the timespan.

- Relational Analytics

1. Using our definition of "repeaters" above, what percentage of total complaints are they responsible for?
2. What is the demographic information (race, age and sex) of the "repeaters" using our definition above?
3. What is the distribution of the categories of the misconducts (illegal search, use of force, etc.) for those "repeaters" we defined above?
4. Among the allegations against the "repeaters", what percentage of these cases lead to the "repeaters" being disciplined?

- Visualization

1. Horizontal Bar chart: each bar represents the total complaints of the repeaters each year. With Horizontal Bar Chart we can clearly visualize the changes in the total amount by year.
2. Pie chart(Composition of misconduct categories in repeater complaints): We want to have a straight-forward view of which types of misconducts they are responsible for.

- Interactive Visualization:

1. Zoomable Circle Packing would be a fit for our theme since it enables us to visualize different combinations of different categories, i.e number of repeaters who were above the age of 40 and were committed to multiple use of force.

2. Sequences sunburst is the other Interactive Visualization of us since it can show the categories. This is good for us since it can show things like categories of misconduct or the number of repeaters who receive complaints they are responsible for. It differs from Zoomable Circle Packing as we can use one type of category as “base” and further analyze other types of categories.

- Machine learning:
  We would like to predict the following using machine learning models:

1. The number of all misconducts committed by the repeaters defined above in the Chicago area in the next year.
2. Given the information of a police officer, is he or she more or less likely to be committed to any kinds of misconduct. (Probability)
   For the first one, we might want to use regression models such as linear regression and Naive Bayes. For the second one, we will try models like SVM and neural networks to classify.

- Natural Language Processing
  In CPDP, the misconduct behaviors of police officers are categorized in various types, such as “Use of Force”, “Illegal Search”, etc. We would like to apply NLP models to help classify the type of misconduct type by feeding in complaint text. For this task, we would like to use the Transformer Model to classify the texts. Transformer model consists of an encoder-decoder architecture. The encoder includes a self-attention mechanism and a feed-forward neural network. And the decoder of Transformer includes a self-attention mechanism, an attention mechanism over the encodings, and a feed-forward neural network.
  And beyond that, because our theme mainly focuses on the behaviors of the “repeaters”, we would like to try to separate the data into “repeater” data and “non-repeater” data, and train NLP models to learn from them. And we are interested to see the outcome of the trained models. For example, we can use the trained model for “non-repeaters” and apply them to “repeater” test data or vice versa, and investigate the results.