Grant, Kartik, Sundar

Checkpoint 1 Findings

NOTE: Integrating with the old dataset was more difficult and took a great deal of time, so we made integration one of our questions. We also encountered a bit of a curveball in allegation #2 and had to adjust some of our proposal questions and expectations.

Relational Analytics (4 questions):

1. Can we acquire old allegation data from the Invisible Institute (Rajiv Sinclair) and add that data to our database as new table(s)?

Findings: Our theme is looking at how allegations have been changed over time (e.g. The type of allegation being changed to something less serious). This requires us to access data not currently available in CPDB. Professor Rogers and the Invisible Institute provided us with access to database files and we found **08_28_2018_case_info** to be the most useful for answering these questions.

2. Which allegations have changed?

- Time permitting: What categories of allegations are changed the most?

Findings: File: **cp1_2.sql**. After creating a table called **case_info_08282018** to store the old allegation data, we joined it with other relevant tables in the DB. We joined the allegation tables and the table containing old allegation data **case_info_08282018** to gives us a chance to see which allegations have changed. See **query1_output.png** for the result set.

3. Can we compare the distribution of allegation categories from the old data set case_info_08282018 to the new data set (using COUNT and GROUP BY)?

Findings: Here we compare the distributions of categories in the old data, case_info_08282018, and the new data in CPDB. The output can be seen in Question3_old_data.png and Question3_new_data.png. Further analysis can be done when normalizing the data. This may show how the distribution of allegation categories has changed over time.

4. Using the data generated in Question 3, can we use the discipline data to see how many allegations were disciplined, grouped by category?

Findings: See ****Question4.png**** for output, showing how many allegations were disciplined in each category. Interestingly, ****Drug / Alcohol Abuse**** had a discipline rate of >50%, which is much higher than most categories. On the other hand, ****Illegal Search**** was hardly ever disciplined, having a dismal discipline rate of .3%, implying that either many allegations of illegal search are unfounded, hard to prove, or just simply not discplined for no good reason. It is likely a combination of both. Further analysis may benefit from normalization (time-permitting).