COMP_SCI 396, 496: Data Science Seminar (Fall 2019)

Students Jacoya Thompson (ID 2992892) and Victor S. Bursztyn (ID 3046613). Professor Jennie Rogers.

Project Proposal

Background: Minority residents of Chicago are up to 14 times more likely to be targeted for excessive and deadly force, according to the latest statistics [1]. Seven black people have been killed by police in 2019, and the fight for justice for people of color does not end after Jason Van Dyke. Our group attempts to investigate how socio-demographic features such as race and ethnicity could play a role in the **outcomes** of police misconduct investigations, rather than focusing on the **occurrences** of such misconducts.

Main Research Question: *When a case of police misconduct victimizes groups that are historically marginalized, how unfavourable is its investigation?*

This single question unfolds over a series of hypothetical, measurable effects on the investigation. When compared to non-marginalized groups, to some extent it could:

Effect 1: Slow down a disciplinary investigation;
Effect 2: Decrease the likelihood of the officer(s) involved being disciplined;
Effect 3: Decrease the likelihood of the state granting compensation.

Methodology: It's challenging to use observational data to study causality. For instance, we could find that a particular socio-demographic feature is highly correlated to more severe cases of police misconduct because — the police would argue — "they live in more violent neighborhoods." Fig. 1 shows how a lurking variable could be used to raise uncertainties and limit public debate. However, if we keep track of the distributions of potential lurking variables (namely, the severity of police misconduct allegations), we can demonstrate that investigation outcomes vary significantly even after controlling for these factors. For this reason, we think it's valuable to bring allegation severity into our analysis.

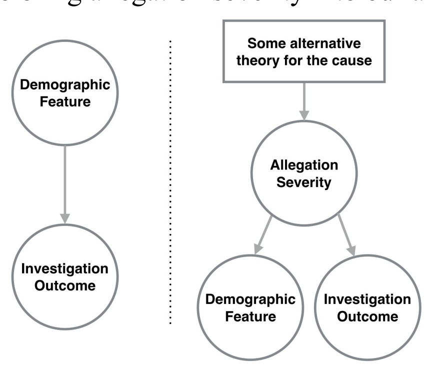

Fig. 1 — A lurking variable confounding the debate of what causes the phenomenon.

Checkpoints and Checkpoint-Specific Questions: This proposal is organized around five checkpoints, each with its own set of questions that contribute to the pursuit of the main research question.

1. **Checkpoint #1 — Relational Analytics:**
   Motivation: to retrieve baselines over the entire population for our hypothetical effects.
   a. *Question #1: What's the percentage of allegations that are sustained?*
   b. *Question #2: What's the average investigation time measured from its start to the first decision?*
   c. *Question #3: What's the most common type of disciplinary action?*
   d. *Question #4: What's the average compensation amount measured from settlements data?*
2. **Checkpoint #2 — Visualization:**
   Motivation: to explore the distributions behind our baselines using victim race as potentially discriminant variable.
   a. *Tableau Chart #1: What's the percentage of allegations sustained across victim races?*
   b. *Tableau Chart #2: What's the distribution of investigation times?*
   c. *Tableau Chart #3: What's the frequency of each disciplinary action?*
   d. *Interactive Visualization #1: What's the influence of victim race on the last two distributions? <u>Implementation details</u>**:** Add an interactive dimension "victim race" to Tableau Charts #2 and #3.*
3. **Checkpoint #3 — Data Cleaning and Integration:**
   Motivation: as we scan Settlements records and try to retrieve victim race from CPDB, we start to assemble tables that can be used in predictive analysis (Checkpoint #5), that is, with victim race (what we are testing as discriminant variable), allegation type (our lurking variable), case's judge (the decision-maker for settlements), and settlement amount (one hypothetical effect).
   a. *Question #1: What's the average compensation amount by victim race?*
   b. *Question #2: What's the average compensation amount by allegation type?*
   c. *Question #3: What's the average compensation amount by victim race and allegation type combined?*
   d. *Question #4: What's the average compensation amount by victim race, allegation type and case's judge combined?*
4. **Checkpoint #4 — Graph Analytics:**
   Motivation: to test whether we can group allegation types (see Checkpoint #3 above) in a way that reflects a severity scale. Having more coarse-grained features is better for fitting predictive models (Checkpoint #5).
   a. *Question #1: How allegation types connect (as vertices) in a graph where edges represent allegations committed by the same officer? Can we identify different levels of severity by looking at either components or communities of allegations? <u>Implementation details</u>**:** Fig. 2 shows how to construct this network, spanning the 415 allegation types.*
   b. *Question #2: If we construct two different networks, one for allegations involving victims of color, and one for white victims, do their topologies match? Can we identify disparities in how allegations connect?*

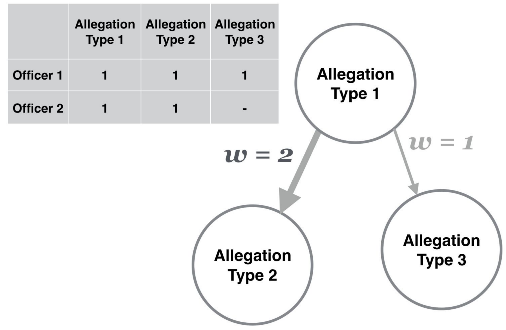

Fig. 2 — A homogeneous network where vertices are allegation types and edges represent co-occurring allegation types in an officer's career. By constructing this network and studying its topology, we may answer Questions #1 and #2.

5. **Checkpoint #5 — Machine Learning and Text Analytics:**
   Motivation is two-fold: (i) to fit models that are better than naive baselines (e.g., the mean or the mode); and (ii) to compare features w.r.t. their predictive power. Two types of Machine Learning models that seem suitable are Decision/Regression Trees and Logistic Regression models. The first allows comparisons of features by looking at their information gains (i.e., the levels of the trees), whereas the latter allows comparisons of features by looking at fitted coefficients.
   Fig. 3 shows how our models will consider at least three independent variables: victim race (what we are testing as discriminant variable), allegation type (our lurking variable), a decision-maker who can be proved biased (either an investigator or a judge, depending on the effect being predicted); and then our effects as the dependent variable (either investigation time, disciplinary action and settlement amount). Finally, Fig. 3 also shows how we can inspect models and compare features w.r.t. their predictive power — in this case, the levels of a Decision Tree.

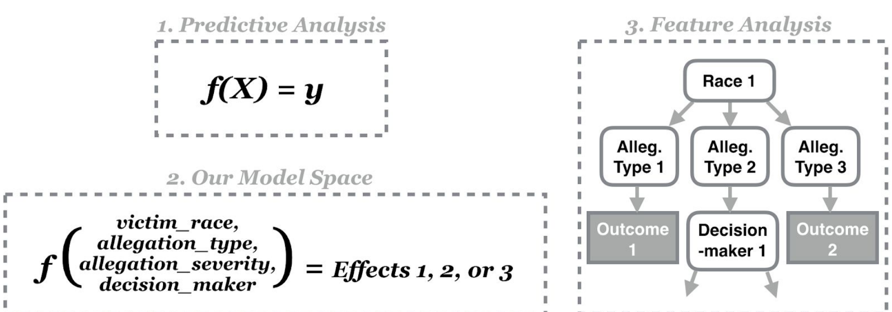

Fig. 3 — How we plan to answer Questions #1, #2, and #3 with our predictive analyses.

   a. *Question #1: How predictive the features* ***victim race***, ***allegation type*** *(with or without a value for allegation severity) and case's investigator (a* ***decision-maker*** *who can be biased) are in predicting the* ***investigation time****?*
   b. *Question #2: How predictive the features* ***victim race***, ***allegation type*** *(with or without a value for allegation severity) and case's investigator (a* ***decision-maker*** *who can be biased) are in predicting the* ***disciplinary action****?*

c. *Question #3: How predictive the features* ***victim race*****,** ***allegation type*** *(with or without a value for allegation severity) and case's judge (a* ***decision-maker*** *who can be biased) are in predicting* ***settlement outcomes****?*
d. *Question #4: Are allegations involving victims of color more likely of being minimized/underestimated in reports filed by investigators?*
e. *Question #5: Does investigator's race affect minimization of allegations involving victims of color?*

Expectations: By the end of Checkpoints #2, #3 and #5, we expect to gather three classes of evidence on our research question:

1. Checkpoint #2 will show how much victim race (the discriminant variable we are set to test) affects the distributions of three potential effects on the investigation, and baselines from Checkpoint #1 should start to be referenced in our qualitative analyses.
2. Checkpoint #3 will indicate potential relationships (or lack thereof) between lurking variables and our hypothesis. In Question #1, we will see how much victim race changes the average settlement amount w.r.t. our baseline from Checkpoint #1; in Question #2, we will see how much allegation type changes the average settlement amount also w.r.t. baseline; and, in Question #3, we will see how the two variables interact by looking at their local averages in a cross product and comparing to results from Questions #2 and #3.
3. Checkpoint #5 is the most important one, as it will show whether victim race is predictive of the three potential effects on the investigation (investigation pace, disciplinary action, or settlement amount); and it will show whether it outweighs what the police could argue to be lurking variables. By narrowing the space to Decision/Regression Trees and Logistic Regression, we can interpret the resulting models in light of our research question.

References:

[1] "Chicago Police are 14 times more likely to use force against black men than against whites." — https://theintercept.com/2018/08/16/chicago-police-misconduct-racial-disparity/