Princeton University

Princeton, New Jersey

**DRAFT** November 10, 2017

**Research project in collaboration with the Manhattan District Attorney’s Office**

**I. Research Statement**

Deep learning, machine learning, and advanced statistical methods are increasingly impacting legal practice. To date, however, this impact has largely been limited to the private sector. Our research, in contrast, represents an attempt at using these methods to increase knowledge and aid decision making in the public legal sector—specifically, in prosecutorial work.

Our research program contains three phases. The first phase involves garnering and analyzing data from closed criminal cases. (For a full description of data points sought, please see Section II *infra*.) Data analysis will explore factors that potentially influence prosecutorial decision making, including but not limited to race and gender. Some analysis will take place through traditional hypothesis testing and some will involve using deep learning applications on unstructured data, with the latter referring primarily to analysis of arresting officers’ narrative reports. In all, the goal of the first phase is to identify and gain insight into the factors impacting prosecutorial decision making.

The second phase involves using predictive analytics and machine learning to develop a model that, given an arresting officer’s narrative report and recommended charges, can predict what the outcome of the case would be if the defendant were treated race- and gender-neutrally.

The third phase involves experimental testing of the model. For a discrete period of time, a random selection of Manhattan prosecutors will use the model (experimental group) while the remaining prosecutors in the office will receive no intervention (control group). If a prosecutor in the experimental group, in drafting a complaint or in the course of plea bargaining, finds that he or she is significantly deviating from the model’s predicted outcome, the reasons for the deviation will be documented. These deviations will be used to refine the model moving forward.

Taken together, the three phases will provide (1) insight into prosecutorial decision making, (2) a model that can anchor prosecutorial decision making to race and gender neutral outcomes, and (3) assessment of whether such a model can be beneficially integrated into legal practice.

**II. Data Overview**

Data sought includes complete case files for all closed cases, beginning with approximately the year 2000 and ending with approximately 2016. For each case, of primary importance are the following:

* The arresting officer’s report, which may include:
  1. Criminal charges suggested by the arresting police officer,
  2. A short narrative of the arrest (often called a “probable cause statement”), and
  3. A longer narrative of the arrest (often drafted a day or two following the arrest);
* The criminal complaint filed by the prosecuting attorney, which may include:
  1. Criminal charges;
* The results of the indictment, if applicable;
* The final disposition of the case, which may include:
  1. Conviction charges,
  2. Charges dismissed,
  3. Sentence details, and
  4. How the case was resolved (dismissal, plea bargain, trial).
* Other information of importance includes:
  1. Defendants’ race, age, gender, criminal history at time of arrest, residential zip code, level of education, and income level,
  2. Prosectuors’ race, age, gender, length of employment by the District Attorney’s Office, position within the office, and department within the office, and
  3. Arresting officers’ race, age, gender, length of service on the police force, position on the force, and precinct;
* Additional information includes:
  1. Transcriptions of or notes concerning interviews with the defendant, the victim(s), and witness(es).

**III. Princeton Research Team**

Joseph Avery, J.D., is the primary contact for the project. Faculty involved include Joel Cooper, Professor of Psychology; Arvind Narayanan, Assistant Professor of Computer Science; and Samuel Wang, Professor of Molecular Biology and Neuroscience. Graduate students involved include Lawrence Yuanda Xu, a PhD candidate in the Program in Applied and Computational Mathematics; and Mauricio Matsumoto, a PhD candidate in the Department of Economics. Vinícius Pantoja, a researcher from Fundação Getulio Vargas in Brazil, also is involved. Lastly, additional graduate students from the Departments of Computer Science and Applied and Computational Mathematics may contribute.

All individuals working on the project will be bound to confidentiality and non-disclosure as set forth *infra*.

**IV. Confidentiality**

All records from this study will be kept confidential. We will not include information that will make it possible to identify defendants, police officers, police precincts, or prosecutors. The identity of the Manhattan District Attorney’s Office itself also will be kept confidential and will not be disclosed to parties beyond the Research Team. Research records will be stored securely on password-protected computers and Princeton University-operated servers, and all data will be encrypted. The Research Team is the only party that will have access to the data.