Government Analytics Using Machine Learning

This version: June 7th, 2022

Authors:
Sandeep Bhupatiraju[1]
Daniel Chen[2]
Slava Jankin[3]
Galileu Kim[4]
Maximilian Kupi[5]
Manuel Ramos Maqueda[6]

Abstract
The use of machine learning (ML) offers new opportunities for improving the productivity of the public sector. Increasing availability of public sector data and algorithmic approaches provide a conducive environment for machine learning for government analytics. However, the successful deployment of machine learning solutions requires first developing data infrastructure of the required quality to feed these algorithms, as well as building the human capital necessary to develop them. Ethical principles regarding the use of ML technologies must be defined and respected, particularly for the justice system. This chapter provides an overview of potential applications of ML in the public sector and in the justice system specifically, as well as the necessary steps to develop them sustainably and ethically. It then analyzes the case of ML deployment in India to illustrate this process in practice.

Acknowledgments: Authors' names are listed alphabetically. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

[1] Consultant, DIME.
[2] Senior Economist, DIME.
[3] Professor of Data Science and Public Policy, Hertie School
[4] ET Consultant, DIME.
[5] Ph.D. candidate, Hertie School.
[6] Research Analyst, DIME.

Practitioner Points:

1. **Machine learning is fundamentally a methodological approach: it defines a performance indicator and trains, using collected data, an algorithm to improve this indicator.** Because of this relatively broad definition, machine learning (ML) includes different algorithms and may be applied in a variety of domains, from payroll fraud detection to court rulings. This flexibility requires practitioners to make key design decisions: what kind of performance indicator will be used? What training data and algorithm will be deployed? These decisions may substantially alter the ML algorithm's results. Making these decisions thus requires close collaboration between ML engineers, domain experts, and the agencies that will use the technology.
2. **While ML can offer efficiency gains in public administration, governments need to be aware of their role in generating and using measurements in public administration.** ML algorithm requires extensive data and measurements on both citizens and civil servants, but this data is often collected without their consent. As a result, governments should be transparent on how this data is being used to enhance public administration.
3. **Ethical principles should be prioritized in the deployment of ML applications.** This requires communication to citizens and civil servants on how these technologies are being used to affect public administration. Care should be taken not to reproduce biases such as racial or gender discrimination in the ML algorithms. If these measures are not taken, governments risk losing the trust of citizens and civil servants who stand to benefit from these technologies.
4. **In order to fully reap the benefits of ML, governments must undertake costly and extended investments in data infrastructure and human capital.** Before machine learning is implemented, a long investment in data infrastructure must take place. Data quality needs to be improved, as well as developing data pipelines to train the algorithm. Additionally, specialized ML engineers must be hired and trained to implement the technology in the public sector. These investments are costly and require long-term planning: governments should not expect ML technologies to be developed overnight.
5. **ML experts must be advised by subject matter experts to guide how the ML technology will benefit civil servants who will use the technology.** Besides the technical knowledge to develop and operate ML technologies, substantial levels of domain and political expertise as well as awareness about potential ethical and legal pitfalls are necessary to ensure the effective use of the ML solution. For example, if the ML technology is to assist judges in reducing racial bias, judicial experts and judges themselves should be consulted to ensure that relevant performance indicators and data are used. The differential uptake of ML recommendations by judges may induce. An evaluation of both the ML algorithm and its actual use should be therefore considered before and during its deployment.
6. **Machine learning is not a panacea, and practitioners should be aware of the limitations of this approach.** Algorithms are limited by that which is measurable by data, and performance indicators may reflect the bias of ML engineers and even subject experts. Improving a particular performance indicator may not necessarily be the best way of achieving a policy goal. As a result, ML applications should not be considered a substitute for policymaking, but a tool to complement and enhance decisions made by government agencies and their civil servants.

1. Introduction

Machine learning (ML) is a discipline that focuses on the development of a computer system (machine) that through the analysis of training data can improve their performance (learn).[7] Recent advances in data collection and processing power have expanded opportunities for machine learning applications in a variety of fields. Advances in machine learning have brought tangible benefits in the world of business, medicine and large-scale systems more generally.[8] However, this growth in opportunities has often led to excessive optimism on what it can accomplish, as well as downplaying the often steep costs of deploying ML technologies.[9]

We start our discussion by offering a general definition of ML. What distinguishes machine learning from other methodological approaches is the definition of a learning problem under a statistical framework. Following Jordan and Mitchell (2015), we define a learning problem as the "problem of improving some measure of performance when executing some task, through […] training experience."[10] The following example illustrates a machine learning approach. Suppose a government is interested in reducing irregularities in its payroll system. One measure of performance would be the proportion of irregular paychecks correctly identified by the machine. The training experience – or data – would be a collection of paychecks manually classified by payroll analysts.[11] The learning problem is thus defining a statistical model that learns how to best predict irregular paychecks, trained on historical payroll data.[12]

In this chapter, we discuss applications of machine learning to public administration. We outline the data infrastructure and human capital requirements for developing ML applications, which are considerable. As noted in chapter 6, Data Infrastructure for Government Analytics, the foundational step for any form of data analytics – including ML – is the development of a robust data infrastructure. We also highlight ethical concerns regarding the development and deployment of ML applications, which relate directly to the discussion on chapter 5 of this Handbook. Despite the focus on machine learning, we consider that the broader shift to a data-driven and statistically informed culture – regardless of the implementation of algorithms – is often already sufficient to bring substantial benefits to public service delivery. These include organizational changes, data literacy and performance monitoring. With a transition to a data-driven policymaking, ML applications become a natural next step in government analytics: automatically leveraging data to improve the performance of public administration through well-defined performance metrics.

Following this broader discussion on ML in public administration, we focus on machine learning applications in justice systems. Within public administration, the justice system generates a large

---

[7] Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), pp.255-260.

[8] E Brynjolfsson and A McAfee, 'The business of artificial intelligence' (2017) Harvard Business Review, 1-20.

[9] JH Chen and SM Asch, 'Machine learning and prediction in medicine—beyond the peak of inflated expectations' (2017) 376 (26) The New England Journal of Medicine 2507.

[10] Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), pp. 225.

[11] For more details, see chapter 6.3.

[12] ML is therefore a methodological approach anchored in a learning framework. It is not a radical departure from more classical approaches to statistics and in fact often builds on canonical models (such as linear or logistic regressions), nor is it exempt from well-known challenges such as bias and model misspecification.

number of case rulings linking legal cases and actors (training data) to ideally fair rulings (performance). A ML learning problem would thus be: training an algorithm in a collection of case rulings, can we identify and reduce racial bias in court rulings? Instead of defining what a fair ruling is, we may define what an unfair ruling is. For instance, the decision of the judge should not be influenced by extraneous factors that are unrelated to the case, such as the time of the day in which the judge rules or the race of the defendant. By identifying cases in which the decision is influenced by such biases, a ML model has the potential to identify and ultimately prevent unfair rulings. This approach thus allows ML to improve the quality and fairness of judicial decisions.[13] Despite this promise, limited data literacy and statistical training inhibits applications of data analytics in the judiciary more generally and machine learning specifically.

Machine learning is not a panacea, requiring significant – and often prohibitively costly – investments before any of its benefits come to fruition. As we detail throughout the chapter, ML applications require expensive investments in data infrastructure and developing the necessary human capital to develop and deploy ML algorithms. Undertaking these investments and often long cycles of development may be out of reach for some practitioners. Additionally, ethical concerns regarding embedded racial or gender biases in training data highlight how technologies can inadvertently reproduce the same human biases they were designed to substitute. Thus, initial optimism regarding the revolutionary potential of ML approach should be balanced by a recognition of its limitations.[14] Practitioners have much to gain from deploying machine learning in public administration but should approach so cognizant of the challenges in doing so. Having said that, a lot of improvement in performance can come simply from embedding a data-driven approach to government functioning. There is not always a need for sophisticated approaches in ML to make progress. In fact, ML generally comes only after more basic steps have been taken on data management and analytics in public administration, as highlighted in other chapters of the book.[15]

The chapter is structured as follows. Section 1 describes applications of machine learning in public administration. Section 2 then outlines a roadmap to applying ML in the public sector, focusing on data infrastructure requirements and human capital needs. Section 3 shifts our focus from public administration more generally to the justice system. In doing so it highlights applications of ML in the justice system, as well as the data infrastructure and human capital requirements to implement them. Section 4 presents a case study of India to illustrate ML approaches to justice in practice. Section 5 moves beyond descriptive analysis, to outline how ML can be used to assist causal inference. Finally, we conclude.

2. Machine Learning for Public Administration

The use of ML is spreading across many functional areas of public administration. While European Union governments focus on service delivery and public engagement, other areas such as internal management or law enforcement are progressively being targeted for the deployment of ML

---

[13] Ramos-Maqueda, Manuel and Daniel Li Chen, "The Role of Justice in Development: The Data Revolution". World Bank Working Paper Series.

[14] https://www.economist.com/technology-quarterly/2020/06/11/an-understanding-of-ais-limitations-is-starting-to-sink-in

[15] See, for instance, chapter 6.0 on HRMIS systems.

solutions to increase their efficiency and effectiveness.[16] The applications are diverse, from detection of Covid-19 outbreaks to simulating the impact of changes in macroeconomic policy. ML applications thus provide novel ways for governments to use their data to improve public administration. One of the chapters in this Handbook, chapter 12, highlights how machine learning can be used to detect similarities between goods in public procurement.[17]

The use of ML provides a few advantages compared to more standard analytical approaches. Standard data analytics provide the analyst with tools bounded by the analyst's capacity to investigate connections between variables in the data – often the coefficients in a regression specification. However, in many public administration settings, the analyst is faced with factors – individual or organizational – that may be influencing a policy outcome without the analysts' knowledge. ML enables the exploration of relationships between variables in a principled, and often unsupervised way. However, causality in ML is a relatively recent development, and presents considerable challenges.[18] Potential applications therefore focus less on causal interventions, and more on predictive applications. of ML can be subsumed under the following three categories:

Detection and prediction:

ML can help policy makers detect and predict destructive events, improving the design and implementation of adequate policy measures. This is the largest application of ML approaches, addressing issues such as Covid-19 outbreaks, fake news, hate speech, tax fraud, military aggression, terrorist activity, cyber-attacks, natural disasters, street crime, and traffic congestion – only to mention a few. While the detection and prediction of destructive events is only the first step towards effective government intervention, it is an important instrument for effective policy making.

For example, in Delhi, over 7,500 CCTV cameras, automatic traffic lights, and a thousand LED signs are equipped with sensors and cameras that collect traffic data, which a ML system processes into real-time insights. Local authorities can then use it to decide how to balance traffic flow in real-time, and identify traffic patterns and congestion trends in order to plan for the long-term to mitigate traffic problems.[19] Besides these benefits geared towards improving the general traffic flow, these systems are also being used by the Delhi police department to track and enforce traffic violations such a speeding or illegal parking.[20]

Simulation and evaluation:

Simulating and evaluating the impact of future policy measures is another widespread application area for machine learning. Being able to simulate the potential costs of a policy measure against its expected benefits has become an increasingly relevant tool for governments. For example, in

---

[16] G Misuraca and C van Noordt 'Overview of the use and impact of ML in public services in the EU' (2020) EUR 30255 EN, Publications Office of the European Union, Luxembourg

[17] The machine learning application is part of a broader study on Bureaucratic allocation, available in Bandiera, Best, Khan and Prat (2021).

[18] For a discussion, see here.

[19] J Devanesan, 'AI-powered traffic management is slashing Asia's congestion problem' (August 2020) Techwire Asia, https://techwireasia.com/2020/08/ai-powered-traffic-management-is-slashing-asias-congestion-problem/

[20] N Lal, 'How traffic cameras issue e-challans' (April 2021) The Times of India, https://timesofindia.indiatimes.com/city/delhi/how-traffic-cameras-issue-e-challans/articleshow/82103731.cms

the United States a simulation known as the National Planning Scenario 1 allows policymakers to simulate what would happen if Washington D.C. were subject to a nuclear attack.[21] Be it policies designed to stimulate the economy or to contain the spread of a virus, simulation and evaluation provide valuable insight to policymakers before implementing them, allowing them to choose which one maximizes intended policy effects.

Personalization and automation:

Machine learning can also be applied in the personalization and automation of government processes and services. For example, policymakers may custom tailor digital government services for parents to every life-stage of their new-born child or fitting the provision of healthcare services to every patient's particular needs. Additionally, automation of repetitive tasks leaves more time for other tasks for public servants. All in all, these novel technologies may assist governments to be more efficient in their use of time and increase their responsiveness to citizens' needs.

A medical example illustrates this approach. There has been growing interest within the United States federal government to use machine learning to improve public health outcomes. A series of pilots to develop such ML solutions have been rolled out. These include use of medical reports in order predict potential adverse drug reactions, the classification of whether a child is likely to have autism based on medical records, or the prediction of unplanned hospital admissions and adverse events.[22] Another study has found, through the application of machine learning techniques, that physicians over test low-risk patients but simultaneously undertest high-risk patients.[23]

2.2. Practical Steps for ML in Public Administration

The implementation of ML in public administration comprises two key steps. The first one is building a high-quality data infrastructure to feed the necessary training data to the machine learning algorithm. Since in public administration, data infrastructures are often developed without ML applications in mind, adaptation is often necessary. New data pipelines need to integrate public sector information systems that previously operated in isolation, such as public procurement and budget data. Data standardization practices, ensuring that variables in different data tables are named consistently, as well as other quality checks need to be in place to ensure that the data fed to the ML system is accurate and comprehensive.

Another key step is developing the human capital necessary to deploy machine learning. Before fulfilling the promise of automated and self-learning algorithms, a team of human developers is necessary to set the system in place. In fact, the entire pipeline, from data infrastructure to the training of the algorithm, to disseminating actionable insights for policymakers has to be designed by humans. Having an in-house team capable of developing and maintaining ML applications is crucial. Continuous collaboration between the ML implementation team and policy colleagues who

---

[21] Waldrop, M.M., 2018. Free agents.

[22] DF Engstrom et al., Appendix to 'Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies' (2020) https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-Appendix.pdf

[23] Sendhil Mullainathan, Ziad Obermeyer, Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care, The Quarterly Journal of Economics, Volume 137, Issue 2, May 2022, Pages 679–727, https://doi.org/10.1093/qje/qjab046

will use its insights ensures that applications are adapted to and stay relevant to public administration's needs.

The following sections dig deeper into these steps. In doing so, we highlight examples of strategies to ensure that both the data infrastructure and human capital requirements are in place to deploy machine learning in public administration.

2.3. Public Sector Data Infrastructure for ML

Machine learning requires large volumes of data. This data should be of high-quality: it should be comprehensive, covering all measurements necessary for the algorithm, and complete, reducing to the extent possible any gaps in measurements that may arise. A robust data infrastructure ensures that these two principles are respected and is a pre-requirement for any ML application. The implementation process may require upgrading legacy information systems or integrating new systems into old ones to process the resulting, oftentimes large, datasets. Practitioners may benefit from referring to chapter 6.0 (HRMIS) for a wider discussion on how to reform data infrastructure for analytical insights, which provides lessons that are applicable to ML settings as well.

Having said that, some types of data structure may be more amenable for machine learning than others. ML applications often require structured data – with well-defined formats and measurements – so policy areas that traditionally deal with structured data such as finance or budget lend themselves particularly well for it (for an overview of using budgetary data for analytics, see chapter 9). At the same time, governments produce unstructured data – which lack a pre-defined data format – such as written documents, meeting recordings and satellite imagery. To take full advantage of this range of data, practitioners should develop a flexible storage solution that accommodates different types of data. This flexibility should be complemented by thorough documentation of data collection and standardization practices, as well as ensuring compliance with data security regulations such as the General Data Protection Regulation (GDPR).

The deployment of this data infrastructure requires long-term and costly investment. Data engineers and IT technicians should partner with the ML implementation team to define data requirements, identify relevant variables, and connect the ML applications to the data infrastructure. The development of a robust data infrastructure is foundational for the effective deployment of ML in public administration and should always precede it. Since the data infrastructure is embedded within public administration, its development requires careful coordination between ML engineers, data engineers, and institutional counterparts who own permissions to the data. Data should be integrated across government agencies to ensure that the largest pool of data is made available for training the application. Open communication between teams and agencies is therefore key.

2.2 Human Capital Requirements for ML in Public Administration

A sustainable ML application is often best achieved by building on in-house human capital. This way it can be ensured that the developed solutions are in line with existing government regulations and policy choices are encoded faithfully. Furthermore, in-house ML experts will be more likely to possess the necessary subject and political expertise required to implement ML solutions in a policy area. Finally, even in case an agency decides to mainly rely on procuring ML solutions from external service providers, a certain level of embedded expertise is required in order to know what is technically possible and feasible, as well as making informed judgements about the quality of contractor-provided solutions.

To build the necessary skills infrastructure in government organizations it is first necessary to understand what competencies are needed for ML developers. Naturally, knowledge about machine learning and deep learning algorithms are necessary. Beyond this basic knowledge, methods for dealing with large-scale data and databases in general and knowledge about distributed computing systems are also key. To successfully develop, deploy and operate machine learning in government, familiarity with human-centered design, as well as acquaintance with the legal and ethical frameworks in public administration are important. Finally, policy area expertise and knowledge about governance and policymaking in general enables ML applications to be anchored in the operational needs of government.

Integrating the necessary skills infrastructure in government organizations oftentimes proves to be rather difficult. Hence, it is advisable for governments to follow one or more of the following best practices. ML talent does usually not follow the classical tenure paths of public sector officials. Lateral entries or dedicated programs that allow entries for a limited amount of time can be an effective method to attract these specialists into government office. Furthermore, adapting job classification schemes to include ML-related job categories and increasing salaries and career prospects to better compete with comparable private sector job placements are advisable. Ultimately, it is important to raise awareness amongst the target talent group of the motivating challenges (e.g., social impact) and rewarding benefits (e.g., job stability and work-life balance) that working for the public sector can have.

Once in government, ML experts' work can benefit greatly from exchange and knowledge sharing with colleagues. Establishing so called communities of practice to cross knowledge boundaries within and across agencies can help gain legitimacy vis-à-vis relevant stakeholders and foster collaboration among different agencies. Including non-tech colleagues in these communities can also ensure that ML applications are developed in a user-friendly manner and integrate well into the daily activities of public servants. Another often applied practice is the establishment of excellence centers that offer research, support and training services and help agencies stay on the cutting edge of ML technology. Finally, open communication, e.g., through blogs or dedicated events, helps other departments take notice and learn from each other's experience.

Collaborations with external experts and research institutions can be another effective approach to bring external expertise into a specific project context while maintaining control and monitoring quality. Besides concrete project collaborations, establishing academic partnerships like mobility or internship programs for the temporary assignment of personnel between government agencies and universities or research centers can help institutionalize such collaborative efforts. Also, tailoring ML-related educational offers of partner academic institutions to the particular needs of government organizations can be a viable way for ensuring inflow of ML talent. Finally, building and sustaining intersectoral and interdisciplinary networking initiatives focused on the use of ML in government can help establish collaborations and foster learning and exchange.

2.4. Ethical Considerations for the Deployment of ML

Ethical considerations should be at the forefront of deployment of ML in public administration, and analytical applications more broadly.[24] The social contract between governments and citizens differs substantially from the one that private sector companies have with their customers. Citizens or civil servants rarely have a choice whether to share their data with their government. This makes data security and privacy particularly sensitive since most ML applications require rather substantial amounts of data for appropriate training. On top of more generally applying regulations like the General Data Protection Regulation (GDPR) of the EU, ensuring the responsible usage and sharing of data, potentially also by applying adequate anonymization techniques, should be a priority for governments to ensure citizen's trust (for a larger discussion, see Chapter 5).

Another factor that can inhibit citizens' trust stems from the rare position that governments have vis-à-vis ML technologies. They unify the role of users and regulators in one entity. This makes public sector's usage of ML a particularly delicate target of public scrutiny. Cases where governmental ML systems violate citizens' rights, like in the recent case of the Dutch automated surveillance system for detecting welfare fraud, pose serious threats to citizen trust (see Box 1). Faithfully encoding legal and political choices and ensuring compliance with international regulatory frameworks are therefore necessary for ensuring ethical machine learning applications in the public sector.

Box 1: The Precedent Case of the Dutch Automated Surveillance System

On February 5, 2020, the District Court of The Hague ruled that the System Risk Indication (SyRI), a ML application used by the Dutch government to detect welfare fraud, violates Article 8 of the European Convention on Human Rights (ECHR), i.e., the right to respect for private and family life. This case is one of the first times a court has stopped a government's use of ML technologies on human rights grounds, and thus is considered an important legal precedent for other courts to follow.

The SyRI system was designed to prevent and combat fraud in areas such as social benefits, allowances, and taxes by linking and analyzing data from various government and public agencies and generating personal risk profiles. It was deployed by the Dutch minister of social affairs and employment upon request of various public agencies, which include, among others, municipalities, the Social Insurance Bank, and the Employee Insurance Agency. The system mainly used a neighborhood-oriented approach, meaning it targeted specific neighborhoods where the linked data indicated an increased risk of welfare fraud.

Although the Court agreed with the Dutch government that the fight against fraud is crucial and thus employing novel technologies which offer more possibilities to prevent and combat fraud generally serves a legitimate purpose, it ruled that the way the SyRI was operated did not strike a "fair balance" between social interests and any violation of the private life of citizens as required by the ECHR. In particular, the Court stated that, due to the lack of insights into the risk indicators and the operation of the risk model, the system violated the transparency principle and discrimination or stigmatization of citizens in problem areas could not be ruled out.

The ruling, which lead to the immediate halt of the SyRI system and caused public uproar far beyond the Netherlands, is a telling example for the potential negative consequences of applying ML systems for government purposes without adequately addressing their potentially ethically adverse side effects.

---

[24] For an overview, see chapter 5, "The Ethics of Measurement in Public Administration", of this Handbook.

Applications of ML in government must consider how citizens can often only rely on government for a public service, e.g., social security. This is a particular challenge to using ML in settings where the algorithm must make a choice. For instance, regarding social security systems, an algorithm may decide whether a citizen is eligible for a particular government benefit. In these situations, the algorithm must compare what would happen if a citizen was granted that benefit versus not. The algorithms that underlie this decision-making have to make assumptions about what would happen in each scenario, and the usefulness of its final decision depends on how appropriate these underlying assumptions are. If a citizen is denied that service due to an algorithm's decision, who holds the algorithm accountable?

Often, modeling assumptions are not directly testable and hence require a substantial level of expertise over both what assumptions the algorithm is making and the suitability of those assumptions for a given public sector setting.[25] Public policy making through machine learning therefore raises important ethical questions. Choices may be made on behalf of government officials about citizen outcomes by machines they do not fully understand. As such, there is a tension between using machine learning technology to improve public administration and the oversight required to ensure that its use is in accordance with ethical principles. This tension becomes particularly salient when using previous administrative data for algorithm training leads to human biases being translated into the system. Not uncovered, these biases can lead to "discrimination at scale" in sensitive areas such as racial profiling.

Finally, most applications of ML for governmental purposes are not static and should be adapted to evolving understandings of ethical principles. For example, algorithms for detecting fraud need to constantly be updated or retrained in order to address new forms of misconduct uncovered by agency employees and avoid excessive focus on past forms of misconduct. Without such updating, algorithm's may be biased towards past versions of criminal conduct. Constant updating by consulting domain experts and ethical advisors is necessary to ensure the effectiveness and ethical compliance of ML technologies in government.

3. Machine Learning for Justice

We now turn our focus on applications of ML in the justice system. The justice system is an institutional setting with high-frequency data, well-documented cases, extensive textual evidence and a host of legal actors. As such, it is a useful setting in which to explore the use of ML to administration in the public service. An example of a core analytical question in justice is how the characteristics of judges impact judicial outcomes such as rulings. This is a formulation of a wider question as to how the individual characteristics of public officials impact the quality of public services provided by the government. It is a question that the analytics of public administration can investigate with the right measurement, data infrastructure and skills for analysis.

[25] S Athey 'Beyond prediction: Using big data for policy problems' (2017) Science, 355(6324), 483-485.

Significant progress has been made to answer this question, using ML.[26] In the United States, ML is already used in the processing of bail applications, DNA analysis of crimes, gunshot detection, and crime forecasting.[27] The large volume of data from surveillance systems, digital payments platforms, newly computerized bureaucratic systems and even social media platforms can be analyzed to detect anomalous activity, investigate potential criminal activity, and improve systems of justice. For example, in the 2021 January 6th Capitol riots, investigators used ML-powered facial detection technologies to identify participants and initiate prosecutions.[28] ML systems can also reduce the barriers to accessing courts by providing users with timely information directly, rather than through lawyers. Sadka, Seira, and Woodruff (2017) find that providing information to litigants in mediation reduces the level of overconfidence of litigants, and nearly doubles the overall settlement rate, but this only occurs when litigants are directly informed rather than their lawyers.[29]

The application of ML systems to justice systems is useful because slight tendencies in human behavior can have significant impacts on judicial outcomes. A growing body of work demonstrates how small external factors, most of which the participants are unaware of, can leave their mark on the outcomes of legal cases. Analysis of the US, French, Israeli, UK, and Chilean courts for example, find in various settings, that the tone of words used in the first three minutes of a hearing, the incidence of birthdays, the outcomes of sporting events, and even the time of day of a hearing or defendant's name, affect the outcome of cases.[30] The analysis of 18,686 judicial rulings, collected over seventy-seven years, by the twelve US circuit courts (also known as courts of appeals or federal appellate court) illustrated that judges demonstrate considerable bias before national elections.[31] Similarly, there is new evidence on sequencing matters in high-stakes decisions: decisions made on previous cases affect the outcomes of current cases, even though the cases have nothing to do with each other. Refugee asylum judges are two percentage points more likely to deny asylum to refugees if their previous decision granted asylum.[32]

Given the abundant evidence of how bias shapes decisions made by officials in the justice system, ML methods can identify these sources of bias and signal when they shape judicial outcomes. The subtlety of different forms of biases requires an approach that searches through a very large number of relationships before for their wider effects to be detected, for which ML may be well suited. This can result in a more streamlined system and reduction in backlog. Such tools can identify discrimination and bias even when these are not evident to the participants in the courts

---

26 DL Chen, 'Judicial analytics and the great transformation of American Law' (2019) 27(1) Artificial Intelligence and Law 15-42; C Rigano, 'Using artificial intelligence to address criminal justice needs' (2019) 280 National Institute of Justice 1; DL Chen, 'Machine learning and the rule of law' in M Livermore and D Rockmore (eds), *Computational Analysis of Law* (Santa Fe Institute Press) (forthcoming).

27 Rigano (n 4);WJ Epps Jr and JM Warren, 'Now Being Deployed in the Field of Law' 59(1) The Judges' Journal 16-39.

28 https://www.washingtonpost.com/technology/2021/04/02/capitol-siege-arrests-technology-fbi-privacy/

29 Sadka, J., Seira, E. and Woodruff, C., 2017. Overconfidence and Settlement: Evidence from Mexican Labor Courts. en. In, p.50.

30 Chen (n 4).

31 C Berdejo and DL Chen, 'Electoral cycles among us courts of appeals judges' (2017) 60(3) The Journal of Law and Economics 479-496.

32 DL Chen, TJ Moskowitz and K Shue, 'Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires' (2016) 131(3) The Quarterly Journal of Economics 1181-1242.

themselves, thereby strengthening the credibility of the judiciary.[33] Moreover, as large backlogs of cases are a significant problem for the efficiency of the judiciary in developing countries, interest is growing around performance metrics that will improve the functioning of the judiciary.

The adoption of ML systems, however, is not an easy-to-implement solution, in particular for the justice system. First, despite the growing availability of judicial data, it is first necessary to process it in a way that is amenable for machine learning. This requires integration of data from different sources, processing of textual data into quantifiable metrics, and the definition of indicators for learning tasks that reflect either performance objectives or operationalizations of concepts of fairness and impartiality. This is not an easy task, requiring substantial investments in data infrastructure and human capital, as well as building a conceptual framework. Therefore, in order to implement ML algorithms successfully, justice systems need to acquire and train a team of machine learning engineers, subject matter experts and legal actors to develop ML algorithms that are operationally relevant. These considerations are similar to the ones highlighted in the broader consideration of public administration.

Finally, there are ethical concerns regarding ML applications on judicial outcomes. Practitioners and citizens may raise question regarding the interpretability of algorithms, since technological sophistication creates a “black box” problem:[34] improvement in technological sophistication makes its operation less interpretable. The challenge of interpretability also raises concerns about the accountability and oversight for these systems. Furthermore, the gap between those who can and cannot access and understand these technologies exacerbates existing social divisions and intensifies polarization. For all these reasons, ML tools should be seen as complements, rather than substitutes for human decision-making, in particular for institutions that make life-altering decisions such as the judiciary.

3.1. Judicial Data Infrastructure for ML

It is increasingly recognized that "the world's most valuable resource is no longer oil, but data".[35] Like oil, raw data is not valuable in and of itself; its value arises when it is cleaned, refined, processed, and connected to other databases that allow for the generation of insights that inform decision-making. This is particularly the case in the field of ML, which requires large amounts of data to build accurate predictive models that provide information on the process, behaviors, and results of any indicators of interest.

Judiciaries collect vast amounts of data daily. Despite the availability of data, judicial data has rarely been analyzed quantitatively. In recent years, the transition from paper trails to e-filing and case management systems has facilitated the systematic analysis of massive amounts of data, generating performance metrics that can be used to evaluate courts and justice actors. Furthermore, with advances on ML, Natural Language Processing (NLP) and processing power, this data creates valuable opportunities to apply ML models to evaluate and improve justice

---

33 K Kannabiran, ‘Judicial meanderings in Patriarchal thickets: Litigating sex discrimination in India’ (2009) 44(44) Economic and Political Weekly 88-98; M Galanter, *Competing equalities: Law and the Backward classes in India* (OUP 1984); P. Bhushan, ‘Misplaced priorities and class bias of the judiciary’ (2009) 44(14) Economic and Political Weekly 32-37.

34 F Pasquale, *The black box society* (HUP 2015).

35 See article [here.]

systems.[36] Nonetheless, the extent to which each of these countries can utilize novel approaches in machine learning and data analytics will depend on the available data infrastructure. The question is then: what data do (and should) judiciaries collect?

In the justice domain, an integrated justice system brings together data of each case and connects this data with information on the actions and decisions of the milestone of each case. For instance, this will include information on the case filing, initial decisions, hearings, rulings, and sentences for each case. This data should also relate to the potential appeal to understand the evolution of the case. By implementing NLP on the text of case filings or judicial decisions, judiciaries can automate the revision of case filings or identify relevant jurisprudence for judges and court actors, for instance. Beyond the information on the justice process itself, judiciaries will gain valuable insights from connecting these data with other information such as Human Resources (HR) data, information on recruitment, or data from judicial training, to understand how best to select, train, and motivate judges depending on their background and experience.

To evaluate the impact of ML interventions in justice, judiciaries will ideally collect information from other agencies involved in the justice process as well as from the economic outcomes of the parties involved in the judicial process. In criminal justice, an interoperable data ecosystem will connect judicial data with data from the prosecutors' office, the police, and the prisons, which will enable them to understand where the case comes from and the implications of judicial sentences. In civil cases, this may include economic data of citizens and firms who participate in the justice system, such as tax data, social insurance data, or procurement data. This way, judiciaries will be able to evaluate the impact of ML applications not only on the judicial process, but also on the lives of citizens and the financial status of firms who use the justice system.

In addition to the external and internal databases, it is also recommended to carry out surveys that complement the administrative information with the experience of the parties and employees involved in the justice system. Administrative data will not capture important elements of user or employee satisfaction, for instance, which is why survey data is a necessary complement to understand the impacts of any new ML models. We recommend also surveying those who are not necessarily part of the justice system through legal needs assessments, who could be potential users themselves in the future.

Finally, there are additional complexities for developing data ecosystems for ML. Data has to be ensured to be of high quality, and large volumes need to be collected and stored in order to make it amenable for AI algorithms that in general presuppose big data. This would require data ETL (Extraction, Transformation and Loading) processes designed to support AI pipelines, and in most use cases dedicated data engineers to maintain it.

3.2. Human Capital Requirements for ML in Justice

Justice systems often have limited in-house access to the necessary human capital to develop machine learning applications. Judicial officers are rarely experts on data analysis – as that is seldom part of their training – and ML engineers generally lack the domain expertise necessary to understand the functioning of the law. In courts without sufficient human capital to take advantage of available data, training bureaucrats to learn even simple data analysis skills may be

[36] Ramos-Maqueda and Chen, 2021.

a valuable long-term investment for improving the functioning of courts. Nevertheless, the development of ML approaches may remain out of reach of public officials whose training does not include statistical modeling or data engineering.

An alternative approach is relying on for non-governmental organizations, international organizations or even private companies to develop ML applications. An example of this approach is COMPAS in the United States, an algorithm that generates recidivism risk scores to aid judges in their ruling decisions. However, outsourcing judiciaries should consider the long-term sustainability of the solution and ethical concerns. COMPAS itself has been the target of controversy due to its proprietary algorithm and inability by public officials and citizens to understand how it operates under the hood.[37] Additionally, reliance on external contractors often substitutes for the in-house development of the necessary human capital to develop ML technologies, reproducing external reliance on non-judicial actors for both maintenance and expansion of ML solutions.

Whether it be in-house or externally sourced, the human capital requirements for ML applications are diverse and costly. The implementation team should include ML engineers, software developers to code the user interface, data engineers to develop the data infrastructure, legal experts and project managers that communicate the judiciary's needs to the implementation team. Because each component of the project often relies on one another – there can be no user interface without a data infrastructure to feed it – the team should ensure that their timelines are aligned. Sufficient budget should be allocated to the project to cover the team's time for both implementation and monitoring of the technology after the first version of the application is developed.

3.3. Ethical concerns

ML applications should carefully consider the ethical implications of their use by judicial actors. Only technologies that aid human decision-making, rather than replace it, should be adopted in the courts. This recommendation is motivated by multiple reasons. As noted earlier, algorithms have the "black-box" problem of interpretability, i.e., it is not easy to trace the output of complex algorithms to the data inputs themselves. Additionally, biases in the decisions of judicial actors are reflected in the algorithm's training data and may be encoded into the algorithm itself. Thus, using ML algorithms to inform judicial decisions without critical oversight raises the risk of replicating these biases elsewhere in the system. The inherent choices of performance metrics can also reinforce existing biases by decision-makers within the system. Addressing these issues requires a participatory and deliberative approach towards the design, implementation, and evaluation of the adoption of ML technologies.

A reasonable demand to guarantee trust and fairness is that algorithms be interpretable. A judge may request a reason for why a particular decision is recommended by the algorithm. This transparency enhances judges' trust in technology, as well as allowing for disagreement with its recommendations. Given the complexities of working with ML algorithms, it is essential that any rollout be preceded by a phase of comprehensive study and rigorous testing of the systems themselves. Randomized controlled trials that carefully estimate the causal impacts of the adoption of these algorithms to properly evaluate their costs and benefits are essential. A carefully

[37] https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/.

constructed trial can provide important benchmarks on cost, efficiency, user satisfaction and impact on key performance metrics, all essential for a justice system to credibly serve citizens.

4. Case study: ML for justice systems in India

This case study illustrates how ML was implemented in the national justice system of India by the Data and Evidence for Justice Reform (DE JURE) team. Due to India's large population and volume of cases, justice officials are often unable to effectively manage cases in a timely fashion. India has just nineteen judges per million people, and twenty-seven million (2.7 crore) pending cases.[38] To address this, the Indian justice system has made considerable advances in adoption of information technology, released large volumes of data to court users and encouraged them to use electronic systems. Yet, legislative, institutional, and resource constraints have limited their full impact.[39]

In this section, we describe how the team engaged in the implementation of ML applications in India. We first highlight the data infrastructure requirements for implementing ML applications, as well as how these applications could enhance the functioning of the justice system.

4.1 Judicial Data Infrastructure in India

In the past fifteen years, considerable efforts have been made to adopt and deploy information technology systems in the courts of India. One of the most significant projects, the e-courts project, was first launched in 2005 by the Supreme Court of India through the "National Policy and Action Plan for Implementation of Information and Communication Technology (ICT) in the Indian Judiciary". The e-courts initiative introduced technology in Indian courts in a variety of ways.

The most obvious feature of the system was the deployment of technology within the court rooms. Judges were provided with LCD touch screen machines, screens and projectors were connected via a local network to disseminate information to lawyers and electronic boards at the courts display the queue of case numbers for hearing scheduled on a particular day. Outside of the courtroom, e-filing procedures were established, and an architecture of data management was created that ranged from scanning of old cases into the electronic system to the creation of digital archives. The ICT plan also established direct electronic communication with litigants and an online case management system.

These investments have eventually paved the way for the creation of the National Judicial Data Grid (NJDG), a database of twenty-seven million cases available to court users to view the status of pending cases and access information on past hearings. For the team's goal to implement ML tools, the most significant resource was the digital archives of cases. We were able to scrape these publicly available digital archives to construct an e-Courts district court dataset of eighty-three

---

[38] VA Kumar, 'Judicial Delays in India: Causes & Remedies' (2012) 4 Journal of Law Policy & Globalization 16; M Chemin, 'Does court speed shape economic activity? Evidence from a court reform in India (2012) 28(3) The Journal of Law, Economics, & Organization 460-485; A Amirapu, 'Justice delayed is growth denied: The effect of slow courts on relationship-specific industries in India' (2020) Economic Development and Cultural Change <https://doi.org/10.1086/711171> accessed on 16 August 2021.

[39] Amirapu (n 9); D Damle and T Anand, 'Problems with the e-Courts data' (2020) National Institute of Public Finance and Policy Working Paper 314 <https://www.nipfp.org.in/media/medialibrary/2020/07/WP_314__2020.pdf> accessed 16 August 2021.

million cases from 3289 court establishments.[40] We were able to curate details like the act under which the case is filed, the case type (criminal or civil), district where it originates, the parties to the case, and the history of case hearings, in a manner that makes the data amenable to large-scale analysis.

A wider data ecosystem was created by joining additional sources to the case data:

1. **Data on Judges:** In order to better understand the impact of specific judges — their identity, training and experience — the team constructed a database of judges for the courts of India. We have begun this task by extracting data from the Judges Handbooks, released by the Supreme Court of India, and appending to it information from various High Court websites. Thus far, we have assembled details of 2,239 judges from the handbooks for the years between 2014 and 2020. Most notably, 93.5% of these judges are males and 6.5% are females and their range of experience covers a period spanning approximately seventy years.

2. **Database of Central Acts:** This auxiliary dataset is intended to give a definitive list of standardised act names. This could then be used to standardize the act names appearing in the various cases. This allows us to analyze all cases filed under a given act. We have, for example, examined all cases related to the Water Act of 1974 and found a total of 978 such cases at the Supreme Court and High Courts of India. The list of central (federal) acts can be viewed on the Legislative Department website of the Ministry of Law and Justice. There is currently no centralized source for all state legislation — this needs to be obtained from the state websites separately.

3. **Other Administrative Data:** Data on other institutions can be linked to the judicial data at the district as well as the state level. For example, data on Indian banks and their branches is available through the Reserve Bank of India. This database contains information on their name, location, license number, license date, address, and other unique identifiers. We have scraped, cleaned and organized this data for further analysis. It contains about 160,000 entries. The unique identifiers and location information allow us to merge this data with banks appearing in litigation in courts that are present in the e-Courts databases. The merging of this data with the legal data allows us to examine a variety of interesting questions about the development of financial markets in an area, participation in the justice system, and the impacts of legal rulings.

The quality of the data varies significantly — there is no nationally standardized system for defining variables or reporting on them. For instance, in some states the legal act name and section numbers are well delineated, but in other states this is not the case. This makes it difficult to compare individual case-types across courts and across states.[41] There are no standardized identifiers within the data to follow a case throughout its potential sequence of appeals in higher courts. In a similar vein, there is no easy way to track a criminal case from its entry into the system as a FIR to its exit as a judgment. There are inconsistencies in identifying information about

[40] The eCourts data is public and can be accessed via the district court websites, the eCourts Android/iOS app, or the district court services webpage <https://services.ecourts.gov.in/ecourtindia_v6/>.

[41] Damle and Anand (n 10).

participants, their attributes and the types of laws or acts that the case relates to. There are also issues of incorrect reporting and spelling mistakes.

4.2. ML Applications in the Courts of India

The quality of data in India's justice system is often compromised: case data is incomplete or litigant's identities not registered. The DE JURE team has constructed a robust data pipeline to collect often incomplete judicial data, as well as ML tools to clean and prepare it for analysis. In this section we contextualize the problem: how data quality issues in judicial data manifest themselves in India. In the following section, we describe the solution: how ML tools have been designed to enhance the quality of judicial data for analysis.

Legal data released by the Indian judiciary is voluminous, messy, and complex.[42] The typical case has clear tags for some key dates (filing date), key actors (petitioner, respondent, and judges) and court name, but information about the type of case, outcome of the deliberations, and pertinent acts cited are often not clearly identifiable in the textual body of the judgment. Cleaning and pre-processing the data is critical for any form of analysis, especially so for supervised algorithms trained on this data. Traditional empirical legal studies have typically addressed this issue by relying on small-scale data sets, where legal variables are manually coded, and the scope of inference is related to a small body of legal cases that is pertinent to a single issue.[43]

These traditional approaches are unable to keep up with the incoming volume of cases. In this context, ML tools provide an alternative approach to detecting errors or gaps in the data, and correcting for them in an automated fashion. Using ML, it is possible to infer identities of participants even when that data is not registered. Additionally, laws used as precedents for the ruling can be identified through text analysis. Beyond data quality itself, ML approaches can help identify biases and discrimination in judge's rulings.

4.2.1 Inference about the identity of participants

Some databases of judgements provide no identifying information on the participants in the cases themselves. To better understand who participates in the courts, we first extract litigant names from the raw text of the judgements, and then using matching algorithms, identify the type of litigant (individuals, companies, or state institutions). Classifying participants can be challenging. If the identification exercise involves government agencies, it is first necessary to compile all the different agencies of the state government and national government. Manually tagging entities is prohibitively time consuming, but the existence of latent patterns in the names makes this fertile ground for ML applications.

---

[42] Damle and Anand (n 10).

[43] V Gauri, 'Public interest litigation in India: Overreaching or Underachieving?' (2009) World Bank Policy Research Working Paper 5109 <https://poseidon01.ssrn.com/delivery.php?ID=709021124002094091083101097022125125019054034003088001030009059006043060039039097126091016105064067026031050057103005023124026030004026113067029027097007105125022065069083094082097017013024&EXT=pdf&INDEX=TRUE> accessed 18 August 2021; S Krishnaswamy, S K Sivakumar and S Bail, 'Legal and judicial reform in India: A Call for Systemic and Empirical Approaches' (2014) 2(1) Journal of National Law University Delhi 1-25; U Baxi, *Towards a sociology of Indian law* (1st edn, New Delhi: Satvahan 1986).

The ML application relies on similarity across names for participants that belong to the same "group" to classify a particular name as belonging to that group. Using pre-labeled data – individual name A belongs to group B – the ML algorithm can extrapolate to non-labeled data, where only individual's B name is available, but not the group itself. Some obvious groups of interest are a participant's gender, caste, and religion, which are not recorded in judicial data, but available in other data sources. Another group of interest may be whether a participant is a government agency or not. We have focused here on people's first and last names, for illustration.

We first format individual names to ensure that each individual could be identified by an honorific title, a first name, and a last name. Honorifics such as Shri, Sri, Smt., Mr., Mrs., and Ms. Enables the algorithm to directly identify gender. To extend this classification to names without an honorific, we train an algorithm on a publicly available corpus of labeled common Indian first names. Training this algorithm, often referred to as training a classifier, is the process of learning patterns within the data related to the group. Here, these patterns are the statistics of co-occurrence of alphabets in names, length of the name, and other features which allows the algorithm to determine whether a name is indeed of that particular group, in this case a gender.[44]

These algorithms formalize intuitive notions of why a name belongs to a given group by identifying frequently occurring patterns within names associated to that group. Caste assignment is more complicated because the same last name can be associated with multiple caste groups. The name Kumar, for example, could be the name of a person belonging to scheduled cast, schedules tribe, or 'Other' category. In the case of such names, we generate the distributions of the last name across the different caste categories. We use this distribution to generate a prediction and then combine this with predictions of other models to ensure a robust prediction. We assign a caste to each household based on a simple majority vote between these models.

4.2.2. Identification of Laws and Acts

Legal texts in India's justice system do not currently employ standardized citation style for referring to acts or laws. For example, the Hindu Marriage act may be referred to in a variety of ways: "u/s 13 clause 2 of Hindu Marriage Act," "u/s 13(b) Hindu Marriage Act," or "u/s 13 of Hindu Marriage Act 1995". Again, ML tools can also be used to address this issue.

In this project, the team is using a set of tools that create mathematical representations of the text in the form of vectors. "Term Frequency - Inverse Document Frequency" (TF-IDF) is one such popular method to represent a string of words as a numerical score that reflects how frequently a word is used within a given text and how infrequently it appears in the corpus. Applying this to act names, we use different clustering algorithms to group together particular act citations, based on how similar they are numerically. This approach groups the underlying act names data in a manner

---

[44] To reduce model overfitting, we use the majority vote from multiple trained classifiers, including a logistic regression model and a random forest classifier to make predictions on gender. A logistic regression models the probability of a binary outcome or event. A random forest classifier will use decision trees (nested if-then statements) on features of the data to make the prediction. We have also made predictions of religion and caste using similar approaches. Muslims can be recognized in the data through the distinctiveness of Muslim names: common names such as Khan and Ahmed can easily be assigned and coded, but for others we utilise the occurrence of specific alphabets (such as Q and Z) through appropriate classifiers, to identify additional names.

that best preserves the coherence within groups (a particular act name) and the distance across groups to make the classification.

Identification of specific acts and how often they are cited opens new opportunities of legal analysis. We can, for example, compare the different types of acts that are cited in different courts of India's justice system. It can allow researchers and practitioners to identify the real time evolution of legal citation – and legal thought – as judges refer to these acts.

4.2.3. Using descriptive analysis and machine learning to identify discrimination and bias

Concerns about stereotyping and discrimination in the courts of India are widespread, and these forms of bias can influence case outcomes and undermine the rule of law. Bias shown by a court, or even a single judge, is difficult to identify and analyze rigorously. This challenge is of course not unique to judiciaries. Many papers in the academic literature have demonstrated that bias by a human decision-maker can have conscious as well as unconscious drivers and may manifest in complex ways than can be difficult to prove in a variety of contexts.[45] In other settings, such as labor markets and educational institutions, algorithms — rules that take "inputs" (like the characteristics of a job applicant) and predict some outcome (like a person's salary) — have been shown to create new forms of transparency and serve as methods to detect discrimination.[46] In the courts of India, algorithms could help judges make some critical decisions about cases (for example, dismissals or bail applications) and reduce bias in their rulings.

Building such algorithms requires a rich dataset typically consisting of variables that include litigant characteristics (caste and gender), lawyer characteristics, court characteristics, case details (filing details and evidence provided), additional variables (time of the event) and case outcomes (such as granting of bail or dismissal of a case). A ML engineer would develop a "learning procedure" that would aim to provide a predicted outcome from a broad range of inputs and modeling approaches. In neural networks, multiple sequential layers of intermediate steps connect input features and output classes, so that the outputs of one layer serve as inputs of the next layer.[47] These models stand in sharp contrast to traditional statistical methods such as linear regression, which is more deductive (presuming a linear fit between a few sets of variables) than inductive (allowing the data to report the best fit between a large set of variables).

These insights could be invaluable not only within the court room itself, but also in judicial education. Experiments are currently underway, in the Judicial Academy of Peru for example, to assess methods to improve case-based teaching by using the history of a judge's past decisions — which can reveal potential bias or error.[48] The data is also suitable for creating personalized dashboards and interfaces that provide judges, mediators and other decision-makers with real-

---

45 A Banerjee and others, 'Labor market discrimination in Delhi: Evidence from a field experiment' (2009) 37(1) Journal of Comparative Economics 14-27; M Bertrand and S Mullainathan, 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination' (2004) 94(4) American Economic Review 991-1013; M Ewens, B Tomlin and LC Wang, 'Statistical discrimination or prejudice? A large sample field experiment' (2014) 96(1) The Review of Economics and Statistics 119-134; J Kleinberg and others, 'Discrimination in the Age of Algorithms' (2018) 10 Journal of Legal Analysis 113.

46 J Kleinberg and others, 'Algorithms as discrimination detectors' (2020) 117(48) Proceedings of the National Academy of Sciences 30096; Y LeCun, Y Bengio and G Hinton, 'Deep learning' (2015) 521(7553) Nature 436-444.

47 Dayhoff, J.E., 1990. *Neural network architectures: an introduction*. Van Nostrand Reinhold Co..

48 JR Kling, 'Incarceration length, employment, and earnings' (2006) 96(3) American Economic Review 863-876.

time information about their own performance relative to their own previous decisions and others who are comparable.[49] This information can be used to augment the capabilities of judges and lawyers, increase their productivity and reducing potential biases in their decisions.

5. Beyond Descriptive Analysis: Impact Evaluation in the Justice System

Moving beyond descriptive analysis and more correlational analysis of data, an underexplored field in the justice system is policy experiments for impact evaluation. Legal scholars and judges have long debated the merits of implementing of various laws and regulations and justified their arguments with theories about the effects of these legal rules. This situation resembles the field of medicine a century ago — prior to clinical trials, medical research focused on theoretical debates rather than rigorous causal evidence.

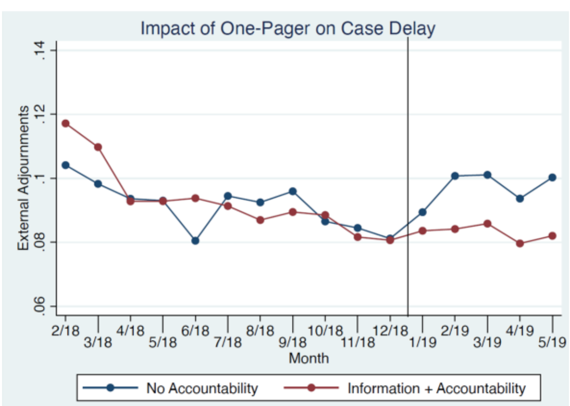
A growing body of empirical research now demonstrates that causal inference is possible in judicial studies. For example, in situations where cases are randomly assigned to judges, the random assignment itself can be used as an exogenous source of variation to evaluate the impact of judicial decisions. Since judges do not choose their cases and end up with them randomly, observed rulings reflect personal characteristics of the judge (ideological preferences, biases) and the case rather than the judicial process as a whole. Additionally, informational treatments can have an impact on the behavior of judges, improving their performance (Box 2).

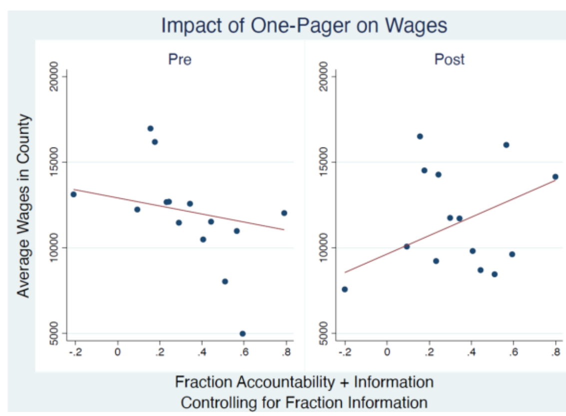
49 Ibid.

Box 2: Leveraging data and technology for a more timely resolution of justice

In partnership with the Kenyan Judiciary and McGill University, the World Bank's DIME Data and Evidence for Justice Reform (DE JURE) team has been leveraging underutilized administrative data to improve judicial efficiency. Through its case management system, the Kenyan Judiciary collects large amounts of administrative data on the characteristics of the case, the dates of hearings and reasons for adjournments, and other important metrics of court performance. This data is readily available to understand and design interventions to address challenges to efficient delivery of justice, such as adjournments of hearings, which cause large delays in court proceedings. Despite the richness of this data, it was not being used for decision-making. Thus, the partnership between DIME and the Kenyan judiciary decided to leverage these data systems to design an algorithm that identifies the greatest sources of court delay for each court, and present recommended actions. The team included such performance information in a one-page feedback report. Then it studied whether this simplified, action-oriented information could reduce adjournments and improve judicial performance.

In a randomized controlled trial across all 124 court stations in Kenya, the team compared the impact of only sharing the one-page feedback reports with judges and supervisors to sharing this report with Court User Committees as well, the latter of which acted as an additional accountability mechanism. The team found that the one-page feedback report with the accountability mechanism reduced the number of adjournments by 20% over a 4-month period, and increased the number of cases resolved. The conclusion was that the report is more effective when both tribunals and Court User Committees receive it. Thus, sharing performance information with courts may be effective to improve efficiency, but it is particularly effective when this information is also shared with civil society and court stakeholders. This study served as proof of concept about how utilizing data to provide information to judicial actors can reduce adjournments and increase speed of justice, which has a downstream impact on the economic outcomes of citizens and firms.





Randomizing cases to judges predicted to be harsh or lenient allows researchers to identify long-run causal impacts of the length of sentences.[50] To identify the causal effect of a sentence length of eight months or eight years, a randomized control trial would need to randomize the sentence,

[50] W Dobbie and J Song, 'Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection' (2015) 105(3) American Economic Review 1272-1311.

which is impossible. However, assigning a defendant to a judge predicted to assign eight months or another judge predicted to assign eight years sentence length allows us to identify the causal impact of sentence length on subsequent life outcomes.

The same conceptual framework can examine the causal effects of debt relief on individuals' earnings, employment, and mortality.[51] This causal approach sheds light on the impact that these judicial decisions can have an impact on individual's welfare outcomes. By applying machine learning to infer biases, lenience and ideological preferences of judges, researchers can identify the causal effect of these variables on the judicial system and life outcomes of those affected by these judge decisions.

6. Conclusion

In this chapter, we argue that ML is a powerful tool for improving public administration more broadly, and the justice system in particular. ML, at its core, emphasizes a methodological approach centered around a learning problem: defining indicators and using evidence to improve them. Under the umbrella of this methodological approach, multiple applications are available to tackle key issues in public administration. Algorithms can be written to draw inferences about the identity of participants and study the deliberative processes they employ within court rooms. ML tools can also convert a high volume of textual data to numerical estimates that can be used for understanding the processes and outcomes of different types of case data, including public procurement, taxes, and the systems of justice themselves.

These tools, however, have several limitations and requirements that need to be addressed before they can be effectively deployed in the courts. At the very outset, there are significant issues related to the privacy related to personally identifiable information, security, and control of legal data. Next, the algorithms require data pre-processing, training on large and high-frequency datasets, and iterative refinement with respect to the actual use-cases where they are deployed. This requires strong pilot programs that are studied as part of randomized control trials (RCTs). Insights on data privacy, costs as well as outcomes require these pilots to be constructed on a reasonable scale.

Public administration officials often execute a range of tasks, from the ordinary to complex, such as adjudication of a trial. The smart application of ML can enhance both levels of automation, productivity and the level of information to be extracted from the data generated in the public sector. If done right, it can help reduce noise, and this can be one step in aiding the impersonal execution of tasks, reducing bias, enhancing predictability and making decision rules more transparent. But none of this is presupposed from the ML approach: it depends on the ethical framework and operational relevance underlying its implementation. ML practitioners are advised to therefore take necessary precautions and develop solutions which are accountable to the public and useful for government officials.

---

51 B Sampat and HL Williams, 'How do patents affect follow-on innovation? Evidence from the human genome' (2019) 109 (1) American Economic Review 203-236.

References:

Bandiera, O., Best, M.C., Khan, A.Q. and Prat, A., 2021. The allocation of authority in organizations: A field experiment with bureaucrats. *The Quarterly Journal of Economics*, *136*(4), pp.2195-2242.

A Banerjee, et al. ‘Labor market discrimination in Delhi: Evidence from a field experiment’ (2009) 37(1) Journal of Comparative Economics 14-27

M Bertrand and S Mullainathan, ‘Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination’ (2004) 94(4) American Economic Review 991-1013

M Ewens, B Tomlin and LC Wang, ‘Statistical discrimination or prejudice? A large sample field experiment’ (2014) 96(1) The Review of Economics and Statistics 119-134

J Kleinberg et al., ‘Discrimination in the Age of Algorithms’ (2018) 10 Journal of Legal Analysis 113.

J Kleinberg and others, ‘Algorithms as discrimination detectors’ (2020) 117(48) Proceedings of the National Academy of Sciences 30096

Y LeCun, Y Bengio and G Hinton, ‘Deep learning’ (2015) 521(7553) Nature 436-444.

Dayhoff, J.E., 1990. *Neural network architectures: an introduction*. Van Nostrand Reinhold Co..

JR Kling, ‘Incarceration length, employment, and earnings’ (2006) 96(3) American Economic Review 863-876.

W Dobbie and J Song, ‘Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection’ (2015) 105(3) American Economic Review 1272-1311.