NYU Center for Data Science and Data Science Justice Collaboratory

Building an Unbiased Robot Judge

Elliott Ash (eash@nyu.edu), ETH Zurich, Chair of Law, Economics, and Data Science
Daniel L. Chen (dlc16@nyu.edu), Institute for Advanced Study at Toulouse, Harvard, NBER

http://users.nber.org/~dlchen/#data-science-justice-collaboratory
http://users.nber.org/~dlchen/teaching/Projects2019.pdf
DS-GA 1012

1 Introduction

We offer a set of projects using AI to learn legal reasoning from a diverse corpus. Our research group is developing the most extensive, comprehensive legal corpora ever produced for academic research. There are a variety of legal datasets that comprise in total 12 TB for the US, India, and Kenya. Nine projects from previous years have resulted in NIPS workshop selections or peer-reviewed publications. Previous students have gone on to top PhD programs at Harvard, MIT, Stanford, and NYU and formed startups like Hunch and Zagaran.

Annotated and linked datasets include the universe of U.S. cases in the Supreme, Circuit, District, and state courts, text-aligned Supreme Court oral arguments, asylum decisions, linked data from arrest to final sentence from a district attorney's office, World War I British military tribunals, and bankruptcy filings. India case records include its 24 High Courts and 3000 subordinate courts with details on over 8.7 million case records and 67 million hearings and the universe of text since before independence.

We will have weekly meetings and share a high performance computing environment. A challenge will be to formulate any of the problems as deep neural net applications, including convolutional NNs, recurrent NNs, and autoencoders. For example, use a convolutional neural network to discover feature mappings for significantly predictive legal language features, or a recurrent neural network to predict legal language sequences given seed language, or an autoencoder to optimally reduce dimensionality of judicial decisions. Jurisprudence of common law judges can also analogize to information retrieval and search on a network of opinions.

An excellent class project assumes a high level of commitment, resulting in journal article publication. A good class project may result in conference submission and publication, e.g., at NIPS, and continuation for Capstone project. Best practices in reproducibility are required (end-to-end code and intermediate data).

2 Natural Language Understanding

Projects more closely related to DS-GA 1012 include projects to:

- Recover logic and dependency from court opinions
- Explain what the rule means, answer questions about cases, automatically summarize cases
- Infer judge's intent from searching for legal precedent (analogize from web or social media search)
- Entity recognition of legal doctrine
- Learn from human-annotated entity summaries or co-occurence to generate a topology of precedent or legal knowledge graph (analogize to learning from wikipedia)
- Automated discovery of the emergence of new legal issues
- Generate a page-rank of judicial influence via the citations and generation of agreement or dissent (on 3-judge panels)
- Automated impact analysis of judicial writing, framing of issues, legal reasoning, narrative generation, surprise and serendipity, creativity (analogize from computational journalism)
- Use law review articles commenting on opinions (analogize from newspaper comment sections)
- Study longitudinal queries of judges to legal precedent
- Model judicial learning (exploration vs. exploitation vs. bias)
- Automated discovery of opinion authorship (in unsigned opinions or judge vs. clerks who transit from Circuit to Supreme Court)
- Mood detection among judges
- Identify parts of opinions that spark cites, for identifying sections that lead to polarized cites, and for thinking of citation recommendation to judges.

3 Debiasing Judicial Decisions

1. Reasoning: ***The Fact-Value Distinction*** is widely considered a source of conflict between science and ethics – the distinction between what can be known to be true and the personal preferences of individuals. An important step for automated moral reasoning is the ability to make this distinction. Court opinions have been previously annotated to distinguish between facts and legal reasoning of a case. We want to train a model to classify text as fact or reasoning. This will be useful for asking whether a decision follows reasoning, or instead judges use reasoning as an ex post rationalization

of a subconscious decision (motivated reasoning). Do differences in fact descriptions explain final decisions? Do judges distort fact descriptions? Can the model be used to ask if legal areas be classified as objective or subjective? Are citation patterns reflective of these differences?

2. Reasoning: Universal grammar is the theory of an innate component of the language faculty, independent of sensory experience. ***The Grammar of Law Project*:** a) exploits parallel multilingual legal databases to identify equivalent legal phrases (to identify the 'molecules' of law); b) uses these molecules to automate the detection of legal inconsistencies; c) uses the molecules, and sentiment/treatment, to automate encoding of moral views. Potential approach is a convolutional neural network to recover feature mappings that are predictive of decisions, and feature mappings that have an impact on higher-court decisions. Is there an optimal deviation in legal consistency? How much legal innovation engenders subsequent (positive) citations? Is there a trade-off between innovation and reversal likelihood? Does the legal text or the structure of its citations predict subsequent treatment in terms of importance (citations), controversy (dissent), and mistake (reversal)?

3. Reasoning: For powerful AI systems to help in domains where human values are ambiguous or underdetermined, it is valuable to formalize what is ambiguous or uncertain. ***The Legal Ambiguity Project*** provides a context to study the use of certainty: a more certain opinion states a clearer policy position, which makes it more attractive to cite by future judges but also more likely to be reversed. A first step is constructing metrics for legal certainty/ambiguity, for example by use of certainty words. ("High-certainty" Supreme Court opinions have been found to be cited more often.) An application is to study the polarizing effect of elections on judge certainty. Other usage scenarios include studying what policy areas have relatively more ambiguous statutory terms over time. For example, consensus issues like highway maintenance might converge on low-ambiguity clauses, while politicized/divisive issues like guns and abortion might sustain high levels of clause ambiguity.

4. Reasoning: Since Dawkin's The Selfish Gene, it has been proposed that ideas are memes that can be analogized to genes as units of replication and propagation. ***The Genealogy of Ideology Project*:** a) predicts the memetic *phrases and citations* that are likely to be passed along the network in forward citation, but do not otherwise appear in a distant case in the citation graph; b) detects propagation (peer effects) when the underlying network is partially observed (for example, in oral arguments); c) identifies influential ideas and thought leaders; d) predicts how the judge votes on the next panel, using only the history of who the judges previously sat with on panels and how the judges' votes aligned with the panelists. We use random assignment of judges to make causal inferences. Then one can test population biology theories like whether greater volatility (be they environmental or economic) in a jurisdiction leads to more legal innovation, more generation of memes, and more citations.

5. Reasoning: *Measuring Moral Reasoning Project* probes classic divides in moral philosophy and in economics and law. To understand human values, AI systems will likely need to be able to detect and annotate when an argument is utilitarian or deontological. One approach might be to classify assignments of obligations and authority: e.g., "the right is vested in party 1", "party 1 has the right", "the duty is assigned to party 1", "party 1 must", etc., or to classify conditional language: "if A, then B" constructions. Another could be to observe whether economics-trained judges use different features of defendants consistent with cost-benefit analysis when obtaining sentences and when writing more generally. Consider building a classifier for political speech seeking to expand/restrict policy and use this on the congressional record floor speech and court opinions to see, for example, if expansions are utilitarian and contractions are deontological.

6. Reasoning: Distinguishing **analogical or logical** modes of reasoning is useful for AI systems. Use this on several hundred years of Confucian examinations or analogically-linked court opinions (arguing from precedent). Then one can model different fields of science or law and build a model of comparative advantage and trade flows.

7. Fairness: Phonology is the study of speech variation beyond word choice, that is, fluctuations in the way one speaks holding the words fixed. *The* ***Vocal Implicit Bias*** *Project* investigates if judge's vocal intonations reflect implicit bias. Using 15 years of Supreme Court oral arguments, it appears that vocal intonation of gendered words (e.g., actor vs. actress) classify vocal intonations of neutral words into stereotypically male (e.g., logical, ability, think) and female (e.g., looking, cook, goodwill). Documenting this in a rarefied setting like the Supreme Court is surprising, and suggests the relevance of people's perceptions of gender being revealed in how people speak. Other validation possibilities are audio data from police stops of motorists. Then one can assess and compare these implicit features with explicit measures such as interruptions in explaining disparities in outcomes in the Supreme Court or police stops.

8. Reasoning: ***The Judge Embeddings Project*** **constructs** inputs for end-to-end machine learning estimates of the causal impacts of law is a widely sought-after goal, whether to simply score judges on an ideological multi-dimensional spectrum or to use in high-dimensional instrumental variables impact analysis of court decisions. Related goal is to analyze the legal text as a high-dimensional treatment for causal analysis and to see what parts of the opinion impact outcomes. Can we score judges on statutory interpretation, like textualism or originalism, critical legal studies, etc.?

9. Reasoning: Whether AI systems can learn philosophical reasoning from a diverse corpus of human text and human dialogue includes querying whether the manner of speech matters. *The Vocal Convergence Project* starts by **predicting ideology from audio** holding fixed the words spoken and a speaker's demographic characteristics. Several studies have documented that short-term convergence in speech (and audio) is predictive of votes and that judicial ideology displays longitudinal

movements over time - is this reflected in their audio beyond the text? How much does audio aid in predictive accuracy relative to a baseline model using text alone? Examine whether the model fits political ideology of televised political speech, sermons, Buddhist talks, etc.

10. Reasoning: *The Demography of Judging Project* considers life tenure of judges and the relationship between aging, health, and output. Aging judges appear to become more lenient in criminal and asylum courts and use simpler language. Predict early death or dementia, or simply retirement, using judicial corpora. Does health covary with judicial predictability or unpredictability?

11. Fairness: ***The Prosecutors Project* examines what prosecutors maximize**. Do they seek to minimize recidivism, maximize conviction rates, maximize sentence lengths, or minimize time until trial completion? How do prosecutors compare to a predicted prosecutor? Developed further, the project could have several important policy implications: it could suggest ways of alleviating criminal caseloads without increasing crime rates; it might identify defendant characteristics that are 'noisy' to prosecutors; and it might provide important insights into how a prosecutor's background relates to the quality and nature of their charging decisions.

4 Impacts of Judicial Text

Predictive judicial analytics holds the promise of increasing efficiency of law. While much theoretical work in economics predict the consequences of law, the advent of machine learning tools can investigate these theoretical claims. The ***Causal Impacts of Court Decisions Project*** uses an empirical framework where predictive analytics is used in the first step of causal inference, where the features employed are exogenous to the case. The project will apply as baseline a method for two-stage least squares estimation ("Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica (80(6), 2369-2429, 2012). We cannot randomize judicial decisions, since doing so would undermine the notion of justice and equal treatment before the law, but judges are randomly assigned and there is substantial variation in how they decide–their habits or legal philosophies. The usage scenario is rapid causal analytics: automatically identify the nearest previous cases when a District Court case comes up to the Circuit Court, fast-decision classification of the prior cases' directionalities, consider the cases as a site of high-dimensional causal treatments to be reduced to a low-dimensional representation of the dicta, reasoning, and citations, and generate deep predictions from judicial corpora of previous decisions. A platform has been developed that employs a machine learning first stage using features that are exogenous (randomly assigned) to simulate a retrospective clinical trial. The second step analyzes the correlation between the predicted court decision and society-wide outcomes. A variety of policy questions are available for analysis in tax, crime, immigration, environment, labor, civil rights, and societal attitudes. Previous examples include sexual harassment, government takings, eminent domain, first amendment, abortion, religious freedom, and piercing the corporate veil.

5 Judge Vectors: Topography of Ideology

Recent work in computational linguistics has made breakthroughs in vector representations of language. The success of Word2Vec is that it "learns" the conceptual relations between words; a trained model can produce synonyms, antonyms, and analogies for any given word. *The Judge Embeddings Project* understands the relations between rulings, and between judges, using vector algebra, which successfully recovers the Circuits, the decade, the legal topic, and the judge's birth cohort, but less so of political party and law school. Vector representations recover judges employing economic concepts and those most similar to Richard Posner more successfully than cosine similarity to economics articles. The project develops a range of potential applications: isolate directions for legal and political concepts, liberal vs. conservative, procedural vs. substantive, originalist vs. pragmatists, constitutional applications of principles; citation embeddings to identify substitutable and complementary cases based on how often they are cited together; understand differences across judges in sentiment toward policies or social groups; judge embeddings based on their predictiveness for case outcomes or downstream economic outcomes, rather than just the language features. This model could then be used to simulate counterfactuals. For example, how would the decision in a case change by switching out the authoring judge? How would the style of language change for a different circuit? This will give a topography of ideology in the US judiciary.

6 Debiasing Judges: AI and Rule of Law

Predictive judicial analytics holds the promise of increasing fairness of law. While much empirical work evaluates judges to observe inconsistencies in their behavior, the advent of machine learning tools offers an approach to automate the detection of inconsistencies and, more novel, difference-in-indifference of judges. *Automated Debiasing of Judges Project* applies a theoretical framework to understand a large set of behavioral findings on judicial decision-making. First, settings where judges are closer to indifference among options are more likely to lead to detectable effects of behavioral biases. Second, inter-judge disparities in prediction accuracy can reveal indifference of judges to the circumstances of cases. Third, algorithms can identify and reduce difference in judicial indifference. Fourth, implicit bias in judicial opinions and oral arguments can be predictive of disparities. Applications include asylum judges, criminal justice, and textual and audio data. A conceptual distinction between inter-judge disparities in predictions and inter-judge disparities in prediction accuracy is articulated as another potentially normatively relevant criteria with regards to fairness. Machine learning may help detect due process violations.

7 Law and Development

The Law and Development Project studies (1) impact of court functioning on economic growth and inequality, (2) impact of economics, political, or psychological factors on court outcomes, (3) impact of court decisions or precedents on individuals' outcomes, (4) artificial intelligence applications, and (5) field experiments (such as personalized nudges) to improve efficiency and fairness of law. Other applications include applications to India and Kenyan courts of US findings on motivated reasoning, economic reasoning, implicit bias, extraneous influence, early predictability, difference in judicial indifference, algorithms as juridical decision-makers, memes, fact discretion, reversal prediction, or personalized nudges of judges.

8 Miscellaneous

- Examine schools of legal thought (originalism, critical legal studies, as well as law and economics) in the judiciary.
- Investigate particular qualitative hypotheses, like whether "mental health" (in capital cases, liberal v. conservative) is a driver of leniency, and see how the spread of thought on institutionalization affected imprisonment.
- Study the discourse around the idea of terrorism; 9/11 caused asylum cases to crowd several court dockets, did it also affect discourse?
- Compute citation embeddings using the shopping cart model
- Predict nature of suit in district/circuit cases based on available labels
- Predict affirm/reverse in state and federal
- Automated topic labeling using document embeddings
- Predict citations (number or treatment) in the circuit case based on the district court case
- Compute Bayesian Surprise in legal documents
- Study cultural evolution in legal documents
- Measure innovation in laws (courts and/or statutes)
- Do judge embeddings with neural nets
- Classify prospective versus retrospective language. Or "maintain status quo" versus "change from status quo" language. Then use this to analyze political speeches and newspaper op-eds. Under what economic conditions do people use prospective vs retrospective language?

- Identify language that expands scope versus restricts scope
- Train Bayesian skip gram embeddings on legal corpus
- Effect of free speech laws or copyright laws on number of books published (and richness of language).
- Do prison openings suddenly lead to sentencing decisions changing, and the court opinion reasoning changing.
- Identify Plaintiff, Defendant, Appellant, Appellee throughout cases
- Identify Circuit splits (two cases in different Circuits that has reached opposing conclusions on the law for the same set of facts)

9 Datasets

The data are extremely expensive in terms of money, time, effort, and resources. More than 10 years have been invested in these datasets, which comprise in total 12 terabytes (cloud computing services and dropbox access provided). The data are confidential and require a non-disclosure agreement for access.

1. U.S. Circuit Courts
   (a) Digitized universe since 1891 of all 380,000 cases, 1 million judge votes, across 94 hand-labeled legal topics, engineered into 2 billion N-grams of length eight, and 5 million citation edges across cases.
   (b) This is merged with the 268 judges who served during this time period, 250 biographical features, a 5% random sample with 400 hand-labeled features (330-paged codebook), and 6000 cases handcoded for meaning in 25 polarized legal areas.
   (c) Also merged with administrative data (date of key milestones, e.g., oral arguments, when was the last brief filed, etc.), publicly available U.S. Supreme Court datasets, U.S. District Court datasets, geocoded judge seats, biographies of judicial clerks, and oral arguments' audio files.
   (d) The identities of randomly assigned judges sitting on 3-judge panels (who is authoring the opinions, writing dissents, or writing concurrences) render a random seating network among the judges.
   (e) 25 polarized legal areas have in addition been collected and hand-coded: sexual harassment, eminent domain, free speech, abortion, church-state separation, affirmative action, gay rights, disability rights, campaign finance, capital punishment, criminal appeals, desegregation, sex discrimination, punitive damages, federalism, National Labor Review Board, environmental protection, National Environmental Policy Act, Federal Communications Commission, Title VII, First Amendment, Eleventh Amendment, standing, contracts, and corporate veil piercing.

2. U.S. District Courts

   (a) Digitized universe of millions of criminal sentencing decisions across 94 U.S. District Courts from 1991 (with randomly assigned judges), hand-labeled biographical data of judges, and 83-paged codebook from the U.S. Sentencing Commission.

   (b) Linkages to judge identity were obtained (not publicly available) and hand-labeled biographical data of judges incorporated. Data linkages have been made to daily weather and local sporting events.

   (c) Text of opinions are available since 1923.

3. U.S. Supreme Court

   (a) Digitized speech patterns in oral arguments since 1955– longitudinal data on speech intonation (linguistic turns) are rare.

   (b) Linked to hand-labeled oral advocates' biographies, lawyers' faces, clipped identical introductory sentences*, ratings of their traits, and publicly available U.S. Supreme Court databases containing dozens of additional features and preceding U.S. Circuit Court data.

   (c) *Lawyers always use the exact same sentence when they introduce themselves to the Supreme Court: "Mr. Chief Justice, (and) may it please the Court." We have clipped this data for 1955-2013 comprising over 8000 audio recordings, spoken by many different lawyers over time. Mechanical turk workers have rated 1999-2013 sample (2000+ recordings) based on whether they sound "confident", "trustworthy", "attractive", "masculine", etc. We also have data on the Mturk workers, the Supreme Court cases, and the Supreme Court oral advocates (including their faces).

   (d) Actual analysis of speech patterns is statistically challenging, since speech is modified dynamically. A common measure for variation in speech patterns considers resonances of vowel sounds. In order to properly measure these, the starting locations of all distinct vowel sounds have been manually flagged. An algorithm then measured vowel resonances and assigned to each vowel sound a multidimensional continuous quantity. Therefore, the size of the oral data set is much larger than the size of the underlying text.

   (e) Text is traditionally treated with discrete models. Speech measurements (for example resonances) by contrast are continuous.

4. U.S. State Supreme Courts

   (a) Digitized universe for 1947-1994 (roughly 400,000 cases), identities of judges sitting on the panels, hand-labeled biographies, citation network, and original text.

   (b) Some of these judges run for election.

5. EOIR Immigration Courts

   (a) Digitized universe of administrative data on 1 million refugee asylum and 15 million hearing sessions and their time of day across 50 courthouses and 20 years (with randomly assigned judges), hand-labeled biographical data of judges, and dozens of features on the case and the defendant.

   (b) We know when the asylum case was assigned, whether the hearing was an individual hearing or whether multiple individuals were scheduled in the same session, how many cases were scheduled for sessions during a day for that judge, whether this was an in person hearing or by audio or video, whether it was a written or oral order, whether there are other related applications for relief filed by the individual and the judge's ruling on each, ethnicity of the applicant, the reason for the case and the judge.

   (c) Data linkages have been made to daily weather and local sporting events.

6. Vertical linkages from arrest to final sentence

   (a) Digitized universe of individuals in a district attorney's office over a decade with many stages of random assignment (screener, federal prosecutor, and judge). New Orleans is the largest city and metropolitan area in the state of Louisiana. The Orleans Parish District Attorney's Office and its prosecuting attorneys are responsible for enforcing state criminal laws and local ordinances to protect and serve the citizens of New Orleans and surrounding areas.

   (b) The current data set is from 1988 to 1999 and provides detailed information on approximately 430,000 charges and 280,000 cases (involving 145,000 defendants) filed or adjudicated during this timeframe. The data collected also contains detailed information regarding each individual offender, such as social security number and the corresponding prosecutor and judge.

   (c) Linkages have been made to voting records, bankruptcy, and foreclosure. The dataset is rare: vertical linkages from the time of arrest, including those sent home without a trial, otherwise do not exist. There is a 594-paged codebook.

7. World War I British courts martial

   (a) Digitized World War I British archival datasets, including universe of deserters (including names and often their birthplace) reported in military diaries, police gazettes, and handwritten military trials, commuted and executed capital sentences (which historians believe was random), geocoded casualties, maps, officer lists, and order of battle.

8. US Medicare

(a) Linked administrative Medicare data to industry-physician relationships cleaned from litigation settlements and through the Affordable Care Act, comprising 30 million payments including the date of payment and affiliated drug code.

(b) Linked to biographical characteristics of physician and some patient demographics.

9. Other datasets include judges' financial disclosures, many social and economic datasets, and other cleaned legal datasets:

   (a) Corporate Filings: The text of disclosures, contracts, and charters submitted to the SEC by publicly listed companies for 1996-2016.

   (b) Federal and State Legislation: We have the full history of federal and state laws enacted up until 2012.

   (c) UN Parallel Texts and Hong Kong Laws.

   (d) Party Manifestos Corpus: Large corpus of political party manifestos across the world, with rich metadata.

   (e) U.S. Congressional Record: All speeches by congressman and senator for 1880-2015, with metadata.

   (f) Teacher's union contracts: Large corpus of 6000 teachers contracts for all districts in the state of Michigan since 2010.

10. Over 1000 legal databases tagged and linked including all federal (supreme, appellate, district, bankruptcy, tax, patent, trade, customs, claims, unpublished) and state (supreme, appellate, district, tax, chancery, family, labor, unpublished) court cases to the earliest available date (some as early as 1778).

    (a) Types of databases include code, statutes, bills, regulations, bulletins and notices, commission decisions, Attorney General opinions, rulings, statements, opinion letters, bill tracking, workers' compensation decisions, municipal codes, physician discipline decisions, market conduct examinations, issuances, directives, public health reports, FTC, IRS, EEOC, Department of Labor, Department of Defense, EPA, SEC, Federal Reserve, contract appeals decisions, legislative service, manuals, etc.

11. Case records collected from 24 High Courts and 3000 subordinate courts in India with details on over 8.7 million case records and 67 million hearings. Study (1) impact of court functioning on economic growth and inequality, (2) impact of economics, political, or psychological factors on court outcomes, (3) impact of court decisions or precedents on individuals' outcomes, and (4) artificial intelligence applications.