NYU Center for Data Science and Data Science Justice Collaboratory

Elliott Ash (eash@nyu.edu), Warwick University, ETH Zurich
Daniel L. Chen (dlc16@nyu.edu), Institute for Advanced Study at Toulouse, Harvard, NBER
Levent Sagun (lds351@nyu.edu), ENS-Paris, CEA-Saclay

http://users.nber.org/~dlchen/#data-science-justice-collaboratory
http://users.nber.org/~dlchen/teaching/Projects2018.pdf

1 Introduction

Powerful AI systems will need to be able to reason about moral and philosophical problems and ethical theories. Can AI systems learn philosophical reasoning from a diverse corpus of human text and dialogue? This would be especially helpful in domains where human values are ambiguous or underdetermined. Little work has been done to make these questions tractable. The 2018 themes for DSGA-1003/1012 are:

- **Reasoning**: learning philosophical reasoning from a diverse corpus of human text and dialogue
- **Fairness**: ethics and machine learning
- **Challenge**: competitive results on legal prediction problems using natural language understanding

Our research group is developing the most extensive, comprehensive legal corpora ever produced for academic research. The topics bring together computer science, economics, and law – and lie at the intersection of machine learning, natural language understanding, behavioral economics, and causal inference. There are a variety of legal datasets that comprise in total 12 TB. 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.

- A good class project may result in conference submission and publication, e.g., at NIPS, and continuation for Capstone project.
- An excellent class project brings together two of the broad themes and assumes a high level of commitment, resulting in article publication (see 3 model papers at the end).
- Best practices in reproducibility are required (end-to-end code and intermediate data).

Annotated and linked datasets include the universe of U.S. Supreme, Circuit, District, and state court cases, 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, U.S. Medicare, and bankruptcy filings. These and other datasets are described below. Discussions can deviate from the proposed topics. We will have weekly meetings and share a high performance computing environment.

2 Predicting Judicial Decisions

1. Challenge: **U.S. Supreme Court Prediction (or generally, predicting appeals court decisions using features of lower-court decisions)** is widely considered a challenge for machine learning in law. Several attempts have been made, including by previous year's students, but the success is scarcely better than guessing the most frequent class. The challenge offers hundreds of years of training data and is obviously resistant to simple rule-based systems. Social science has a poor understanding when sets of features provide the greatest lift, be they political (ideology), legal (precedent), economic (considerations), psychological (biases), or sociological (group interactions). Students might improve on previous efforts with a deep learning approach. Contexts: U.S. Supreme Court, U.S. Circuit Courts, State Supreme Courts, Bankruptcy Appeals Boards, Immigration Appeals Boards.

2. Challenge: **Predict Criminal Sentence Decisions.** This challenge involves millions of sentencing decisions, with the goal of predicting these decisions given judge and defendant characteristics. A particularly interesting dimension is to see how features of written opinions in non-crime domains are reflected in their sentencing decisions. Interpretable machine learning may be feasible from the random assignment of cases to judges. Some interesting questions include whether extraneous factors predict decisions: (i) characteristics of the individuals the judges recently saw, (ii) sporting events, (iii) weather, and (iv) which groups of defendants or judges are more predictably affected.

3. Challenge: **Predict a physician's chance of being sued (successfully) for medical malpractice** using microdata on patient characteristics and treatment choices. The challenge is policy-motivated and is made feasible through special access to the U.S. Medicare database. Then one can ask how malpractice risk changes with medical liability law. Auxiliary analysis includes predicting treatment choices from patient characteristics. Does this predictive model change in response to new laws mandating disclosure of gifts to doctors from pharmaceutical companies? Does this predictive model change in response to receipt of gifts from pharmaceutical companies?

3 Reconstructing Judicial Decisions

1. Reasoning: Characterizing judicial decisions as **Liberal or Conservative** is important as a measure for empirical analyses and as a step toward AI understanding of human values. Powerful AI systems will need to be able to reason about moral and philosophical problems and ethical theories. We have a 5% sample for 400 hand-coded case features (along with interrater reliability scores); we want to train a model to form predictions for these features in the 95% unlabeled sample. Other samples exist with hand-coded features, e.g., case category, or, politically salient meaning in 25+ salient legal areas.

2. 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?

3. 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)?

4. 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. **Legal Ambiguity** provide 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.

5. 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.

6. Reasoning: **Utilitarian vs. Deontological** modes of reasoning is a classic divide 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.

7. 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.

8. Fairness: Implicit bias refers to the attitudes or stereotypes that affect our understanding, actions, and decisions in an unconscious manner. Recent advances in natural language understanding suggest AI's ability for **Measuring Implicit Bias From Semantics** derived automatically from corpora. The goal of this project is to conduct an in-depth investigation on the properties of measuring judicial implicit bias using text. How robust are the measurements? Can they be validated against actual judicial decisions (e.g., in discrimination or market competition cases)? What is the causal impact of randomly assigning a woman (black) judge to a panel on the verdict's gender (race) implicit bias? What is the causal impact of court-made law on the text of society?

9. Fairness: Phonology is the study of speech variation beyond word choice, that is, fluctuations in the way one speaks holding the words fixed. Do judge's vocal intonations reflect implicit bias? The goal of this project is also to assess the robustness of **Measurements of Vocal 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.

10. Reasoning: **Judge Embeddings** as 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.?

11. 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. A starting point is to **Predict 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.

4 Predictability of Judicial Decisions

1. Fairness: Why do individuals want laws to be predictable but eschew replacing judicial decision-makers by algorithms? We distinguish between **Early Predictability** (using features available to a judge prior to the opening of a case) and predictability (using features available to a statistician over the course of a case but not observing the decision itself). Early predictability suggests ignoring information about the case. The project has a behavioral component of analyzing information acquisition of judges (for example, using textual data from Wikileaks). Part of the challenge is to set it up in a proper sequential way, to simulate building a model that can then be used for prediction of future trials. and then to contrast it with a model with full information.

2. Fairness: A perennial question is how much law clerks affect judicial opinions. Several prominent studies suggest traces of the writing styles of clerks are reflected in the final opinions. This project examines **Legal Schools of Thought** using the law school training of clerks. Several episodes of unraveling in the judicial clerkship market (where judges had to make hires with less information about clerks) allow studying the efficiency and equity consequences of the current clerkship market system. A parallel question is whether judges' financial conflicts of interest is an important, predictive feature of how judges' vote, write, and cite legal precedent.

3. Challenge: **Judge Health and Output.** Predict early death or dementia, or simply retirement, using judicial corpora. Does health covary with judicial predictability or unpredictability?

4. Fairness: **What Do 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.

5. Fairness: Construct topographical map of WW1 loyalty using universe of courts martial and see if it explains deterrent or anti-deterrent effect of death penalty across different military crimes. Investigate the long-run impacts of "missing men" (geocoded casualties) on demographic and socio-economic outcomes in the UK. Analyze judicial response to executions: **Endogenous Justice**. Analyze physical descriptions of soldiers (youth, complexion, looks, occupation), especially outcome heterogeneity, e.g., do young soldiers respond more to executions? Model geographic spillover / news of executions using geo-location of military units.

5 Impacts of Judicial Text

1. Fairness: A platform for analyzing the **Causal Impacts of Court Decisions** 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.

2. Fairness: A platform for analyzing the **Stock Market Impacts of Text** (e.g., in Wikileaks, SEC corporate filings, etc.). Predict changes in state GDP, employment, profits, and wages from text of session laws, and use topic models to interpret the prediction. Compare with predicting changes in the same outcomes from the text of legal precedent.

6 Deep Learning Applications

- Formulate any of the previous problems as deep neural net applications, including convolutional NNs, recurrent NNs, and autoencoders:

  - Use a convolutional neural network to discover feature mappings for significantly predictive legal language features.
  - Use a recurrent neural network to predict legal language sequences given seed language.
  - Use an autoencoder to optimally reduce dimensionality of judicial decisions.

- Explain what the rule means, answer questions about cases, automatically summarize cases.

7 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.

8 Model Papers

1. Human Decisions and Machine Predictions, The Quarterly Journal of Economics, Volume 133, Issue 1, 1 February 2018, Pages 237–293 (https://doi.org/10.1093/qje/qjx032)

2. Improving Refugee Integration Through Data-Driven Algorithmic Assignment, Science 19 Jan 2018: Vol. 359, Issue 6373, pp. 325-329 (http://science.sciencemag.org/content/359/6373/325)

3. Semantics Derived Automatically From Language Corpora Contain Human-like Biases, Science 14 Apr 2017: Vol. 356, Issue 6334, pp. 183-186 (http://science.sciencemag.org/content/356/6334/183)

9 Prior Years' Projects

Prior to conference submission, previous years' groups ran the end-to-end code from another group to check reproducibility.

1. Algorithms as Prosecutors: Lowering Rearrest Rates Without Disparate Impacts and Identifying Defendant Characteristics 'Noisy' to Human Decision-Makers (D. Amaranto, E. Ash, D. Chen, L. Ren, C. Roper)
   (a) NIPS workshop paper in 2017
   (b) Conference on Law and STEM
2. The Genealogy of Ideology: Identifying Persuasive Memes and Predicting Agreement in U.S. Circuit Courts (*Proceedings of the ACM Conference on AI and the Law*, forthcoming; D. Chen, A. Parthasarathy, S. Verma)
   (a) Cited in 2017 AEA presidential address by Nobelist Bob Shiller
   (b) ACM Conference on AI and the Law
   (c) Conference on Empirical Legal Studies (CELS)
3. What Matters: Agreement Between U.S. Courts of Appeals Judges (*Journal of Machine Learning Research*; TSE Working Paper No. 16-747; D. Chen, X. Cui, L. Shang, J. Zheng)
   (a) NIPS workshop paper in 2016
   (b) Conference on Empirical Legal Studies (CELS)
4. Early Predictability of Asylum Court Decisions (*Proceedings of the ACM Conference on AI and the Law*, forthcoming; D. Chen, M. Dunn, L. Sagun, H. Sirin)
   (a) ACM Conference on AI and the Law
5. Can Machine Learning Help Predict the Outcome of Asylum Adjudications? (*Proceedings of the ACM Conference on AI and the Law*, forthcoming; D. Chen, J. Eagel)
   (a) ACM Conference on AI and the Law
6. Full List