Data Science Justice Collaboratory &
World Bank DE JURE (Data and Evidence for Justice Reform)

Daniel L. Chen (dlchen@nber.org, http://users.nber.org/~dlchen)
Institute for Advanced Study at Toulouse, Toulouse School of Economics, Harvard Medical School,
NYU Center for Data Science, World Bank DE JURE

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 themes for research topics 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.

The topics bring together computer science, economics, and law – and lie at the intersection of **deep learning, natural language understanding, behavioral economics, and causal inference**. Nine projects from previous years have resulted in 3 papers accepted at NIPS (machine learning and law; interpretable machine learning; and causalML) and 6 other peer-reviewed publications. Previous students have received job offers from Facebook AI, Google Brain, Nvidia, Twitter, and Simons Collaboration on Cracking the Glass Problem, and become Chief Data Scientist at LivePerson.

Our research group is developing the most extensive, comprehensive legal corpora ever produced for academic research, comprising over 12 TB. Annotated and linked datasets include the available universe of U.S. Supreme, Circuit, and District court cases, judicial biographies, citation network, 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, bankruptcy filings, and linked U.S. Medicare prescribing, patient health, industry physician payments, and physician malpractice. Growing data ecosystems also include textual or administrative court records linked to firms or household outcomes in Chile, Peru, Brazil, India, Pakistan, Bangladesh, Kenya, Croatia, US, and more. Some projects include collaborations with governments and implementation of machine learning and randomized control trials.

A challenge can be to formulate any of the problems as deep learning applications. 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. Best practices in reproducibility are required (end-to-end code and intermediate data). There will be at least weekly

meetings and Slack will be used to facilitate communication.

2 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.

   (a) 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.

   The usage scenario is rapid causal analytics: automatically identify the nearest previous cases when a District Court case comes up to the Circuit Court (automated topic labeling using document embeddings), 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. Recent econometric theory bridge analysis of text as multi-valued treatments. In addition, recent work propose minimizing prediction error of causal effects.

   (a) Reasoning: 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 appli-

cations 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. Do judge embeddings with neural nets. Likewise, compute citation embeddings using the shopping cart model can render a topography of law. Generate a page-rank of judicial influence via the citations and generation of agreement or dissent (on 3-judge panels).

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: ***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?

4. 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.

   (a) 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.

5. 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.

   (a) 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. Effect of free speech laws or copyright laws on number of books published (and richness of language).

6. 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. Study cultural evolution in legal documents. Predict citations (number, or positive/negative treatment) in the circuit case based on the district court case. Measure innovation in laws (courts and/or statutes). Compute Bayesian Surprise in legal documents.

7. 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?

   (a) 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: 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.

9. 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.

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?

3 Challenges: Judicial Analytics

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. Nowcasting during oral arguments: predict the final ruling using audio recordings, transcribed text, and judge information. The initial goal of this work will be to determine the upper bound of a model's predictive power using the complete oral argument information, but subsequently research will focus on what percent of the oral argument is required to predict the court ruling. Additionally, ablation studies will focus on interpreting what features determine predictions and whether models are utilizing spurious correlations to make predictions. Textio for judges: whether judicial writing can be assisted in real-time, whether to reduce error, reversal rates, or increase citation impact. Automated impact analysis of judicial writing, framing of issues, legal reasoning, narrative generation, surprise and serendipity, creativity (analogize from computational journalism). Identify parts of opinions that spark cites, for identifying sections that lead to polarized cites, and for thinking of citation recommendation to judges. Infer judge's intent from searching for legal precedent (analogize from web or social media search). Study longitudinal queries of judges to legal precedent. Model judicial learning (exploration vs. exploitation vs. bias).

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. 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 surrounding criminals and asylum seekers? Do prison openings suddenly lead to sentencing decisions changing, and the court opinion reasoning changing.

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?

4. Challenge: ***The Legal Summaries Project***: Using a new benchmark data set focused on legal summaries, the work will be primarily focused on building baseline models to evaluate the task difficulty. Evaluation will be done using BLEU and other similar metrics leveraging SOTA Natural Language Generation (NLG) techniques to generate line by line predictions of case summaries. Explain what the rule means, answer questions about cases.

5. 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. Automated discovery of opinion authorship (in unsigned opinions or judge vs. clerks who transit from Circuit to Supreme Court).

6. Challenge: ***Original Meaning Project*** attempts to assist judges and legal scholars in determining the "original meaning" of a concept. Auxiliary work intends to increase access to justice by offering users the possibility of highlighting a part of an argument in a case and then showing all the related cases. Use case - someone can highlight a part of the court opinion and trace its original meaning or see the other cases that use similar arguments. Identify Circuit splits (two cases in different Circuits that has reached opposing conclusions on the law for the same set of facts). Automated discovery of the emergence of new legal issues.

7. Challenge: ***CausalML.*** Adapt or deploy state of the art causal machine learning techniques to benchmark against the true causal effect estimated using random assignment. Predict causal effects using the rich ecosystem of random assignment to cases and to peers.

8. Challenge: ***Legal Knowledge Representation.*** Adapt or deploy state of the art knowledge representation embeddings and benchmark against expert labeled legal outlines. Leverage data on text and citations. Develop a question and answering chatbot.

9. Challenge: ***Generative Model of Judicial Writing and Legal Reasoning.*** Adapt or deploy state of the art generative models of writing on judicial opinions. Likewise, deploy state of the art information retrieval models to generate a model of how judges cite (legal reasoning).

10. Fairness: Automated analysis of disparities. (1) Using data before the date of the decision, does adding controls affect the correlation. (2) Using data with decision-maker identifier, how much of aggregate disparities is driven by how many of the (or which) decision-makers? (3) Analysis of disparities in sequential setting. Using linked data across subsequent or antecedent decision-makers (police), how much of disparities increase/decrease at each decision node? (4) Analysis of disparities, but using data after the date of the decision and (allowing for) different thresholds for decision-makers, are decision-makers more often 'wrong' for certain groups? (5) Using data after the date of the decision, how does the predicted decision-maker compare with the actual decision-maker (if the algorithm lacks data that a decision-maker usefully employs, the predicted decision-maker does worse than the actual decision-maker; if the human is affected by extraneous factors, then the predicted self performs better). How does the algorithm compare with the actual decision-maker? [And this can be broken into different groups of defendants] (6) Analysis of disparities due to decision-maker inattention. Using extraneous factors, see if decision-makers are more inattentive/affected by extraneous factors for certain groups. (7) Analysis of consequences of decision-maker disparities using linked administrative data after the decision. (8) Analysis of internal drivers of decision-maker disparities. Using the text of their decisions, see if stereotypes reflected in one's writings are predictive of decisions and disparities.

11. Fairness: Incremental AI: The usual narrative is backlash to AI. A recent study found that in Kentucky, when judges were given decision-support, it ended up increasing disparities - not because the algorithm was biased - in fact the algorithm would have resulted in lower disparities. But the judges selectively paid attention to the algorithm, which resulted in greater disparities. Consider an incremental approach leveraging recent theoretical insights from social preference economics. The core insight is that judges are moral decision-makers, you're right or wrong, good or bad, and to understand what motivates these decision-makers, one might turn to self-image motives - "I think I'm a good person - a good judge" - a topic of active research in recent years. Each stage leverages motives of self-image, self-improvement, self-understanding, and ego. In stage 1, people use AI as a

support tool, speeding up existing processes (for example, by prefilling forms). Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies and reminding the human of her prior choices in similar situations. Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns. Then, in stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.

4 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.

12. Comparable datasets with Chile and Brazil. What is unique about the Chile opportunity is a rich data ecosystem. Through our relationship with the Chilean Judiciary, we have access to case-level data for civil cases, criminal cases, and appeals. These data sets include every characteristic of the case and parties that is submitted to the electronic system. The data also includes the outcome of the case, as well as whether the case was appealed and the outcome of appeal. Moreover, we have

data on the logins to a dashboard used by judges and other court staff to check statistics on their performance as well as their peers' performance. We also have access to the judiciary's Human Resources Data, which includes biographical, contact and professional information on every court staff member. According to this dataset, there are 13,368 people employed by the judiciary, among which 1,588 are judges, 283 are Court managers, 167 are Ministers of Courts and the rest occupy different staff positions. Other relevant variables include age, gender, position, professional scale, pay grade, type of appointment, status, date of appointment, court, jurisdiction, administrative section. We have firms' data linkable by tax identifier to cases, as well as court user and staff satisfaction surveys. More data is available on request to our Chilean counterparts (text, arbitration, labor employment datasets). For Brazil, we have 200 million cases linked to employment registers - 100 million workers and 3.5 million firms. Ask about Sweden, Croatia, Peru, Kenya, Bangladesh, and Pakistan. Some are linked to registries or censuses of individuals and firms.