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

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**.

Our research group is developing the most extensive, comprehensive legal corpora ever produced for academic research, comprising over 12 TB.

- List of countries with legal data/corpora: Chile, Peru, Brazil, India, Pakistan, Bangladesh, Kenya, Croatia, US, and more
- Not all countries have the same level of detail or availability. We anticipate a conversation to find the maximum overlap between interests.
- Each section has a few sample papers.

1 Legal Predictions

This project type involves predictions that are binary, continuous, or multi-class - outcomes (hearings, cases, trials), decisions (sentencing, appeals, reversals), or categories (case type, biographical features like schools or teachers, expert labels). Ablation studies will focus on interpreting what features determine predictions and whether models are utilizing spurious correlations to make predictions.
Sample papers include

- Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." Econometrica 87.4 (2019): 1307-1340.
- Ash, Elliott, and Daniel L. Chen. "Case vectors: Spatial representations of the law using document embeddings." Law as Data, Santa Fe Institute Press, ed. M. Livermore and D. Rockmore, 2019(11).
- Dunn, Matt, Sagun Levant, Hale Sirin, Daniel L. Chen. "Early predictability of asylum court decisions." Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law. ACM, 2017.
- Lu, Wei, Elliott Ash, and Daniel L. Chen, "Motivated Reasoning in the Field: Polarization of Precedent, Prose, and Policy in U.S. Circuit Courts, 1930-2013"

2 High-Dimensional Causal Inference

This project type involves causal inference using inputs into legal predictions that are randomly assigned and causal inference when there are high-dimensional treatments (like text).
Sample papers include

- Lee, Sokbae, and Bernard Salanié. "Identifying effects of multivalued treatments." Econometrica 86.6 (2018): 1939-1963.
- Heckman, James J., and Rodrigo Pinto. "Unordered monotonicity." Econometrica 86, no. 1 (2018): 1-35.
- Belloni, Alexandre, Daniel L. Chen, Victor Chernozhukov, and Christian Hansen. "Sparse models and methods for optimal instruments with an application to eminent domain." Econometrica 80.6 (2012): 2369-2429.
- Chen, Daniel L., and Jasmin K. Sethi. "Insiders, outsiders, and involuntary unemployment: Sexual harassment exacerbates gender inequality." (2018).

One usage scenario is rapid causal analytics: automatically identify the nearest precedents when a lower court case appears in at appeal, and implement high-dimensional causal inference of the impacts of past verdicts and precedents (represented as text or legal reasonings).

3 Disparities

This project type involves machine learning for rule of law (i.e., equal treatment before the law and equality based on recognition of difference). One usage scenario is automated analysis of disparities.
Sample steps include:

1. Using data before the date of the decision, does adding controls affect the correlation.
   (a) Fryer Jr, Roland G. "An empirical analysis of racial differences in police use of force." Journal of Political Economy 127.3 (2019): 000-000.
2. Using data with decision-maker identifier, how much of aggregate disparities is driven by how many of the (or which) decision-makers?
   (a) Goncalves, Felipe, and Steven Mello. A Few Bad Apples?: Racial Bias in Policing. Industrial Relations Section, Princeton University, 2017. American Economic Review, revise and resubmit.
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?

(a) Rehavi, M. Marit, and Sonja B. Starr. "Racial disparity in federal criminal sentences." Journal of Political Economy 122.6 (2014): 1320-1354.

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?

   (a) Arnold, David, Will Dobbie, and Crystal S. Yang. "Racial bias in bail decisions." The Quarterly Journal of Economics 133.4 (2018): 1885-1932.

5. Using data after the date of the decision, how does the predicted decision-maker compare with the actual decision-maker; How does the algorithm compare with the actual decision-maker? [And b) this can be broken into different groups of defendants]

   (a) Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human decisions and machine predictions." The Quarterly Journal of Economics 133, no. 1 (2017): 237-293.

   (b) Amaranto, Daniel, Elliott Ash, Daniel L. Chen, Lisa Ren, and Caroline Roper. "Algorithms as Prosecutors: Lowering Rearrest Rates Without Disparate Impacts and Identifying Defendant Characteristics 'Noisy' to Human Decision-Makers." Available at SSRN 2993003 (2017).

   (c) Use structural model to estimate what prosecutors maximize: Do they seek to minimize recidivism, maximize conviction rates, maximize sentence lengths, or minimize time until trial completion? See e.g., Lim, Claire SH. "Preferences and incentives of appointed and elected public officials: Evidence from state trial court judges." American Economic Review 103, no. 4 (2013): 1360-97.

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.

   (a) Eren, Ozkan, and Naci Mocan. "Emotional judges and unlucky juveniles." American Economic Journal: Applied Economics 10, no. 3 (2018): 171-205.

   (b) Chen, Daniel L. and Markus Loecher. "Mood and the Malleability of Moral Reasoning: The Impact of Irrelevant Factors on Judicial Decision Making" Available at SSRN 2993003 (2017).

7. Analysis of consequences of decision-maker disparities using linked administrative data after the decision.

   (a) Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad. "Incarceration, recidivism and employment." Journal of Political Economy, forthcoming.

   (b) Mogstad, Magne and A. Torgovitsky. "Identification and Extrapolation with Instrumental Variables" 2018, 08, 577-613, Annual Review of Economics.

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.

   (a) Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356, no. 6334 (2017): 183-186.

   (b) Ornaghi, Arianna, Elliott Ash, and Daniel L. Chen "Stereotypes in High Stake Decisions: Evidence from U.S. Circuit Courts"

4 Memes

This project type studies narratives. 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. Predict the memetic *phrases and citations* that are likely to be passed along the network, but do not otherwise appear in a distant case in the citation graph. Predict influence and how the judge votes on the next panel, using only the history of who the judges previously sat with and how the judges' votes aligned with the panelists.
Sample papers:

- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.
- Shiller, Robert J. "Narrative economics." American Economic Review 107, no. 4 (2017): 967-1004.
- Parthasarathy, A., S. Verma, and Daniel L. Chen. "The Genealogy of Ideology: Identifying Persuasive Memes and Predicting Agreement in the U.S. Courts of Appeals" (Proceedings of the ACM Conference on AI and the Law, 2017)

Usage scenarios can include testing theoretical predictions from population biology.

5 Audio

Phonology is the study of speech variation beyond word choice, that is, fluctuations in the way one speaks holding the words fixed. Several studies have documented that short-term convergence in speech is predictive of votes and that judicial ideology displays longitudinal movements over time. This project predicts court decisions using audio data in short and long time spans. Ablation studies will focus on how much time is needed ("nowcasting") and interpreting features in isolation (recordings, transcriptions, case and judge history). Does the model fit with other data (audio of other speakers' ideologies, or recordings from police stops)? Is the vocal intonation of gendered words (e.g., actor vs. actress) able to classify vocal intonations of neutral words into stereotypically male (e.g., logical, ability, think) and female (e.g., looking, cook, goodwill).
Sample papers:

- Mayew, William J., and Mohan Venkatachalam. "The power of voice: Managerial affective states and future firm performance." The Journal of Finance 67, no. 1 (2012): 1-43.

- Voigt, Rob, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. "Language from police body camera footage shows racial disparities in officer respect." Proceedings of the National Academy of Sciences 114, no. 25 (2017): 6521-6526.

- Chen, Daniel, Yosh Halberstam, and C. L. Alan. "Perceived masculinity predicts US Supreme Court outcomes." PloS one 11, no. 10 (2016): e0164324.

- Epstein, Lee, William M. Landes, and Richard A. Posner. "Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument." The Journal of Legal Studies 39, no. 2 (2010): 433-467.

- Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'Brien. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." American Political Science Review 113, no. 4 (2019): 941-962.

- Pérez-Rosas, Verónica, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. "Deception detection using real-life trial data." In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 59-66. ACM, 2015.

6 Natural Language Processing and Information Retrieval

This project type applies recent advances in NLP, information retrieval, and deep learning for possible application in economics. Recent work in computational linguistics has made breakthroughs in vector representations of language. In this work students will focus on understanding whether the BERT architecture actually understands legal documents or is simply using spurious relationships in the document to make predictions. The first step in this work will be to develop an evaluation method that utilizes features of the legal documents. Possible avenues for such a metric may be next sentence prediction or next word prediction.
Sample papers include

- Niven, Timothy, and Hung-Yu Kao. "Probing neural network comprehension of natural language arguments." arXiv preprint arXiv:1907.07355 (2019). summarized here https://bheinzerling.github.io/post/clever-hans/

- Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon

Htut, Paloma Jeretic and Samuel R. Bowman. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. Proceedings of EMNLP. 2019.

- Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Other usage scenarios include isolating directions for legal and political concepts, liberal vs. conservative, procedural vs. substantive, originalist vs. pragmatists, constitutional applications of principles; identifying substitutable and complementary cases based on how often they are cited together; detection of legal innovation. Use citation data to identify parts of opinions that spark cites, for identifying sections that lead to polarized cites (e.g., by Republicans but not Democrats), and for citation recommendation to judges. Other recommendations include offering users the possibility of highlighting a part of an argument in a case and then showing all the related cases, trace original meaning, or identify inconsistencies reached opposing conclusions on the law for the same set of facts.

7 Summarization

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 (bilingual evaluation understudy) and other similar metrics leveraging state-of-the-art Natural Language Generation (NLG) techniques to generate line by line predictions of case summaries. Usage scenarios would be explaining what the rule means, answer questions about cases, automated chatbots.
Sample papers for summarization (of scientific papers) include

- http://super-ms.mit.edu/rum-final-version.pdf (summarized at https://scienmag.com/can-science-writing-be-automated/)

Then build a task from the legal domain other than summarization. Sample tasks include:

- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353-355. 2018.

- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill,

Omer Levy, and Samuel R. Bowman. "Superglue: A stickier benchmark for general-purpose language understanding systems." arXiv preprint arXiv:1905.00537 (2019).

8 Cognitive/Behavioral/Medical

Predict a physician's chance of being sued and successfully sued for medical malpractice using microdata on patient characteristics and treatment choices. Then one can ask how malpractice risk changes with medical liability law, with receiving gifts from pharmaceutical companies, or with mandated disclosure of these gifts?
Sample papers:

- Jena, Anupam B., Seth Seabury, Darius Lakdawalla, and Amitabh Chandra. "Malpractice risk according to physician specialty." New England Journal of Medicine 365, no. 7 (2011): 629-636.
- Currie, Janet, and W. Bentley MacLeod. "First do no harm? Tort reform and birth outcomes." The Quarterly Journal of Economics 123, no. 2 (2008): 795-830.
- Chen, Daniel L., Vardges Levonyan, S. Eric Reinhart, and Glen Taksler. "Mandatory disclosure: Theory and evidence from industry-physician relationships." The Journal of Legal Studies 48, no. 2 (2019): 409-440.

Predict early death or dementia or simply retirement using judicial corpora. Does health covary with judicial predictability or unpredictability?

9 AI Incremental

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.
Related papers:

- Mullainathan, Sendhil, and Ziad Obermeyer. Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error. No. w26168. National Bureau of Economic Research, 2019.
- Karlinsky-Shichor, Yael, and Oded Netzer. "Automating the B2B Salesperson Pricing Decisions: Can Machines Replace Humans and When?." Available at SSRN 3368402 (2019).
- Babic, B., D. L. Chen, T. Evgeniou, A. Fayard. "Onboarding AI" Harvard Business Review, forthcoming. http://users.nber.org/~dlchen/papers/Onboarding_AI.pdf

10 Reconstructing Judicial Decisions

This project entails studying entailment, analogies, values, ambiguity, and scope. To understand human values, AI systems will likely need to be able to detect and annotate 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. AI systems will also need to understand the fact-value distinction: between what can be known to be true and the personal preferences of individuals. Distinguishing **between analogical or logical** modes of reasoning may also be useful for AI systems. It may also be valuable to formalize what is ambiguous or uncertain. An application may be to study the polarizing effect of elections on judge certainty. Classify prospective versus retrospective language, or "maintain status quo" versus "change from status quo" language.
Sample papers:

- Ash, E., MacLeod, W. B., & Naidu, S. (2018). Optimal Contract Design in the Wild: Promises and Amenities in Canadian Labor Contracts, 1986-2015. Available at SSRN 3204832.
- Cao, Y., Ash, E., and Chen, Daniel L. "Automated Fact-Value Distinction in Court Opinions" European Journal of Law and Economics, invited to resubmit.
- Gennaioli, Nicola, and Andrei Shleifer. "The evolution of common law." Journal of Political Economy 115, no. 1 (2007): 43-68.
- Berdejo, Carlos, and Daniel L. Chen. "Electoral cycles among us courts of appeals judges." The Journal of Law and Economics 60, no. 3 (2017): 479-496.
- Anderson, Jill C. "Misleading Like a Lawyer: Cognitive Bias in Statutory Interpretation." Harv. L. Rev. 127 (2013): 1521.

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. 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).