NYU Center for Data Science

The topics are at the intersection of machine learning and causal inference. There are a variety of legal datasets that comprise in total 4TB. You may need to get help to access the appropriate AWS or NYU High Performance Computing services (please check on this early). Some projects may coordinate with someone local.

- An excellent class project results in journal publication, tackles one of the asterisked projects, and assumes a high level of commitment.
- A good class project may result in conference submission and publication, e.g., at NIPS, and continuation for Capstone project. This route or above is recommended for those seeking letters for PhD programs.
- An ambitious project may utilize recent developments in orthogonalized machine learning, creative use of datasets described below, and join new datasets. This route may also result in presentation-only conference submission.
- A moderately ambitious project may build on prior projects to estimate heterogenous treatment effects through machine learning. This route may involve thinking about social science mechanisms.
- An adequate project entails pure application of classroom tools, revisiting prior years' projects for statistical inference, and evaluating why alternative methods render different conclusions. The last two routes may involve building data wikis.

The data are extremely expensive in terms of money, time, effort, and resources. More than 10 years have been invested in these datasets. The data is strictly confidential. Anyone who wants to access needs to sign a non-disclosure agreement in terms of safeguarding the confidentiality of the data. A portion of one dataset forbids direct contact.

1 Methods

- machine vision - machine listening - NLP (classical / deep networks) - feature importance / feature selection - time series prediction - econometrics - causal inference - orthogonalized machine learning - heterogenous treatment effects - natural experiments - law - moral reasoning - social justice - human rights - judicial disparities

2 Datasets

1. 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. Linked to biographical characteristics of physician and some patient demographics. To examine the impact of pharmaceutical company payments to doctors on prescribing, patient outcomes, and patient adherence and the impact of disclosure laws.

2. Judges are randomly assigned in the U.S. Courts of Appeals. Data on 25 polarized legal areas have already 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. To study the channels through which legal regulations have their effects.

3. Digitized universe of U.S. Courts of Appeals cases from 1880 to 2013 (roughly 380,000 cases), identities of randomly assigned judges sitting on the panels (who is authoring the opinions, writing dissents, or writing concurrences), hand-labeled judges' biographies, hand-labeled legal topic, citation network among the cases, seating network among the judges, 2 billion N-grams of up to length eight, and original text (for use with neural nets). Linked to publicly available Supreme Court datasets, U.S. District Court docket datasets, geocoded judge seats, (some) biographies of judicial clerks, 5% random sample that was hand-labeled for hundreds of features including vote ideology, presence of oral arguments, and administrative data from Administrative Office of the US Courts (date of key milestones, e.g., oral arguments, when was the last brief filed, etc.). To study equal treatment before the law and equality based on recognition of difference and to study the transmission of legal thought. We also have the digitized universe of U.S. State Supreme Court cases from 1947-1994 (roughly 400,000 cases), identities of judges sitting on the panels, hand-labeled biographies, citation network, and original text.

4. Digitized World War I British archival datasets, including universe of deserters (including

names and often their birthplace) reported in military diaries, police gazettes, and hand-written military trials, commuted and executed capital sentences (which historians believe was random), geocoded casualties, maps, officer lists, and order of battle. To study the role of legitimacy in legal compliance and differential effects of the death penalty.

5. 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. 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. Data linkages have been made to daily weather and local sporting events.

6. Digitized universe of 1 million criminal sentencing decisions across 94 U.S. District Courts from 1992-2009 (with randomly assigned judges), hand-labeled biographical data of judges, and 83-paged codebook on features of the case and the defendant. Data comes from U.S. Sentencing Commission. 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.

7. Digitized universe of individuals in a federal prosecutor's office over a decade with many stages of random assignment. 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. 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. 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.

8. Digitized speech patterns in U.S. Supreme Court oral arguments since 1955–longitudinal data on speech intonation (linguistic turns) are rare. Linked to hand-labeled oral advocates' biographies, (some) 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. 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. A machine 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. Text is traditionally treated with discrete models. Speech measurements (for example resonances) by contrast are continuous. Recordings of U.S. Circuit Court oral arguments for a limited time period.

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.

3 Judicial Decisions

1. Refugees seeking asylum are assigned to a randomly chosen judge. 400K decisions of 350 judges in over 40 courts. The refugee is either granted asylum or not. For each refugee, we

have a record of the judge they were assigned to and the timestamp of the trial, as well as information about the refugee (e.g. nationality, defensive/affirmative, lawyer).

   (a) The obvious prediction problem here is whether or not the refugee is granted asylum. Part of the challenge of this problem 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. The rest of the challenge is to find all the things (via feature generation) that could be predictive (e.g. judge ID, presumed skintone of refugee based on nationality, time-of-day, most recent decisions by the same judge, recent decisions in the courthouse, recent successes of the lawyer, characteristics of the individuals the judges recently saw, etc.), and whether these factors affect disparities.

   (b) Study information acquisition of judges who have discretion in scheduling hearing sessions.

   (c) Judges negatively autocorrelate their decisions. One interpretation is gambler's fallacy and another is sequential contrast effects. Are there racial contrast effects?

2. Criminal defendants are assigned to a randomly chosen District Court judge.

   (a) The obvious prediction problem here is the sentence length. Part of the challenge of this problem 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. The rest of the challenge is to find all the things (via feature generation) that could be predictive (e.g. judge ID, most recent decisions by the same judge, recent decisions in the courthouse, characteristics of the individuals the judges recently saw, etc.). What is interesting from a social and legal perspective is whether extraneous factors like (i) characteristics of the individuals the judges recently saw, (ii) sporting events, and (iii) weather ends up predicting decisions and if so, what factors are relevant and by how much, and whether these factors affect disparities.

3. We have data on arrested individuals constituting the universe for the New Orleans District Attorney office for over a decade with many stages of random assignment (screener, federal prosecutor, and judge).

   (a) The obvious prediction problem here is the sentence length, whether the individual was guilty, and whether the individual was sent home without trial. Part of the challenge of

this problem is to set it up in a proper sequential way, to simulate building a model that can then be used for prediction of future cases. The rest of the challenge is to find all the things (via feature generation) that could be predictive (e.g. judge ID, most recent decisions by the same judge, recent decisions in the courthouse, characteristics of the individuals the judges recently saw, characteristics of the screener, what the screener recently saw, characteristics of the prosecutor, etc.). What is interesting from a social and legal perspective is whether extraneous factors like (i) the screener identity, (ii) prosecutor identity, and (iii) judge ends up predicting decisions and predicts inequality of treatment of individuals by race and gender, and if so, what factors are relevant and by how much, and whether these factors affect disparities.

4 Court Opinions

1. For a structured prediction problem on cases,

   (a) Automate the coding of cases as liberal or conservative.

      i. Automate the coding of case category.

      ii. A sample of 6,500 cases have been hand-coded for meaning, like pro-plaintiff or pro-defendant, pro-business or pro-environment, pro-criminal defendant rights or pro-prosecutor, etc., in 25 politically salient legal areas; a 5% sample has been hand-coded for hundreds of features

      iii. Using the hand-coded label, predict the label in the larger dataset using the text of the 300,000 cases over a pre-defined set of legal topics.

      iv. Using information other than the judge identity, predict the hand-coded label.

      v. Do word choice co-appearance predict decisions better than words by themselves?

   (b) Automate the partitioning of cases.

      i. Predict where the fact pattern is located using only the text of the opinions; i.e., use hand-coded cases from law students identifying the fact patterns from law, then predict where fact pattern is.

      ii. Predict where there is discussion of a previous case, versus discussion of the current case.

   (c) *Predict reversals by the U.S. Supreme Court.

i. Predict higher court decisions from the way the lower court opinion's fact pattern is summarized.

ii. Predict subsequent treatment by courts (citation or flavor of citation); i.e., predict the type of citation (not just presence of citation) as a continuous measure of reversal. Does the legal text or its citations predict future treatment by courts?

iii. Automate the detection of legal inconsistencies.

iv. Study the use of certainty by judges. One idea is that a more certain opinion states a clearer policy position. This makes it more attractive to cite by future judges. However, it makes it more likely to be reversed. ("High-certainty" Supreme Court opinions have been found to be cited more often.) Use assignment of high-certainty judges as an instrument for the certainty of opinions, with citations and reversals as outcomes. In the state courts data, examine whether judges become more certain in response to election incentives.

(d) *Predict the memetic phrases or memetic citations, for example, using information on the citation network or randomly assigned seating network.

i. Using only the citation network, predict the memetic *phrases* that are likely to be passed along the network in forward citation but not appear in a distant case in the citation graph.

ii. Using only the citation network, predict the memetic *citations* that are likely to be passed along the network in forward citation but not appear in a distant case in the citation graph.

iii. *Using only the seating network, predict the memetic phrases or citations that are likely to be passed along the network after sitting together but not appear in a distant seat in the seating graph.

iv. Identify memes in different legal topics.

v. *Apply and compare with genomic techniques on judicial N-grams and citations to infer history of ideas.

vi. Classify the writer of an opinion based on the style for each circuit-year.

vii. The behavior of judges varies over the presidential election cycle. Do their words? Do campaign advertisements get reflected in the words of judges?

(e) Structured prediction of who agrees with whom, jointly, memes that predict agreement

or memes that transmit, knowing the exact identities of all three judges.

(f) Predict which judges lead and which follow; i.e., which judges are more likely to influence how others cite and write after sitting with them.

   i. Using only the history of who the judges previously sat with on panels and how the judges' votes aligned with the panelists, predict how the judge votes on the next panel.

   ii. Rank judges as more or less distinct/influential based on how much their language style is pushed into the opinions for panels where they are assigned.

   iii. Predict early death or onset of dementia using text of judge written opinions.

(g) *Measuring judges' implicit biases from semantics derived automatically from their corpora.

   i. Apply techniques in 1, 2, and 3 to judicial opinions.

   ii. Examine the causal impacts of implicit biases on their votes.

   iii. Inferring is-ought distinctions in judicial word embeddings.

(h) Can we infer something like ideology from judges' written decisions (in addition to their votes); how many dimensions does ideology have?

   i. Form predictors of types of statutory interpretation; i.e., textualism, etc.

   ii. Create measures of utilitarian vs. deontological thinking among state supreme court judges (we also have a comprehensive dataset on state supreme court judges).

      A. Some judges are elected and some are appointed. There is also substantial variation in their wages.

      B. What does it mean textually for judges to be more duty-oriented in their thinking?

   iii. Do judges who attend the law and economics conferences use different features of defendants in obtaining sentences?

(i) Data on judges' financial disclosures is somewhat pre-processed.

   i. Do financial payments to judges influence how they vote, how they write, how they cite?

   ii. Generally, the influence of markets on development of law / schools of thought.

iii. Compare the judge embedding before and after they attended an important law and economics education conference.

(j) Predict paradigm shifts; i.e., predict the boundaries of citation clusters using only data on backwards citations.

i. Citation patterns distinguish objective (gradual, cumulative) and subjective (paradigm shifts) knowledge in the sciences and humanities; do they also reveal differences in the law (e.g., corporate law vs. civil liberties)?

(k) *Automating causal analysis of judicial decisions through machine learning. Examples include 1 and 2. Practical impact is an app that can be used.

i. *Implement high-dimensional instrumental variables for text from the corpus of a judge's previous opinions. Using deep instrumental variables to predict the impact of judges decisions in heterogenous populations. Alternative project is to interpret the legal text as a situation of many treatments.

ii. Is the hand-coded label of the text the dominant predictor of outcomes?

iii. Which part of the opinion is predictive of outcomes?

iv. Does dicta matter beyond the holding (the court ruling)?

v. Measure effect of republican judge assignment to circuit court economics cases, and subsequent effect on state profit share in BEA accounts.

vi. Effect of judge assignment on:

A. IRS/tax decisions and tax collections by circuit

B. criminal cases and criminal activity

C. immigration cases and immigration/wages

D. environmental cases and pollution

E. labor law cases and union density

F. civil rights cases and attitudes towards rights

G. (more)

(l) Interpreting document similarity.

i. Visualize which words are most important in driving similarity between documents.

A. http://mccormickml.com/2016/11/04/interpreting-lsi-document-similarity/

(m) Measuring ambiguity in legal text.

i. http://web.law.columbia.edu/sites/default/files/microsites/tax-policy/files/2013/anderson_-_misreading_like_a_lawyer_nov_2013_cls.pdf

ii. Look at 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.

(n) Classifying conditional language: "if A, then B" constructions.

(o) 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.

(p) Entity recognition and coreference resolution in legal texts across domains. Identify the same entities in statutes, cases, and contracts. For businesses, link to dun bradstreet ID.

2. Predict judicial quality using clerk data.

(a) Predict judicial writings using law schools of clerks.

5 Legal Documents

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

(a) Predict future changes in stock price and returns using the text of filing documents.

(b) Predict the volume of mergers using the text of filing documents.

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

(a) Predict changes in state GDP, employment, profits, and wages from text of session laws, and use topic models to interpret the prediction.

(b) Compare with predicting changes in the same outcomes from the text of legal precedent.

3. UN Parallel Texts and Hong Kong Laws.

(a) Generate equivalent legal phrases using parallel multilingual corpora.

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

   (a) Use the hand-coded political statements (as seeking to expand policy versus restrict policy) to build a classifier for political speech seeking to expand/restrict policy.

   (b) Then use this on the congressional record floor speech and court opinions.

5. Teacher's union contracts: Large corpus of 6000 teachers contracts for all districts in the state of michigan since 2010.

   (a) Extract features of teacher collective bargaining agreements associated with better or worse student test scores.

   (b) Predict test scores and other outcomes out of sample.

   (c) Analyze differences in contracts across areas with different socioeconomic status and school density.

6 Oral Arguments

1. 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).

   (a) Using the hand-coded label, predict the perceptions on the 1955-2013 set of audio files.

   (b) We also have the previous best prediction model of Supreme Court outcomes, which incorporates features such as the judges, case characteristics, characteristics of the lower court of origin, and historical trends at the lower courts and of the judges; do addition of features generated from the raw sound file or lawyers' faces predict

outcomes? Does relevance of audio and facial features in best predictive model differ for men and women?

2. We have the entire 1-hour audio files and transcripts for oral arguments in the Supreme Court dating back to 1955 along with the opinions. Features have also already been generated for some raw sound files.

   (a) Comparing against the previous best prediction model of Supreme Court outcomes, do addition of transcript features predict voting?

   (b) Can convergence in how judges speak predict voting together over time?

   (c) *Modeling social influence via phonetic accomodation in oral arguments. One project involves applying 1 to the audio data (not just textual) data in oral arguments.

   (d) Do linguistic features of oral arguments predict political valence? Are there political dialects in intonation (in addition to political polarization in the use of words)?

3. Circuit Court oral arguments are also available but have not been pre-processed.

   (a) Predict who is speaking.

   (b) Nothing is known about whether linguistic features are relevant, whether they predict stock market outcomes, etc.

   (c) Nothing is known about whether judges' questions or how lawyers' respond to questions are predictive of outcomes.

7 Conflicts of Interest, Attitudes, and Historical Data

1. Construct a predictive model of physician payments from pharmaceutical companies.

   (a) The effects of disclosure laws on what is being disclosed are typically unknown since data on disclosed activity rarely exist in the absence of disclosure laws.

   (b) Does the predictive model change before and after mandatory disclosure?

   (c) Predict malpractice based on patterns in payments, prescribing, patient outcomes.

2. We have data on social attitudes of Americans since 1972 and the social attitudes of a global sample of individuals.

(a) Predict the attitudes using demographic characteristics or any other social, political, economic feature. What is interesting from a social and legal perspective is whether major Supreme Court or Circuit Court cases affect attitudes in the direction the law intended, in the opposite direction the law intended, or polarizes individuals, (ii) the extraneous factors like judicial biographies randomly assigned to cases affect attitudes. Compare the predictive model with one for world sample.

(b) We also have population-representative U.S. data on labor markets, marriage markets, voting outcomes, crime outcomes, economic growth, property prices, etc., for alternative outcome analyses.

3. Predict hot-spots of empathy gaps in multiculturalist societies (conflict, violence, crime, anxiety) using data from developed or developing world.

4. Predict contemporary UK outcomes using geographic variation in WWI casualties, etc.