Legal Studies 190.2
Data, Prediction, and Law
Spring 2018
TuTh 10 a.m.-12:00 p.m.
Barrows 122

Jon Marshall
jdmarshall [at] berkeley [dot] edu
2240 Piedmont Ave. Room 114
510-642-3670
Office hours TU 2-4 PM, W 2-4 PM (or by appt.)

GSI: Aniket Kesari
akesari [at] berkeley [dot] edu
Office Hours: Th 2-3 PM (or by appt)
Location: TBA

Data, Prediction, and Law

Description

Data, Prediction, and Law allows students to explore different data sources that scholars and government officials use to make generalizations and predictions in the realm of law. The course will also introduce critiques of predictive techniques in law. Students will apply the statistical and Python programming skills from Foundations of Data Science to examine a traditional social science dataset, "big data" related to law, and legal text data.

Note: students should complete Foundations of Data Science, or complete equivalent preparation in Python and statistics, before enrolling in this course.

Learning objectives

By the end of Data, Prediction, and Law, students will be able to

1. use common statistical and computational techniques to analyze different types of data (traditional survey data, big data, and text data) related to law; and
2. critique the use of data and predictive tools in sociolegal processes, including the identification and punishment of crime.

Assessment

The instructors will assess student progress using problem sets, a data investigation final project that will be a team effort, an online final exam, and class participation.

problem sets / lab exercises	40%
data investigation project	30%
final examination	15%
class participation	15%

Texts

The readings for the course are entirely electronic, and will either be available as a public document somewhere on the Internets or on the bCourses site for the course, or both.

Policies

The course requires you to read the reading assignments, participate in discussion and lab, do your homework, complete a team project, and take a final. Please feel free to come to office hours (or use the bCourses discussion or email tools) with ideas and questions. It has never been easier to talk to your instructor and GSI, so take advantage.

Please be on time. You are expected to prepare for each class. Take notes as you read (and in class). If you want to use social media, send text messages, or communicate with friends, do it outside of class. Drinking coffee, water, etc., in class is fine, but eating is a distraction to your fellow students, so do not eat in class. Basically, we are all adults here, so the expectation is that we will treat one another with respect.

Finally, please refer to Berkeley's Academic Integrity policy (http://sa.berkeley.edu/conduct/integrity). *I take academic integrity and honesty seriously. If you plagiarize, cheat, or are otherwise dishonest, you will at fail* ***at least*** *the assignment in question (or more likely the course), and I will file an academic dishonesty report.* If you have any questions about this, please ask.

Students requiring accommodation for disability should also make sure that I get the official accommodation notice from DSP ***by the third week of the semester*** (or as soon as possible after they have been to DSP). Make sure to check bCourses daily, since that will be our medium of communication.

Course Structure

The course will be divided into three units, each of which focuses on a different type of data and the tools, techniques, and problems associated with that type of data. Some readings are perhaps yet to be determined.

I. Social Science Data and Generalization

By the end of Unit I, students should be able to

1. explain the features of structured social data
2. use Python to analyze social science survey data
3. show familiarity with critical perspectives on the role prediction in the field of law (esp. probation and parole)

	date	class meeting topic	reading to have prepared before class
1	1/16	data, prediction, law student questionnaire	Gebru et al 2017 (http://www.pnas.org/content/114/50/13108.abstract) [bCourses] look at all three data sources for class (ANES 2016, SFPD Incident Reports, Old Bailey Proceedings)
2	1/18	data types (incl. UC Berkeley's Tables), functions in Python, collection and cleaning of traditional survey data	Adhikari and DeNero, chs. 3-5 https://ds8.gitbooks.io/textbook/content/ ANES 2016 Codebook (pp. 3-7) [bCourses]
3	1/23	summary stats (mean, s.d., distributions…) Team question for the ANES data	Adhikari and DeNero chs. 7, 9 Harcourt, *Against Prediction* ch. 1 [bCourses] *supplementary*: Feeley and Simon 1992 "The New Penology" *Criminology* (**30**:4) pp. 449-474 [bCourses]
4	1/25	hypothesis testing	Adhikari and DeNero ch. 10

5	1/30	estimation & uncertainty, large N Team answer for ANES data	Adhikari and DeNero chs. 11-12 Skeem & Lowenkamp 2016 "Risk, Race, & Recidivism" [bCourses]
6	2/1	correlation, OLS regression regression and causal inference	Adhikari and DeNero ch. 13 *suggested*: Introduction to Statistical Learning, chs. 2-3 [bCourses]
7	2/6	more on predicting recidivism lab: mapping	Adhikari and DeNero ch. 14 Angwin et al 2016 "Machine Bias" with Larson et al 2017 "How We Analyzed the COMPAS Recidivism Algorithm" (appendix) [bCourses] data: https://github.com/propublica/compas-analysis
8	2/8	wrap up on structured social data prediction & parole lab: mapping 2	Flores et al 2017 "False Positives, False Negatives" (rejoinder to Angwin) [bCourses] **DATA INVESTIGATION PROJECT PROPOSAL DUE**

II. Big Data and the Police

By the end of Unit II, students should be able to

1. explain the power and pitfalls of "big data" in making predictions
2. use Python to make predictions, as well as data visualizations and maps, from large datasets
3. show critical understanding of machine prediction

	date	**class meeting topic**	**reading to have prepared before class**
9	2/13	predictive instruments and the decision to punish lab: mapping: heat maps + time	Adhikari and DeNero chs. 15-16 check out SFPD Incident report data https://data.sfgov.org/Public-Safety/Police-Department-Incidents-Current-Year-2017-/9v2m-8wqu/data Bay Area News Group materials on Scribd from Brock Turner case (survivor's statement (ex. 16), probation report) (see also police report, character letters, complaint, sentencing memo); material also at L.A. Times *suggested*: Introduction to Statistical Learning ch. 4 [bCourses]
10	2/15	Bayes and updating priors machine versus human predictions	Adhikari and DeNero ch. 17

		lab: folium mapping plugins	Kleinberg et al. 2017 "Human Decisions and Machine Predictions" [bCourses] Kleinberg et. al. 2015, "Prediction Policy Problems" (don't get hung up on the math notation!) [bCourses]
11	2/20	SFPD incident report data and its application machine learning models lab: math operations numpy/scipy	SFPD Incident report data Ang et al 2015 "San Francisco Crime Classification" (grad student project) [bCourses]
12	2/22	modeling risk litigating predictive models (more about COMPAS) lab: intro to scikit-learn	Barry-Jester, "Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?" [bCourses] State of Wisconsin v. Loomis (pp. 1-31, *supplementary 31-48*) [bCourses] Dressel and Farid (2018), "The accuracy, fairness, and limits of predicting recidivism," Science Advances 4:1 (17 Jan) http://advances.sciencemag.org/content/4/1/eaao5580 **PROBLEM SET 1 DUE**
13	2/27	a new phrenology? thinking about what models are actually doing lab: model selection	Wu and Zhang 2016 "Automated Inference on Criminality" [bCourses] http://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html [bCourses] Wu and Zhang 2017 "Responses to Critiques on Machine Learning of Criminality Perceptions" [bCourses]
14	3/1	allocation of resources and models lab: preprocessing text	http://www.pbs.org/wgbh/frontline/film/policing-the-police/ Washington *Post*, "Sessions Orders Justice Department" (3 Apr 2017) [bCourses]; Atlantic, "Can Trump's Justice Department" (4 Apr 2017) [bCourses]
15	3/6	surveillance, selection, and the ratchet effect lab: intro to text analysis	*Floyd v. City of New York* ("stop and frisk" decision), pp. 1-15 (and whatever else interests you) [bCourses] Harcourt 2007 *Against Prediction* ch. 5 [bCourses]
16	3/8	remaining questions and discussion on predictions from "big data"	Justin Grimmer and Brandon Stewart, "Text as Data: The Promise and Pitfalls of Automatic

		lab: data investigation project	Content Analysis Methods for Political Texts," *Political Analysis* pp. 1-31 [bCourses]
17	3/13	selecting into a dataset: thinking critically about what data are collected lab: parsing XML data	Fryer (2016) "An Empirical Analysis of Racial Differences in Police Use of Force" (pp. 1-7) and its discussion and follow-up on Andrew Gelman's blog [bCourses] **PROBLEM SET 2 DUE**

III. Law as Text as Data

By the end of Unit III, students should be able to

1. identify and explain what questions can be asked of text data;
2. use Python and other tools to prepare and analyze text computationally;
3. demonstrate understanding of historical context in which text was produced.

	date	class meeting topic	reading to have prepared before class
18	3/15	understanding how Old Bailey Proceedings data got made lab: regular expressions and dictionary methods	"About the Proceedings," "Historical Background to the Proceedings of the Old Bailey (esp. "Crime, Justice, and Punishment)" on Old Bailey Corpus site https://www.oldbaileyonline.org/ [bCourses] Hall & Wright 2008, "Systematic Content Analysis of Judicial Opinions," 96 Cal. L. Rev. 63
19	3/20	Marx, history, and law as indicator or constitutive lab: text classification (cont.)	Hay, "Property, Authority, and the Criminal Law" *Albion's Fatal Tree* (New York: Pantheon, 1975, 17-63) [bCourses] Langbein, "Albion's Fatal Flaw" *Past & Present* 98 (Feb. 1983), 96-120
20	3/22	outline of computational text analysis techniques lab: topic modeling	Programming Historian on extracting and using Old Bailey Corpus https://programminghistorian.org/lessons/naive-bayesian and Beautiful Soup for scraping https://programminghistorian.org/lessons/intro-to-beautiful-soup
21	4/3	the Old Bailey in its legal-historical context lab: word to vector modeling	Tim Hitchcock and William J. Turkel. 2016. "The Old Bailey Proceedings, 1674–1913: Text Mining for Evidence of Court Behavior," *Law and History Review* 34:4, 929-955. [bCourses] Randall McGowen 2002 "Making the 'Bloody Code'? Forgery Legislation in Eighteenth-

			Century England" *Law, Crime and English Society 1660-1830* (Cambridge University Press), 117-138. [Berkeley Libraries: campus only][bCourses] *supplementary*: Klingenstein, Hitchcock, and DeDeo, 2014. "The Civilizing Process in London's Old Bailey," *PNAS* 111:26, 9419-9424. [bCourses] *supplementary*: Lieberman, "Mapping Criminal Law: Blackstone and the Categories of English Jurisprudence" *Law, Crime and English Society 1660-1830* (Cambridge University Press), 139-162. [bCourses]
22	4/5	wrap up on Old Bailey in its historical context lab: more word to vec	http://guides.lib.berkeley.edu/text-mining API link for Court Listener data https://www.courtlistener.com/api/ **PROBLEM SET 3 DUE**
23	4/10	text as social science evidence lab: other word embedding models	Iris Hui 2017 "Shaping the Coast with Permits: Making the State Regulatory Permitting Process Transparent with Text Mining" Coastal Management 45:3, 179-198. [bCourses]
24	4/12	text as social science evidence 2 lab: other word embedding models (continued)	Oard & Webber, Information Retrieval for E-Discovery, "Introduction" & "The E-Discovery Process" (Provides a helpful overview of the E-Discovery process in the US from the perspective of information retrieval experts) Grossman & Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient that Exhaustive Manual Review," Richmond Journal of Law & Technology (Academic article that put technology-assisted review on the map, cited in almost all TAR-related court cases and articles) Rio Tinto (Judge Peck Opinion): "Predictive Coding a.k.a Computer Assisted Review a.k.a. Technology Assisted Review (TAR) – Da Silva Moore Revisited" (Opinion that provides helpful, short overview of TAR-related case law) In re: Broiler Chicken Antitrust Litigation: "Order Regarding Search Methodology for Electronically Stored Information" (good

			example of scope and content of contemporary E-Discovery activities and negotiations)
25	4/17	text as social science evidence 3 lab: feature selection in machine learning	Alice Wu 2017 "Gender Stereotyping in Academia: Evidence from Economic Job Market Rumors Forum" [bCourses]
26	4/19	lab: ensemble methods	
27	4/24	Data Investigation Project team presentations	
28	4/26	Data Investigation Project team presentations	**PROBLEM SET 4 DUE**

DATA INVESTIGATION PROJECT DUE FRIDAY 4 MAY 5 P.M.

TAKE HOME FINAL EXAMINATION DUE FRIDAY 11 MAY 5 P.M.

Data Investigation Project

The DIP is intended to let students <nyuk>get their feet wet</nyuk> in a data analysis project of their own choosing. It will be a team project. Students will work in pairs (the plan is pair up students who have more coding experience with students who have less), propose what they would like to find out from what data, decide how they will get the data, decide how they will answer their question, and then go about answering their question and documenting (using a Jupyter notebook) how they got their answers.

proposal	10%
collection and use of data	20%
modelling/analysis	30%
oral presentation of findings	10%
written report (including visualizations)	30%

Project teams should consult with the instructor and GSI early and often. The research question in the proposal is key, so project teams should think carefully about what interesting question they want to answer using data. The data may be from one of the datasets we explore in class (i.e., the National Crime Victimization Survey, San Francisco Police Incident Reports, Old Bailey Proceedings) or from another source selected by the project team. Be sure that you follow best practice for acquiring publicly available data. Please see the Library's guide here http://guides.lib.berkeley.edu/text-mining, and remember that Berkeley has access to a huge variety of data through sources like HathiTrust, Inter-University Consortium for Political and Social Research, and so on. If we as a university community violate terms of service, then UC Berkeley could be cut off from these valuable sources of data. Practice safe scraping and acquire data properly (see the Library's flowchart). If you are in doubt, ask the instructors, and if they are in doubt, we will all ask the research librarians.

The DIP will be created as a Jupyter notebook, so that other people can see what you have done and can have the opportunity to replicate your work. Your DIP may serve as the basis

for future work you do at Berkeley, and so the ability to document and repeat what you have done could come in handy for a senior thesis, job interview, or other project.

Proposal

Your proposal should be about two to three paragraphs long and should present, in reasonable detail (for 10 points each)

1. your research question
2. the data you plan to use and how it will help you answer your question
3. how you will gain access to the data you need and put it into a form you can analyze

Be sure to come up with a research question that is interesting to you and to other people too. This is a team project so each team will get one grade; if you have problems with slacking, free-riding, sniping, or general lack of cooperation please see your instructors as soon as possible so we can work on a corrective. Please upload your proposal here (note that there are restrictions on file type so be sure you have uploaded a readable file, and I enabled Turnitin just because).

Collection and Use of Data, Modelling/Analysis

This portion of the project will be evaluated through what each team presents in the oral presentation and the written report. These elements of the project will be evaluated separately from the quality of the oral presentation or written report. Collection and use of data will be evaluated based on adherence to ethical data collection standards, appropriateness to question, inventiveness in acquiring and combining data, and clarity in explanation of data gathering methods and dataset content. Modelling and analysis will be evaluated based on appropriateness to question, clarity in explanation of model and the reasons for using it, and the exposition of the relationship between the team's modelling efforts and the conclusions the team draws.

Oral Presentation

Each team will be responsible for reporting the results of their Data Investigation Project to the rest of the class during the last week of classes. The oral presentation will be evaluated on delivery, use of visualizations, clarity, and audience engagement.

Written Report

The written portion of the Data Investigation Project includes both a Python notebook that allows interested readers to reproduce the team's analysis (and so incorporates sufficient comments and markup cells to explain what is going on) and a brief (no more than six pages, double-spaced, in 12-point type) report of the question under investigation, why the question is important, how you went about answering the question, and the conclusions you were able to draw from the data. The report will be evaluated based on completeness in covering the points above, organization, clarity, and (as a bonus) originality and inventiveness.