Legal Studies 123	Jon Marshall
Data, Prediction, and Law	jdmarshall [at] berkeley [dot] edu
Spring 2021	Office hours: TU 2-3:30 PM (or by appt.)
MW 10 AM - noon	GSI: Ilya Akdemir
	CA's: Anna Gueorguieva, Akhila Kandaswamy, Ananya Raghavan

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 also introduces critiques of predictive techniques in law. Students 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. The course proceeds along two tracks; the readings and discussion cover the critique, and the labs and problem sets enable the exploration.

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 and produce visualizations for 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, written reflections on the class readings, a data investigation final project that will be a team effort, and class participation (which includes attending and speaking up but also written opportunities via the Zoom chat, Piazza, or bCourses, on a 0-3 scale for each session). Although the in-person version of the class also has a take-home final exam, the remote version will treat the data investigation project and last problem set as a final assessment of what students have learned in the class.

problem sets (3 total)	50%
data investigation project	35%
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. A book you may find useful in discussing the foundations of machine learning is James et al., *An Introduction to Statistical Learning, with Applications in R* (Springer, 2013). The book is available in electronic form from Berkeley campus IP addresses (or campus VPN) and via

Oskicat, and I will put a couple of relevant chapters on bCourses. Don't worry about R (since we will be using Python) but rather the conceptual content.

Policies

The course requires you to read the reading assignments, participate in discussion (even if that is asynchronous) and lab, do your homework problem sets, and complete a team project. Please feel free to come to office hours (or use the bCourses and Piazza discussion or email tools) with ideas and questions. It has never been easier to talk to your instructor and GSI, so take advantage. Let me know if you would prefer a class Piazza group and I will set one up.

Please be on time to synchronous activities and meetings. 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 time. It improves the discussion experience if we can see one another but given the nature of Zoom and working from home, I understand that it may not always be possible. Still, please try. 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 primary 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 may be subject to change.

Each class meeting features a lab exercise. The labs are not graded but we will discuss them in class, typically at the end of class. **The problem sets rely on the techniques you learn in the labs,** though, so do them; don't just look at the solutions, since that will make the work on the problem sets harder. I plan to ask that you **start the labs before the class** meeting so that we can work on questions and discuss what is in each lab during the class period. Some students may be more familiar with the Python, or the methods, that each lab features, so my hope is that you help each other out and teach one another what you know.

Labs are available in a Git repository at https://github.com/ds-modules/Legalst-123. We will run each lab on Datahub to ensure that all the dependencies work, so you can use this interact link to pull the labs into your directory on Datahub. That will sync what is in your directory with what is in the Git repository, so if you want to save your work, change its filename or it will be overwritten the next time you click. You should also save a local copy that you can refer to once the semester is over. If you are going to work locally, use big data files, etc., you will need to install Anaconda. Many projects have relied on Google Colab, which allows you to run your notebook and collaborate (on Google's servers).

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 critical understanding of prediction in policing

	date	class meeting topic	prepare before class
1	1/20	data, prediction, law student questionnaire examples of big data inference Lab 1: Anaconda setup if you want to run things locally Lab 2: Intro to Jupyter Notebooks	Gebru et al 2017 "Using Deep Learning and Google Street View to Estimate the Demographic Makeup of Neighborhoods" PNAS 114:50 [bCourses] Chang et al 2020, "Mobility Network Models of Covid-19" Nature 589, 82-87 (2021) [bCourses] look at all three data sources for class (ANES 2016, SFPD Incident Reports, Old Bailey Proceedings)
2	1/25	data structures (incl. Pandas dataframes) Lab 3: Dataframe Operations & Simple Visualizations	Adhikari and DeNero, *Computational and Inferential Thinking* chs. 3-5 (review) https://www.inferentialthinking.com/chapters/intro ANES 2016 Codebook (pp. 3-7) [bCourses]
3	1/27	summary stats (mean, s.d., distributions…), collection and cleaning of traditional survey data Lab 4: Probability Distributions, Bootstrap, and Confidence Intervals	Adhikari and DeNero chs. 7, 9-10 (review) *suggested*: Introduction to Statistical Learning, ch. 2 [bCourses]
4	2/1	estimation & uncertainty, large N hypothesis testing Guest: Prof. Sarah Brayne, University of Texas Sociology Lab 5: Large n and hypothesis testing	Adhikari and DeNero ch. 11-14 (review) Brayne 2017, "Big Data Surveillance: The Case of Policing," Amer. Soc. Review 82:5, 977-1008 [bCourses] *suggested*: Steinberger 2020, "Does Palantir See Too Much?" NY Times Magazine 21 Oct. [bCourses] [optional] Brayne 2018, "The Criminal Law and Law Enforcement Implications of Big Data," *Ann. Rev. of Law Soc. Sci.*, 14:293–308 [bCourses]
5	2/3	correlation, OLS regression regression and causal inference Lab 6: OLS for Causal Inference	Adhikari and DeNero ch. 15-16 (review) *for future reference*: CTA labs (courtesy Ilya)

			Wexler, "How Data Privacy Laws," LA *Times* 31 Jul 2019 [bCourses] *suggested*: Introduction to Statistical Learning, ch. 3 [bCourses]
6	2/8	Prediction in crime control Lab 7: Introduction to Folium (mapping)	Kleinberg et. al. 2015, "Prediction Policy Problems" (don't get hung up on the math notation!) [bCourses] Harcourt, *Against Prediction* ch. 1 [bCourses] *supplementary*: Feeley & Simon 1992 "The New Penology" *Criminology* (**30**:4) pp. 449-474 [bCourses]
7	2/10	SFPD incident report data and its application machine learning models Lab 8: Folium Choropleth Maps	SFPD Incident report data Ang et al 2015 "San Francisco Crime Classification" (grad student project) [bCourses] https://bids.berkeley.edu/news/predicting-protest-policing-research
8	2/17	wrap up on structured social data prediction & supervision Bayes and updating priors Lab 9: Folium Heat Maps	Adhikari and DeNero chs. 17-18 (review) 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); also at L.A. Times Static 99 actuarial risk assessment instrument *suggested*: Introduction to Statistical Learning ch. 4 [bCourses]

II. Big Data and the Problem of Crime

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 familiarity with critical perspectives on the role of prediction in the field of law (including probation and parole)

	date	**class meeting topic**	**reading to have prepared before class**
9	2/22	predictive instruments and the decision to punish Lab 10: Folium Plugins	Skeem & Lowenkamp 2016 "Risk, Race, & Recidivism" [bCourses] Barry-Jester, "Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?" [bCourses] *suggested*: Medium AUC-ROC explainer

10	2/24	litigating predictive models and predictive bias--COMPAS Lab 11: Math in Scipy	State of Wisconsin v. Loomis (pp. 1-31, *supplementary 31-48*) [bCourses] 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 http://andrewgelman.com/2018/06/06/average-predictive-comparisons-else-equal-fallacy/ **DATA INVESTIGATION PROJECT PROPOSAL (DUE FRI 2/26)**
11	3/1	more on COMPAS Lab 12: Regression for Prediction, Data Splitting	Flores et al 2017 "False Positives, False Negatives" (rejoinder to Angwin) [bCourses] Dressel and Farid (2018), "The accuracy, fairness, and limits of predicting recidivism," *Science Advances* 4:1 (17 Jan)
12	3/3	modeling risk machine versus human predictions Lab 13: Model Selection	Kleinberg et al. 2017 "Human Decisions and Machine Predictions" [bCourses] **PSET 1 [DUE FRI 3/5]**
13	3/8	a new physiognomy? thinking about what models are actually doing Lab 14: Feature 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]; NY Times 10 Jul 19 *for reference*: Daylight Security Research Lab "Machine Learning Failures" Python notebook collection
14	3/10	allocation of resources and models Lab 15: Text Preprocessing	Mohler et al. 2015 "Randomized Controlled Field Trials of Predictive Policing," *Jnl. Amer. Stat. Assn.* 110:1399-1411 [bCourses] 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	3/15	surveillance, selection, and the ratchet effect Lab 16: Intro to Text Analysis (BOW)	*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	3/17	remaining questions and discussion on predictions from "big data" Lab 17: Parse XML (Beautiful Soup)	Justin Grimmer and Brandon Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* pp. 1-31 [bCourses] Programming Historian on extracting and using Old Bailey Corpus and Beautiful Soup for scraping Computational Journalism on parsing HTML
17	3/29	selecting into a dataset: thinking critically about what data are collected Guest: Isaac Dalke, UC Berkeley Sociology Lab 18: Regular Expressions	Dalke 2020 "Insight and the Reconfiguration of Penal Practice in California" (ms), attention to methods appendices [bCourses] *optional*: 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]

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	3/31	understanding how Old Bailey Proceedings data got made content analysis of cases outline of computational text analysis techniques Lab 19: TF-IDF and Classification	"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] *optional*: Hall & Wright 2008, "Systematic Content Analysis of Judicial Opinions," 96 Cal. L. Rev. 63 [bCourses] **DIP EXPLORATORY DATA ANALYSIS (DUE 2 APRIL)**
19	4/5	Marx, history, and law as indicator or constitutive Guest: Professor David Lieberman	Hay, "Property, Authority, and the Criminal Law" *Albion's Fatal Tree* (New York: Pantheon, 1975, 17-63) [bCourses] Langbein, "Albion's Fatal Flaws" *Past & Present* 98 (Feb. 1983), 96-120

		Lab 20: Exploratory Data Analysis (feature extraction, visualizations, principal components analysis)	
20	4/7	TAR and the legal profession Guest: Dr. Lon Troyer, H5	Oard & Webber 2013, Information Retrieval for E-Discovery, "Introduction" & "The E-Discovery Process" [bCourses] Grossman & Cormack 2011, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient that Exhaustive Manual Review" [bCourses] *Rio Tinto v. Vale* (Peck Opinion) 2015 [bCourses] *In re: Broiler Chicken Antitrust Litigation* 2018 [bCourses] *supplementary:* Kluttz & Mulligan 2019 "Automated Decision Support Technologies and the Legal Profession" **PSET 2 (DUE 9 APRIL)**
21	4/12	the Old Bailey in its legal-historical context Lab 21: Neural Nets	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] *optional*: Klingenstein, Hitchcock, and DeDeo, 2014. "The Civilizing Process in London's Old Bailey," *PNAS* 111:26, 9419-9424. [bCourses] *optional*: 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	4/14	text as social science evidence 1 Lab 22: Word Embedding	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]

			Arseniev 2018 "Conceptual Intro to Word2Vec" [bCourses] *technical*: Rong 2014 "word2vec Parameter Learning Explained" [bCourses] *optional*: Bolukbasi et al 2016 "Man is to Computer as Woman is to Homemaker?" [bCourses] http://guides.lib.berkeley.edu/text-mining Harvard Case Law Project https://case.law/
23	4/19	text as social science evidence 2 Guest: Dr. Aniket Kesari, D-Lab Lab 23: Topic Models	Kesari "Privacy Legislation Text Re-Use Across State Legislatures" [bCourses]
24	4/21	text as social science evidence 3 Lab 24: Sentiment: Morality	Wu 2019 "MARMOT: A Deep Learning Framework for Constructing Multimodal Representations for Vision-and-Language Tasks," [bCourses] **DIP MODEL & EXPLANATION (DUE 4/23)**
25	4/26	text as social science evidence 4 Lab 25: Ensemble Methods	Alice Wu 2017 "Gender Stereotyping in Academia: Evidence from Economic Job Market Rumors Forum" [bCourses] (optional) Ethan Michelson 2019 "A Look Back at the Heyday of Political Activism" (paper prepared for WI Int'l Law J. Annual Symposium, 5 April) [bCourses]
26	4/28	Questions and wrap up Guest: Ilya Akdemir, Berkeley Law	Chen, Daniel L., 2019, "Machine Learning and the Rule of Law [bCourses] Bartlett, Robert P. et al. 2020, "Algorithmic Discrimination and Input Accountability under the Civil Rights Acts" [bCourses] Gordley, J., 2013 The jurists: a critical history. Chapter 7, "Mos Geometricus – the coming of rationalism" Oxford University Press [bCourses]

27	5/3	Instructor office hrs	*RRR week*
28	5/5	Instructor office hrs	*RRR week* **PROJECT NOTEBOOK (DUE MAY 7)** **PROJECT WEB PRESENTATIONS (DUE MAY 7)** **PSET 3 (DUE MAY 14)**

DATA INVESTIGATION PROJECT NOTEBOOK DUE FRIDAY 7 MAY

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 teams (the instructors will group 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. The project has a number of graded drafts before the final Jupyter notebook is due.

proposal	10%	**2/26**
exploratory data analysis	20%	**4/2**
modelling & explanation	20%	**4/23**
web presentation of findings	15%	**5/7**
complete notebook (including text and visualizations)	35%	**5/7**

Project teams should consult with the instructor, GSI, and Project 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. Example notebooks, posted with student permission, are here. Sample web presentations are here. You may use datasets that teams have used in the past, but please use them to ask a novel question.

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 to the bCourses assignments page.

Exploratory Data Analysis

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

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.

Web Presentation

Each team will be responsible for reporting the results of their Data Investigation Project in a markdown file that will be uploaded to a Github website accessible to the public (project groups can choose to remain anonymous if they wish). Project groups will hand the markdown file and associated files in a zipped folder on bCourses. The web presentation will be evaluated on writing quality, use of visualizations, clarity, and audience engagement.

Project Notebook

The written portion of the Data Investigation Project is a Python notebook that allows interested readers to reproduce the team's analysis (and so incorporates sufficient comments and markdown cells to explain what is going on) and includes a brief (no more than 2000 words total, in markdown cells) 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 (including clarity and usefulness of visualizations), and (as a bonus) originality and inventiveness.