Legal Studies 123
Data, Prediction, and Law
Spring 2020
MW 12 p.m..-2:00 p.m.
Barrows 110
GSI: Ilya Akdemir
Project UGSI: Wilson Berkow

Jon Marshall
jdmarshall [at] berkeley [dot] edu
2240 Piedmont Ave. Room 114
510-642-3670
Office hours M 2-4 PM, TU 2-3:30 PM (or by appt.)
CA's: Iland Leigh, Violet Yao

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 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, an online final exam, and class participation (which includes both attending and speaking).

problem sets (3 total)	45%
data investigation project	35%
reading reflections (2)	10%
class participation	10%

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 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 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. Let me know if you would prefer a class Piazza group and I will set one up.

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.

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. Labs are available in a Git repository at https://github.com/ds-modules/Legal-123-Sp20. 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.

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	have prepared before class
1	1/22	data, prediction, law	Gebru et al (2017) [bCourses]

		student questionnaire Lab 1: Anaconda setup if you want to run things locally (before class) Lab 2: Intro to Jupyter Notebooks	look at all three data sources for class (ANES 2016, SFPD Incident Reports, Old Bailey Proceedings)
2	1/27	data structures (incl. Pandas dataframes) Lab 3: Dataframe Operations & Simple Visualizations	Adhikari and DeNero, *Computational and Inferential Thinking*. chs. 3-5 https://www.inferentialthinking.com/chapters/intro ANES 2016 Codebook (pp. 3-7) [bCourses]
3	1/29	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 *suggested*: Introduction to Statistical Learning, ch. 2 [bCourses]
4	2/3	estimation & uncertainty, large N hypothesis testing Lab 5: Large n and hypothesis testing	Adhikari and DeNero ch. 11-14 Ferguson, *The Rise of Big Data Policing*, ch. 4 (pp. 62-83) [bCourses]
5	2/5	Guest: Rebecca Wexler, Berkeley Law litigating predictive models (COMPAS) correlation, OLS regression regression and causal inference Lab 6: OLS for Causal Inference	Adhikari and DeNero ch. 15-16 State of Wisconsin v. Loomis (pp. 1-31, *supplementary 31-48*) [bCourses] Wexler, "How Data Privacy Laws," LA *Times* 31 Jul 2019 [bCourses] *suggested*: Introduction to Statistical Learning, ch. 3 [bCourses]
6	2/10	Prediction in policing 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	2/12	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	2/19	wrap up on structured social data prediction & parole Bayes and updating priors Lab 9: Folium Heat Maps	Adhikari and DeNero chs. 17-18 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 *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 critical understanding of machine prediction

	date	**class meeting topic**	**reading to have prepared before class**
9	2/24	Guest: Stephanie Croft, HRC Investigations Lab 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]
10	2/26	predictive bias--COMPAS Lab 11: Math in Scipy	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 2/28)**
11	3/2	more on COMPAS Guest: Prof. Hany Farid 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	3/4	modeling risk machine versus human predictions Lab 13: Model Selection	Kleinberg et al. 2017 "Human Decisions and Machine Predictions" [bCourses] **PSET 1 [DUE FRI 3/6]**
13	3/9	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
14	3/11	allocation of resources and models Lab 15: Text Preprocessing	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	3/16	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	3/18	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
17	3/30	Guest: Isaac Dalke, Sociology selecting into a dataset: thinking critically about what data are collected Lab 18: Regular Expressions	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] Dalke 2020 "I Come Before You a Changed Man" (ms), pp. 1-14, 33-42, plus whatever else [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	4/1	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] Hall & Wright 2008, "Systematic Content Analysis of Judicial Opinions," 96 Cal. L. Rev. 63 [bCourses] **DIP Exploratory Data Analysis (due 7 April)**
19	4/6	Marx, history, and law as indicator or constitutive Lab 20: Exploratory Data Analysis (feature extraction, visualizations, principal components analysis)	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	4/8	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 16 April)**
21	4/13	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	4/15	wrap up on Old Bailey in its historical context Lab 22: Word Embedding	Arseniev 2018 "Conceptual Intro to Word2Vec" [bCourses] Rong 2014 "word2vec Parameter Learning Explained" [bCourses] http://guides.lib.berkeley.edu/text-mining Harvard Case Law Project https://case.law/
23	4/20	Guest: Aniket Kesari, JSP text as social science evidence Lab 23: Topic Models	Kesari SEC chapter [bCourses] *optional*: 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	4/22	Guest: Matt Cannon, JSP text as social science evidence 2 Lab 24: Sentiment: Morality	Cannon patent text mining piece [bCourses]
25	4/27	text as social science evidence 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] **DIP Model & Explanation (due 4/27)**
26	4/29	Questions and wrap up	
27	5/4	Instructor office hrs	*RRR week*

28	5/6	Instructor office hrs	*RRR week* **PROJECT NOTEBOOK (DUE MAY 8)** **PROJECT WEB PRESENTATIONS (DUE MAY 8)** **PSET 3 (DUE MAY 15)**

DATA INVESTIGATION PROJECT NOTEBOOK DUE FRIDAY 8 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.

proposal checkpoint	10%	**2/28**
exploratory data analysis checkpoint	20%	**4/3**
modelling checkpoint	20%	**4/27**
web presentation of findings	15%	**5/8**
write-up (including visualizations)	35%	**5/8**

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. Examples, posted with student permission, are here.

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 word total, in markup 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, and (as a bonus) originality and inventiveness.

Reading Reflections

Each student will be responsible for doing two Reading Reflections. Each Reading Reflection should be no more than 800 words and should be formatted in 12 pt type, double spaced, preferably as a .docx file. Students will upload their Reading Reflections files to bCourses, which will run them through Turnitin. Students will write the Reading Reflection paper on all of the required readings for the class meetings to which they have been assigned. Students may also refer to optional readings and outside readings if they choose to (as long as they cite those). Students assigned a Reading Reflection for a class meeting should turn it in on bCourses by noon on the day of class. In class, the students who had a Reading Reflection that day should be prepared to talk about it, but it will not be a formal presentation.

Since the readings for each class meeting are different, and sometimes have two opposing points of view, the following are to be understood as guidelines for the questions a Reading Reflection should answer.

1. What argument does each author make?
2. What evidence or supporting authority does each author use to support her/their/his argument?
3. Develop a critique of at least one author's argument (getting help from other readings is fine). Is that author's argument strong enough to withstand the criticism, or does it require revision?
4. Can you think of a way, or an additional way, to test the author's argument? Explain.

Reading Reflections will be evaluated based on how thoroughly they answer these questions and how thorough the critical evaluation of the readings is, as well as on organization, mechanics, and style.