|  |  |
| --- | --- |
| Legal Studies 123 | Jon Marshall |
| Data, Prediction, and Law | jdmarshall [at] berkeley [dot] edu |
| Spring 2024 | Office hours: Weds 2-3:30 pm (or by appt.) |
| MW 10 am - noon | GSI: Maura Lievano, office hrs MW 12-1 (or by appt.) |
| 110 Social Sciences Bldg. | CA’s: Ruby Chan Frey, Sashvitha Machani |

**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 and develop insights from publicly available data. The course proceeds along two tracks; the readings and discussion cover the use of prediction and its 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. Without this background, you may not enroll in this course.** Please see Dr. Marshall or Maura if you have not taken Foundations of Data Science to see if you have the background to do well in this class. This course pairs well with Data 100, and students may wish to take Data 100 before or concurrently with Data, Prediction, and Law.

**Learning objectives**

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

1. answer questions by using common statistical and computational techniques to analyze different types of “real world” data related to law (especially criminal law);
2. explain those answers using text and visualizations; and
3. 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 that are *individual* assessments of the student’s grasp of the course material,
* a data investigation project that is a *team* effort,
* lab exercises that teach techniques needed for the problem sets (evaluated for completion on a 0-2 scale)
* class participation (on a 0-3 scale for each session, which includes attending and speaking up but may also include other participation).

The third problem set will also serve as a final assessment for the class, and so will be worth 50% more than the other two problem sets. Work that is late (that is, without an extension in writing from the instructor in advance) will be penalized 3% per day. It will always pay off to turn in work that you have done to a decent level of quality. When you anticipate that you will not make a deadline, contact the instructor **and** GSI to request an extension.

problem sets (3 total) 45%

data investigation project 35%

class participation 10%

labs 10%

**Texts**

The readings for the course are entirely electronic and will either be available as a public document online or on the bCourses site for the course. 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 Python* (Springer, 2023). The book is available in [electronic form](https://www.statlearning.com/) and via the Library, and I will put a couple of relevant chapters on bCourses. The Data 100 website <https://ds100.org/> also has useful information about the data science tools and concepts we will cover in the labs. Questions to guide you through the readings will be posted on Ed Discussion.

**Policies**

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

Please be on time to class 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. 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 (including using the output of LLMs without attribution), cheat, or are otherwise dishonest, the default penalty will be a failing grade in the class, and I will file an academic dishonesty report.* I discourage the use of Large Language Models for answering questions or writing code; you will be more accurate and less wordy if you produce your own text, and you will understand your code if you look to human output on Stack Overflow, for example. If you have any questions about this, please ask.

Students requiring [accommodation](http://www.dsp.berkeley.edu/students) 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 and Piazza daily, especially since office hours may need occasional adjustments.

**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. I ask that you **start the labs before the class** meeting so that we can work on questions and problems and then discuss what we can take away from each lab. The labs will be graded for completion and will be due by 6 p.m. on class days. **The problem sets (and your projects) rely on the techniques you learn in the labs.** Lab solutions will be posted after the lab due date. Some students may be more familiar with the Python code, or the methods, that each lab features, so my hope is that you help each other out and teach one another what you know. One person in each project team will be “on call” to talk about the lab, either to the whole class at the start of the lab period or to the other students in their project team during lab. This will part of the participation grade.

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, and there will be a page on bCourses with the interact links to pull the labs into your directory on Datahub. **That will sync what is in your Datahub directory with what is in the Git repository, so if you want to save your work and your notes therein, it is a good idea to change its filename or it may be overwritten if you click the link again.** Note that some labs are still being revised, so you will need to pull the labs from Github again once revisions are done. 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 Data Investigation Projects have relied on Google Colab, which allows you to run your notebook and collaborate (on Google’s servers). This class has an expanded quota on Datahub, so our labs should run fine there.

**I. Social Science Data, Generalization, and Policing**

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

1. use Jupyter notebooks to write clear code in Python
2. use Pandas and plotting tools to clean, organize, summarize, and visualize data
3. show critical understanding of prediction in policing

|  |  |  |  |
| --- | --- | --- | --- |
|  | **date** | **class meeting topic** | **prepare before class** |
| 1 | 1/17 | Data, prediction, law  Examples of big data inference  Lab 1: Anaconda setup to run things locally  Lab 2: Intro to Jupyter Notebooks | ripped from the headlines: [regulating AI by Executive Order](https://www.nytimes.com/2023/10/30/us/politics/biden-ai-regulation.html) [bCourses], [US missing a chance to prevent AI harm](https://www.nytimes.com/2023/12/06/technology/ai-regulation-policies.html) [bCourses], [EU regulating AI](https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html) [bCourses]  [Buolamwini, J. and Gebru, T](http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline). (2018). “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification” *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77-91. <https://proceedings.mlr.press/v81/buolamwini18a.html> [bCourses]  look at data sources for class ([SFPD Incident Reports](https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783), [NY State DCJS](https://ww2.nycourts.gov/pretrial-release-data-33136), [Old Bailey Proceedings](https://www.oldbaileyonline.org/))  *optional*: with your data you need a [side of social science](https://statmodeling.stat.columbia.edu/2023/01/03/explanation-and-reproducibility-in-data-driven-science-new-course/) [bCourses] |
| 2 | 1/22 | Pandas dataframes and data types, flow of control, state of objects, etc.  Prediction in the legal domain  Lab 3: Pandas Dataframe Operations & Data Cleaning | student questionnaire (Google Form)  Harcourt, *Against Prediction* ch. 1 [bCourses]  *optional*: Adhikari and DeNero, *Computational and Inferential Thinking* chs. 3-5 (review) <https://www.inferentialthinking.com/chapters/intro> |
| 3 | 1/24 | Summary stats and data visualization  Lab 4: Visualizing, Summarizing, Aggregating Data | Adhikari and DeNero chs. 7, 9-10 (review)  *suggested***:** [Introduction to Statistical Learning](https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html), ch. 2 [bCourses] |
| 4 | 1/29 | Estimation & uncertainty, large N  Classical hypothesis testing  Lab 5: Central Limit Theorem, Confidence Intervals, Hypothesis Testing | Kleinberg et. al. 2018, “Discrimination in the Age of Algorithms” [bCourses]  Adhikari and DeNero ch. 11-14 (review)  *optional*: Steinberger 2020, “[Does Palantir See Too Much?](https://www.nytimes.com/interactive/2020/10/21/magazine/palantir-alex-karp.html)” NY Times Magazine 21 Oct. [bCourses]  *optional*: Brayne 2018, “[The Criminal Law and Law Enforcement Implications of Big Data](https://www.annualreviews.org/doi/pdf/10.1146%2Fannurev-lawsocsci-101317-030839),” *Ann. Rev. of Law Soc. Sci*., 14:293–308 [bCourses] |
| 5 | 1/31 | Police, prosecutors, and crime control  Correlation, OLS regression, causal inference, levels of measurement, units of observation/analysis  Uniform Crime Reports dataset (Maura)  Lab 6: OLS for Causal Inference | National Academy of Science 2018 “[Proactive Policing: Effects on Crime and Communities](http://nap.nationalacademies.org/24928)” [Summary](https://nap.nationalacademies.org/read/24928/chapter/2) (pp. 1-13) [bCourses]  Chalfin and McCrary 2013 “The Effect of Police on Crime” NBER Working Paper 18815 (pp.1-4, 10-15, 22-29, 37-45) [bCourses]  Gelman 2023 “[Debate over effect of reduced prosecutions on urban homicides](https://statmodeling.stat.columbia.edu/2023/10/12/debate-over-effect-of-reduced-prosecutions-on-urban-homicides-also-larger-questions-about-synthetic-control-methods-in-causal-inference/)”  Adhikari and DeNero ch. 15-16 (review)  *suggested***:** Introduction to Statistical Learning, ch. 3 [bCourses]  *text methods for future reference*: [CTA labs](https://github.com/ilyaakdemir/ls123-intro-to-cta) (courtesy Ilya) |
| 6 | 2/5 | Surveillance with Big Data  Guests: Professor Sarah Brayne (UT Austin) and Professor Karen Levy  Lab 7: Introduction to Folium (mapping) | Brayne, Lageson, and Levy 2023 “Surveillance deputies: When ordinary people surveil for the state” *Law Soc Rev*. 57(4):462-488 DOI: 10.1111/lasr.12681  *optional*: Brayne 2017, “[Big Data Surveillance: The Case of Policing](https://doi.org/10.1177%2F0003122417725865),” Amer. Soc. Review 82:5, 977-1008 [bCourses] |
| 7 | 2/7 | Police allocation of resources and reporting  SFPD incident report data and its application  Lab 8: Folium [Choropleth](https://towardsdatascience.com/creating-choropleth-maps-with-pythons-folium-library-cfacfb40f56a) Maps | <https://www.pbs.org/wgbh/frontline/film/policing-the-police-2020/>  [SFPD Incident report data](https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry)  Ang et al 2015 “San Francisco Crime Classification” (grad student project) [bCourses]  *optional*: Mohler et al. 2015 “[Randomized Controlled Field Trials of Predictive Policing](https://doi.org/10.1080/01621459.2015.1077710),” *Jnl. Amer. Stat. Assn.* 110:1399-1411 [bCourses]  *optional*: O’Flaherty and Sethi 2010 “The Racial Geography of Street Vice” *Jnl Experimental Criminology* |
| 8 | 2/12 | Surveillance, selection, and the ratchet effect  Data selection: thinking critically about what data are collected  Lab 9: Folium Heat Maps | Harcourt 2007 *Against Prediction* ch. 5 [bCourses]  [*Floyd v. City of New York*](https://scholar.google.com/scholar_case?case=390056199313197546&q=floyd+v.+city+of+new+york&hl=en&as_sdt=2006&as_vis=1) 959 F. Supp. 2d 540 - Dist. Court, SD New York 2013), pp. 556-576 [bCourses corresponding pages 1-37] (and whatever else interests you)  Sankin et al 2021 “[Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them](https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them)” Markup 2 Dec.  *optional*: Khorshidi 2021, Tucker 2021  *optional*: [methods and data](https://github.com/the-markup/investigation-prediction-bias) for Markup piece on PredPol; [Twitter discussion](https://twitter.com/JuliaAngwin/status/1466397179992543236) from Julia Angwin  *suggested*: Adhikari and DeNero chs. 17-18 (review)  *suggested***:** Introduction to Statistical Learning ch. 4 [bCourses]  *optional*: Fryer (2016) “[An Empirical Analysis of Racial Differences in Police Use of Force](http://www.nber.org/papers/w22399)” (pp. 1-7) and its [discussion](http://andrewgelman.com/2016/07/14/about-that-claim-that-police-are-less-likely-to-shoot-blacks-than-whites/) and [follow-up](http://andrewgelman.com/2016/08/03/30833/) on Andrew Gelman’s blog [bCourses] |

**II. Data and the Decision to Punish**

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

1. explain the power and pitfalls of data in making predictions in the criminal adjudication and punishment processes
2. show an understanding of and be able to apply different classification models
3. apply critical perspectives to data generating processes and to uses of prediction in the field of law (including pretrial release, probation, and parole)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **date** | **class meeting topic** | **reading to have prepared before class** |
| 9 | 2/14 | Using and litigating predictive models in deciding punishment  Lab 10: Folium Plugins | Skeem & Lowenkamp 2016 “Risk, Race, & Recidivism” [bCourses]  [State of Wisconsin v. Loomis](https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690) (pp. 1-31, *suggested 31-48*) [bCourses]  *optional*: Feeley & Simon 1992 “The New Penology” *Criminology* (**30**:4) pp. 449-474 [bCourses] |
| 10 | 2/21 | Predictive instruments and bias, and detecting bias  Lab 11: Data cleaning and EDA for classification | Angwin et al 2016 “[Machine Bias](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)” ProPublica 23 May [bCourses]  Larson et al 2017 “[How We Analyzed the COMPAS Recidivism Algorithm](https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm)” (appendix) [bCourses] (their [data](https://github.com/propublica/compas-analysis))  *suggested*: [Medium AUC-ROC explainer](https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb)  *suggested*: <http://andrewgelman.com/2018/06/06/average-predictive-comparisons-else-equal-fallacy/>  **Data Investigation Project Proposal (due Fri 2/23)** |
| 11 | 2/26 | Yet more on COMPAS  Lab 12: Classification using logistic regression | Flores et al 2017 “False Positives, False Negatives” (rejoinder to Angwin) [bCourses]  Chouldechova 2017 [bCourses]  [Dressel and Farid](http://advances.sciencemag.org/content/4/1/eaao5580) 2018, “The accuracy, fairness, and limits of predicting recidivism,” *Science Advances* 4:1 (17 Jan) |
| 12 | 2/28 | Question and work it out time  Lab 13: Performance metrics for classifiers | **PSet 1 [due Fri 3/1]** |
| 13 | 3/4 | Modeling risk  Machine versus human predictions  Lab 14: More classification models & feature selection | Kleinberg et al. 2017 “Human Decisions and Machine Predictions” [bCourses]  New York City Criminal Justice Agency [Release Assessment Instrument](https://www.nycja.org/ra-court-forms)  *optional*: Dobbie and Yang 2019 “Proposals for Improving the U.S. Pretrial System” Brookings [bCourses] |
| 14 | 3/6 | A new physiognomy?  Thinking about what models are actually doing  Lab 15: Text Preprocessing | 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](https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html)  *for reference*: Daylight Security Research Lab “[Machine Learning Failures](https://daylight.berkeley.edu/mlfailures/)” Python notebook collection |
| 15 | 3/11 | Prediction & supervision decisions—the Brock Turner case *(trigger warning: the case materials discuss a rape case on the Stanford campus)*  digital stigma  Lab 16: Intro to Text Analysis (BOW) | Bay Area News Group materials on Scribd from Brock Turner case ([survivor’s statement (ex. 16)](https://www.scribd.com/doc/315399999/People-v-Brock-Turner-Document-6), [probation report](https://www.documentcloud.org/documents/2858997-Probation-officer-s-report-in-Brock-Turner-case.html)) (see also [police report](https://www.scribd.com/doc/315399976/People-v-BrockTurner-Documents1), [character letters](https://www.scribd.com/doc/315400014/People-v-Brock-Turner-Document-8), [complaint](https://www.scribd.com/doc/315399984/People-v-Brock-Turner-Document-2), [sentencing memo](https://www.scribd.com/doc/315428494/People-vs-Brock-Turner-Documents-7)); Judge Persky [sentencing memo](https://www.theguardian.com/us-news/2016/jun/14/stanford-sexual-assault-read-sentence-judge-aaron-persky); also at [L.A. Times](http://documents.latimes.com/stanford-brock-turner/)  [Static 99](http://www.static99.org/) actuarial risk [assessment instrument](http://www.static99.org/pdfdocs/Static-99R_coding_form.pdf) [bCourses]  *optional*: Lageson 2020, [*Digital Punishment* ch. 3](https://academic.oup.com/book/33571/chapter/288035546) [bCourses] |
| 16 | 3/13 | Guest: Dorna Mohaghegh, JSP  Algorithmic neutrality and the law  Lab 17: Parse XML (Beautiful Soup) | Programming Historian on [extracting and using Old Bailey Corpus](https://programminghistorian.org/lessons/naive-bayesian) and [Beautiful Soup](https://programminghistorian.org/lessons/intro-to-beautiful-soup) for scraping  Mohaghegh 2023 “Jurisprudence in the Age of Algorithms” [bCourses]  *optional:* [Computational Journalism on parsing HTML](mailto:http://www.compjour.org/warmups/govt-text-releases/intro-to-bs4-lxml-parsing-wh-press-briefings/)  *optional*: Halevy et al. 2009 “The Unreasonable Effectiveness of Data” [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 software packages (e.g. scikit-learn) 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** |
| 17 | 3/18 | text as data  Lab 18: Regular Expressions | Grimmer, Roberts, and Stewart 2022 *Text as Data* chs. 2-4 (pp. 14-47) [bCourses] |
| 18 | 3/20 | Understanding how Old Bailey Proceedings data got made  Lab 19: TF-IDF and Classification | “[About the Proceedings](https://www.oldbaileyonline.org/static/Proceedings.jsp),” “[Historical Background to the Proceedings of the Old Bailey](https://www.oldbaileyonline.org/static/History.jsp) (esp. “[Crime, Justice, and Punishment](https://www.oldbaileyonline.org/static/Crime.jsp)”) on Old Bailey Corpus site <https://www.oldbaileyonline.org/> [bCourses]  **DIP Exploratory Data Analysis (due 22 March)** |
| 19 | 4/1 | 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 Flaws”](https://libproxy.berkeley.edu/login?qurl=https://www.jstor.org/stable/650689) *Past & Present* 98 (Feb. 1983), 96-120 |
| 20 | 4/3 | 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]  *optional*: 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](https://ebookcentral.proquest.com/lib/berkeley-ebooks/reader.action?ppg=131&docID=201627&tm=1515624823288)] [bCourses]  **PSet 2 (due Friday 5 April)** |
| 21 | 4/8 | Guest: Prof. David Hausman, Berkeley Law | Hausman 2023. “Risk Assessment Software Change as Policy Change in Immigration Detention Decisions”  *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 | 4/10 | Text as social science evidence 1  Lab 22: Word Embedding | Arseniev 2018 “Conceptual Intro to Word2Vec” [bCourses]  Dalke 2020 “Insight and the Reconfiguration of Penal Practice in California” (ms) with attention to methods appendices [bCourses]  <http://guides.lib.berkeley.edu/text-mining>  *technical & optional*: Rong 2014 “word2vec Parameter Learning Explained” [bCourses]  *optional*: Bolukbasi et al 2016 “[Man is to Computer as Woman is to Homemaker?](https://arxiv.org/abs/1607.06520)” [bCourses]  *resources*: Harvard Case Law Project <https://case.law/>, <http://guides.lib.berkeley.edu/text-mining> |
| 23 | 4/15 | Text as social science evidence 2  Lab 23: Topic Models | Data Investigation Project check-in |
| 24 | 4/17 | Text as social science evidence 3  Lab 24: Sentiment Analysis: Moral Foundations Dictionary | Grimmer, Roberts, and Stewart 2022 Text as Data, chs 25-27 (pp. 272-293)  *optional*: Wu 2019 “[MARMOT: A Deep Learning Framework for Constructing Multimodal Representations for Vision-and-Language Tasks](https://arxiv.org/abs/2109.11526),” [bCourses]  **DIP Model & Explanation (due Mon. 22 April)** |
| 25 | 4/22 | Text as social science evidence 4 | 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 | 4/24 | Questions and wrap up  Guest: Maura Lievano, GSPP | Lievano TBD |
|  | 4/29 | Instructor office hrs in regular class room & time | *RRR week* |
|  | 5/1 | Instructor office hrs in regular class room & time | *RRR week*  **Project Notebook (due May 3)**  **Project Web Presentations (due May 3)**  **PSet 3 (due May 10)** |

**Data Investigation Project**

The DIP gives students the opportunity to undertake 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 pieces before the final Jupyter notebook is due.

|  |  |  |
| --- | --- | --- |
| proposal | 10% | **2/23** |
| exploratory data analysis | 15% | **3/22** |
| modelling & explanation | 20% | **4/22** |
| web presentation of findings | 15% | **5/3** |
| complete notebook (including text and visualizations) | 40% | **5/3** |

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 or (even better) 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](https://lucid.app/publicSegments/view/62dcb4af-7b22-4831-9dc6-01f0ef8c9e61/image.jpeg)). 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](https://github.com/jdmarshl/LS_123_Projects). Sample web presentations are [here](https://legalst123.github.io/) and [here](https://legalst123.github.io/sp21/). 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

1. your research question
2. a list of at least five references that deal with the same or similar questions; each item should have a brief summary (no more than five sentences) of the empirical methods and findings
3. the data you plan to use and how it will help you answer your question (with a discussion of the unit of analysis for the project)
4. 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. The EDA must include a) a summary table that clearly identifies the outcome variable(s) and gives descriptive statistics for the variables in the dataset; b) a correlation matrix or contingency table that describes the relationship between the variables in the dataset; c) a brief list of the main takeaways from the EDA. The EDA should also include at least three visualizations that will help your audience understand your data.

**Modeling**

Modeling 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 final product 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. Include any visualizations that will help readers understand the data and your analysis. The final notebook should also have an attribution list at the end which details which team member was responsible for each part of the project. 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.