Algorithm Biases: Fall 2021 Special Study Proposal

Form Questions:

1. **Description of proposed special studies: (Provide a one to two paragraph description of the proposed study.)**

This special study will give me the opportunity to learn the skills to determine if discrimination against different subgroups exist in specific machine learning algorithms, learn how that bias is formed into common machine learning algorithms, and ways to mitigate that bias. This has the real-world benefit of developing the skills to notice and respond if an algorithm treats protected classes or other population subgroups differently than the majority population. The special study begins with researching different definitions of fairness since there currently is not a consensus in the machine learning community on what it means for an algorithm to be "fair."

Just like in the Statistical and Data Sciences capstone, the mid-semester assignments for the special study will include seven "blog post" style writings. I will post these writings on my personal website alongside my existing Statistical and Data Sciences capstone blog posts. The final project will be a teachable unit on algorithm ethics, written for students in SDS 293 (Modeling for Machine Learning).

2. **Readings and Source Materials**

See the week-by-week plan (Table 1) below for the title of each week's academic article readings. Each article's abstract is included in Table 2, in no particular order.

3. **Nature of Final Project(s) to be Evaluated/Graded: (Bibliography and/or paper and/or creative work and/or poster, etc.)**

The final project will be a teachable unit on ethics written at an undergraduate machine learning class level. This would include a lab (and an answer key) written in R or Python. This lab will be written at the challengingness level of an SDS 293 class and turned in on the last day of classes.

Table 1: Week-by-week plan

9/2	**Definitions of fairness** Academic paper reading: On Formalizing Fairness in Prediction with Machine Learning Deliverable: submit the special study application
9/6	**Definitions of fairness** Academic paper reading: 50 Years of Test (Un)fairness: Lessons for Machine Learning Deliverable: Blog 1 a blog post breaking down what I believe are the most foundational and important definitions of algorithm fairness. Pick three fairness definitions to use to look at the different algorithm types. Acknowledge how my own background influences my view and values.
9/13	**How algorithms can create discrimination as a result of missing data** Academic paper reading: How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications Deliverable: Blog 2 Answer how algorithms can create discrimination as a result of missing data, including times when it may be advantageous to include or omit sensitive data.
9/20	Catch up week.
9/27	**Alg type: k-means, k-means, and k-medians** Academic paper reading: Socially Fair k-Means Clustering; Fair Clustering Through Fairlets
10/4	**Alg type: Nearest Neighbors** Academic paper reading: Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning Deliverable: Blog 3A summarize the main points of how fairness interacts with the very popular k-means and related k-algorithms. Include the real-world impact and if any solutions exist.
10/11	**Alg Type: Deeper into to DkNN idea** Review the DkNN since I really struggled with my initial reading about it. Deliverable: Presentation about DkNN;

	Blog Post 3B about DkNN
10/18	**Alg Type: Deeper into to DkNN idea** Review the DkNN since I really struggled with my initial reading about it. Deliverable: Presentation about DkNN; Blog Post 3B about DkNN
10/25	**Halfway point check-in.** **Alg type: Making sure you have quality data to feed into your algorithm** Academic paper reading: Mitigating Bias in Set Selection with Noisy Protected Attributes; Mitigating Unwanted Biases with Adversarial Learning Deliverable: Blog 4 talk about how quality data collection and acknowledging the weaknesses of your data can be a tool to prevent an algorithm's discrimination against subgroups. Brainstorm the teachable unit on algorithm ethics.
11/1	**Alg type: Anomaly detection** Academic paper reading: Towards Fair Deep Anomaly Detection
11/8	**Review ICML 2021.** Search for papers that fit in the Kiri Wagstaff model for ML. Skim those papers. Work on the teachable unit on algorithm ethics. Perhaps do the postponed Nonparametric kernel regression/weighted sum: **Alg type: Nonparametric kernel regression / weighted sum** Academic paper reading: Deep Weighted Averaging Classifiers; Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models
11/15	Deliverable: Blog 6 blog detailing what to look for when auditing algorithms. Consider a blog post about how to check if your algorithm is treating certain groups unfairly based on simple statistics.
11/22	Work on the teachable unit on algorithm ethics.
11/29	Work on the teachable unit on algorithm ethics.
12/6	Work on the teachable unit on algorithm ethics. Be done with it, ideally.

	Further ideas that didn’t have time during the semester: a blog post about the cost to implement fairness and explaining what papers are referring to when they discuss the "cost" of implementing fairness measures. “Cost” could be additional computing time, additional number of calculations, financial price to run a data warehouse training a model, extra weeks of salary paid to data scientists to develop a more fair model, or opportunity cost of implementing a more accurate but less fair model.

Table 2: Reading abstracts for all the papers listed above as well as some related potentially supplemental readings

Differential Privacy Has Disparate Impact on Model Accuracy	Differential privacy (DP) is a popular mechanism for training machine learning models with bounded leakage about the presence of specific points in the training data. The cost of differential privacy is a reduction in the model's accuracy. We demonstrate that in the neural networks trained using differentially private stochastic gradient descent (DP-SGD), this cost is not borne equally: accuracy of DP models drops much more for the underrepresented classes and subgroups. For example, a gender classification model trained using DP-SGD exhibits much lower accuracy for black faces than for white faces. Critically, this gap is bigger in the DP model than in the non-DP model, i.e., if the original model is unfair, the unfairness becomes worse once DP is applied. We demonstrate this effect for a variety of tasks and models, including sentiment analysis of text and image classification. We then explain why DP training mechanisms such as gradient clipping and noise addition have disproportionate effect on the underrepresented and more complex subgroups, resulting in a disparate reduction of model accuracy.
On Formalizing Fairness in Prediction with Machine Learning	Machine learning algorithms for prediction are increasingly being used in critical decisions affecting human lives. Various fairness formalizations, with no firm consensus yet, are employed to prevent such algorithms from systematically discriminating against people based on certain attributes protected by law. The aim of this article is to survey how fairness is formalized in the machine learning literature for the task of prediction and present these formalizations with their corresponding notions of distributive justice from the social sciences literature. We provide theoretical as well as empirical critiques of these notions from the social sciences literature and explain how these critiques limit the suitability of the corresponding fairness formalizations to certain domains. We also suggest two notions of distributive justice which address some of these critiques and discuss avenues for prospective fairness formalizations.
Equal Protection Under the Algorithm: A Legal-Inspired Framework for Identifying Discrimination in Machine Learning	Within the field of ethical machine learning, an area of special concern is the possibility of machine learning algorithms discriminating against groups of people in unethical ways, such as targeting advertisements based on race. In this paper, we propose a framework based on long-standing U.S. legal principles to determine whether the targeting of a group should be viewed with suspicion. Unlike existing work, we are focused on the case when the group is not correlated with known

	‘protected features’, or such data is unavailable.
Human Decisions and Machine Predictions [see PDF emailed to you by the library]	Can machine learning improve human decision making? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released. The concreteness of the prediction task combined with the volume of data available makes this a promising machine-learning application. Yet comparing the algorithm to judges proves complicated. First, the available data are generated by prior judge decisions. We only observe crime outcomes for released defendants, not for those judges detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. Second, judges may have a broader set of preferences than the variable the algorithm predicts; for instance, judges may care specifically about violent crimes or about racial inequities. We deal with these problems using different econometric strategies, such as quasi-random assignment of cases to judges. Even accounting for these concerns, our results suggest potentially large welfare gains: one policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates. Moreover, all categories of crime, including violent crimes, show reductions; these gains can be achieved while simultaneously reducing racial disparities. These results suggest that while machine learning can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.
Socially Fair k-Means Clustering	We show that the popular k-means clustering algorithm (Lloyd’s heuristic), used for a variety of scientific data, can result in outcomes that are unfavorable to subgroups of data (e.g., demographic groups). Such biased clusterings can have deleterious implications for human-centric applications such as resource allocation. We present a fair k-means objective and algorithm to choose cluster centers that provide equitable costs for different groups. The algorithm, Fair-Lloyd, is a modification of Lloyd’s heuristic for k-means, inheriting its simplicity, efficiency, and stability. In comparison with standard Lloyd’s, we find that on benchmark datasets, Fair-Lloyd exhibits unbiased performance by ensuring that all groups have equal costs in the output k-clustering, while incurring a negligible increase in running time, thus making it a viable fair option wherever k-means is currently used.
LOGAN: Local Group Bias Detection by Clustering	Machine learning techniques have been widely used in natural language processing (NLP). However, as revealed by many recent studies, machine learning models often inherit and amplify the societal biases in data. Various

	metrics have been proposed to quantify biases in model predictions. In particular, several of them evaluate disparity in model performance between protected groups and advantaged groups in the test corpus. However, we argue that evaluating bias at the corpus level is not enough for understanding how biases are embedded in a model. In fact, a model with similar aggregated performance between different groups on the entire data may behave differently on instances in a local region. To analyze and detect such local bias, we propose LOGAN, a new bias detection technique based on clustering. Experiments on toxicity classification and object classification tasks show that LOGAN identifies bias in a local region and allows us to better analyze the biases in model predictions.
Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning	Deep neural networks (DNNs) enable innovative applications of machine learning like image recognition, machine translation, or malware detection. However, deep learning is often criticized for its lack of robustness in adversarial settings (e.g., vulnerability to adversarial inputs) and general inability to rationalize its predictions. In this work, we exploit the structure of deep learning to enable new learning-based inference and decision strategies that achieve desirable properties such as robustness and interpretability. We take a first step in this direction and introduce the Deep k-Nearest Neighbors (DkNN). This hybrid classifier combines the k-nearest neighbors algorithm with representations of the data learned by each layer of the DNN: a test input is compared to its neighboring training points according to the distance that separates them in the representations. We show the labels of these neighboring points afford confidence estimates for inputs outside the model's training manifold, including on malicious inputs like adversarial examples– and therein provides protections against inputs that are outside the models understanding. This is because the nearest neighbors can be used to estimate the nonconformity of, i.e., the lack of support for, a prediction in the training data. The neighbors also constitute human-interpretable explanations of predictions. We evaluate the DkNN algorithm on several datasets, and show the confidence estimates accurately identify inputs outside the model, and that the explanations provided by nearest neighbors are intuitive and useful in understanding model failures.
How Algorithms Discriminate Based	Organizations often employ data-driven models to inform decisions that can have

on Data They Lack: Challenges, Solutions, and Policy Implications	a significant impact on people's lives (e.g., university admissions, hiring). In order to protect people's privacy and prevent discrimination, these decision-makers may choose to delete or avoid collecting social category data, like sex and race. In this article, we argue that such censoring can exacerbate discrimination by making biases more difficult to detect. We begin by detailing how computerized decisions can lead to biases in the absence of social category data and in some contexts, may even sustain biases that arise by random chance. We then show how proactively using social category data can help illuminate and combat discriminatory practices, using cases from education and employment that lead to strategies for detecting and preventing discrimination. We conclude that discrimination can occur in any sociotechnical system in which someone decides to use an algorithmic process to inform decisionmaking, and we offer a set of broader implications for researchers and policymakers.
Fair Algorithms for Clustering	We study the problem of finding low-cost {\em fair clusterings} in data where each data point may belong to many protected groups. Our work significantly generalizes the seminal work of Chierichetti \etal (NIPS 2017) as follows. - We allow the user to specify the parameters that define fair representation. More precisely, these parameters define the maximum over- and minimum under-representation of any group in any cluster. - Our clustering algorithm works on any ℓp-norm objective (e.g. k-means, k-median, and k-center). Indeed, our algorithm transforms any vanilla clustering solution into a fair one incurring only a slight loss in quality Our algorithm also allows individuals to lie in multiple protected groups. In other words, we do not need the protected groups to partition the data and we can maintain fairness across different groups simultaneously. Our experiments show that on established data sets, our algorithm performs much better in practice than what our theoretical results suggest.
Fair Clustering Through Fairlets	We study the question of fair clustering under the disparate impact doctrine, where each protected class must have approximately equal representation in every cluster. We formulate the fair clustering problem under both the k-center and the k-median objectives, and show that even with two protected classes the problem is challenging, as the optimum solution can violate common conventions---for instance a point may no longer be assigned to its nearest cluster center! En route we introduce the concept of fairlets, which are

	minimal sets that satisfy fair representation while approximately preserving the clustering objective. We show that any fair clustering problem can be decomposed into first finding good fairlets, and then using existing machinery for traditional clustering algorithms. While finding good fairlets can be NP-hard, we proceed to obtain efficient approximation algorithms based on minimum cost flow. We empirically demonstrate the \emph{price of fairness} by quantifying the value of fair clustering on real-world datasets with sensitive attributes.
Towards Fair Deep Anomaly Detection	Anomaly detection aims to find instances that are considered unusual and is a fundamental problem of data science. Recently, deep anomaly detection methods were shown to achieve superior results particularly in complex data such as images. Our work focuses on deep one-class classification for anomaly detection which learns a mapping only from the normal samples. However, the non-linear transformation performed by deep learning can potentially find patterns associated with social bias. The challenge with adding fairness to deep anomaly detection is to ensure both making fair and correct anomaly predictions simultaneously. In this paper, we propose a new architecture for the fair anomaly detection approach (Deep Fair SVDD) and train it using an adversarial network to de-correlate the relationships between the sensitive attributes and the learned representations. This differs from how fairness is typically added namely as a regularizer or a constraint. Further, we propose two effective fairness measures and empirically demonstrate that existing deep anomaly detection methods are unfair. We show that our proposed approach can remove the unfairness largely with minimal loss on the anomaly detection performance. Lastly, we conduct an in-depth analysis to show the strength and limitations of our proposed model, including parameter analysis, feature visualization, and run-time analysis.
Mitigating Bias in Set Selection with Noisy Protected Attributes	Subset selection algorithms are ubiquitous in AI-driven applications, including, online recruiting portals and image search engines, so it is imperative that these tools are not discriminatory on the basis of protected attributes such as gender or race. Currently, fair subset selection algorithms assume that the protected attributes are known as part of the dataset. However, protected attributes may be noisy due to errors during data collection or if they are imputed (as is often the case in real-world settings). While a wide body of work addresses the effect of noise on the performance of machine learning algorithms, its effect on fairness remains largely unexamined. We find that in the presence of noisy protected attributes, in attempting to increase fairness without considering noise, one can, in fact, decrease the fairness of the result! Towards addressing this, we consider an existing noise model in which there is probabilistic information about the protected attributes (e.g., [19, 32, 44, 56]), and ask is fair selection possible under noisy conditions? We formulate a "denoised" selection problem which

	functions for a large class of fairness metrics; given the desired fairness goal, the solution to the denoised problem violates the goal by at most a small multiplicative amount with high probability. Although this denoised problem turns out to be NP-hard, we give a linear-programming based approximation algorithm for it. We evaluate this approach on both synthetic and real-world datasets. Our empirical results show that this approach can produce subsets which significantly improve the fairness metrics despite the presence of noisy protected attributes, and, compared to prior noise-oblivious approaches, has better Pareto-tradeoffs between utility and fairness.
Leave-one-out Unfairness	We introduce leave-one-out unfairness, which characterizes how likely a model's prediction for an individual will change due to the inclusion or removal of a single other person in the model's training data. Leave-one-out unfairness appeals to the idea that fair decisions are not arbitrary: they should not be based on the chance event of any one person's inclusion in the training data. Leave-one-out unfairness is closely related to algorithmic stability, but it focuses on the consistency of an individual point's prediction outcome over unit changes to the training data, rather than the error of the model in aggregate. Beyond formalizing leave-one-out unfairness, we characterize the extent to which deep models behave leave-one-out unfairly on real data, including in cases where the generalization error is small. Further, we demonstrate that adversarial training and randomized smoothing techniques have opposite effects on leave-one-out fairness, which sheds light on the relationships between robustness, memorization, individual fairness, and leave-one-out fairness in deep models. Finally, we discuss salient practical applications that may be negatively affected by leave-one-out unfairness.
Deep Weighted Averaging Classifiers	Recent advances in deep learning have achieved impressive gains in classification accuracy on a variety of types of data, including images and text. Despite these gains, however, concerns have been raised about the calibration, robustness, and interpretability of these models. In this paper we propose a simple way to modify any conventional deep architecture to automatically provide more transparent explanations for classification decisions, as well as an intuitive notion of the credibility of each prediction. Specifically, we draw on ideas from nonparametric kernel regression, and propose to predict labels based on a weighted sum of training instances, where the weights are determined by distance in a learned instance-embedding space. Working within the framework of conformal methods, we propose a new measure of nonconformity suggested by our model, and experimentally validate the accompanying theoretical expectations, demonstrating improved transparency, controlled error rates, and robustness to out-of-domain data, without compromising on accuracy or calibration.
50 Years of Test	Quantitative definitions of what is unfair and what is fair have been

(Un)fairness: Lessons for Machine Learning	introduced in multiple disciplines for well over 50 years, including in education, hiring, and machine learning. We trace how the notion of fairness has been defined within the testing communities of education and hiring over the past half century, exploring the cultural and social context in which different fairness definitions have emerged. In some cases, earlier definitions of fairness are similar or identical to definitions of fairness in current machine learning research, and foreshadow current formal work. In other cases, insights into what fairness means and how to measure it have largely gone overlooked. We compare past and current notions of fairness along several dimensions, including the fairness criteria, the focus of the criteria (e.g., a test, a model, or its use), the relationship of fairness to individuals, groups, and subgroups, and the mathematical method for measuring fairness (e.g., classification, regression). This work points the way towards future research and measurement of (un)fairness that builds from our modern understanding of fairness while incorporating insights from the past.
Mitigating Unwanted Biases with Adversarial Learning	Machine learning is a tool for building models that accurately represent input training data. When undesired biases concerning demographic groups are in the training data, well-trained models will reflect those biases. We present a framework for mitigating such biases by including a variable for the group of interest and simultaneously learning a predictor and an adversary. The input to the network X, here text or census data, produces a prediction Y, such as an analogy completion or income bracket, while the adversary tries to model a protected variable Z, here gender or zip code. The objective is to maximize the predictor's ability to predict Y while minimizing the adversary's ability to predict Z. Applied to analogy completion, this method results in accurate predictions that exhibit less evidence of stereotyping Z. When applied to a classification task using the UCI Adult (Census) Dataset, it results in a predictive model that does not lose much accuracy while achieving very close to equality of odds (Hardt, et al., 2016). The method is flexible and applicable to multiple definitions of fairness as well as a wide range of gradient-based learning models, including both regression and classification tasks.
Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models	Predictive models are increasingly deployed for the purpose of determining access to services such as credit, insurance, and employment. Despite potential gains in productivity and efficiency, several potential problems have yet to be addressed, particularly the potential for unintentional discrimination. We present an iterative procedure, based on orthogonal projection of input attributes, for enabling interpretability of black-box predictive models. Through our iterative procedure, one can quantify the relative dependence of a black-box model on its input attributes.The relative significance of the inputs to a predictive model can then be used to assess the fairness (or discriminatory extent) of such a model.