CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Fairness Mini-Bootcamp Course Plan

Developed by the University of California, Berkeley's Center for Long-Term Cybersecurity (CLTC), this document details a curriculum for an Algorithmic Fairness Bootcamp, a week-long curriculum designed to teach students how to detect, identify, discuss and address bias in real-world machine learning algorithms.

ABOUT

The course includes two lectures introducing machine learning bias and fairness, along with labs that delve into how these algorithms are situated in larger social contexts, prompting students to discuss who designs these algorithms, who uses them, and who gets to decide what it means for algorithms to be working properly.

The Algorithmic Fairness Mini-Bootcamp has been designed to be integrated into any existing course on machine learning. The course consists of two lectures and one lab (as well as an optional second lab), all focused on teaching students how to identify and ameliorate bias in machine-learning algorithms. Links to slides, videos, and other resources are included.

The curriculum draws upon real-world examples: Lab 1 focuses on a medical risk-scoring algorithm that exhibits a bias against Black patients, while Lab 2 walks students through a hiring algorithm that exhibits a bias against women. Each Lab comes in two versions: one that requires knowledge of the Python programming language, and one that is programming-free.

These labs address a shortcoming in current computer science training, as many students may graduate and begin work as data scientists without having learned about bias or fairness in machine learning. The Algorithmic Fairness Mini-Bootcamp is designed to train the next generation of students to identify, discuss, and address the risks posed by machine learning algorithms in a variety of contexts.

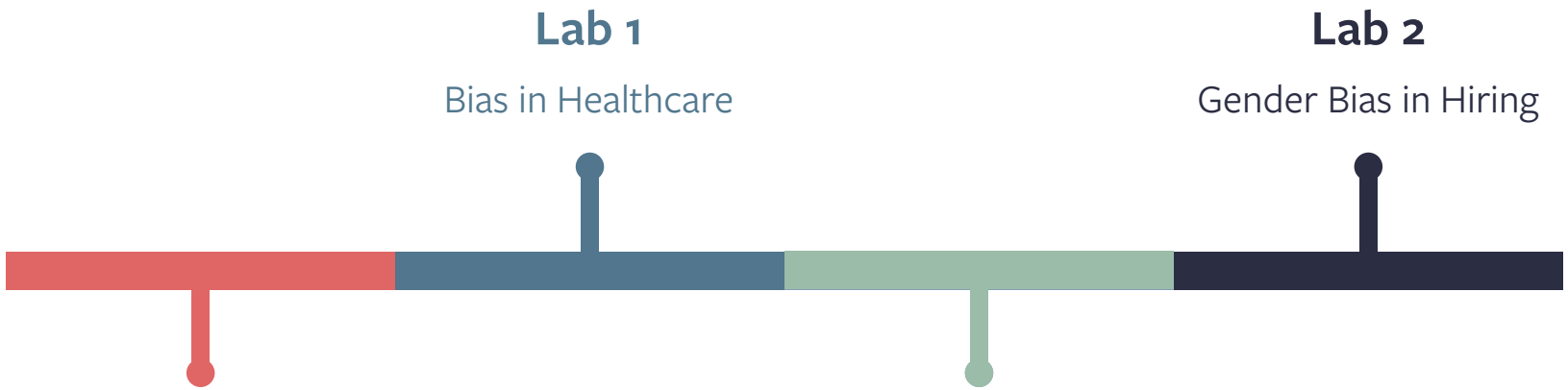
This course plan is designed to help students answer such questions such as: If you have an algorithm in front of you, how do you know if that algorithm is biased? How is it biased? And what do those biases mean in practice?

We encourage you to review these materials, and customize them to your liking. (The labs are licensed under the Creative Commons CC BY-NC-SA 4.0 license, so you can copy and modify them for non-commercial use; simply attribute the original source.)

We would be grateful if you would let us know how you end up using these materials, and we welcome your questions, ideas, and other feedback. We are eager to improve this offering.

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Bootcamp Timeline



CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Algorithmic Fairness Bootcamp Curriculum

Lecture 1: Identifying bias

The initial lecture provides an overview of machine learning failures, with examples of how algorithms can make unanticipated or undesirable decisions. The slides include examples of how protected classes, such as race, may be violated by ML Failures.

OVERVIEW

Machine learning algorithms have endless potential benefits, as they can automate the work of programming for a variety of complex tasks, from self-driving cars to medical diagnoses. However, as numerous research studies have shown, such algorithms are prone to bias (including racial and gender bias), often stemming from biased training data. These errors can be difficult to detect, as AI is often based on "black boxes" that lack explainability. Currently, not all programmers, computer scientists, or even machine learning practitioners are equipped to identify machine learning failures.

Click to access Lecture 1 slides

Topics addressed: Understanding Algorithmic Bias:

- What is the problem?
- Why is it a big deal?
- How does it happen?
- What can we do?

Key ideas:

- **Machine learning algorithms can exhibit bias against people whose characteristics have served as the basis for systematically unjust treatment in the past.** This bias can emerge for a variety of reasons, and can be so severe as to be illegal.
- **Bias in machine learning algorithms is both a social and a technical problem.** There are no technical "fixes," though technical tools can help us identify bias and reduce its harmfulness.
- **Do NOT remove sensitive data (like race and gender) from the training set.** That makes it difficult to know when your algorithm is biased.

Topics for breakout discussions & activities:

- What are some examples of machine learning bias you've heard about?
- What are some uses for machine learning where bias is a particular concern?
- In the healthcare example, who got to decide when the algorithm was fair enough?
- Why is it dangerous to remove race (or other sensitive attributes) from your training set?

Recommended assigned reading before lecture:

- **Obermeyer, Powers, Vogeli and Mullainathan. 2019.** Dissecting racial bias in an algorithm used to manage the health of populations. Science.

Consider requiring students to post a reaction to the reading (e.g., on Canvas) before the lecture.

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Lab 1: Identifying racial bias in a health care algorithm

To effectively manage patients, health systems often need to estimate particular patients' health risks. Using quantitative measures, or "risk scores," healthcare providers can prioritize patients and allocate resources to patients who need them most. In this lab, students examine an algorithm widely-used in industry to establish quantitative risk scores for patients. They will discover how this algorithm embeds a bias against Black patients, undervaluing their medical risk relative to White patients.

RESOURCES

- Watch a video walkthrough of this lab with a student. This video serves as a primer on the lab and its main takeaways, and may be a useful resource for TAs or others who are leading lab sections.
- Get the labs
  - Version with Python programming
    - See this primer, which covers the programming knowledge required to complete this lab.
  - Version with no Python programming
  - Answer key

  Recommendation: Assign Lab 1 after Lecture 1. Host a lab section covering it. Make Lab 1 due before Lecture 2.

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Lecture 2: Ameliorating bias

Approaches to ameliorating bias.

Click to access Lecture 2 slides

Topics covered:

- What's the lingo? (Terminology for discussing bias and its impact)
- What metrics can describe bias? (Quantitative indicators of bias)
- How do we make biased algorithms less harmful? (Technical strategies to ameliorate bias)
- What's next? (Issues in machine learning beyond fairness and bias)

Key ideas:

- Terminology of **privileged/unprivileged groups.**
- **Disparate impact** (and other metrics) can help describe the extent of bias toward privileged groups.
- **Fairness constraints** (and other methods) can help ameliorate the impact of bias, even with a fixed dataset.
- **Fairness is not a solely technical problem.** No technical "fix" can "solve" bias in machine learning. **These strategies are tools for making bias less harmful.**
- **Issues outside of bias and fairness** include algorithmic accountability, transparency, environmental consequences.

Topics for breakout discussions & activities:

- What algorithms do you personally interact with that might exhibit bias?
- How does disparate impact compare to the 4/5ths rule in US law (discussed in lecture 1)? How can we use disparate impact to evaluate the legality of an algorithm? What are some limitations of the 4/5ths rule?
- For lab 2: if we remove gender from the training set, the algorithm will still earn a bias against women. How? Why?
- For lab 2: it is often said that there is a tradeoff between fairness and accuracy. However, one can also say that the data is incorrectly labeled due to managers' bias. Compare and contrast these two ways of framing our ways for correcting bias.

Recommended reading:

- **Reading: Mulligan, Kroll, Kohli & Wong. 2019.** This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction.

Consider requiring students to post a reaction to the reading (e.g., on Canvas or BCourses) before the lecture.

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

(Optional) Lab 2: Ameliorating gender bias in a hiring algorithm

Various companies (including Amazon) have attempted to use machine learning to automate hiring decisions. However, when these algorithms are trained on past hiring decisions, they are likely to learn human biases: in this case, encoding a pay gap between men and women.From a social and ethical standpoint, we want to remove or minimize this bias so that our models are not perpetuating harmful stereotypes or injustices. In this lab, we take a dataset in which prior hiring decisions have adversely impacted women, and show how applying fairness constraints can ameliorate the effect of this impact, making it less harmful. We also prompt students to consider when, whether, and to what extent machine learning ought to be applied to hiring decisions.

RESOURCES

- Get the labs
  - Version with Python programming
    - See this primer, which covers the programming knowledge required to complete this lab.
  - Version with no Python programming
  - Answer key

Recommendation: Assign Lab 2 after Lecture 2. Due date can be flexible to your course's needs..