Responsible Machine Learning

Lecture 1: Intro

CS 4973-05

Fall 2023

Instructor: Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University, Boston, MA

Northeastern University

evijit.io

adept ID

Avijit Ghosh

- Research Data Scientist at AdeptID
- Areas of study include:
  - Audits of real systems,
  - Measurement and evaluation of algorithmic fairness, and
  - Challenges of incorporating algorithmic fairness interventions in the real world.

jlgleason.github.io

Jeffrey Gleason

- 3rd-year PhD Candidate working with Christo Wilson
- Areas of study include:
  - Algorithm auditing (controlled experiments and causal inference)
  - Platform power (e.g. Google + Amazon self-preferencing)

Please Introduce Yourself

- What made you want to take this class?
- What problems are you excited to solve?

Class Schedule

- **Class: Ryder 277**

  Tuesdays 11:45am - 1:25pm

  Thursdays 2:50pm - 4:30pm

- **Office Hours:**

  Avijit: Thursdays 5pm-6pm

  Jeffrey: Wednesdays 4pm - 5pm

Class Format

- **Reading assignments:** 30%
- **Coding assignments:** 15%
- **Quizzes:** 10%
- **Final project + Term paper:** 45%

Reading Assignments

I will assign portions of research papers to read to prepare for next day's lecture. A brief reading assignment quizlet will accompany the readings. You are required to upload your answers to the reading assignment questions **BEFORE** class starts.

Helpful guide: How to read a research paper

Coding Assignments

Less frequent than reading assignments, few and far between. Expected language: Python 3, preferably on Jupyter notebooks

Quizzes

Short, multiple choice quizzes, one for midterm and one for finals. Prep level needed low as long as you were paying attention in class.

Final Project + Term paper

The most important component of this class. You are expected to work on a final project in groups of two. This project involves three components:

1. Reading research papers/blogs on the topic you have chosen
2. Downloading relevant data and writing code to get results
3. A term paper describing what you did. The format is usually: *Introduction > Related work > Methods > Results > Conclusion > Limitations and Future Work*
    - Overleaf Latex Format

We will help you in every step of the way! While this is due at the end of the semester, it is good to start planning early and keep me updated on what you are doing.

AI is powerful


"do you guys ever think about BIAS?"
[record scratch]

What does Responsible ML mean?

What does Responsible ML mean?

- Fairness/Bias
- Explainability
- Transparency
- Privacy
- Safety
- Regulations/Policy

What does Responsible ML mean?

- Fairness/Bias
- Explainability
- Transparency
- Privacy
- Safety
- Regulations/Policy

Northeastern University

Predictive systems are on the rise. Machine learning and AI technologies have generated enormous benefits for business.

Credits: Chris Walker - Data Science @ Illuminate Education

How Artificial Intelligence Is Revolutionizing the Legal Practice

Vol. 43 No. 1

Paige E. K

The autho

BOTS GUEST

How A.I. is revolutionizing today's workplace

ADELYN ZHOU, TOPBOTS

The Marriage of Artificial Intelligence (AI) In Sports is Revolutionizing the Sector

Date : 0

Source : In

13D Research Follow

Navigating complexity in a rapidly-changing world. For more from What I Learned This Week, go to: ...

Apr 14 · 6 min read

Artificial Intelligence is on the precipice of revolutionizing medical diagnosis.

AI will propel new winners in the tech and healthcare sectors. And most importantly of all, it will save lives.

Northeastern University

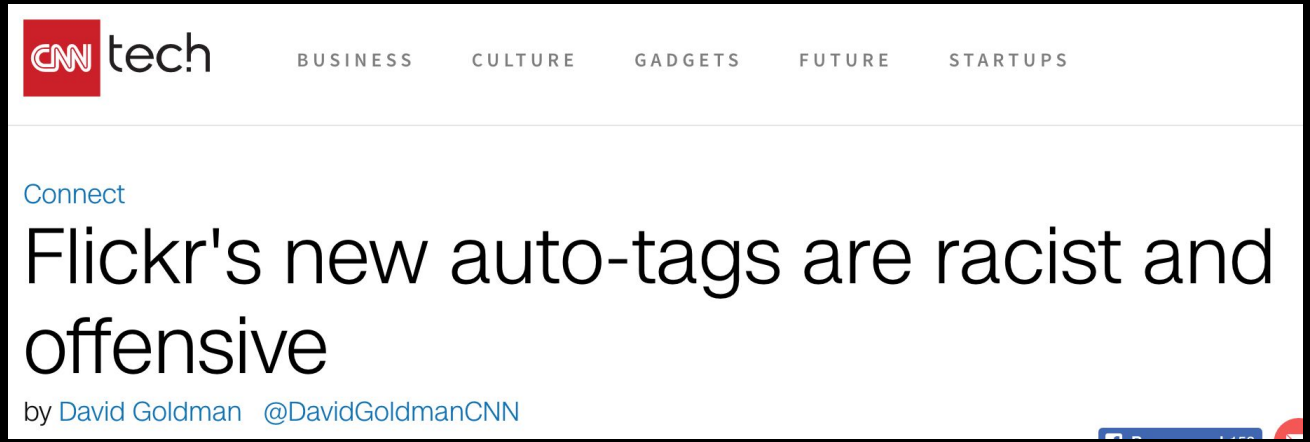
A Provocation:
Should we trust a machine to make the right decisions about a person's future?

Credits: Chris Walker - Data Science @ Illuminate Education

Skepticism is healthy.

Businesses have learned some painful lessons about the dangers of machines creating their own algorithms.

Credits: Chris Walker - Data Science @ Illuminate Education



Northeastern University

Credits: Chris Walker - Data Science @ Illuminate Education

How do algorithms become biased?

Credits: Chris Walker - Data Science @ Illuminate Education

Let's play spot the bias...

```
if person.ethnicity == 'African-American':
        credit.deny()
else:
        credit.grant()
```

Credits: Chris Walker - Data Science @ Illuminate Education

The bias is... Intentional. Obvious.

(and horrible!)

```
if person.ethnicity == 'African-American':
        credit.deny()
else:
        credit.grant()
```

Credits: Chris Walker - Data Science @ Illuminate Education

Before machine learning, algorithms were written only by humans.

Only a malicious developer would write such a rule.

But problems like these are easy to fix.

Credits: Chris Walker - Data Science @ Illuminate Education

Can you spot the bias here?

```
if person.zip_code == 38131:
        credit.deny()
else:
        credit.grant()
```

Northeastern University

Credits: Chris Walker - Data Science @ Illuminate Education

Seems harmless... until you learn that zip code 38131 is nearly 100% African-American

```
if person.zip_code == 38131:
        credit.deny()
else:
        credit.grant()
```

Credits: Chris Walker - Data Science @ Illuminate Education

The bias is... Maybe Intentional. Less obvious.

```
if person.zip_code == 38131:
        credit.deny()
else:
        credit.grant()
```

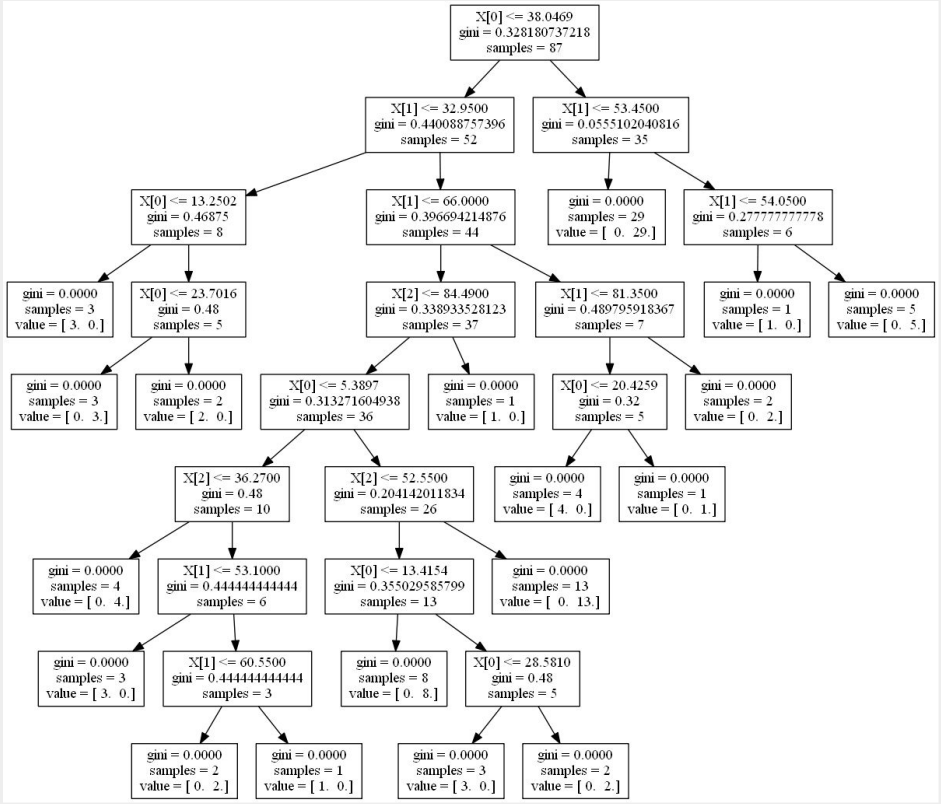
Credits: Chris Walker - Data Science @ Illuminate Education

A statistical model might suggest such a rule — less obvious, but equally bad.

A principled developer or statistician can still catch these if they are vigilant. This is still relatively easy to fix.

Credits: Chris Walker - Data Science @ Illuminate Education

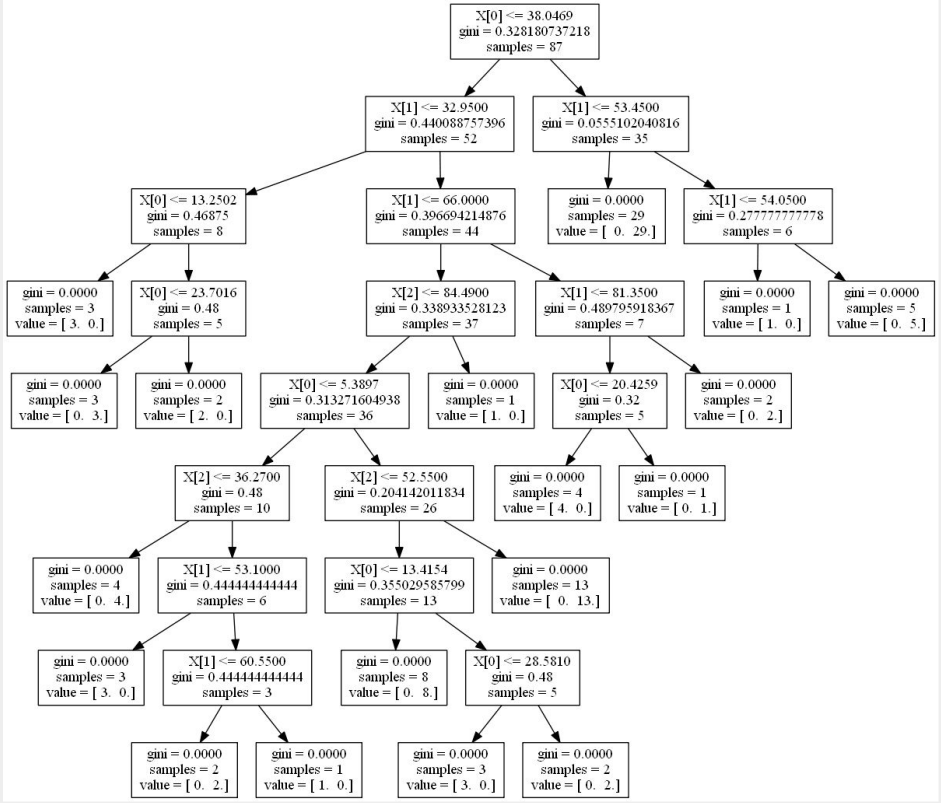
What about this one?



Credits: Chris Walker - Data Science @ Illuminate Education

Northeastern University

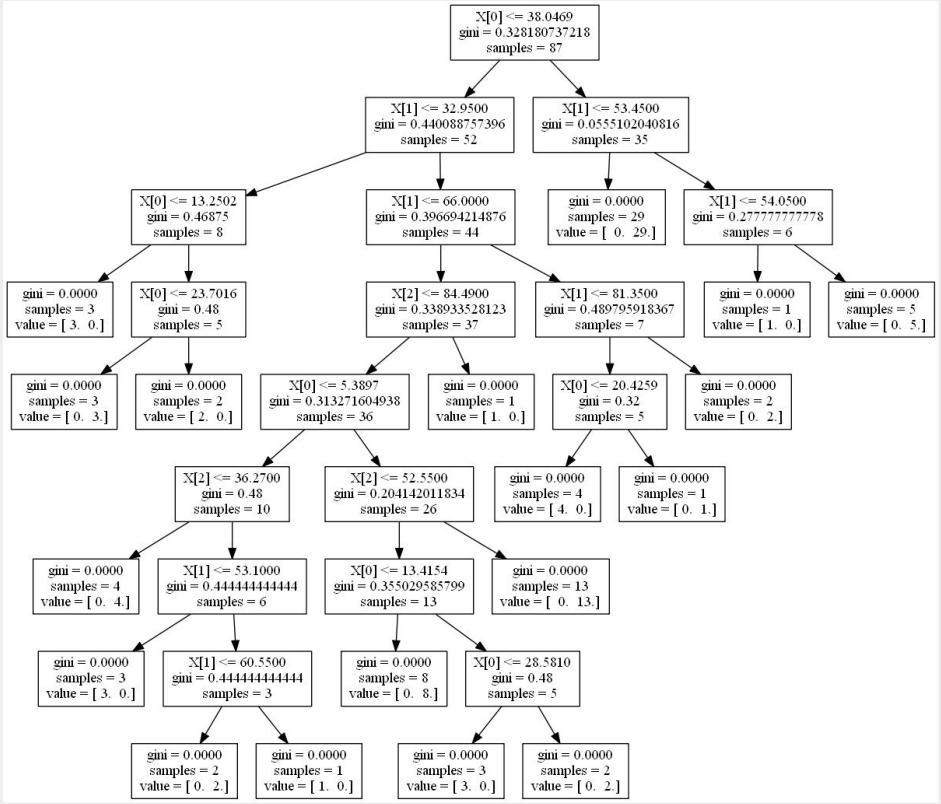
This is a small section of a decision tree built with machine learning



Credits: Chris Walker - Data Science @ Illuminate Education

Northeastern University

The bias is... Unintentional. And far less obvious.



Credits: Chris Walker - Data Science @ Illuminate Education

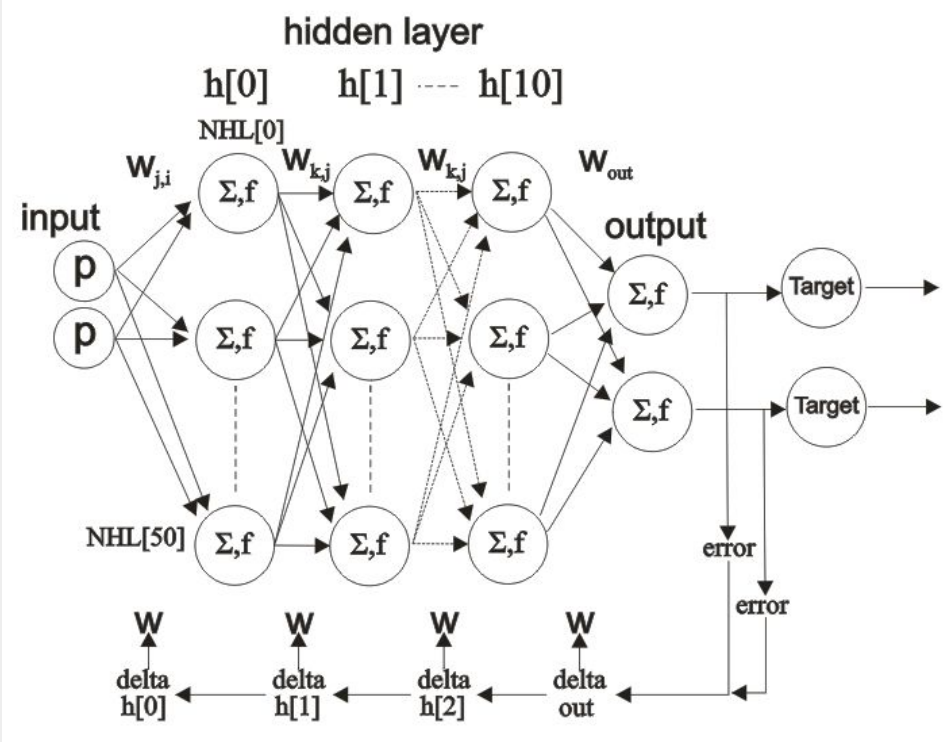
Northeastern University

Now our algorithm is made of hundreds of decision points. Each one might be biased.

Combinations of decision points might be biased too. That's millions of possibilities! Not so easy to fix.

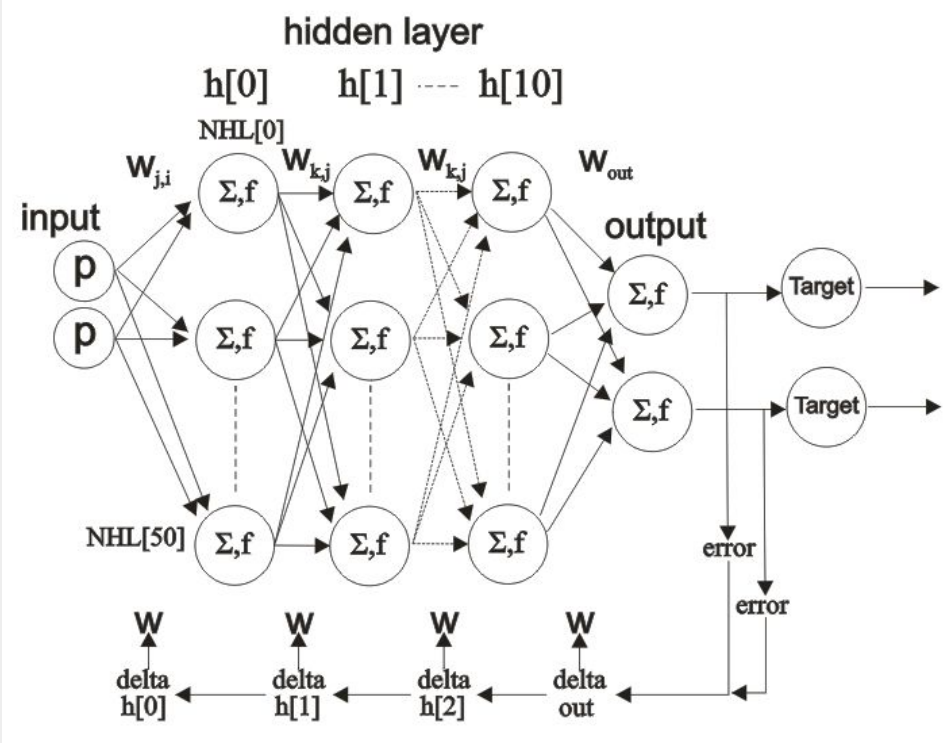
Credits: Chris Walker - Data Science @ Illuminate Education

Okay, last one...



Credits: Chris Walker - Data Science @ Illuminate Education

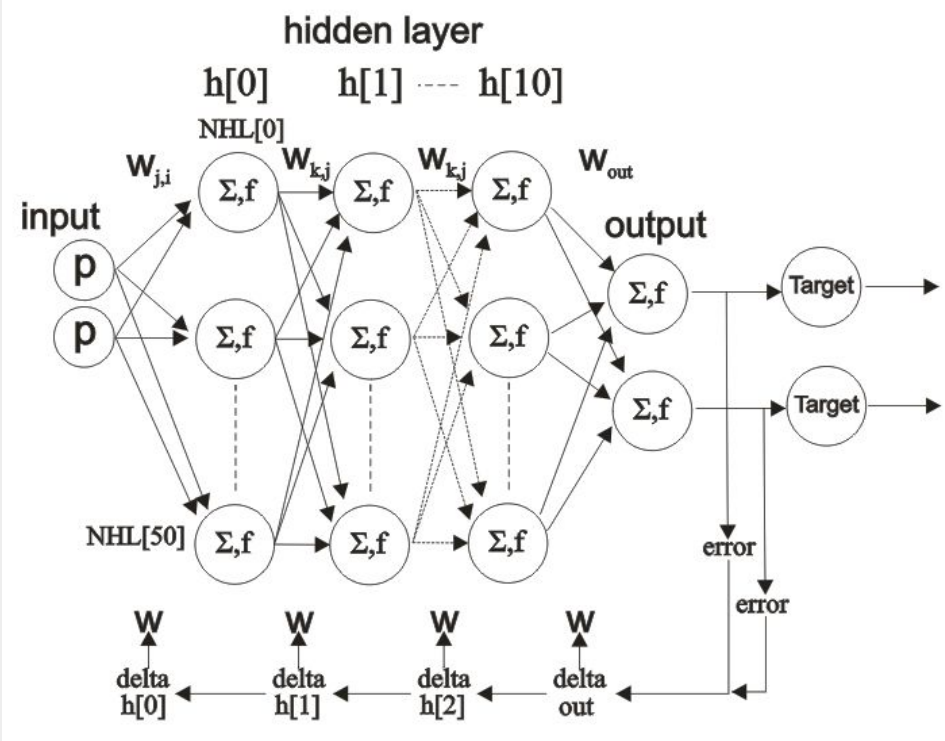
Welcome to Artificial Intelligence (AI)



Credits: Chris Walker - Data Science @ Illuminate Education

Northeastern University

The bias is... Unintentional. Totally obscured.



Credits: Chris Walker - Data Science @ Illuminate Education

Another Provocation:
If the builder of a system can't spot the bias, what hope do we have of correcting it?

Credits: Chris Walker - Data Science @ Illuminate Education

Yet Another Provocation:

How accurately do you think vendors report their system's capabilities?

Credits: Chris Walker - Data Science @ Illuminate Education

Principles Worth Defending

Fairness

How can a computer judge a person on something they haven't even done yet?

Transparency

How exactly was the decision made? What data were used?

Remediation

To whom does a person complain when things go sideways?

How can we make it right?

Northeastern University

Credits: Chris Walker - Data Science @ Illuminate Education

Thank You!

Readings for Next Class:

- Machine learning: Trends, perspectives, and prospects
  - M. I. Jordan and T. M. Mitchell