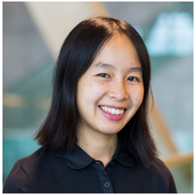
Human-Centered Machine Learning

Dong Nguyen

2021

Utrecht University

Hello!



Dong Nguyen

Heysem Kaya

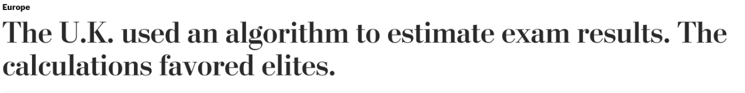
Yupei Du

Announcements

This lecture will be *recorded*.

Questions or comments? Raise your hand or use the chat.

ML in society



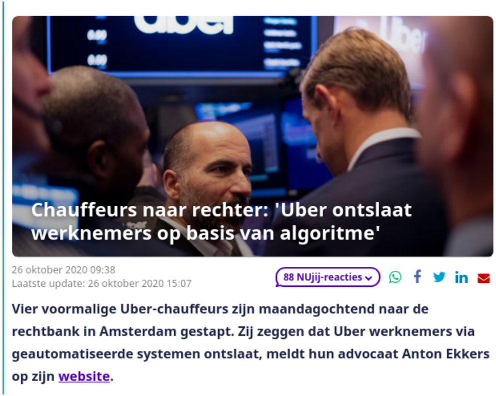
Europe

The U.K. used an algorithm to estimate exam results. The calculations favored elites.

Figure: Washington Post, 18 August 2020

Figure: The Guardian, 20 August 2020, Photograph: Matthew Chattle/Rex/Shutterstock

ML in society



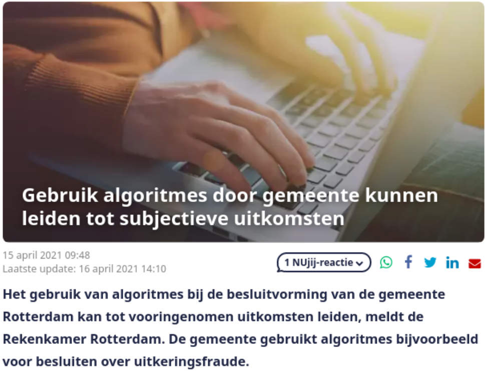
Chauffeurs naar rechter: 'Uber ontslaat werknemers op basis van algoritme'

26 oktober 2020 09:38
Laatste update: 26 oktober 2020 15:07

Vier voormalige Uber-chauffeurs zijn maandagochtend naar de rechtbank in Amsterdam gestapt. Zij zeggen dat Uber werknemers via geautomatiseerde systemen ontslaat, meldt hun advocaat Anton Ekkers op zijn website.

Figure: NU.nl, 26 October 2020

ML in society



Gebruik algoritmes door gemeente kunnen leiden tot subjectieve uitkomsten

15 april 2021 09:48
Laatste update: 16 april 2021 14:10

Het gebruik van algoritmes bij de besluitvorming van de gemeente Rotterdam kan tot vooringenomen uitkomsten leiden, meldt de Rekenkamer Rotterdam. De gemeente gebruikt algoritmes bijvoorbeeld voor besluiten over uitkeringsfraude.

Figure: NU.nl, 16 April 2021

New AI regulation from the EU?

Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)—April 21, 2021 [link]

“*The Commission is proposing the first ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.*”

“Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.”

New AI regulation from the EU?

Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)—April 21, 2021 [link]

“*The Commission is proposing the first ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.*”

“High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of this Title.”

New AI regulation from the EU?

Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)—April 21, 2021 [link]

“*The Commission is proposing the first ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.*”

“High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations (‘feedback loops’) are duly addressed with appropriate mitigation measures”

New AI regulation from the EU?

Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)—April 21, 2021 [link]

What are high-risk AI systems?

- AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;
- AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;
- AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending or the risk for potential victims of criminal offences;
- ...

Human-Centered Machine Learning

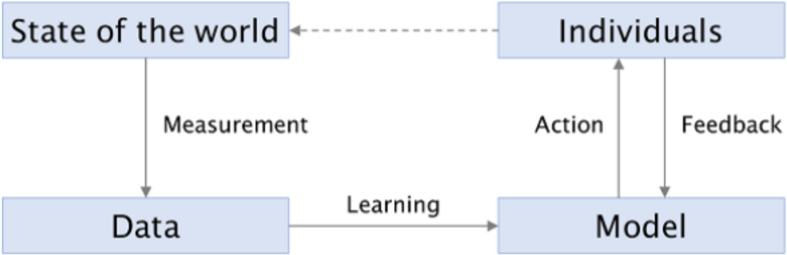
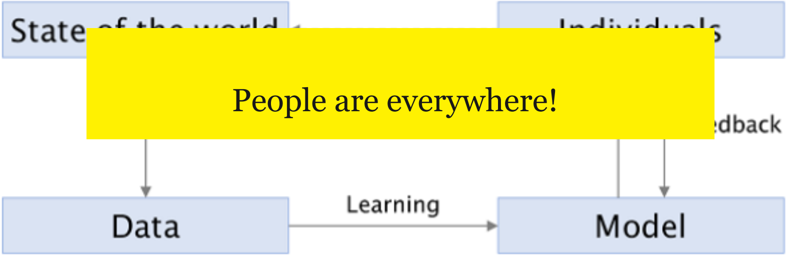


Figure: Fig 1 from Fairness ML book [link]

Human-Centered Machine Learning



People are everywhere!

Figure: Fig 1 from Fairness ML book [link]

Human-Centered Machine Learning

We'll focus on two themes:

- Fairness (Dong)
- Explainability (Heysem)

Course Organisation

Background information

Prerequisites

- Proficiency in Python (exercises are in Python, project: your choice!)
- Basic machine learning and statistics

And…

- You will read academic articles, in depth.
- Be “comfortable” with engaging with a fast-paced, emerging research field and research challenges for which there no definite solutions.

Schedule

First part:

- Fairness
  - Lectures, programming exercises, paper presentation
- Explainability
  - Lectures, programming exercises, paper presentation
- Midterm 10th of June (mark your calendars!)

Schedule

First part:

- Fairness
  - Lectures, programming exercises, paper presentation
- Explainability
  - Lectures, programming exercises, paper presentation
- Midterm 10th of June (mark your calendars!)

Second part:

- Group project
- Guest lectures

Schedule

First part:

- Fairness
  - Lectures, programming exercises, paper presentation
- Explainability
  - Lectures, programming exercises, paper presentation
- Midterm 10th of June (mark your calendars!)

Second part:

- Group project
- Guest lectures

Check the syllabus (Blackboard) for all important dates!

Paper presentations

- There will be two paper presentations: one on fairness, one on explainability.
- Each time you will be in a different group.
- Groups of 4. *Randomly* assigned.
- Papers can be selected from a list provided by us. Groups can indicate preferences.

Programming assignments

- Two programming assignments (fairness, explainability)
- Pass or fail
- Pairs
- Python
- The first exercise will be released next Tuesday. For that one, we'll be using AI Fairness 360.
- For the second programming assignment we'll be using interpret.ml.

Project

- *Self-assigned* groups
- Majority of the work will be *after* the midterm, but group composition and project idea needs to be submitted end of week 5.
- Project updates + project office hours.
- Project presentations (Gather town) + project report.

Feedback

It’s the first time we’re teaching this course. Feedback is very welcome. Help us improve the course!

As usual, there’ll be a **Caracal evaluation** at the end. In addition:

- We’ll send around a short feedback form halfway of the course.

Todo (now!)

Fill in the short background questionnaire. **Deadline: Tomorrow 5pm**

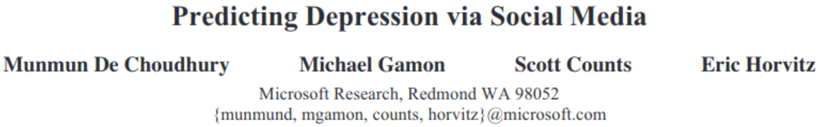
Go to Blackboard → Course Content → Programming Assignments → Background questionnaire

Fairness

Roadmap

- **Today**: Introduction, sources of unfairness, harms
  - Role of data
  - Model development
- **Lecture 2**: Measuring the fairness of ML systems
- **Lecture 3**: Making ML systems more fair + broader perspective

Dual use: Should I build this system?



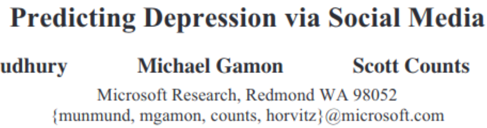
Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

“We explore the potential to use social media to detect and diagnose major depressive disorder in individuals.”

Predicting Depression via Social Media, De Choudhury et al., 2013 [link]

Dual use: Should I build this system?



Predicting Depression via Social Media

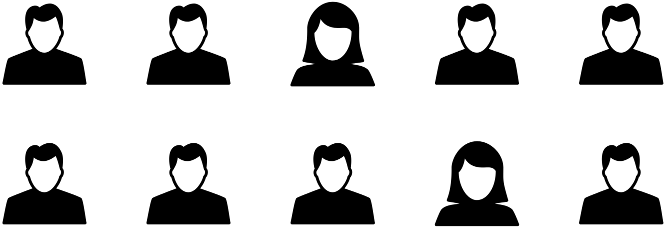
Munmun De Choudhury Michael Gamon Scott Counts Eric Horvitz
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

How can such a system be used for a beneficial purpose?
How can such a system be used for a harmful purpose?

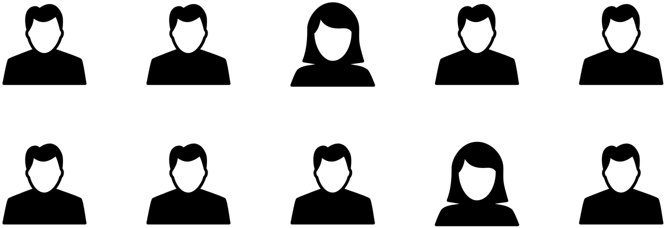
“We ose major
depr

Predicting Depression via Social Media, De Choudhury et al., 2013 [link]

Suppose you do an image search for “*CEO*” …



Suppose you do an image search for “*CEO*” …



Do you think these results are biased?

If so, do you think Google should try to address it?

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations, Kay et al, CHI 2015 [link]
Algorithms of Oppression: How Search Engines Reinforce Racism, Safiya Noble, 2018

Types of harms

- **Allocative harms**: “*when a system withholds certain groups an opportunity or a resource*”
- **Representational harms**: “*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*”

See also the keynote by Kate Crawford: The trouble with bias (Youtube, 50 min.)

Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

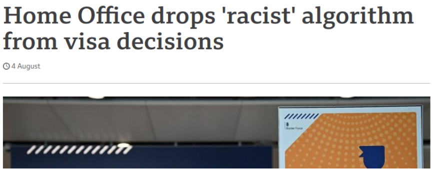
See also the keynote by Kate Crawford: The trouble with bias (Youtube, 50 min.)

Should I hire this person?

Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

See also the keynote by Kate Crawford: The trouble with bias (Youtube, 50 min.)



Home Office drops 'racist' algorithm from visa decisions

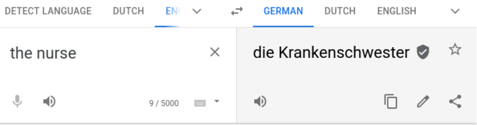
4 August

Figure: www.bbc.com/news/technology-53650758

Types of harms

- **Allocative harms**: “*when a system withholds certain groups an opportunity or a resource*”
- **Representational harms**: “*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*”

See also the keynote by Kate Crawford: The trouble with bias (Youtube, 50 min.)



the nurse

die Krankenschwester

Figure: Google Translate

Types of harms

- **Allocative harms**: *“when a system withholds certain groups an opportunity or a resource”*
- **Representational harms**: *“when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.”*

See also the keynote by Kate Crawford: The trouble with bias (Youtube, 50 min.)

Suppose you do an image search for “CEO” …

Feedback loops

You work at a bank and you build a system to

Classify loan applicants into:

- High-risk: They receive a higher interest rate (e.g., 15%)
- Low-risk: They receive a lower interest rate (e.g., 5%).

Feedback loops

You work at a bank and you build a system to

Classify loan applicants into:

- High-risk: They receive a higher interest rate (e.g., 15%)
- Low-risk: They receive a lower interest rate (e.g., 5%).

This makes it more likely that high-risk applications are not able to pay back their loan.

A model trained on this data may assess particular groups to be even *more* high risk.

Terminology I

Linear regression:

$$y = b + w_1 * x_1 + w_2 * x_2 + ... + w_d * x_d$$

where b is called the bias term. Some of you may also know about the *bias-variance* trade off.

Terminology I

Linear regression:

$$y = b + w_1 * x_1 + w_2 * x_2 + ... + w_d * x_d$$

where b is called the bias term. Some of you may also know about the *bias-variance* trade off.

Note: in the fairness ML literature, “bias” is used in a different way!

Terminology II

But:

- Fair machine learning is just getting started! There is no single definition for “bias” or “fairness”.
- Research articles often don’t define what they mean with these terms. Different studies have different conceptualizations of bias.

Terminology II

But:

- Fair machine learning is just getting started! There is no single definition for “bias” or “fairness”.
- Research articles often don’t define what they mean with these terms. Different studies have different conceptualizations of bias.

In the next two lectures, we will look at criteria to assess whether ML systems are fair. These criteria formalize the relevant concepts in a more precise way.

Data

Garbage in, garbage out

Bias in, bias out

Dong Nguyen (2021)

Garbage in, garbage out

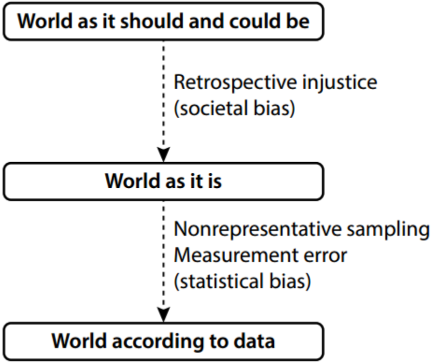
Bias in, bias out

Yes! But also:

- Which problem do you end up solving?
- Datasets shape the course of academic communities

Datasets from “Patterns, Predictions, and Actions“ by Hardt and Recht, 2021 [link]

“Biased” data



World as it should and could be

Retrospective injustice
(societal bias)

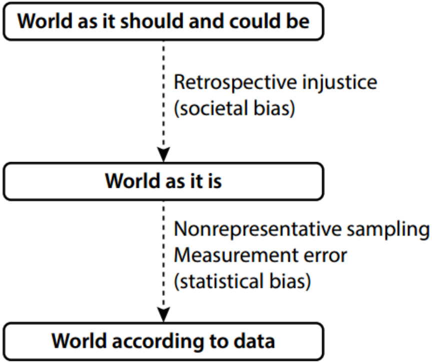
World as it is

Nonrepresentative sampling
Measurement error
(statistical bias)

World according to data

Figure: Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021 [link]

“Biased” data



If we would have *a perfect representation of the world*, we would only address the statistical bias problem.
There are no real-world datasets *free of societal biases*

Figure: Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021 [link]

Statistical Bias

Non-representative sampling: *Commonly used image datasets are often US/European centered.*

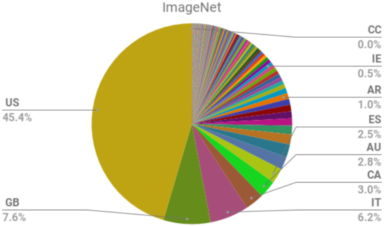


Figure: Source: from Figure 1, Shankar et al. 2017

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Shankar et al, NIPS 2017 workshop: Machine Learning for the Developing World [link]

Statistical Bias

Non-representative sampling: *Commonly used image datasets are often US/European centered.*

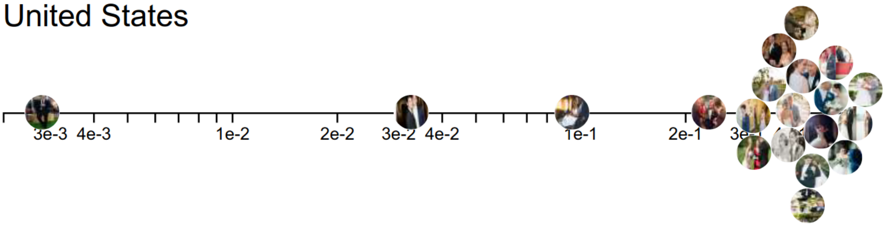


Figure: Source: from Figure 5, Shankar et al. 2017; the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Shankar et al, NIPS 2017 workshop: Machine Learning for the Developing World [link]

Statistical Bias

Non-representative sampling: *Commonly used image datasets are often US/European centered.*

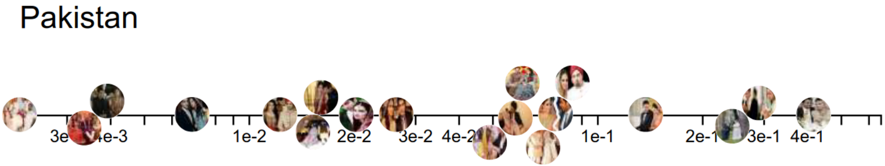


Figure: Source: from Figure 5, Shankar et al. 2017; the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Shankar et al, NIPS 2017 workshop: Machine Learning for the Developing World [link]

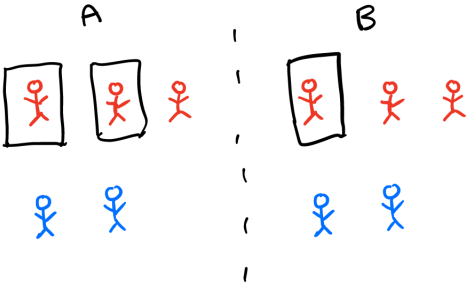
Statistical Bias

Measurement error

Your boss wants you to make a system to predict crime rates in neighborhoods to improve the efficiency of police efforts…

Statistical Bias

Measurement error

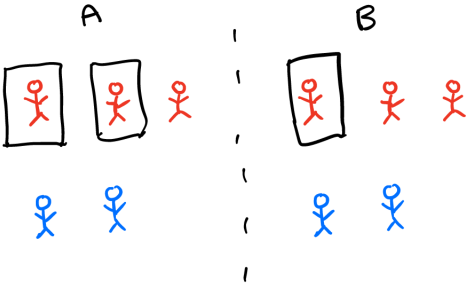


In both neighborhoods, 3/5 of the people *commit a crime*.

In A: 2/5 are *arrested*.
In B: 1/5 are *arrested*.

Statistical Bias

Measurement error



In both neighborhoods, 3/5 of the people *commit a crime*.

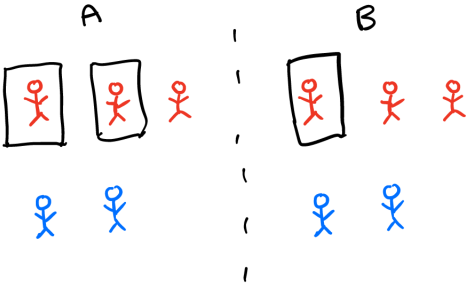
In A: 2/5 are *arrested*.
In B: 1/5 are *arrested*.

Now what happens when we train on a ML model on arrest data?

Dong Nguyen (2021)

Statistical Bias

Measurement error



In both neighborhoods, 3/5 of the people *commit a crime*.

In A: 2/5 are *arrested*.
In B: 1/5 are *arrested*.

Feedback loop!
Why overpolicing? societal bias.

There’s often a disconnect between the target variable and our overall goal!

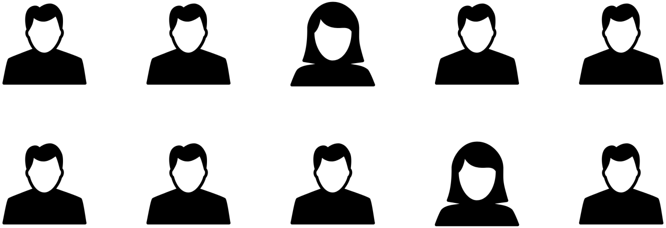
Being re-arrested vs. re-offending vs. risk to society

Repayment of loans vs. better lending policies

Often different stakeholders have different overarching goals.

Statistical bias vs societal bias

“CEO” …



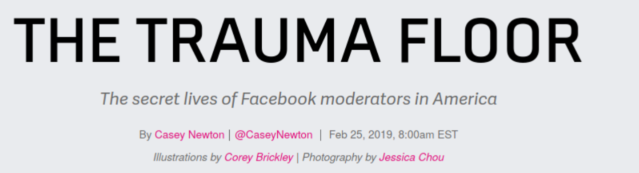
Data collection and labeling practices I

Check if your Flickr photos were used to build face recognition Exposing.AI

Would you be ok with your *public* images (Flickr, Twitter, Instagram, ...) being used for learning to:

- detect objects (chair, house, ...)
- detect locations
- detect people
- detect emotions
- rate attractiveness

Data collection and labeling practices II



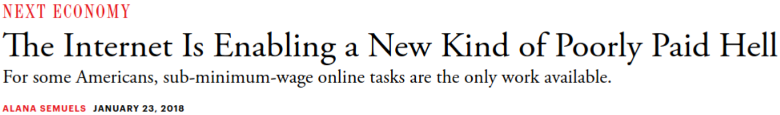
THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST

Illustrations by Corey Brickley | Photography by Jessica Chou

Figure: Source: The Verge



NEXT ECONOMY

The Internet Is Enabling a New Kind of Poorly Paid Hell

For some Americans, sub-minimum-wage online tasks are the only work available.

ALANA SEMUELS JANUARY 23, 2018

Figure: Source: The Atlantic

Datasheets for Datasets

Datasheets for Datasets, by Gebru et al.

- **Motivation:** e.g., *for what purpose was the dataset created?*
- **Composition:** e.g., *does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, ...)?*
- **Collection process:** e.g., *what mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?*
- **Uses:** e.g., *are there tasks for which the dataset should not be used?*
- **Distribution:** e.g., *how will the dataset will be distributed (e.g., tarball on website, API, GitHub)?*
- **Maintenance:** *will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?*

Development of ML models

It can happen with the best intentions!

A study found that fewer women saw ads promoting job opportunities in STEM (science, technology, engineering, math)—even though the delivery was intended to be gender neutral. Why?

It can happen with the best intentions!

A study found that fewer women saw ads promoting job opportunities in STEM (science, technology, engineering, math)—even though the delivery was intended to be gender neutral. Why?

Younger women were more expensive to show ads to. Thus: optimizing for cost-effectiveness led to ad delivery that can be seen as biased.

Anja Lambrecht, Catherine Tucker (2019) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science 65(7):2966-2981 [link]

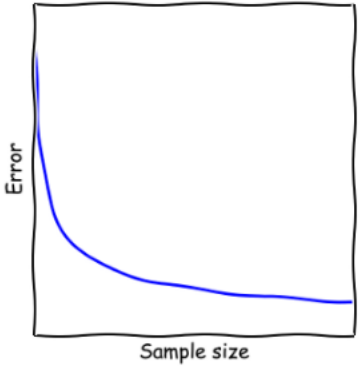
Machine *learning*

Generalization

- Training versus test examples
- Memorization is not enough!

We don’t want just to memorize. We want to **generalize**.

Machine learning: Sample size



Performance tends to be lower for minority groups. Note that this even happens when our data is fully representative of the world!

Figure: Figure from Moritz Hardt 2014 [link]

Machine learning: Optimization

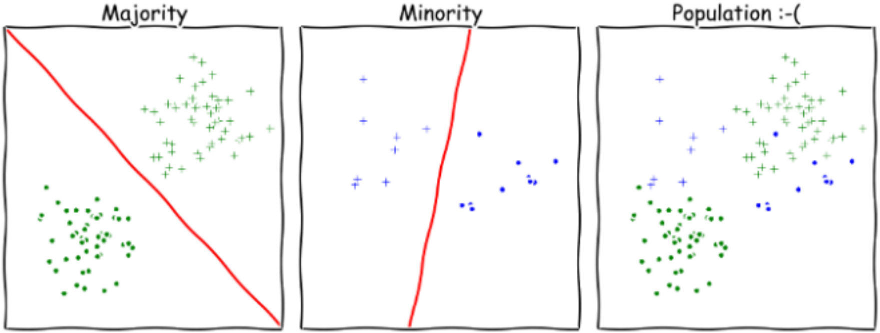


Figure: Figure from Moritz Hardt 2014 [link]

Machine Learning: Optimization

When we optimize to find the model with the min. average error, our model will work better for the majority group in our data.

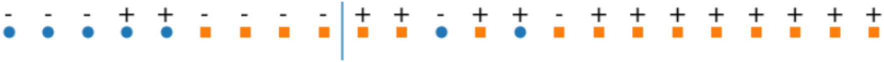


Figure: Example based on “the ethical algorithm” by Kearns and Roth, “fairness fighting accuracy”

19/23 correct

Machine Learning: Optimization

When we optimize to find the model with the min. average error, our model will work better for the majority group in our data.

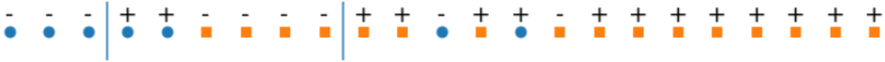
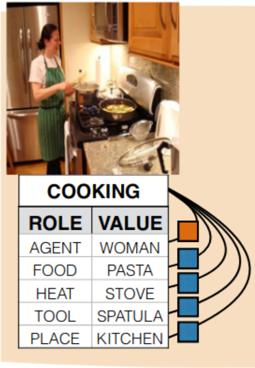


Figure: Example based on “the ethical algorithm” by Kearns and Roth, “fairness fighting accuracy”

Group specific thresholds?

ML models can amplify biases in the data



COOKING	
ROLE	**VALUE**
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Figure: From Fig 1 from Zhao et al.

33% of the *cooking* images have *man* in the agent role. But during test time, only 16% of the agent roles are filled with *man*.

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, Zhao et al., EMNLP 2017 [link]

Features

Instances are represented by features:

- House: number of bedrooms, neighborhood, has garden?, etc.
- Person: education, number of years experience with skill X, etc.

Which features are informative for a prediction may differ between different groups. A particular feature set may lead to high accuracy for the majority group, but not for a minority group.

The quality of the features may differ between different groups.

Features

Instances are represented by features:

- House: number of bedrooms, neighborhood, has garden?, etc.
- Person: education, number of years experience with skill X, etc.

What about the inclusion of *sensitive attributes* as feature? gender, race, ...
What if including such a feature:

- Improves overall accuracy but lowers accuracy for specific groups
- Improves overall accuracy, for all groups

What if we need such information to evaluate the fairness of systems?

Evaluation

In machine learning, the evaluation often makes strong assumptions (e.g., *i.i.d*).

- Outcomes are not affected by decisions on others.
    - Denying someone's loan can impact the ability of a family member to repay their loan.
- We don't look at the type and distribution of errors.
- Decisions are evaluated simultaneously
    - Feedback loops. Long-term effects.

Algorithmic Fairness: Choices, Assumptions, and Definitions, Mitchell et al., Annual Review of Statistics and Its Application, 2021 [link]

Model Cards for Model Reporting

Model Cards for Model Reporting, by Mitchell et al. FAT* 2019 for *transparant model reporting*, such as:

- Model details (e.g., version, type, license, features)
- Intended use (e.g., primary intended uses and users, out-of-scope use cases)
- Training data
- Evaluation data
- Ethical considerations
- …

An example online Model Card for Face Detection can be found here.

Outlook

What can we do?

Three types of responses (Wachter et al.):

- ~~Nothing~~
- Correct for “technical” bias so that the system reflects the status quo. Make society not more unequal than it currently is. Example: Equal error rates across groups.
  - aligns with the concept of “formal equality” in EU non-discrimination law
- Acknowledge that the status quo is a result of existing inequalities.
  - align with the concept of “substantive equality” in EU non-discrimination law

Wachter et al.: “*While legal scholars broadly agree that the aim of EU nondiscrimination law is substantive equality, they disagree about how best to achieve the necessary structural, institutional, and societal change in practice.* ”

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, Forthcoming [link]

Literature for today

Required reading:

- Chapter 1 "*Introduction*" of `https://fairmlbook.org/` "Fairness and machine learning" book, by Solon Barocas, Moritz Hardt, Arvind Narayanan
- Chapter 8 "*Datasets*", "Harms associated with data" and up (p19-29) of `https://mlstory.org/` "Patterns, Predictions, and Actions" book by Moritz Hardt and Benjamin Recht

Optional but recommended:

- Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, Forthcoming [link]

Next lectures

We will mostly take a **machine learning perspective**: *How can we measure fairness and make our ML models more fair?*

- Lecture 2: Measuring fairness
- Lecture 3: Making ML systems more fair

Next lectures

We will mostly take a **machine learning perspective**: *How can we measure fairness and make our ML models more fair?*

- Lecture 2: Measuring fairness
- Lecture 3: Making ML systems more fair

But... machine learning can't solve everything! Interdisciplinary approaches are needed (ethics, philosophy, psychology, human computer interaction, etc...)

Preparation for next lecture

Do the short quiz on Blackboard by **Monday 10am**.

Revisit:

- Evaluation (e.g., confusion matrix, precision, recall, true positives, true negatives, etc)
- Vector representations
- Cosine similarity
- kNN

Take a look at the following webpage:

https://research.google.com/bigpicture/attacking-discrimination-in-ml