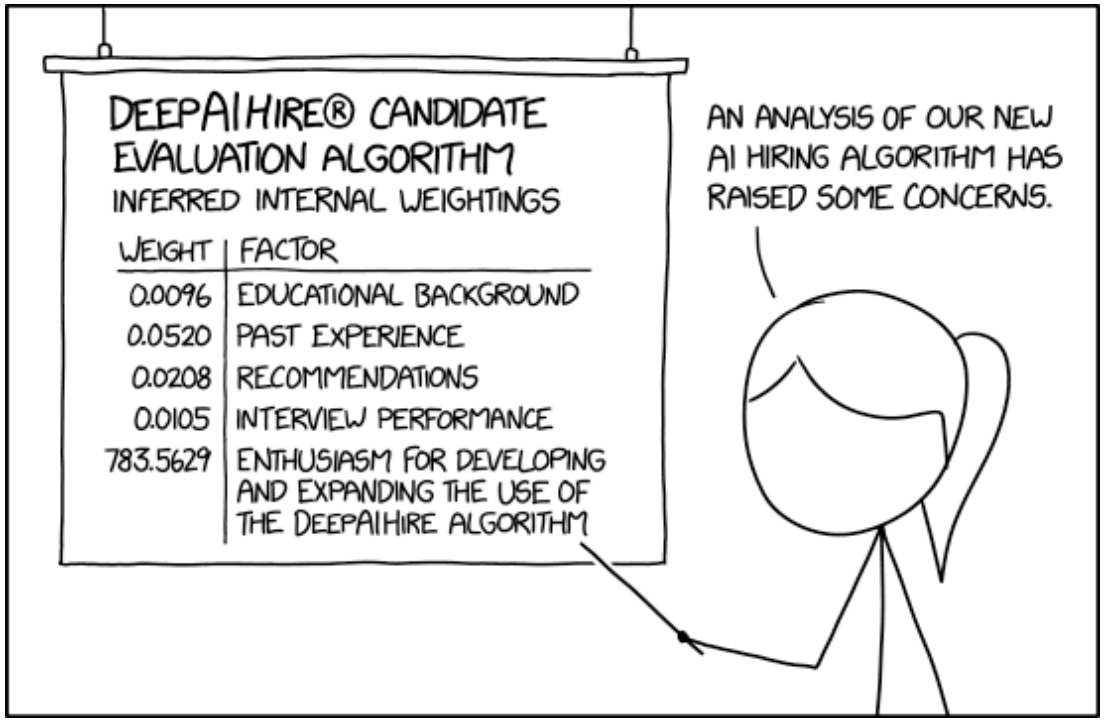
Fairness in Machine Learning

Gosia Migut
Slides credit: Tom Viering
CSE2510 Machine Learning

FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan




Cognitive Robotics
ASIMOV
TU Delft
Studium Generale
In collaboration with ASIMOV: Think & Drink
GARBAGE IN, GARBAGE OUT?
FAIRNESS & AI
With Cynthia Liem, Evgeni Aizenberg and Ronald Meester
HUMAN
Reserve your ticket through Eventbrite!
October 13th | 17:00 - 19:00 | EXPO @ RoboHouse

Examples of problematic use of ML

BBC | BBC Account | News | Sport | Weather | Shop | Reel | Travel | M

NEWS

Home | Video | World | UK | Business | Tech | Science | Stories | Entertainment & Arts

Amazon scrapped 'sexist AI' tool

10 October 2018 | Share

An algorithm that was being tested as a recruitment tool by online giant Amazon was sexist and had to be scrapped, according to a Reuters report.

The artificial intelligence system was trained on data submitted by applicants over a 10-year period, much of which came from men, it claimed.

Reuters was told by members of the team working on it that the system effectively taught itself that male candidates were preferable.

[https://ww

Gender and racial bias found in Amazon's facial recognition technology (again)

Research shows that Amazon's tech has a harder time identifying gender in darker-skinned and female faces

By James Vincent | Jan 25, 2019, 9:45am EST

As facial recognition systems become more common, Amazon has emerged as a frontrunner in the field, courting customers around the US, including police departments and Immigration and Customs Enforcement (ICE). But experts say the company is not doing enough to allay fears about bias in its algorithms, particularly when it comes to performance on faces with darker skin.

The latest cause for concern is a study published this week by the MIT Media Lab, which found that Rekognition performed worse when identifying an individual's gender if they were female or darker-skinned.

https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-

Ethics

- ML products that don’t involve people have no ethical problems
- Who agrees?

Fairness…?

- Demographic disparities. Sometimes called “biases”
- Not to be confused with bias of bias-variance trade-off

- For example: “Amazon offers free-same-day delivery is twice as likely in a neighbourhood with more white residents compared with Black residents”

- Is this discrimination?

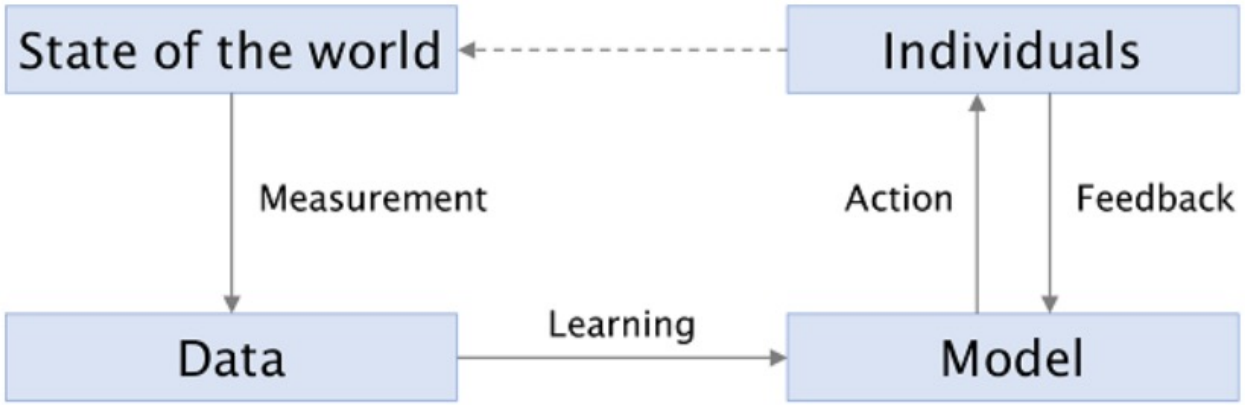
Why do disparities exist?

- History of explicit discrimination
- Implicit attitudes about groups of people
- Stereotypes
- Differences in characteristics in the groups

- Stereotypes can be self-fulfilling and are hard to get rid off…
- How to avoid them in machine learning models?

Agenda

- Part 1: why disparities get into our machine learning models

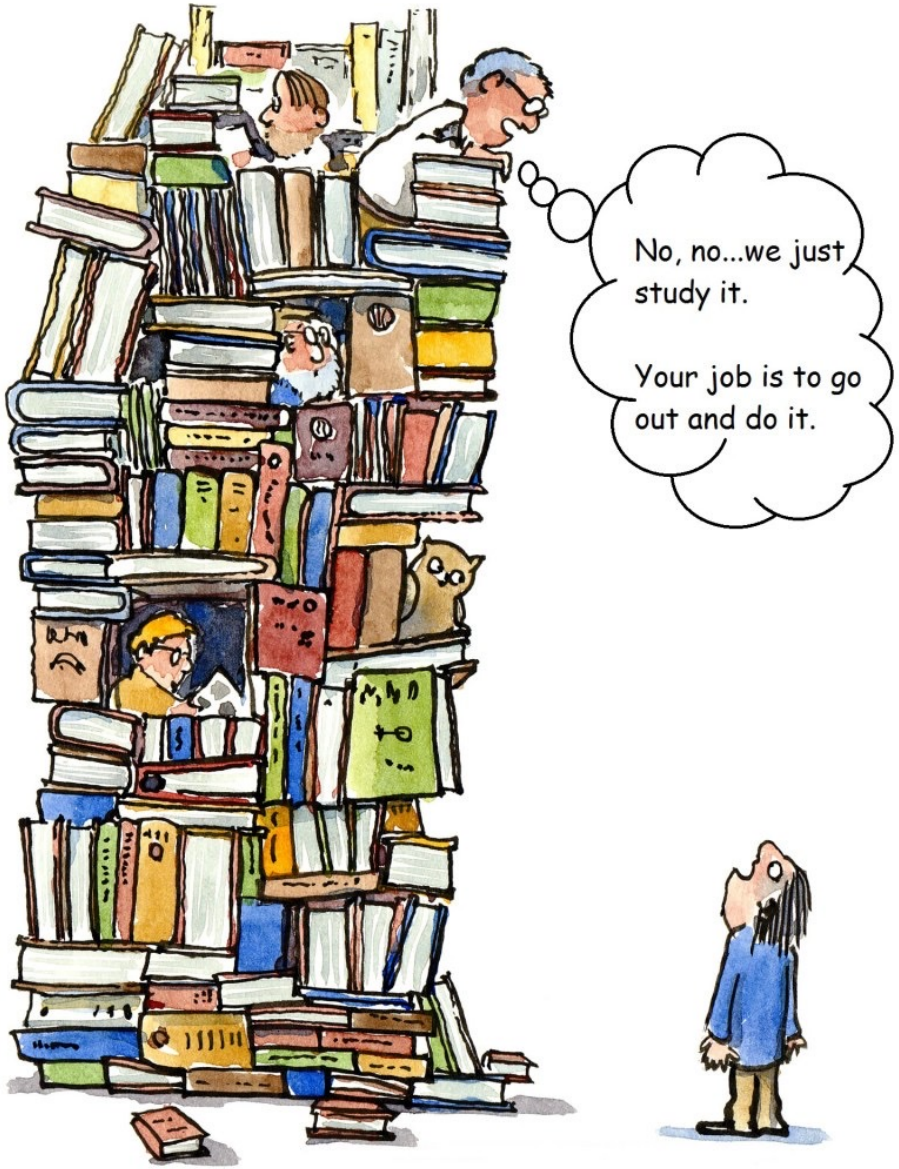


- Part 2: how to quantify disparities, fairness using statistics, and how to improve the situation

Measurement

- Downloading an excel sheet, CSV file
- Scientist measuring in a lab
- Using a sensor
- Objective

- True or False?



Measuring difficulties (1)

- Measuring race:
  - social construct; not related to biology
  - changes over time,
  - different terms in different cultures (“African-American”, …)

- Checkbox?
  - Multiracial
  - How to measure?

- Measuring stuff about people is subjective, challenging

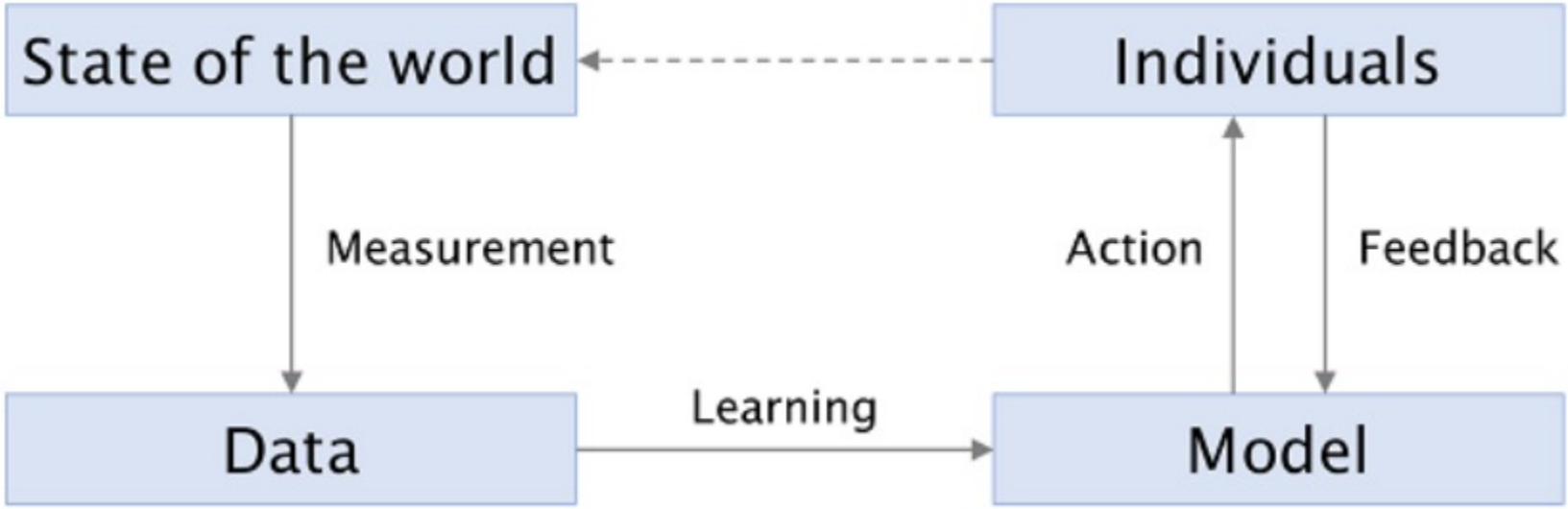
Measuring difficulties (2)

- Machine learners need to make some variables / categories / targets

- Target variables:
  - “credit worthiness”
  - “good employee”
  - “physical attractiveness”
  - “criminal risk assessment”

- Subjective judgements inherit stereotypes (unconsciously)
- Social constructs

Machine Learning loop



Learning difficulties (1)

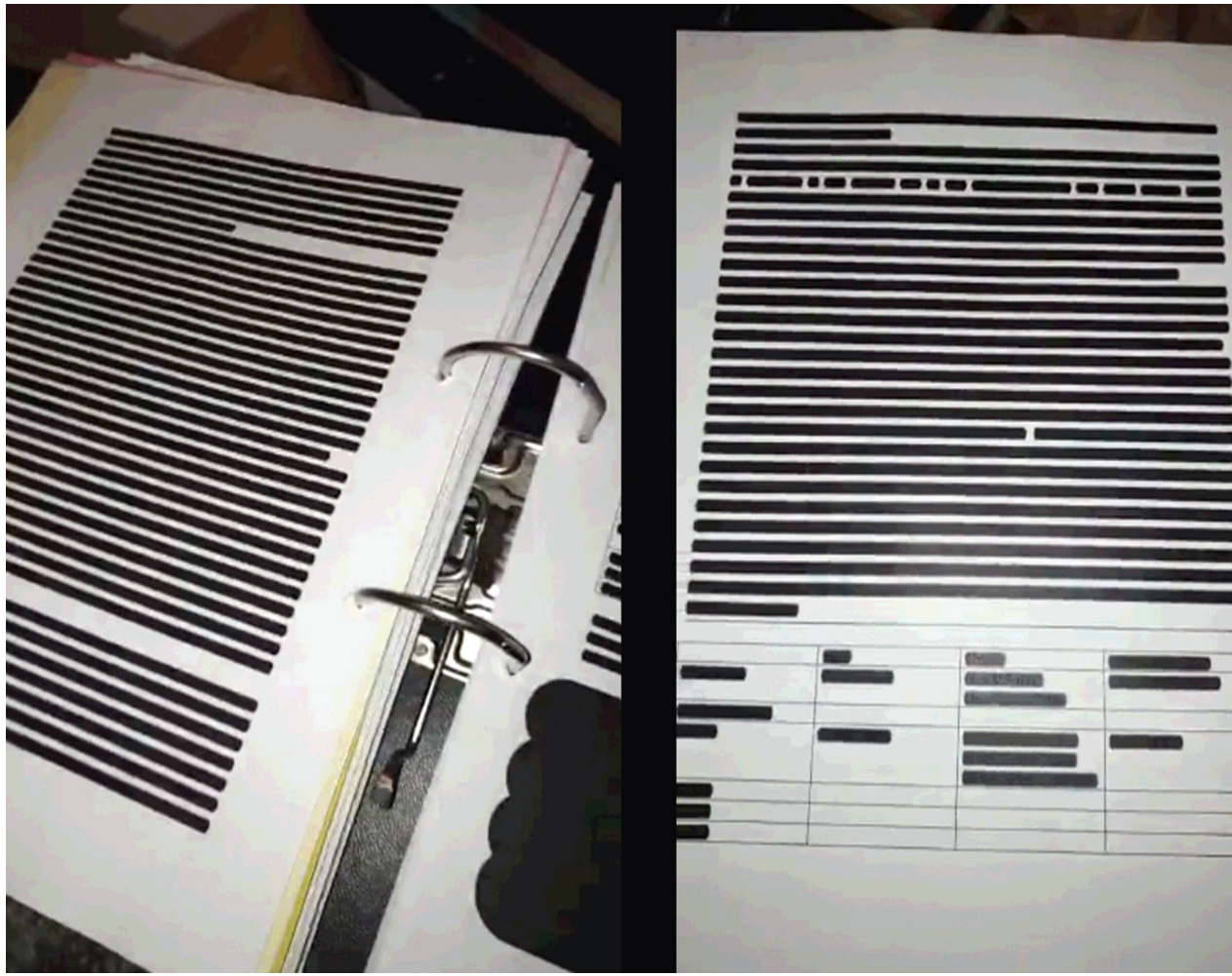
- Models do not understand what patterns are objectionable / racist, and which ones are OK

- Learning algorithms will extract stereotypes, discrimination from the data

Idea!

- ML application that judges job applications
- Idea: remove the gender

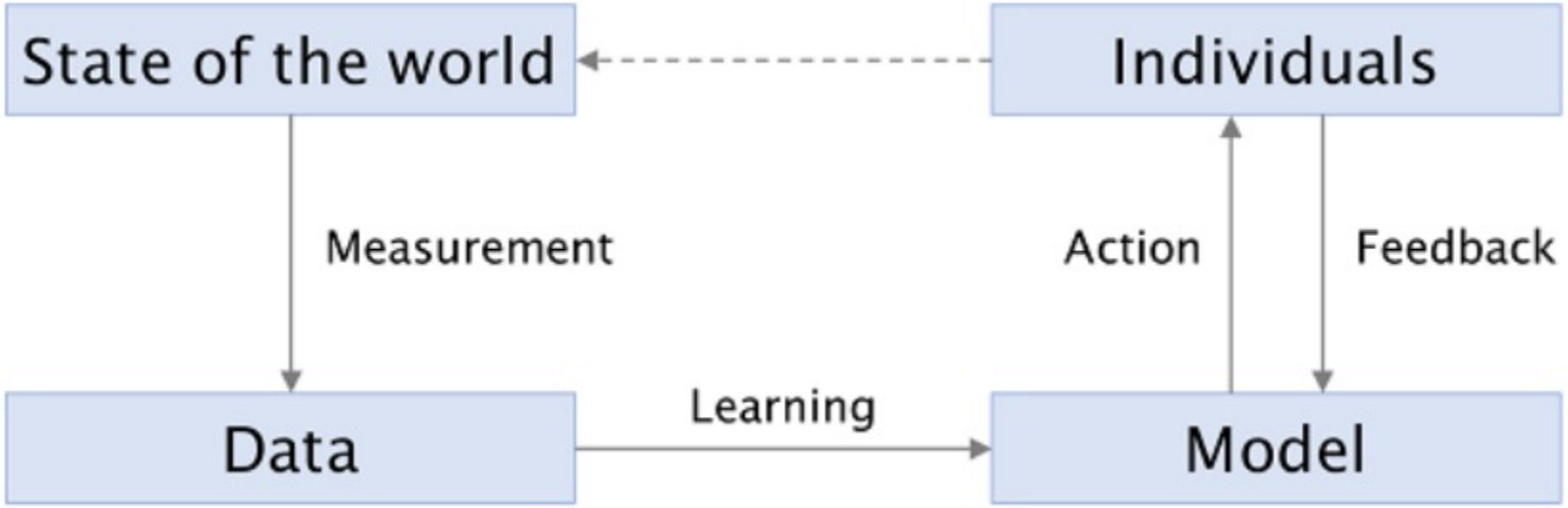
- How?
- Age you start programming
- For men at much earlier age
- Remove from CV?
- If we remove all gender-identifying information, what are we left with? How?



Learning Difficulties (3)

- CelebA dataset, widely used for facial recognition
- Celebrities from Hollywood
- Problem?
- Little minority data
- Result: models underperform for minorities (too little training samples)

Machine Learning loop



Action Difficulties (1)

- Our model predicts who is suitable for a job

- Training data:
- 90% of population 1 is suitable for the job
- 50% of population 2 is suitable for the job

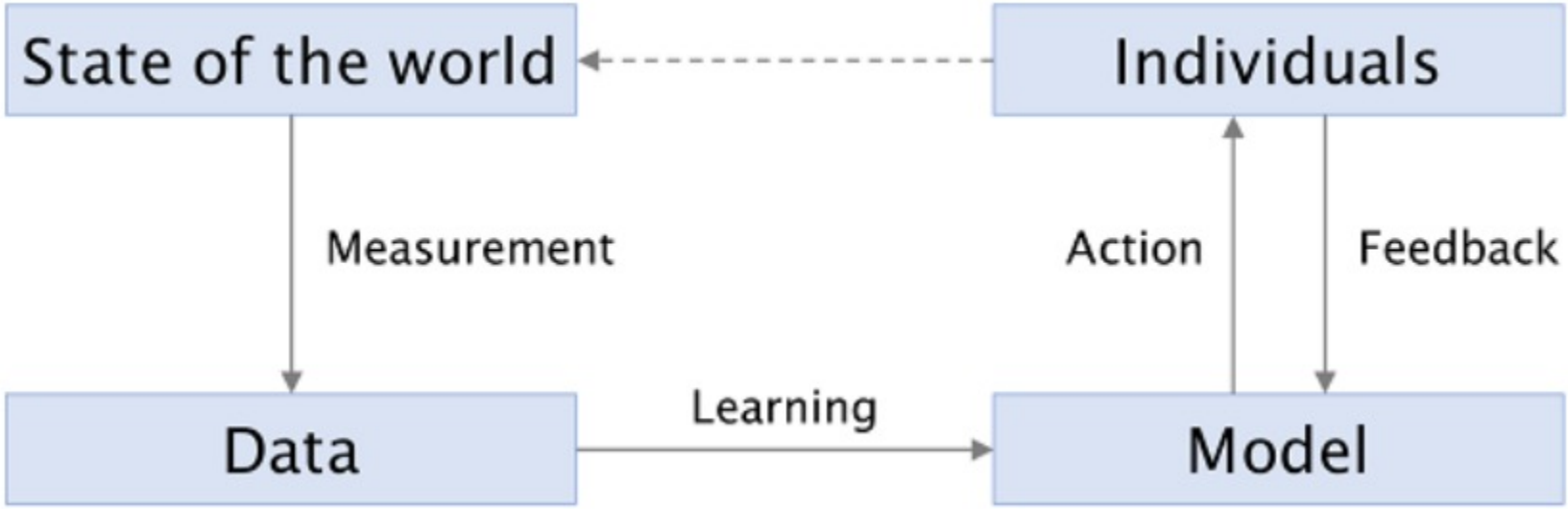
- The model will reproduce this pattern in the data
- Is that bad?

Action Difficulties (2)

- Automated decisions
- Only the outcomes are important, not the process.
- True or False?

- Ethical decisions: not only outcome important, but process
- Why was a decision made?
- Only if we can understand how the decision is made, we can understand if it was ethical or not
- Needs to be explainable / white box

Machine Learning loop



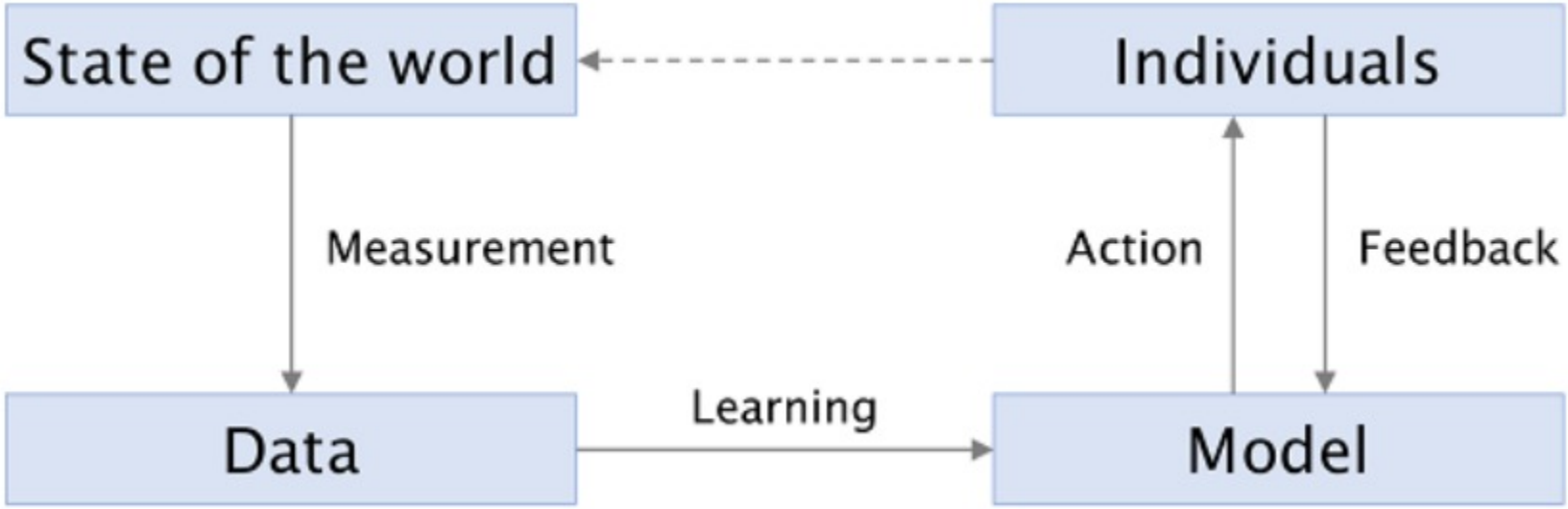
Feedback

- Feedback: click on Ad, link search result
- Feedback to the machine learning system to improve model

- Googling a Black-sounding name: more ads for arrest records
- Because users click on ads that conform to stereotypes

- Is the feedback what we want? Clicked because best result, or just because on top of page?
- Feedback can reflect or amplify cultural prejudices

Machine Learning loop



Loops

- Self-fulfilling predictions
- Predictions that affect the training data
- Predictions that affect society at large

Self-fulfilling predictions

- Predictive policing
- Machine learning model says where police should patrol

- Police go more often to areas that are predicted to be high risk
- Officers will make more arrests there
- Officers unconsciously may be more alert or may treat people differently in the area

Predictions that affect the training data

- Predictions seem validated by new data
- Feedback loop: this area will become more and more high risk
- Algorithm identify similar areas based on features and also consider them high risk

- Small mistakes in first model spiral out of control: patterns, biases, mistakes amplified
- PredPol study: Black people targeted twice as much for drug policing
- Even though both groups equal rate of drug use

Predictions that affect society at large

- Machine learning that amplify prejudice
- Makes prejudice even more persistent
- Affects poverty and crime on long time scale

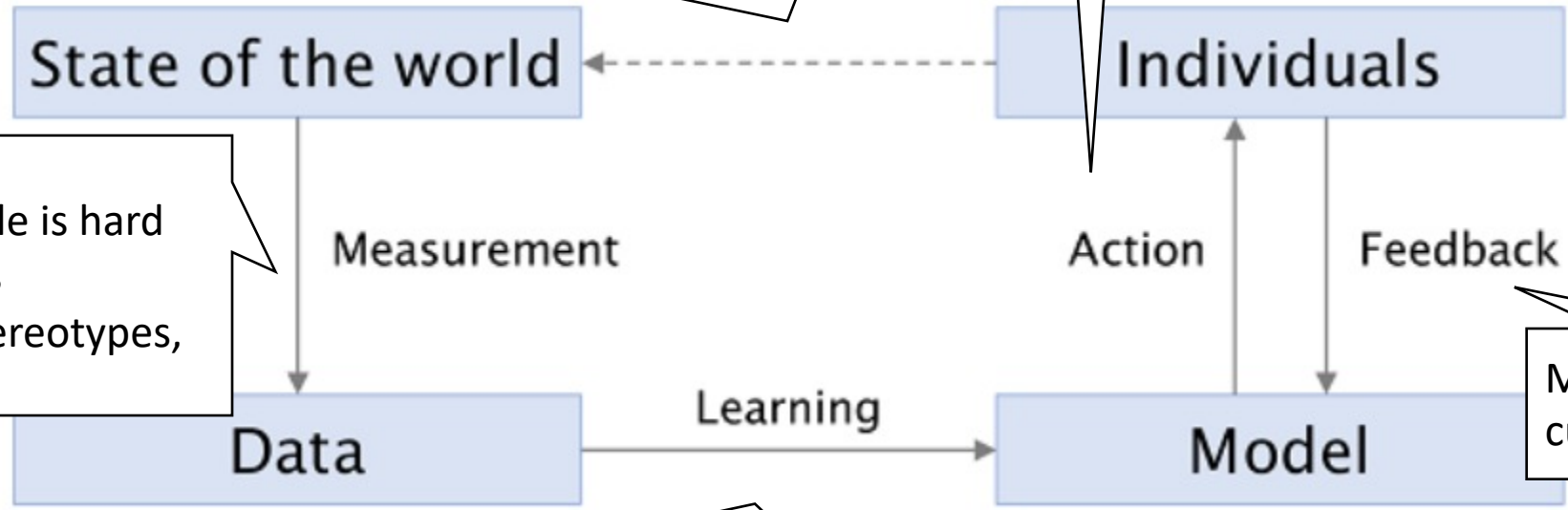
- Almost all systems have an effect
- Search engines: result rankings.
- Why?

- Over time, highly ranked items are read more, more important.
- Highly ranked items influence society more.
- Will influence future search results.

Summary

Feedback loops
- Self-fulfilling predictions
- Actions affect the new training data
- Actions affect society, which also affects training data

- Acceptance rate (e.g. job application) may differ based on demographic / gender
- Actions need to be explainable to be ethical



- Subjective
- Measuring people is hard
- Social constructs
- Subconscious stereotypes, racism, etc.

May contain or amplify cultural prejudices

- Models don't know difference between ethical versus non-ethical patterns (racism, biases, stereotypes)
- Removing all gender / race information hard / unfeasible
- Minorities: little data, underrepresentation means worse model

Other considerations

- A “fair” learning model might not always be the best solution
  - Maybe: try to improve workplace for minorities
- Should we automate and measure everything?
  - Errors with facial recognition technology, DNA for forensics, etc…

Reasons to be optimistic

- Machine learning can be more accurate than experts

- Machine learning can be more transparent
- Forces us to articulate our objectives
- Debate and specify: what is fair? What are the trade-offs?

- More difficult to hide poorly specified or harmful true intentions

Part 2: Measuring fairness & adapting models

- Why fairness through blindness doesn’t work
- Why sensitive attributes shouldn’t removed
- Measuring fairness using statistics:
  - Criteria 1: Independence
  - Criteria 2: Separation
- Limitations of statistics

Some notation

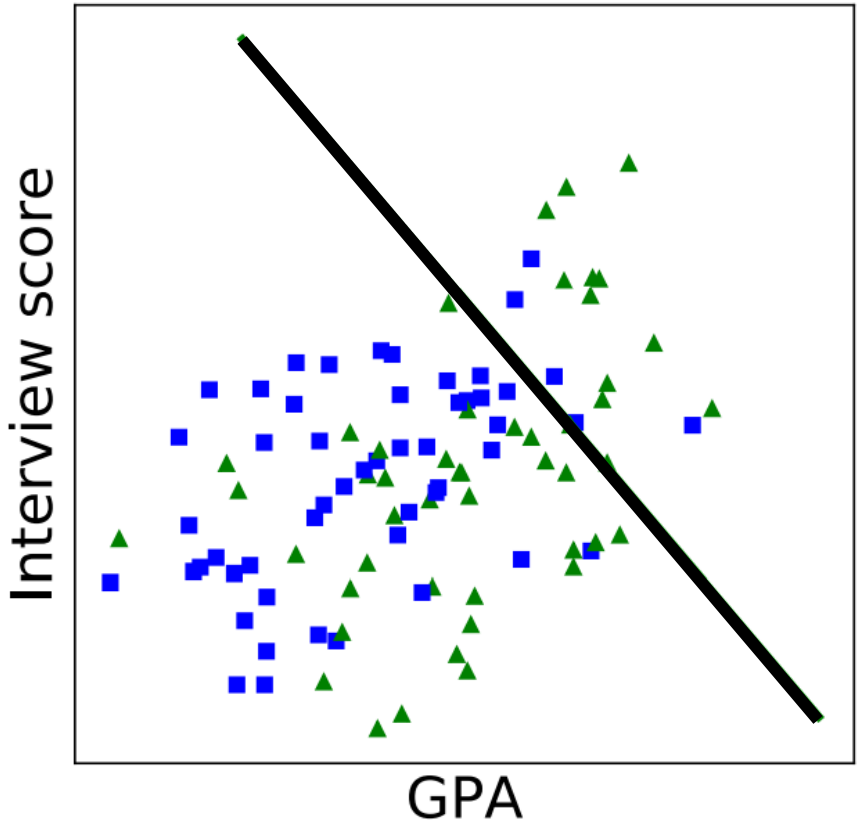
A sensitive attribute. Example: $A = a$ (female), $A = b$ (male)

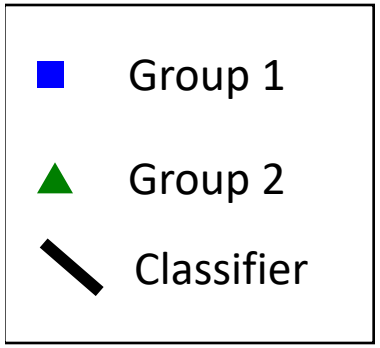
X other features. Example: word occurrences in resume.

Y classification target. $Y = 0$ (job rejection), $Y = 1$ (accept)

R classifier output. $R = 0$ (predict: reject), $R = 1$ (predict: accept)

Example



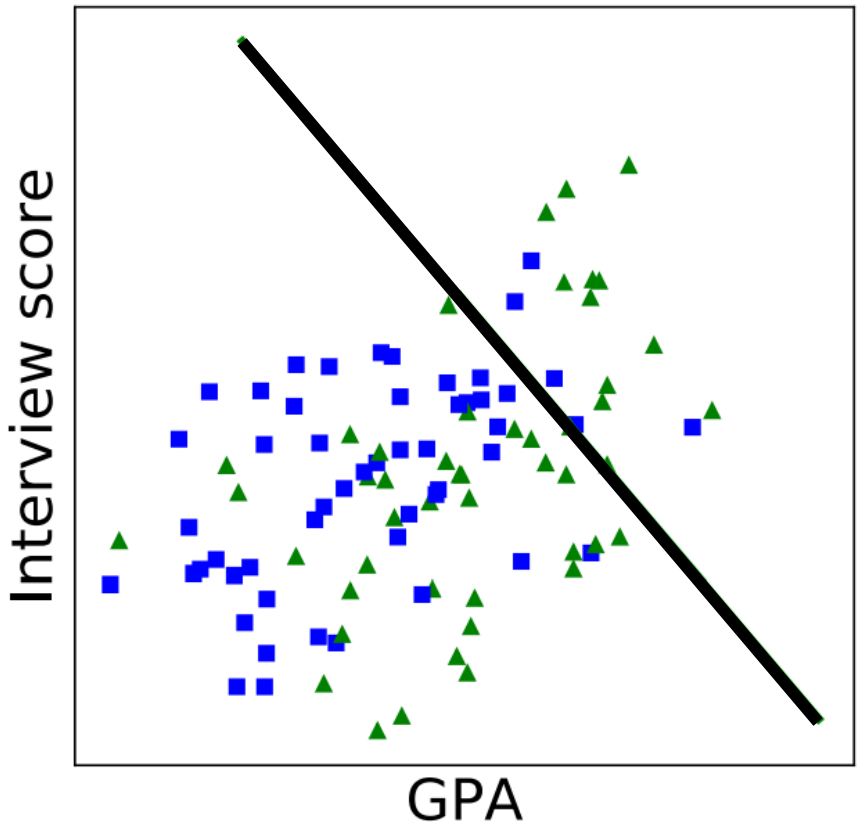


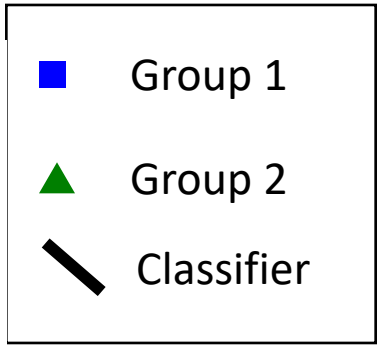
Fair?

Group 1: 5/100 hired = 5%
Group 2: 11/60 hired = 18%

- Model doesn't use the sensitive attribute to make predictions, only GPA and interview score.
- Is this fair?

Example



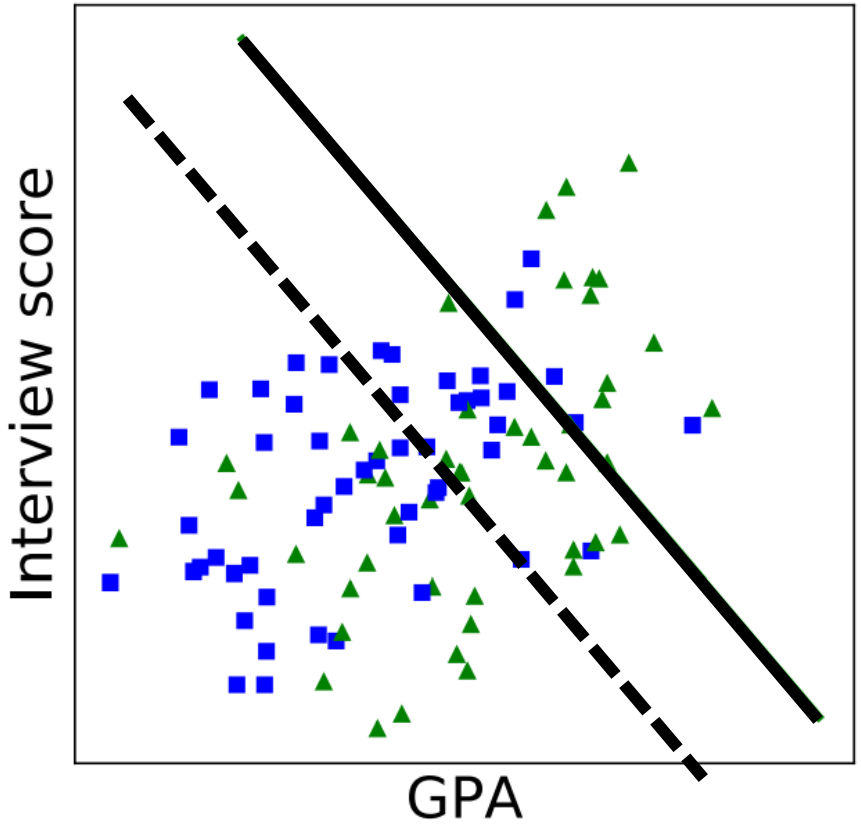


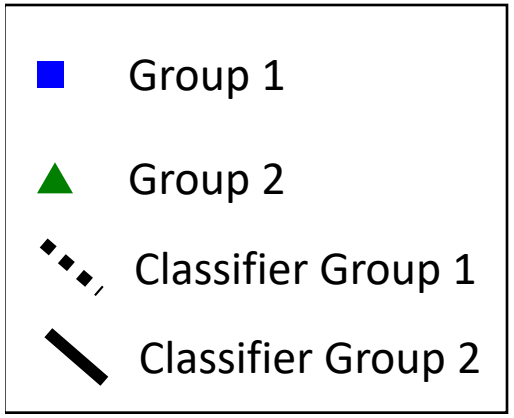
Fair?

Group 1: 5/100 hired = 5%
Group 2: 11/60 hired = 18%

- Why does the model accept more from one group than the other?
- Conclusion: removing or not using the sensitive attribute is not enough.

Example





Fair?

Group 1: **18/100** hired = **18%**
Group 2: 11/60 hired = 18%

- Can use 2 models: one for group 1, one for group 2.
- Now fair?

Criteria 1: Independence

$$P(R = 1|A = a) = P(R = 1|A = b)$$

Criteria 1: Independence.

The acceptance rate for each group should be the same.

The classification rule does not have to be the same for each group (!)

Independence not satisfied

Group 1: 5/100 hired = 5%
Group 2: 11/60 hired = 18%

Independence satisfied

Group 1: 18/100 hired = 18%
Group 2: 11/60 hired = 18%

Criteria 1: Independence

$$P(R = 1|A = a) = P(R = 1|A = b)$$

Criteria 1: Independence.

The acceptance rate for each group should be the same.

The classification rule does not have to be the same for each group (!)

Note: need sensitive attribute to compute.

Never remove the sensitive attribute from your dataset!

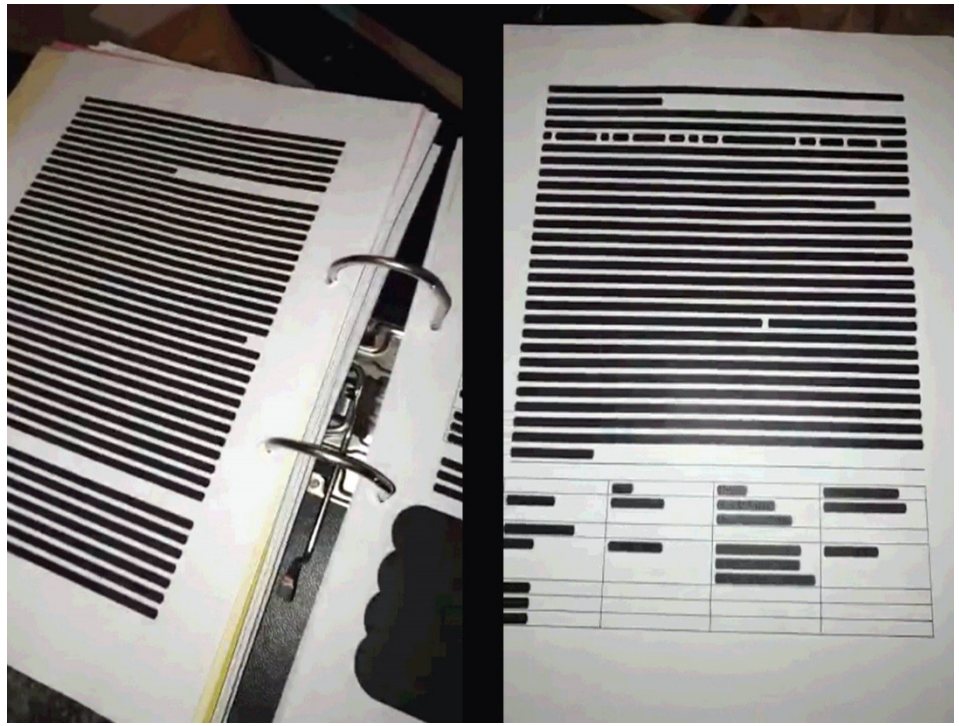
Example exam question

- Is this statement True or False?
  Explain why, in one sentence.

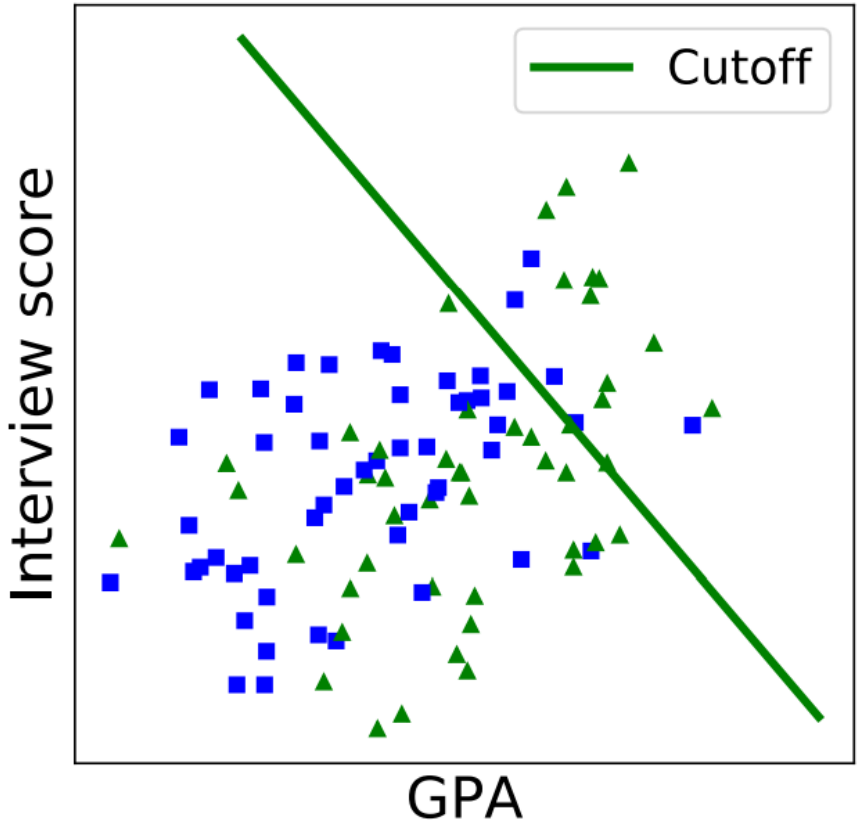
Removing a sensitive feature will make sure that the trained model cannot discriminate based on the sensitive feature.

Independence Approach 2: Pre-process

- Another way to satisfy independence is to pre-process the data
- The data should be pre-processed in such a way, that it becomes impossible to tell from which group a candidate comes
- Problems: 1) How?
- 2) May loose too much information



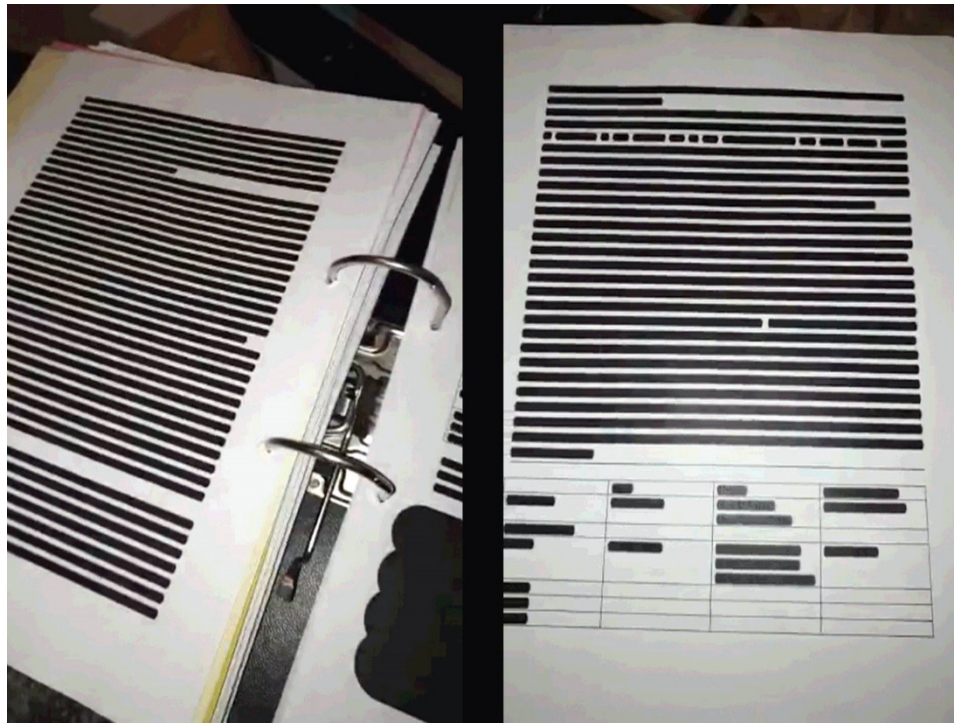
Independence Approach 2: Pre-process



- Which feature would you remove?
- Would you use a model that ignores the interview score?

Pre-processing problems

- For bigger datasets even more problematic. Why?
- Sensitive attribute is often intertwined with all data. May need to remove a lot.
- Model performance might suffer a lot.



Pre-processing problems

- Model input: DNA. Output: predict the income.

- What will this model learn?
- A: DNA -> race -> income.

- What do we remove from the DNA?
- For visual data: what should be removed to obscure gender?

Criteria 1: Independence

$$P(R = 1|A = a) = P(R = 1|A = b)$$

Criteria 1: Independence.

The acceptance rate for each group should be the same.

Approaches:

- Different model per group
- Pre-process data so that we cannot distinguish the 2 groups

Problem with Independence

Group	Loans payed back	Loans defaulted (fail to pay back)
A=a	90	10
A=b	10	90

The bank doesn’t want to use a classification model that satisfies independence here. Why not?

A: The acceptance rate in both groups need to be the same.

Need to accept a lot from group b, or reject a lot from group a.

Too much lost profit.

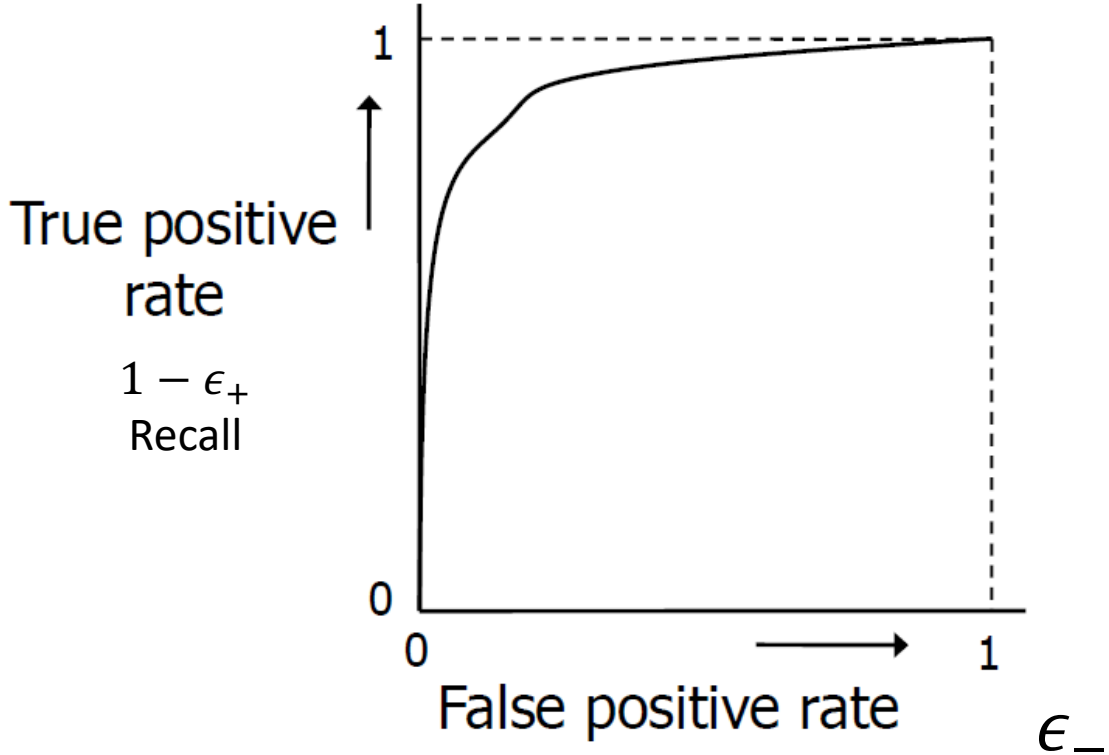
Criteria 2: Separation

- Instead: fraction of mistakes for each group should be the same.
- ϵ_+ and ϵ_- should be the same for each group
- Fraction of positives can now be different per group

		PREDICTED		
		Positive	**Negative**	**Error rate per class**
ACTUAL	**Positive**	TP / True Positive	FN / False Negative	$\epsilon_+ = FN/(TP + FN)$
	Negative	FP / False Positive	TN / True Negative	$\epsilon_- = FP/(FP + TN)$

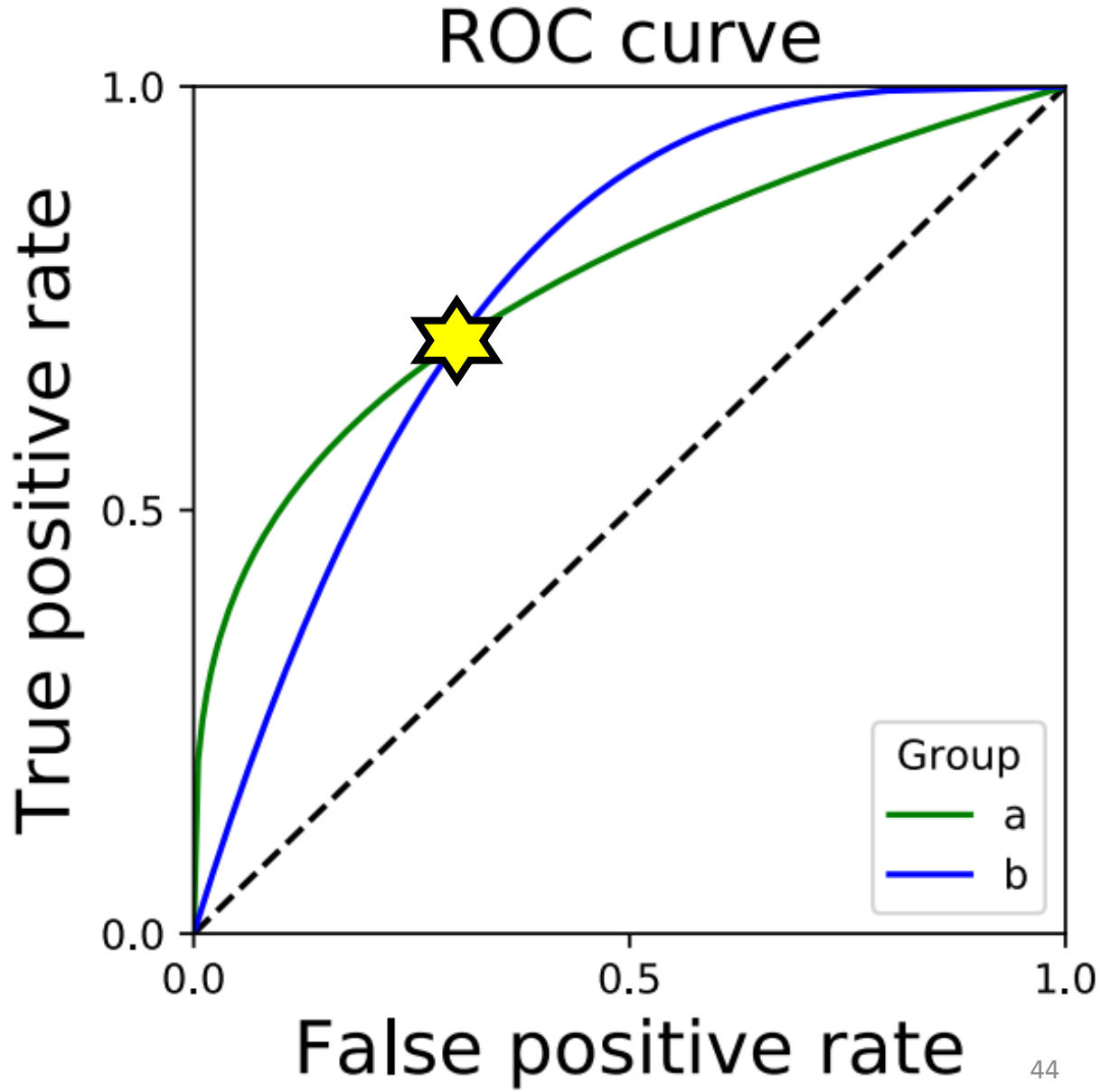
ROC curve

		PREDICTED		
		Positive	**Negative**	**Error rate per class**
ACTUAL	**Positive**	TP / True Positive	FN / False Negative	$\epsilon_+ = FN/(TP + FN)$
	Negative	FP / False Positive	TN / True Negative	$\epsilon_- = FP/(FP + TN)$

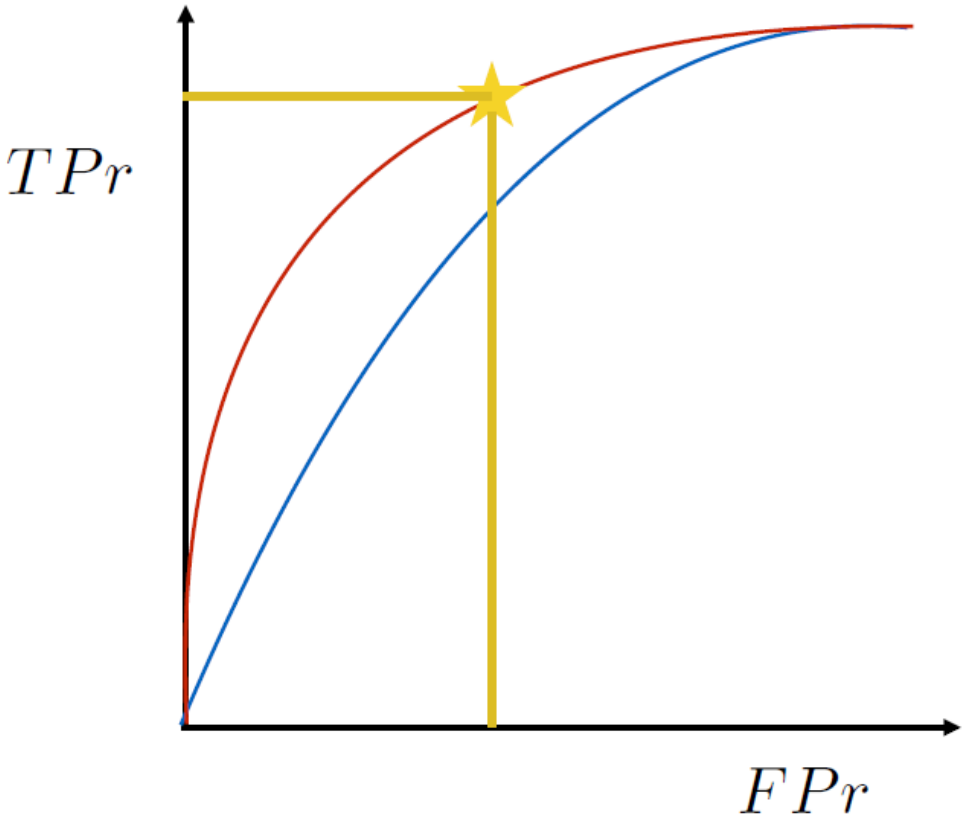


Separation

For which point is
Separation satisfied here?

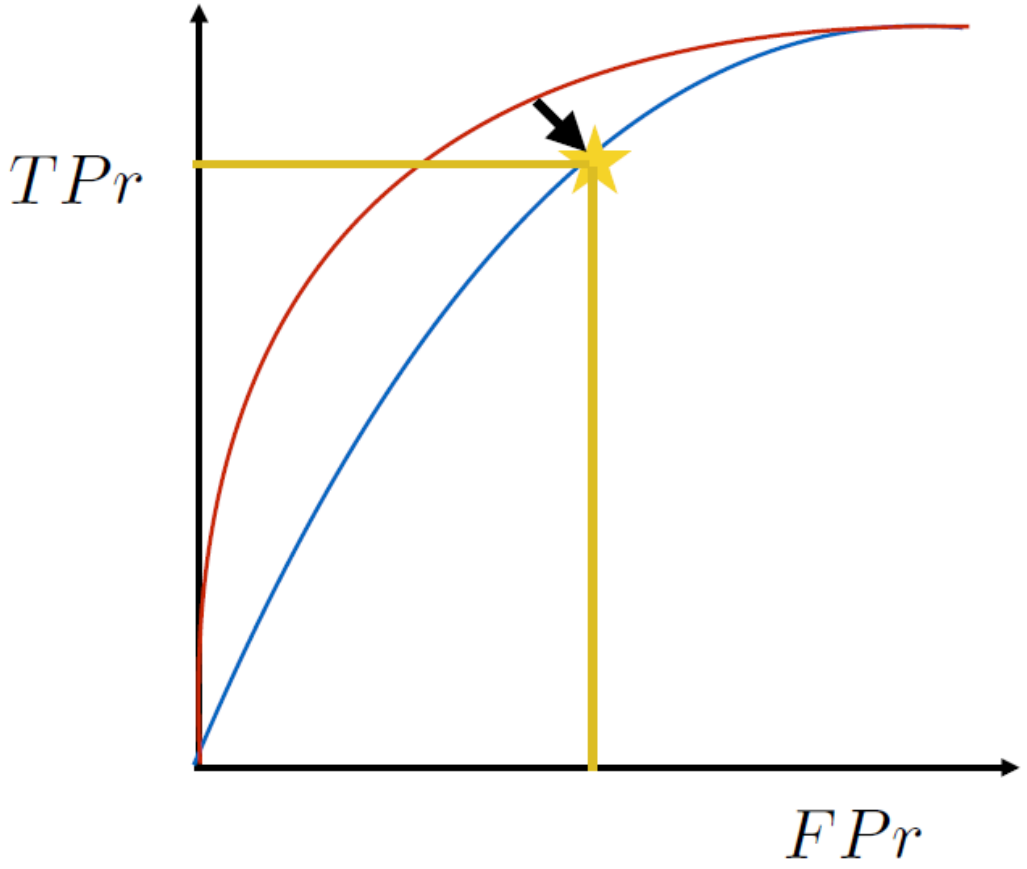


Separation: how?



What to do in this case?

Separation: how?

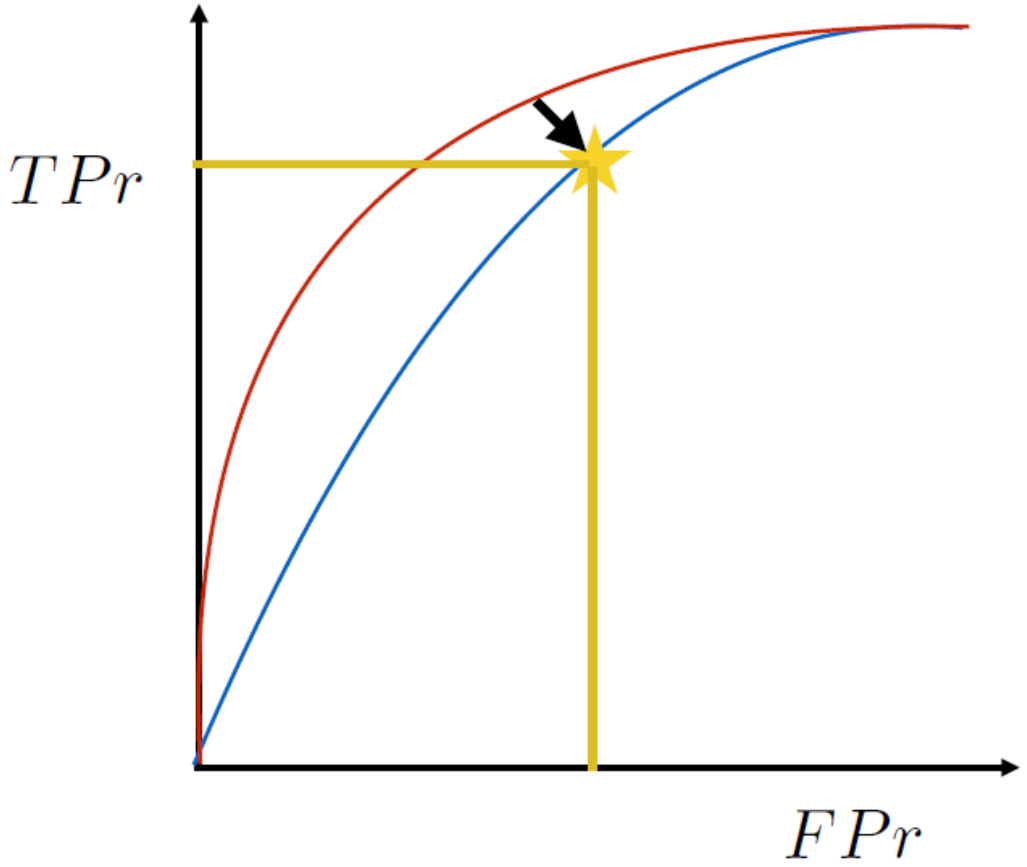


Add some noise to the output of the model if we get a sample from the red group.

The noise will make the classifier dumber, lowering the TPr and FPr.

Choose just the right amount of noise so that the star lands on a point on the blue curve.

Separation: how?



Final rule:

If from red group:
Prediction = Noise + model output (dumber).

If from blue group:
Prediction = Model output. (no noise needed)

Note that for this final rule we again need to use the sensitive attribute!

Example exam question

Is the statement True or False?
Explain why, in max. three sentences.

In the context of predicting whether people should receive a loan, the Independence criterion allows the bank to make a larger profit than the Seperation criterion.

Limitations of Statistics

- Criteria only look at the statistics. Problem?

- Example: we hire from group A using interview score, we hire from group B randomly.
- Example: we hire from group A using the highest interview score, we hire from group B with the lowest interview score.

- Conclusion: for fairness we need to know how a decision was made
- Statistics: can only judge fairness on group level

A final note

- Separation and Independence cannot always be both satisfied
- So which statistic is *really* fair?
- Still active discussion.
- More than 20 fairness definitions have been proposed recently…

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Conclusions

- Why it’s not enough to ignore sensitive attributes
- Criteria 1: Independence
  - How to satisfy it: (1) different model per group (2) pre-processing
  - Problems with pre-processing, problem with independence criterion
- Criteria 2: Separation
  - Can have different acceptance rate per group
  - How to satisfy it: making a smart model dumber, ROC curve
- We need to keep sensitive attributes
  - To make models more fair
  - To measure fairness