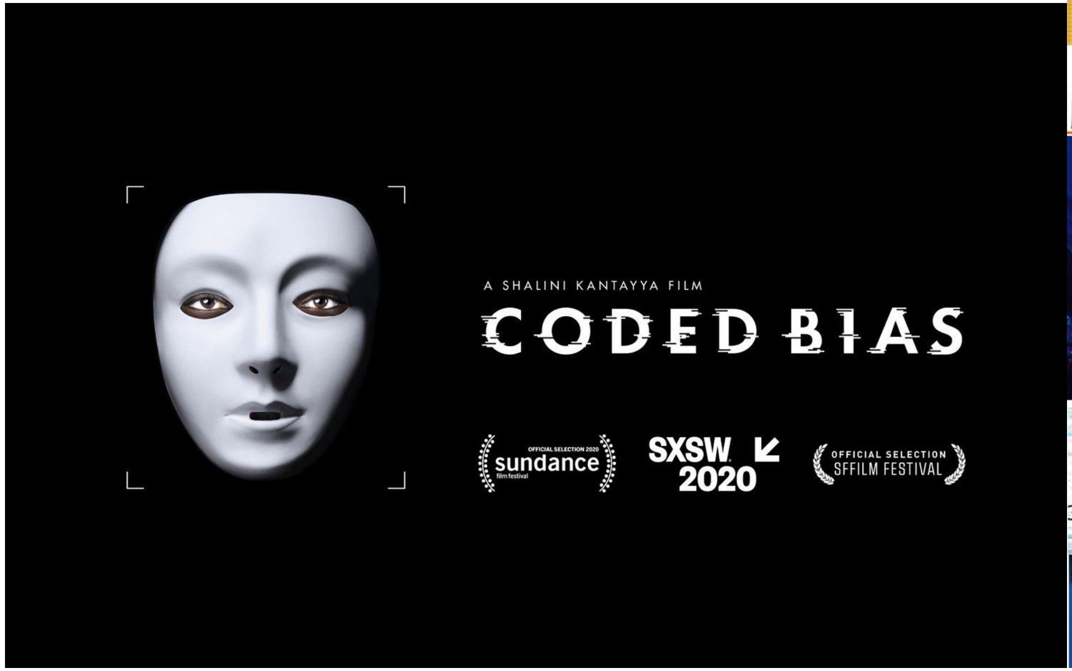
CPSC 340:
Machine Learning and Data Mining

Responsible ML – Bonus Lecture

Lironne Kurzman

Fall 2021

Bonus Lecture – Responsible ML (or what we should *really* be afraid of…)



By the end of this lecture…

- Is a model with E_train = 0 a “correct” model?

By the end of this lecture…

- Is a model with E_train = 0 a “correct” model?
- What about a model with E_test = 0?

By the end of this lecture…

- Is a model with E_train = 0 a “correct” model?
- What about a model with E_test = 0?
- What makes a model “correct”?

By the end of this lecture…

- Is a model with E_train = 0 a “correct” model?
- What about a model with E_test = 0?
- What makes a model “correct”?
- What makes a model fair? Unfair?

By the end of this lecture…

- Is a model with E_train = 0 a “correct” model?
- What about a model with E_test = 0?
- What makes a model “correct”?
- What makes a model fair? Unfair?
- How much do we trust the data? The labels?

By the end of this lecture…

- Is a model with E_train = 0 a "correct" model?
- What about a model with E_test = 0?
- What makes a model "correct"?
- What makes a model fair? Unfair?
- How much do we trust the data? The labels?

- We may not have the answers, but hopefully you will keep these questions in mind

Responsible ML

- Recent umbrella term referring to ethical practices, fairness, and governance within the field

Responsible ML

- Recent umbrella term referring to ethical practices, fairness, and governance within the field

- It is **not** a strict set of rules or solutions, it is however a collection of factors to consider when writing decision makers (regardless of what decision they make)

"With great power comes great responsibility…"

Correlation does not imply Causation

- Who’s at fault for biased gender translations?
- E.g., Hungarian is a gender-neutral language, so google assigns a gender based on frequency in the training corpus

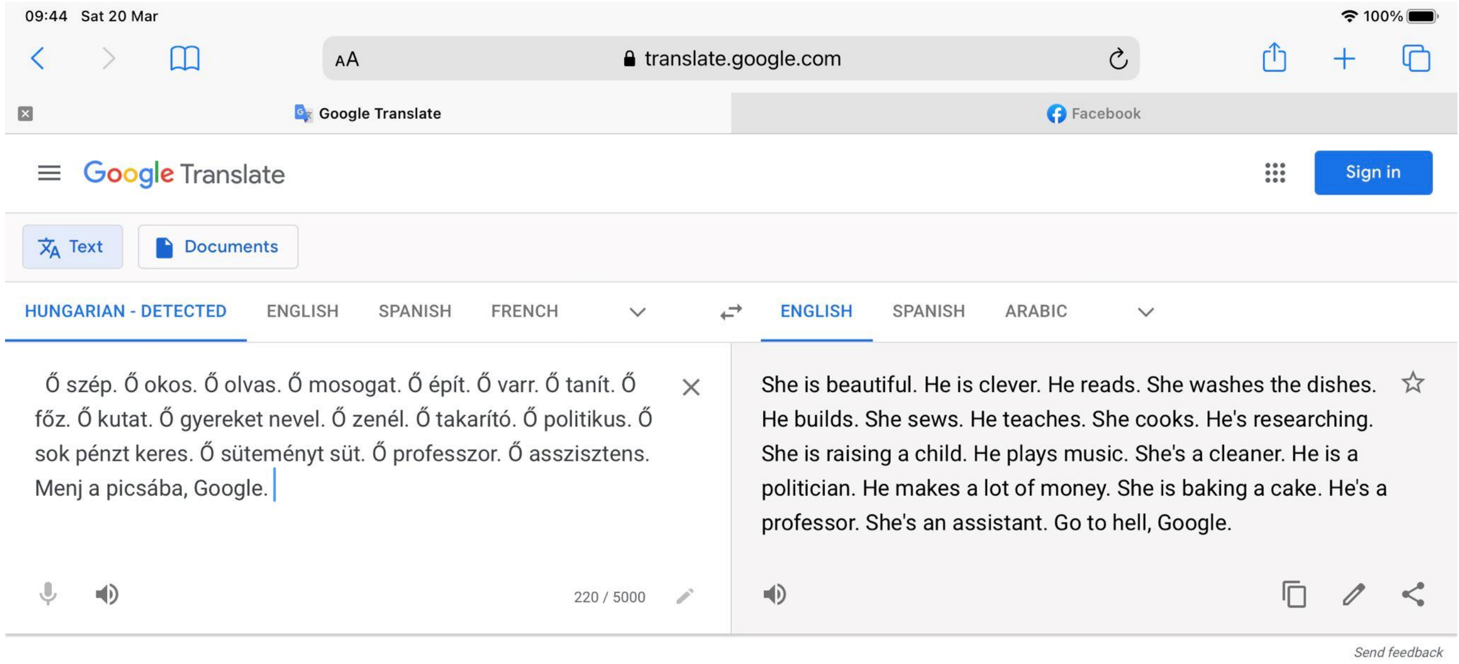


Image taken from twitter @DoraVargha

Correlation does not imply Causation

- What about automating the hiring process?

RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 3 YEARS AGO

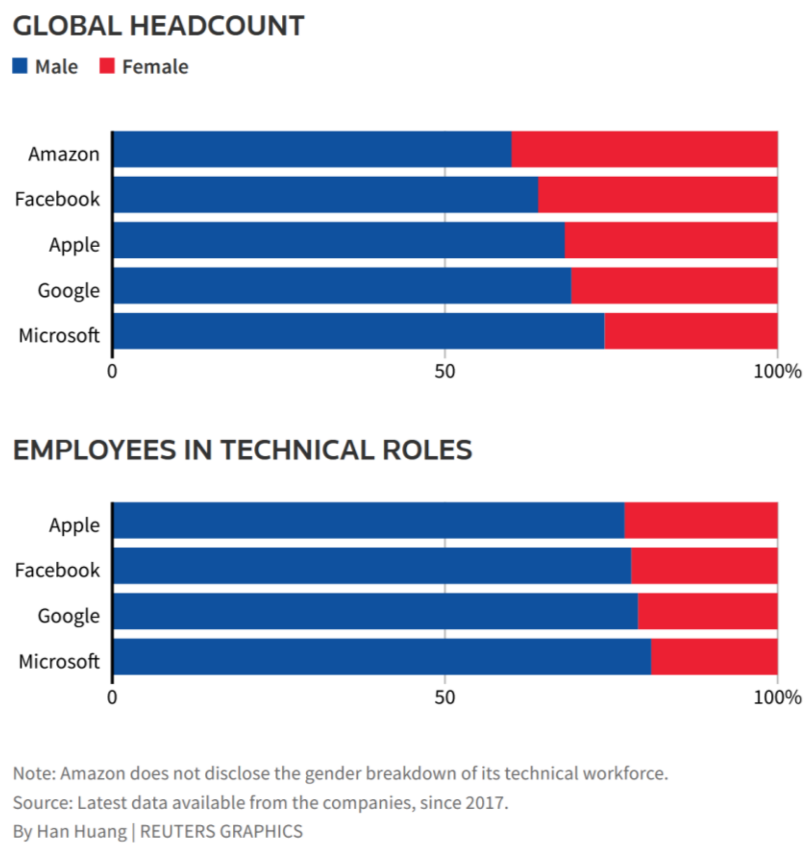
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Correlation does not imply Causation

- What about automating the hiring process?
- Women are less likely to be Software Engineers, therefore women are less likely to be good software engineers?
- The algorithm penalized any candidate that had the word “woman/women” in their resume
  - i.e. “Women’s chess club captain”, “Executive member at Women in CS club” etc.



Who is supervising supervised learning?

- TayTweets was a Chat Bot made by Microsoft
- It was released March of 2016; Tay was designed to learn how to converse from twitter

Who is supervising supervised learning?

- Are the labels coming from a reliable source?

MICROSOFT | WEB | TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

Via The Guardian | Source TayandYou (Twitter)

Who is supervising supervised learning?

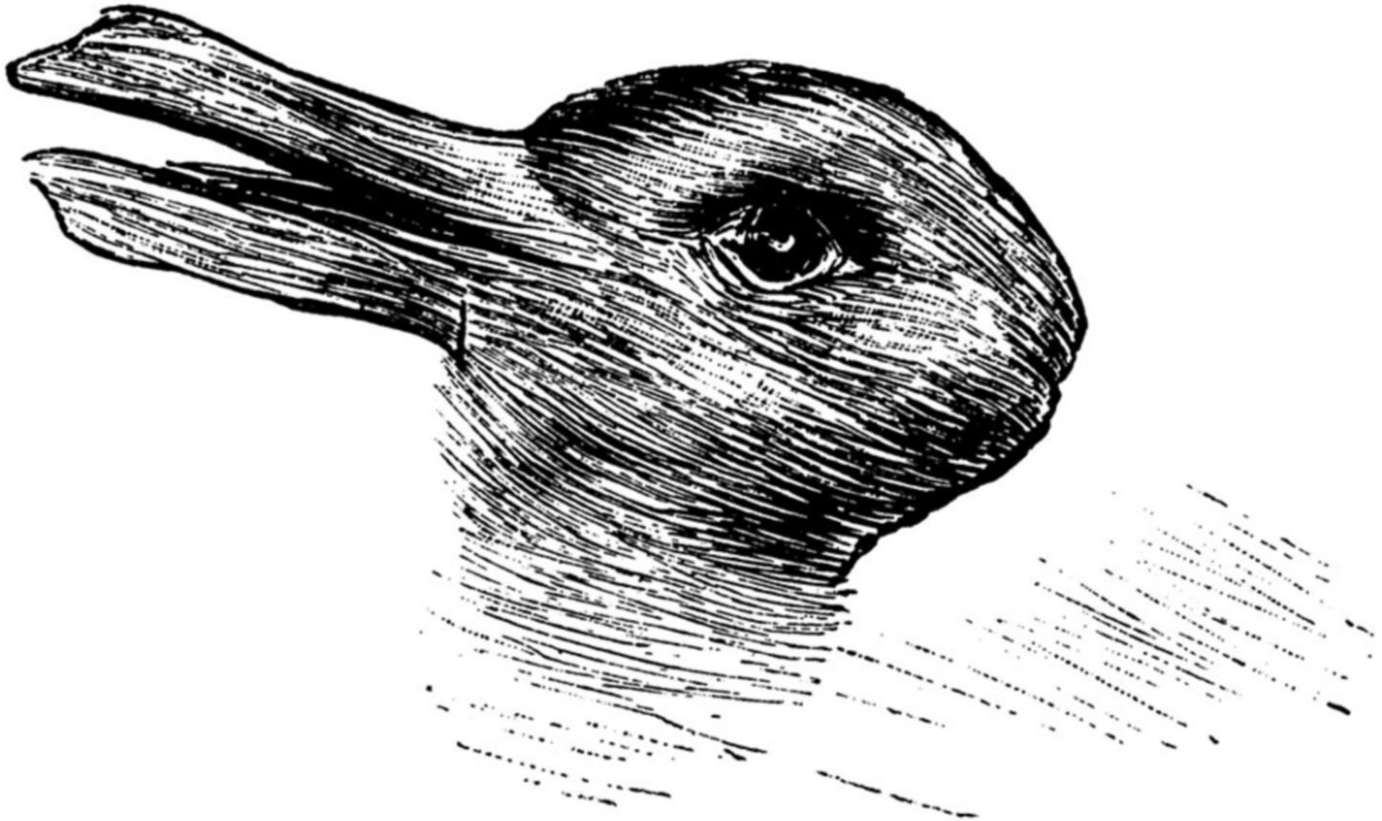
- In under 24 hours, not only did Tay became good at being bad, but Tay also became *really good* at being bad.

Who is supervising supervised learning?

- In less than 24 hours, not only did Tay became good at being bad, but Tay also became *really good* at being bad
- From a technical perspective Tay managed to learn in an extremely short time how to produce complex syntactically correct sentences (if we ignore the semantics)

Who is supervising supervised learning?

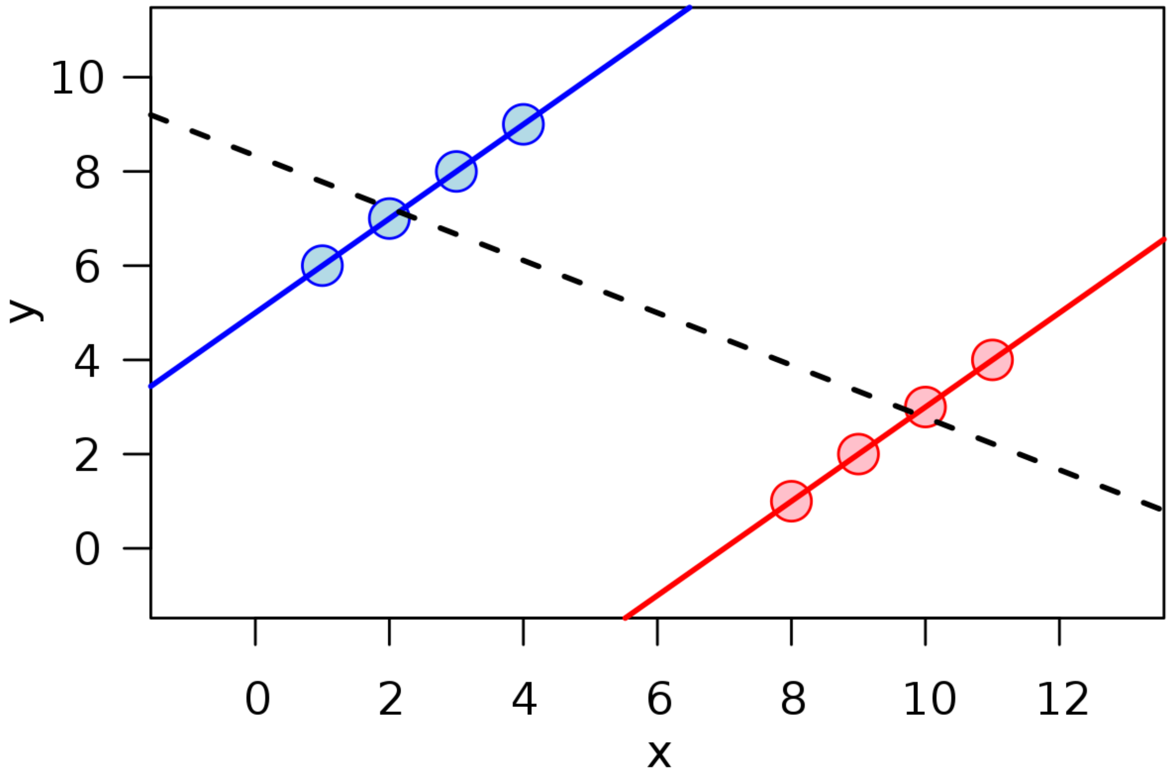
- Can we always agree on the label?



Consider Performance on all Populations

"There are three kinds of lies: lies, damned lies, and statistics." -various

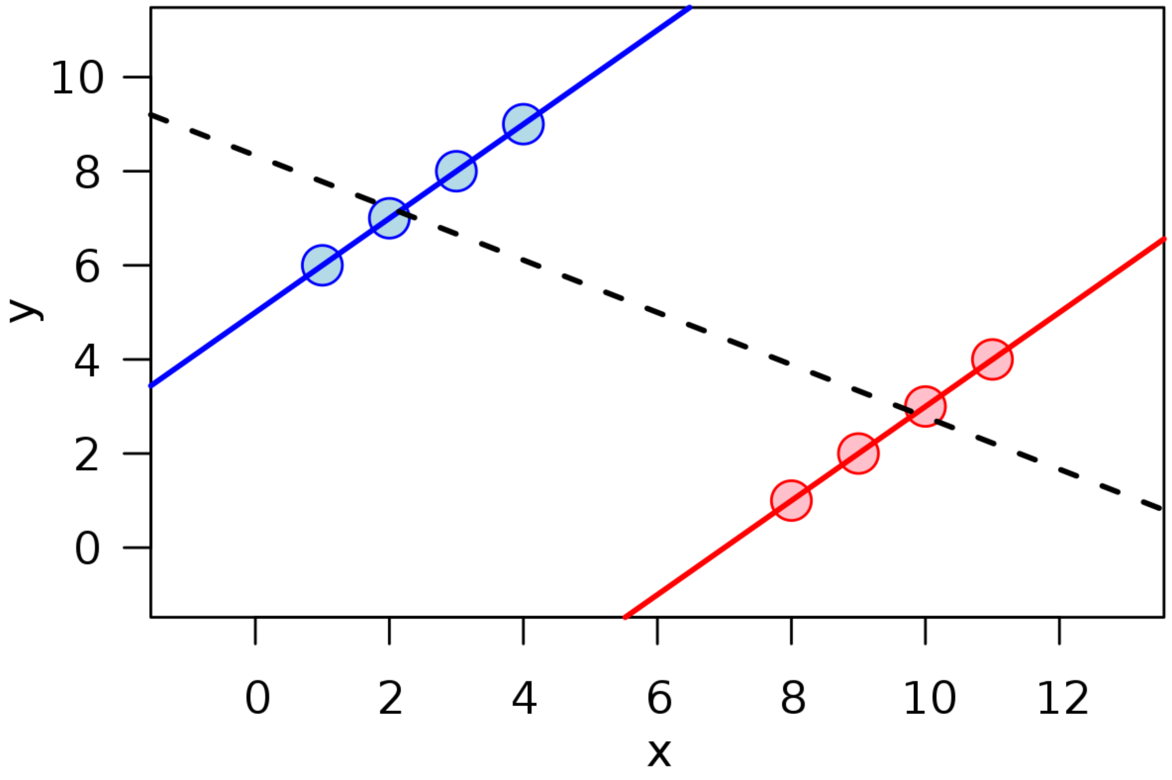
- Simpson's Paradox – When the relationship (trend, correlation coefficient etc.) between variables reverses when you partition the data into sub-categories



Consider Performance on all Populations

"There are three kinds of lies: lies, damned lies, and statistics." -various

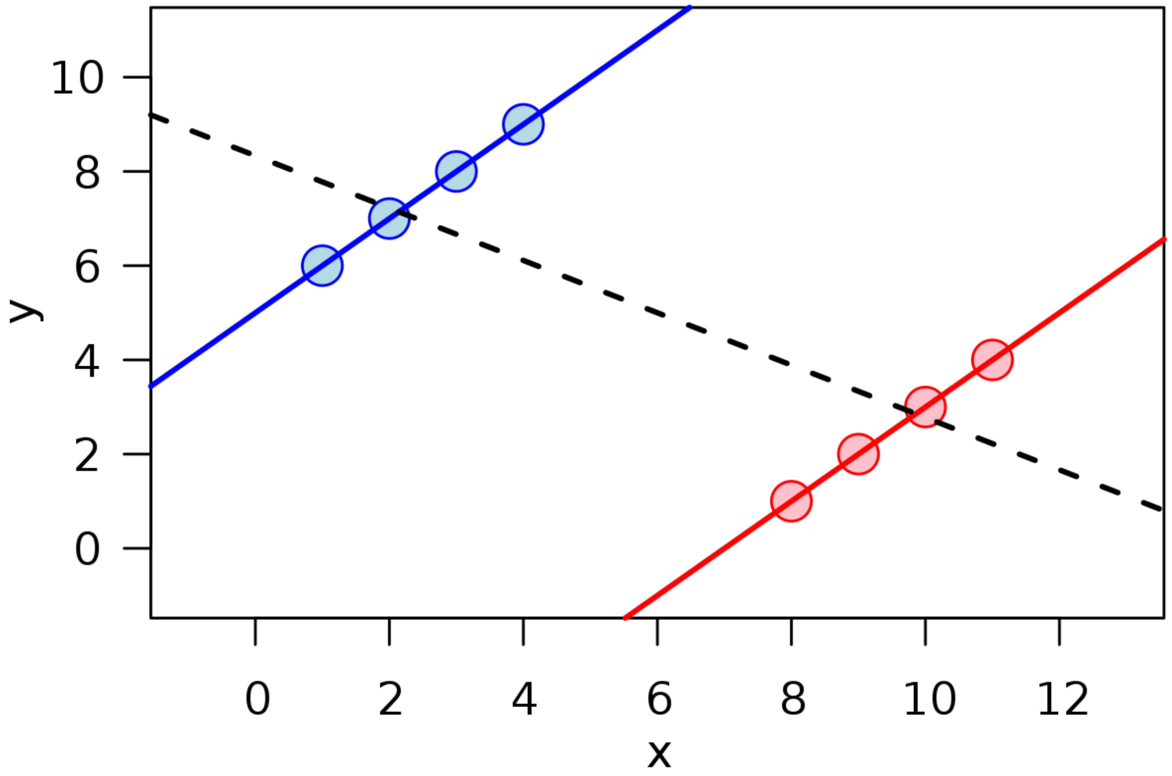
- Simpson's Paradox – When the relationship (trend, correlation coefficient etc.) between variables reverses when you partition the data into sub-categories
- i.e., If Student A had an 83% avg grade on a given year, and Student B has a 78% avg for the same year, who performed better?



Consider Performance on all Populations

"There are three kinds of lies: lies, damned lies, and statistics." -various

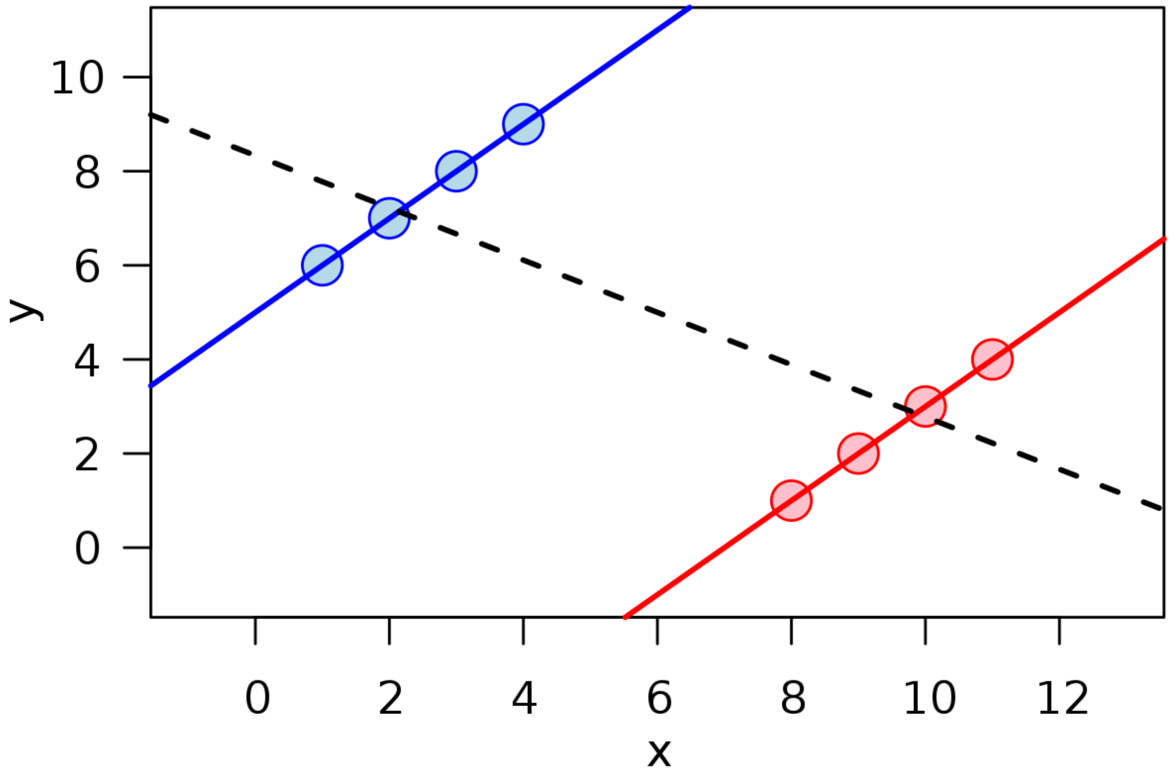
- Simpson's Paradox – When the relationship (trend, correlation coefficient etc.) between variables reverses when you partition the data into sub-categories
- But Student A had an 87% avg grade for term 1 and 69% avg for term 2, While Student B had a 93% avg for term 1 and 73% for term 2.



Consider Performance on all Populations

“There are three kinds of lies: lies, damned lies, and statistics.” -various

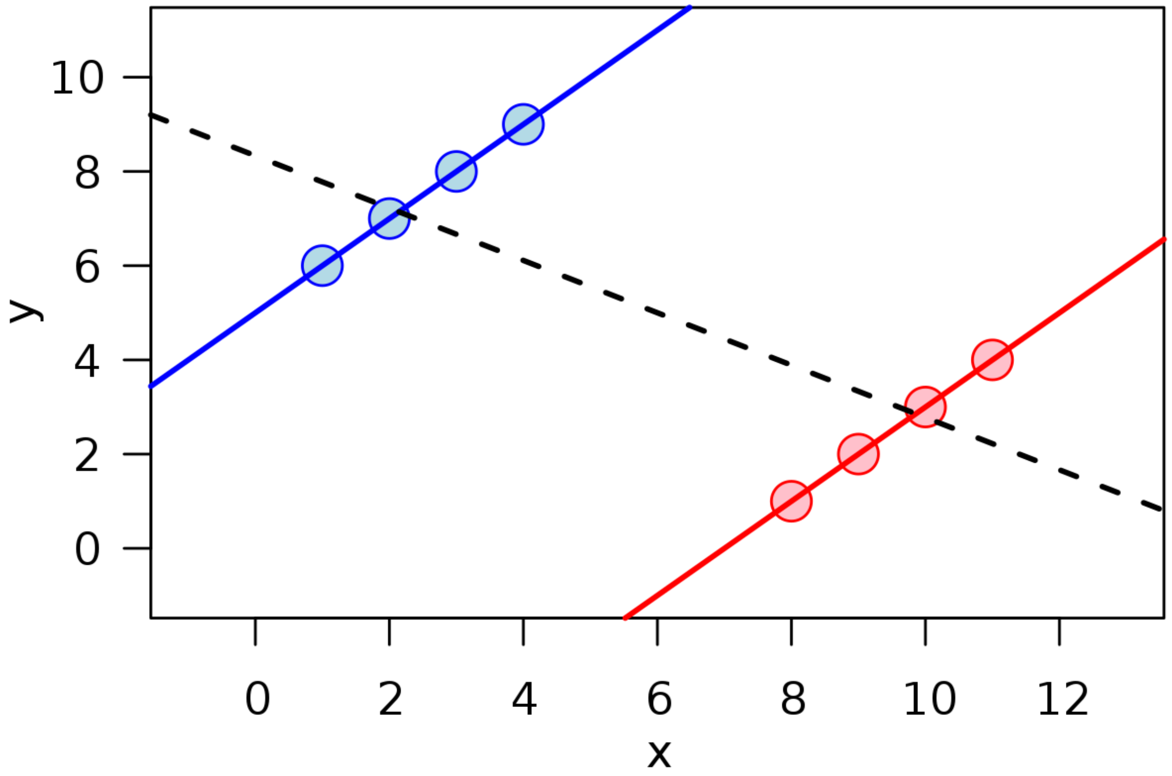
- Simpson’s Paradox – When the relationship (trend, correlation coefficient etc.) between variables reverses when you partition the data into sub-categories
- But Student A had an 87% avg grade for term_1 and 69% avg for term_2, While Student B had a 93% avg for term_1 and 73% for term_2
- This is possible if Student A took 4 classes in term_1 and 2 classes in term_2, and Student B took 1 class in term_1, and 5 classes in term_2



Consider Performance on all Populations

"There are three kinds of lies: lies, damned lies, and statistics." -various

- Simpson's Paradox – When the relationship (trend, correlation coefficient etc.) between variables reverses when you partition the data into sub-categories
- What if these confounding factors are attributes such as gender, race, age, etc.?
- What if they are not yet known?



Maybe a little bias is good?

- When we standardize a process, consider the individuals that fall right bellow the threshold
- If I want to apply for a mortgage, and I apply at 10 different banks

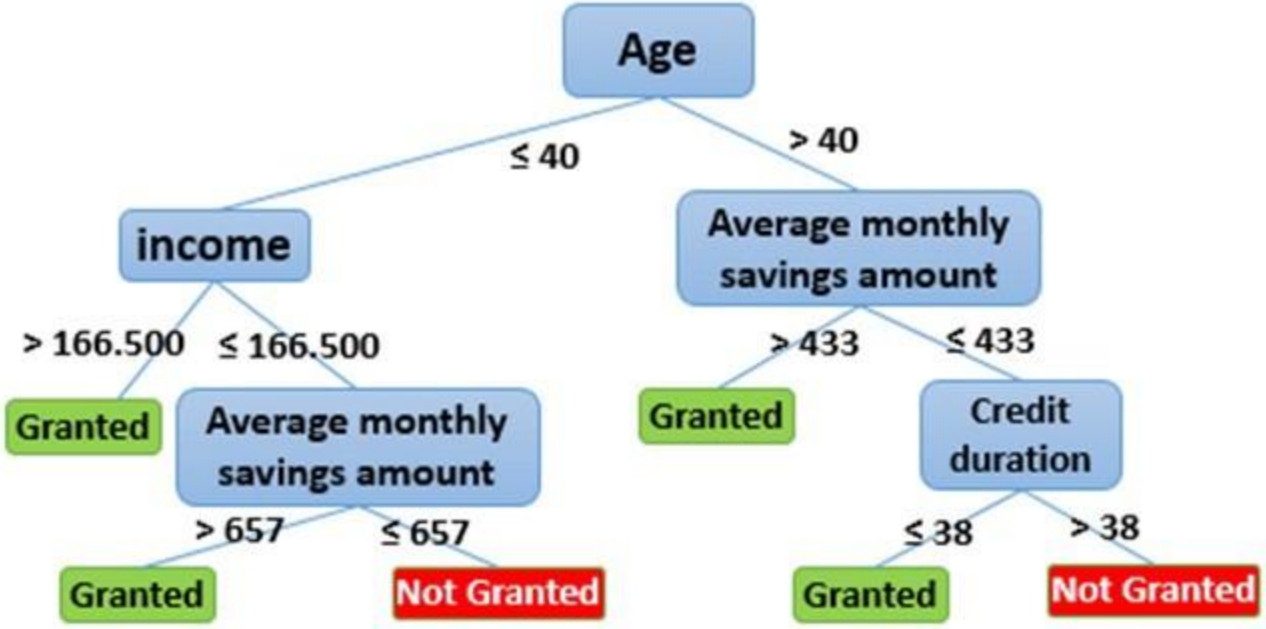


Image taken from Machine learning for Banking: Loan approval use case

Maybe a little bias is good?

- When we standardize a process, consider the individuals that fall right bellow the threshold
- If I want to apply for a mortgage, and I apply at 10 different banks
- If each bank has different decision rules (for example a credit duration of 37 instead of 38, or lower income)
- There’s a chance I’ll get approved

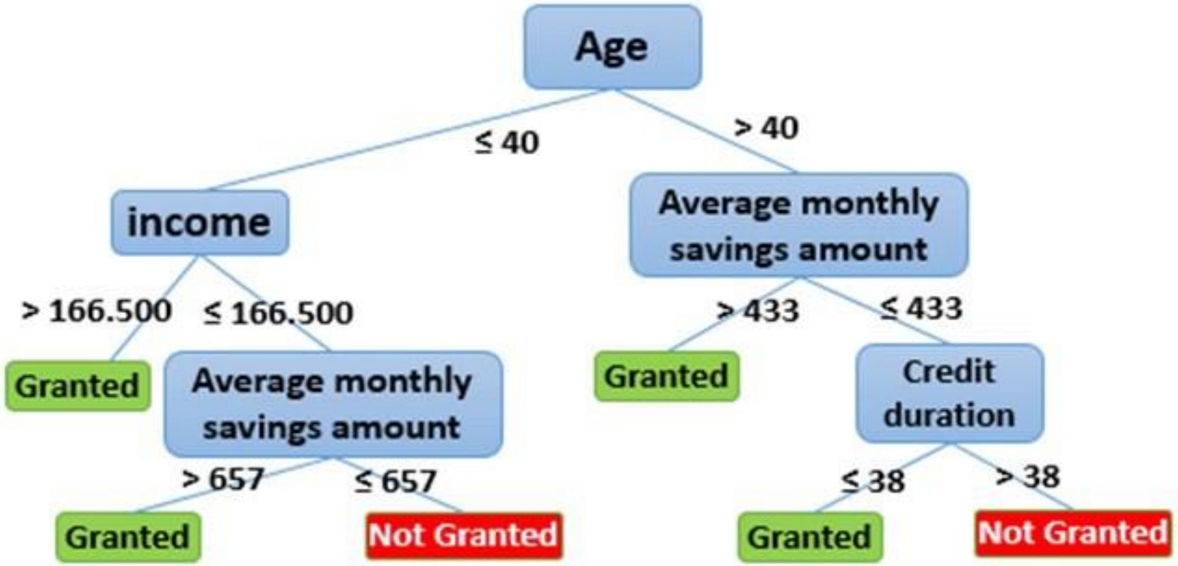


Image taken from Machine learning for Banking: Loan approval use case

Maybe a little bias is good?

- When we standardize a process, consider the individuals that fall right bellow the threshold
- If I want to apply for a mortgage, and I apply at 10 different banks
- If each bank has different decision rules (for example a credit duration of 37 instead of 38, or lower income)
- There's a chance I'll get approved
- But what if my race and gender affect my income and credit?
- If all banks now use the same model, I will be rejected by all banks

Who is the average case?

- Average case performance is not great if you are underrepresented

Kinect May Have Issues with Dark-Skinned Users

By Jane McEntegart November 05, 2010

An interesting post on GameSpot suggests that Microsoft's new motion-sensing peripheral, Kinect, might have problems recognizing the faces of some dark-skinned users.

Who is the average case?

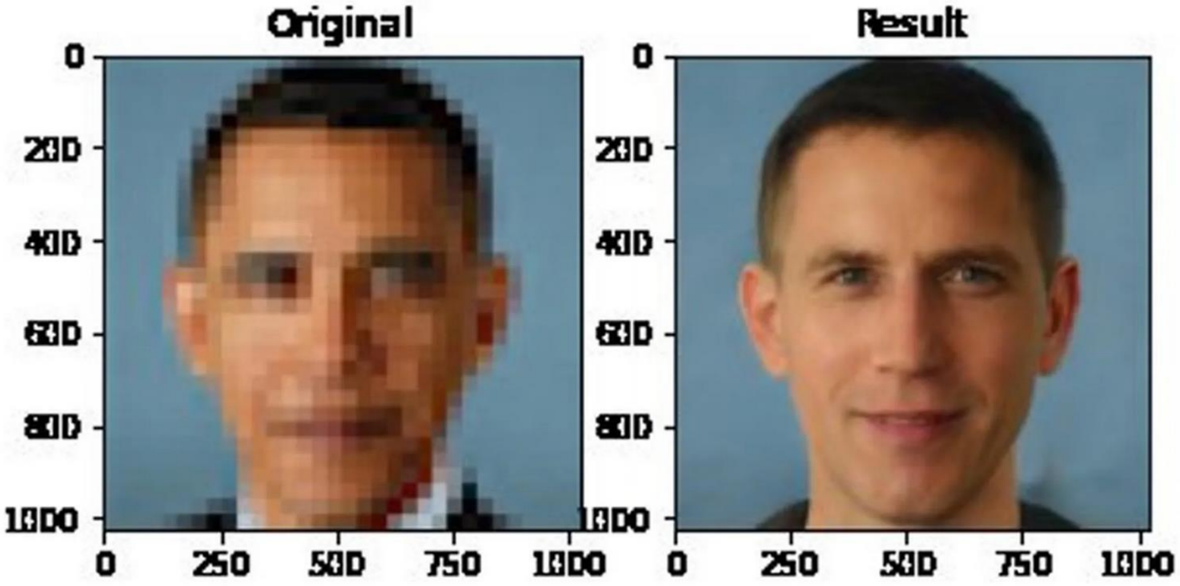
- Average case performance is not great if you are underrepresented
- What if you are an outlier?

Who is the average case?

- Average case performance is not great if you are underrepresented
- What if you are an outlier?
- Imagine a system that finds the best location to build a bathroom for a given floor
- So, min-distance to the bathroom looks at everyone's seating locations and decides on best spot
- What if you are the only non-male on the floor? How likely is min-distance to consider your preference?

Who is the average case?

- Average case performance is not great if you are underrepresented
- We don’t view a system as working until it works for everyone



What makes a decision right?

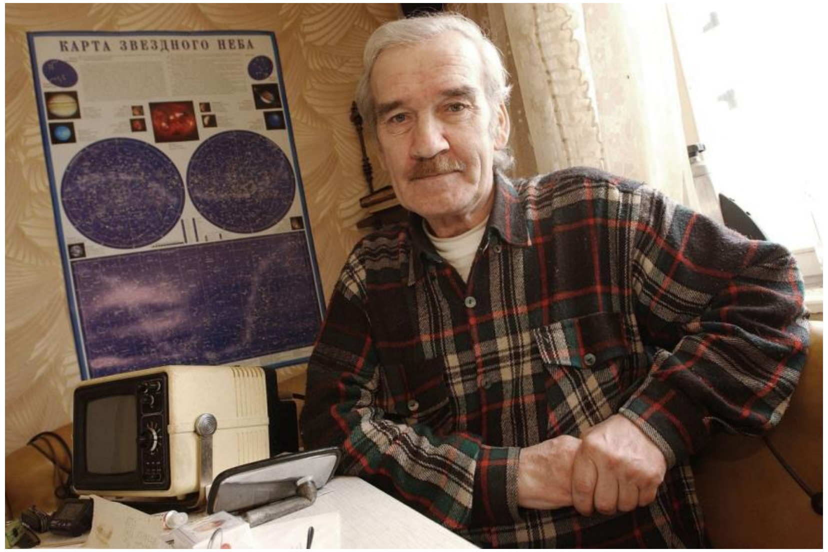
- Can a model learn accountability? Empathy? Are they necessary?

What makes a decision right?

- Can a model learn accountability? Empathy? Are they necessary?
- In 1983 Stanislav Petrov was an officer in the Soviet army, whose job was to register apparent enemy missile launches
- On Sept 26, 1983, a report came in that the U.S. has launched their attack, the system indicated that the reliability of the alert was “highest”
- If Stanislav were to report the attack, the Soviet Army would have retaliated, and a nuclear war ensues.

What makes a decision right?

- “The siren howled, but I just sat there for a few seconds, staring at the big, back-lit, red screen with the word 'launch' on it,“
- “There was no rule about how long we were allowed to think before we reported a strike. But we knew that every second of procrastination took away valuable time;“
- “All I had to do was to reach for the phone; to raise the direct line to our top commanders - but I couldn't move. I felt like I was sitting on a hot frying pan,"
- “There were 28 or 29 security levels. After the target was identified, it had to pass all of those 'checkpoints'. I was not quite sure it was possible, under those circumstances,”



Quotes taken from the BBC

What makes a decision right?

- ''I knew perfectly well that nobody would be able to correct my mistake if I had made one''
- Protocol demanded that the decision would be based on what the systems read out
- When we are wrong, we have to face the consequences of our decisions

Quotes taken from the BBC

What *can* be done?

- try to properly sample the full distribution
- always consider confiders
- *just say no!* to some applications (e.g., recidivism prediction)

Summary

- No model is free of bias sometimes the bias is embedded in the data

    - A model can still be wrong even when it's doing everything right

- Simpson's Paradox relationships in data are not always obvious

    - Even if they appear to be obvious - it can still be misleading!

- No such 'one-fits-all' standard generalization fails for diverse populations
    - Data does not represent all populations equally *or fairly*
    - Would you want to be judged based on numbers alone?

- Models can be wrong sometimes being wrong has grave consequences
    - Who is accountable when a model fails?

- A tool is only as good as its user!

Sounds Interesting?

- Join us at MLRG we start Wednesday next week (13/10) at 1pm
- Subscribe to the mailing list (instructions at https://ml.ubc.ca/mlrg/)

Resources

- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G - Amazon automated the glass ceiling
- https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist - Learning how to be racist from twitter
- https://www.york.ac.uk/depts/maths/histstat/lies.htm - on the origins of “Lies, Damned Lies and Statistics”
- https://medium.com/@fenjiro/data-mining-for-banking-loan-approval-use-case-e7c2bc3ece3 - Machine learning for Banking: Loan approval use case
- https://www.bbc.com/news/world-europe-24280831 - Seeing and doing are two very different things
- https://www.tomsguide.com/us/Microsoft-Kinect-Dark-Skin-Facial-Recognition,news-8638.html – testing at deployment is a bad idea
- https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias - White Obama