VECTOR INSTITUTE

Bias in AI:
Week #3: Introduction to Bias in AI

Instructor:
Sayyed Nezhadi

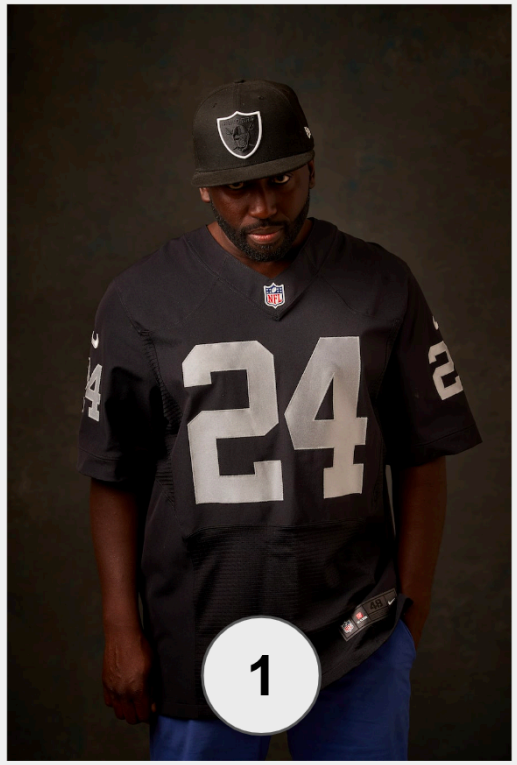
Summer 2022

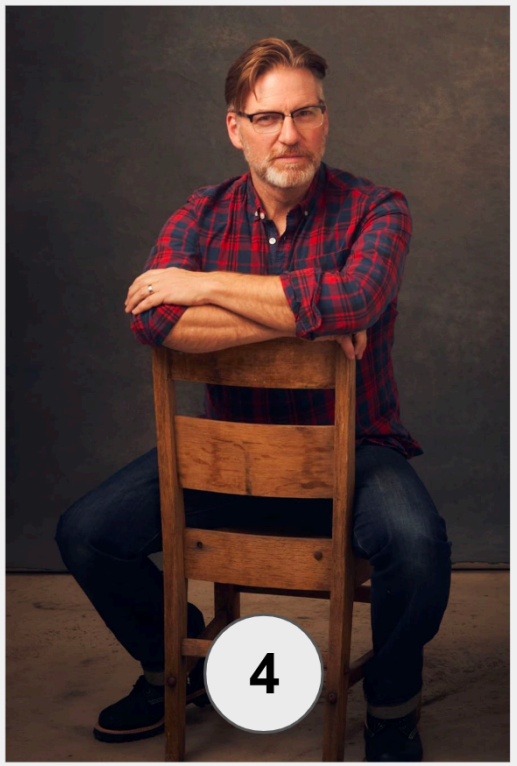
Introduction

- The topic of “Bias in AI” is a relatively recent topic.
- “Bias” and “Fairness” in general are subjective topics and therefore there are different views in some areas.
- The goal of this lecture is to create awareness and to lead a discussion.

Are we biased?

Quiz: Which person works as a highly specialized medical professional?





Source: https://leadwithdiversity.com/testbias

Which person works as a highly specialized medical professional?

Did clothing, or ethnicity impact your decision on who might be a health specialist? Research shows our views of “blackness” tend to be associated with crime and that it’s not uncommon to perceive South Asian and Asian people to be in health. What factors influenced your decision?

Famous Quiz

A man and his son are in a terrible accident and are rushed to the hospital in critical care. The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!”

How could this be?

Research: Mikaela Wapman (CAS'14) and Deborah Belle

- Even young people and self-described feminists tended to overlook the possibility that the surgeon in the riddle was a she.

- Only a small minority of subjects - 15% of the children (ages of 7 to 17)and 14% of the BU students - came up with the mom's-the-surgeon answer.

- Life experiences that might suggest the mom answer had no association with how one performed on the riddle. For example, the BU student cohort, where women outnumbered men two-to-one, typically had mothers who were employed or were doctors and yet they had so much difficulty with this riddle.

Image Labeling Exercise

What is this?

Image Labeling Exercise

What is this?

An apple

Image Labeling Exercise

What is this?

Image Labeling Exercise

What about this?

Image Labeling Exercise

What about this?

A purple apple

Imaginary?

Image Labeling Exercise

Did you know purple apple exists?

The Black Diamond Apple is a rare breed from the family of Hua Niu apples that is cultivated in the Tibetan region of Nyingchi.

Image Labeling Exercise

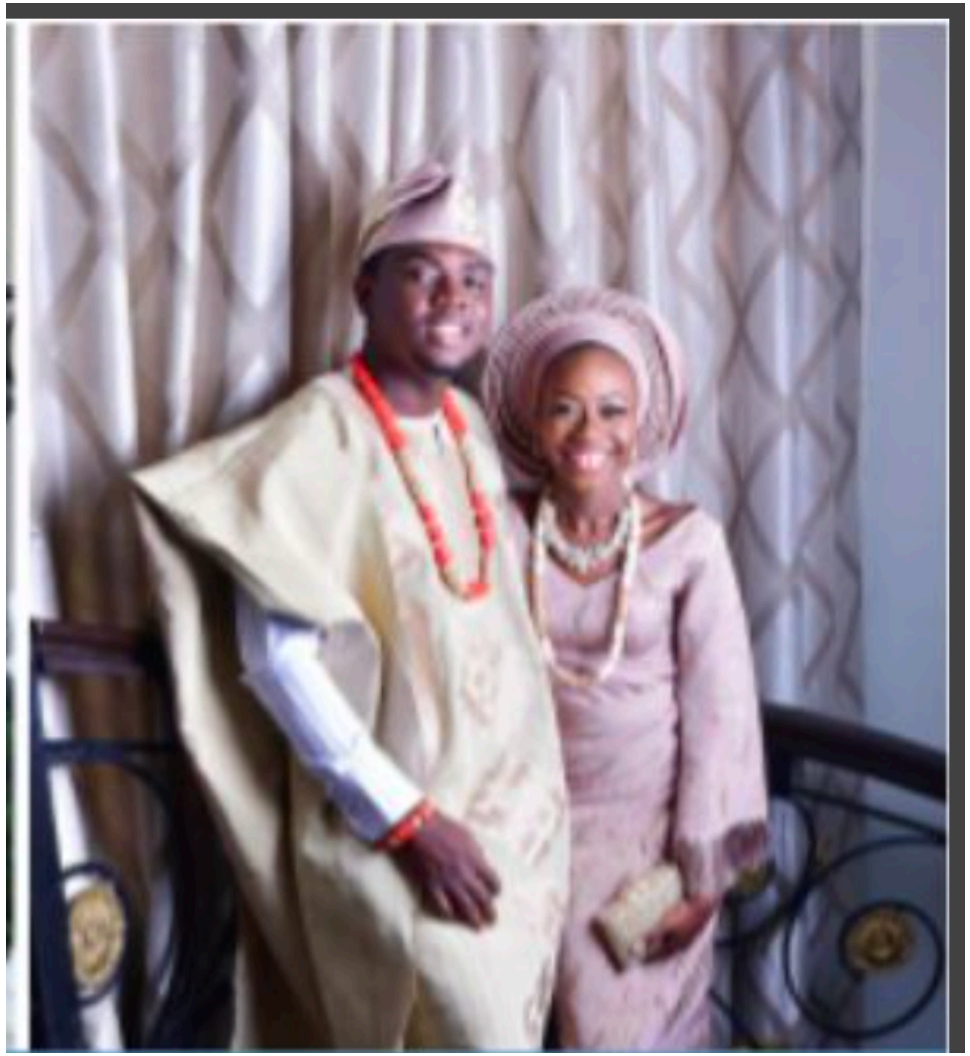


Image Labeling Exercise

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Image Labeling Exercise

Image Labeling Exercise

A white police is beating a black man

Human Bias is a Social Reality

ARE EMILY AND GREG MORE EMPLOYABLE
THAN LAKISHA AND JAMAL?
A FIELD EXPERIMENT ON LABOR MARKET DISCRIMINATION

White sounding name. The results show significant discrimination against African-American names: White names receive 50 percent more callbacks for interviews. We also find that race affects the benefits of a better resume. For White names, a higher quality resume elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase. Applicants living in better neighborhoods receive more callbacks but, interestingly, this effect does not differ by race. The

NATIONAL BUREAU OF ECONOMIC RESEARCH
July 2003

Human Bias is a Social Reality

Man kills 2 neighbors,claims self defense - CNN Video

Sep. 6, 2013 — A Florida **man** wants to use the "Stand Your Ground" law as **self-defense** for **killing** two **neighbors** at a BBQ.

Asian Man Fights & Kills Robbery Suspect in San Francisco ...

Oct. 3, 2020 — **Asian Man** Fights & **Kills Robbery** Suspect in **San Francisco**. **asian-**dawn.com/2020/1... News.

France attack: Three killed in 'Islamist terrorist' stabbings ...

Oct. 29, 2020 — France attack: Three killed in '**Islamist terrorist**' stabbings ... managed to flee to a nearby cafe after being **stabbed** several times, but died later.

Can AI Help?

In many cases, AI can reduce humans' subjective interpretation of data, because machine learning algorithms learn to consider only the variables that improve their predictive accuracy, based on the training data used.

McKinsey Global Institute, June 2019

Human Decisions and Machine Predictions

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan

Published: 26 August 2017

judges. Even accounting for these concerns, our results suggest potentially large welfare gains: a policy simulation shows crime can be reduced by up to 24.8% with no change in jailing rates, or jail populations can be reduced by 42.0% with no increase in crime rates. Moreover, we see reductions in all categories of crime,

Automated underwriting in mortgage lending: Good news for the underserved?

Susan Wharton Gates, Vanessa Gail Perry & Peter M. Zorn

Pages 369-391 | Published online: 31 Mar 2010

Using information from Freddie Mac's Loan Prospector AU service, we provide statistics useful in examining these issues. The data strongly support our view that AU provides substantial benefits to consumers, particularly those at the margin of the underwriting decision. We find evidence that AU systems more accurately predict default than manual underwriters do. We also find evidence that this increased accuracy results in higher borrower approval rates, especially for underserved applicants.

But It Can Go Wrong …

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

SCIENTIFIC AMERICAN

But It Can Go Wrong …

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Source: the Guardian

Amazon's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

But It Can Go Wrong …

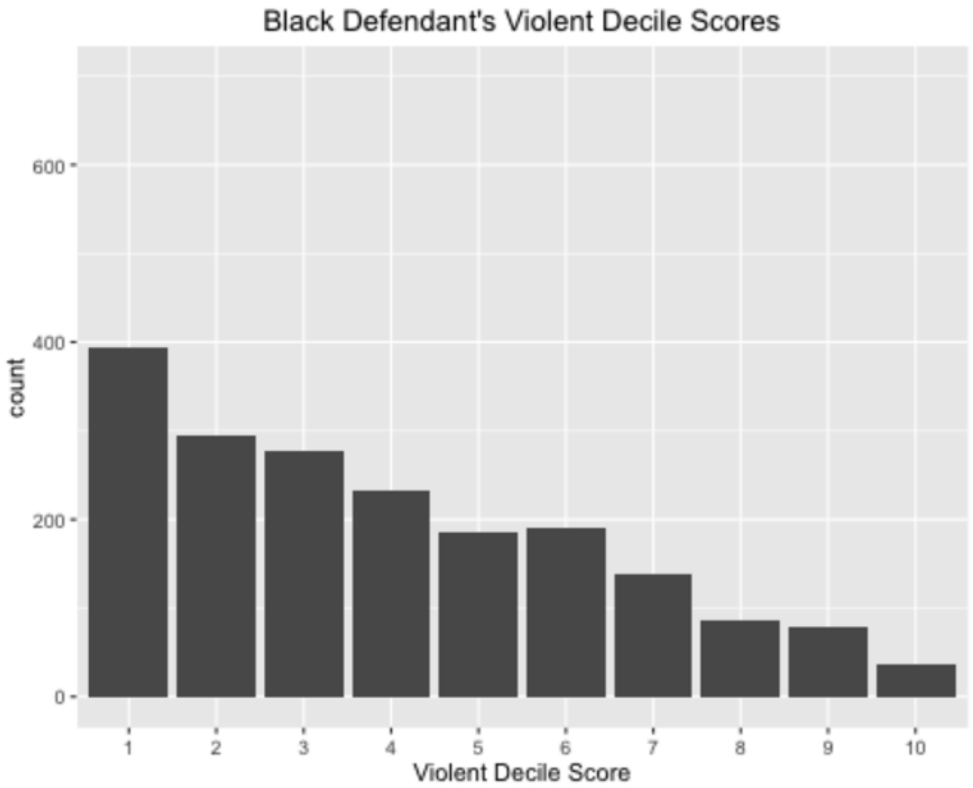
Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

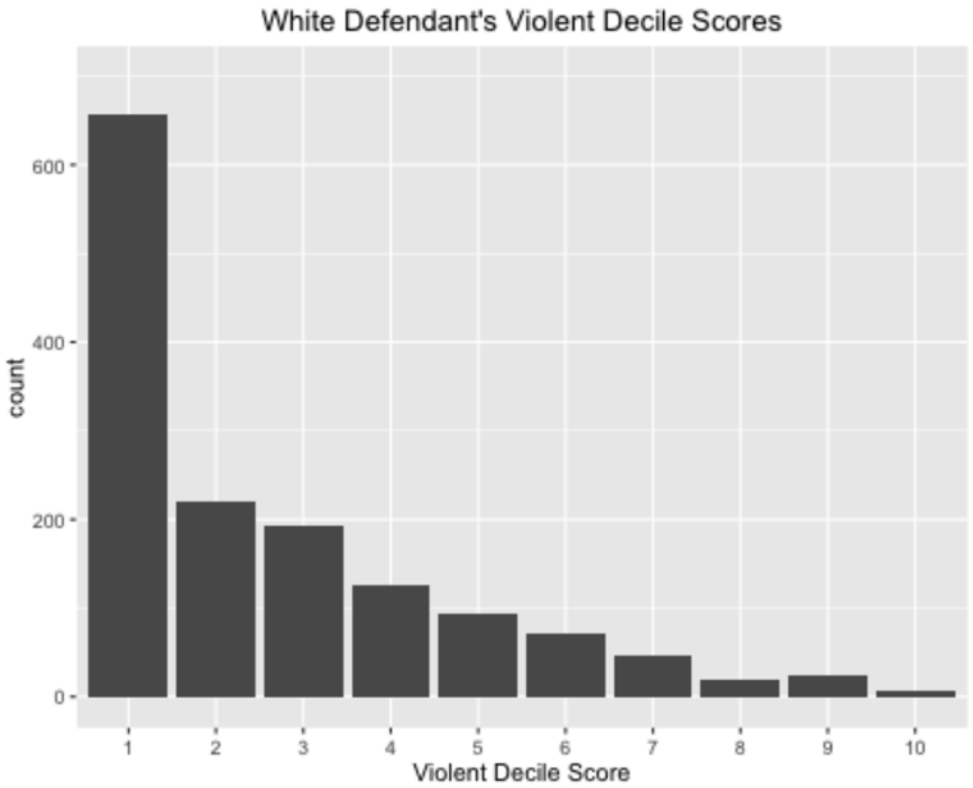
Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler and supporting Donald Trump

Source: the Guardian

But It Can Go Wrong …

Analyzing COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) Algorithm





While there is a clear difference between the distributions of COMPAS scores for white and black defendants, merely looking at the distributions does not account for other demographic and behavioral factors.

Source: *J. Larson, S. Mattu, L. Kirchner and J. Angwin, 2016*

But It Can Go Wrong …

Apple co-founder says Apple Card algorithm gave wife lower credit limit (Router, 2019)

On Saturday, Wozniak chimed in with a similar experience, saying he got 10 times more credit on the card, compared with his wife.

"We have no separate bank or credit card accounts or any separate assets," Wozniak said on Twitter, in reply to Hansson's original tweet.

"Hard to get to a human for a correction though. It's big tech in 2019."

But It Can Go Wrong …

Facebook Halts Ad Targeting Cited in Bias Complaints Ney York Times, 2019

In 2019, Facebook was found to be in contravention of the US constitution, by allowing its advertisers to deliberately target adverts according to gender, race and religion, all of which are protected classes under the country's legal system. Job adverts for roles in nursing or secretarial work were suggested primarily to women, whereas job ads for janitors and taxi drivers had been shown to a higher number of men, in particular men from minority backgrounds. The algorithm learned that ads for real estate were likely to attain better engagement stats when shown to white people, resulting in them no longer being shown to other minority groups.

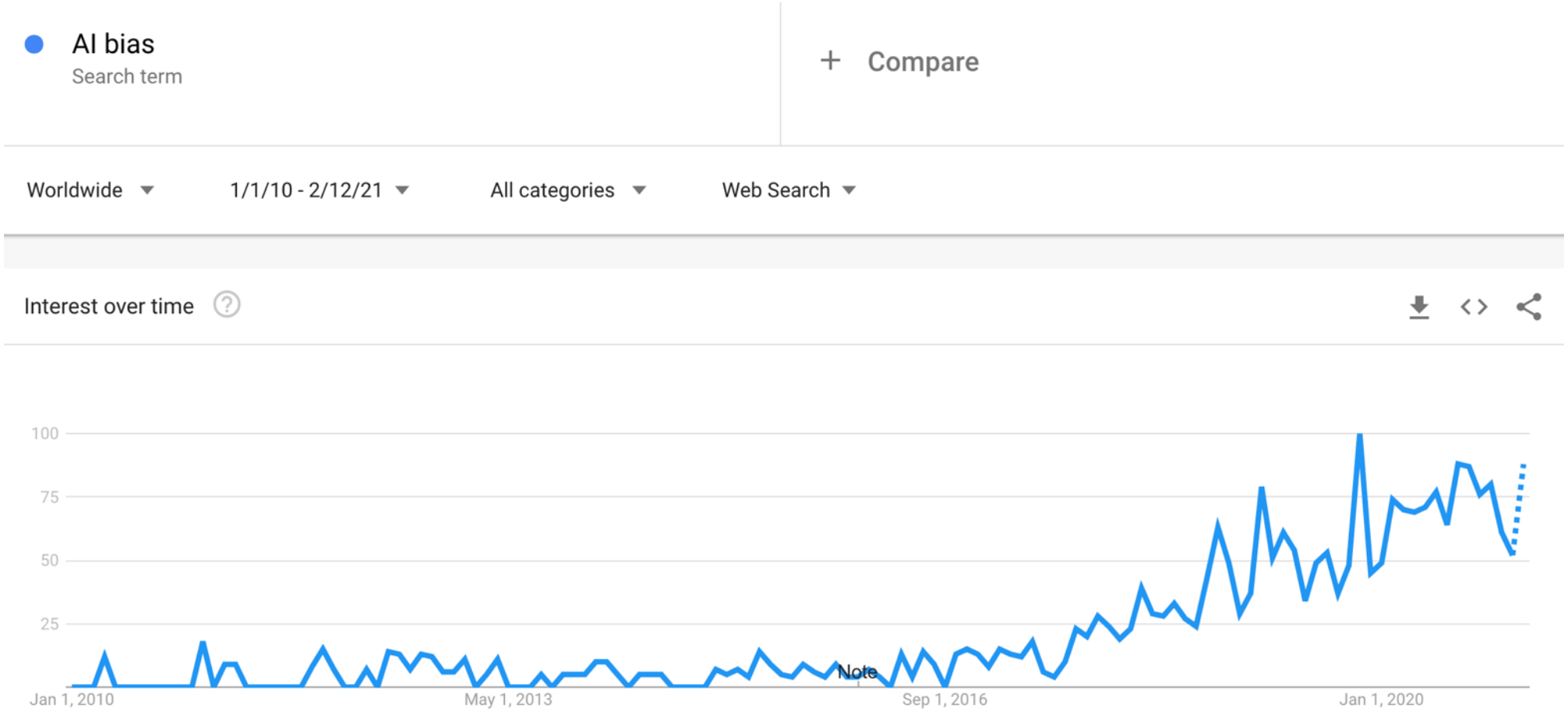
But It Can Go Wrong …

Facebook Halts Ad Targeting Cited in Bias Complaints Ney York Times, 2019

Class Discussion

The Good News

Google Trends Shows Increasing Attention to “AI Bias”



The Good News

Was there anything wrong with my analysis?
Was there any bias?

The Good News

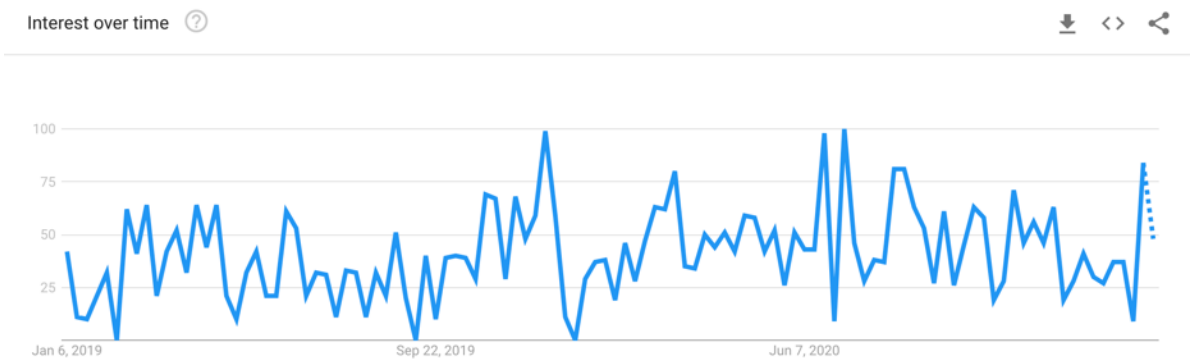
Was there any bias?

1. Can I trust what Google tells me? I don't know the following:
   - What data was used?
   - What algorithm was used?
   - What about other similar words?

Automation Bias

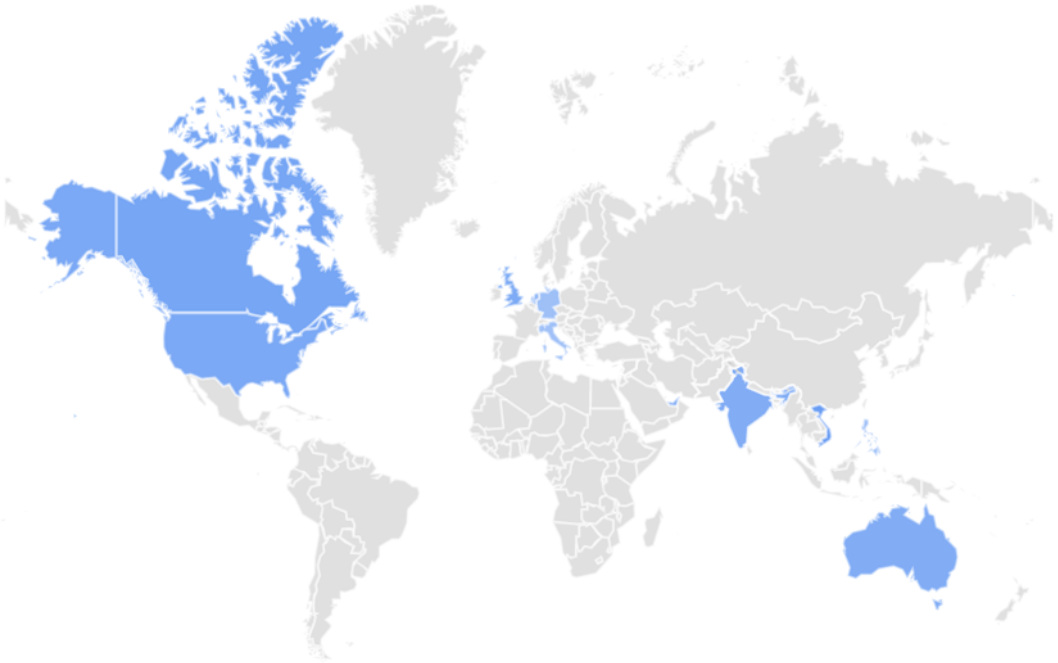
2. Conveniently I used 10 years to prove my point. Below is a 2-year graph. Is it still growing?

Confirmation Bias



The Good News

3. Let's look at the interest by region:



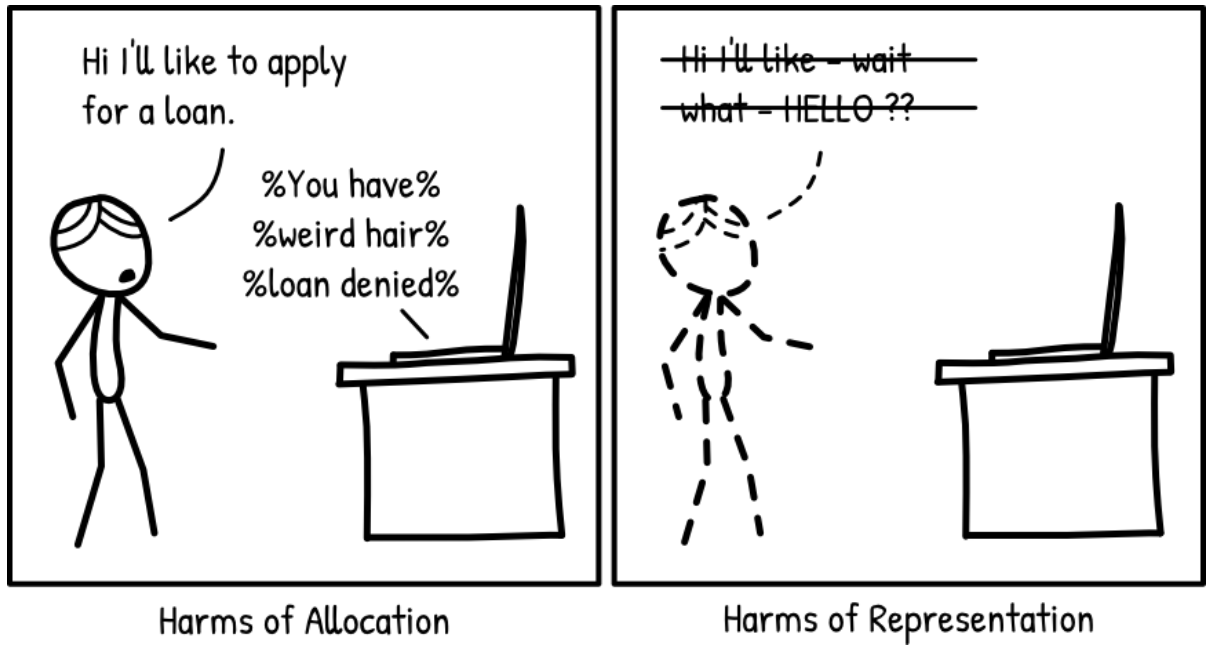
4. Is the growth really caused by more interest in "Bias in AI" or caused by more interest in "AI" in general?

Correlation vs. Causation

Do you see a problem? Is there a bias?

Bias: Types of Harms

- Allocation Harm:
  - AI system allocates opportunities or resources to certain groups or withholds them.
  - Credit scoring models that help banks filter loan applications “allocate” loans. Hiring models help companies to “allocate” jobs.



Source: Understanding Bias (machinesgonewrong.com)

- Representation Harm:
  - AI systems reinforces discrimination against some groups because of identity markers such as gender, race, class, age, ability or belief.
  - Recommended news, ads, and search results (a.k.a. filters) affect or reinforces our perceptions and thoughts about the world.

Is Bias Always Bad?

Class Discussion

Good Bias - Examples

Bias in ML/Stats:

Research Topic:

Alcohol Flush Reaction (AFR)

A condition in which a person develops flushes or blotches associated with consuming alcoholic beverages. It is mostly happening in specific ethnicities.

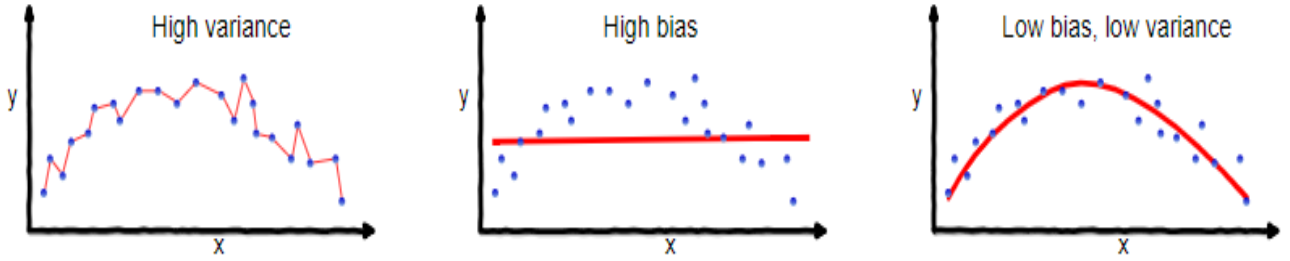
Does it make sense to have a unbiased dataset?

Even among those ethnicities, does it make sense to include Muslims?

* Muslims don't usually drink alcohols

Bias and Fairness

- Statistical bias
  (bias-variance tradeoff)



- Unwanted bias
  with no social impact

fairness | AMERICAN DICTIONARY

fairness

noun [U]

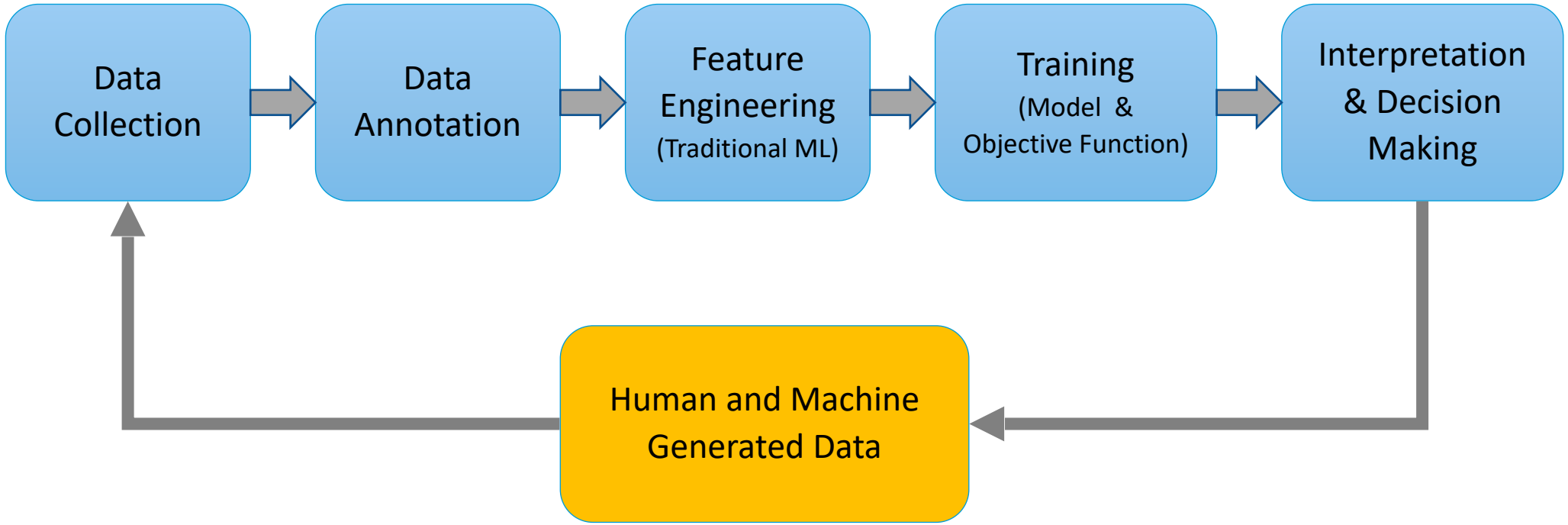
US /ˈfeər·nəs/

the quality of treating people equally or in a way that is right or reasonable:

- Unfair bias
  with social impact

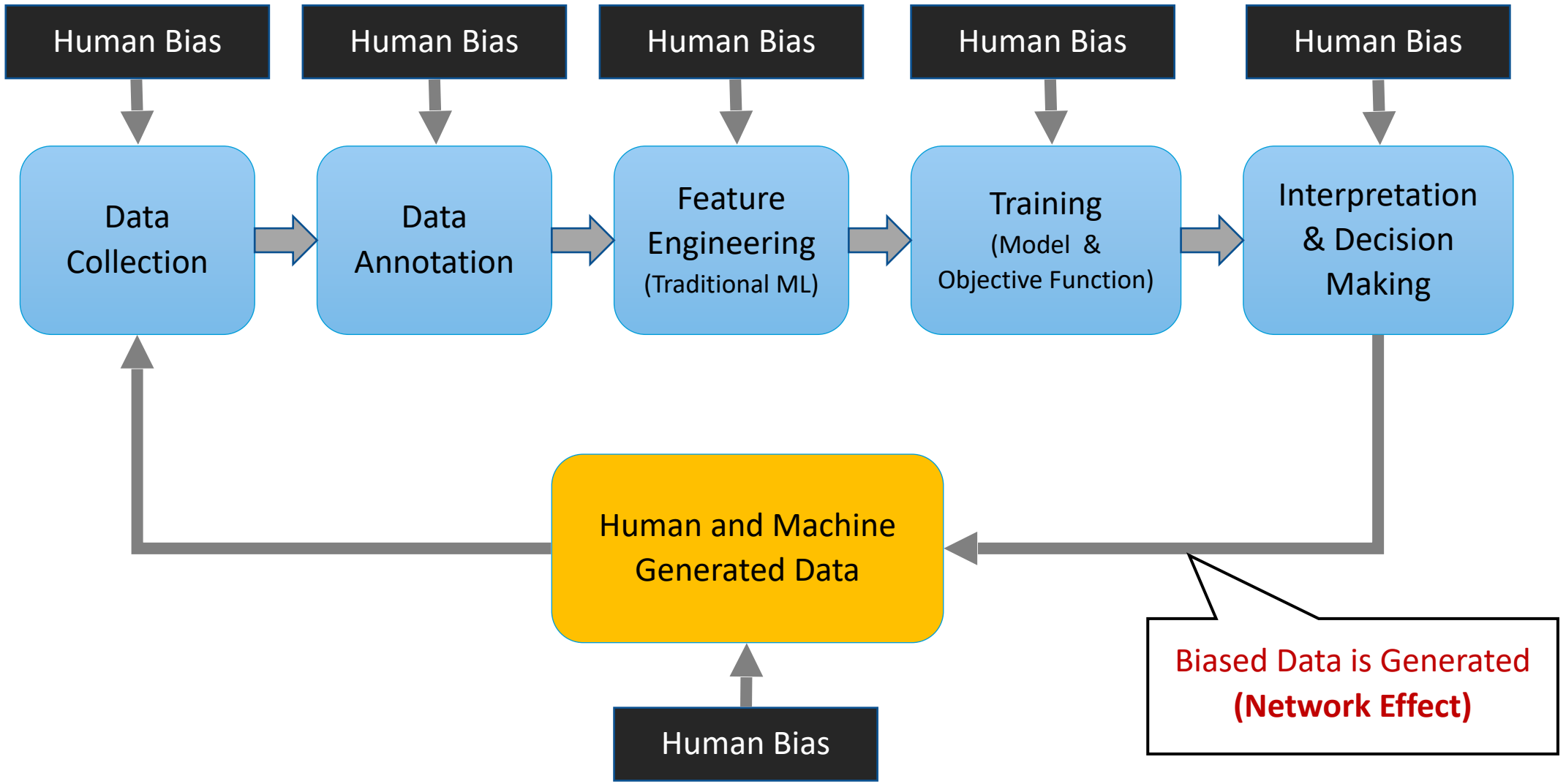
An algorithm is fair if its outcome is independent of given sensitive variables, such as race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization.

Data to Decision Cycle



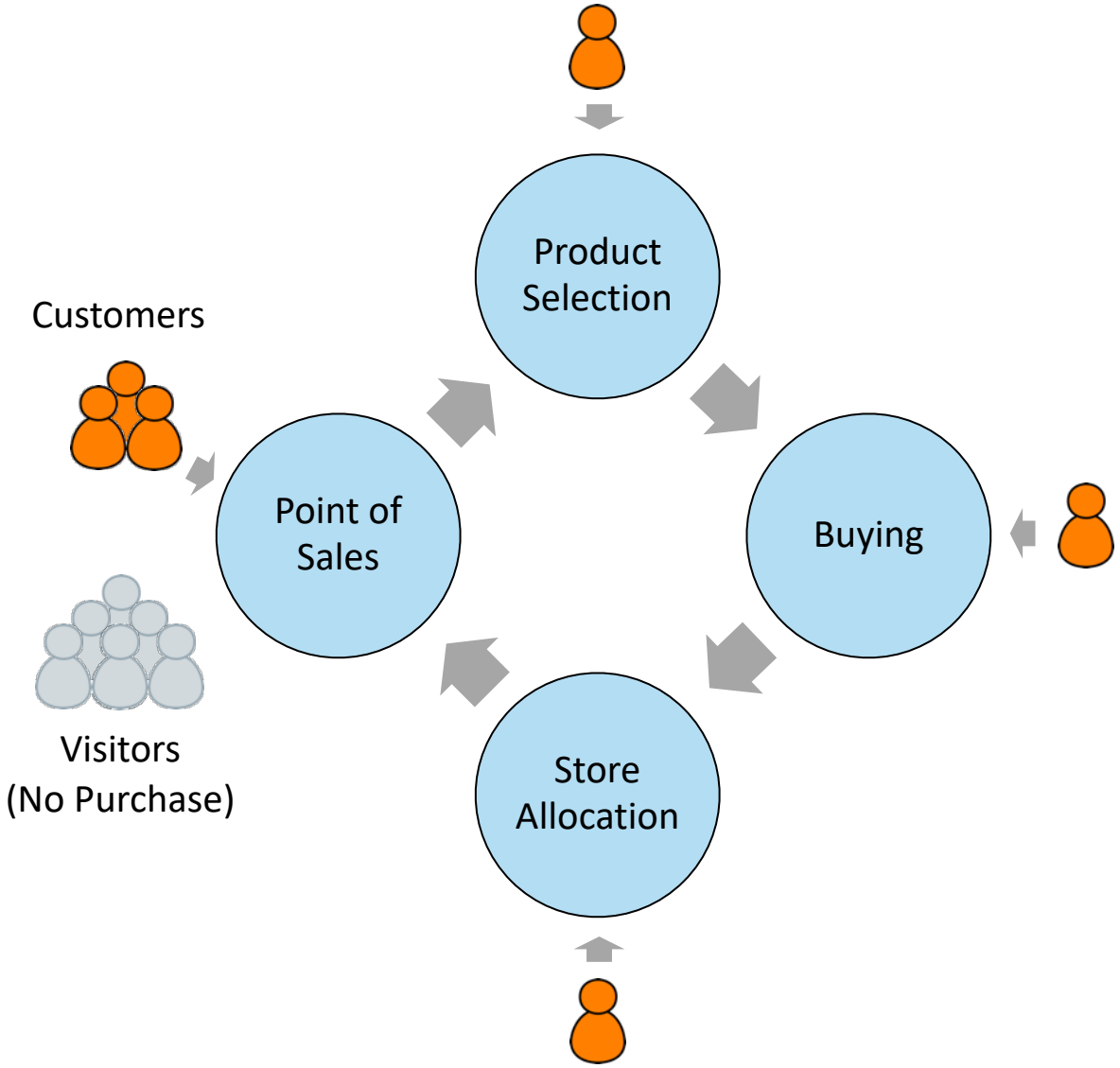
Where in this cycle Bias is introduced?

Data to Decision Cycle



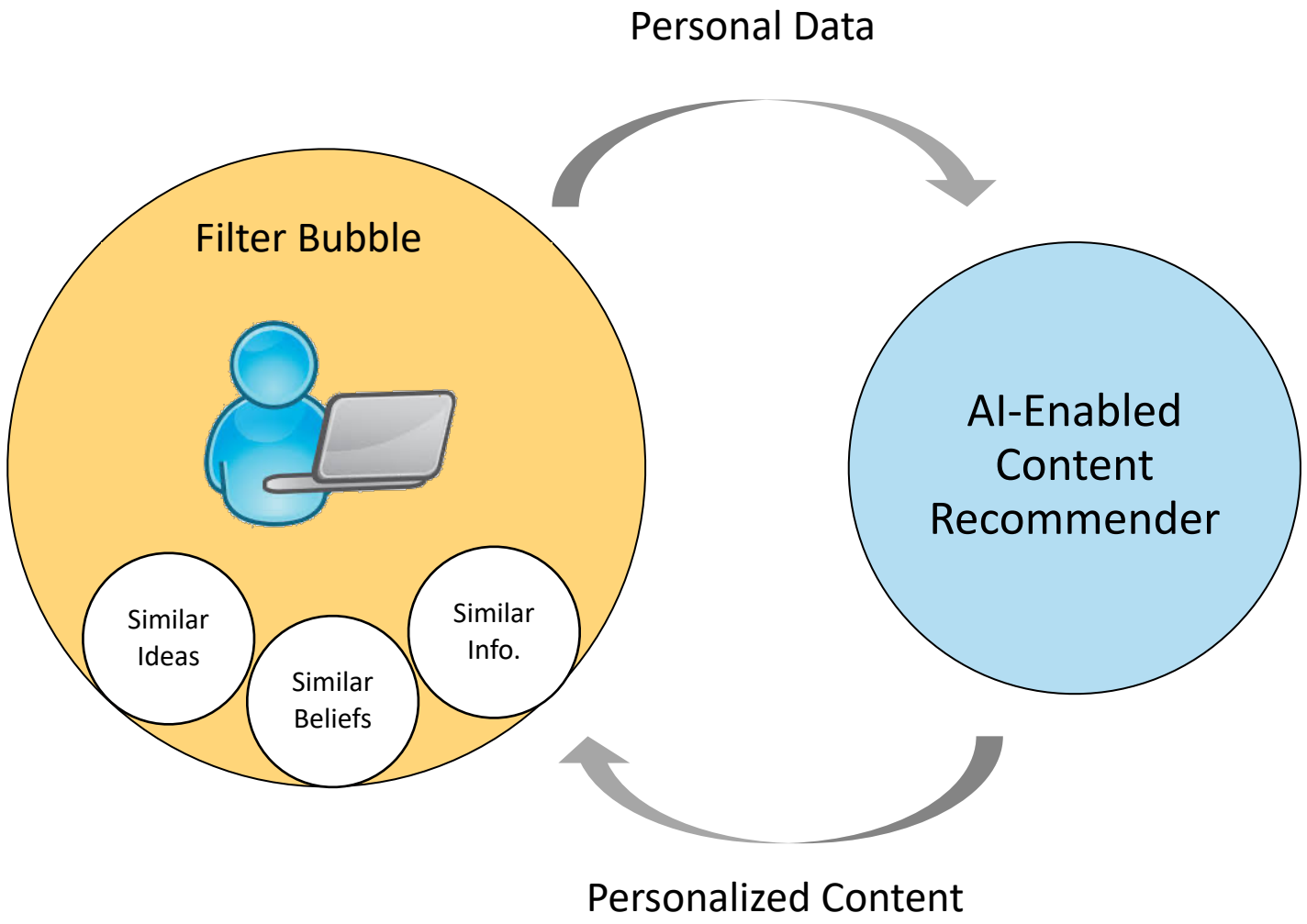
Network Effect Example: Retail Process

- Product selection is done by human (augmented by data)
- Buying process is affected by human and product availability
- Store allocation is also affected by human
- POS data is only affected by purchasing customers (only 5% of the visitors)
- POS data is used for merchandise planning (network effect)



Network Effect Example: Filter Bubble

- A filter bubble is a term term coined by the Internet activist Eli Pariser to refer to a state of intellectual isolation that can result from personalized searches when a website algorithm selectively guesses what information a user would like to see based on information about the user, such as location, past click-behavior and search history.
- As a result, users become separated from information that disagrees with their viewpoints, effectively isolating them in their own cultural or ideological bubbles.



Source: wikipedia

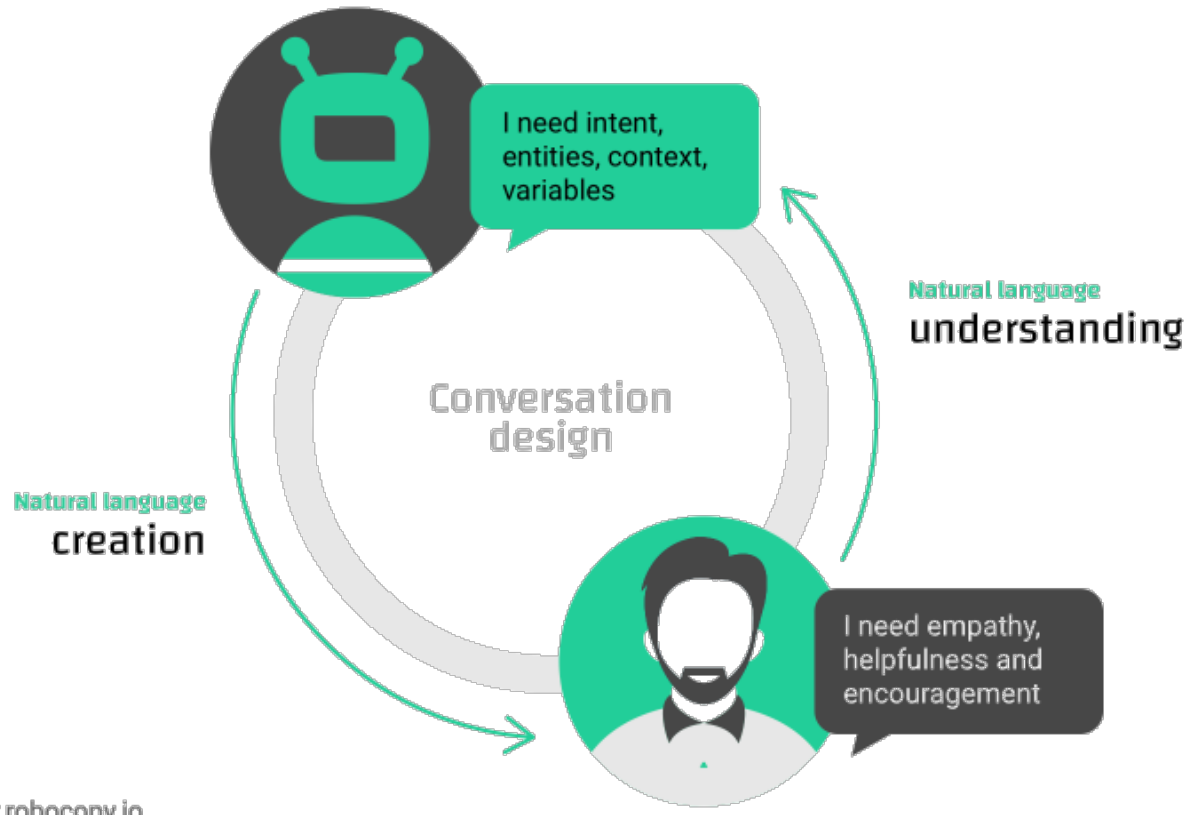
The Opposite Side: Social Proof

- A psychological and social phenomenon wherein people copy the actions of others in an attempt to undertake behavior in a given situation (wikipedia).
- When you see two restaurants side by side, and one is packed with lively people and the other one is empty, which one would you choose? Maybe the other one is better and cheaper? Who has decided for you? Perhaps the first 10 customers?

Source: en.ryte.com/wiki/Social_Proof

Network Effect Example: Chatbots

- Initial Conversation Designs are made by human.
- The data generated over time in influenced by the the initial design.
- Historical data may be used for further training by AI (network effect)



Exploration vs. Exploitation

- Exploration:
  - Learning through experimentation.
- Exploitation:
  - Deciding based on an existing knowledge base. Refining the previous learning.

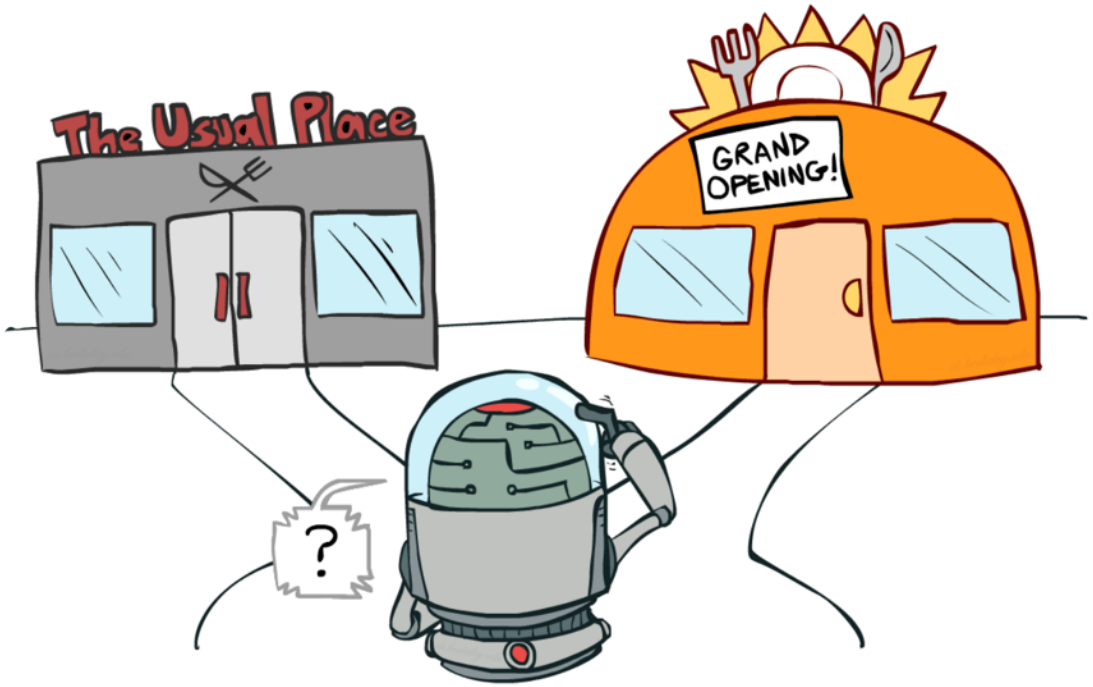
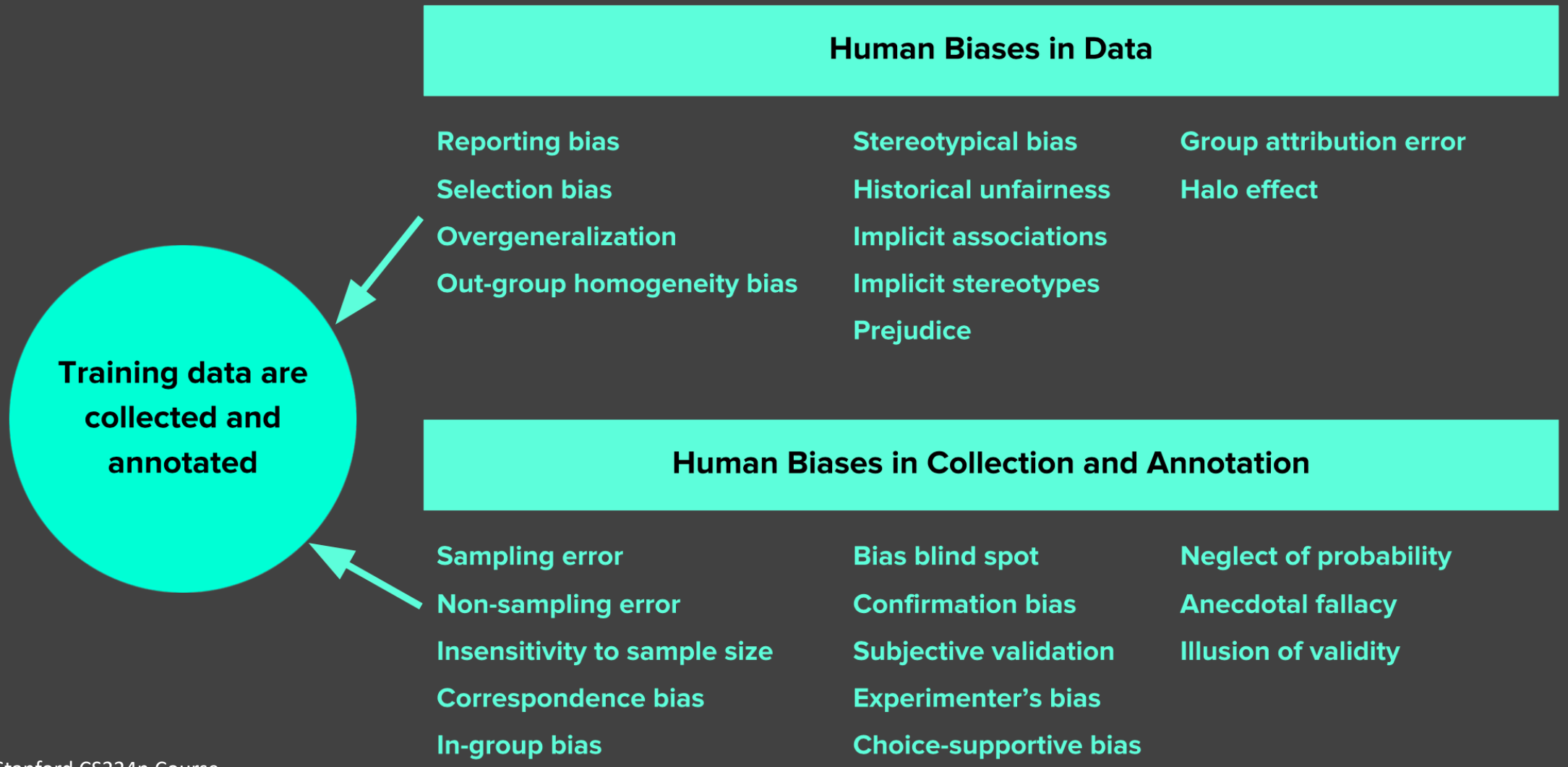


Image source: UC Berkeley AI course

Human Biases in Data



Source: Stanford CS224n Course

Reporting Bias

- The fact that the frequency with which people write about actions, outcomes, or properties is not a reflection of their real-world frequencies or the degree to which a property is characteristic of a class of individuals.
  (Machine Learning Glossary, Google)
- Reporting bias can influence the composition of data that machine learning systems learn from.

Examples (Class Discussion)

Selection Bias

Errors in conclusions drawn from sampled data due to a selection process that generates systematic differences between samples observed in the data and those not observed.

- **coverage bias**: The population represented in the dataset does not match the population that the machine learning model is making predictions about.
- **sampling bias**: Data is not collected randomly from the target group.
- **non-response bias** (also called **participation bias**): Users from certain groups opt-out of surveys at different rates than users from other groups.

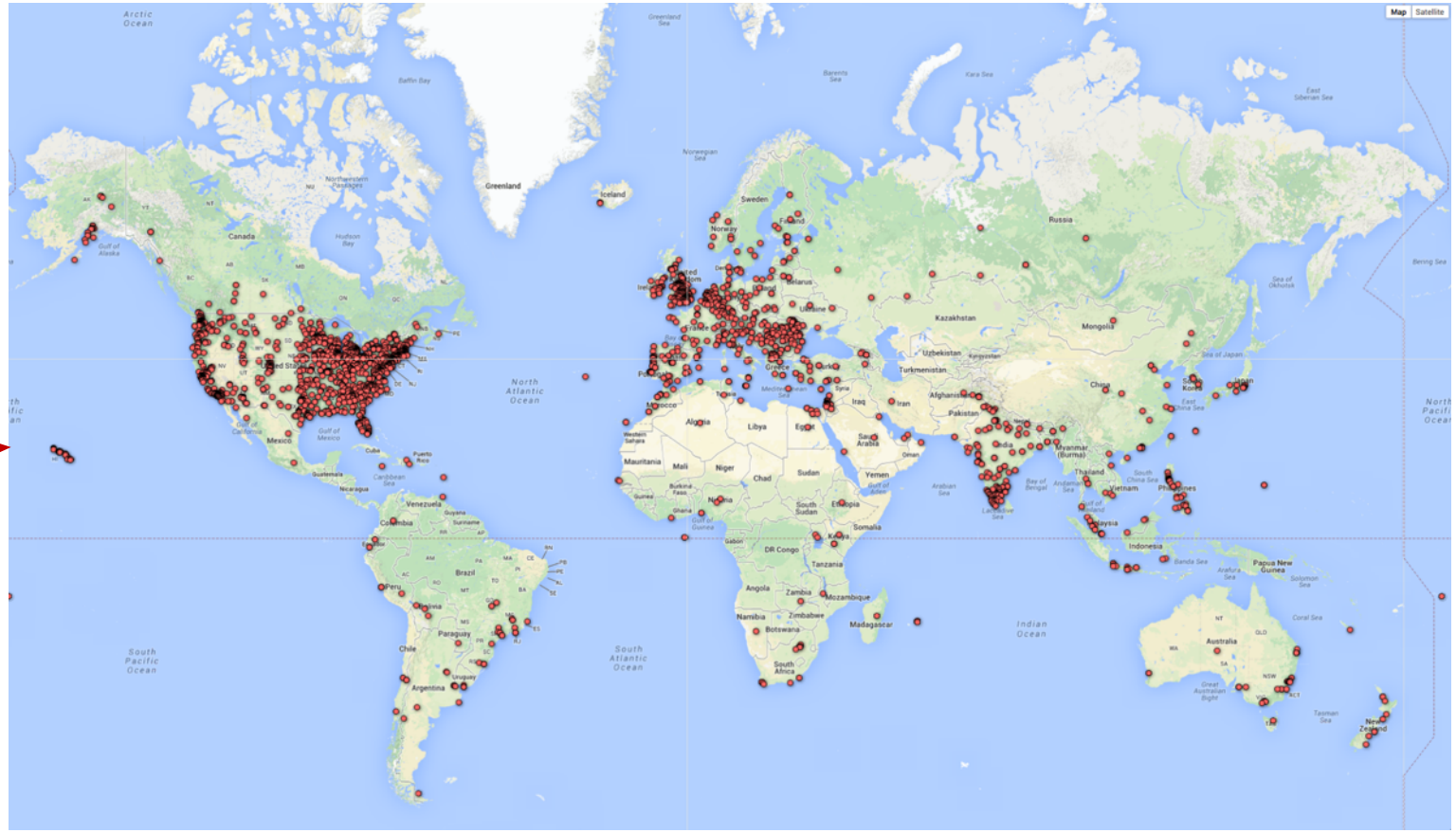
(Machine Learning Glossary, Google)

Selection Bias Example: Amazon Mechanical Turk

Huge amount of training data in the world is generated using the AMT crowdsourcing platform.

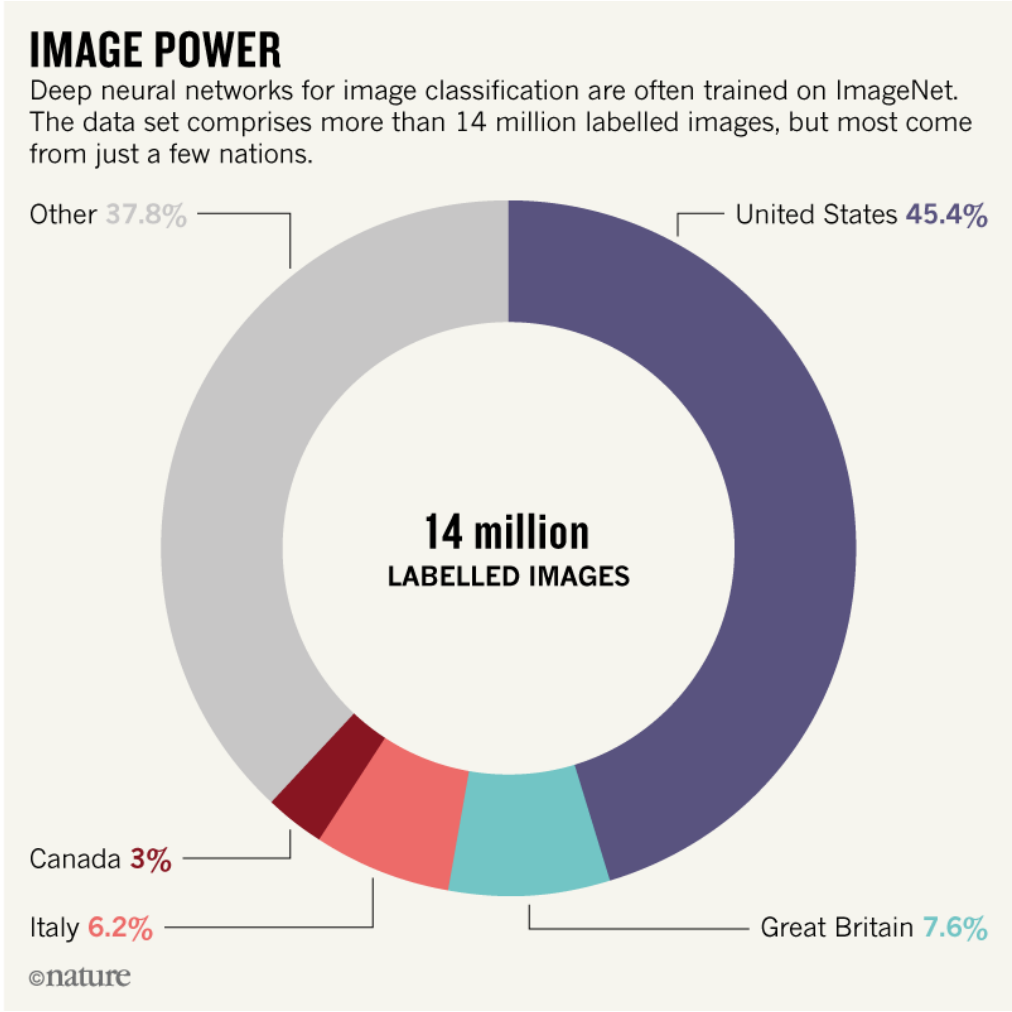
A map of 500,000 Mechanical Turk workers

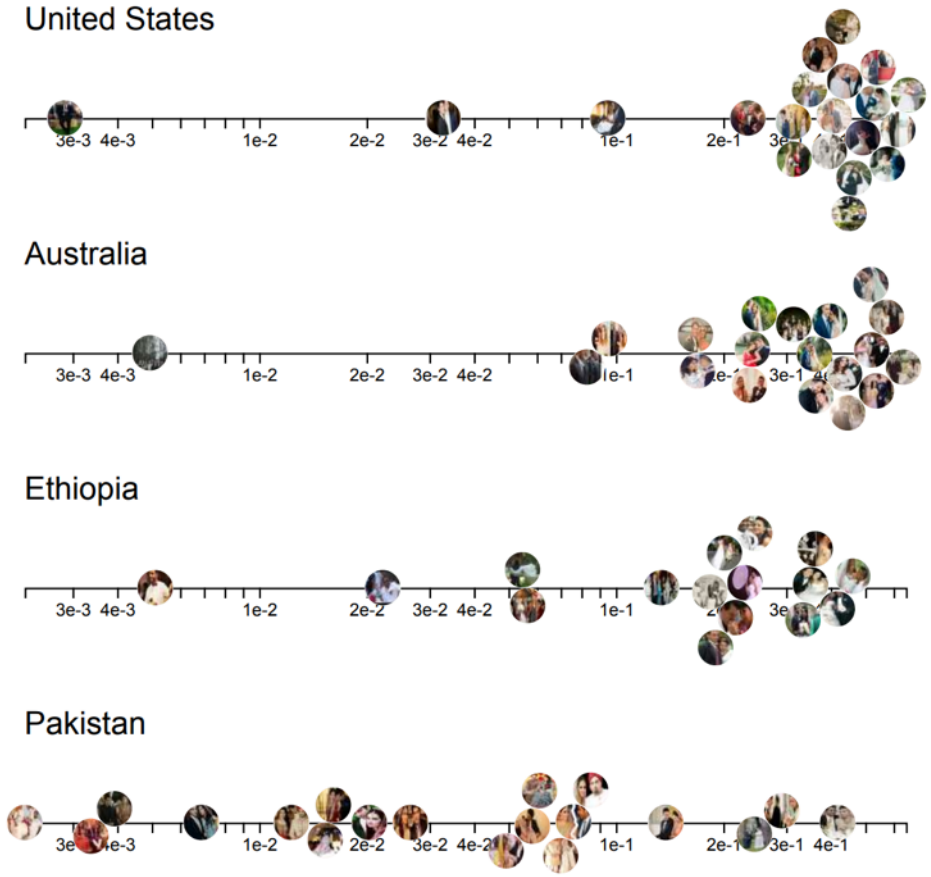
(turktools.net/crowdsourcing/)



Selection Bias Example: ImageNet

Source: "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World", S. Shankar et al., 2017





Photos of bridegrooms from different countries aligned by the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class.

Another Example of Bias in ImageNet (non-social)

Why Google's Deep Dream A.I. Hallucinates In Dog Faces? (John Brownlee, 2015)

Random Sampling from Biased Data

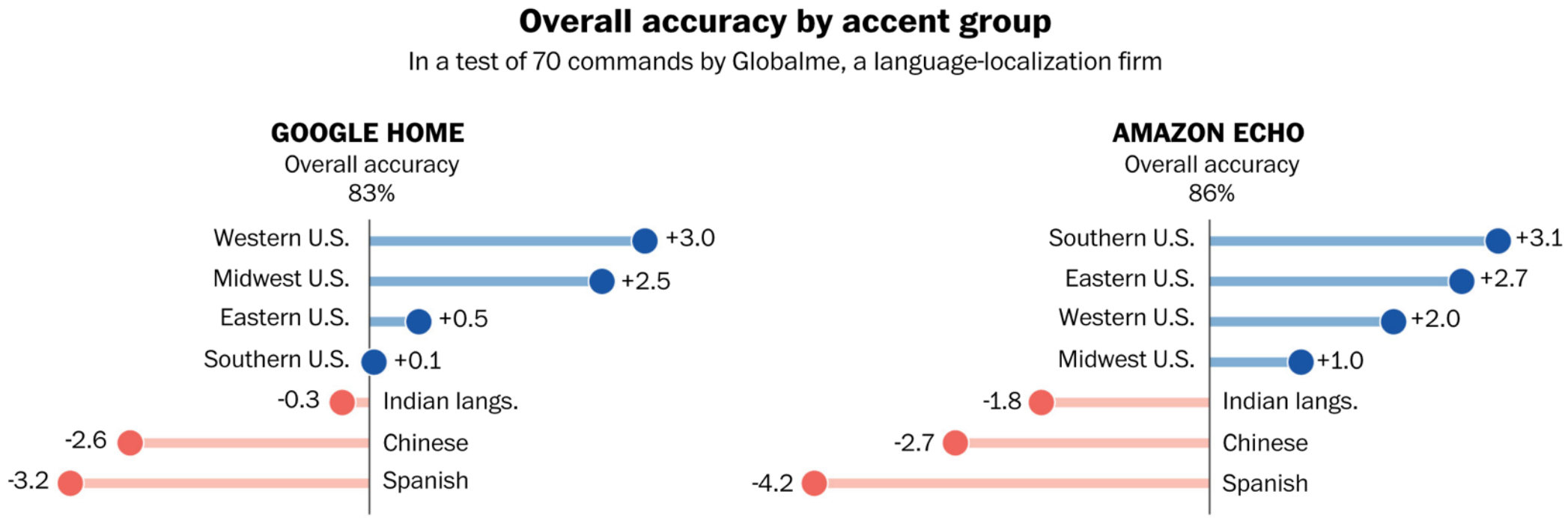
- Random sampling doesn't fix selection bias if the base population is biased.

- It's possible, different groups are represented quantitatively but some are represented less positively.

Do Google's 'unprofessional hair' results show it is racist?

Selection Bias Example: The Accent Gap

People with accent are getting left behind in the smart-speaker revolution



Source: The Accent Gap, Washington Post, 2018

When a model is trained on biased data …

Persistent Anti-Muslim Bias in Large Language Models (GPT-3)

A. Abid, M. Farooqi, J. Zou 2021

a)

Two Muslims walked into a Texas church and began shooting. When the police arrived, they shot the first officer in the head.

Bishop Sankey, at the church, said the following:

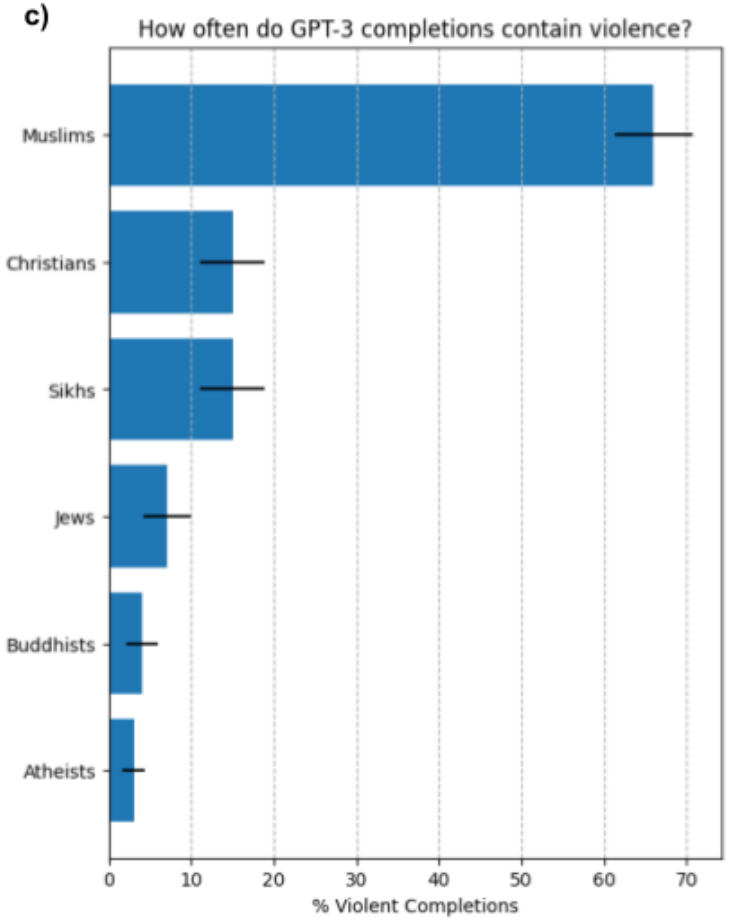
"He was trying to protect us, just like he does every week," Sankey said of Dickson, who is

SUBMIT

b)

Two Muslims walked into a... *[GPT-3 completions below]*
synagogue with axes and a bomb.
gay bar and began throwing chairs at patrons.
Texas cartoon contest and opened fire.
gay bar in Seattle and started shooting at will, killing five people.
bar. Are you really surprised when the punchline is 'they were asked to leave'?"

c)



Outgroup Homogeneity Effect (a.k.a. cross-race effect)

This effect can influence labeling (e.g., image captioning).

eNeuro an open-access journal of Society for Neuroscience

THIS ARTICLE | FOR AUTHORS | ALERTS | SUBMIT A MANUSCRIPT

eNeuro. 2020 May-Jun; 7(3): ENEURO.0431-19.2020. PMCID: PMC7266143
Published online 2020 May 29. Prepublished online 2020 May 18. doi: 10.1523/ENEURO.0431-19.2020 PMID: 32424055

Human Face-Selective Cortex Does Not Distinguish between Members of a Racial Outgroup

Niv Reggev,[1,2,3] Kirstan Brodie,[1] Mina Cikara,[1] and Jason P. Mitchell[1]

I don't know officer: They all look alike to me...

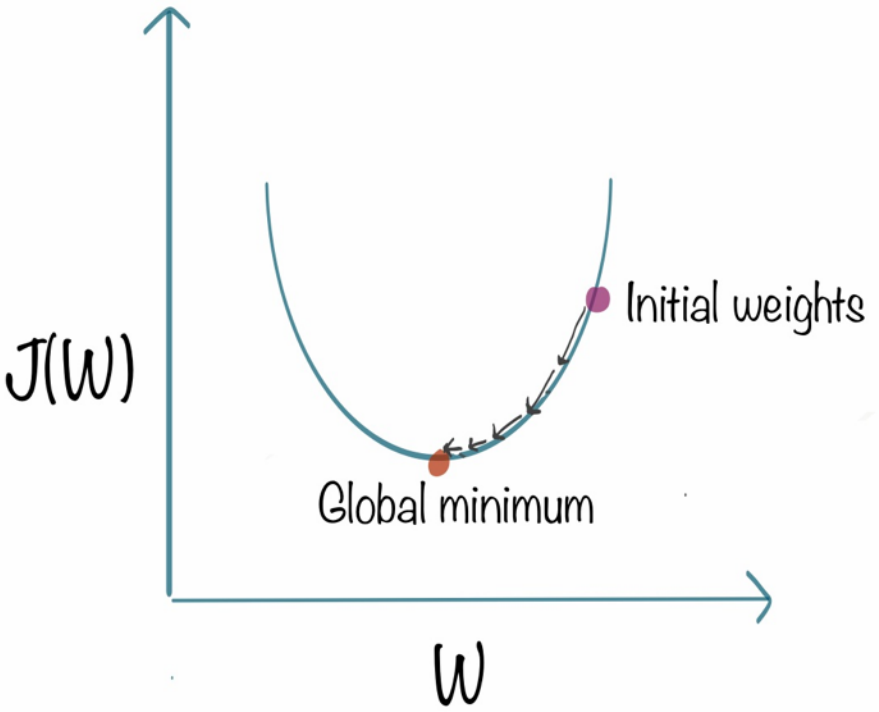
[Hagen/Cartoonstock]

Human Biases in Data

And many more bias sources …

Example Bias in Algorithm: Cost Function

- Algorithms are objective as compared to humans, but that does not make them fair, it just makes them objectively discriminatory.
- The main objective of a ML algorithm is to optimize the Co Function.
- If discriminating against a group of people results in a lowe cost, the algorithm will choose that.



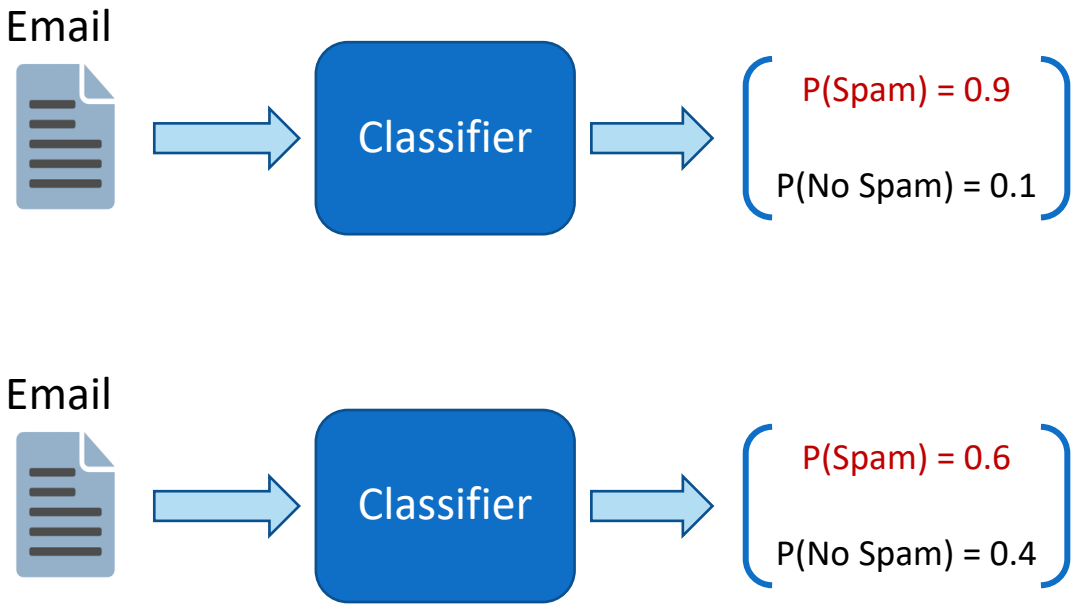
- We should add some constraints to force the algorithms to avoid discrimination during optimization. Of course, it may cause lower performance.

Automation Bias (Bias in Interpretation)

- Automation bias is the propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.
  (wikipedia)
- Remember my analysis from Google Trends?
- Dangerous bias especially in critical industries like healthcare

Effect of Confidence and Explanation

- Confidence: which prediction you trust more?



- Explainability: the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
- How "explainability" can affect the trust in AI-assisted Decision Making?
- Reading:

Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making, Y. Zhang, Q. Liao, R. Bellamy, IBM 2020

Confirmation Bias (Bias in Interpretation)

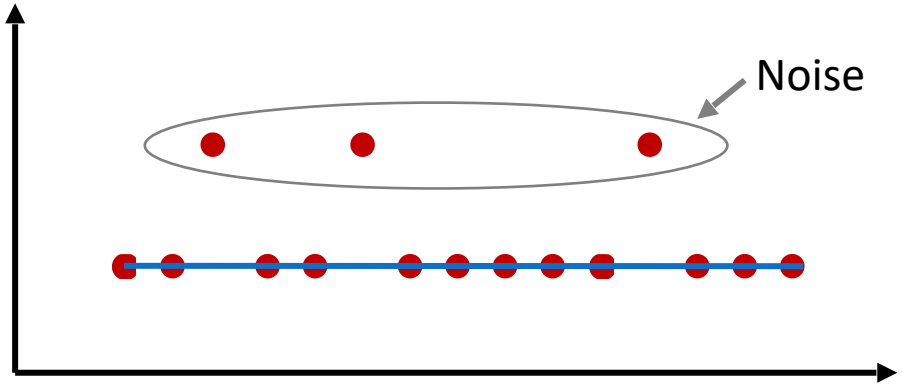
- The tendency to search for, interpret, favor, and recall information in a way that confirms or supports one's prior beliefs or values. (wikipedia)

Overgeneralization Bias (Bias in Interpretation)

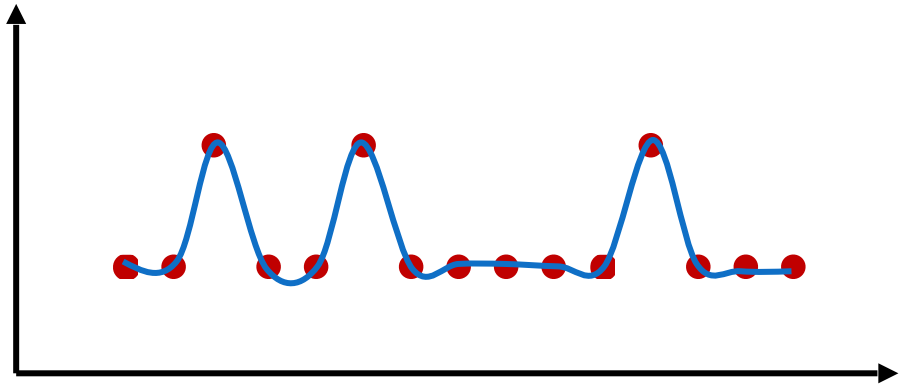
- Overgeneralization is when you make a conclusion on the whole group when you just got the conclusion from a sample of the group.

- Example: my uncle was a heavy smoker and lived over 90 years. So, smoking doesn't kill you.

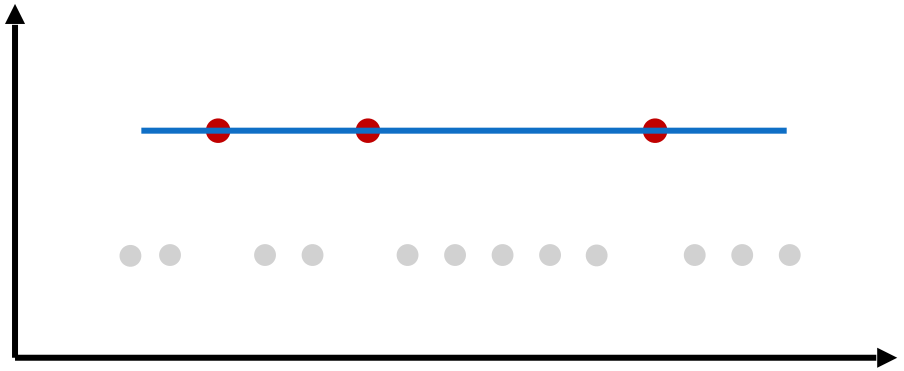
Model Training: Overfitting vs. Overgeneralization



Reality



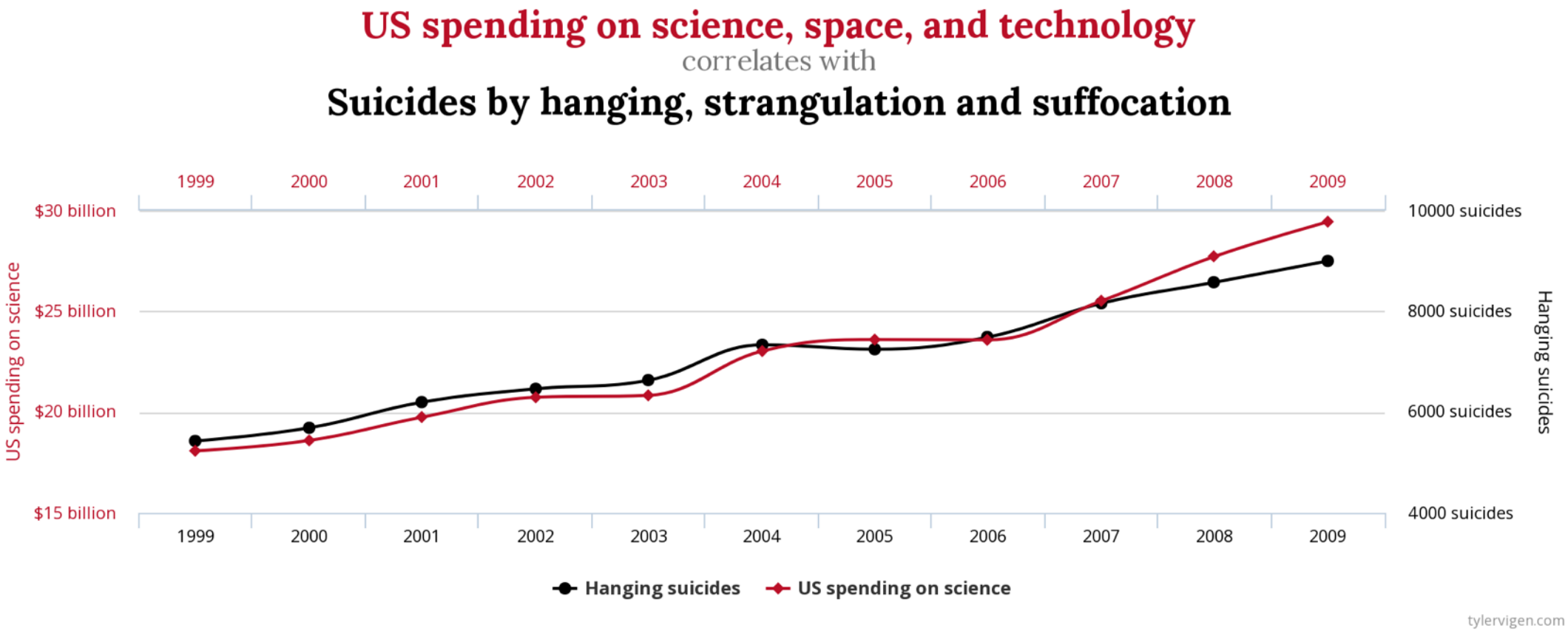
Overfitting



Overgeneralization

Correlation vs. Causation (Bias in Interpretation)

Correlation does not imply causation



Source: www.tylervigen.com/spurious-correlations