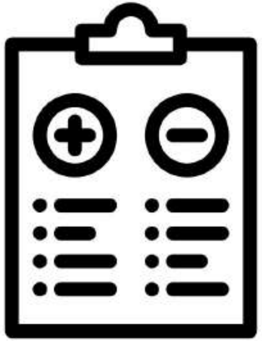
Copyright Notice

deeplearning.ai

Disadvantages of GANs

Outline

- Advantages of GANs
- Disadvantages of GANs



Advantages of GANs

- Amazing empirical results - especially with fidelity

Advantages of GANs

- Amazing empirical results - especially with fidelity
- Fast inference (image generation during testing)

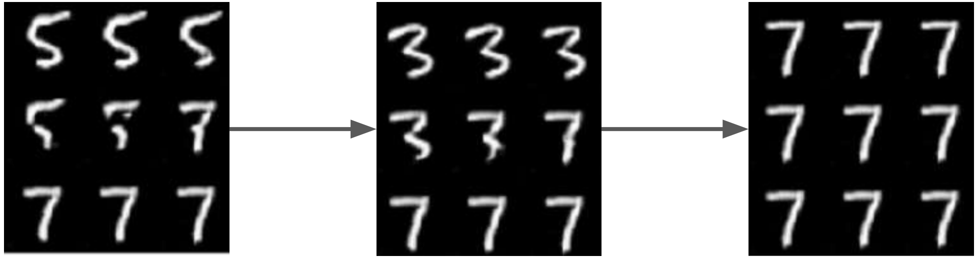
Disadvantages of GANs

- Lack of intrinsic evaluation metrics

Looks pretty
real..?

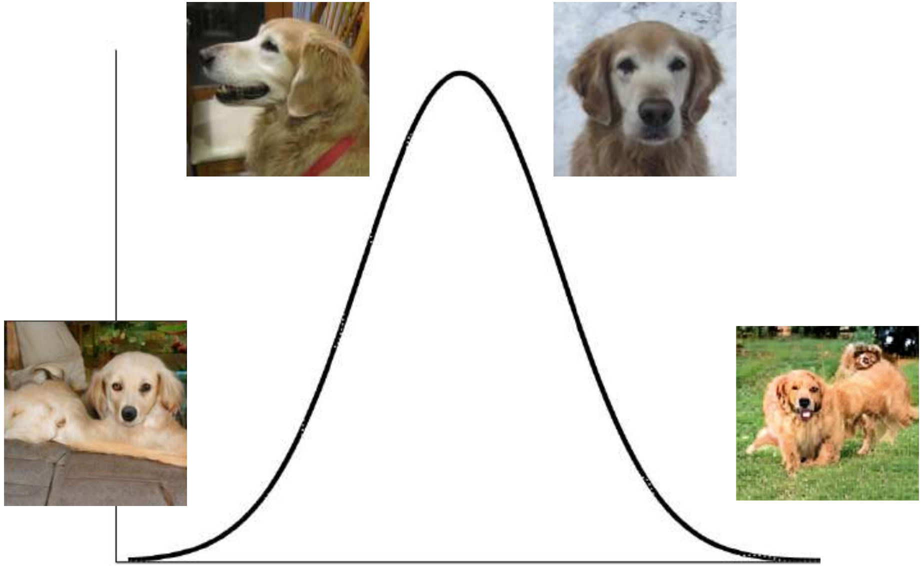
Disadvantages of GANs

- Lack of intrinsic evaluation metrics
- Unstable training



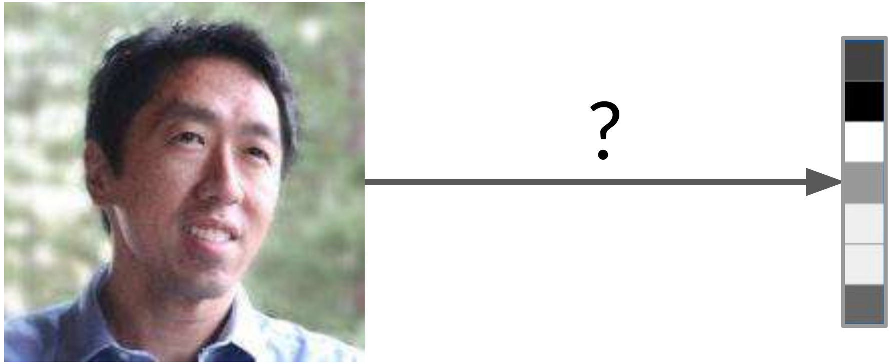
Disadvantages of GANs

- Lack of intrinsic evaluation metrics
- Unstable training
- No density estimation



Disadvantages of GANs

- Lack of intrinsic evaluation metrics
- Unstable training
- No density estimation
- Inverting is not straightforward



Summary

Advantages

- Amazing empirical results
- Fast inference

Disadvantages

- Lack of intrinsic evaluation metrics
- Unstable training
- No density estimation
- Inverting is not straightforward

Summary

Advantages

- Amazing empirical results
- Fast inference

GANs have **amazing results**, but shortcomings as well.

Disadvantages

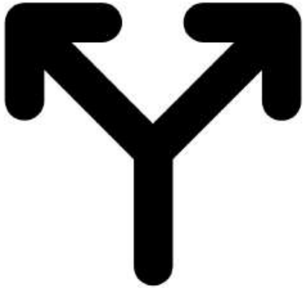
- Lack of intrinsic evaluation metrics
- Unstable training
- No density estimation
- Inverting is not straightforward

deeplearning.ai

Alternatives to GANs

Outline

- Overview of generative models
- VAEs and other alternatives



Generative Models

Noise Class Features

$$\xi, Y \rightarrow X$$

$$P(X|Y)$$

Generative Models

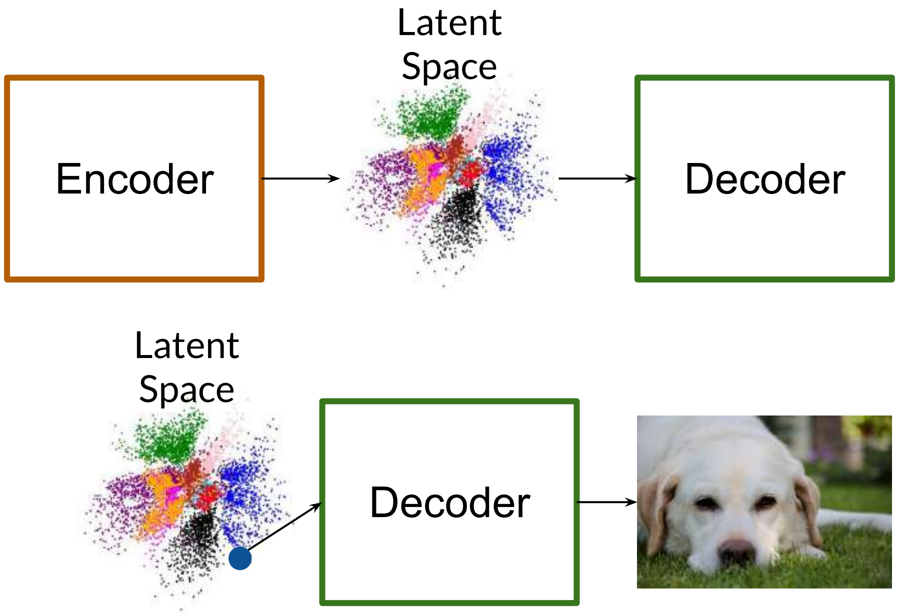
There are other generative models than just GANs!

Noise Class Features

$$\xi, Y \rightarrow X$$

$$P(X|Y)$$

Variational Autoencoders (VAEs)



Available from: https://arxiv.org/abs/1804.00891

Variational Autoencoders (VAEs)

Advantages

- Has density estimation
- Invertible
- Stable training

Disadvantages

- Lower quality results

Variational Autoencoders (VAEs)

VQ-VAE (Proposed) BigGAN deep

Available from: https://arxiv.org/abs/1906.00446

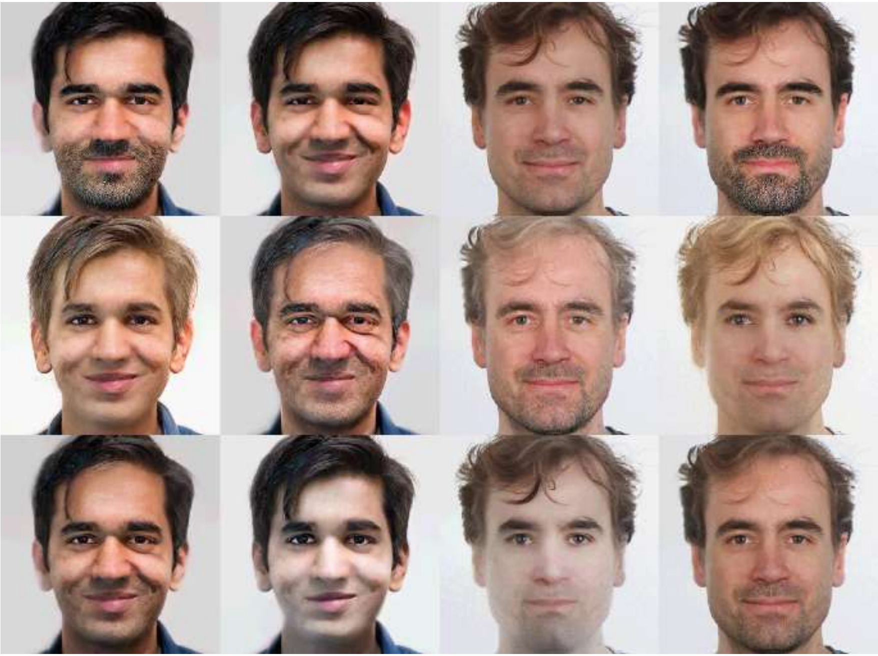
Autoregressive Models

Left: source image Right: new portraits generated from high-level latent representation

Relies on previous pixels to generate next pixel

Available from: https://arxiv.org/abs/1606.05328

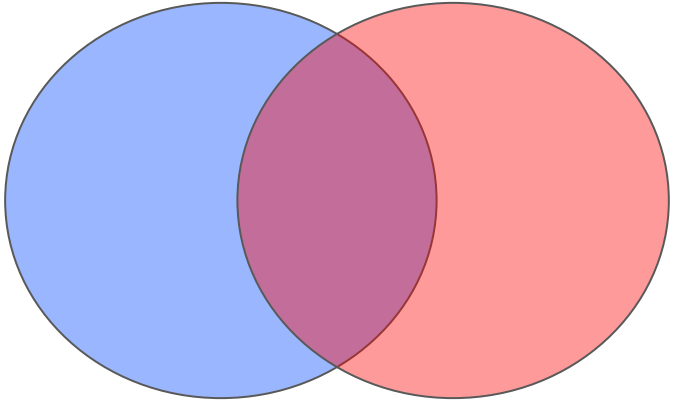
Flow Models



Uses invertible mappings

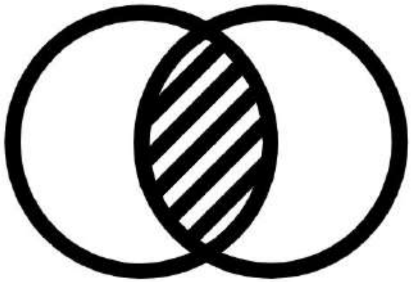
Available from: https://openai.com/blog/glow/

Hybrid Models



Summary

- VAEs have the opposite pros/cons as GANs
  - Often lower fidelity results
  - Density estimation, inversion, stable training
- Other alternative generative models:
  - Autoregressive models
  - Flow models
  - Hybrid models



deeplearning.ai

Intro to Machine Bias

Outline

- *Machine Bias* (ProPublica)
- Racial disparity in AI for risk assessments
- Impacts of biased AI

Machine Bias

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine Bias

Risk assessment = likelihood of committing a crime in the future

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS Algorithm

- One of two leading commercial tools used by the legal system

COMPAS Algorithm

- One of two leading commercial tools used by the legal system
- Used in pretrial hearings and criminal sentencing to assess risk of re-offense (recidivism)

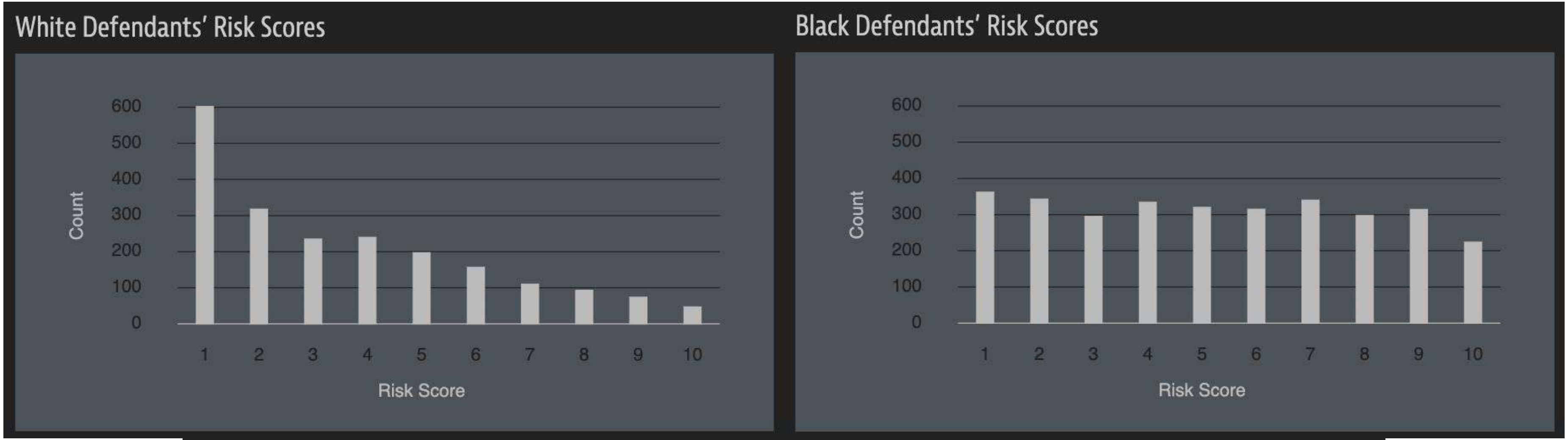
COMPAS Algorithm

- One of two leading commercial tools used by the legal system
- Used in pretrial hearings and criminal sentencing to assess risk of re-offense (recidivism)
- Score based on proprietary calculations
  - Not available to the public
  - Unvalidated

COMPAS Algorithm

- One of two leading commercial tools used by the legal system
- Used in pretrial hearings and criminal sentencing to assess risk of re-offense (recidivism)
- Score based on proprietary calculations
  - Not available to the public
  - Unvalidated
- Predicts recurrence of violent crime correctly only 20% of the time

Biased Risk Assessment



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Biased Risk Assessment

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Consequences of a Higher Score

Paul Zilly

Plea deal overturned and sentenced to two years in state prison.

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Consequences of a Higher Score

Paul Zilly

Plea deal overturned and sentenced to two years in state prison.

"Had I not had the COMPAS, I believe it would likely be that ***I would have given one year, six months***"

- Appeals judge

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Consequences of a Higher Score

Sade Jones

Bond was raised from the recommended $0 to $1000

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Consequences of a Higher Score

Sade Jones

Bond was raised from the recommended $0 to $1000

"I went to McDonald's and a dollar store, and they all said no ***because of my background***"

- Jones

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Prediction Failure

Prediction Fails Differently for Black Defendants

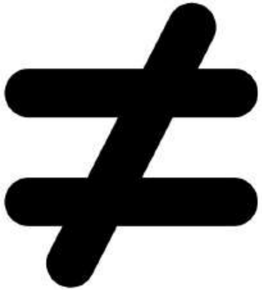
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Summary

- Machine learning bias has a disproportionately negative effect on historically underserved populations
- Proprietary risk assessment software:
  - Difficult to validate
  - Misses important considerations about people

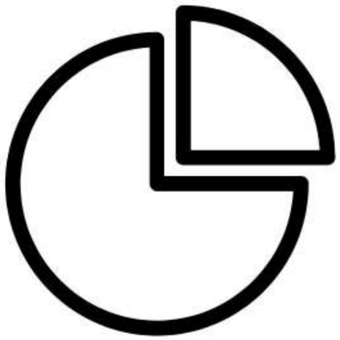


deeplearning.ai

Defining Fairness

Outline

- What is fairness?
- Complexity of defining fairness



Fairness in Machine Learning

Reading 1: [Fairness Definitions Explained]

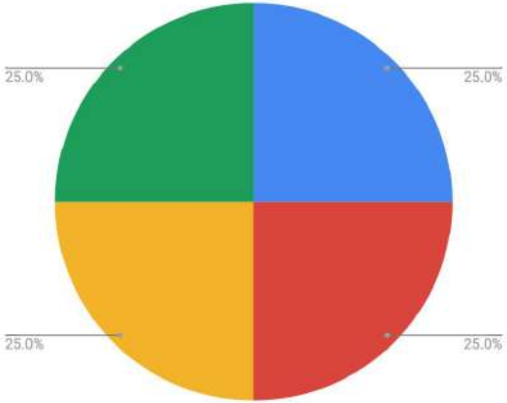
Reading 2: [A Survey on Bias and Fairness in Machine Learning]

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

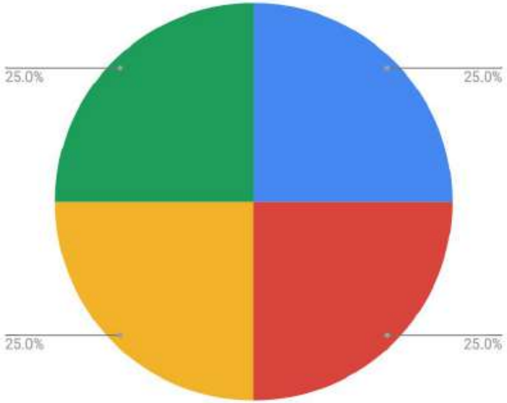
Table 1: Considered Definitions of Fairness

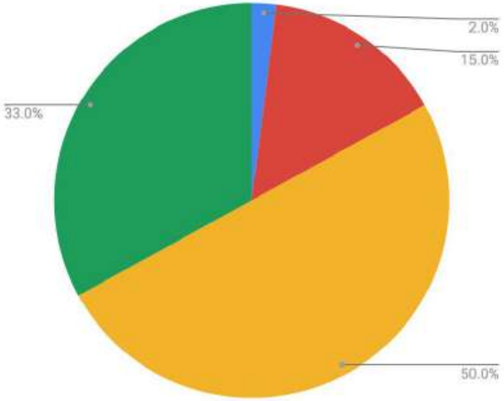
Available from: https://fairware.cs.umass.edu/papers/Verma.pdf

Defining Fairness

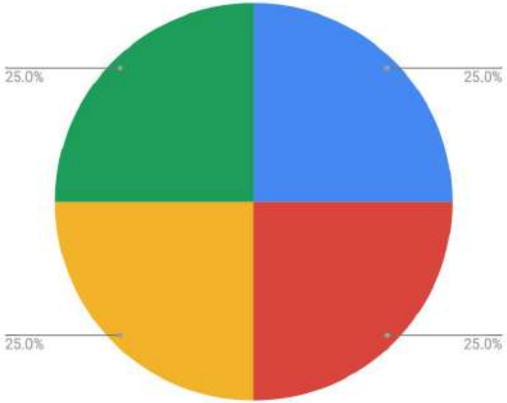


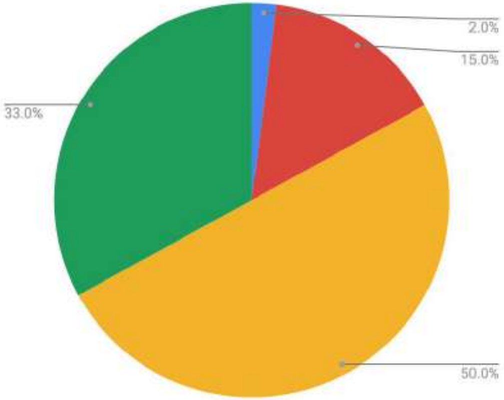
Defining Fairness





Defining Fairness



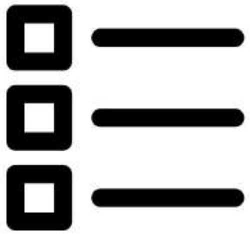


	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Available from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Summary

- Fairness is difficult to define
- There is no single definition of fairness
- Important to explore these before releasing a system into production

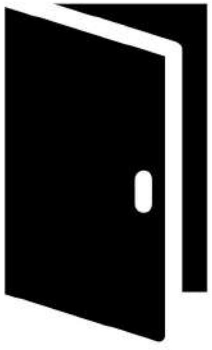


Ways Bias is Introduced

deeplearning.ai

Outline

- A few ways bias can enter a model
- PULSE: A case study with a biased GAN



Training Bias

Training data

- **No variation** in who or what is represented

Training Bias

Training data

- **No variation** in who or what is represented
- Bias in **collection methods**

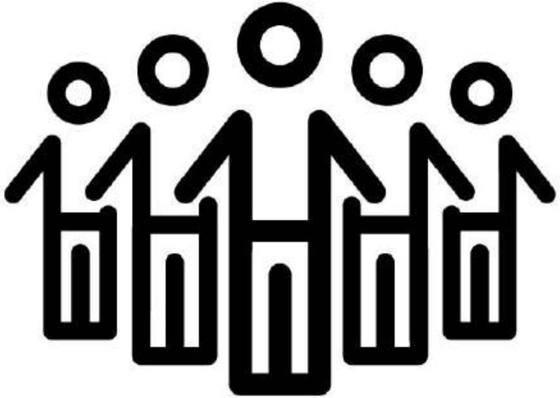
Training Bias

Training data

- **No variation** in who or what is represented
- Bias in **collection methods**

Data labelling

- **Diversity** of the labellers

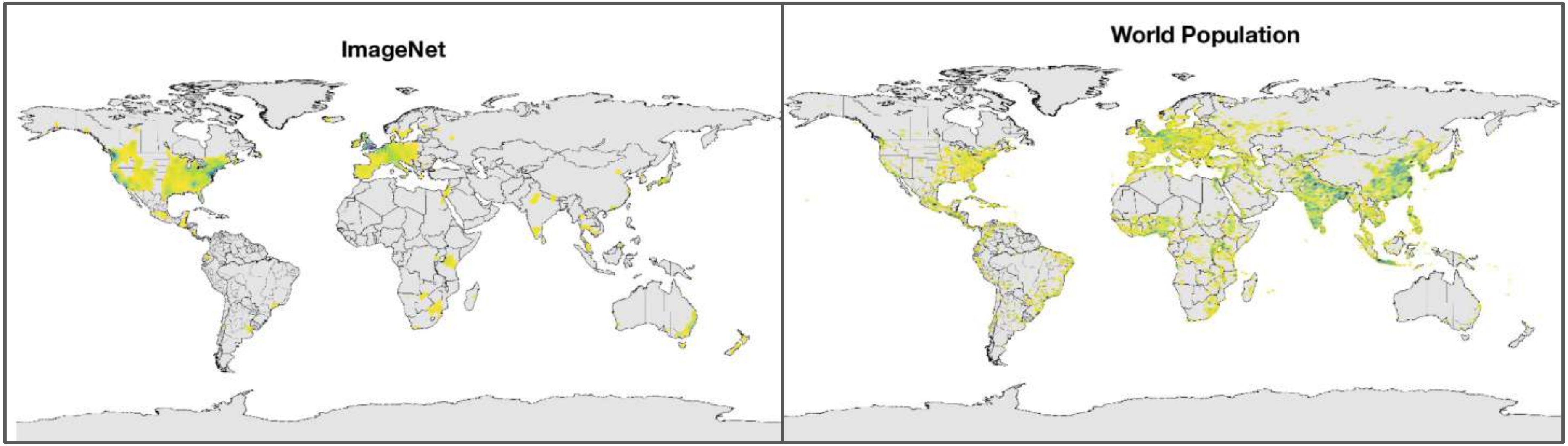


Evaluation Bias

- Images can be biased to reflect “correctness” in the dominant culture

Evaluation Bias

- Images can be biased to reflect “correctness” in the dominant culture



Available from: https://arxiv.org/abs/1906.02659

Evaluation Bias

- Images can be biased to reflect “correctness” in the dominant culture

Available from: https://arxiv.org/abs/1906.02659

Evaluation Bias

- Images can be biased to reflect “correctness” in the dominant culture

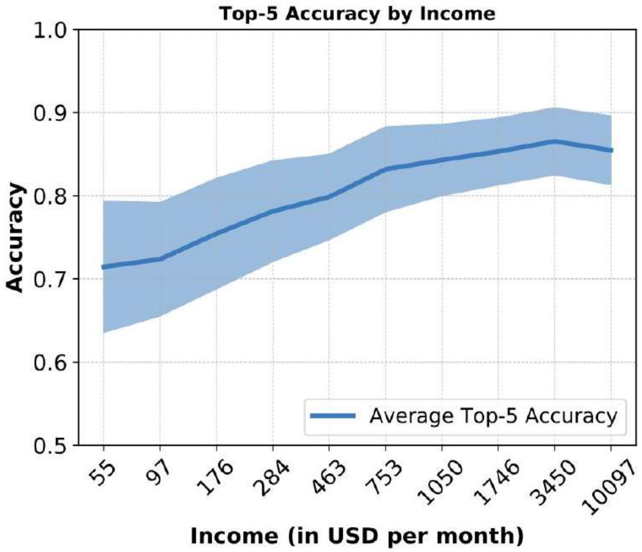
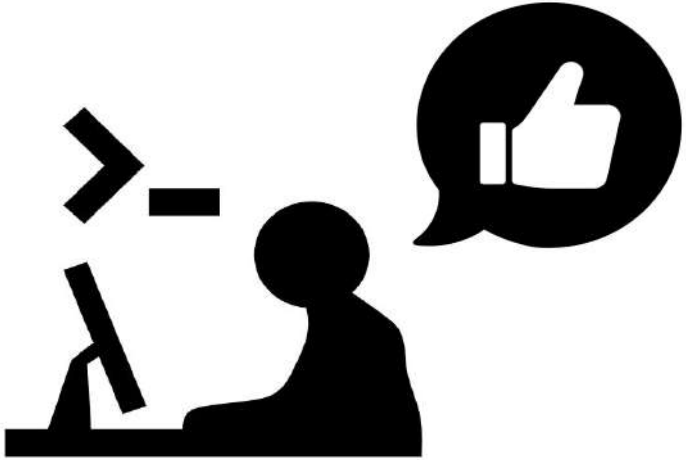


Figure 3: Average accuracy (and standard deviation) of six object-recognition systems as a function of the normalized consumption income of the household in which the image was collected (in US$ per month).

Available from: https://arxiv.org/abs/1906.02659

Model Architecture Bias

- Can be influenced by the coders who designed the architecture or optimized the code

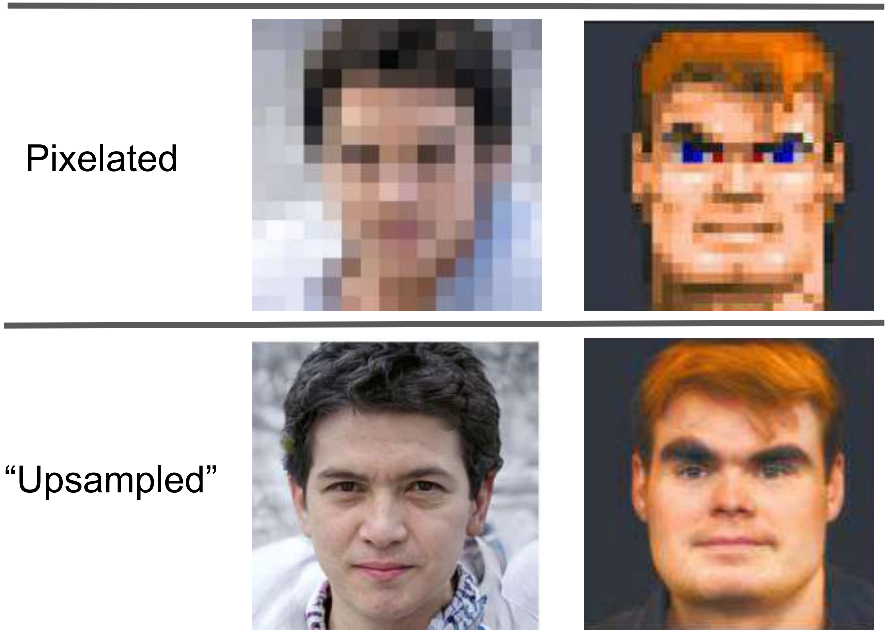


Other Avenues for Bias Introduction

Bias can appear at any step:

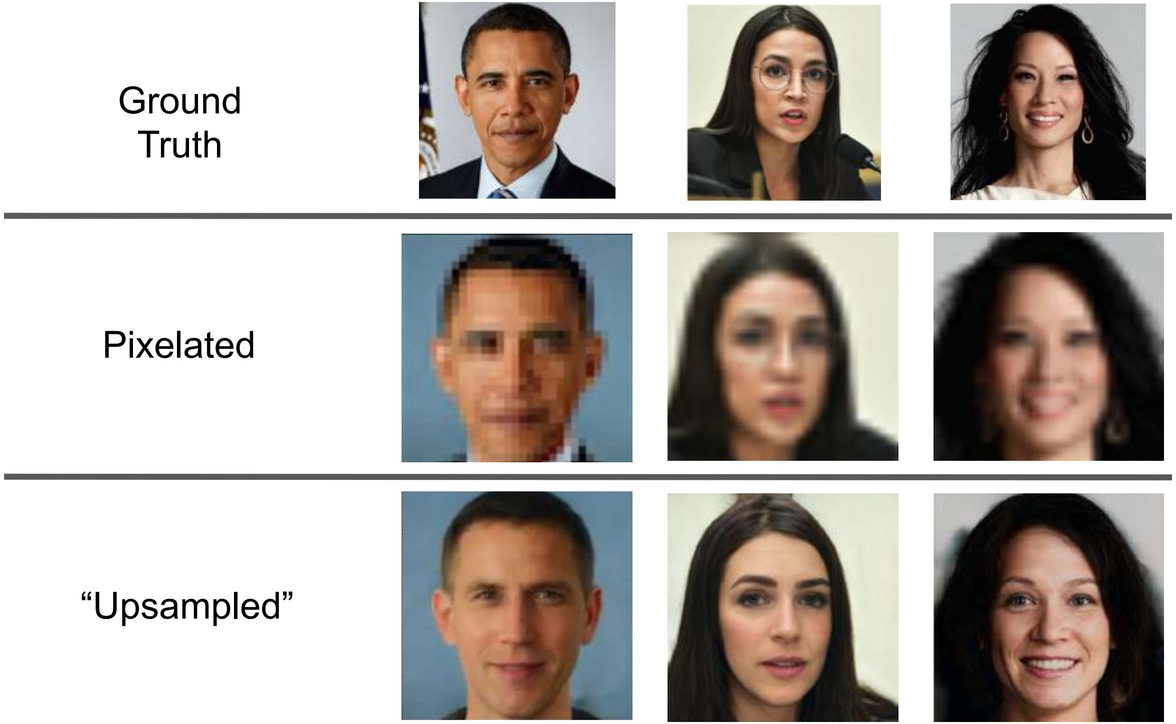
- Research
- Design
- Engineering
- Anywhere a person was involved

PULSE



(Left) Available from: https://arxiv.org/abs/2003.03808
(Right) Available from: https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

PULSE



Available from: https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

Summary

- Bias can be introduced into a model at each step of the process
- Awareness and mitigation of bias is vital to responsible use of AI and, especially, state-of-the-art GANs