22 - Bias and Fairness

Ludwig Krippahl

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Summary

- Bias and its ethical problems.
- Sources of bias
  - Sampling, population, assumptions
- Mitigating undesirable biases

Bias

Inductive Bias

- In machine learning, inductive bias is the set of assumptions that constrain hipotheses and allow generalization
- All learning systems need bias for generalization, including ourselves.

Bias

- In general, bias is any correlation found in data. Is what we learn

Bias

Bias

- In general, bias is any correlation found in data. Is what we learn

Bias

- In general, bias is any correlation found in data. Is what we learn
  - Stereotypes, assumptions, prototypes
- But bias is not always desirable
- E.g. NIST review of face recognition products
  - False match for American Indian women 68 times higher for American Indian women than for white men
  - Also 47 times higher for American Indian men and 10 times higher for black women
- Good bias: regularities that we can learn
- Bad bias: correlations that lead to unfair results

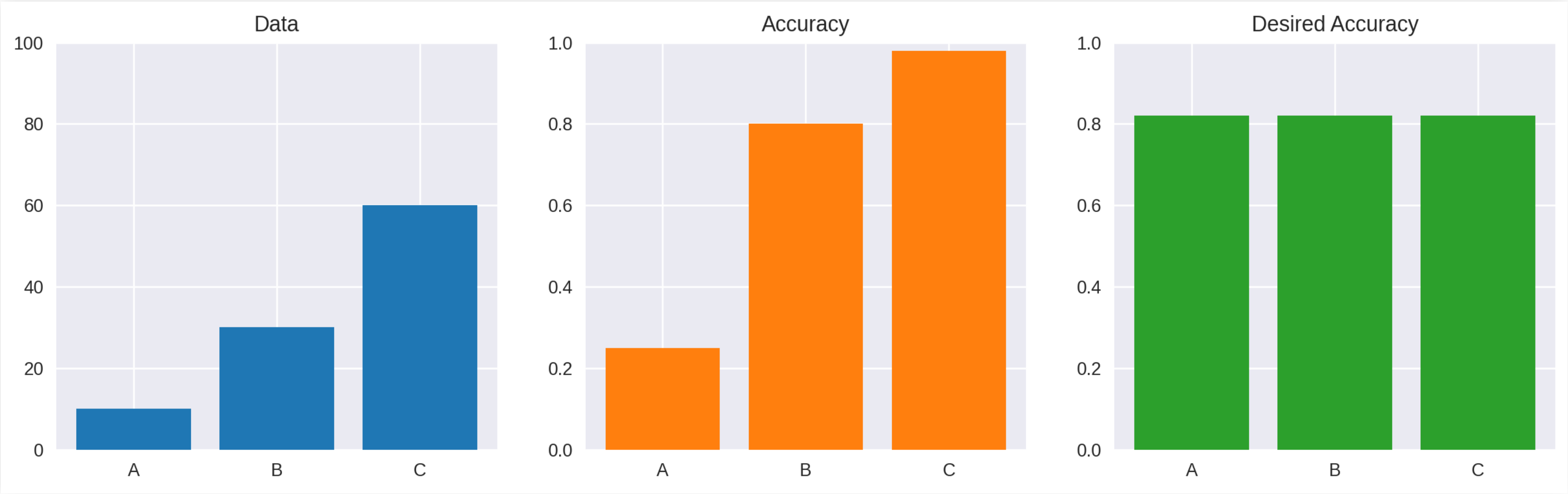
Fairness

Intuituion

- A system is fair if its results do not depend on certain features
  - E.g. sex, race, religious beliefs, political ideology, sexual orientation.
  - We can ommit such features from structured data
  - But with unstructured data this is harder
- And also if it works equally well on all groups
  - E.g. Classify photos to identify CEO of important companies
  - We get 95% accuracy in our test set.
  - But only about 5% of the CEO of large companies are women.
  - The classifier may be discarding all women as negative examples
  - (and have 0% accuracy on women CEO)

Intuituion

- A system is fair if its results do not depend on certain features
- And also if it works equally well on all groups
- Class imbalance can be a problem:



Fairness

Intuituion

- A system is fair if its results do not depend on certain features
- And also if it works equally well on all groups

- This is important if we are developing models that impact people

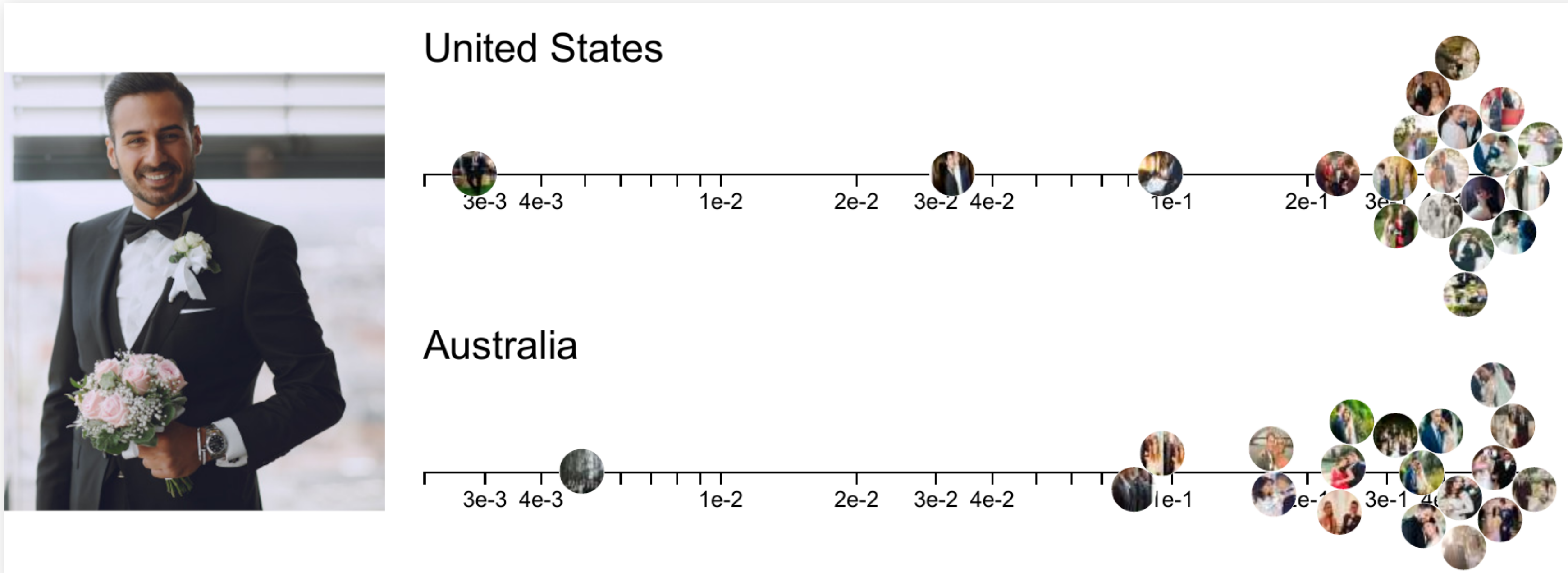
Sources of Bias

Sources of Bias

Bias can have several sources

- In data:

- Inadequate sampling. E.g. ImageNet biased for western countries



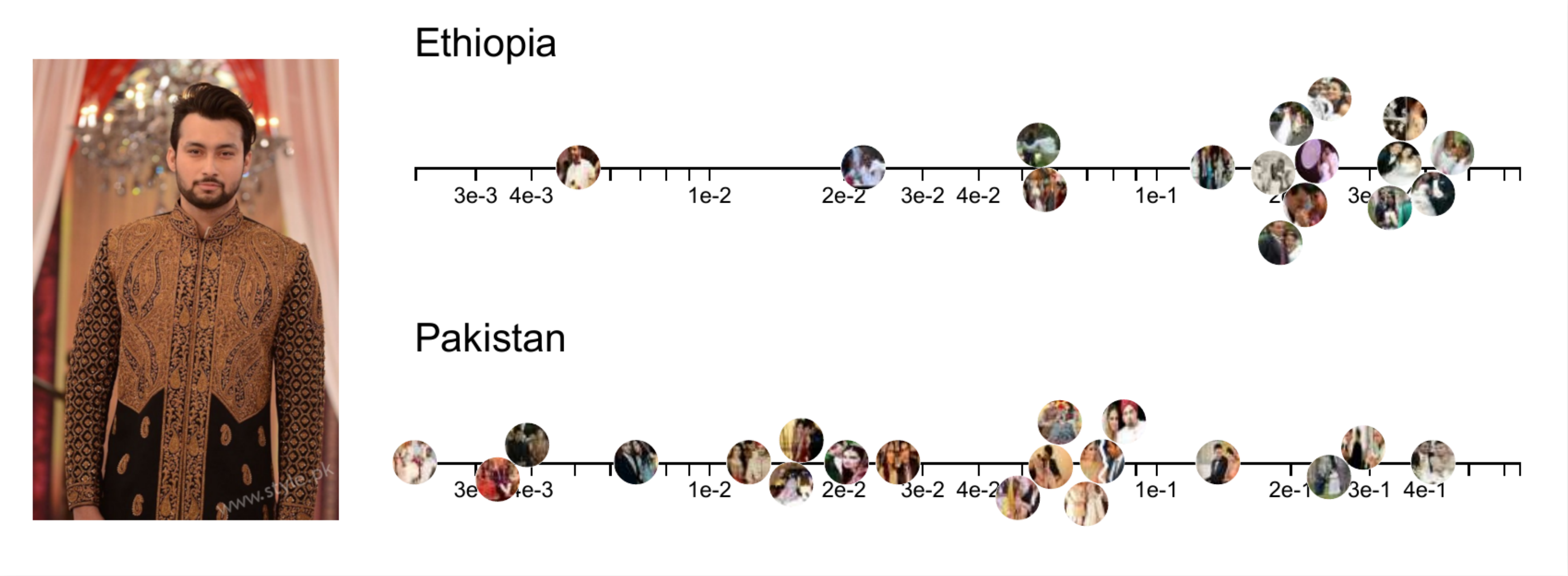
Shankar et al, No Classification without Representation, 2017

Sources of Bias

Bias can have several sources

- In data:

• Inadequate sampling. E.g. ImageNet biased for western countries



Shankar et al, No Classification without Representation, 2017

Sources of Bias

Bias can have several sources

- In data:
  - Inadequate sampling
  - The universe is biased
  - Gender imbalances in professions like nursing, construction, engineering or sociology

Sources of Bias

Bias can have several sources

- In data:
  - Inadequate sampling
  - The universe is biased
- Due to feature selection:
  - Nearly all violent criminals are men
  - We will not use sex as a feature for prediction (protected characteristic)
  - But height and weight are strongly correlated with sex

Sources of Bias

Bias can have several sources

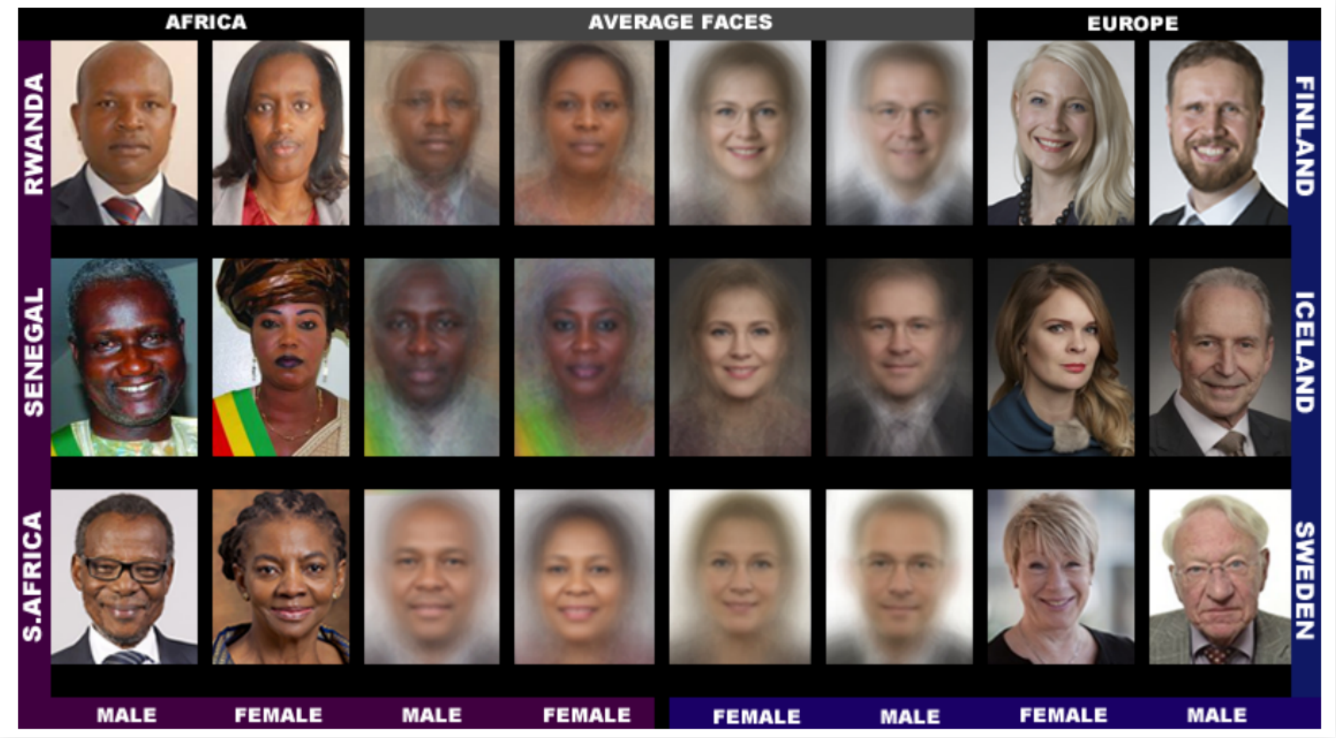
- In data:
  - Inadequate sampling
  - The universe is biased
- Due to feature selection
- Aggregation effects
  - Haemoglobin A1c is an indicator for high blood glucose levels and diabetes
  - However, levels differ between ethnic groups, so using an average value will not work in some groups

Mitigating undesirable bias

Mitigating Bias

Sampling Bias

- If the problem is sampling, the best solution is better sampling
- Pilot Parliaments Benchmark



Shankar et al, No classification without representation, 2017

Mitigating Bias

Bias in population (or cannot get better data)

- Resample the data we have
- Representation Bias Removal (REPAIR)
- We have a DNN classifier

- All layers up to last extract features
- The last layer is a linear classifier with softmax output
- We can retrain this last classifier with cross-entropy loss:
- For the dataset D and parameteres θ:

$$L(D,\theta) = \mathbb{E}\left(-\log P(Y \mid X)\right) = -\frac{1}{|D|}\sum_{(x,y)\in D} \log P(y \mid x)$$

Mitigating Bias

Representation Bias Removal (REPAIR)

- Bias is the reduction in uncertainty, normalized

$$B(D,\theta) = 1 - \frac{L(D,\theta)}{H(Y)} \qquad H(Y) = -\frac{1}{|D|}\sum_{(x,y)\in D} \log p_y$$

- If loss is low bias is high (that is what we larn)
- Goal: adjust sampling probability to make classification harder

$$L(D',\theta) = -\frac{1}{\sum_{i=1}^{|D|} w_i}\sum_{i=1}^{|D|} w_i \log P(y_i \mid x_i)$$

$$B(D',\theta) = 1 - \frac{L(D',\theta)}{H(Y')} \qquad H(Y') = -\frac{1}{\sum_{i=1}^{|D|} w_i}\sum_{i=1}^{|D|} w_i \log \frac{\sum_{i:y_i=y} w_i}{\sum_i w_i}$$

Representation Bias Removal (REPAIR)

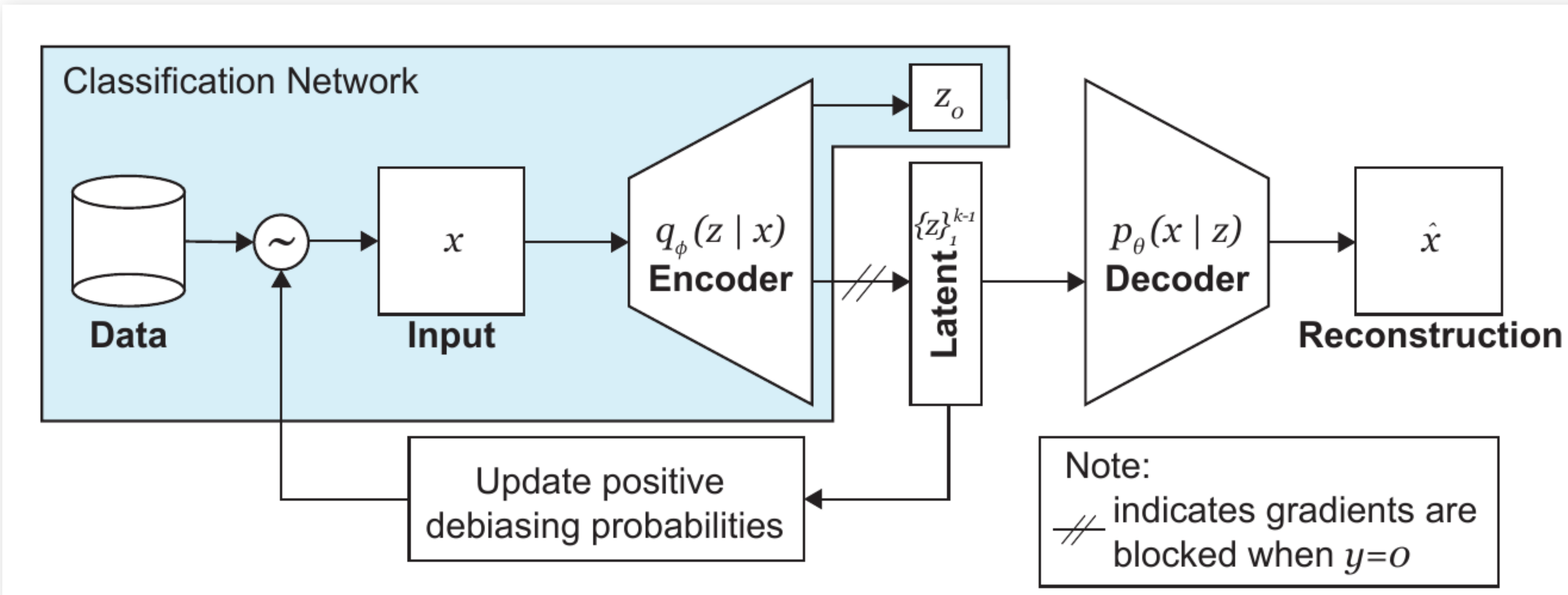
- Bias is the reduction in uncertainty, normalized

$$B(D,\theta) = 1 - \frac{L(D,\theta)}{H(Y)} \qquad H(Y) = -\frac{1}{|D|}\sum_{(x,y)\in D} \log p_y$$

- If loss is low bias is high (that is what we larn)
- Goal: adjust sampling probability to make classification harder
- Minimize $L(D',\theta)$ with respect to θ
- Minimize $B(D',\theta)$ with respect to W
- (This is done with adversarial training)
- Intuition: eliminate imbalances that facilitate classification

Debiasing Variational Autoencoder

- Use a variational autoencoder to learn a latent representation

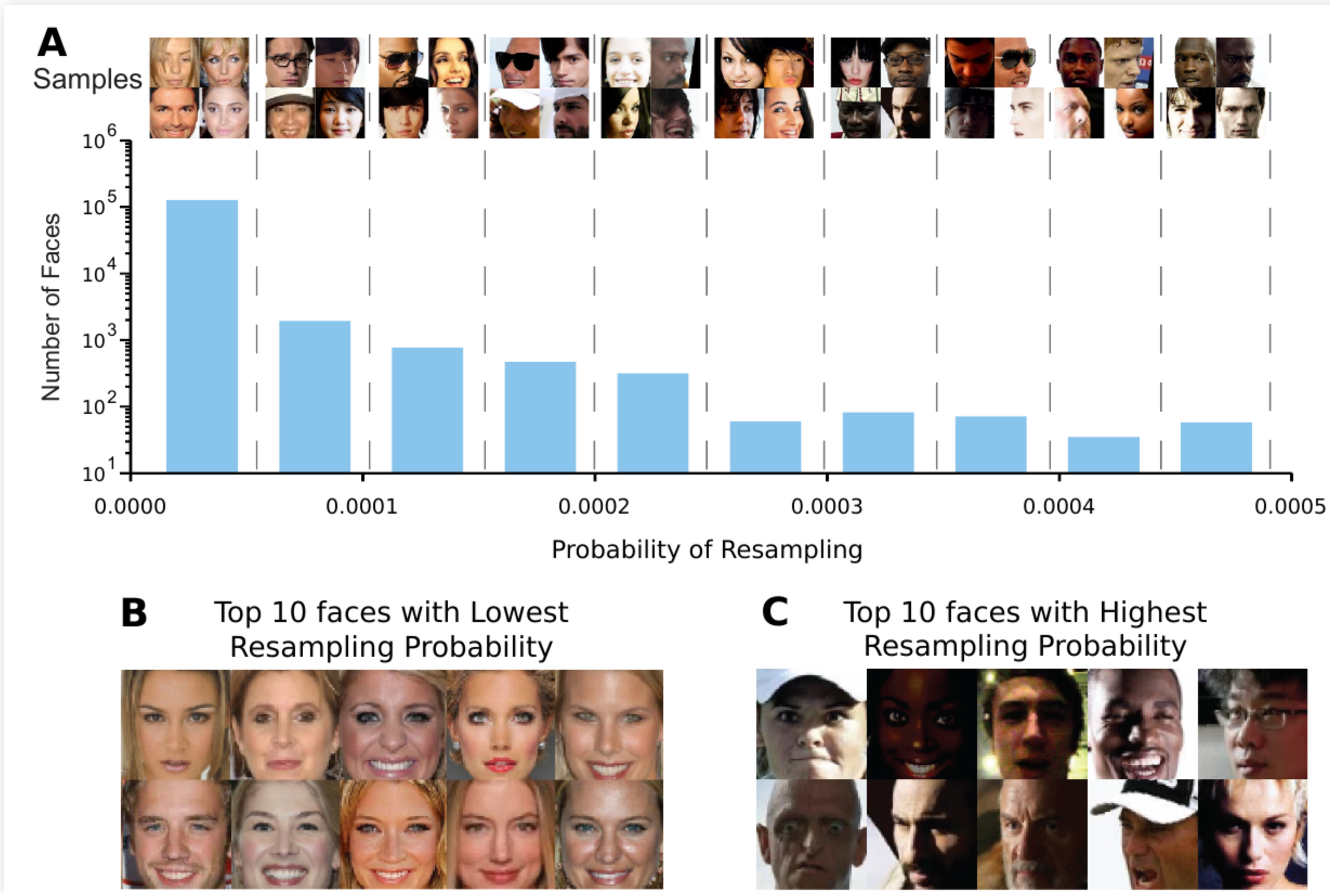


Amini et al, Uncovering and mitigating algorithmic bias through learned latent structure, 2019

- Minimize weighted cross-entropy, KL divergence and reconstruction

Mitigating Bias

- Sampling probability is inverse of density in manifold region



Amini et al, Uncovering and mitigating algorithmic bias through learned latent structure, 2019

Conclusion

Conclusion

Bias is good

- Biases are fundamental for learning. They are what we learn
  - Correlations between features and values to predict

But bias is bad

- Whenever these correlations arise from mistakes (e.g. sampling errors)
- Or lead to unfairness:
  - Results that depend on characteristics that should not be used
  - Or systems that perform poorly for some groups

Conclusion

Best practices

- Consider appliction and impact of models
- Beware of discrimination based on protected characteristics
  - And correlation with other attributes
- Be critical of the data used for training
- Evaluate performance on different groups
- If in doubt, mitigate by actively reducing biases

Summary

- Bias: fundamentally an ethical problem
- Sources of bias
  - Sampling, population, assumptions
- Mitigating undesirable biases