Explainable and Fair AI

27 November 2020

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

Artificial Intelligence

- Human decision makers are susceptible to prejudice and bias; e.g., gender and racial stereotypes.

  Machine Learning is free from such bias as it is no longer based on our gut-feeling, but on facts and statistics.

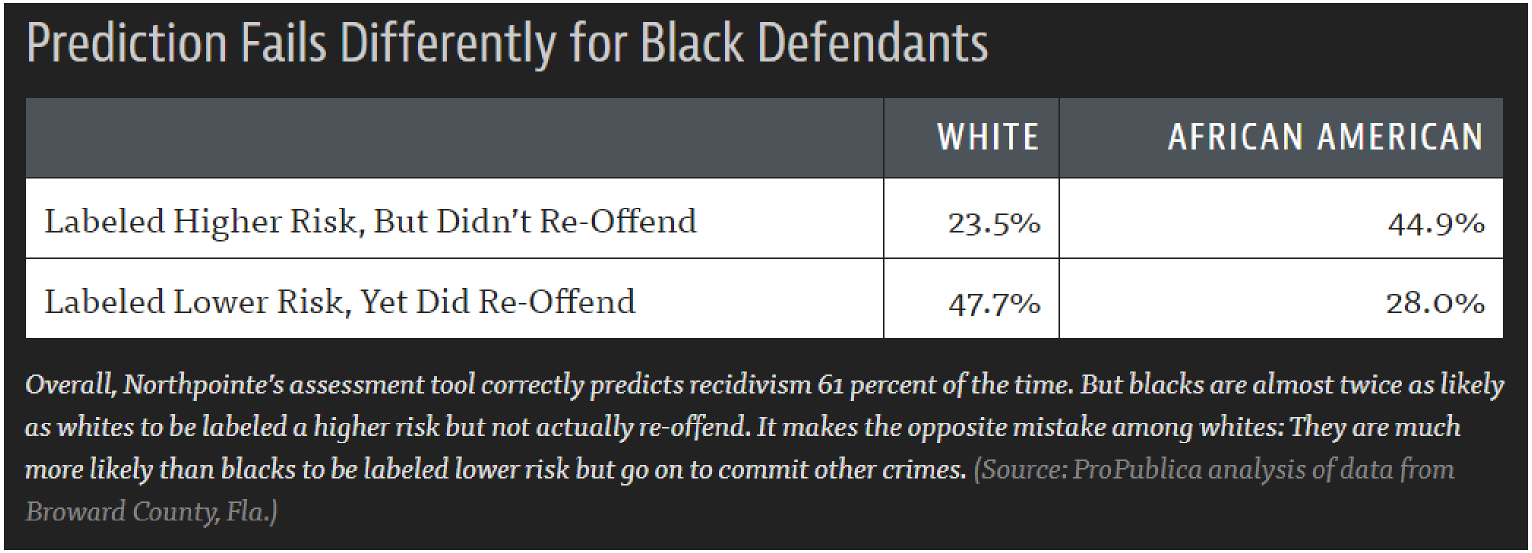
However …

- AI uses *correlations* learned from *historical* data to make predictions about the future
- It has been shown that models may unintentionally pickup bias from training data or introduce new biases

- Particularly problematic because many models are black boxes and decisions are hard to explain

Example: ProPublica Study (2016)

- Predicting risk for recidivism



Prediction Fails Differently for Black Defendants

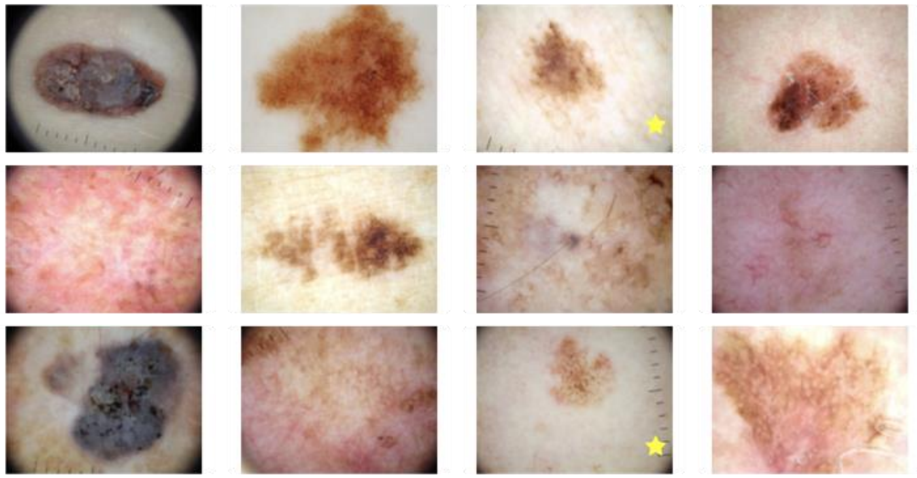
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

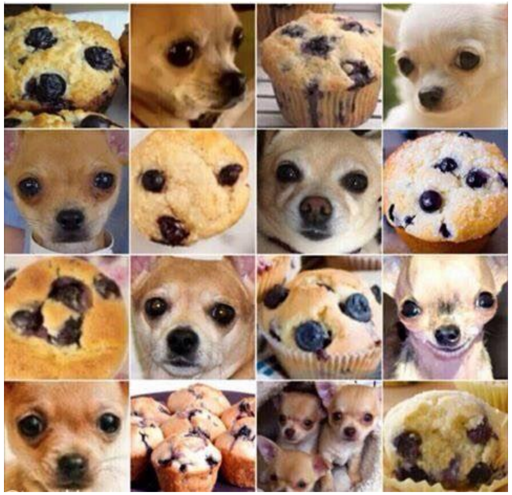
Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk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

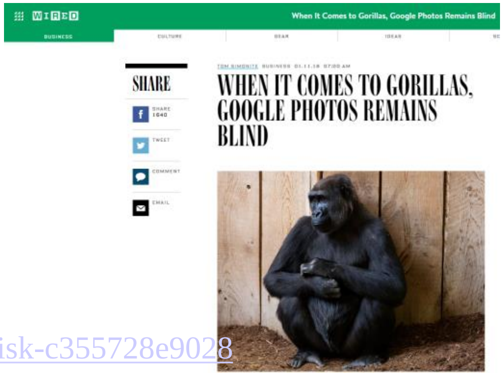
Need for Explanations

Explain

Trust | Compliance | Improve







WHEN IT COMES TO GORILLAS, GOOGLE PHOTOS REMAINS BLIND

https://towardsdatascience.com/is-the-medias-reluctance-to-admit-ai-s-weaknesses-putting-us-at-risk-c355728e9028

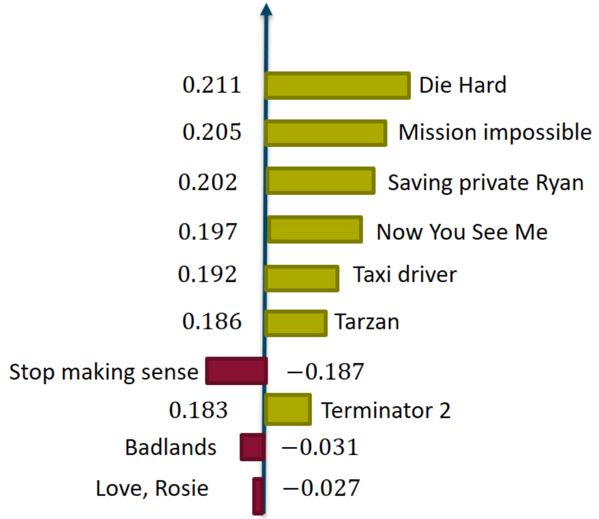
Instance-Based Explanations

User: Sam

WHY?

Sam watched 120 movies
Sam is predicted as *male*

LIME: **L**inear **I**nterpretable **M**odel-Agnostic **E**xplainer (k=10)



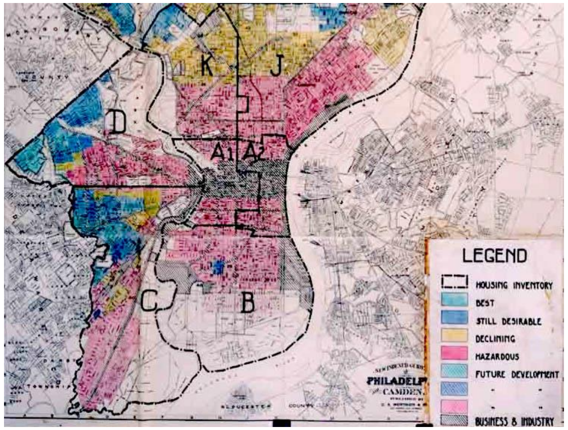
EDC: **E**vi**D**ence **C**ounterfactual

IF Sam would not have watched ***{Taxi driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me, Interstellar}***, **THEN** his predicted class would change from male to female

Need for Methods avoiding Bias

- Based on explanations we may identify biases that need to be removed from predictions

- Biases based on:
  - Causality vs correlation
  - Historical biases
  - Lack of information



Prof. dr. Toon Calders
Toon.calders@uantwerpen.be
University of Antwerp
Antwerp Tax Academy, DigiTax Research Center