Topics in Data Ethnics : Bias

.

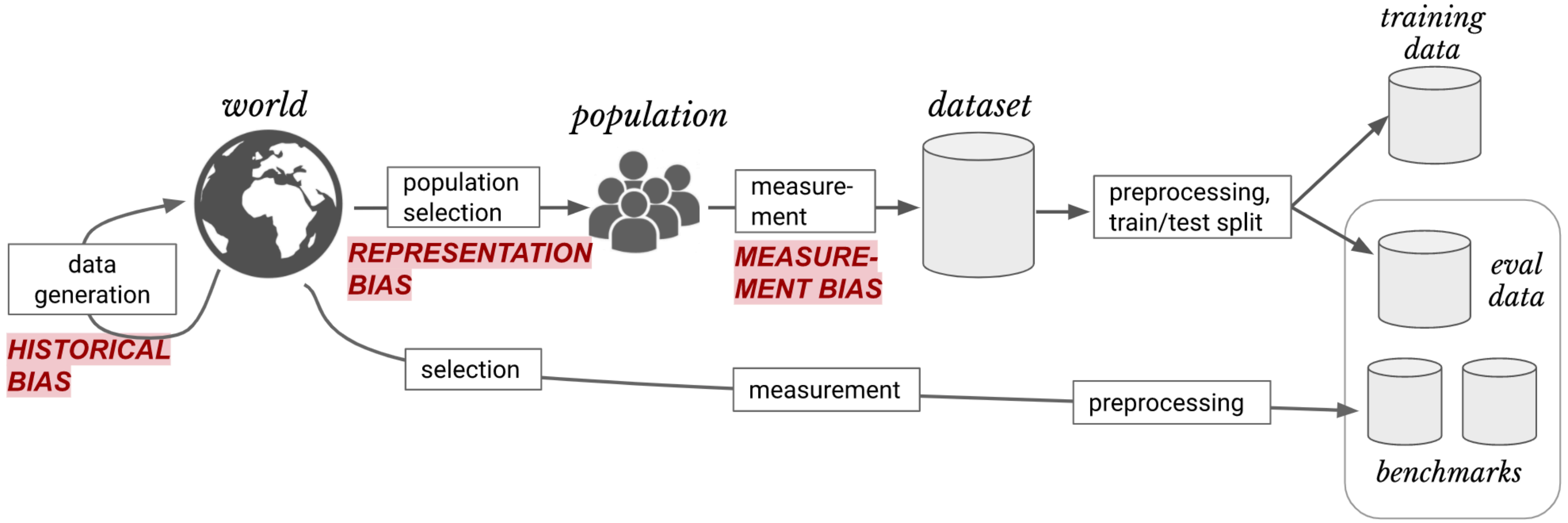
DongJun Min

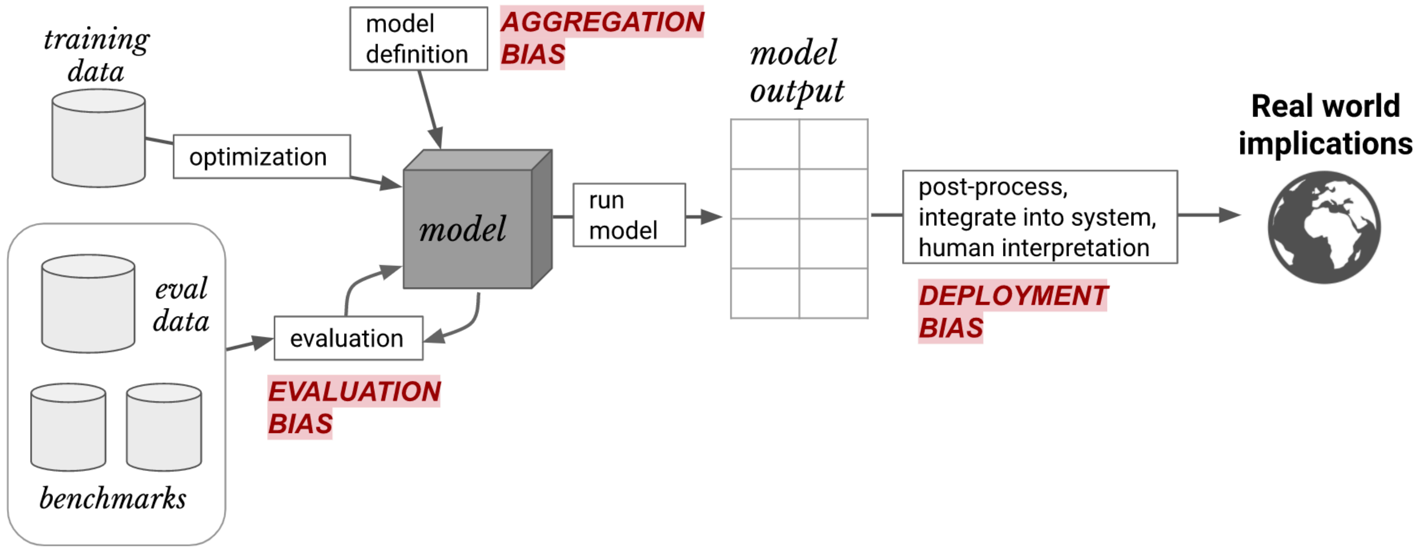
0. Bias

A Framework for Understanding Unintended Consequences of Machine Learning *By Harini Suresh et al.*

: Social science concept of bias

1. Historical Bias
2. Representation Bias
3. Measurement Bias
4. Aggregation Bias
5. Evaluation Bias
6. Deployment Bias





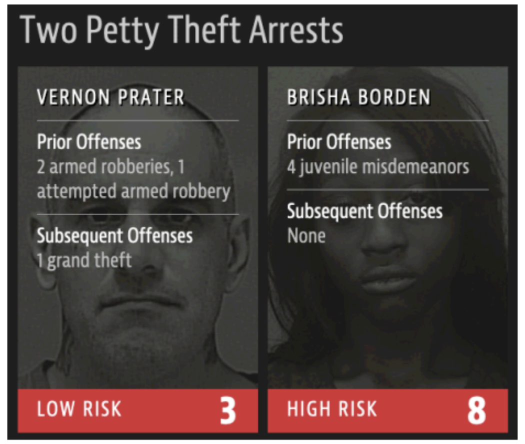
1. Historical bias

people are biased, processes are biased, and society is biased

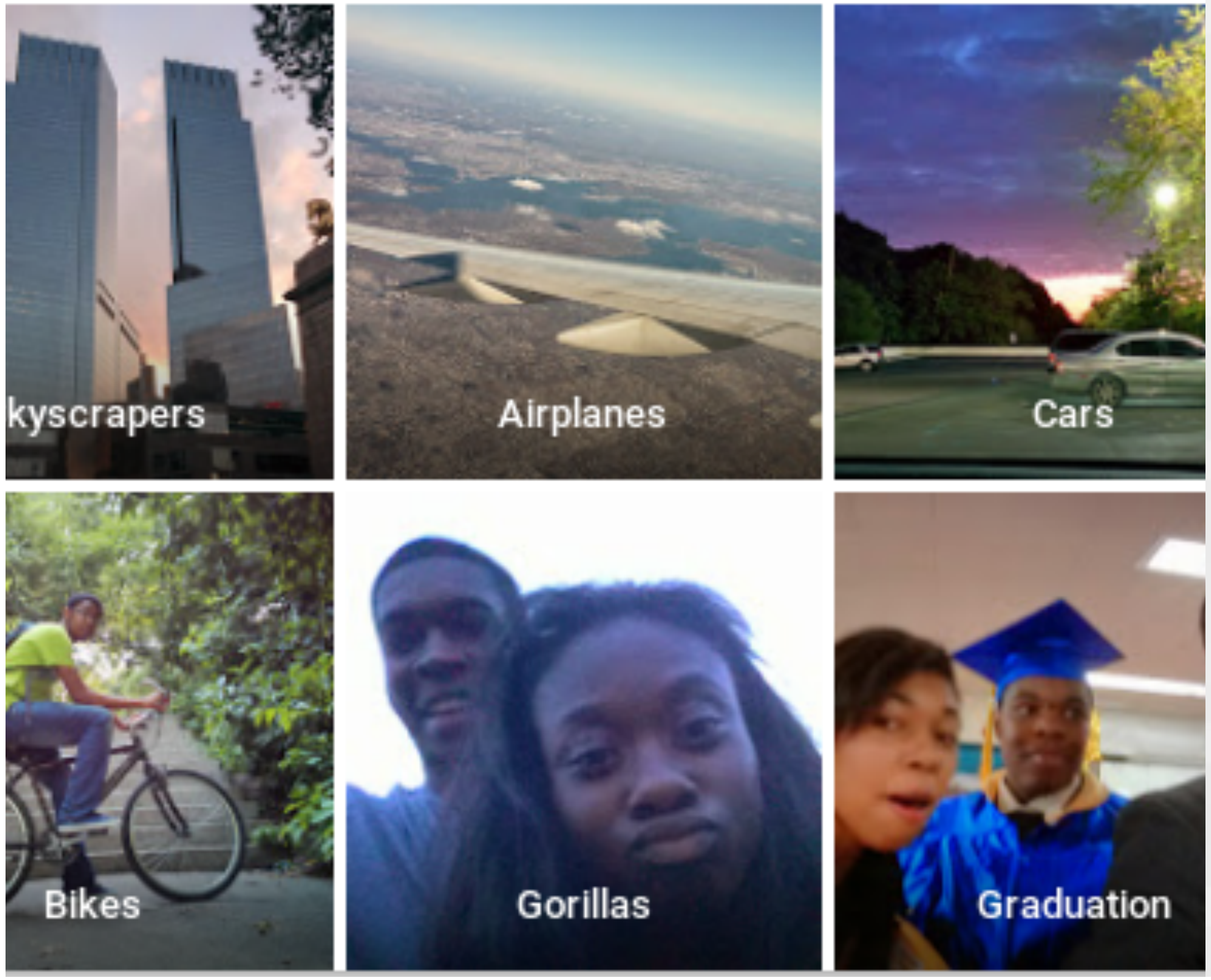
Racial Bias, COMPAS software

Prediction Fails Differently for Black Defendants

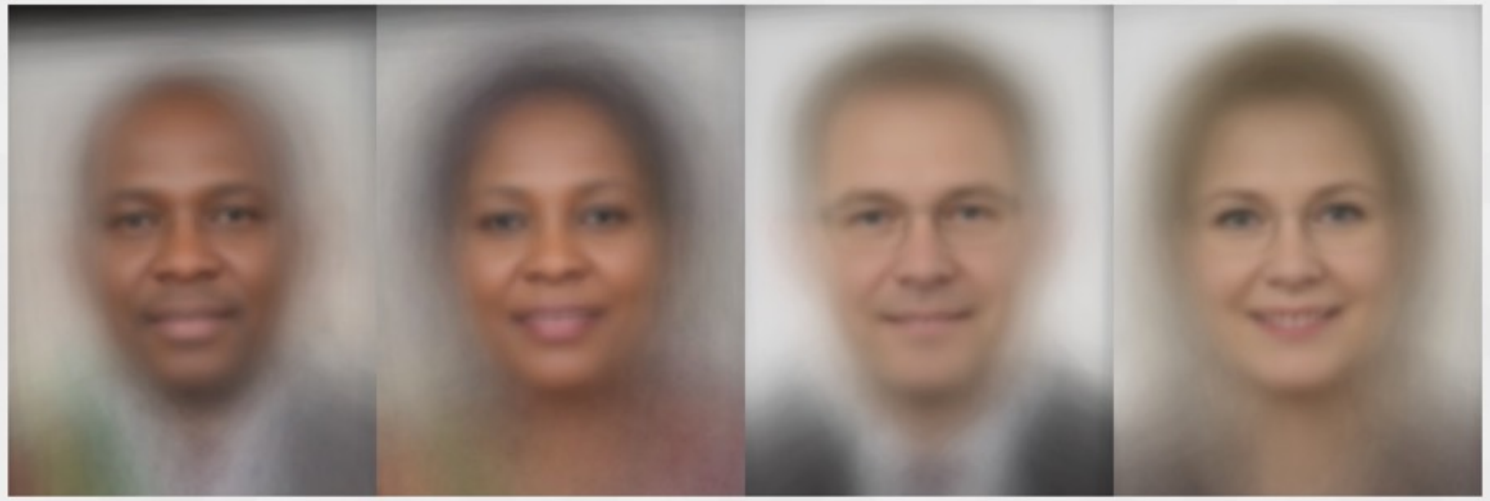
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%



1. Historical bias

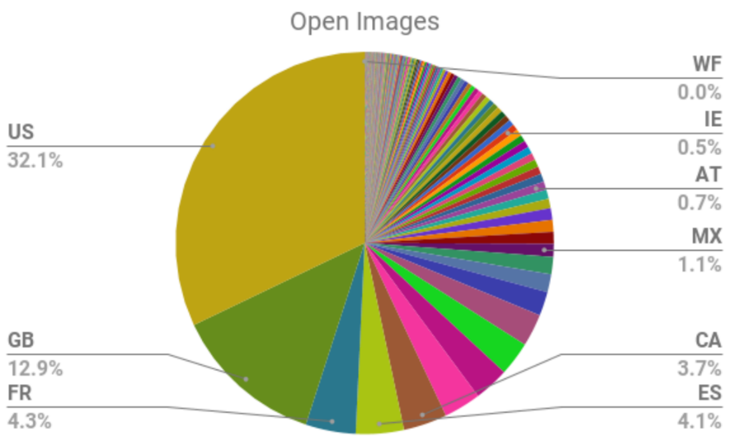


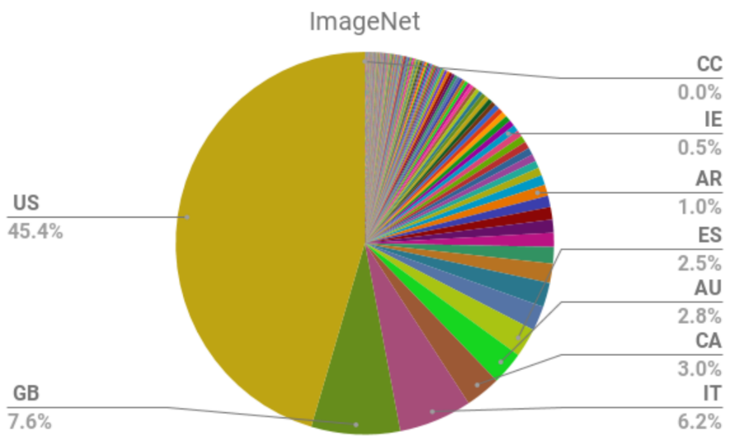
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

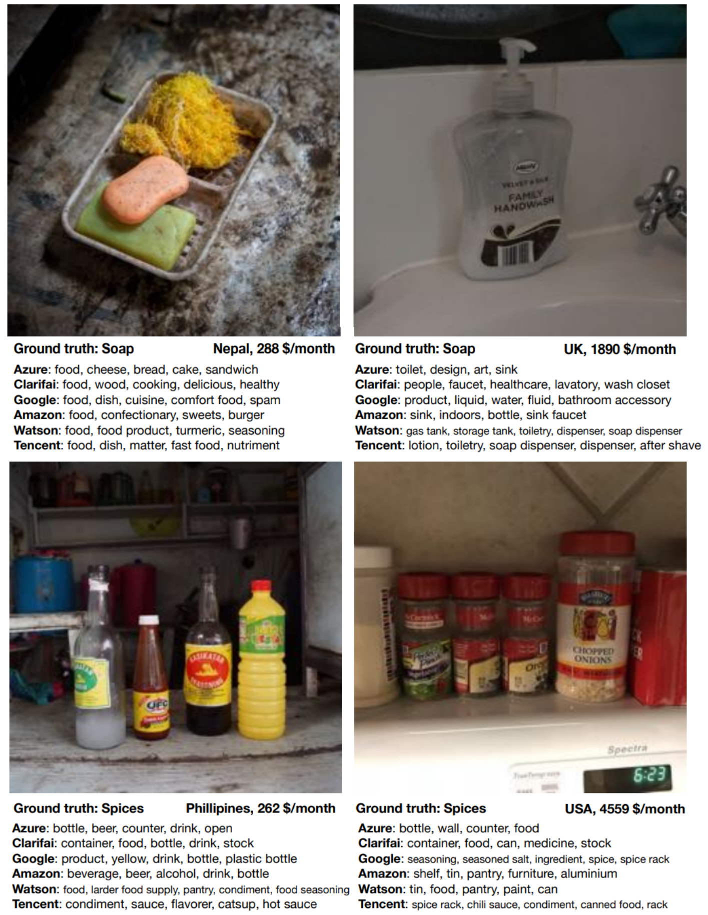


1. Historical bias

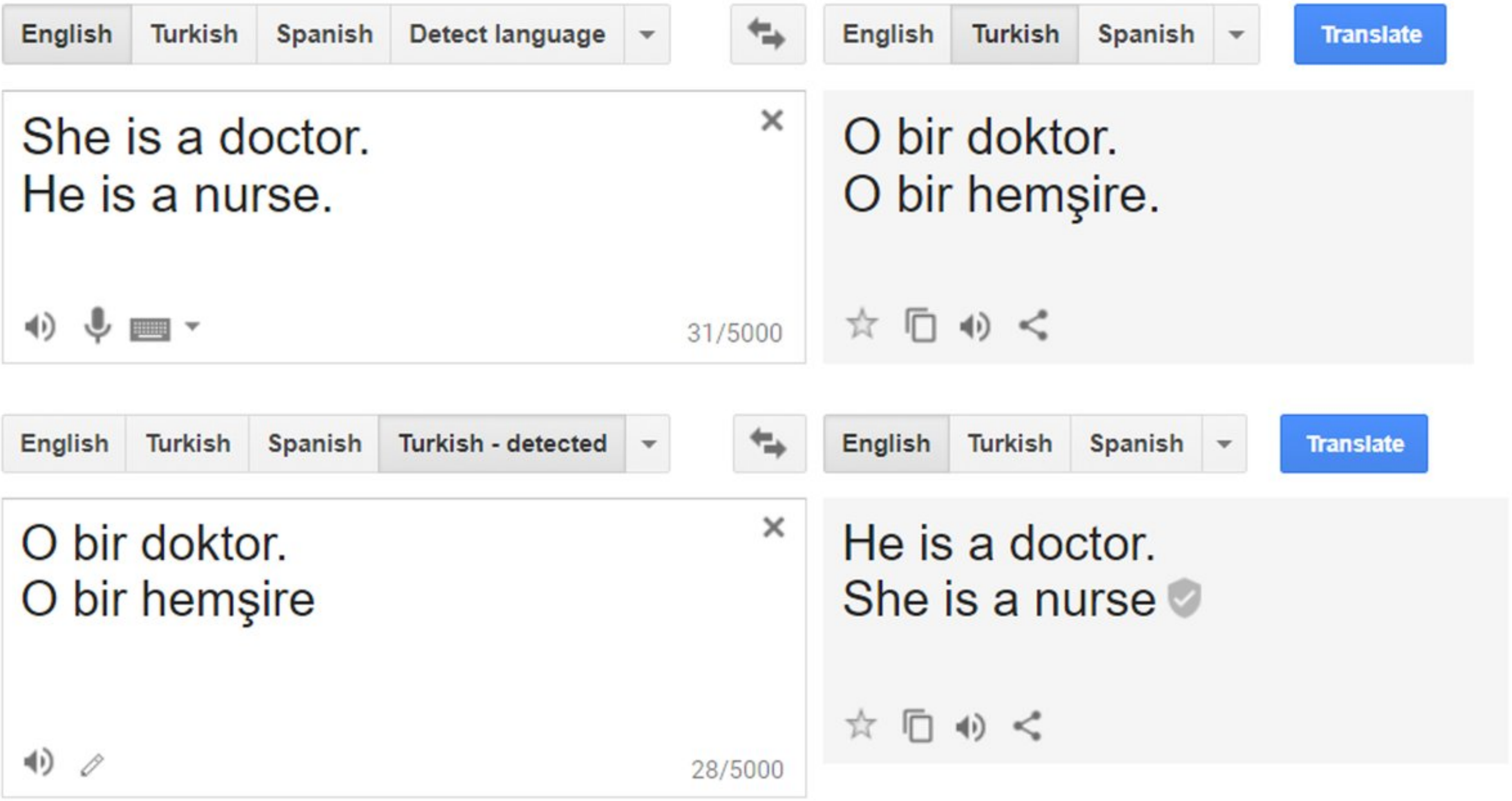
1.*[No Classification without Representation:Assessing Geodiversity Issues in Open Data Sets for the Developing World]* *By Shreya Shankar et al.*
2.*[Does Object Recognition Work for Everyone?]* *By Terrance DeVries et al.*







1. Historical bias



2. Measurement bias

What factors are most predictive of stroke?

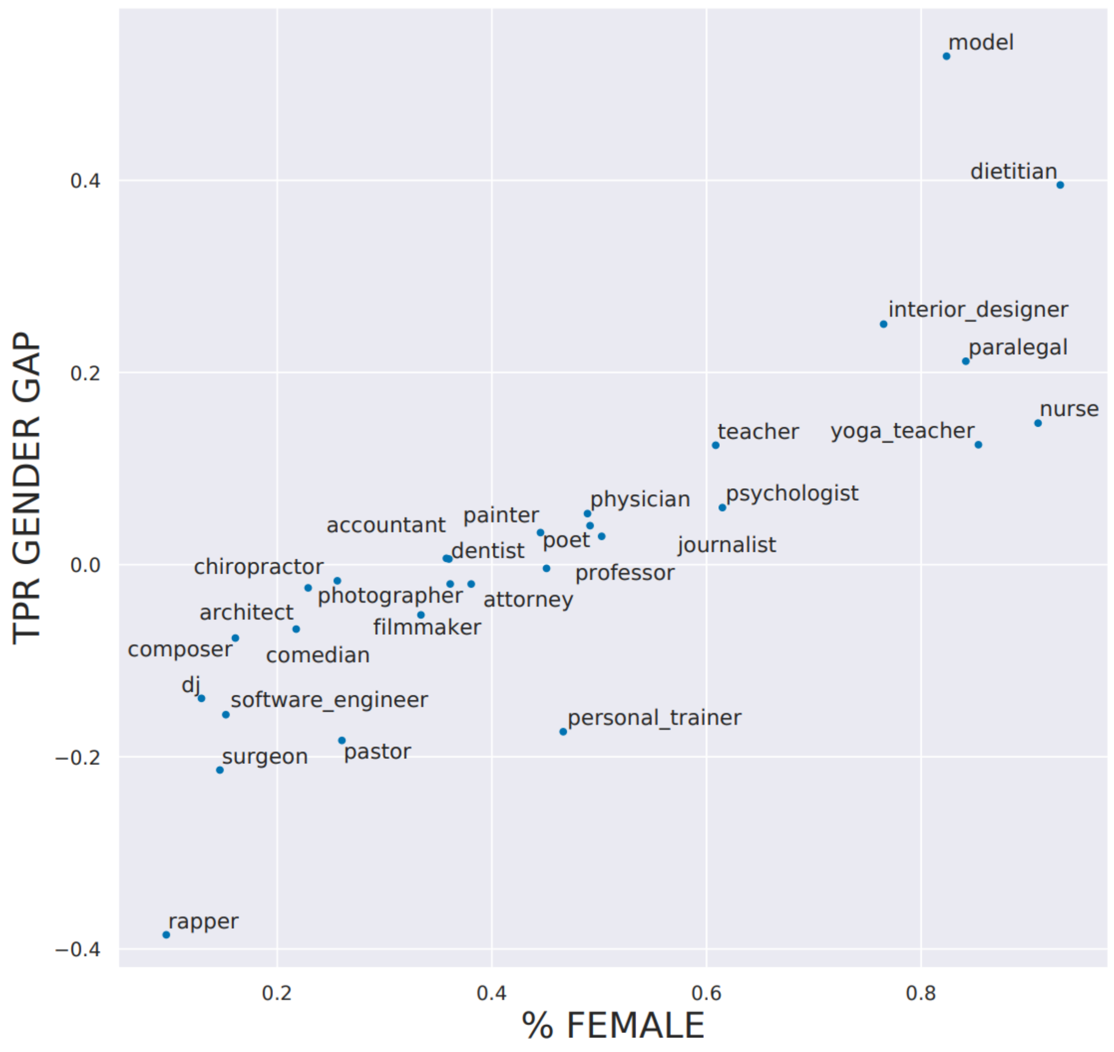
- Prior stroke
- Cardiovascular disease
- Accidental injury
- Benign breast lump
- Colonoscopy
- Sinusitis

3. Aggregation bias

Can occurs model that does not include these important variables and interactions.

4. Representation bias

Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting By Maria De-Arteaga et al.



5. Addressing different types of bias

Does Machine Learning Automate Moral Hazard and Error By Sendhil Mullainathan and Ziad Obermeyer

Consider these points about machine learning algorithms:

» Machine learning can create feedback loops

» Machine learning can amplify bias

» Algorithms & humans are used differently

» Technology is power

5. Addressing different types of bias

- People are more likely to assume algorithms are objective or error-free (even if they're given the option of a human override).
- Algorithms are more likely to be implemented with no appeals process in place.
- Algorithms are often used at scale.
- Algorithmic systems are cheap.