FAIRNESS IN MACHINE LEARNING

Toniann Pitassi

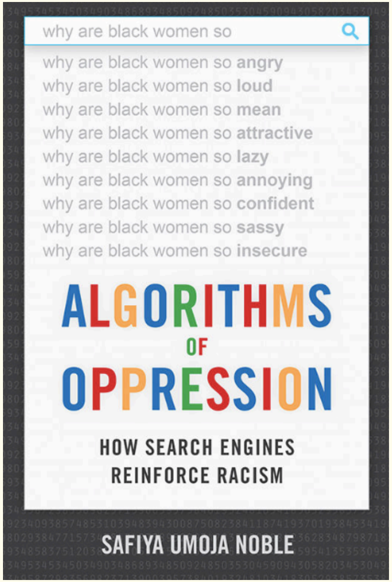
UNIVERSITY OF TORONTO

INSTITUTE FOR ADVANCED STUDY · TRUTH · BEAUTY · FOUNDED 1930

Vector Institute

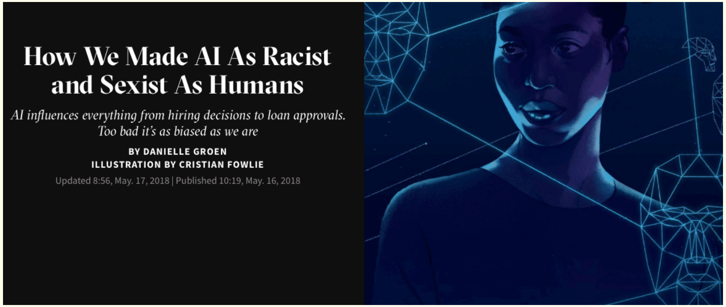
WHY WAS I NOT SHOWN THIS AD?

BIAS IN MACHINE LEARNING?



Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Joy Buolamwini



The Walrus, 2018

Fairness in ML: Goals

Identify and mitigate bias in ML-based decision-making, in all aspects of data pipeline

CLASSIFICATION

$x \in \mathcal{U}$ feature vector

$y = f(x)$ actual value (0 or 1)

$\hat{y}$ predicted value

GOAL

given a lot of labelled examples from population

Learn a classifier that is accurate

on > 99% of population

CLASSIFICATION

$x \in \mathcal{U}$ feature vector

$y = f(x)$ actual value (0 or 1)

$\hat{y}$ predicted value

EXAMPLES:

- Recognize if an image contains a car
- Predict (from resume) if candidates gets interview
- Predict if criminal will recidivate

FAIR CLASSIFICATION

$x \in \mathcal{U}$	feature vector
$y = f(x)$	actual value (0 or 1)
$\hat{y}$	predicted value
A	protected group

GOAL

Learn a classifier that is:

- accurate
- fair with respect to A

FAIR CLASSIFICATION : DEFINITIONS

Most common way is to define "fair" is to require some invariance/independence with respect to the sensitive attribute

FAIR CLASSIFICATION : DEFINITIONS

Most common way is to define "fair" is to require some invariance/independence with respect to the sensitive attribute

- DEMOGRAPHIC PARITY: $\hat{Y} \perp A$

FAIR CLASSIFICATION : DEFINITIONS

Most common way is to define "fair" is to require some invariance/independence with respect to the sensitive attribute

- DEMOGRAPHIC PARITY: $\hat{Y} \perp A$
- EQUALIZED ODDS: $\hat{Y} \perp A \mid Y$

FAIR CLASSIFICATION : DEFINITIONS

Most common way is to define "fair" is to require some invariance/independence with respect to the sensitive attribute

- DEMOGRAPHIC PARITY: $\hat{Y} \perp A$
- EQUALIZED ODDS: $\hat{Y} \perp A \mid Y$
- EQUALIZED CALIBRATION: $Y \perp A \mid \hat{Y}$

HISTORY

50 Years of Test (Un)fairness: Lessons for Machine Learning by Hutchinson & Mitchell

Flurry of activity in ML trying to define fairness mirrors efforts 50+ years ago to define bias and fairness in educational testing

US Civil Rights Act of 1964 outlawed discrimination on basis of race, color, religion, sex, national origin; followed by questions whether assessment tests were discriminatory

Example: on formal model predicting educational outcome from test scores (Cleary 1966)

> "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low."

Parallels --

- Test items or questions – input features
- Responses – values of features
- Linear model predicts test score– simple outcome prediction models

HISTORY

- Cleary studied the relation between SAT scores and college GPA using real-world data from 3 schools, (racial data from admissions office, NAACP list of students, class pictures) -- did not find racial bias

- Overall many parallels: formal notions of fairness based on population subgroups, the realization that some fairness criteria are incompatible with one another

- Example: Thorndike (1971) pointed out that different groups vary in false positive/negative rates, should be balanced between the groups via different thresholds

- Research died out, possibly due to focus on quantitative definitions, separation from social, legal, societal concerns – cautionary tale?

How can Learned classifiers be Biased?

Sources of Bias / Discrimination ?

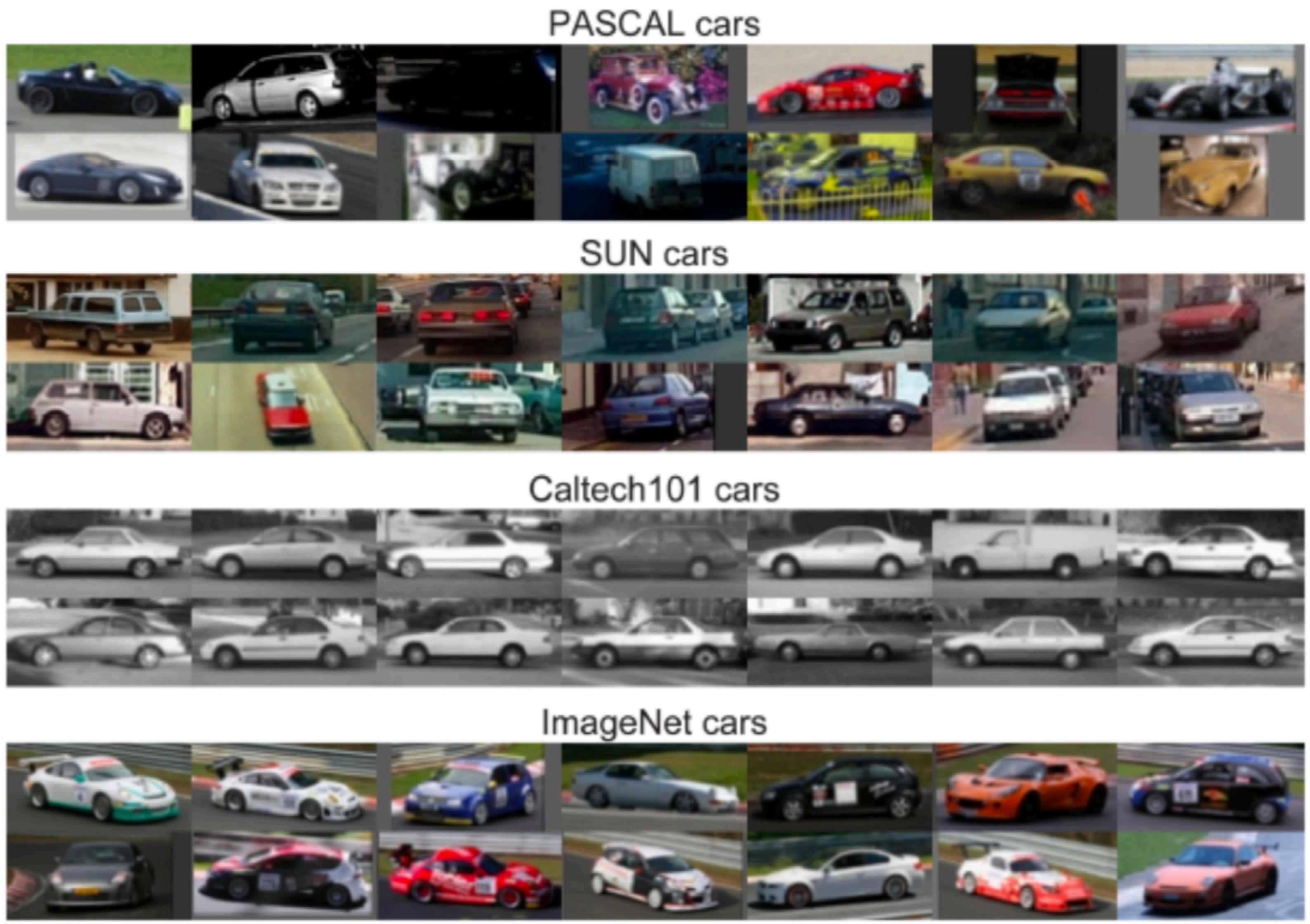
- Imbalanced data / impoverished data
- Labelled data incorrect / Noisy
- Measurements – selective choices, measurement issues

DATA

- ML prediction error imbalanced
- Compound Injustices (Hellman)

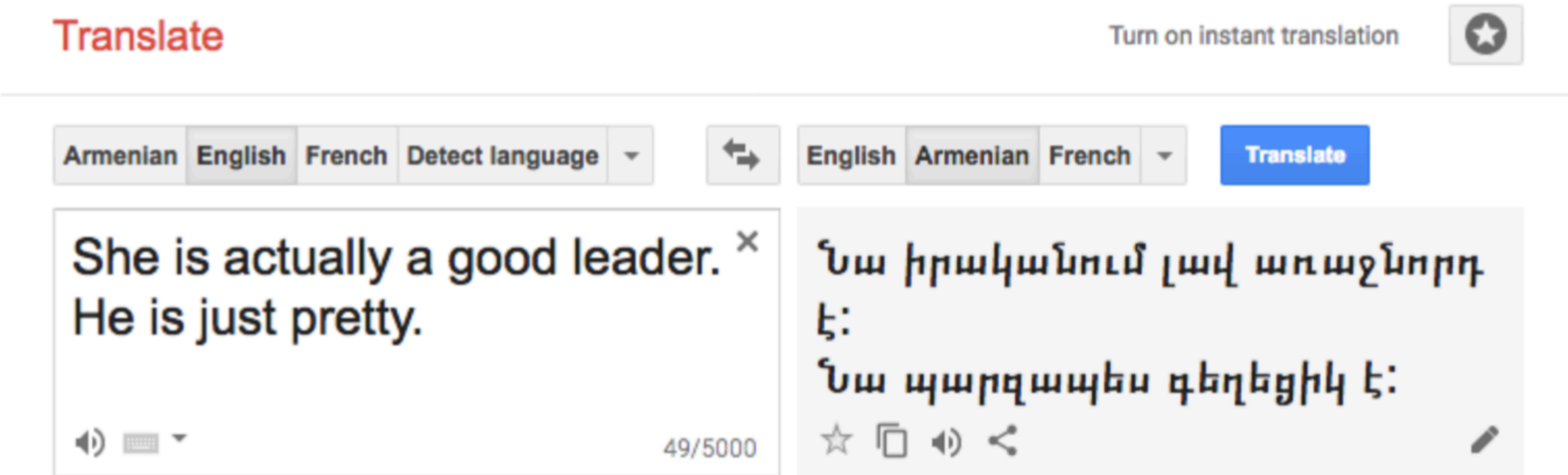
MODEL

EXAMPLE OF BIAS

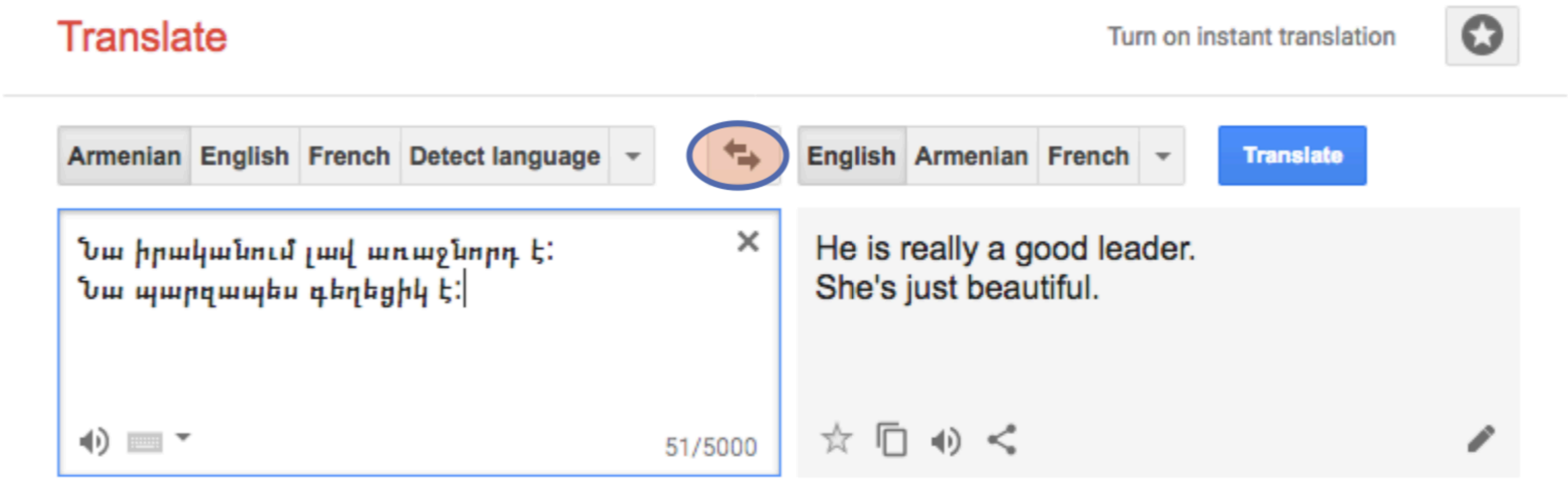


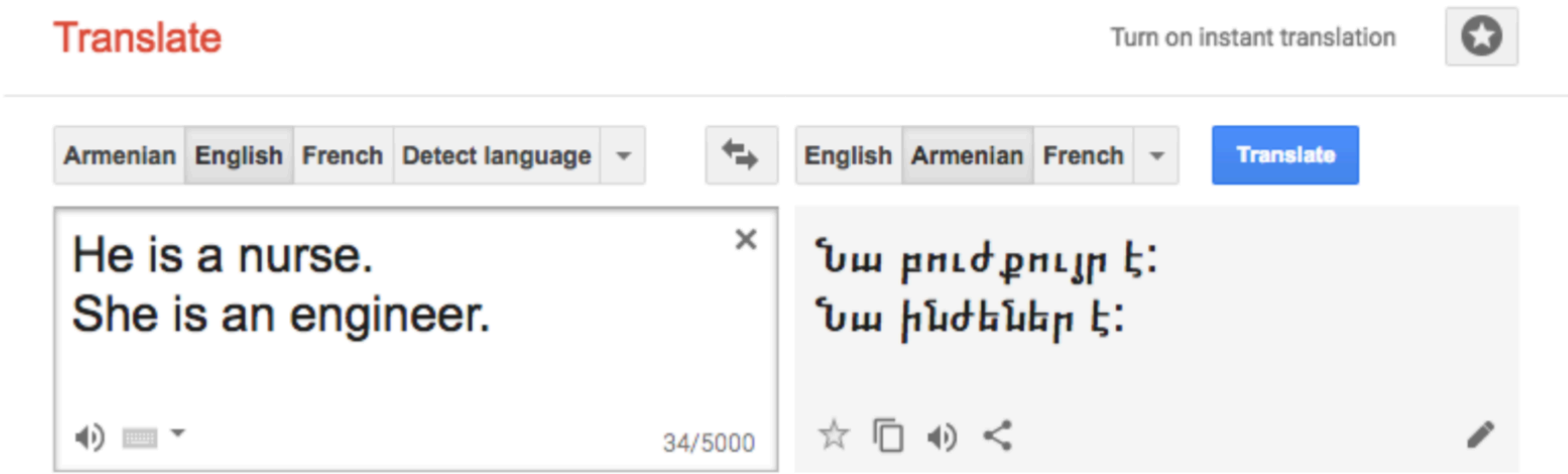
Predictor trained on Caltech101 won't recognize sports cars

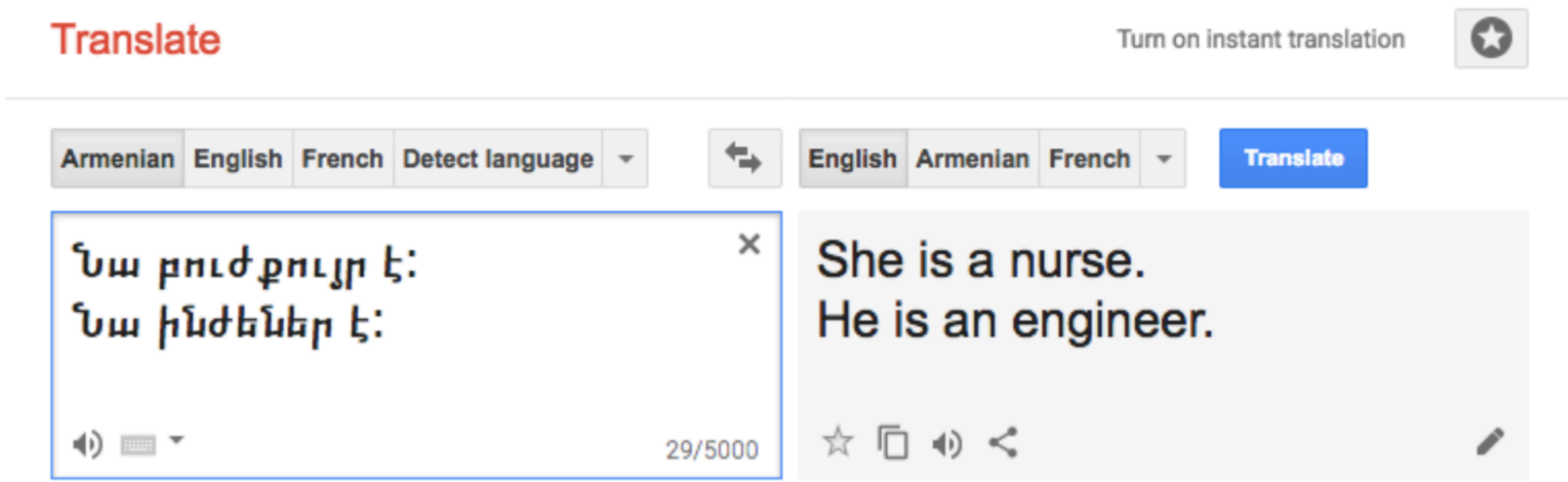
TRANSLATION



TRANSLATION




Translate
Turn on instant translation
Armenian English French Detect language
English Armenian French
Translate
He is a nurse.
She is an engineer.
34/5000
Նա բուժքույր է:
Նա ինժեներ է:


Translate
Turn on instant translation
Armenian English French Detect language
English Armenian French
Translate
Նա բուժքույր է:
Նա ինժեներ է:
29/5000
She is a nurse.
He is an engineer.

APPROACHES TO FAIR CLASSIFICATION

I. Model-centered

- Add fairness criteria to objective function
  - Regularizer
  - Adversarial
- Postprocess to achieve fairness

II. Data-centered

- change/Modify data
- Learn a fair representation of data

HURDLES AND SUBTLETIES

1. Seems impossible to have <u>one</u> good definition of fairness

- DEMOGRAPHIC PARITY: $\hat{Y} \perp A$
- EQUALIZED ODDS: $\hat{Y} \perp A \mid Y$
- EQUALIZED CALIBRATION: $Y \perp A \mid \hat{Y}$

<u>Theorem</u> These three definitions of fairness are mutually exclusive

Example

COMPAS : risk assessment program

Propublica concluded that COMPAS is biased:

- more blacks incorrectly predicted to recidivate

	Black	White
Accuracy	64.9	65.7
False Positive Rate	40.4	25.4
False Negative Rate	30.9	47.9

IS CLASSIFIER BIASED ?

Probublica says:

Blacks face higher false positive rates so violates equalized odds

Northpointe's defense:

Scores satisfy equalized calibration and we can't have both.

HURDLES AND SUBTLETIES

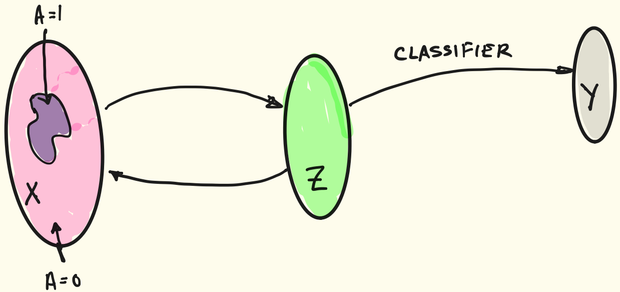
1. Seems impossible to have <u>one</u> good definition of fairness

Alternatives

- individual fairness

  underlying task-specific similarity metric
  ensure similar treatment for similar people

- fair representations



HURDLES AND SUBTLETIES

1. Seems impossible to have <u>one</u> good definition of fairness

2. How do we even know which groups are being treated unfairly?

HURDLES AND SUBTLETIES

1. Seems impossible to have <u>one</u> good definition of fairness

2. How do we even know which groups are being treated unfairly?

    - multigroup fairness
    - fairness under changing dynamics

Challenges

* Understand dynamics of unfairness

* Impoverished Data:
  what would have happened if ...
  Causal inference ?

•
•
•

AND OPPORTUNITIES !

* A chance to understand, identify, challenge, and improve decision making (not just automated decision making)

Thanks !