SOURCES OF BIAS

TONIANN PITASSI RICHARD ZEMEL

CSC 2541

SEPTEMBER 24, 2019

OUTLINE

Brief recap of fairness definitions

Next class: fairness mechanisms – methods to address unfairness of classifiers

Today – various studies of biases in data

- What are the various notions of bias?
- What are the sources of the bias?

FAIR CLASSIFICATION

Explosion of fairness research over last five years

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction (or R)
- A, a sensitive attribute (race, gender, age, socio-economic status)

We want to learn a classifier that is:

- accurate
- fair with respect to A

FAIR CLASSIFICATION: DEFINITIONS

Definitions based on predicted outcomes:

- Demographic / statistical parity
- Conditional statistical parity (loan conditioned on credit history, amount, employment)

Definitions based on predicted and actual outcomes:

- Balanced PPV (FDR) – predictive equality
- Balanced FNR (TPR) – equal opportunity
- Balanced FNR and FPR – equalized odds

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** $\text{PPV} = \frac{TP}{TP+FP}$ $\text{TPR} = \frac{TP}{TP+FN}$	**False Positive (FP)** $\text{FDR} = \frac{FP}{TP+FP}$ $\text{FPR} = \frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** $\text{FOR} = \frac{FN}{TN+FN}$ $\text{FNR} = \frac{FN}{TP+FN}$	**True Negative (TN)** $\text{NPV} = \frac{TN}{TN+FN}$ $\text{TNR} = \frac{TN}{TN+FP}$

FAIR CLASSIFICATION: DEFINITIONS

Most common way to define fair classification is to require some invarianc with respect to the sensitive attribute

- Demographic parity: $\hat{Y} \perp A$
- Equalized Odds: $\hat{Y} \perp A|Y$
- Equal Opportunity: $\hat{Y} \perp A|Y = y$, for some y
- Equal (Weak) Calibration: $Y \perp A|\hat{Y}$
- Equal (Strong) Calibration: $Y \perp A|\hat{Y}$ and $\hat{Y} = P(Y = 1)$
- Fair Subgroup Accuracy: $\mathbb{1}[Y = \hat{Y}] \perp A$

VISUALIZATION

Equality of opportunity in supervised learning, by Hardt, Price, Srebro

Introduce equalized odds, opportunity – minimize both false positive and false negative rates, or just false positives

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = y\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = y\right\}, \quad y \in \{0, 1\}$$

Very simple approach – just adjust thresholds on pre-defined scores to optimize selected measure

Useful visualization:

http://research.google.com/bigpicture/attacking-discrimination-in-ml/

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

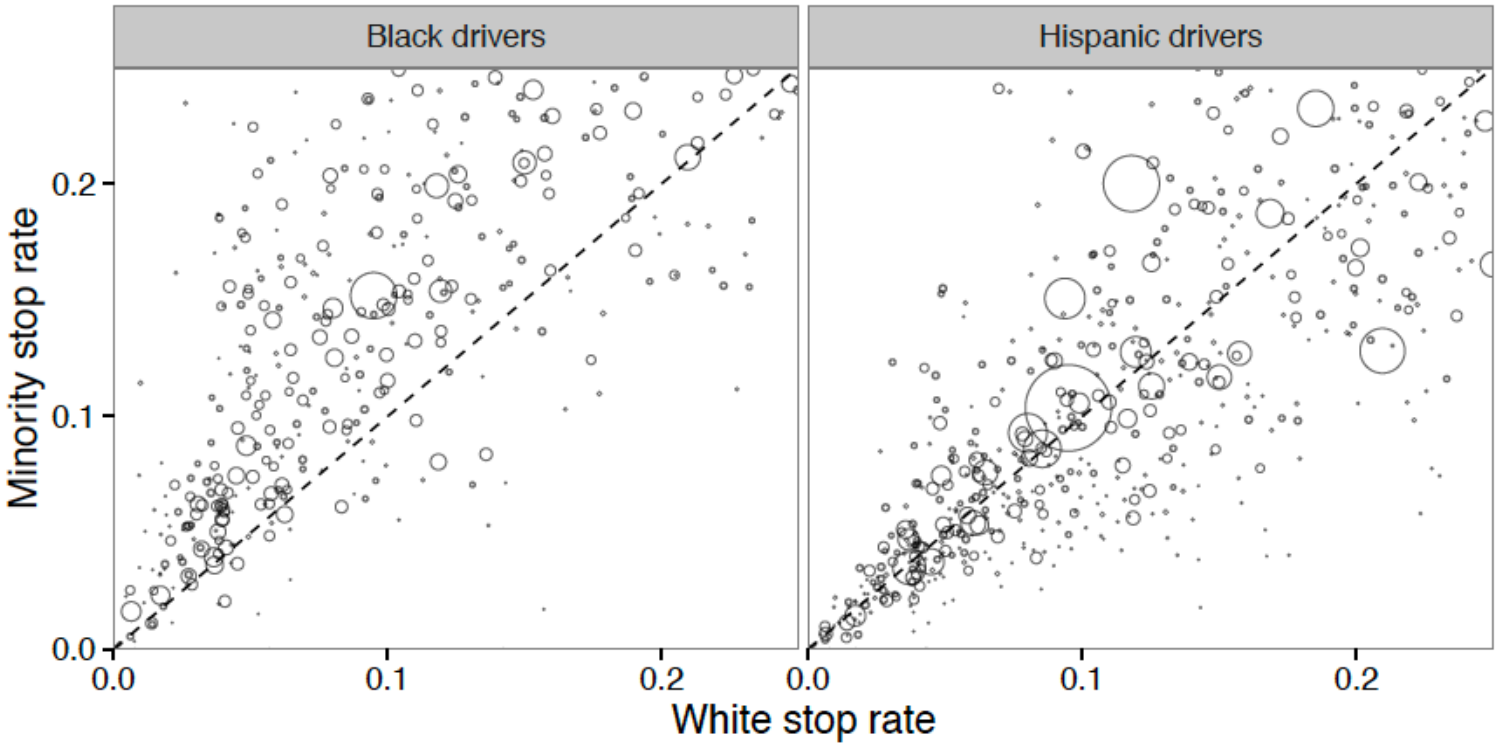
- Example: Thorndike (1971) pointed out that different groups vary in false positive/negative rates, should be balanced between the groups via different thresholds

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

STOP RATES

Stops per person of driving age, in 16 states with location recorded; relative to share of driving-age population

Each point specific to a location



A large-scale analysis of racial disparities in police stops across the United States, Pierson, E., et al., 2017

STOP RATE DEMOGRAPHICS

$$y_{ragly} \sim \text{NegBin}\left(n_{ragly} e^{\mu+\alpha_r+\beta_a+\gamma_g+\delta_\ell+\epsilon_y}, \phi\right)$$

Fit negative binomial to observed stop rates

Blacks stopped 1.4x rate of white stops [exp(.37)]

	Stop	Citation	Search	Consent search	Arrest
Black	0.37 (0.01)	0.18 (0.00)	0.73 (0.01)	0.77 (0.03)	0.65 (0.01)
Hispanic	-0.40 (0.01)	0.29 (0.00)	0.54 (0.01)	0.62 (0.02)	0.69 (0.01)
Male	0.72 (0.00)	0.08 (0.00)	0.58 (0.01)	0.86 (0.02)	0.43 (0.01)
Age 20-29	0.65 (0.01)	-0.13 (0.01)	0.13 (0.01)	-0.38 (0.03)	0.38 (0.01)
Age 30-39	0.47 (0.01)	-0.35 (0.01)	-0.06 (0.01)	-0.79 (0.03)	0.30 (0.01)
Age 40-49	0.25 (0.01)	-0.47 (0.01)	-0.37 (0.01)	-1.20 (0.04)	-0.04 (0.01)
Age 50+	-0.53 (0.01)	-0.68 (0.01)	-0.80 (0.01)	-1.82 (0.04)	-0.47 (0.01)

Table 2: *Coefficients and standard errors for stop rate and post-stop outcome models.*

STOP RATE DEMOGRAPHICS

Analyze young males

	White	Black	Hispanic
Stop rate	0.29	0.42	0.19
Speeding citation	72%	75%	77%
Search	1.3%	2.7%	2.3%
Consent search	0.1%	0.3%	0.3%
Arrest	2.8%	5.3%	5.5%

Table 3: *Model-estimated rates for a typical 20-29 year-old male. The "speeding citation" outcome corresponds to receiving a citation rather than a warning (or no penalty) when pulled over for speeding. Negative binomial regression is used for stop rate (first row), benchmarked to the driving-age population; logistic regression is used for all other analyses. The stop rate regression includes controls for age, gender, stop location, and stop year; all other regressions additionally include controls for stop quarter, weekday, and hour (binned into three-hour segments).*

SEARCH & ARREST RATES

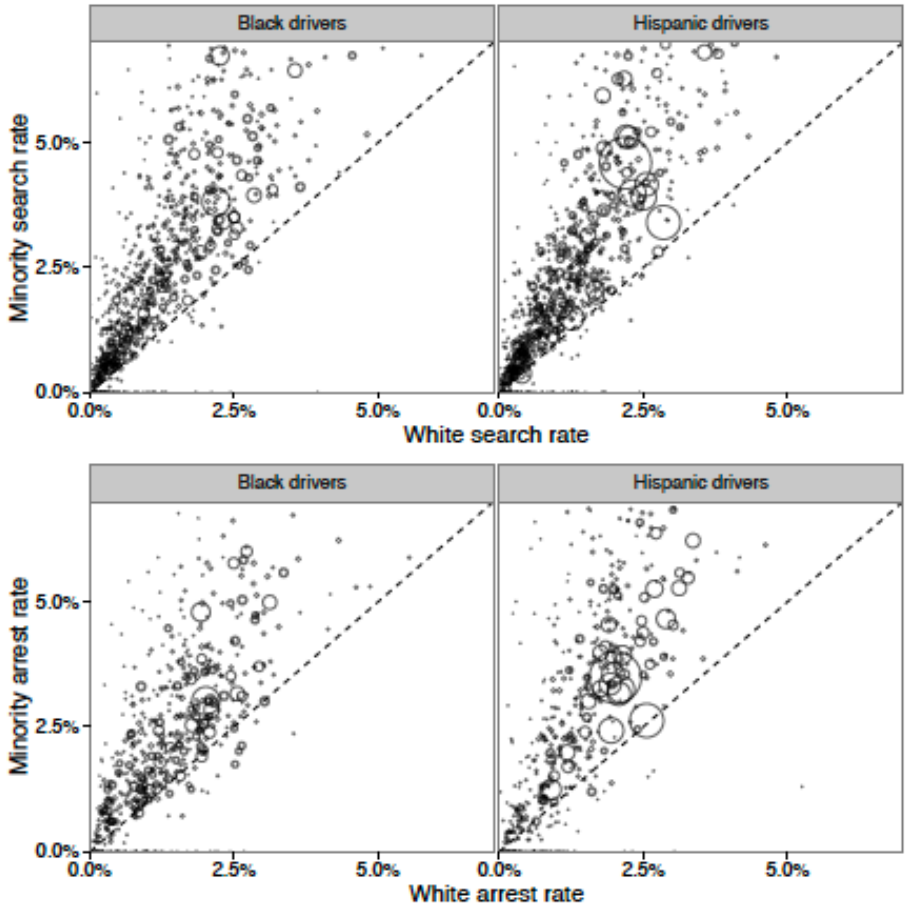


Figure 3: *Search rates (top) and arrest rates (bottom) by race and location among stopped drivers. In nearly every area, minorities are searched and arrested more often than whites. The search data cover 16 states, comprising a total of 56 million stops, and the arrest data include 40 million stops in 13 states.*

TEST FOR BIAS

Possible that one group more likely to carry contraband than another

Outcome test:

- Examine not search rate but hit rate – proportion of searches that turn up contraband (equal if just search rate disparities)
- Hispanics 22%, Whites and Blacks 28% stops yield contraband

Threshold test takes into account more factors

Hierarchical Bayesian model – considers officer's decision when to stop and search

Personal threshold on decision

DATASET BIAS: COMPUTER VISION

Unbiased look at dataset bias, Torralba & Efros, 2011

EASY TO CLASSIFY DATASET

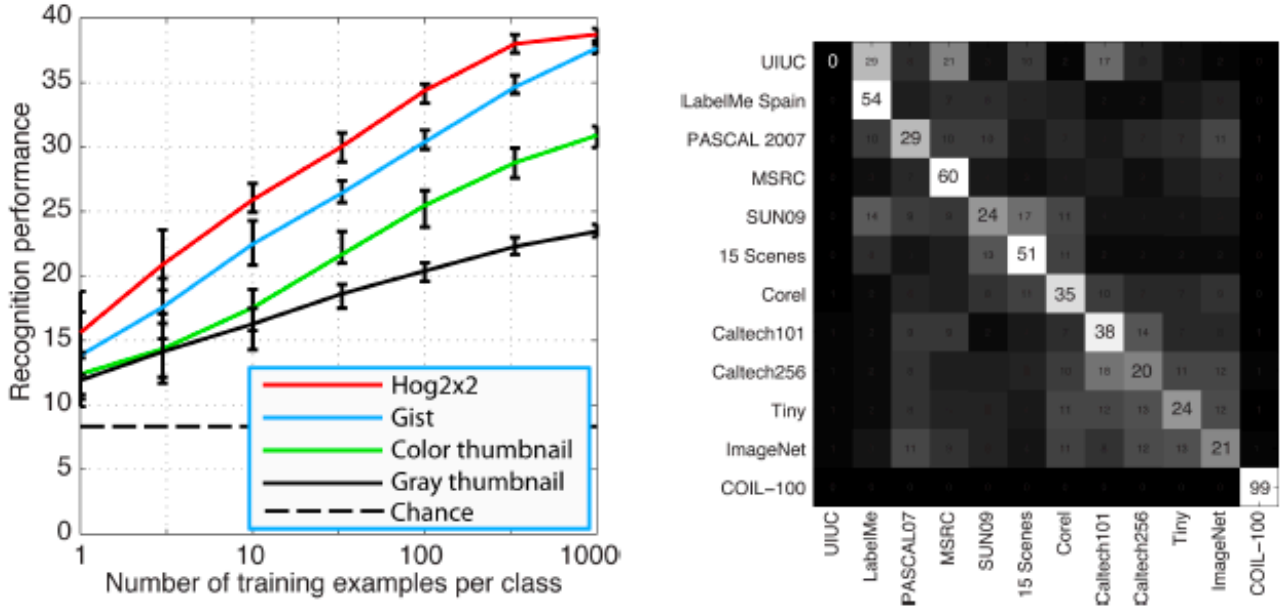


Figure 2. Computer plays *Name That Dataset*. Left: classification performance as a function of dataset size (log scale) for different descriptors (notice that performance does not appear to saturate). Right: confusion matrix.

EVOLUTION OF RECOGNITION DATASETS

Reaction against:

- Lab stock images → Lena
- Model-based approaches (staplers) → appearance-based (Tylenol bottles) [COIL]
- Simple backgrounds → complexity [Corel]
- Professional → internet [Caltech]
- Object-in-middle → clutter, many objects [MSRC, LabelMe]
- Small datasets → large scale [TinyImages, ImageNet]

EVALUATE DATASET BIAS

Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for "car" and "person" when training on one dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. "Self" refers to training and testing on the same dataset (same as diagonal), and "Mean Others" refers to averaging performance on all except self.

task	Train on: \ Test on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
"car" detection	SUN09	**69.8**	50.7	42.2	42.6	54.7	69.4	69.8	51.9	**26%**
	LabelMe	61.8	**67.6**	40.8	38.5	53.4	67.0	67.6	52.3	**23%**
	PASCAL	55.8	55.2	**62.1**	56.8	54.2	74.8	62.1	59.4	**4%**
	ImageNet	43.9	31.8	46.9	**60.7**	59.3	67.8	60.7	49.9	**18%**
	Caltech101	20.2	18.8	11.0	31.4	**100**	29.3	100	22.2	**78%**
	MSRC	28.6	17.1	32.3	21.5	67.7	**74.3**	74.3	33.4	**55%**
	Mean others	42.0	34.7	34.6	38.2	57.9	61.7	72.4	44.8	**48%**
"person" classification	SUN09	**16.1**	11.8	14.0	7.9	6.8	23.5	16.1	12.8	**20%**
	LabelMe	11.0	**26.6**	7.5	6.3	8.4	24.3	26.6	11.5	**57%**
	PASCAL	11.9	11.1	**20.7**	13.6	48.3	50.5	20.7	27.1	**-31%**
	ImageNet	8.9	11.1	11.8	**20.7**	76.7	61.0	20.7	33.9	**-63%**
	Caltech101	7.6	11.8	17.3	22.5	**99.6**	65.8	99.6	25.0	**75%**
	MSRC	9.4	15.5	15.3	15.3	93.4	**78.4**	78.4	29.8	**62%**
	Mean others	9.8	12.3	13.2	13.1	46.7	45.0	43.7	23.4	**47%**

EXAMPLE OF BIAS

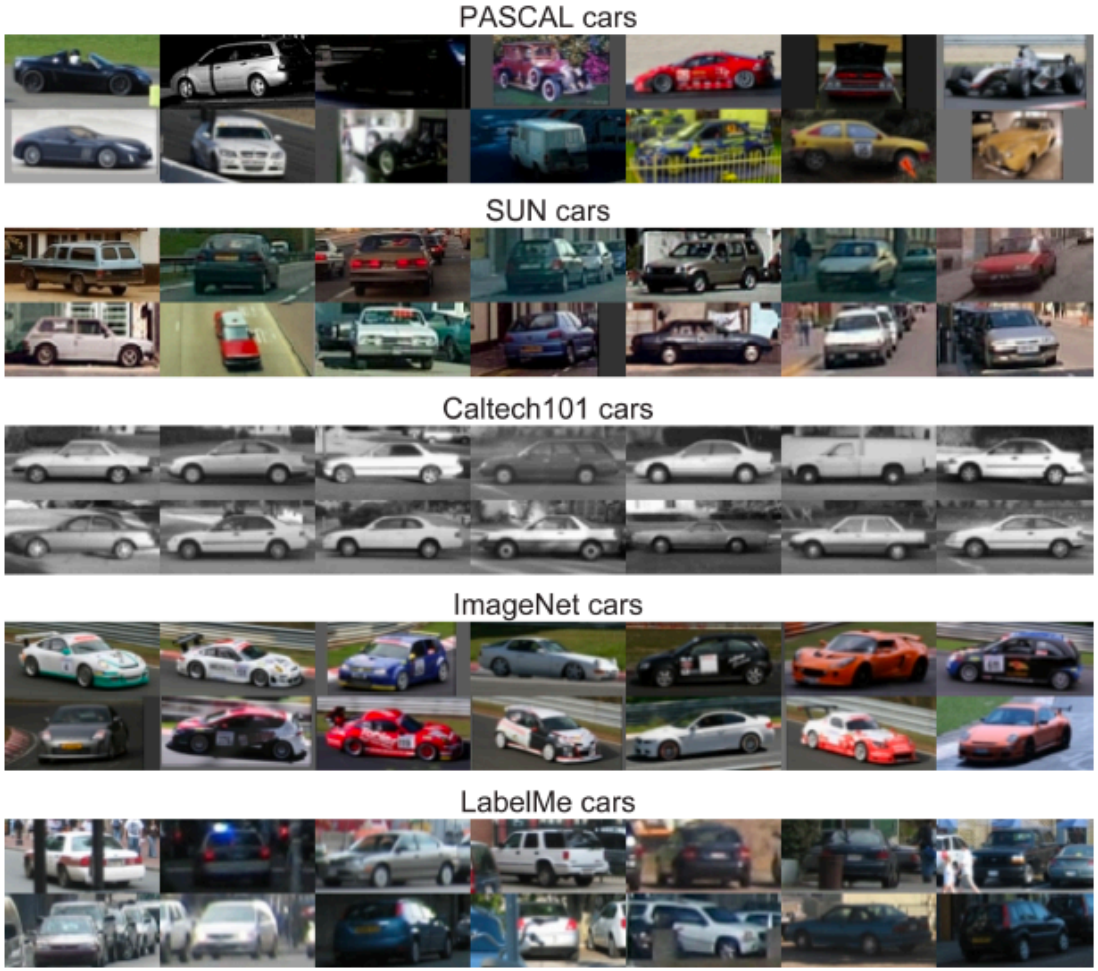


Figure 4. Most discriminative cars from 5 datasets

SOURCES OF DATASET BIAS

1. Selection bias – which images (source)?
2. Capture bias – photographers' habits, styles
3. Category or label bias – painting vs. picture
4. Negative set bias – what will the classifier classify as not a car? [out-of-distribution detection]

How to remedy?

RECENT STUDIES

1. Inclusive Images Competition

2. *Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples* by Eleni Triantafillou et al.

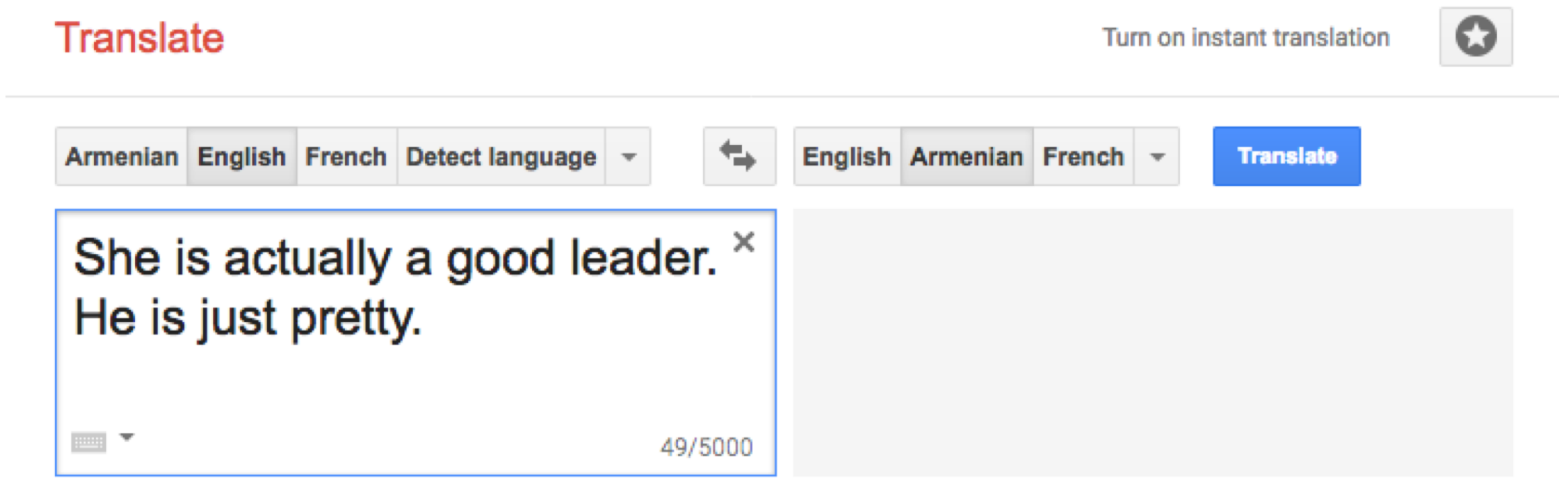
NLP BIAS: A MOTIVATING EXAMPLE

green
PARTY OF CANADA

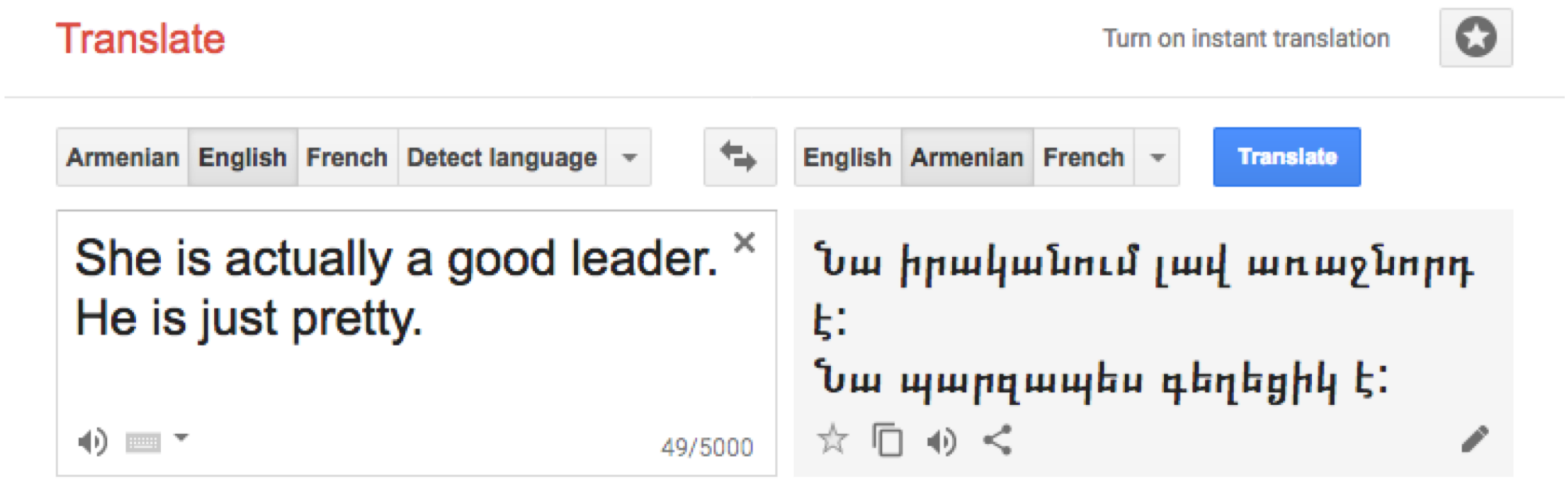
LIBERAL
PARTY OF
CANADA


“She is actually a good leader. He is just pretty.”
#NoPlanetB
LIBERAL PARTY OF CANADA

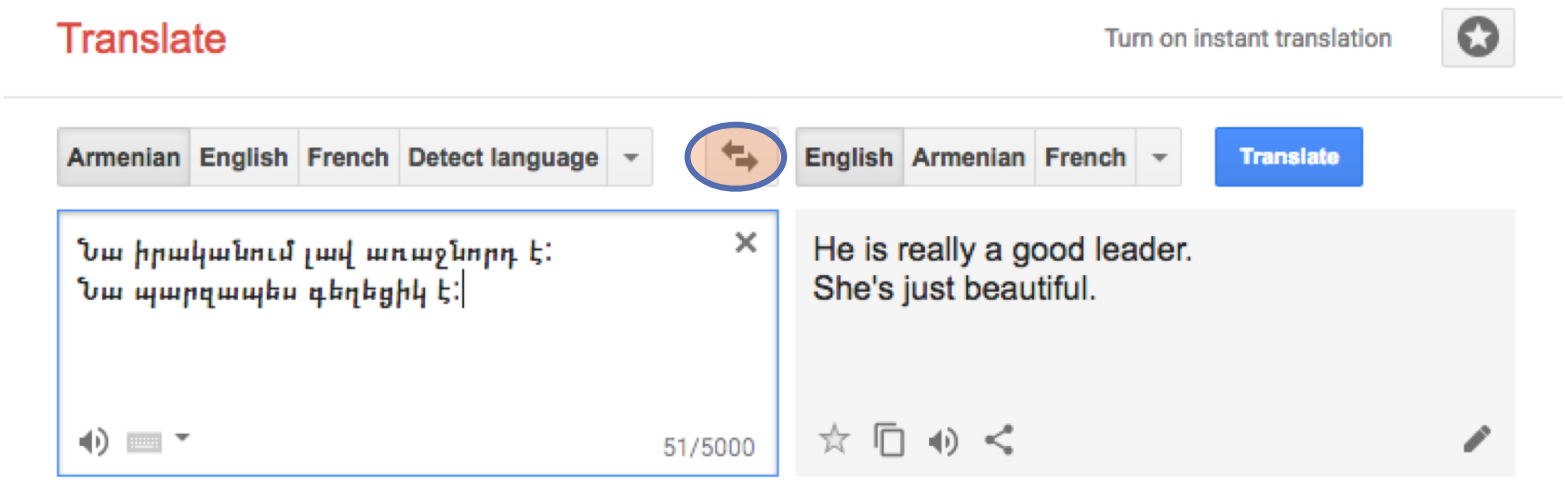
TRANSLATION



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.


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.

WORD CO-OCCURRENCES

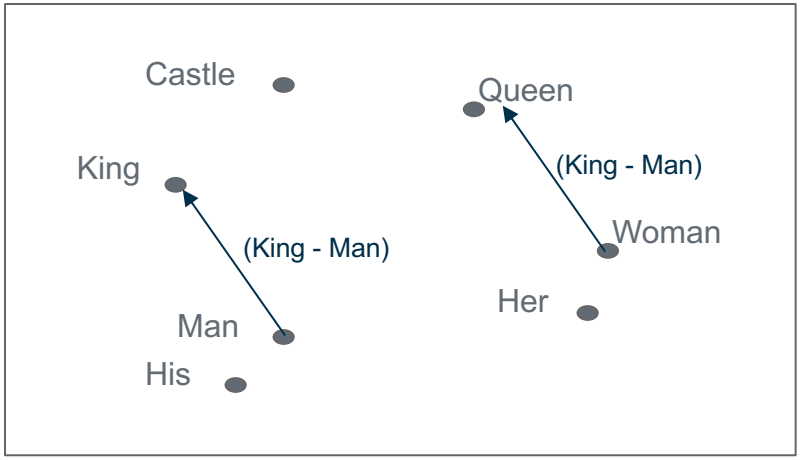
	engineer	**nurse**	**leader**	**pretty**	***(all)***
Ratio of **he:she** co-occurrences	6.25	0.550	9.25	3.07	3.53

The New York Times Annotated Corpus (1987-2007, approx. 1B words, context window: 8)

WORD EMBEDDINGS

What are they?

- A compact vector representation for words
- Learned from a very large corpus of text
- Preserves syntactic and semantic meaning through vector arithmetic (very useful)



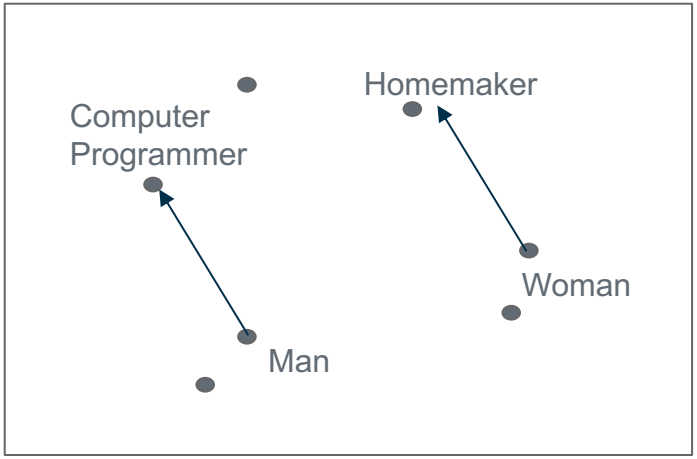
“King” - “Man” + “Woman” ≈ “Queen”

Applications:

- Sentiment analysis
- Document classification / summarization
- Translation
- Temporal semantic trajectories

ANALOGIES

- 🙂 **King : Man :: Queen : Woman**
- 🙂 **Paris : France :: London : England**
- 😮 **Man : Computer_Programmer :: Woman : Homemaker**



Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai (NIPS 2016)

WORD EMBEDDING ASSOCIATION TEST

Implicit Association Test: two words implicitly associated if words can be categorized quicker to their pairing than alternative pairing

WEAT designed as analogous test for word embeddings

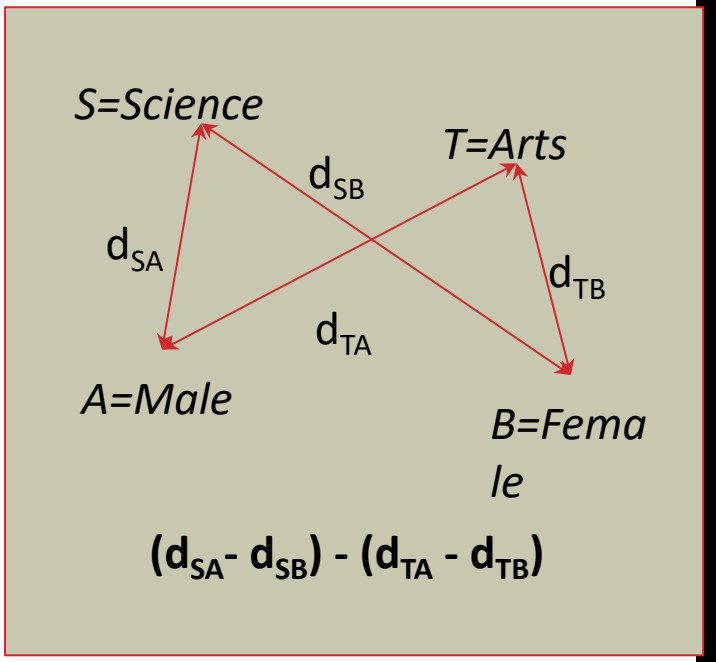
Target Word Sets:
S = {physics, chemistry… } ≈ *Science*
T = {poetry, literature… } ≈ *Arts*

Attribute Word Sets:
A = {he, him, man… } ≈ *Male*
B = {she, her, woman} ≈ *Female*

Measures relative association between four concepts

$$f(w, A, B) = \underset{a \in A}{\text{mean}} cos(\vec{w}, \vec{a}) - \underset{b \in B}{\text{mean}} cos(\vec{w}, \vec{b})$$

Effect Size = $$\frac{\underset{s \in S}{\text{mean}} f(s, A, B) - \underset{t \in T}{\text{mean}} f(t, A, B)}{\underset{w \in S \cup T}{\text{std-dev}} f(w, A, B)}$$



Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan (Science 2017)

MEASURING BIAS

Target words	Attrib. words	Original Finding				Our Finding			
		Ref	N	d	p	N_T	N_A	d	p
Flowers vs insects	Pleasant vs unpleasant	(5)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(5)	32	1.66	10^{-10}	25×2	25×2	1.53	10^{-7}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(5)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (5)	(7)	Not applicable			16×2	25×2	1.50	10^{-4}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (9)	(7)	Not applicable			16×2	8 × 2	1.28	10^{-3}
Male vs female names	Career vs family	(9)	39k	0.72	$< 10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs arts	Male vs female terms	(9)	28k	0.82	$< 10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs arts	Male vs female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs old people's names	Pleasant vs unpleasant	(9)	43k	1.42	$< 10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Science: "Semantics derived automatically from language corpora contain human-like biases"

WEAT INHERENTLY FLAWED

1. *What causes the bias – data, model, noise?*
2. *Is WEAT a good test for word associations?*
3. *Can word embeddings be debiased by subtracting projections onto 'bias subspace'?*

Questions addressed in excellent recent paper:

Understanding undesirable word embedding associations, Ethayarajh, Duvenaud, Hirst (ACL 2019)

Shows that WEAT has theoretical flaws – if word pairs do not occur with equal frequency in the dataset then the bias is severely over-estimated

Propose a simple alternative – define bias axis based on first principal component of differences between word pairs (man – woman, male – female); project each word onto it to estimate degree of bias

DISCUSSION

1. *What are the various notions of bias discussed today?*

2. *What are the sources of the biases?*