

# An Investigation of Representation & Allocation Harms in Contrastive Learning

Subha Maity<sup>1</sup>, Mayank Agarwal<sup>2</sup>, Mikhail Yurochkin<sup>2</sup>,  
& Yuekai Sun<sup>1</sup>

<sup>1</sup>Department of Statistics, University of Michigan

<sup>2</sup>IBM Research, MIT-IBM Watson AI Lab

# Representation harm due to underrepresentation

Representation: contrastive learning (CL) (Chen et al., 2020)

» Controlled study on CIFAR10 dataset (Krizhevsky et al., 2009)

» Q: Biases/harms from underrepresentation of “automobile” class in CL?

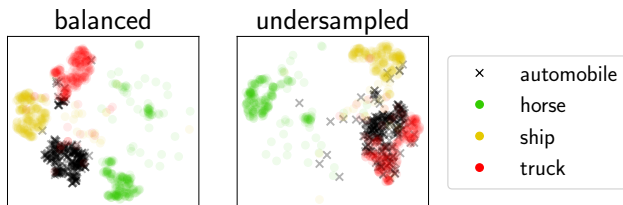
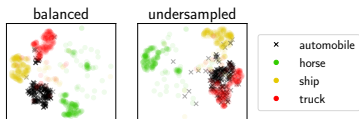


Fig: Two-dimensional t-SNE embeddings

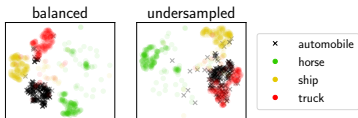
**Stereotyping:** a simplified representation of a (heterogeneous) group, sometimes causing errors in judgment (Bordalo et al., 2016)

# Stereotyping due to underrepresentation



$$\text{RH}(y, y') = \frac{\sum_{i,j} \text{cos-dist}(\varphi_{y,i}^{(\text{under})}, \varphi_{y',j}^{(\text{under})})}{\sum_{i,j} \text{cos-dist}(\varphi_{y,i}^{(\text{bal})}, \varphi_{y',j}^{(\text{bal})})} (< 1?)$$

# Stereotyping due to underrepresentation



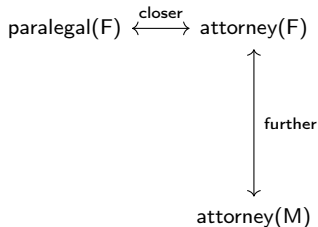
$$RH(y, y') = \frac{\sum_{i,j} \text{cos-dist}(\varphi_{y,i}^{(\text{under})}, \varphi_{y',j}^{(\text{under})})}{\sum_{i,j} \text{cos-dist}(\varphi_{y,i}^{(\text{bal})}, \varphi_{y',j}^{(\text{bal})})} (< 1?)$$

vehicles with mostly blue (water/sky) background	airplane	1.077 ±0.013	0.94 ±0.01	0.974 ±0.01	0.962 ±0.008	0.973 ±0.007	0.975 ±0.007
	ship	0.887 ±0.008	1.154 ±0.013	0.945 ±0.008	0.952 ±0.011	1.004 ±0.006	1.007 ±0.007
vehicles with mostly road/land background	automobile	0.935 ±0.006	0.914 ±0.011	1.041 ±0.01	0.78 ±0.009	1.012 ±0.008	1.008 ±0.009
	truck	0.95 ±0.013	0.916 ±0.013	0.832 ±0.012	1.001 ±0.016	1.001 ±0.008	0.982 ±0.009
animals with mostly green background	deer	0.996 ±0.008	1.0 ±0.008	1.01 ±0.009	1.006 ±0.006	1.126 ±0.012	0.953 ±0.009
	horse	0.99 ±0.008	1.001 ±0.009	1.01 ±0.01	0.993 ±0.009	0.852 ±0.008	1.068 ±0.009
	airplane		ship	automobile	truck	deer	horse

**Key obs:** undersampling leads to harms of representation among similar classes!

# Stereotyping in text representation

Bias in Bios(De-Arteaga et al., 2019): “attorney” (~ 38% female) vs. “paralegal” (~ 85% female).



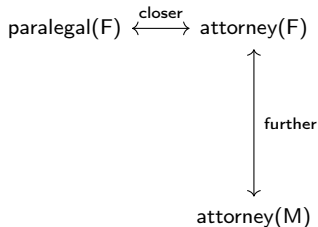
$$RH_F = \frac{\text{cos-dist}\{\text{attorney(F)}, \text{paralegal(F)}\}}{\text{cos-dist}\{\text{attorney(F)}, \text{attorney(M)}\}}$$

professor (F)	0.96	1.00	1.01	0.98	0.98	1.00	1.00	1.02	0.98	1.02
attorney (F)	1.02	0.95	0.94	1.00	0.98	1.01	1.02	1.03	1.02	1.03
paralegal (M)	1.03	0.98	0.94	1.02	0.99	1.00	1.00	1.01	1.04	1.02
surgeon (F)	1.03	1.03	1.02	0.96	0.95	1.00	1.03	1.04	1.02	1.03
dentist (F)	1.09	1.06	1.05	1.00	0.92	1.04	1.05	1.06	1.05	1.05
physician (M)	1.02	1.02	1.00	0.96	0.97	0.95	1.02	1.02	1.02	1.02
dj (F)	1.18	1.18	1.17	1.15	1.12	1.18	0.95	0.99	1.09	1.09
rapper (F)	1.23	1.22	1.22	1.19	1.16	1.21	1.01	0.95	1.18	1.12
composer (F)	1.10	1.12	1.12	1.09	1.06	1.11	1.03	1.10	0.94	1.10
model (M)	1.03	1.02	1.00	1.01	0.99	1.00	0.95	0.97	1.00	0.96

» Theoretical analysis of stereotyping in our paper.

# Stereotyping in text representation

Bias in Bios(De-Arteaga et al., 2019): “attorney” (~ 38% female) vs. “paralegal” (~ 85% female).



$$RH_F = \frac{\text{cos-dist}\{\text{attorney(F)}, \text{paralegal(F)}\}}{\text{cos-dist}\{\text{attorney(F)}, \text{attorney(M)}\}}$$

professor (F)	0.96	1.00	1.01	0.98	0.98	1.00	1.00	1.02	0.98	1.02
attorney (F)	1.02	0.95	0.94	1.00	0.98	1.01	1.02	1.03	1.02	1.03
paralegal (M)	1.03	0.98	0.94	1.02	0.99	1.00	1.00	1.01	1.04	1.02
surgeon (F)	1.03	1.03	1.02	0.96	0.95	1.00	1.03	1.04	1.02	1.03
dentist (F)	1.09	1.06	1.05	1.00	0.92	1.04	1.05	1.06	1.05	1.05
physician (M)	1.02	1.02	1.00	0.96	0.97	0.95	1.02	1.02	1.02	1.02
dj (F)	1.18	1.18	1.17	1.15	1.12	1.18	0.95	0.99	1.09	1.09
rapper (F)	1.23	1.22	1.22	1.19	1.16	1.21	1.01	0.95	1.18	1.12
composer (F)	1.10	1.12	1.12	1.09	1.06	1.11	1.03	1.10	0.94	1.10
model (M)	1.03	1.02	1.00	1.01	0.99	1.00	0.95	0.97	1.00	0.96

» Theoretical analysis of stereotyping in our paper.

# Stereotyping leads to harms of allocation

Q: Does stereotyping lead to an increase in misclassification rate? Fit a linear classifier with CL embedding.

$$\text{Metric: } AH(y, y') = P\{\hat{y}^{(y)} = y' \mid y\} - P\{\hat{y}^* = y' \mid y\}$$

airplane	-0.091 ±0.001	0.032 ±0.001	0.001 ±0.0	0.009 ±0.0	0.009 ±0.001	0.001 ±0.0
ship	0.056 ±0.001	-0.118 ±0.001	0.016 ±0.0	0.019 ±0.0	0.002 ±0.0	0.001 ±0.0
automobile	0.01 ±0.001	0.015 ±0.001	-0.095 ±0.002	0.062 ±0.001	0.002 ±0.0	0.0 ±0.0
truck	0.017 ±0.0	0.009 ±0.001	0.045 ±0.001	-0.079 ±0.001	0.0 ±0.0	0.001 ±0.0
deer	0.012 ±0.0	0.002 ±0.0	0.001 ±0.0	0.001 ±0.0	-0.155 ±0.002	0.055 ±0.001
horse	0.004 ±0.0	0.001 ±0.0	0.001 ±0.0	0.001 ±0.0	0.056 ±0.001	-0.128 ±0.002
	airplane	ship	automobile	truck	deer	horse

(a) allocation harm (AH)

airplane	1.077 ±0.013	0.94 ±0.01	0.974 ±0.01	0.962 ±0.008	0.973 ±0.007	0.975 ±0.007
ship	0.887 ±0.008	1.154 ±0.013	0.945 ±0.008	0.952 ±0.011	1.004 ±0.006	1.007 ±0.007
automobile	0.935 ±0.006	0.914 ±0.011	1.041 ±0.01	0.78 ±0.009	1.012 ±0.008	1.008 ±0.009
truck	0.95 ±0.013	0.916 ±0.013	0.832 ±0.012	1.001 ±0.016	1.001 ±0.008	0.982 ±0.009
deer	0.996 ±0.008	1.0 ±0.008	1.01 ±0.009	1.006 ±0.006	1.126 ±0.012	0.953 ±0.009
horse	0.99 ±0.008	1.001 ±0.009	1.01 ±0.01	0.993 ±0.009	0.852 ±0.008	1.068 ±0.009
	airplane	ship	automobile	truck	deer	horse

(b) representation harm (RH)

Key obs: stereotyping leads to harms of allocation among similar classes.

# Thank you!!!

**Paper:** An Investigation of Representation and Allocation Harms in Contrastive Learning

**Poster:** Session 7; Friday, May 10th; 10:45am - 12:45pm local time

**Contact info:** Subha Maity; [smaity@umich.edu](mailto:smaity@umich.edu)