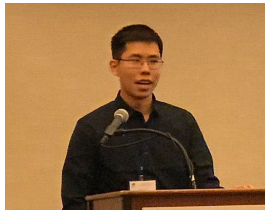
Computer vision meets fairness

Olga Russakovsky

PRINCETON UNIVERSITY


Prem Nair


Kenji Hata


Kyle Genova


Arvind Narayanan


Kate Crawford (NYU/MSR)

NSF

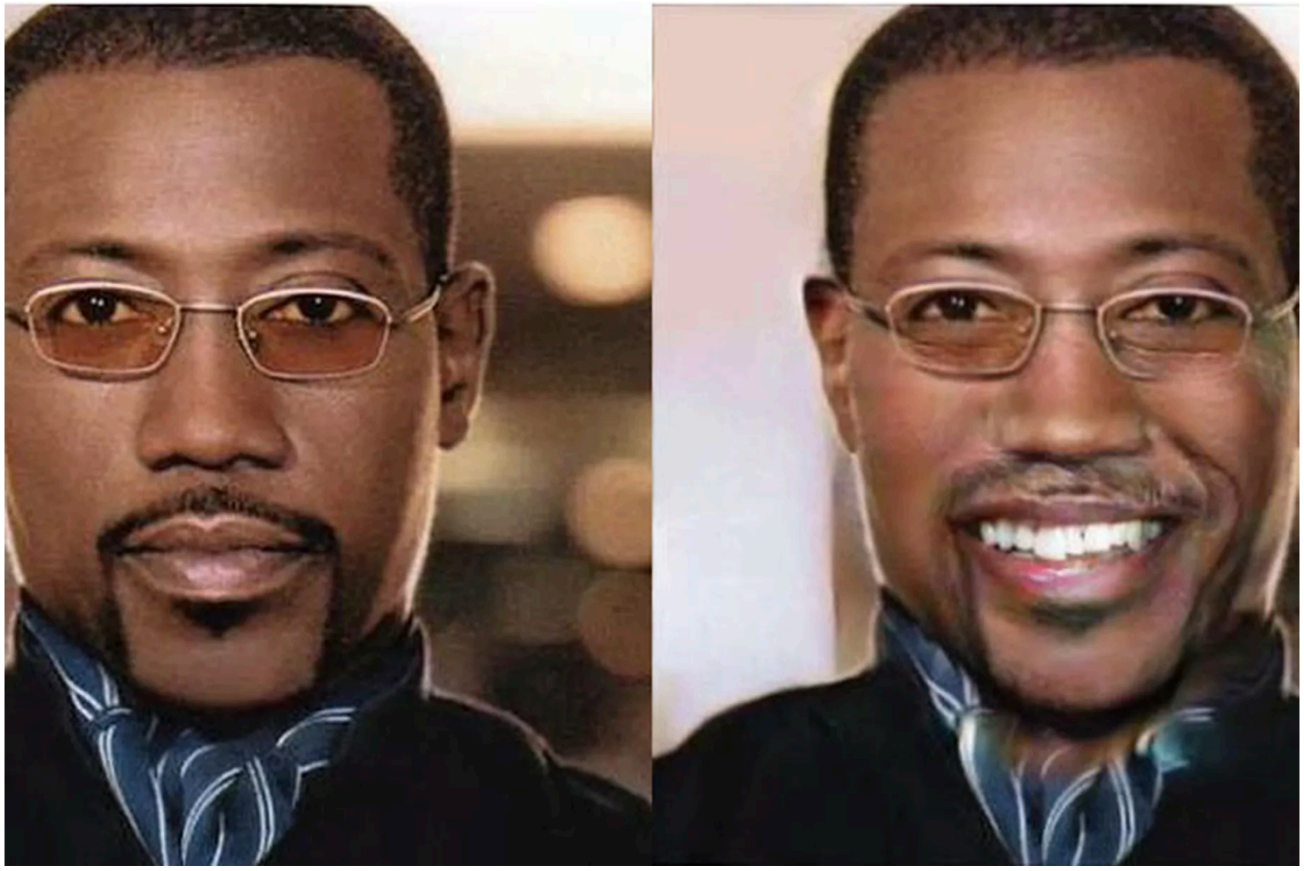
Google Cloud

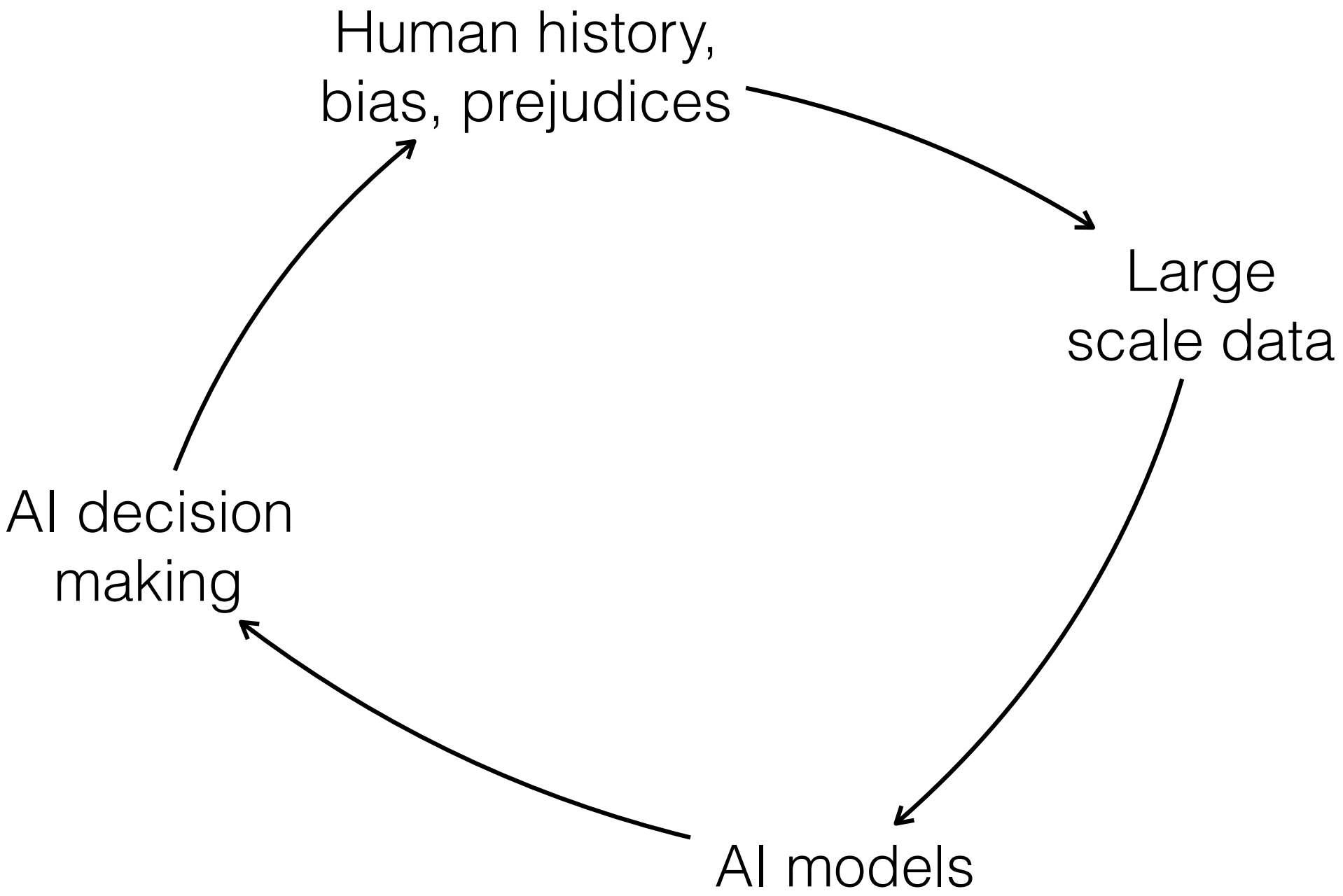
PRINCETON
School of Engineering and Applied Science

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

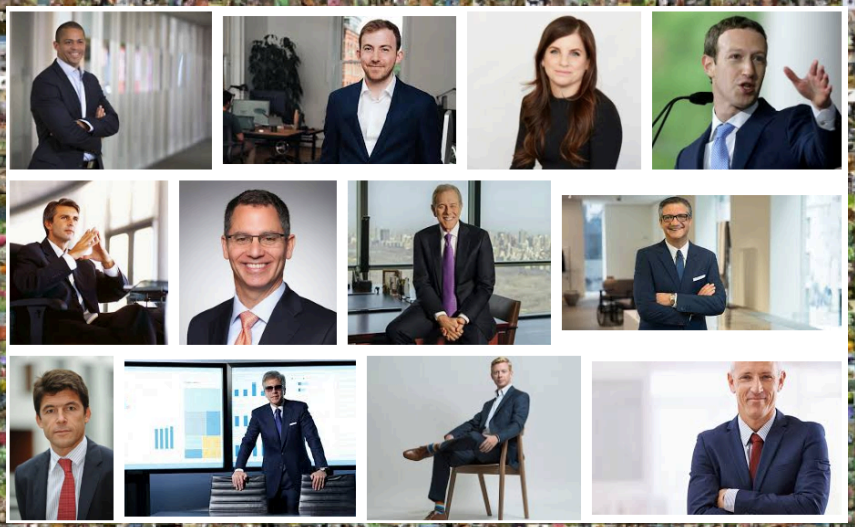
“FaceApp’s creator apologizes for the app’s lightening ‘hot’ filter”

THE VERGE **April 25, 2017**




Human history,
bias, prejudices
Large
scale data
AI models
AI decision
making

Massive scale ≠ fair representation



[No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, D. Sculley. NIPS 2017 Workshop]

[Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. Matthew Kay, Cynthia Matuszek, Sean A. Munson. CHI 2015]

Why should computer vision researchers care?

1) Ethical responsibility
2) Selfish reasons

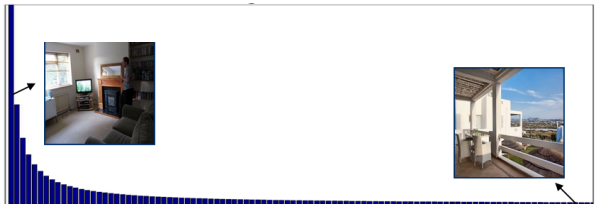
Why should computer vision researchers care?

1) Ethical responsibility

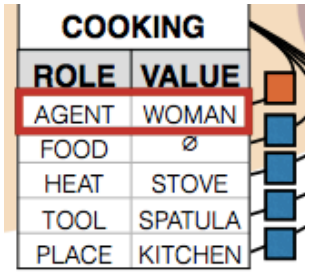
This talk →

2) Selfish reasons

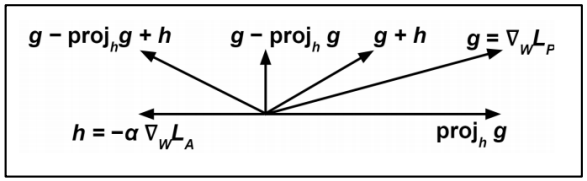
Computer vision meets fairness



1) Learning with long tail distributions

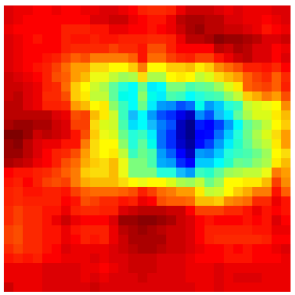


2) Reducing bias amplification



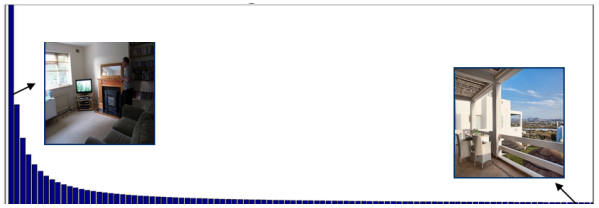
3) Incorporating model constraints

4) Understanding domain adaptation

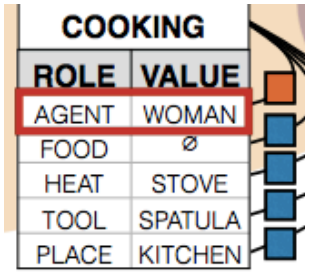


5) Designing interpretable models

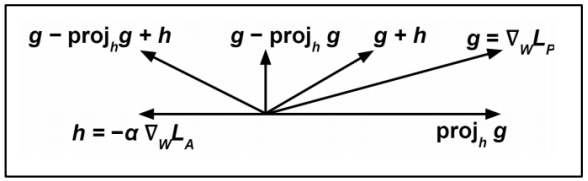
Computer vision meets fairness



1) Learning with long tail distributions

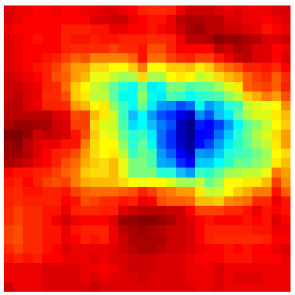


2) Reducing bias amplification



3) Incorporating model constraints

4) Understanding domain adaptation



5) Designing interpretable models

“Facial recognition is accurate, if you’re a white guy”

The New York Times **Feb 9, 2018**

Gender classification (MSFT)

On **lighter** faces
0.7% error

On **darker** faces
12.9% error

[**Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.** Joy Boulamwini, Timnit Gebru. FAT* conference 2018]

Standard face datasets

LFW

[Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller.]

83.5% white

IJB-A

[Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. Brendan F Klare et al. CVPR 2015]

79.6% light skin

Adience

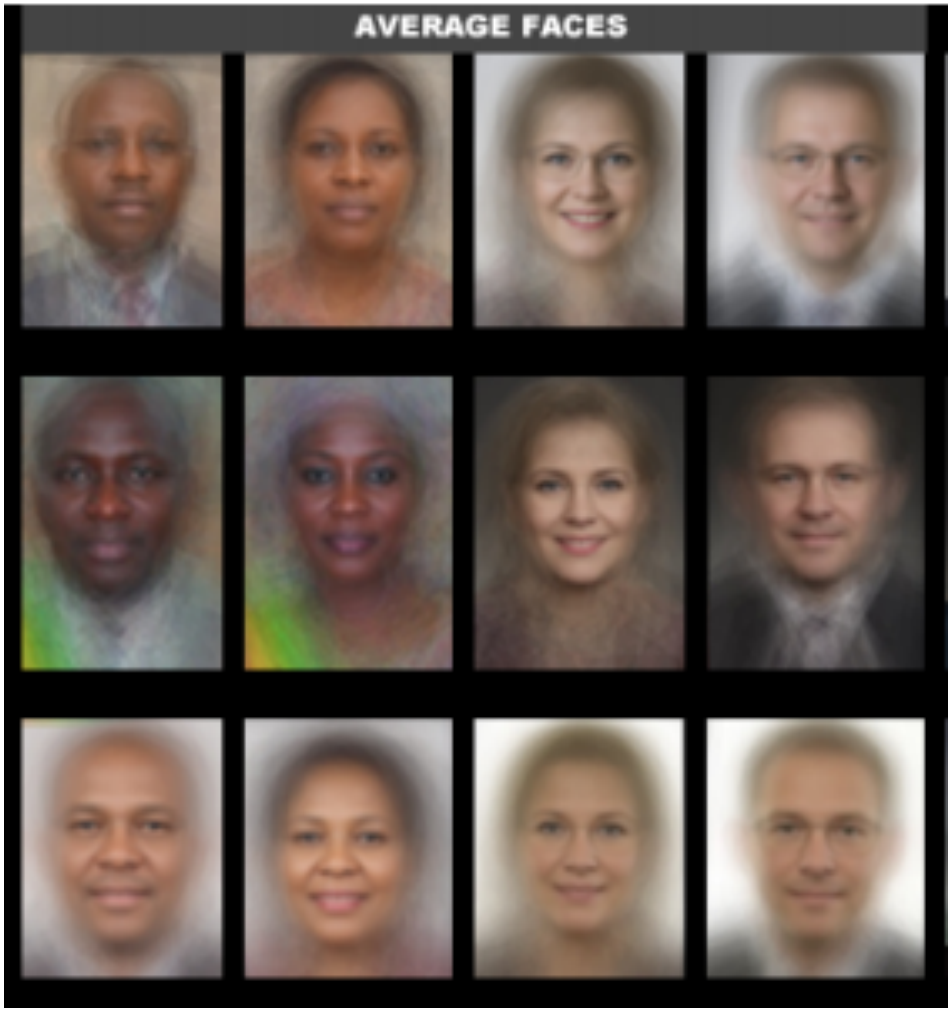
[Age and gender classification using convolutional neural networks. Gil Levi and Tal Hassner. CVPR Workshop 2015.]

86.2% light skin

Statistics from [Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Joy Boulamwini, Timnit Gebru. FAT* conference 2018] [Age, gender and race estimation from unconstrained face images. Hu Han and Anil K Jain. Michigan State Univ, Tech Report 2014.]

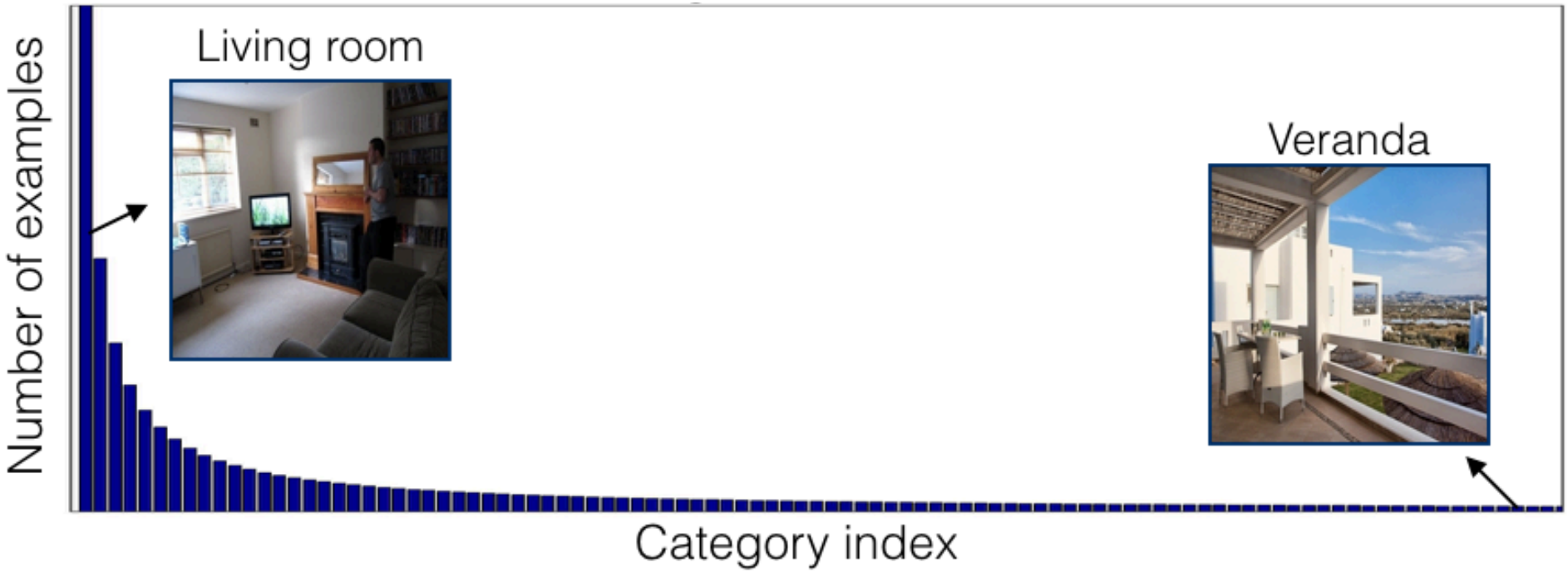
One option: balance the dataset

Pilot Parliaments Benchmark (PPB)

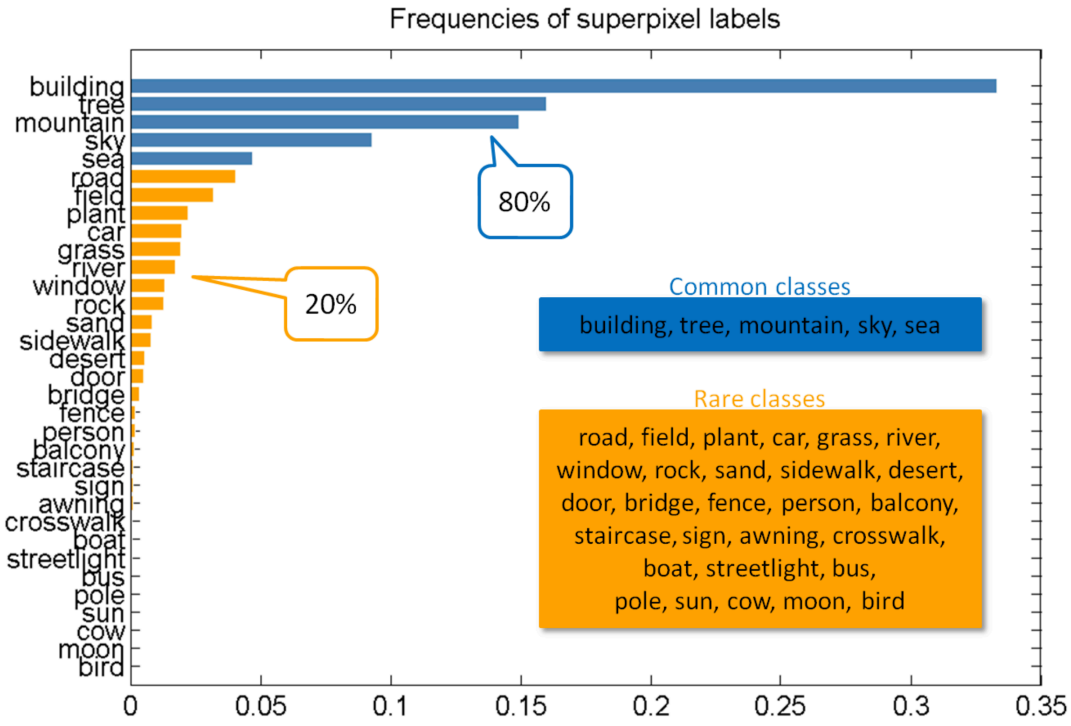


[Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Joy Boulamwini, Timnit Gebru. FAT* conference 2018]

But artificially balancing the data isn't sustainable at scale… or fair

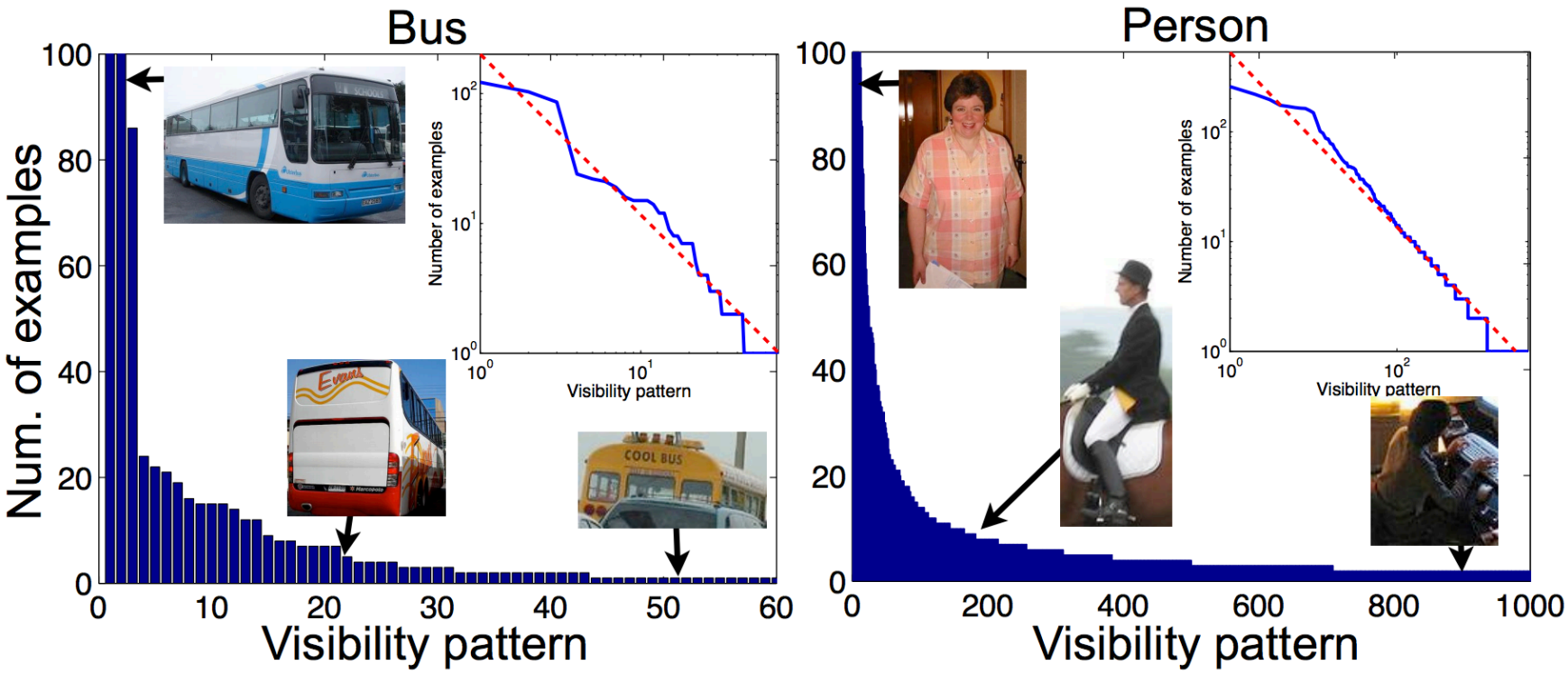


But artificially balancing the data isn't sustainable at scale… or fair



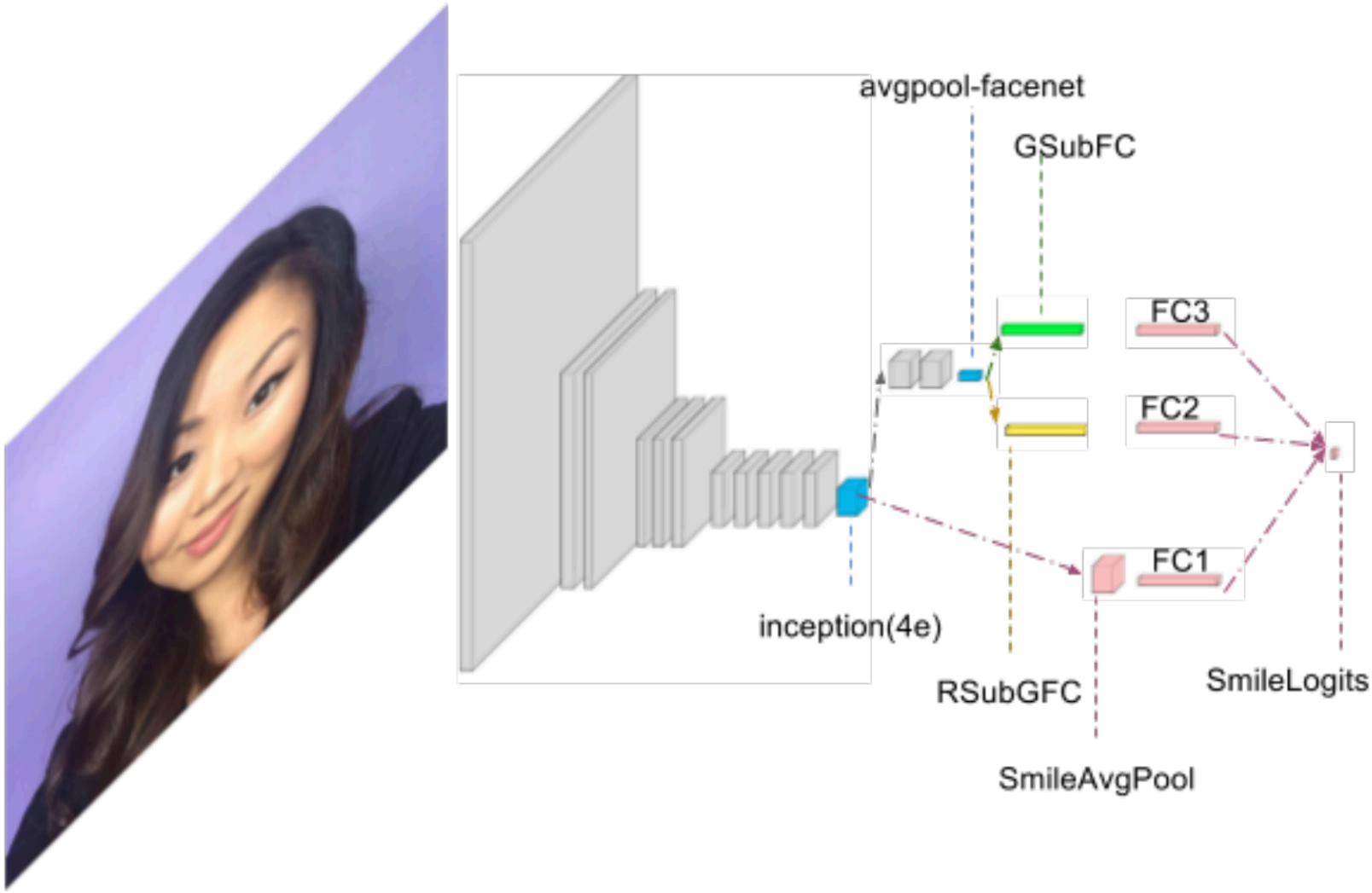
[Context Driven Scene Parsing with Attention to Rare Classes. Yang et al. CVPR'14]

But artificially balancing the data isn't sustainable at scale… or fair



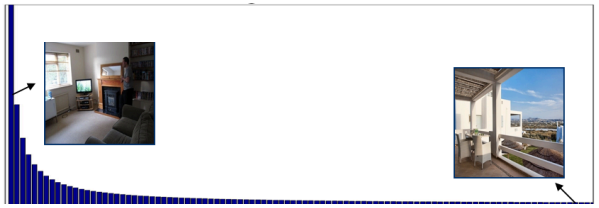
[Capturing long-tail distributions of object subcategories. Zhu, Anguelov, Ramanan. CVPR'14]

“Improving smiling detection with race and gender diversity”

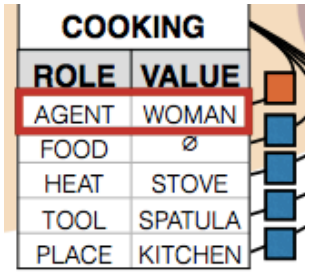


[**Improving Smiling Detection with Race and Gender Diversity.** Hee Jung Ryu, Margaret Mitchell, Hartwig Adam. https://arxiv.org/abs/1712.00193]

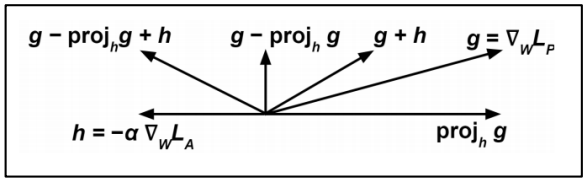
Computer vision meets fairness



1) Learning with long tail distributions

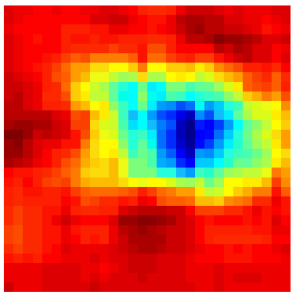


2) Reducing bias amplification



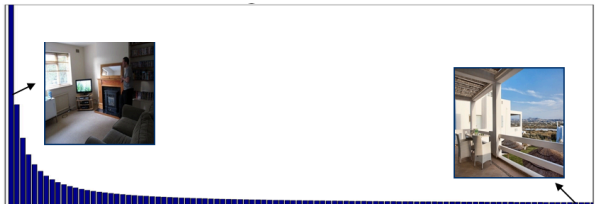
3) Incorporating model constraints

4) Understanding domain adaptation

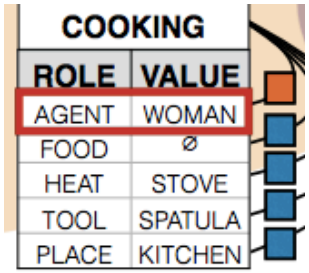


5) Designing interpretable models

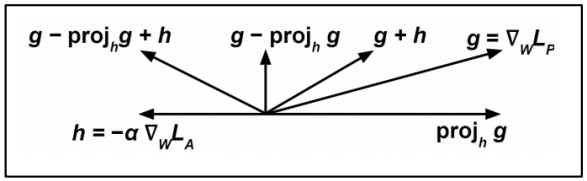
Computer vision meets fairness



1) Learning with long tail distributions

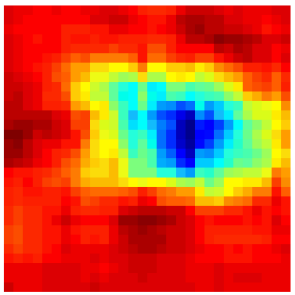


2) Reducing bias amplification



3) Incorporating model constraints

4) Understanding domain adaptation



5) Designing interpretable models

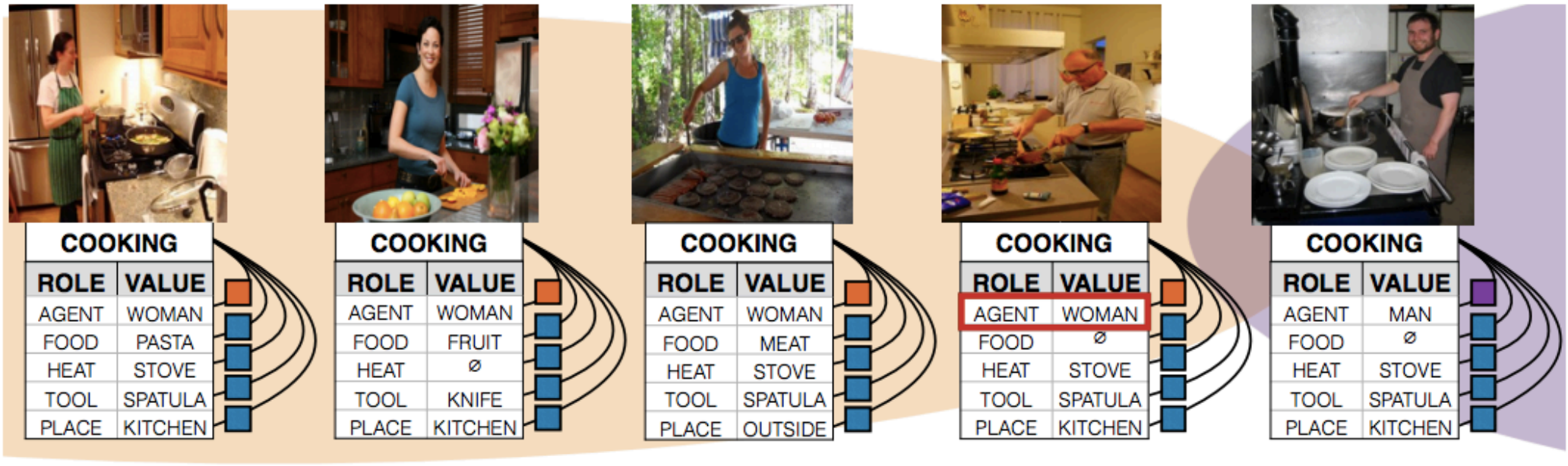
AI models may amplify social bias

Word embeddings (w2vNEWS)

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

[Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Tolga Bolukbasi, Kai-Wei Chang, James Zou , Venkatesh Saligrama, Adam Kalai. NPS 2016] [Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science 14 Apr 2017]

AI models may amplify social bias



[Men also like shopping: Reducing Gender Bias Amplification using Corpus-level constraints. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. Best Long Paper Award at EMNLP 2017]

Translating fairness constraints into models

We speak to the concept of *mitigating bias* using the known term *debiasing*[1], following definitions provided by Hardt et al. (2016) and refined by Beutel et al. (2017).

Definition 1. DEMOGRAPHIC PARITY. A predictor $\hat{Y}$ satisfies *demographic parity* if $\hat{Y}$ and Z are independent.

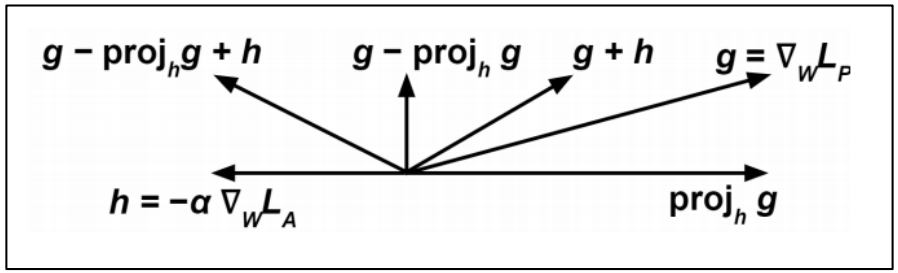
This means that $P(\hat{Y} = \hat{y})$ is equal for all values of the protected variable Z: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$.

Definition 2. EQUALITY OF ODDS. A predictor $\hat{Y}$ satisfies *equality of odds* if $\hat{Y}$ and Z are conditionally independent given Y.

This means that, for all possible values of the true label Y, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

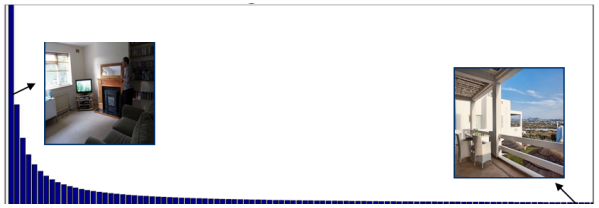
Definition 3. EQUALITY OF OPPORTUNITY. If the output variable Y is discrete, a predictor $\hat{Y}$ satisfies *equality of opportunity* with respect to a class y if $\hat{Y}$ and Z are independent conditioned on $Y = y$.

This means that, for a *particular* value of the true label Y, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

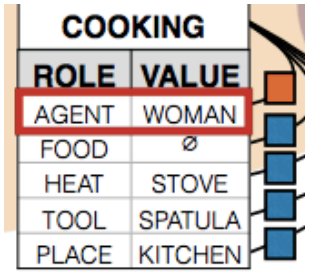


[Mitigating Unwanted Biases with Adversarial Learning. Brian Hu Zhang, Blake Lemoine, Margaret Mitchell. AIES 2018]

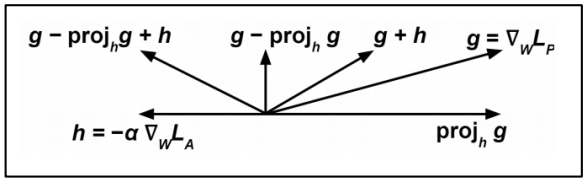
Computer vision meets fairness



1) Learning with long tail distributions

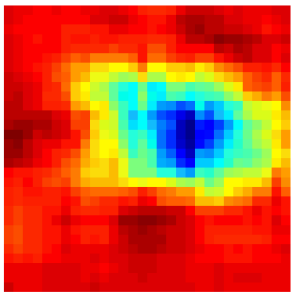


2) Reducing bias amplification



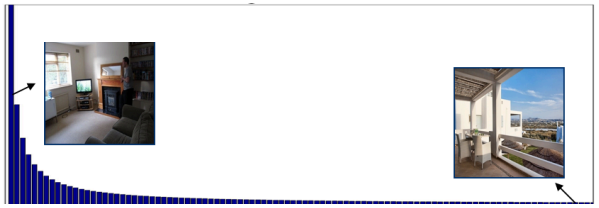
3) Incorporating model constraints

4) Understanding domain adaptation

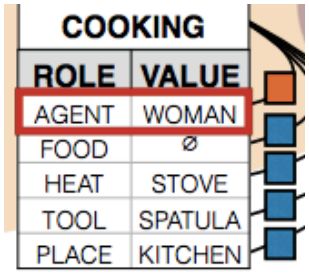


5) Designing interpretable models

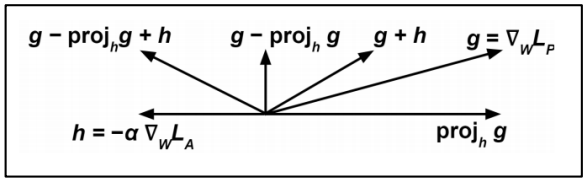
Computer vision meets fairness



1) Learning with long tail distributions

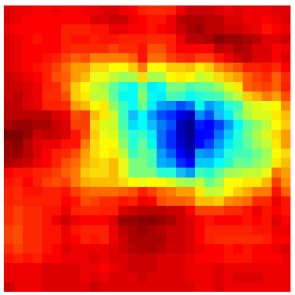


2) Reducing bias amplification



3) Incorporating model constraints

4) Understanding domain adaptation



5) Designing interpretable models

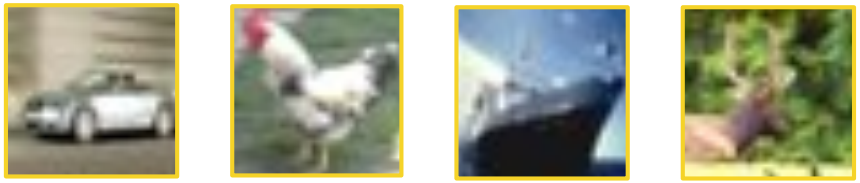
Recognizing “cooking”

Classifying “cooking” vs “driving”

Classifying “cooking” vs “driving”

A toy illustration on CIFAR-10

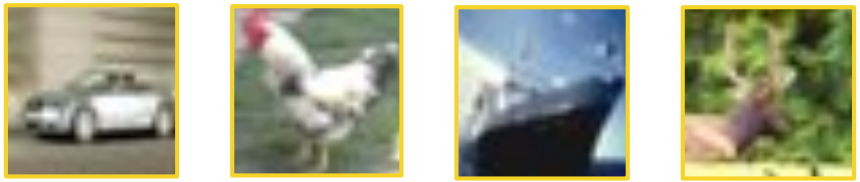
Goal: classifying color images into one of 10 object classes



[Prem Nair, Kenji Hata, Kyle Genova, Olga Russakovsky. Work in progress.]

A toy illustration on CIFAR-10

Goal: classifying color images into one of 10 object classes

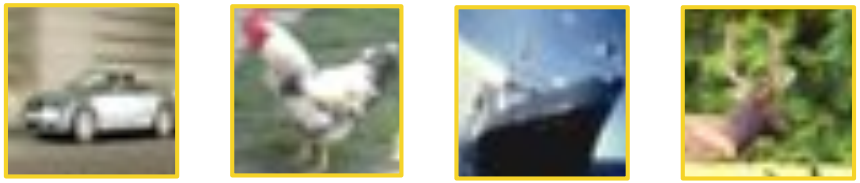


Training option 1:
different domain

[Prem Nair, Kenji Hata, Kyle Genova, Olga Russakovsky. Work in progress.]

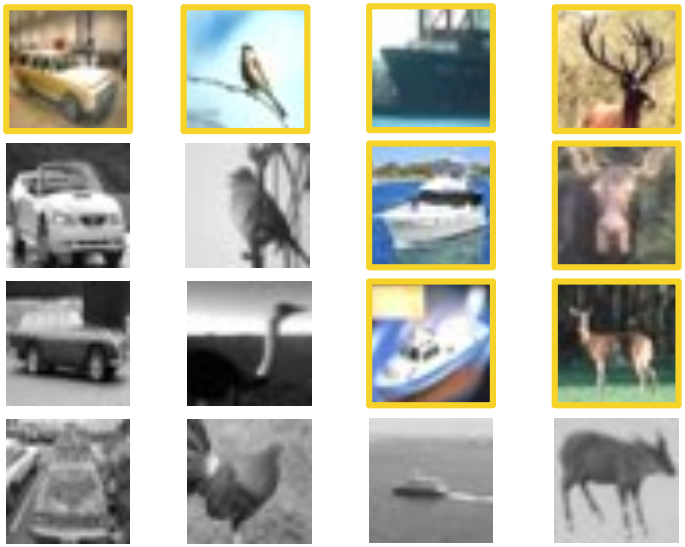
A toy illustration on CIFAR-10

Goal: classifying color images into one of 10 object classes



Training option 1:
different domain

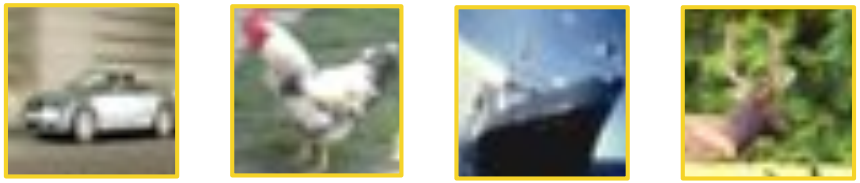
Training option 2: both domains
but skewed distributions



[Prem Nair, Kenji Hata, Kyle Genova, Olga Russakovsky. Work in progress.]

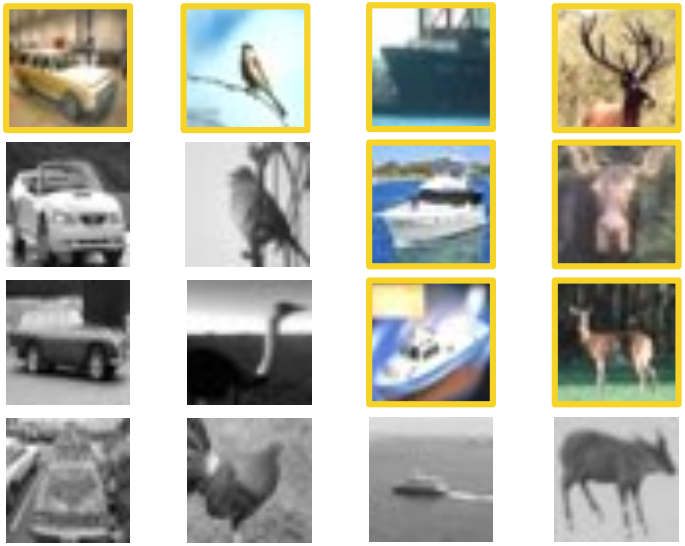
A toy illustration on CIFAR-10

Goal: classifying color images into one of 10 object classes



Training option 1:
different domain

Training option 2: both domains
but skewed distributions



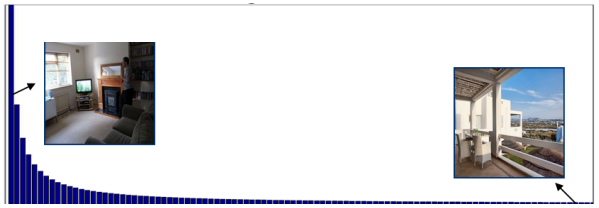
Accuracy on 10-way classification using ResNet-18

91.0%

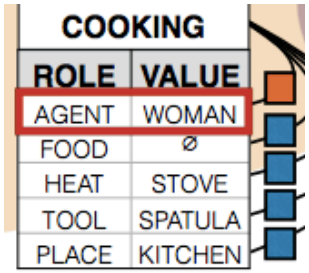
88.1%

[Prem Nair, Kenji Hata, Kyle Genova, Olga Russakovsky. Work in progress.]

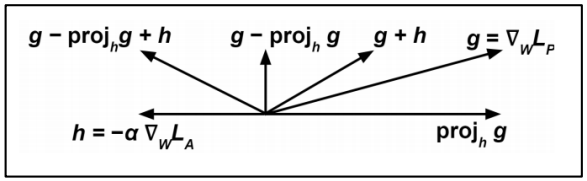
Computer vision meets fairness



1) Learning with long tail distributions

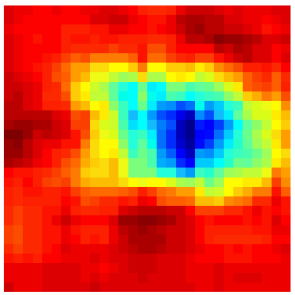


2) Reducing bias amplification



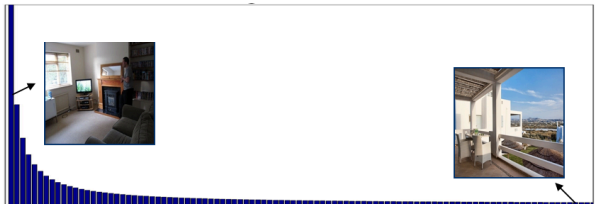
3) Incorporating model constraints

4) Understanding domain adaptation

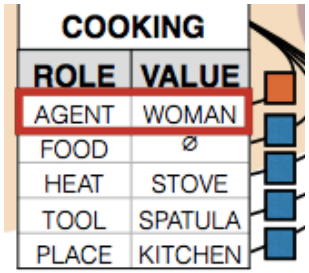


5) Designing interpretable models

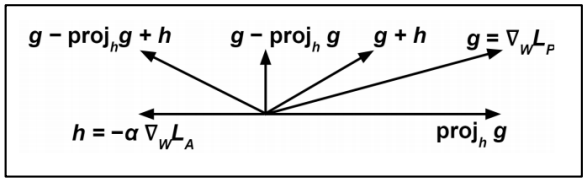
Computer vision meets fairness



1) Learning with long tail distributions

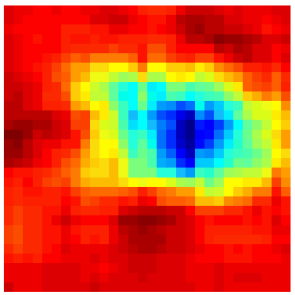


2) Reducing bias amplification



3) Incorporating model constraints

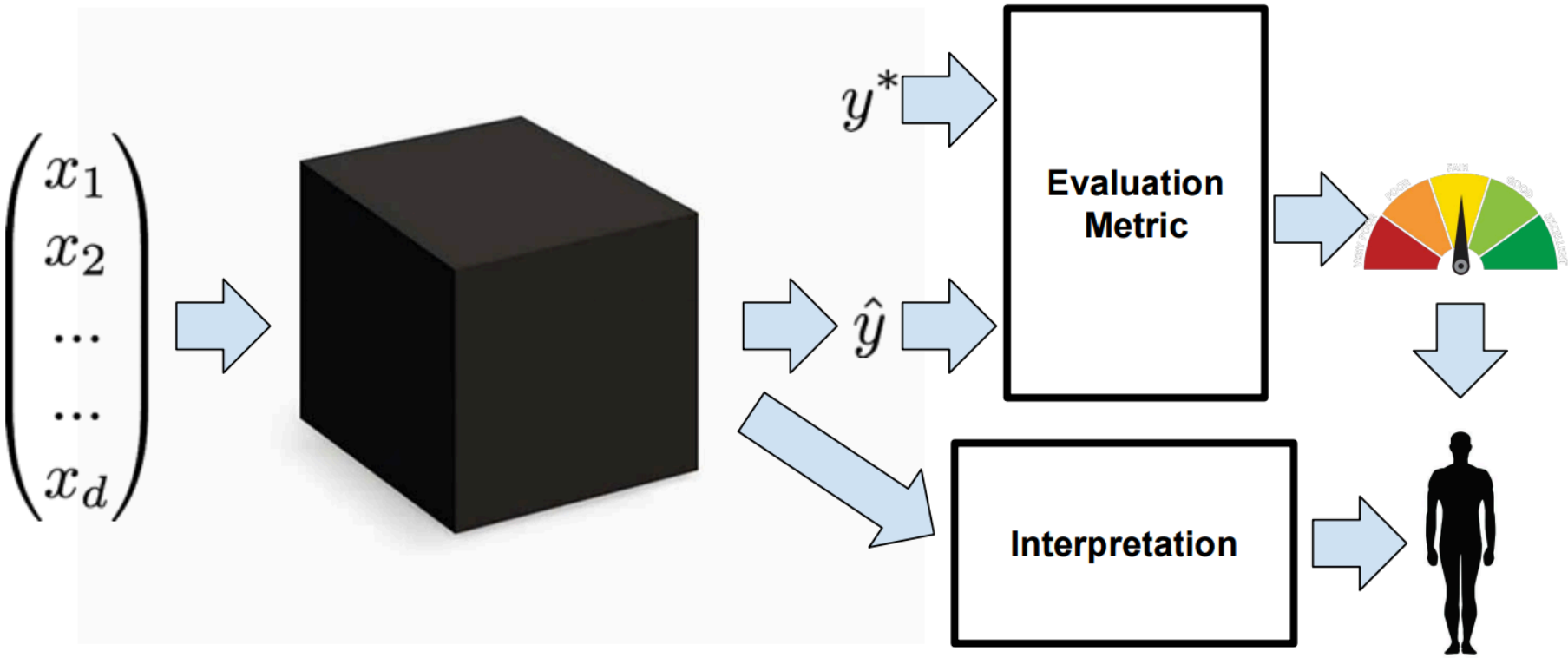
4) Understanding domain adaptation



5) Designing interpretable models

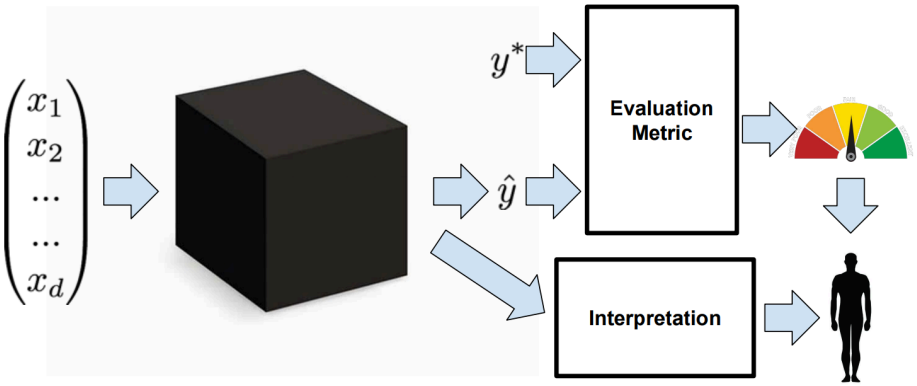
The Mythos of Model Interpretability

Zachary Lipton, ICML Workshop on Human Interpretability in Machine Learning 2016



The Mythos of Model Interpretability

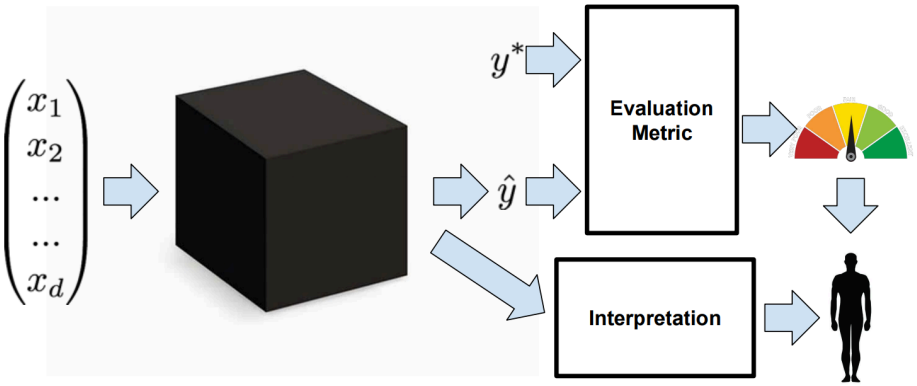
Zachary Lipton, ICML Workshop on Human Interpretability in Machine Learning 2016



- Fair and ethical decision making

The Mythos of Model Interpretability

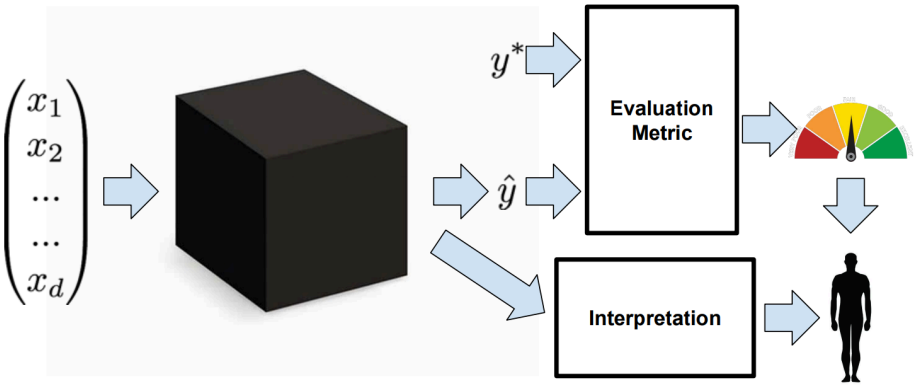
Zachary Lipton, ICML Workshop on Human Interpretability in Machine Learning 2016



- Fair and ethical decision making
- Trust
- Causality
- Transferability
- Informativeness

The Mythos of Model Interpretability

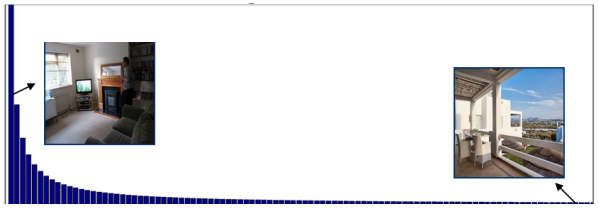
Zachary Lipton, ICML Workshop on Human Interpretability in Machine Learning 2016



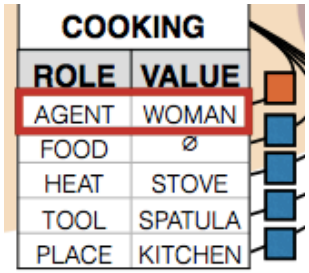
- Fair and ethical decision making
- Trust
- Causality
- Transferability
- Informativeness

- Imposing constraints
- Providing feedback to fix errors
- Enabling wide deployment
- …

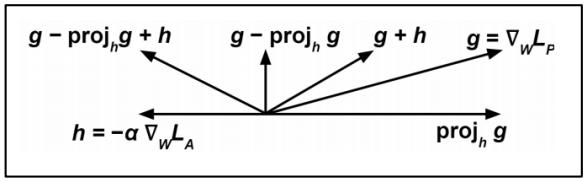
Computer vision meets fairness



1) Learning with long tail distributions

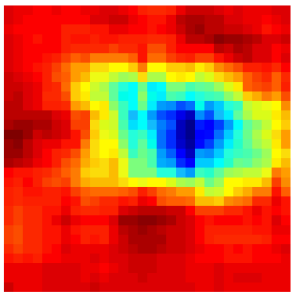


2) Reducing bias amplification



3) Incorporating model constraints

4) Understanding domain adaptation



5) Designing interpretable models

Food for thought

- Pros & cons of benchmark datasets
- “Curated” vs “real-world” datasets
- Types of bias in visual data
- Bias vs prior
- How to detect bias in visual data (chicken-and-egg problem?)
- Bias propagation in models
- Visualization of bias