Fairness in visual recognition

Olga Russakovsky


Vikram Ramaswamy


Angelina Wang


Zeyu Wang


Kaiyu Yang

PRINCETON UNIVERSITY

@VisualAILab
@ai4allorg

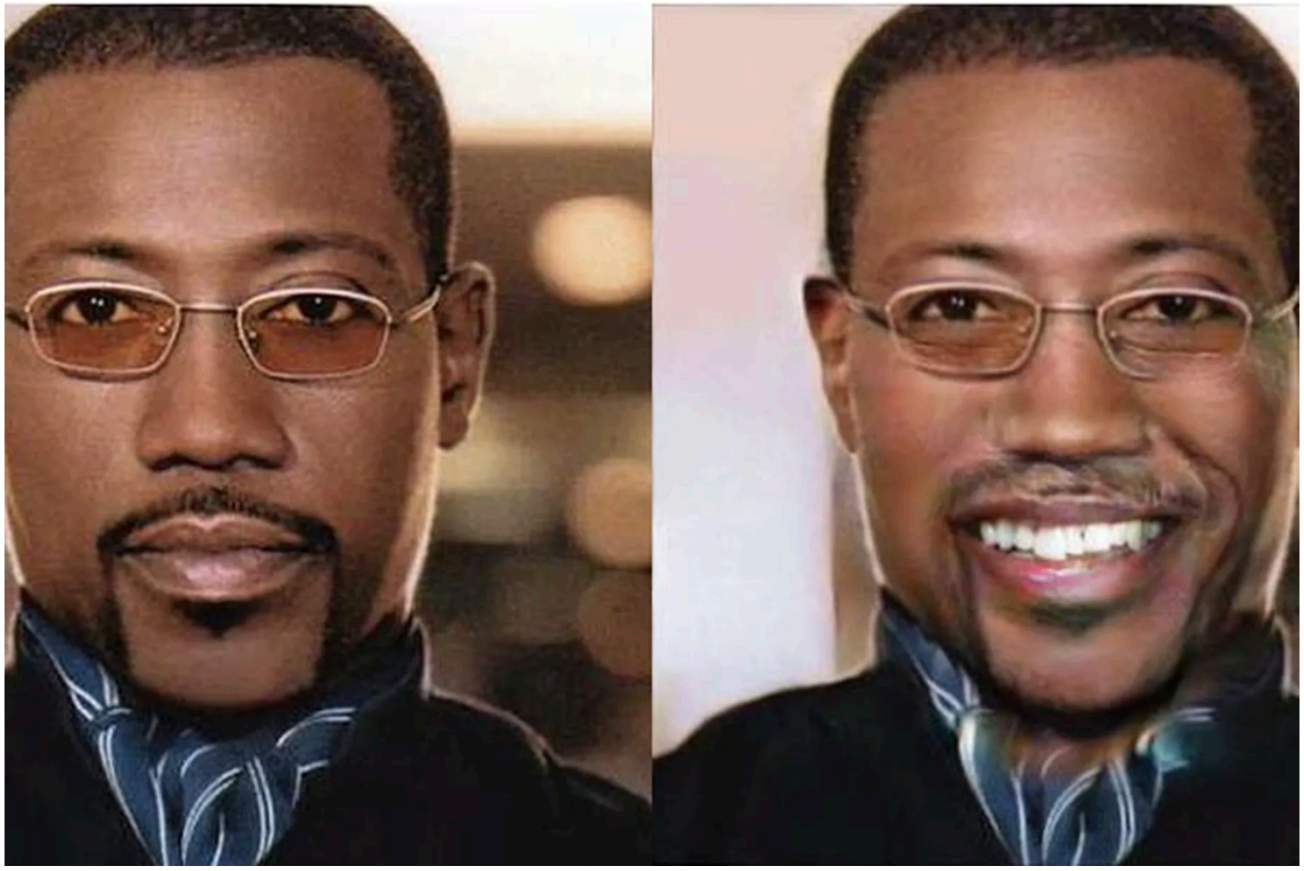
NSF

Google Cloud

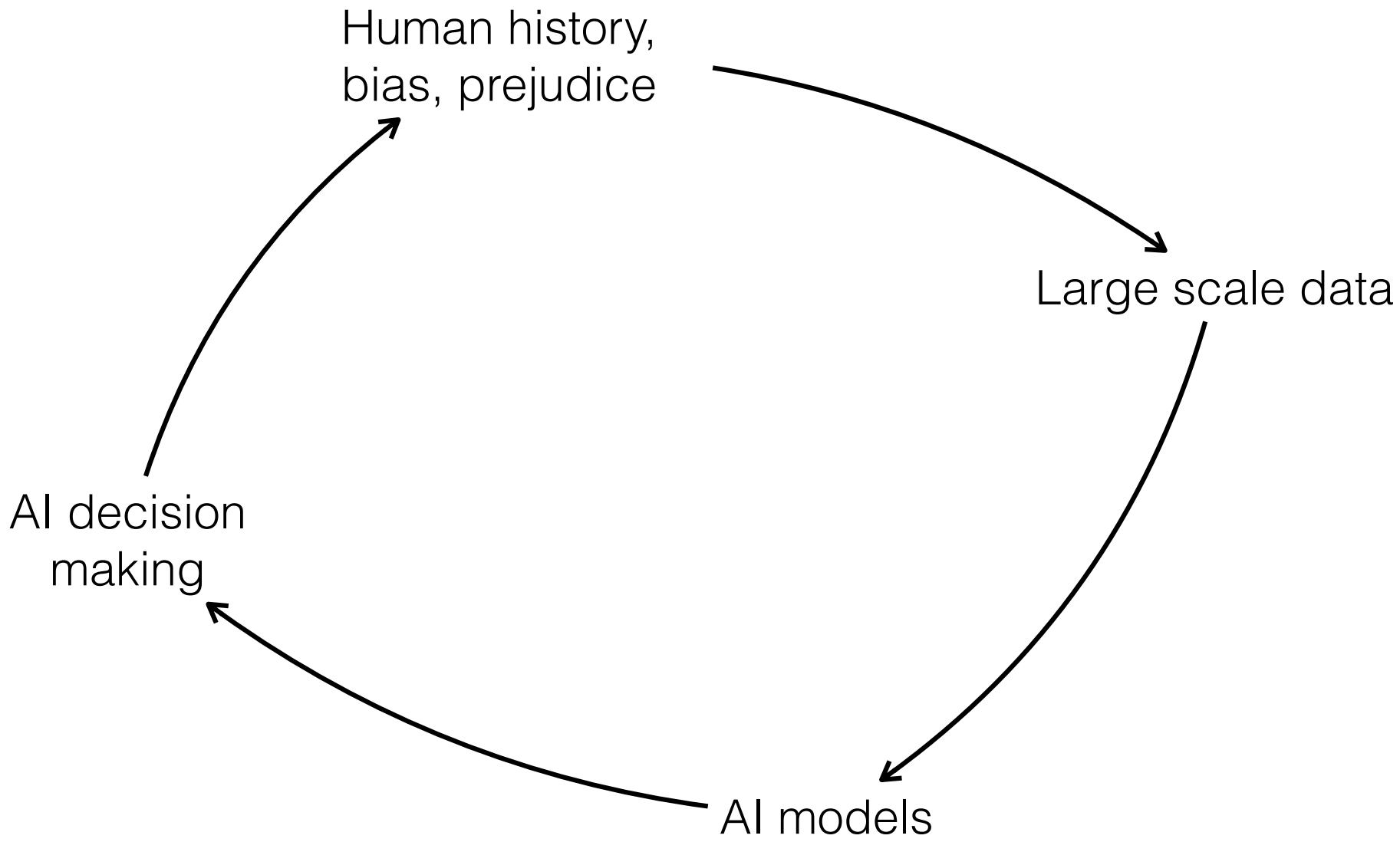
PRINCETON
School of Engineering and Applied Science

Computer vision model learns to “increase attractiveness” by manipulating skin color

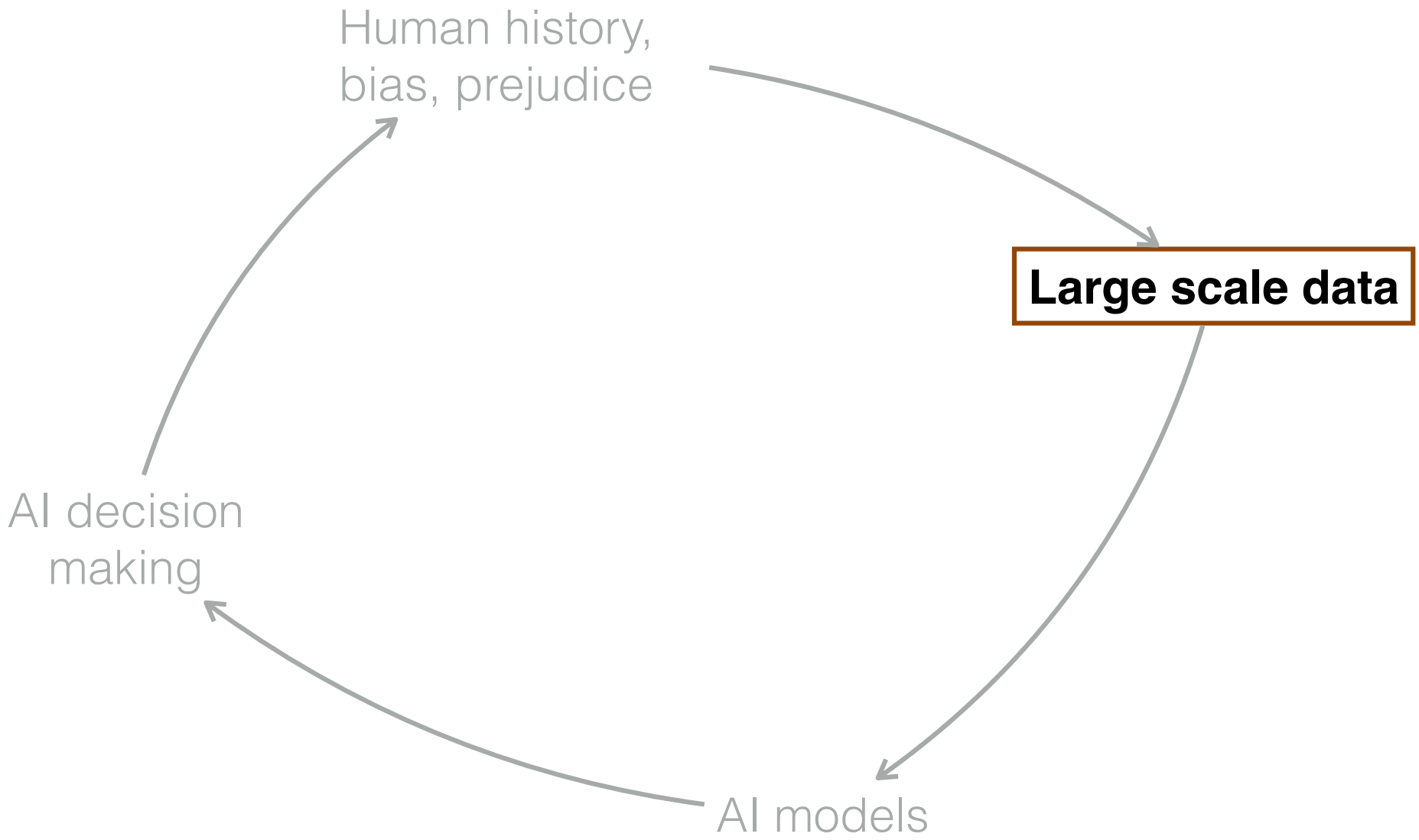
THE VERGE **April 25, 2017**



Can we adjust the AI design to mitigate these effects?



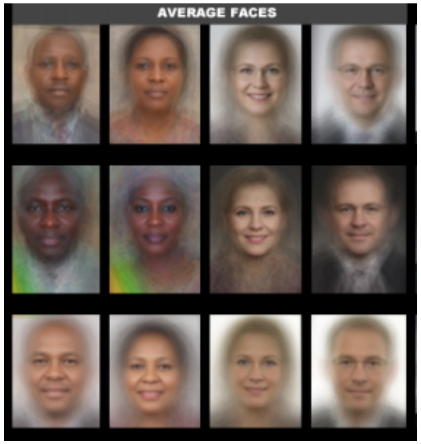
Can we adjust the AI design to mitigate these effects?



Large scale ≠ fair representation

Race diversity in face datasets

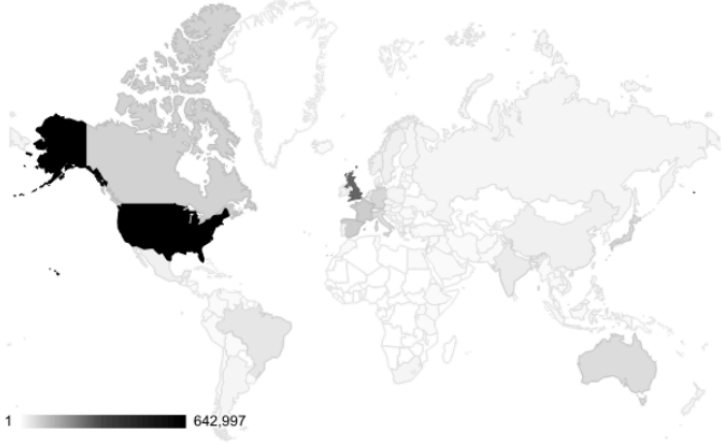
Dataset	Lighter
IJB-A	79.6%
Adience	86.2%



[Joy Buolamwini & Timnit Gebru. FAT* 2018]

Geographic diversity

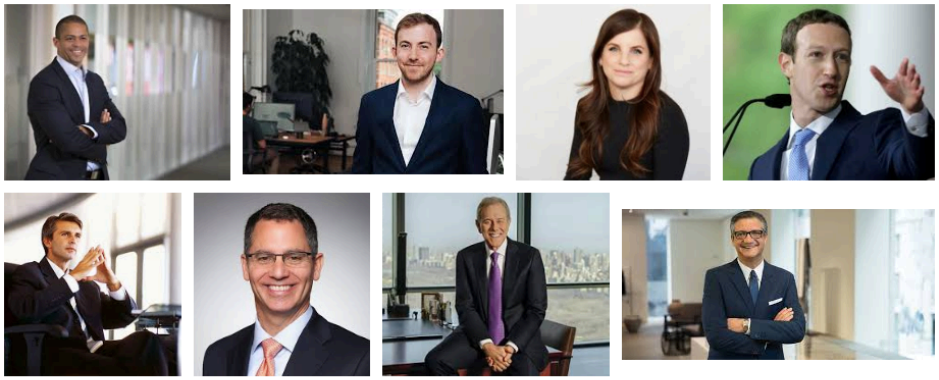
(in ImageNet and OpenImages)



[Shreya Shankar et al. NeurIPS 2017 Workshop]

Diversity in image search results

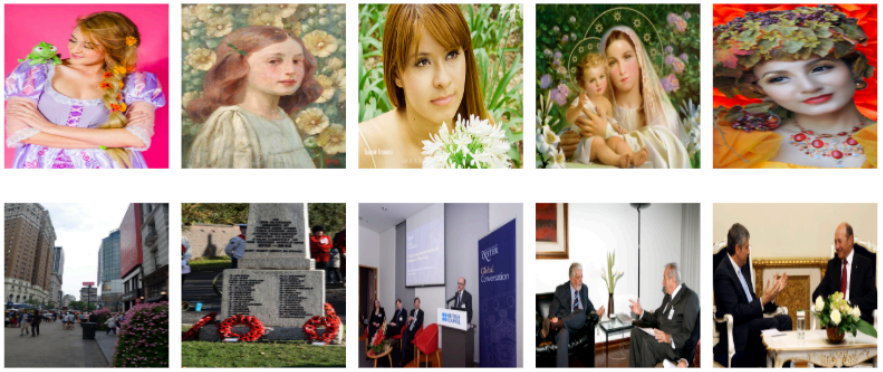
CEO



[Matthew Kay et al. CHI 2015]

Stereotyped representation in datasets

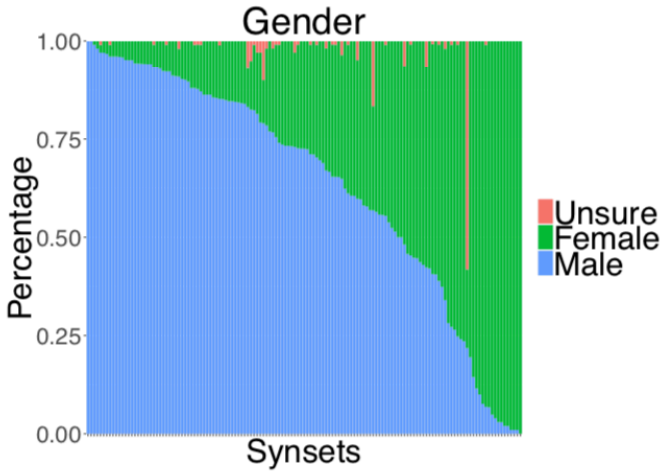
person+flower

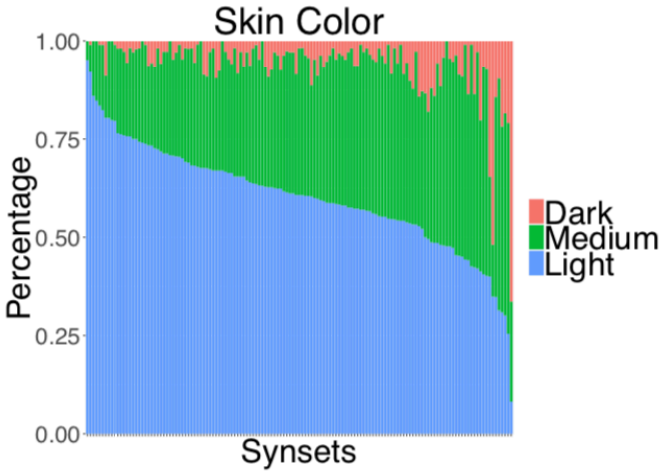


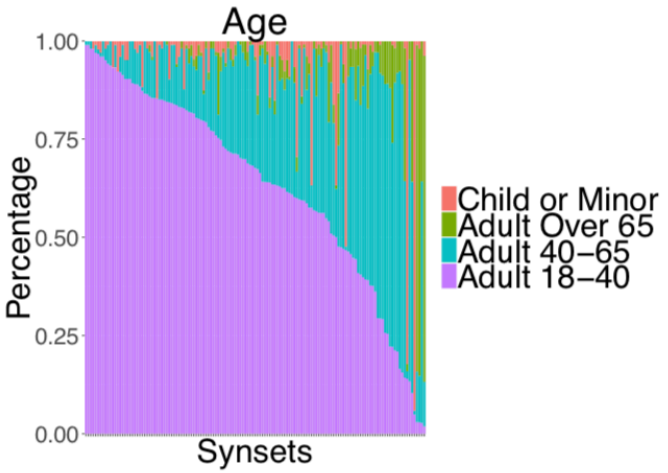
[Angelina Wang et al. ECCV 2020]

Counteracting the disparities by annotating demographics

Annotated demographics on 139 people synsets (categories) in ImageNet
13,900 images; 109,545 worker judgments.



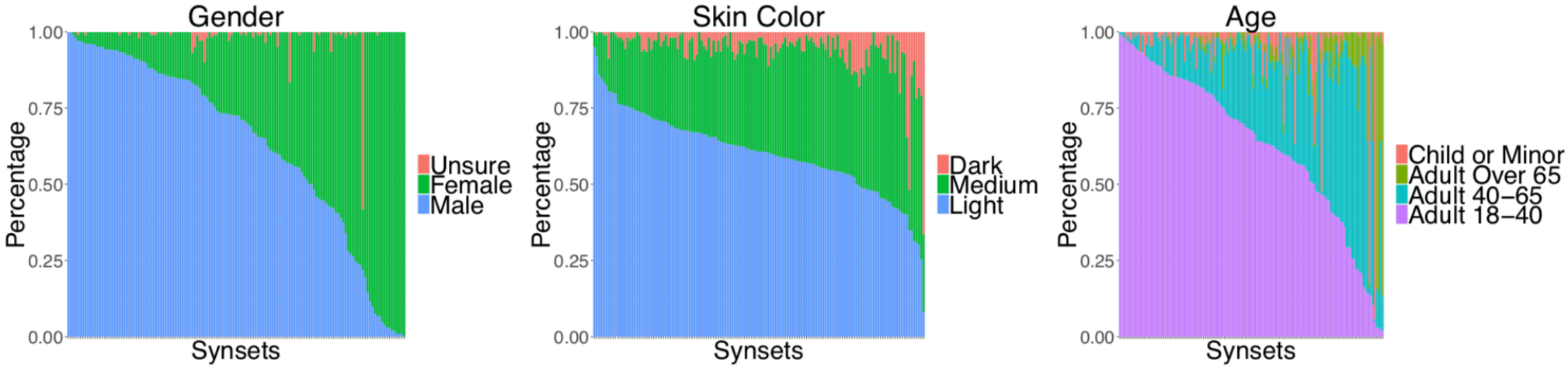




["Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy." Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, Olga Russakovsky. FAT* 2020. http://image-net.org/filtering-and-balancing]

Counteracting the disparities by annotating demographics

Annotated demographics on 139 people synsets (categories) in ImageNet
13,900 images; 109,545 worker judgments.



Subtleties:

- Rebalancing $\Longrightarrow$ removing data, changing the original distribution
- Accuracy/validity of these labels
- The implication of including people categories in a dataset (cf. the FAT* paper)
- Privacy of subjects, esp. minors; consent of content creators (working on this)
- The representation of folks of different genders (skin colors, ages) within a synset

["Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy."
Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, Olga Russakovsky. FAT* 2020. http://image-net.org/filtering-and-balancing]

Revealing and mitigating dataset biases with REVISE: REvealing VIsual biaSEs tool

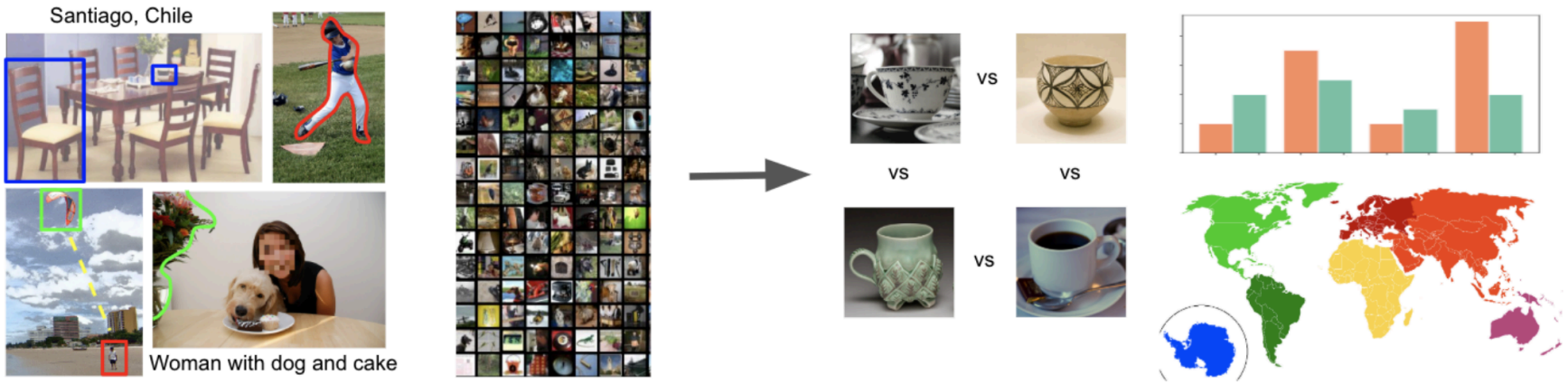


Fig. 1: Our tool takes in as input a visual dataset and its annotations, and outputs metrics, seeking to produce insights and possible actions.

Key contributions:

- Goes beyond *underrepresentation* to analyzing differences in *portrayal*
- Allows for semi-automatic analysis of *large-scale* datasets
- *Aids* dataset creators&users: fairness ultimately requires manual intervention
- Integrates bias mitigation throughout the dataset *construction* process

[“REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets.”
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

Inner workings of the REVISE tool

Implementation:

- Freely available Python notebooks
- Analyzes portrayal of objects, people and geographic regions
- Uses provided annotations, pre-trained models, and models trained on the data

In this talk:

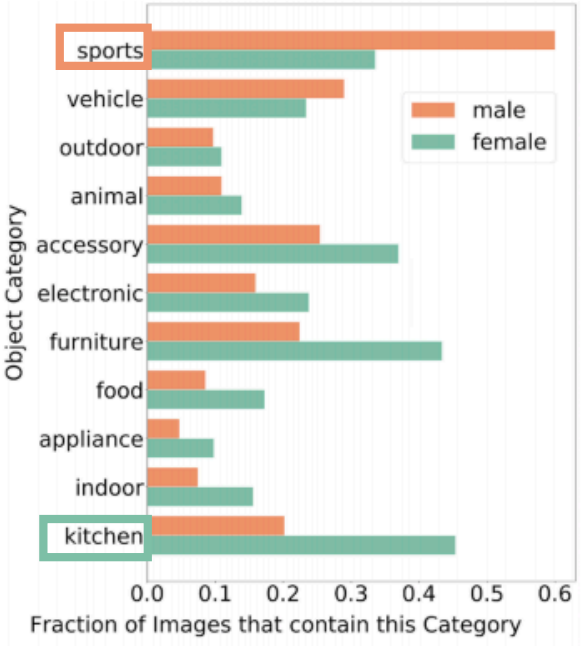
- Focus specifically on portrayal of different *genders*
- Caveat: use of binarized socially-perceived gender expression
- Analysis on COCO [T. Y. Lin et al. ECCV '14] and OpenImages [I. Krasin et al. '17]
- Gender annotations derived from image captions [J. Zhao et al. EMNLP'17]

["REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets."
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

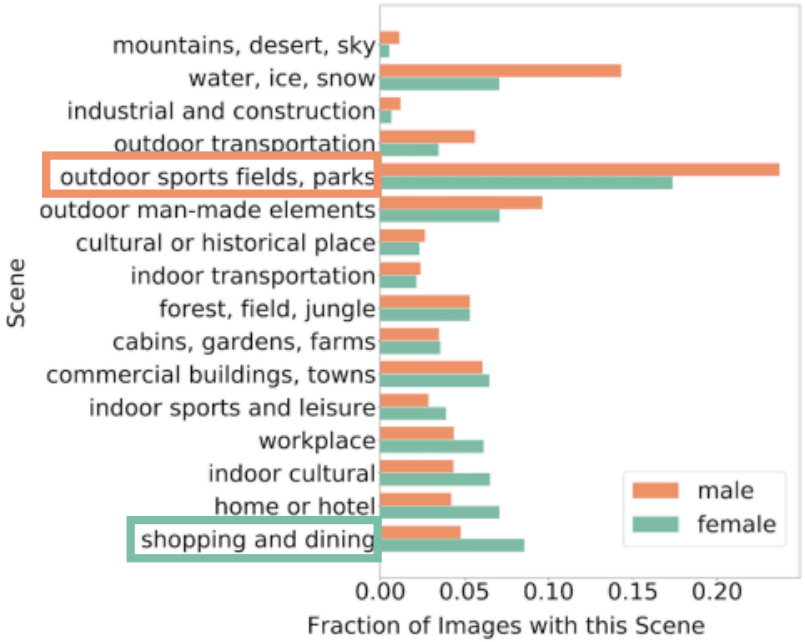
Co-occurrence of males and females with different objects and in different scenes

Analysis: correlate the presence of different genders in COCO with

Object categories, using ground truth object annotations grouped manually into super-categories



Scene categories, computed with pre-trained Places network [B. Zhou et al. TPAMI '17]

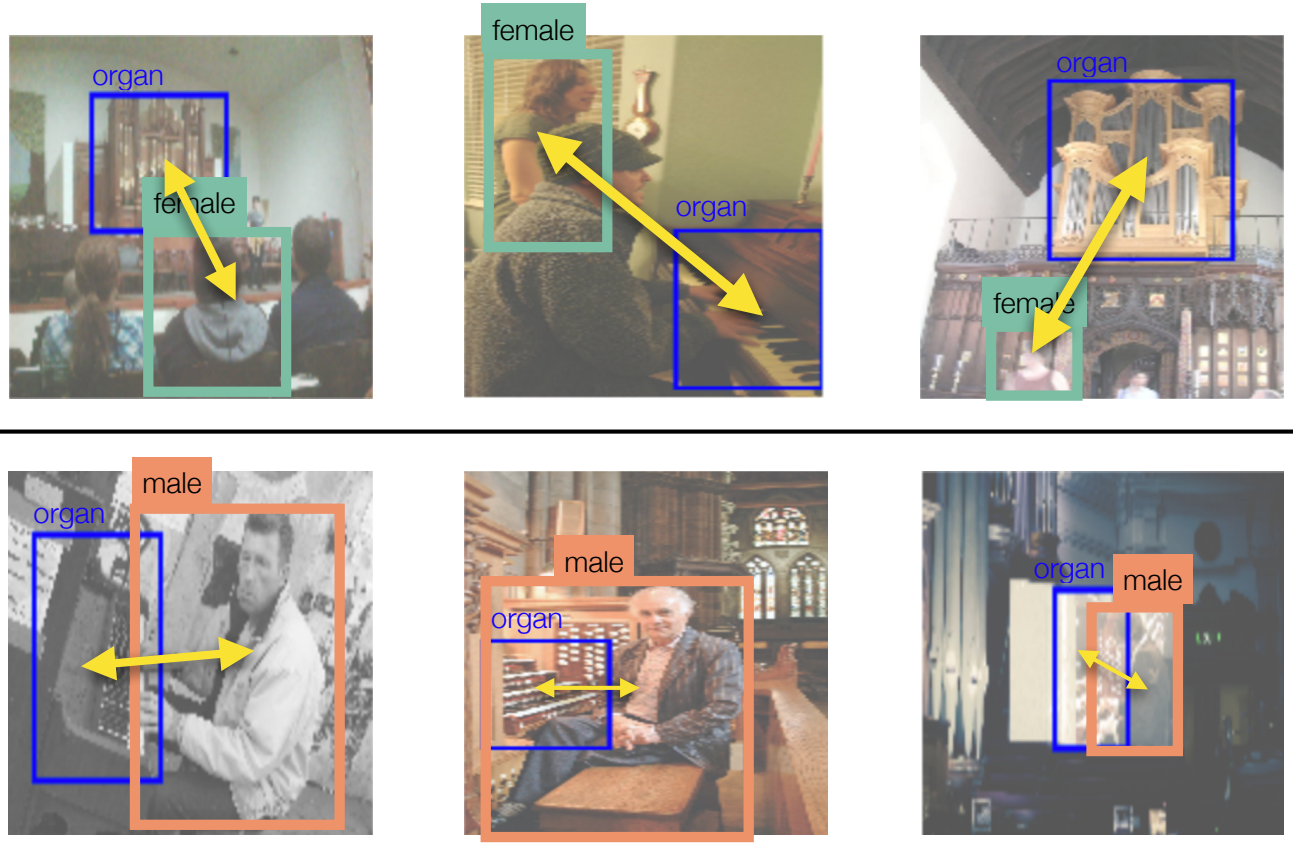


Actionable insight: collect images of the underrepresented gender with the corresponding objects and scenes

["REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets."
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

Interaction between objects and people of different genders

Analysis: use the person-object distance as a proxy for interaction



Actionable insight: consider equalizing the level of interaction with the object (if warranted)

["REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets."
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

Differences in *portrayal* of different genders

Analysis:

- for each object class, learn visual classifiers for recognizing this object when it's present with females vs present with males
- identify classes with most stark differences between genders

Actionable insight: collect more images of each gender with the particular object in more diverse situations

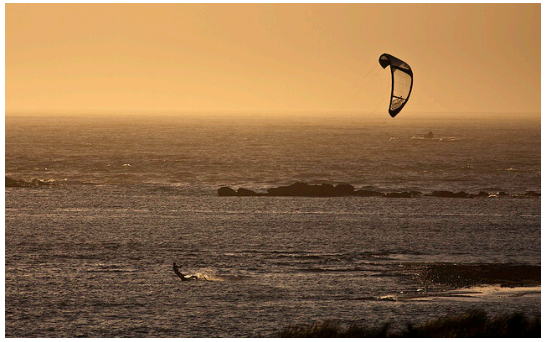
["REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets."
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

Annotated gender in datasets defaults to “male”

Analysis: investigate occurrences where gender is annotated but the person is too small or no face is detected in the image

Man and boats on the sand in low tide.

The group of buses are parked along the city street as a **man** crosses the street in the background.



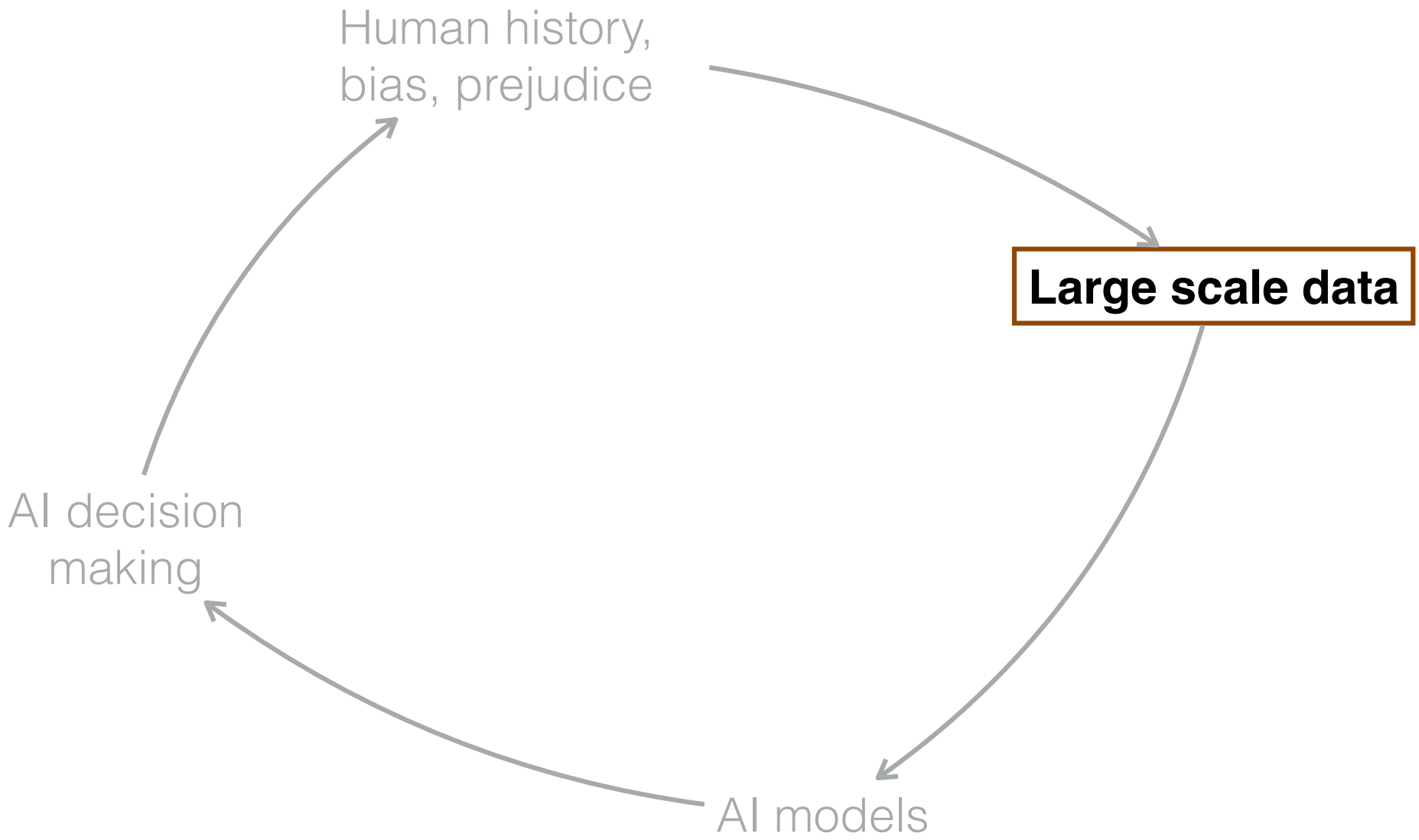
A **man** is kiteboarding in the open ocean.

A **man** riding a kiteboard on top of a wave in the ocean.

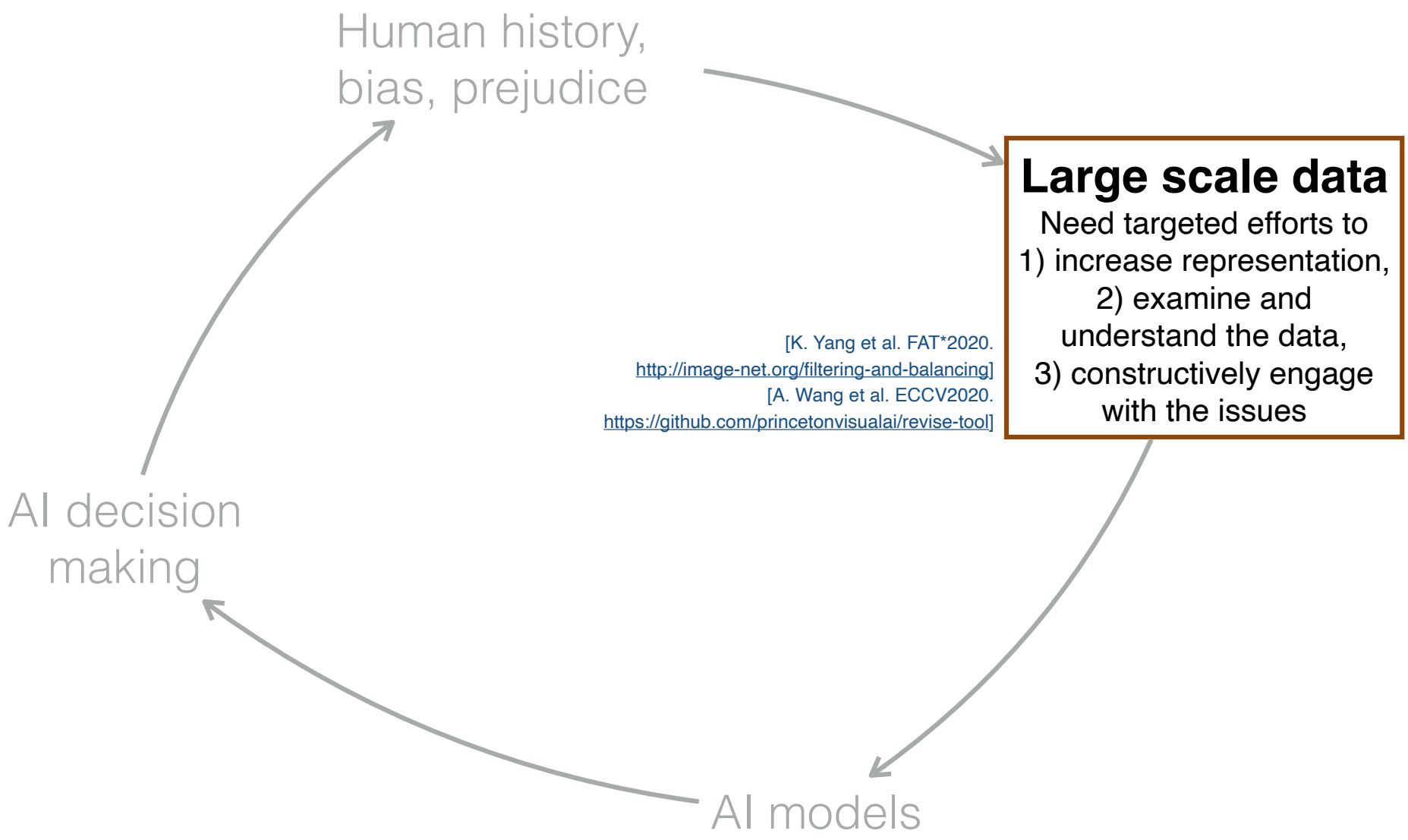
Actionable insight: prune these gender labels

[“REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets.”
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

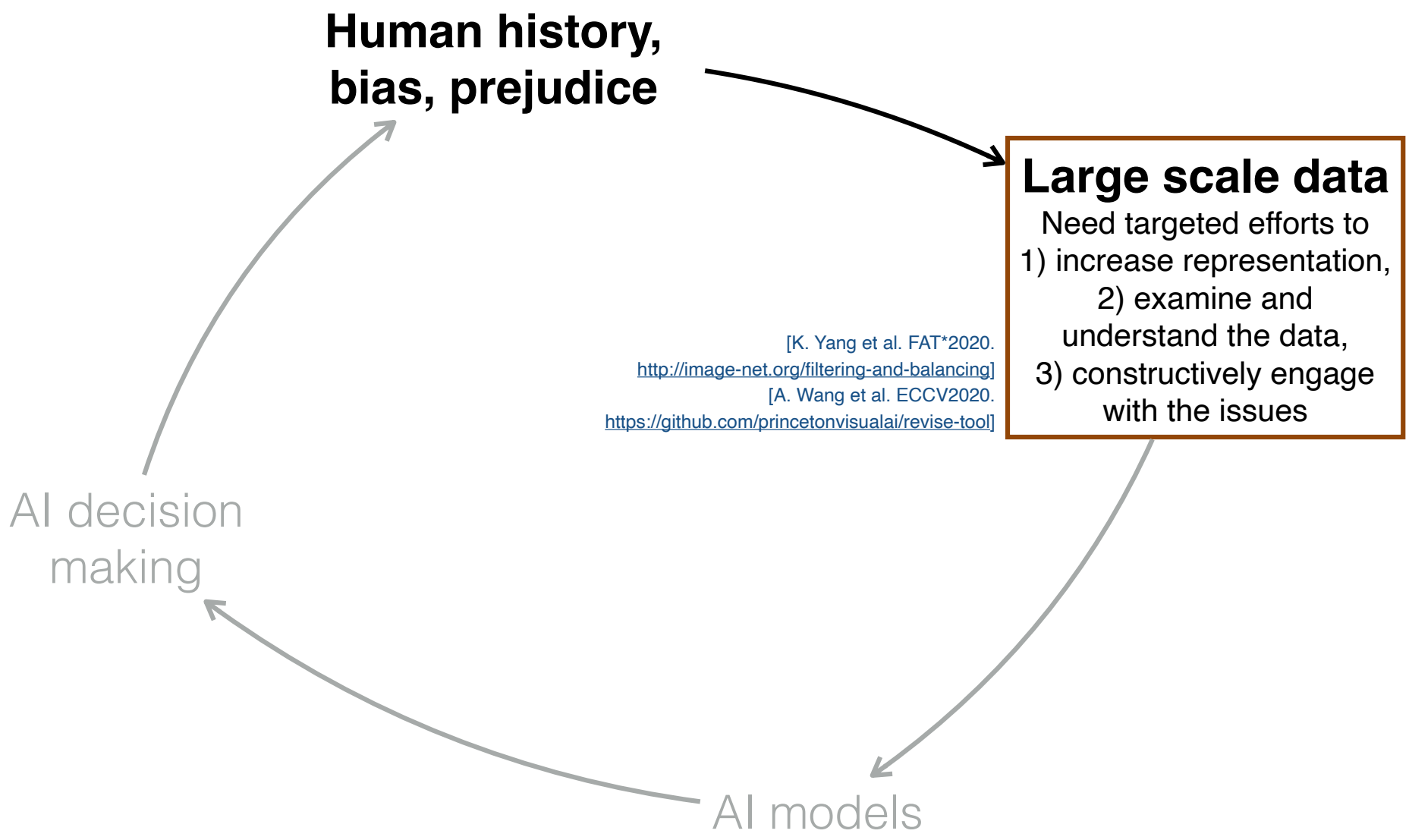
Can we adjust the AI design to mitigate these effects?



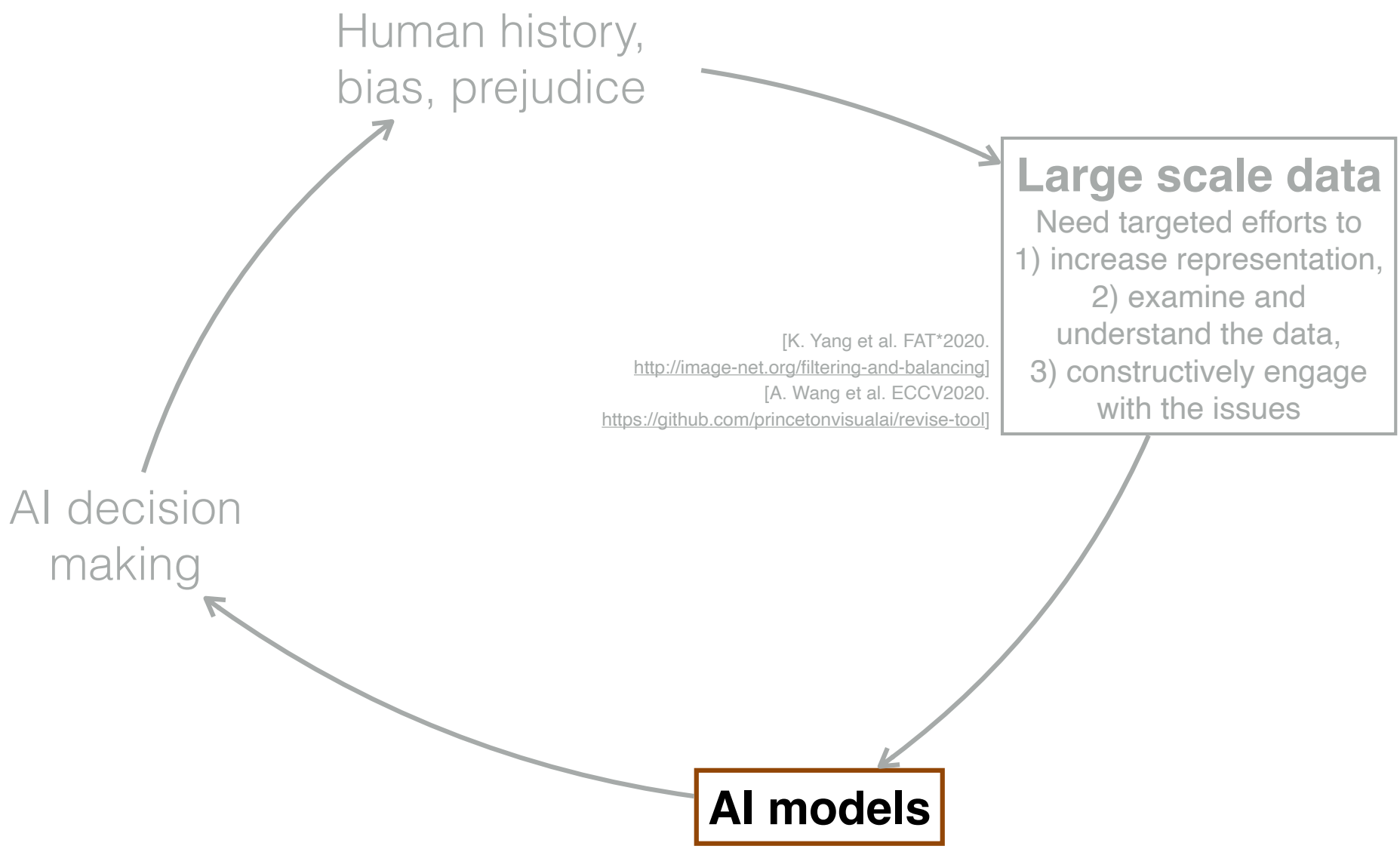
Can we adjust the AI design to mitigate these effects?



Can we adjust the AI design to mitigate these effects?

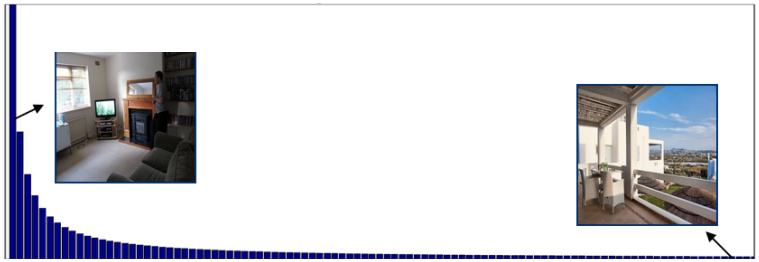


Can we adjust the AI design to mitigate these effects?

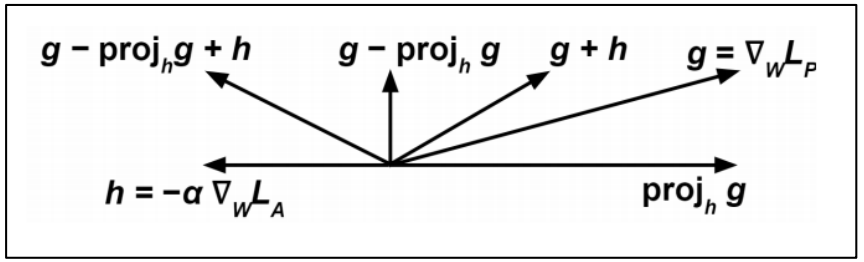


Re-visiting many existing problems in this context

Long tail distributions

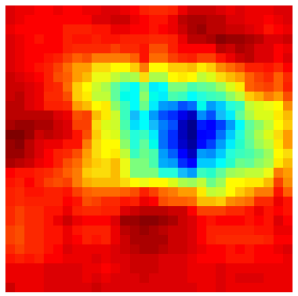


Learning with constraints



Domain adaptation

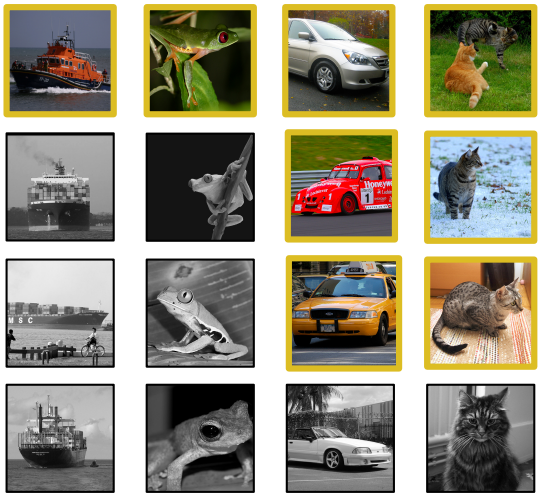
Interpretability



Our problem: teaching a classifier to ignore a known spurious correlation in the data

Toy illustration on CIFAR, to temporarily simplify the exploration

Training: skewed distributions
(correlates class with color/grayscale)



Testing: classifying images into one of
10 object classes (no correlation)

Testing on color images

Training on skewed data:
89% accuracy

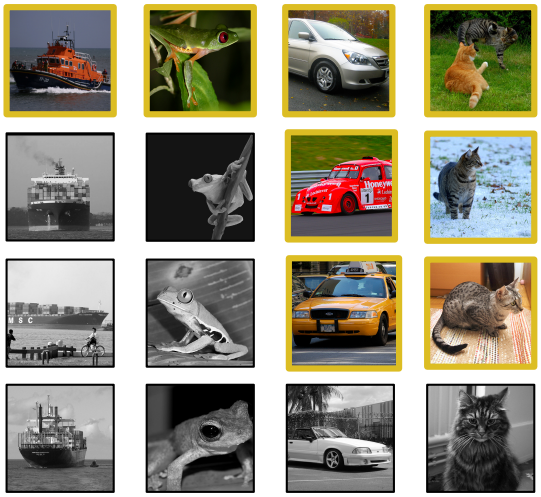
Training on all-grayscale:
93% accuracy

["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

Our problem: teaching a classifier to ignore a known spurious correlation in the data

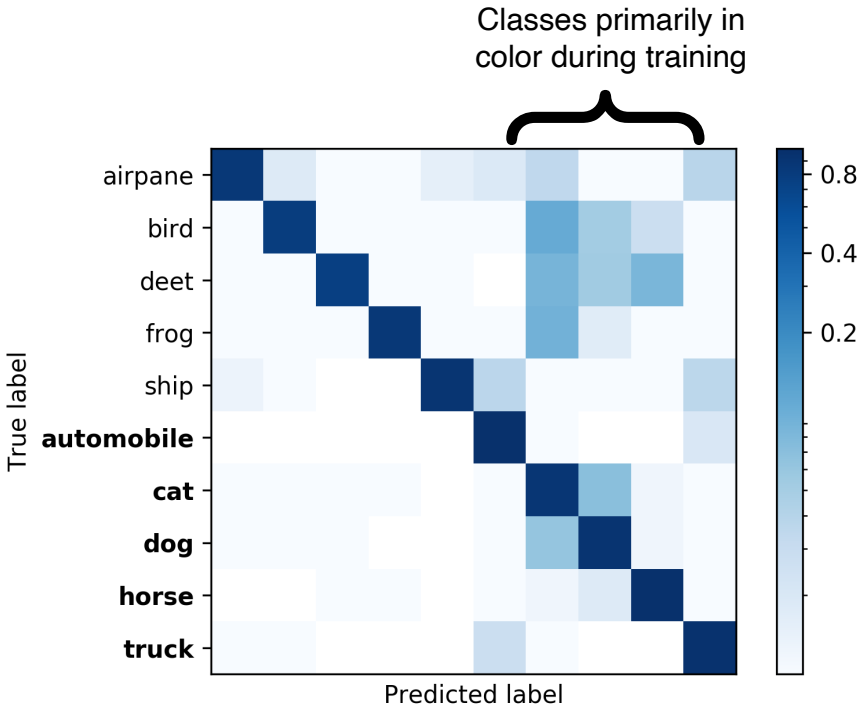
Toy illustration on CIFAR, to temporarily simplify the exploration

Training: skewed distributions
(correlates class with color/grayscale)



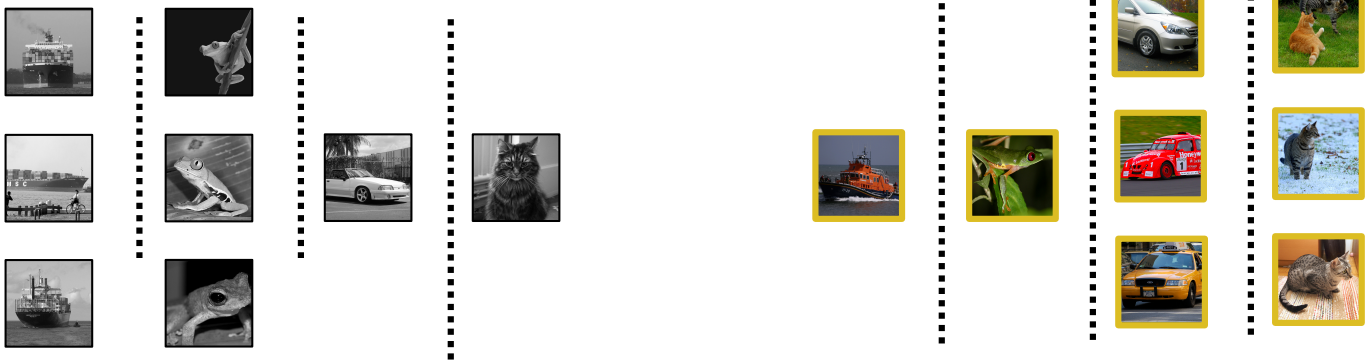
Testing: classifying images into one of
10 object classes (no correlation)

Testing on color images

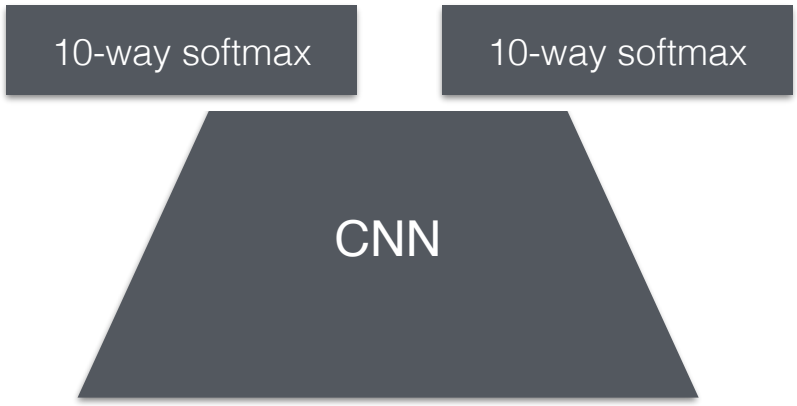


["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

Domain-independent training works very well



$$\mathscr{L} = -\sum_i log P(y_i \mid d_i, x_i)$$



x_i = image i

y_i = object class for image i

d_i = domain (c or g) for image i

Inference:

$$\arg\max_y \sum_d s(y, d, x)$$

s = pre-softmax score

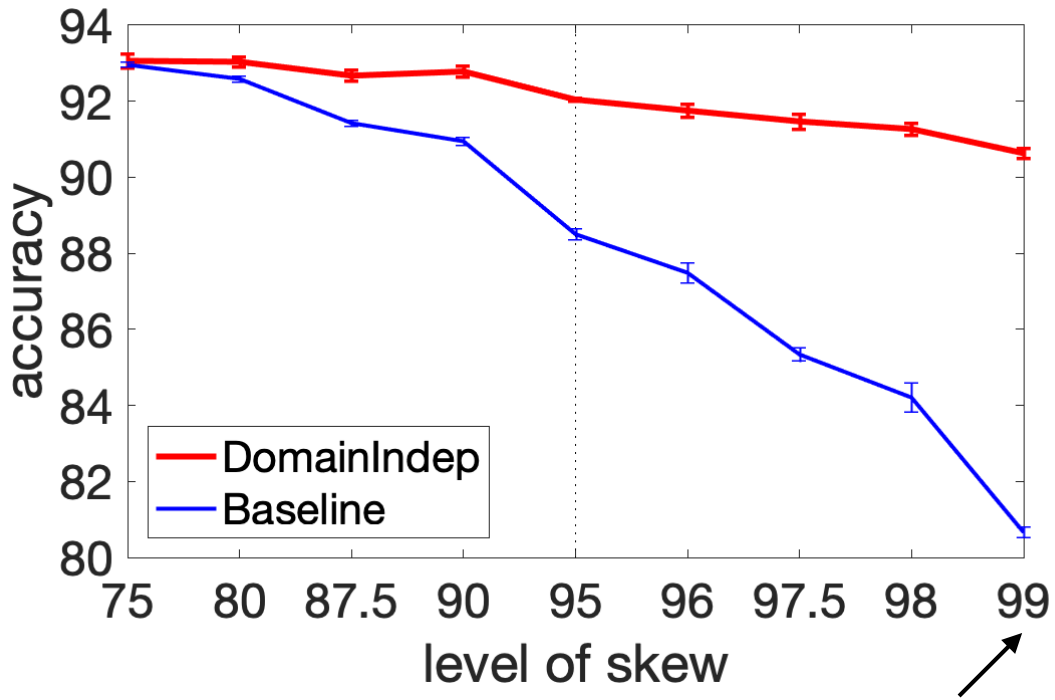
["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

Domain-independent training works very well, especially with high class/domain correlation

Training data: CIFAR-10, skewed color/grayscale distribution

Architecture: ResNet-18

Testing metric: Mean per-class per-domain accuracy (i.e., equal color/grayscale distribution within classes)



Every object class is either 99% color images or 99% grayscale images during training

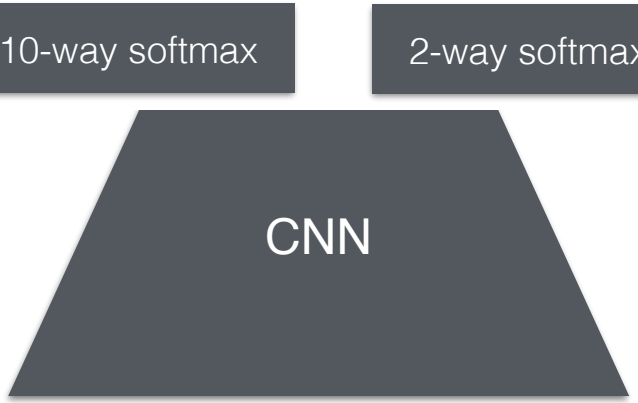
["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

Adversarial de-biasing does not work

Want to classify the object classes

But not be able to classify the domain

$$L_{cls} = -\sum_i \left(logP(y_i \mid x_i) - \alpha \frac{1}{2} \sum_{d=1,2} logP(d \mid x_i) \right)$$



$$L_{domain} = -\sum_i logP(d_i \mid x_i)$$

†[Alvi et al. "Explicit removal of biases…" ECCVW' 18]

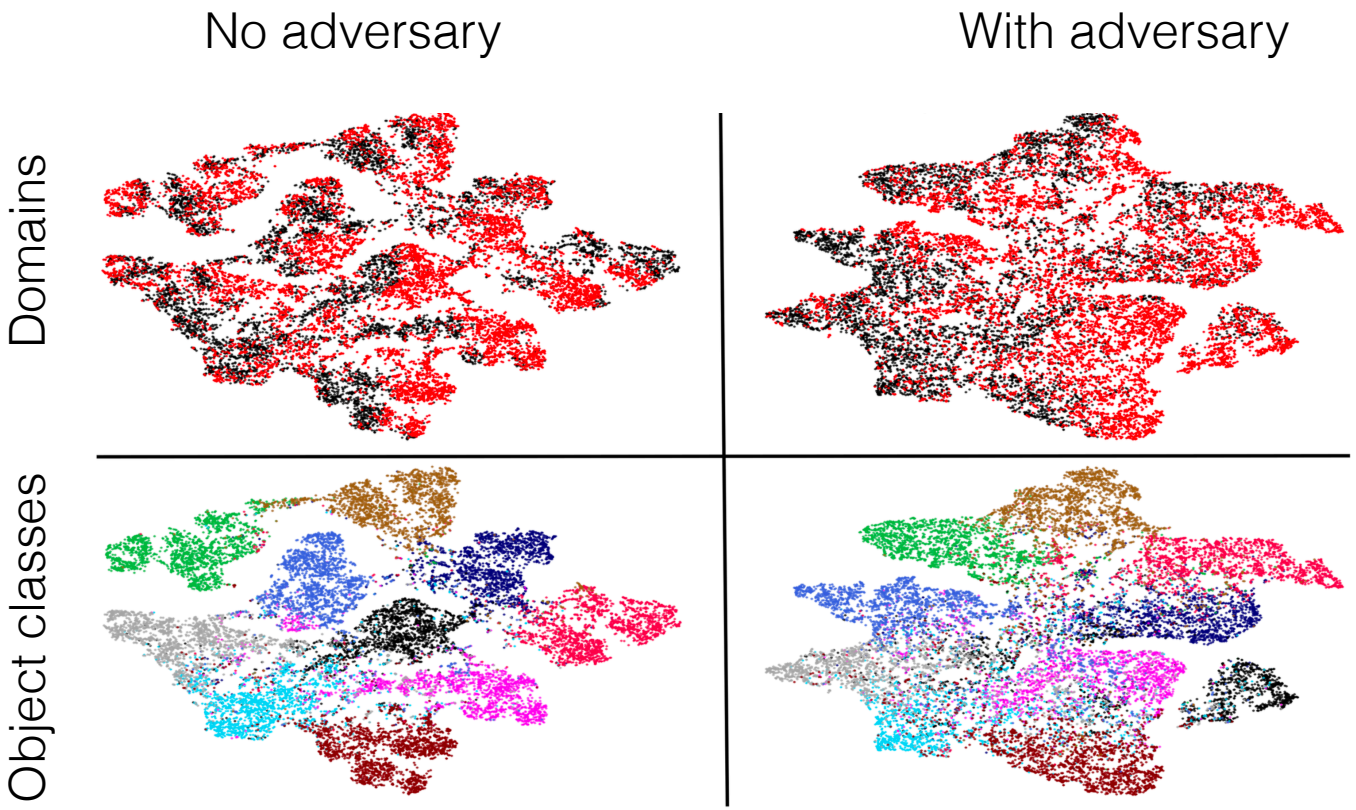
x_i = image i

y_i = object class for image i

d_i = domain (c or g) for image i

["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

Adversarial de-biasing does not work



- It is very difficult to train a visual representation that can't classify the domain (visual representations are powerful!)
- In cases of high correlation between classes and domains, one can be inferred from the other during training so adversarial de-biasing may not be appropriate
- Beware of fairness literature that reports equal error rates but not overall accuracy

["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

These findings hold across a variety of settings

MODEL	CIFAR-grayscale	CIFAR-ImageNet	CelebA
BASELINE	88.5 ± 0.3	79.4 ± 0.4	74.7
ADVERSARIAL	83.8 ± 1.1	74.1 ± 0.6	71.9
DOMINDEP	$\mathbf{92.0 \pm 0.1}$	$\mathbf{83.5 \pm 0.3}$	**76.3**

Training data: CIFAR-10, skewed color/grayscale distribution

Architecture: ResNet-18

Testing metric: Mean per-class per-domain accuracy (i.e., equal color/grayscale distribution within classes)

Same, except more subtle domain shift (substituting in images of similar classes from ImageNet instead of converting to grayscale)

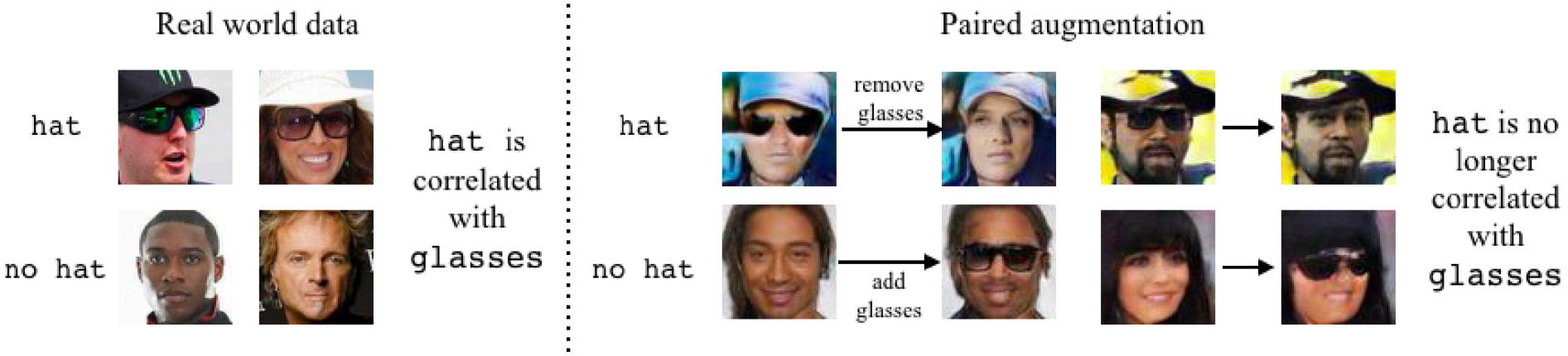
Task: multi-label face attribute recognition, where presence&appearance of an attribute may be correlated with gender

Architecture: ResNet-50, pre-trained on ImageNet

Testing metric: Weighted mean average precision (i.e., equal gender distribution within classes)

["Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. CVPR 2020. https://github.com/princetonvisualai/DomainBiasMitigation]

Coming soon: can we *generate* the data we need?



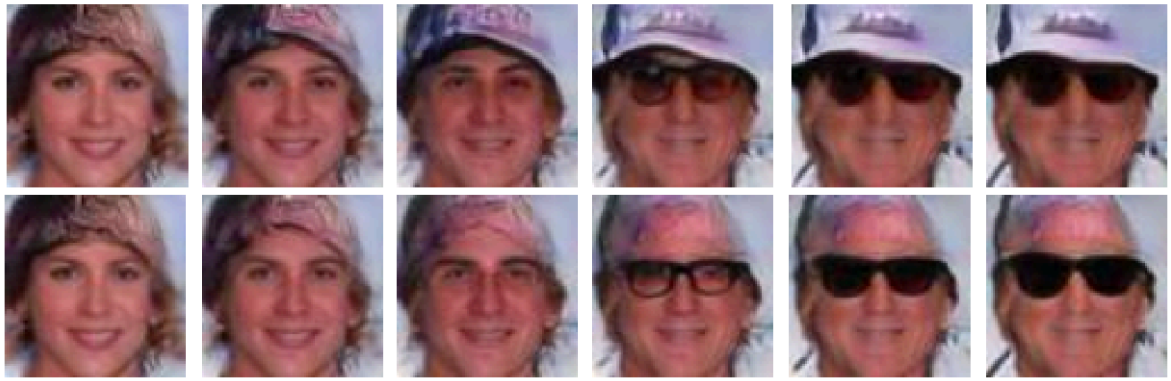
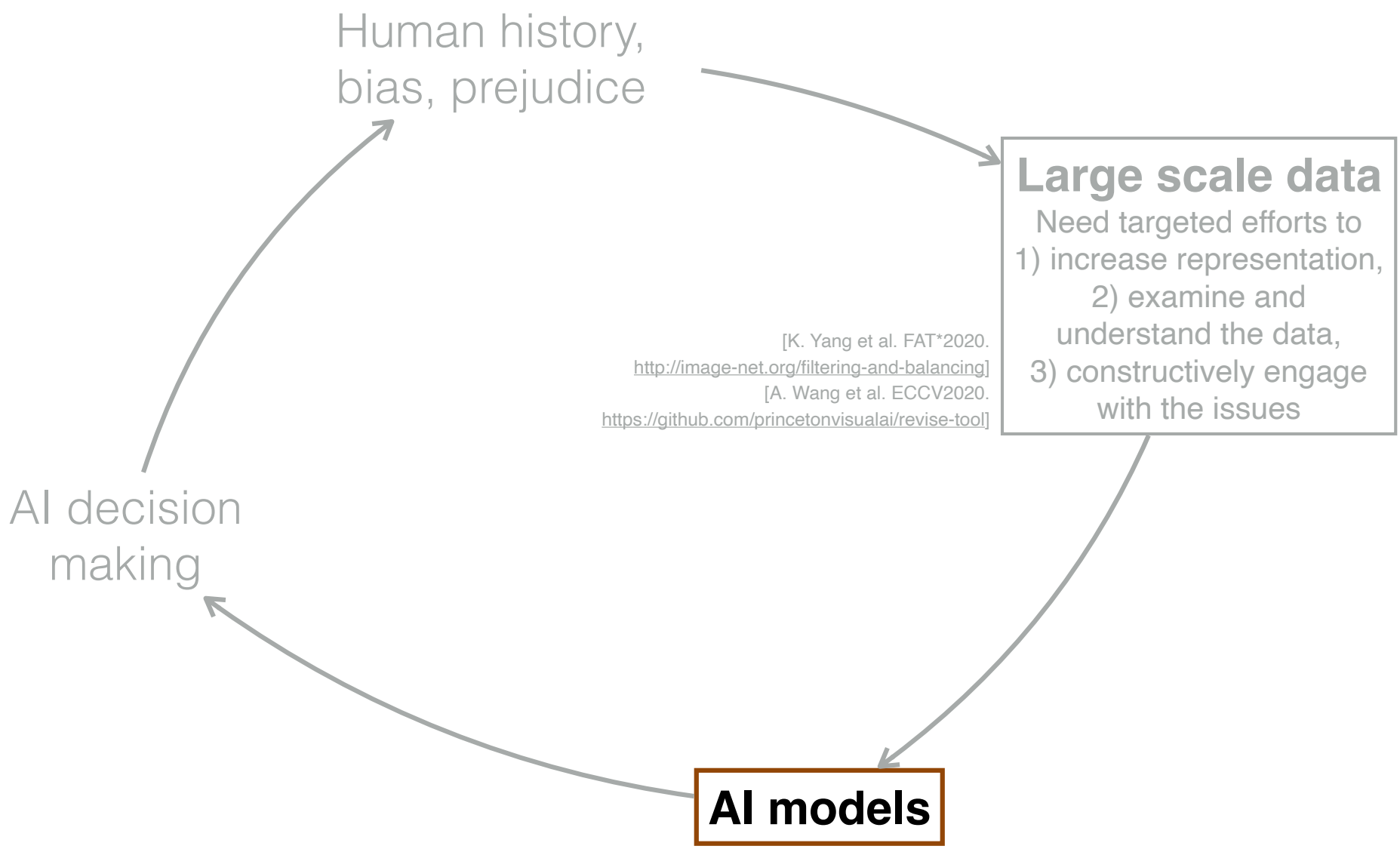


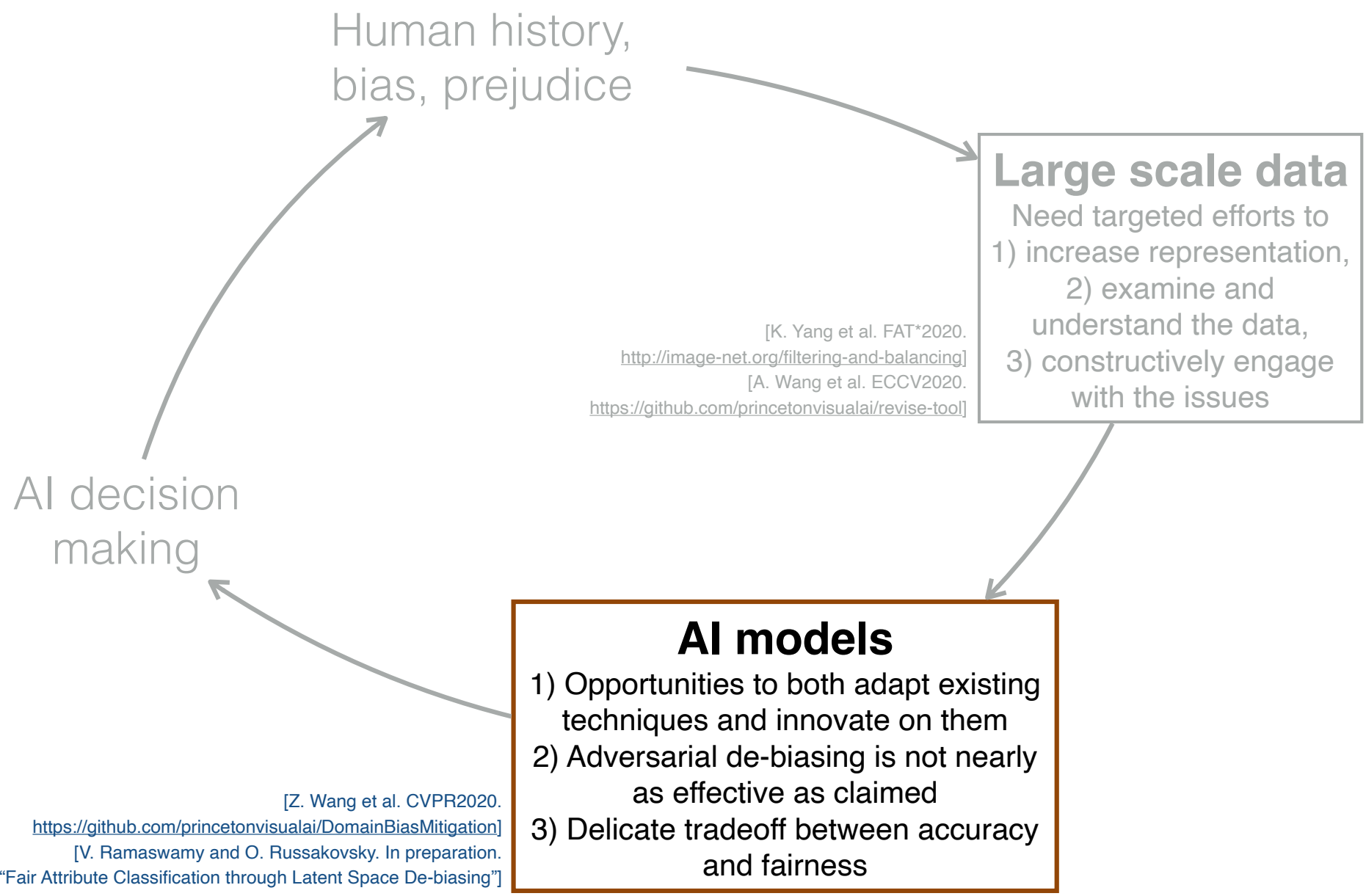
Fig. 2: Consider a GAN trained on a biased dataset of faces where the presence of hats is correlated with the presence of sunglasses. Moving in the latent space towards adding glasses additionally adds a hat (*top*). In contrast, our method allows for adding glasses without including the correlated hat attribute (*bottom*). Note that other attributes (apart from the target attribute) can change

["Fair Attribute Classification through Latent Space De-Biasing." Vikram Ramaswamy and Olga Russakovsky. In preparation.]

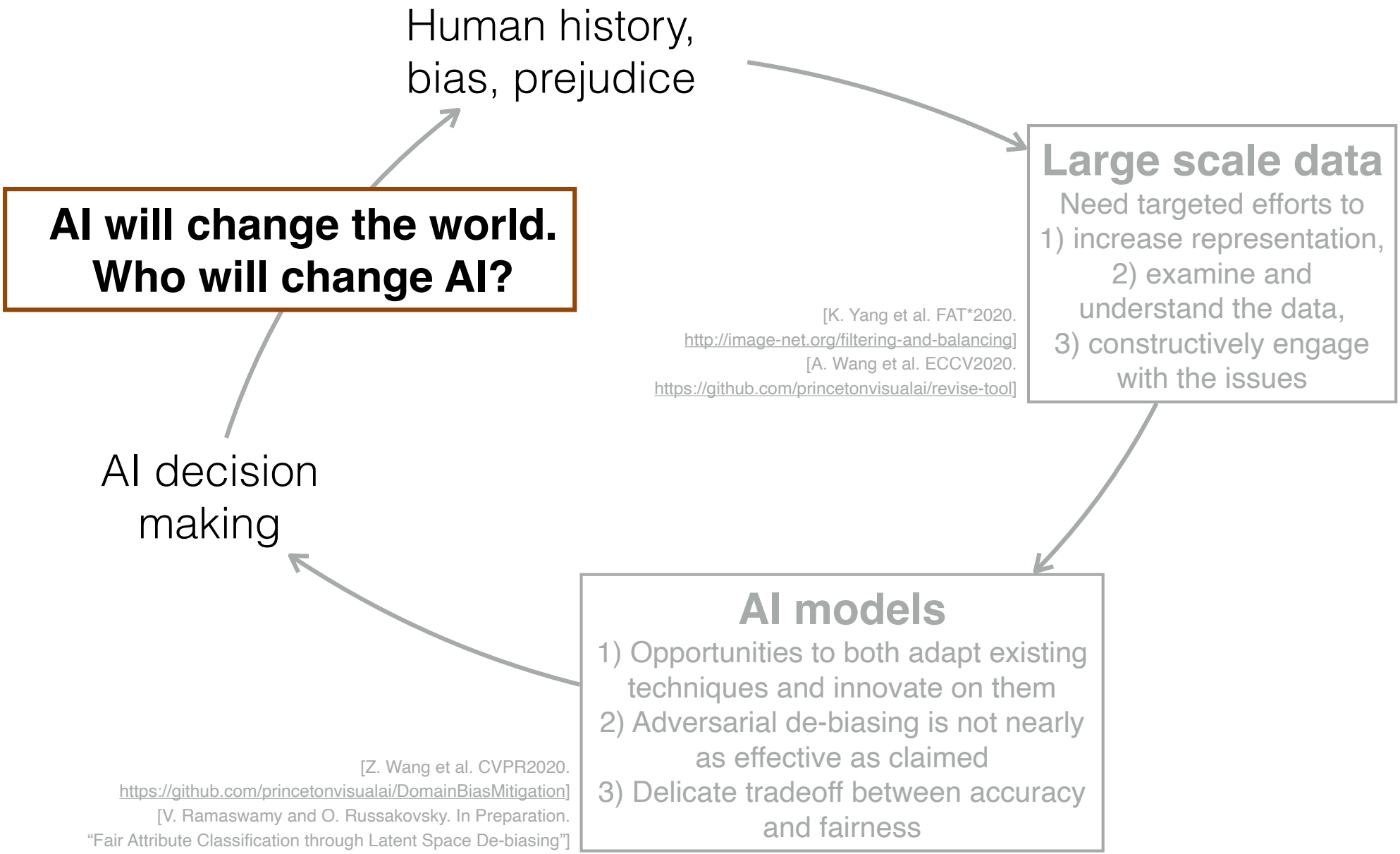
Can we adjust the AI design to mitigate these effects?



Can we adjust the AI design to mitigate these effects?



Can we adjust the AI design to mitigate these effects?



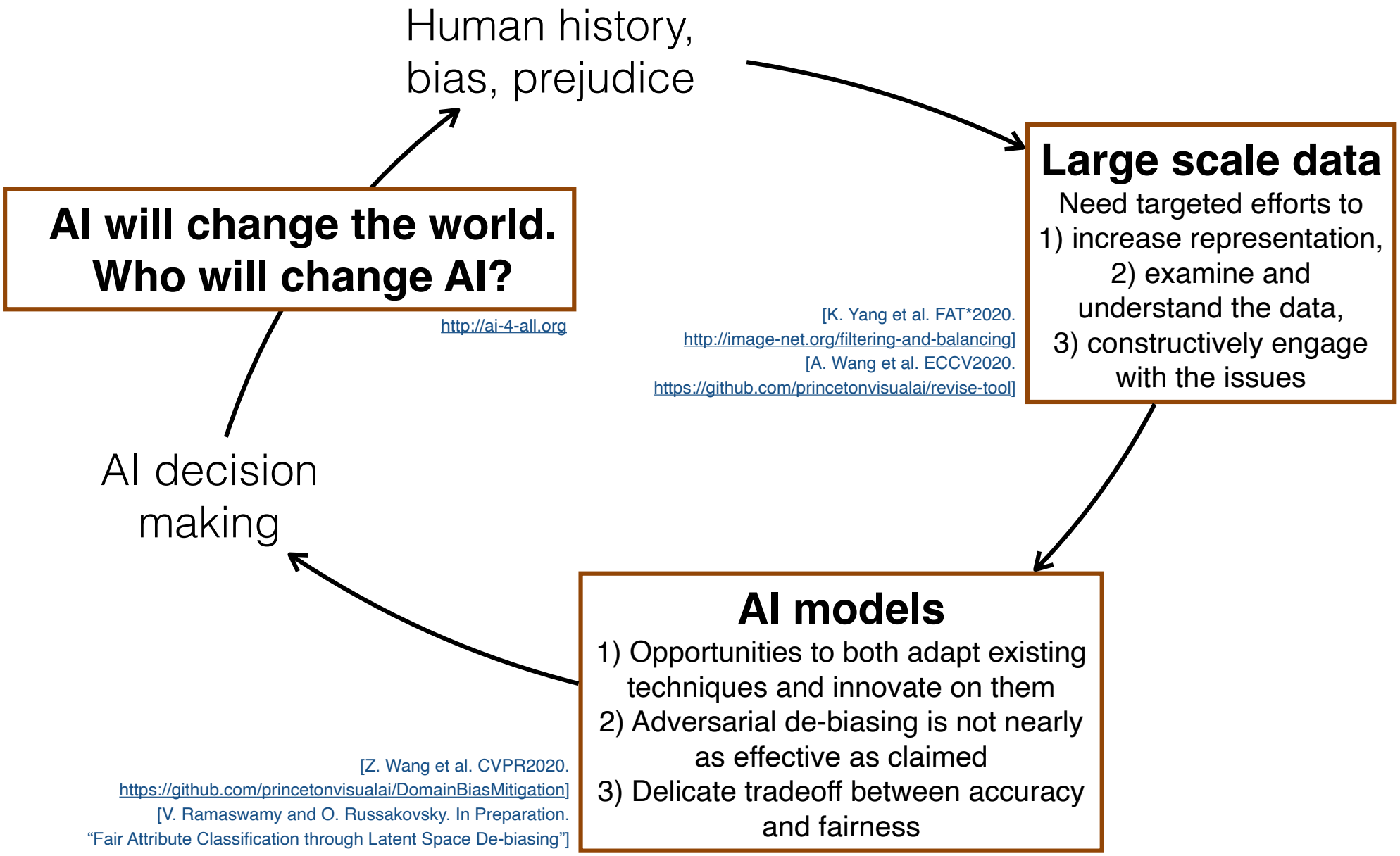
AI4ALL: a non-profit dedicated to increasing diversity and inclusion in AI

- Celebrated our 3rd birthday on March 8, 2020
- Partnered with 16 universities to run summer programs for high school students from underrepresented groups
  - https://ai-4-all.org/summer-programs/
- Launched a free online OpenLearning platform
  - https://ai-4-all.org/open-learning
- Summer program alumni have started AI research projects, internships, working groups, panels, clubs, … (while still in high school/early college)
  - https://medium.com/ai4allorg/alumni/
- Long-term vision is to foster a community of diverse ***leaders*** in AI

"Until this program, I never thought that people who look like me could succeed in computer science and AI."

- AI4ALL 2016 student

Can we adjust the AI design to mitigate these effects?



Can we adjust the AI design to mitigate these effects?

@VisualAILab, @ai4allorg
http://visualai.princeton.edu
olgarus@cs.princeton.edu

Human history, bias, prejudice

AI will change the world. Who will change AI?

http://ai-4-all.org


[K. Yang et al. FAT*2020. http://image-net.org/filtering-and-balancing]
[A. Wang et al. ECCV2020. https://github.com/princetonvisualai/revise-tool]

Large scale data

Need targeted efforts to
1) increase representation,
2) examine and understand the data,
3) constructively engage with the issues

AI decision making


[Z. Wang et al. CVPR2020. https://github.com/princetonvisualai/DomainBiasMitigation]
[V. Ramaswamy and O. Russakovsky. In Preparation. "Fair Attribute Classification through Latent Space De-biasing"]

AI models

1) Opportunities to both adapt existing techniques and innovate on them
2) Adversarial de-biasing is not nearly as effective as claimed
3) Delicate tradeoff between accuracy and fairness

NSF
Google Cloud
PRINCETON
School of Engineering and Applied Science