Fairness in visual recognition

Olga Russakovsky

http://visualai.princeton.edu

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

THE VERGE **April 25, 2017**

“Machines taught by photos learn a sexist view of women”

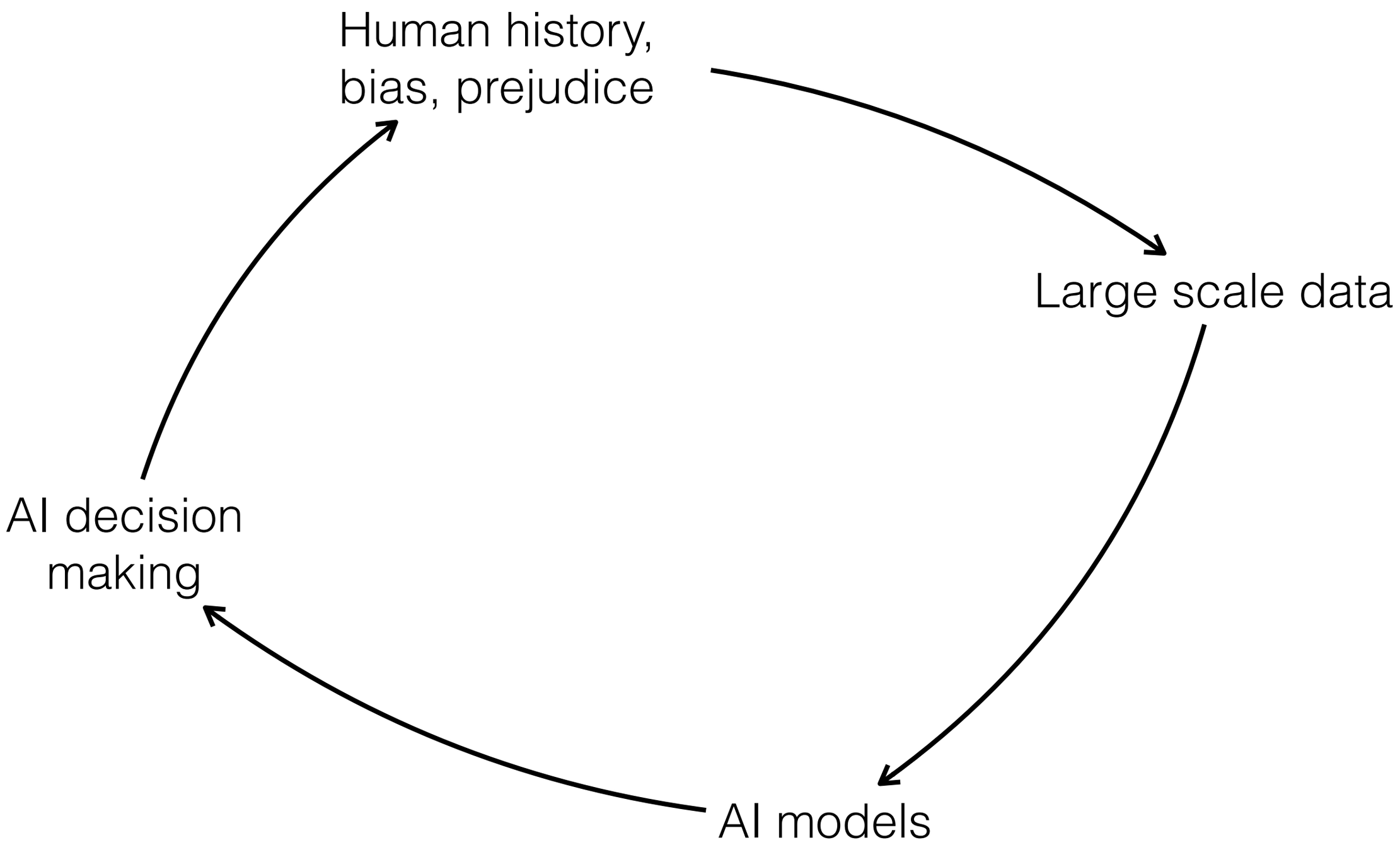
WIRED **Aug 21, 2017**

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

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

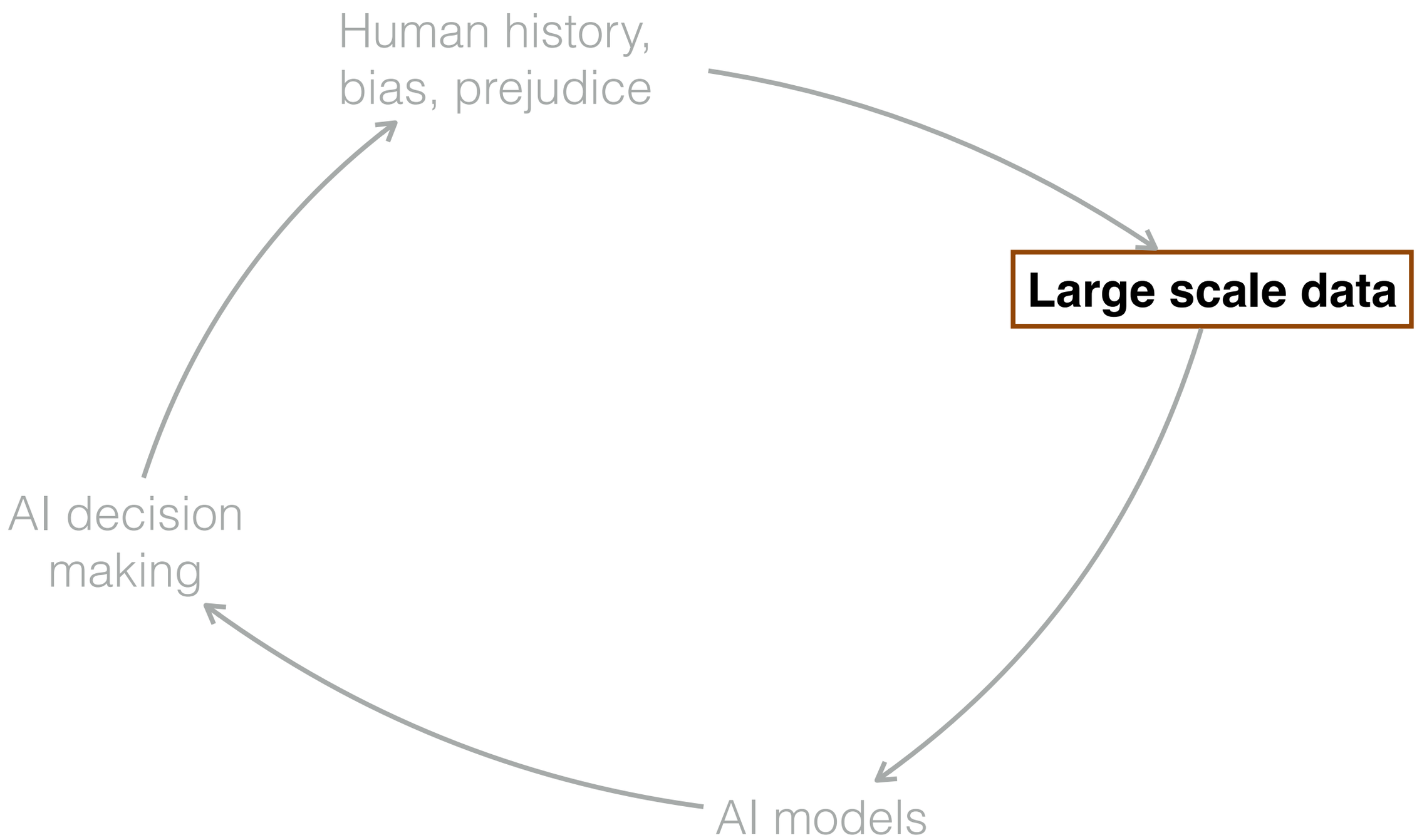
See: Ruha Benjamin’s “Race after technology” — an excellent book

Can we adjust the AI design to mitigate these effects?



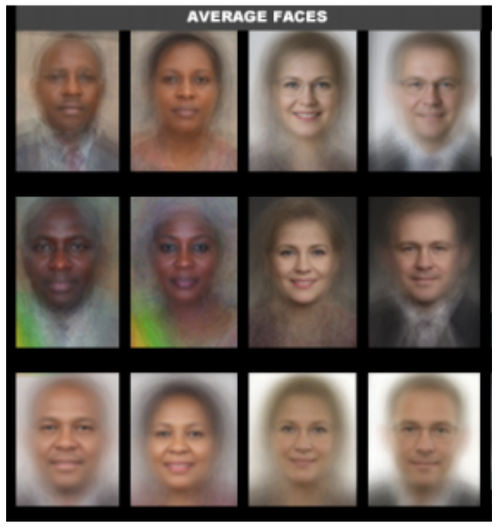
See: Ruha Benjamin's "Race after technology" — an excellent book

Can we adjust the AI design to mitigate these effects?



Large scale ≠ fair representation

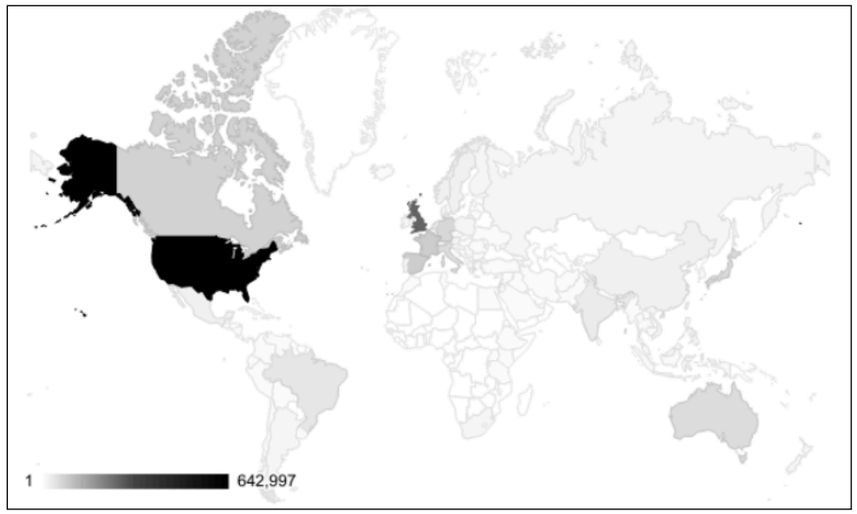
Race diversity in face datasets



[Joy Buolamwini & Timnit Gebru. FAT* 2018]

Geographic diversity

(in ImageNet and OpenImages)



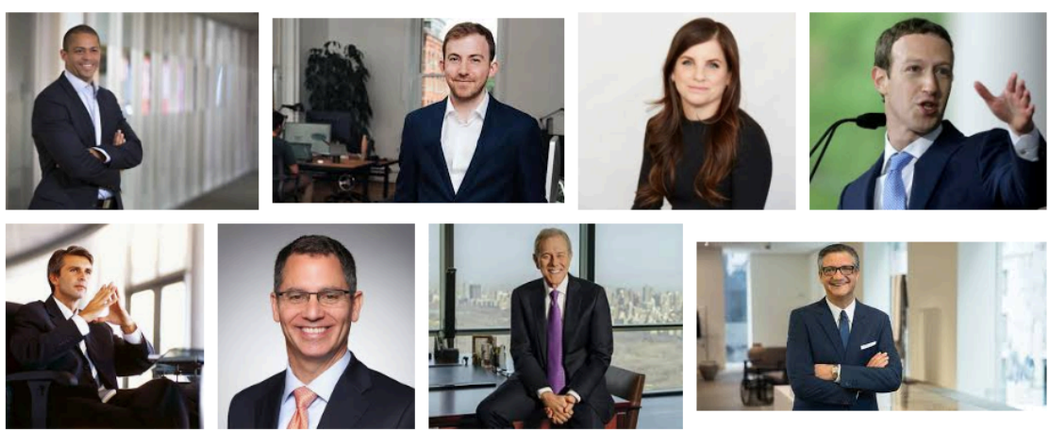
[Shreya Shankar et al. NeurIPS 2017]

Object diversity

[Terrance DeVries et al. CVPRW 2019]

Diversity in image search results

CEO



[Matthew Kay et al. CHI 2015]

Safiya Umoja Noble's "Algorithms of Oppression" — another excellent book

REVISE: REvealing VIsual biaSEs tool

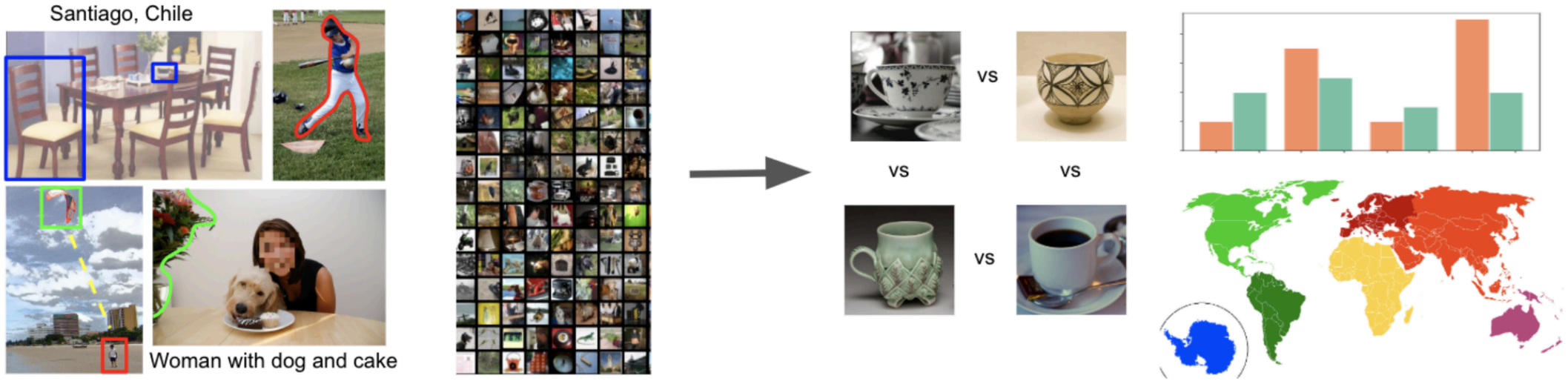


Fig. 1: Our tool takes in as input a visual dataset and its annotations, and outputs metrics, seeking to produce insights and possible actions.

Key contributions:

- Provides transparency into *large-scale* datasets
- *Aids* dataset creators&users: fairness ultimately requires manual intervention
- Leverages the (1) available annotations, (2) existing pre-trained models, *and* (3) models trained on the data
- Goes beyond noting *underrepresentation* to analyzing differences in *portrayal* of different objects, genders and geographic regions

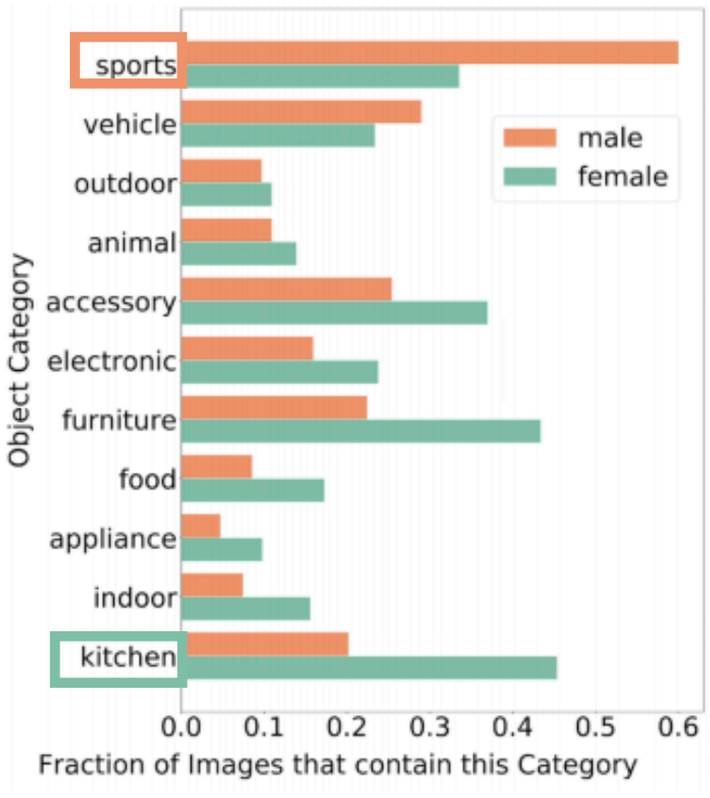
["REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets."
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

Context in images depicting males and females*

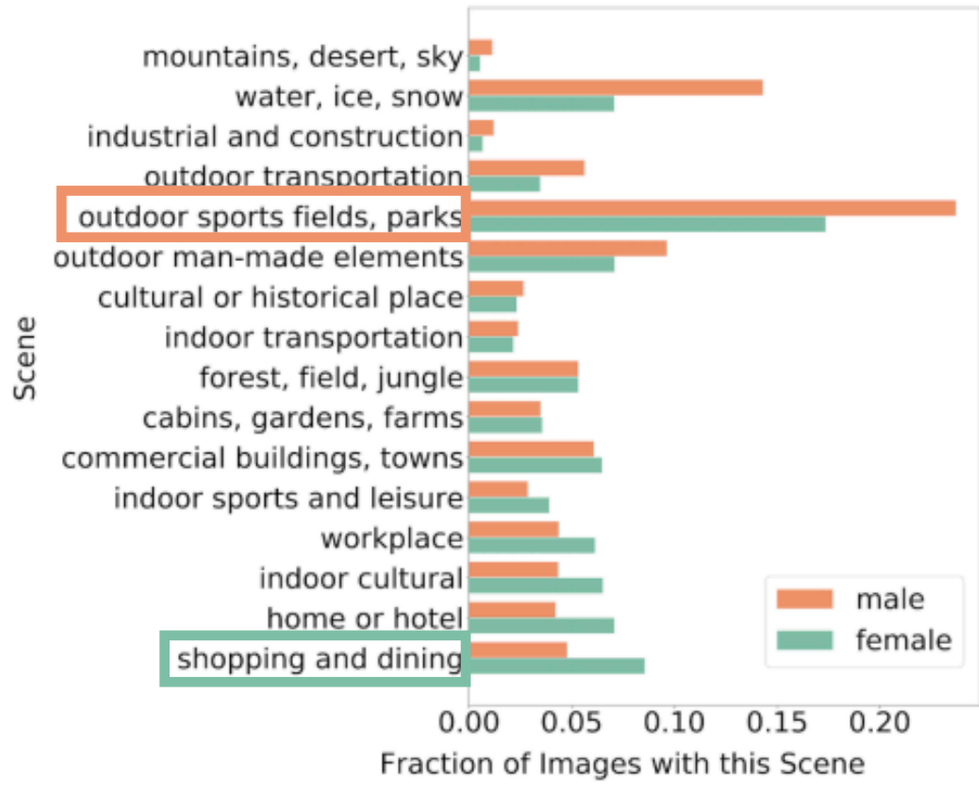
*Gender annotations of *binarized socially-perceived* gender *expression* derived from image captions [J. Zhao et al. EMNLP'17]

Analysis: correlate the presence of different genders in COCO [Lin et al. ECCV'14] with

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



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



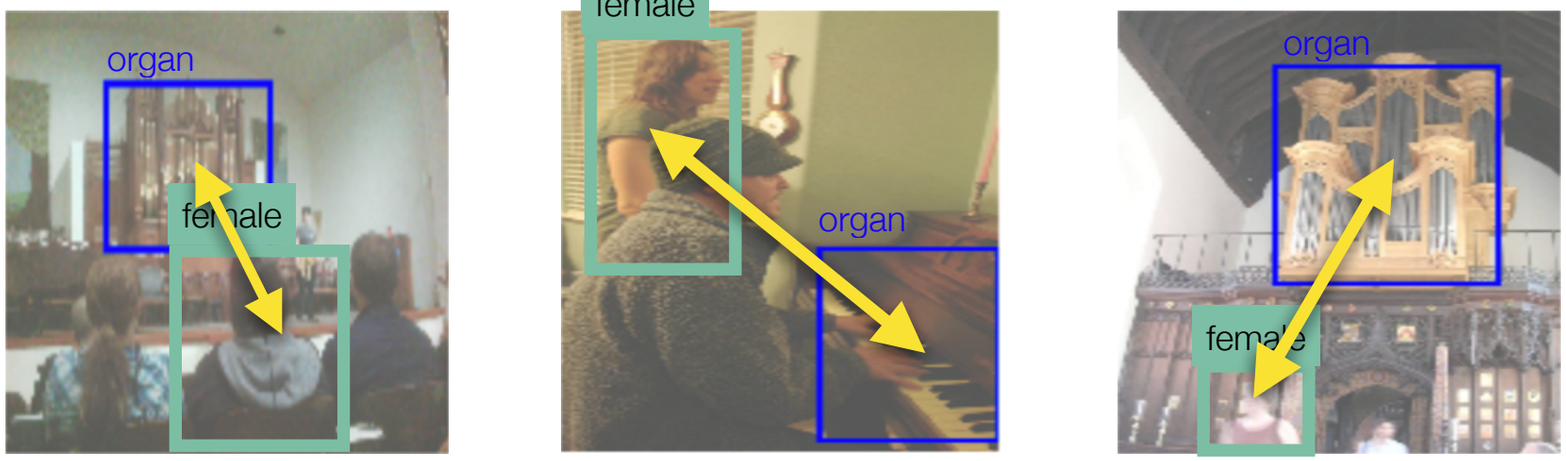
Actionable insight: collect images of the underrepresented gender with the corresponding objects and scenes

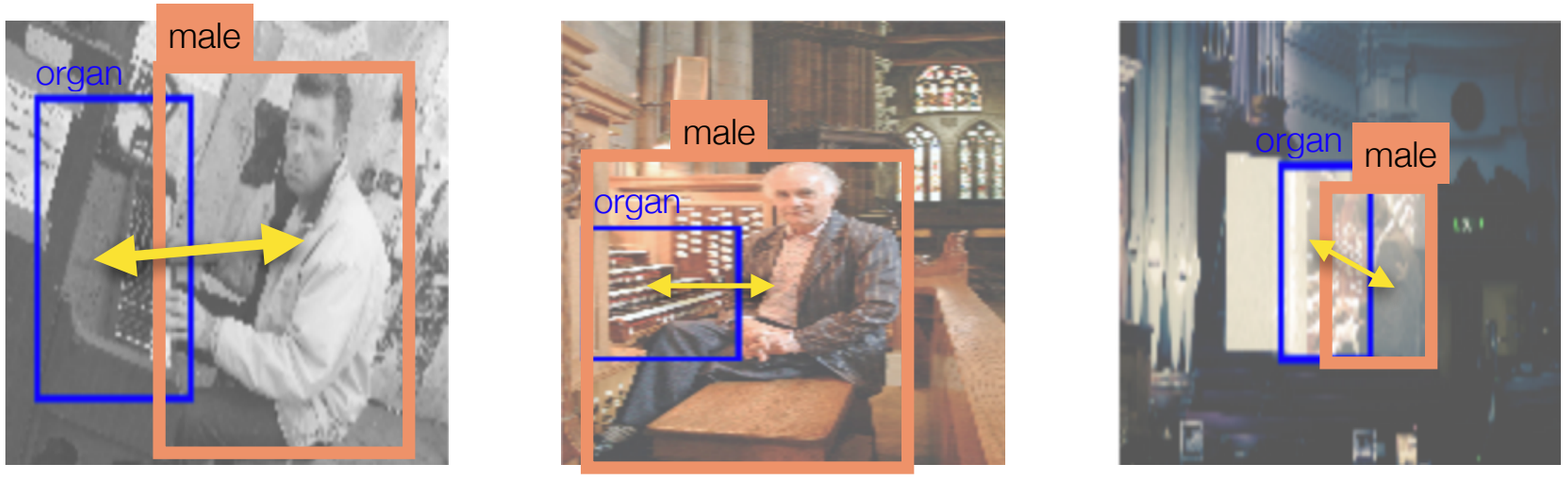
["REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets."
Angelina Wang, Arvind Narayanan, Olga Russakovsky. ECCV 2020 (spotlight). https://github.com/princetonvisualai/revise-tool]

Interaction between objects and people of different genders*

*Gender annotations of *binarized socially-perceived* gender *expression* derived from image captions [J. Zhao et al. EMNLP'17]

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*

- *Gender annotations of *binarized socially-perceived* gender *expression* derived from image captions [J. Zhao et al. EMNLP'17]

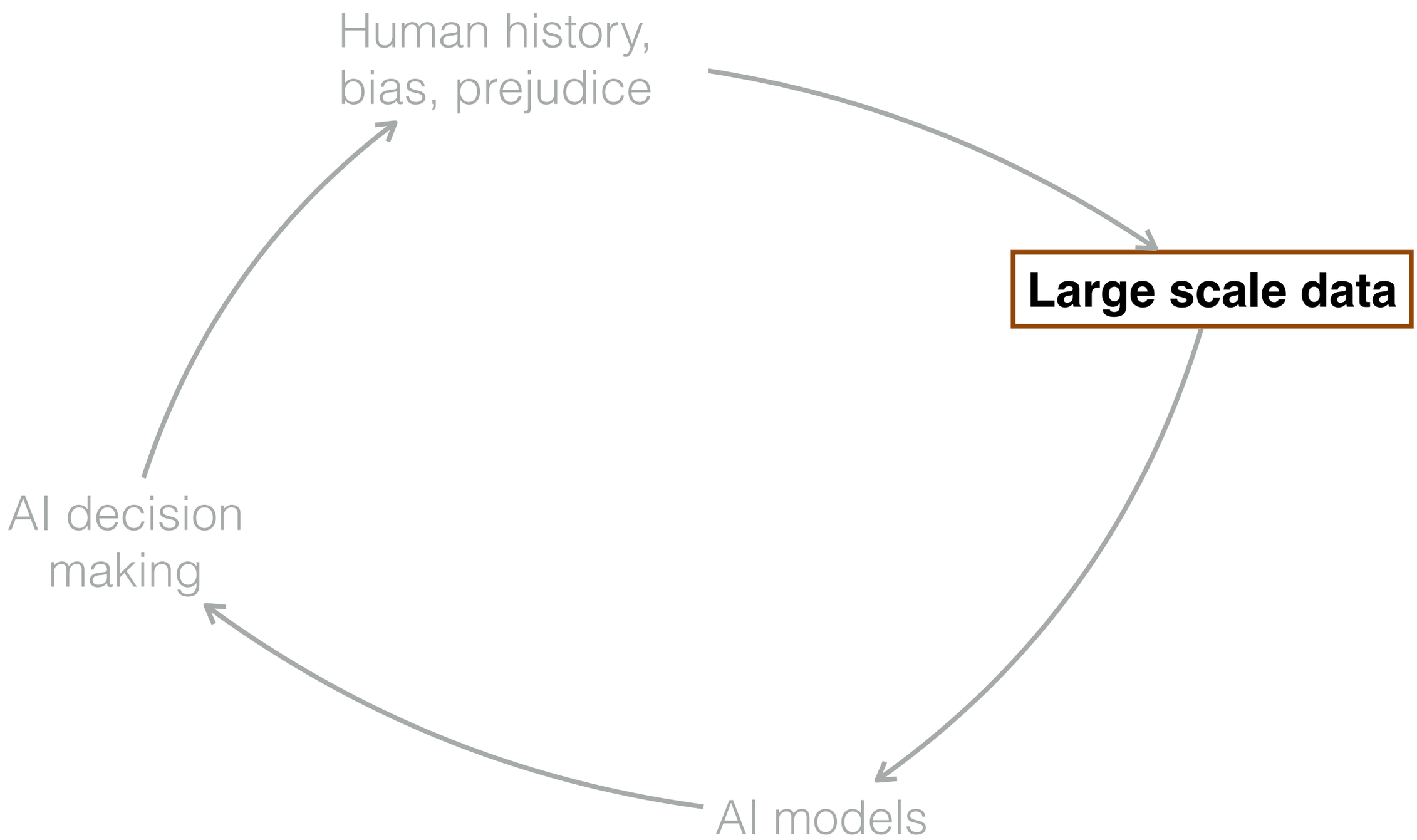
Analysis:

- 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]

Can we adjust the AI design to mitigate these effects?



Can we adjust the AI design to mitigate these effects?

Human history, bias, prejudice

Large scale data

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

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

AI models

AI decision making

Can we adjust the AI design to mitigate these effects?

Human history, bias, prejudice

Large scale data

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

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

AI models

AI decision making

Can we adjust the AI design to mitigate these effects?

Human history,
bias, prejudice

Large scale data

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

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

AI models

AI decision
making

Can we adjust the AI design to mitigate these effects?

Human history,
bias, prejudice

[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 models

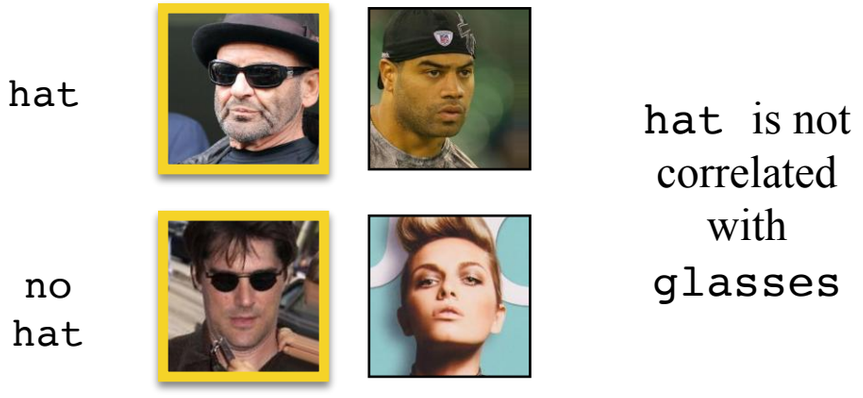
1) Fairness research presents unique challenges in visual recognition
2) Opportunities to both adapt existing techniques and innovate on them
3) Delicate tradeoff between accuracy and fairness

AI decision
making

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

Training: skewed distribution
(e.g., the target class is correlated with a protected attribute like gender or race)

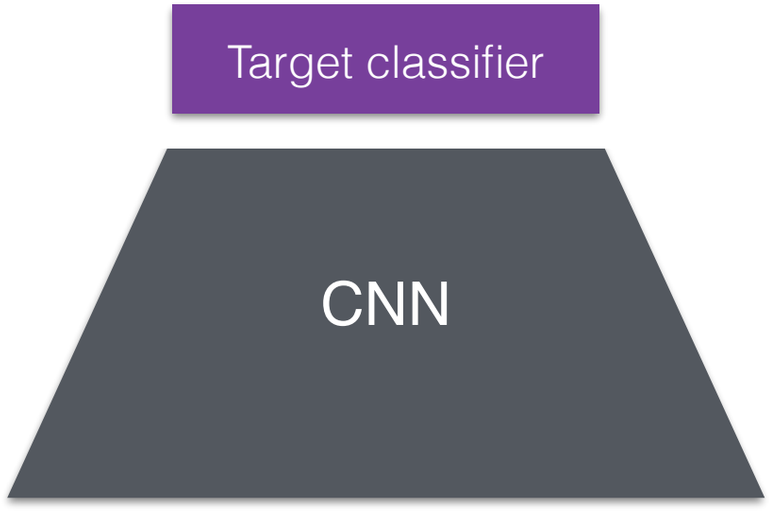
Testing: classifying images while ignoring this undesirable correlation



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

Baseline may learn this undesirable correlation

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



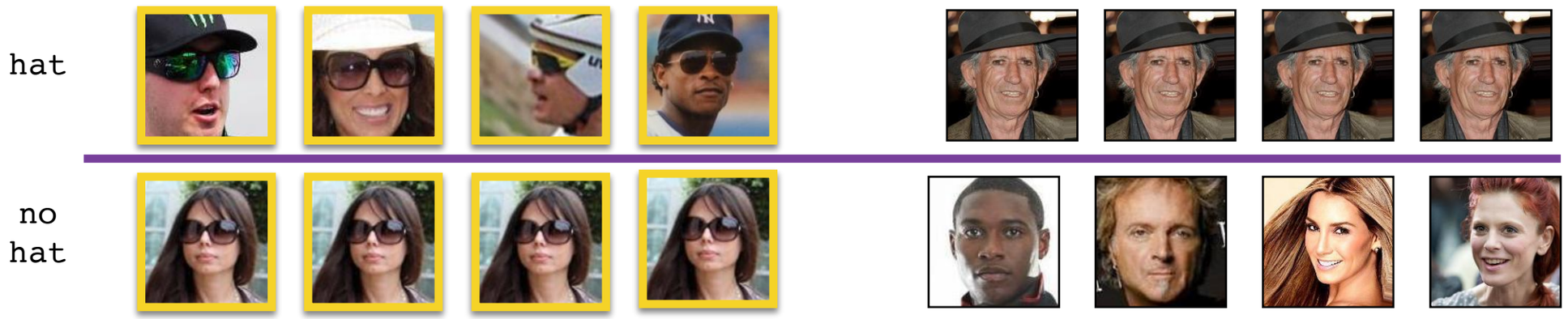
x_i = image i

y_i = target class for image i

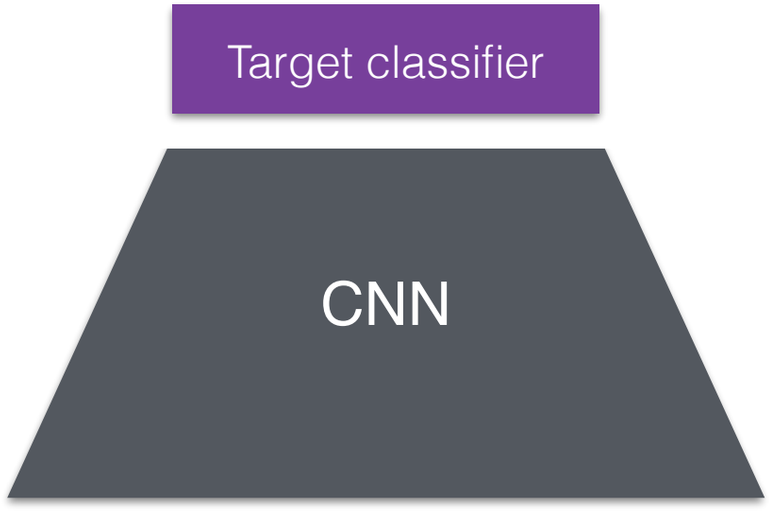
- Will learn the most discriminative feature(s) in the training dataset

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

(1) Over/under-sampling may not work well



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



x_i = image i

y_i = target class for image i

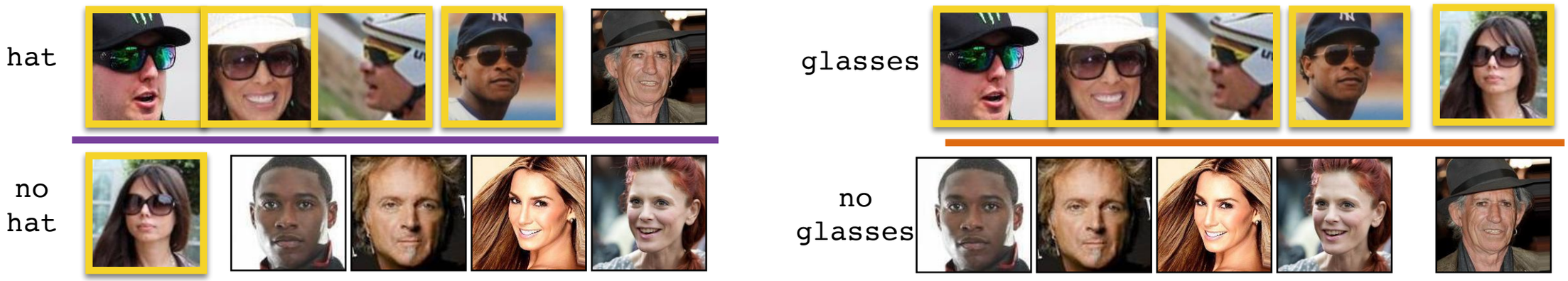
- Undersampling discards data
- Oversampling/reweighting is prone to overfitting and numerical instability

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

(2) Adversarial de-biasing does not work

Want to classify the target class

But not be able to classify the protected attribute



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

Target classifier

Protected classifier

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

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

x_i = image i

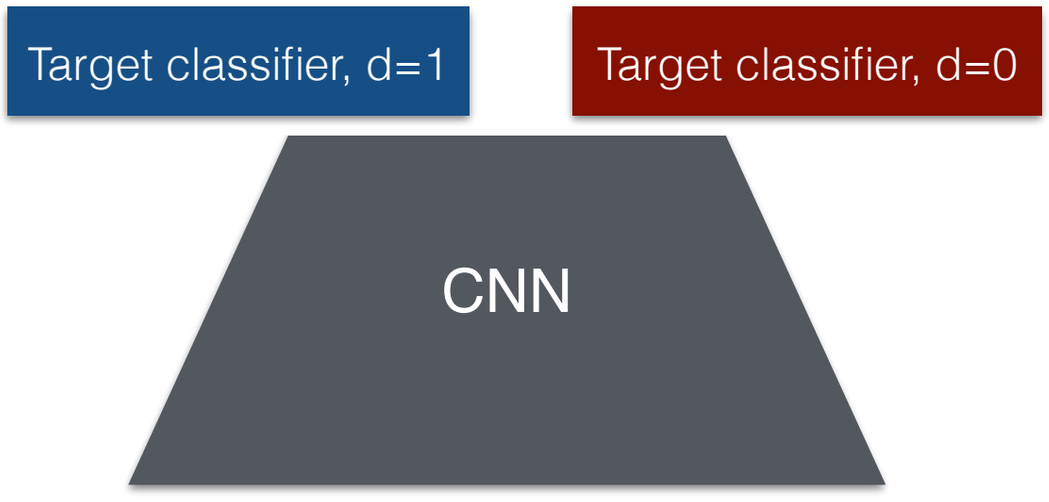
y_i = target class for image i

d_i = protected class for image i

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

(3) 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 = target class for image i

d_i = protected attribute for image i

Inference:

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

s = class score

- Works especially well in settings with high skew (high correlation between target and protected attributes)

[Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, Olga Russakovsky. "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation." 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
Oversampling	89.1 ± 0.4	78.6 ± 0.4	**77.6**
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

Data: Modified CIFAR-10 [Krizhevsky et al. 2009]

Target: 10-way object classification

Protected attribute: color/grayscale

Architecture: ResNet-18

Metric: Mean per-class per-attribute accuracy

Data: Modified CIFAR-10 + ImageNet [Deng et al. CVPR 2009]

Target: 10-way object classification

Protected attribute: CIFAR vs ImageNet distribution

Architecture: ResNet-18

Metric: Mean per-class per-attribute accuracy

Data: CelebA dataset [Liu et al. ICCV 2015]

Target: multi-label face attribute recognition

Protected attribute: gender

Architecture: ResNet-50, pre-trained

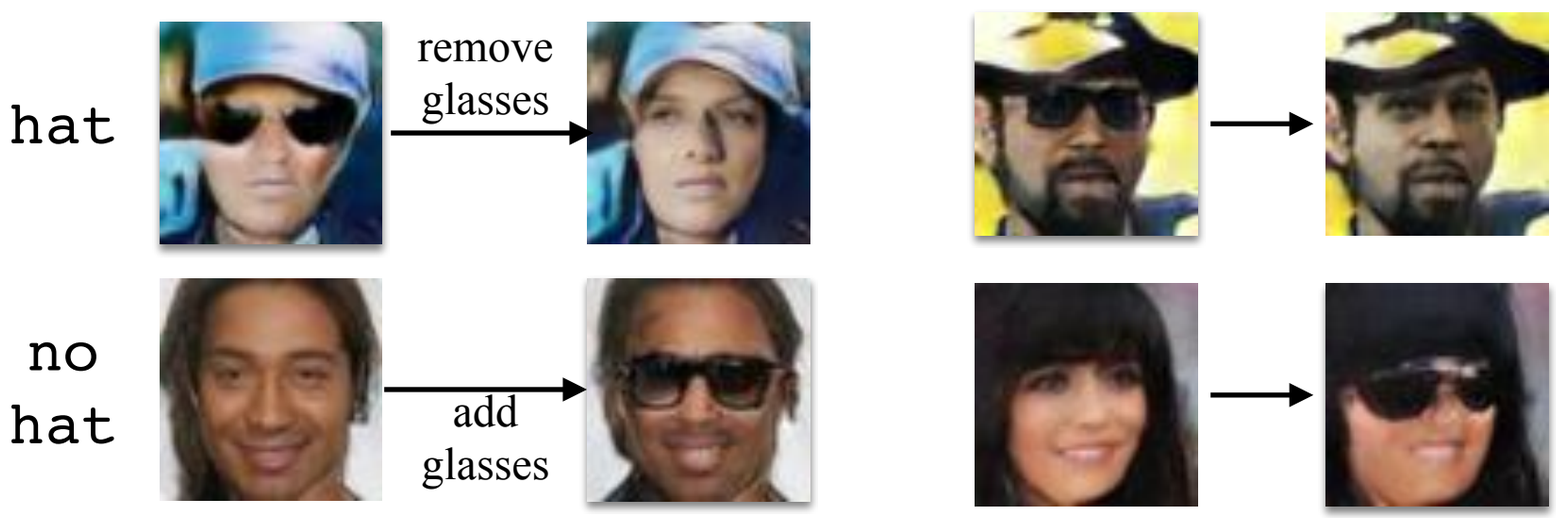
Metric: Weighted mean average precision

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

Can we *generate* the data we need?

Real-world data

Paired augmentation

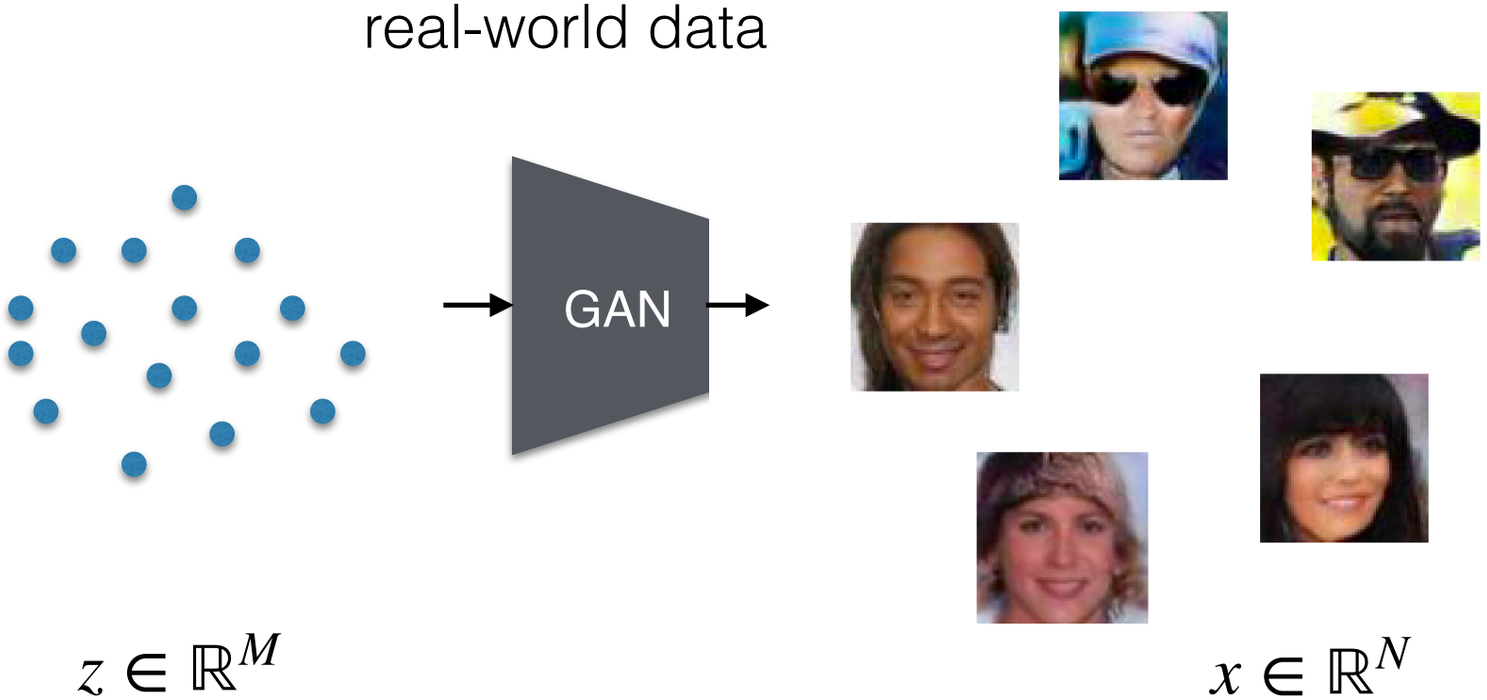


[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

[I. Goodfellow et al. NeurIPS 2014]
[T. Karras et al. ICLR 2018]

GAN
trained on
real-world data

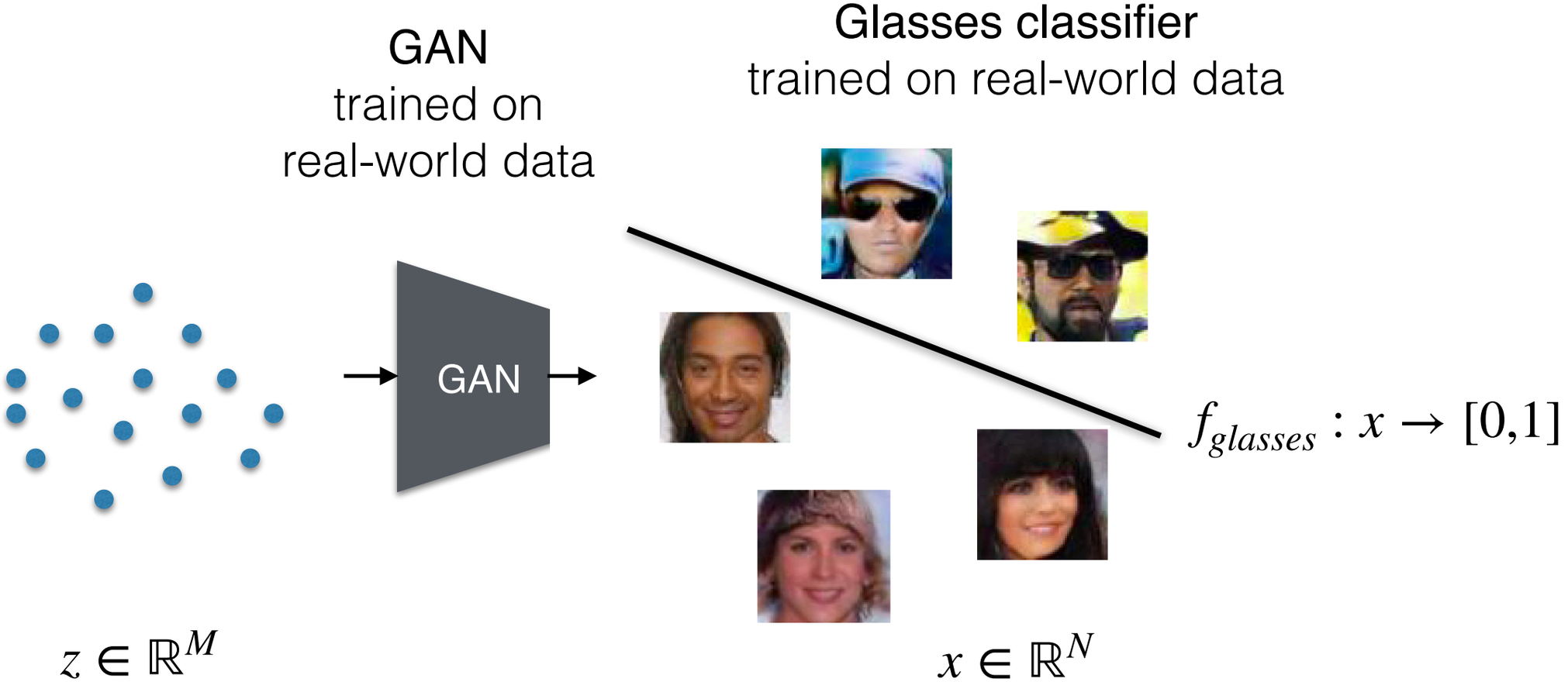


$z \in \mathbb{R}^M$

$x \in \mathbb{R}^N$

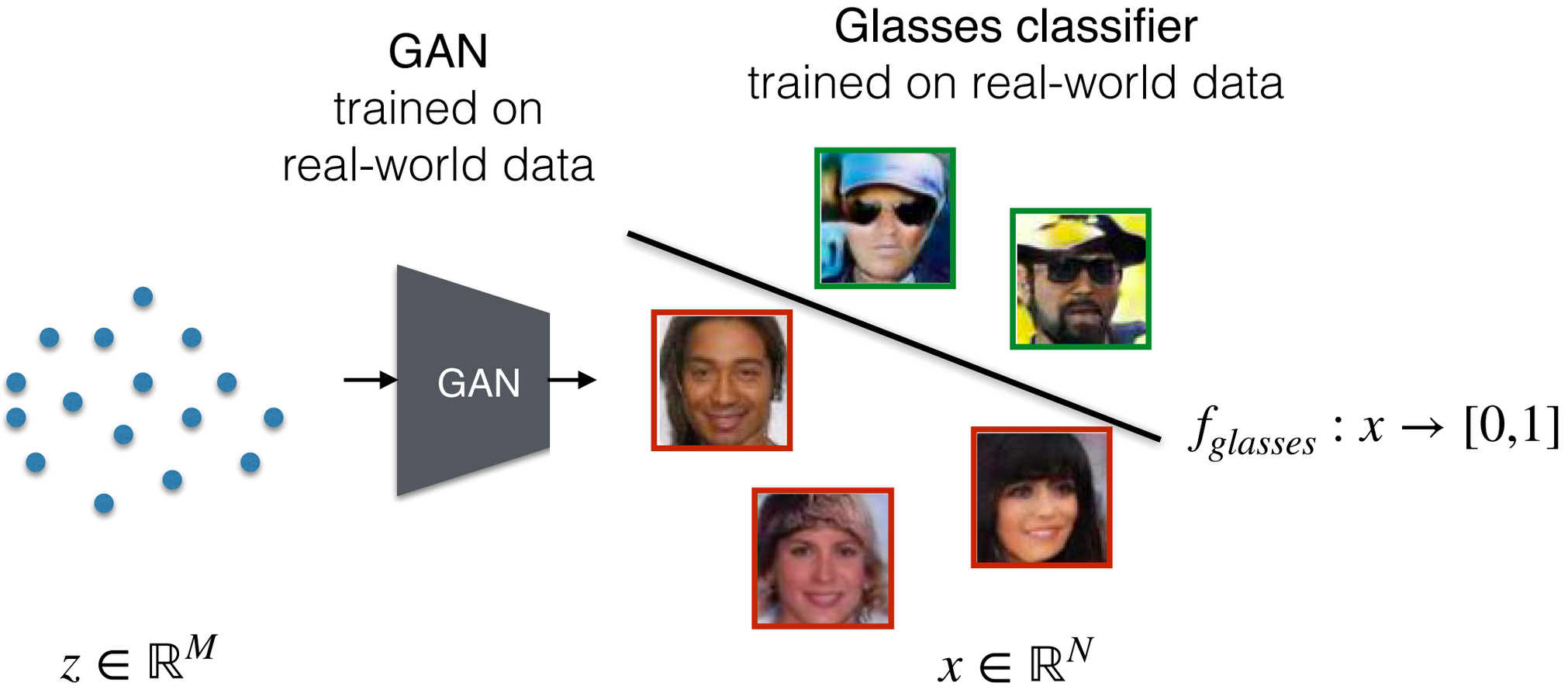
[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes



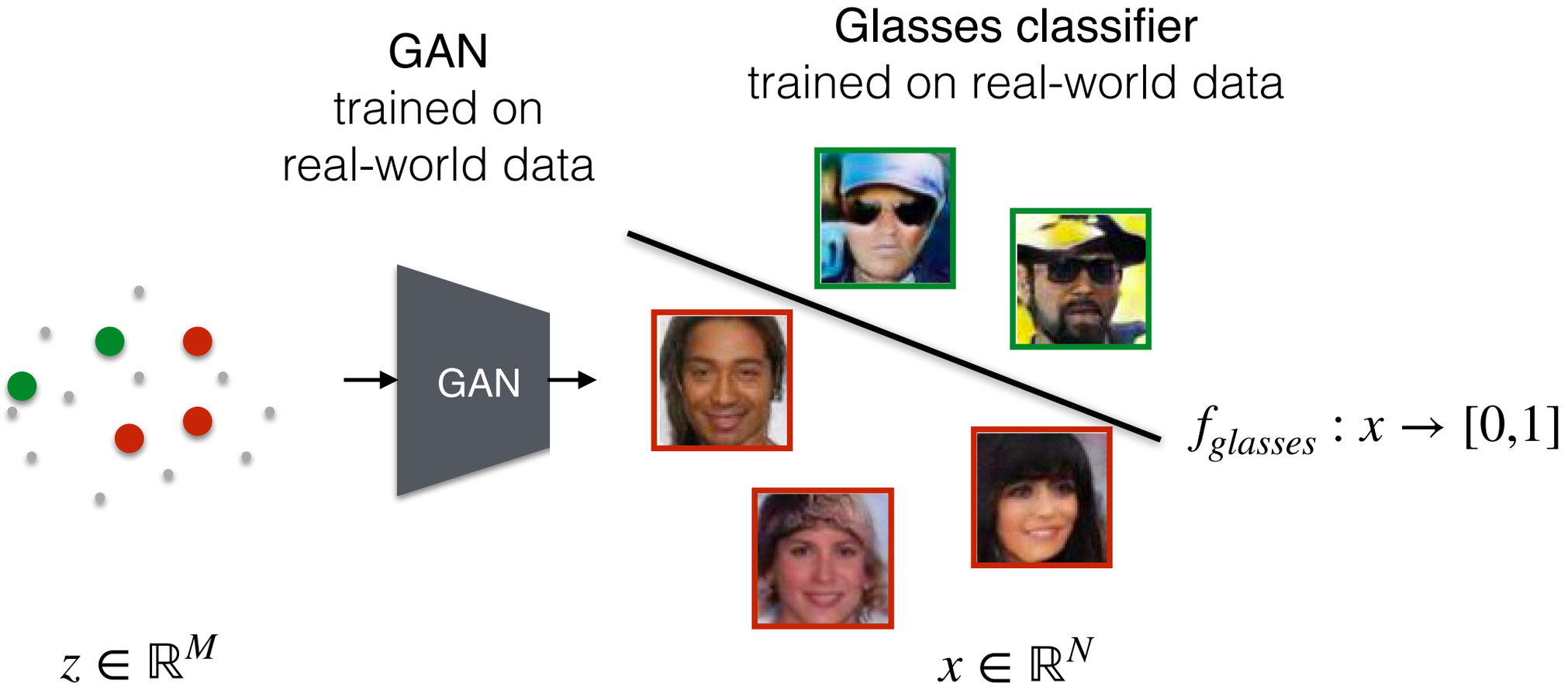
[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes



[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

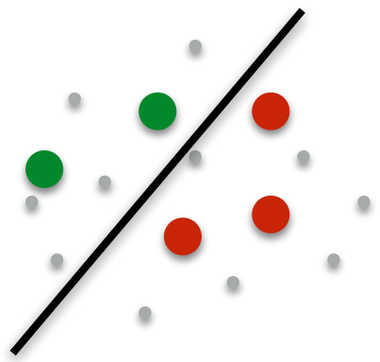
Generating images with de-correlated attributes



[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

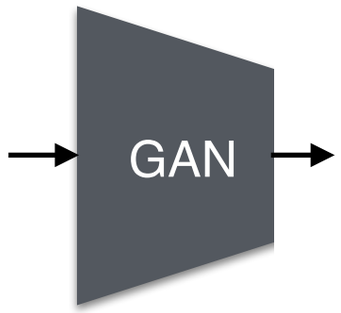
Allows us to train a **glasses classifier** in the latent space



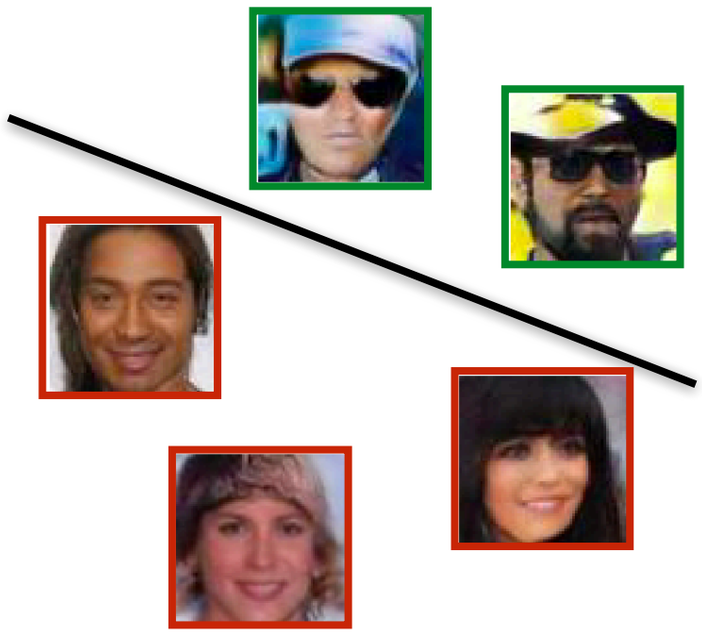
$h_{glasses} : z \rightarrow [0,1]$

$z \in \mathbb{R}^M$

GAN
trained on
real-world data



Glasses classifier
trained on real-world data



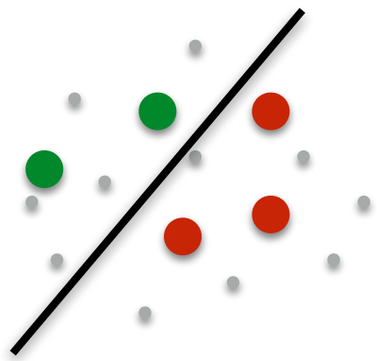
$f_{glasses} : x \rightarrow [0,1]$

$x \in \mathbb{R}^N$

[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

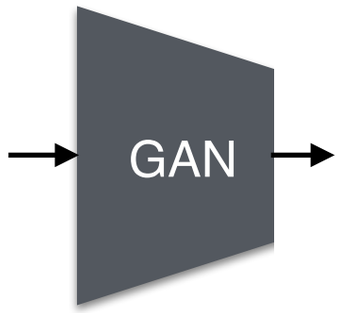
Allows us to train a **glasses classifier** in the latent space



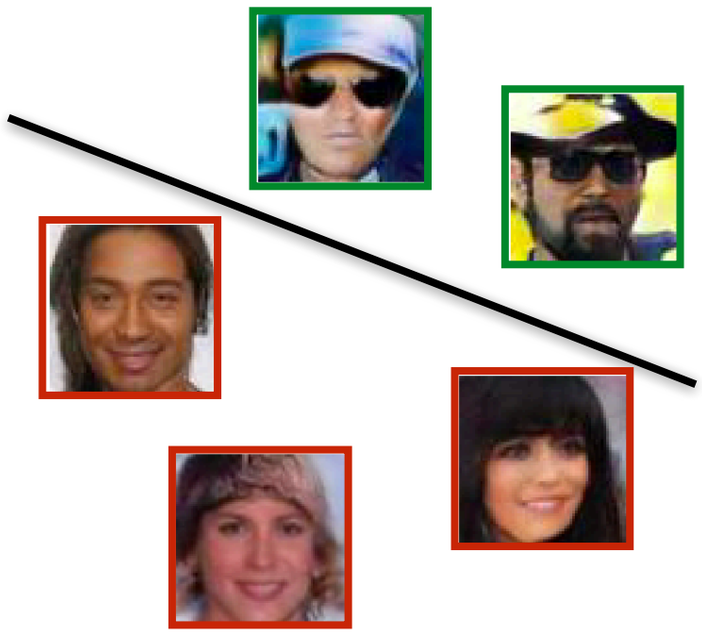
$h_{glasses} : z \rightarrow [0,1]$

$z \in \mathbb{R}^M$

GAN
trained on
real-world data



Glasses classifier
trained on real-world data



$f_{glasses} : x \rightarrow [0,1]$

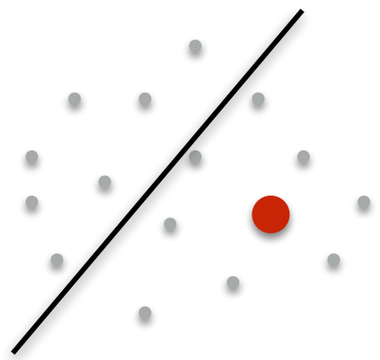
$x \in \mathbb{R}^N$

[E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. “Image counterfactual sensitivity analysis for detecting unintended bias.” In CVPRW 2019.]

[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
“Fair Attribute Classification through Latent Space De-biasing.” https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

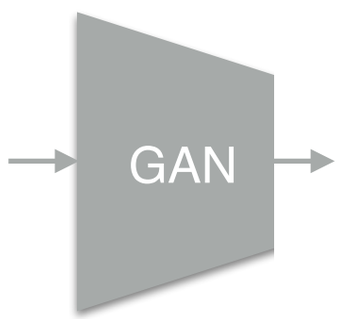
Allows us to train a **glasses classifier** in the latent space



$h_{glasses} : z \rightarrow [0,1]$

$z \in \mathbb{R}^M$

GAN
trained on
real-world data



Glasses classifier
trained on real-world data

$f_{glasses} : x \rightarrow [0,1]$

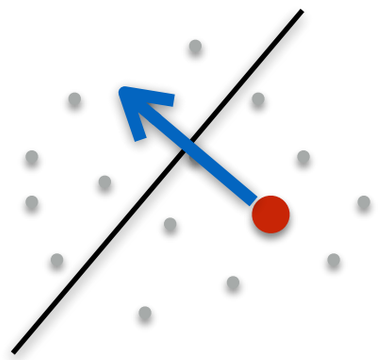
$x \in \mathbb{R}^N$

[E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. “Image counterfactual sensitivity analysis for detecting unintended bias.” In CVPRW 2019.]

[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
“Fair Attribute Classification through Latent Space De-biasing.” https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

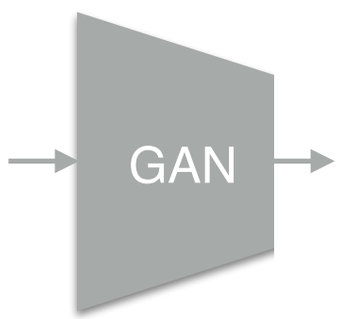
Allows us to train a **glasses classifier** in the latent space



$h_{glasses} : z \rightarrow [0,1]$

$z \in \mathbb{R}^M$

GAN
trained on
real-world data



Glasses classifier
trained on real-world data

$f_{glasses} : x \rightarrow [0,1]$

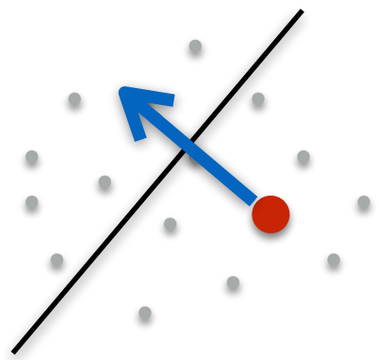
$x \in \mathbb{R}^N$

[E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. “Image counterfactual sensitivity analysis for detecting unintended bias.” In CVPRW 2019.]

[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
“Fair Attribute Classification through Latent Space De-biasing.” https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

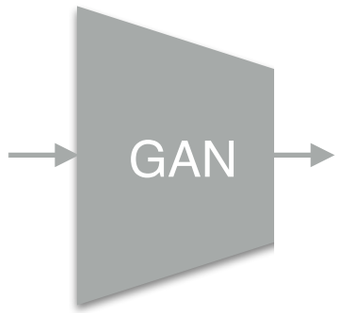
Allows us to train a **glasses classifier** in the latent space



$h_{glasses} : z \rightarrow [0,1]$

$z \in \mathbb{R}^M$

GAN
trained on
real-world data



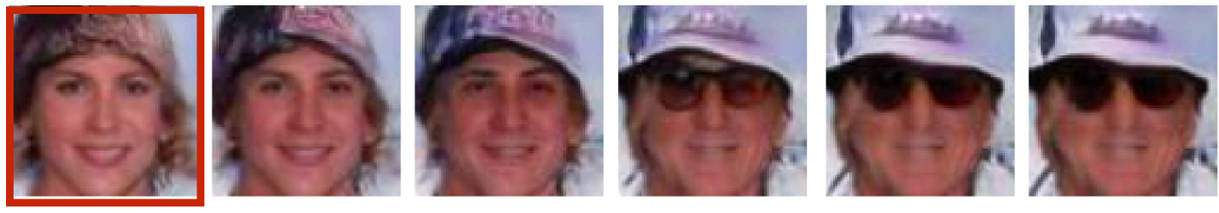
Glasses classifier
trained on real-world data

$f_{glasses} : x \rightarrow [0,1]$

$x \in \mathbb{R}^N$

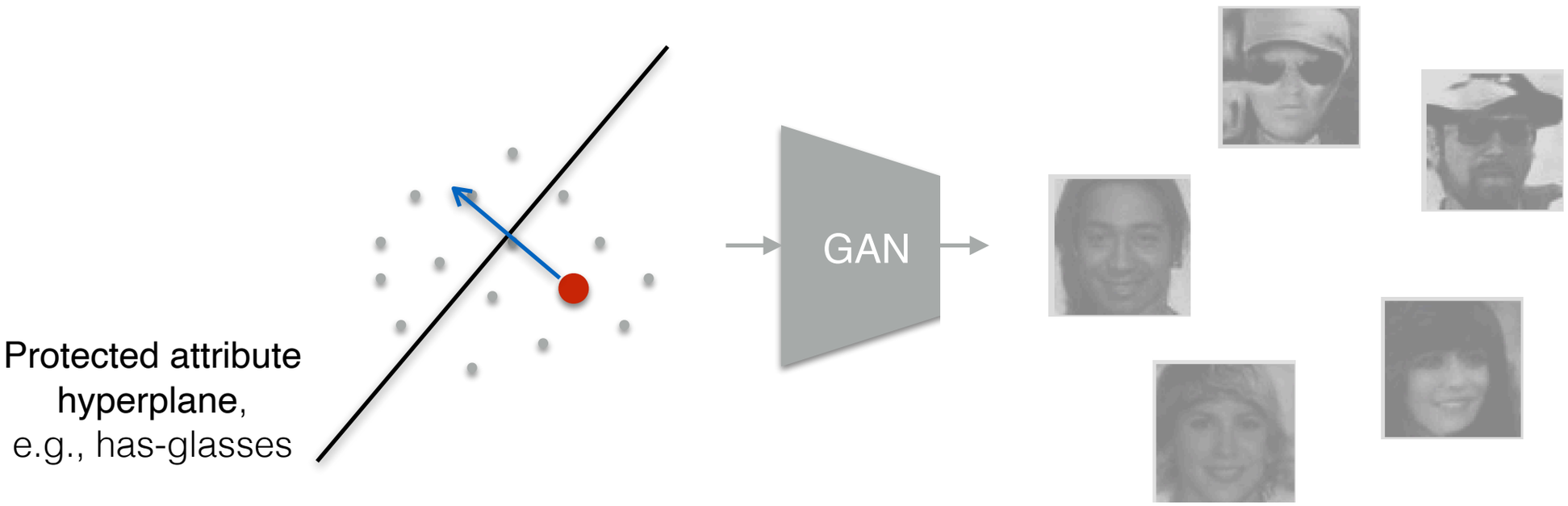
[E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. "Image counterfactual sensitivity analysis for detecting unintended bias." In CVPRW 2019.]

Generated:

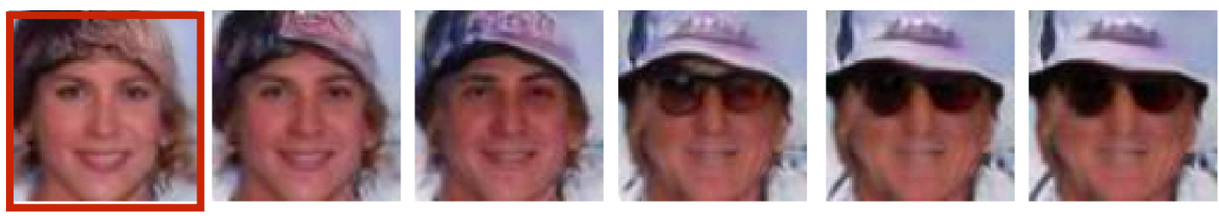


[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

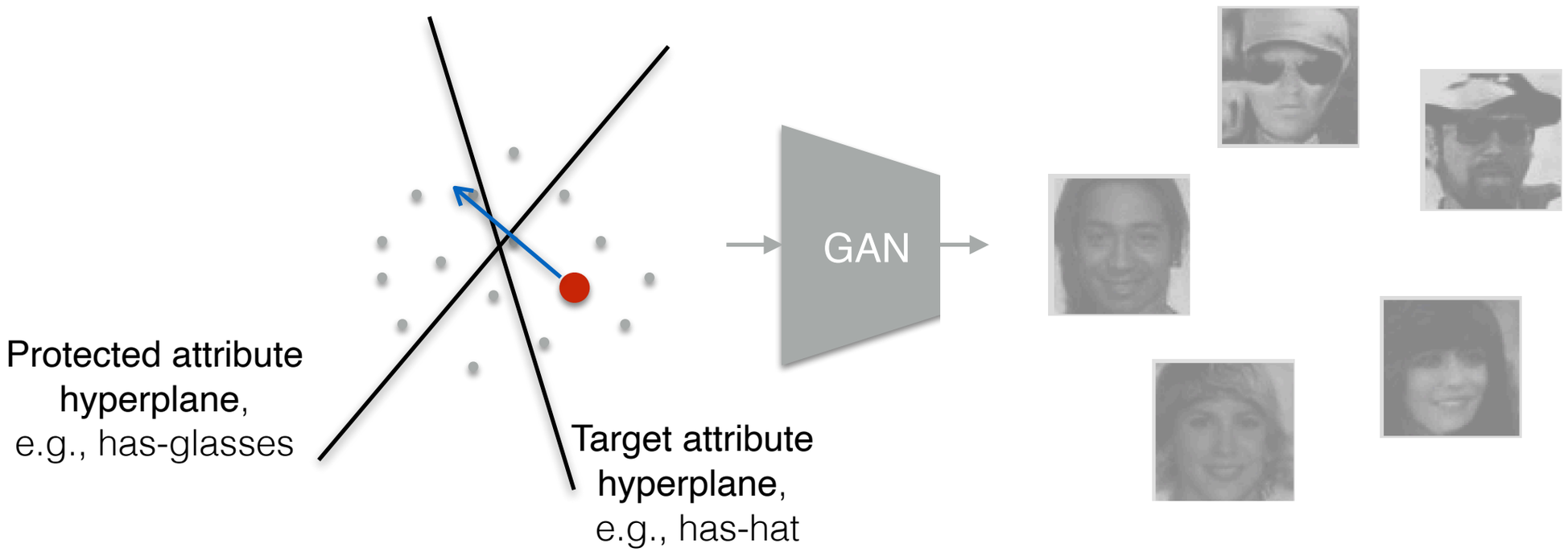


Original:

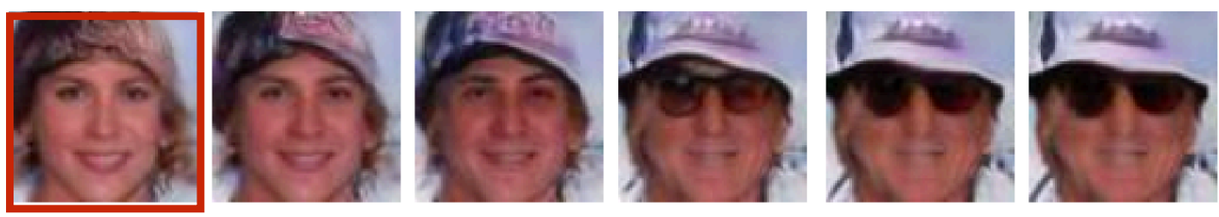


[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

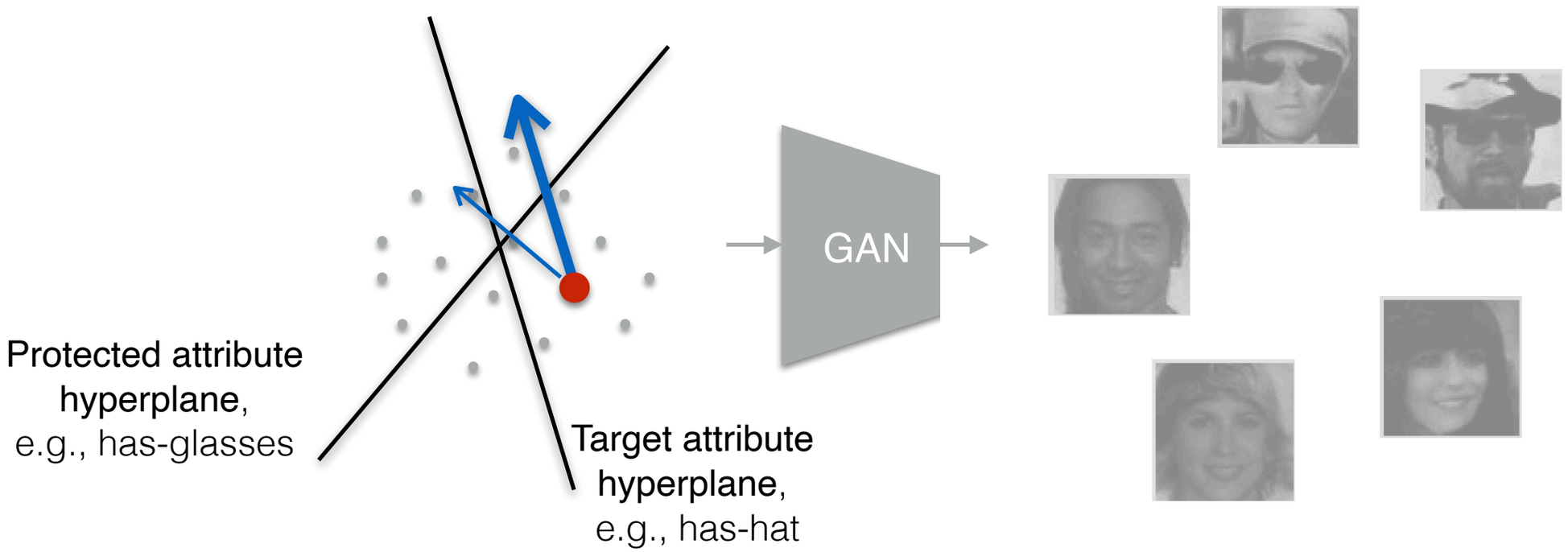


Original:

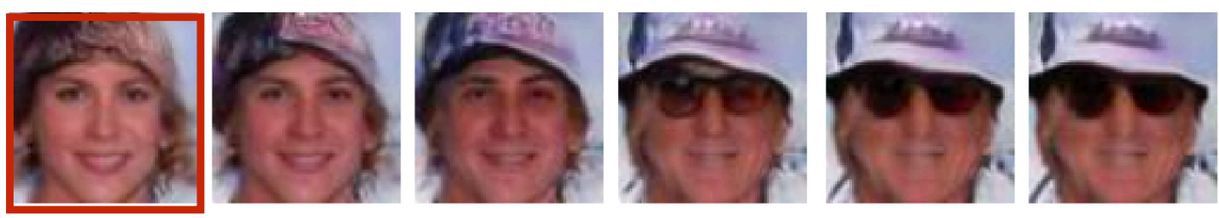


[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

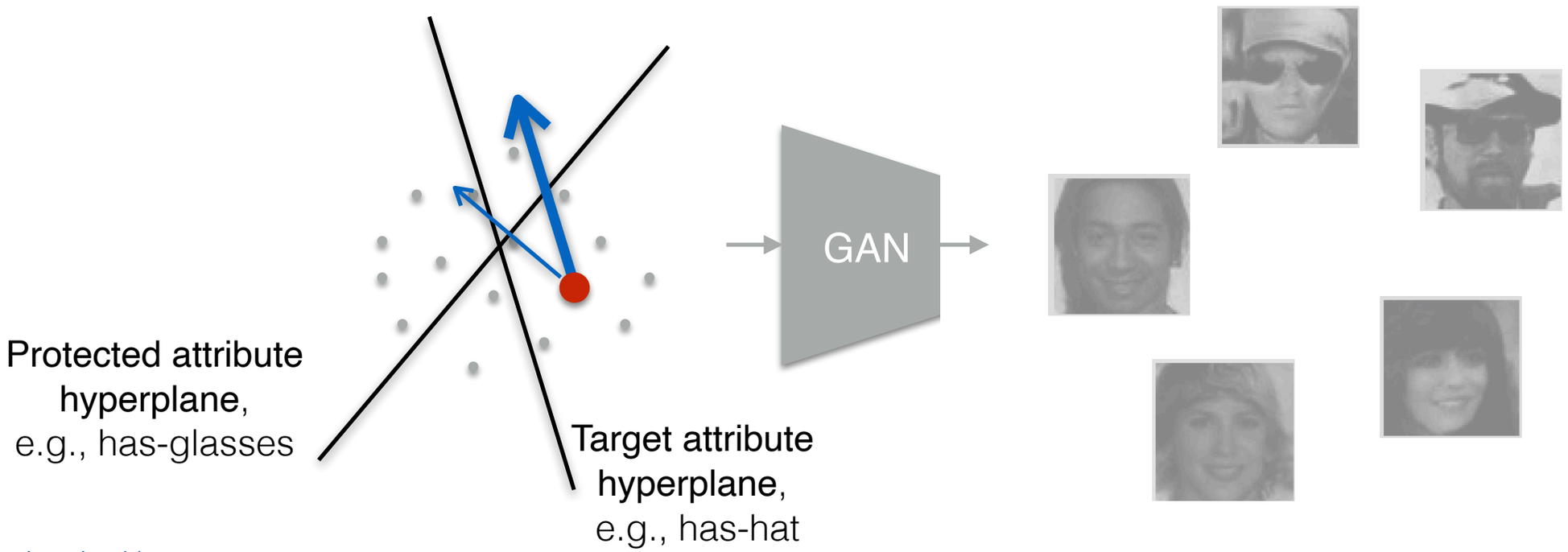


Original:



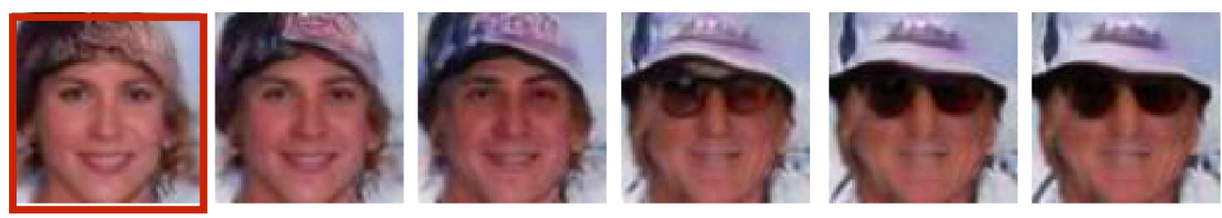
[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes



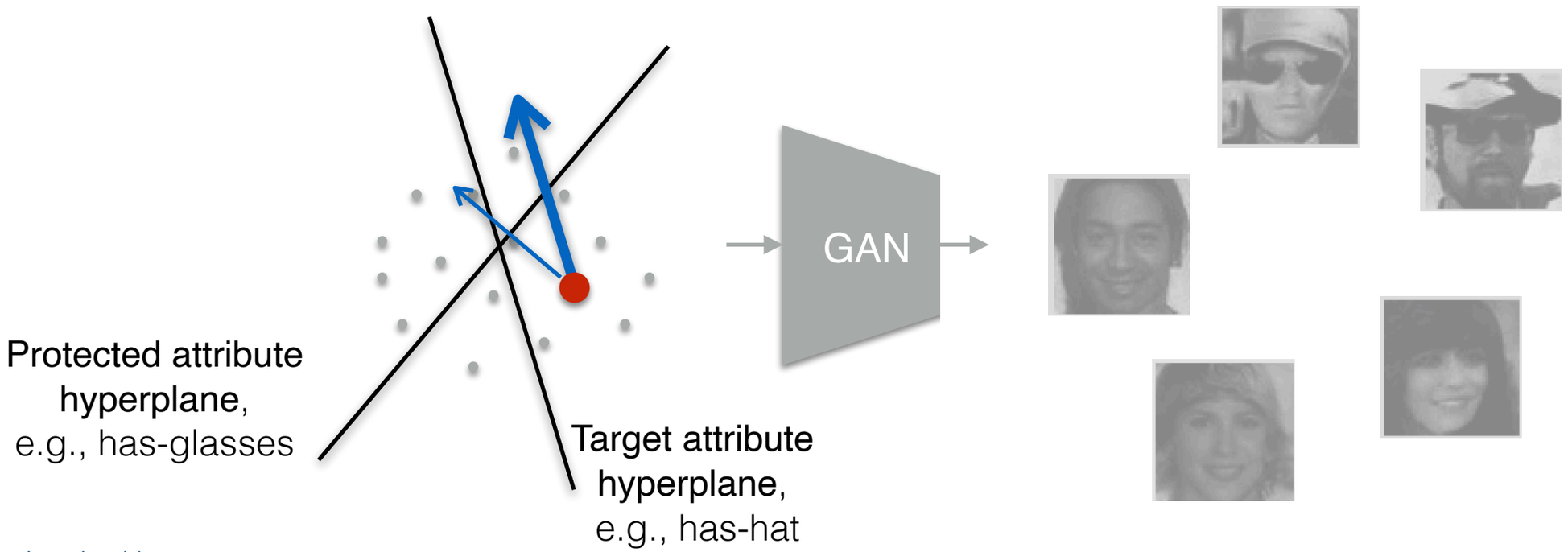
Inspired by:
[B. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In Conference on AI, Ethics, and Society (AIES), 2018]

Original:

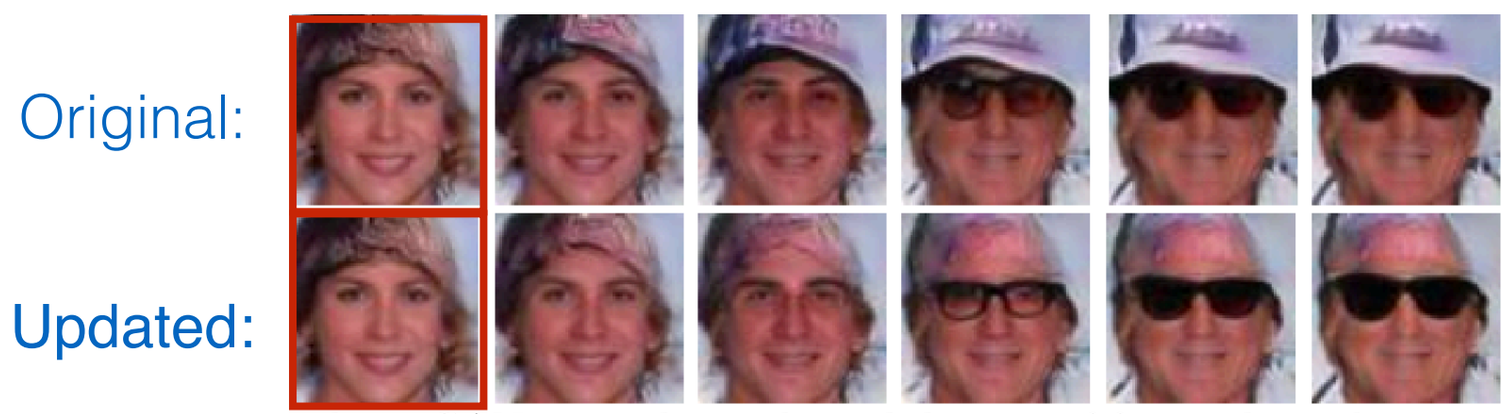


[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

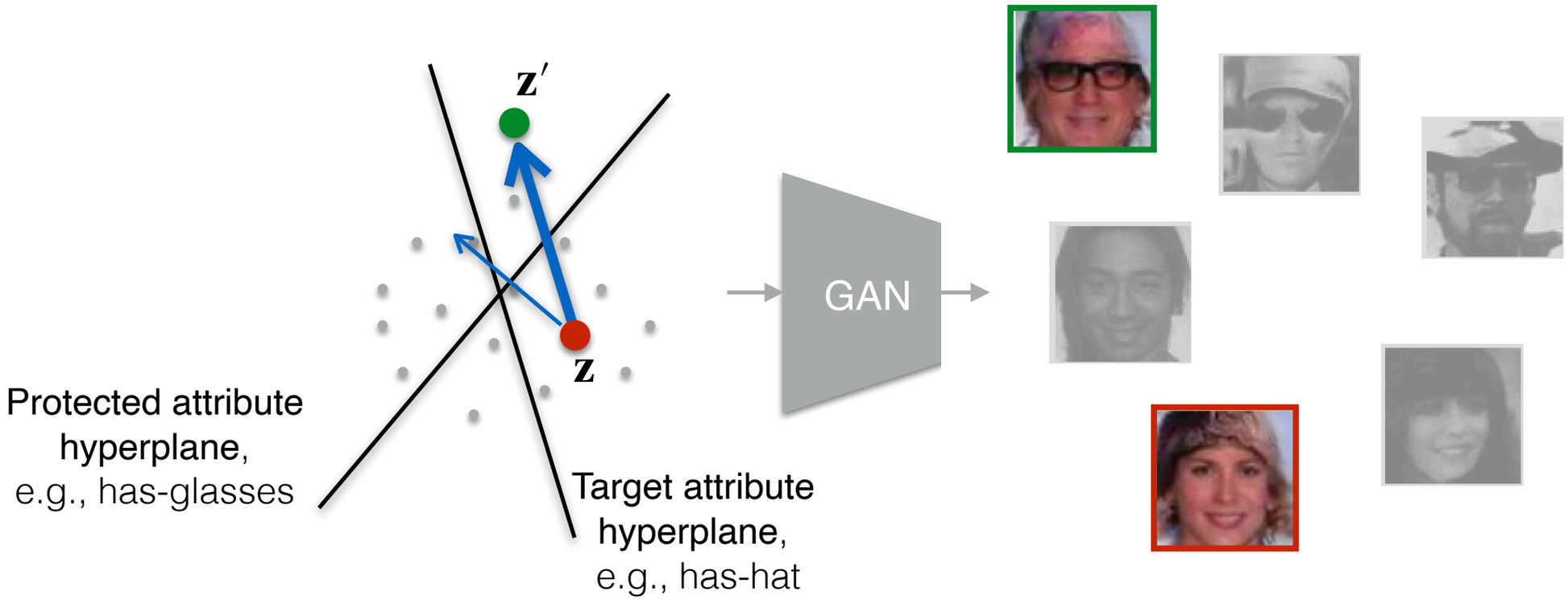


Inspired by:
[B. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In Conference on AI, Ethics, and Society (AIES), 2018]



[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes



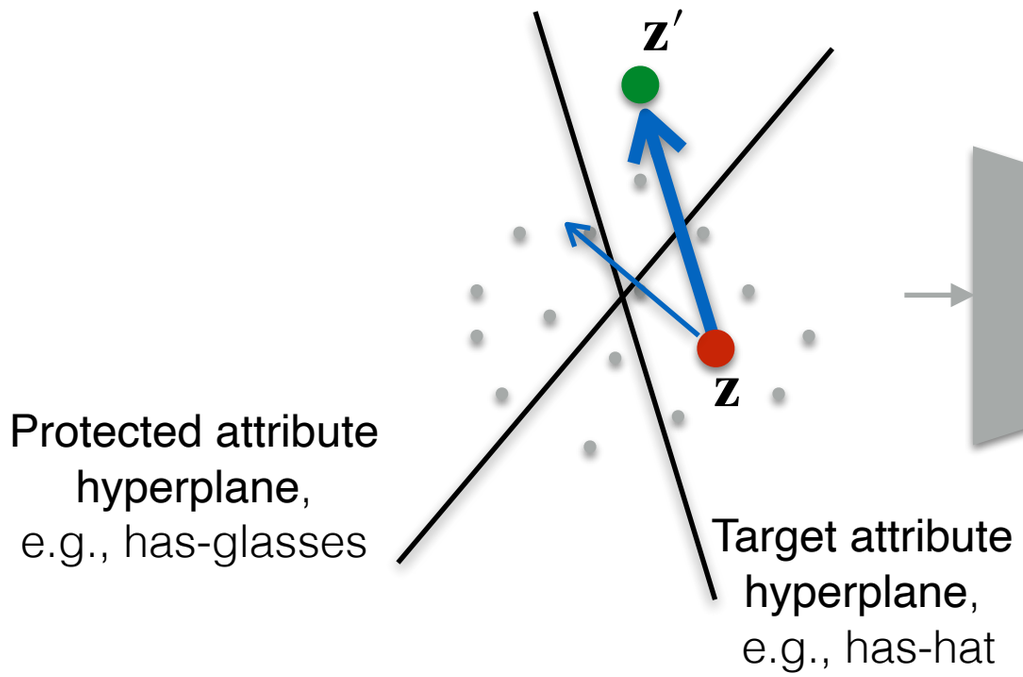
Original:

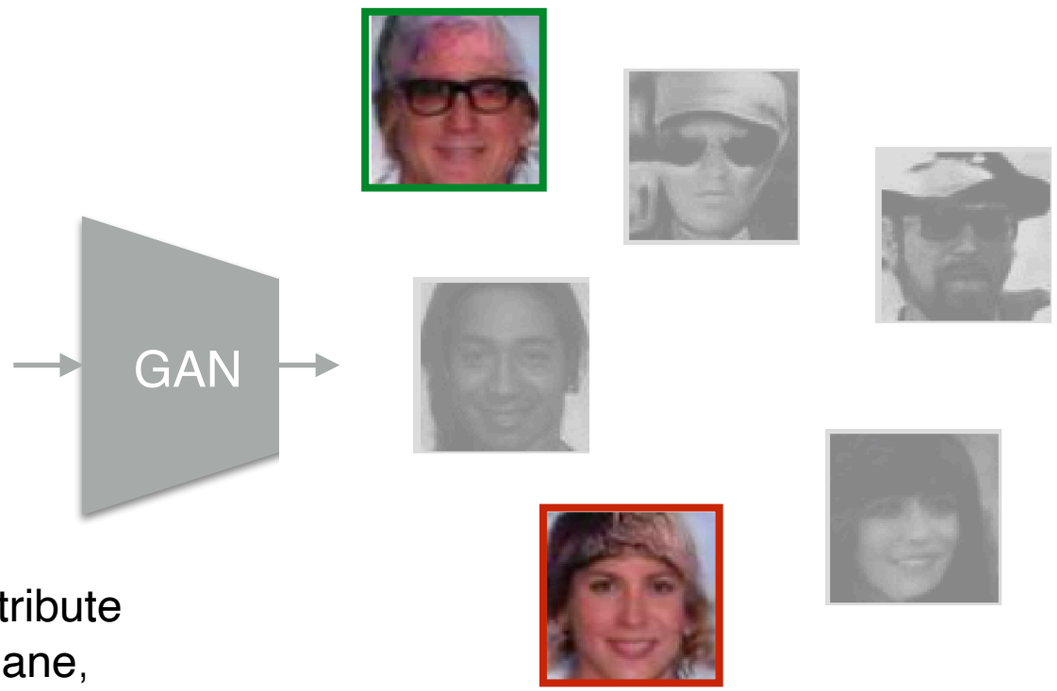
Updated:

[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Generating images with de-correlated attributes

$$\mathbf{z}' = \mathbf{z} - 2\left(\frac{\mathbf{w_p}\mathbf{z} + b_p}{1 - (\mathbf{w_p}^T\mathbf{w_t})^2}\right)\left(\mathbf{w_p} - (\mathbf{w_p}^T\mathbf{w_t})\mathbf{w_t}\right)$$





Original:

Updated:

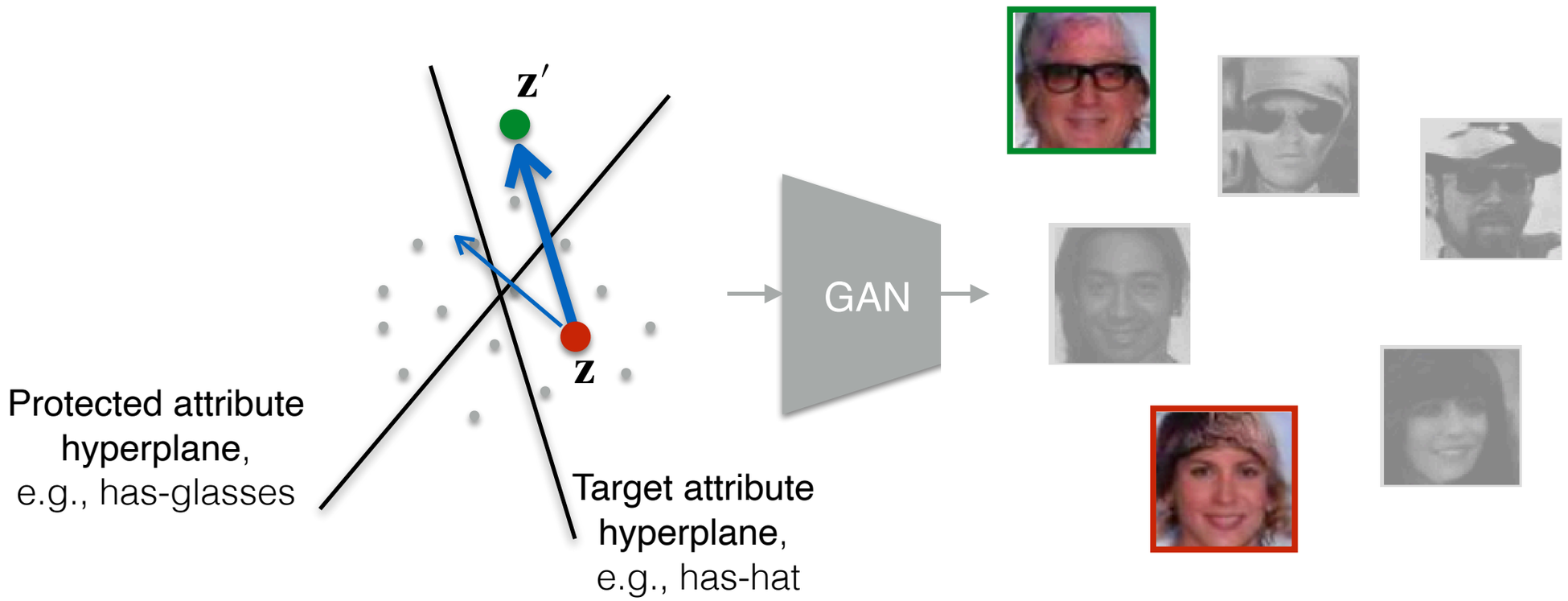
[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

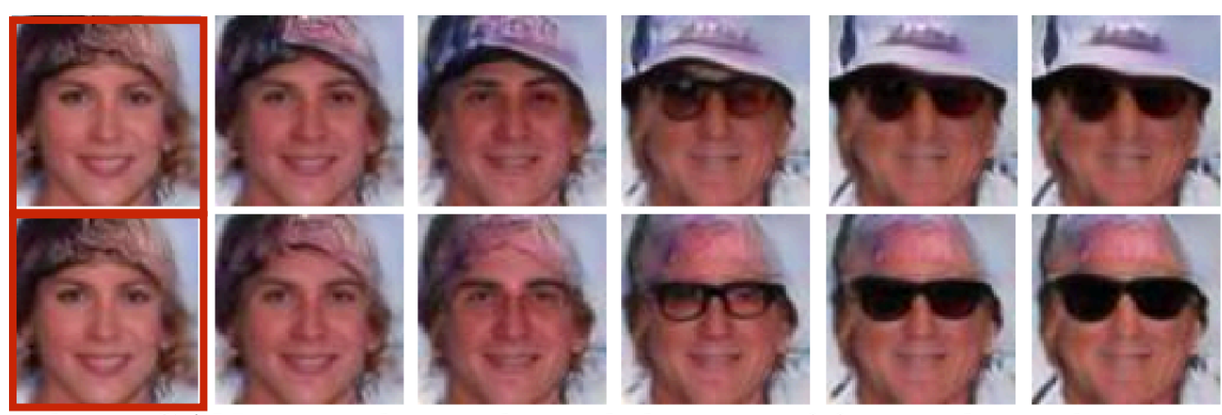
Pros:

1. Single GAN for multiple target/protected attributes
2. Preserves intra-class variation
3. Ensures independence between attributes in the generated data
4. Can be generalized to multiple protected attributes

Cons:

1. The classifiers need to be at least reasonable
2. Might add other correlations
3. GANs are tricky…





[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Evaluating classifiers trained with our generated data

Setup: CelebA dataset, protected attribute = gender, target attributes = 26 other face attributes (e.g., has-hat)

vs baseline:
much better on all 3 fairness metrics

Attr. type	Model	DEO ↓	BA ↓	KL ↓
Incons.	Baseline	21.5 ± 4.4	2.1 ± 0.6	1.7 ± 0.3
	Ours	$\mathbf{16.5 \pm 4.2}$	$\mathbf{0.5 \pm 0.6}$	$\mathbf{1.3 \pm 0.4}$
G-dep.	Baseline	25.7 ± 3.5	2.3 ± 0.5	1.3 ± 0.2
	Ours	$\mathbf{23.4 \pm 3.6}$	$\mathbf{1.6 \pm 0.5}$	$\mathbf{1.2 \pm 0.2}$
G-indep.	Baseline	16.7 ± 5.0	0.3 ± 0.6	1.1 ± 0.5
	Ours	$\mathbf{13.9 \pm 5.2}$	$\mathbf{0.0 \pm 0.5}$	$\mathbf{0.9 \pm 0.6}$

vs weighted and adversarial:
generally better than adversarial;
not as good as weighted
(but weighted strongly amplifies bias in the negative direction)

Method	DEO ↓	BA ↓	KL ↓
Weighted	$\mathbf{5.7 \pm 4.2}$	$\mathbf{-2.8 \pm 0.5}$	$\mathbf{0.5 \pm 0.4}$
Adversarial	23.9 ± 4.4	1.5 ± 0.5	0.6 ± 0.5
Ours	16.7 ± 4.7	0.5 ± 0.5	1.0 ± 0.5

vs domain-independent: better when skew is low, but domain-independent still best when skew is high

Skew	Method	DEO ↓	BA ↓	KL ↓
Low/ Mod.	Dom. Ind.	7.0 ± 3.1	$\mathbf{-0.1 \pm 0.5}$	0.8 ± 0.7
	Ours	$\mathbf{6.0 \pm 3.0}$	$\mathbf{-0.1 \pm 0.5}$	$\mathbf{0.3 \pm 0.1}$
High	Dom. Ind.	$\mathbf{14.9 \pm 5.6}$	$\mathbf{-0.4 \pm 0.5}$	$\mathbf{0.8 \pm 1.0}$
	Ours	23.9 ± 5.5	0.9 ± 0.4	1.5 ± 0.6

vs others like Fairness GAN: better

	Fairness GAN [35]				Ours (Synthetic only)	
	Dem. Par.		Eq. Opp.			
Gender exp. g	$g=-1$	$g=1$	$g=-1$	$g=1$	$g=-1$	$g=1$
FPR ↓	0.52	0.26	0.42	**0.17**	**0.22**	0.39
FNR ↓	0.18	0.41	0.21	0.44	**0.06**	**0.27**
Error ↓	0.30	0.28	0.29	0.23	**0.21**	**0.18**
Error Rate ↓	0.22		0.29		**0.20**	

[P. Sattigeri, et al. IBM Journal of Research and Development, 2019]

[Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky.
"Fair Attribute Classification through Latent Space De-biasing." https://github.com/princetonvisualai/gan-debiasing]

Can we adjust the AI design to mitigate these effects?

Human history,
bias, prejudice

Large scale data

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

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

AI models

1) Fairness research presents unique challenges in visual recognition
2) Opportunities to both adapt existing techniques and innovate on them
3) Delicate tradeoff between accuracy and fairness

[Z. Wang et al. CVPR2020.
https://github.com/princetonvisualai/DomainBiasMitigation]
[V. Ramaswamy et al.
https://github.com/princetonvisualai/gan-debiasing]

AI decision
making

Can we adjust the AI design to mitigate these effects?

Human history,
bias, prejudice

AI will change the world.
Who will change AI?

AI decision
making

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

[Z. Wang et al. CVPR2020.
https://github.com/princetonvisualai/DomainBiasMitigation]
[V. Ramaswamy et al.
https://github.com/princetonvisualai/gan-debiasing]

Large scale data

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

AI models

1) Fairness research presents unique challenges in visual recognition
2) Opportunities to both adapt existing techniques and innovate on them
3) Delicate tradeoff between accuracy and fairness

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

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

- AI4ALL 2016 student

- Our goal is to educate and support a diverse next generation of AI ***leaders***
- 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
- Providing life-long training, mentorship, opportunities and community for our alumni
  - Amazingly, many alumni have **taken on AI leadership roles while still in high school/early college** by starting research projects, working groups, panels, clubs, outreach events, …
  - https://medium.com/ai4allorg/alumni/

http://ai-4-all.org

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

Can we adjust the AI design to mitigate these effects?

Human history, bias, prejudice

AI will change the world. Who will change AI?

http://ai-4-all.org

AI decision making

Large scale data

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

[Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. **FAT*2020**. http://image-net.org/filtering-and-balancing]

[Angelina Wang, Arvind Narayanan, and Olga Russakovsky. **ECCV2020**. https://github.com/princetonvisualai/revise-tool]

AI models

1) Fairness research presents unique challenges in visual recognition
2) Opportunities to both adapt existing techniques and innovate on them
3) Delicate tradeoff between accuracy and fairness

[Zeyu Wang, Klint Qinami, Ioannis C. Karakozis, Kyle Genova, Prem Nair, Kenji Hata and Olga Russakovsky. **CVPR2020**. https://github.com/princetonvisualai/DomainBiasMitigation]

[Vikram V. Ramaswamy, Sunnie S. Y. Kim and Olga Russakovsky. https://github.com/princetonvisualai/gan-debiasing]

Recommended books:

Ruha Benjamin's "Race after technology"

Safiya Umoja Noble's "Algorithms of oppression"

Solon Barocas, Moritz Hardt and Arvind Narayanan's "Fairness and machine learning"