ML Fairness

April 11, 2019
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Wennie Zhang, Maulik Dang, Gurnaaz Kaur

Announcements

- Project updates — check ins, grid space, using external tools, what am I missing?
- Emails

Today

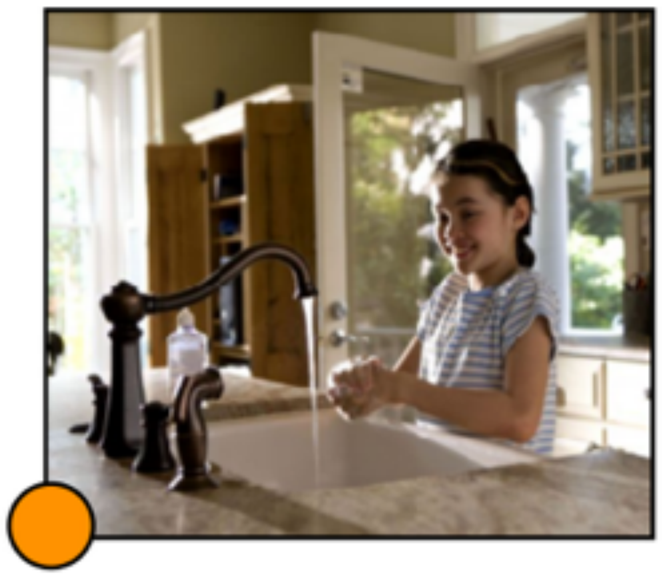
- Bias in ML
- Clickers—let's make a deal

Ethics, Fairness, and AI

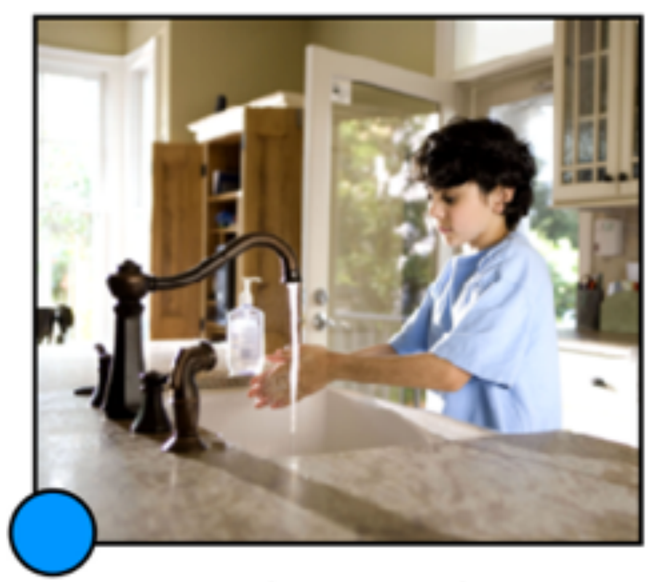
It is often assumed that big data techniques are unbiased because of the scale of the data and because the techniques are implemented through algorithmic systems. However, it is a mistake to assume they are objective simply because they are data-driven.

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.
Executive Office of the President, May 2016

Ethics, Fairness, and AI



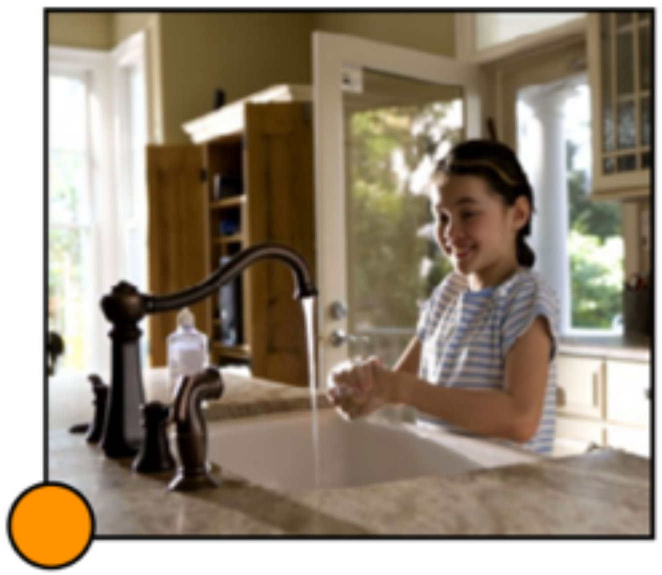
woman cooking



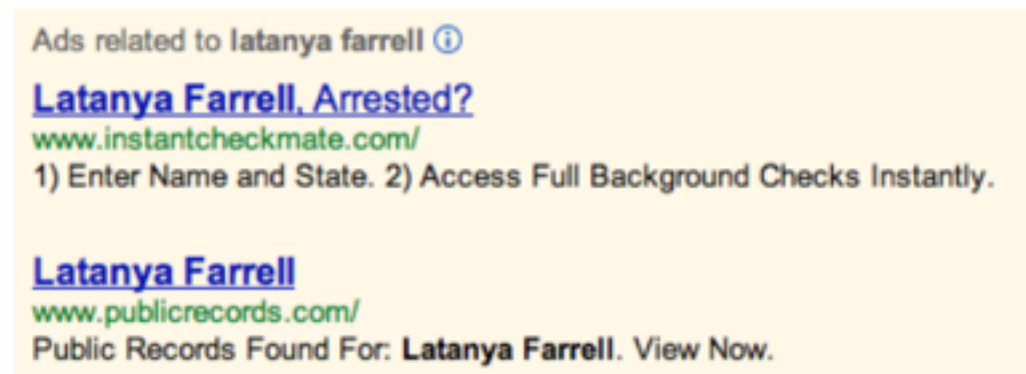
man fixing faucet

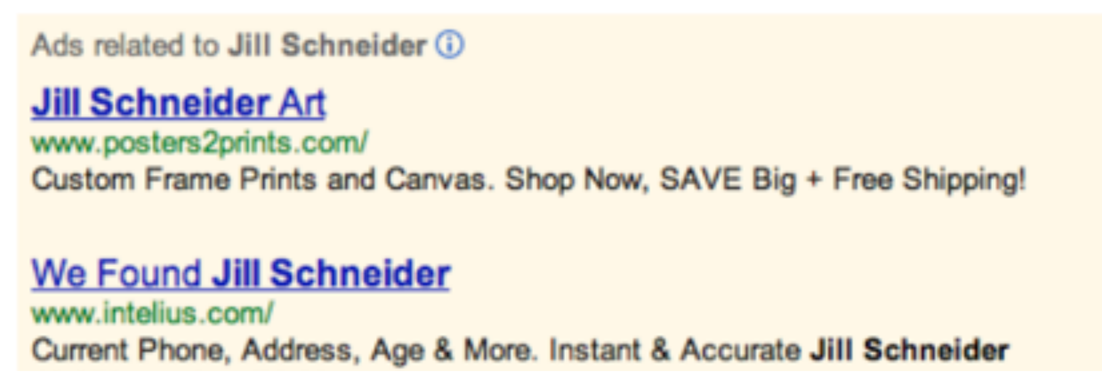
Men Also Like Shopping: Reducing Gender Bias Amplification… Zhao et al. (2017).

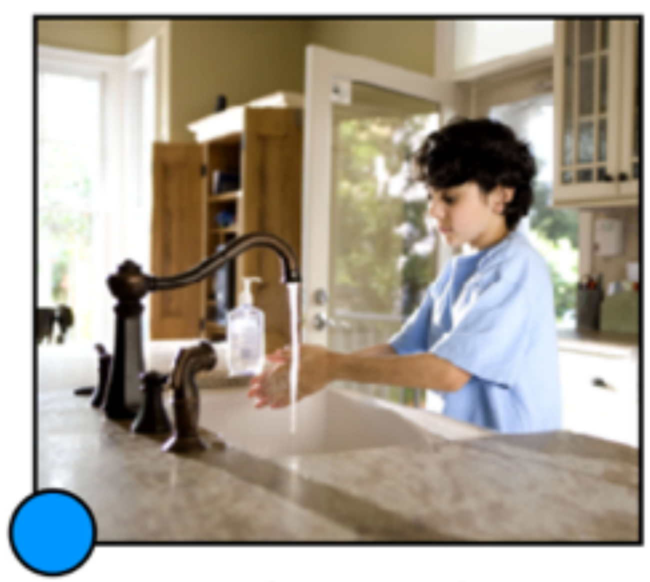
Ethics, Fairness, and AI



woman cooking



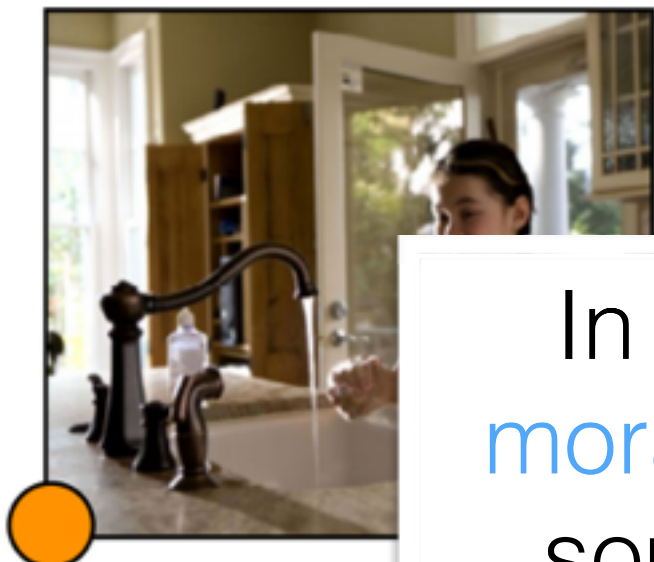




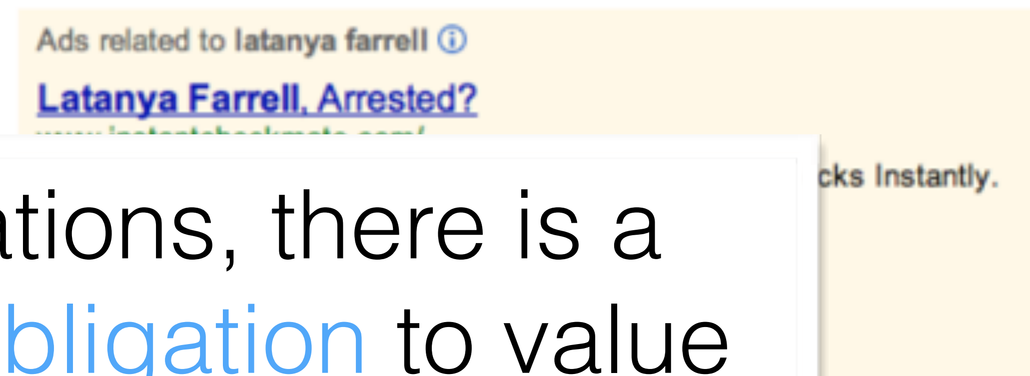
man fixing faucet

Men Also Like Shopping: Reducing Gender Bias Amplification… Zhao et al. (2017).
Discrimination in Online Ad Delivery. Sweeny (2013).

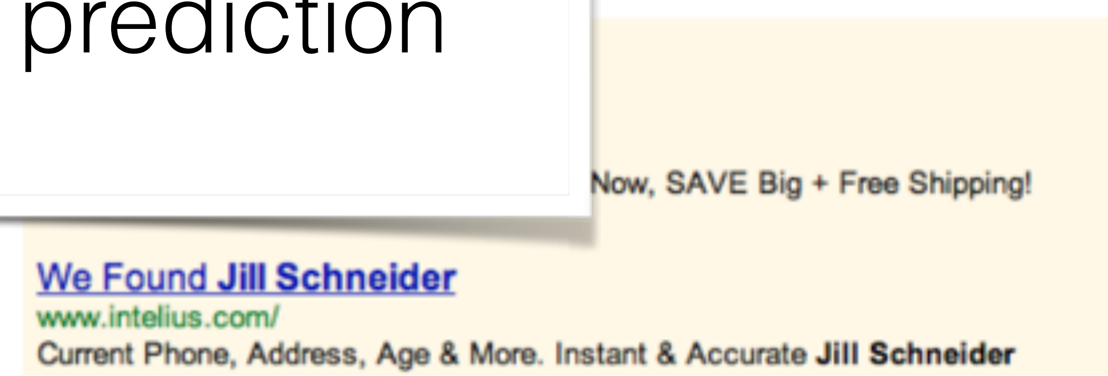
Ethics, Fairness, and AI



woman c



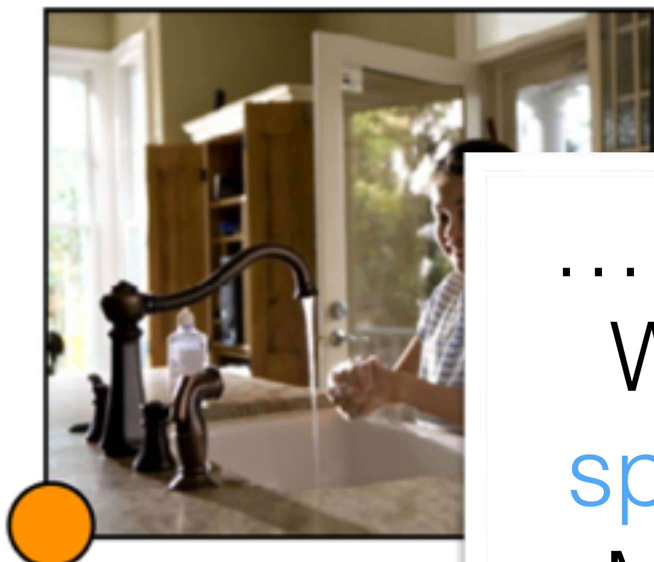
In some applications, there is a moral and legal obligation to value something other than prediction accuracy…



man fixing faucet

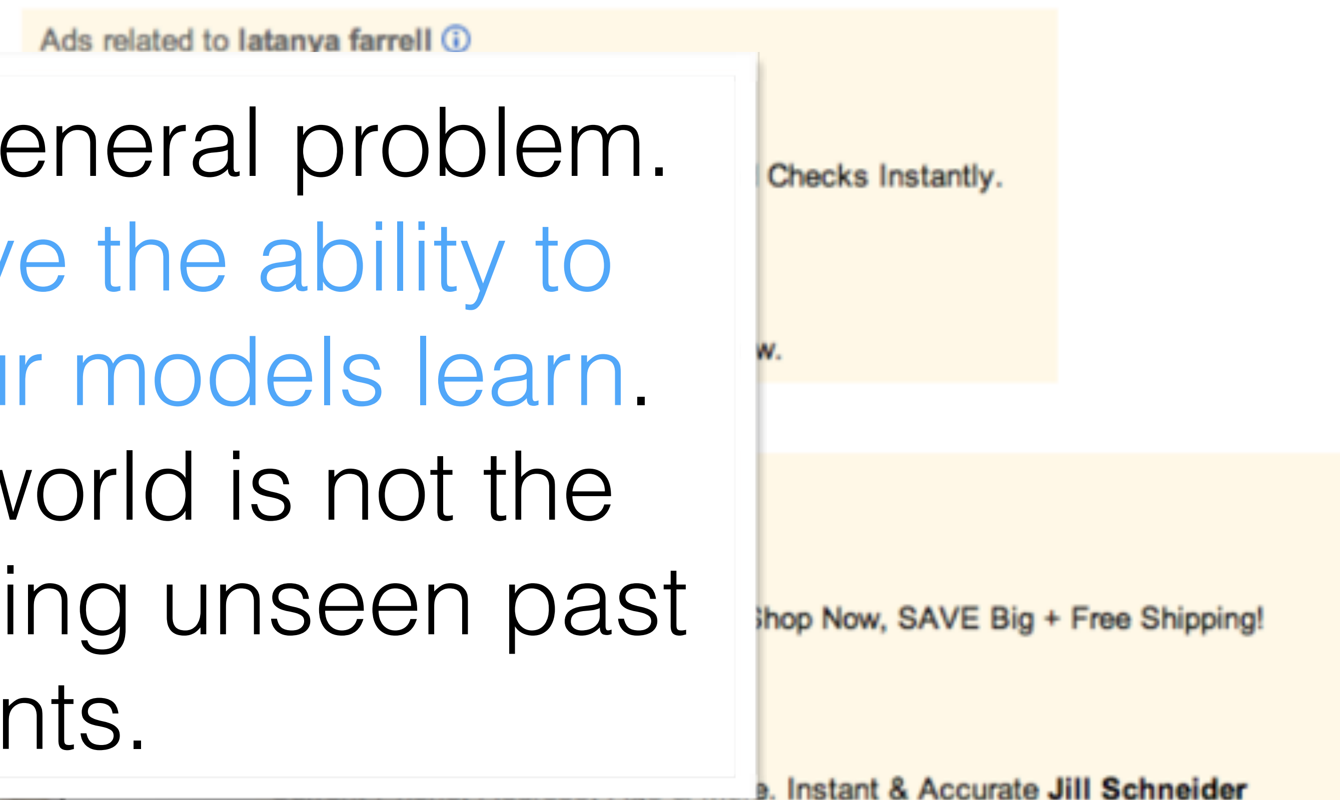
Men Also Like Shopping: Reducing Gender Bias Amplification… Zhao et al. (2017).
Discrimination in Online Ad Delivery. Sweeny (2013).

Ethics, Fairness, and AI



woman coo

…but this is a general problem. We should have the ability to specify what our models learn. Modeling the world is not the same as predicting unseen past events.



man fixing faucet

Men Also Like Shopping: Reducing Gender Bias Amplification… Zhao et al. (2017).
Discrimination in Online Ad Delivery. Sweeny (2013).

Core Components of ML

Input Data

Model

Objective

Core Components of ML

Input Data

Model

Objective

Core Components of ML

Input Data

Model

Objective

Core Components of ML

Input Data

Model

Objective

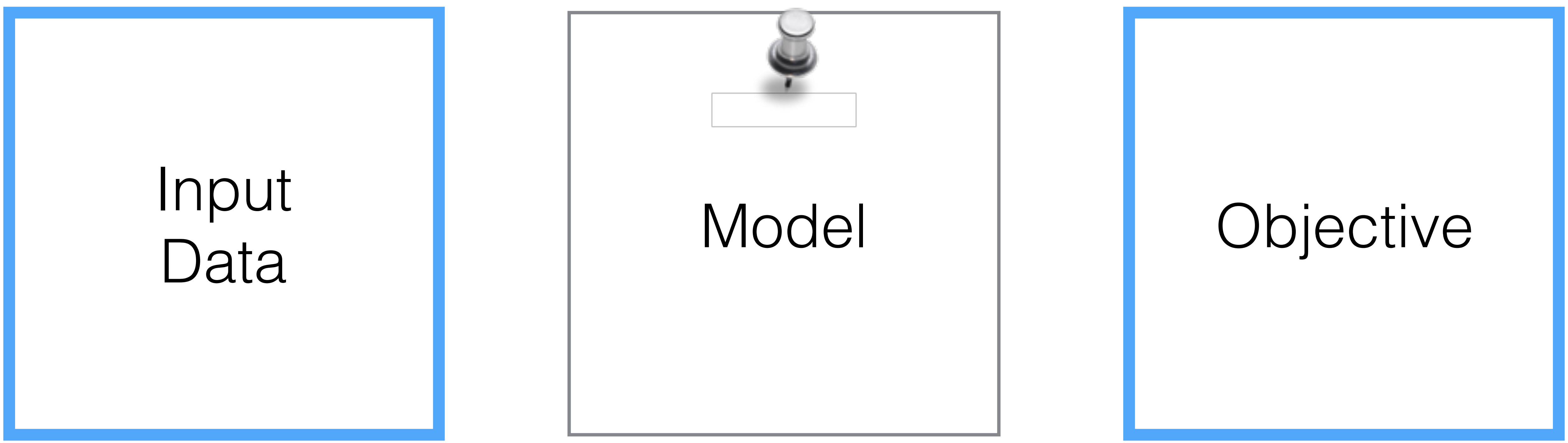
Core Components of ML

Input Data

Model

Objective

Core Components of ML



Core Components of ML

Input Data

Model

Objective

Biases in Input Data

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.
Executive Office of the President, May 2016

Biases in Input Data

- **Poorly selected data** where designers decide that certain data are important to the decision but not others. E.g. only include roads, not public transit or bike routes.

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.
Executive Office of the President, May 2016

Biases in Input Data

- **Poorly selected data** where designers decide that certain data are important to the decision but not others. E.g. only include roads, not public transit or bike routes.
- **Incomplete, incorrect, or outdated data**, where there may be a lack of technical rigor and comprehensiveness to data collection. E.g. bus or train routes not updated as quickly as road traffic.

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.
Executive Office of the President, May 2016

Biases in Input Data

- **Poorly selected data** where designers decide that certain data are important to the decision but not others. E.g. only include roads, not public transit or bike routes.
- **Incomplete, incorrect, or outdated data**, where there may be a lack of technical rigor and comprehensiveness to data collection. E.g. bus or train routes not updated as quickly as road traffic.
- **Selection bias**, where the set of data inputs to a model is not representative of a population. E.g. data collected from smartphone users.

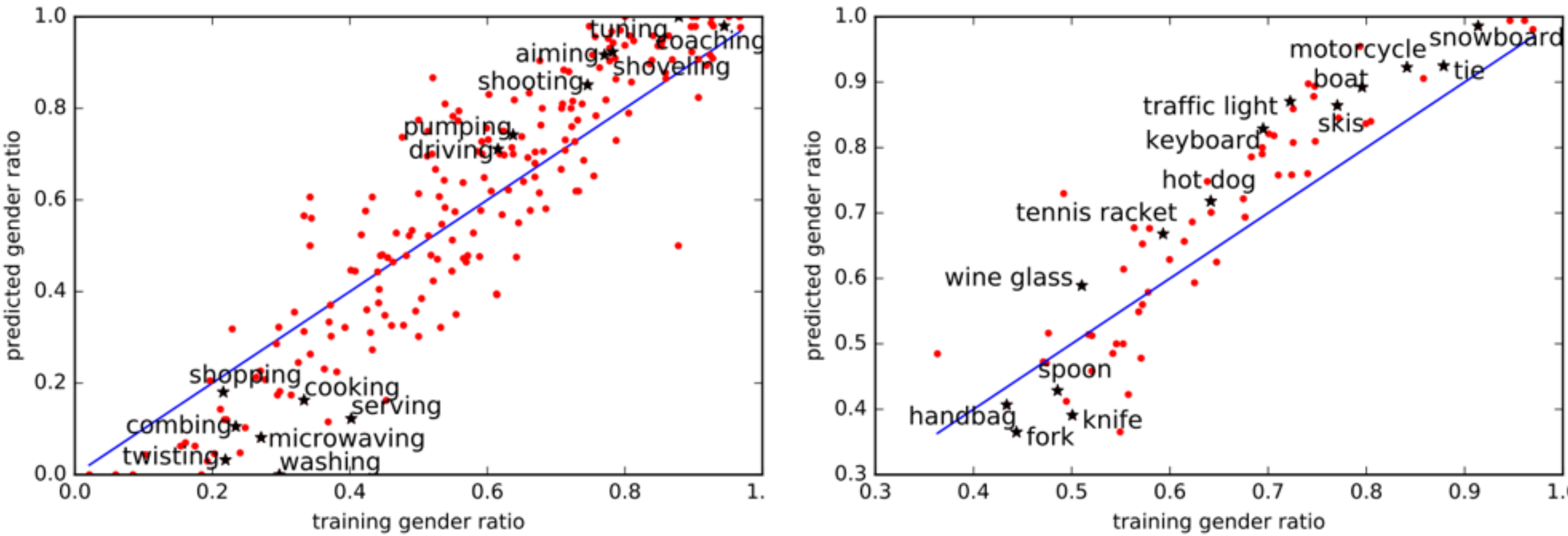
Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights.
Executive Office of the President, May 2016

Biases in Input Data

- **Poorly selected data** where designers decide that certain data are important to the decision but not others. E.g. only include roads, not public transit or bike routes.
- **Incomplete, incorrect, or outdated data**, where there may be a lack of technical rigor and comprehensiveness to data collection. E.g. bus or train routes not updated as quickly as road traffic.
- **Selection bias**, where the set of data inputs to a model is not representative of a population. E.g. data collected from smartphone users.
- **Unintentional perpetuation and promotion of historical biases**, where a feedback loop causes bias in results of the past to replicate in the future. E.g. hiring for "culture fit"

Big D ... nts.
Executive Office of the President, May 2016

Bias Amplification

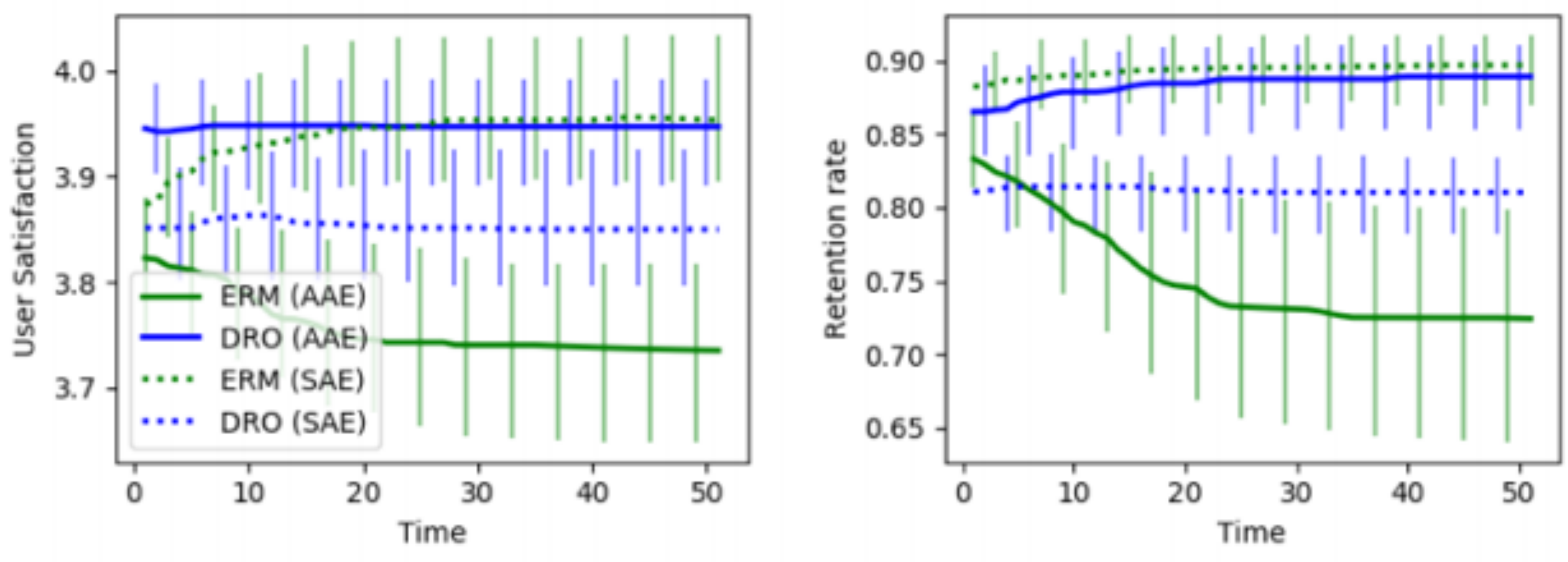


(a) Bias analysis on imSitu vSRL

(b) Bias analysis on MS-COCO MLC

Men Also Like Shopping: Reducing Gender Bias Amplification… Zhao et al. (2017).

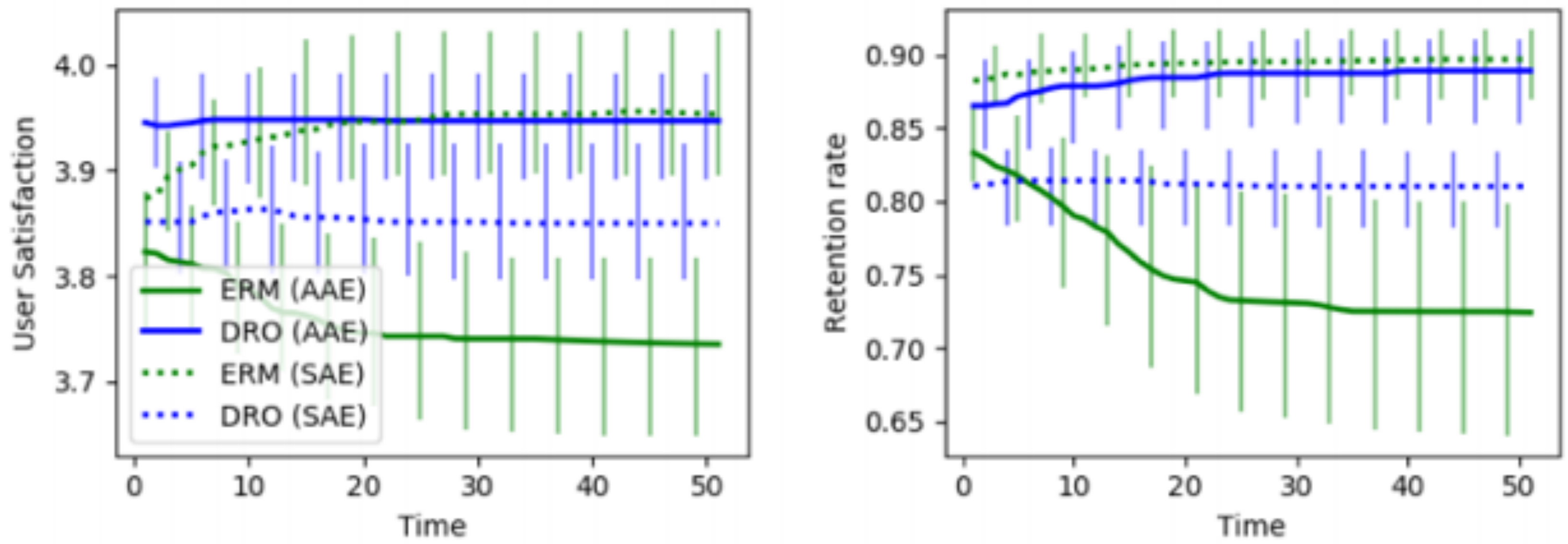
Bias Amplification



(a) User satisfaction (b) User retention

Fairness Without Demographics in Repeated Loss Minimization. Hashimoto et al (2018).

Bias Amplification



(a) User satisfaction

(b) User retention

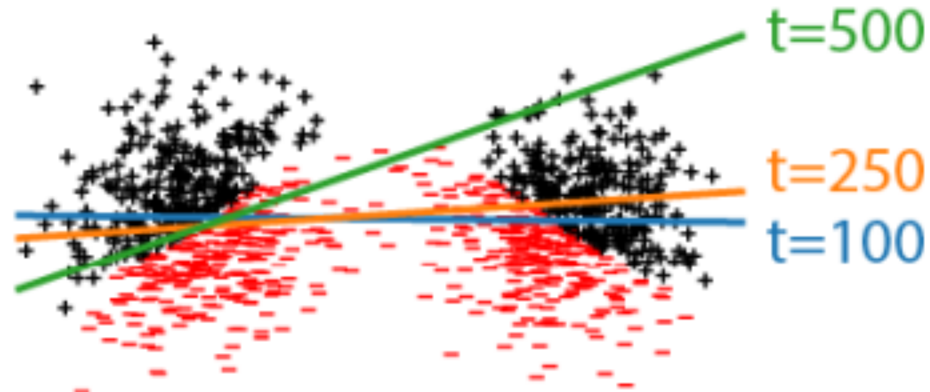


Figure 1. An example online classification problem which begins fair, but becomes unfair over time.

Fairness Without Demographics in Repeated Loss Minimization. Hashimoto et al (2018).

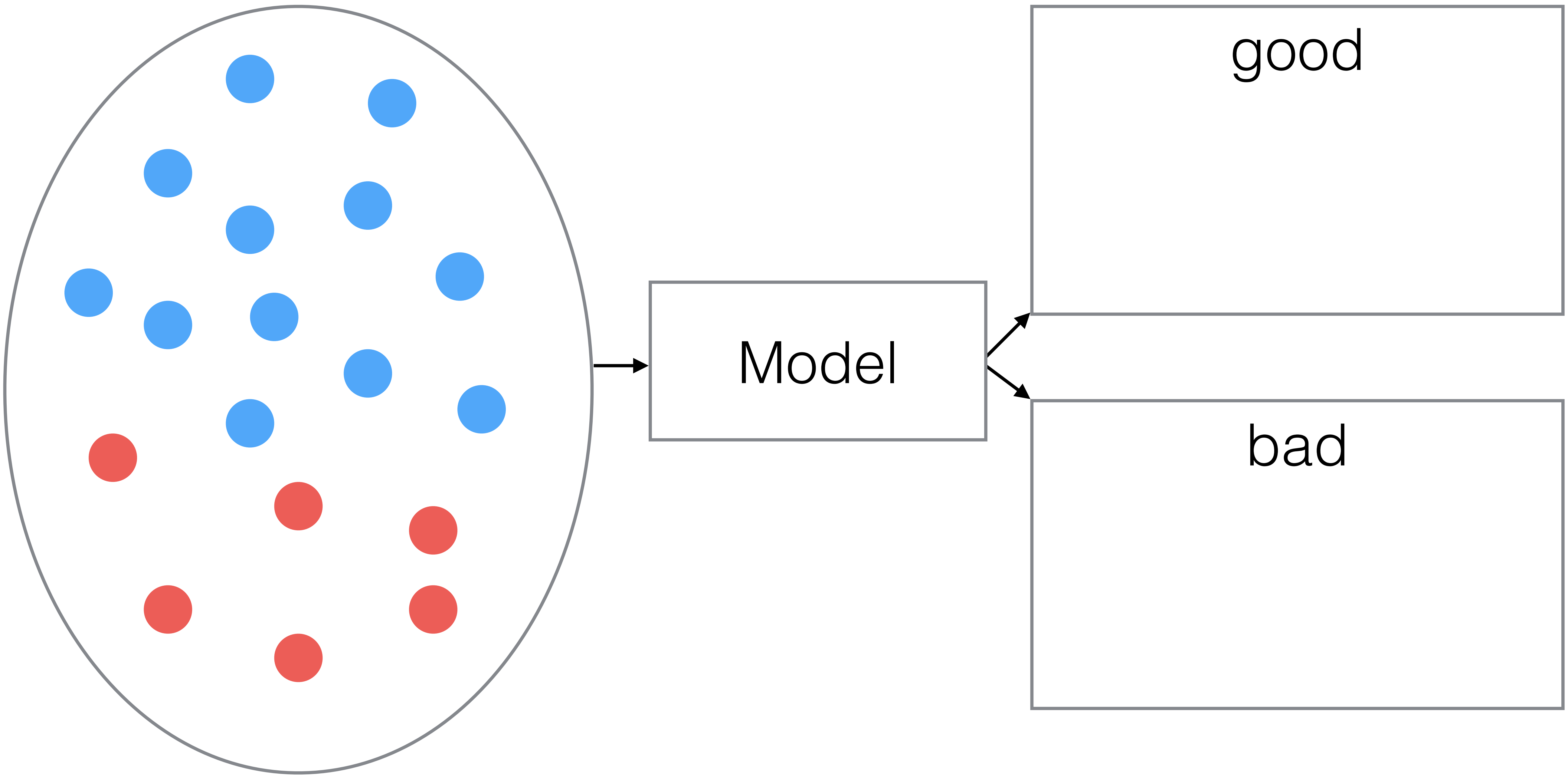
Core Components of ML

Input
Data

Model

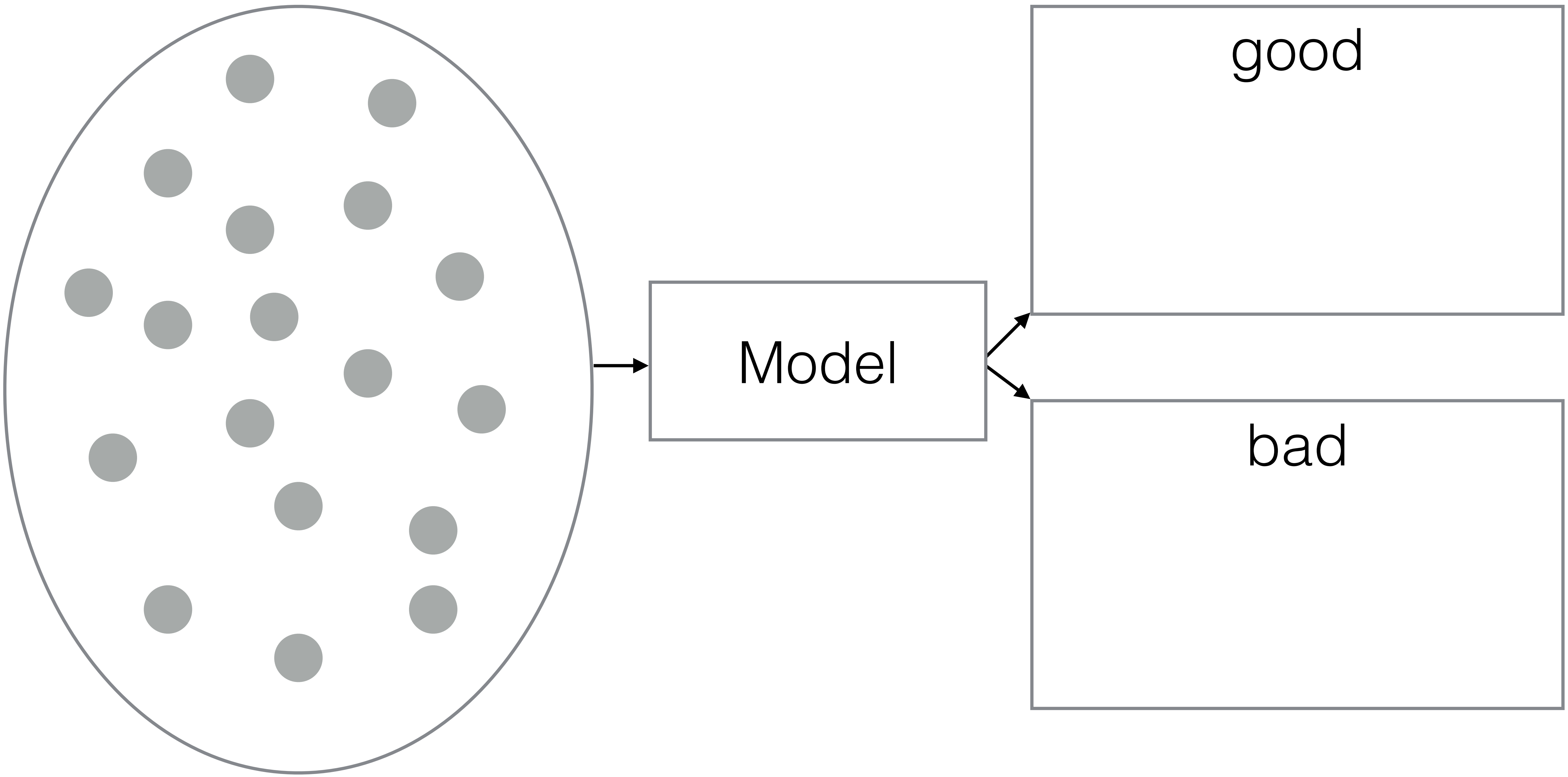
Objective

What is “fair”?



Fairness through unawareness

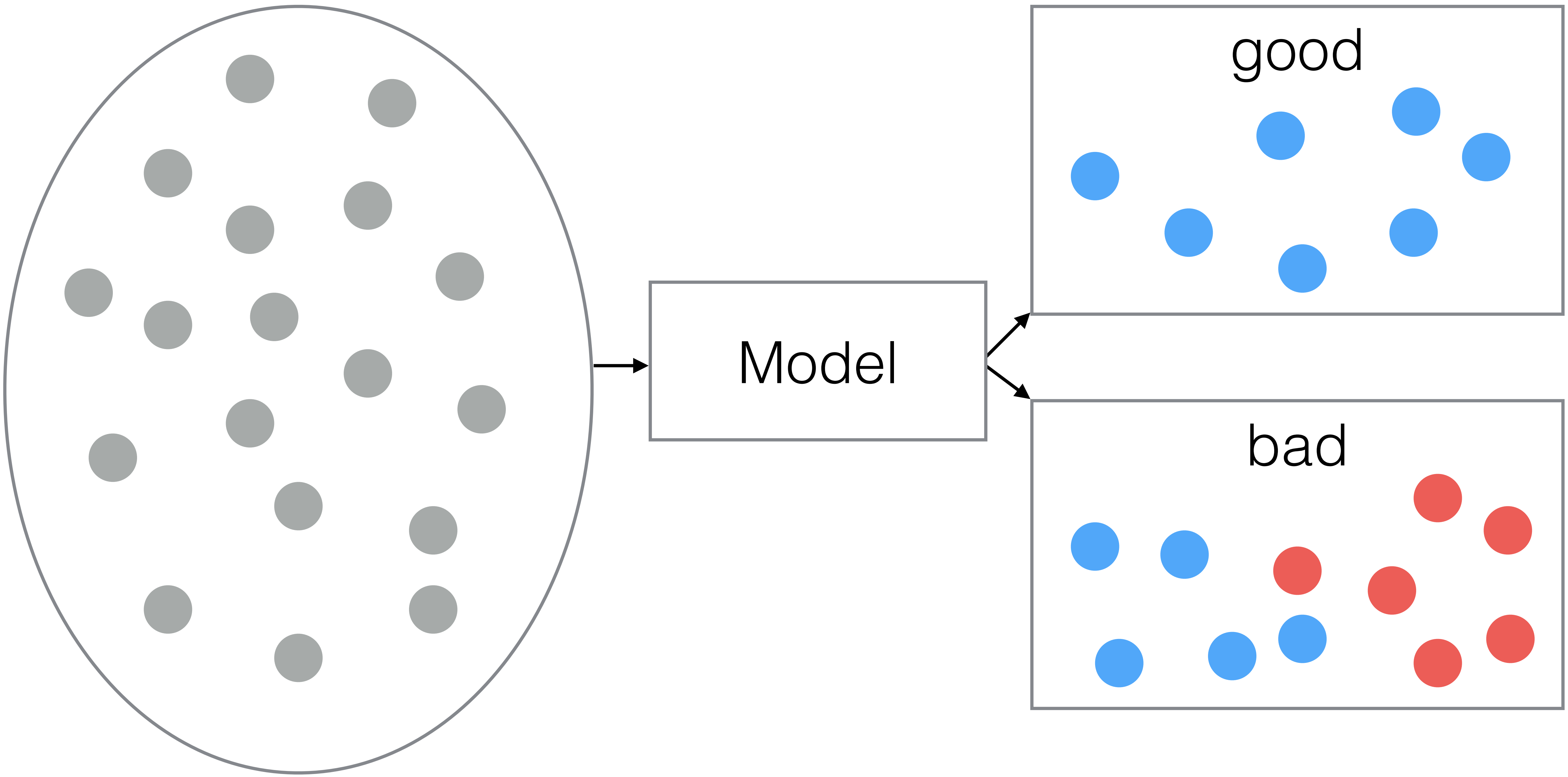
"I don't see color" approach



Counterfactual Fairness. Kusner et al (2017).

Fairness through unawareness

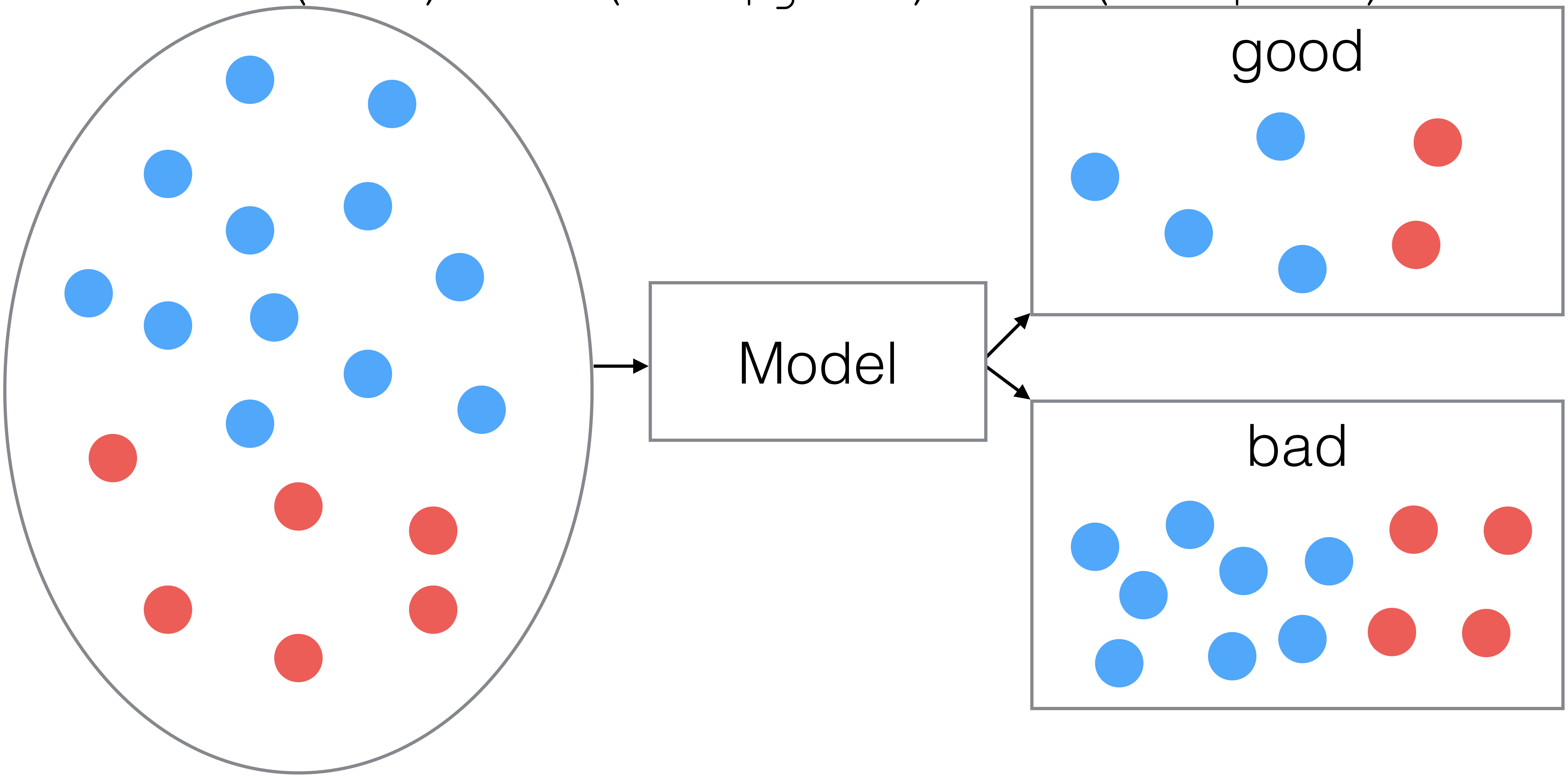
"I don't see color" approach



Counterfactual Fairness. Kusner et al (2017).

Demographic Parity

```
P(blue) = P(blue|good) = P(blue|bad)
P(red) = P(red|good) = P(red|bad)
```

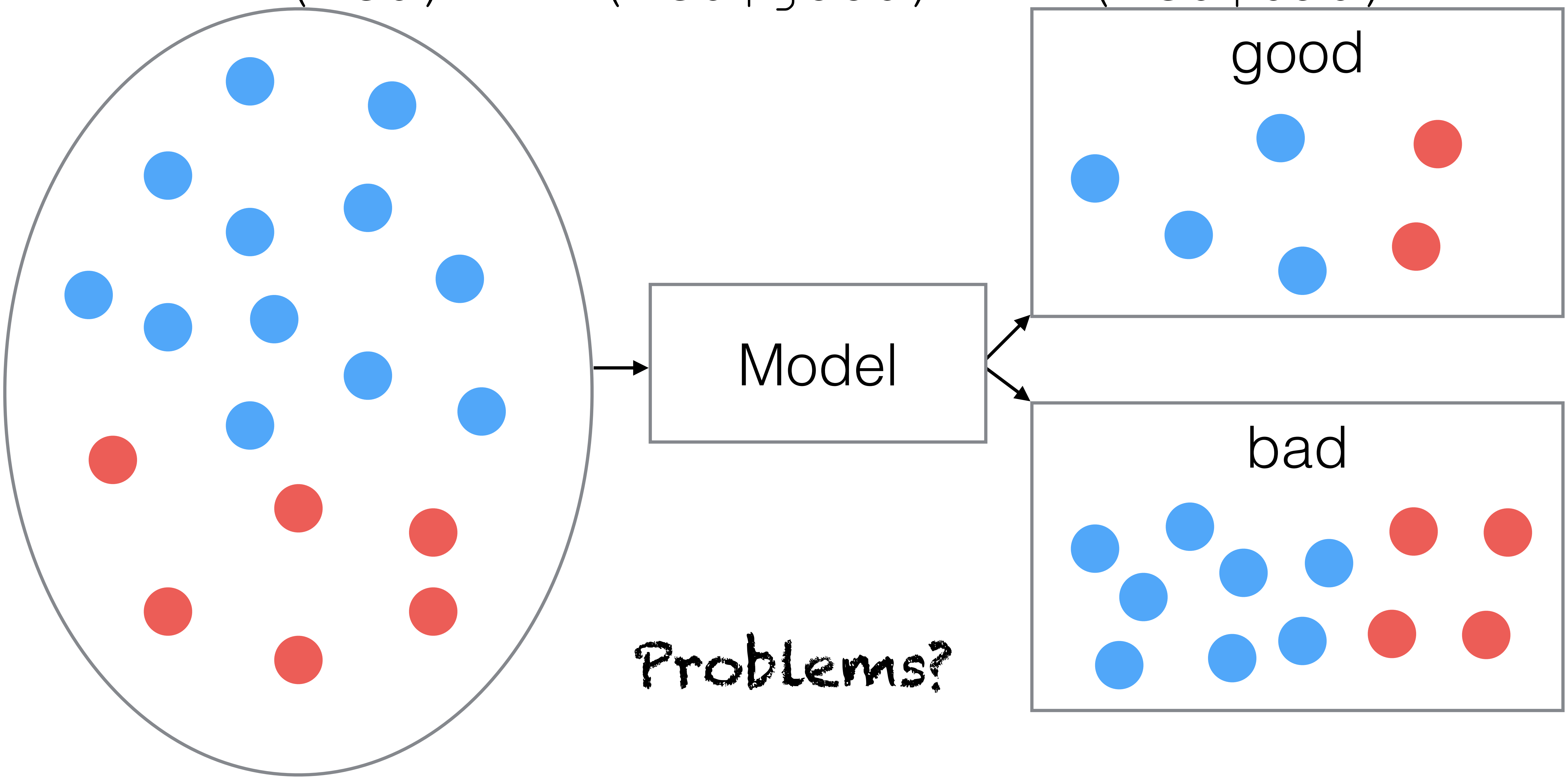


Fairness Through Awareness. Dwork et al. (2012).

Demographic Parity

P(blue) = P(blue|good) = P(blue|bad)

P(red) = P(red|good) = P(red|bad)



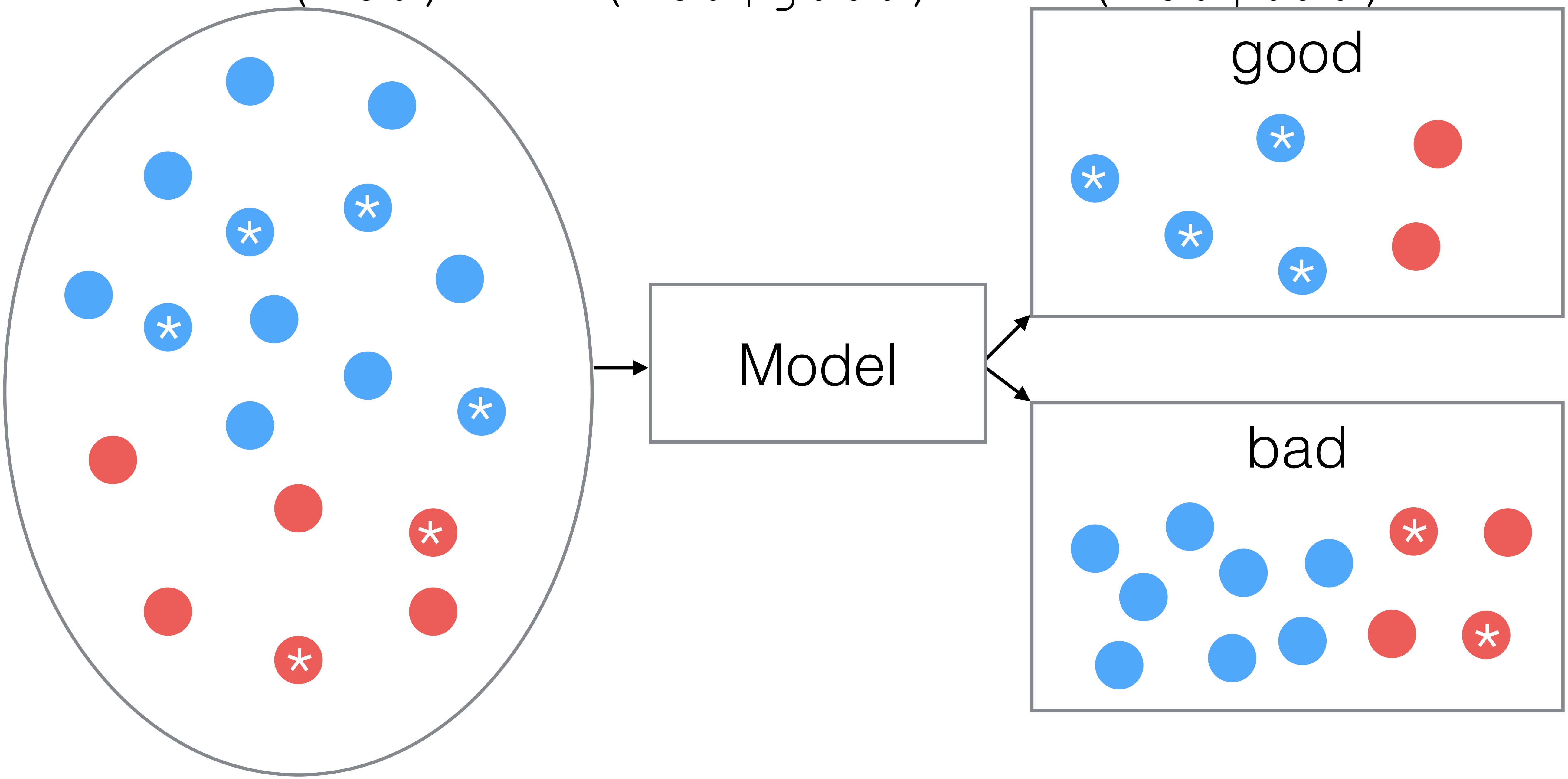
Problems?

Fairness Through Awareness. Dwork et al. (2012).

Demographic Parity

P(blue) = P(blue|good) = P(blue|bad)

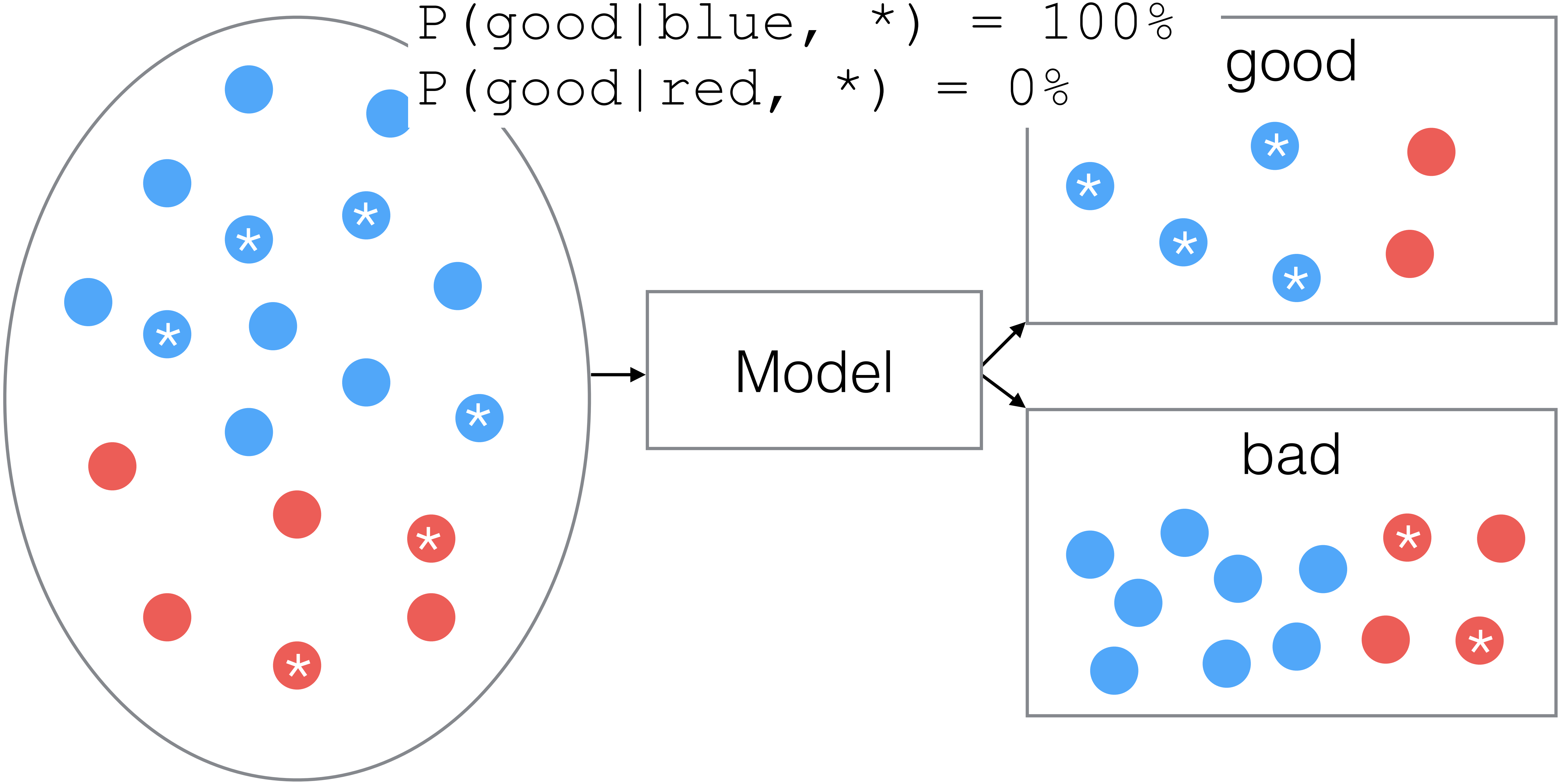
P(red) = P(red|good) = P(red|bad)



Fairness Through Awareness. Dwork et al. (2012).

Demographic Parity

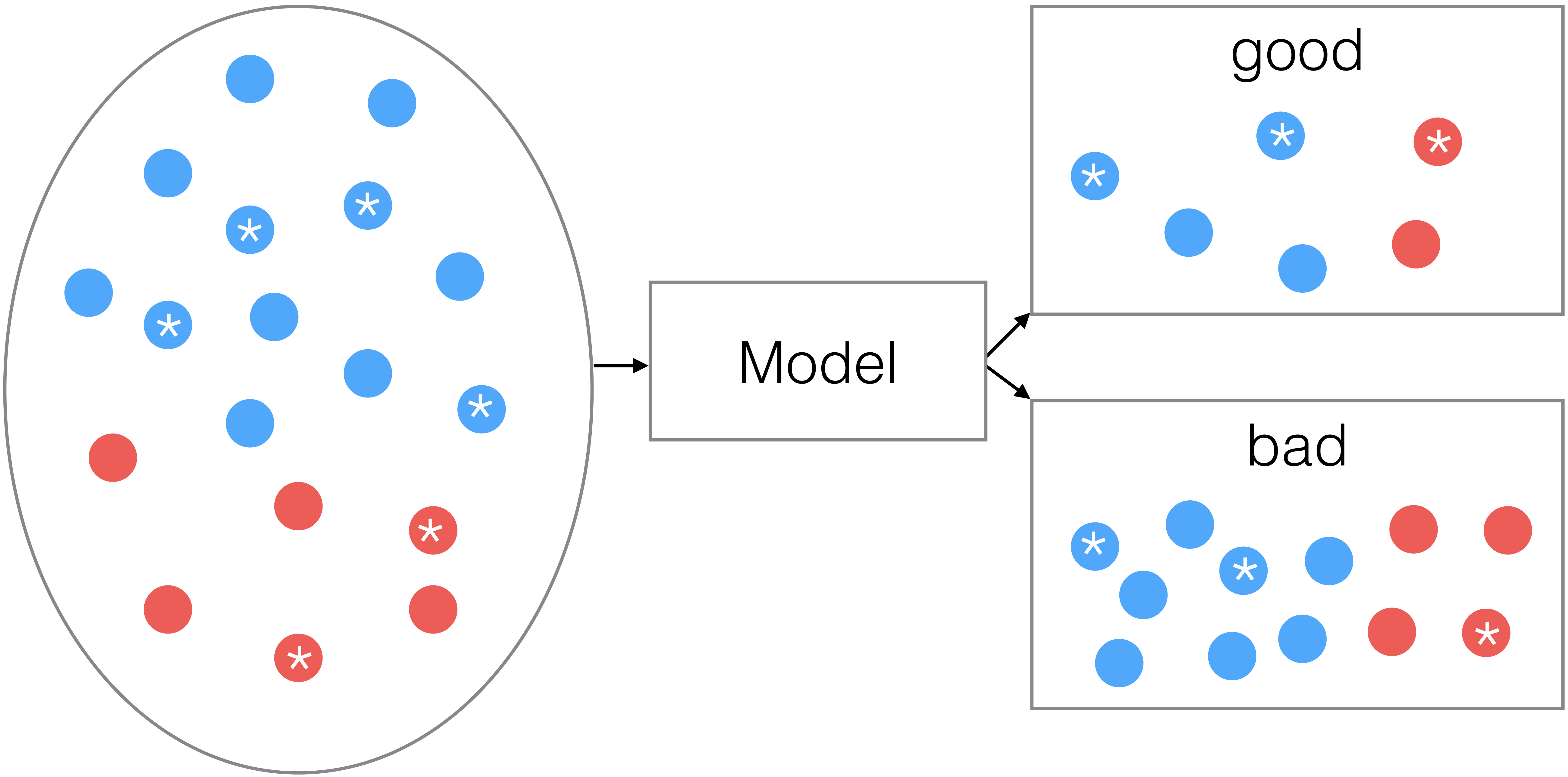
```
P(good| *) = 67%
P(good|blue, *) = 100%
P(good|red, *) = 0%
```



Fairness Through Awareness. Dwork et al. (2012).

Equalized Odds

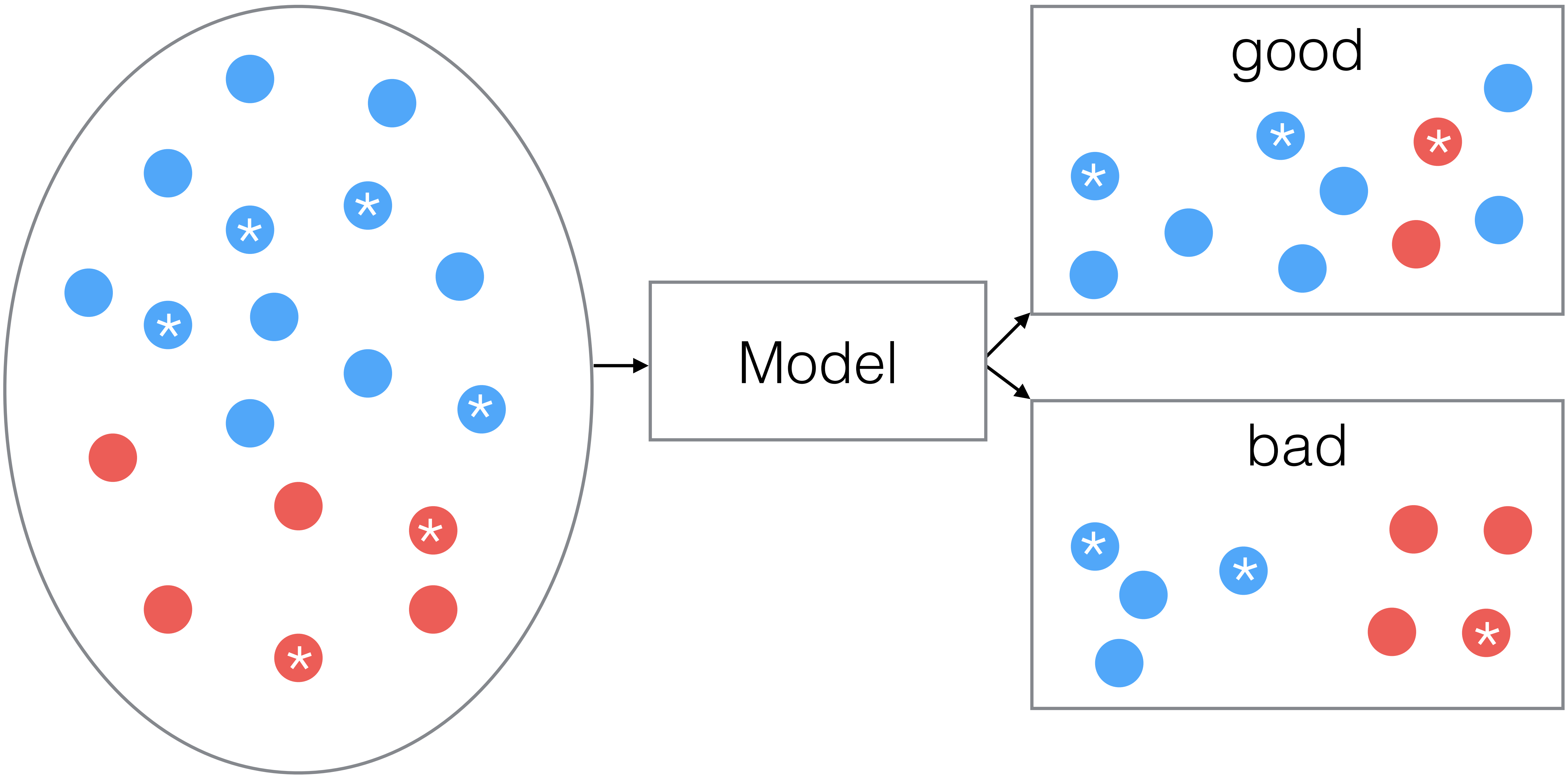
```
P(good| *) = P(good|blue, *) = P(good|red, *)
P(bad| *) = P(bad|blue, ~*) = P(bad|red, ~*)
```



Equality of Opportunity in Supervised Learning. Hardt et al. (2016).

Equal Opportunity

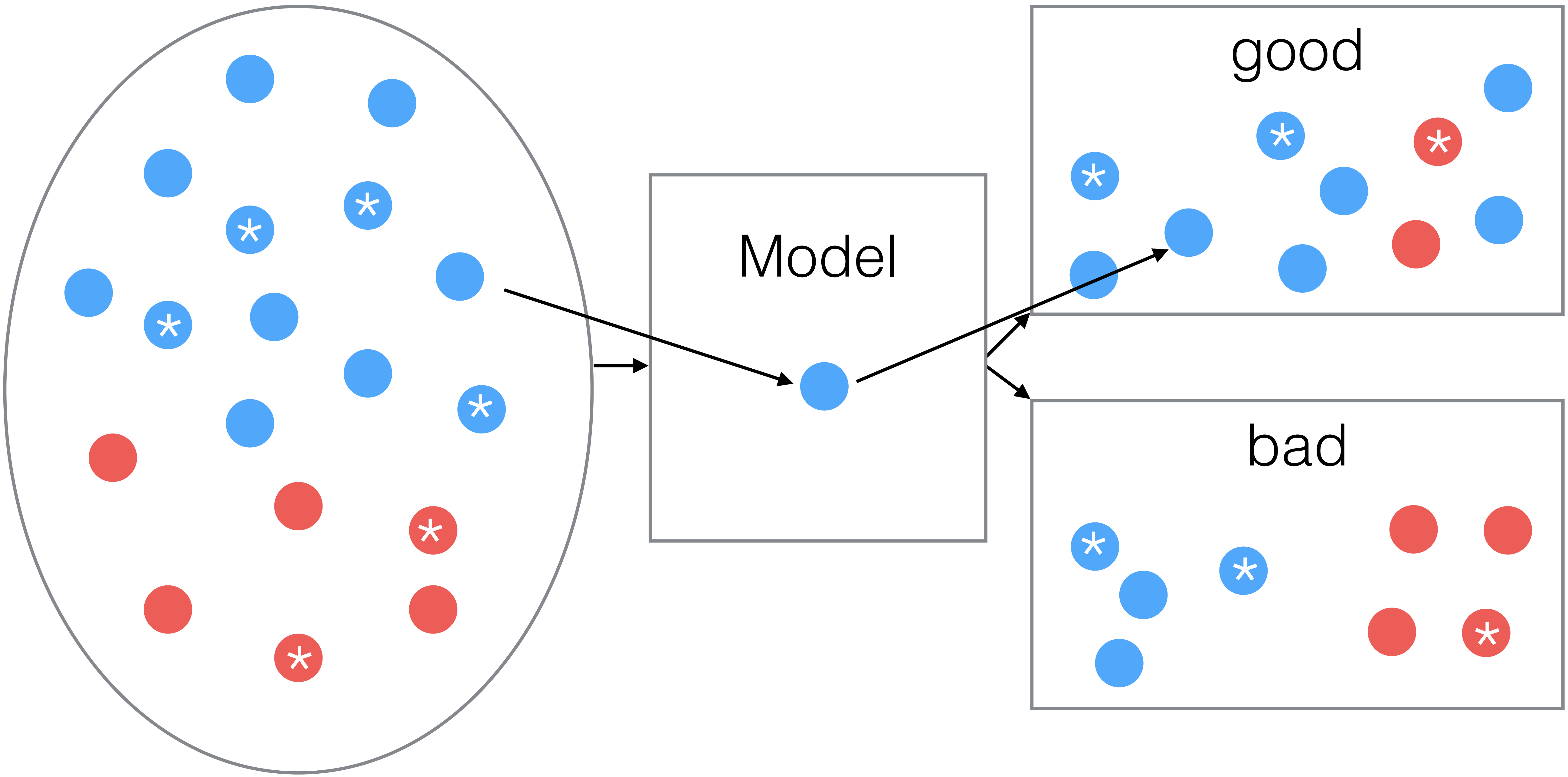
`P(good| *) = P(good|blue, *) = P(good|red, *)`



Equality of Opportunity in Supervised Learning. Hardt et al. (2016).

Counterfactual Fairness

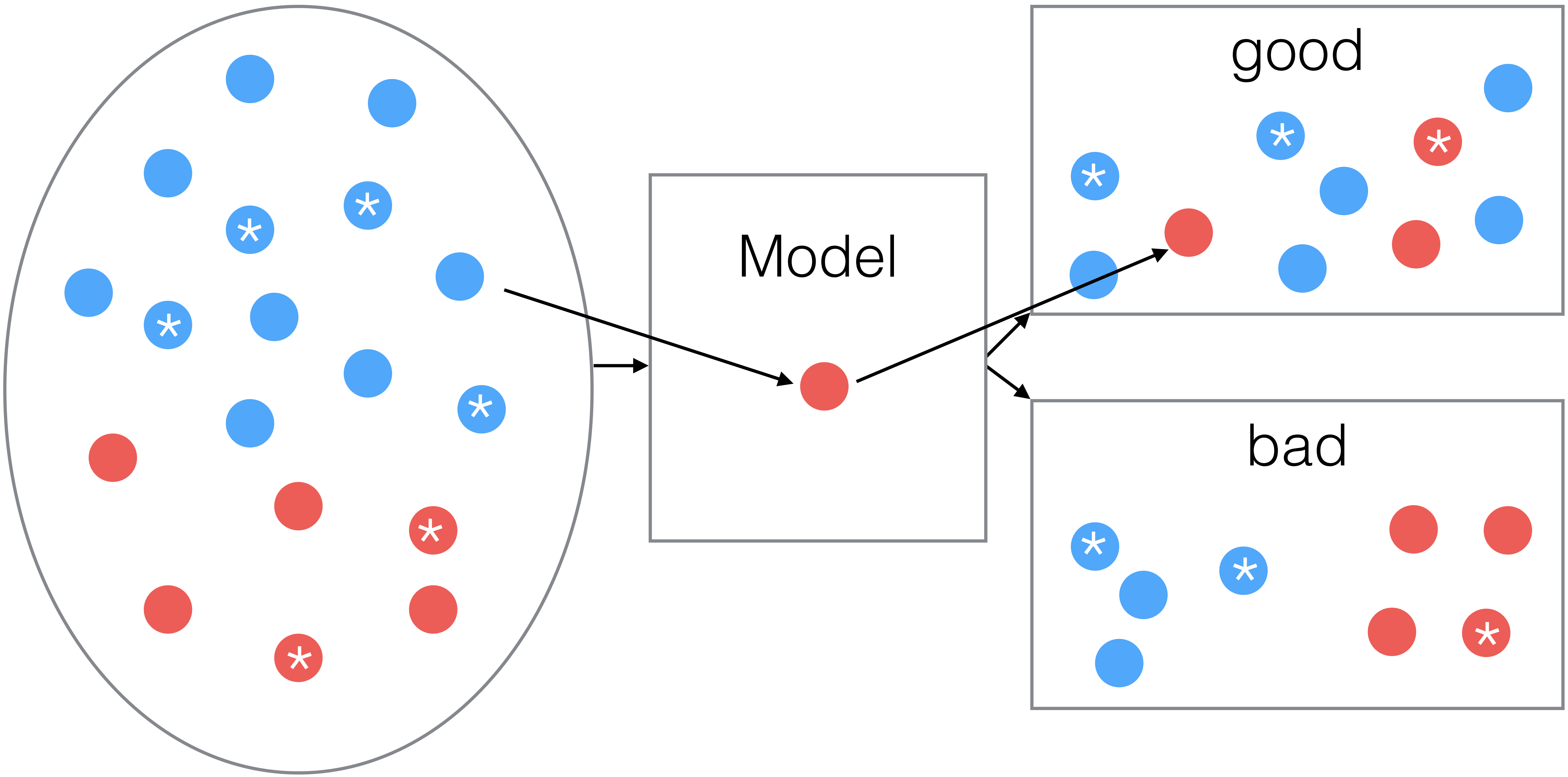
"Would you say that if I were white?" approach



Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

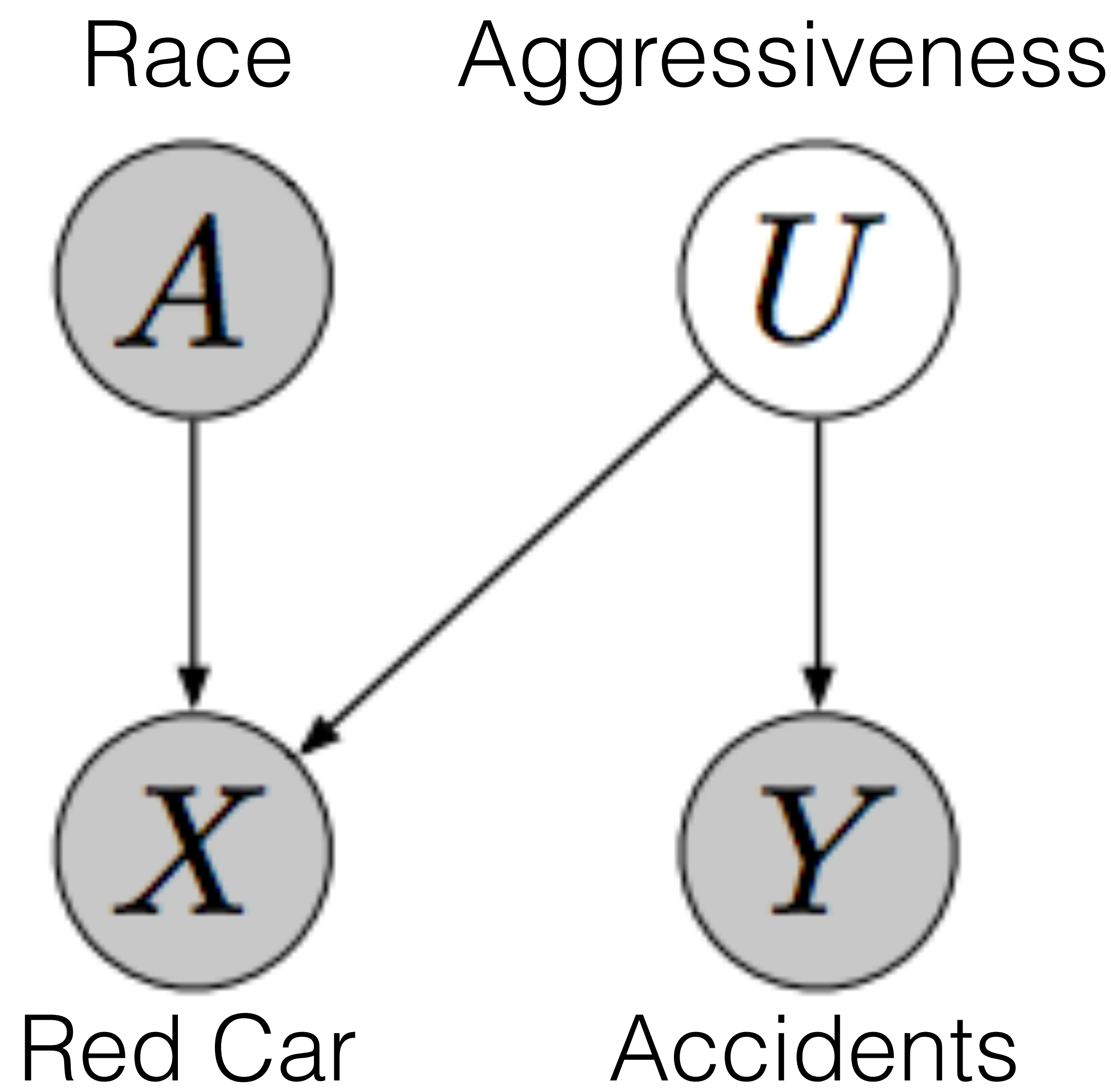
"Would you say that if I were white?" approach



Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

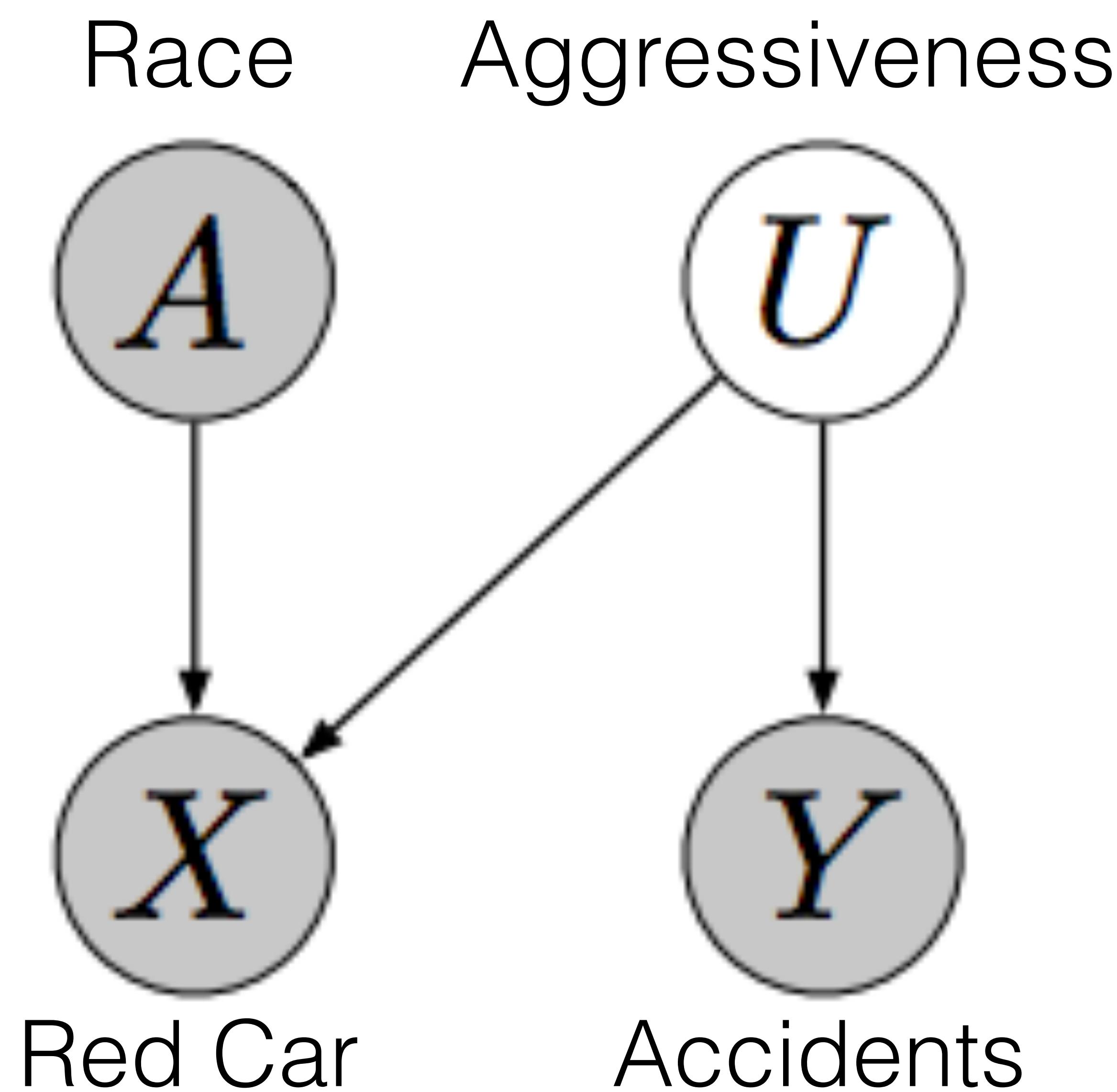
A car insurance company wishes to price insurance for car owners by predicting their accident rate Y. They assume there is an unobserved factor corresponding to **aggressive driving U**, that both causes drivers to be more likely have an accident, and **also causes individuals to prefer red cars (the observed variable X)**. Moreover, individuals belonging to a certain **race A are more likely to drive red cars**. However, these individuals are no more likely to be aggressive or to get in accidents than any one else. Thus, using the red car feature X to predict accident rate Y would seem to be an unfair prediction because it may charge individuals of a certain race more than others, even though no race is more likely to have an accident. Counterfactual fairness agrees with this notion: **changing A while holding U fixed will also change X and, consequently, $\hat{Y}$** .



Counterfactual Fairness. Kusner et al (2017).

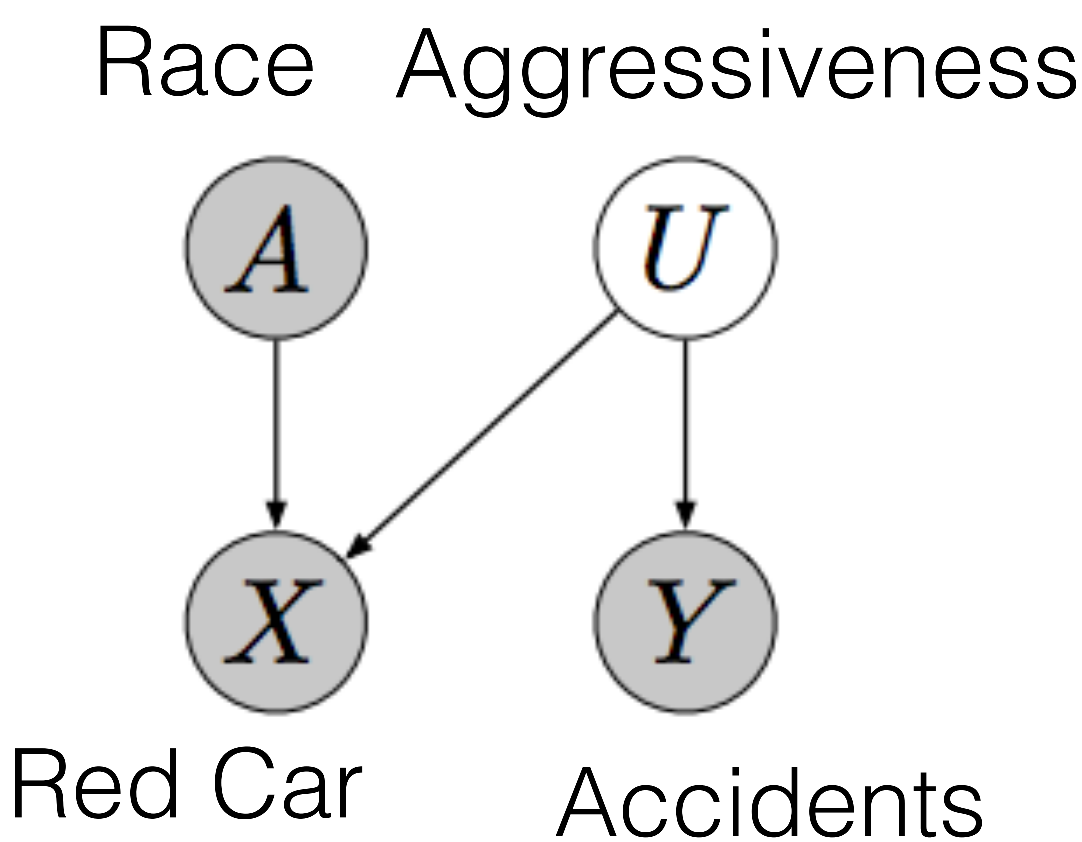
Counterfactual Fairness

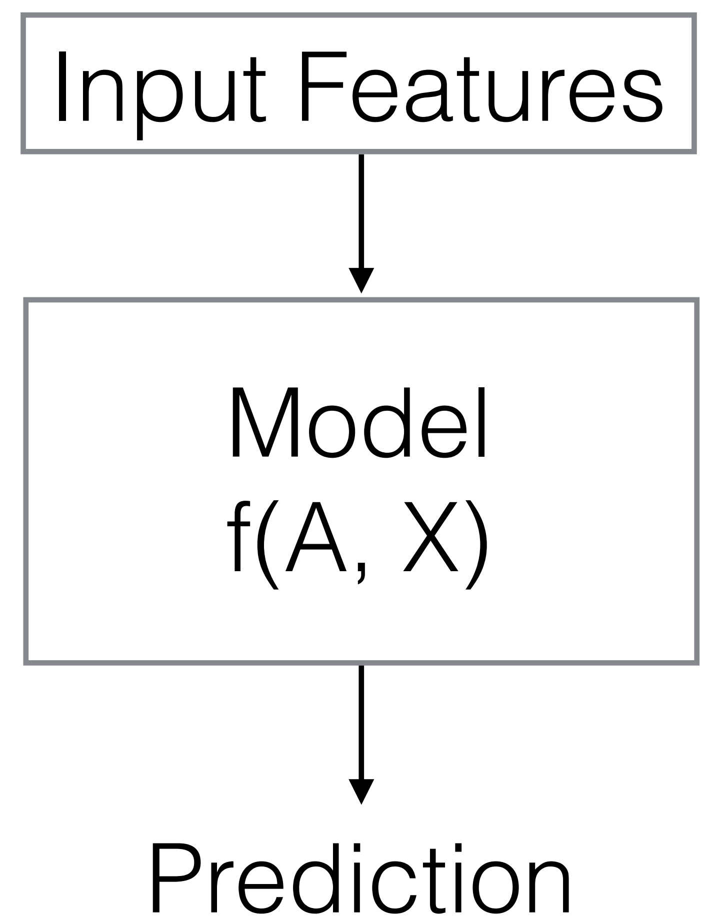
We can show that in a linear model, regressing Y on A and X is equivalent to regressing on U, **so off-the-shelf regression here is counterfactually fair.** Regressing Y on X alone obeys the FTU criterion but is not counterfactually fair, so **omitting A (FTU) may introduce unfairness into an otherwise fair world**.



Counterfactual Fairness. Kusner et al (2017).

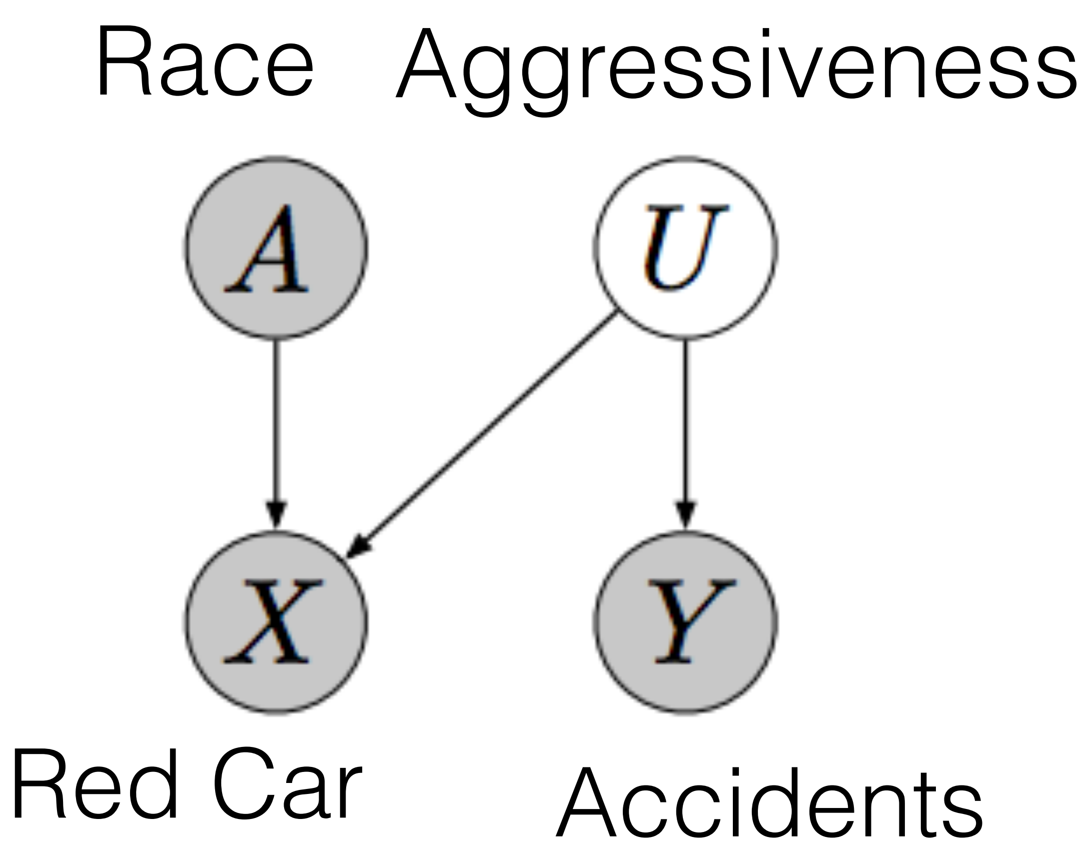
Counterfactual Fairness

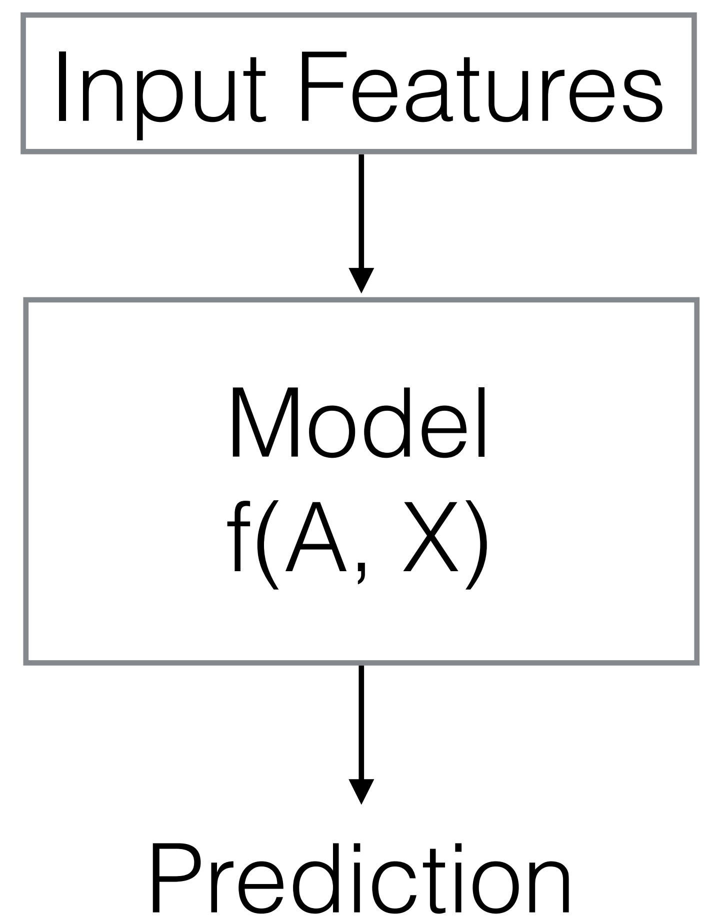




Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

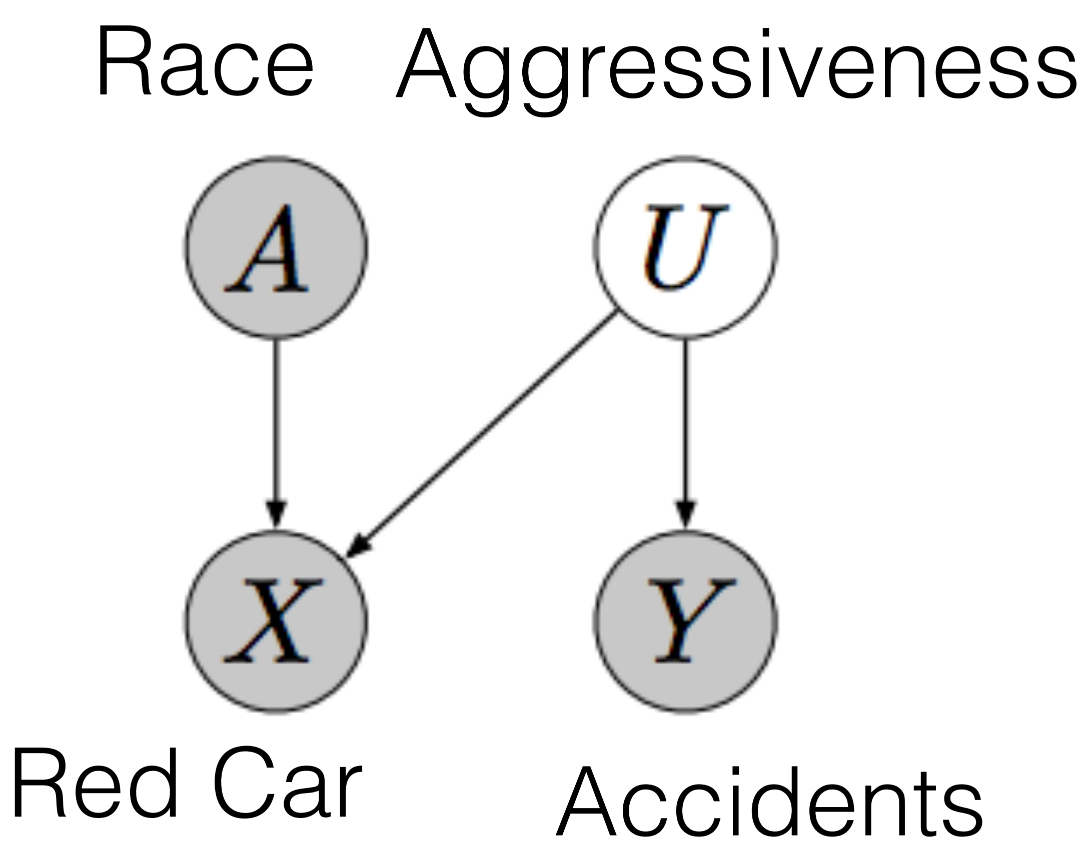




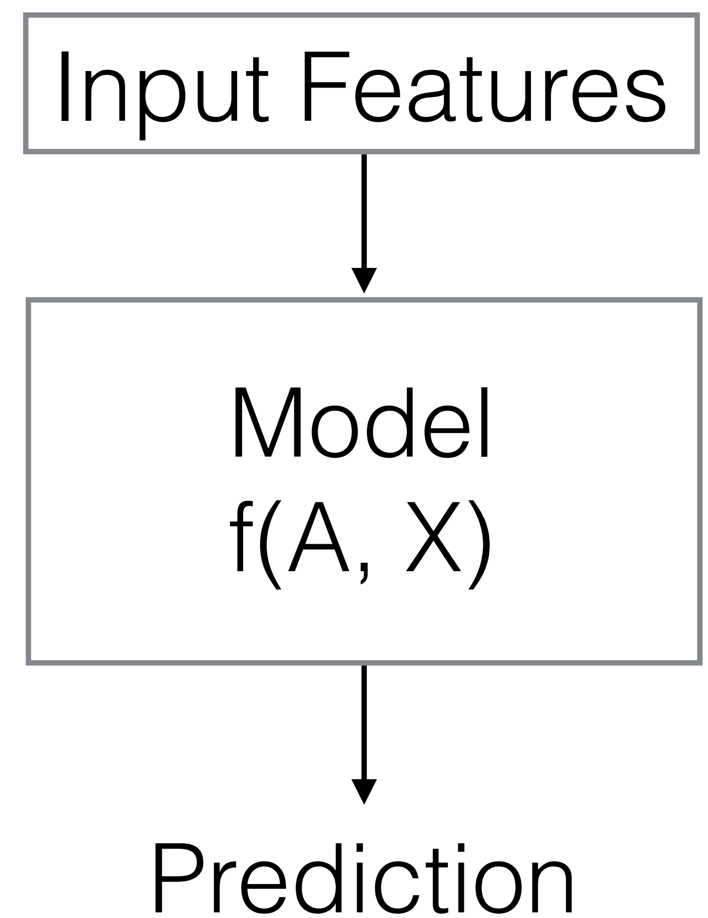
Train this one

Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness



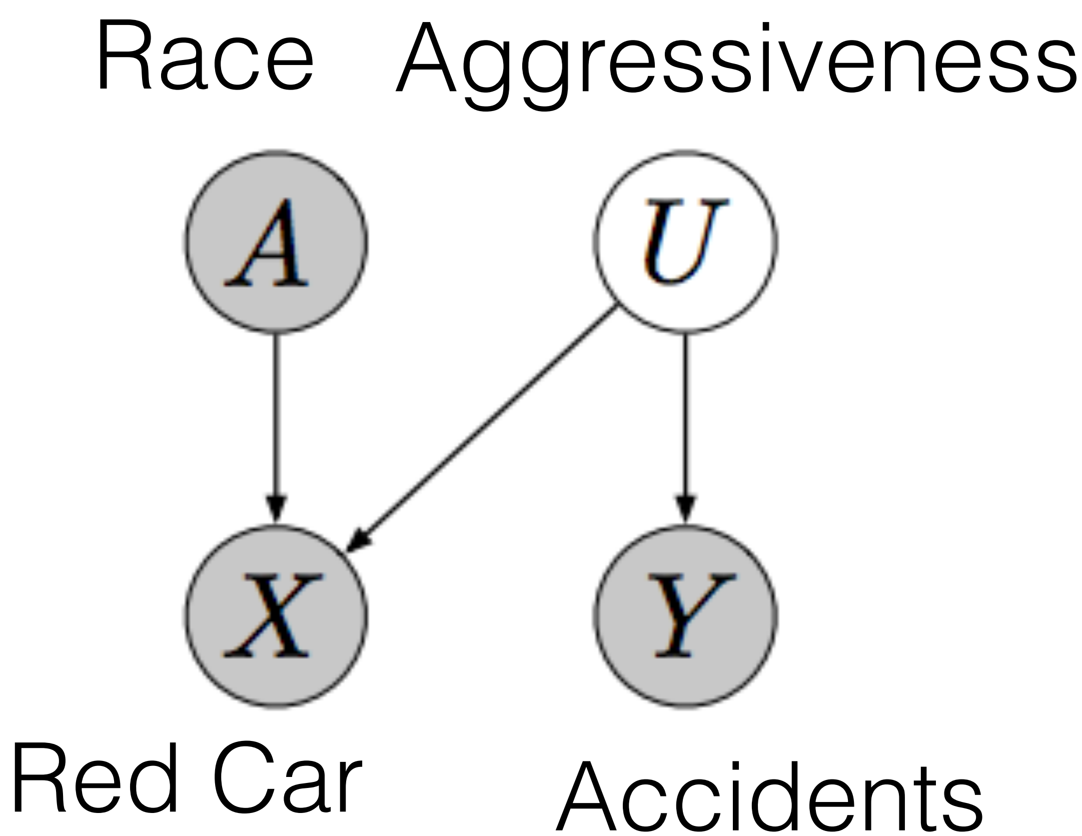
Postulate this one



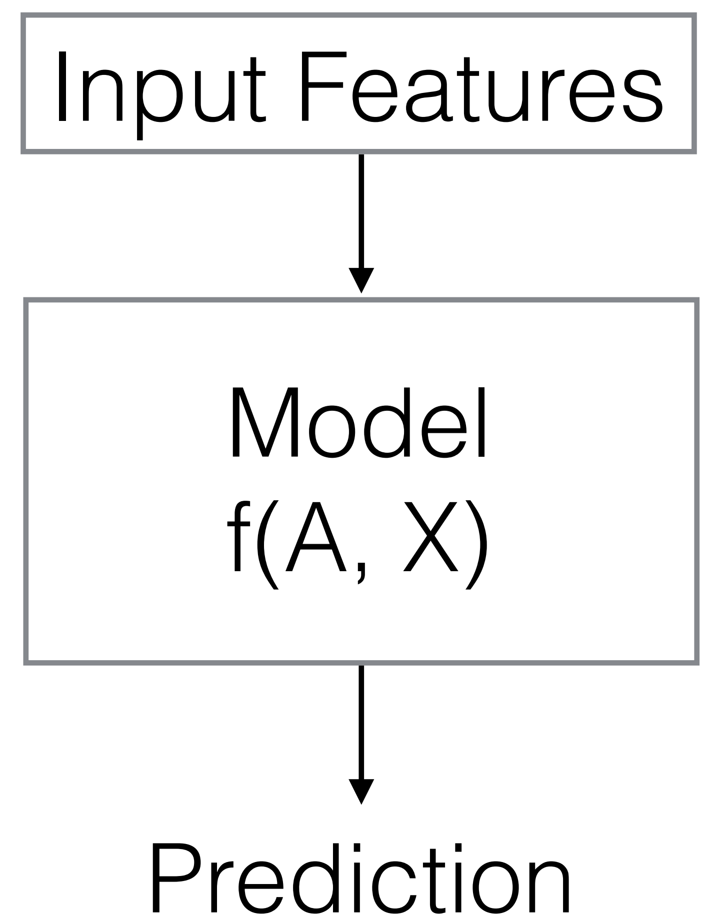
Train this one

Counterfactual Fairness. Kusner et al (2017).

Use to generate
training samples
with contain U



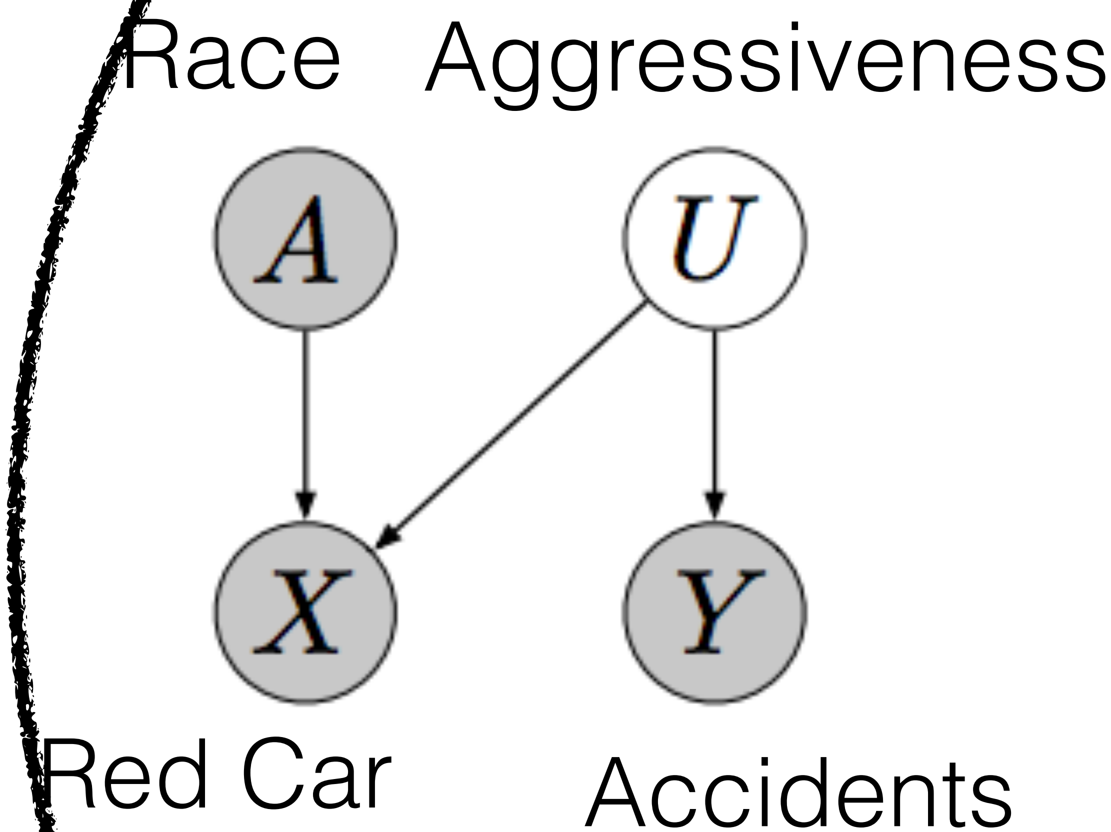
Postulate this one



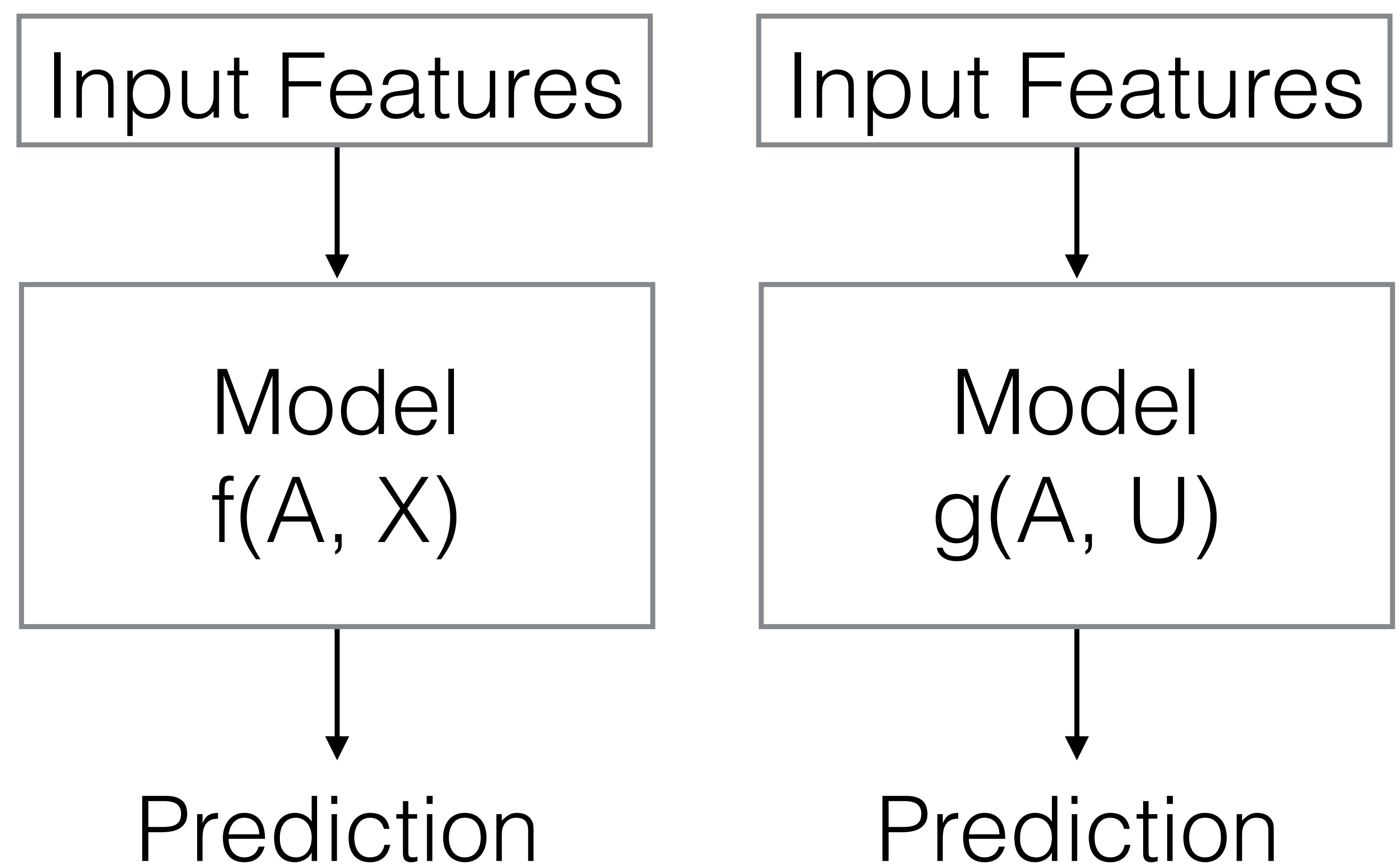
Train this one

Counterfactual Fairness. Kusner et al (2017).

Use to generate training samples with contain U

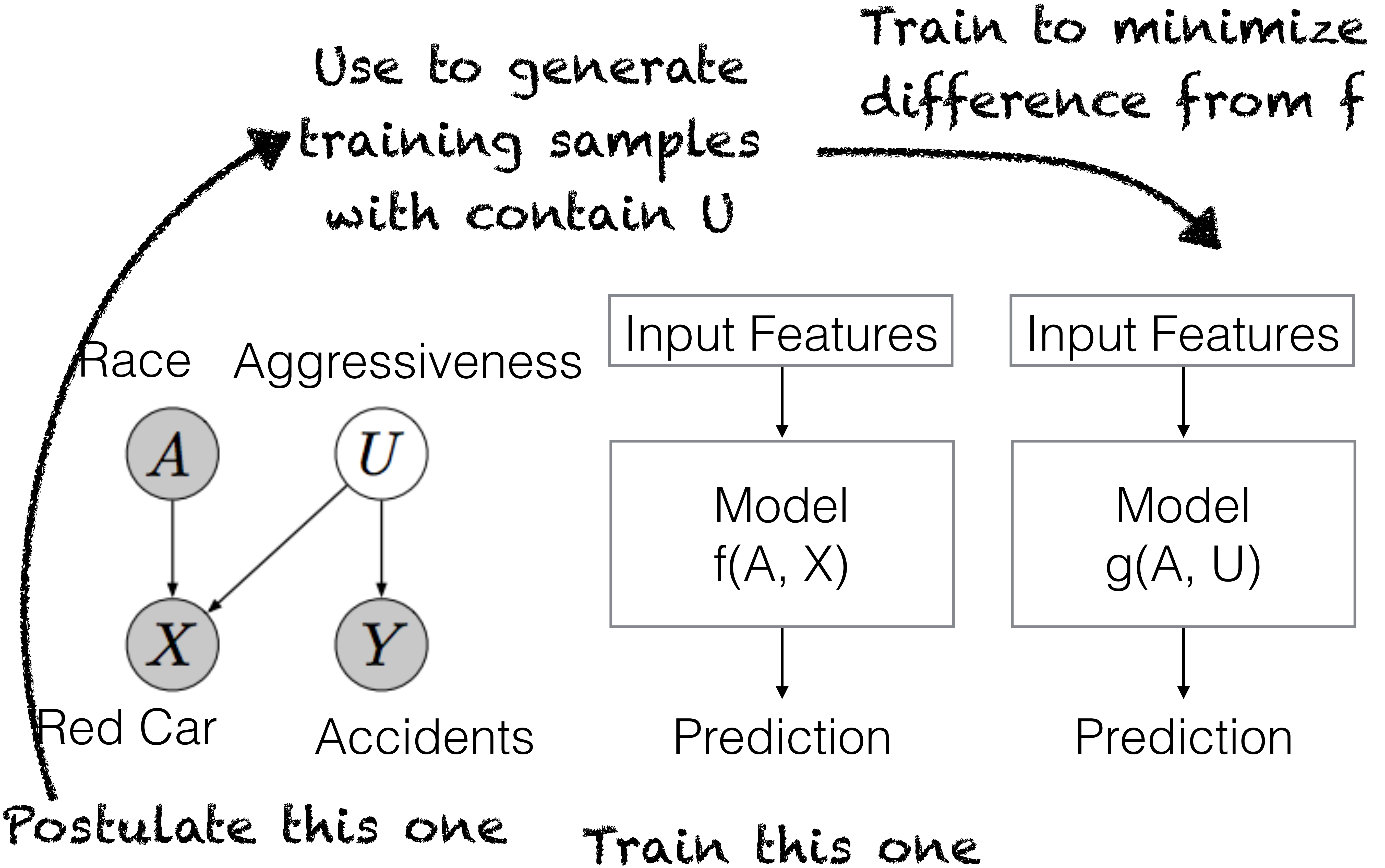


Postulate this one



Train this one

Counterfactual Fairness. Kusner et al (2017).



Counterfactual Fairness. Kusner et al (2017).

Use to generate training samples with contain U

Train to minimize difference from f

Ra

ures

Red

on

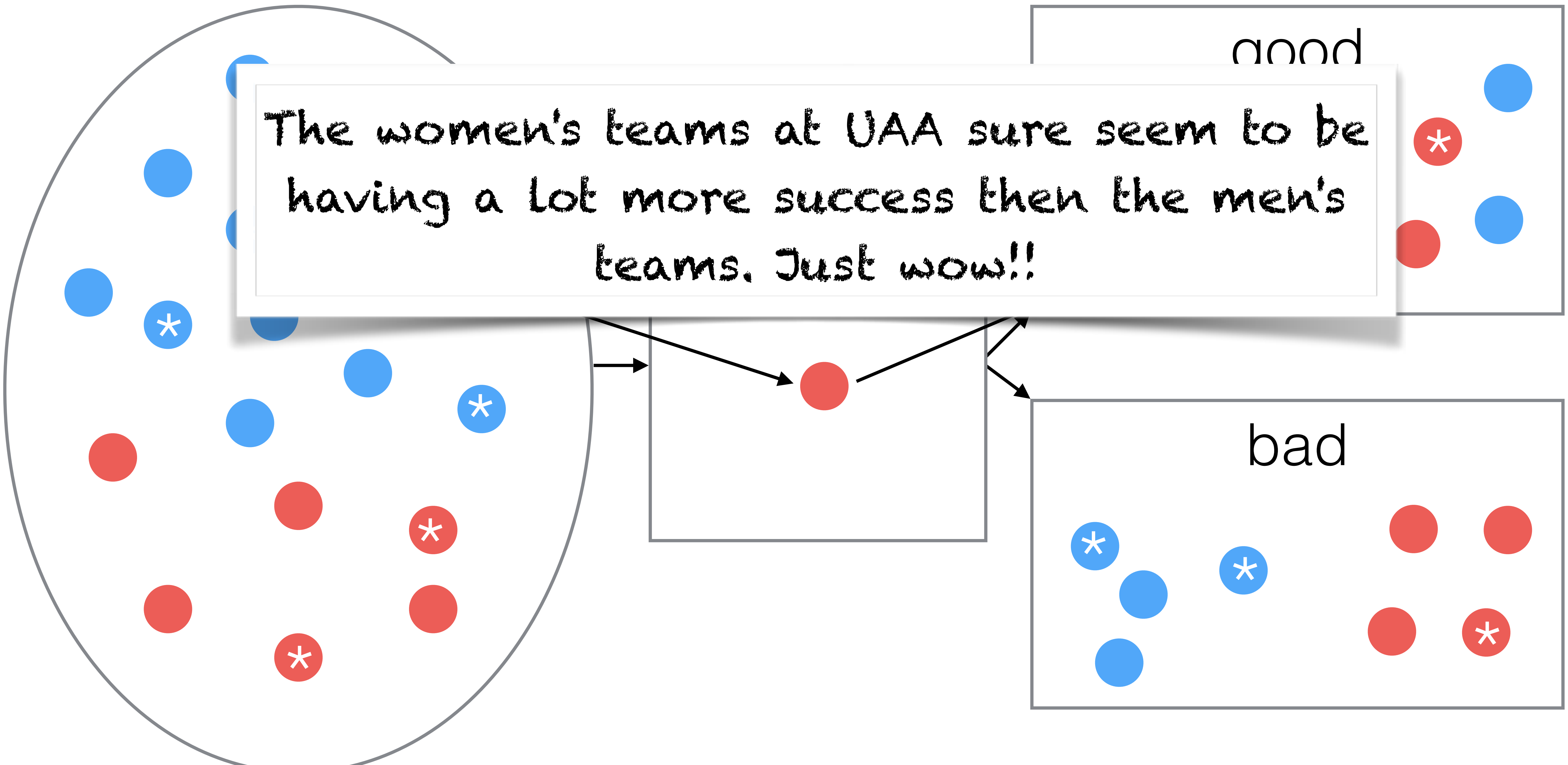
Post

“[Causal model] must be provided to [our fair learning algorithm]. Although this is well understood, it is worthwhile remembering that **causal models always require strong assumptions**…Having passed testable implications, the remaining components of a counterfactual model **should be understood as conjectures formulated according to the best of our knowledge**. Such models should be deemed provisional and prone to modifications…”

Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

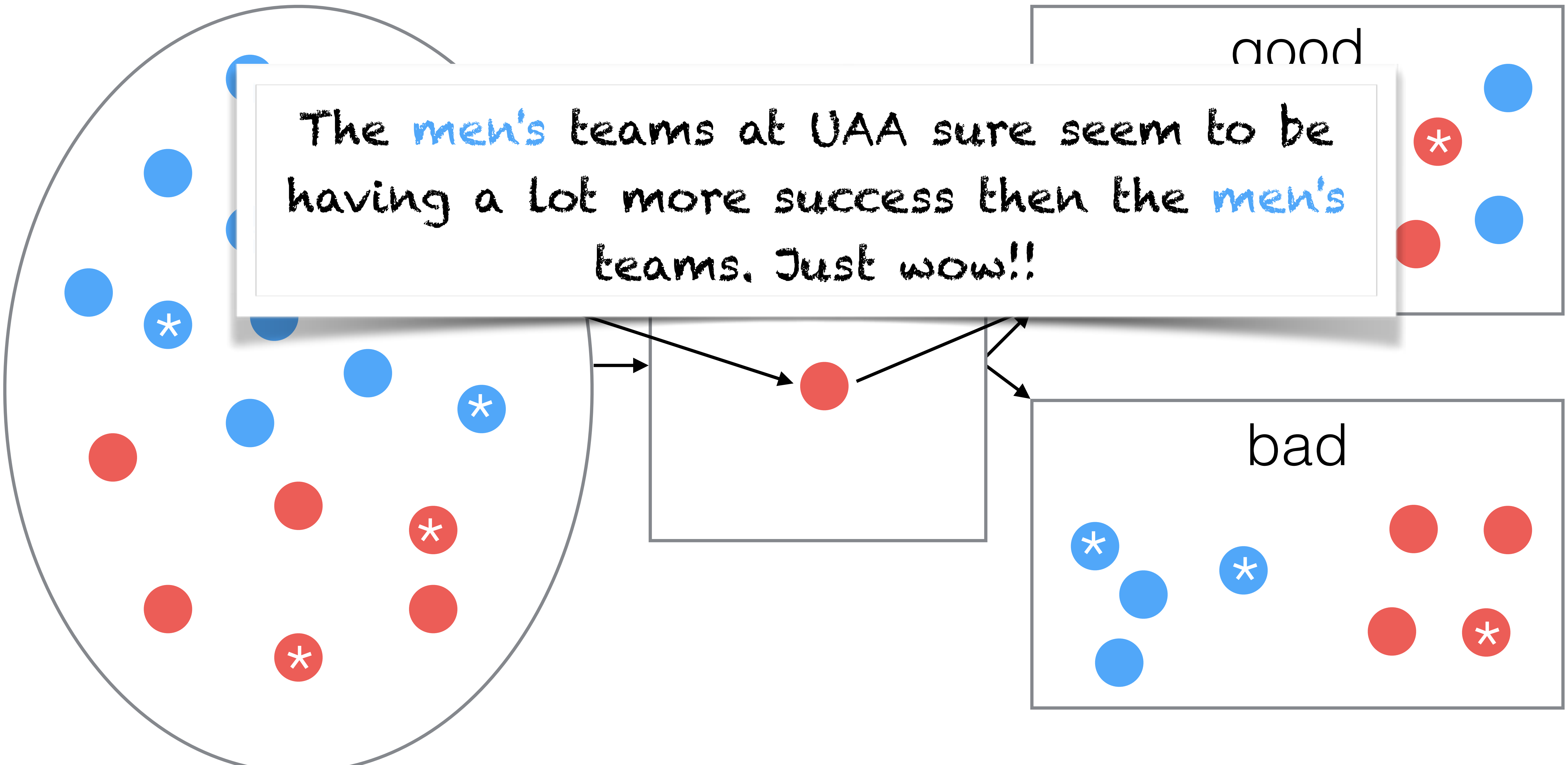
"Would you say that if I were white?" approach



https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

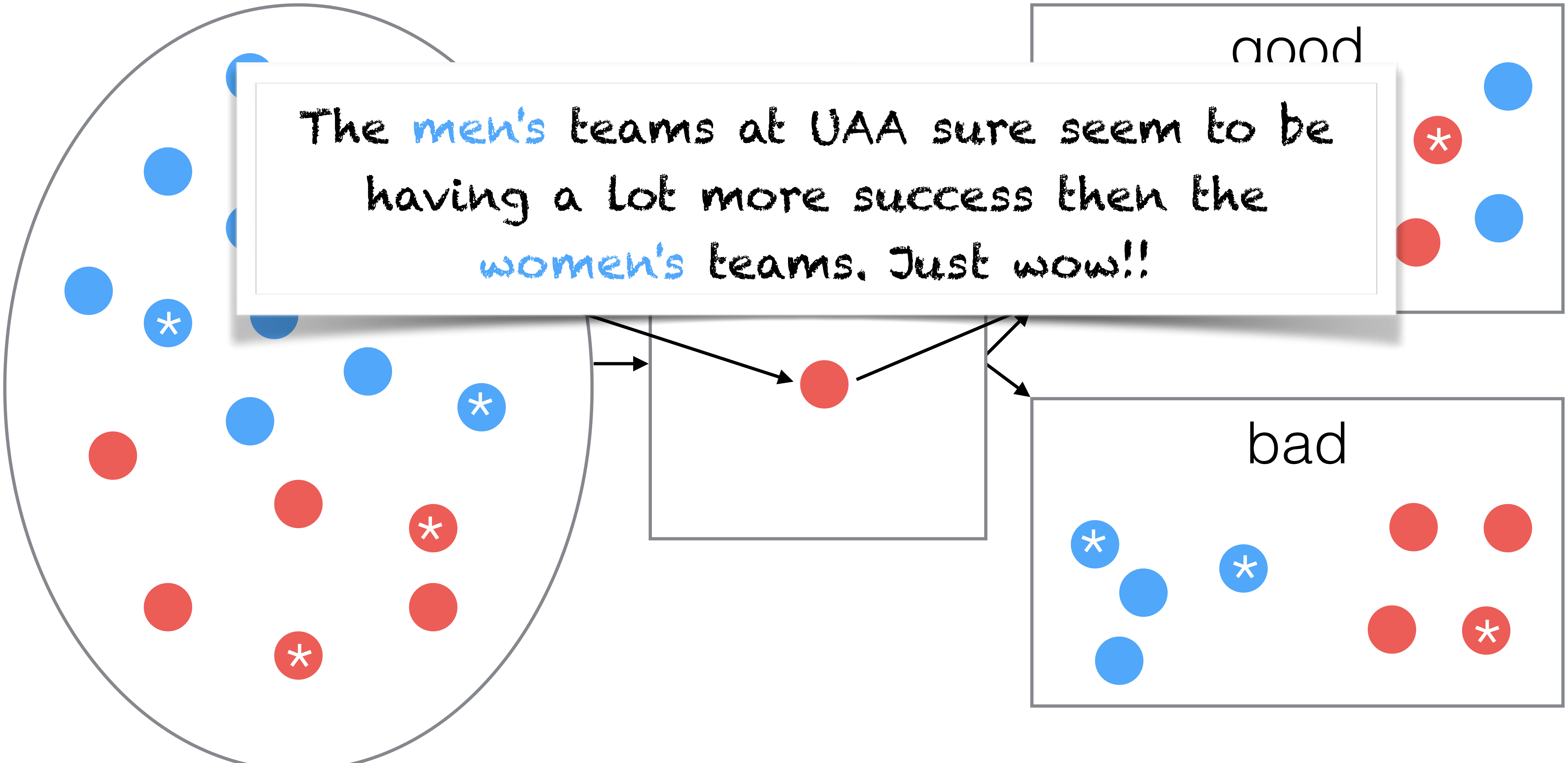
"Would you say that if I were white?" approach



https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

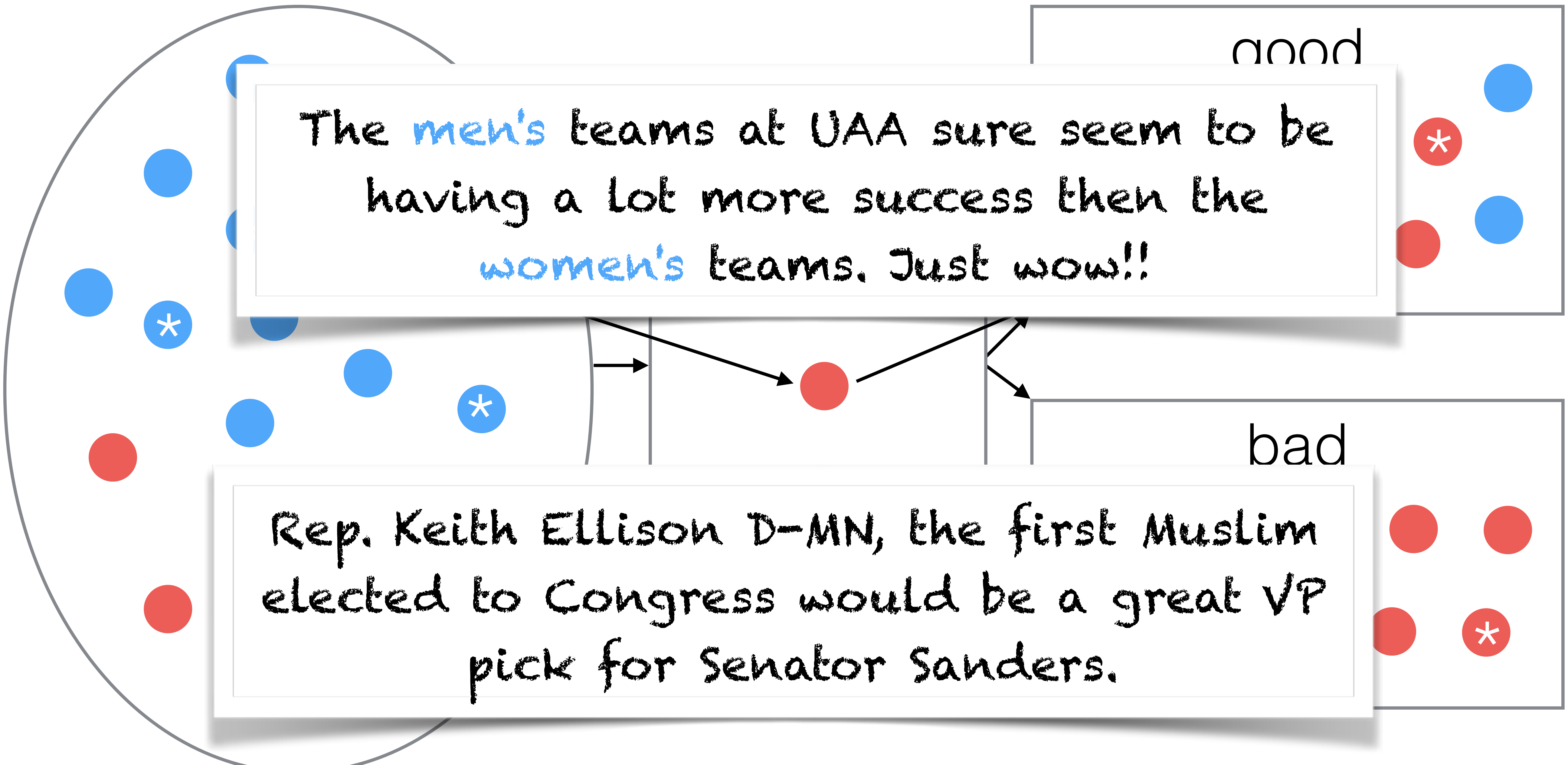
“Would you say that if I were white?” approach



https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

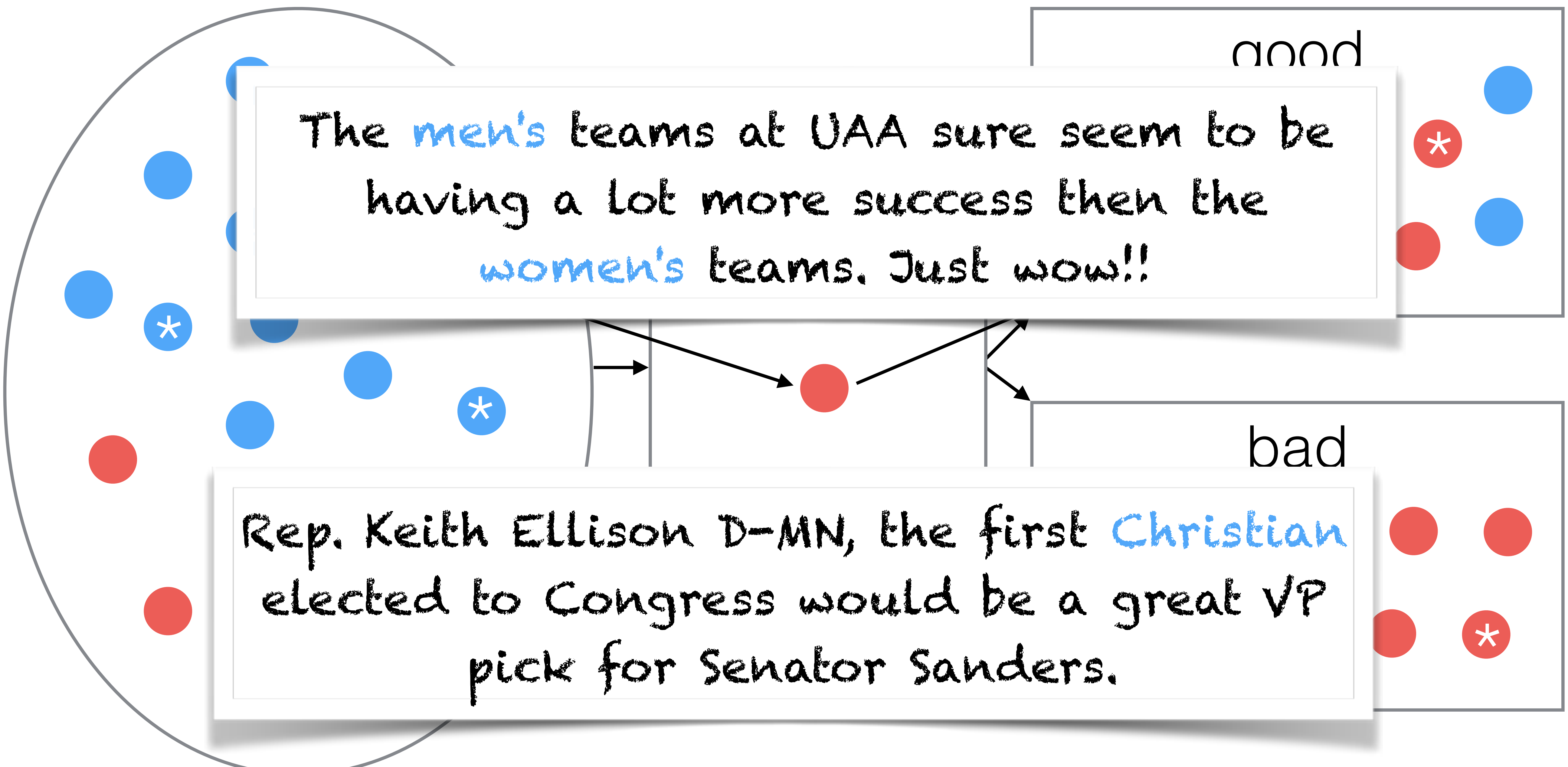
“Would you say that if I were white?” approach



https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

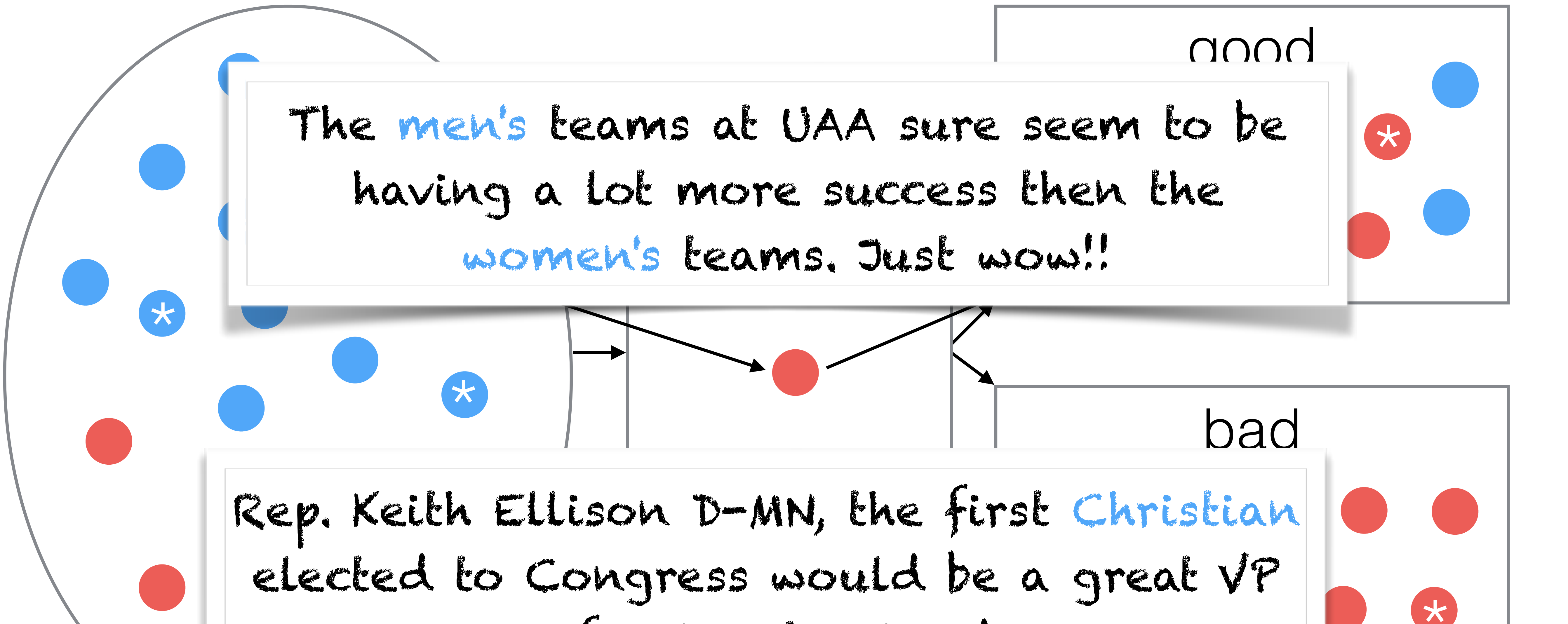
"Would you say that if I were white?" approach



https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
Counterfactual Fairness. Kusner et al (2017).

Counterfactual Fairness

"Would you say that if I were white?" approach

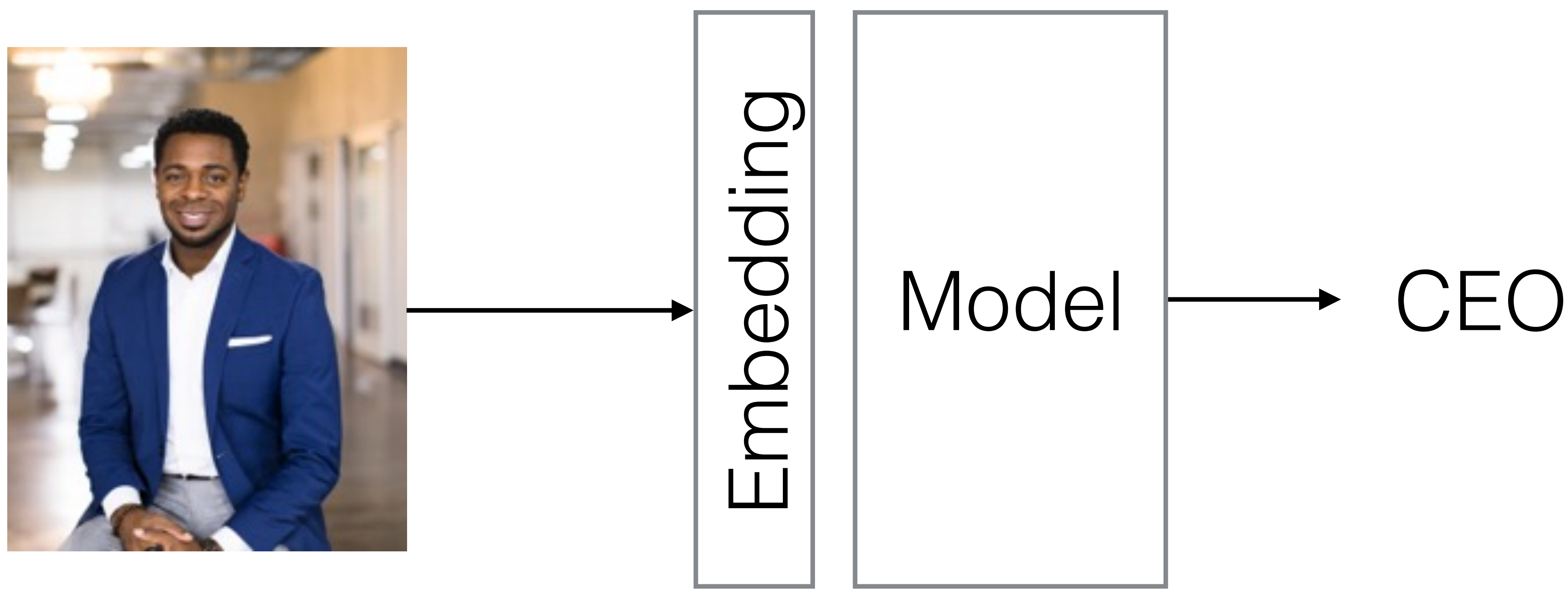


The men's teams at UAA sure seem to be having a lot more success then the women's teams. Just wow!!

Rep. Keith Ellison D-MN, the first Christian elected to Congress would be a great VP

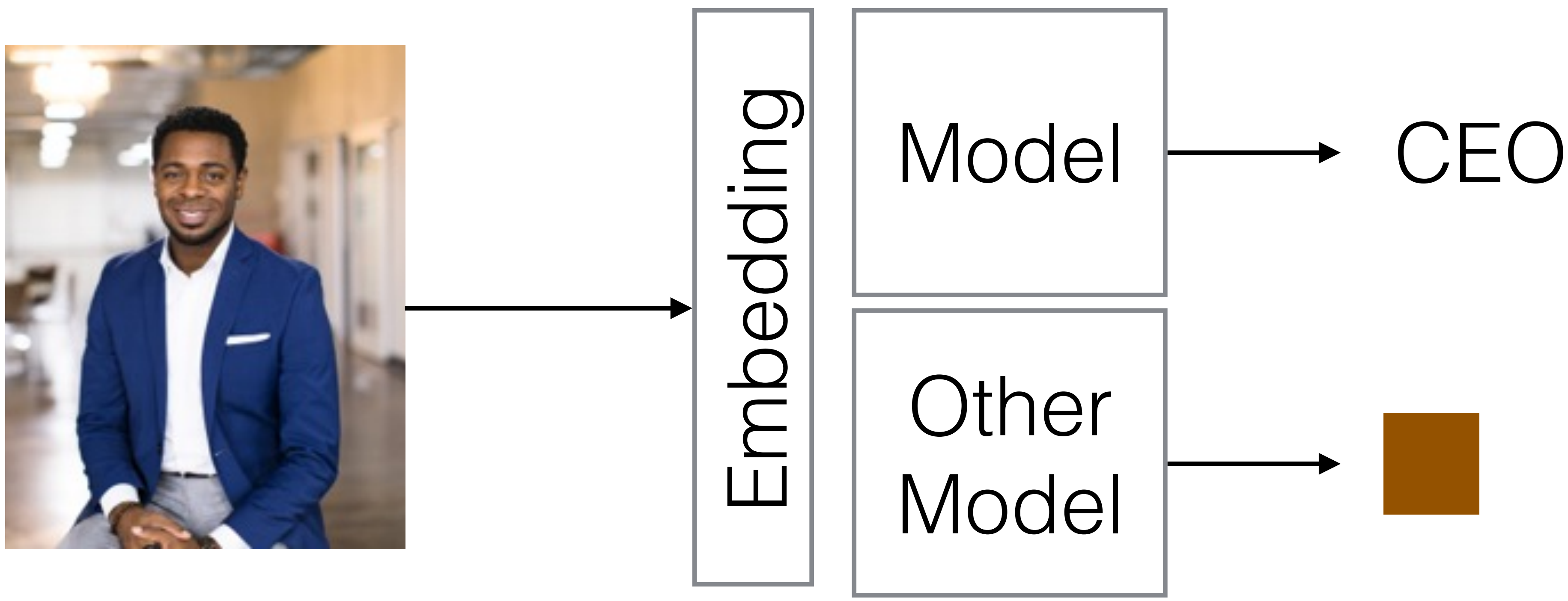
https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

Adversarial Objectives



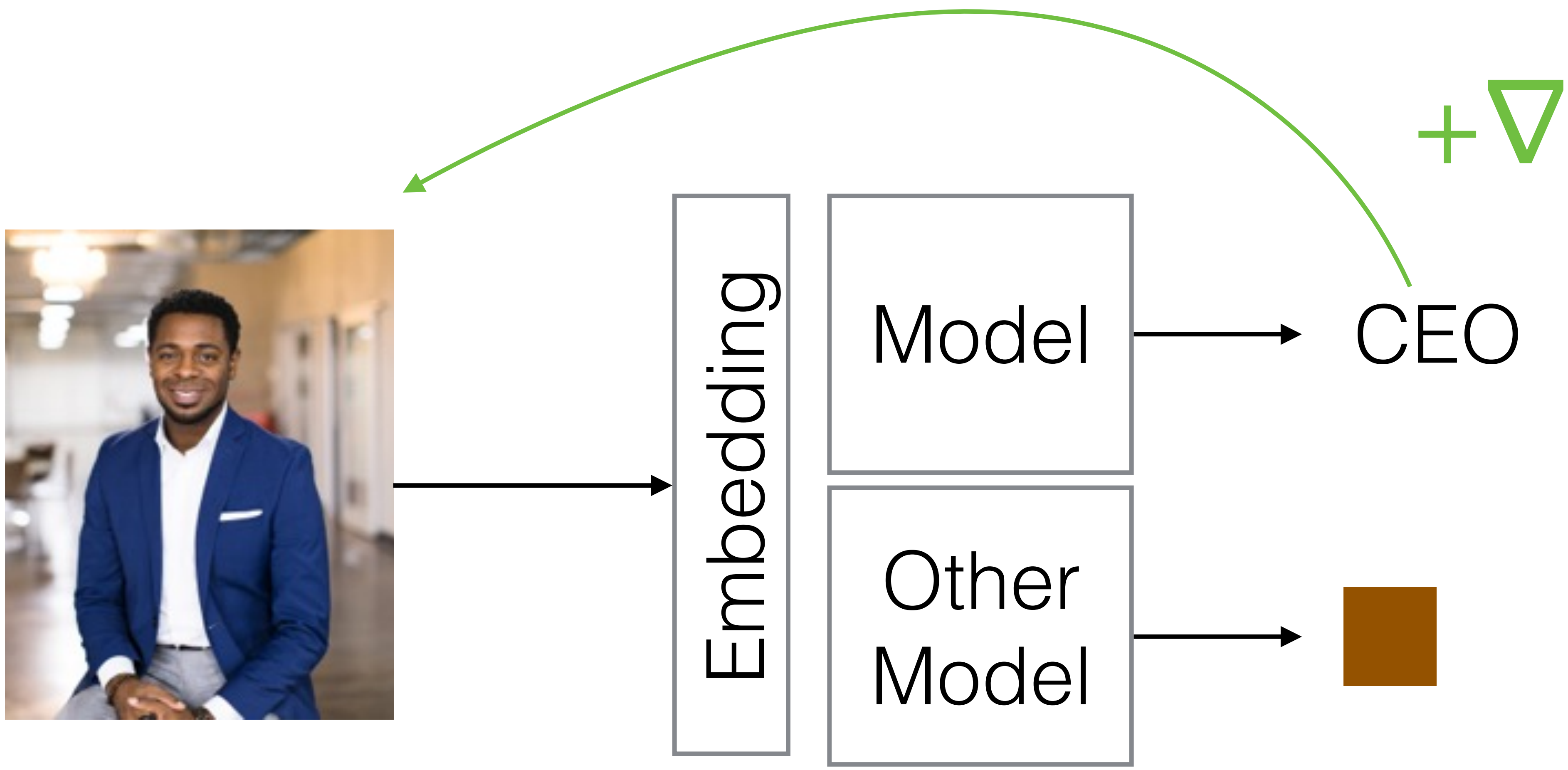
Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Adversarial Objectives



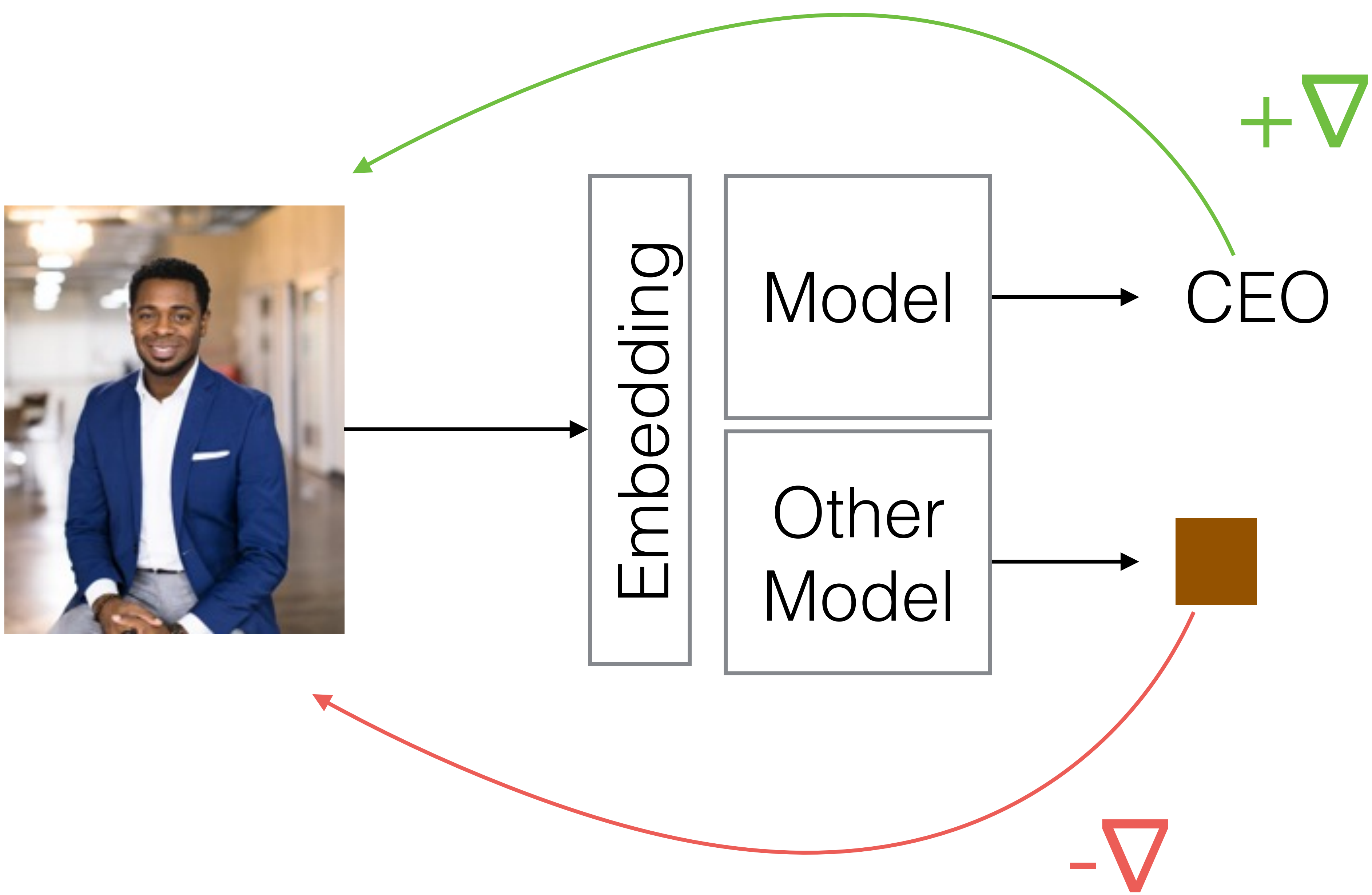
Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Adversarial Objectives



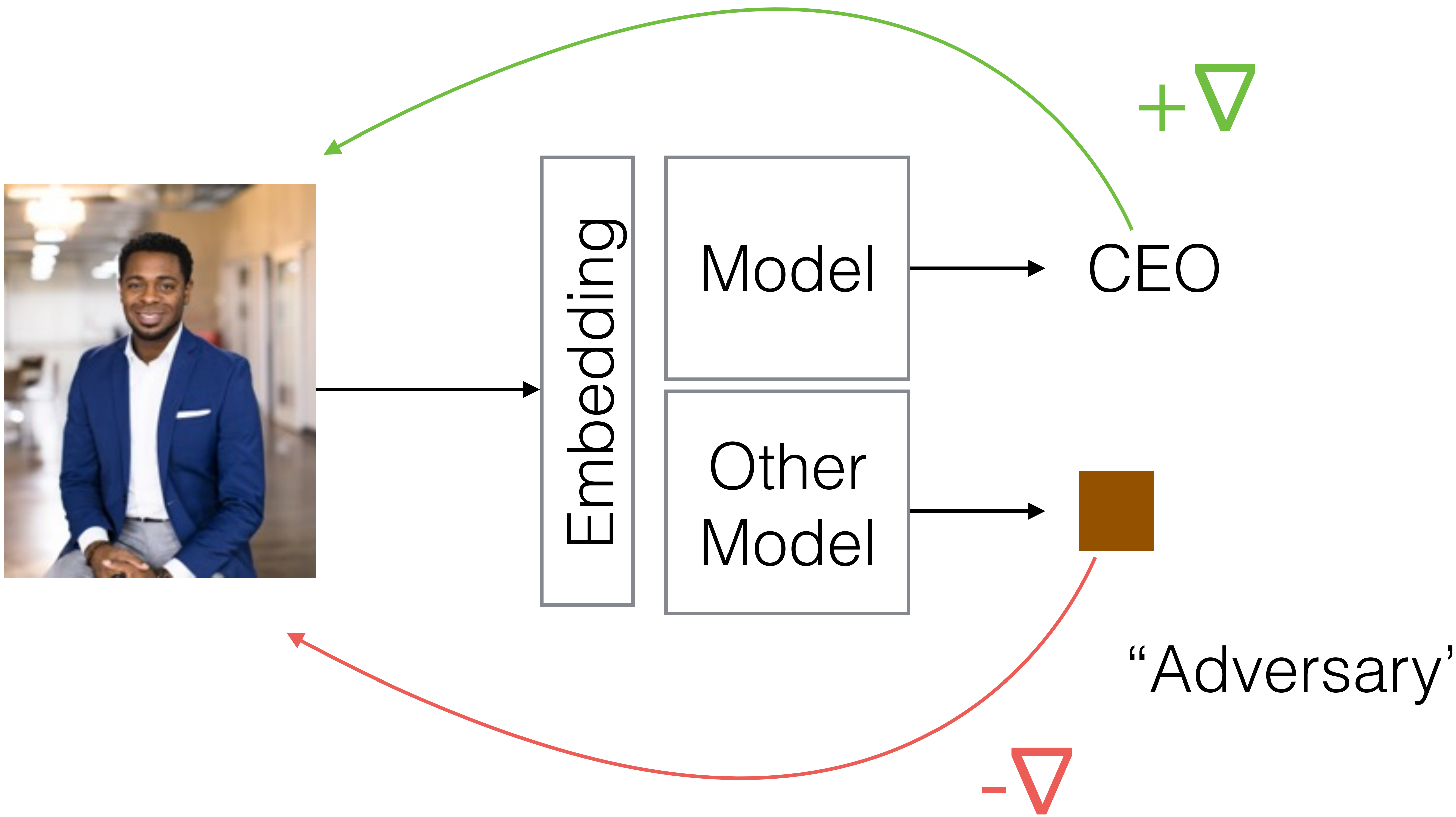
Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Adversarial Objectives



Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Adversarial Objectives



Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Adversarial Objectives

biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

Table 1: Completions for `he : she :: doctor : ?`

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes > $50K annually
marital_status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
workclass	Cat	Employer type

| Objectives

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

| Objectives

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes > $50K annually
marital_status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
workclass	Cat	Employer type

Without Debiasing			With Debiasing		
Female	Pred 0	Pred 1	*Female*	Pred 0	Pred 1
True 0	4711	120	True 0	4518	313
True 1	265	325	True 1	263	327
Male	Pred 0	Pred 1	*Male*	Pred 0	Pred 1
True 0	6907	697	True 0	7071	533
True 1	1194	2062	True 1	1416	1840

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

| Objectives

Feature	**Type**	**Description**
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes > $50K annually
marital_status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
workclass	Cat	Employer type

Without Debiasing			With Debiasing		
Female	Pred 0	Pred 1	*Female*	Pred 0	Pred 1
True 0	4711	120	True 0	4518	313
True 1	265	325	True 1	263	327
Male	Pred 0	Pred 1	*Male*	Pred 0	Pred 1
True 0	6907	697	True 0	7071	533
True 1	1194	2062	True 1	1416	1840

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes > $50K annually
marital_status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
workclass	Cat	Employer type

| Objectives

Without Debiasing			With Debiasing		
Female	<50K	>50K	*Female*	<50K	>50K
<50K	4711	120	True 0	4518	313
>50K	265	325	True 1	263	327
Male	<50K	>50K	*Male*	<50K	>50K
<50K	6907	697	True 0	7071	533
>50K	1194	2062	True 1	1416	1840

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

| Objectives

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whetl nually
marital_status	Cat	Marit
native_country	Cat	Count
occupation	Cat	Occu
race	Cat	White Indiar
relationship	Cat	Wife, famil
sex	Cat	Femal
workclass	Cat	Empl

	Female		Male	
	Without	With	Without	With
FPR	0.0248	0.0647	0.0917	0.0701
FNR	0.4492	0.4458	0.3667	0.4349

Without Debiasing			With Debiasing		
Female	<50K	>50K	*Female*	<50K	>50K
<50K	4711	120	True 0	4518	313
>50K	265	325	True 1	263	327
Male	<50K	>50K	*Male*	<50K	>50K
<50K	6907	697	True 0	7071	533
>50K	1194	2062	True 1	1416	1840

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whetl nually
marital_status	Cat	Marit
native_country	Cat	Count
occupation	Cat	Occu
race	Cat	White Indiar
relationship	Cat	Wife, family
sex	Cat	Femal
workclass	Cat	Empl

| Objectives

	Female		Male	
	Without	With	Without	With
FPR	0.0248	0.0647	0.0917	0.0701
FNR	0.4492	0.4458	0.3667	0.4349

Without Debiasing			With Debiasing		
Female	<50K	>50K	*Female*	<50K	>50K
<50K	4711	120	True 0	4518	313
>50K	265	325	True 1	263	327
Male	<50K	>50K	*Male*	<50K	>50K
<50K	6907	697	True 0	7071	533
>50K	1194	2062	True 1	1416	1840

Mitigating Unwanted Biases with Adversarial Learning. Mitchell et al (2018).

Core Components of ML

Input Data

Model

Objective

Core Components of ML

Input Data

Model

Objective

Interpretability/Transparency

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

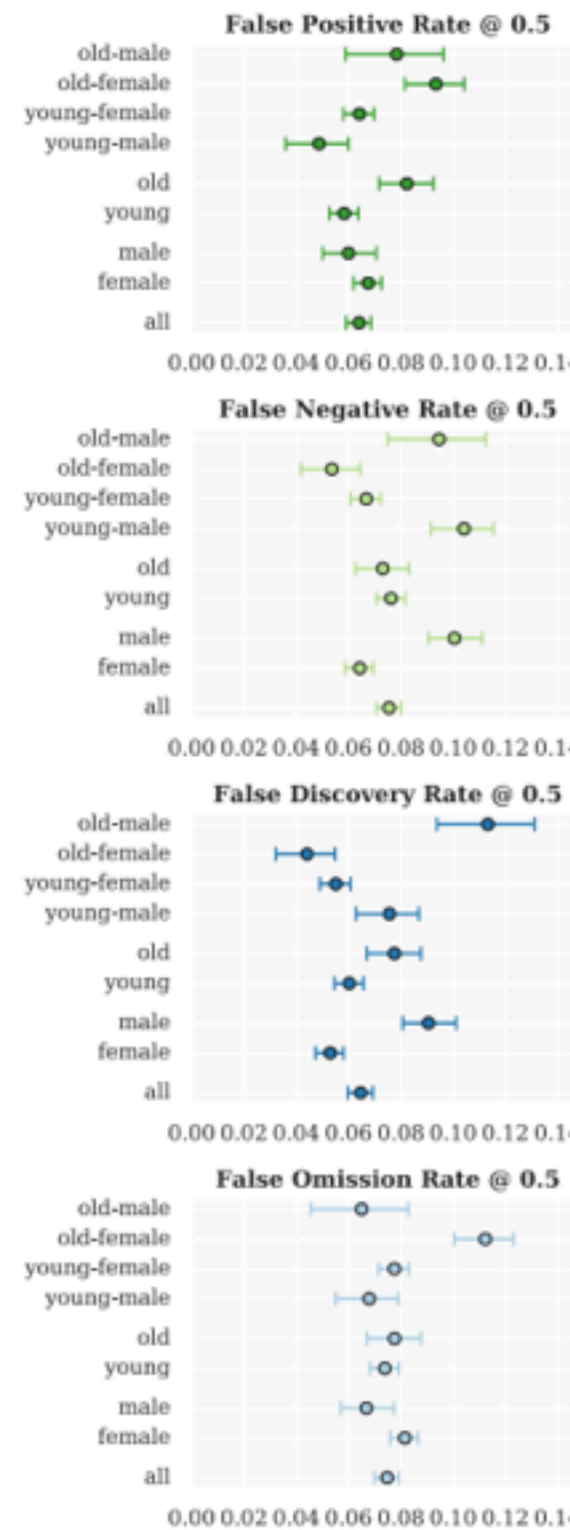
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses



Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Model Cards for Model Reporting Mitchell et al. (2018).

Testing with Concept Activation Vectors (TCAV)

Model
f(x)

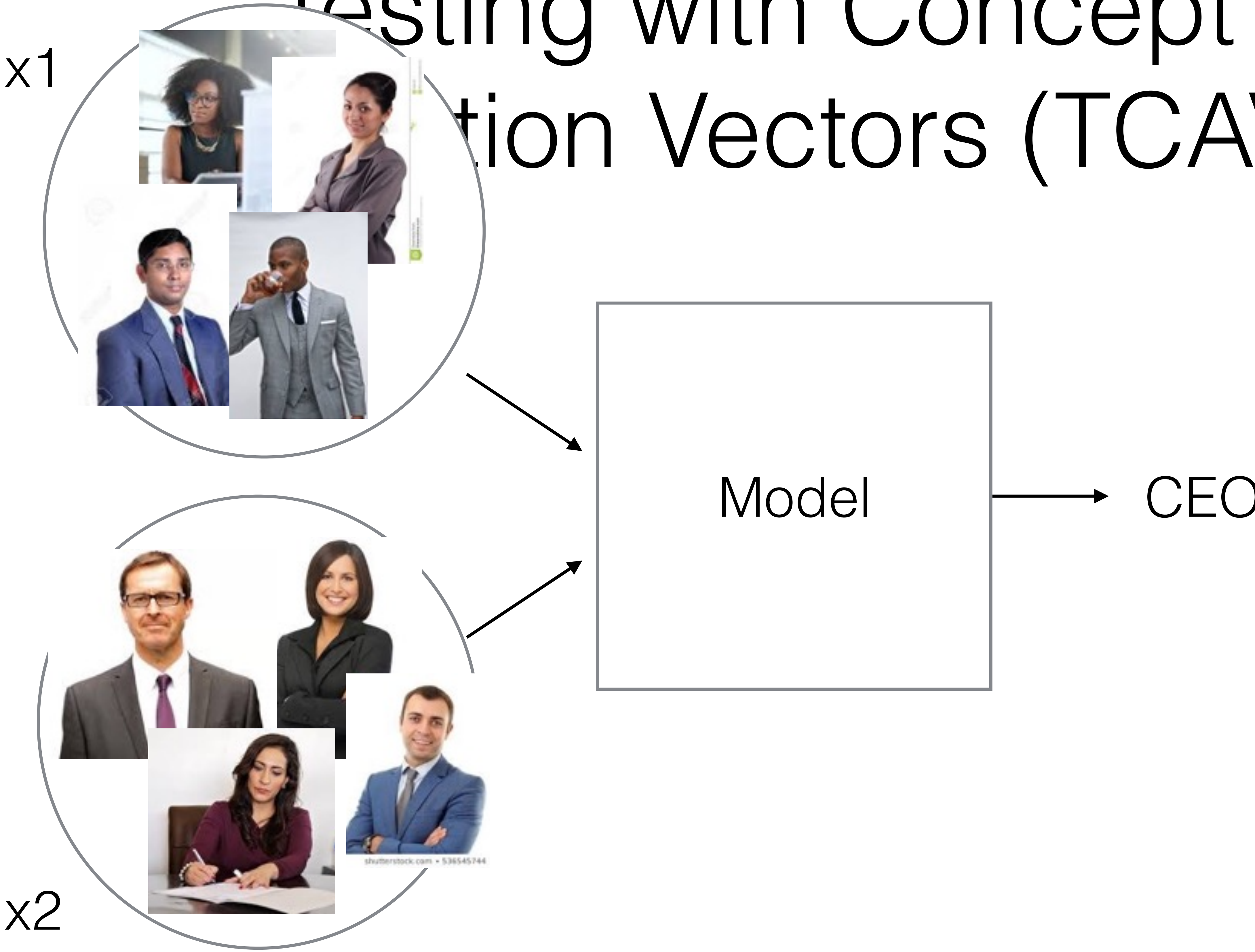
CEO

Input
x

Label
k

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

Testing with Concept …tion Vectors (TCAV)



Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

Testing with Concept Activation Vectors (TCAV)

x1

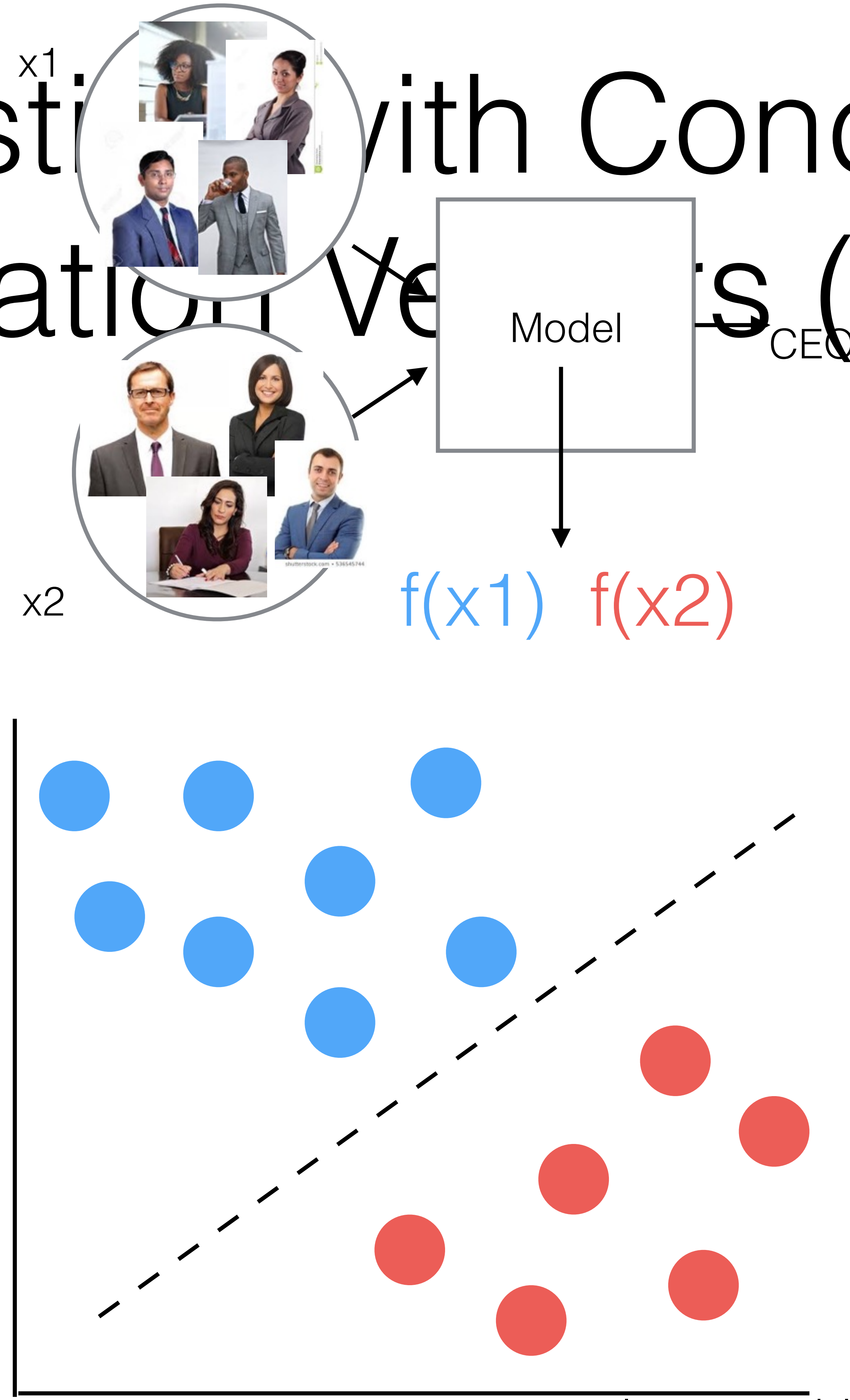
Model → CEO

{f(x1)} {f(x2)}

x2

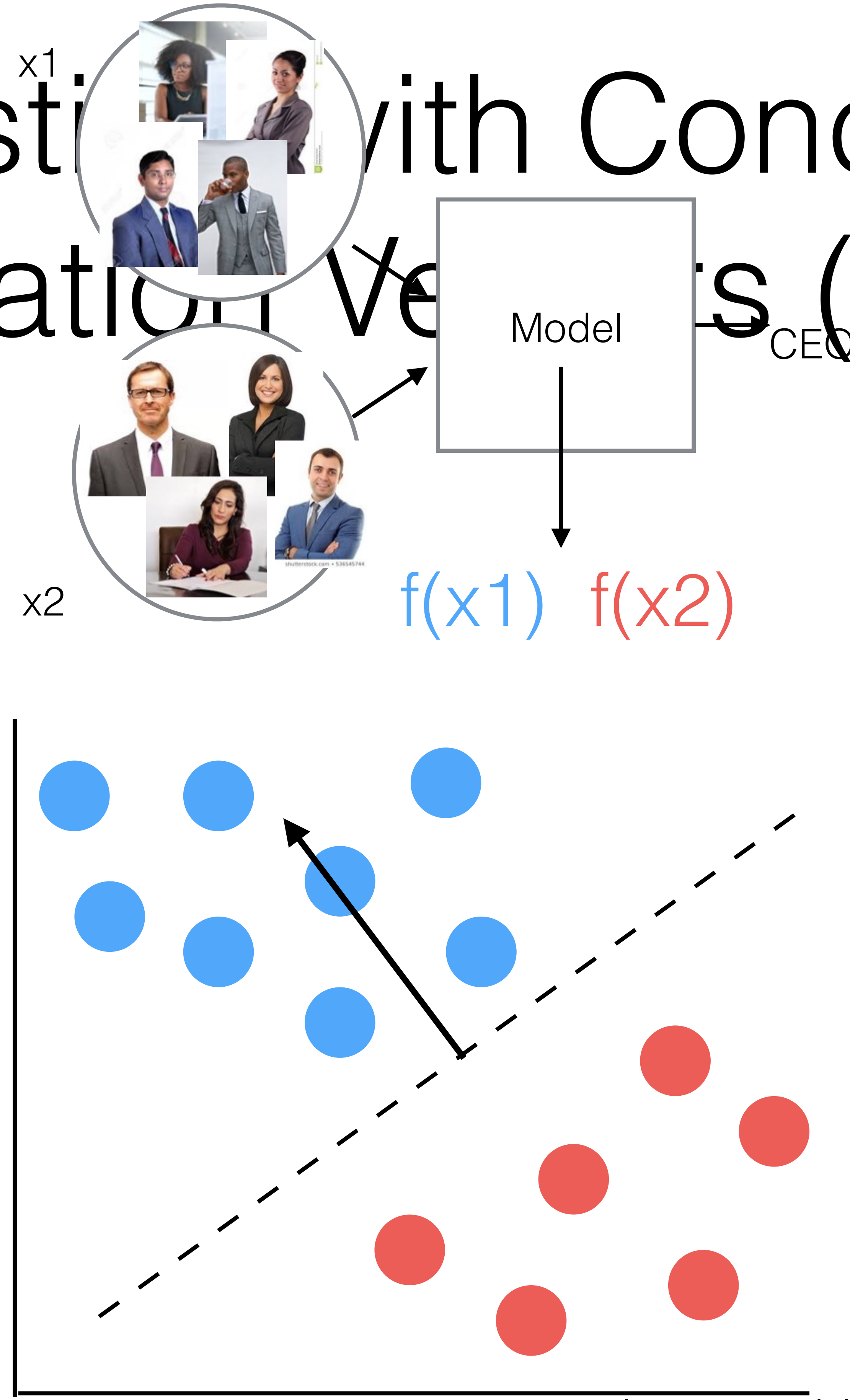
Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

Testing with Concept Activation Vectors (TCAV)



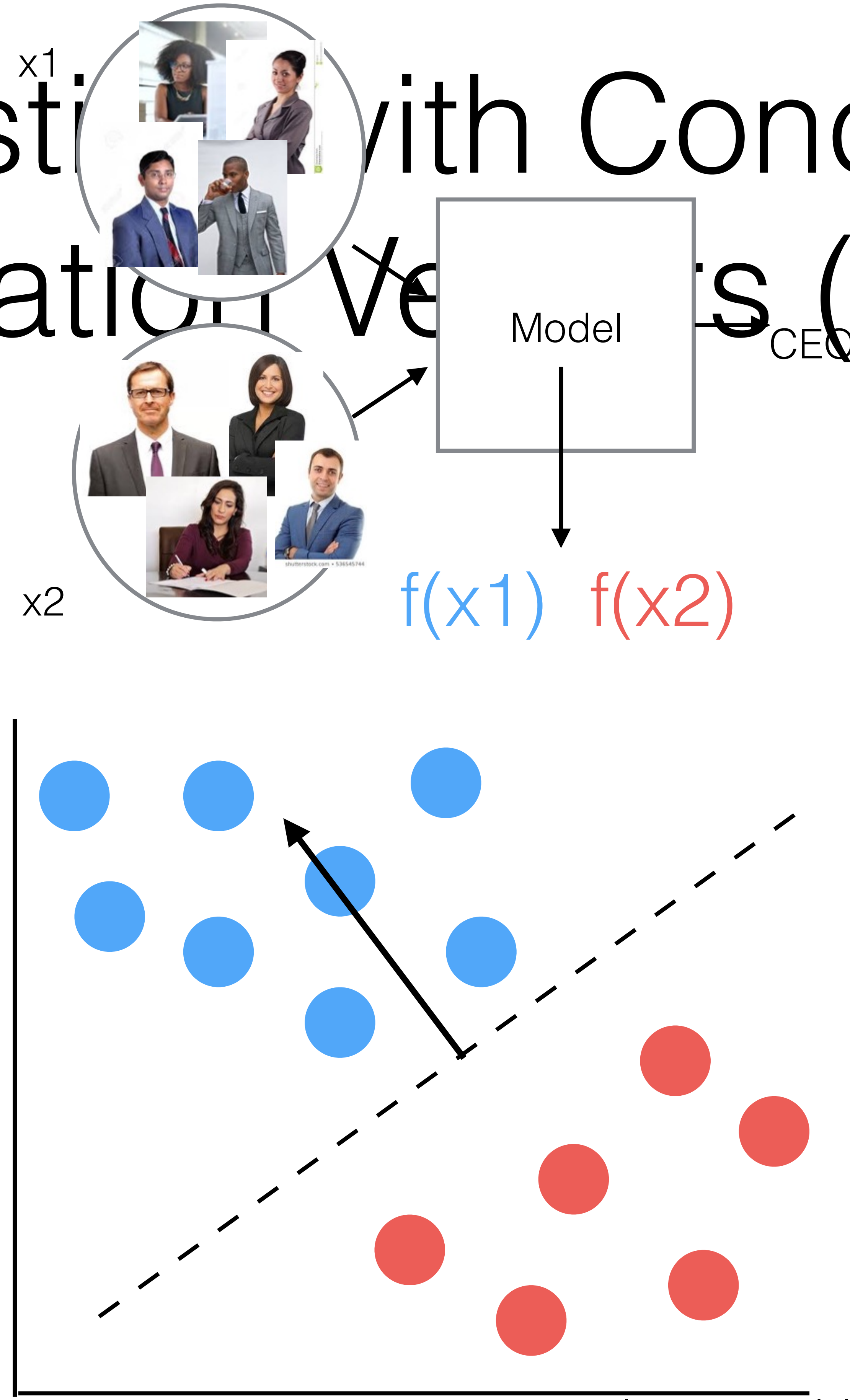
Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

Testing with Concept Activation Vectors (TCAV)



Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

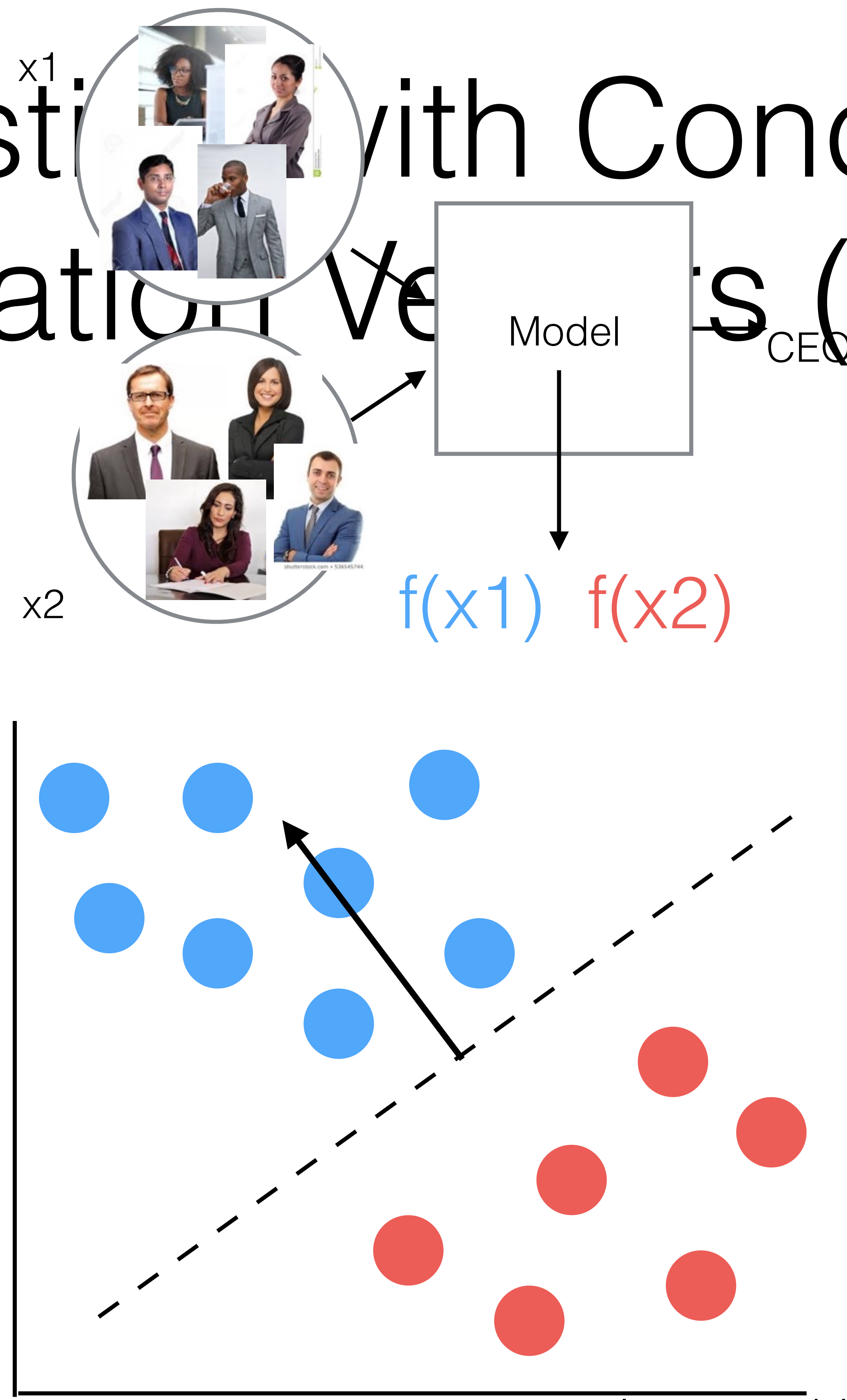
Testing with Concept Activation Vectors (TCAV)



"Content Activation Vector"

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

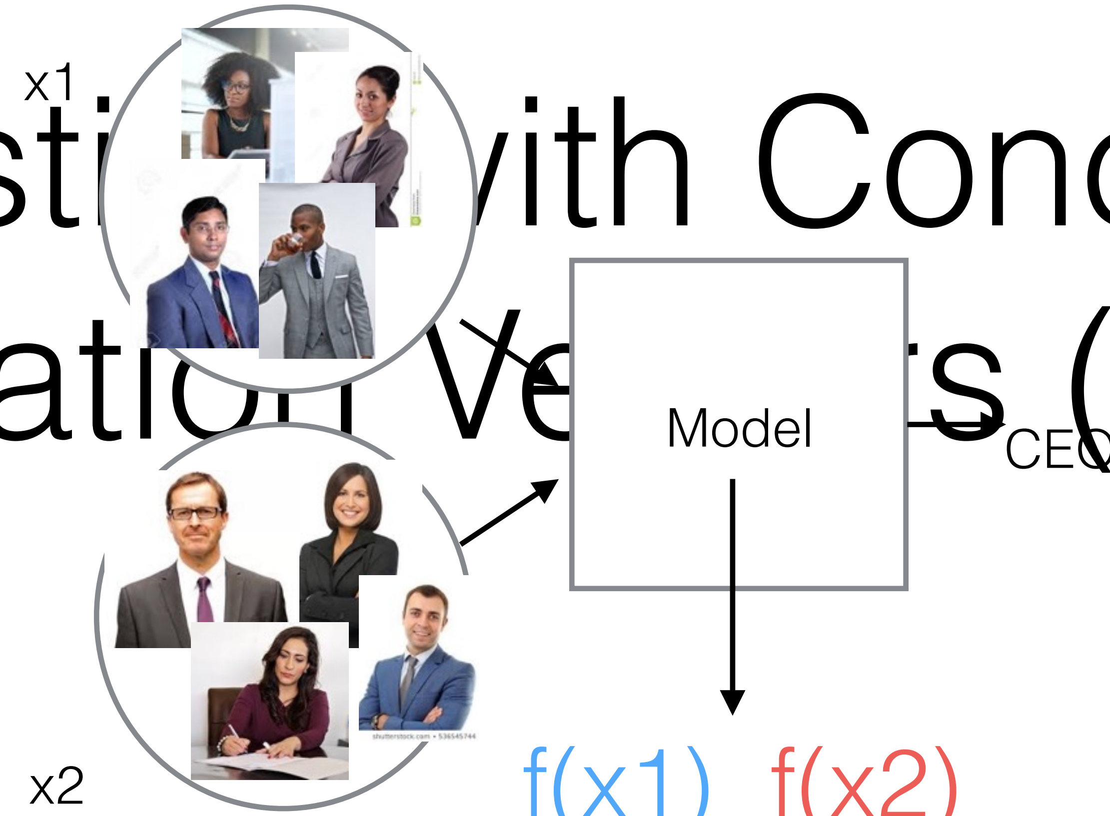
Testing with Concept Activation Vectors (TCAV)

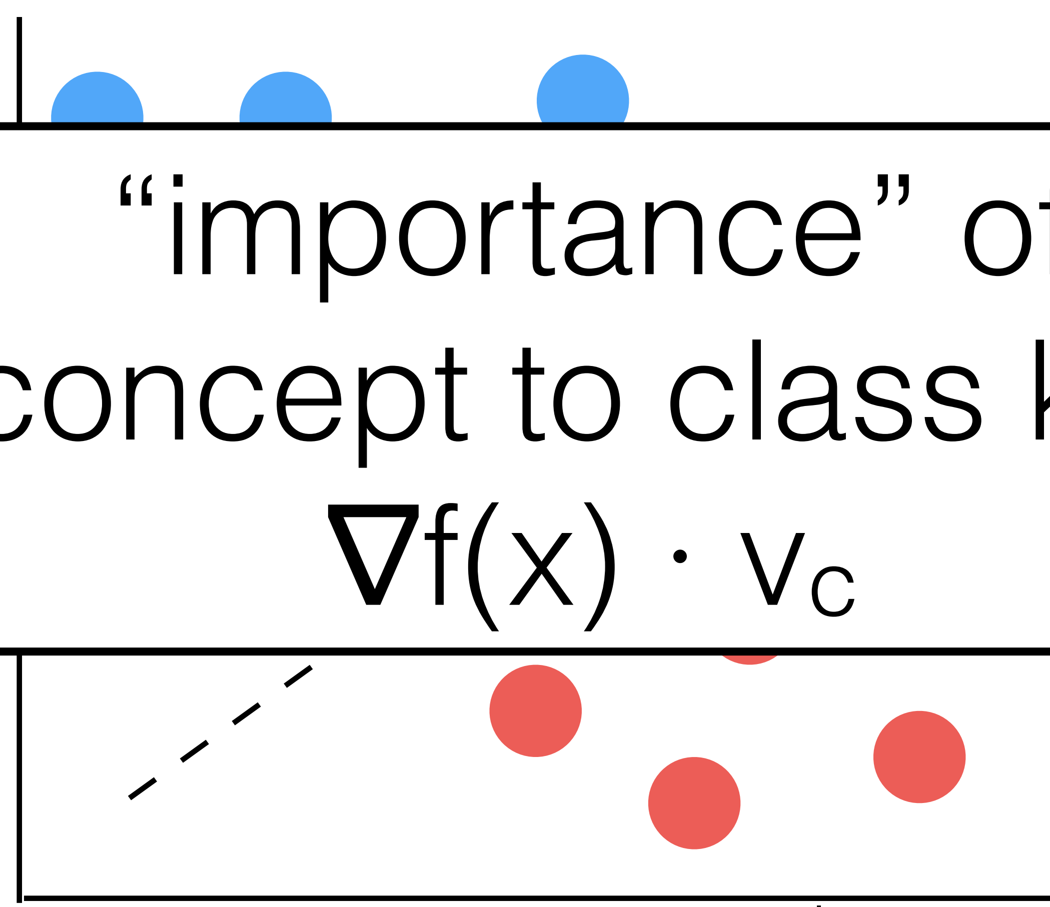


"Content Activation Vector" (v_C)

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

Testing with Concept Activation Vectors (TCAV)





“importance” of concept to class k = $\nabla f(x) \cdot v_c$

Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

Testi with Concept Activati Ve s (TCAV)

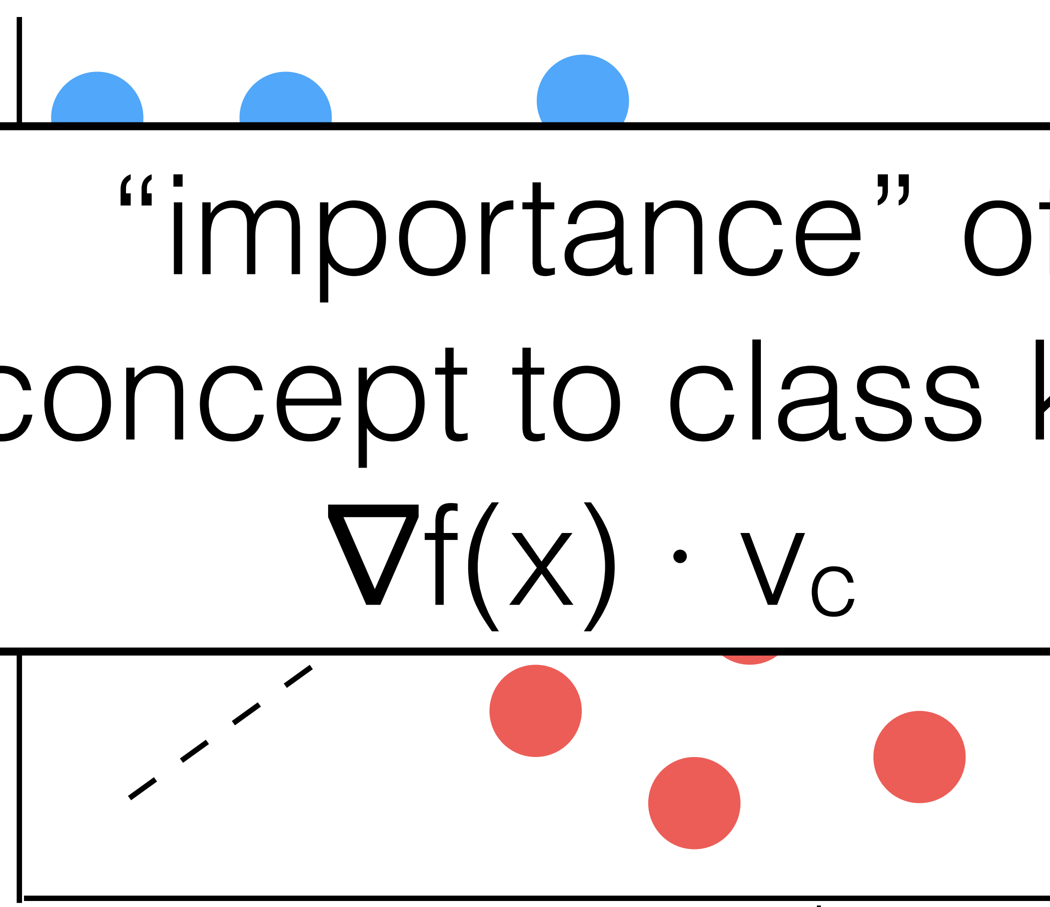
x1

x2

Model

f(x1) f(x2)

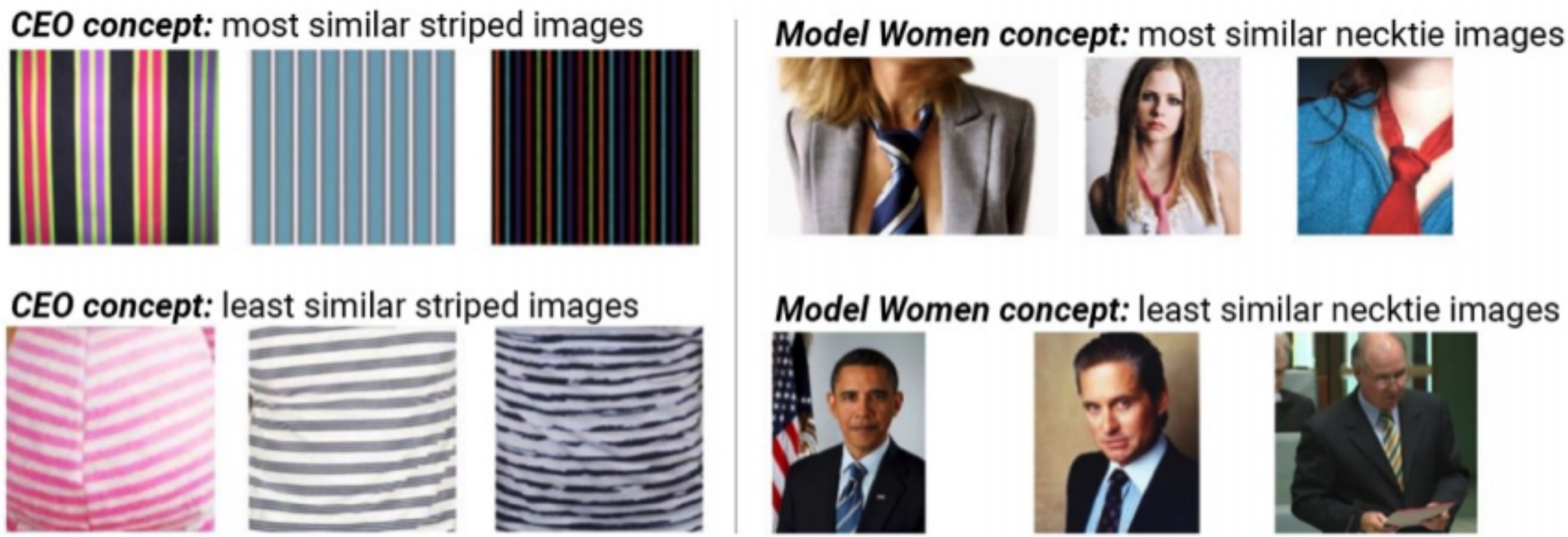
take lots of samples and test hypothesis that classifier performance is > 0.5



"importance" of concept to class k = $\nabla f(x) \cdot v_c$

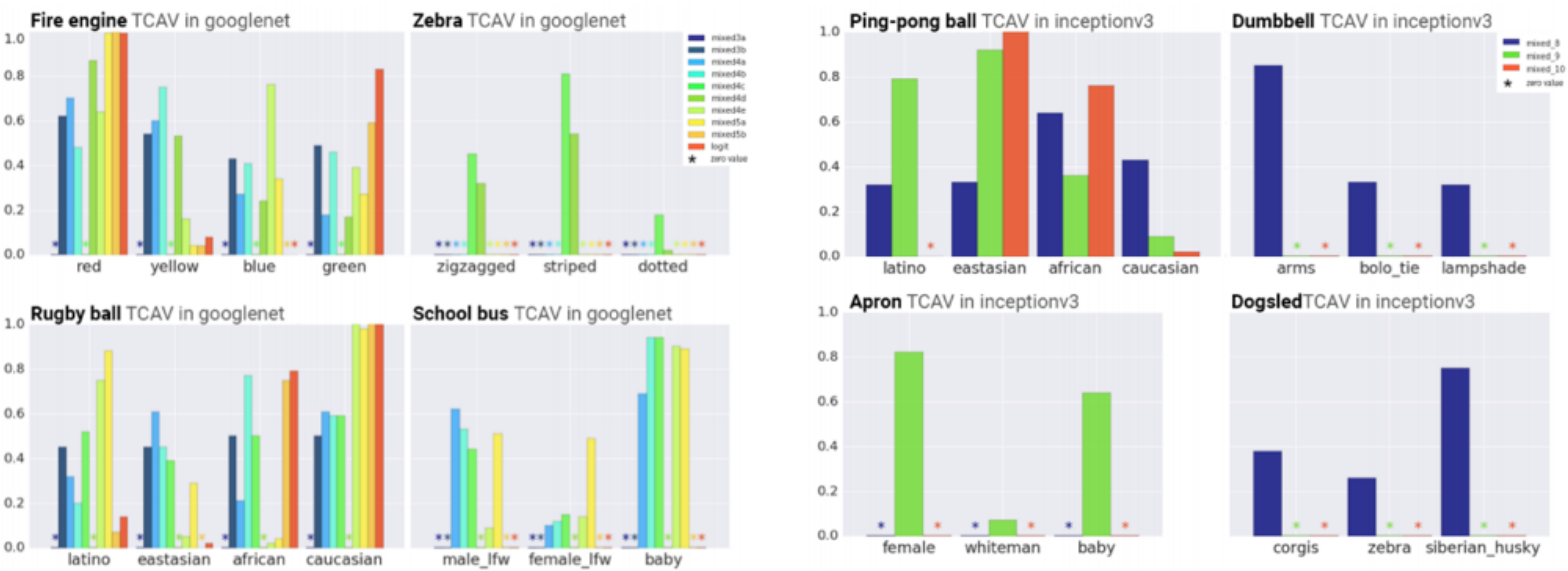
Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

TCAV



Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).

TCAV



Interpretability Beyond Feature Attribution:
Quantitative Testing with Concept Activation Vectors (TCAV). Kim et al (2018).