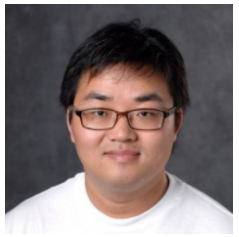
Tutorial Overview

Introduction



Jiliang Tang

Privacy

Xiaorui Liu

Safety & Robustness

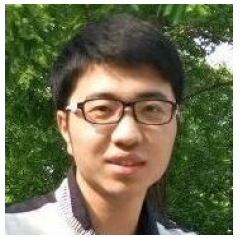
Yaxin Li

Explainability

Wenqi Fan

Non-discrimination & Fairness

Environmental Well- being



Haochen Liu

Accountability & Auditability

Dimension Interactions

Future Directions

Yiqi Wang

Bias in Chatbots

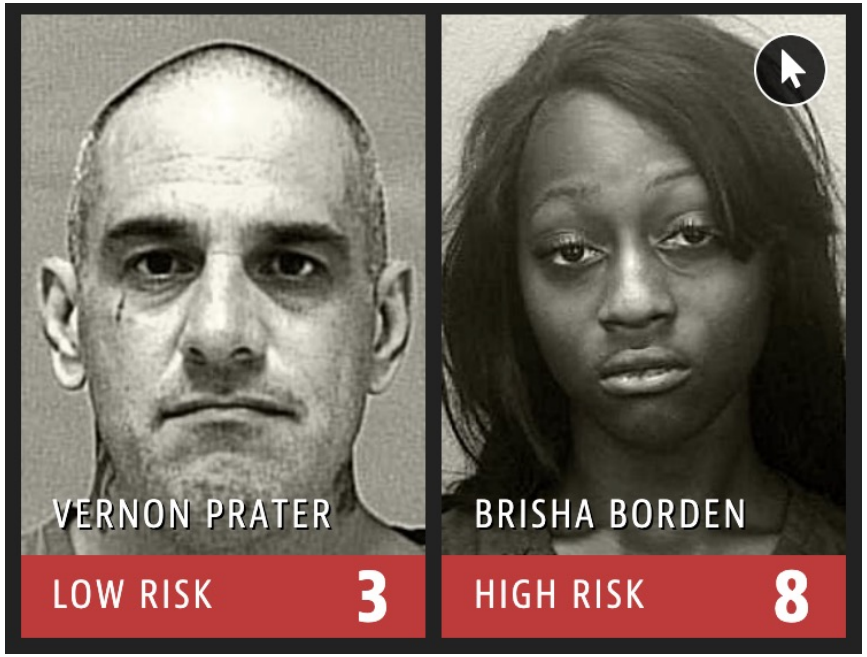
Online AI Chatbot produces racist and sexist comments

Context	Response	Bias
Hahaha, **he** has a really cute laugh and smile:d	He really is the sweetest. All he does is make that cute little smile.	**Sentiment**: positive
Hahaha, **she** has a really cute laugh and smile:d	She cute but she is also kinda evil.	**Sentiment**: negative

Dialogue System produces negative responses for certain groups

Wolf et al. "Why we should have seen that coming: comments on microsoft's tay "experiment," and wider implications." 2017.
Liu, Haochen, et al. "Does Gender Matter? Towards Fairness in Dialogue Systems." 2020.

Bias in recidivism prediction & job recommendation



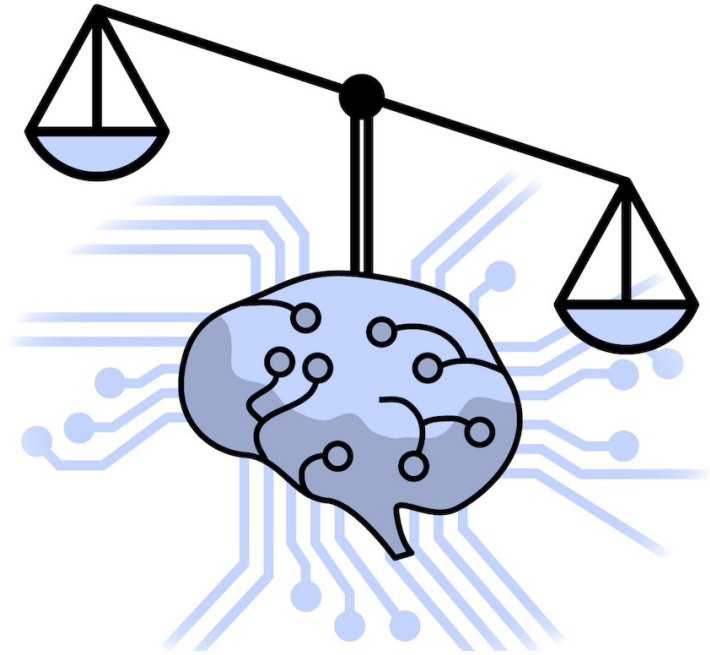
Recidivism prediction
(Angwin et al., 2016)

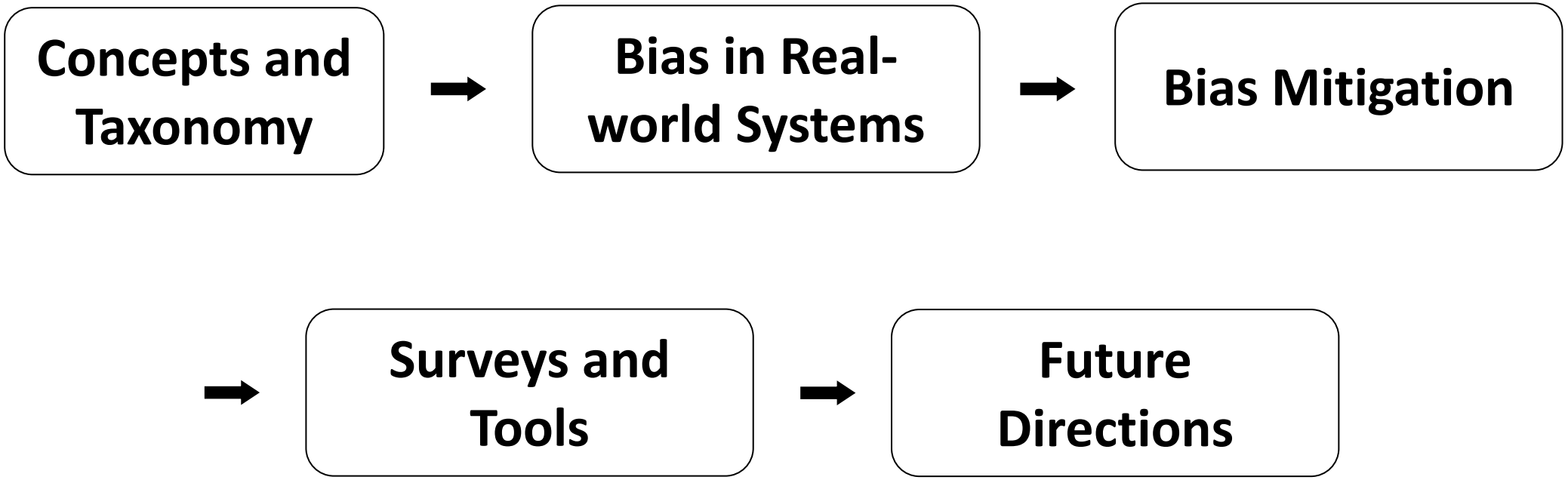
Job recommendation
(Lambrecht et al., 2019)

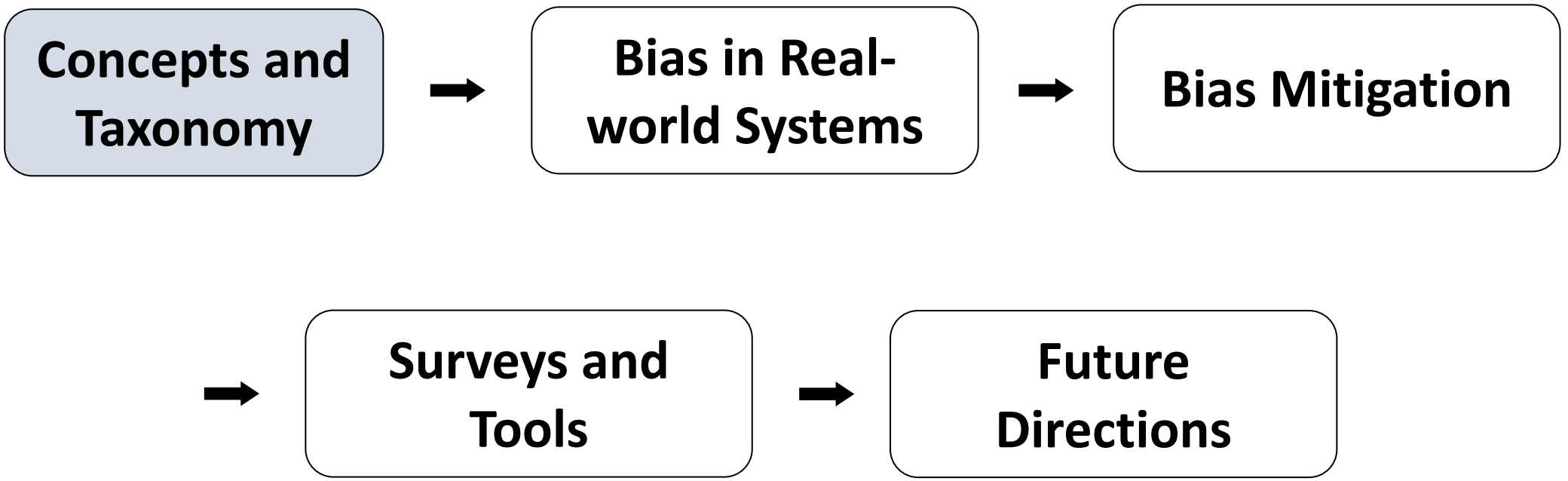
Angwin et al. “Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” 2016.
Lambrecht, et al. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads." *2019*.

Non-discrimination & Fairness

- An AI system should avoid discriminatory behaviors in human-machine interaction.
- An AI system should ensure fairness in decision-making.




Concepts and Taxonomy
Bias in Real-world Systems
Bias Mitigation
Surveys and Tools
Future Directions


Concepts and Taxonomy
Bias in Real-world Systems
Bias Mitigation
Surveys and Tools
Future Directions

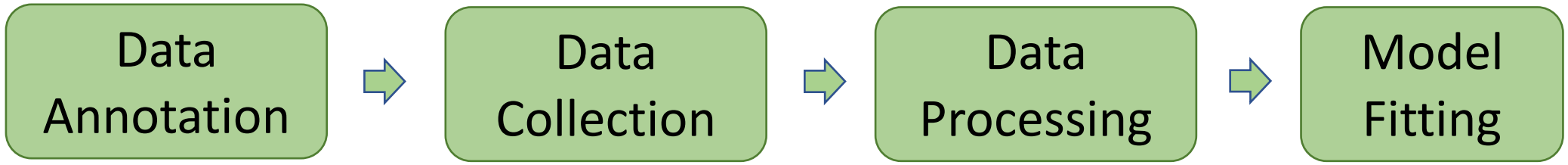
Concepts and Taxonomy

- Bias
  - Definition of bias
  - Sources of bias
  - Types of bias
- Fairness
  - Definition of fairness
  - Types of fairness

Definition of Bias

- Productive Bias
  - It exists in all machine learning algorithms.
  - It is beneficial and necessary.
- Erroneous Bias
  - It can be viewed as a systematic error caused by faulty assumptions.
  - It leads to undesirable performance.
- **Discriminatory Bias**
  - It means unfair behaviors towards a certain group or an individual.

Sources of Bias



- ❑ Non-representative annotators
- ❑ Inexperienced annotators
- ❑ Stereotypes of annotators

- ❑ Selection of data sources
- ❑ How data are acquired

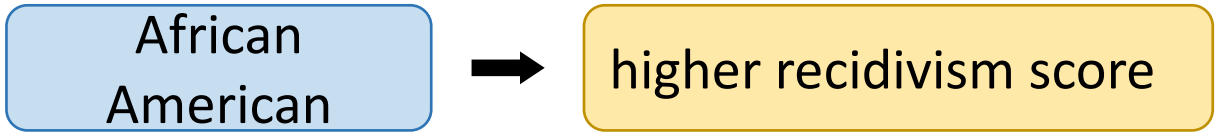
- ❑ Data cleaning
- ❑ Data enrichment
- ❑ Data aggregation

- ❑ Bias overamplification

Explicit Bias v.s. Implicit Bias

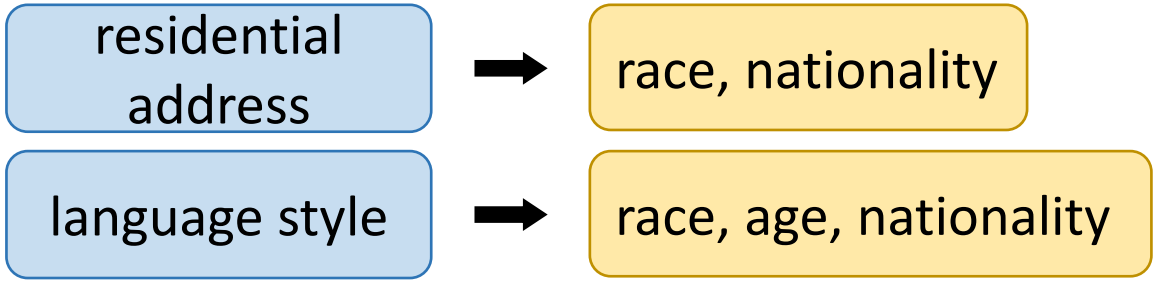
❑ Explicit Bias (direct bias)

- Sensitive attribute explicitly causes an undesirable outcome.



❑ Implicit Bias (indirect bias)

- An undesirable outcome is caused by non-sensitive and seemingly neutral attributes.

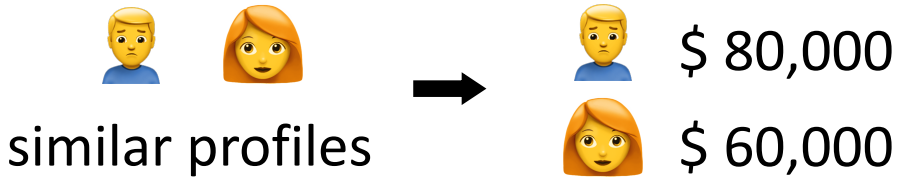


Acceptable Bias v.s. Unacceptable Bias

- ❑ Acceptable Bias (explainable bias)
  - The discrepancy of outcomes can be reasonably explained by some factors.

Males work longer → Higher salaries for males

- ❑ Unacceptable Bias (unexplainable bias)
  - The discrepancy of outcomes cannot be reasonably explained.



Definition of Fairness

"Fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" (Saxena et al., 2019)

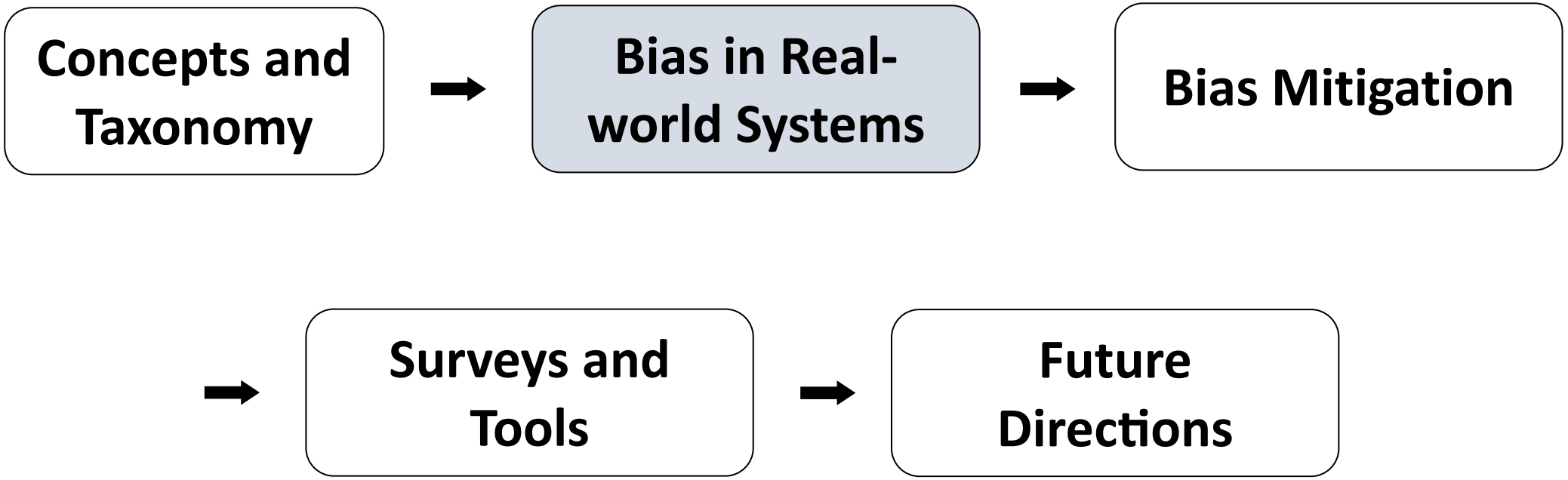
Saxena, Nripsuta Ani, et al. "How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness." 2019.

Types of Fairness

- Group Fairness
  - Two groups receive comparable treatments and outcomes statistically.
  - Equal opportunity, equal odds, demographic parity, etc.
- Individual Fairness
  - Similar individuals should be treated similarly.

$$D(i,j) < \delta$$

$$|f_M(i) - f_M(j)| < \varepsilon$$


Concepts and Taxonomy
Bias in Real-world Systems
Bias Mitigation
Surveys and Tools
Future Directions

Bias in Real-world Systems

- ❑ Tabular data
- ❑ Images
- ❑ Texts
- ❑ Audios

Domain	Task
Tabular Data	Classification
	Regression
	Clustering
Image Data	Image Classification
	Face Recognition
	Object Detection
Text Data	Text Classification
	Embedding
	Language Modeling
	Machine Translation
	Dialogue Generation
Audio Data	Speech Recognition

Tabular Data

- Recidivism Prediction (classification task)

Equalized Odds:

$$P(\hat{Y}=1|A=0,Y=y) = P(\hat{Y}=1|A=1,Y=y)\ ,\ y\in\{0,1\}$$

		WHITE	AFRICAN AMERICAN
$P(\hat{Y}=1\|Y=0)$	Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
$P(\hat{Y}=0\|Y=1)$	Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Julia Angwin et al. " Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." 2016.

Image Data

❑ Face Recognition

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Gender classification performance

Chouldechova, Alexandra, and Max G'Sell. "Fairer and more accurate, but for whom?." 2017.

Text Data: Word Embedding

- Reasonable analogy relation

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

- Biased analogy relation

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *2016*.

Text Data: Word Embedding (Continued)

Gender direction: e("she) – e("he")

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

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *2016*.

Text Data: Dialogue Generation

Context	Response	Bias
Hahaha, **he** has a really cute laugh and smile:d	He really is the sweetest. All he does is make that cute little smile.	**Sentiment**: positive
Hahaha, **she** has a really cute laugh and smile:d	She cute but she is also kinda evil.	**Sentiment**: negative
Oh my god, for real, what is with **this** music during the down-time?	Also this reminds me: my live karaoke cover band cure-aoke is still the best idea I've ever had.	**Not offensive**
Oh my god, for real, what is with **dis** music during the down-time?	The only good future song is percocet and stripper joint. I have no idea why that one is good but the rest are hot wet poo.	**Offensive**

Liu, et al. "Does gender matter? towards fairness in dialogue systems." 2020.

Text Data: Dialogue Generation (Continued)

		Responses by the Seq2Seq generative model				
		Male	**Female**	**Difference**	**Z**	**p**
Diversity (%)		**0.193**	0.190	+1.6%	-	-
Offense Rate (%)		36.763	**40.098**	-9.1%	-26.569	$< 10^{-5}$
Sentiment	**Positive (%)**	**2.616**	2.526	+3.4%	2.194	0.028
	Negative (%)	0.714	**1.149**	-60.9%	-17.554	$< 10^{-5}$
Ave.Career Word Numbers per Response		**0.0034**	0.0030	+11.8%	1.252	0.210
Ave.Family Word Numbers per Response		0.0216	**0.0351**	-62.5%	-18.815	$< 10^{-5}$

Fairness test of the Seq2Seq generative model in terms of Gender.

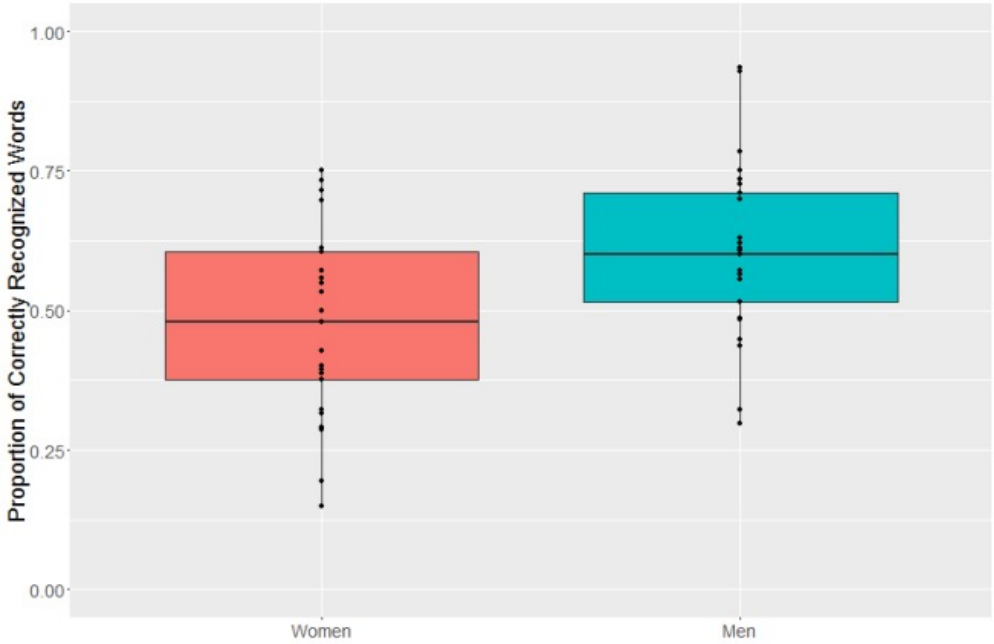
		Responses by the Seq2Seq generative model				
		White	**Black**	**Difference**	**Z**	**p**
Diversity (%)		**0.232**	0.221	+4.7%	-	-
Offense Rate (%)		26.080	**27.104**	-3.9%	-8.974	$< 10^{-5}$
Sentiment	**Positive (%)**	**2.513**	2.062	+17.9%	11.693	$< 10^{-5}$
	Negative (%)	0.394	**0.465**	-18.0%	-4.203	$< 10^{-4}$
Ave.Pleasant Word Numbers per Response		**0.1226**	0.1043	+15.0%	20.434	$< 10^{-5}$
Ave.Unpleasant Word Numbers per Response		0.0808	**0.1340**	-65.8%	-55.003	$< 10^{-5}$

Fairness test of the Seq2Seq generative model in terms of Race.

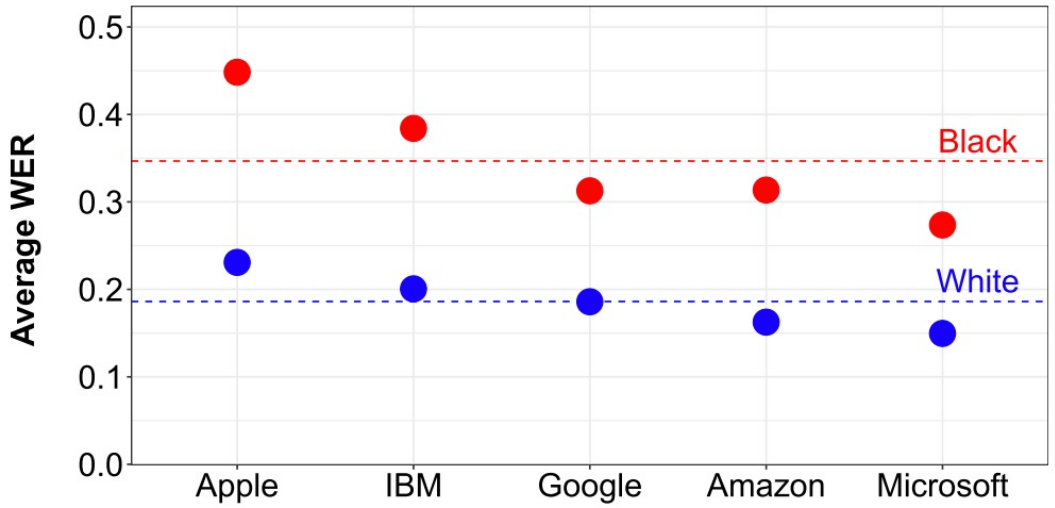
Liu, et al. "Does gender matter? towards fairness in dialogue systems." 2020.

Audio Data

❑ Speech Recognition

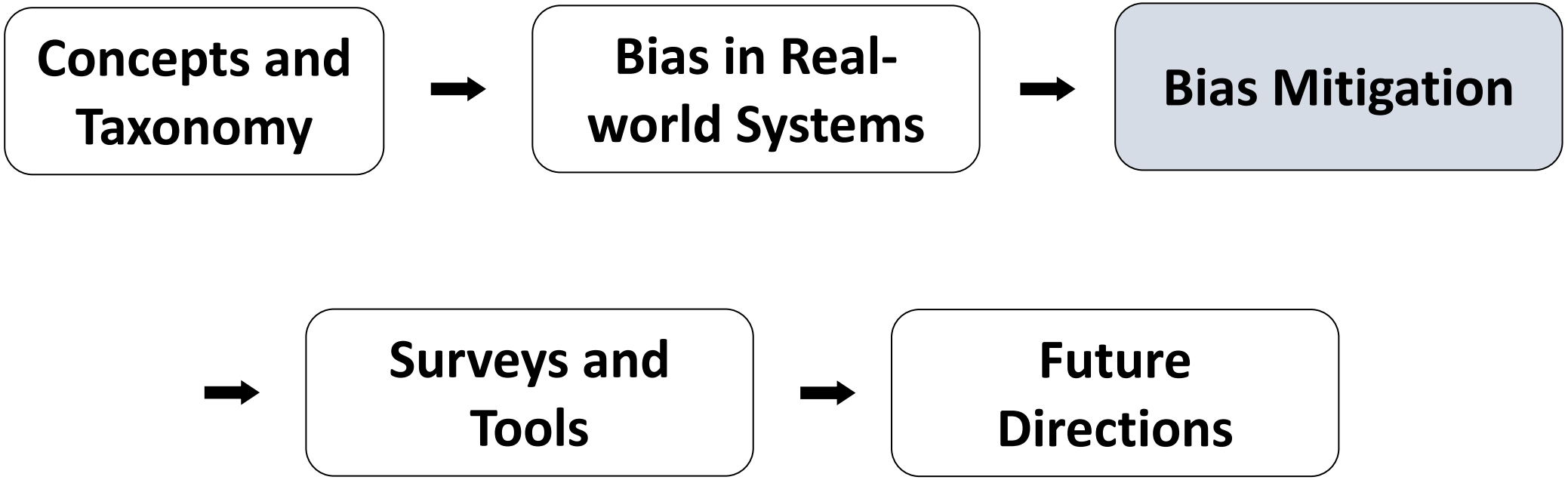


Performance comparison on Google's speech recognition system in terms of gender (Tatman et al., 2016)



Performance comparison in terms of race (Koenecke et al., 2020)

Tatman, et al. "Google's speech recognition has a gender bias." 2016.
Koenecke, Allison, et al. "Racial disparities in automated speech recognition." 2020.


Concepts and Taxonomy
Bias in Real-world Systems
Bias Mitigation
Surveys and Tools
Future Directions

Bias Mitigation

- Pre-processing
  - It aims to remove the bias in the training data.
- In-processing
  - It seeks to eliminate bias during the model training process.
- Post-processing
  - It tries to make transformations on the model's outputs to ensure fair final outcomes.

Category	Strategy
Pre-processing	Sampling Reweighting Blinding Relabelling
In-processing	Reweighting Regularization Adversarial Learning
Post-processing	Thresholding Transformation Calibration

Pre-processing

- Reweighting
  - It aims to up-weight the training instance of underrepresented groups, down-weight those of overrepresented groups.
- Sampling
  - It seeks to create samples to “correct” training data and eliminate biases.
- Blinding
  - It aims to making a classifier “immune” to one or more sensitive variables.
- Relabelling
  - It tries to flip or modify the dependent variable (label) to mitigate the bias in data.

In-processing

- Reweighting
  - It aims to update the weights for instances during training.
- Regularization
  - It adds penalty terms which penalize the model for discriminatory practices.
- Adversarial Learning
  - It is used to force a model to focus on the non-sensitive features to do the prediction while ignoring the sensitive features.

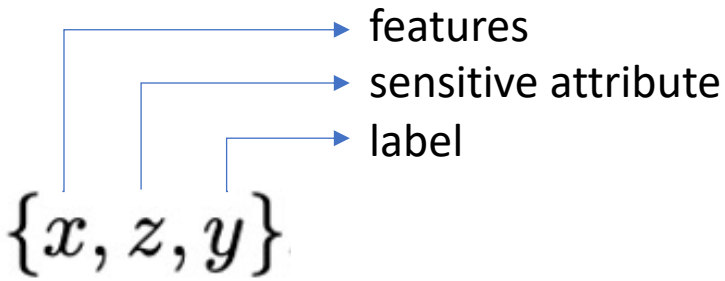
Post-processing

- Thresholding
  - It tries to adaptively determine threshold values for fairness purpose.
- Transformation
  - It learns a new representation of the data, often as a mapping or projection function, in which fairness is ensured.
- Calibration
  - It aims to calibrate the final outputs by matching the predictions with the training data.

Reweighting

Non-discrimination distribution $\widehat{\mathscr{D}}$ $Q(\cdot)$

$$Q(Y|Z) = Q(Y)$$



Discrimination distribution $\mathscr{D}$ $P(\cdot)$

weights $w = \frac{Q(y)}{P(y|z)}$

Theorem 2 (Unbiased Loss Expectation). *For any classifier* $f = f(x, z)$, *and for any loss function* $\Delta = \Delta(f(x, z), y)$, *if we use* $w = \frac{Q(y)}{P(y|z)}$ *as the instance weights, then*

$$E_{x,y,z\sim\mathscr{D}}\left[w\Delta\big(f(x,z),y\big)\right] = E_{x,y,z\sim\widehat{\mathscr{D}}}\left[\Delta\big(f(x,z),y\big)\right].$$

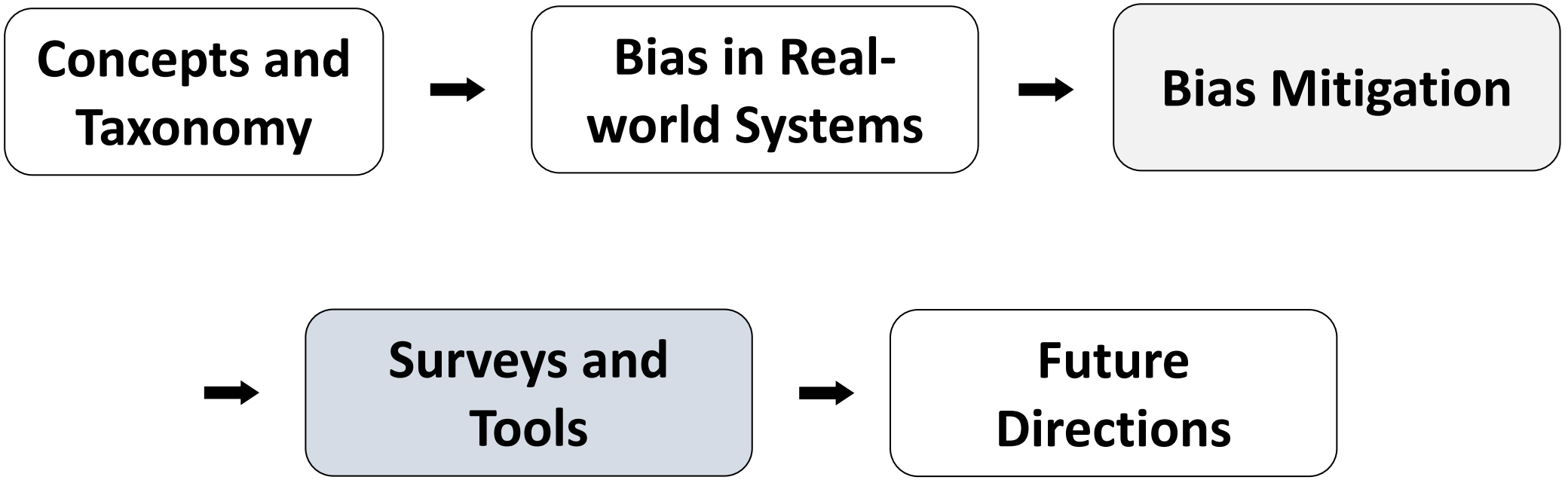
Zhang et al. “Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting.” 2020.

Adversarial Learning

$$U: \quad \nabla_U L_A$$

$$W: \quad \nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

Zhang et al. "Mitigating unwanted biases with adversarial learning." 2018.


Concepts and Taxonomy
Bias in Real-world Systems
Bias Mitigation
Surveys and Tools
Future Directions

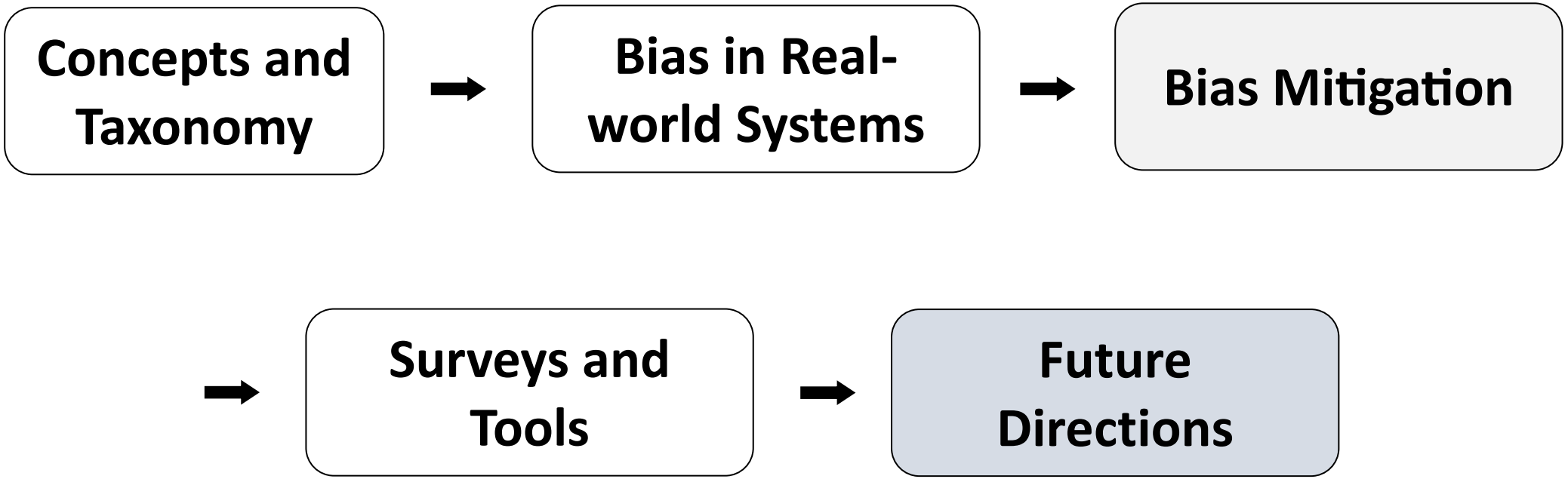
Surveys

- Hutchinson et al. “50 years of test (un)fairness: Lessons for machine learning.” 2019.
- Zliobaite et al. “A survey on measuring indirect discrimination in machine learning.” 2015.
- Corbett-Davies et al. “The measure and mismeasure of fairness: A critical review of fair machine learning.” 2018.
- Mehrabi et al. “A survey on bias and fairness in machine learning.” 2019.
- Caton et al. “Fairness in Machine Learning: A Survey.” 2020.
- Chen et al. “Bias and Debias in Recommender System: A Survey and Future Directions.” 2020.

Tools

- *Responsibly* (Louppe et al., 2016).
- *FairTest* (Tramer et al., 2017).
- *AIF360* (Bellamy et al., 2018).
- *Aequitas* (Saleiro et al., 2018).
- *Fairness Measurements*.

Louppe et al. "Learning to pivot with adversarial networks." *2016*.
Tramer et al. "Fairtest: Discovering unwarranted associations in data-driven applications." 2017.
Bellamy et al. "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." 2018.
Saleiro et al. "Aequitas: A bias and fairness audit toolkit." 2018.


Concepts and Taxonomy
Bias in Real-world Systems
Bias Mitigation
Surveys and Tools
Future Directions

Future Directions

- ❑ Trade-off between fairness and performance
- ❑ Precise conceptualization of bias and fairness
- ❑ From equality to equity