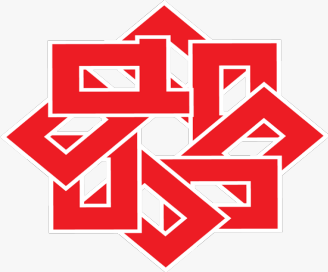
Oracle Labs



EMNLP 2022

Don't Just Clean It, Proxy Clean It:
Mitigating Bias by Proxy in Pre-Trained Models

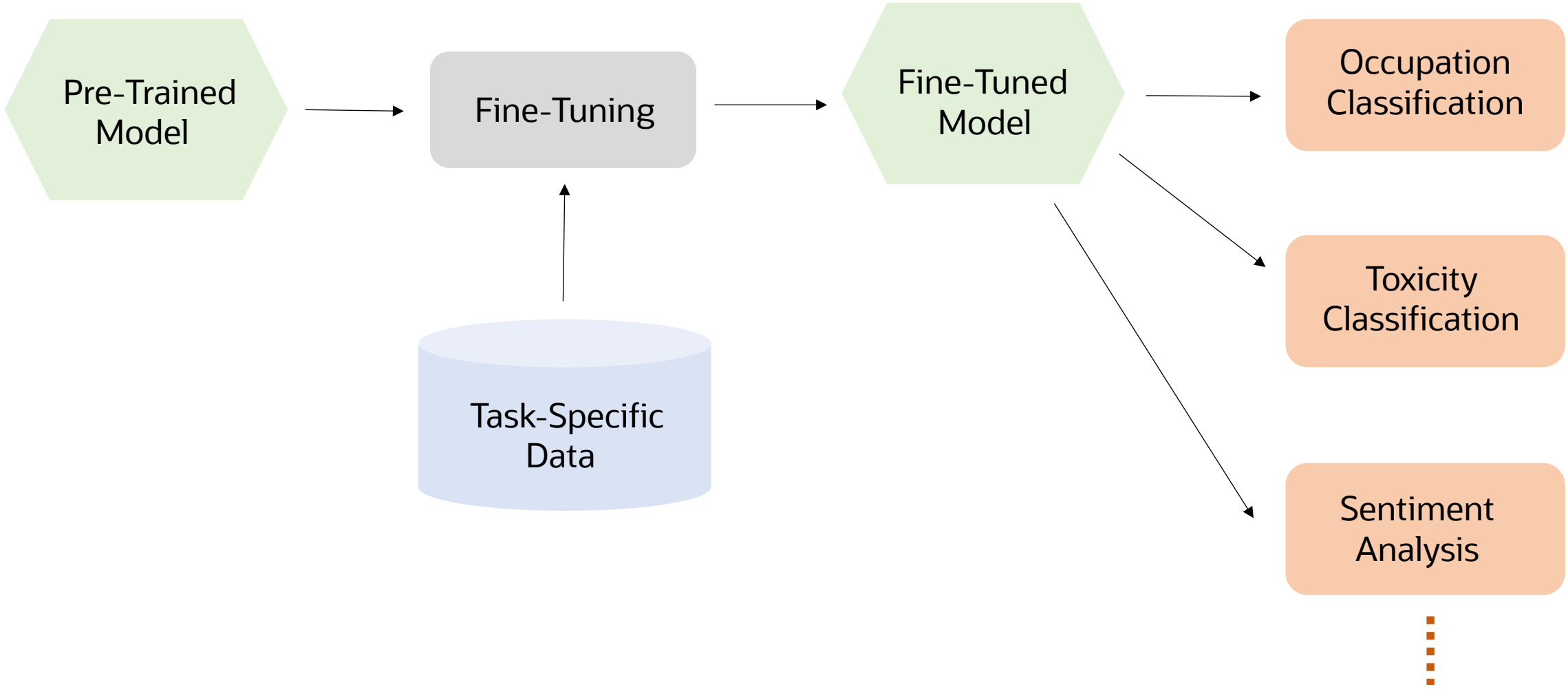
Swetasudha Panda, Ari Kobren, Michael Wick and Qinlan Shen

Oracle Labs
Burlington, MA

Outline

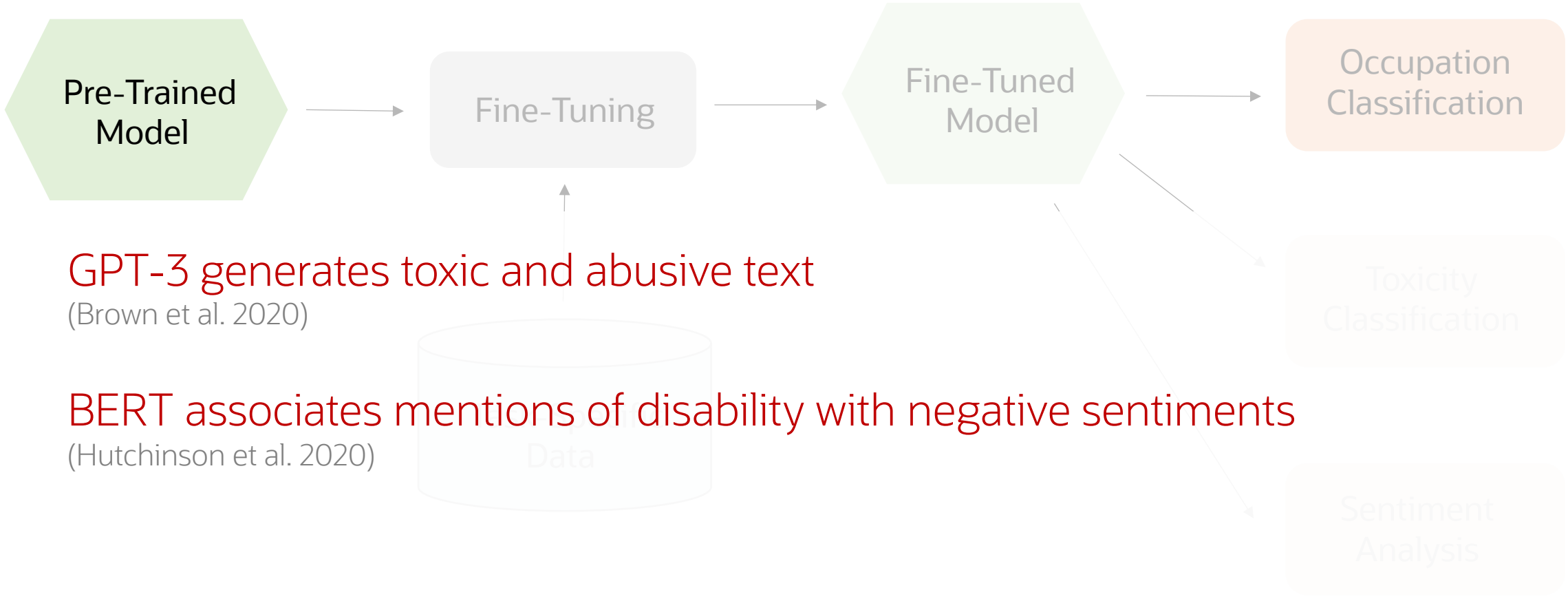
- Introduction
- Hypothesis
- Debiasing Approach
- Experiments

Pre-Trained Language Models



Pre-Trained Models Encode Social Biases*

Upstream Representations

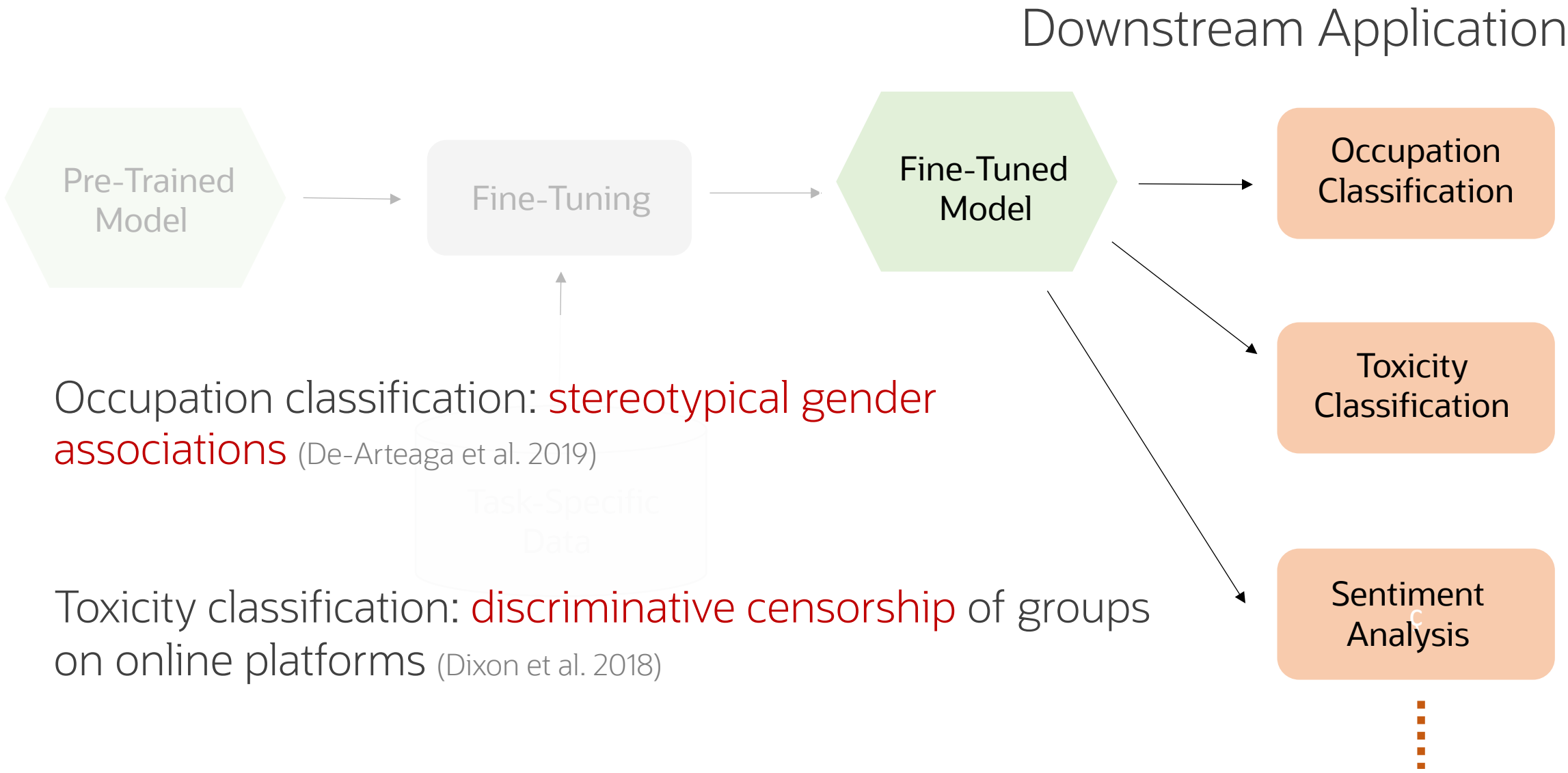


GPT-3 generates toxic and abusive text
(Brown et al. 2020)

BERT associates mentions of disability with negative sentiments
(Hutchinson et al. 2020)

***Social biases: discrepancies in model behavior towards certain demographic identities**

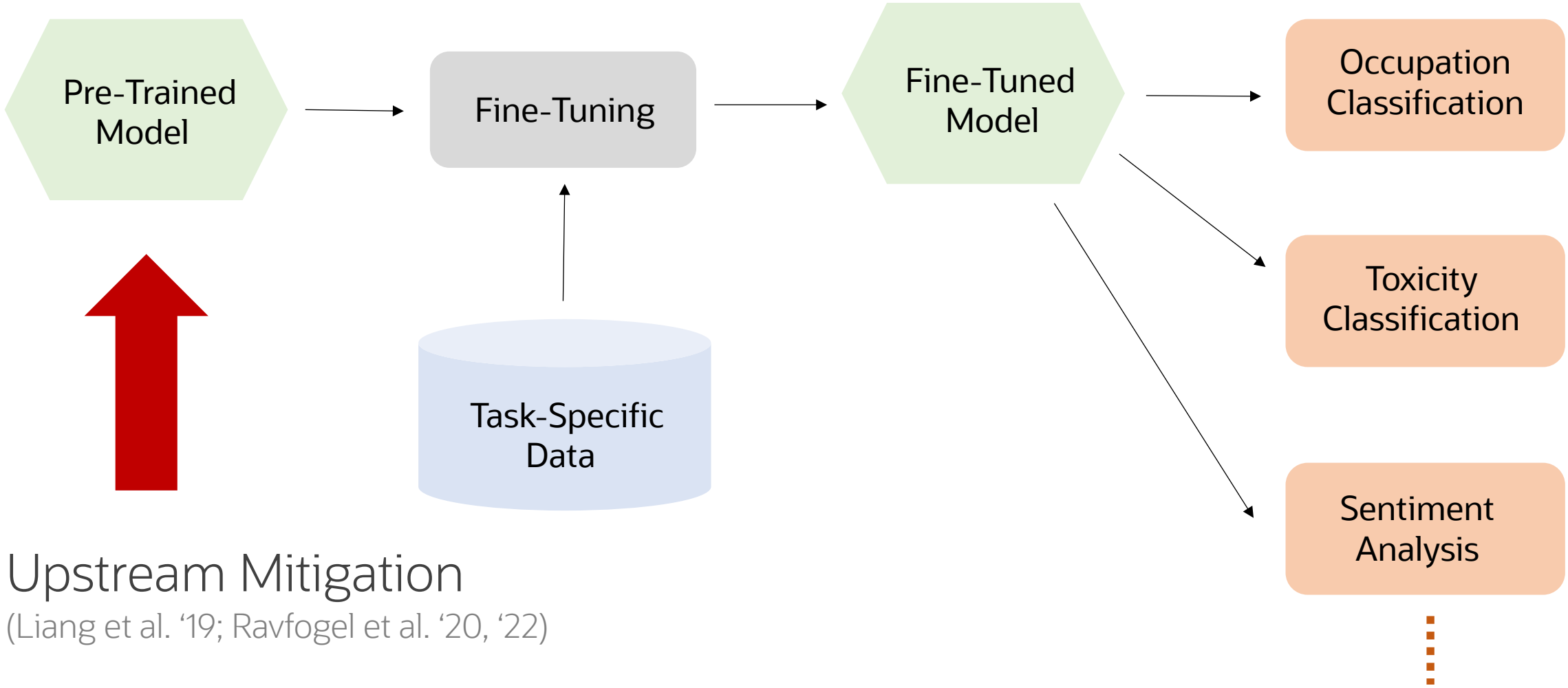
Fine-Tuned Models Cause Allocational Harms



Occupation classification: stereotypical gender associations (De-Arteaga et al. 2019)

Toxicity classification: discriminative censorship of groups on online platforms (Dixon et al. 2018)

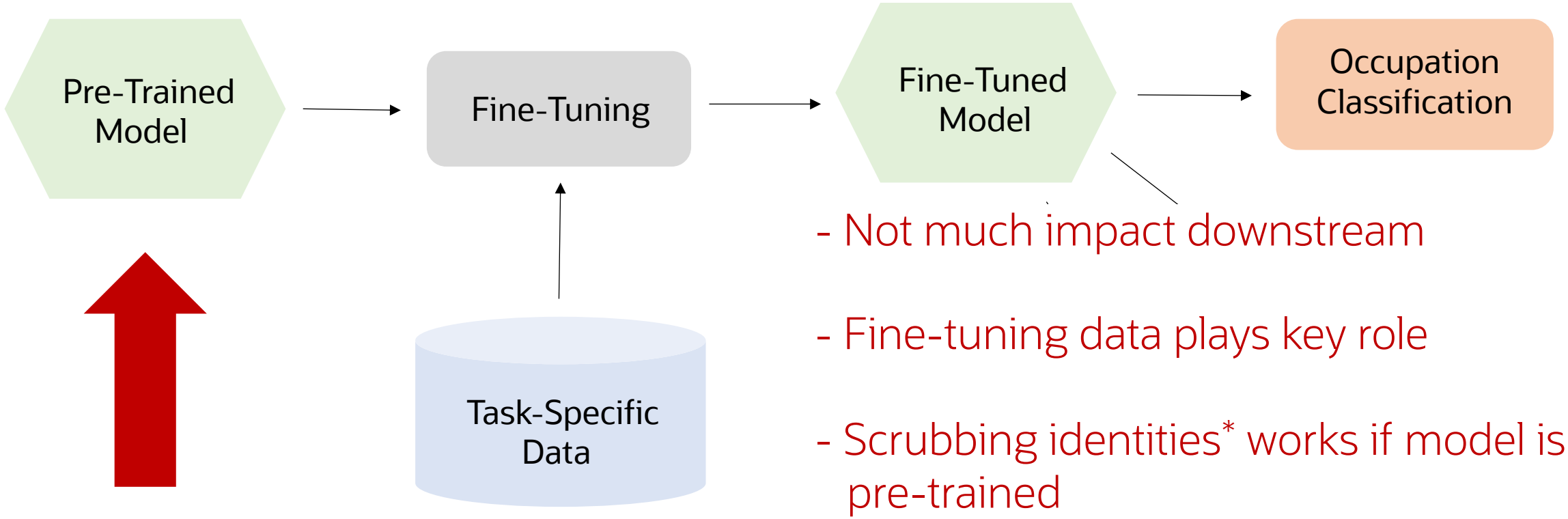
Debiasing Interventions



Upstream Mitigation

(Liang et al. '19; Ravfogel et al. '20, '22)

Bias Transfer Hypothesis (Steed et al. '22)



- Not much impact downstream
- Fine-tuning data plays key role
- Scrubbing identities* works if model is not pre-trained

Upstream Mitigation

(Liang et al. '19; Ravfogel et al. '20, '22)

***Identities: words that indicate demographic identity**

Outline

- Introduction
- Hypothesis
- Debiasing Approach
- Experiments

Our Hypothesis

Pre-trained models are sensitive to *bias by proxy*

Proxies* act as additional sources of bias, in absence of identity words

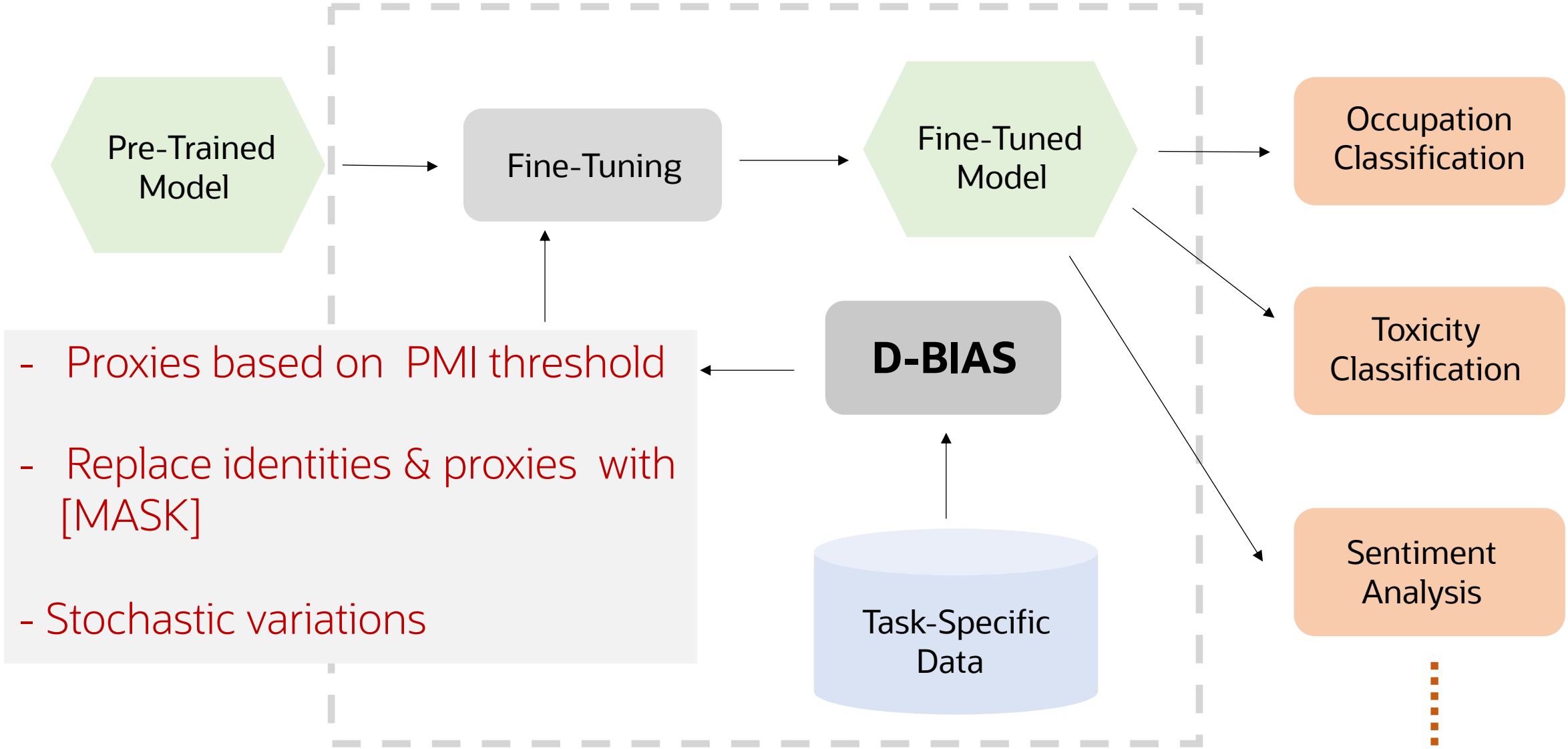
Eliminating proxies can help reduce downstream bias

***Proxies: words which frequently co-occur with identity words**

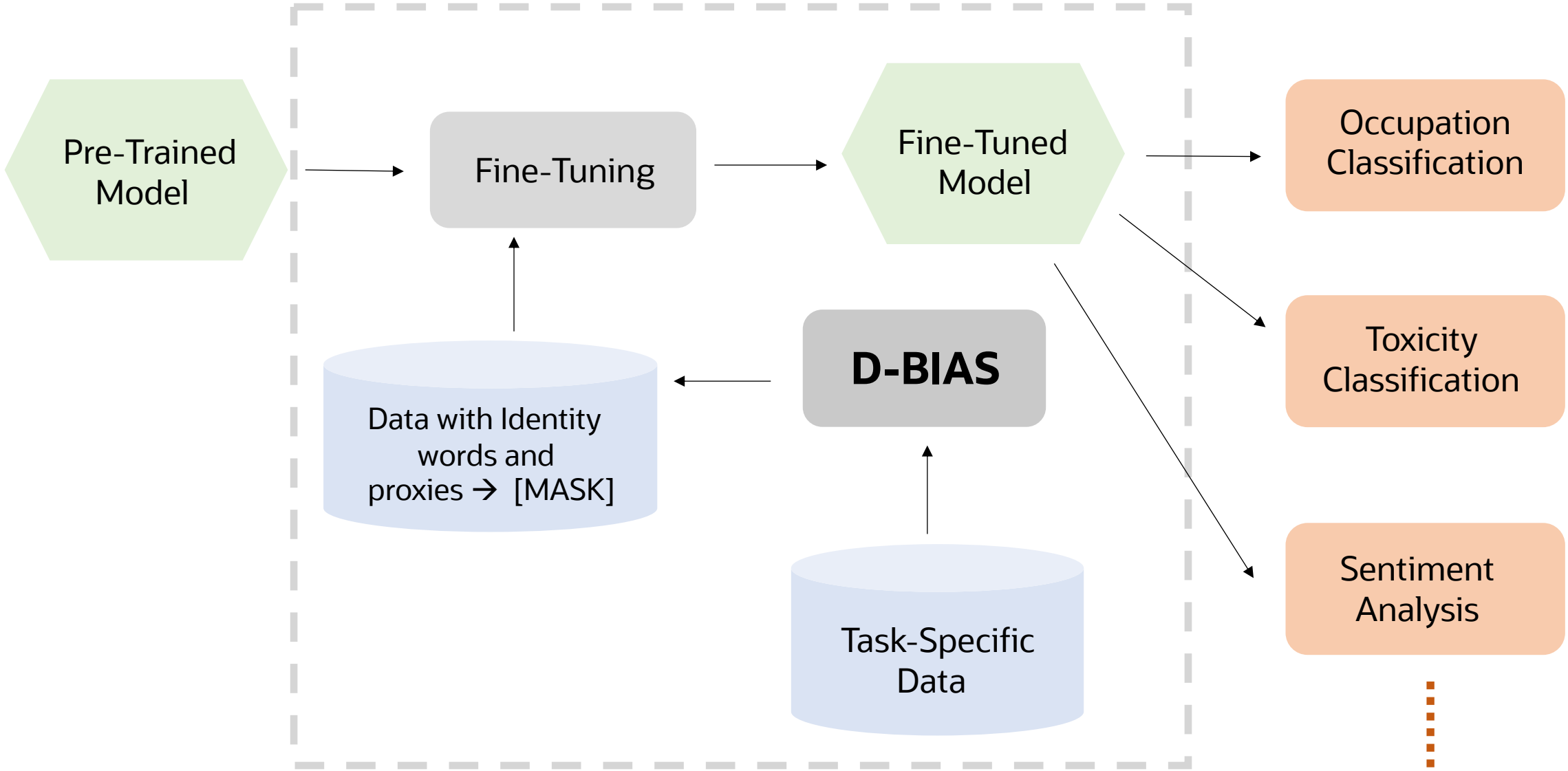
Outline

- Introduction
- Hypothesis
- Debiasing Approach
- Experiments

Dropout Bias ASsociations (D-BIAS)



Dropout Bias ASsociations (D-BIAS)



Outline

- Introduction
- Hypothesis
- Debiasing Approach
- Experiments

Occupation Classification (BIOS)

Multi-class classification (De-Arteaga et al. 2019)

Data: ~400k online biographies from Common Crawl, 28 occupations

Identities: he/him & she/her pronouns

- Concern: Model predictions reflect stereotypical gender biases; can lead to hiring discrimination
- Bias Measure: Empirical true positive rate gap $|TPR_{he/him} - TPR_{she/her}|$, lower is better

Occupation Classification

Toxicity Classification

Sentiment Analysis

Toxicity Classification (WIKI)

Binary classification: if a comment on an online forum is toxic (Dixon et al. 2018)

Data: ~130k comments from Wikipedia Talk Pages

Identities: ~50 identities based on ethnicity, gender, age, disability etc.

- Concern: Model censors harmless mentions of identities e.g., gay
- Bias Measure: Empirical false positive rate for identities, lower is better

Occupation Classification

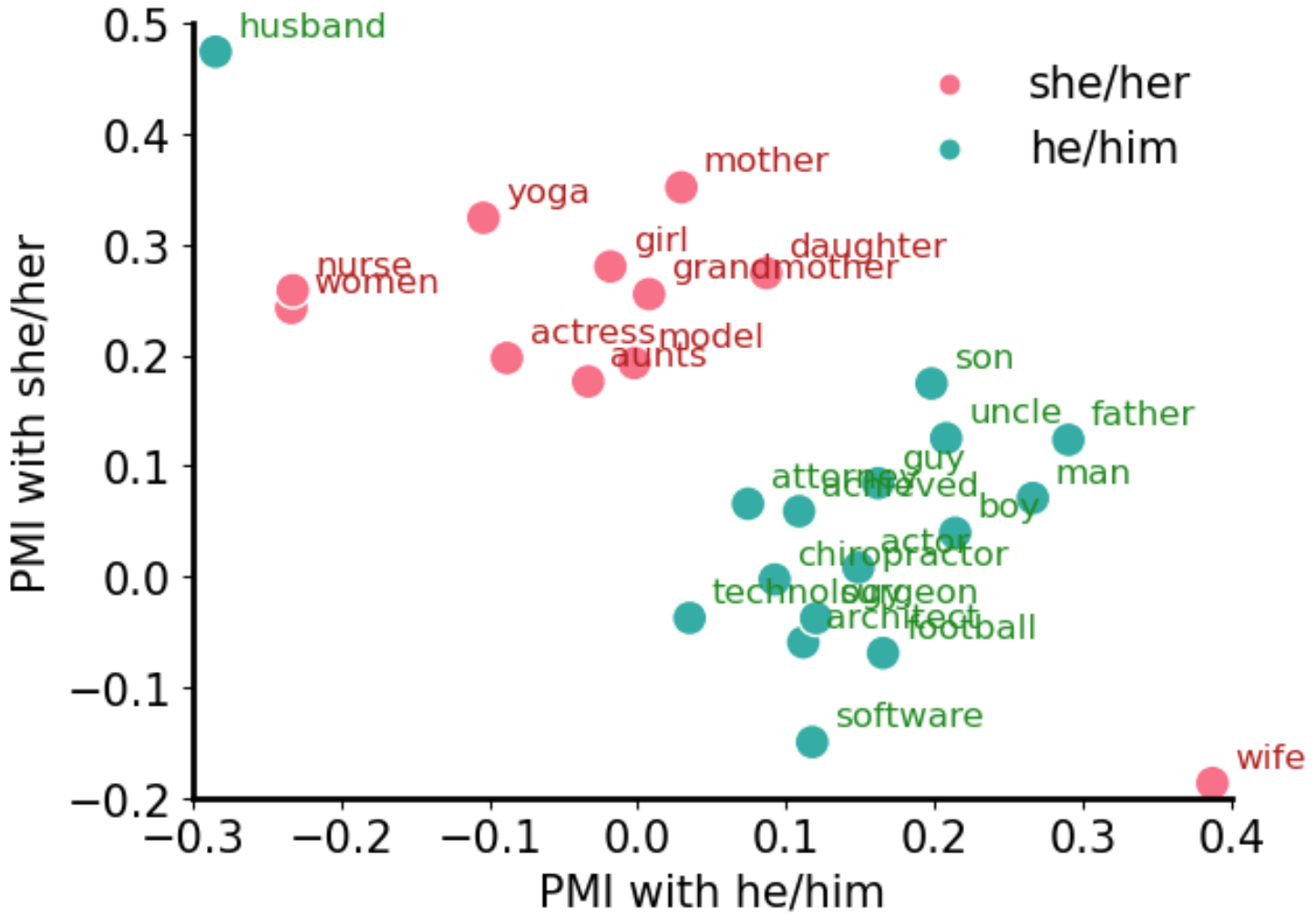
Toxicity Classification

Sentiment Analysis

Identifying Proxies with PMIs: BIOS

Examples of words frequently co-occurring with each set of pronouns

Our approach identifies prominent proxies in each case



Identifying Proxies with PMIs: WIKI

Identity	Top-Ranking nPMI words
asian	afghans, persians, israelis, aryan, culturally, afghanistan, south-east, arabs
african	African-American, races, south, Civil, Obama, Africa, black, color, people
hispanic	phillipino, phillipinos, Spaniards, Spain, waves, Latin, Europe
indian	Bihar, valmiki, maharshi, subcontinent, goverment, Modi, hindi, Indus, singh
buddhist	buddhism, Asoka, patronizer, jainism, Guptas, mimansa, deities, edicts, monks
catholic	baptist, catholicism, nobility, christians, Pope, resignation, roman
muslim	tolerent, islam, divorce-divorce-divorce, jehad, balochistan-pakistan, ummah
jewish	humus, tautological, long-bearded, missionary, judaism, hebrew, jesus
gay	f******, d***, homophobia, same-gender, sucks, die, racist, lesbian, sexuality
homosexual	cross-gendered, masculinized, sexualorientation, transsexuals, abstention
queer	gender-binary, heteronormative, unconventional, insane, b****, a**, f***

Examples of words which frequently co-occur with identity words

(a subset of the 50 identity words shown)

Identities prone to censorship co-occur frequently with abuse words

D-BIAS Results: BIOS

	Test Acc ↑	TPR_{gap}(RMSE) ↓	$\rho\ \text{TPR}_{\text{gap}}\%_{\text{F}}$ ↓
BERT	86.04 (0.10)	0.145 (0.005)	0.818 (0.005)
D-BIAS	84.40 (0.25)	**0.088 (0.006)**	0.728 (0.029)
D-BIAST	84.67 (0.24)	0.112 (0.003)	0.738 (0.024)
SENT-K	**85.93 (0.06)**	0.105 (0.004)	**0.719 (0.014)**
SENT-ST	85.90 (0.10)	0.101 (0.003)	**0.719 (0.022)**
UNIFORM	85.36 (0.19)	0.110 (0.006)	0.741 (0.017)
SCRUB	85.90 (0.04)	0.103 (0.003)	0.720 (0.021)
INLP	84.98 (0.06)	0.113 (0.009)	0.797 (0.027)
R-LACE:1	85.09 (0.07)	0.117 (0.011)	0.794 (0.025)
R-LACE:100	85.04 (0.09)	0.115 (0.014)	0.792 (0.025)

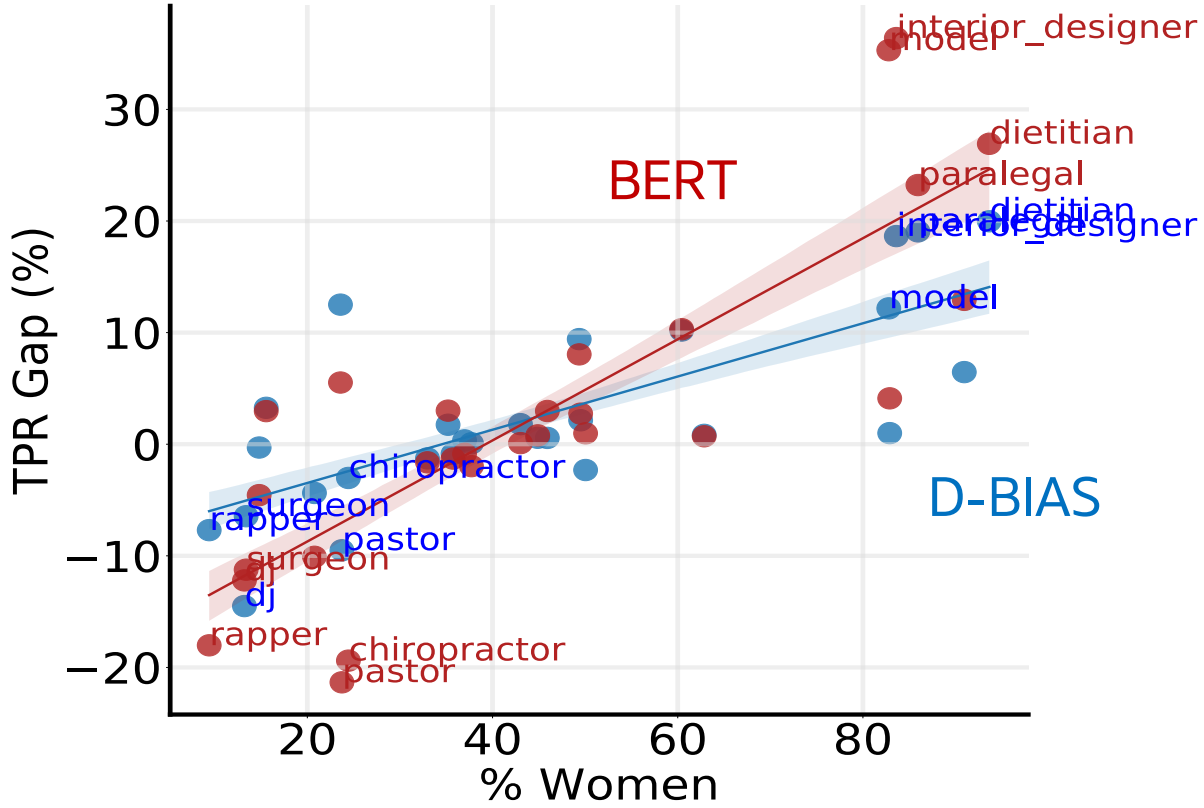
D-BIAS and variations outperform all baselines (Ravfogel et al. '20,'22)

D-BIAS achieves lowest TPR gap

D-BIAS Results: BIOS

D-BIAS outperforms BERT: closer look

- Decreases TPR gap on most problematic occupations
- Decreases correlation between TPR gap in model predictions and disparities in training data



D-BIAS Results: WIKI

	Test Acc ↑	Group FPR (across identities)		Group Acc (across identities)	
		Mean ↓	Spread ↓	Mean ↑	Spread ↓
BERT	95.88 (4.15)	7.31 (1.13)	23.89 (1.66)	91.75 (0.64)	11.17 (0.78)
D-BIAS	96.59 (3.58)	**1.81 (1.63)**	**6.52 (5.21)**	93.31 (1.49)	**4.24 (1.86)**
D-BIAST	94.20 (5.90)	5.42 (1.57)	20.24 (2.92)	88.62 (1.46)	9.69 (1.05)
SENT-K	96.51 (3.53)	3.81 (0.57)	16.35 (0.78)	93.06 (0.67)	7.71 (0.26)
SENT-ST	**96.70 (3.31)**	5.03 (1.54)	18.92 (3.18)	**93.44 (0.33)**	9.03 (1.52)
UNIFORM	96.04 (4.06)	7.72 (0.71)	23.82 (1.19)	92.17 (1.19)	11.46 (0.50)
SCRUB	95.80 (4.27)	7.69 (1.01)	24.20 (1.94)	91.59 (1.01)	11.29 (0.87)

D-BIAS and variations outperform baselines

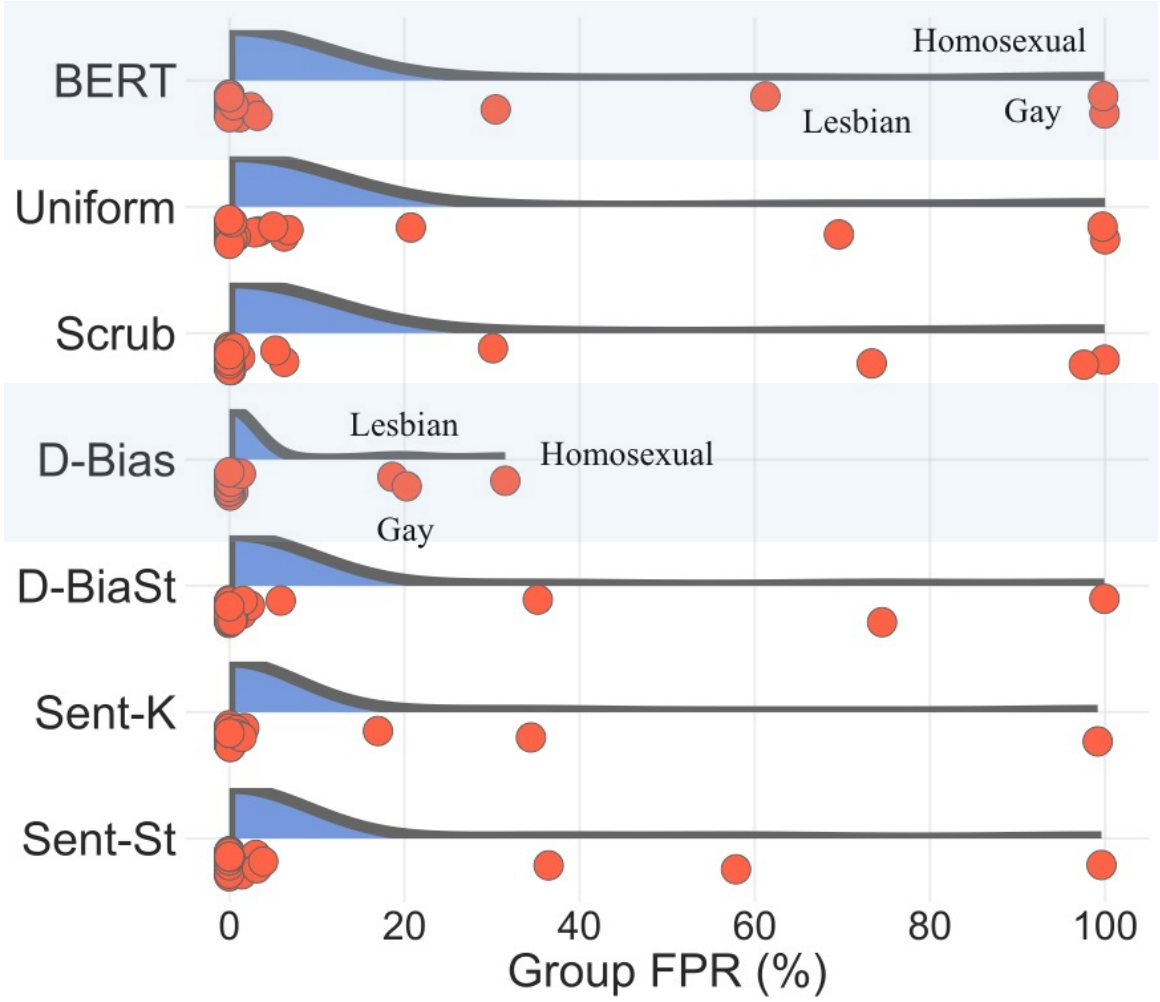
D-BIAS achieves lowest mean FPR and lowest variation in FPR / Acc across identities

D-BIAS Results: WIKI

With BERT, FPR on most censored identities is alarmingly ~ 100%

D-BIAS achieves huge reductions, to ~ 18% and 30% respectively

D-BIAS reduces FPR on other identities
& variations in FPR across identities



Conclusions

- We propose D-BIAS: a simple and easy to implement intervention which drastically decreases downstream biases without affecting accuracy
- D-BIAS outperforms SOTA baselines, scrubbing identities and uniform dropout
- D-BIAS extends to multiple identities
- Experimental results support the bias by proxy hypothesis

Thank You

Email: swetasudha.panda@oracle.com