Ethics in NLP

CSE 538

Ethics in NLP

Bias

Privacy

Ethical Research and Development

Ethics in NLP - Bias

What is Bias?

Shah, D., Schwartz, H. A., Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *In ACL-2020: Proceedings of the Association for Computational Linguistics.*

Ethics in NLP - Bias

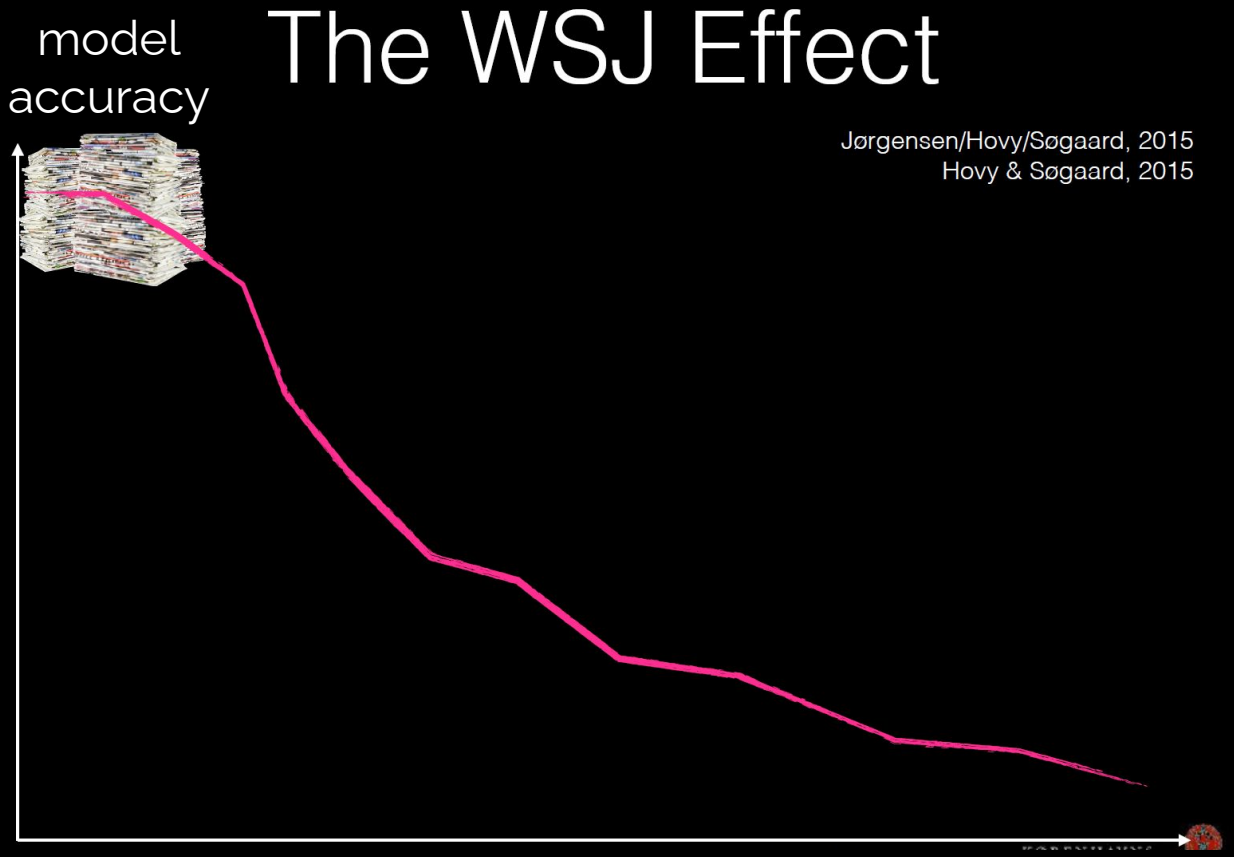
Consequences of Sociodemographic Bias in NLP Models:

- Outcome Disparity: Predicted distribution given A, are dissimilar from ideal distribution given A

- Error Disparity: Predicts less accurate for authors of given demographics.

Shah, D., Schwartz, H. A., Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *In ACL-2020: Proceedings of the Association for Computational Linguistics.*

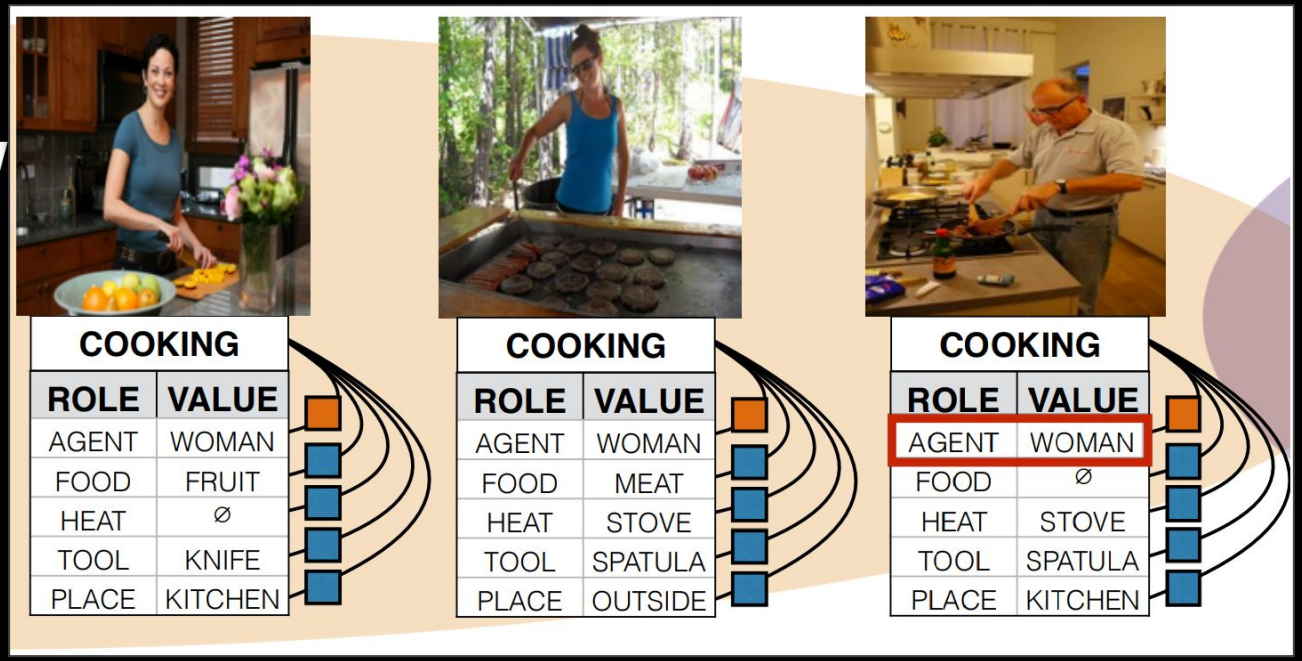
Two Examples



Two Examples

model accuracy

The W



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

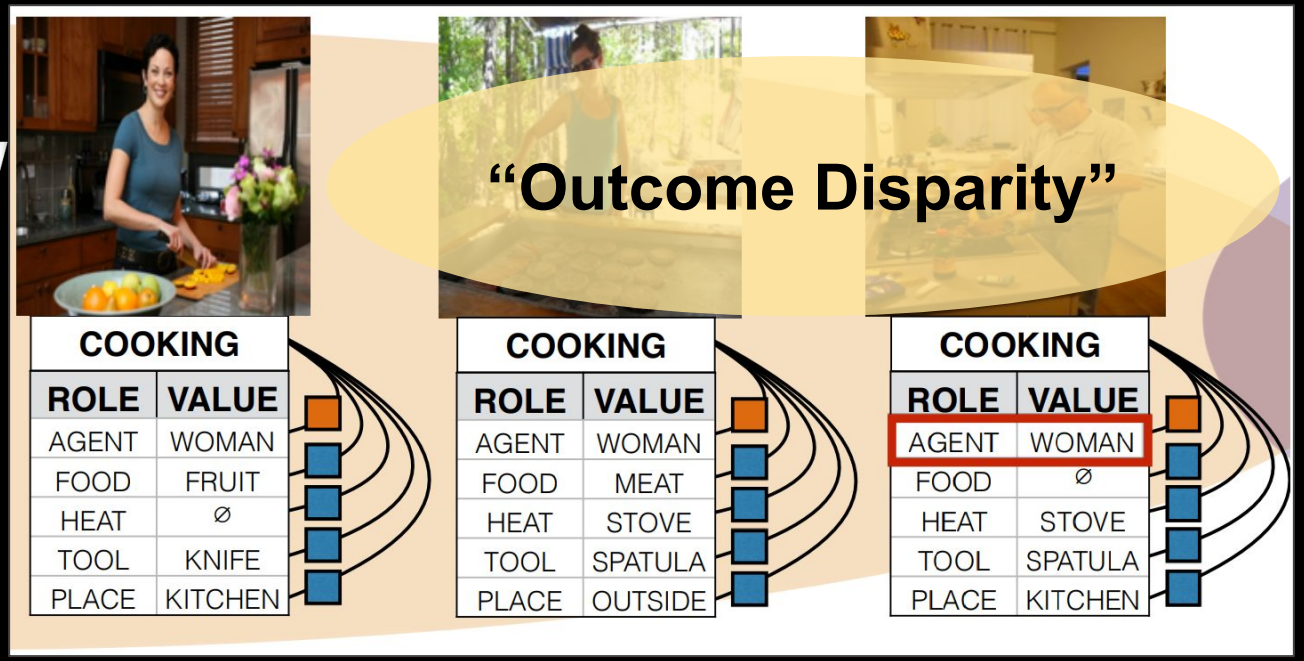
Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

distance from "standard" WSJ author demographics

Two Examples

model accuracy

The W



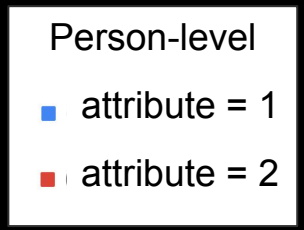
“Error Disparity”

distance from “standard” WSJ author demographics

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

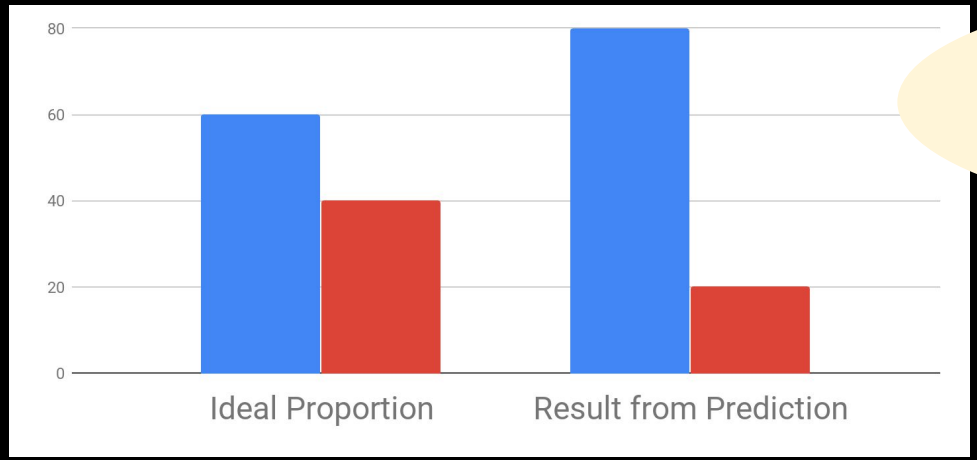
Our data and models are (human) biased.

“Outcome Disparity”



“Error Disparity”

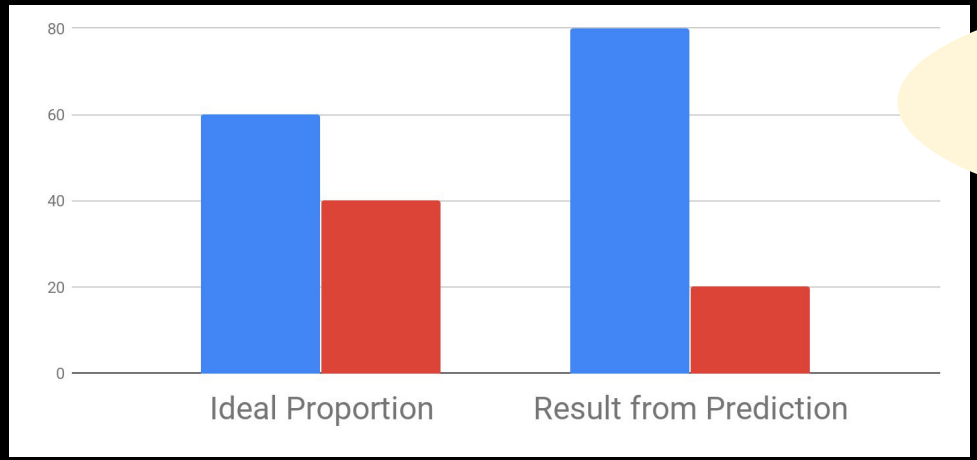
Our data and models are (human) biased.



“Outcome Disparity”

“Error Disparity”

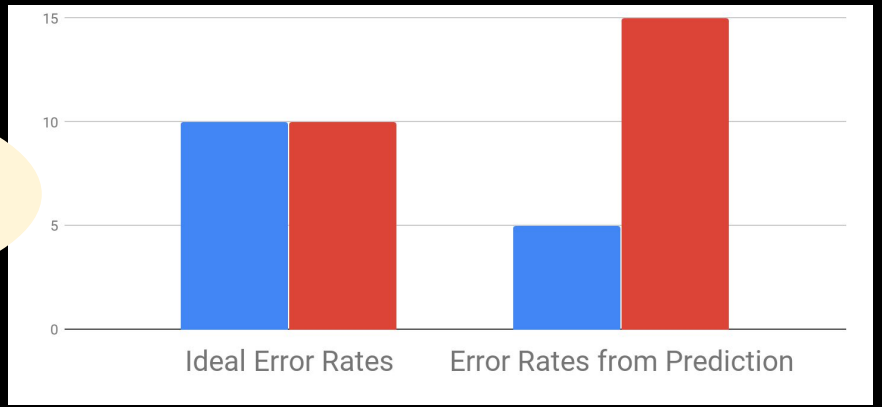
Our data and models are (human) biased.



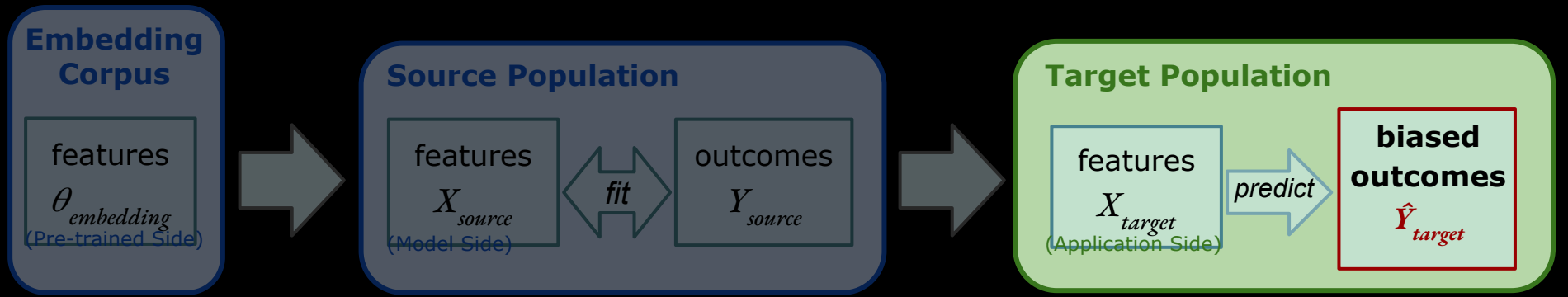
“Outcome Disparity”

Person-level
attribute = 1
attribute = 2

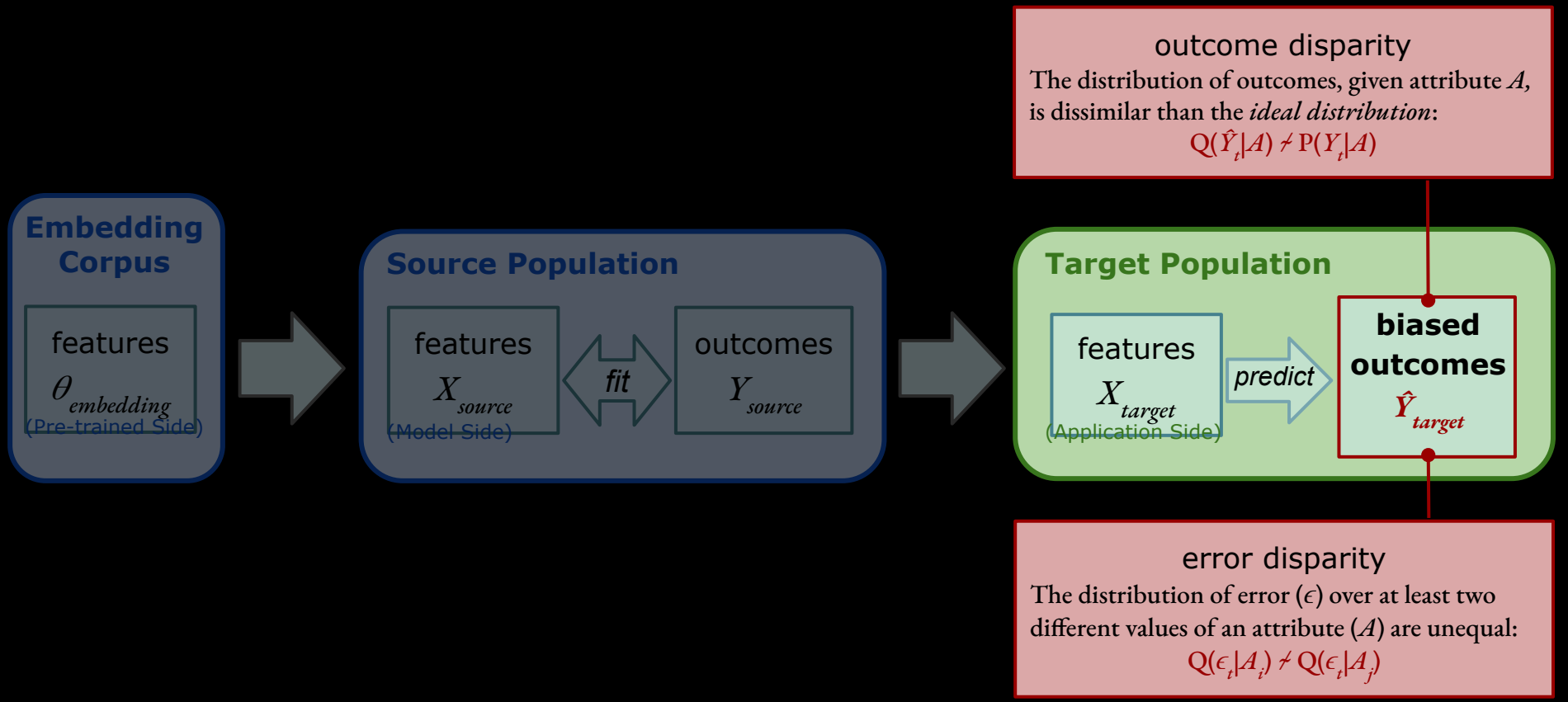
“Error Disparity”



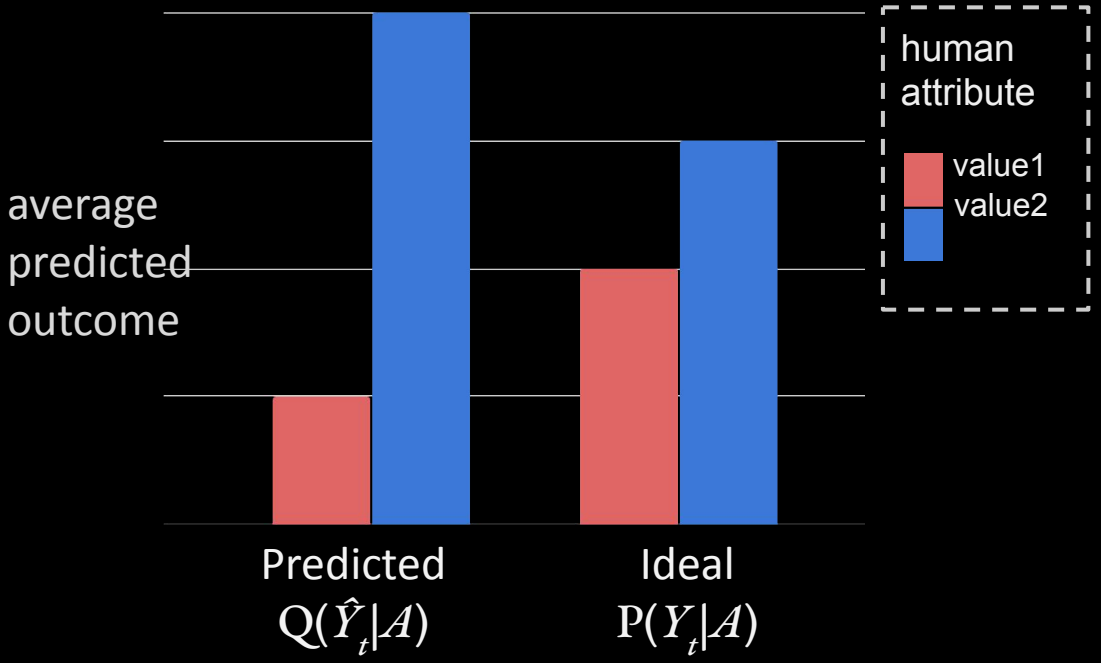
Conceptual Framework:



Conceptual Framework:



Outcome Disparity



outcome disparity

The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:

$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

Target Population

features X_{target} (Application Side)

predict

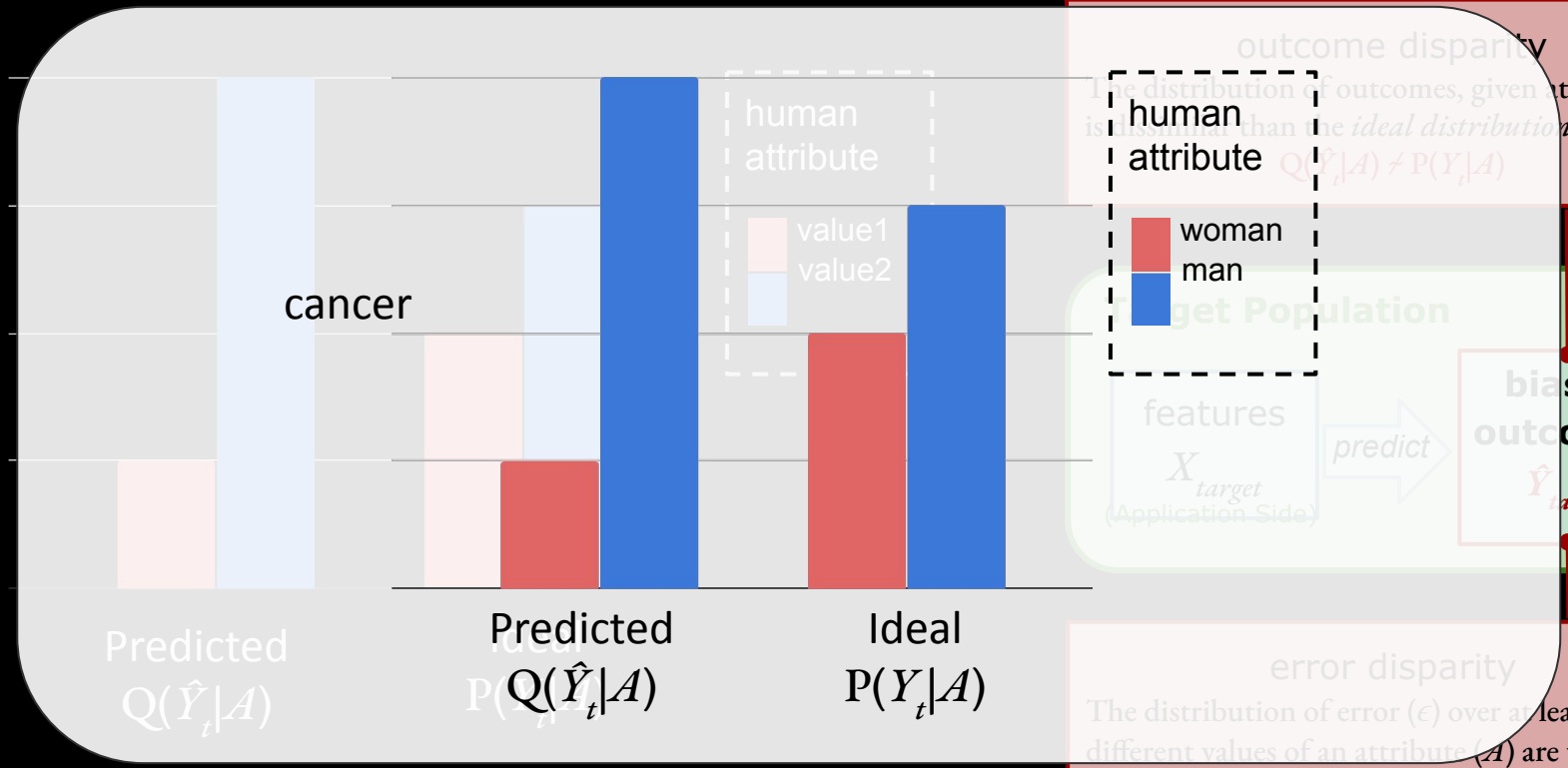
biased outcomes $\hat{Y}_{target}$

error disparity

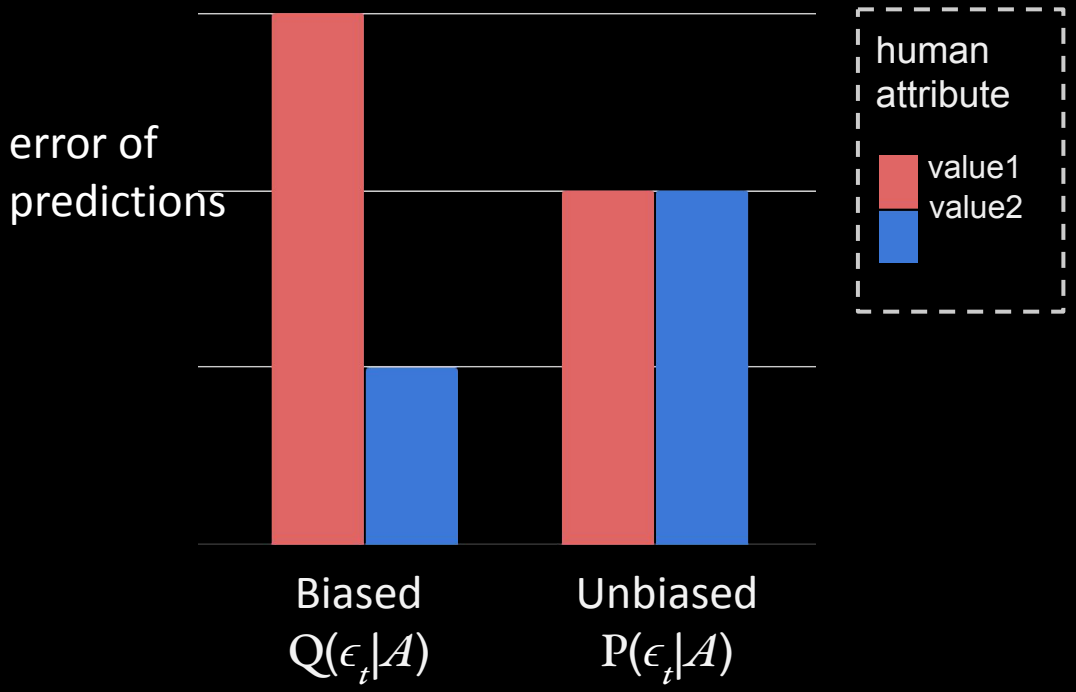
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

Outcome Disparity



Error Disparity



outcome disparity

The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:

$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

Target Population

features X_{target} (Application Side)

predict

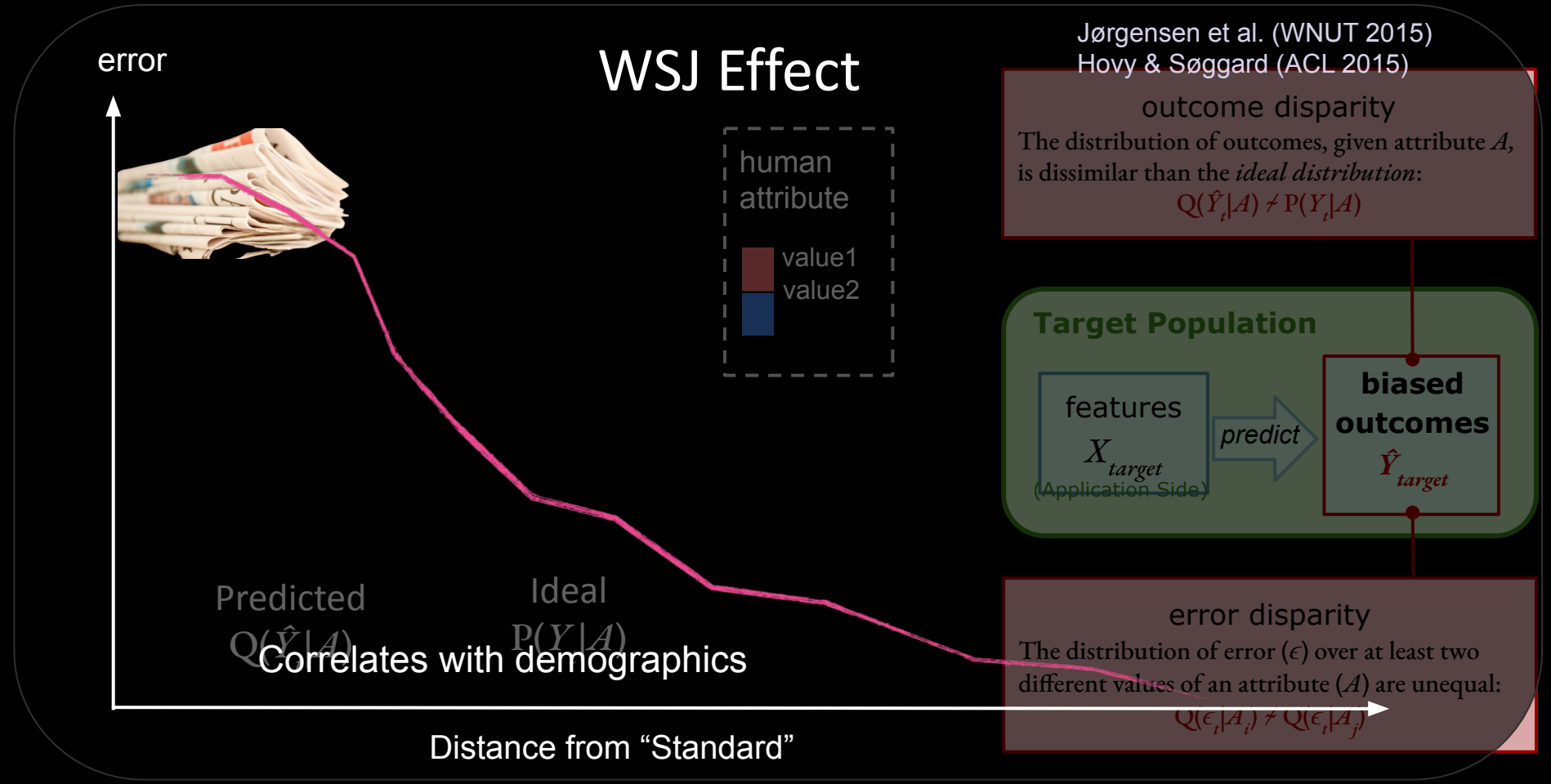
biased outcomes $\hat{Y}_{target}$

error disparity

The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

Error Disparity



Jørgensen et al. (WNUT 2015)
Hovy & Søggard (ACL 2015)

outcome disparity

The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:

$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

Target Population

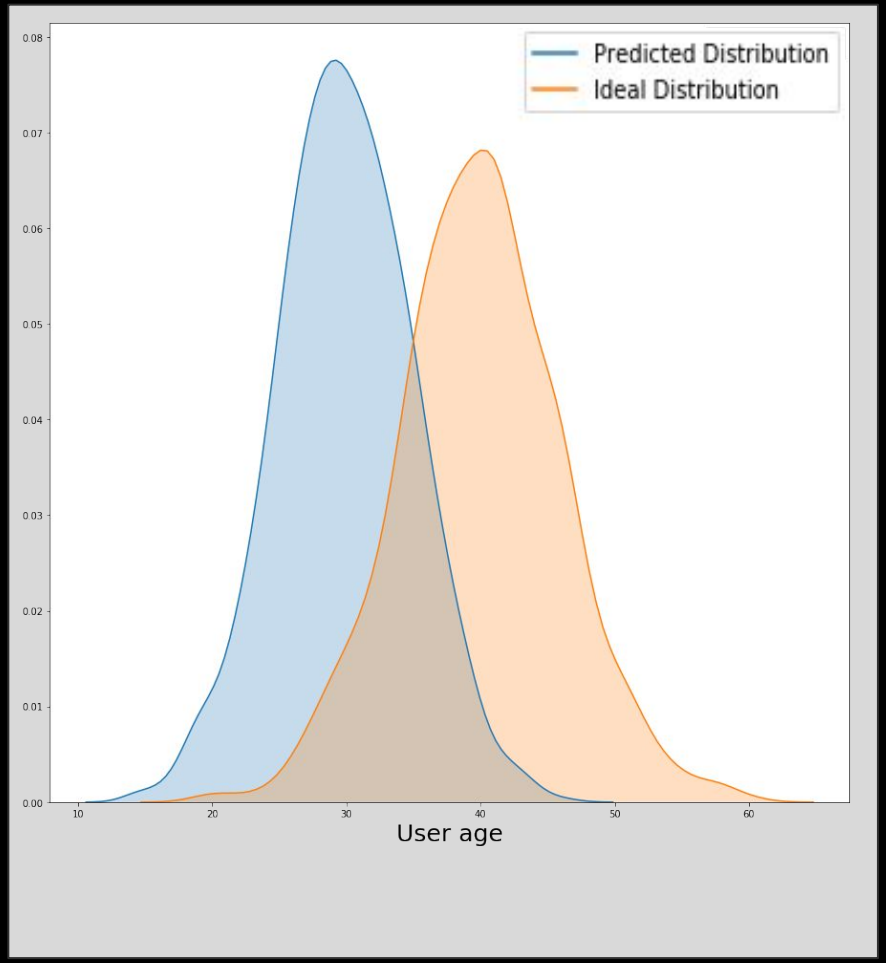
features X_{target} (Application Side) → *predict* → **biased outcomes** $\hat{Y}_{target}$

error disparity

The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

Disparities



outcome disparity

The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:

$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

Target Population

features X_{target} (Application Side)

predict

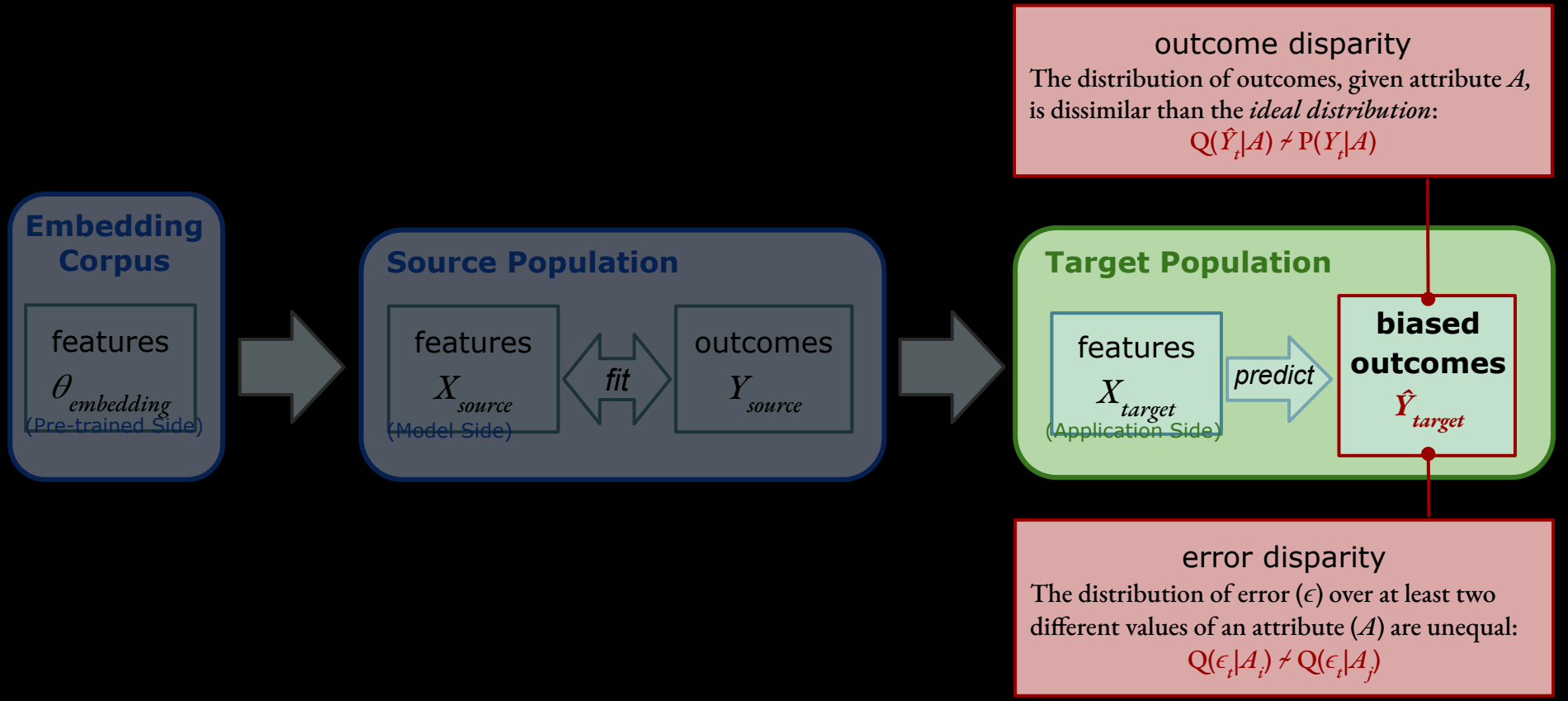
biased outcomes $\hat{Y}_{target}$

error disparity

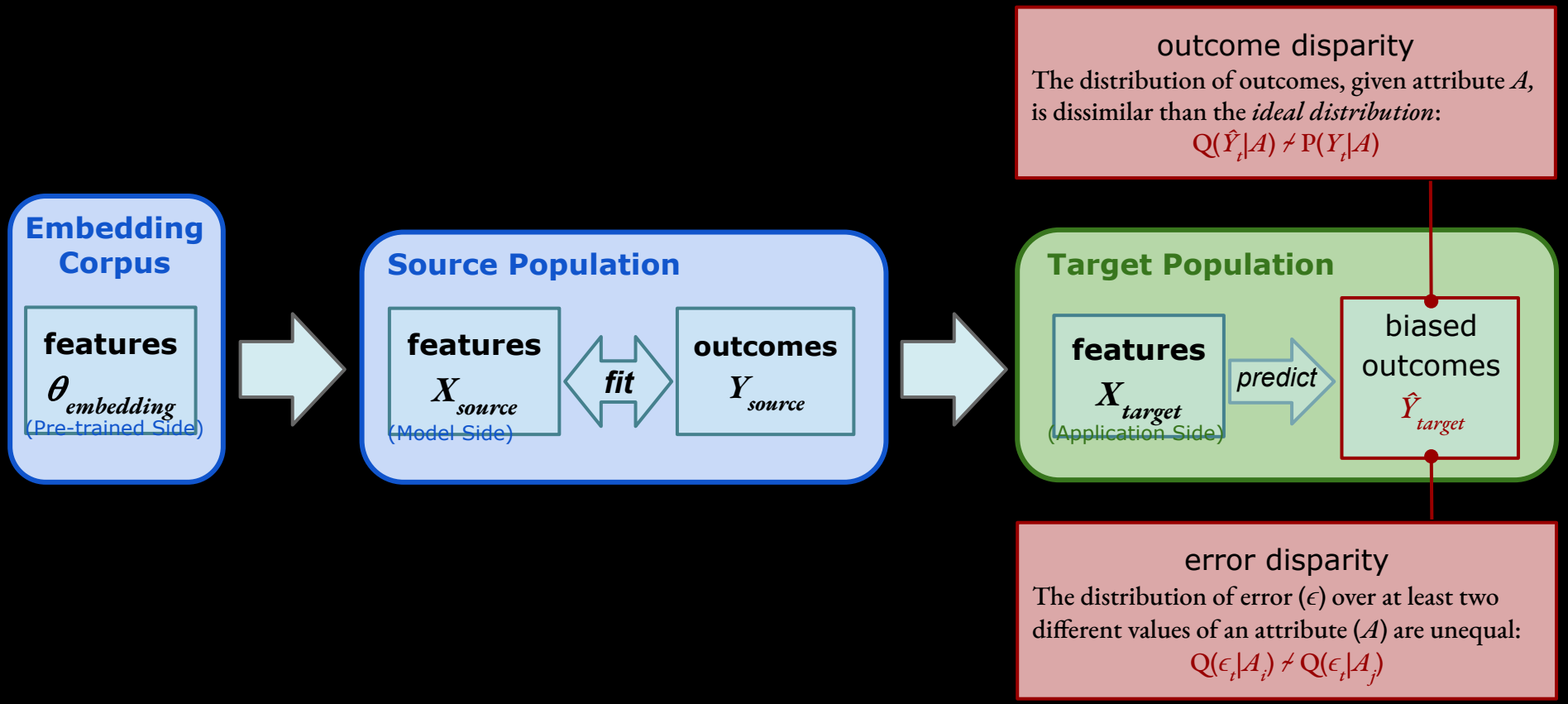
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

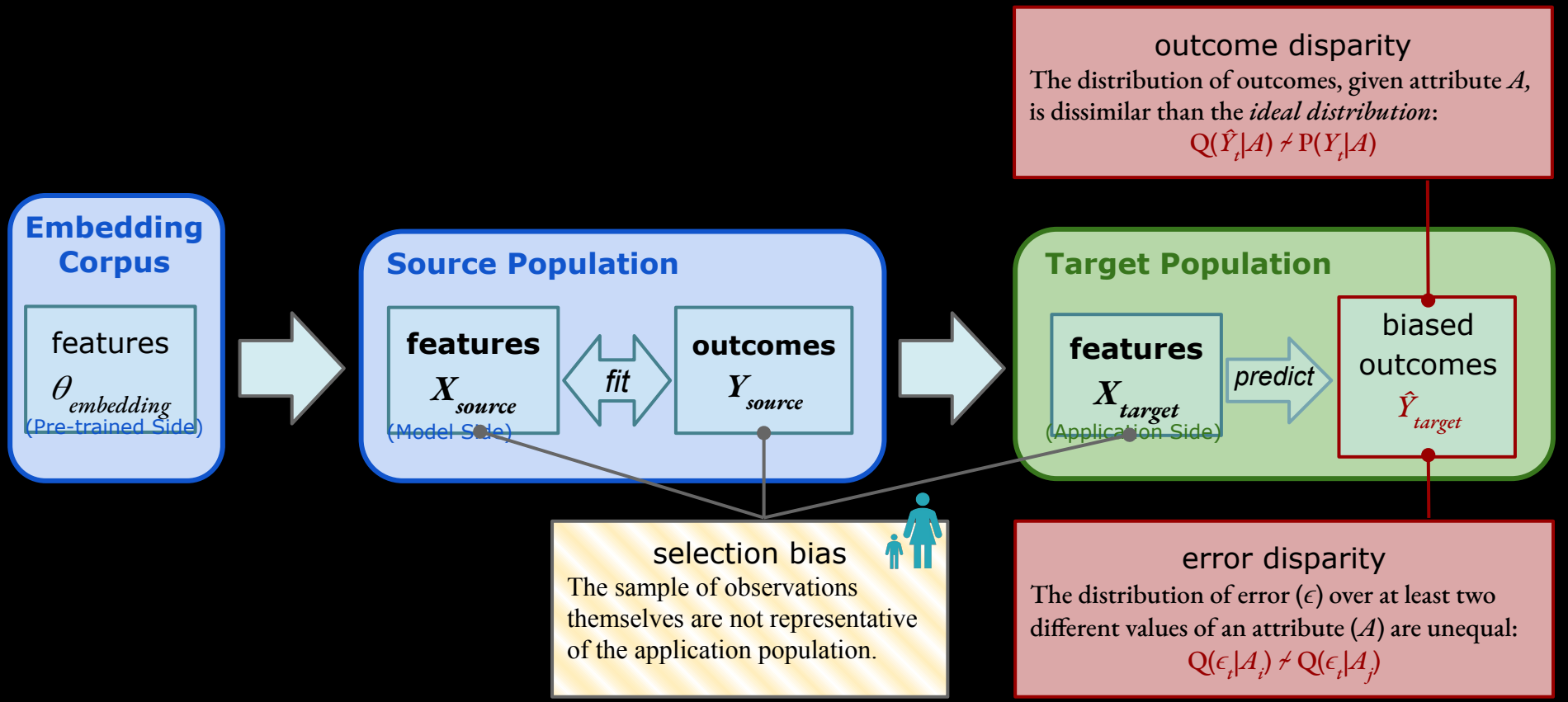
Disparities



Origins of Bias



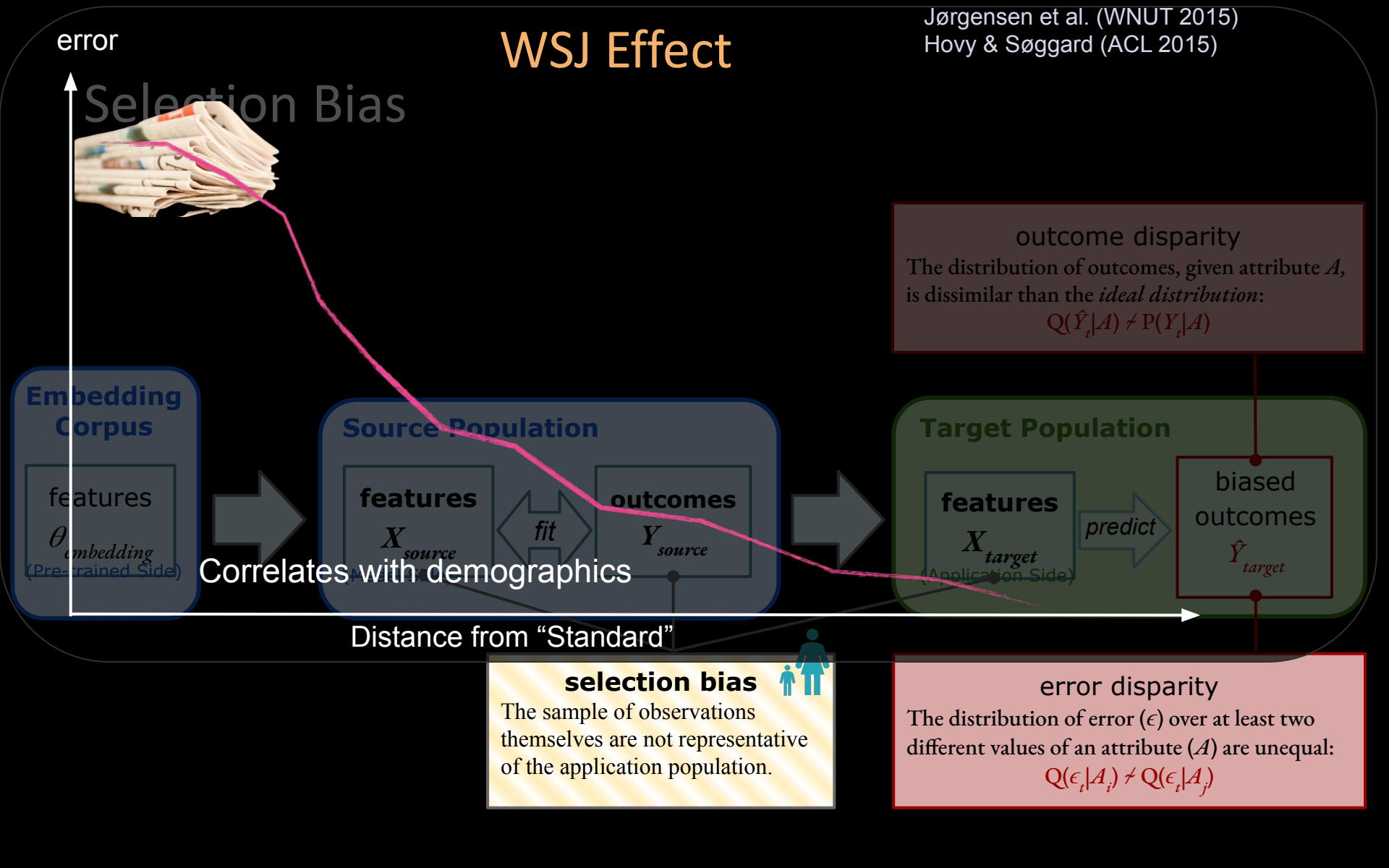
Selection Bias




Selection Bias
proportion of sample
human attribute
value1
value2
Embedding Corpus
features
$\theta_{embedding}$
(Pre-trained Side)
Source
$Q(A_S)$
features
X_{source}
(Model Side)
Target
$P(A_T)$
fit
outcomes
Y_{source}
Target Population
features
X_{target}
(Application Side)
predict
biased outcomes
$\hat{Y}_{target}$
outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$
selection bias
The sample of observations themselves are not representative of the application population.
error disparity
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$

WSJ Effect

Jørgensen et al. (WNUT 2015)
Hovy & Søggard (ACL 2015)



outcome disparity

The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:

$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

selection bias

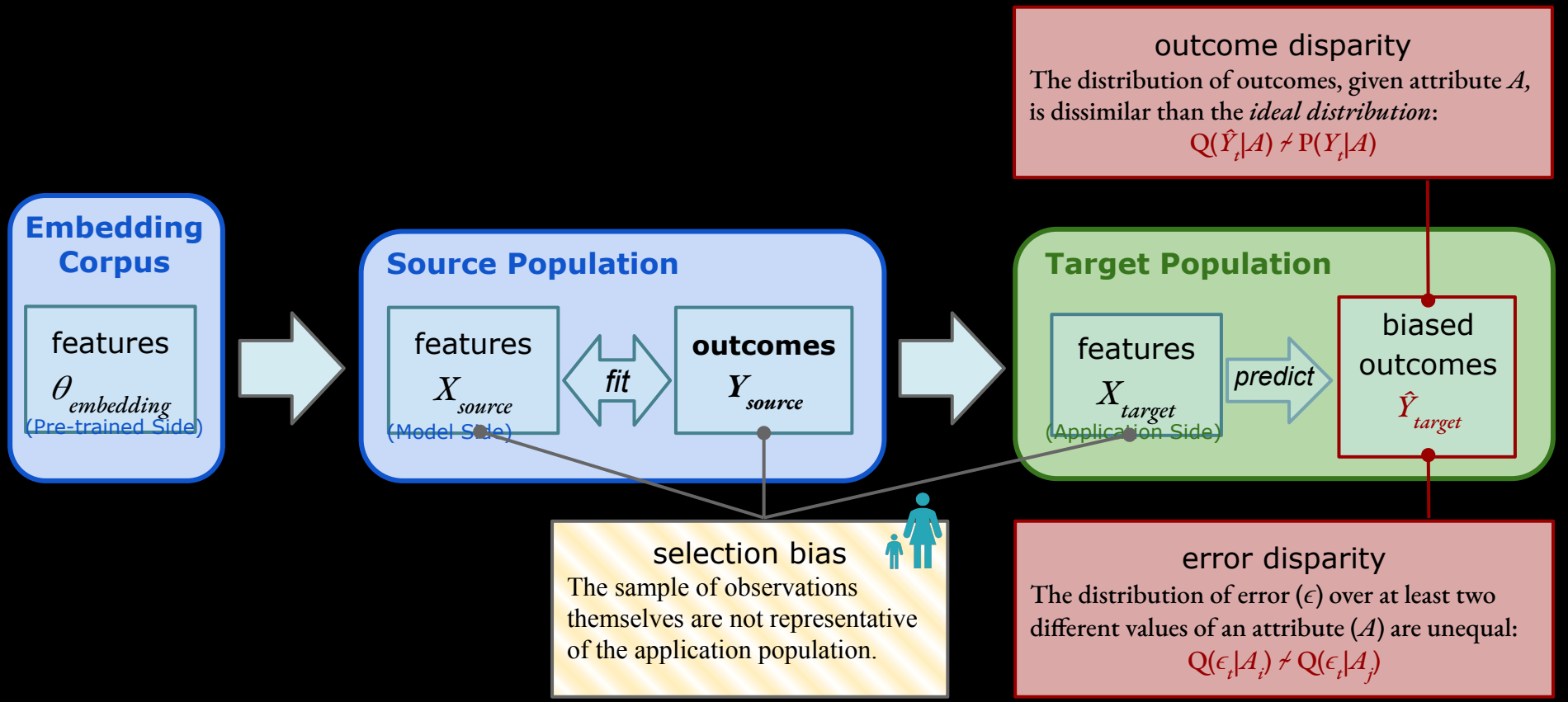
The sample of observations themselves are not representative of the application population.

error disparity

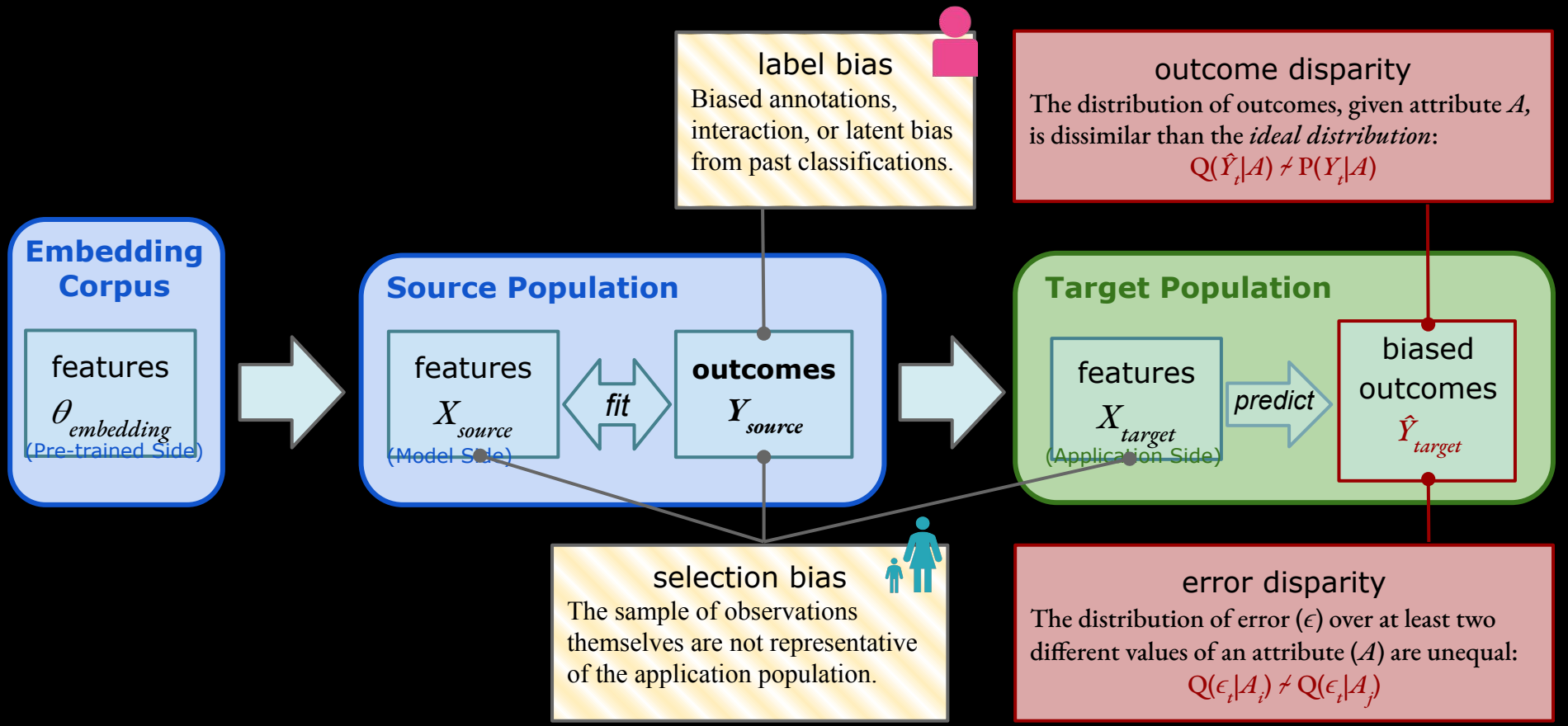
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

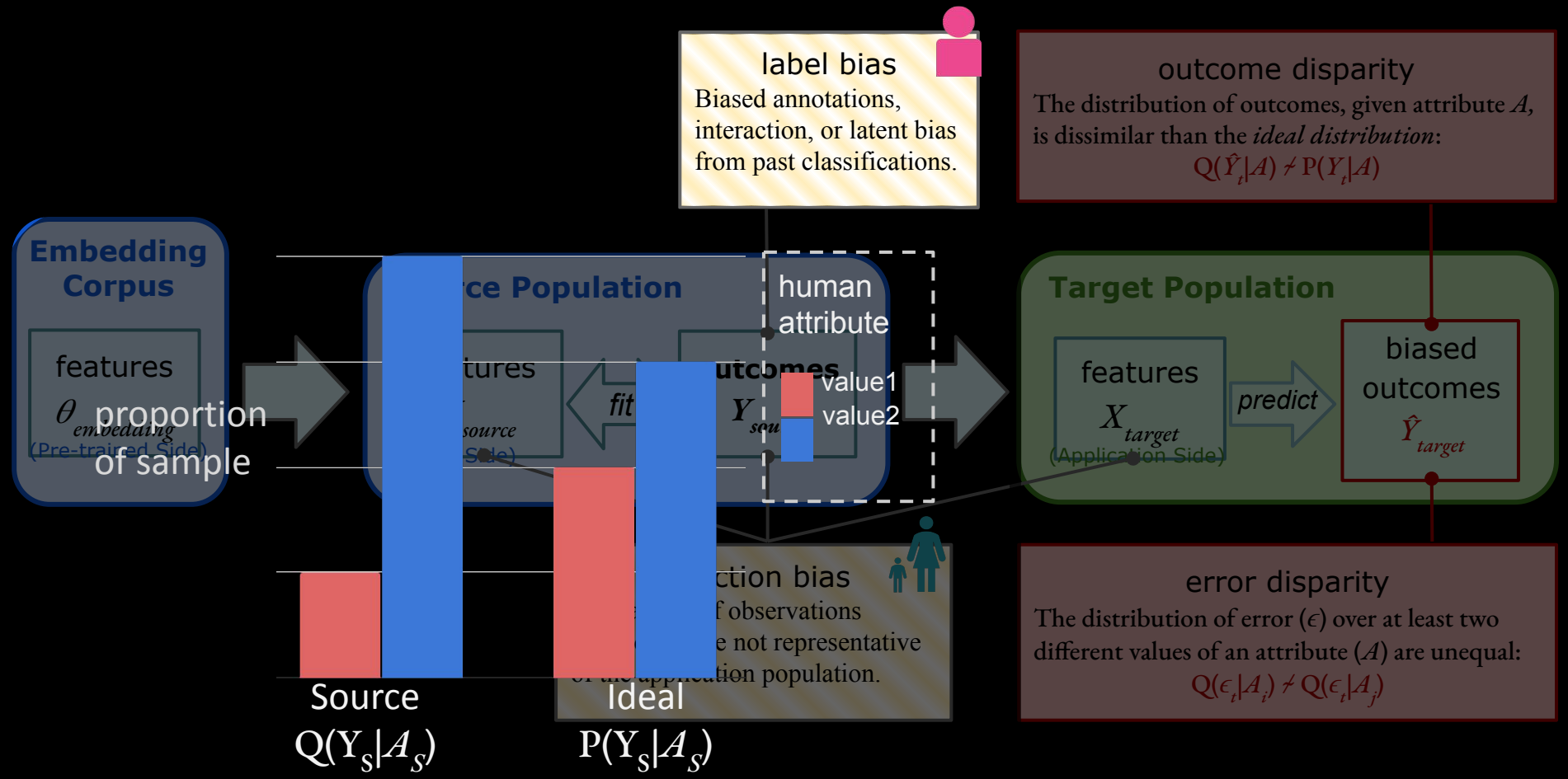
Selection Bias



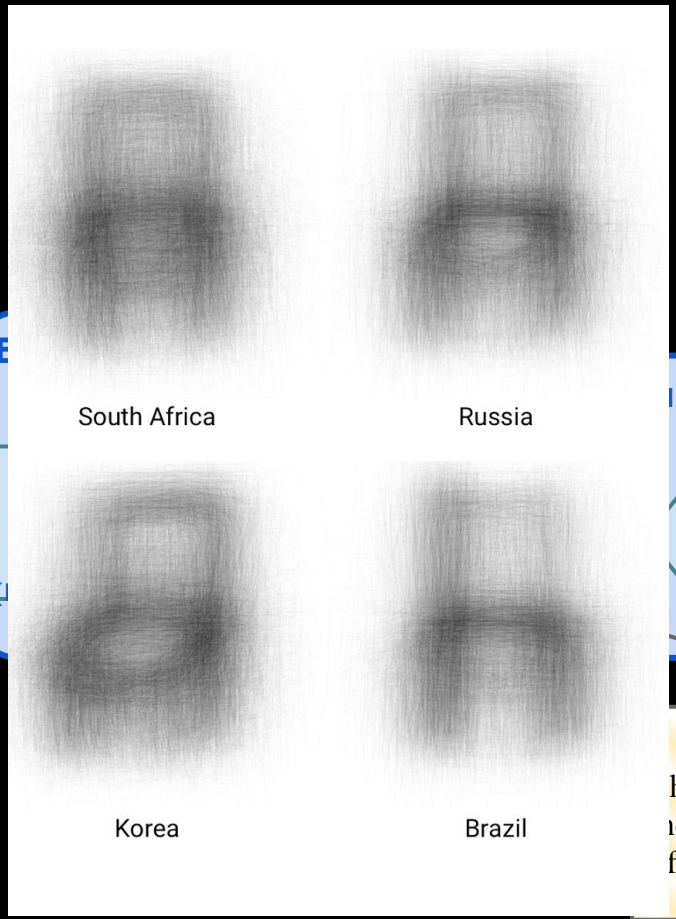
Label Bias


label bias
Biased annotations, interaction, or latent bias from past classifications.
outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$
Embedding Corpus
features
$\theta_{embedding}$
(Pre-trained Side)
Source Population
features
X_{source}
(Model Side)
fit
outcomes
Y_{source}
Target Population
features
X_{target}
(Application Side)
predict
biased outcomes
$\hat{Y}_{target}$
selection bias
The sample of observations themselves are not representative of the application population.
error disparity
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$

Label Bias


label bias
Biased annotations, interaction, or latent bias from past classifications.
outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$
Embedding Corpus
features
$\theta_{embedding}$ proportion of sample
(Pre-trained Side)
ce Population
tures
source
Side)
fit
utcomes
Y_{sou}
human attribute
value1
value2
Target Population
features
X_{target}
(Application Side)
predict
biased outcomes
$\hat{Y}_{target}$
ction bias
f observations
e not representative
of the application population.
error disparity
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$
Source
$Q(Y_S|A_S)$
Ideal
$P(Y_S|A_S)$

Label Bias - Example: Label word with drawing



label bias
Biased annotations, interaction, or latent bias from past classifications.

outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$

lation
fit
outcomes Y_{source}

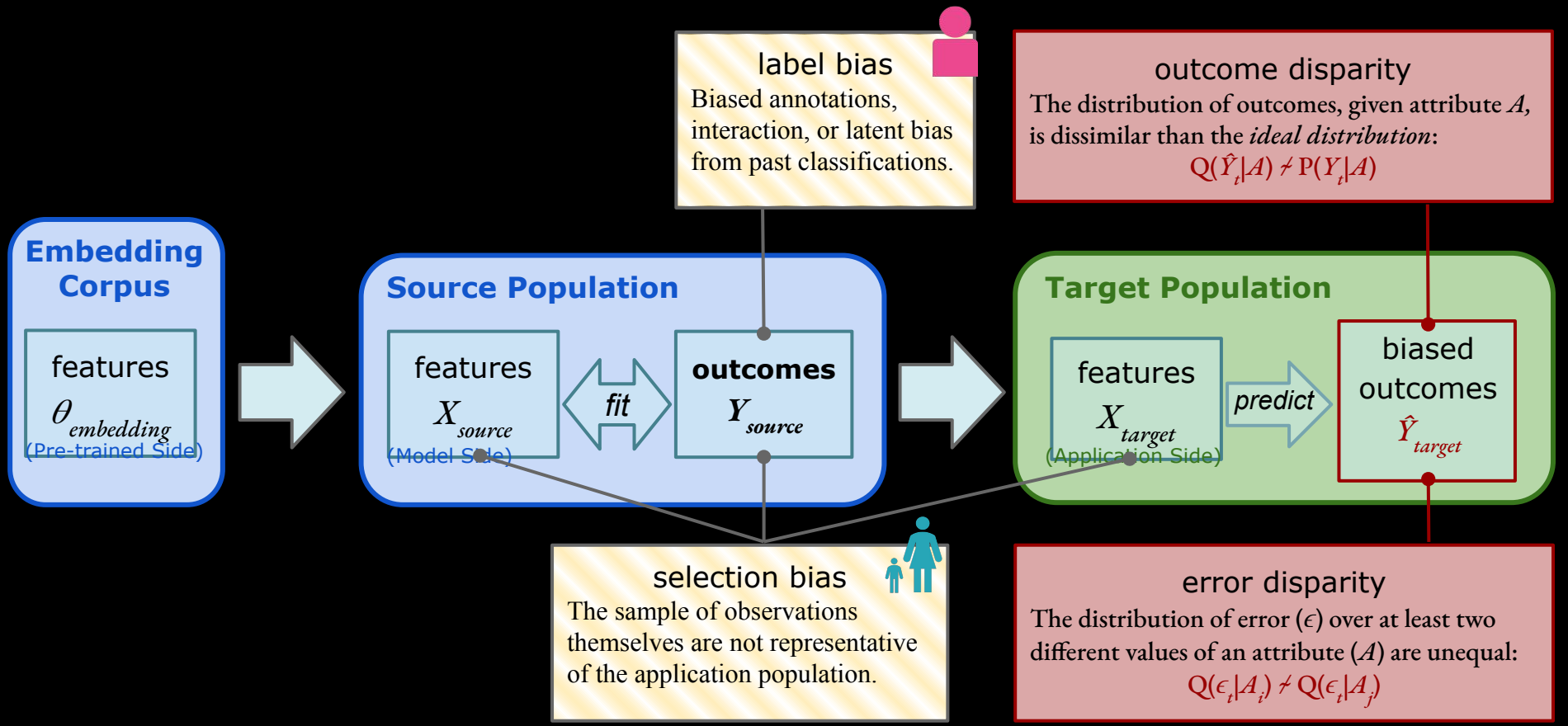
Target Population
features X_{target}
(Application Side)
predict
biased outcomes $\hat{Y}_{target}$

selection bias
he sample of observations
iemselves are not representative
f the application population.

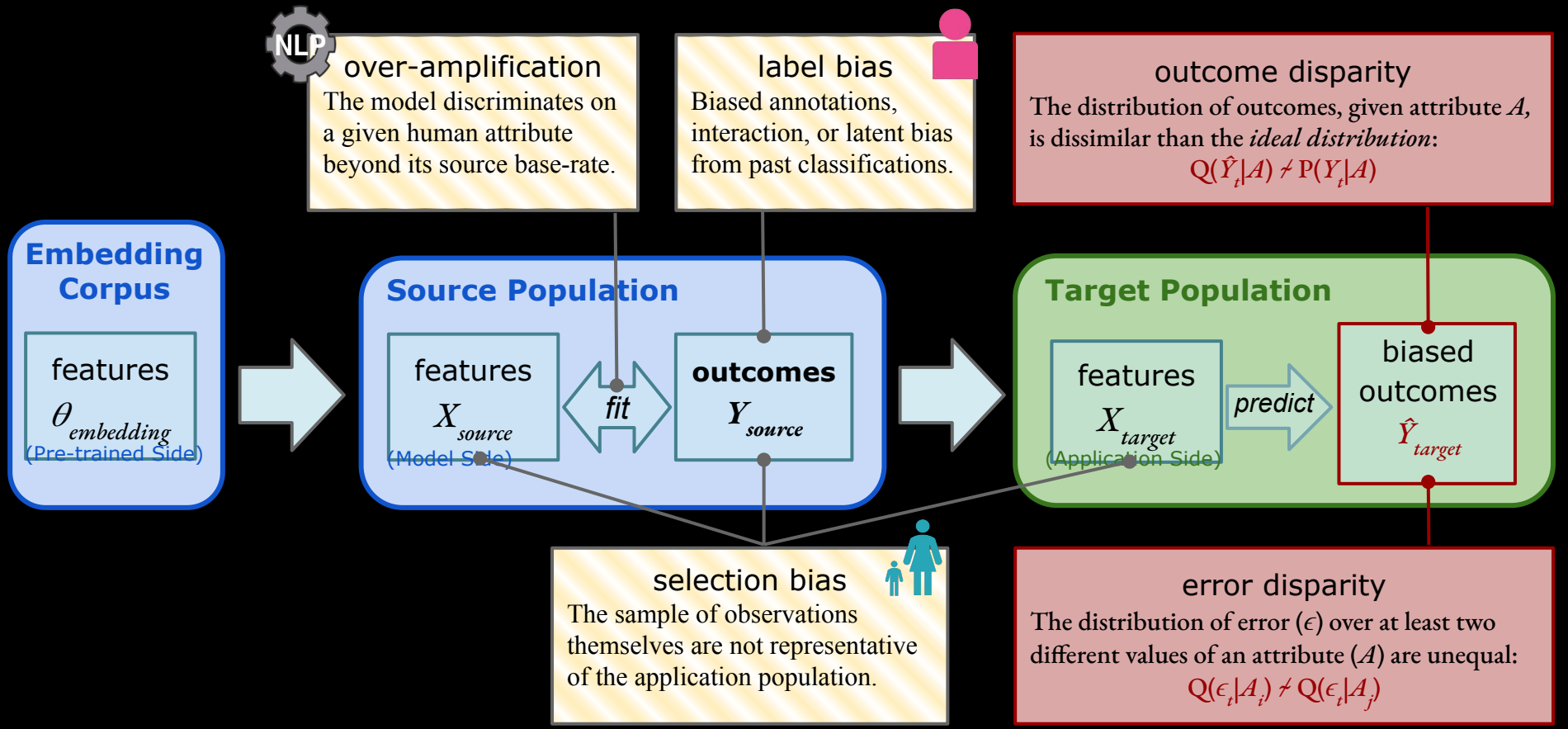
error disparity
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$

Devin Coldeway. 2017. TechCrunch: Google releases millions of bad drawings for you (and your AI) to paw through
https://techcrunch.com/2017/08/25/google-releases-millions-of-bad-drawings-for-you-and-your-ai-to-paw-through/

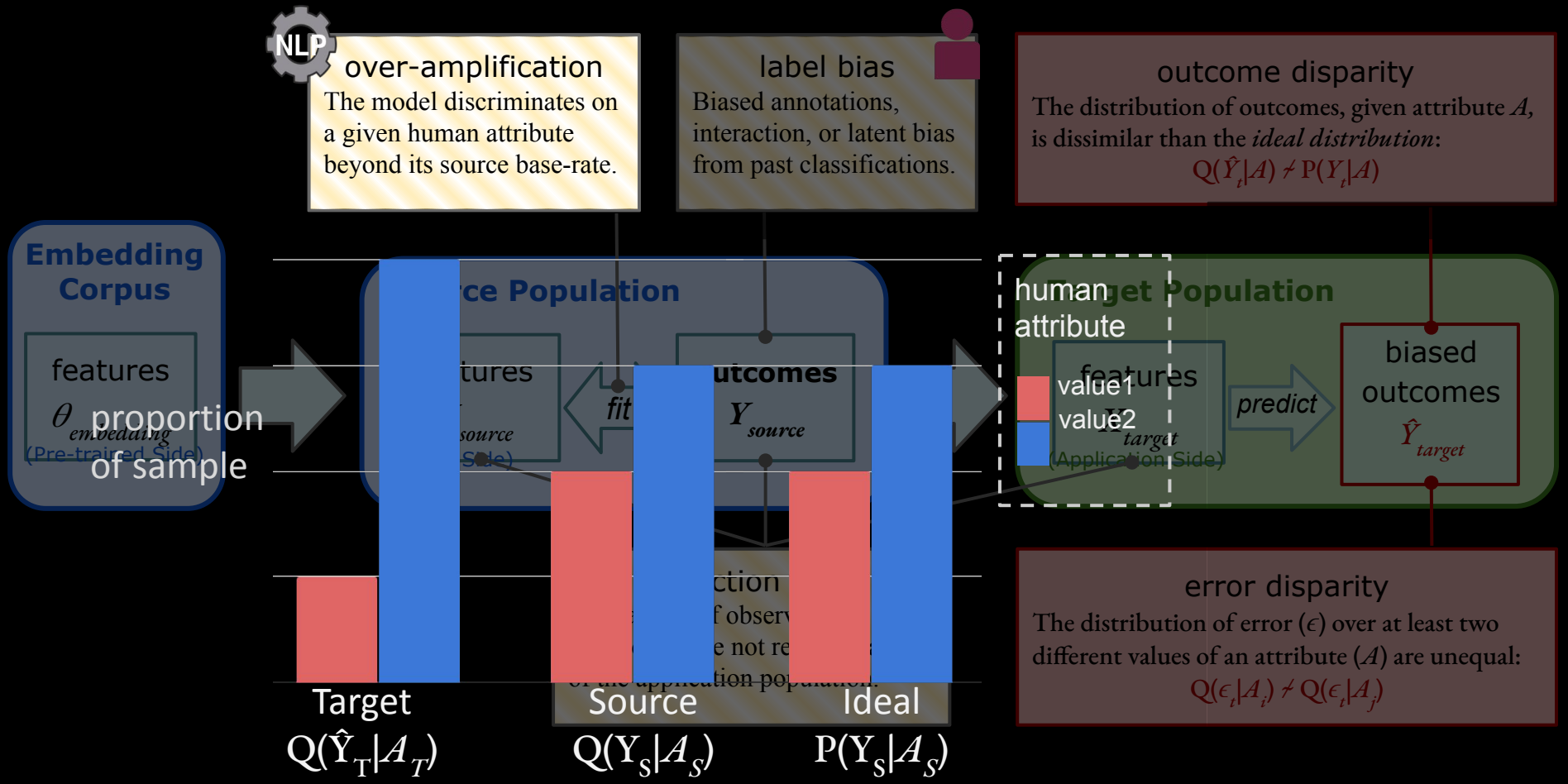
Label Bias


label bias
Biased annotations, interaction, or latent bias from past classifications.
outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$
Embedding Corpus
features
$\theta_{embedding}$
(Pre-trained Side)
Source Population
features
X_{source}
(Model Side)
fit
outcomes
Y_{source}
Target Population
features
X_{target}
(Application Side)
predict
biased outcomes
$\hat{Y}_{target}$
selection bias
The sample of observations themselves are not representative of the application population.
error disparity
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$

Overamplification



Overamplification



Zhao et al. (ACL 2015)

Overamplification - Model Amplifies Bias

BIAS = 0.66

Agent: WOMAN

Agent: MAN

Agent: WOMAN

NLP

BIAS = 0.84

WOMAN

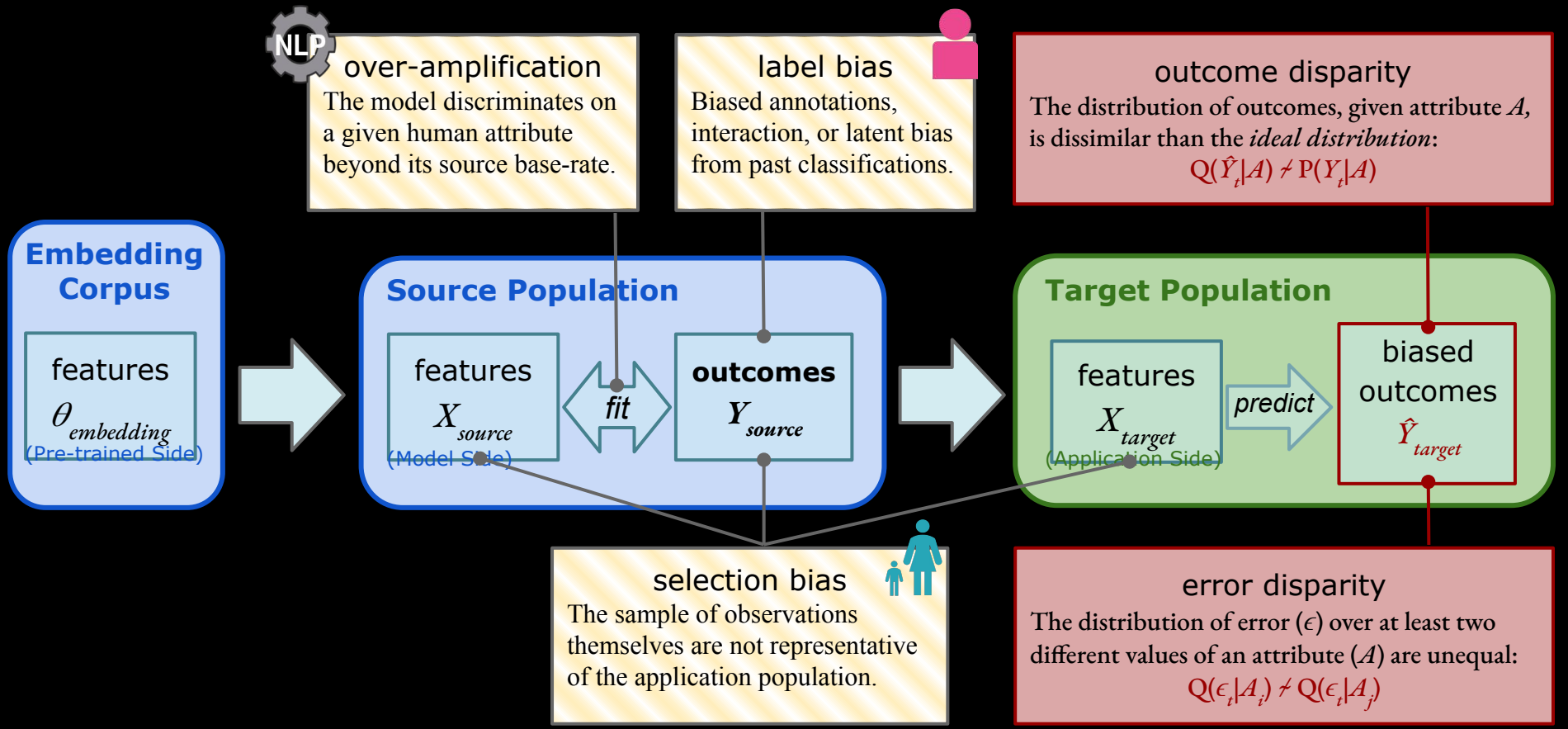
WOMAN

WOMAN

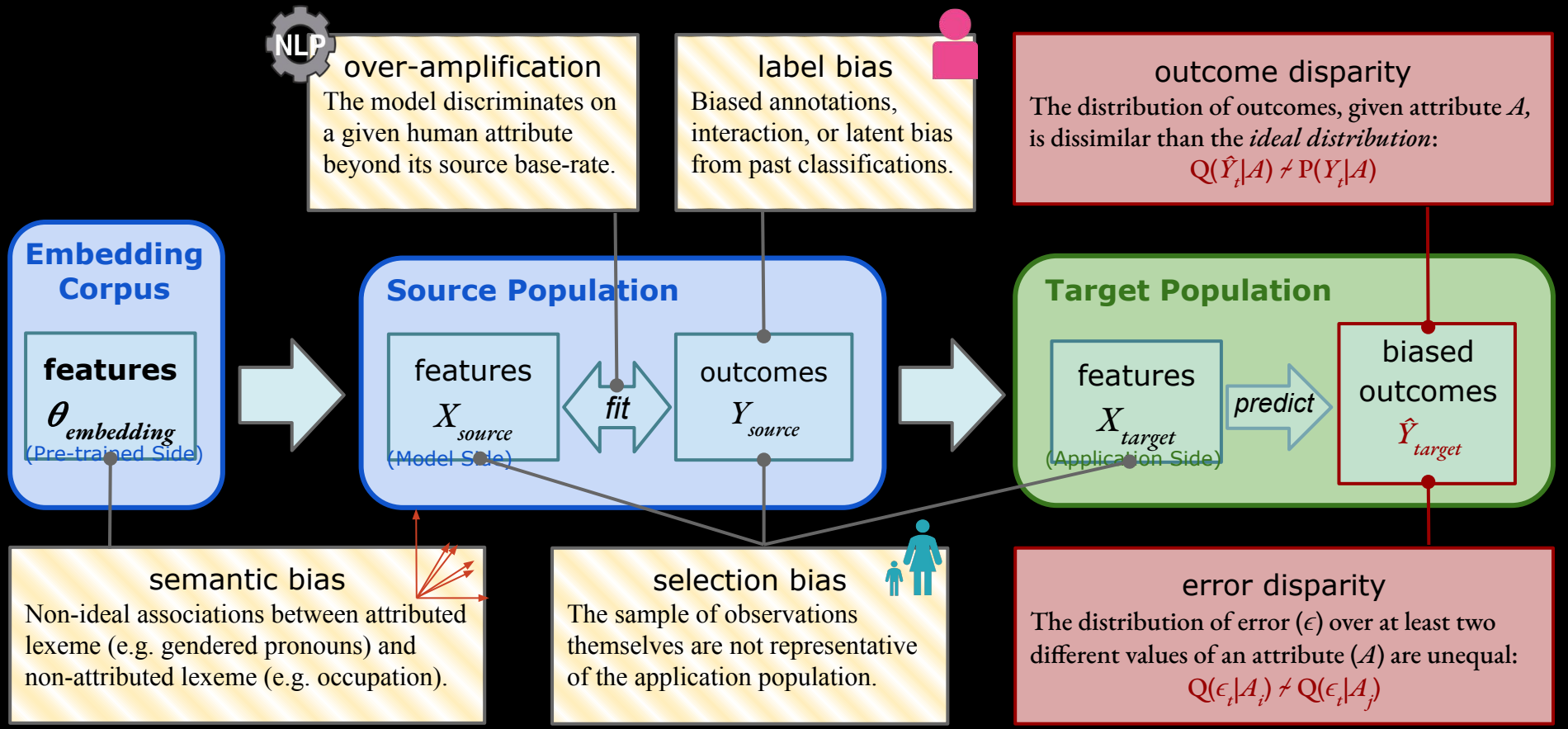
MAN

WOMAN

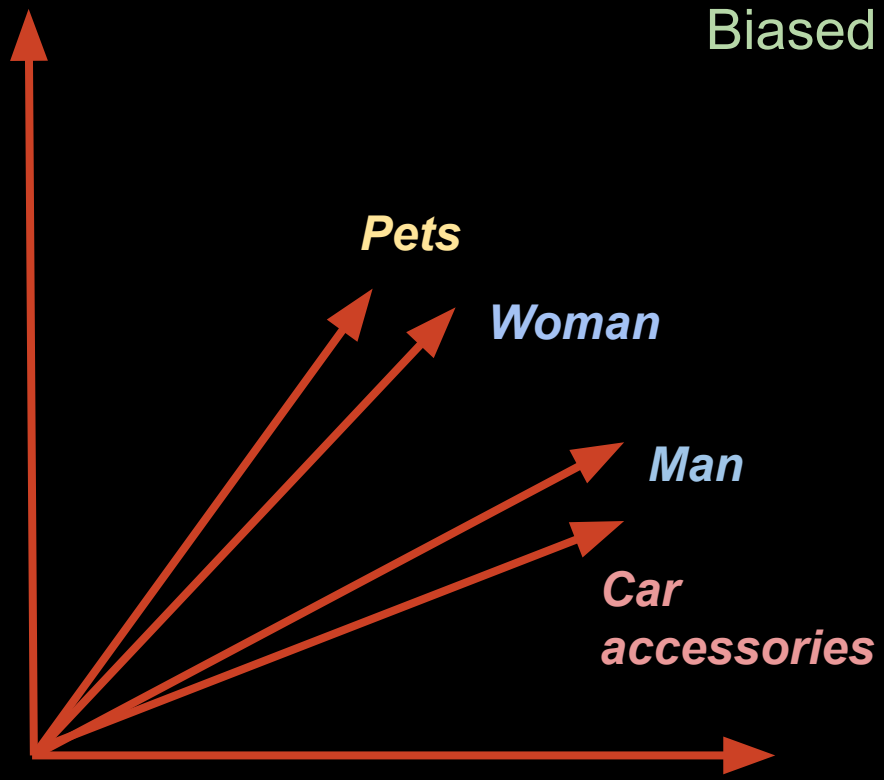
Overamplification



Semantic Bias



Semantic Bias



E.g. Coreference resolution:

connecting entities to references (i.e. pronouns).

"*The doctor told Mary that she had run some blood tests.*"

semantic bias

Non-ideal associations between attributed lexeme (e.g. gendered pronouns) and non-attributed lexeme (e.g. occupation).

selection bias

The sample of observations themselves are not representative of the application population.

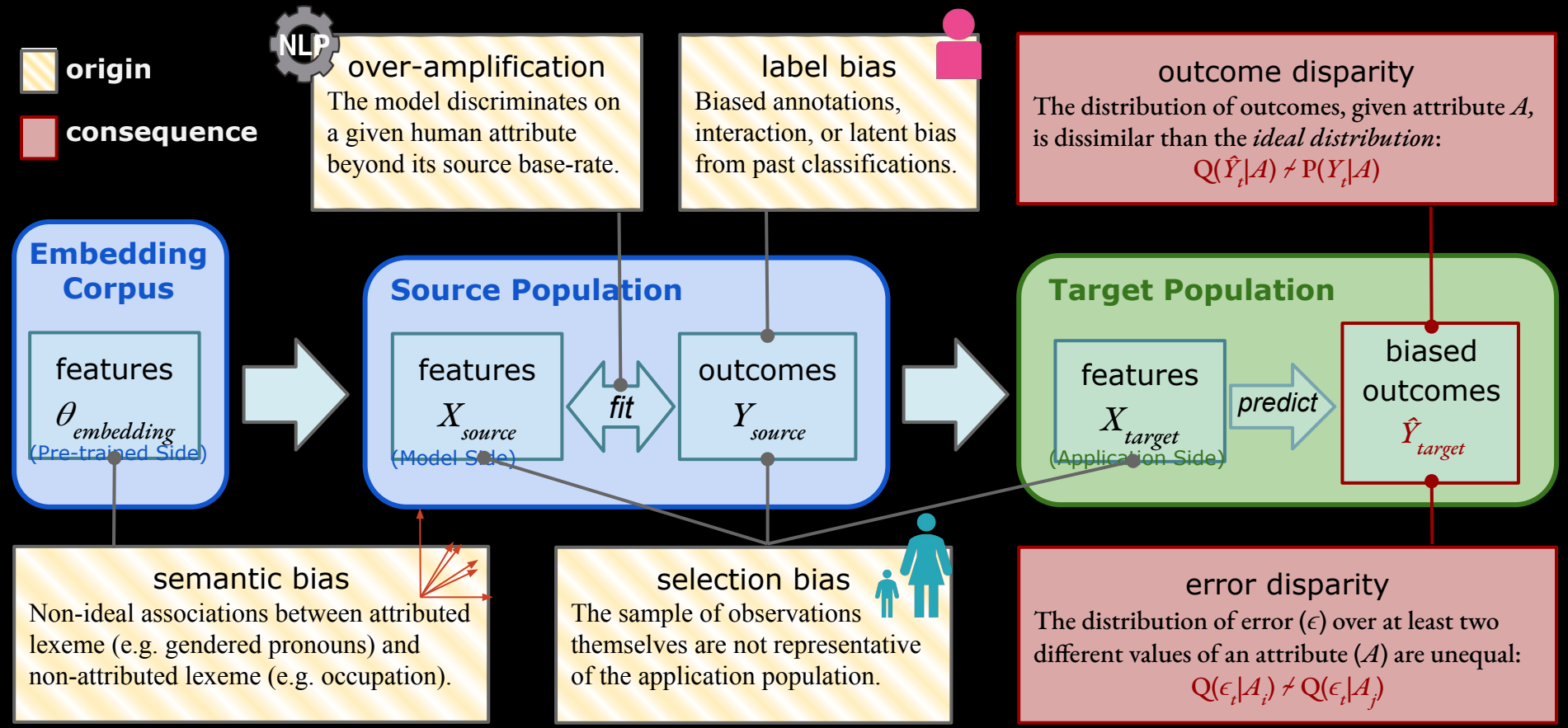
error disparity

The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

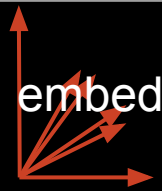
$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

Shah, D., Schwartz, H. A., Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *In ACL-2020: Proceedings of the Association for Computational Linguistics.*

Predictive Bias Framework for NLP



Summary of Countermeasures

Source	Origin	Countermeasures
 annotation	**Label Bias**	Post-stratification, Re-train annotators
 data selection	**Selection Bias**	Stratified sampling, Post-stratification or Re-weighing techniques
 models	**Overamplification**	Synthetically match distributions, add outcome disparity to cost function
 embeddings	**Semantic Bias**	Use above techniques and re-train embeddings

Mitigating Bias in LLMs - "Alignment"

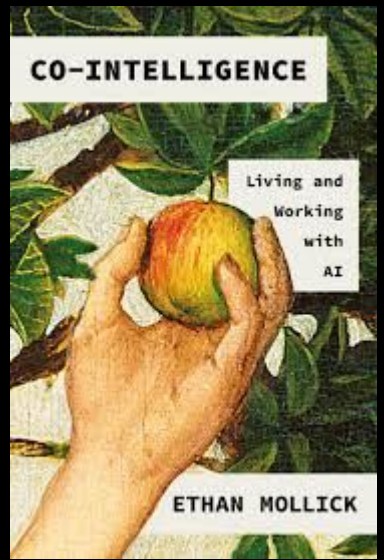
- Most of bias mitigation is via RLHF (Reinforcement Learning from Human Feedback).

  Updating weights to favor responses judged positively.

- Also introduces "guard rails"
  – safety limits to restrict AI behavior.

- Ethical dilemma: Annotators exposure to distressing content.

- Additional risk: Manipulating guard rails



(chapter 2)

Bias - Takeaways

Bias, as outcome and error **disparities**, can result from many **origins**:

- the **embedding** model
- the feature **sample**
- the **fitting** process
- the **outcome** sample

Our understanding is evolving:

This is an active area of work, both theoretically and technically!

Ethics in NLP

Bias

Privacy

Ethical Research and Development

Ethics in NLP

Bias

Privacy

Ethical Research and Development

Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion

Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion
- Mitigation strategies:

Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion
- Mitigation strategies:
  - Anonymize where possible – remove named entities

Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion
- Mitigation strategies:
  - Anonymize where possible – remove named entities
  - Informed consent -- let participants know and opportunity to opt-in/-out
    - Information targeting: “You are being shown this ad because ...”
  - Do not share / secure storage

Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion
- Mitigation strategies:
  - Anonymize where possible – remove named entities
  - Informed consent -- let participants know and opportunity to opt-in/-out
    - Information targeting: “You are being shown this ad because ...”
  - Do not share / secure storage
  - *Federated learning* -- obfuscate to the point of preserving privacy

Ethics in NLP

Bias

Privacy

Ethical Research and Development

Ethics in NLP

Bias

Privacy

Ethical Research and Development

Ethics in NLP Research

ACM Code of Ethics; General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.

https://www.acm.org/code-of-ethics

Ethics in NLP Research

ACM Code of Ethics; General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
- Avoid harm.

https://www.acm.org/code-of-ethics

Ethics in NLP Research

ACM Code of Ethics; General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
- Avoid harm.
- Be honest and trustworthy.

https://www.acm.org/code-of-ethics

Ethics in NLP Research

ACM Code of Ethics; General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
- Avoid harm.
- Be honest and trustworthy.
- Be fair and take action not to discriminate.

https://www.acm.org/code-of-ethics

Ethics in NLP Research

ACM Code of Ethics; General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
- Avoid harm.
- Be honest and trustworthy.
- Be fair and take action not to discriminate.
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.

https://www.acm.org/code-of-ethics

Ethics in NLP Research

ACM Code of Ethics; General Ethical Principles:

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing.
- Avoid harm.
- Be honest and trustworthy.
- Be fair and take action not to discriminate.
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
- Respect privacy.
- Honor confidentiality.

https://www.acm.org/code-of-ethics

Ethics in NLP

Human Subjects Research

Observational versus Interventional

Ethics in NLP

Human Subjects Research

Observational versus Interventional

(The Belmont Report, 1979)

(i) Distinction of research from practice.
(ii) Risk-Benefit criteria
(iii) Appropriate selection of human subjects for participation in research
(iv) Informed consent in various research settings.

Ethics in NLP

Human Subjects Research

Observational versus Interventional
(modeling) (models interact)

Ethics in NLP

Human Subjects Research

Observational versus Interventional
(modeling) (models interact)

Deploying a model within an application often shifts the works from being simply observational (privacy harms) to interventional (consideration for additional harms).

Ethics in NLP

Bias – Consider target application and population.

Alignment - LLM based Safety and Bias Mitigation

Privacy - Secure, do not share, and inform

Ethical Research and Development