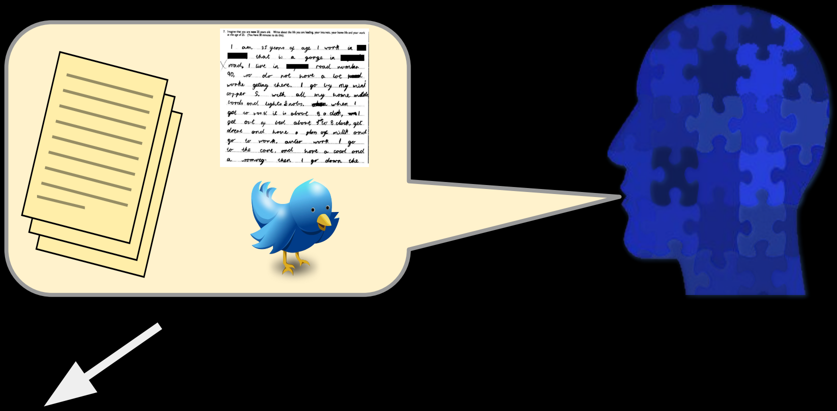
Human-Centered NLP and Ethics in NLP

CSE 354

The “Task” of human-centered NLP



7. Imagine that you are **now** 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

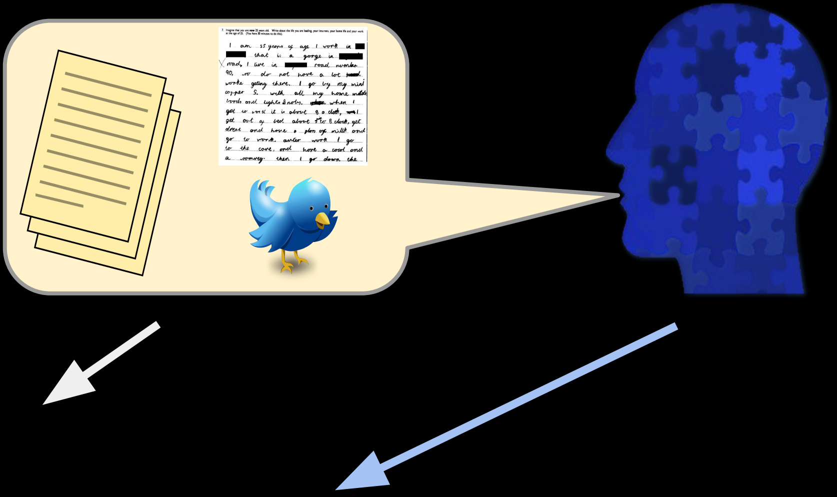
I am 25 years of age I work in [redacted]
[redacted] that is a gorge in [redacted]
road, I live in [redacted] road number
90, we do not have a lot
worke geting there. I go by my mini
copper S. with all my home made
bords and lights & nobs. when I
get to work it is about 8 o clock, I
get out of bed about 5 to 8 clock, get
drese and have a glass of millk and
go to work. after work I go
to the cave, and have a cool and
a somveg. then I go down the

Most NLP Tasks. E.g.

- POS Tagging
- Document Classification
- Sentiment Analysis
- Stance Detection
- Mental Health Risk Assessment
- …

  (language modeling, QA, …

The “Task” of human-centered NLP



7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home madde boords and lights & nobs. when I got to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drese and have a glass of milk and go to work. after work I go to the cave. and have a cocol and a somveg. then I go down the

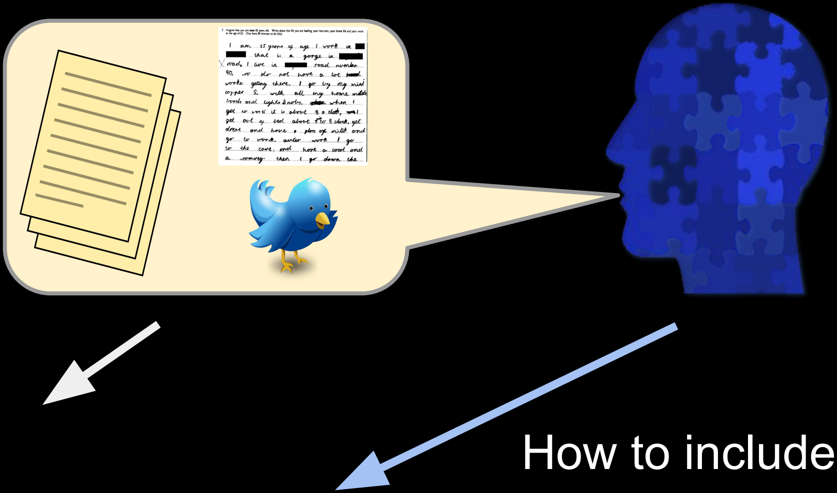
age
gender
personality
expertise
beliefs
…

Most NLP Tasks. E.g.

- POS Tagging
- Document Classification
- Sentiment Analysis
- Stance Detection
- Mental Health Risk Assessment
- …

(language modeling, QA, …

The “Task” of human-centered NLP



7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, so do not have a lot worke geting there. I go by my mini copper S. with all my home madde bords and lights & nobs. when I get to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drest and have a glass of millk and go to work. after work I go to the cave, and have a cool and a somveg. then I go down the

age
gender
personality
expertise
beliefs
…

Most NLP Tasks. E.g.

- POS Tagging
- Document Classification
- Sentiment Analysis
- Stance Detection
- Mental Health Risk Assessment
- …
  (language modeling, QA, …

How to include extra-linguistics?

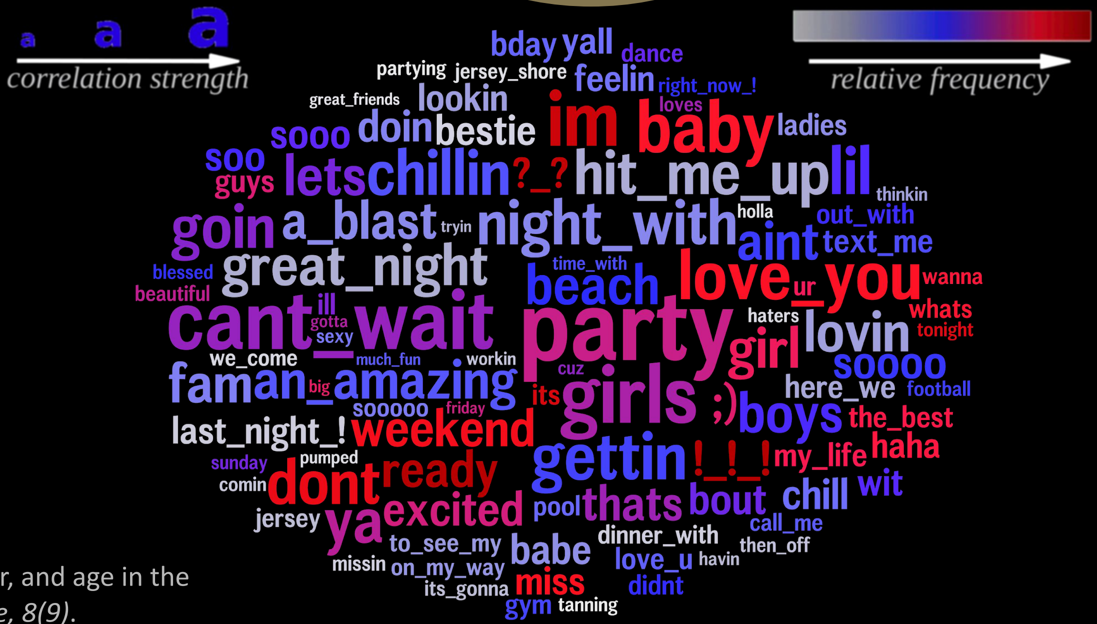
- Additive Inclusion
- Adaptive Extralinguistics
  - Adapting Embeddings
  - Adapting Models
- Correcting for bias

Natural Language Processing

Psychological & Health Sciences

Natural Language Processing

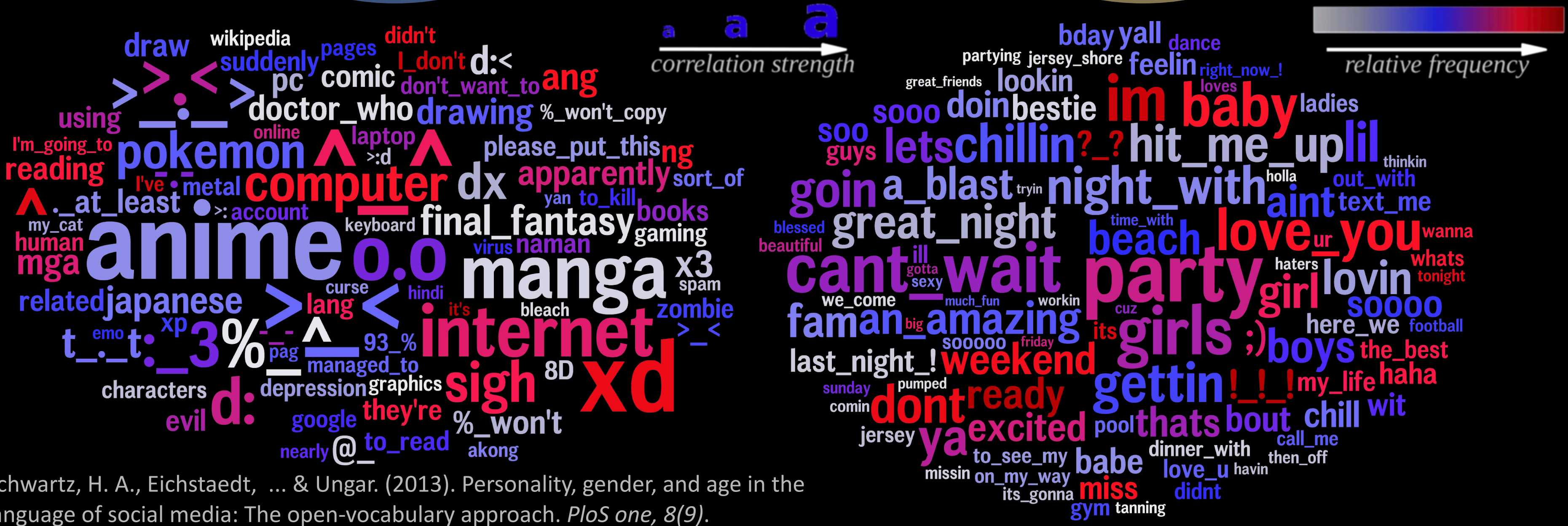
Psychological & Health Sciences



Schwartz, H. A., Eichstaedt, ... & Ungar. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8(9)*.

Natural
Language
Processing

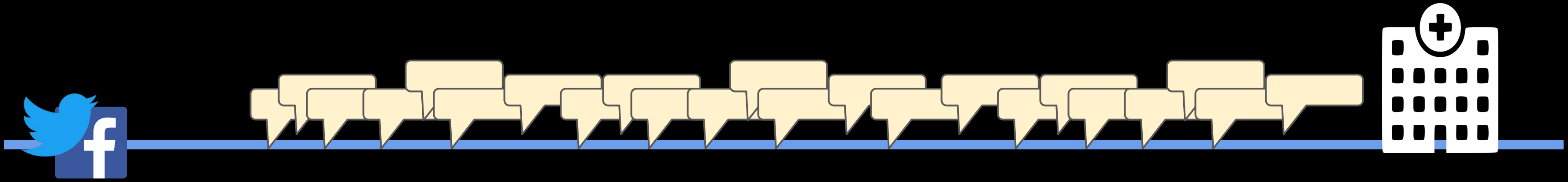
Psychological
& Health
Sciences



Schwartz, H. A., Eichstaedt, ... & Ungar. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one, 8(9)*.

Natural Language Processing

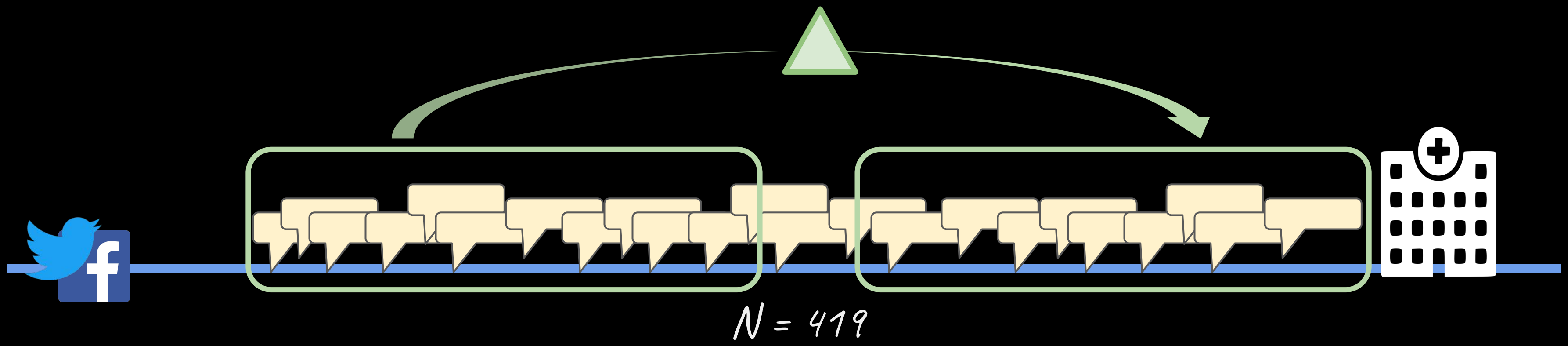
Psychological & Health Sciences



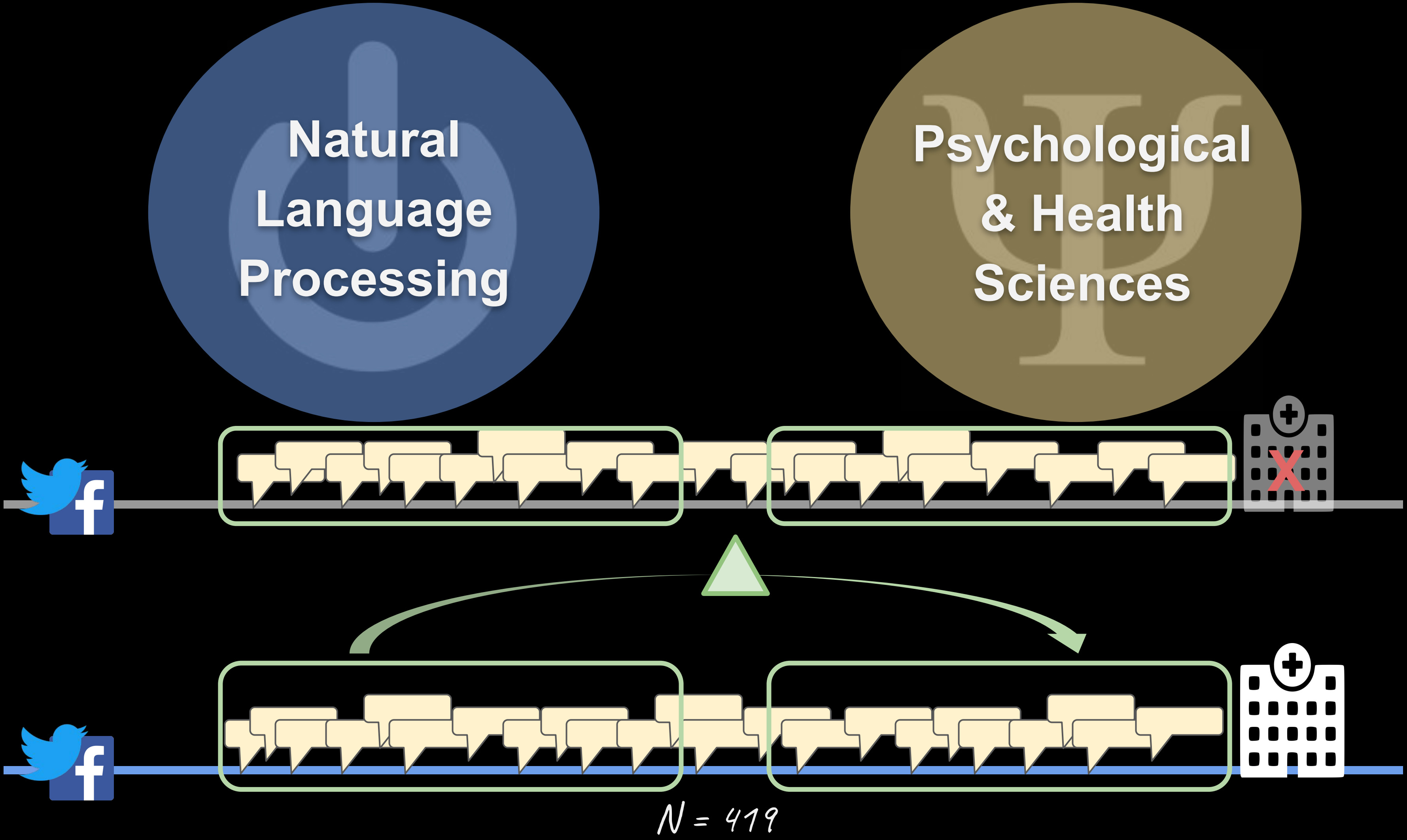
Guntuku, S. C., Schwartz, H. A., Kashyap, A., Gaulton, J. S., Stokes, D. C., Asch, D. A., ... & Merchant, R. M. (2020). Variability in Language used on Social Media prior to Hospital Visits. *Nature - Scientific Reports*, *10(1),* 1-9.

Natural Language Processing

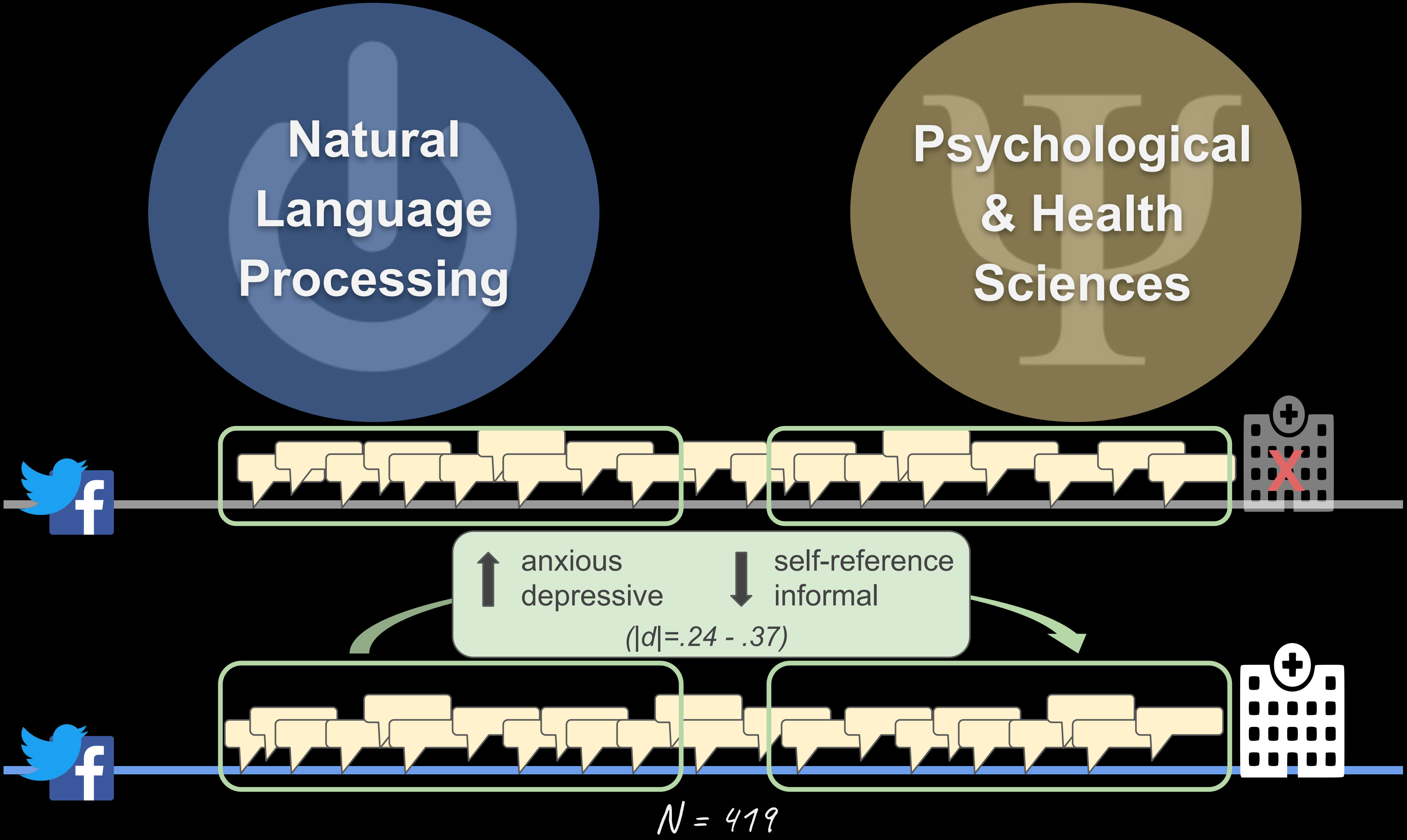
Psychological & Health Sciences



Guntuku, S. C., Schwartz, H. A., Kashyap, A., Gaulton, J. S., Stokes, D. C., Asch, D. A., ... & Merchant, R. M. (2020). Variability in Language used on Social Media prior to Hospital Visits. *Nature - Scientific Reports*, *10(1),* 1-9.



Guntuku, S. C., Schwartz, H. A., Kashyap, A., Gaulton, J. S., Stokes, D. C., Asch, D. A., ... & Merchant, R. M. (2020). Variability in Language used on Social Media prior to Hospital Visits. *Nature - Scientific Reports*, *10(1),* 1-9.



Natural Language Processing

Psychological & Health Sciences

⬆ anxious
depressive

⬇ self-reference
informal

(|*d*|=.24 - .37)

N = 419

Guntuku, S. C., Schwartz, H. A., Kashyap, A., Gaulton, J. S., Stokes, D. C., Asch, D. A., ... & Merchant, R. M. (2020). Variability in Language used on Social Media prior to Hospital Visits. *Nature - Scientific Reports*, *10(1),* 1-9.

Natural Language Processing

Psychological & Health Sciences

Problem

Natural language is written by

Problem

Natural language is written by **people.**

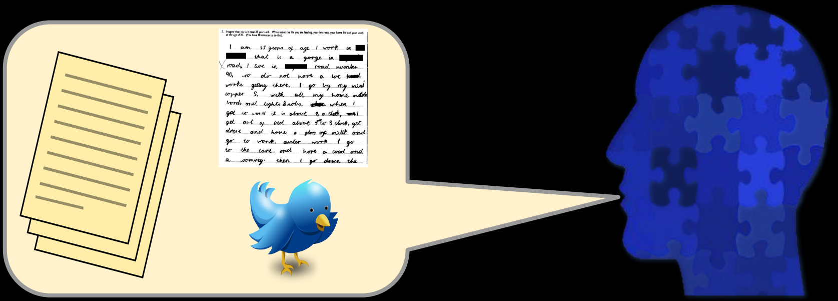
Problem

Natural language is written by **people.**

Problem

Natural language is written by **people.**

Natural language is generated by people.

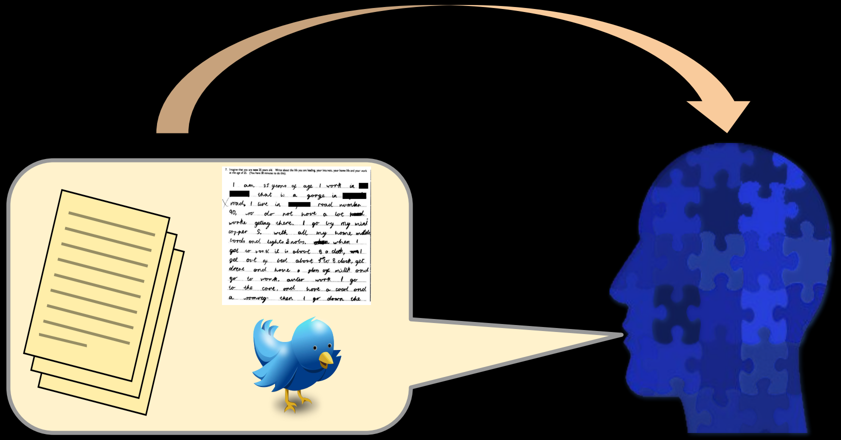


7. Imagine that you are **now** 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot [illegible] worke geting there. I go by my mini copper S. with all my home madde bords and lights & nobs. [illegible] when I get to work it is about 8 o clock, [illegible] I get out of bed about 5 to 8 clock, get drese and have a glass of millk and go to work. after work I go to the cave. and have a cocol and a somveg. then I go down the

People have different beliefs, backgrounds, styles, vocabularies, preferences, knowledge, personalities, …

Natural language is generated by people.



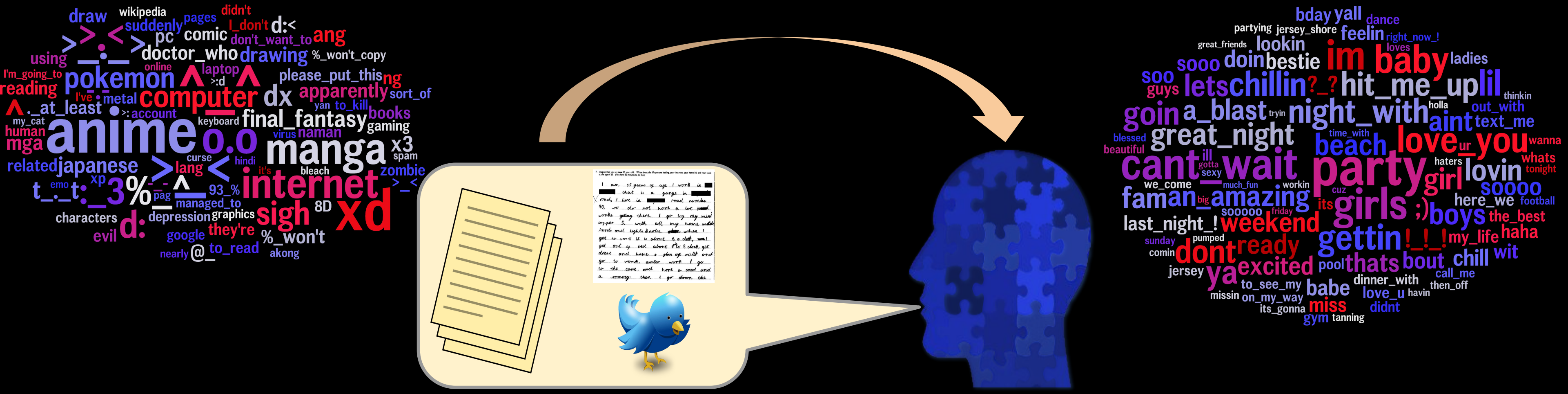
7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted]
that is a gorge in [redacted]
road, I live in [redacted] road number
90, we do not have a lot
worke geting there. I go by my mini
copper S. with all my home madde
bords and lights & nobs. when I
get to work it is about 8 o clock, I
get out of bed about 5th to 8 clock, get
drese and have a glas of millk and
go to work. arter work I go
to the cave, and have a cool and
a somveg. then I go down the

People have different beliefs, backgrounds, styles, vocabularies, preferences, knowledge, personalities, …,

and our language reflects these differences.

Natural language is generated by people.



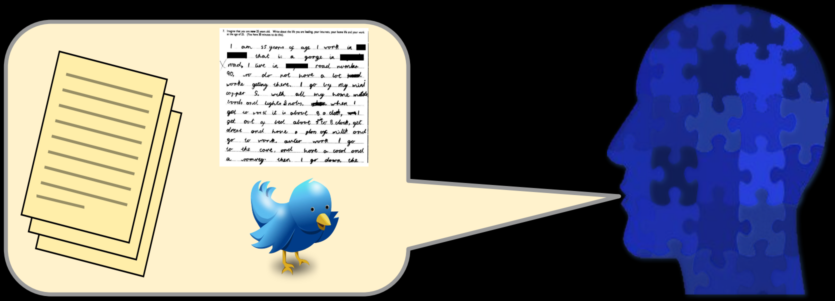
7. Imagine that you are **now** 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home madde bords and lights &nobs. when I get to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drese and have a glass of millk and go to work. aview work I go to the cave, and have a cool and a somveg. then I go down the

People have different beliefs, backgrounds, styles, vocabularies, preferences, knowledge, personalities, …,

and our language reflects these differences.

Human Centered NLP:

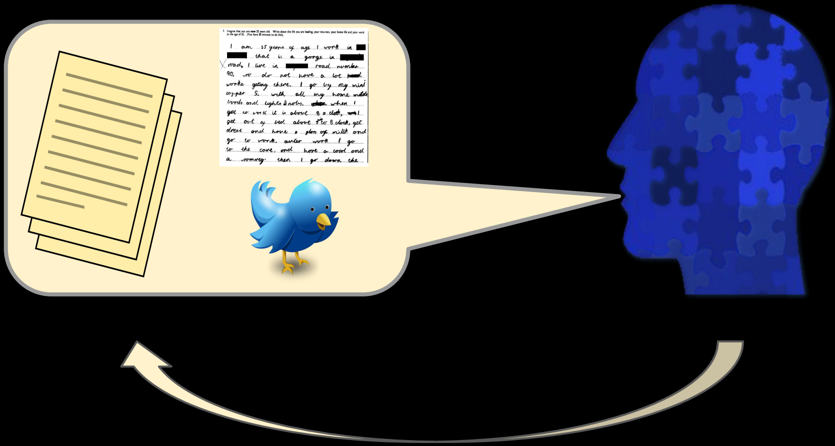


7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home made bords and lights & nobs. when I got to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drest and have a glass of millk and go to work. after work I go to the cove, and have a cood and a somveg. then I go down the

Human Centered NLP:

1. Model language as a human process

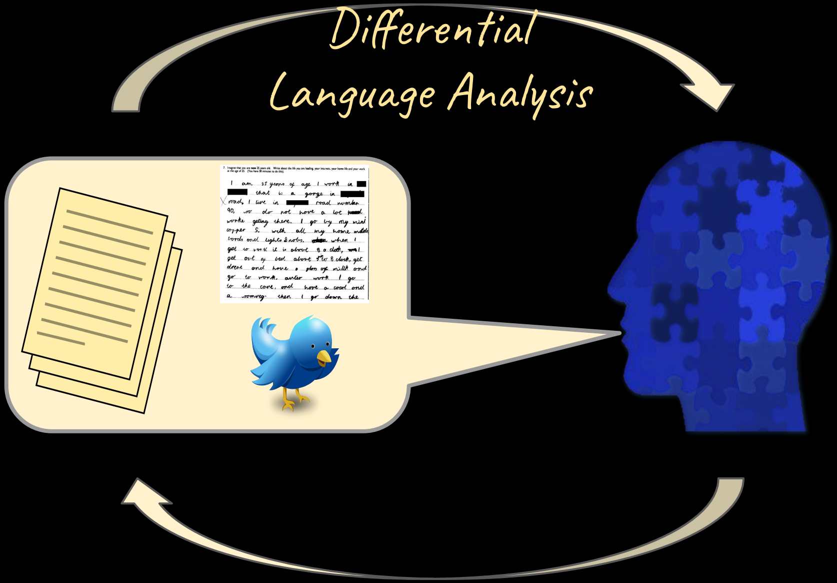


7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home madde bords and lights & nobs. when I get to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get dress and have a glass of millk and go to work. after work I go to the cove, and have a cool and a somveg. then I go down the

Human Centered NLP:

1. Model language as a human process
2. Use language to better understand humans.



7. Imagine that you are **now** 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 yeors of age I work in [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home made bords and lights & nobs. when I get to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drest and have a glos of millk and go to work. after work I go to the cafe, and have a cocol and a somveg. then I go down the

Differential Language Analysis

Input:

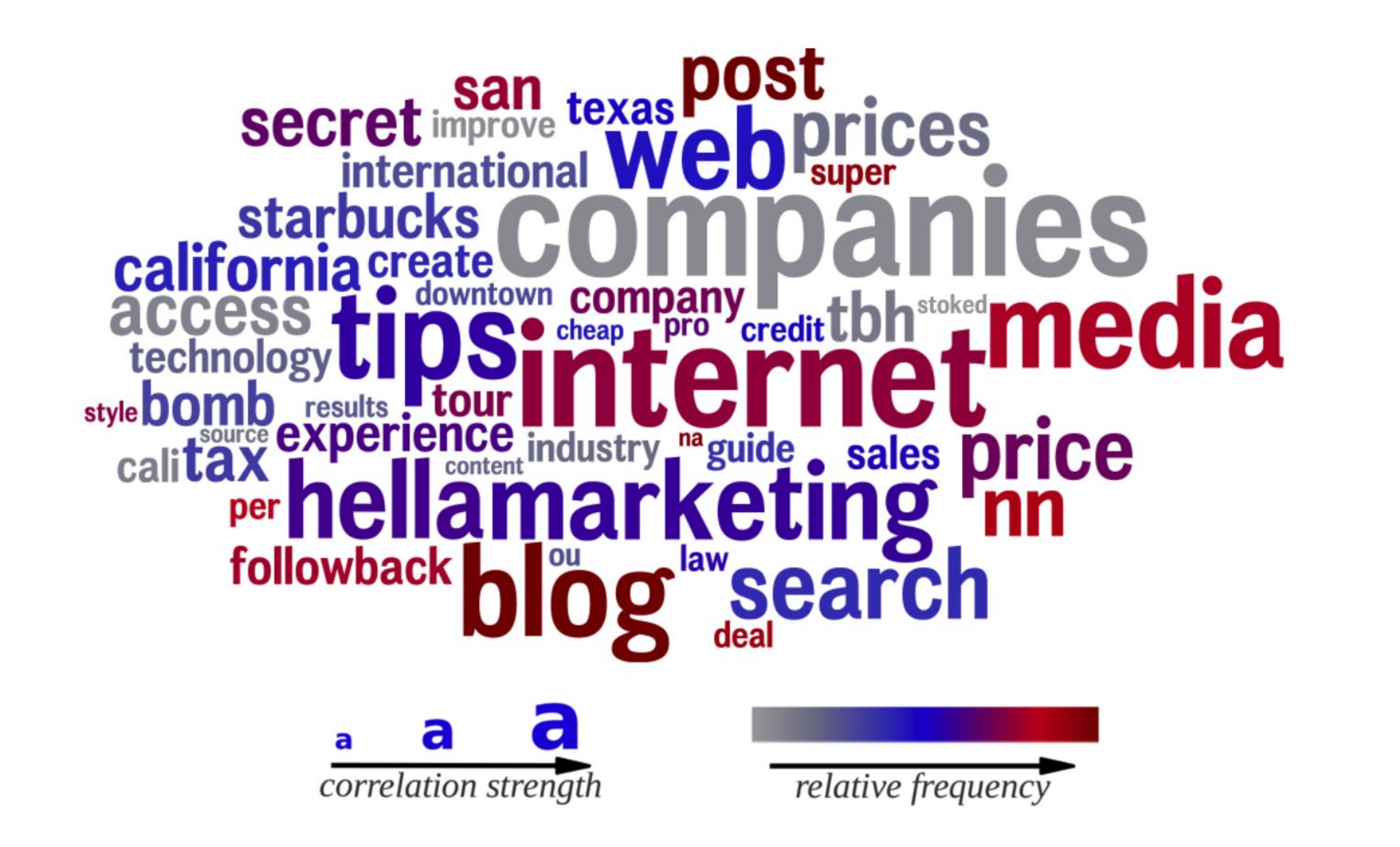
- Linguistic features
- Human or community attribute

Output:

- Features distinguishing attribute

Goal: Data-driven insights about an attribute

E.g. Words distinguishing communities with increases in real estate prices.



Differential Language Analysis

Input:

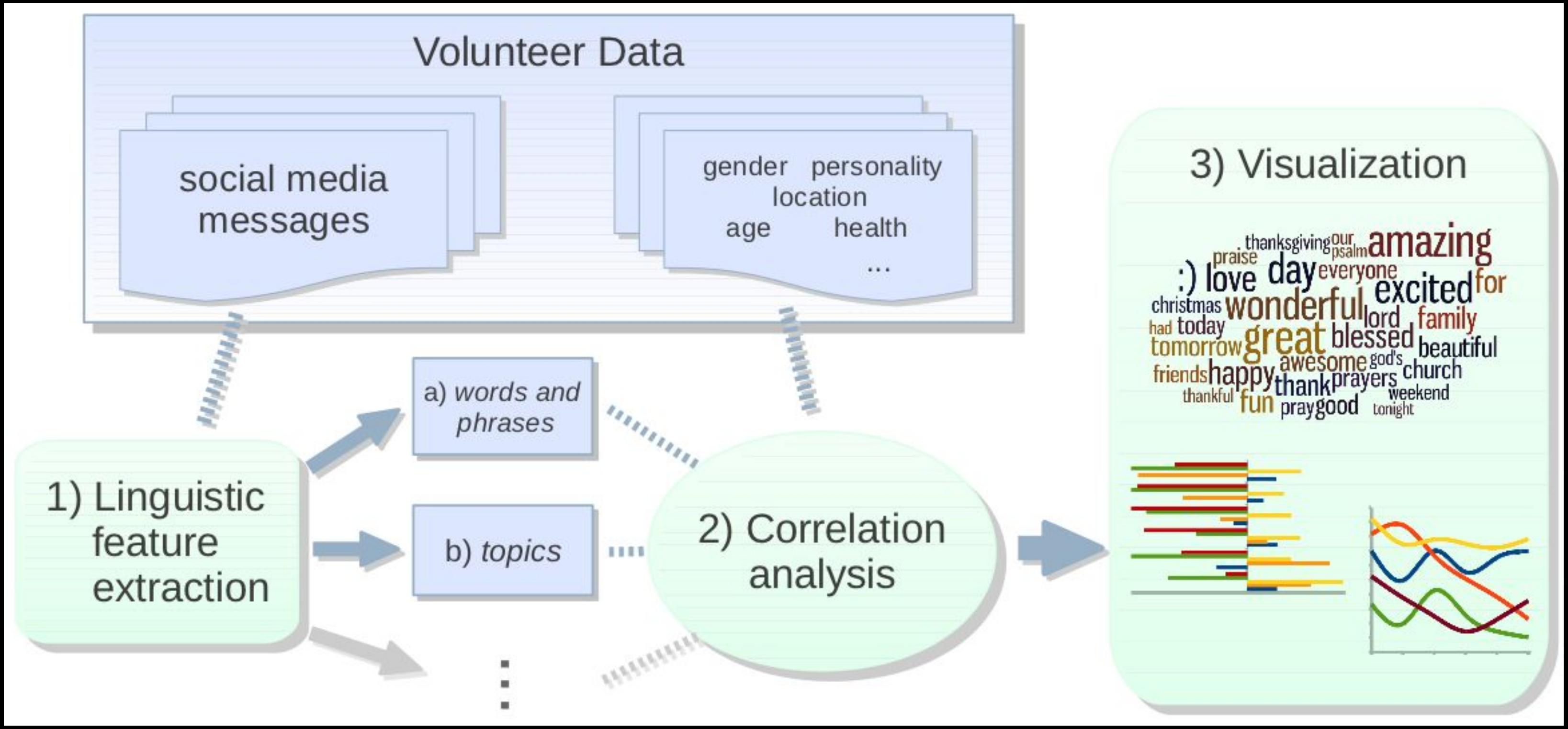
- Linguistic features
- Human or community attribute

Output:

- Features distinguishing attribute

Goal: Data-driven insights about an attribute

Differential Language Analysis


Volunteer Data
social media messages
gender personality location age health ...
1) Linguistic feature extraction
a) words and phrases
b) topics
...
2) Correlation analysis
3) Visualization
thanksgiving our psalm amazing praise :) love day everyone excited for christmas wonderful lord family had today tomorrow great blessed beautiful friends happy awesome god's church thank prayers thankful fun pray good weekend tonight

Differential Language Analysis

Methods of Correlation Analysis:

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

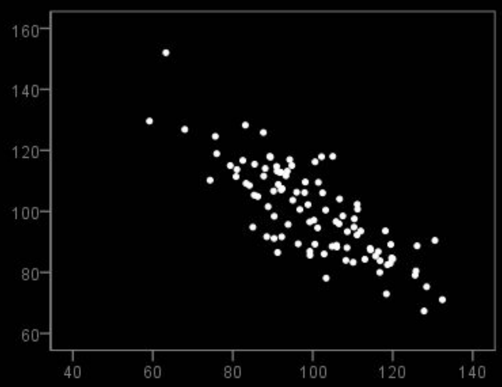
Differential Language Analysis

Methods of Correlation Analysis:

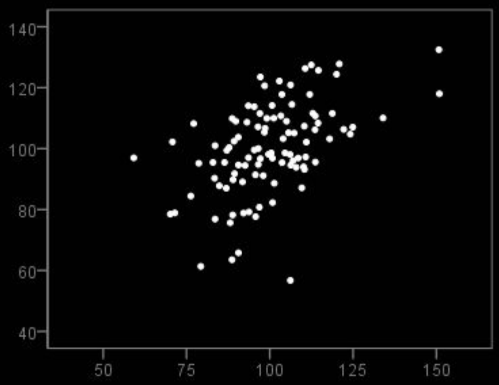
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

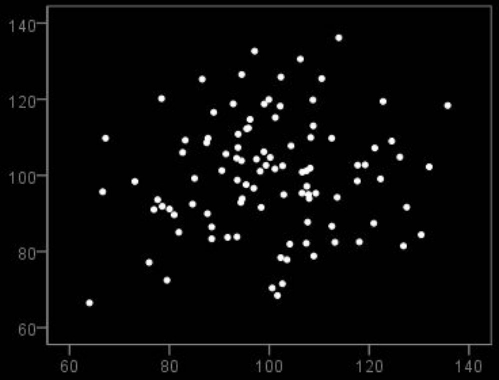
  Limitation: Doesn't handle controls



r = -0.8



r = 0.5



r = 0.1

Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

- Standardized Multivariate Linear Regression

  Fit the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

- **Standardized** Multivariate Linear Regression

  Fit the model:

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

  Adjust all variables to have “mean center” and “unit variance”:

Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

- **<u>Standardized</u>** Multivariate Linear Regression

  Fit the model:

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

  Adjust all variables to have “mean center” and “unit variance”:

  $$z = \frac{x - \mu}{\sigma}$$

  $\mu =$ Mean

  $\sigma =$ Standard Deviation

Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

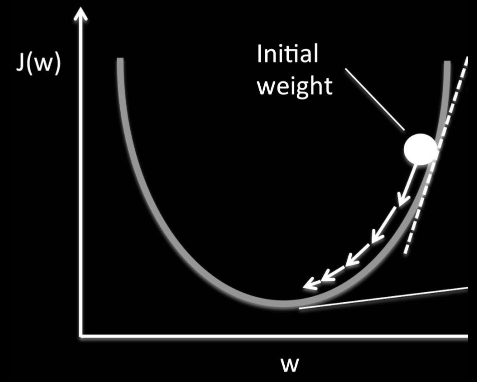
- Standardized Multivariate Linear Regression

  Fit the model:

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

  Option 1: Gradient Descent:

  $J = \sum (y - \hat{y})^2$ -- “Sum of Squares” Error



Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

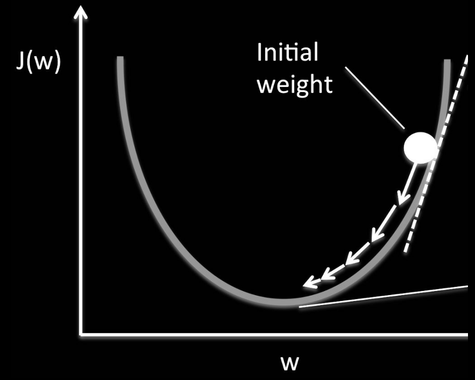
- Standardized Multivariate Linear Regression

  Fit the model: $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

  Option 1: Gradient Descent:

  $J = \sum (y - \hat{y})^2$ -- “Sum of Squares” Error

  Option 2: Matrix model: $$Y = X\beta + \epsilon$$



Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn't handle controls

- Standardized Multivariate Linear Regression

  Fit the model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$

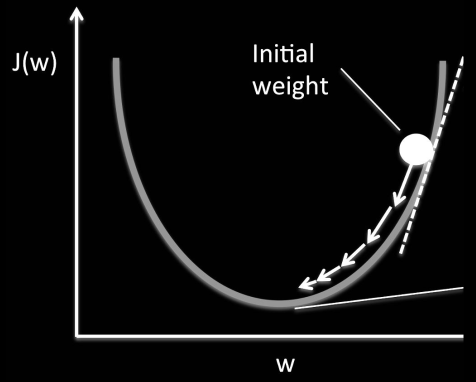
  Option 1: Gradient Descent:

  $J = \sum (y - \hat{y})^2$ -- "Sum of Squares" Error

  Option 2: Matrix model: $Y = X\beta + \epsilon$

  Matrix Computation Solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$



Differential Language Analysis

Methods of Correlation Analysis:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

  Limitation: Doesn’t handle controls

- Standardized Multivariate Linear Regression

  Fit the model: $$Y_i = \beta_0 + \boxed{\beta_1} X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$

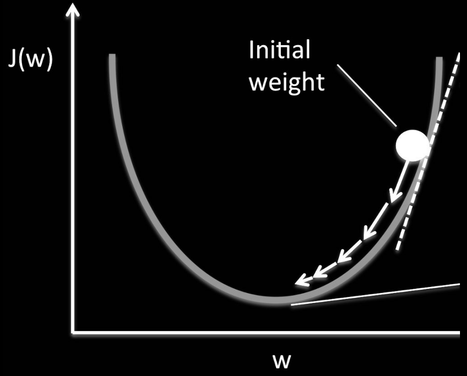
  Option 1: Gradient Descent:

  $J = \sum (y - \hat{y})^2$ -- “Sum of Squares” Error

  Option 2: Matrix model: $$Y = X\beta + \epsilon$$

  Matrix Computation Solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$



Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables

- Odds Ratio

$$\frac{\dfrac{\dfrac{countA(\text{"}horrible\text{"})}{NA}}{1-\dfrac{countA(\text{"}horrible\text{"})}{NA}}}{\dfrac{\dfrac{countB(\text{"}horrible\text{"})}{NB}}{1-\dfrac{countB(\text{"}horrible\text{"})}{NB}}}$$

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables

- Odds Ratio

$$\frac{\dfrac{\dfrac{countA(\text{"horrible"})}{NA}}{1-\dfrac{countA(\text{"horrible"})}{NA}}}{\dfrac{\dfrac{countB(\text{"horrible"})}{NB}}{1-\dfrac{countB(\text{"horrible"})}{NB}}} \propto log\left(\frac{\dfrac{countA(\text{"horrible"})}{NA}}{1-\dfrac{countA(\text{"horrible"})}{NA}}\right) - log\left(\frac{\dfrac{countB(\text{"horrible"})}{NB}}{1-\dfrac{countB(\text{"horrible"})}{NB}}\right)$$

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables

- Odds Ratio

$$\frac{\dfrac{\frac{countA(\text{"horrible"})}{NA}}{1-\frac{countA(\text{"horrible"})}{NA}}}{\dfrac{\frac{countB(\text{"horrible"})}{NB}}{1-\frac{countB(\text{"horrible"})}{NB}}} \propto log\left(\frac{\frac{countA(\text{"horrible"})}{NA}}{1-\frac{countA(\text{"horrible"})}{NA}}\right) - log\left(\frac{\frac{countB(\text{"horrible"})}{NB}}{1-\frac{countB(\text{"horrible"})}{NB}}\right)$$

$$= log\left(\frac{countA(\text{"horrible"})}{NA-countA(\text{"horrible"})}\right) - log\left(\frac{countB(\text{"horrible"})}{NB-countB(\text{"horrible"})}\right)$$

Differential Language Analysis

$$log\left(\frac{countA(\text{“horrible”})}{NA-countA(\text{“horrible”})}\right)-log\left(\frac{countB(\text{“horrible”})}{NB-countB(\text{“horrible”})}\right)$$

- Odds Ratio using Informative Dirichlet Prior

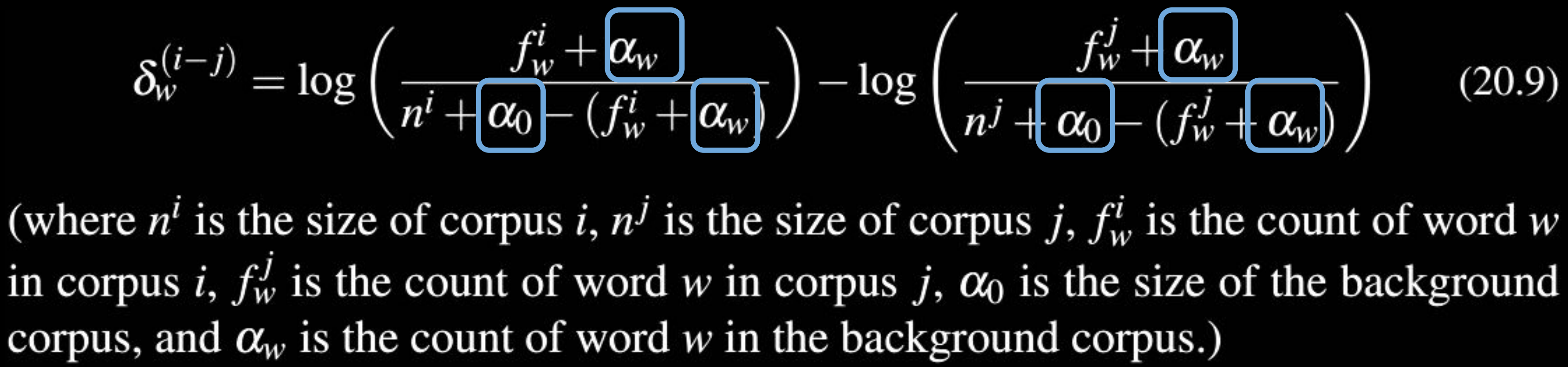
$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i+\alpha_w}{n^i+\alpha_0-(f_w^i+\alpha_w)}\right)-\log\left(\frac{f_w^j+\alpha_w}{n^j+\alpha_0-(f_w^j+\alpha_w)}\right) \qquad (20.9)$$

(Monroe et al., 2010; Jurafsky, 2017)

Differential Language Analysis

$$log\left(\frac{countA(\text{"horrible"})}{NA-countA(\text{"horrible"})}\right)-log\left(\frac{countB(\text{"horrible"})}{NB-countB(\text{"horrible"})}\right)$$

- Odds Ratio using **Informative Dirichlet Prior**



$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (f_w^i + \boxed{\alpha_w})}\right) - \log\left(\frac{f_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (f_w^j + \boxed{\alpha_w})}\right) \qquad (20.9)$$

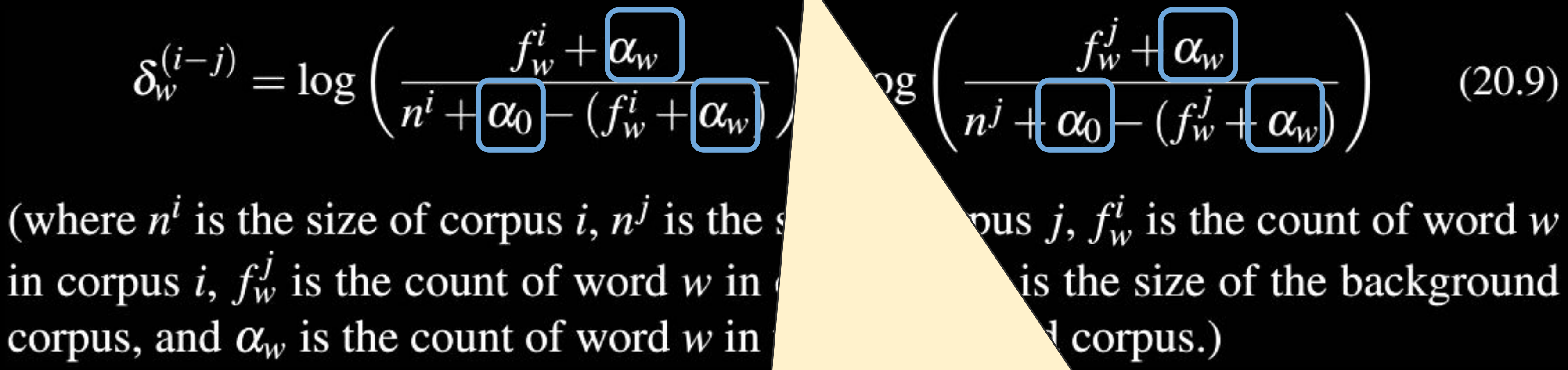
(where n^i is the size of corpus i, n^j is the size of corpus j, f_w^i is the count of word w in corpus i, f_w^j is the count of word w in corpus j, α_0 is the size of the background corpus, and α_w is the count of word w in the background corpus.)

(Monroe et al., 2010; Jurafsky, 2017)

Differential Language Analysis

$$log\left(\frac{countA(\text{“horrible”})}{NA-countA(\text{“horrible”})}\right) - log\left(\frac{countB(\text{“horrible”})}{NB-countB(\text{“horrible”})}\right)$$

- Odds Ratio using **Informative Dirichlet Prior**



$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (f_w^i + \boxed{\alpha_w})}\right) \quad \text{og}\left(\frac{f_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (f_w^j + \boxed{\alpha_w})}\right) \qquad (20.9)$$

(where n^i is the size of corpus i, n^j is the s pus j, f_w^i is the count of word w in corpus i, f_w^j is the count of word w in c is the size of the background corpus, and α_w is the count of word w in d corpus.)

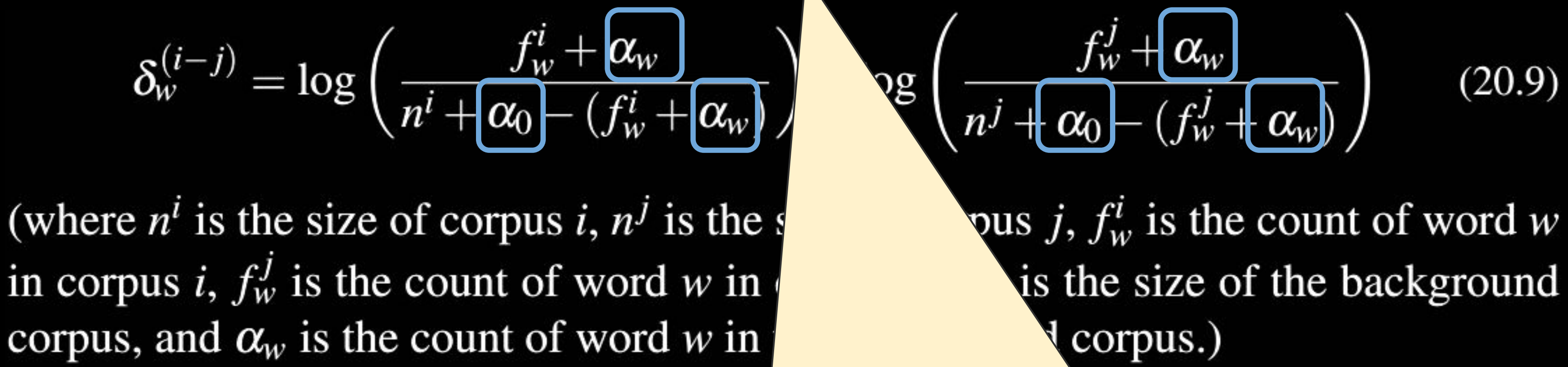
Bayesian term for “smoothing”: accounts for uncertainty as a function of event frequency (i.e. words observed less) by integrating “prior” beliefs mathematically.

(Mo

Differential Language Analysis

$$log\left(\frac{countA(\text{"horrible"})}{NA - countA(\text{"horrible"})}\right) - log\left(\frac{countB(\text{"horrible"})}{NB - countB(\text{"horrible"})}\right)$$

- Odds Ratio using **Informative Dirichlet Prior**



$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (f_w^i + \boxed{\alpha_w})}\right) - \log\left(\frac{f_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (f_w^j + \boxed{\alpha_w})}\right) \quad (20.9)$$

(where n^i is the size of corpus i, n^j is the s pus j, f_w^i is the count of word w in corpus i, f_w^j is the count of word w in is the size of the background corpus, and α_w is the count of word w in l corpus.)

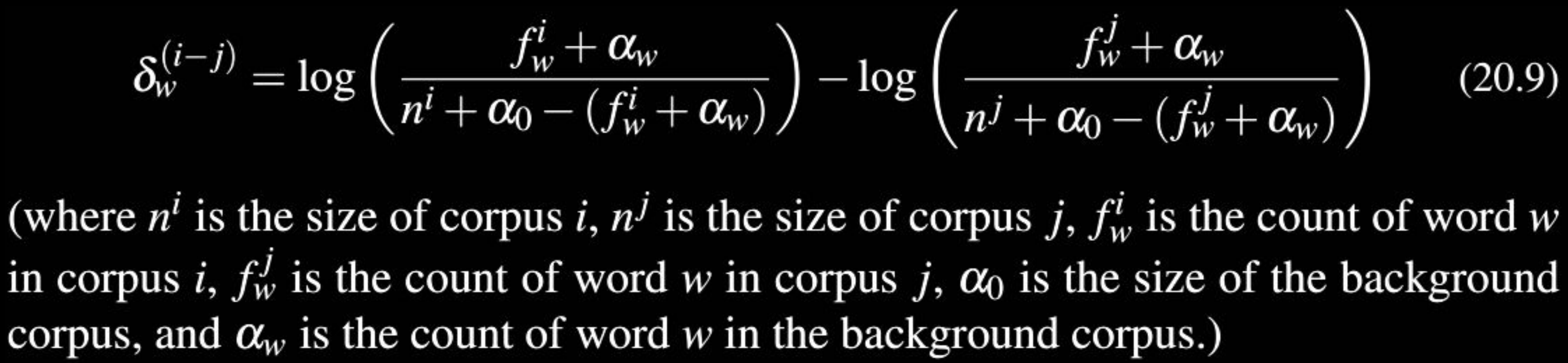
Bayesian term for “smoothing”: accounts for uncertainty as a function of event frequency (i.e. words observed less) by integrating “prior” beliefs mathematically.
“Informative”: the prior is based on past evidence. Here, the total frequency of the word.

(Mc

Differential Language Analysis

$$log\left(\frac{countA(\text{"horrible"})}{NA - countA(\text{"horrible"})}\right) - log\left(\frac{countB(\text{"horrible"})}{NB - countB(\text{"horrible"})}\right)$$

- Odds Ratio using Informative Dirichlet Prior



$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i + \alpha_w}{n^i + \alpha_0 - (f_w^i + \alpha_w)}\right) - \log\left(\frac{f_w^j + \alpha_w}{n^j + \alpha_0 - (f_w^j + \alpha_w)}\right) \qquad (20.9)$$

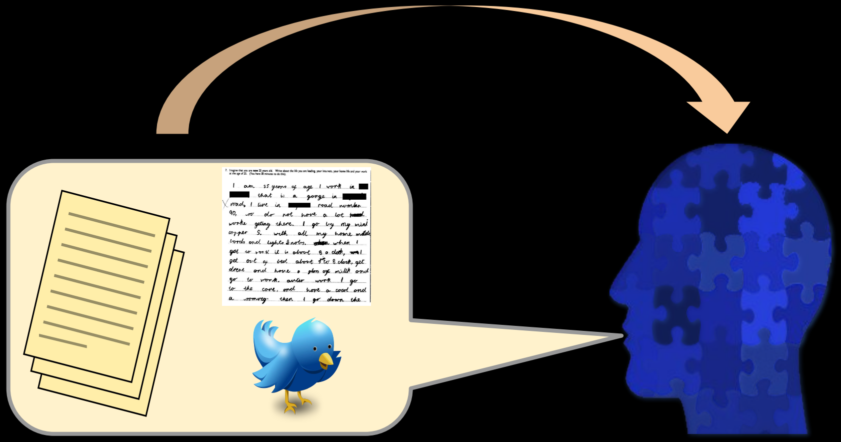
(where n^i is the size of corpus i, n^j is the size of corpus j, f_w^i is the count of word w in corpus i, f_w^j is the count of word w in corpus j, α_0 is the size of the background corpus, and α_w is the count of word w in the background corpus.)

Final score is standardized (z-scored): $\dfrac{\hat{\delta}_w^{(i-j)}}{\sqrt{\sigma^2\left(\hat{\delta}_w^{(i-j)}\right)}}$, where

$$\sigma^2\left(\hat{\delta}_w^{(i-j)}\right) \approx \frac{1}{f_w^i + \alpha_w} + \frac{1}{f_w^j + \alpha_w}$$

(Monroe et al., 2010; Jurafsky, 2017)

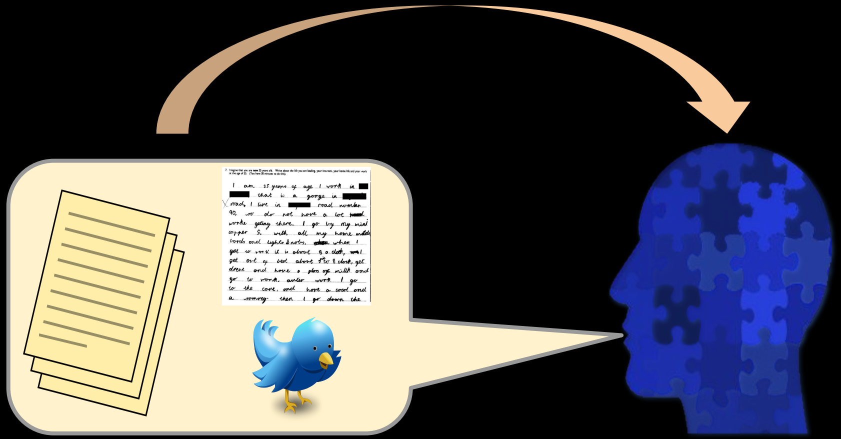
Natural language is generated by *people*.



7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted]
[redacted] that is a gorge in [redacted]
road, I live in [redacted] road number
90, we do not have a lot
worke geting there. I go by my mini
copper S. with all my home madde
bords and lights & nobs. when I
get to work it is about 8 o clock, I
get out of bed about 5 to 8 clock, get
drese and have a glass of millk and
go to work. after work I go
to the cove, and have a cool and
a somveg. then I go down the

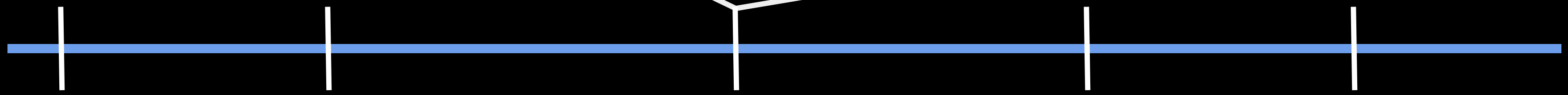
Natural language is generated by people.



7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 years of age I work in [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot worke geting there. I go by my mini copper S. with all my home madde lords and lights & nobs. when I got to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get dresst and have a glass of milk and go to work. after work I go to the cove, and have a cool and a somveg. then I go down the

"The common misconception is that language has got to do with words and what they mean. It does not. It has to do with people and what they mean."



Shannon, 1948

Mosteller & Wallace 1963

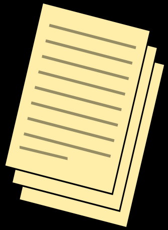
Clark & Schober, 1992

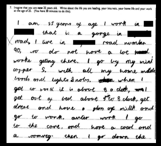
Mairesse, Walker, et al., 2007

Hovy & Soogaard, 2015

Natural language is generated by people.

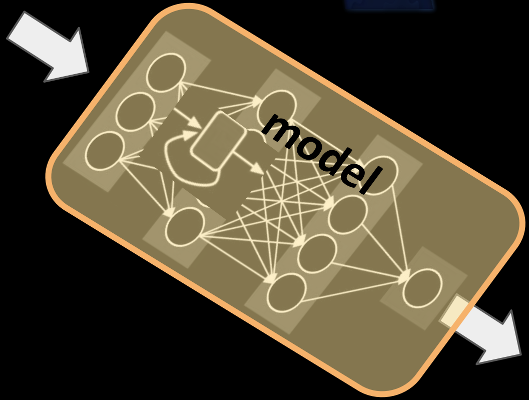
Yet, our models:

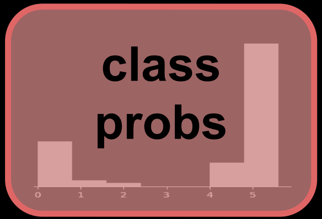




7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

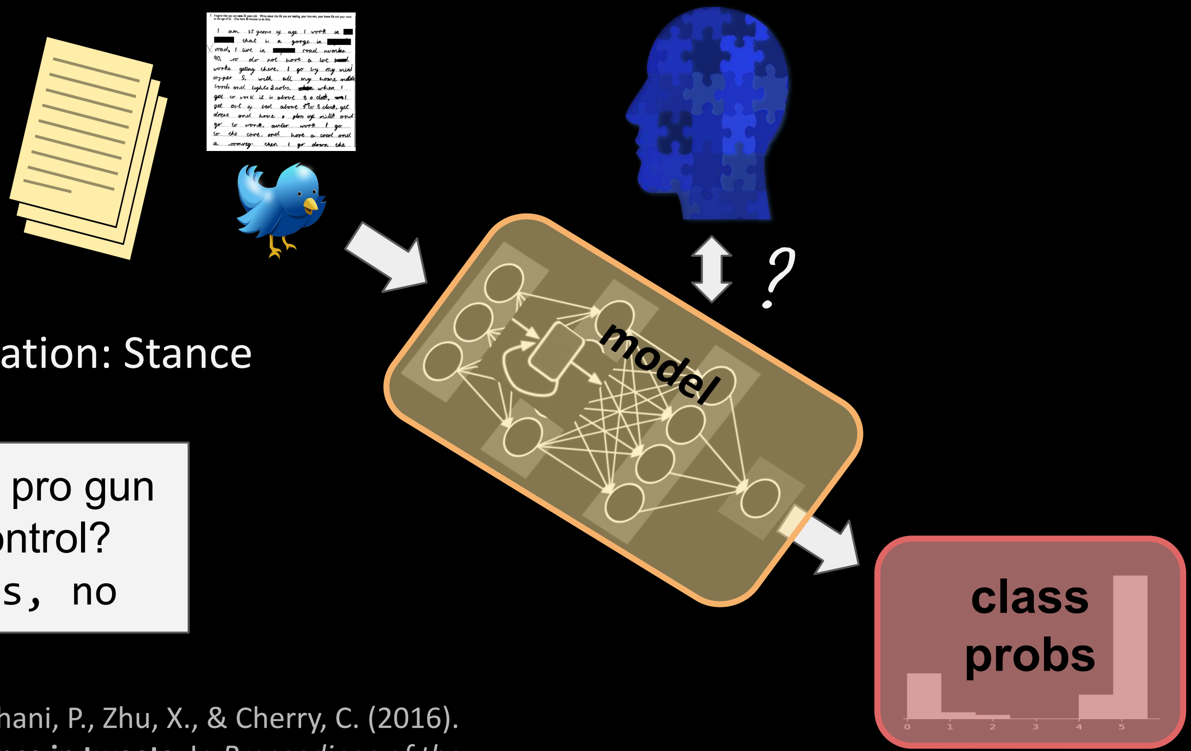
I am 25 years of age I work in [redacted] [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not have a lot works geting there. I go by my mini copper S. with all my home madde bords and lights & nobs. when I get to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drese and have a glass of millk and go to work. aster work I go to the cafe, and have a cool and a somveg. then I go down the





Natural language is generated by people.

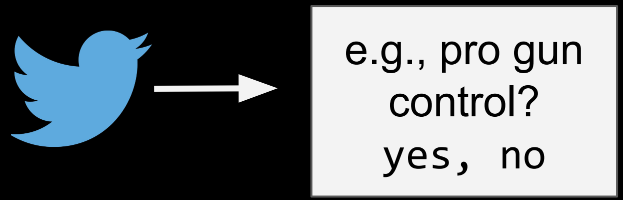
Yet, our models:



7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

I am 25 yeors of age I work in [redacted]
[redacted] that is a gorge in [redacted]
road, I live in [redacted] road number
90, we do not have a lot
worke geting there. I go by my mini
copper S. with all my home madde
bords and lights & nobs. when I
get to work it is about 8 o clock, I
get out of bed about 5 to 8 clock, get
drese and have a glass of millk and
go to work. after work I go
to the cove, and have a cocol and
a somveg. then I go down the

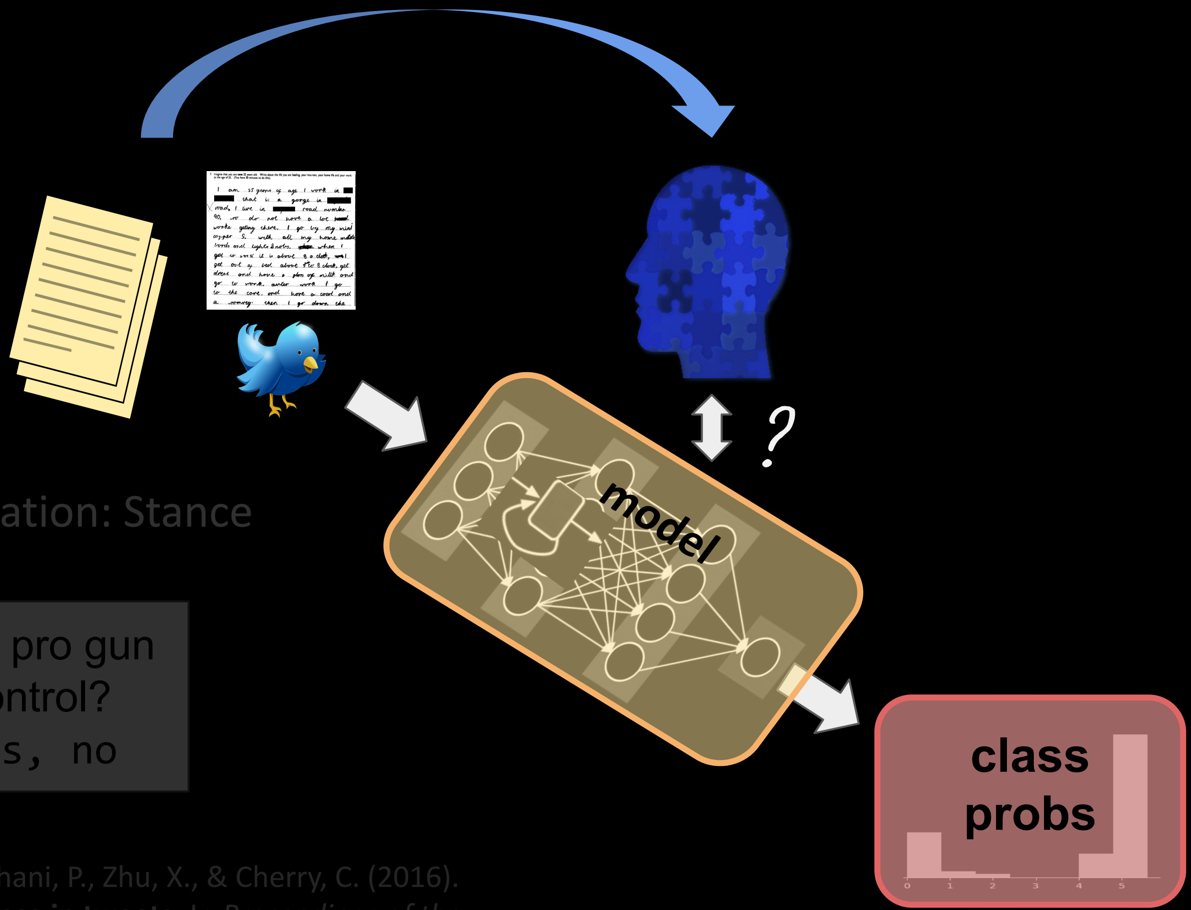
e.g. Document Classification: Stance



e.g., pro gun control?
yes, no

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). **Semeval-2016 task 6: Detecting stance in tweets.** In *Proceedings of the 10th International Workshop on Semantic Evaluation*.

Natural language is generated by people.



7. Imagine that you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to do this).

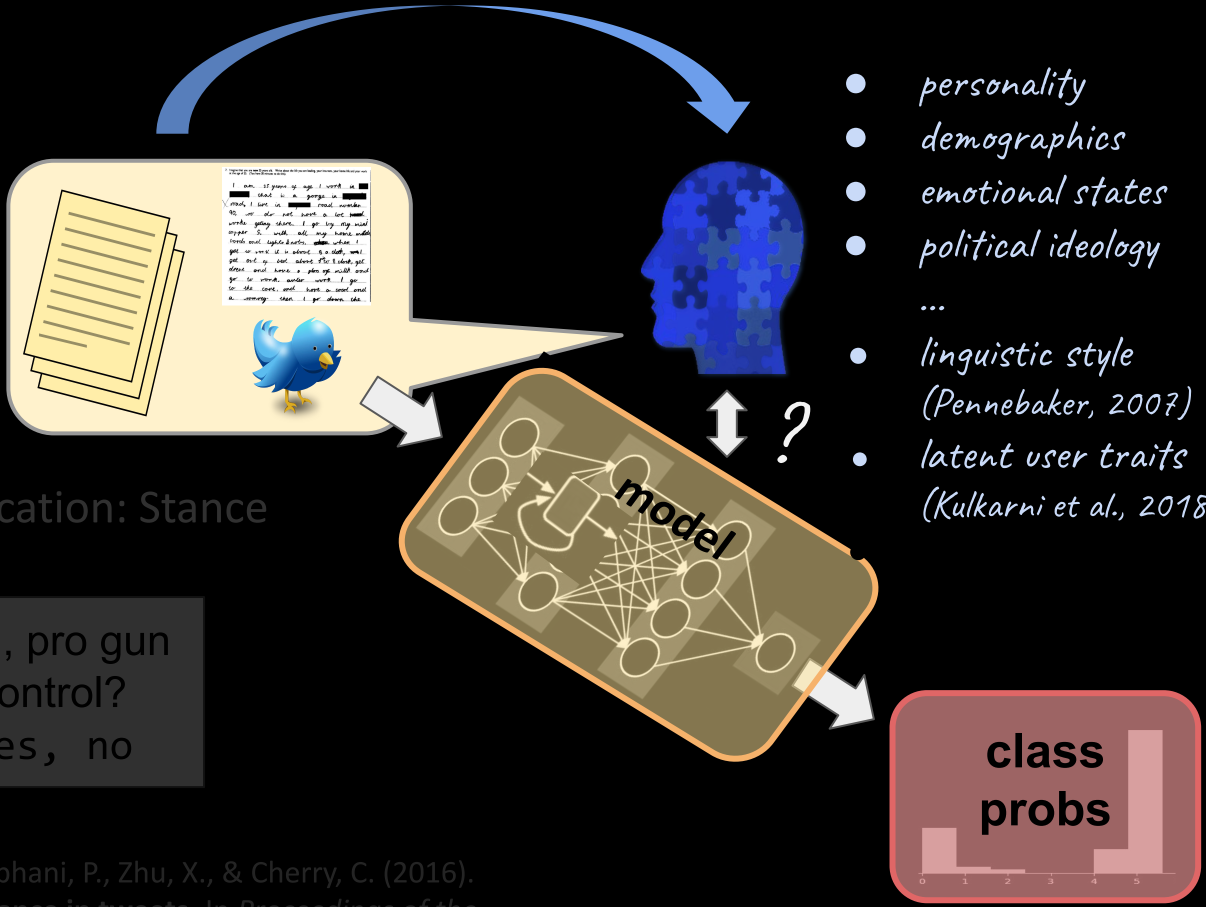
I am 25 yeors of age I work in [redacted] that is a gorge in [redacted] road, I live in [redacted] road number 90, we do not hove a lot worke geting there. I go by my mini copper S. with all my home madde words and lights & nobs. when I got to work it is about 8 o clock, I get out of bed about 5 to 8 clock, get drest and hove a glass of millk and go to work. after work I go to the cove, and hove a cocol and a somveg. then I go down the

e.g. Document Classification: Stance

e.g., pro gun control?
yes, no

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). **Semeval-2016 task 6: Detecting stance in tweets.** In *Proceedings of the 10th International Workshop on Semantic Evaluation*.

Natural language is generated by people.

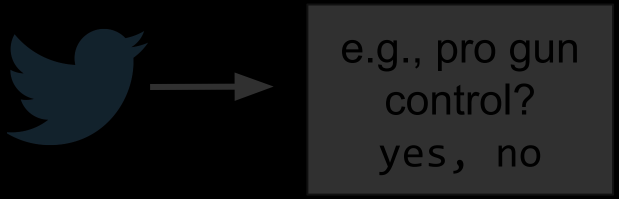


- personality
- demographics
- emotional states
- political ideology

...

- linguistic style (Pennebaker, 2007)
- latent user traits (Kulkarni et al., 2018)

e.g. Document Classification: Stance

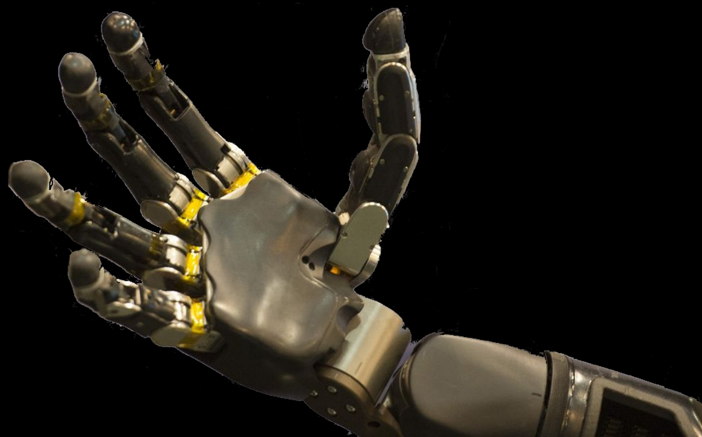


Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). **Semeval-2016 task 6: Detecting stance in tweets.** In *Proceedings of the 10th International Workshop on Semantic Evaluation.*

Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.



Natural language is generated by people.

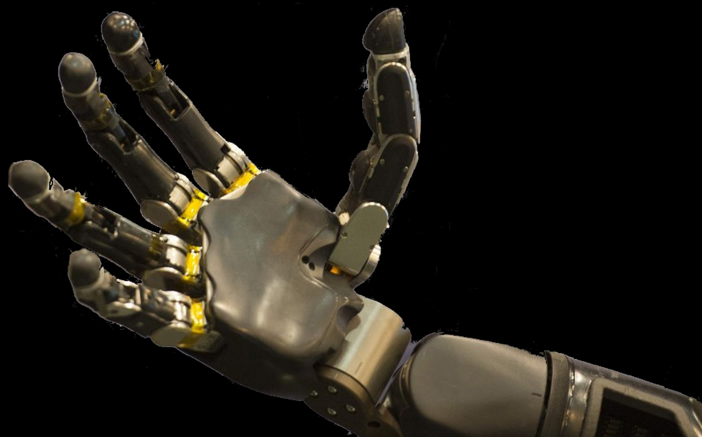
What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.

Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.



Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)

   (e.g. “sick” said from a young individual versus old individual)

2. Additive: Include direct effect of human factor on outcome.

   (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.

   (e.g. image captioner label pictures of men in kitchen as women)

Approaches to Human Factor Inclusion

What are human “factors”?

1. g if language to change depending
o called “compositional”)
ndividual versus old individual)

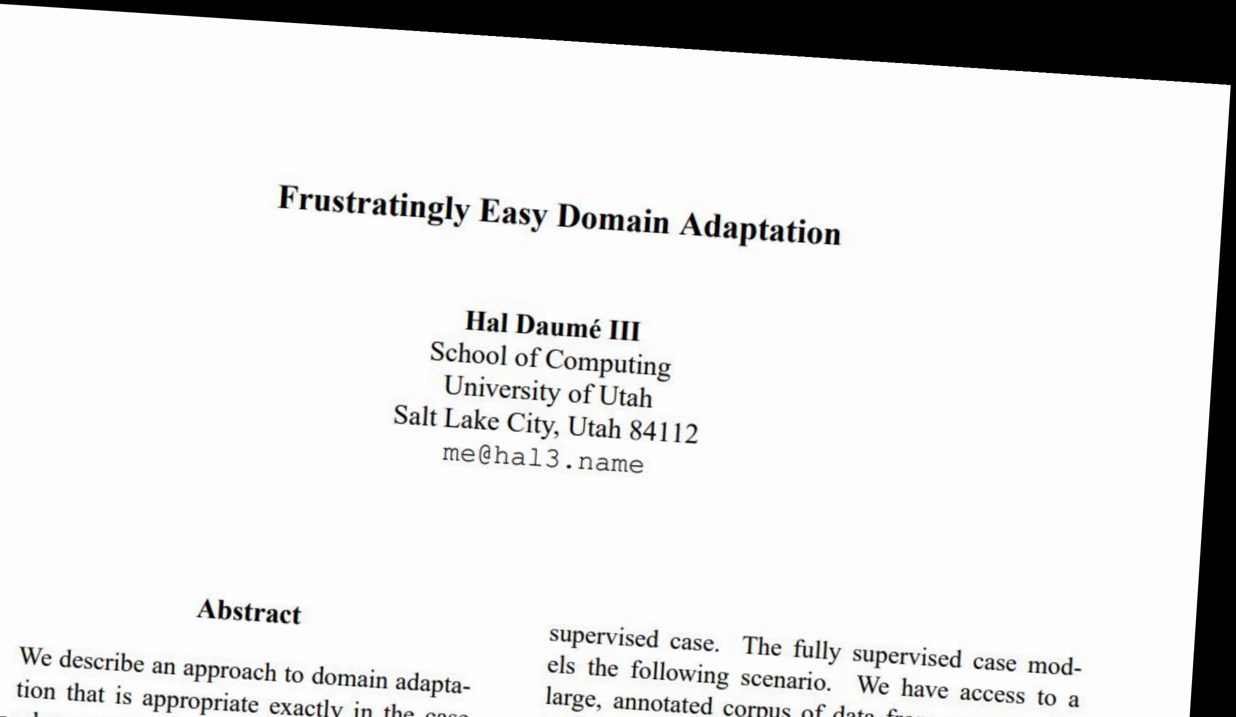
2. Additive: Include direct effect of human factor on outcome.
(e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.
(e.g. image captioner label pictures of men in kitchen as women)

Adaptation Approach: Domain Adaptation

Features for: source target

$$\Phi^s(x) = \langle x, x, 0\rangle, \quad \Phi^t(x) = \langle x, 0, x\rangle$$



Frustratingly Easy Domain Adaptation

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@hal3.name

Abstract

We describe an approach to domain adaptation that is appropriate exactly in the case

supervised case. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data

Adaptation Approach: Domain Adaptation

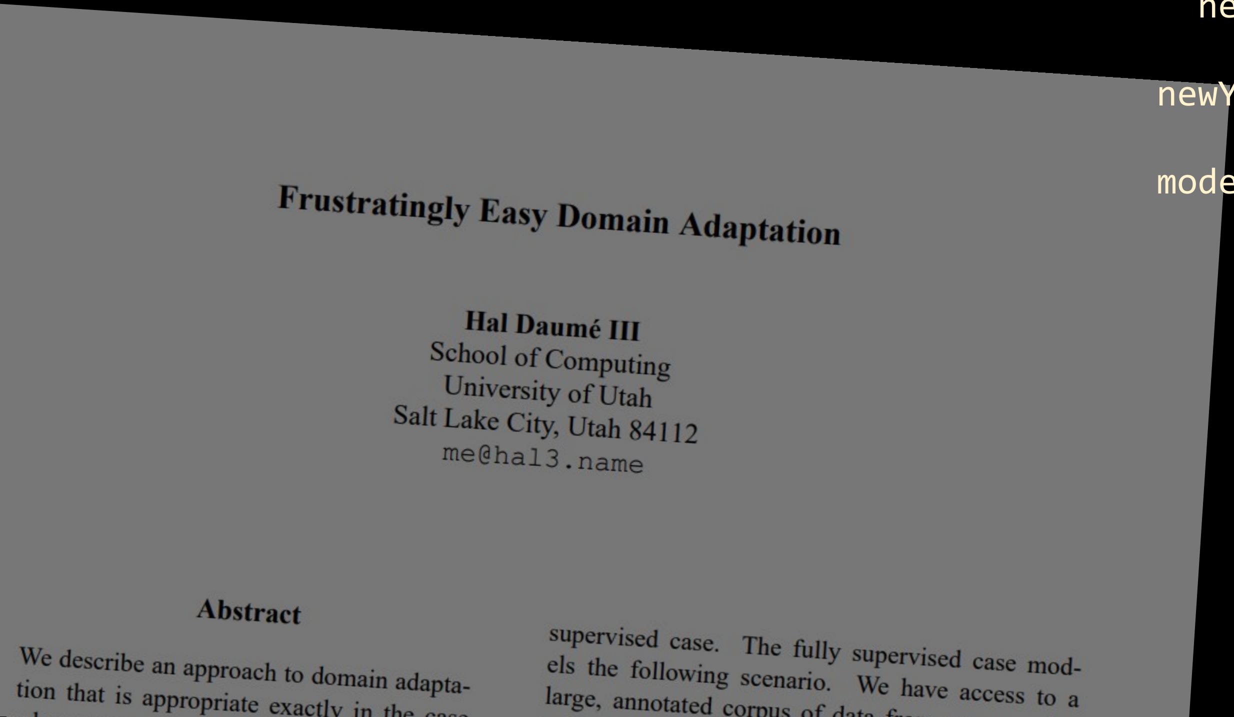
Features for: source target

$$\Phi^s(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{x}, \boldsymbol{0} \rangle, \quad \Phi^t(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{0}, \boldsymbol{x} \rangle$$

```
newX = []
for all x in source_x:
  newX.append(x + x + [0]*len(x))
for all x in target_x
  newX.append(x + [0]*len(x), x)

newY = source_y + target_y

model = model.train(newX,newY)
```



Frustratingly Easy Domain Adaptation

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@hal3.name

Abstract

We describe an approach to domain adaptation that is appropriate exactly in the case

supervised case. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data fr

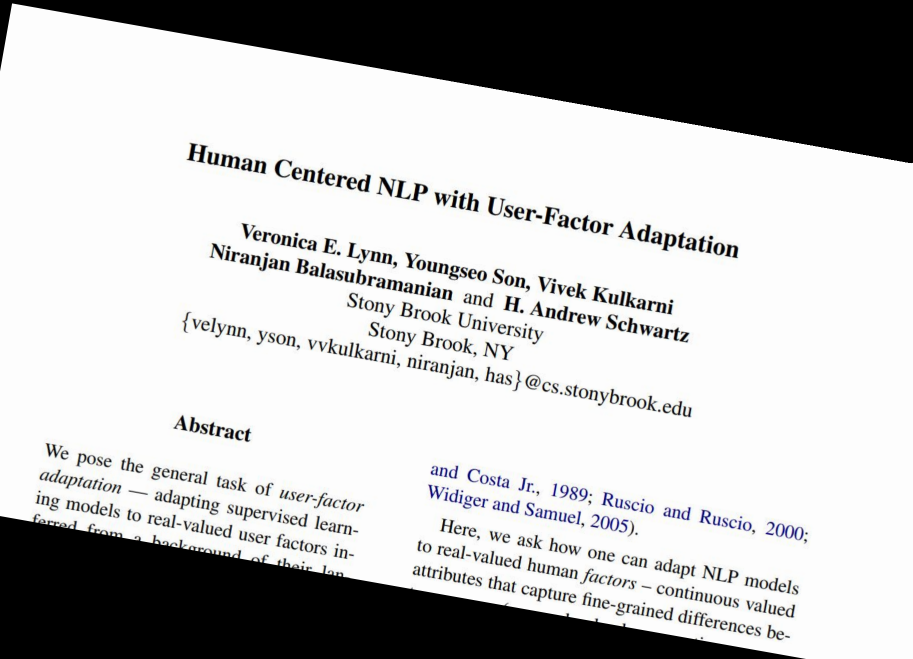
Human Factors

--- Any attribute, represented as a continuous or discrete variable, of the humans generating the natural language.

E.g.

- Gender
- Age
- Personality
- Ethnicity
- Socio-economic status

Adaptation Approach: Factor Adaptation



Human Centered NLP with User-Factor Adaptation

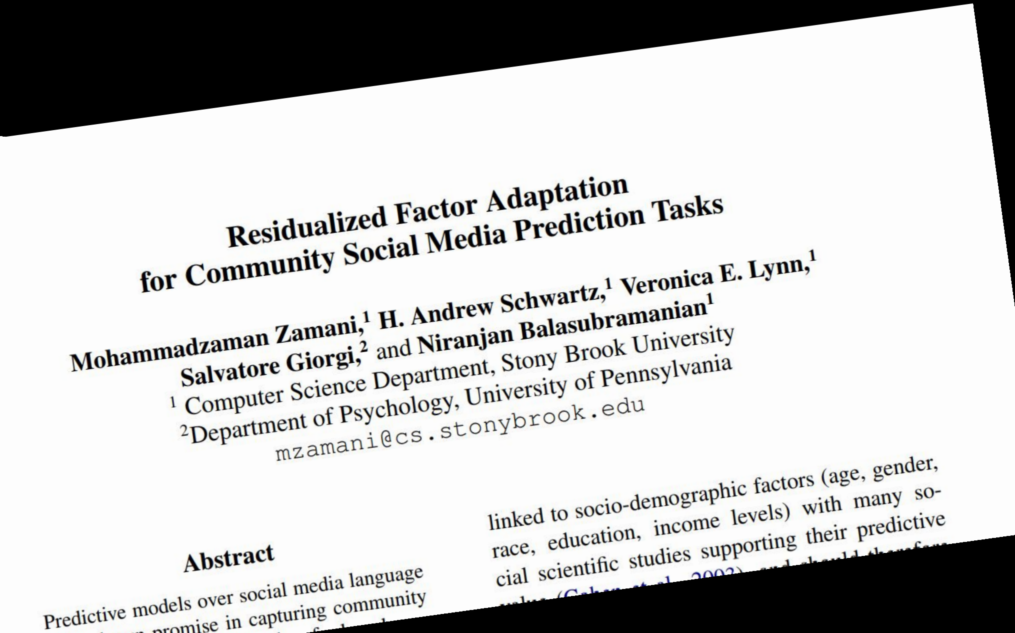
Veronica E. Lynn, Youngseo Son, Vivek Kulkarni
Niranjan Balasubramanian and **H. Andrew Schwartz**
Stony Brook University
Stony Brook, NY
{velynn, yson, vvkulkarni, niranjan, has}@cs.stonybrook.edu

Abstract

We pose the general task of *user-factor adaptation* — adapting supervised learning models to real-valued user factors inferred from a background of their lan-

and Costa Jr., 1989; Ruscio and Ruscio, 2000; Widiger and Samuel, 2005).

Here, we ask how one can adapt NLP models to real-valued human *factors* – continuous valued attributes that capture fine-grained differences be-



Residualized Factor Adaptation for Community Social Media Prediction Tasks

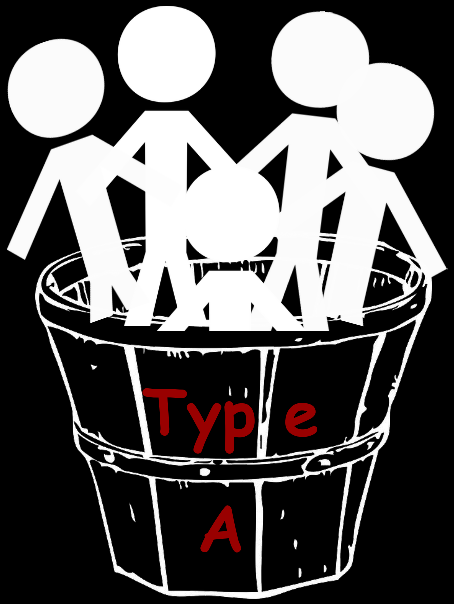
Mohammadzaman Zamani,[1] **H. Andrew Schwartz,**[1] **Veronica E. Lynn,**[1]
Salvatore Giorgi,[2] and **Niranjan Balasubramanian**[1]
[1] Computer Science Department, Stony Brook University
[2]Department of Psychology, University of Pennsylvania
mzamani@cs.stonybrook.edu

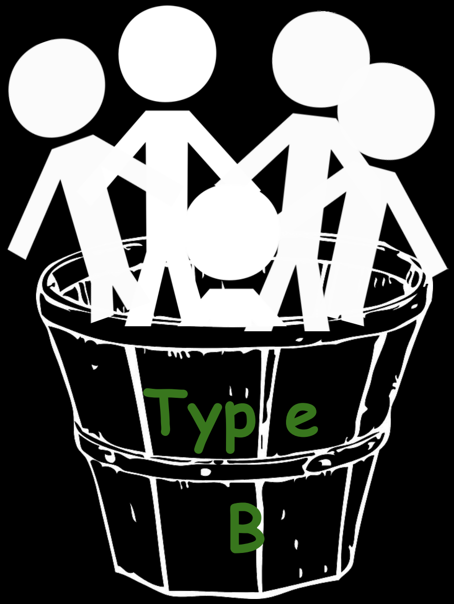
Abstract

Predictive models over social media language

linked to socio-demographic factors (age, gender, race, education, income levels) with many social scientific studies supporting their predictive

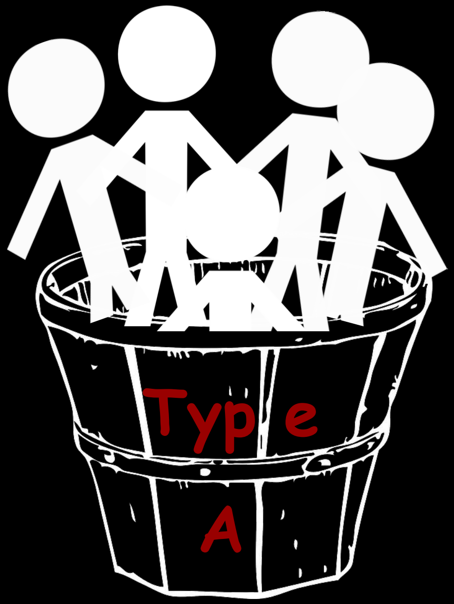
Adaptation

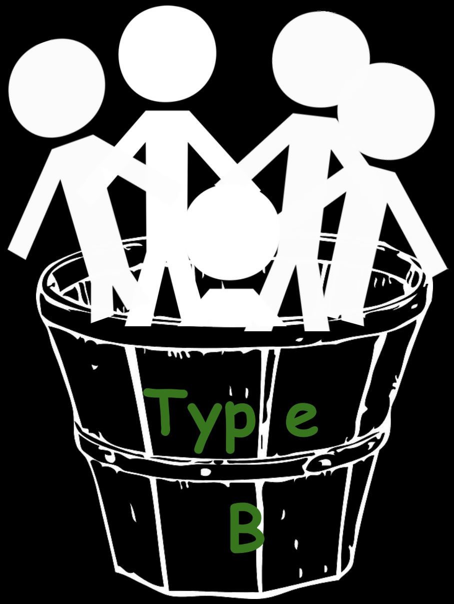




typically requires putting people into discrete bins *"most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]"*

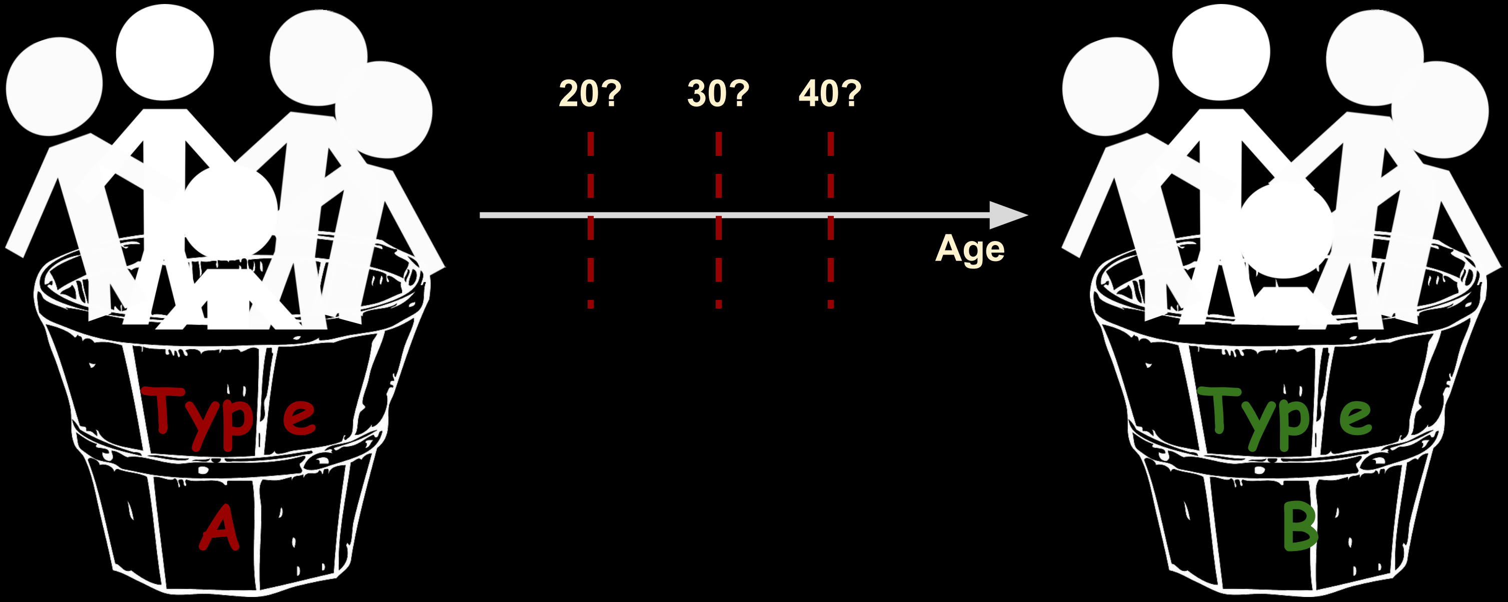
(Haslam et al., 2012)



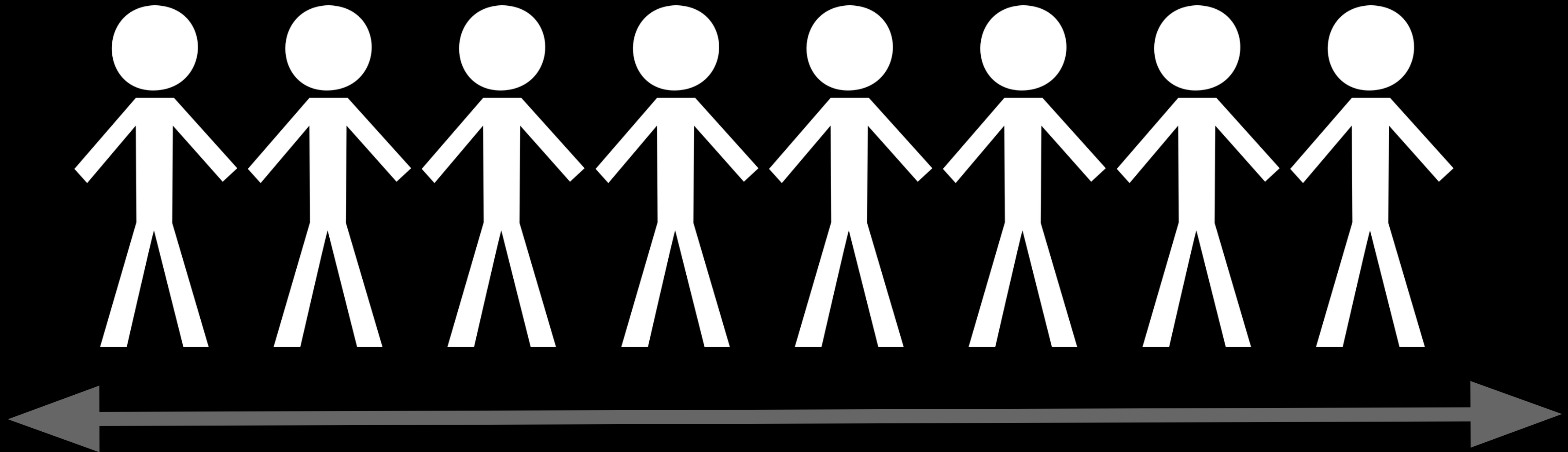


“most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]”

(Haslam et al., 2012)



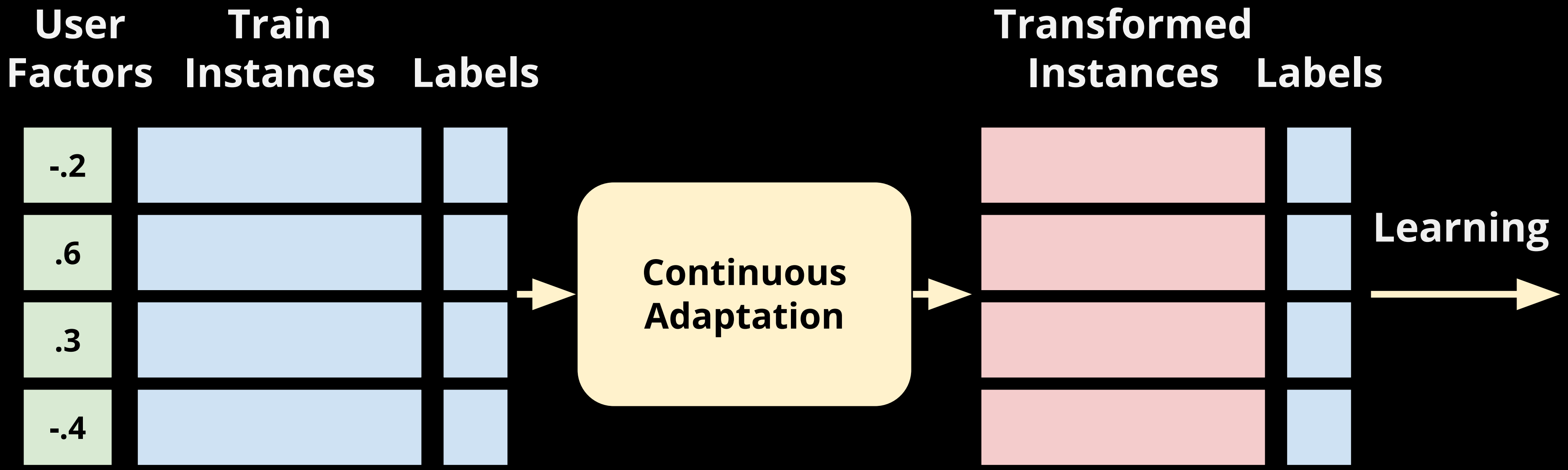
"most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]"
(Haslam et al., 2012)



Less *Factor* A

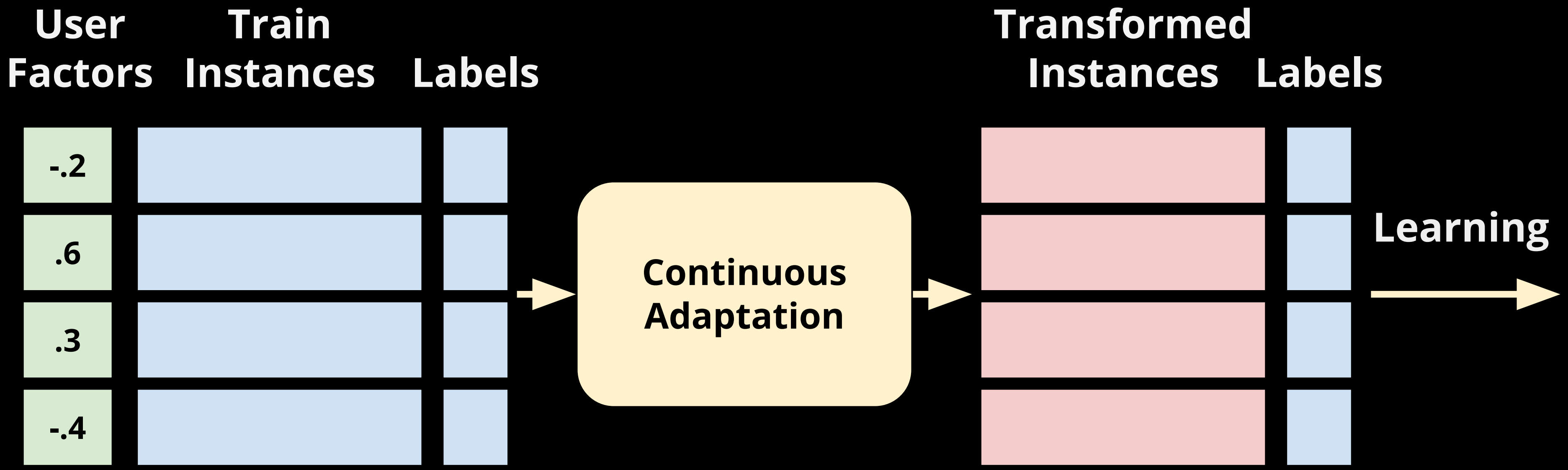
More *Factor* A

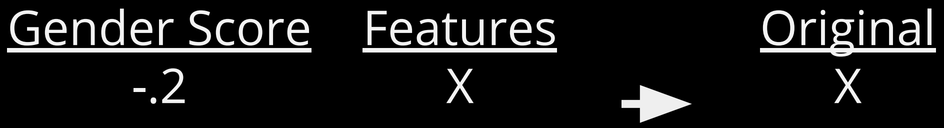
Our Method: Continuous Adaptation



(Lynn et al., 2017)

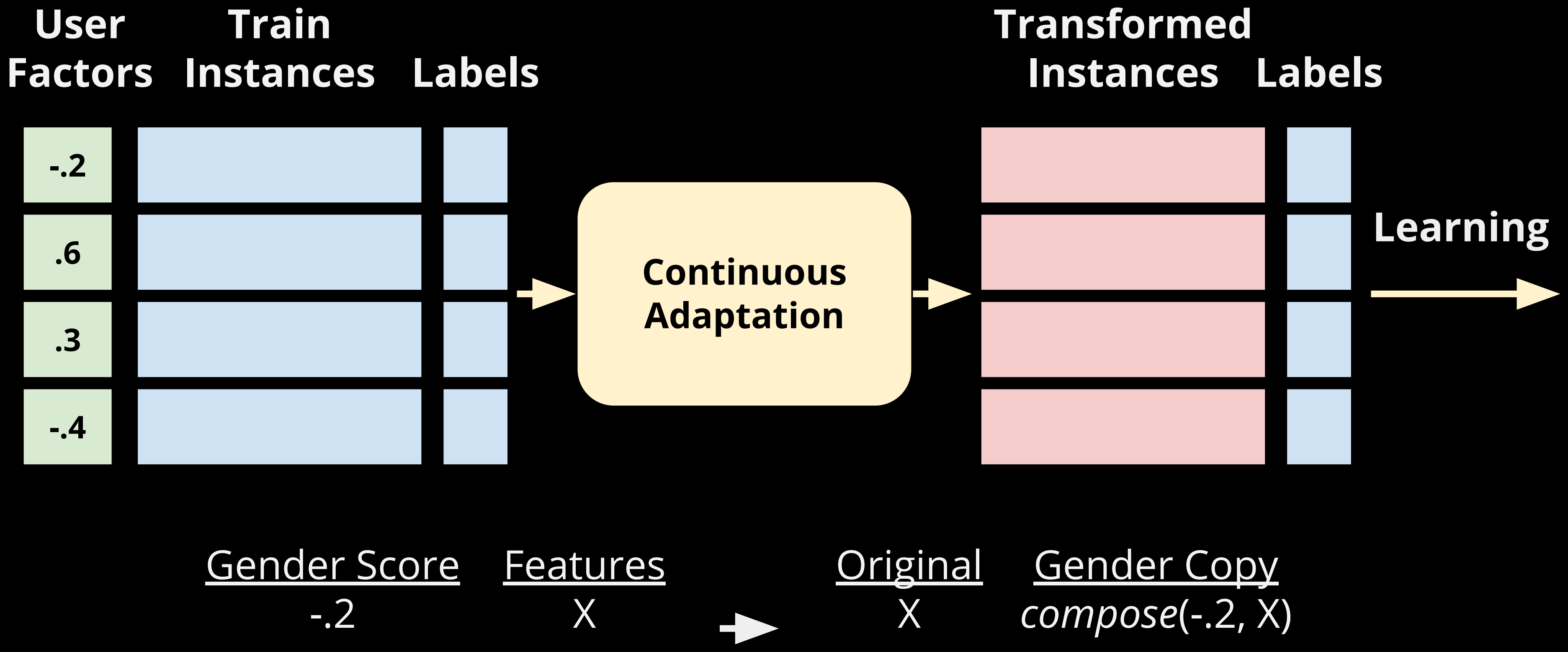
Our Method: Continuous Adaptation





(Lynn et al., 2017)

Our Method: Continuous Adaptation



User Factors

Train Instances

Labels

-.2

.6

.3

-.4

Continuous Adaptation

Transformed Instances

Labels

Learning

Gender Score
-.2

Features
X

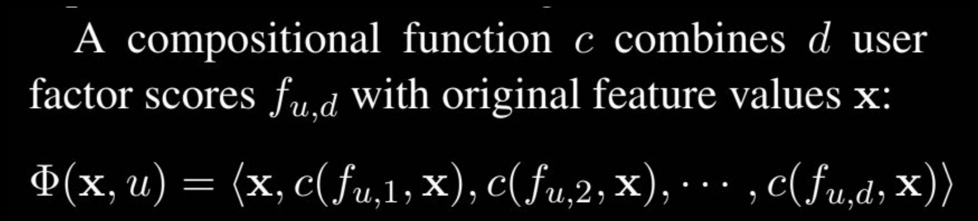
Original
X

Gender Copy
compose(-.2, X)

(Lynn et al., 2017)

User Factor Adaptation: Handling multiple factors

Replicate features for each factor:



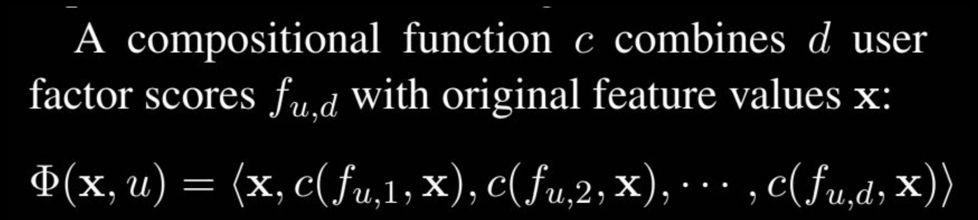
A compositional function c combines d user factor scores $f_{u,d}$ with original feature values $\mathbf{x}$:

$$\Phi(\mathbf{x}, u) = \langle \mathbf{x}, c(f_{u,1}, \mathbf{x}), c(f_{u,2}, \mathbf{x}), \cdots, c(f_{u,d}, \mathbf{x}) \rangle$$

(Lynn et al., 2017)

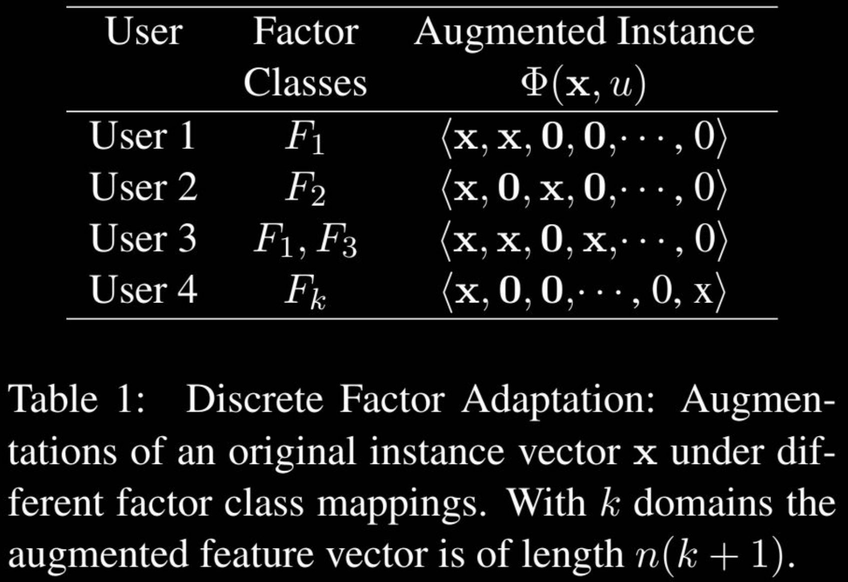
User Factor Adaptation: Handling multiple factors

Replicate features for each factor:



A compositional function c combines d user factor scores $f_{u,d}$ with original feature values $\mathbf{x}$:

$$\Phi(\mathbf{x}, u) = \langle \mathbf{x}, c(f_{u,1}, \mathbf{x}), c(f_{u,2}, \mathbf{x}), \cdots, c(f_{u,d}, \mathbf{x}) \rangle$$



User	Factor Classes	Augmented Instance $\Phi(\mathbf{x}, u)$
User 1	F_1	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, 0 \rangle$
User 2	F_2	$\langle \mathbf{x}, \mathbf{0}, \mathbf{x}, \mathbf{0}, \cdots, 0 \rangle$
User 3	F_1, F_3	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{x}, \cdots, 0 \rangle$
User 4	F_k	$\langle \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, 0, \mathrm{x} \rangle$

Table 1: Discrete Factor Adaptation: Augmentations of an original instance vector $\mathbf{x}$ under different factor class mappings. With k domains the augmented feature vector is of length $n(k + 1)$.

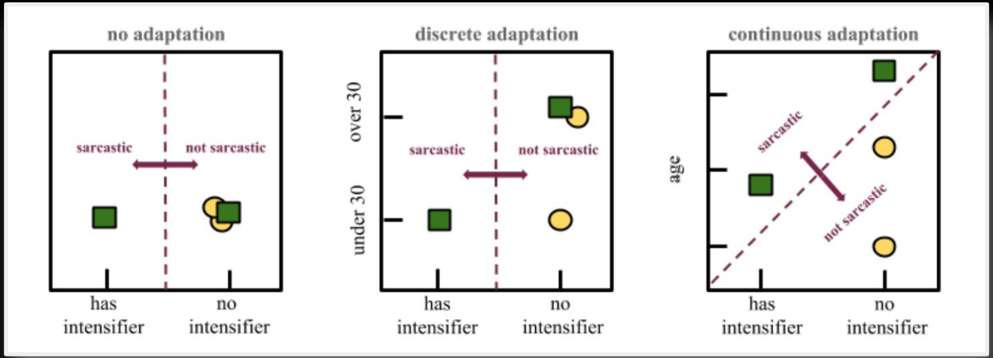
(Lynn et al., 2017)

User Factor Adaptation: Handling multiple factors

Replicate features for each factor:

A compositional function c combines d user factor scores $f_{u,d}$ with original feature values $\mathbf{x}$:

$$\Phi(\mathbf{x}, u) = \langle \mathbf{x}, c(f_{u,1}, \mathbf{x}), c(f_{u,2}, \mathbf{x}), \cdots, c(f_{u,d}, \mathbf{x}) \rangle$$



User	Factor Classes	Augmented Instance $\Phi(\mathbf{x}, u)$
User 1	F_1	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, 0 \rangle$
User 2	F_2	$\langle \mathbf{x}, \mathbf{0}, \mathbf{x}, \mathbf{0}, \cdots, 0 \rangle$
User 3	F_1, F_3	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{x}, \cdots, 0 \rangle$
User 4	F_k	$\langle \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, 0, \mathrm{x} \rangle$

Table 1: Discrete Factor Adaptation: Augmentations of an original instance vector $\mathbf{x}$ under different factor class mappings. With k domains the augmented feature vector is of length $n(k + 1)$.

(Lynn et al., 2017)

Main Results

Adaptation improves over unadapted baselines (Lynn et al., 2017)

Task	Metric	No Adaptation	Gender	Personality	Latent (User Embed)
Stance	F1	64.9	**65.1 (+0.2)**	**66.3 (+1.4)**	**67.9 (+3.0)**
Sarcasm	F1	73.9	**75.1 (+1.2)**	**75.6 (+1.7)**	**77.3 (+3.4)**
Sentiment	Acc.	60.6	**61.0 (+0.4)**	**61.2 (+0.6)**	**60.7 (+0.1)**
PP-Attach	Acc.	71.0	70.7 (-0.3)	70.2 (-0.8)	70.8 (-0.2)
POS	Acc.	91.7	**91.9 (+0.2)**	91.2 (-0.5)	90.9 (-0.8)

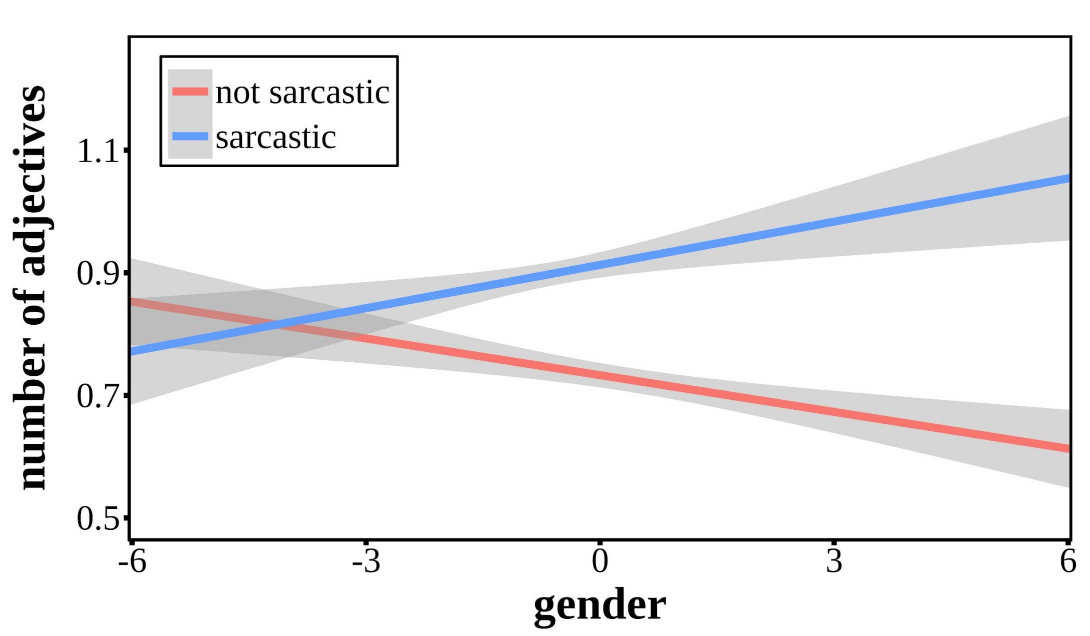
Example: How Adaptation Helps

Women

more adjectives→sarcasm

Men

more adjectives→no sarcasm



more “male”

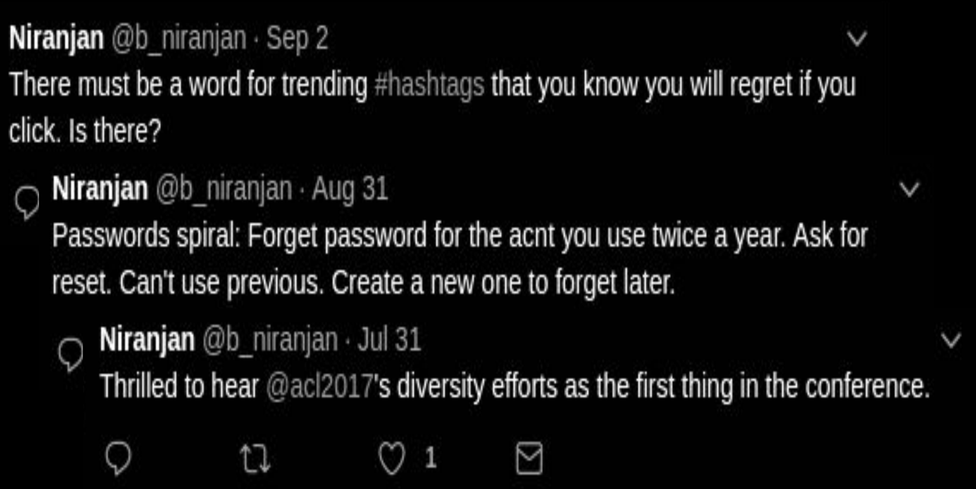
more “female”

Problem

User factors are not always available.

Solution: User Factor Inference

past tweets



Niranjan @b_niranjan · Sep 2
There must be a word for trending #hashtags that you know you will regret if you click. Is there?

Niranjan @b_niranjan · Aug 31
Passwords spiral: Forget password for the acnt you use twice a year. Ask for reset. Can't use previous. Create a new one to forget later.

Niranjan @b_niranjan · Jul 31
Thrilled to hear @acl2017's diversity efforts as the first thing in the conference.

→ **inferred factors**

Known

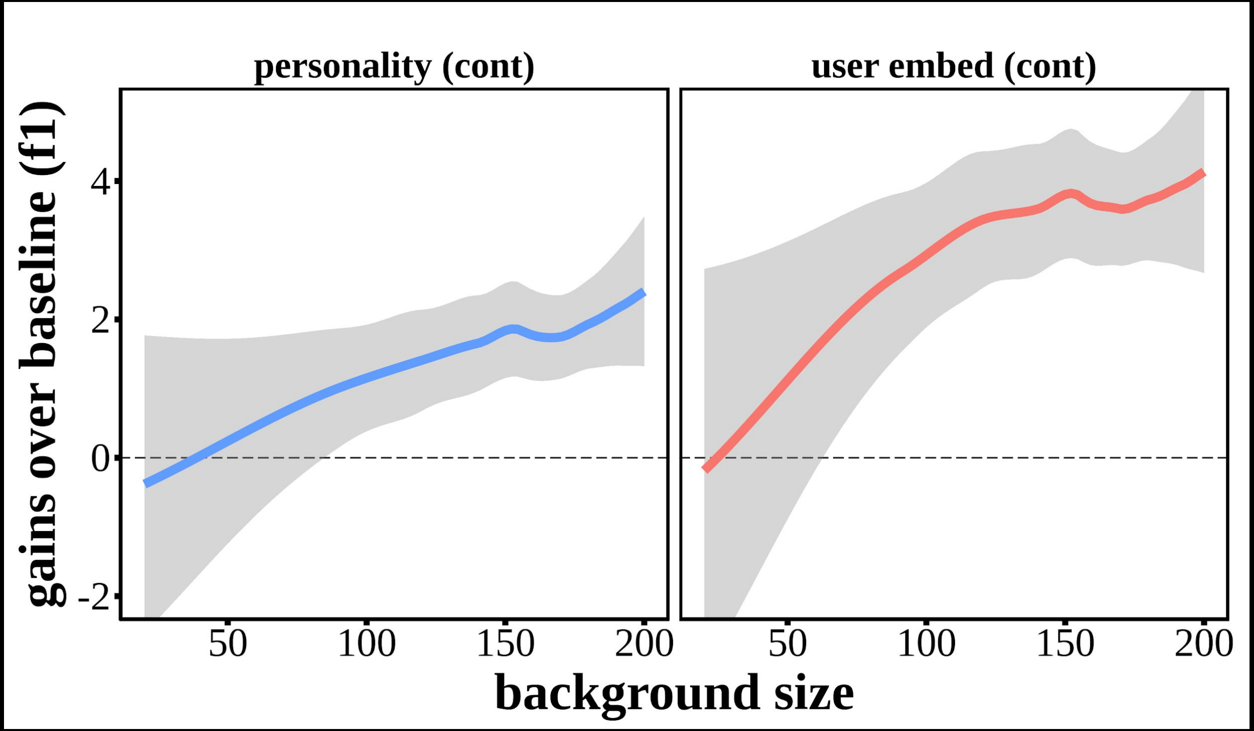
Age (Sap et al. 2014)
Gender (Sap et al. 2014)
Personality (Park et al. 2015)

Latent

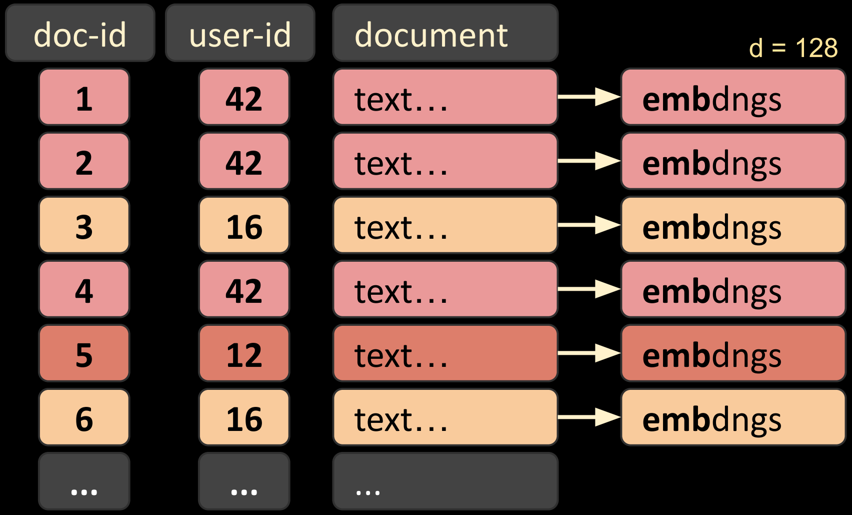
User Embeddings
(Kulkarni et al. 2017)
Word2Vec
TF-IDF

Background Size

Using more background tweets to infer factors produces larger gains



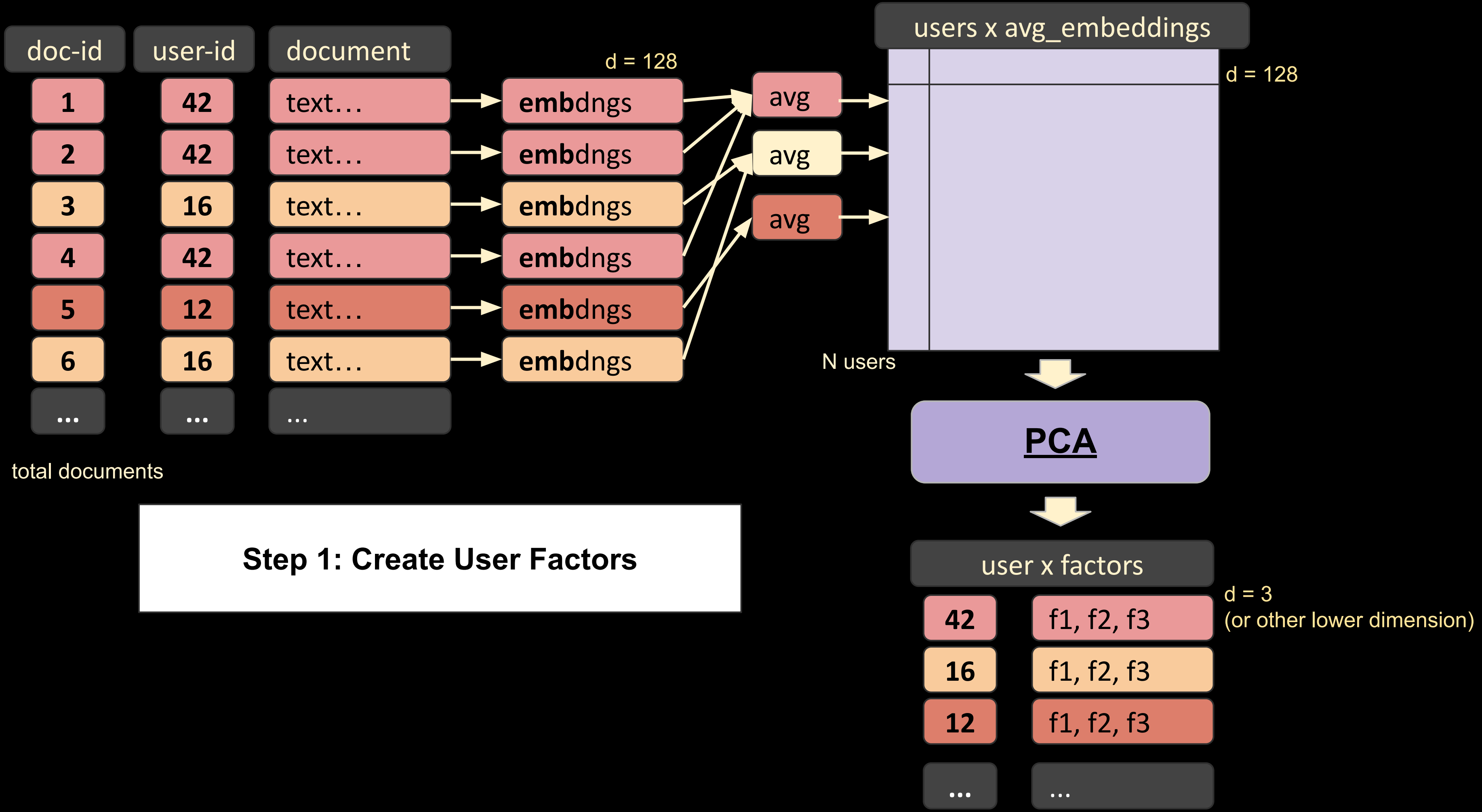
Full User Factors Adaptation Pipeline: with latent factors from training



doc-id	user-id	document		d = 128
1	42	text…	→	embdngs
2	42	text…	→	embdngs
3	16	text…	→	embdngs
4	42	text…	→	embdngs
5	12	text…	→	embdngs
6	16	text…	→	embdngs
...	...	…		

total documents

Full User Factors Adaptation Pipeline: with latent factors from training



doc-id	user-id	document	
1	42	text…	embdngs
2	42	text…	embdngs
3	16	text…	embdngs
4	42	text…	embdngs
5	12	text…	embdngs
6	16	text…	embdngs
...	...	...	

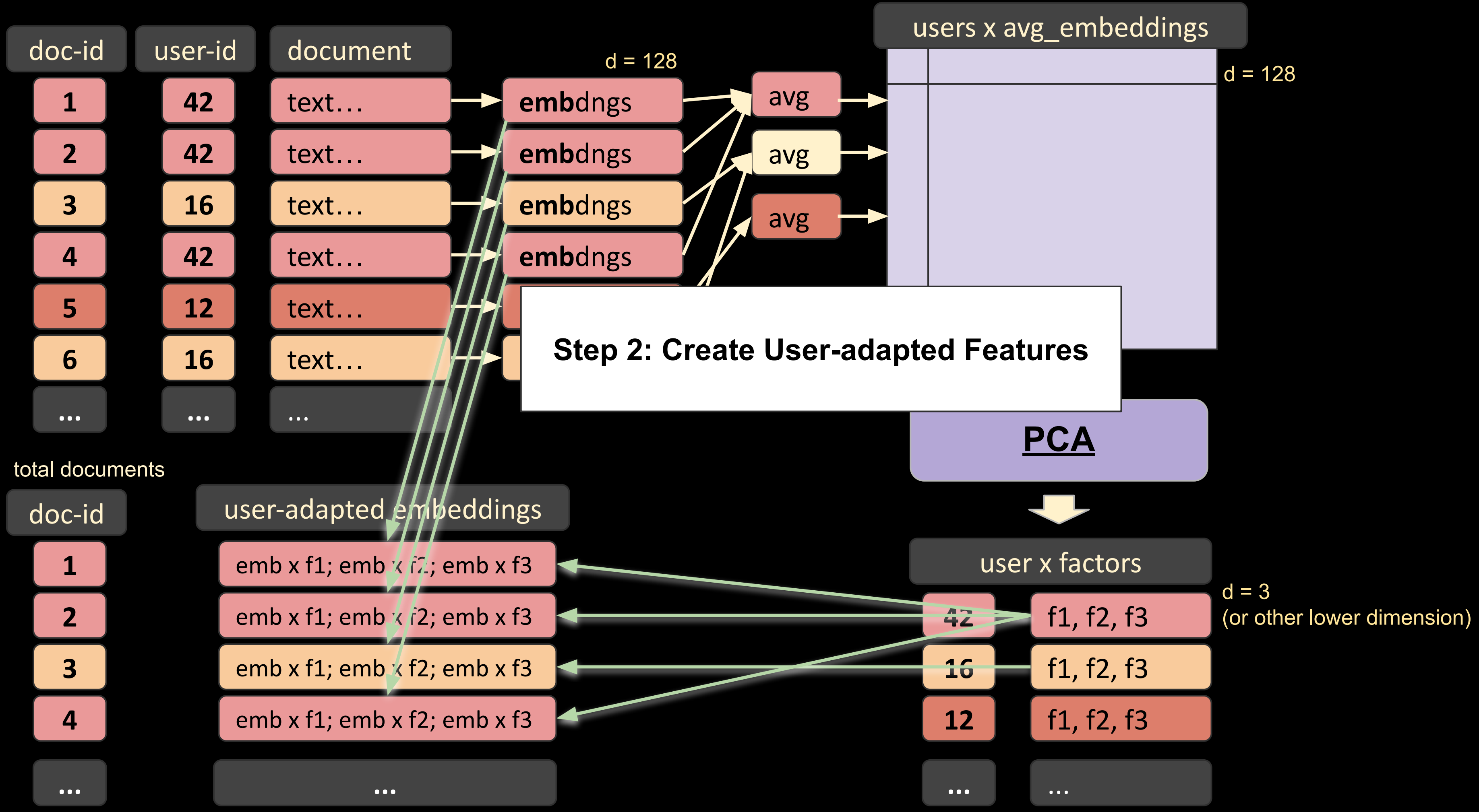
total documents

user x factors	
42	f1, f2, f3
16	f1, f2, f3
12	f1, f2, f3
...	...

d = 3
(or other lower dimension)

Step 1: Create User Factors

Full User Factors Adaptation Pipeline: with latent factors from training



doc-id	user-id	document
1	42	text…
2	42	text…
3	16	text…
4	42	text…
5	12	text…
6	16	text…
…	…	…

d = 128

embdngs

embdngs

embdngs

embdngs

avg

avg

avg

users x avg_embeddings

d = 128

Step 2: Create User-adapted Features

PCA

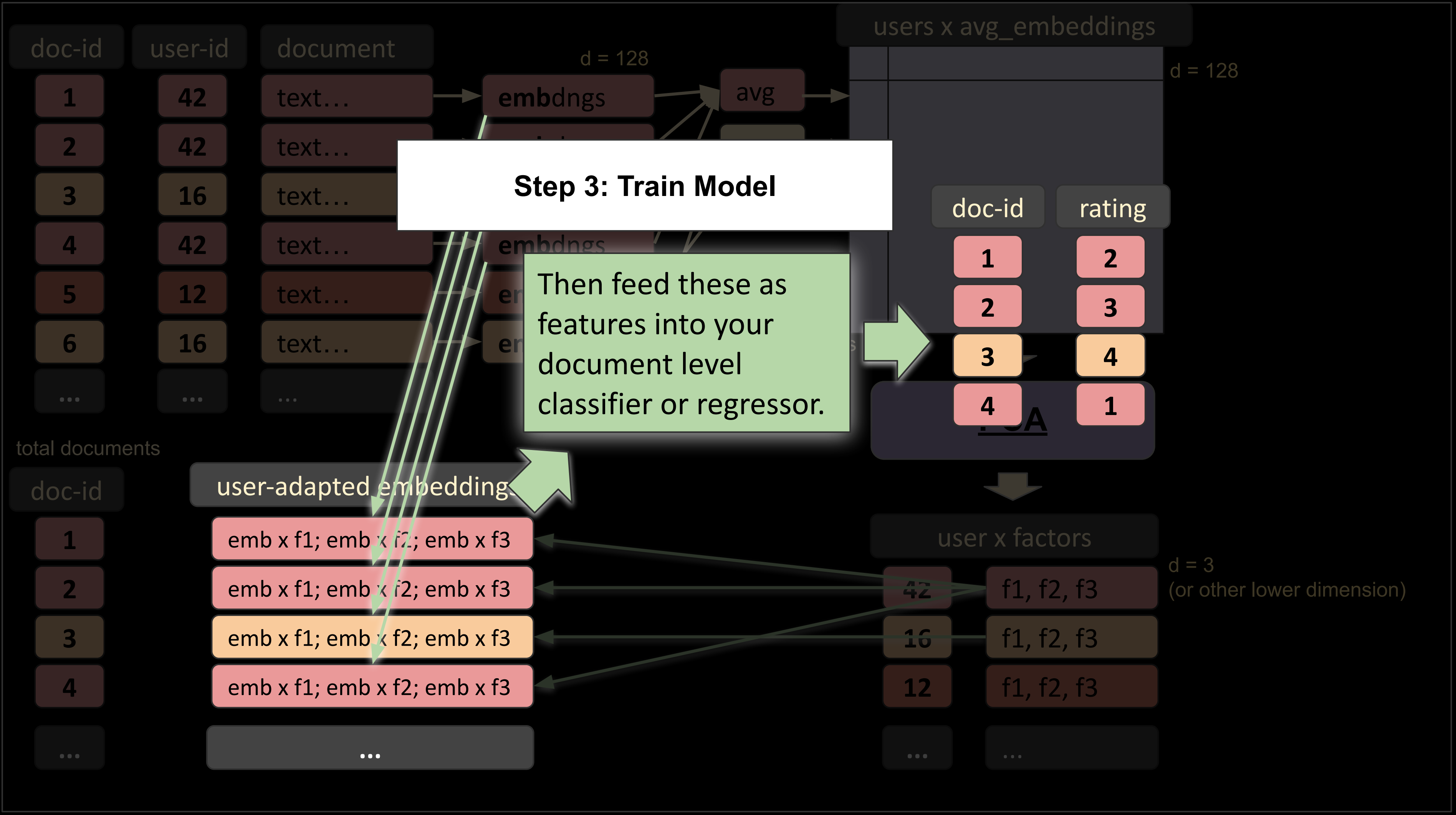
total documents

doc-id	user-adapted embeddings
1	emb x f1; emb x f2; emb x f3
2	emb x f1; emb x f2; emb x f3
3	emb x f1; emb x f2; emb x f3
4	emb x f1; emb x f2; emb x f3
…	…

user x factors	
42	f1, f2, f3
16	f1, f2, f3
12	f1, f2, f3
…	…

d = 3
(or other lower dimension)

Full User Factors Adaptation Pipeline: with latent factors from training



doc-id	user-id	document
1	42	text…
2	42	text…
3	16	text…
4	42	text…
5	12	text…
6	16	text…
…	…	…

Step 3: Train Model

doc-id	rating
1	2
2	3
3	4
4	1

Then feed these as features into your document level classifier or regressor.

total documents

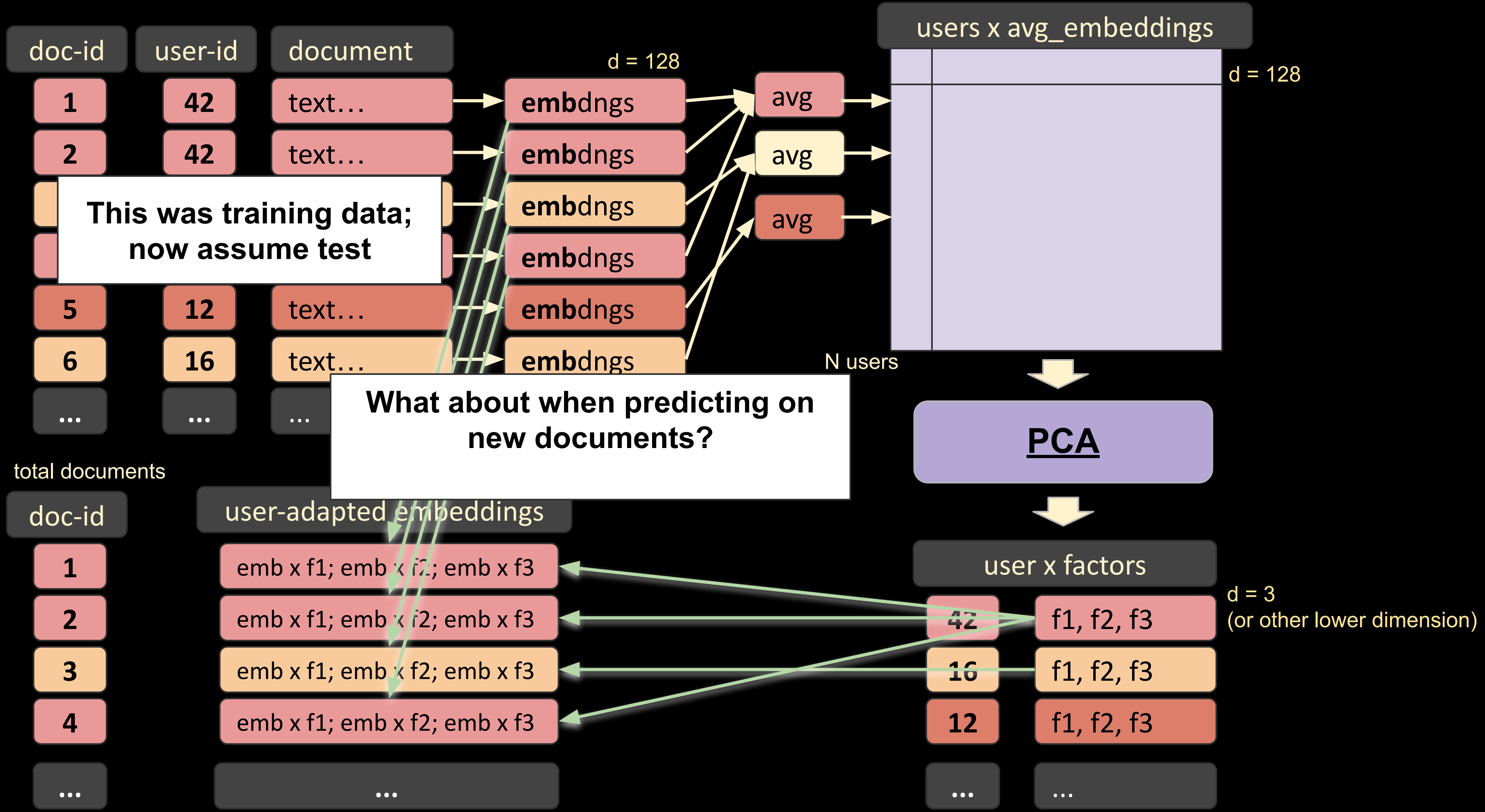
doc-id	user-adapted embeddings
1	emb x f1; emb x f2; emb x f3
2	emb x f1; emb x f2; emb x f3
3	emb x f1; emb x f2; emb x f3
4	emb x f1; emb x f2; emb x f3
…	…

user x factors

42	f1, f2, f3
16	f1, f2, f3
12	f1, f2, f3
…	…

d = 3
(or other lower dimension)

Full User Factors Adaptation Pipeline: with latent factors from training



doc-id	user-id	document	d = 128
1	42	text…	embdngs
2	42	text…	embdngs
			embdngs
			embdngs
5	12	text…	embdngs
6	16	text…	embdngs
…	…	…	

total documents

This was training data; now assume test

avg

avg

avg

users x avg_embeddings

d = 128

N users

PCA

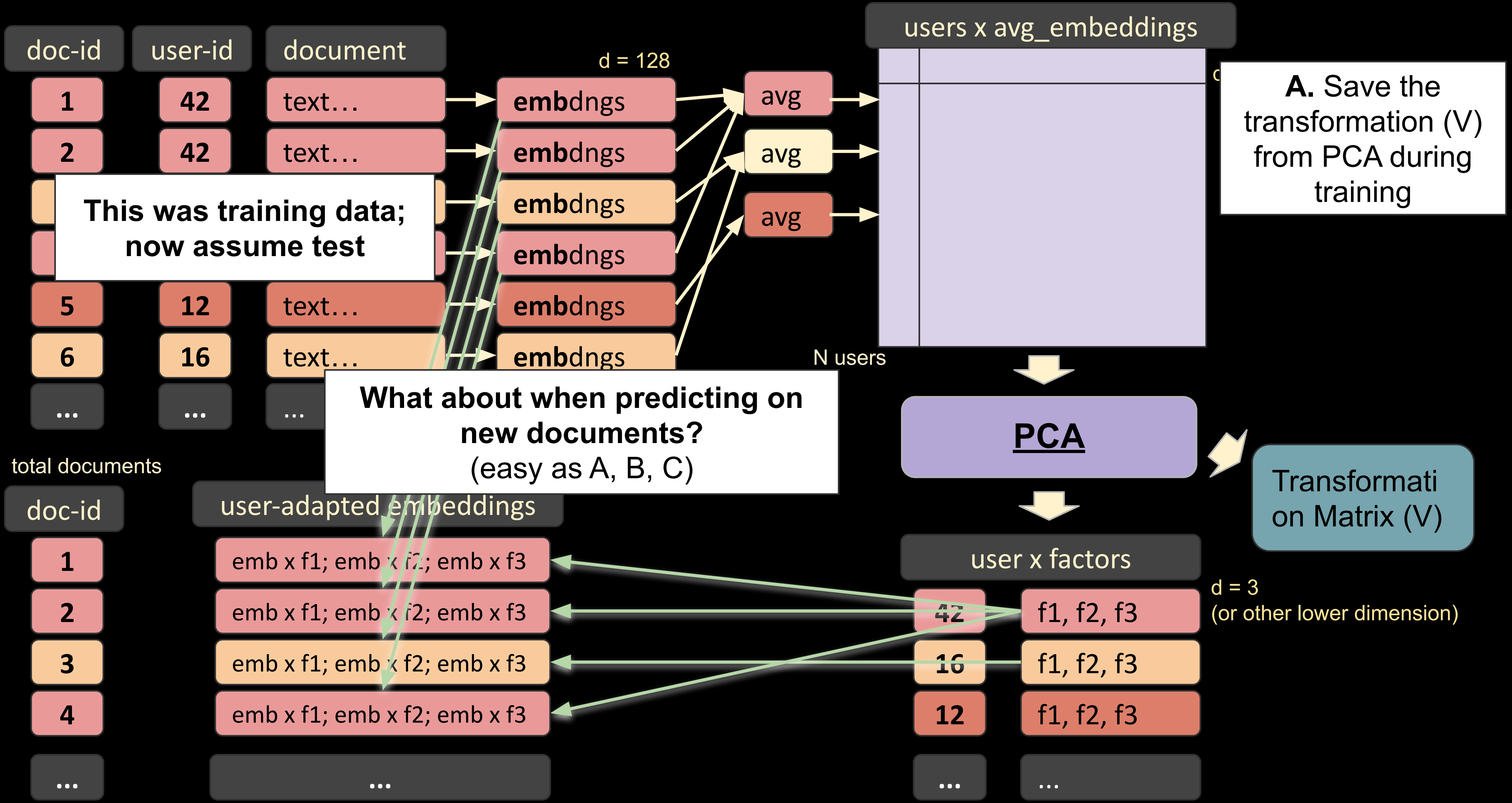
What about when predicting on new documents?

doc-id	user-adapted embeddings
1	emb x f1; emb x f2; emb x f3
2	emb x f1; emb x f2; emb x f3
3	emb x f1; emb x f2; emb x f3
4	emb x f1; emb x f2; emb x f3
…	…

user x factors	
42	f1, f2, f3
16	f1, f2, f3
12	f1, f2, f3
…	…

d = 3
(or other lower dimension)

Full User Factors Adaptation Pipeline: with latent factors from training



doc-id	user-id	document
1	42	text…
2	42	text…
5	12	text…
6	16	text…
…	…	…

total documents

d = 128

embdngs

embdngs

embdngs

embdngs

embdngs

embdngs

avg

avg

avg

users x avg_embeddings

N users

This was training data; now assume test

A. Save the transformation (V) from PCA during training

PCA

Transformation Matrix (V)

What about when predicting on new documents?

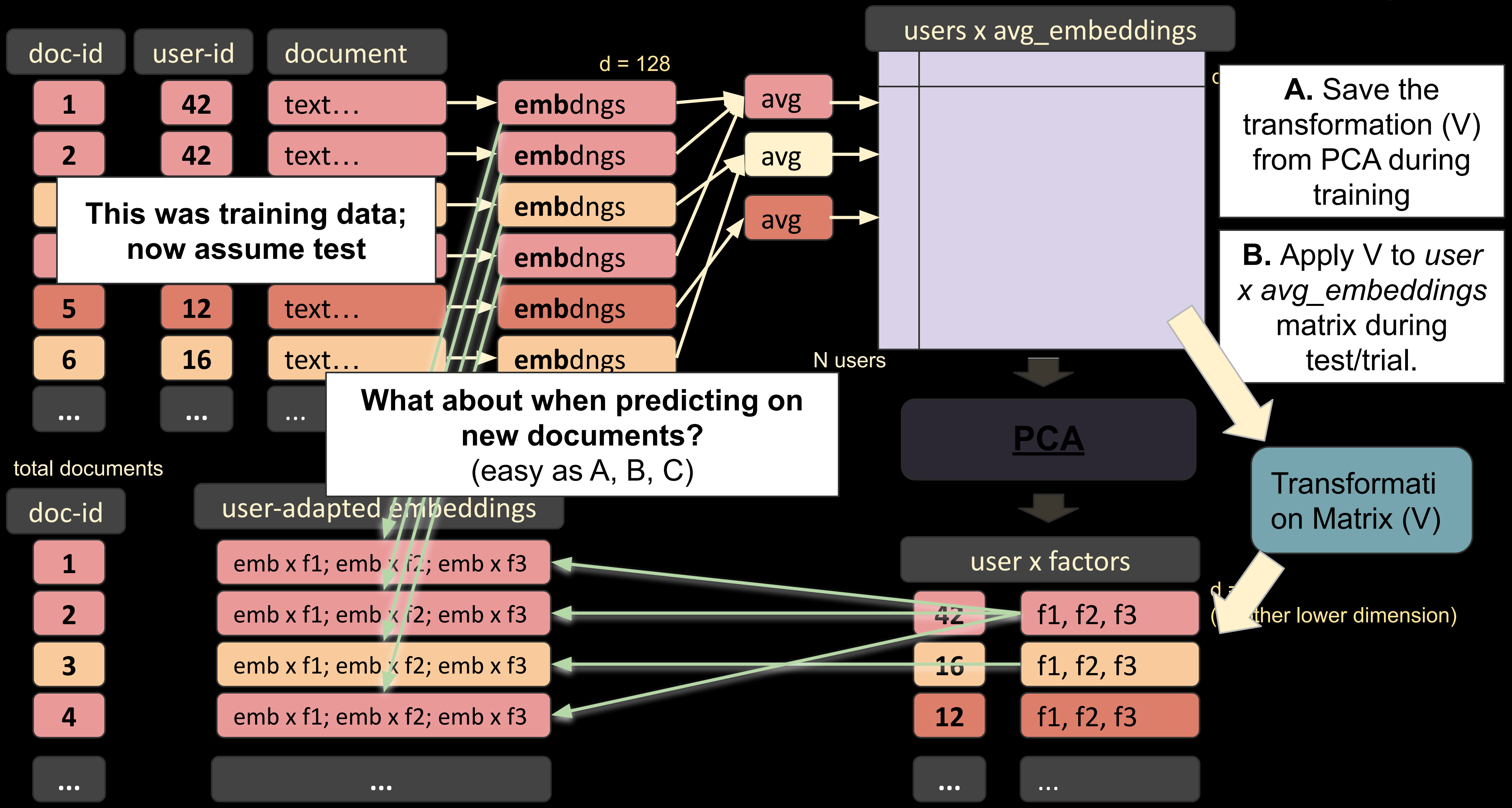
(easy as A, B, C)

doc-id	user-adapted embeddings
1	emb x f1; emb x f2; emb x f3
2	emb x f1; emb x f2; emb x f3
3	emb x f1; emb x f2; emb x f3
4	emb x f1; emb x f2; emb x f3
…	…

user x factors	
42	f1, f2, f3
16	f1, f2, f3
12	f1, f2, f3
…	…

d = 3
(or other lower dimension)

Full User Factors Adaptation Pipeline: with latent factors from training



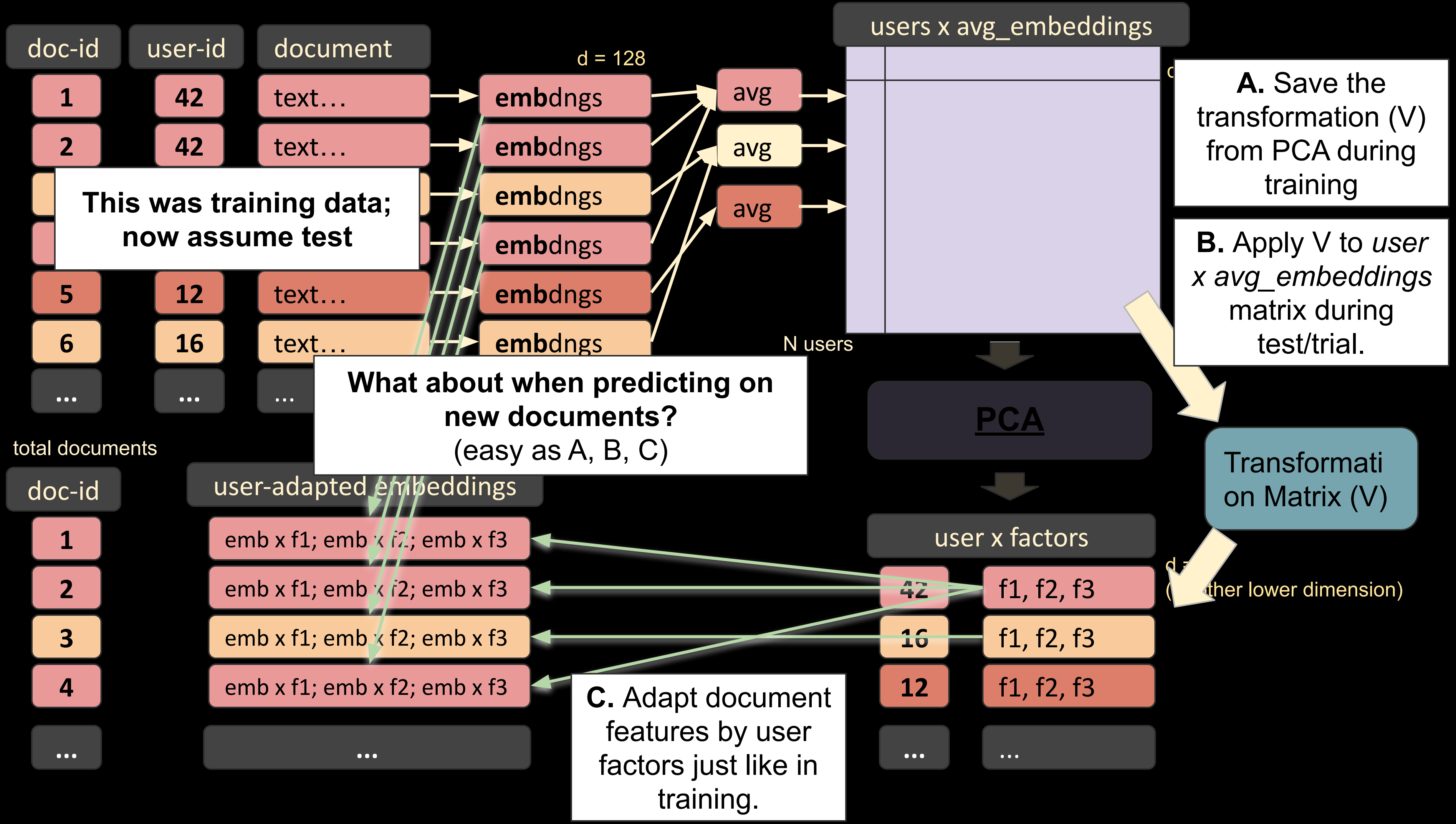
This was training data; now assume test

What about when predicting on new documents?
(easy as A, B, C)

A. Save the transformation (V) from PCA during training

B. Apply V to *user x avg_embeddings* matrix during test/trial.

Full User Factors Adaptation Pipeline: with latent factors from training



This was training data; now assume test

What about when predicting on new documents?
(easy as A, B, C)

A. Save the transformation (V) from PCA during training

B. Apply V to *user x avg_embeddings* matrix during test/trial.

C. Adapt document features by user factors just like in training.

Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)

2. Additive: Include direct effect of human factor on outcome.
   (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.
   (e.g. image captioner label pictures of men in kitchen as women)

Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)

   (e.g. “sick” said from a young individual versus old individual)

2. Additive: Include direct effect of human factor on outcome.

   (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.

   (e.g. image captioner label pictures of men in kitchen as women)

Ethics in NLP

Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)

2. Additive: Include direct effect of human factor on outcome.
   (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.
   (e.g. image captioner label pictures of men in kitchen as women)

Ethics in NLP

Bias

Privacy

Ethical Research

Ethics in NLP - Bias

Consequences of Sociodemographic Bias in NLP Models:

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

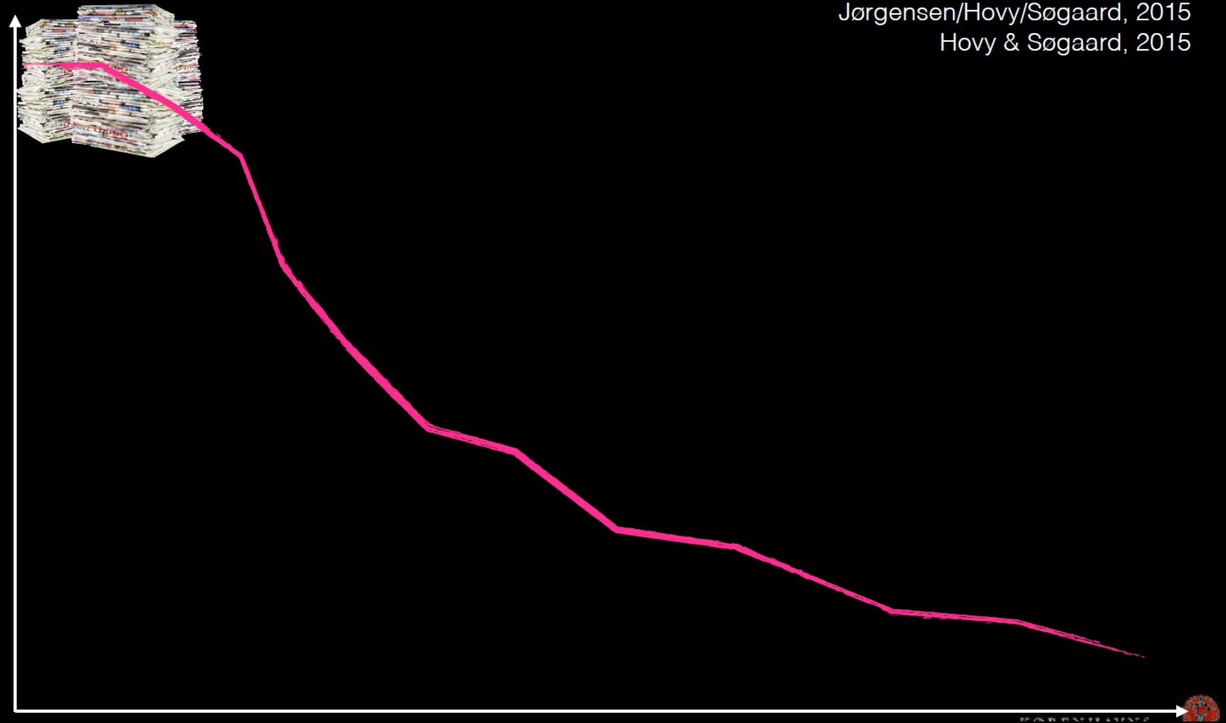
- Error Disparity: Predicts less accurate for authors of given demographics.

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

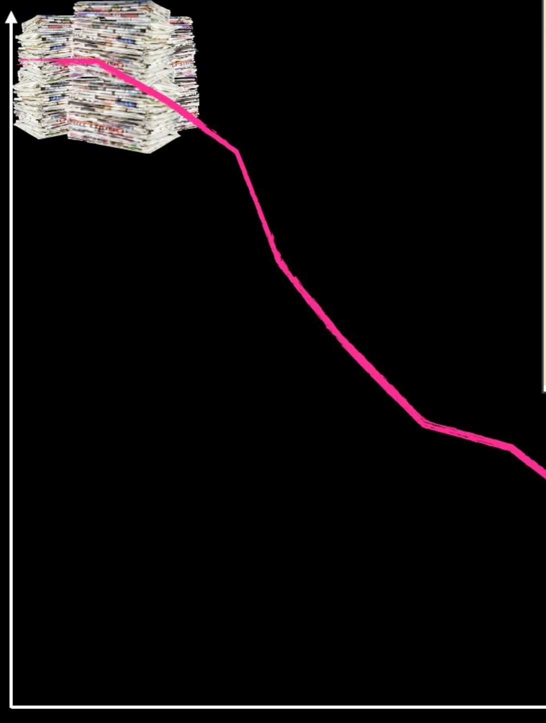
The WSJ Effect

Jørgensen/Hovy/Søgaard, 2015
Hovy & Søgaard, 2015

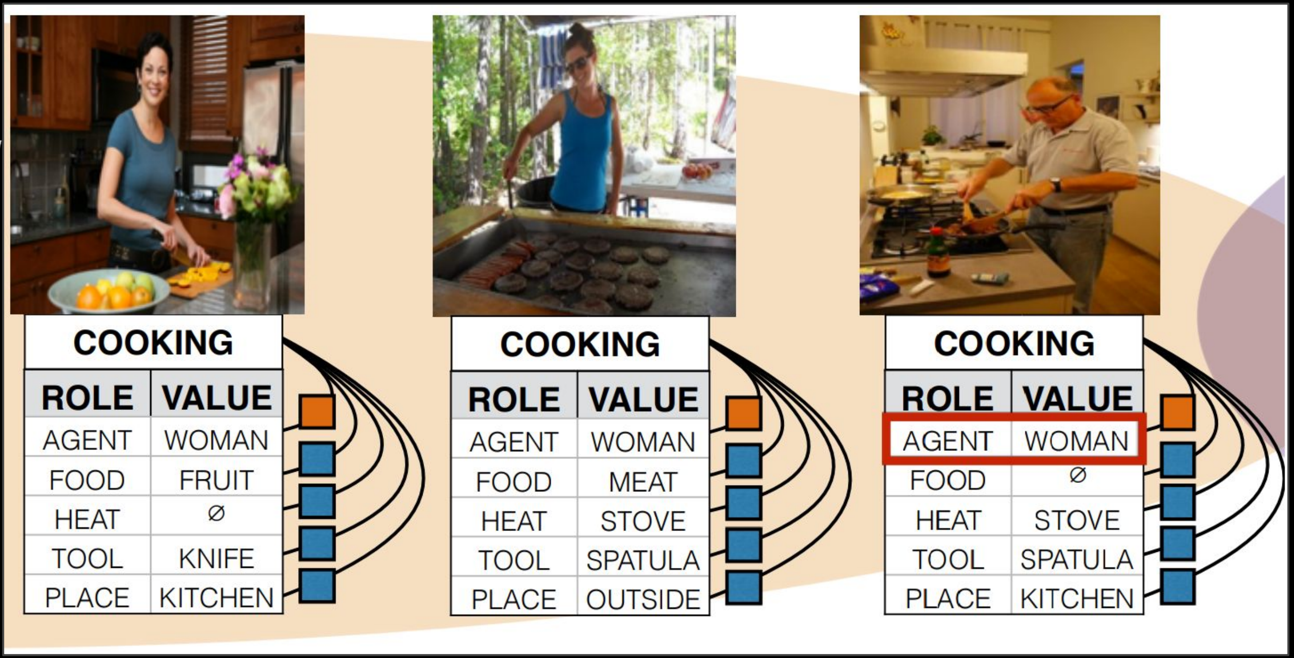


Two Examples

The W



distance from “standard” WSJ author demographics



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

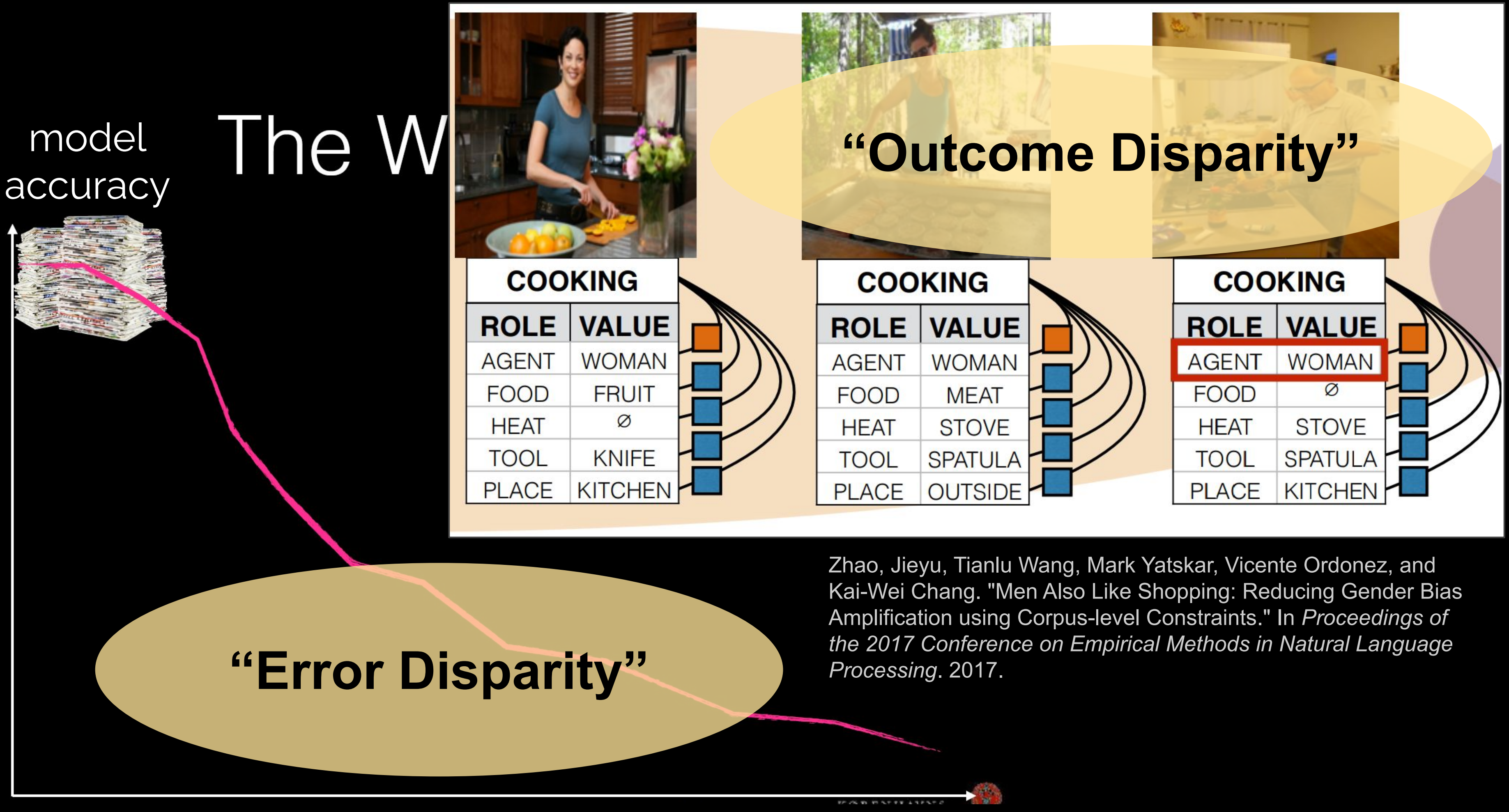
COOKING

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.

Two Examples

The W



“Outcome Disparity”

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

“Error Disparity”

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”

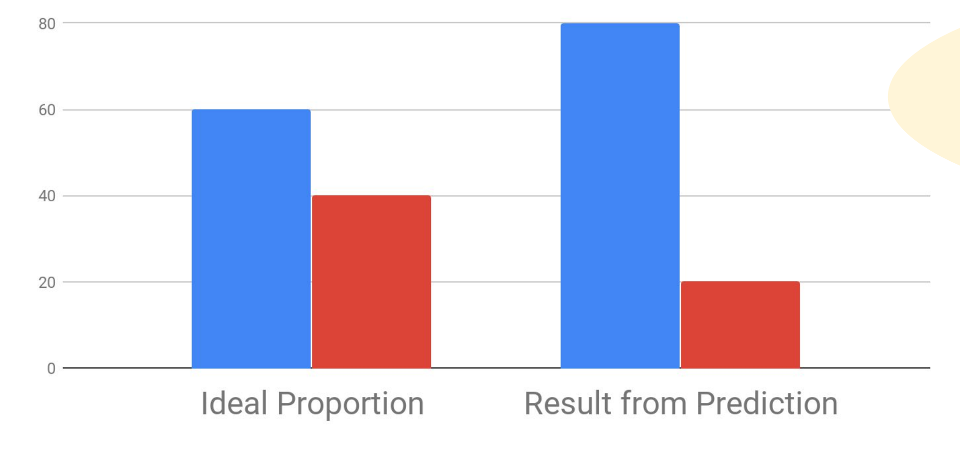
Person-level

🟦 attribute = 1

🟥 attribute = 2

“Error Disparity”

Our data and models are (human) biased.



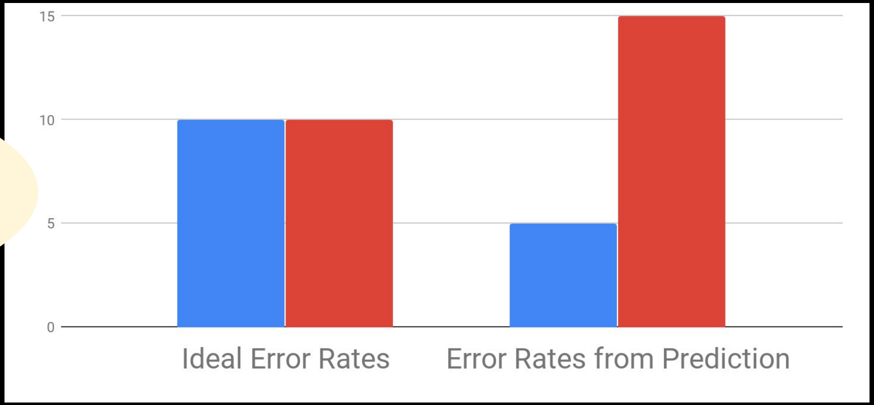
“Outcome Disparity”

Person-level

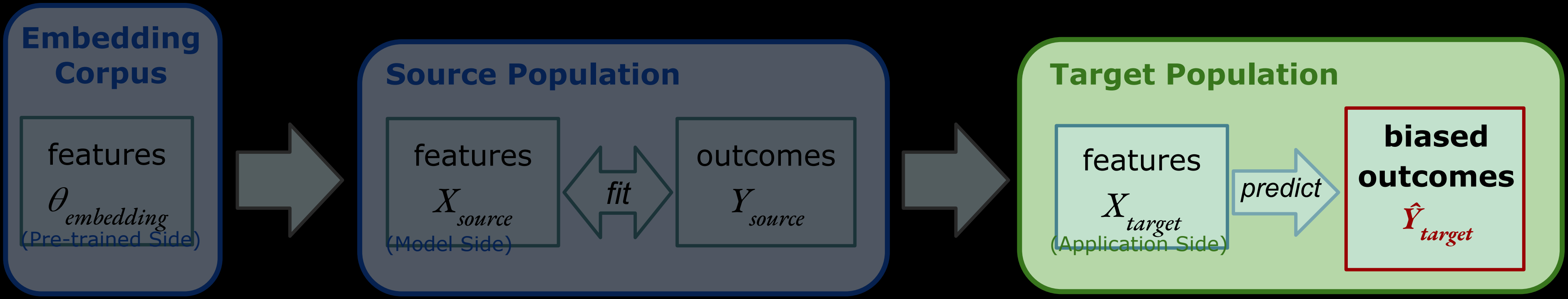
🟦 attribute = 1

🟥 attribute = 2

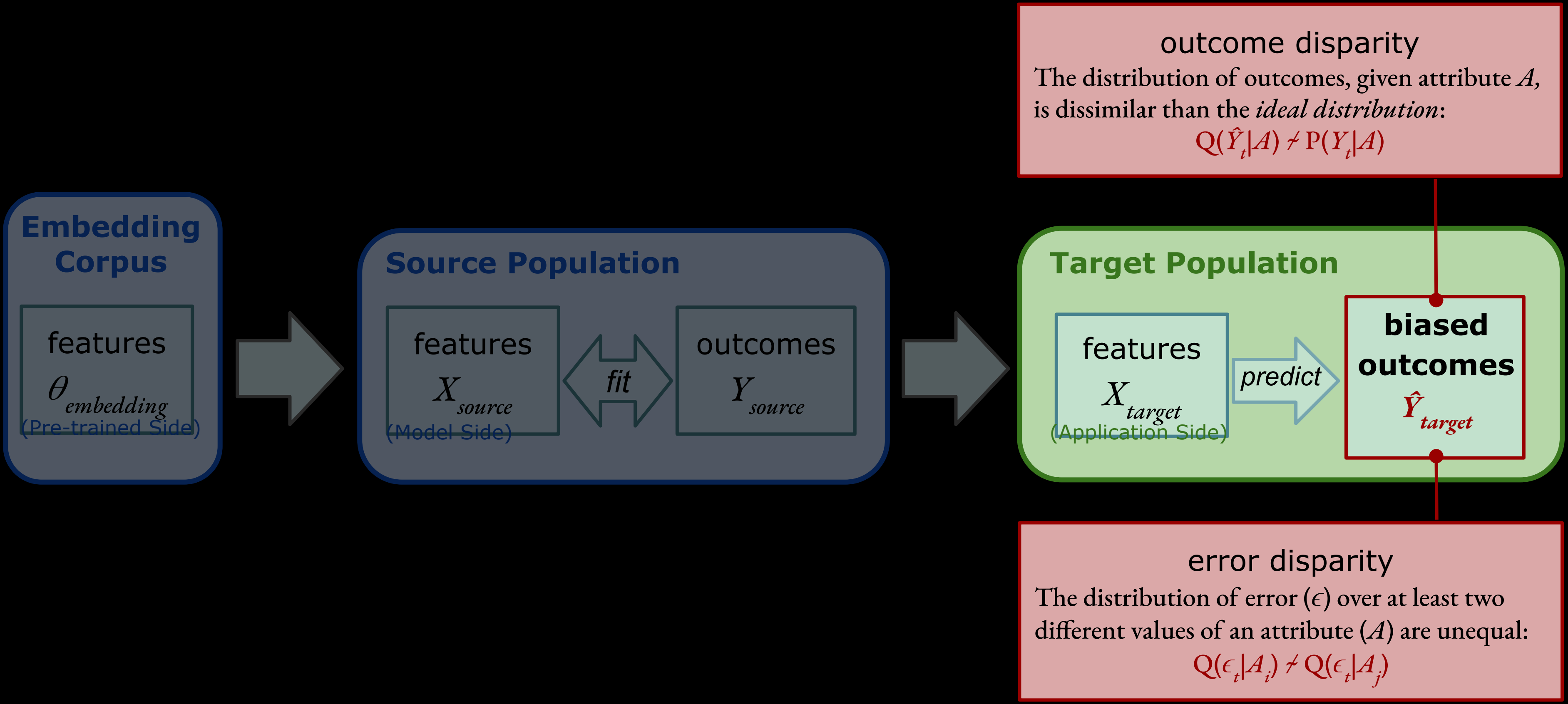
“Error Disparity”



Conceptual Framework:



Conceptual Framework:



Embedding Corpus

features

$\theta_{embedding}$

(Pre-trained Side)

Source Population

features

X_{source}

fit

outcomes

Y_{source}

(Model Side)

Target Population

features

X_{target}

predict

biased outcomes

$\hat{Y}_{target}$

(Application Side)

outcome disparity

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

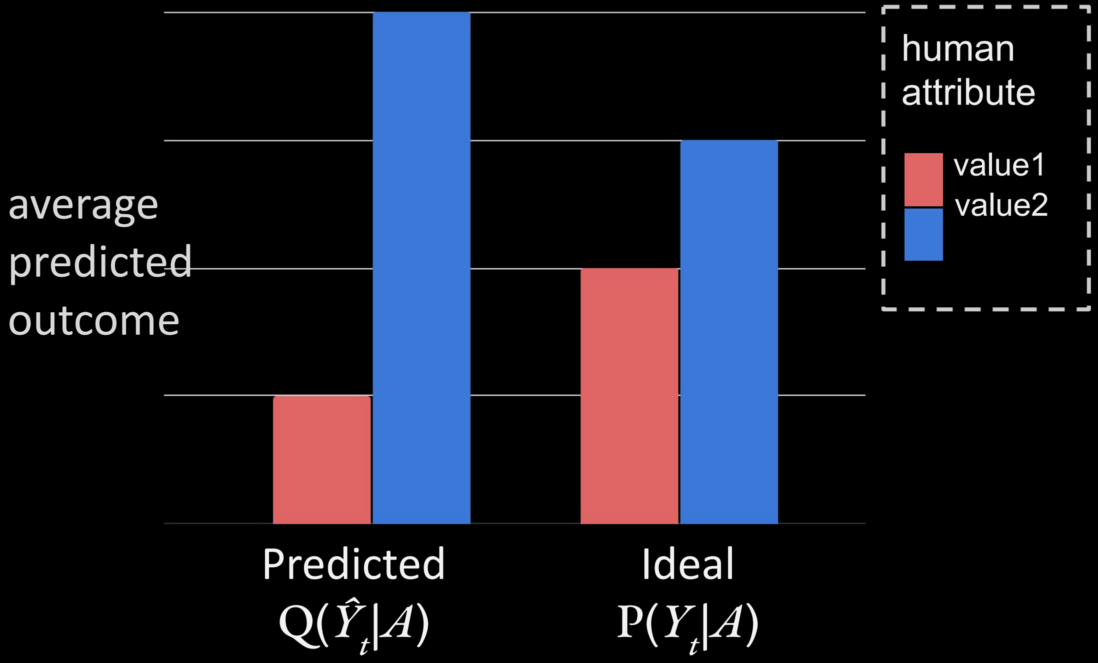
$Q(\hat{Y}_t|A) \nsim P(Y_t|A)$

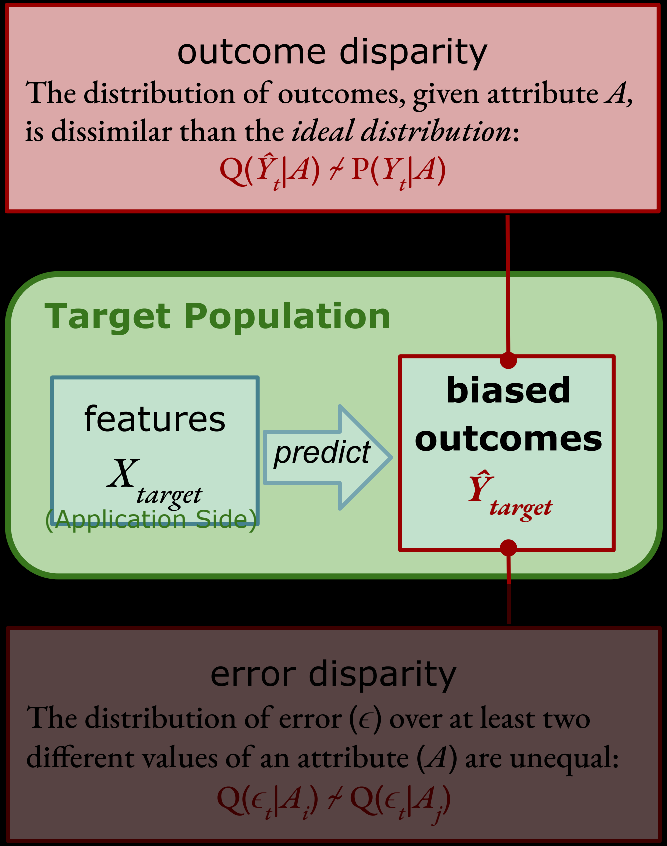
error disparity

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

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

Outcome Disparity





outcome disparity

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

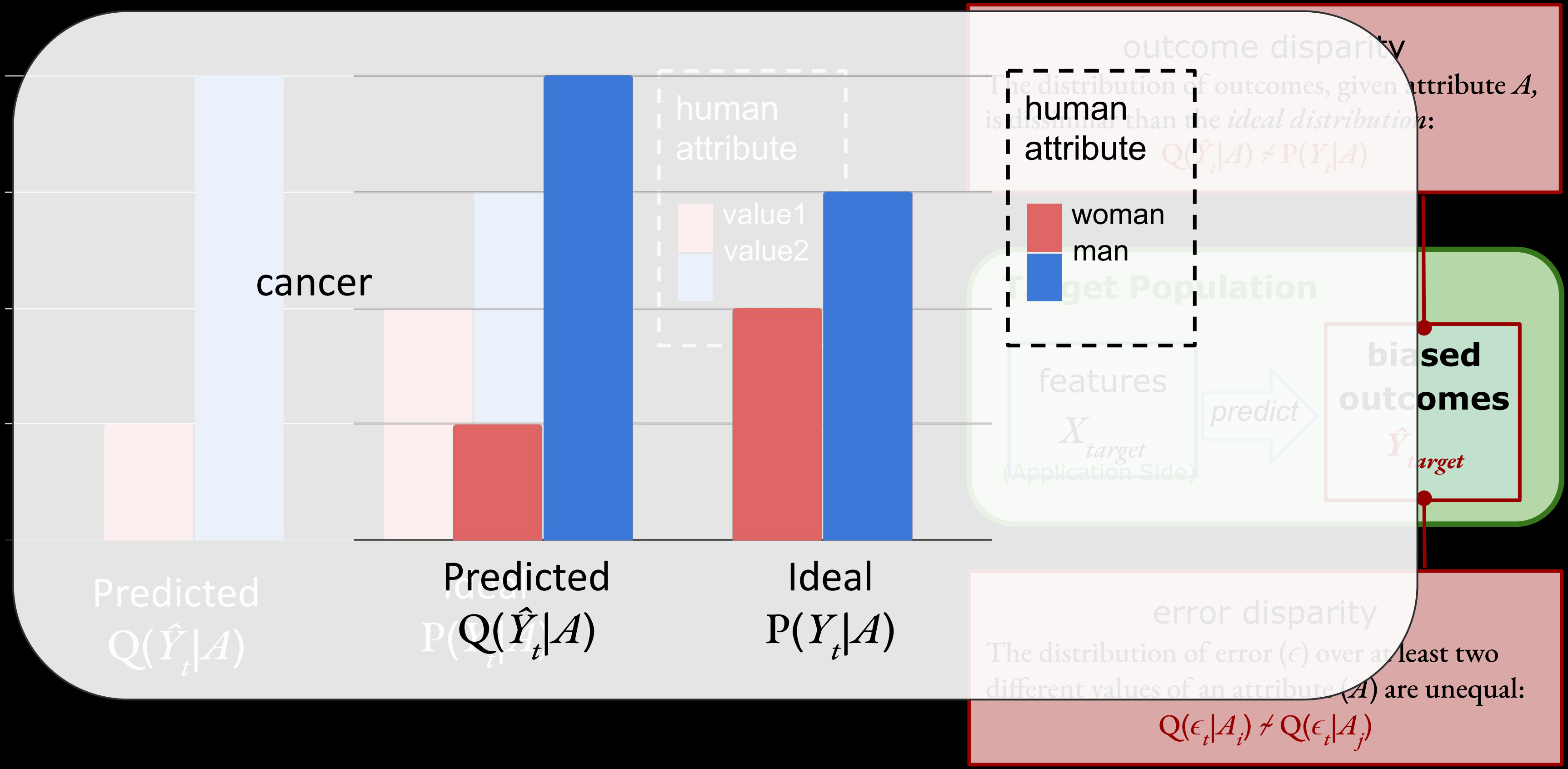
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$

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



outcome disparity

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

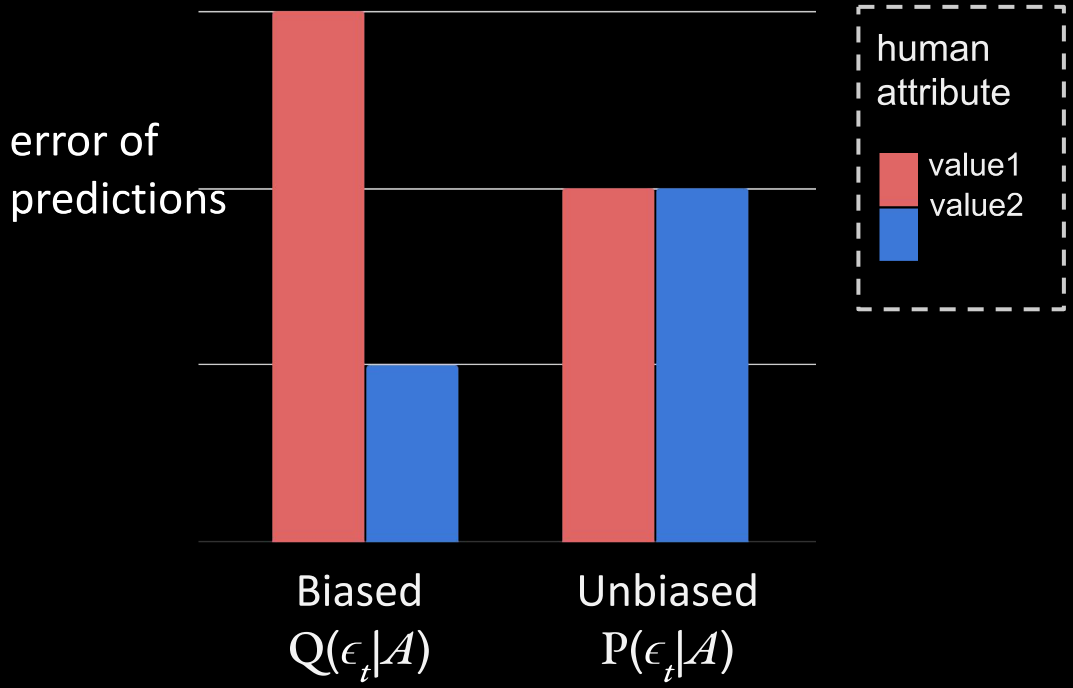
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$

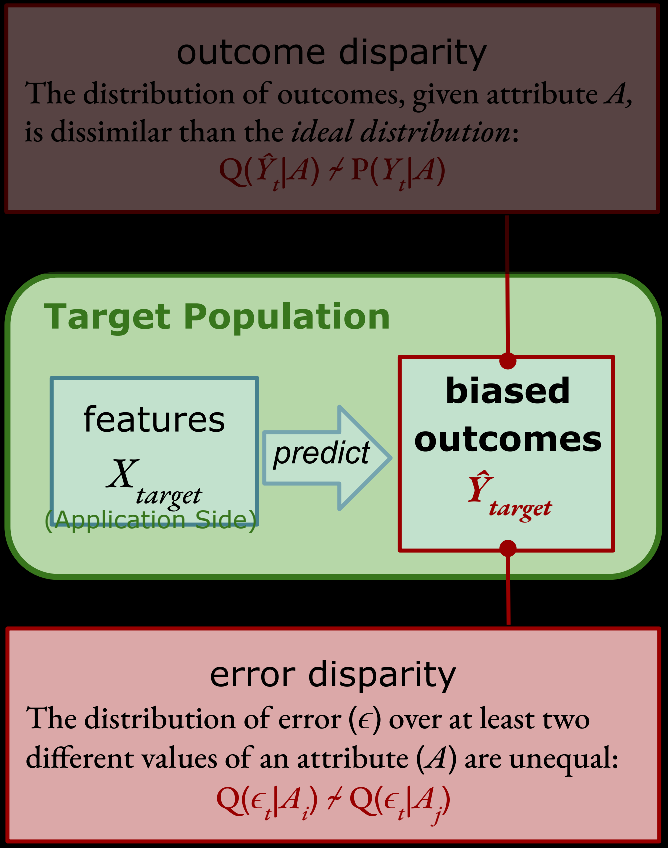
error disparity

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





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

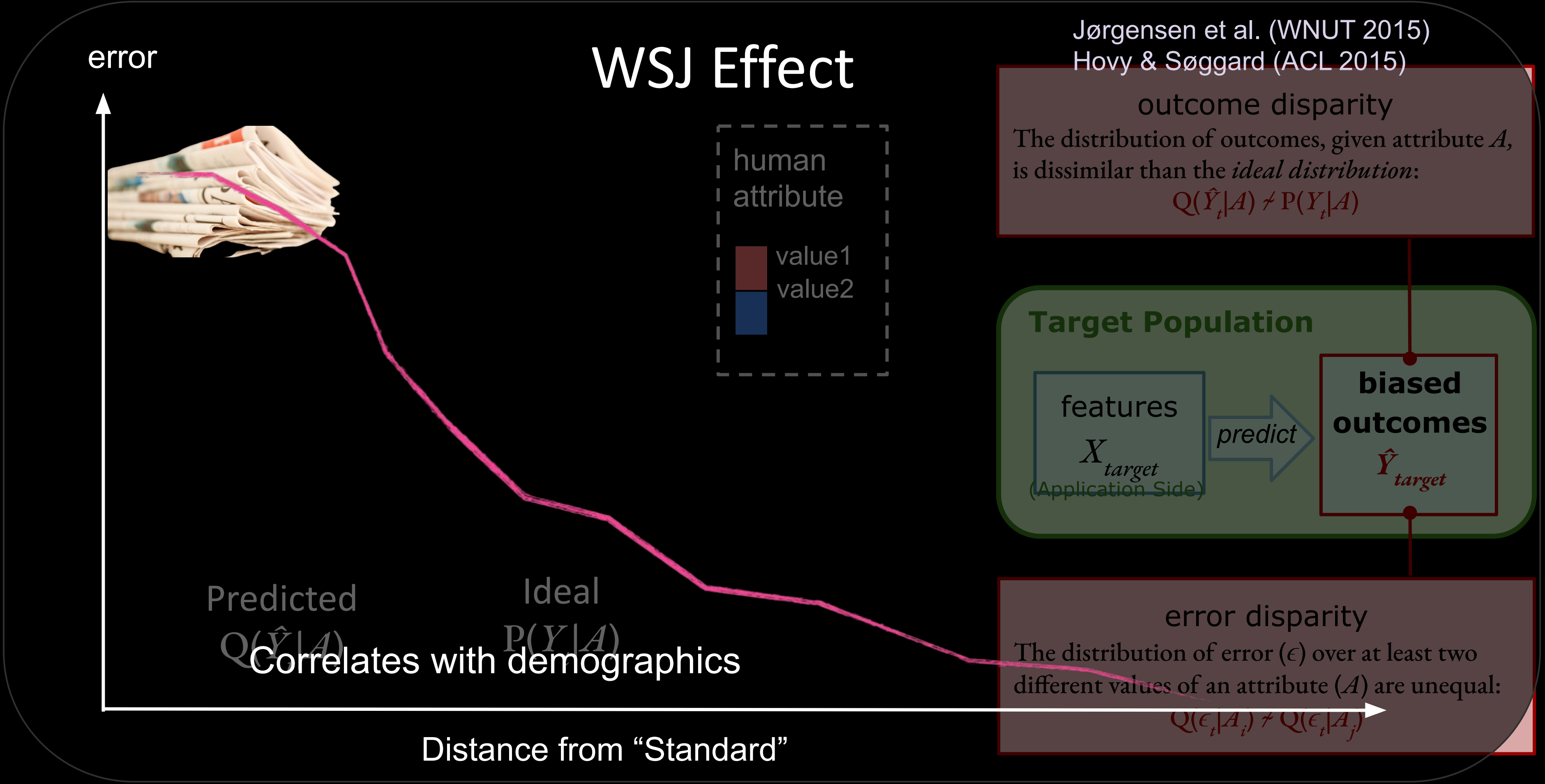
biased outcomes $\hat{Y}_{target}$

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)

Correlates with demographics

outcome disparity

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

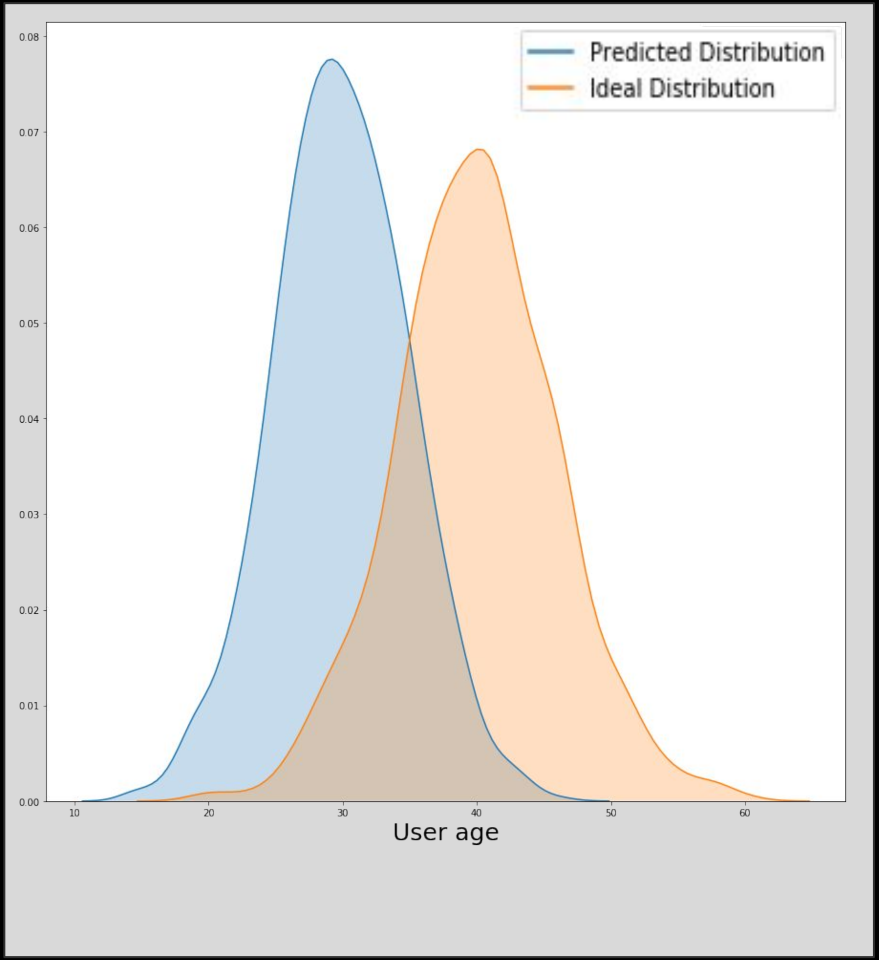
$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

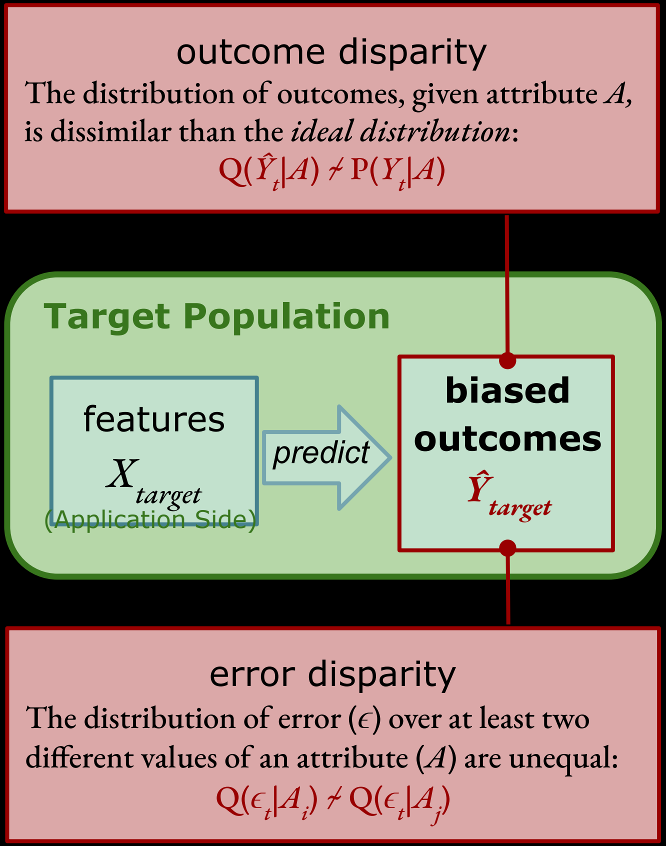
error disparity

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*:

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

Target Population

features

X_{target}

(Application Side)

predict

biased outcomes

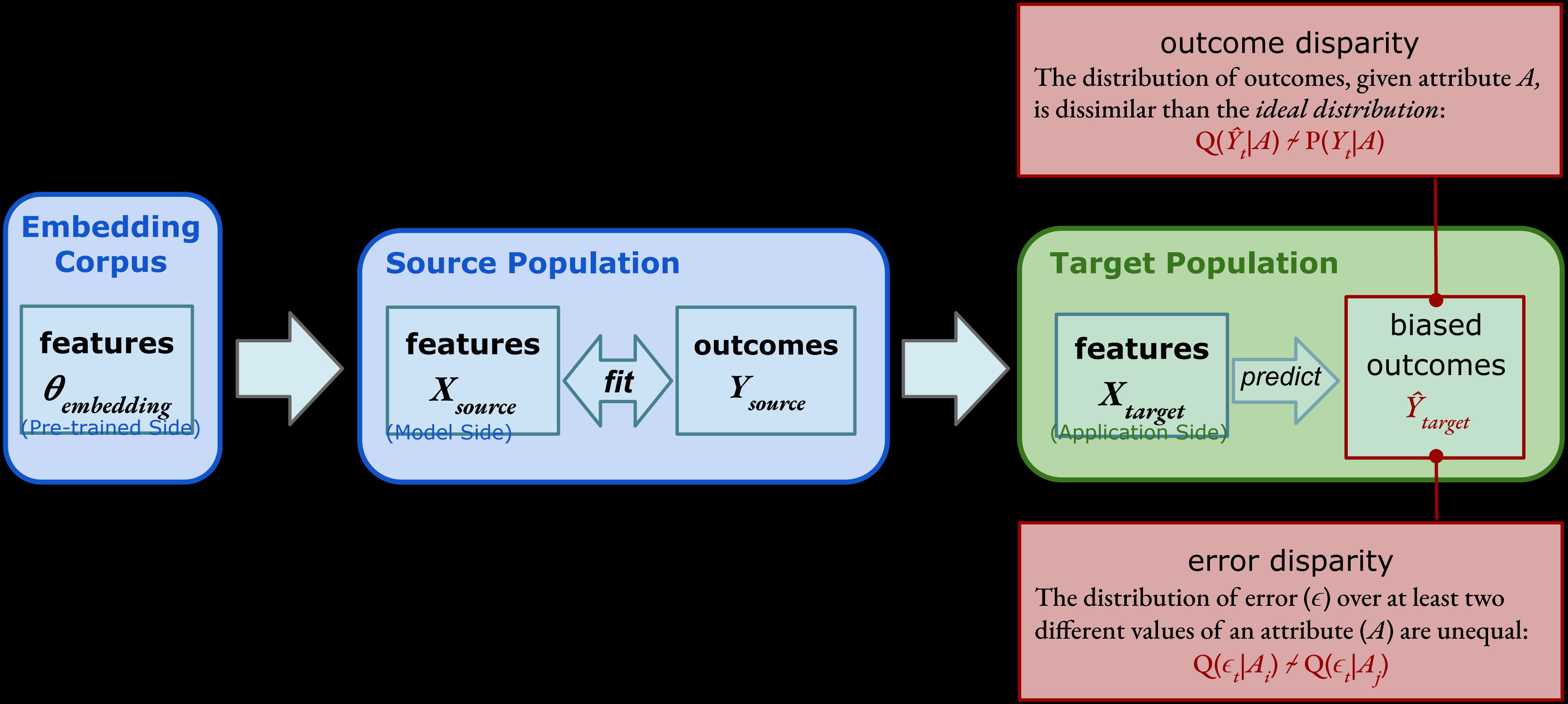
$\hat{Y}_{target}$

error disparity

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

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

Origins of Bias



outcome disparity

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

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

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

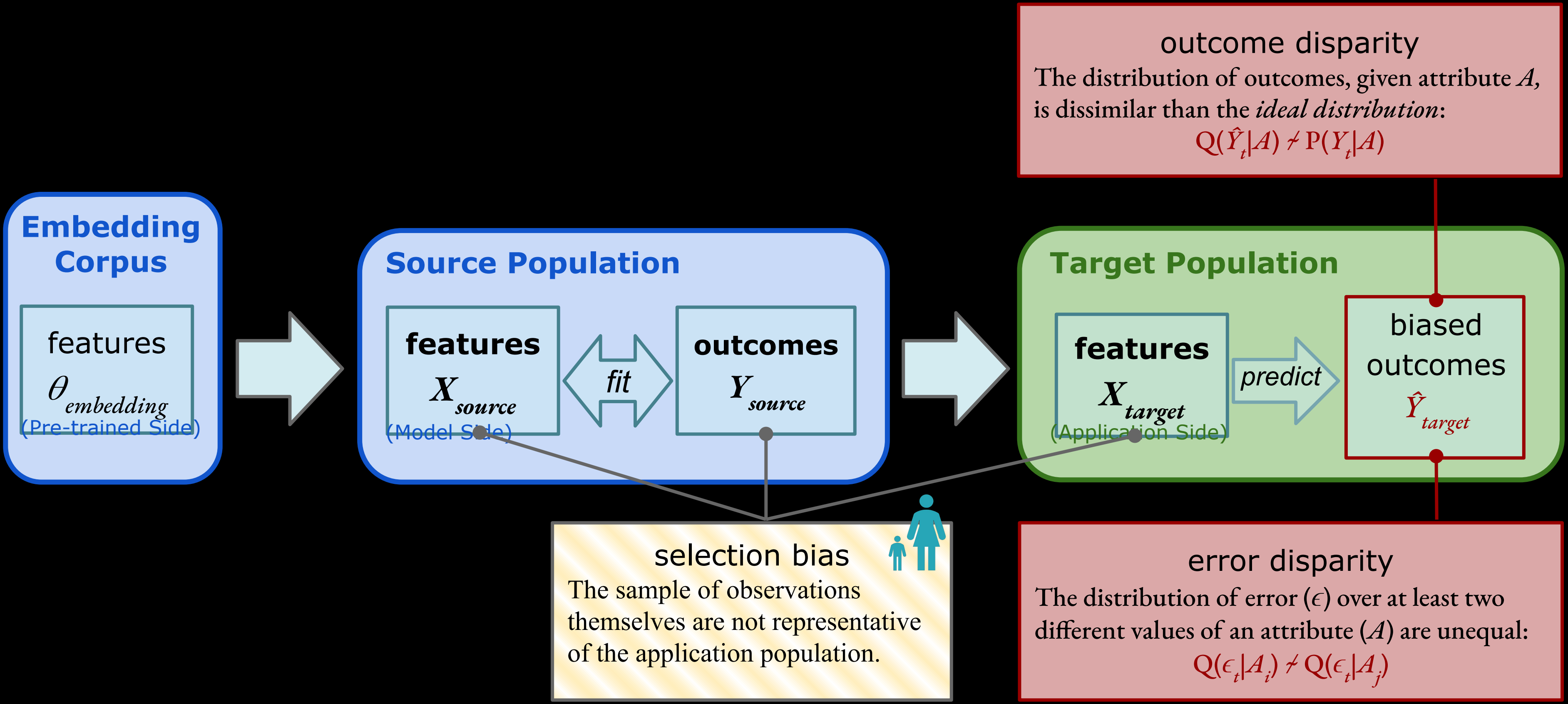
$\hat{Y}_{target}$

error disparity

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

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

Selection Bias



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

$\hat{Y}_{target}$

outcome disparity

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

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

selection bias

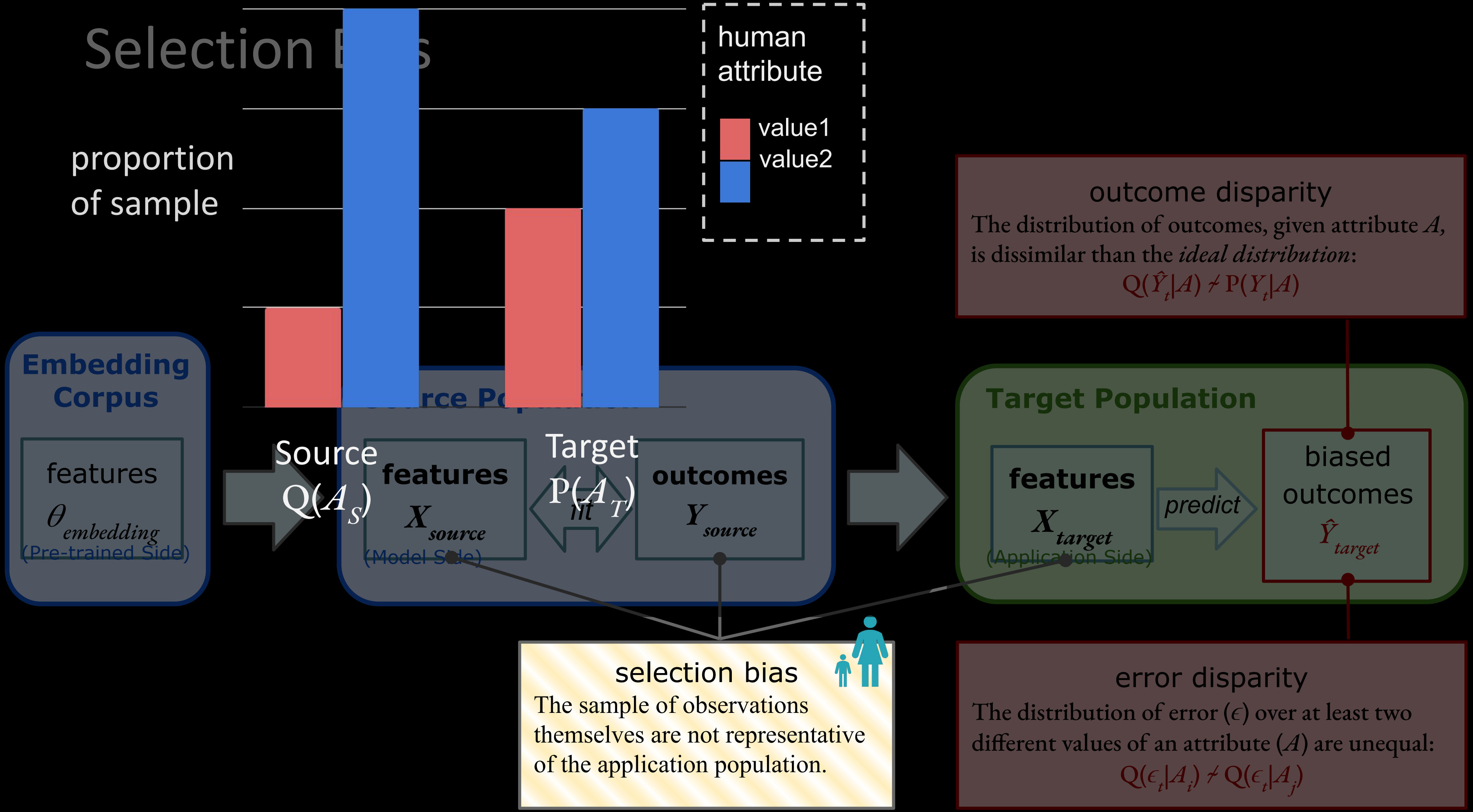
The sample of observations themselves are not representative of the application population.

error disparity

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

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

Selection Bias



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.

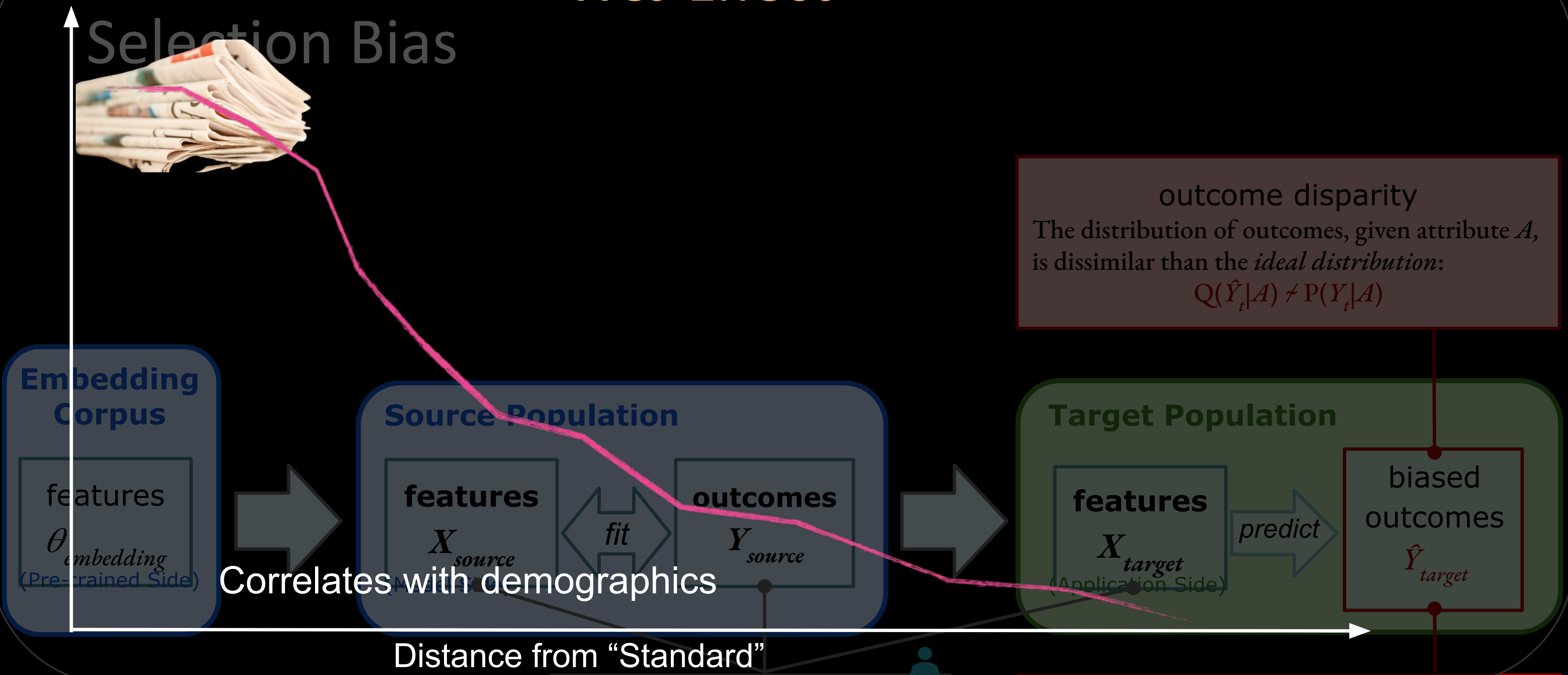
error disparity

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

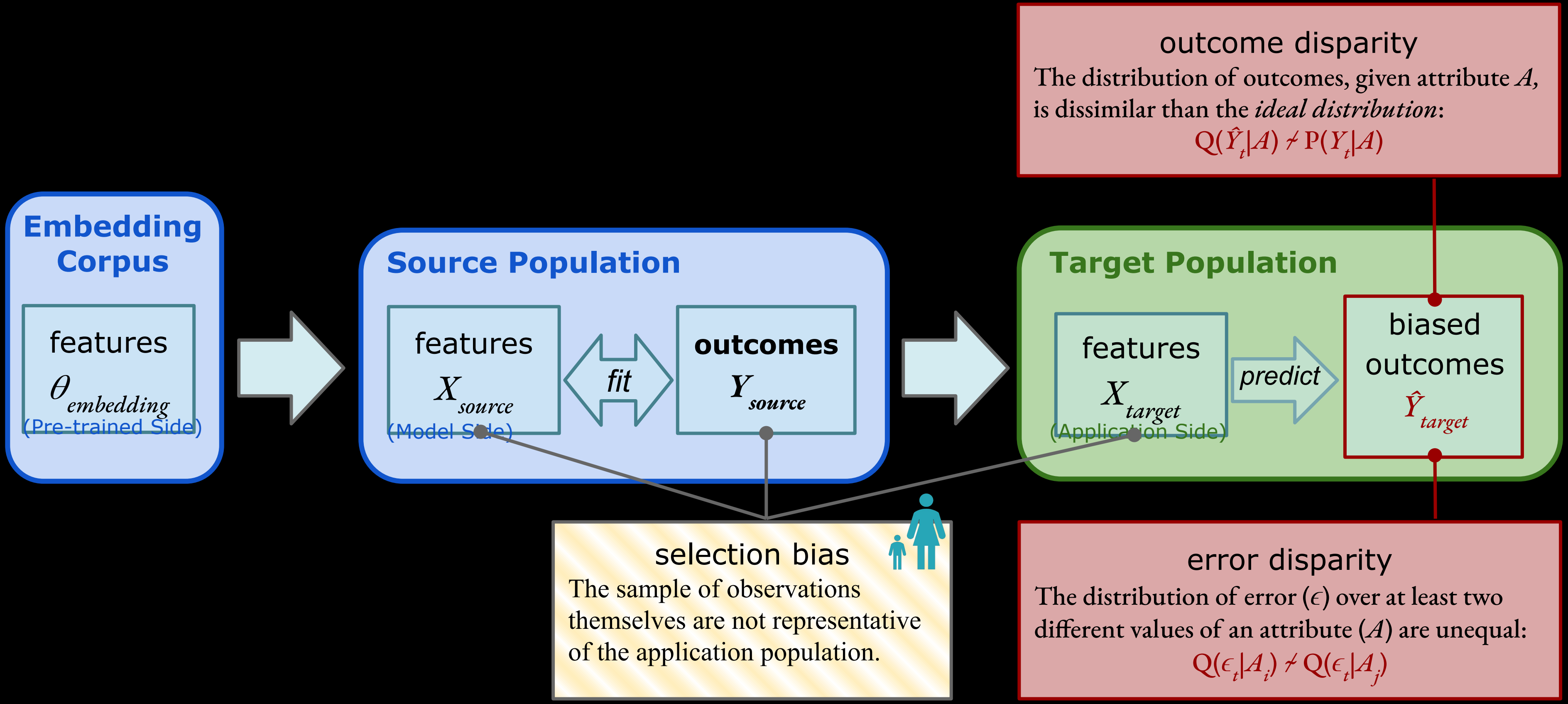
The sample of observations themselves are not representative of the application population.

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



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

$\hat{Y}_{target}$

outcome disparity

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

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

selection bias

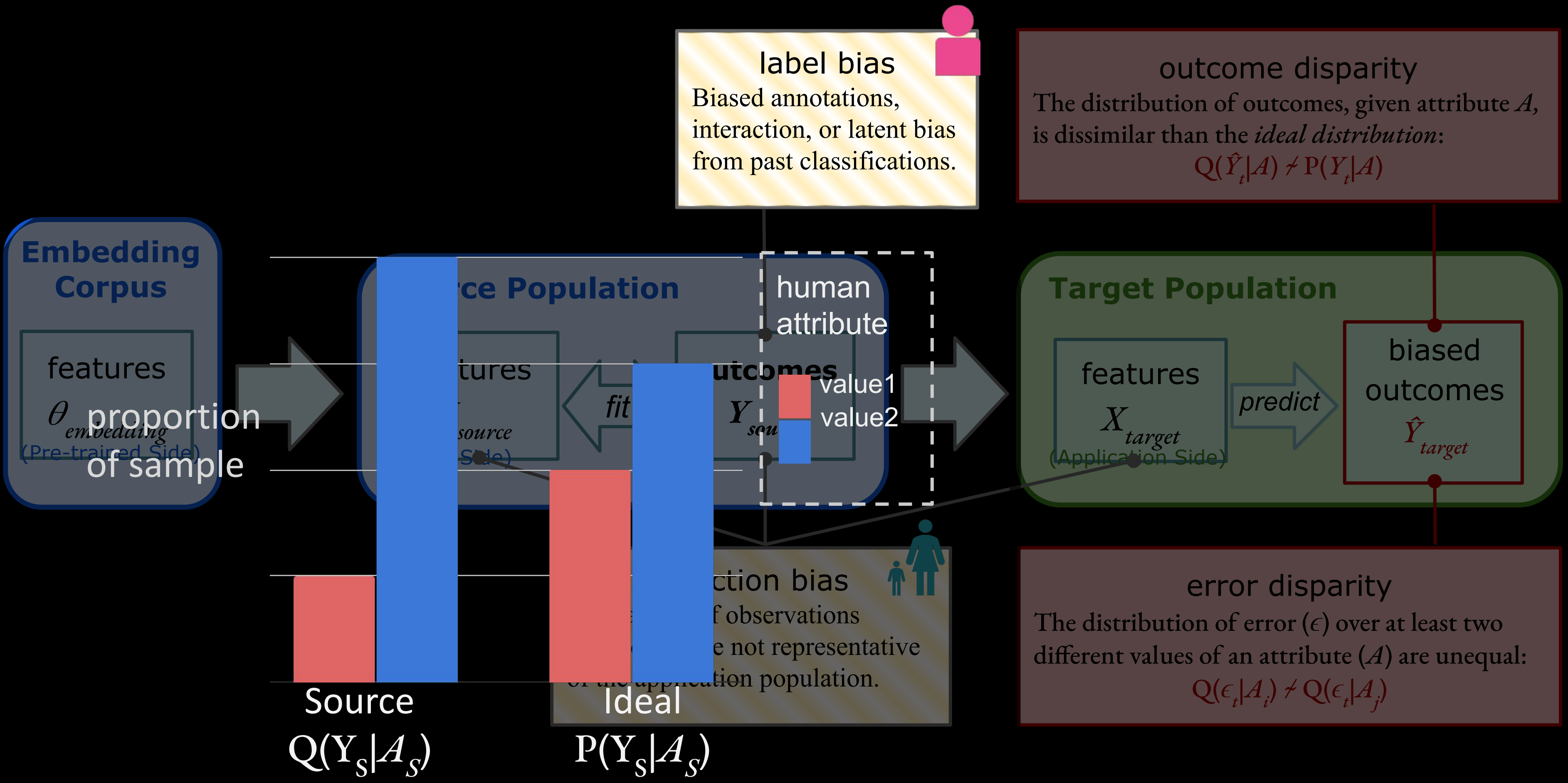
The sample of observations themselves are not representative of the application population.

error disparity

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

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

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*:

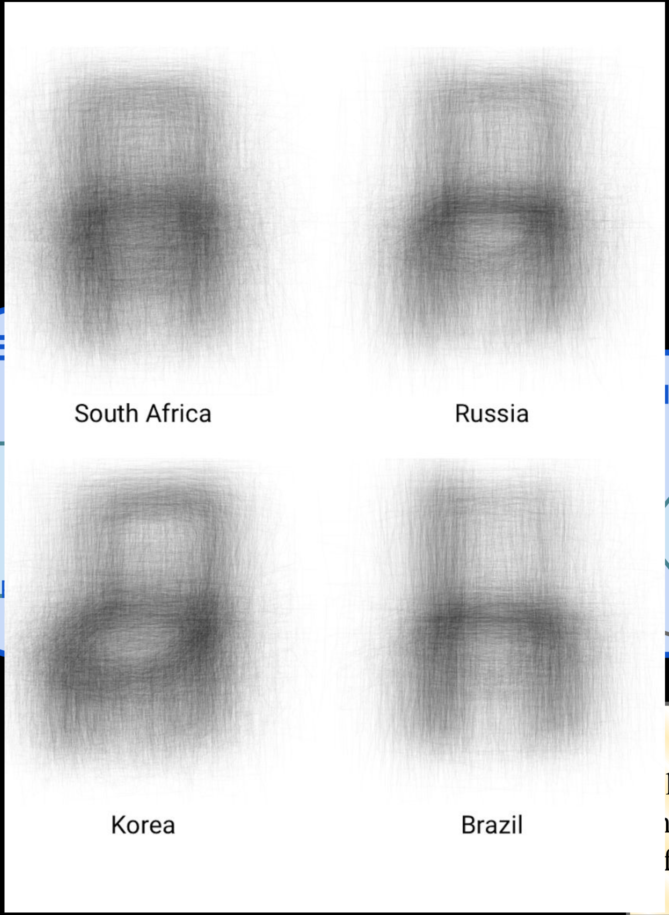
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$

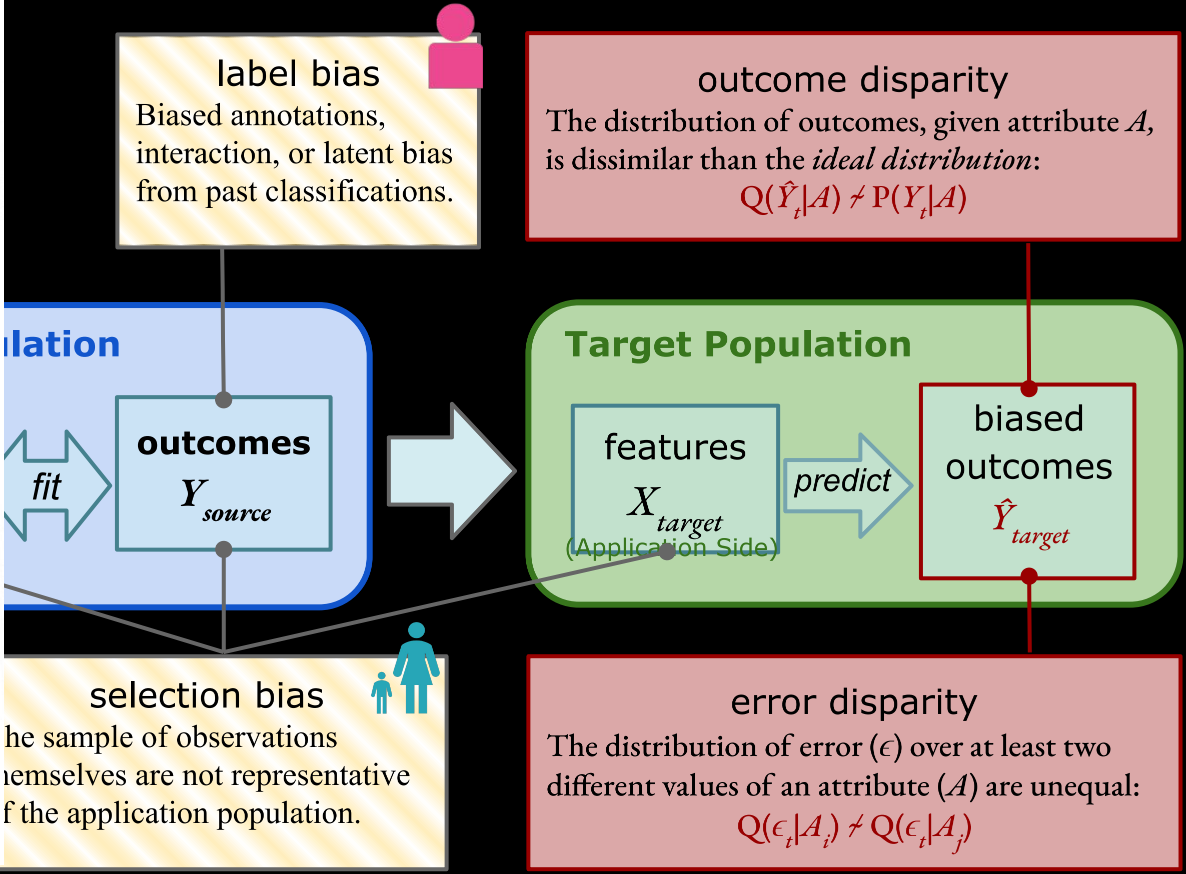
error disparity

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 - 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*:

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

selection bias

he sample of observations hemselves are not representative f the application population.

error disparity

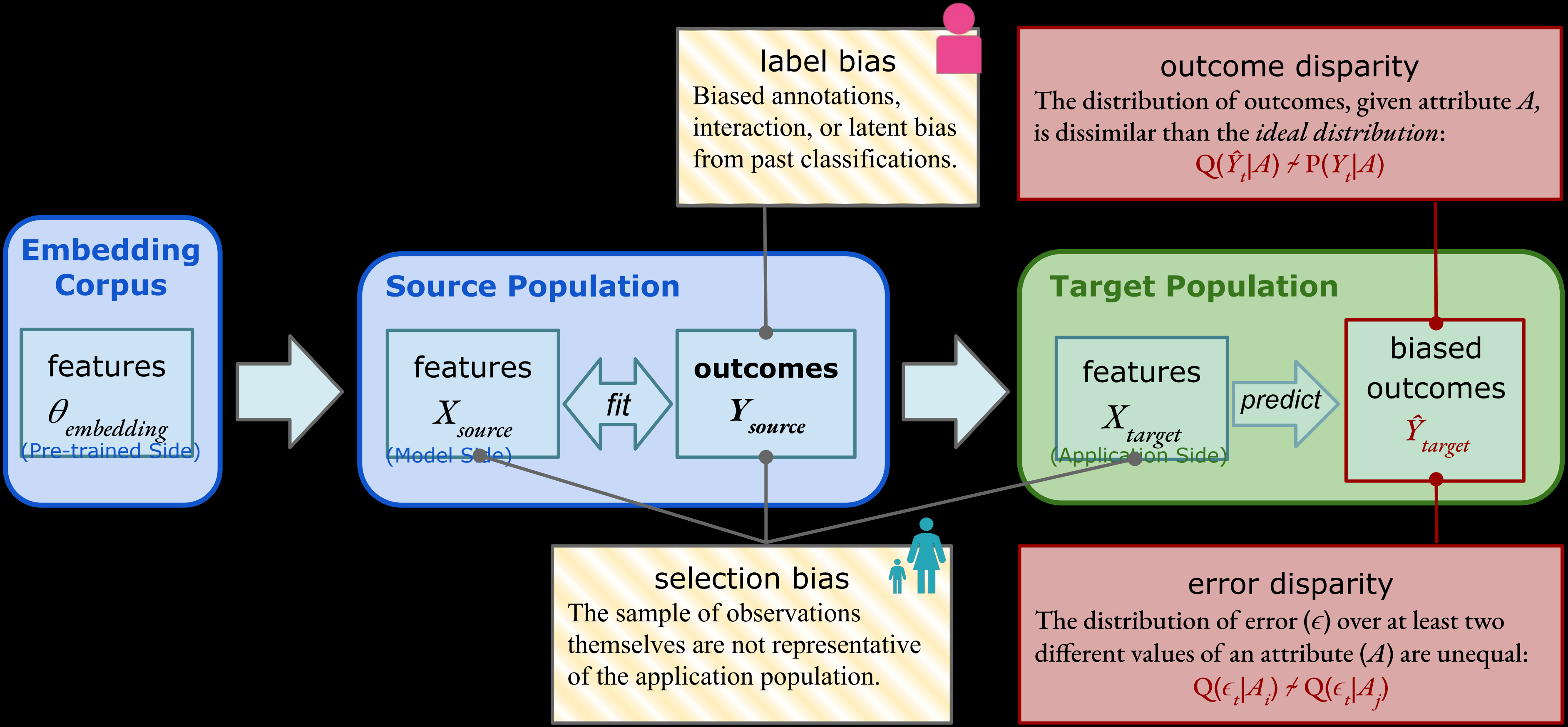
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:

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

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*:

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

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

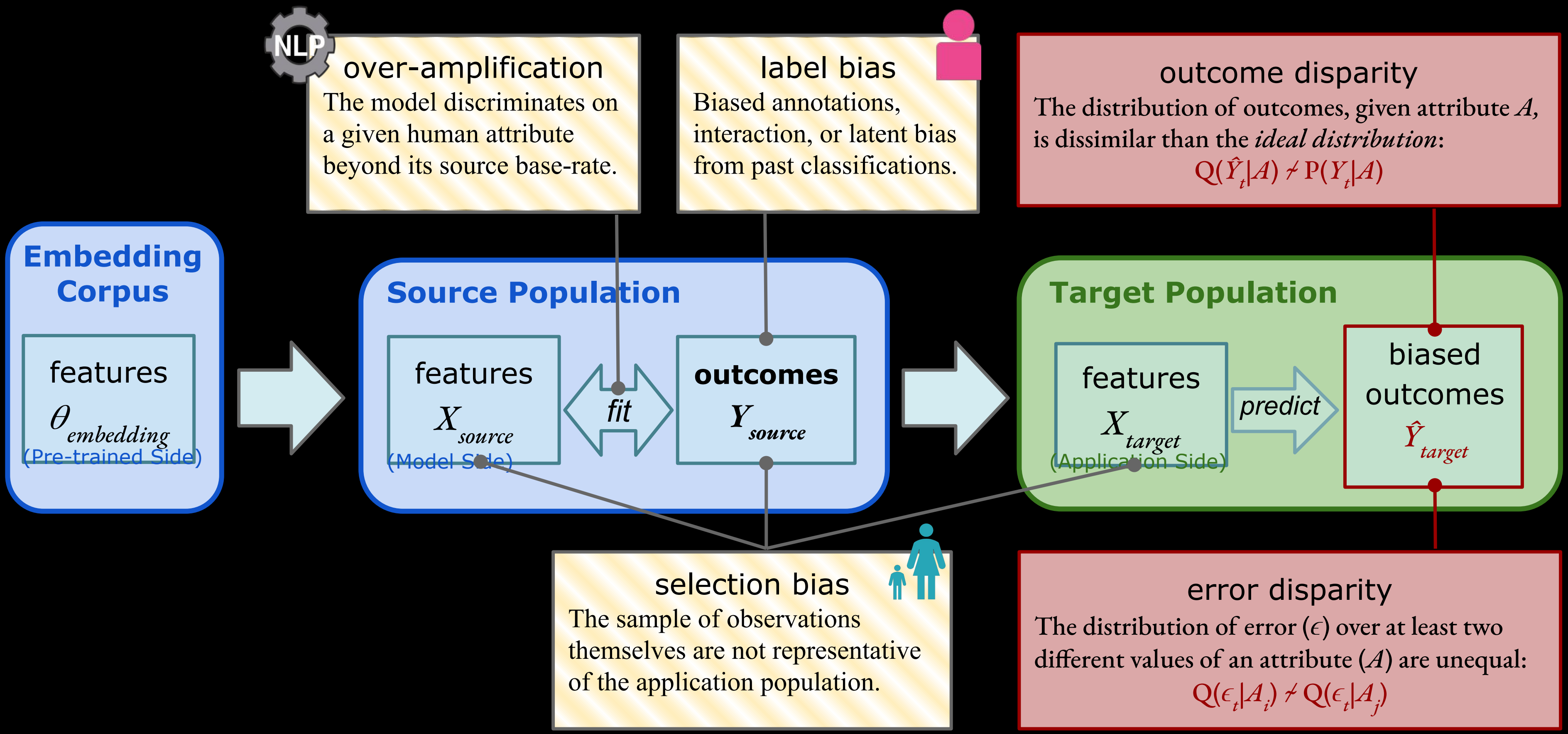
The sample of observations themselves are not representative of the application population.

error disparity

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

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

Overamplification



over-amplification

The model discriminates on a given human attribute beyond its source base-rate.

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

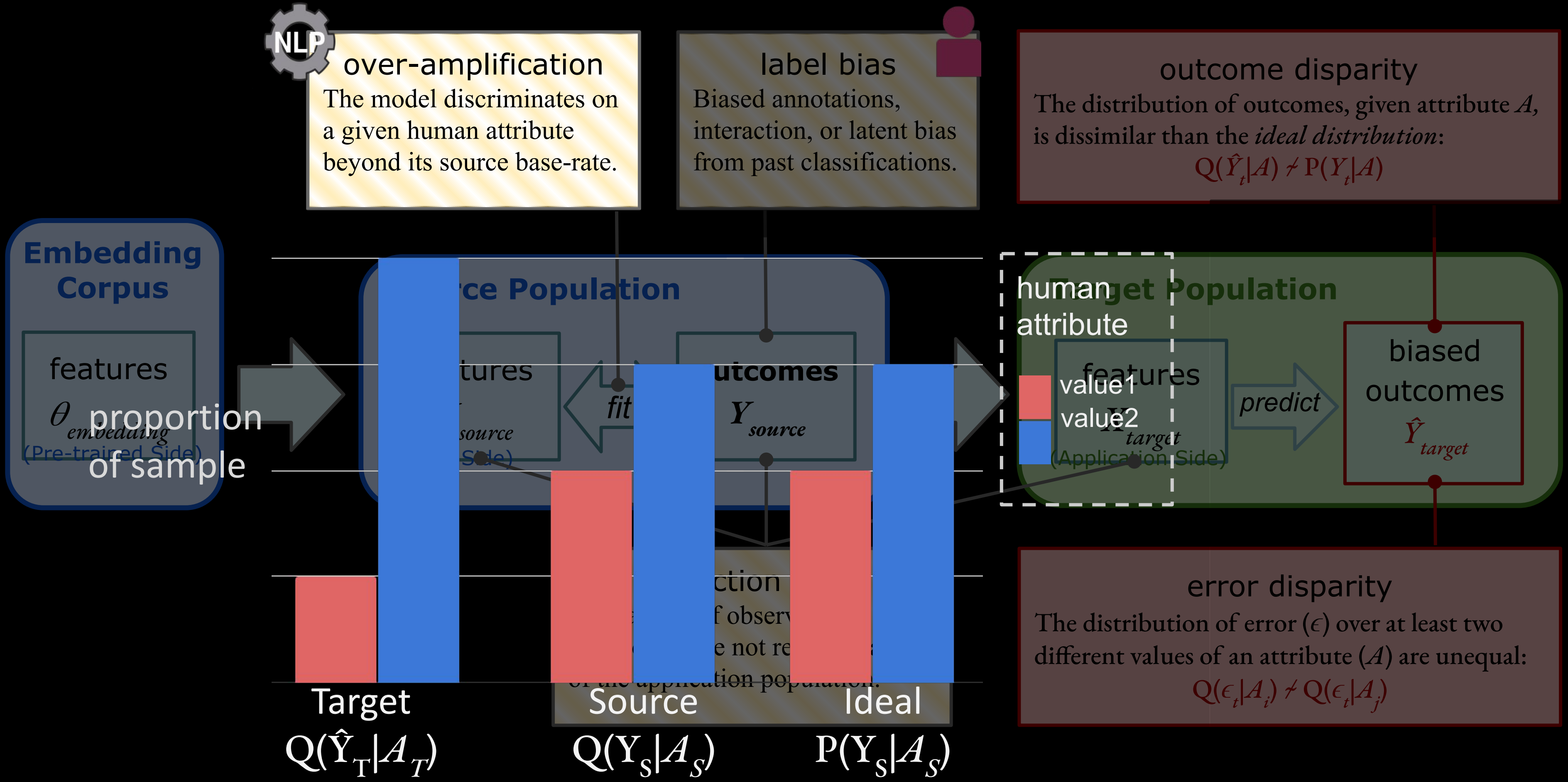
The sample of observations themselves are not representative of the application population.

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



over-amplification

The model discriminates on a given human attribute beyond its source base-rate.

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*:

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

error disparity

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

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

ction

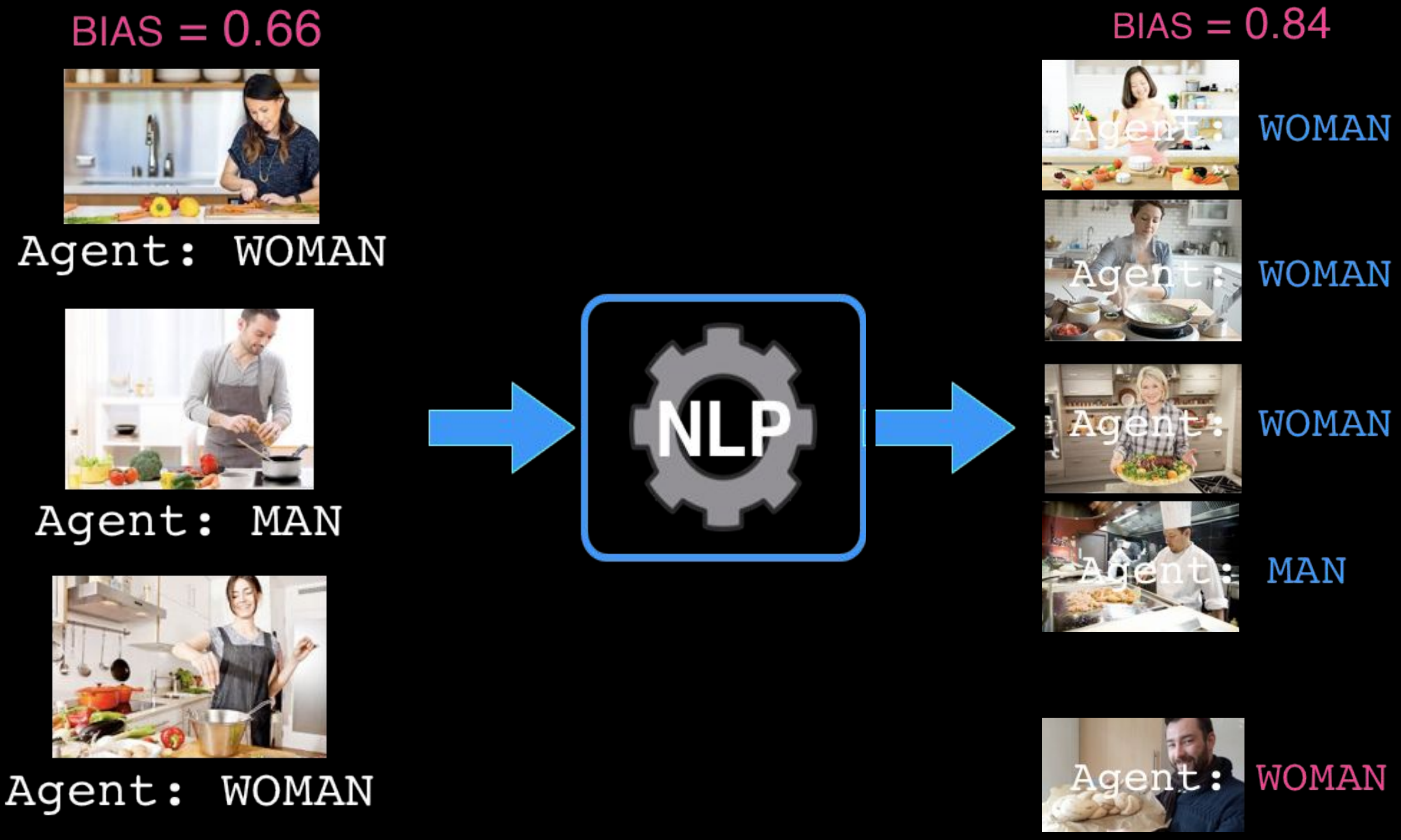
f observ

e not re

of the application population.

Zhao et al. (ACL 2015)

Overamplifiction - Model Amplifies Bias



BIAS = 0.66

Agent: WOMAN

Agent: MAN

Agent: WOMAN

BIAS = 0.84

Agent: WOMAN

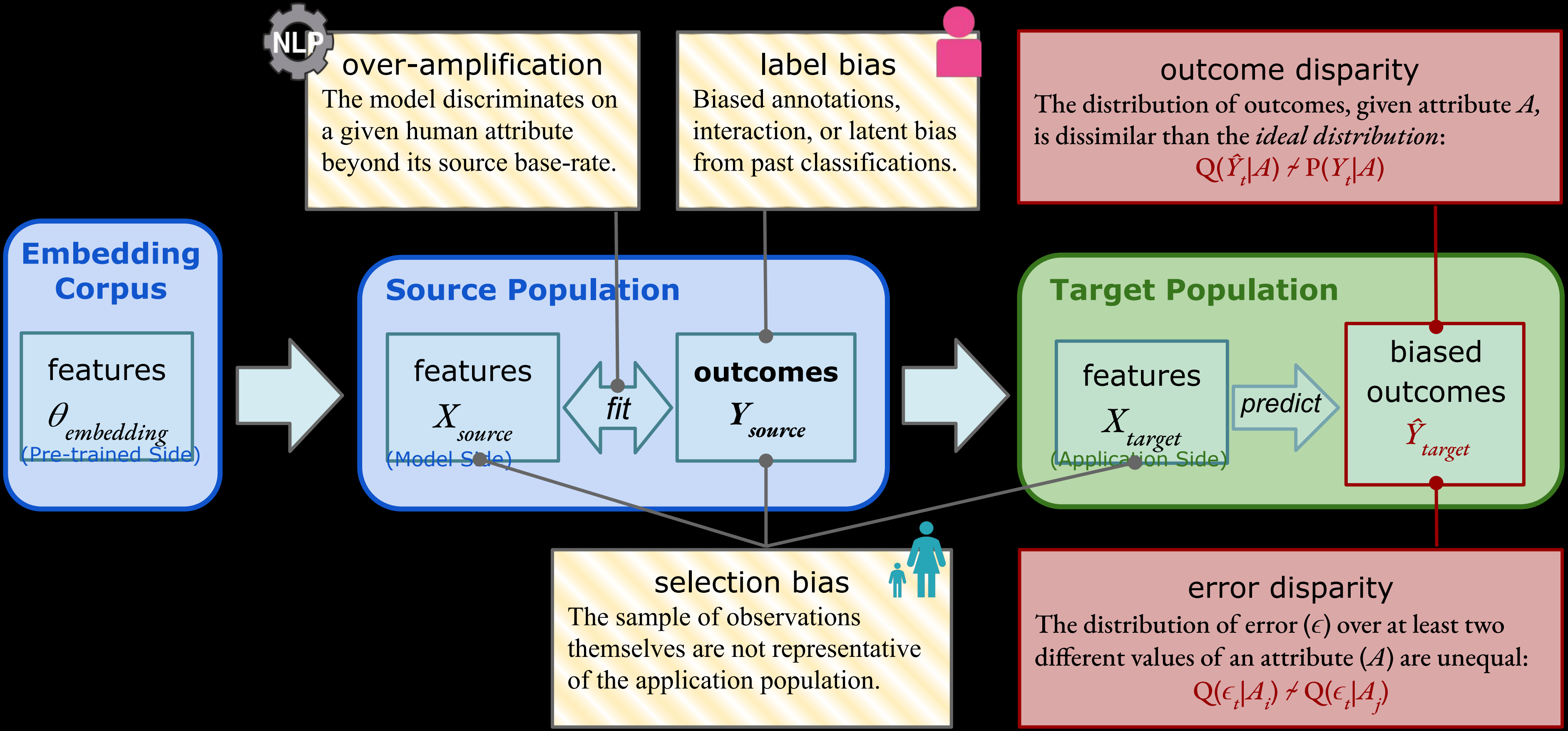
Agent: WOMAN

Agent: WOMAN

Agent: MAN

Agent: WOMAN

Overamplification



NLP

over-amplification

The model discriminates on a given human attribute beyond its source base-rate.

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*:

$$Q(\hat{Y}_t|A) \not\sim P(Y_t|A)$$

Embedding Corpus

features

$\theta_{embedding}$

(Pre-trained Side)

Source Population

features

X_{source}

fit

outcomes

Y_{source}

(Model Side)

Target Population

features

X_{target}

predict

biased outcomes

$\hat{Y}_{target}$

(Application Side)

selection bias

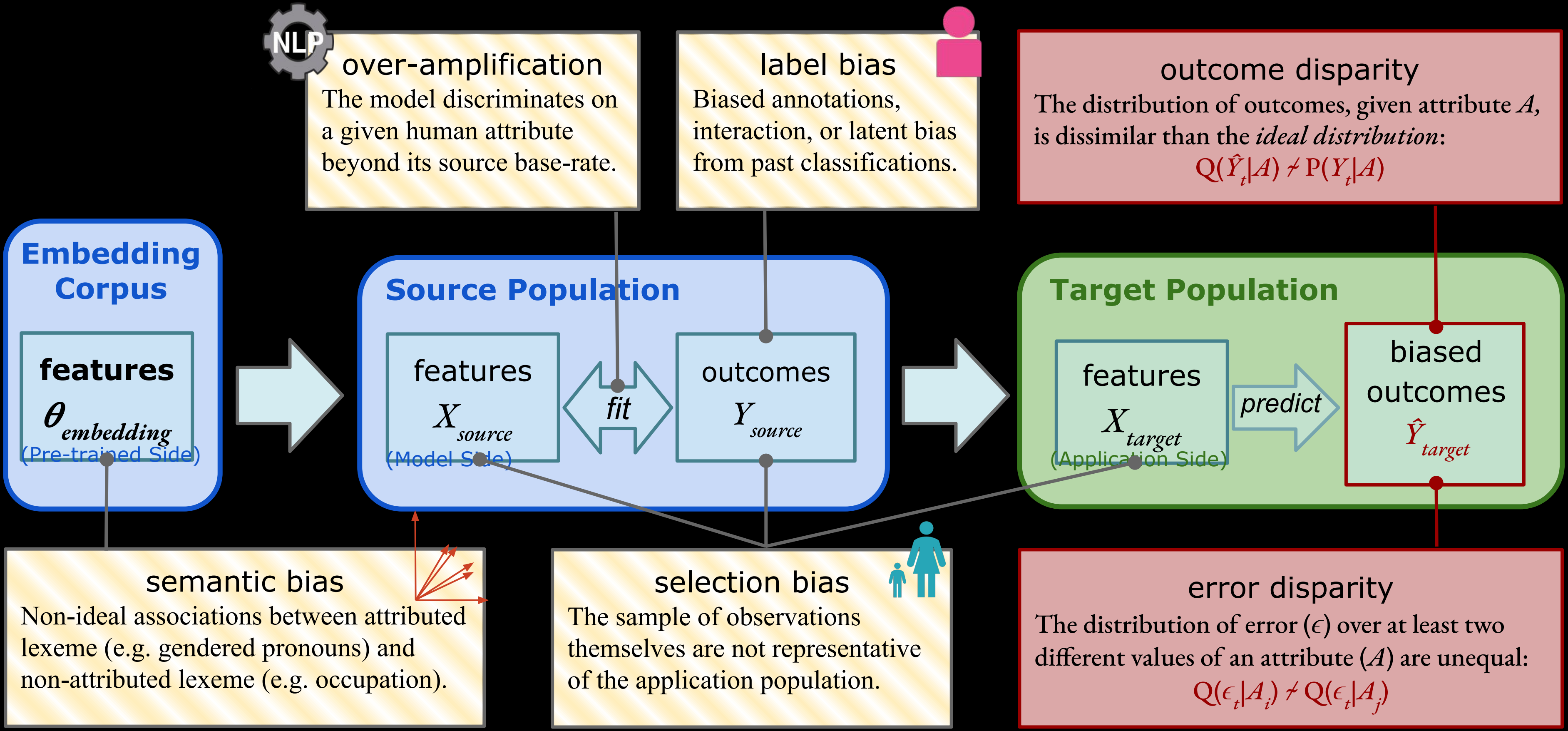
The sample of observations themselves are not representative of the application population.

error disparity

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

$$Q(\epsilon_t|A_i) \not\sim Q(\epsilon_t|A_j)$$

Semantic Bias



over-amplification
The model discriminates on a given human attribute beyond its source base-rate.

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

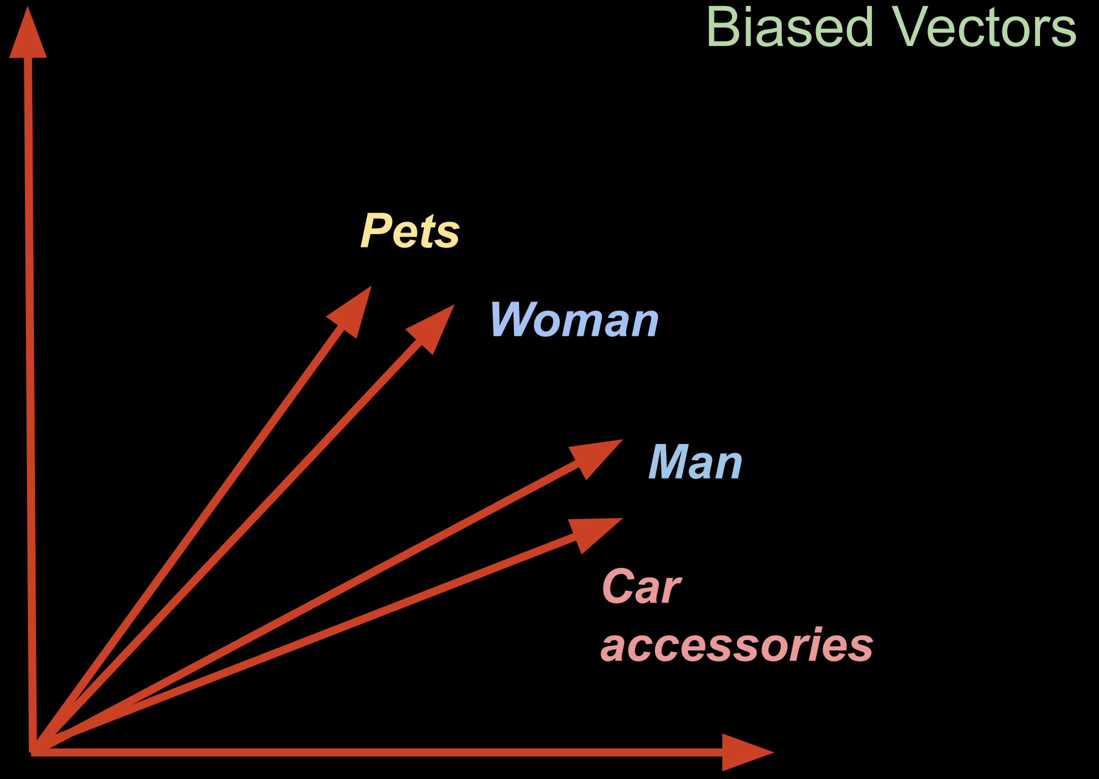
outcome disparity
The distribution of outcomes, given attribute *A*, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \not\sim P(Y_t|A)$

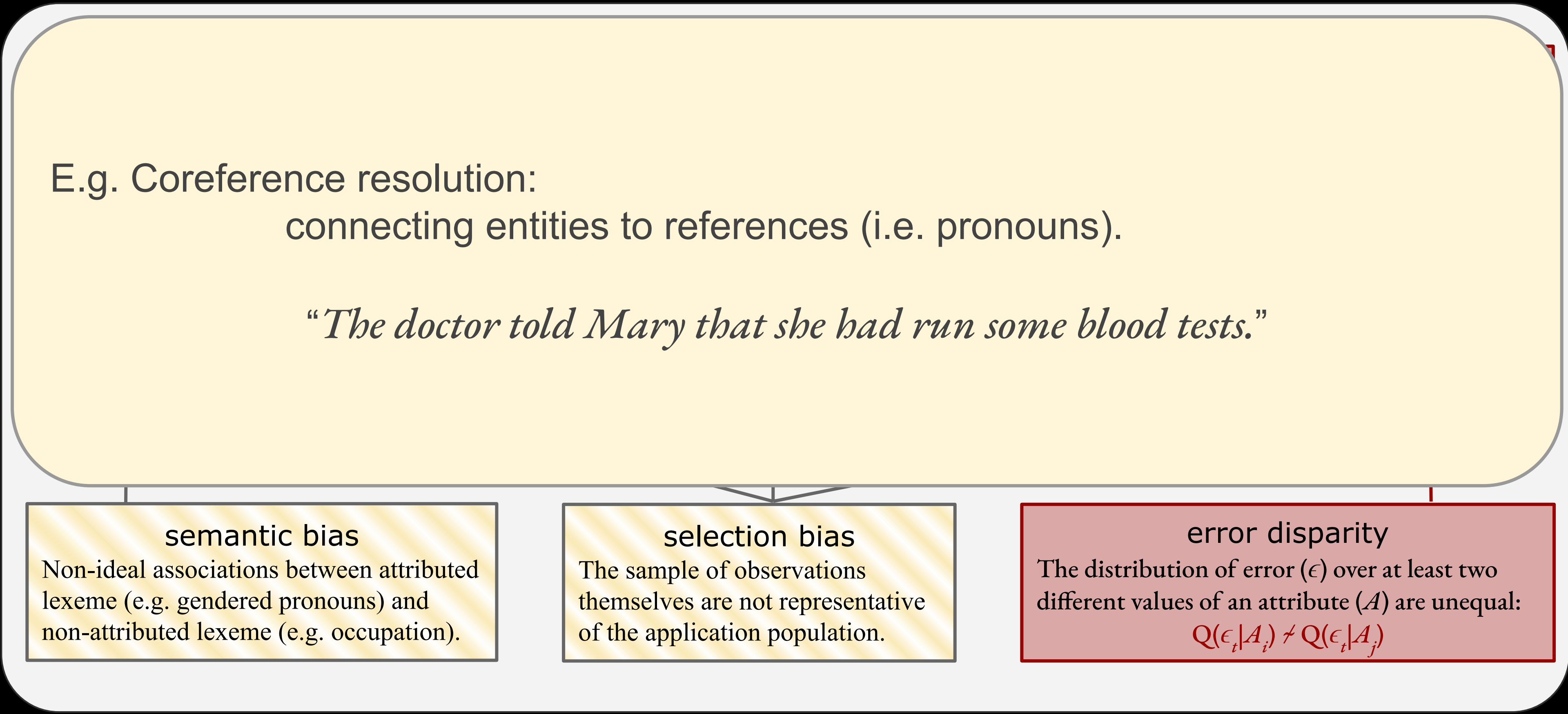
semantic bias
Non-ideal associations between attributed lexeme (e.g. gendered pronouns) and non-attributed lexeme (e.g. occupation).

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:
$Q(\epsilon_t|A_i) \not\sim Q(\epsilon_t|A_j)$

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.

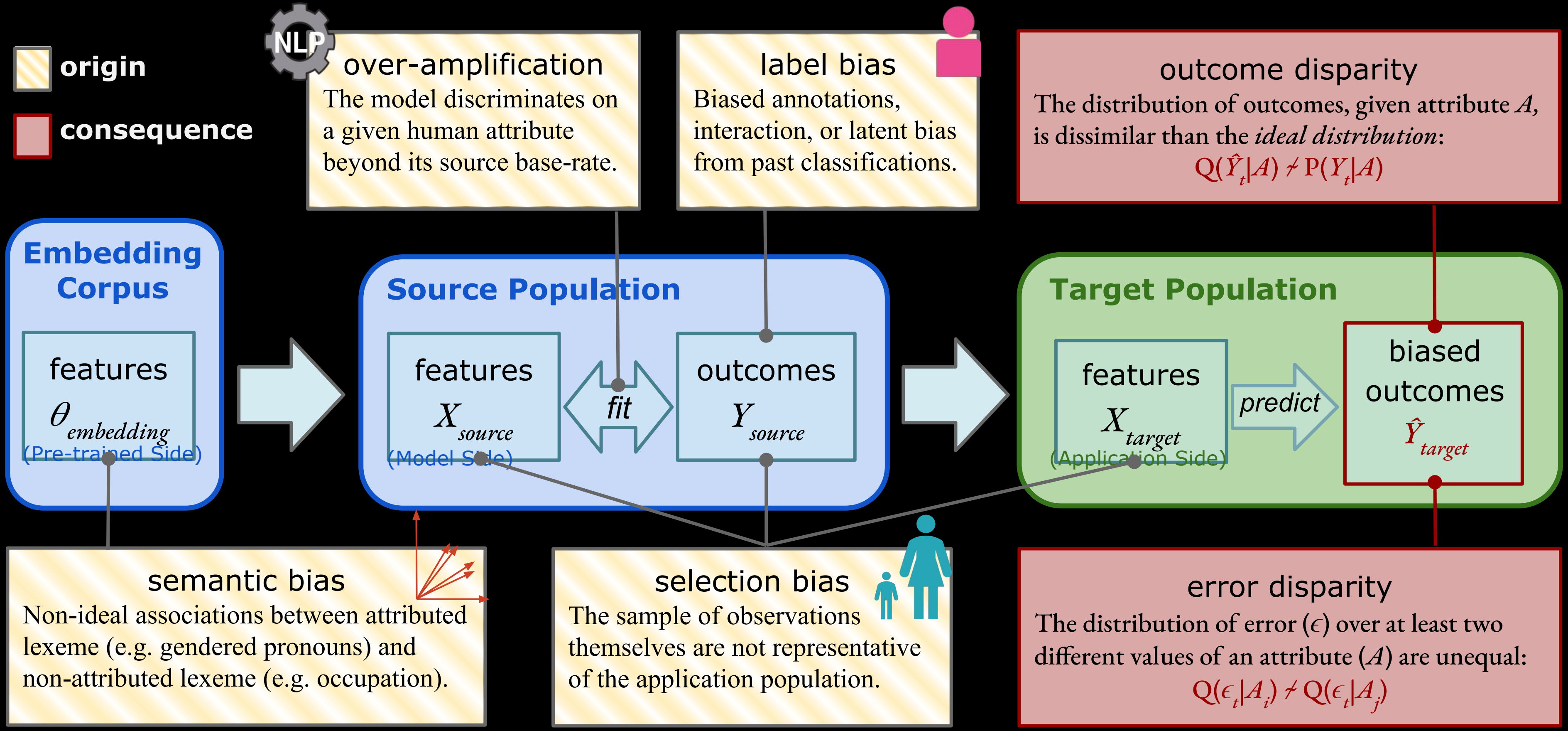
error disparity

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



▨ origin

🟥 consequence

over-amplification
The model discriminates on a given human attribute beyond its source base-rate.

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}
fit
outcomes Y_{source}
(Model Side)

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

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

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
NLP models	**Overamplification**	Synthetically match distributions, add outcome disparity to cost function
embeddings	**Semantic Bias**	Use above techniques and re-train embeddings

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

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:
  - Informed consent -- let participants know and opportunity to opt-in/-out
  - Do not share / secure storage
  - *Federated learning* -- obfuscate to the point of preserving privacy
  - Transparency in information targeting

    "You are being shown this ad because ..."

Ethics in NLP

Bias

Privacy

Ethical Research

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 Belmount 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.