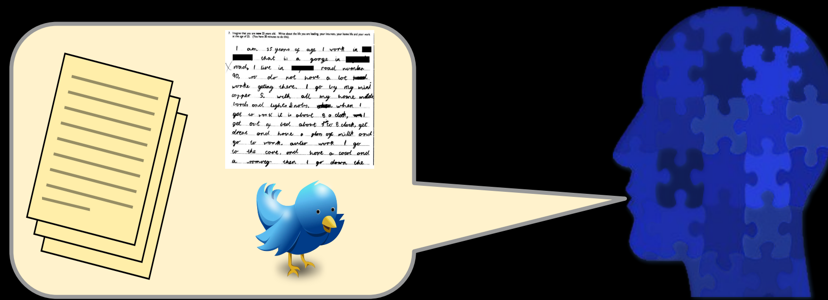
Human-Centered Natural Language Processing

CSE392 - Spring 2019
Special Topic in CS

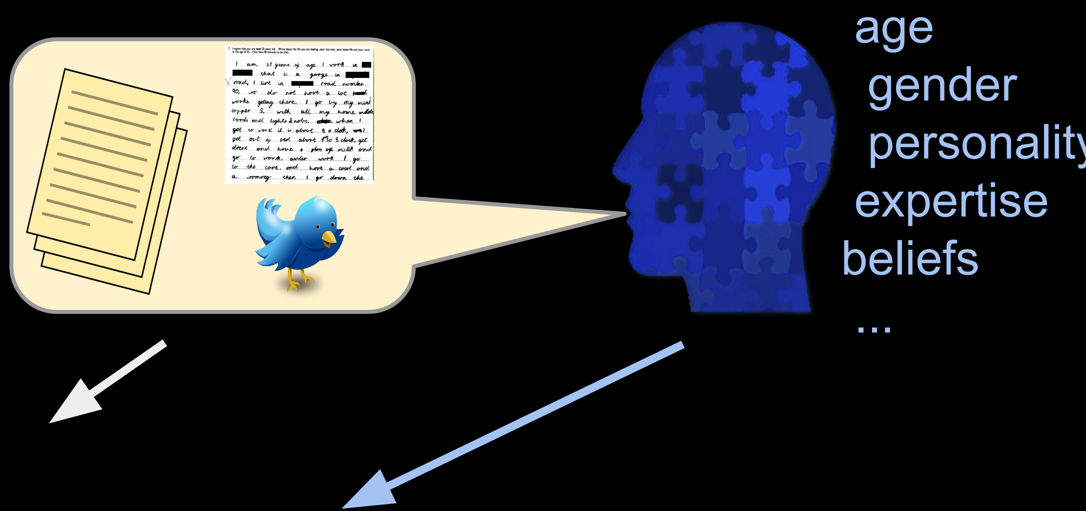
The "Task" of human-centered NLP



Most NLP Tasks. E.g.

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

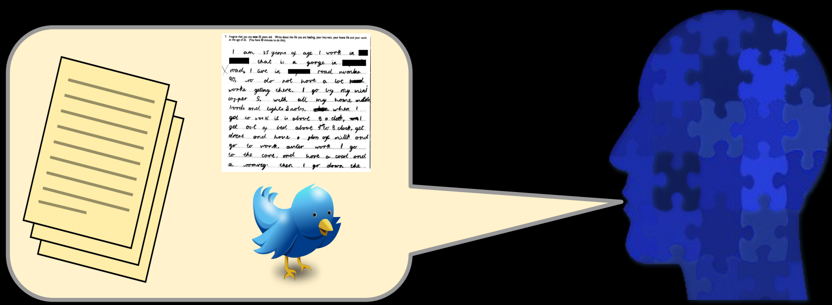
The "Task" of human-centered NLP



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



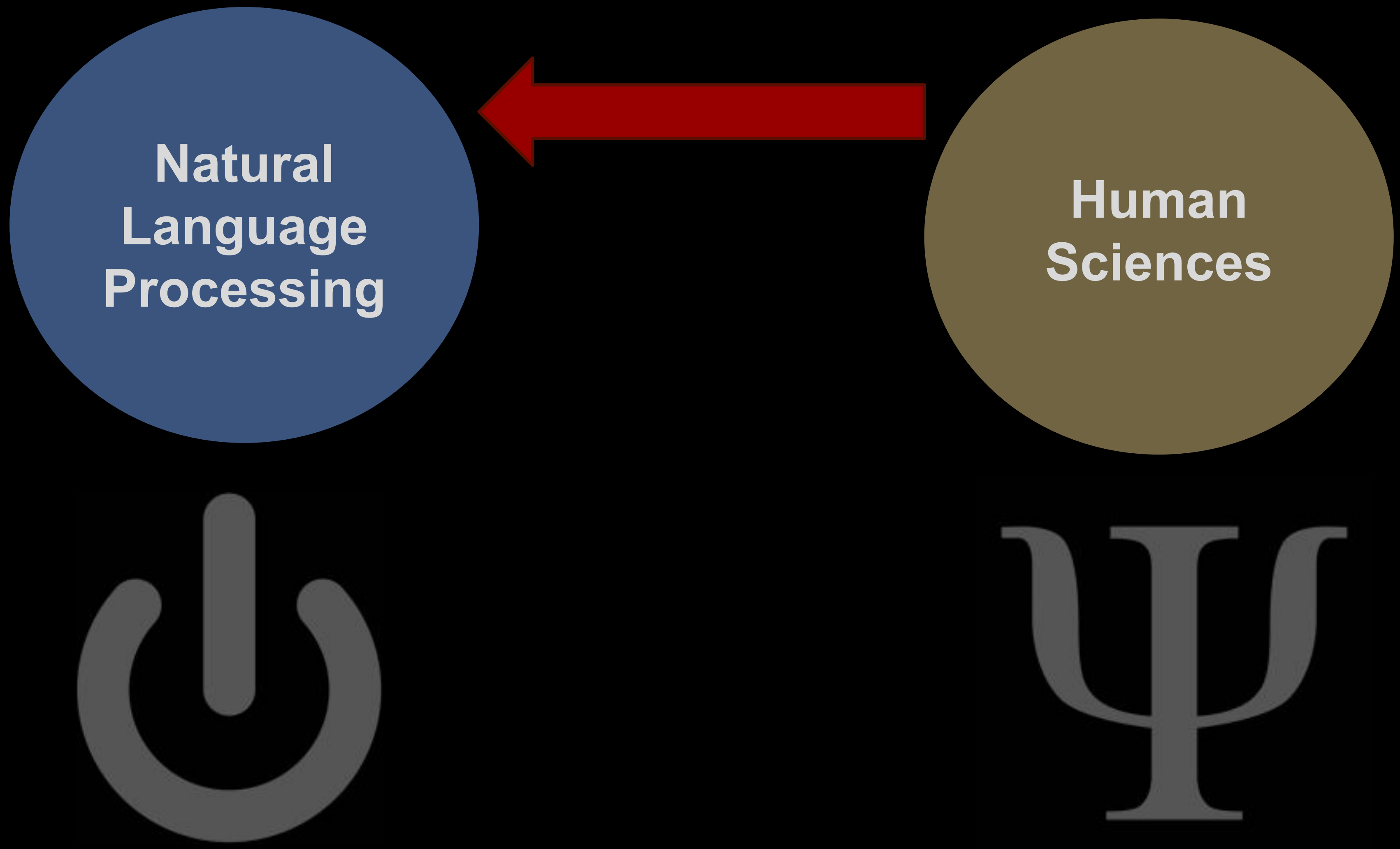
age
gender
personality
expertise
beliefs
...

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
Human 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.**

Problem

Natural language is written by **people.**

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

Practical Implication:

- Our NLP models are biased

Proble

Natural language is w

People have different beliefs, b
vocabularies, preferences, knowledge

Practical Implication:

- Our NLP models are biased


"The WSJ Effect"

Tagging Performance Correlates with Author Age

Dirk Hovy[1] and Anders Søgaard[1]
Center for Language Technology
University of Copenhagen, Denmark

Problem

Natural language is written by **people.**

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

Practical Implication:

- Our NLP models are biased
- Sometimes our predictions are invalid

Task: PTSD or Depression?
AUC = 0.80

…ounds, styles,

…owledge, personalities, …

…n:

…LP models are biased

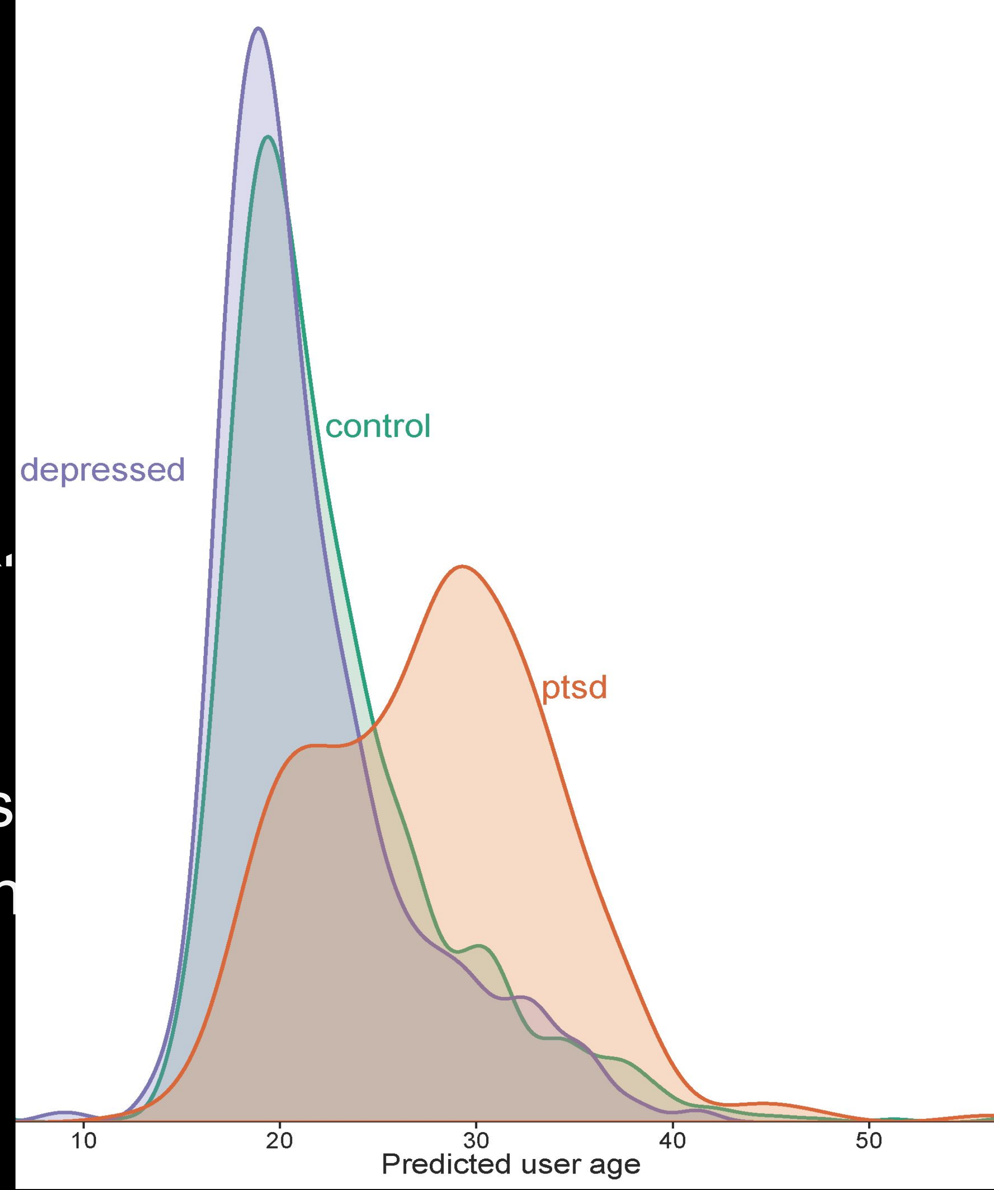
Sometimes our predictions are invalid

Task: PTSD or Depression?
AUC = 0.80

on:

NLP models are bias

Sometimes our prediction



Problem

Natural language is written by **people.**

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

Practical Implication:

- Our NLP models are biased
- Sometimes our predictions are invalid

Put language in the context of the person who wrote it
=> Greater Accuracy

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)

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)

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)

Human Factors

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

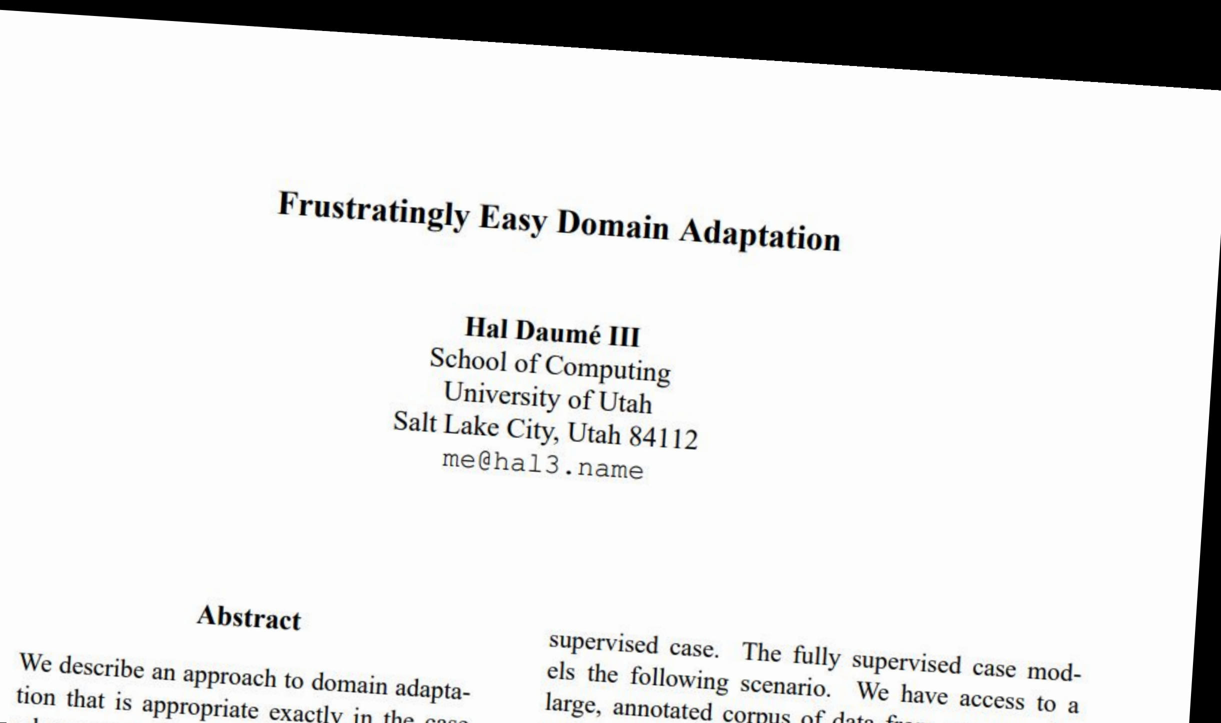
E.g.

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

Adaptation Approach: Domain Adaptation

Features for: source target

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



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 adapta-
tion that is appropriate exactly in the case

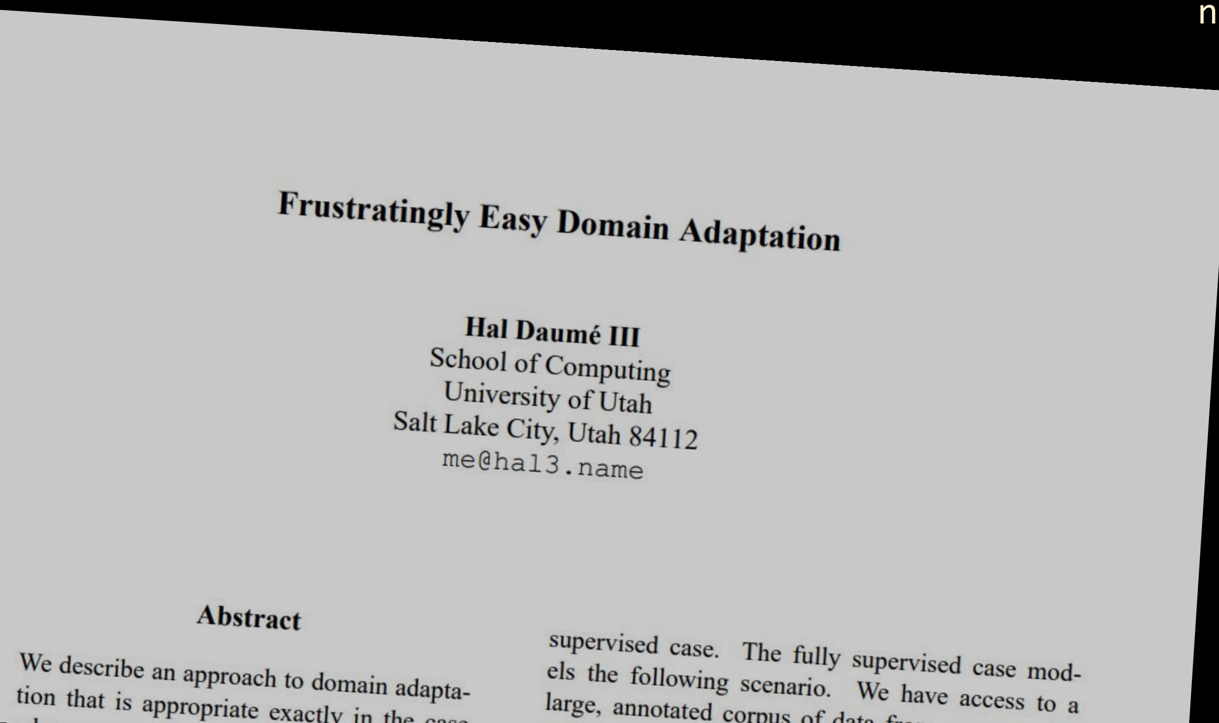
supervised case. The fully supervised case mod-
els the following scenario. We have access to a
large, annotated corpus of data from

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)
```



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 adapta-
tion that is appropriate exactly in the

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

Adaptation Approach: Domain Adaptation

Features for: source target

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

```
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

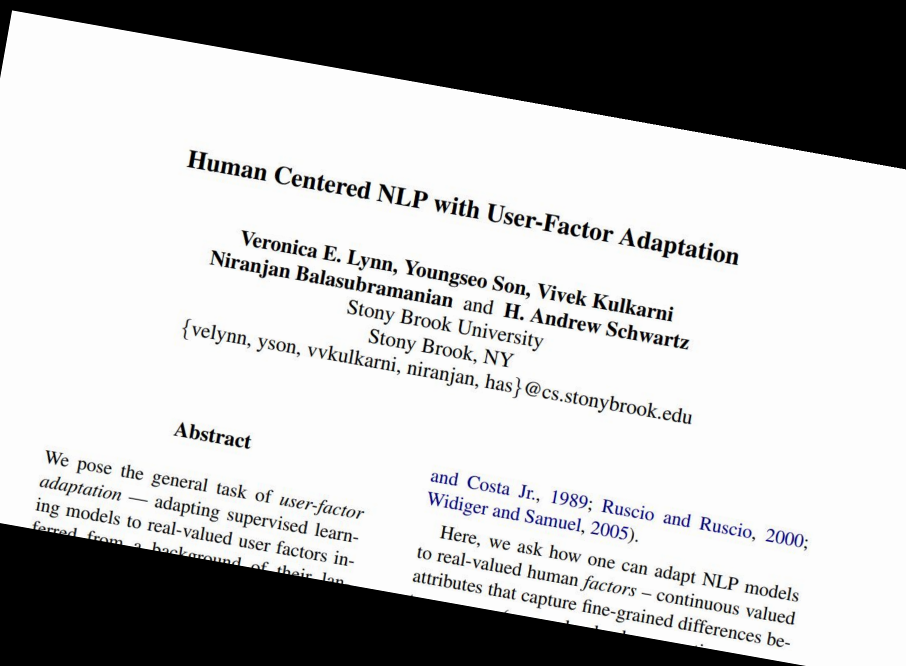
Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@hal3.name

Abstract

We describe an approach to domain adapta-
tion that is appropriate exactly in the case

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

Adaptation Approach: Factor Adaptation



Human Centered NLP with User-Factor Adaptation

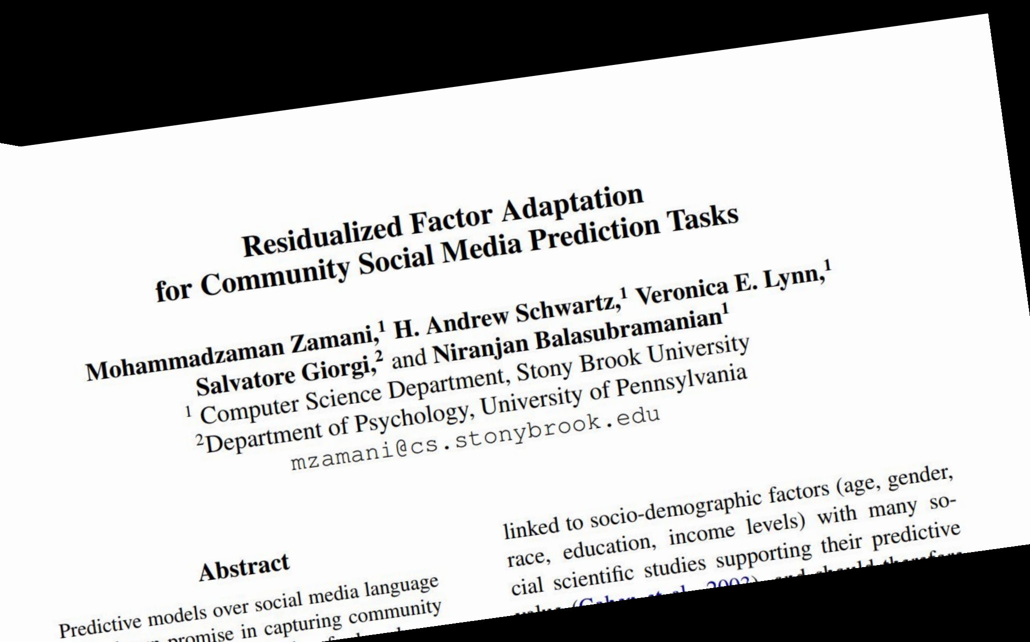
Veronica E. Lynn, Youngseo Son, Vivek Kulkarni
Niranjan Balasubramanian and H. Andrew Schwartz
Stony Brook University
Stony Brook, NY
{velynn, yson, vvkulkarni, niranjan, has}@cs.stonybrook.edu

Abstract

We pose the general task of *user-factor adaptation* — adapting supervised learning models to real-valued user factors in-

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

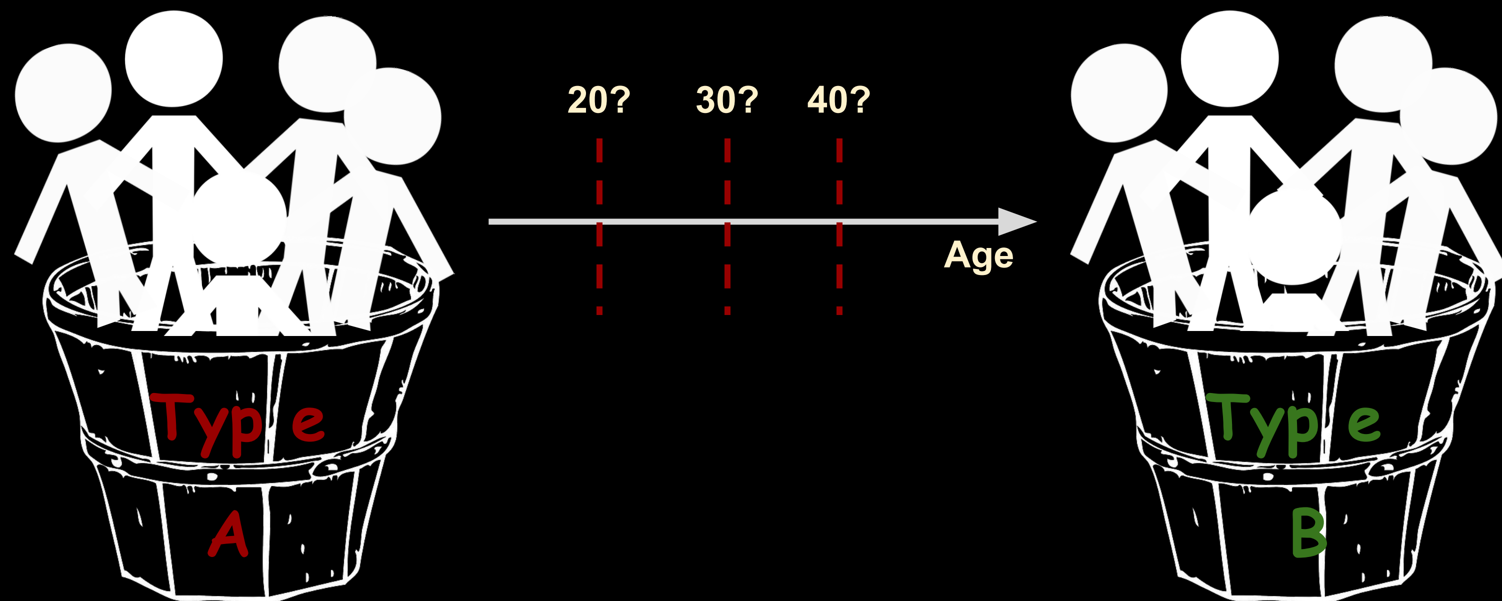
typically requires putting people into discrete bins

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

(Haslam et al., 2012)

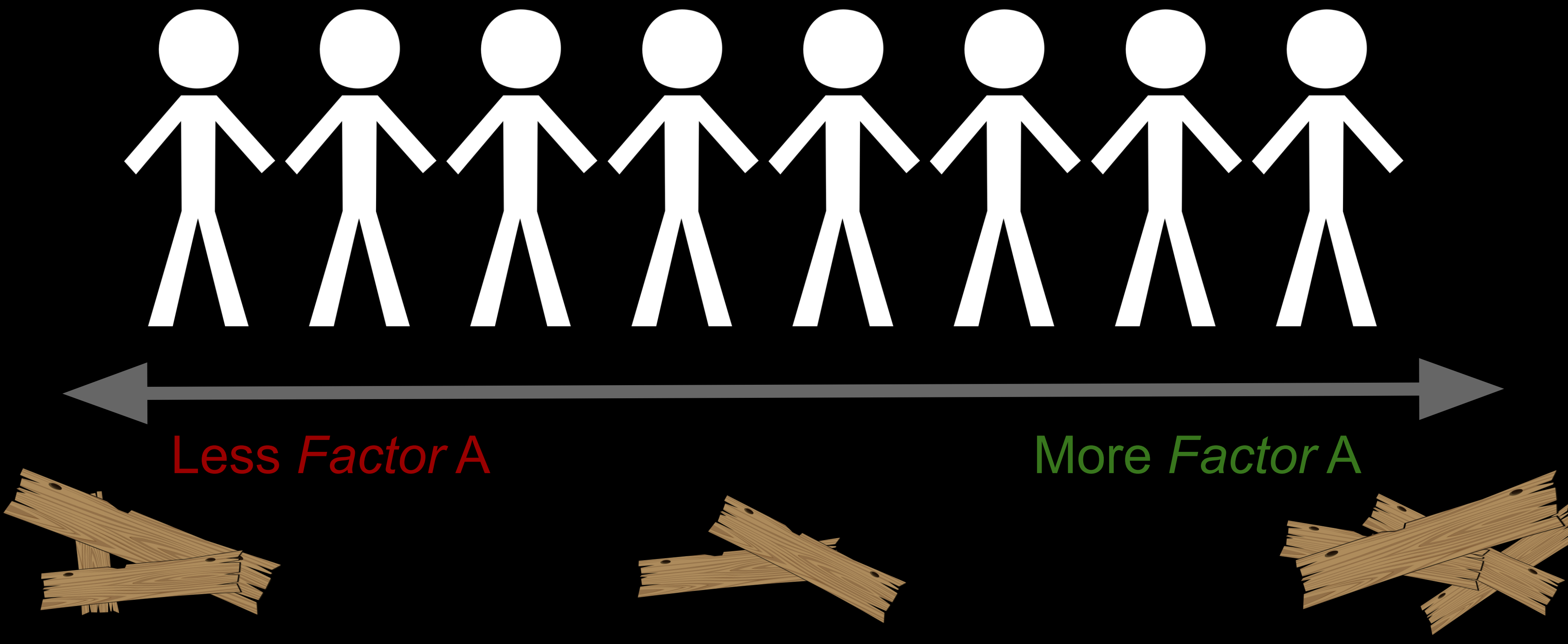
"most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]"

(Haslam et al., 2012)

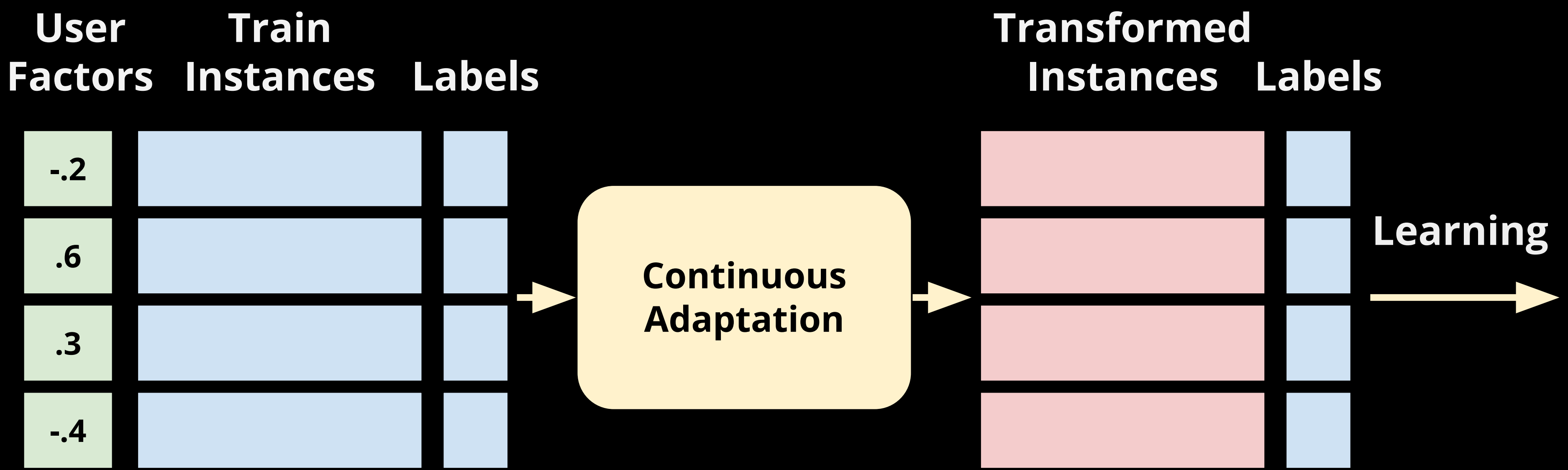


"most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]"

(Haslam et al., 2012)

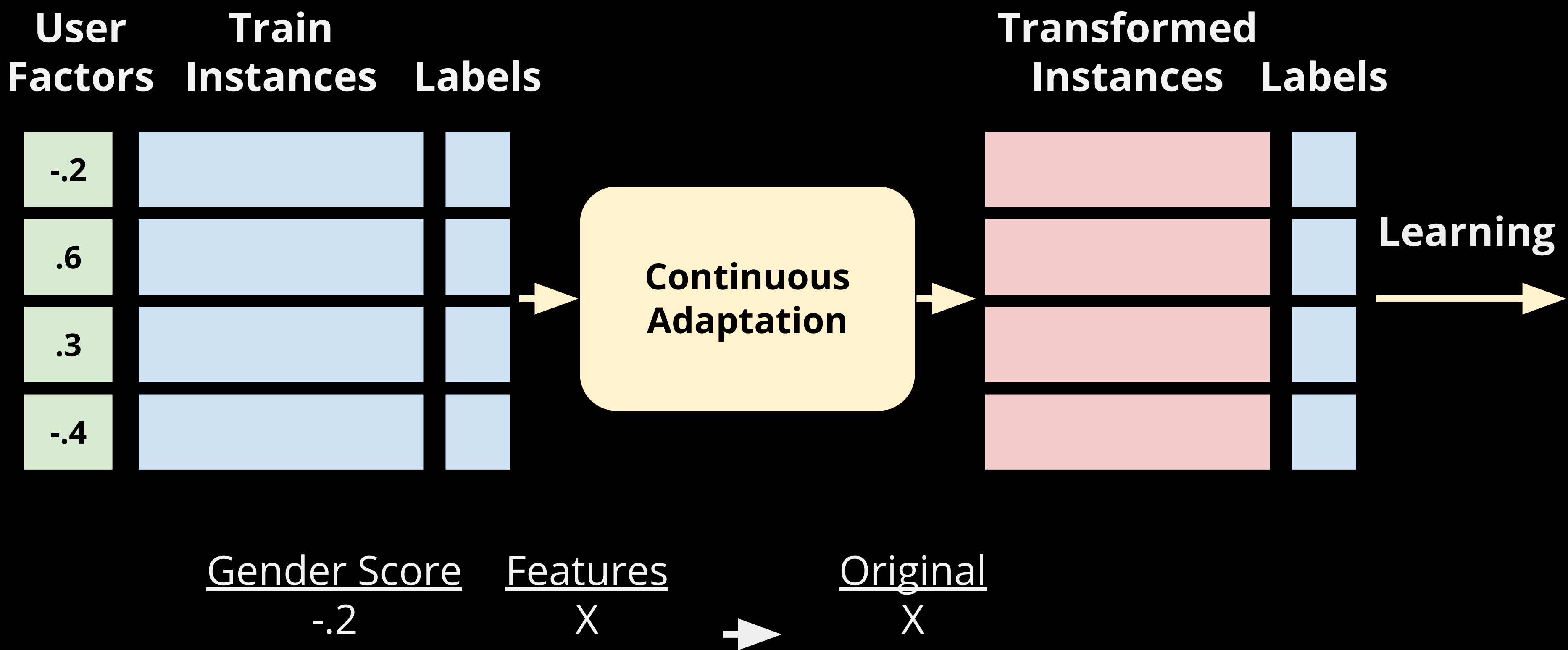


Our Method: Continuous Adaptation



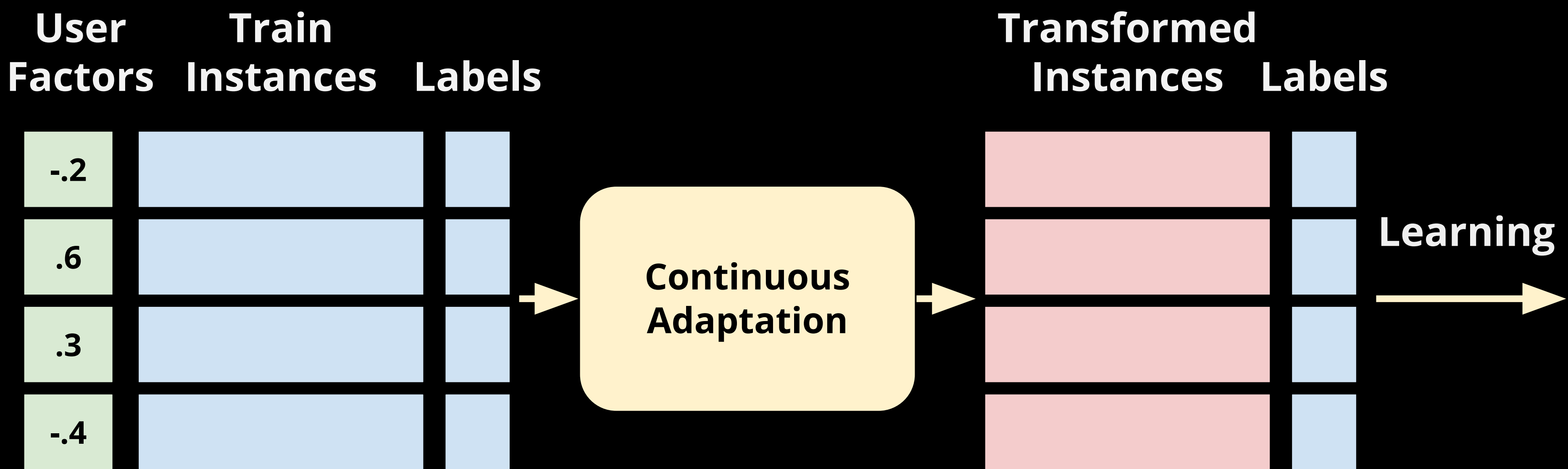
(Lynn et al., 2017)

Our Method: Continuous Adaptation



(Lynn et al., 2017)

Our Method: Continuous Adaptation



<u>Gender Score</u> -.2 <u>Features</u> X → <u>Original</u> X <u>Gender Copy</u> *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 $n(k+1)$.

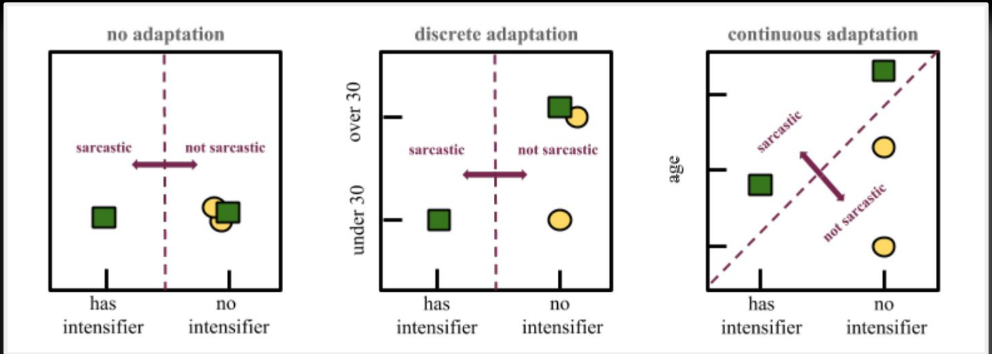
(Lynn et al., 2017)

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	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0} \rangle$
User 2	F_2	$\langle \mathbf{x}, \mathbf{0}, \mathbf{x}, \mathbf{0}, \cdots, \mathbf{0} \rangle$
User 3	F_1, F_3	$\langle \mathbf{x}, \mathbf{x}, \mathbf{0}, \mathbf{x}, \cdots, \mathbf{0} \rangle$
User 4	F_k	$\langle \mathbf{x}, \mathbf{0}, \mathbf{0}, \cdots, \mathbf{0}, \mathrm{x} \rangle$

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

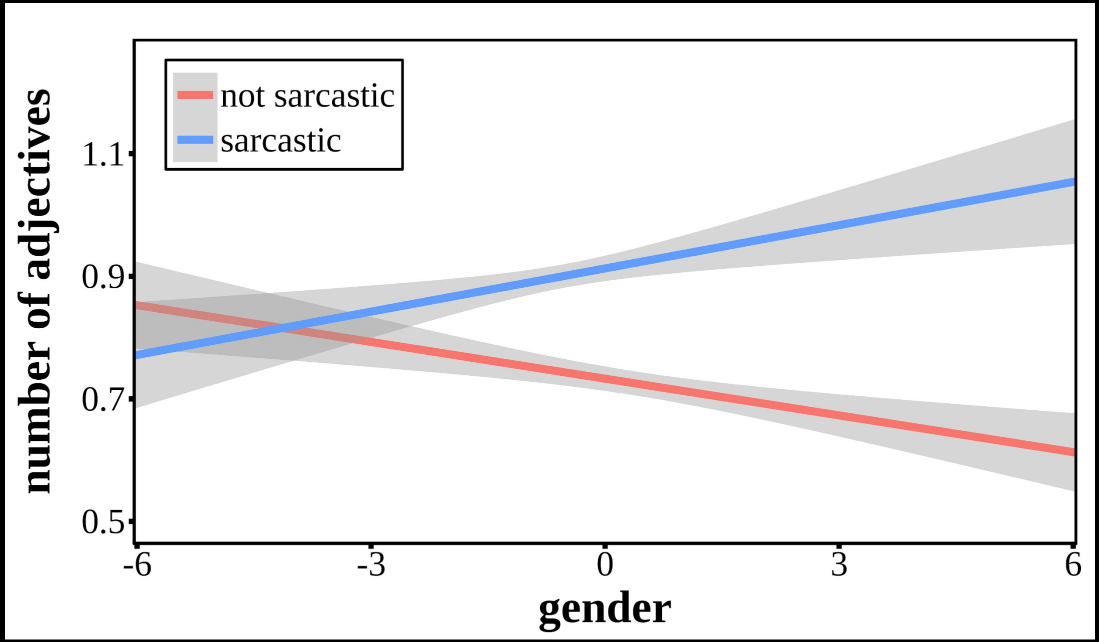
Adaptation improves over unadapted baselines (Lynn et al., 2017)

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

<u>Women</u>
more adjectives→sarcasm

<u>Men</u>
more adjectives→no sarcasm



more “male”

more “female”

Problem

User factors are not always available.

Solution: User Factor Inference

past tweets

inferred factors

Known

Age (Sap et al. 2014)

Gender (Sap et al. 2014)

Personality (Park et al. 2015)

Latent

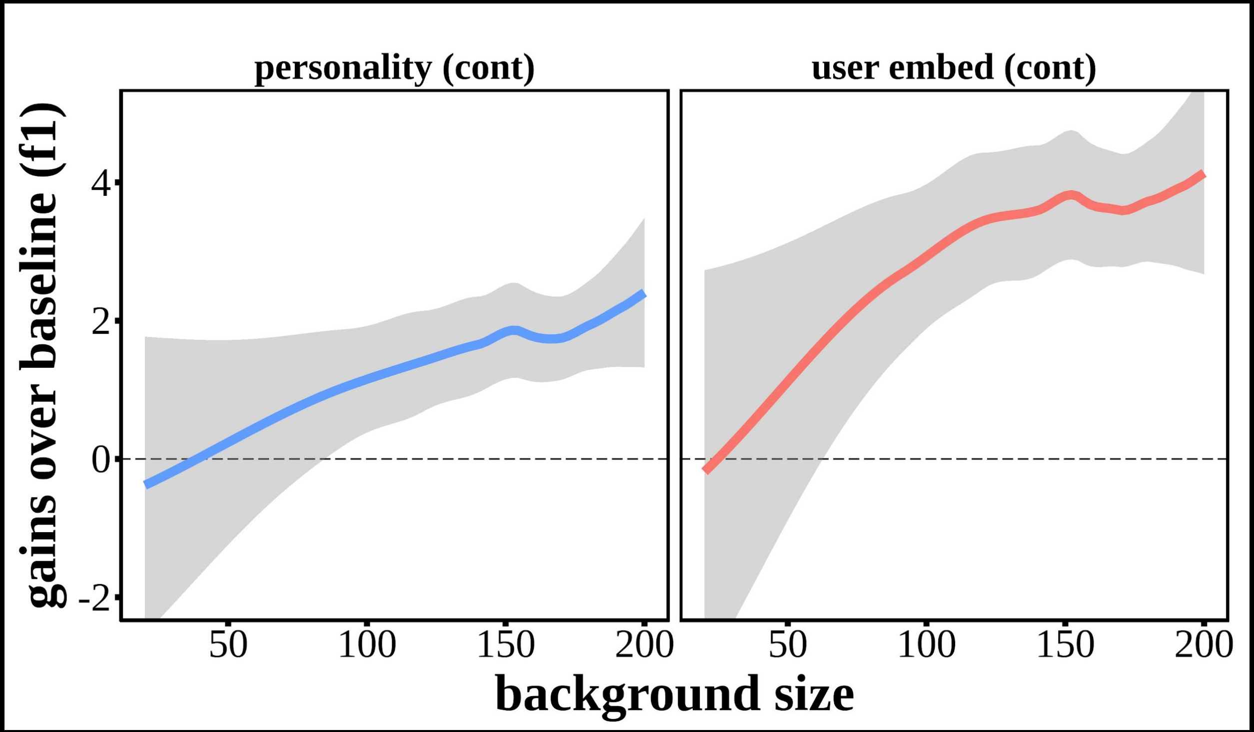
User Embeddings
(Kulkarni et al. 2017)

Word2Vec

TF-IDF

Background Size

Using more background tweets to infer factors produces larger gains



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)

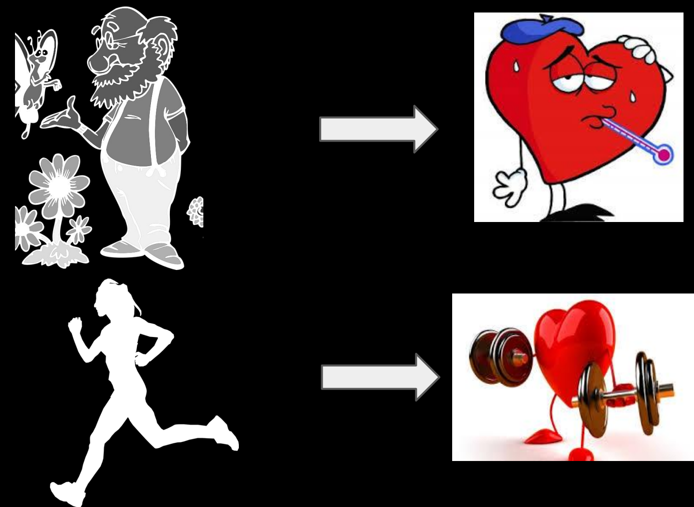
Approaches to Human Factor Inclusion

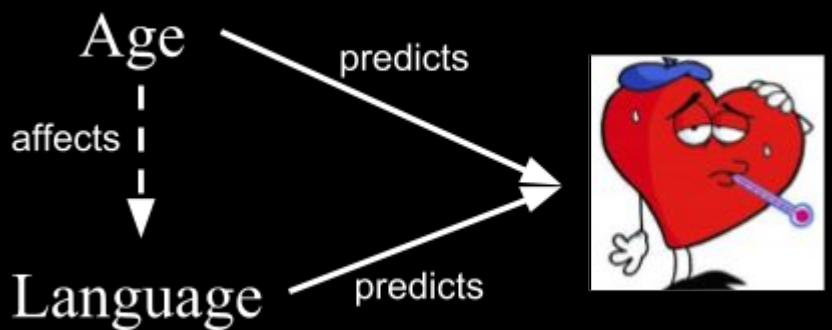
1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)

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

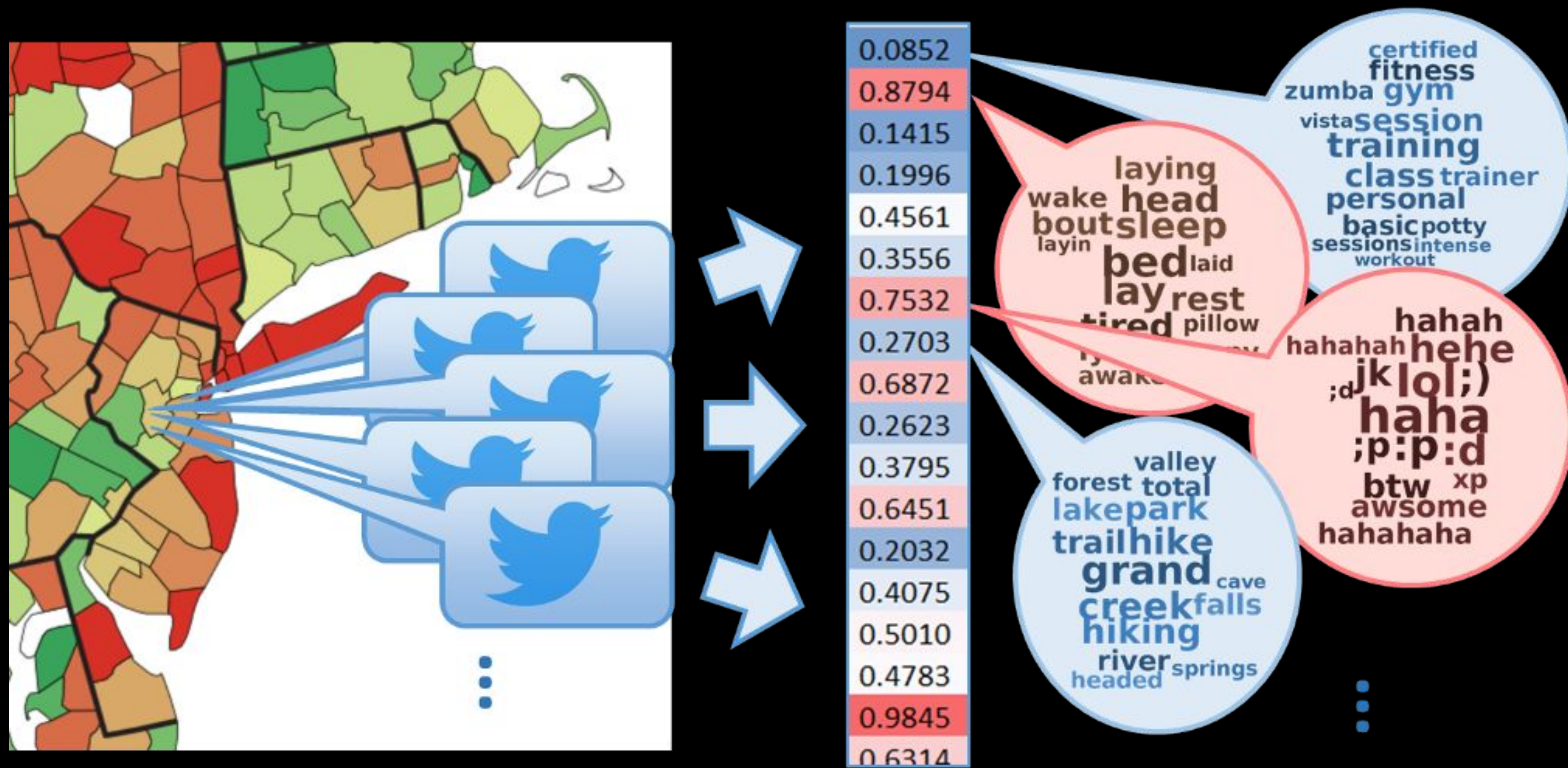
3. Bias Correction: Optimize so as not to pick up on unwanted relationships.
   (e.g. image captioner label pictures of men in kitchen as women)

Example 1: Individual Heart Disease

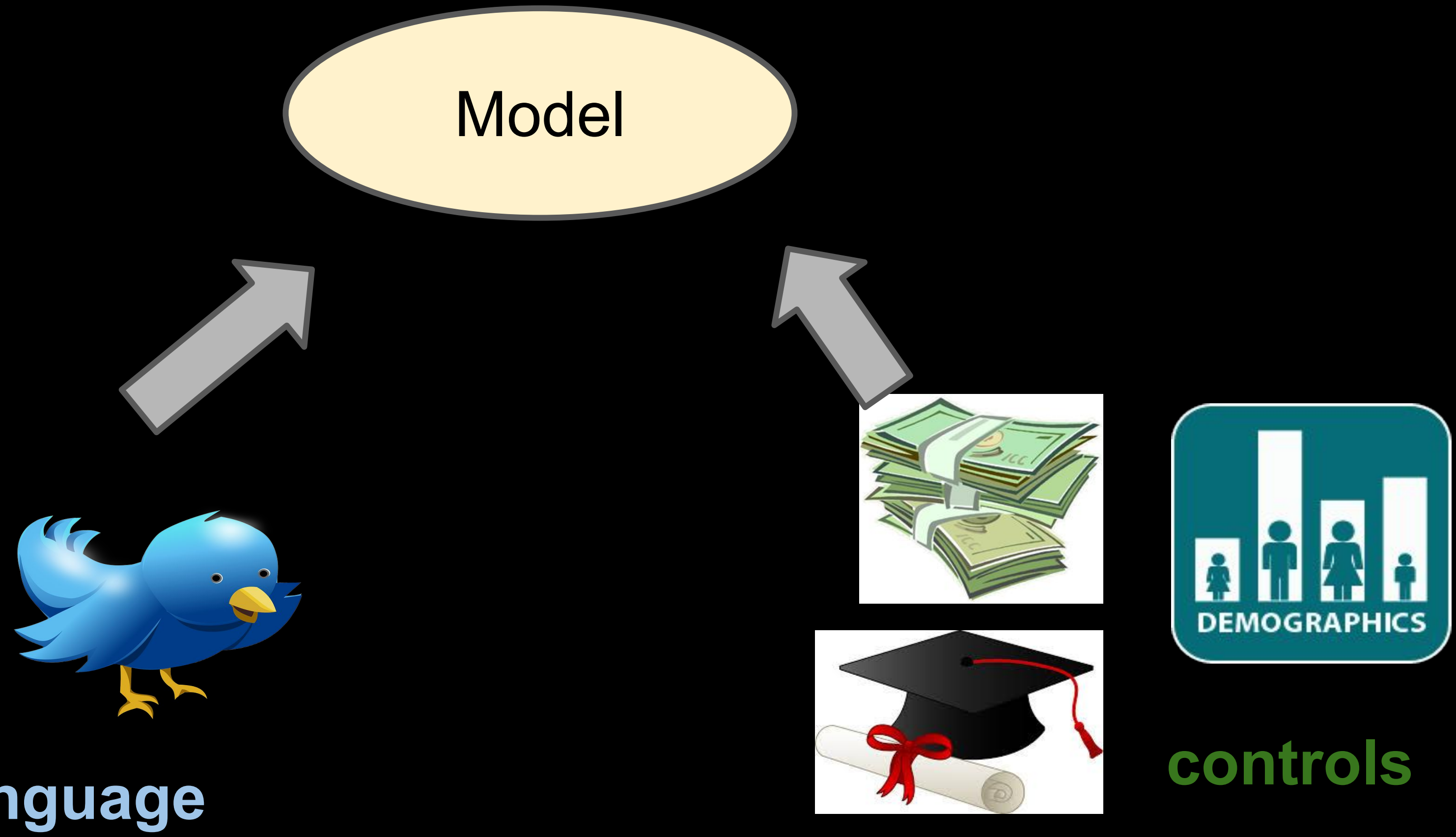




Example 2: Twitter Language + Socioeconomics



Additive (Residualized Control)



Additive (Residualized Control)

Challenges:

High-dimensional, sparse, and noisy.

few and well estimated

language

controls

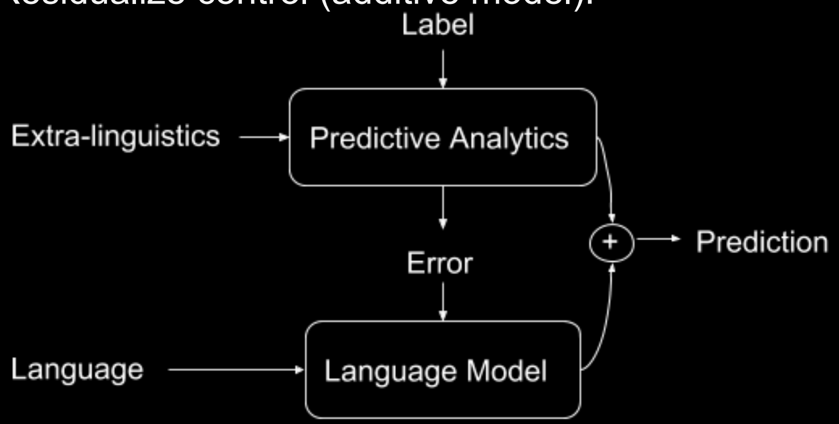
Additive (Residualized Control)

Effectively use both low dimensional control features and high-dimensional, noisy language features:

1. **Train a control model** using the control values
2. **Calculate the residual** error and consider it as the new label
3. **Train a language model over the new labels**

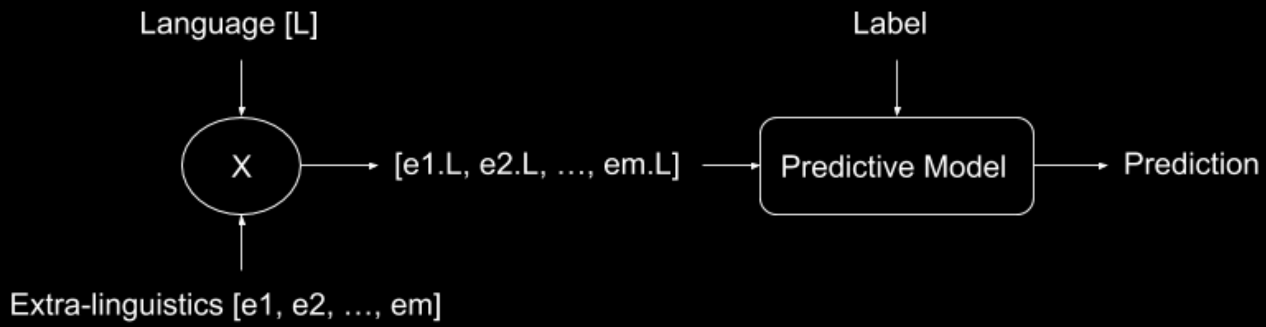
Additive (Residualized Control)

Residualize control (additive model):



(Zamani et al., EACL 2017)

Adaptive model:



Additive (Residualized Control)

Effectively use both low dimensional control features and high-dimensional, noisy language features:

1. **Train a control model** using the control values
2. **Calculate the residual** error and consider it as the new label
3. **Train a language model over the new labels**

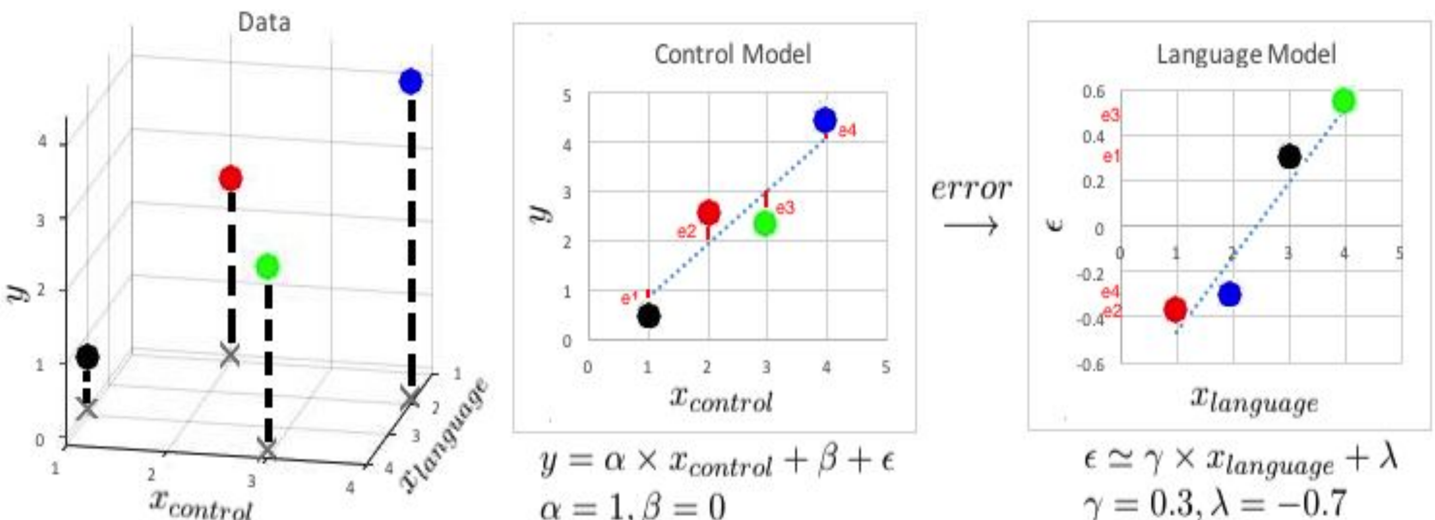
Model:

$$Y = \alpha\, x_1 + \beta x_2 + \gamma$$

Both learn the same linear model above, but

- Different learning algorithms per variable type.
- Different penalization methods

Residualized Control Model



$$y = \alpha \times x_{control} + \beta + \epsilon$$
$$\alpha = 1, \beta = 0$$

$$\epsilon \simeq \gamma \times x_{language} + \lambda$$
$$\gamma = 0.3, \lambda = -0.7$$

$$\Rightarrow y \simeq \alpha \times x_{control} + \gamma \times x_{language} + \beta + \lambda$$
$$\beta + \lambda = -0.7$$

Zamani M, Schwartz HA. Using Twitter Language to Predict the Real Estate Market. EACL 2017. 2017 Apr 3:28.

	Foreclosure	Increased-price
language	0.38	0.48
combined	0.40	0.49

	Foreclosure	Increased-price
language	0.38	0.48
combined	0.40	0.49
residualized control	**0.42**	**0.59**


post
san
texas
secret
improve
web
prices
international
super
starbucks
companies
california
create
downtown
company
access
tips
cheap
pro
credit
tbh
stoked
media
technology
internet
style
bomb
results
tour
source
experience
industry
na
guide
sales
price
cali
tax
content
per
hellamarketing
nn
followback
ou
law
blog
search
deal
a
a
a
correlation strength
relative frequency

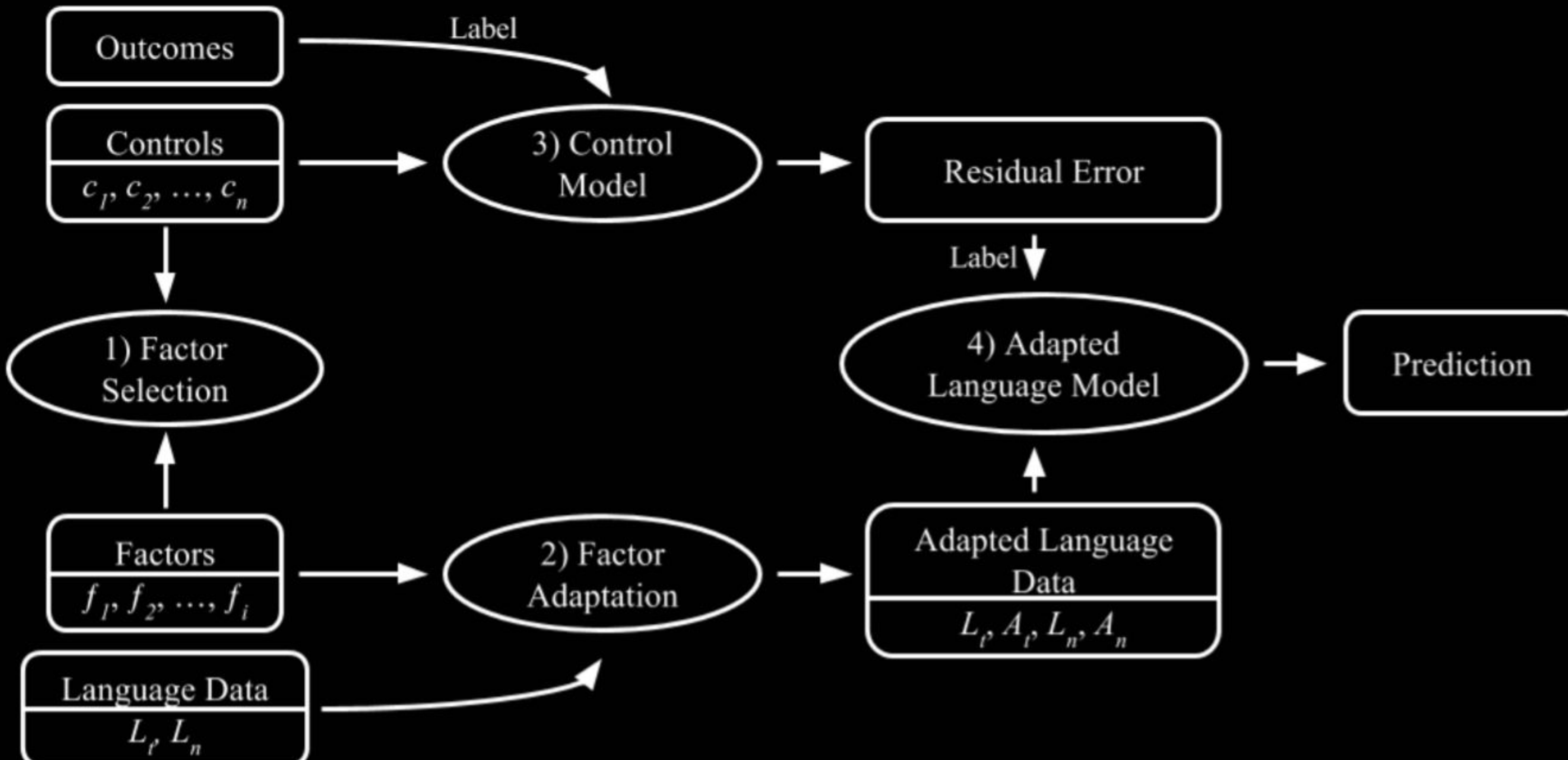
Combining Adaptive and Additive

Two Goals:

1. **Adaptive:** adapt to given human attributes
   (*user factor adaptation;*
   Lynn, Balasubramanian, Son, Kulkarni & Schwartz,
   EMNLP 2017)

2. **Additive:** predict beyond given attributes
   (*residualized control*; Zamani & Schwartz, *EACL* 2017)

Solution: Residualized Factor Adaptation



Results: County Health Predictions

	Lang.	Controls Only	Added-Controls
Heart Dis	0.585	0.514	0.608
Suicide	0.414	0.307	0.431
Poor Health	0.602	0.609	0.641
Life Satis.	0.209	0.329	0.335
Avg.	0.453	0.440	0.503

Results: County Health Predictions

	Lang.	All Factors		
		Controls Only	*Added-Controls*	*Res-Control*
Heart Dis	0.585	0.514	0.608	0.628
Suicide	0.414	0.307	0.431	0.460
Poor Health	0.602	0.609	0.641	0.661
Life Satis.	0.209	0.329	0.335	0.372
Avg.	0.453	0.440	0.503	0.530

Results: County Health Predictions

	Lang.	All Factors			
		Controls Only	*Added-Controls*	*Res-Control*	*FA*
Heart Dis	0.585	0.514	0.608	0.628	0.635
Suicide	0.414	0.307	0.431	0.460	0.494
Poor Health	0.602	0.609	0.641	0.661	0.674
Life Satis.	0.209	0.329	0.335	0.372	0.352
Avg.	0.453	0.440	0.503	0.530	0.539

Results: County Health Predictions

	Lang.	All Factors				
		Controls Only	*Added-Controls*	*Res-Control*	*FA*	*RFA*
Heart Dis	0.585	0.514	0.608	0.628	0.635	**0.655**
Suicide	0.414	0.307	0.431	0.460	0.494	**0.510**
Poor Health	0.602	0.609	0.641	0.661	0.674	0.682
Life Satis.	0.209	0.329	0.335	0.372	0.352	**0.396**
Avg.	0.453	0.440	0.503	0.530	0.539	0.560

Results: County Health Predictions

	Lang.	All Factors				
		Controls Only	*Added-Controls*	*Res-Control*	*FA*	*RFA*
Heart Dis	0.585	0.514	0.608	0.628	0.635	**0.655**
Suicide	0.414	0.307	0.431	0.460	0.494	**0.510**
Poor Health	0.602	0.609	0.641	0.661	0.674	0.682
Life Satis.	0.209	0.329	0.335	0.372	0.352	**0.396**
Avg.	0.453	0.440	0.503	0.530	0.539	0.560

variance explained (R^2)

Implications

a. Data is inherently multi-level: person-document

b. Often need control for “already-available” attributes

c. Linguistic features *interact* with human attributes

d. Language also has longitudinal context

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

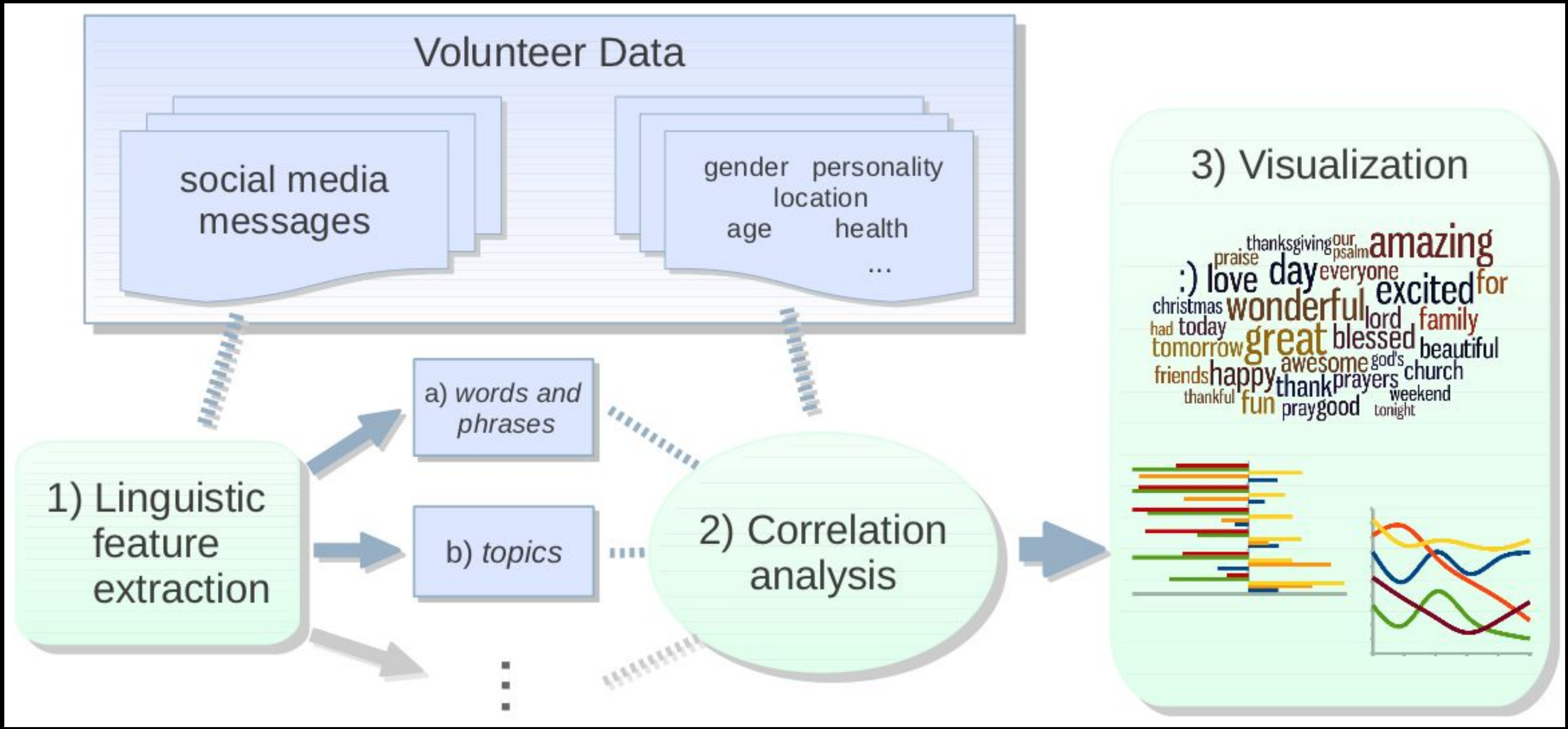
Human or community attribute

Output:

Features distinguishing attribute

Goal: Data-driven insights about an attribute

Differential Language Analysis



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

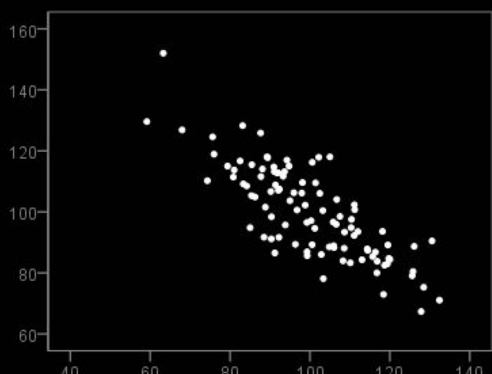
Differential Language Analysis

Methods of Correlation Analysis:

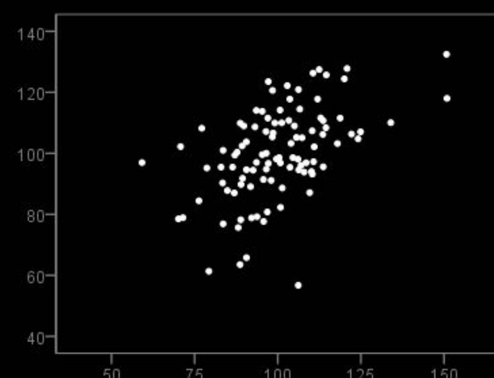
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson Product-Moment Correlation

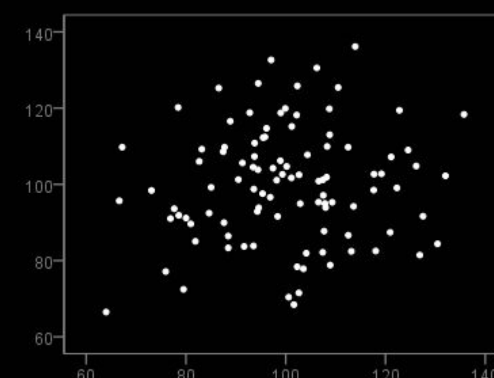
    Limitation: Doesn’t handle controls



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

- **<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":

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

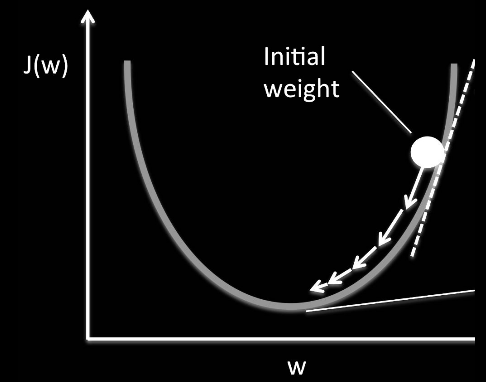
Differential Language Analysis

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



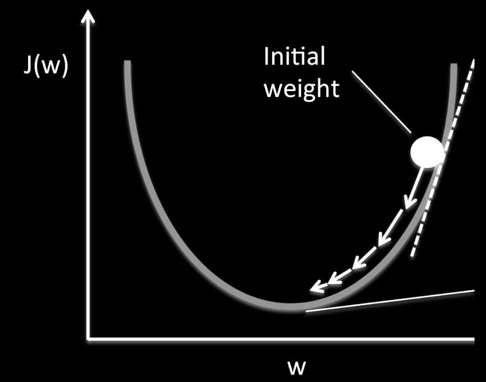
Differential Language Analysis

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: $$Y = X\beta + \epsilon$$



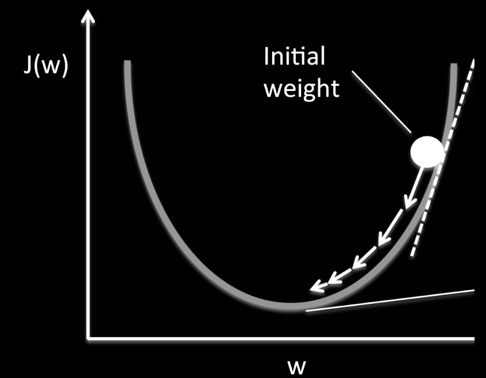
Differential Language Analysis

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

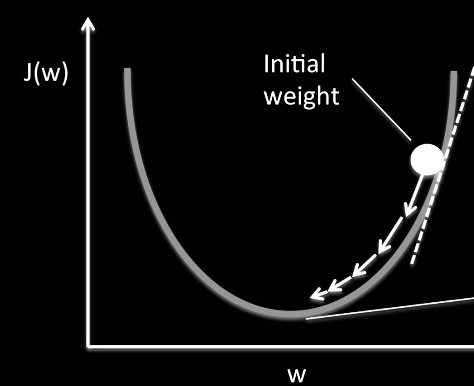
  Option 1: Gradient Descent:

  $J = \sum (y - \hat{y})^2$ -- "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

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

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

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables

- Odds Ratio

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

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

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

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using Informative Dirichlet Prior $log\left(\frac{countA(\text{"horrible"})}{NA-countA(\text{"horrible"})}\right) - log\left(\frac{countB(\text{"horrible"})}{NB-countB(\text{"horrible"})}\right)$

$$\hat{\delta}_w^{(i-j)} = log\left(\frac{y_w^i + \alpha_w}{n^i + \alpha_0 - (y_w^i + \alpha_w)}\right) - log\left(\frac{y_w^j + \alpha_w}{n^j + \alpha_0 - (y_w^j + \alpha_w)}\right)$$

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

Differential Language Analysis

Methods of "Correlation" Analysis for binary outcomes:

- Logistic Regression over Standardized variables

- Odds Ratio using Informative Dirichlet Prior

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

$$\hat{\delta}_w^{(i-j)} = log\left(\frac{y_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (y_w^i + \boxed{\alpha_w})}\right) - log\left(\frac{y_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (y_w^j + \boxed{\alpha_w})}\right)$$

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

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using Informative Dirichlet Prior

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

$$\hat{\delta}_w^{(i-j)} = log\left(\frac{y_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (y_w^i + \boxed{\alpha_w})}\right) - log\left(\frac{y_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (y_w^j + \boxed{\alpha_w})}\right)$$

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

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

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables

- Odds Ratio using Informative Dirichlet Prior

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

$$\hat{\delta}_w^{(i-j)} = log\left(\frac{y_w^i + \boxed{\alpha_w}}{n^i + \boxed{\alpha_0} - (y_w^i + \boxed{\alpha_w})}\right) - log\left(\frac{y_w^j + \boxed{\alpha_w}}{n^j + \boxed{\alpha_0} - (y_w^j + \boxed{\alpha_w})}\right)$$

Bayesian term for “smoothing”: accounts for uncertainty as a function of less events (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.

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

Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using Informative Dirichlet Prior $log\left(\frac{countA(\text{"horrible"})}{NA-countA(\text{"horrible"})}\right) - log\left(\frac{countB(\text{"horrible"})}{NB-countB(\text{"horrible"})}\right)$

$$\hat{\delta}_w^{(i-j)} = log\left(\frac{y_w^i + \alpha_w}{n^i + \alpha_0 - (y_w^i + \alpha_w)}\right) - log\left(\frac{y_w^j + \alpha_w}{n^j + \alpha_0 - (y_w^j + \alpha_w)}\right)$$

(n^i is the size of corpus i, n^j is the size of corpus j, y_w^i is the count of word w in corpus i, y_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.)

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

- Final statistic for a word: z-score of its log-odds-ratio:

$$\frac{\hat{\delta}_w^{(i-j)}}{\sqrt{\sigma^2\left(\hat{\delta}_w^{(i-j)}\right)}}$$

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

Ethics in NLP

Types of bias in NLP tasks:

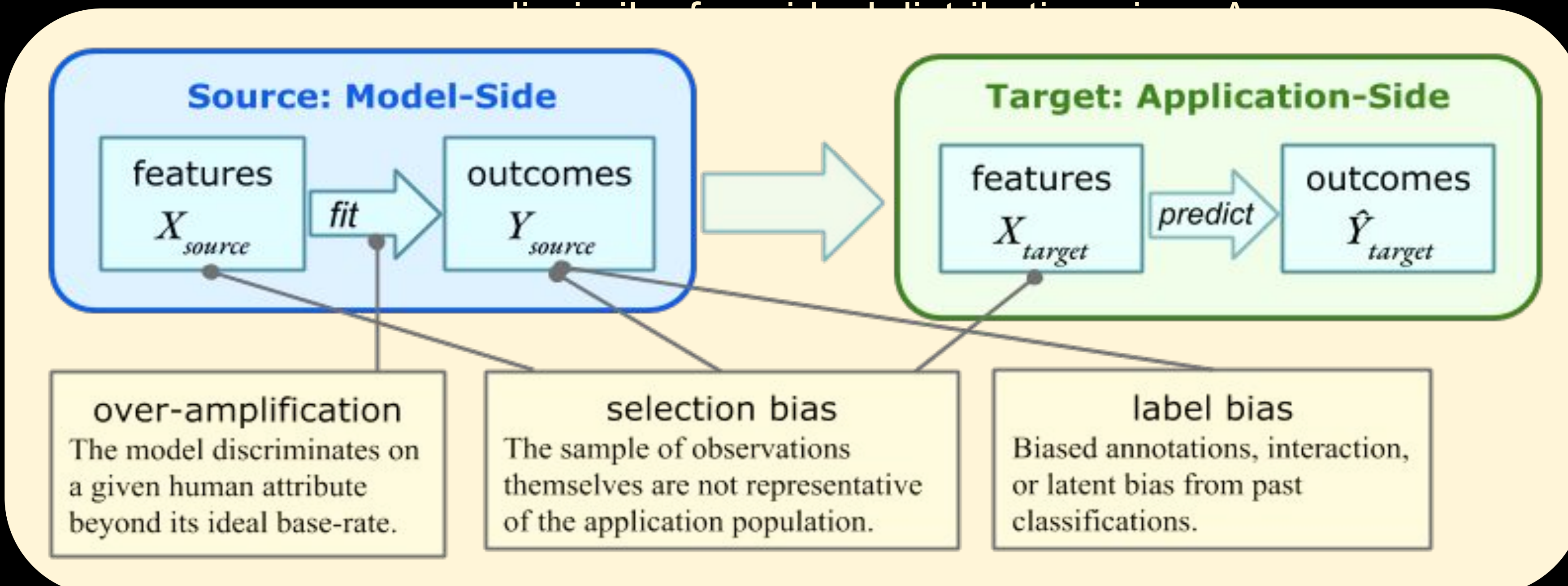
- Predictive Bias: Predicted distribution given A,
  are dissimilar from ideal distribution given A
  - Selection bias
  - Label bias
  - Over-amplification

Work in progres; Hovy et al., 2019

Ethics in NLP

Types of bias in NLP tasks:

- Predictive Bias: Predicted distribution given A,



Ethics in NLP

Types of bias in NLP tasks:

- Predictive Bias: Predicted distribution given A, are dissimilar from ideal distribution given A
    - Selection bias
    - Label bias
    - Over-amplification

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

Work in progres; Hovy et al., 2019

Ethics in NLP

Types of bias in NLP tasks:

- Predictive Bias: Predicted distribution given A,
  are dissimilar from ideal distribution given A
  - Selection bias
  - Label bias
  - Over-amplification

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

- Semantic Bias: Representations of meaning store demographic associations.

Work in progres; Hovy et al., 2019

Ethics in NLP

Types of bias in NLP tasks:

E.g. Coreference resolution: connecting entities to references (i.e. pronouns).

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

- Semantic Bias: Representations of meaning store demographic associations.

Work in progres; Hovy et al., 2019

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:
  - Informed consent -- let participants know
  - Do not share / secure storage
  - *Federated learning* -- separate and obfuscate to the point of preserving privacy
  - Transparency in information targeting
    "You are being shown this ad because ..."

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.