MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

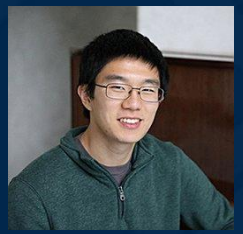
You Are What You Annotate

Towards Better Modeling Through Annotator Representations


Naihao Deng


Xinliang Zhang


Siyang Liu


Winston Wu


Lu Wang


Rada Mihalcea

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

- **Motivation**
- Factors
- Our Approach
- Dataset Selection
- Experiment Set-ups
- Results
- Analysis
- Related Works

The ubiquitous annotator disagreement

The ubiquitous annotator disagreement

Friends QIA

Question: Did Rachel tell you we hired a male nanny?

Answer: I think that's great!

ANN Answer (1), Not the Answer (2), Answer Subject to Some Conditions (3), Neither (4), Other (5): 1, 1, 4

Pejorative

Text: @WORSTRAPLYRICS Everything Jay-Z writes is trash.

ANN Pejorative (1) <–> Non-Pejorative (0): 1, 0, 0

HS-Brexit

Text: RT <user>: Islam has no place in Europe #Brexit.

ANN No Hate (1) <–> Hate (0): 1, 1, 1, 0, 0, 0

MultiDomain Agreement

Text: Please lost you yelling insanely at the sky on Nov 3 losers

ANN Offensive (1) <–> Not Offensive (0): 1, 1, 1, 0, 0

Go Emotions

Text: This is how I feel when I use a crosswalk on a busy street

ANN Positive (1), Neutral (0), Ambiguous (-1), Negative (-2): 1, 0

Humor

Text A: Being crushed by large objects can be very depressing.

Text B: As you make your bed, so you will sleep on it.

ANN which is funnier, X means a tie: A, A, B, X, X

CommitmentBank

Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.

Hypothesis: Carolyn had borrowed a book from Clare.

ANN Entail (3) <–>Contradict (-3): 3, 3, 3, 2, 0, -3, -3, -3

Sentiment Analysis

Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.

ANN Positive (2) <–>Negative (-2) : 2, 2, 0, -1

Is it acceptable to ignore such disagreement?

- Hate Speech Detection
  - Aggregating / omitting labels

HS-Brexit

Text: RT <user>: Islam has no place in Europe #Brexit.

ANN NO HATE (1) <–> HATE (0): 1, 1, 1, 0, 0, 0

Is it acceptable to ignore such disagreement?

- Hate Speech Detection
  - Aggregating / omitting labels → “standard” way of how people feel towards those texts → ignores the under-represented groups.
    Is this acceptable?

Is it acceptable to ignore such disagreement?

- Humor Detection / Sentiment Analysis
  - Difficult to reach a consensus on such subjective tasks.
  - **Is this acceptable?**

Sentiment Analysis

Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.

ANN POSITIVE (2) <–>NEGATIVE (-2) : 2, 2, 0, -1

Is it acceptable to ignore such disagreement?

- Natural Language Inference (NLI)
  - Inherent disagreements in people's judgments [1].

CommitmentBank

Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.

Hypothesis: Carolyn had borrowed a book from Clare.

ANN ENTAIL (3) <–>CONTRADICT (-3): 3, 3, 3, 2, 0, -3, -3, -3

[1] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. Transactions of the Association for Computational Linguistics, 7:677–694.

Is it acceptable to ignore such disagreement?

- Natural Language Inference (NLI)
  - Inherent disagreements in people's judgments [1].
  - Aggregating labels in NLI tasks
    Is this acceptable?

[1] Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. Transactions of the Association for Computational Linguistics, 7:677–694.

Is it acceptable to ignore such disagreement?

Convention: Aggregating / Omitting

Problematic

Is it acceptable to ignore such disagreement?

Convention: Aggregating / Omitting
Problematic

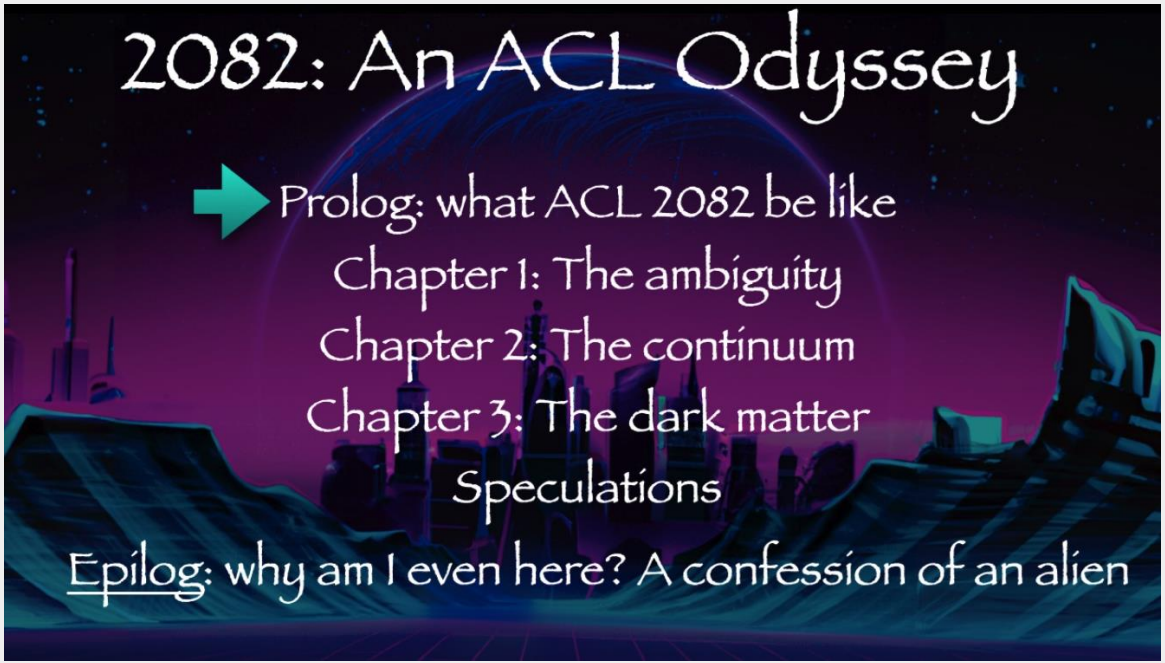
Let models learn from data that has inherent disagreement

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

- Motivation
- **Factors**
- Our Approach
- Dataset Selection
- Experiment Set-ups
- Results
- Analysis
- Related Works

The deep factors that cause annotator disagreement

- Ambiguity



The deep factor that causes annotator disagreement

- Differences in interpretation
- Certain preferences
- Difficult cases or multiple plausible answers

Barbara Plank. 2022. The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

The deep factor that causes annotator disagreement: Qualia

Internal experience

Sometimes we actually know that the same input is perceived differently by different people. (Dress on the right): some people see it as white and gold and other people see it as black and blue

From Jeff Hawkins "A thousand brains"

The deep factor that causes annotator disagreement: Qualia

I know what a pickle tastes like to me, but I can't know if a pickle tastes the same to you. Even if we use the **same words** to describe the taste of a pickle.

From Jeff Hawkins "A thousand brains"

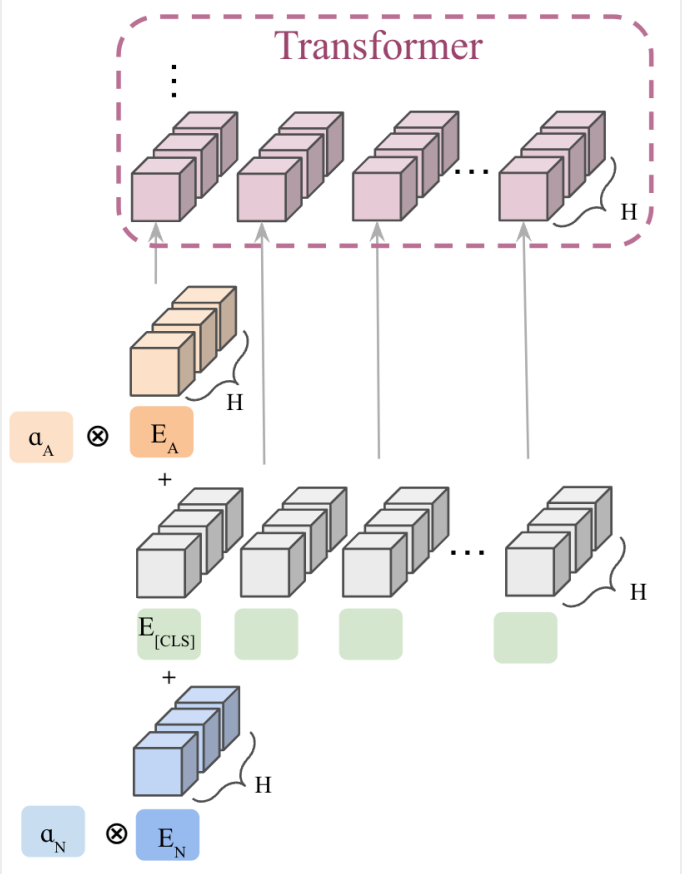
MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

- Motivation
- Factors
- **Our Approach**
- Dataset Selection
- Experiment Set-ups
- Results
- Analysis
- Related Works

Our Approach

- Annotator Embedding: represent each annotator
- Annotation Embedding: represent their annotations

Conceptual Diagram



MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

- Motivation
- Factors
- Our Approach
- **Dataset Selection**
- Experiment Set-ups
- Results
- Analysis
- Related Works

The ubiquitous annotator disagreement

Friends QIA
Question: Did Rachel tell you we hired a male nanny?
Answer: I think that's great!
ANN Answer (1), Not the Answer (2), Answer Subject to Some Conditions (3), Neither (4), Other (5): 1, 1, 4

Pejorative
Text: @WORSTRAPLYRICS Everything Jay-Z writes is trash.
ANN Pejorative (1) <–> Non-Pejorative (0): 1, 0, 0

HS-Brexit
Text: RT <user>: Islam has no place in Europe #Brexit.
ANN No Hate (1) <–> Hate (0): 1, 1, 1, 0, 0, 0

MultiDomain Agreement
Text: Please lost you yelling insanely at the sky on Nov 3 losers
ANN Offensive (1) <–> Not Offensive (0): 1, 1, 1, 0, 0

Go Emotions
Text: This is how I feel when I use a crosswalk on a busy street
ANN Positive (1), Neutral (0), Ambiguous (-1), Negative (-2): 1, 0

Humor
Text A: Being crushed by large objects can be very depressing.
Text B: As you make your bed, so you will sleep on it.
ANN which is funnier, X means a tie: A, A, B, X, X

CommitmentBank
Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.
Hypothesis: Carolyn had borrowed a book from Clare.
ANN Entail (3) <–>Contradict (-3): 3, 3, 3, 2, 0, -3, -3, -3

Sentiment Analysis
Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.
ANN Positive (2) <–>Negative (-2) : 2, 2, 0, -1

Dataset Categories

- NLI
  - FIA: Friends QIA
  - COM: CommitmentBank

The ubiquitous annotator disagreement

Friends QIA
Question: Did Rachel tell you we hired a male nanny?
Answer: I think that's great!
ANN Answer (1), Not the Answer (2), Answer Subject to Some Conditions (3), Neither (4), Other (5): 1, 1, 4

Pejorative
Text: @WORSTRAPLYRICS Everything Jay-Z writes is trash.
ANN Pejorative (1) <–> Non-Pejorative (0): 1, 0, 0

HS-Brexit
Text: RT <user>: Islam has no place in Europe #Brexit.
ANN No Hate (1) <–> Hate (0): 1, 1, 1, 0, 0, 0

MultiDomain Agreement
Text: Please lost you yelling insanely at the sky on Nov 3 losers
ANN Offensive (1) <–> Not Offensive (0): 1, 1, 1, 0, 0

Go Emotions
Text: This is how I feel when I use a crosswalk on a busy street
ANN Positive (1), Neutral (0), Ambiguous (-1), Negative (-2): 1, 0

Humor
Text A: Being crushed by large objects can be very depressing.
Text B: As you make your bed, so you will sleep on it.
ANN which is funnier, X means a tie: A, A, B, X, X

CommitmentBank
Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.
Hypothesis: Carolyn had borrowed a book from Clare.
ANN Entail (3) <–>Contradict (-3): 3, 3, 3, 2, 0, -3, -3, -3

Sentiment Analysis
Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.
ANN Positive (2) <–>Negative (-2) : 2, 2, 0, -1

Dataset Categories

- NLI
- Sentiment Analysis
  - GOE: Go Emotions
  - SNT: Sentiment Analysis

The ubiquitous annotator disagreement

Friends QIA
Question: Did Rachel tell you we hired a male nanny?
Answer: I think that's great!
ANN Answer (1), Not the Answer (2), Answer Subject to Some Conditions (3), Neither (4), Other (5): 1, 1, 4

Pejorative
Text: @WORSTRAPLYRICS Everything Jay-Z writes is trash.
ANN Pejorative (1) <–> Non-Pejorative (0): 1, 0, 0

HS-Brexit
Text: RT <user>: Islam has no place in Europe #Brexit.
ANN No Hate (1) <–> Hate (0): 1, 1, 1, 0, 0, 0

MultiDomain Agreement
Text: Please lost you yelling insanely at the sky on Nov 3 losers
ANN Offensive (1) <–> Not Offensive (0): 1, 1, 1, 0, 0

Go Emotions
Text: This is how I feel when I use a crosswalk on a busy street
ANN Positive (1), Neutral (0), Ambiguous (-1), Negative (-2): 1, 0

Humor
Text A: Being crushed by large objects can be very depressing.
Text B: As you make your bed, so you will sleep on it.
ANN which is funnier, X means a tie: A, A, B, X, X

CommitmentBank
Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.
Hypothesis: Carolyn had borrowed a book from Clare.
ANN Entail (3) <–>Contradict (-3): 3, 3, 3, 2, 0, -3, -3, -3

Sentiment Analysis
Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.
ANN Positive (2) <–>Negative (-2) : 2, 2, 0, -1

Dataset Categories

- NLI
- Sentiment Analysis
- Hate Speech Detection
  - PEJ: Pejorative
  - MDA: Multidomain Agreement
  - HSB: HS Brexit

The ubiquitous annotator disagreement

Friends QIA
Question: Did Rachel tell you we hired a male nanny?
Answer: I think that's great!
ANN Answer (1), Not the Answer (2), Answer Subject to Some Conditions (3), Neither (4), Other (5): 1, 1, 4

Pejorative
Text: @WORSTRAPLYRICS Everything Jay-Z writes is trash.
ANN Pejorative (1) <–> Non-Pejorative (0): 1, 0, 0

HS-Brexit
Text: RT <user>: Islam has no place in Europe #Brexit.
ANN No Hate (1) <–> Hate (0): 1, 1, 1, 0, 0, 0

MultiDomain Agreement
Text: Please lost you yelling insanely at the sky on Nov 3 losers
ANN Offensive (1) <–> Not Offensive (0): 1, 1, 1, 0, 0

Go Emotions
Text: This is how I feel when I use a crosswalk on a busy street
ANN Positive (1), Neutral (0), Ambiguous (-1), Negative (-2): 1, 0

Humor
Text A: Being crushed by large objects can be very depressing.
Text B: As you make your bed, so you will sleep on it.
ANN which is funnier, X means a tie: A, A, B, X, X

CommitmentBank
Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.
Hypothesis: Carolyn had borrowed a book from Clare.
ANN Entail (3) <–>Contradict (-3): 3, 3, 3, 2, 0, -3, -3, -3

Sentiment Analysis
Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.
ANN Positive (2) <–>Negative (-2) : 2, 2, 0, -1

Dataset Categories

- NLI
- Sentiment Analysis
- Hate Speech Detection
- Humor:
  - HUM: Humor

Select Dataset with High Quality: Get Rid of Noise

E.g. CommitmentBank dataset

Control examples to assess annotators' attention

Filter data from annotators giving other responses.

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

- Motivation
- Factors
- Our Approach
- Dataset Selection
- **Experiment Setups**
- Results
- Analysis
- Related Works

Experiment Setups

- Model types: BERT [1], RoBERTa [2], DeBERTa V3 [3]
- Model sizes: base; large
- We tuned each model for 3 epochs with lr of 1e-5.

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
[2] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
[3] He, Pengcheng, Jianfeng Gao, and Weizhu Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing." arXiv preprint arXiv:2111.09543 (2021).

Experiment Setups

Other baselines:

- R: random
- MV_{ind}: majority vote for each individual
- MV_{macro}: majority vote for all the training labels
- NC: naively concatenate text with annotator ID

Experiment Setups

- Evaluation metrics:
  - EM accuracy
  - Macro F1 scores
- Annotation split:
  - Same set of annotators in train and test

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

- Motivation
- Factors
- Our Approach
- Dataset Selection
- Experiment Set-ups
- Results
- **Analysis**
- Related Works

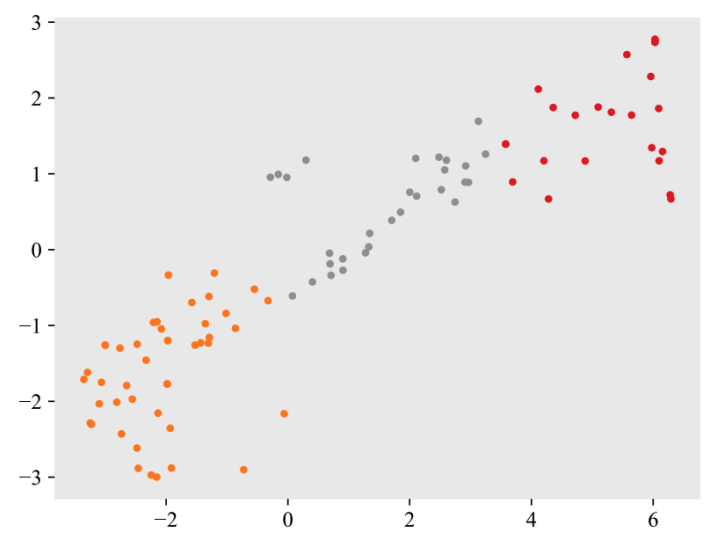
Further Analysis: Case 1: Annotation Embed the Best

	R	MV_{ind}	MV_{macro}	T	NC	E_n	E_a	$E_n + E_a$
MDA	50.03	61.71	63.58	75.06	64.81	**76.70**	75.72	75.76
				75.67	65.88	76.13	74.67	74.97
				75.65	68.36	76.02	75.14	75.28
				76.24	69.73	**77.18**	75.99	75.68
				76.45	70.43	**77.78**	76.93	77.26
				76.38	73.02	77.22	74.75	77.19

Domains of MDA: English tweets about Black Lives Matter, Elections and Covid-19

Uniform agreement within groups that have same ideology

Further Analysis: Case 1: Annotation Embed the Best



TSNE plots for MDA dataset. Annotation Embed seems to align with the “group tendencies”.

Further Analysis: Case 2: Annotator Embed the Best

	R	MV_{ind}	MV_{macro}	T	NC	E_n	E_a	$E_n + E_a$
GOE	25.05	41.27	36.71	63.04	60.88	68.49	**69.98**	69.90
				62.90	62.12	68.39	**69.92**	66.32
				63.22	60.49	67.42	**69.22**	68.54
				63.19	58.86	64.41	65.71	**68.46**
				63.59	62.28	68.58	**69.70**	69.60
				62.94	58.60	65.18	66.52	**69.74**
HUM	33.30	45.65	41.55	54.26	52.05	56.72	**58.15**	53.89
				54.11	51.07	56.67	**58.19**	54.35
				54.43	47.16	55.07	**56.31**	53.31
				54.40	52.55	54.26	51.97	50.02
				54.71	53.63	56.33	**57.70**	53.31
				54.67	54.81	57.18	**58.76**	51.86

GOE and HUM: emotion and humor are highly personalized feelings

Annotator Embeddings seem to better capture these individual differences.

Further Analysis: Case 3: Synergy of the Two Embeds

	R	MV_{ind}	MV_{macro}	T	NC	E_n	E_a	$E_n + E_a$
SNT	20.04	49.47	37.49	47.09	39.20	62.88	60.23	**64.61**
				47.32	36.91	61.88	56.20	**63.65**
				46.40	43.32	**60.30**	45.57	59.65
				47.88	43.82	**58.19**	46.50	55.16
				45.75	43.62	**61.21**	52.57	60.83
				48.76	43.78	67.37	68.39	**69.77**

SNT: designed specifically for age-related bias.

Annotator-based predictions

Text-only predicts only a single label for different annotators

Embeddings-based methods can accommodate annotator differences.

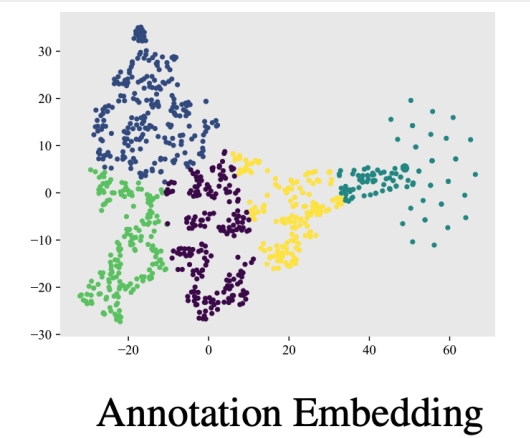
Text: We know it anecdotally from readers we've heard from who've been blatantly discriminated against because they're older.

POSITIVE (2) <–> NEGATIVE (-2)

Annotator ID	1	2	3	4
Gold	-1	0	-2	-2
T	-1	-1	-1	-1
$\mathbf{E}_n + \mathbf{E}_a$	-1	0	-1	-2

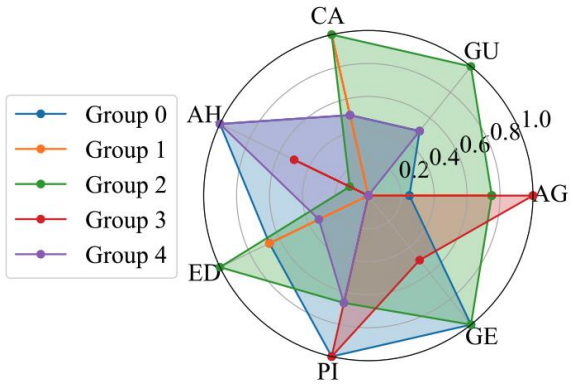
Annotation Embeddings can be Grounded to Real-World Demographic Features

Naturally emerged clusters for annotation embed (K-means)



Annotation Embeddings can be Grounded to Real-World Demographic Features

We can map these clusters back to demographic features



CA: current living area
GU: grew up area
AG: age
GE: gender
PI: political identification
ED: education
AH: annual household income

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Thanks!