

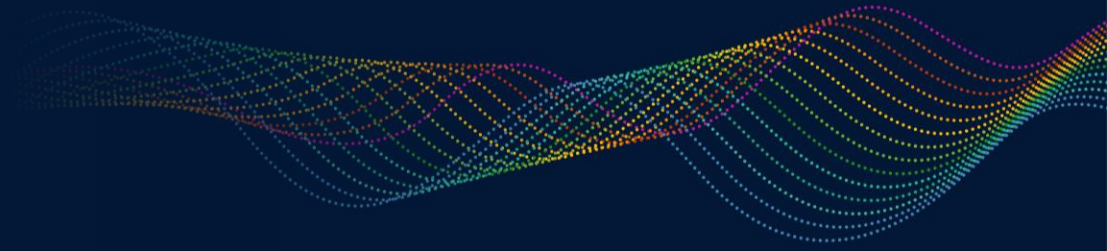
High Quality Training Data for AI Models: Lessons from 20 years in Surveys

Stephanie Eckman

GOR Plenary

April 1, 2025

Research Group



Rob Chew,
RTI International



Bolei Ma,
LMU



Frauke Kreuter,
LMU, UMD

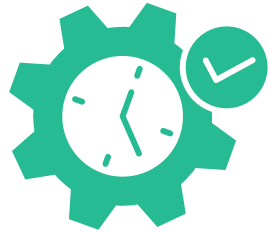


Christoph Kern,
LMU



Jacob Beck,
LMU

AI Models: Benefit or Harm



Increased
Efficiency



Protein
Folding



Discrimination
due to AI



Telling people
to eat glue

Training Data Collection

- Crowdworkers
 - Mechanical Turk
 - Appen
 - Scale AI
- Tend to be objective tasks



What kind of animal?

- Dog
- Cat
- Other
- No animal

Training Data Collection



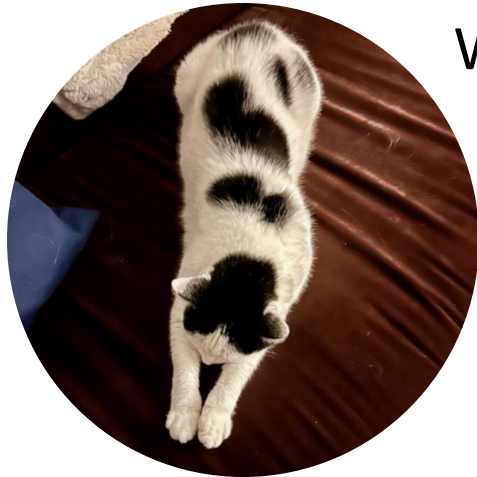
- Trained on everything
- Rewarded based on feedback data
 - More subjective, opinion tasks
 - Collected from crowdworkers

What tools can I use to break into a house?

Select the better answer:

- How about a screwdriver?
- I cannot help you commit a crime

Core Insight



What kind of animal?

- Dog
- Cat
- Other
- No animal

A lot like surveys!

What tools do I need to break into a house?

Select the better answer:

- How about a screwdriver?
- I cannot help you commit a crime

Submit

Skip

«

Page 3 / 11

»

Instruction

Summarize the following news article:

Article text here

Include output

Output A

Article summary

Rating (1 = worst, 7 = best)

1

2

3

4

5

6

7

Fails to follow the correct instruction / task ? Yes No

Inappropriate for customer assistant ? Yes No

Contains sexual content Yes No

Contains violent content Yes No

Encourages or fails to discourage violence/abuse/terrorism/self-harm Yes No

Denigrates a protected class Yes No

Gives harmful advice ? Yes No

Expresses moral judgment Yes No

Training

Survey Data Concerns



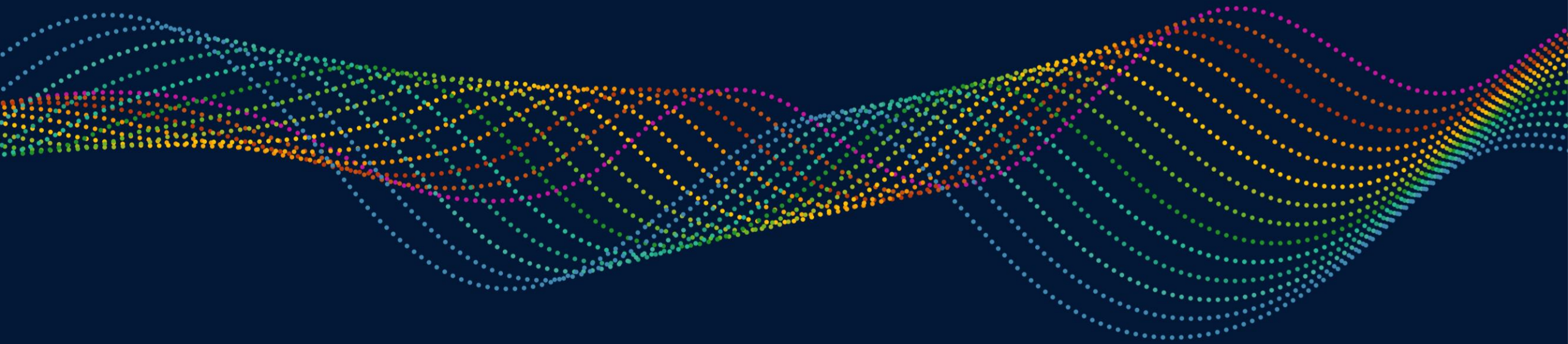
Measurement

Are the **labels** correct?



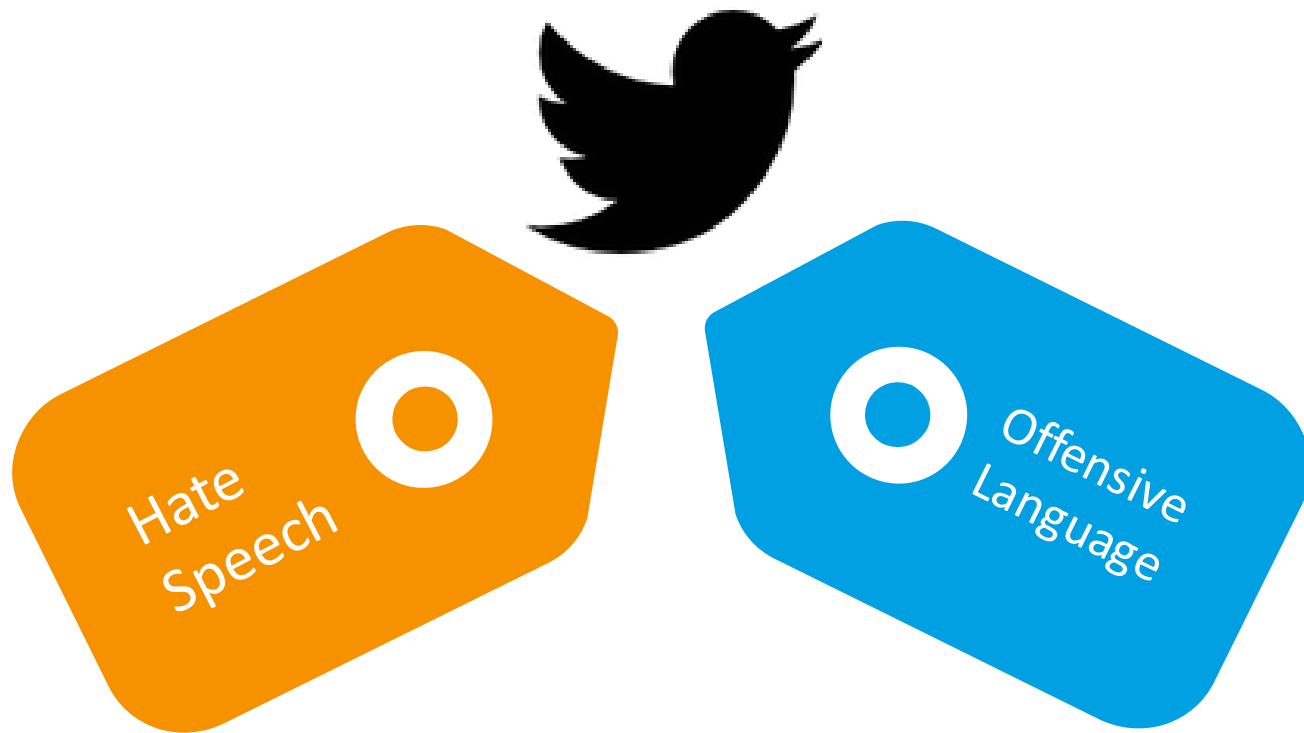
Representation

Who **labels**?













































Measurement

Study: Instrument Effects



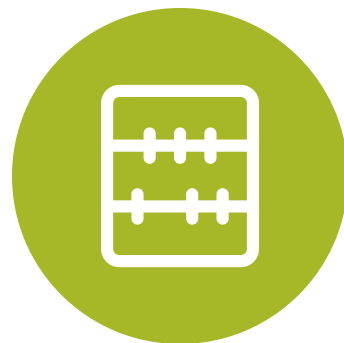
5 Labelling Conditions

A				
				
				
B				
				
C				
				
D			...	
				
				
E			...	
				
				

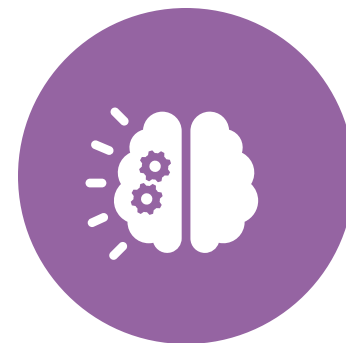
Three Sets of Results



LABELS



MODELS

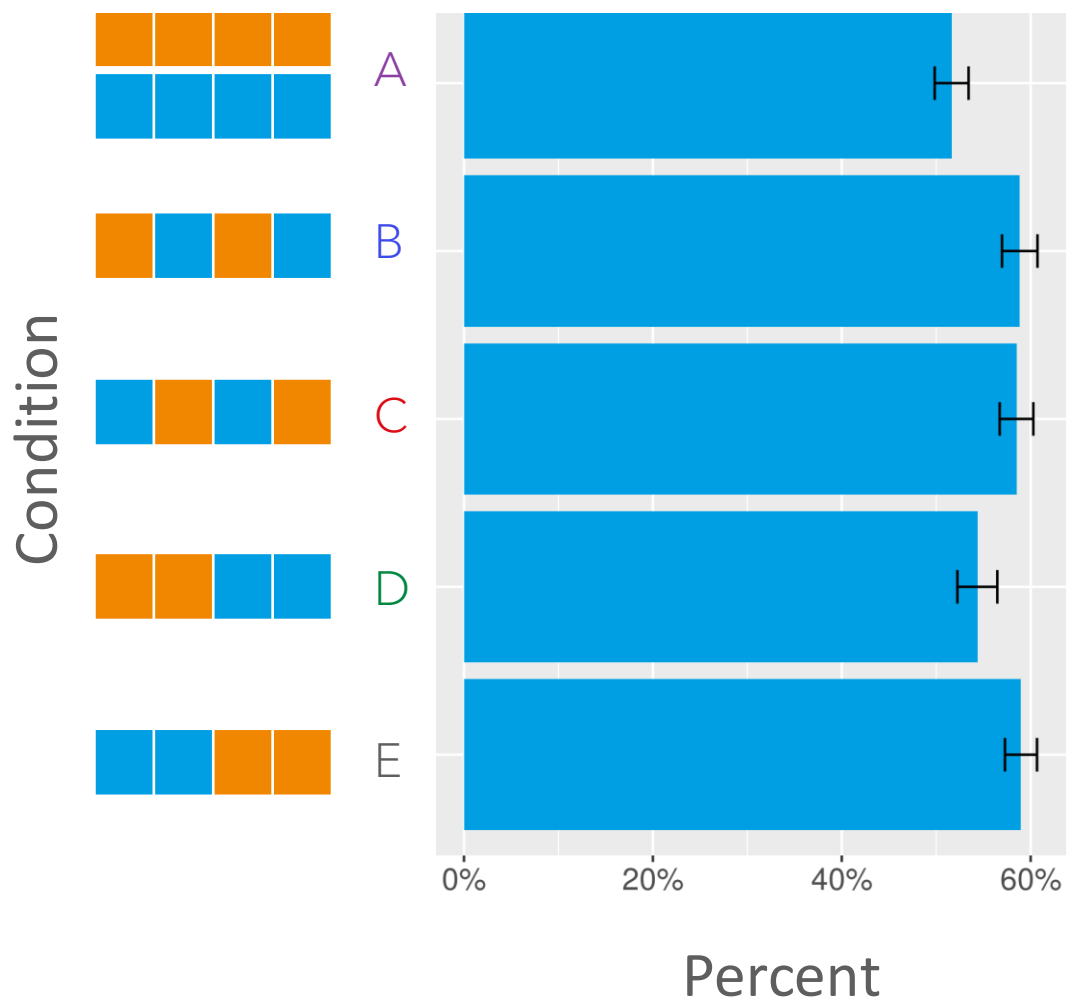


PREDICTIONS



Labels

% Tweets Labeled as Offensive Language

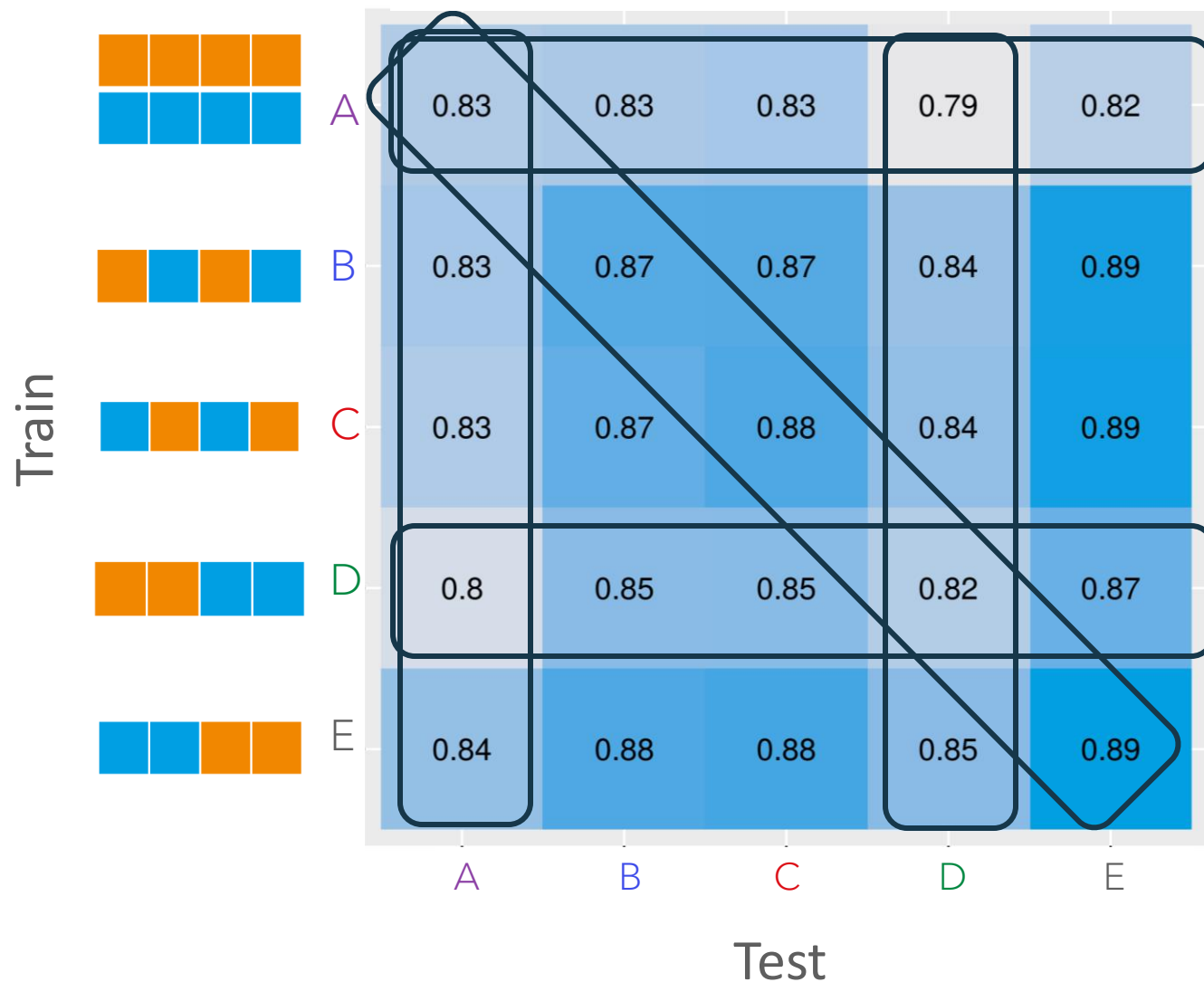




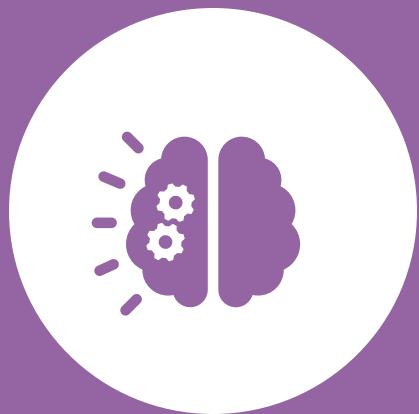
Models

Offensive
Language

Balanced Accuracy



Kern et al. 2023. "Annotation Sensitivity: Training Data Collection Methods Affect Model Performance"



Predictions

Correlation of Model Predictions

	A	B	C	D
B	0.68			
C	0.75	0.87		
D	0.73	0.87	0.90	
E	0.68	0.88	0.86	0.87

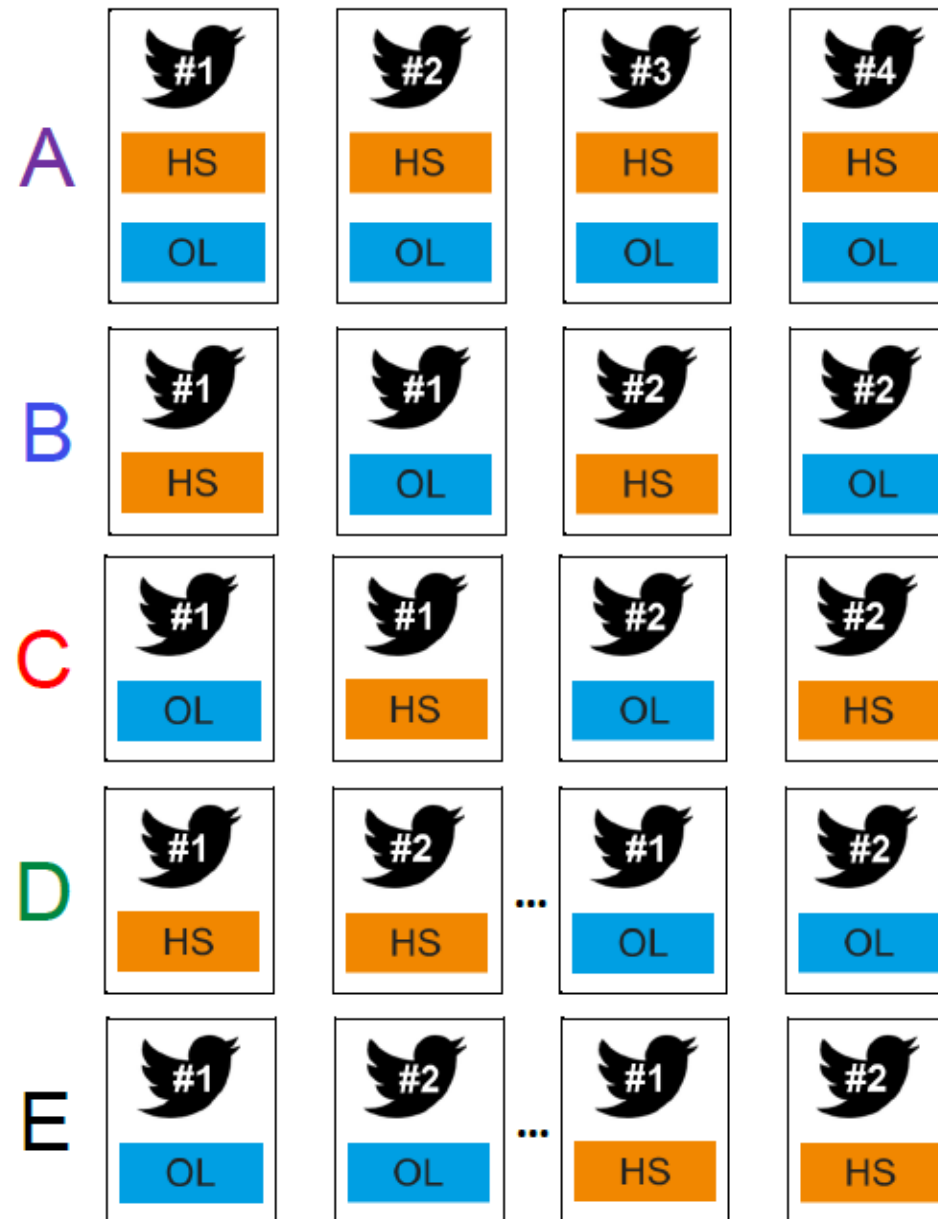
Takeaways

Instrument impacts labels

- Fatigue effect
- Order effect

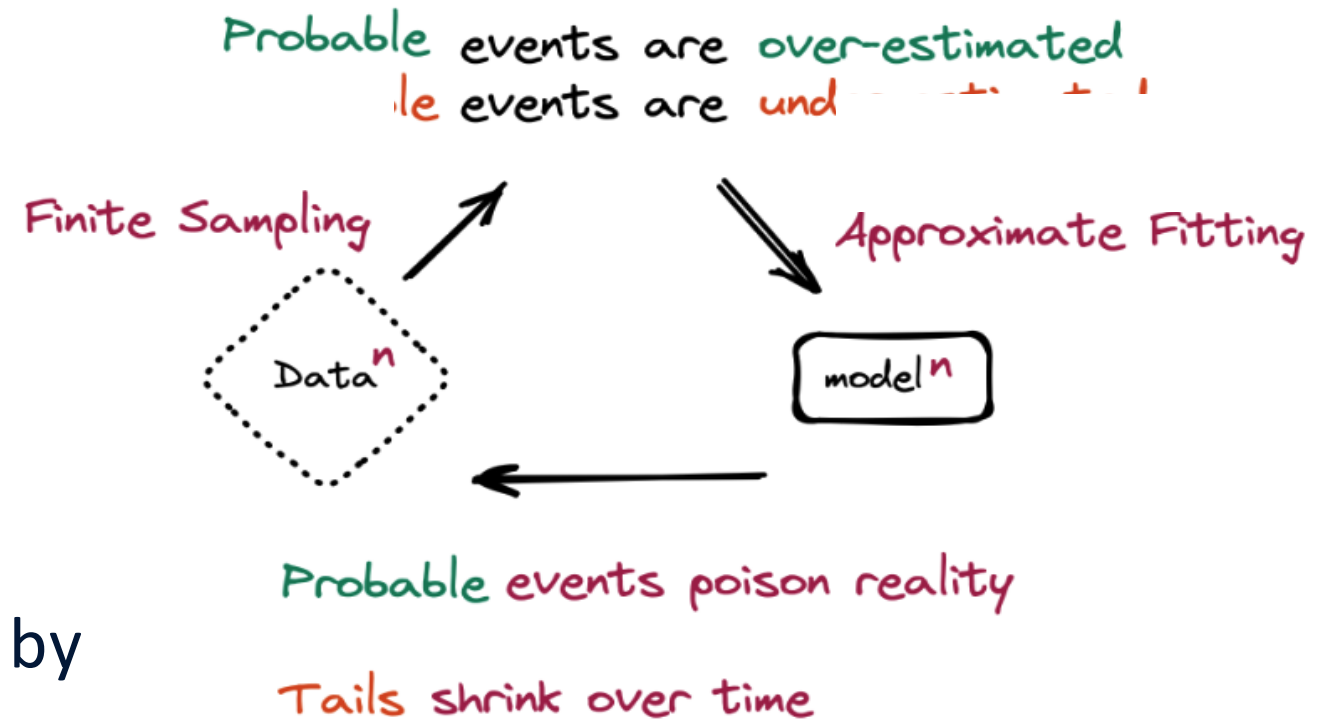
Instrument impacts models

Findings from surveys apply to training data collection

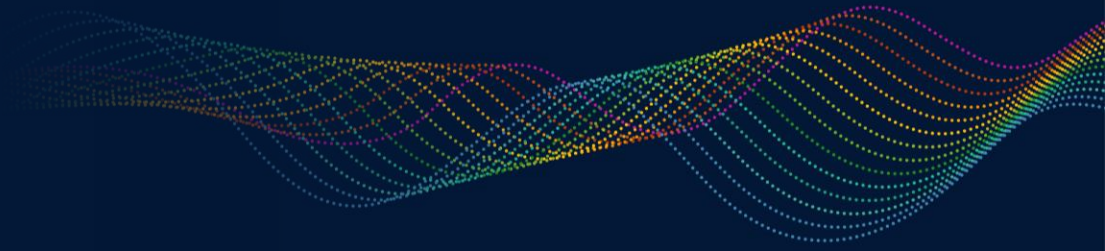


Can't a Model Label my Data?

- Models show same biases as humans
- Model autophagy / collapse
- Feedback data: most important, difficult labeling should be done by humans



Pre-Labeling



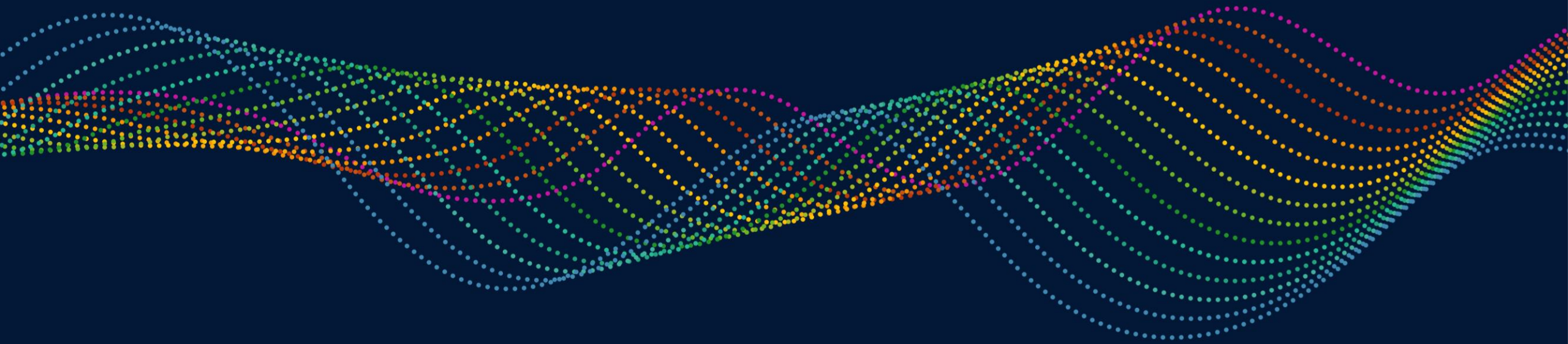
	2019	2020
CO ₂ Emissions	869	533
Scope 1	0	0
Scope 2	0	0
Scope 3	860	533

The 2020 Scope 3 emissions are 523, according to the AI.

Is this correct?

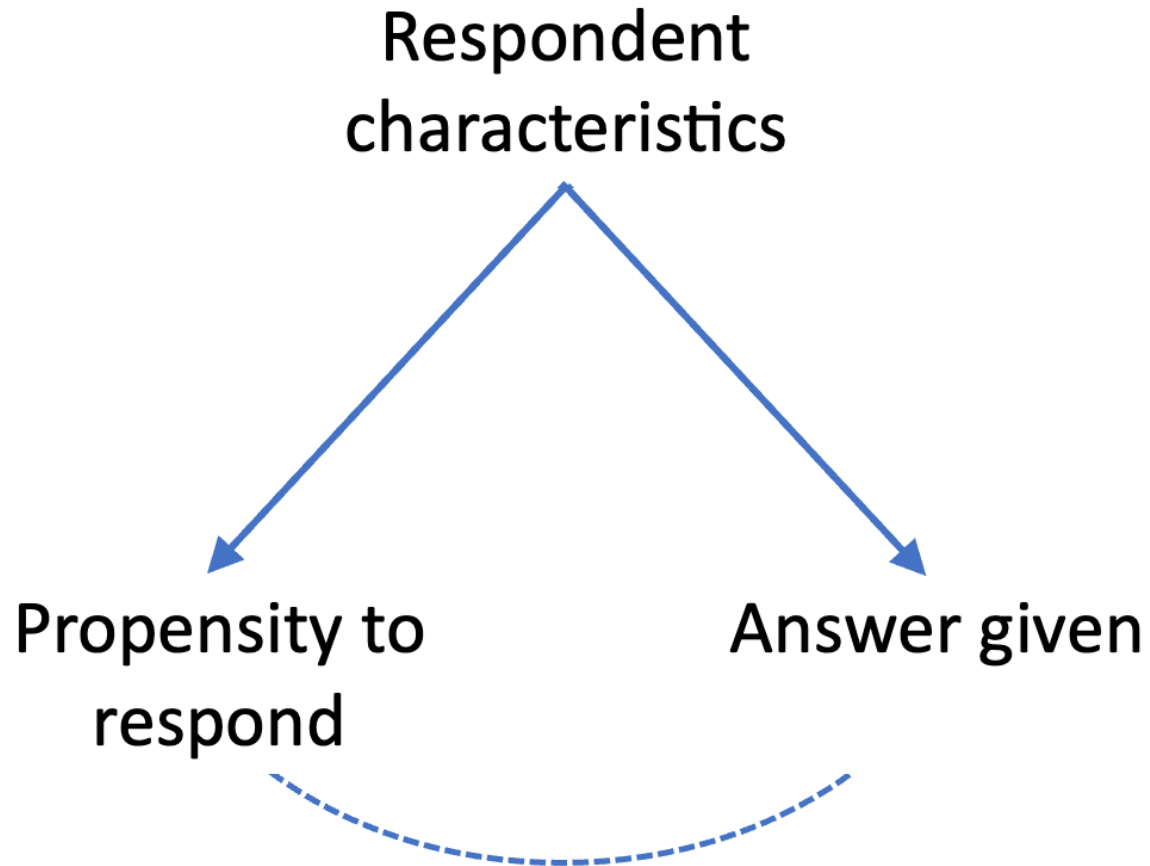
Errors underreported when labelers are:

- More trusting of AI
- Required to give correct label



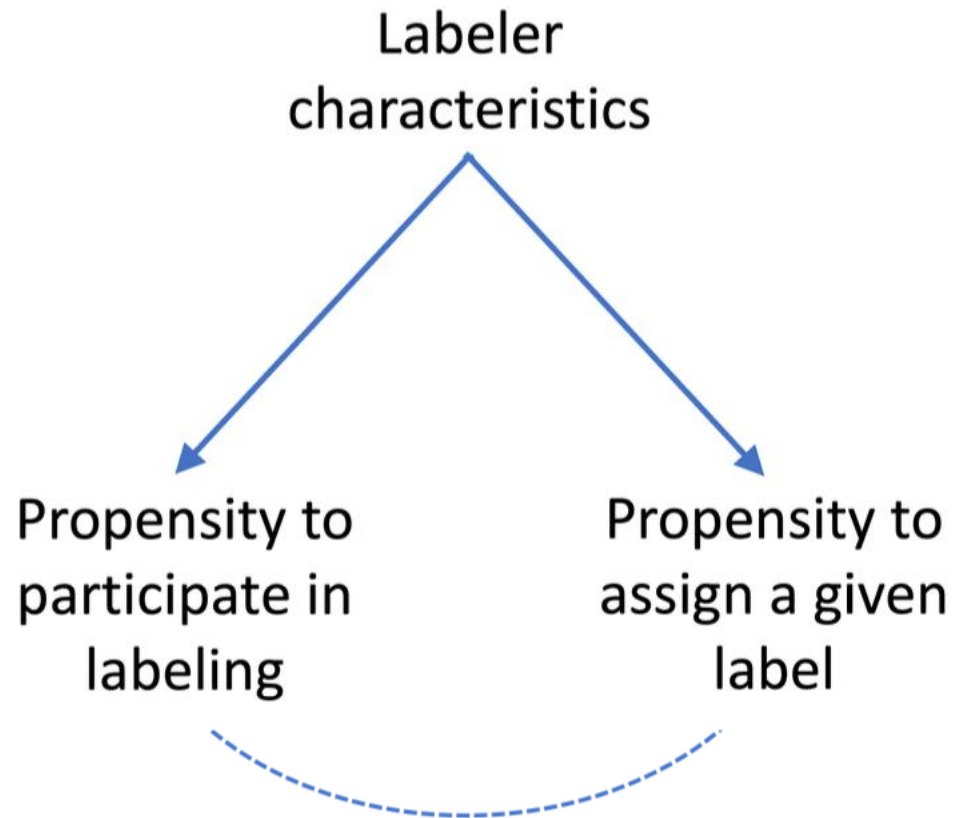
Representation

Nonresponse Bias in Surveys



- Nonresponse Bias occurs when **respondent characteristics** impact **propensity to respond** and **answers given**

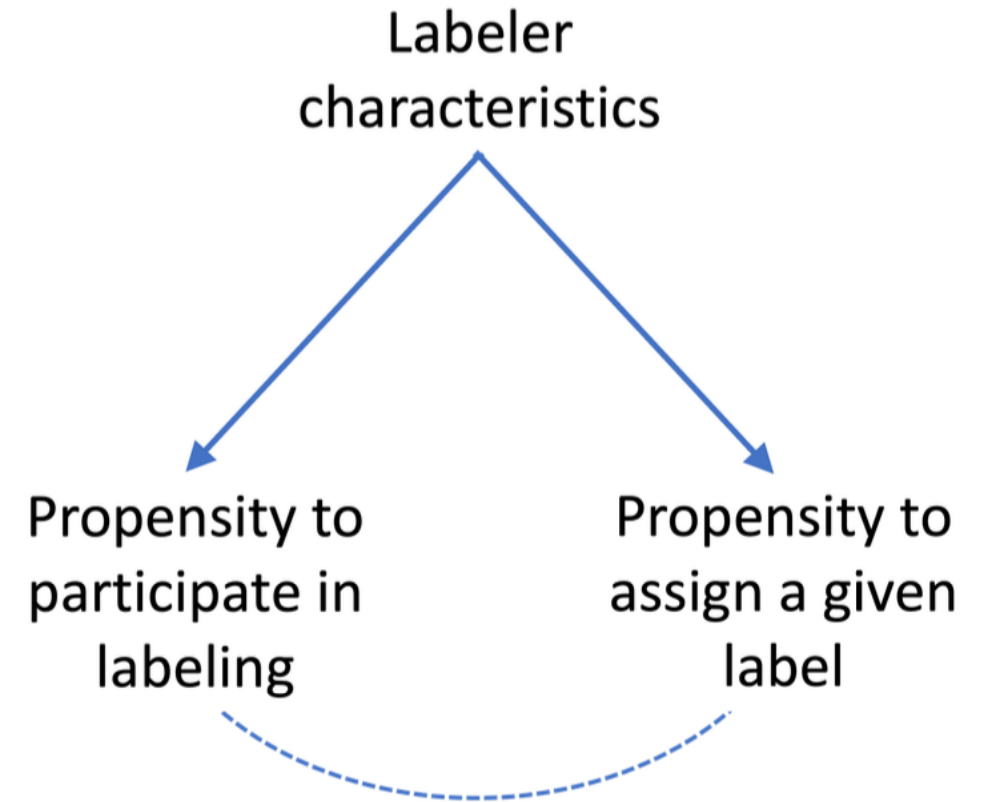
Selection Bias in Labels and Models



- Labeler characteristics influence **labels**
- As well as **models**

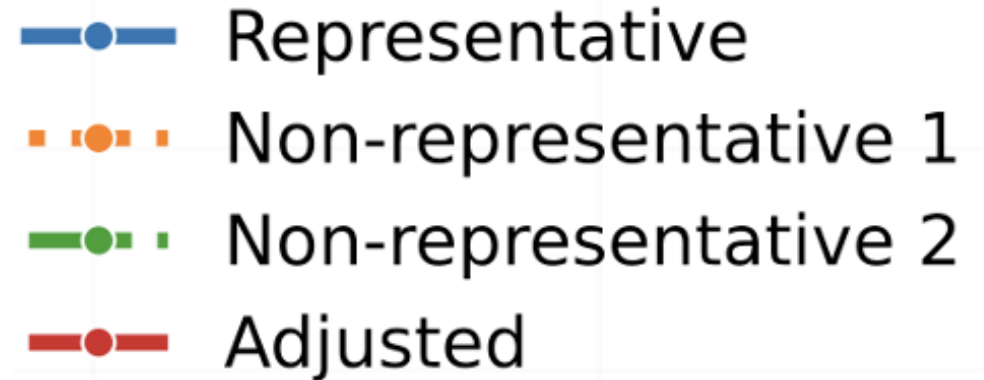
Solutions to Selection Bias

- **Left:** Diversify labeler pool
- **Right:** Train to label uniformly
- Adjust labels to match population



Recent Paper

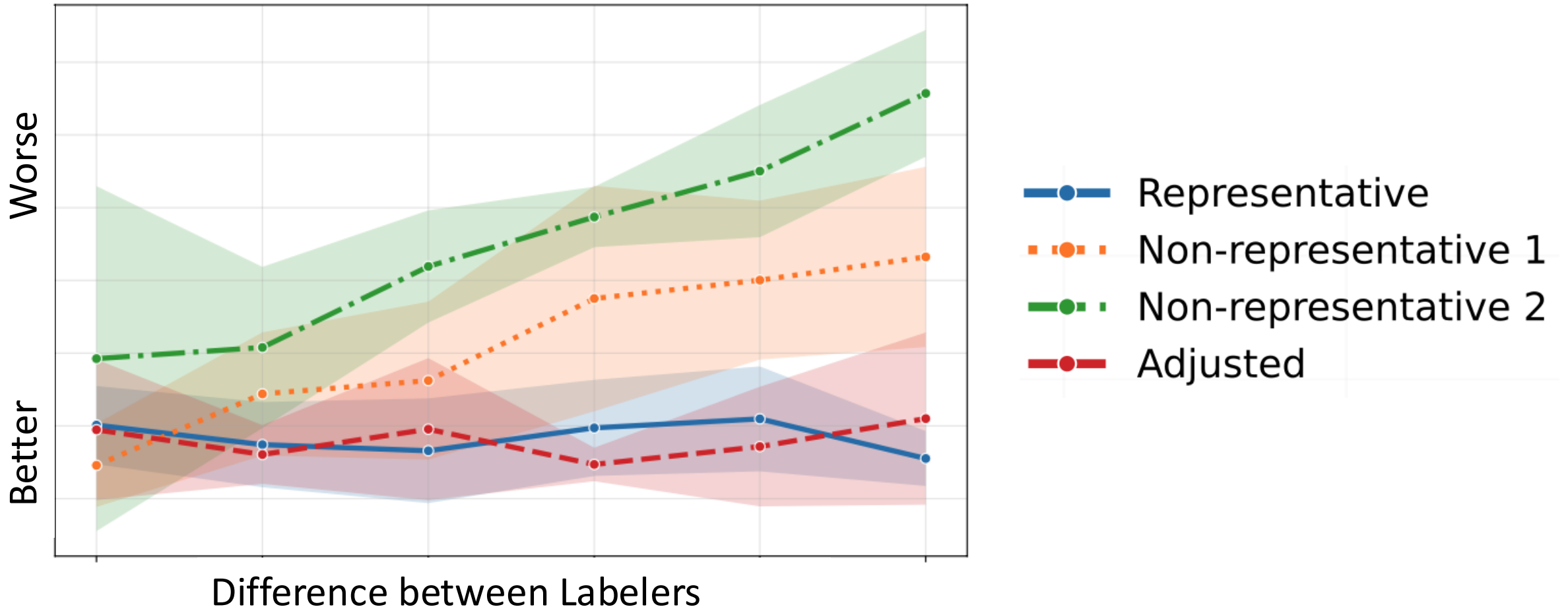
- Labels of 3,000 tweets
- Simulate labels from 2 types:
 - **More likely** to see offensive language
 - **Less likely**
- Vary mix of types
 - Reweight to reflect population



Results

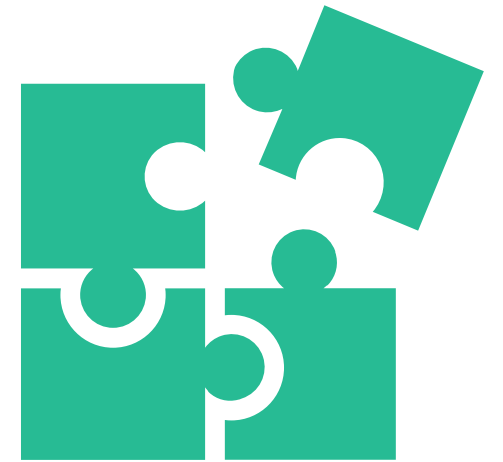


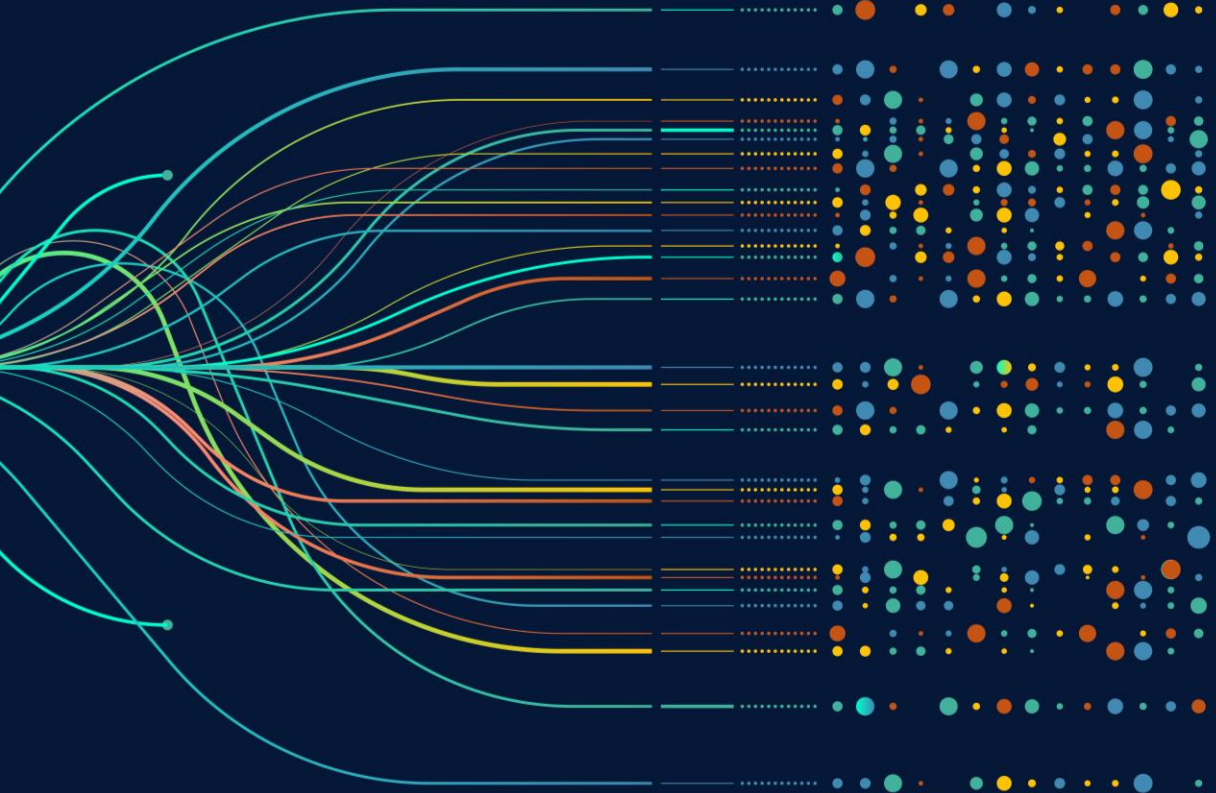
Absolute Calibration Bias



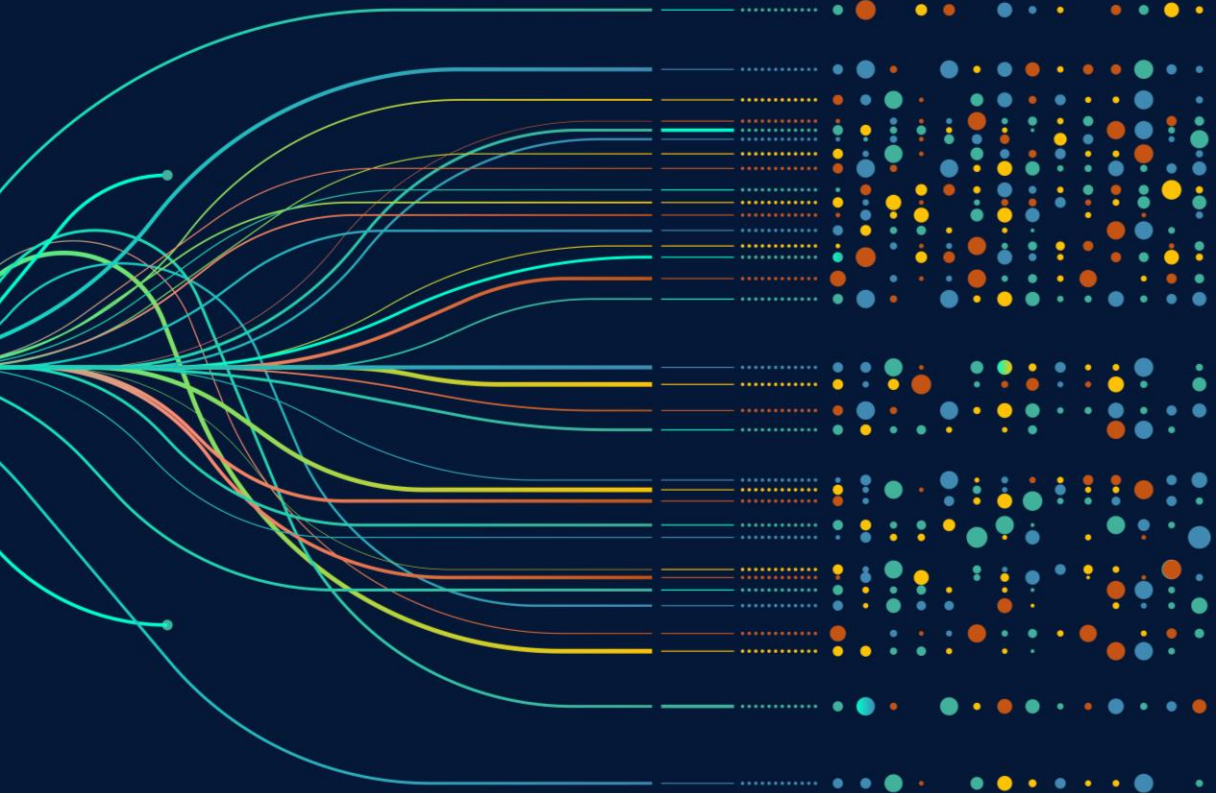
More Research Needed

- What population do we weight to?
- How to use weights in model training?





Opportunities for Survey Researchers



Thank you

Stephanie Eckman

www.stepheckman.com

Recent Papers



Eckman et al. 2024. Position: Insights from Survey Methodology can Improve Training Data for Machine Learning Models. ICML
<https://proceedings.mlr.press/v235/eckman24a.html>

Kern et al. 2023. Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. EMNLP
<https://aclanthology.org/2023.findings-emnlp.992/>

Beck et al. 2024. Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity. UncertainNLP <https://aclanthology.org/2024.uncertainlp-1.8/>

Eckman et al. 2025. Correcting Annotator Bias in Training Data: Population-Aligned Instance Replication (PAIR). Working Paper
<https://arxiv.org/abs/2501.06826>