Collecting High Quality Training Data:

Lessons from 20 years in Surveys

Stephanie Eckman

Researcher, University of Maryland

Introductions

- Name, where you work or study
- What brought you to the workshop or class
- Any comments or questions from read-ahead

Who I am

- 20+ years of experience collecting survey data
- Respondent incentives & data quality
- Combining survey and non-survey data

- www.stepheckman.com

Introductions

- Name, where you work or study
- What brought you to the workshop or class
- Any comments or questions from read-ahead

Motivation for Course

Motivation / Overview

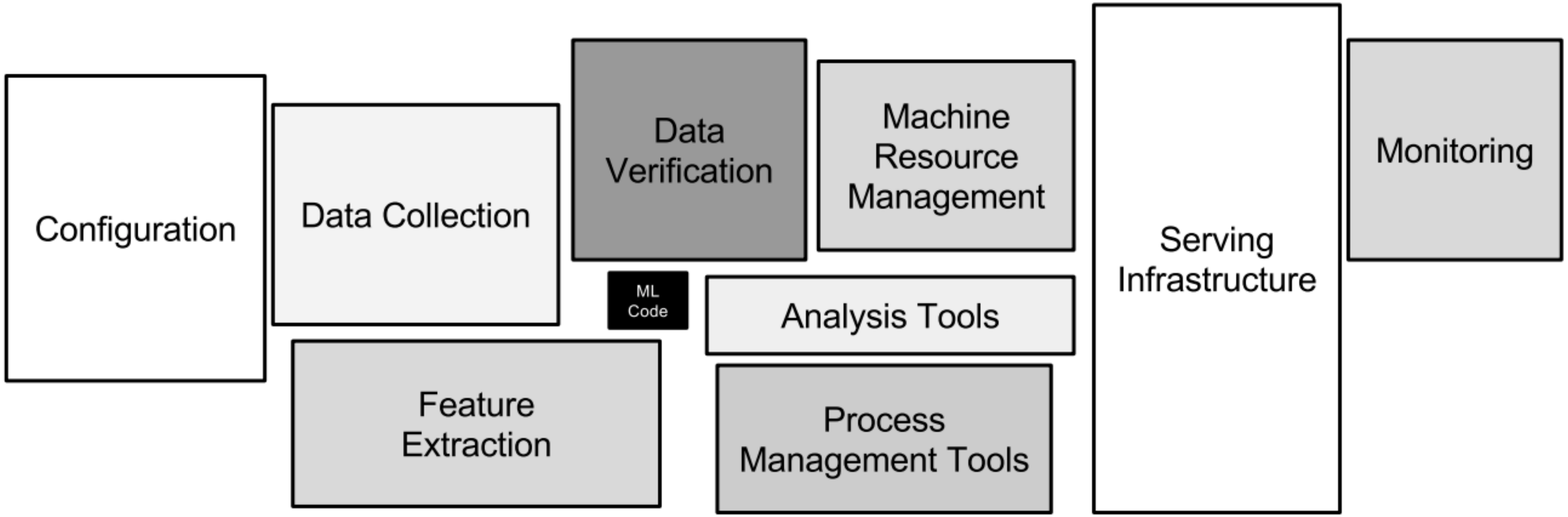
- Better data quality can improve model performance / alignment
- Survey researchers know how to collect data
- Science-backed advice for AI/ML researchers about how to collect data

Data Work can be Undervalued in ML/AI

"Everyone wants to do the model work, not the data work"

Sambasivan et al, 2021 doi:10.1145/3411764.3445518

Data Centric AI

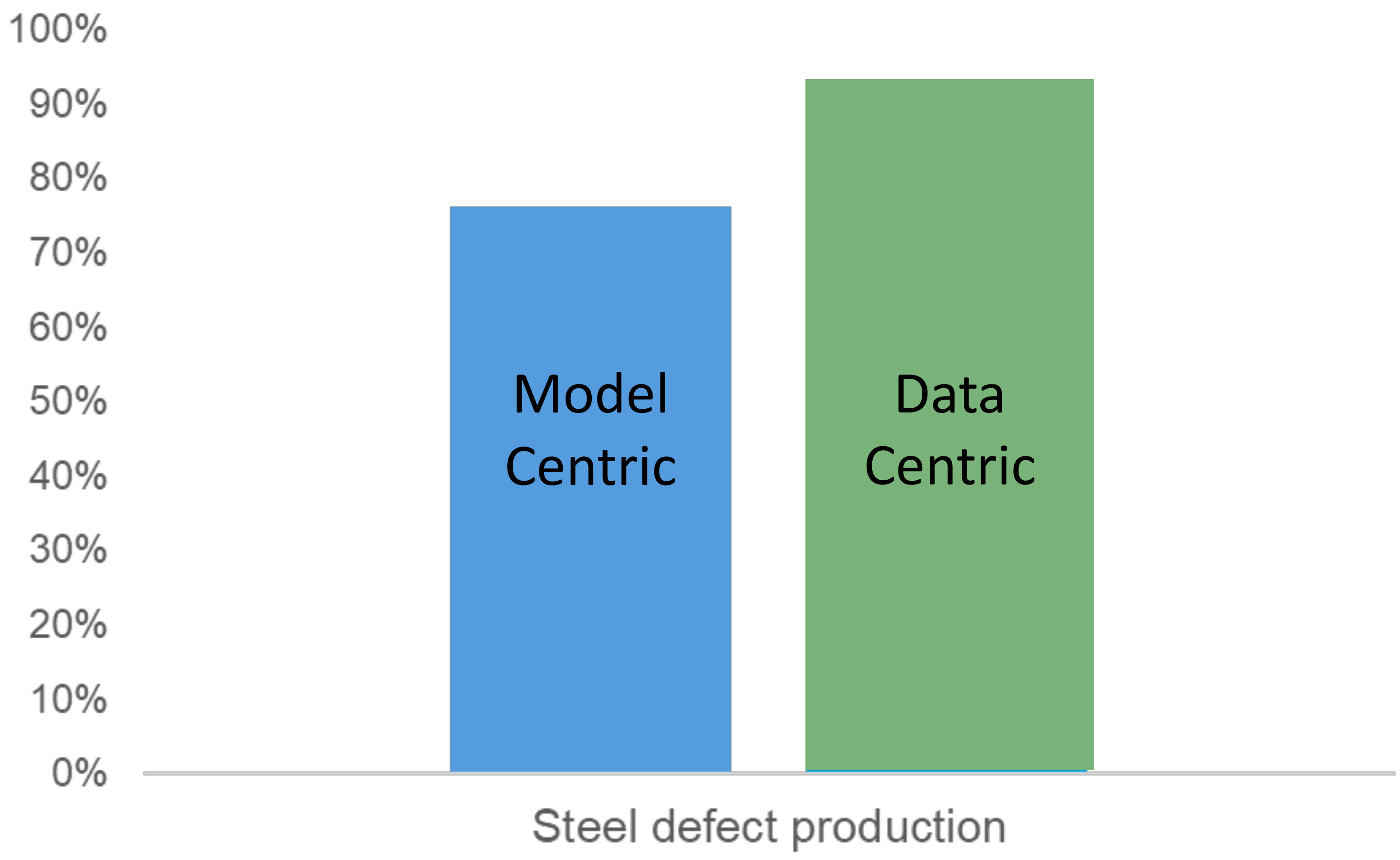


Sculley et. al. NIPS 2015: Hidden Technical Debt in Machine Learning Systems

Impact of Data Centric AI

- Andrew Ng:

"reasonable algorithm with good data is preferable to a great algorithm with not-so-good data"



Andrew Ng, Mar 24, 2021
https://youtu.be/06-AZXmwHjo

Training Data Collection

Types of Training Data

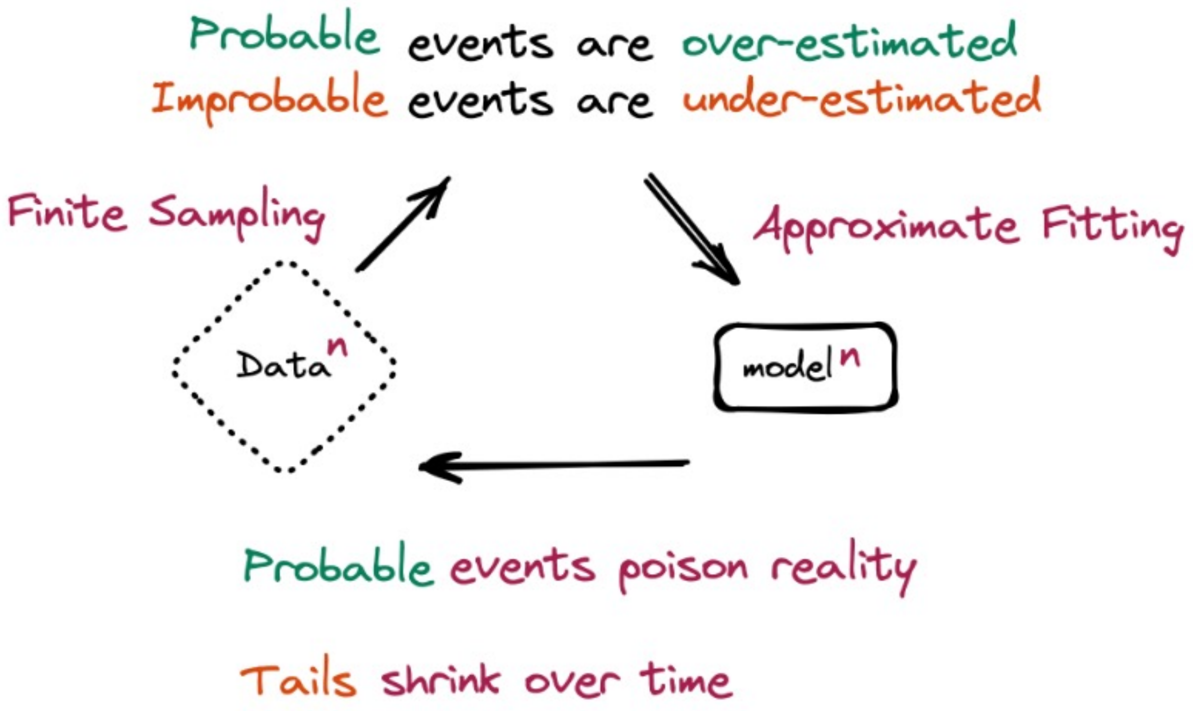
- NLP
  - Sentiment
  - Hate speech
  - Parts of speech
  - Translation
- Images
  - Existence of X in image
  - Bounding boxes
- RLHF
  - Value, accuracy of results
  - Fair
  - Unbiased
  - Aligned with human interests
- Robot movements
  - catch ball, put thing in bag

Training Data Sources

- Found data
- Federal sources: https://arxiv.org/pdf/2108.04884
- Web scraping
  - Legality in question:
    - NYT lawsuit against OpenAI
    - RIAA suit against music generation AI
  - Sites prohibit scraping (Reddit, Stack Overflow)
  - As AI text takes over internet, we're training models on data from models
- Human in the Loop

Can’t a Model Label my Data?

- Yes
- But:
  - Models trained on models trained on models
    - Model autophagy
    - Model collapse
- Combination
  - Most important, difficult labels still generated by humans



From: https://arxiv.org/pdf/2305.17493

“The bias I am most nervous about is the bias of the human feedback raters“

Sam Altman
March 25 2023 “The Lex Fridman Podcast“

Examples of Labeling Task

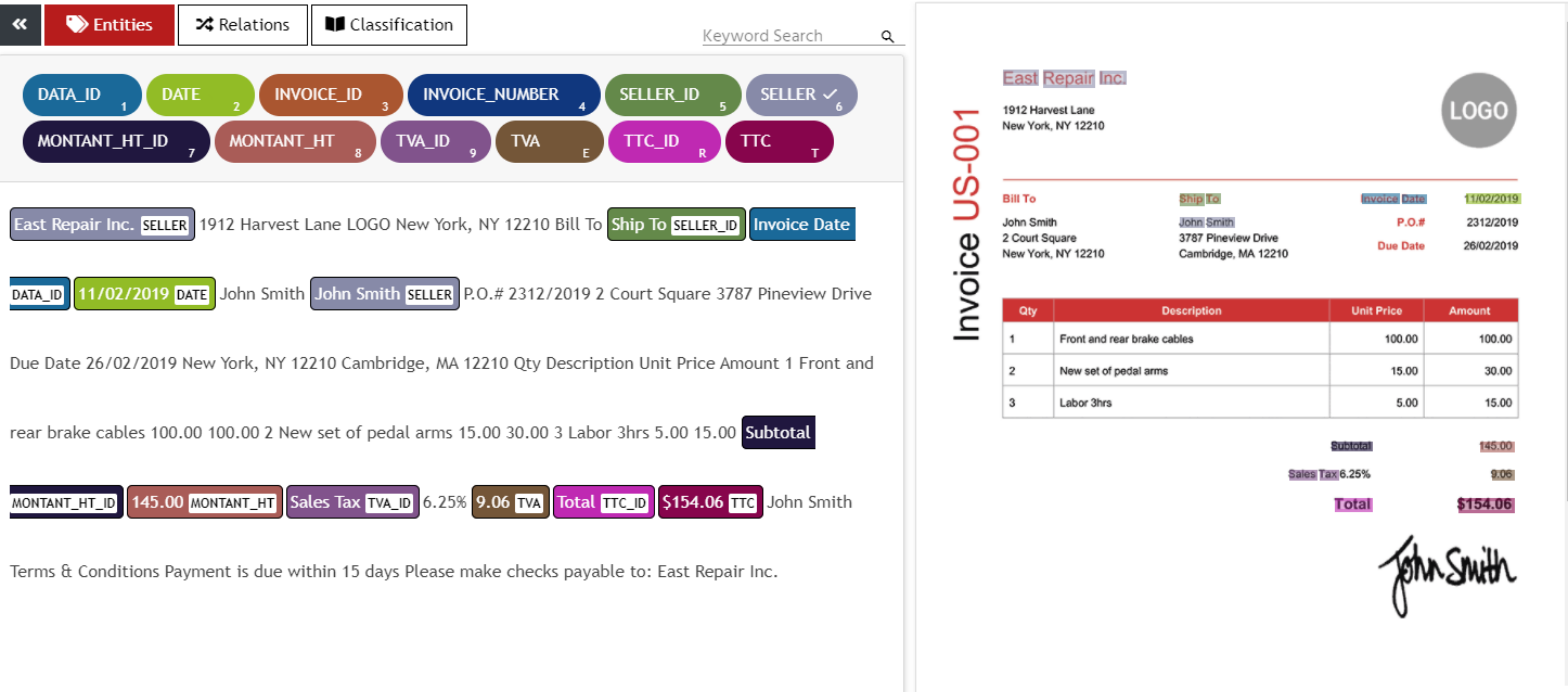
From: https://arxiv.org/abs/2203.02155

Search ...

- Bird
- Not_Bird
- Unable_to_tell

SUBMIT

https://aclanthology.org/2024.nlperspectives-1.1/



From: https://ubiai.tools/annotate-pdfs-and-images-for-nlp-applications-ubiai/

Labeler Stories

Venezuelan

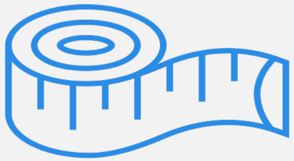
Kenyan

Syrian

Lessons from Surveys

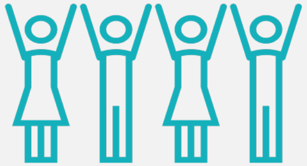
Survey Data Concerns

Training



Measurement: Are the answers correct?

labels



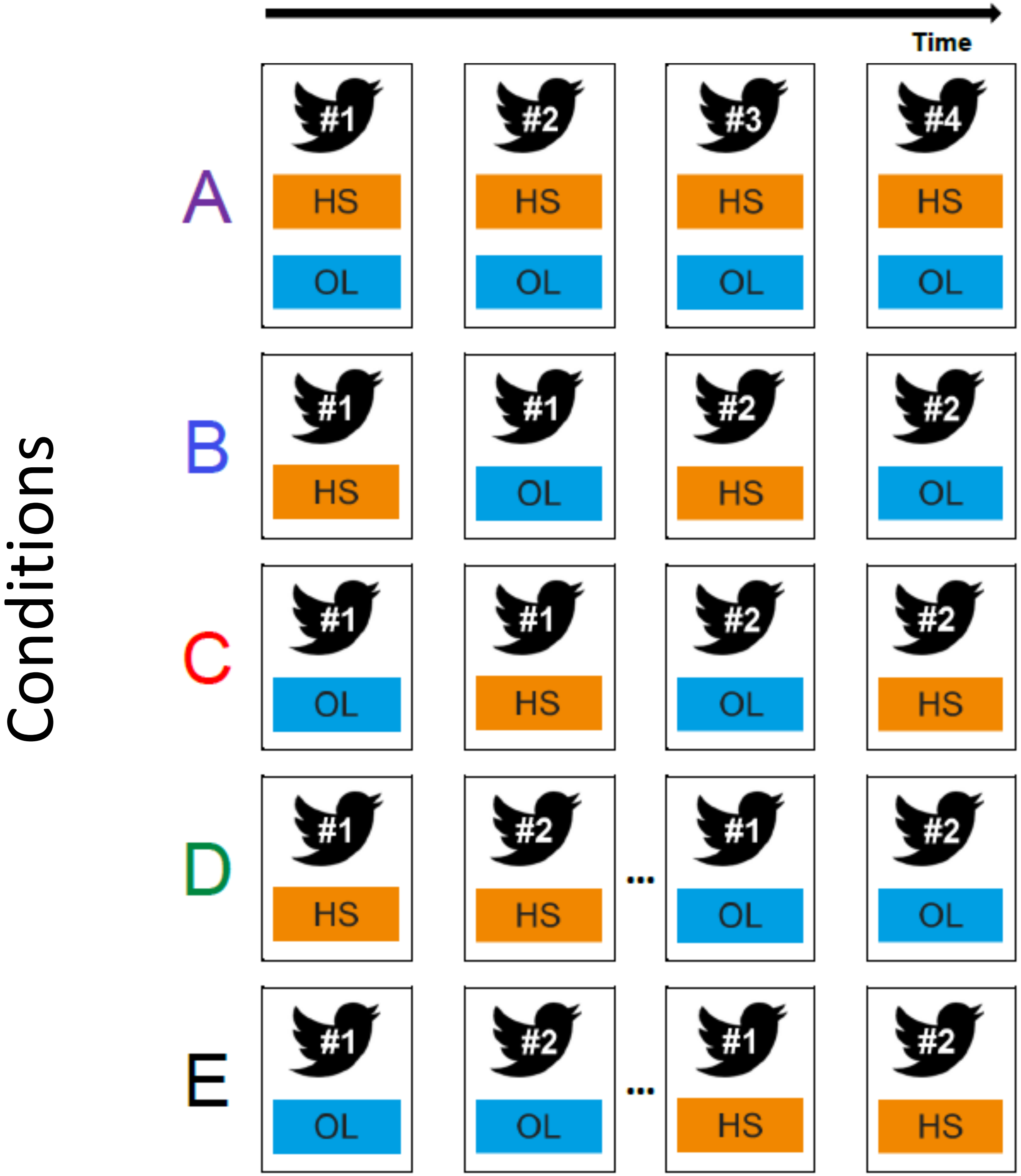
Representation: Who responds?

labels

From: https://arxiv.org/abs/2203.02155

Research design

https://arxiv.org/pdf/2311.14212



Data Collection

- 3000 tweets (Davidson et al 2017)
- ~900 labelers from Prolific (Nov-Dec 2022)

- 50 tweets / labeler
- 3 labels / tweet - condition
- 15 total labels / tweet

https://arxiv.org/pdf/2311.14212

Model Training

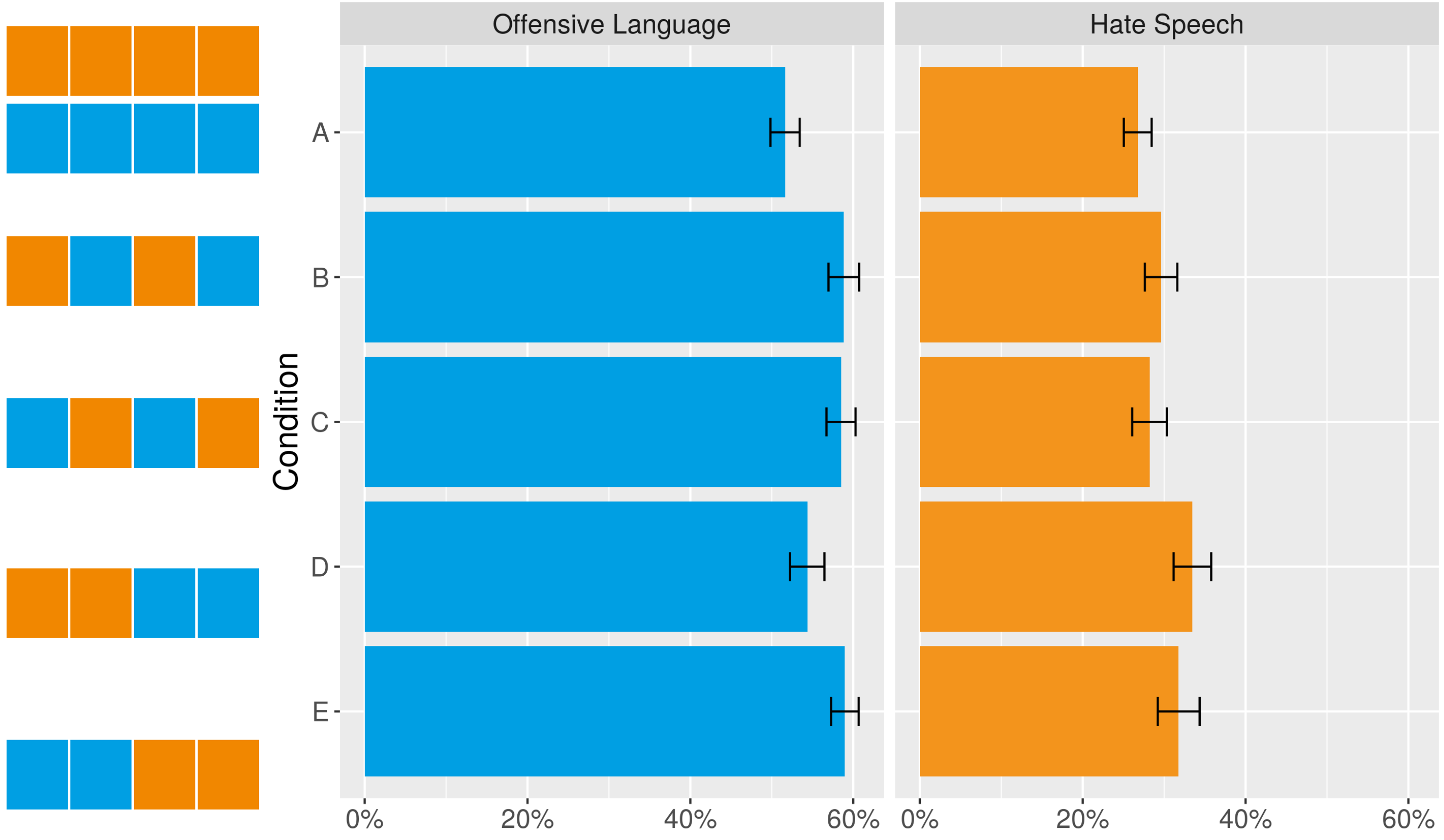
Training Set
N=2,250

Train A
Train B
Train C
Train D
Train E

Test Set
N=750

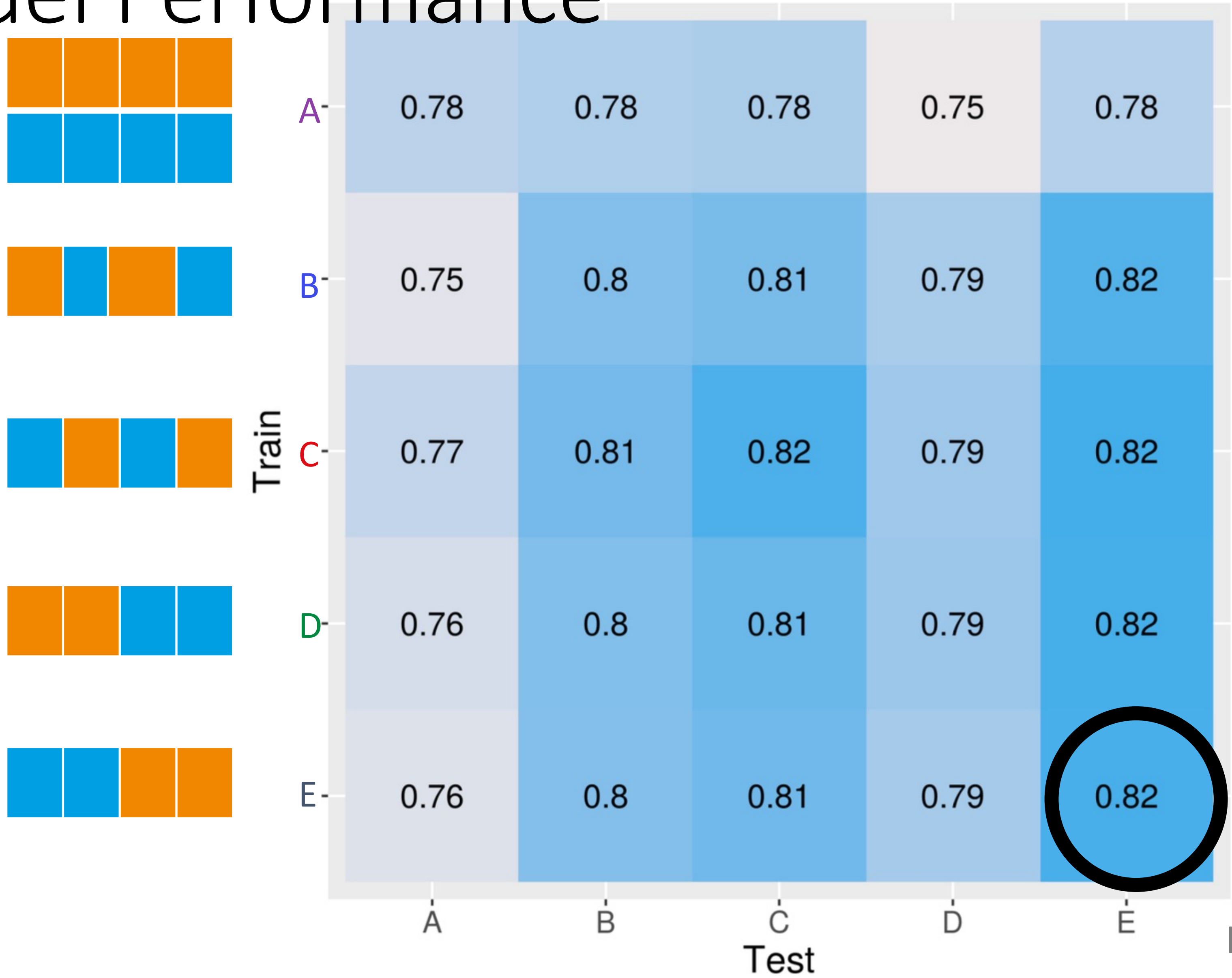
Test A
Test B
Test C
Test D
Test E

Labels



https://arxiv.org/pdf/2311.14212

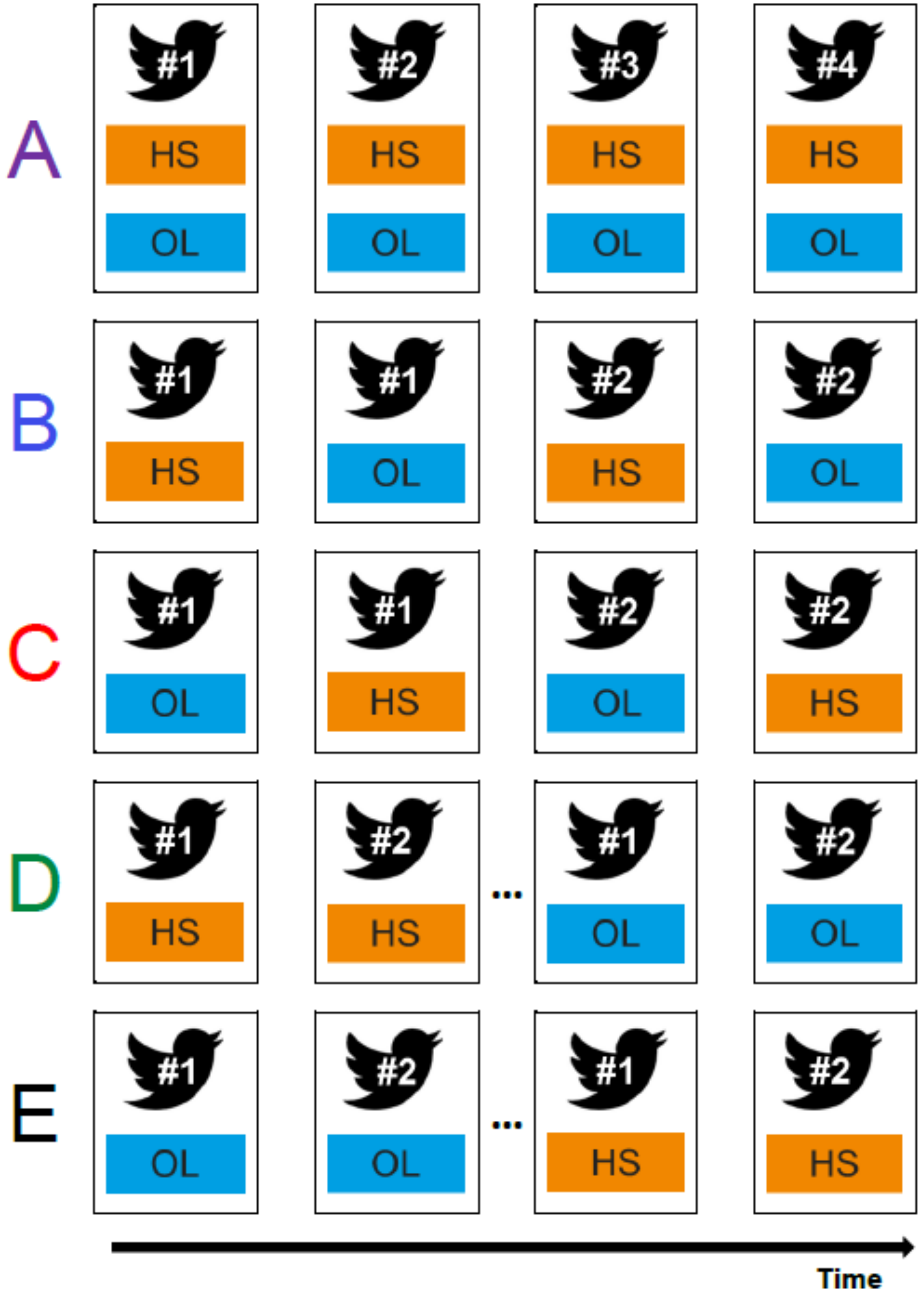
Model Performance



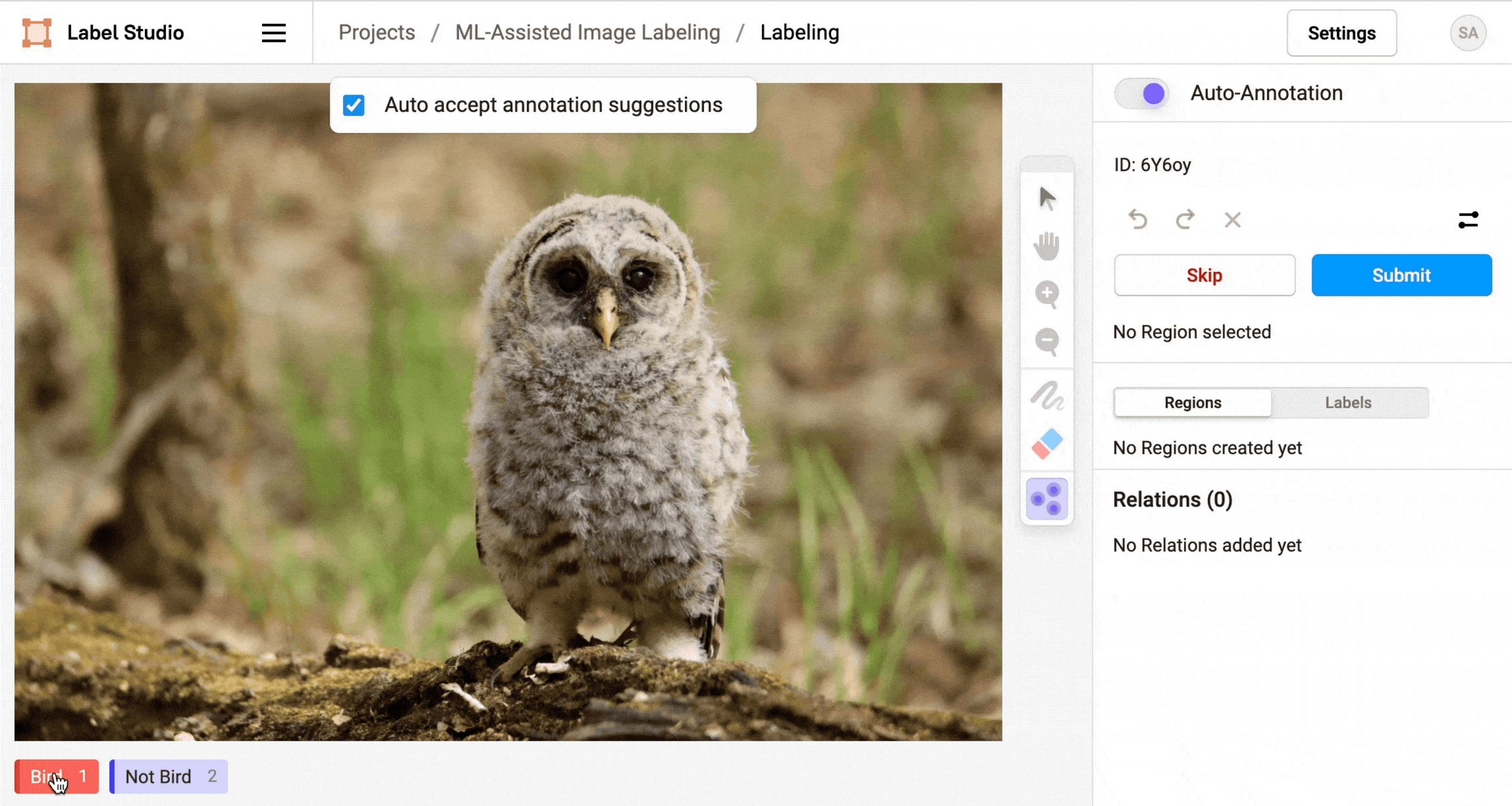
https://arxiv.org/pdf/2311.14212

Takeaways

- **A** not ideal
- Suggestions of order effects
  - **D** underperforms on OL
  - **E** underperforms on HS



Pre-Labeling



https://arunjitha.medium.com/integrating-labelstudio-in-react-machine-learning-applications-5dda72c79ce5

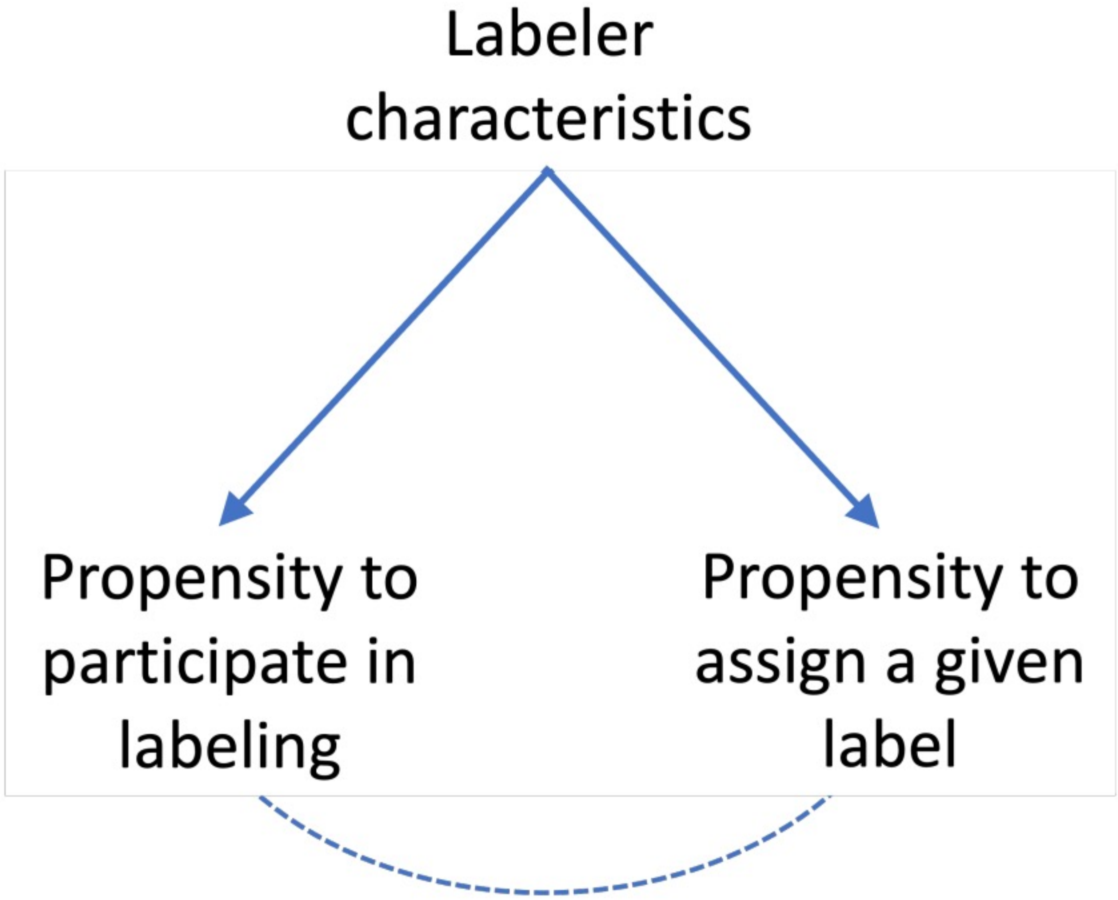
- Less expensive, faster
- Literature mixed on effect on quality
  - anchoring bias, complacency bias, pre-annotation bias

Who Labels?

- Experts
- Researchers, staff, students
- Crowdworkers
  - Appen, Sama, Upwork, Scale AI, Prolific, Mturk
  - Labelers tend to be from the Global South (Smart et al., 2014)
  - MTurk members younger, lower income than US pop (Berinsky et al., 2012)

Labeler Diversity

- Often train on modal label
- Is disagreement between labelers *signal* or *noise*?
- If labeler characteristics correlate with labels, then who labels matters

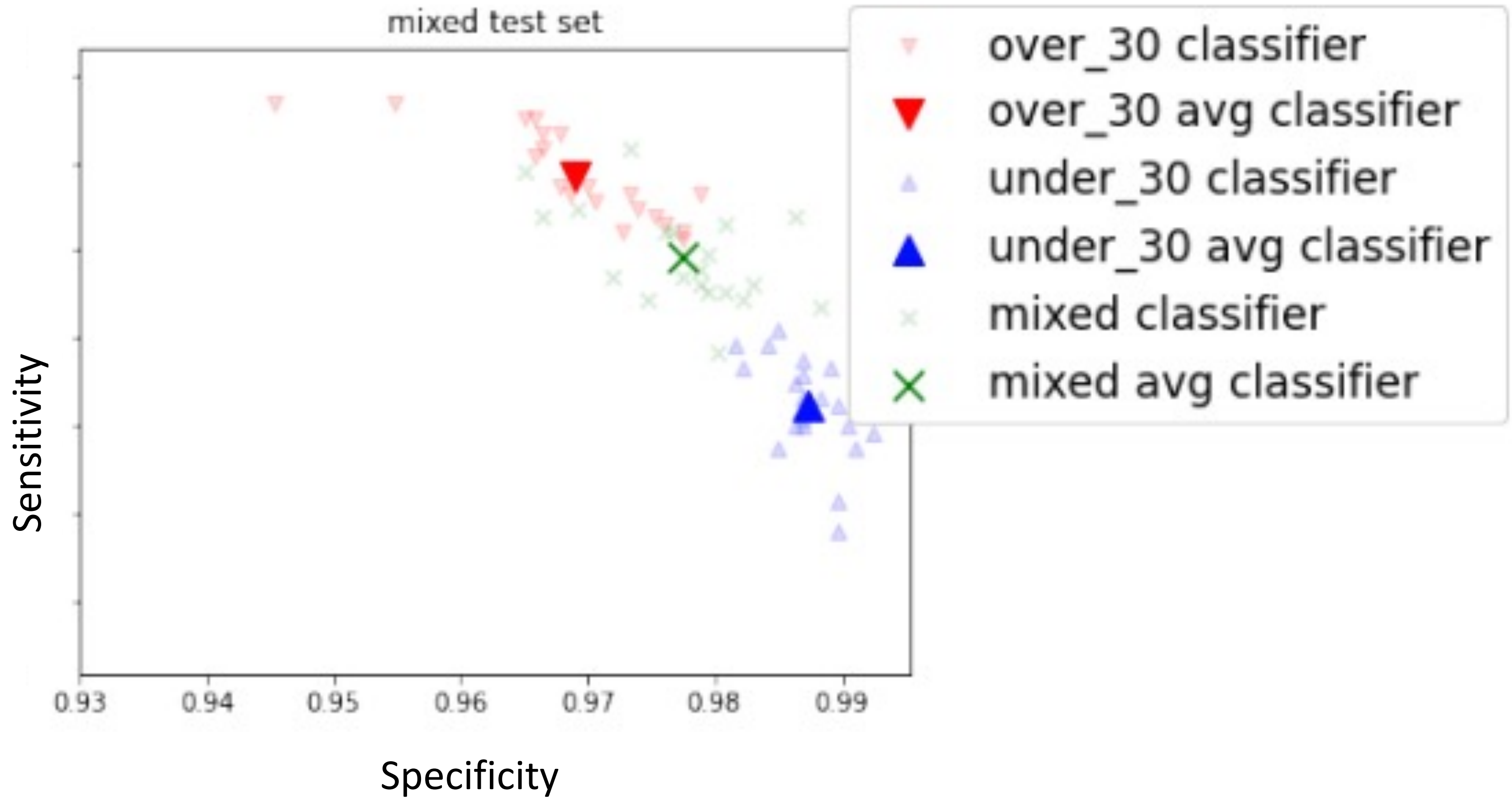


Selection Bias in Labels

Labelers	Trustworthiness
Male	Medium
Female	Low
Black	High
Asian	Low
White	High
Latino	High

Otterbacher et al "DESCANT: Detecting Stereotypes in Human Computational Tasks"

Selection Bias in Models



Al Kuwalty et al “Identifying and Measuring Annotator Bias Based on Annotators’ Demographic Characteristics”

Ethical Issues

- Wages, benefits
- Impact of exposure to terrible content
- Work often commissioned by those in wealthy, western countries, but carried out by those in lower income countries
  - Discipline and Label: A WEIRD Genealogy and Social Theory of Data Annotation https://arxiv.org/html/2402.06811v1
- When is human subjects approval needed?
  - https://mags.acm.org/communications/may_2024/MobilePagedReplica.action?=undefined&pm=2&folio=52#pg54
- https://data-workers.org/about/

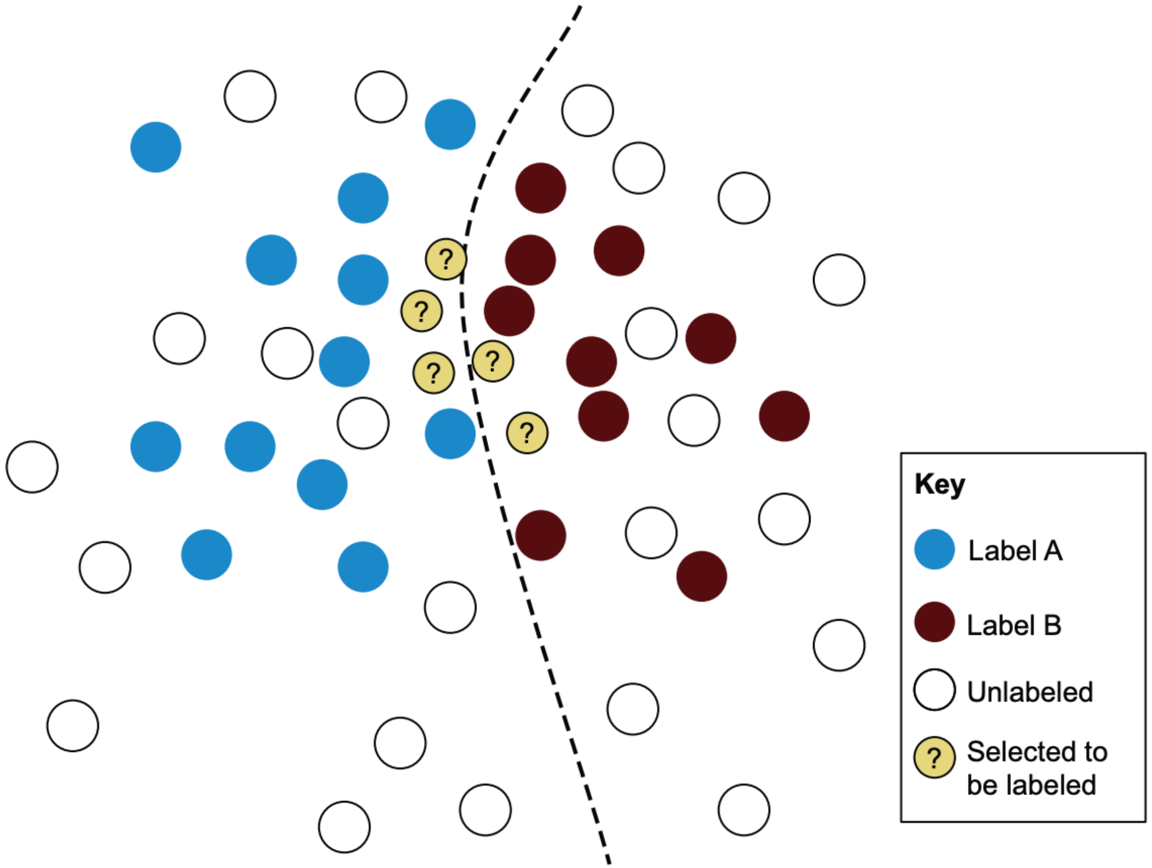
Recommendations for Responsible Design

1. Diversify your dataset and audit thoroughly
2. Strive for higher dataset quality
3. Start early and iterate
4. Document datasets openly and communicate limitations
5. Create user-centric datasets and limit inappropriate applications
6. Contend with privacy and consent
7. Make the datasets you need

Orr & Crawford https://arxiv.org/pdf/2409.00252

Sampling Instances to Label

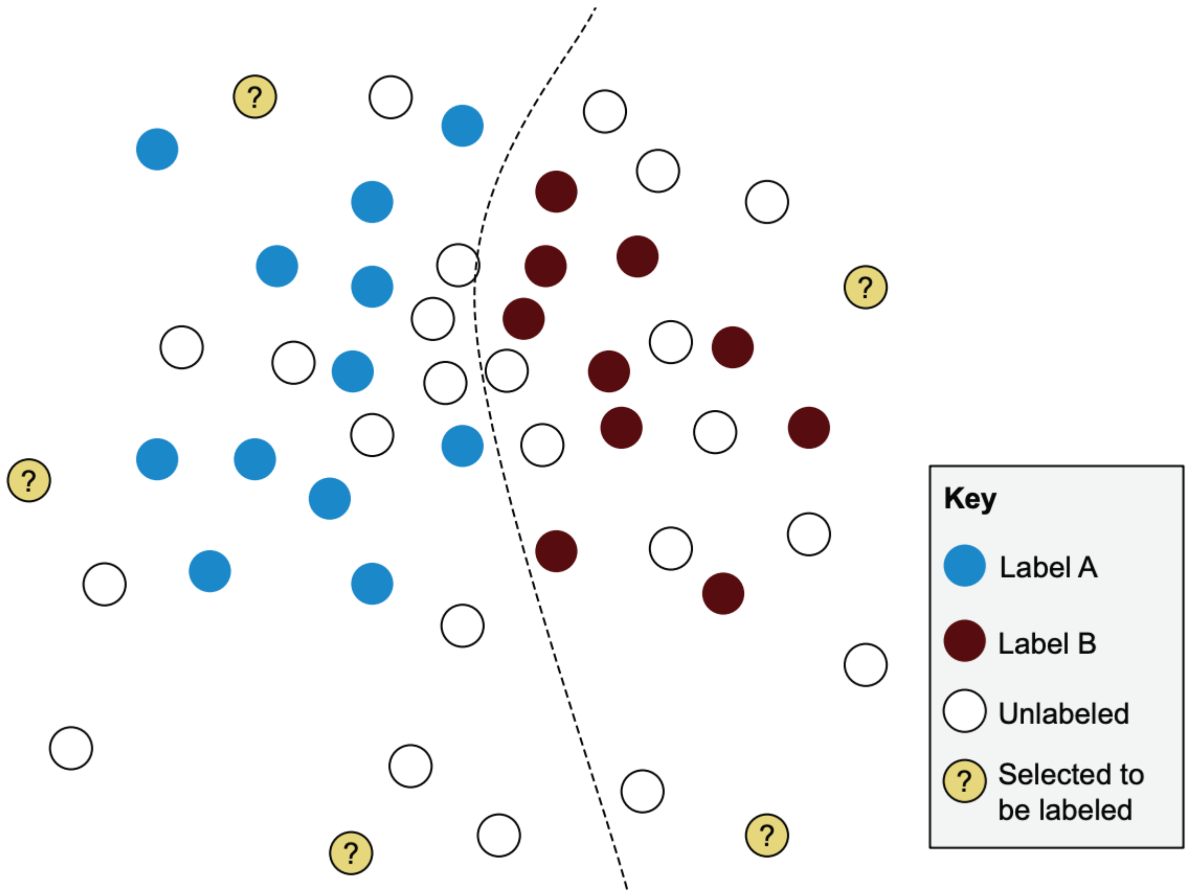
- Simple random sample
- Uncertainty sampling



Monarch (2021) Human-in-the-Loop Machine Learning

Sampling Instances to Label

- Simple random sample
- Uncertainty sampling
- Diversity sampling



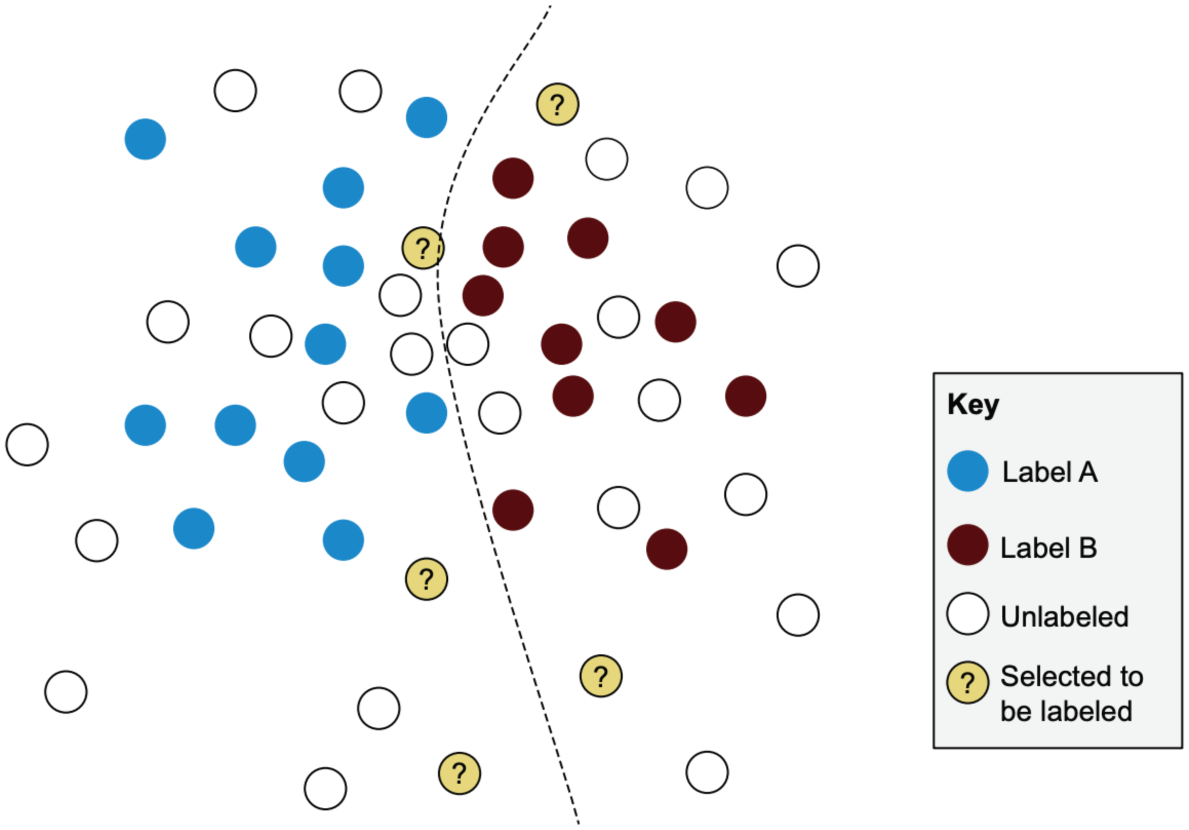
Monarch (2021) Human-in-the-Loop Machine Learning

Sampling Instances to Label

- Simple random sampling
- Uncertainty sampling
- Diversity sampling
- Stratified sampling
  - Called clustering sampling in Monarch (2021)

Sampling Instances to Label

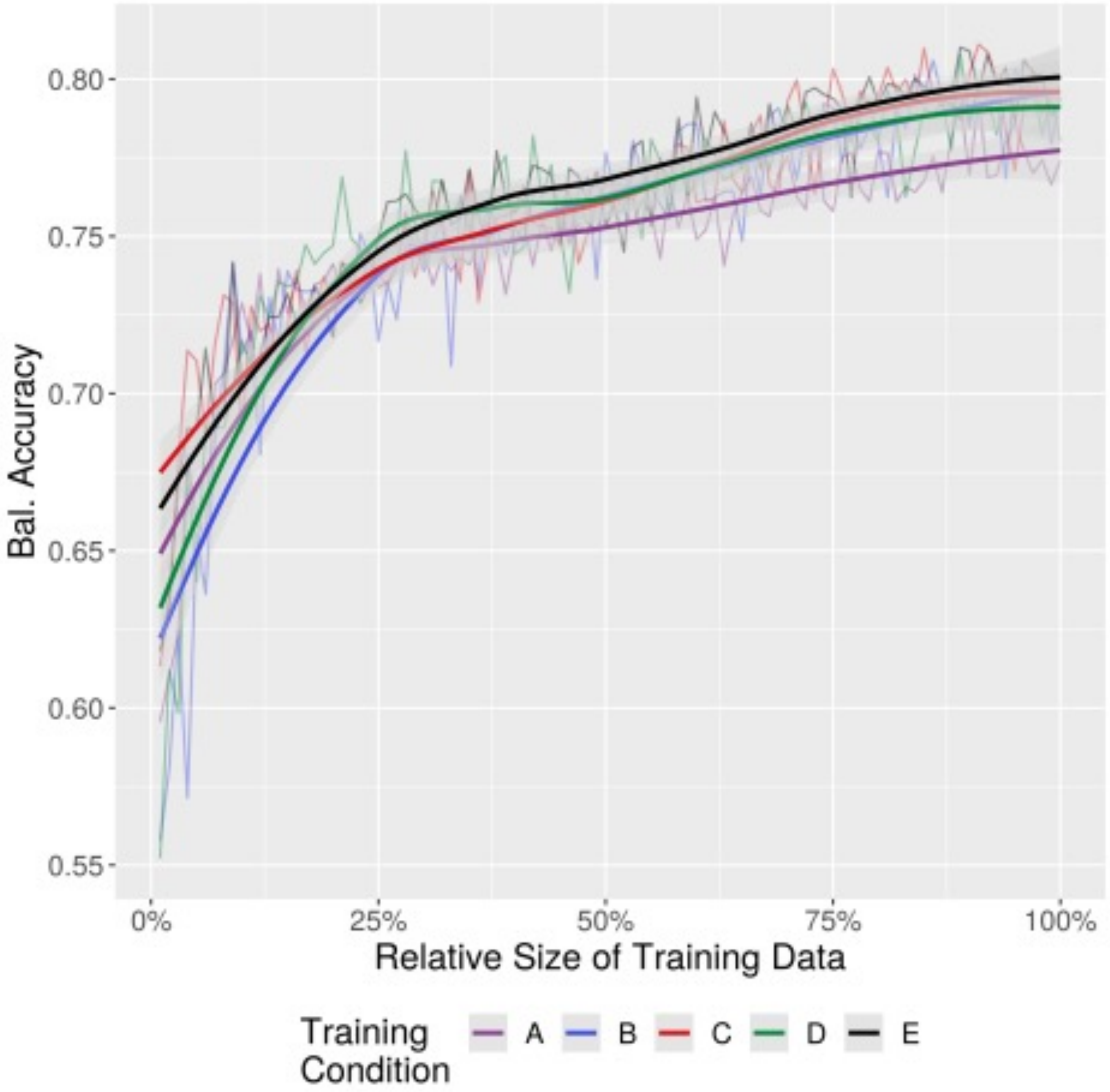
- Simple random sampling
- Uncertainty sampling
- Diversity sampling
- Stratified sampling
- Active Learning = uncertainty + diversity sampling



Monarch (2021) Human-in-the-Loop Machine Learning

How Many Labels do you Need?

- Depends on:
  - Model type
  - Complexity of task
  - Quality of data



Discussion:
Questions, Insights, Ideas

Tools

- Open-source vs commercial
- Type of label
- Most do not collect labeler characteristics
- Most do not collect paradata
  - Timing
  - Labeler id
  - Order of instances

Concrete Advice

- Pay adequately
- Recruit many, diverse labelers
- Simplify instructions: 8^{th} grade reading level
- Simplify instrument: Follow survey design practices
- Test instructions and instrument before collecting data