

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

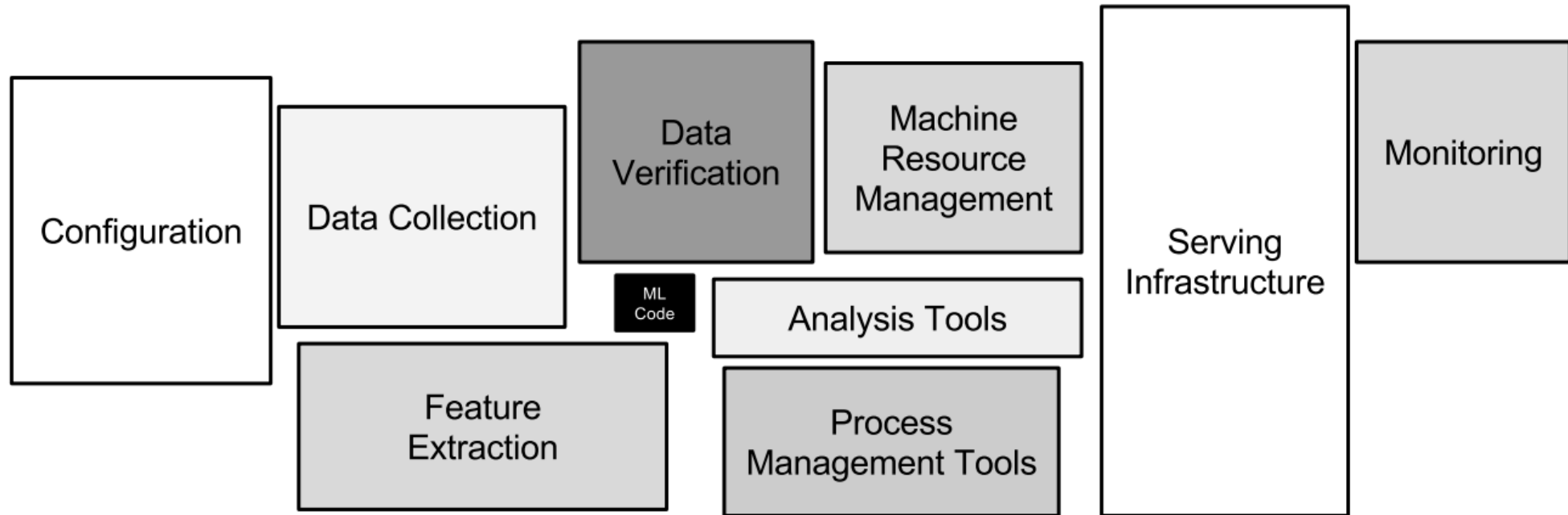
- Better data quality can improve model performance / alignment
- Survey researchers know how to collect data
- Science-backed advice for AI/ML researchers about how to collect data

Data Work can be Undervalued in ML/AI

“Everyone wants to do
the model work, not the
data work”

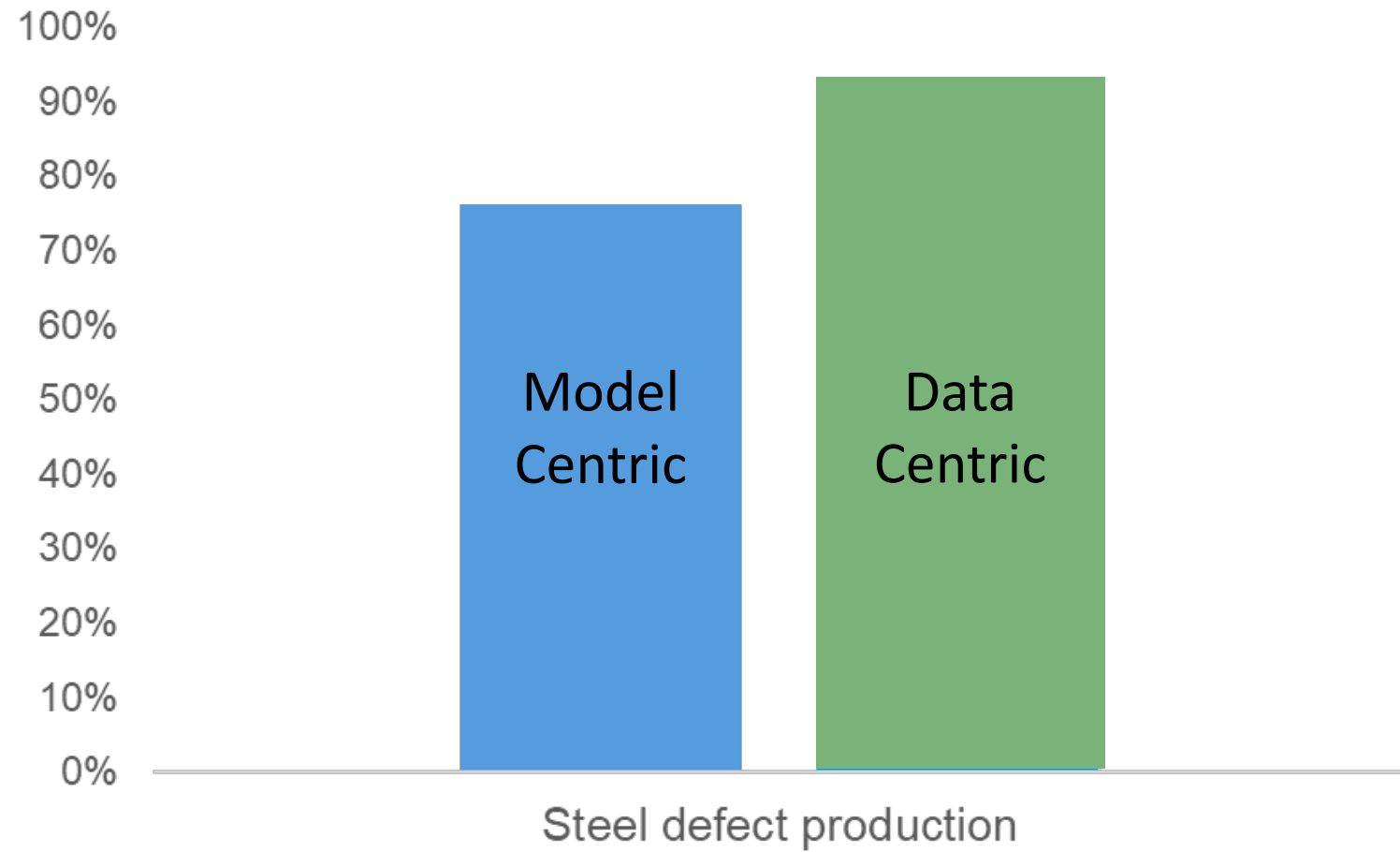
Sambasivan et al, 2021 doi:10.1145/3411764.3445518

Data Centric AI



Impact of Data Centric AI

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



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

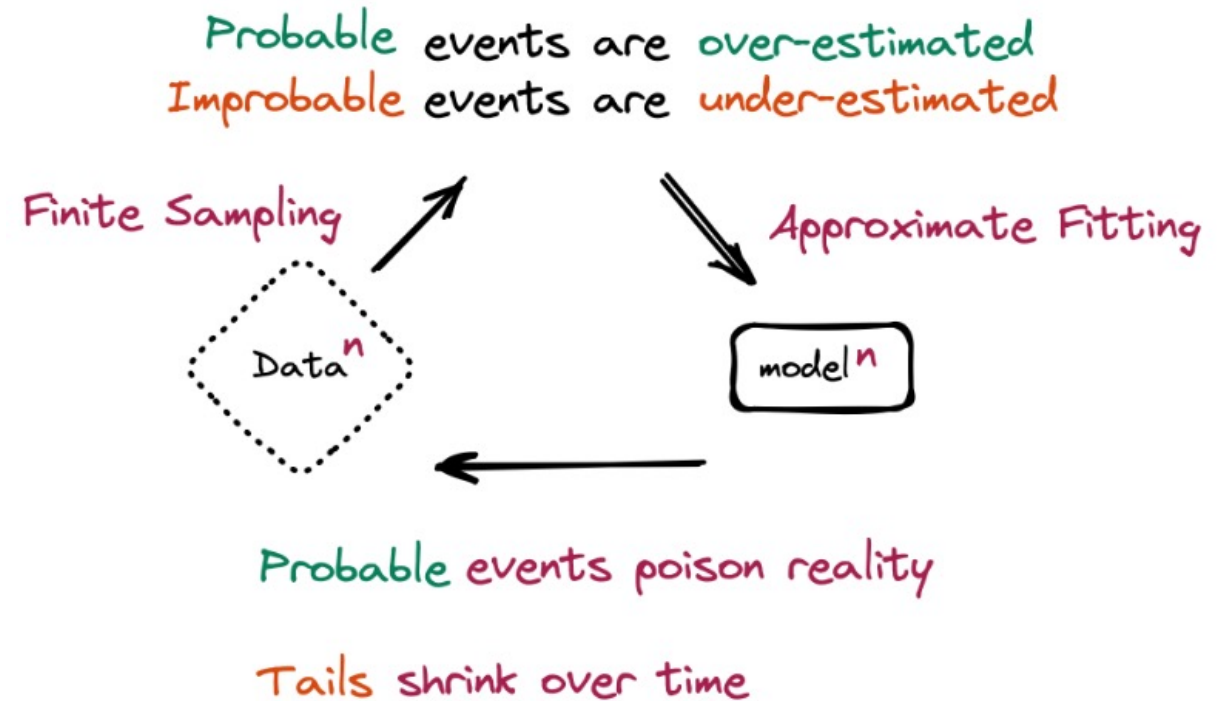
- catch ball, put thing in bag

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

Submit

Skip

«

Page 3 / 11

»

Total time: 05:39

Instruction

Summarize the following news article:

```
====  
{article}  
====
```

Article text here

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

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

Notes

(Optional) notes



Search ...

- Bird
- Not_Bird
- Unable_to_tell

SUBMIT

DATA_ID 1
DATE 2
INVOICE_ID 3
INVOICE_NUMBER 4
SELLER_ID 5
SELLER ✓ 6
MONTANT_HT_ID 7
MONTANT_HT 8
TVA_ID 9
TVA E
TTC_ID R
TTC T

East Repair Inc. SELLER 1912 Harvest Lane LOGO New York, NY 12210 Bill To Ship To SELLER_ID Invoice Date

DATA_ID 11/02/2019 DATE John Smith John Smith SELLER P.O.# 2312/2019 2 Court Square 3787 Pineview Drive


Due Date 26/02/2019 New York, NY 12210 Cambridge, MA 12210 Qty Description Unit Price Amount 1 Front and

rear brake cables 100.00 100.00 2 New set of pedal arms 15.00 30.00 3 Labor 3hrs 5.00 15.00 Subtotal

MONTANT_HT_ID 145.00 MONTANT_HT Sales Tax TVA_ID 6.25% 9.06 TVA Total TTC_ID \$154.06 TTC John Smith

Terms & Conditions Payment is due within 15 days Please make checks payable to: East Repair Inc.


Invoice US-001



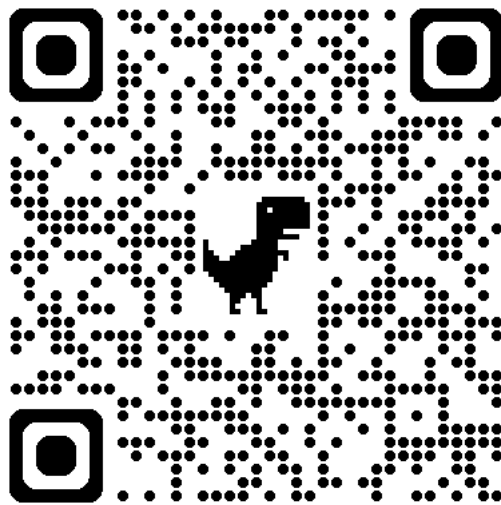
East Repair Inc.
1912 Harvest Lane
New York, NY 12210

Bill To	Ship To	Invoice Date 11/02/2019
John Smith 2 Court Square New York, NY 12210	John Smith 3787 Pineview Drive Cambridge, MA 12210	P.O.# 2312/2019 Due Date 26/02/2019

Qty	Description	Unit Price	Amount
1	Front and rear brake cables	100.00	100.00
2	New set of pedal arms	15.00	30.00
3	Labor 3hrs	5.00	15.00
Subtotal			145.00
Sales Tax 6.25%			9.06
Total			\$154.06



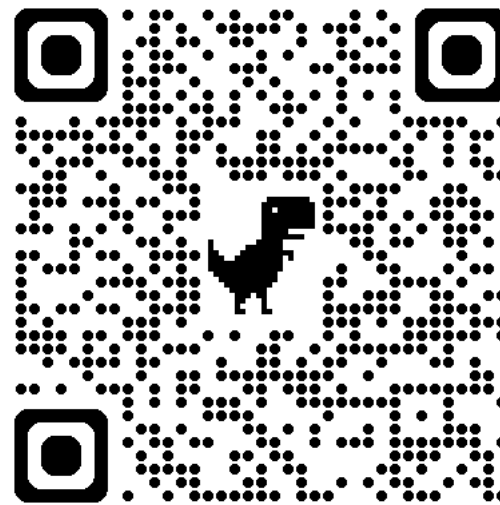
Labeler Stories



Venezuelan



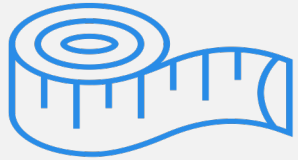
Kenyan



Syrian

Lessons from Surveys

Training Survey Data Concerns



Measurement: Are the **labels** correct?



Representation: Who **labels**?

Submit

Skip

«

Page 3 / 11

»

Total time: 05:39

Instruction

Summarize the following news article:

```
====  
{article}  
====
```

Article text here

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

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

Notes

(Optional) notes

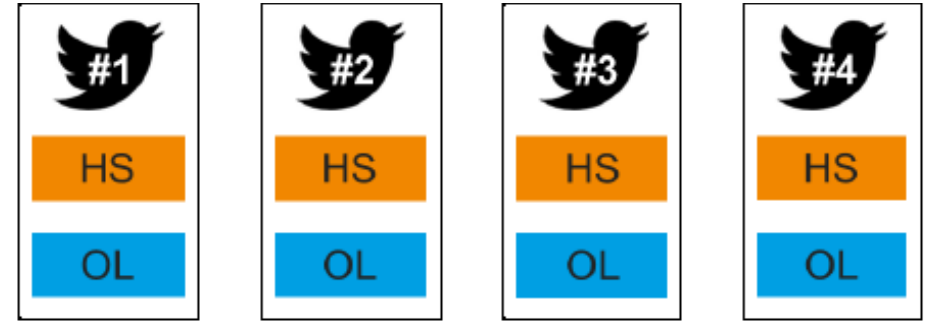
Research design



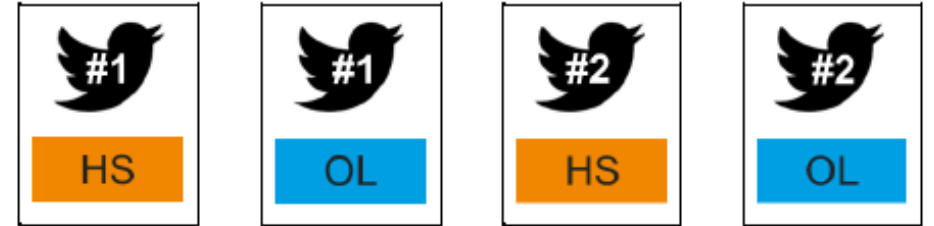
<https://arxiv.org/pdf/2311.14212>

Conditions

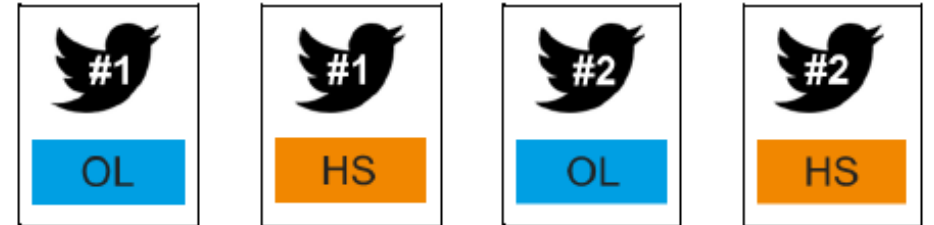
A



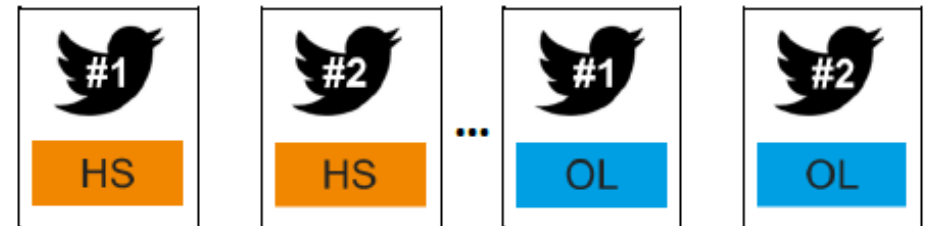
B



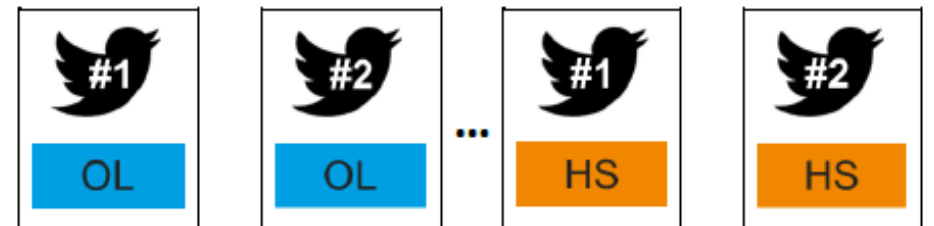
C



D



E



Data Collection

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

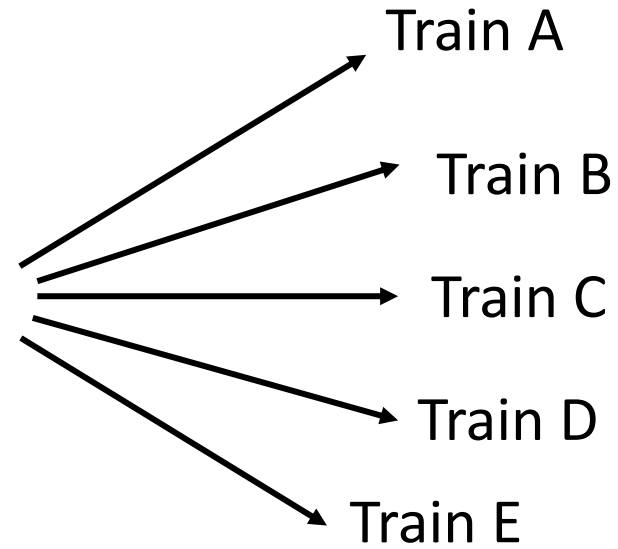
- 50 tweets / labeler
- 3 labels / tweet - condition
- 15 total labels / tweet

<https://arxiv.org/pdf/2311.14212>

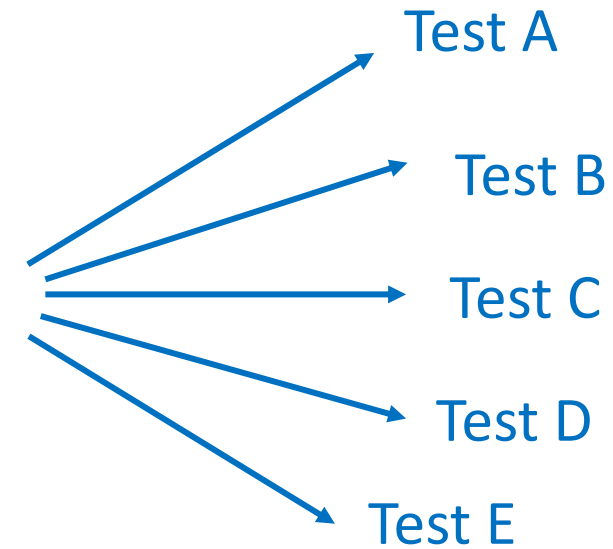
Model Training



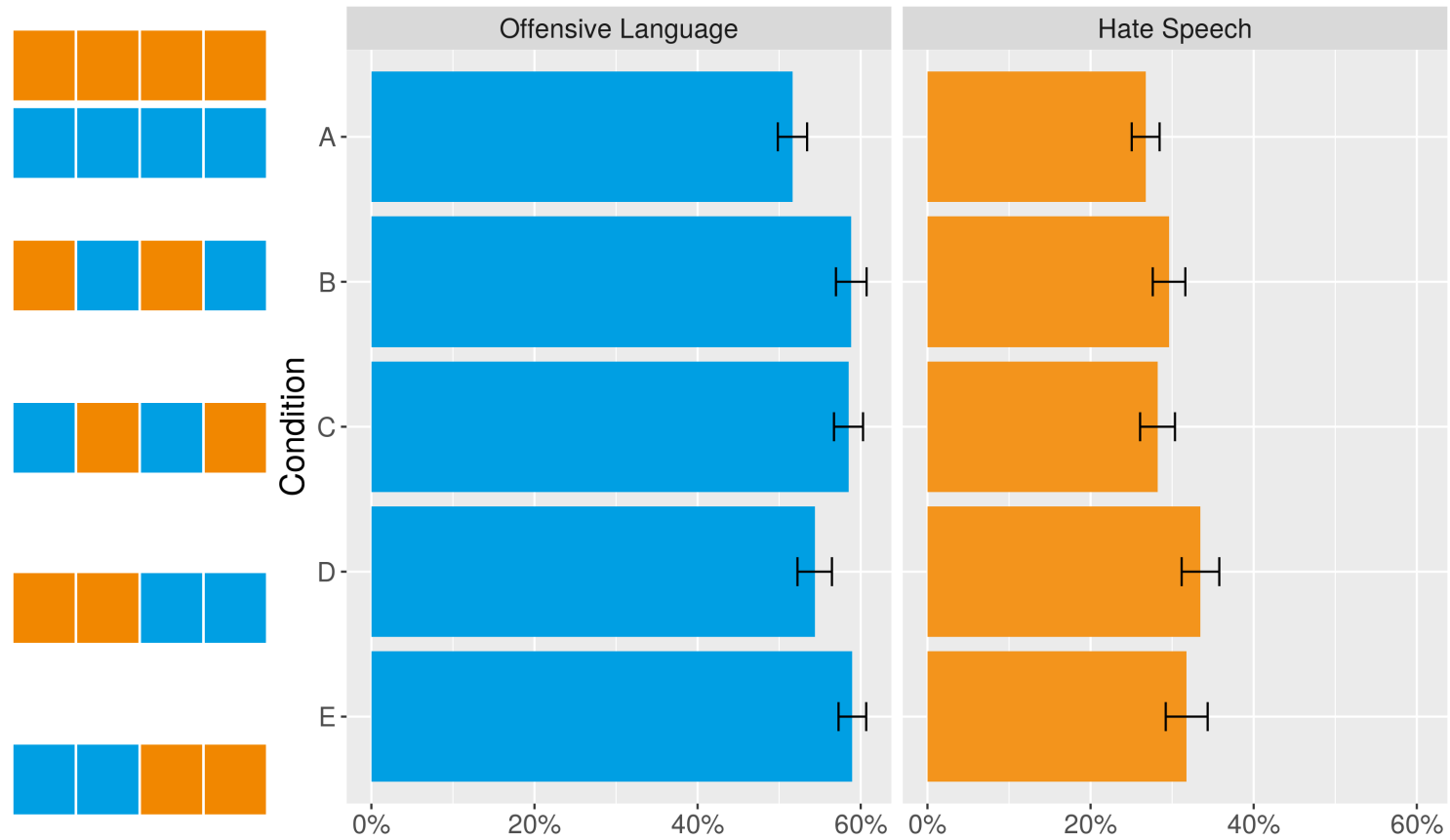
Training Set
N=2,250



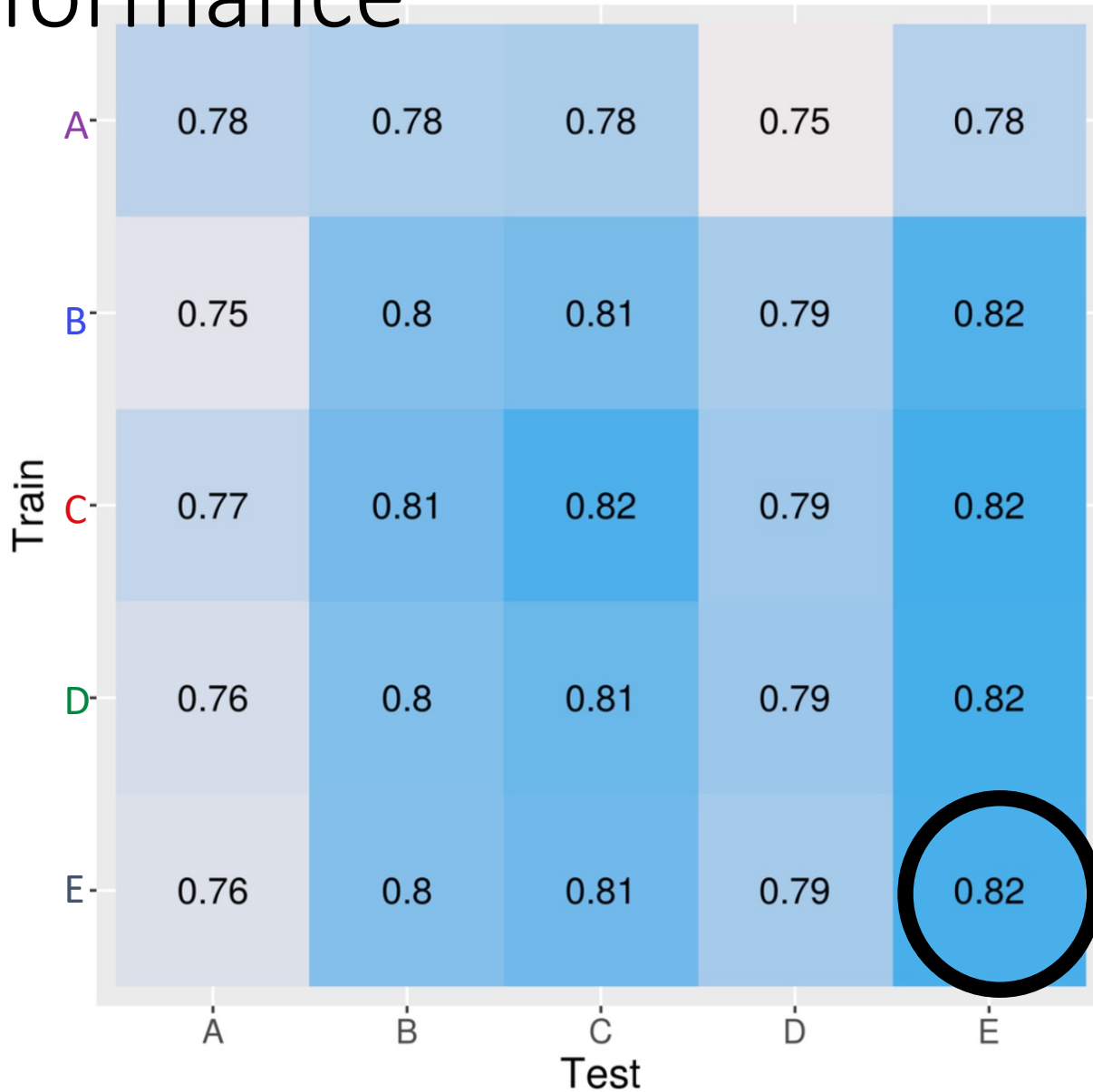
Test Set
N=750



Labels

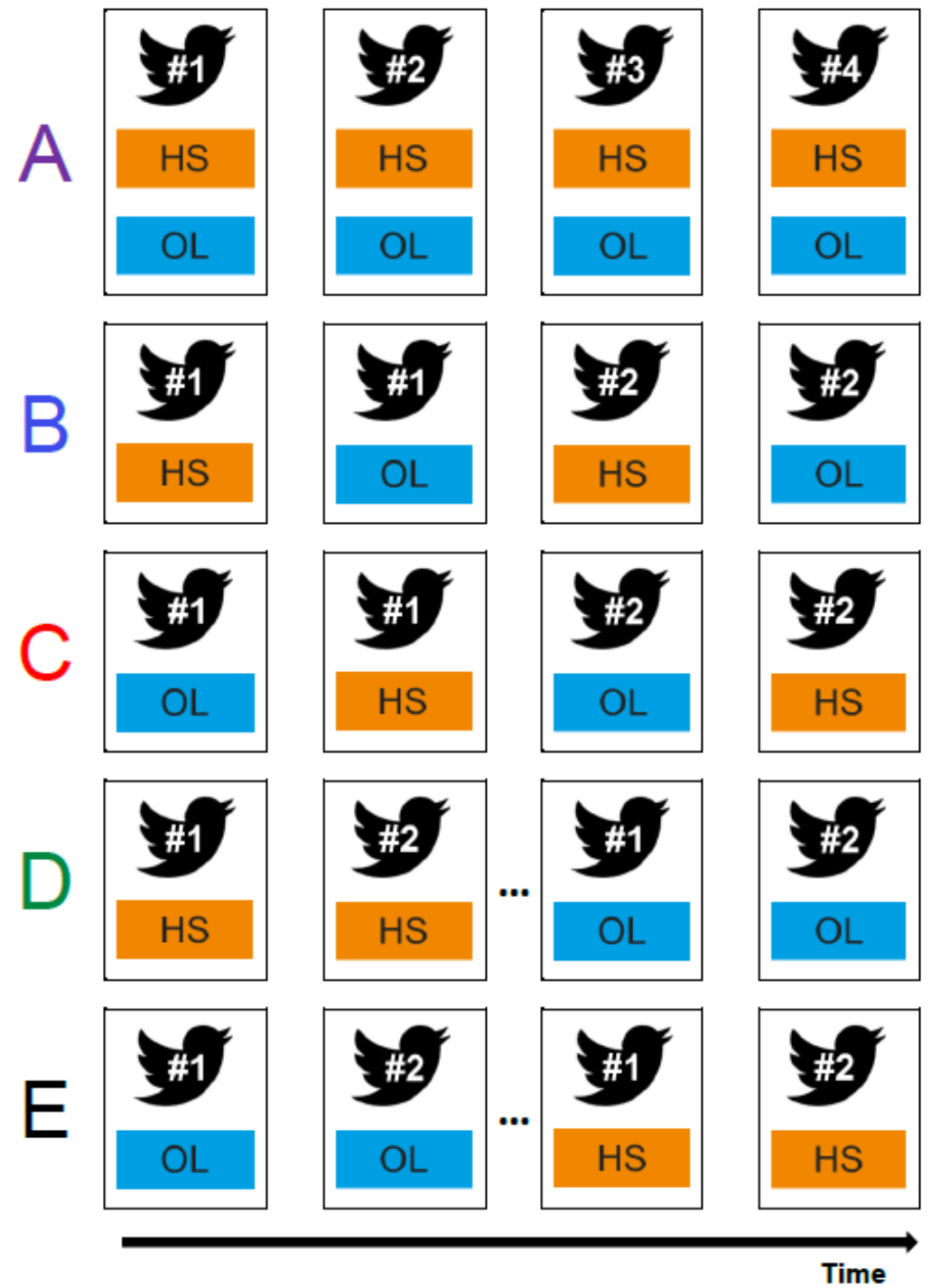


Model Performance

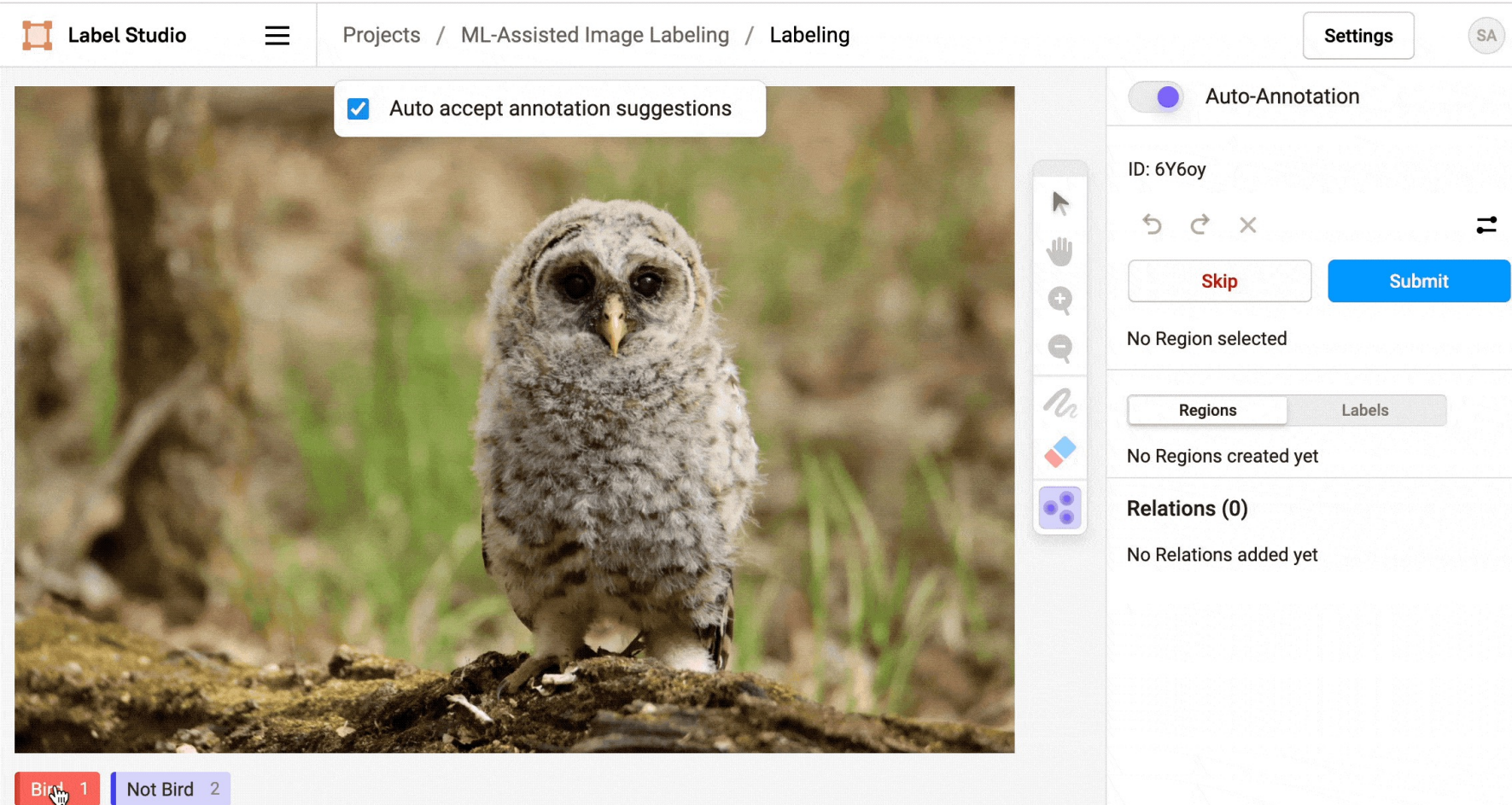


Takeaways

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



Pre-Labeling



The screenshot displays the Label Studio web interface. At the top, the navigation bar includes the Label Studio logo, a menu icon, and the breadcrumb path: Projects / ML-Assisted Image Labeling / Labeling. On the right side of the navigation bar, there are buttons for 'Settings' and a user profile icon labeled 'SA'. The main workspace is divided into three sections. The left section shows a large image of a fluffy, grey bird chick perched on a rock. Above the image is a white notification box with a blue checkmark and the text 'Auto accept annotation suggestions'. Below the image, there are two buttons: a red 'Bird 1' button and a blue 'Not Bird 2' button. The middle section is a vertical toolbar with icons for various actions: a mouse cursor, a hand, a plus sign, a minus sign, a magnifying glass, a diamond, and a cluster of four dots. The right section is a control panel. It features a toggle switch for 'Auto-Annotation' which is currently turned on. Below this, the image ID '6Y6oy' is displayed. There are navigation icons (back, forward, close) and a refresh icon. Two buttons, 'Skip' (red) and 'Submit' (blue), are present. Below these buttons, it says 'No Region selected'. There are two tabs, 'Regions' and 'Labels', with 'Regions' selected. Below the tabs, it says 'No Regions created yet'. At the bottom of the control panel, it says 'Relations (0)' and 'No Relations added yet'.

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

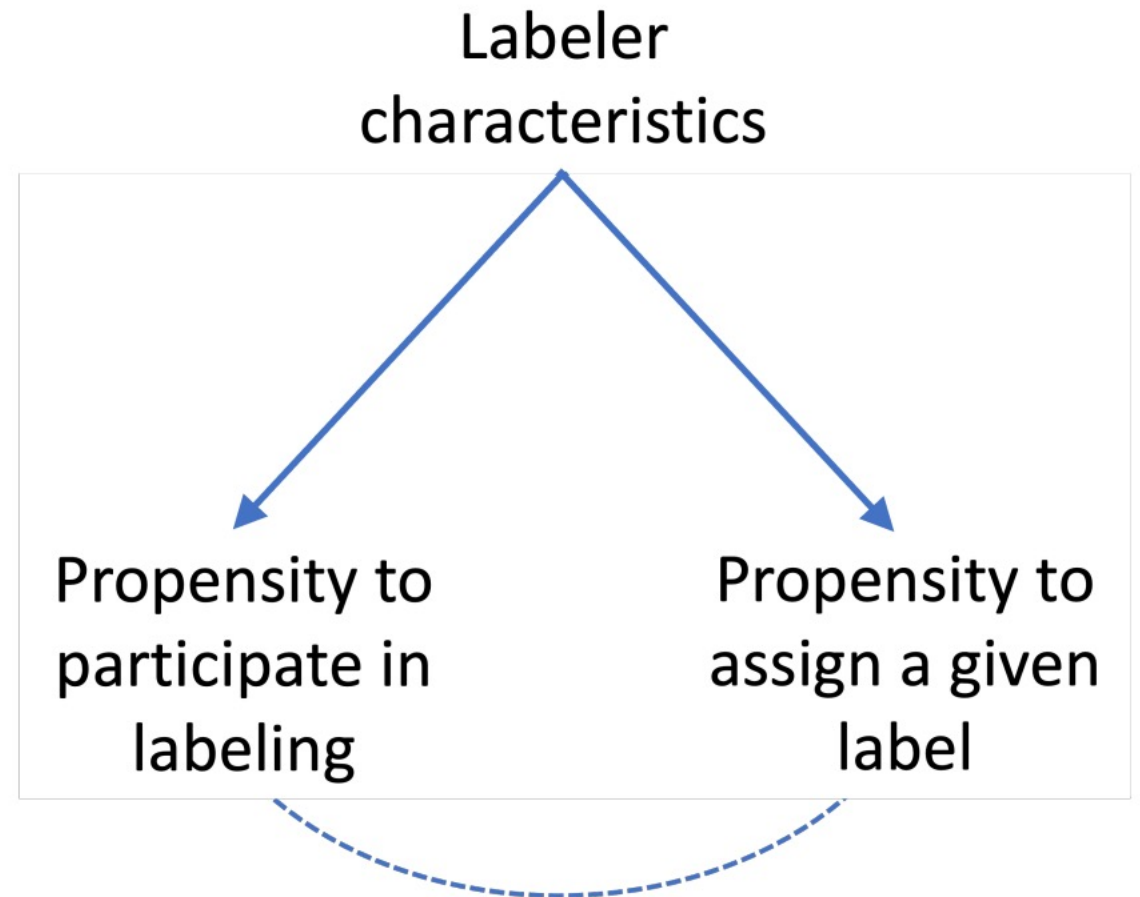
<https://arunjitha.medium.com/integrating-labelstudio-in-react-machine-learning-applications-5dda72c79ce5>

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

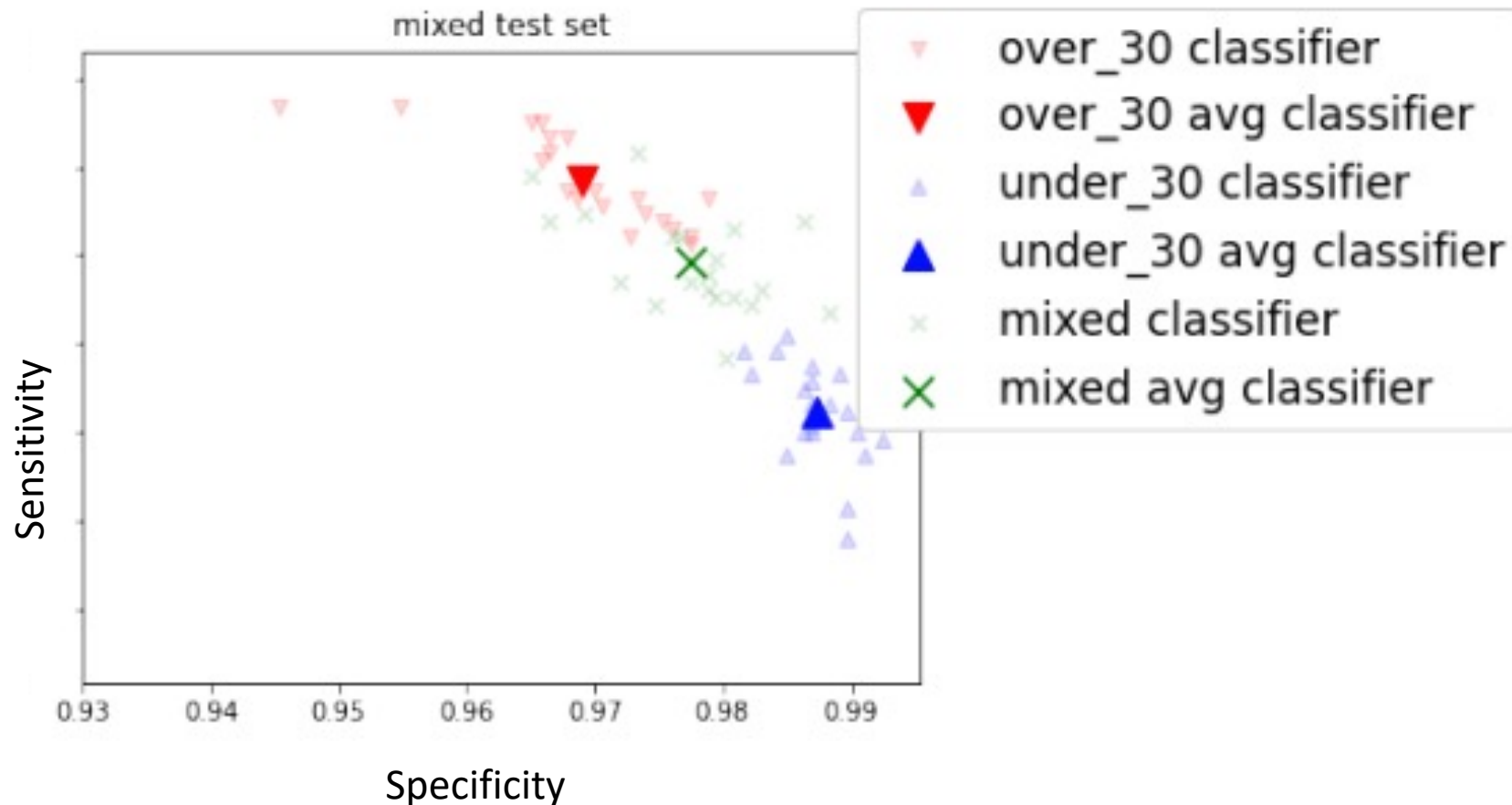


Selection Bias in Labels



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

Selection Bias in Models



Ethical Issues

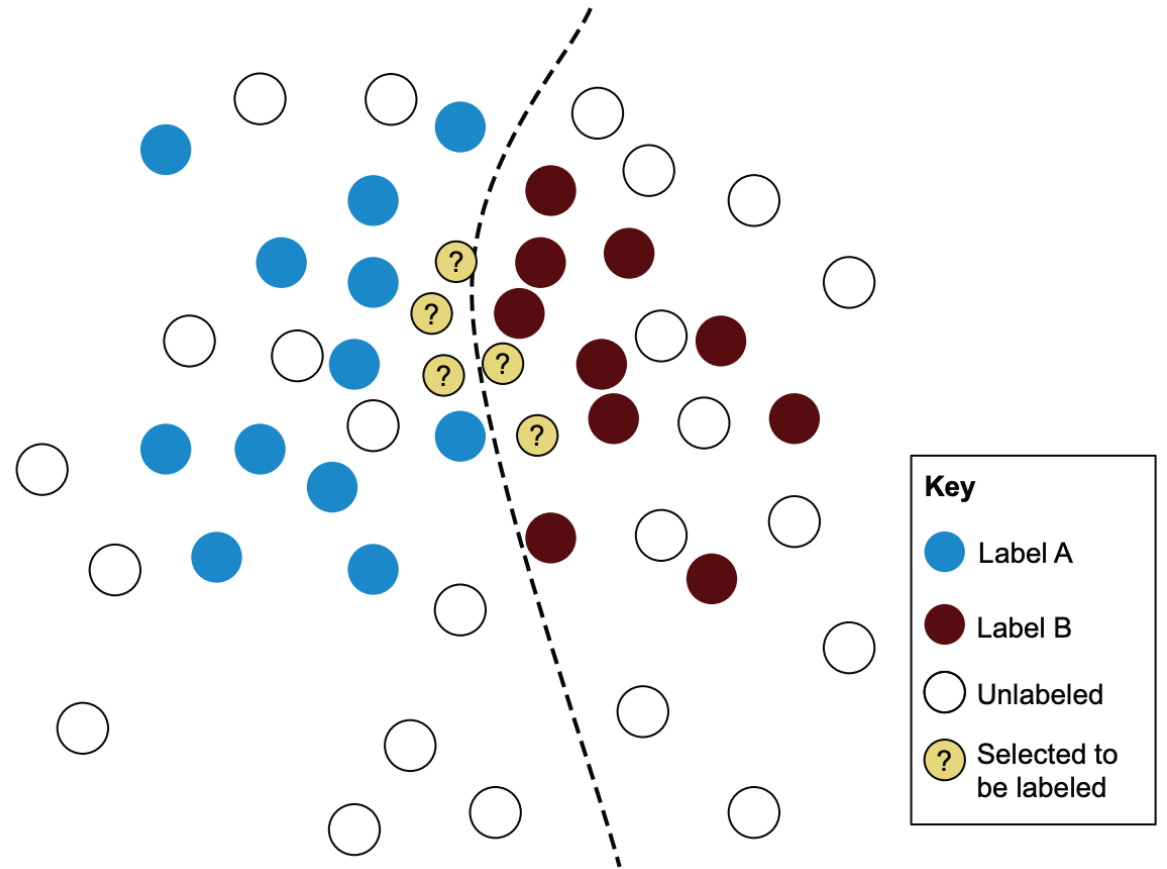
- Wages, benefits
- Impact of exposure to terrible content
- Work often commissioned by those in wealthy, western countries, but carried out by those in lower income countries
 - Discipline and Label: A WEIRD Genealogy and Social Theory of Data Annotation <https://arxiv.org/html/2402.06811v1>
- When 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

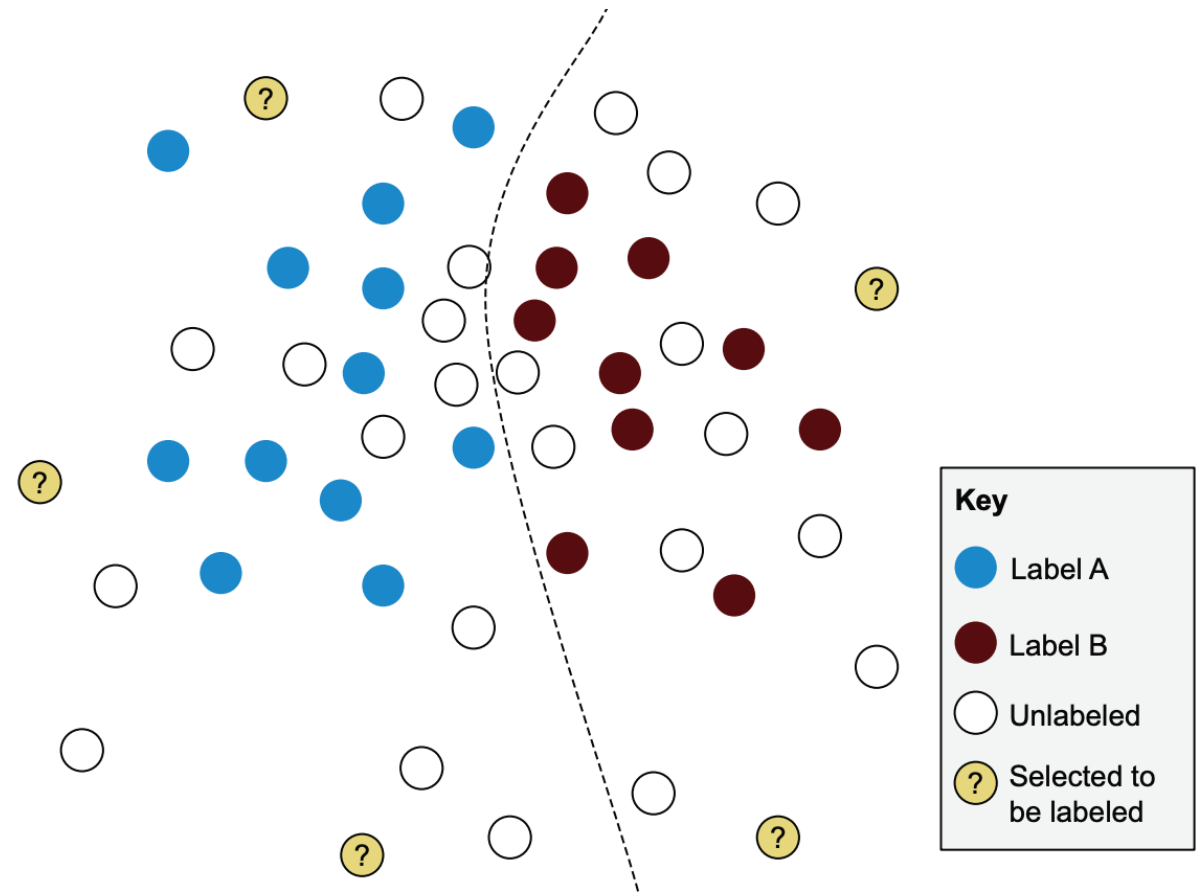
Sampling Instances to Label

- Simple random sample
- Uncertainty sampling



Sampling Instances to Label

- Simple random sample
- Uncertainty sampling
- Diversity sampling

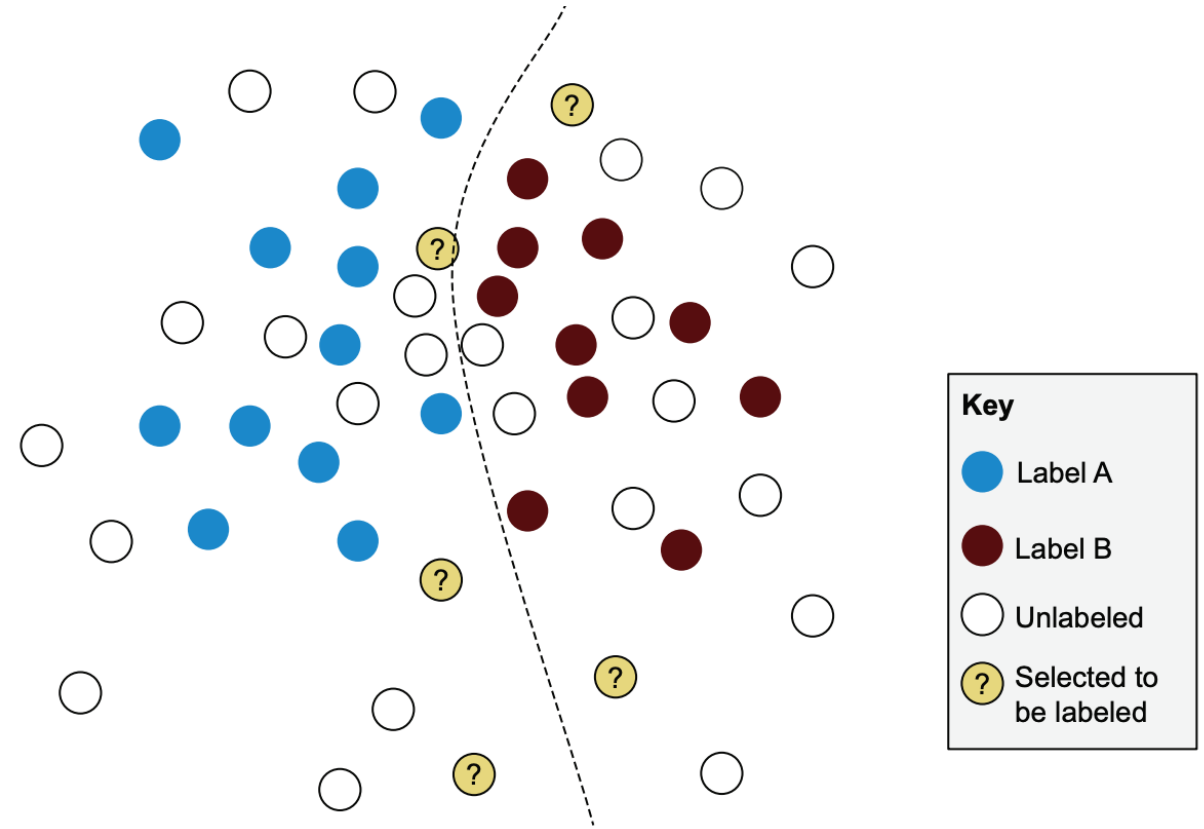


Sampling Instances to Label

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

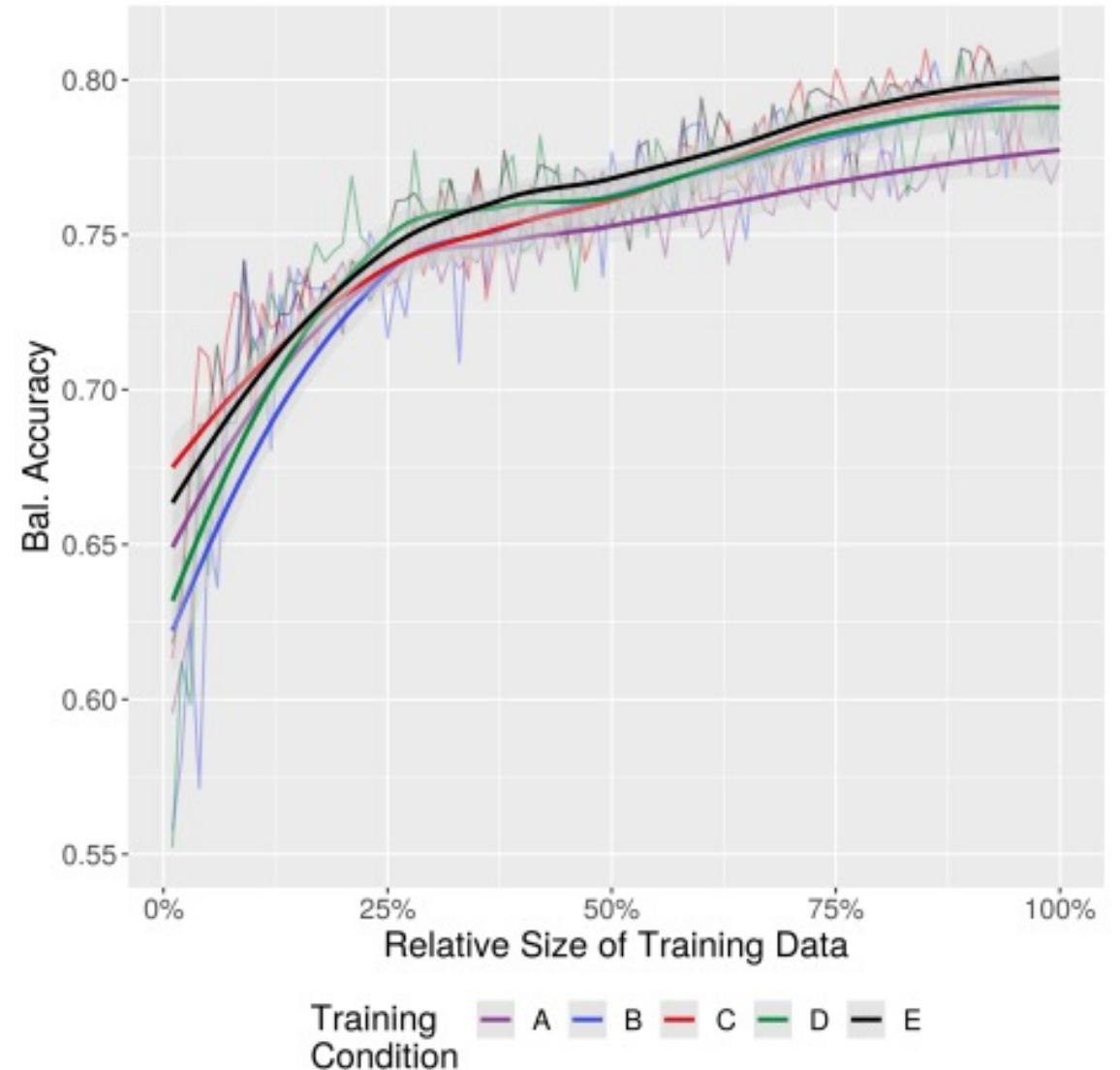
Sampling Instances to Label

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



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: 8th grade reading level
- Simplify instrument: Follow survey design practices
- Test instructions and instrument before collecting data