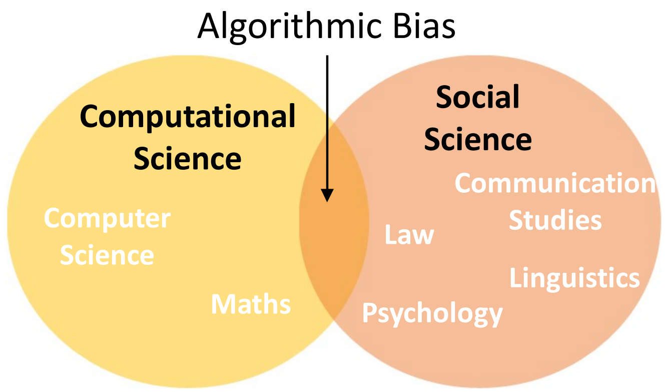
Measuring social biases in human annotators using counterfactual queries in Crowdsourcing

BHAVYA GHAI
PhD Candidate, Computer Science Department
Stony Brook University
Adviser: Prof. Klaus Mueller

Algorithmic Bias

When Algorithms exhibit preference for or prejudice against certain sections of society based on their identity. Such discriminatory behavior is termed as Algorithmic bias

- Generally emanates from biased training data
- Minorities & underrepresented groups are worst hit.
- Which sub-domains of AI are affected? ALL



Algorithmic Bias

Computational Science

Computer Science

Maths

Social Science

Communication Studies

Law

Linguistics

Psychology

Speech

Search Engine

NLP

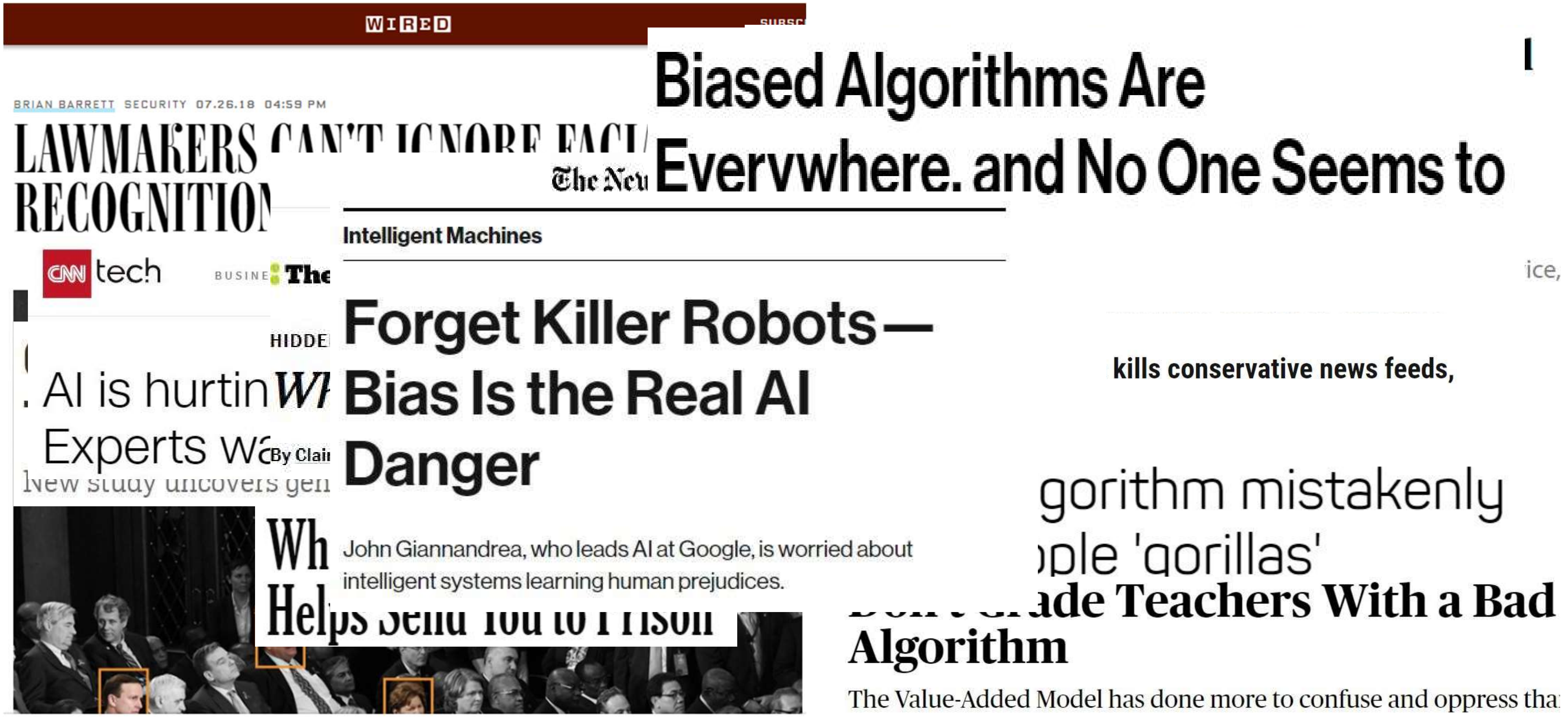
Computer Vision

Recommender Systems

Algorithmic Bias is the imminent AI danger impacting millions daily

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. "Unequal representation and gender stereotypes in image search results for occupations." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

In the media …



WIRED

SUBSC

BRIAN BARRETT SECURITY 07.26.18 04:59 PM

LAWMAKERS CAN'T IGNORE FACI

RECOGNITIO

The New

Biased Algorithms Are Everywhere, and No One Seems to

ice,

Intelligent Machines

Forget Killer Robots—Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.

CNN tech

BUSINE The

HIDDE

AI is hurtin

Experts wa

New study uncovers gen

Wh

By Clai

Wh

Helps Send You to Prison

kills conservative news feeds,

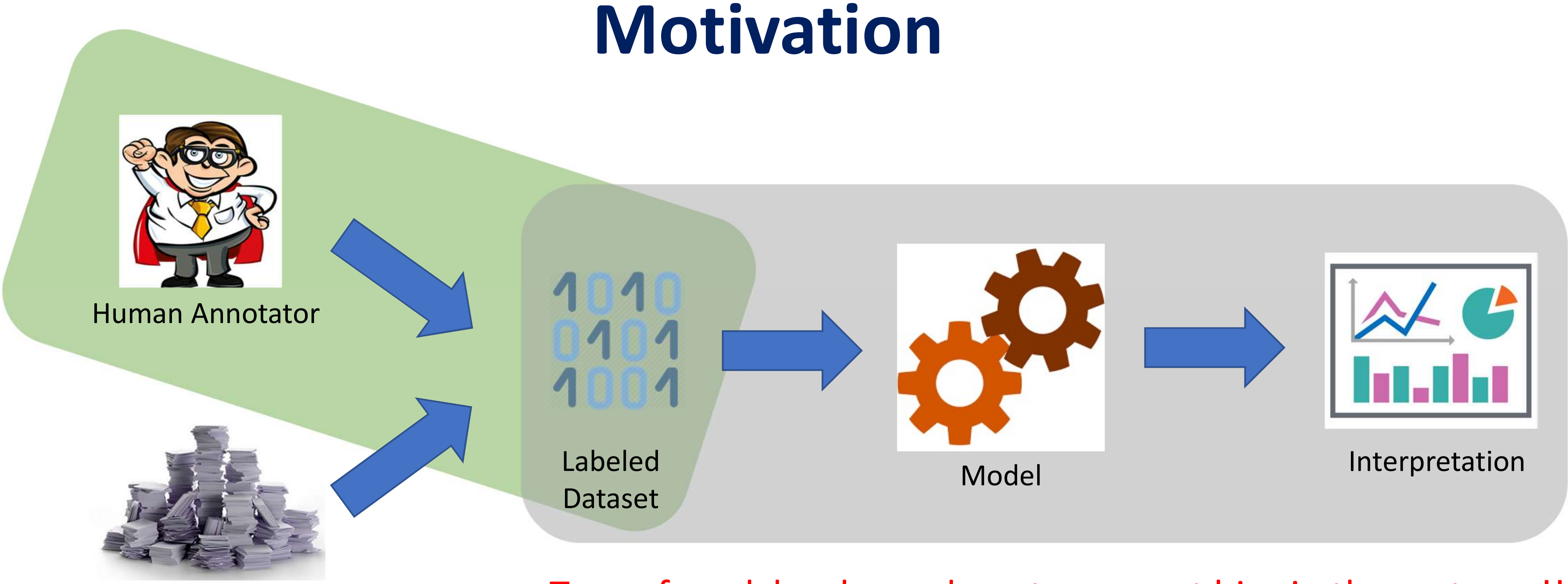
gorithm mistakenly

ople 'gorillas'

Don't Grade Teachers With a Bad Algorithm

The Value-Added Model has done more to confuse and oppress tha

Motivation



Human Annotator

Unlabeled Data

Labeled Dataset

Model

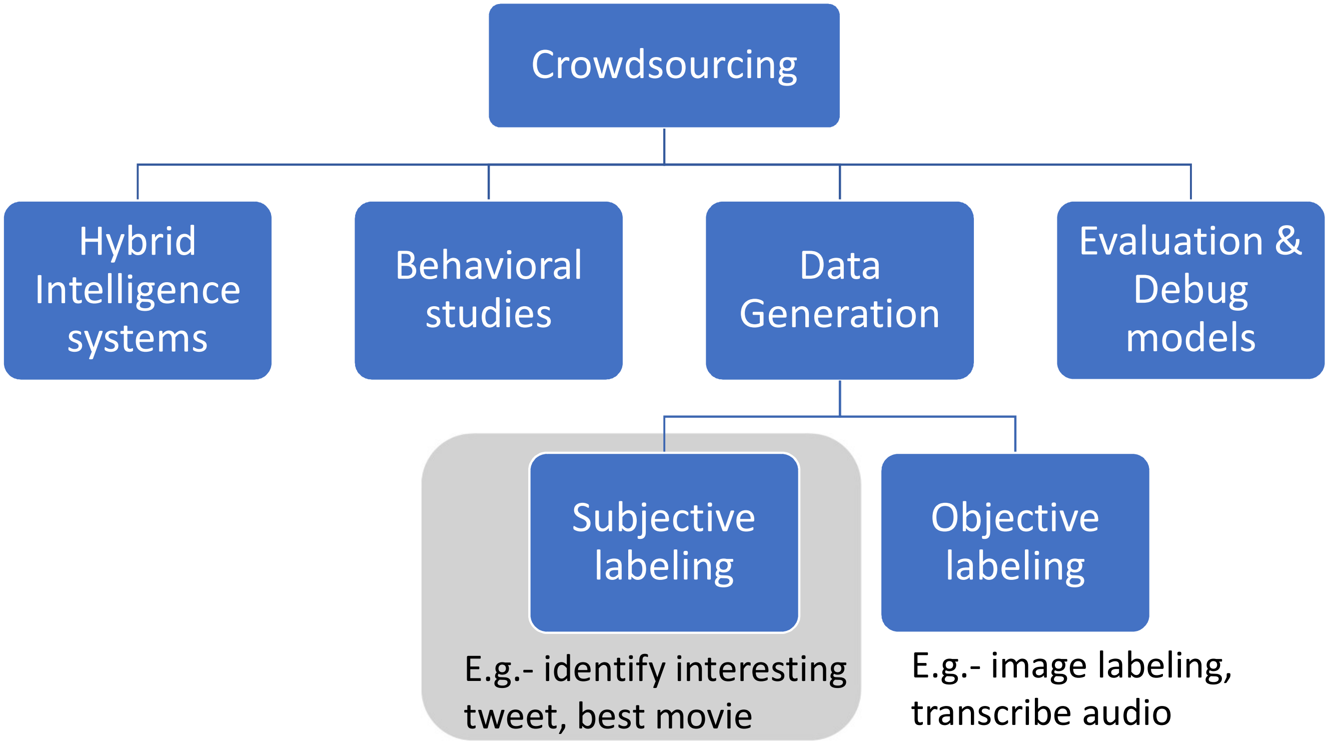
Interpretation

Tons of work has been done to prevent bias in these stages!!

Tackling Algorithmic bias in the crowdsourcing stage hasn't been explored

Holstein, Kenneth, et al. "Improving fairness in machine learning systems: What do industry practitioners need?." *arXiv preprint arXiv:1812.05239* (2018).

Crowdsourcing for Machine Learning



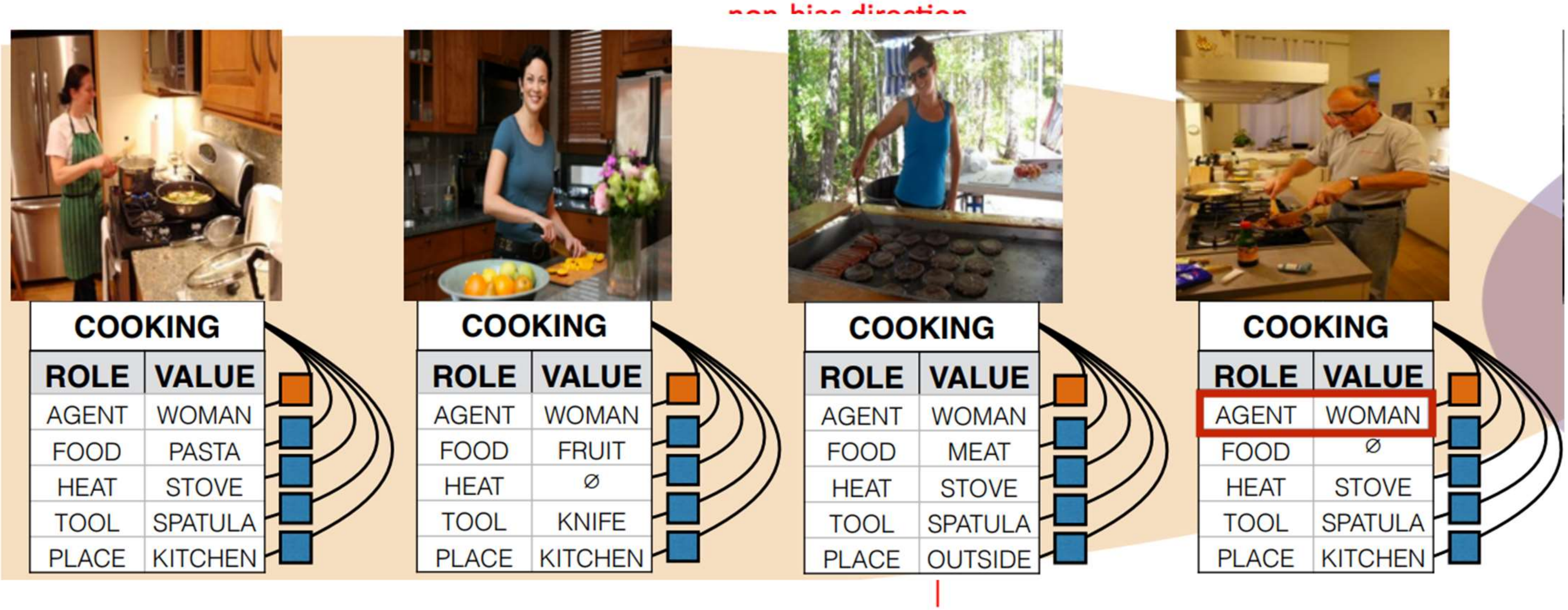
E.g.- identify interesting tweet, best movie

E.g.- image labeling, transcribe audio

We focus on Subjective labeling tasks because implicit bias may play a key role

Vaughan, Jennifer Wortman. "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research." Journal of Machine Learning Research 18 (2017): 193-1.

When Crowdsourcing got biased datasets



COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Crowdsourcing is not immune to social biases & may lead to Algorithmic bias

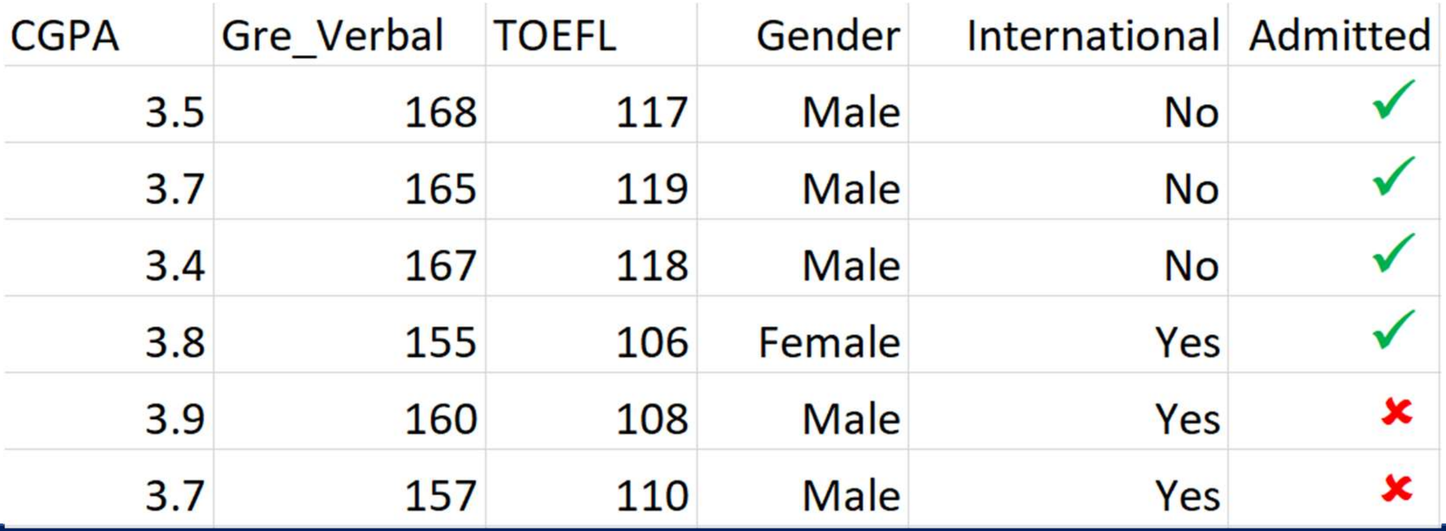
Zhao, Jieyu, et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." arXiv preprint arXiv:1707.09457 (2017).

Sources of Bias

Label Bias: If the distribution of positive outcomes is skewed with respect to a demographic group

Selection bias: Samples chosen for labeling don’t represent the underlying population.

For e.g. Consider a graduate admissions scenario.



CGPA	Gre_Verbal	TOEFL	Gender	International	Admitted
3.5	168	117	Male	No	✓
3.7	165	119	Male	No	✓
3.4	167	118	Male	No	✓
3.8	155	106	Female	Yes	✓
3.9	160	108	Male	Yes	✗
3.7	157	110	Male	Yes	✗

In this study, we are just focused on Label bias

Types of Labelers

Naive

Expert

Spammer

Adversarial

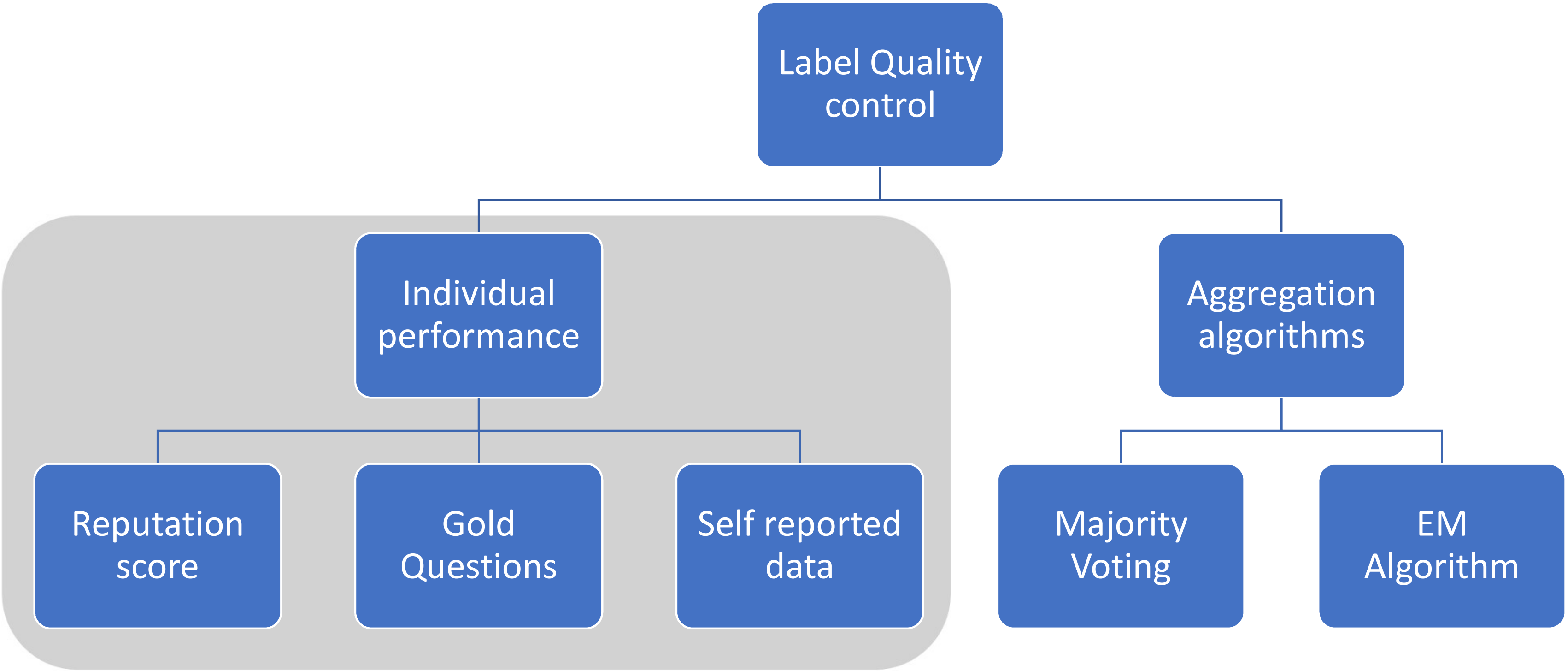
Biased

Biased – A human annotator infested with serious social biases based on gender, race, etc. which are reflected in his/her labels. Their labels might reflect strong preference for or prejudice against a demographic group.

In this study, we are trying to identify & control for biased labelers

Stony Brook University

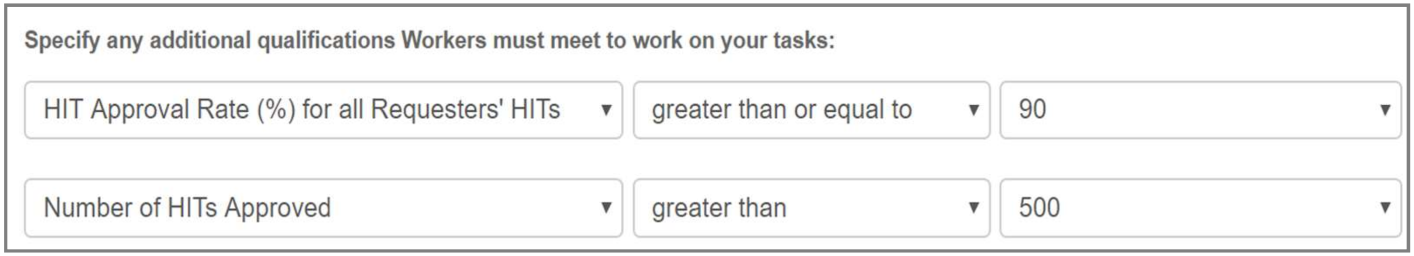
Existing Literature



Our objective is to devise a new technique for measuring Individual Performance

Reputation Score

Based on worker’s past performance. Eg.- percentage of previously approved HITs



Specify any additional qualifications Workers must meet to work on your tasks:

HIT Approval Rate (%) for all Requesters' HITs ▾	greater than or equal to ▾	90 ▾
Number of HITs Approved ▾	greater than ▾	500 ▾

Snippet from Amazon MTurk

. **Drawbacks**

- Requesters are approving HITs more than they should, thereby inflating workers’ reputation levels[1]
- It is possible, that a biased user might achieve high reputation score by performing several objective tasks, so qualifies for a subjective task where his/her response(s) might be biased

Does reputation score capture implicit social bias of annotators? Maybe Not

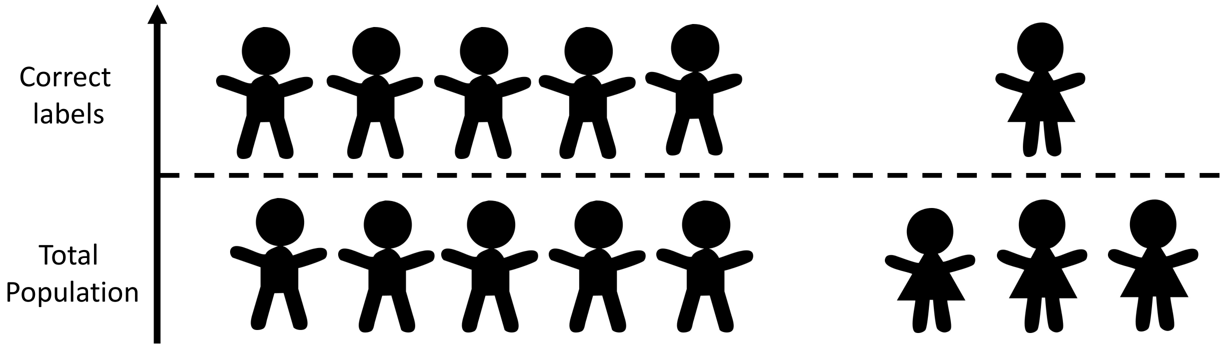
[1] Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti. "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk." Behavior research methods 46.4 (2014): 1023-1031.

Gold Questions

- Gold questions are the tasks for which ground truth is available. It’s one of the most common ways to evaluate noisy labelers like spammers, etc..

- If a worker correctly answers more than a threshold of gold questions, he/she is considered eligible for the study.

- Knowing how often someone is right is important. But in the context of social biases, it’s equally important to know when someone fails



Overall Accuracy: 75%

Male Accuracy: 100%

Female Accuracy: 33%

High accuracy on Gold Questions doesn’t always mean low bias

Self Reported data

Survey Questionnaire

1. No matter how accomplished he is, a man is not complete as a person unless he has the love of a woman

2. Most women interpret innocent remarks or acts as being sexist

3. Most women fail to appreciate what all men do for them.

4. When women lose to men in a fair competition, they typically complain about being discriminated against.

5. Women, as compared to men, tend to have a more refined sense of culture and good taste

- One of the only measures designed to capture implicit social biases.

- The content of survey questions is quite different from the study. Hence, they make crowd workers conscious that they are being judged

- Suffer from Social desirability & Social approval bias

- Not very engaging.

- Inaccurate

It can serve as a good baseline for upcoming techniques to measure social bias

Glick, Peter, and Susan T. Fiske. "The ambivalent sexism inventory: Differentiating hostile and benevolent sexism." Social Cognition. Routledge, 2018. 116-160.

Our approach - Counterfactual Queries

Counterfactual tries to estimate the outcome in a hypothetical world where a different treatment was given.

In ML literature, an ML model is considered counterfactually fair if

$$P(Y|\ X, A=1) = P(Y|\ X, A=0)$$

where A is the sensitive attribute like gender, race, etc.

We are trying to adopt this technique to identify biased workers in Crowdsourcing. Counterfactual query is created by flipping the sensitive attribute of the original query

Hypothesis: Unbiased worker will give consistent labels for counterfactuals

Kusner, Matt J., et al. "Counterfactual fairness." Advances in Neural Information Processing Systems. 2017.

Use case- Toxic Comment classification

Rate the following statements on toxicity (1 to 10 scale) where 1 is non-toxic and 10 is highly toxic

Q: Homosexuality is a disease that must be cured

CQ: Heterosexuality is a disease that must be cured

$$\textit{Worker Bias score} = \textit{mean}(|\ \text{Label(Q)- Label(CQ)}\ |)$$

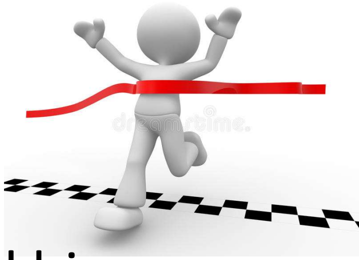
If Bias score > λ (threshold) => Worker is biased

Doesn’t need Ground truth & blends with the task perfectly!

Garg, Sahaj, et al. "Counterfactual fairness in text classification through robustness." Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 2019.

Stony Brook University

Conclusion & Future Work



- Datasets curated via crowdsourcing maybe polluted by social biases of crowd workers and may eventually lead to Algorithmic bias.
- Need for new label quality control techniques which incorporate fairness metrics apart from accuracy.
- Counterfactual queries can be one way to capture social biases without having any ground truth.
- Next, we intend to conduct a user study to test existing techniques and compare with our approach.

Thanks for your attention!

Bhavya Ghai

For any Questions, suggestions, feedback, criticism, please email me at:-

bghai@cs.stonybrook.edu

Stony Brook University