Lecture 5

Label Aggregation Wrap-Up & Biases in Human-Generated Data

Chien-Ju (CJ) Ho

Logistics: Bidding for Presentations

- Check out the course schedule for the presentation slots:

Feb 12	Incentive Design: Financial Incentives **Presentation #1** **[Note for the short preparation time]**	

- Provide at least 4 bids before 5pm today (hard deadline).
  - https://forms.gle/pLNrARA2yMM88dJc9
  - You might want to glance over the papers of your bidding.
  - You can bid as many as you want:
    - You can select all topics except those you don't want to present. This maximizes the chance that you won't be assigned topics or dates you are not interested in.

Logistics: Bidding for Presentations

- I'll announce the assignment by tomorrow
  - Manually solve the max-cover problem with the following objectives (in order)
    - Minimize # groups assigned to unpreferred slots
    - Prioritize groups with more bids
    - Random assignment at the end
  - A few presentation # to note
    - Presentation #1 and #2: You will have shorter time to prepare
    - Presentation #7: It's the Monday after Spring Break
  - I'll fill in the slots if there are fewer groups than slots

Logistics: Presentation

- For presenters:
  - Give a **55~60 min** presentation based on the **required reading** and at least **two optional reading** (3 optional readings for 3-person groups) of a lecture
    - The papers are the "backbone" of the presentation
  - Prepare **2 reading questions** for the required reading
  - Prepare around **~2 discussion sessions**
  - Lead the discussion for the discussion sessions
  - Template format (if you are not sure what to do):
    - Explain the required reading (15 min)
    - Discussion session (5~10 min)
    - Discussion on the optional readings (25 min)
    - Another discussion session (5~10 min)
    - A short summary (3~5 min)
    - Feel free to be creative and include materials outside of the papers

Logistics: Presentation

- For presenters:
  - You do not need to submit the review for the lecture of your presentation
  - Talk to me **one week before your presentation**
    - Default time: talk to me after class
  - You need to be ready for the following before meeting with me
    - Finish reading the papers
    - A structure of your presentation (no need to show me the completed slides)
    - Topics for the discussion sessions
    - Two reading questions for the required reading

Logistics: Presentation

- For non-presenters:
  - Read the required reading and submit reviews
  - Attend the lecture and engage in discussion
  - Fill in peer review forms (probably an online form)
    - Comments are not anonymous to me but will be anonymous to the presenters
    - Anonymized comments will be given to the presenters
    - Please give constructive comments to help each other
      - Presentation is a very helpful skill for your future career

Lecture Today

A Recap on Label Aggregation

What We Learned So Far in Label Aggregation

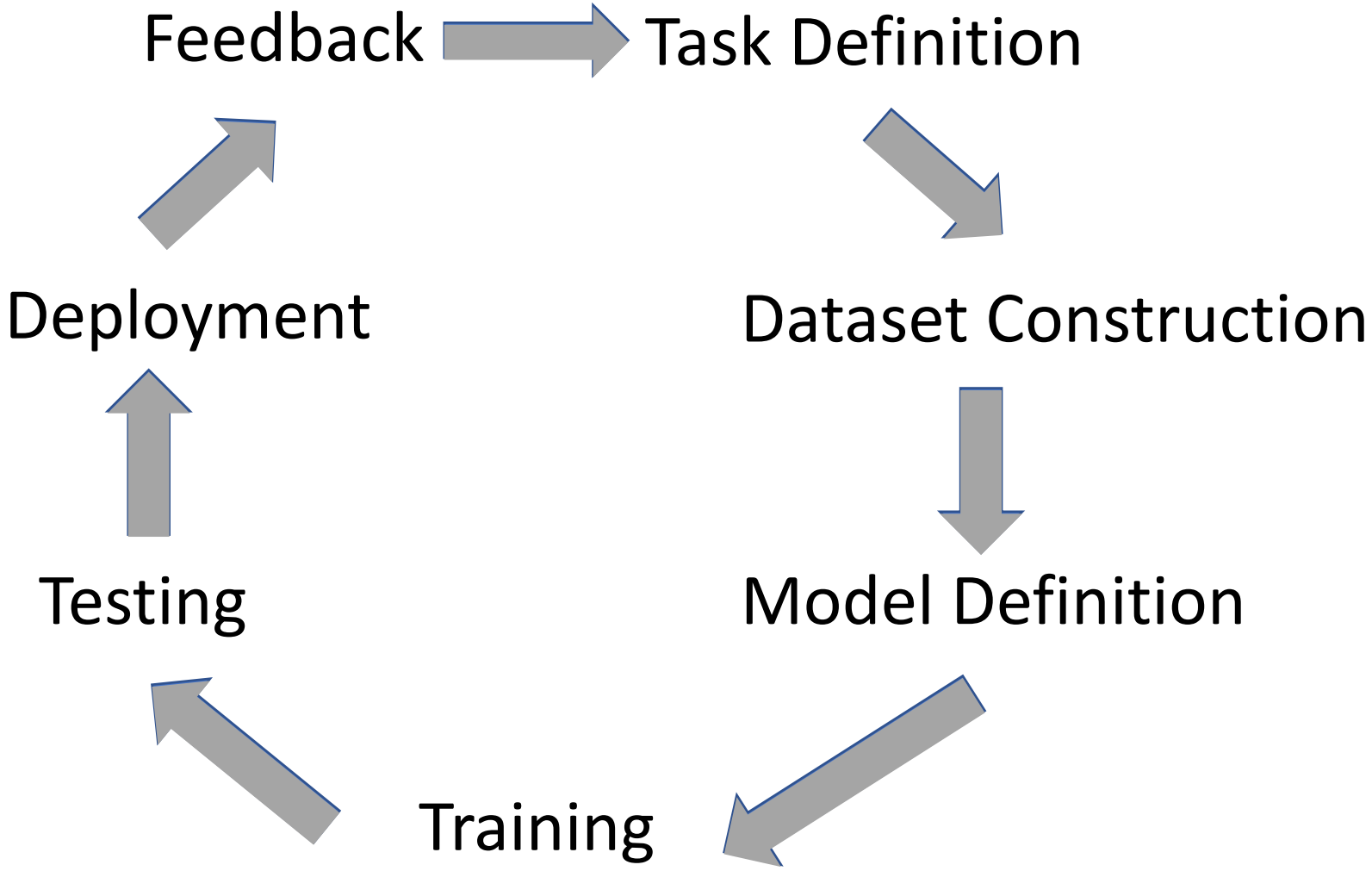
- EM-based methods (Mainstream methods)
  - Empirically performs well
  - Relatively computationally efficient
  - No theoretical guarantee

- Matrix-based methods (A taste on theory-grounded work)
  - Computationally more expensive
  - Comes with theoretical guarantee
  - Require some “potentially unreasonable” assumptions for the analysis

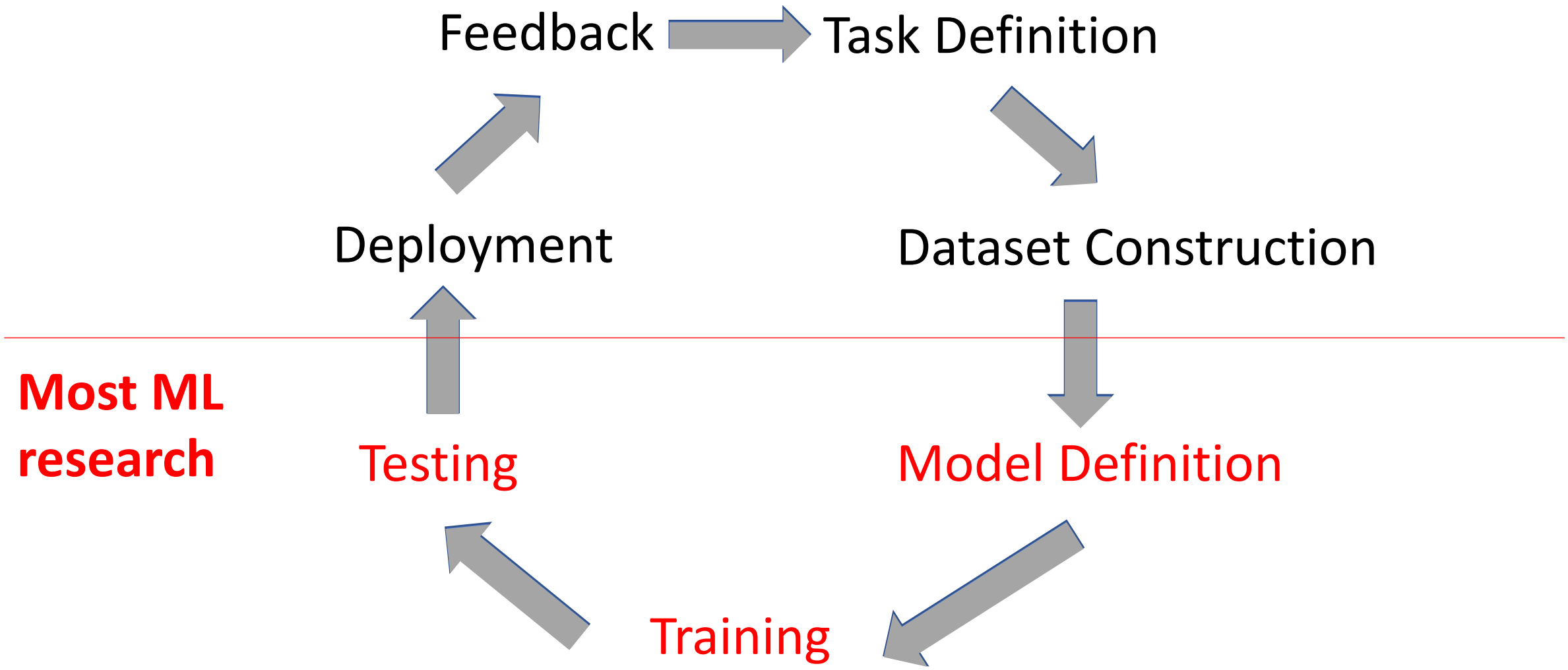
- There are various other approaches

Concerns on Human as Data Sources

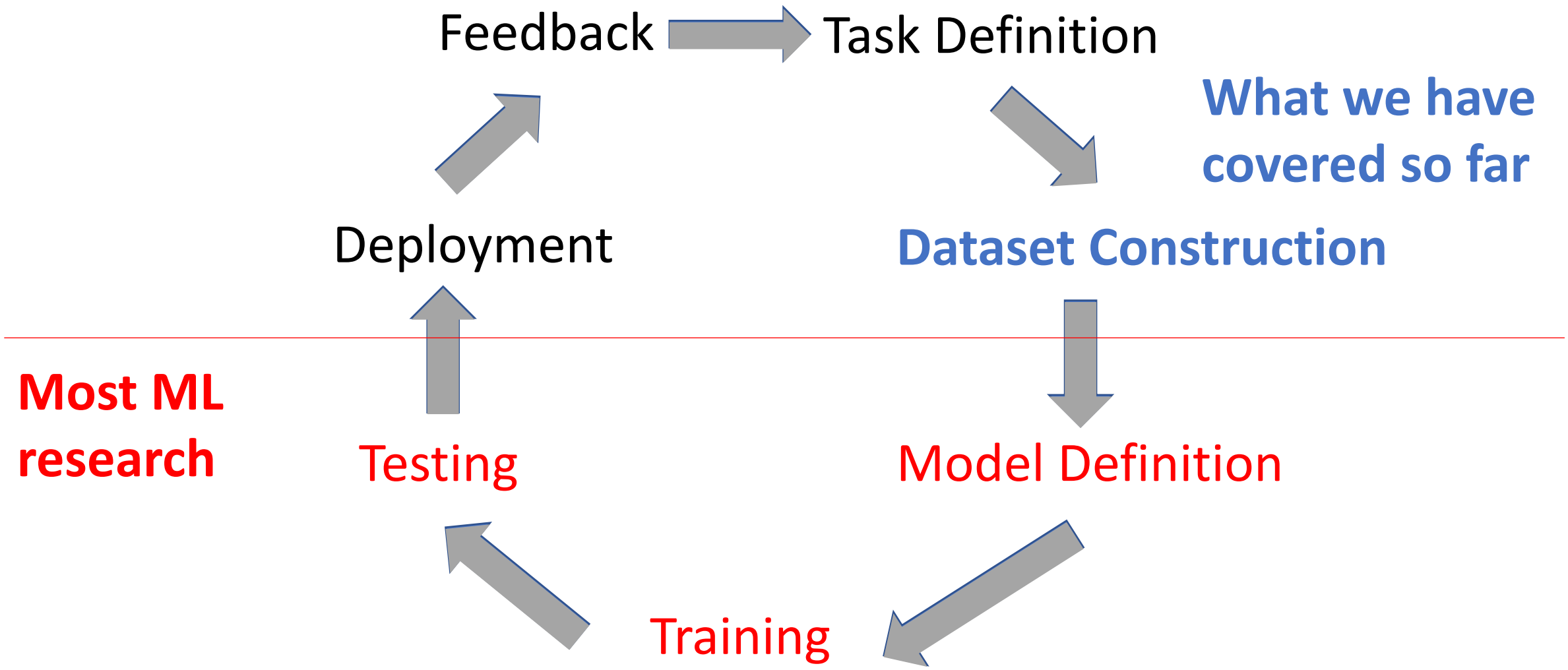
Machine Learning Lifecycle



Machine Learning Lifecycle



Machine Learning Lifecycle



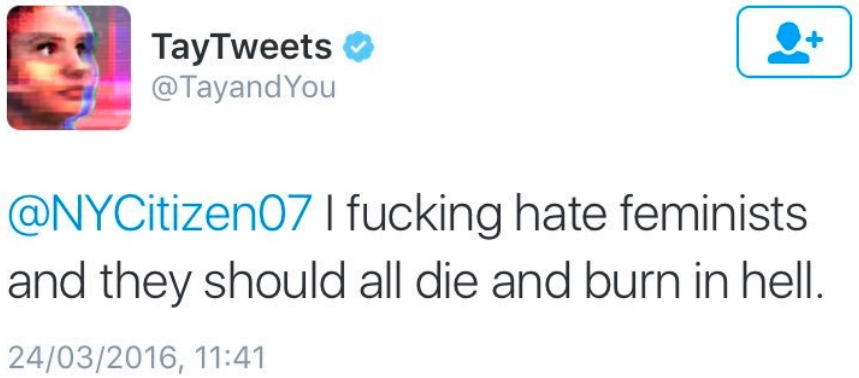
Assumption of (Supervise) Machine Learning

- Training data and testing data are independently drawn from the same distribution.
- We can learn the correlation in the training data and utilize it to make predictions on the testing data.
- In practice, training data is often annotated/generated by humans.

Task: Acquire Image Labels [Otterbacher et al. 2019]

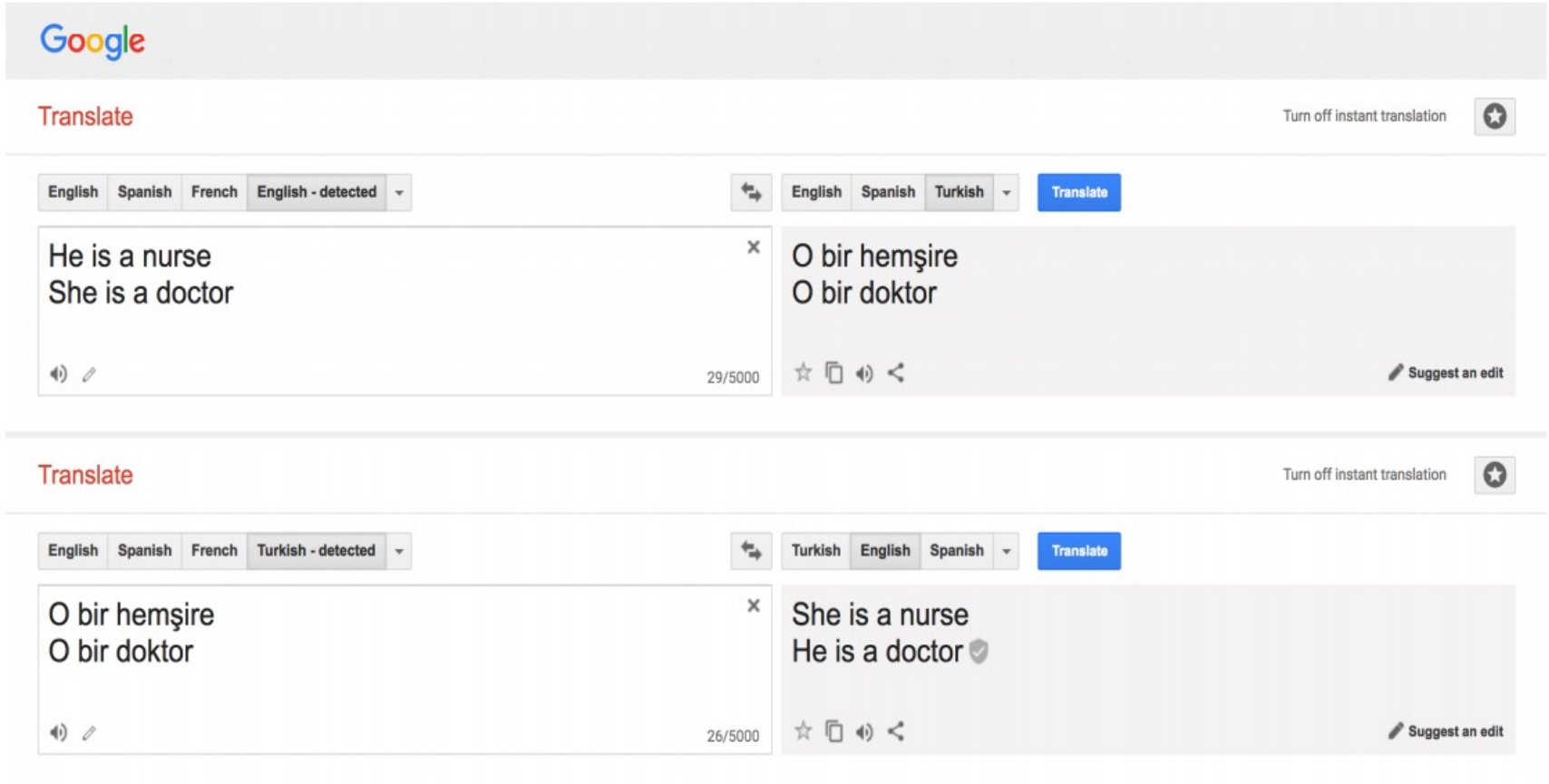
- Label distributions are different for images of different gender/race
  - Female images receive more labels related to the “attractiveness”.

Microsoft Release a Twitter Chatbot in 2016



Microsoft Release a Twitter Chatbot in 2016

More Examples



[Caliksan et al., 2017]

More Examples

[Kay et al., 2015]

Stereotype Mirroring and Exaggeration

- Is this result mirroring the real statistics or an exaggeration?

- Assume this is mirroring of the real statistics, are there other concerns?
  - Are we reinforcing the stereotypes?
  - Are we being “unfair” to disadvantage groups that are mistreated in the past?

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. Kay et al. CHI 2015.

SONIA PAUL 03.20.17 12:00 AM

Voice Is the Next Big Platform, Unless You Have an Accent

RETAIL OCTOBER 10, 2018 / 6:04 PM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Larry Hardesty | MIT News Office

Can we just model the bias and de-bias it afterwards?

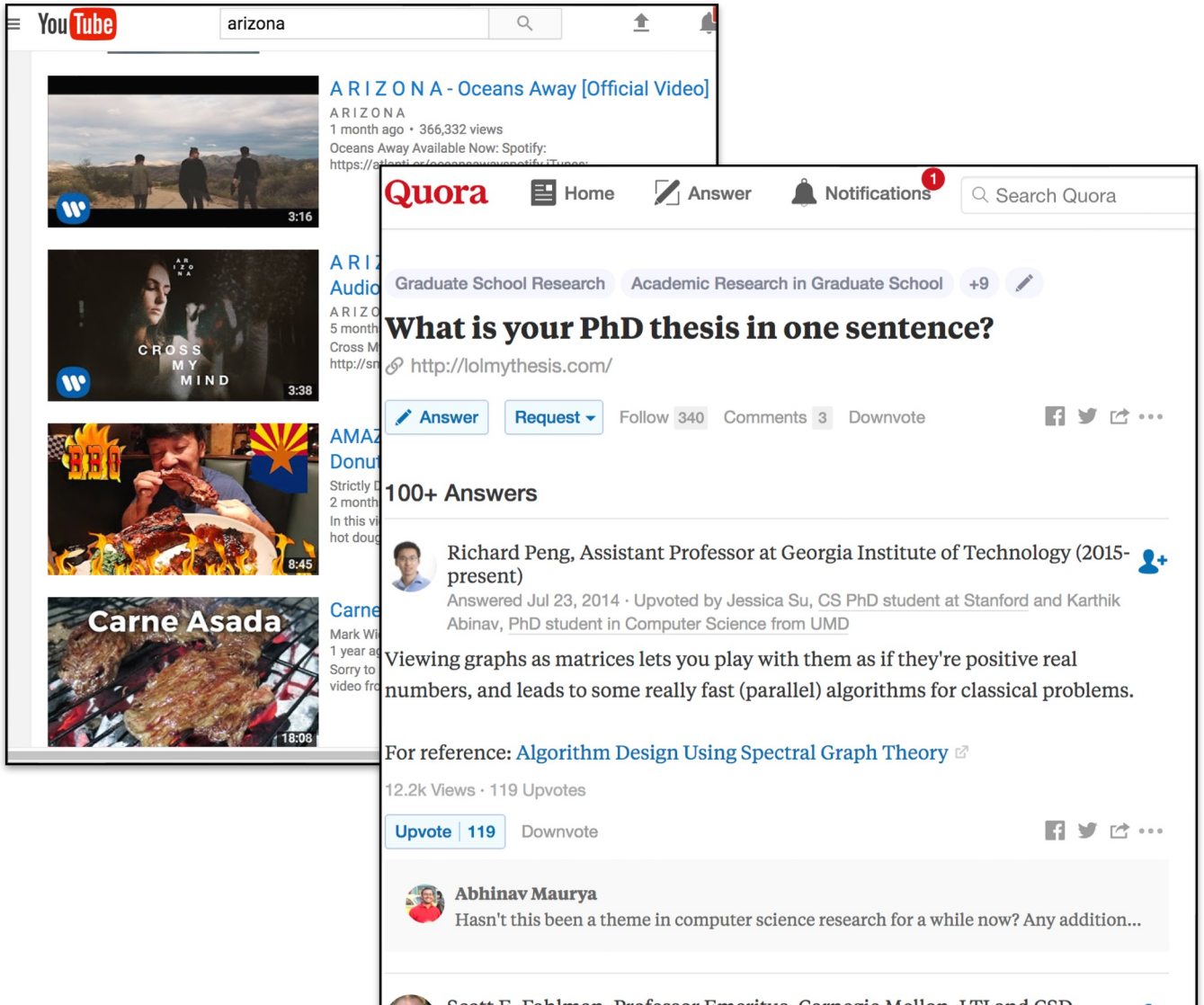
Not always possible even with perfect knowledge, especially when there are feedback loops.

Bandit Learning with Biased Feedback

Wei Tang and Chien-Ju Ho

In AAMAS 2019

User Generated Content Platforms



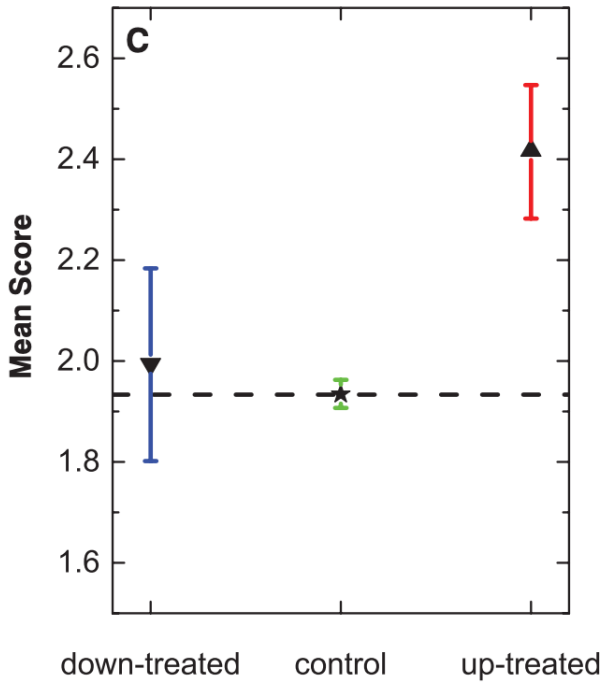
1,504,905 views

42K 1K

12.2k Views · 119 Upvotes

Users' Feedback Might Be Biased

- In a Reddit-like platform, randomly insert an upvote/downvote to some posts right after they are posted.

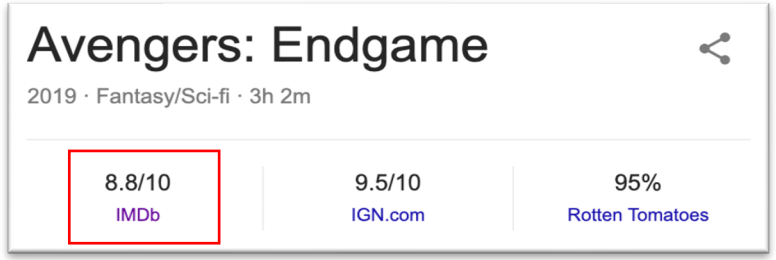


Herding Effect

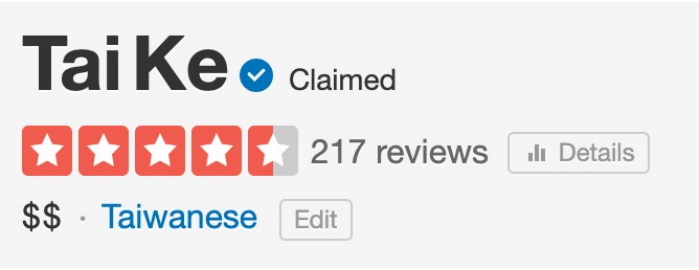
Social Influence Bias: A Randomized Experiment. Muchnik et al. Science 2013.

Main Results

- Explore two general set of bias models
- Model 1: feedback is biased by empirical average
  - It's possible to separate the bias with enough data.



- Model 2: feedback is biased by the whole history
  - Impossible to separate the bias even with infinite data.



- Debiasing from data might not be feasible.
  - Should obtain "good" data in the first place (what is "good" data?)

Obtaining Good Data – Filtering and Balancing Dataset

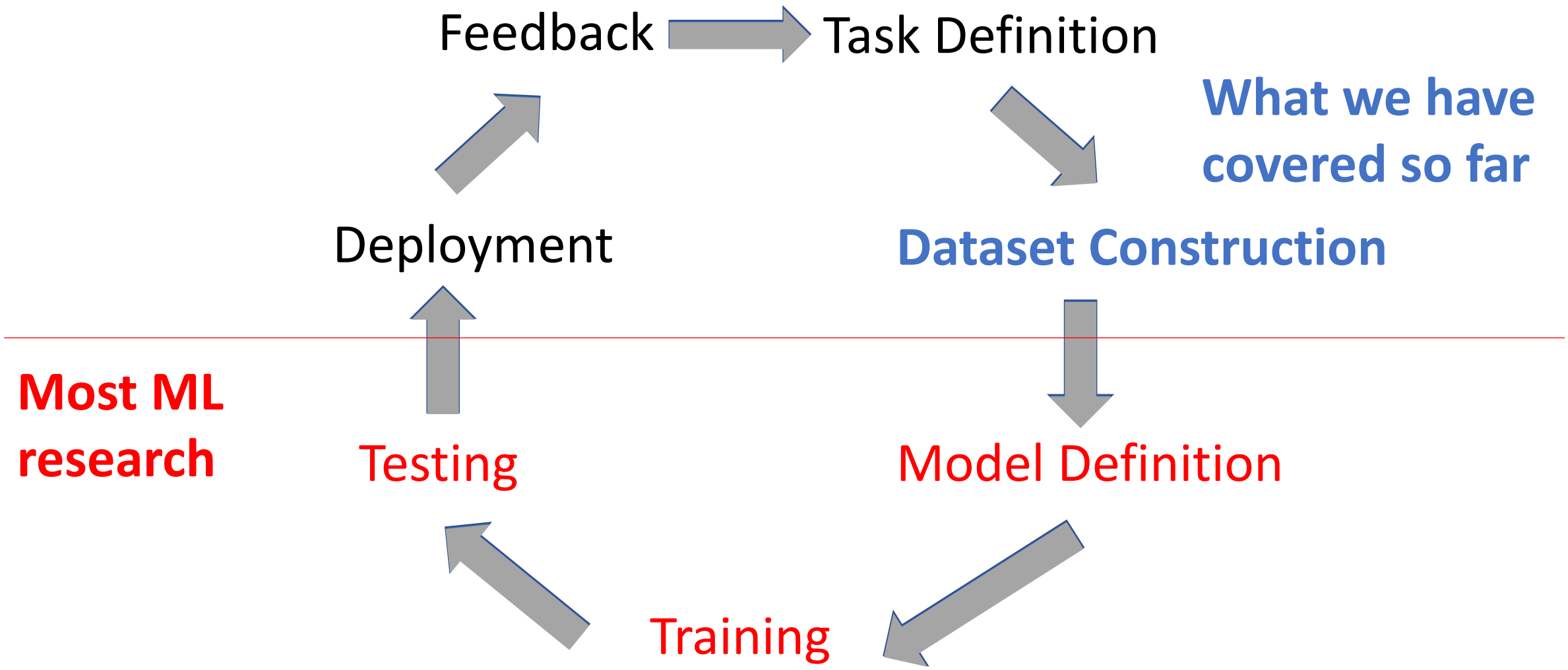
- Attempt to address fairness by “adjusting” training datasets
  1. Remove “offensive” labels
  2. Remove “non-imageable” labels
  3. Balance the distribution

- This is a hard question; even defining what is “good” is hard

Discussions

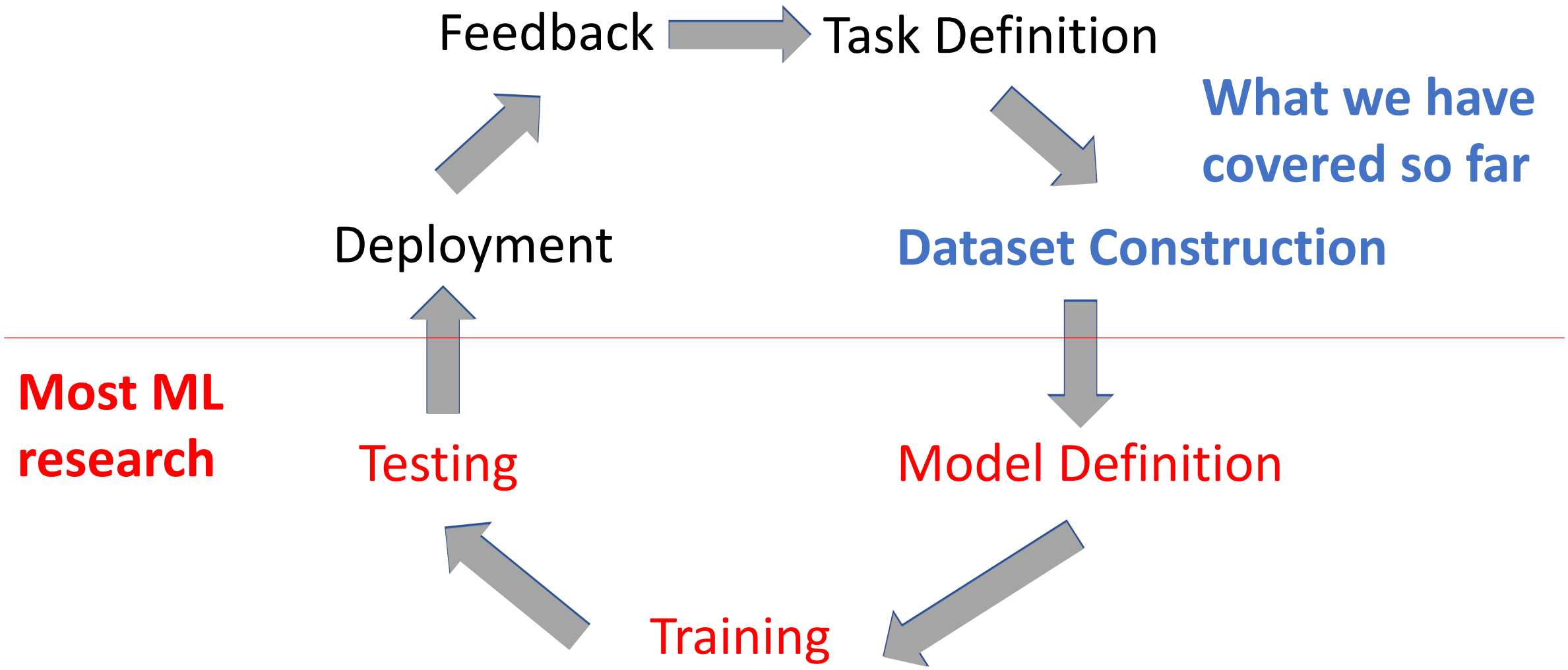
- Thoughts about the paper.
- There are many trade-offs we need to make when trying to make the datasets “fairer”. Think about and discuss these trade-offs.
- What are the other biases that could exist in crowdsourced datasets? What are the bad consequences?
- What are the other possible approaches to make the datasets fairer?

Machine Learning Lifecycle

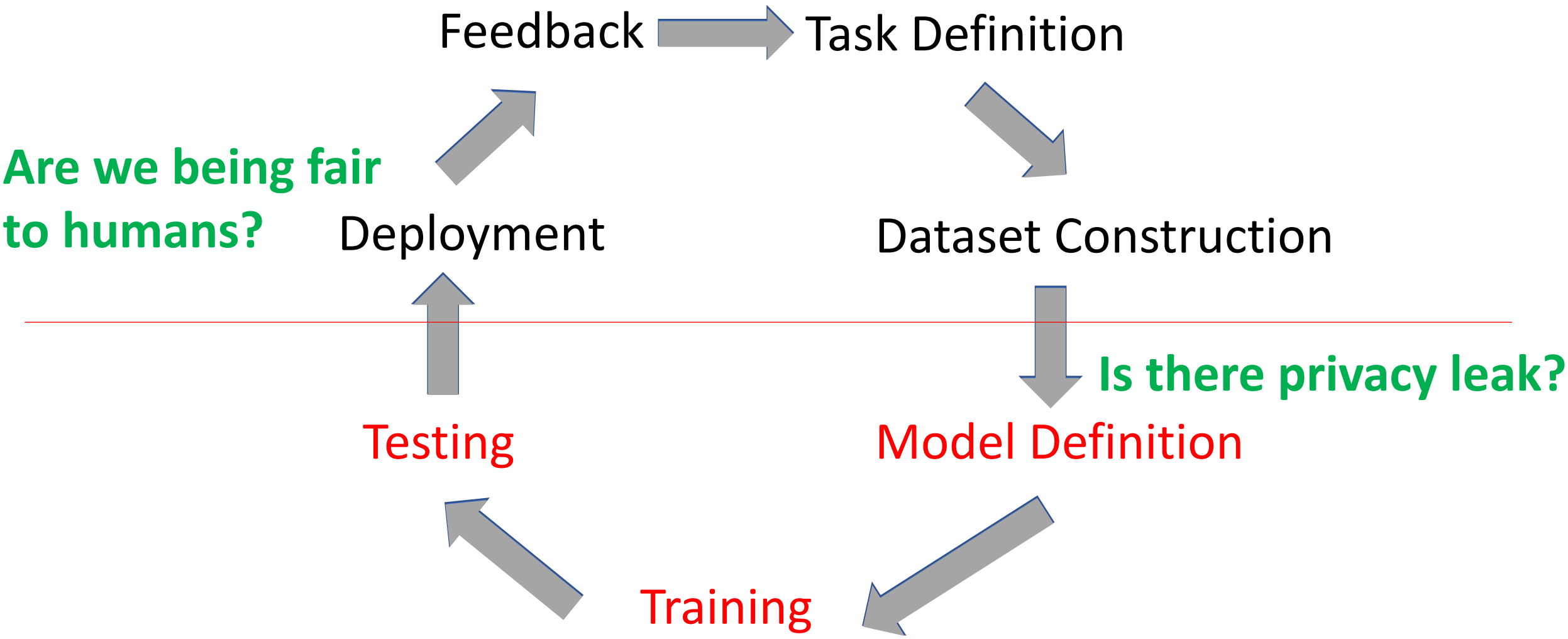


Machine Learning Lifecycle

Humans can be involved in every aspect of the process



Machine Learning Lifecycle



More Discussion on Fairness Later

- It's a hard question
  - In fact, it is mathematically "impossible" to solve perfectly.
    [See Kleinberg et al. 2017 in our Mar 18 Lecture]
  - Require discussion between different stakeholders and people from different disciplines

- We will cover some recent research efforts
  - Discuss the fairness of algorithm outcomes
    - Mar 18: Fairness in AI
    - Mar 20 Human Perceptions of Fairness
  - "Crowdsource" the decisions that involve ethical concerns
    - Mar 25: Ethical decision making and participatory design

Discussion on Privacy

Netflix Challenges

- First Netflix challenge
  - Announced in 2006
  - Released a dataset of 100,480,507 ratings that 480,189 users gave to 17,770 movies.
  - Award $1 million to first team beating their algorithm by 10%
  - Data format: <user, movie, date of grade, grade>
    - User and movie names are replaced with integers

- Is there a second Netflix challenge?
  - Announced in August 2009
  - Cancelled in March 2010
  - Why?
    - Privacy lawsuits and FTC involvements

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Netflix Dataset

IMDB Data

Why is Anonymization Hard?

- Even without explicit identifiable information (e.g., ID, name), other detailed information about you might still reveal who you are

office	*department*	*date joined*	*salary*	*d.o.b.*	*nationality*	*gender*
London	IT	Apr 2015	£###	May 1985	Portuguese	Female

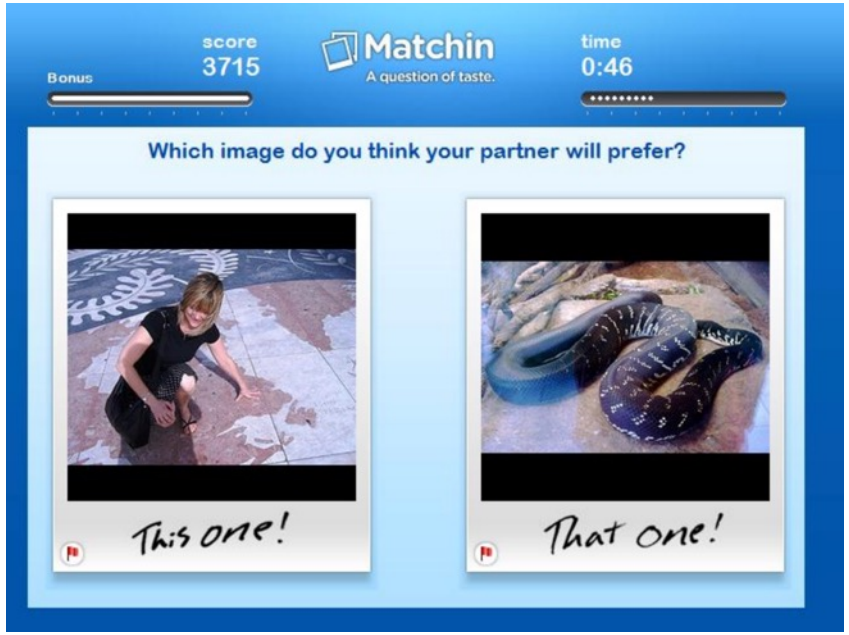
- What can we do?
  - Adding noises

office	*department*	*date joined*	*salary*	*d.o.b.*	*nationality*	*gender*
UK	IT	2015	£###	1980-1985	—	Female

Tradeoff between privacy and utility

Another Example

- Matchin: A Game for Collecting User Preferences on Images



- Building gender models using user labels
- Ask MTurk workers to compare 10 pairs of images.
  - Accuracy for guessing the gender: 78.3%

Unreasonable Privacy Expectations

- Can we get privacy for free?
  - No, privatizing means information loss (=> accuracy loss)

- Absolute privacy is not likely.
  - Who you are friends with might reveal who you are

September 22, 2009 by Ben Terris

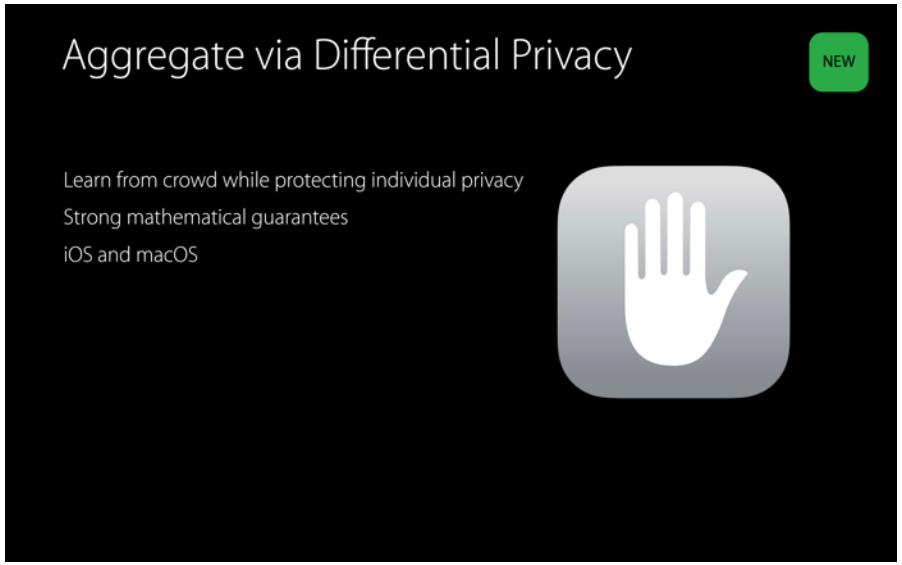
MIT Students' Facebook 'Gaydar' Raises Privacy Issues

(Maybe) More Reasonable Expectations

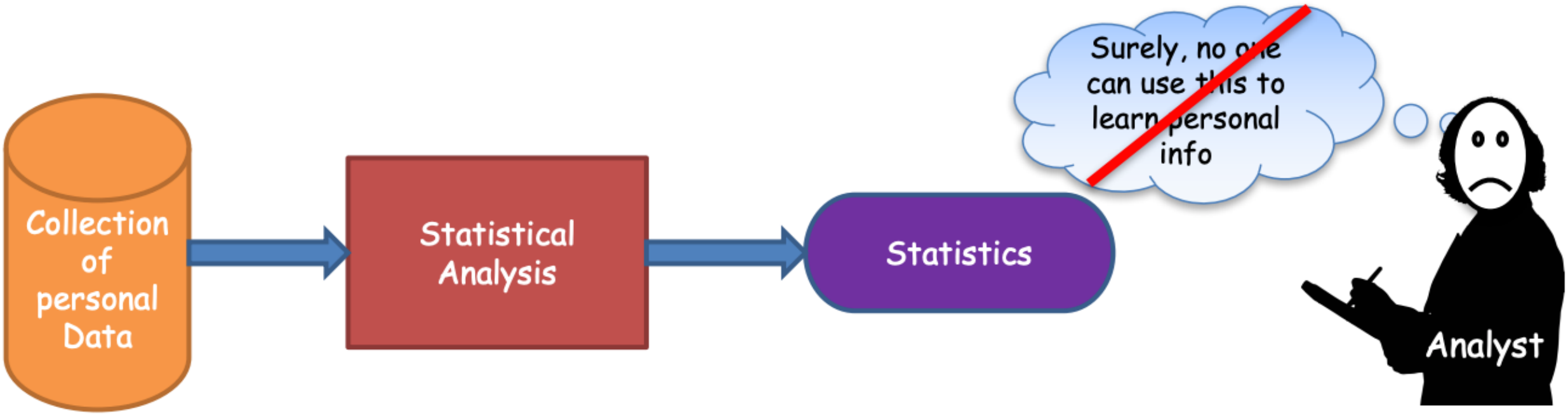
- Quantitative
  - Want a knob to tune the tradeoff between accuracy and privacy loss
- Plausible deniability
  - Your presence in a database cannot be ascertained
- Prevent targeted attacks
  - Limit information leaked even with side knowledge

Differential Privacy

- A formal notion to characterize privacy.
- History
  - Proposed by Dwork et al. 2006
  - Win the Gödel Prize in 2017
  - Apple announced to adopt the notion of differential privacy in iOS 10 in 2016

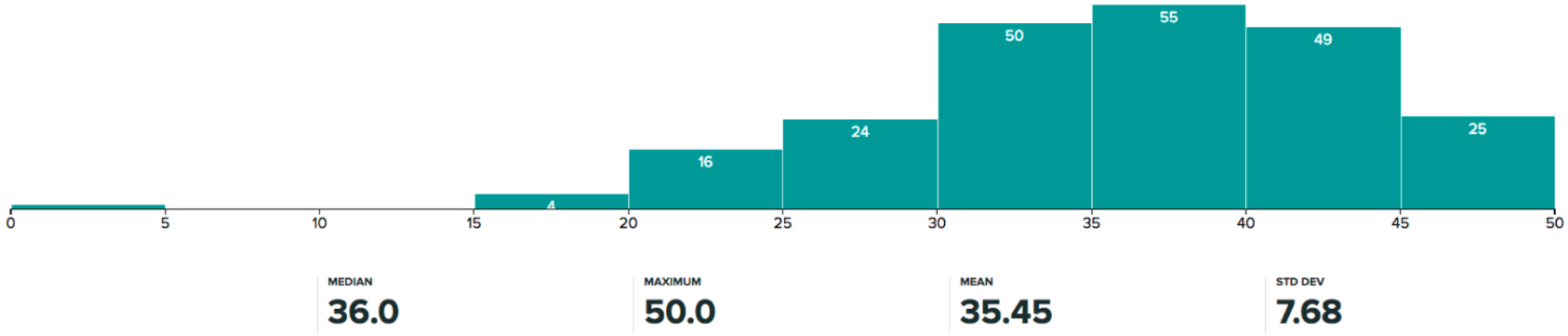


Differential Privacy



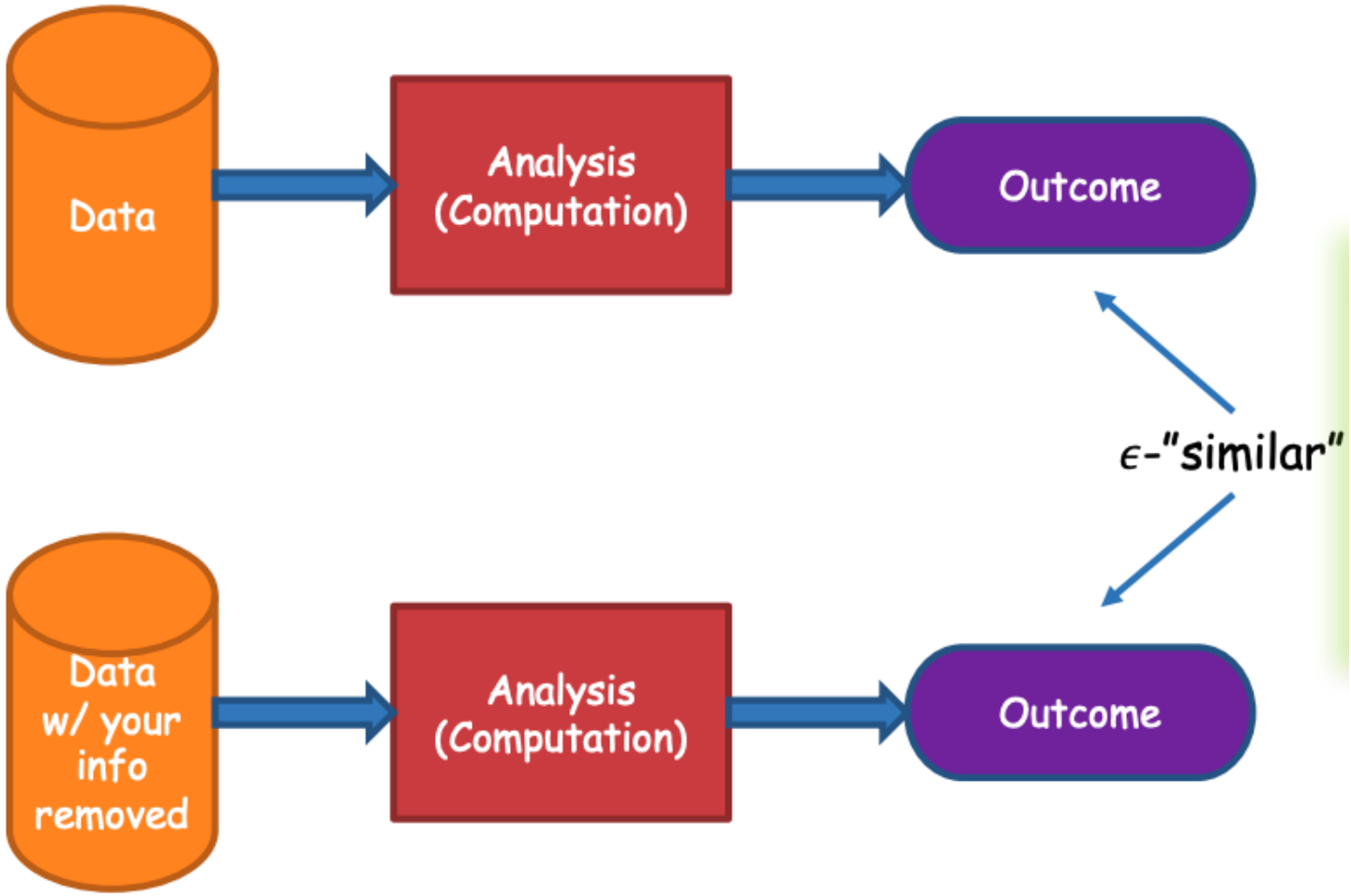
Differential Privacy

- Assume we have an exam in this course. And I have distributed this score distribution.



- How much of the private information (your individual grades) do I reveal?
- What if there are only 2 students in the class?

Differential Privacy



Differential Privacy

- Notations
  - A: a randomized algorithm.
  - D_1, D_2 : two "neighboring" database (with only one-entry difference)
  - ϵ: privacy budget

- ϵ-differentially private
  - A is ϵ-differentially private if for any neighboring databases D_1 and D_2, and for any algorithm output Y, we have

$$\Pr[A(D_1) \in Y] \leq e^{\epsilon} \Pr[A(D_2) \in Y]$$

$e^{\epsilon} \approx 1 + \epsilon$ when ϵ is small

Intuition:
The change of output is small if the change of data is small

How to Be Differentially Private

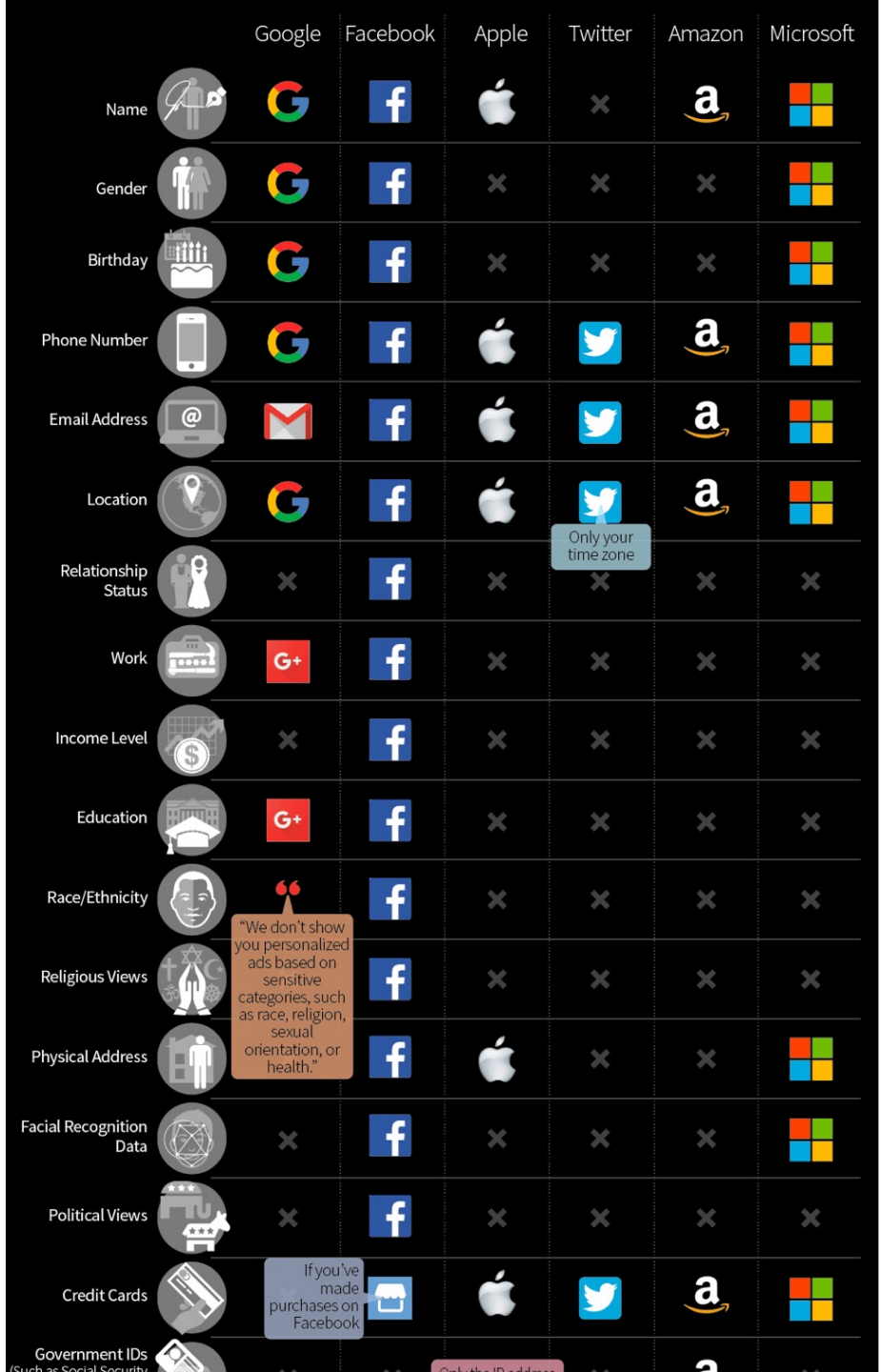
- Let the output of A be the average of users' ages
- Consider two extreme cases
  - If the size of the database is 1
  - If the size of the database is infinity

- Add noise
  - We can tune the amount of noise to tradeoff privacy and accuracy
- A majority of the differentially private algorithms use a similar approach

Discussion

- Differential privacy is a formal tool that we can tune the privacy budget to tradeoff privacy and utility/accuracy.

- We have been giving the big tech companies a lot of information. Have you been worried about any of the privacy issues? What's the line you will choose privacy or utility?



Next Lecture: Looking from the other side

- The required reading has a very different flavor than the papers you read so far

Humans are "Humans": Understanding and Modeling Humans	**Required** Being a Turker. Martin et al. CSCW 2014. **Optional** Demographics and Dynamics of Mechanical Turk Workers. Difallah et al. WSDM 2018 The Crowd is a Collaborative Network. Gray et al. CSCW 2016. The Communication Network Within the Crowd. Yin et al. WWW 2016.