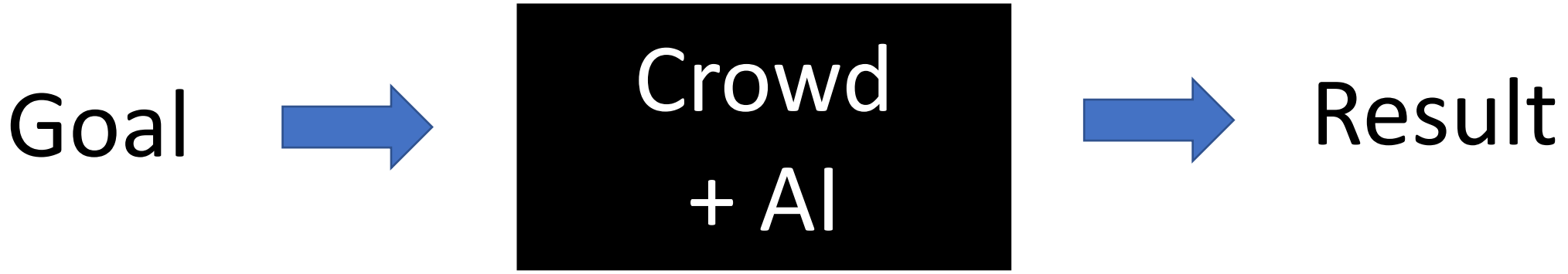
Ethics in AI / ML

Chien-Ju (CJ) Ho

Crowdsourcing: From a Task Solver's Perspective



- Understanding humans, developing realistic ***human models***, and incorporating them into the ***computation framework***.
- Multidisciplinary in nature

Machine Learning | Economics Theory | Computational Social Science | Optimization | Human-Computer Interaction | and more…

Beyond Solving Objective Tasks

- Fair division among the crowd
- Crowd research: open and scalable lab
- Crowdsourcing democracy
- Incentives in Blockchain
- Ethics issues of AI and ML (Today's Focus)

Crowdsourcing:
From Workers' Perspective

Short Discussion

- From workers' perspectives, what do you think are "wrong" for the current crowdsourcing platforms? How do you think we can fix it?

- To put these questions into context, consider the following question

"Can we foresee a future crowd workplace in which we would want our children to participate?"

Crowdsourcing =/= Mechanical Turk

Can We Learn from Traditional Workplaces

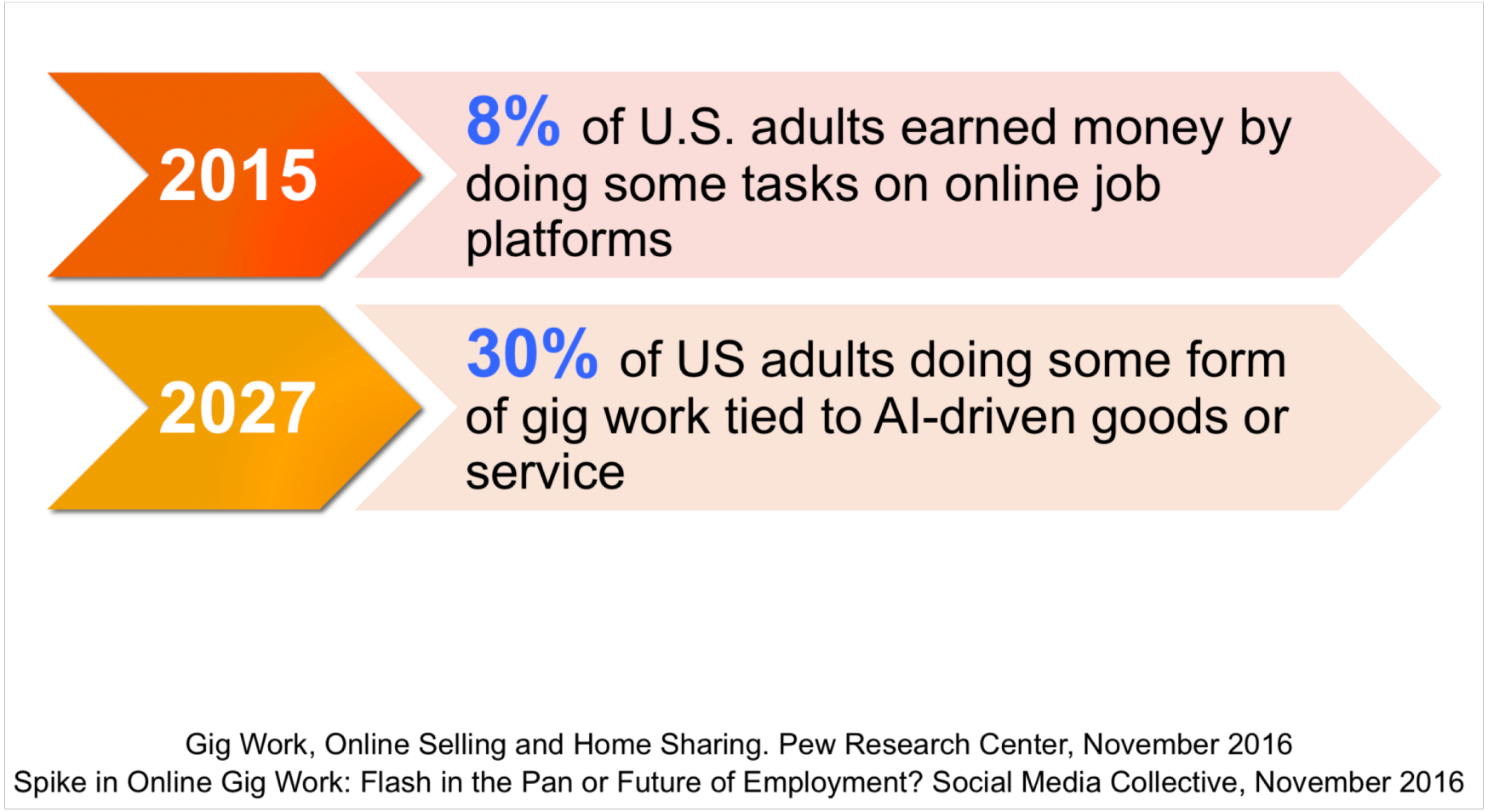
Traditional	On-demand
	
Centralized	*Distributed*
	
Career	*Piecework*
	
Management Strategies	*???*

Improving Worker Welfare

- Discussion among crowdsourcing researchers [Kittur et. CSCW 2013]
  - Create career ladders
    - Motivation, job design, reputation, hierarchy
  - Improve task design through better communications
    - Quality assurance, job design, task assignment, real-time crowd work, synchronous collaboration, platforms
  - Facilitate learning
    - Reputation and credentials, AIs guiding crowds, crowds guiding AIs, task assignment, quality assurance

- Various guidelines for conducting crowdsourcing
  - Guidelines for Academic Requesters
  - Responsible Research with Crowds: Pay Crowdworkers at Least Minimum Wage

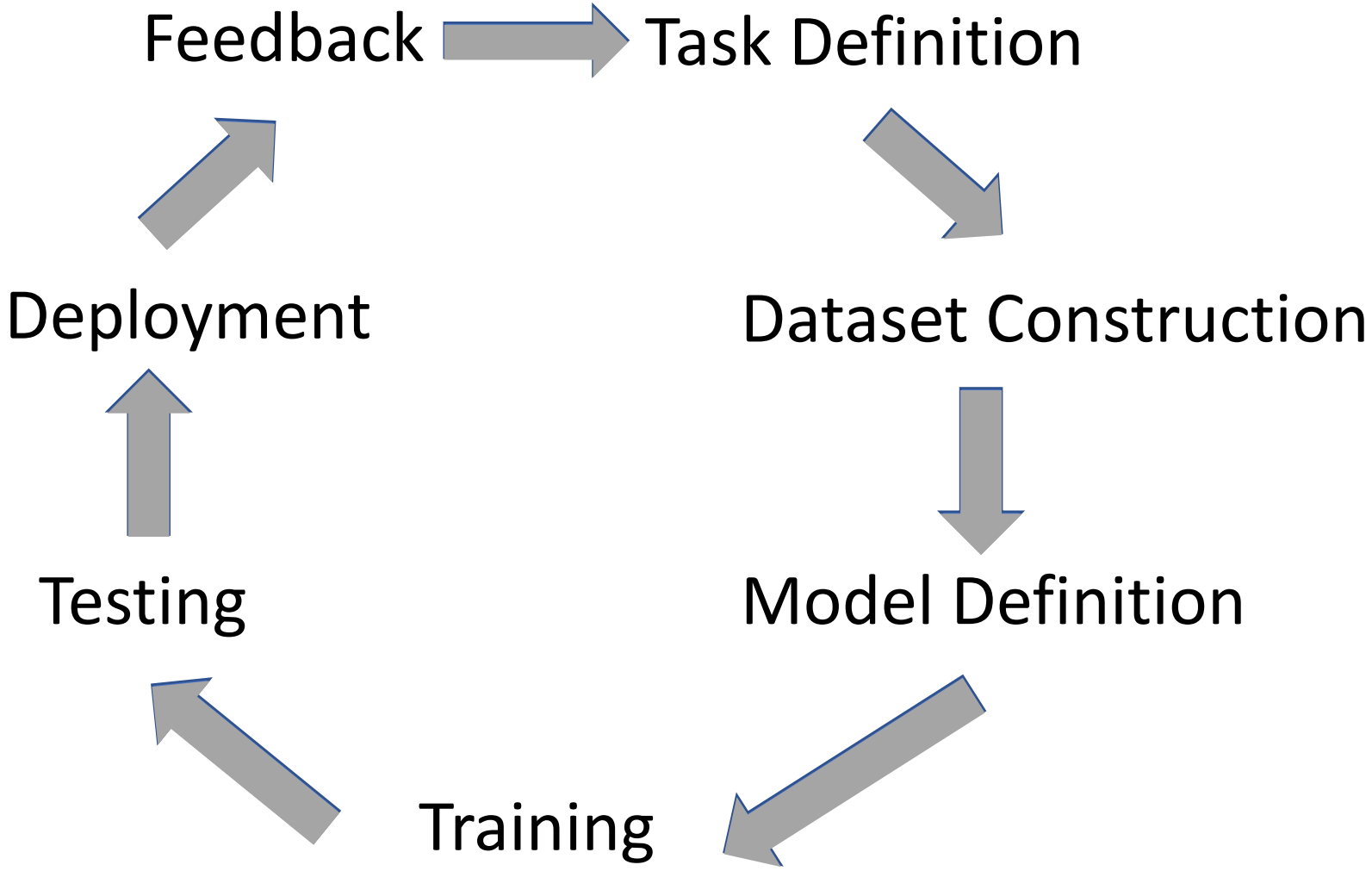
More than Crowdsourcing Markets: Gig-Economy



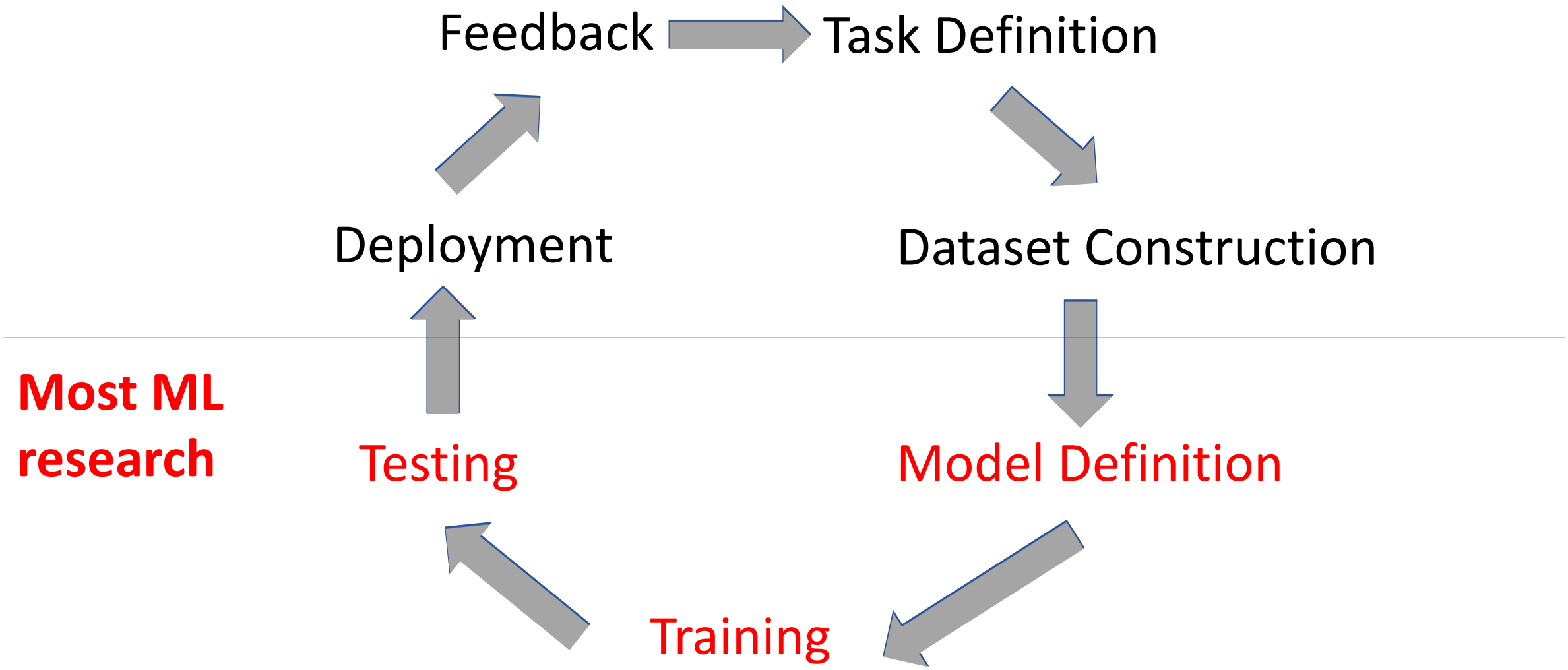
Gig Work, Online Selling and Home Sharing. Pew Research Center, November 2016
Spike in Online Gig Work: Flash in the Pan or Future of Employment? Social Media Collective, November 2016

Crowdsourcing:
From Machine Learning Perspective

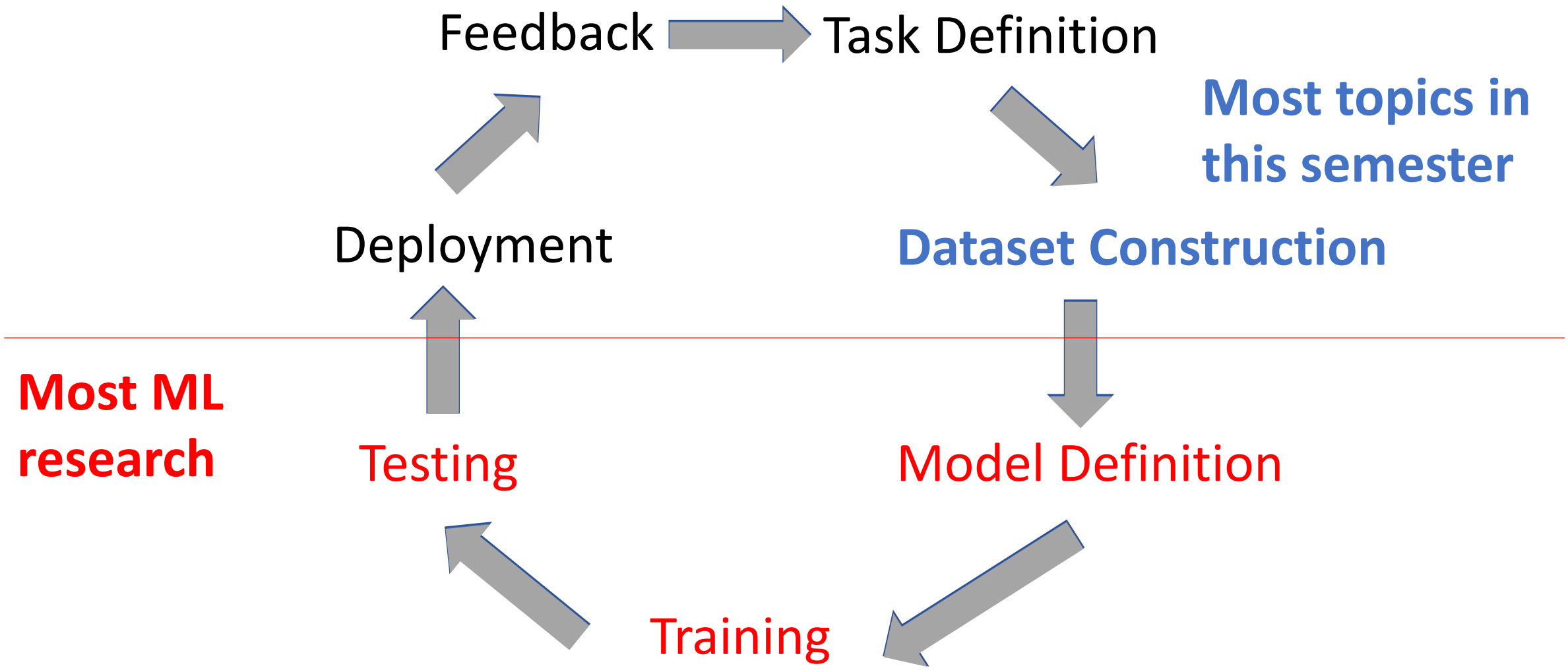
Machine Learning Lifecycle



Machine Learning Lifecycle

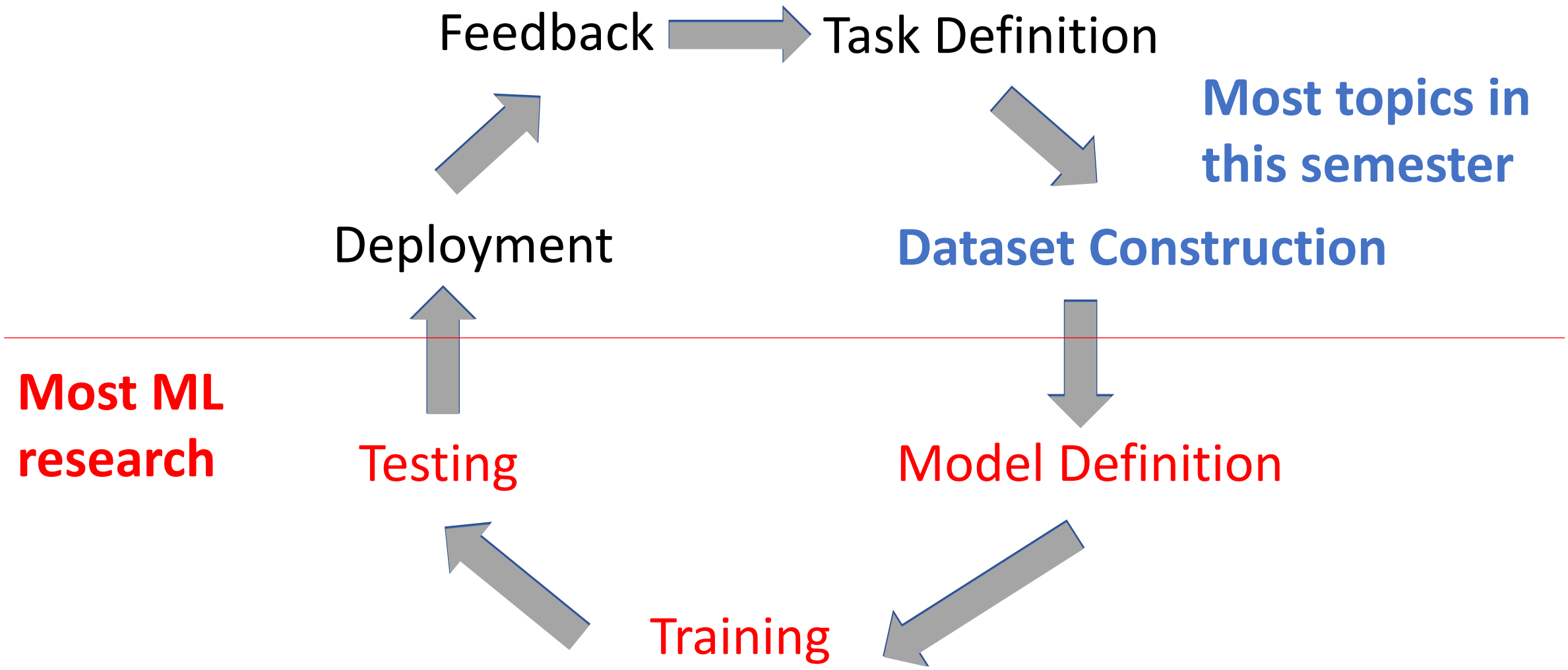


Machine Learning Lifecycle

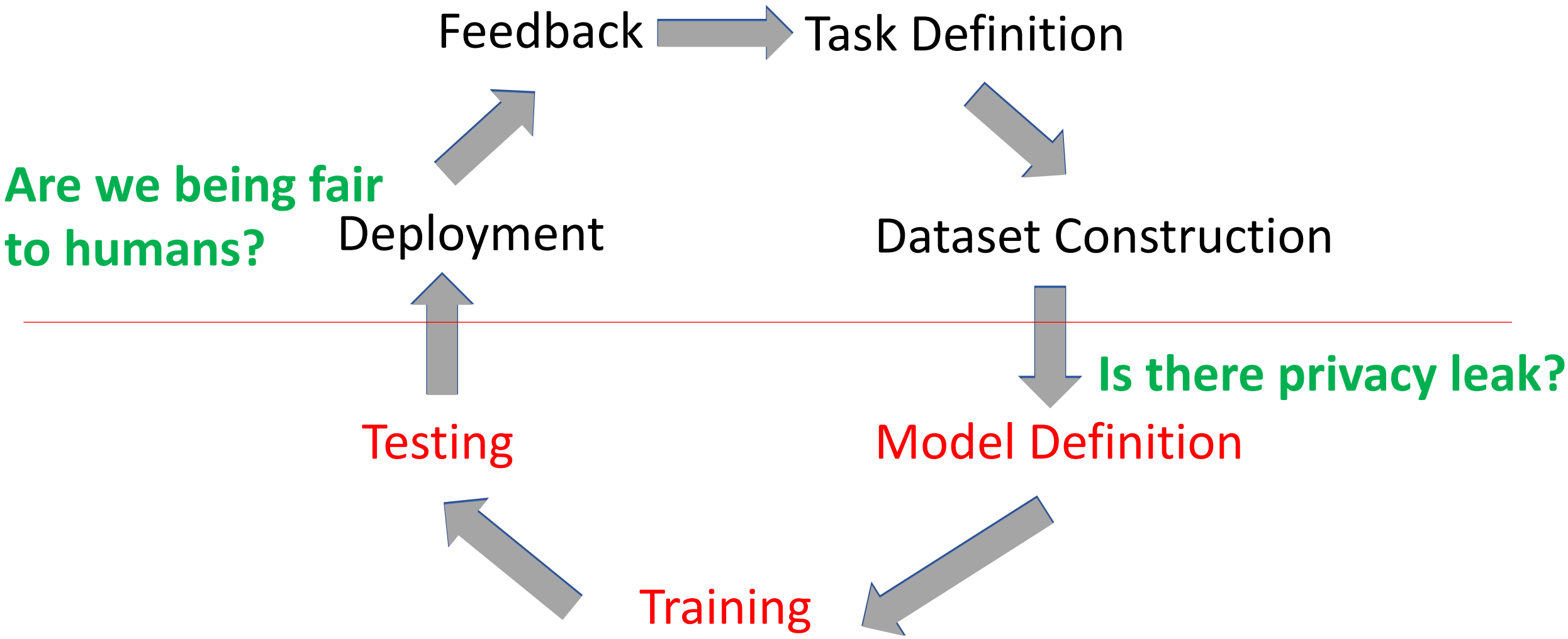


Machine Learning Lifecycle

Humans can be involved in every aspect of the process



Machine Learning Lifecycle



Ethical Issues in AI / ML

Focus on privacy and fairness

Netflix Challenges

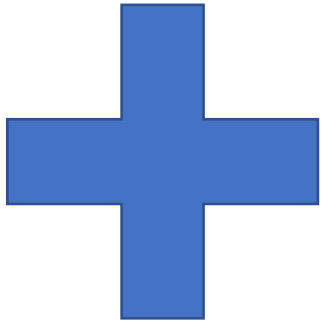
- First Netflix challenge
  - Announced in 2006
  - Released a dataset of 100,480,507 ratings that 480,189 users gave to 17,770 movies.
  - Award $1 million to first team beating their algorithm by 10%
  - Data format: <user, movie, date of grade, grade>
    - User and movie names are replaced with integers

- 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), 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

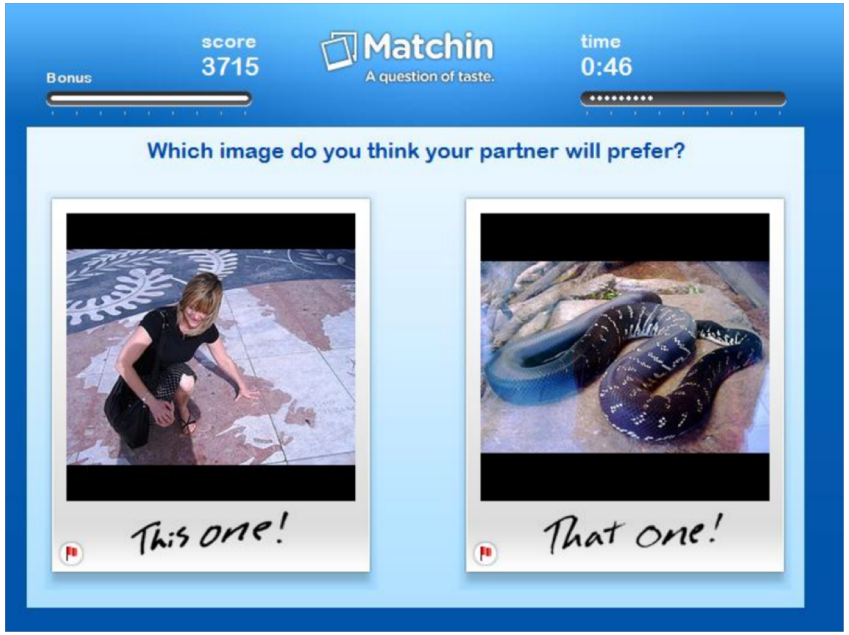
- What can we do?

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)

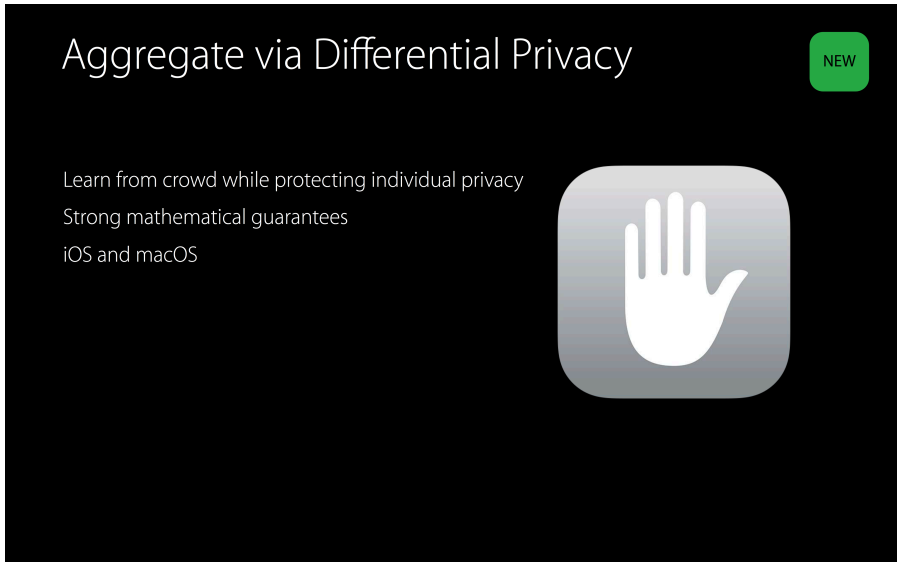
- Absolute privacy is not likely.
  - Who you are friends with might reveal who you are

(Maybe) More Reasonable Expectations

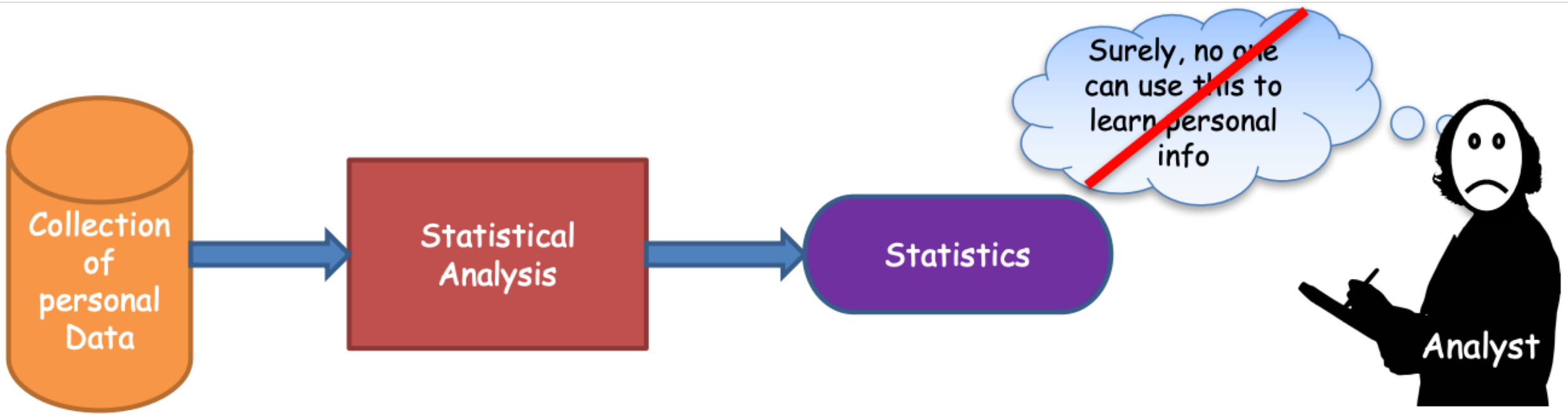
- Quantitative
  - Want a knob to tune the tradeoff between accuracy and privacy loss
- Plausible deniability
  - You presence in a database cannot be ascertained
- Prevent targeted attacks
  - Limit information leaked even with side knowledge

Differential Privacy

- A formal “notation” 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.

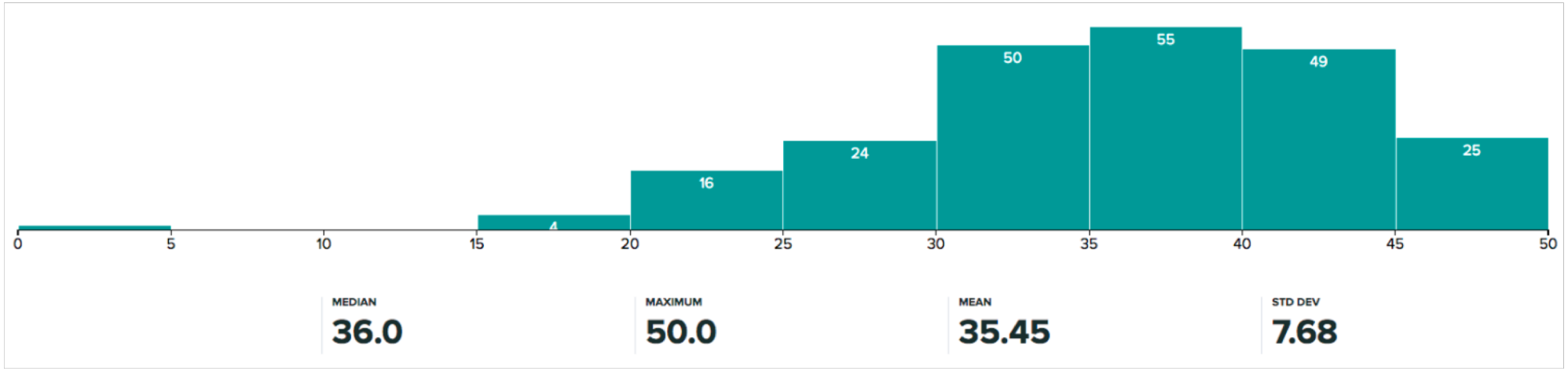


Differential Privacy



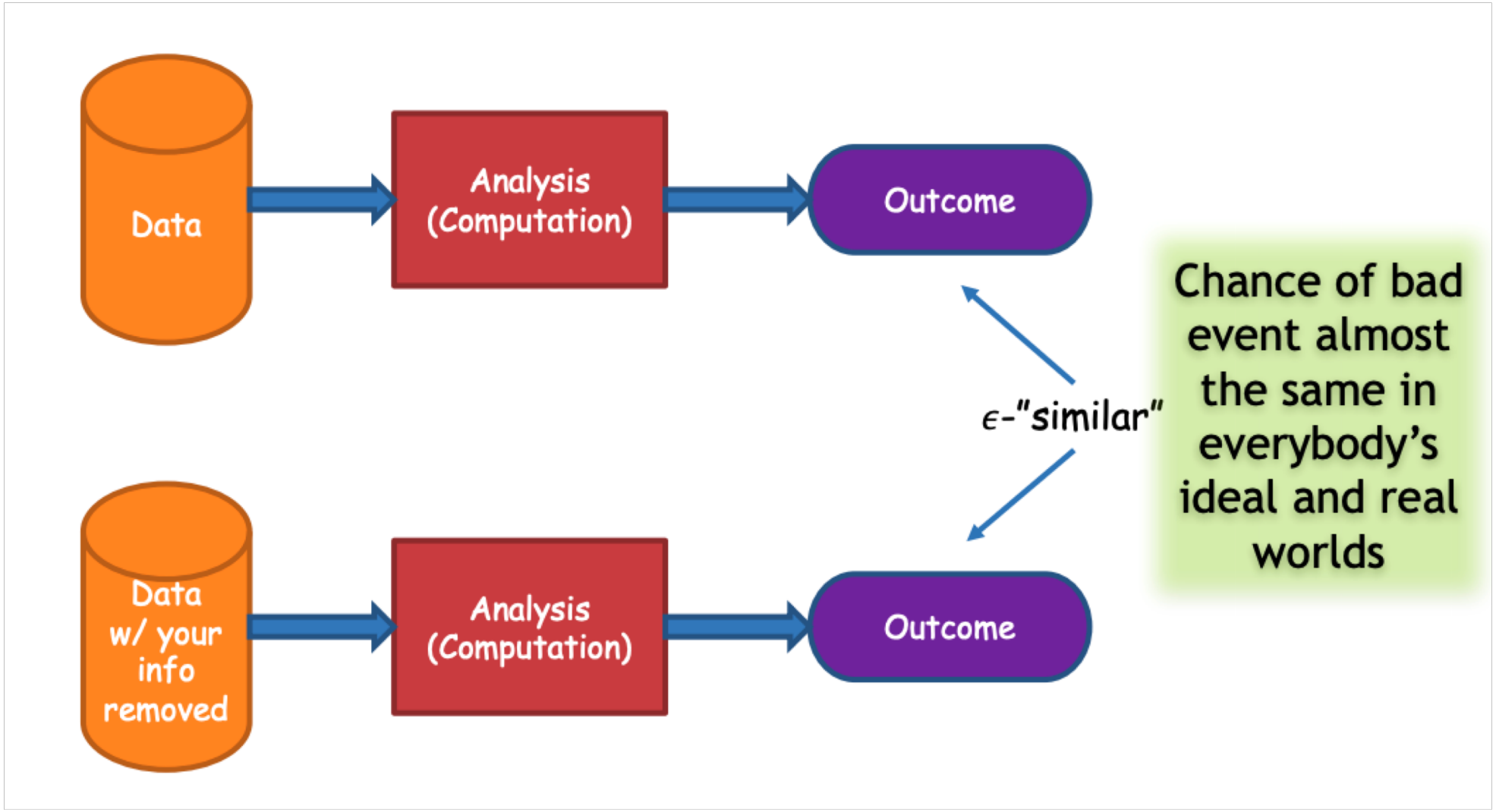
Differential Privacy

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



- How much of the privacy 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

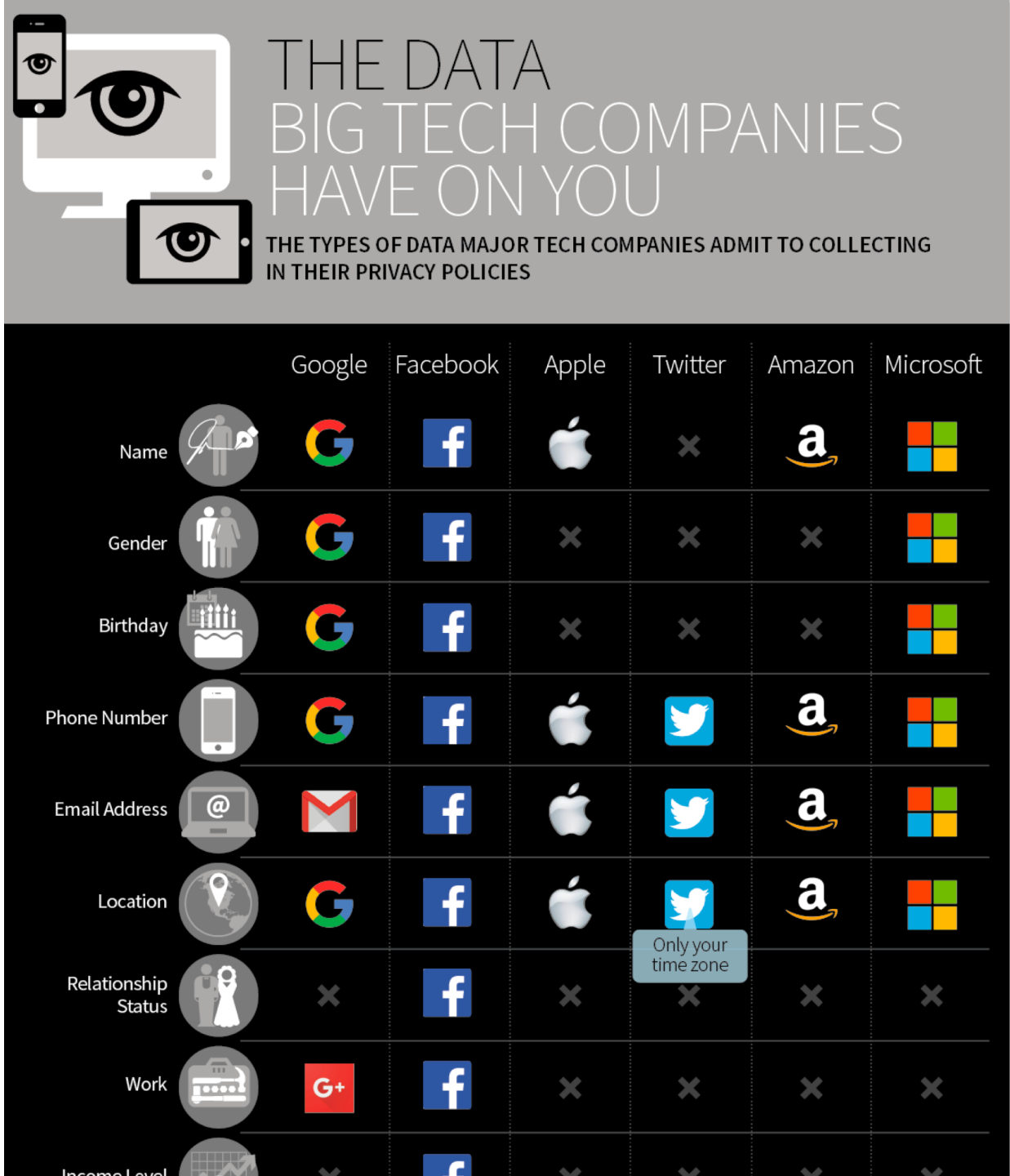
- Let the output of A be the average of users' ages
- Consider two extreme cases
  - If the size of the database is 1
  - If the size of the database is infinity

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

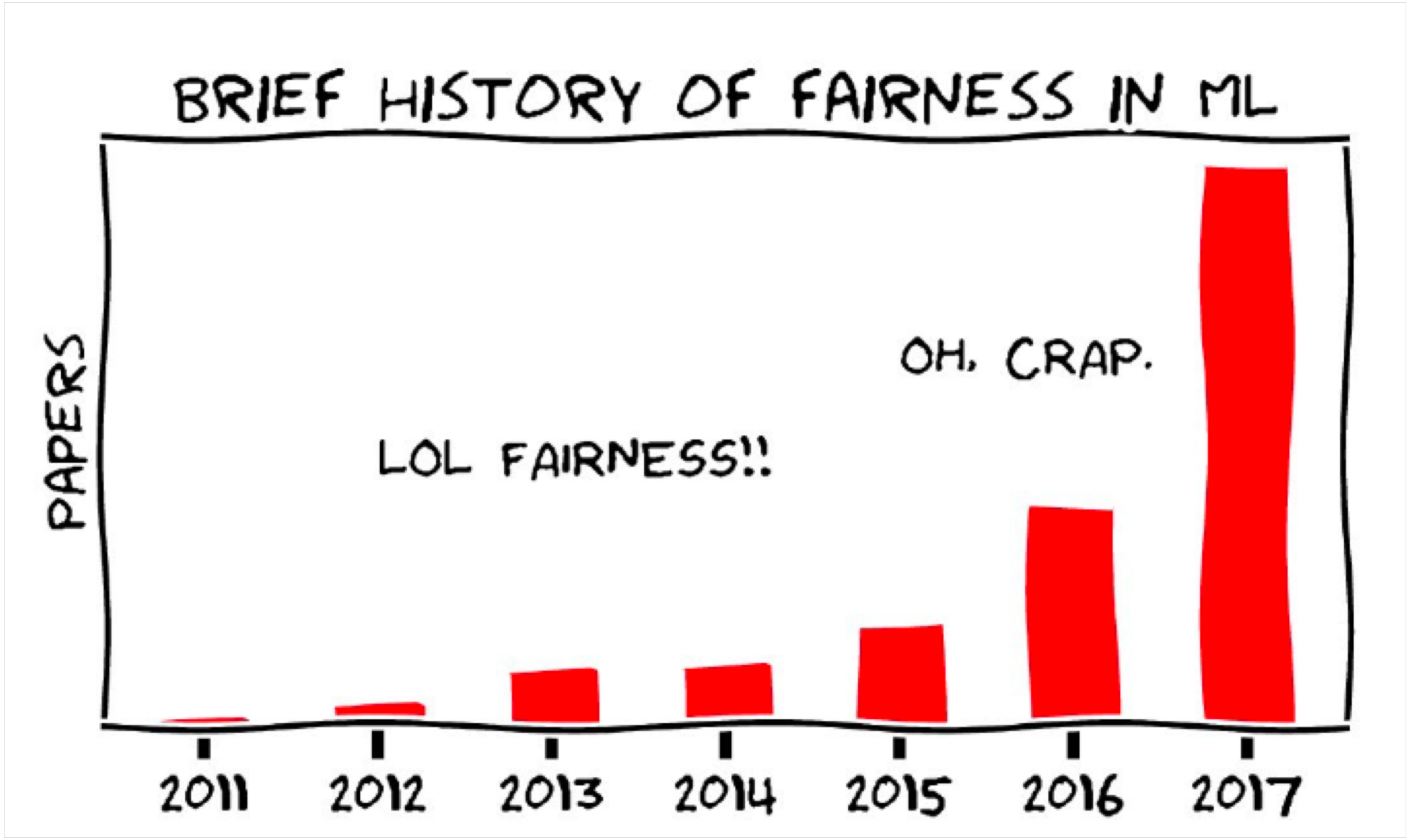
Discussion

- 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?



Fairness


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.
2011
2012
2013
2014
2015
2016
2017

Isn’t the point of ML to discriminate?

Want to avoid “unjustified” discrimination.

Example: Loan Applications

- By law, the banks can’t discriminate people according to their race.
- First natural approach (fairness through blindness)
  - remove the race attribute from the data
- Guess what happened?
  - Redlining

What should we do?

- From computer scientists / engineers' point of view….
- Give me an operational definition of fairness, I'll implement a system that satisfy it!

- How should we define fairness?

Another Example: Probation Decisions

- Using a ML classifier to predict whether the prisoner will commit a crime after probation.

- Consider the case with two sub-groups (e.g., two races), what does it mean to be fair if we apply this to two different races?

	Labeled Low-Risk	Labeled High-Risk
Won't Recidivate	TN1	FP1
Will Recidivate	FN1	TP1

- Defendant: the probability that I'm incorrectly classified high-risk is independent of my race.
  - Equal False Positive Rate: FP1 = FP2

- Defendant: the probability that I'm incorrectly classified as low-risk is independent of my race.
  - Equal False Negative Rate: FN1 = FN2

	Labeled Low-Risk	Labeled High-Risk
Won't Recidivate	TN2	FP2
Will Recidivate	FN2	TP2

- Decision-maker: the ratio of people who recidivated, among the ones labeled high-risk, is independent of race.
  - Equal predictive value: $\frac{\mathrm{TP1}}{\mathrm{TP1}+\mathrm{FP1}} = \frac{\mathrm{TP2}}{\mathrm{TP2}+\mathrm{FP2}}$

Discussion:

- Can your mathematically represent the above fairness criteria?
- Are there other possible math definitions?

	Labeled Low-Risk	Labeled High-Risk
Won't Recidivate	TN1	FP1
Will Recidivate	FN1	TP1

- Defendant: the probability that I'm incorrectly classified high-risk is independent of my race.
  - Equal False Positive Rate: FP1 = FP2

- Defendant: the probability that I'm incorrectly classified as low-risk is independent of my race.
  - Equal False Negative Rate: FN1 = FN2

	Labeled Low-Risk	Labeled High-Risk
Won't Recidivate	TN2	FP2
Will Recidivate	FN2	TP2

- Decision-maker: the ratio of people recidivated, among the ones I labeled high-risk, is independent of race.
  - Equal predictive value: $\frac{\mathrm{TP1}}{\mathrm{TP1}+\mathrm{FP1}} = \frac{\mathrm{TP2}}{\mathrm{TP2}+\mathrm{FP2}}$

Impossibility Result [Kleignberg et al. 2016]

The above three conditions cannot be satisfied simultaneously, unless the predictor is perfect or the two groups are the same.

There are more possible definitions of fairness

- Society (maybe think of hiring instead of probations): is the selected set demographically balanced?

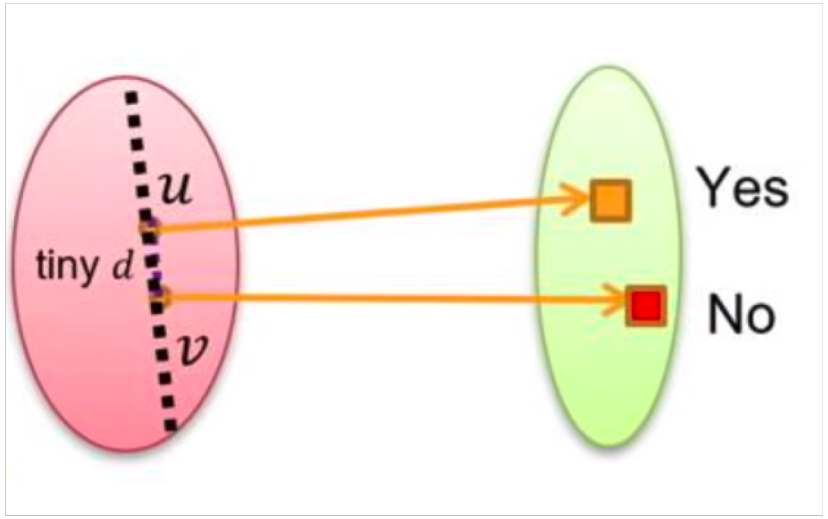
- FAT*18 Tutorial: 21 Definitions of Fairness and Their Politics. Arvind Narayanan.

Individual Fairness

- So far we have been mostly talked about “group fairness”. There is another notion of “individual fairness”

- Cucumbers and Grapes
  - https://youtu.be/-KSryJXDpZo

- One way of thinking about fairness:
  - Similar people should be treated similarly
  - Need to impose some “smooth” notion.
  - How do we define similarity.

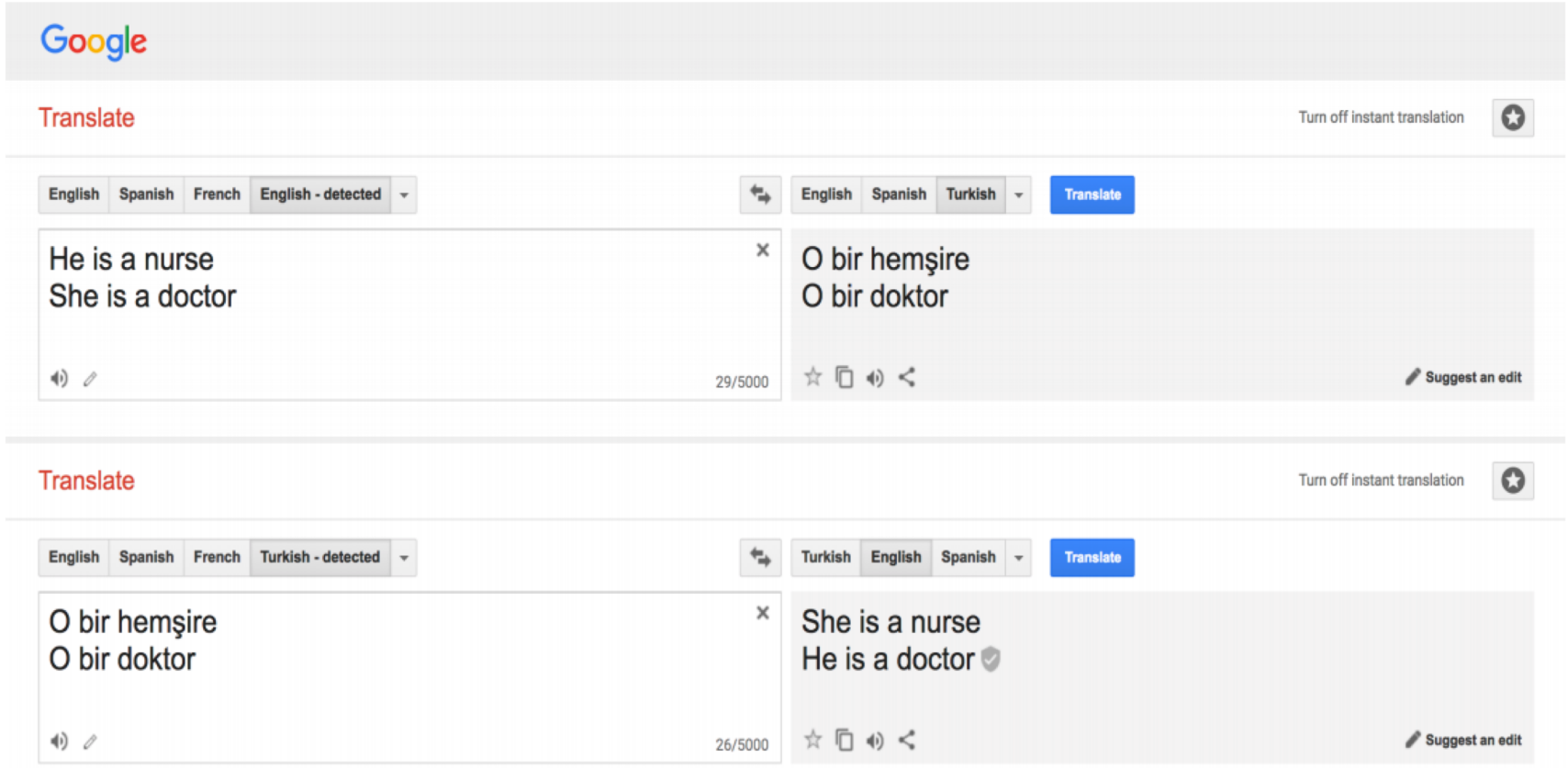


Take-Aways

- As AI/ML becomes more ubiquitous in our daily decision making, ethical issues are getting more important as well.
- **Being aware** of the issues is the important first step!
- ”Solving” the issues (if at all possible) requires communications among people in different disciplinaries.

- More references:
  - NIPS17 Tutorial: https://vimeo.com/248490141
  - FAT*18 Tutorial: https://www.youtube.com/watch?v=jIXIuYdnyyk
  - FAT*19 Putting fairness in practice: https://algorithmicbiasinpractice.wordpress.com

More Examples on Bias



[Caliksan et al., 2017]

More Examples on Bias

[Kay et al., 2015]

Course Recap