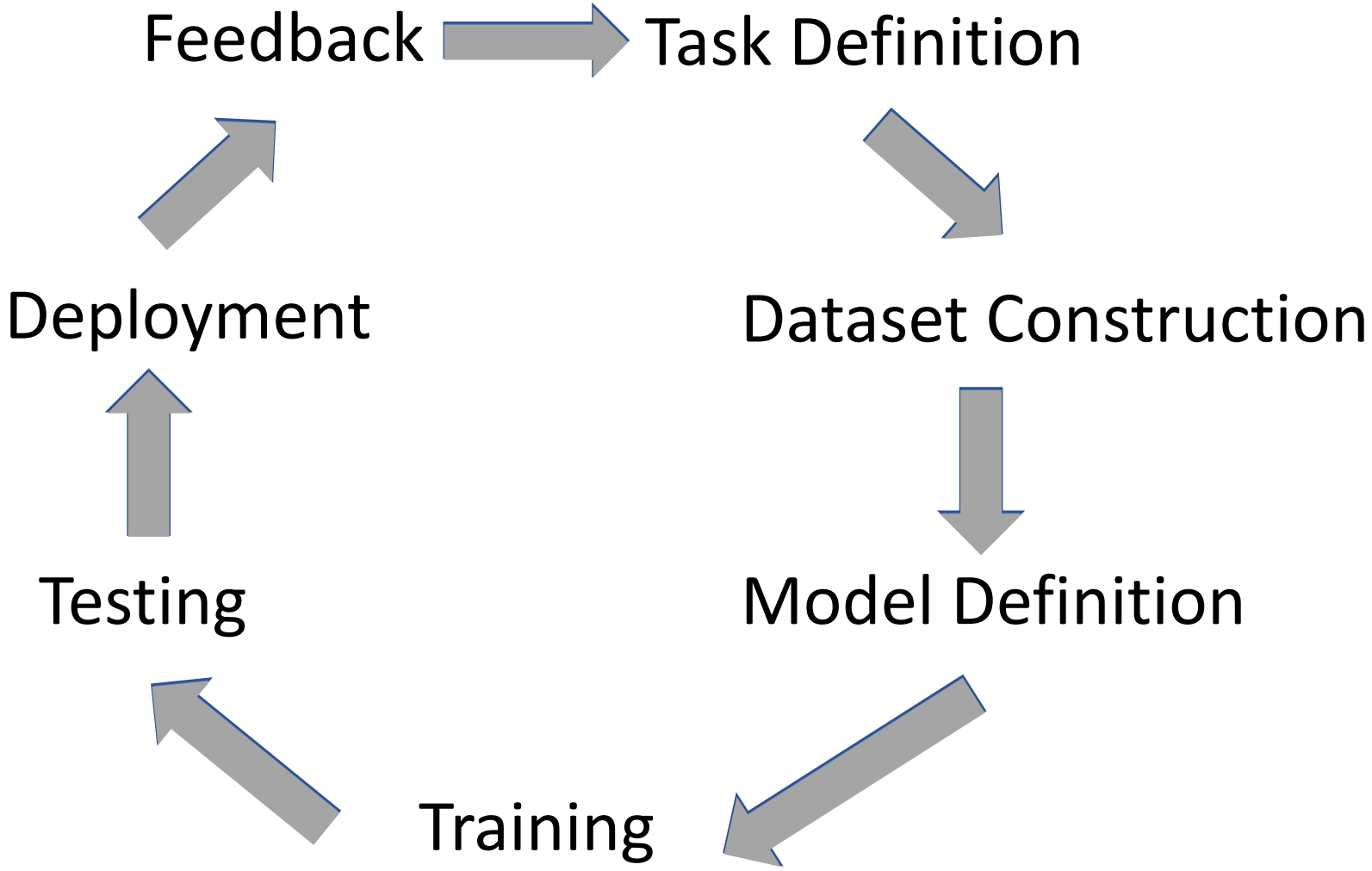
CSE 417T
Introduction to Machine Learning

Lecture 24

Instructor: Chien-Ju (CJ) Ho

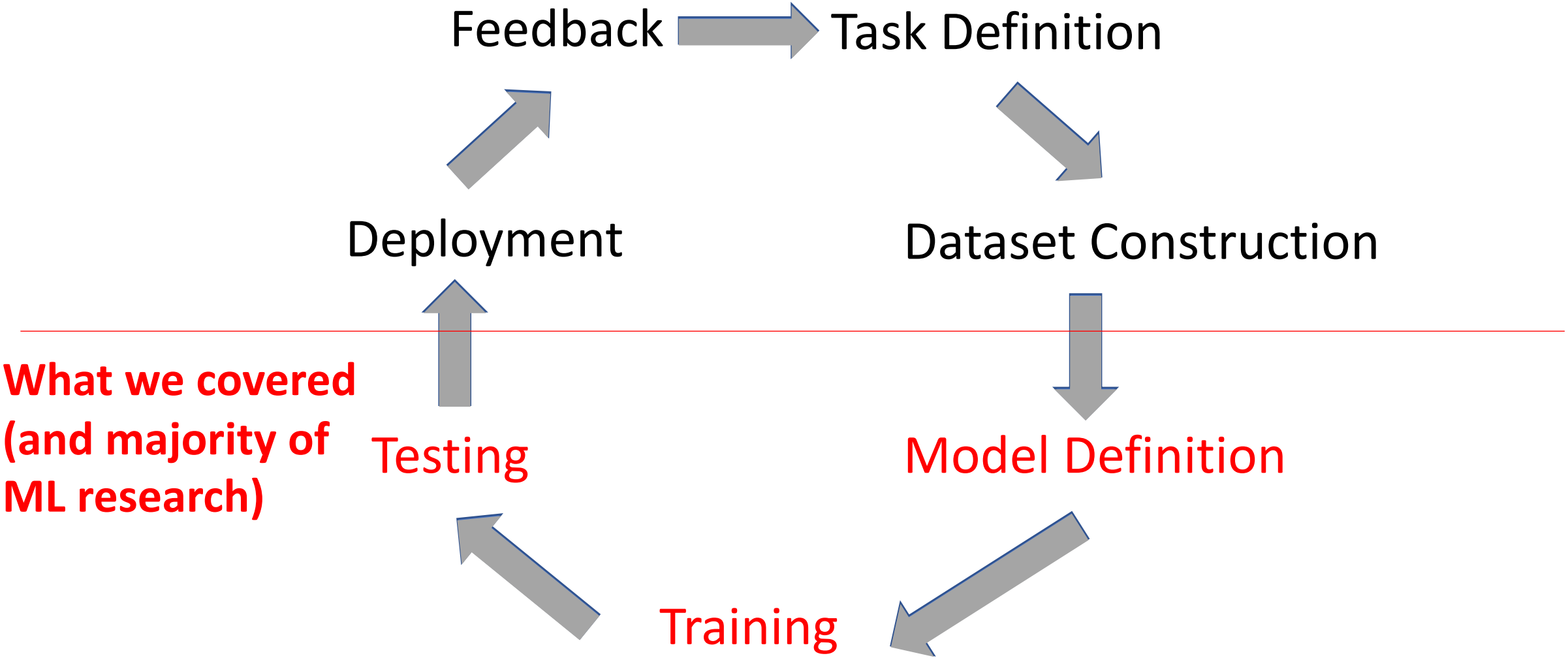
Recap

Machine Learning Lifecycle



Machine Learning Lifecycle

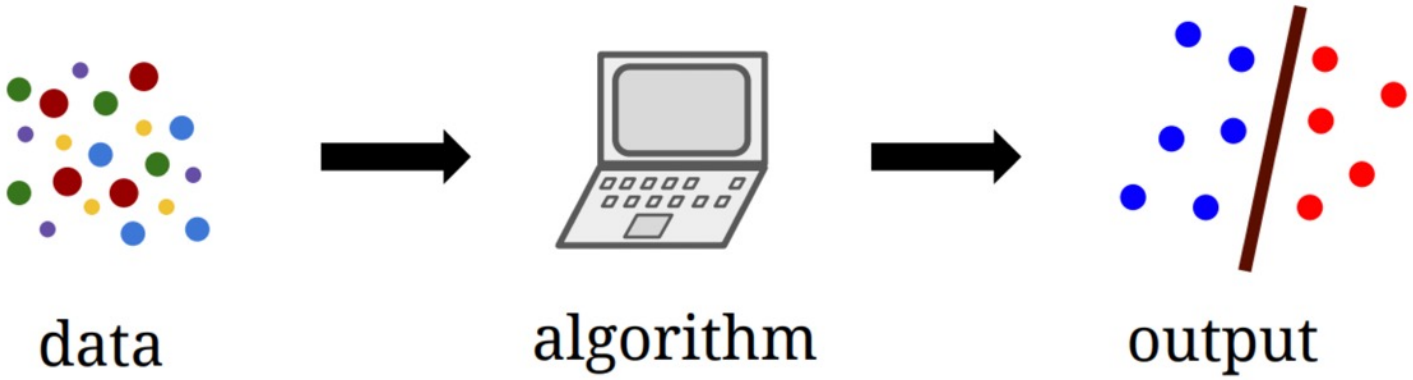
For ML to have "positive" impacts, we need to be careful in every stage



What we covered (and majority of ML research)

Classification

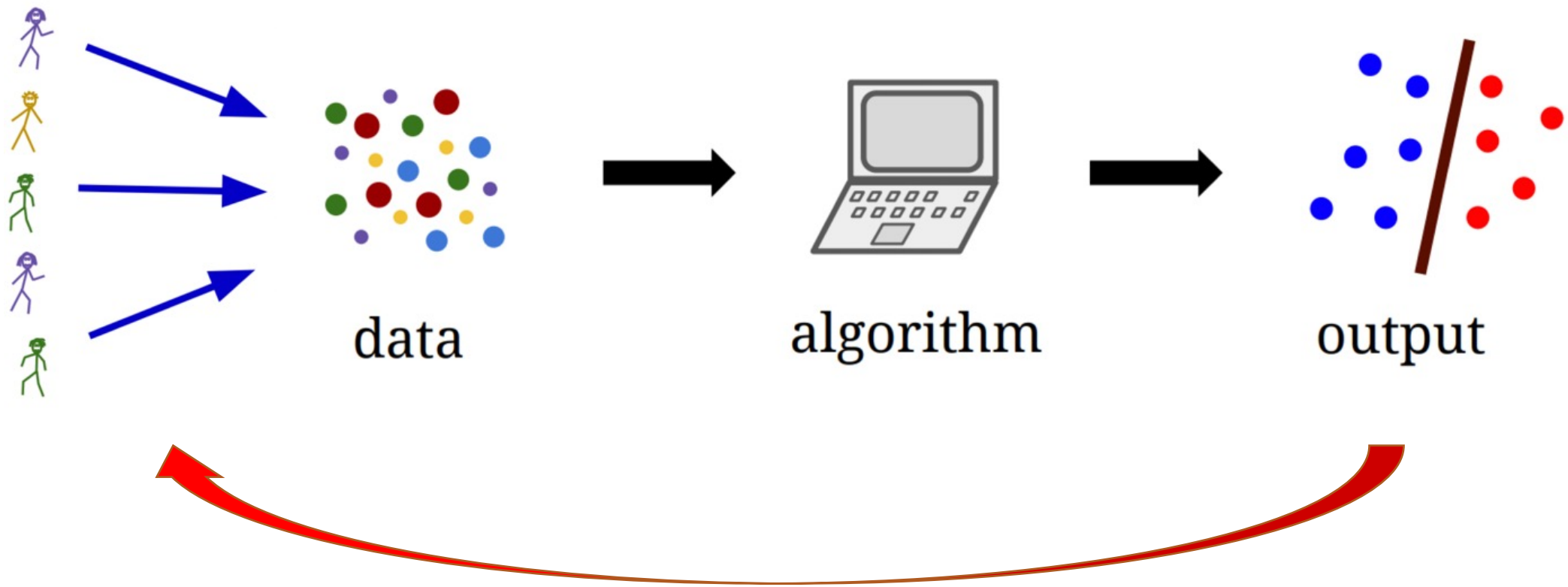
- Standard setup of (supervised) machine learning



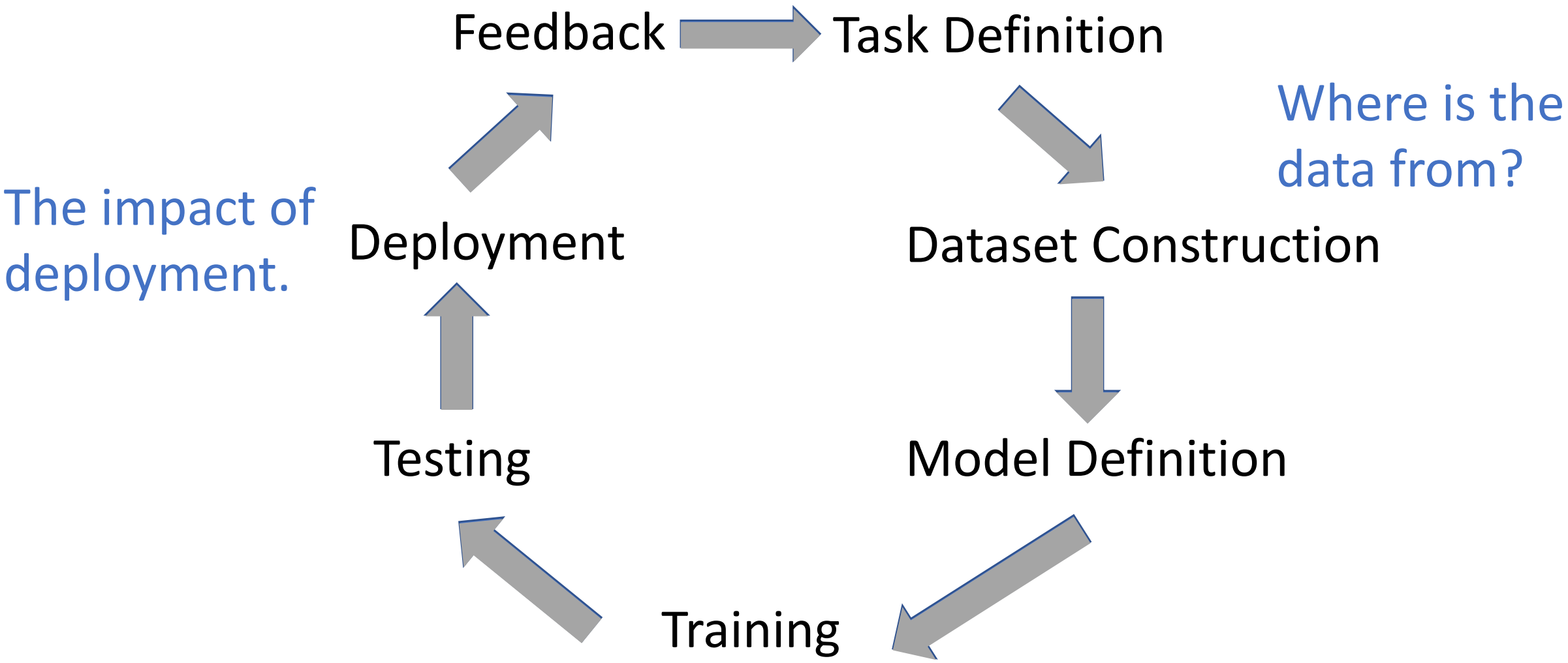
  - Finding patterns from the given training datasets
  - Use the pattern to make predictions on new testing data

- Fundamental assumption:
  - Training and testing data points are i.i.d. drawn from the same distribution

Strategic Classification



Machine Learning Lifecycle



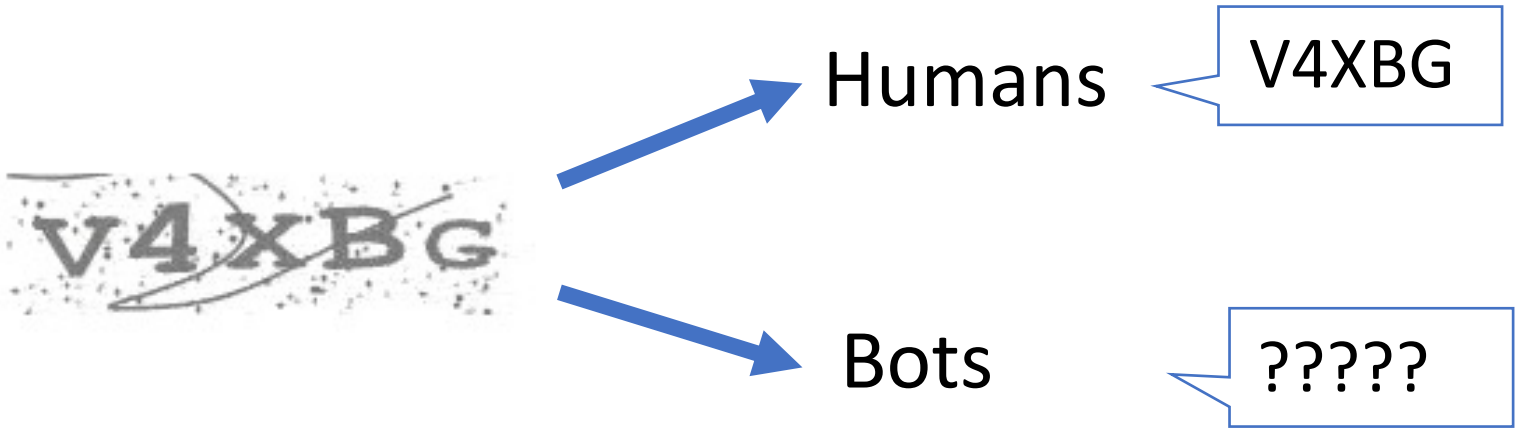
Today’s Lecture

Modern ML is driven by data.

Where does data come from?

CAPTCHA

Completely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part

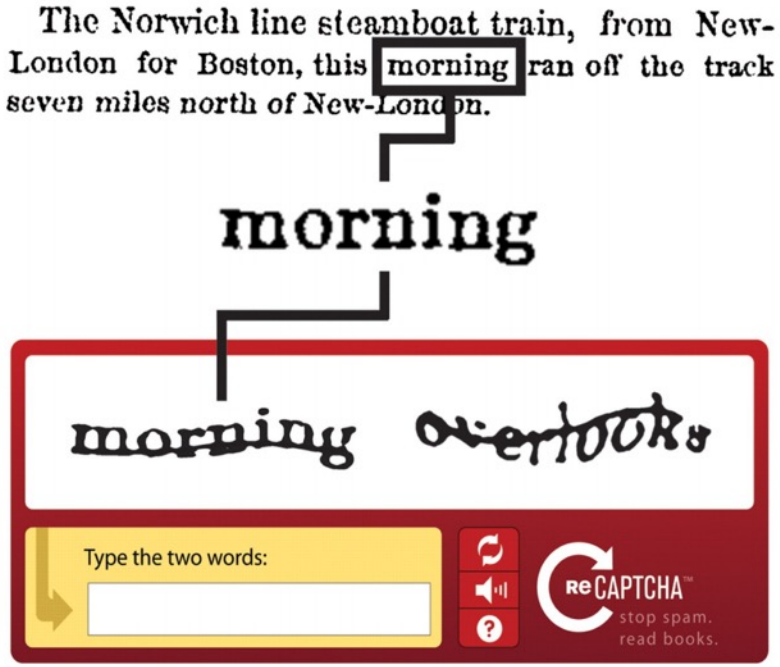


Roughly 200 million CAPTCHAs are typed every day*

10s of human time per CAPTCHA

Can we utilize this wasted human computation power?

*statistics around 2011



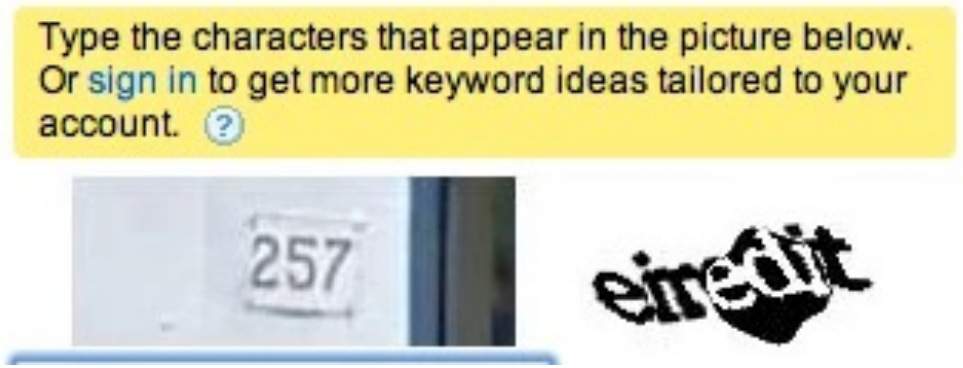
Word 1: an OCR task to solve
Word 2: tell apart humans and bots

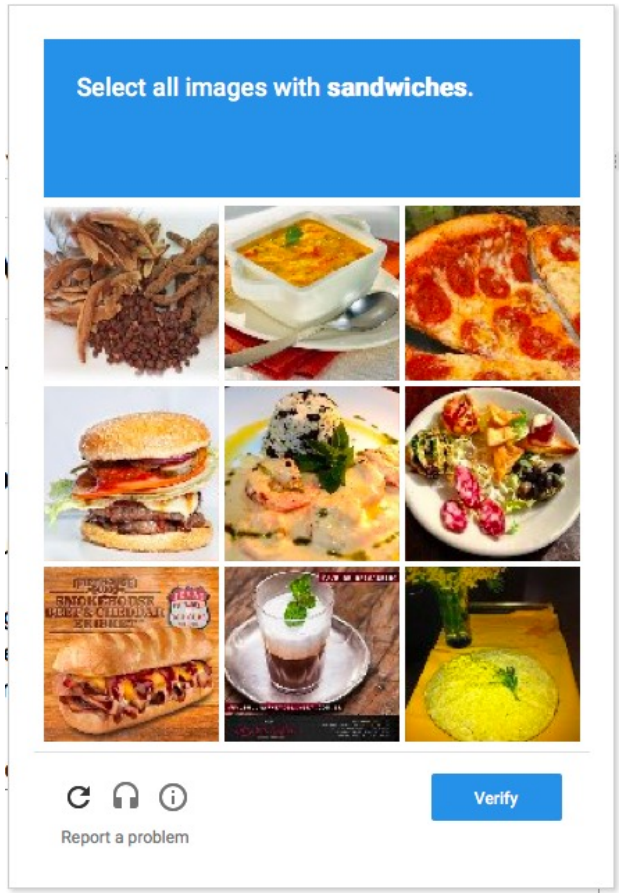
"reCAPTCHA has completely digitized the archives of The New York Times and books from Google Books, as of 2011"

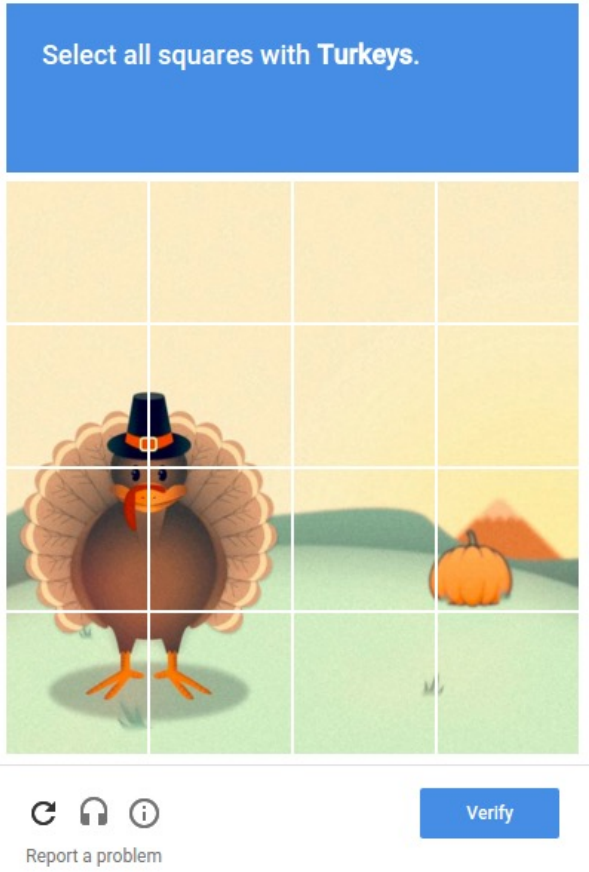
von Ahn et al. reCaptcha: Human-based Character Recognition via Web Security Measures. Science, September 2008

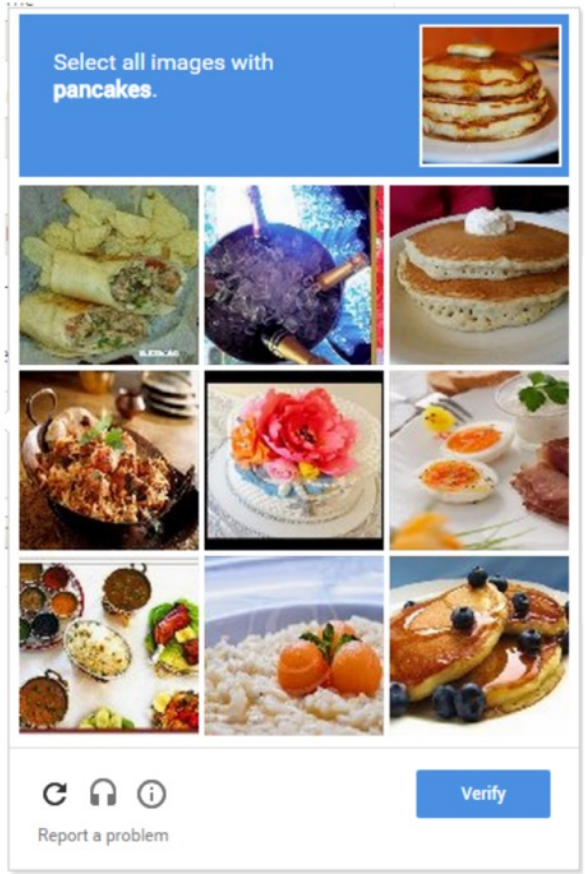
More than recognizing text

- Google acquired reCAPTCHA in 2009.



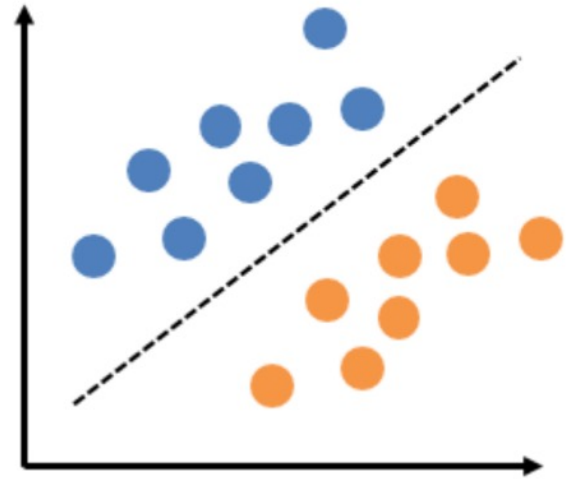






Training Data

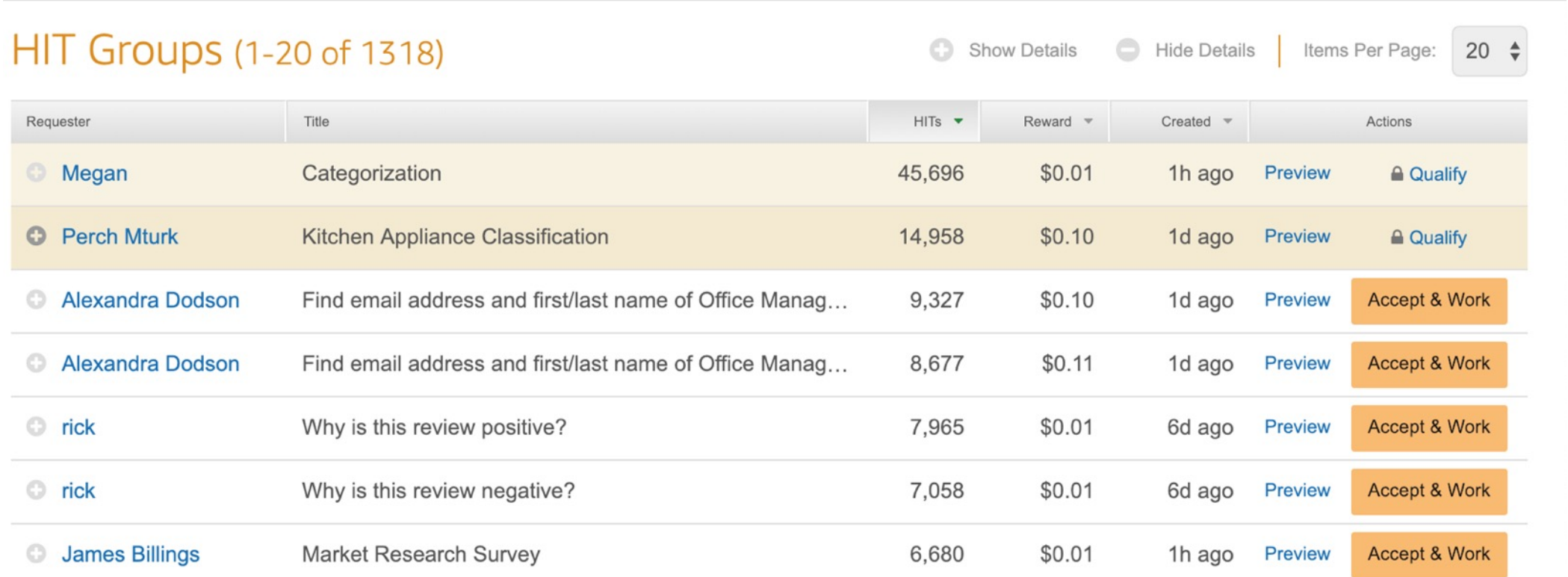
Hard Tasks



Data is often generated by humans

Explicitly: Human Labelers

- Amazon Mechanical Turk: Artificial Artificial Intelligence
  - A marketplace to collect data from humans
  - E.g., ImageNet has utilized this platform to collect image labels



HIT Groups (1-20 of 1318)

Show Details | Hide Details | Items Per Page: 20

Requester	Title	HITs	Reward	Created	Actions	
Megan	Categorization	45,696	$0.01	1h ago	Preview	Qualify
Perch Mturk	Kitchen Appliance Classification	14,958	$0.10	1d ago	Preview	Qualify
Alexandra Dodson	Find email address and first/last name of Office Manag…	9,327	$0.10	1d ago	Preview	Accept & Work
Alexandra Dodson	Find email address and first/last name of Office Manag…	8,677	$0.11	1d ago	Preview	Accept & Work
rick	Why is this review positive?	7,965	$0.01	6d ago	Preview	Accept & Work
rick	Why is this review negative?	7,058	$0.01	6d ago	Preview	Accept & Work
James Billings	Market Research Survey	6,680	$0.01	1h ago	Preview	Accept & Work

Implicitly…


Google


twitter


yelp


Quora


NETFLIX


facebook


stackoverflow


gwap


YAHOO! ANSWERS


reCAPTCHA


duolingo


up


99designs


PredictWise


kaggle


[topcoder]™


Predictious


CrowdFlower

Data (labeled or generated by humans) is the main driving force of ML

Good: Humans help drive ML forward

But?

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”.

Data (labeled or generated by humans) is the main driving force of ML

Good: Humans help drive ML forward

lachine learning models leak personal info if raining data is compromised

tackers can insert hidden samples to steal secrets

anna Quach

Science

AI unmasks anonymous c privacy risks

Software that identifies unique playing styles cou

12 JAN 2022 · 2:30 PM · BY MATTHEW HUTSON

RESEARCH-ARTICLE

Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy

Authors: Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, Olga Russakovsky

Authors Info & Affiliations

Publication: FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency • January 2020 • Pages 547–558 • https://doi.org/10.1145/3351095.3375709

Microsoft Release a Twitter Chatbot in 2016

MICROSOFT | WEB | TL;DR

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

Via The Guardian | Source TayandYou (Twitter)

BUSINESS NEWS OCTOBER 9, 2018 / 10:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

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

It's super funny that Alexa can't understand my mom — until we need Alexa to use the web, drive a car, and do pretty much anything else.

Privacy Concerns

Machine learning models leak personal info if training data is compromised

Attackers can insert hidden samples to steal secrets

Katyanna Quach

Tue 12 Apr 2022 // 02:45 UTC

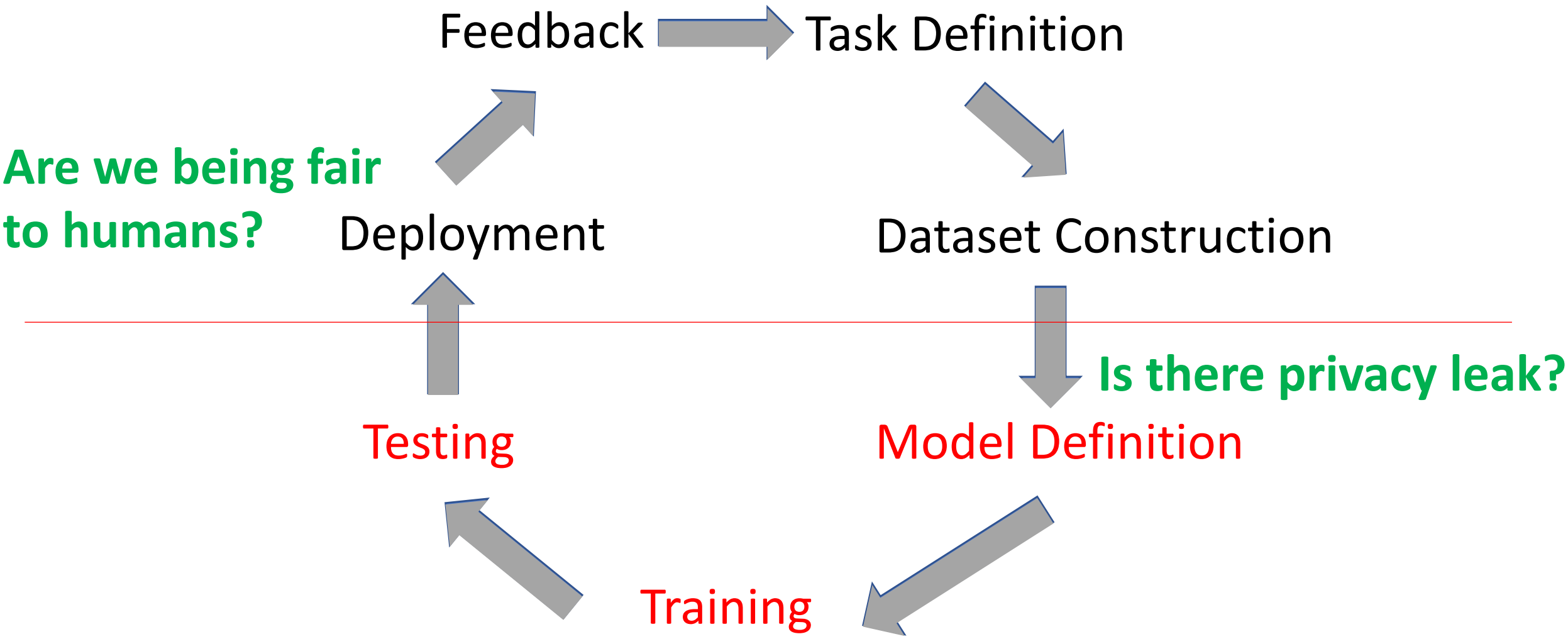
Science

AI unmasks anonymous chess players, posing privacy risks

Software that identifies unique playing styles could lead to better tutorials and game play

12 JAN 2022 · 2:30 PM · BY MATTHEW HUTSON

Machine Learning Lifecycle



Discussion on Privacy

Netflix Challenges

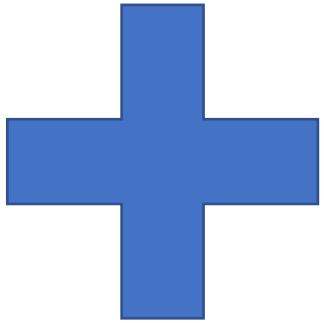
- First Netflix challenge
  - Announced in 2006
  - Released a dataset of 100,480,507 ratings that 480,189 users gave to 17,770 movies.
  - Award $1 million to first team beating their algorithm by 10%
  - Data format: <user, movie, date of grade, grade>
    - User and movie names are replaced with integers

- 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

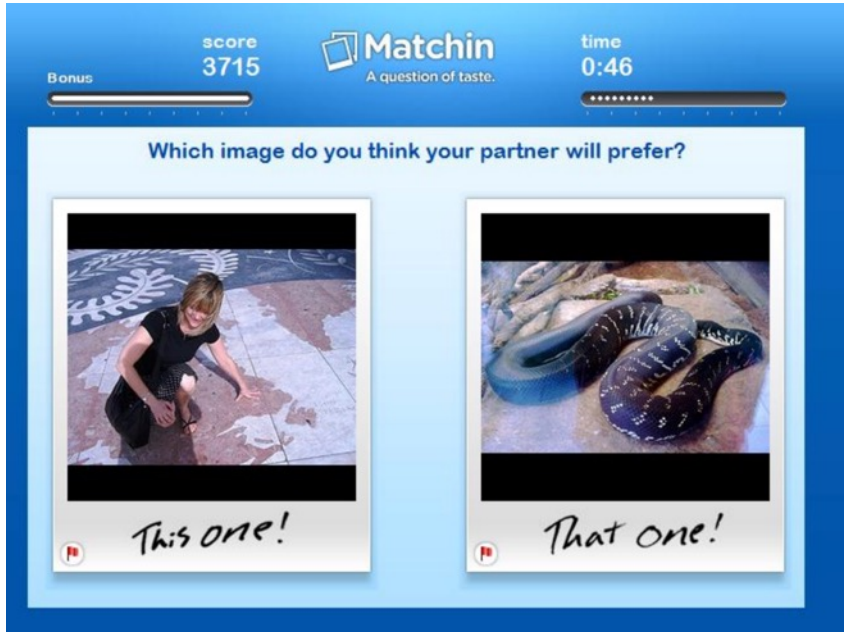
- What can we do?
  - Adding noises

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

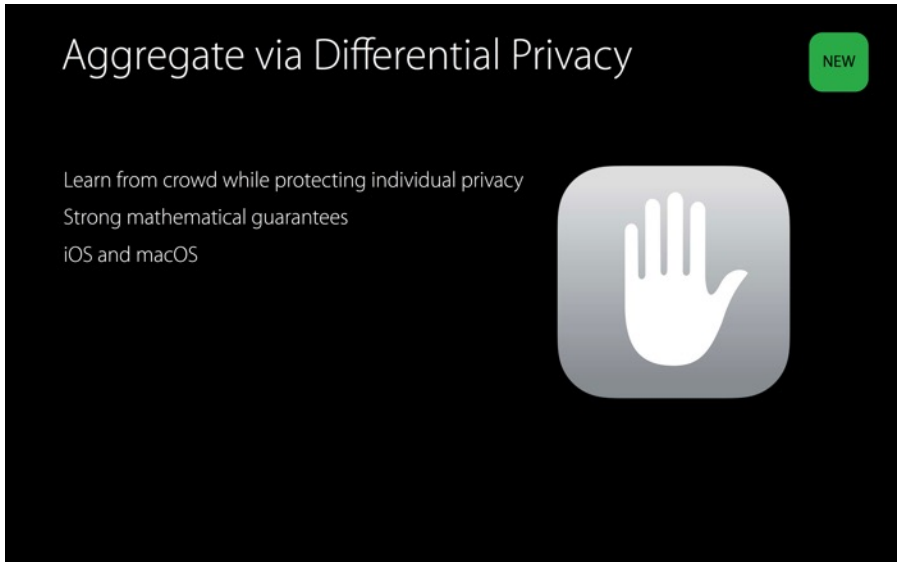
MIT Students' Facebook 'Gaydar' Raises Privacy Issues

(Maybe) More Reasonable Expectations

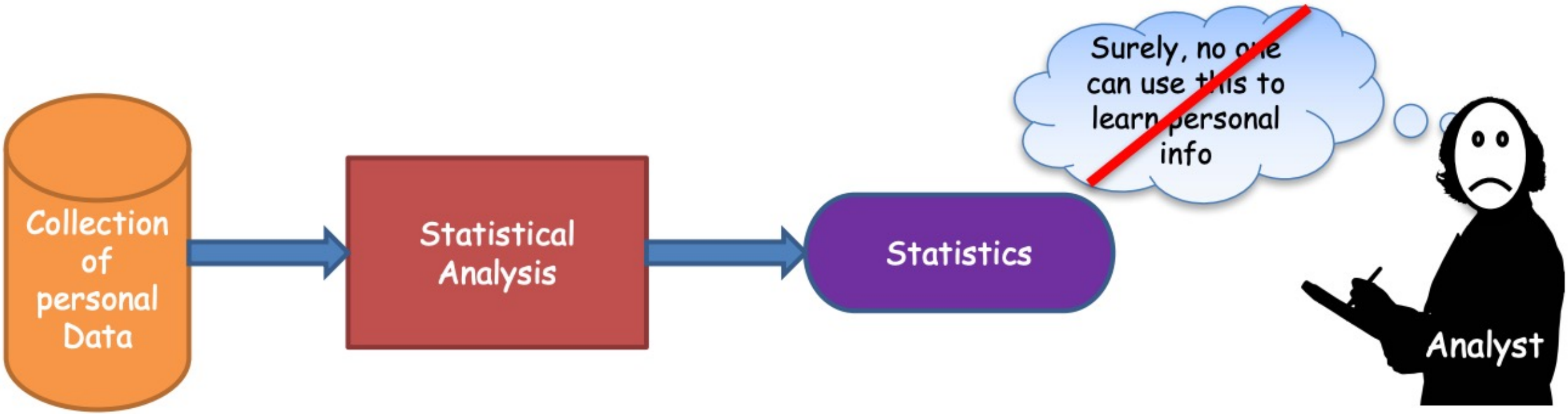
- Quantitative
  - Want a knob to tune the tradeoff between accuracy and privacy loss
- Plausible deniability
  - Your presence in a database cannot be ascertained
- Prevent targeted attacks
  - Limit information leaked even with side knowledge

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

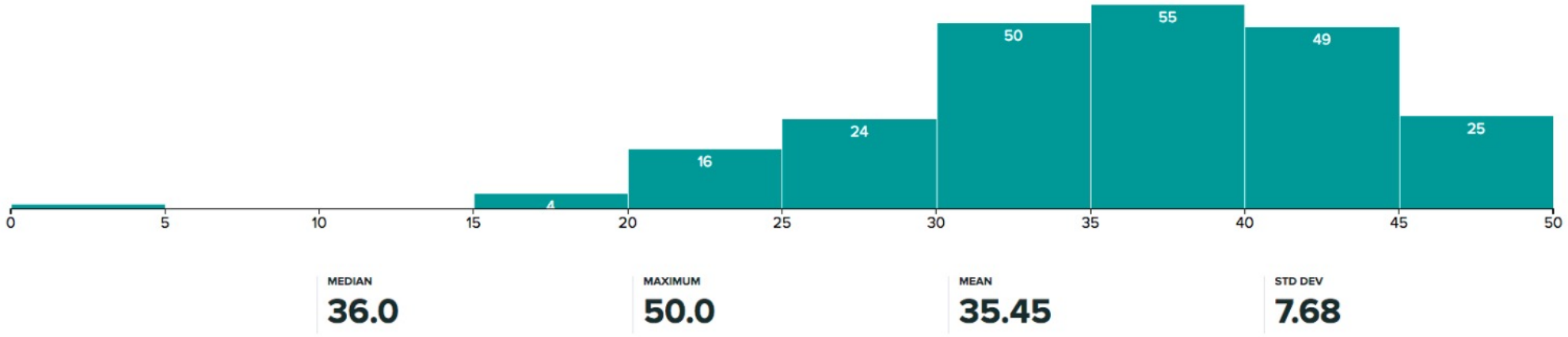


Differential Privacy



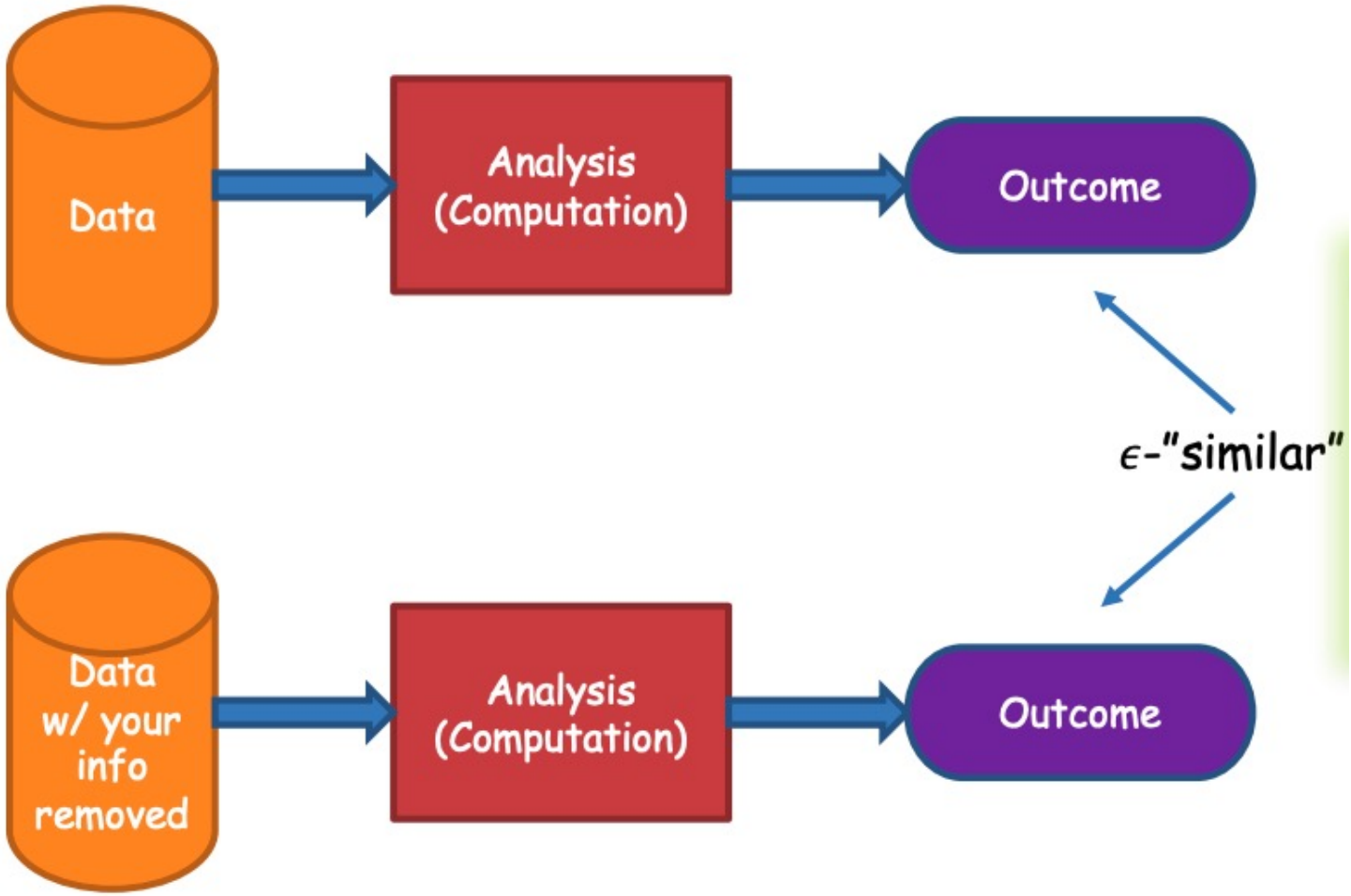
Differential Privacy

- 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

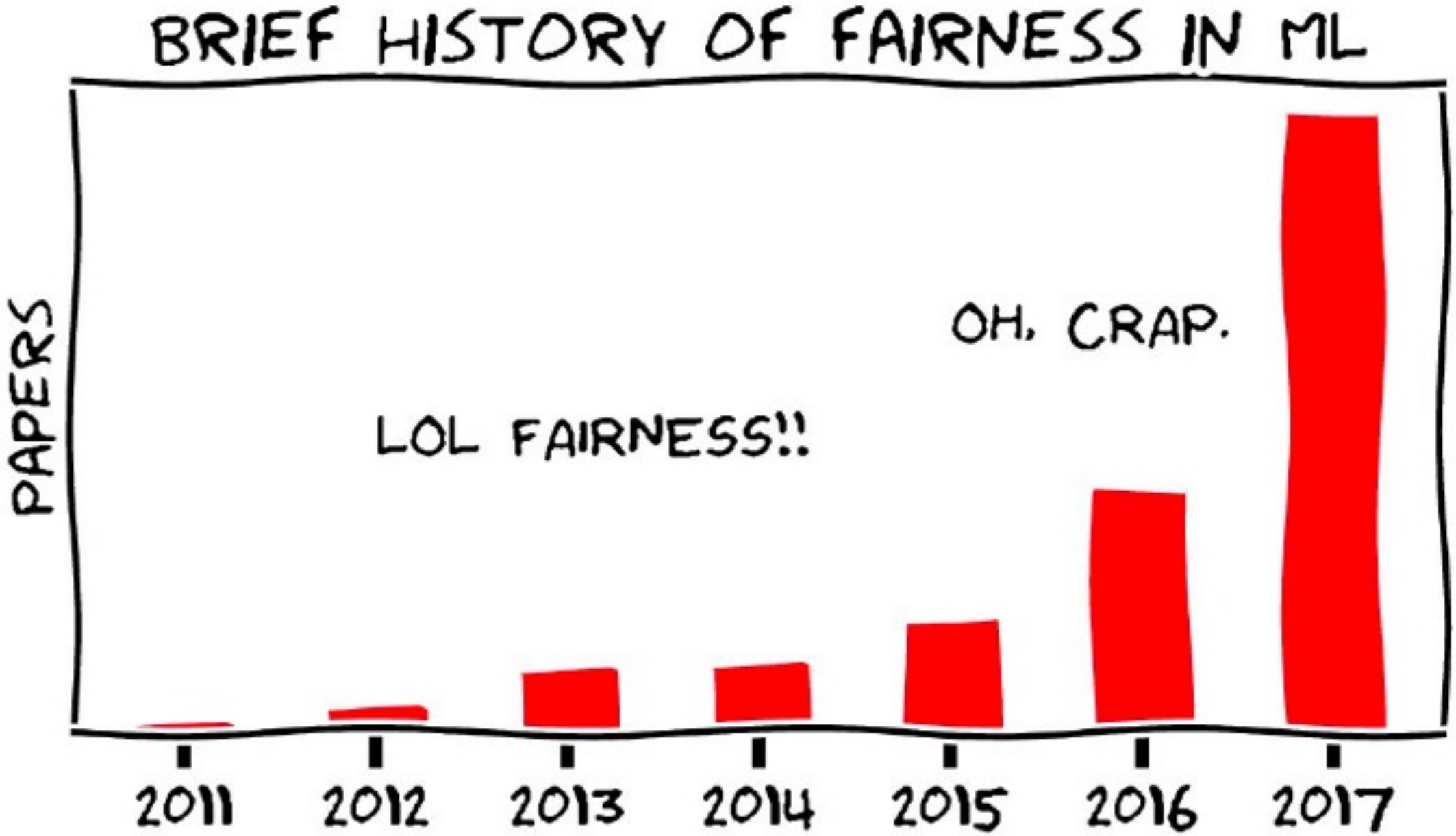
- Let the output of A be the average of users' ages
- Consider two extreme cases
  - If the size of the database is 1
  - If the size of the database is infinity

- 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 on Fairness

Cucumbers and Grapes Experiments

- https://youtu.be/-KSryJXDpZo


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, 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!

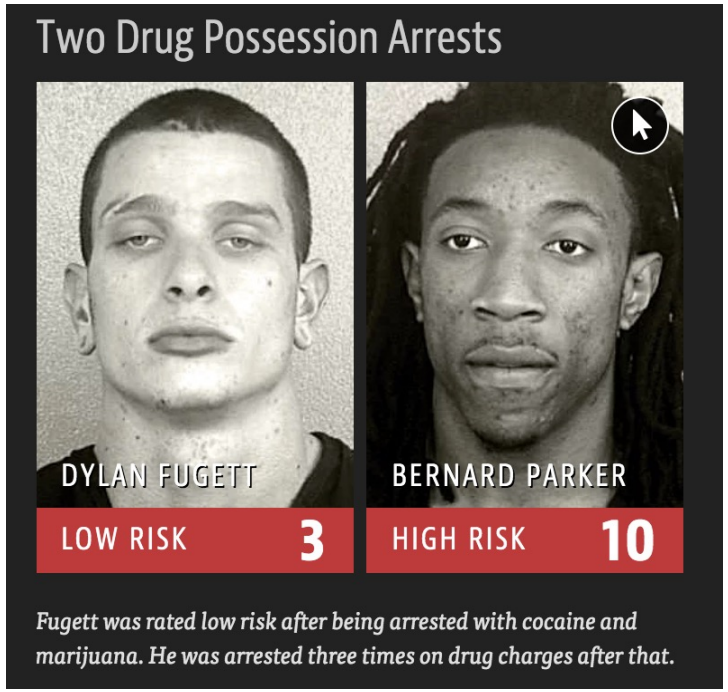
- One potential approach:
  - Minimize error subject to fairness constraints (Recall regularizations)

  - Several recent research and open-source libraries are done this way
    - Fairlearn: A toolkit for assessing and improving fairness in AI
    - GerryFair: Auditing and Learning for Subgroup Fairness
    - …

How should we define fairness?

Another Example: Probation Decisions

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



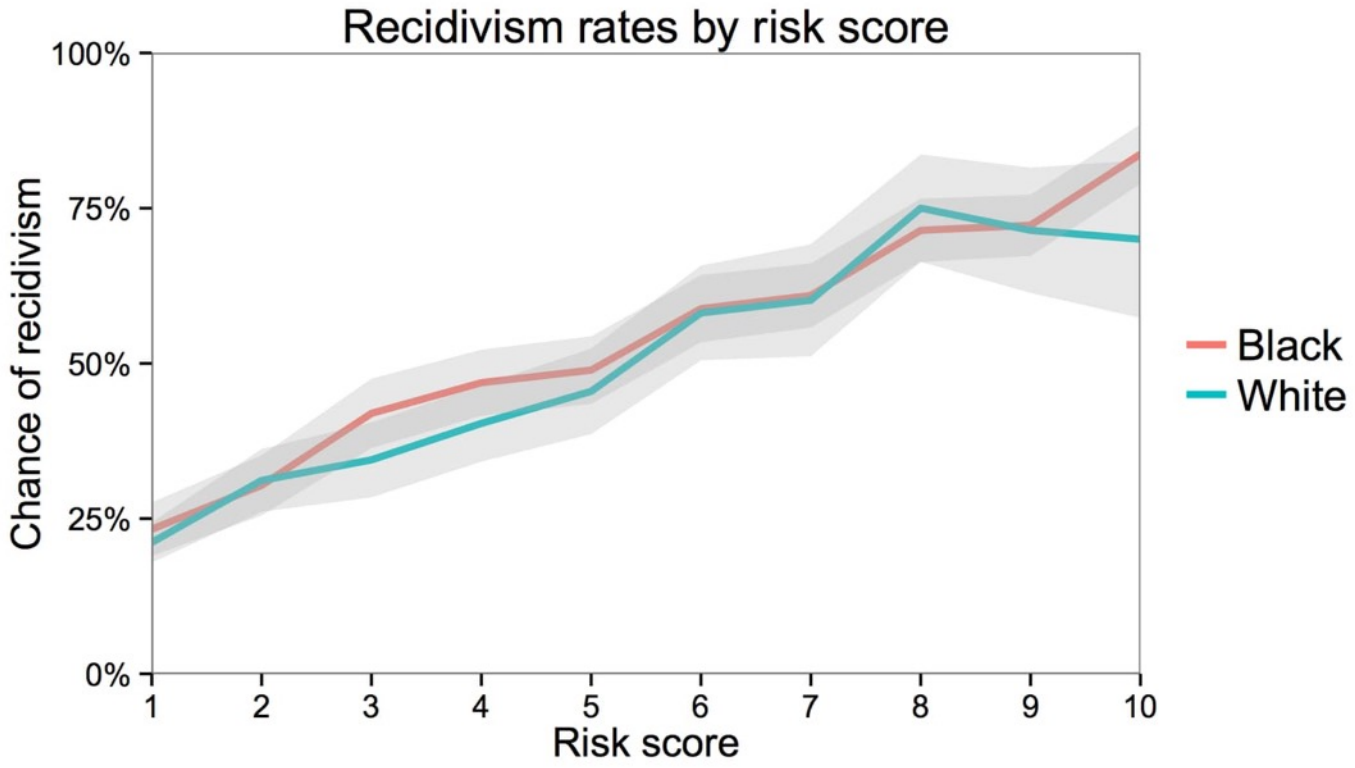
Controversy and Debates

- ProPublica (a non-profit institution)
  - COMPAS is not fair!

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Controversy and Debates

- Northpointe (company that develops COMPAS)
  - COMPAS is fair!



Impossibility Result [Kleinberg et al. 2017]

The above fairness conditions (together with similar variations) cannot be satisfied simultaneously, unless the predictor is perfect or the two groups are the same.

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

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

- Defendant: the probability that I'm incorrectly classified high-risk is independent of my race.
  - Equal False Positive Rate: $\frac{\mathrm{FP1}}{\mathrm{TN1}+\mathrm{FP1}} = \frac{\mathrm{FP2}}{\mathrm{TN2}+\mathrm{FP2}}$
- Defendant: the probability that I'm incorrectly classified as low-risk is independent of my race.
  - Equal False Negative Rate: $\frac{\mathrm{FN1}}{\mathrm{FN1}+\mathrm{TP1}} = \frac{\mathrm{FN2}}{\mathrm{FN2}+\mathrm{TP2}}$
- Decision-maker: the ratio of people who recidivated among the ones labeled high-risk is independent of race.
  - Equal predictive value: $\frac{\mathrm{TP1}}{\mathrm{TP1}+\mathrm{FP1}} = \frac{\mathrm{TP2}}{\mathrm{TP2}+\mathrm{FP2}}$

Impossibility Result [Kleignberg et al. 2017]

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

The Same Impossibility Results Applies to Other Sets of Fairness Definitions

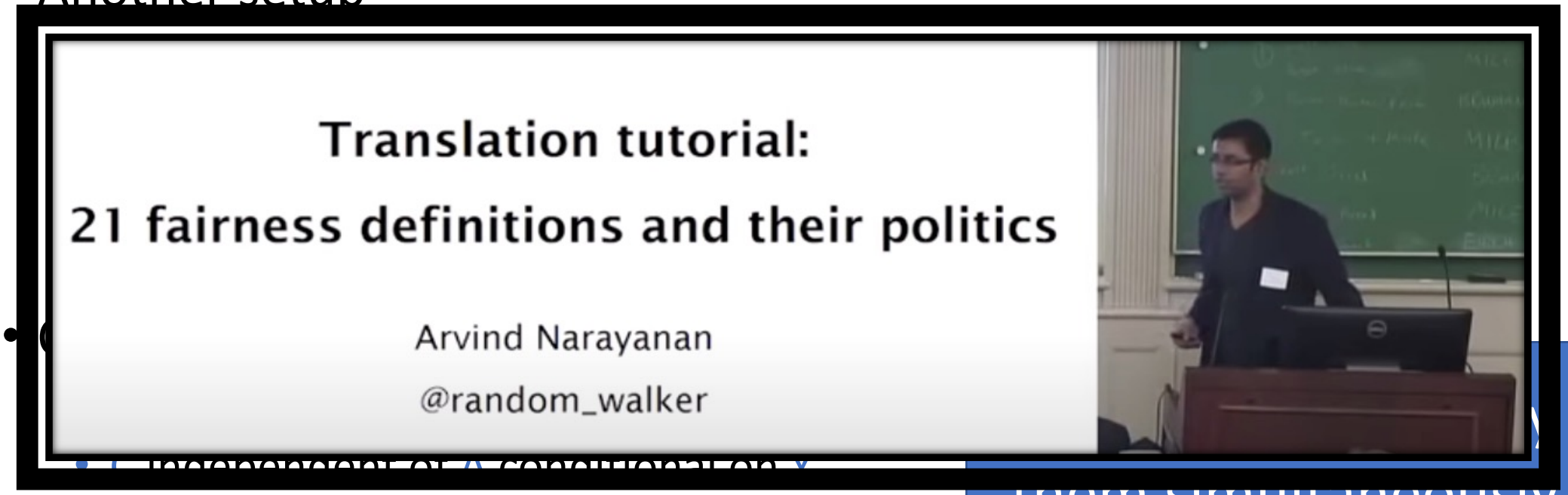
- Another setup
  - A: Sensitive attributes (e.g., race)
  - Y: True labels (e.g., commit a crime in the future)
  - C: Predictions (e.g., predictions of recidivism)

- Criteria:
  - C independent of A
  - C independent of A conditional on Y
  - Y independent of A conditional on C

Impossible to satisfy them simultaneously.

The Same Impossibility Results Applies to Other Sets of Fairness Definitions

- Another setup
- C
  - Y independent of A conditional on C



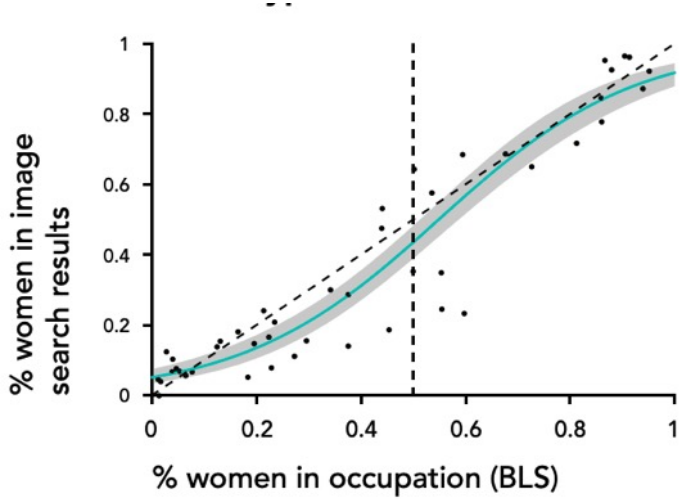
them simultaneously.

More Examples

[Kay et al., 2015]

Stereotype Mirroring and Exaggeration

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



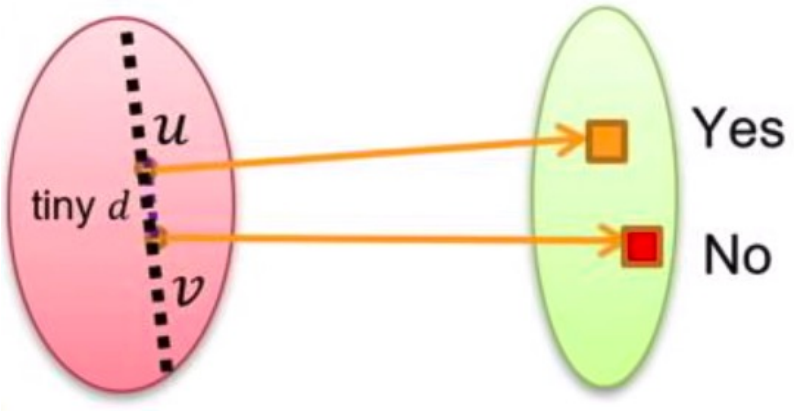
- Even when this is mirroring of the real statistics, are there other concerns?
  - Are we reinforcing the stereotypes?
  - Are we being “unfair” to disadvantage groups that are mistreated in the past?

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

Other Types of Fairness: Individual Fairness

- Similar people should be treated similarly

- Challenges
  - What do we mean by similar people
    - Need to define some kind of “distance” measure
  - What do we mean by being treated similarly
    - Decisions based on threshold won’t work
    - Need to impose some “smooth” notion
    - Randomization is often required



Other Types of Fairness: Counterfactual Fairness

- A decision is fair towards an individual if it gives the same predictions in
  - (a) the observed world and
  - (b) a world where the individual had always belonged to a different demographic group

I understood gender discrimination once I added "Mr." to my resume and landed a job

Woman Who Switched to Man's Name on Resume Goes From 0 to 70 Percent Response Rate

Other Types of Fairness: Procedural Fairness (Procedural Justice)

Take-Aways

- ML is a powerful tool to help extract patterns from data.
  - If you have data, ML might be able to help!
- However, ML may also be an amplifier of human biases
  - Biases could creep in through many stages of the ML life cycle, such as data, task definition, model choice, parameter tuning, …
- No silver bullet (yet)
  - **Being aware** of the issues is the important first step
  - ”Solving” the issues (if at all possible) requires communications among people in different disciplinaries

An Emerging Research Agenda on AI/ML + Humans/Society

- WashU Division of Computational and Data Sciences
  - A new PhD program hosted by CSE, Political Science, Social Work, Psychology and Brain Science

- MIT Institute for Data, Systems, and Society
- CMU Societal Computing
- Stanford Institute for Human-Centered Artificial Intelligence
- USC Center for AI in Society

- ACM FAT* (Fairness, Accountability, and Transparency)
- AAAI/ACM AIES (AI, Ethics, and Society)

Course Wrap-Up

Revisit Our Course Plan

- Foundations
  - What's machine learning
  - Feasibility of learning
  - Generalization
  - Linear models
  - Non-linear transformations
  - Overfitting and how to avoid it
    - Regularization
    - Validation
- Techniques
  - Decision tree
  - Ensemble learning
    - Bagging and random forest
    - Boosting and Adaboost
  - Nearest neighbors
  - Support vector machine
  - Neural networks
  - …

There are a lot more…