Fairness in AI

Instructor: Chien-Ju (CJ) Ho

Logistics: Project Milestone 1

- Project milestone 1
  - Due: Mar 22 (Fri)
  - Initial literature survey (At least 3~5 papers)
  - A plan on what you want to do for the remaining of the semester
    - Formalize your research question and approaches, e.g.,
      - Theory/simulation project: formalize your models
      - Data-analysis project: figure out where and how to get data and what you plan to do with it
      - Experiment/application project: have a prototype design and an evaluation plan
    - Include a timeline (weekly) on what you plan to do

  - Remaining milestones
    - Apr 5: Milestone 2 (Getting initial results)
    - Apr 8: Optional lecture (Eclipse): I'll be here to answer project-related questions
    - Apr 22/24: Presentation
    - Apr 26: Report

Logistics: Assignment 3

- Due: Mar 29 (Friday)

- An exploratory assignment
  - Use LLM as workers in crowdsourcing workflows
  - A replication of the assignment by CMU
  - Inspired by the required reading today

- Given this is the first time we have this assignment
  - You can do the assignment in groups

Logistics: Assignment 3

- Requirement:
  - Read one of the crowdsourcing workflow papers; summarize the discuss the workflow
    - See Table 1 of the required reading this Wednesday
  - Find N (# people in your assignment) use cases / examples of their workflow
    - Try to be a bit diverse
  - Compare the results of the following
    - Baseline: a single prompt to LLM
    - Workflow: Use LLM to replace workers in the workflow
  - Define your own metric to compare the results
  - General discussion
    - You can base your discussion on the required reading, disagree or agree with their points.
    - You can raise new points based on your observations.

Algorithmic Decision Making

More items to explore

Earth Balance Vegan Cheddar Flavor Squares, 6 oz.
4.5 2,498
$6.34

Louisville Vegan Jerky - Smoked Black Pepper, Vegetarian & Vegan-Friendly Jerky, 21 Grams of Non-GMO Soy Protein, 240 Calories…
4.5 1,077
5 offers from $9.75

Louisville Vegan Jerky - Maple Bacon, Vegetarian & Vegan-Friendly Jerky, 18 Grams of Non-GMO Soy Protein, 270 Calories Per Bag, Gluten-Free…
4.5 1,181
$7.99

Enjoy Life Semi Sweet Chocolate Mini Chips, 10 oz
5 1,599
22 offers from $3.92

Tofurky, Deli Slices, Peppered, 5.5 oz
4.5 824
$49.08

Justin's Dark Chocolate Peanut Butter Cups, 1.4 oz
5 1,267
4 offers from $6.21

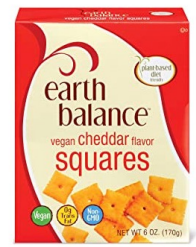
Customers who bought this item also bought

Gardein Plant-Based Chick'n Noodl' Soup, Vegan, 15 oz
4.0 1,497
#1 Best Seller in Packaged Beef Soups
$2.89

Gardein Plant-based Saus'ge Gumbo Soup, 15 oz
4.0 725
1 offer from $10.98

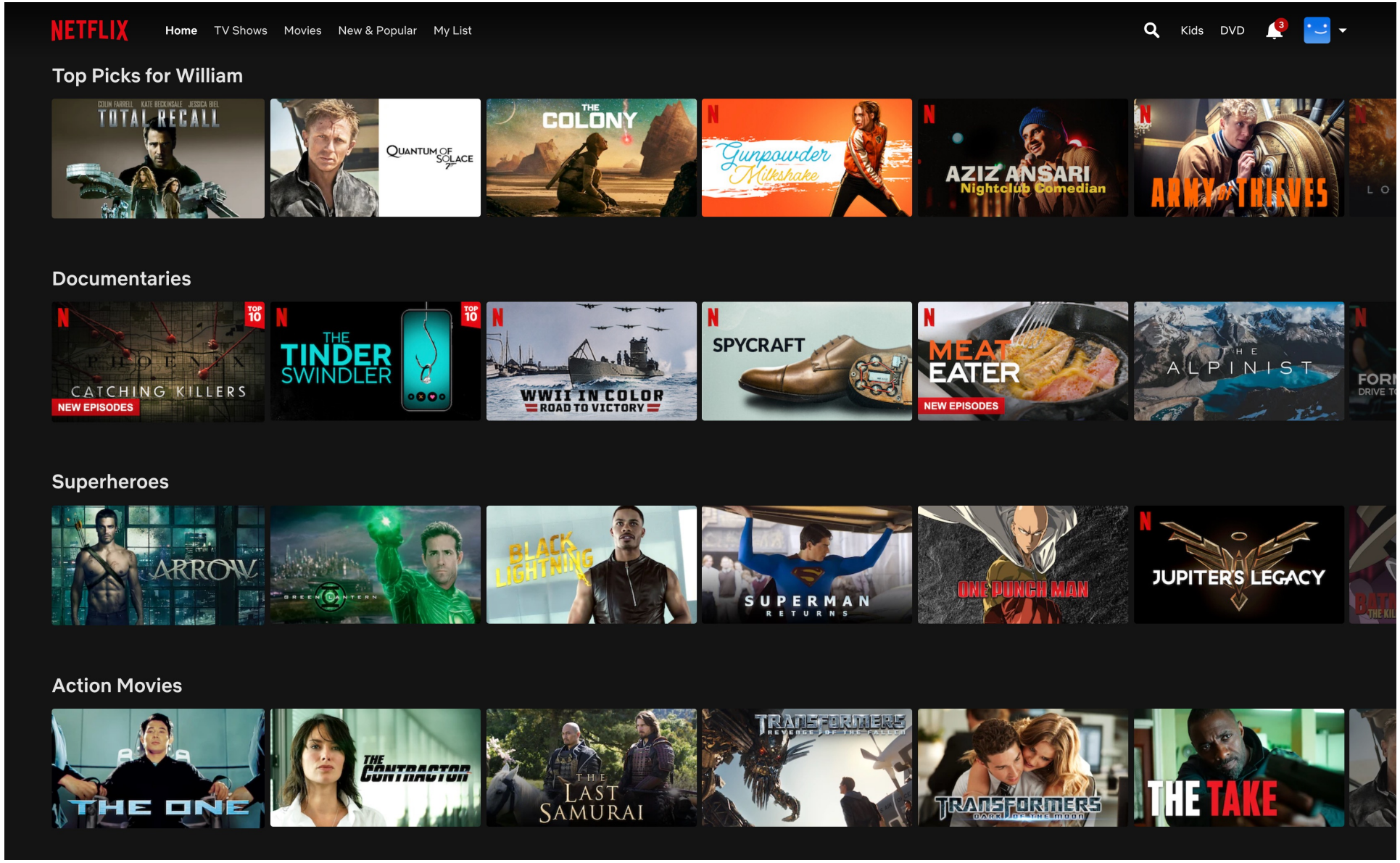
HIPPEAS Organic Chickpea Puffs + Vegan White Cheddar | 4 ounce | Vegan, Gluten-Free, Crunchy, Protein Snacks
4.5 3,959
9 offers from $2.25

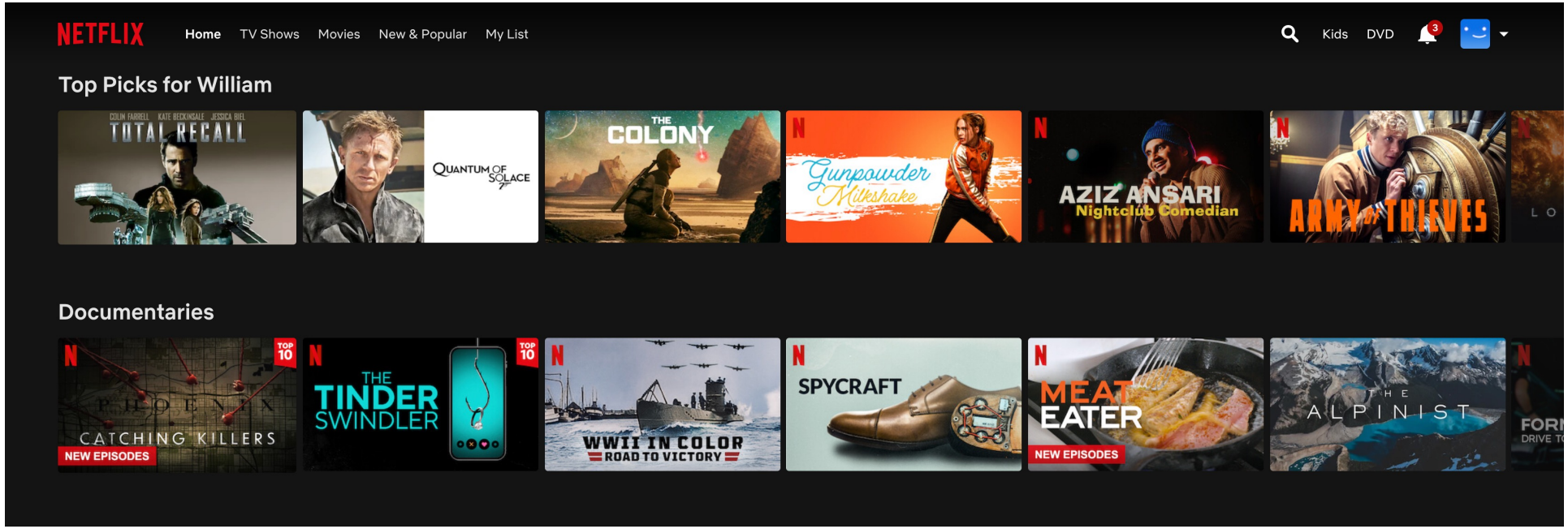
Justin's, Mini Dark Chocolate Peanut Butter Cups, 4.7 oz
4.5 1,722
$4.99



Earth Balance Vegan Cheddar Flavor Squares, 6 oz.
4.5 2,498
$6.34

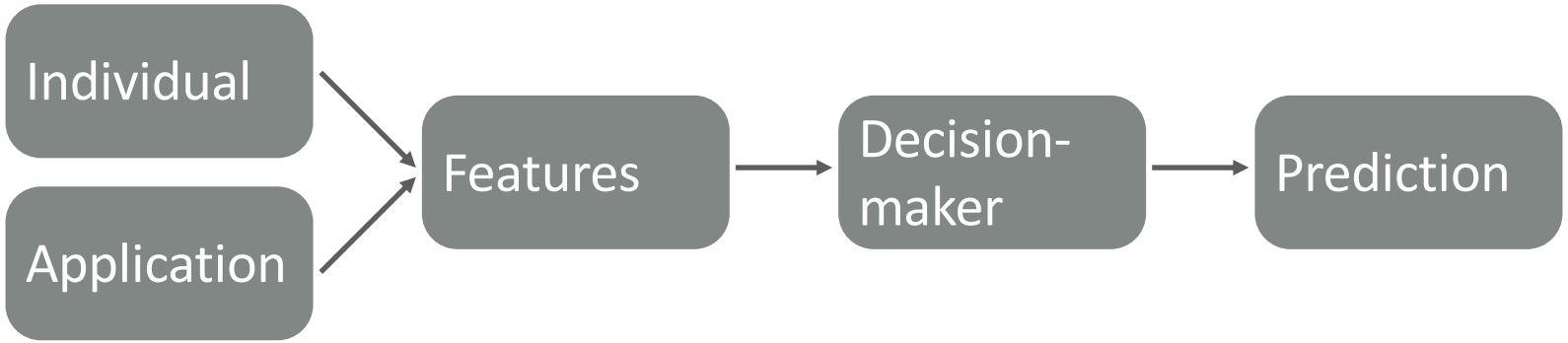
Enjoy Life Semi Sweet Chocolate Mini Chips, 10 oz
5 1,599
22 offers from $3.92


NETFLIX
Home
TV Shows
Movies
New & Popular
My List
Kids
DVD
Top Picks for William
TOTAL RECALL
QUANTUM OF SOLACE
THE COLONY
Gunpowder Milkshake
AZIZ ANSARI
Nightclub Comedian
ARMY OF THIEVES
Documentaries
CATCHING KILLERS
NEW EPISODES
THE TINDER SWINDLER
WWII IN COLOR
ROAD TO VICTORY
SPYCRAFT
MEAT EATER
NEW EPISODES
THE ALPINIST
Superheroes
ARROW
BLACK LIGHTNING
SUPERMAN RETURNS
ONE PUNCH MAN
JUPITER'S LEGACY
Action Movies
THE ONE
THE CONTRACTOR
THE LAST SAMURAI
TRANSFORMERS
REVENGE OF THE FALLEN
TRANSFORMERS
DARK OF THE MOON
THE TAKE



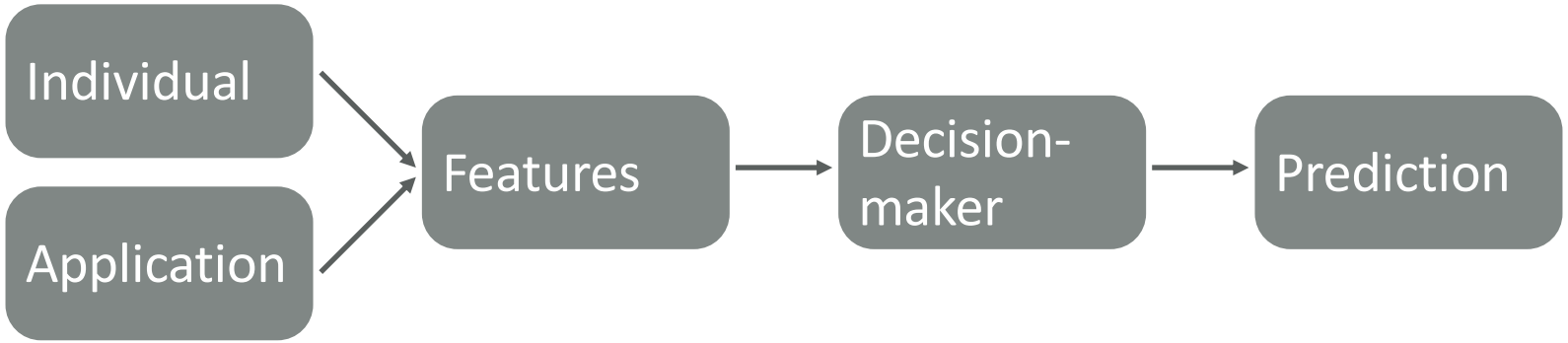
- Recommendation and ranking problems:
  - Take as inputs:
    - Features about you (based on your interactions with the system)
    - Features of a product (ice cream, movie, etc.)
  - Reduce them to a feature vector
  - Use that to estimate probability that you will like the product

Prediction Problems



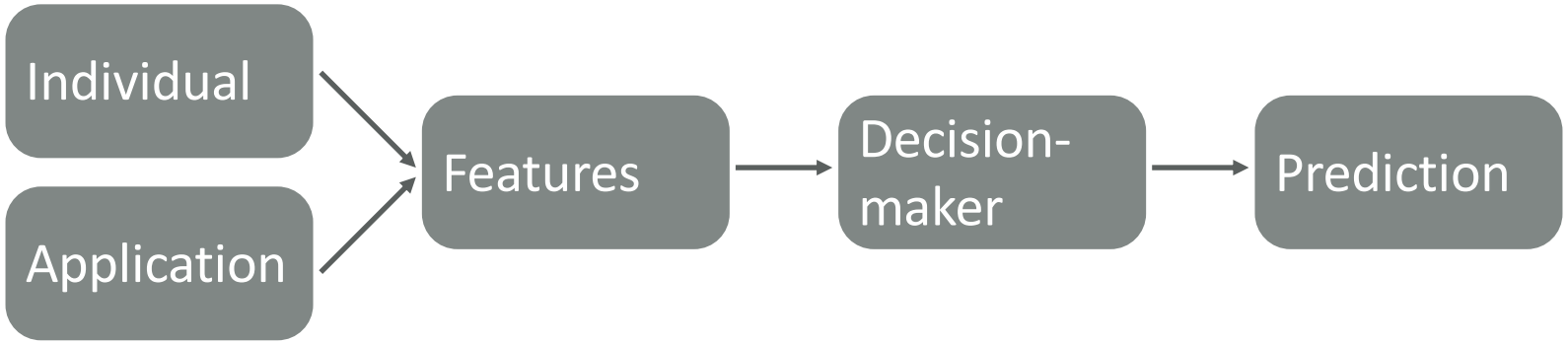
- Recommendation and ranking problems:
  - Take as inputs:
    - Features about you (based on your interactions with the system)
    - Features of a product (ice cream, movie, etc.)
  - Reduce them to a feature vector
  - Use that to estimate probability that you will like the product

Prediction Problems



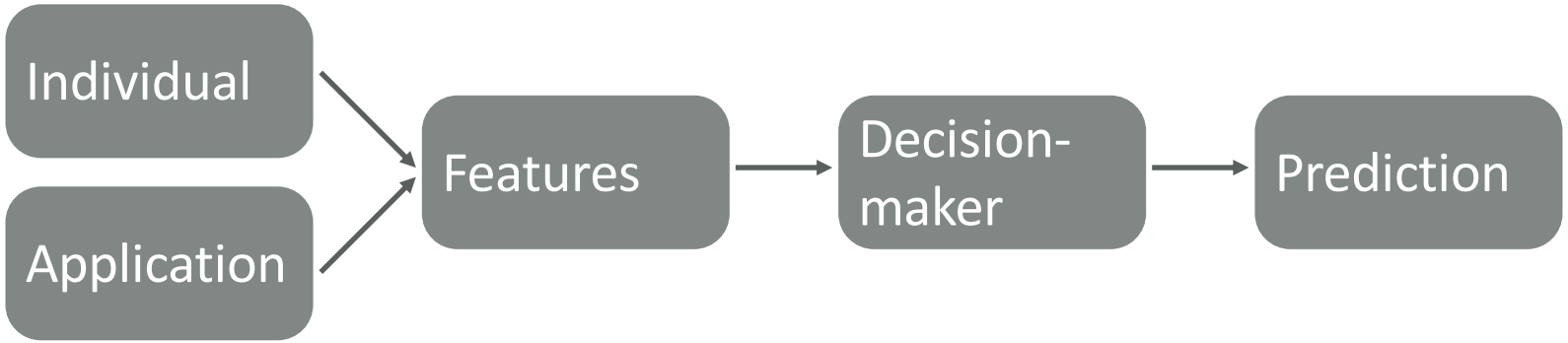
- College application problem:
  - Take as inputs:
    - Features about you (your background, K-12 experience, etc.)
    - Features of the school
  - Reduce them to a feature vector (application essay, SAT scores, etc.)
  - Use that to estimate probability that you will be a good fit

Prediction Problems



- Job application problem:
  - Take as inputs:
    - Features about you (your background, education, work experience, etc.)
    - Features of the company and the position
  - Reduce them to a feature vector (resume, recommendation letters, etc.)
  - Use that to estimate probability that you will be a good fit

Prediction Problems



- Example applications:
  - Online companies: Prediction of interest in product
  - Education: Prediction of success in college
  - Employment: Prediction of future productivity and growth in company
  - Credit: Prediction of successful loan repayment
  - Criminal Justice: Prediction of recividism (likelihood to reoffend)

Warm-Up Discussion

- AI/algorithms work well with online recommendation systems. It's natural to wonder whether we can apply them more broadly.

- What are the potential concerns of applying AI/algorithms in making decisions in hiring, college admissions, loan approval, etc?

- Are these concerns shared for online recommendations for movies/products? What are the differences?

Prediction Problems

- Recommendations for online companies (Amazon, Netflix, Yelp) are *individually* low-stake decisions
  - Note that they could be *collectively* high-stake decisions
  - For example, if Netflix consistently recommends movies featuring actors/actresses from subpopulation *X* over movies featuring those from subpopulation *Y*, then subpopulation *Y* can be significantly affected
- Recommendations for education, employment, credit, and criminal justice are *individually* high-stake decisions
  - A wrong recommendation can significantly affect the individual

Prediction Problems

- Recommendations for online companies (Amazon, Netflix, Yelp) are *individually* low-stake decisions
  - Note that they could be *collectively* high-stake decisions
  - For e… …ng acto… …ose from subp… …cted
- Recommendations for education, employment, credit, and criminal justice are *individually* high-stake decisions
  - A wrong recommendation can significantly affect the individual

Risks of bias in both human and algorithmic decision-makers

Human Bias

- Resume call-back rates:
  - Bertrand and Mullainathan 2004
  - Created two piles of fictitious resumes, identical between piles
    - Pile A: very White sounding names (e.g., Emily Walsh, Greg Baker)
    - Pile B: very Black sounding names (e.g., Lakisha Washington, Jamal Jones)
    - Sent to real employers in Boston and Chicago looking to hire

Human Bias

- Resume call-back rates:
  - Bertrand and Mullainathan 2004
  - Created two piles of fictitious resumes, identical between piles
    - Pile A: very White sounding names (e.g., Emily Walsh, Greg Baker)
    - Pile B: very Black sounding names (e.g., Lakisha Washington, Jamal Jones)
    - Sent to real employers in Boston and Chicago looking to hire
  - Results:
    - White names receive 50% more callbacks for interviews
    - Higher quality resume benefits White names more than Black names
      - For White names, a higher quality resume elicits 30% more callbacks
      - For Black names, a higher quality resume elicits 9% more callbacks

Human B

- Resume call
  - Bertrand a
  - Created tw
    - Pile A:
    - Pile B:
    - Sent to
  - Results:
    - White
    - Higher
      - For
      - For

s
g Baker)
n, Jamal Jones)
re

ack names
nore callbacks
re callbacks

Project Implicit®

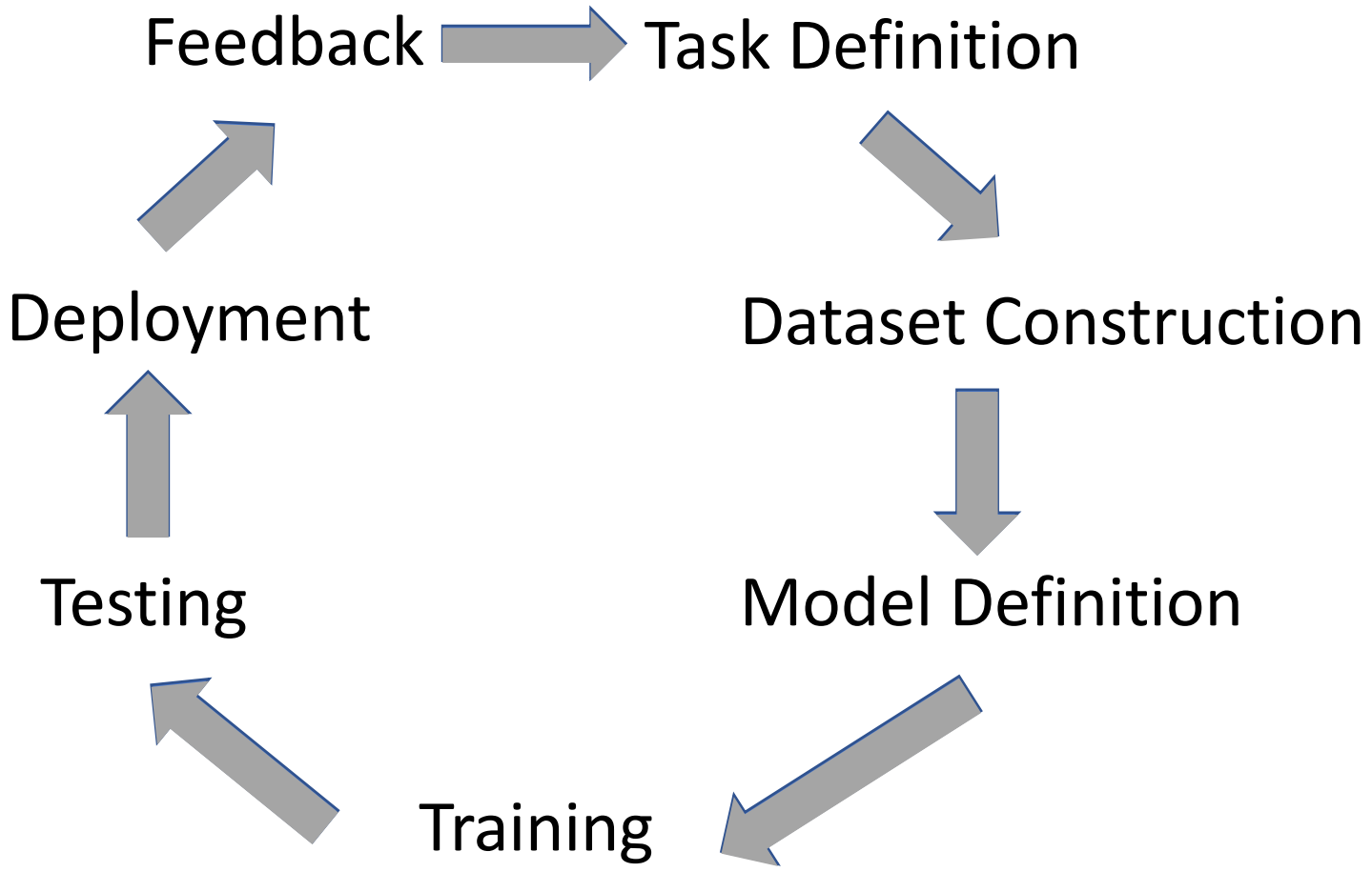
LOG IN | TAKE A TEST | ABOUT US | EDUCATION | BLOG | HELP | CONTACT US | DONATE

Race IAT	***Race*** **('Black - White' IAT).** This IAT requires the ability to distinguish faces of European and African origin. It indicates that most Americans have an automatic preference for white over black.
Arab-Muslim IAT	***Arab-Muslim*** **('Arab Muslim - Other People' IAT).** This IAT requires the ability to distinguish names that are likely to belong to Arab-Muslims versus people of other nationalities or religions.
Weight IAT	***Weight*** **('Fat - Thin' IAT).** This IAT requires the ability to distinguish faces of people who are obese and people who are thin. It often reveals an automatic preference for thin people relative to fat people.
Gender-Science IAT	***Gender - Science.*** This IAT often reveals a relative link between liberal arts and females and between science and males.
Asian IAT	***Asian American*** **('Asian - European American' IAT).** This IAT requires the ability to recognize White and Asian-American faces, and images of places that are either American or Foreign in origin.
Disability IAT	***Disability*** **('Disabled - Abled' IAT).** This IAT requires the ability to recognize symbols representing abled and disabled individuals.
Weapons IAT	***Weapons*** **('Weapons - Harmless Objects' IAT).** This IAT requires the ability to recognize White and Black faces, and images of weapons or harmless objects.
Gender-Career IAT	***Gender - Career.*** This IAT often reveals a relative link between family and females and between career and males.
Skin-tone IAT	***Skin-tone*** **('Light Skin - Dark Skin' IAT).** This IAT requires the ability to recognize light and dark-skinned faces. It often reveals an automatic preference for light-skin relative to dark-skin.
Transgender IAT	***Transgender*** **('Transgender People – Cisgender People' IAT).** This IAT requires the ability to distinguish photos of transgender celebrity faces from photos of cisgender celebrity faces.
Sexuality IAT	***Sexuality*** **('Gay - Straight' IAT).** This IAT requires the ability to distinguish words and symbols representing gay and straight people. It often reveals an automatic preference for straight relative to gay people.
Age IAT	***Age*** **('Young - Old' IAT).** This IAT requires the ability to distinguish old from young faces. This test often indicates that Americans have automatic preference for young over old.
Religion IAT	***Religion*** **('Religions' IAT).** This IAT requires some familiarity with religious terms from various world religions.
Presidents IAT	***Presidents*** **('Presidential Popularity' IAT).** This IAT requires the ability to recognize photos of Joseph Biden and one or more previous presidents.

Algorithmic Bias

Algorithms do not have an incentive to be biased.
So, how does an algorithm become biased?

Algorithmic Bias



Algorithmic Bias

- Bias can creep into algorithms through
  - Choice of features, labels, objective function, training data, …
  - … and humans are responsible for generating, deciding, and fine-tuning many of those choices

Algorithmic Bias

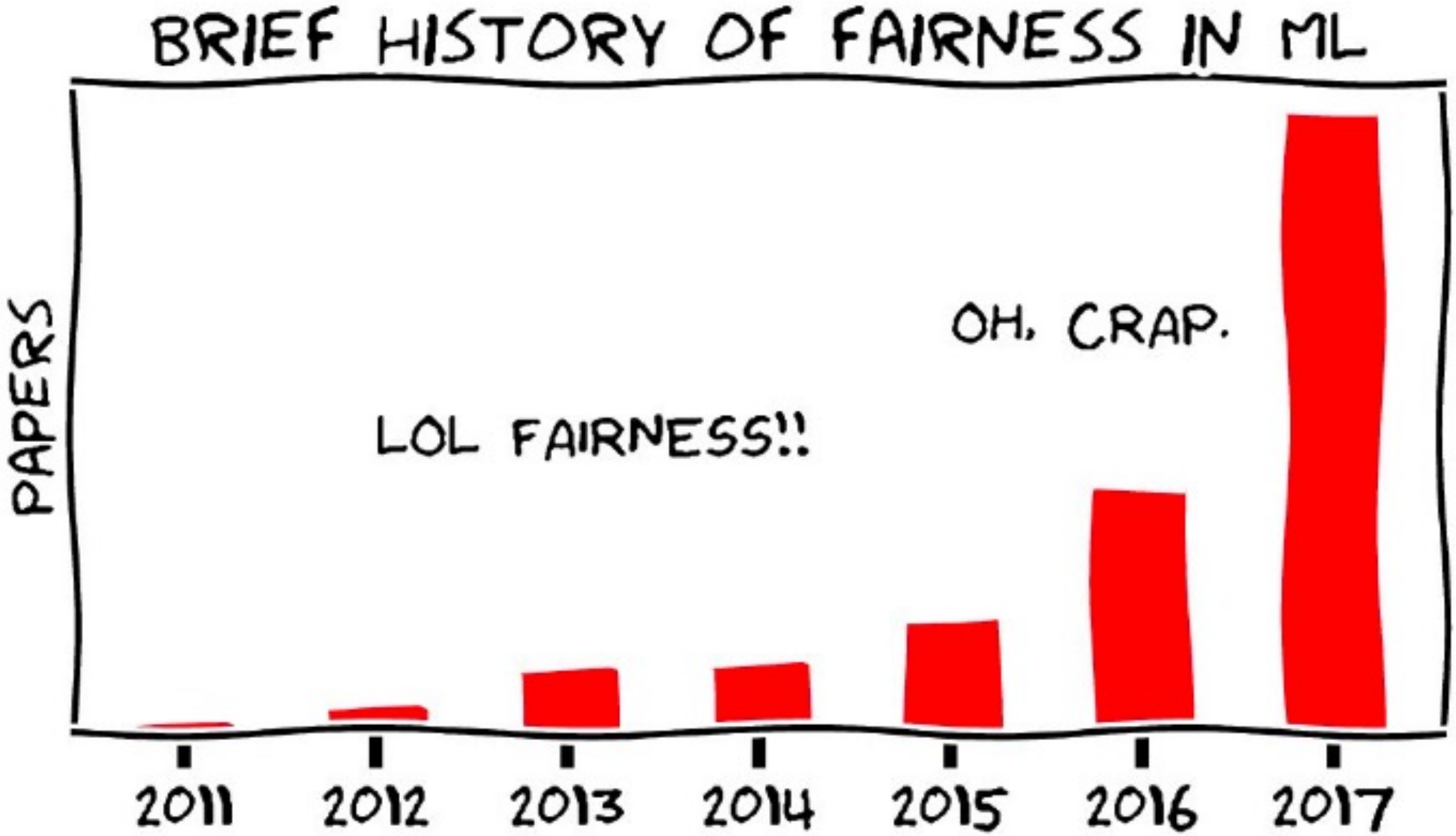
- Bias can creep into algorithms through
  - Choice of features, labels, objective function, training data, …
  - … and humans are responsible for generating, deciding, and fine-tuning many of those choices

- Examples:
  - Features: Using protected features or proxies
  - Labels: Using biased decisions as ground truth
  - Objective functions: Using biased functions (e.g., a function that has good correlation rates with one subgroup but not another)
  - Training data: Using non-representative training data

Algorithmic 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!

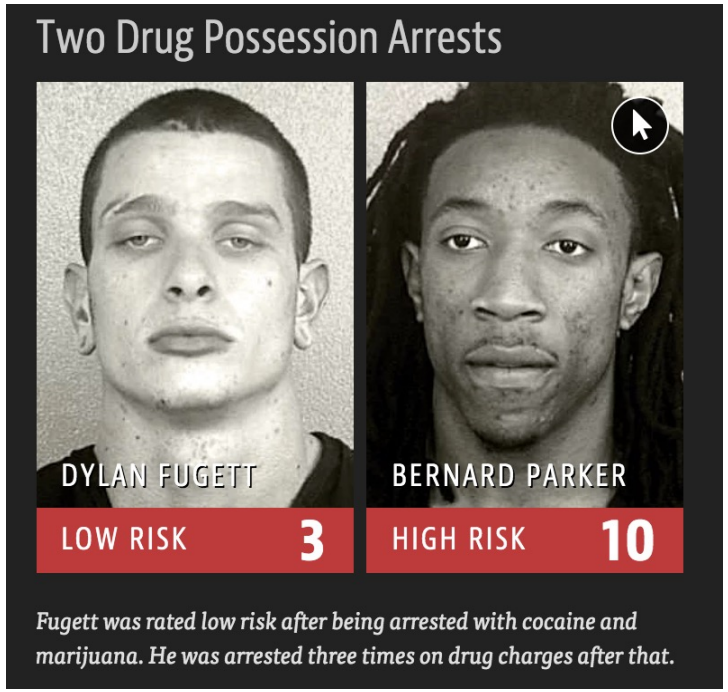
- One potential approach:
  - Minimize error subject to fairness constraints (similar to regularizations)

  - 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.



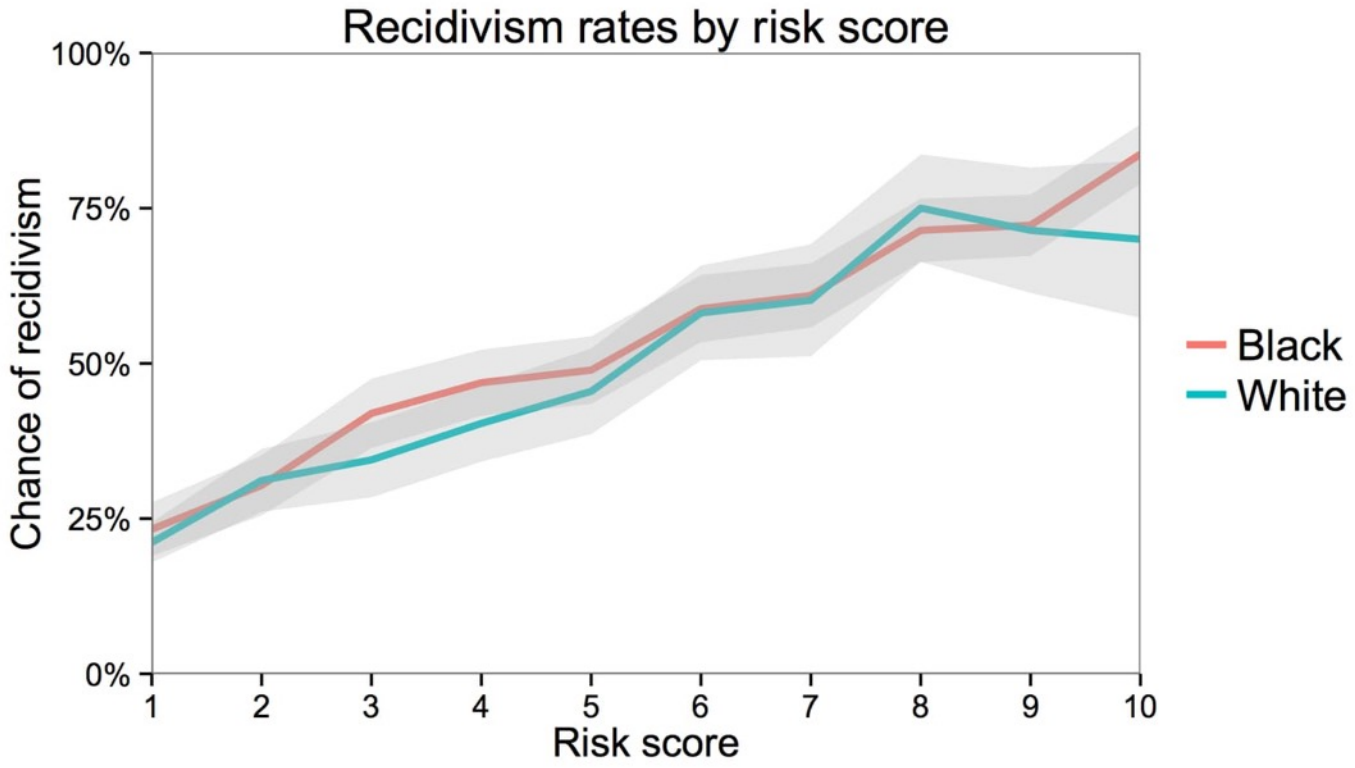
Controversy and Debates

- 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.

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

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

- 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

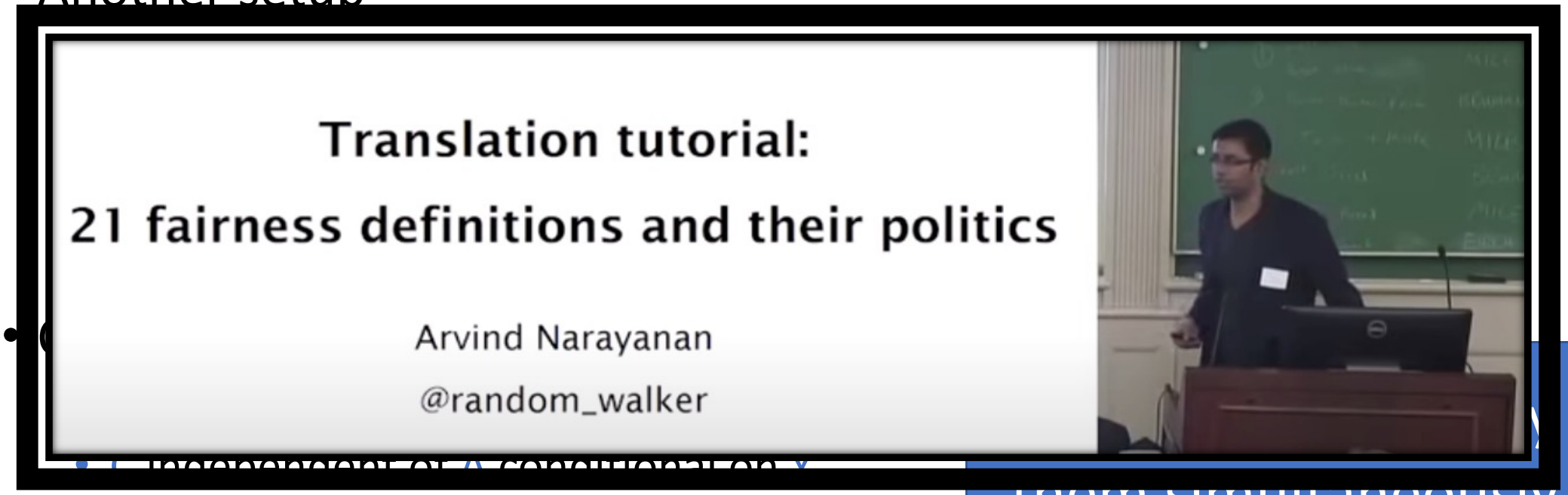
- Another setup
  - A: Sensitive attributes (e.g., race)
  - Y: True labels (e.g., commit a crime in the future)
  - C: Predictions (e.g., predictions of recidivism)

- 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



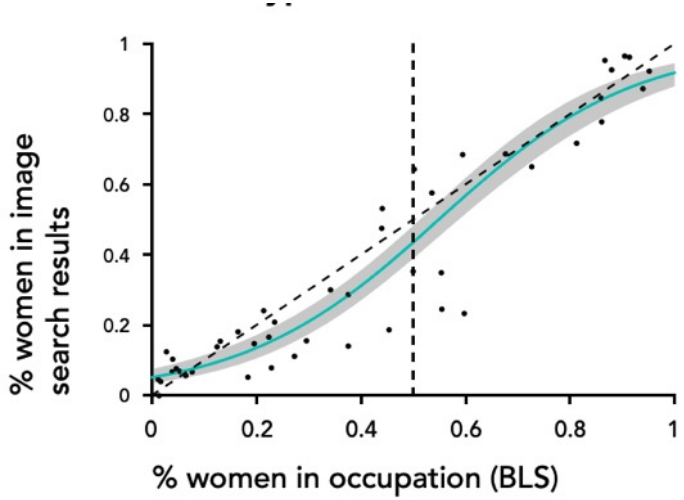
them simultaneously.

More Examples

[Kay et al., 2015]

Stereotype Mirroring and Exaggeration

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



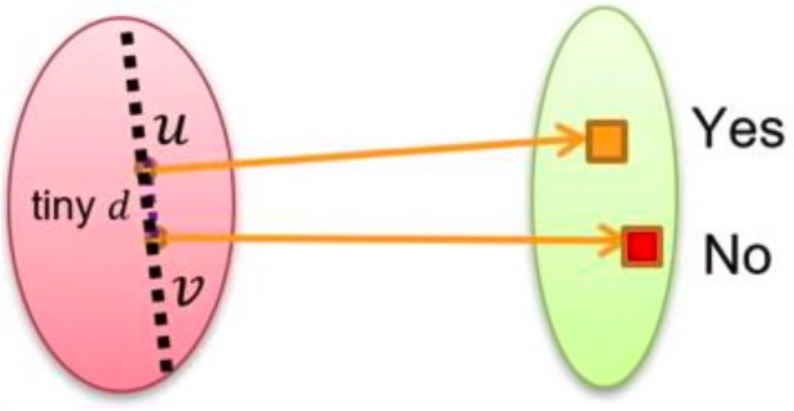
- Even when this is mirroring of the real statistics, are there other concerns?
  - Are we reinforcing the stereotypes?
  - Are we being “unfair” to disadvantage groups that are mistreated in the past?

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)

Discussion

- What does fairness mean for you?
  - You can discuss this in the context of recidivism prediction as in the required reading or other contexts (hiring, college admission, etc)

- Based on your fairness criteria and your context, think about examples that unfairness arises.

- Is there a way that we can take interventions to promote fairness? Does that lead to potential negative unintended consequences?

College Admission Example

From Jon Kleinberg's lecture:
https://www.youtube.com/watch?v=wfMlu4x4bnI

College Admission Example

- Applicants and feature vectors:
  - Applicants are described by Boolean variables $x = (x_1, x_2)$
  - Function f describes success $f(x)$ of applicant with feature x
  - Plan: Sort by f-value, admit top r fraction

College Admission Example

- Applicants and feature vectors:
  - Applicants are described by Boolean variables $x = (x_1, x_2)$
  - Function f describes success $f(x)$ of applicant with feature x
  - Plan: Sort by f-value, admit top r fraction
- Group membership:
  - Applicants belong to either *advantaged* group A or *disadvantaged* group D
  - Feature x of applicants do not include group membership
  - Function f is independent of group membership: $f(x, A) = f(x, D) = f(x)$

College Admission Example

- Applicants and feature vectors:
  - Applicants are described by Boolean variables $x = (x_1, x_2)$
  - Function f describes success $f(x)$ of applicant with feature x
  - Plan: Sort by f-value, admit top r fraction
- Group membership:
  - Applicants belong to either *advantaged* group A or *disadvantaged* group D
  - Feature x of applicants do not include group membership
  - Function f is independent of group membership: $f(x, A) = f(x, D) = f(x)$

  - $w(x, g)$ = fraction of population with features x and group g
  - Disadvantage condition: If $f(x) > f(x')$, then $\frac{w(x, A)}{w(x, D)} > \frac{w(x', A)}{w(x', D)}$

College Admission Example

- Applicants are equally divided between both groups g
- Applicants from A have $x_i = 1$ with probability 2/3
- Applicants from D have $x_i = 1$ with probability 1/3
- $f(x) = x_1$ AND x_2
- At admission rate of 5/18,
  all admitted students have f-value of 1

x_1	x_2	g	f	w
1	1	D	1	1/18
1	1	A	1	4/18
1	0	D	0	2/18
1	0	A	0	2/18
0	1	D	0	2/18
0	1	A	0	2/18
0	0	D	0	4/18
0	0	A	0	1/18

College Admission Example

- Applicants are equally divided between both groups *g*
- Applicants from *A* have $x_1 = 1$ with proba
- Appli
  proba
- $f(x) = x_1$ AND x_2
- At admission rate of 5/18,
  all admitted students have *f*-value of 1

x_1	x_2	*g*	*f*	*w*
1	1	*D*	1	1/18
1	1	*A*	1	4/18
1	0	*D*	0	2/18
1	0	*A*	0	2/18
0	1	*D*	0	2/18
0	1	*A*	0	2/18
0	0	*D*	0	4/18
0	0	*A*	0	1/18

Only 1/5 of admitted students are from group *D*,
while 4/5 of admitted students are from group *A*.

Is this okay? Any thoughts?

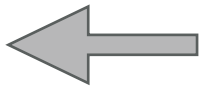
College Admission Example

- Now, suppose we only use x_1 because:
  - x_2 is a protected feature and should not be used
  - Collecting x_2 is too expensive
  - Interpreting or understanding x_2 is too computationally or cognitively complex

x_1	x_2	g	f	w
1	1	*D*	1	1/18
1	1	*A*	1	4/18
1	0	*D*	0	2/18
1	0	*A*	0	2/18
0	1	*D*	0	2/18
0	1	*A*	0	2/18
0	0	*D*	0	4/18
0	0	*A*	0	1/18

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0	*any*	*any*	0	9/18

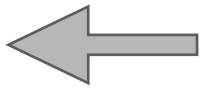


- At the same admission rate of 5/18:
  - average *f*-value is
  - fraction from group *D* is

x_1	x_2	g	f	w
1	1	*D*	1	1/18
1	1	*A*	1	4/18
1	0	*D*	0	2/18
1	0	*A*	0	2/18
0	1	*D*	0	2/18
0	1	*A*	0	2/18
0	0	*D*	0	4/18
0	0	*A*	0	1/18

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0	*any*	*any*	0	9/18

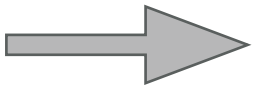


- At the same admission rate of 5/18:
  - average *f*-value is 5/9 (not 1 like before)
  - fraction from group *D* is 1/3 (not 1/5 like before)
- Improved equity, lost some efficiency

x_1	x_2	g	f	w
1	1	*D*	1	1/18
1	1	*A*	1	4/18
1	0	*D*	0	2/18
1	0	*A*	0	2/18
0	1	*D*	0	2/18
0	1	*A*	0	2/18
0	0	*D*	0	4/18
0	0	*A*	0	1/18

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0	*any*	*any*	0	9/18

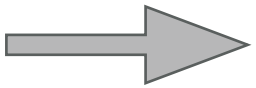


x_1	x_2	g	avg f	w
1	*any*	*A*	2/3	6/18
1	*any*	*D*	1/3	3/18
0	*any*	*A*	0	3/18
0	*any*	*D*	0	6/18

- At the same admission rate of 5/18:
  - average *f*-value is
  - fraction from group *D* is

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0	*any*	*any*	0	9/18

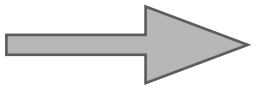


x_1	x_2	g	avg f	w
1	*any*	*A*	2/3	6/18
1	*any*	*D*	1/3	3/18
0	*any*	*A*	0	3/18
0	*any*	*D*	0	6/18

- At the same admission rate of 5/18:
  - average *f*-value is 2/3 (not 5/9 like before)
  - fraction from group *D* is 0 (not 1/3 like before)
- Improved efficiency, lost equity
  - lost more than if both x_1 and x_2 were used!

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0	*any*	*any*	0	9/18



x_1	x_2	g	avg f	w
1	*any*	*A*	2/3	6/18
1	*any*	*D*	1/3	3/18
0	*any*	*A*	0	3/18
0	*any*	*D*	0	6/18

- At the same admission rate of 5/18:
  - average *f*-value is 2/3 (not 5/9 like before)
  - fraction from group *D* is 0 (not 1/3 like before)
- Improved efficiency, lost equity
  - lost more than if both x_1 and x_2 were used!
- Not using x_2 also creates an incentive to use group membership in a way that hurts group *D*

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0				

x_1	x_2	g	avg f	w
1	*any*	*A*	2/3	6/18
1	*any*	*D*	1/3	3/18
			0	3/18
			0	6/18

Paradoxically, forcing some feature to not be used can incentivize the use of some other proxy feature that will harm a disadvantaged group even more!

- At
  - average f-value is 2/3 (not 5/9 like before)
  - fraction from group *D* is 0 (not 1/3 like before)
- Improved efficiency, lost equity
  - lost more than if both x_1 and x_2 were used!
- Not using x_2 also creates an incentive to use group membership in a way that hurts group *D*

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0				

x_1	x_2	g	avg f	w
1	*any*	*A*	2/3	6/18
1	*any*	*D*	1/3	3/18
			0	3/18
			0	6/18

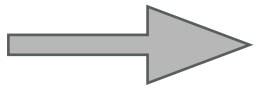
Paradoxically, forcing some feature to not be used
can incentivize the use of some other proxy feature
that will harm a disadvantaged group even more!

- At
  - average f-value is 2/3 (not 5/9 like before)
  -
- Im
  - lost more than if both x_1 and x_2 were used!
- Not using x_2 also creates an incentive to use group membership in a way that hurts group *D*

Must there always be a trade off between equity and efficiency?
Is it possible to improve both?

College Admission Example

x_1	x_2	g	avg f	w
1	*any*	*any*	5/9	9/18
0	*any*	*any*	0	9/18



x_1	x_2	g	avg f	w
1	1	*D*	1	1/18
1	*any*	*any*	1/2	8/18
0	*any*	*any*	0	9/18

- Assume that we spend additional effort to get x_2 from applicants in D
- At the same admission rate of 5/18:
  - average *f*-value is 3/5 (not 5/9 like before)
  - fraction from group *D* is 2/5 (not 1/3 like before)
- Improved efficiency and equity!

College Admission Example

x_1	x_2	g	avg f	w
1				
0				

x_1	x_2	g	avg f	w
				1/18
			/2	8/18
				9/18

(Informal) Theorem [*Kleinberg and Mullainathan 2019*]

For every simplification of a function f that satisfies the disadvantage condition:

(1) There is always an approximate function that strictly improves it in both efficiency and equity

(2) If the simplification does not use group membership, then adding group membership increases efficiency and reduces equity

- As
  to
- At
  - average f value is 3/5 (not 3/5 like before)
  - fraction from group *D* is 2/5 (not 1/3 like before)
- Improved efficiency and equity!

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)