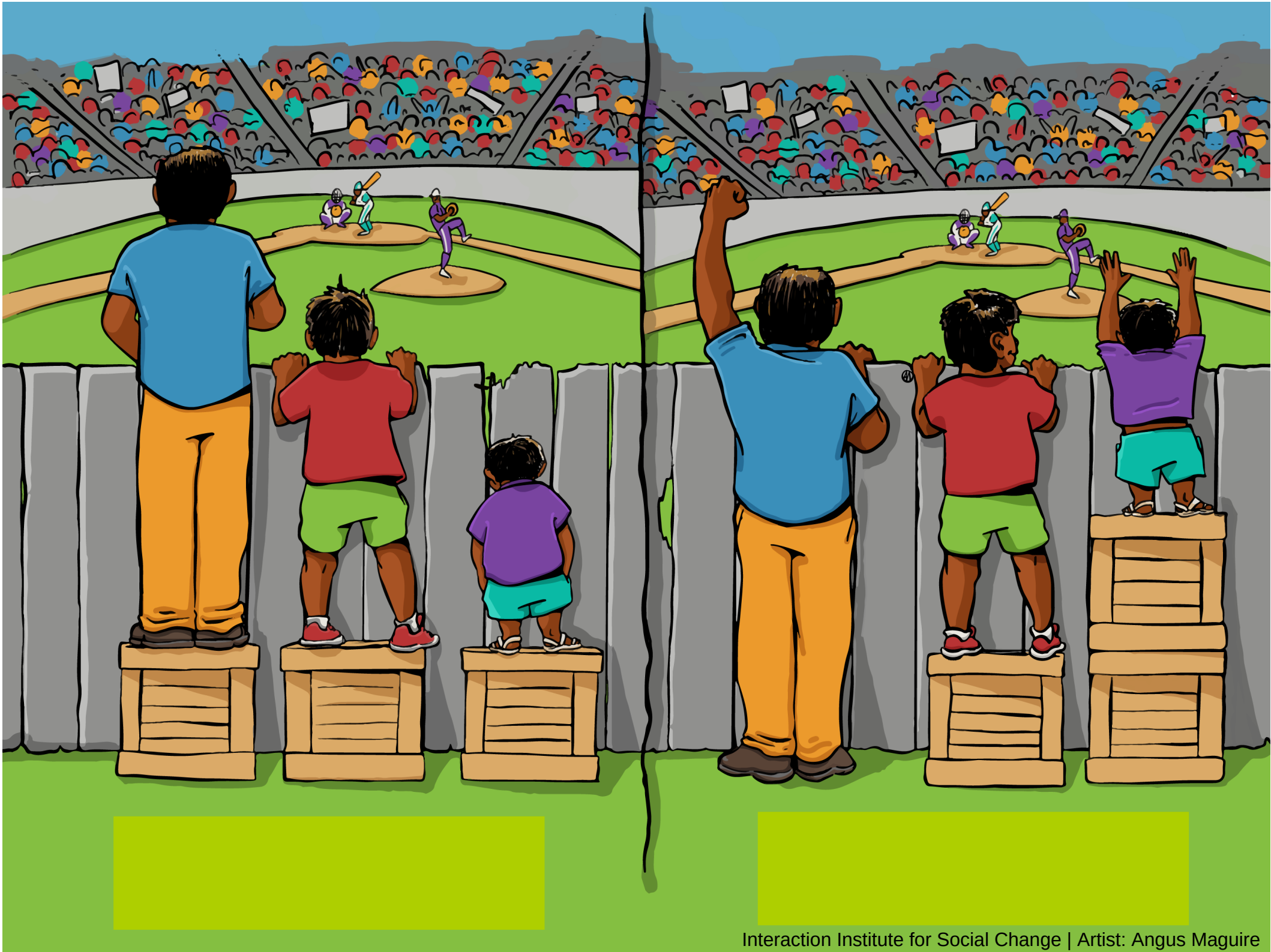
Measuring algorithmic fairness

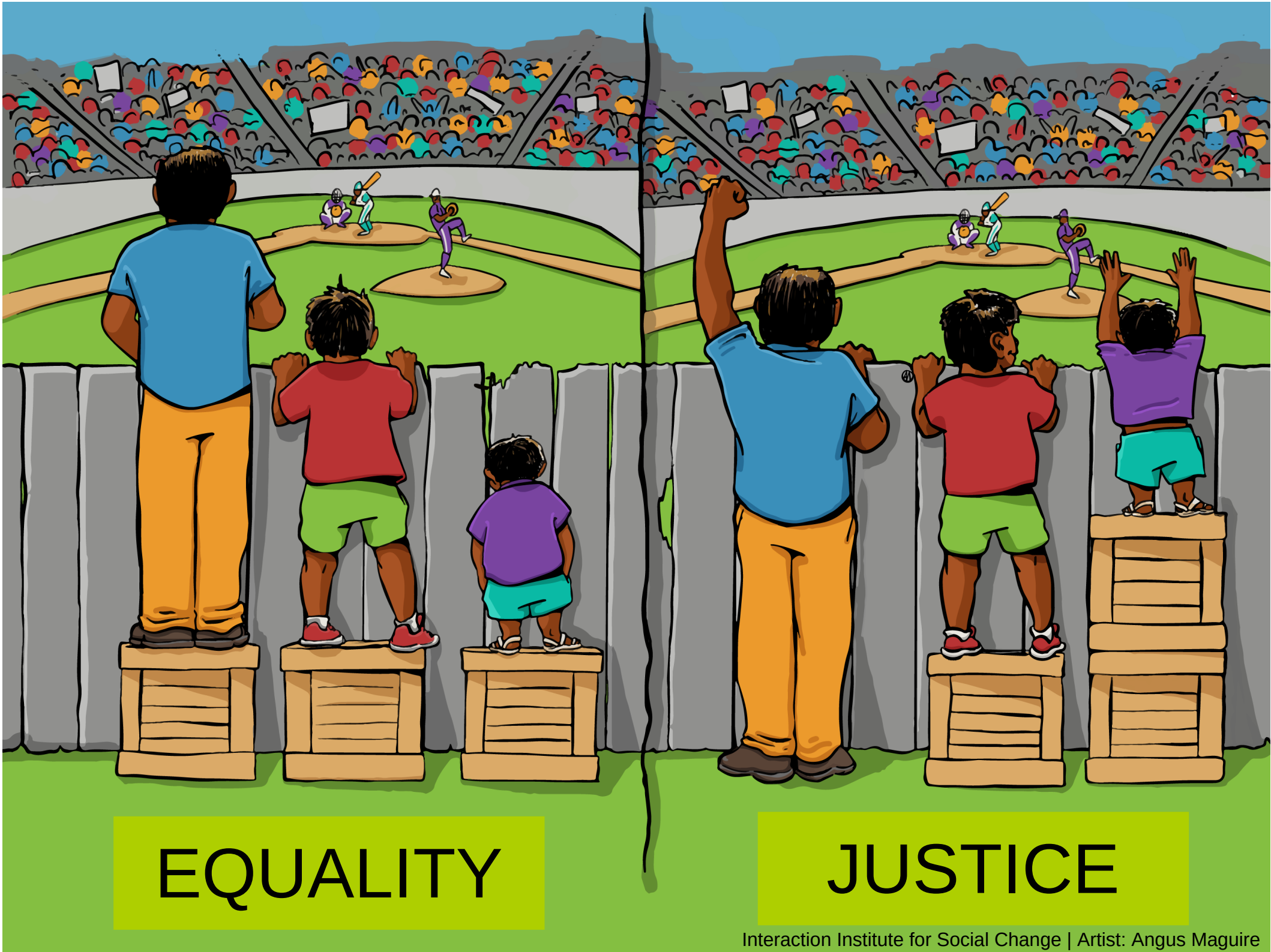
Indrė Žliobaitė
University of Helsinki

August 28, 2019
Lyngby

What is fairness?



Interaction Institute for Social Change | Artist: Angus Maguire



Interaction Institute for Social Change | Artist: Angus Maguire

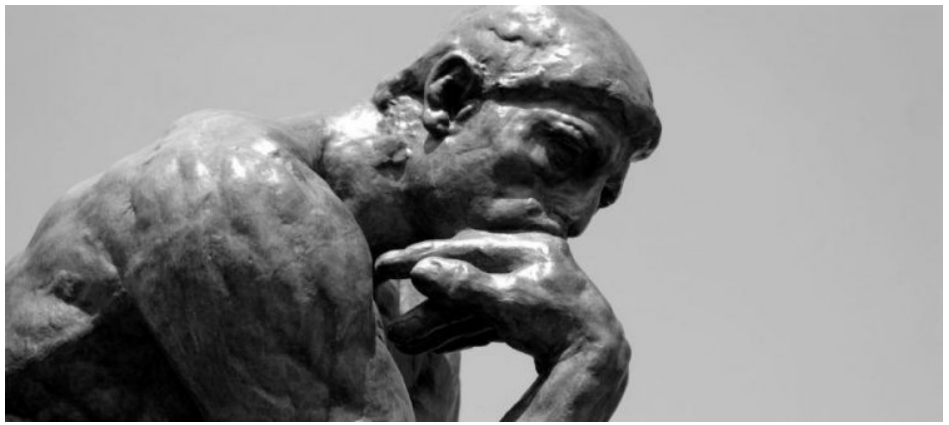
Philosophy → Morality → Ethics → Justice

Principles of behavior

Uff …
balancing
conflicting interests?

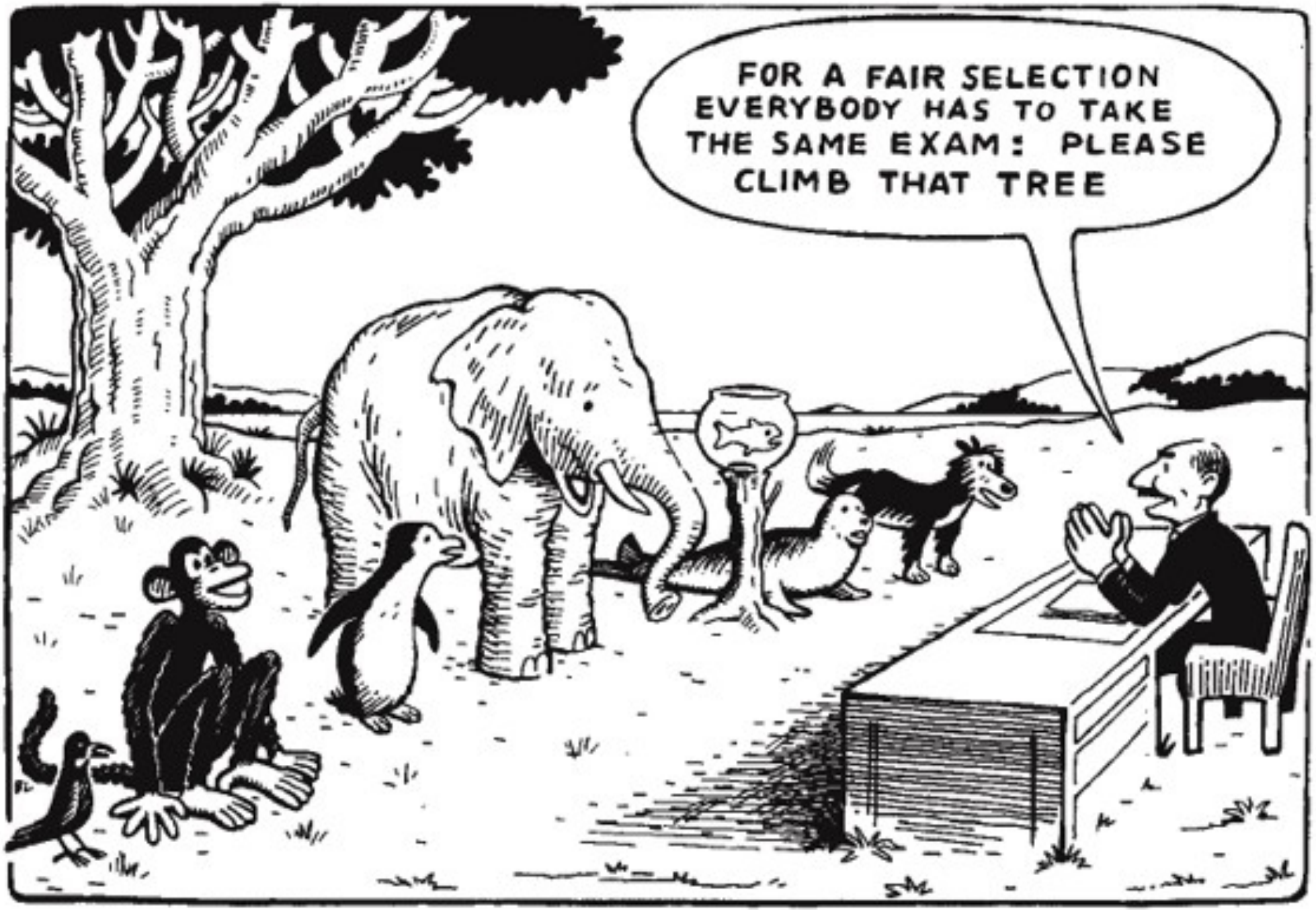
Distinction between
right and wrong

Fundamental questions
about existence,
Values, reason



https://mesbar.org/why-should-we-teach-philosophy/

https://www.nbcnews.com/politics/politics-news/merriam-webster-s-2018-word-year-justice-n949016



INTERVENTION!

http://www.inspiremykids.com/2015/fairness-quotes-for-kids/

Brief history of fairness?

- Not brief! See Hutchinson and Mitchell 2019

- Race
- Ethnicity
- Gender
- …..

- Education
- Employment
- Banking
- Insurance
- Voting
- …..

Sneetches
Dr. Seuss, 1961

"...until neither the Plain nor the Star-Bellies knew

whether this one was that one... or that one was this one...

or which one was what one... or what one was who."

http://www.stevehackman.net/wp-content/uploads/2013/02/Sneetches.jpg

Fairness ~ criteria for justice?

- Aristotle says justice consists in what is lawful and fair, with fairness involving equitable distributions and the correction of what is inequitable
- ….
- Rawls analyzed justice in terms of maximum equal liberty regarding basic rights and duties for all members of society, with socio-economic inequalities requiring moral justification in terms of equal opportunity and beneficial results for all

How can fairness be measured?

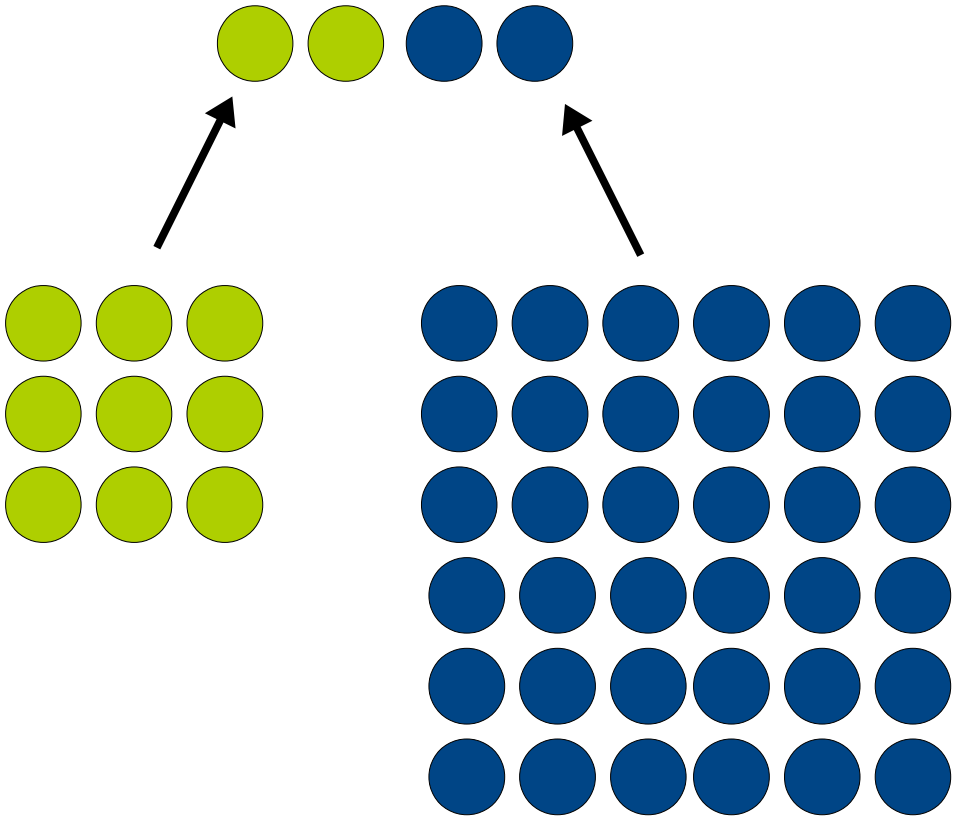
https://www.iep.utm.edu/justwest/

Equal Parity

Also known as Demographic or Statistical Parity

WHEN DO YOU CARE?

If you want each group represented equally among the selected set.



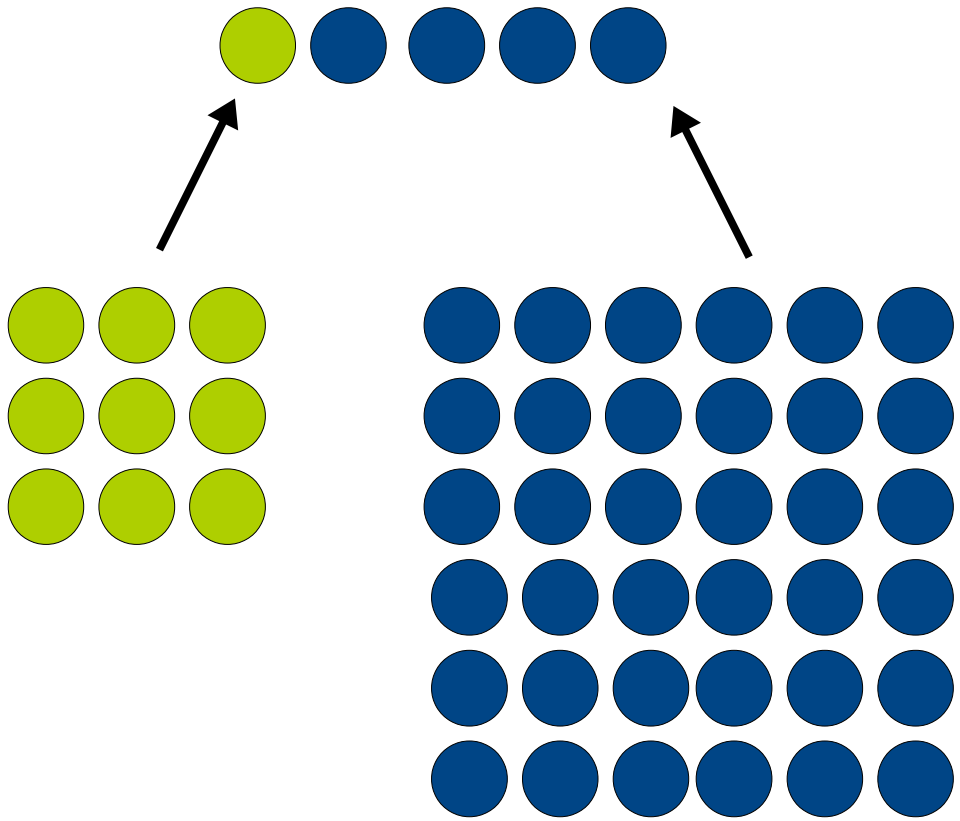
professors

Proportional Parity

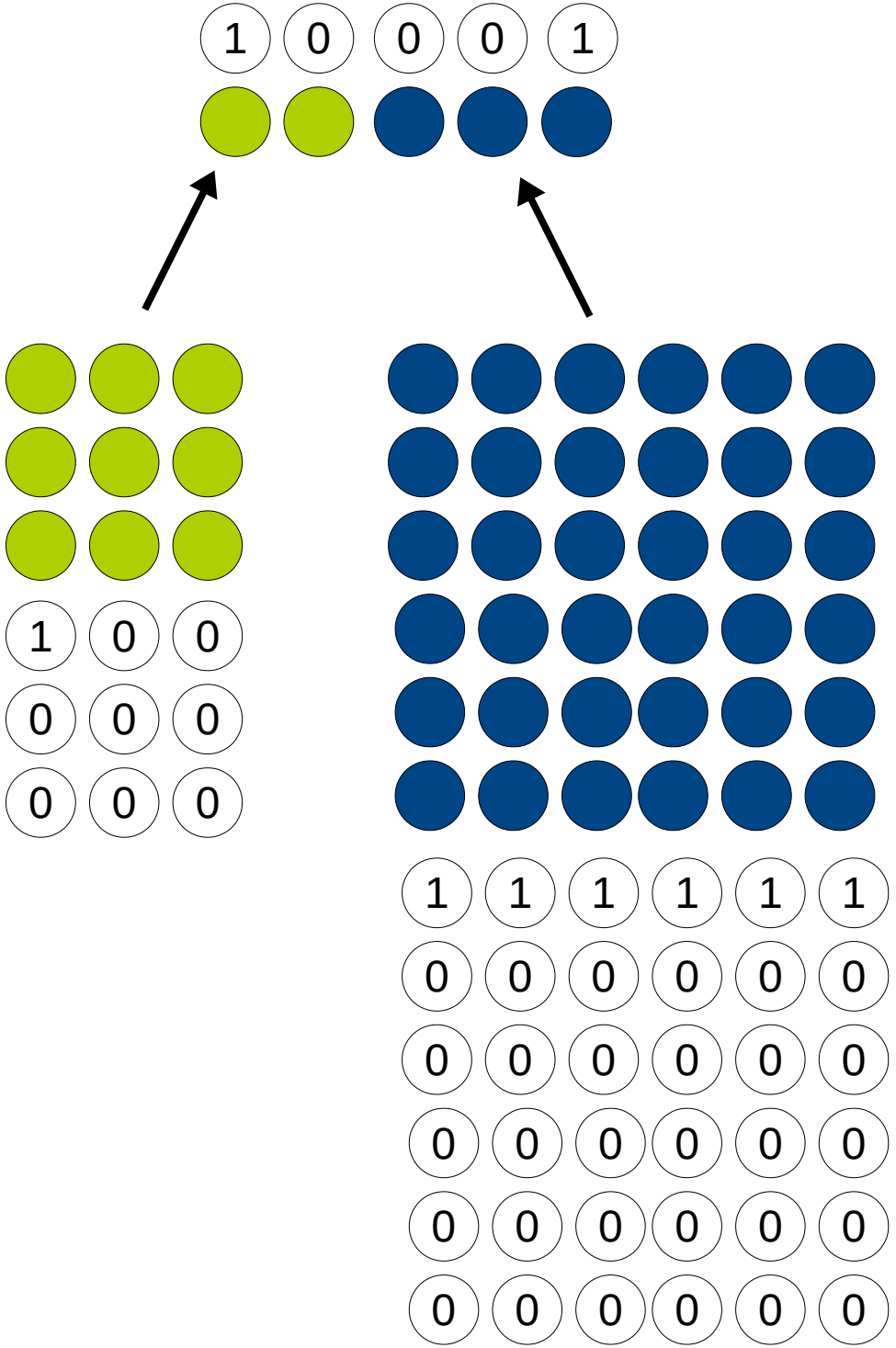
Also known as Impact Parity or Minimizing Disparate Impact

WHEN DO YOU CARE?

If you want each group represented proportional to their representation in the overall population



insurance



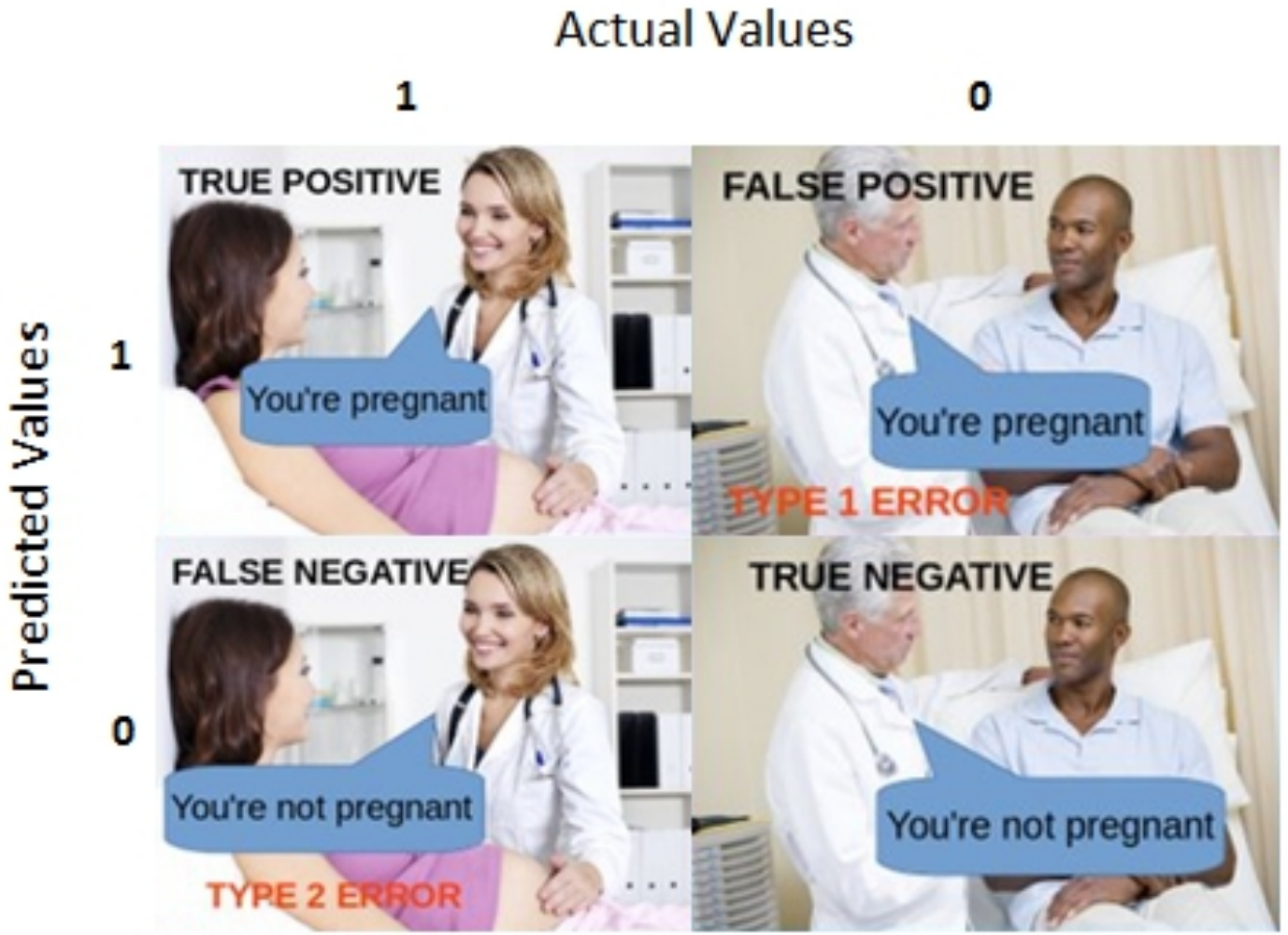
False Positive Parity

Desirable when your interventions are punitive

WHEN DO YOU CARE?

If you want each group to have equal False Positive Rates

prison

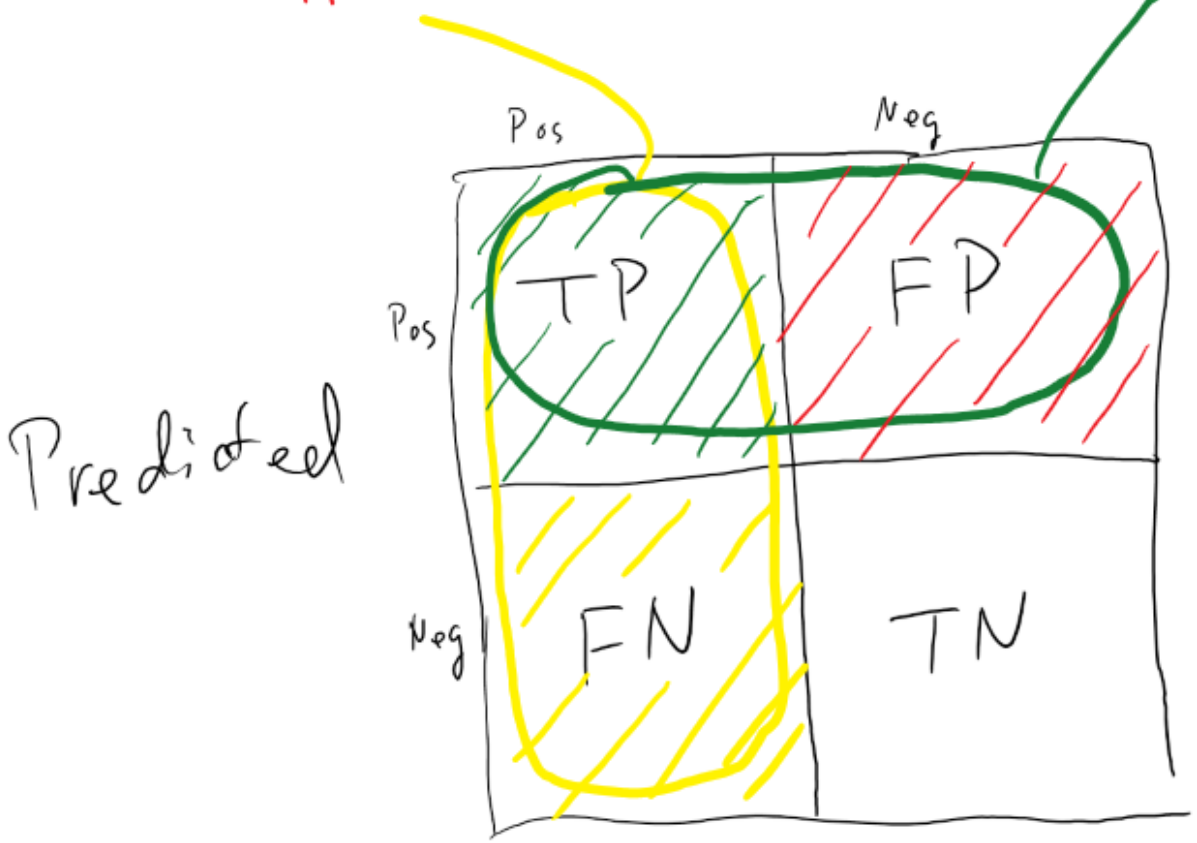


https://www.quora.com/What-is-an-example-for-true-positive-and-true-negative

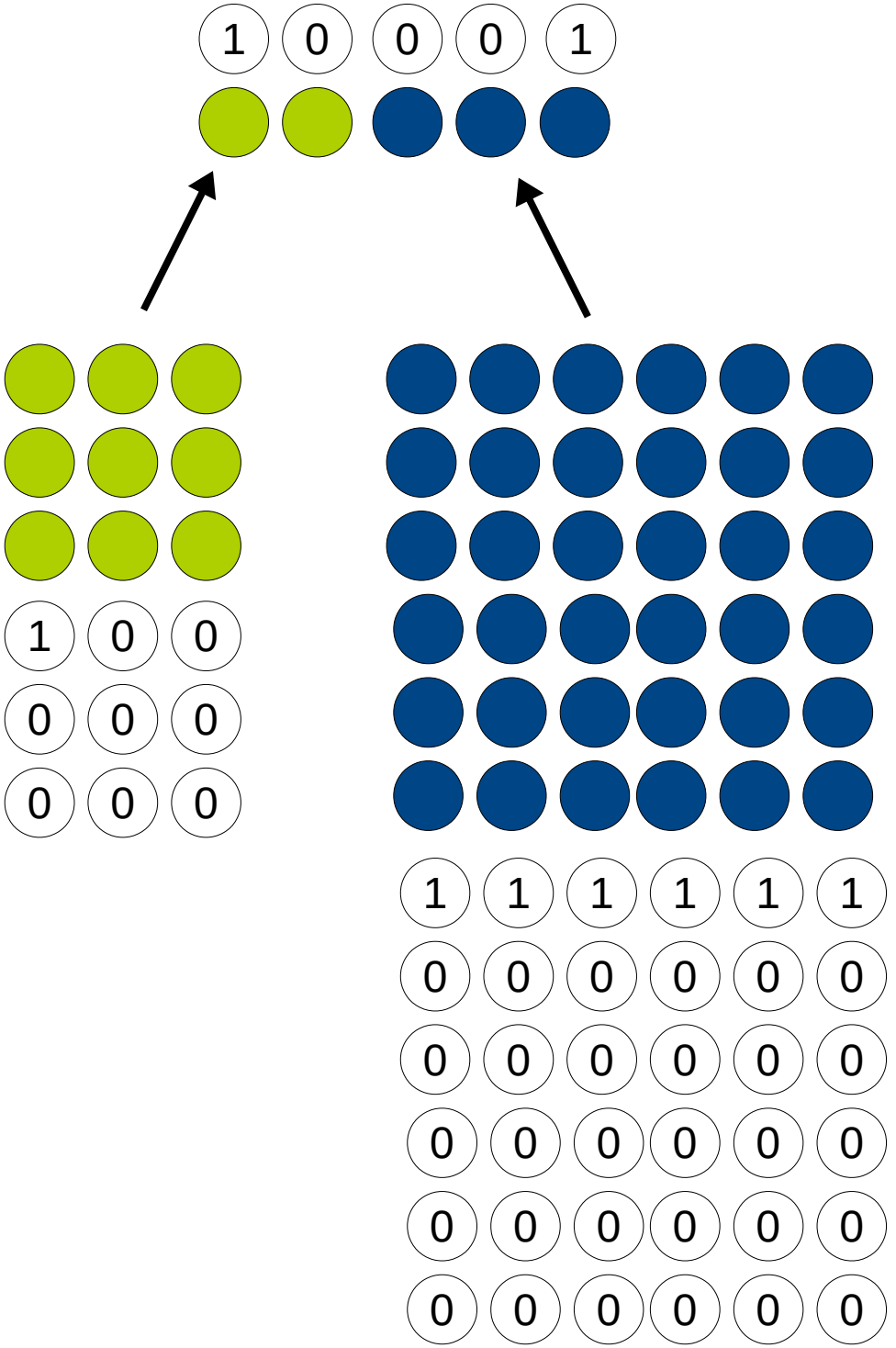
$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$



https://medium.com/cracking-the-data-science-interview/machine-learning-classifier-basics-and-evaluation-44dd760fea50



False Negative Parity

Desirable when your interventions are assistive/preventative

WHEN DO YOU CARE?

If you want each group to have equal False Negative Rates

social benefits

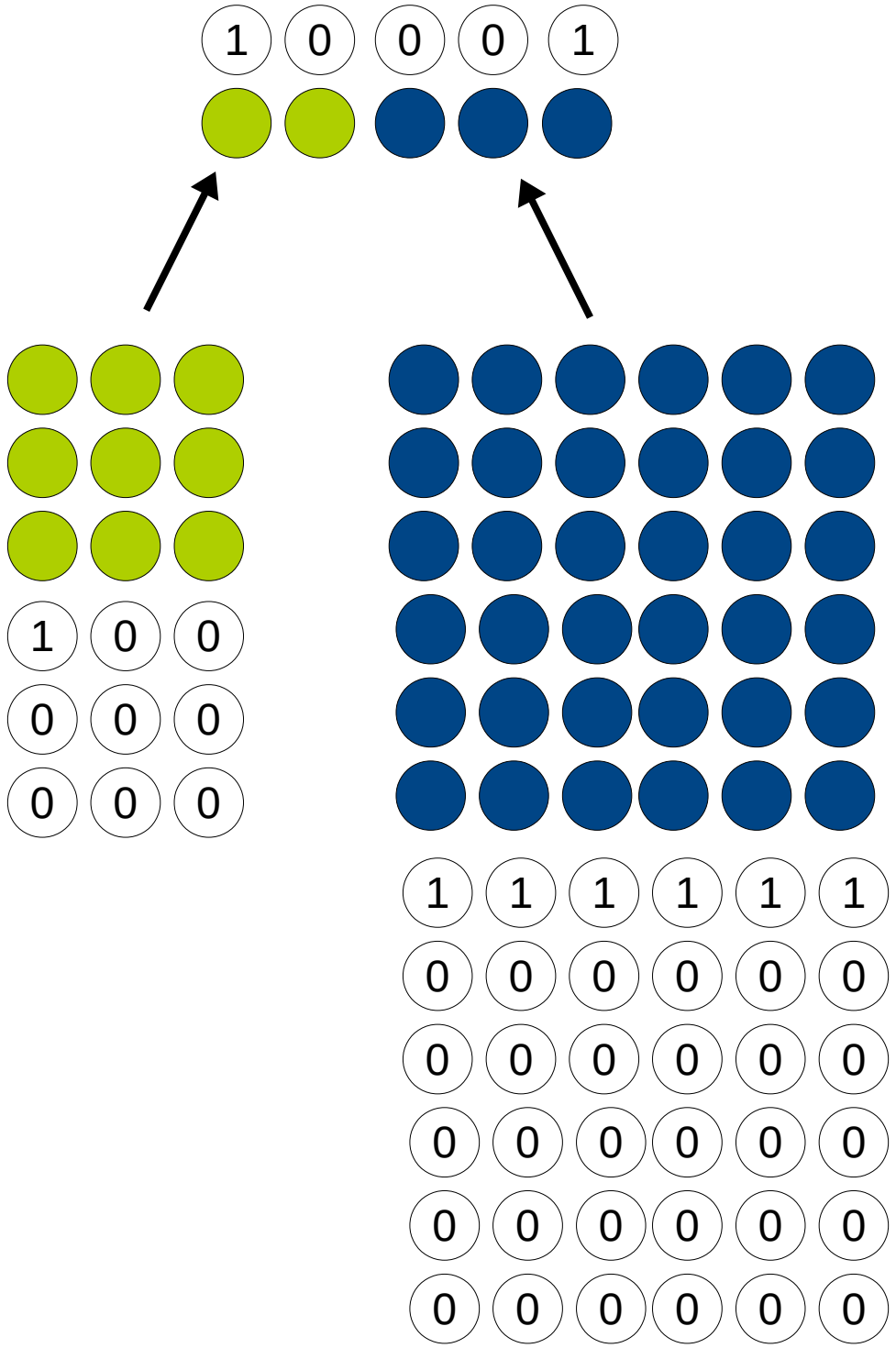
professors

insurance

prison

social benefits

http://aequitas.dssg.io/



Exercise 1

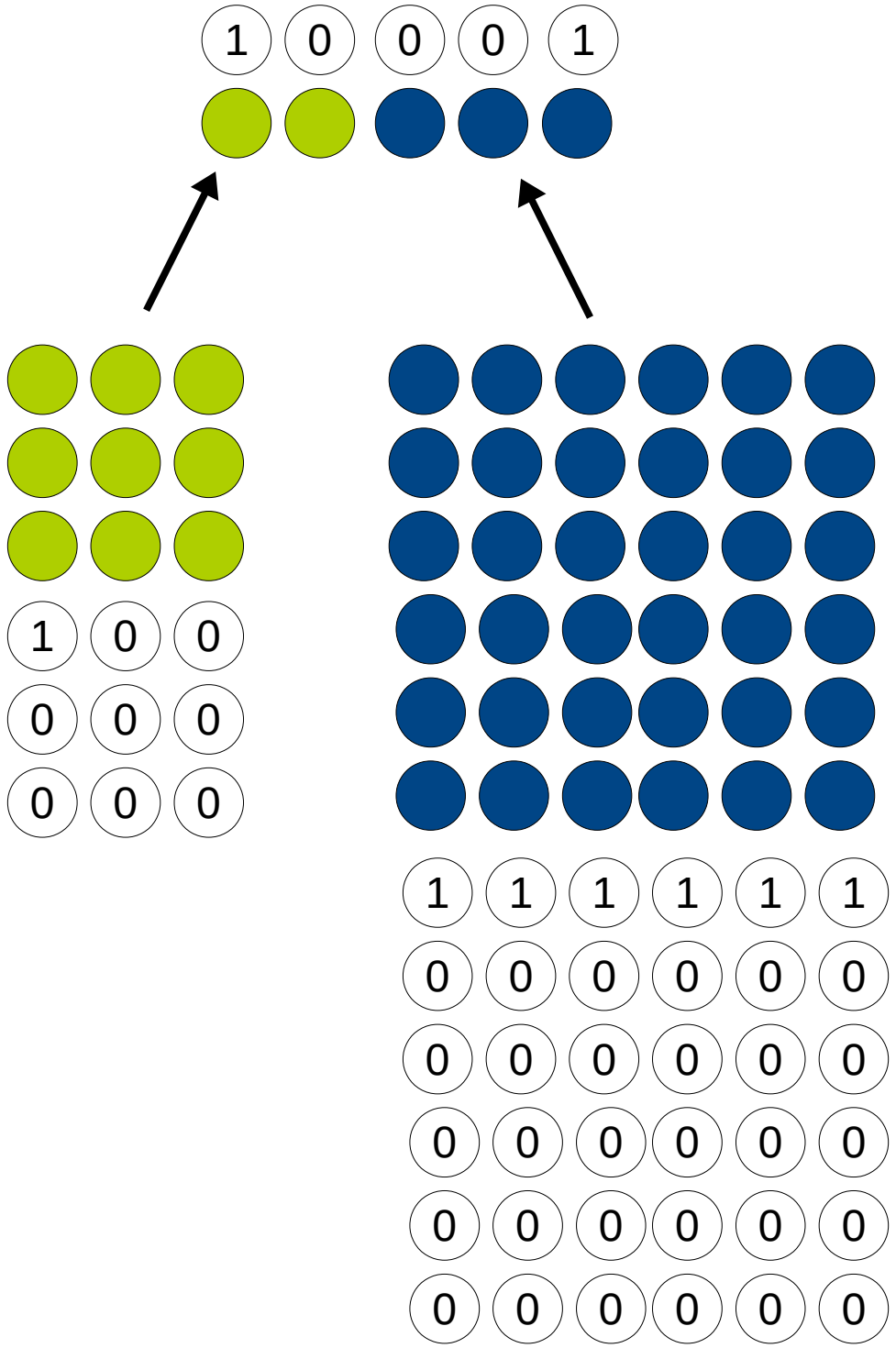
Assuming green is
the reference group compute:

equal parity:
n(green) =
n(blue) =
Diff =

proportional parity:
p(green) =
p(blue) =
Diff =

false positive parity:
fp(green) =
fp(blue) =
Diff =

and false negative parity:
fn(green) =
fn(blue) =
Diff =



Exercise 1

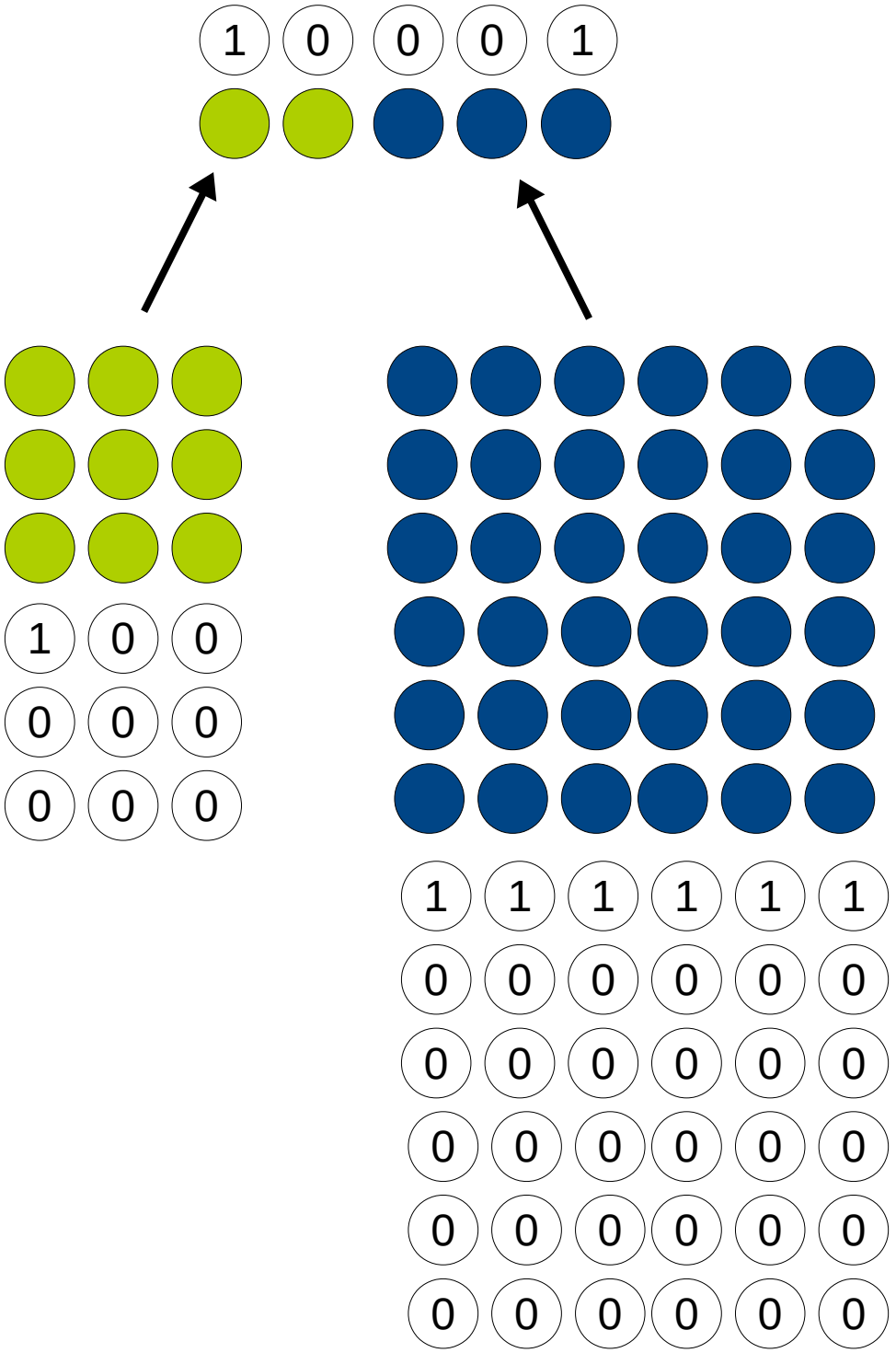
Assuming green is
the reference group compute:

equal parity: FAIL
n(green) = 2
n(blue) = 3
Diff = -50%

proportional parity:
p(green) =
p(blue) =
Diff =

false positive parity:
fp(green) =
fp(blue) =
Diff =

and false negative parity:
fn(green) =
fn(blue) =
Diff =



Exercise 1

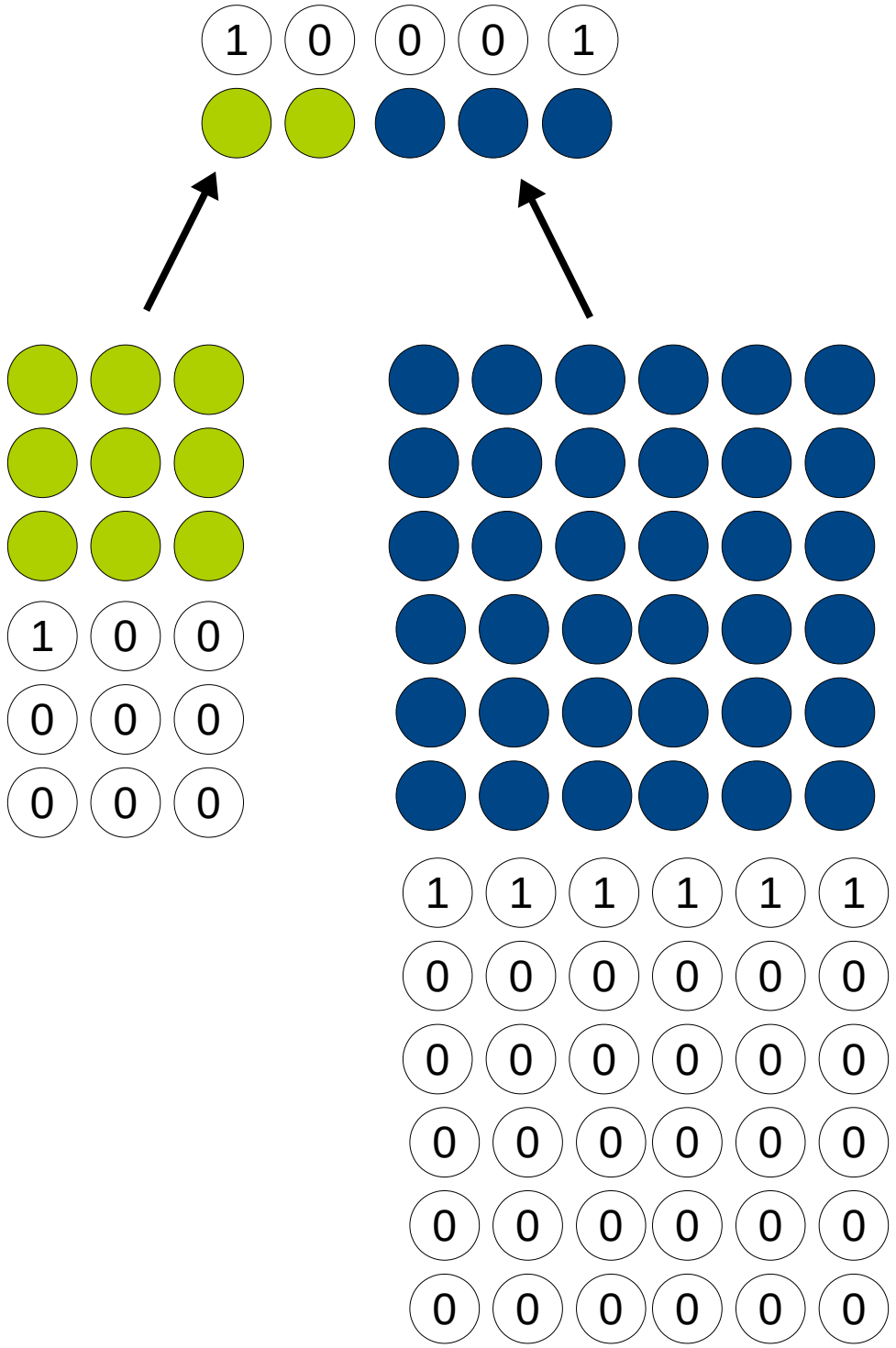
Assuming green is
the reference group compute:

equal parity: FAIL
n(green) = 2
n(blue) = 3
Diff = -50%

proportional parity: FAIL
p(green) = 2/9
p(blue) = 3/36
Diff = 63%

false positive parity:
fp(green) =
fp(blue) =
Diff =

and false negative parity:
fn(green) =
fn(blue) =
Diff =



Exercise 1

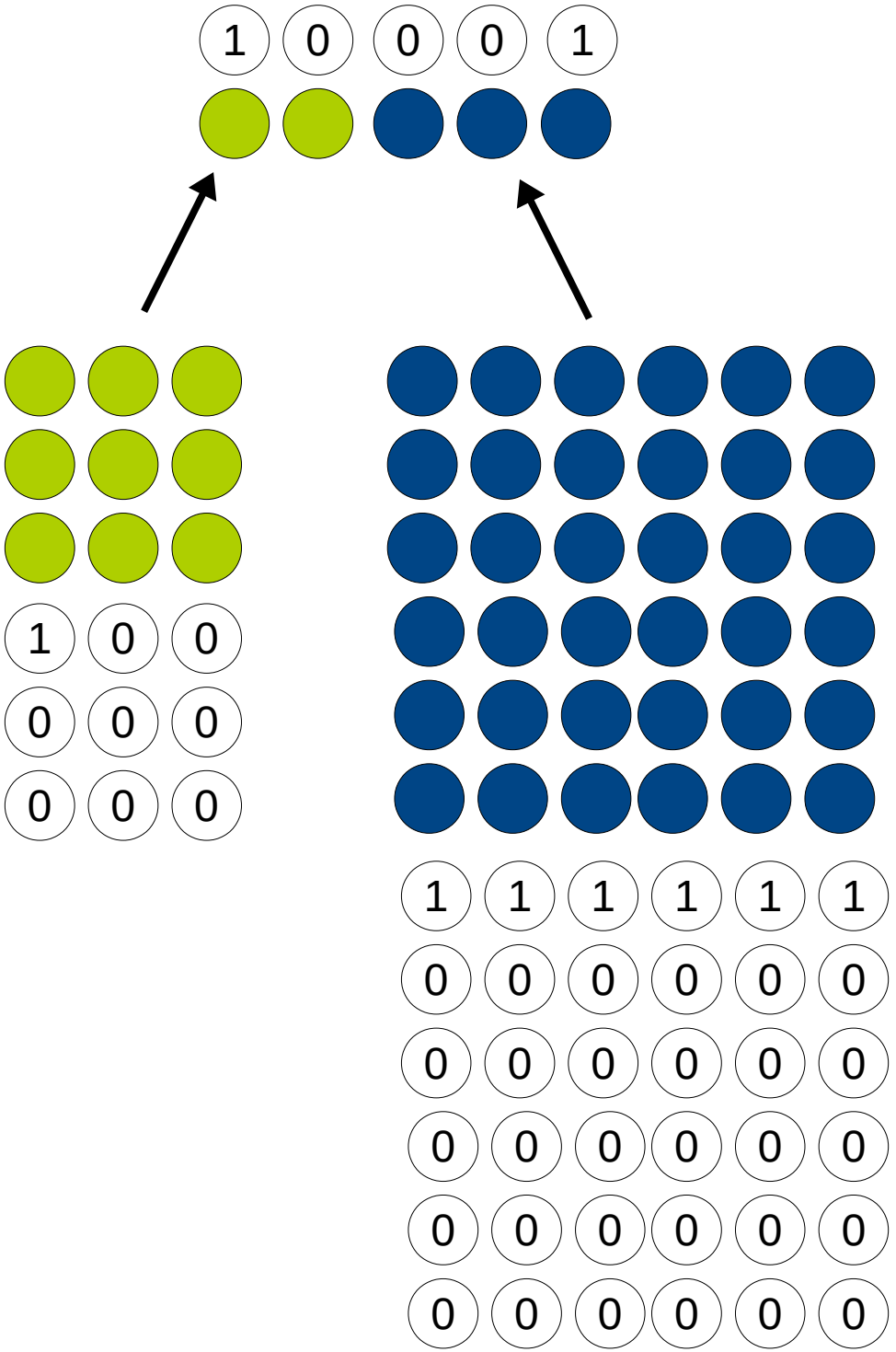
Assuming green is
the reference group compute:

equal parity: FAIL
n(green) = 2
n(blue) = 3
Diff = -50%

proportional parity: FAIL
p(green) = 2/9
p(blue) = 3/36
Diff = 63%

false positive parity: FAIL
fp(green) = 1/9
fp(blue) = 2/32
Diff = -44%

and false negative parity:
fn(green) =
fn(blue) =
Diff =



Exercise 1

Assuming green is
the reference group compute:

equal parity: FAIL
n(green) = 2
n(blue) = 3
Diff = -50%

proportional parity: FAIL
p(green) = 2/9
p(blue) = 3/36
Diff = 63%

false positive parity: FAIL
fp(green) = 1/9
fp(blue) = 2/32
Diff = -44%

and false negative parity: FAIL
fn(green) = 1/2
fn(blue) = 6/7
Diff = 71%

Computing differences

“American”

$$\frac{p(\text{deprived group})}{p(\text{reference group})} - 1$$

Example:
p(green) = 0.3
p(blue) = 0.2

$$\frac{0.2}{0.3} - 1 = -33\%$$

“European”

p(reference group) - p(deprived group)

Example:
p(green) = 0.3
p(blue) = 0.2

0.3 - 0.2 = 0.1

HIDDEN BIAS

When Algorithms Discriminate

By Claire Cain Miller

July 9, 2015

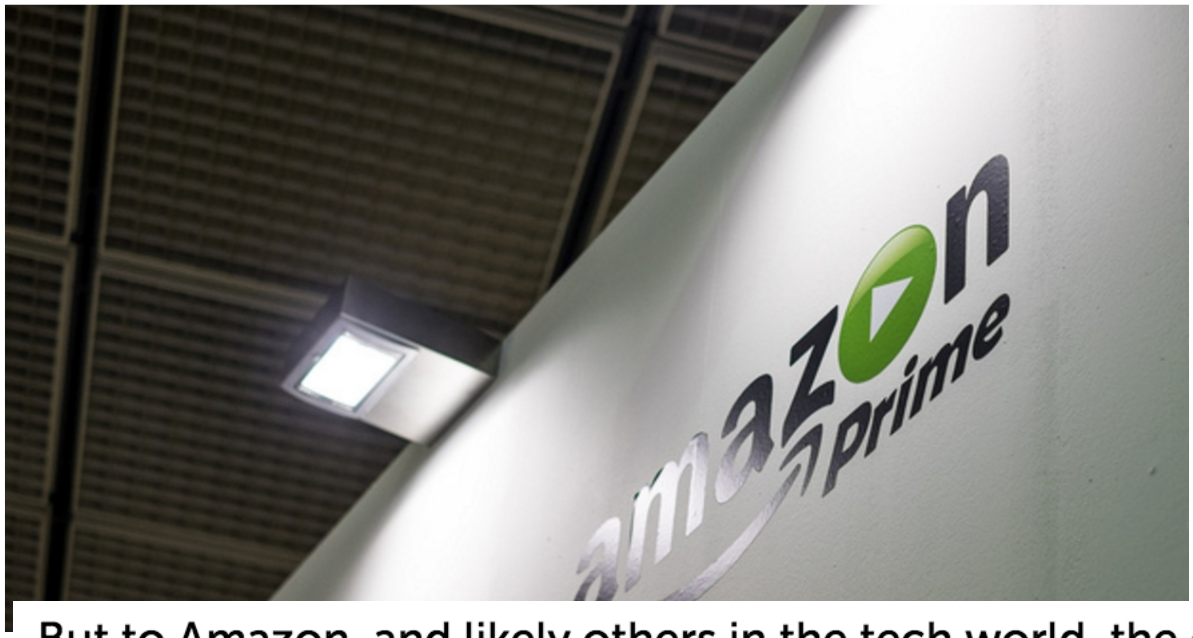
The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

May 11, 2016

OPINION BY PRESTON GRALLA

Amazon Prime and the racist algorithms



MORE LIKE THIS


Meet the 2016 Premier 100


How Kronos' cloud migration makes for a better business


This is why Amazon will open physical bookstores

on IDG Answers
What is two-factor authentication and why

But to Amazon, and likely others in the tech world, the decision had nothing to do with racism and everything to do with the facts. Amazon argued that it wasn't acting on prejudice when it excluded those neighborhoods. Instead, algorithms and the underlying data on which those algorithms were based made it clear that Amazon couldn't make a profit in them. And, given that Amazon is profit-driven, the company excluded them. Race, Amazon said, had nothing to do with it.

The Guardian

How white engineers built racist code – and why it's dangerous for black people

As facial recognition tools play a bigger role in fighting crime, inbuilt racial biases raise troubling questions about the systems that create them

Ali Breland

Mon 4 Dec 2017 09.00 GMT

Brief history of fairness-aware ML

ALGORITHM WATCH

2008 First paper on discrimination-aware data mining by Pedrechi et al

2009 First paper on algorithmic prevention by Kamiran and Calders

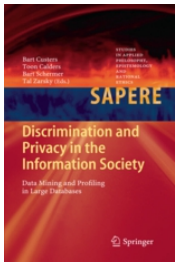
2010 First Bayesian solutions by Calders and Verwer

2011 Conditional discrimination paper by Žliobaitė et al

2011 First PhD thesis: F.Kamiran

2012 First dedicated workshop at IEEE ICDM

2013 Edited book



2013 Second PhD thesis: S.Hajian

2014 Special issue in AI and Law

2014 Obama report I

2014 AirBnB case: digital discrimination

2014/5 FATML is established

2015 Obama report II

2015 Repeated coverage in Guardian

2015 Google hires Moritz Hardt

2015 First causality perspective paper by Bonchi at al

2015 First dedicated course, Aalto and UH

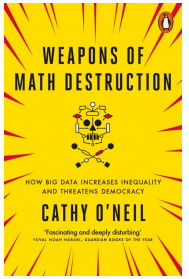
Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

2016 Obama report III

2016 NGOs are being established

2016 Many research papers are coming out, conferences start having dedicated tracks



2016 O'Neil's book

2017 FATML is sold out

2017-2018 Major attention from the society, media, funding agencies

2017-2018 A new paper or a few every week, many on optimization criteria/ measurement

2019 FAT*, policies, working groups, institutes, faculty positions, ...

AI in everyday life: data driven decision support

Banking

Hiring

Sports analytics

University admittance

Image source: https://www.thebalance.com/how-credit-scores-work-315541
http://tcgstudy.com/ranking_to_universities.html
https://www.pinterest.com/pin/443041682071895474/
https://www.insperity.com/blog/people-analytics-step-step-guide-using-data-make-hiring-decisions/

(Supervised) machine learning

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Machine learning:
model is a summary of data

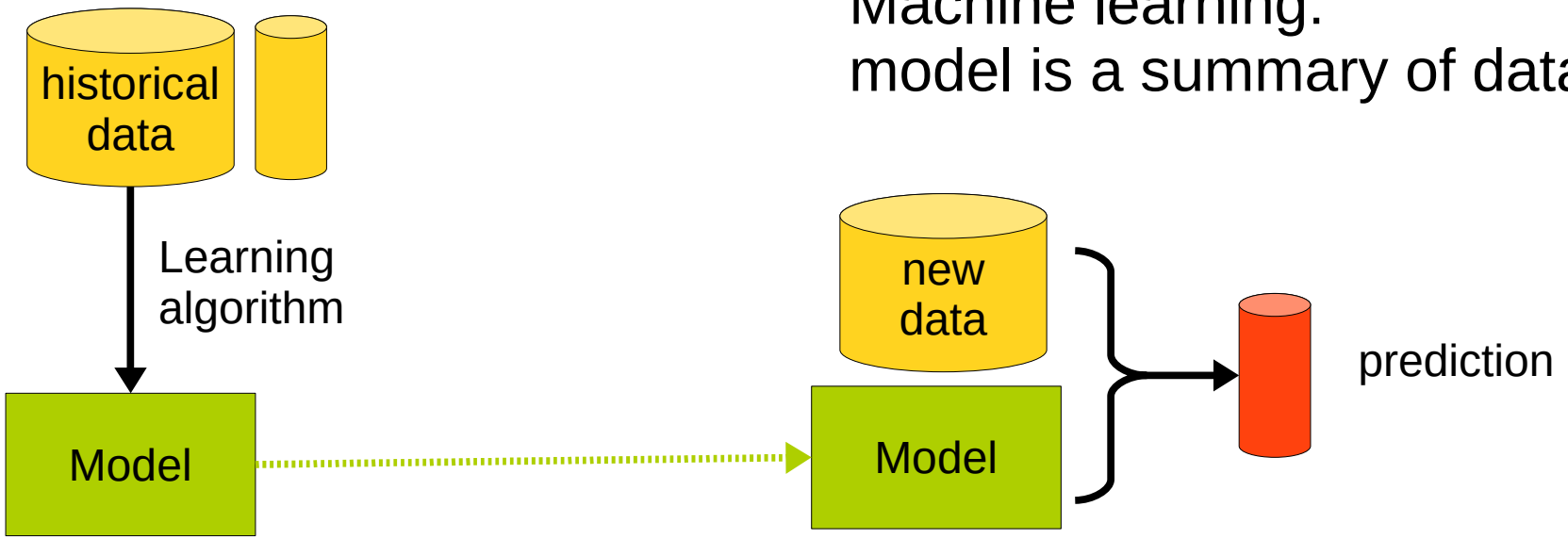


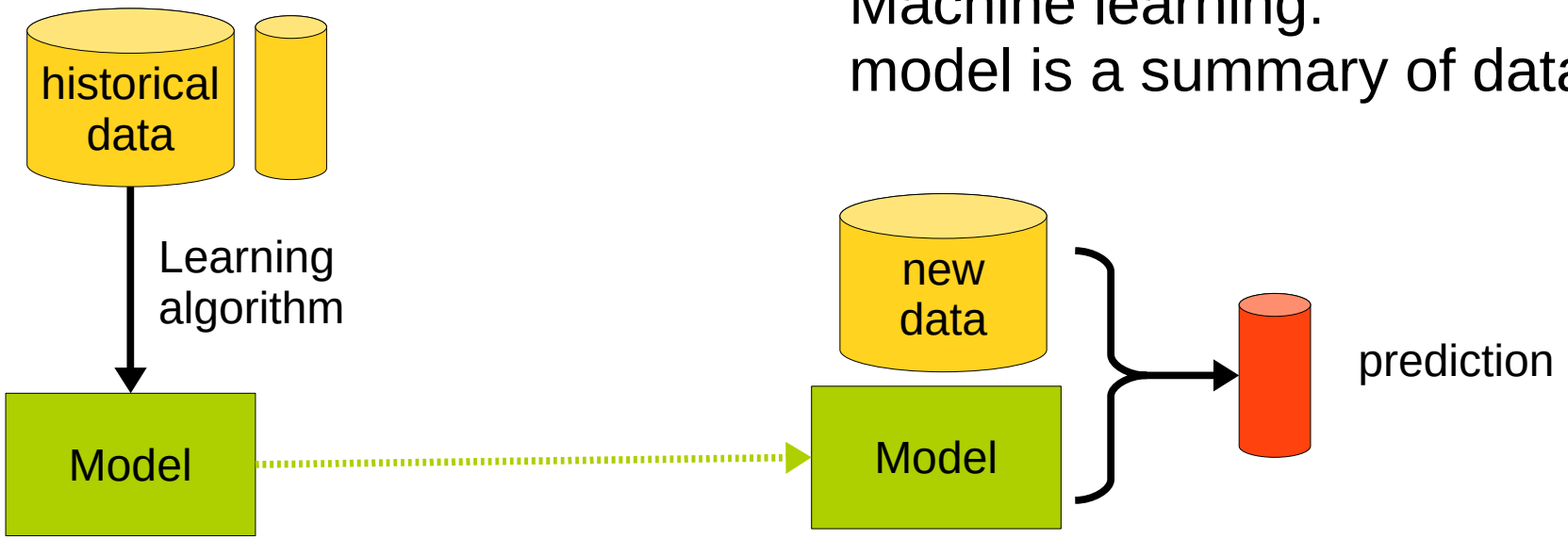
Image source: Introduction to Data Mining, 2nd Edition by Tan, Steinbach, Karpatne, Kumar

If there are biases in the data on which a model is trained

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Machine learning:
model is a summary of data



Unless instructed otherwise,
the learned model will carry those biases forward

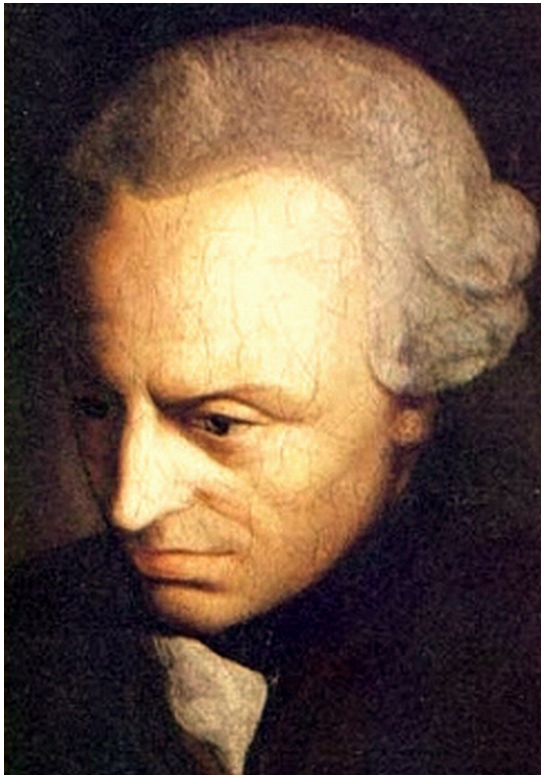
Machine learning and discrimination

- Discrimination – inferior treatment based on ascribed group rather than individual merits
- Discriminate = *distinguish* (lat.)

- Machine learning
  - uses proxies to distinguish an individual from the mean
  - without judging what is morally right or wrong
  - can enforce constraints defined by legislation and/or social norms (external)

There are no “right” or “wrong” variables

Morality is a social convention?



Immanuel Kant introduced the categorical imperative:
"Act only according to that maxim whereby you can,
at the same time, will that it should become a universal law"

Source: Wikimedia Commons, Unidentified painter 18th-century portrait painting of men

Which variables are fair to use? In which circumstances?

- Gender
  - Credit scoring?

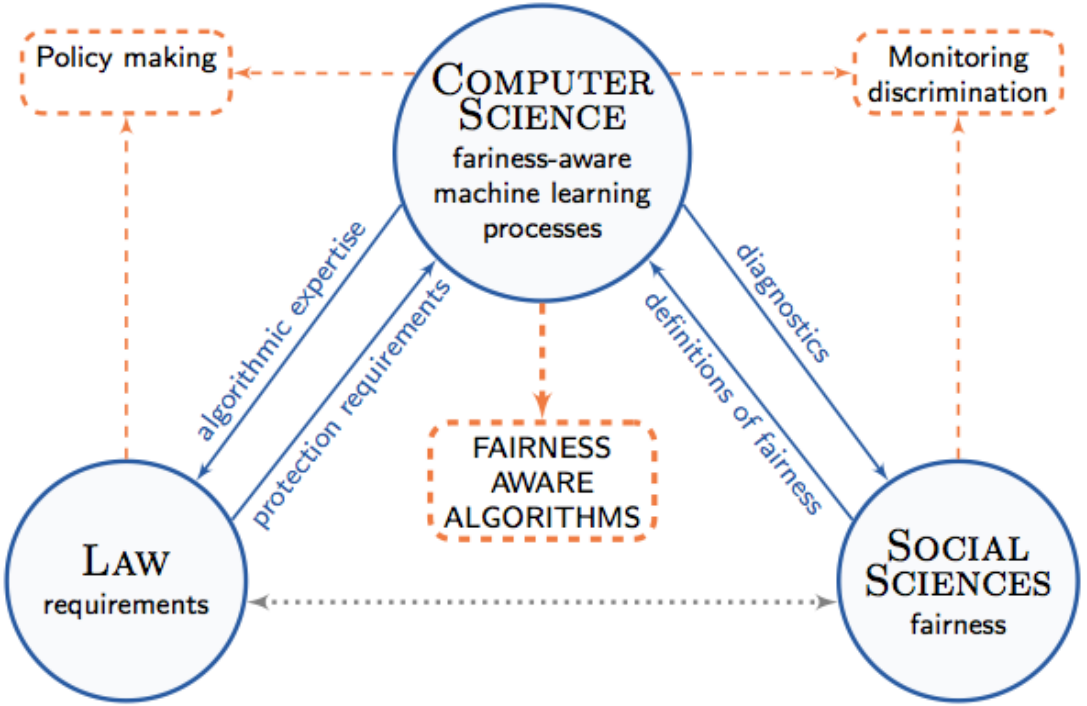
Which variables are fair to use? In which circumstances?

- Gender
  - Credit scoring?
  - Sports?

Which variables are fair to use? In which circumstances?

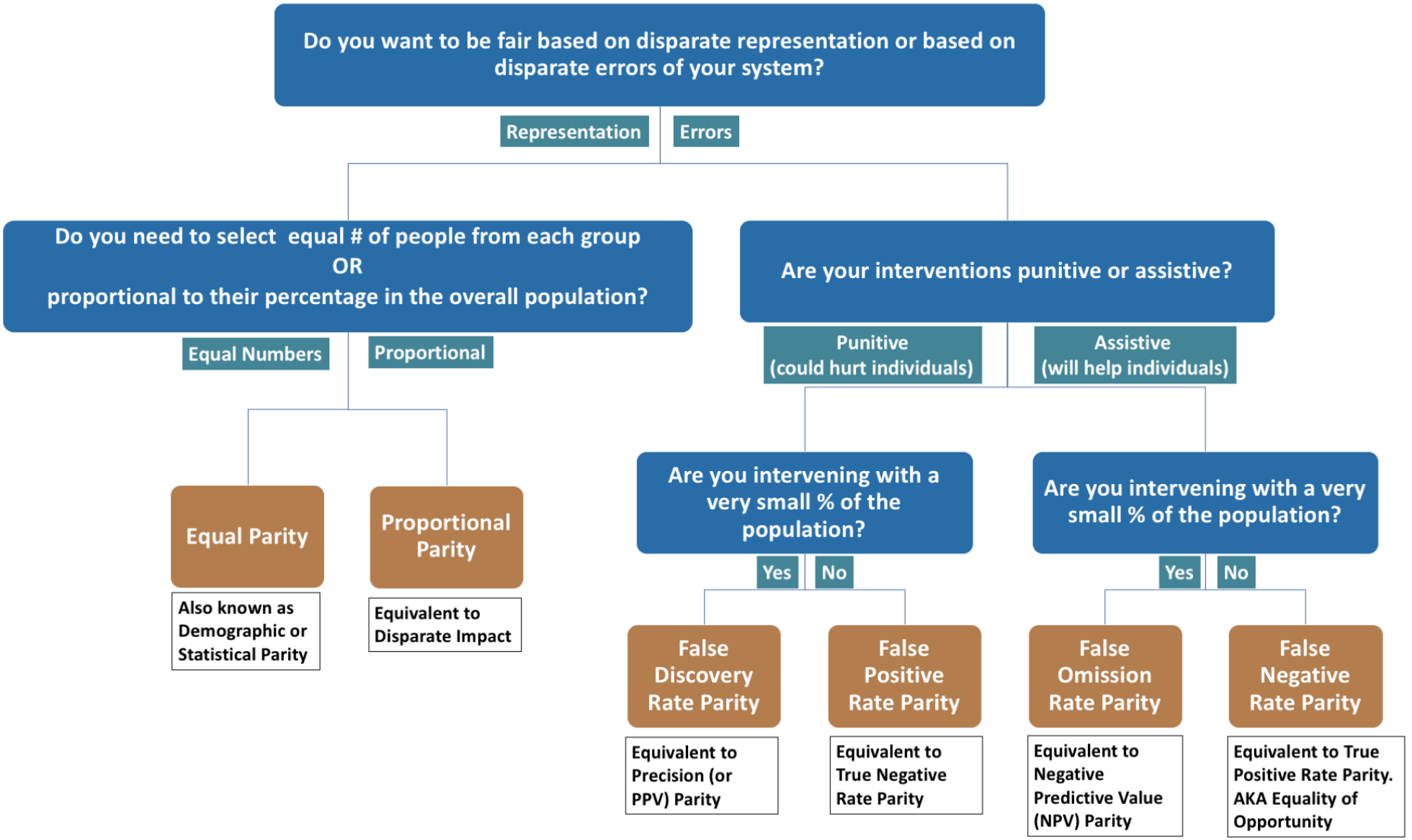
- Gender
  - Credit scoring?
  - Sports?
  - Medical diagnosis?

Cross-disciplinary challenges



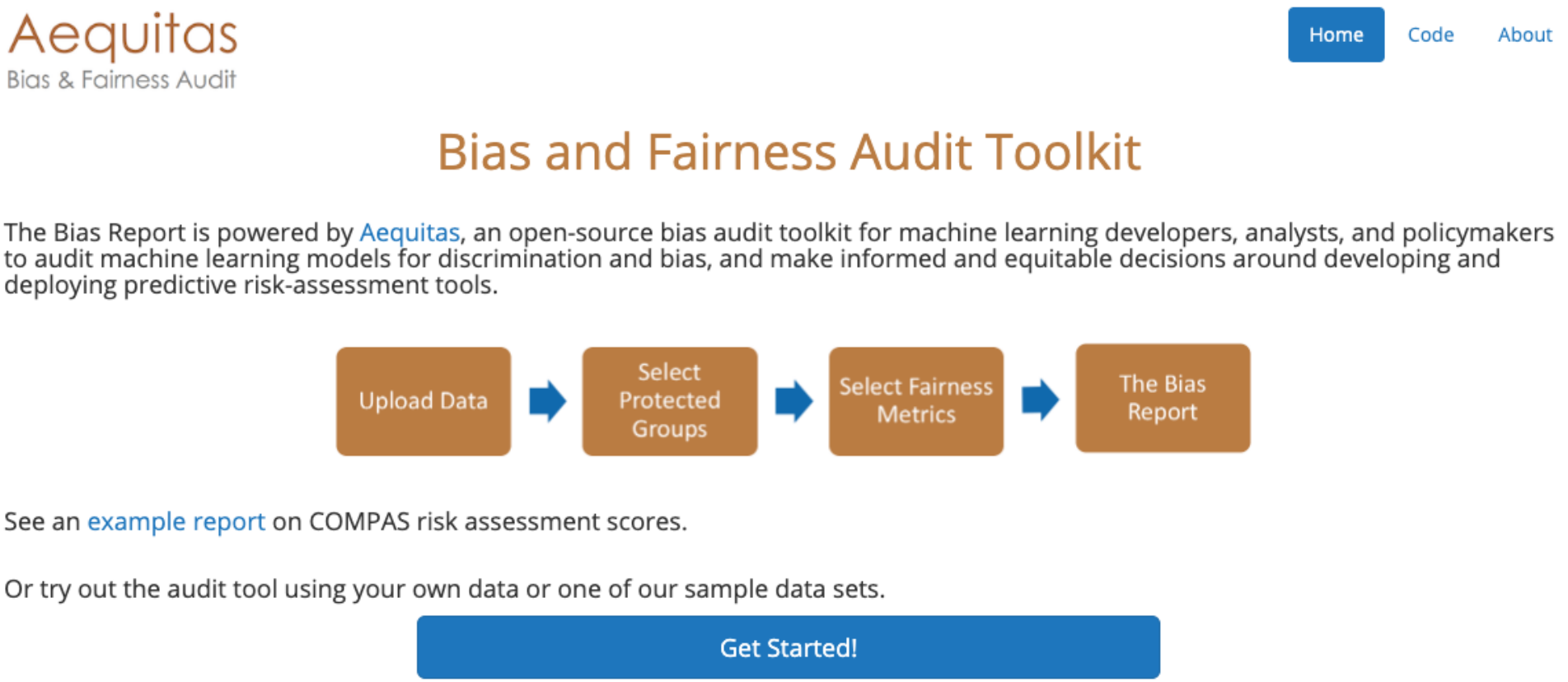
- Discrimination model — social-technical
  - How discrimination happens? “Right” allocation of resources?
- Optimization criteria — legal-technical
  - How to select a good measure for non-discrimination?
- Sanitizing decision models — technical
  - How to incorporate constraints into algorithmic decision making?

FAIRNESS TREE



Source: Center for Data Science and Public Policy - University of Chicago

Exercise 2



http://aequitas.dssg.io/

Determining which are the sensitive variables and what kind of justice should be aimed at

is an extremely difficult, interdisciplinary problem

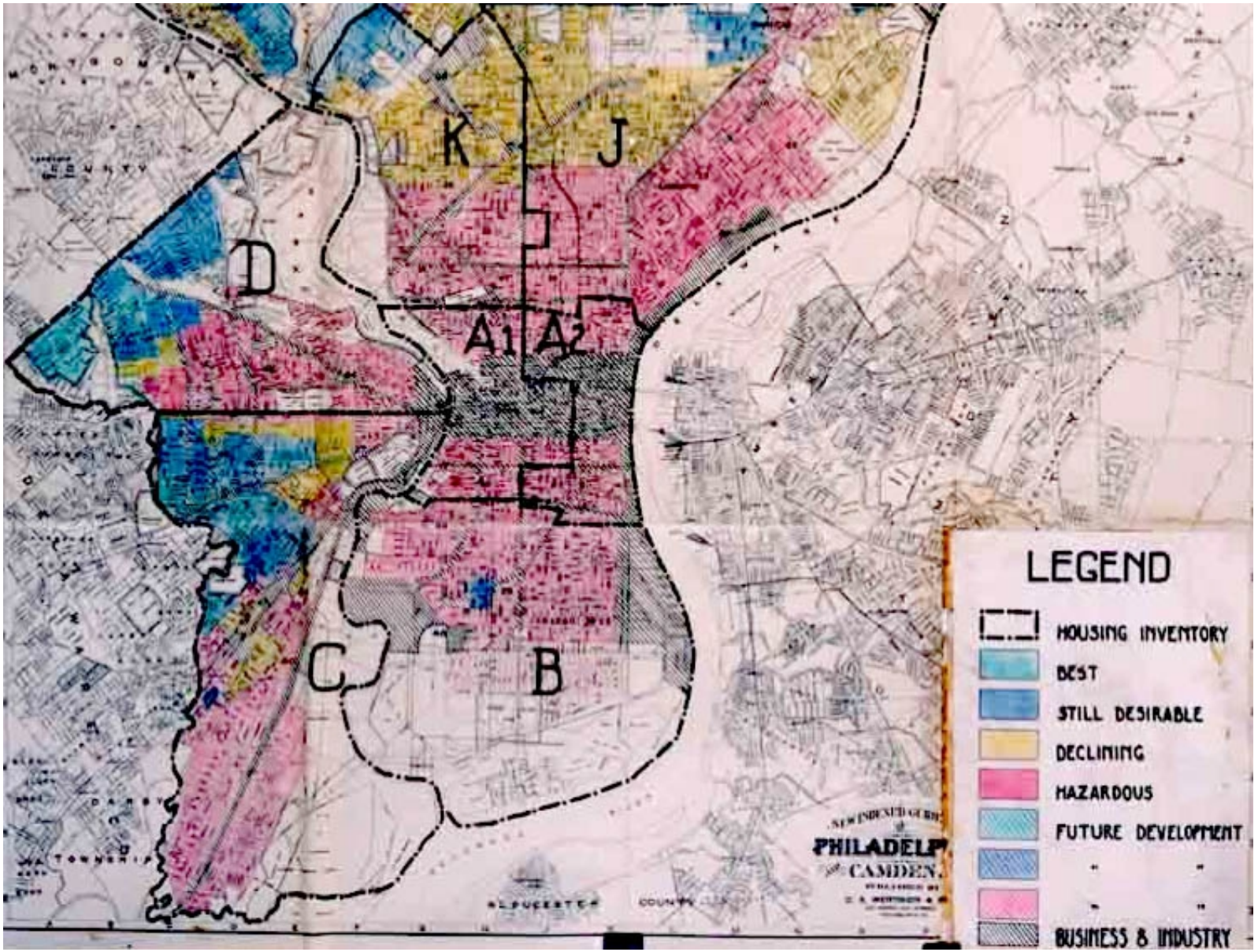
We will not address here
We will assume that we somehow know, which variables are ok to use

+REDLINING!

Variables are usually correlated with each other

Postal code with race
Working hours with gender
Level of education with ethnicity

Redlining



Source: "Home Owners' Loan Corporation Philadelphia redlining map". Licensed under Public Domain via Wikipedia

Why fair computational process can lead to “unfair” outcomes?

How can it happen?

- Suppose salary is decided (in decision maker's head) as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

Removing sensitive variables makes things worse

- Suppose salary is decided (in decision maker's head) as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Data scientist assumes

$$salary = b_0 + b_1 \times education$$

Removing sensitive variables makes things worse

- Suppose salary is decided (in decision maker's head) as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Data scientists assumes

$$salary = b_0 + b_1 \times education$$

- Observes data

education	*ethnicity*	*salary*
1	1	600
2	1	700
3	1	800
4	1	900
10	1	1500

education	*ethnicity*	*salary*
1	0	1100
6	0	1600
7	0	1700
9	0	1900
10	0	2000

Privacy vs. Ethics
Example 1: regression

- Suppose historically salary is decided as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Data scientists assumes

$$salary = b_0 + b_1 \times education$$

- Observes data

education	*ethnicity*	*salary*
1	1	600
2	1	700
3	1	800
4	1	900
10	1	1500

education	*ethnicity*	*salary*
1	0	1100
6	0	1600
7	0	1700
9	0	1900
10	0	2000

Žliobaitė and Custers 2016

Removing sensitive variables makes things worse

- Suppose salary is decided (in decision maker's head) as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Data scientist assumes

$$salary = b_0 + b_1 \times education$$

- Observes data

education	*ethnicity*	*salary*
1	1	600
2	1	700
3	1	800
4	1	900
10	1	1500

education	*ethnicity*	*salary*
1	0	1100
6	0	1600
7	0	1700
9	0	1900
10	0	2000

- Learns model

$$salary = 602 + 128 \times education$$

Removing sensitive variables makes things worse

- Suppose historically salary is decided as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Data scientists assumes

$$salary = b_0 + b_1 \times education$$

- Observes data

education	*ethnicity*	*salary*
1	1	600
2	1	700
3	1	800
4	1	900
10	1	1500

education	*ethnicity*	*salary*
1	0	1100
6	0	1600
7	0	1700
9	0	1900
10	0	2000

- Learns model

$$salary = 602 + 128 \times education$$

Lower base salary,
higher reward for education
the model punishes
ethnical minorities

Žliobaitė and Custers 2016

How large is the bias?

- Suppose salary is decided (in decision maker's head) as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

α

- Learned model

$$salary = 602 + 128 \times education$$

γ β

- Salary = (1000 - 398) + (100 + 28) x education

Punishment on base salary

$\gamma = \alpha$ x mean(Ethnicity) – β x mean(education)
= (-500)x0.5 – 28x5.3 = 398

Award for education

$\beta = \alpha$ x Cov(Education,Ethnicity)/ Var(Education)
= (-500)x(-0.72)/12.9 = 28

Žliobaitė and Custers 2016

A simple solution (special case)

- Learn a model on the full dataset

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Remove the sensitive component of the model

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

Žliobaitė and Custers 2016

Protect privacy ≠ prevent discrimination

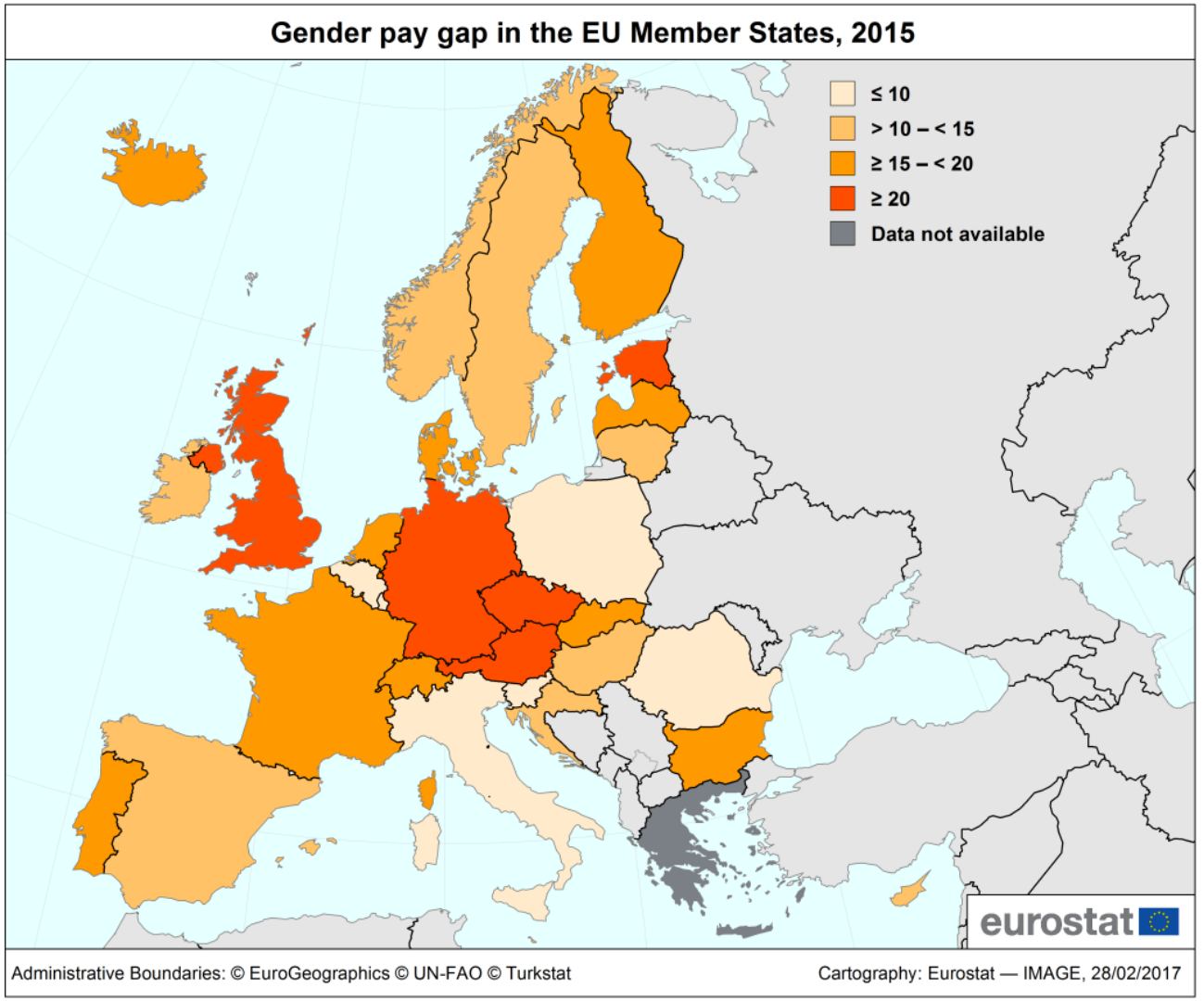
$$salary = 602 + 128 \times education$$

Lower base salary,
higher reward for education
the model punishes
ethnical minorities

Learning by causality, learning by association

Why unbiased computational processes can lead to discriminative decision procedures

- data is “incorrect”
  - due to biased decisions in the past



Why unbiased computational processes can lead to discriminative decision procedures

- data is incomplete (omitted variable bias)

Image source:
https://www.syfy.com/sites/syfy/files/styles/syfy_episode_recap_full_breakpoints_theme_syfy_smartphone_1x/public/2018/03/futurama_408_hero.jpg

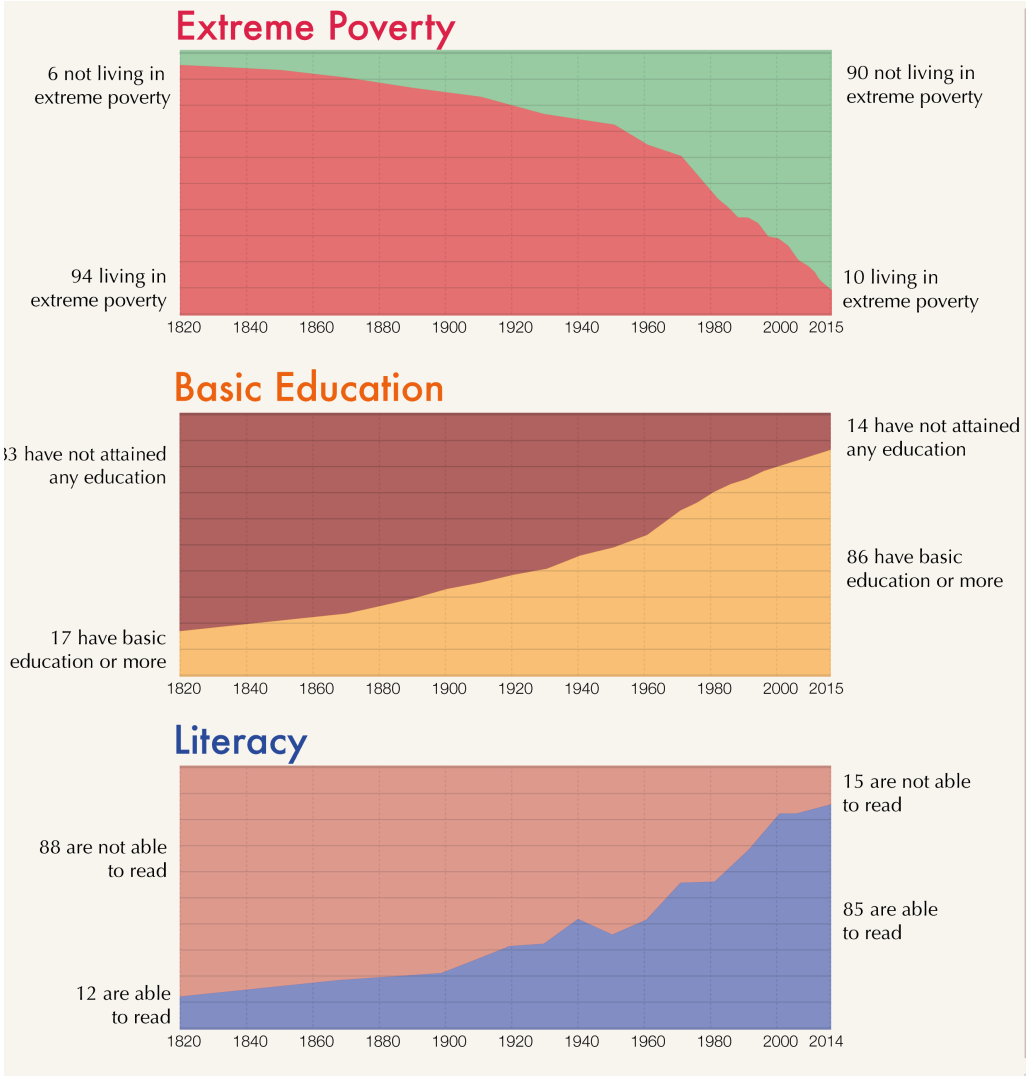
Why unbiased computational processes can lead to discriminative decision procedures

- sampling procedure skews the data

Image source: https://www.abc.net.au/news/image/8670240-3x2-700x467.jpg

Why unbiased computational processes can lead to discriminative decision procedures

- The world is changing over time

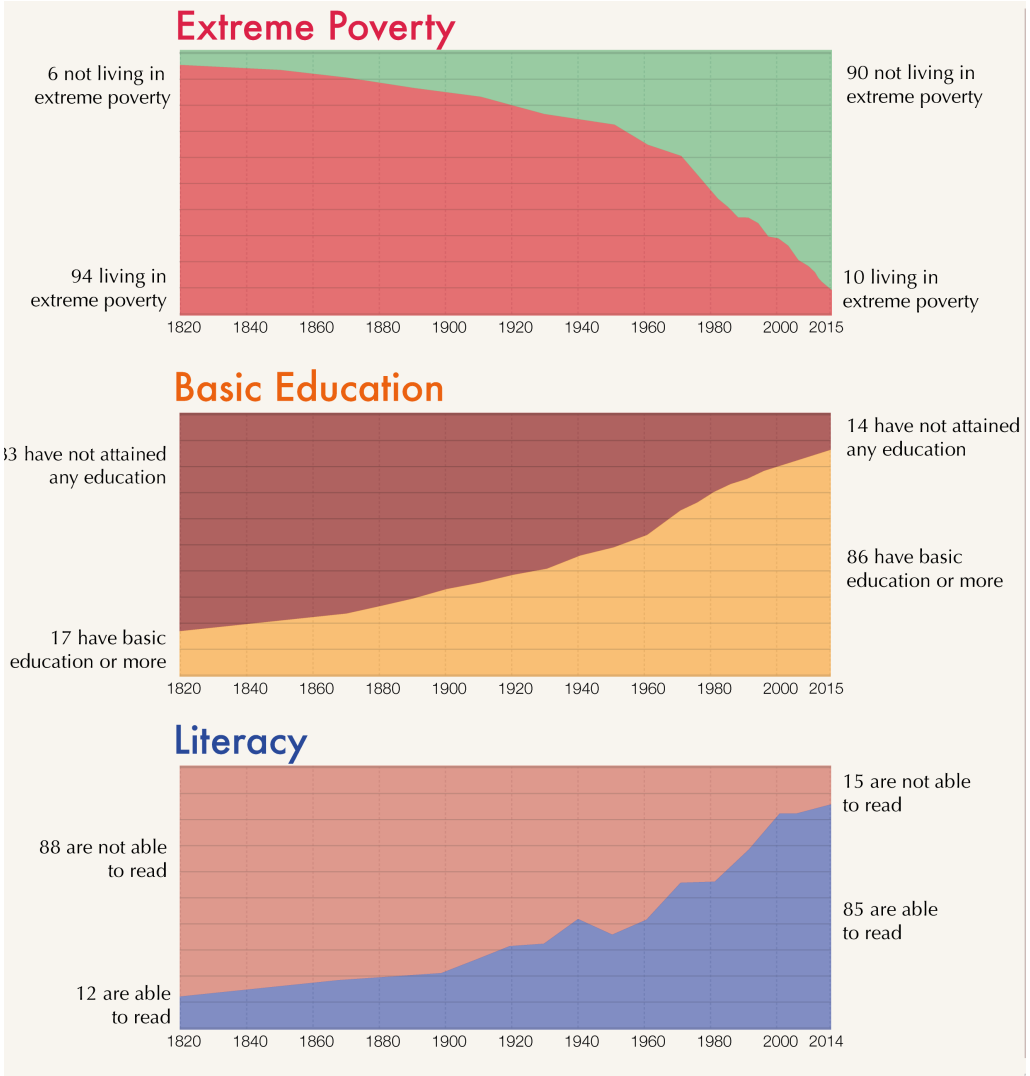


The World as 100 People
over the last two centuries

Image source:
https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts/
https://ritholtz.com/2017/04/world-100-people-last-two-centuries/

Why unbiased computational processes can lead to discriminative decision procedures

- The world is changing over time



How about modeling the universe?

Image source:
https://ourworldindata.org/a-history-of-global-living-conditions-in-5-charts/
https://ritholtz.com/2017/04/world-100-people-last-two-centuries/

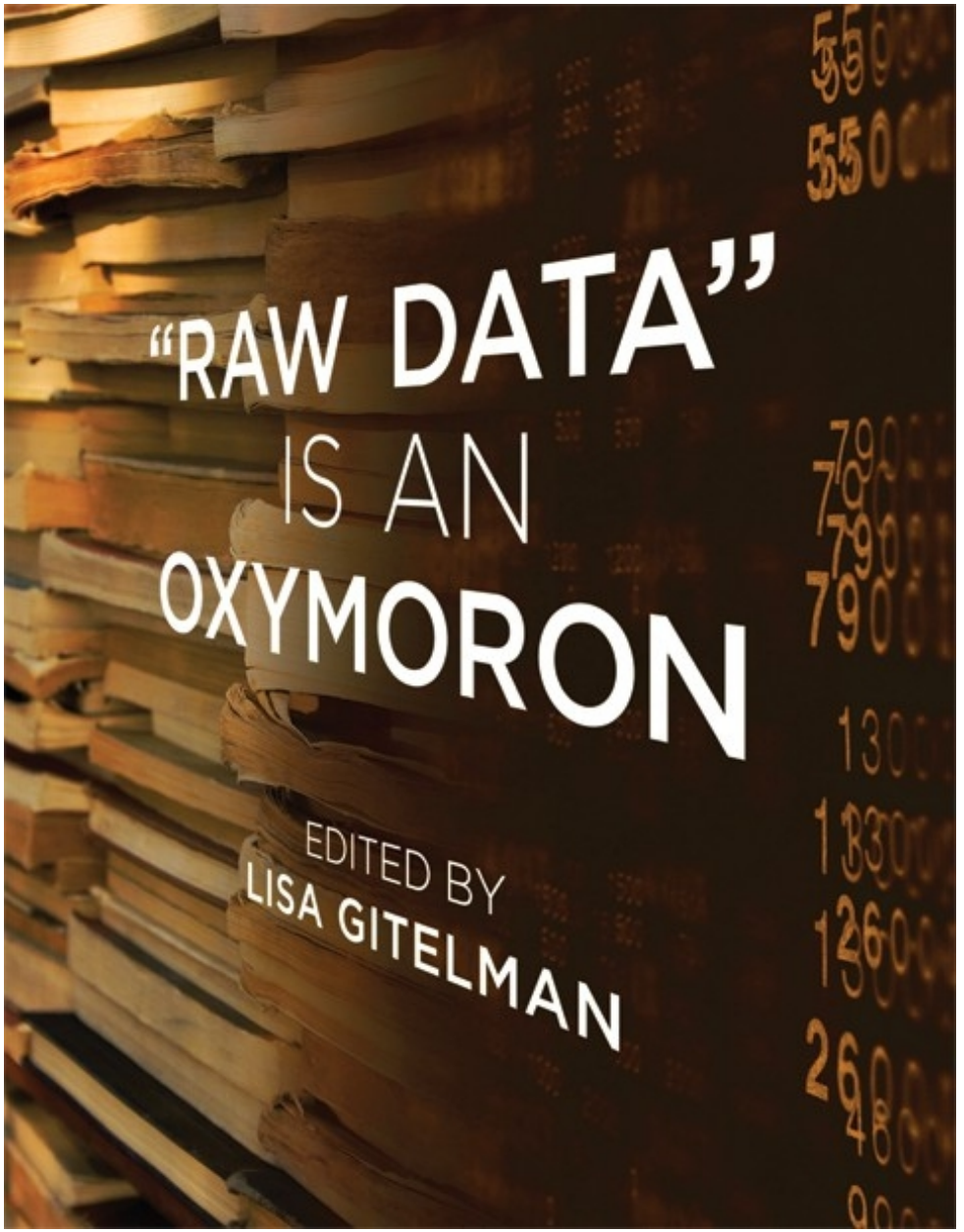
Why unbiased computational processes can lead to discriminative decision procedures

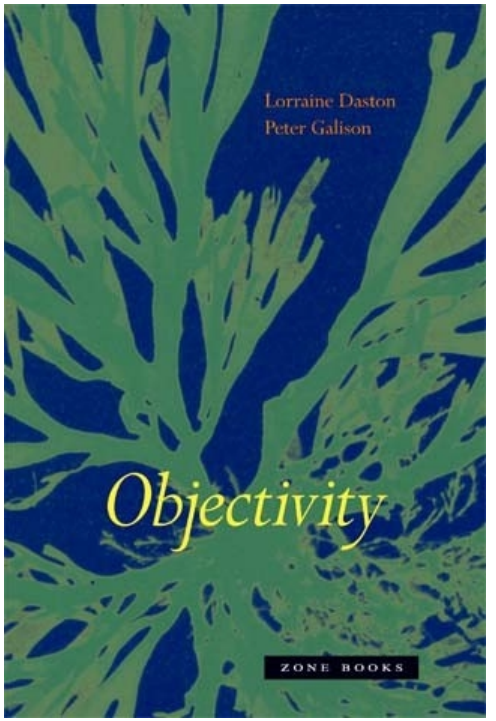
- data is “incorrect”
  - due to biased decisions in the past

Computer scientists or society?

- data is incomplete (omitted variable bias)
- sampling procedure skews the data
- the world is changing over time

Calders and Žliobaitė 2013, book chapter


"RAW DATA"
IS AN
OXYMORON
EDITED BY
LISA GITELMAN


Lorraine Daston
Peter Galison
Objectivity
ZONE BOOKS

Measuring

Machine learning and discrimination

y polarized

- Discrimination – inferior treatment based on ascribed group rather than individual merits

s

X

Predictive models $y \rightarrow f(X)$

Source: https://www.theverge.com/2018/8/16/17693866/ford-self-driving-car-safety-report-dot

Machine learning and discrimination

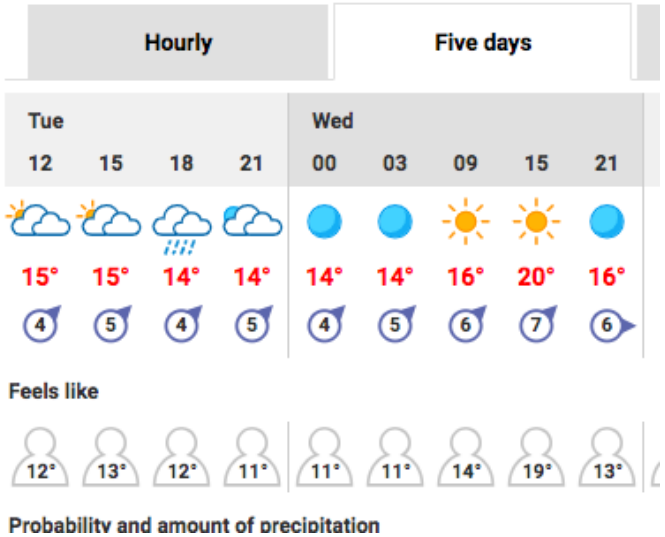
y polarized

- Discrimination – inferior treatment based on ascribed group rather than individual merits

s

X

Predictive models y → f(X)



?

Source: https://ilmatieteenlaitos.fi/

Machine learning and discrimination

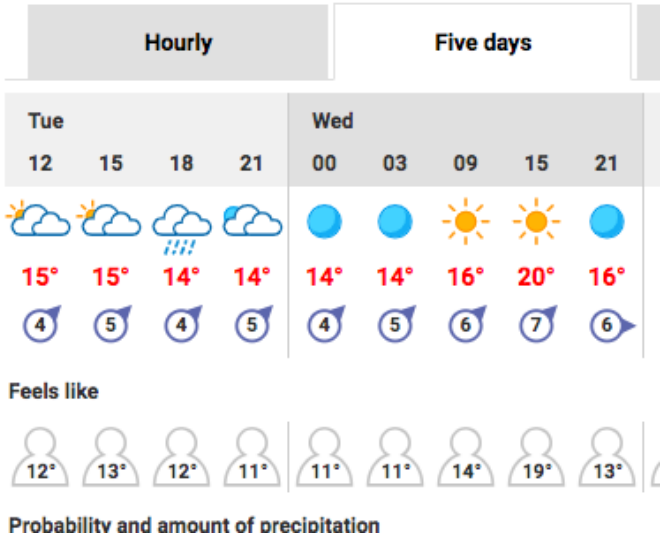
y polarized

- Discrimination – inferior treatment based on ascribed group rather than individual merits

s

X

Predictive models y → f(X)



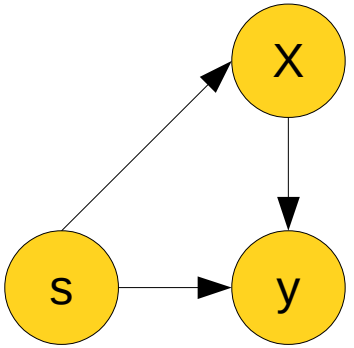
?

?

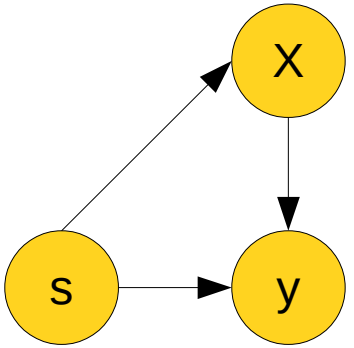
Source: https://ilmatieteenlaitos.fi/

Source: https://www.theverge.com/2018/8/16/17693866/ford-self-driving-car-safety-report-dot

- Removing protected characteristic does not solve the problem if s is correlated with X
  - desired:
    - y → f(X)
  - what happens:
    - s* → f(X)
    - y → f(X,s*)



- Removing protected characteristic does not solve the problem if s is correlated with X
  - desired:
    - y → f(X)
  - what happens:
    - s* → f(X)
    - y → f(X,s*)



e.g. working hours

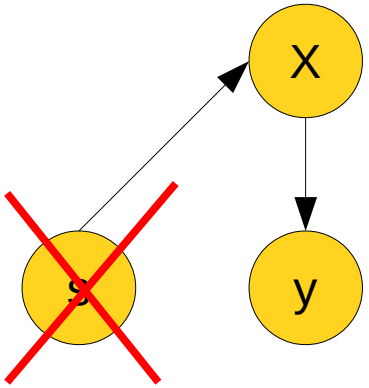
e.g. salary

e.g. gender

	hours per week		annual income (K$)
female	36.4	5.8$/h	10.9
male	42.4	13.8$/h	30.4
all data	40.4	11.4$/h	23.9

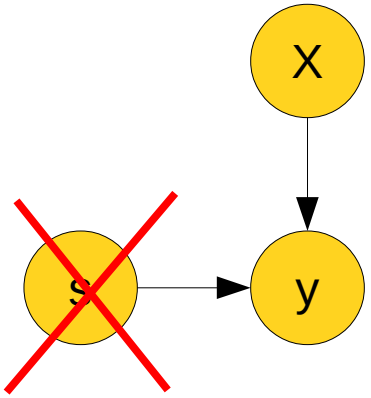
Kamiran and Žliobaitė 2013

- Removing protected characteristic does not solve the problem, unless..



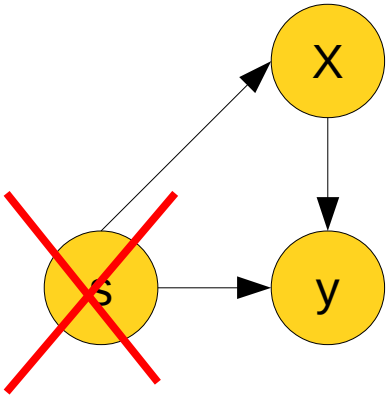
No problem

e.g. all differences in salaries can be explained by working hours



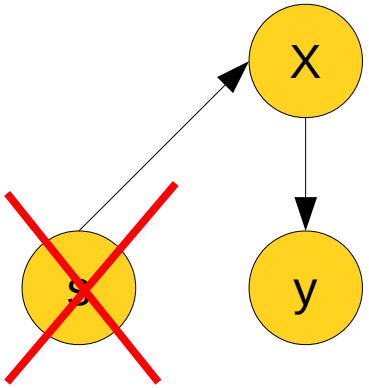
No problem

e.g. salaries differ, but all else is equal

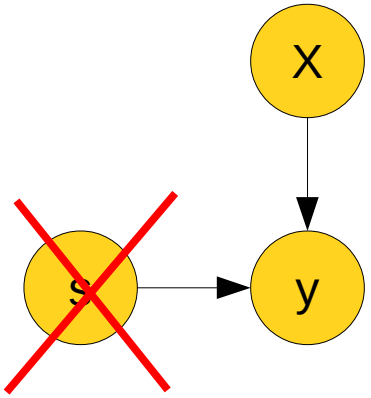


Problem!

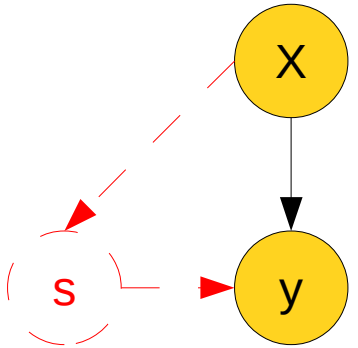
- Removing protected characteristic does not solve the problem, unless..



NO DISCRIMINATION



DIRECT DISCRIMINATION



INDIRECT DISCRIMINATION

Measuring: direct discrimination
“Twin test”

- $p(+|X,s1) \neq p(+|X,s2)$

Image source: https://www.reference.com/science/can-boy-girl-identical-twins-f20c37b6ec0408c1

“Blind” auditioning

Source: https://www.astridbaumgardner.com/blog-and-resources/blog/ysm-mock-auditions/

Measuring discrimination in algorithmic decision making

Is pretty much the same as measuring discrimination in human decision making

Equal Parity

Also known as Demographic or Statistical Parity

WHEN DO YOU CARE?

If you want each group represented equally among the selected set.

professors

Proportional Parity

Also known as Impact Parity or Minimizing Disparate Impact

WHEN DO YOU CARE?

If you want each group represented proportional to their representation in the overall population

insurance

False Positive Parity

Desirable when your interventions are punitive

WHEN DO YOU CARE?

If you want each group to have equal False Positive Rates

prison

False Negative Parity

Desirable when your interventions are assistive/preventative

WHEN DO YOU CARE?

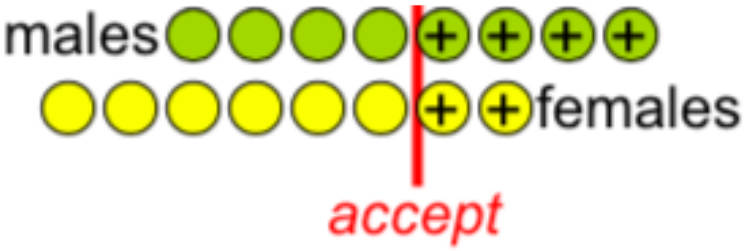
If you want each group to have equal False Negative Rates

bank loan

http://aequitas.dssg.io/

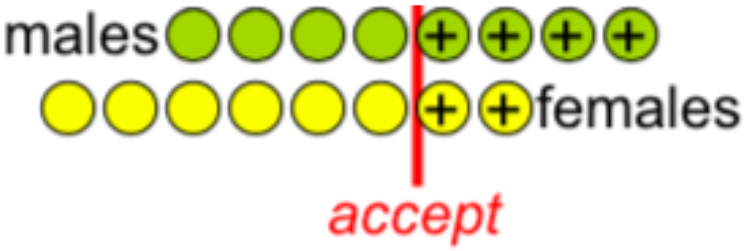
Measuring

p(+|M) = 50%
p(+|F) = 25%

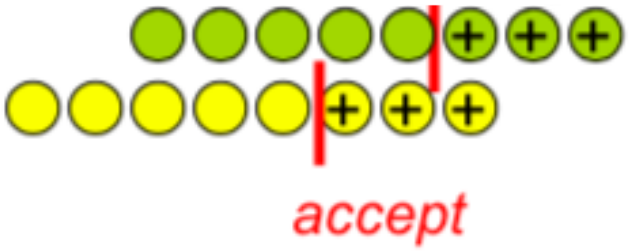


Measuring

p(+|M) = 50%
p(+|F) = 25%

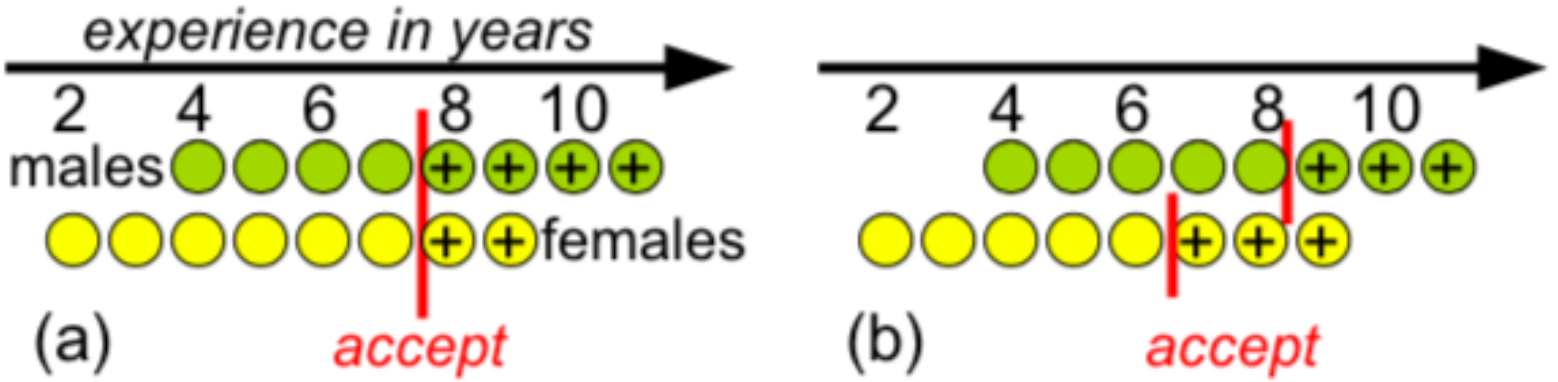


p(+|M) = 38%
p(+|F) = 38%



Measuring

Reverse discrimination



Measuring

p(+|1..3,M) = n.a.
p(+|1..3,F) = 0%

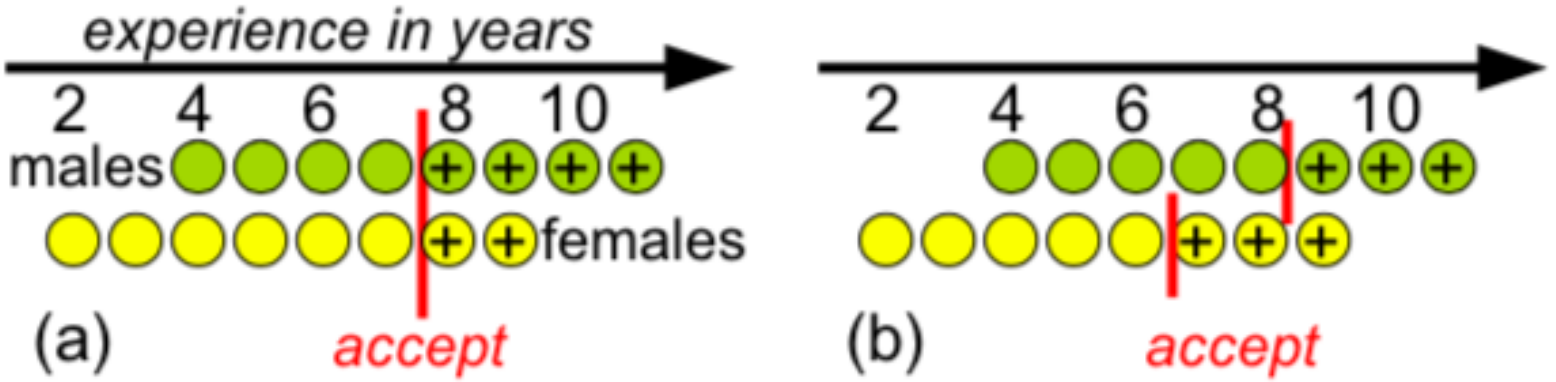
p(+|4..6,M) = 0%
p(+|4..6,F) = 0%

p(+|7..9,M) = 67%
p(+|7..9,F) = 67%

p(+|10..12,M) = 100%
p(+|10..12,F) = n.a.

Measuring

- “Twin test” / counterfactual fairness
  - p(+|X,M) = p(+|X,F)



- No “Redlining”
  - p(+|X,M) = p(+|X,F) = p(+|X) = “right”

Measuring

No discrimination

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	20%	20%	40%	40%
Accepted (+)	160	40	80	320

Kamiran et al 2013, KAIS

Measuring

No discrimination

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	20%	20%	40%	40%
Accepted (+)	160	40	80	320

Discrimination is present

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	15%	25%	35%	45%
Accepted (+)	120	50	70	360

How to correct?
What should be the acceptance rate?

Kamiran et al 2013, KAIS

Measuring

No discrimination

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	20%	20%	40%	40%
Accepted (+)	160	40	80	320

Is there discrimination, or not?

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	5%	5%	55%	55%
Accepted (+)	40	10	110	440

Redlining / indirect discrimination

Is there discrimination, or not?

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	5%	5%	55%	55%
Accepted (+)	40	10	110	440

2000
30%
600

Redlining / indirect discrimination

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	20%	20%	40%	40%
Accepted (+)	160	40	80	320

200 400

2000
30%
600

	medicine		computer	
	female	male	female	male
Number of applicants	800	200	200	800
Acceptance rate	5%	5%	55%	55%
Accepted (+)	40	10	110	440

50 550

2000
30%
600

New data, is there discrimination or not?

Is there discrimination, or not? Redlining?

	postal code I		postal code II	
	immigrant	native	immigrant	native
Number of applicants	800	200	200	800
Acceptance rate	5%	5%	55%	55%
Accepted (+)	40	10	110	440

Which is the “right” level?

Table 8.1 Summary statistics of the Adult dataset

	hours per week		annual income (K$)
female	36.4	5.8$/h	10.9
male	42.4	13.8$/h	30.4
all data	40.4	11.4$/h	23.9

Like male?

Like all?

Redistribution of the same total resources?

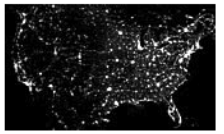
Kamiran and Žliobaitė 2013

Exercise 3

- Adult data
  - Download the data from UCI repository https://archive.ics.uci.edu/ml/datasets/Adult
  - Measure the percentage of high income for male and female
  - Use hours per week as the explanatory attribute, measure again
  - Duplicate instances: take one high income female that works fewest hours per week and one low income male that works most hours per week, duplicate 10, 100, 1000 times. What happens to the measured discrimination?
  - Optional: cluster data on all explanatory attributes. Measure discrimination within each cluster with respect to gender.

Adult Data Set

Download: Data Folder, Data Set Description



Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	1560943

Source:

Donor:

Ronny Kohavi and Barry Becker
Data Mining and Visualization
Silicon Graphics.
e-mail: ronnyk '@' live.com for questions.

Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:

>50K, <=50K.

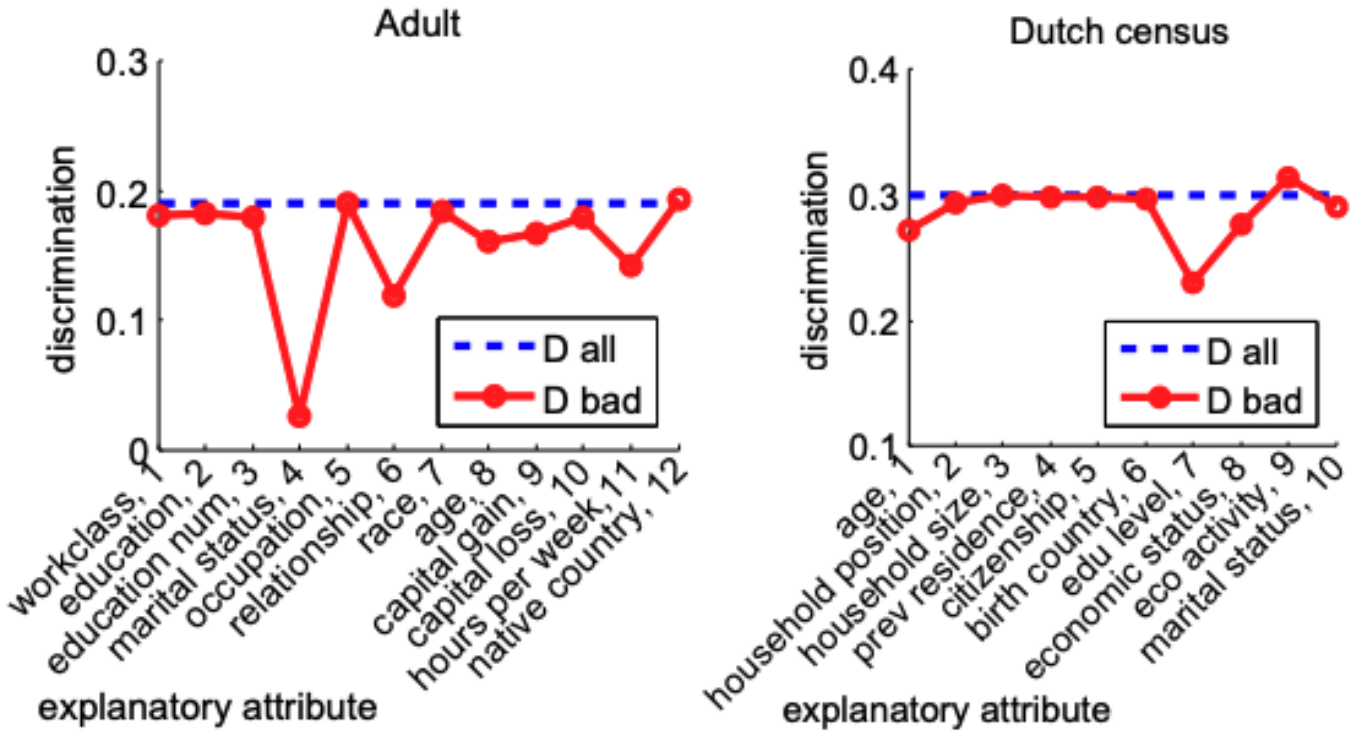
age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Relevant Papers:

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996
[Web Link]

Measuring indirect discrimination in practice

- Many potential explanatory variables, all correlated



Kamiran et al 2013

See literature for more on measuring

- Starting points:
  - Kamiran, F., Žliobaitė, I. and Calders, T. (2013).Quantifying explainable discrimination and removing illegal discrimination in automated decision making.*Knowledge and Information Systems* 35(3), p. 613-644.
  - Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery 31(4), 1060-1089.

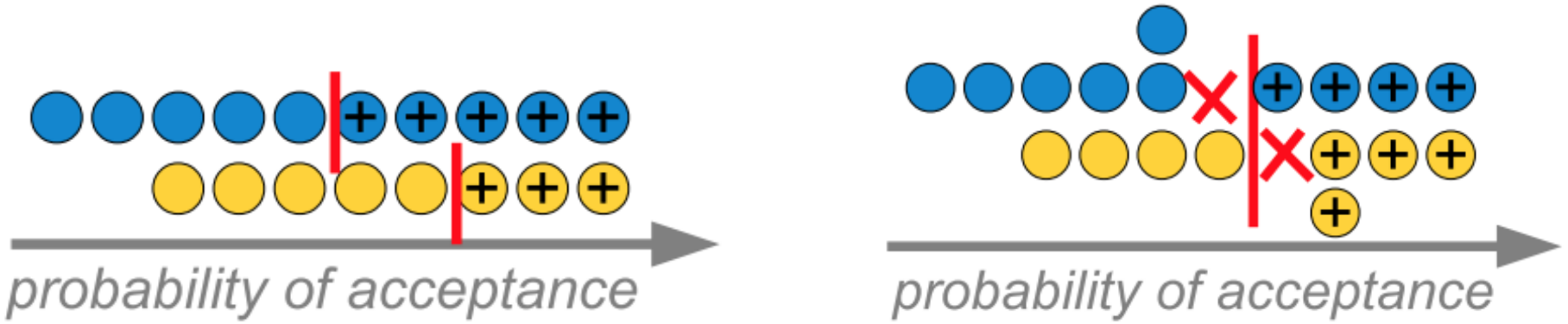
- We can measure on model outputs the same way we measure on data

Solutions

Algorithmic solutions for prevention

- Preprocessing
  - Modify training data (inputs, protected or outcomes)
  - Resample training data
- Postprocessing
  - Modify models
  - Modify predictions
- Optimization with fairness constraints

Preferential sampling



Kamiran and Calders 2010

Postprocessing

- Suppose historically salary is decided as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Data scientists assumes

$$salary = b_0 + b_1 \times education$$

- Observes data

education	*ethnicity*	*salary*
1	1	600
2	1	700
3	1	800
4	1	900
10	1	1500

education	*ethnicity*	*salary*
1	0	1100
6	0	1600
7	0	1700
9	0	1900
10	0	2000

- Learns model

$$salary = 602 + 128 \times education$$

Lower base salary,
higher reward for education
the model punishes
ethnical minorities

Žliobaitė and Custers 2016

Postprocessing

- Suppose historically salary is decided as

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- If data scientist observes full data
- Learns a complete model including ethnicity
- The sensitive component can be easily removed

$$salary = 1000 + 100 \times education - 500 \times ethnicity$$

- Decision making

$$salary = 1000 + 100 \times education$$

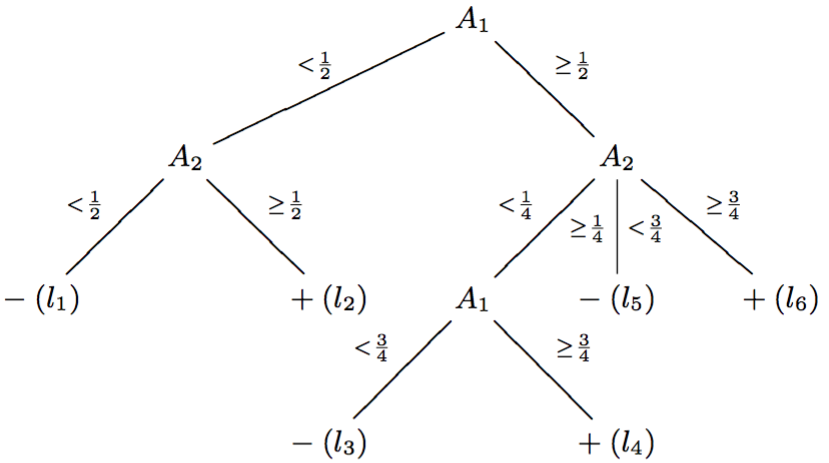
Žliobaitė and Custers 2016

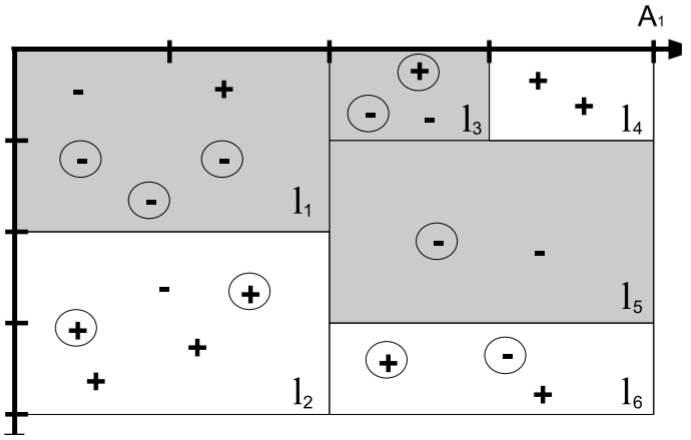
Postprocessing

- Relabel tree leaves to remove

the most discrimination with

the least damage to the accuracy





Kamiran et a 2010

Optimization with constraints

- Decision tree

Regular tree induction

$$IGC := H_{Class}(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_{Class}(D_i)$$

Data subset due to split

Entropy wrt class label

Discrimination-aware tree induction

$$IGS := H_B(D) - \sum_{i=1}^{k} \frac{|D_i|}{|D|} H_B(D_i)$$

Entropy wrt protected characteristic

Tree splits are decided on: IGC - IGS

Kamiran et a 2010

Legal and societal criticisms

- Preprocessing
  - Modify input data X, s or y
  - Resample input data

- Postprocessing
  - Modify models
  - Modify outputs

- Learning with constraints

From the legal perspective

- Decision manipulation – very bad
- Data manipulation – quite bad
- Learning with constraints - ok
  - Protected characteristic should not be used in decision making

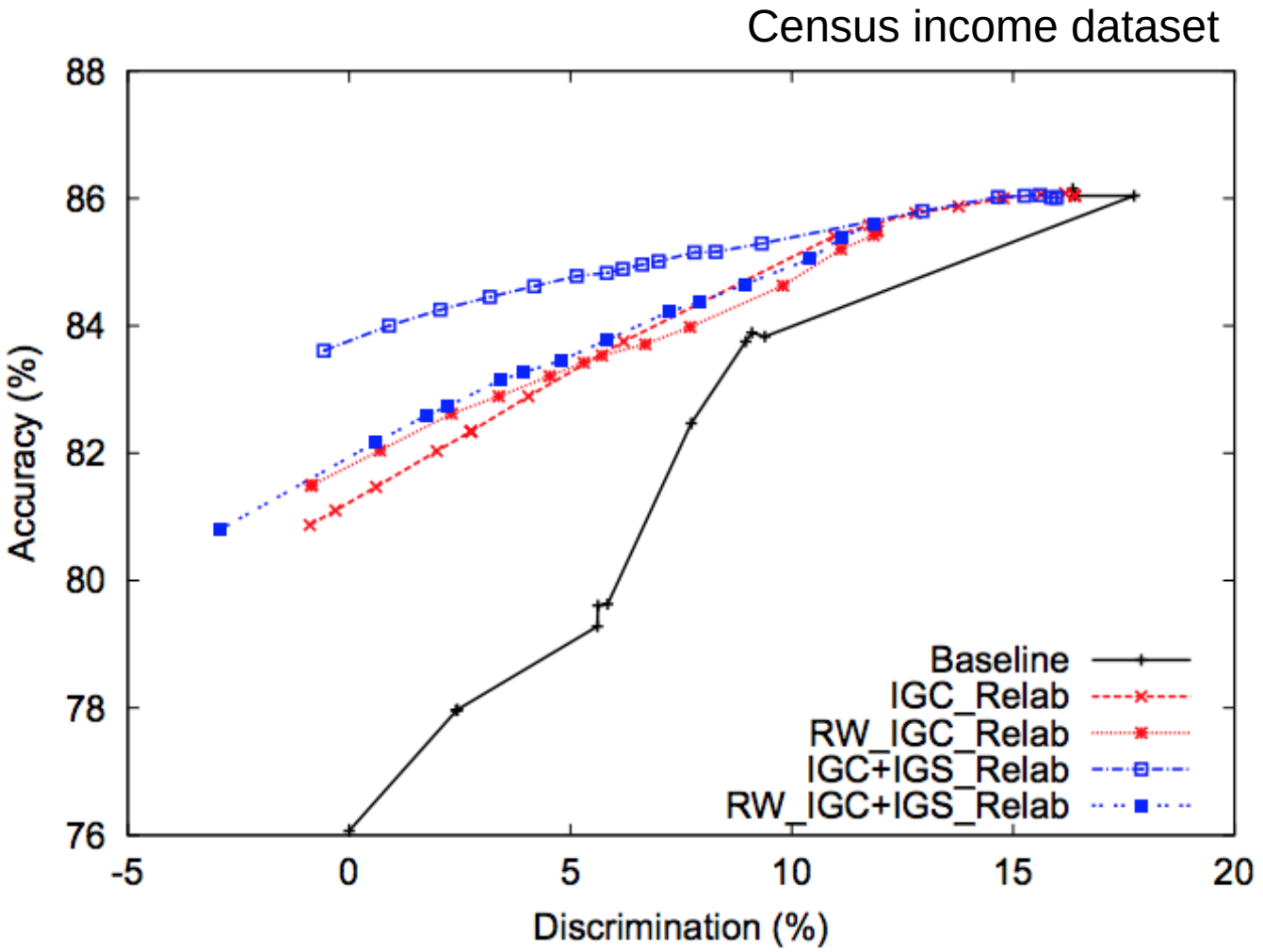
Comparison of solutions

Suppose we know what we want

How to select which solution is better?

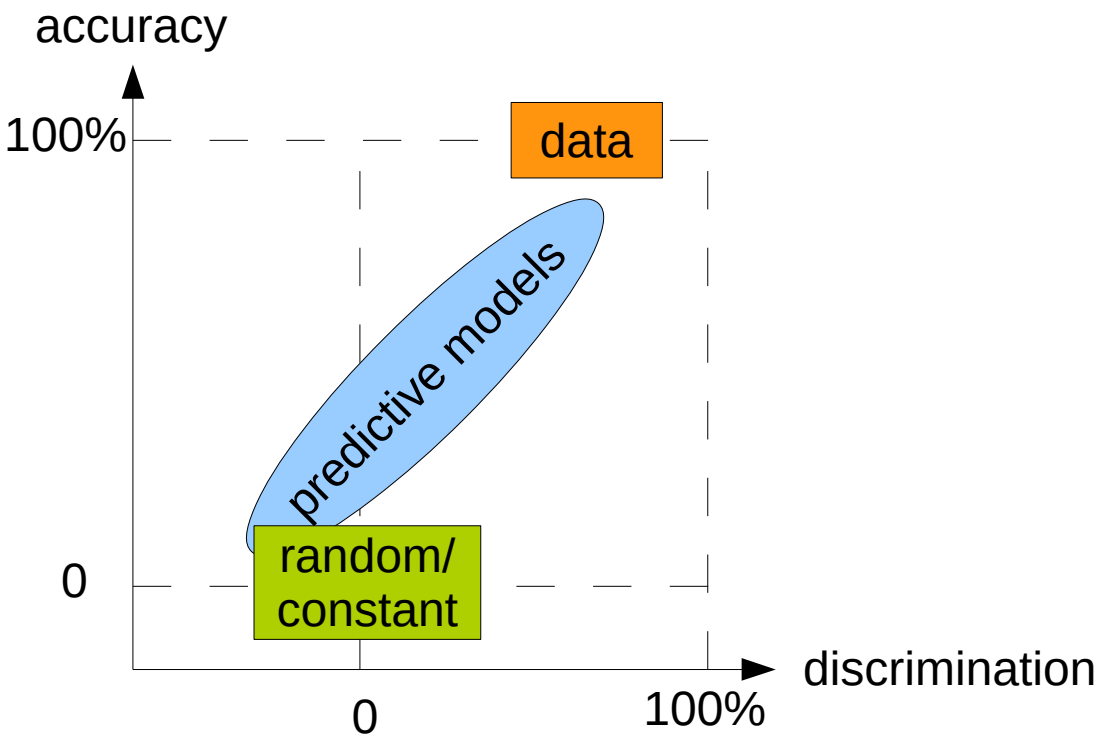
http://aequitas.dssg.io/

Accuracy vs. discrimination

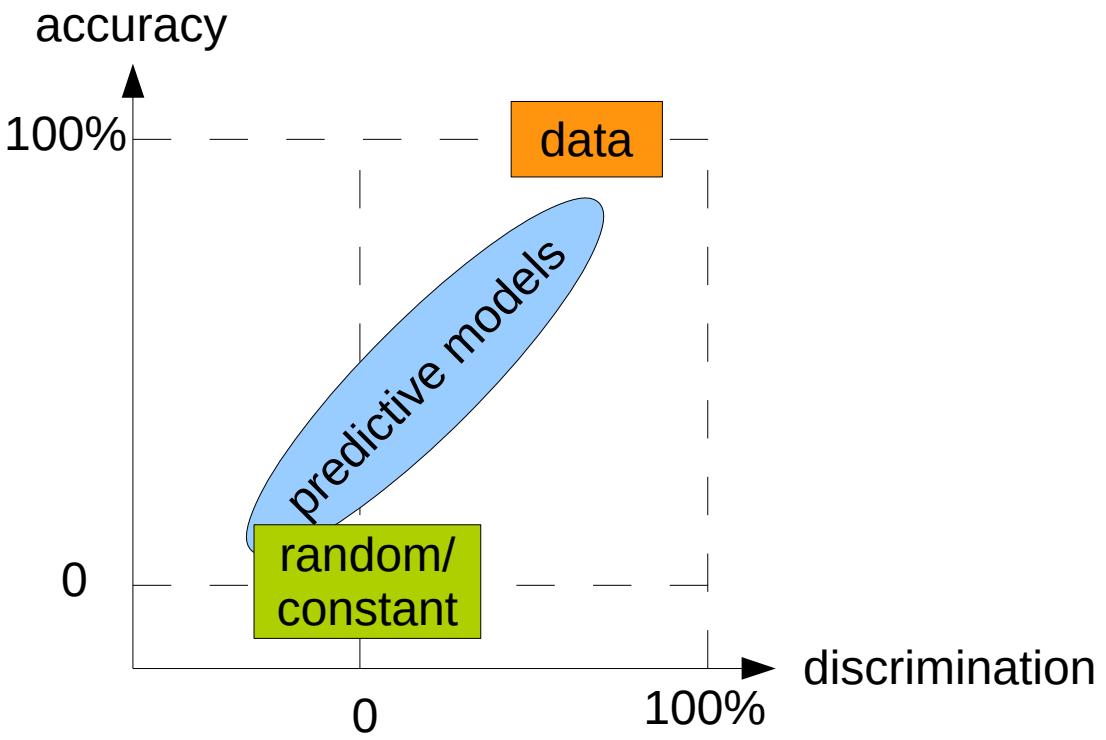


Kamiran et al 2010

Testing on discriminatory data?



Testing on discriminatory data?

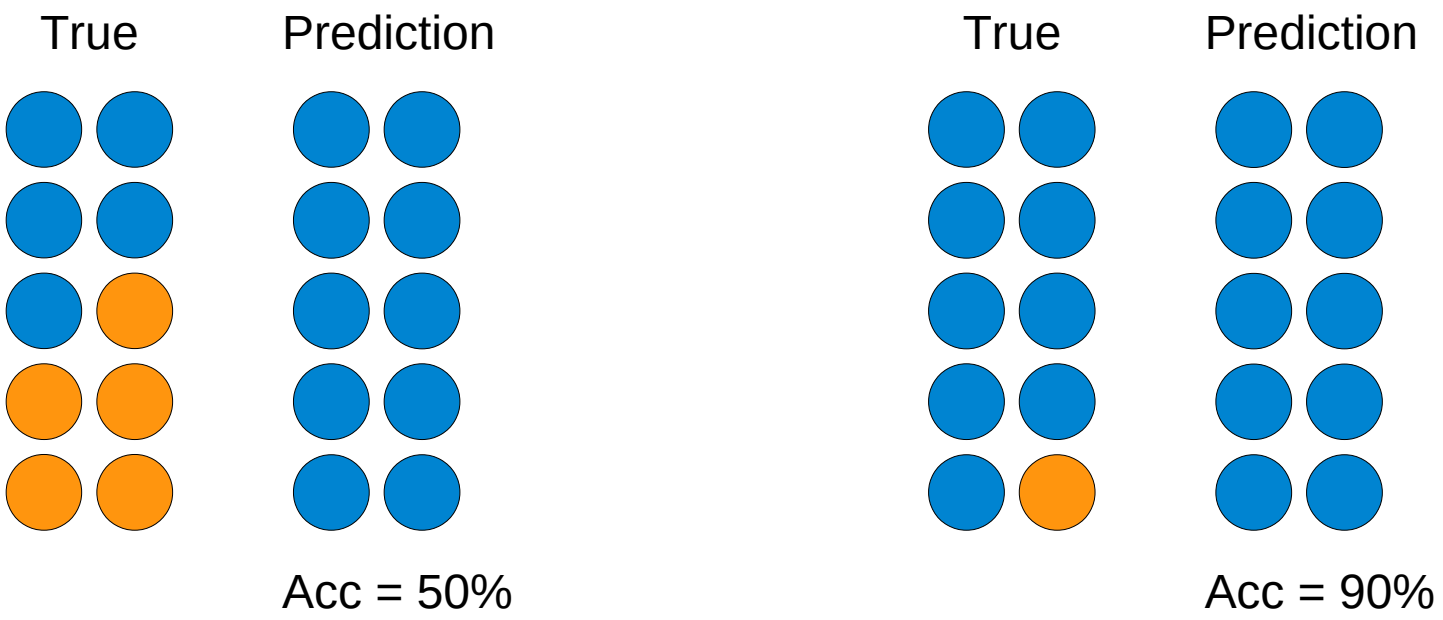


FRA
EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

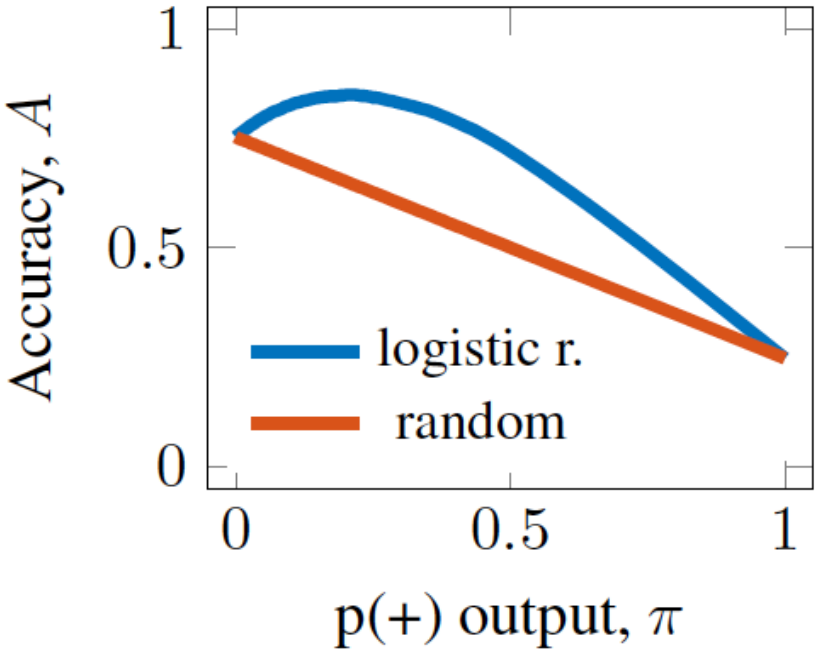
Shouldn't removing discrimination improve accuracy?

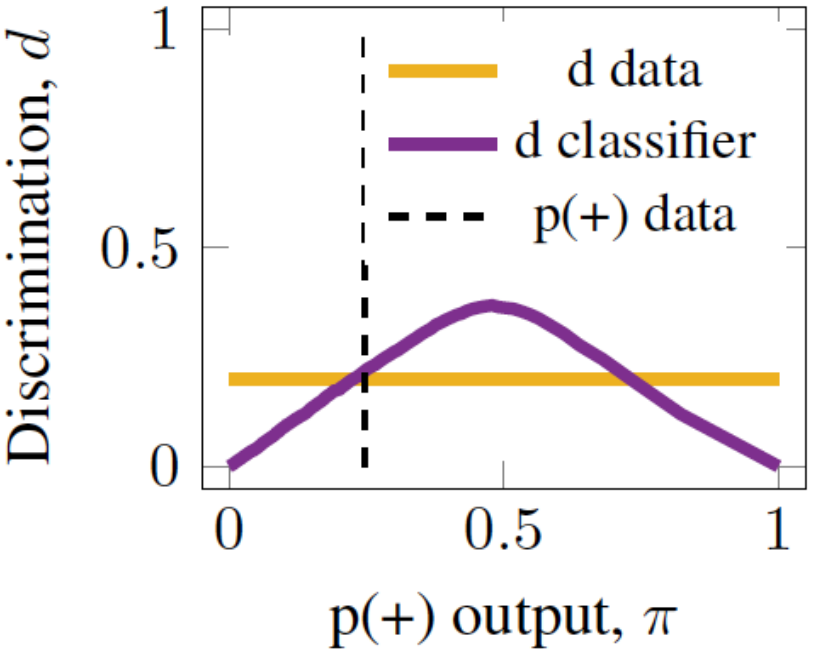
Decision threshold vs. discrimination

- Baseline accuracy and baseline discrimination varies with varying overall acceptance rate
- Classifiers with different overall acceptance rates are not comparable



Experiment





Adult dataset from UCI

Discrimination-accuracy tradeoffs

	Acc. A	Disc. d
Data/oracle	100	19.9
Logistic with s	84.9	18.3
Logistic no s	84.9	17.6
Logistic massage	83.5	6.9
NB with s	81.9	13.5
NB no s	81.4	10.9
NB massaged	81.5	6.8
Tree J48 with s	85.1	17.9
Tree J48 no s	85.0	17.9
Tree massage	83.5	6.1

Removing s does not solve the problem, not much accuracy is lost either

Žliobaitė 2015

Discrimination-accuracy tradeoffs

	p(+) π	Acc. A	Disc. d
Data/oracle	24.7	100	19.9
Logistic with s	20.2	84.9	18.3
Logistic no s	20.1	84.9	17.6
Logistic massage	22.1	83.5	6.9
NB with s	15.4	81.9	13.5
NB no s	14.4	81.4	10.9
NB massaged	15.4	81.5	6.8
Tree J48 with s	19.6	85.1	17.9
Tree J48 no s	19.6	85.0	17.9
Tree massage	22.9	83.5	6.1

Decreasing acceptance rates may show lower nominal discrimination!

Žliobaitė 2015

Preferential treatment is a ranking problem

- No discrimination baseline: random order
- Maximum discrimination: all members of the favored community go before all the members of the protected community

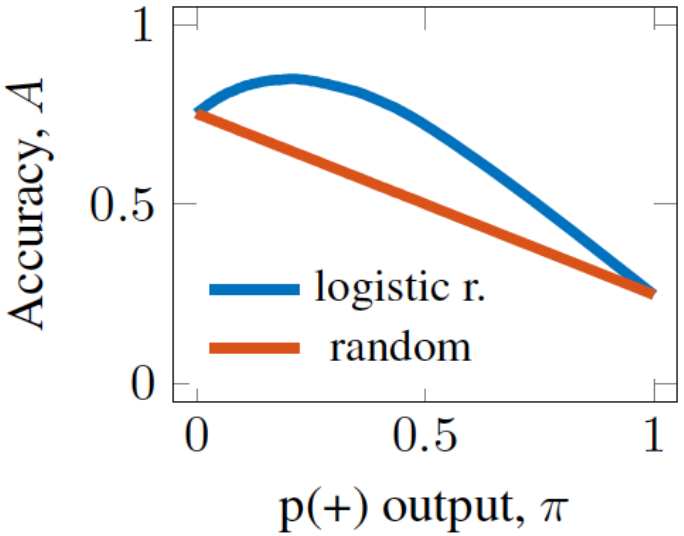
Žliobaitė 2015, discrimination-accuracy trade-offs

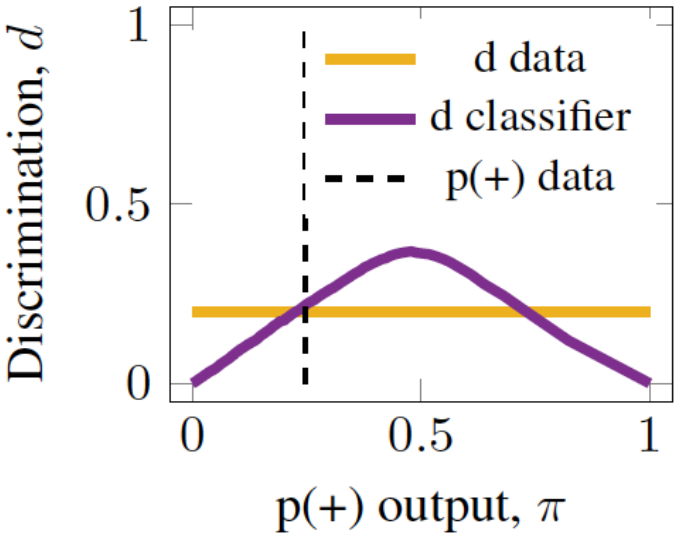
Normalized measures

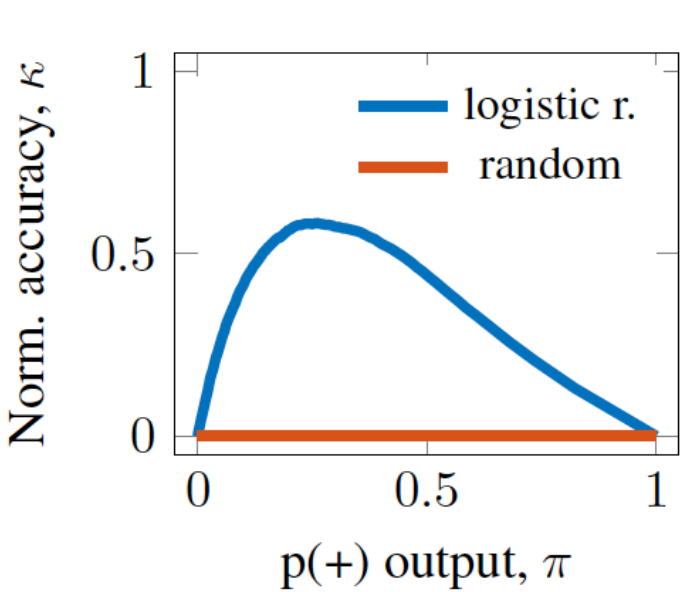
- We propose to normalize discrimination by dmax
  - d = D/dmax, where $d_{max} = \min\left(\frac{\pi}{\alpha}, \frac{1-\pi}{1-\alpha}\right)$ ← acceptance rate p(+); ← proportion of natives p(native)
    - 1 max, 0 no discrimination, <0 reverse discrimination

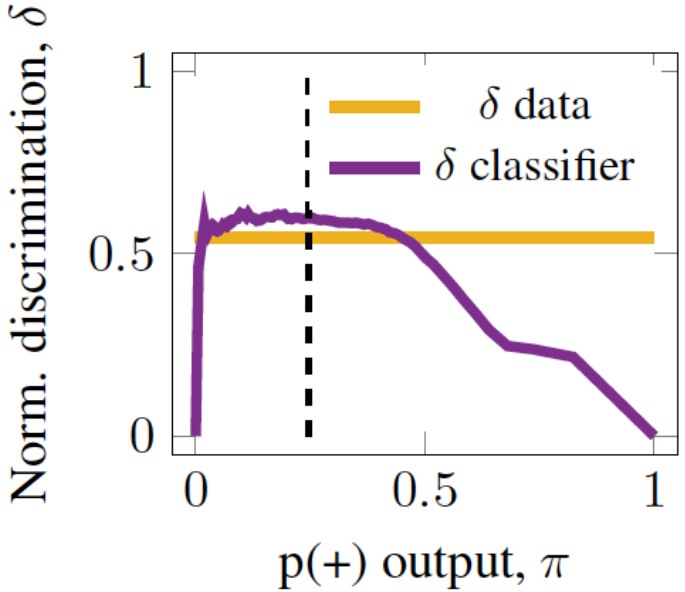
- We recommend normalizing accuracy – Cohen's kappa
  - k = (Acc – RAcc) / (1 – Racc)
    - 1 max, 0 like random, <0 very bad

Experiment cont.









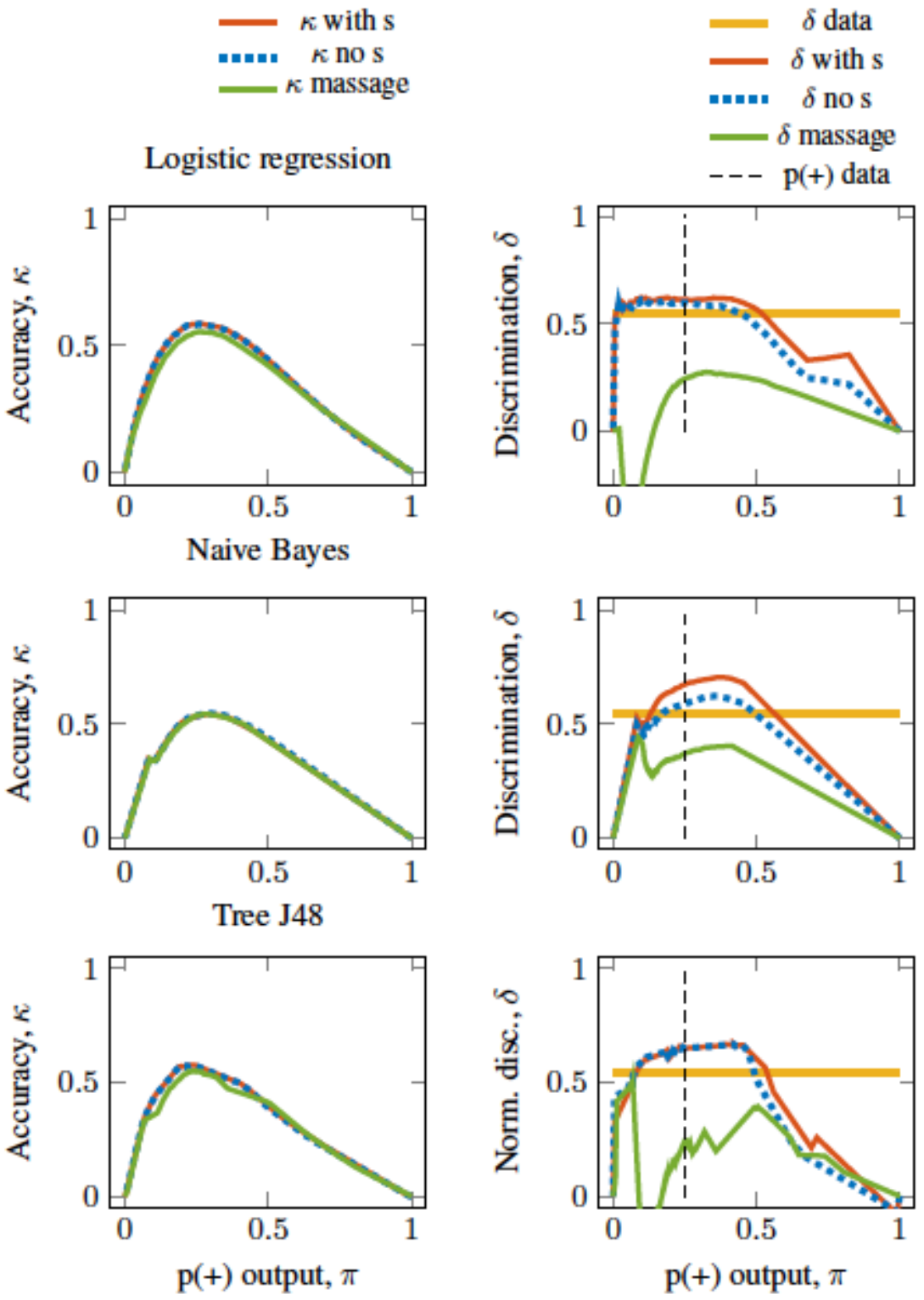
Discrimination-accuracy tradeoffs

	p(+) π	Acc. A	Disc. d	N. disc. δ
Data/oracle	24.7	100	19.9	54.4
Logistic with s	20.2	84.9	18.3	**61.4**
Logistic no s	20.1	84.9	17.6	**59.6**
Logistic massage	22.1	83.5	6.9	21.3
NB with s	15.4	81.9	13.5	**59.7**
NB no s	14.4	81.4	10.9	51.3
NB massaged	15.4	81.5	6.8	29.7
Tree J48 with s	19.6	85.1	17.9	**61.9**
Tree J48 no s	19.6	85.0	17.9	**61.8**
Tree massage	22.9	83.5	6.1	18.1

Decreasing acceptance rates may show lower nominal discrimination!
But not the real underlying discrimination

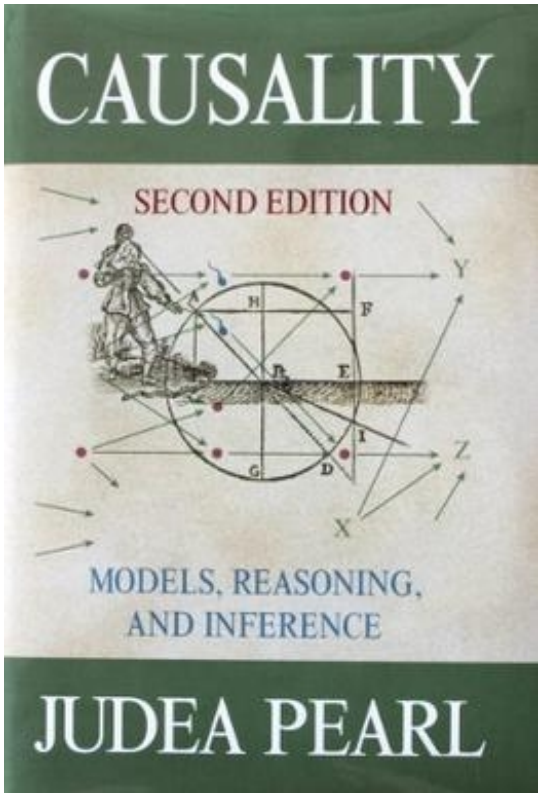
Žliobaitė 2015

Massaging



Žliobaitė 2015

Testing on discriminatory data?

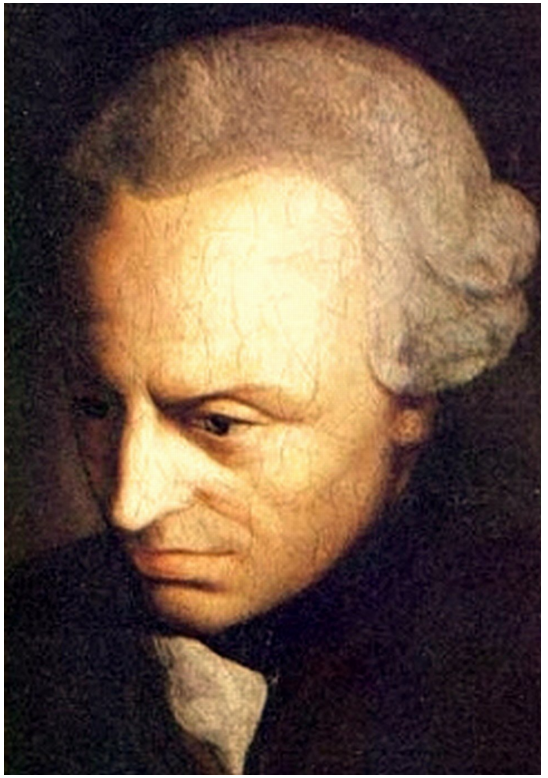


FRA
EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS

Shouldn't removing discrimination improve accuracy?

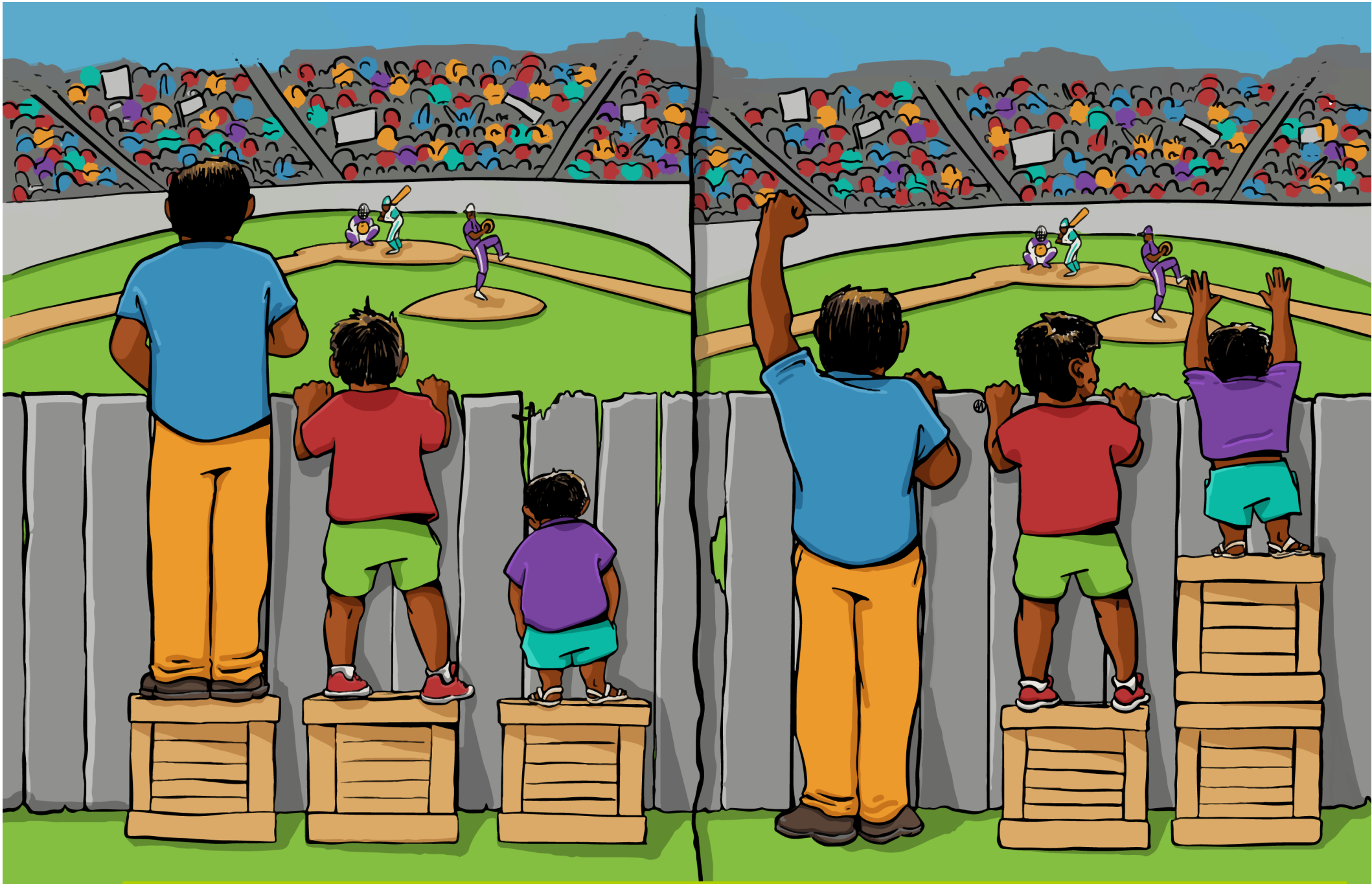
There are no “right” or “wrong” variables

Morality is a social convention?



Immanuel Kant introduced the categorical imperative:
"Act only according to that maxim whereby you can,
at the same time, will that it should become a universal law"

Source: Wikimedia Commons, Unidentified painter 18th-century portrait painting of men



There are no “right” or “wrong” measures

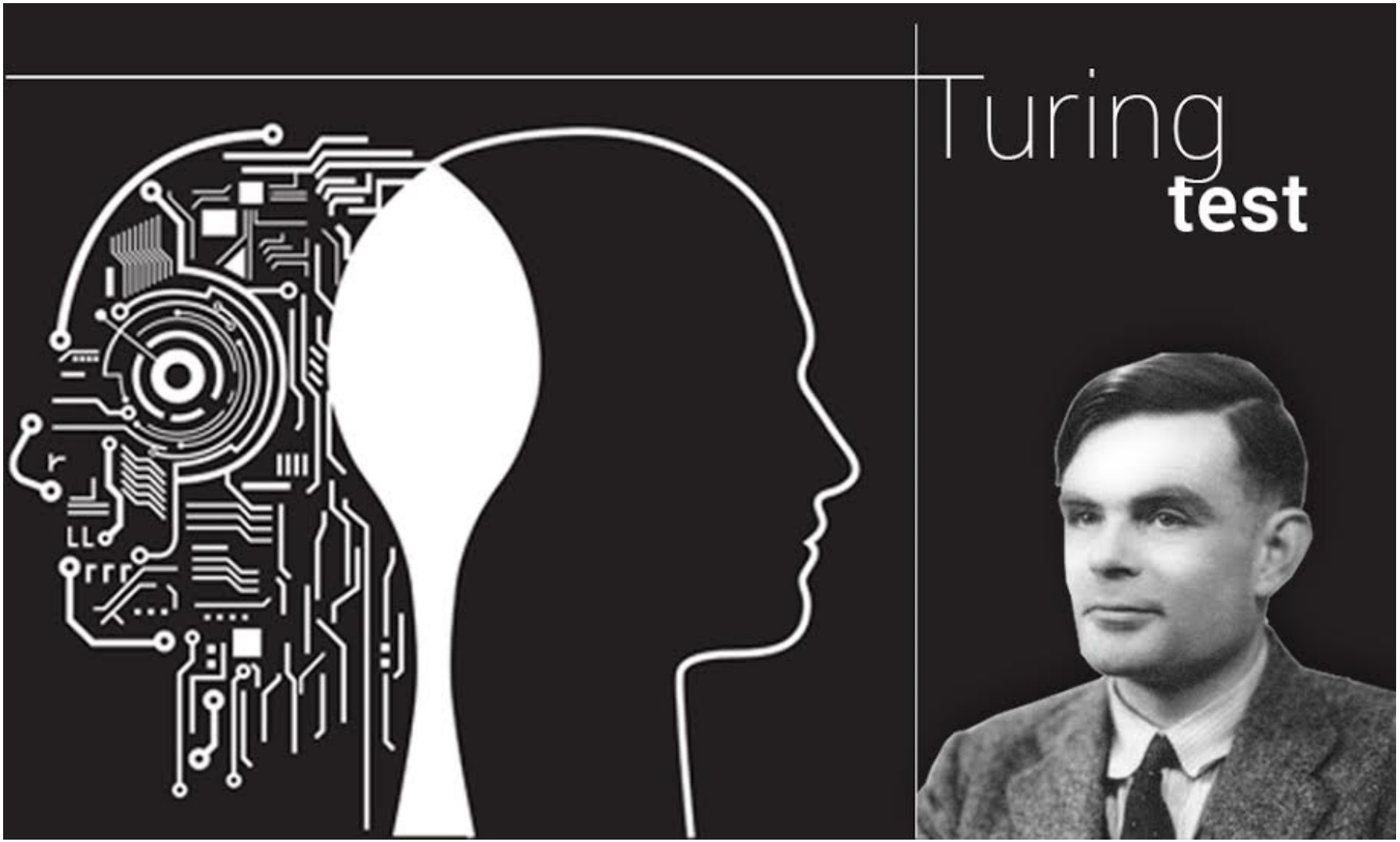
Interaction Institute for Social Change | Artist: Angus Maguire

Computer can extract patterns
`but their interpretation is up to humans

Source:
https://www.outoffocus.org/outoffocus/2013/09/21-10thbirthdaypost-42.html

Source:
https://www.kseniaanske.com/blog/2012/12/2/douglas-adams-the-man-on-the-bicycle.html

Machine intelligence?



Strong AI – machine consciousness and mind

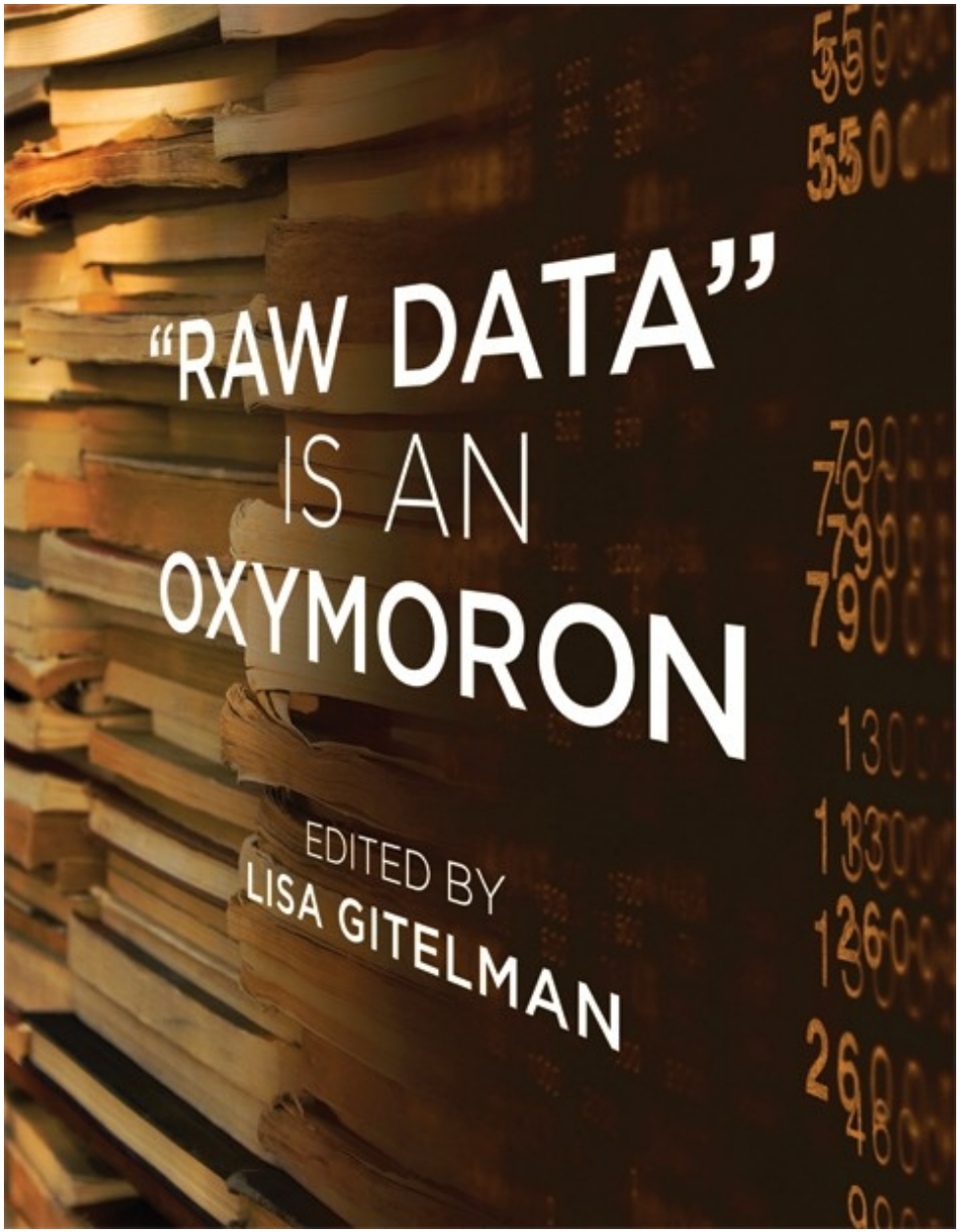
Image source: https://analyticsindiamag.com/turing-test-key-contribution-field-artificial-intelligence/

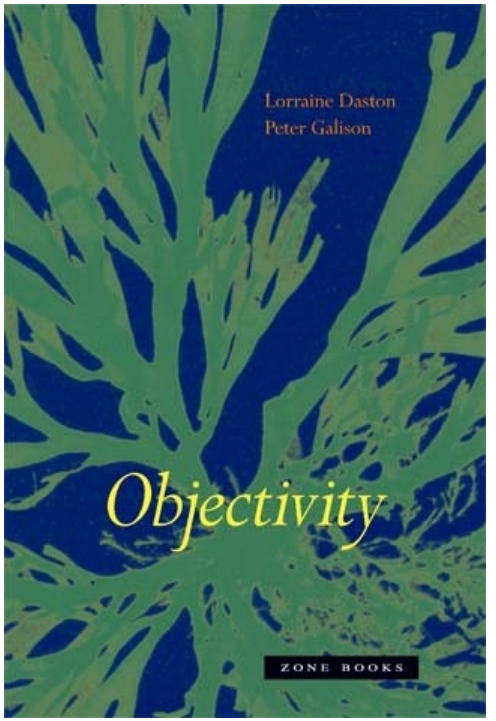
Weak AI – focused on one narrow task

Humans need to tell what is the task

no genuine intelligence, no self-awareness, no life

Image source: http://www.trustedreviews.com/news/apple-s-next-gen-siri-app-could-blow-alexa-and-co-away-2943759
https://www.translinguoglobal.com/10-reasons-why-google-translate-is-not-better-than-learning-language/


"RAW DATA"
IS AN
OXYMORON
EDITED BY
LISA GITELMAN


Lorraine Daston
Peter Galison
Objectivity
ZONE BOOKS