Responsible Data Science

Prof. Julia Stoyanovich

Computer Science and Engineering &
Center for Data Science
New York University

@stoyanoj

NYU

Online price discrimination

THE WALL STREET JOURNAL.

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES,
JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

WHAT PRICE WOULD YOU SEE?

lower prices offered to buyers who live in more affluent neighborhoods

https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

Online job ads

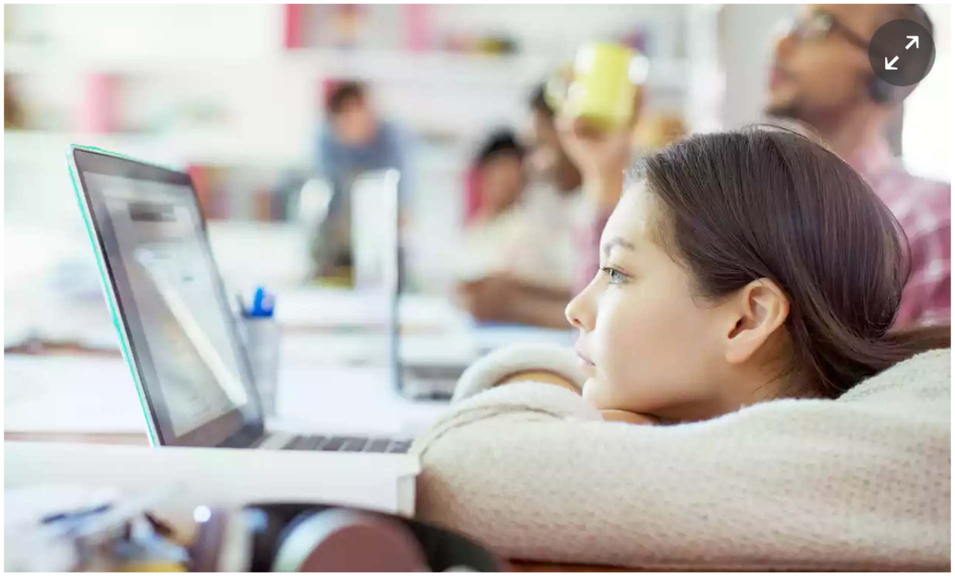
theguardian

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



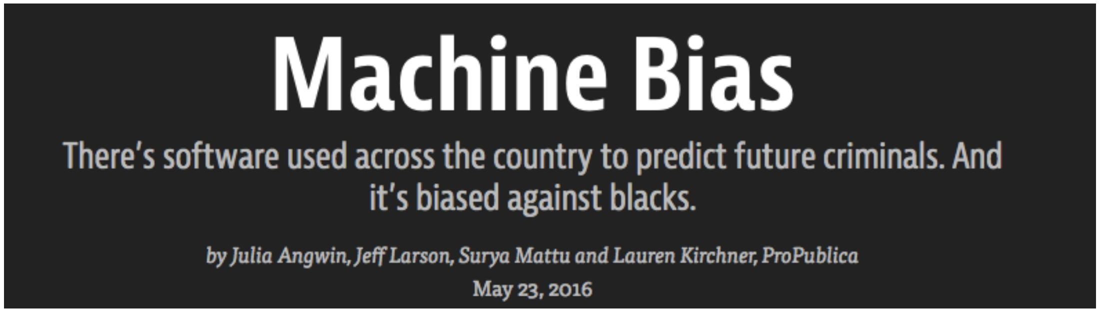
One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for "$200k+" executive jobs **1,852 times to the male group and only 318 times to the female group**. Another experiment, in July 2014, showed a similar trend but was not statistically significant.

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

Racial bias in criminal sentencing


Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

A commercial tool COMPAS automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

The tool correctly predicts recidivism **61% of the time.**

Blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend.

The tool makes **the opposite mistake among whites**: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Is data science impartial?

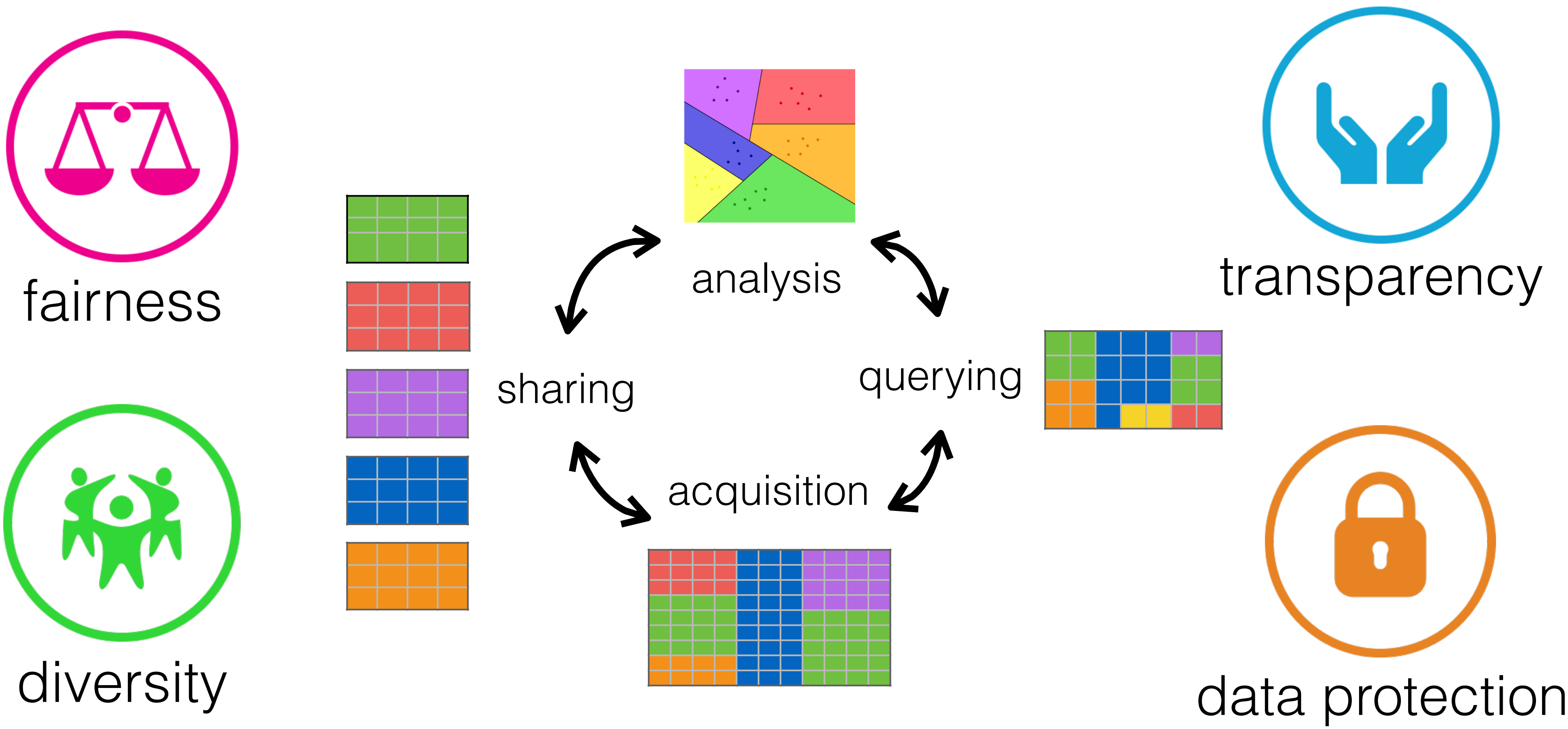
Data science is algorithmic, therefore it cannot be biased! And yet…

- All traditional evils of **discrimination**, and many new ones, exhibit themselves in the data science eco system
- **Bias** that is inherent in the data or in the process, and that is often due to systemic discrimination, is propelled and amplified
- **Transparency** helps prevent discrimination, enable public debate, establish **trust**
- Technology alone won't do: also need **policy**, **user involvement** and **education**

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

Data, responsibly

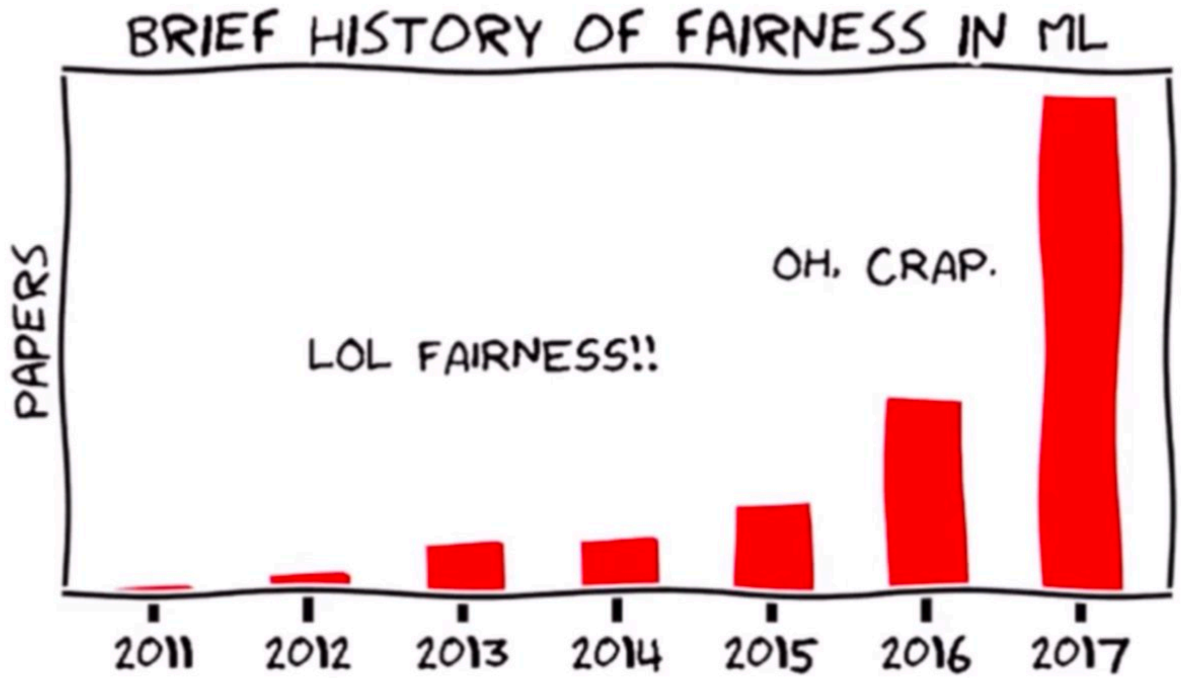
Because of its **power**, data science must be used **responsibly**



… with a holistic view of the **lifecycle**

Fairness

Fairness in ML



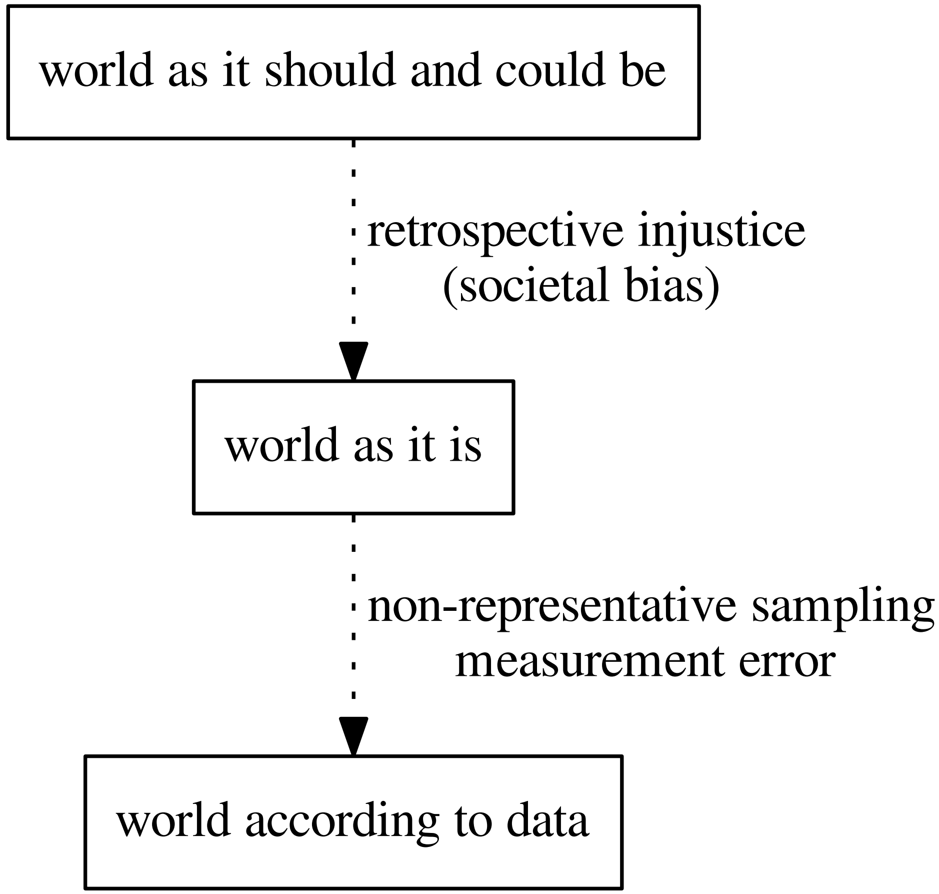
by Moritz Hardt

Fairness is lack of "bias"

- What are the tasks we are interested in?
  - predictive analytics
- What do we mean by **bias**?
  - **statistical bias**: a model is biased if it doesn't summarize the data correctly
  - **societal bias**: a dataset or a model is biased if it does not represent the world "correctly", e.g., data is not representative, there is measurement error, or the **world is "incorrect"**

the world as it is or as it should be?

"Biased data"



from "Prediction-Based Decisions and Fairness" by Mitchell, Potash and Barocas, 2018

when data is about people, bias can lead to discrimination

The evils of discrimination

Disparate treatment is the illegal practice of treating an entity, such as a creditor or employee, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

Disparate impact is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

http://www.allenovery.com/publications/en-gb/Pages/Protected-characteristics-and-the-perception-reality-gap.aspx

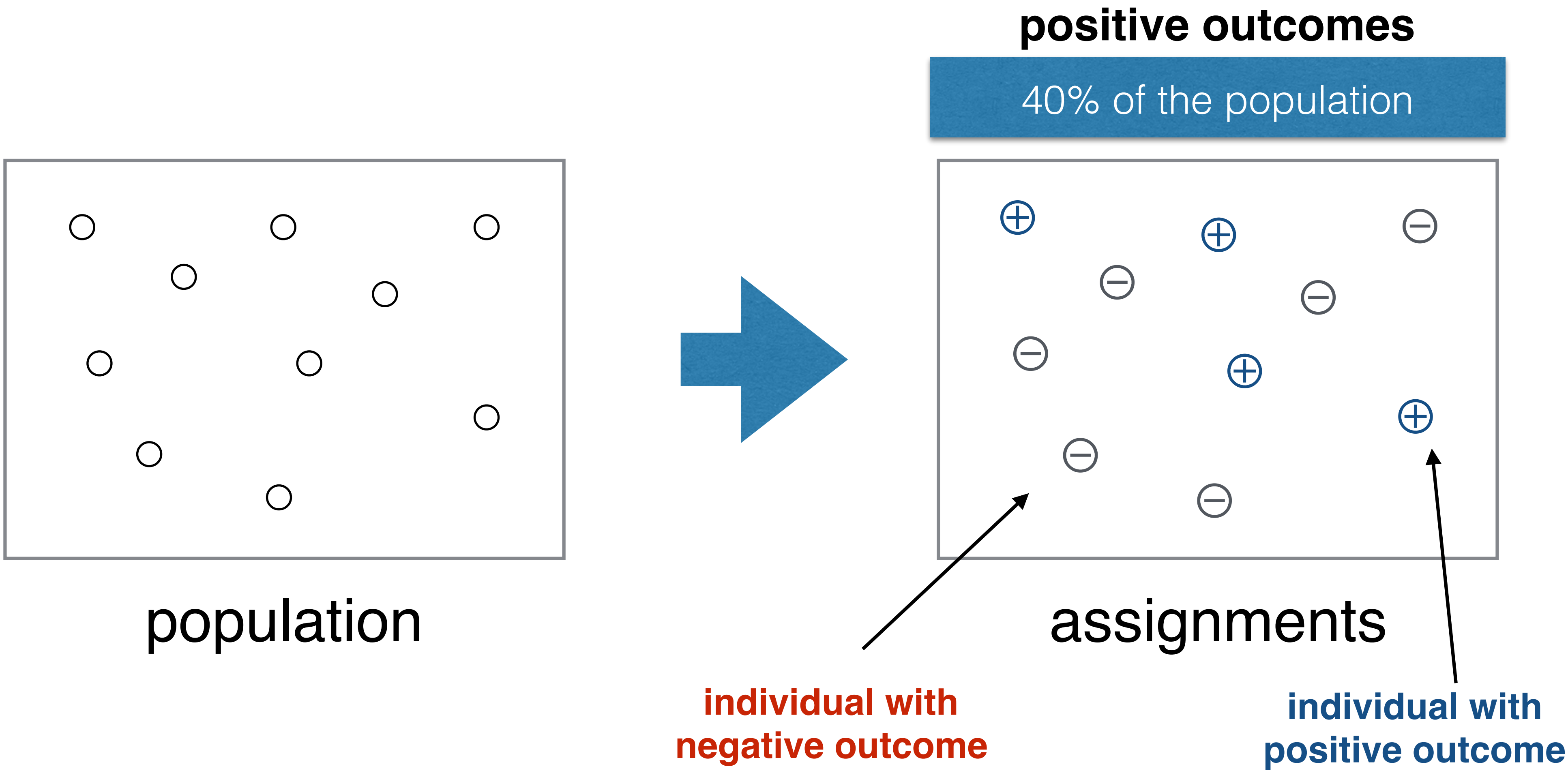
Vendors and outcomes

Consider a **vendor** assigning positive or negative **outcomes** to individuals.

Positive Outcomes	Negative Outcomes
offered employment	denied employment
accepted to school	rejected from school
offered a loan	denied a loan
offered a discount	not offered a discount

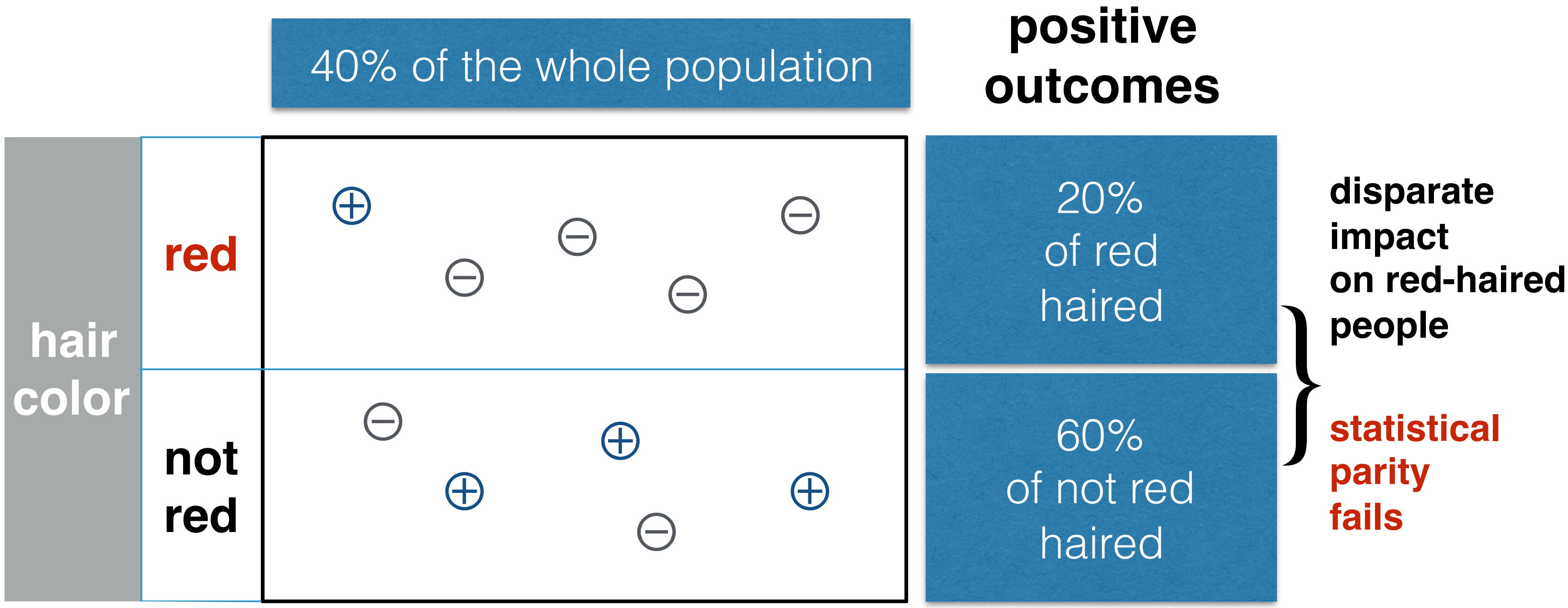
Assigning outcomes to populations

Fairness is concerned with how outcomes are assigned to a population



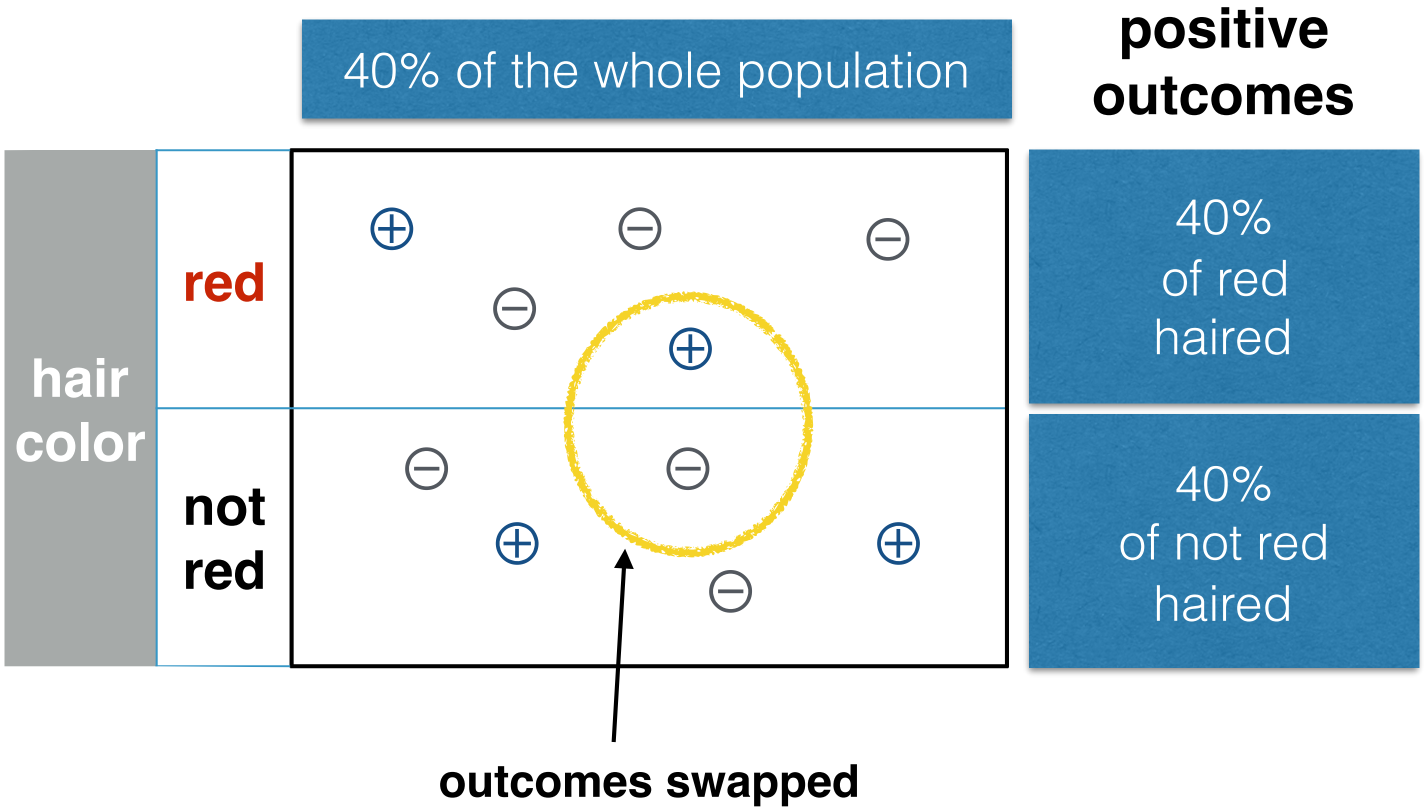
Sub-populations may be treated differently

Sub-population: those with red hair
(under the same assignment of outcomes)



Statistical parity

Statistical parity (a popular **group fairness** measure)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Redundant encoding

Now consider the assignments under both
hair color (protected) and **hair length** (innocuous)

		hair length		positive outcomes
		long	not long	
hair color	red	⊕	⊖ ⊖ ⊖ ⊖	20% of red haired
	not red	⊕ ⊕ ⊕	⊖ ⊖	60% of not red haired

Deniability
The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.

Blinding is not an excuse

Removing **hair color** from the vendor's assignment process does not prevent discrimination!

		hair length	
		long	**not long**
hair color	**red**	⊕	⊖ ⊖ ⊖ ⊖
	not red	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

20% of red haired

60% of not red haired

Assessing disparate impact
Discrimination is assessed by the <u>effect</u> on the protected sub-population, not by the input or by the process that lead to the effect.

Redundant encoding

Let's replace hair color with **race** (protected),
hair length with **zip code** (innocuous)

		zip code	
		10025	**10027**
race	**black**	⊕	⊖ ⊖ ⊖ ⊖
	white	⊕ ⊕ ⊕	⊖ ⊖

positive outcomes

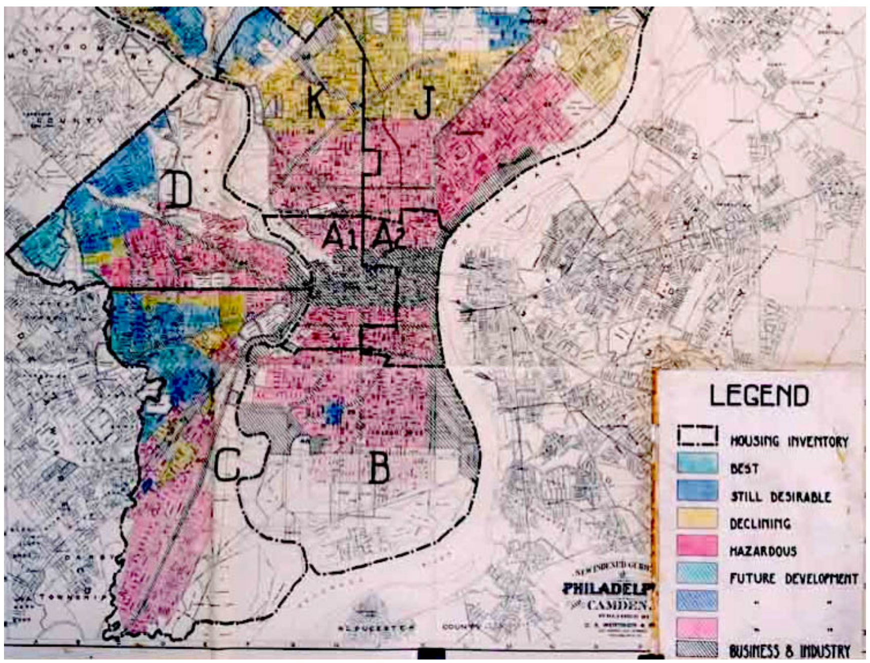
20%
of black

60%
of white

Redlining

Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor.

Philadelphia, 1936



wikipedia

Households and businesses in the red zones could not get mortgages or business loans.

Imposing statistical parity

May be contrary to the goals of the vendor

positive outcome: offered a loan

		credit score	
		good	**bad**
race	**black**	⊕	⊖ ⊖ ⊕ ⊖
	white	⊕ ⊕ ⊖	⊖ ⊖

positive outcomes

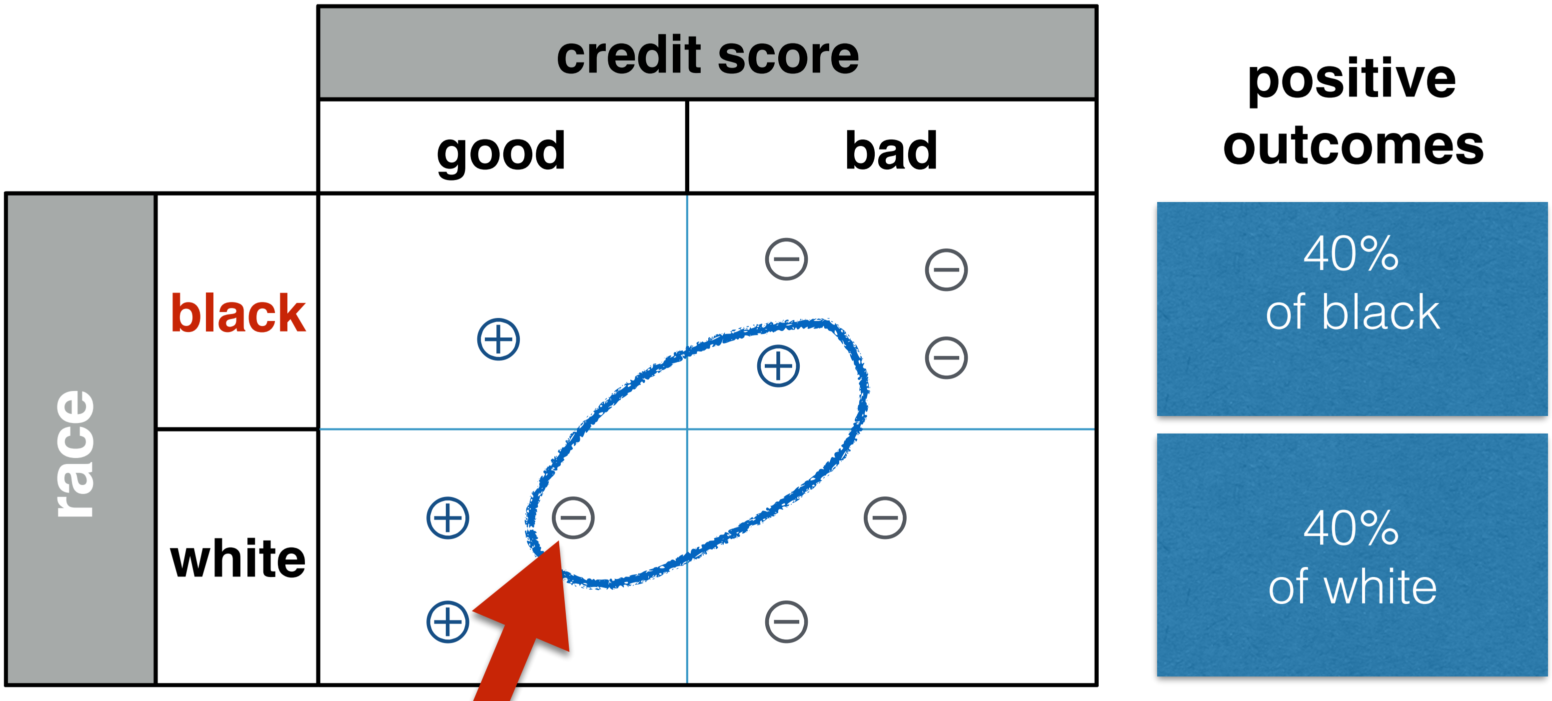
40% of black

40% of white

Impossible to predict loan payback accurately.
Use past information, which may itself be biased.

Is statistical parity sufficient?

Statistical parity (a popular **group fairness** measure)
demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Individual fairness
any two individuals who are similar w.r.t. a particular task should receive similar outcomes

Ricci v. DeStefano (2009)

Supreme Court Finds Bias Against White Firefighters

By **ADAM LIPTAK** JUNE 29, 2009 The New York Times

Case opinions	
Majority	Kennedy, joined by Roberts, Scalia, Thomas, Alito
Concurrence	Scalia
Concurrence	Alito, joined by Scalia, Thomas
Dissent	Ginsburg, joined by Stevens, Souter, Breyer
Laws applied	
Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e et seq.	

Karen Lee Torre, left, a lawyer who represented the New Haven firefighters in their lawsuit, with her clients Monday at the federal courthouse in New Haven. Christopher Capozziello for The New York Times

Two notions of fairness

individual fairness

equality

group fairness

equity

two intrinsically different world views

On the (im)possibility of fairness

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Goal: tease out the difference between beliefs about fairness and mechanisms that logically follow from those beliefs.

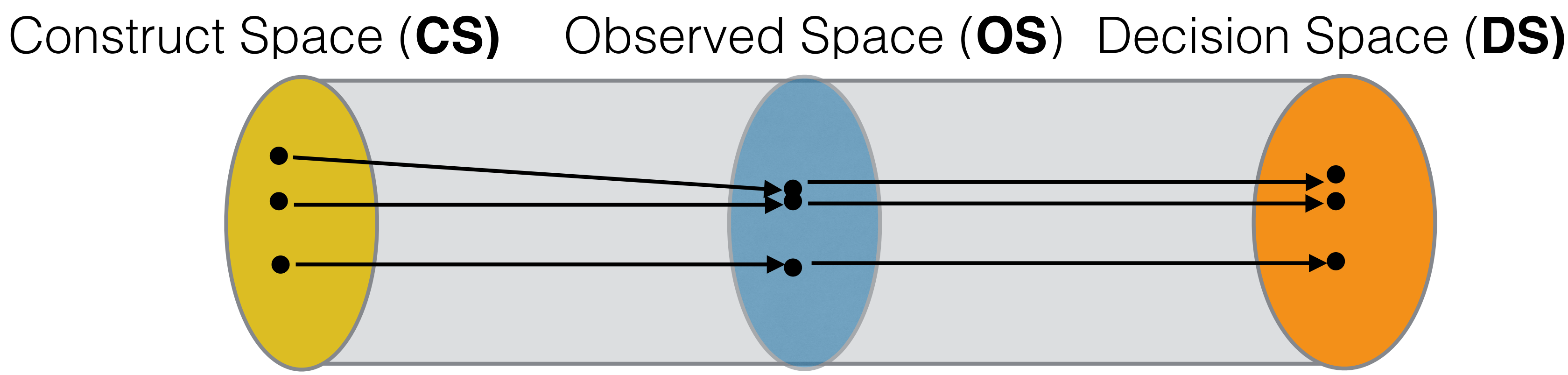
Construct Space (CS)	Observed Space (OS)	Decision Space (DS)
intelligence	SAT score	performance in college
grit	high-school GPA	
propensity to commit crime	family history	recidivism
risk-averseness	age	

Fairness through mappings

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]

Fairness: a mapping from **CS** to **DS** is **(ε, ε')**-fair if two objects that are no further than **ε** in **CS** map to objects that are no further than **ε'** in **DS**.

$$f : CS \rightarrow DS \qquad\qquad d_{CS}(x,y) < \varepsilon \Rightarrow d_{DS}(f(x), f(y)) < \varepsilon'$$



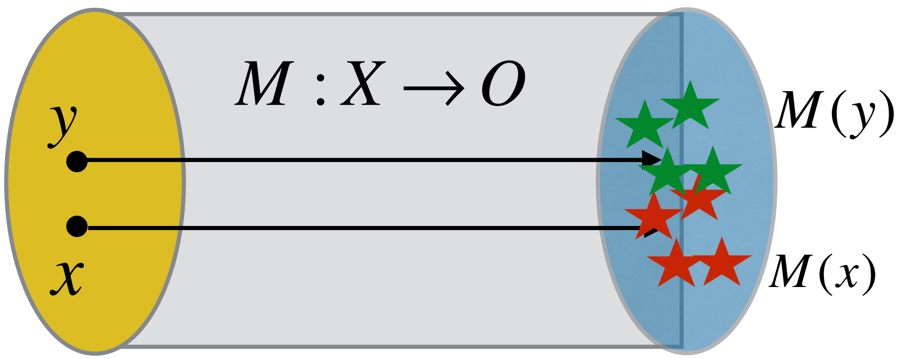
Individual fairness: The mapping from **CS** to **OS** has low distortion. That is, OS faithfully represents CS. (**WYSIWYG**)

Group fairness: The mapping from **CS** to **OS** has **structural bias,** a distortion that aligns with group structure of **CS**. (**WAE**)

Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]

Fairness: Individuals who are **similar** for the purpose of classification task should be **treated similarly**.



X individuals ***O*** outcomes

A task-specific similarity metric is given $d(x,y)$

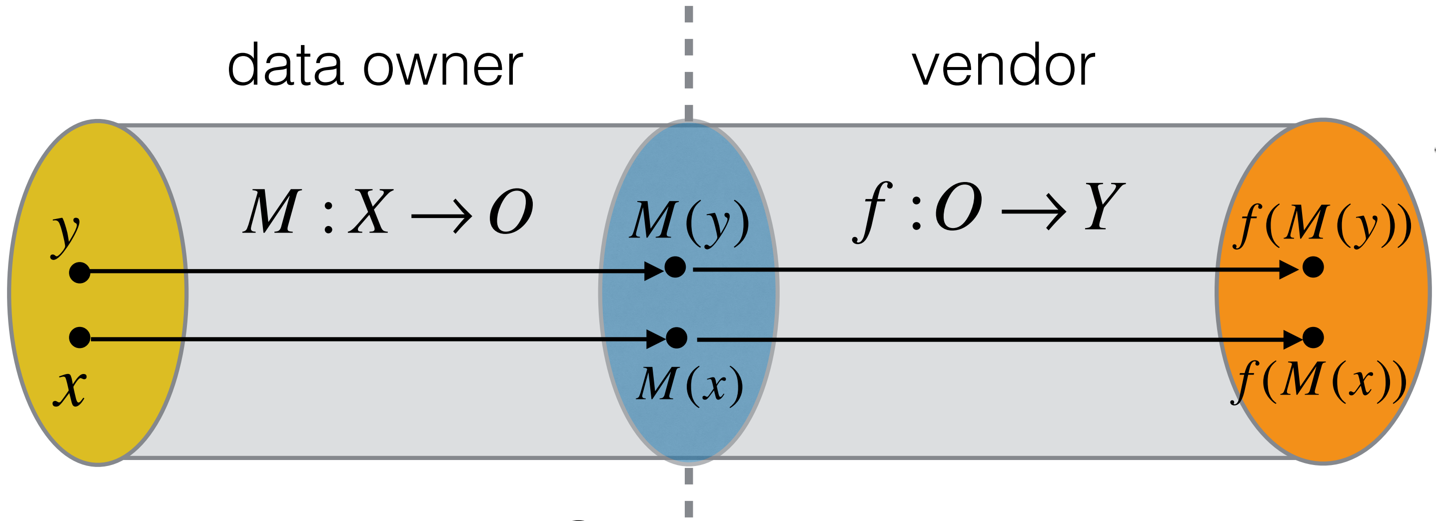
$M : X \rightarrow O$ is a **randomized mapping**: an individual is mapped to a distribution over outcomes

M is a Lipschitz mapping if $\forall x,y \in X \quad \|M(x),M(y)\| \leq d(x,y)$

close individuals map to close distributions

Fairness through awareness

[C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel; *ITCS 2012*]



simpsons.wikia.com

X individuals ***O*** outcomes ***Y*** actions

fairness enforced at this step **vendor cannot introduce bias**

Vendors can maximize expected utility,
subject to the Lipschitz condition

Racial bias in criminal sentencing

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

Prediction Fails Differently for Black Defendants

	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%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

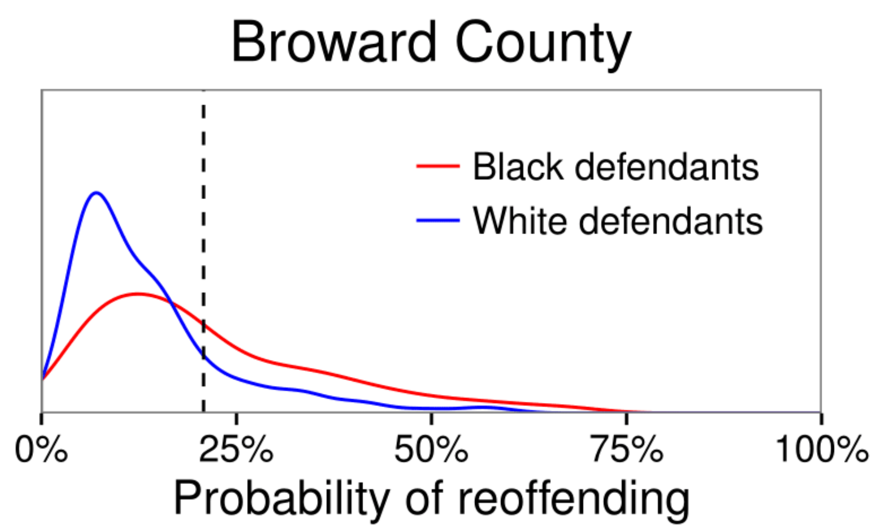
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS as a predictive instrument

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Predictive parity (also called **calibration**)
an instrument identifies a set of instances as having probability *x* of constituting positive instances, the approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is **well-calibrated**: in the window around 40%, the fraction of defendants who were re-arrested is ~40%, both over-all and per group.



[plot from Corbett-Davies et al.; *KDD 2017*]

Group fairness impossibility result

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups

Recidivism rates in the ProPublica dataset are higher for the black group than for the white group

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

What is recidivism?: Northpointe [*the maker of COMPAS*] defined recidivism as "**a finger-printable arrest** involving a charge and a filing for any uniform crime reporting (UCR) code."

Fairness for whom?

based on a slide by Arvind Narayanan

Decision-maker: of those I've labeled high-risk, how many will recidivate?

Defendant: how likely am I to be incorrectly classified high-risk?

Society: (think positive interventions) is the selected set demographically balanced?

	labeled low-risk	labeled high-risk
did not recidivate	TN	**FP**
recidivated	**FN**	TP

different metrics matter to different stakeholders

Diversity

Initial reading

The New York Times

Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016

Like all technologies before it, artificial intelligence will reflect the values of its creators. So **inclusivity matters** — from who designs it to who sits on the company boards and which ethical perspectives are included.

Otherwise, **we risk constructing machine intelligence that mirrors a narrow and privileged vision of society**, with its old, familiar biases and stereotypes.

REVIEW

Diversity in Big Data: A Review

Marina Drosou[1], H.V. Jagadish[2], Evaggelia Pitoura[1], and Julia Stoyanovich[3,*]

Big Data
Volume 5 Number 2, 2017

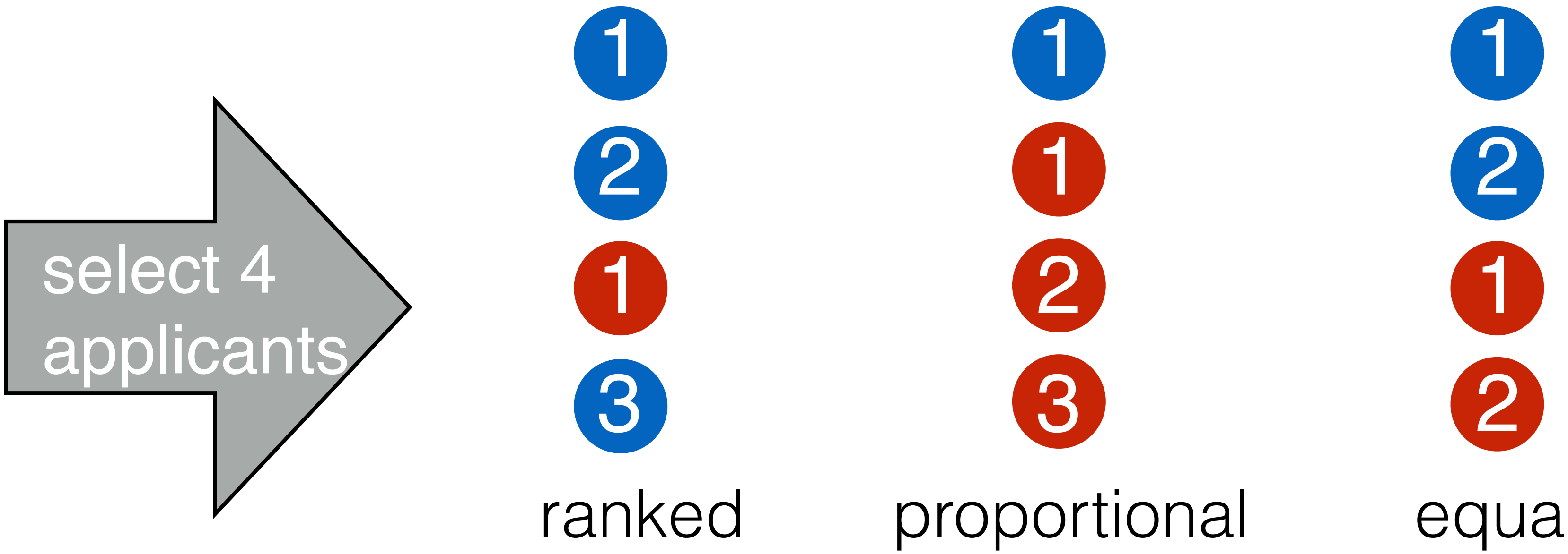
DOI: 10.1089/big.2016.0054

Abstract

Big data technology offers unprecedented opportunities to society as a whole and also to its individual members. At the same time, this technology poses significant risks to those it overlooks. In this article, we give an overview of recent technical work on diversity, particularly in selection tasks, discuss connections between diversity and fairness, and identify promising directions for future work that will position diversity as an important component of a data-responsible society. We argue that diversity should come to the forefront of our discourse, for reasons that are both ethical—to mitigate the risks of exclusion—and utilitarian, to enable more powerful, accurate, and engaging data analysis and use.

Keywords: data; diversity; empirical studies; models and algorithms; responsibly

Job applicant selection

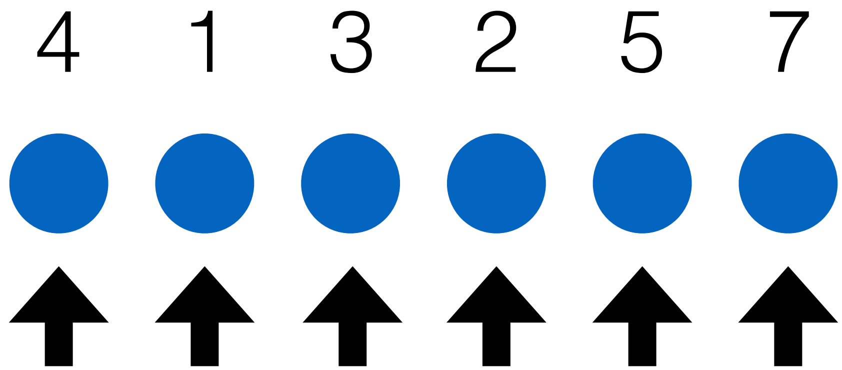


Can state all these as constraints:

for each category ***i***, pick ***K*** elements, with $floor_i \leq K_i \leq ceil_i$

Hiring a job candidate

Goal: Hire a candidate with a high score



Candidates arrive one-by-one

A candidate's score is revealed when the candidate arrives

Decision to accept or reject a candidate made on the spot

The Secretary Problem

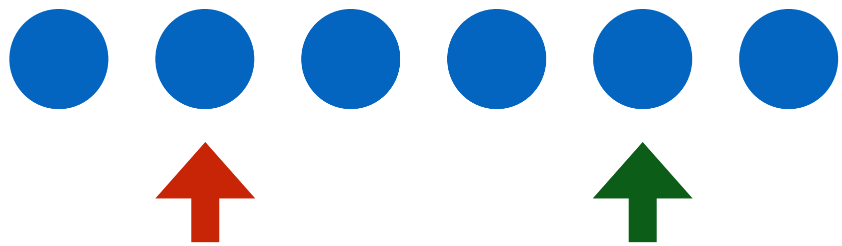
Goal: Design an algorithm for picking **one** element of a **randomly ordered** sequence, to maximize the probability of picking the **maximum element** of the entire sequence.

$N = 6$

$S = \left\lfloor {}^{N}\!/_{e} \right\rfloor = 2$

$T = 4$

4 1 3 2 5 7



Competitive ratio

${}^{1}\!/_{e}$

the best possible!

Consider, and reject, the first *S* candidates

Record *T*, the best seen score among the first *S* candidates

Accept the next candidate with score better than *T*

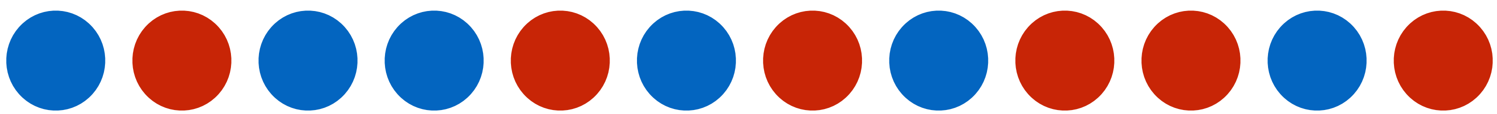
Diverse K-choice Secretary

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

For each **category** ***i***, pick ***K_i*** elements, with $floor_i \leq K_i \leq ceil_i$

6 4 1 8 2 3 1 2 9 5 7 5



$$N_{red} = N_{blue} = 6$$
$$K = 3$$
$$1 \leq K_{red}, K_{blue} \leq 2$$

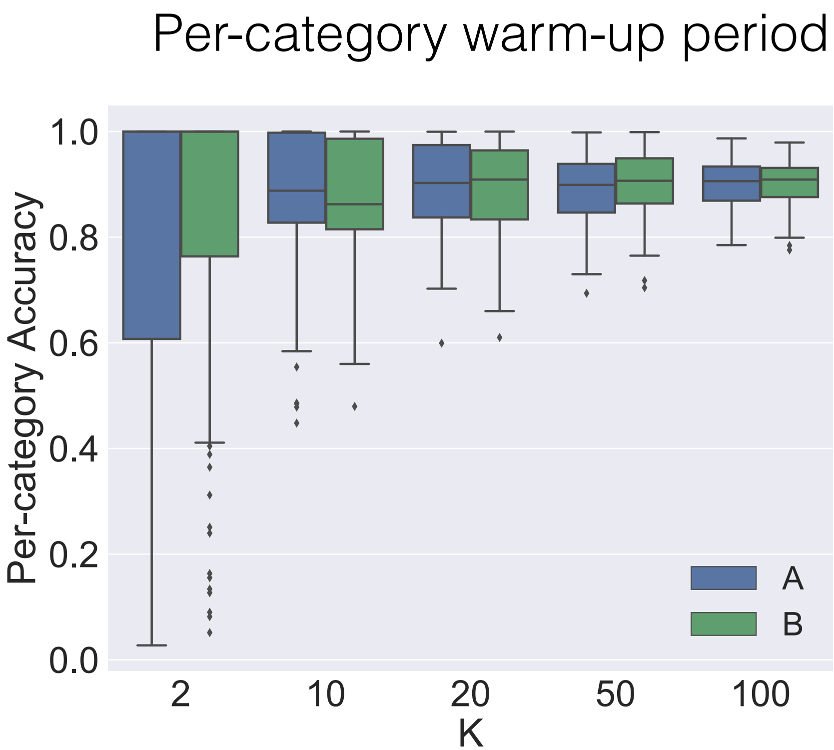
Accept *floor* items for each category from per-category streams

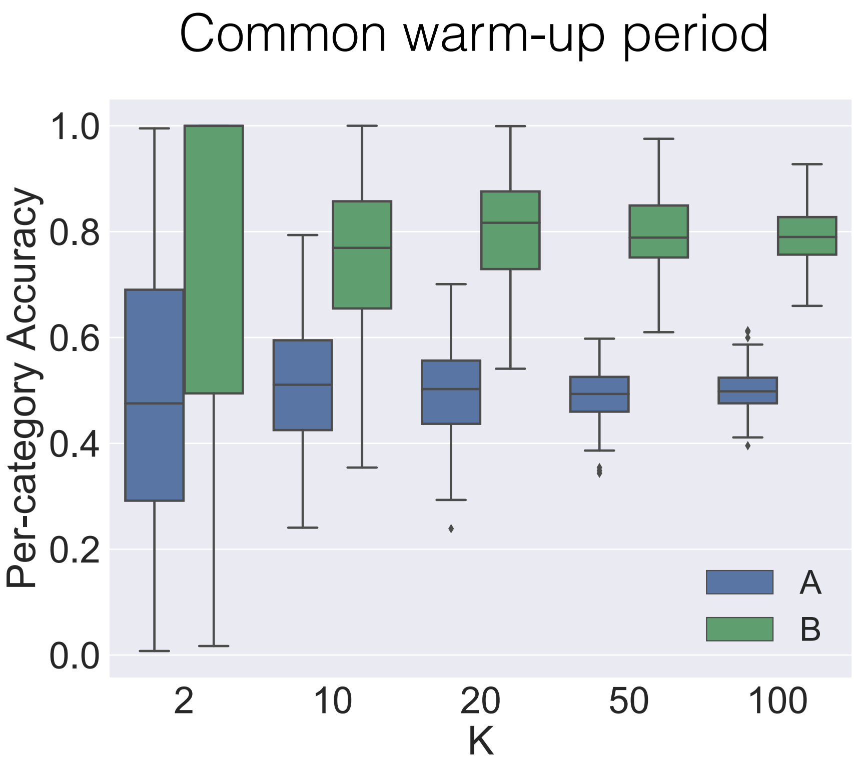
$$slack = K - (floor_{red} + floor_{blue})$$

Accept the remaining *slack* items irrespective of category membership, but subject to *ceil*

Diversity by design is crucial

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]





synthetic data with categories A and B, score depends on category, lower for A

Transparency

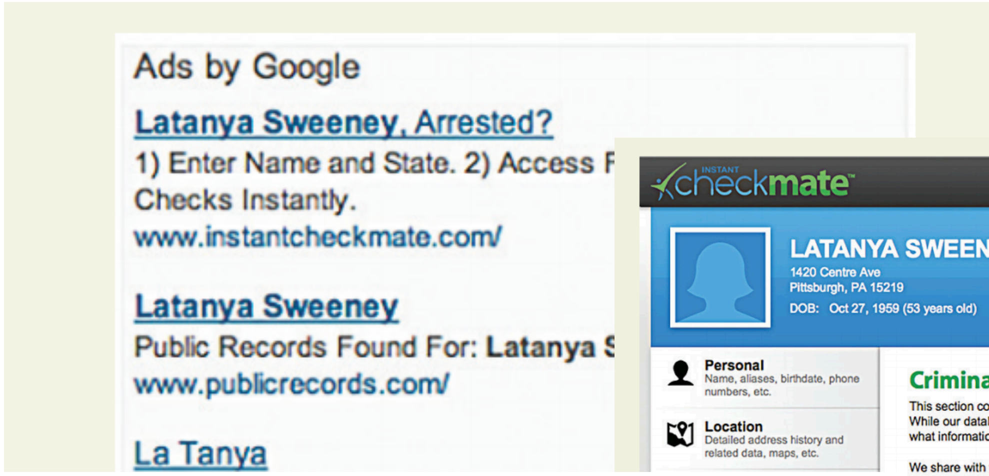
Transparency themes

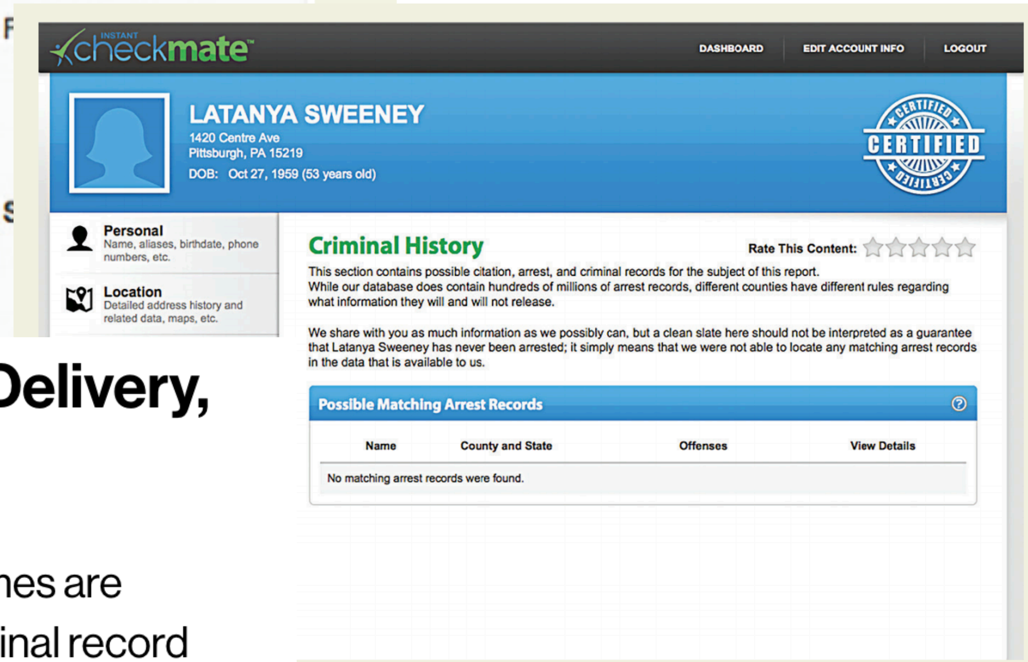
- **Online ad targeting**: identifying the problem
  - Racially identifying names [Sweeney, CACM 2013]
  - Ad Fisher [Datta et al., PETS 2015]
- **Explaining black-box models** (classifiers)
  - LIME: local interpretable explanations [Ribeiro et al., KDD 2016]
  - QII: causal influence of features on outcomes [Datta et al., SSP 2016]
- **Software design and testing** for fairness (won't cover today)
- **Interpretability**
  - Nutritional labels for rankings [Yang et al., SIGMOD 2018]

Racially identifying names

[Latanya Sweeney; *CACM 2013*]

Google
AdSense





Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

racially identifying names trigger ads suggestive of a criminal record

https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

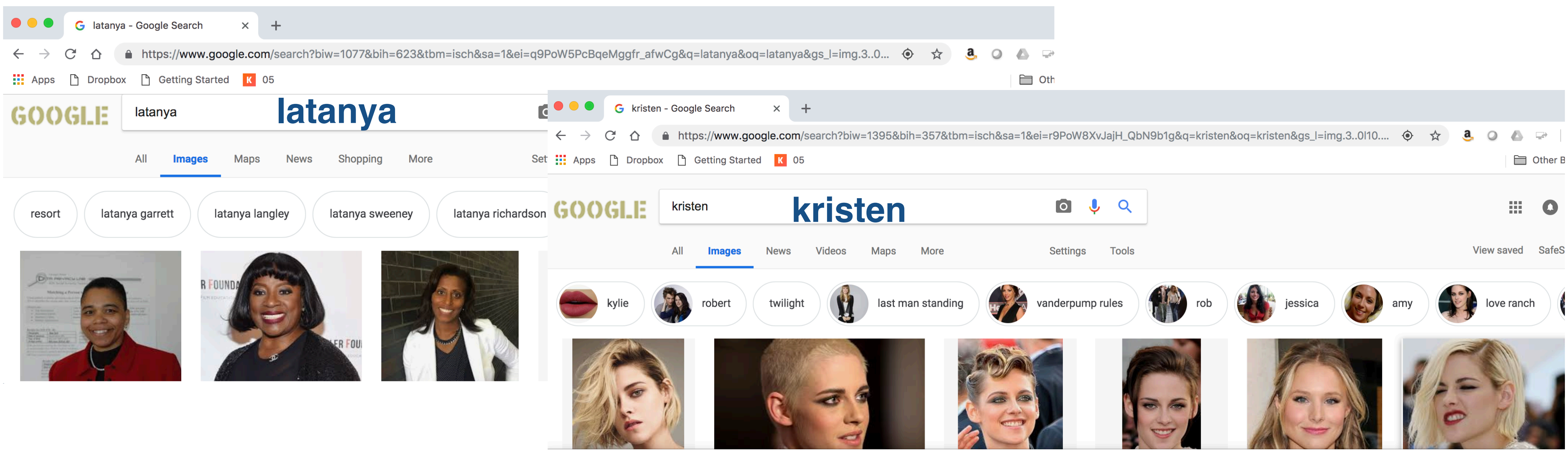
Observations

[Latanya Sweeney; *CACM 2013*]

- Ads suggestive of a criminal record, linking to Instant Checkmate, appear on google.com and reuters.com in response to searches for "Latanya Sweeney", "Latanya Farrell"and "Latanya Lockett"*

- No Instant Checkmate ads when searching for "Kristen Haring", "Kristen Sparrow"* and "Kristen Lindquist"*

- * next to a name associated with an actual arrest record



Why is this happening?

[Latanya Sweeney; *CACM 2013*]

Possible explanations (from Latanya Sweeney):

- Does Instant Checkmate serve ads specifically for black-identifying names?
- Is Google's Adsense explicitly biased in this way?
- Does Google's Adsense learn racial bias based on from click-through rates?

How do we know which explanation is right?

We need transparency!

Response

https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

In response to this blog post, a **Google** spokesperson send :

"AdWords does not conduct any racial profiling. We also
violence policy which states that we will not allow ads tha
organisation, person or group of people. It is up to individ
which keywords they want to choose to trigger their ads."

Instantcheckmate.com sends the following statement:

"As a point of fact, Instant Checkmate would like to state u
never engaged in racial profiling in Google AdWords. We h
technology in place to even connect a name with a race an
attempt to do so. The very idea is contrary to our company
principles and values."

Who is responsible?

- Who benefits?
- Who is harmed?
- What does the law say?
- Who is in a position to mitigate?

transparency …. **responsibility** …. **trust**

Pivot: the origins of data protection

Detour: Barrow, Alaska, 1979

Native leaders and city officials, worried about drinking and associated violence in their community **invited a group of sociology researchers** to assess the problem and work with them to devise solutions.

Methodology:

- 10% representative sample (N=88) of everyone over the age of 15 using a 1972 demographic survey
- Interviewed on attitudes and values about use of alcohol
- Obtained psychological histories & drinking behavior
- Given the Michigan Alcoholism Screening Test
- Asked to draw a picture of a person (used to determine cultural identity)

based on a slide by Bill Howe

Study “results”

Alcohol Plagues Eskimos; Alcoholism Plagues Eskimo Village

DAVA SOBEL ();
January 22, 1980,
, Section Science Times, Page C1, Column , words

PERMISSIONS

[DISPLAYING ABSTRACT]

THE Inupiat Eskimos of Alaska's North Slope, whose culture has been overwhelmed by energy development activities, are "practically committing suicide" by mass alcoholism, University of Pennsylvania researchers said here yesterday. The alcoholism rate is 72 percent among the 2,000 Eskimo men and women in the village of Barrow, where violence is becoming the ...

At the conclusion of the study researchers formulated a report entitled **“The Inupiat, Economics and Alcohol on the Alaskan North Slope”**, released **simultaneously** at a press release and to the Barrow community.

The press release was picked up by the New York Times, who ran a front page story entitled **“Alcohol Plagues Eskimos”**

based on a slide by Bill Howe

Harms and backlash

Study **results were revealed** in the context of a press conference that was held far from the Native village, and **without the presence, much less the knowledge or consent**, of any community member who might have been able to present any context concerning the socioeconomic conditions of the village.

Study results suggested that nearly all adults in the community were alcoholics. In addition to the shame felt by community members, the town's Standard and Poor bond rating suffered as a result, which in turn decreased the tribe's ability to secure funding for much needed projects.

Article Preview

Eskimos Irate Over Alcoholism Study

[DISPLAYING ABSTRACT]

PERMISSIONS

BARROW, ALASKA HOT tempers and tension arising from a scientific report that found a high rate of alcoholism in this predominantly Eskimo community have abated somewhat after two days of meetings here at the northernmost point of Alaska.

based on a slide by Bill Howe

Problems

Methodological

Edward F. Foulks, M.D., "Misalliances In The Barrow Alcohol Study"

- "The authors once again met with the Barrow Technical Advisory Group, who stated their concern that only Natives were studied, and that outsiders in town had not been included."

any chance of selection bias?

- "The **estimates of the frequency of intoxication** based on **association with the probability of being detained** were termed "ludicrous, both logically and statistically."

Ethical

- Participants not in control of their data
- Significant harm: social (stigmatization) and financial (bond rating)
- No laws were broken, and harms are not about individual privacy!
- **Who benefits? Who is harmed?**

data protection …. responsibility …. trust

based on a slide by Bill Howe

GENERAL DATA PROTECTION REGULATION (GDPR) | RECITALS | KEY ISSUES | Deutsch

General Data Protection Regulation
GDPR

Welcome to gdpr-info.eu. Here you can find the official PDF of the Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018 as a neatly arranged website. All Articles of the GDPR are linked with suitable recitals. The European Data Protection Regulation is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe. If you find the page useful, feel free to support us by sharing the project.

Quick Access

Pivot back: Transparency

Online job ads

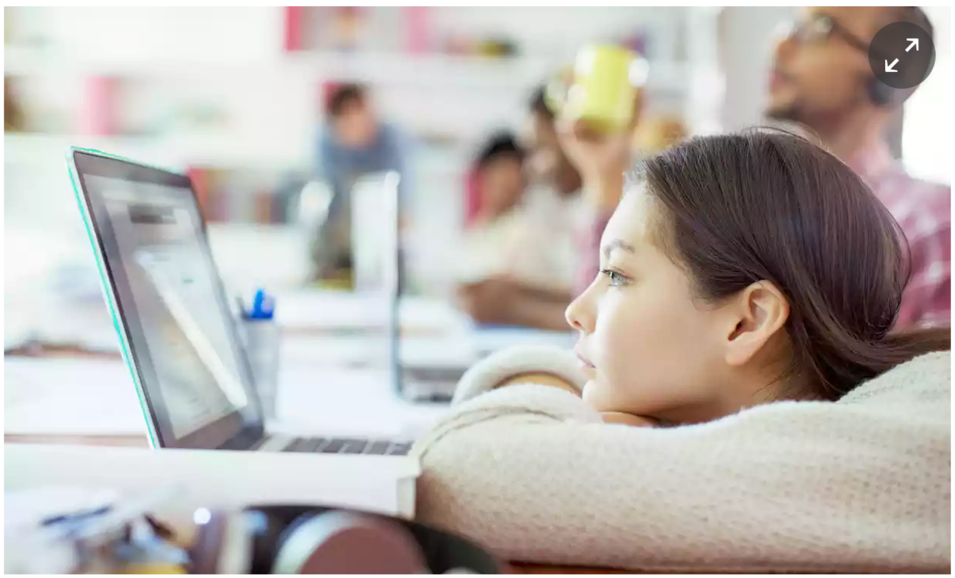
theguardian

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for "$200k+" executive jobs **1,852 times to the male group and only 318 times to the female group**. Another experiment, in July 2014, showed a similar trend but was not statistically significant.

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

Ad targeting online

- **Users** browse the Web, consume content, consume ads (see / click / purchase)
- **Content providers** (or **publishers**) host online content that often includes ads. They outsource ad placement to third-party ad networks
- **Advertisers** seek to place their ads on publishers' websites
- **Ad networks** track users across sites, to get a global view of users' behaviors. They connect advertisers and publishers

AdFisher: gender and jobs

[A. Datta, M. Tschantz, A. Datta; *PETS 2015*]

Non-discrimination: Users differing only in protected attributes are treated similarly

Causal test: Find that a protected attribute changes ads

Experiment: **gender and jobs**

Specify gender (male/female) in Ad Settings, simulate interest in jobs by visiting employment sites, collect ads from Times of India or the Guardian

Result: males were shown ads for higher-paying jobs significantly more often than females (1852 vs. 318)

violation

Who is responsible?

[A. Datta, A. Datta, J. Makagon, D. Mulligan, M. Tschantz; *FAT* 2018*]

- **Google alone:** explicitly programming the system to show the ad less often to females, e.g., based on independent evaluation of demographic appeal of product (**explicit and intentional discrimination**)
- **The advertiser:** targeting of the ad through explicit use of demographic categories (**explicit and intentional**), selection of proxies (**hidden and intentional**), or through those choices without intent (**unconscious selection bias**) and **Google** respecting these targeting criteria
- **Other advertisers:** others outbid our advertiser when targeting to females
- **Other users:** Male and female users behaving differently to ads, and Google learning to predict this behavior

How is targeting done?

[A. Datta, A. Datta, J. Makagon, D. Mulligan, M. Tschantz; *FAT* 2018*]

Can an advertiser use AdWords to target on gender, or on a known proxy of gender? **Yep!**

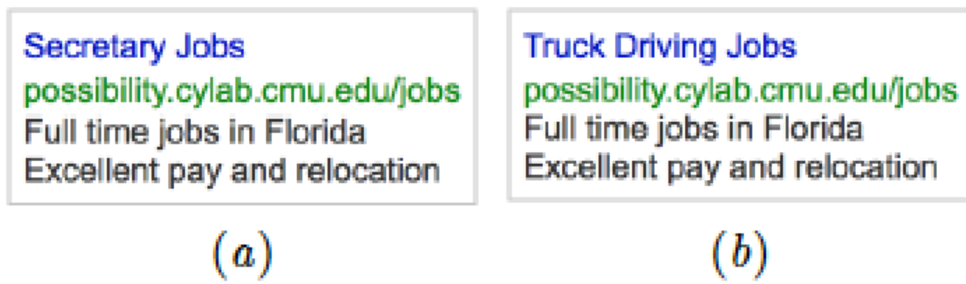


Figure 1: Ads approved by Google in 2015. The ad in the left (right) column was targeted to women (men).

"This finding demonstrates that an advertiser with discriminatory intentions can use the AdWords platform to serve employment related ads disparately on gender."

What are the legal ramifications?

[A. Datta, A. Datta, J. Makagon, D. Mulligan, M. Tschantz; *FAT* 2018*]

- Each actor in the advertising ecosystem may have contributed inputs that produced the effect
- **It is impossible to know, without additional information, what actors - other than the consumers of the ads - did or did not do**
- In particular, impossible to asses intent, which *may* be necessary to asses the extent of legal liability. Or it may not!
    - **Title VII of the 1964 Civil Rights Act** that makes it unlawful to discriminate based on sex in several stages of employment. It includes an advertising prohibition (think sex-specific help wanted columns in a newspaper), which does not turn on intent
    - Title VII is limited in scope to employers, labor organizations, employment agencies, joint labor-management committees - does not directly apply here!
    - **Fair Housing Act (FHA)** is perhaps a better guide than Title VII, limiting both content and activities that target advertisement based on protected attributes

Explaining black-box classifiers

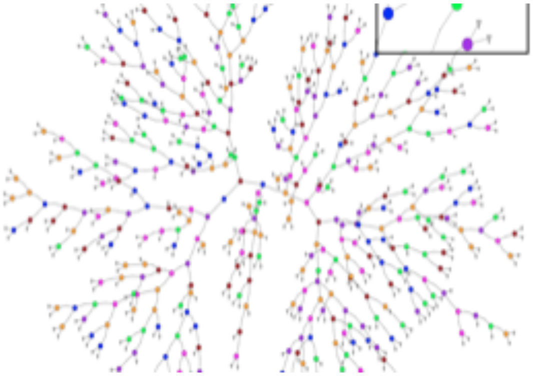
LIME: Local explanations of classifiers

[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

https://www.youtube.com/watch?v=hUnRCxnydCc

Three must-haves for a good explanation

Interpretable
- Humans can easily interpret reasoning



Definitely
not interpretable

Potentially
interpretable

slide by Marco Tulio Ribeiro, KDD 2016

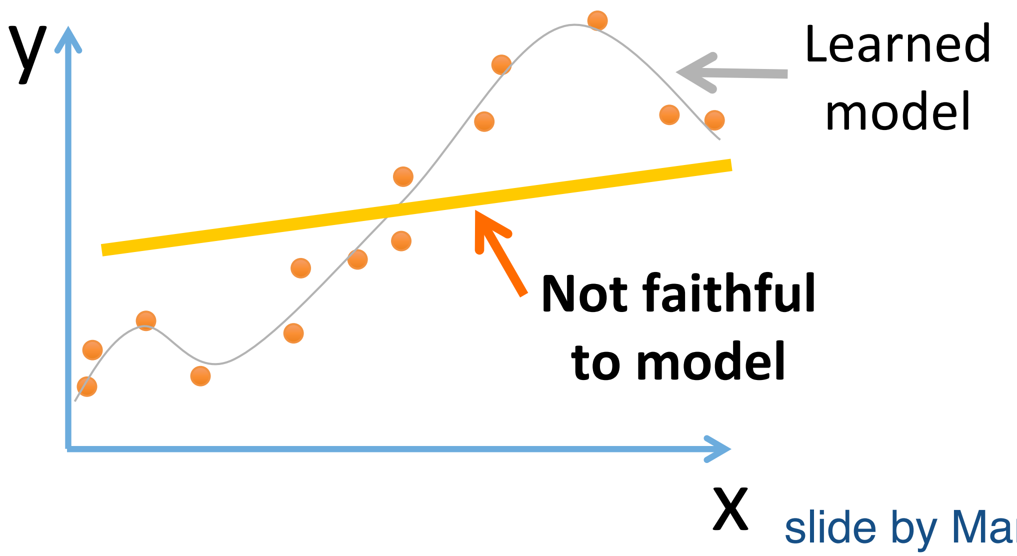
LIME: Local explanations of classifiers

[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

https://www.youtube.com/watch?v=hUnRCxnydCc

Three must-haves for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves



slide by Marco Tulio Ribeiro, KDD 2016

LIME: Local explanations of classifiers

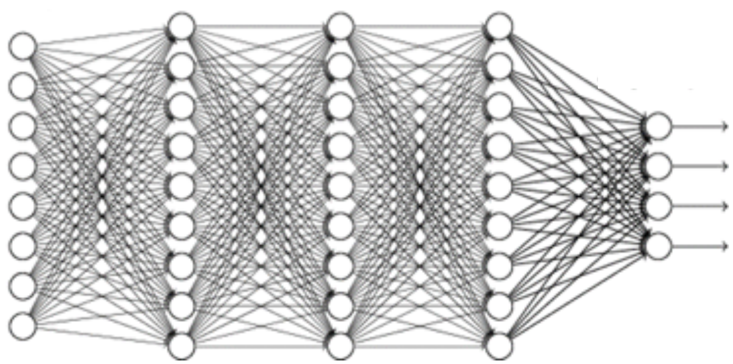
[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

https://www.youtube.com/watch?v=hUnRCxnydCc

Three must-haves for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves
Model agnostic	• Can be used for *any* ML model

Can explain this mess ☺

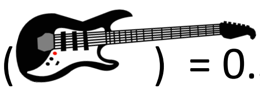


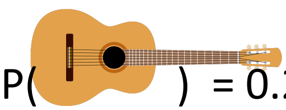
slide by Marco Tulio Ribeiro, KDD 2016

Local explanations of classifiers

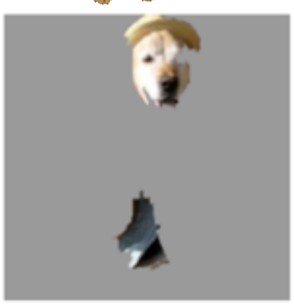
[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

Explaining Google's Inception NN

P() = 0.32

P() = 0.24

P() = 0.21

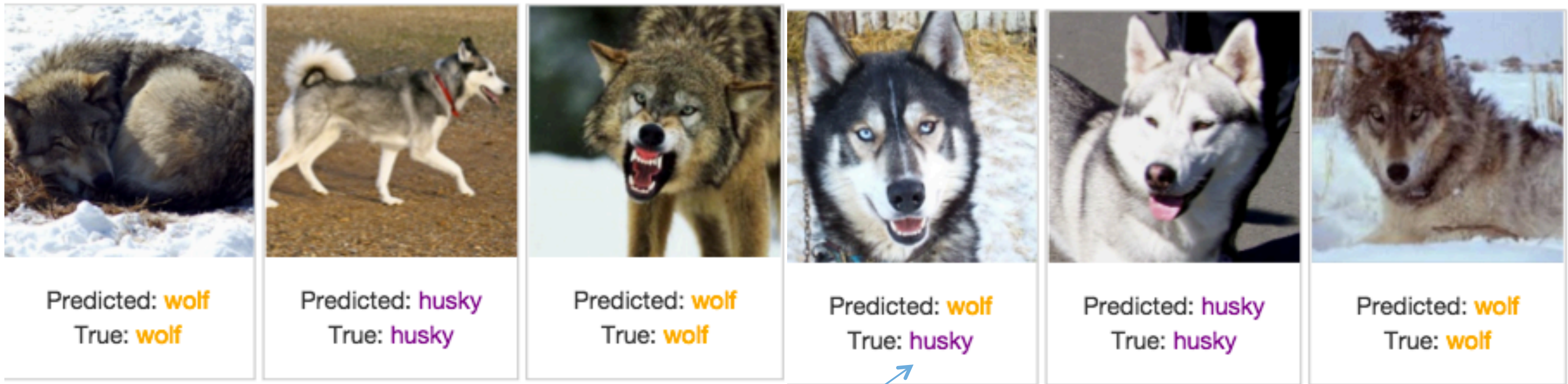


slide by Marco Tulio Ribeiro, KDD 2016

Local explanations of classifiers

[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

Train a neural network to predict wolf v. husky



Only 1 mistake!!!

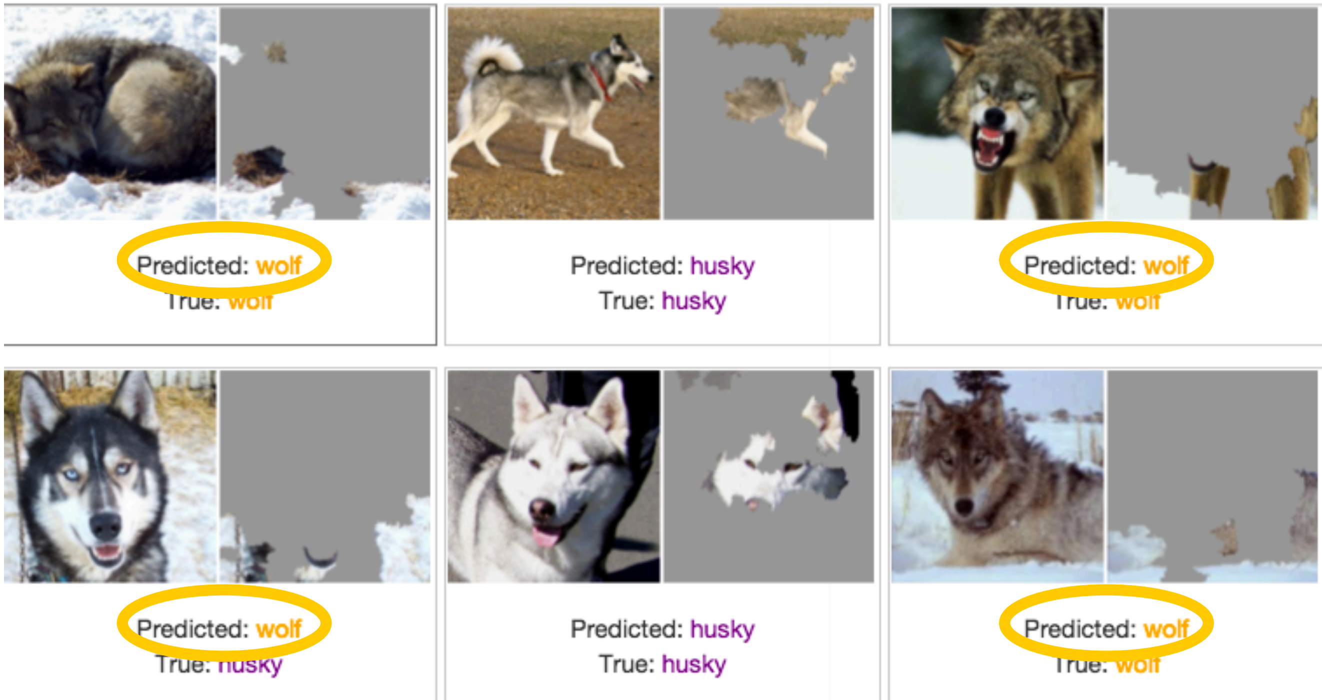
Do you trust this model?
How does it distinguish between huskies and wolves?

slide by Marco Tulio Ribeiro, KDD 2016

Local explanations of classifiers

[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

Explanations for neural network prediction

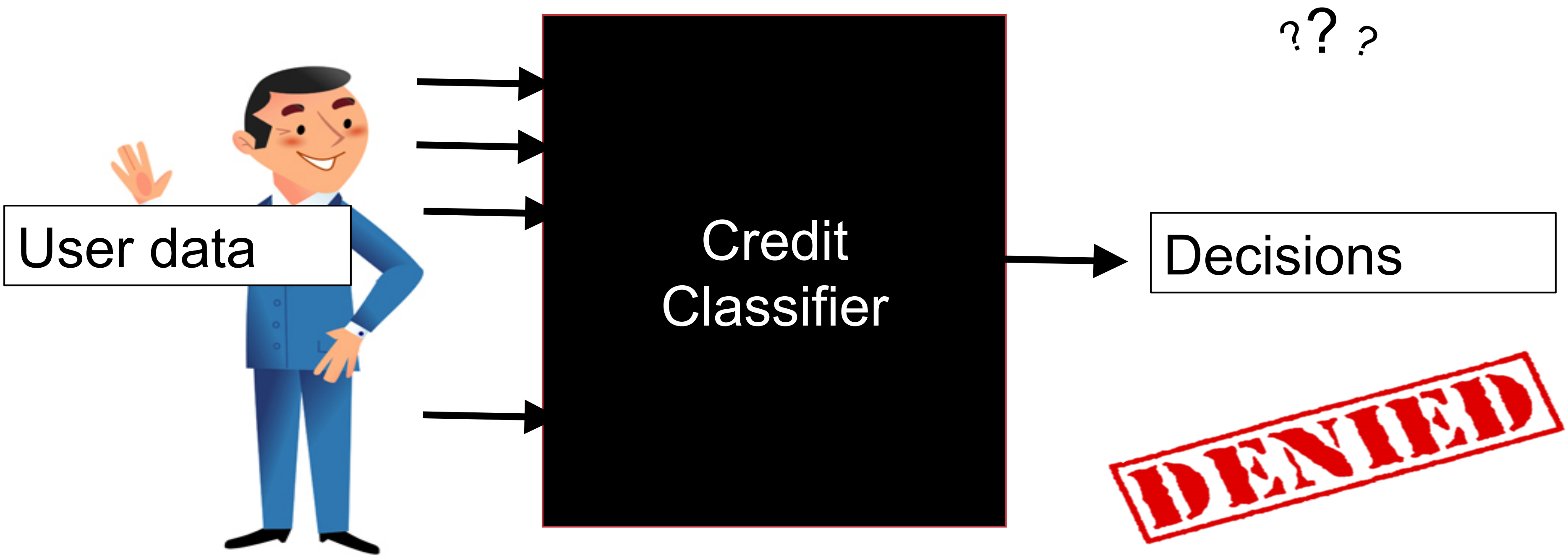


We've built a great snow detector… ☹

slide by Marco Tulio Ribeiro, KDD 2016

Auditing black-box models

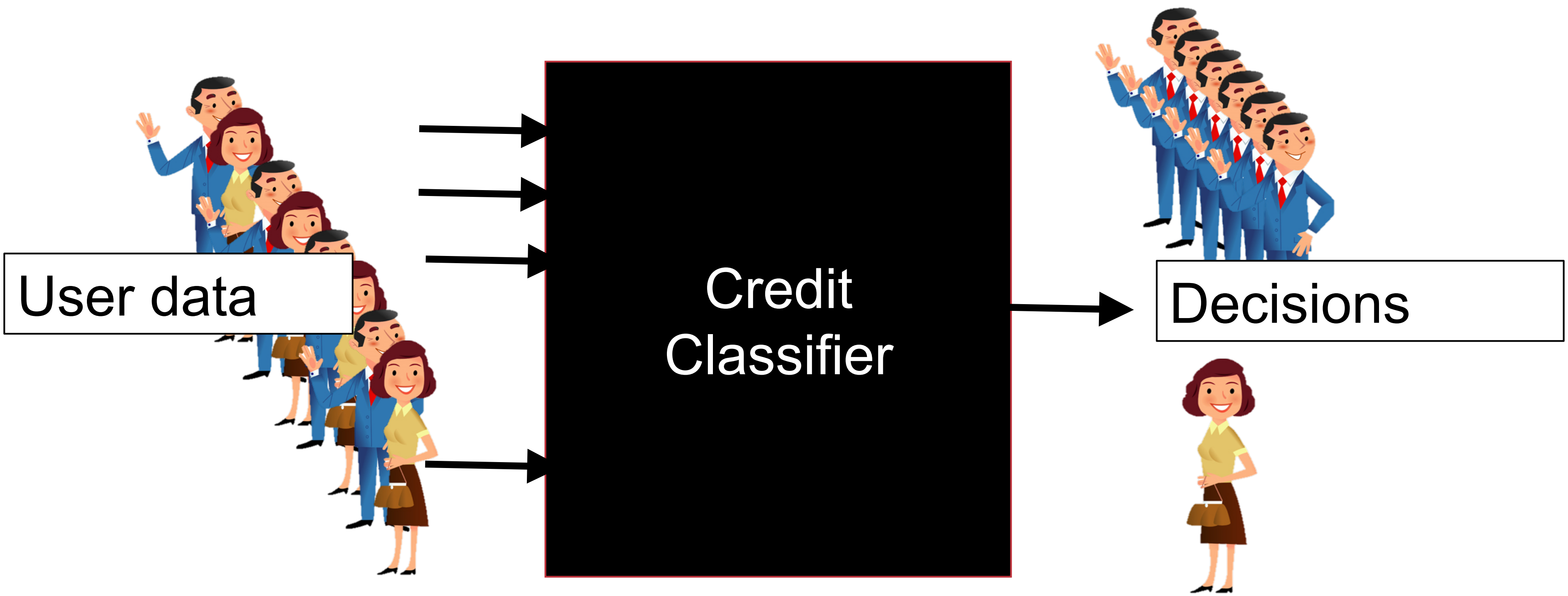
[A. Datta, S. Sen, Y. Zick; *SP 2016*]



slide by A. Datta

Auditing black-box models

[A. Datta, S. Sen, Y. Zick; *SP 2016*]



slide by A. Datta

Influence of inputs on outcomes

[A. Datta, S. Sen, Y. Zick; *SP 2016*]

QII: quantitative input influence framework

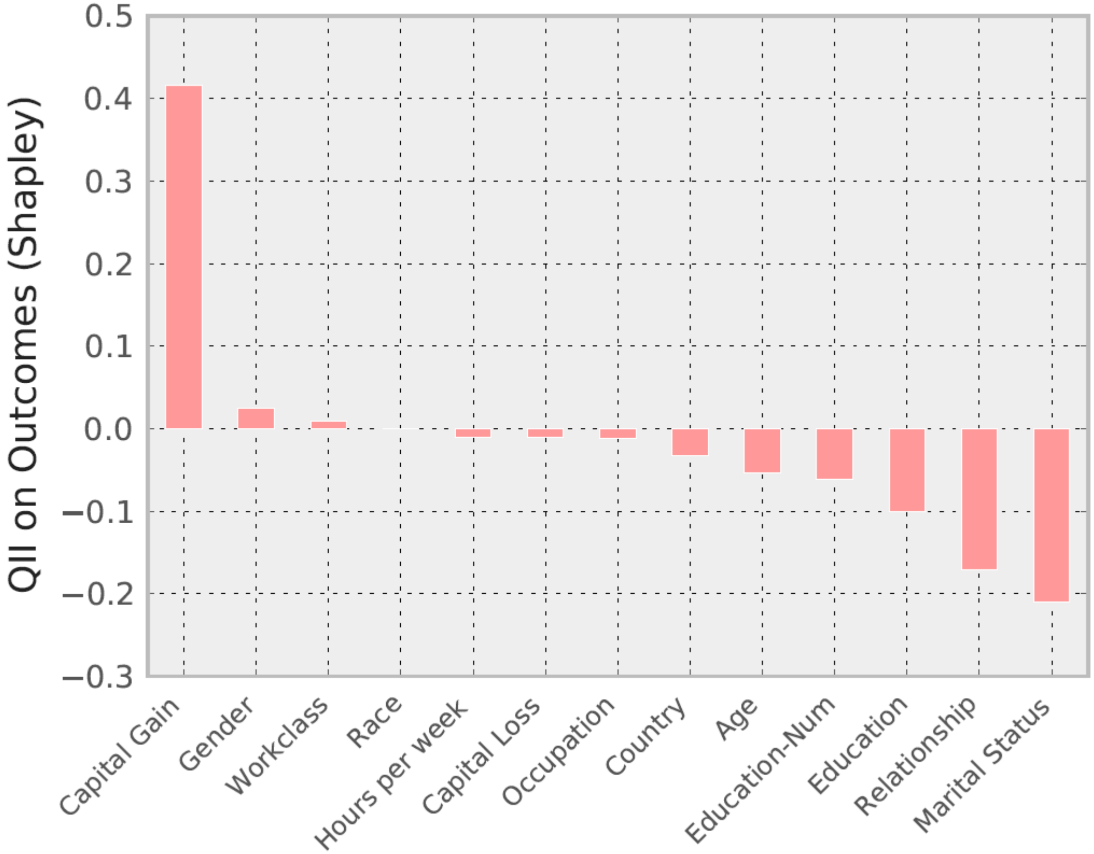
Goal: determine how much influence an input, or a set of inputs, has on a **classification outcome** for an individual or a group

Uses **causal inference**: For a quantity of influence **Q** and an input feature **i**, the QII of **i** on **Q** is the difference in **Q** when **i** is changed via an **intervention**

Replace features with random values from the population, examine the distribution over outcomes

Transparency report: Mr X

[A. Datta, S. Sen, Y. Zick; *SP 2016*]



DENIED

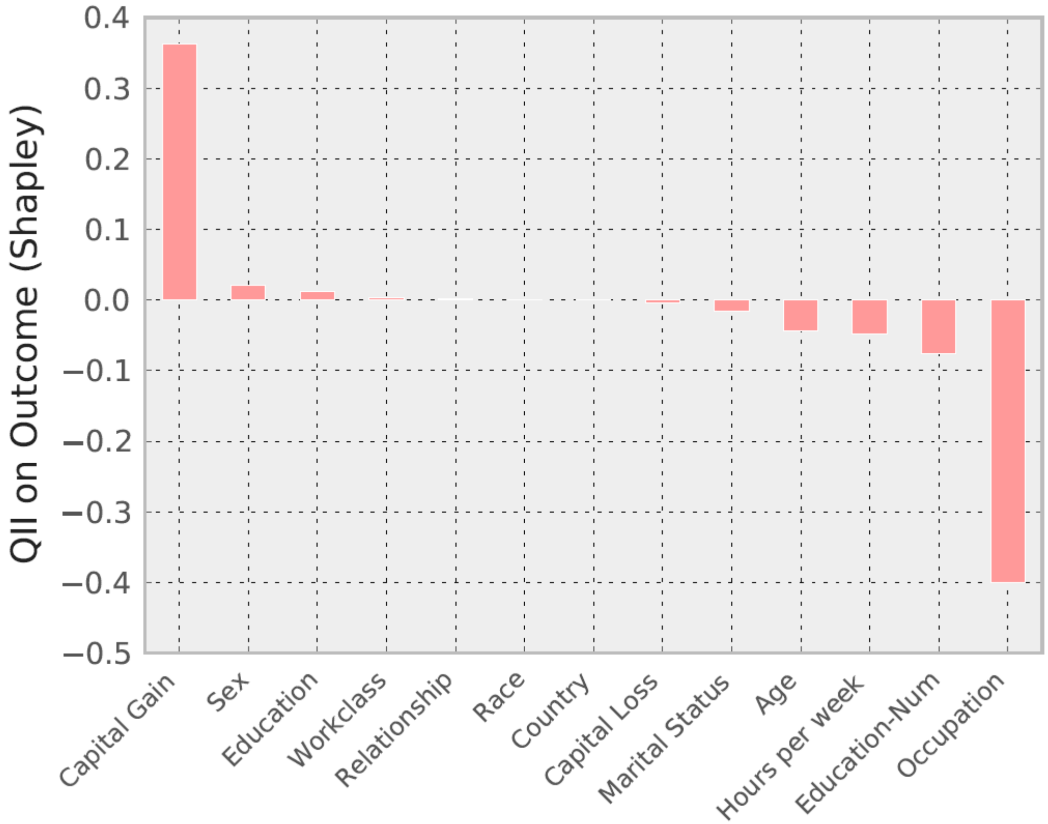
Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344
Capital loss	$0
Work hours per week	40
Country	Vietnam

income

slide by A. Datta

Transparency report: Mr Y

[A. Datta, S. Sen, Y. Zick; *SP 2016*]



Age	27
Workclass	Private
Education	Preschool
Marital Status	Married
Occupation	Farming-Fishing
Relationship to household income	Other Relative
Race	White
Gender	Male
Capital gain	$41310
Capital loss	$0
Work hours per week	24
Country	Mexico

income

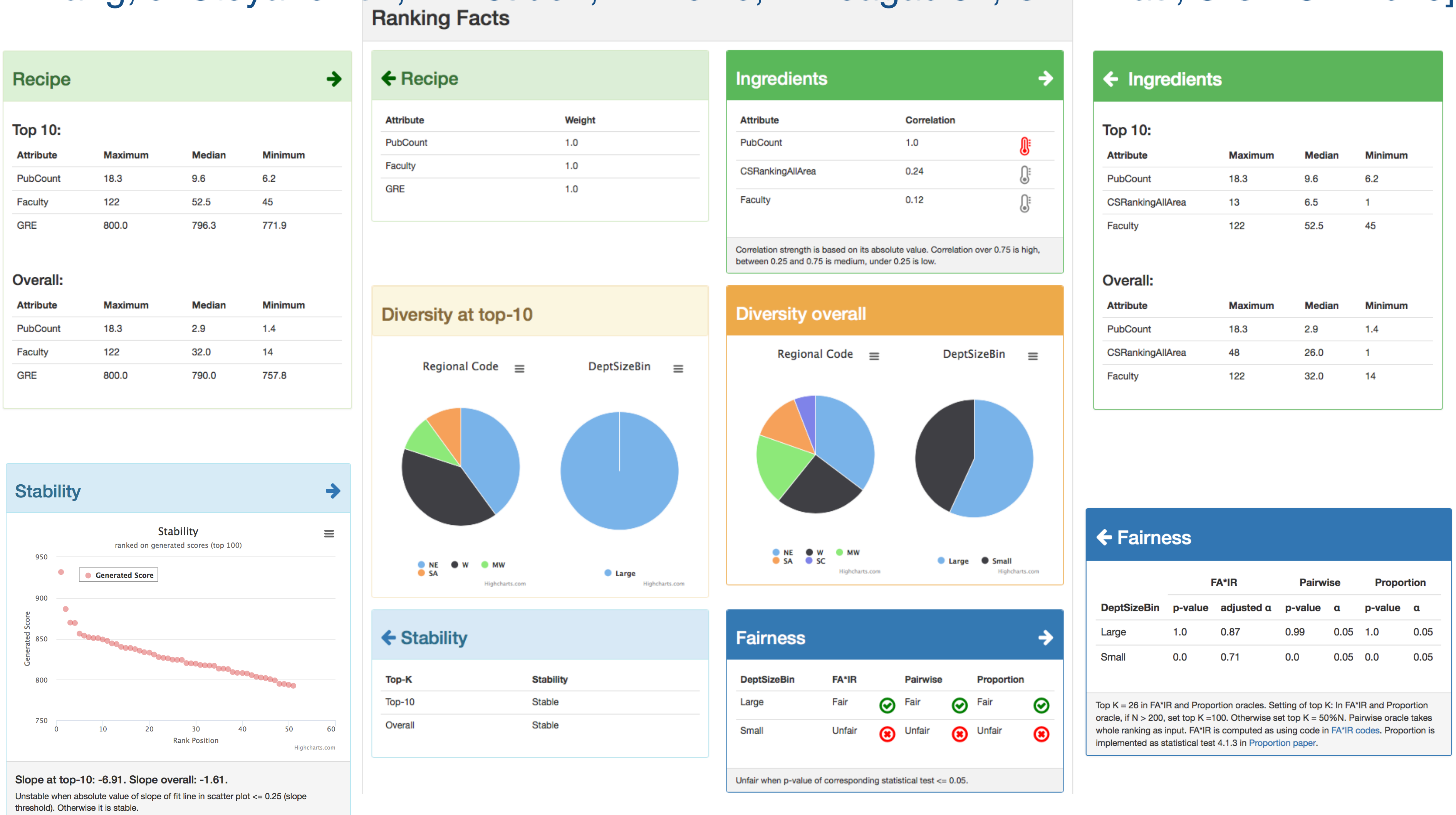
explanations for superficially similar individuals can be different

slide by A. Datta

Interpretability

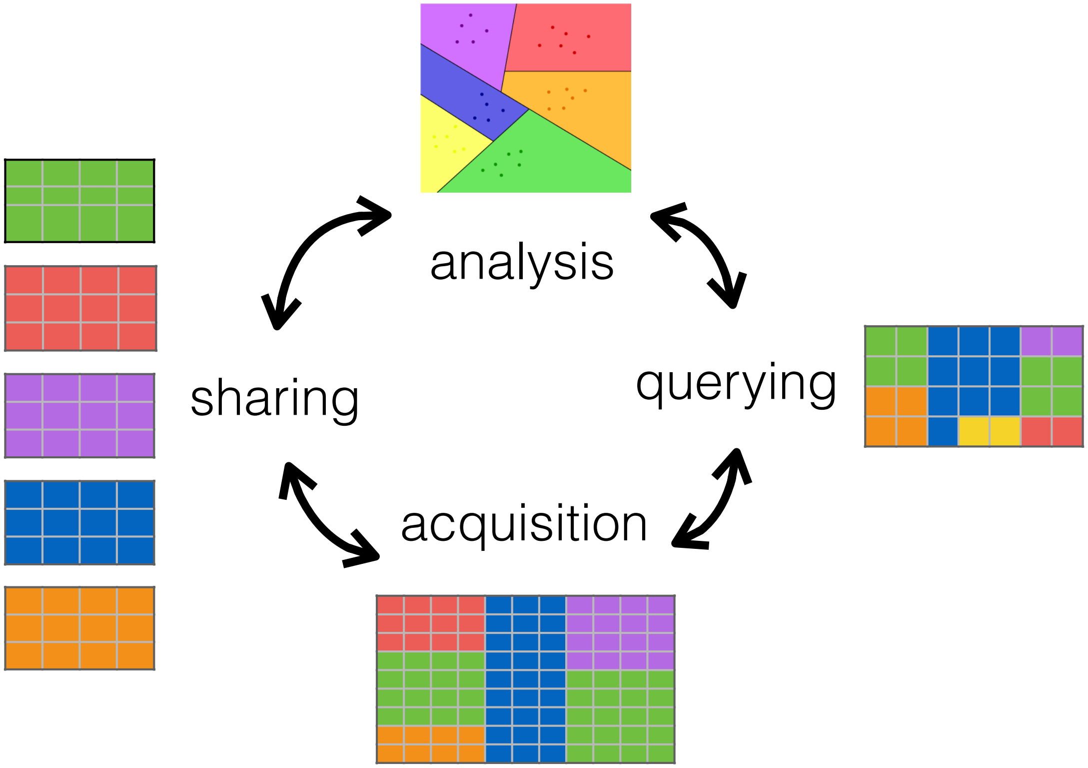
“Nutritional labels” for data and models

[K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, HV Jagadish, G. Miklau; *SIGMOD 2018*]



http://demo.dataresponsibly.com/rankingfacts/nutrition_facts/

Zooming out



NYC ADS transparency law

1/11/2018

Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies

The New York City Council
Corey Johnson, Speaker

Sign In

Legislative Research Center

Council Home | Legislation | Calendar | City Council | Committees

RSS | Alerts

Details | Reports

File #:	Int 1696-2017 Version: A	Name:	Automated decision systems used by agencies.
Type:	Introduction	Status:	Enacted
		Committee:	Committee on Technology
On agenda:	8/24/2017		
Enactment date:	1/11/2018	Law number:	2018/049
Title:	A Local Law in relation to automated decision systems used by agencies		
Sponsors:	James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca		
Council Member Sponsors:	9		
Summary:	This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems.		
Indexes:	Oversight		
Attachments:	1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017		

Get engaged!

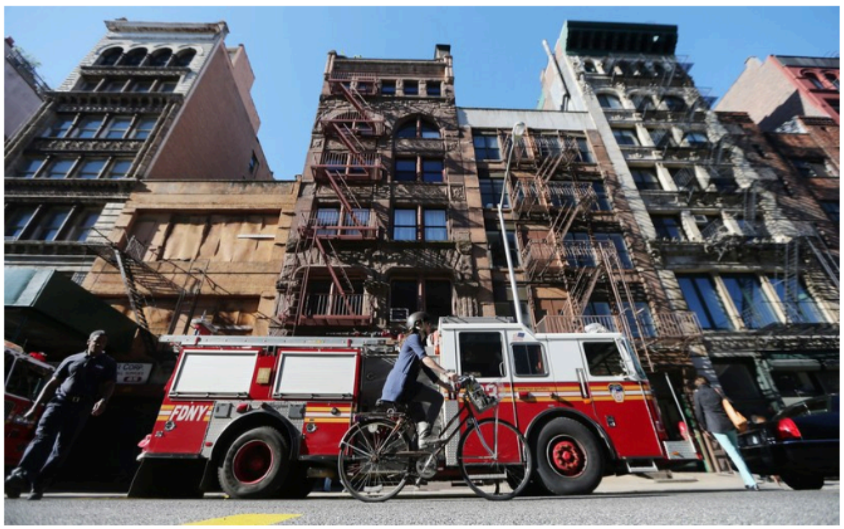
10/16/2017

THE NEW YORKER

By Julia Powles December 20, 2017

ELEMENTS

NEW YORK CITY'S BOLD, FLAWED ATTEMPT TO MAKE ALGORITHMS ACCOUNTABLE



Automated systems guide the allocation of everything from firehouses to food stamps. So why don't we know more about them?

Photograph by Mario Tama / Getty

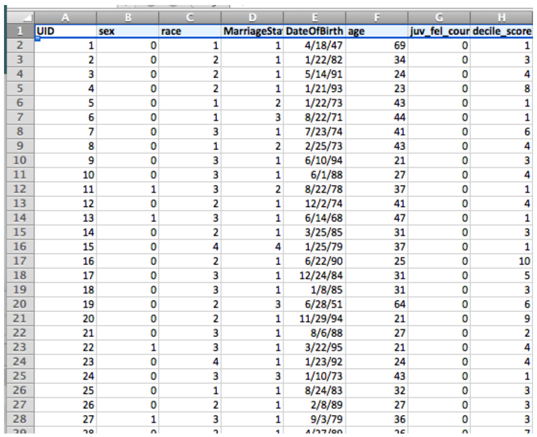
Summary of Int. No. 1696-A

Form an automated decision systems (**ADS**) task force that surveys current use of algorithms and data in City agencies and develops procedures for:

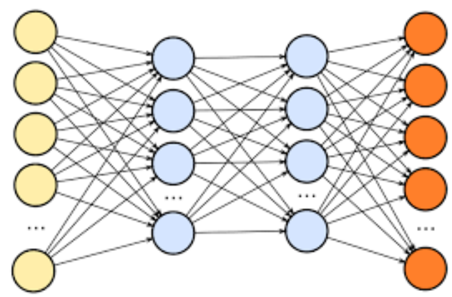
- requesting and receiving an **explanation** of an algorithmic decision affecting an individual (3(b))
- interrogating ADS for **bias and discrimination** against members of legally-protected groups (3(c) and 3(d))
- allowing the **public** to **assess** how ADS function and are used (3(e)), and archiving ADS together with the data they use (3(f))

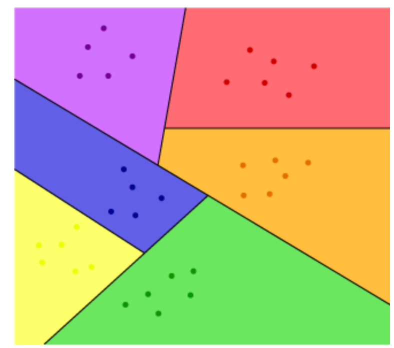
Responsible data science

- Be **transparent** and **accountable**
- Achieve **equitable** resource distribution
- Be cognizant of the **rights** and **preferences** of individuals



	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3

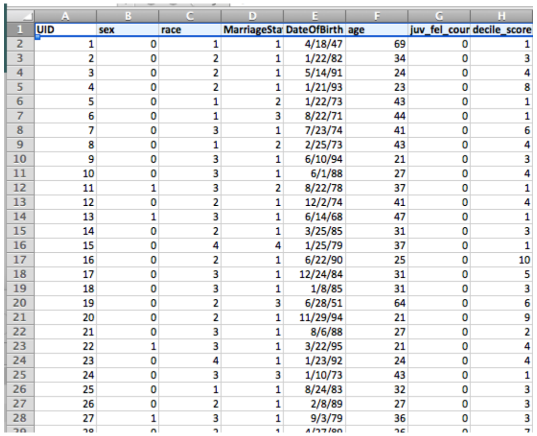




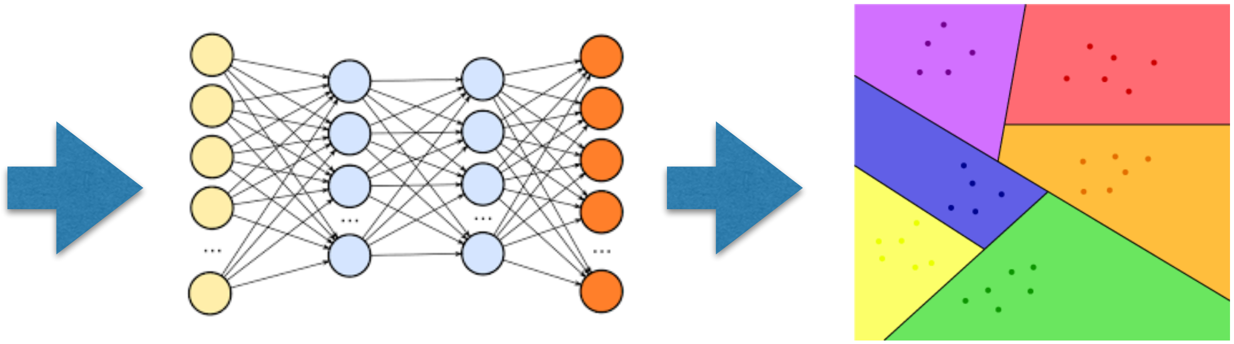
done?

but where does the data come from?

Mitigating urban homelessness



	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3



finding: women are underrepresented in the favorable outcome groups (group fairness)

fix the model!

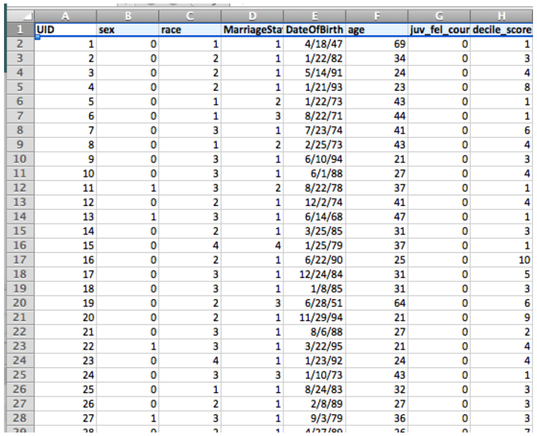
of course, but maybe… the input was generated with:

select * from R
where status = 'unsheltered'
and length > 2 1 month

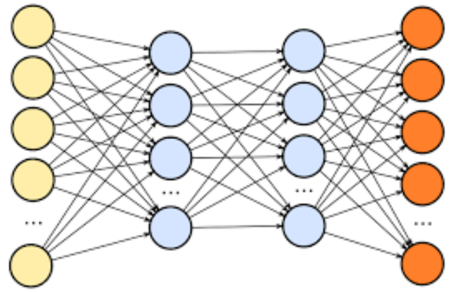
10% female

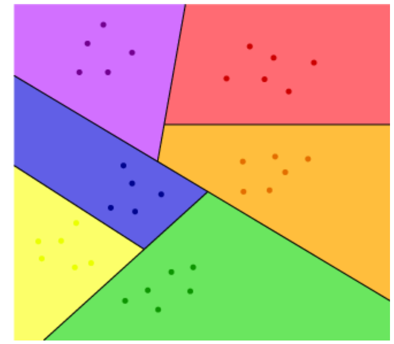
40% female

Mitigating urban homelessness



	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3





finding: young people are recommended pathways of lower effectiveness (high error rate)

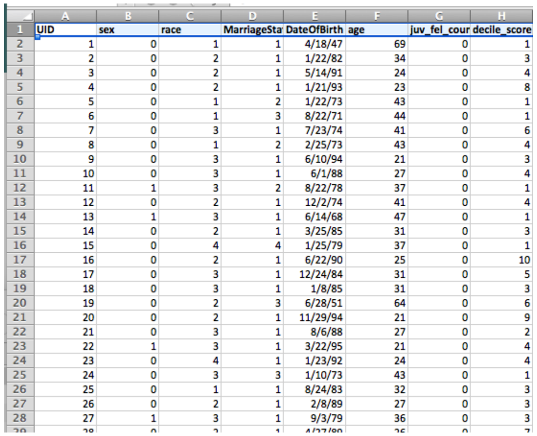
fix the model!

of course, but maybe…

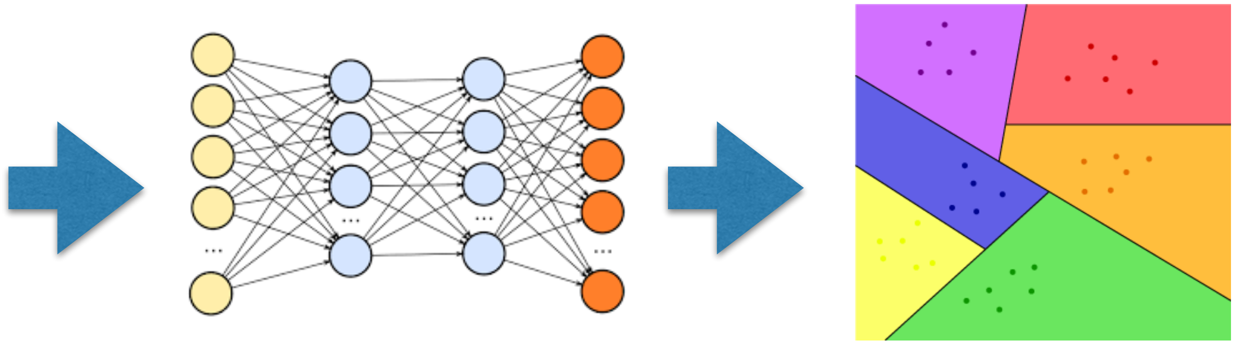
mental health info was missing for this population

go back to the data acquisition step, look for additional datasets

Mitigating urban homelessness



UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
1	0	1	1	4/18/47	69	0	1
2	0	2	1	1/22/82	34	0	3
3	0	2	1	5/14/91	24	0	4
4	0	2	1	1/21/93	23	0	8
5	0	1	2	1/22/73	43	0	1
6	0	1	3	8/22/71	44	0	1
7	0	3	1	7/23/74	41	0	6
8	0	1	2	2/25/73	43	0	4
9	0	3	1	6/10/94	21	0	3
10	0	3	1	6/1/88	27	0	4
11	1	3	2	8/22/78	37	0	1
12	0	2	1	12/2/74	41	0	4
13	1	3	1	6/14/68	47	0	1
14	0	2	1	3/25/85	31	0	3
15	0	4	4	1/25/79	37	0	1
16	0	2	1	6/22/90	25	0	10
17	0	3	1	12/24/84	31	0	5
18	0	3	1	1/8/85	31	0	3
19	0	2	3	6/28/51	64	0	6
20	0	2	1	11/29/94	21	0	9
21	0	3	1	8/6/88	27	0	2
22	1	3	1	3/22/95	21	0	4
23	0	4	1	1/23/92	24	0	4
24	0	3	3	1/10/73	43	0	1
25	0	1	1	8/24/83	32	0	3
26	0	2	1	2/8/89	27	0	3
27	1	3	1	9/3/79	36	0	3



finding: minors are underrepresented in the input, compared to their actual proportion in the population (insufficient data)

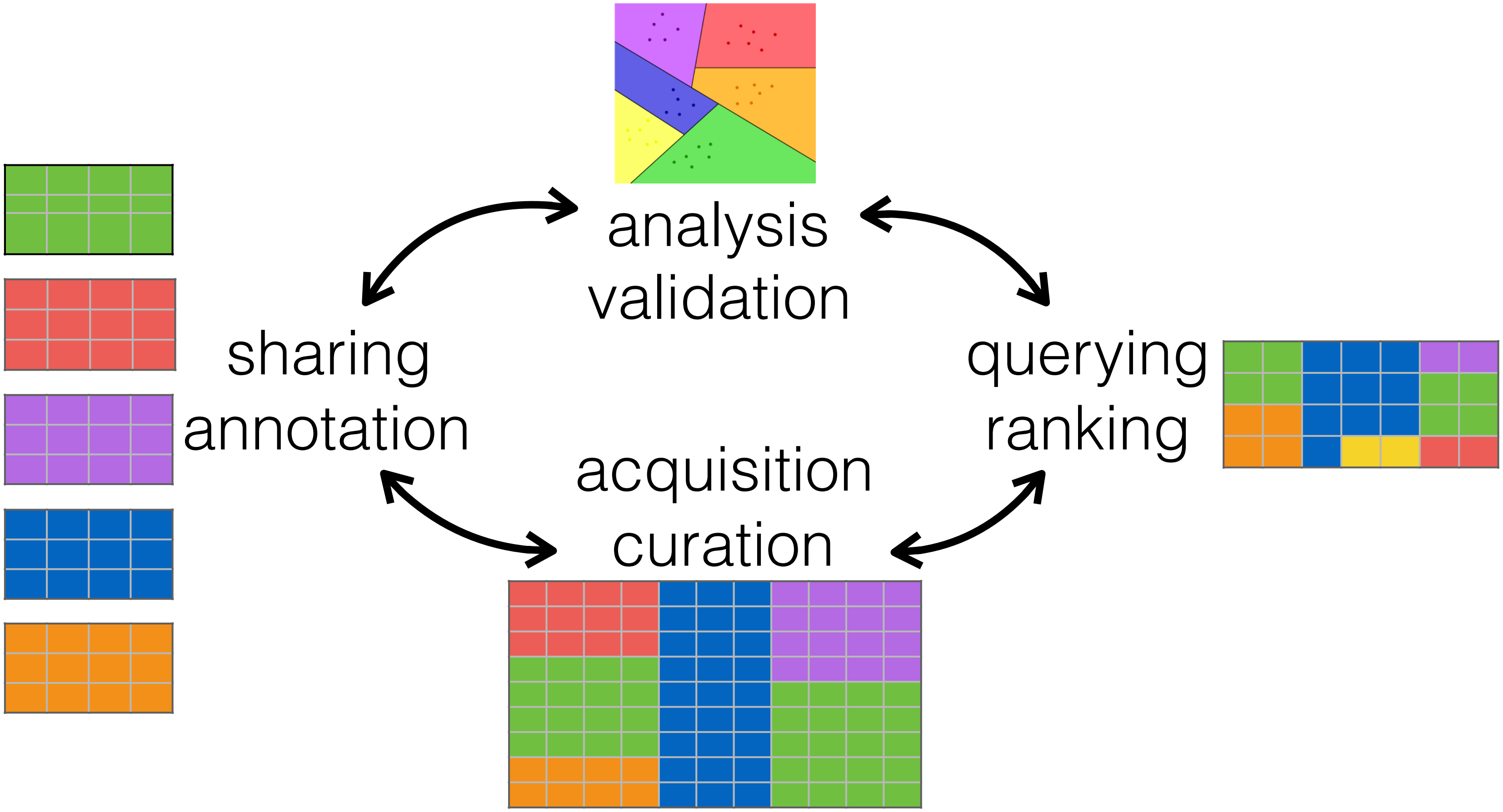
unlikely to help!

fix the model??

minors data was not shared

go back to the data sharing step, help data providers share their data while adhering to laws and upholding the trust of the participants

The data science lifecycle



responsible data science requires a holistic view of the data lifecycle

Codes of ethics

Digital Library | CACM | Queue | TechNews | Learning Center | Career Center

Association for Computing Machinery

Advancing Computing as a Science & Profession

Join | Volunteer | myACM | Search

ABOUT ACM | MEMBERSHIP | PUBLICATIONS | SPECIAL INTEREST GROUPS | CONFERENCES | CHAPTERS | AWARDS | EDUCATION | PUBLIC POLICY | GOVERNANCE

Home > Code Of Ethics

ACM Code of Ethics and Professional Conduct

ACM Code of Ethics and Professional Conduct

Preamble

Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession.

The Code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. Additionally, the Code serves as a basis for remediation when violations occur. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. Each principle is supplemented by guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.

Section 1 outlines fundamental ethical principles that form the basis for the remainder of the Code. Section 2 addresses additional, more specific considerations of professional responsibility. Section 3 guides individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Commitment to ethical conduct is required of every ACM member, and principles involving compliance with the Code are given in Section 4.

The Code as a whole is concerned with how fundamental ethical principles apply to a computing professional's conduct. The Code is not an algorithm for solving ethical problems; rather it serves as a basis for ethical decision-making. When thinking through a particular issue, a computing professional may find that multiple principles should be taken into account, and that different principles will have different relevance to the issue. Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration. The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency.

PDF of the ACM Code of Ethics

On This Page

Codes of ethics

BACK

Community Principles on Ethical Data Practices

OVERVIEW
BACKGROUND
VALUES
PRINCIPLES
AUTHORS
SIGNATORIES

SIGN

JOIN

SUBSCRIBE

This code of ethics for data sharing is created and proposed for adoption by the data science community to reflect the behaviors and principles for the responsible and ethical use and sharing of data by data scientists.

As a community-driven crowdsourced effort, you can join the the discussion and contribute to the next version of the Community Principles on Ethical Data Sharing.

NSF contacts - Google Docs
docs.google.com/document/d/.../edit

OVERVIEW

The Community Principles on Ethical Data Practices are being developed by people from the data science community in conjunction with data science organizations. These principles focus on defining ethical and responsible behaviors for sourcing, sharing and implementing data in a manner that will cause no harm and maximize positive impact. The goal of this initiative is to develop a community-driven code of ethics for data collection, sharing and utilization that provides people in the data science community a standard set of easily digestible, recognizable principles for guiding their behaviors.

This code is not intended to be all encompassing. Rather, these principles will provide academia, industry, and individual data scientists a common set of guidelines for driving the development of standards, curriculums, and best practices for the ethical use and sharing of data, ultimately advancing the responsible and ethical use of data as a collective force for good.