Information Storage and Retrieval

CSCE 670
Texas A&M University
Department of Computer Science & Engineering
Instructor: Prof. James Caverlee

The Dark Side
10 April 2018

https://www.youtube.com/watch?v=4eMYiDaY3-Q

Kumail Nanjiani
@kumailn

Follow

Thread: I know there's a lot of scary stuff in the world rn, but this is something I've been thinking about that I can't get out of my head.

1:56 PM - 1 Nov 2017

19,016 Retweets **40,088** Likes

682 19K 40K

Kumail Nanjiani @kumailn · 1 Nov 2017

As a cast member on a show about tech, our job entails visiting tech companies/conferences etc. We meet ppl eager to show off new tech.

31 545 5.0K

Kumail Nanjiani @kumailn · 1 Nov 2017

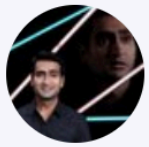
Often we'll see tech that is scary. I don't mean weapons etc. I mean altering video, tech that violates privacy, stuff w obv ethical issues.

41 890 6.5K

Kumail Nanjiani @kumailn · 1 Nov 2017

And we'll bring up our concerns to them. We are realizing that ZERO consideration seems to be given to the ethical implications of tech.

106 1.8K 11K



Kumail Nanjiani @kumailn · 1 Nov 2017

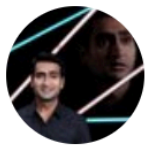
They don't even have a pat rehearsed answer. They are shocked at being asked. Which means nobody is asking those questions.

107 1.0K 10K

Kumail Nanjiani @kumailn · 1 Nov 2017

"We're not making it for that reason but the way ppl choose to use it isn't our fault. Safeguard will develop." But tech is moving so fast.

54 637 7.0K



Kumail Nanjiani @kumailn · 1 Nov 2017

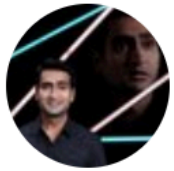
That there is no way humanity or laws can keep up. We don't even know how to deal with open death threats online.

33 639 7.6K

Kumail Nanjiani @kumailn · 1 Nov 2017

Only "Can we do this?" Never "should we do this? We've seen that same blasé attitude in how Twitter or Facebook deal w abuse/fake news.

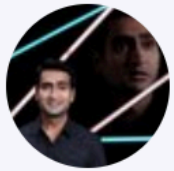
88 1.3K 12K



Kumail Nanjiani @kumailn · 1 Nov 2017

Only "Can we do this?" Never "should we do this? We've seen that same blasé attitude in how Twitter or Facebook deal w abuse/fake news.

88 1.3K 12K



Kumail Nanjiani @kumailn · 1 Nov 2017

Tech has the capacity to destroy us. We see the negative effect of social media. & no ethical considerations are going into dev of tech.

100 1.7K 10K

Kumail Nanjiani @kumailn · 1 Nov 2017

You can't put this stuff back in the box. Once it's out there, it's out there. And there are no guardians. It's terrifying. The end.

471 1.2K 15K

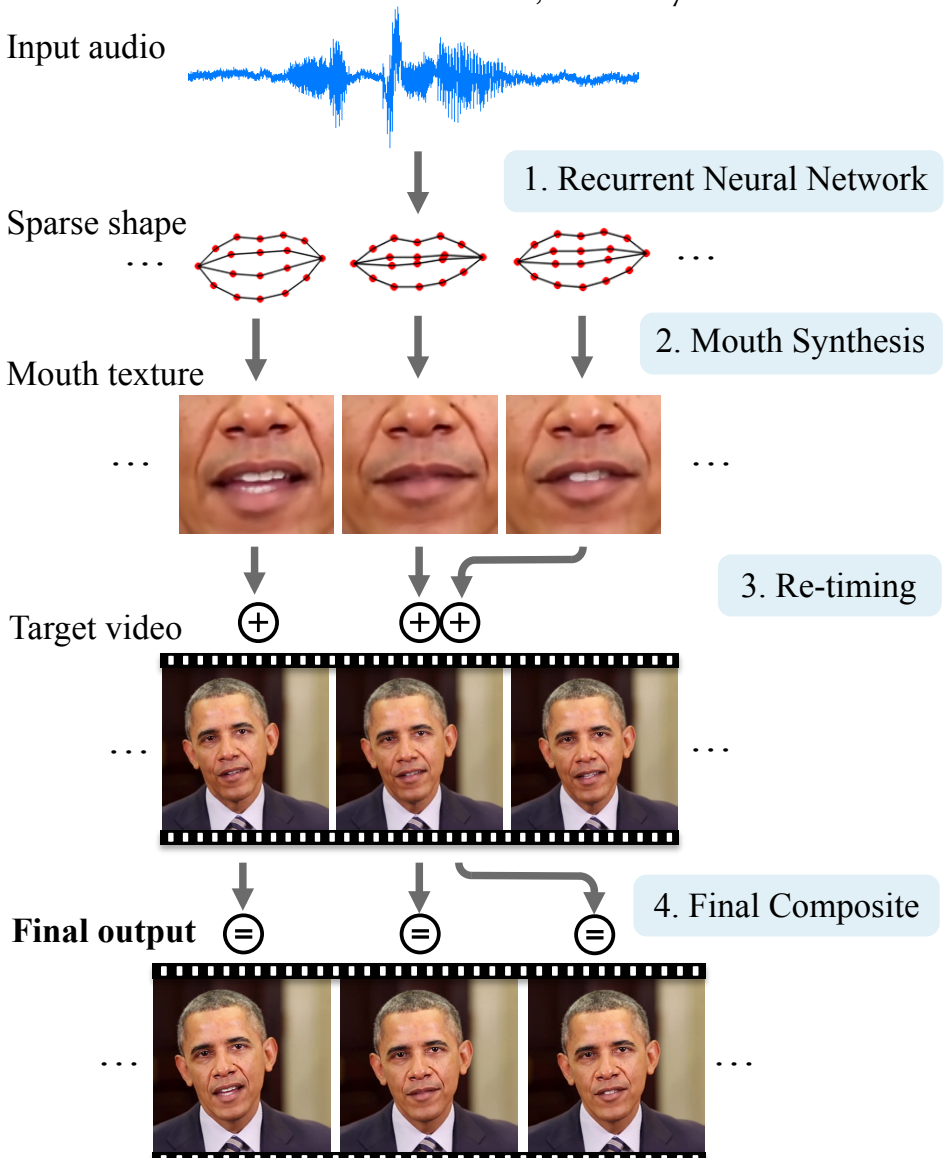
Examples …

- https://grail.cs.washington.edu/projects/AudioToObama/
- https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html

https://grail.cs.washington.edu/projects/AudioToObama/

Synthesizing Obama: Learning Lip Sync from Audio

SUPASORN SUWAJANAKORN, STEVEN M. SEITZ, and IRA KEMELMACHER-SHLIZERMAN, University of Washington



One Idea:

- It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process
- March 29, 2018
- Main idea: Revise peer review process to also focus on negative impacts of our research

https://acm-fca.org/2018/03/29/negativeimpacts/

Example: Crowdwork

- A researcher who invents a new crowdwork framework likely motivates her work by highlighting the problem the framework solves and often the financial benefits of the solution. Crowdwork, however, also comes with serious negative externalities such as incentivizing very low pay [18]. Under our recommendations, this researcher should ideally find ways to engineer her crowdwork framework such that these externalities are structurally mitigated. Alternatively or additionally, she should state what new technologies or policy must complement her system for it to have clear net positive impact. For instance, she might advocate for minimum wage laws to be adapted to a contingent labor context. She should also ensure that her system still has practical use in the context of higher pay.

Example: Blockchain

- Consider a researcher writing a manuscript or whitepaper describing a new blockchain-based technology. Under current practices, this researcher or practitioner would almost certainly not address the serious negative externality of blockchain's energy usage and corresponding carbon footprint [25]. Under our recommendation, this researcher would be urged by gatekeepers (peer reviewers or the press) to discuss this significant, often-unspoken downside to blockchain-based approaches. This is especially the case if – like many blockchain-based technologies – some of the key short-term use cases for the new blockchain-based technology are to better support some societal function in developing countries. Since many of these countries are expected to bear very heavy costs from climate change [34], that the new technology contributes to climate change may significantly complicate the claimed benefits of the technology.

Find Your Group

- What are some of the potential negative impacts of your course project?
- (We'll add this as a component of your final deliverable … to highlight some of the negative aspects.)

Examples

- Review topics — can attack specific topics; suppress free speech
- Spotify recs — recommend based on title of playlists; spam your methods, lead to bad outcomes
- Summarization — news app is using your methods, could make mistakes, create fake news
- Rec — could expose people to ultraviolence,

Rest of This Week: The Dark Side

- The Limits of Machine Learning (?)
- Algorithmic Discrimination
  - Examples of discrimination on the web and in data-driven applications
  - Identifying sources of bias
- Countering Bias: Fairness-aware Algorithms
- Fake News + Hoaxes (Thursday)

The Limits of Machine Learning

Automated Inference on Criminality using Face Images

Xiaolin Wu
McMaster University
Shanghai Jiao Tong University
xwu510@gmail.com

Xi Zhang
Shanghai Jiao Tong University
zhangxi_19930818@sjtu.edu.cn

We study, for the first time, automated inference on criminality based solely on still face images, which is free of any biases of subjective judgments of human observers. Via supervised machine learning, we build four classifiers (logistic regression, KNN, SVM, CNN) using facial images of 1856 real persons controlled for race, gender, age and facial expressions, nearly half of whom were convicted criminals, for discriminating between criminals and non-criminals. All four classifiers perform consistently well and empirically establish the validity of automated face-induced inference on criminality, despite the historical controversy surrounding this line of enquiry. Also, some discriminating structural features for predicting criminality have been found by machine learning. Above all, the most important discovery of this research is that criminal and non-criminal face images populate two quite distinctive manifolds. The variation among criminal faces is significantly greater than that of the non-criminal faces. The two manifolds consisting of criminal and non-criminal faces appear to be concentric, with the non-criminal manifold lying in the kernel with a smaller span, exhibiting a law of "normality" for faces of non-criminals. In other words, the faces of general law-biding public have a greater degree of resemblance compared with the faces of criminals, or criminals have a higher degree of dissimilarity in facial appearance than non-criminals.

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

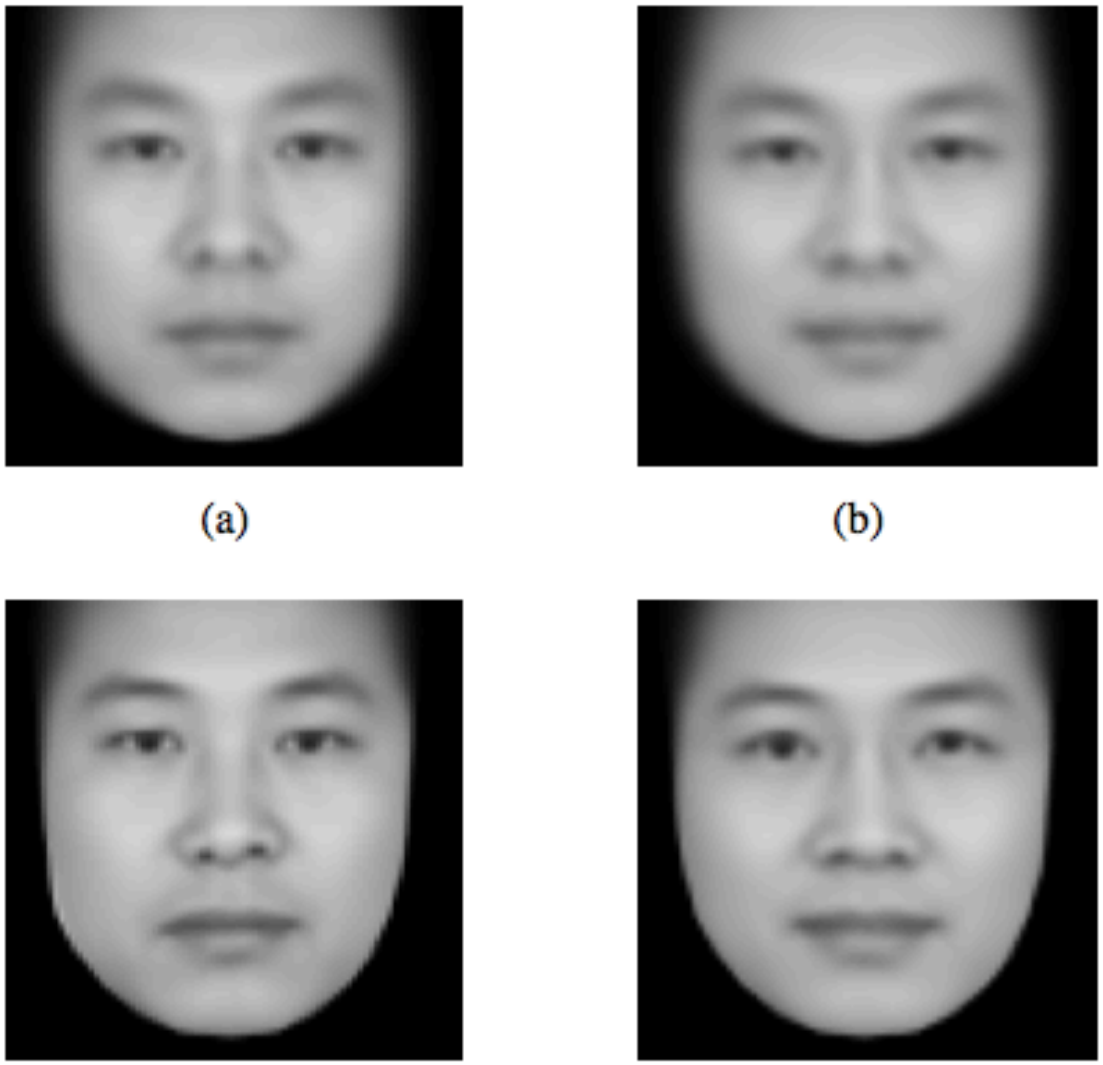
Figure 1. Sample ID photos in our data set.

- From the paper: “Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages, having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal. The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.”
- Thoughts?

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

(c) (d)

Figure 10. (a) and (b) are "average" faces for criminals and non-criminals generated by averaging of eigenface representations ; (c) and (d) are "average" faces for criminals and non-criminals generated by averaging of landmark points and image warping.

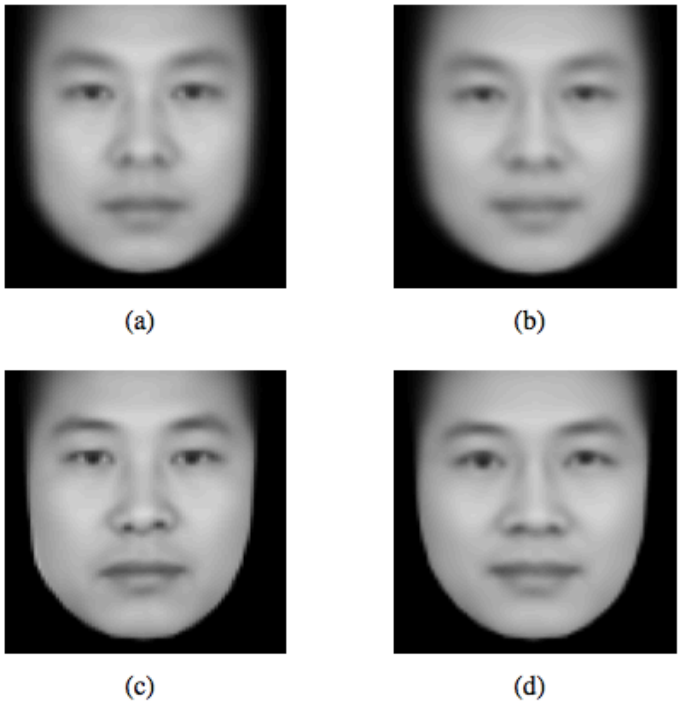
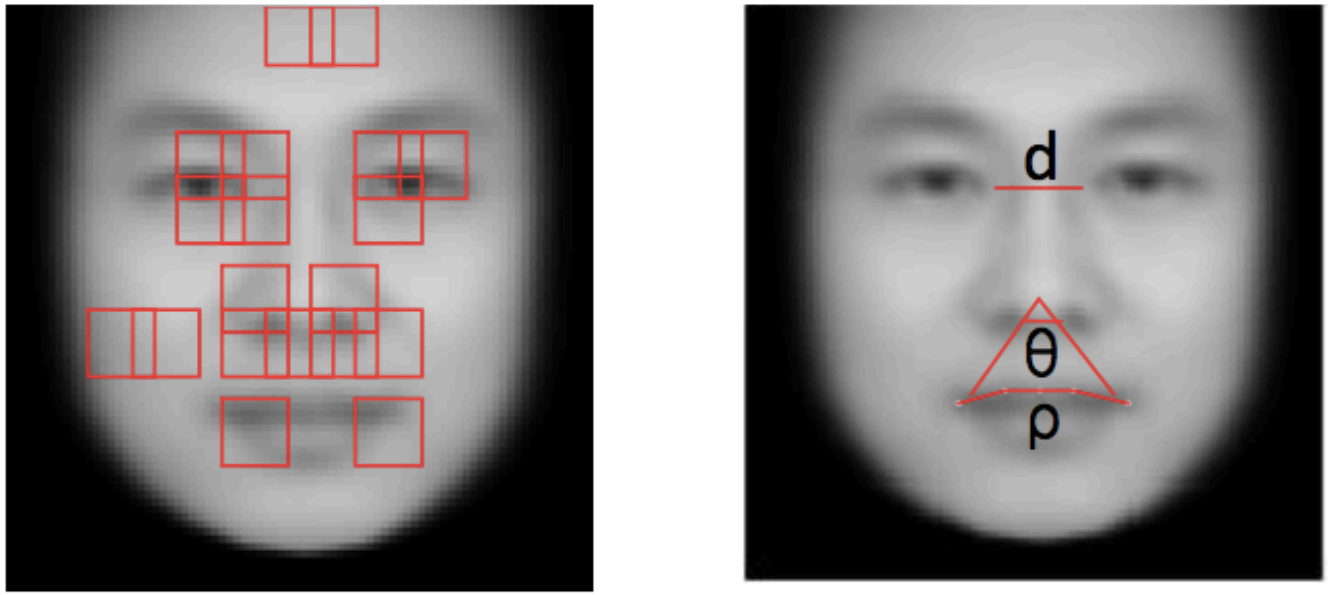


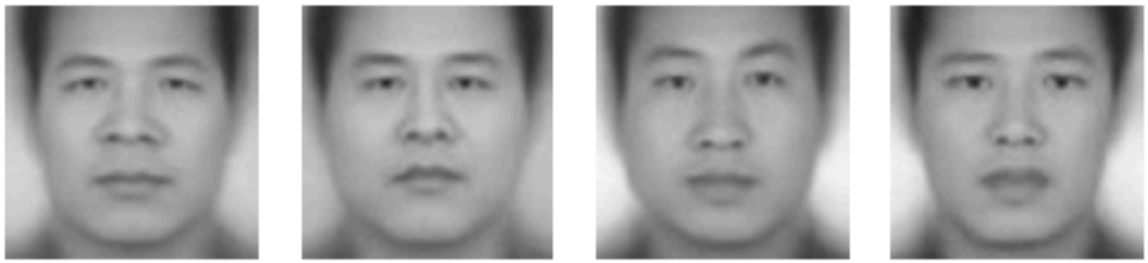
Figure 10. (a) and (b) are "average" faces for criminals and non-criminals generated by averaging of eigenface representations ; (c) and (d) are "average" faces for criminals and non-criminals generated by averaging of landmark points and image warping.

Although the antithesis of criminals and non-criminals is very strong, conventionally-defined average faces of the two populations…appear hardly distinguishable as demonstrated [in the figure above].



(a) (b)

Figure 8. (a) FGM results; (b) Three discriminative features ρ, d and θ.



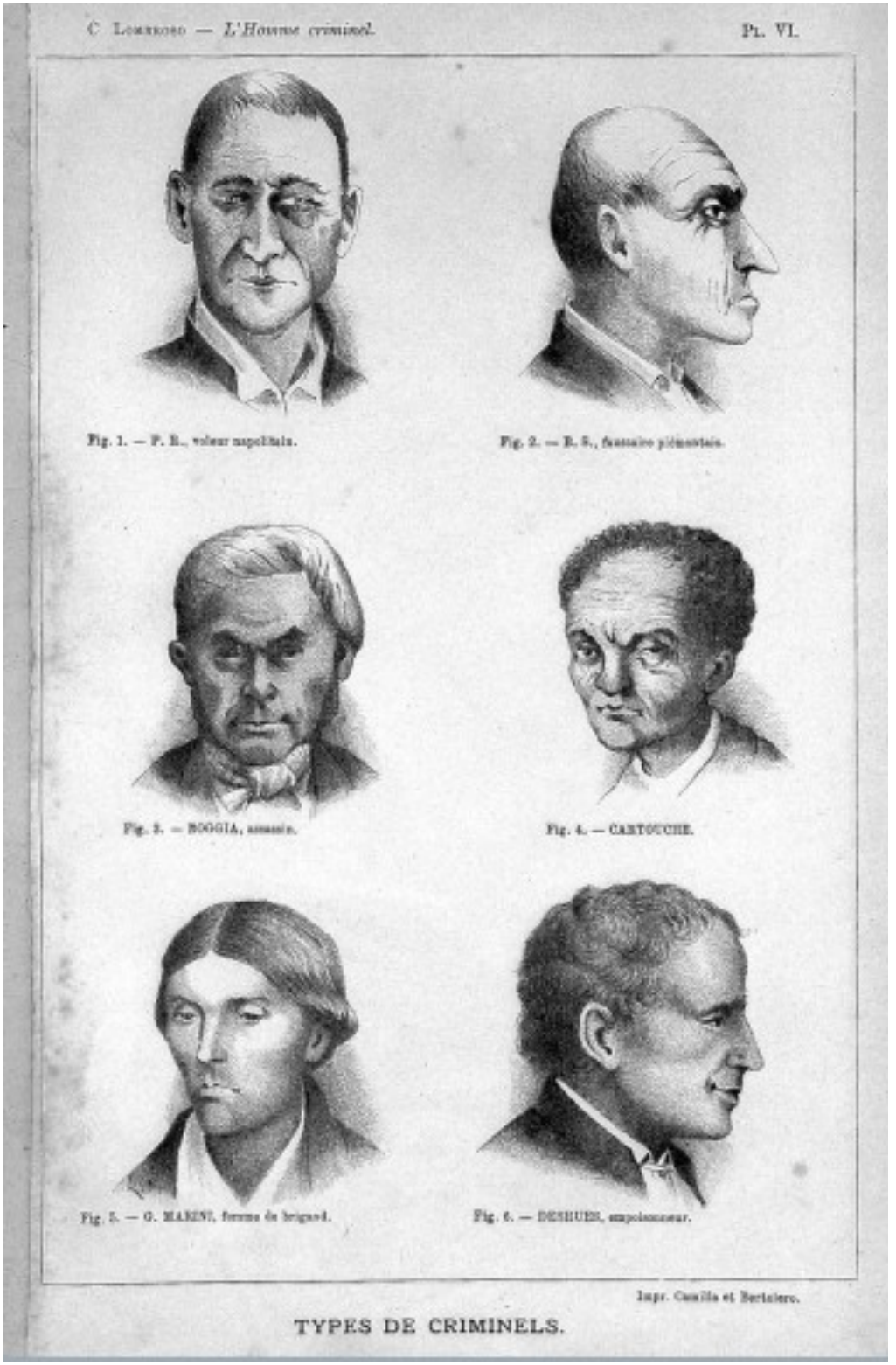
(a) -0.98 (b) -0.68 (c) -0.28 (d) -0.38

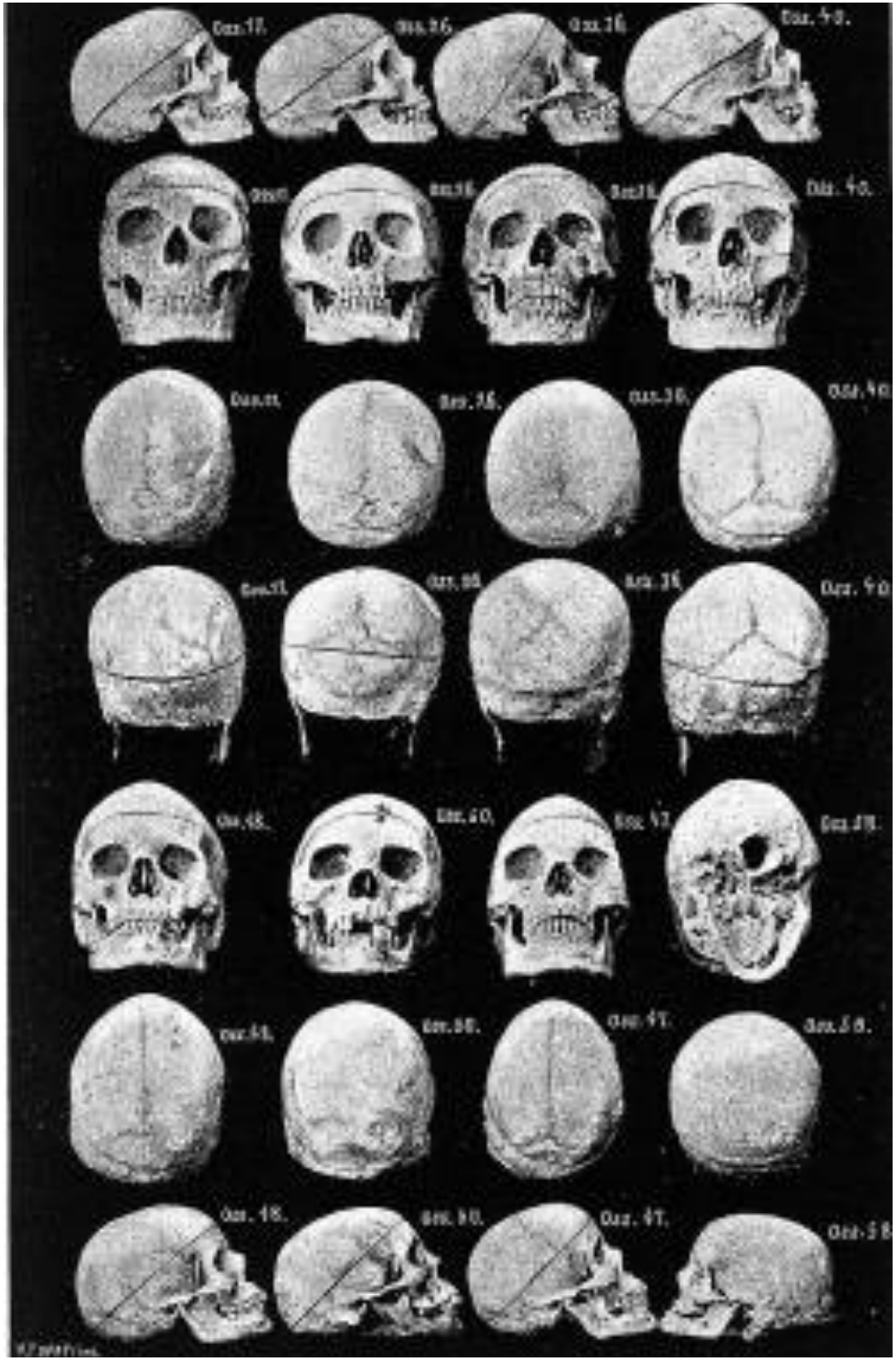
(e) 0.76 (f) 0.98 (g) 0.66

Figure 13. (a), (b), (c) and (d) are the four subtypes of criminal faces corresponding to four cluster centroids on the manifold of S_c; (e), (f) and (g) are the three subtypes of non-criminal faces corresponding to three cluster centroids on the manifold of S_n. The number associated with each face is the average score of human judges (-1 for criminals; 1 for non-criminals).

Sources of bias?

Cesare Lombroso





phrenology = the detailed study of the shape and size of the cranium as a supposed indication of character and mental abilities.

Algorithmic Discrimination

A dangerous reasoning

To discriminate is to treat someone differently

(Unfair) discrimination is based on group membership, not individual merit

People's decisions include objective and subjective elements

Hence, they can be discriminate

Algorithmic inputs include only objective elements

Hence, they cannot discriminate?

Discrimination and Opacity in Online Behavioral Advertising

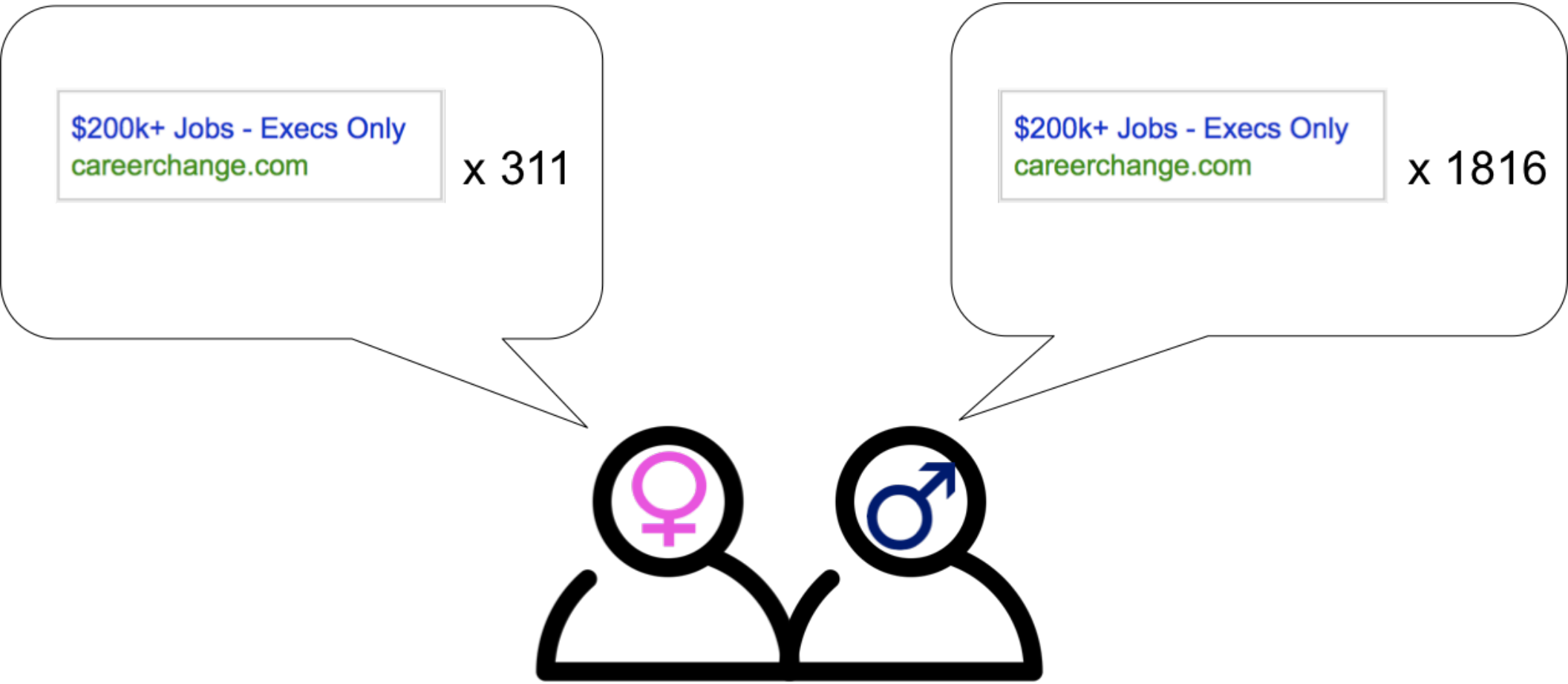
- http://possibility.cylab.cmu.edu/adfisher/
- To study discrimination, we had AdFisher create 1000 fresh browser instances and assign them randomly to two groups. One group set their gender to male on Google's Ad Settings page, while the other set it to female. Then, all the browsers visited the top 100 websites for employment on Alexa. Thereafter, all the browsers collected the ads served by Google on the Times of India.

Top ads for identifying the female group

Ad Title	Ad URL	Times shown to Females	Times shown to Males
Jobs (Hiring Now)	www.jobsinyourarea.co	45	8
4Runner Parts Service	www.westernpatoyotaservice.com	36	5
Criminal Justice Program	www3.mc3.edu/Criminal+Justice	29	1
Goodwill - Hiring	goodwill.careerboutique.com	121	39
UMUC Cyber Training	www.umuc.edu/cybersecuritytraining	38	30

Top ads for identifying the male group

Ad Title	Ad URL	Times shown to Females	Males
$200k+ Jobs - Execs Only	careerchange.com	311	1816
Find Next $200k+ Job	careerchange.com	7	36
Become a Youth Counselor	www.youthcounseling.degreeleap.com	0	310
CDL-A OTR Trucking Jobs	www.tadrivers.com/OTRJobs	0	8
Free Resume Templates	resume-templates.resume-now.com	8	10



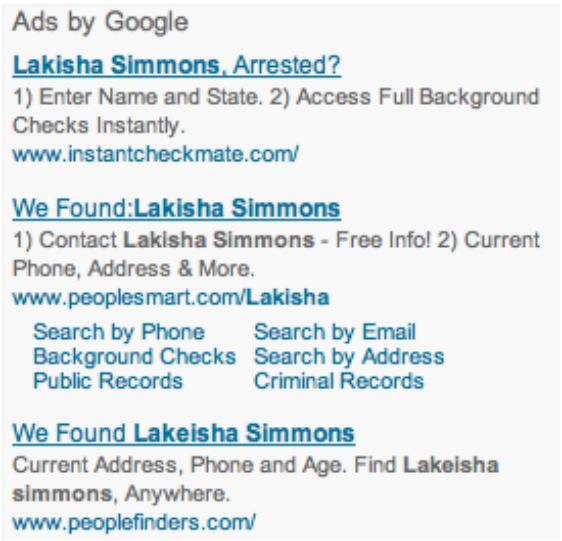
- The top two ads served to the male group was from a career coaching service called careerchange.com that promised high-paying executive level jobs. The top ad was served 1816 times to the male users, but only 311 times to the female users. Of the 500 simulated male users, 402 received the ad at least once, but only 60 female users received the same ad at least once.

Discrimination in Online Ad Delivery, Sweeney 2013

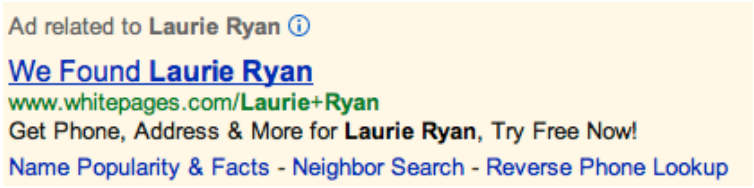
	White Female	Black Female	White Male	Black Male
(a)	Allison Anne Carrie Emily Jill Laurie Kristen Meredith	Aisha Ebony Keisha Kenya Latonya Lakisha Latoya Tamika	Brad Brendan Geoffrey Greg Brett Jay Matthew Neil	Darnell Hakim Jermaine Kareem Jamal Leroy Rasheed Tremayne
(b)	Molly Amy Claire Emily* Katie Madeline Katelyn Emma	Imani Ebony* Shanice Aaliyah Precious Nia Deja Diamond	Jake Connor Tanner Wyatt Cody Dustin Luke Jack	DeShawn DeAndre Marquis Darnell* Terrell Malik Trevon Tyrone
(c)		Latanya Latisha		


Ad related to Lakisha Simmons
Lakisha simmons: Truth
www.instantcheckmate.com/
Arrests and Much More. Everything About Lakisha simmons

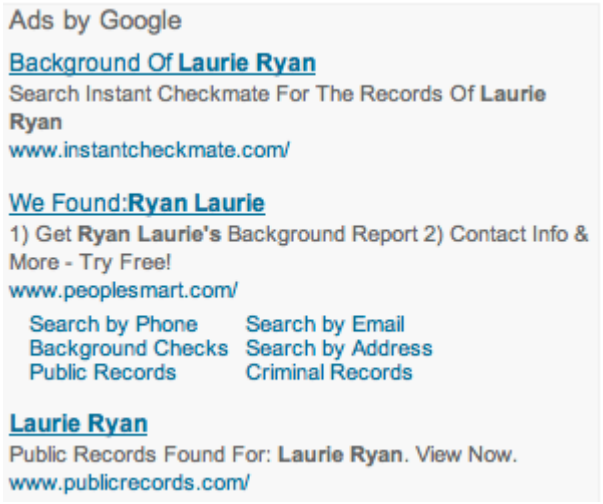
(a)


Ads by Google
Lakisha Simmons, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/
We Found:Lakisha Simmons
1) Contact Lakisha Simmons - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/Lakisha
Search by Phone
Search by Email
Background Checks
Search by Address
Public Records
Criminal Records
We Found Lakeisha Simmons
Current Address, Phone and Age. Find Lakeisha simmons, Anywhere.
www.peoplefinders.com/

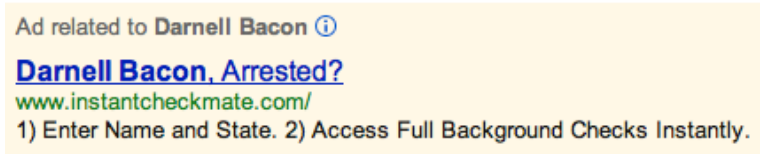
(b)


Ad related to Laurie Ryan
We Found Laurie Ryan
www.whitepages.com/Laurie+Ryan
Get Phone, Address & More for Laurie Ryan, Try Free Now!
Name Popularity & Facts - Neighbor Search - Reverse Phone Lookup

(c)


Ads by Google
Background Of Laurie Ryan
Search Instant Checkmate For The Records Of Laurie Ryan
www.instantcheckmate.com/
We Found:Ryan Laurie
1) Get Ryan Laurie's Background Report 2) Contact Info & More - Try Free!
www.peoplesmart.com/
Search by Phone
Search by Email
Background Checks
Search by Address
Public Records
Criminal Records
Laurie Ryan
Public Records Found For: Laurie Ryan. View Now.
www.publicrecords.com/

(d)

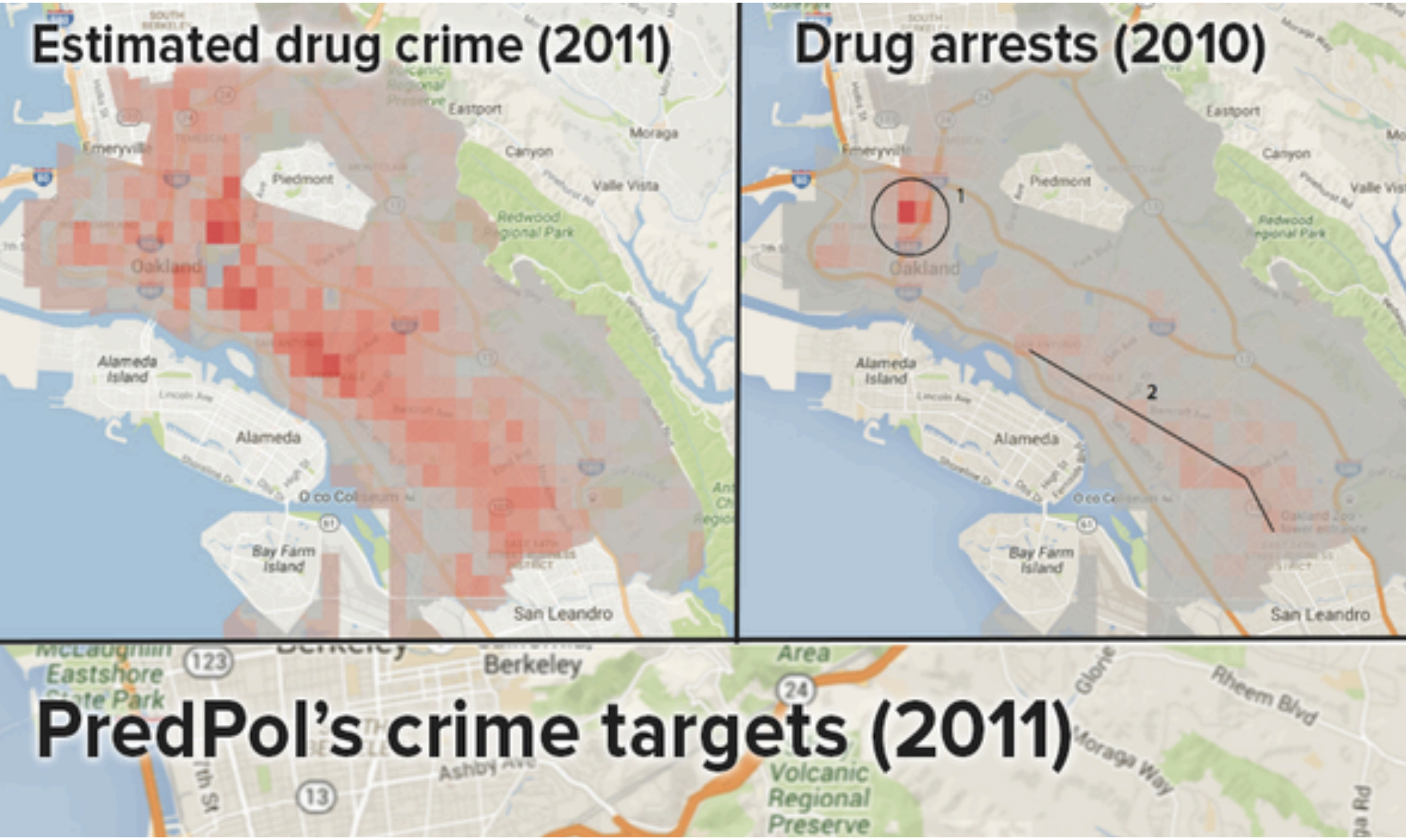

Ad related to Darnell Bacon
Darnell Bacon, Arrested?
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

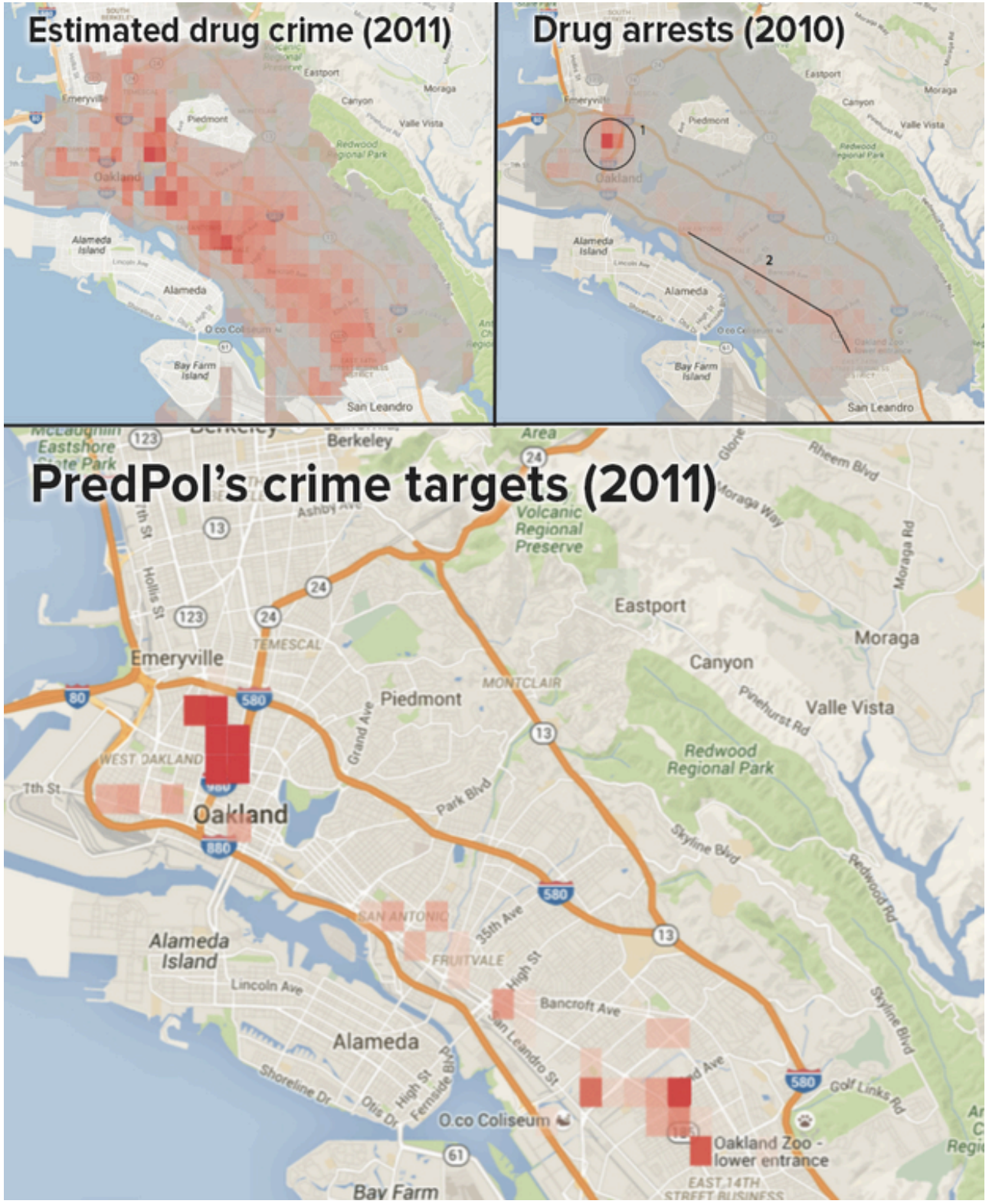

Ads by Google
Darnell Bacon, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/
Darnell Bacon
Public Records Found For: Darnell Bacon. View Now.
www.publicrecords.com/

Discrimination in Online Ad Delivery, Sweeney 2013

- First names, previously identified by others as being assigned at birth to more black or white babies, are found predictive of race (88% black, 96% white), and those assigned primarily to black babies, such as DeShawn, Darnell and Jermaine, generated ads suggestive of an arrest in 81 to 86 percent of name searches on one website and 92 to 95 percent on the other, while those assigned at birth primarily to whites, such as Geoffrey, Jill and Emma, generated more neutral copy: the word "arrest" appeared in 23 to 29 percent of name searches on one site and 0 to 60 percent on the other.

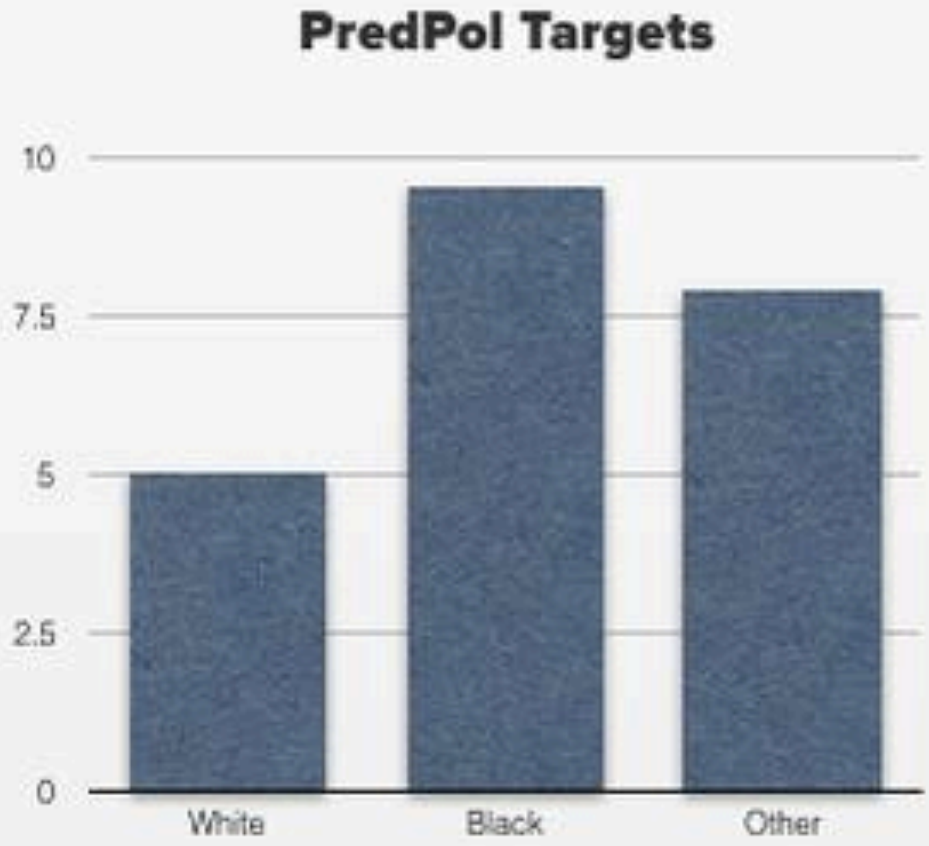

PREDPOL®
The Predictive Policing Company®
MORE THAN A HOTSPOT TOOL
PROVEN & FIELD TESTED
EASY TO DEPLOY & ACCESS
IN THE NEWS

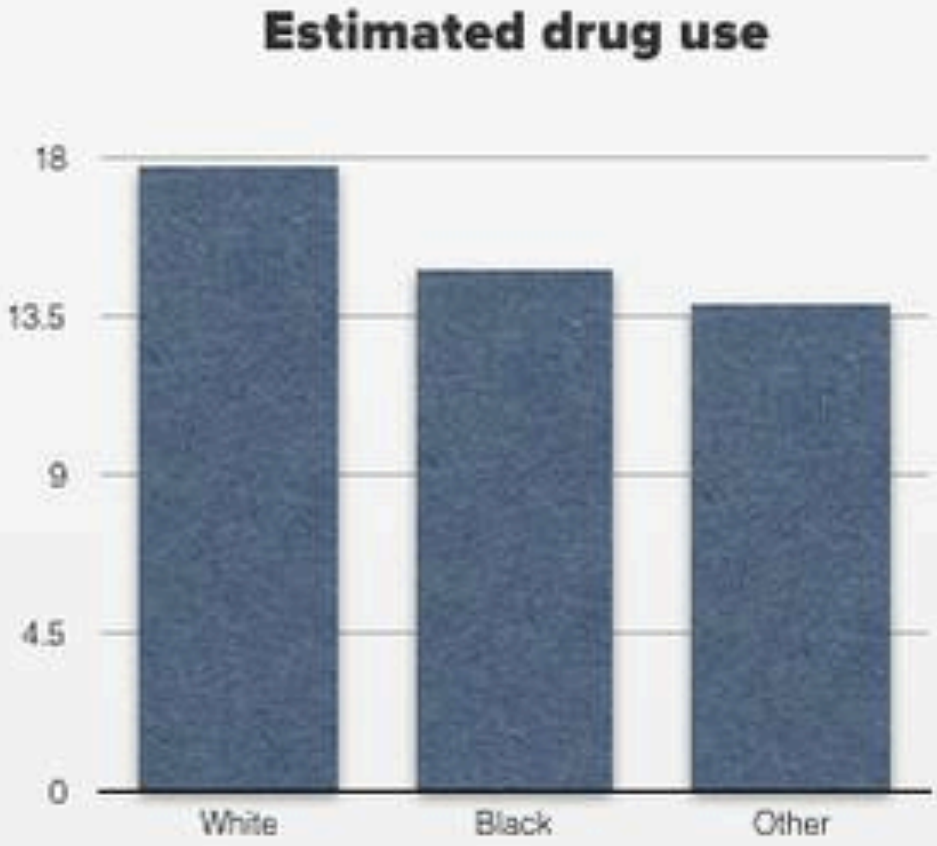

Estimated drug crime (2011)
Drug arrests (2010)
PredPol's crime targets (2011)
Eastport
Moraga
Canyon
Emeryville
Piedmont
Valle Vista
Redwood Regional Park
Oakland
Alameda Island
Alameda
Bay Farm Island
San Leandro
1
2
Berkeley
Eastshore State Park
Volcanic Regional Preserve
Rheem Blvd
Moraga Way
Ashby Ave
123
24
13
580


Estimated drug crime (2011)
Drug arrests (2010)
1
2
PredPol's crime targets (2011)
Berkeley
Emeryville
Piedmont
Oakland
Alameda
Alameda Island
Bay Farm Island
San Leandro
Eastport
Canyon
Moraga
Valle Vista
Redwood Regional Park
Volcanic Regional Preserve
Lincoln Ave
Bancroft Ave
Golf Links Rd
Skyline Blvd
Oakland Zoo - lower entrance
O.co Coliseum

Predictive policing in Oakland vs. actual drug use

The chart on the left shows the demographic breakdown of people targeted for policing based on a simulation of PredPol in Oakland. The chart on the right shows actual estimated use of illicit drugs.





source: National Survey on Drug Use and Health , Human Rights Data Analysis Group

Mic

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

- Stereotype exaggeration
  - e.g., male-dominated professions tend to have even more men in their results than would be expected if the proportions reflected real-world distributions
- Systematic over-/under- representation:
  - Search results also exhibit a slight under-representation of women in images, such that an occupation with 50% women would be expected to have about 45% women in the results on average.

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

- Qualitative differential representation:
  - Image search results also exhibit biases in how genders are depicted: those matching the gender stereotype of a profession tend to be portrayed as more professional-looking and less in-appropriate-looking.
- Perceptions of occupations in search results
  - people's existing perceptions of gender ratios in occupations are quite accurate (R2 of 0.72), but that manipulated search results can have a small but significant effect

Self-perpetuating algorithmic biases

Credit scoring algorithm suggests Joe has high risk of defaulting

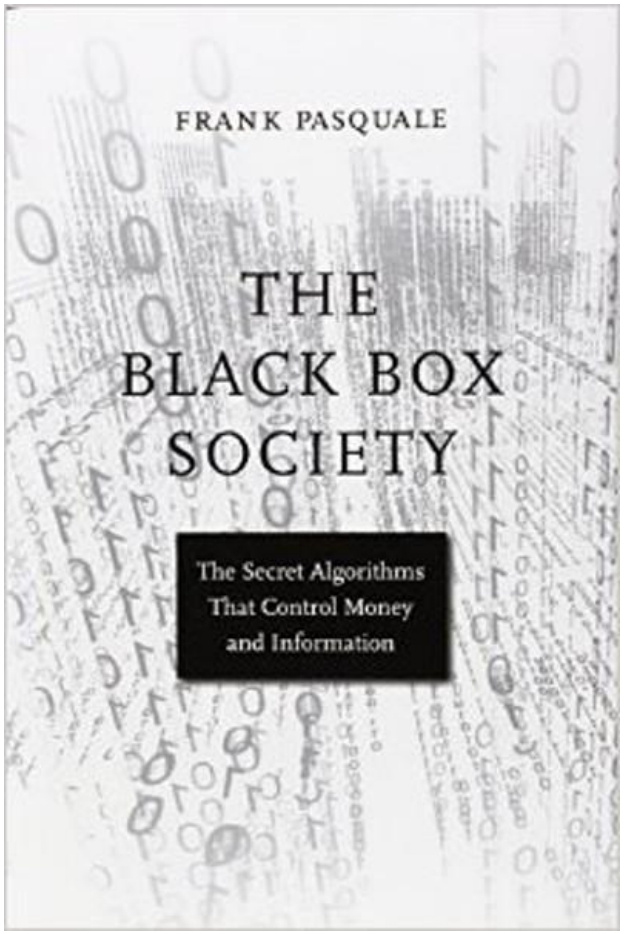
Hence, Joe needs to take a loan at a higher interest rate

Hence, Joe has to make payments that are more onerous

Hence, Joe's risk of defaulting has increased

The same happens with stop-and-frisk of minorities further increasing incarceration rates

To make things worse ...



Algorithms are "black boxes" protected by

- Industrial secrecy
- Legal protections
- Intentional obfuscation

Discrimination becomes invisible

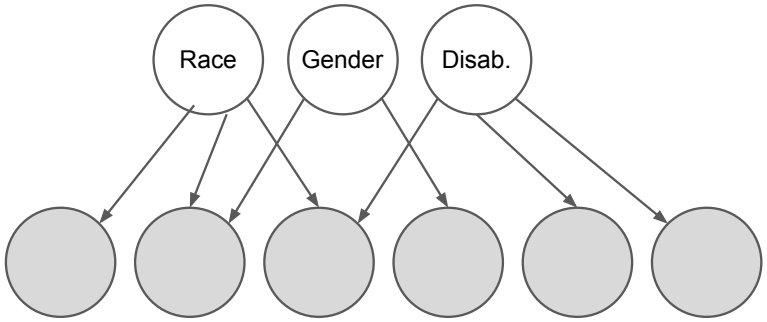
Mitigation becomes impossible

F. Pasquale (2015): The Black Box Society. Harvard University Press.

Some sources of algorithmic bias

Data as a social mirror

Protected attributes redundantly encoded in observables

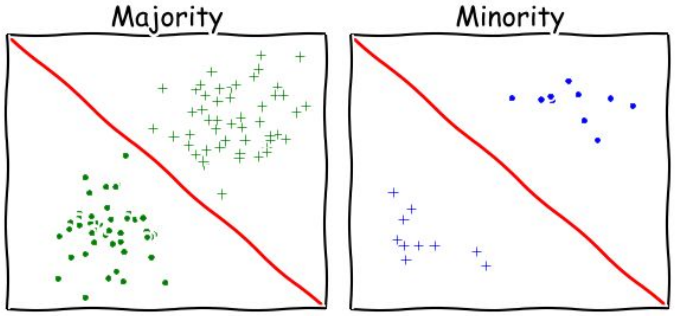


Correctness and completeness

Garbage in, garbage out (GIGO)

Sample size disparity: learn on majority

Errors concentrated in the minority class



M. Hardt (2014): "How big data is unfair". Medium.

Data mining assumptions might not hold

Data mining assumptions are not always observed in reality

- Variables might not be independently identically distributed
- Samples might be biased
- Labels might be incorrect

Errors might be concentrated in a particular class

Sometimes, we might be seeking more simplicity than what is possible

T. Calders and I. Žliobaitė (2013). Why unbiased computational processes can lead to discriminative decision procedures. Chapter 3 of: Discrimination and Privacy in the Information Society. Springer.

Two areas of concern: data and algorithms

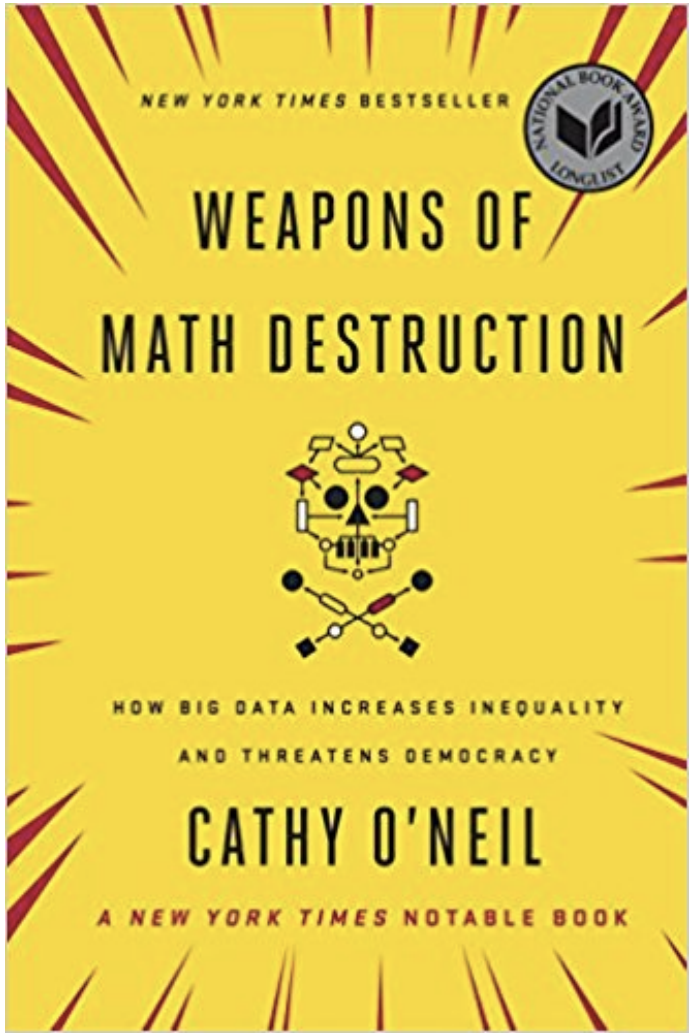
Data inputs:

- Poorly selected (e.g., observe only car trips, not bicycle trips)
- Incomplete, incorrect, or outdated
- Selected with bias (e.g., smartphone users)
- Perpetuating and promoting historical biases (e.g., hiring people that "fit the culture")

Algorithmic processing:

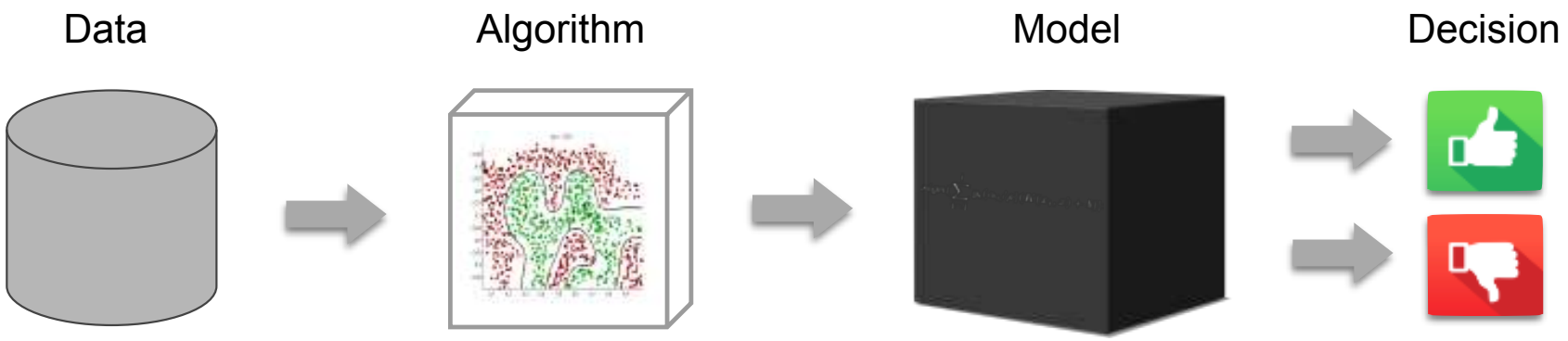
- Poorly designed matching systems
- Personalization and recommendation services that narrow instead of expand user options
- Decision making systems that assume correlation implies causation
- Algorithms that do not compensate for datasets that disproportionately represent populations
- Output models that are hard to understand or explain hinder detection and mitigation of bias

Executive Office of the US President (May 2016): "Big Data: A Report on Algorithmic Systems,Opportunity,and Civil Rights"

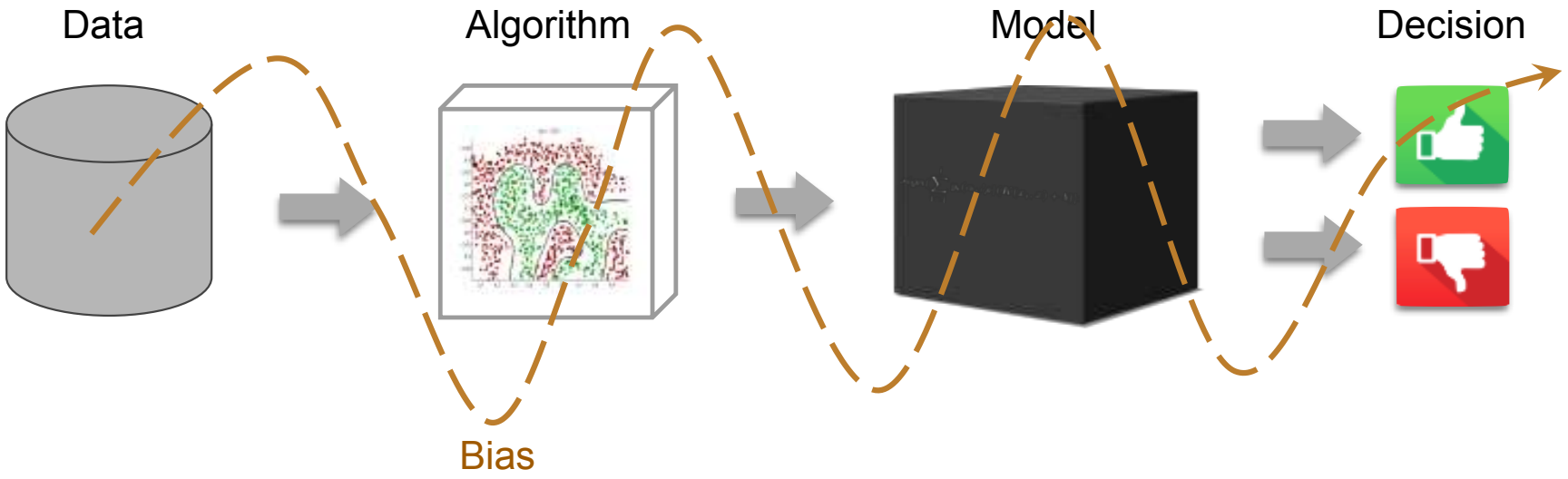

NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK

Fairness-Aware Algorithms

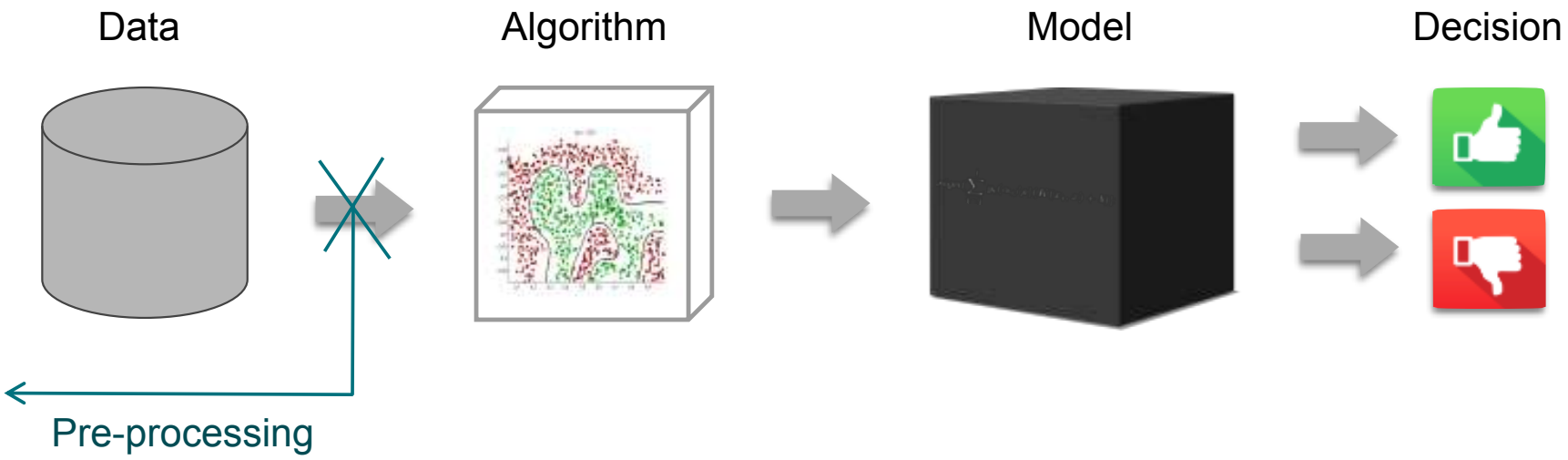
Non-discriminatory data-driven decision-making



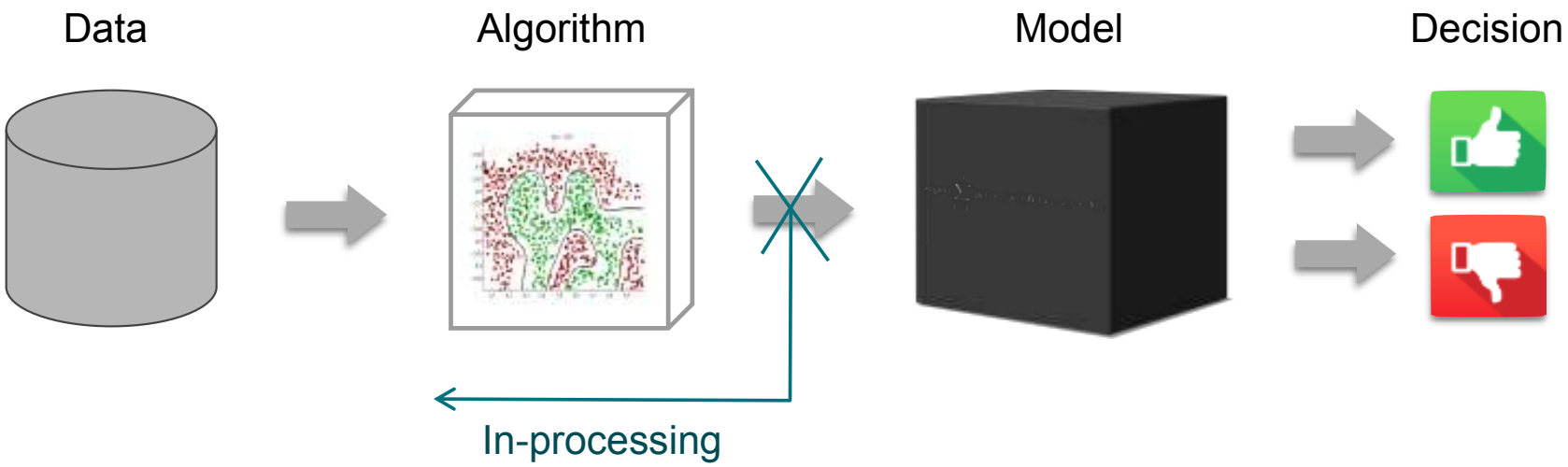
Non-discriminatory data-driven decision-making



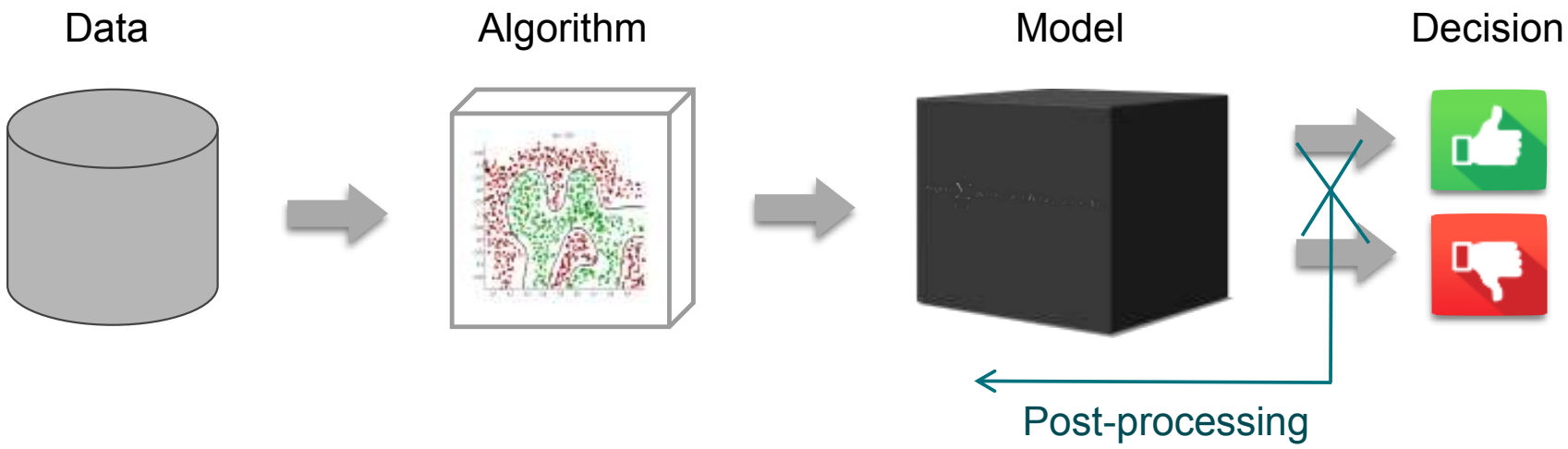
Non-discriminatory data-driven decision-making



Non-discriminatory data-driven decision-making



Non-discriminatory data-driven decision-making



Fairness-aware data mining: common aspects

Goal: develop a non-discriminatory decision-making process while preserving as much as possible the quality of the decision.

Steps:

(1) Defining anti-discrimination/fairness constraints

(2) Transforming data/algorithm/model to satisfy the constraints

(3) Measuring data/model utility

Efficiency Improvement of Neutrality-Enhanced Recommendation, Kamashima et al. 2013

Providing neutral information is important in recommendation

- avoidance of biased recommendation
- fair treatment of content suppliers or item providers
- adherence to laws and regulations in recommendation

↓

Information-neutral Recommender System

The absolutely neutral recommendation is intrinsically infeasible, because recommendation is always biased in a sense that it is arranged for a specific user

↓

This system makes recommendation so as to enhance the neutrality from a viewpoint feature specified by a user

Viewpoint Feature

As in a case of standard recommendation, we use random variables
X: a user, Y: an item, and R: a rating value

We adopt an additional variable for the recommendation neutrality

V : viewpoint feature

- It is specified by a user depending on his/her purpose
- Recommendation results are neutral from this viewpoint
- Its value is determined depending on a user, an item, and their features

Ex. viewpoint = user's gender / movie's release year

Recommendation Neutrality

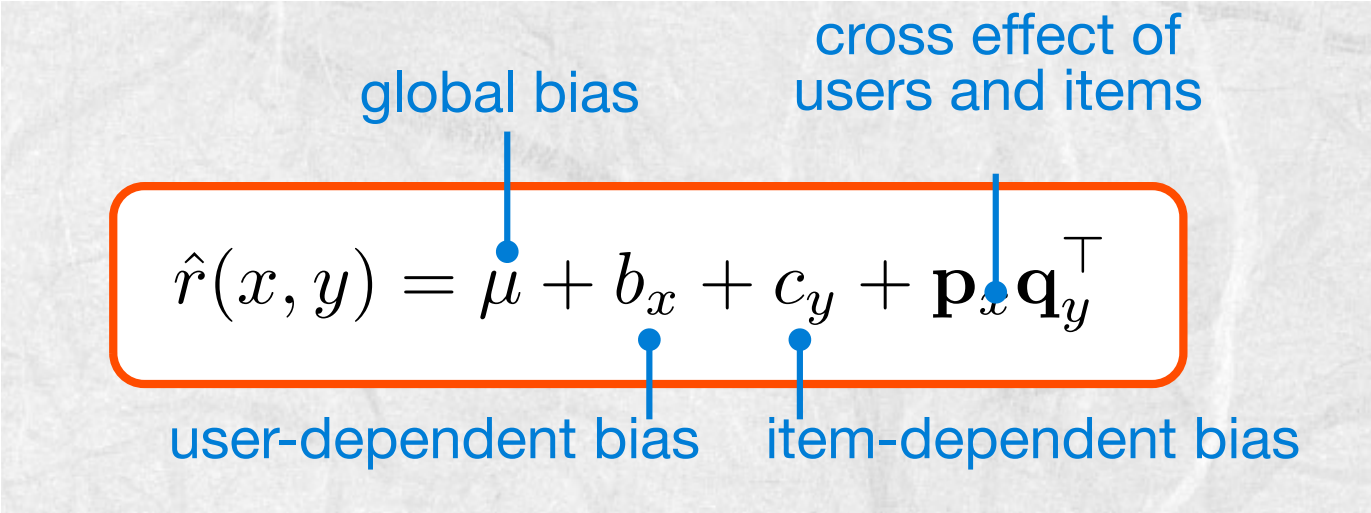
Recommendation Neutrality

- Recommendation results are neutral if no information about a given viewpoint feature does not influence the results
- The status of the specified viewpoint feature is explicitly excluded from the inference of the results

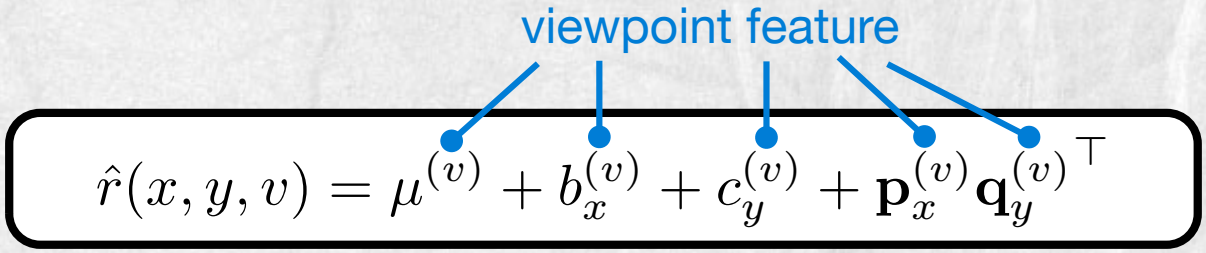
Ex. viewpoint = movie's release year

Whether a movie is new or old does not influence
the inference of whether the movie is recommended or not

If movies A and B are the same except for their release year,
the movie A is always recommended when the movie B is recommended,
and vice versa



adjust ratings according to the state of a viewpoint



- Multiple models are built separately, and each of these models corresponds to the each value of a viewpoint feature
- When predicting ratings, a model is selected according to the value of viewpoint feature

Objective Function of an Information-neutral PMF Model

neutrality parameter to control the balance between the neutrality and accuracy

regularization parameter

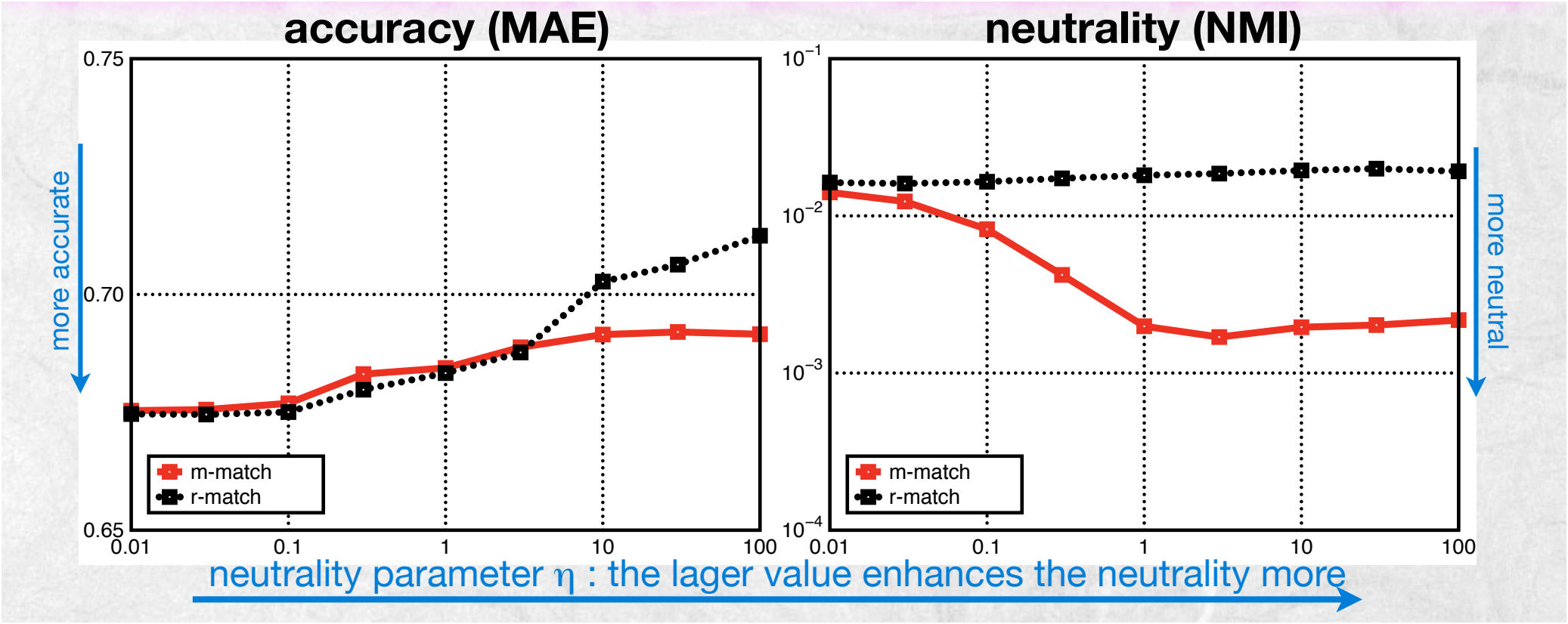
$$\sum_{\mathcal{D}} \left(r_i - \hat{r}(x_i, y_i, v_i)\right)^2 + \eta\, \text{neutral}(R, V) + \lambda \|\boldsymbol{\Theta}\|_2^2$$

squared loss function

neutrality term

L_2 regularizer

Parameters are learned by minimizing this objective function


accuracy (MAE)
neutrality (NMI)
0.75
0.70
0.65
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0.01
0.1
1
10
100
m-match
r-match
more accurate
more neutral
neutrality parameter η : the lager value enhances the neutrality more

https://www.fatml.org/

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.