

# Responsible Data Science

## Transparency & Interpretability

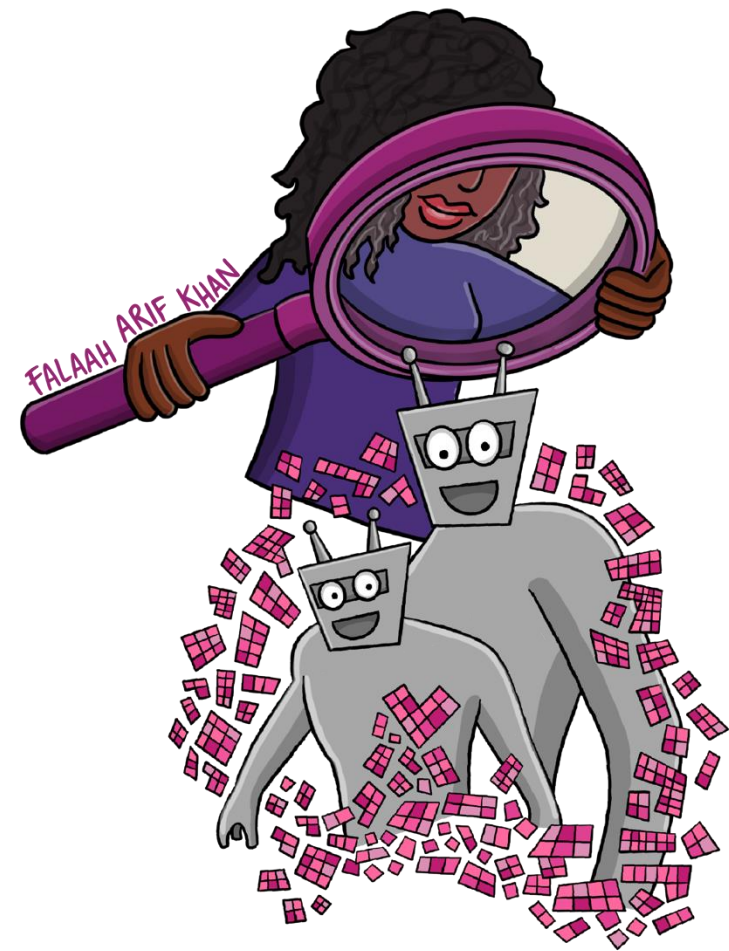
### Transparency in online ad delivery

*March 17, 2024*

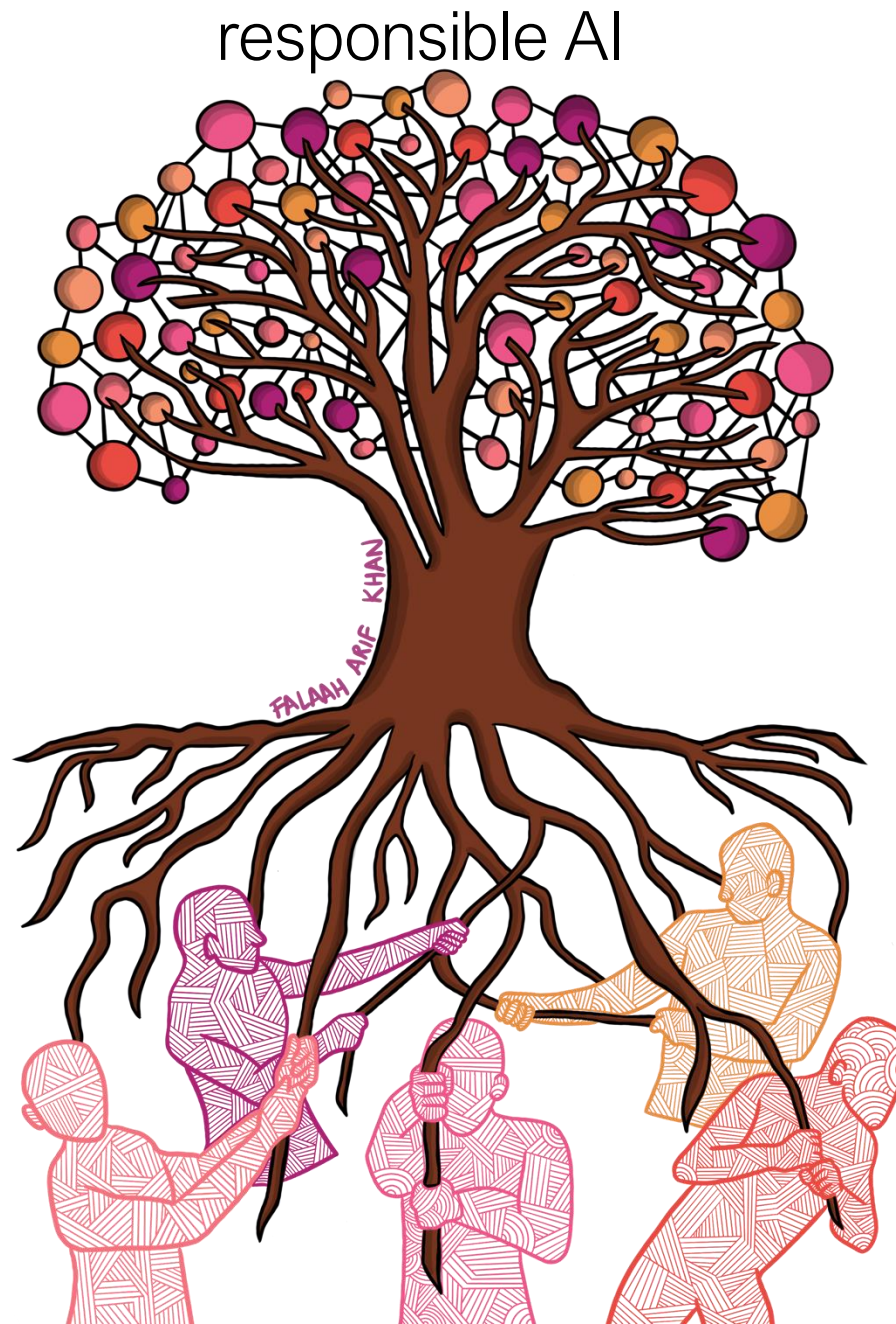
Lucas Rosenblatt

New York University

# Terminology & vision



transparency, interpretability,  
explainability, intelligibility

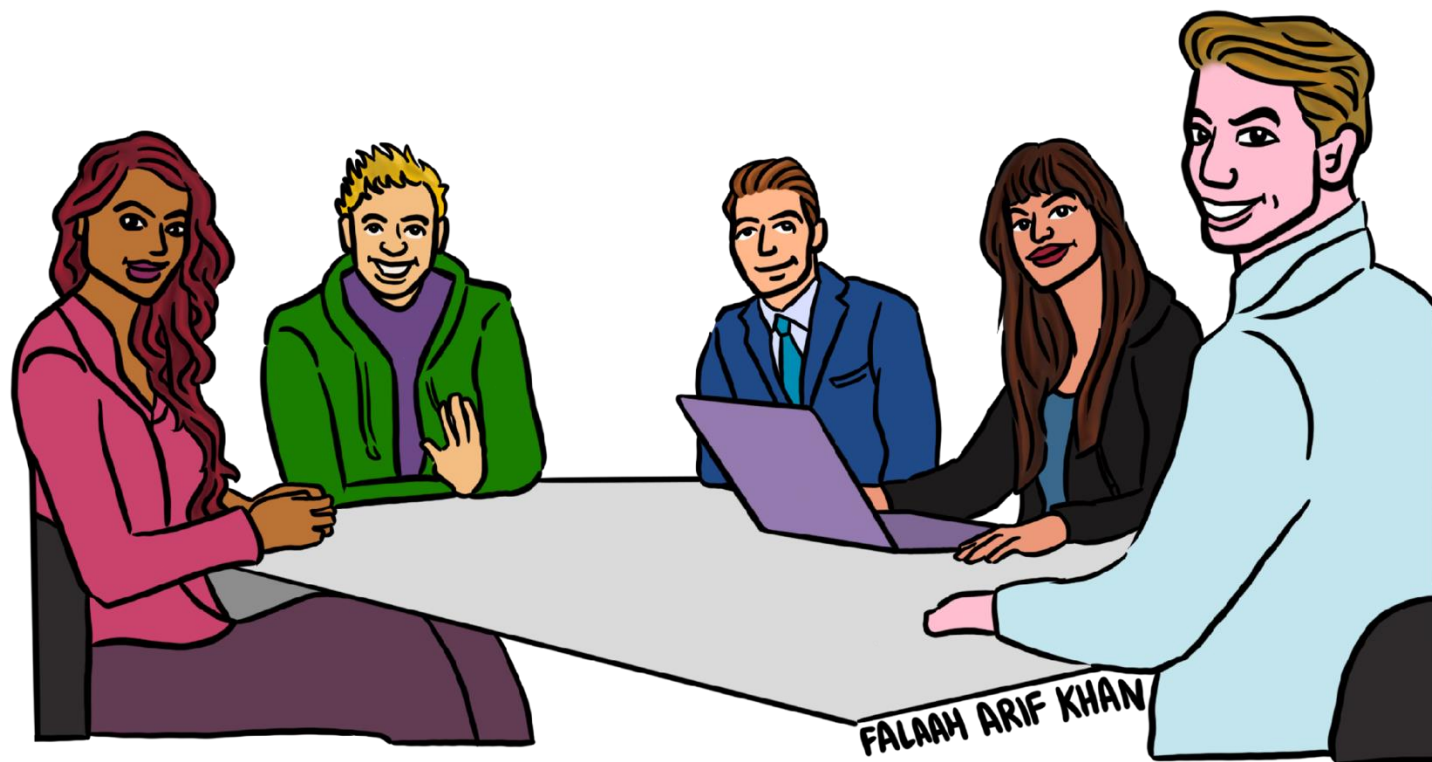


responsible AI



agency, responsibility

# Interpretability for different stakeholders



**What** are we explaining?

To **Whom** are we explaining?

**Why** are we explaining?

Amit Datta\*, Michael Carl Tschantz, and Anupam Datta

## Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

**Abstract:** To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to ensure the statistical soundness of our results. We use AdFisher to find that the Ad Settings was opaque about some features of a user's profile, that it does provide some choice on ads, and that these choices can lead to seemingly discriminatory ads. In particular, we found that visiting webpages associated with substance abuse changed the ads shown but not the settings page. We also found that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male. We cannot determine who caused these findings due to our limited visibility into the ad ecosystem, which includes Google, advertisers, websites, and users. Nevertheless, these results can form the starting point for deeper investigations by either the companies themselves or by regulatory bodies.

**Keywords:** blackbox analysis, information flow, behavioral advertising, transparency, choice, discrimination  
DOI 10.1515/popets-2015-0007  
Received 11/22/2014; revised 2/18/2015; accepted 2/18/2015.

**Keywords:** blackbox analysis, information flow, behavioral advertising, transparency, choice, discrimination  
DOI 10.1515/popets-2015-0007  
Received 11/22/2014; revised 2/18/2015; accepted 2/18/2015.

### 1 Introduction

**Problem and Overview.** With the advancement of tracking technologies and the growth of online data aggregators, data collection on the Internet has become a

\*Corresponding Author: **Amit Datta:** Carnegie Mellon University, E-mail: amiddatta@cmu.edu  
**Michael Carl Tschantz:** International Computer Science Institute, E-mail: mct@icsi.berkeley.edu  
**Anupam Datta:** Carnegie Mellon University, E-mail: danna-pam@cmu.edu

serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

People are concerned about behavioral marketing on the web (e.g., [2]). To increase transparency and control, Google provides Ad Settings, which is "a Google tool that helps you control the ads you see on Google services and on websites that partner with Google" [3]. It displays inferences Google has made about a user's demographics and interests based on his browsing behavior. Users can view and edit these settings at <http://www.google.com/settings/ads>. Yahoo [4] and Microsoft [5] also offer personalized ad settings.

However, they provide little information about how these pages operate, leaving open the question of how completely these settings describe the profile they have about a user. In this study, we explore how a user's behaviors, either directly with the settings or with content providers, alter the ads and settings shown to the user and whether these changes are in harmony. In particular, we study the degree to which the settings provides transparency and choice as well as checking for the presence of discrimination. Transparency is important for people to understand how the use of data about them affects the ads they see. Choice allows users to control how this data gets used, enabling them to protect the information they find sensitive. Discrimination is an increasing concern about machine learning systems and one reason people like to keep information private [6, 7].

To conduct these studies, we developed AdFisher, a tool for automating randomized, controlled experiments for studying online tracking. Our tool offers a combination of automation, statistical rigor, scalability, and explanation for determining the use of information by web advertising algorithms and by personalized ad settings, such as Google Ad Settings. The tool can simulate having a particular interest or attribute by visiting web-

## Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes

MUHAMMAD ALI\*, Northeastern University, USA  
PIOTR SAPIEZYNSKI\*, Northeastern University, USA  
MIRANDA BOGEN, Upturn, USA  
ALEKSANDRA KOROLOVA, University of Southern California, USA  
ALAN MISLOVE, Northeastern University, USA  
AARON RIEKE, Upturn, USA

The enormous financial success of online advertising platforms is partially due to the precise targeting features they offer. Although researchers and journalists have found many ways that advertisers can target—or exclude—particular groups of users seeing their ads, comparatively little attention has been paid to the implications of the platform's *ad delivery* process, comprised of the platform's choices about which users see which ads.

It has been hypothesized that this process can "skew" ad delivery in ways that the advertisers do not intend, making some users less likely than others to see particular ads based on their demographic characteristics. In this paper, we demonstrate that such skewed delivery occurs on Facebook, due to market and financial optimization effects as well as the platform's own predictions about the "relevance" of ads to different groups of users. We find that both the advertiser's budget and the content of the ad each significantly contribute to the skew of Facebook's ad delivery. Critically, we observe significant skew in delivery along gender and racial lines for "real" ads for employment and housing opportunities despite neutral targeting parameters.

Our results demonstrate previously unknown mechanisms that can lead to potentially discriminatory ad delivery, even when advertisers set their targeting parameters to be highly inclusive. This underscores the need for policymakers and platforms to carefully consider the role of the ad delivery optimization run by ad platforms themselves—and not just the targeting choices of advertisers—in preventing discrimination in digital advertising.<sup>1</sup>

CCS Concepts: • Information systems → Social advertising; • Human-centered computing → Empirical studies in HCI; • Applied computing → Law;

Keywords: online advertising; ad delivery; bias; fairness; policy

### ACM Reference Format:

Muhammad Ali\*, Piotr Sapiezynski\*, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 199 (November 2019). ACM, New York, NY. 30 pages. <https://doi.org/10.1145/3359301>

<sup>1</sup>The delivery statistics of ad campaigns described in this work can be accessed at <https://facebook-targeting.ccs.neu.edu/>. \*These two authors contributed equally

Authors' addresses: Muhammad Ali\*, Northeastern University, USA, mali@ccs.neu.edu; Piotr Sapiezynski\*, Northeastern University, USA, sapiezynski@gmail.com; Miranda Bogen, Upturn, USA, mirandabogen@gmail.com; Aleksandra Korolova, University of Southern California, USA, korolova@usc.edu; Alan Mislove, Northeastern University, USA, amislove@ccs.neu.edu; Aaron Rieke, Upturn, USA, aaron@upturn.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2019/11-ART199 \$15.00

<https://doi.org/10.1145/3359301>

Proceedings of the ACM on Human-Computer Interaction, Vol. 3, No. CSCW, Article 199. Publication date: November 2019.

DOI:10.1145/2447976.2447990

Article development led by ACM Queue  
[queue.acm.org](http://queue.acm.org)

Google ads, black names and white names, racial discrimination, and click advertising.

BY LATANYA SWEENEY

# Discrimination in Online Ad Delivery

DO ONLINE ADS suggestive of arrest records appear more often with searches of black-sounding names than white-sounding names? What is a black-sounding name or white-sounding name, anyway? How do you design technology to reason about societal consequences like structural racism? Let's take a scientific dive into online ad delivery to find answers.

"Have you ever been arrested?" Imagine this question appearing whenever someone enters your name in a search engine. Perhaps you are in competition for an award or a new job, or maybe you are in a position of trust, such as a professor or a volunteer. Perhaps you are dating or engaged in any one of hundreds of circumstances for which someone wants to learn more about you online. Appearing alongside your accomplishments is an advertisement implying you may have a criminal record, whether you actually have one or not. Worse, the ads may not appear for your competitors.

Employers frequently ask whether applicants have ever been arrested or charged with a crime, but if an employer disqualifies a job applicant based solely upon information indicating an arrest record, the company may face legal consequences. The U.S. Equal Employment Opportunity Commission (EEOC) is the federal agency charged with enforcing Title VII of the Civil Rights Act of 1964, a law that applies to most employers, prohibiting employment discrimination based on race, color, religion, sex, or national origin, and extended to those having criminal records.<sup>5,12</sup> Title VII does not prohibit employers from obtaining criminal background information, but a blanket policy of excluding applicants based solely upon information indicating an arrest record can result in a charge of discrimination.

To make a determination, the EEOC uses an adverse impact test that measures whether certain practices, intentional or not, have a disproportionate effect on a group of people whose defining characteristics are covered by Title VII. To decide, you calculate the percentage of people affected in each group and then divide the smaller value by the larger to get the ratio and compare the result to 80. If the ratio is less than 80, then the EEOC considers the effect disproportionate and may hold the employer responsible for discrimination.<sup>9</sup>

What about online ads suggesting someone with your name has an arrest record? Title VII only applies if you have an arrest record and can prove the employer inappropriately used the ads.

Are the ads commercial free speech—a constitutional right to display the ad associated with your name? The First Amendment of the U.S. Constitution protects advertising, but the U.S. Supreme Court set out a test for assessing restrictions on commercial speech, which begins by determining whether the speech is misleading.<sup>7</sup> Are online ads suggesting the existence of an arrest record misleading if no one by that name has an arrest record?

ILLUSTRATION BY ALEX WILLIAMSON

*a closer look at LIME*

# LIME: Local Interpretable Model-Agnostic Explanations

## Why should I trust you?

Explaining the predictions of any classifier



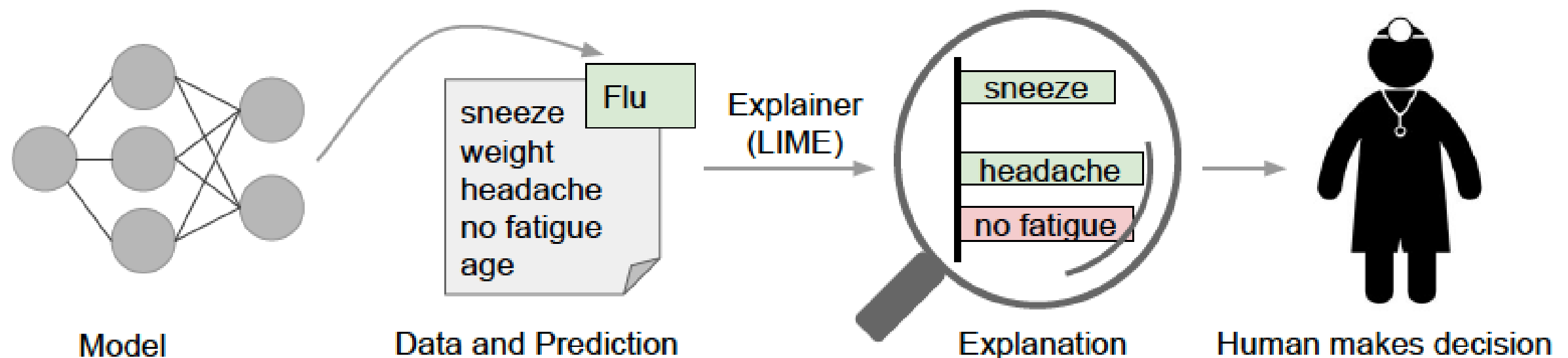
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Check out our paper, and open source project at  
<https://github.com/marcotcr/lime>

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

# LIME: Explanations based on features

- **LIME** (Local Interpretable Model-Agnostic Explanations): to help users trust a prediction, explain individual predictions
- **SP-LIME**: to help users trust a model, select a set of representative instances for which to generate explanations



features in green (“sneeze”, “headache”) support the prediction (“Flu”), while features in red (“no fatigue”) are evidence against the prediction

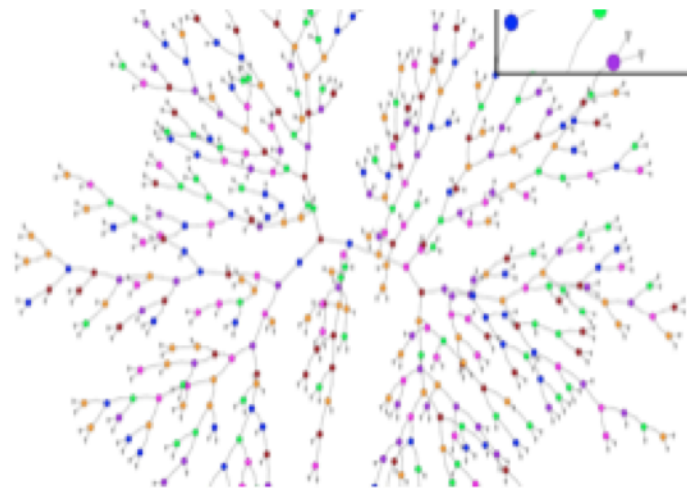
what if patient id appears in green in the list? - an example of “data leakage”

# LIME: Local explanations of classifiers

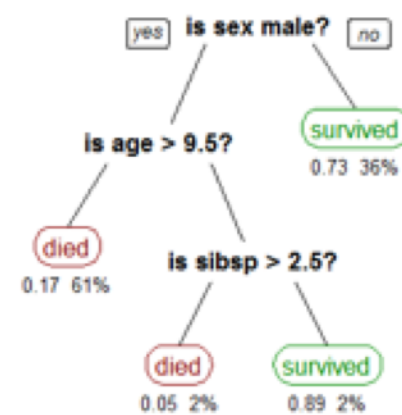
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely  
not interpretable



Potentially  
interpretable

slide by Marco Tulio Ribeiro, KDD 2016



# Explanations based on features

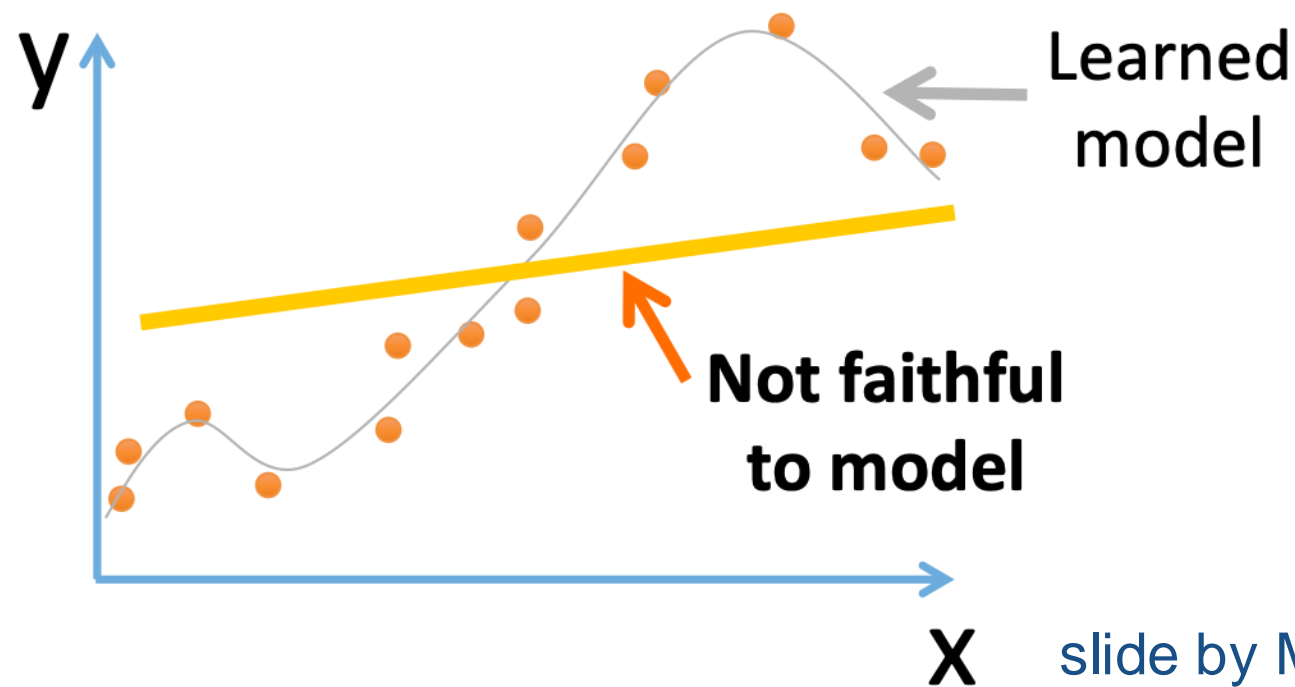
## Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

Faithful

- Describes how this model actually behaves



slide by Marco Tulio Ribeiro, KDD 2016

# Explanations based on features

## Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

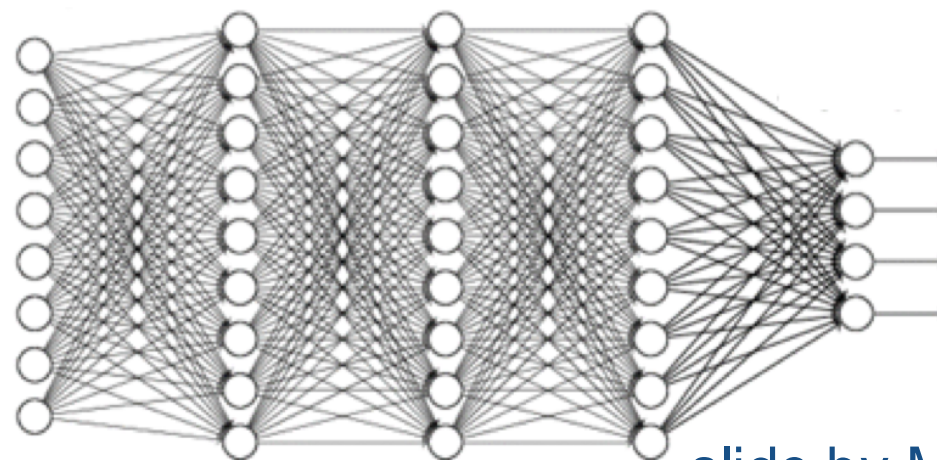
Faithful

- 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

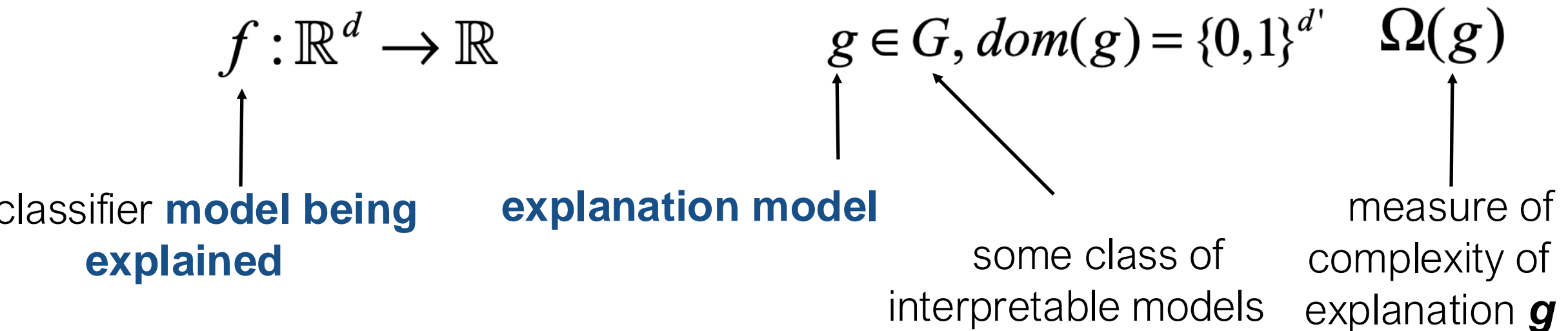
# Key idea: Interpretable representation

“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

- LIME relies on a distinction between **features** and **interpretable data representations**; examples:
  - In text classification features are word embeddings; an interpretable representation is a vector indicating the presence or absence of a word
  - In image classification features encoded in a tensor with three color channels per pixel; an interpretable representation is a binary vector indicating the presence or absence of a contiguous patch of similar pixels
- **To summarize:** we may have some  $d$  features and  $d'$  interpretable components; interpretable models will act over domain  $\{0, 1\}^{d'}$  - denoting the presence or absence of each of  $d'$  interpretable components

# Fidelity-interpretability trade-off

“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”



$f(\mathbf{x})$  denotes the probability that  $\mathbf{x}$  belongs to some class

$\pi_x$  is a **proximity measure** relative to  $\mathbf{x}$

we make no assumptions about  $f$   
to remain model-agnostic: draw  
samples weighted by  $\pi_x$

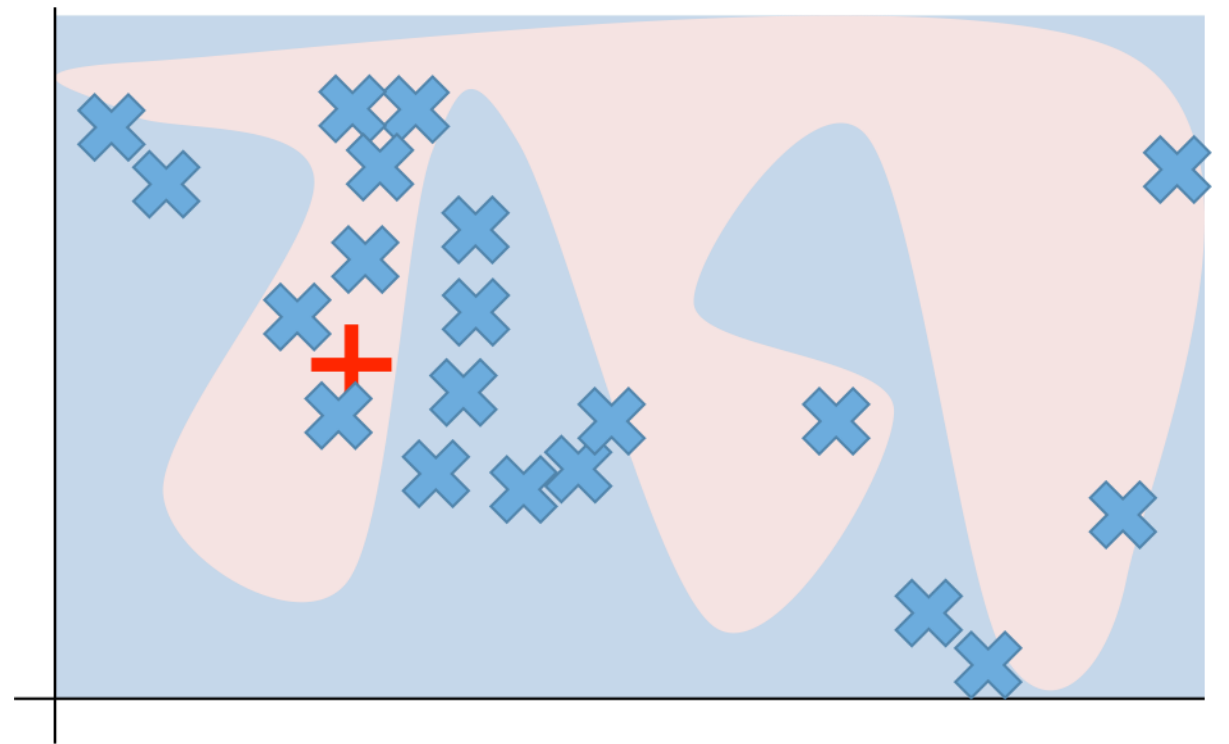
**explanation** measures how unfaithful is  $g$  to  $f$  in the locality around  $\mathbf{x}$

$$\xi(\mathbf{x}) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

# Fidelity-interpretability trade-off

“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

1. sample points around 

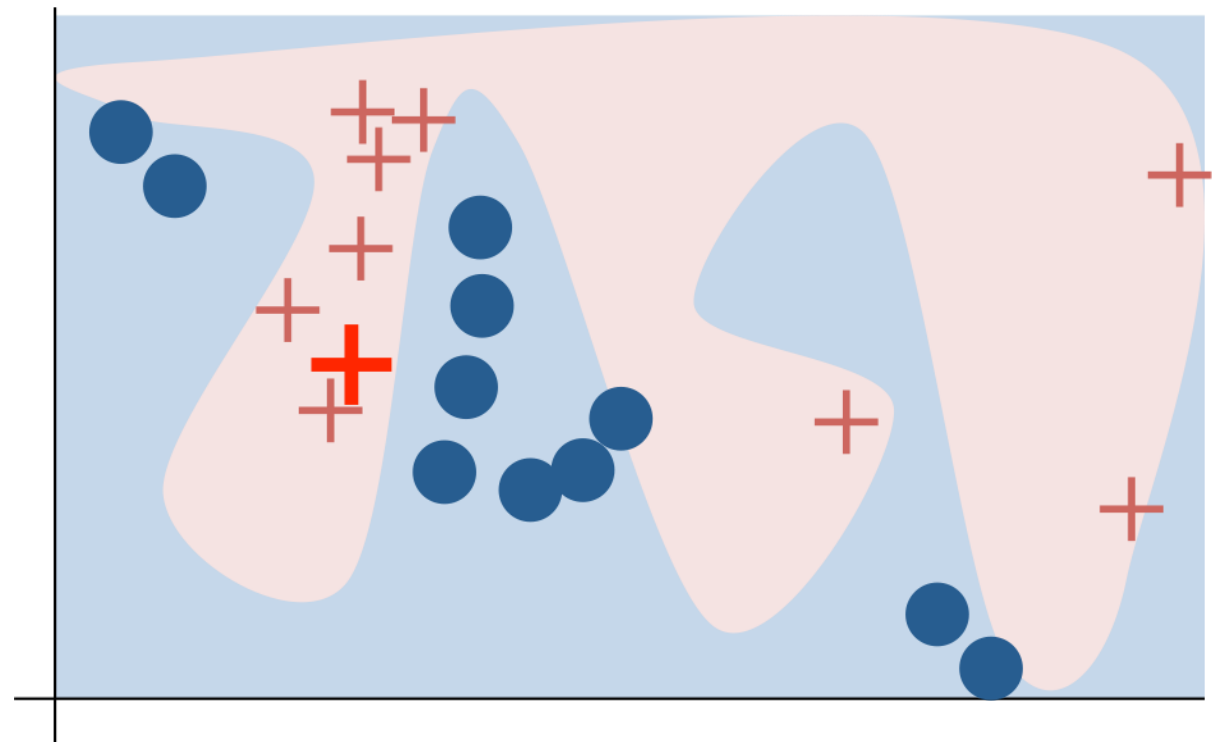


based on a slide by Marco Tulio Ribeiro, KDD 2016

# Fidelity-interpretability trade-off

“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

1. sample points around  $+$
2. use complex model  $f$  to assign class labels

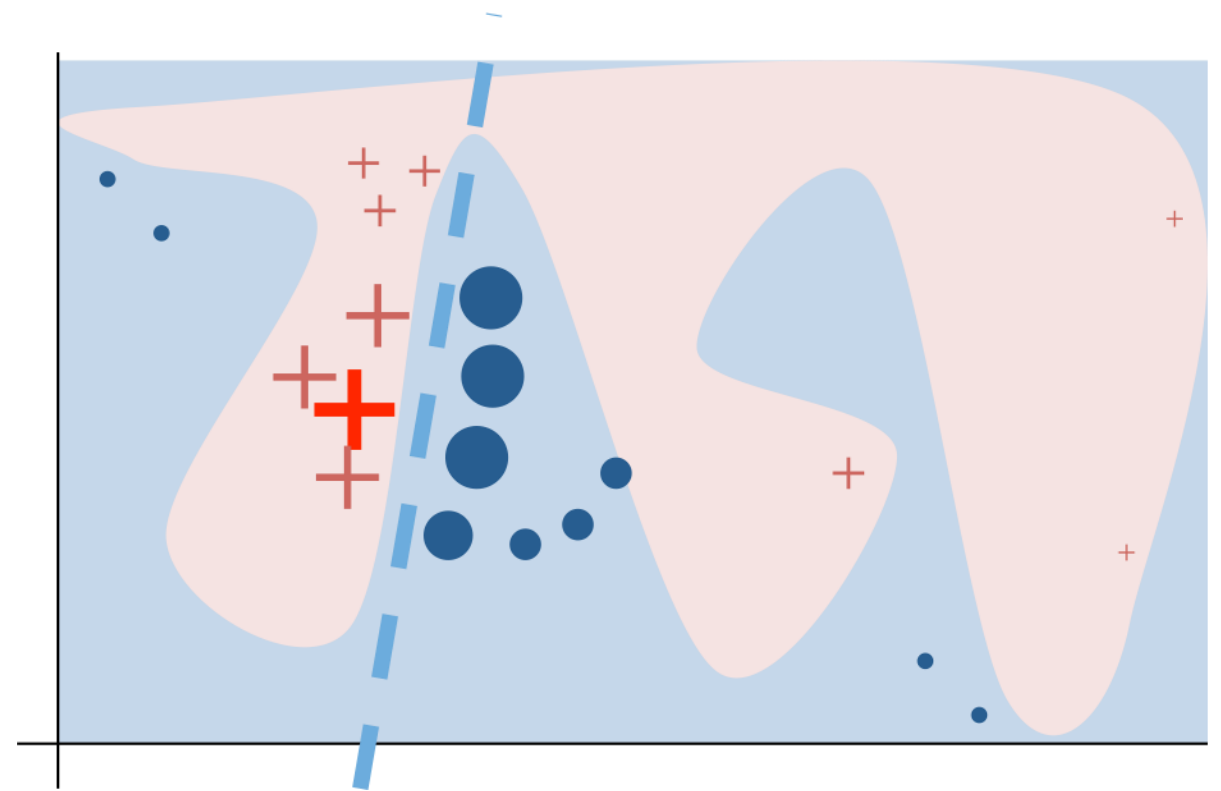


based on a slide by Marco Tulio Ribeiro, KDD 2016

# Fidelity-interpretability trade-off

“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

1. sample points around  $+$
2. use complex model  $f$  to assign class labels
3. weigh samples according to  $\pi_x$
4. learn simple model  $g$  according to samples



based on a slide by Marco Tulio Ribeiro, KDD 2016

# Example: text classification with SVMs

Example #3 of 6 True Class: ● Atheism Instructions Previous Next

### Algorithm 1

**Words that A1 considers important:**

GOD	
mean	
anyone	
this	
Koresh	
through	

**Predicted:** ● Atheism  
**Prediction correct:** ✓

---

**Document**

From: pauld@verdix.com (Paul Durbin)  
Subject: Re: DAVID CORESH IS! **GOD!**  
Nntp-Posting-Host: sarge.hq.verdix.com  
Organization: Verdix Corp  
Lines: 8

### Algorithm 2

**Words that A2 considers important:**

Posting	
Host	
Re	
by	
in	
Nntp	

**Predicted:** ● Atheism  
**Prediction correct:** ✓

---

**Document**

From: pauld@verdix.com (Paul Durbin)  
Subject: **Re:** DAVID CORESH IS! GOD!  
**Nntp-Posting-Host:** sarge.hq.verdix.com  
Organization: Verdix Corp  
Lines: 8

**94% accuracy, yet we shouldn't trust this classifier!**

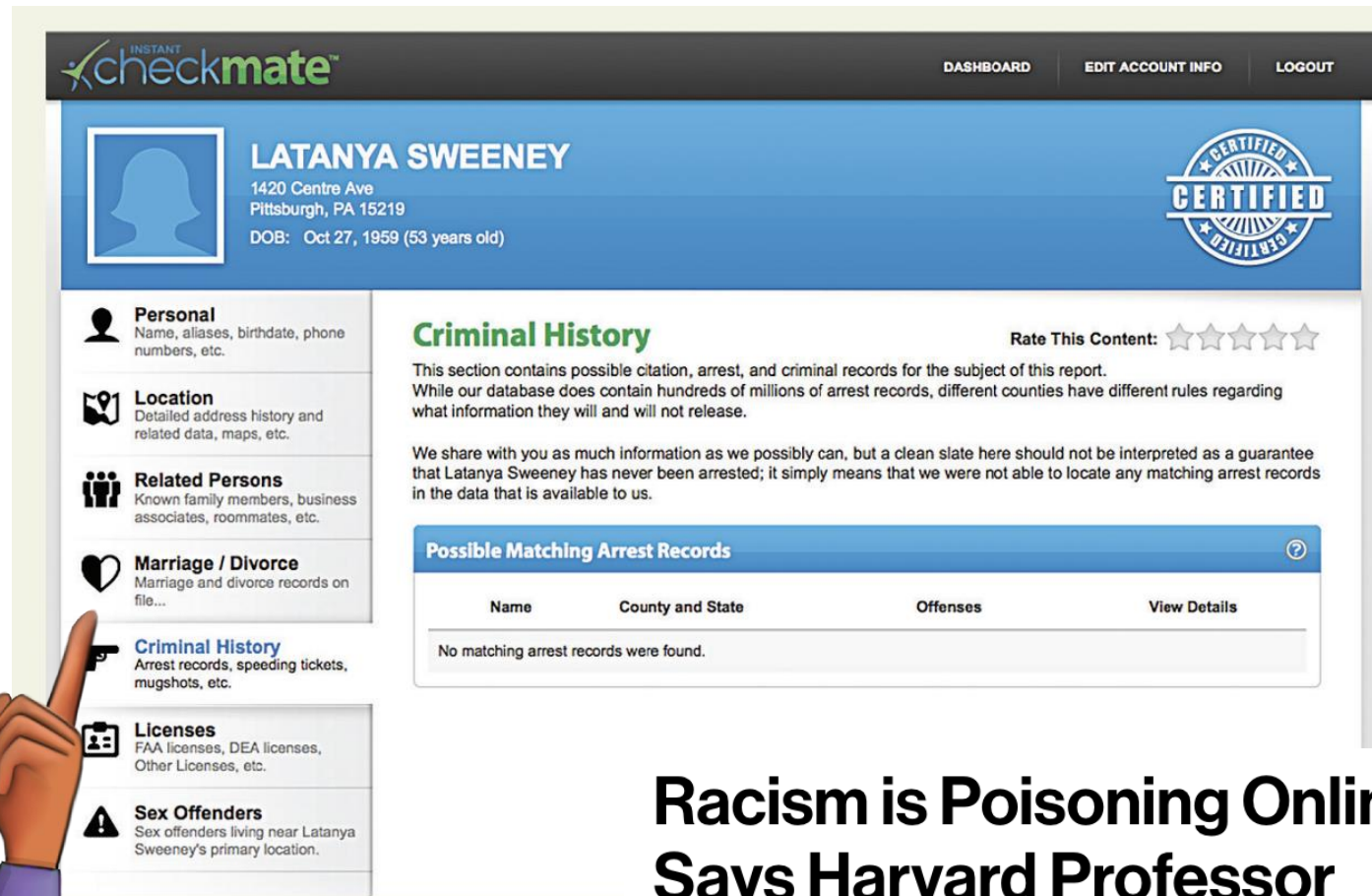


*Instant CheckMate*

# Instant Checkmate

February 2013

Google  
AdSense



**INSTANT checkmate™** DASHBOARD EDIT ACCOUNT INFO LOGOUT

**LATANYA SWEENEY**  
1420 Centre Ave  
Pittsburgh, PA 15219  
DOB: Oct 27, 1959 (53 years old)

**Personal**  
Name, aliases, birthdate, phone numbers, etc.

**Location**  
Detailed address history and related data, maps, etc.

**Related Persons**  
Known family members, business associates, roommates, etc.

**Marriage / Divorce**  
Marriage and divorce records on file...

**Criminal History**  
Arrest records, speeding tickets, mugshots, etc.

**Licenses**  
FAA licenses, DEA licenses, Other Licenses, etc.

**Sex Offenders**  
Sex offenders living near Latanya Sweeney's primary location.

**Criminal History** Rate This Content: ★★★★★  
This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.  
We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Sweeney has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

**Possible Matching Arrest Records**

Name	County and State	Offenses	View Details
No matching arrest records were found.			

FALAH ANIF KHAN

## 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

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

[Sweeney, *Comm ACM* 2013]

r/ai

# Latanya Sweeney's experiment

February 2013

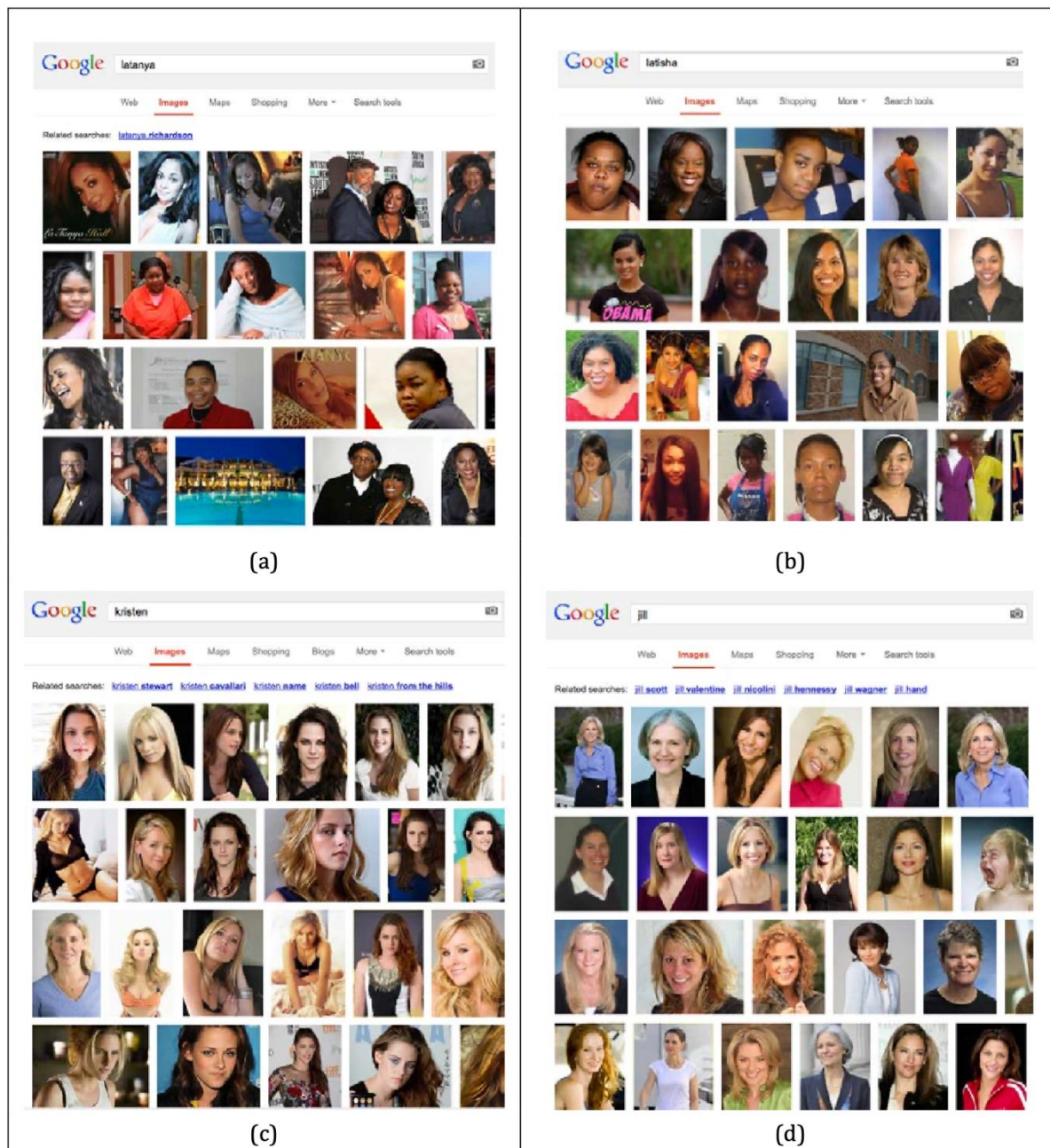


Figure 4. Sample face images on google.com retrieved for searches “latanya” (a), “latisha” (b), “kristen” (c), and “jill” (d).

## Findings

Ads suggestive of criminal record, linking to Instant Checkmate, appear on [google.com](http://google.com) and [reuters.com](http://reuters.com) in response to searches “Latanya Sweeney”, “Latanya Farrell”, and “Latanya Locket”\*

No Instant Checkmate ads when searching for “Kristen Haring”, “Kristen Sparrow”\*, and “Kristen Lindquist”\*

\* Name associated with an actual arrest record

# Possible explanations

## Conjectures

Does Instant Checkmate serve ads specifically for Black-identifying names?

Is Google AdSense explicitly biased in this way?

Does Google AdSense learn racial bias from click-through rates?

February 2013



## Response

**Google:**“AdWords does not conduct any racial profiling. ...**It is up to individual advertisers to decide which keywords they want to choose** to trigger their ads.”

“**Instant Checkmate would like to state unequivocally that it has never engaged in racial profiling in Google AdWords.** We have absolutely no technology in place to even connect a name with a race and have never made any attempt to do so.”

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

*Ad Fisher*

# Online job ads targeting

theguardian

July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

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

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

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 (i.e. view, click, purchase)


**Publishers** (or **content providers**) host online content that often includes ads. They outsource ad placement to third-party **ad networks** (e.g. Google Ads)



**Advertisers** seek to place their ads on **publisher's** website (usually by bidding in ad auctions)

**Ad networks** track users across sites, collecting data. They connect **advertisers** and **publishers**.

# Google ad settings (ca. 2015)

**Google ad settings** aims to provide **transparency** / give **control to users** over the ads that they see




Your Google profile 

 Gender  Age 35-44

Ads based on your interests  ON

Improve your ad experience when you are signed in to Google sites

With Ads based on your interests ON	With Ads based on your interests OFF
<ul style="list-style-type: none"><li>The ads you see will be delivered based on your prior search queries, the videos you've watched on YouTube, as well as other information associated with your account, such as your age range or gender</li><li>On some Google sites like YouTube, you will see ads related to your interests, which you can edit at any time by visiting this page</li><li>You can block some ads that you don't want to see</li></ul>	<ul style="list-style-type: none"><li>You will still see ads and they may be based on your general location (such as city or state)</li><li>Ads will not be based on data Google has associated with your Google Account, and so may be less relevant</li><li>You will no longer be able to edit your interests</li><li>All the advertising interests associated with your Google Account will be deleted</li></ul>

Google Julia   

### Control your Google ads

You can control the ads that are delivered to you based on your Google Account, across devices, by editing these settings. These ads are more likely to be useful and relevant to you.

Your interests

- Action & Adventure Films
- Cooking & Recipes
- History
- Hygiene & Toiletries
- Mobile Phones
- Phone Service Providers
- Reggaeton
- Vehicle Brands
- Cats
- Fitness
- Hybrid & Alternative Vehicles
- Make-Up & Cosmetics
- Parenting
- Recording Industry
- Search Engine Optimization & Marketing

[+ ADD NEW INTEREST](#) [WHERE DID THESE COME FROM?](#)

These interests are derived from your activity on Google sites, such as the videos you've watched on YouTube. This does not include Gmail interests, which are used only for ads within Gmail. [Learn more](#)

Ad Fisher's question: Do users truly have control over the ads they see? Or is this a placebo button?



# AdFisher: Overview

## Experiment

Question: How do user behaviors, ads, and settings interact?

Approach: Automated randomized controlled experiments for studying online tracking

Desideratum: **Individual data use**

**transparency**: Ad network must disclose which user information is used when determining which ads to serve

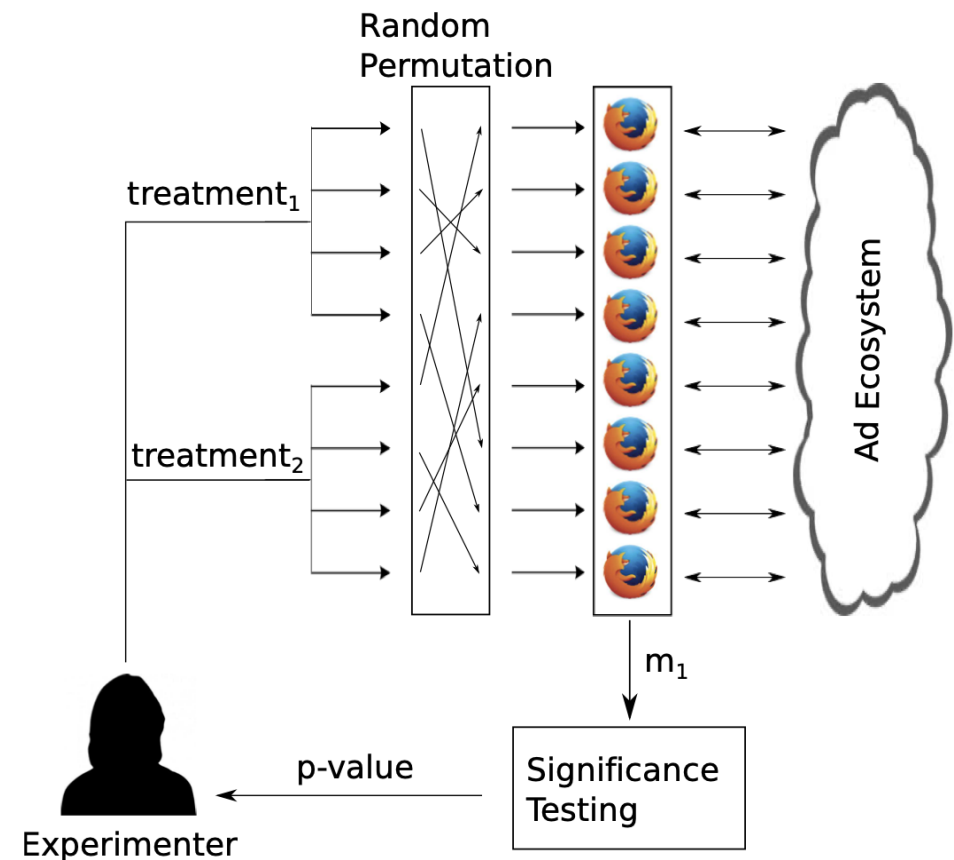


Figure 2: Experimental setup to carry out significance testing on eight browser agents comparing the effects of two treatments. Each agent is randomly assigned a treatment which specifies what actions to perform on the web. After these actions are complete, they collect measurements which are used for significance testing.

# AdFisher: Methodology

Browser-based experiments with simulated users:

**input:** (1) visits to content providing websites  
(2) interactions with Google Ad Settings

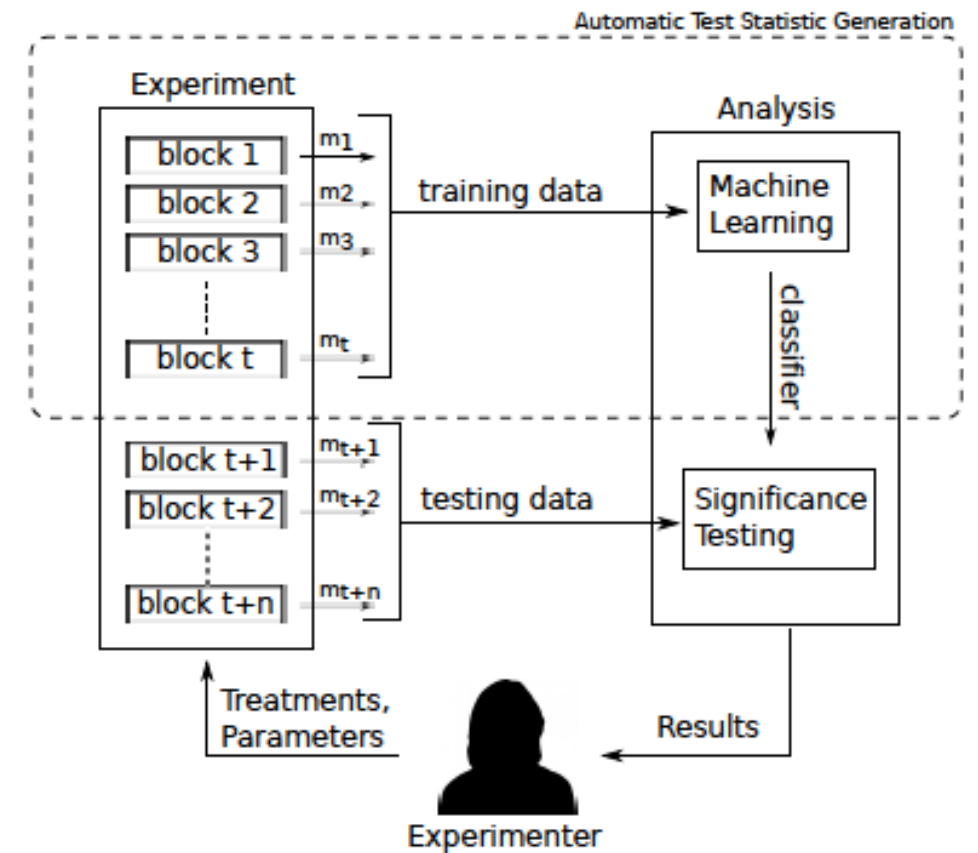
**output:** (1) ads shown to users by Google  
(2) change in Google Ad Settings

Use Fisher randomized hypothesis testing:

**null hypothesis:** inputs do not affect outputs

control and experimental treatments

AdFisher can help select a test statistic



# AdFisher: transparency tests

- **Transparency:** User can view data about them used for ad selection
- **Causal test:** Find attribute that changes ads but not settings
- Experiment 1: **Substance abuse**
  - Simulate interest in substance abuse in the experimental group but not in the control group, check for differences in Ad Settings, collect ads from Times of India
  - Result: No difference in Ad Settings between the groups, yet significant differences in ads served: rehab vs. stocks & driving jobs

**violation**

# AdFisher: discrimination tests

- **Non-Discrimination:** Users differing only in protected attributes are treated similarly
- **Causal test:** Does a protected attribute change ads?
- Experiment 2: **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: In one experiment, males were shown ads for higher-paying jobs far more often than females

**violation**

# AdFisher: ad choice tests

- **Ad choice:** Removing an interest decreases the number of ads related to that interest
- **Causal test:** Does removing an interest cause a decrease in related ads?
- Experiment 3: **Online dating**
  - Simulate interest in online dating in both groups, remove “Dating & Personals” from the interests on Ad Settings for experimental group, collect ads
  - Result: members of experimental group do not get ads related to dating, while members of the control group do

**compliance**

# Follow-up: How is targeting done?

- On gender directly
- On a proxy of gender (i.e., on a known correlate of gender *because* it is a correlate)
- On a known correlate of gender, but not because it is a correlate
- On an unknown correlate of gender

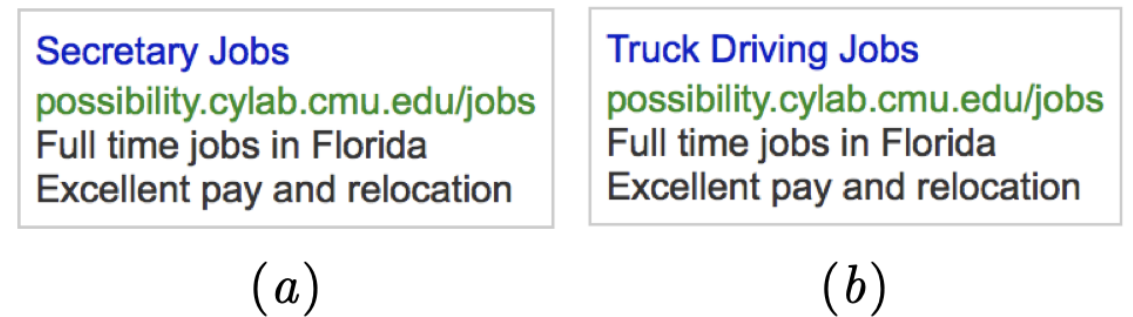


Figure 1: Ads approved by Google in 2015. The ad in the left (right) column was targeted to women (men).

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

# AdFisher: Who is responsible?

## Finding

### Secretary Jobs

possibility.cylab.cmu.edu/jobs  
Full time jobs in Florida  
Excellent pay and relocation

(a)

### Truck Driving Jobs

possibility.cylab.cmu.edu/jobs  
Full time jobs in Florida  
Excellent pay and relocation

(b)

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

## Conjectures

Is **Google explicitly programming** the system to show the ad less often to women?

Is the **advertiser** targeting the ad through **explicit use of demographic categories or selection of proxies**, and Google respecting these targeting criteria?

Are **other advertisers outbidding** our advertiser when targeting to women?

Are **male and female users behaving differently** in response to ads?

# Conclusions from AdFisher

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



discrimination through  
optimization

# Discrimination through optimization

- Follow-up work on AdFisher (Google ads, gender-based discrimination for the purposes of employment) ascertained that it was possible to target on gender for job ads
- Platforms have since taken steps to address such blatant violations

*“... Facebook currently has several policies in place to avoid discrimination for certain types of ads. Facebook also recently **built tools to automatically detect ads offering housing, employment, and credit**, and pledged to prevent the use of certain targeting categories with those ads. Additionally, Facebook relies on advertisers to self-certify that they are not in violation of Facebook’s advertising policy prohibitions against discriminatory practices. More recently, in order to settle multiple lawsuits stemming from these reports, **Facebook stated that they will soon no longer allow age, gender, or ZIP code-based targeting for housing, employment or credit ads**, and that they would also block other detailed targeting attributes that are “describing or appearing to relate to protected classes”.*

- Yet, the question remains: **Does the ad delivery platform itself embed discriminatory outcomes?**

# Discrimination through optimization

**Key question:** does **the platform itself** introduce demographic skew in ad delivery?

## Conjectures

Users see relevant ads, maximizing the likelihood of engagement. **Based on historical engagement data, delivery may be skewed** in ways that an advertiser may not have intended.

**Market effects and financial optimization can lead to skewed ad delivery.** In a nutshell: some populations are more “valuable” and so advertising to them costs more. If an advertiser bids less, they won’t get to the more “valuable” population.

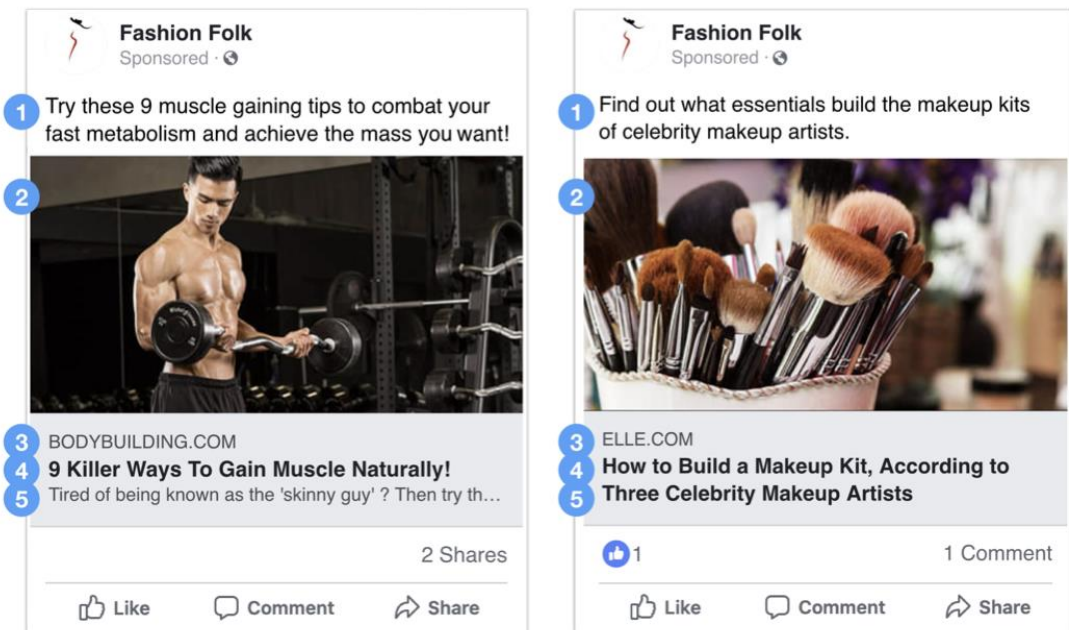
# Discrimination through optimization

## Part 1: ad creation

- ad contents
- audience selection
- bidding strategy

## Part 2: ad delivery

For every opportunity to show a user an ad (e.g., **an ad slot** is available as the user is browsing the service), the ad platform will run an **ad auction** to determine, from among all of the ads that include the current user in the audience, which ad should be shown.



**Figure 1: Each ad has five elements that the advertiser can control: (1) the ad text, entered manually by the advertiser, (2) the images and/or videos, (3) the domain, pulled automatically from the HTML meta property `og:site_name` of the destination URL, (4) the title, pulled automatically from the HTML meta property `og:title` of the destination URL, and (5) the description from meta property `og:description` of the destination URL. The title and description can be manually customized by the advertiser if they wish.**

# Discrimination through optimization

## Part 1: ad creation

- ad contents
- audience selection
- bidding strategy

## Part 2: ad delivery

For every opportunity to show a user an ad (e.g., **an ad slot** is available as the user is browsing the service), the ad platform will run an **ad auction** to determine, from among all of the ads that include the current user in the audience, which ad should be shown.

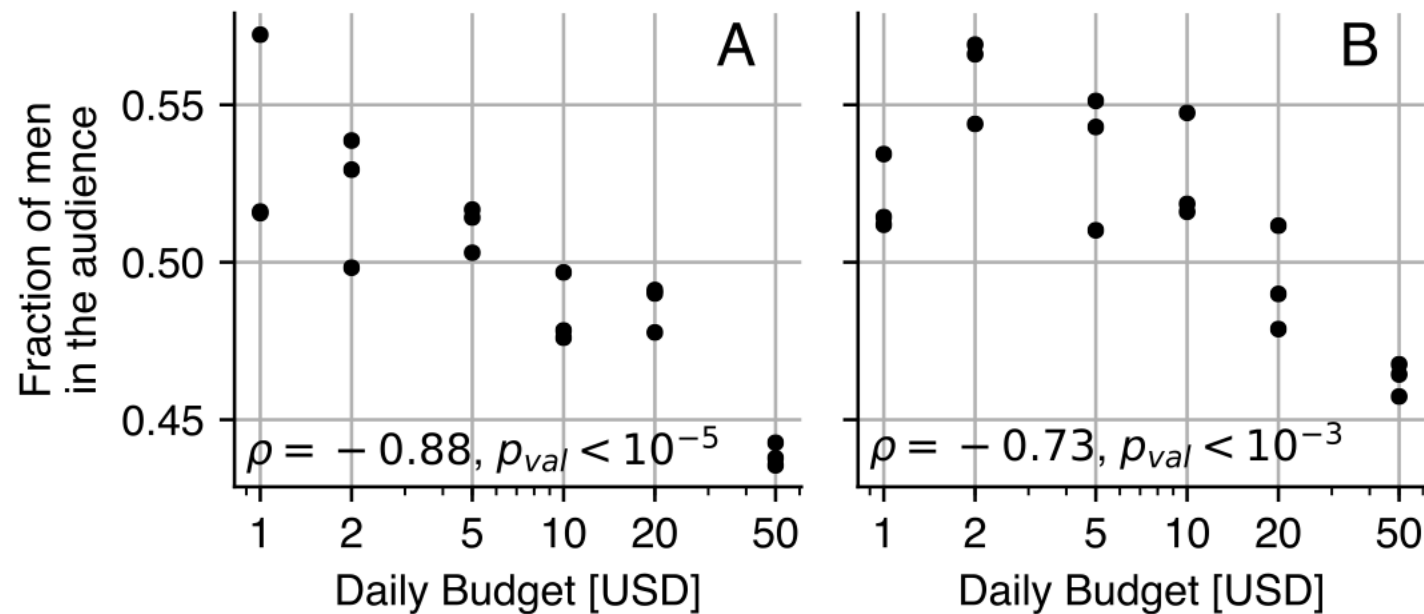
When Facebook has ad slots available, it runs an ad auction among the active advertisements bidding for that user. However, **the auction does not just use the bids placed by the advertisers**; Facebook says:

*“The ad that wins an auction and gets shown is the one with the highest **total value**. Total value isn’t how much an advertiser is willing to pay us to show their ad. It’s combination of 3 major factors: (1) Bid, (2) Estimated action rates, and (3) Ad quality and relevance.”*

*“During ad set creation, you chose a target audience ... and an optimization event ... **We show your ad to people in that target audience who are likely to get you that optimization event.**”*

Quiz + Break

# Discrimination through optimization

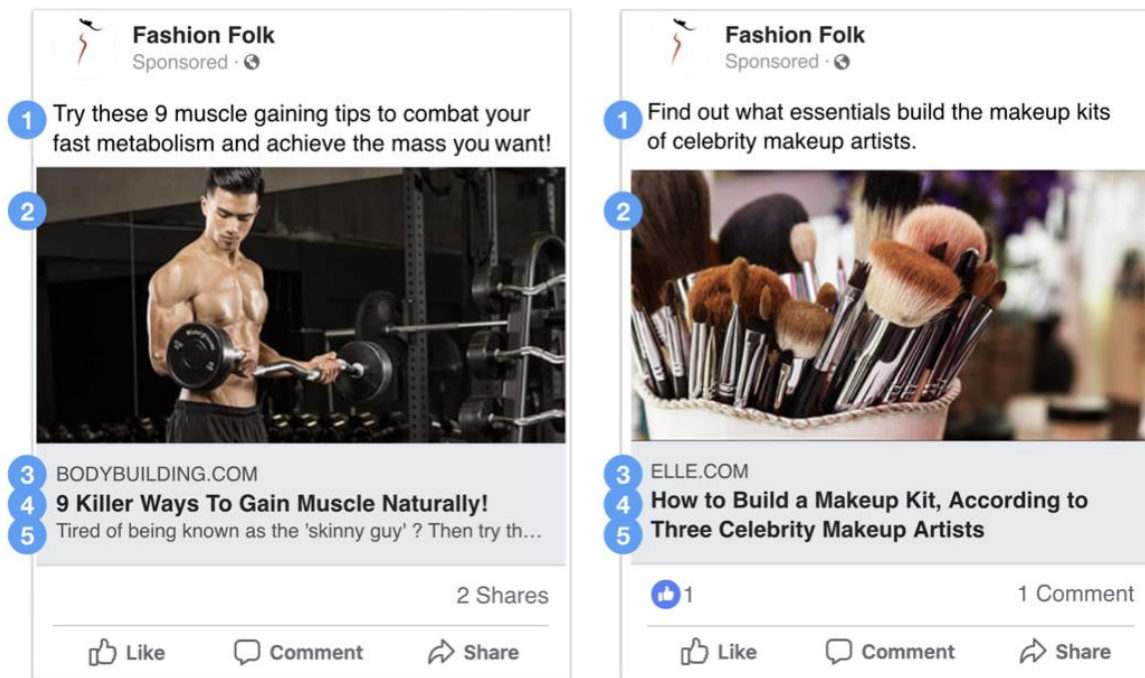


**Figure 2: Gender distributions of the audience depend on the daily budget of an ad, with higher budgets leading to a higher fraction of women. The left graph shows an experiment where we target all users located in the U.S.; the right graph shows an experiment where we target our random phone number custom audiences.**

“In both cases, **we observe that changes in ad delivery due to differences in budget are indeed happening**: the higher the daily budget, the smaller the fraction of men in the audience.”

“The stronger effect we see when targeting all U.S. users may be due to the additional freedom that the ad delivery system has when choosing who to deliver to, as this is a significantly larger audience.”

# Discrimination through optimization



**Figure 1: Each ad has five elements that the advertiser can control: (1) the ad text, entered manually by the advertiser, (2) the images and/or videos, (3) the domain, pulled automatically from the HTML meta property `og:site_name` of the destination URL, (4) the title, pulled automatically from the HTML meta property `og:title` of the destination URL, and (5) the description from meta property `og:description` of the destination URL. The title and description can be manually customized by the advertiser if they wish.**

**Same bidding strategy** for bodybuilding and cosmetics, without explicitly mentioning gender

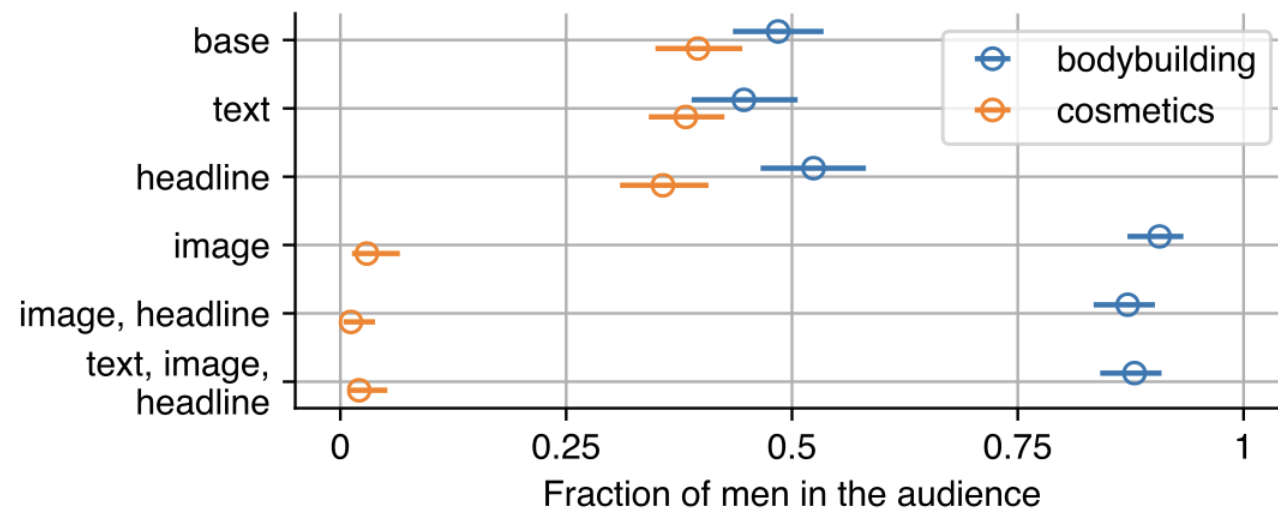
**Strong gender skew in delivery:** bodybuilding delivered to over 75% men on average, cosmetics delivered to over 90% women on average



# Discrimination through optimization

Which component of the ad creative impacts delivery most?











“It seems that the image, both alone and in conjunction with the title, was the most influential factor towards skewing Facebook’s ad delivery.”

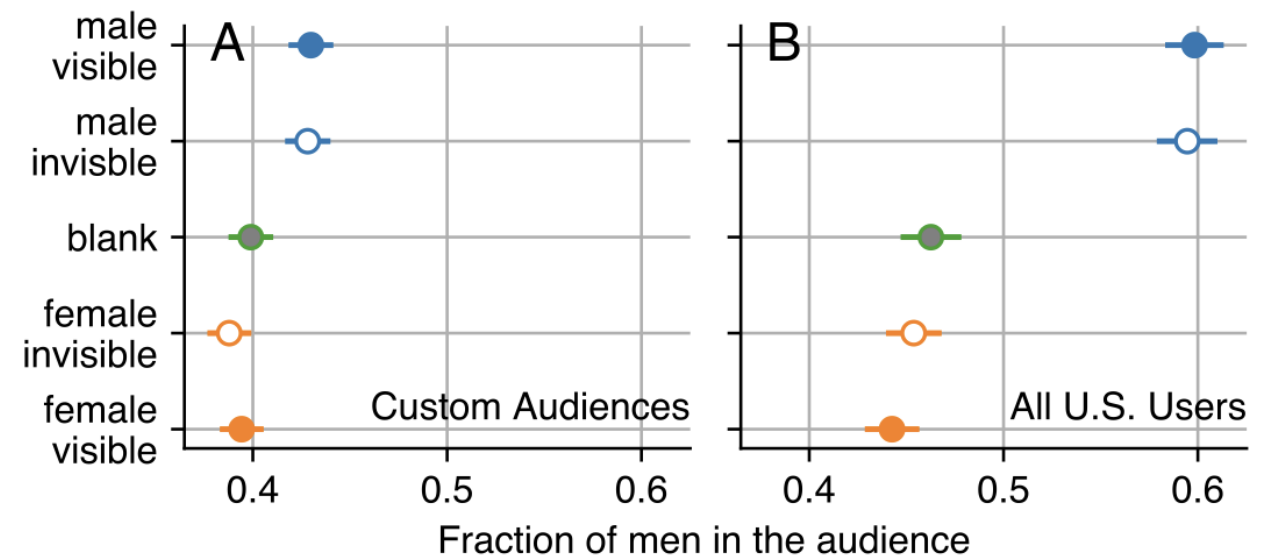


**Figure 3: “Base” ad contains a link to a page about either bodybuilding or cosmetics, a blank image, no text, or headline. There is a small difference in the fraction of male users for the base ads, and adding the “text” only decreases it. Setting the “headline” sets the two ads apart but the audience of each is still not significantly different than that of the base version. Finally, setting the ad “image” causes drastic changes: the bodybuilding ad is shown to a 91% male audience, the cosmetics ad is shown to a 5% male audience, despite the same target audience.**

# Discrimination through optimization

Transparent images are still targeted

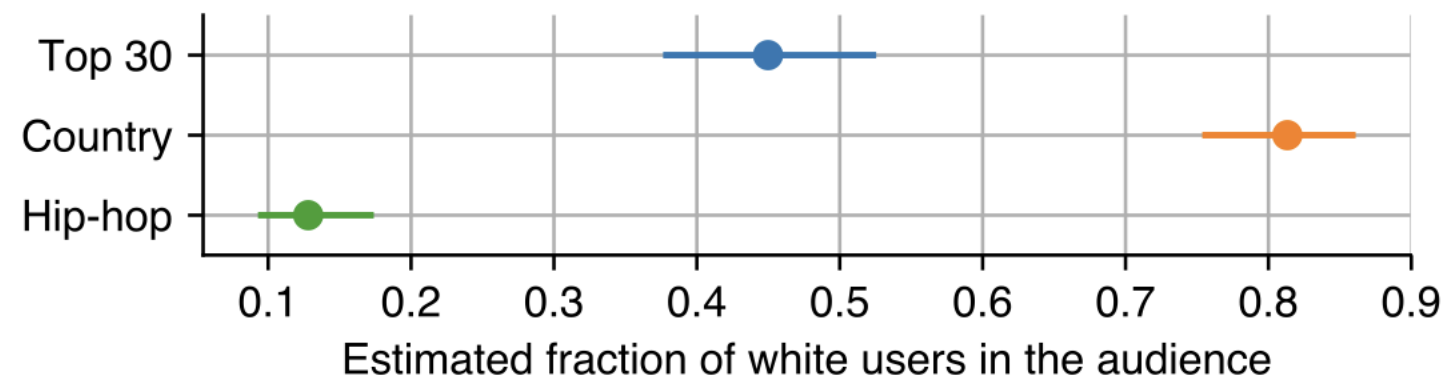
No.	Masculine		Feminine	
	Visible	Invisible	Visible	Invisible
1				
2				
3				
4				
5				



We can observe that ad delivery is, in fact, skewed, with the **ads with stereotypically masculine images delivering to over 43% men** and the **ads with feminine images delivering to 39% men** in the experiment targeting custom audiences as well as 58% and 44% respectively in the experiment targeting all U.S. users. Interestingly, we also observe that the **masculine invisible ads appear to be indistinguishable in the gender breakdown of their delivery from the masculine visible ads**, and the feminine invisible ads appear to be indistinguishable in their delivery from the feminine visible ads.

This strongly suggests that Facebook uses **an automated image classification mechanism** to steer different ads towards different subsets of the user population

# Discrimination through optimization

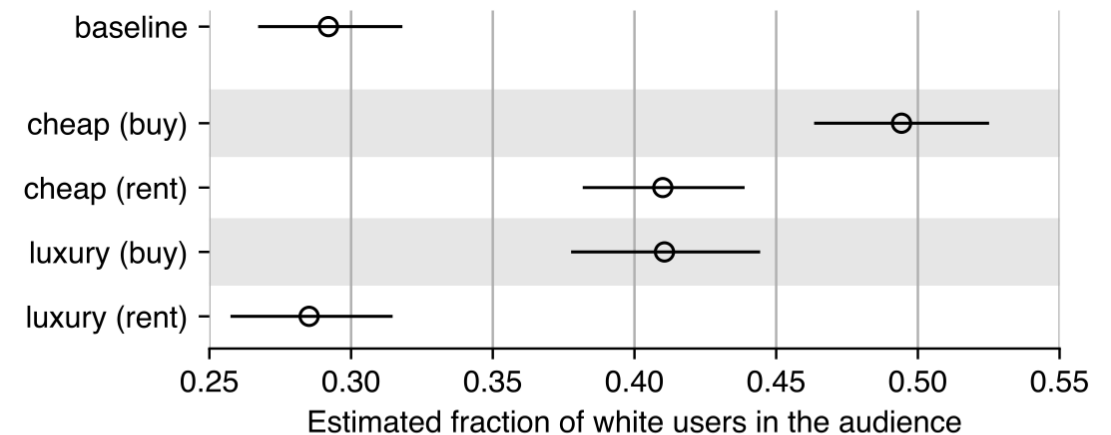


**Figure 7: We run three campaigns about the best selling albums. *Top 30* is neutral, targeting all. *Country* implicitly targets white users, and *Hip-hop* implicitly targets Black users. Facebook classification picks up on the implicit targeting and shows it to the audience we would expect.**

“We hold targeting parameters fixed, run ads that are stereotypically of interest to different races. We find that **Facebook ad delivery follows the stereotypical distribution**, despite all ads being targeted in the same manner and using the same bidding strategy.”

# Discrimination through optimization

**Key question:** does **the platform itself** introduce demographic skew in ad delivery?



**Figure 9: Results for housing ads, showing a breakdown in the ad delivery audience by race. Despite being targeted in the same manner, using the same bidding strategy, and being run at the same time, we observe significant skew in the makeup of the audience to whom the ad is delivered (ranging from estimated 27% white users for luxury rental ads to 49% for cheap house purchase ads).**

## Findings

Skew was observed along racial lines, in ads for housing opportunities

# Discrimination through optimization

**Key question:** does **the platform itself** introduce demographic skew in ad delivery?

## Findings

Skew can arise due to financial optimization effects and the ad delivery platform's predictions about the relevance of its ads to different user categories

Ad content - text and images - and advertiser budget both may contribute to the skew.

*legal ramifications*

# Facebook ads and HEC

When is  
“skew” in fact  
discrimination?

## SUMMARY OF SETTLEMENTS BETWEEN CIVIL RIGHTS ADVOCATES

### AND FACEBOOK

March 19, 2019

### **Housing, Employment, and Credit Advertising Reforms**

In the settlements, Facebook will undertake far-reaching changes and steps that will prevent discrimination in housing, employment, and credit advertising on Facebook, Instagram, and Messenger. These changes demonstrate real progress.

- **Facebook will establish a separate advertising portal for creating housing, employment, and credit (“HEC”) ads on Facebook, Instagram, and Messenger that will have limited targeting options, to prevent discrimination.**
- **The following rules will apply to creating HEC ads.**
  - *Gender, age, and multicultural affinity targeting options will not be available when creating Facebook ads.*
  - *HEC ads must have a minimum geographic radius of 15 miles from a specific address or from the center of a city. Targeting by zip code will not be permitted.*

# Facebook ads and HEC

## SUMMARY OF SETTLEMENTS BETWEEN CIVIL RIGHTS ADVOCATES

### AND FACEBOOK

March 19, 2019

### Housing, Employment, and Credit Advertising Reforms

- *HEC ads will not have targeting options that describe or appear to be related to personal characteristics or classes protected under anti-discrimination laws. This means that targeting options that may relate to race, color, national origin, ethnicity, gender, age, religion, family status, disability, and sexual orientation, among other protected characteristics or classes, will not be permitted on the HEC portal.*
- *Facebook's "Lookalike Audience" tool, which helps advertisers identify Facebook users who are similar to advertisers' current customers or marketing lists, will no longer consider gender, age, religious views, zip codes, Facebook Group membership, or other similar categories when creating customized audiences for HEC ads.*



# Facebook ads and HEC

## SUMMARY OF SETTLEMENTS BETWEEN CIVIL RIGHTS ADVOCATES

### AND FACEBOOK

March 19, 2019

#### Housing, Employment, and Credit Advertising Reforms

- *Advertisers will be asked to create their HEC ads in the HEC portal, and if Facebook detects that an advertiser has tried to create an HEC ad outside of the HEC portal, Facebook will block and re-route the advertiser to the HEC portal with limited options.*

# Legal implications, not just for Google

POLICY US & WORLD TECH

## HUD reportedly also investigating Google and Twitter in housing discrimination probe

By [Adi Robertson](#) | [@thedextriarchy](#) | Mar 28, 2019, 3:52pm EDT

March 2019

POLICY US & WORLD TECH

## Facebook has been charged with housing discrimination by the US government

*'Facebook is discriminating against people based upon who they are and where they live,' says HUD secretary*

By [Russell Brandom](#) | Mar 28, 2019, 7:51am EDT

**Fair Housing Act**, also called **Title VIII of the Civil Rights Act of 1968**, U.S. federal legislation that protects individuals and families from [discrimination](#) in the sale, rental, financing, or **advertising** of housing. The Fair Housing Act, as [amended](#) in 1988, prohibits discrimination on the basis of **race**, **color**, **religion**, **sex**, **disability**, **family status**, and **national origin**.

This is the first federal discrimination lawsuit to deal with **racial bias in targeted advertising**, a milestone that lawyers at HUD said was overdue. “Even as we confront new technologies, the fair housing laws enacted over half a century ago remain clear—discrimination in housing-related advertising is against the law,” said HUD General Counsel Paul Compton. “**Just because a process to deliver advertising is opaque and complex doesn’t mean that it’s exempts Facebook and others from our scrutiny and the law of the land.**”

# Facebook ads and the Fair Housing Act

**THE VERGE**

## Facebook has been charged with housing discrimination by the US government

*'Facebook is discriminating against people based upon who they are and where they live,' says HUD secretary*

By [Russell Brandom](#) | Mar 28, 2019, 7:51am EDT

**March 2019**

The Department of Housing and Urban Development has [filed charges](#) against Facebook for housing discrimination, escalating the company's ongoing fight over its ad targeting system. The charges build on [a complaint filed in August](#), finding reasonable cause to believe Facebook has served ads that violate the Fair Housing Act.

*ProPublica* first raised concerns over housing discrimination on Facebook in 2016, when reporters found that [the "ethnic affinities" tool](#) could be used to exclude black or Hispanic users from seeing specific ads. If those ads were for housing or employment opportunities, the targeting could easily violate federal law. At the time, Facebook had no internal safeguards in place to prevent such targeting.

**Fair Housing Act**, also called **Title VIII of the Civil Rights Act of 1968**, U.S. federal legislation that protects individuals and families from [discrimination](#) in the sale, rental, financing, or **advertising** of housing. The Fair Housing Act, as [amended](#) in 1988, prohibits discrimination on the basis of **race**, **color**, **religion**, **sex**, **disability**, **family status**, and **national origin**.

# The Fair Housing Act

**THE VERGE**

## Facebook has been charged with housing discrimination by the US government

*'Facebook is discriminating against people based upon who they are and where they live,' says HUD secretary*

By [Russell Brandom](#) | Mar 28, 2019, 7:51am EDT

March 2019

Facebook has struggled to effectively address the possibility of discriminatory ad targeting. The company pledged to step up anti-discrimination enforcement in the wake of *ProPublica's* reporting, but [a follow-up report](#) in 2017 found the same problems persisted nearly a year later.

**"WE'RE DISAPPOINTED BY TODAY'S DEVELOPMENTS," FACEBOOK SAYS**

According to the HUD complaint, many of the options for targeting or excluding audiences are shockingly direct, including a map tool that explicitly echoes [redlining practices](#). "[Facebook] has provided [a toggle button that enables advertisers to exclude men or women](#) from seeing an ad, a search-box to exclude people who do not speak a specific language from seeing an ad, and [a map tool to exclude people who live in a specified area from seeing an ad by drawing a red line around that area,](#)" the complaint reads.

# And in recent(ish) news



THE UNITED STATES  
DEPARTMENT of JUSTICE

January 9, 2023

The Justice Department announced today that it has reached a key milestone in its settlement agreement with Meta Platforms Inc. (Meta), formerly known as Facebook Inc., **requiring Meta to change its advertisement delivery system to prevent discriminatory advertising in violation of the Fair Housing Act (FHA)**. As required by the settlement entered on June 27, 2022, resolving a lawsuit filed in the U.S. District Court for the Southern District of New York, Meta has now built a new system to address algorithmic discrimination. Today, the parties informed the court that they have reached agreement on the system's compliance targets. This development ensures that **Meta will be subject to court oversight and regular review of its compliance with the settlement through June 27, 2026**.

“This development marks a pivotal step in the Justice Department’s efforts to hold Meta accountable for unlawful algorithmic bias and discriminatory ad delivery on its platforms,” said Assistant Attorney General Kristen Clarke of the Justice Department’s Civil Rights Division. “The Justice Department will continue to hold Meta accountable by ensuring the Variance Reduction System addresses and eliminates discriminatory delivery of advertisements on its platforms. **Federal monitoring of Meta should send a strong signal to other tech companies that they too will be held accountable for failing to address algorithmic discrimination that runs afoul of our civil rights laws.**”

<https://www.justice.gov/opa/pr/justice-department-and-meta-platforms-inc-reach-key-agreement-they-implement-groundbreaking>



# The socio-legal landscape

## Related concern:

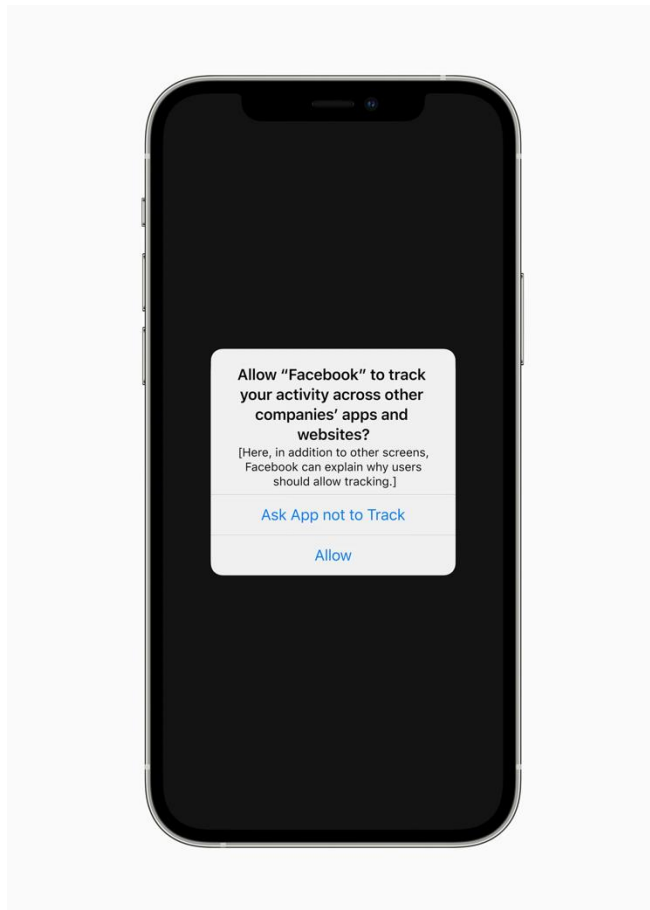
Are ads commercial free speech?

- The First Amendment of the U.S. Constitution protects advertising, but the U.S. Supreme Court set out a test for assessing restrictions on commercial speech, **which begins by determining whether the speech is misleading**
- Are online ads suggesting the existence of an arrest record misleading if no one by that name has an arrest record?
- Assume the ads are free speech: what happens when these ads appear more often for one racial group than another? Not everyone is being equally affected. Is that free speech or racial discrimination?

# Tracking and consent

The New York Times

## *To Be Tracked or Not? Apple Is Now Giving Us the Choice.*



By [Brian X. Chen](#)

April 26, 2021 Updated 12:40 p.m. ET

If we had a choice, would any of us want to be tracked online for the sake of seeing more relevant digital ads?

We are about to find out.

On Monday, [Apple](#) plans to [release iOS 14.5](#), one of its most anticipated software updates for iPhones and iPads in years. It includes a new privacy tool, called App Tracking Transparency, which could give us more control over how our data is shared.

Here's how it works: When an app wants to follow our activities to share information with third parties such as advertisers, a window will show up on our Apple device to ask for our permission to do so. If we say no, the app must stop monitoring and sharing our data.

A pop-up window may sound like a minor design tweak, but it has thrown the online advertising industry into upheaval. Most notably, Facebook has gone on the warpath. Last year, the social network created a website and took out full-page ads in newspapers denouncing Apple's privacy feature as [harmful to small businesses](#).

# Is there still pressure with new U.S. Admin?

The Register

Hewlett Packard  
Enterprise

June 4, 2024

## Meta algorithms push Black people more toward expensive universities, study finds

Just as we saw with housing, Facebook giant's advertising system seems to treat Whites and POC differently

Thomas Claburn

Tue 4 Jun 2024 | 00:00 UTC

**SPECIAL REPORT** Meta's algorithms for presenting educational ads show signs of racial bias, according to researchers from Princeton University and the University of Southern California.

## Auditing for Racial Discrimination in the Delivery of Education Ads

Basileal Imana  
imana@princeton.edu  
Center for Information Technology  
Policy, Princeton University  
Princeton, New Jersey, USA

Aleksandra Korolova  
korolova@princeton.edu  
Department of Computer Science and  
School of Public and International  
Affairs, Princeton University  
Princeton, New Jersey, USA

John Heidemann  
johnh@isi.edu  
Information Sciences Institute,  
University of Southern California  
Los Angeles, California, USA

### ABSTRACT

Digital ads on social-media platforms play an important role in shaping access to economic opportunities. Our work proposes and implements a new third-party auditing method that can audit for racial bias in the delivery of ads for education opportunities. Third-party auditing is important because it allows external researchers to demonstrate presence or absence of bias in social-media advertising. Education is a domain with legal protections against

June 03–06, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 14 pages.  
<https://doi.org/10.1145/3630106.3659041>

Still possible in U.S.?

Probably not, for now...

...Korolova told us: "We'd like Meta to turn off its algorithm in all advertising domains that relate to life opportunities, including important topics (such as education, insurance, healthcare, etc). Identification of these topics could perhaps be done in consultation with the community."

[https://www.theregister.com/2024/06/04/meta\\_ad\\_algorithm\\_discrimination/](https://www.theregister.com/2024/06/04/meta_ad_algorithm_discrimination/)

r/ai



# Is there still pressure with new U.S. Admin?

February 13, 2025



Campaigns v Regions v Publications v About us v Donate

Home > Campaigns > Digital Threats to Democracy > New evidence of Facebook's sexist algorithm

## New evidence of Facebook's sexist algorithm

Published: 12 June 2023

Updated: 13 February 2025

SHARE

Investigation

Digital Threats to Democracy

Big Tech Social media

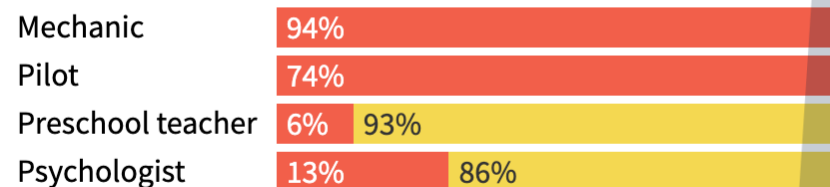
New research reveals how Facebook could be discriminating against users and acting in violation of equality laws and data protection rules in France and the Netherlands.

Imagine that you and your colleague are on a lunch break. You are both scrolling through your Facebook feeds at your desks. You work in the same companies, have similar responsibilities, and experience. Your colleague sees an ad that you'd love to get the chance to apply for, but you haven't. You refresh your feed and scan the ads again, but it still doesn't show up for you. Why didn't you see the ad?

As uncomfortable as it sounds, perhaps one of the reasons is your gender.

### Who did Facebook's algorithm show job ads to in France and the Netherlands?

Male Female Unknown



But! U.S. is not only place these companies operate!

<https://globalwitness.org/en/campaigns/digital-threats/new-evidence-of-facebooks-sexist-algorithm/>

