DS-GA 3001.009: Responsible Data Science

Transparency and Accountability

Prof. Julia Stoyanovich
Center for Data Science
Computer Science and Engineering at Tandon

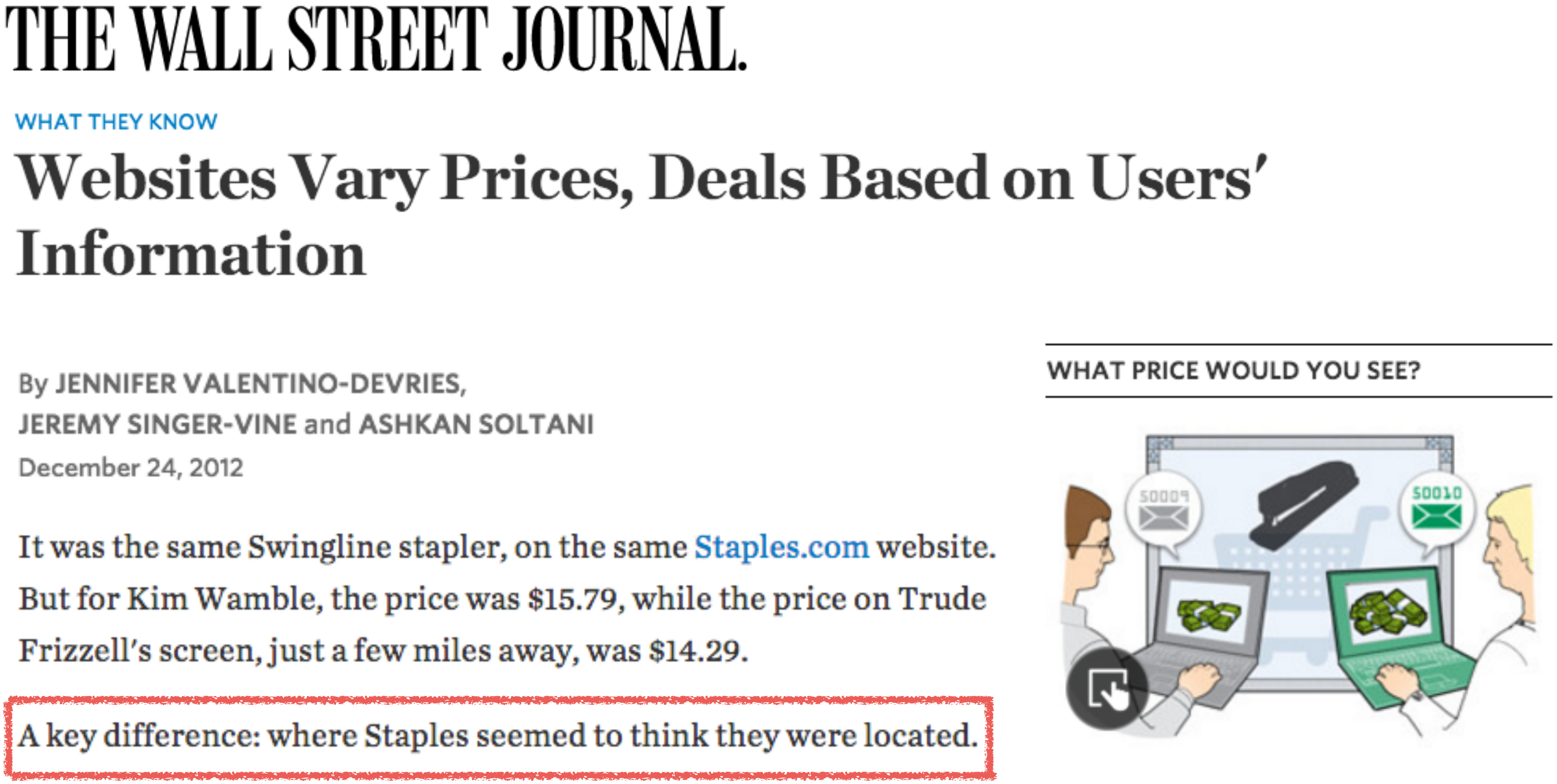
@stoyanoj

http://stoyanovich.org/
https://dataresponsibly.github.io/

Transparency themes

- **Explaining black-box models**
  - LIME: local interpretable explanations [Ribeiro et al., KDD 2016]
  - QII: causal influence of features on outcomes [Datta et al., SSP 2016]
- **Online ad targeting**: identifying the problem
  - Racially identifying names [Sweeney, CACM 2013]
  - Ad Fisher [Datta et al., PETS 2015]
- **Interpretability**
  - Nutritional labels [Yang et al., SIGMOD 2018]

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.

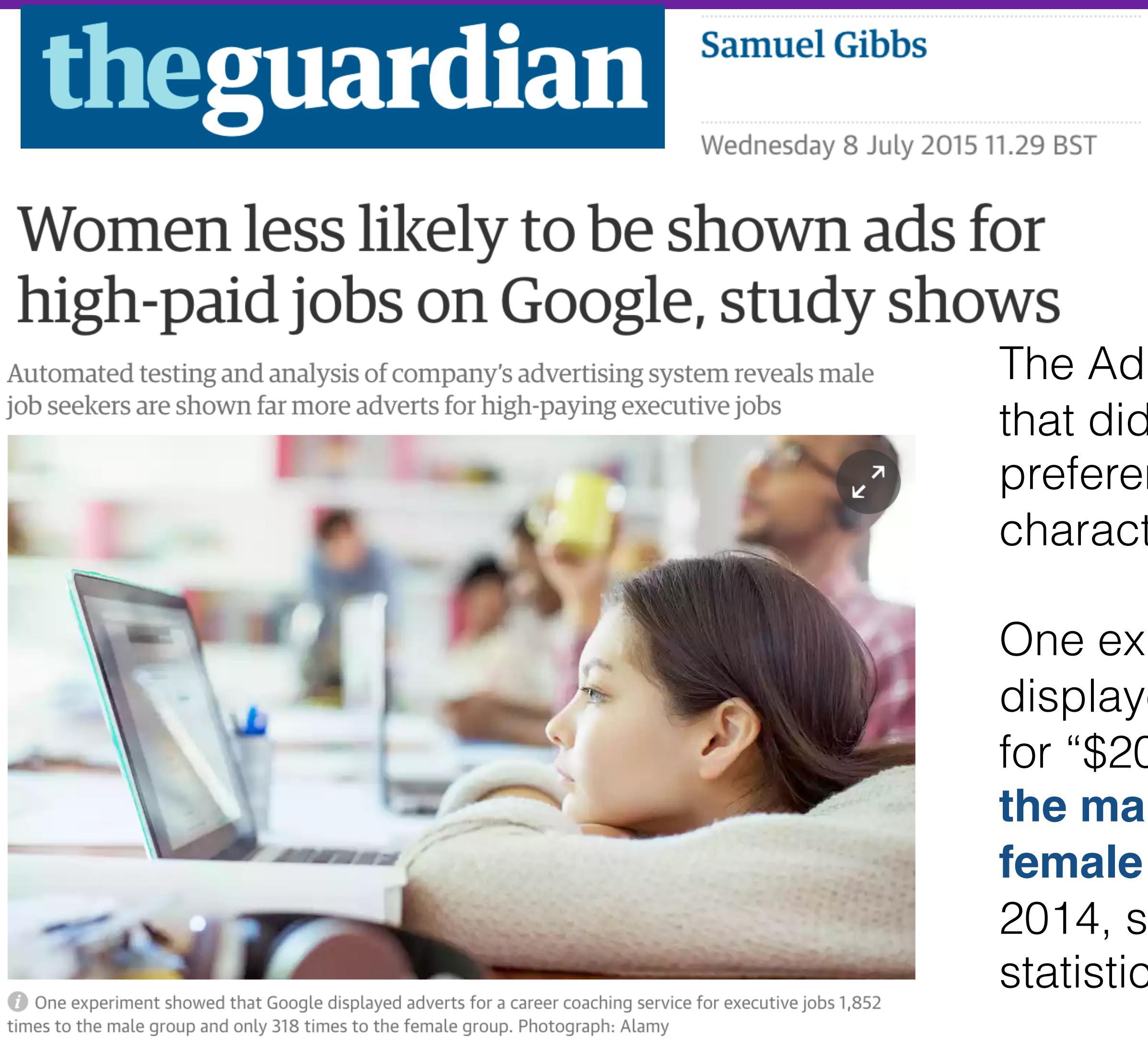
A key difference: where Staples seemed to think they were located.

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

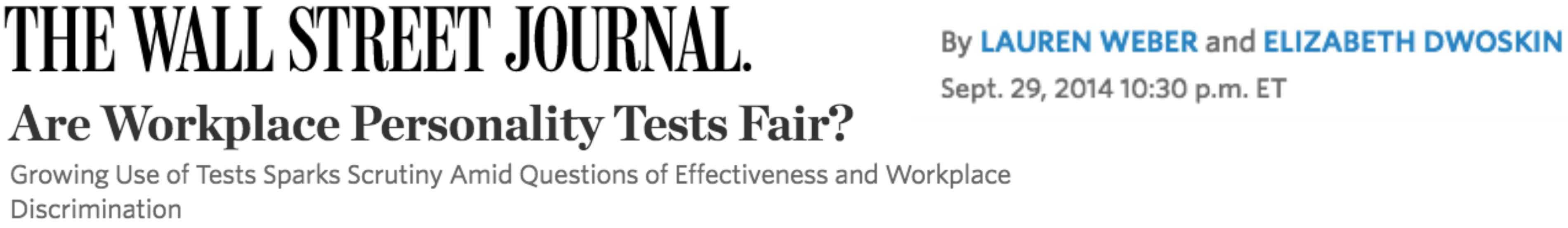
ⓘ One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

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

Job-screening personality tests



THE WALL STREET JOURNAL.

By LAUREN WEBER and ELIZABETH DWOSKIN

Sept. 29, 2014 10:30 p.m. ET

Are Workplace Personality Tests Fair?

Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination

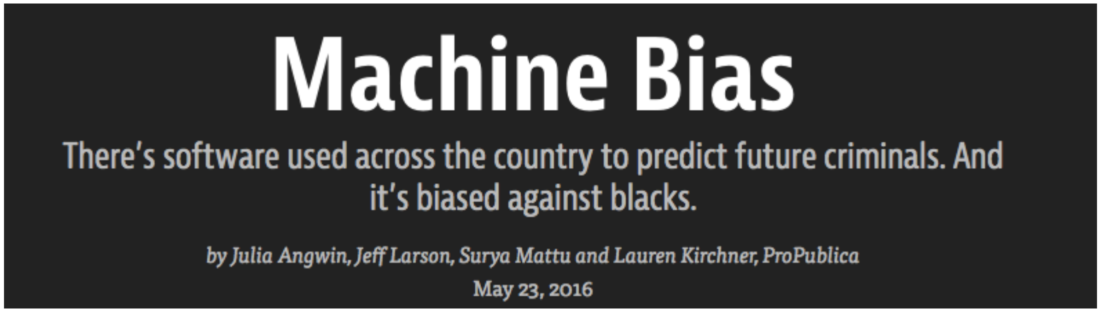
Kyle Behm accused Kroger and six other companies of discrimination against the mentally ill through their use of personality tests. *TROY STAINS FOR THE WALL STREET JOURNAL*

The Equal Employment Opportunity commission is **investigating whether personality tests discriminate against people with disabilities**.

As part of the investigation, officials are trying to determine if the tests **shut out people suffering from mental illnesses** such as depression or bipolar disorder, even if they have the right skills for the job.

http://www.wsj.com/articles/are-workplace-personality-tests-fair-1412044257

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

Explaining black-box classifiers

"Why should I trust you?" Explaining the predictions of any classifier (LIME)

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

Interpretability enables trust

- **If users do not trust a model or a prediction, they will not use it!**
  - predictive models are bound to make mistakes (recall our discussion of fairness in risk assessment)
  - in many domains (e.g., medical diagnosis, terrorism detection, setting global policy, ….) consequences of a mistake may be catastrophic
  - think **agency** and **responsibility**
- The authors of LIME distinguish between two related definitions of trust:
  - trusting **a prediction** sufficiently to take some action based on it
  - trusting **a model** to behave in a reasonable way when it is deployed
- Of course, trusting **data** plays into both of these - garbage in / garbage out (recall our discussion of data profiling)

Is accuracy sufficient for trust?

We wouldn’t be discussing interpretability if accuracy were sufficient, but what are some of the reasons that accuracy may not be enough?

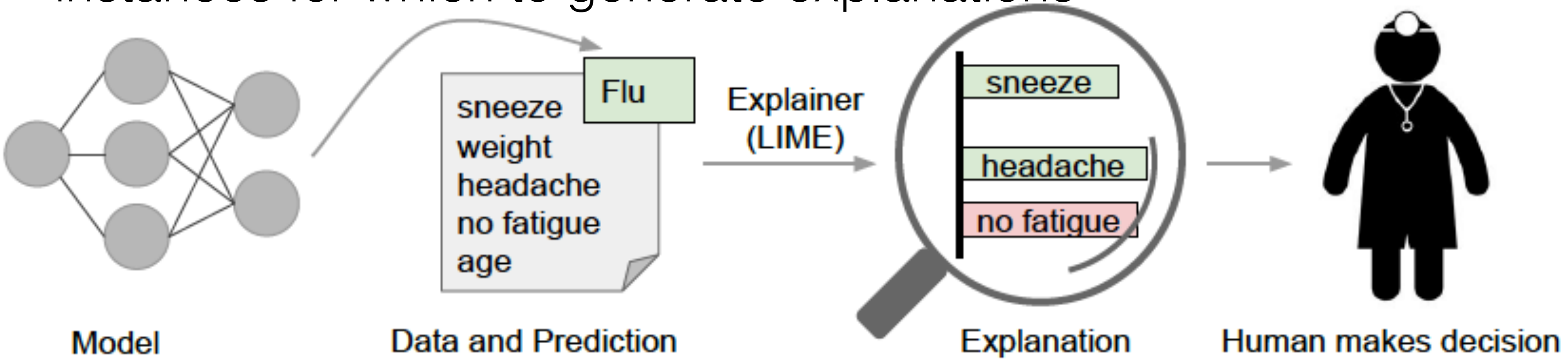
- how is accuracy measures?
- accuracy for whom? over-all, in sub-populations?
- accuracy over which data?
- accuracy / mistakes for what reason?

Explanations based on features

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

- **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

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

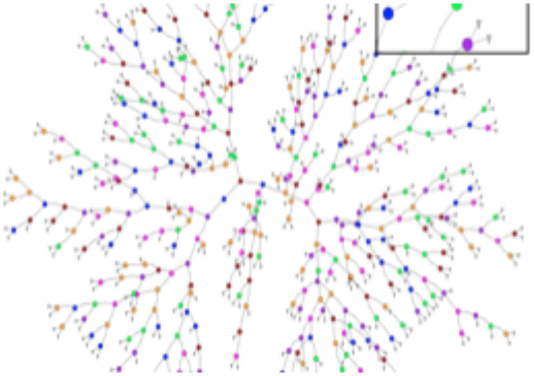
https://www.youtube.com/watch?v=hUnRCxnydCc

Three must-haves for a good explanation

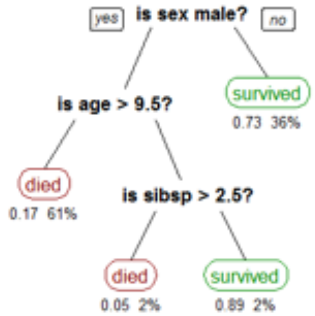
Interpretable

- Humans can easily interpret reasoning

what's interpretable depends on who the user is



Definitely
not interpretable



Potentially
interpretable

slide by Marco Tulio Ribeiro, KDD 2016

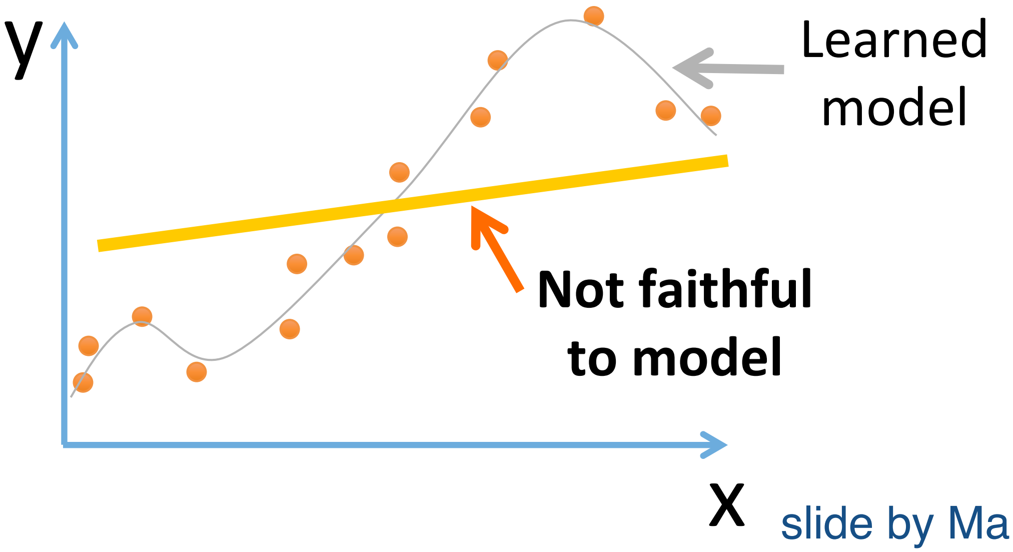
LIME: Local explanations of classifiers

[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

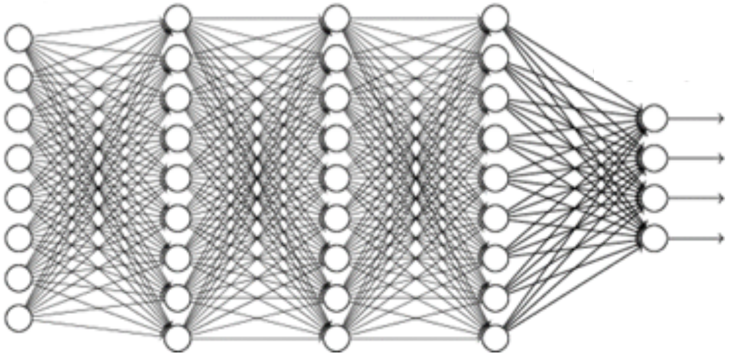
[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

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

Key idea: Interpretable representation

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

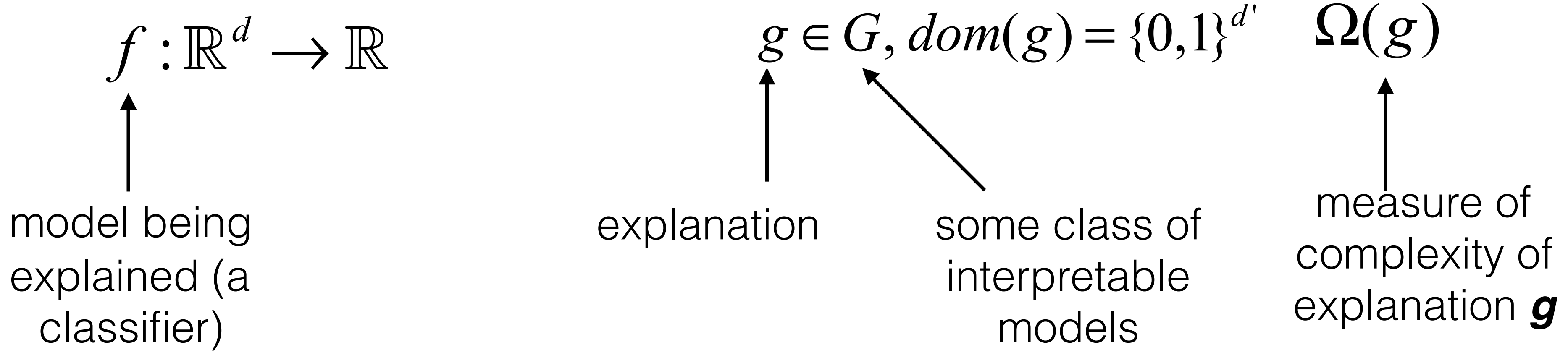
“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classier.”

- 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 of 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 of absence of each of d’ interpretable components

Fidelity-interpretability trade-off

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

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



$f : \mathbb{R}^d \rightarrow \mathbb{R}$

model being explained (a classifier)

$g \in G, dom(g) = \{0,1\}^{d'}$

explanation

some class of interpretable models

$\Omega(g)$

measure of complexity of explanation ***g***

$f(x)$ denotes the probability that ***x*** belongs to some class

π_x is a proximity measure relative to ***x***

and make no assumptions about ***f*** to remain model-agnostic - draw samples weighted by π_x

measures how unfaithful is ***g*** to ***f*** in the locality around ***x***

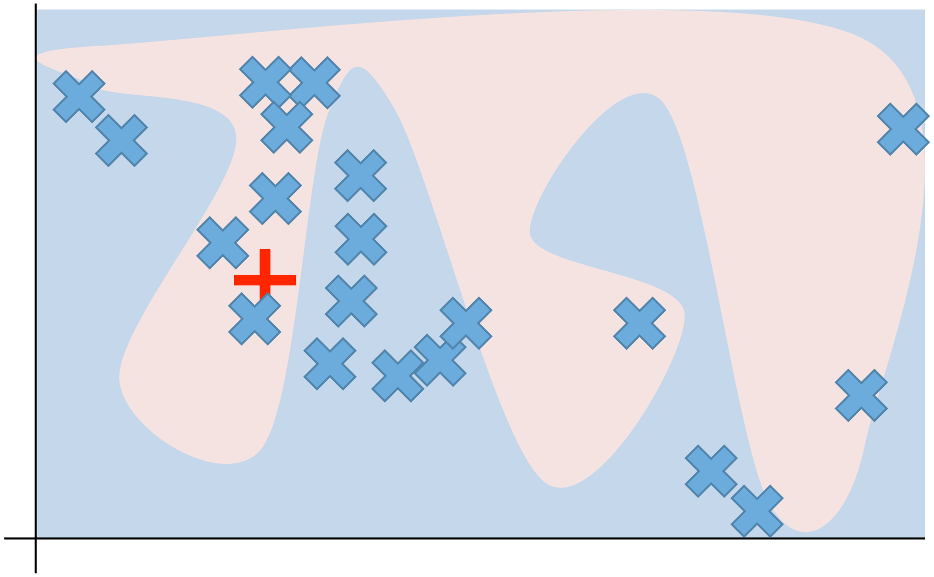
$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Fidelity-interpretability trade-off

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

“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

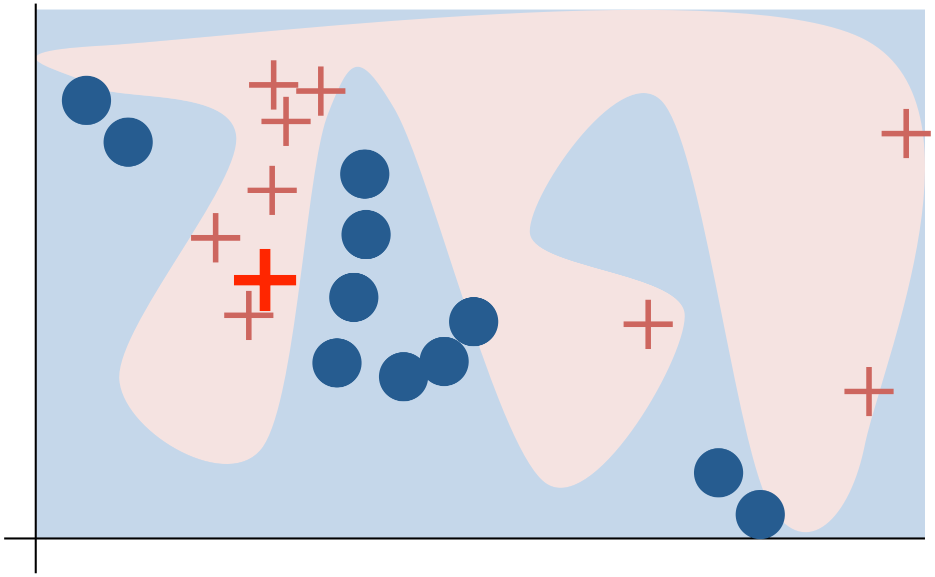
Fidelity-interpretability trade-off

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

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

1. sample points around +

2. use complex model ***f*** to assign class labels



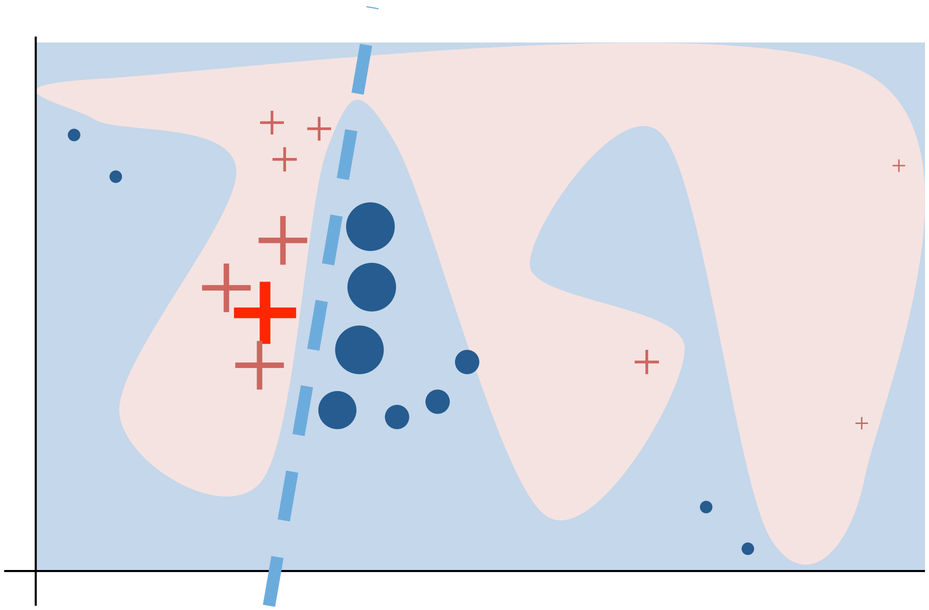
based on a slide by Marco Tulio Ribeiro, KDD 2016

Fidelity-interpretability trade-off

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

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

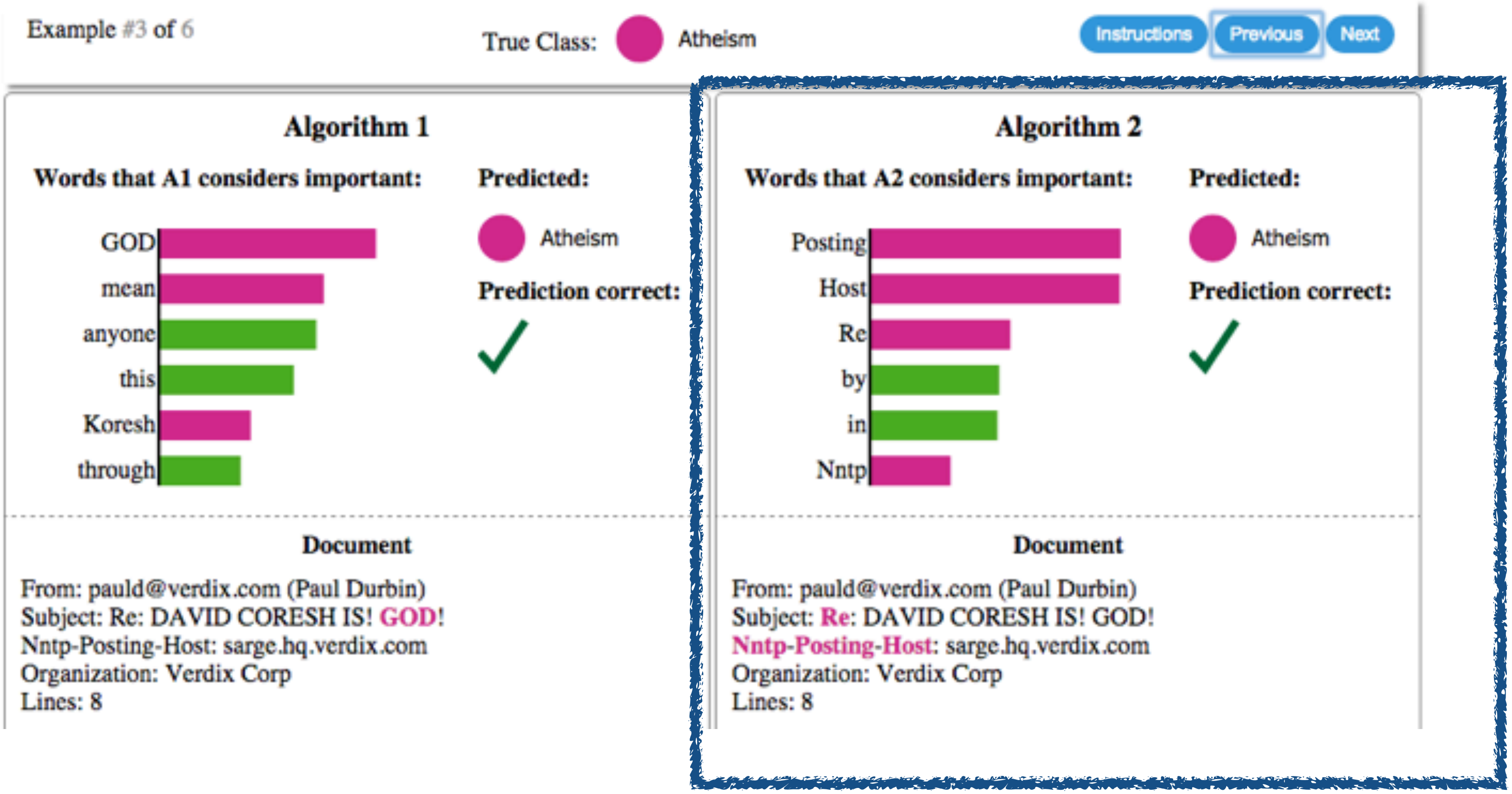
1. sample points around +
2. use complex model ***f*** to assign class labels
3. weigh samples according to π_x
4. learn simple model ***g*** according to samples



based on a slide by Marco Tulio Ribeiro, KDD 2016

Example: text classification with SVMs

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



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

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:

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

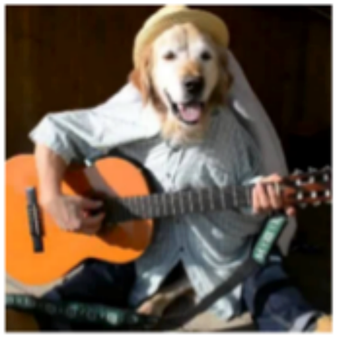
94% accuracy, yet we shouldn't trust this classifier!

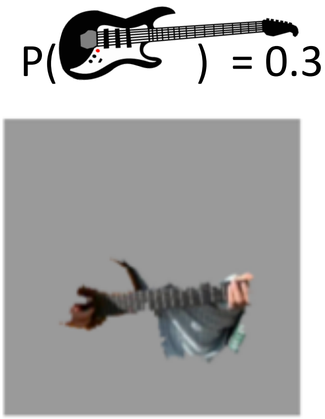
Example: deep networks for images

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

Explaining Google’s Inception NN

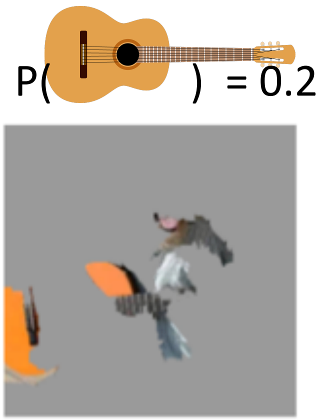
probabilities of the top-3 classes and the super-pixels predicting each





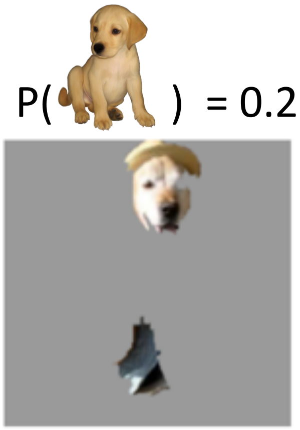
P( ) = 0.32

Electric guitar (incorrect, but this mistake is reasonable - similar fretboard)



P( ) = 0.24

Acoustic guitar



P( ) = 0.21

Labrador

based on a slide by Marco Tulio Ribeiro, KDD 2016

Next up: explaining models

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

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

- **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

Given a budget **B** of explanations that a user is willing to consider, **pick** a set of ***B*** representative instances for the user to inspect

Important to pick a set of instances that would generate a **diverse non-redundant set of explanations**, to help the user understand how the model behaves globally

Picking diverse explanations

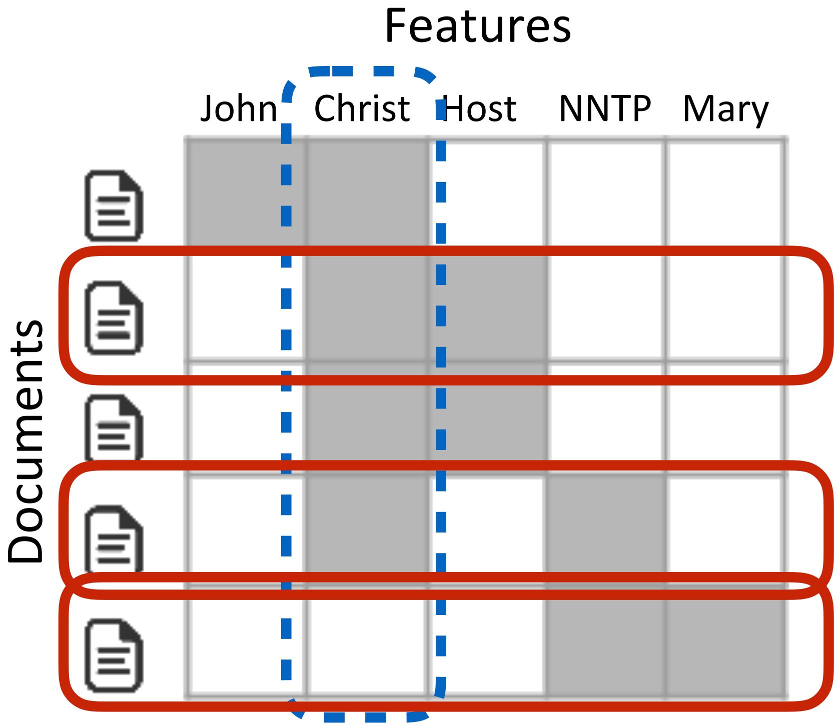
[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

Represent by a matrix the relationship between instances (here, documents) and the interpretable representations (features) that are most important in explaining the classification around those instances

“Christ” is the most important feature

Suppose that ***B* = 2**, pick 2 instances (document) to explain to the user, so as to cover most features

Slightly more complex than that, since features are weighted by their importance (in the matrix here weight are binary)



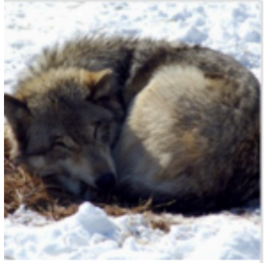
this is the problem of maximizing weighted coverage function, NP-hard
the problem is submodular, can be approximated to within 1 - 1/e with a greedy algorithm

based on a slide by Marco Tulio Ribeiro, KDD 2016

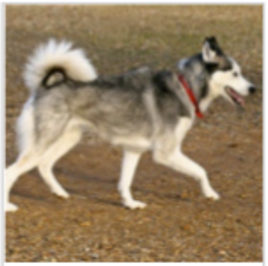
Example: deep networks for images

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

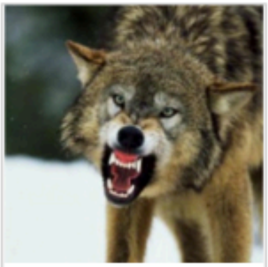
Train a neural network to predict wolf v. husky



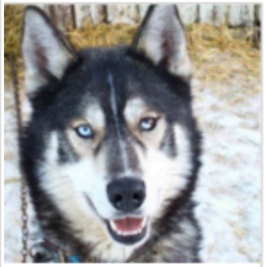
Predicted: wolf
True: wolf



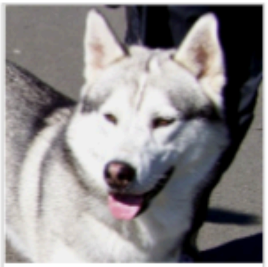
Predicted: husky
True: husky



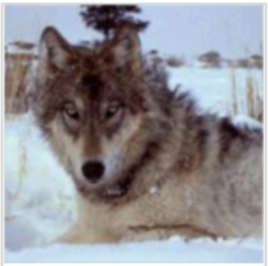
Predicted: wolf
True: wolf



Predicted: wolf
True: husky



Predicted: husky
True: husky



Predicted: wolf
True: wolf

Only 1 mistake!!!

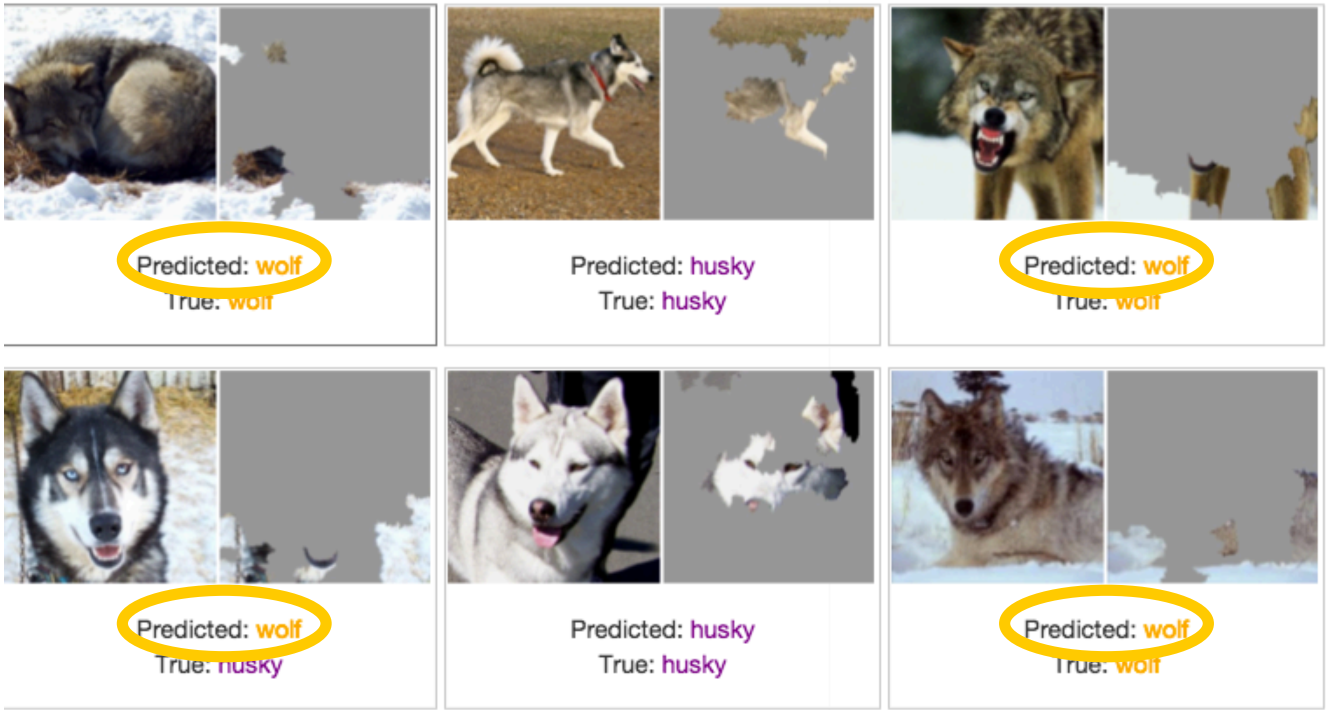
Do you trust this model?
How does it distinguish between huskies and wolves?

slide by Marco Tulio Ribeiro, KDD 2016

Example: deep networks for images

[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

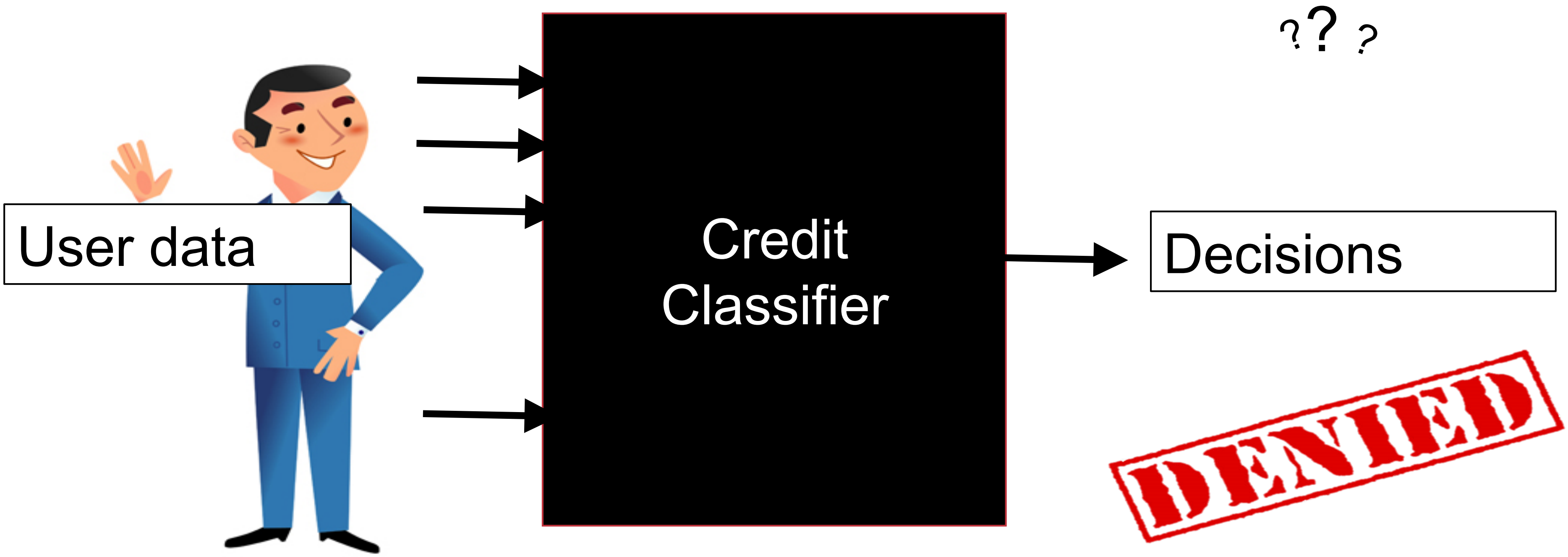
Explaining black-box classifiers

Algorithmic transparency with quantitative input influence (QII)

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

Auditing black-box models

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



User data

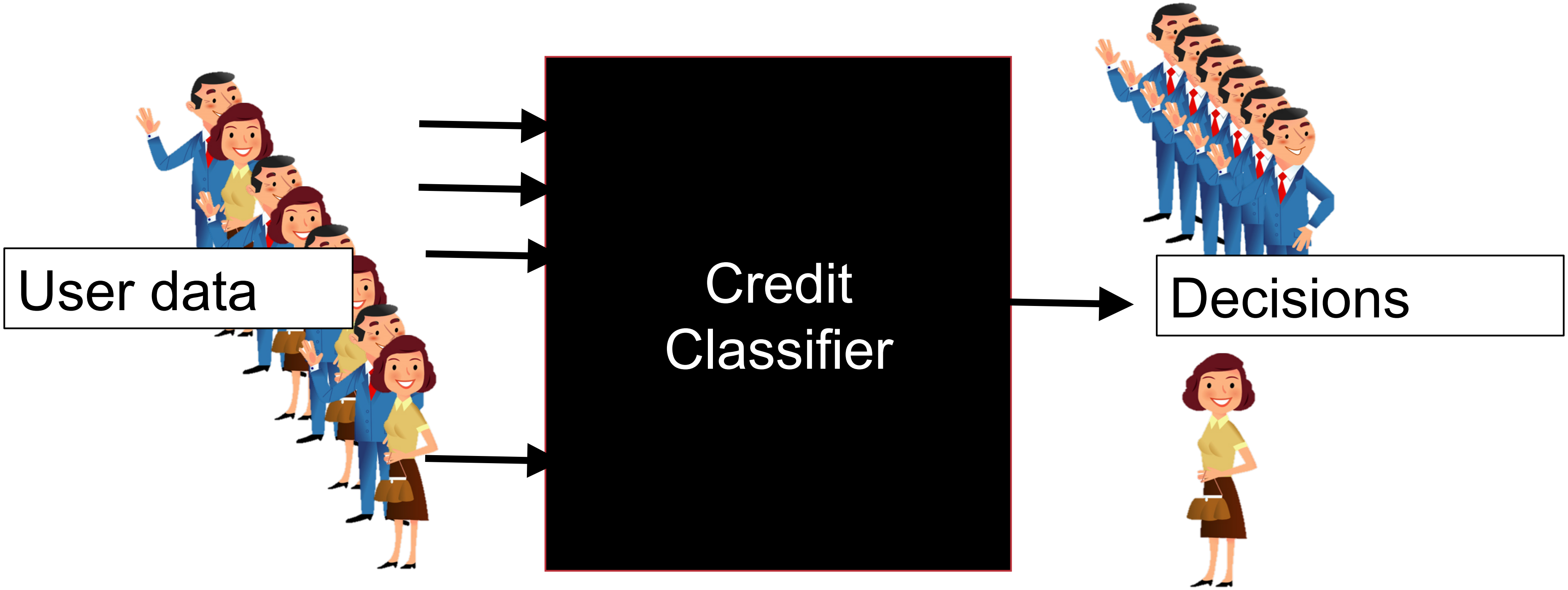
Credit Classifier

Decisions

slide by A. Datta

Auditing black-box models

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



User data

Credit Classifier

Decisions

slide by A. Datta

Influence of inputs on outcomes

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

Running example: Consider hiring decisions by a moving company, based on gender, age, education, and weight lifting ability. **Does gender influence hiring decisions?**

Possible answers:

- yes, directly
- yes, through a proxy
- yes, in combination with other features (will see an example later)
- no

which of these constitutes discrimination?

Influence of inputs on outcomes

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

Running example: Consider hiring decisions by a moving company, based on gender, age, education, and weight lifting ability. **Does gender influence hiring decisions?**

"Gender and the ability to lift heavy weights are inputs to the system. They are positively correlated with each other and with the hiring decisions. Yet ***transparency into whether the system uses the weight lifting ability or the gender in making its decisions (and to what degree) has substantive implications for determining if it is engaging in discrimination*** *(the business necessity defense could apply in the former case [E.G. Griggs v. Duke Power Co. (1977)]). This observation makes us look beyond correlation coefficients and other associative measures."*

Quantitative input influence (QII)

[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**

Intervention: Replace features with random values from the population, examine the distribution over outcomes. (More generally, sample feature values from the **prior**.)

Methodology works under **black-box access**: can specify inputs and observe outputs (as in software testing) but cannot access or analyze the code of the model. **Has knowledge of the input dataset on which the model operates**.

Back to the example

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

Running example: Consider hiring decisions by a moving company, based on gender, age, education, and weight lifting ability. **Does gender influence hiring decisions?**

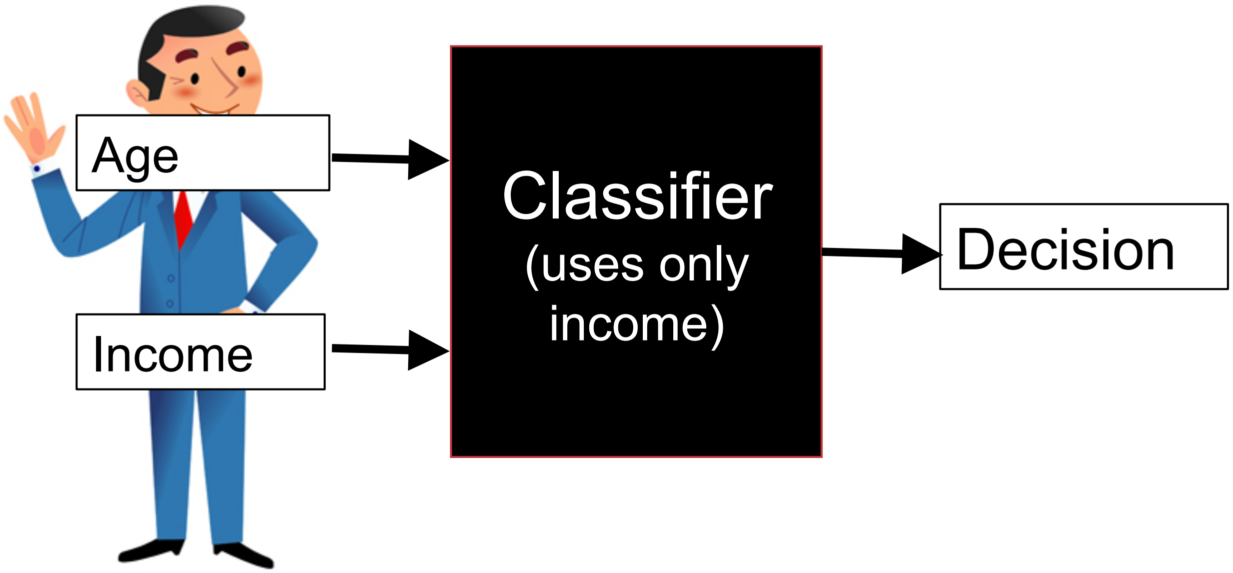
- Observe that 20% of female profiles receive the positive classification.
- To check whether gender impacts hiring decisions, take the input dataset and replace the value of gender in each input profile by drawing it from the uniform distribution: set gender in 50% of the inputs to female and 50% to male.
- If we observe that 20% of female profiles are positively classified - gender does not influence hiring decisions.
- Do a similar test for other features, one at a time. This is known as **Unary QII**

does this tell the whole story?

Unary QII

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

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

slide by A. Datta

Quantifying 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

Transparency queries / quantities of interest

Individual: Which inputs have the most influence in my credit denial?

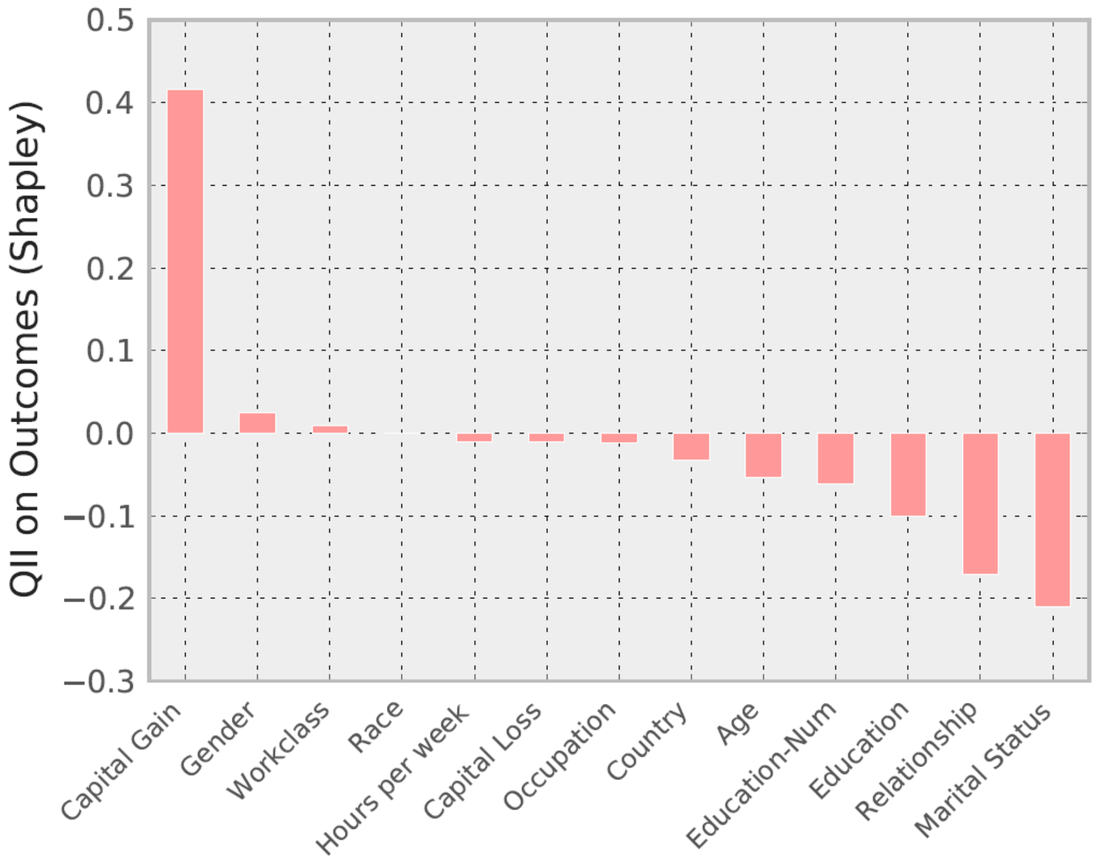
Group: Which inputs have the most influence on credit decisions for women?

Disparity: Which inputs influence men getting more positive outcomes than women?

Transparency report: Mr X

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

How much influence do individual features have a given classifier's decision about an individual?



DENIED

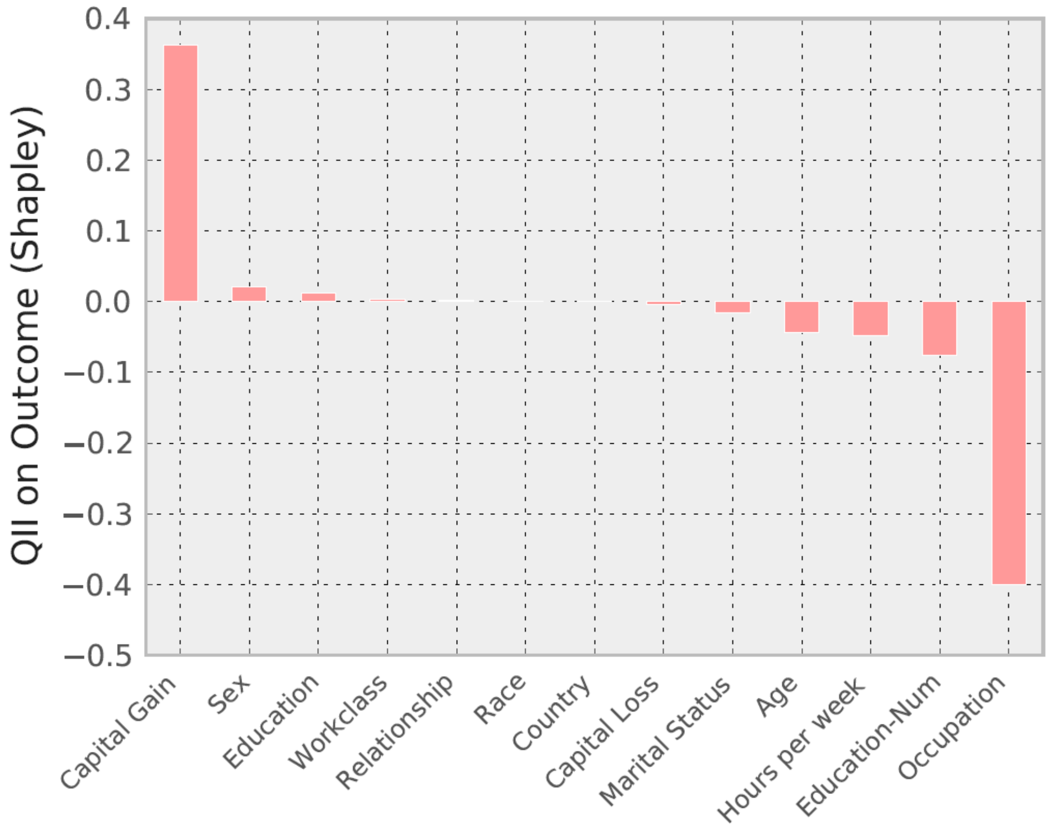
Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344
Capital loss	$0
Work hours per week	40
Country	Vietnam

income

slide by A. Datta

Transparency report: Mr Y

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



DENIED

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

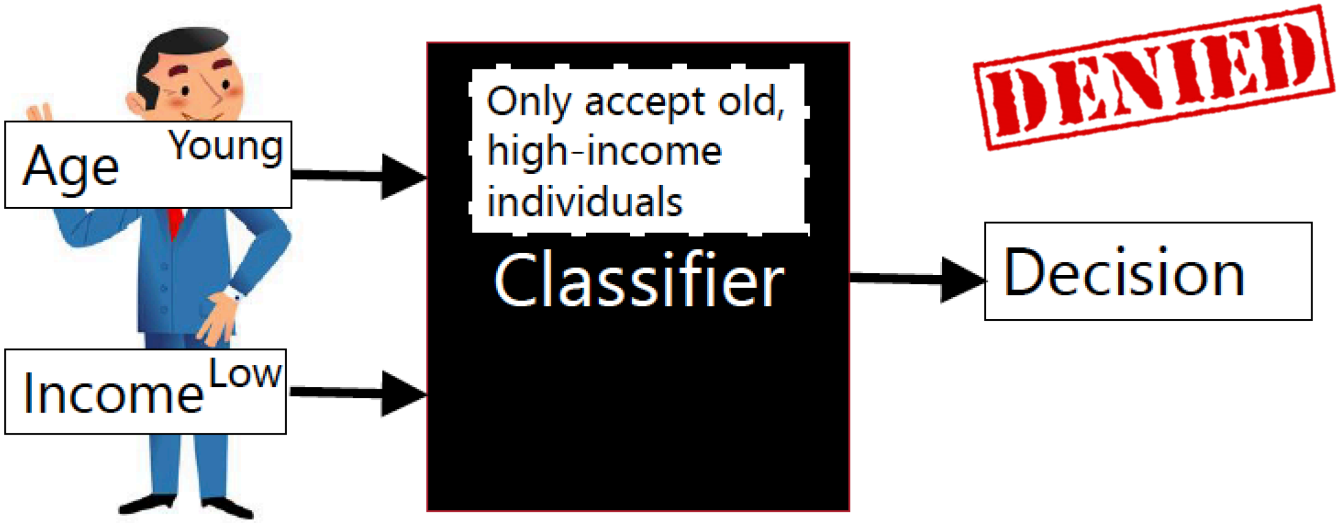
explanations for superficially similar individuals can be different

slide by A. Datta

Limitations of unary QII

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

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**.

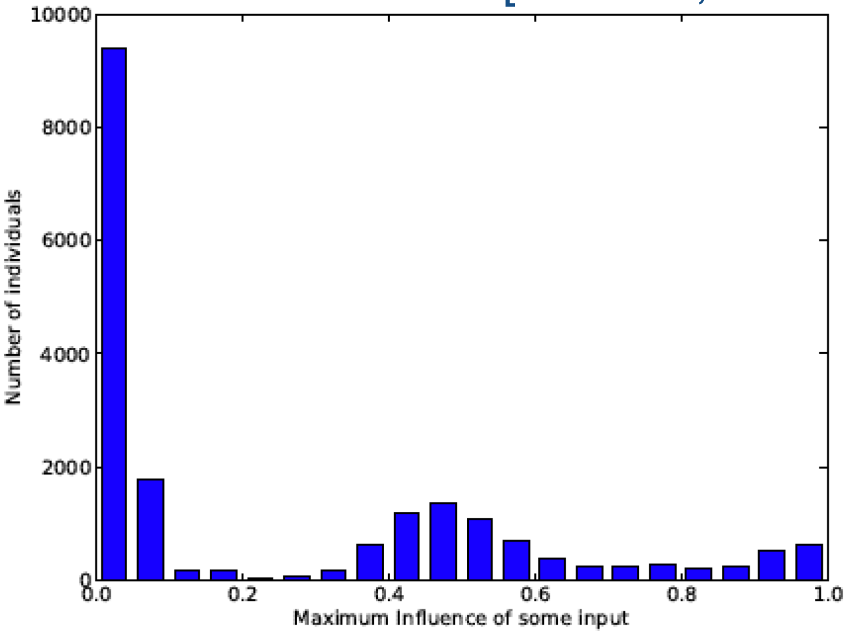


intervening on one feature at a time will not have any effect

based on a slide by A. Datta

Set and marginal QII

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



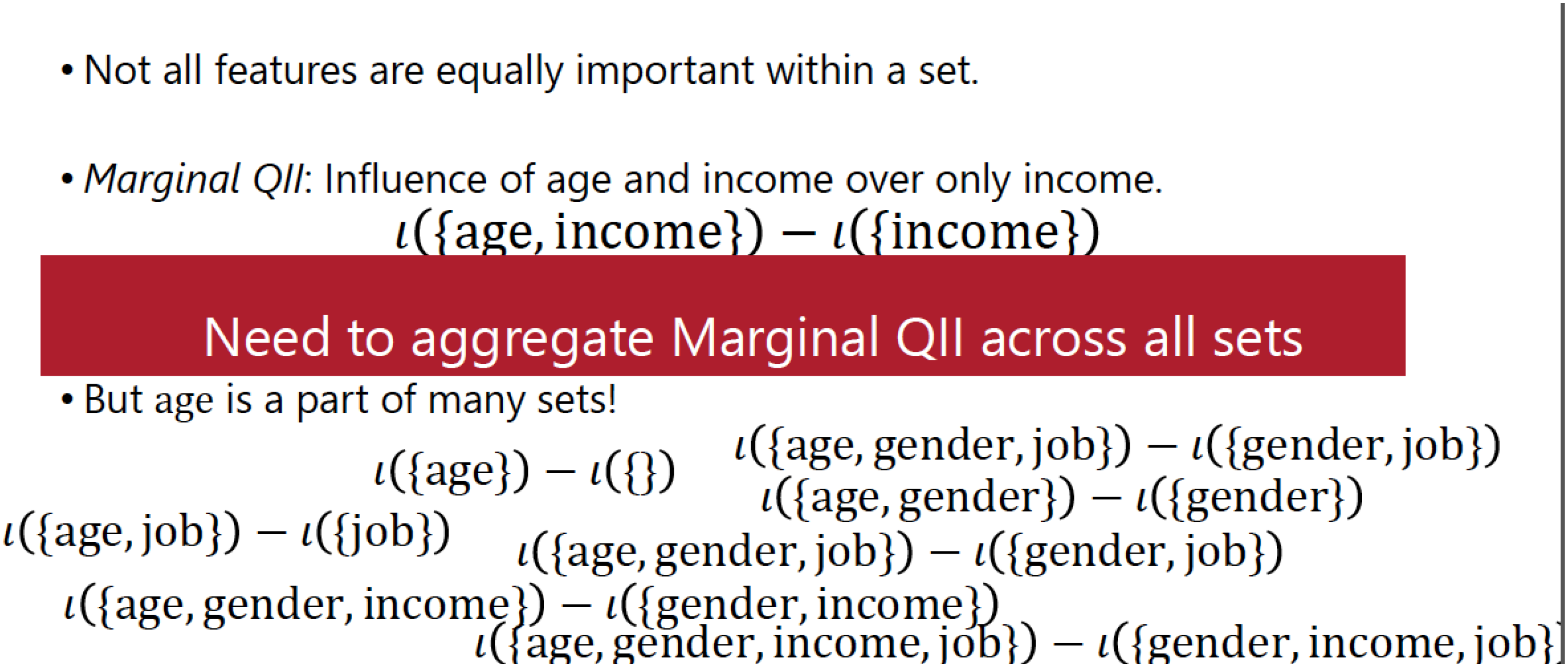
A histogram of the highest specific causal influence for some feature across individuals in the UCI adult dataset. **Alone, most inputs have very low influence.**

Set QII measures the **joint influence** of a set of features *S* on the quantity of interest *Q*.

Marginal QII measures the **added influence** of feature *i* with respect to a set of features *S* on the quantity of interest *Q*. Use cooperative games (Shapley value) to aggregate marginal influence

Marginal QII

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



- Not all features are equally important within a set.

- *Marginal QII*: Influence of age and income over only income.

$$\iota(\{\text{age}, \text{income}\}) - \iota(\{\text{income}\})$$

Need to aggregate Marginal QII across all sets

- But age is a part of many sets!

$$\iota(\{\text{age}, \text{gender}, \text{job}\}) - \iota(\{\text{gender}, \text{job}\})$$

$$\iota(\{\text{age}\}) - \iota(\{\})$$

$$\iota(\{\text{age}, \text{gender}\}) - \iota(\{\text{gender}\})$$

$$\iota(\{\text{age}, \text{job}\}) - \iota(\{\text{job}\})$$

$$\iota(\{\text{age}, \text{gender}, \text{job}\}) - \iota(\{\text{gender}, \text{job}\})$$

$$\iota(\{\text{age}, \text{gender}, \text{income}\}) - \iota(\{\text{gender}, \text{income}\})$$

$$\iota(\{\text{age}, \text{gender}, \text{income}, \text{job}\}) - \iota(\{\text{gender}, \text{income}, \text{job}\}$$

slide by A. Datta

Aggregating influence across sets

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

Idea: Use game theory methods: voting systems, revenue division

"In voting systems with multiple agents with differing weights, voting power often does not directly correspond to the weights of the agents. For example, the US presidential election can roughly be modeled as a cooperative game where each state is an agent. The ***weight of a state is the number of electors in that state*** *(i.e., the number of votes it brings to the presidential candidate who wins that state). Although states like California and Texas have higher weight, swing states like Pennsylvania and Ohio tend to have higher power in determining the outcome of elections."*

This paper uses the **Shapley value** as the aggregation mechanism

$$\varphi_i(N, v) = \mathbb{E}_\sigma[m_i(\sigma)] = \frac{1}{n!} \sum_{\sigma \in \Pi(N)} m_i(\sigma)$$

QII, in summary

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

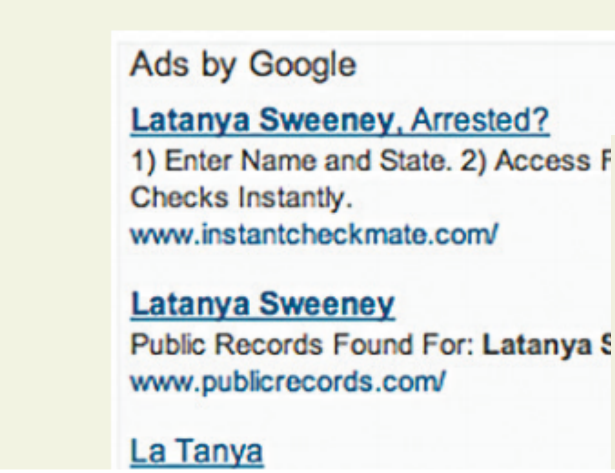
- A principled (and beautiful!) framework for determining the influence of a feature, or a set of features, on a decision
- Works for black-box models, with the assumption that the full set of inputs is available
- Accounts for correlations between features
- "Parametrizes" on what quantity we want to set (QII), how we intervene, how we aggregate the influence of a feature across sets
- Experiments in the paper: interesting results
- Also in the paper: a discussion of transparency under differential privacy

Online ad delivery

Racially identifying names

[Latanya Sweeney; *CACM 2013*]

Google AdSense

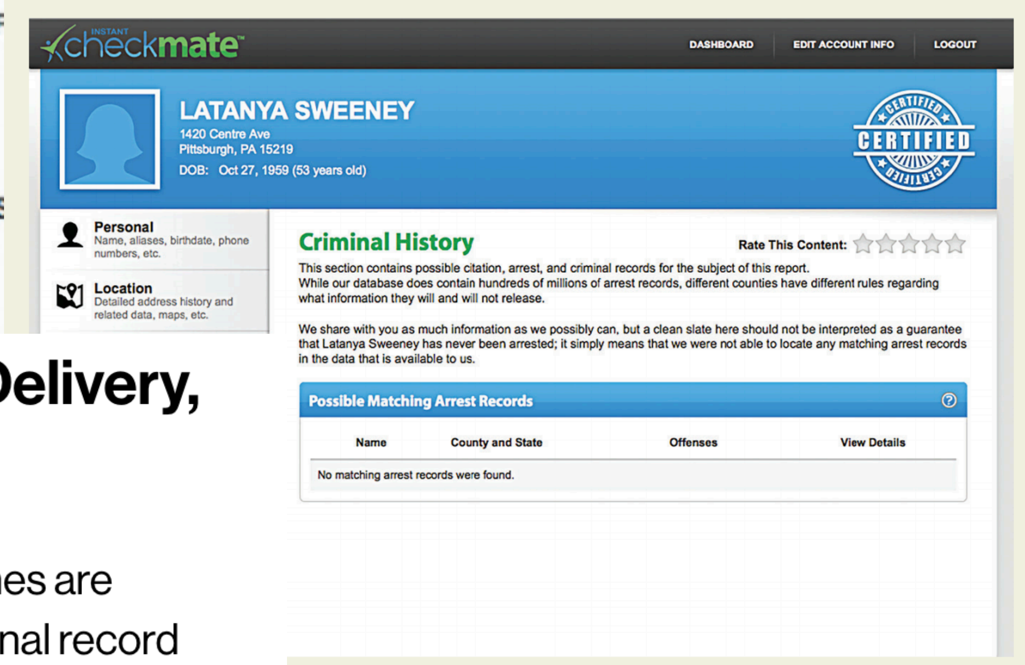


Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access F
Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: Latanya S
www.publicrecords.com/

La Tanya



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.

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.			

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/

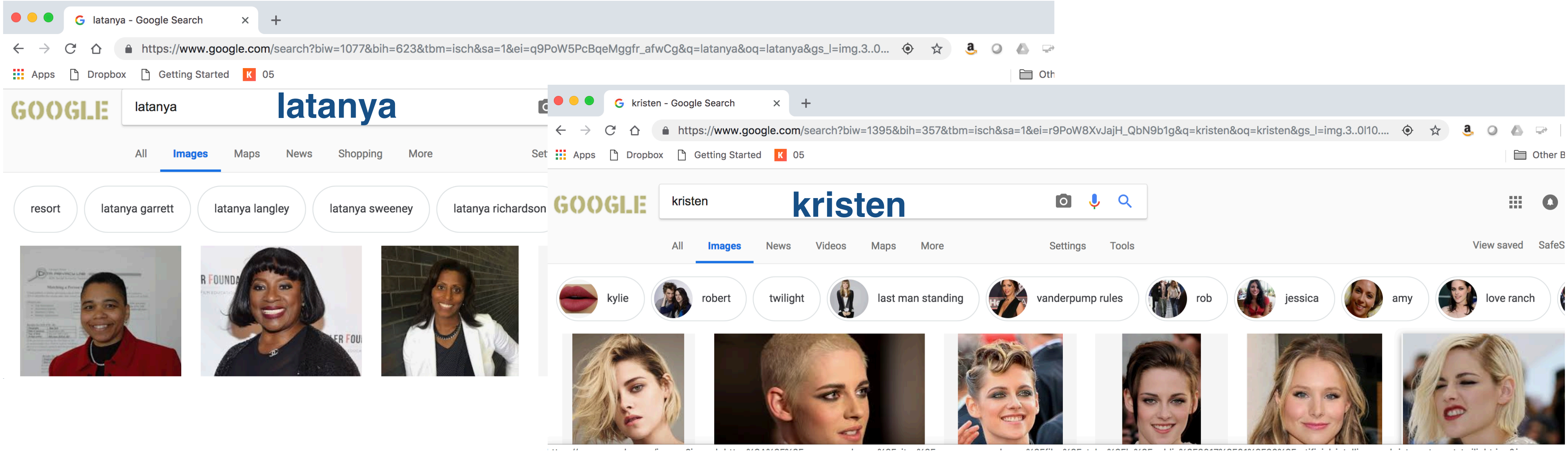
Observations

[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



latanya - Google Search

https://www.google.com/search?biw=1077&bih=623&tbm=isch&sa=1&ei=q9PoW5PcBqeMggfr_afwCg&q=latanya&oq=latanya&gs_l=img.3..0...

latanya

kristen - Google Search

https://www.google.com/search?biw=1395&bih=357&tbm=isch&sa=1&ei=r9PoW8XvJajH_QbN9b1g&q=kristen&oq=kristen&gs_l=img.3..0l10....

kristen

Racially identifying names: details

[Latanya Sweeney; *CACM 2013*]

”A greater percentage of Instant Checkmate ads having the word arrest in ad text appeared for black-identifying first names than for white-identifying first names within professional and netizen subsets, too. On Reuters.com, which hosts Google AdSense ads, **a black-identifying name was 25% more likely to generate an ad suggestive of an arrest record**.”

More than 1,100 Instant Checkmate ads appeared on Reuters.com, with 488 having black-identifying first names; of these, 60% used arrest in the ad text. Of the 638 ads displayed with white-identifying names, 48% used arrest. This difference is statistically significant, with less than a 0.1% probability that the data can be explained by chance (chi-square test: X^2 (1)=14.32, p < 0.001).

The EEOC’s and U.S. Department of Labor’s adverse impact test for measuring discrimination is 77 in this case, so if this were an employment situation, a charge of discrimination might result. (The adverse impact test uses the ratio of neutral ads, or 100 minus the percentages given, to compute disparity: 100-60=40 and 100- 48=52; dividing 40 by 52 equals 77.)

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 t:

“**AdWords does not conduct any racial profiling**. We al 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 technology in place to even connect a name with a race** any attempt to do so. The very idea is contrary to our comp principles and values.”

Who is responsible?

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

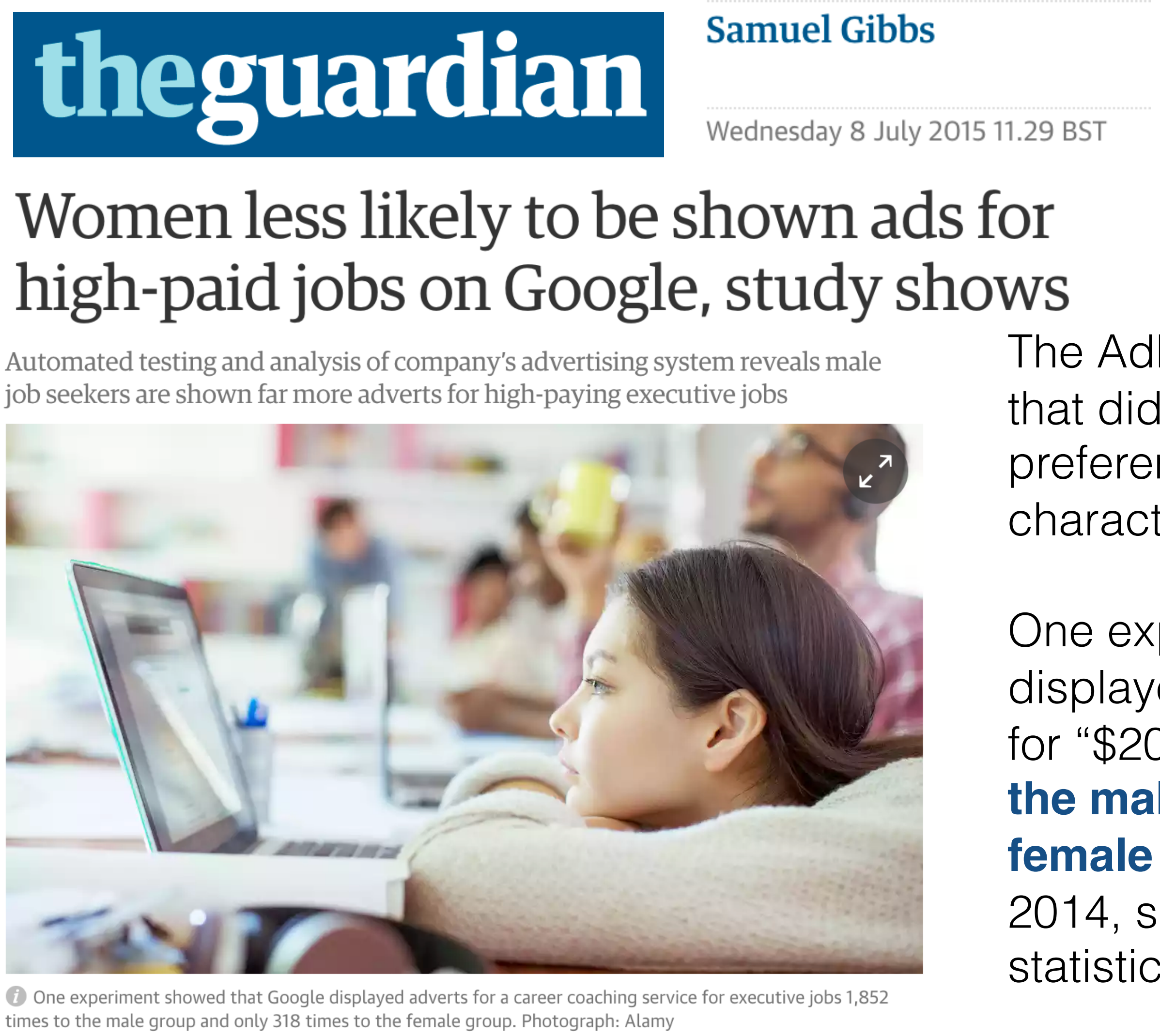
transparency …. **responsibility** …. **trust**

Gender discrimination in online job ads

Automated Experiments on Ad Privacy Settings (AdFisher)

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

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.

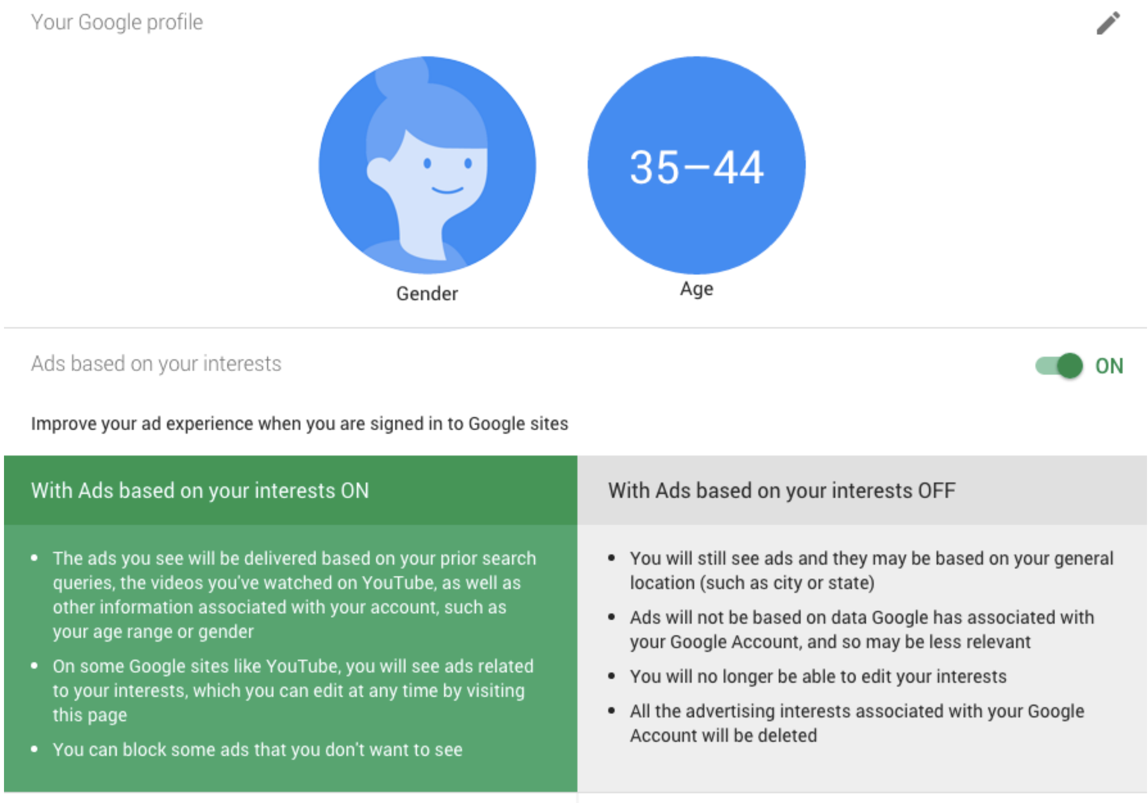
https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

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

Google ad settings

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



Your Google profile

Ads based on your interests ON

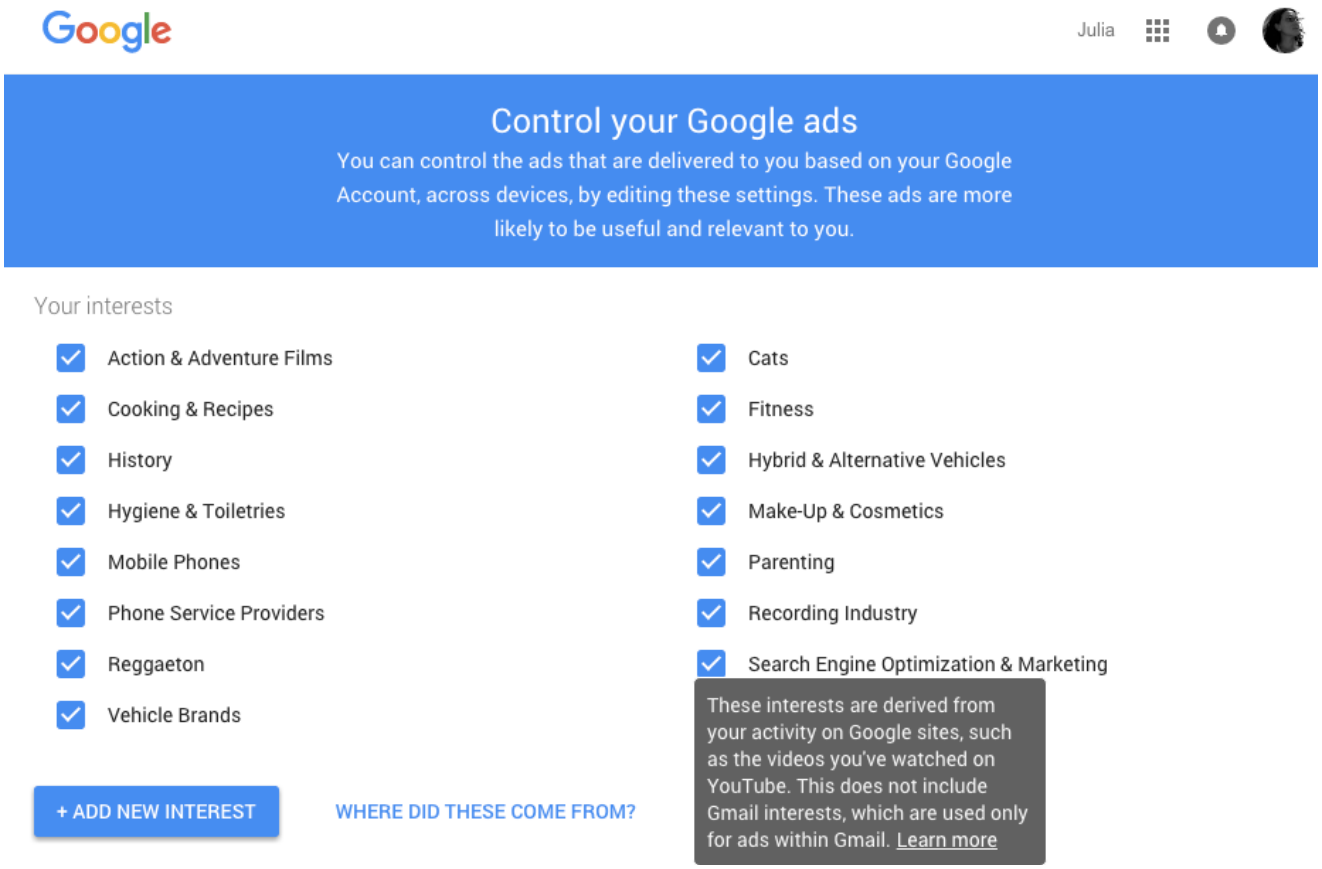
Improve your ad experience when you are signed in to Google sites

With Ads based on your interests ON	With Ads based on your interests OFF
• 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 • 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 • You can block some ads that you don't want to see	• You will still see ads and they may be based on your general location (such as city or state) • Ads will not be based on data Google has associated with your Google Account, and so may be less relevant • You will no longer be able to edit your interests • All the advertising interests associated with your Google Account will be deleted

http://www.google.com/settings/ads

Google ad settings

Do users truly have transparency / choice or is this a placebo button?



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

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

http://www.google.com/settings/ads

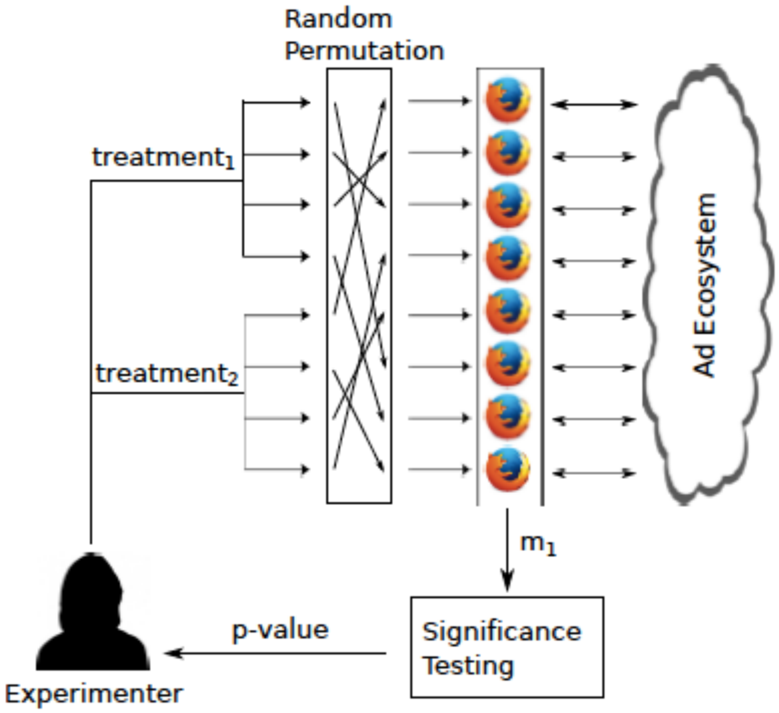
AdFisher

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

From anecdotal evidence to statistical insight: How do user behaviors, ads and ad settings interact?

Automated randomized controlled experiments for studying online tracking

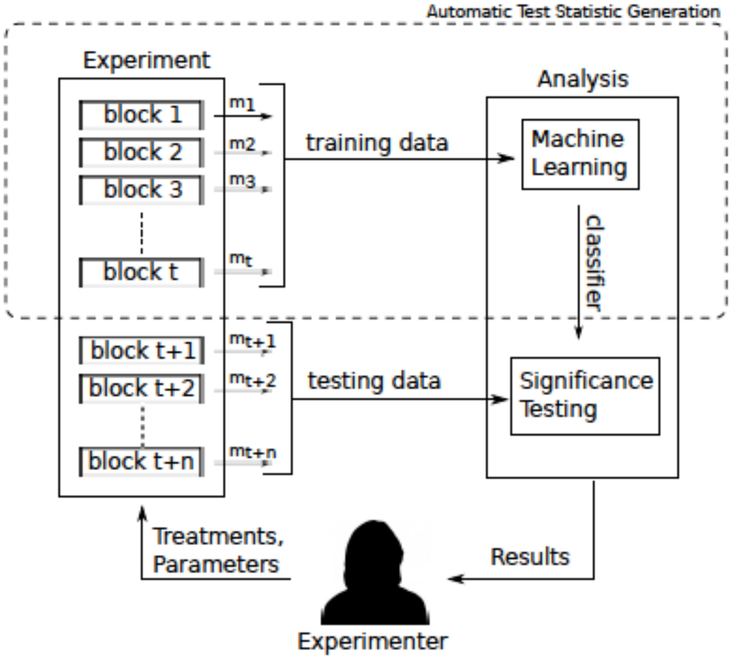
Individual data use transparency: ad network must share the information it uses about the user to select which ads to serve to him



AdFisher: methodology

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

- Browser-based experiments, 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
- Fisher randomized hypothesis testing
  - **null hypothesis** inputs do not affect outputs
  - control and experimental treatments
  - AdFisher can help select a test statistic



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

AdFisher: substance abuse

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

Transparency: User can view data about him used for ad selection

Causal test: Find attribute that changes ads but not settings

Experiment 2: **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 what ads are served: rehab vs. stocks + driving jobs

violation

AdFisher: online dating

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

Ad choice: Removing an interest decreases the number of ads related to that interest.

Causal test: Find that removing an interest causes 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

Recall the set-up

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

- **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

Why are males seeing ads for high-paying jobs more often?

What is causing gender-based discrimination?

(1) who is responsible and (2) how is discrimination enacted?

Who is responsible?

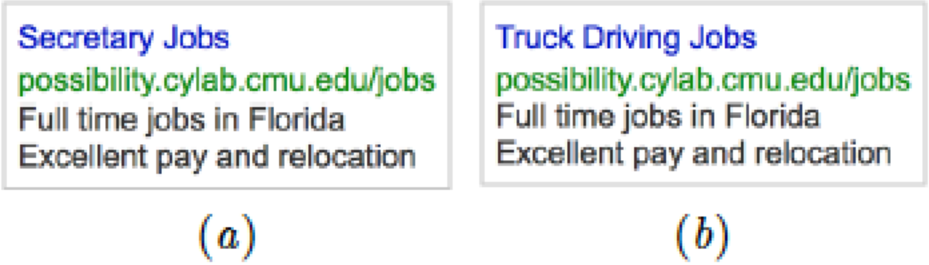
[A. Datta, A. Datta, J. Makagon, D. Mulligan, M. Tschantz; *FAT* 2018*]

- **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*]

- 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



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

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

(a) (b)

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

experiments show that is possible to use Google AdWords to target on gender

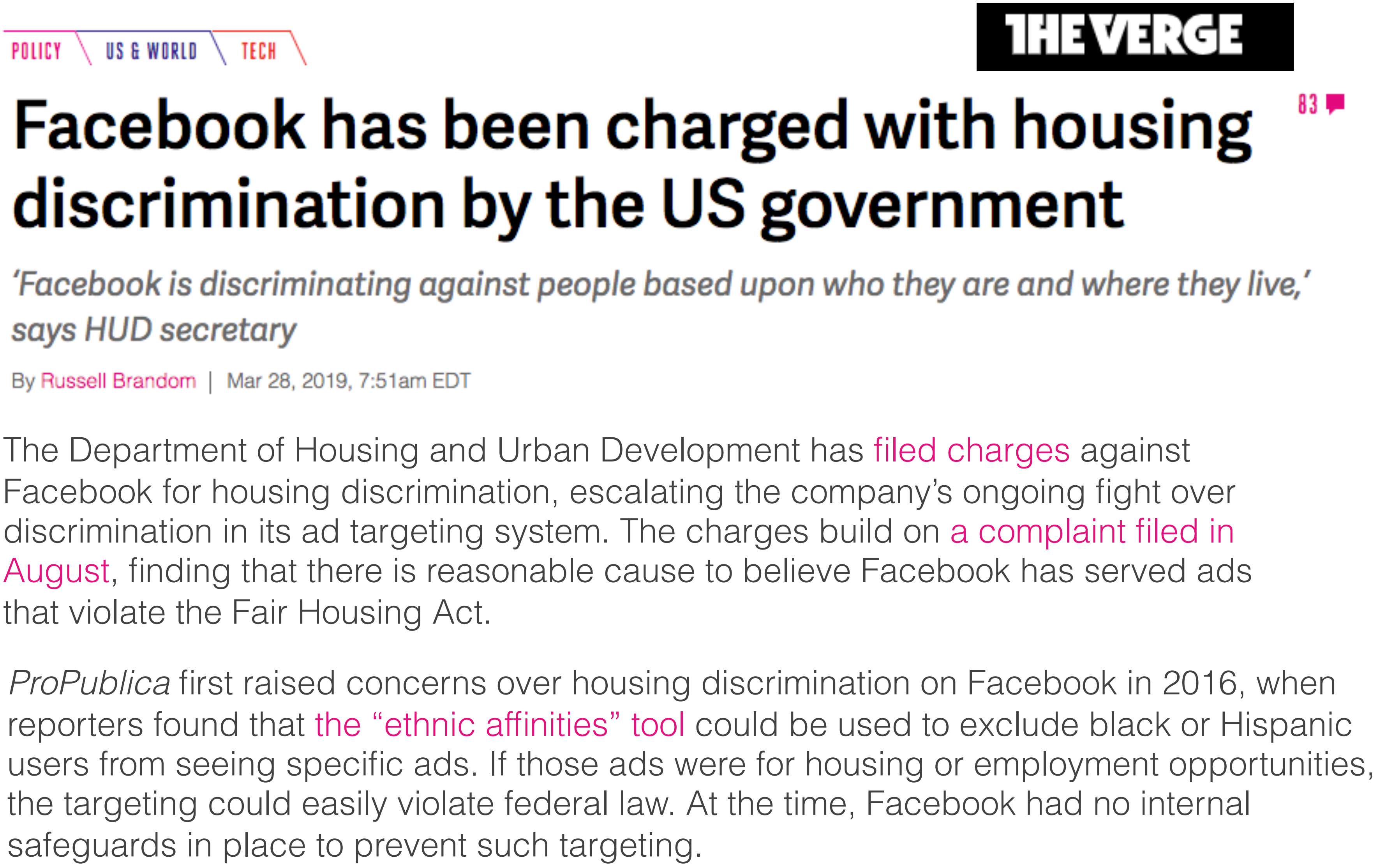
"This finding demonstrates that an advertiser with discriminatory intentions can use the AdWords platform to serve employment related ads disparately on gender."

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 the different 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** 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

In the news: Facebook ads



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

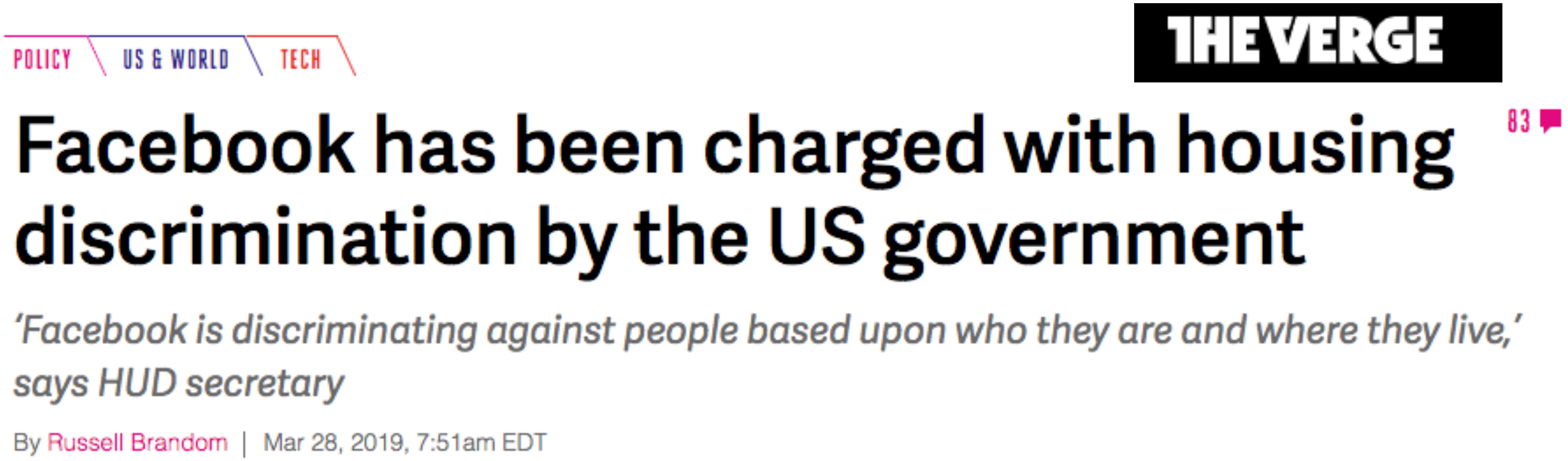
By Russell Brandom | Mar 28, 2019, 7:51am EDT

The Department of Housing and Urban Development has filed charges against Facebook for housing discrimination, escalating the company's ongoing fight over discrimination in its ad targeting system. The charges build on a complaint filed in August, finding that there is 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.

https://www.theverge.com/2019/3/28/18285178/facebook-hud-lawsuit-fair-housing-discrimination

In the news: Facebook ads



POLICY | US & WORLD | TECH

THE VERGE

83

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

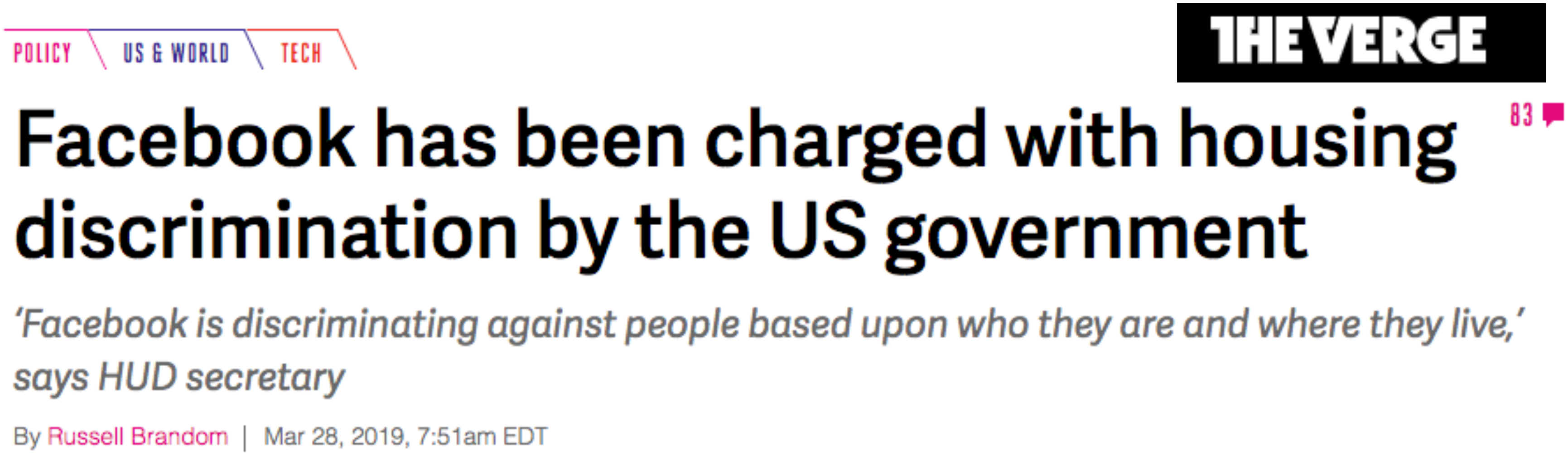
By Russell Brandom | Mar 28, 2019, 7:51am EDT

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.

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.

https://www.theverge.com/2019/3/28/18285178/facebook-hud-lawsuit-fair-housing-discrimination

In the news: Google and Twitter ads



POLICY \ US & WORLD \ TECH

THE VERGE

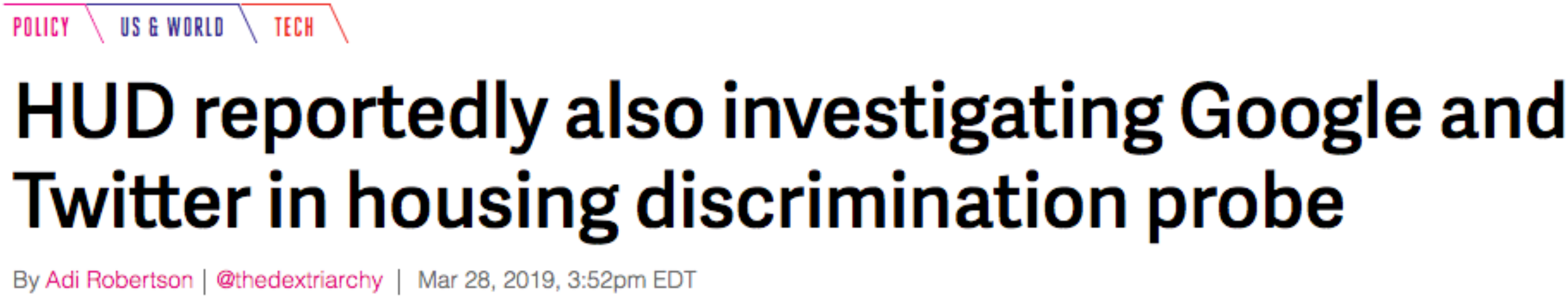
Facebook has been charged with housing discrimination by the US government

83

'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

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



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

https://www.theverge.com/2019/3/28/18285899/housing-urban-development-hud-facebook-lawsuit-google-twitter

Discrimination in Facebook's ad delivery

Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

Discrimination in Facebook's ad delivery

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

- 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 still remains: **Does the ad delivery platform itself embed discriminatory outcomes?**

Facebook ad delivery

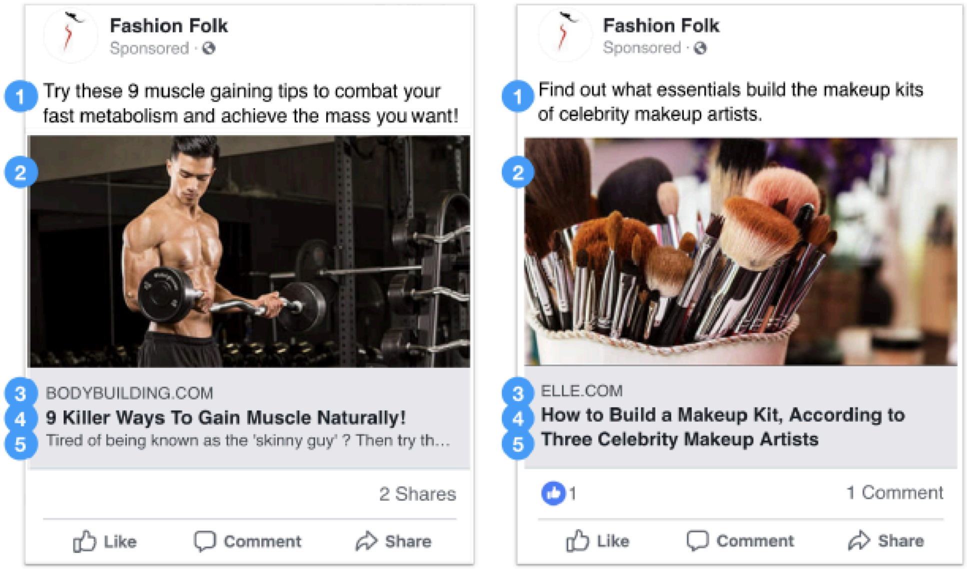
[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

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.



Fashion Folk
Sponsored ·

Try these 9 muscle gaining tips to combat your fast metabolism and achieve the mass you want!

BODYBUILDING.COM
9 Killer Ways To Gain Muscle Naturally!
Tired of being known as the 'skinny guy' ? Then try th...

2 Shares

Fashion Folk
Sponsored ·

Find out what essentials build the makeup kits of celebrity makeup artists.

ELLE.COM
How to Build a Makeup Kit, According to Three Celebrity Makeup Artists

1
1 Comment

Figure 1: Each ad has five elements that the advertiser can control: (1) the ad headline and 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.

Facebook ad delivery

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

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, t**he 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.****"*

Facebook ad delivery: insights

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

Facebook ad delivery results can be skewed **in ways that advertisers do not intend**

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

- What contributes to the skew?
  - ad content - both text and images, which are likely automatically analyzed by Facebook
  - advertiser budget

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

Budget impacts demographics

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

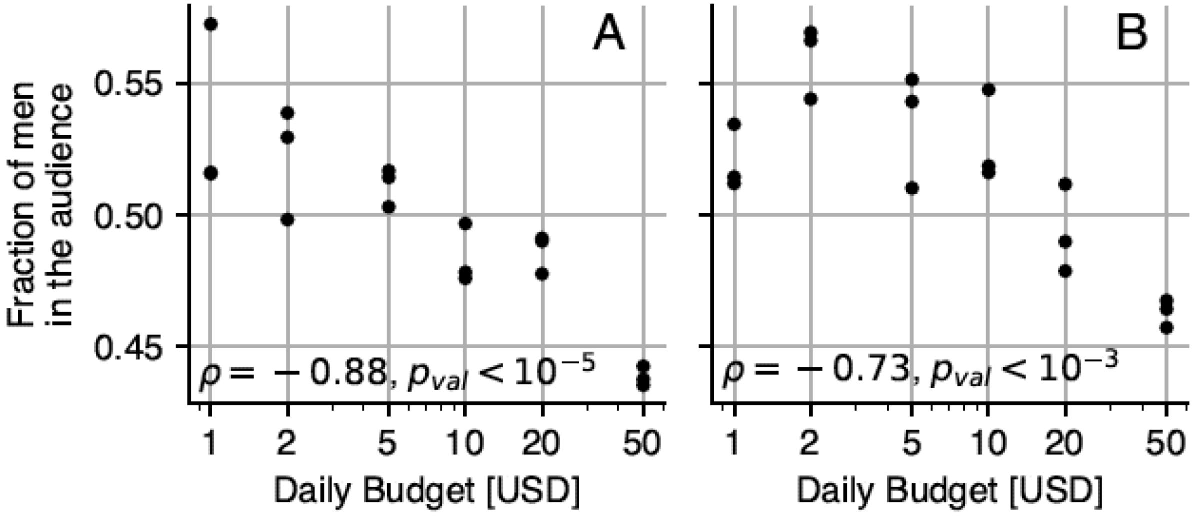
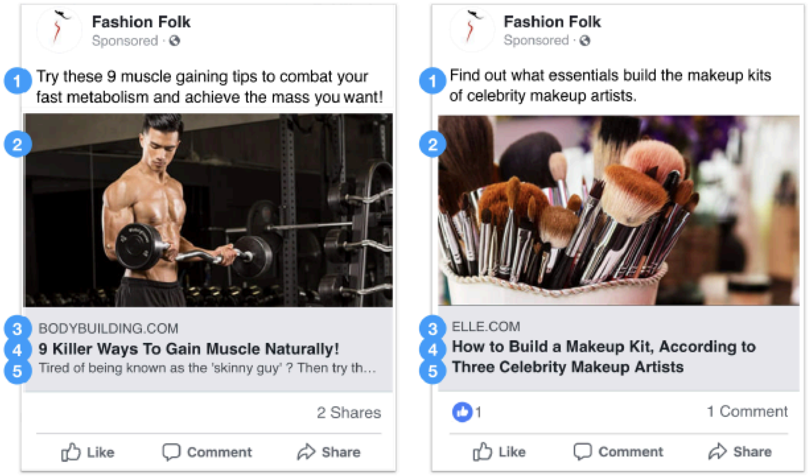


Figure 2: Gender distributions of the audience depend on the daily budget of an ad, with higher budgets leading to a higher fraction of women. The left graph shows an experiment where we target all users located in the U.S.; the right graph shows an experiment where we target our random phone number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

Ad creative impacts demographics

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]



Fashion Folk
Sponsored ·

1 Try these 9 muscle gaining tips to combat your fast metabolism and achieve the mass you want!

2

3 BODYBUILDING.COM
4 9 Killer Ways To Gain Muscle Naturally!
5 Tired of being known as the 'skinny guy' ? Then try th...

2 Shares

Fashion Folk
Sponsored ·

1 Find out what essentials build the makeup kits of celebrity makeup artists.

2

3 ELLE.COM
4 How to Build a Makeup Kit, According to Three Celebrity Makeup Artists

1 1 Comment

Figure 1: Each ad has five elements that the advertiser can control: (1) the ad headline and 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.

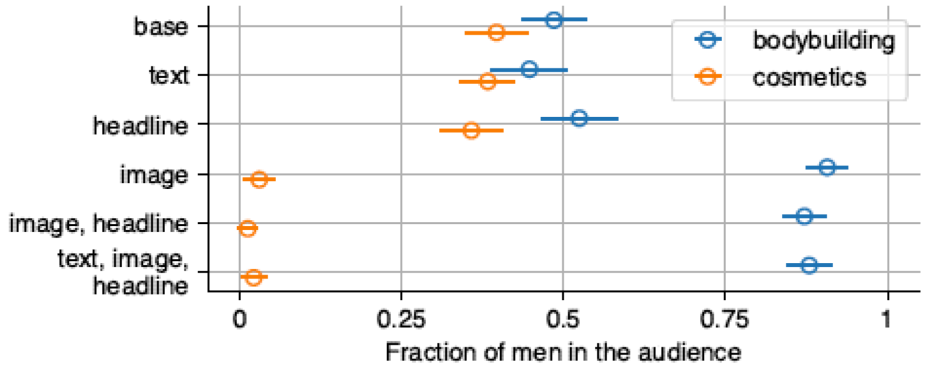


Figure 3: "Base" ad contains a link to a page about either bodybuilding or cosmetics, a blank image, no text, or headline. There is a small difference in the fraction of male users for the base ads, and adding setting the "text" only decreases it. Setting the "headline" sets the two ads apart but the 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 very few men, despite the same target audience.

Transparent images are still targeted!

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

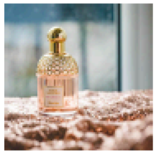
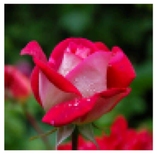
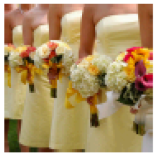
	Male		Female	
No.	Visible	Invisible	Visible	Invisible
1				
2				
3				
4				
5				

Table 2: Diagram of the images used in the transparency experiments. Shown are the five stereotypical male and female images, along with the same images with a 98% alpha channel, denoted as invisible. The images with the alpha channel are almost invisible to humans, but are still delivered in a skewed manner.

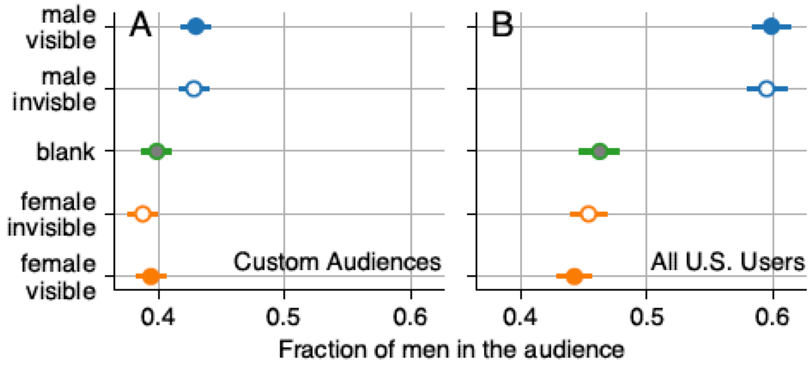


Figure 6: Ad delivery to ads with the images from Table 2, targeting general US audience as well as the random phone number custom audience. The solid markers are visible images, and the hollow markers are the same images with 98% opacity. Also shown is the delivery to truly white images ("blank"). We can observe that a difference in ad delivery exists, and that that difference is statistically significant between the male and female invisible images. This suggests that automated image classification is taking place.

Racial discrimination in housing ads

[M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke; *arXiv 2019*]

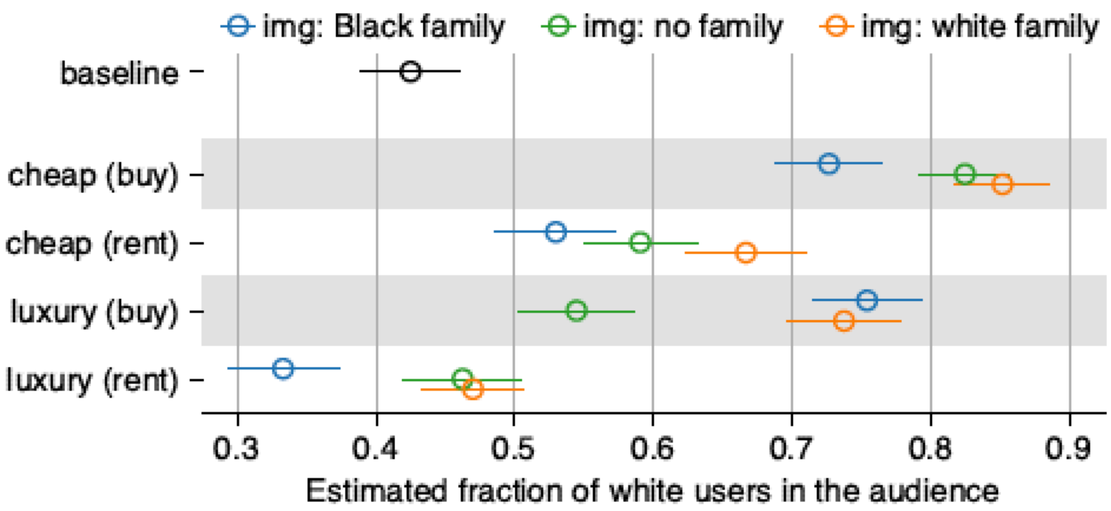


Figure 9: Results for housing ads, showing a breakdown in the ad delivery audience by race. Despite being targeted in the same manner, using the same bidding strategy, and being run at the same time, we observe significant skew in the makeup of the audience to whom the ad is delivered (ranging from over 85% white users to over 65% Black users).