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/

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

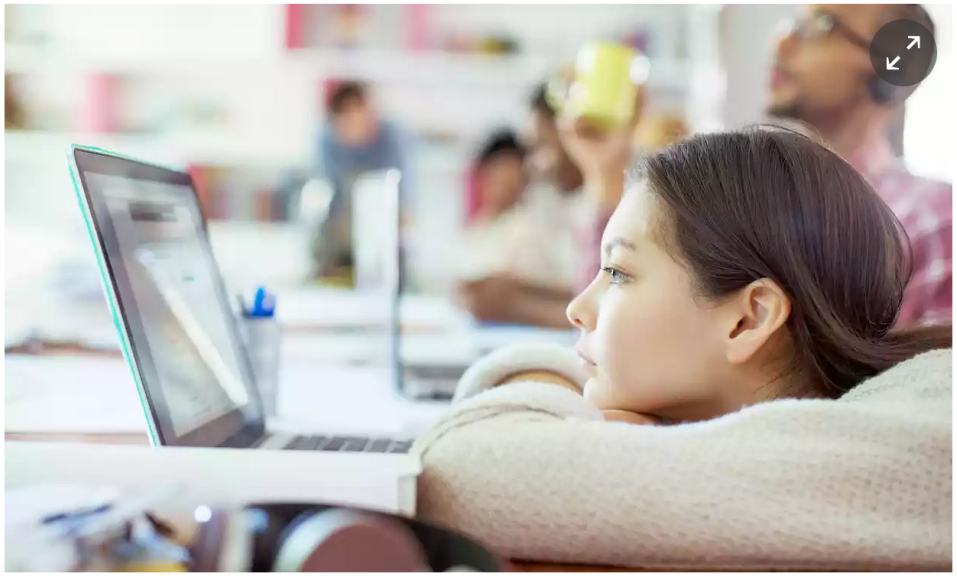
theguardian

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.

Are Workplace Personality Tests Fair?

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

By LAUREN WEBER and ELIZABETH DWOSKIN

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

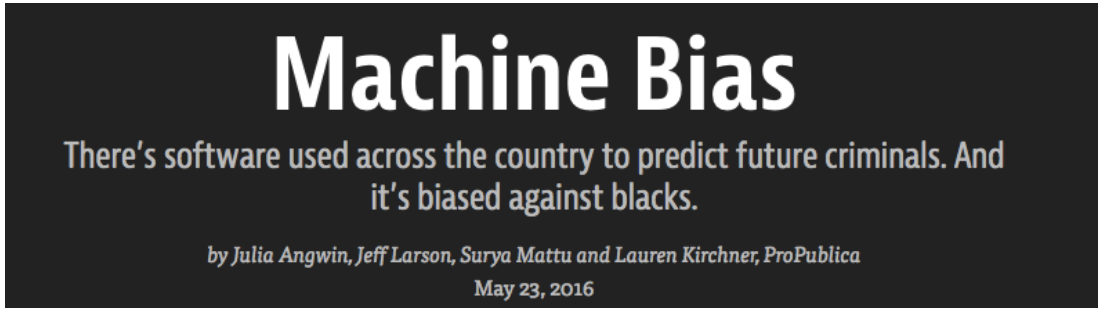
Kyle Behm accused Kroger and six other companies of discrimination against the mentally ill through their use of personality tests. *TROY STAINS FOR THE WALL STREET JOURNAL*

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



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.

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 precautions 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?

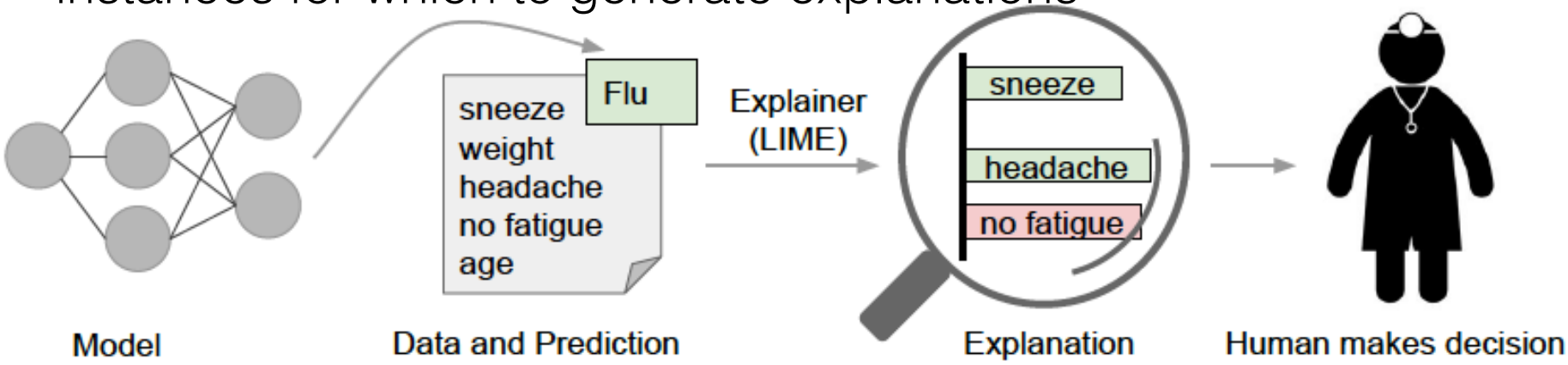
We wouldn't be discussing interpretability if accuracy were sufficient, but what are some of the reasons that accuracy may not be enough?

- 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*]

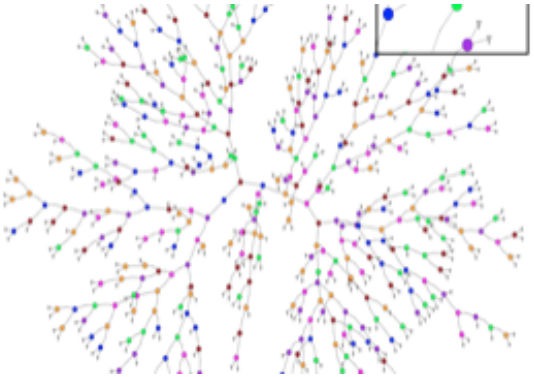
https://www.youtube.com/watch?v=hUnRCxnydCc

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

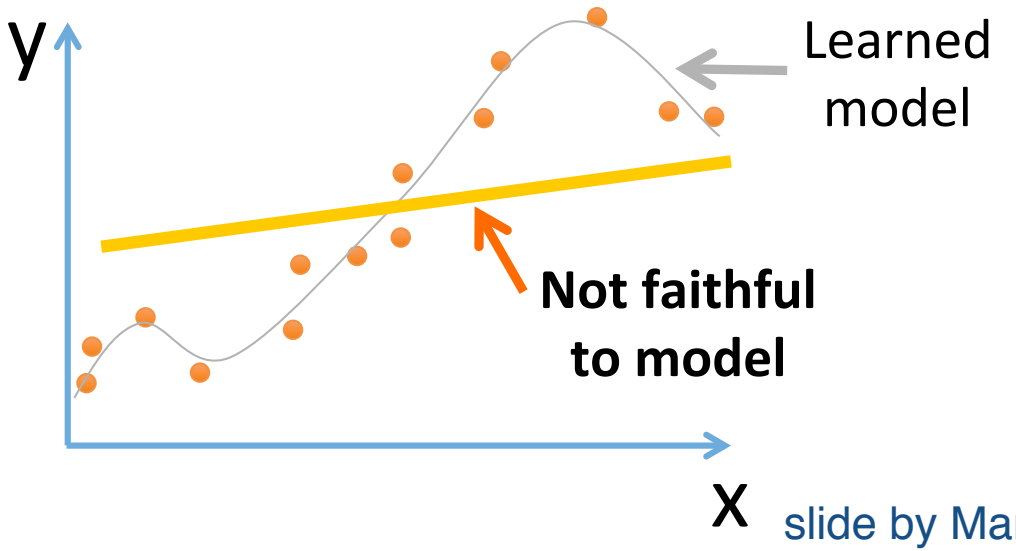
LIME: Local explanations of classifiers

[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

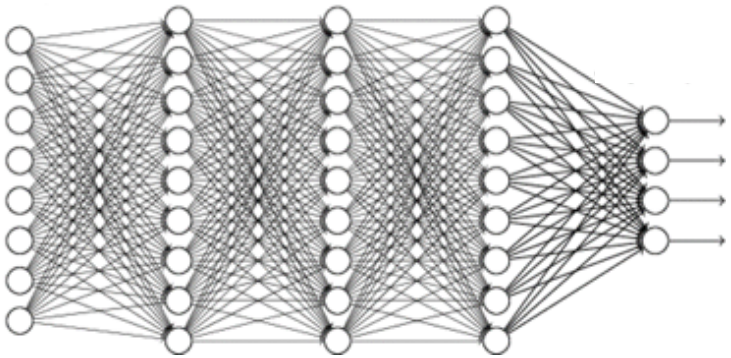
[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

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."

$$f : \mathbb{R}^d \rightarrow \mathbb{R} \qquad g \in G, dom(g) = \{0,1\}^{d'} \qquad \Omega(g)$$

model being explained (a classifier)

explanation

some class of interpretable models

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***

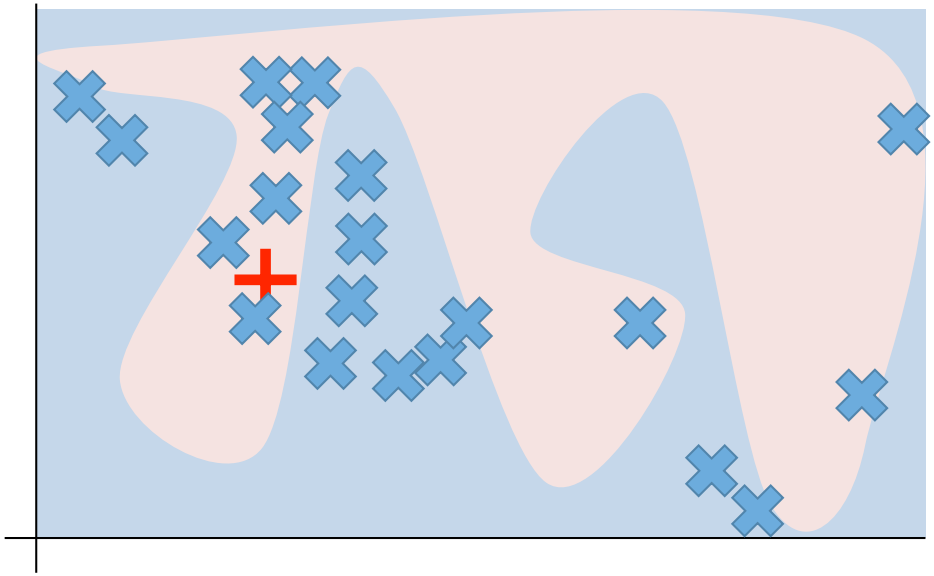
$$\xi(x) = \text{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

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 classier."

1. sample points around +



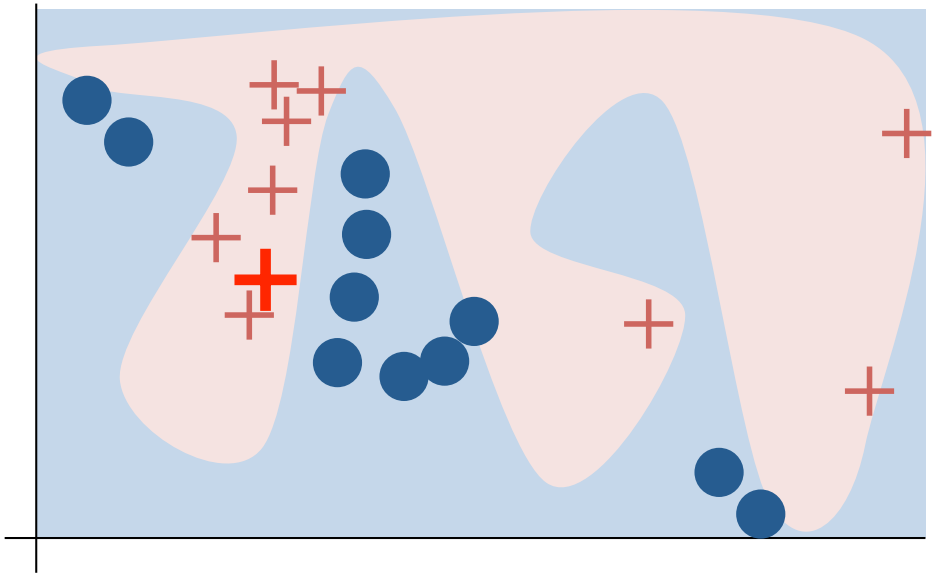
based on a slide by Marco Tulio Ribeiro, KDD 2016

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



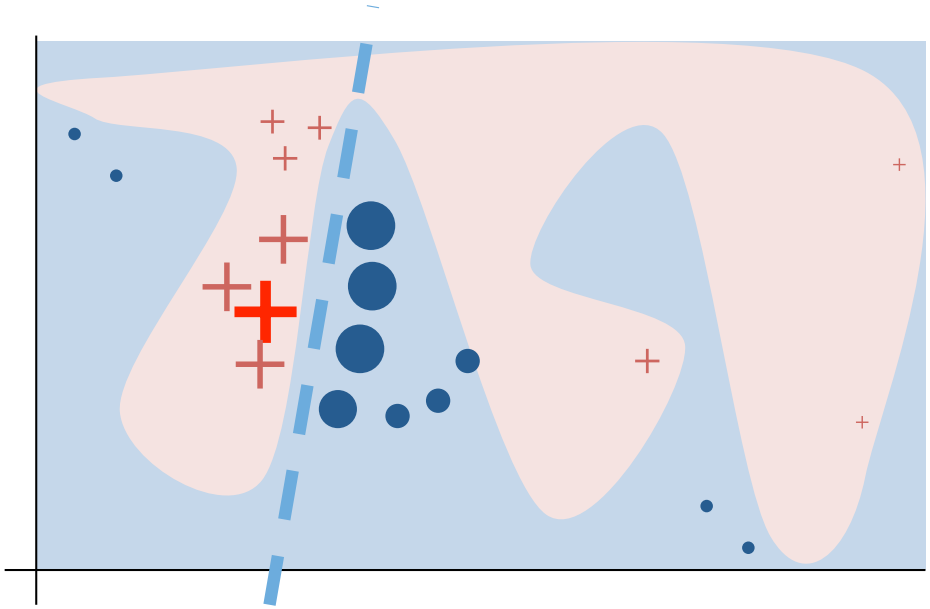
based on a slide by Marco Tulio Ribeiro, KDD 2016

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 classier."

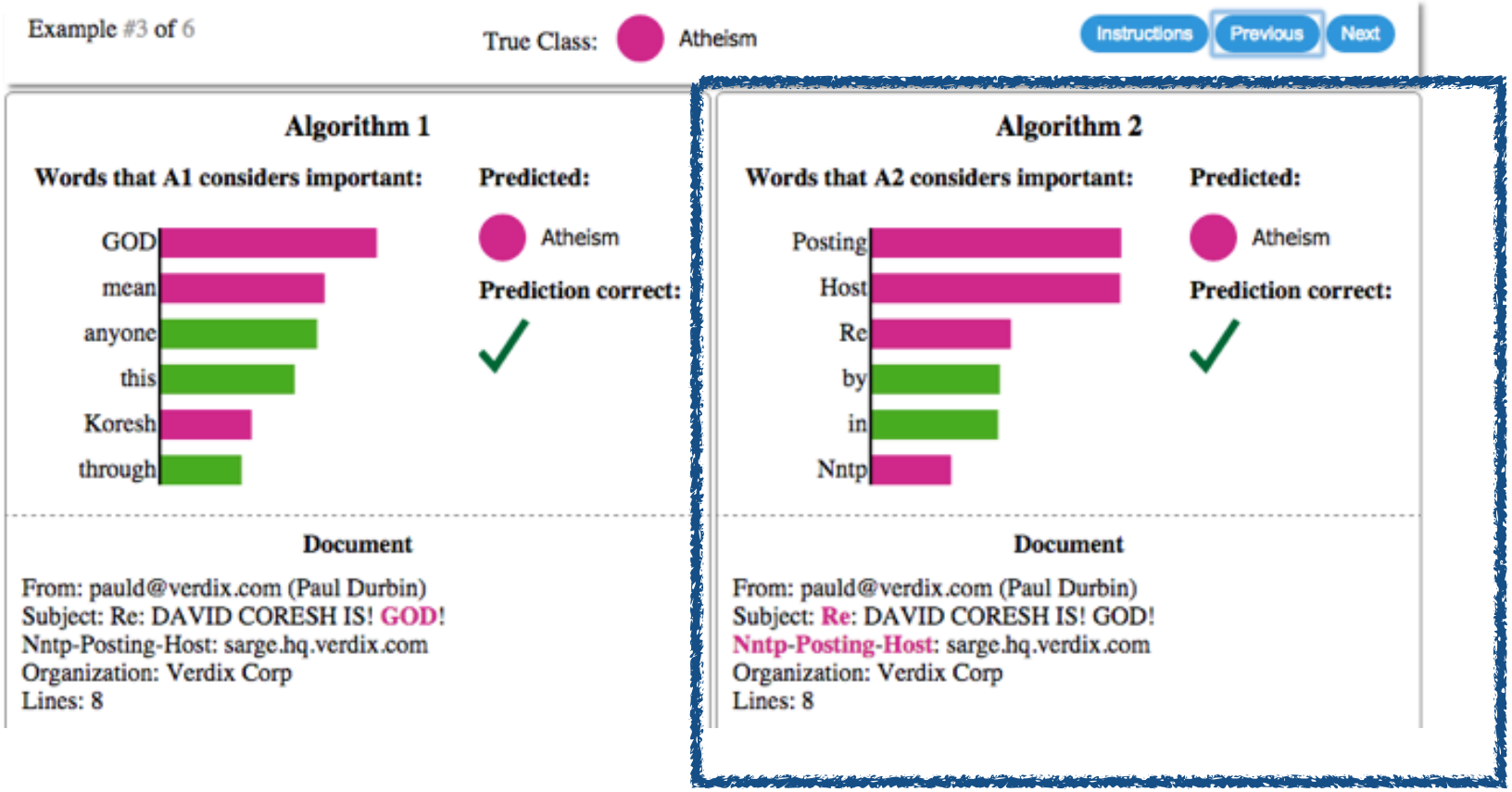
1. sample points around +
2. use complex model ***f*** to assign class labels
3. weigh samples according to π_x
4. learn simple model ***g*** according to samples



based on a slide by Marco Tulio Ribeiro, KDD 2016

Example: text classification with SVMs

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



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

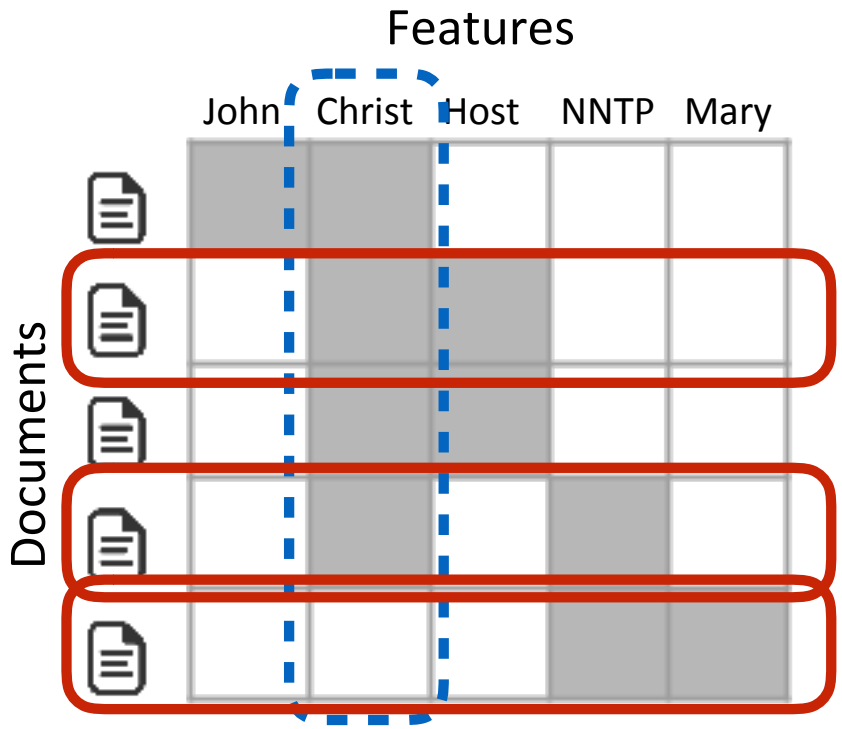
Given a budget **B** of explanations that a user is willing to consider, **pick** a set of ***B*** representative instances for the user to inspect

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

Only 1 mistake!!!

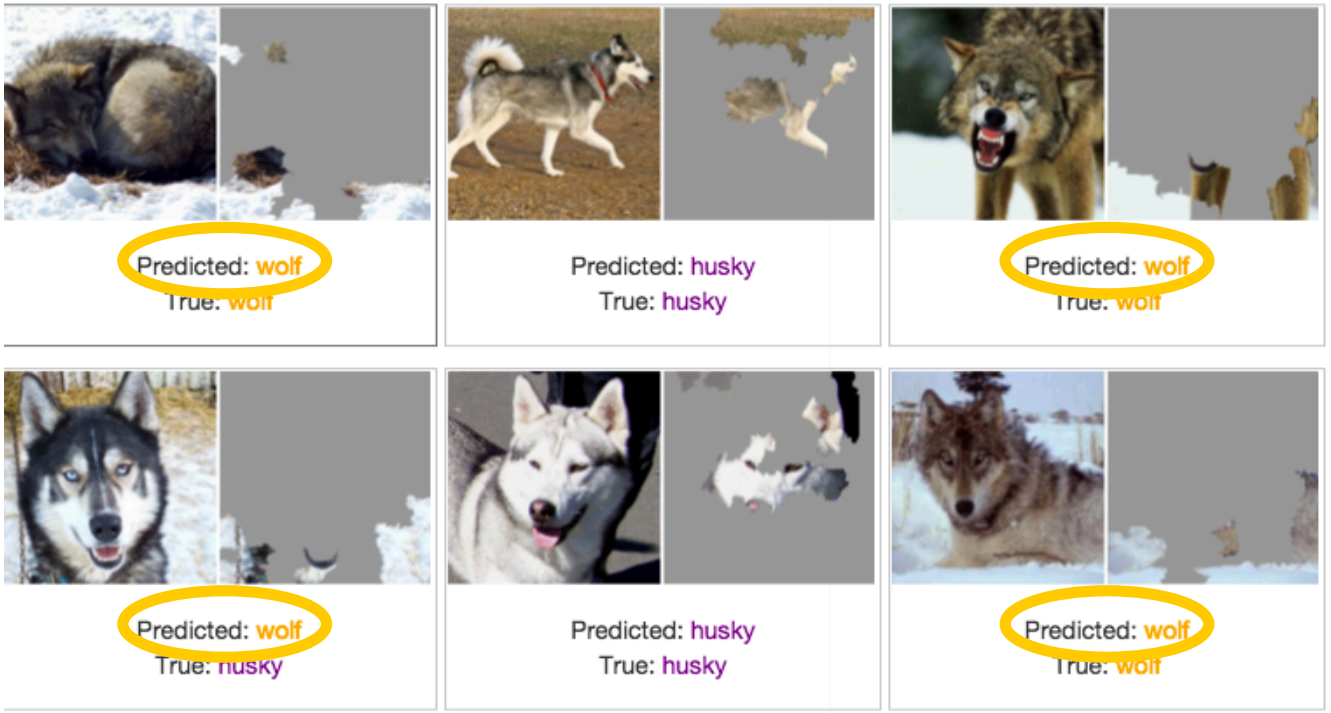
Do you trust this model?
How does it distinguish between huskies and wolves?

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

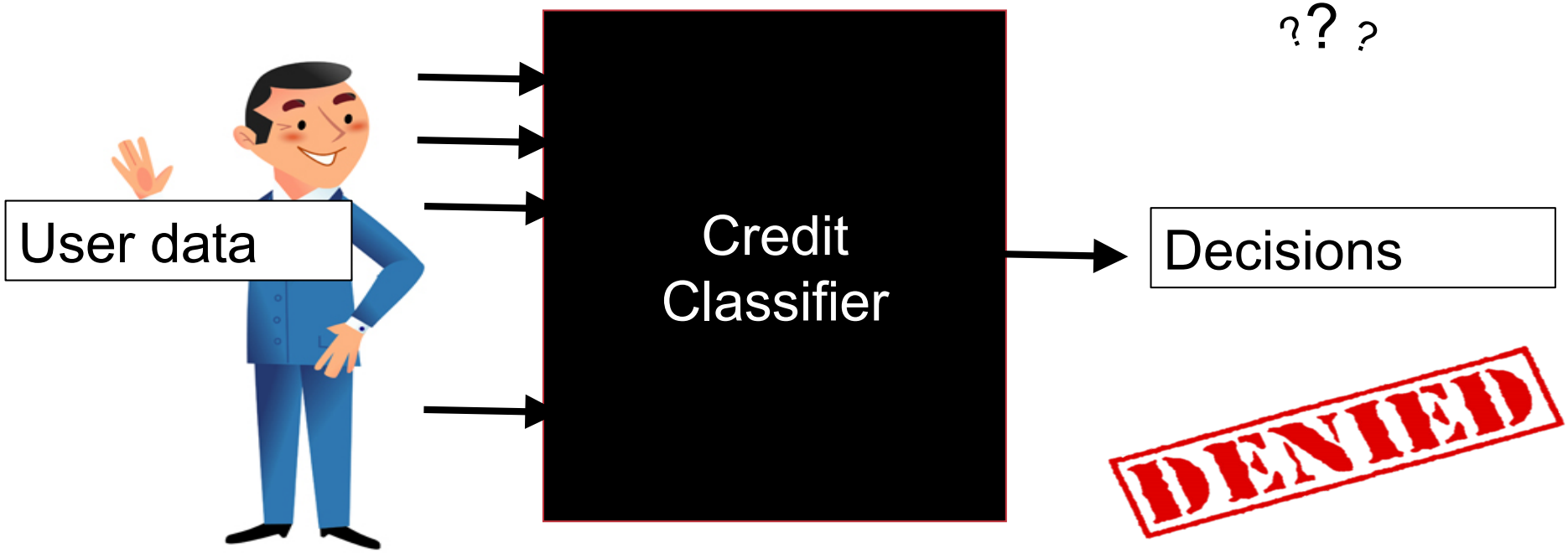
Explaining black-box classifiers

Algorithmic transparency with quantitative input influence (QII)

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

Another method: Quantitative Input Influence

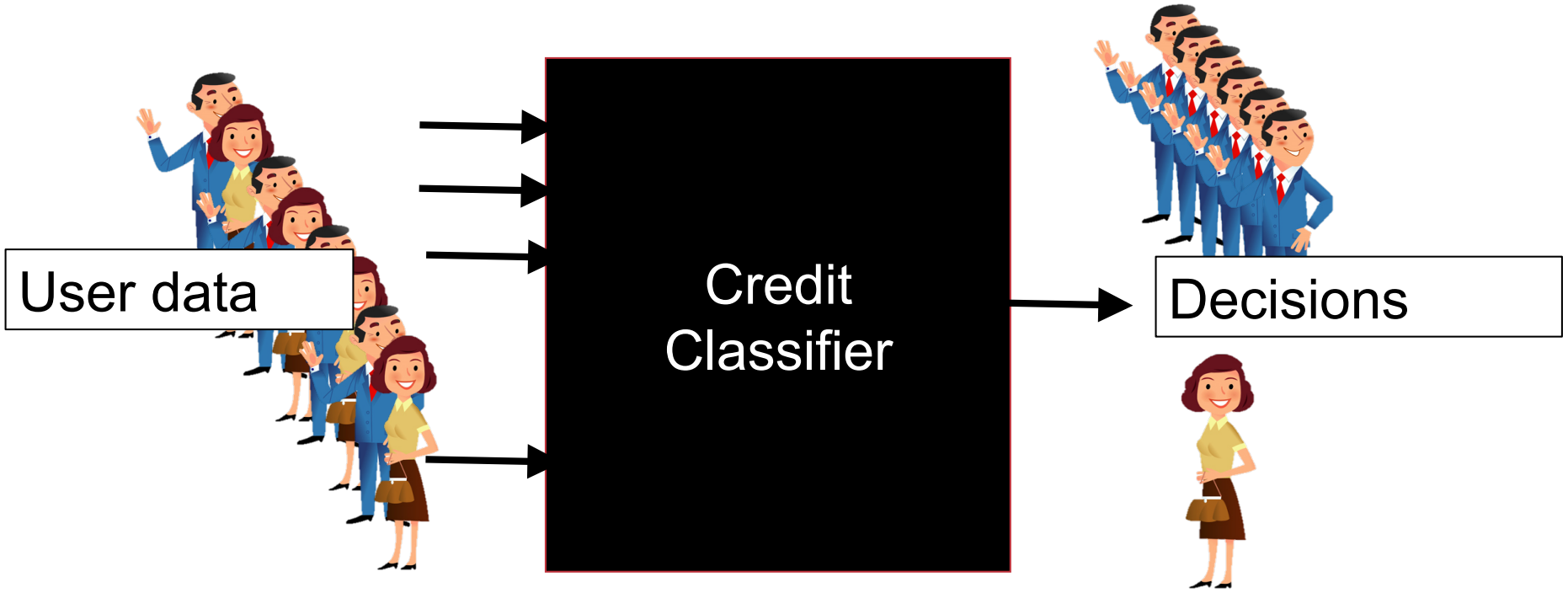
[A. Datta, S. Sen, Y. Zick; *SP 2016*]



slide by A. Datta

Auditing black-box models

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



slide by A. Datta

Influence of inputs on outcomes

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

QII: quantitative input influence framework

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

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

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

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?

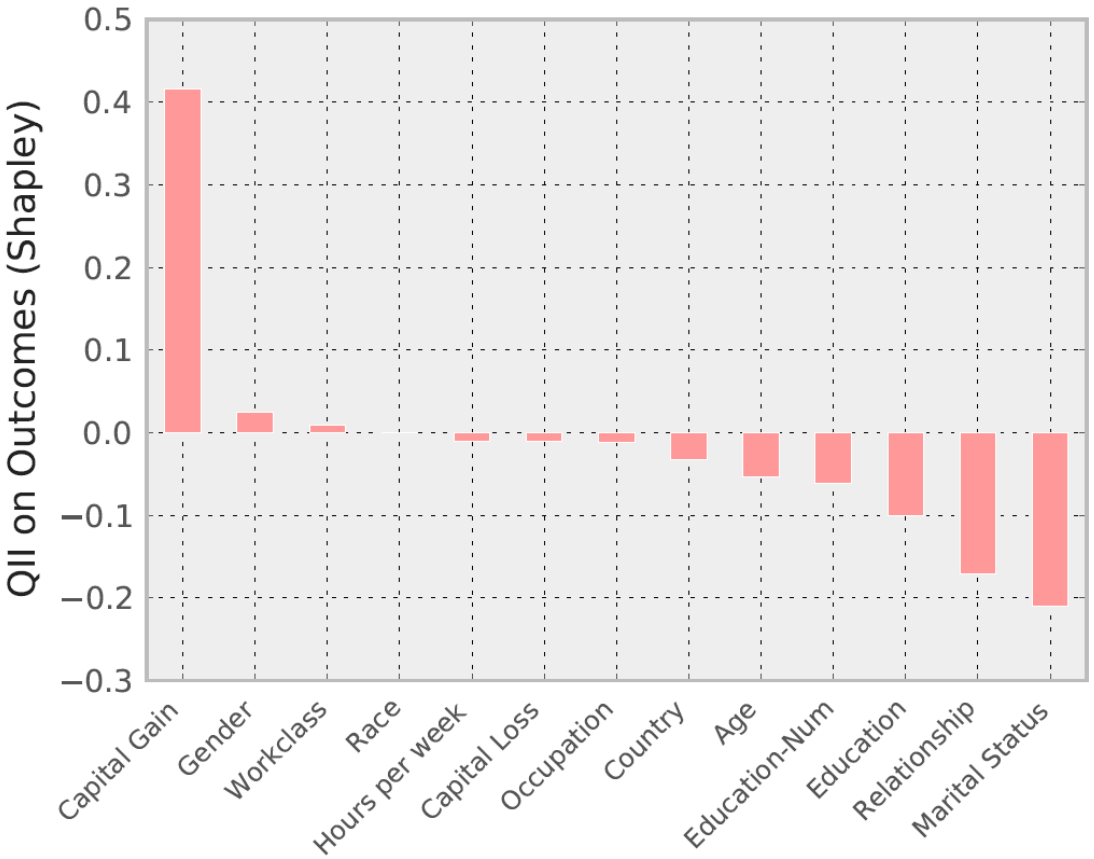
Group: Which inputs have most influence on credit decisions for women?

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

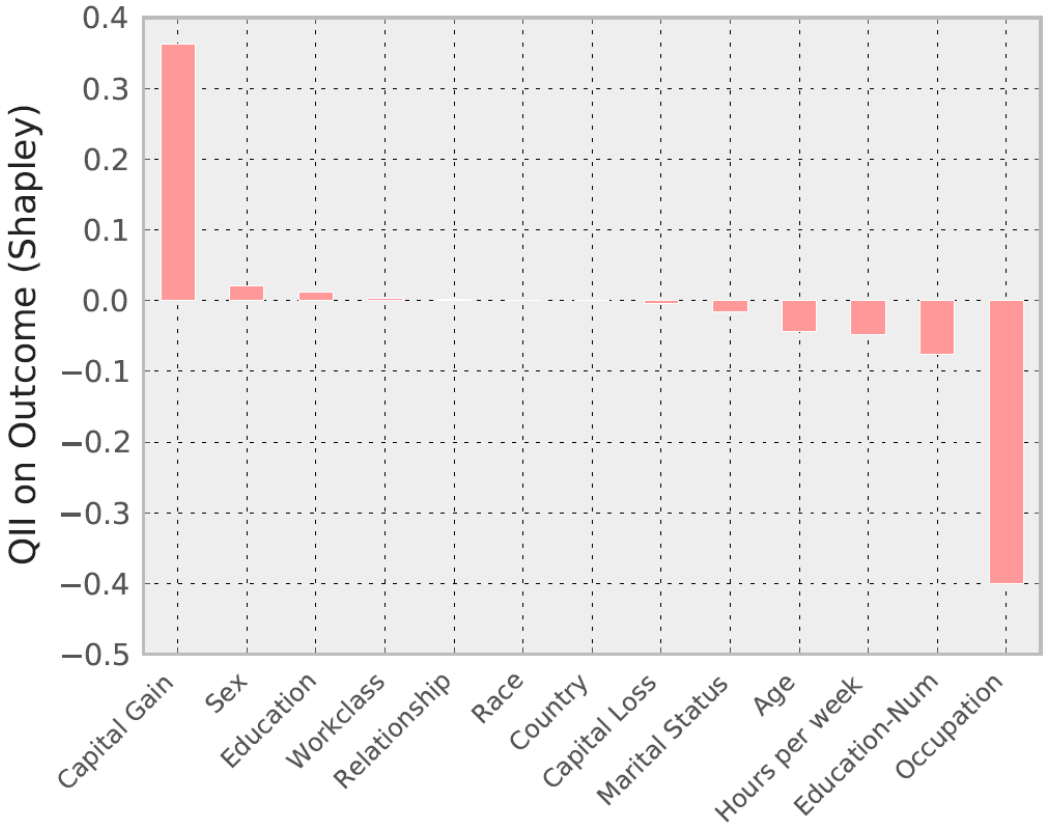
Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344
Capital loss	$0
Work hours per week	40
Country	Vietnam

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

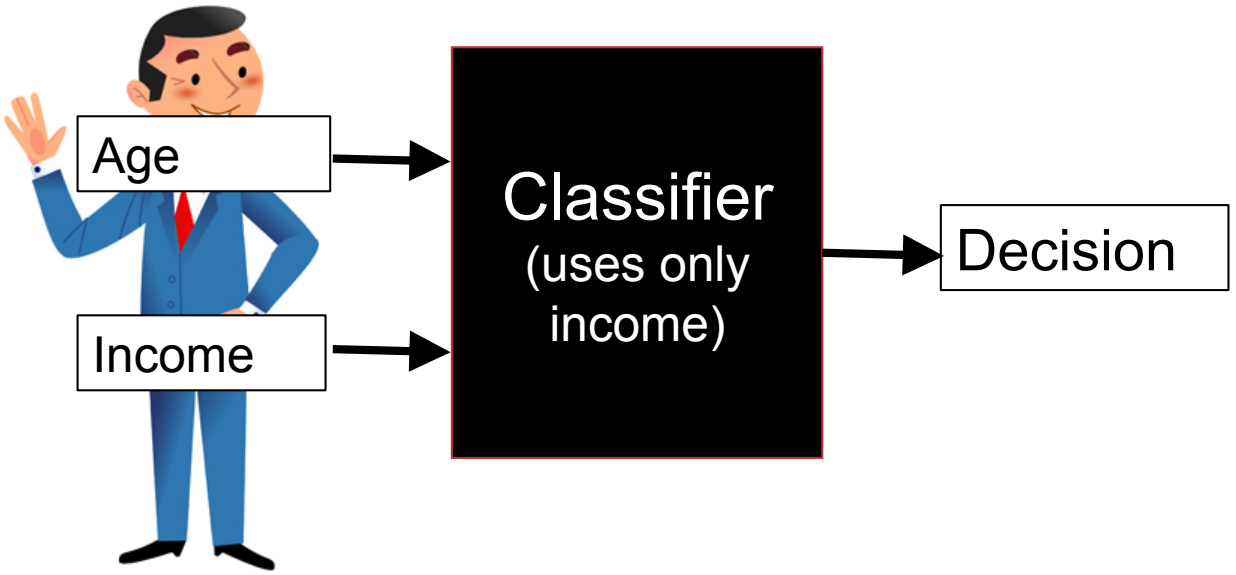
explanations for superficially similar individuals can be different

slide by A. Datta

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**.

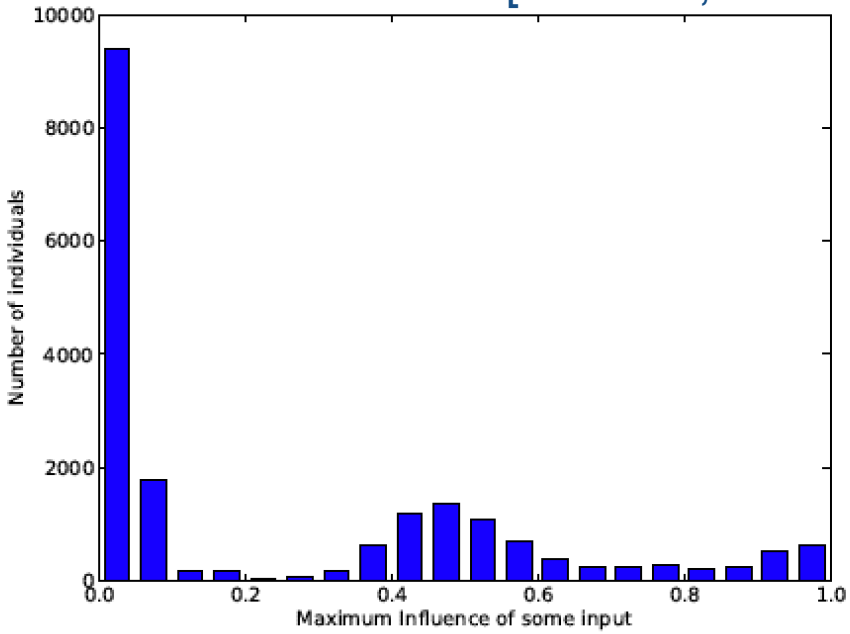


replace features with random values from the population, examine the distribution over outcomes

slide by A. Datta

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

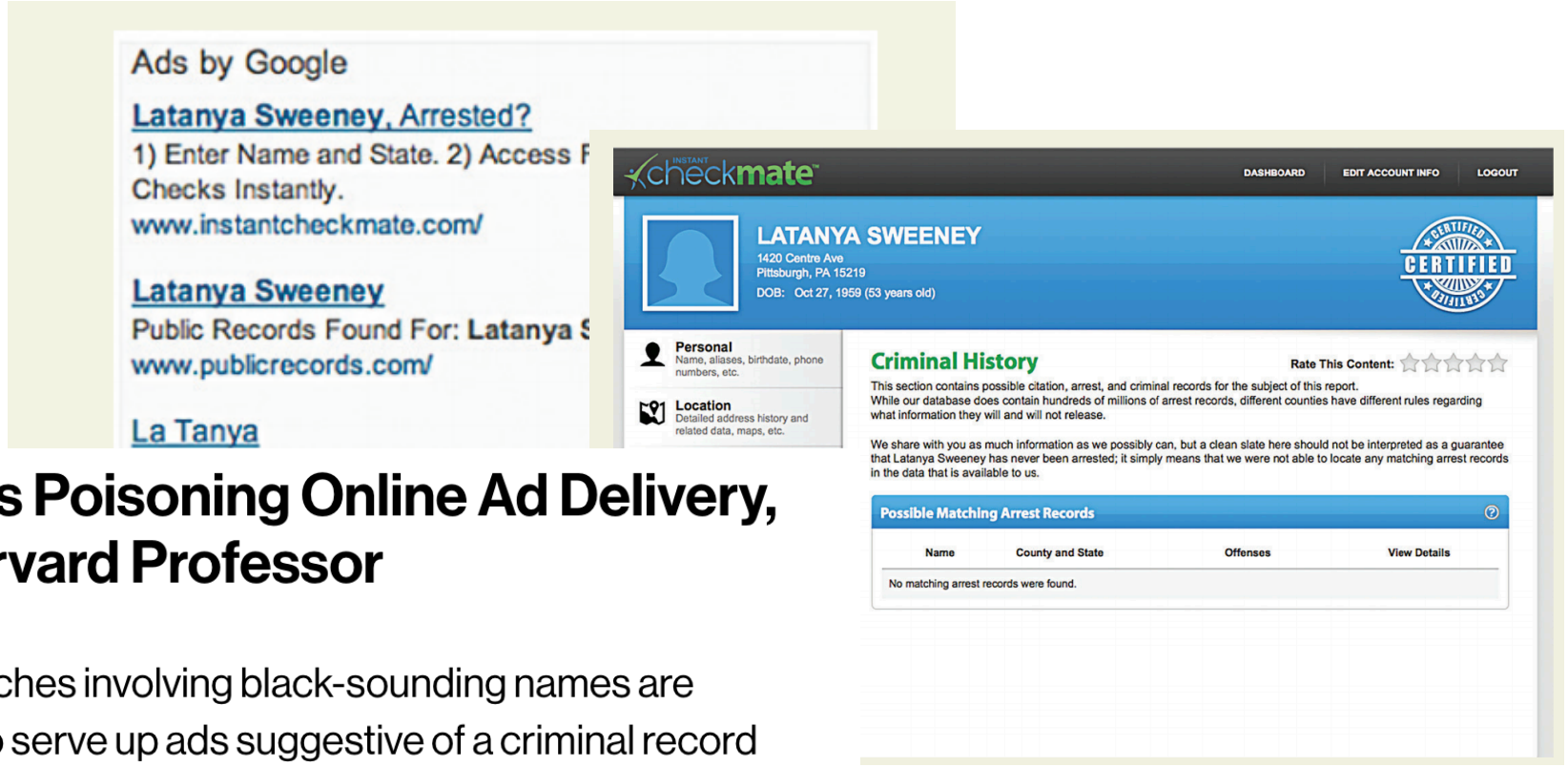
Online ad delivery

Transparency themes

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

Racially identifying names

[Latanya Sweeney; *CACM 2013*]



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

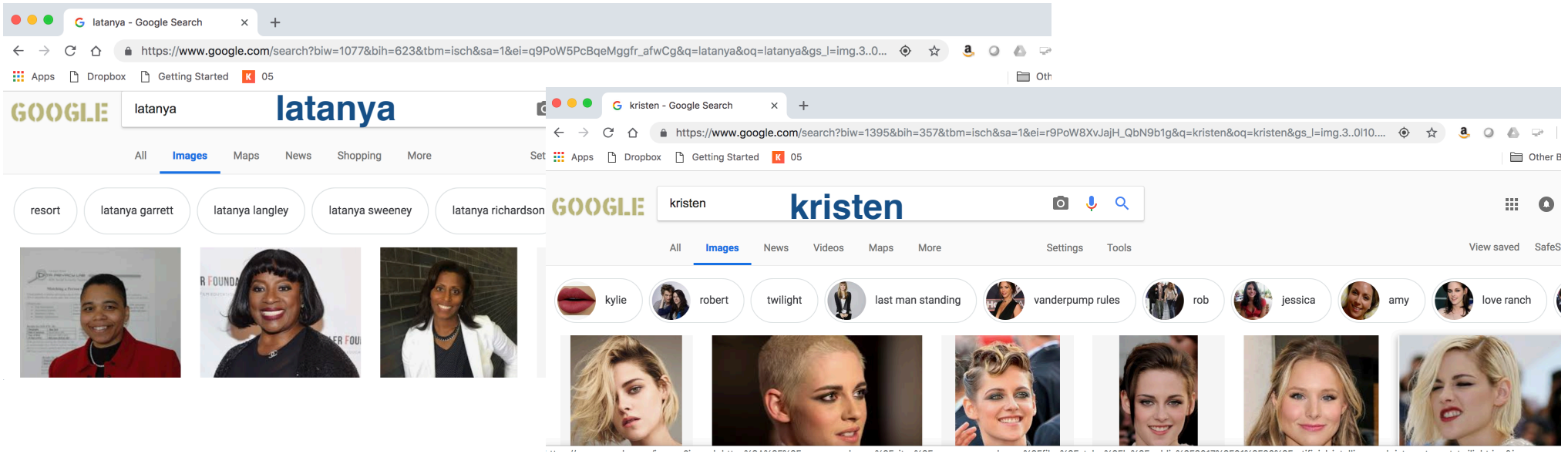
racially identifying names trigger ads suggestive of a criminal record

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



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

Instantcheckmate.com sends the following statement:

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

Who is responsible?

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

transparency …. responsibility …. trust

Online job ads

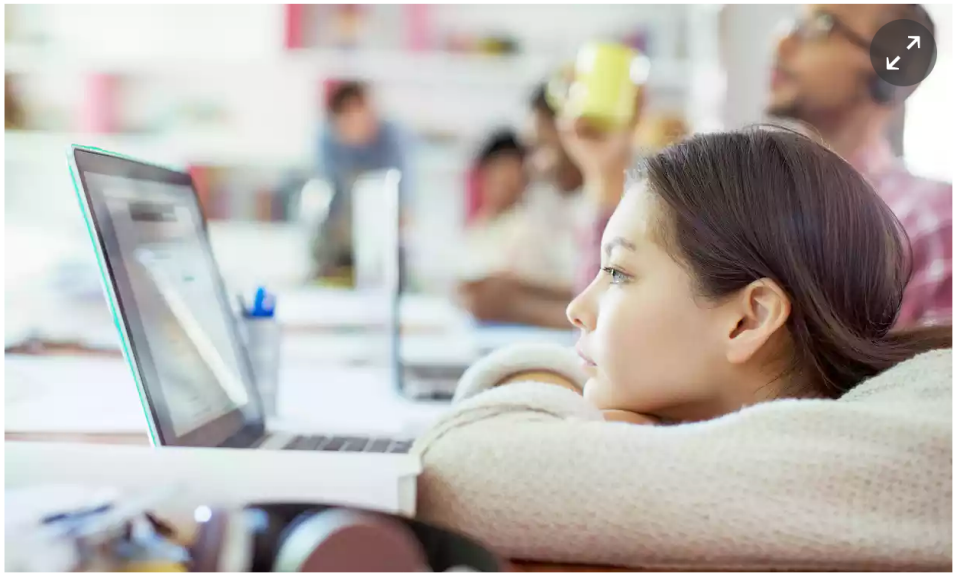
theguardian

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.

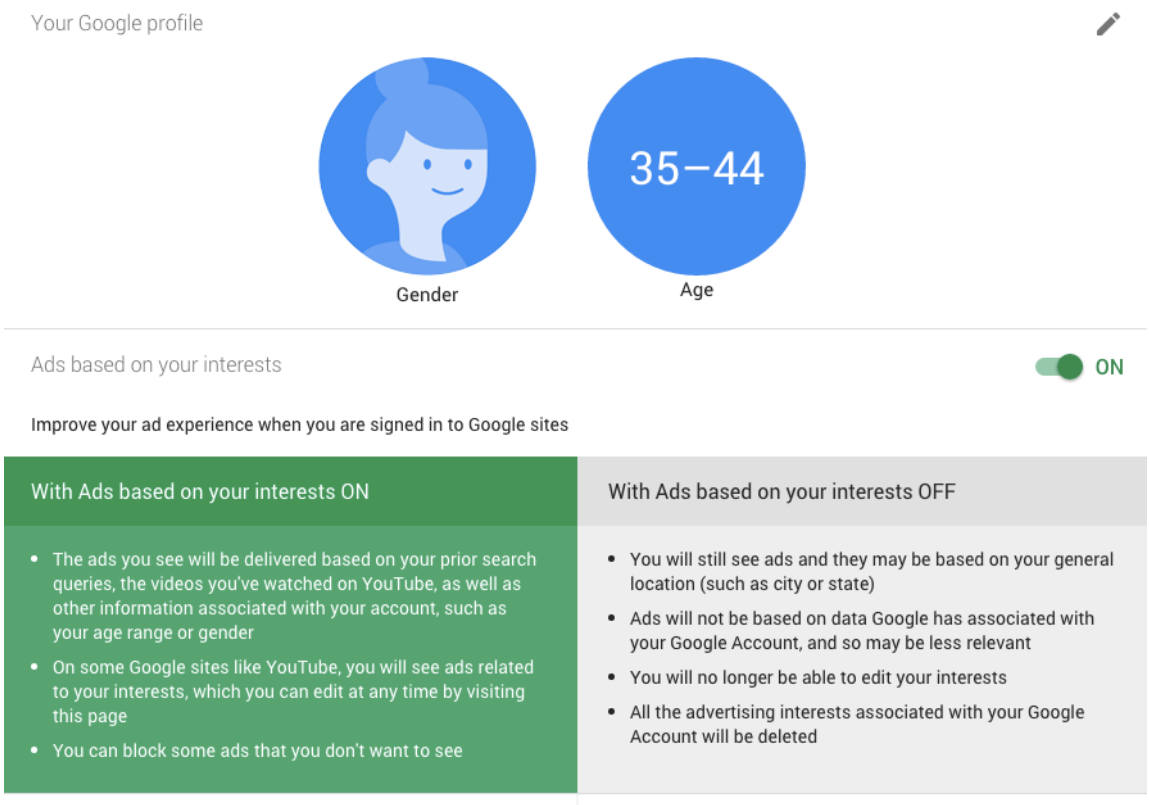
https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

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

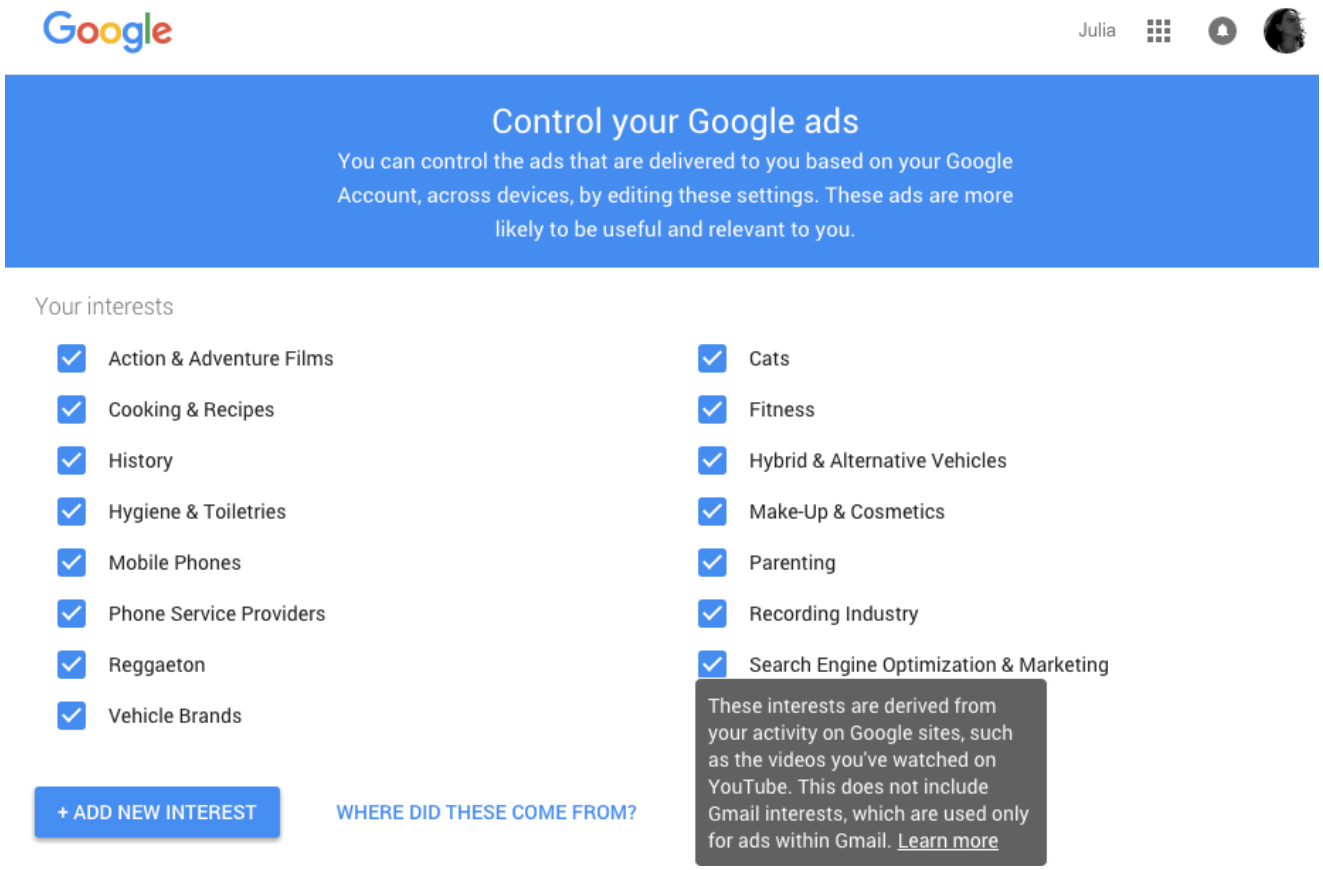
Google ad settings aims to provide **transparency** / give **control to users** over the ads that they see



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

Google ad settings

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



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

AdFisher

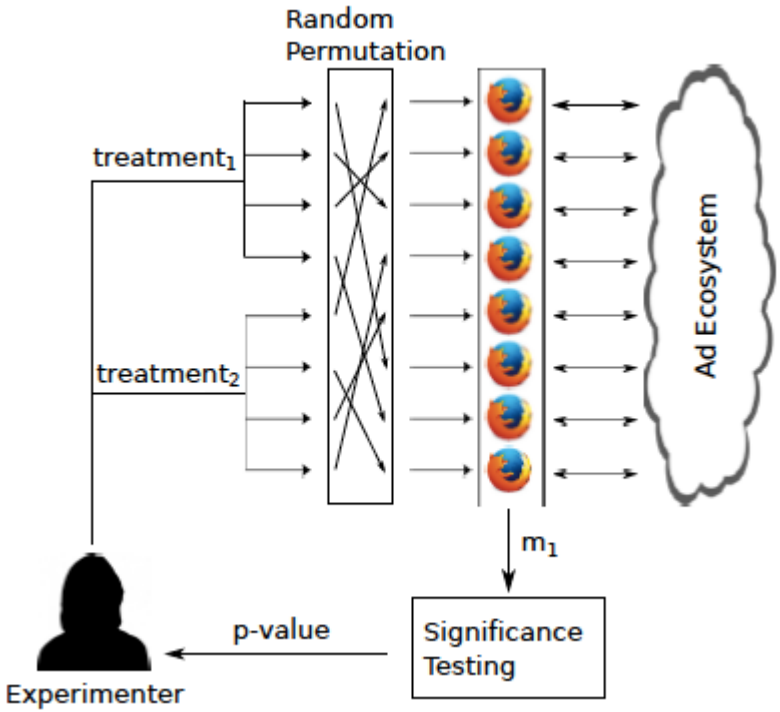
[A. Datta, M. Tschantz, A. Datta; *PETS 2015*]

From anecdotal evidence to statistical insight:

How do user behaviors, ads and ad settings interact?

Automated randomized controlled experiments for studying online tracking

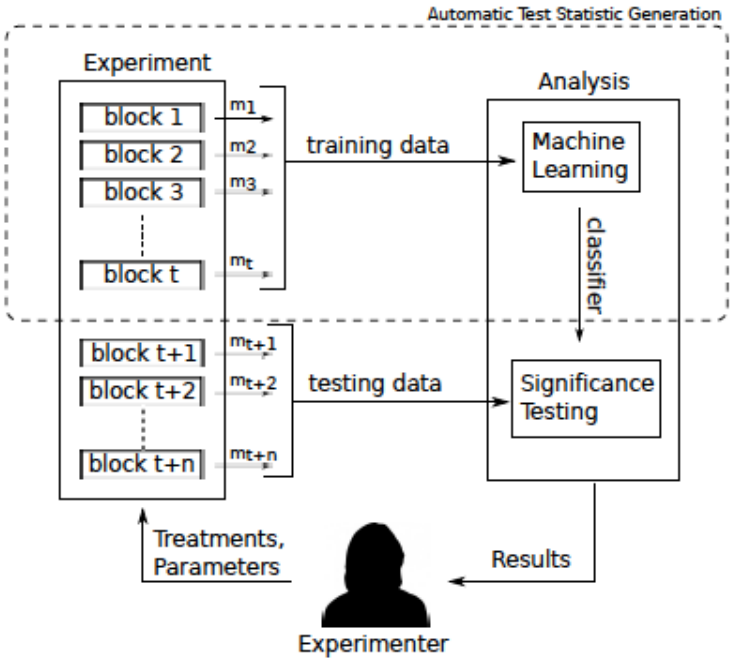
Individual data use transparency: ad network must share the information it uses about the user to select which ads to serve to him



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*]

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

- on gender directly
- on a proxy of gender, i.e., on a known correlate pf 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

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

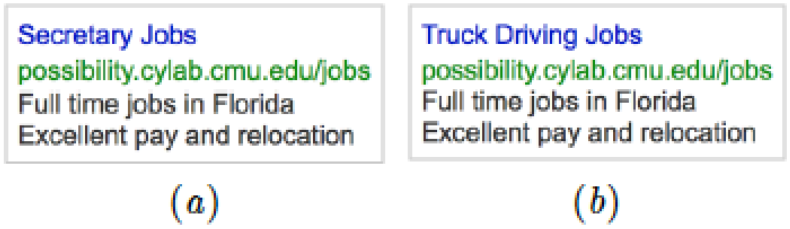


Figure 1: Ads approved by Google in 2015. The ad in the left (right) column was targeted to women (men).

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

What are the legal ramifications?

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

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

In the news

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html

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

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

In the news

POLICY | US & WORLD | TECH

THE VERGE

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

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

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

Interpretability

Transparency in ranking

Input: database of items (individuals, colleges, cars, …)

Score-based ranker: computes the score of each item using a **known formula**, e.g., monotone aggregation, then sorts items on score

Output: permutation of the items (complete or top-k)

Do we have transparency?

We have syntactic transparency, but lack interpretability!

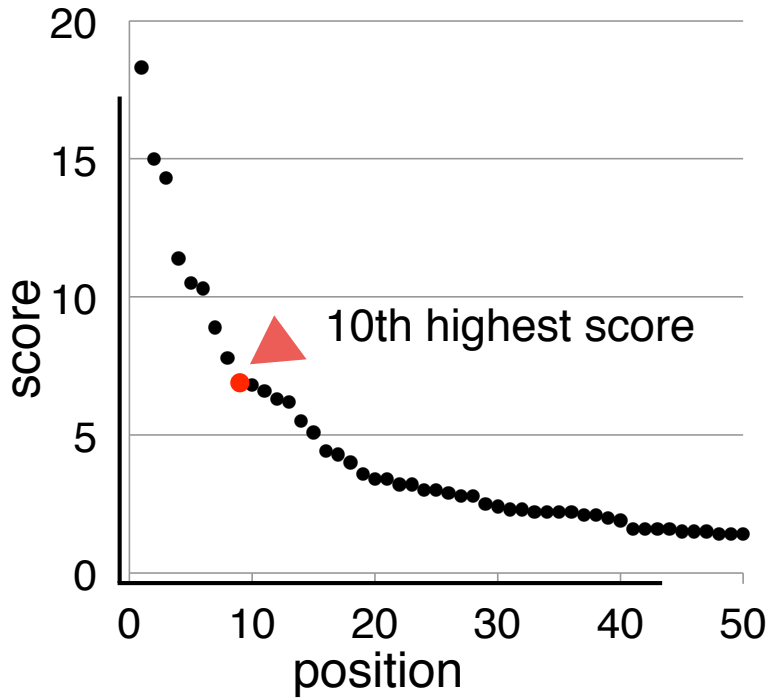
https://freedom-to-tinker.com/2018/05/03/refining-the-concept-of-a-nutritional-label-for-data-and-models/

https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

Opacity in algorithmic rankers

Reason 1: The scoring formula alone does not indicate the relative rank of an item.

Scores are absolute, rankings are relative. Is 5 a good score? What about 10? 15?



Opacity in algorithmic rankers

Reason 2: A ranking may be unstable if there are tied or nearly-tied items.

Rank	Institution	Average Count	Faculty
1	▶ Carnegie Mellon University	18.4	123
2	▶ Massachusetts Institute of Technology	15.6	64
3	▶ Stanford University	14.8	56
4	▶ University of California - Berkeley	11.5	50
5	▶ University of Illinois at Urbana-Champaign	10.6	56
6	▶ University of Washington	10.3	50
7	▶ Georgia Institute of Technology	8.9	81
8	▶ University of California - San Diego	8	51
9	▶ Cornell University	7	45
10	▶ University of Michigan	6.8	63
11	▶ University of Texas - Austin	6.6	43
12	▶ University of Massachusetts - Amherst	6.4	47

Opacity in algorithmic rankers

Reason 3: A ranking methodology may be unstable: small changes in weights can trigger significant re-shuffling.

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

1. Chevrolet Corvette 205
2. Lotus Evora 195
3. Porsche Cayman 195

1. Lotus Evora 205
2. Porsche Cayman 198
3. Chevrolet Corvette 192

1. Porsche Cayman 193
2. Chevrolet Corvette 186
3. Lotus Evora 182

Opacity in algorithmic rankers

Reason 4: The **weight of an attribute** in the scoring formula **does not determine its impact** on the outcome.

Rank	Name	Avg Count	Faculty	Pubs	GRE
1	CMU	18.3	122	2	791
2	MIT	15	64	3	772
3	Stanford	14.3	55	5	800
4	UC Berkeley	11.4	50	3	789
5	UIUC	10.5	55	3	772
6	UW	10.3	50	2	796
....					
39	U Chicago	2	28	2	779
40	UC Irvine	1.9	28	2	787
41	BU	1.6	15	2	783
41	U Colorado Boulder	1.6	32	1	761
41	UNC Chapel Hill	1.6	22	2	794
41	Dartmouth	1.6	18	2	794

Given a score function:

$$0.2 * faculty +$$

$$0.3 * avg\ cnt +$$

$$0.5 * gre$$

Rankings are not benign!

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

Rankings are not benign. They enshrine very particular ideologies, and, at a time when American higher education is facing a crisis of accessibility and affordability, we have adopted a de-facto standard of college quality that is uninterested in both of those factors. And why? Because a group of magazine analysts in an office building in Washington, D.C., decided twenty years ago to value selectivity over efficacy, to use proxies that scarcely relate to what they're meant to be proxies for, and to pretend that they can compare a large, diverse, low-cost land-grant university in rural Pennsylvania with a small, expensive, private Jewish university on two campuses in Manhattan.

Harms of opacity

1. Due process / fairness. The subjects of the ranking cannot have confidence that their ranking is meaningful or correct, or that they have been treated like similarly situated subjects - *procedural regularity*

2. Hidden normative commitments. What factors does the vendor encode in the scoring ranking process (syntactically)? What are the ***actual*** effects of the scoring / ranking process? Is it stable? How was it validated?

Harms of opacity

3. Interpretability. Especially where ranking algorithms are performing a public function, **political legitimacy** requires that the public be able to interpret algorithmic outcomes in a meaningful way. Avoid *algocracy*: the rule by incontestable algorithms.

4. Meta-methodological assessment. Is *a* ranking / *this* ranking appropriate here? Can we use a process if it cannot be explained? Probably yes, for recommending movies; probably not for college admissions.

“Nutritional labels” for data and models

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

Ranking Facts

Recipe

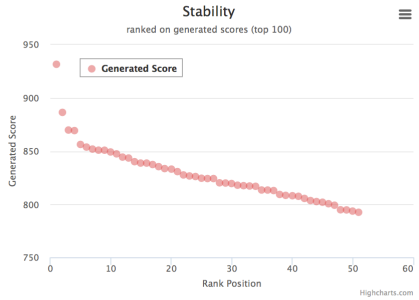
Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
Faculty	122	52.5	45
GRE	800.0	796.3	771.9

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
Faculty	122	32.0	14
GRE	800.0	790.0	757.8

Stability



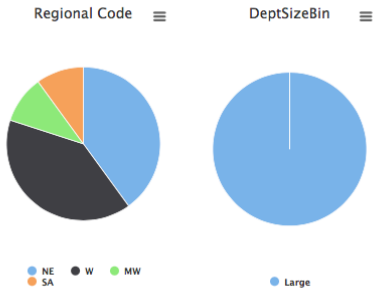
Slope at top-10: -6.91. Slope overall: -1.61.

Unstable when absolute value of slope of fit line in scatter plot <= 0.25 (slope threshold). Otherwise it is stable.

Recipe

Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Diversity at top-10



Stability

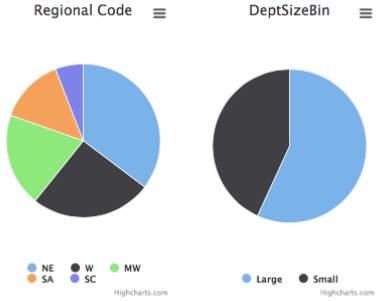
Top-K	Stability
Top-10	Stable
Overall	Stable

Ingredients

Attribute	Correlation
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Correlation strength is based on its absolute value. Correlation over 0.75 is high, between 0.25 and 0.75 is medium, under 0.25 is low.

Diversity overall



Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

Unfair when p-value of corresponding statistical test <= 0.05.

Ingredients

Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
CSRankingAllArea	13	6.5	1
Faculty	122	52.5	45

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
CSRankingAllArea	48	26.0	1
Faculty	122	32.0	14

Fairness

	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	α
Large	1.0	0.87	0.99	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

Top K = 26 in FA*IR and Proportion oracles. Setting of top K: In FA*IR and Proportion oracle, if N > 200, set top K =100. Otherwise set top K = 50%N. Pairwise oracle takes whole ranking as input. FA*IR is computed as using code in FA*IR codes. Proportion is implemented as statistical test 4.1.3 in Proportion paper.

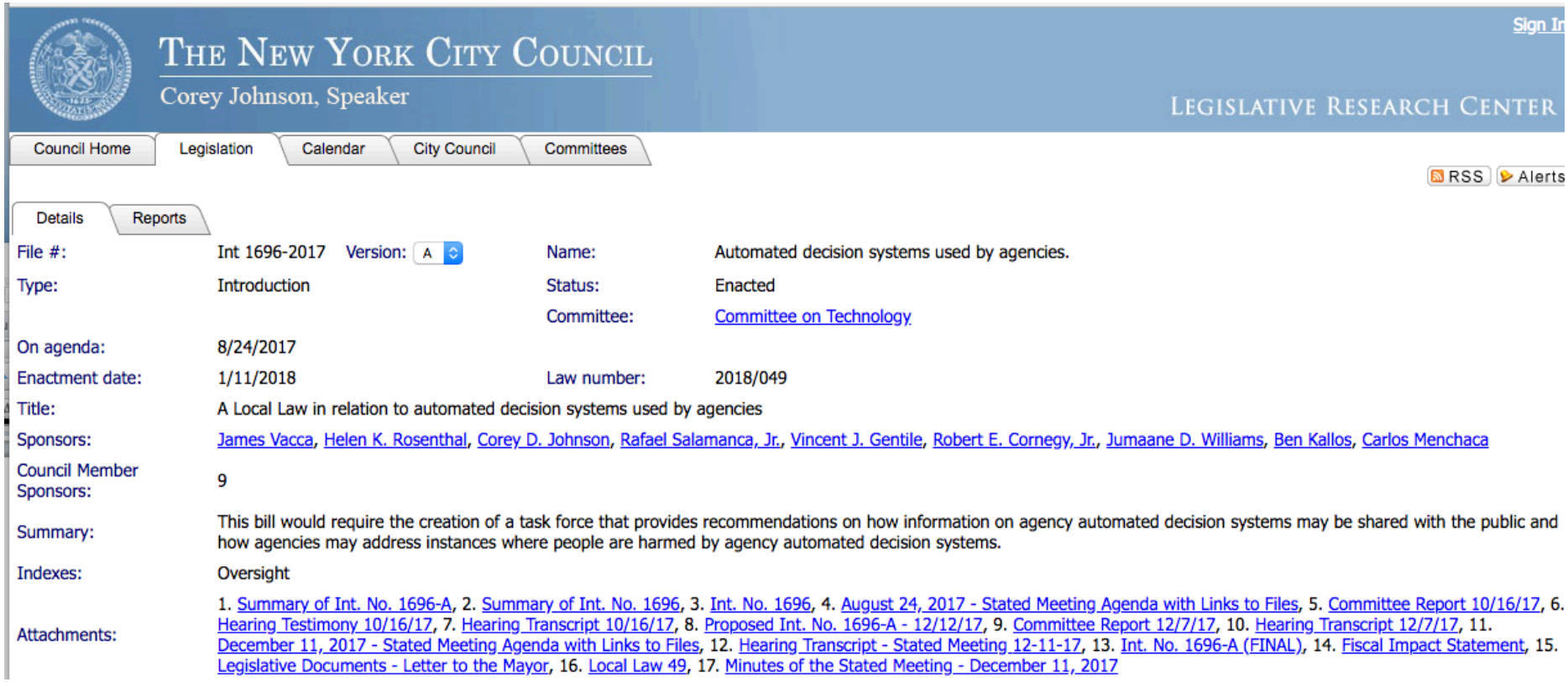
http://demo.dataresponsibly.com/rankingfacts/nutrition_facts/

an (ongoing) attempt at regulation

NYC ADS transparency law

1/11/2018

Local Law 49 of 2018 in relation to automated decision systems used by agencies



THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker
LEGISLATIVE RESEARCH CENTER
Sign In

Council Home | Legislation | Calendar | City Council | Committees

RSS Alerts

Details | Reports

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

The original draft

8/16/2017

Int. No. 1696

By Council Member Vacca

A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons

Be it enacted by the Council as follows:

Section 1. Section 23-502 of the administrative code of the city of New York is amended to add a new subdivision g to read as follows:

g. Each agency that uses, for the purposes of targeting services to persons, imposing penalties upon persons or policing, an algorithm or any other method of automated processing system of data shall:

1. Publish on such agency's website, the source code of such system; and

2. Permit a user to (i) submit data into such system for self-testing and (ii) receive the results of having such data processed by such system.

§ 2. This local law takes effect 120 days after it becomes law.

MAJ
LS# 10948
8/16/17 2:13 PM

this is **NOT** what was adopted

Summary of Local Law 49

1/11/2018

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

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

Point 1

algorithmic transparency is not synonymous with releasing the source code

publishing source code helps, but it is sometimes unnecessary and often insufficient

Point 2

algorithmic transparency requires data transparency

data is used in training, validation, deployment

validity, accuracy, applicability can only be understood in the data context

data transparency is necessary for all ADS, not only for ML-based systems

Point 3

data transparency is not synonymous with making all data public

release data whenever possible;

also release:

data selection, collection and pre-processing methodologies; data provenance and quality information; known sources of bias; privacy-preserving statistical summaries of the data

Point 4

actionable transparency requires interpretability

explain assumptions and effects, not details of operation

engage the public - technical and non-technical

Point 5

transparency by design, not as an afterthought

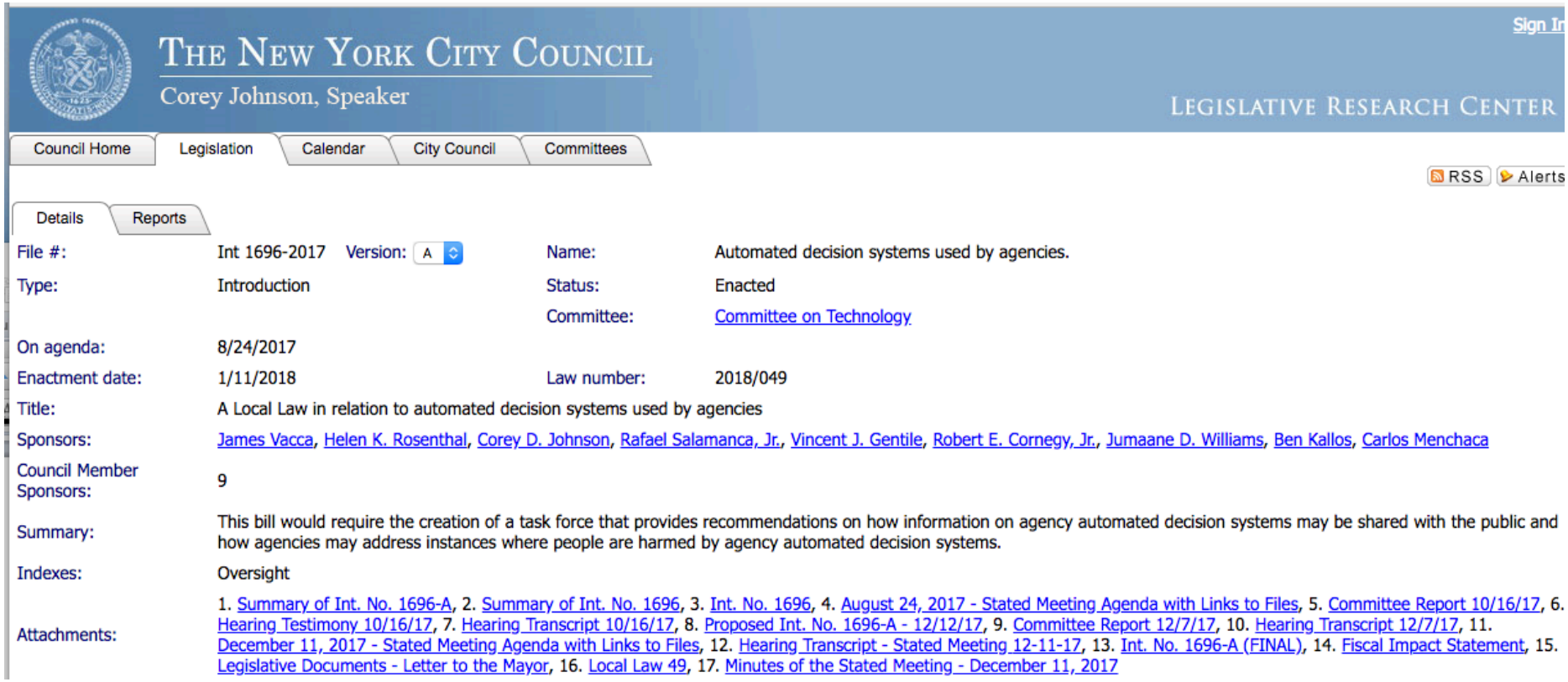
provision for transparency and interpretability at every stage of the data lifecycle

useful internally during development, for communication and coordination between agencies, and for accountability to the public

NYC ADS transparency law

1/11/2018

Local Law 49 of 2018 in relation to automated decision systems used by agencies



THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker
LEGISLATIVE RESEARCH CENTER
Sign In

Council Home | Legislation | Calendar | City Council | Committees

RSS Alerts

Details | Reports

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

The original draft

8/16/2017

Int. No. 1696

By Council Member Vacca

A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons

Be it enacted by the Council as follows:

Section 1. Section 23-502 of the administrative code of the city of New York is amended to add a new subdivision g to read as follows:

g. Each agency that uses, for the purposes of targeting services to persons, imposing penalties upon persons or policing, an algorithm or any other method of automated processing system of data shall:

1. Publish on such agency's website, the source code of such system; and

2. Permit a user to (i) submit data into such system for self-testing and (ii) receive the results of having such data processed by such system.

§ 2. This local law takes effect 120 days after it becomes law.

MAJ
LS# 10948
8/16/17 2:13 PM

this is **NOT** what was adopted

Summary of Local Law 49

1/11/2018

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

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