DS-GA 3001.009: Responsible Data Science

Transparency: Explaining Black-box Models

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

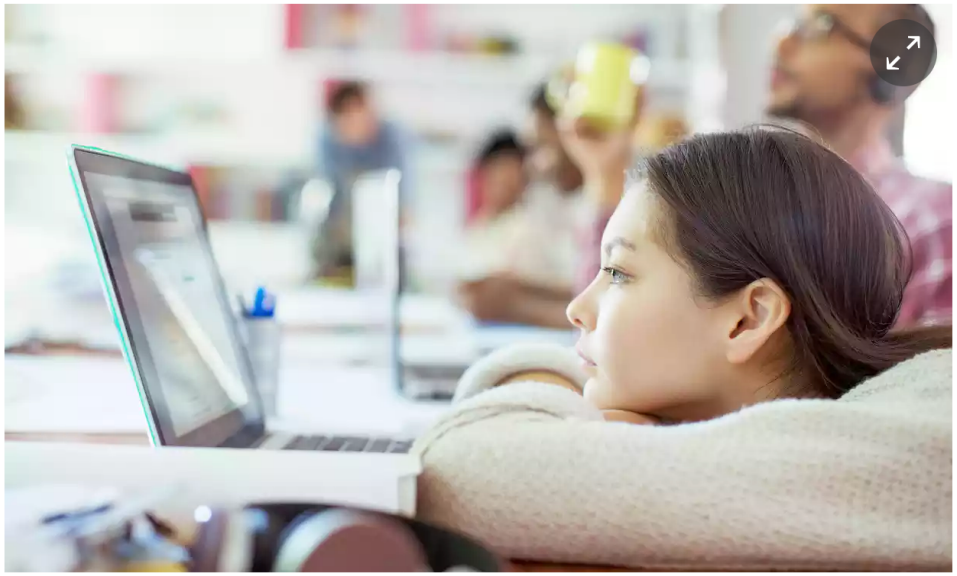
theguardian

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*

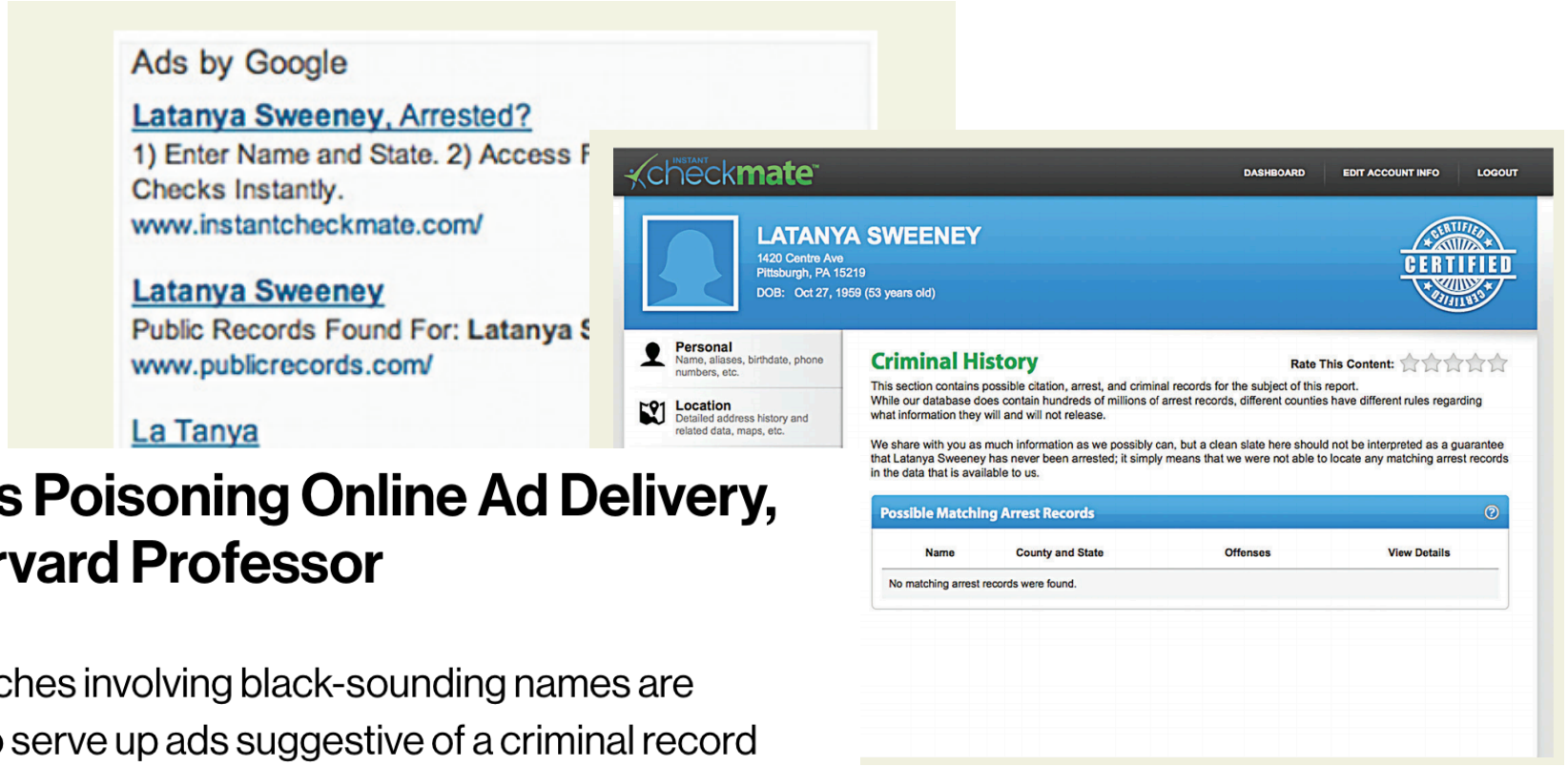
The Equal Employment Opportunity commission is **investigating whether personality tests discriminate against people with disabilities**.

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

Racially identifying names

[Latanya Sweeney; *CACM 2013*]



Google AdSense

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

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

racially identifying names trigger ads suggestive of a criminal record

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

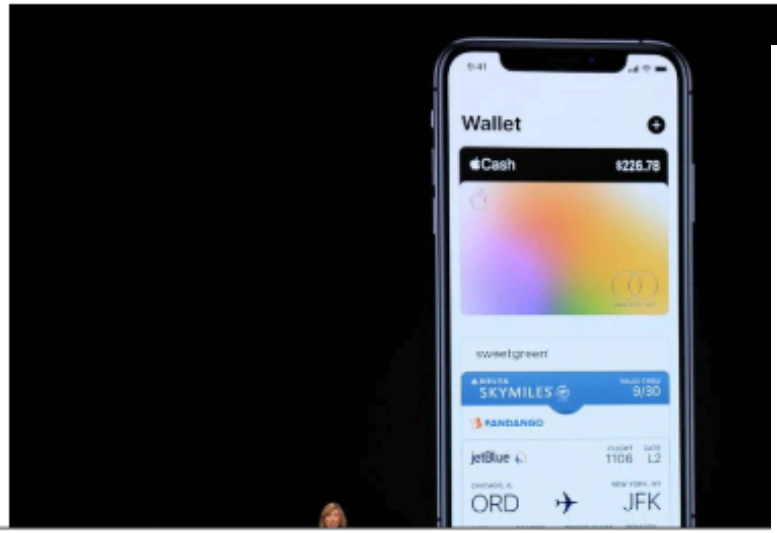
Access to credit

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

November 10, 2019

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.



David Heinemeier Hansson vented on Twitter that even though his spouse, Jamie Hansson, had a better credit score and other factors in her favor, her application for a credit line increase had been denied.

Mr. Hansson, a prominent software developer, wondered how his credit line could be 20 times higher, referring to Apple Card as a "sexist program" (with an expletive added for emphasis).

The card, a partnership between Apple and Goldman Sachs, made its debut in the United States in August.

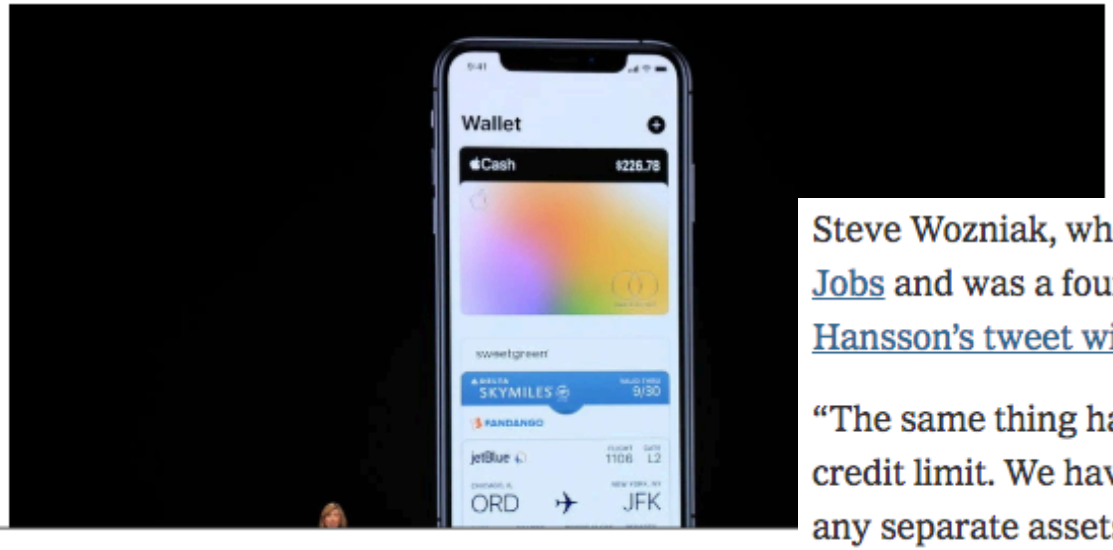
Access to credit

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

November 10, 2019

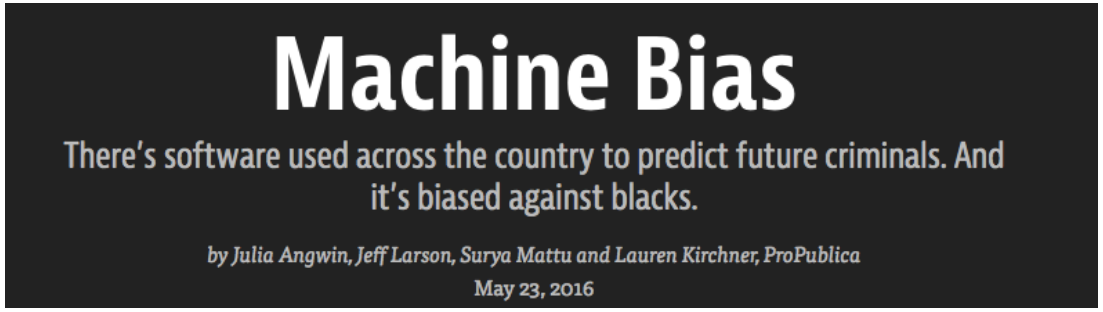
A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.



Steve Wozniak, who invented the Apple-1 computer with Steven P. Jobs and was a founder of the tech giant, responded to Mr. Hansson's tweet with a similar account.

"The same thing happened to us," Mr. Wozniak wrote. "I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019."

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

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]
  - **SHAP**: Shapley additive explanations [Lundberg and Lee, NeurIPS 2017]
- **Online ad targeting**
  - Racially identifying names [Sweeney, CACM 2013]
  - Ad Fisher [Datta et al., PETS 2015]
- **Interpretability**
  - Nutritional labels [Yang et al., SIGMOD 2018]

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**

Interpretability enables trust

- 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?

Detour: Facebook's real-name policy

Shane Creepingbear is a member of the Kiowa Tribe of Oklahoma

October 13, 2014

TIME

Facebook Thinks Some Native American Names Are Inauthentic

BY JOSH SANBURN FEBRUARY 14, 2015

February 14, 2015

If you're Native American, Facebook might think your name is fake.

The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.

The latest flap involves Native Americans, including Dana Lone Hill, who is Lakota. Lone Hill recently wrote in a blog post that Facebook told her her name was not "authentic" when she attempted to log in.

When is accuracy insufficient for trust?

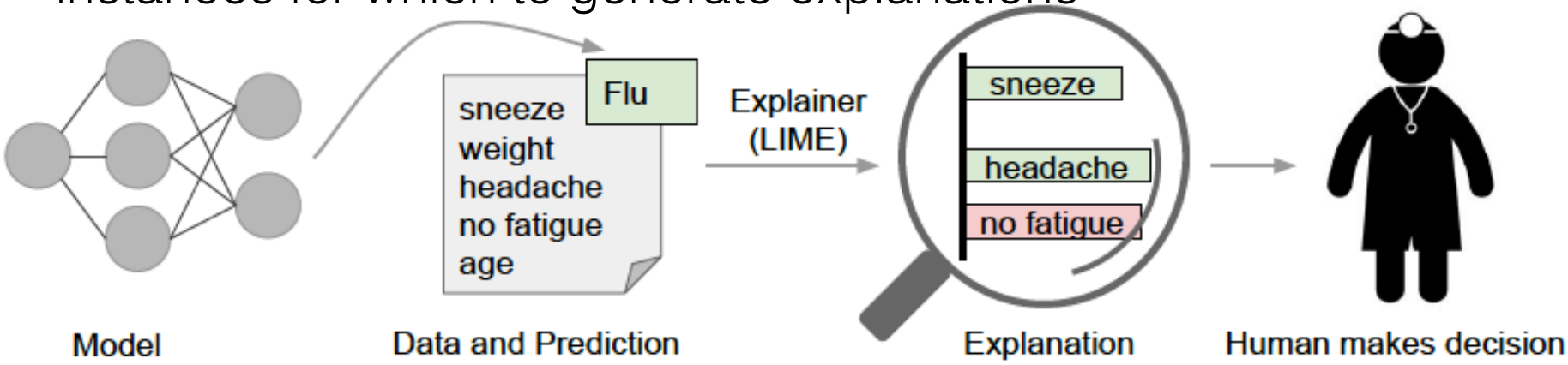
Important questions

- how is accuracy measured?
- accuracy for whom? over-all or in sub-populations?
- accuracy over which data?
- 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*]

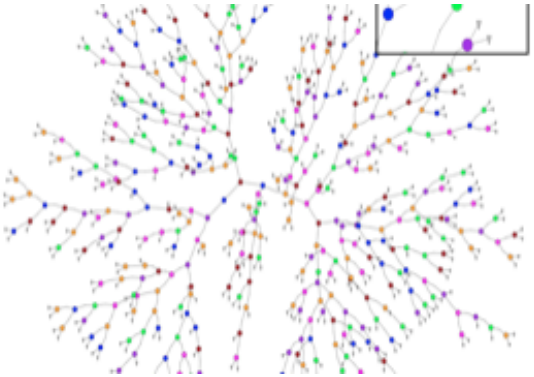
https://www.youtube.com/watch?v=hUnRCxnydCc

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

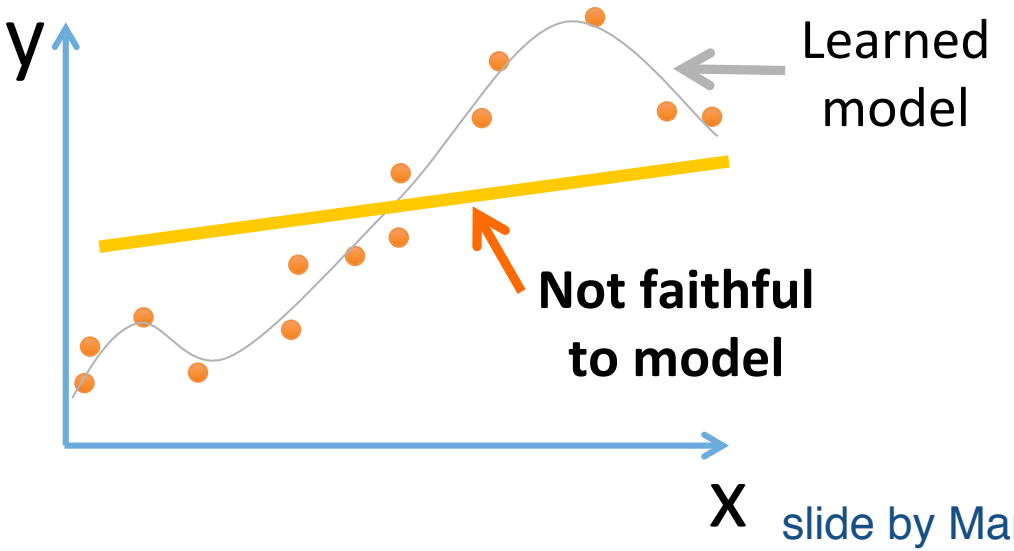
LIME: Local explanations of classifiers

[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

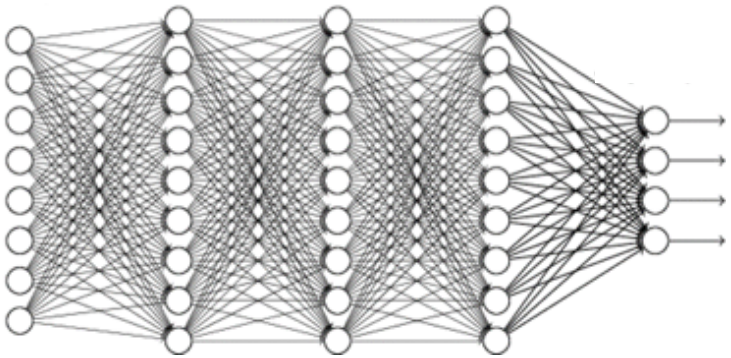
[M. T. Ribeiro, S. Singh, C. Guestrin; *KDD 2016*]

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."

- LIME relies on a distinction between **features** and **interpretable data representations**; examples:
  - In text classification features are word embeddings; an interpretable representation is a vector indicating the presence 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."

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

classifier **model being explained**

explanation model

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***

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

explanation

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

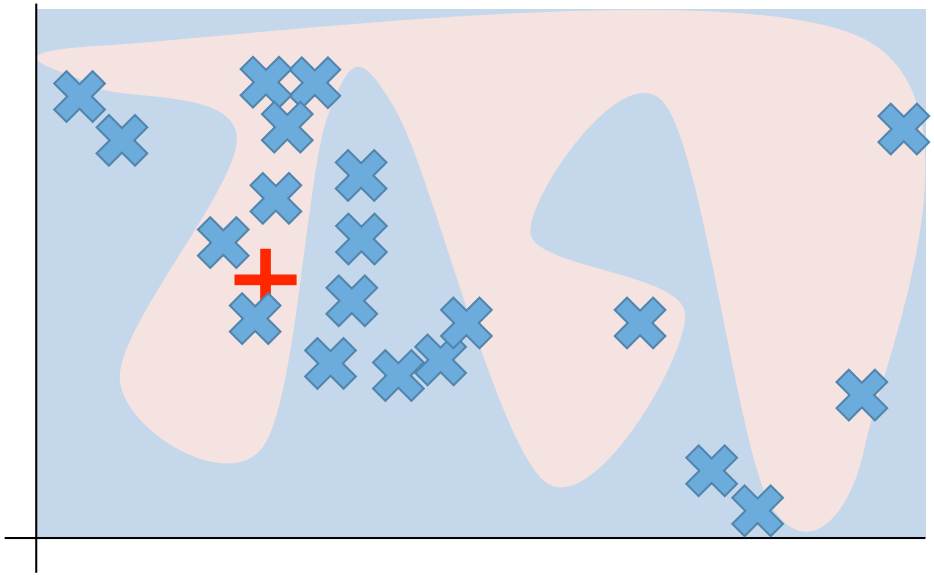
$$\xi(x) = \text{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

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 +



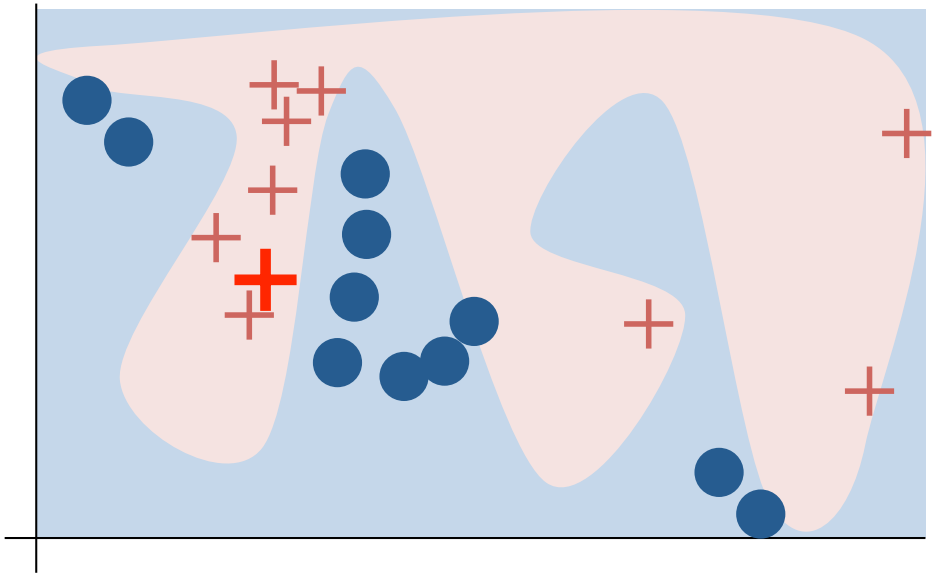
based on a slide by Marco Tulio Ribeiro, KDD 2016

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



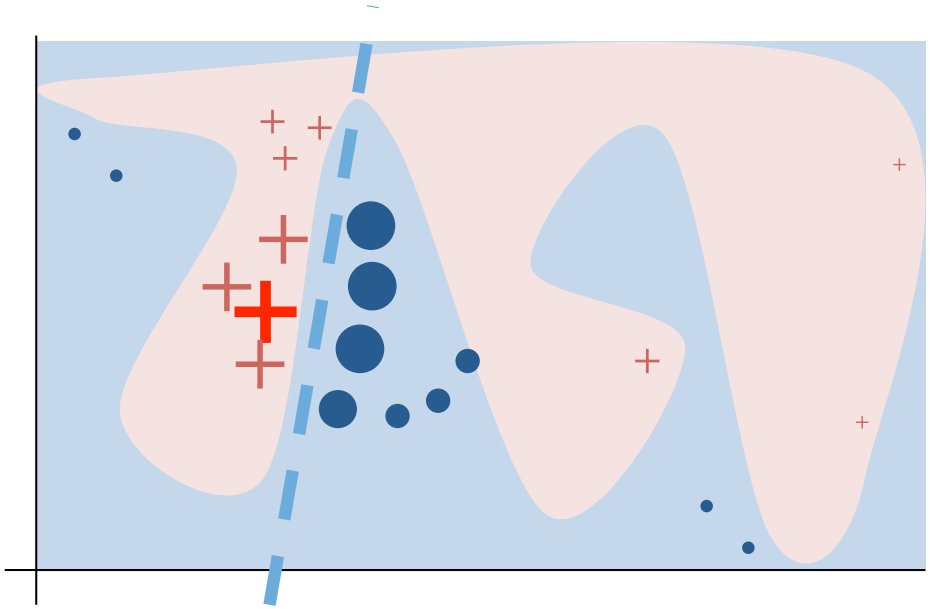
based on a slide by Marco Tulio Ribeiro, KDD 2016

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."

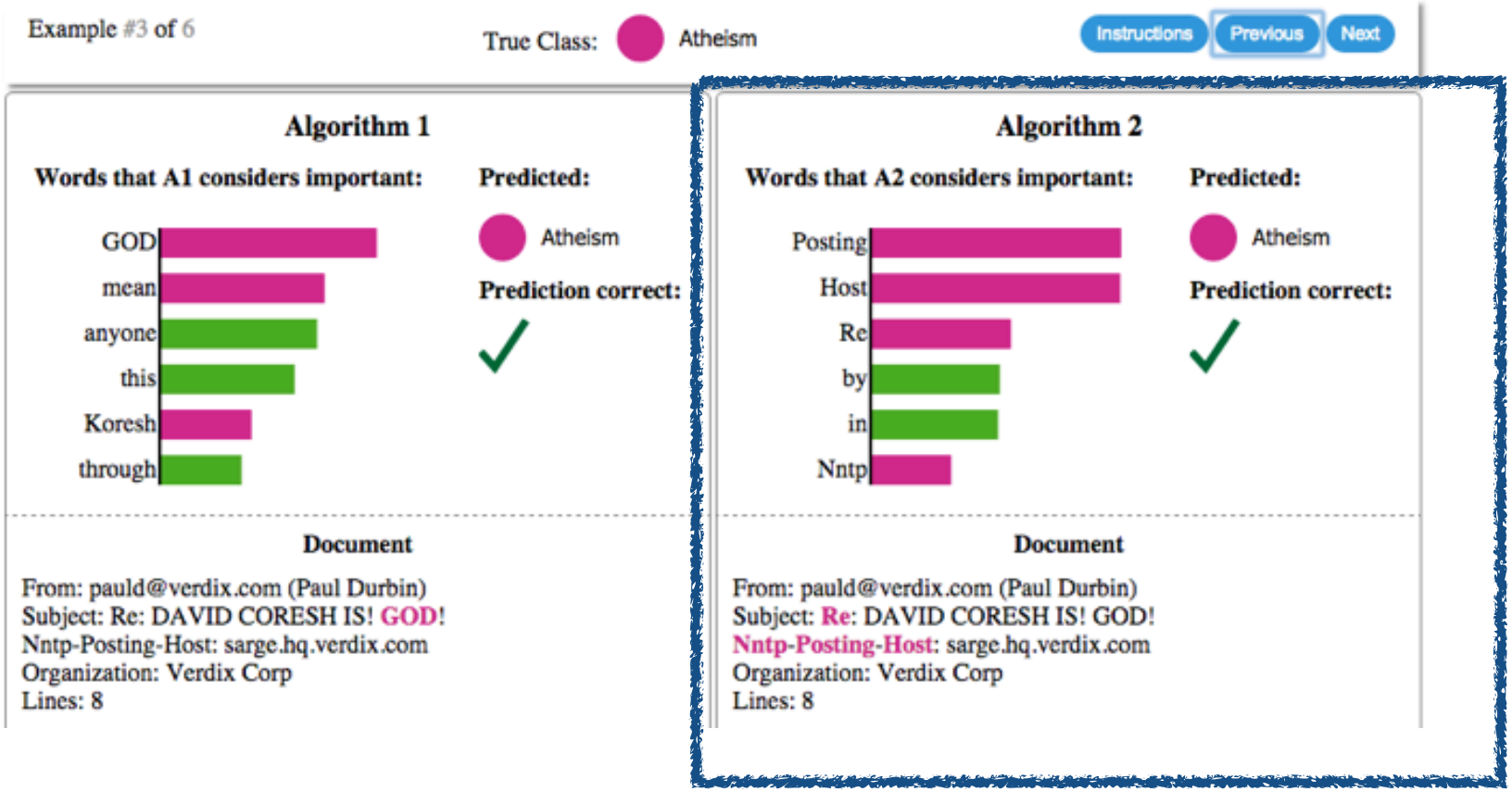
1. sample points around +
2. use complex model ***f*** to assign class labels
3. weigh samples according to π_x
4. learn simple model ***g*** according to samples



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

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!!!

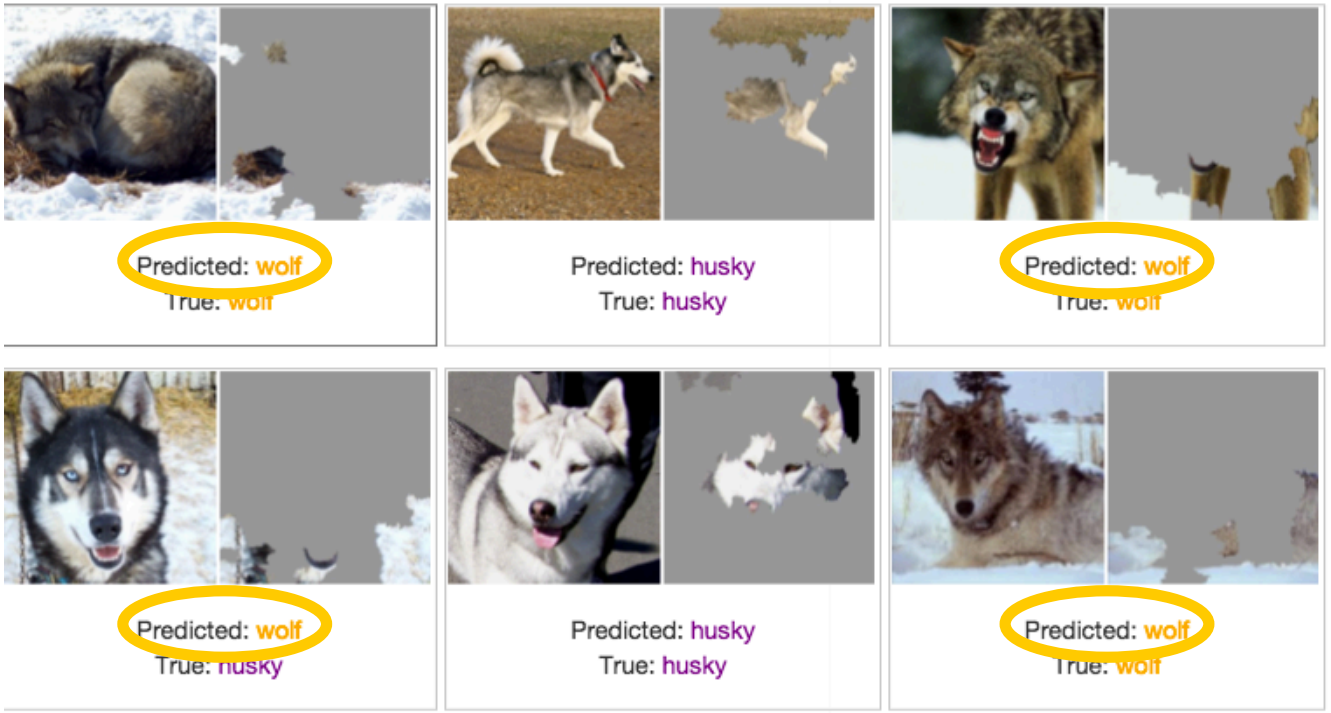
Do you trust this model?
How does it distinguish between huskies and wolves?

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

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

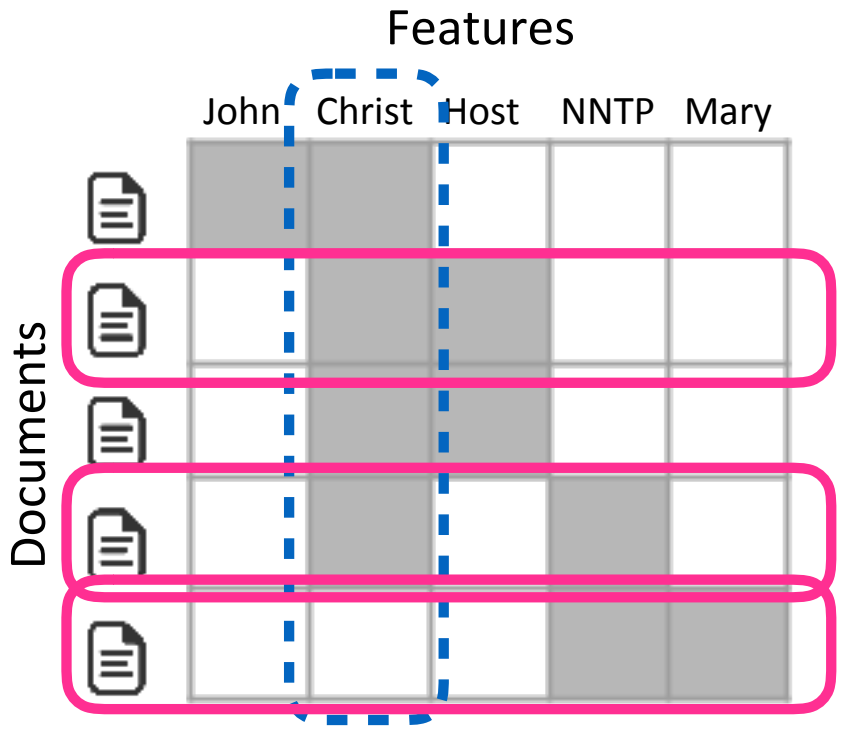
Given a budget **B** of explanations that a user is willing to consider, **pick** a set of ***B*** representative instances for the user to inspect

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

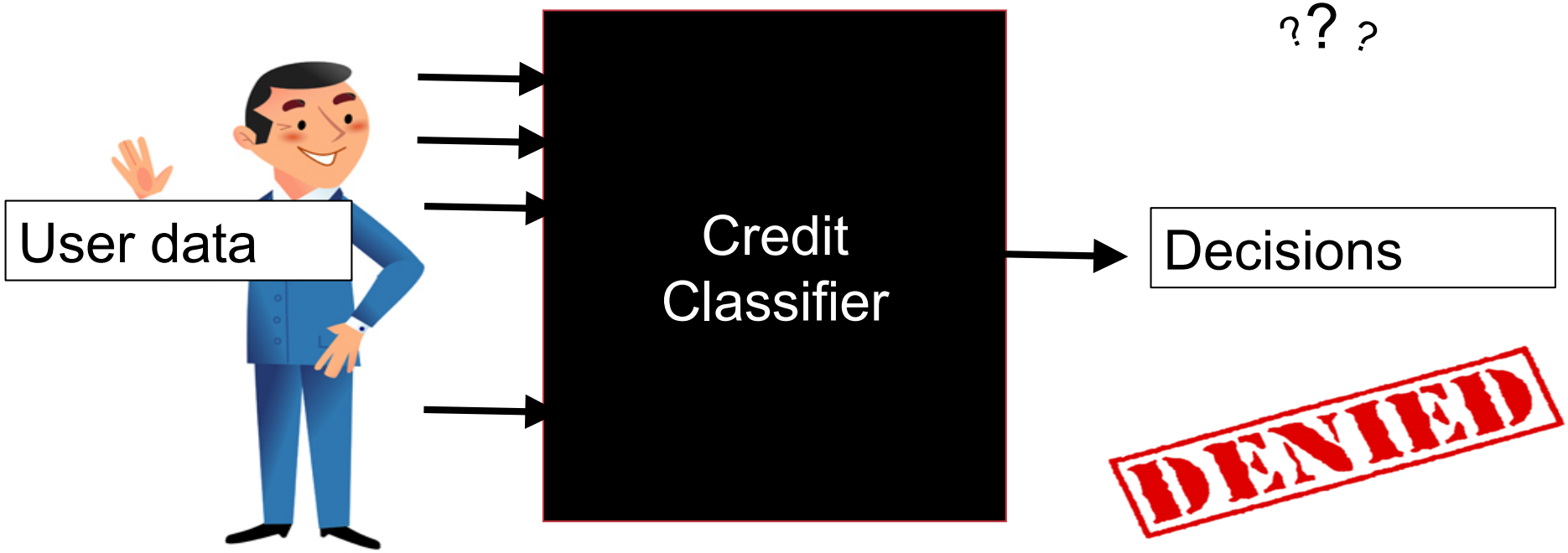
Explaining black-box classifiers

Algorithmic transparency with quantitative input influence (**QII**)

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

Auditing black-box models

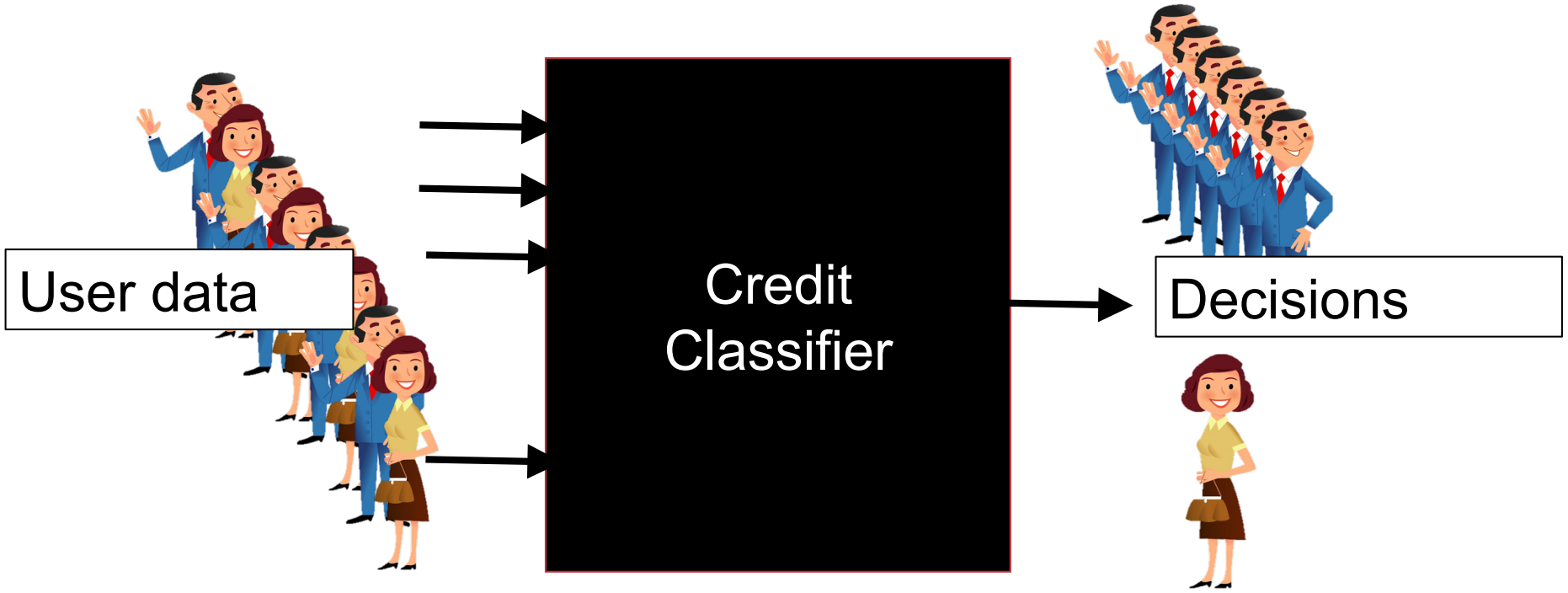
[A. Datta, S. Sen, Y. Zick; *SP 2016*]



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*]

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. **Must have knowledge of the input dataset on which the model operates**.

Back to the example

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

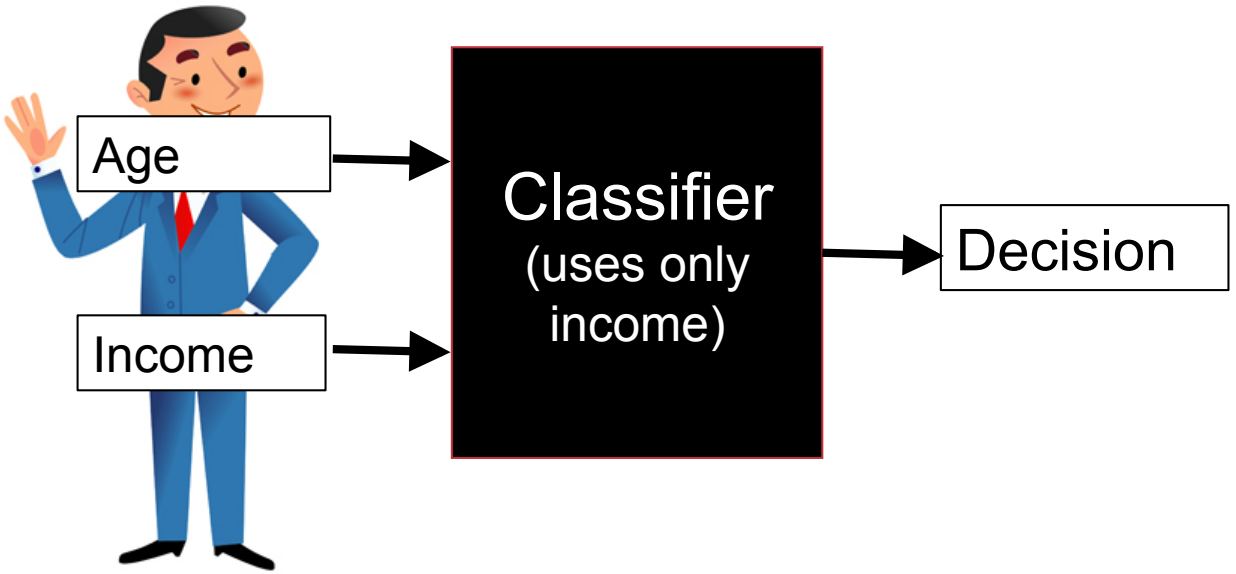
Running example: Consider hiring decisions by a moving company, based on gender, age, education, and weight lifting ability. **Does gender influence hiring decisions?**

- 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 **after the intervention** - we conclude that gender does not influence hiring decisions.
- Do a similar test for other features, one at a time. This is known as **Unary QII**

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?

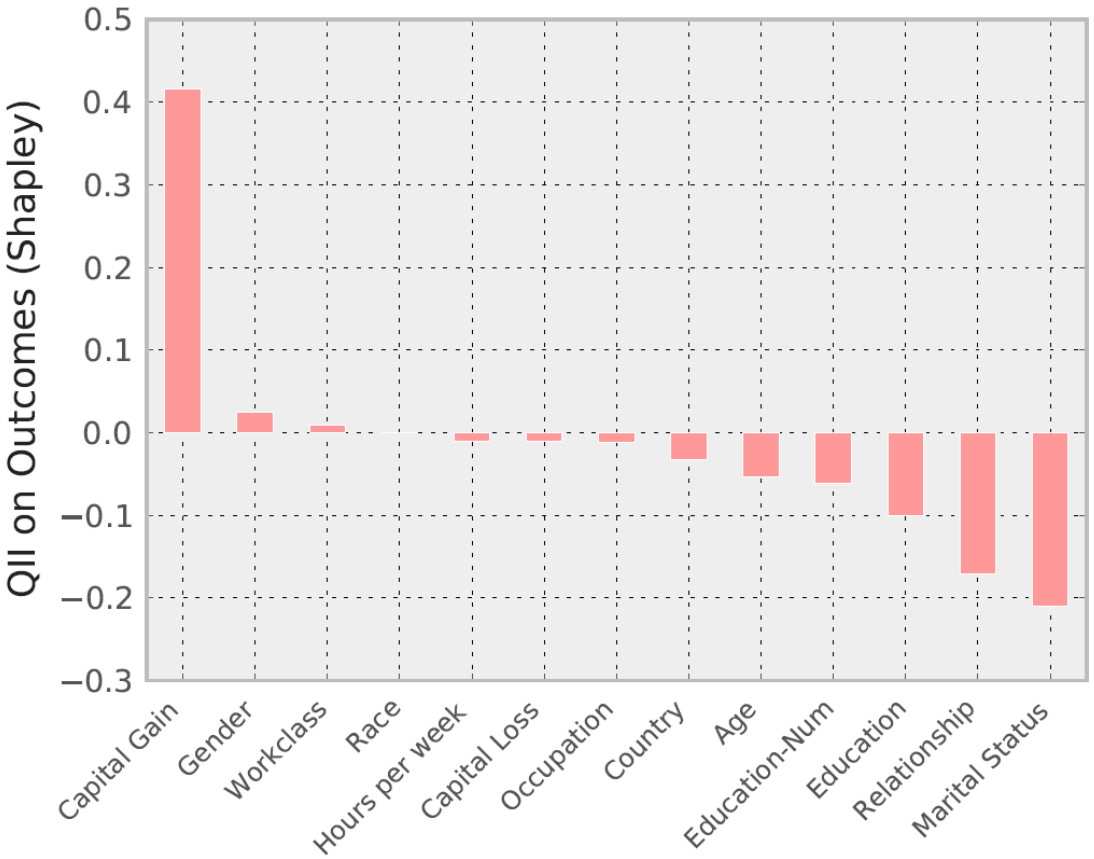
Group: Which inputs have the most influence on credit decisions for women?

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

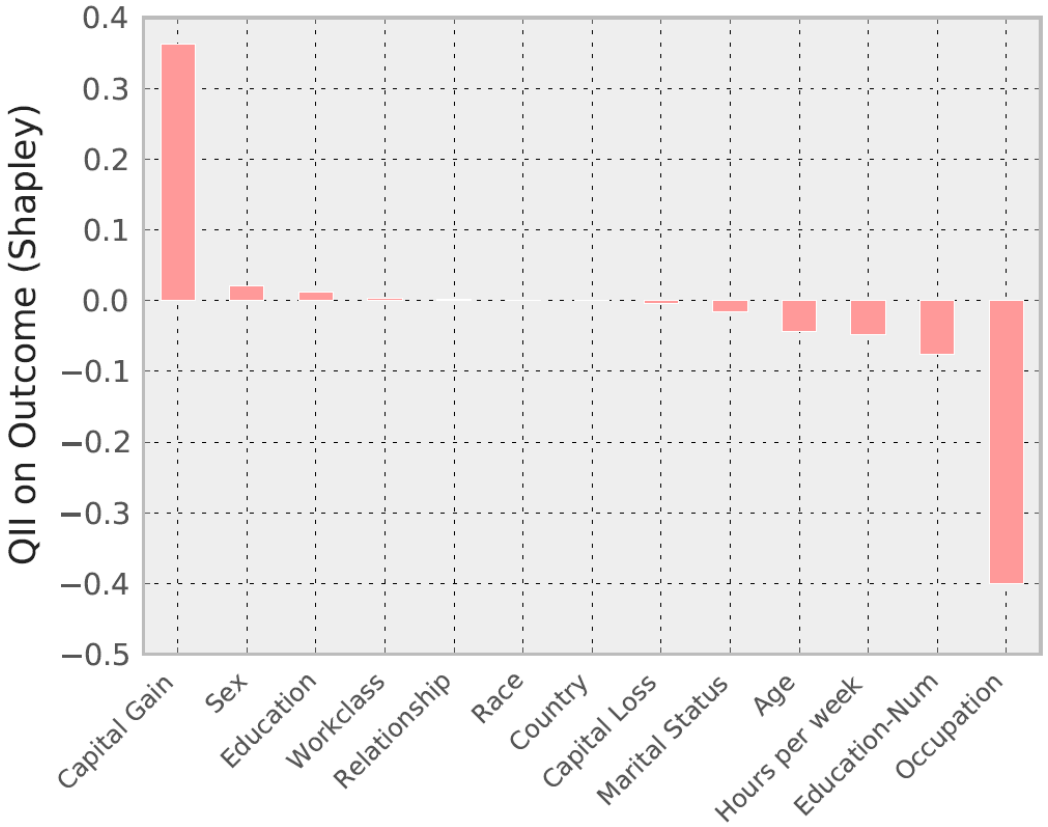
Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344
Capital loss	$0
Work hours per week	40
Country	Vietnam

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

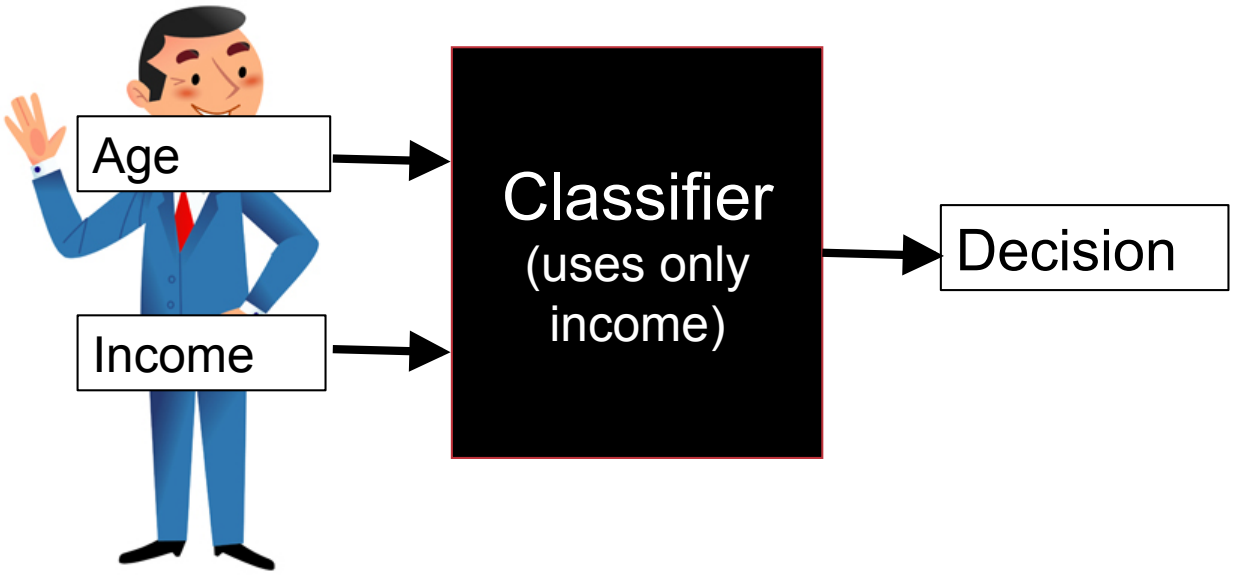
explanations for superficially similar individuals can be different

slide by A. Datta

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

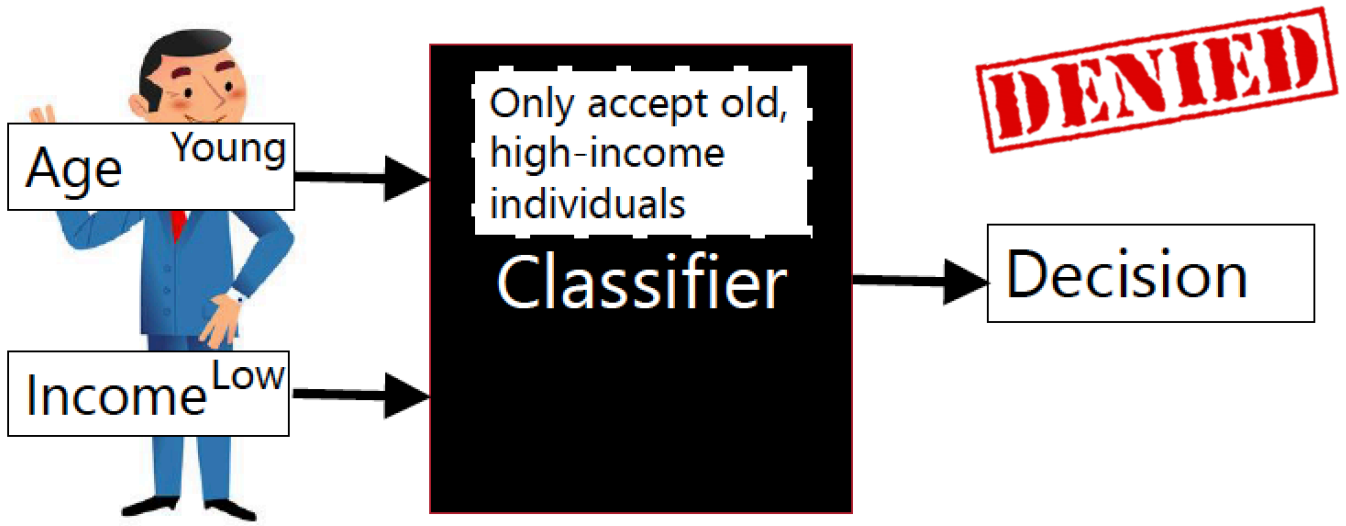
does this tell the whole story?

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**.

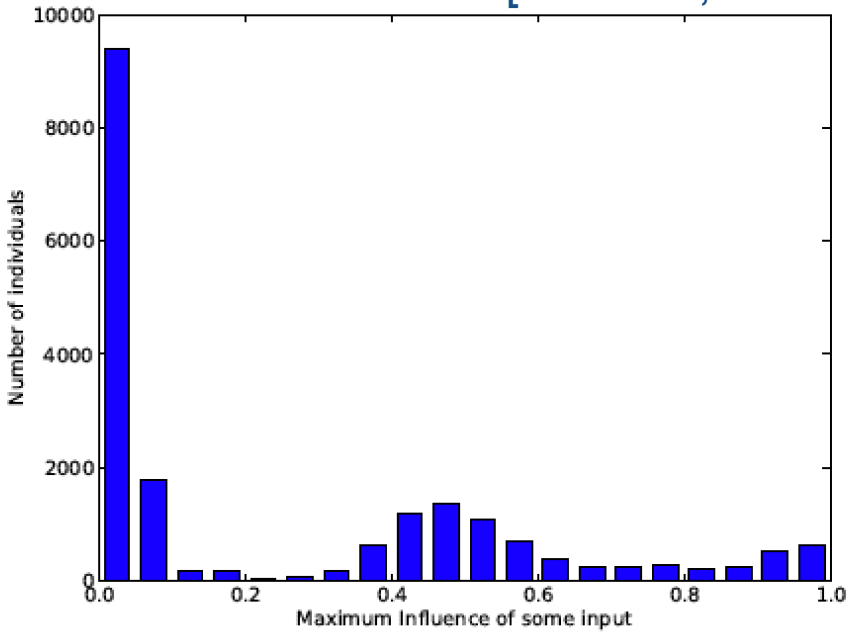


intervening on one feature at a time will not have any effect

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!

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

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

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

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

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

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

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

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

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

Aggregating influence across sets

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

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

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

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

$v : 2^N \to \mathbb{R}$ influence of a set of features ***S*** on the outcome

$\varphi_i(N,v)$ influence of feature ***i***, given the set of features ***N*** = {***1***,…, ***n***}

$\sigma \in \Pi(N)$ a permutation over the features in set ***N***

$m_i(\sigma)$ payoff corresponding to this permutation

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**

Explaining black-box classifiers

A unified approach to interpreting model predictions (**SHAP**)

[Scott Lundberg and Su-In Lee; *NeurIPS 2017*]

SHAP, in summary

[S. Lundberg and S. Lee; *NeurIPS 2017*]

- **SHAP** stands for **SH**apley **A**dditive ex**P**lanations
- Claim: A unifying framework for interpreting predictions with "**additive feature attribution methods**", including LIME and QII, for **local explanations**
- The best explanation of a **simple model** is the model itself: the explanation is both accurate and interpretable. For complex models we must use a simpler **explanation model** — an interpretable approximation of the original model.

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

model being explained

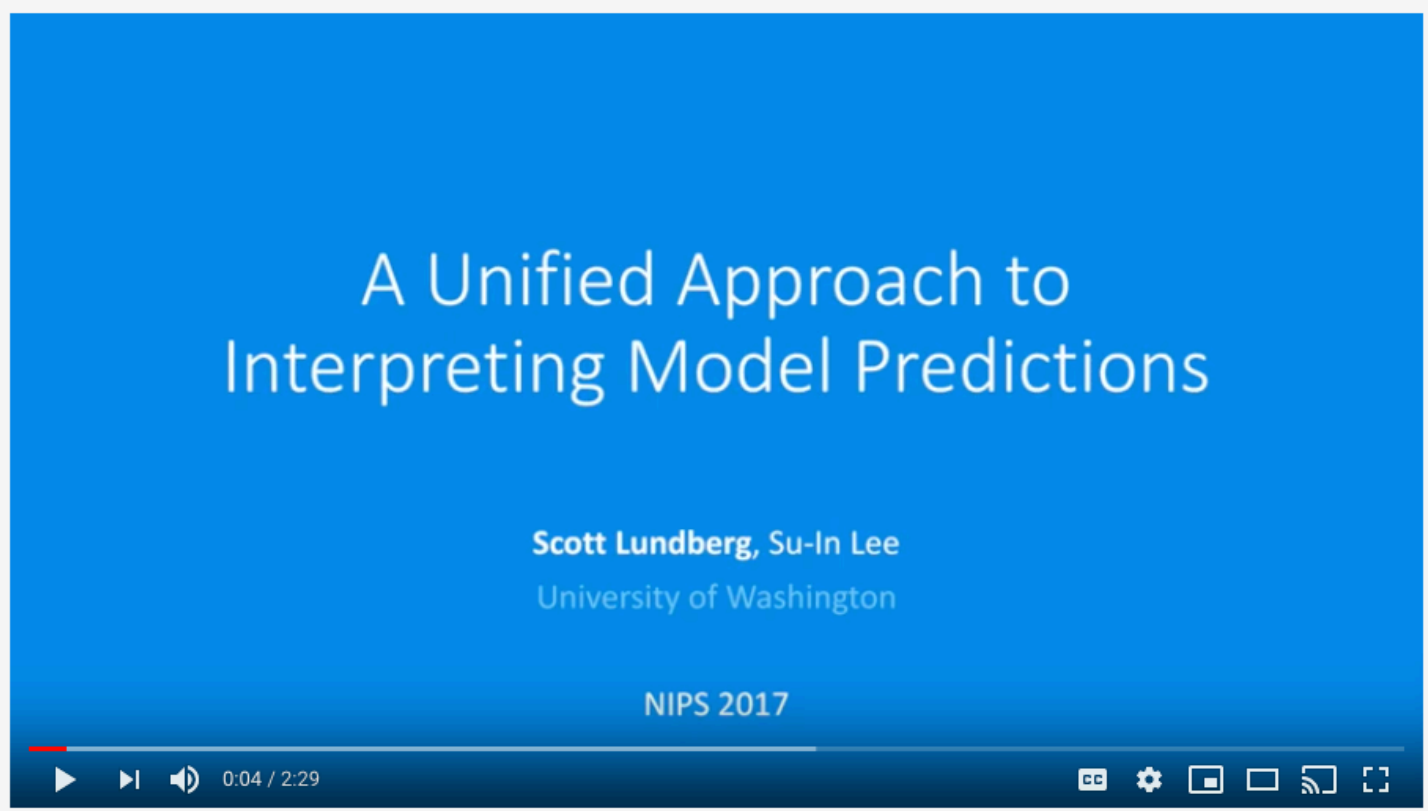
$$g \in G, dom(g) = \{0,1\}^{d'}$$

explanation model from a class of interpretable models, over a set of **simplified features**

- **Additive feature attribution methods** have an explanation model that is a linear function of binary variables

SHAP, in summary

[S. Lundberg and S. Lee; *NeurIPS 2017*]



https://www.youtube.com/watch?v=wjd1G5bu_TY

Additive feature attribution methods

[S. Lundberg and S. Lee; *NeurIPS 2017*]

Additive feature attribution methods have an explanation model that is a linear function of binary variables (simplified features)

$$g(x') = \phi_0 + \sum_{i=1}^{d'} \phi_i x'_i \quad \text{where } x' \in \{0,1\}^{d'}, \text{and } \phi_i \in \mathbb{R}$$

Three properties guarantee a single unique solution — a unique allocation of Shapley values to each feature

1. **Local accuracy**: ***g(x')*** matches the original model ***f(x)*** when ***x'*** is the **simplified input** corresponding to x.

2. **Missingness**: if ***x'$_i$*** — the ith feature of simplified input **x'**— is missing, then it has no attributable impact for ***x***
$$x'_i = 0 \Rightarrow \phi_i = 0$$

3. **Consistency** (**monotonicity**): if toggling off feature ***i*** makes a bigger (or the same) difference in model ***f'(x)*** than in model ***f(x)***, then the weight (attribution) of ***i*** should be no lower in ***f'(x)*** than in ***f(x)***

Additive feature attribution methods

[S. Lundberg and S. Lee; *NeurIPS 2017*]

README.md

SHAP

https://github.com/slundberg/shap