Responsible Data Science

Transparency & Interpretability

Auditing black-box models

April 3 & 5, 2022

Prof. George Wood

Center for Data Science
New York University

NYU Center for Data Science

r/ai

Reading for this module

Week 10: Auditing black-box models

SHAP and LIME

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

NB: Paper is technical: don't worry too much about the details, try to get the intuition

r/ai

Reading for this module

Week 11: Discrimination in on-line ad delivery

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

BY LATANYA SWEENEY

Discrimination in Online Ad Delivery

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

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

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

r/ai

Reading for this module

Weeks 12, 13: Interpretability

THE INTUITIVE APPEAL OF EXPLAINABLE MACHINES

Andrew D. Selbst & Solon Barocas***

Nutritional Labels for Data and Models *

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

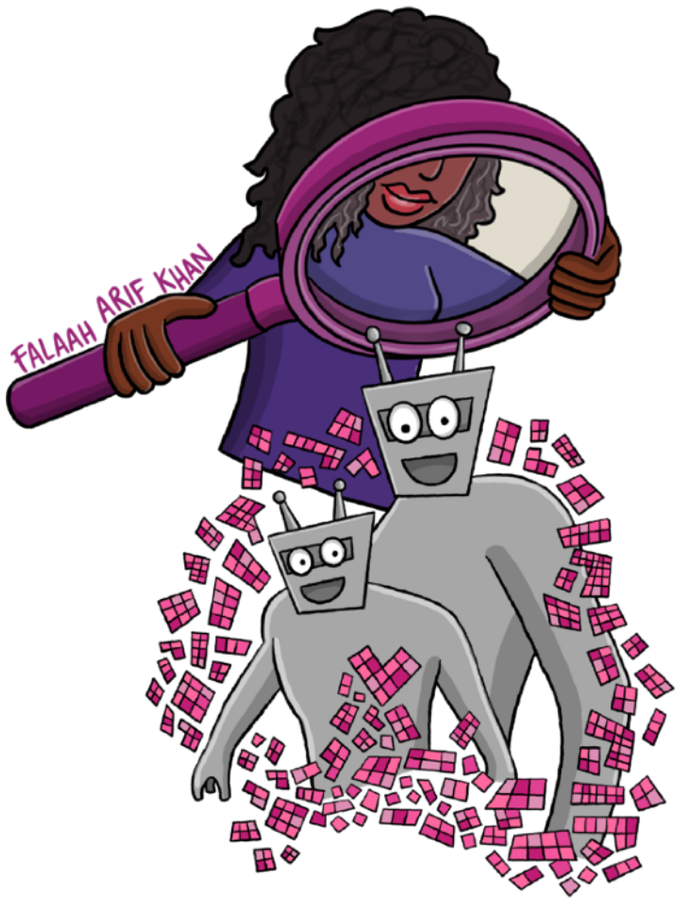
The imperative of interpretable machines

As artificial intelligence becomes prevalent in society, a framework is needed to connect interpretability and trust in algorithm-assisted decisions, for a range of stakeholders.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

r/ai

Terminology & vision



transparency, interpretability,
explainability, intelligibility

responsible AI

agency, responsibility

r/ai

Interpretability for different stakeholders

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

r/ai

Staples discounts

THE WALL STREET JOURNAL.

December 2012

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani

December 24, 2012

WHAT PRICE WOULD YOU SEE?

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.

A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, **Staples appeared to consider the person's distance from a rival brick-and-mortar store**, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

https://www.wsj.com/articles/SB10001424127887323772045781893918138815 34

r/ai

Staples discounts

December 2012

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

WHAT PRICE WOULD YOU SEE?

It was the same Sv
same Staples.com
was $15.79, while
a few miles away,

A key difference:
located.

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

A Wall Street Journal investigation found that th
Inc. website displays different prices to people after estimating their locations. More than that, **Staples appeared to consider the person's distance from a rival brick-and-mortar store**, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

https://www.wsj.com/articles/SB10001424127887323772045781893918138815 34

r/ai

Online job ads

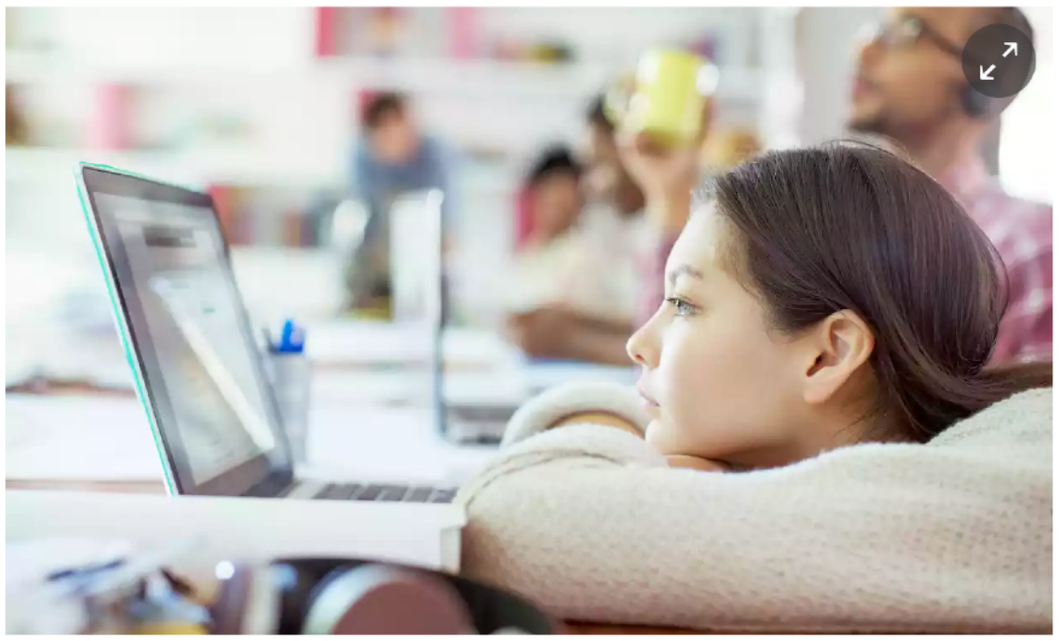
theguardian

July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



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

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

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for "$200k+" executive jobs **1,852 times to the male group and only 318 times to the female group**. Another experiment, in July 2014, showed a similar trend but was not statistically significant.

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

r/ai

Online job ads

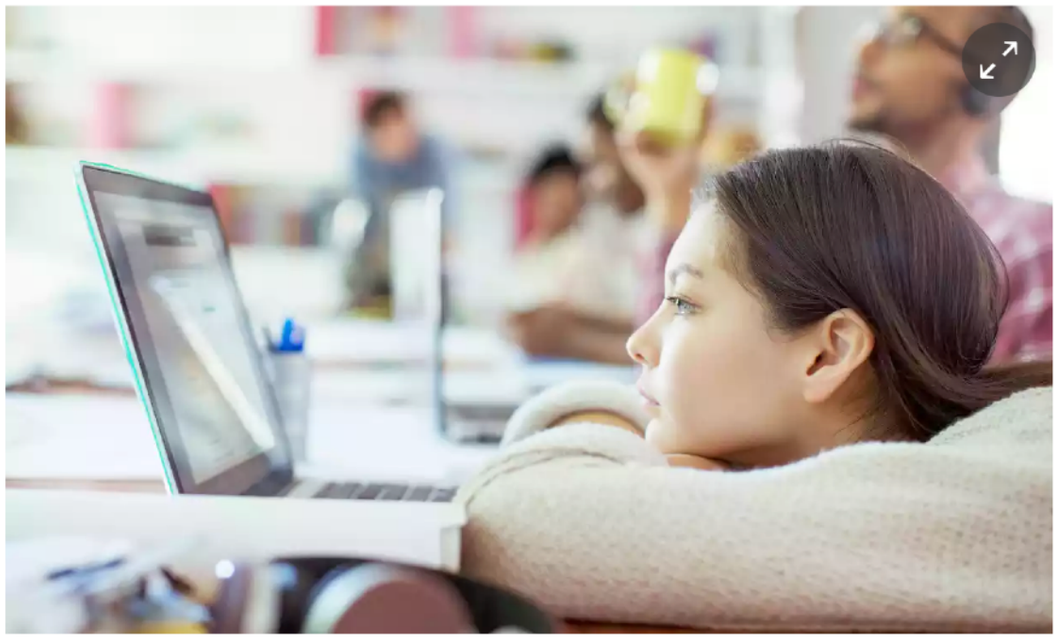
theguardian

July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



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

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

The AdFisher tool simulated job seekers that did not differ in browsing beha
demographic

One experim
ads for a car
executive job
and only 31
Another exp
similar trend

What are we explaining?

To **Whom** are we explaining?

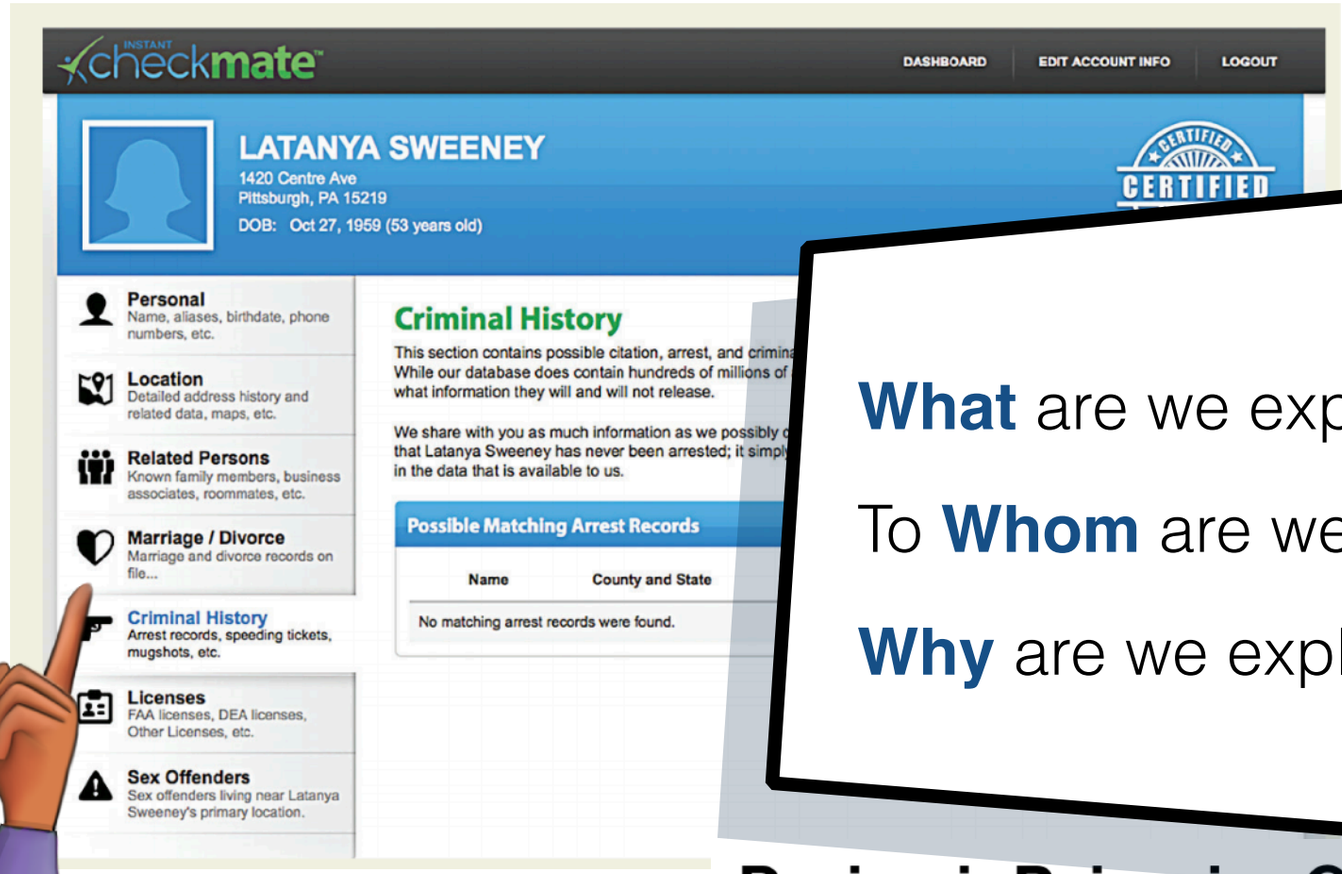
Why are we explaining?

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

r/ai

Instant Checkmate

Google AdSense



February 2013

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

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

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

r/ai

Nutritional labels

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

SIDE-BY-SIDE COMPARISON

Original Label		New Label	
Nutrition Facts		**Nutrition Facts**	
Serving Size 2/3 cup (55g)		8 servings per container	
Servings Per Container About 8		**Serving size**	**2/3 cup (55g)**
Amount Per Serving		**Amount per serving**	
Calories 230	Calories from Fat 72	**Calories**	**230**
	% Daily Value*		% Daily Value*
Total Fat 8g	12%	**Total Fat** 8g	10%
Saturated Fat 1g	5%	Saturated Fat 1g	5%
Trans Fat 0g		*Trans* Fat 0g	
Cholesterol 0mg	0%	**Cholesterol** 0mg	0%
Sodium 160mg	7%	**Sodium** 160mg	7%
Total Carbohydrate 37g	12%	**Total Carbohydrate** 37g	13%
Dietary Fiber 4g	16%	Dietary Fiber 4g	14%
Sugars 1g		Total Sugars 12g	
Protein 3g		Includes 10g Added Sugars	20%
		Protein 3g	
Vitamin A	10%	Vitamin D 2mcg	10%
Vitamin C	8%	Calcium 260mg	20%
Calcium	20%	Iron 8mg	45%
Iron	45%	Potassium 235mg	6%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs.

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

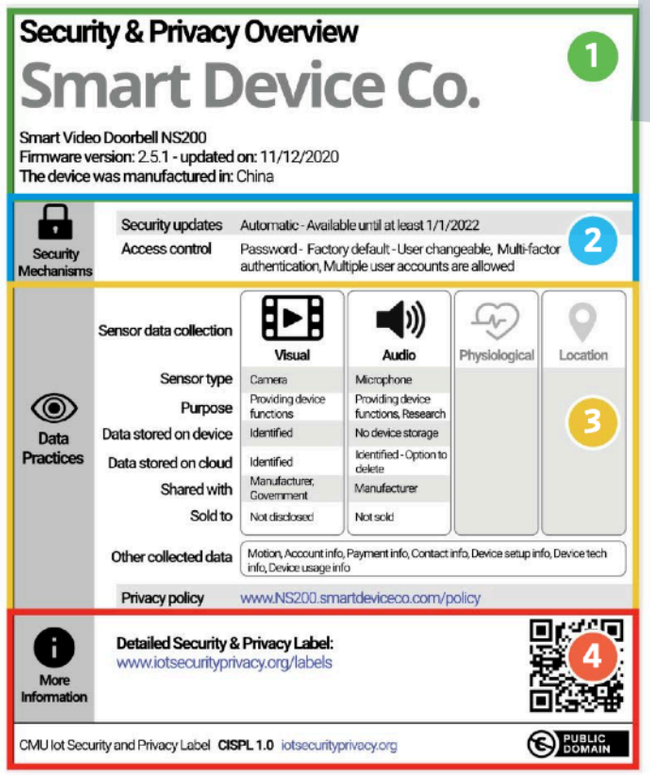
* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Note: The images above are meant for illustrative purposes to show how the new Nutrition Facts label might look compared to the old label. Both labels represent fictional products. When the original hypothetical label was developed in 2014 (the image on the left-hand side), added sugars was not yet proposed so the "original" label shows 1g of sugar as an example. The image created for the "new" label (shown on the right-hand side) lists 12g total sugar and 10g added sugar to give an example of how added sugars would be broken out with a % Daily Value.

An example of the old nutrition labels, left, and the new one. The new nutrition labels will display calories and serving size more prominently, and include added sugars for the first time.

PHOTO: FOOD AND DRUG ADMINISTRATION/ASSOCIATED PRESS

https://www.wsj.com/articles/why-the-labels-on-your-food-are-changing-or-



Security & Privacy Overview

Smart Device Co. 1

Smart Video Doorbell NS200
Firmware version: 2.5.1 - updated on: 11/12/2020
The device was manufactured in: China

Security Mechanisms 2

Security updates	Automatic - Available until at least 1/1/2022
Access control	Password - Factory default - User changeable, Multi-factor authentication, Multiple user accounts are allowed

Data Practices 3

	Visual	Audio	Physiological	Location
Sensor data collection				
Sensor type	Camera	Microphone		
Purpose	Providing device functions	Providing device functions, Research		
Data stored on device	Identified	No device storage		
Data stored on cloud	Identified	Identified - Option to delete		
Shared with	Manufacturer, Government	Manufacturer		
Sold to	Not disclosed	Not sold		

Other collected data: Motion, Account info, Payment info, Contact info, Device setup info, Device tech info, Device usage info

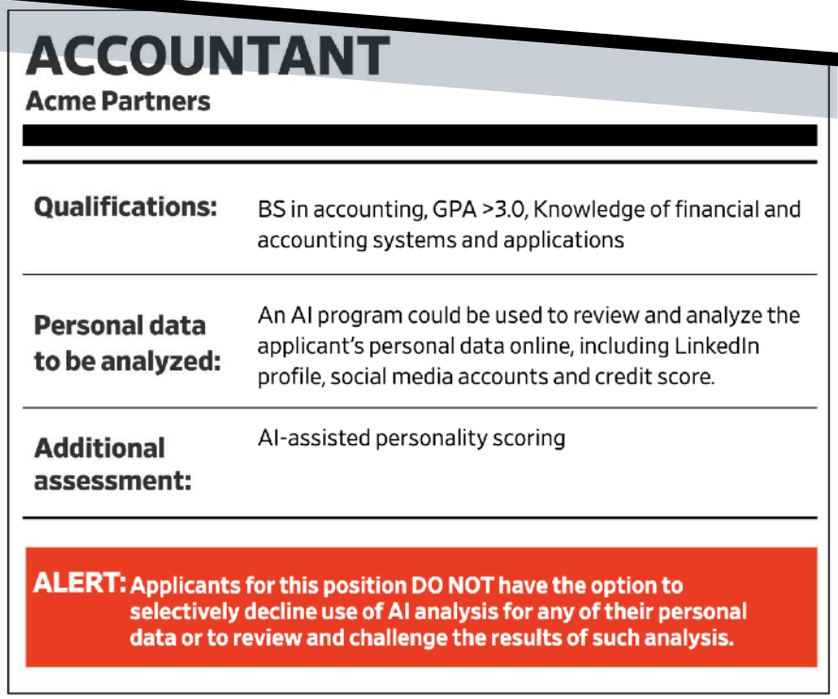
Privacy policy: www.NS200.smartdeviceco.com/policy

More Information

Detailed Security & Privacy Label:
www.iotsecurityprivacy.org/labels 4

CMU Iot Security and Privacy Label **CISPL 1.0** iotsecurityprivacy.org PUBLIC DOMAIN

https://www.wsj.com/articles/imagine-a-nutrition-labelfor-



ACCOUNTANT

Acme Partners

Qualifications:	BS in accounting, GPA >3.0, Knowledge of financial and accounting systems and applications
Personal data to be analyzed:	An AI program could be used to review and analyze the applicant's personal data online, including LinkedIn profile, social media accounts and credit score.
Additional assessment:	AI-assisted personality scoring

ALERT: Applicants for this position DO NOT have the option to selectively decline use of AI analysis for any of their personal data or to review and challenge the results of such analysis.

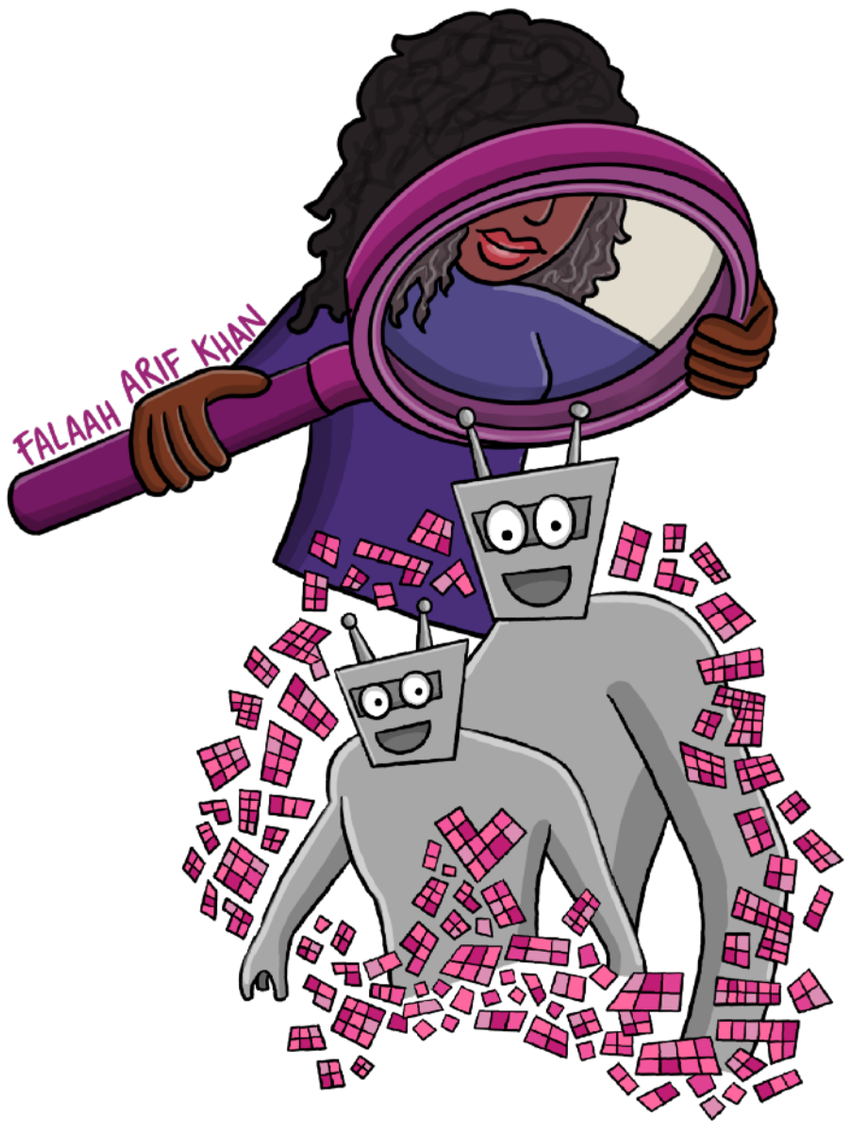
https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

r/ai

explaining black box models

r/ai

What are we explaining?



How does a system work?

How **well** does a system work?

What does a system do?

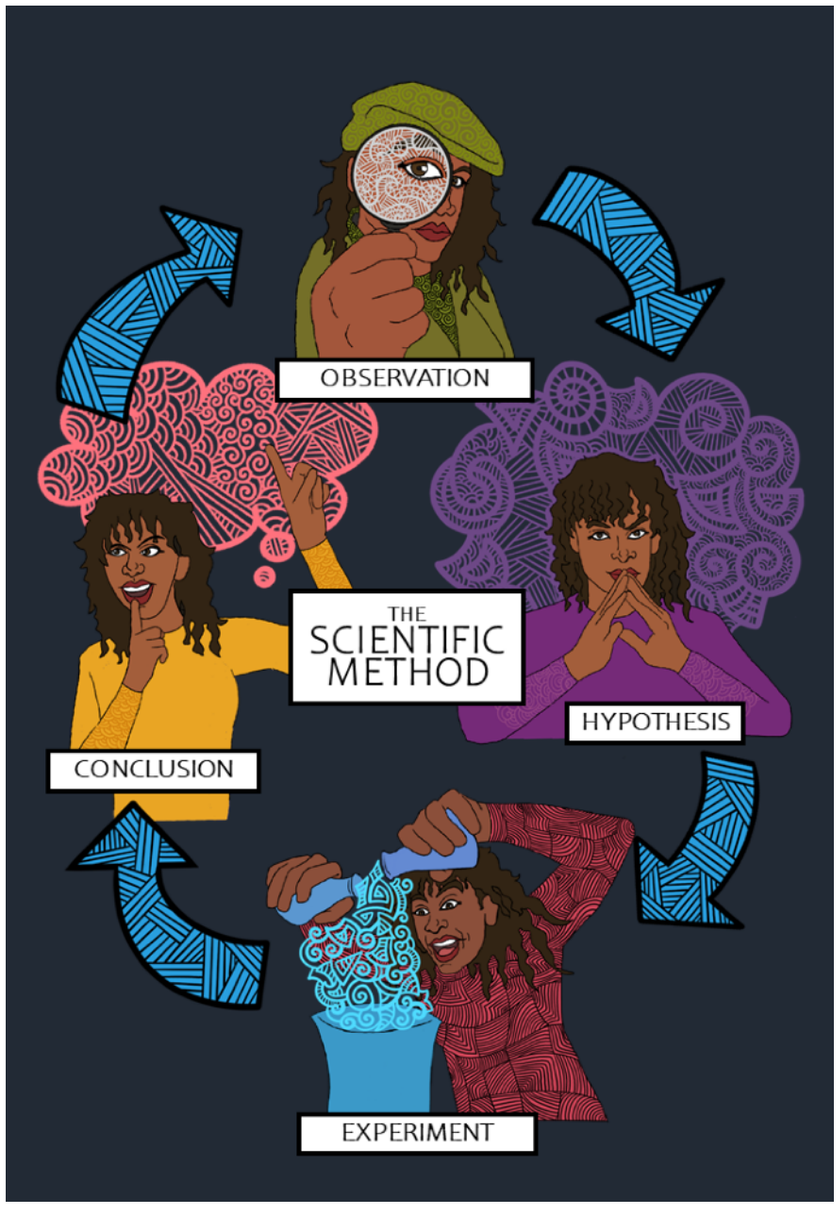
Why was I ___ (mis-diagnosed / not offered a discount / denied credit) ?

Are a system's decisions discriminatory?

Are a system's decisions illegal?

r/ai

But isn’t accuracy sufficient?



How is accuracy measured? FPR / FNR / …

Accuracy for whom: over-all or in sub-populations?

Accuracy over which data?

There is never 100% accuracy. Mistakes for what reason?

r/ai

Facebook's real-name policy

Shane Creepingbear is a member of the Kiowa Tribe of Oklahoma

October 13, 2014

Tweet

Shane Creepingbear @Creepingbear · Oct 13, 2014

Hey yall today I was kicked off of Facebook for having a fake name. Happy Columbus Day great job #facebook #goodtiming #racist #ColumbusDay

17

TIME

Facebook Thinks Some Native American Names Are Inauthentic

BY JOSH SANBURN FEBRUARY 14, 2015

February 14, 2015

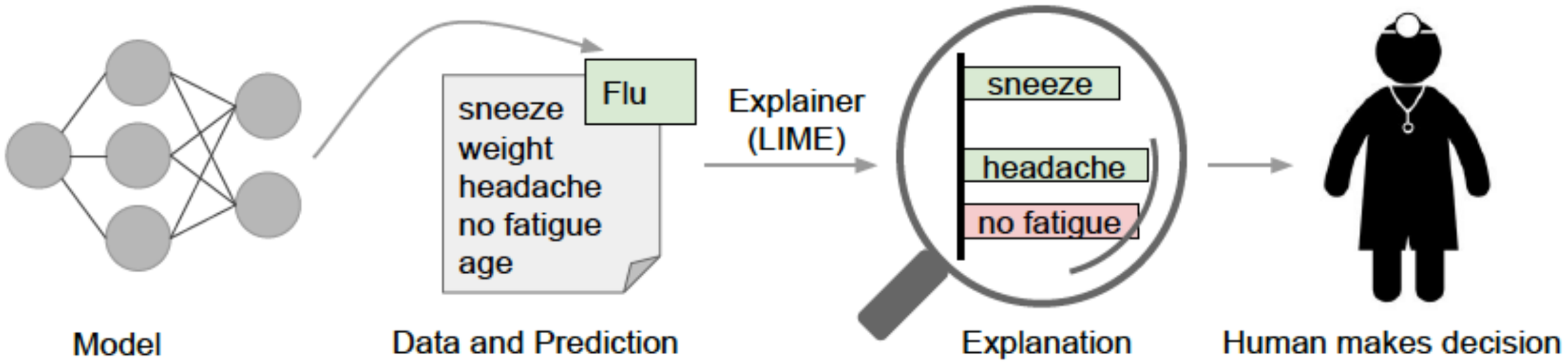
If you're Native American, Facebook might think your name is fake.

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.

Explanations based on features

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



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

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

[Ribeiro, Singh & Guestrin, 2016]

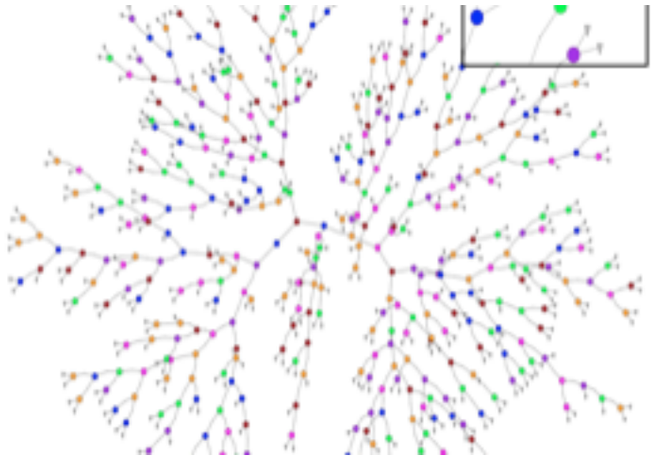
r/ai

LIME: Local explanations of classifiers

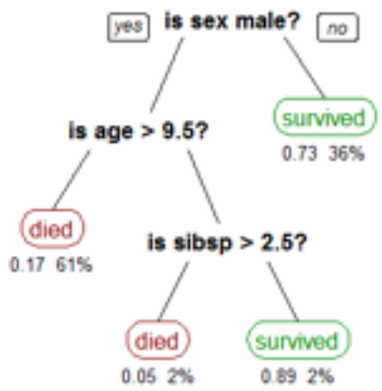
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely
not interpretable



Potentially
interpretable

slide by Marco Tulio Ribeiro, KDD 2016

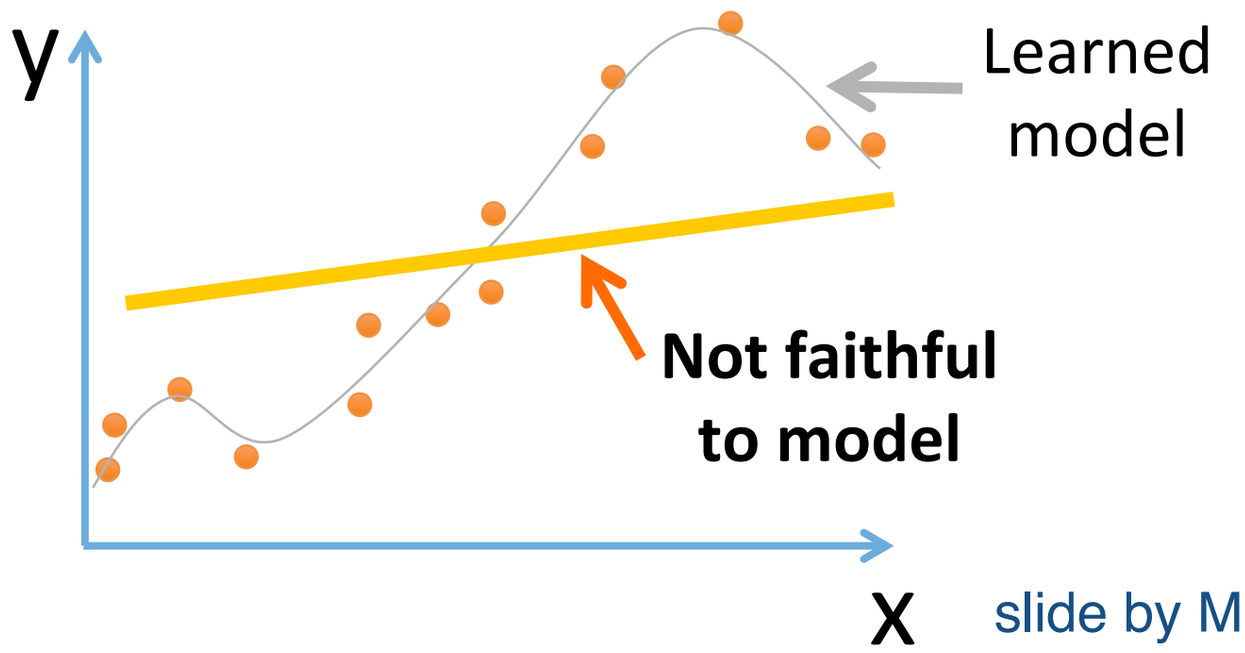
[Ribeiro, Singh & Guestrin, 2016]

r/ai

Explanations based on features

Three must-haves for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves



slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

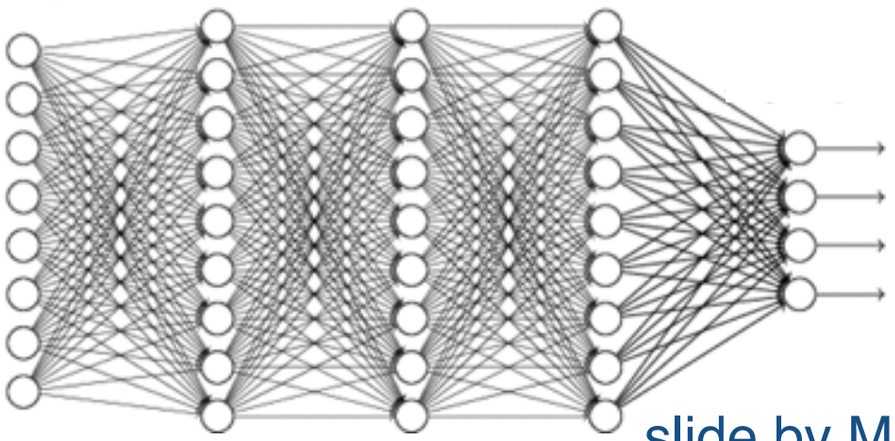
r/ai

Explanations based on features

Three must-haves for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves
Model agnostic	• Can be used for *any* ML model

Can explain this mess ☺



slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

r/ai

Key idea: Interpretable representation

"The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the 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

[Ribeiro, Singh & Guestrin, 2016]

r/ai

Fidelity-interpretability trade-off

"The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the 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***

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

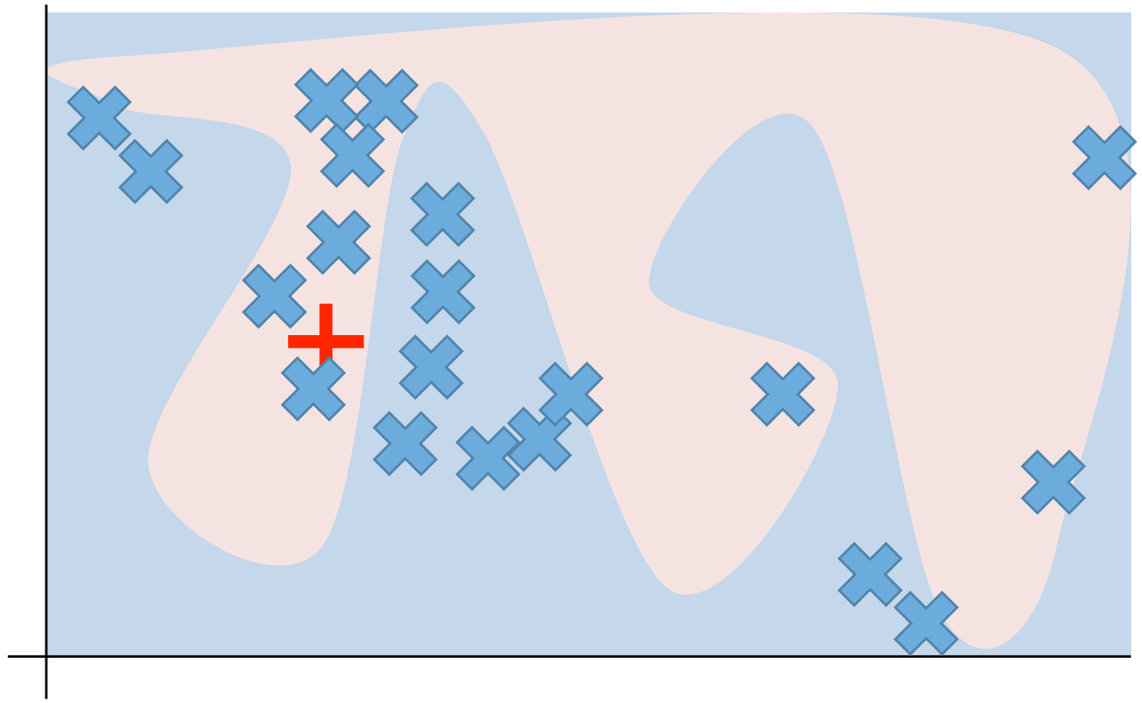
[Ribeiro, Singh & Guestrin, 2016]

r/ai

Fidelity-interpretability trade-off

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

1. sample points around +



based on a slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

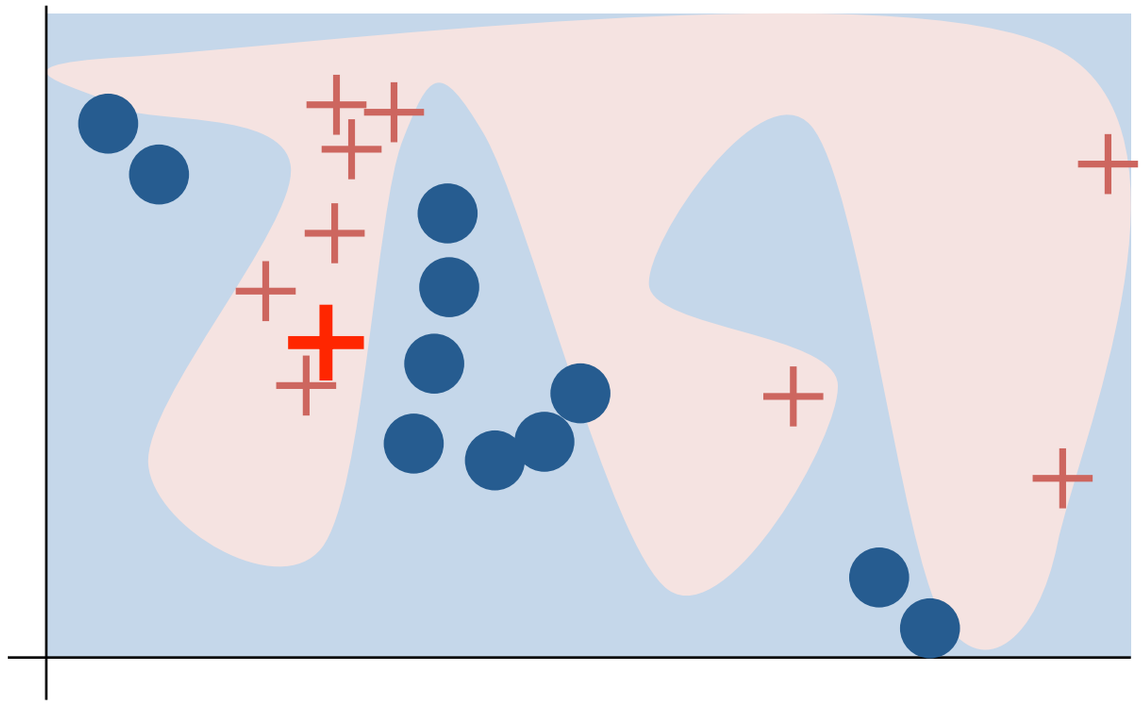
r/ai

Fidelity-interpretability trade-off

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

1. sample points around +

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



based on a slide by Marco Tulio Ribeiro, KDD 2016

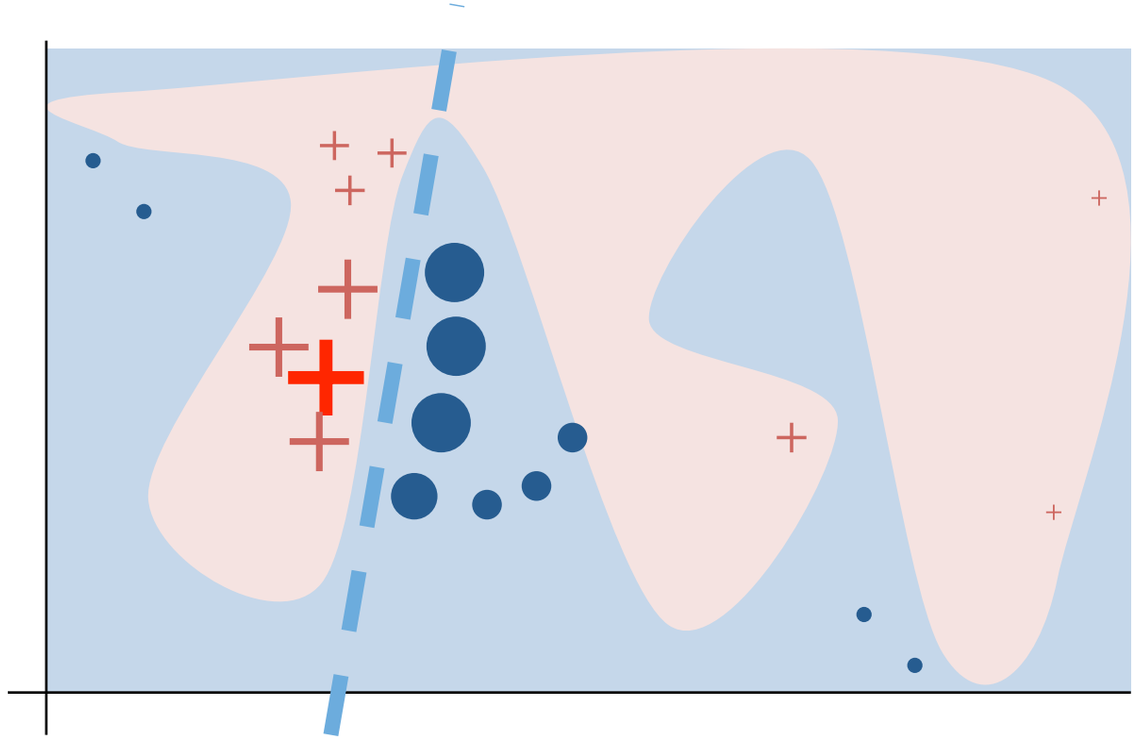
[Ribeiro, Singh & Guestrin, 2016]

r/ai

Fidelity-interpretability trade-off

"The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the 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

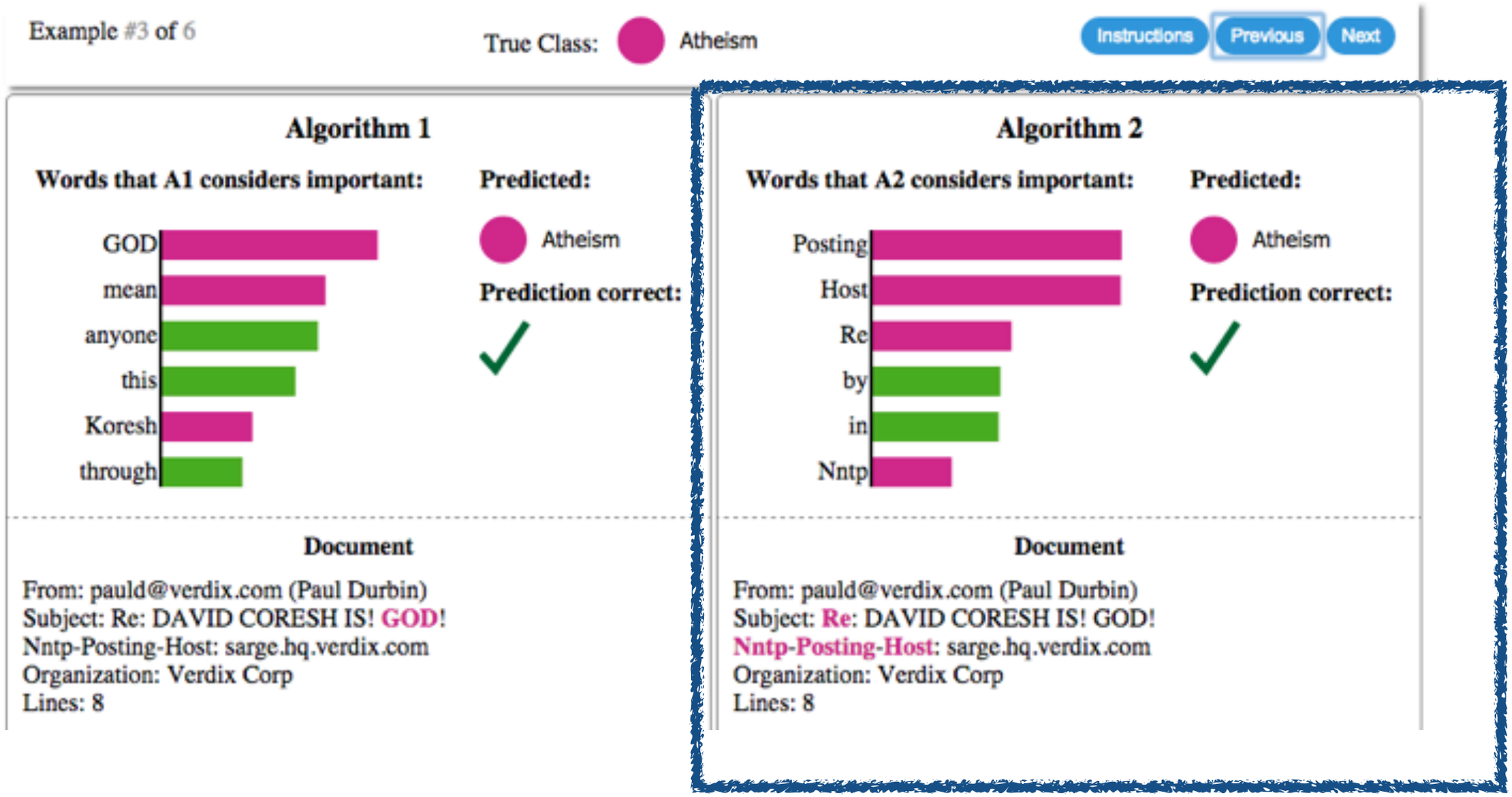


based on a slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

r/ai

Example: text classification with SVMs



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

[Ribeiro, Singh & Guestrin, 2016]

r/ai

When accuracy is not enough

Explaining Google's Inception NN

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

P( ) = 0.32

Electric guitar - incorrect but reasonable, similar fretboard

P( ) = 0.24

Acoustic guitar

P( ) = 0.21

Labrador

[Ribeiro, Singh & Guestrin, 2016]

r/ai

When accuracy is not enough

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?

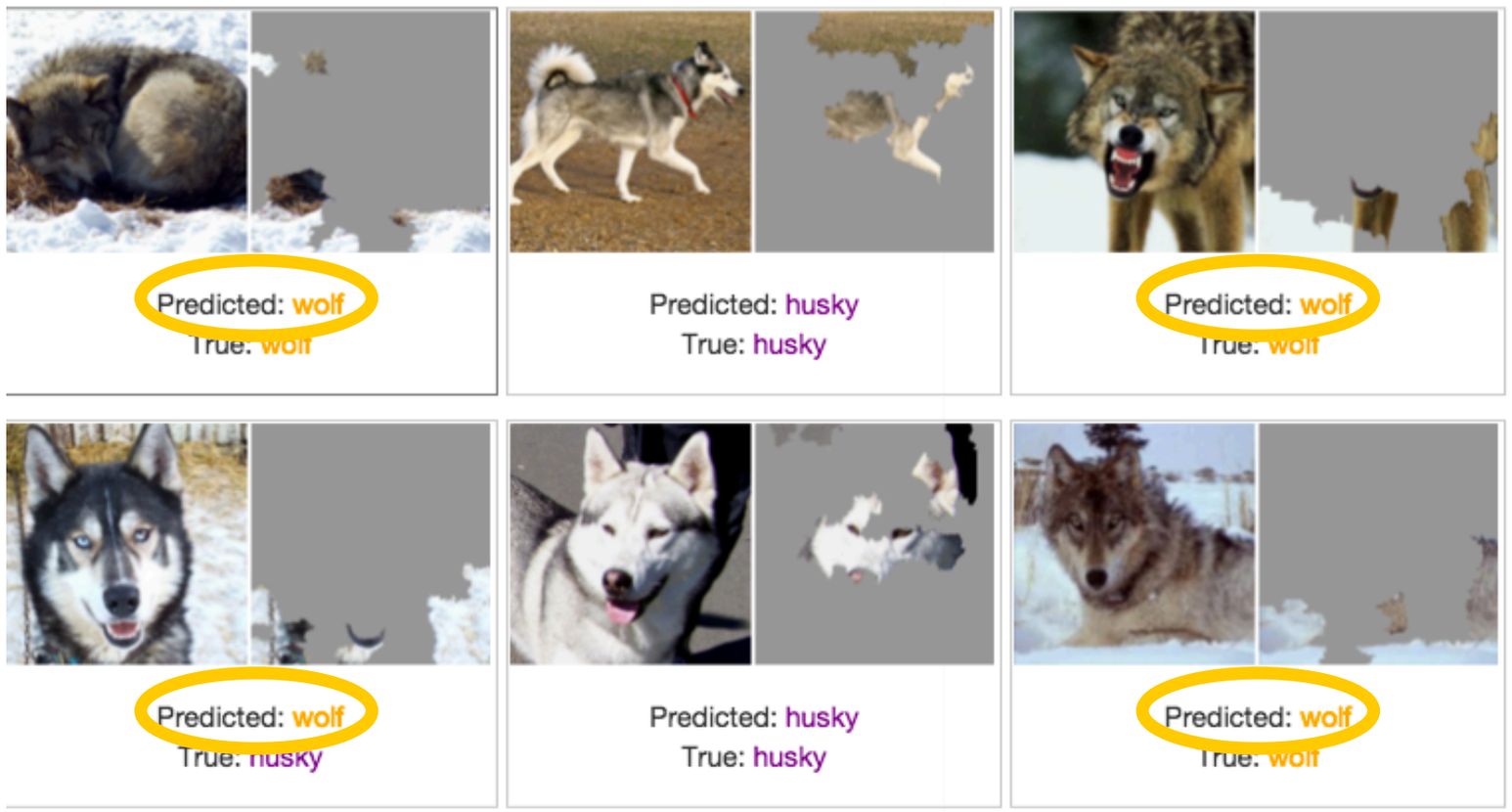
slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

r/ai

When accuracy is not enough

Explanations for neural network prediction



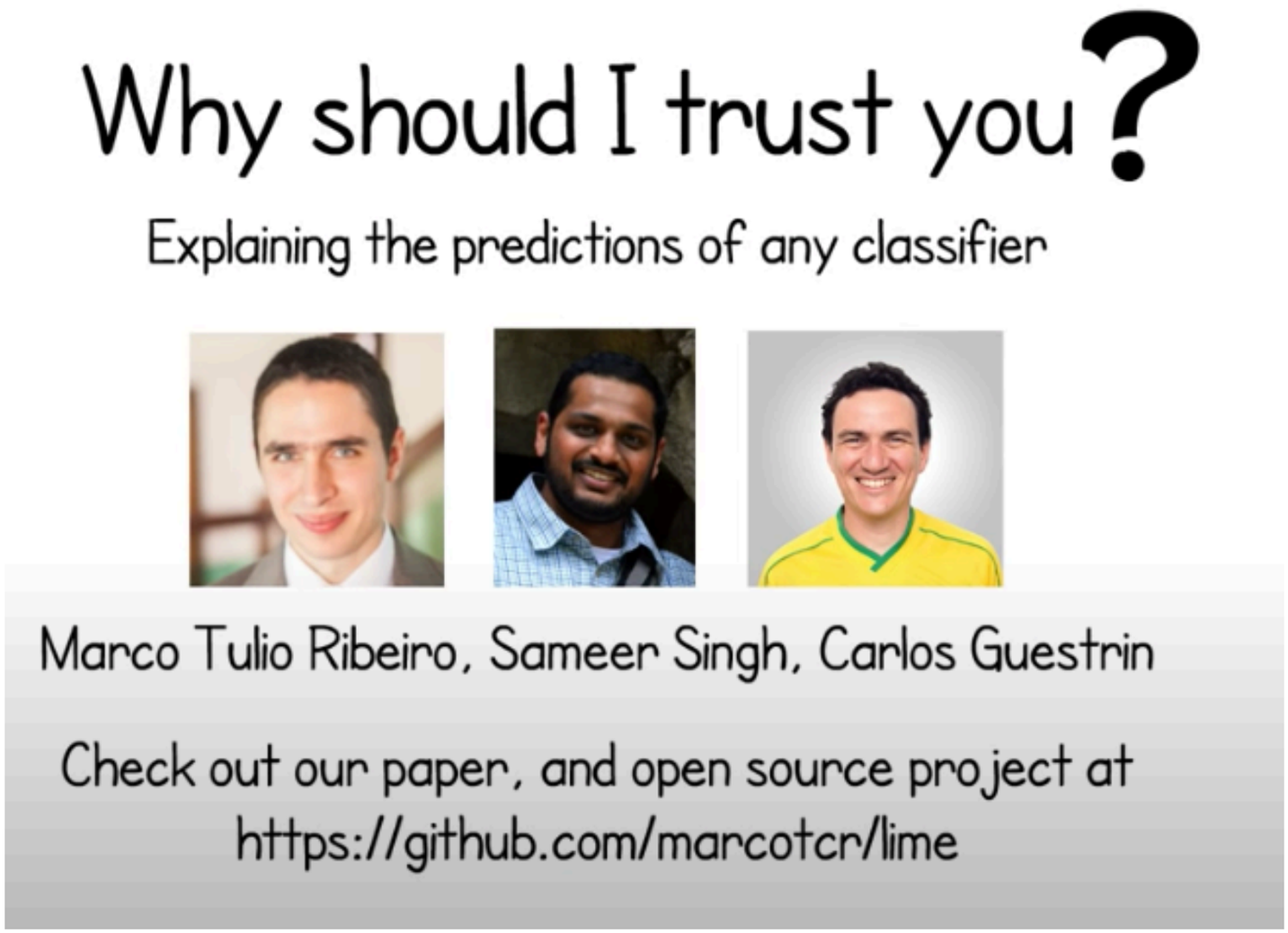
We've built a great snow detector… ☹

slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

r/ai

LIME: Recap

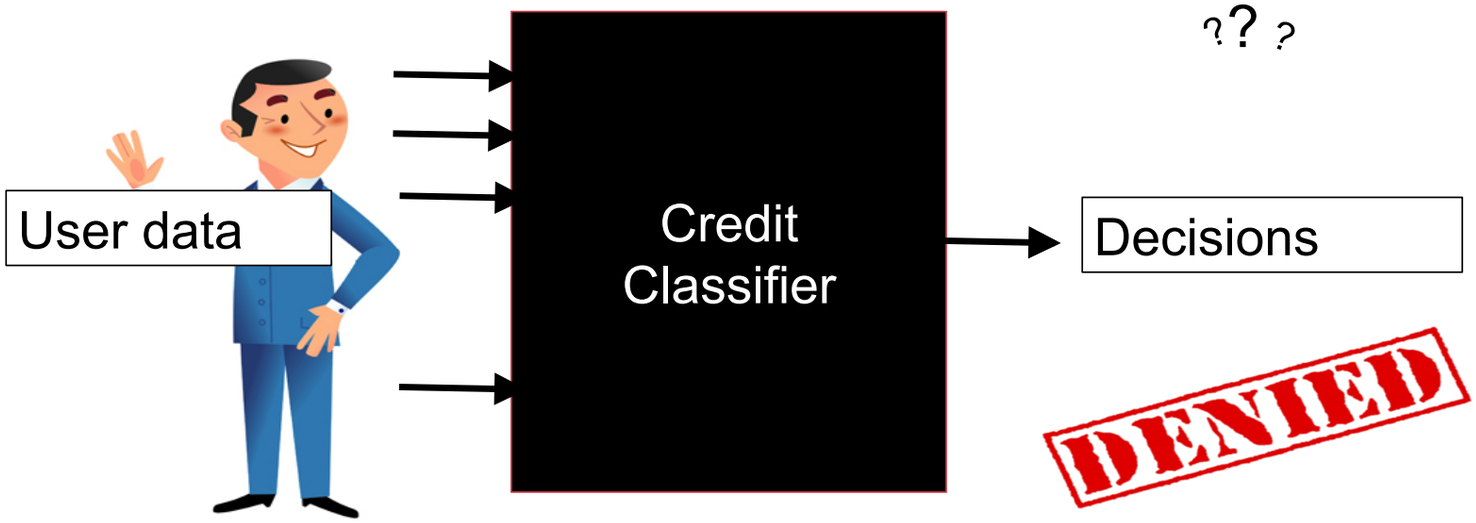


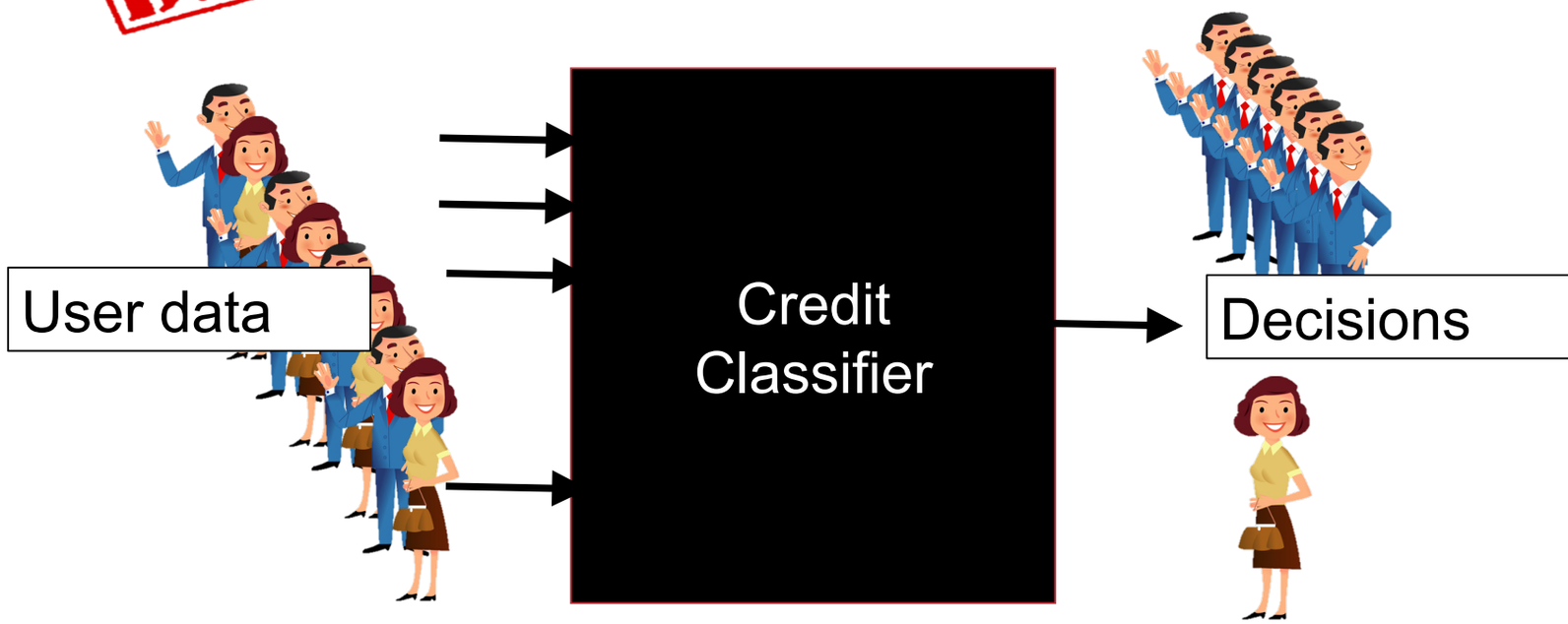
https://www.youtube.com/watch?v=hUnRCxnydCc

[Ribeiro, Singh & Guestrin, 2016]

r/ai

Auditing black-box models





images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

QII: Quantitative Input Influence

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?

[Datta, Sen & Zick, 2016]

r/ai

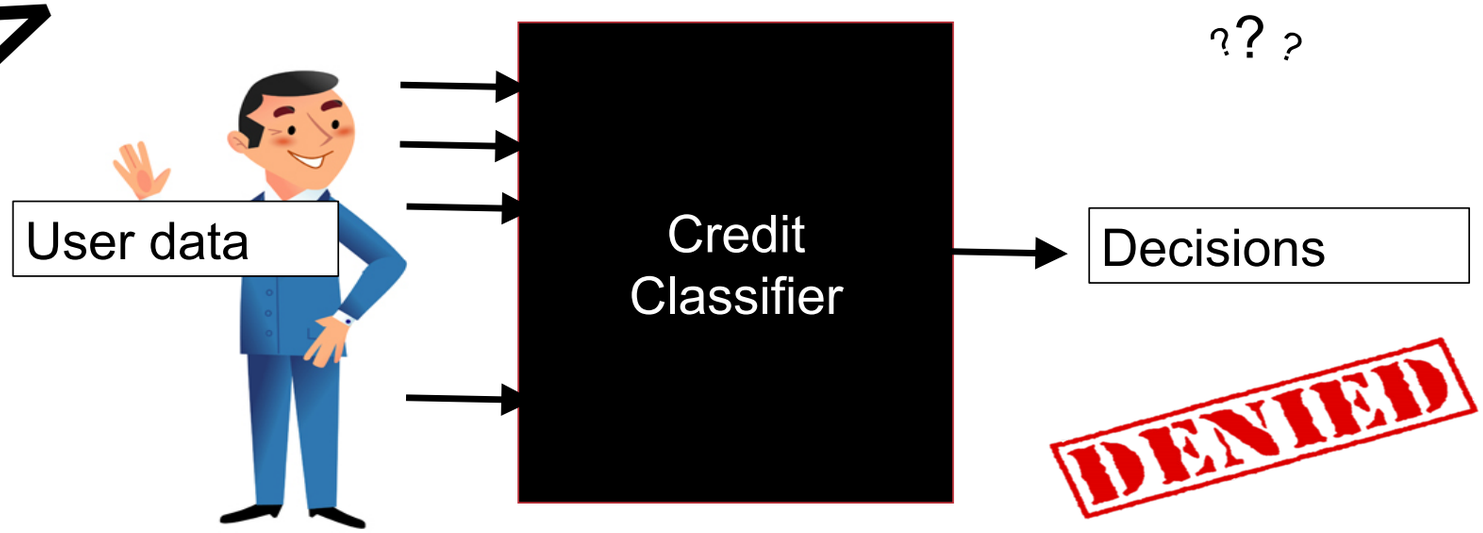
QII: Quantitative Input Influence

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**.

Key ideas

intervene on an input feature, measure its **importance**

aggregate feature importance using its **Shapley value**



images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

Running example

Consider lending decisions by a bank, based on gender, age, education, and income. **Does gender influence lending decisions?**

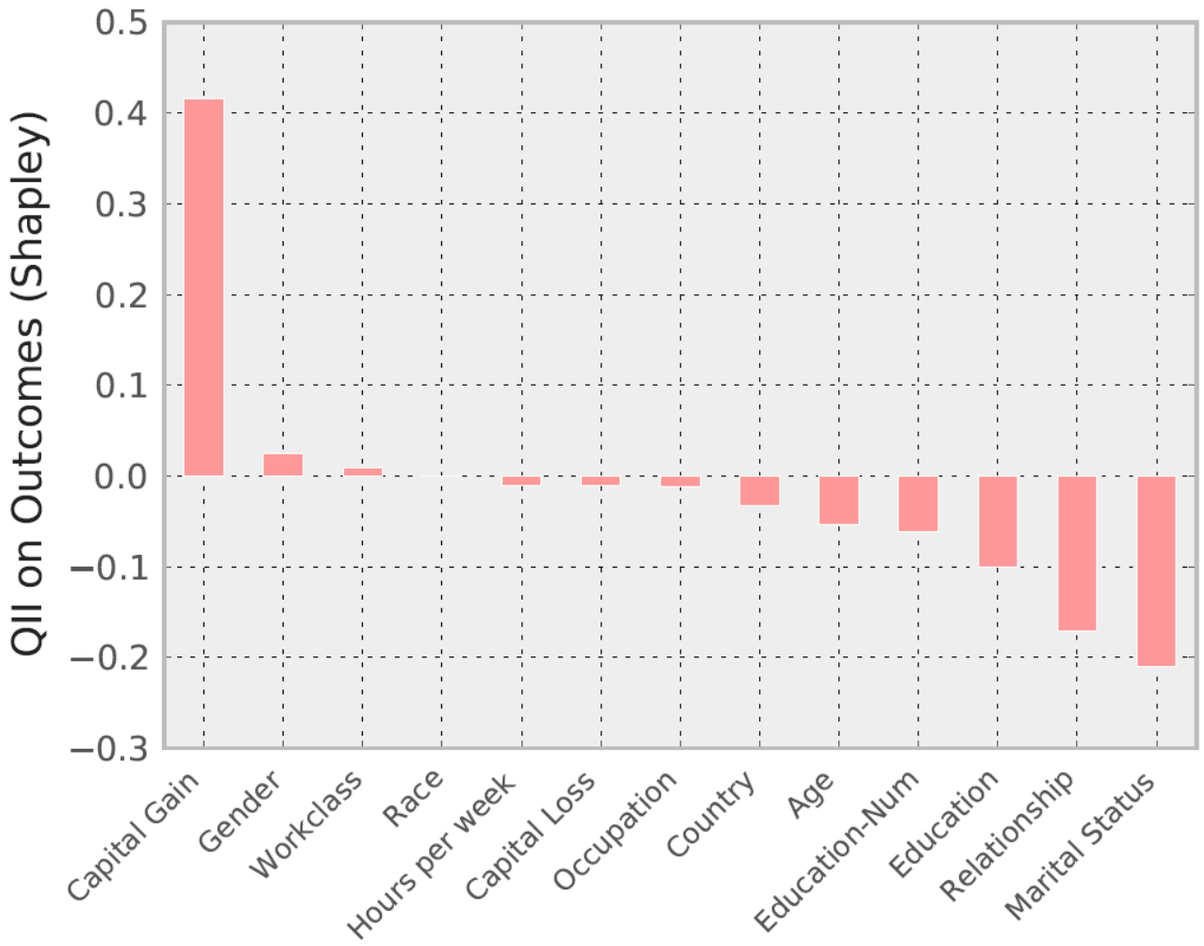
- Observe that 20% of women receive the positive classification.
- To check whether gender impacts decisions, take the input dataset and replace the value of gender in each input profile by drawing it from the uniform distribution: set gender in 50% of the inputs to female and 50% to male.
- If we still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

[Datta, Sen & Zick, 2016]

r/ai

Transparency report: Mr. X

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

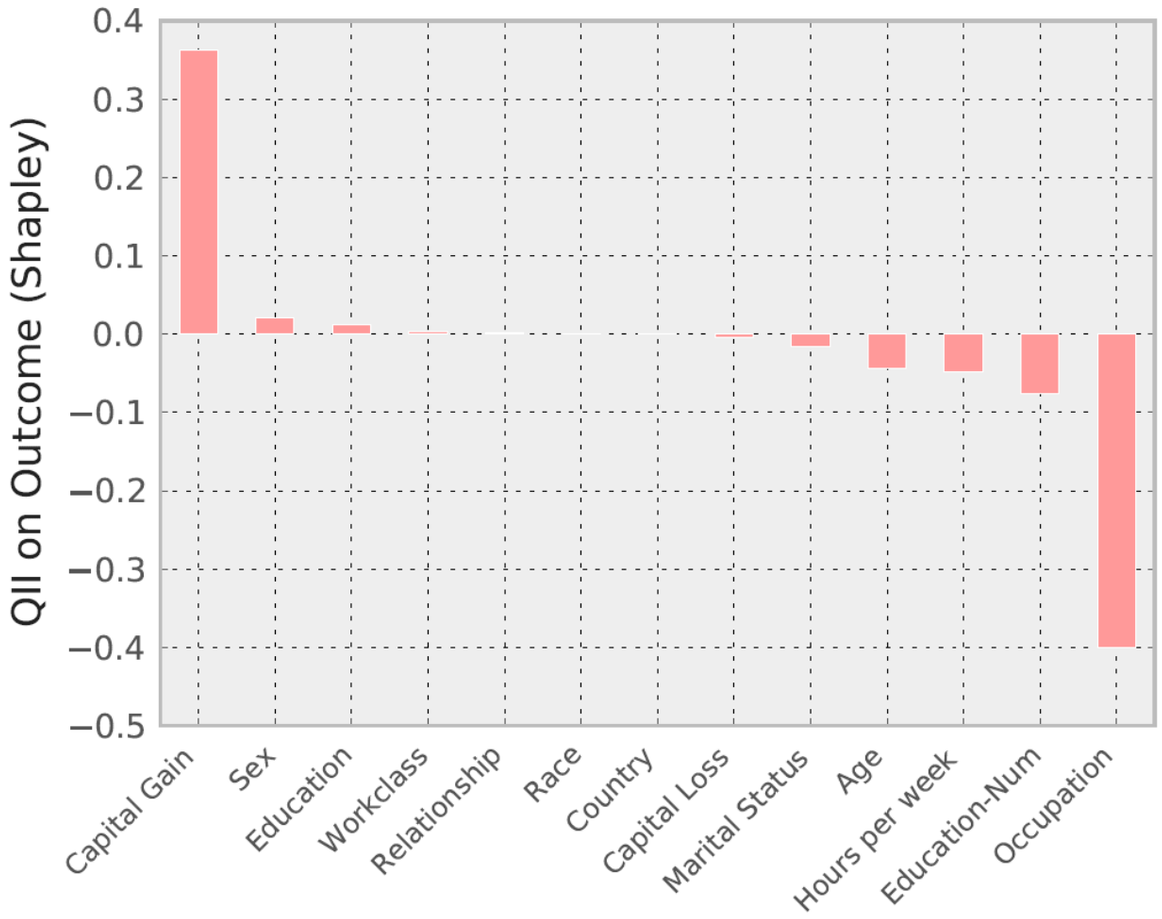
images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

Transparency report: Mr. Y

Explanations for superficially similar individuals can be different



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

images by Anupam Datta

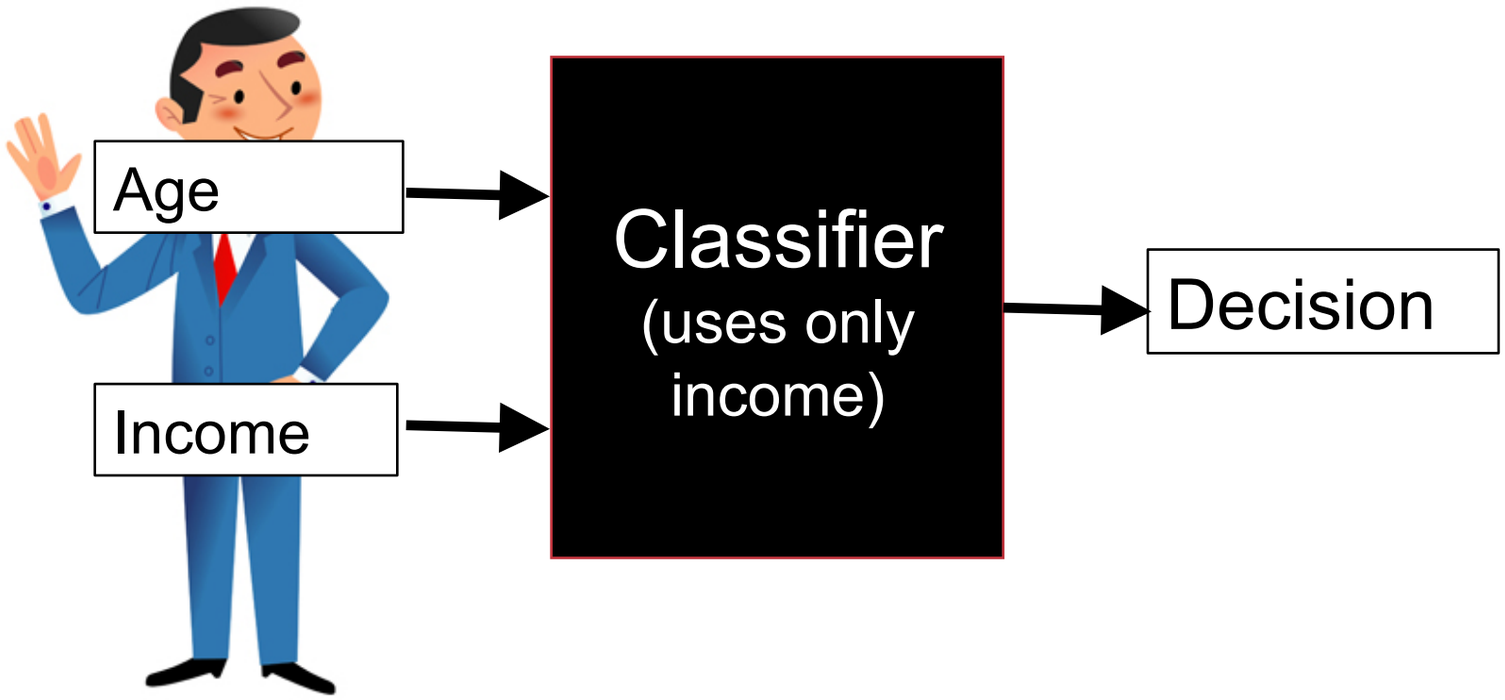
[Datta, Sen & Zick, 2016]

r/ai

Unary QII

images by Anupam Datta

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**.



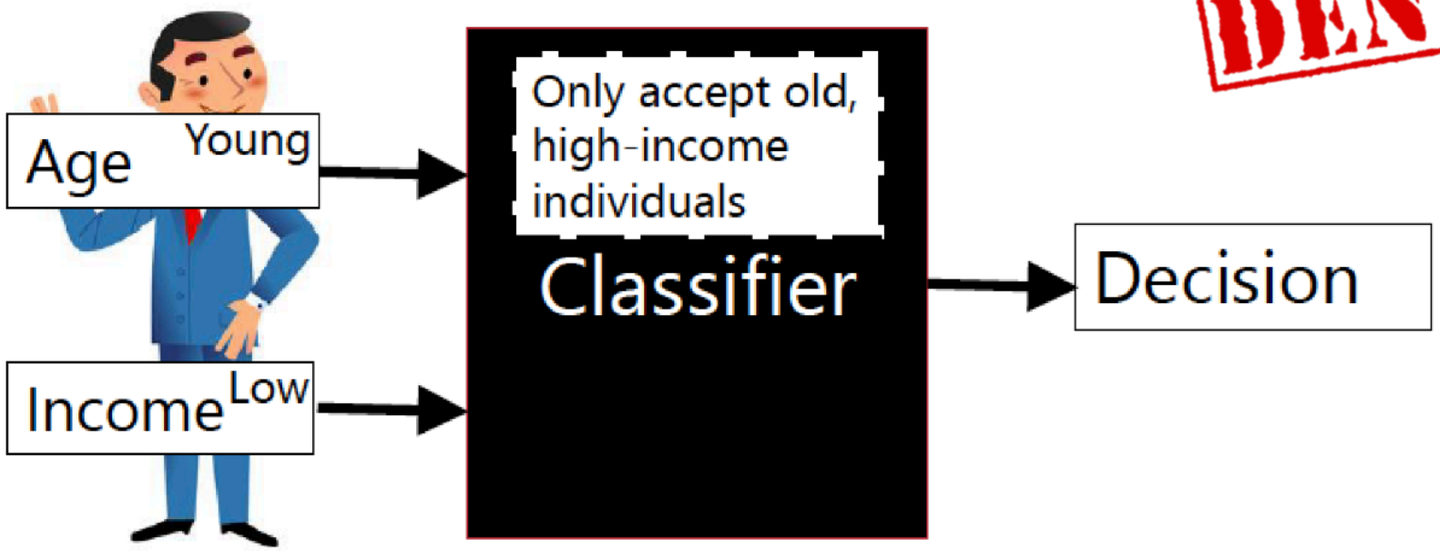
replace features with random values from the population, examine the distribution over outcomes

[Datta, Sen & Zick, 2016]

r/ai

Unary QII

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

images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

Marginal QII

- 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!

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

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

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

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

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

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

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

[Datta, Sen & Zick, 2016]

r/ai

Aggregating influence across sets

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

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

[Datta, Sen & Zick, 2016]

r/ai

Aggregating influence across sets

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

[Datta, Sen & Zick, 2016]

r/ai

QII summary

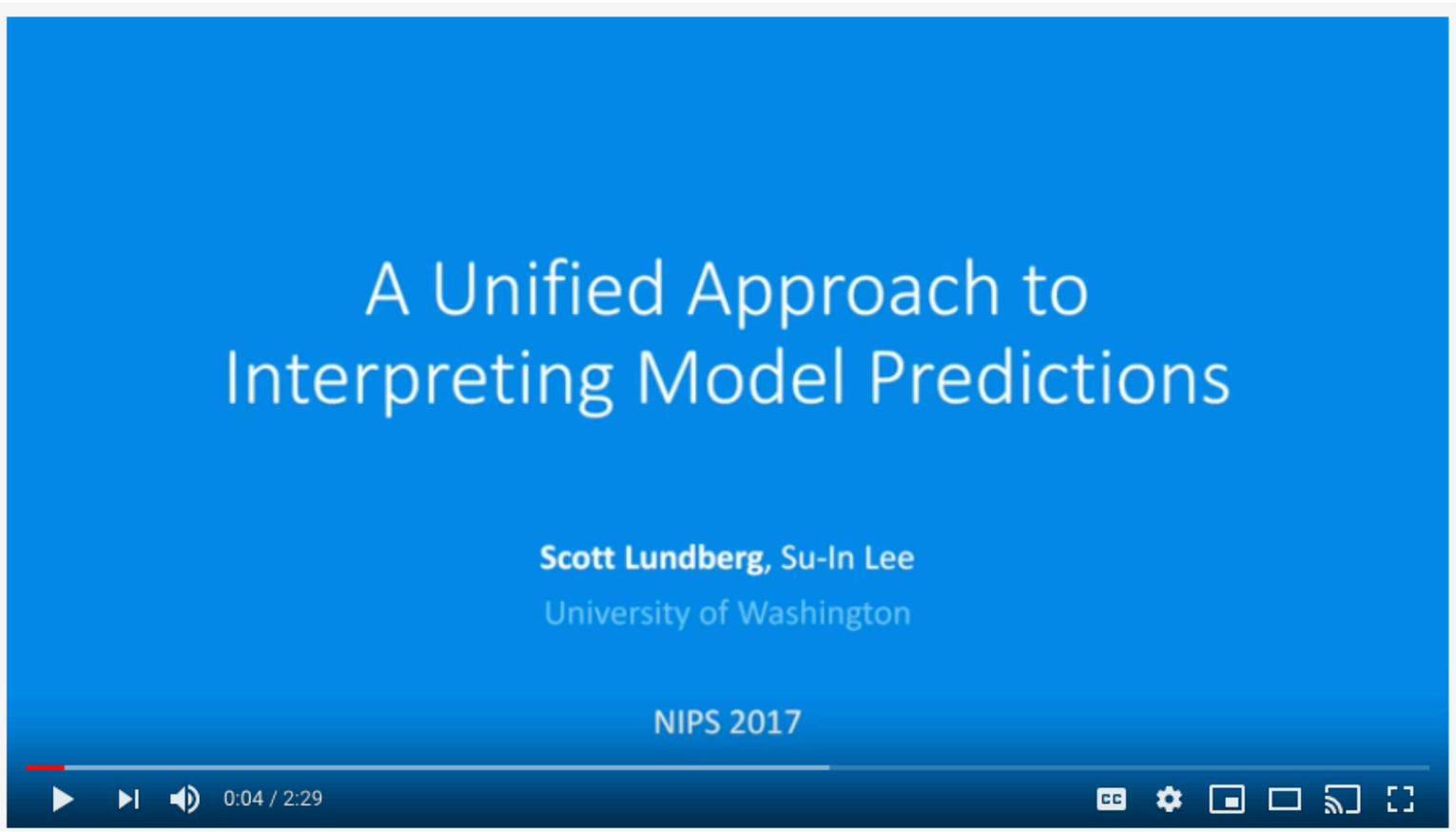
- A principled (and beautiful!) framework for determining the influence of a feature, or a set of features, on a decision
- Works for black-box models, with the assumption that the full set of inputs is available
- Accounts for correlations between features
- “Parametrizes” on what quantity we want to 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**

[Datta, Sen & Zick, 2016]

r/ai

SHAP: Shapley Additive Explanations

A unifying framework for interpreting predictions with “additive feature attribution methods”, including LIME and QII, for **local explanations**



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

[Lundberg & Lee, 2017]

SHAP: Shapley Additive Explanations

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

[Lundberg & Lee, 2017]

r/ai

Additive feature attribution methods

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 i^{th} feature of simplified input **x'**— is missing, then it has no attributable impact for ***x***

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)***

[Lundberg & Lee, 2017]

r/ai

Additive feature attribution methods

README.md

SHAP

https://github.com/slundberg/shap

[Lundberg & Lee, 2017]

r/ai