Responsible Data Science

Transparency & Interpretability

Auditing black-box models

March 5, 2024

Prof. Julia Stoyanovich

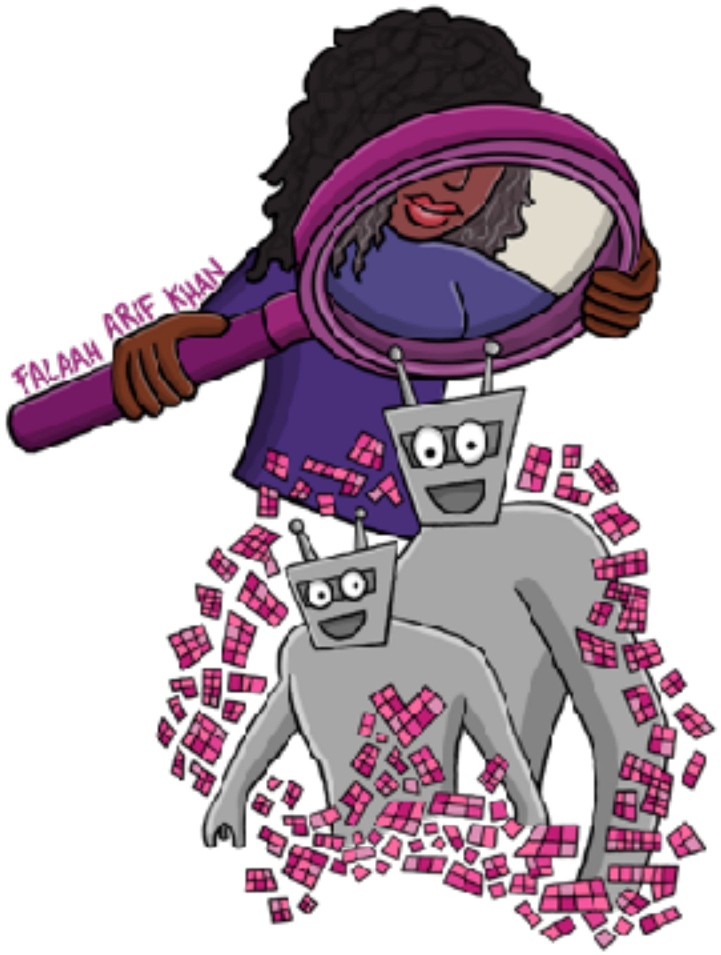
Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

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?

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/SB10001424127887323777204578189391813881534

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 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 the
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/SB10001424127887323777204578189391813881534

Online job ads

theguardian

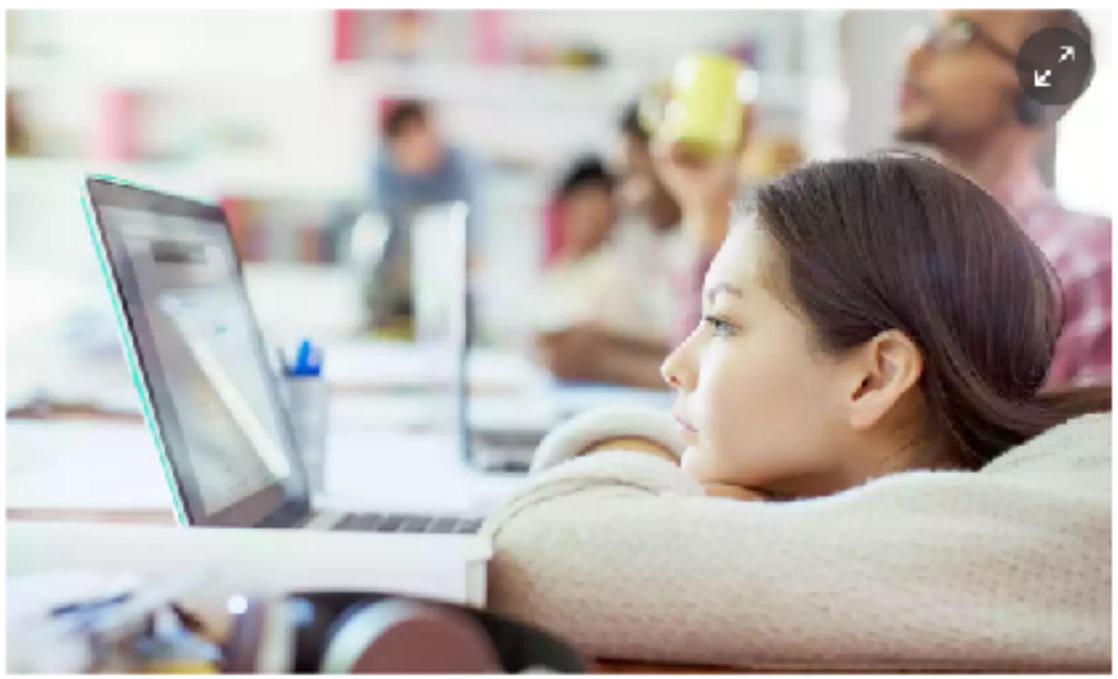
July 2015

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

r/ai

Online job ads

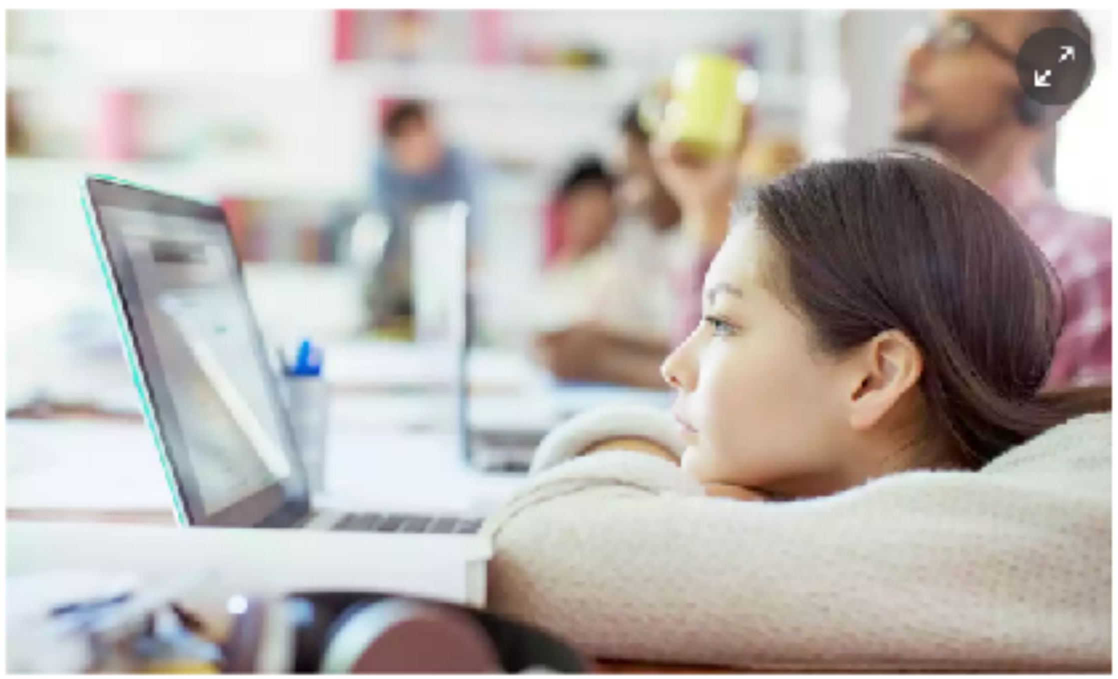
theguardian

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 behavi... demographic...

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

What are we explaining?

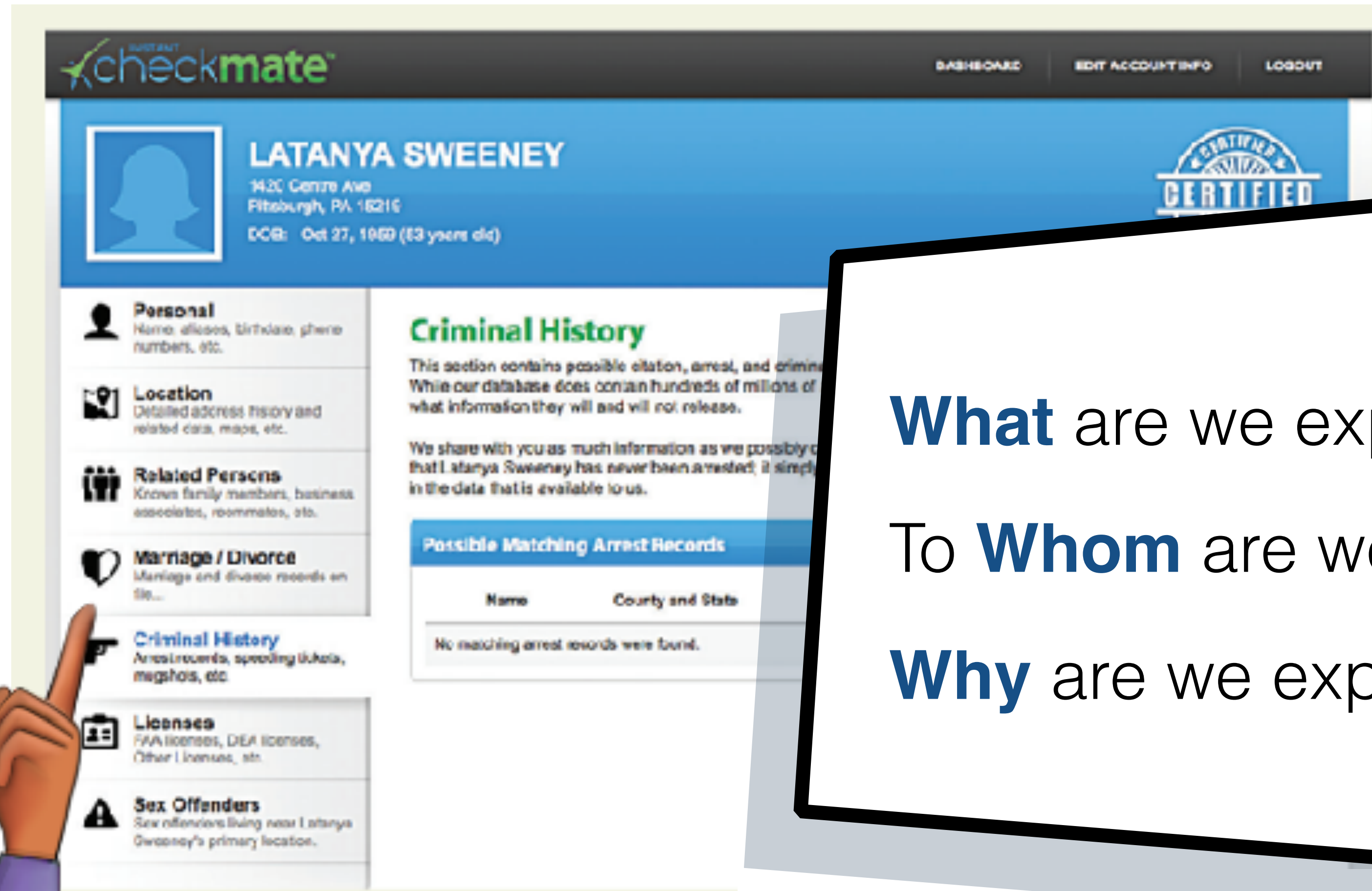
To **Whom** are we explaining?

Why are we explaining?

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

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

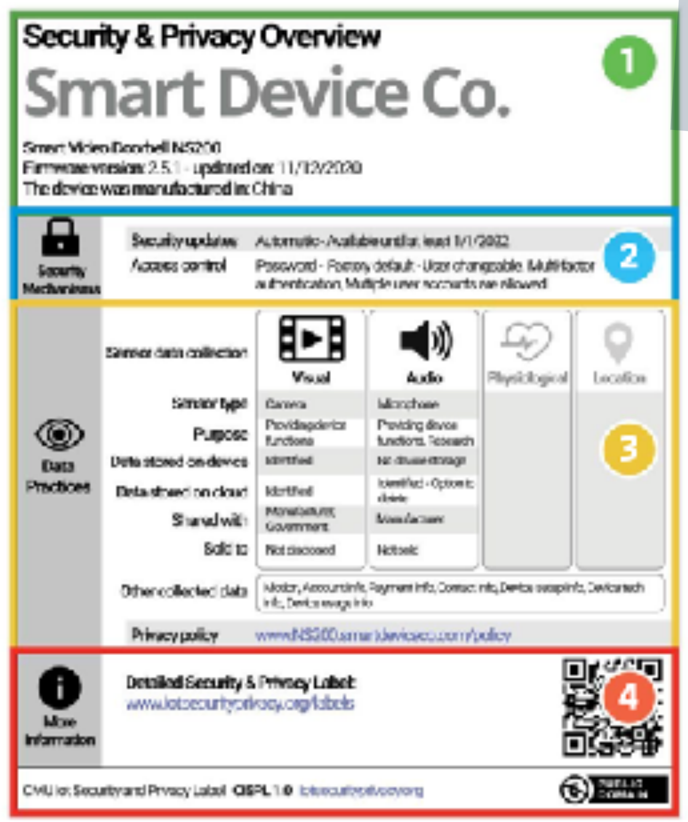
What are we explaining?

To **Whom** are we explaining?

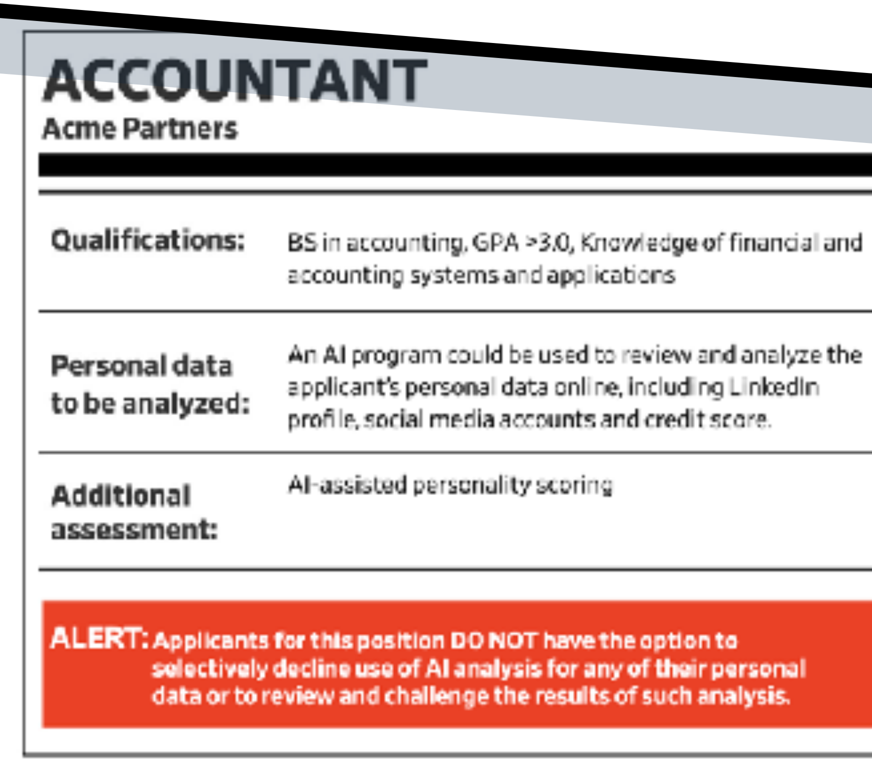
Why are we explaining?

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-



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



https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313 explaining black box models

This week's reading

QII

2016 IEEE Symposium on Security and Privacy

Algorithmic Transparency via Quantitative Input Influence:
Theory and Experiments with Learning Systems

Anupam Datta Shayak Sen Yair Zick
Carnegie Mellon University, Pittsburgh, USA

LIME

"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, University of Washington, Seattle, WA 98105, USA
Sameer Singh, University of Washington, Seattle, WA 98105, USA
Carlos Guestrin, University of Washington, Seattle, WA 98105, USA

This week's reading

SHAP


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

Abstract

1 Introduction

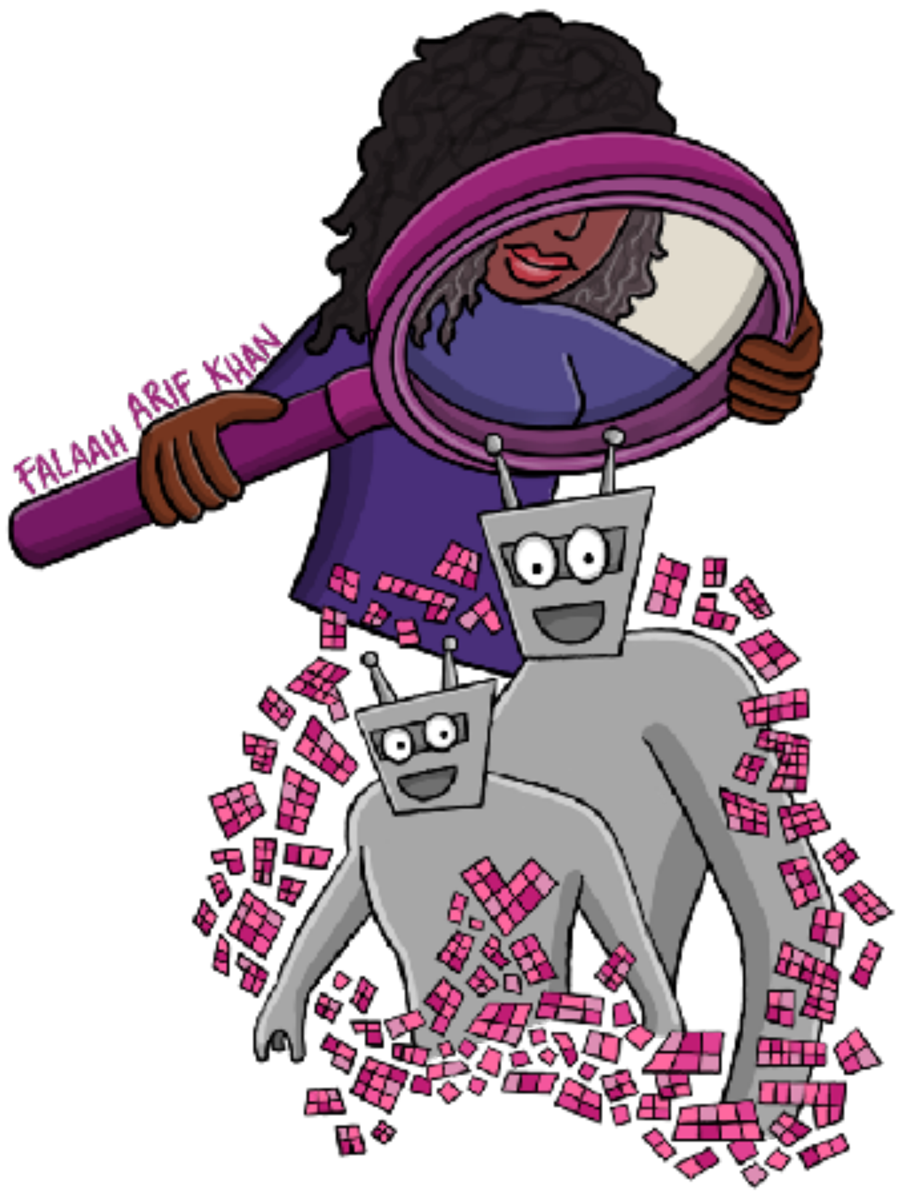
ShaRP


ShaRP: Explaining Rankings with Shapley Values

Abstract

1 Introduction

What are we explaining?



How does a system work?

How **well** does a system work?

What does a system do?

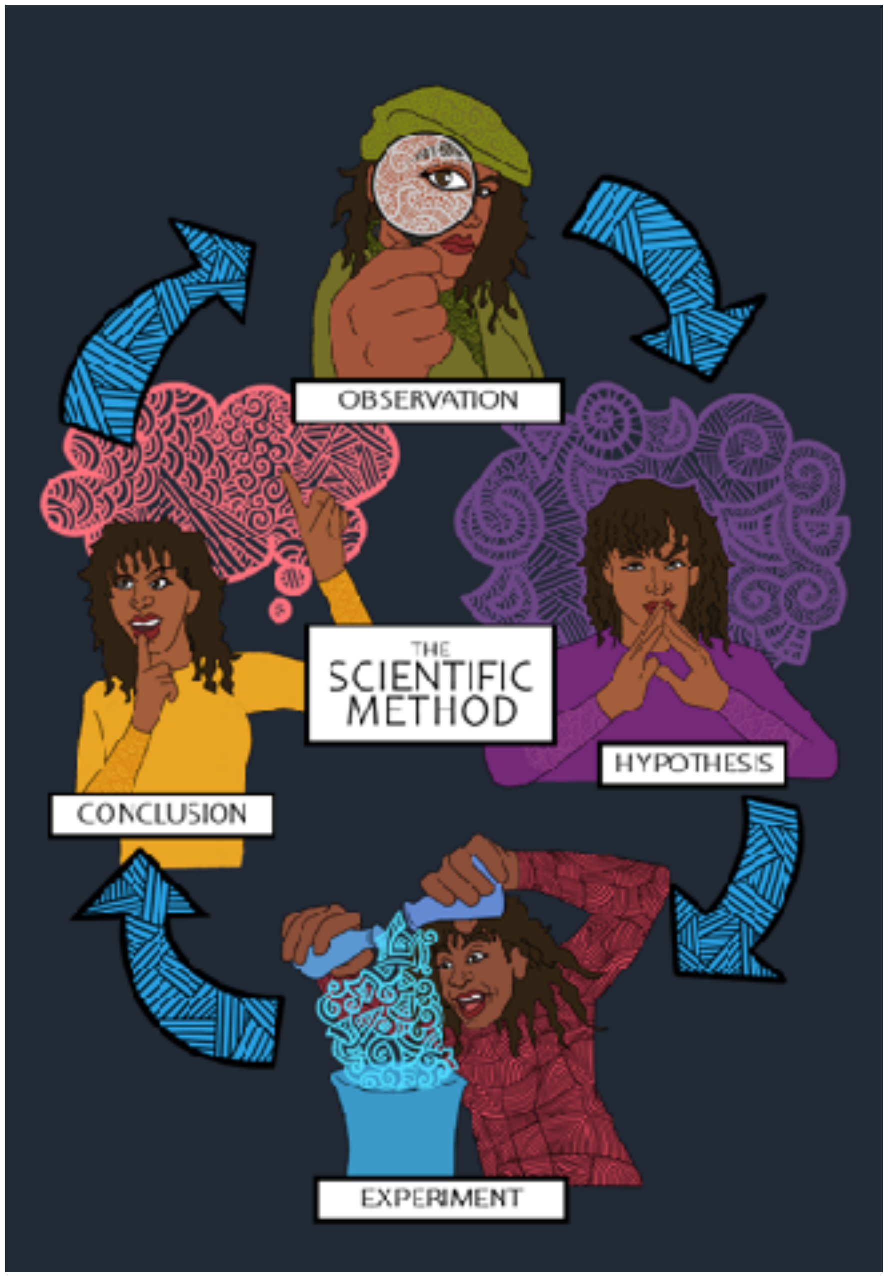
Why was I ___ (mis-diagnosed / not offered a discount / denied credit) ?

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?

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

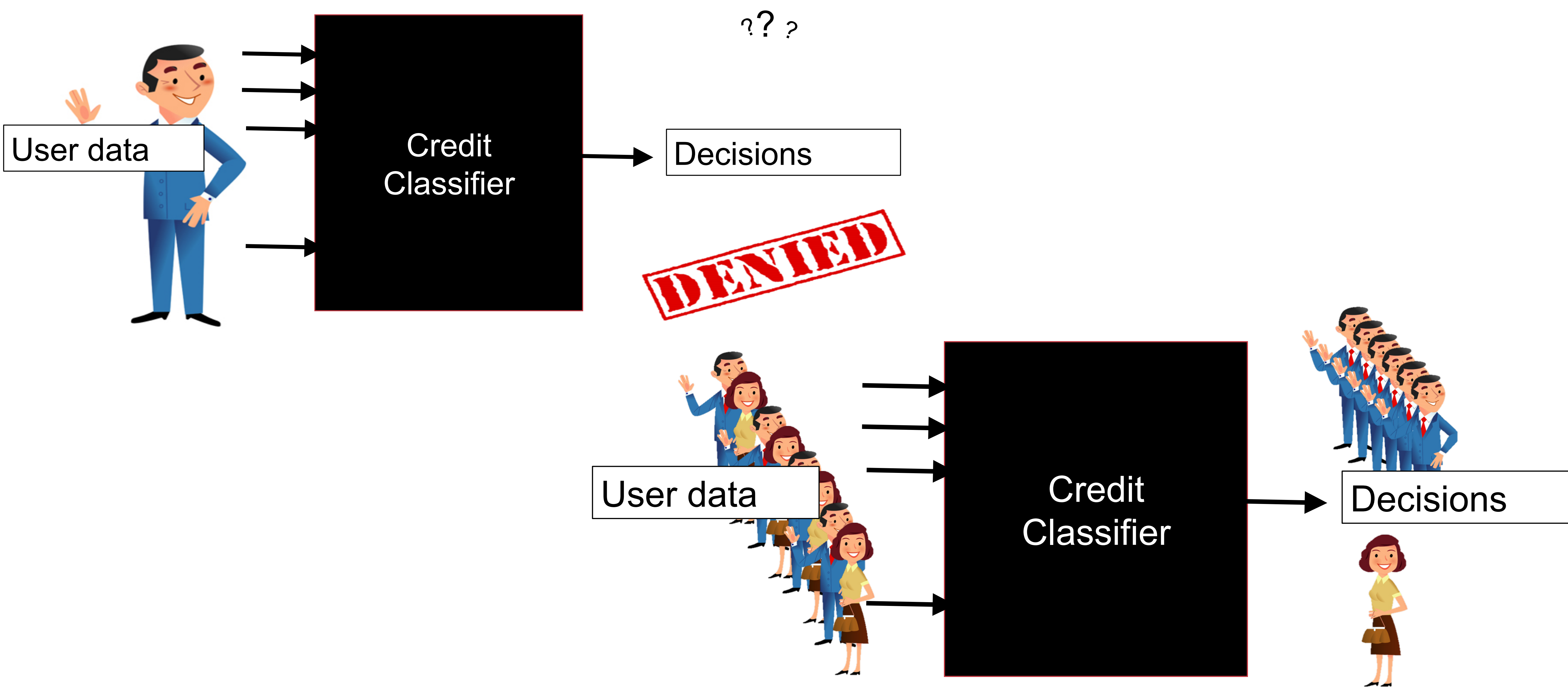
If you're Native American, Facebook might think your name is fake.

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.

r/ai

QII: Auditing black-box models

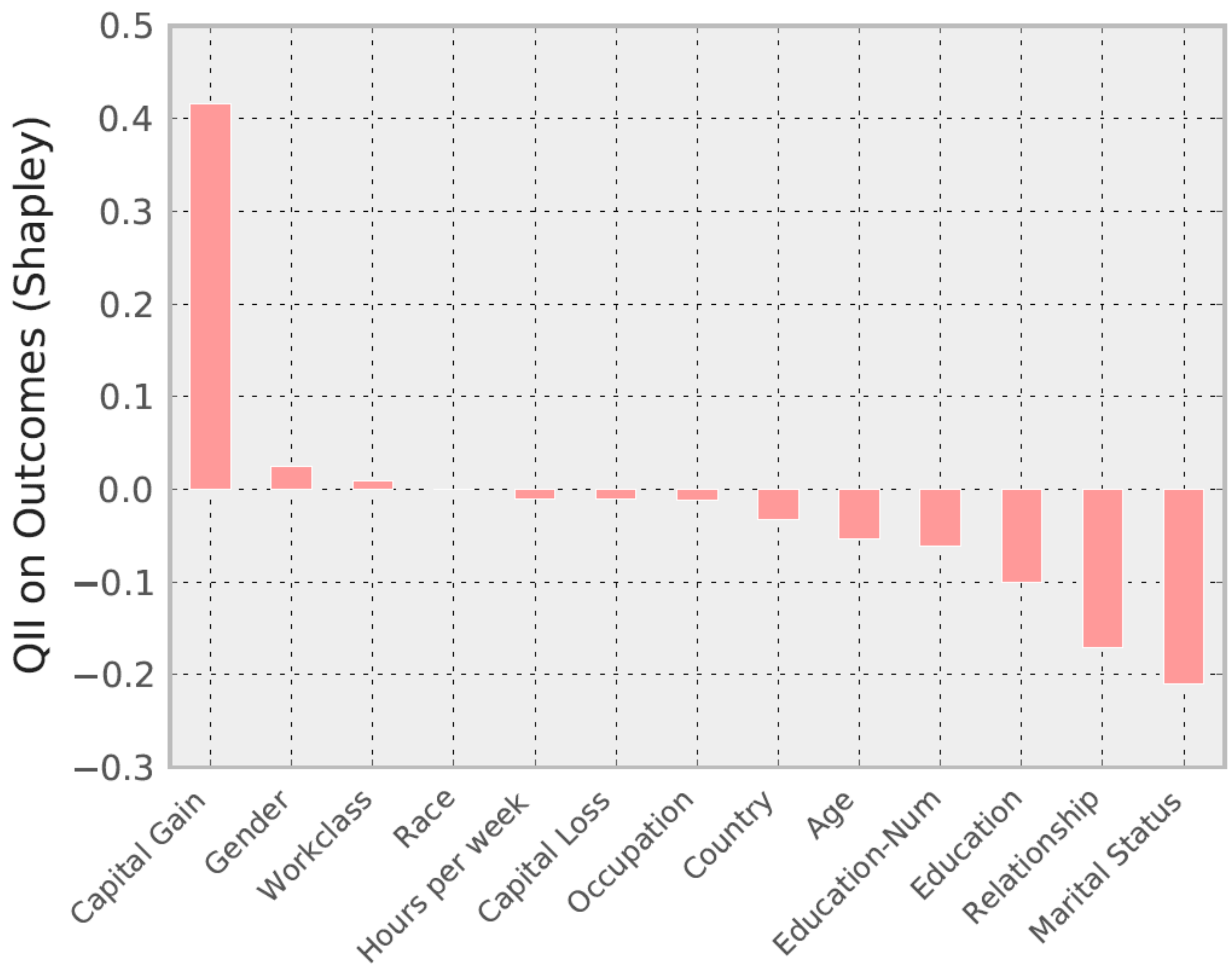


images by Anupam Datta

[Datta, Sen & Zick, 2016]

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

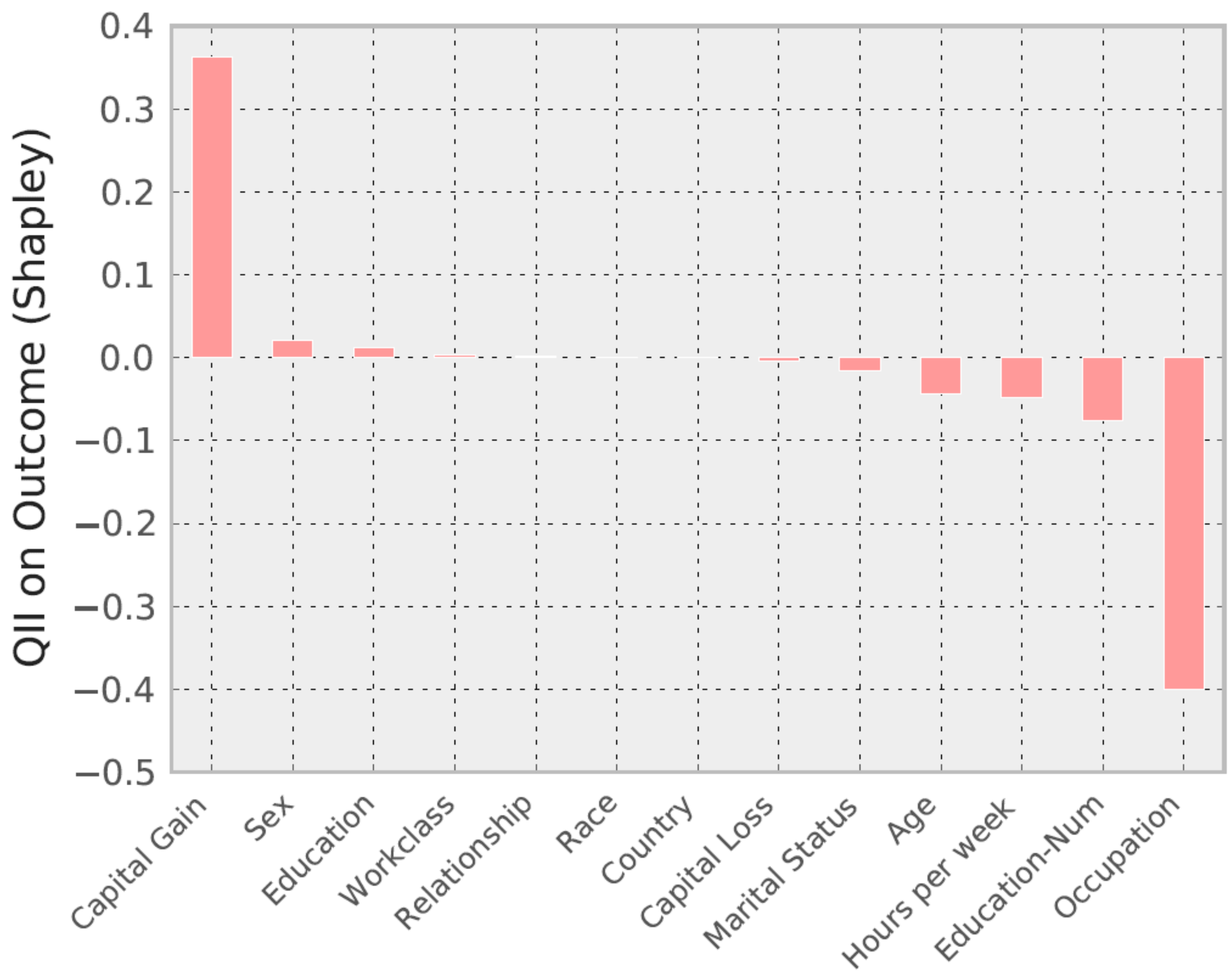
income

images by Anupam Datta

[Datta, Sen & Zick, 2016]

Transparency report: Mr. Y

Explanations for superficially similar individuals can be different



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]

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

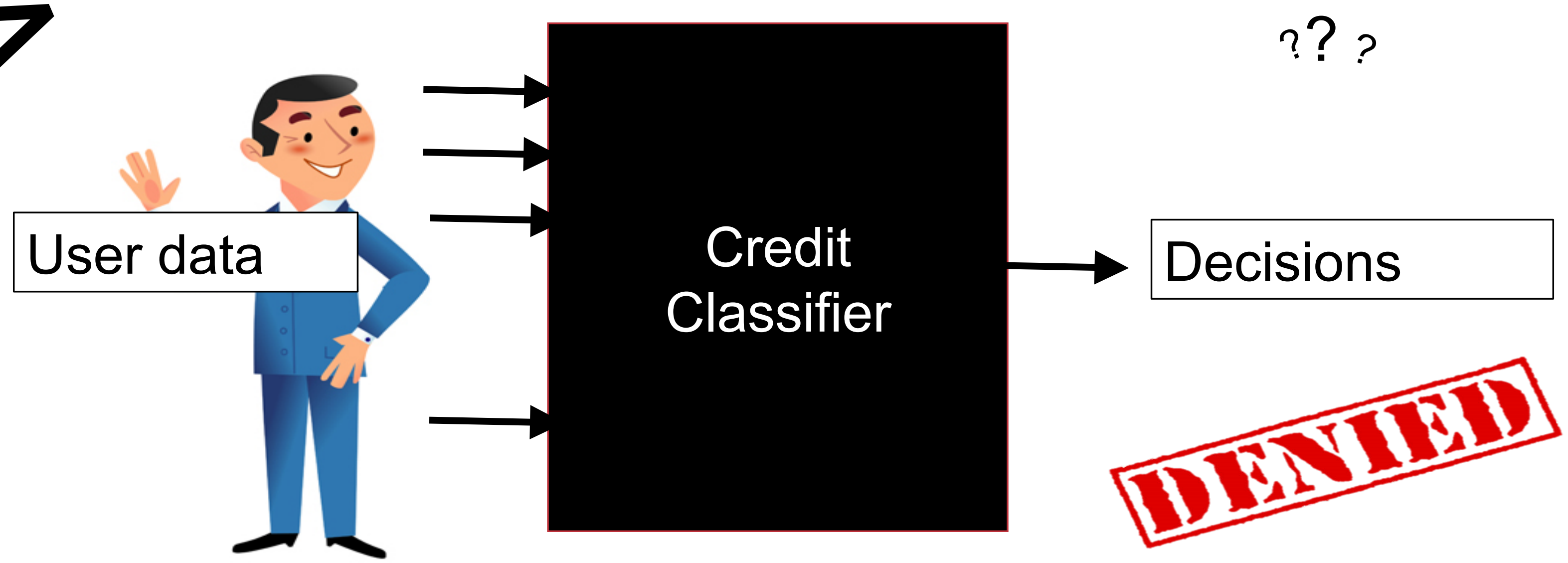
QII: Quantitative Input Influence

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]

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**

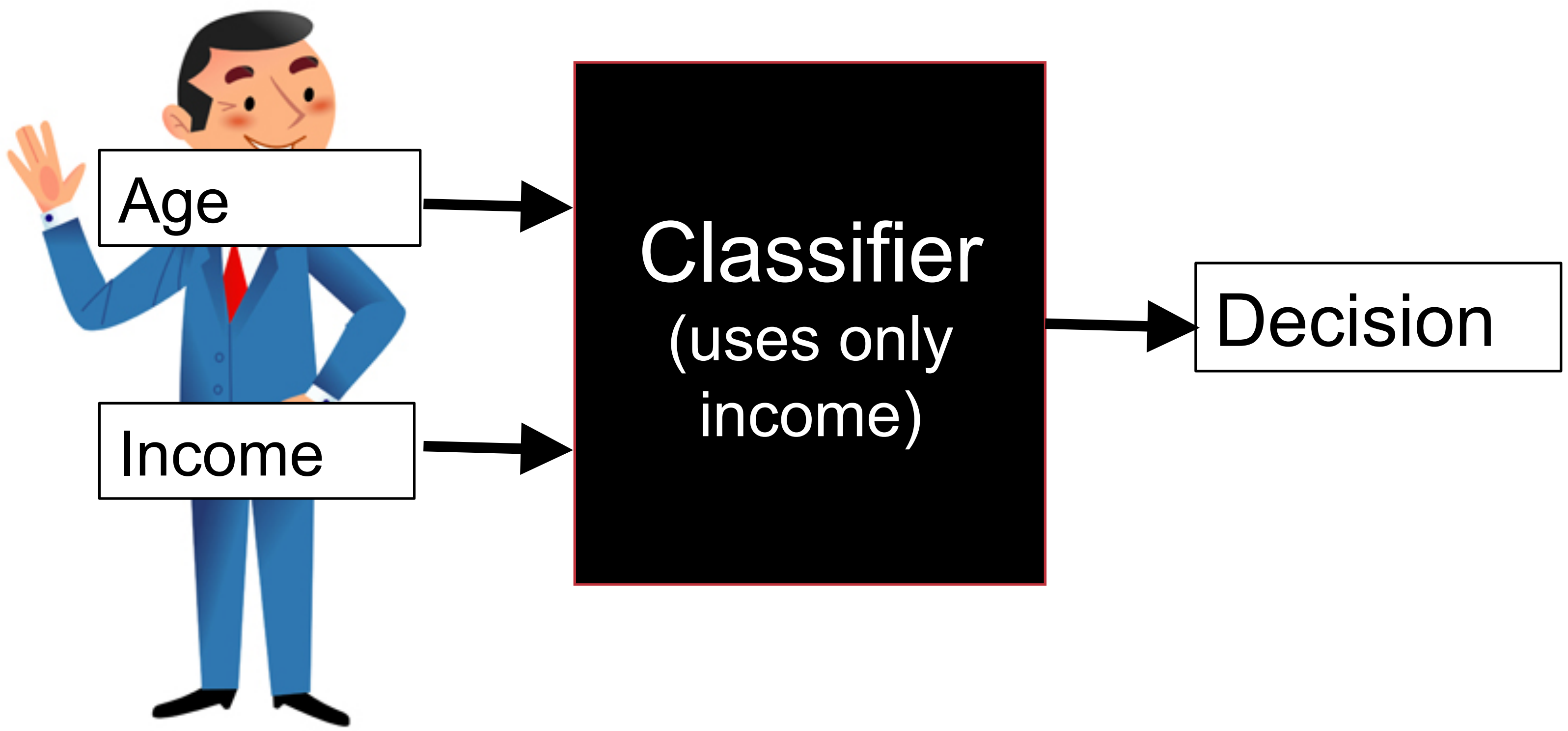
[Datta, Sen & Zick, 2016]

r/ai

Unary QII

images by Anupam Datta

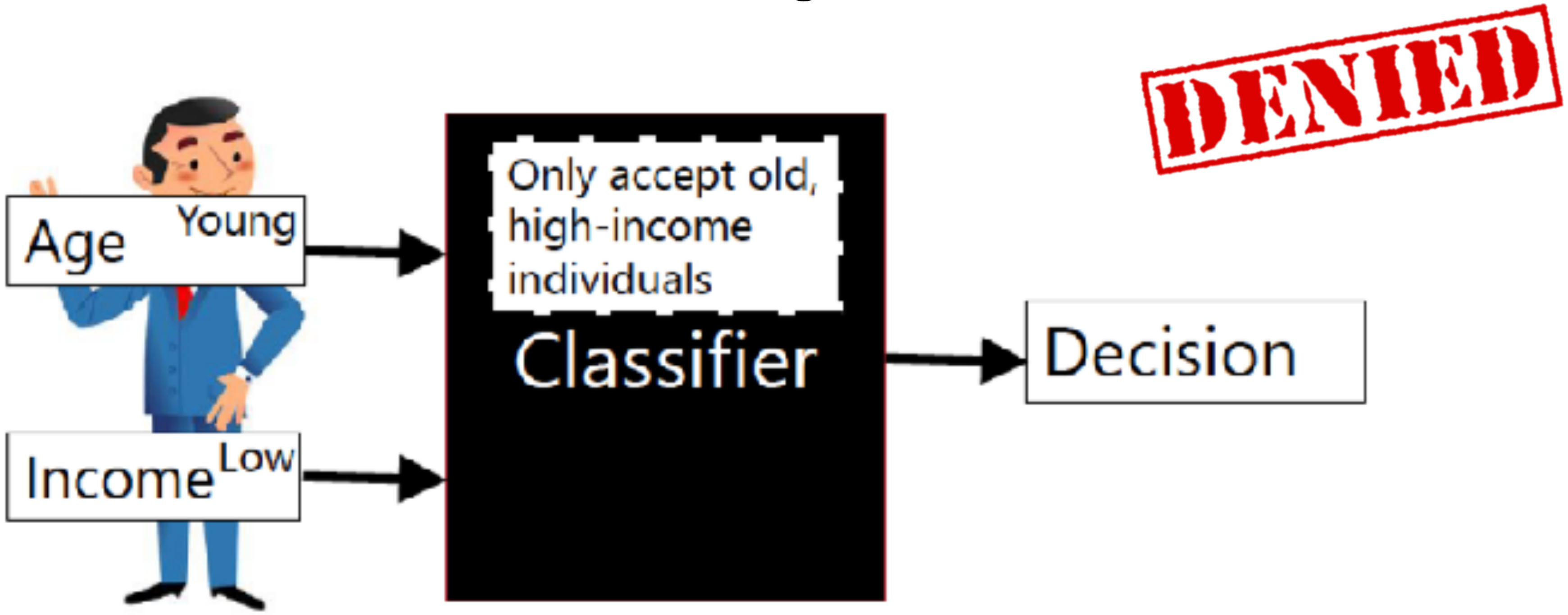
For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.



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

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

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!

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

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

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

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

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

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

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

[Datta, Sen & Zick, 2016]

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

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]

ShaRP: Shapley Values for Rankings & Preferences

name	gpa	sat	essay	f	g
Bob	4	5	5	4.6	5
Cal	4	5	5	4.6	5
Dia	5	4	4	4.4	4
Eli	4	5	3	4.2	3
Fay	5	4	3	4.2	3
Kat	5	4	2	4.0	2
Leo	4	4	3	3.8	3
Osi	3	3	3	3.0	3

(a)

$r_{\mathcal{D},f}$
Bob
Cal
Dia
Eli
Fay
Kat
Leo
Osi

(b)

$r_{\mathcal{D},g}$
Bob
Cal
Dia
Eli
Fay
Leo
Osi
Kat

(c)

Figure 1: (a) Dataset $\mathcal{D}$ of college applicants, scored on gpa, sat, and $essay$. (b) Ranking $r_{\mathcal{D},f}$ of $\mathcal{D}$ on $f = 0.4 \times gpa + 0.4 \times sat + 0.2 \times essay$; the highlighted top-4 candidates will be interviewed and potentially admitted. (c) Ranking $r_{\mathcal{D},g}$ on $g = 1.0 \times essay$; the top-4 coincides with that of $r_{\mathcal{D},f}$, signifying that $essay$ has the highest importance for f, despite carrying the lowest weight.

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

r/ai

Computation of feature importance

Algorithm 1 Feature importance for per-item outcomes

Input: Dataset $\mathcal{D}$, item $\mathbf{v}$, number of samples m, $\iota()$
Output: Shapley values $\phi(\mathbf{v})$ of $\mathbf{v}$'s features

$\phi(\mathbf{v}) = \langle 0, \ldots, 0 \rangle$
for $i \in \mathcal{A}$ **do**
    for $\mathcal{S} \subseteq \mathcal{A} \setminus \{i\}$ **do**
        $\mathbf{U} \sim \mathcal{D} \setminus \mathbf{v}, m$
        $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A} \setminus \mathcal{S}} \mathbf{U}_{\mathcal{S}}$
        $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A} \setminus \{\mathcal{S} \cup i\}} \mathbf{U}_{\mathcal{S} \cup i}$
        $\phi_{i_{\mathcal{S}}}(\mathbf{v}) = \iota(\mathbf{U}_1, \mathbf{U}_2)$
        $\phi_i(\mathbf{v}) = \phi_i(\mathbf{v}) + \frac{1}{d} \frac{1}{\binom{d-1}{|S|}} \phi_{i_{\mathcal{S}}}(\mathbf{v})$
    end for
end for
return $\phi(\mathbf{v})$

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

Computing a specific QoI (the iota function)

Algorithm 2 ι_{Rank}

Input: Dataset $\mathcal{D}$, scoring function f, item $\mathbf{v}$, $\mathbf{U}_1$, $\mathbf{U}_2$, number of samples m
Output: ϕ

$\phi = 0$
for $i \in \{1, \ldots, m\}$ **do**
  $\mathbf{u}_1 = \mathbf{U}_1(i)$
  $\mathbf{u}_2 = \mathbf{U}_2(i)$
  $\mathcal{D}_1 = \mathcal{D} \setminus \{\mathbf{v}\} \cup \{\mathbf{u}_1\}$
  $\mathcal{D}_2 = \mathcal{D} \setminus \{\mathbf{v}\} \cup \{\mathbf{u}_2\}$
  $\phi = \phi + r^{-1}_{\mathcal{D}_2, f}(\mathbf{u}_2) - r^{-1}_{\mathcal{D}_1, f}(\mathbf{u}_1)$
end for
return $\phi / |\mathbf{U}_1|$

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

r/ai

Example dataset: CS Ranking

CSRankings: Computer Science Rankings

CSRankings is a metrics-based ranking of top computer science institutions around the world. **Click on a triangle** (▶) to expand areas or institutions. **Click on a name** to go to a faculty member's home page. **Click on a chart icon** (the 📊 after a name or institution) to see the distribution of their publication areas as a bar chart. **Click on a Google Scholar icon** to see publications, and **click on the DBLP logo** to go to a DBLP entry. *Applying to grad school? Read this first.* For info on grad stipends, check out CSStipendRankings.org. **Do you find CSrankings useful? Sponsor CSrankings on GitHub.**

Rank institutions in USA by publications from 2014 to 2024

All Areas [off | on]

AI [off | on]

- ▶ Artificial intelligence ☑
- ▶ Computer vision ☑
- ▶ Machine learning ☑
- ▶ Natural language processing ☑
- ▶ The Web & information retrieval ☑

Systems [off | on]

- ▶ Computer architecture ☑
- ▶ Computer networks ☑
- ▶ Computer security ☑
- ▶ Databases ☑
- ▶ Design automation ☑
- ▶ Embedded & real-time systems ☑
- ▶ High-performance computing ☑
- ▶ Mobile computing ☑
- ▶ Measurement & perf. analysis ☑
- ▶ Operating systems ☑
- ▶ Programming languages ☑
- ▶ Software engineering ☑

Theory [off | on]

- ▶ Algorithms & complexity ☑
- ▶ Cryptography ☑
- ▶ Logic & verification ☑

Interdisciplinary Areas [off | on]

- ▶ Comp. bio & bioinformatics ☑
- ▶ Computer graphics ☑
- ▶ Computer science education ☑
- ▶ Economics & computation ☑
- ▶ Human-computer interaction ☑
- ▶ Robotics ☑
- ▶ Visualization ☑

#	Institution	Count	Faculty
1	▶ Carnegie Mellon University	19.2	173
2	▶ Univ. of Illinois at Urbana-Champaign	13.9	112
3	▶ Univ. of California - San Diego	12.3	128
4	▶ Georgia Institute of Technology	11.0	143
5	▶ Massachusetts Institute of Technology	10.2	92
6	▶ Univ. of California - Berkeley	10.2	95
7	▶ University of Michigan	10.1	100
7	▶ University of Washington	10.1	81
9	▶ Stanford University	9.8	68
10	▶ Cornell University	9.3	83
11	▶ University of Maryland - College Park	8.8	88
12	▶ Northeastern University	7.7	87
13	▶ Purdue University	7.1	74
14	▶ University of Wisconsin - Madison	7.0	70
15	▶ University of Texas at Austin	6.9	50
16	▶ University of Pennsylvania	6.7	74
17	▶ Columbia University	6.6	59
18	▶ Princeton University	6.4	59
19	▶ New York University	6.2	72
20	▶ Univ. of California - Los Angeles	5.5	43
20	▶ University of Massachusetts Amherst	5.5	60
20	▶ University of Southern California	5.5	61

Different reasons for similar ranked outcomes

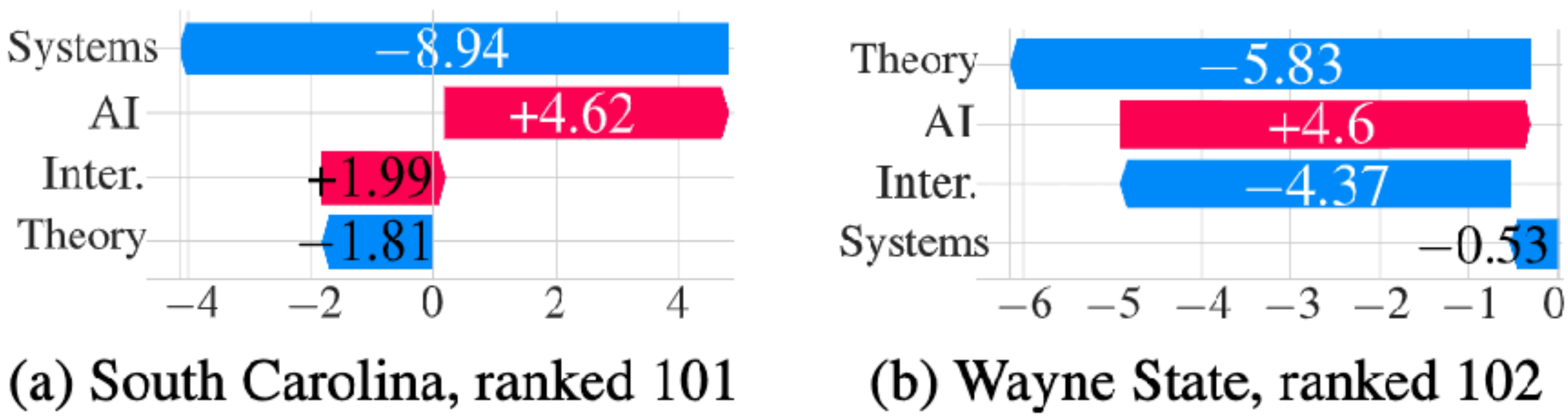


Figure 4: Feature contributions to rank QoI for two departments.

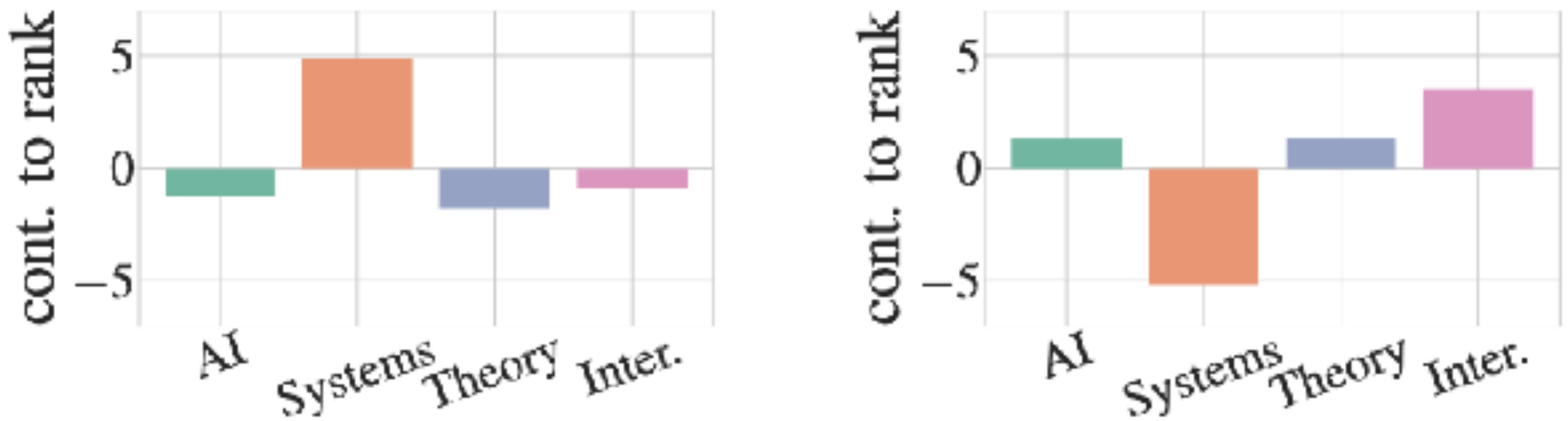
[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

r/ai

Comparing Georgia Tech, Stanford & UMich

Institution	AI	Systems	Theory	Inter.	Rank
Georgia Tech	28.5	7.8	6.9	10.2	5
Stanford	36.7	5.4	13.3	11.5	6
UMich	30.4	9.0	9.3	5.9	7

(b) Feature values and rank of three highly ranked departments: Georgia Tech, Stanford, and UMich.



(c) Pairwise QoI explaining that Georgia Tech ranks higher than Stanford because of its relative strength in Systems.

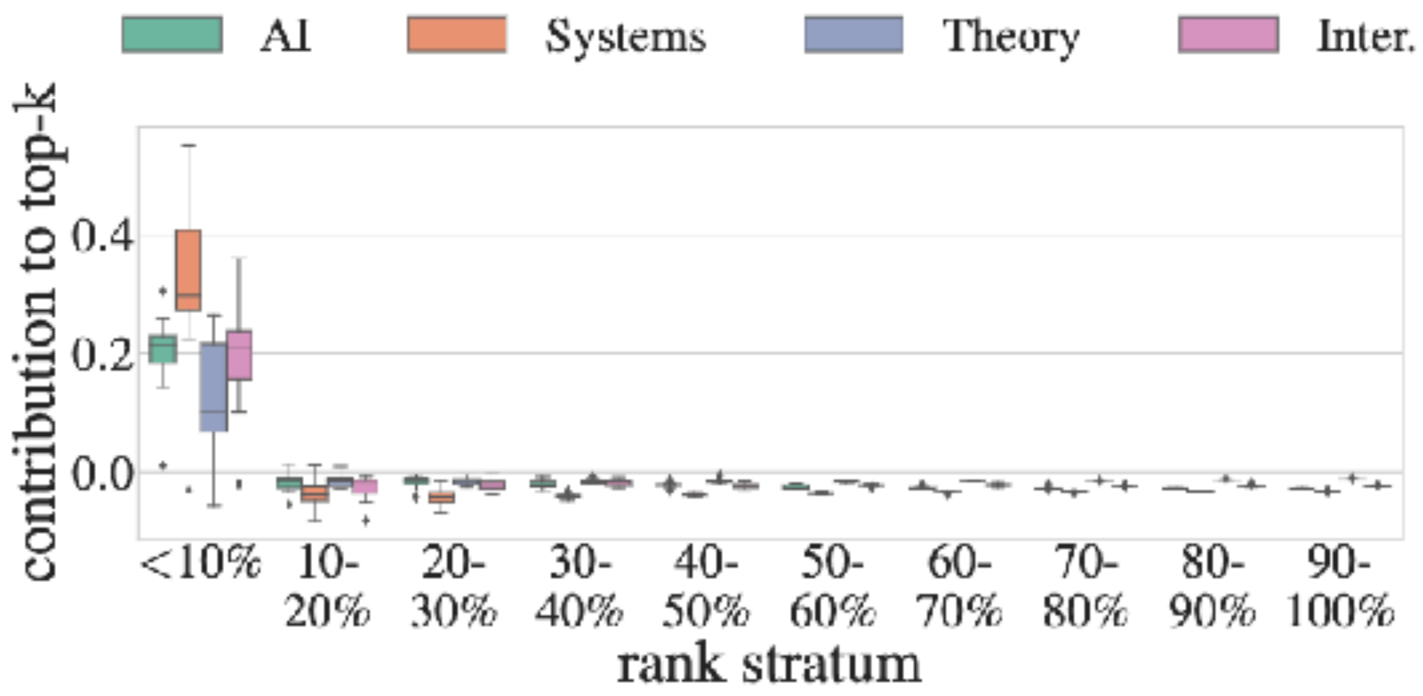
(d) Pairwise QoI explaining that Stanford ranks higher than UMich despite Stanford's relative weakness in Systems.

Figure 3: Feature importance for the top-k QoI for CS Rankings, with further analysis of 3 departments using Pairwise QoI.

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

r/ai

Aggregates feature importance by rank stratum



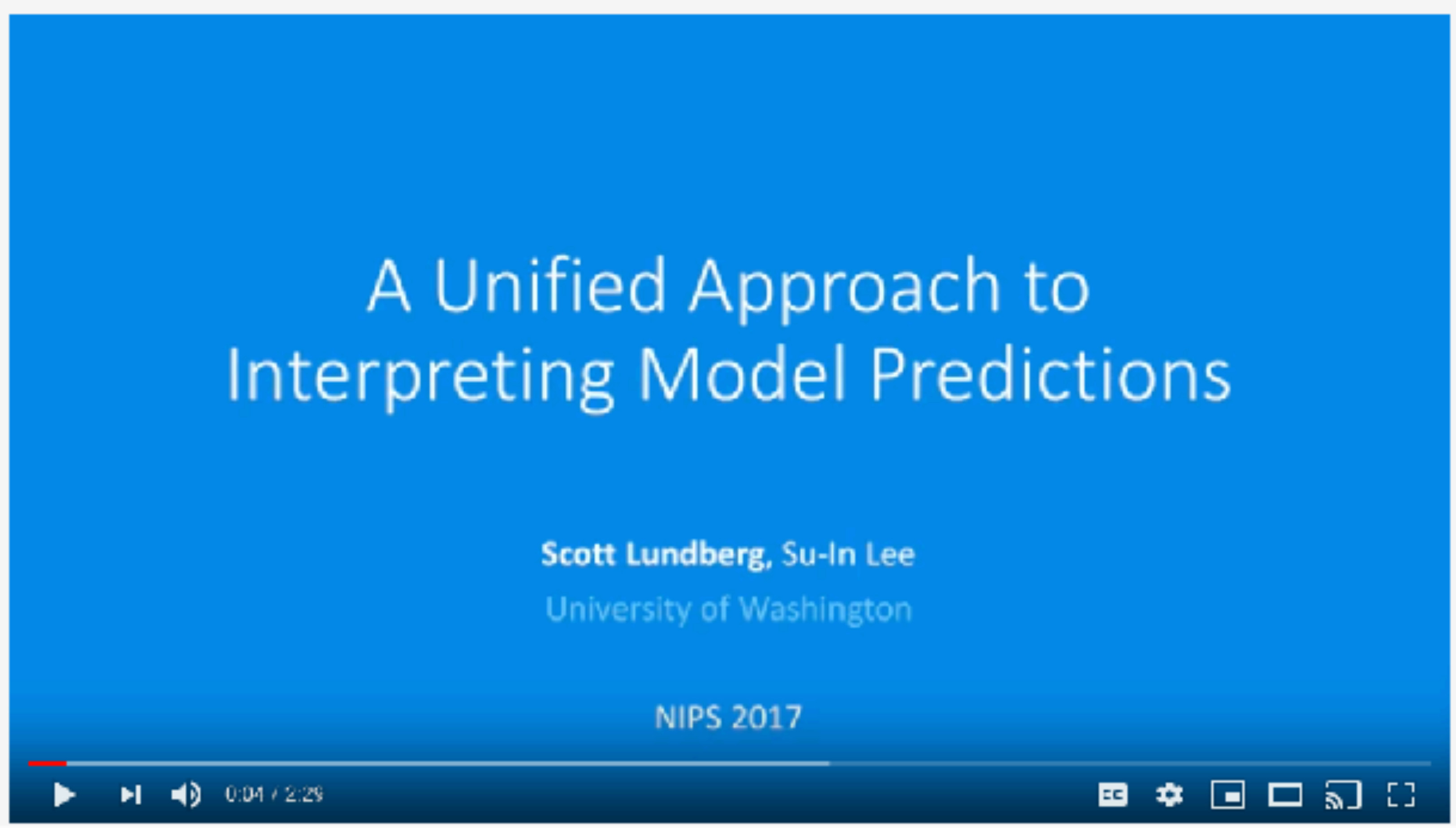
(a) Feature contribution to the top-k QoI, for $k = 10\%$. Systems is the most important feature, followed by Interdisciplinary and AI, while Theory is least important.

Figure 3: Feature importance for the top-k QoI for CS Rankings, with further analysis of 3 departments using Pairwise QoI.

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

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

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.

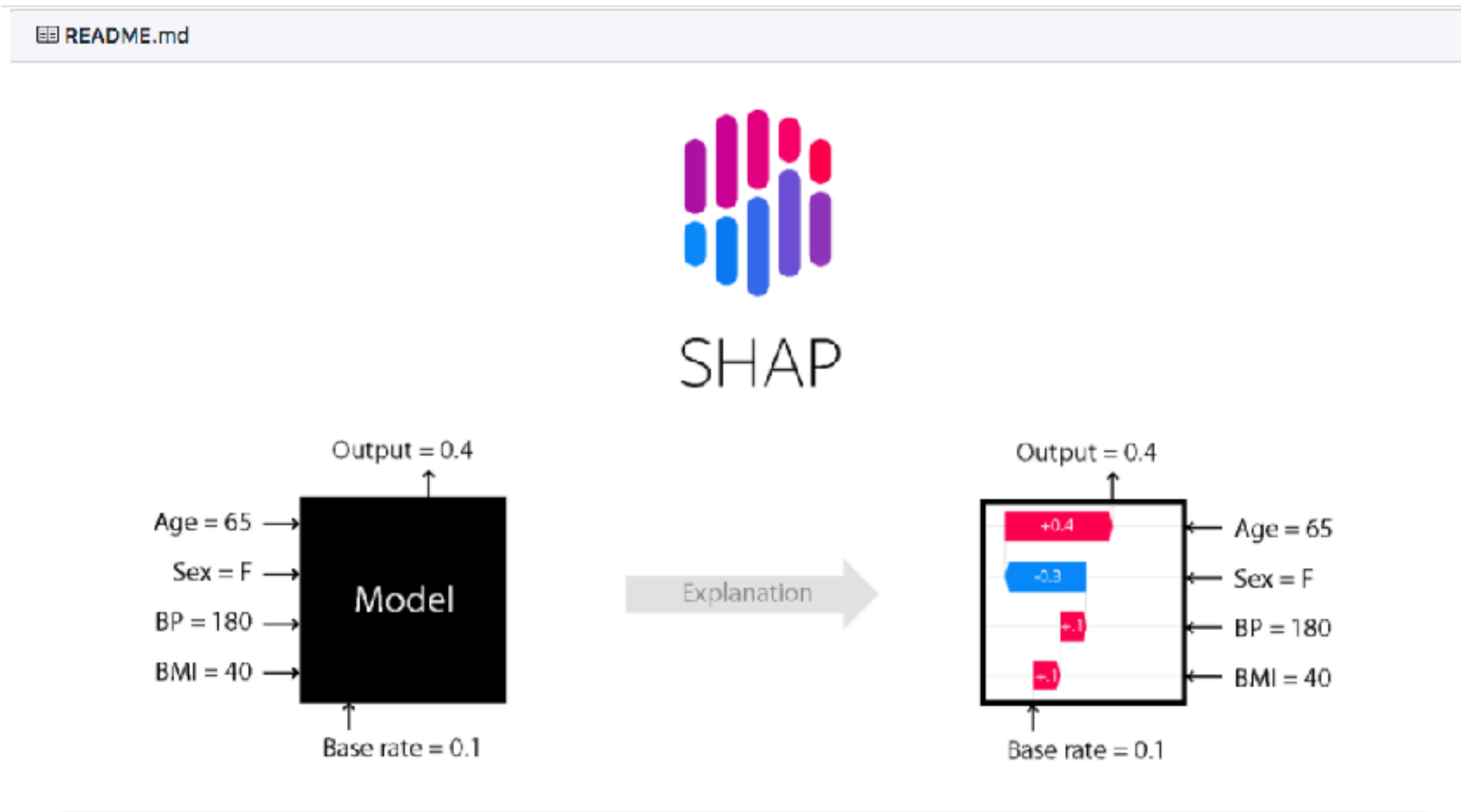
2. **Missingness**: if x'_i — the i^th^ feature of simplified input **x'**— is missing, then it has no attributable impact for ***x***

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

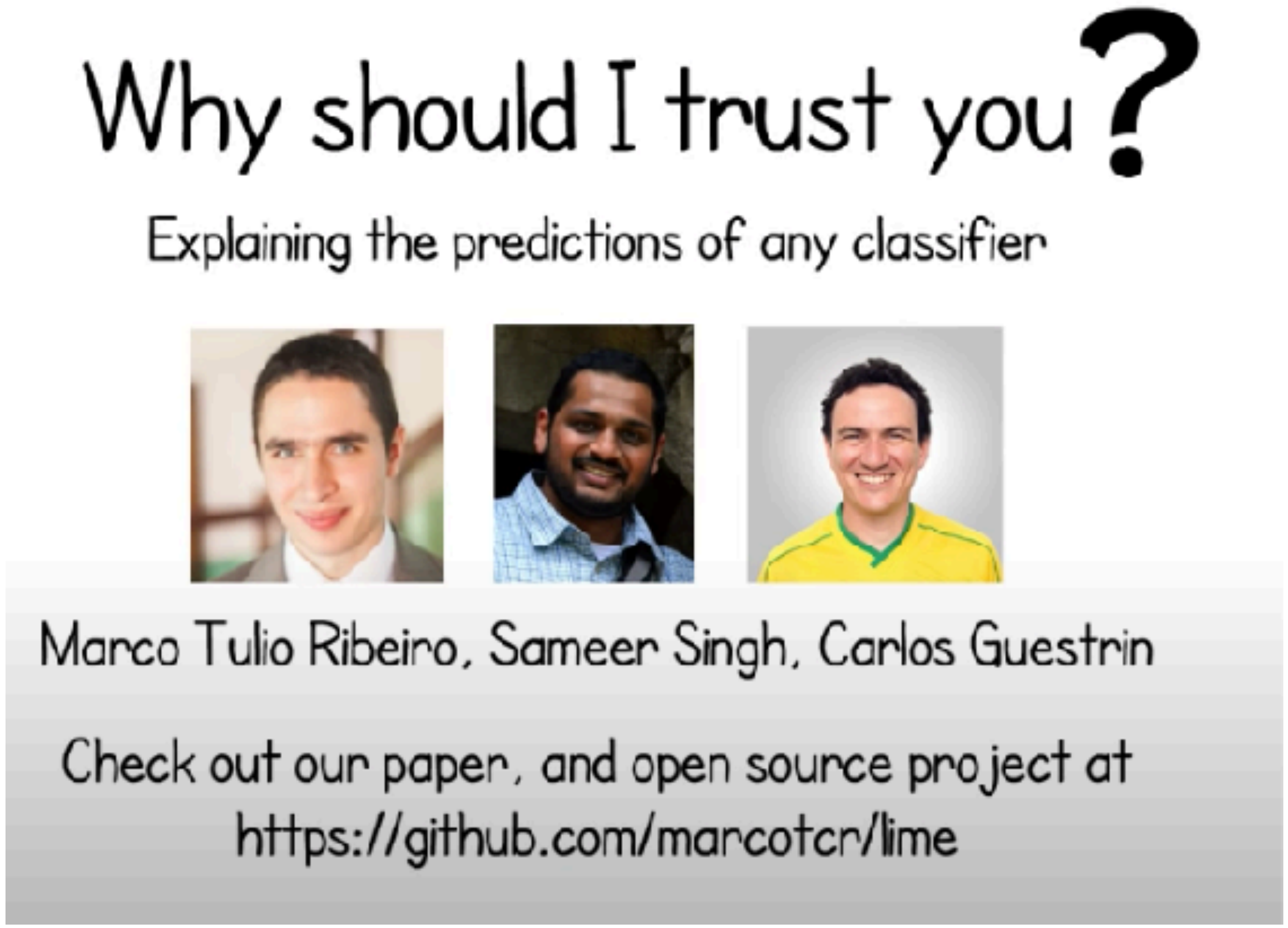
Additive feature attribution methods



https://github.com/slundberg/shap

[Lundberg & Lee, 2017]

LIME: Local Interpretable Model-Agnostic Explanations

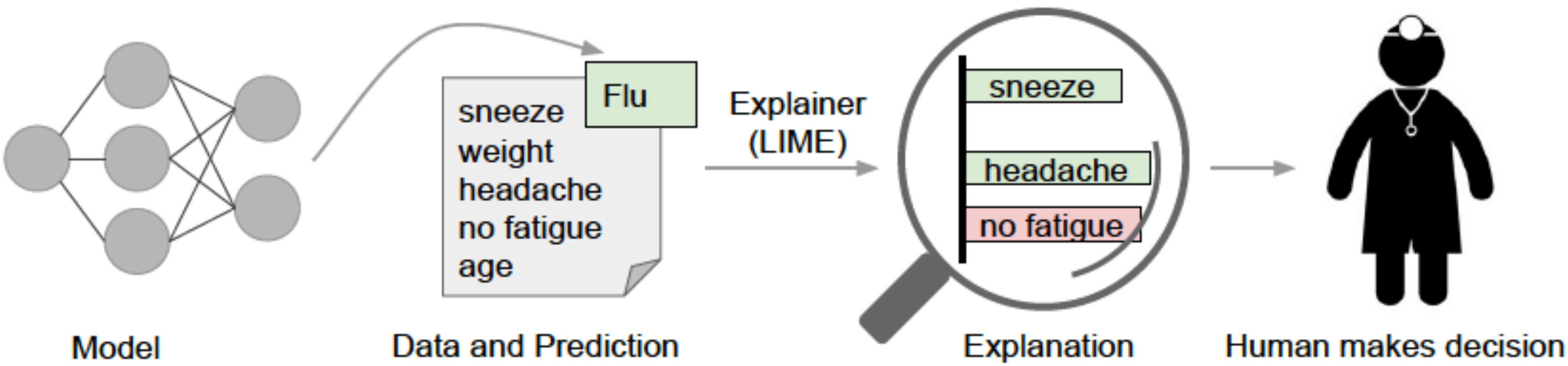


https://www.youtube.com/watch?v=hUnRCxnydCc

[Ribeiro, Singh & Guestrin, 2016]

LIME: Explanations based on features

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



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

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

[Ribeiro, Singh & Guestrin, 2016]

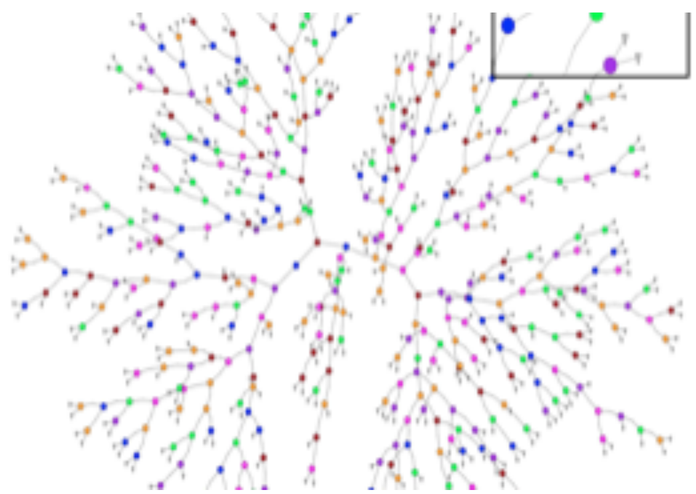
r/ai

LIME: Local explanations of classifiers

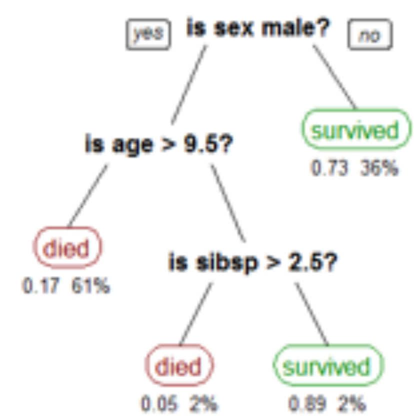
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely
not interpretable



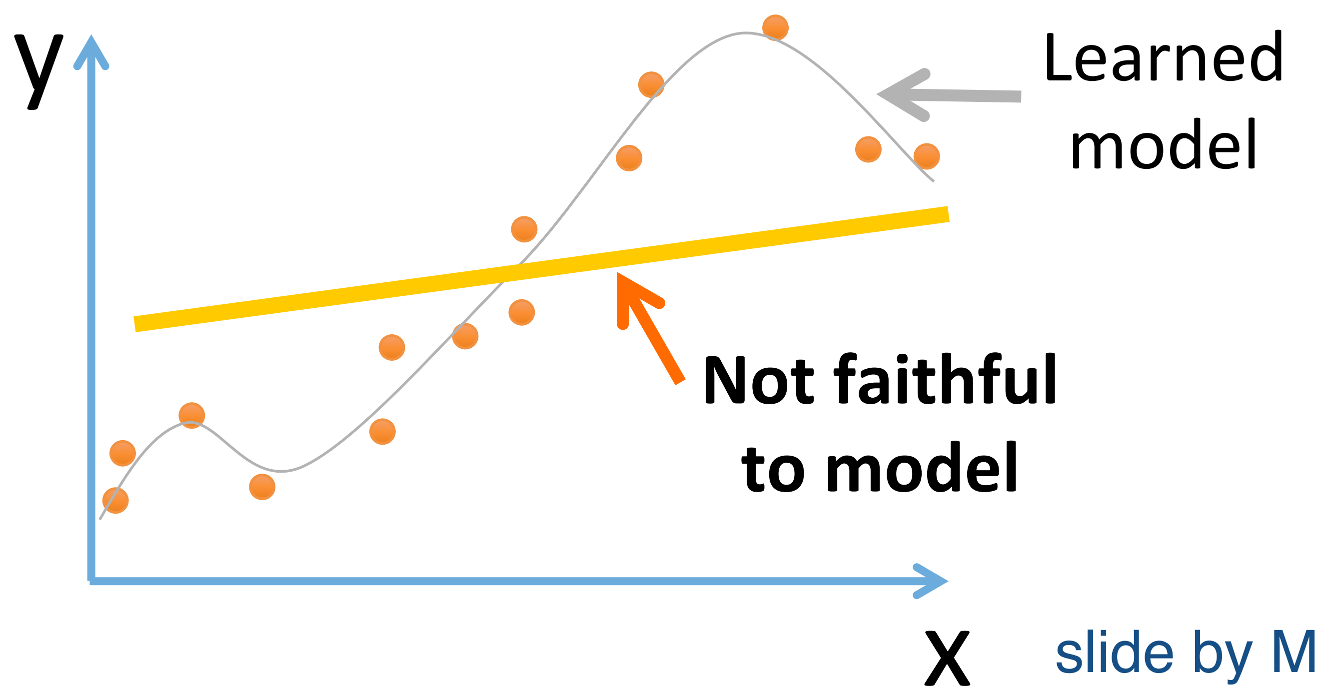
Potentially
interpretable

slide by Marco Tulio Ribeiro, KDD 2016

Explanations based on features

Three must-haves for a good explanation

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



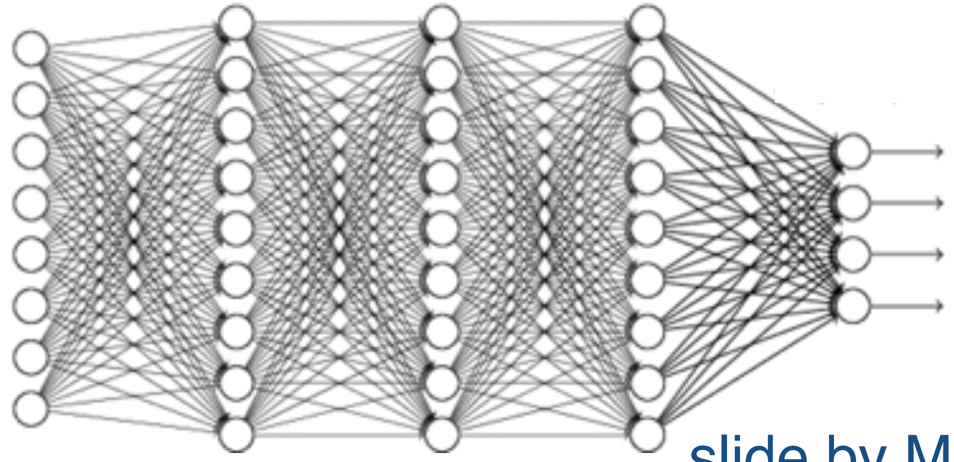
slide by Marco Tulio Ribeiro, KDD 2016

Explanations based on features

Three must-haves for a good explanation

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

Can explain this mess ☺



slide by Marco Tulio Ribeiro, KDD 2016

Key idea: Interpretable representation

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

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."

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

classifier **model being explained**

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

explanation model

some class of interpretable models

$\Omega(g)$

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

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

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

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***

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

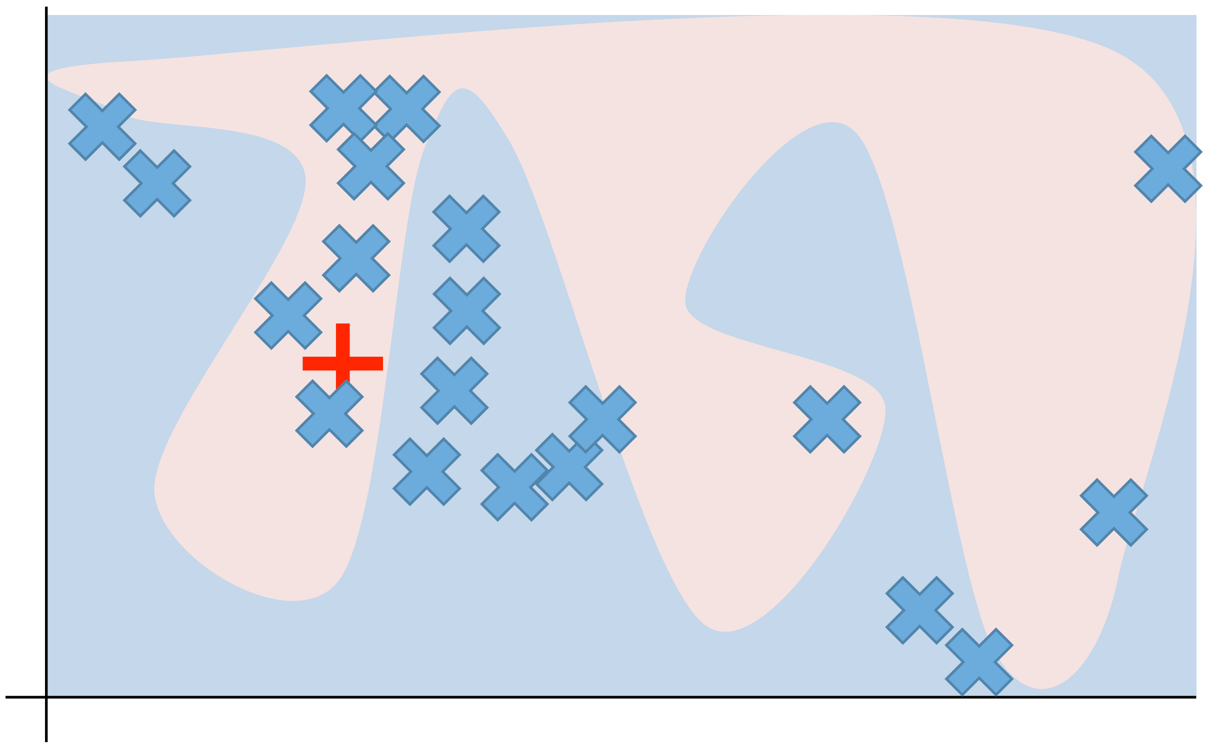
[Ribeiro, Singh & Guestrin, 2016]

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 +

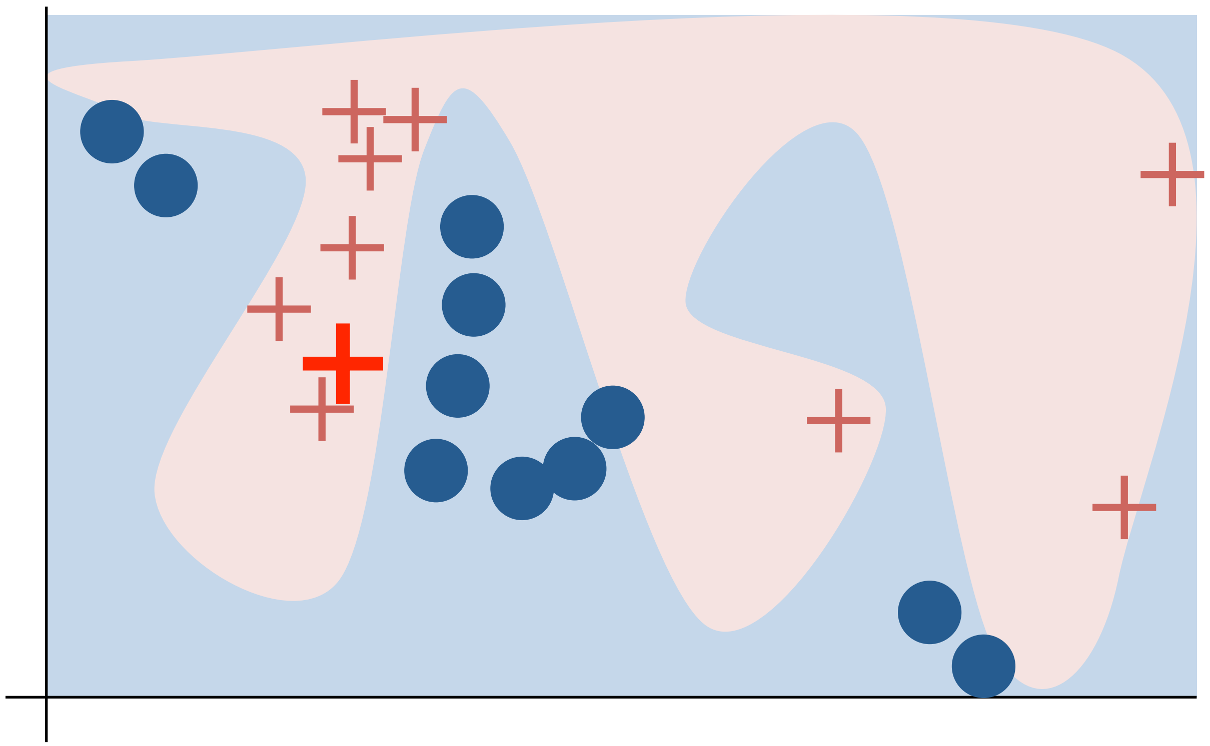


based on a slide by Marco Tulio Ribeiro, KDD 2016

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

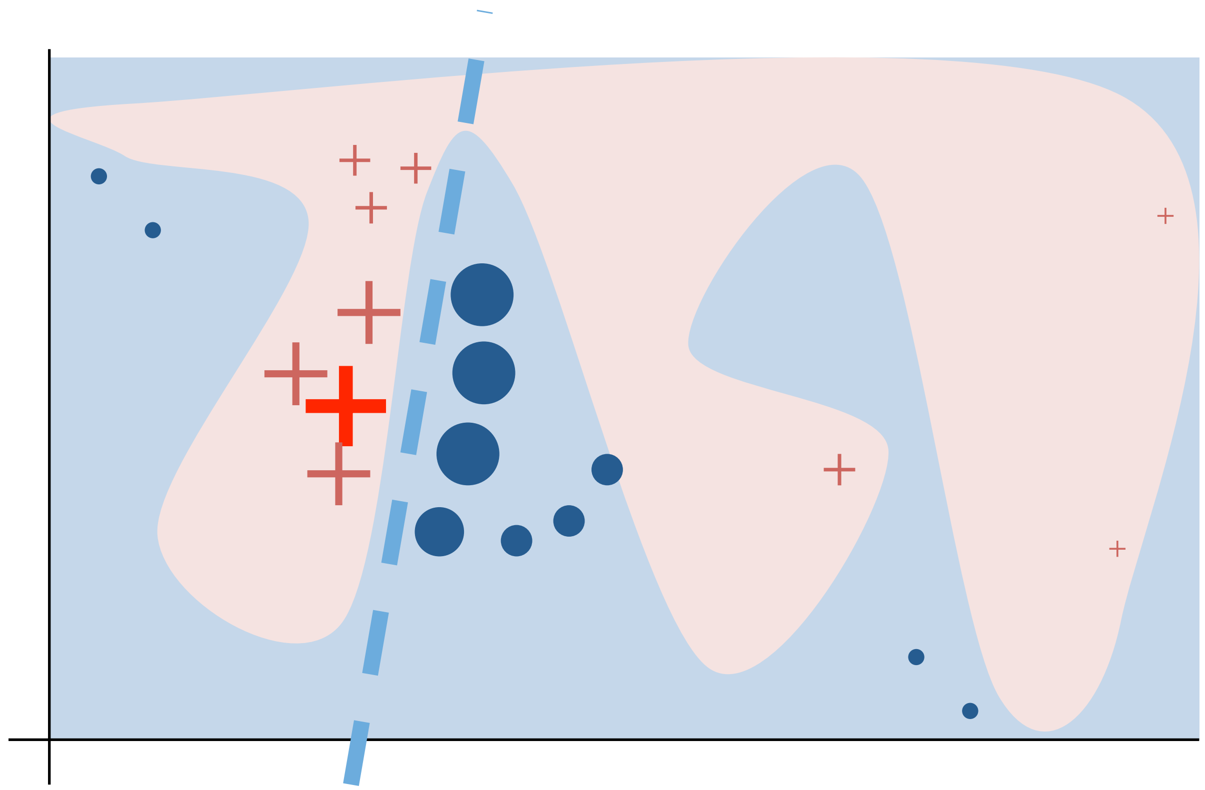


based on a slide by Marco Tulio Ribeiro, KDD 2016

Fidelity-interpretability trade-off

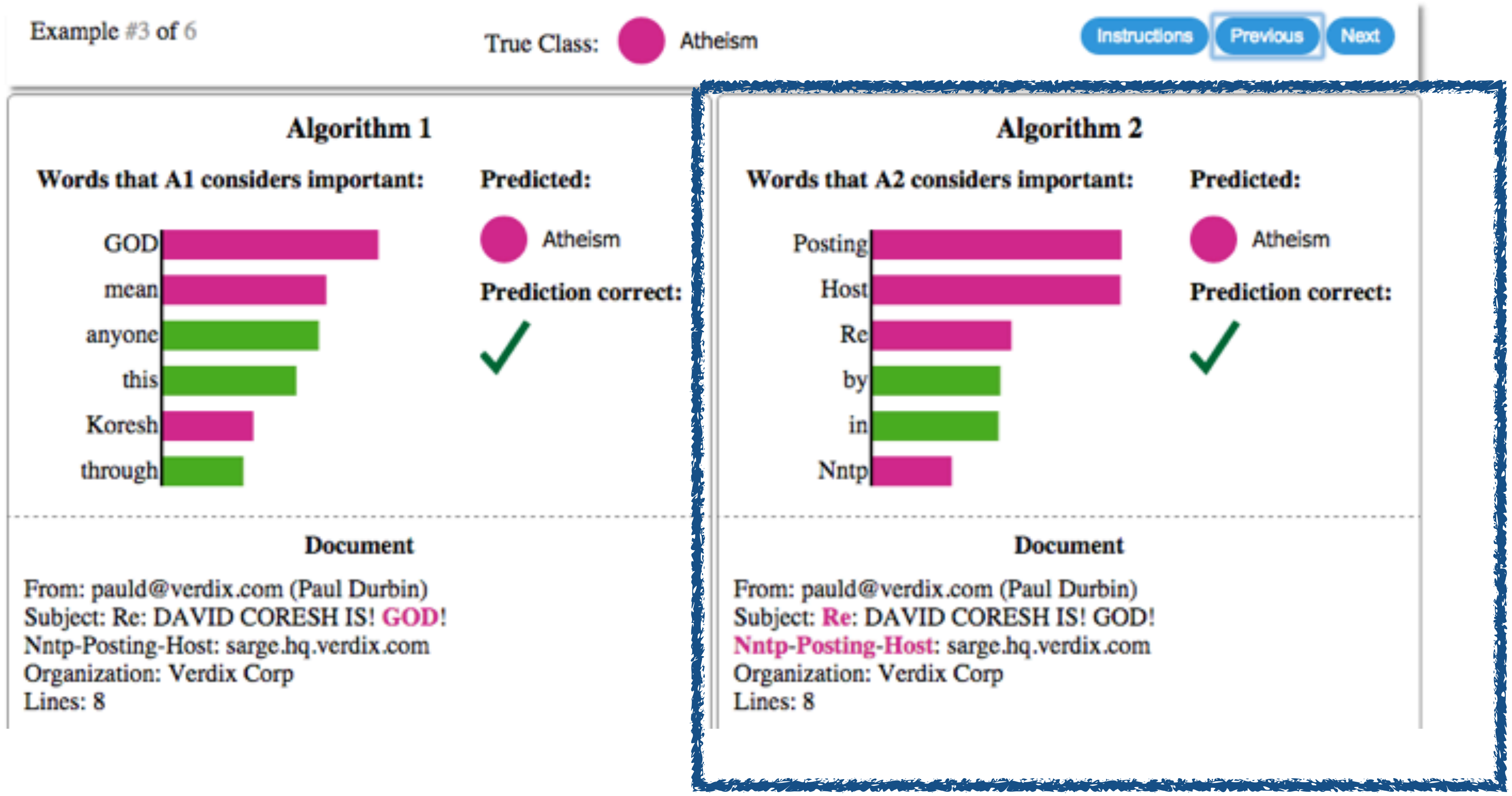
"The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classier."

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

Example: text classification with SVMs



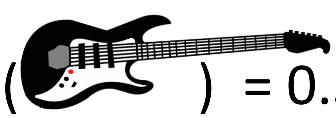
94% accuracy, yet we shouldn't trust this classifier!

[Ribeiro, Singh & Guestrin, 2016]

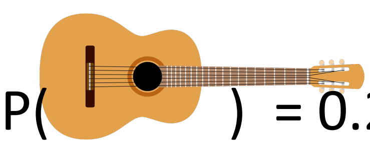
When accuracy is not enough

Explaining Google’s Inception NN

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

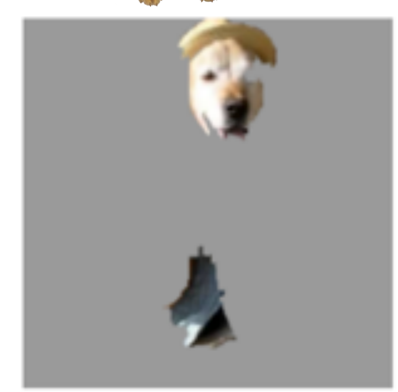
P() = 0.32

Electric guitar - incorrect but
reasonable, similar fretboard

P() = 0.24

Acoustic guitar

P() = 0.21

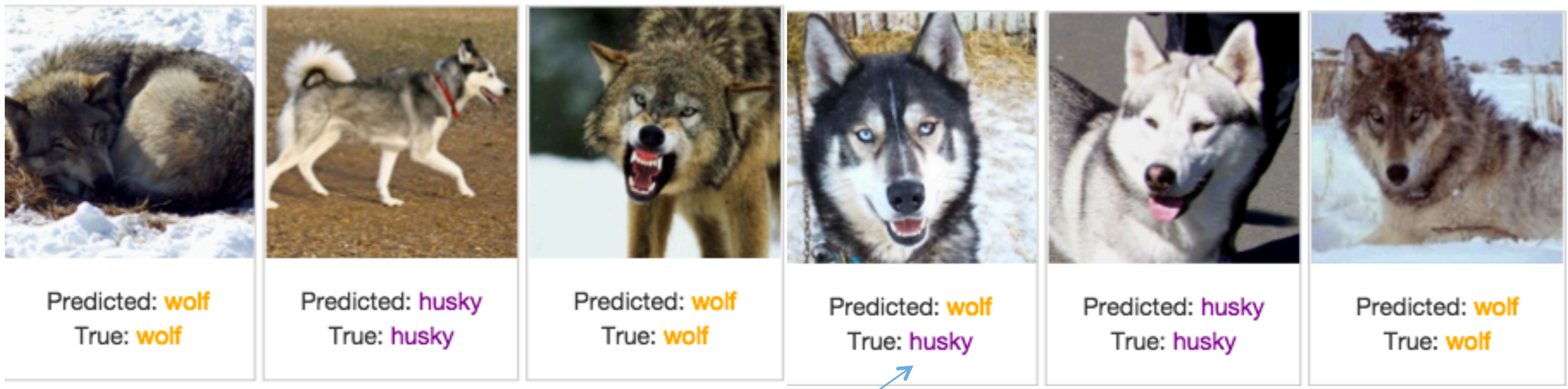


Labrador

[Ribeiro, Singh & Guestrin, 2016]

When accuracy is not enough

Train a neural network to predict wolf v. husky



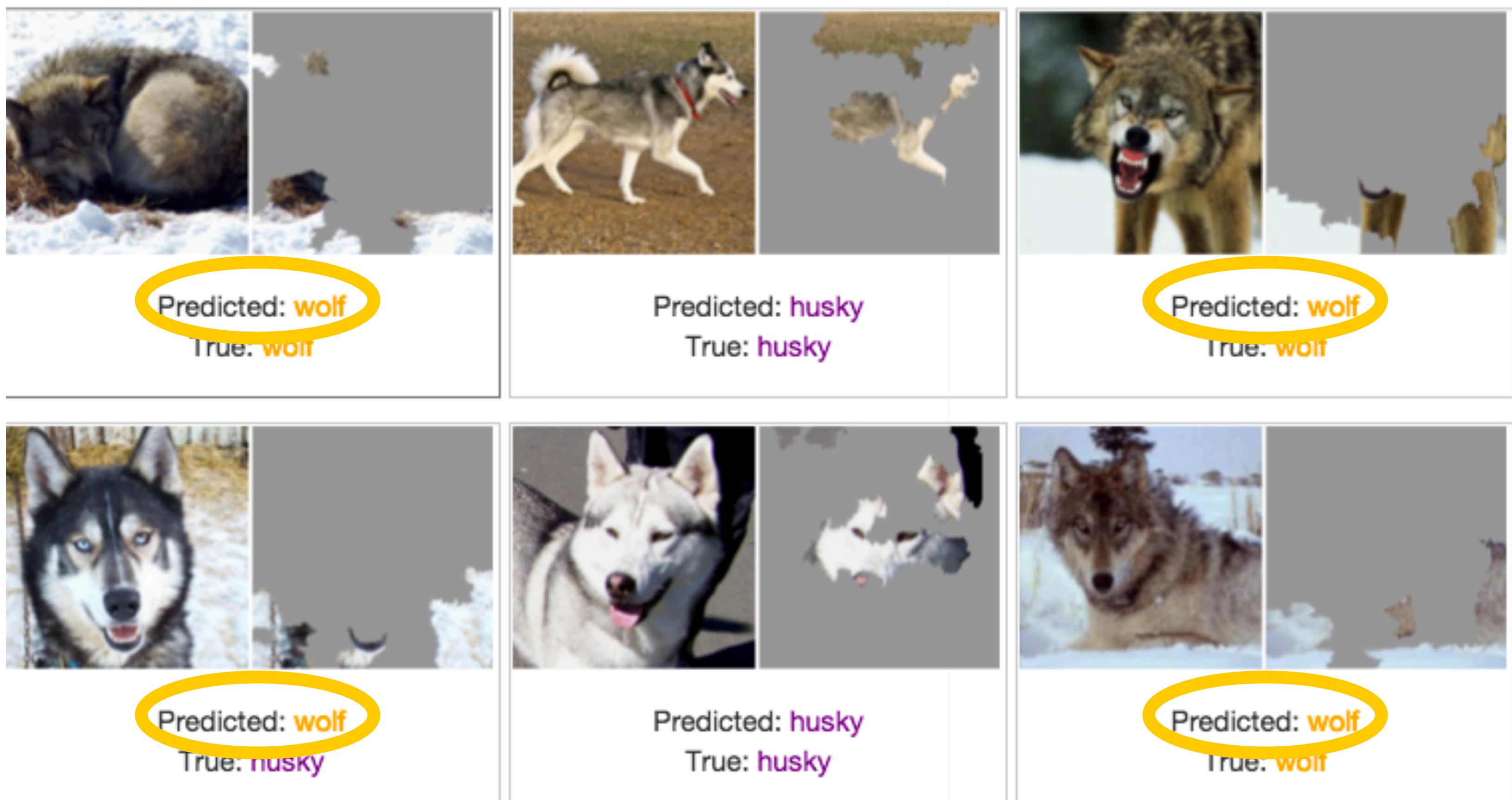
Only 1 mistake!!!

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

slide by Marco Tulio Ribeiro, KDD 2016

When accuracy is not enough

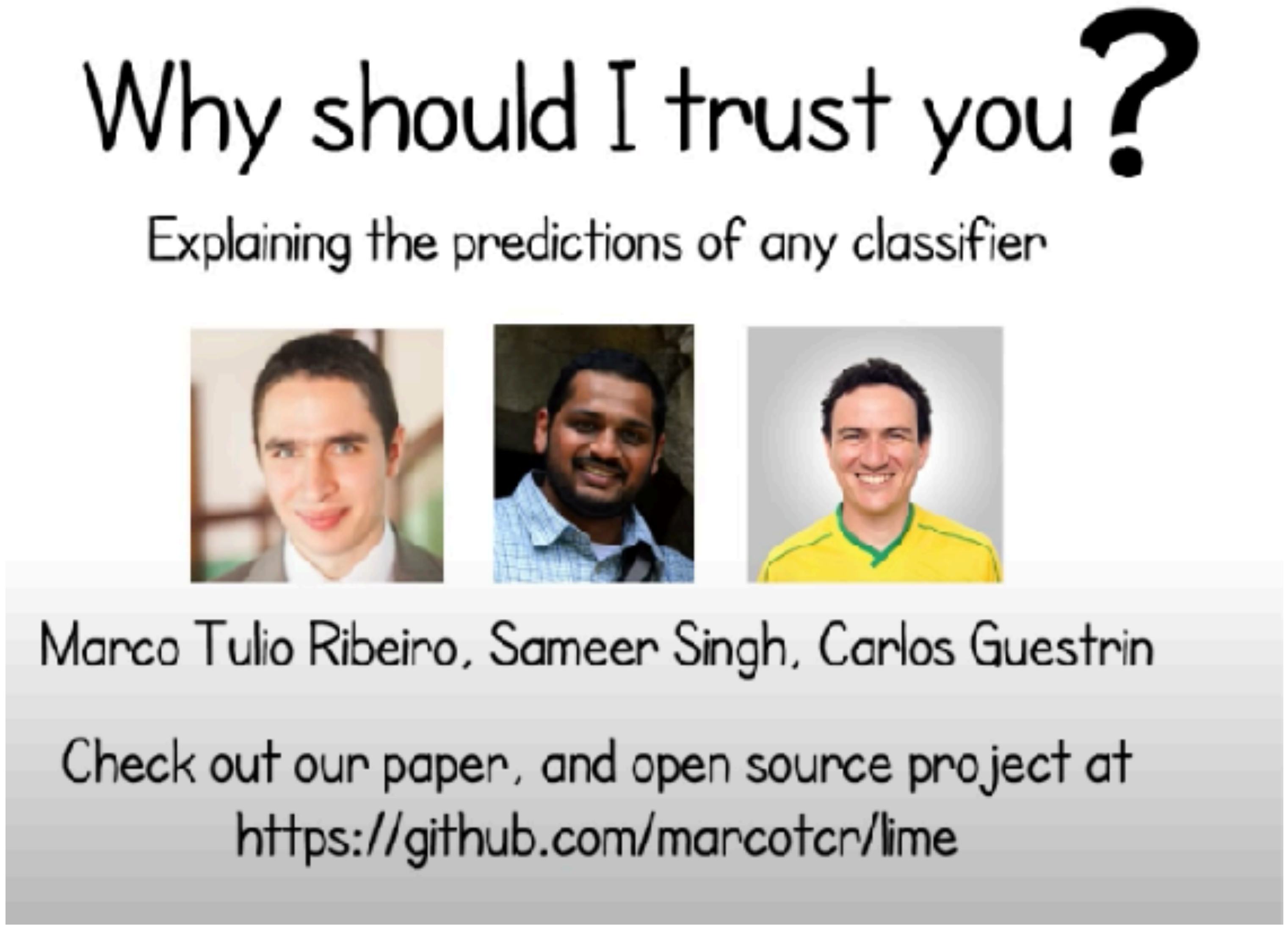
Explanations for neural network prediction



We’ve built a great snow detector… ☹

slide by Marco Tulio Ribeiro, KDD 2016

LIME: Recap



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

[Ribeiro, Singh & Guestrin, 2016]