Responsible Data Science

Transparency & Interpretability

Auditing black-box models

March 5 and 7, 2024

Professor Umang Bhatt

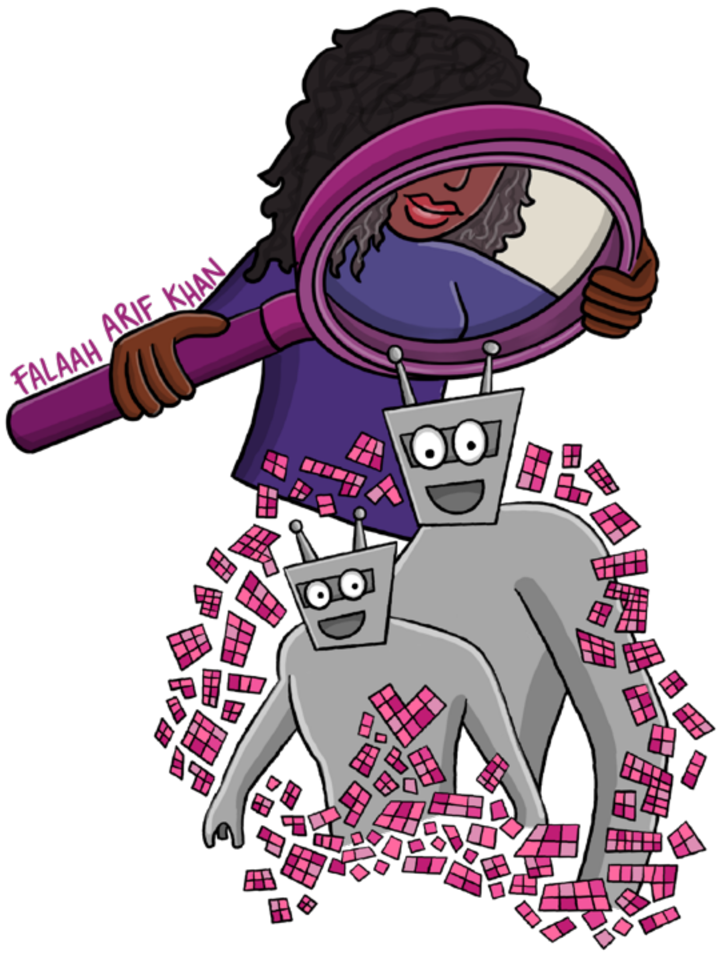
Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

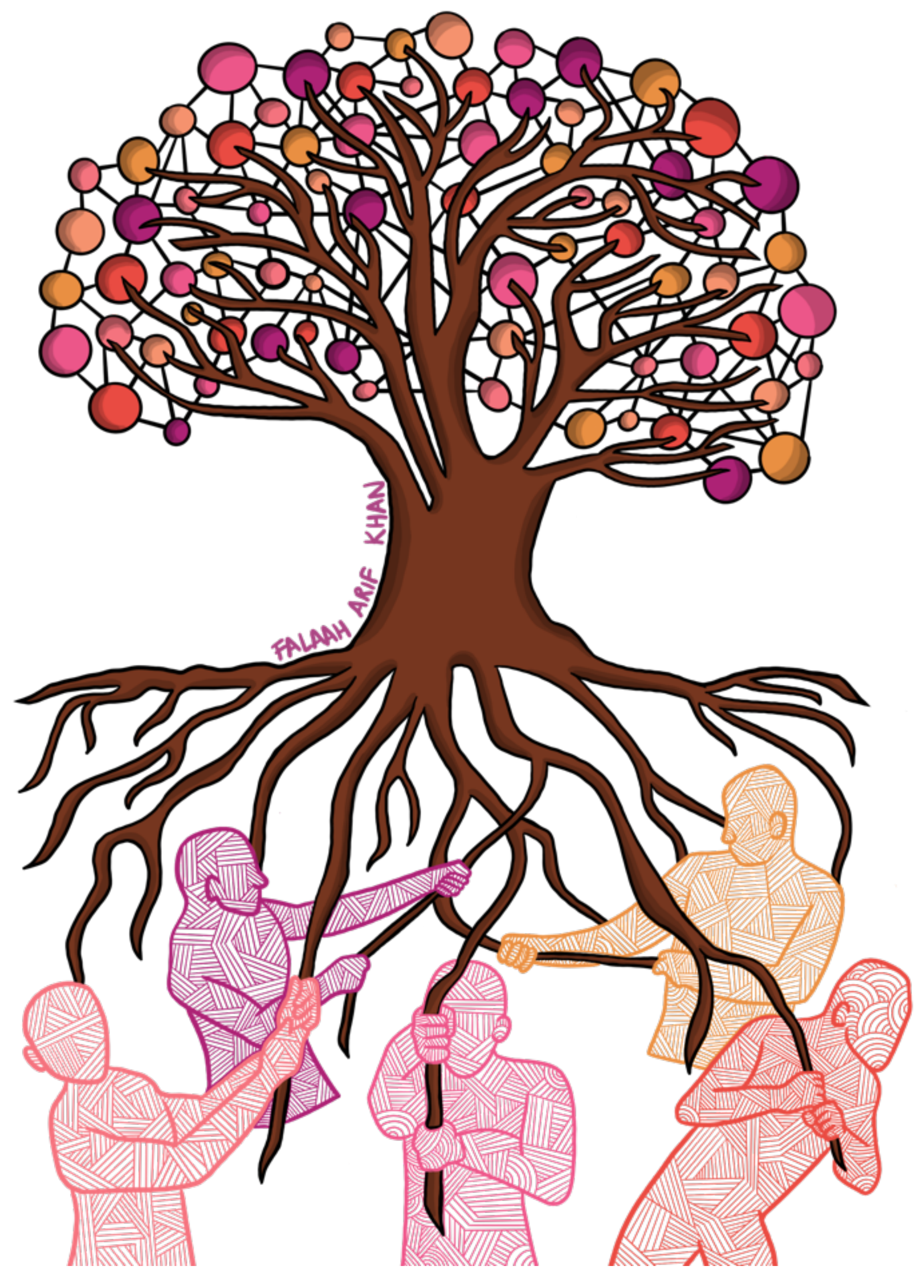
r/ai

Terminology & vision



transparency, interpretability, explainability, intelligibility

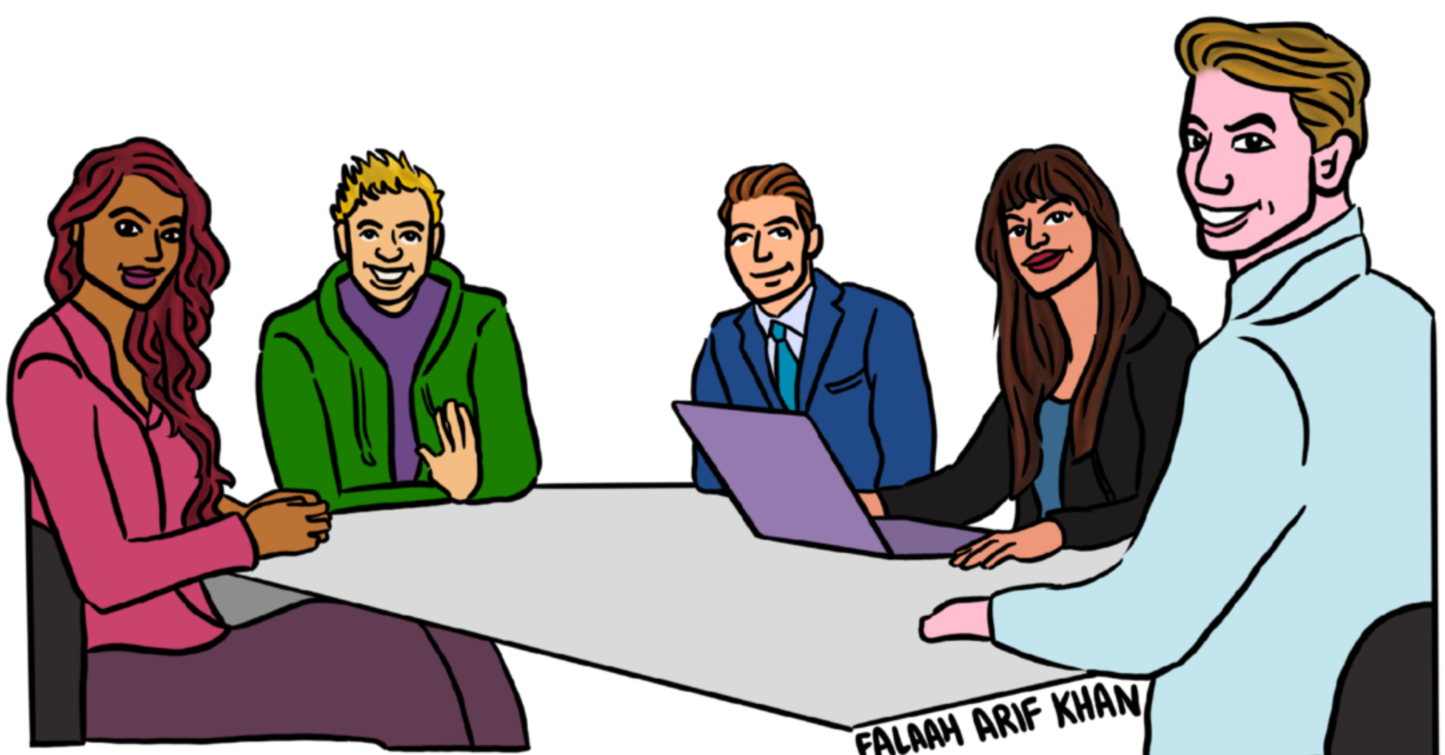
responsible AI



agency, responsibility

r/ai

Interpretability for different stakeholders



What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

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/SB10001424127887323777204578189391813881534

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 Sw
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

r/ai

Online job ads

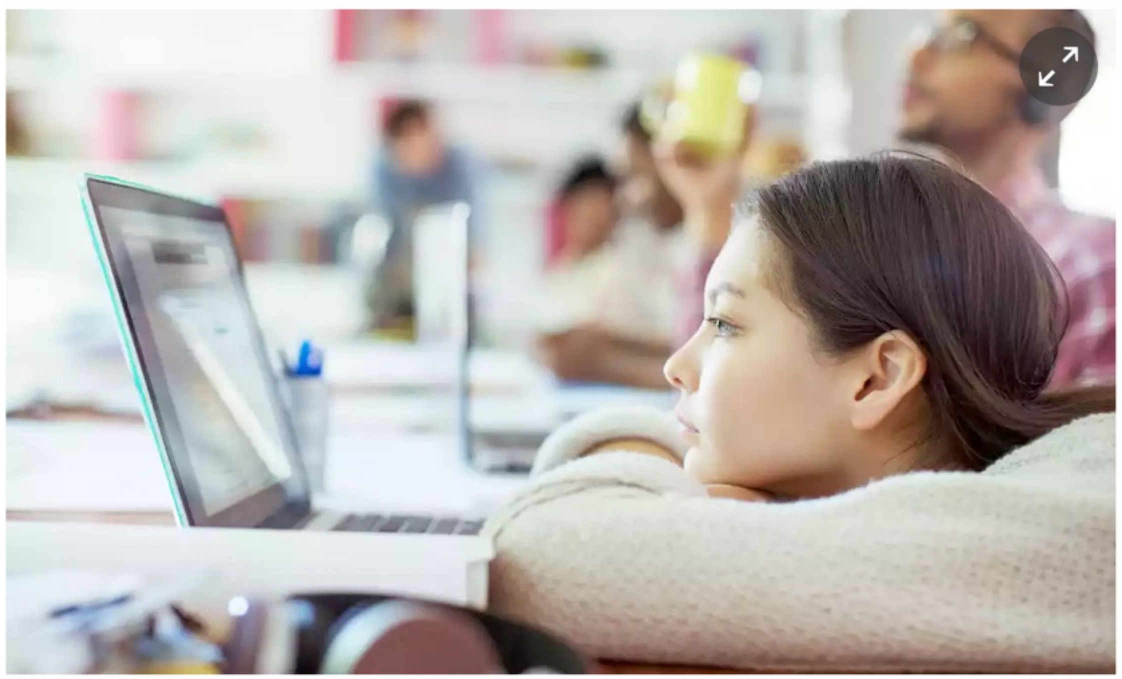
theguardian

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

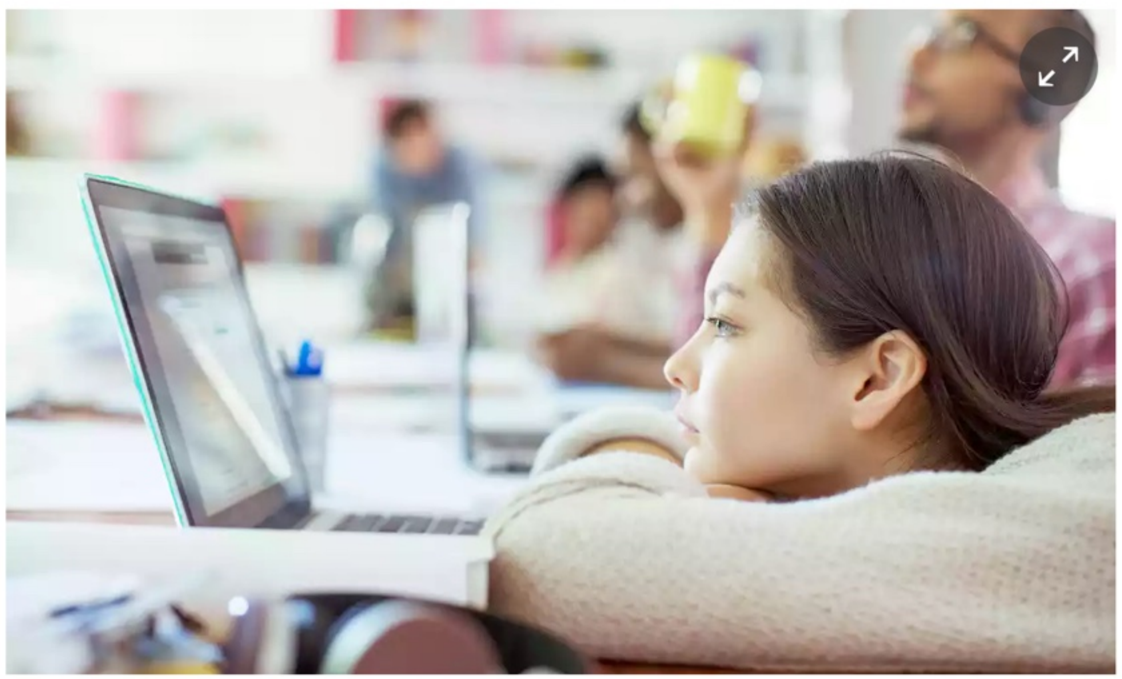
theguardian

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
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?

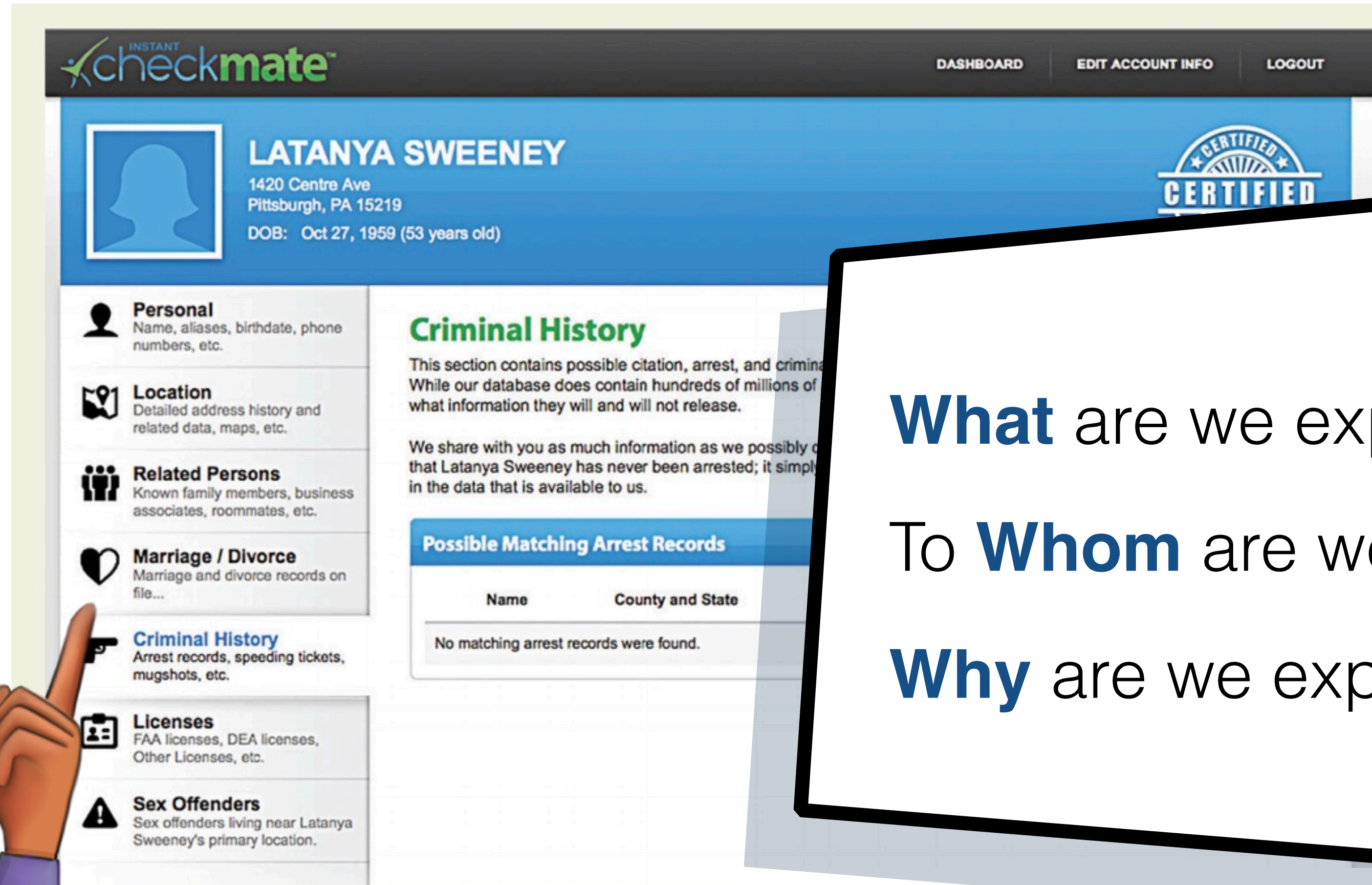
Why are we explaining?

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?

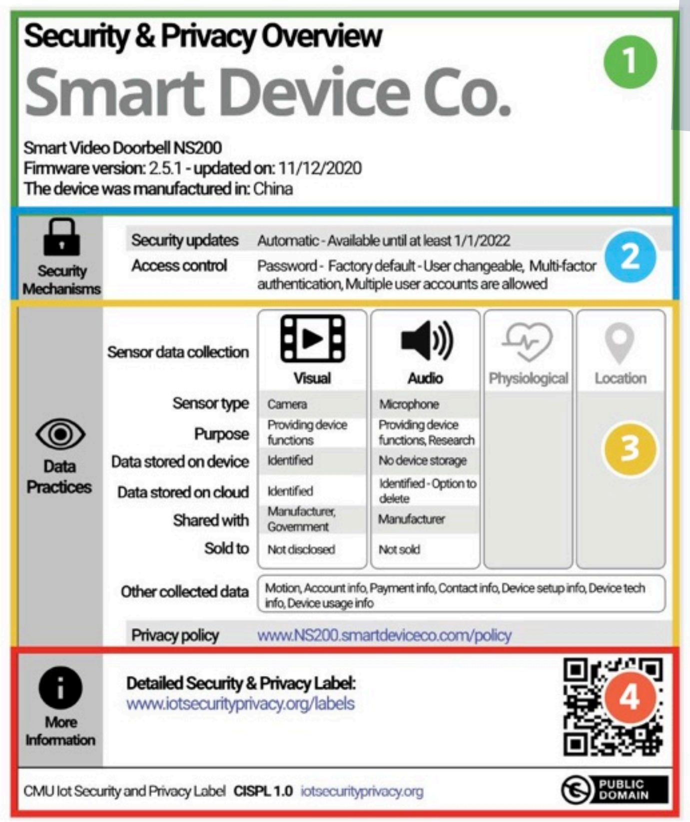
To **Whom** are we explaining?

Why are we explaining?

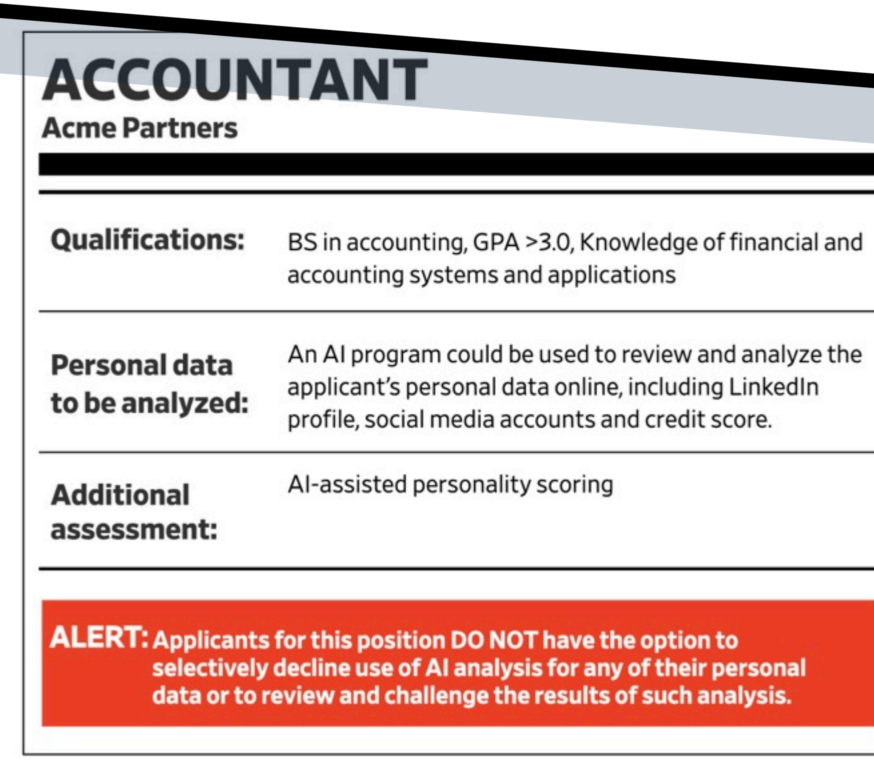
An example of the old nutrition labels, left, and the new one. The new nutrition labels will display calories and serving size more prominently, and include added sugars for the first time.

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
{danupam, shayaks, yairzick}@cmu.edu

***Abstract*—Algorithmic systems that employ machine learning play an increasing role in making substantive decisions in modern society, ranging from online personalization to insurance and credit decisions to predictive policing. But their decision-making processes are often opaque—it is difficult to explain why a certain decision was made. We develop a formal foundation to improve the transparency of such decision-making systems. Specifically, we introduce a family of *Quantitative Input Influence (QII)* measures that capture the degree of influence of inputs on outputs of systems. These measures provide a foundation for the design of transparency reports that accompany system decisions (e.g., explaining a specific credit decision) and for testing tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).**

Distinctively, our *causal* QII measures carefully account for correlated inputs while measuring influence. They support a *general* class of transparency queries and can, in particular, explain decisions about individuals (e.g., a loan decision) and groups (e.g., disparate impact based on gender). Finally, since single inputs may not always have high influence, the QII measures also quantify the *joint influence* of a set of inputs (e.g., age and income) on outcomes (e.g. loan decisions) and the *marginal influence* of individual inputs within such a set (e.g., income). Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using principled aggregation measures, such as the Shapley value, previously applied to measure influence in voting. Further, since transparency reports could compromise privacy, we explore the transparency-privacy tradeoff and prove that a number of useful transparency reports can be made differentially private with very little addition of noise.

Our empirical validation with standard machine learning algorithms demonstrates that QII measures are a useful transparency mechanism when black box access to the learning system is available. In particular, they provide better explanations than standard associative measures for a host of scenarios that we consider. Further, we show that in the situations we consider, QII is efficiently approximable and can be made differentially private while preserving accuracy.

I. Introduction

Algorithmic decision-making systems that employ machine learning and related statistical methods are ubiquitous. They drive decisions in sectors as diverse as Web services, health-care, education, insurance, law enforcement and defense [1], [2], [3], [4], [5]. Yet their decision-making processes are often opaque. *Algorithmic transparency* is an emerging research area aimed at explaining decisions made by algorithmic systems.

The call for algorithmic transparency has grown in intensity as public and private sector organizations increasingly use large volumes of personal information and complex data analytics systems for decision-making [6]. Algorithmic transparency provides several benefits. First, it is essential to enable identification of harms, such as discrimination, introduced by algorithmic decision-making (e.g., high interest credit cards targeted to protected groups) and to hold entities in the decision-making chain accountable for such practices. This form of accountability can incentivize entities to adopt appropriate corrective measures. Second, transparency can help detect errors in input data which resulted in an adverse decision (e.g., incorrect information in a user's profile because of which insurance or credit was denied). Such errors can then be corrected. Third, by explaining why an adverse decision was made, it can provide guidance on how to reverse it (e.g., by identifying a specific factor in the credit profile that needs to be improved).

Our Goal. While the importance of algorithmic transparency is recognized, work on computational foundations for this research area has been limited. This paper initiates progress in that direction by focusing on a concrete algorithmic transparency question:

How can we measure the influence of inputs (or features) on decisions made by an algorithmic system about individuals or groups of individuals?

Our goal is to inform the design of transparency reports, which include answers to transparency queries of this form. To be concrete, let us consider a predictive policing system that forecasts future criminal activity based on historical data; individuals high on the list receive visits from the police. An individual who receives a visit from the police may seek a transparency report that provides answers to *personalized transparency queries* about the influence of various inputs (or features), such as race or recent criminal history, on the system's decision. An oversight agency or the public may desire a transparency report that provides answers to *aggregate transparency queries*, such as the influence of sensitive inputs (e.g., gender, race) on the system's decisions concerning the entire population or about systematic differences in decisions

DOI 10.1109/SP.2016.42

IEEE computer society

LIME

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

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans is an oft-overlooked aspect in the field. Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: *if the users do not trust a model or a prediction, they will not use it.* It is important to differentiate between two different (but related) definitions of trust: (1) *trusting a prediction*, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) *trusting a model*, i.e. whether the user trusts a model to behave in reasonable ways if deployed. Both are directly impacted by how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the "trusting a prediction" problem, and selecting multiple such predictions (and explanations) as a solution to the "trusting the model" problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- SP-LIME, a method that selects a set of representative instances with explanations to address the "trusting the model" problem, via submodular optimization.
- Comprehensive evaluation with simulated and human subjects, where we measure the impact of explanations on trust and associated tasks. In our experiments, non-experts using LIME are able to pick which classifier from a pair generalizes better in the real world. Further, they are able to greatly improve an untrustworthy classifier trained on 20 newsgroups, by doing feature engineering using LIME. We also show how understanding the predictions of a neural network on images helps practitioners know when and why they should not trust a model.

2. THE CASE FOR EXPLANATIONS

By "explaining a prediction", we mean presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction. We

KDD 2016 San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939778

r/ai

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

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

1 Introduction

The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. In some applications, simple models (e.g., linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones. However, the growing availability of big data has increased the benefits of using complex models, so bringing to the forefront the trade-off between accuracy and interpretability of a model's output. A wide variety of different methods have been recently proposed to address this issue [5, 8, 9, 3, 4, 1]. But an understanding of how these methods relate and when one method is preferable to another is still lacking.

Here, we present a novel unified approach to interpreting model predictions.[1] Our approach leads to three potentially surprising results that bring clarity to the growing space of methods:

1. We introduce the perspective of viewing *any* explanation of a model's prediction as a model itself, which we term the *explanation model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unifies six current methods.

[1] https://github.com/slundberg/shap

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

ShaRP

ShaRP: Explaining Rankings with Shapley Values

Venetia Pliatsika[1], **Joao Fonseca**[*2], **Tilun Wang**[1] and **Julia Stoyanovich**[1]
[1]New York University, NY, USA [2]NOVA University, Lisbon, Portugal
[1]{venetia, tw2221, stoyanovich}@nyu.edu, [2]jpfonseca@novaims.unl.pt

arXiv:2401.16744v1 [cs.AI] 30 Jan 2024

Abstract

Algorithmic decisions in critical domains such as hiring, college admissions, and lending are often based on rankings. Because of the impact these decisions have on individuals, organizations, and population groups, there is a need to understand them: to know whether the decisions are abiding by the law, to help individuals improve their rankings, and to design better ranking procedures.

In this paper, we present ShaRP (Shapley for Rankings and Preferences), a framework that explains the contributions of features to different aspects of a ranked outcome, and is based on Shapley values. Using ShaRP, we show that even when the scoring function used by an algorithmic ranker is known and linear, the weight of each feature does not correspond to its Shapley value contribution. The contributions instead depend on the feature distributions, and on the subtle local interactions between the scoring features. ShaRP builds on the Quantitative Input Influence framework, and can compute the contributions of features for multiple Quantities of Interest, including score, rank, pair-wise preference, and top-k. Because it relies on black-box access to the ranker, ShaRP can be used to explain both score-based and learned ranking models. We show results of an extensive experimental validation of ShaRP using real and synthetic datasets, showcasing its usefulness for qualitative analysis.

1 Introduction

Algorithmic rankers are broadly used to support decision-making in critical domains, including critical domains such as hiring and employment, school and college admissions, credit and lending, and college ranking. Because of the impact rankers have on individuals, organizations, and population groups, there is a need to understand them: to know whether the decisions are abiding by the law, to help individuals improve their rankings, and to design better ranking procedures. In this paper, we present ShaRP (Shapley for Rankings and Preferences), a framework that explains the contributions of features to different aspects of a ranked outcome.

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.

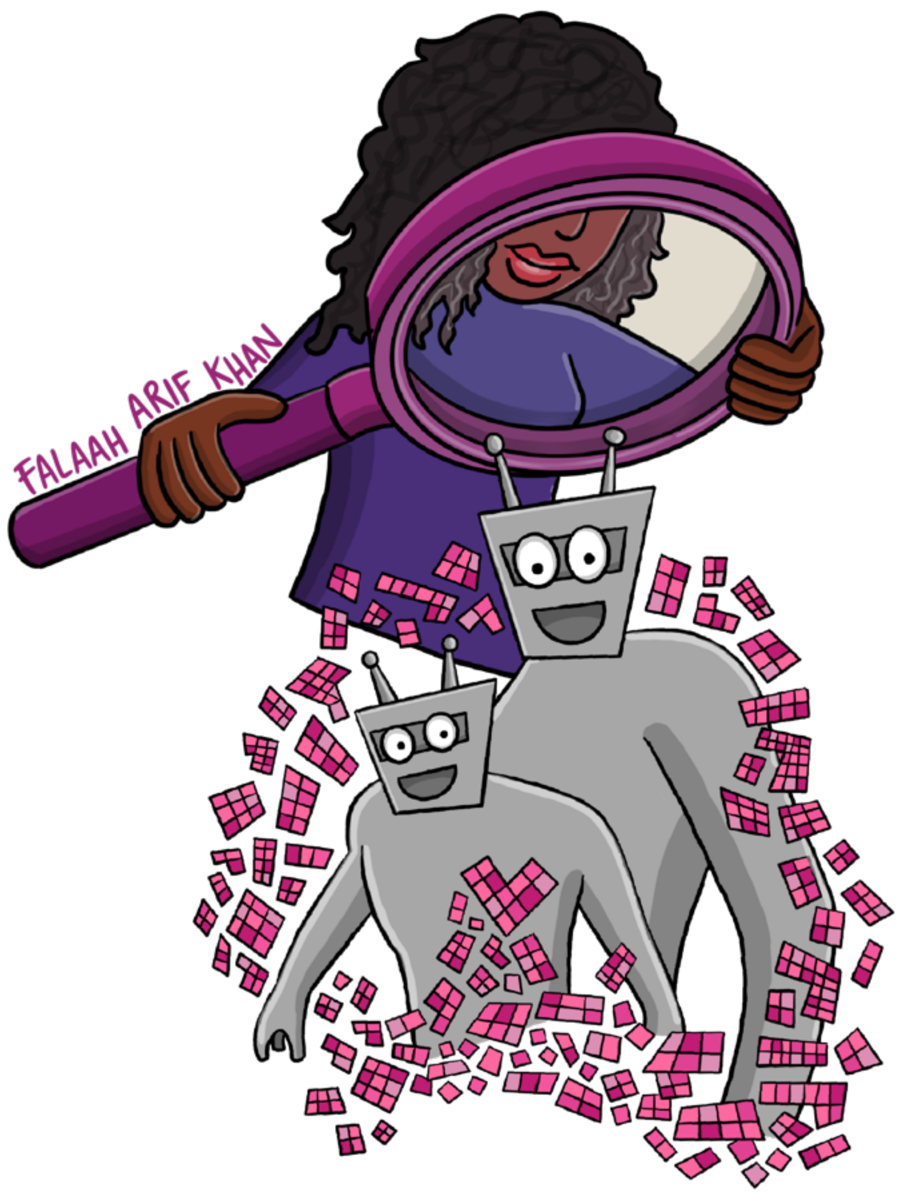
There are two types of rankers: score-based and learned. In score-based ranking, a given set of candidates is sorted on a score, which is typically computed using a simple formula, such as a sum of attribute values with non-negative weights Zehlike *et al.* [2022a]. In supervised learning-to-rank, a preference-enriched set of candidates is used to train a model that predicts rankings of unseen candidates Li [2014]. To motivate our work, let us start with score-based rankers that are often preferred in critical domains, based on the premise that they are easier to design, understand, and justify than complex learning-to-rank models Berger *et al.* [2019]. In fact, score-based rankers are a prominent example of the so-called "interpretable models" Rudin [2019]: the scoring function, such as $Y_1 = 0.4 \times gpa + 0.4 \times sat + 0.2 \times essay$ in a college admissions scenario, is based on a (normative) a priori understanding of what makes for a good candidate.

And yet, despite being syntactically "interpretable", score-based rankers may not be "explainable," in the sense that the designer of the ranker or the decision-maker who uses it, may be unable to accurately predict and understand their output (Miller [2019]; Molnar [2020]). We now illustrate this with a simple example.

Example 1. *Consider a dataset $\mathcal{D}$ of college applicants in Figure 1, with scoring features gpa, sat, and essay. Very different scoring functions $f = 0.4 \times gpa + 0.4 \times sat + 0.2 \times essay$ and $g = 1.0 \times essay$ induce very similar rankings $r_{\mathcal{D},f}$ and $r_{\mathcal{D},g}$, with the same top-4 items appearing in the*

r/ai

What are we explaining?



How does a system work?

How **well** does a system work?

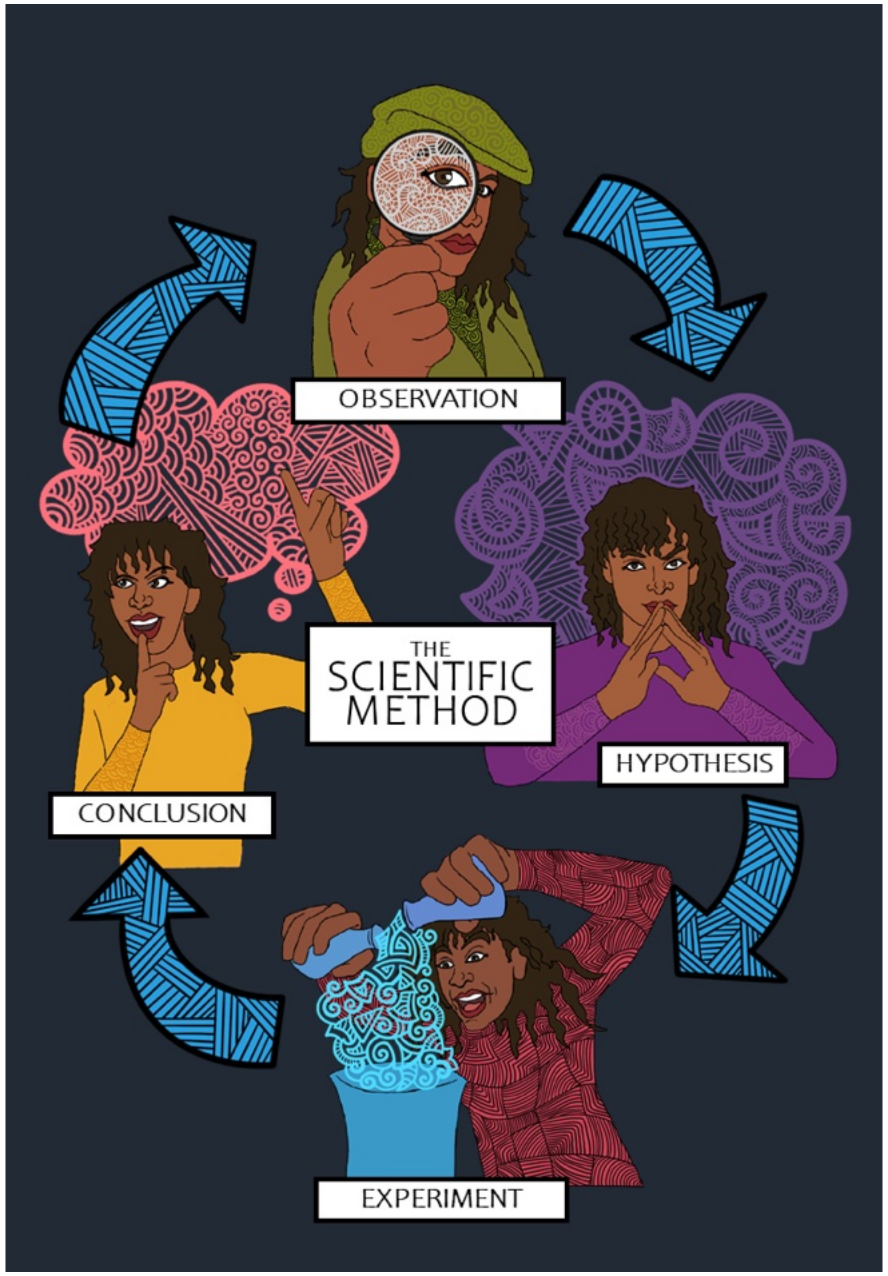
What does a system do?

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

Are a system's decisions discriminatory?

Are a system's decisions illegal?

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

Tweet

Shane Creepingbear is a member of the Kiowa Tribe of Oklahoma

October 13, 2014

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

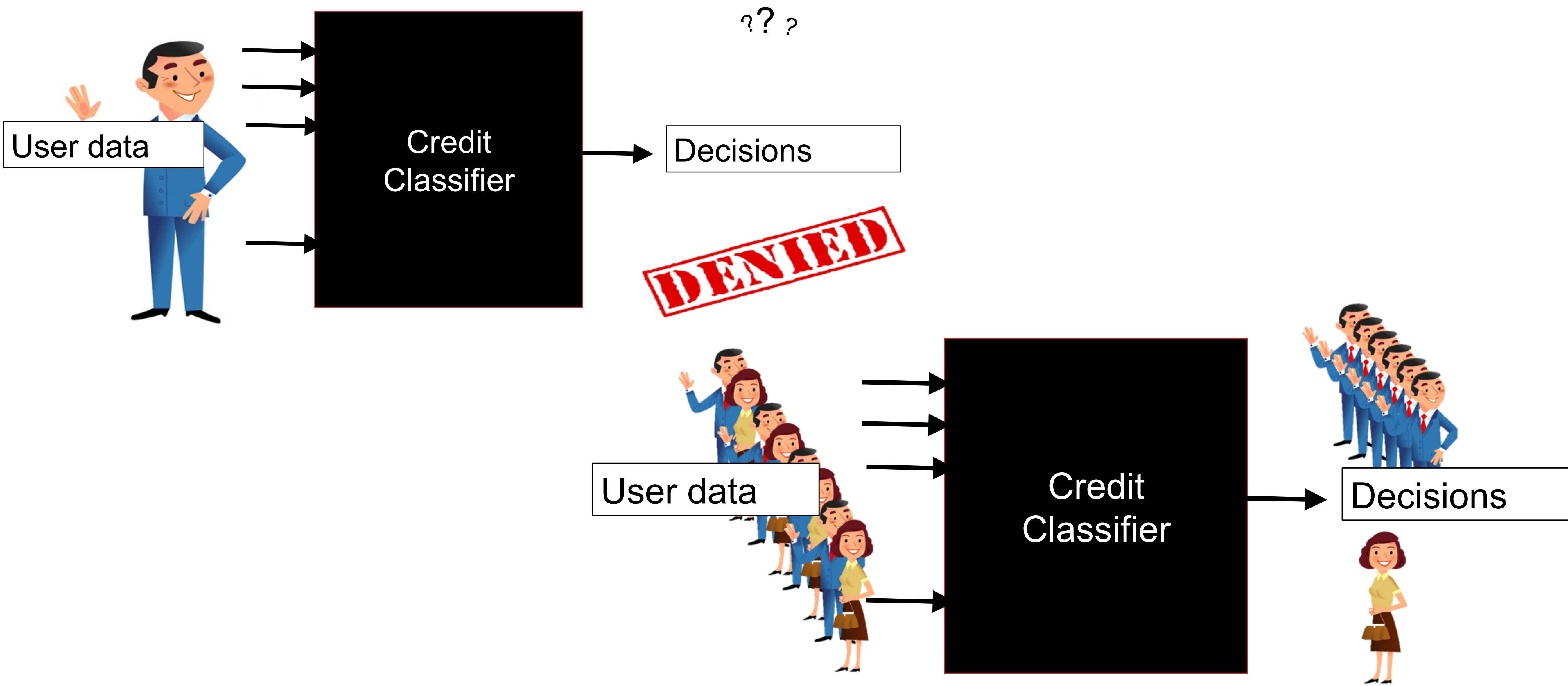
February 14, 2015

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.

QII: Auditing black-box models

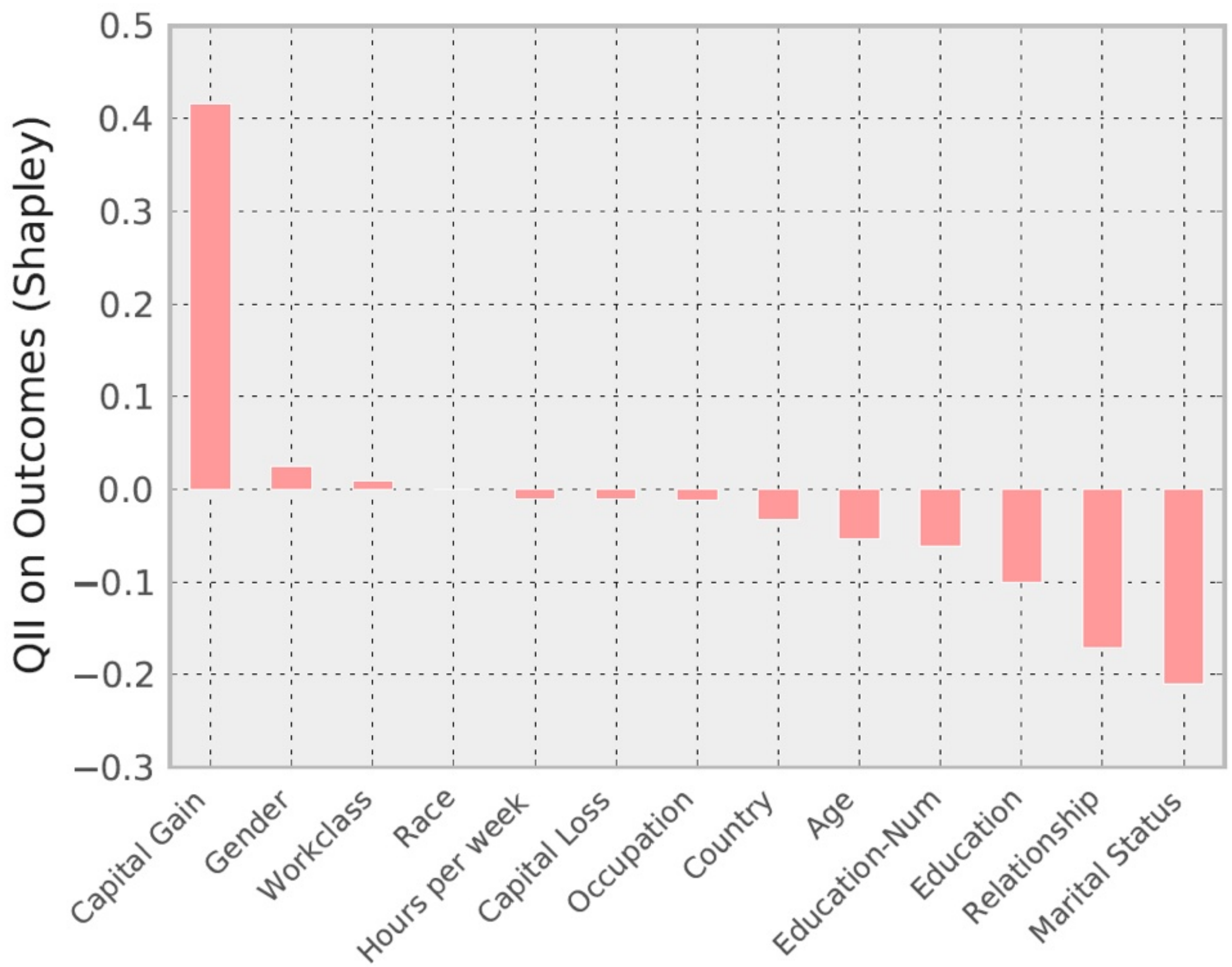


images by Anupam Datta

[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

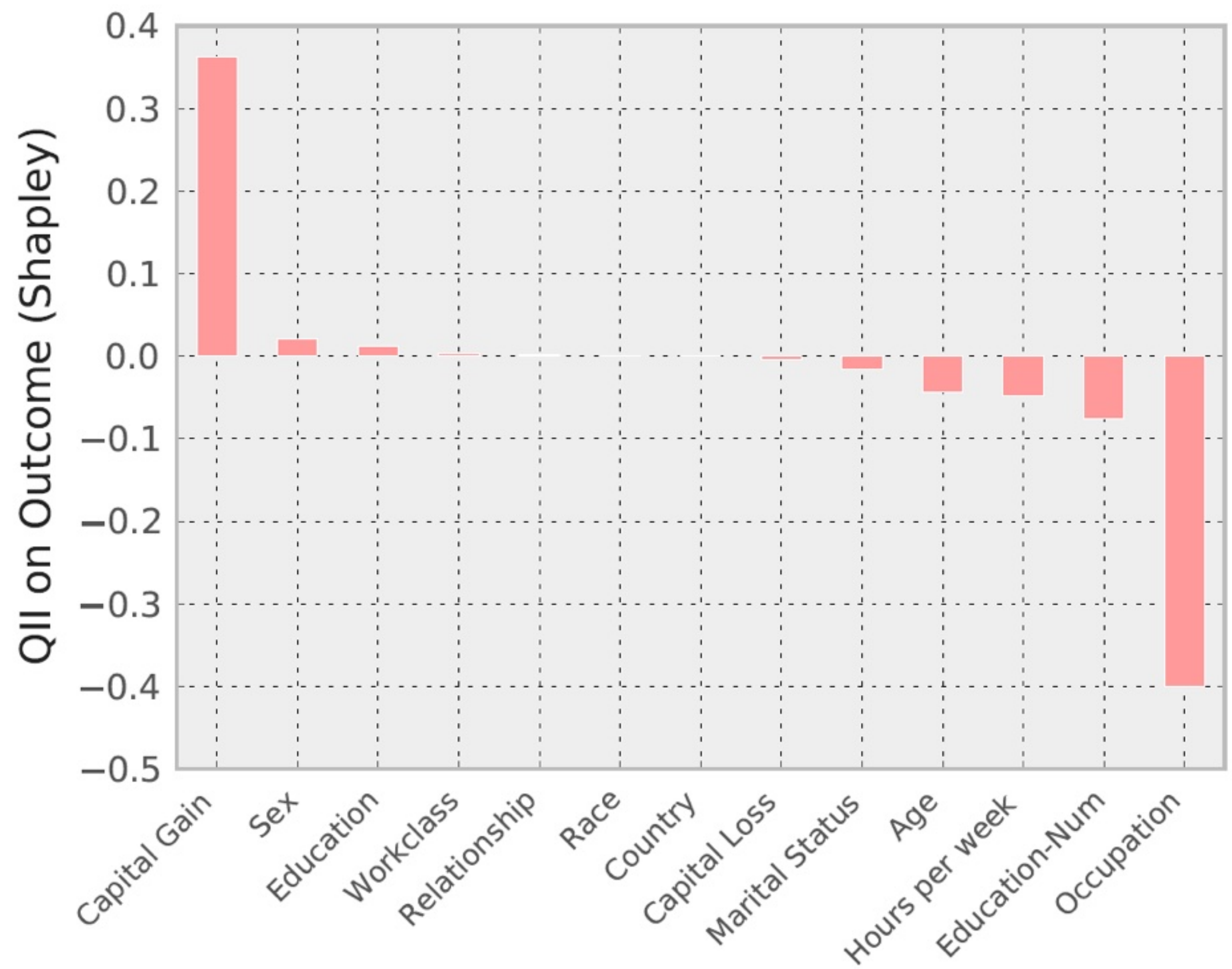
images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

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?

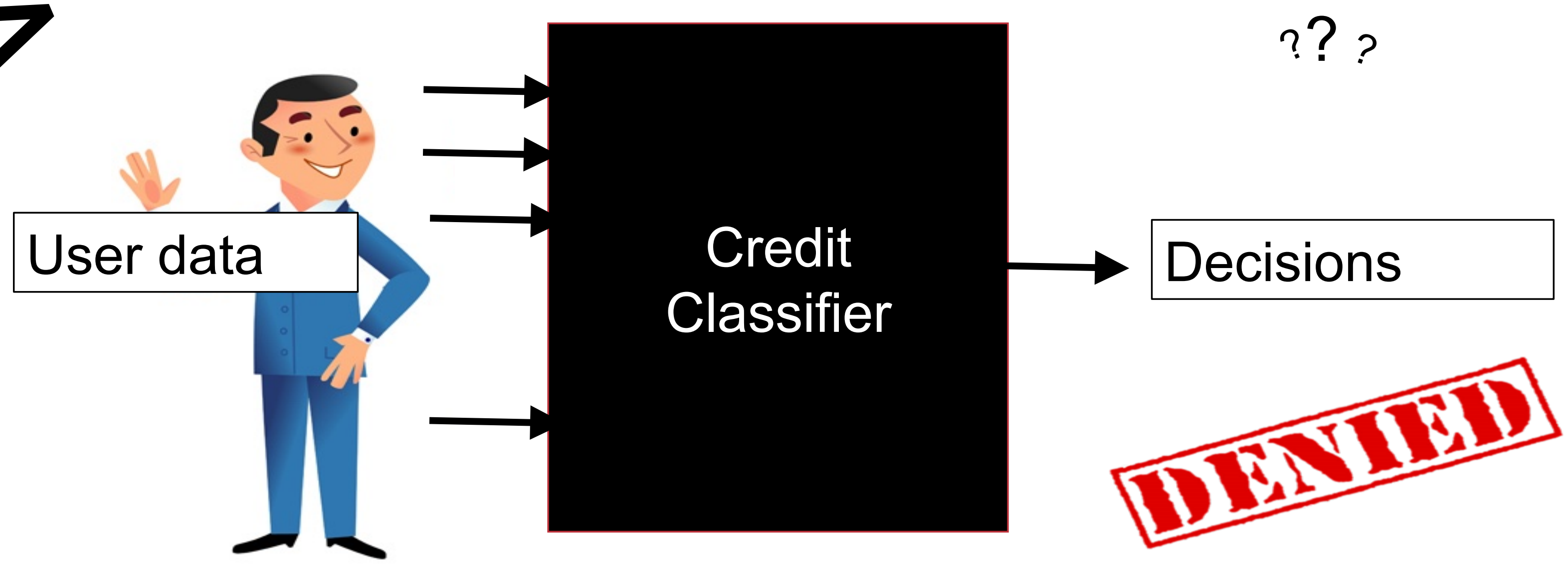
QII: Quantitative Input Influence

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**

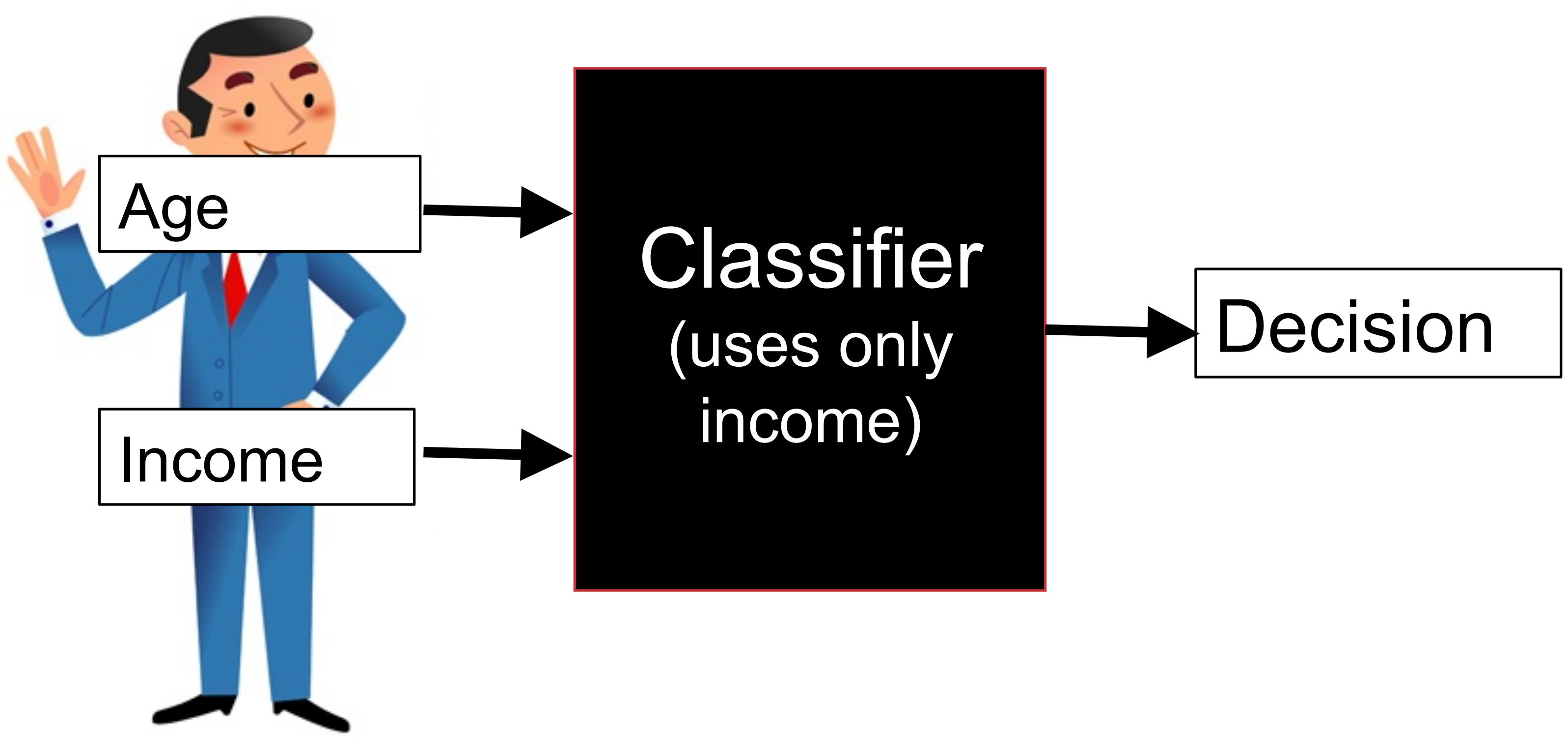
[Datta, Sen & Zick, 2016]

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**.

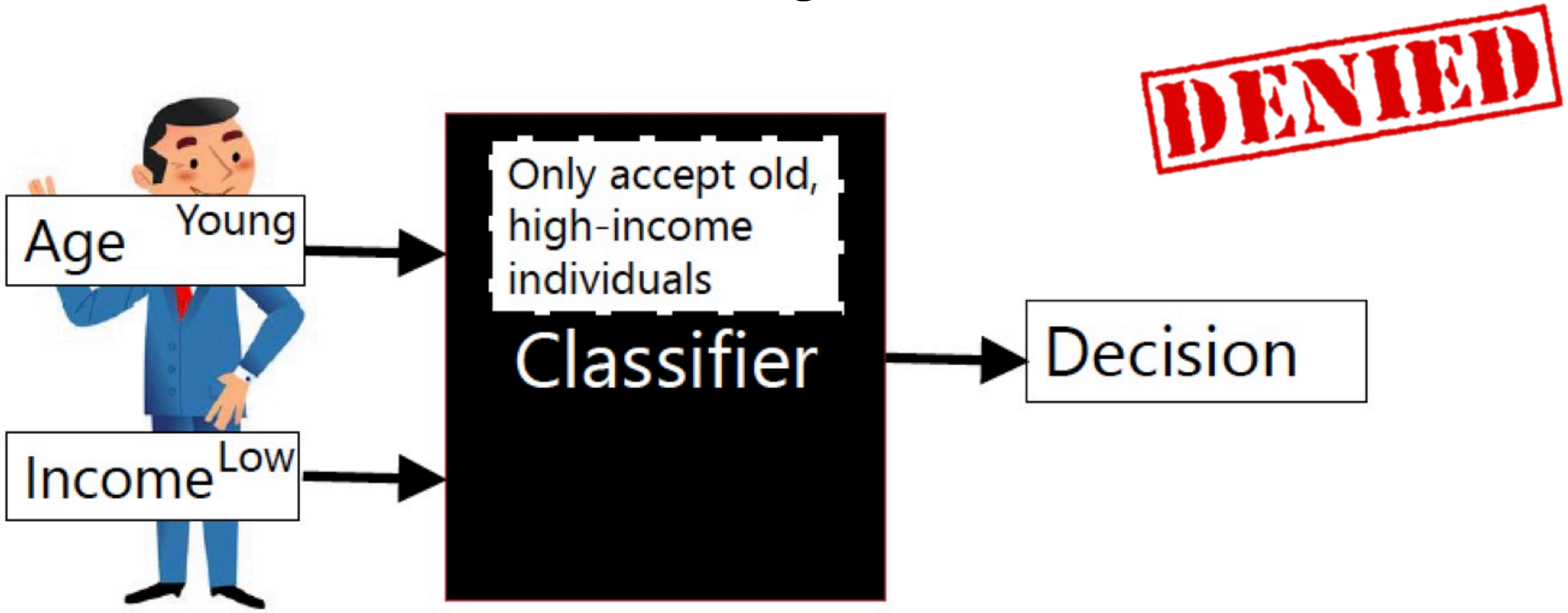


replace features with random values from the population, examine the distribution over outcomes

[Datta, Sen & Zick, 2016]

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]

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!

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

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

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

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

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

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

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

[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

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**
        $\mathbf{U} \sim \mathcal{D} \backslash \mathbf{v}, m$
        $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A} \backslash \mathcal{S}} \mathbf{U}_{\mathcal{S}}$
        $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A} \backslash \{\mathcal{S} \cup i\}} \mathbf{U}_{\mathcal{S} \cup i}$
        $\phi_{i_{\mathcal{S}}}(\mathbf{v}) = \iota(\mathbf{U}_1, \mathbf{U}_2)$
        $\phi_i(\mathbf{v}) = \phi_i(\mathbf{v}) + \frac{1}{d} \frac{1}{\binom{d-1}{|\mathcal{S}|}} \phi_{i_{\mathcal{S}}}(\mathbf{v})$
    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

1: $\phi = 0$
2: **for** $i \in \{1, \ldots, m\}$ **do**
3: $\quad \mathbf{u}_1 = \mathbf{U}_1(i)$
4: $\quad \mathbf{u}_2 = \mathbf{U}_2(i)$
5: $\quad \mathcal{D}_1 = \mathcal{D} \setminus \{\mathbf{v}\} \cup \{\mathbf{u}_1\}$
6: $\quad \mathcal{D}_2 = \mathcal{D} \setminus \{\mathbf{v}\} \cup \{\mathbf{u}_2\}$
7: $\quad \phi = \phi + r^{-1}_{\mathcal{D}_2, f}(\mathbf{u}_2) - r^{-1}_{\mathcal{D}_1, f}(\mathbf{u}_1)$
8: **end for**
9: **return** $\phi / |\mathbf{U}_1|$

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

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]

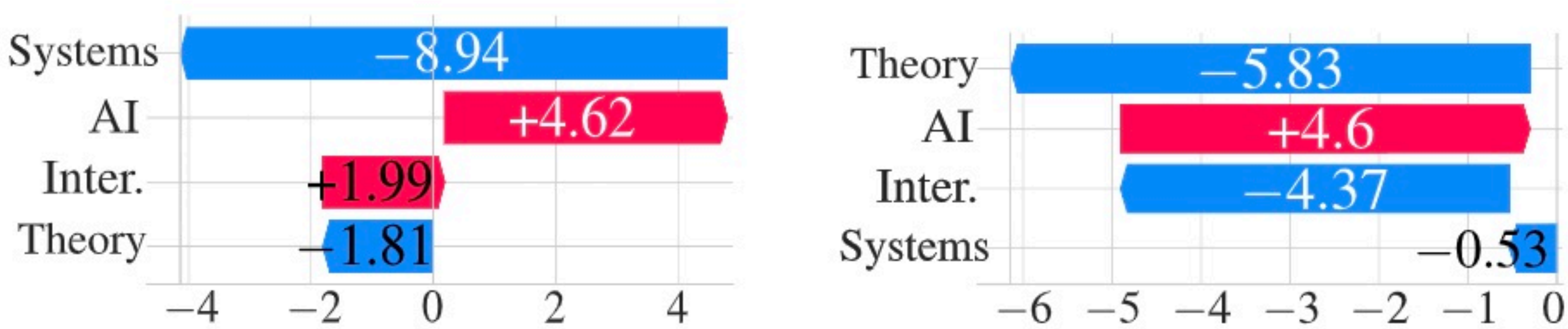
- ▶ Algorithms & complexity
- ▶ Cryptography
- ▶ Logic & verification

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
5	▶ Univ. of California - Berkeley	10.2	95
7	▶ University of Michigan	10.1	100
7	▶ University of Washington	10.1	81
9	▶ Stanford University	9.6	68
10	▶ Cornell University	9.3	83
11	▶ University of Maryland - College Park	8.6	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



(a) South Carolina, ranked 101 (b) Wayne State, ranked 102

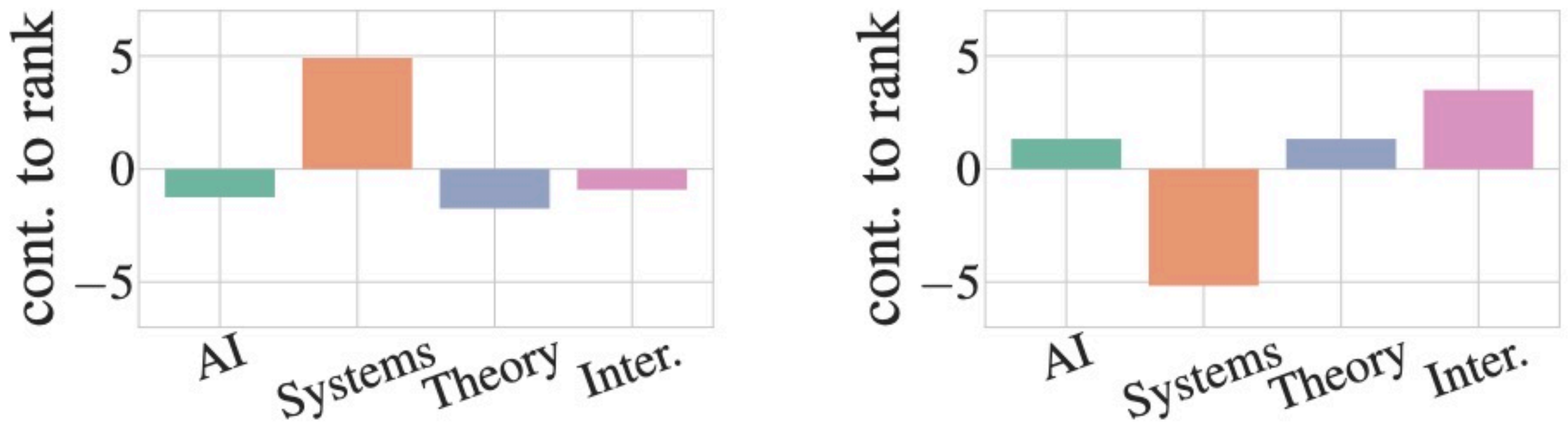
Figure 4: Feature contributions to rank QoI for two departments.

[Pliatsika, Fonseca, Wang, Stoyanovich, 2024]

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.



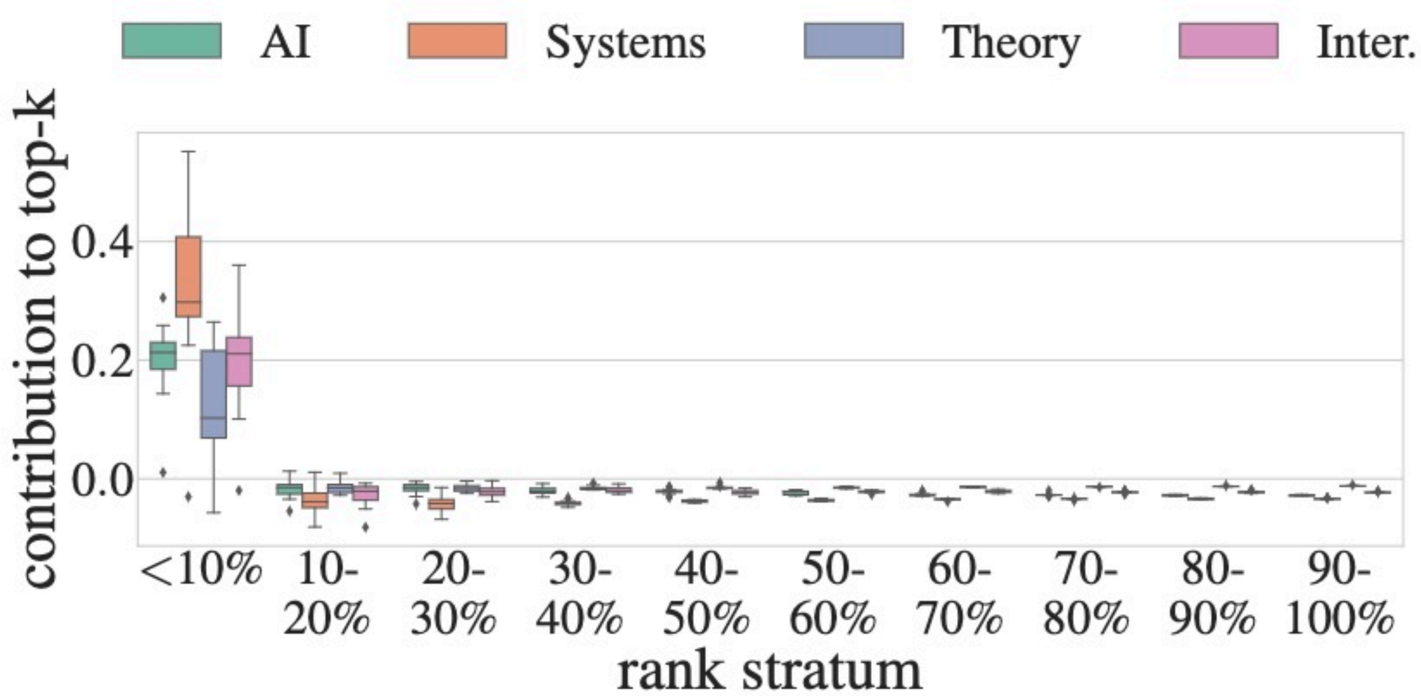
(c) Pairwise QoI explaining that Georgia Tech ranks higher than Stanford because of its relative strength in Systems.

(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]

Aggregates feature importance by rank stratum



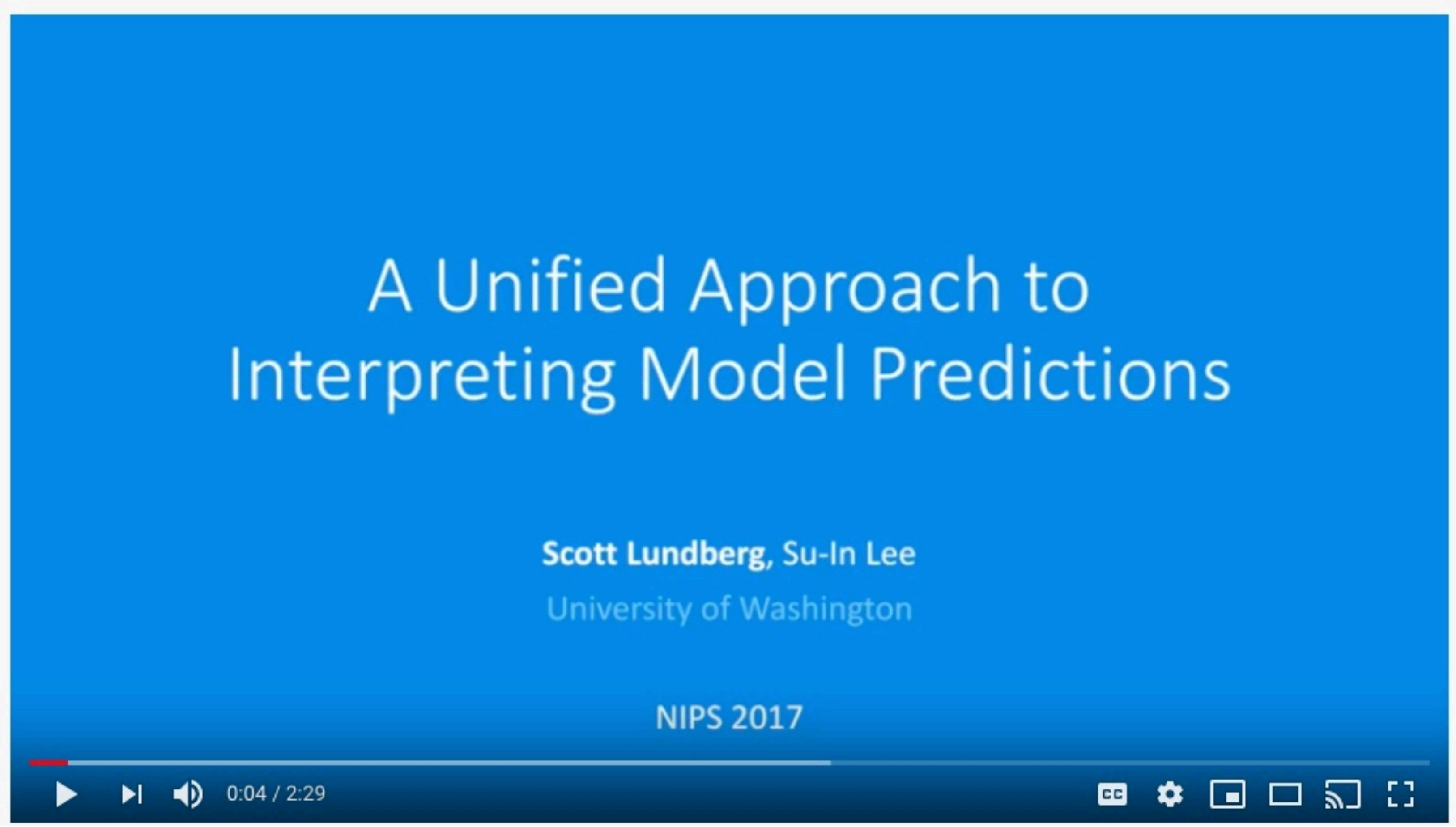
(a) Feature contribution to the top-k QoI, for $k = 10\%$. Systems is the most important feature, followed by Interdisciplinary and AI, while Theory is least important.

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.

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

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

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.

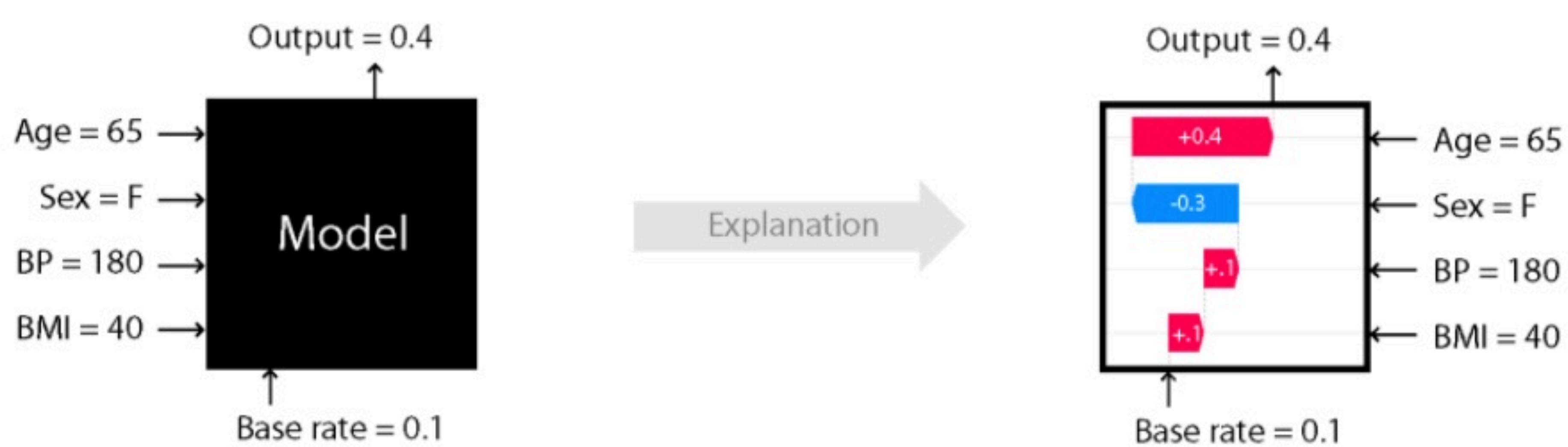
2. **Missingness**: if x'_i — the i^{th} feature of simplified input **x'**— is missing, then it has no attributable impact for ***x***

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)***

Additive feature attribution methods

README.md

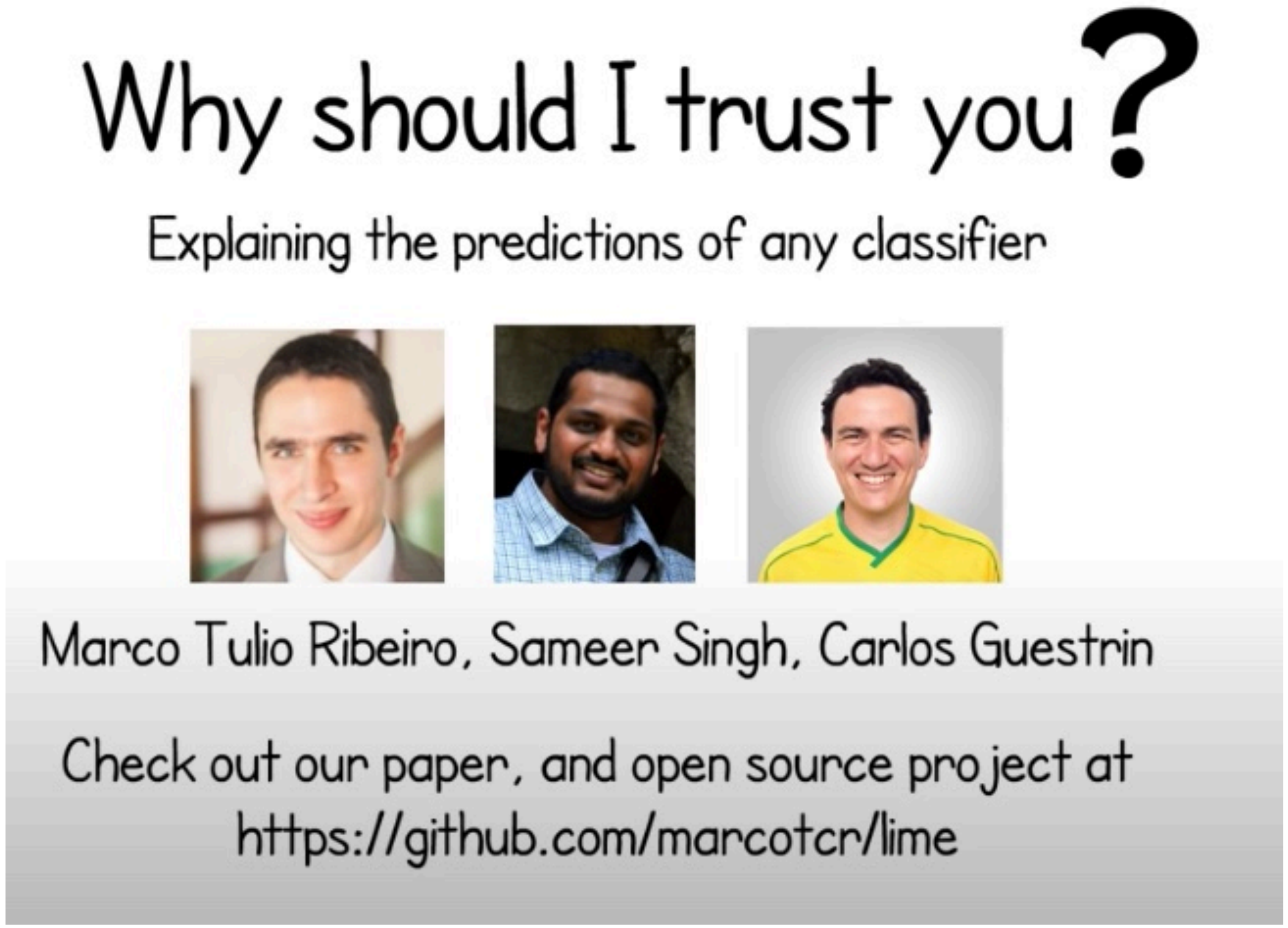
SHAP



https://github.com/slundberg/shap

[Lundberg & Lee, 2017]

LIME: Local Interpretable Model-Agnostic Explanations

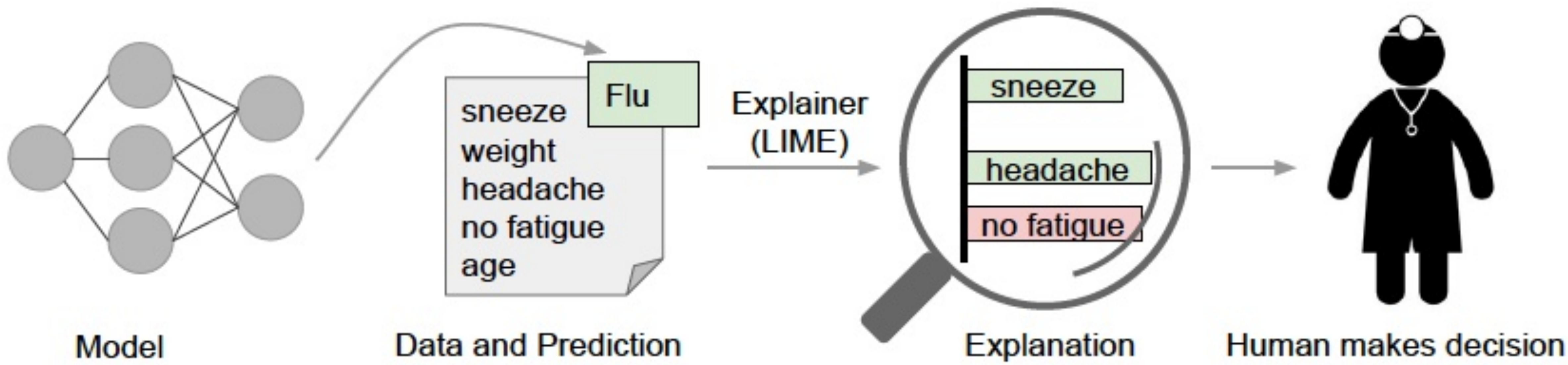


https://www.youtube.com/watch?v=hUnRCxnydCc

[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”

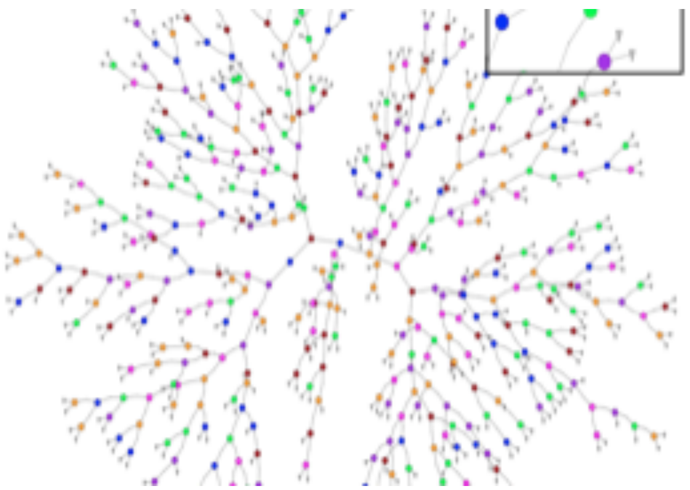
[Ribeiro, Singh & Guestrin, 2016]

LIME: Local explanations of classifiers

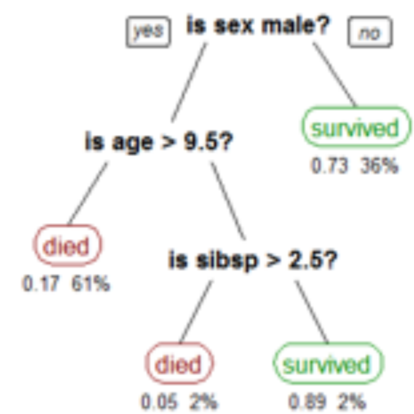
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely
not interpretable



Potentially
interpretable

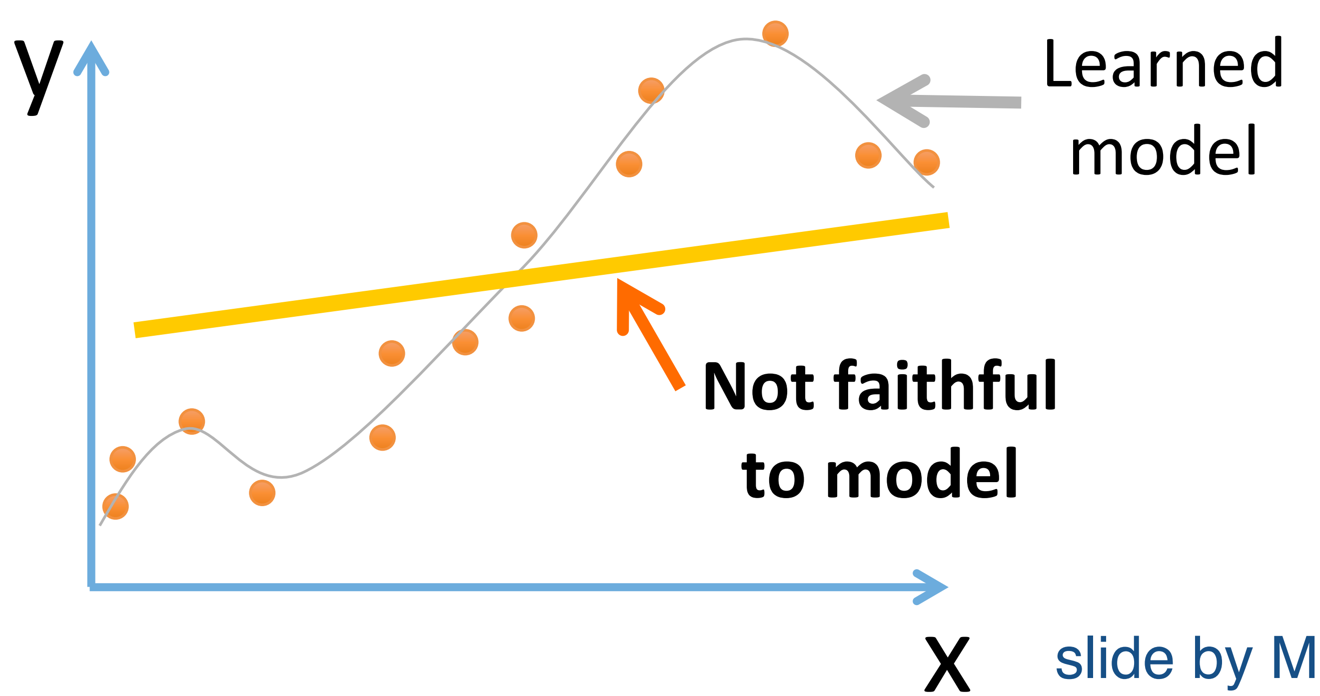
slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

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

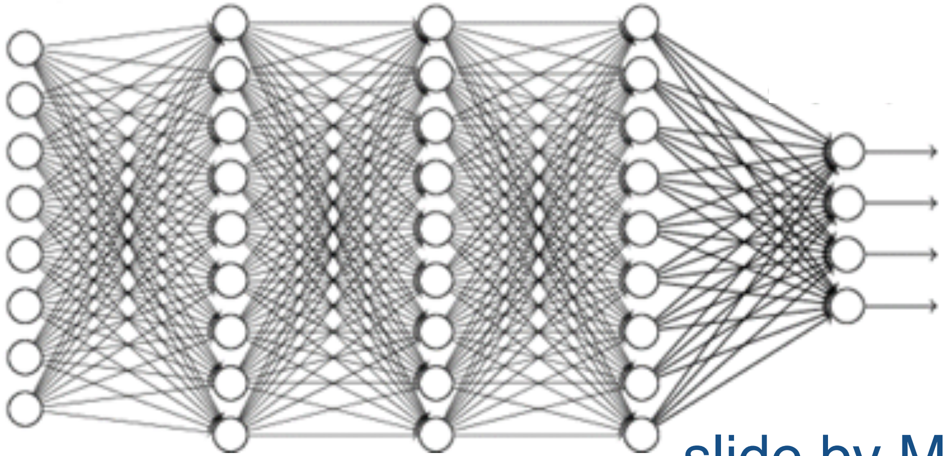
[Ribeiro, Singh & Guestrin, 2016]

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 $\mathbf{\{0, 1\}^{d'}}$ - 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."

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

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

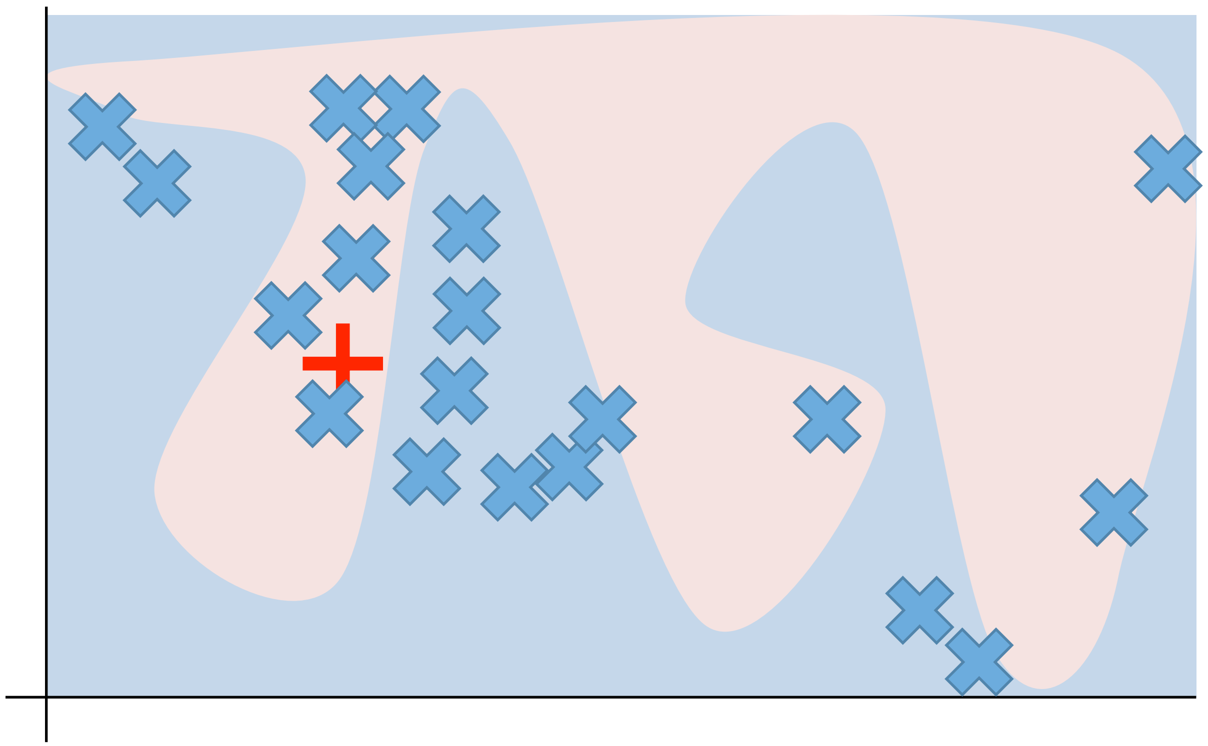
measures how unfaithful is ***g*** to ***f*** in the locality around ***x***

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

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 +

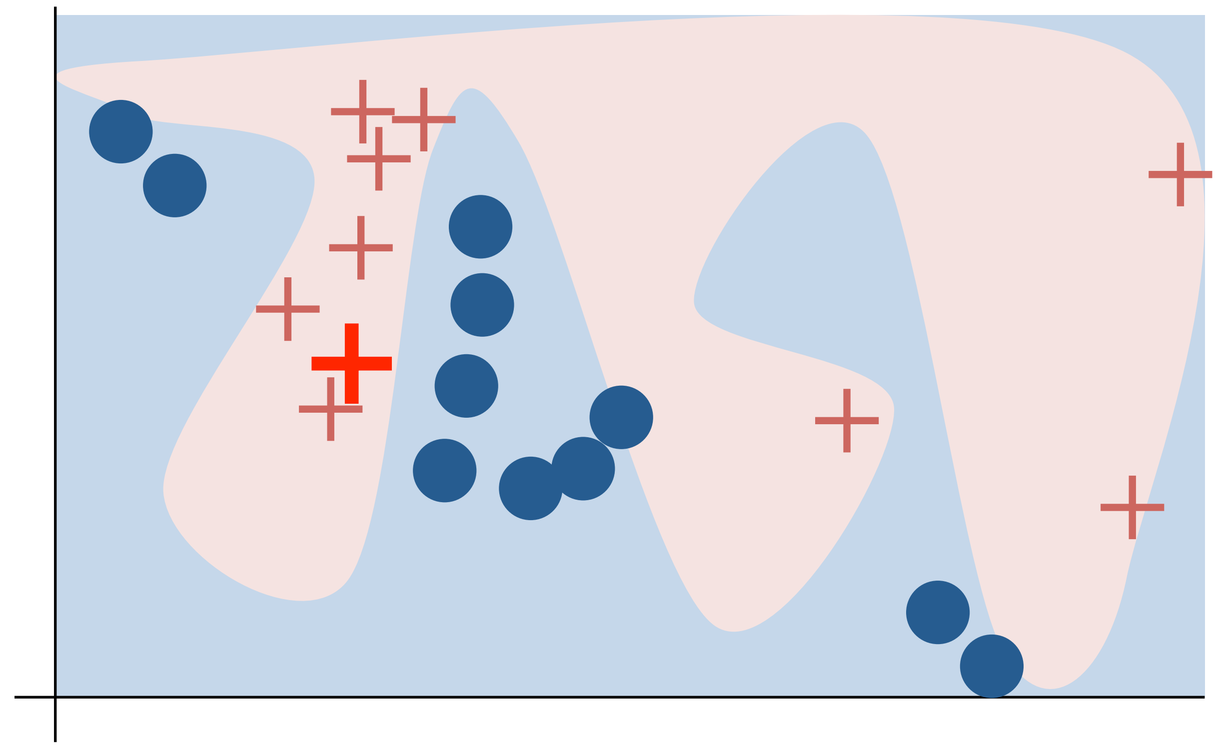


based on a slide by Marco Tulio Ribeiro, KDD 2016

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

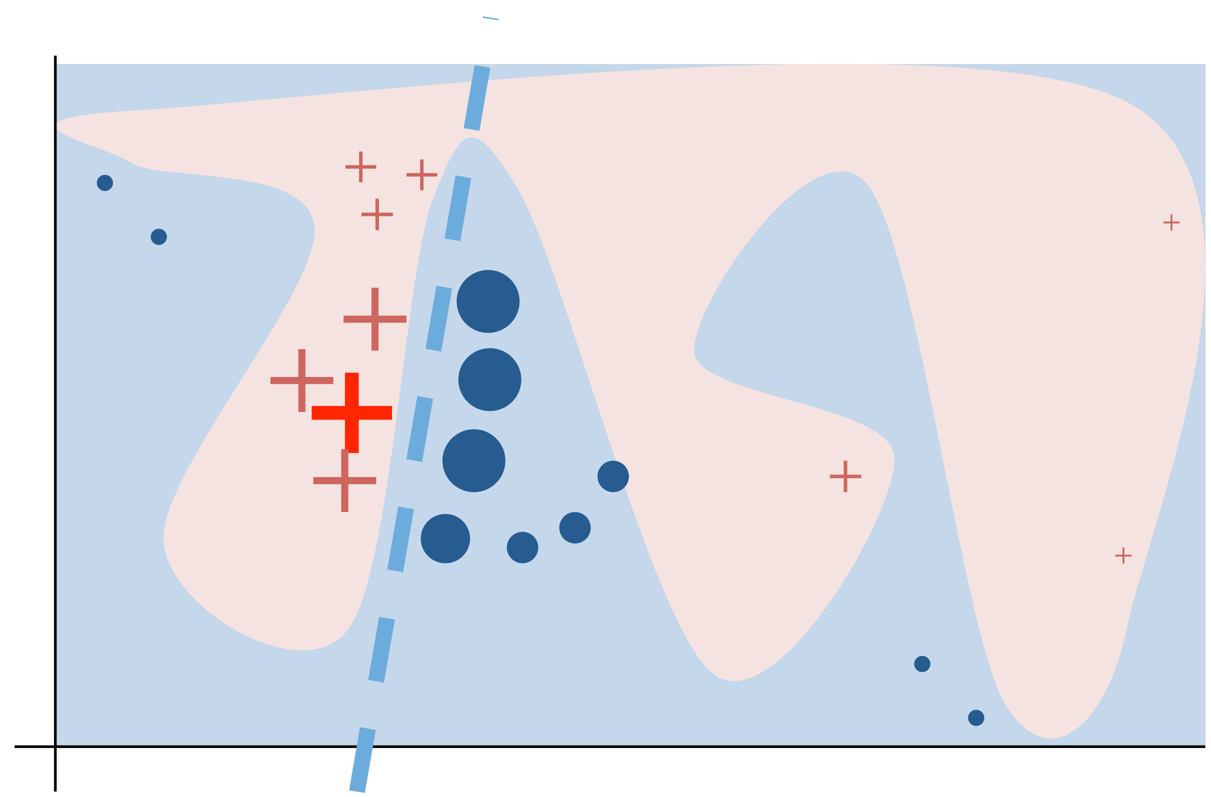


based on a slide by Marco Tulio Ribeiro, KDD 2016

Fidelity-interpretability trade-off

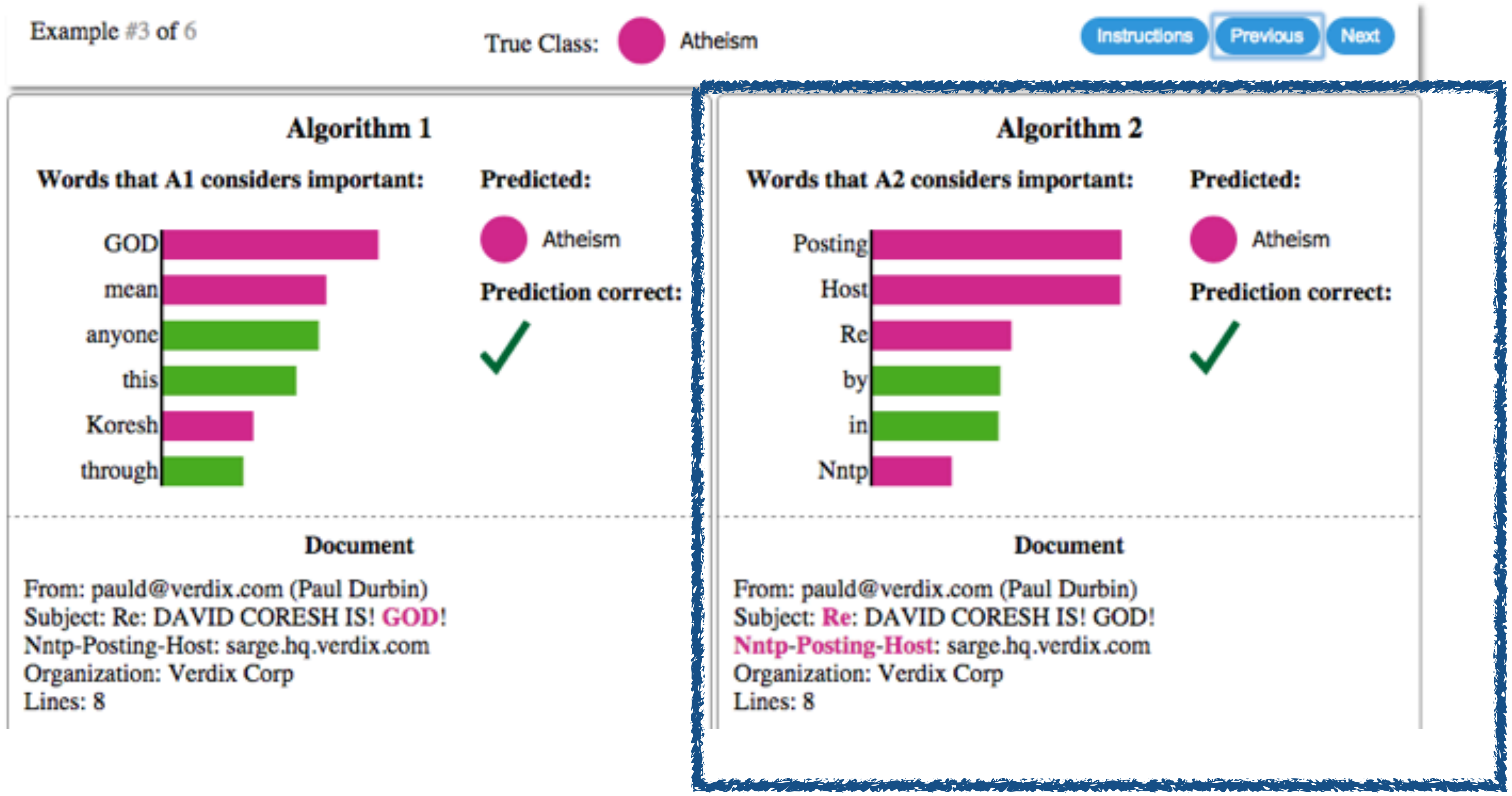
"The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classier."

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



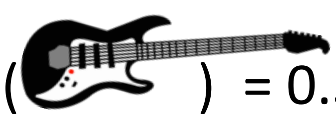
94% accuracy, yet we shouldn't trust this classifier!

[Ribeiro, Singh & Guestrin, 2016]

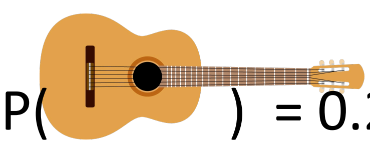
When accuracy is not enough

Explaining Google's Inception NN

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

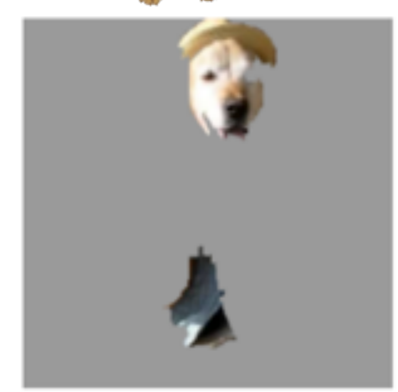
P() = 0.32

Electric guitar - incorrect but
reasonable, similar fretboard

P() = 0.24

Acoustic guitar

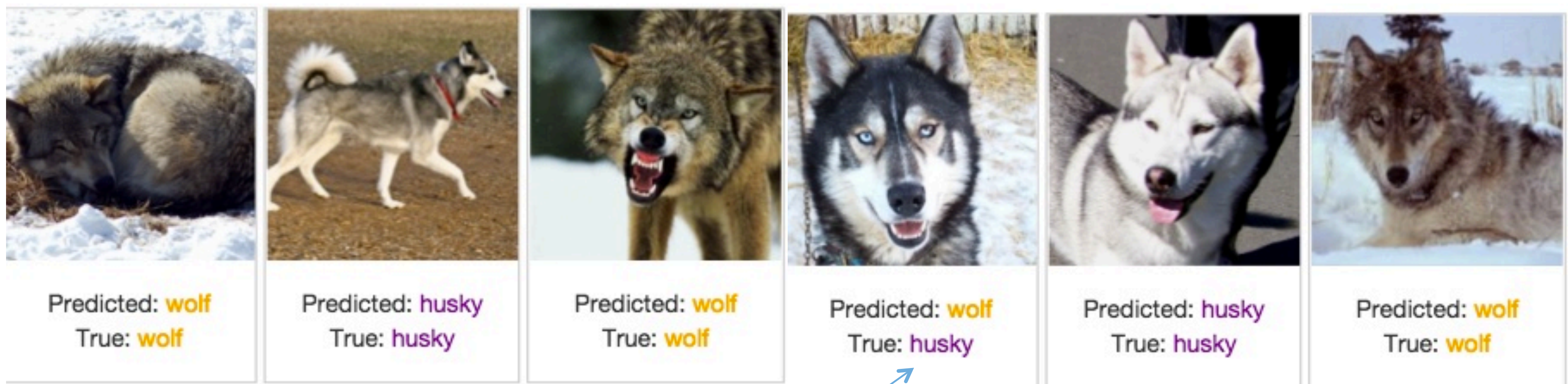
P() = 0.21



Labrador

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?

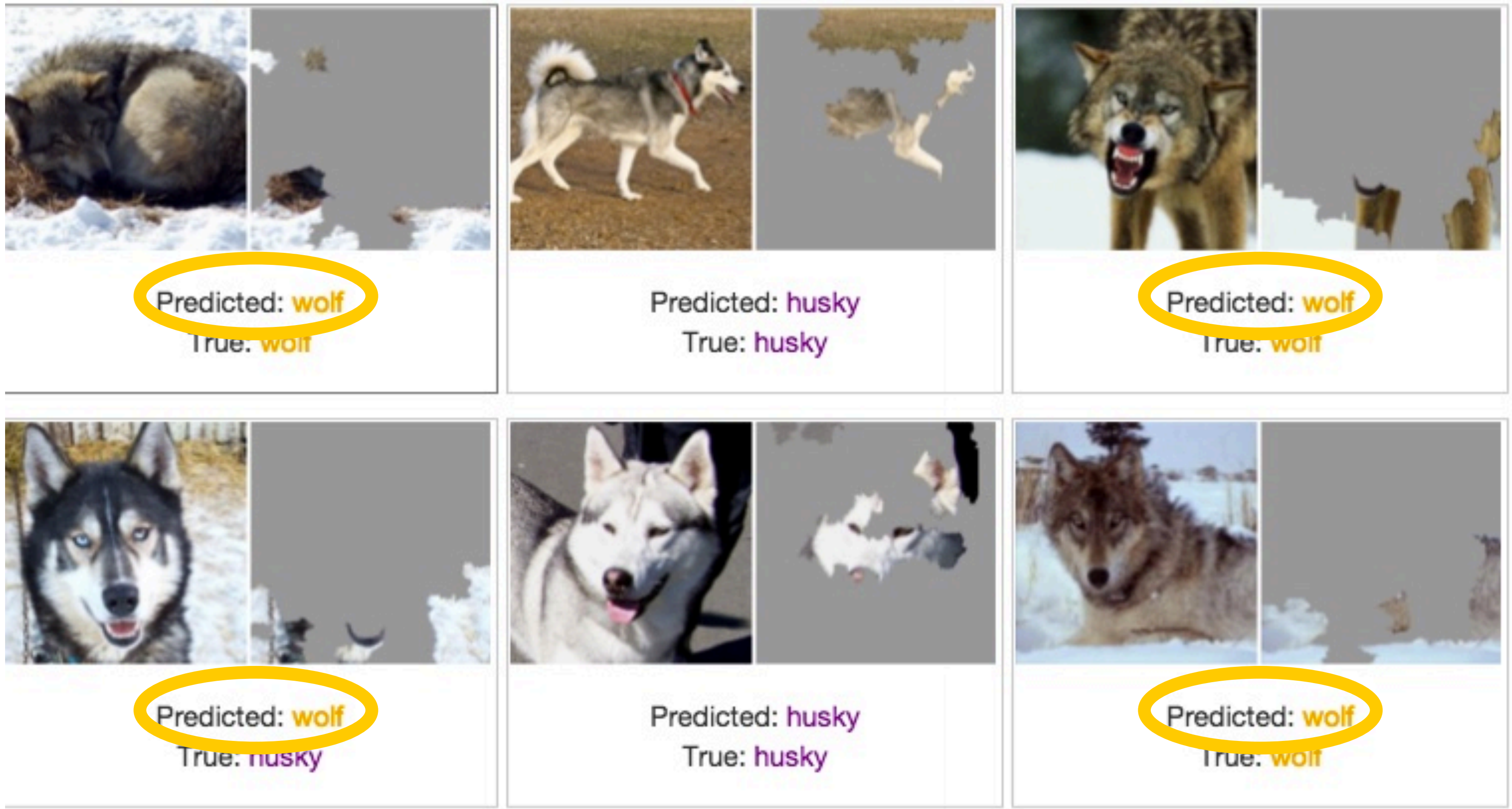
slide by Marco Tulio Ribeiro, KDD 2016

[Ribeiro, Singh & Guestrin, 2016]

r/ai

When accuracy is not enough

Explanations for neural network prediction



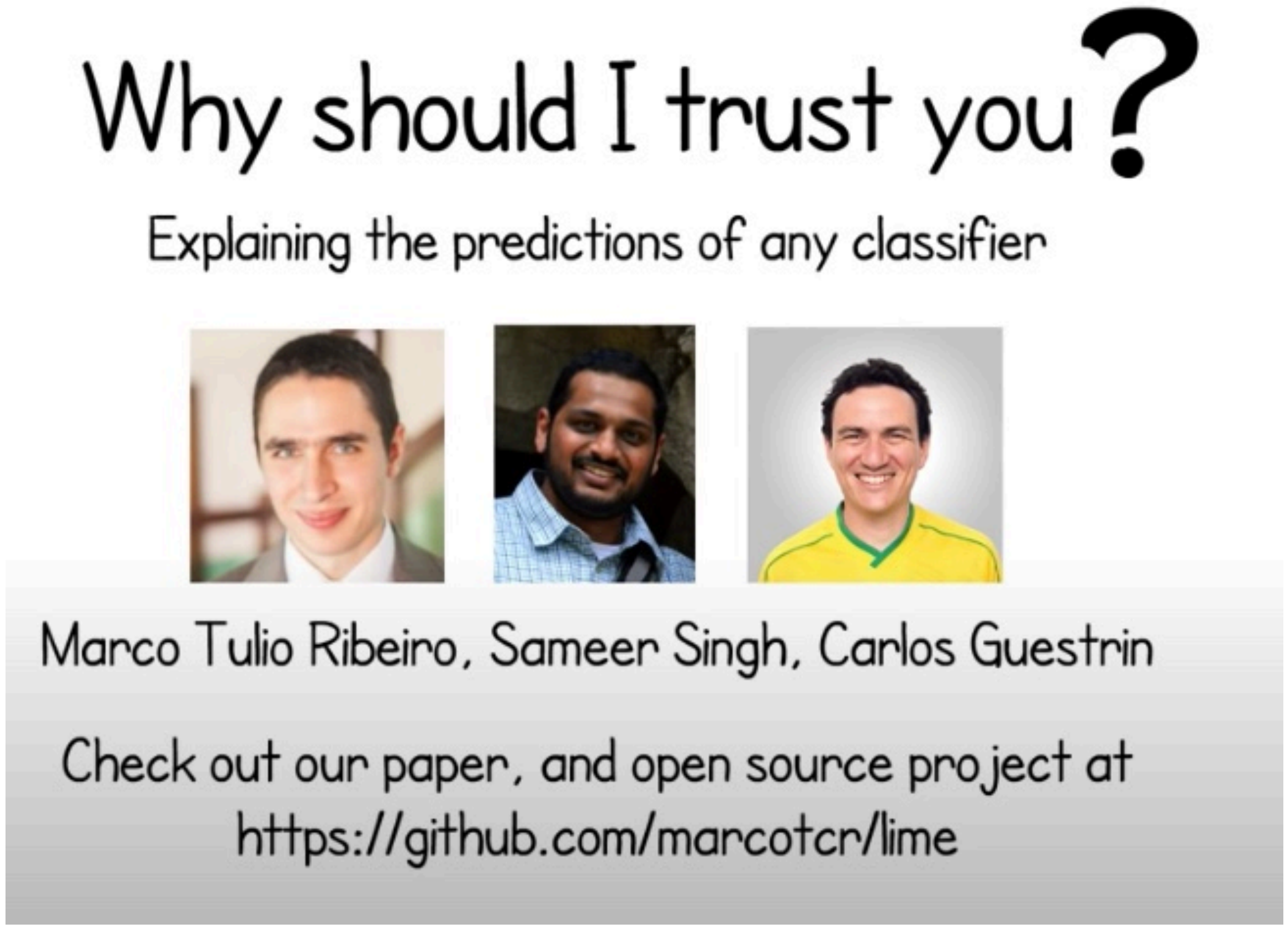
We've built a great snow detector… ☹

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]