

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

Terminology & vision

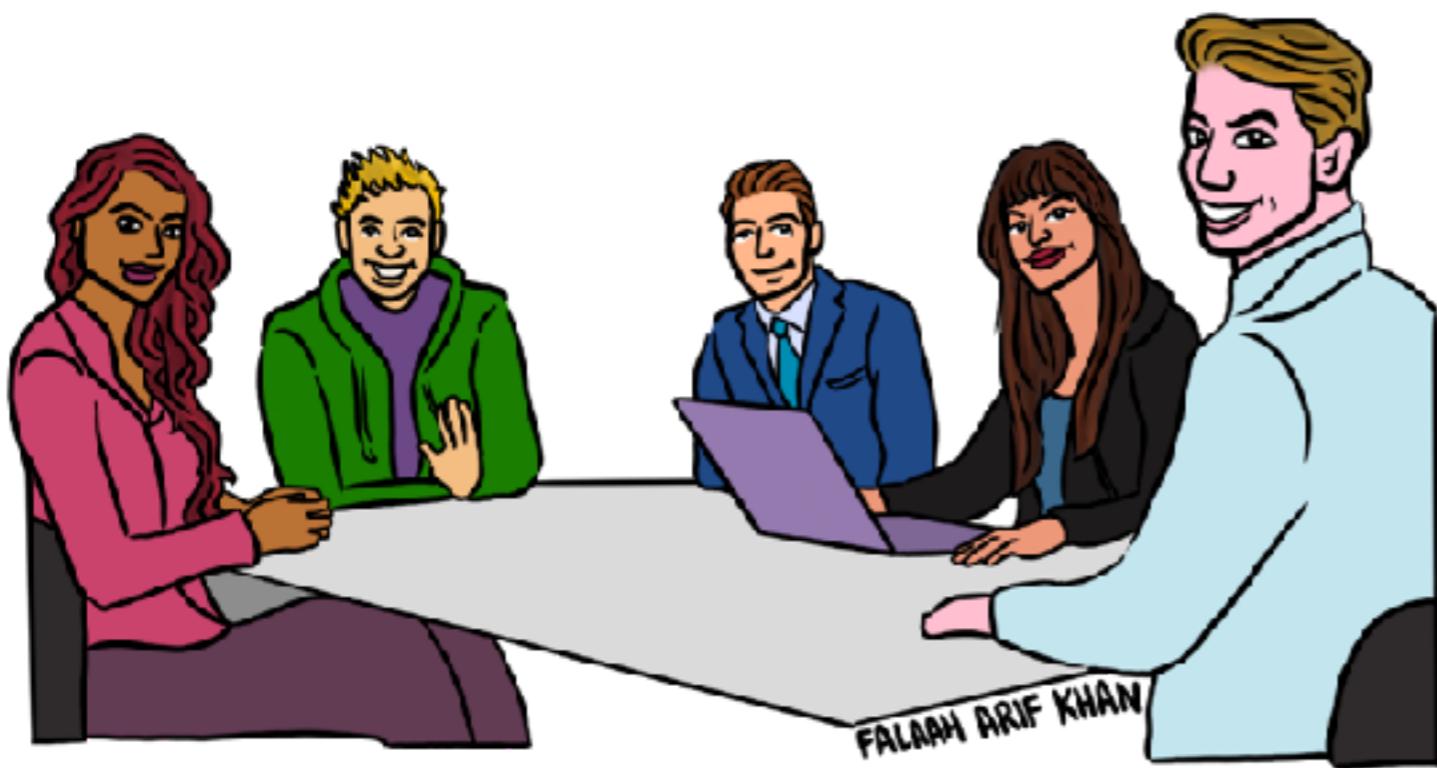


transparency, interpretability,
explainability, intelligibility



agency, responsibility

Interpretability for different stakeholders



What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

Staples discounts

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?



<https://www.wsj.com/articles/SB10001424127887323777204578189391813881534>

December 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was \$15.79, while the price on Trude Frizzell's screen, just a few miles away, was \$14.29.

A key difference: where Staples seemed to think they were located.

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.

r/ai

Staples discounts

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?



<https://www.wsj.com/articles/SB10001424127887323777204578189391813881534>

December 2012

It was the same Staples.com website, but the price for a basic printer was \$15.79, while a few miles away,

A key difference: Staples.com's servers are located.

A Wall Street Journal investigation found that the OfficeMax 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.

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

Online job ads

the guardian

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>

Online job ads

the guardian

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 habits, education, demographic

One experiment showed that Google displayed ads for a career coaching service for executive jobs

What are we explaining?
To **Whom** are we explaining?

and only 318 times to the female group.

Another experiment found a similar trend

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



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

LATANYA SWEENEY
3420 Centre Ave
Pittsburgh, PA 15216
DOB: Oct 27, 1969 (13 years old)

Criminal History
This section contains possible citation, arrest, and criminal history. While our database does contain hundreds of millions of pieces of information they will not release.

We share with you as much information as we possibly can that Latanya Sweeney has never been arrested; it simply is in the data that is available to us.

Possible Matching Arrest Records

Name	County and State
No matching arrest records were found.	

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

Nutritional labels

SIDE-BY-SIDE COMPARISON

Original Label

Nutrition Facts

Serving Size 2/3 cup (55g)	Calories from Fat: 72
Amount Per Serving	
Calories 230	Calories from Fat: 72
Total Fat 8g	% Daily Value*
Saturated Fat 1g	12%
Trans Fat 0g	5%
Cholesterol 0mg	0%
Sodium 150mg	7%
Total Carbohydrate 35g	12%
Dietary Fiber 4g	16%
Sugars 1g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

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

Total Fat	Less than 40g	40g
Sat. Fat	Less than 90g	90g
Cholesterol	Less than 300mg	300mg
Sodium	Less than 1,400mg	1,400mg
Total Carbohydrate	300g	375g
Dietary Fiber	25g	30g

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 label, 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

New Label

Nutrition Facts

6 servings per container	Serving size 2/3 cup (55g)
Amount per serving	
Calories 230	Calories from Fat: 72
Total Fat 8g	% Daily Value*
Saturated Fat 1g	12%
Trans Fat 0g	5%
Cholesterol 0mg	0%
Sodium 150mg	7%
Total Carbohydrate 35g	12%
Dietary Fiber 4g	16%
Sugars 1g	
Protein 3g	
Vitamin A	10%
Vitamin C	8%
Calcium	20%
Iron	45%

*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.

Security & Privacy Overview Smart Device Co.

Smart: Model: Bonshell NSG200
Firmware version: 2.5.1 - updated on: 11/12/2020
The device was manufactured in China

Security Mechanisms	Security updates: Automatic - Available until at least 1/1/2022	Access control: Password - Policy default - User changeable. Multi-factor authentication. Multiple user accounts are allowed.
Data Practices	Sensor data collection: Sensor type: Visual, Audio, Physiological, Location Purposes: Data stored on devices: Data shared on cloud: Sharing with: Sold to: Other collected data: Privacy policy: www.NSG200.smartdeviceco.com/policy	1
2	3	4

What are we explaining?

To Whom are we explaining?

Why are we explaining?

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.

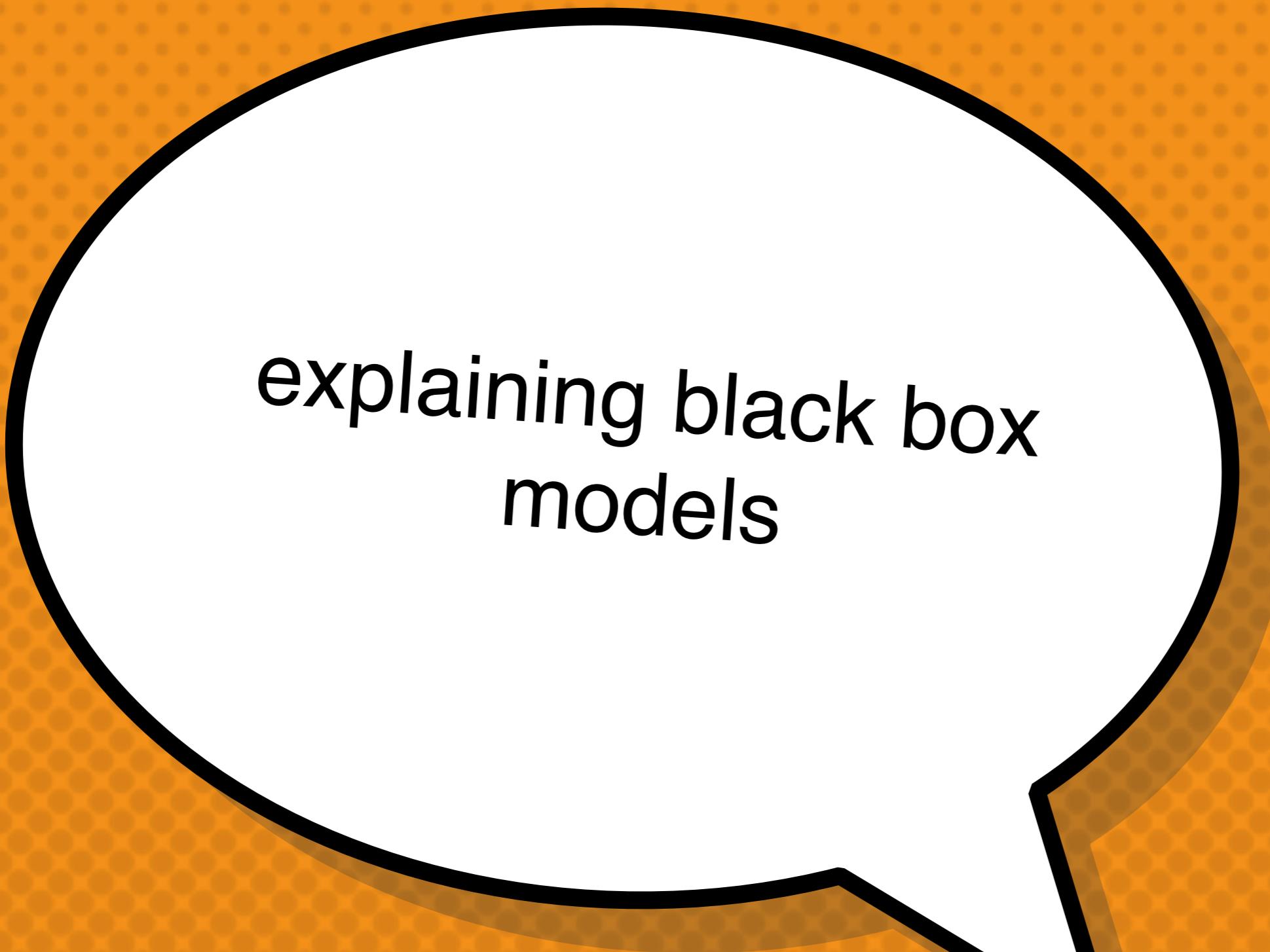
Additional assessment: AI-assisted personality scoring

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/why-the-labels-on-your-food-are-changing-or->

<https://www.wsj.com/articles/imagine-a-nutrition-label-for->

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



explaining black box
models

This week's reading

2016 IEEE Symposium on Security and Privacy

QII

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

Anupam Datta Shyukl Jain Yair Zick
Carnegie Mellon University, Pittsburgh, USA
{datta, shyukl, yairick}@cs.cmu.edu

Abstract Algorithmic systems that employ machine learning play an increasing role in making substantive decisions in modern society, ranging from online personalization in 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 measure 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 and audit for internal and external oversight (e.g., to detect algorithmic discrimination).

bio Anupam Datta (JII) measures carefully account for correlated inputs while measuring influence. They support a general class of transparency stories 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) or extremes (e.g., loan拒绝) and the marginal influence of individual inputs within such a set of inputs. Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using weighted aggregated influences, such as the Shapley values, previously applied to measure influence in voting. Further, since transparency reports could compromise privacy, we explore the transparency privacy tradeoff and consider a notion of useful transparency reports that are made differentially private with very little additional noise.

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?

1. 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, healthcare, 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.

© 2016, Anupam Datta. Published by IEEE.
DOI 10.1109/SP.2016.75

299

@carnegie
Society

LIME

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

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98195, USA
marcotri@cs.washington.edu

Sameer Singh
University of Washington
Seattle, WA 98195, USA
sameer@cs.washington.edu

Carlos Guestrin
University of Washington
Seattle, WA 98195, USA
guestrin@cs.washington.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 uninterpretable model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the prediction 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 concise redundant way, keeping the task over a suitable quadratic scale. We demonstrate the feasibility of these methods by exploring a linear model 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 we should trust a prediction, choosing between models, depending on automatically classified, and identifying why a classifier should not trust itself.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of learning is an oft-neglected aspect in the field. At other times we directly use existing learning classifiers as tools, or re-implement models within other products, a vital concern remains: if the user does not trust a model or a prediction, they will not use it. It is important to differentiate between two different (but related) dimensions of trust: (1) a user's preference, i.e., whether a user trusts an individual and thus 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

Promotion use only digital or hard copies of all or part of this work for personal reference is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise or republish, permission must be obtained in writing from IEEE. Requests for permission to copy should be addressed to: IEEE Copyrights and Permissions Department, 3 Park Avenue, New York, NY 10016-5994, USA.

DOI 10.1109/SP.2016.75
© 2016, Copyright held by the author(s). Publication rights licensed to IEEE. ISBN 978-1-5090-1322-2/16/05 \$31.00
http://dx.doi.org/10.1109/SP.2016.75

how much the human understands a model's behavior, as opposed to seeing it as a black box.

Demanding 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 terrorist detection, for example, predictions cannot be void and must be valid, 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 separate metrics on an available validation dataset. However, real-world data is often significantly different, and hence, 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 each metric. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose LIME, a novel explanation technique that explains the prediction 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 concise redundant way, keeping the task over a suitable quadratic scale. We demonstrate the feasibility of these methods by exploring a linear model 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 we should trust a prediction, choosing between models, depending on automatically classified, and identifying why a classifier should not trust itself.

1.1. LIME as an algorithm that can explain the prediction of any classifier or regressor in a faithful way by approximating it locally with an interpretable model.

1.2. SIMPLICE, a method that selects a set of representative instances with explanations to address the "trust me" problem, via enhanced local optimization.

1.3. Comparative 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 generates a better fit in the real world. Further, they are able to greatly improve an uninterpretable classifier trained on 20 newsgroups, by doing feature engineering using LIME. We also show how understanding the prediction of a new model or images help practitioners know what and why they should trust a model.

2. THE CASE FOR EXPLANATIONS

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

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
slund@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
sulinlee@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 model databases 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 which 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 *explainer model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unites six current methods.

<https://github.com/slundberg/shap>

¹3rd Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

ShaRP

ShaRP: Explaining Rankings with Shapley Values

Venelin Plotsilka¹, Jone Fonseca^{2,3}, Tithi Wang¹ and Julia Stoyanovich¹

¹New York University, NY, USA ²NOVA University, Lisbon, Portugal
³{venelin, tw2221, stoyanovich}@nyu.edu, jfonseca@novau.edu.pt

Abstract

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

In this paper, we present ShaRP (Shapley in Rankings and Preferences), a framework that explores the contributions of features to different aspects of a ranked outcome, and is based on Shapley values. Using Shapley, we show that even when the scoring function used by an algorithmic maker is known and linear, the weight of each feature does not correspond to 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 Quantiles of interest, including score, rank, pairwise preference, and log-lik. 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, showing its usefulness for qualitative analysis.

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

name	gpa	sat	score	f	p	rank	label
Bob	4	5	5	43.5	0.0	Bob	Bob
Tom	4	5	5	43.5	0.0	Tom	Tom
Dan	2	4	4	44.4	0.1	Dan	Dan
Kira	4	2	3	42.3	0.1	Kira	Kira
Edu	2	4	3	42.3	0.1	Edu	Edu
Kat	2	4	3	42.3	0.1	Kat	Kat
Lori	4	4	3	48.3	0.9	Lori	Lori
Sara	2	3	3	33.3	0.1	Sara	Sara

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

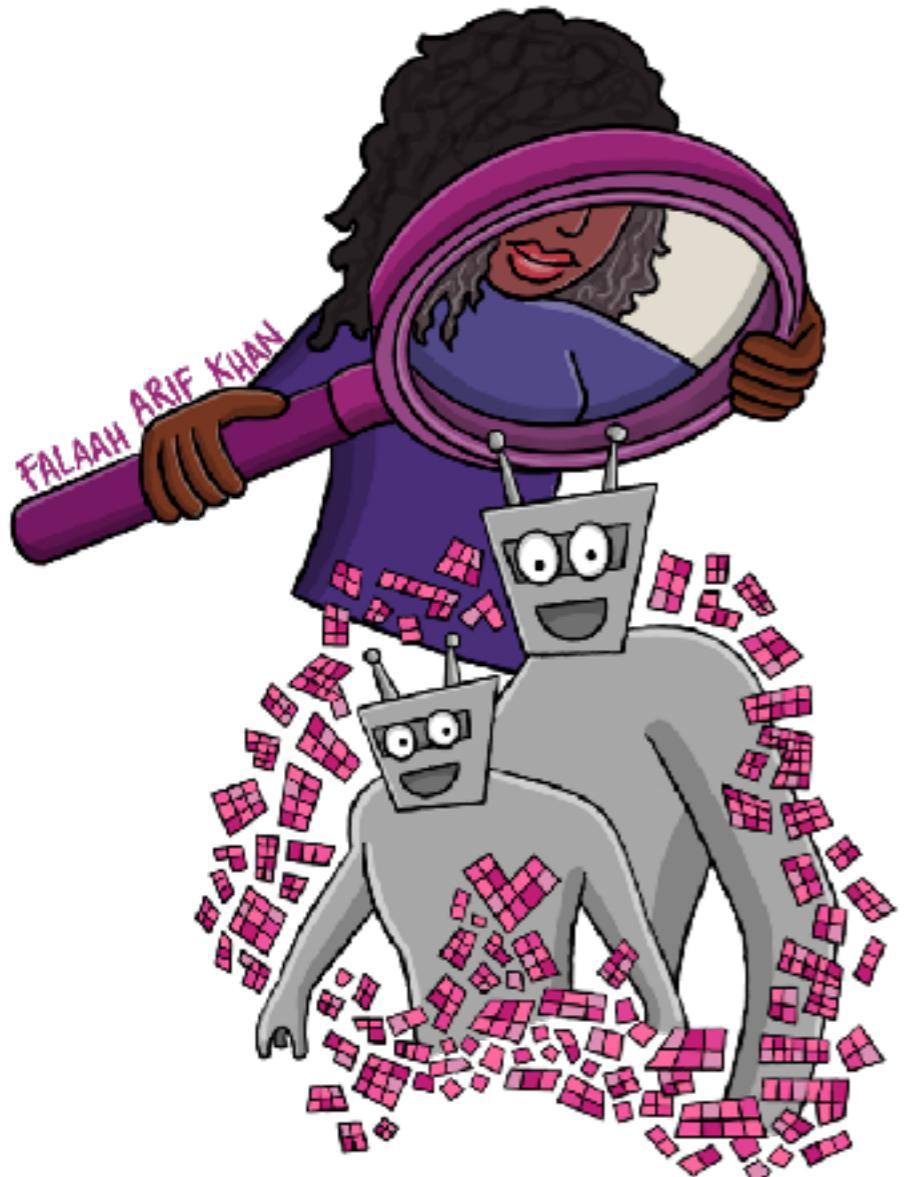
There are two types of ranking: score-based and learned. In score-based ranking, a given set of candidates is sorted on a score, which is typically computed using a simple formula, such as a sum of attribute values with non-negative weights (Stoyanovich et al. (2022)). In supervised learning-to-rank, 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” Radu (2019); the scoring function, such as $f_i = 0.4 \times \text{gpa} + 0.4 \times \text{sat} + 0.2 \times \text{score}$ at a college admissions domain, is based on (intuitive) a priori understanding of what makes for a good candidate.

Alas, just despite being practically “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); Werner (2020)). We now illustrate this with a simple example.

Example 1. Consider a dataset \mathcal{D} of college applicants at Figure 1, with scoring function $f = 0.4 \times \text{gpa} + 0.4 \times \text{sat} + 0.2 \times \text{score}$ and a $p = 1.0 \times \text{score}$ factor very similar numbers $\pi_{\mathcal{D}}$ and $\pi_{\mathcal{D}}^f$ with the same top-4 items appearing in the $\pi_{\mathcal{D}}$ and $\pi_{\mathcal{D}}^f$.

r/ai

What are we explaining?



How does a system work?

How **well** does a system work?

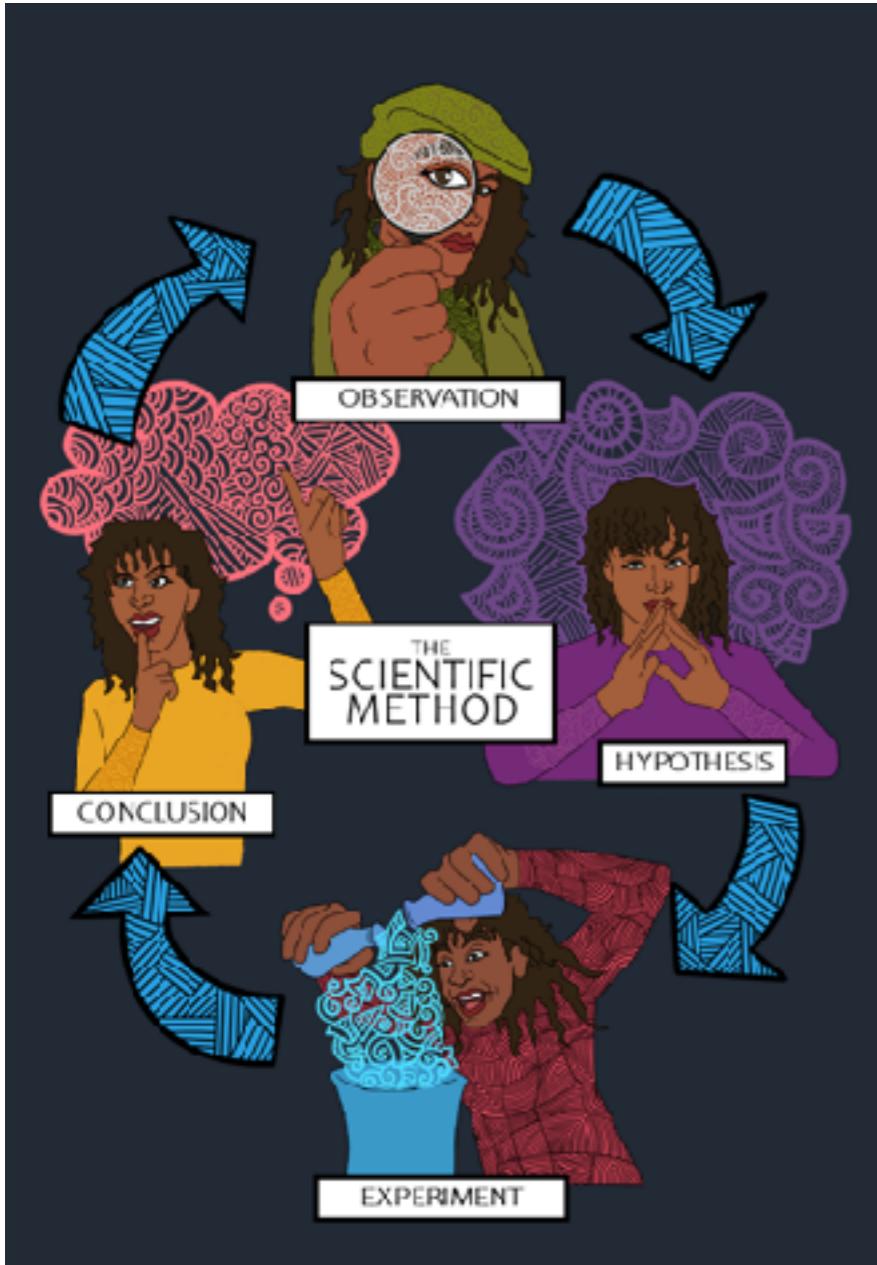
What does a system do?

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?

Facebook's real-name policy

← **Tweet**

Shane Creepingbear is a member of the Kiowa Tribe of Oklahoma



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

October 13, 2014

≡ **TIME**

17

Facebook Thinks Some Native American Names Are Inauthentic

BY JOSH SANBURN FEBRUARY 14, 2015

February 14, 2015

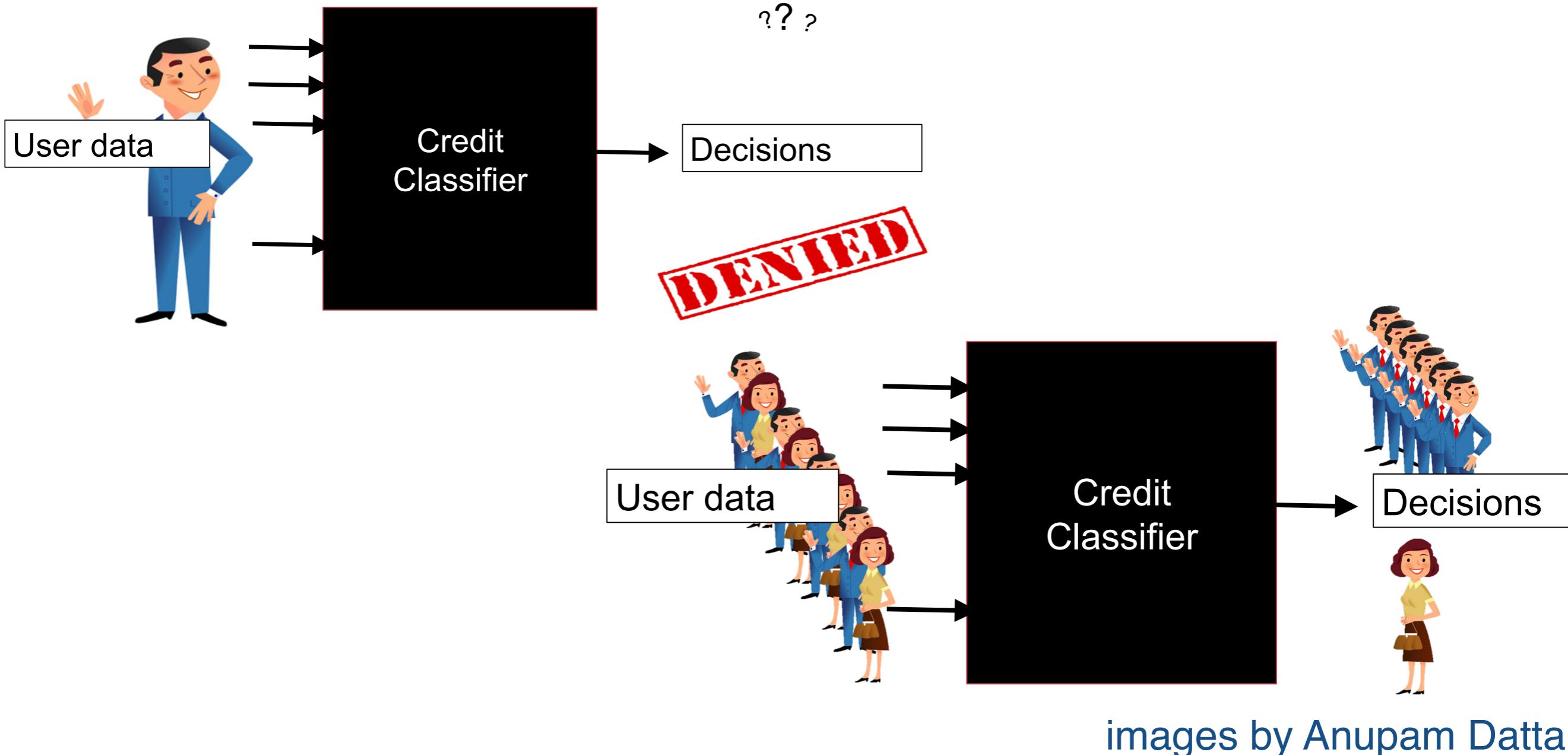
If you're Native American, Facebook might think your name is fake.

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

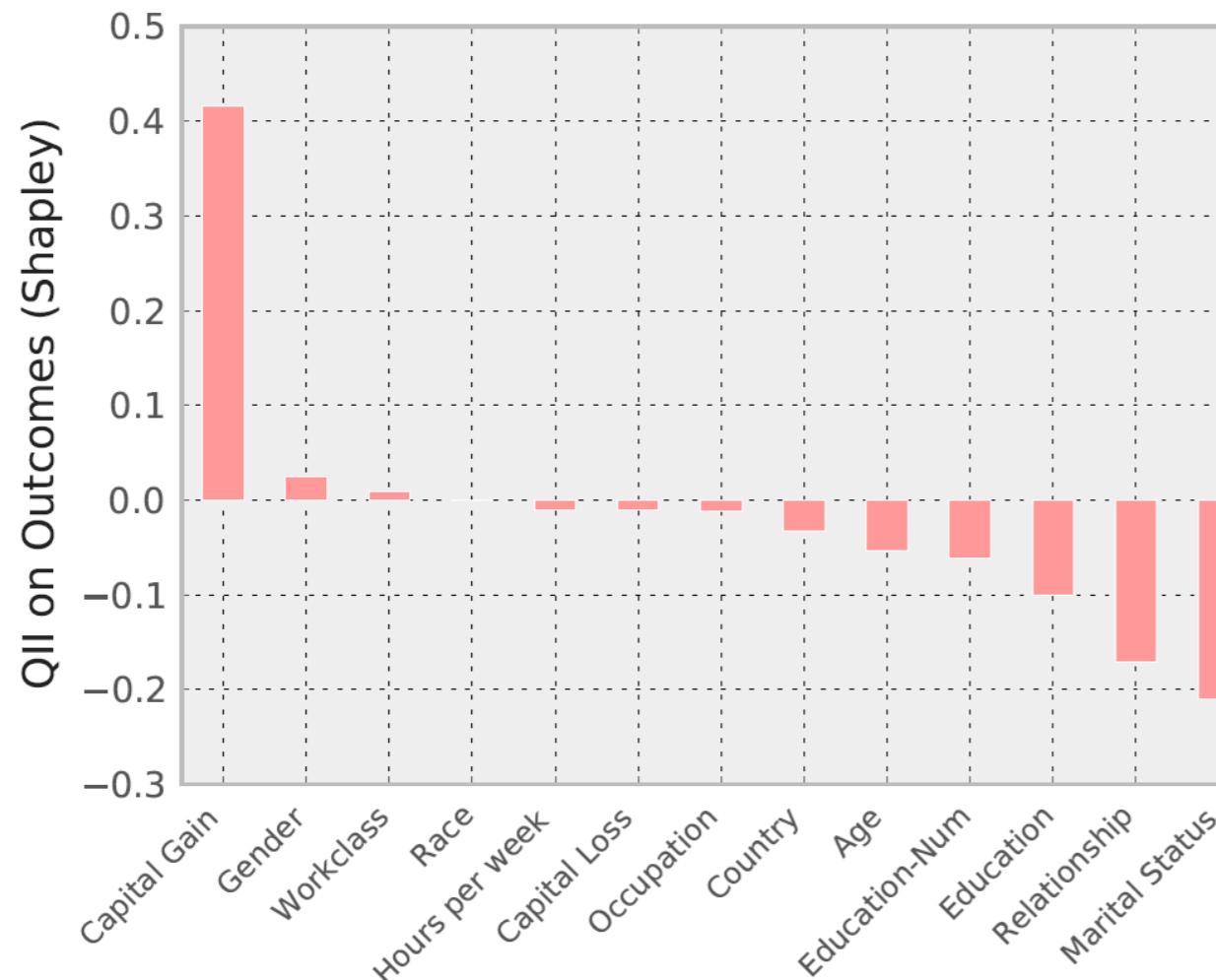
QII: Auditing black-box models



images by Anupam Datta

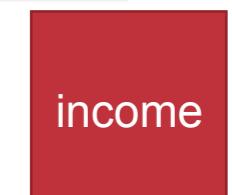
Transparency report: Mr. X

How much influence do individual features have a given classifier's decision about an individual?



Age	23
Workclass	Private
Education	11 th
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

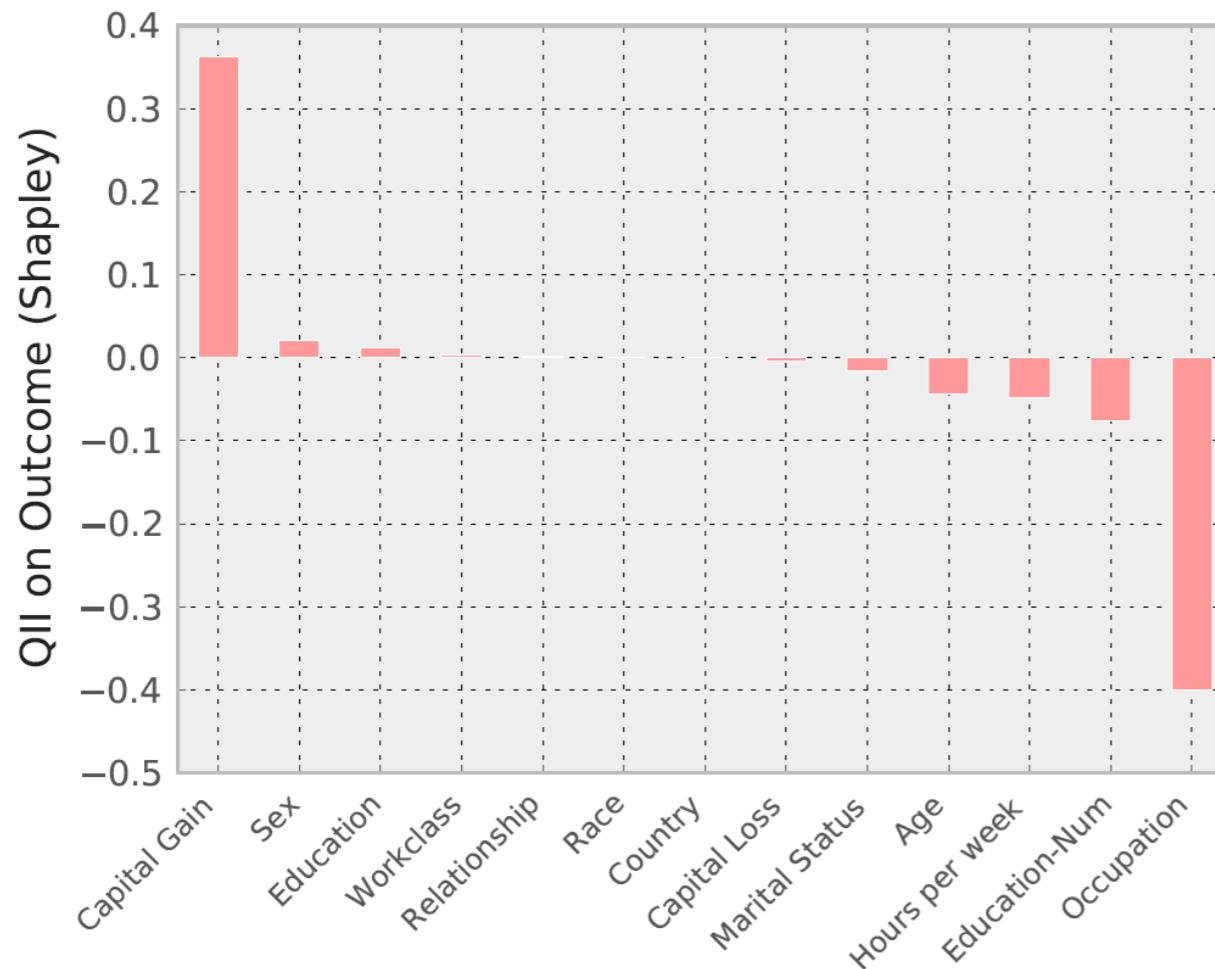
DENIED



images by Anupam Datta

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



images by Anupam Datta

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?

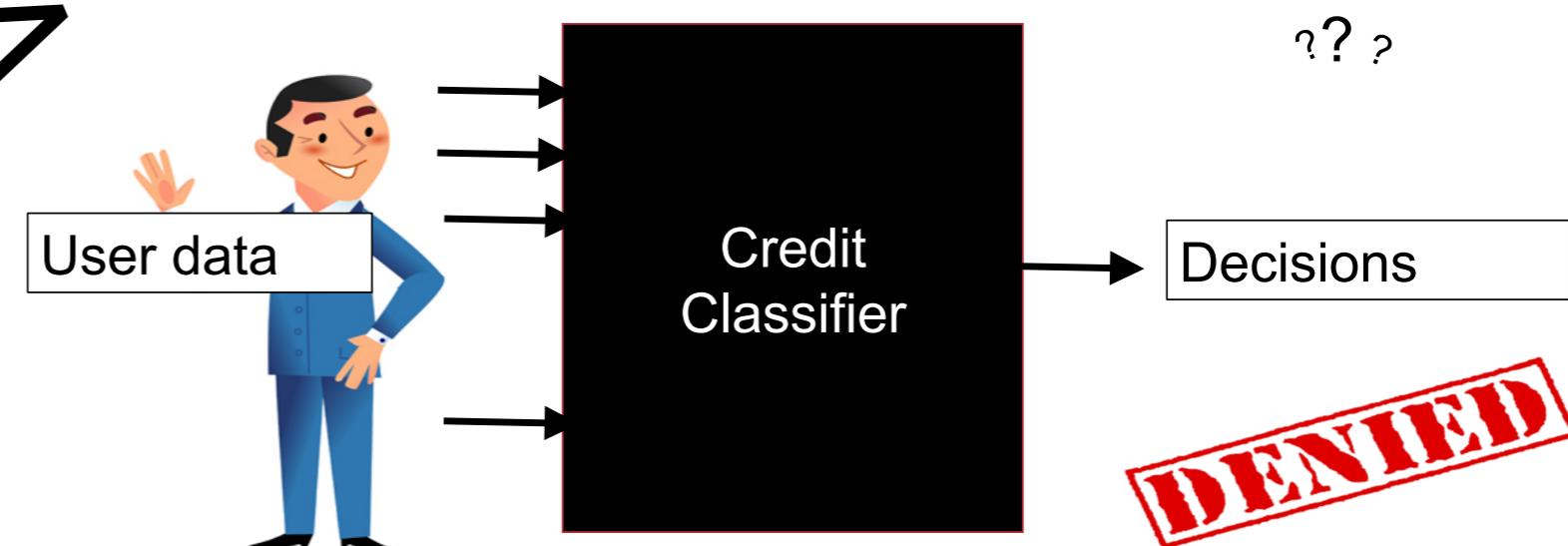
Disparity: Which inputs influence men getting more positive outcomes than women?

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

Running example

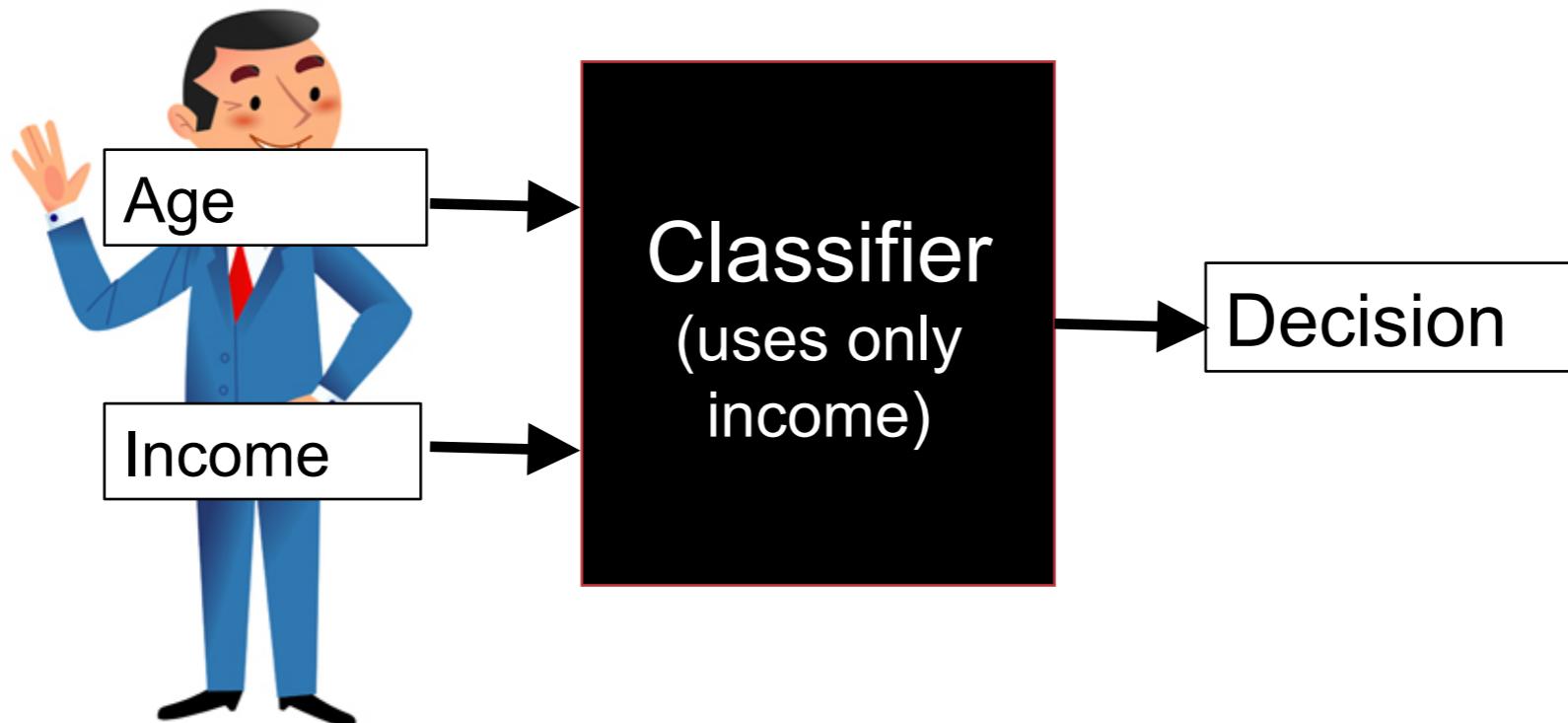
Consider lending decisions by a bank, based on gender, age, education, and income. **Does gender influence lending decisions?**

- 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**

Unary QII

images by Anupam Datta

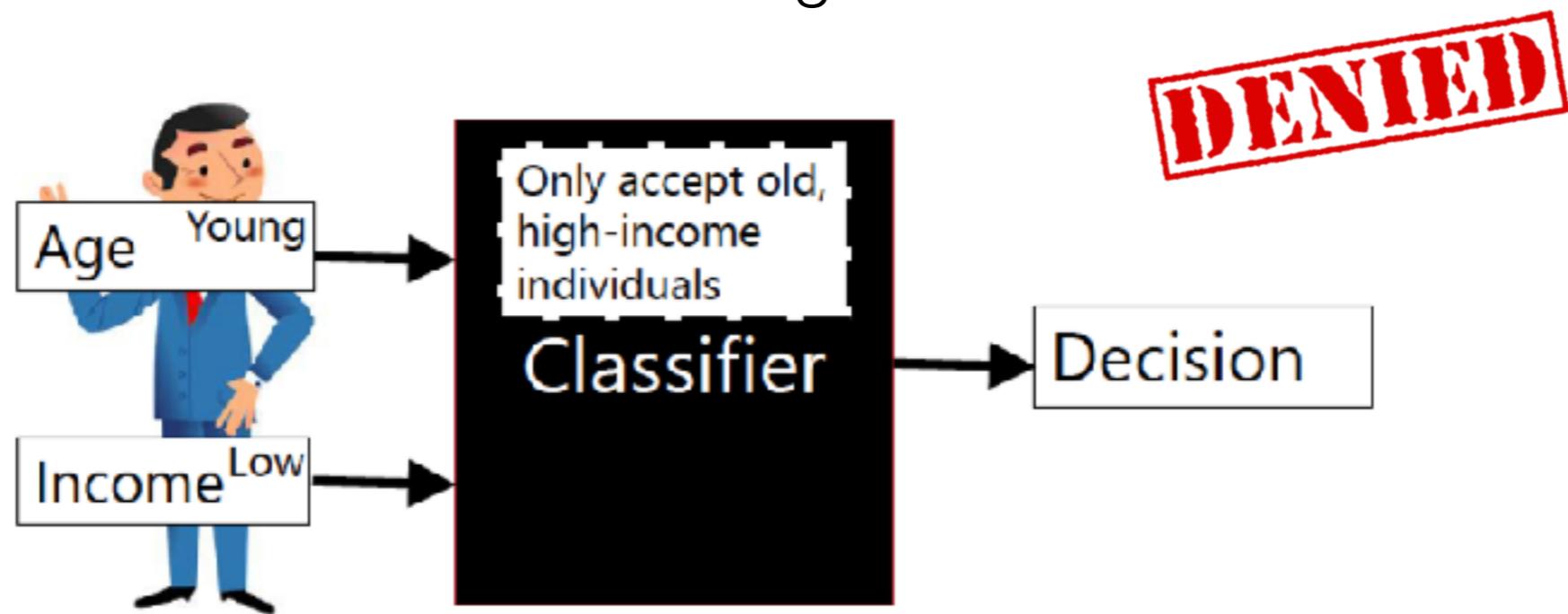
For a quantity of influence Q and an input feature i , the QII of i on Q is the difference in Q when i is changed via an **intervention**.



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!

$$\begin{array}{ll} \iota(\{\text{age}\}) - \iota(\{\}) & \iota(\{\text{age}, \text{gender}, \text{job}\}) - \iota(\{\text{gender}, \text{job}\}) \\ \iota(\{\text{age}, \text{job}\}) - \iota(\{\text{job}\}) & \iota(\{\text{age}, \text{gender}\}) - \iota(\{\text{gender}\}) \\ \iota(\{\text{age}, \text{gender}, \text{income}\}) - \iota(\{\text{gender}, \text{income}\}) & \iota(\{\text{age}, \text{gender}, \text{job}\}) - \iota(\{\text{gender}, \text{job}\}) \\ & \iota(\{\text{age}, \text{gender}, \text{income}, \text{job}\}) - \iota(\{\text{gender}, \text{income}, \text{job}\}) \end{array}$$

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

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

Aggregating influence across sets

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

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

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

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

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

$\sigma \in \Pi(N)$ 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**

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.

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

```
1:  $\phi(\mathbf{v}) = \langle 0, \dots, 0 \rangle$ 
2: for  $i \in \mathcal{A}$  do
3:   for  $\mathcal{S} \subseteq \mathcal{A} \setminus \{i\}$  do
4:      $\mathbf{U} \sim \mathcal{D} \setminus \mathbf{v}, m$ 
5:      $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A} \setminus \mathcal{S}} \mathbf{U}_{\mathcal{S}}$ 
6:      $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A} \setminus \{\mathcal{S} \cup i\}} \mathbf{U}_{\mathcal{S} \cup i}$ 
7:      $\phi_{i_{\mathcal{S}}}(\mathbf{v}) = \iota(\mathbf{U}_1, \mathbf{U}_2)$ 
8:      $\phi_i(\mathbf{v}) = \phi_i(\mathbf{v}) + \frac{1}{d} \frac{1}{\binom{d-1}{|\mathcal{S}|}} \phi_{i_{\mathcal{S}}}(\mathbf{v})$ 
9:   end for
10: end for
11: return  $\phi(\mathbf{v})$ 
```

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

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 members 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 CSGradRankings.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]

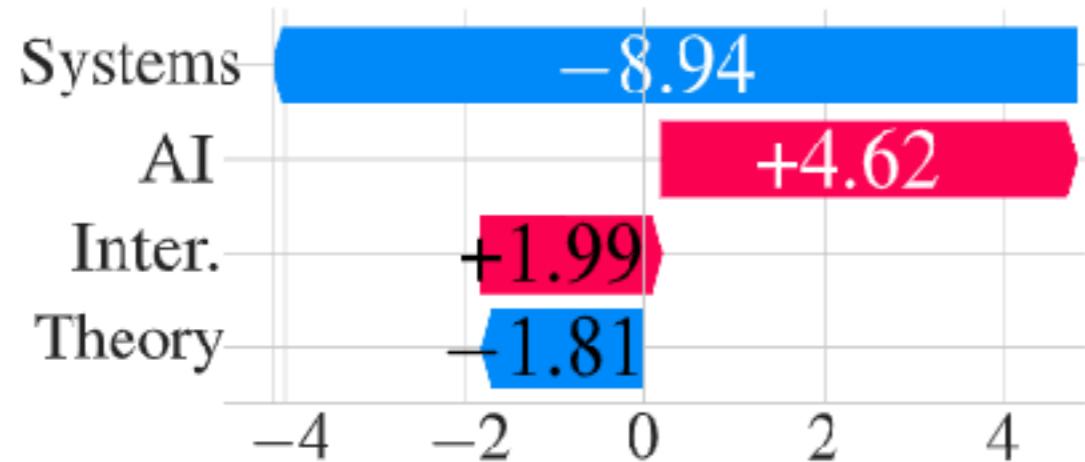
- Algorithms & complexity
- Cryptography
- Logic & verification

Interdisciplinary Areas [off | on]

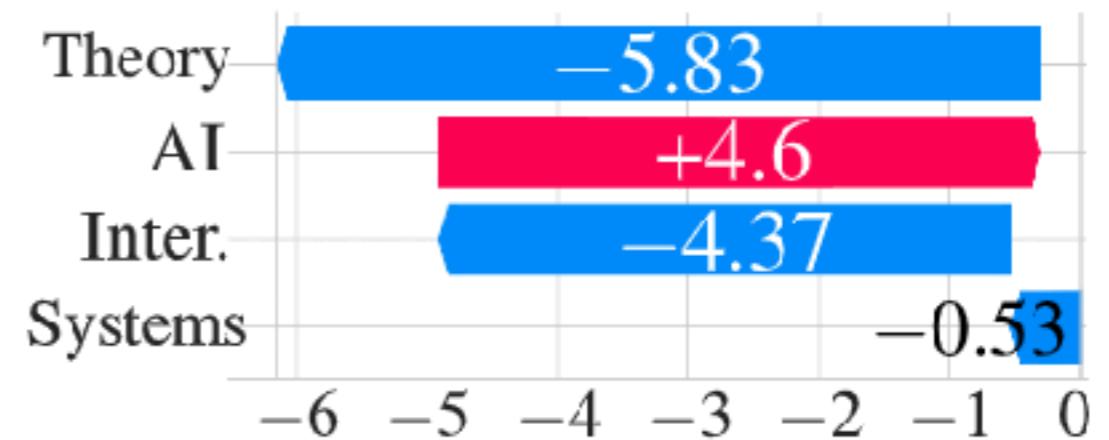
- Comp. bio & bioinformatics
- Computer graphics
- Computer science education
- Economics & computation
- Human-computer interaction
- Robotics
- Visualization

#	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
8	University of Washington	10.1	81
9	Stanford University	9.6	68
10	Cornell University	9.3	83
11	University of Maryland - College Park	8.8	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



(a) South Carolina, ranked 101



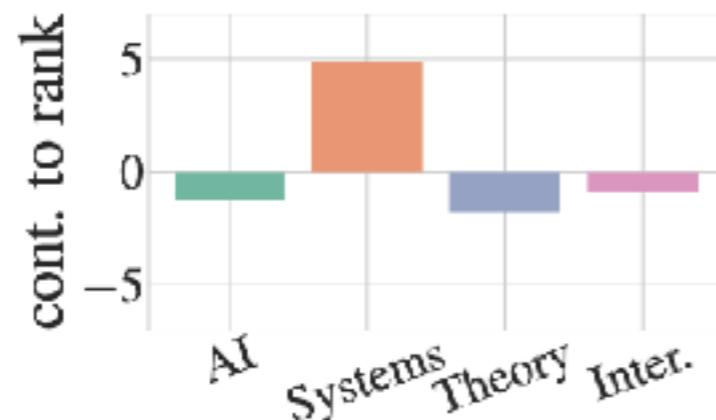
(b) Wayne State, ranked 102

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

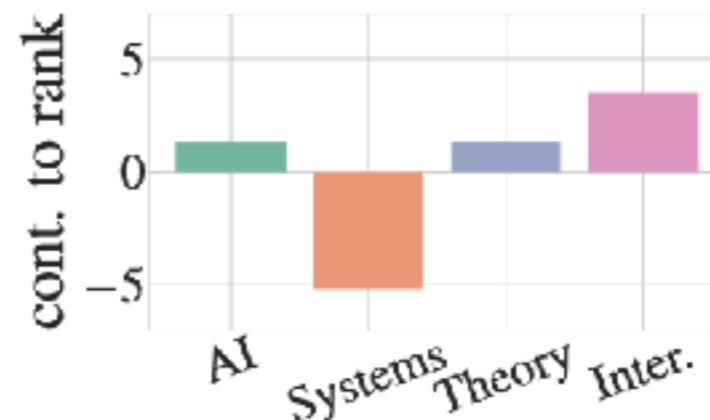
Comparing Georgia Tech, Stanford & UMich

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.



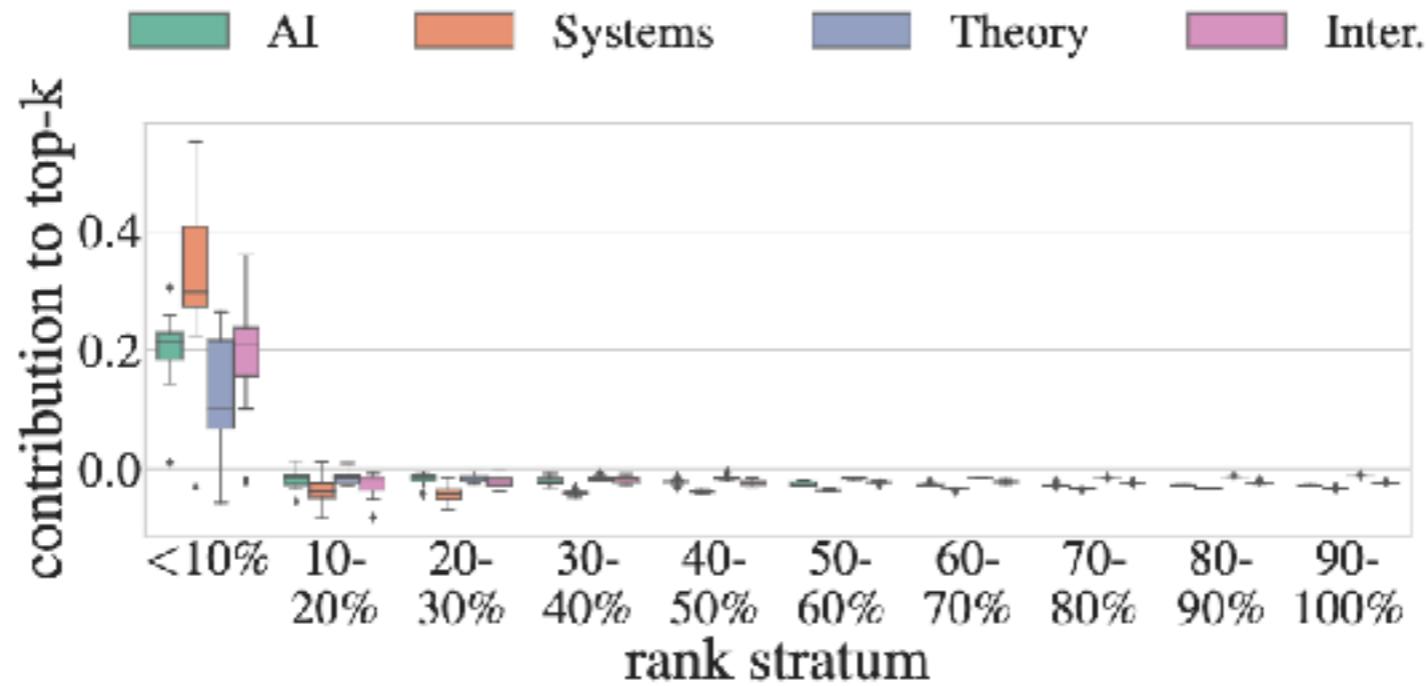
(c) Pairwise QoI explaining that Georgia Tech ranks higher than Stanford because of its relative strength in Systems.



(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.

Aggregates feature importance by rank stratum

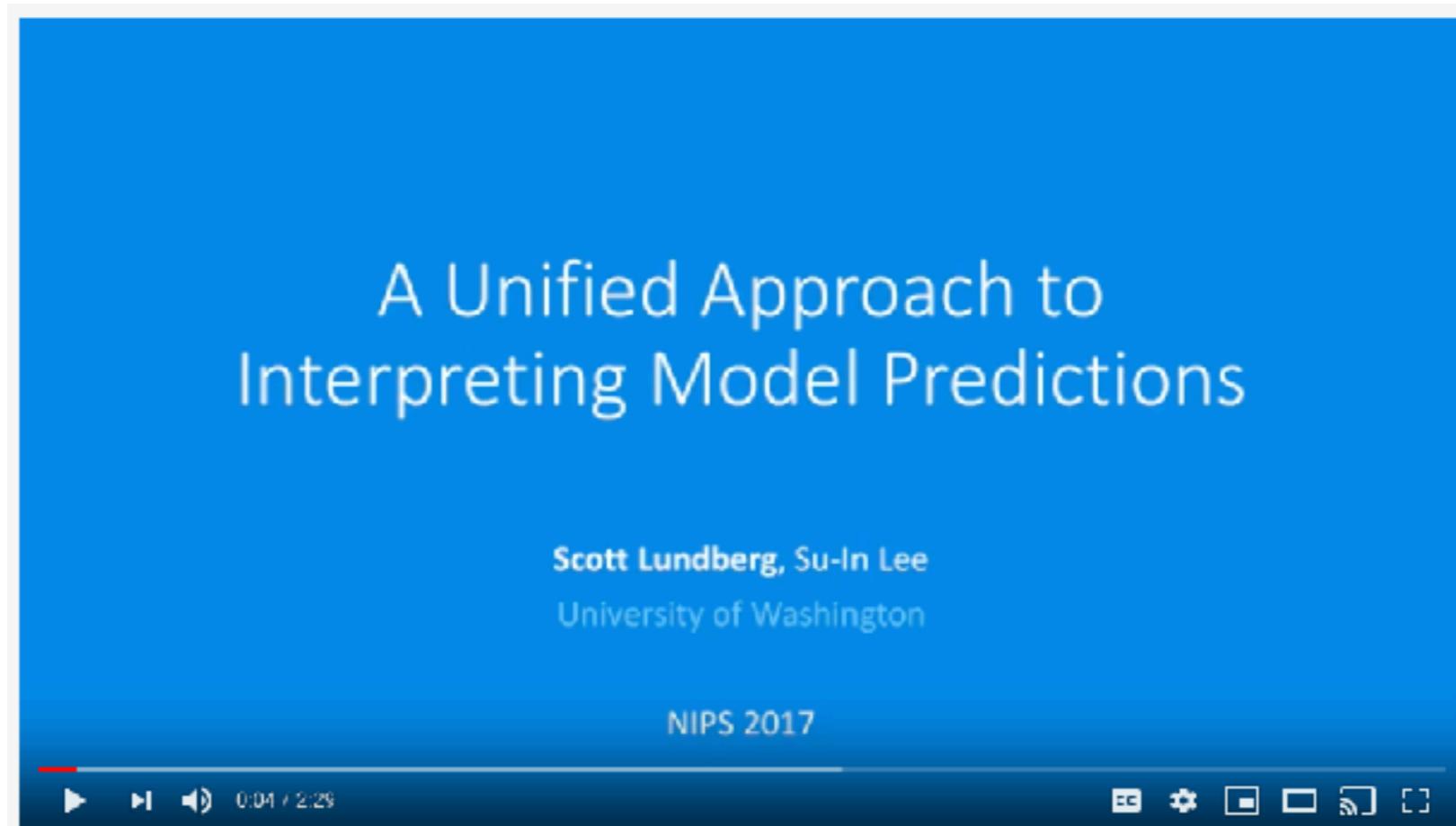


(a) Feature contribution to the top- k QoI, for $k = 10\%$. Systems is the most important feature, followed by Interdisciplinary and AI, while Theory is least important.

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

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

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

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

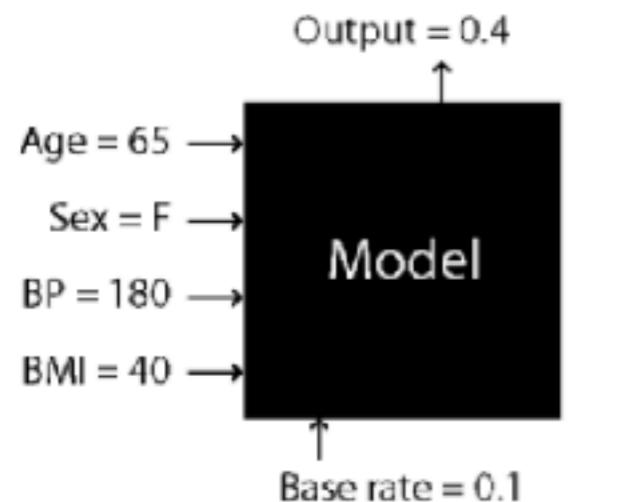
1. **Local accuracy**: $g(x')$ matches the original model $f(x)$ when x' is the **simplified input** corresponding to x .
2. **Missingness**: if x'_i — the i^{th} feature of simplified input x' — is missing, then it has no attributable impact for x
3. **Consistency (monotonicity)**: if toggling off feature 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

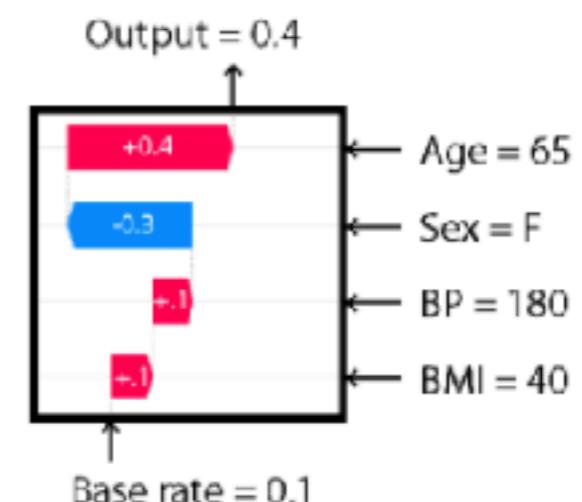
README.md



SHAP



Explanation →



<https://github.com/slundberg/shap>

LIME: Local Interpretable Model-Agnostic Explanations

Why should I trust you?

Explaining the predictions of any classifier



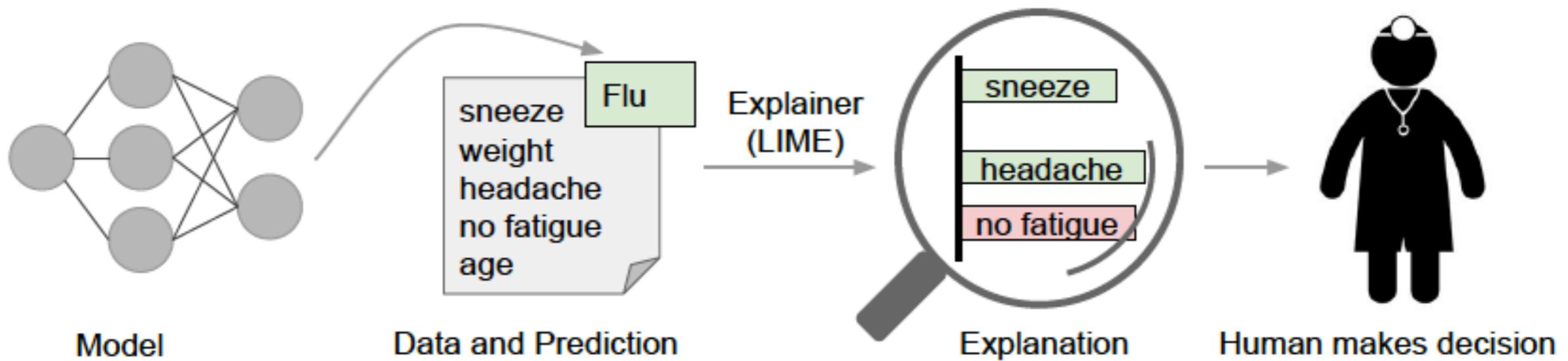
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Check out our paper, and open source project at
<https://github.com/marcotcr/lime>

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

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

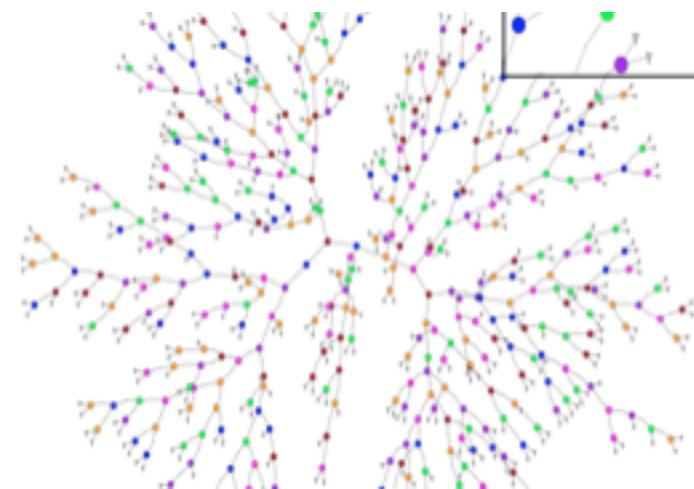
what if patient id appears in green in the list? - an example of “data leakage”

LIME: Local explanations of classifiers

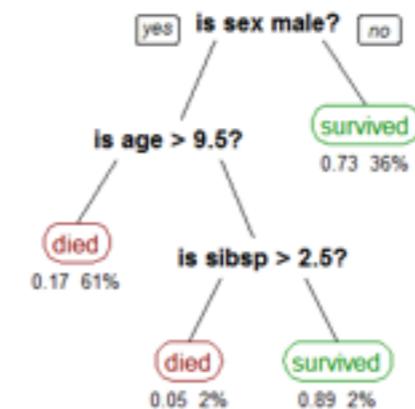
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely
not interpretable



Potentially
interpretable

Explanations based on features

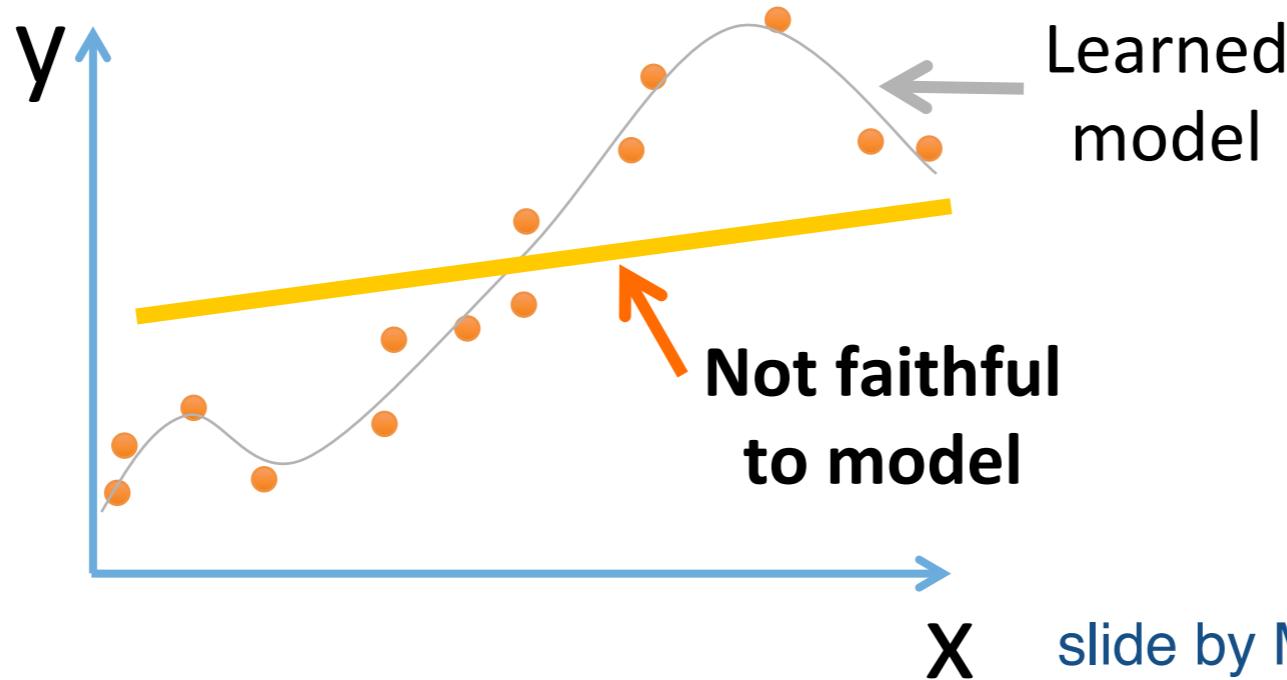
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

Faithful

- Describes how this model actually behaves



X slide by Marco Tulio Ribeiro, KDD 2016

Explanations based on features

Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

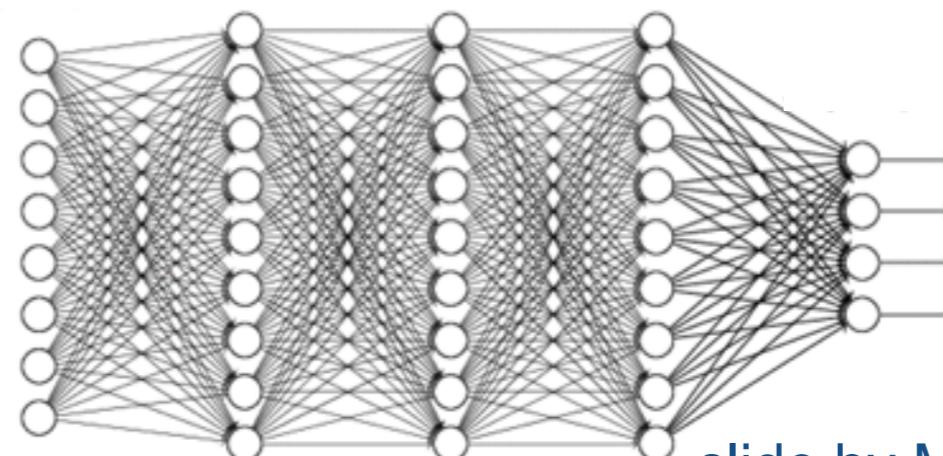
Faithful

- 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

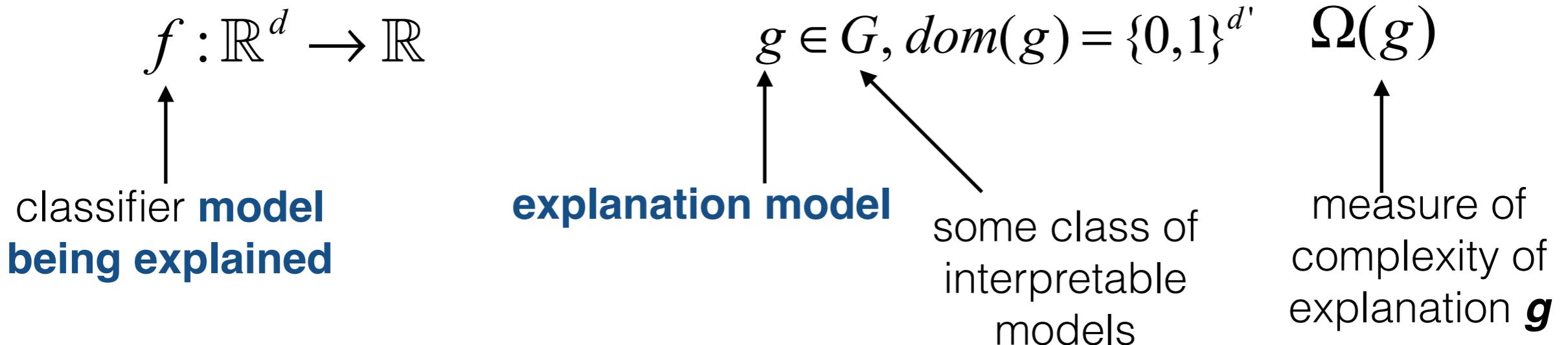
Key idea: Interpretable representation

“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

- 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 or 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 or absence of each of d' interpretable components

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 classifier.”



$f(x)$ denotes the probability that \mathbf{x} belongs to some class

π_x is a **proximity measure** relative to \mathbf{x}

we make no assumptions about \mathbf{f} to remain model-agnostic: draw samples weighted by π_x

explanation

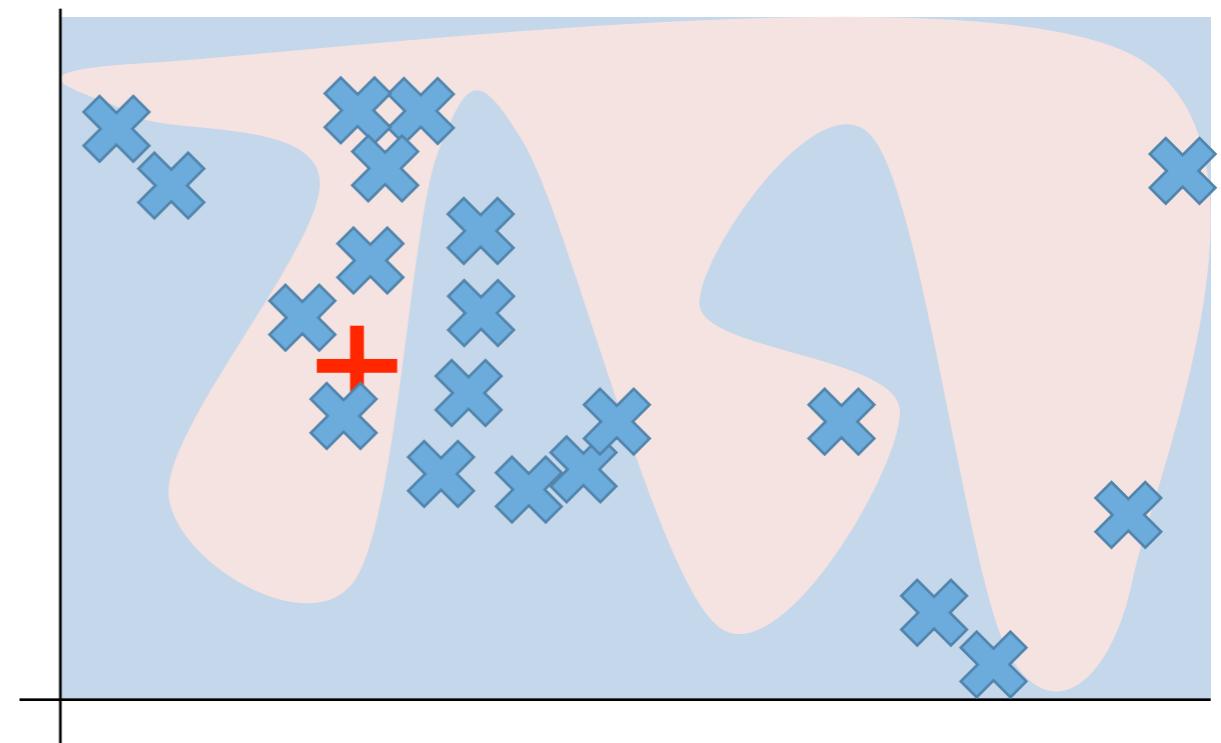
$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

measures how unfaithful is g to \mathbf{f} in the locality around \mathbf{x}

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 classifier.”

1. sample points around 

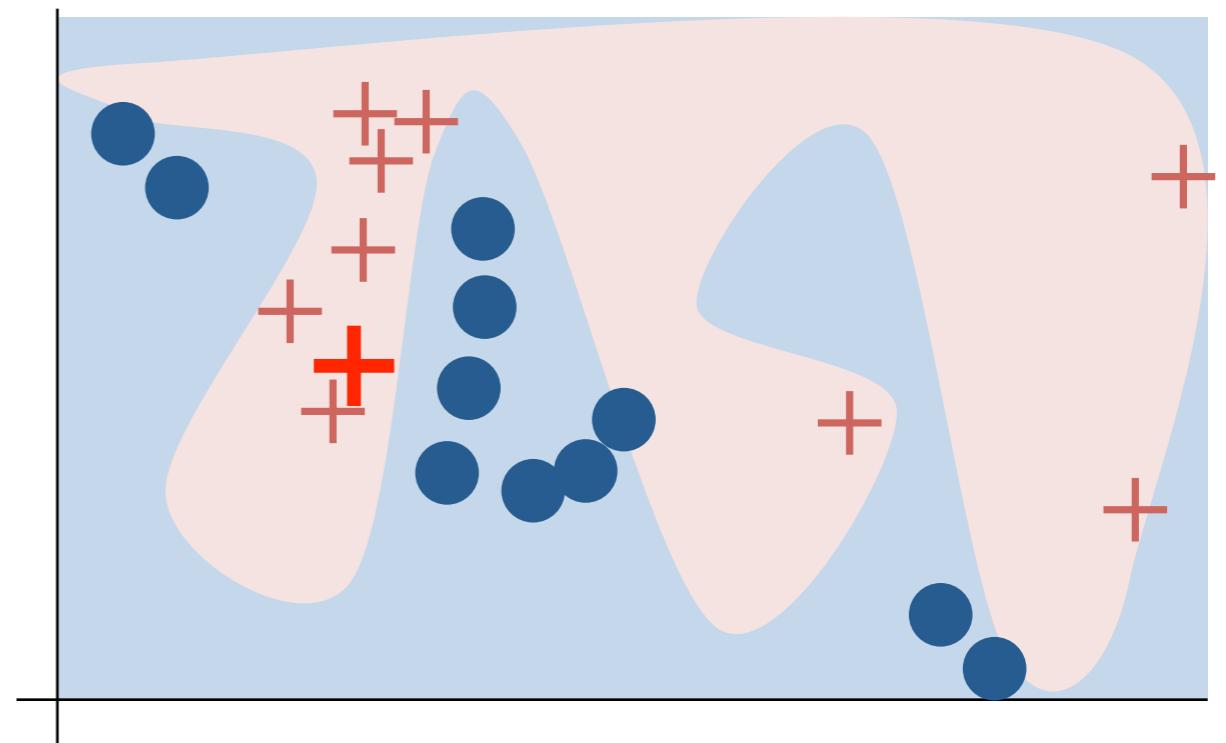


based on a slide by Marco Tulio Ribeiro, KDD 2016

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 classifier.”

1. sample points around 
2. use complex model f to assign class labels

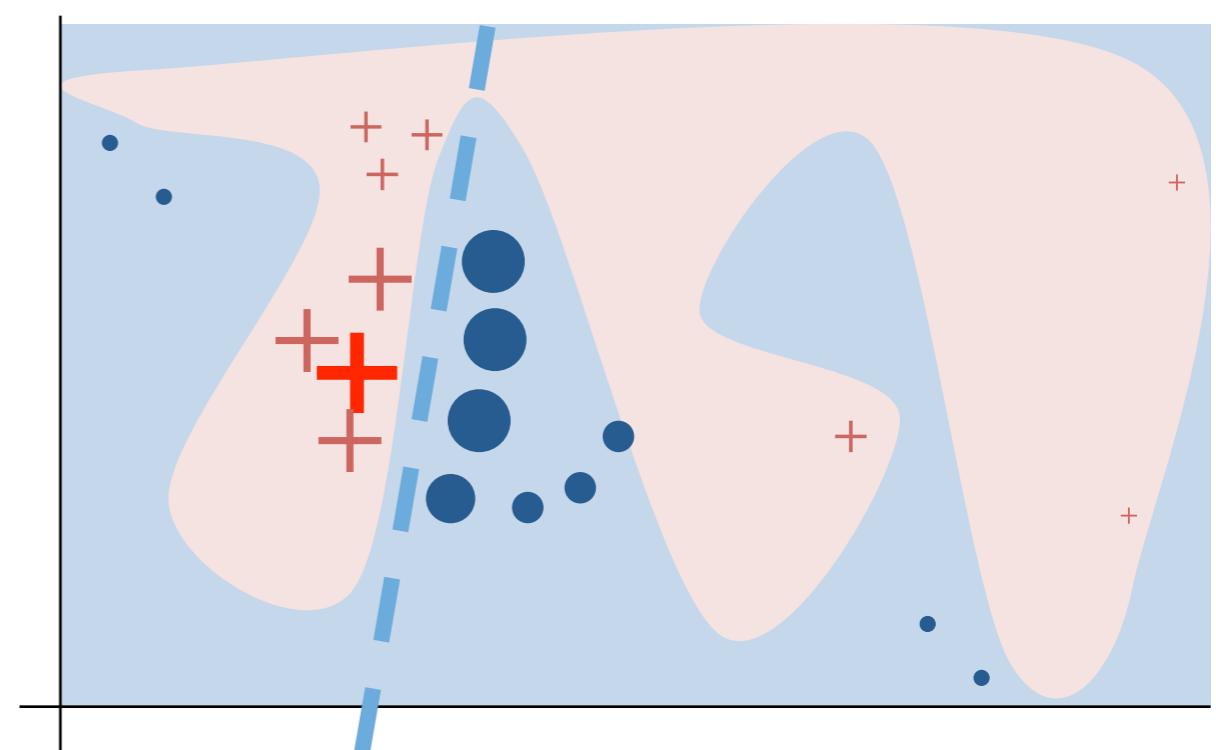


based on a slide by Marco Tulio Ribeiro, KDD 2016

Fidelity-interpretability trade-off

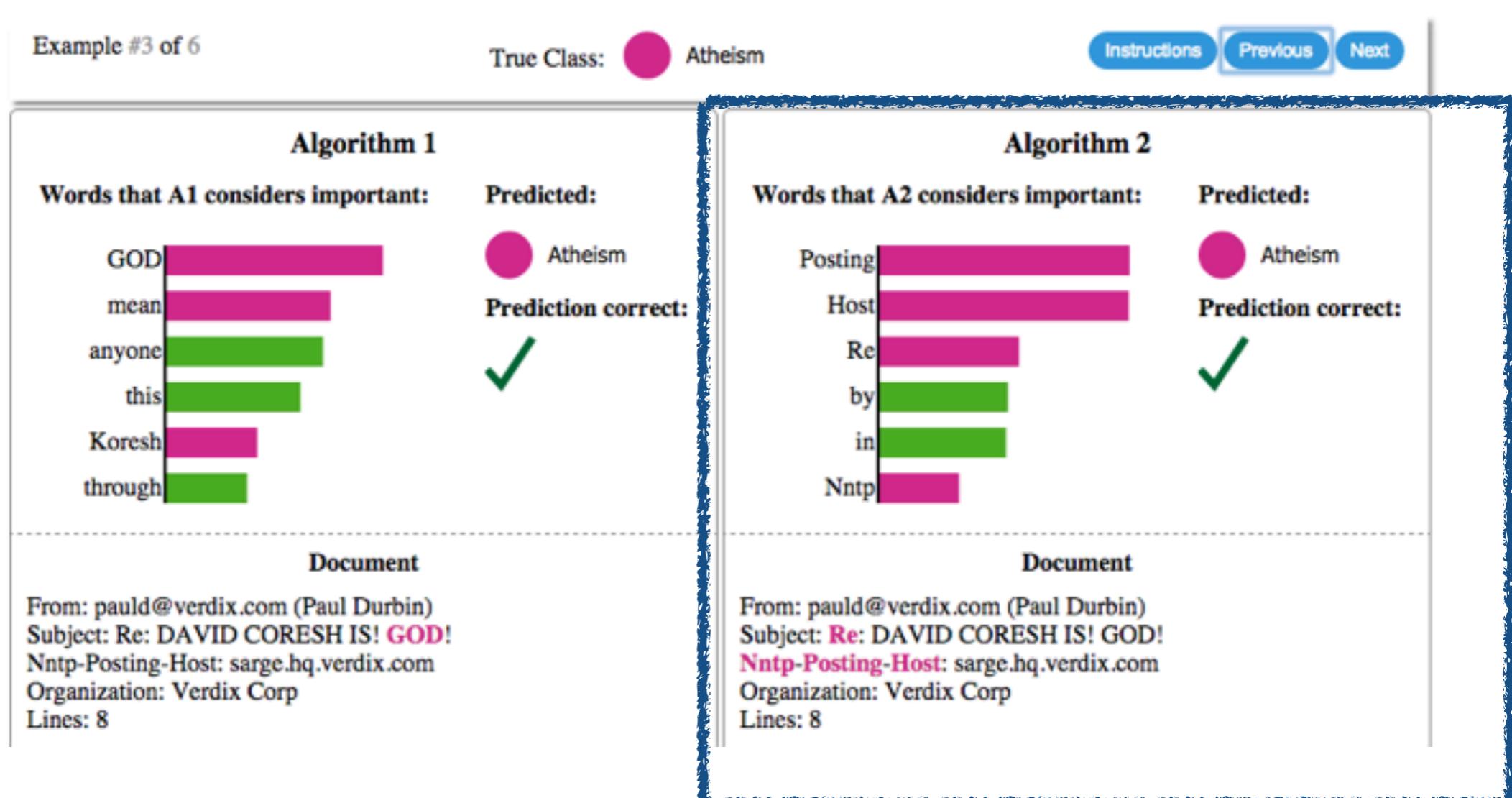
“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

1. sample points around $\textcolor{red}{+}$
2. use complex model \mathbf{f} to assign class labels
3. weigh samples according to $\boldsymbol{\pi}_x$
4. learn simple model \mathbf{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!

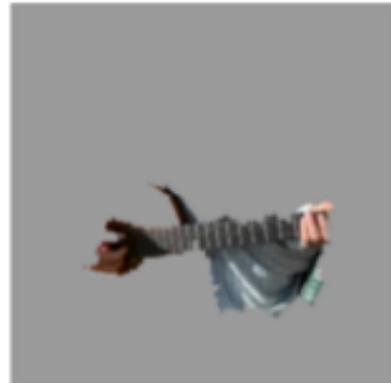
When accuracy is not enough

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



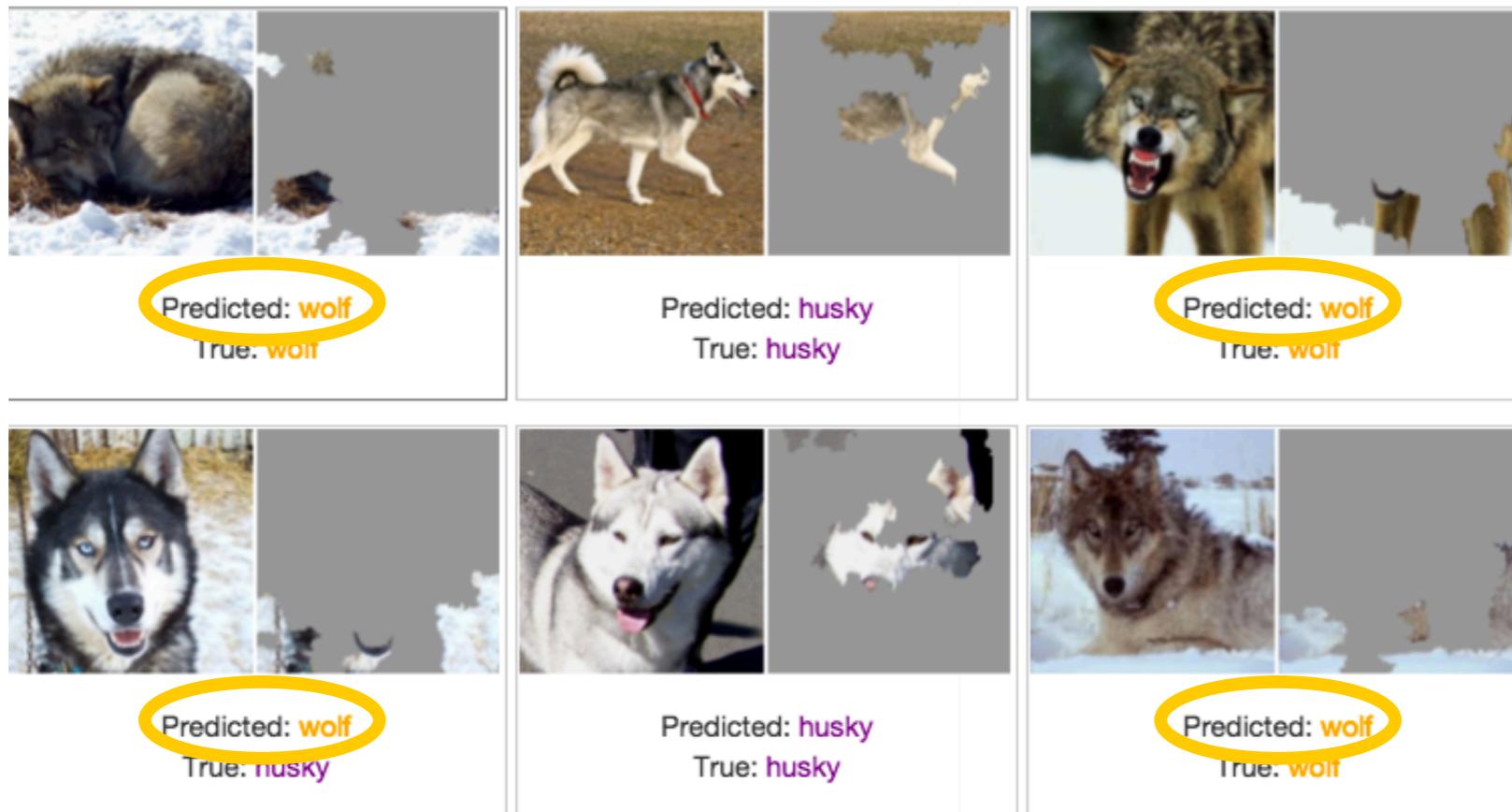
Only 1 mistake!!!

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

slide by Marco Tulio Ribeiro, KDD 2016

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

Why should I trust you?

Explaining the predictions of any classifier



Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Check out our paper, and open source project at
<https://github.com/marcotcr/lime>

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