

Responsible Data Science

Transparency & Interpretability

Auditing black-box models

April 17, 2023

Prof. Julia Stoyanovich

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

Terminology & vision

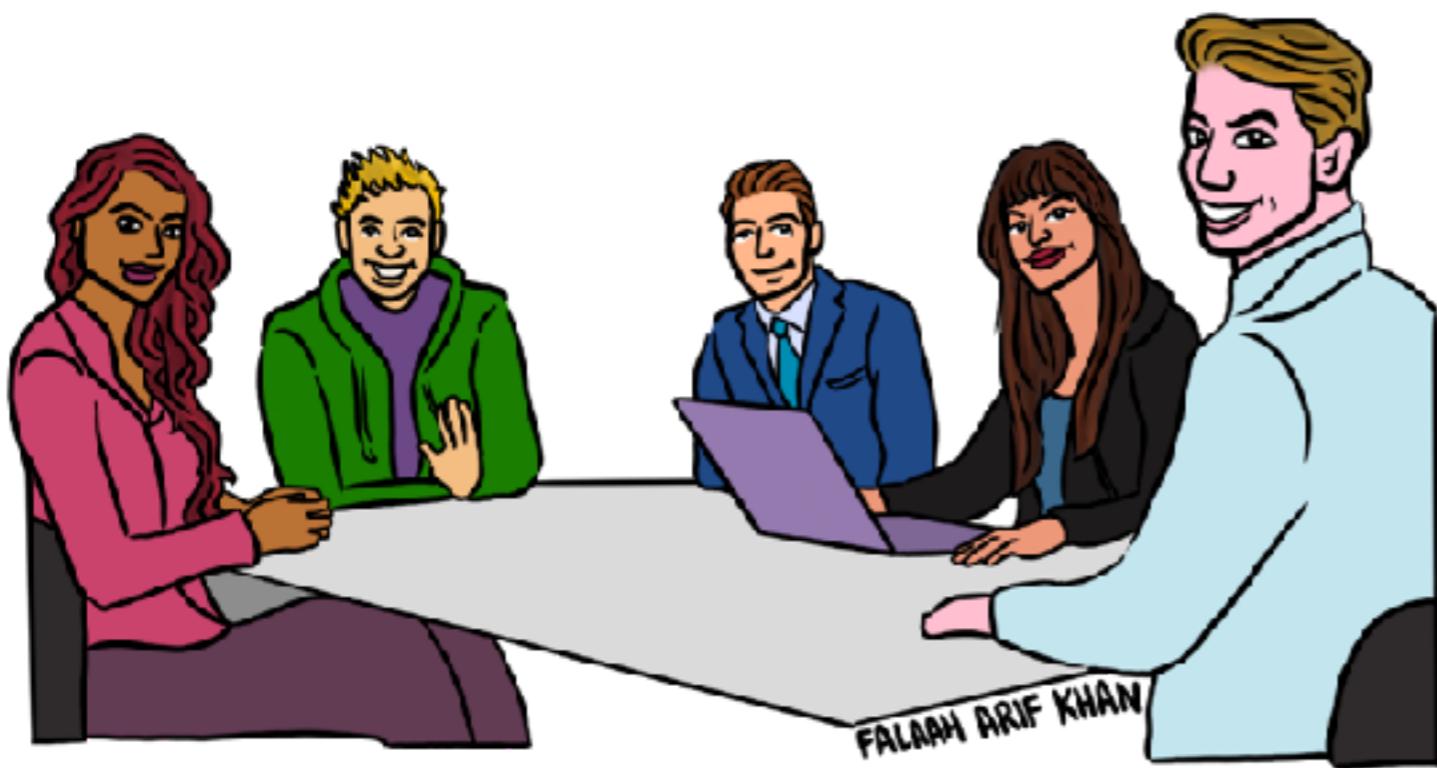


transparency, interpretability,
explainability, intelligibility



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 1,852 times to the male group and only 318 times to the female group. **What** are we explaining? To **Whom** are we explaining? **and only 318 times**

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



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

SIDE-BY-SIDE COMPARISON

Original Label

Nutrition Facts

Serving Size 2/3 cup (55g)	Calories from Fat: 72
servings per container About 8	
Amount Per Serving	% Daily Value*
Calories 230	10%
Total Fat 8g	12%
Saturated Fat 1g	5%
Trans Fat 0g	
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

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

New Label

Nutrition Facts

6 servings per container	Serving size 2/3 cup (55g)
Amount per serving	% Daily Value*
Calories 230	10%
Total Fat 8g	16%
Saturated Fat 1g	5%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 150mg	7%
Total Carbohydrate 35g	12%
Dietary Fiber 4g	14%
Added Sugars 10g	20%
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.	1
Data Practices	Canner data collection: Visual, Audio, Physiological, Location	Similarity type: Device ID, Device functions, Research, No device message, Identified - Opt-in, Device identifier, Manufacturer, Not disclosed	2
Similarity type: Device ID, Device functions, Research, No device message, Identified - Opt-in, Device identifier, Manufacturer, Not disclosed	3	Other collected data: Motor, Account info, Payment info, Contact info, Device settings, Device info, Device usage info	4
Privacy policy: www.NSG200.smartdeviceco.com/policy			
More Information	Detailed Security & Privacy Label: www.iotsecurityandprivacy.org/labels		

CMU IoT Security and Privacy Label - OSPL 1.0 www.iotsecurityandprivacy.org

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

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/hiring-job-candidates-ai-11632244313>

This week's reading

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
{dattan, shayks, yairz}@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 ascertain 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 input features 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 rules used for historical and external oversight (e.g., to detect algorithmic discrimination).

Blockquote **our recent 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, these simple inputs may not always have high influence, the QII measures also quantify the joint influence of a set of inputs (e.g., age and location or outcomes log loss decisions) 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 aggregation measures, such as the Shapley value, previously applied to measure influence in voting. Further, since transparency reports must compute outputs, we capture 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 tool, 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 for purposes of transparency, such as an algorithmic system above 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 or the 1st mobility 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 derive a transparency report that provides answers to aggregate transparency queries, such as the influence of sensitive inputs (e.g., gender), used on the system's decisions concerning the entire population or about systematic differences in decisions

"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
ssingh@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. But, understanding also provides insights into the model, which can be used to transfer its understanding to other predictions from 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 user-friendly way, leaving the task as a sensible application problem. We demonstrate the feasibility of these methods by exploring a range of models for text (e.g., natural language) and image classification (e.g., medical informatics). 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, bypassing an untrustworthy classifier, and identifying why a classifier should not be trusted.

In this paper, we present a novel framework for interpreting predictions. SHAP (SHapley Additive exPlanations) 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, while some several recent methods in the class lack the proposed definitive properties. Based on insights from this articulation, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans in an automation aspect in both of either directly developing machine learning classifiers as tools, or employing models within other products, is often concern because 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

Footnote 1: Marco Tulio Ribeiro is grateful to all co-authors of this work for permission to reuse some of their work without the need to cite them or ask for specific permission. In particular, this work builds upon the work of [1] and [2]. Copyright for components of this work owned by others than the author(s) must be honored. Authoring with non-honorables is granted to copy, distribute, or republish material in certain circumstances. Please refer to the publisher's Terms & Conditions for details. Requests for further permission or rights should be directed to the publisher.

ICML 2016 San Francisco, CA, USA
© 2016, PMLR. All Rights Reserved. Published under license from the ACM.
978-1-4503-4296-3/16/06...\$15.00
https://doi.org/10.1145/2939672.2949978

how much the human understands a model's behavior, 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 terrorist detection, for example, predictions cannot be total ignore or total faith, as the consequences can be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it. "Is the wif?" To make this decision, users need to be confident that the model will perform well on real-world data according to the metric 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 problem's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to model metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we present a novel framework for interpreting predictions. SHAP (SHapley Additive exPlanations) 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, while some several recent methods in the class lack the proposed definitive properties. Based on insights from this articulation, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

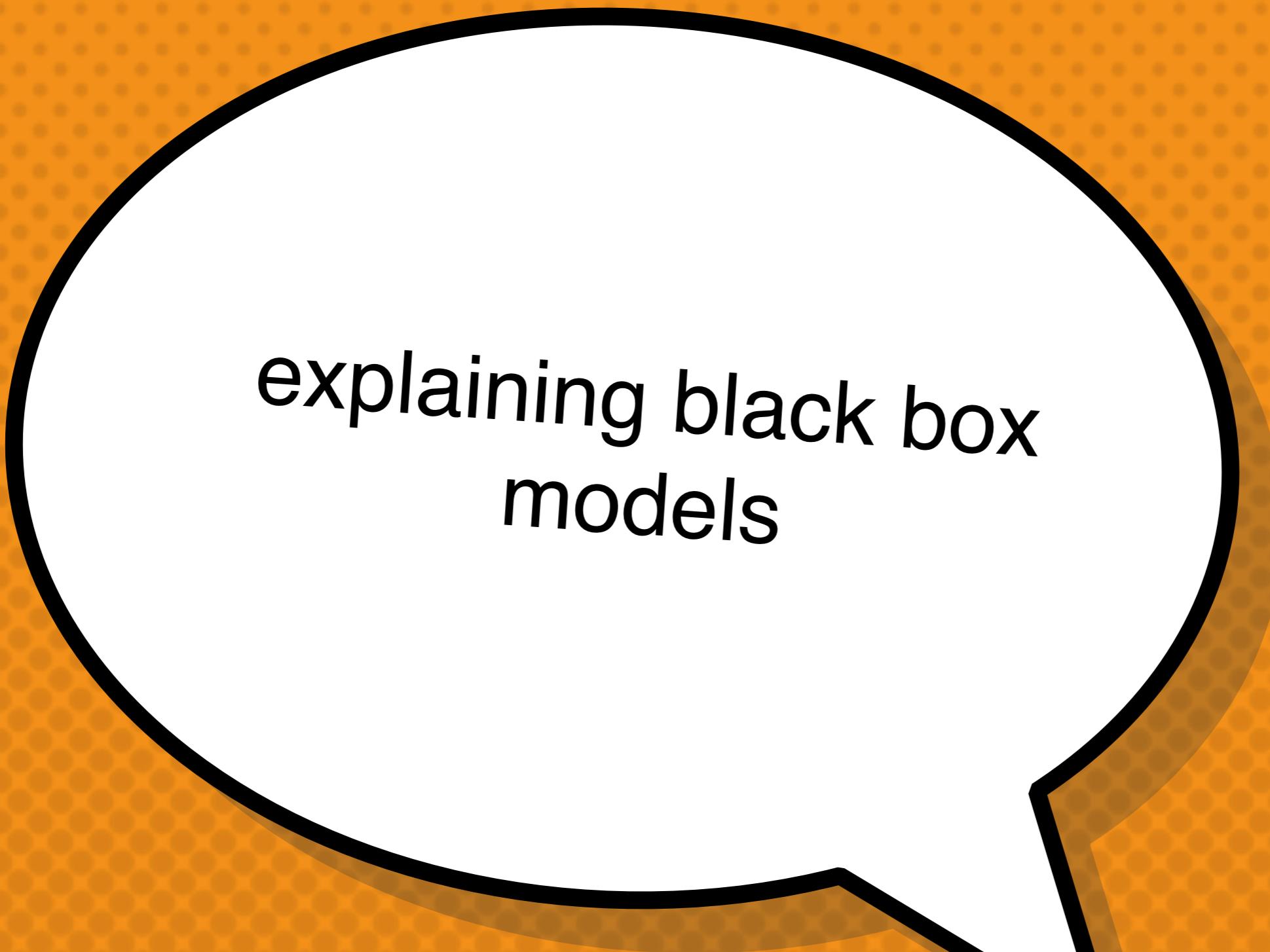
1. Introduction

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, as 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, 11].

Here, we present a novel unified approach to interpreting model predictions.² Our approach leads to three potentially surprising results: that bring clarity to the growing space of methods:

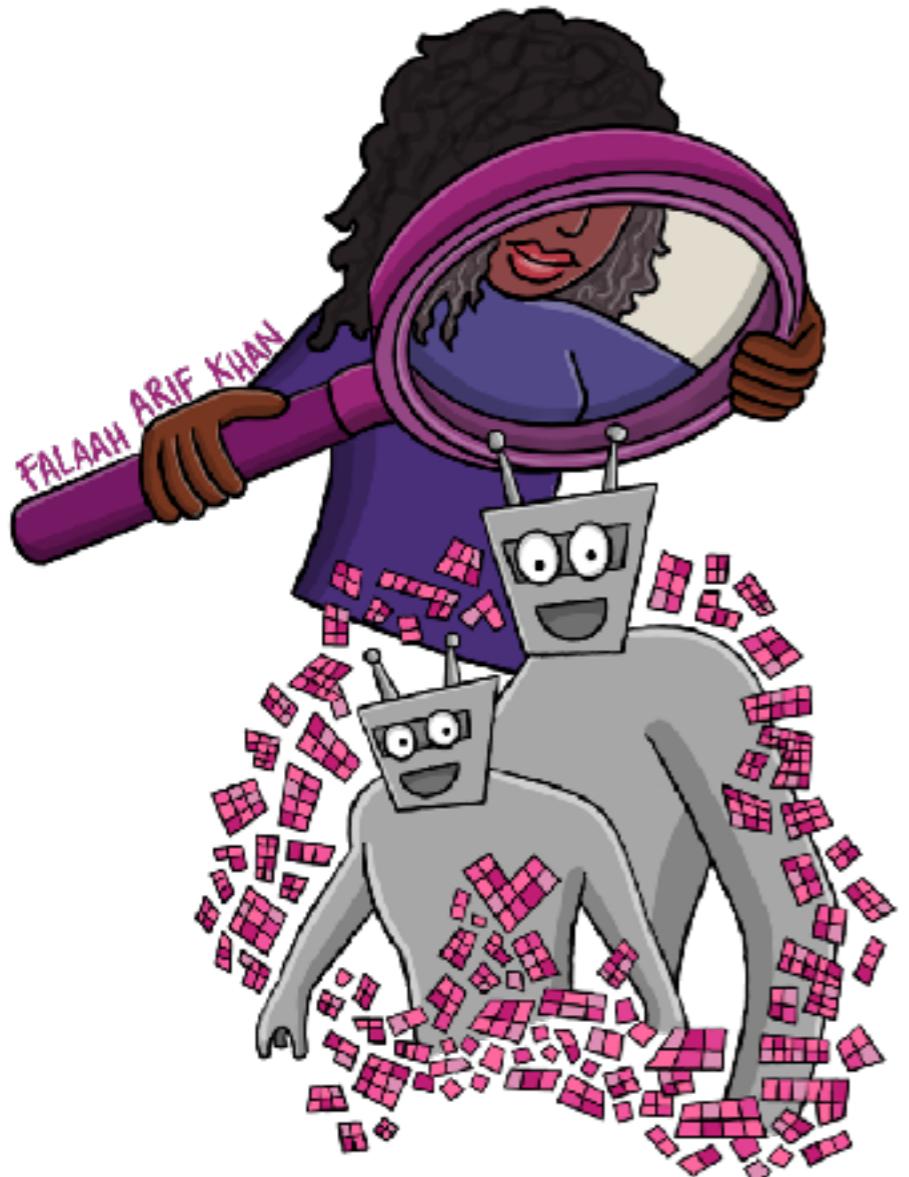
1. We introduce the perspective of viewing any explanation of a model's prediction as a model itself, which we term the explainer model. This lets us define the class of additive feature attribution methods (Section 2), which unifies six current methods.

²<https://github.com/slundberg/shap>



explaining black box
models

What are we explaining?



How does a system work?

How **well** does a system work?

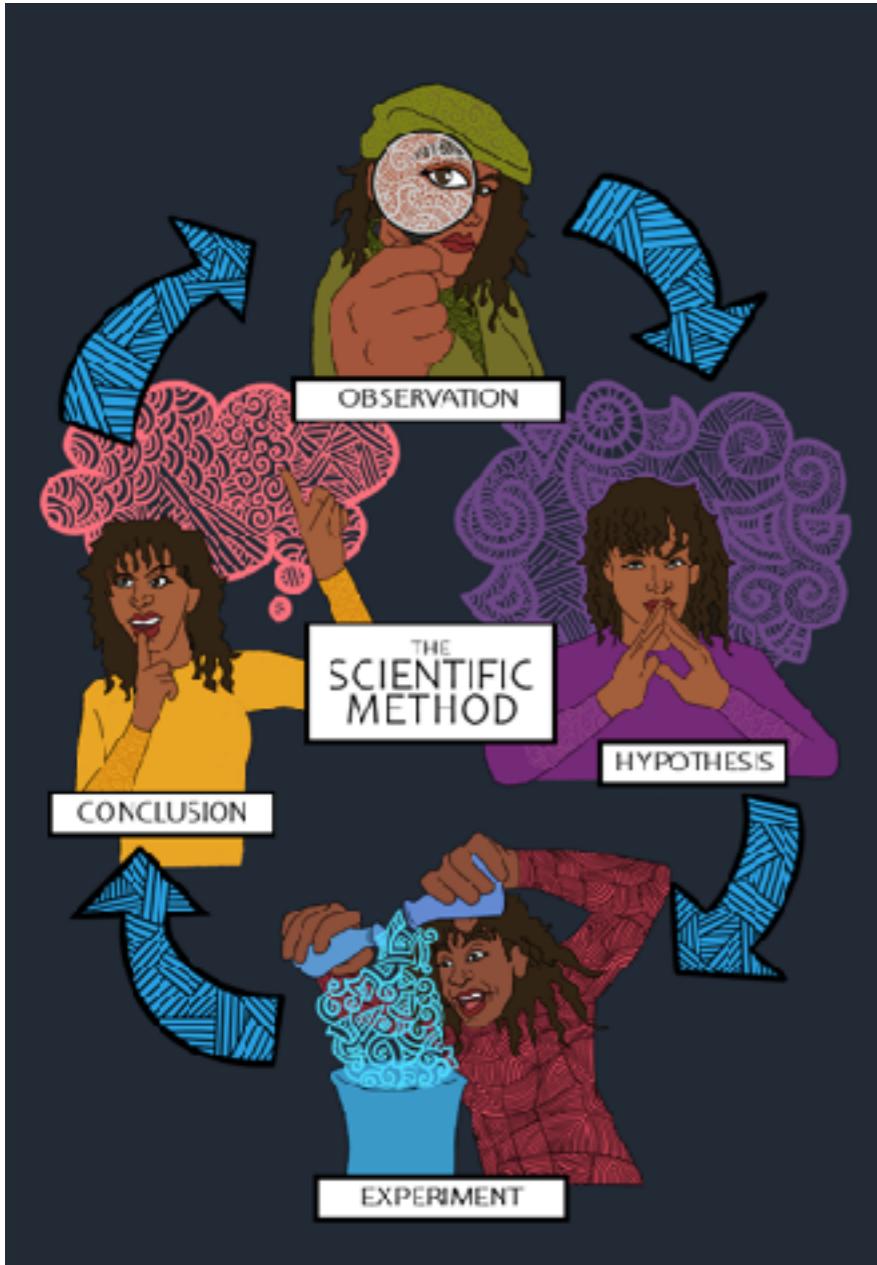
What does a system do?

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.

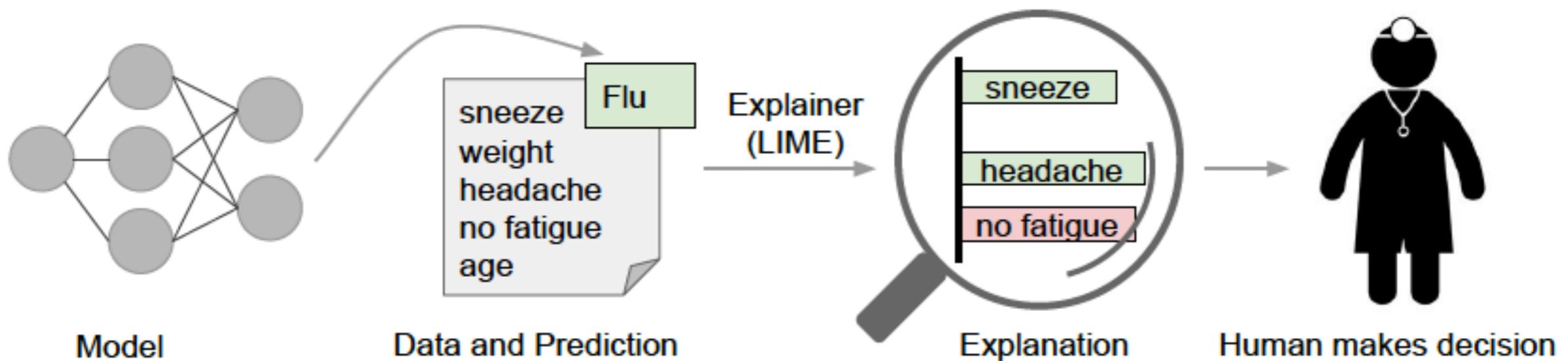
The social network has a history of telling its users that the names they're attempting to use aren't real. Drag queens and overseas human rights activists, for example, have experienced error messages and problems logging in in the past.

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

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

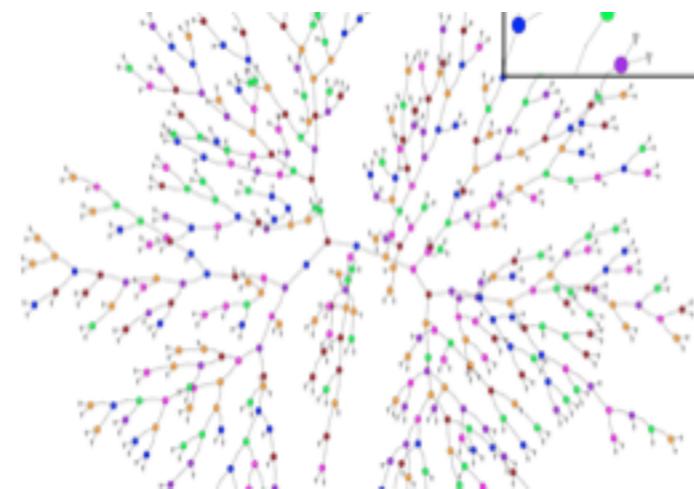
what if patient id appears in green in the list? - an example of “data leakage”

LIME: Local explanations of classifiers

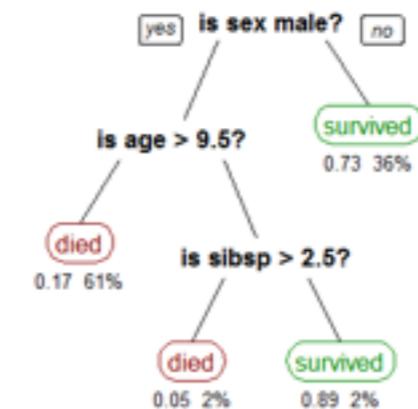
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely
not interpretable



Potentially
interpretable

slide by Marco Tulio Ribeiro, KDD 2016

Explanations based on features

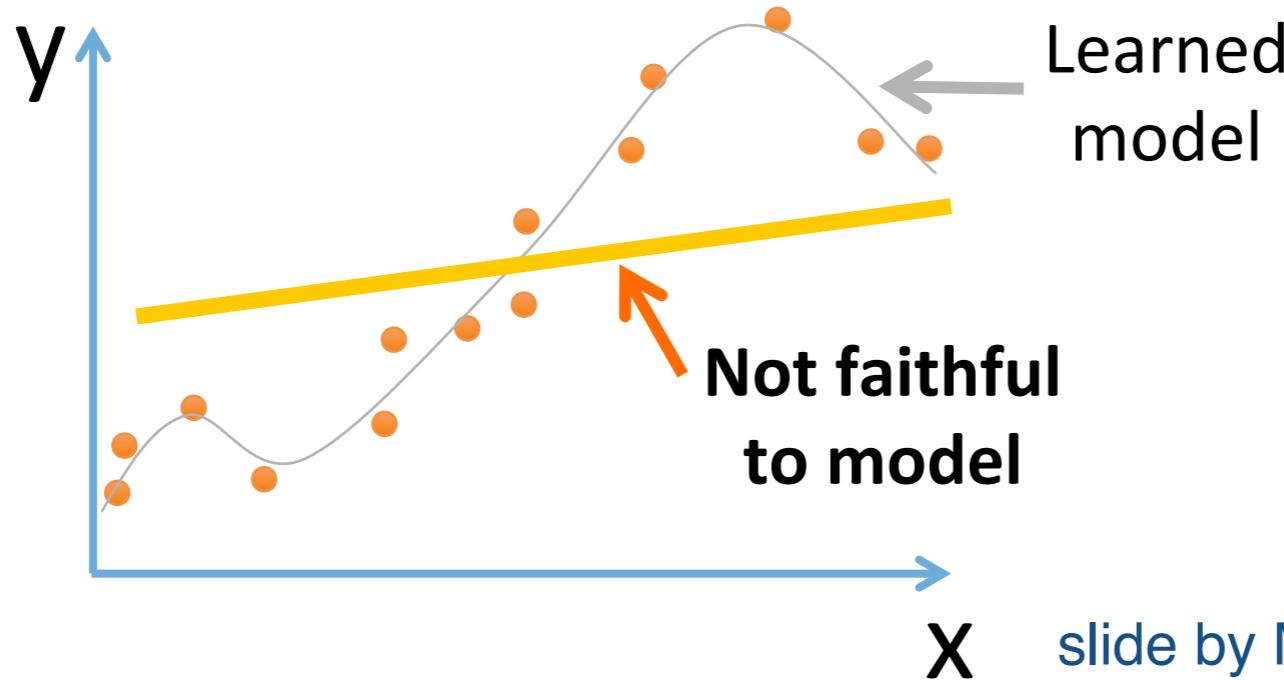
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

Faithful

- Describes how this model actually behaves



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

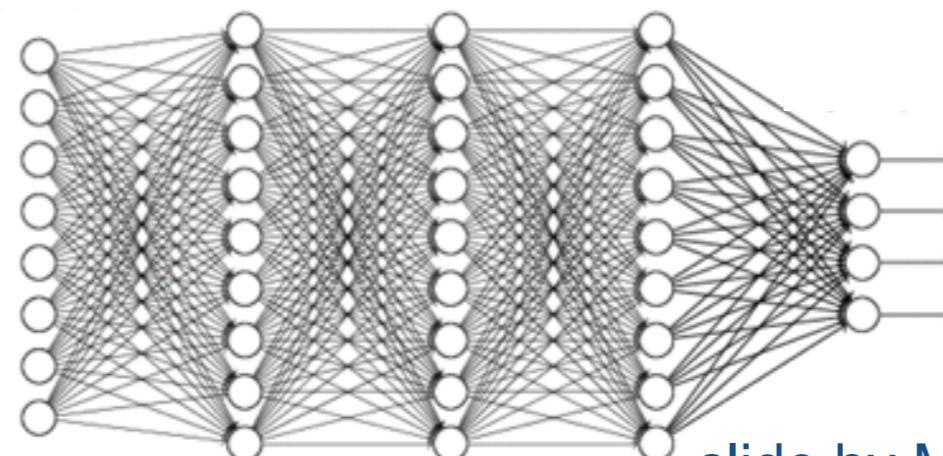
Faithful

- 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

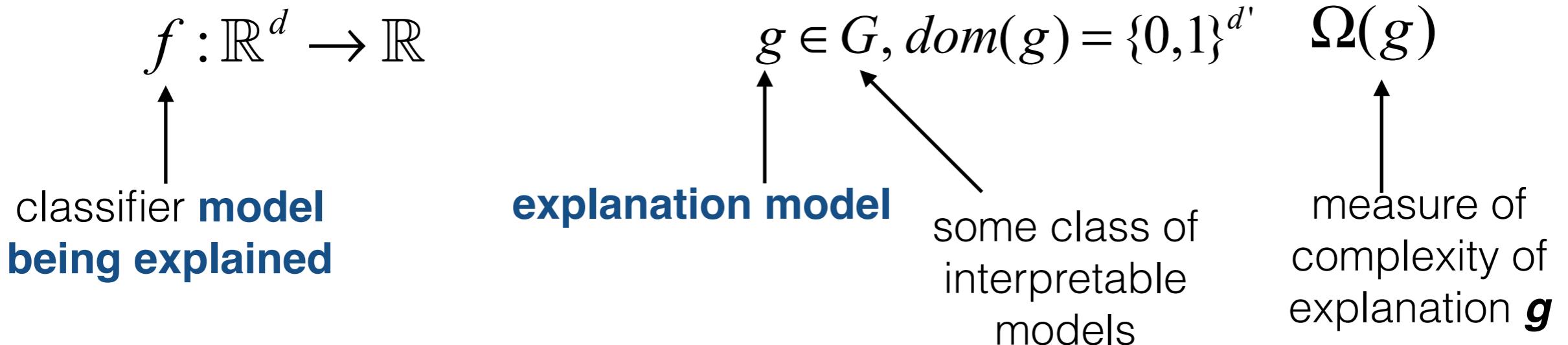
Key idea: Interpretable representation

“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 are 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 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

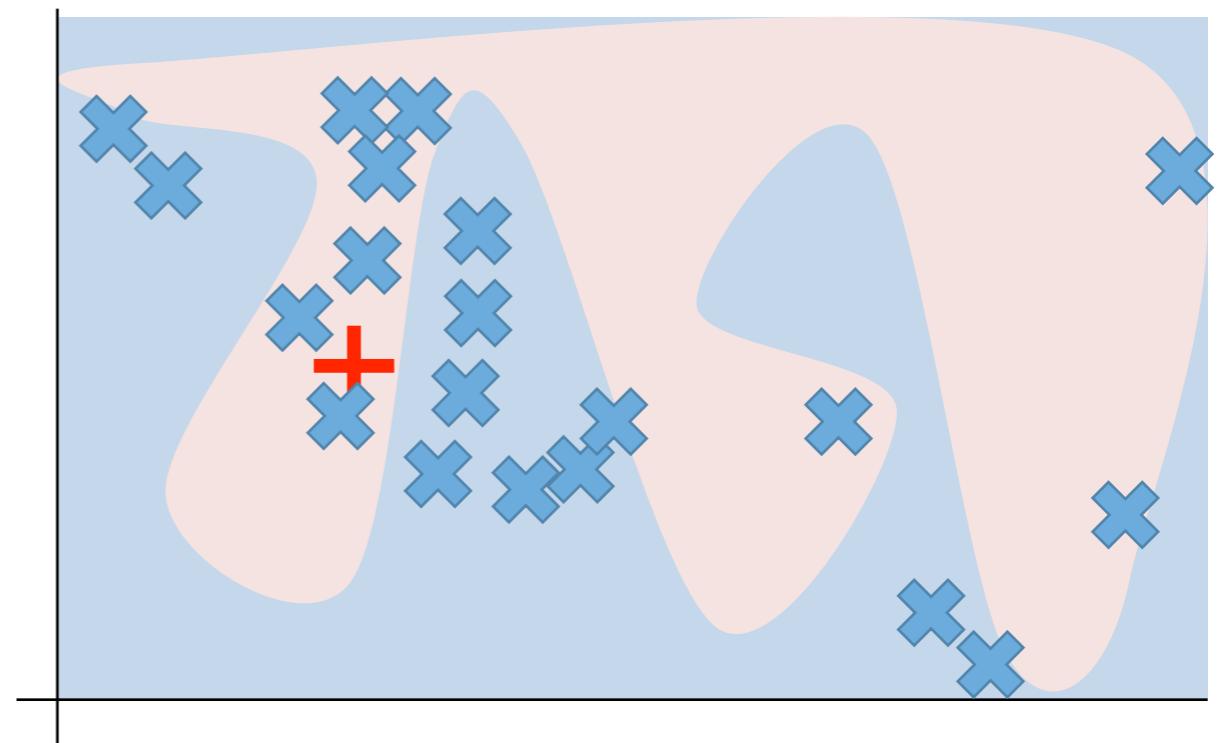
$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

measures how unfaithful is g to f in the locality around 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 

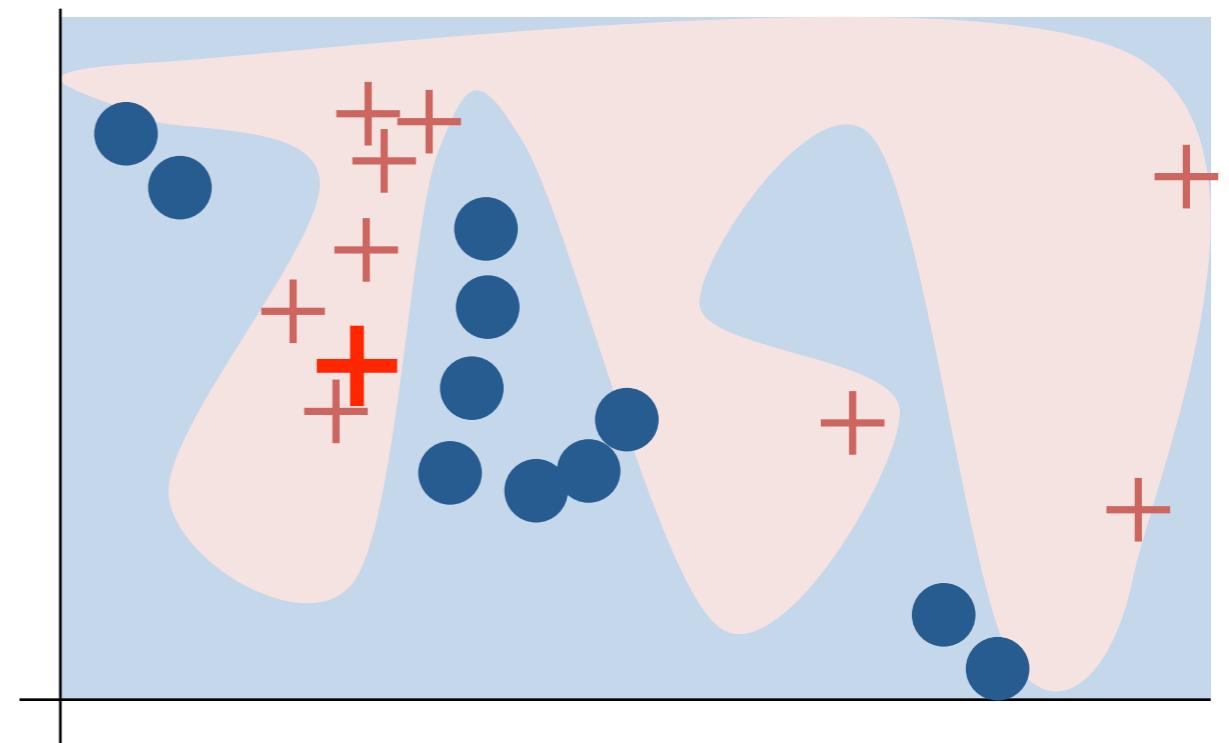


based on a slide by Marco Tulio Ribeiro, KDD 2016

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

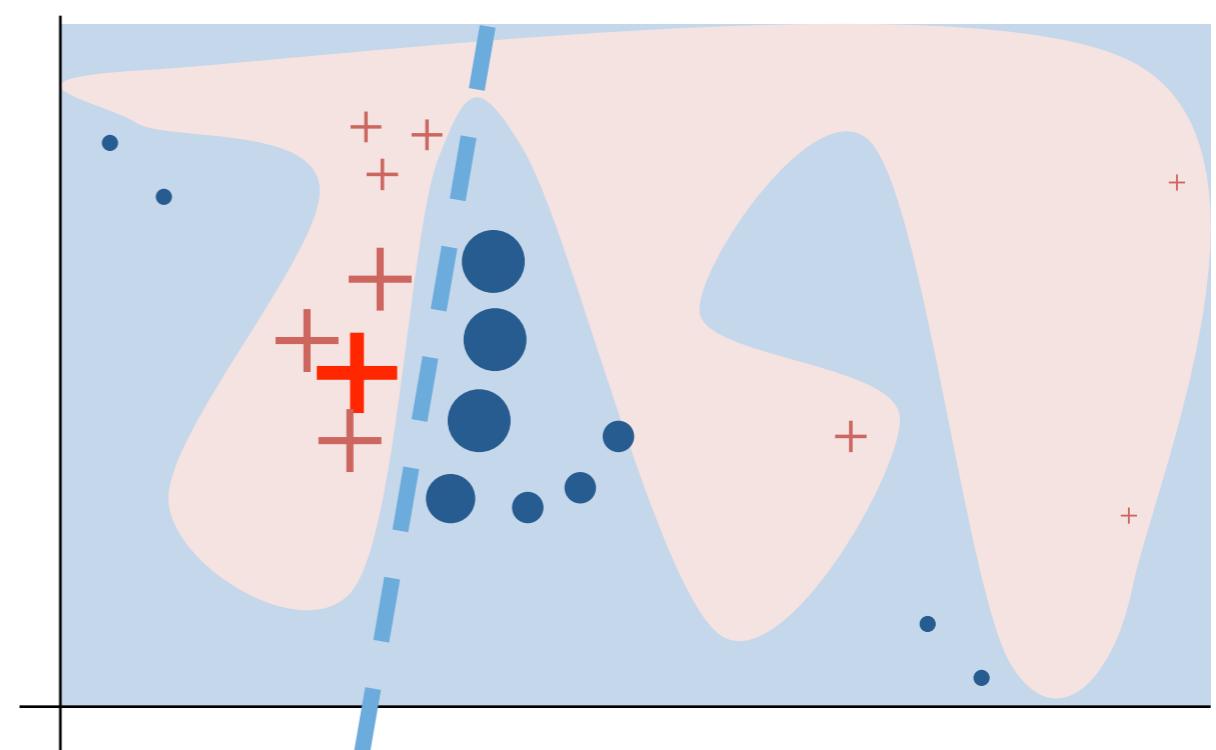


based on a slide by Marco Tulio Ribeiro, KDD 2016

Fidelity-interpretability trade-off

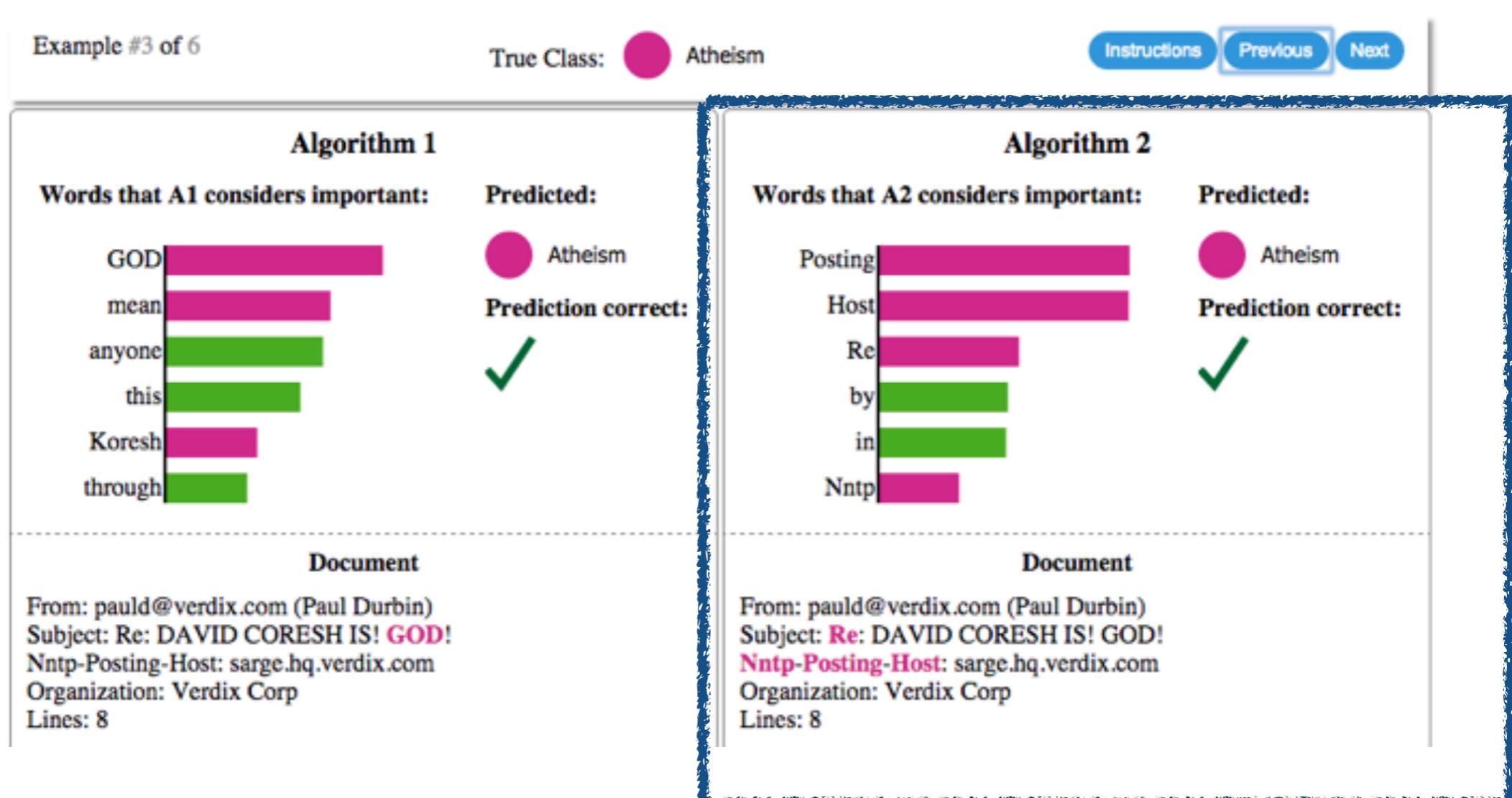
“The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.”

1. sample points around 
2. use complex model \mathbf{f} to assign class labels
3. weigh samples according to π_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!

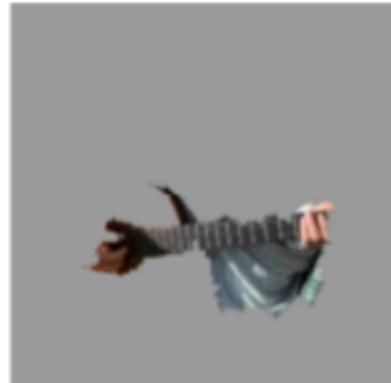
When accuracy is not enough

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



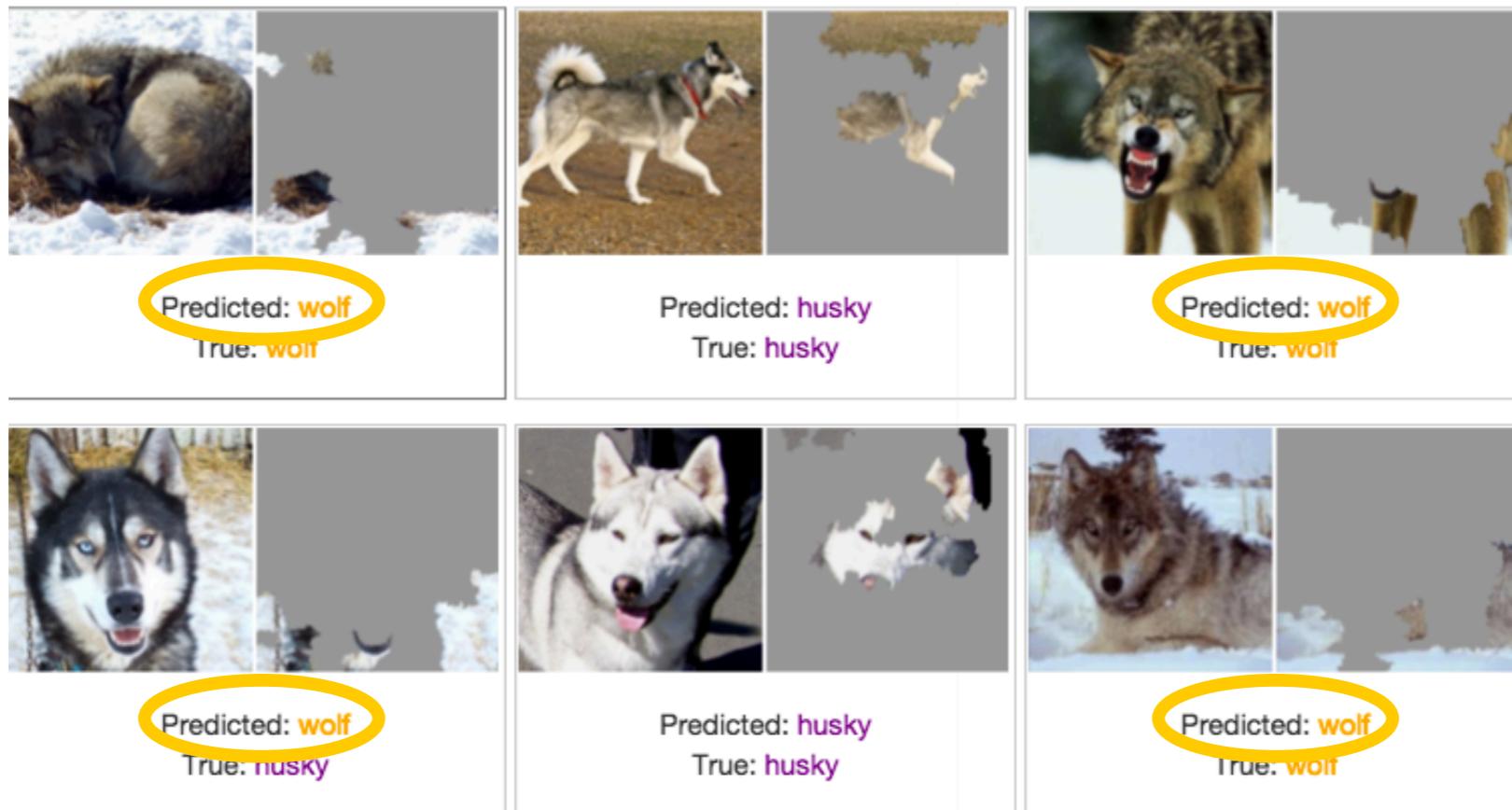
Only 1 mistake!!!

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

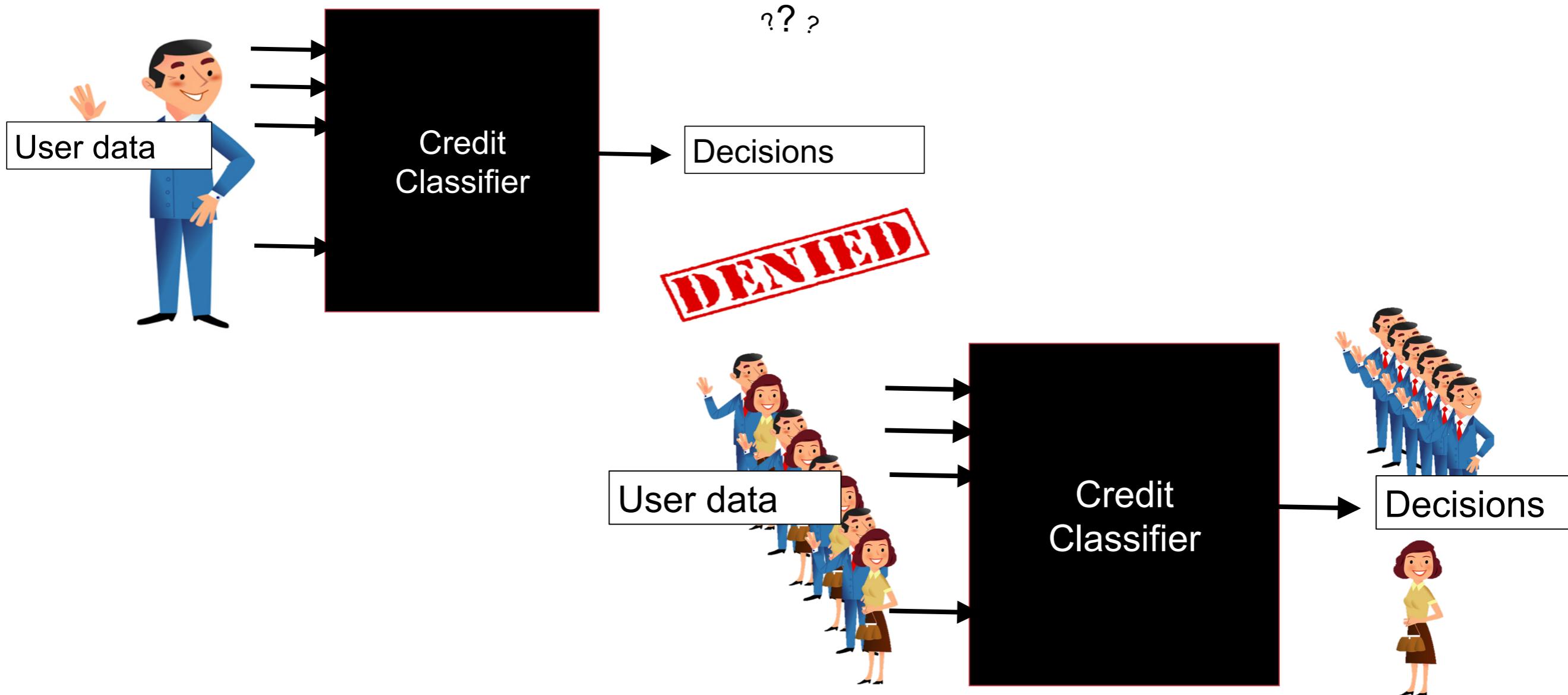


Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

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

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

Auditing black-box models



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?

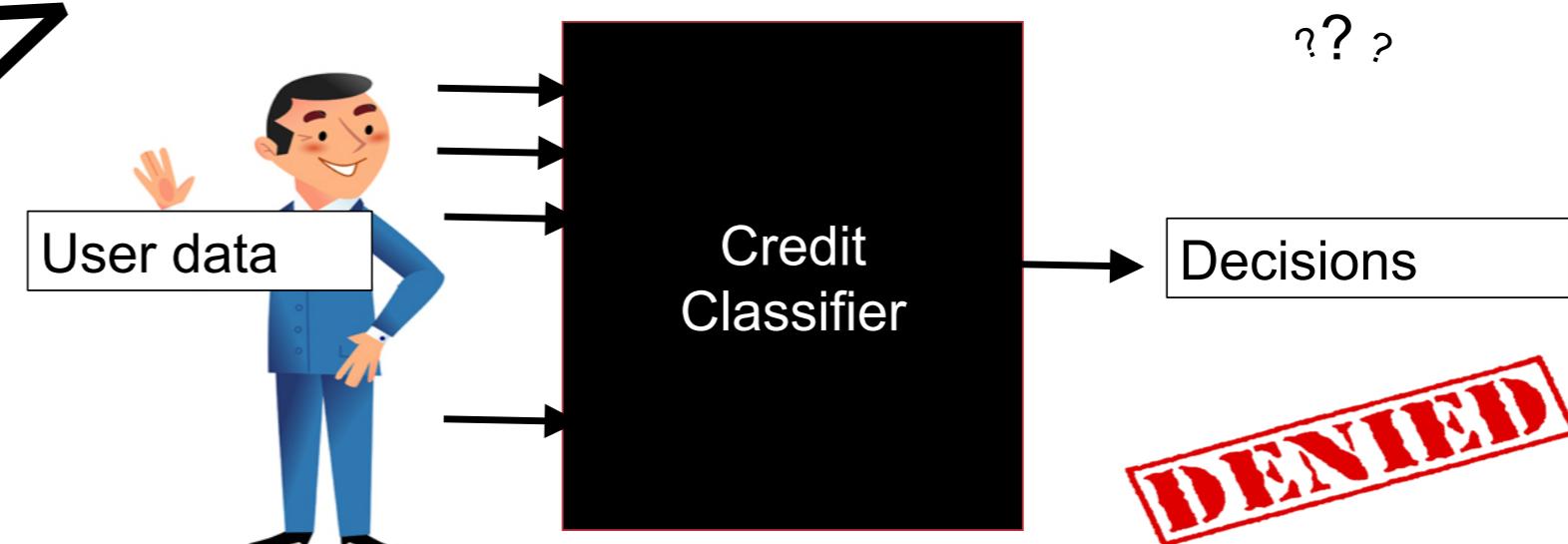
Disparity: Which inputs influence men getting more positive outcomes than women?

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

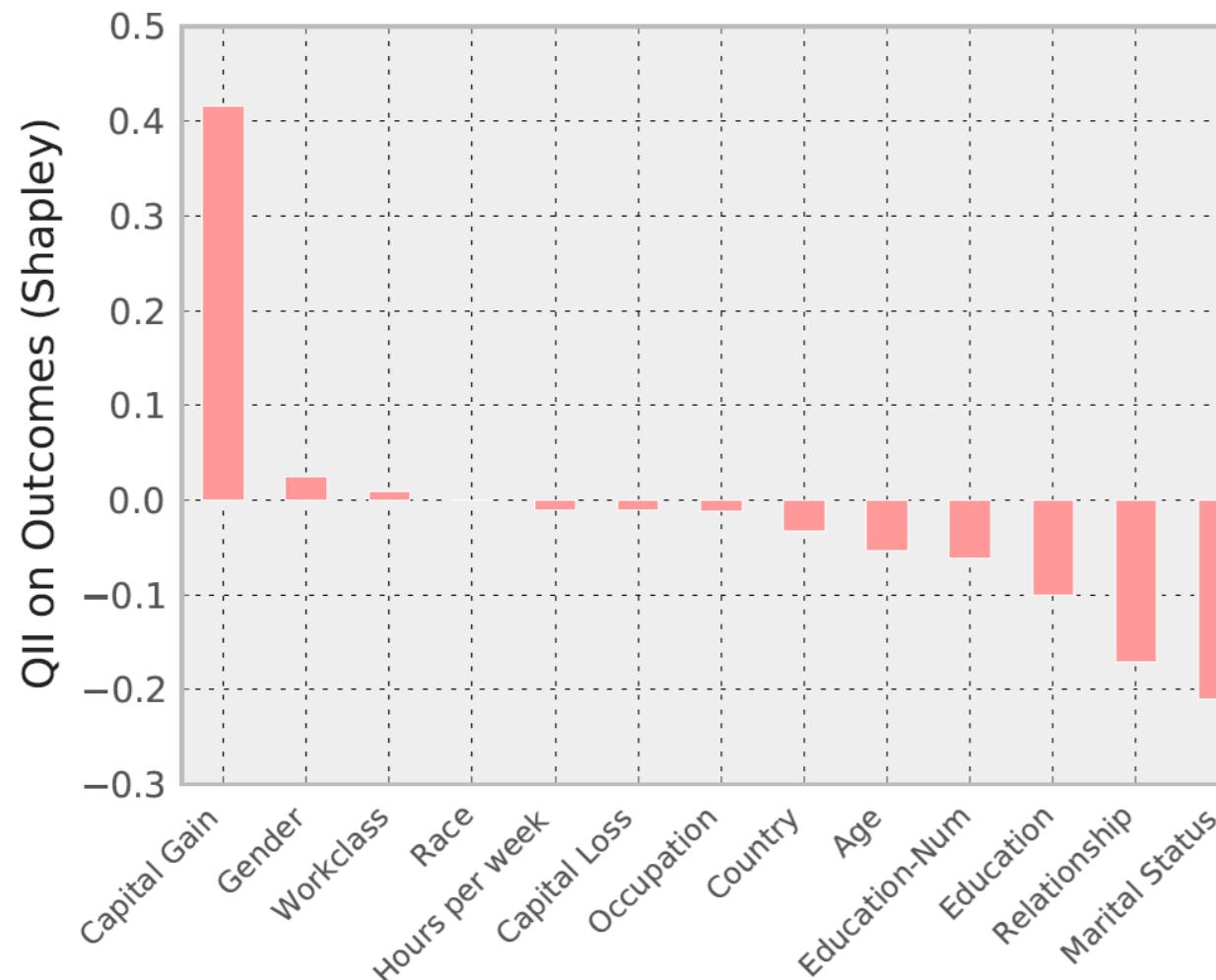
Running example

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**

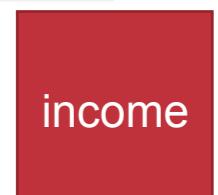
Transparency report: Mr. X

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

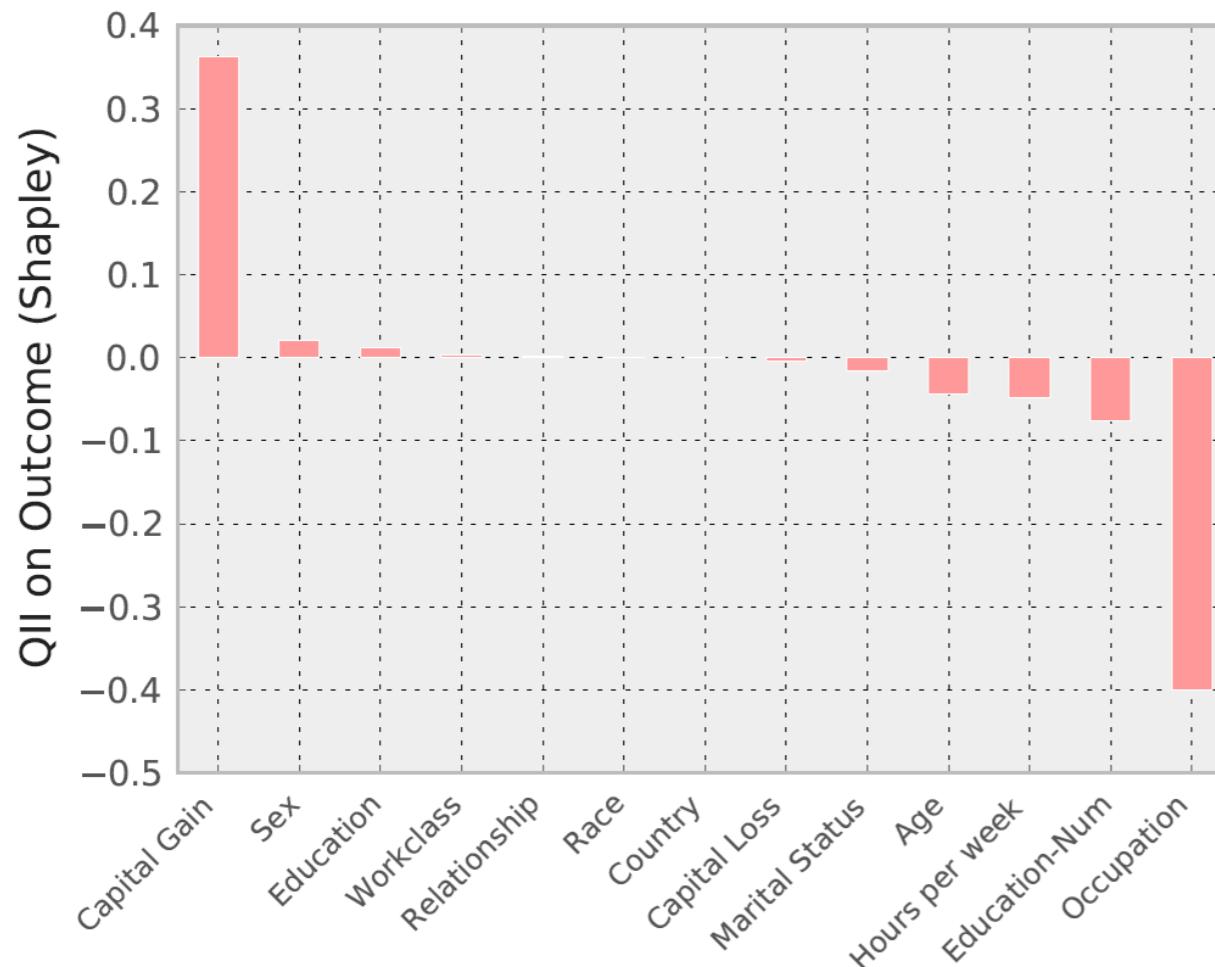
DENIED



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

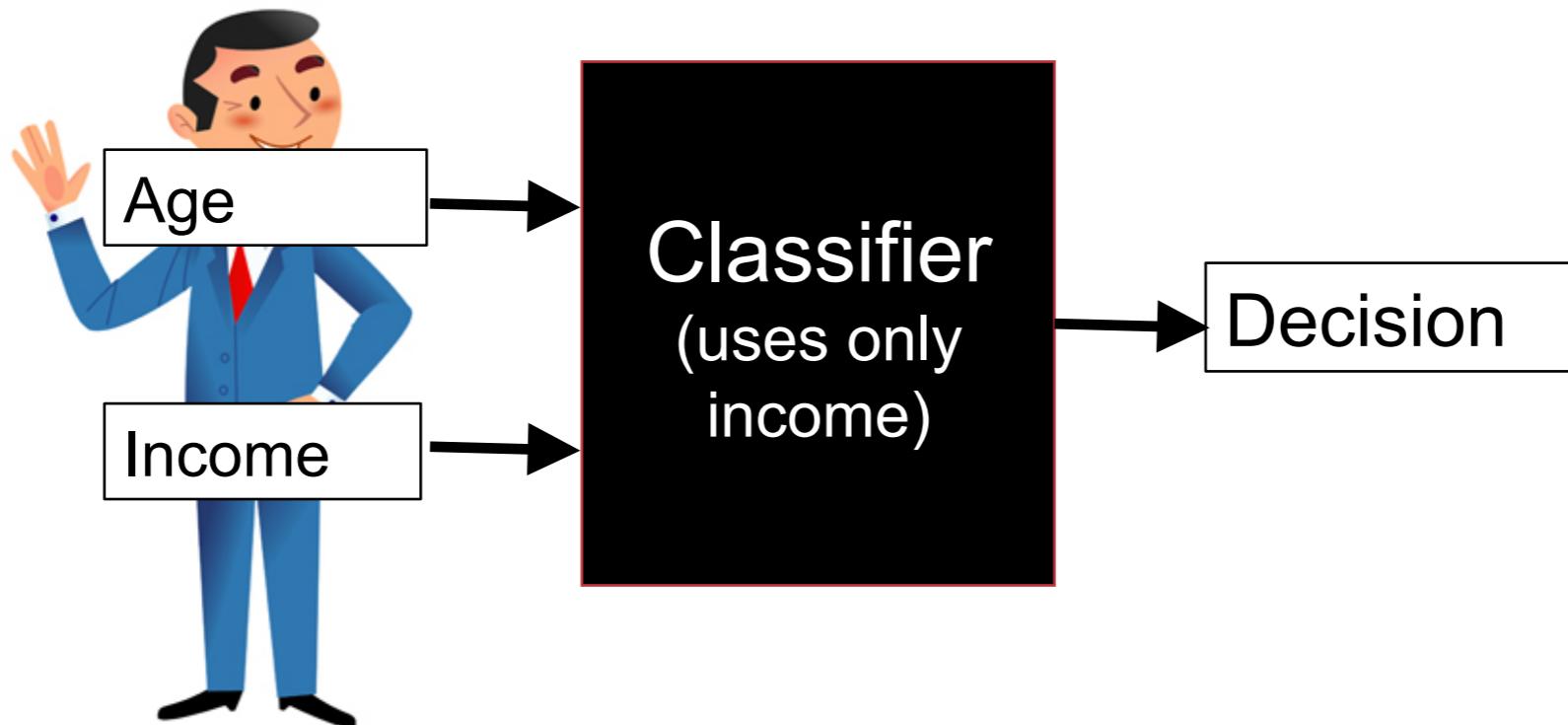


images by Anupam Datta

Unary QII

images by Anupam Datta

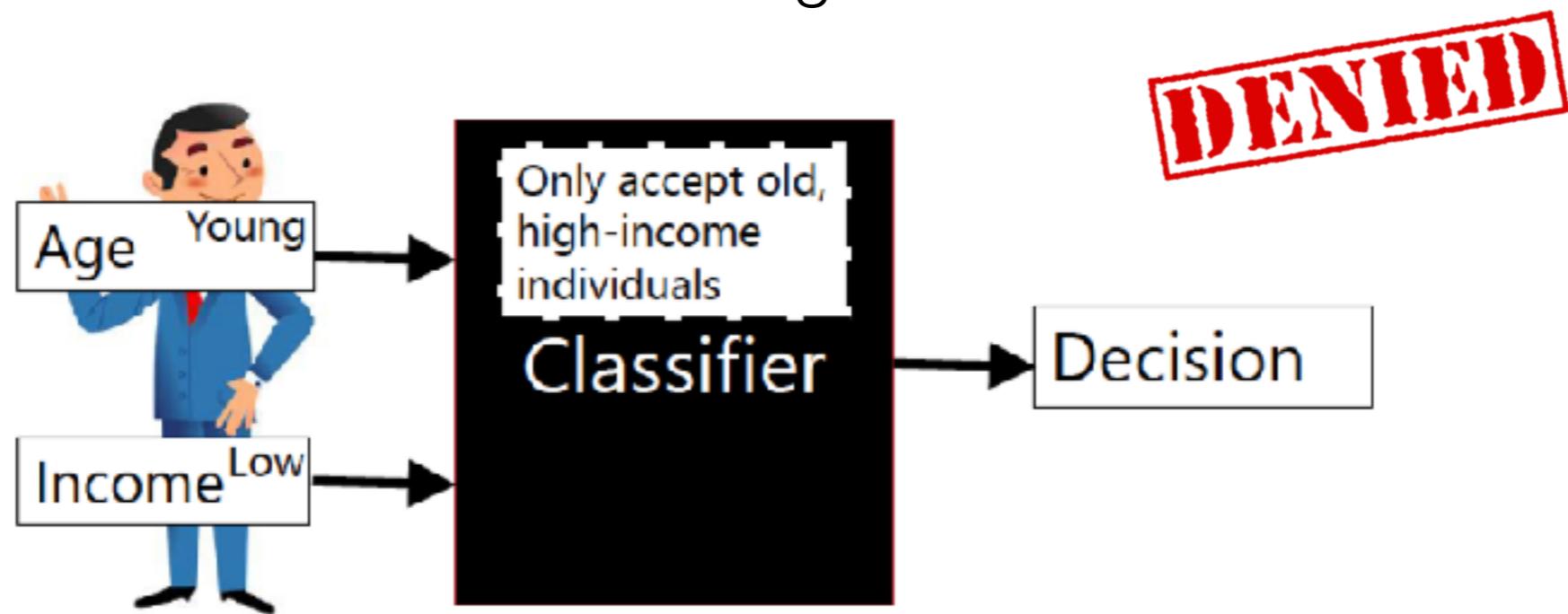
For a quantity of influence Q and an input feature i , the QII of i on Q is the difference in Q when i is changed via an **intervention**.



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

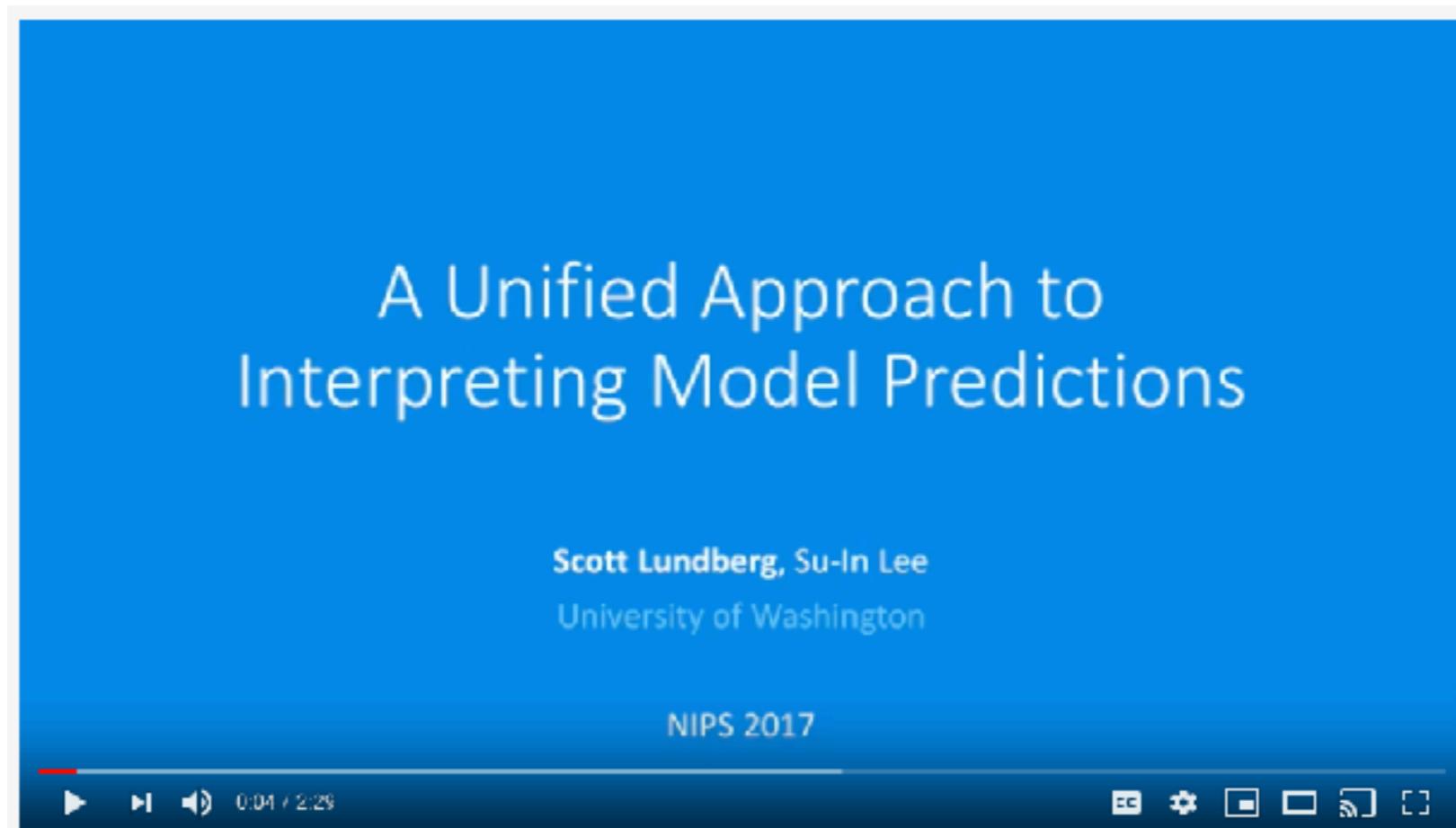
$m_i(\sigma)$ payoff corresponding to this permutation

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**

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

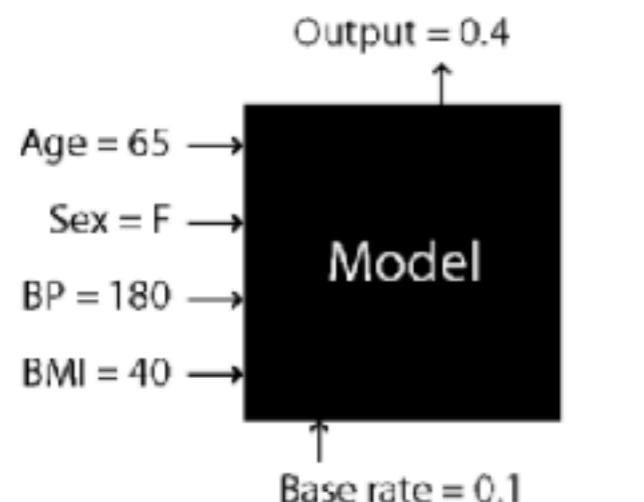
1. **Local accuracy**: $g(x')$ matches the original model $f(x)$ when x' is the **simplified input** corresponding to x .
2. **Missingness**: if x'_i — the i^{th} feature of simplified input x' — is missing, then it has no attributable impact for x
3. **Consistency (monotonicity)**: if toggling off feature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

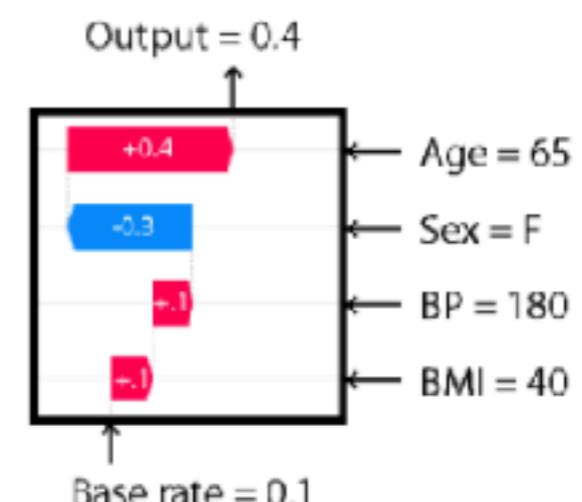
README.md



SHAP



Explanation →



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