

# Responsible Data Science

## Transparency & Interpretability

Auditing black-box models

*April 5, 2022*

---

**Prof. Julia Stoyanovich**

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



NYU

TANDON SCHOOL  
OF ENGINEERING

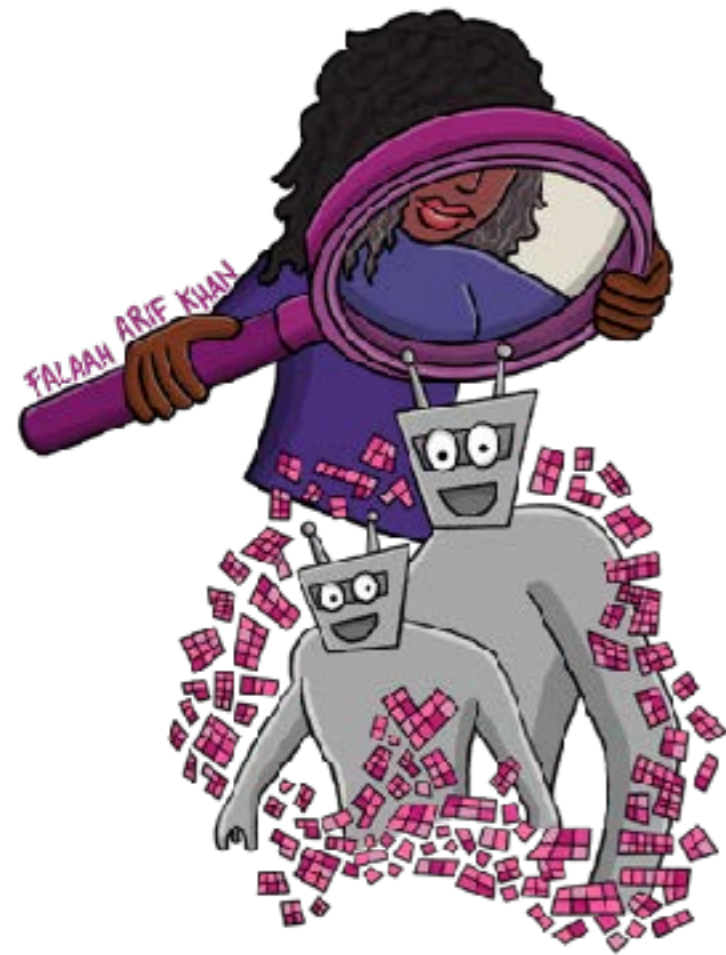


NYU

Center for  
Data Science

r/ai

# Terminology & vision



transparency, interpretability,  
explainability, intelligibility

responsible AI



agency, responsibility

# 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.

# Staples discounts

December 2012

## THE WALL STREET JOURNAL.

WHAT THEY KNOW

### Websites Vary Prices, Deals Based on Users' Information

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

December 24, 2012

WHAT PRICE WOULD YOU SEE?



It was the same Staples.com price for the same Staples.com product. The price in New York was \$15.79, while in Los Angeles, a few miles away, it was \$14.99.

A key difference: the user's location.

A Wall Street Journal investigation found that the Staples.com 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?

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

# Online job ads

theguardian

July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

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



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

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

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

One experiment showed that Google displayed ads for a career coaching service for “\$200k+” executive jobs **1,852 times to the male group and only 318 times to the female group.**

Another experiment, in July 2014, showed a similar trend but was not statistically significant.

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

# Online job ads

theguardian

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

July 2015

## 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 or demographic

One experiment showed ads for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Another experiment showed similar trends

**What** are we explaining?

To **Whom** are we explaining?

**Why** are we explaining?

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

# Instant Checkmate

February 2013

Google  
AdSense

The screenshot shows a user profile on the 'checkmate' website. The profile is for 'LATANYA SWEENEY' with a 'CERTIFIED' badge. The profile includes sections for Personal information, Location, Related Persons, Marriage/Divorce, Criminal History, Licenses, and Sex Offenders. A 'Criminal History' section is highlighted, showing a table for 'Possible Matching Arrest Records' with columns for Name and County and State. The table currently shows 'No matching arrest records were found.'

**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/>

FALAH ANIF KHAN



# Nutritional labels

**SIDE-BY-SIDE COMPARISON**

Original Label	New Label																								
<p><b>Nutrition Facts</b> Serving Size 2/3 cup (55g) Servings Per Container About 6</p> <p>Amount Per Serving</p> <p><b>Calories 230</b>    Calories from Fat 72</p> <p><b>Total Fat 8g</b>    12% Saturated Fat 1g    5% Trans Fat 0g</p> <p><b>Cholesterol 0mg</b>    0% <b>Sodium 150mg</b>    7% <b>Total Carbohydrate 35g</b>    12% Dietary Fiber 4g    16% Sugars 1g</p> <p><b>Protein 3g</b></p> <p>Vitamin A 10% Vitamin C 8% Calcium 50% Iron 45%</p> <p><small>*Percent Daily Values are based on a diet of other people's secrets.</small></p> <table border="1"> <tr> <td></td> <td>Less than</td> <td>4g</td> <td>8g</td> </tr> <tr> <td>Total Fat</td> <td>Less than</td> <td>30g</td> <td>75g</td> </tr> <tr> <td>Cholesterol</td> <td>Less than</td> <td>300mg</td> <td>300mg</td> </tr> <tr> <td>Sodium</td> <td>Less than</td> <td>2,400mg</td> <td>2,400mg</td> </tr> <tr> <td>Total Carbohydrate</td> <td></td> <td>30g</td> <td>17g</td> </tr> <tr> <td>Dietary Fiber</td> <td></td> <td>2g</td> <td>3g</td> </tr> </table>		Less than	4g	8g	Total Fat	Less than	30g	75g	Cholesterol	Less than	300mg	300mg	Sodium	Less than	2,400mg	2,400mg	Total Carbohydrate		30g	17g	Dietary Fiber		2g	3g	<p><b>Nutrition Facts</b> 6 servings per container <b>Serving size 2/3 cup (55g)</b></p> <p>Amount per serving</p> <p><b>Calories 230</b></p> <p><b>Total Fat 8g</b>    16% Saturated Fat 1g    5% Trans Fat 0g</p> <p><b>Cholesterol 0mg</b>    0% <b>Sodium 100mg</b>    7% <b>Total Carbohydrate 37g</b>    13% Dietary Fiber 4g    14% Total Sugars 12g Includes 10g Added Sugars    20%</p> <p><b>Protein 3g</b></p> <p>Vitamin D 2mg    10% Calcium 250mg    20% Iron 8mg    45% Potassium 235mg    9%</p> <p><small>*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 as a general nutrition guide.</small></p>
	Less than	4g	8g																						
Total Fat	Less than	30g	75g																						
Cholesterol	Less than	300mg	300mg																						
Sodium	Less than	2,400mg	2,400mg																						
Total Carbohydrate		30g	17g																						
Dietary Fiber		2g	3g																						

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

An example of the old nutrition labels, left, and the new one. The new nutrition labels will display calories and serving size more prominently, and include added sugars for the first time. PHOTO: FOOD AND DRUG ADMINISTRATION/ASSOCIATED PRESS

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

**Security & Privacy Overview**  
Smart Device Co.

Smart Vision Gearbell NS200  
Firmware version: 2.5.1 - updated on: 11/10/2020  
The device was manufactured in: China

**Security updates:** Automatic - Available until: next 1/1/2022

**Access control:** Password - Force by default - Use of changeable Multi-factor authentication. Multiple user accounts are allowed.

**Camera data collection:** Visual, Audio, Physiological, Location

**Smart type:** Camera, Microphone

**Purpose:** Provide device functions, research for device storage

**Data stored on device:** Identified, Identified - Opt-out: allow

**Data shared on cloud:** Identified, Government, Manufacturer

**Shared with:** Not disclosed, Not set

**Sold to:** Not disclosed, Not set

**Other collected data:** Motor, Account info, Payment info, Contact info, Device usage info, Device usage info

**Privacy policy:** [www.NS200.com/SmartVisionGearbellNS200/privacy](http://www.NS200.com/SmartVisionGearbellNS200/privacy)

**Detailed Security & Privacy Label:** [www.iotsecurityprivacy.org/labels](http://www.iotsecurityprivacy.org/labels)

CVU for Security and Privacy Label - OESPL 1.0 - [www.iotsecurityprivacy.org](http://www.iotsecurityprivacy.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

Amrita Dasgupta, Shiyak Jee, Yuh Zick  
Carnegie Mellon University, Pittsburgh, USA  
{dasgupta, shiyakj, yzick}@cmu.edu

**Abstract**—Algorithmic systems that employ machine learning play an increasing role in making substantive decisions in modern society, ranging from online personalization to insurance and credit decisions to predictive policing. But their decision-making processes are often opaque—it is often not obvious why a certain decision was made. We develop a formal foundation to improve the transparency of such decision-making systems. Specifically, we introduce a family of quantitative input influence (QII) measures that capture the degree of influence of inputs on outputs of systems. These measures provide a foundation for the design of transparency reports that accompany system decisions (e.g., explaining a specific credit decision) and for better tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).

Collectively, our reports (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., separate loan rates for graduates). Finally, showing single inputs may not always have high influence, the QII measures also quantify the *total* influence of a set of inputs (e.g., age and income) on outcomes (e.g., loan decisions) and the *average* influence of individual inputs within such a set (e.g., income). Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using multiplicative aggregation measures, such as the Shapley value, previously applied to measure influence in voting. Further, since transparency reports could compromise privacy, we explore the transparency queries it should and must not answer if such transparency reports can be made differentially private with very little addition of noise.

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

### 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, health-care, education, insurance, law enforcement, and defense [1], [2], [3], [4], [5]. Yet their decision-making processes are often opaque. Algorithmic transparency is an emerging research area aimed at explaining decisions made by algorithmic systems.

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

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

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

Our goal is to inform the design of transparency reports, which include answers to transparency queries of this form. To be precise, let us consider a predictive modeling system that forecasts (a)re critical activity based on historical data, individuals high or low risk (e.g., visit to 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 credit criminal history, on the system's decision. An oversight agency or the public may desire a transparency report that provides answers to aggregate transparency queries, such as the influence of sensitive inputs (e.g., gender, race) on the system's decisions concerning the entire population or about systematic differences in decisions

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

Marco Tulo Ribeiro  
University of Washington  
Seattle, WA 98195, USA  
marco.tulo@cs.washington.edu

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

Carla 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 predictor, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an automatically learned predictor into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier as an interpretable and faithful model, by learning an interpretable model locally around the prediction. We also propose a method to explain models by providing approximate individual predictions and their explanations in a user-oriented way, leaving the task on a reasonable explanation problem. We demonstrate the flexibility of these methods by explaining a forest model for text (e.g. random forests) and image classification (e.g. neural networks). We show the ability of explanations to novel experiments, both conducted with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, exposing an unwanted filter, and identifying why a classifier should not be trusted.

### 1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans is an oft-underestimated aspect in the field. Whether humans are directly using machine learning classifiers to make, or are deploying models within other products, a vital concern remains: if the users do not trust a model or a prediction, they will not use it. It is important to differentiate between two different (and related) definitions of trust: (1) a *static* trust in a predictor, i.e. whether the user trusts a predictor regardless of the prediction it will make when used; (2) a *dynamic* trust in a predictor, i.e. whether the user trusts a model to behave in a reasonable way if deployed. Both are directly impacted by

Provision is made slightly and copies of all in part of this work for personal or internal use only, without the payment of any fee, except for those that are specifically stated otherwise. This permission is given on the basis that the requester is an individual or an organization that is not a competitor of the requester. This permission is not to be used for any other purpose, including for advertising or promotional purposes, for creating new collective works, or for resale.

IEEE, 2016. This work is licensed under a Creative Commons Attribution 4.0 International License. For more information, see <http://creativecommons.org/licenses/by/4.0/>.

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 critical decisions. When one makes a loan rate for a particular client [6] or a credit decision, for example, such decisions cannot be made automatically, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to trust the model as a whole before deploying it "in the wild". To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often statistically different, and human- and evaluation metrics may not be indicators of the model's good. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this sense, it is important to build tools for supporting which instances to inspect, especially for large datasets.

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

- LIME, an algorithm that can explain the predictions of any classifier or regressor in a faithful way by approximating it locally with an interpretable model.
- SHAP, a method that selects a set of representative instances with explanations to address the "trusting the model" problem, via combinatorial optimization.
- Comprehensive evaluation with standard and human-subject data, where we measure the impact of explanations on trust and awareness levels. In our experiments, non-experts using LIME are able to pick which classifier from a pair generalizes better in the real world, further, they are able to greatly improve an automatically classifier trained on 20 image classes, by using feature engineering using LIME. We also show how understanding the prediction of a neural network on images helps practitioners know when and why they should not trust a model.

### 2. THE CASE FOR EXPLANATIONS

In "explaining a prediction", we mean providing context or clues a user can provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction. We

## A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg  
Paul G. Allen School of Computer Science  
University of Washington  
Seattle, WA 98195  
lund@cs.washington.edu

Su-In Lee  
Paul G. Allen School of Computer Science  
Department of Genome Sciences  
University of Washington  
Seattle, WA 98195  
mlu@cs.washington.edu

### Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive Explanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing that it 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 predictor 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 being able to forecast 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 [1, 8, 9, 7, 4, 1], but an understanding of how these methods relate and when one method is preferable to another is still lacking.

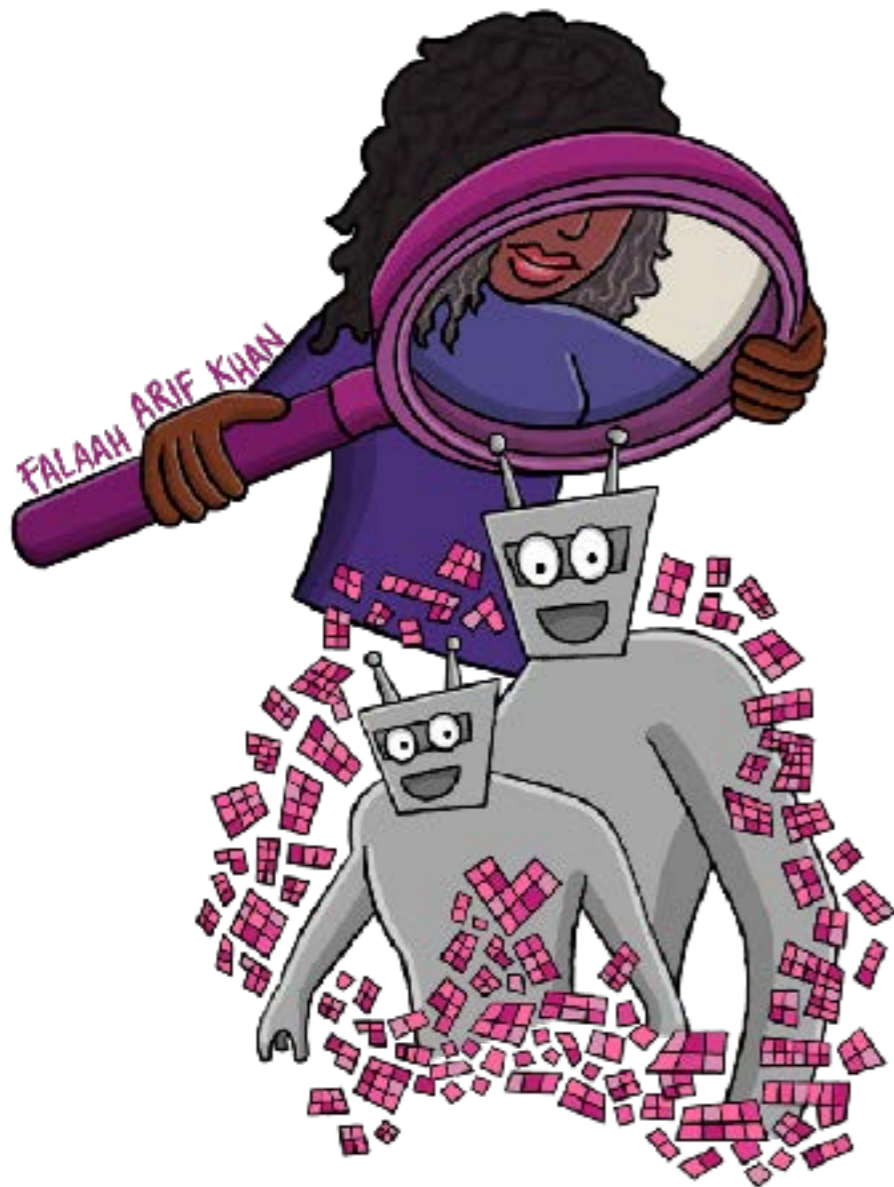
Here, we present a novel unified approach to interpreting model predictions.<sup>1</sup> Our approach leads to three potentially surprising results that bring clarity to the growing space of methods:

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

<sup>1</sup><https://github.com/lundberg/shap>

explaining black box  
models

# What are we explaining?



How does a system work?

How **well** does a system work?

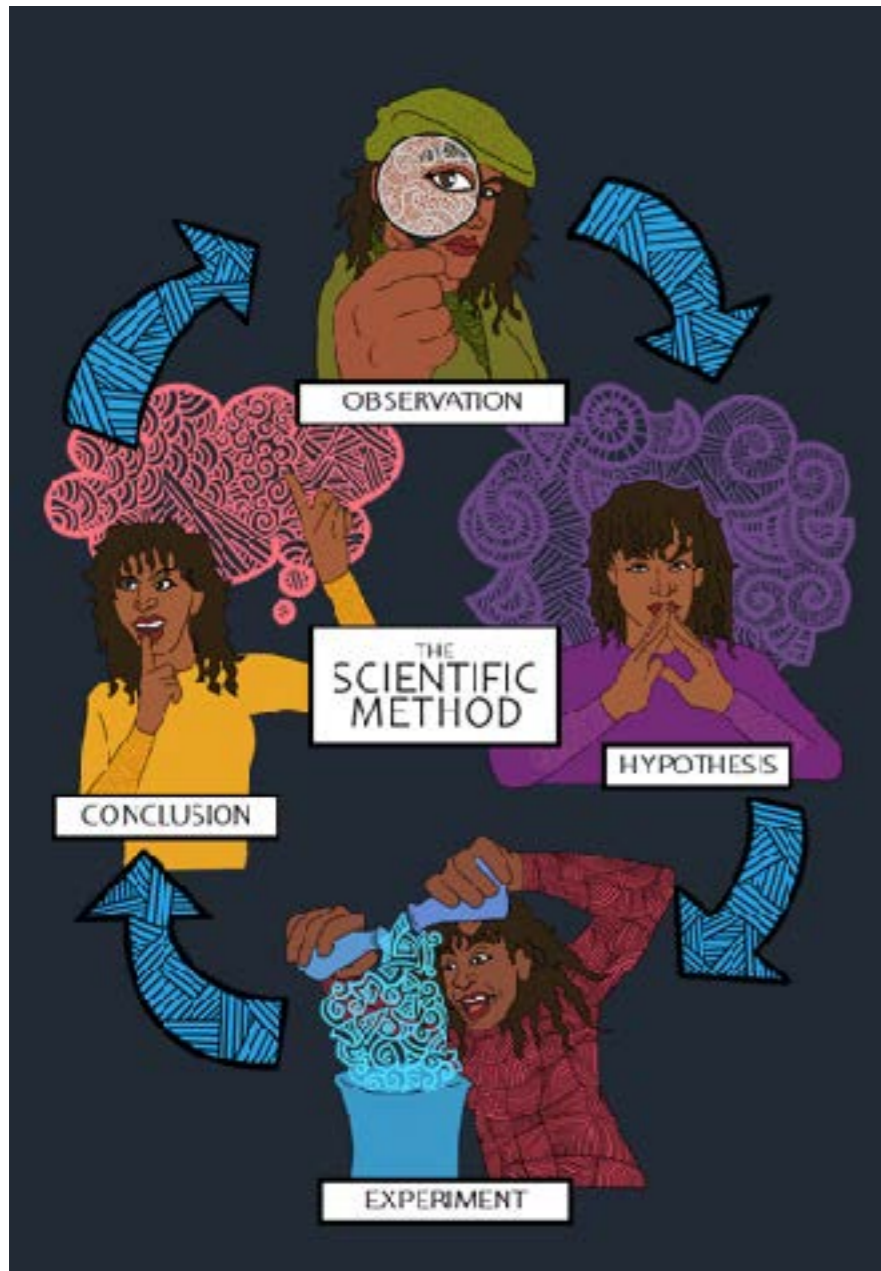
What does a system do?

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

October 13, 2014



Shane Creepingbear @Creepingbear · Oct 13, 2014

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



TIME

↻ 17

## Facebook Thinks Some Native American Names Are Inauthentic

BY JOSH SANBURN FEBRUARY 14, 2015

February 14, 2015

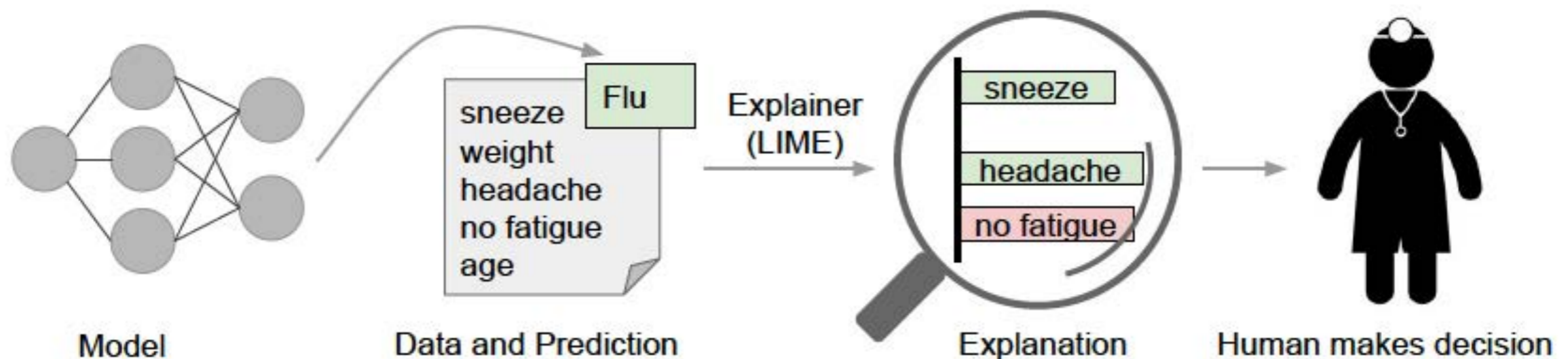
If you're Native American, Facebook might think your name is fake.

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.

# 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

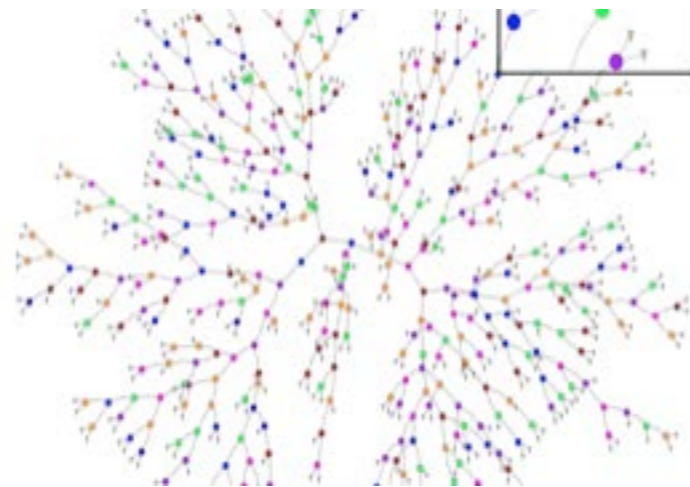
**what if patient id appears in green in the list? - an example of “data leakage”**

# LIME: Local explanations of classifiers

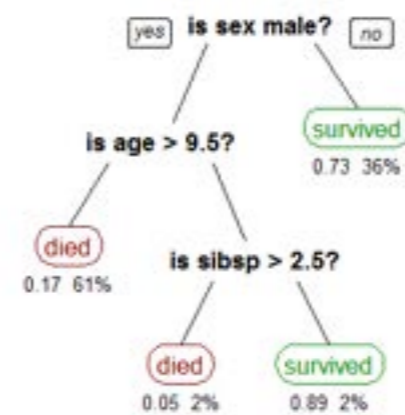
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



Definitely  
not interpretable



Potentially  
interpretable

slide by Marco Tulio Ribeiro, KDD 2016



# Explanations based on features

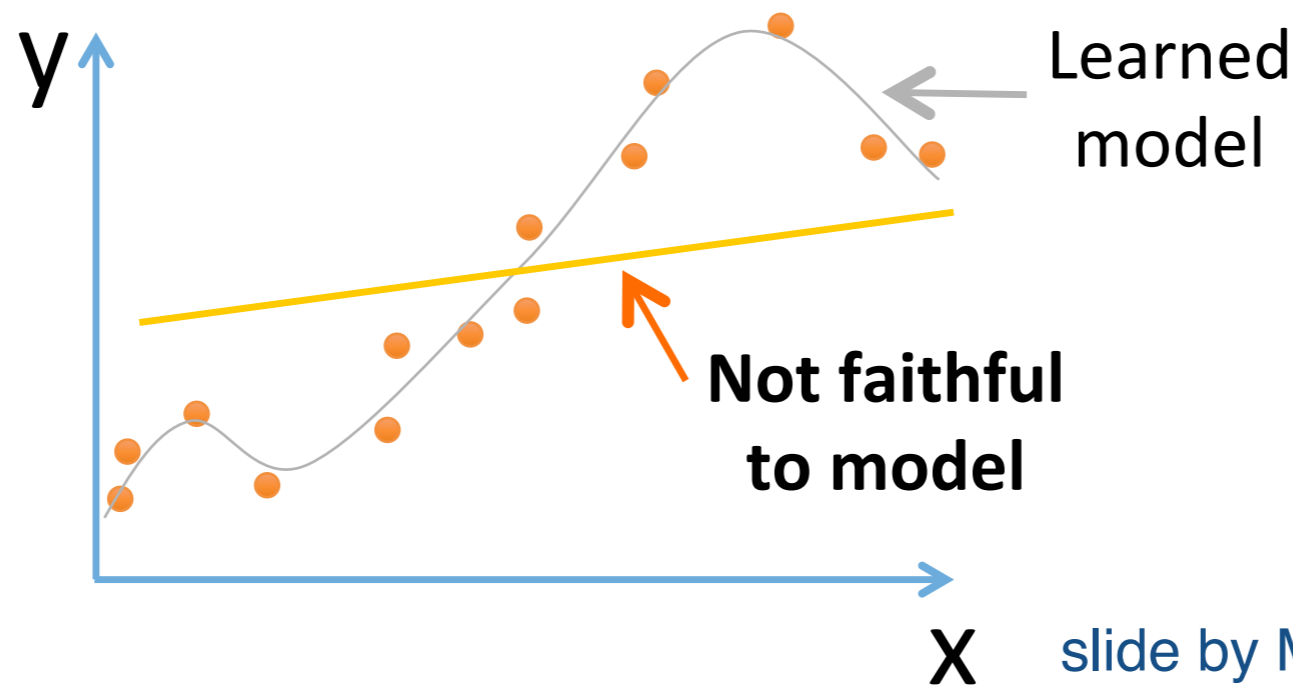
## Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

Faithful

- Describes how this model actually behaves



slide by Marco Tulio Ribeiro, KDD 2016

# Explanations based on features

## Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning

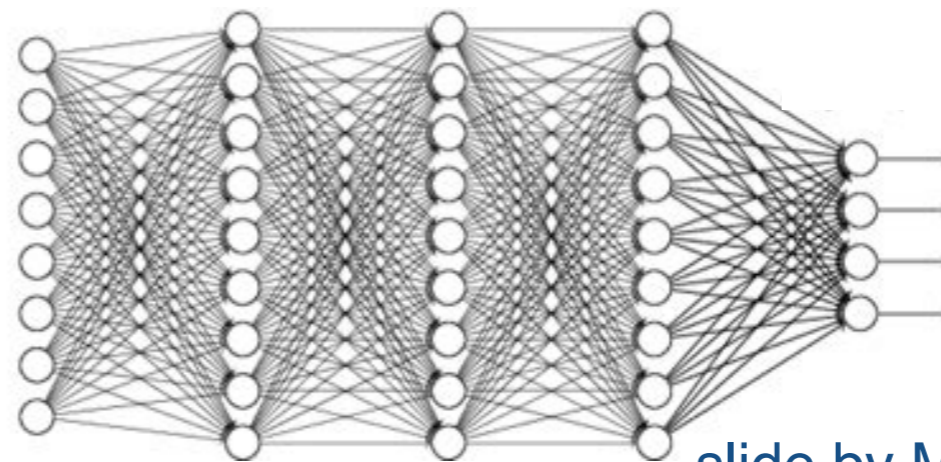
Faithful

- 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

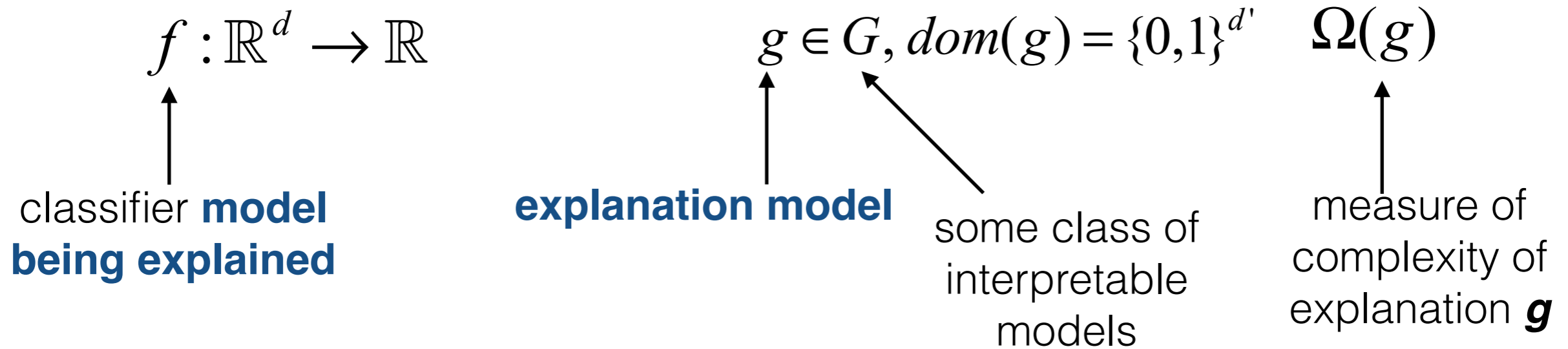
# Key idea: Interpretable representation

“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  $\mathbf{x}$  belongs to some class

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

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

**explanation**

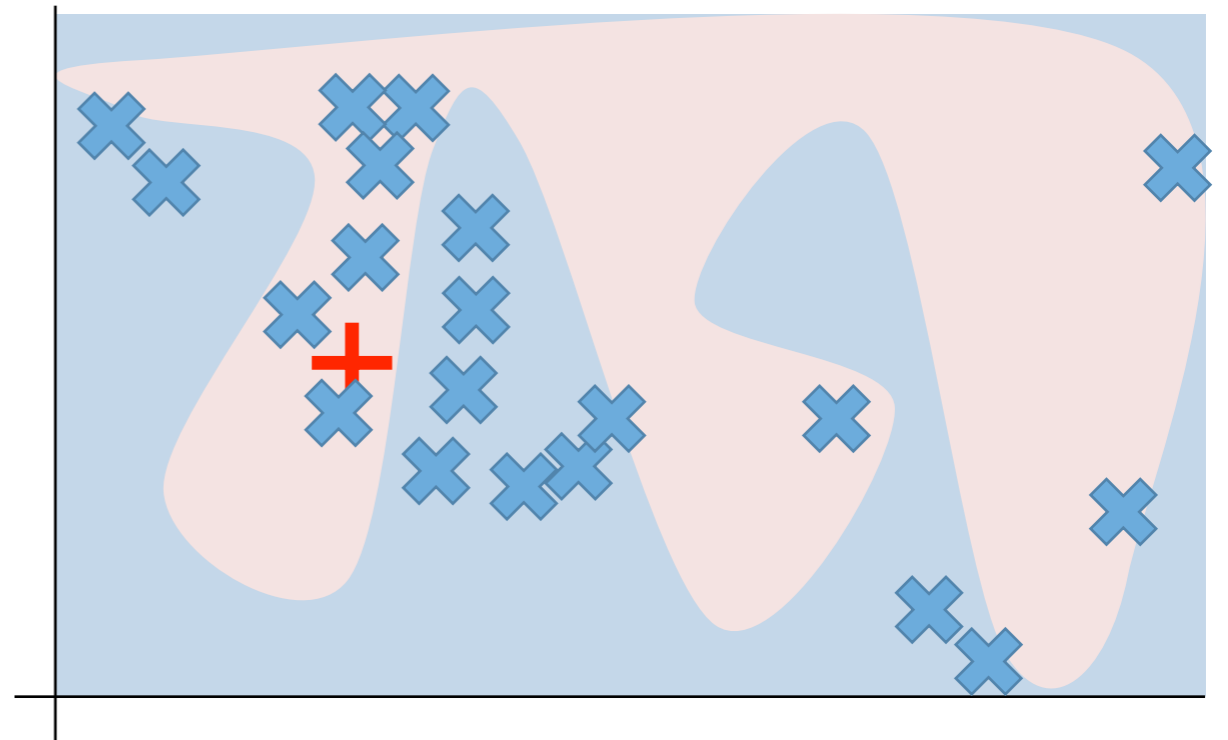
measures how unfaithful is  $\mathbf{g}$  to  $\mathbf{f}$  in the locality around  $\mathbf{x}$

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

# Fidelity-interpretability trade-off

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

1. sample points around +

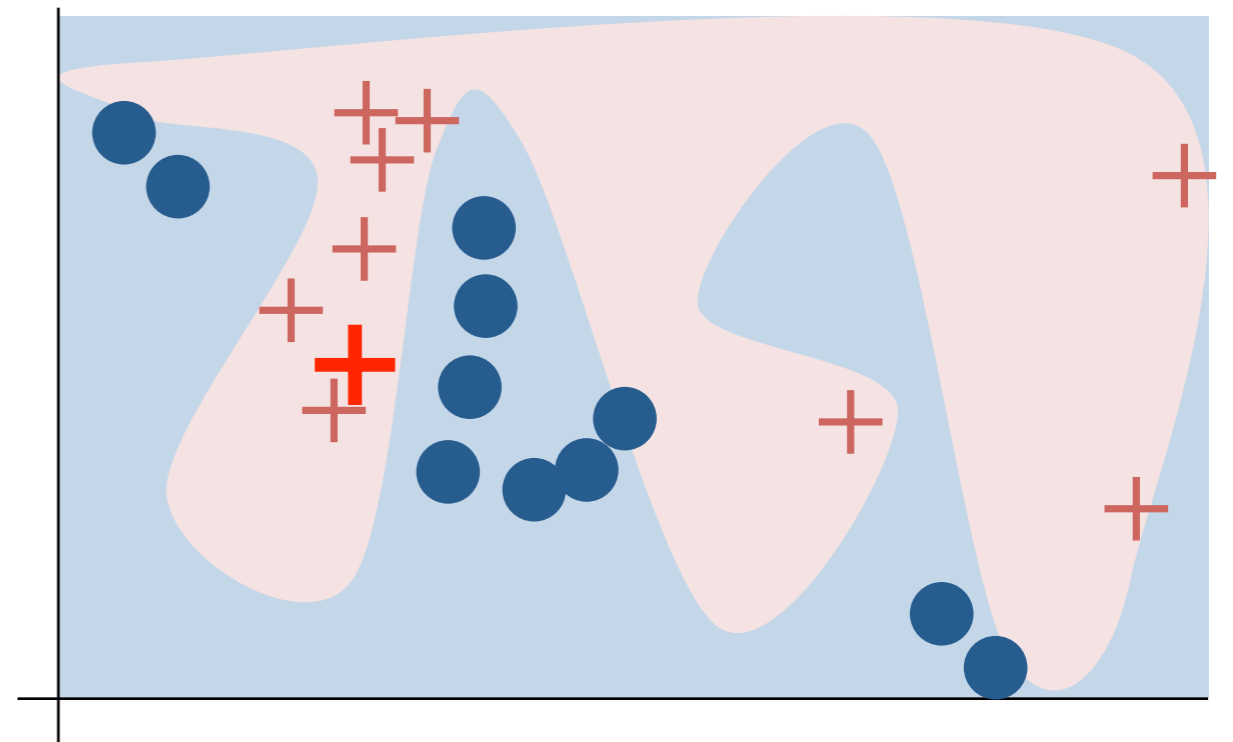


based on a slide by Marco Tulio Ribeiro, KDD 2016

# 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

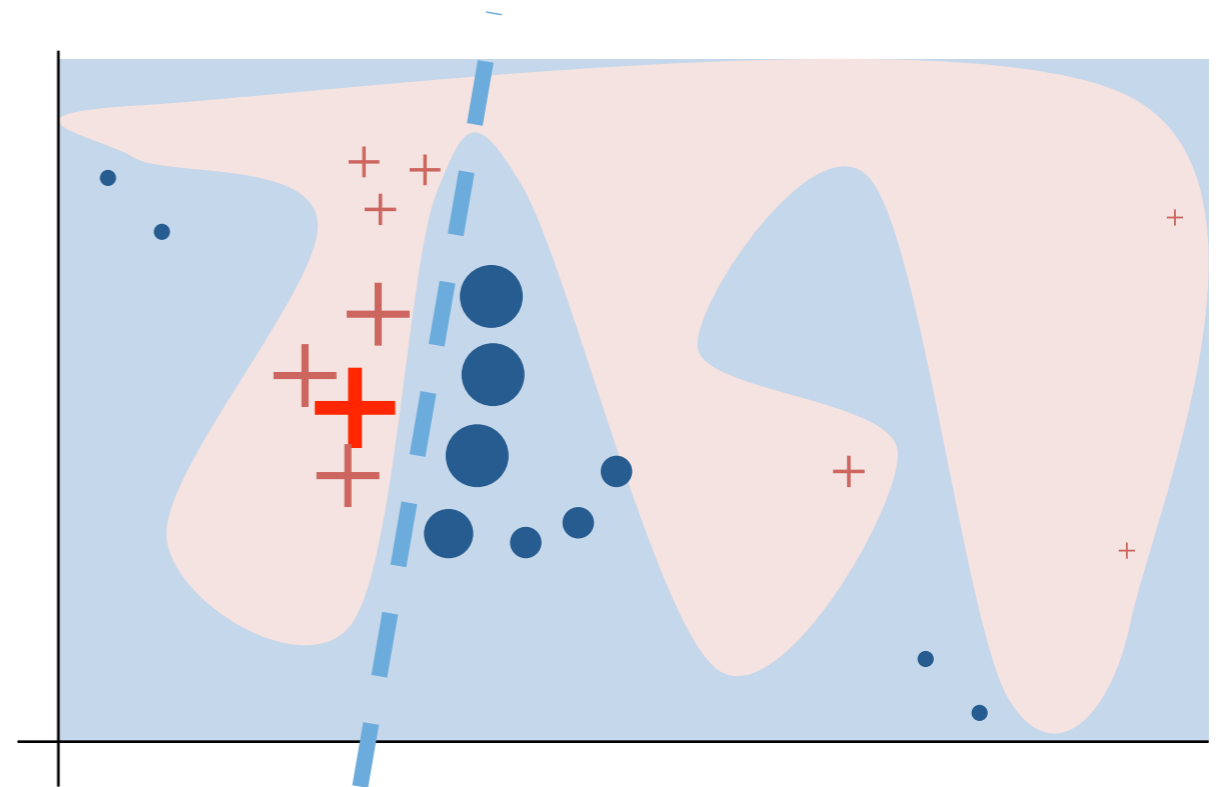


based on a slide by Marco Tulio Ribeiro, KDD 2016

# 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 **f** to assign class labels
3. weigh samples according to  $\pi_x$
4. learn simple model **g** according to samples



based on a slide by Marco Tulio Ribeiro, KDD 2016

# Example: text classification with SVMs

Example #3 of 6

True Class: ● Atheism

[Instructions](#) [Previous](#) [Next](#)

### Algorithm 1

**Words that A1 considers important:**

GOD	
mean	
anyone	
this	
Koresh	
through	

**Predicted:** ● Atheism

**Prediction correct:** ✓

---

**Document**

From: pauld@verdix.com (Paul Durbin)  
Subject: Re: DAVID CORESH IS! **GOD!**  
Nntp-Posting-Host: sarge.hq.verdix.com  
Organization: Verdix Corp  
Lines: 8

### Algorithm 2

**Words that A2 considers important:**

Posting	
Host	
Re	
by	
in	
Nntp	

**Predicted:** ● Atheism

**Prediction correct:** ✓

---

**Document**

From: pauld@verdix.com (Paul Durbin)  
Subject: **Re:** DAVID CORESH IS! GOD!  
**Nntp-Posting-Host:** sarge.hq.verdix.com  
Organization: Verdix Corp  
Lines: 8

**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



$$P(\text{Electric guitar}) = 0.32$$



Electric guitar - incorrect but  
reasonable, similar fretboard

$$P(\text{Acoustic guitar}) = 0.24$$



Acoustic guitar

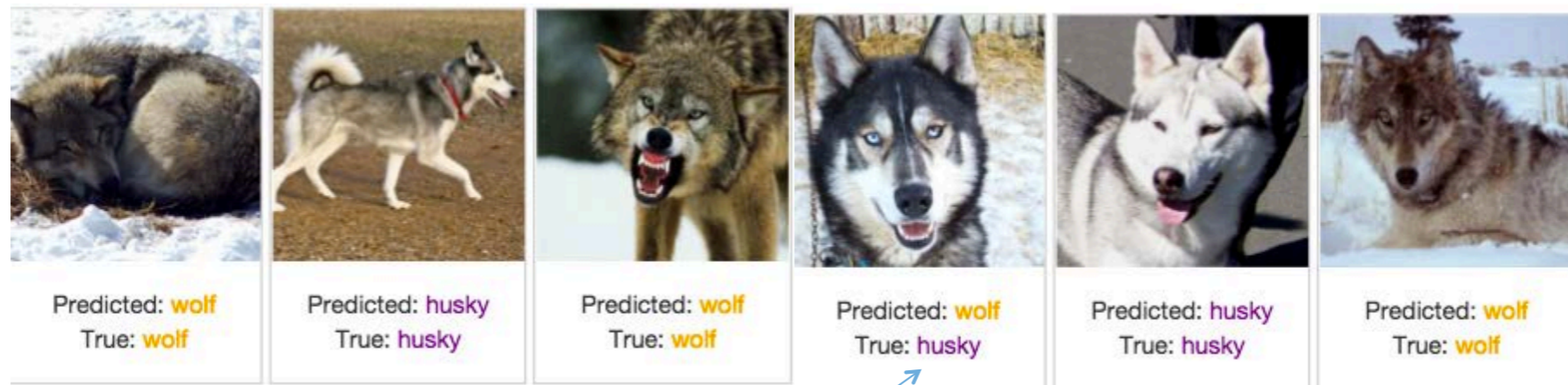
$$P(\text{Labrador}) = 0.21$$



Labrador

# When accuracy is not enough

Train a neural network to predict **wolf** v. **husky**



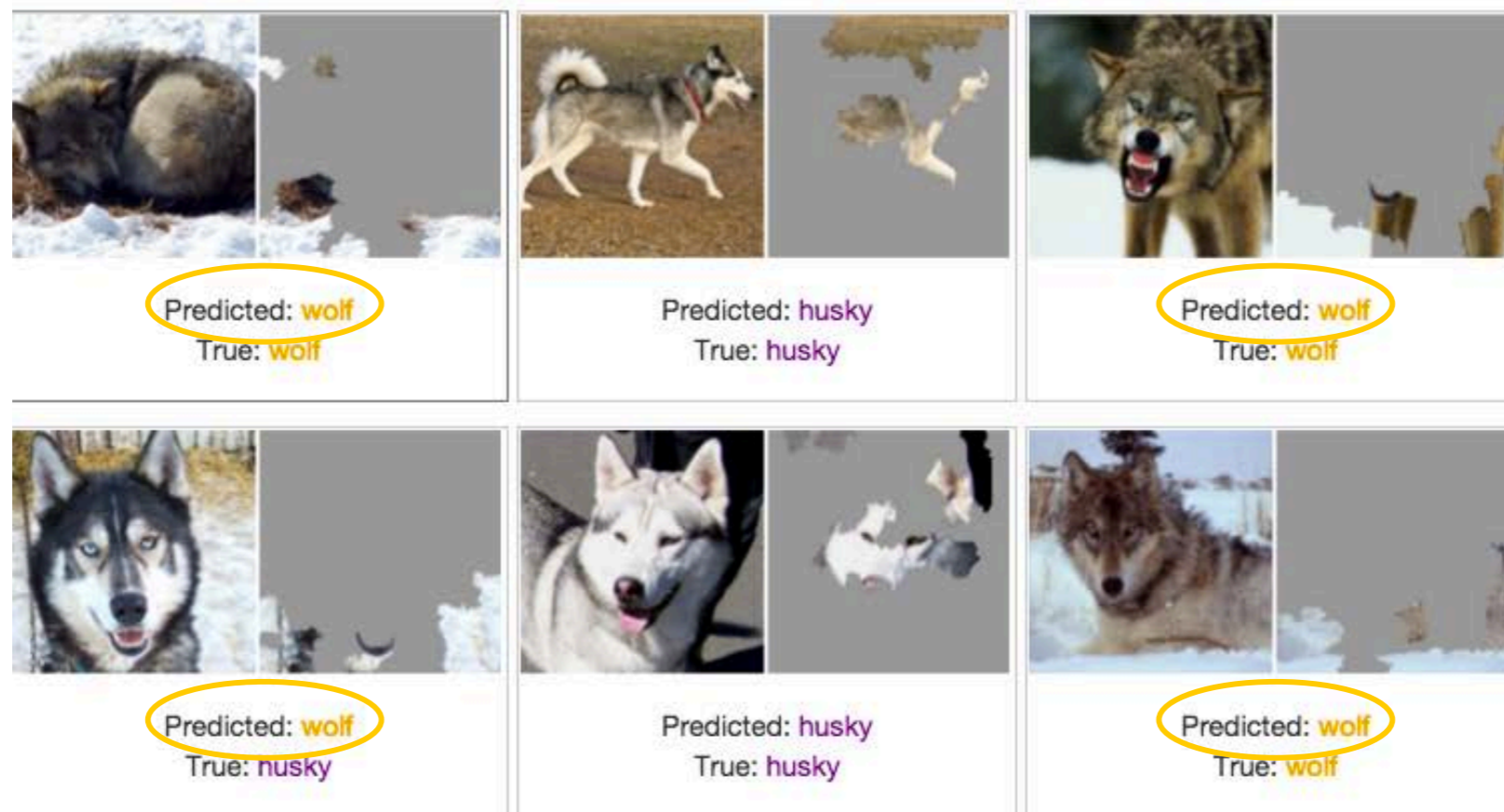
Only 1 mistake!!!

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

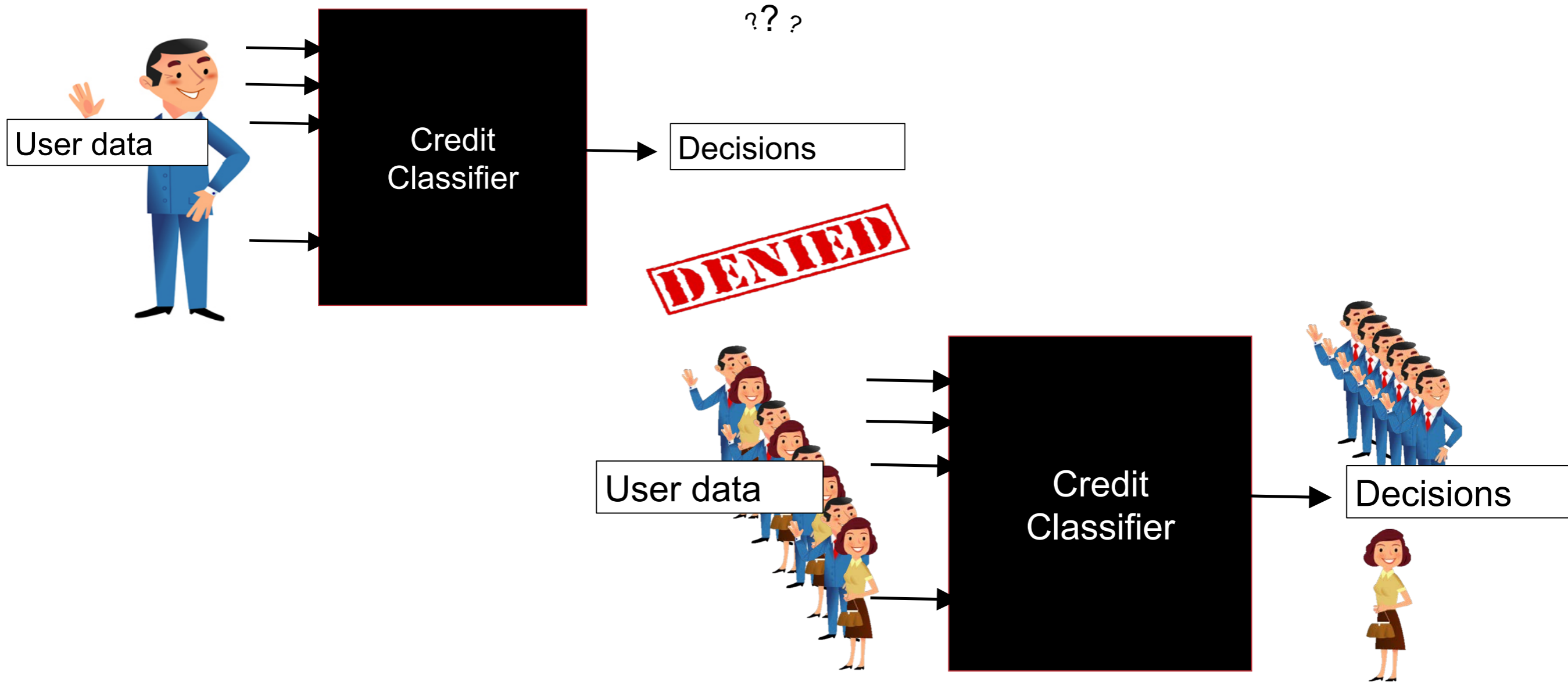


Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

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?

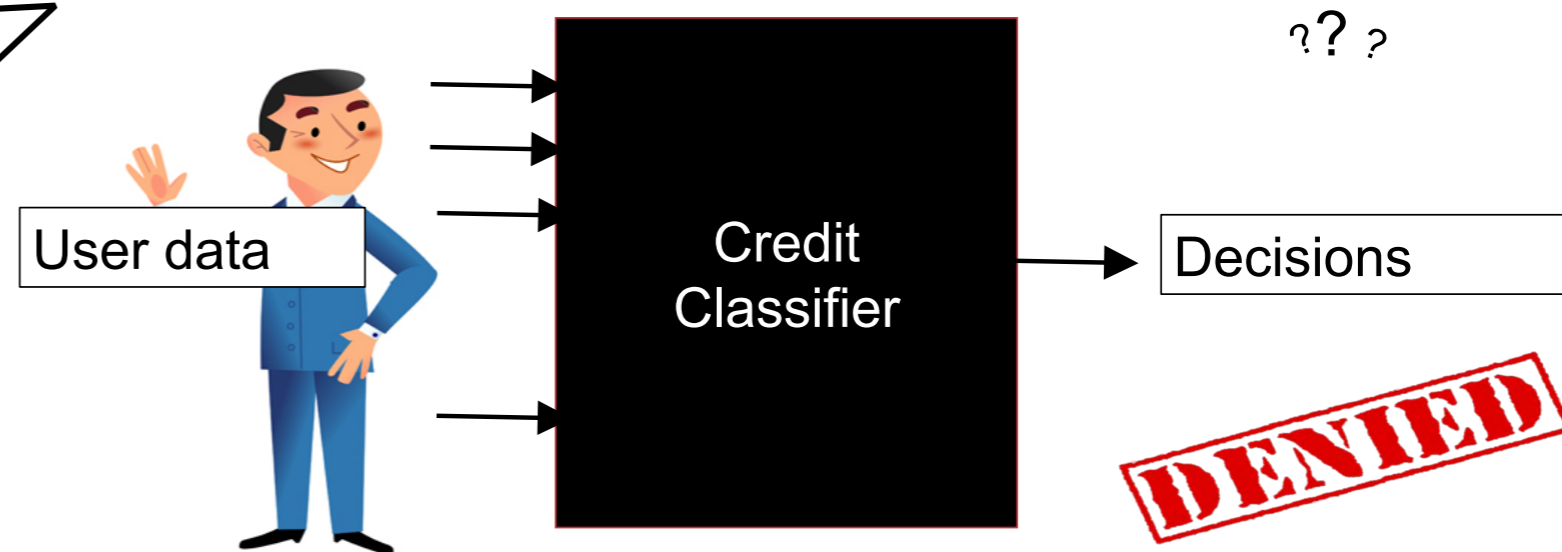
# QII: Quantitative Input Influence

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

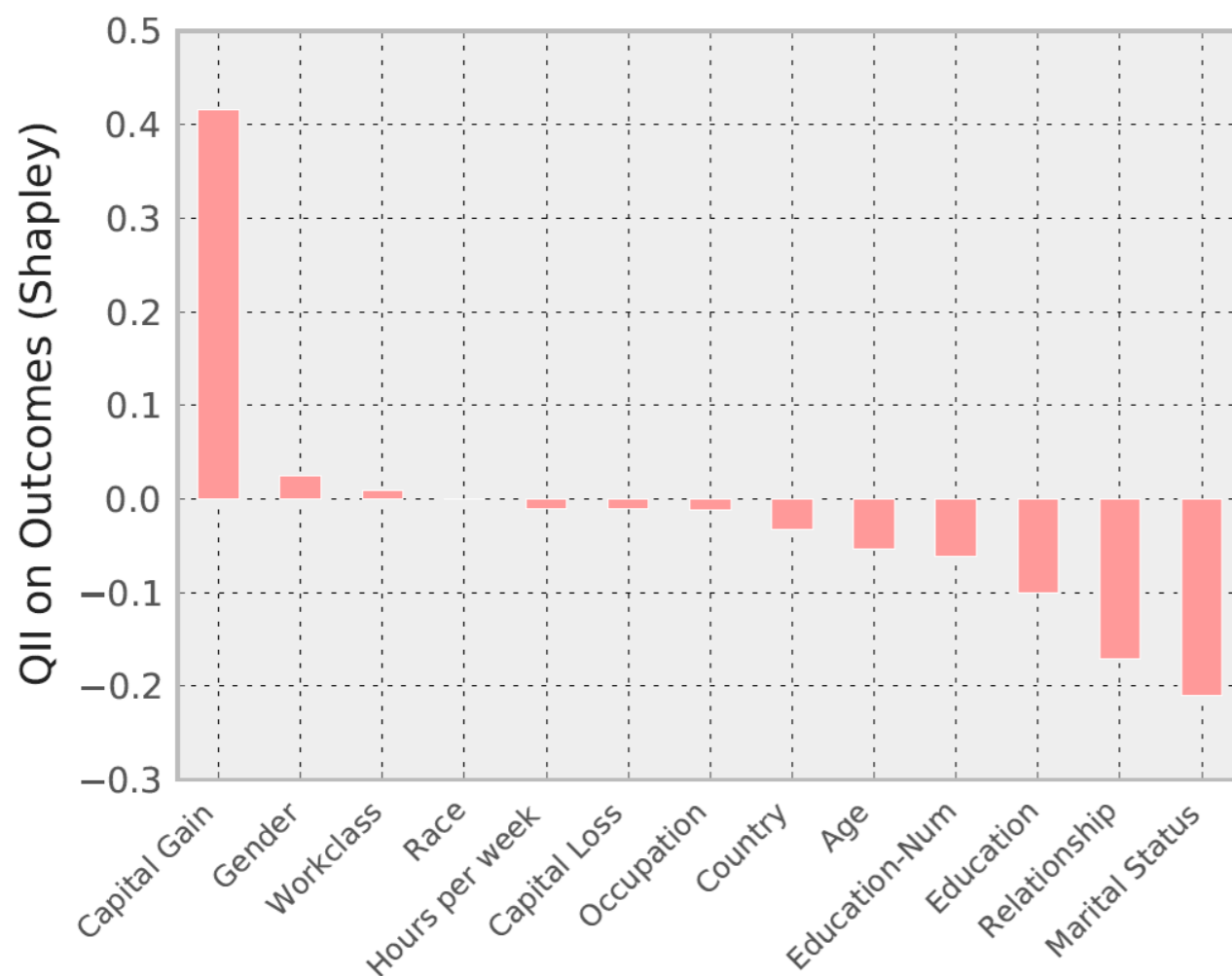
Consider lending decisions by a bank, based on gender, age, education, and income. **Does gender influence lending decisions?**

- 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 <sup>th</sup>
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

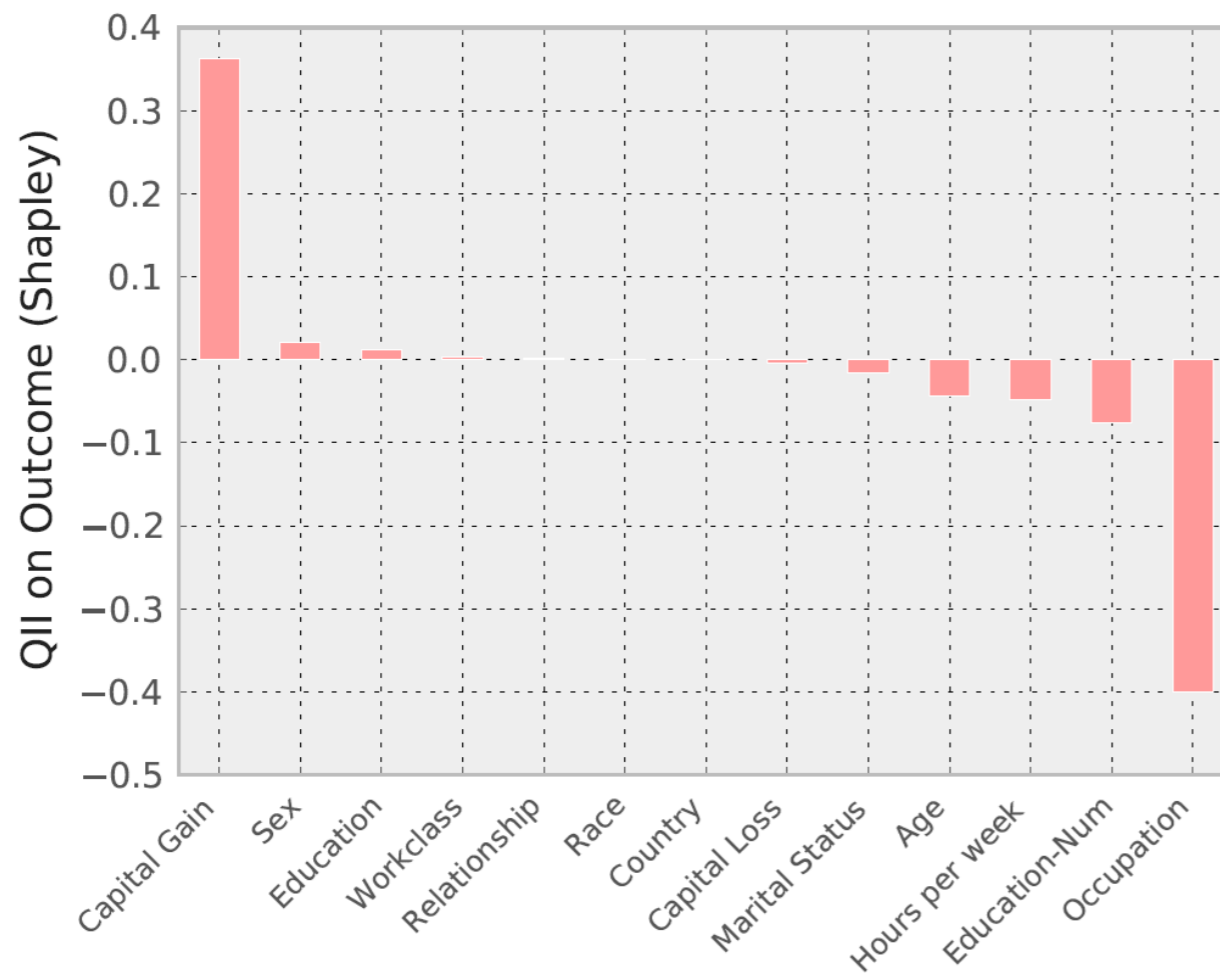
**DENIED**

income

images by Anupam Datta

# Transparency report: Mr. Y

Explanations for superficially similar individuals can be different



Age	27
Workclass	Private
Education	Preschool
Marital Status	Married
Occupation	Farming-Fishing
Relationship to household income	Other Relative
Race	White
Gender	Male
Capital gain	\$41310
Capital loss	\$0
Work hours per week	24
Country	Mexico

**DENIED**

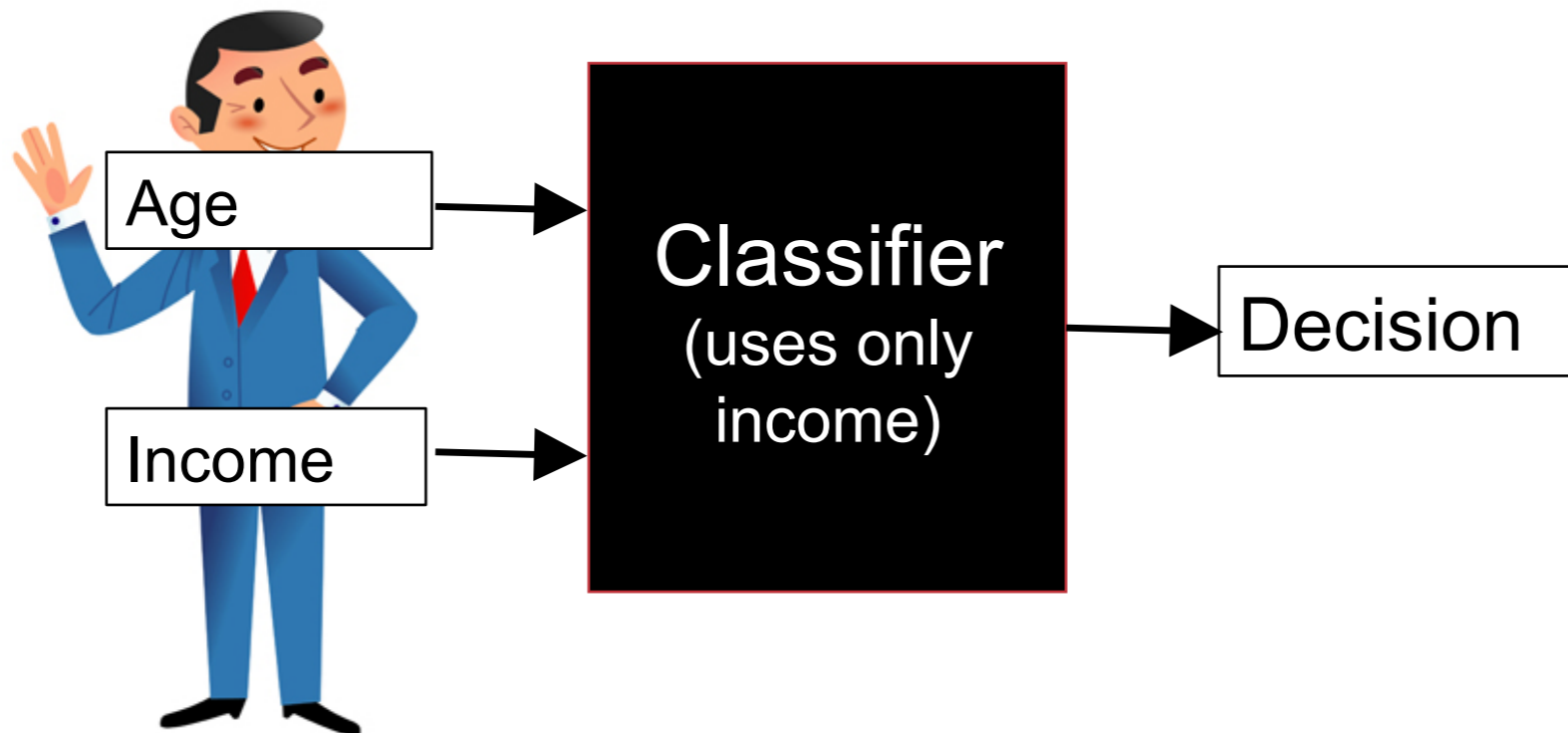
income

images by Anupam Datta

# Unary QII

images by Anupam Datta

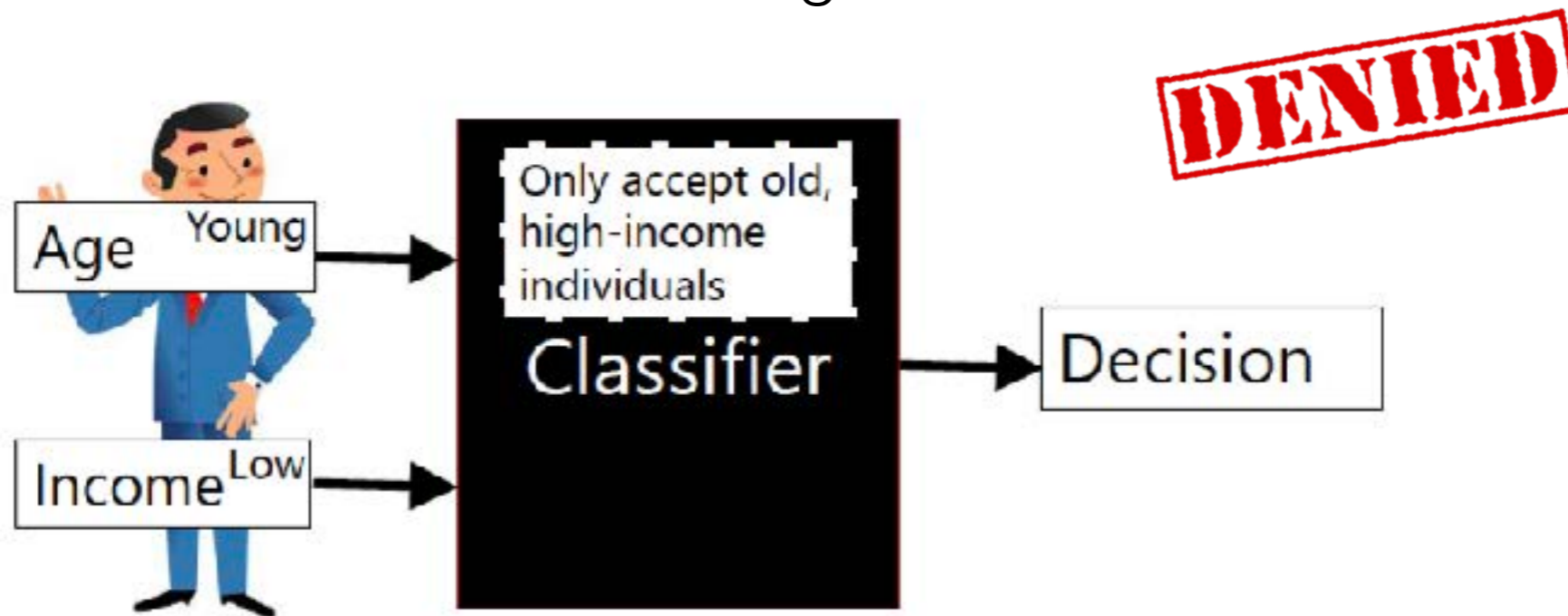
For a quantity of influence  $Q$  and an input feature  $i$ , the QII of  $i$  on  $Q$  is the difference in  $Q$  when  $i$  is changed via an **intervention**.



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.

$$I(\{\text{age, income}\}) - I(\{\text{income}\})$$

Need to aggregate Marginal QII across all sets

- But age is a part of many sets!

$$\begin{array}{l} I(\{\text{age}\}) - I(\{\}) \\ I(\{\text{age, job}\}) - I(\{\text{job}\}) \\ I(\{\text{age, gender, income}\}) - I(\{\text{gender, income}\}) \\ I(\{\text{age, gender, job}\}) - I(\{\text{gender, job}\}) \\ I(\{\text{age, gender, income, job}\}) - I(\{\text{gender, income, job}\}) \\ I(\{\text{age, gender, job}\}) - I(\{\text{gender, job}\}) \\ I(\{\text{age, gender, job}\}) - I(\{\text{gender, job}\}) \\ I(\{\text{age, gender, job}\}) - I(\{\text{gender, job}\}) \end{array}$$

# 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

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

- $\nu: 2^N \rightarrow \mathbb{R}$  influence of a set of features  $\mathbf{S}$  on the outcome
- $\varphi_i(N, \nu)$  influence of feature  $i$ , given the set of features  $\mathbf{N} = \{1, \dots, n\}$
- $\sigma \in \Pi(N)$  a permutation over the features in set  $\mathbf{N}$
- $m_i(\sigma)$  payoff corresponding to this permutation

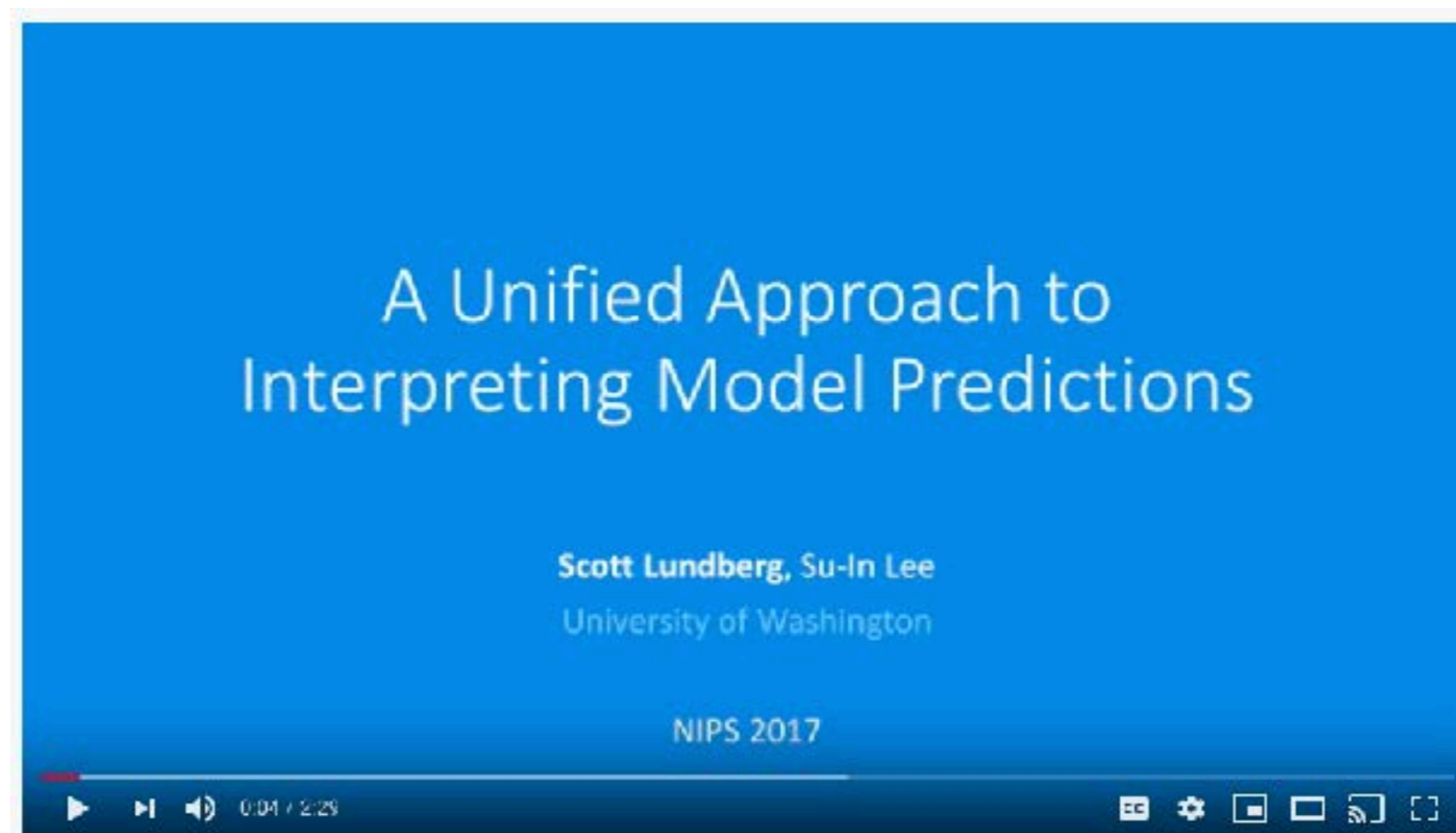
# QII summary

- 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](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)

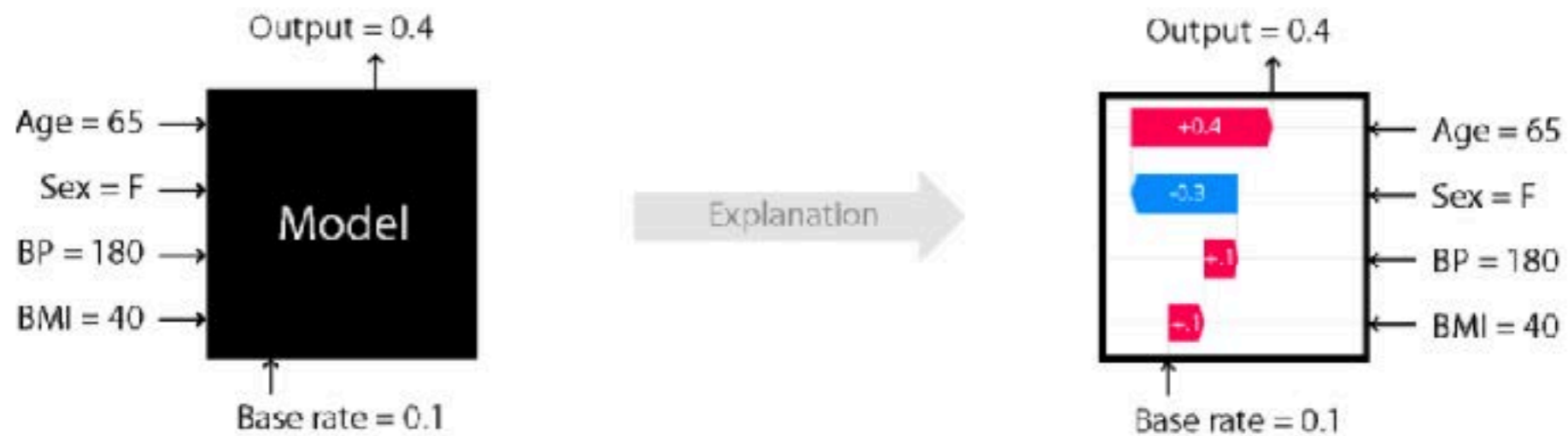
$$g(x') = \phi_0 + \sum_{i=1}^{d'} \phi_i x'_i \quad \text{where } x' \in \{0,1\}^{d'}, \text{ and } \phi_i \in \mathbb{R}$$

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  $i^{\text{th}}$  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



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