Responsible Data Science

Transparency & Interpretability

Auditing black-box models

March 10, 2024

Lucas Rosenblatt

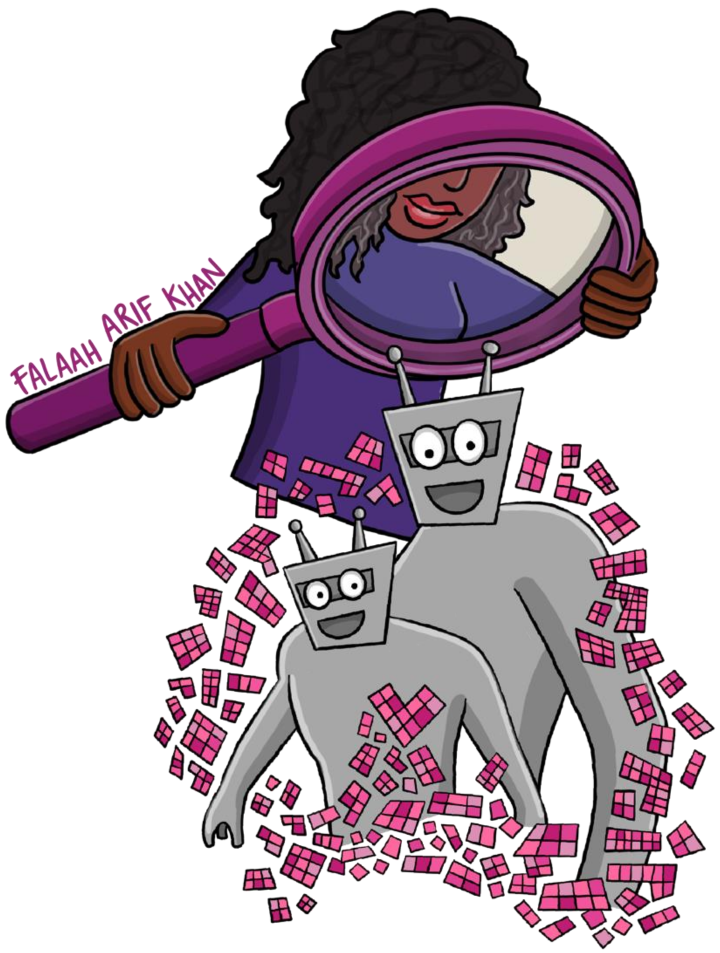
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

Terminology & vision



transparency, interpretability,

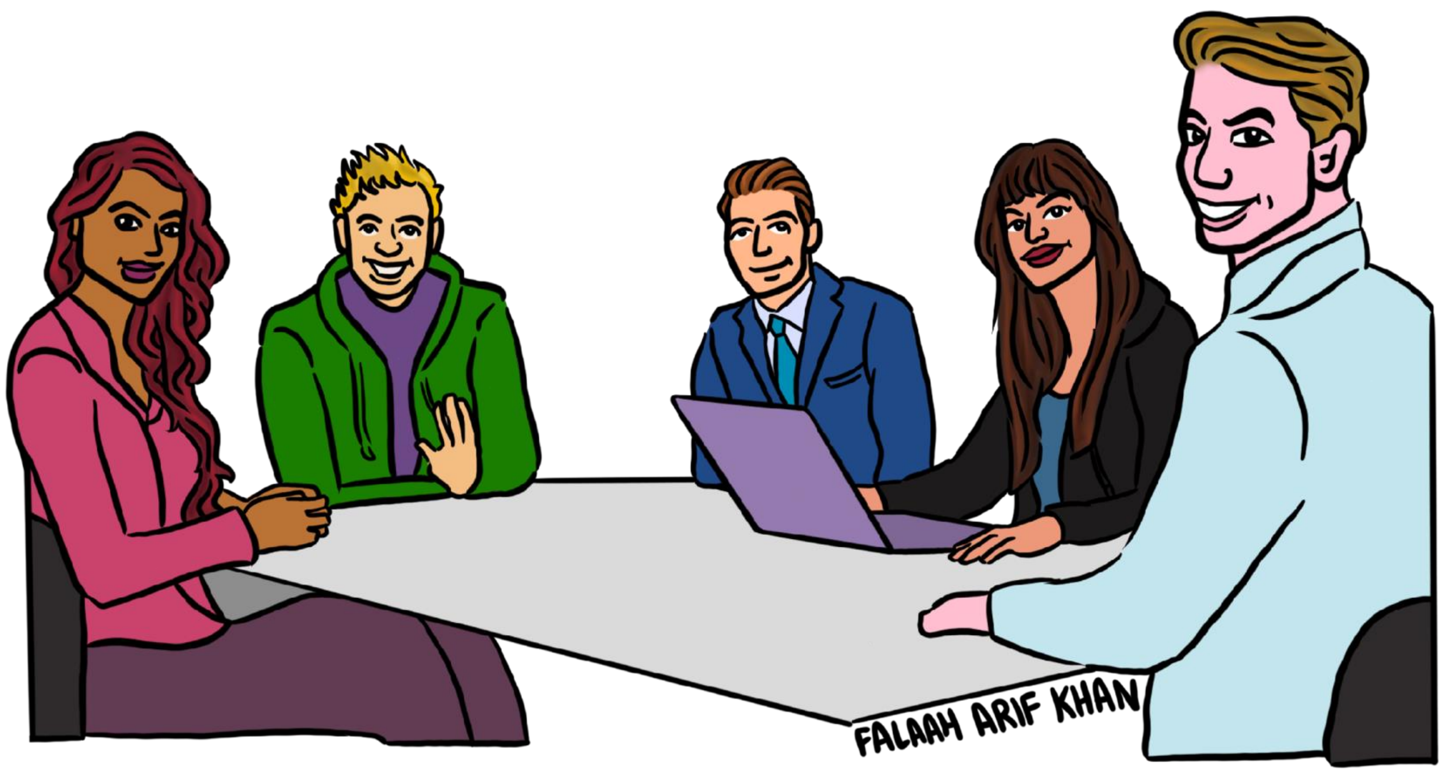
explainability, intelligibility

responsible AI

agency, responsibility

r/ai

Interpretability for different stakeholders



What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

Staples discounts

THE WALL STREET JOURNAL.

December 2012

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani
December 24, 2012

WHAT PRICE WOULD YOU SEE?

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

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

A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, **Staples appeared to consider the person's distance from a rival brick-and-mortar store**, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

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

Staples discounts

THE WALL STREET JOURNAL.

December 2012

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani
December 24, 2012

WHAT PRICE WOULD YOU SEE?

It was the same Sw

for Kim Wamble, th

screen, just a few

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

A key difference: where Staples seemed to think they were lo

A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, **Staples appeared to consider the person's distance from a rival brick-and-mortar store**, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

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

r/ai

Online job ads

theguardian

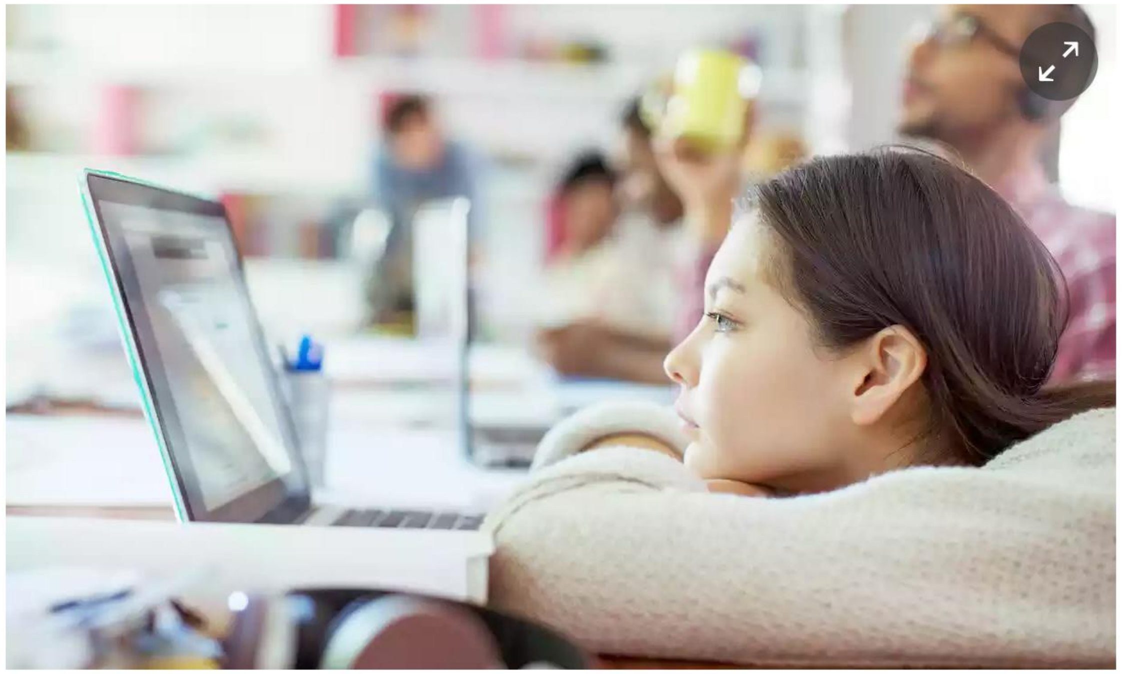
July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

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

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



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

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

One experiment showed that Google displayed ads for a career coaching service for "$200k+" executive jobs **1,852 times to the male group and only 318 times to the female group**. Another experiment, in July 2014, showed a similar trend but was not statistically significant.

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

r/ai

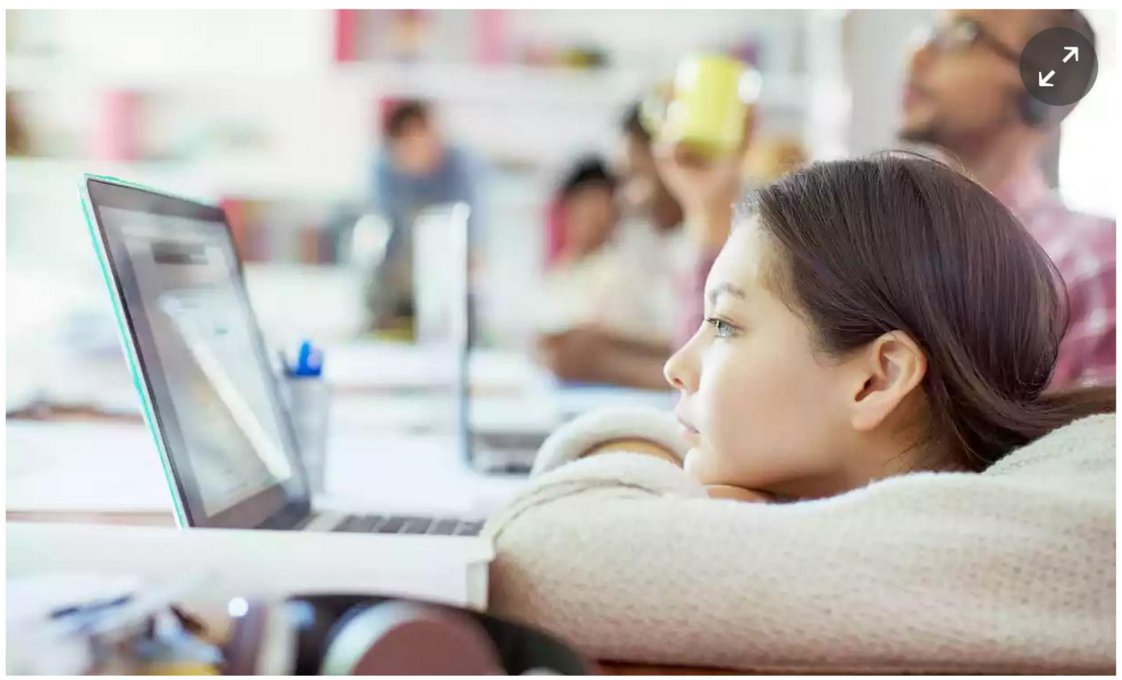
Online job ads

theguardian

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
high-pa

The AdFis
not differ
demogra

One expe
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.

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?

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

r/ai

Instant Checkmate

Google
AdSense

February

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/

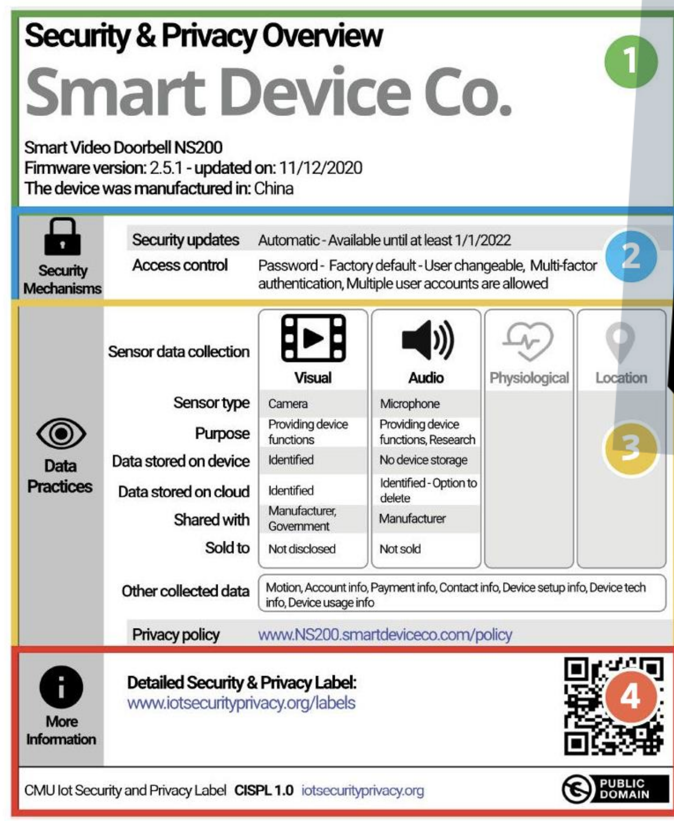
r/ai

Nutritional labels

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-not-1501758003

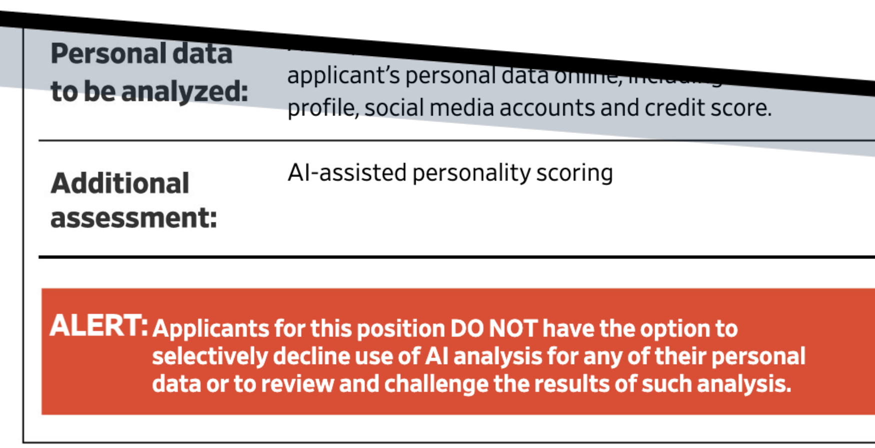


https://www.wsj.com/articles/imagine-a-nutrition-labelfor-cybersecurity-11607436000

What are we explaining?

To **Whom** are we explaining?

Why are we explaining?



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

r/ai

explaining black box models

This week's reading

QII

2016 IEEE Symposium on Security and Privacy

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

Anupam Datta Shayak Sen Yair Zick
Carnegie Mellon University, Pittsburgh, USA
{danupam, shayaks, yairzick}@cmu.edu

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

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

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

I. Introduction

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

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

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

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

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

© 2016, Anupam Datta. Under license to IEEE.
DOI 10.1109/SP.2016.42

LIME

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

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

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

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

ABSTRACT

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

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

1. INTRODUCTION

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

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

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

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

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

2. THE CASE FOR EXPLANATIONS

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use 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. Copyrights 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, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD 2016 San Francisco, CA, USA
© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939778

This week's reading

SHAP

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

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

1 Introduction

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

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

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

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

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

LevSHAP

PROVABLY ACCURATE SHAPLEY VALUE ESTIMATION VIA LEVERAGE SCORE SAMPLING

Christopher Musco & R. Teal Witter
New York University
{cmusco,rtealwitter}@nyu.edu

ABSTRACT

Originally introduced in game theory, Shapley values have emerged as a central tool in explainable machine learning, where they are used to attribute model predictions to specific input features. However, computing Shapley values exactly is expensive: for a general model with n features, $O(2^n)$ model evaluations are necessary. To address this issue, approximation algorithms are widely used. One of the most popular is the Kernel SHAP algorithm, which is model agnostic and remarkably effective in practice. However, to the best of our knowledge, Kernel SHAP has no strong non-asymptotic complexity guarantees. We address this issue by introducing *Leverage SHAP*, a light-weight modification of Kernel SHAP that provides provably accurate Shapley value estimates with just $O(n \log n)$ model evaluations. Our approach takes advantage of a connection between Shapley value estimation and agnostic active learning by employing *leverage score sampling*, a powerful regression tool. Beyond theoretical guarantees, we show that Leverage SHAP consistently outperforms even the highly optimized implementation of Kernel SHAP available in the ubiquitous SHAP library [Lundberg & Lee, 2017].

1 INTRODUCTION

While AI is increasingly deployed in high-stakes domains like education, healthcare, finance, and law, increasingly complicated models often make predictions or decisions in an opaque and uninterpretable way. In high-stakes domains, transparency in a model is crucial for building trust. Moreover, for researchers and developers, understanding model behavior is important for identifying areas of improvement and applying appropriate safe guards. To address these challenges, Shapley values have emerged as a powerful game-theoretic approach for explaining even opaque models (Shapley, 1951; Štrumbelj & Kononenko, 2014; Datta et al., 2016; Lundberg & Lee, 2017). These values can be used to effectively quantify the contribution of each input feature to a model's output, offering at least a partial, principled explanation for why a model made a certain prediction.

Concretely, Shapley values originate from game-theory as a method for determining fair 'payouts' for a cooperative game involving n players. The goal is to assign higher payouts to players who contributed more to the cooperative effort. Shapley values quantify the contribution of a player by measuring how its addition to a set of other players changes the value of the game. Formally, let the *value function* $v : 2^{[n]} \to \mathbb{R}$ be a function defined on sets $S \subseteq [n]$. The Shapley value for player i is:

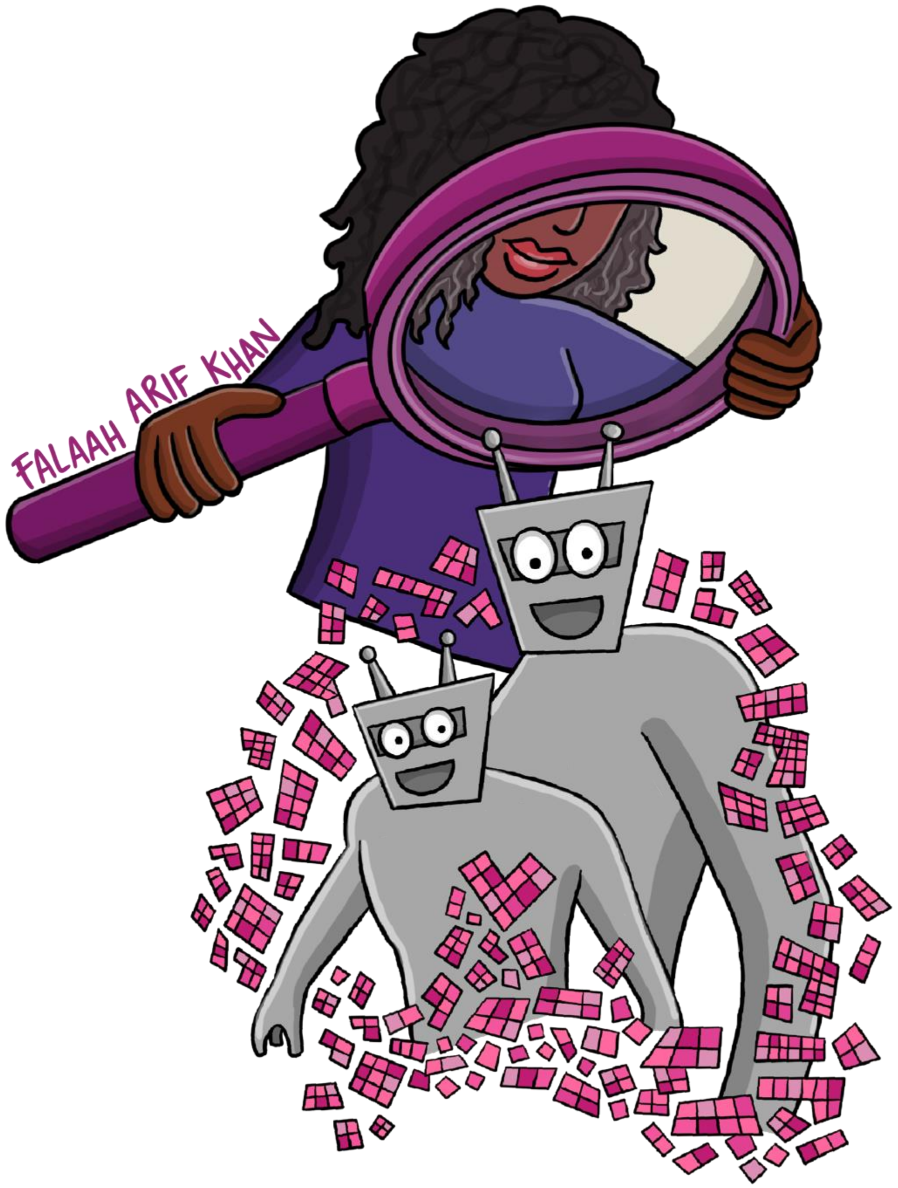
$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}. \quad (1)$$

The denominator weights the marginal contribution of player i to set S by the number of sets of size $|S|$, so that the marginal contribution to sets of each size are equally considered. With this weighting, Shapley values are known to be the unique values that satisfy four desirable game-theoretic properties: Null Player, Symmetry, Additivity, and Efficiency (Shapley, 1951). For further details on Shapley values and their theoretical motivation, we refer the reader to Molnar (2024).

A popular way of using Shapley values for explainable AI is to attribute predictions made by a model $f : \mathbb{R}^n \to \mathbb{R}$ on a given input $\mathbf{x} \in \mathbb{R}^n$ compared to a baseline input $\mathbf{y} \in \mathbb{R}^n$ (Lundberg & Lee, 2017). The players are the features and $v(S)$ is the prediction of the model when using the features

r/ai

What are we explaining?



How does a system work?

How **well** does a system work?

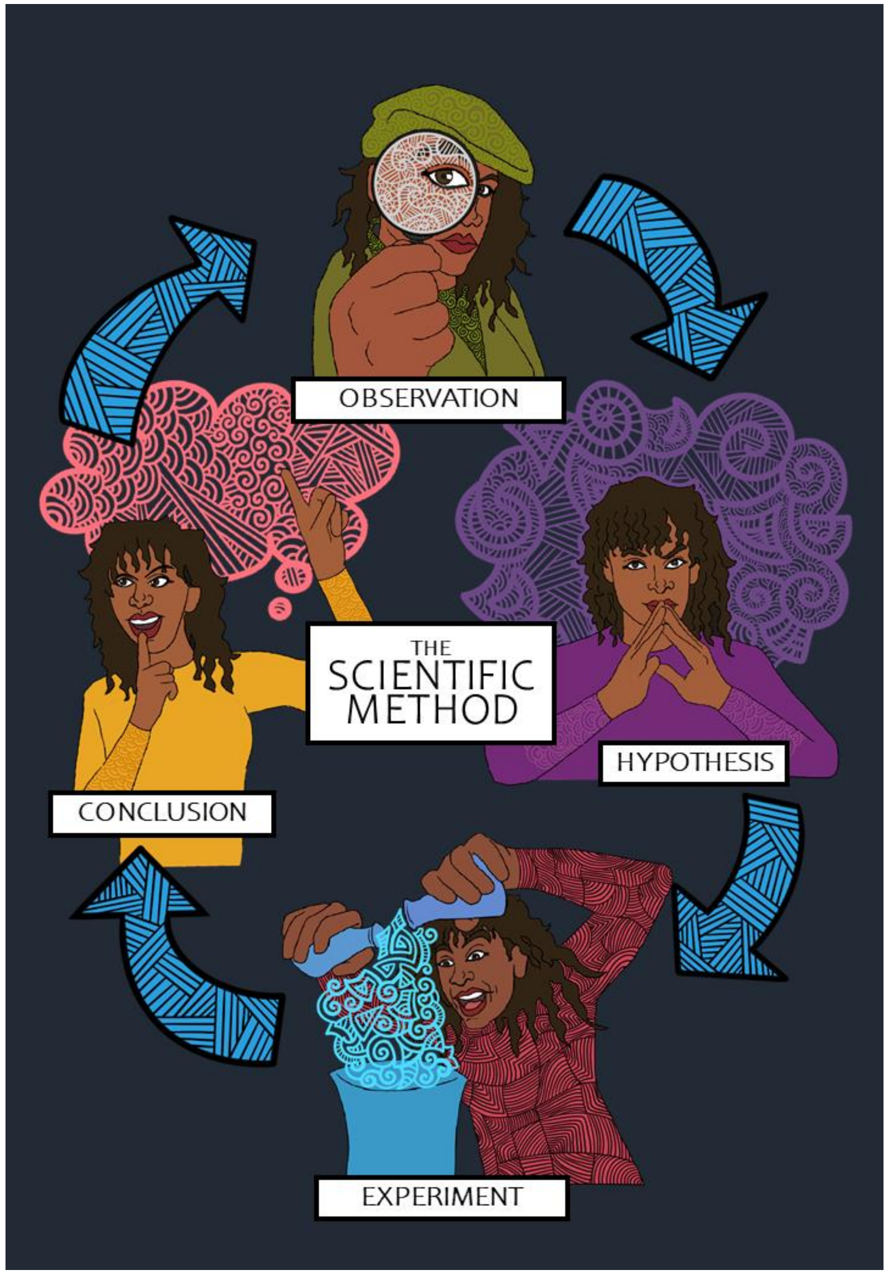
What does a system do?

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

Shane Creepingbear is a member of the Kiowa Tribe of Oklahoma

October 13, 2014

← **Tweet**

Shane Creepingbear @Creepingbear · Oct 13, 2014
Hey yall today I was kicked off of Facebook for having a fake name. Happy Columbus Day great job #facebook #goodtiming #racist #ColumbusDay

17

TIME

Facebook Thinks Some Native American Names Are Inauthentic

BY JOSH SANBURN FEBRUARY 14, 2015

February 14, 2015

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

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

The latest flap involves Native Americans, including Dana Lone Hill, who is Lakota. Lone Hill recently wrote in a blog post that Facebook told her her name was not "authentic" when she attempted to log in.

When accuracy is not enough

Explaining Google's Inception NN

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

P() = 0.32

Electric guitar - incorrect but reasonable, similar fretboard

P() = 0.24

Acoustic guitar

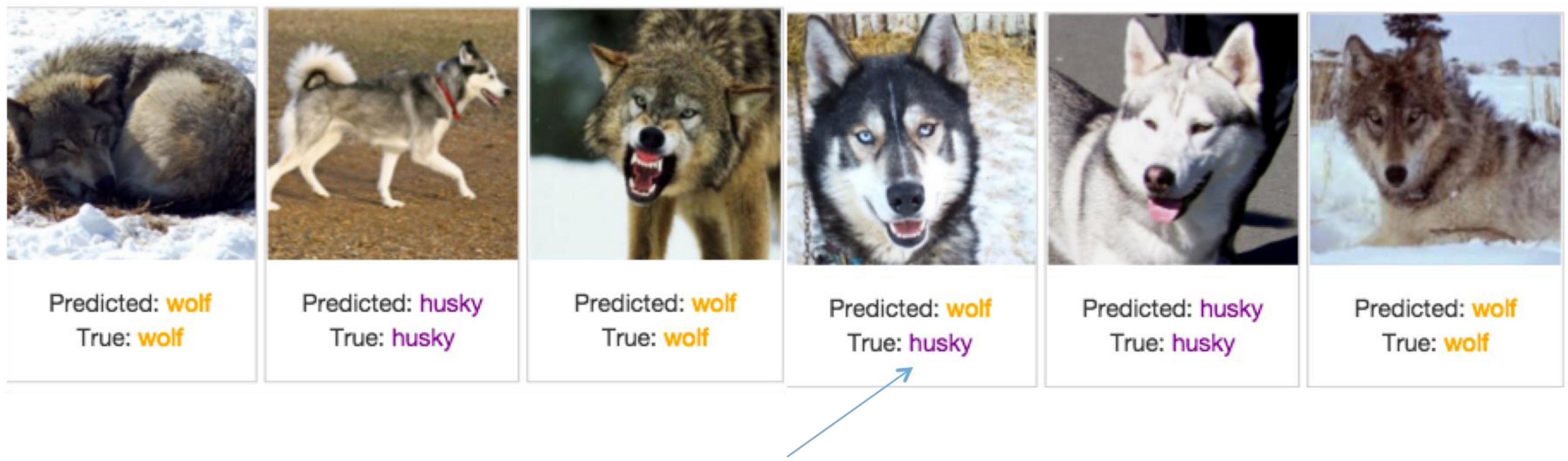
P() = 0.21

Labrador

[Ribeiro, Singh & Guestrin, 2016]

When accuracy is not enough

Train a neural network to predict wolf v. husky



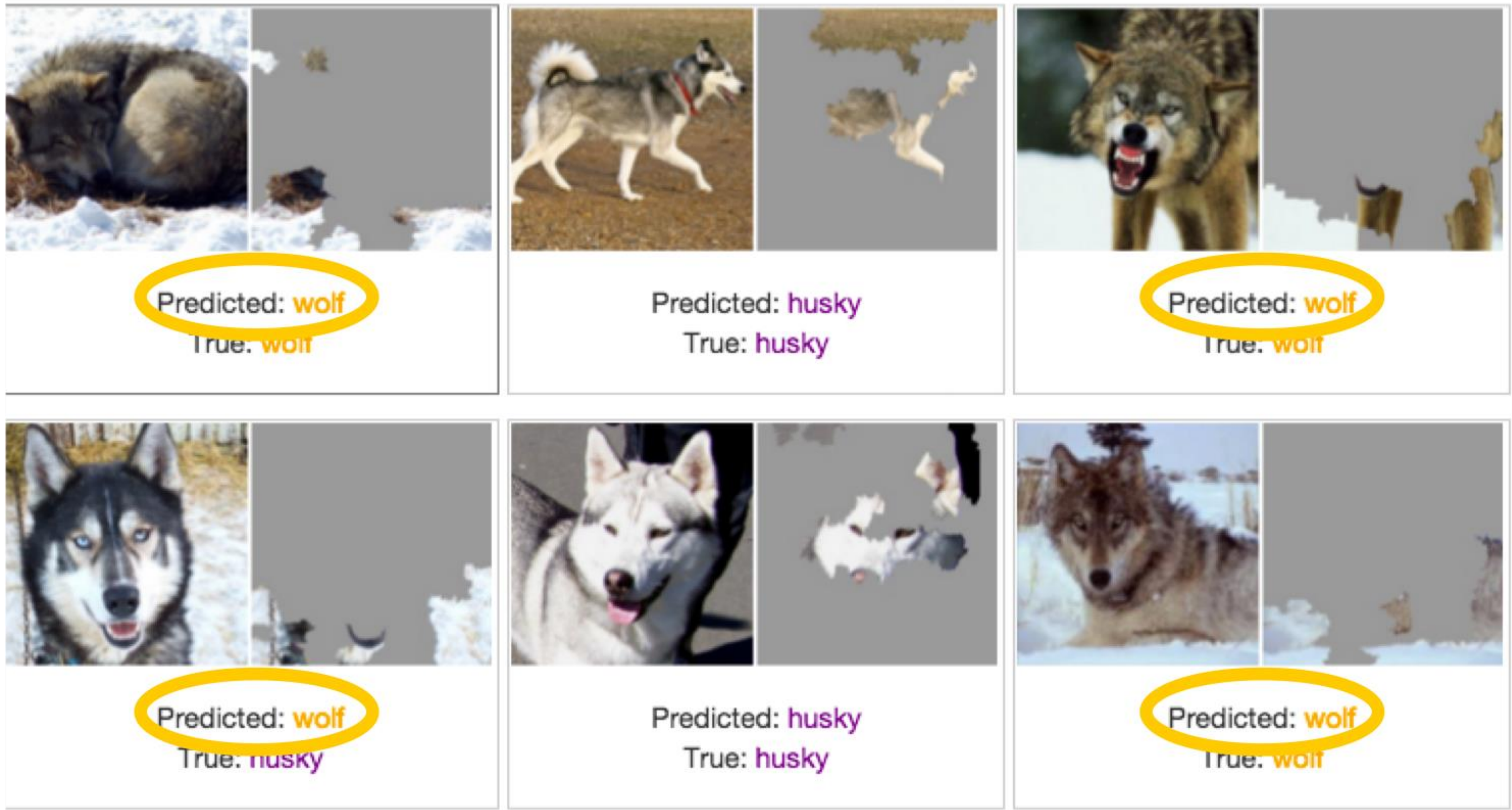
Only 1 mistake!!!

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

slide by Marco Tulio Ribeiro, KDD 2016

When accuracy is not enough

Explanations for neural network prediction



We’ve built a great snow detector… ☹

slide by Marco Tulio Ribeiro, KDD 2016

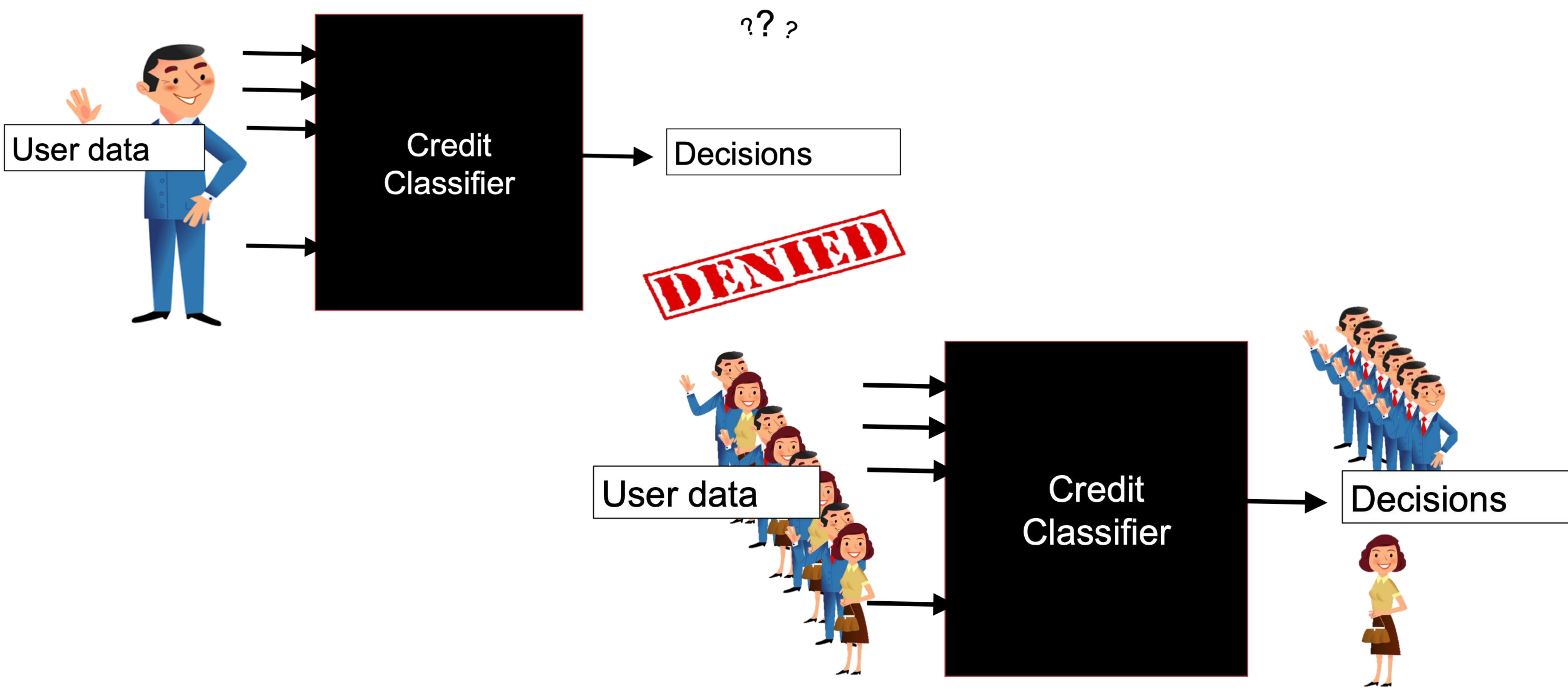
[Ribeiro, Singh & Guestrin, 2016]

LIME: Recap

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

r/ai

QII: Auditing black-box models



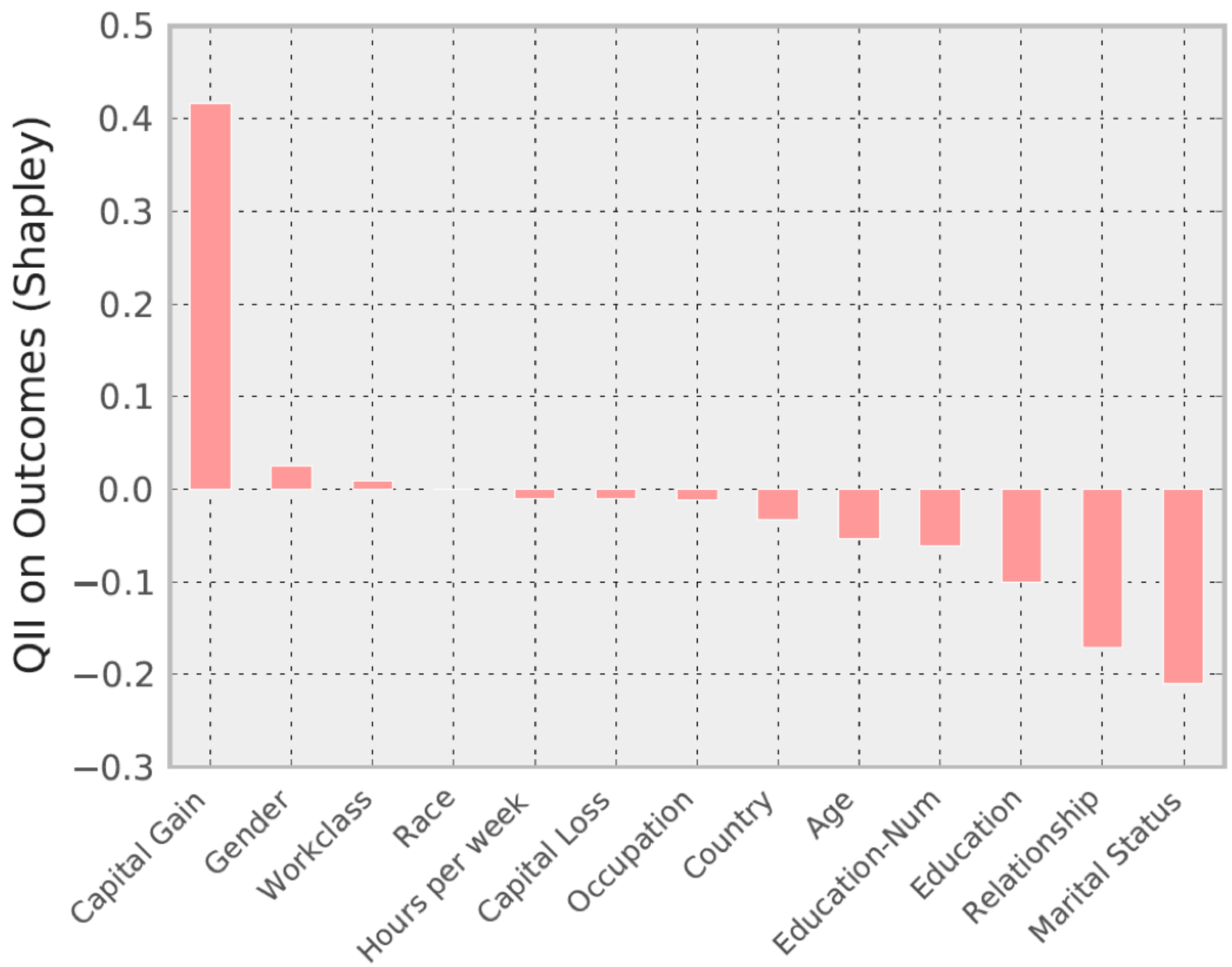
images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

Transparency report: Mr. X

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



DENIED

Age	23
Workclass	Private
Education	11th
Marital Status	Never married
Occupation	Craft repair
Relationship to household income	Child
Race	Asian-Pac Island
Gender	Male
Capital gain	$14344
Capital loss	$0
Work hours per week	40
Country	Vietnam

income

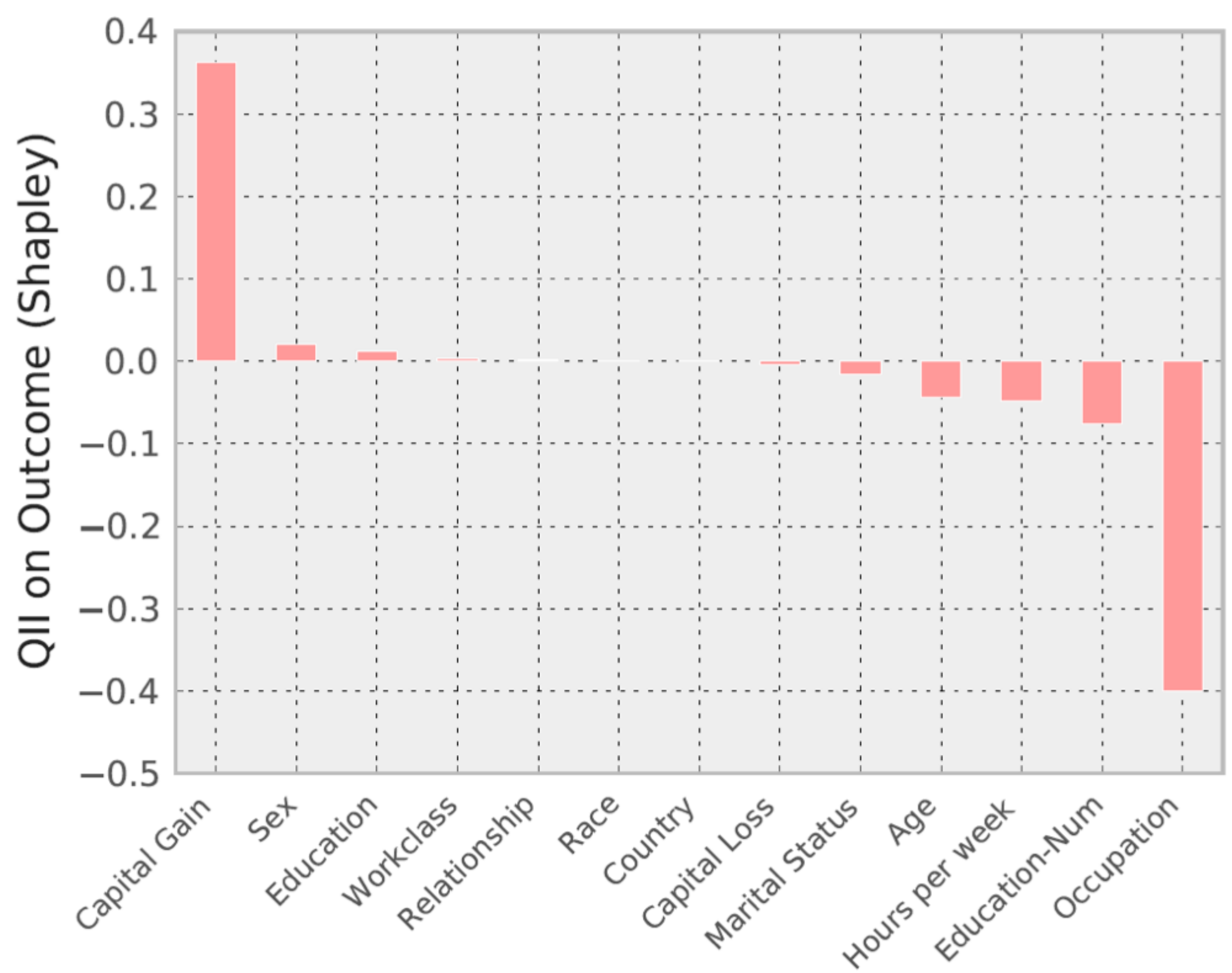
images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

Transparency report: Mr. Y

Explanations for superficially similar individuals can be different



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

income

images by Anupam Datta

[Datta, Sen & Zick, 2016]

r/ai

QII: Quantitative Input Influence

Goal: determine how much influence an input, or a set of inputs, has on a **classification outcome** for an individual or a group

Transparency queries / quantities of interest

Individual: Which inputs have the most influence in my credit denial?

Group: Which inputs have the most influence on credit decisions for women?

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

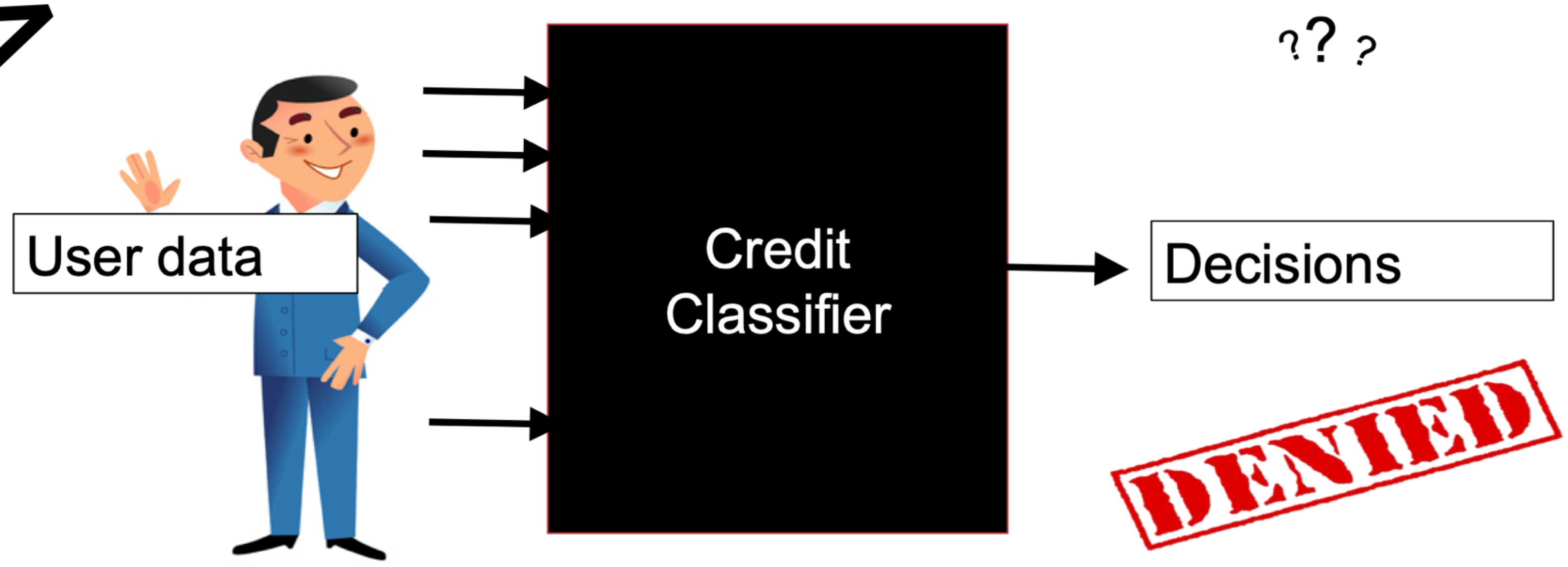
QII: Quantitative Input Influence

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

Key ideas

intervene on an input feature, measure its **importance**

aggregate feature importance using its **Shapley value**



images by Anupam Datta

[Datta, Sen & Zick, 2016]

Running example

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

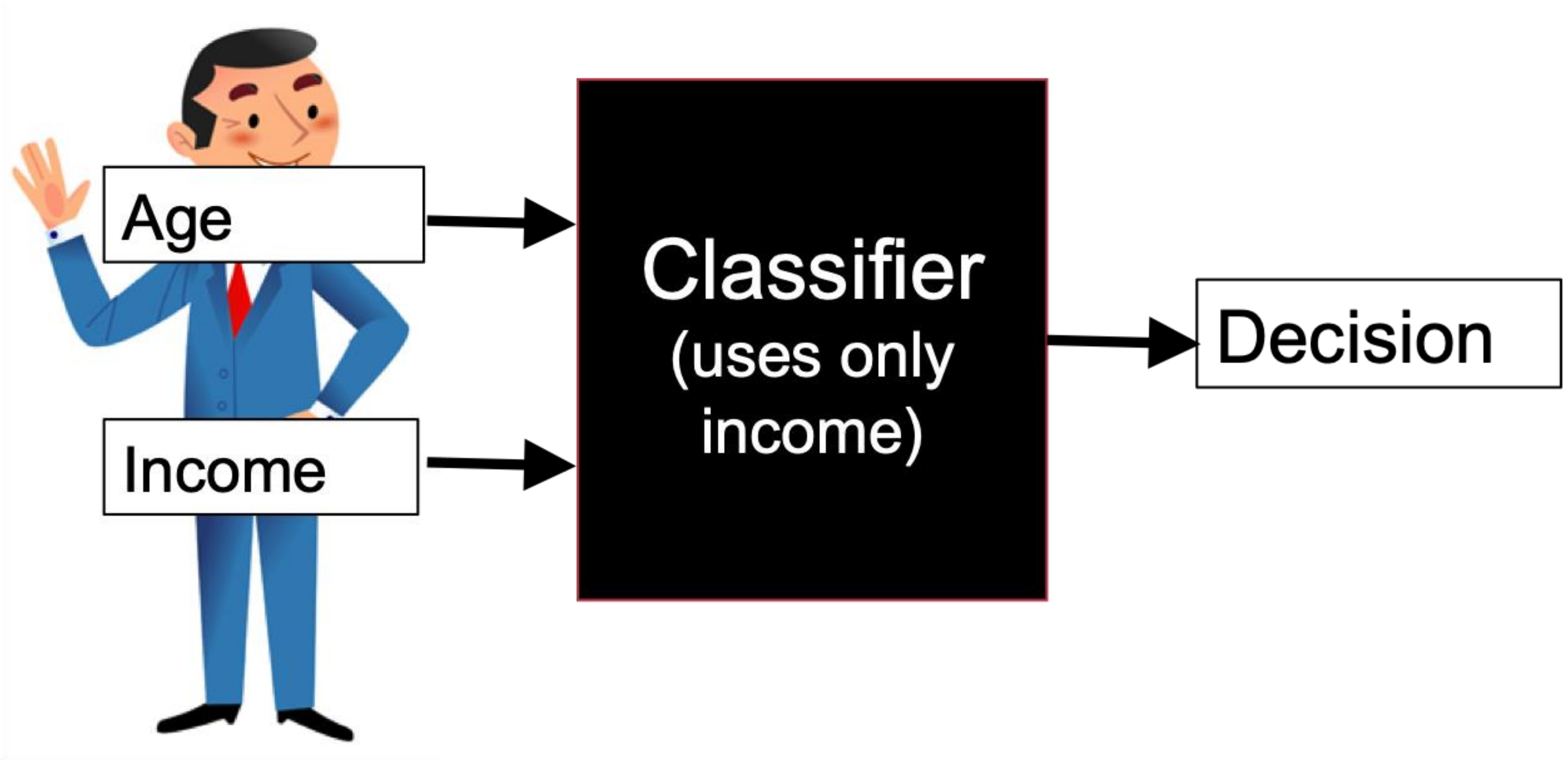
- Observe that 20% of women receive the positive classification.
- To check whether gender impacts decisions, take the input dataset and replace the value of gender in each input profile by drawing it from the uniform distribution: set gender in 50% of the inputs to female and 50% to male.
- If we still observe that 20% of female profiles are positively classified **after the intervention** - we conclude that gender does not influence lending decisions.
- Do a similar test for other features, one at a time. This is known as **Unary QII**

[Datta, Sen & Zick, 2016]

Unary QII

images by Anupam Datta

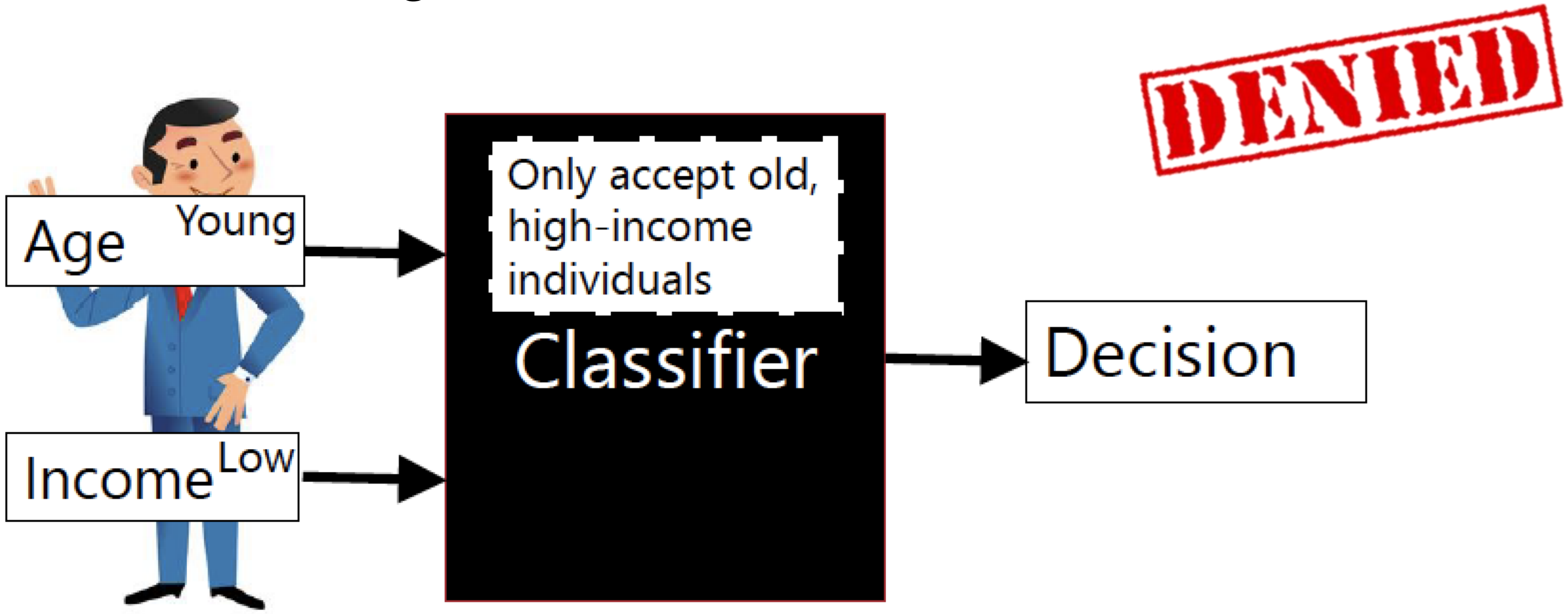
For a quantity of influence *Q* and an input feature *i*, the QII of *i* on *Q* is the difference in *Q* when *i* is changed via an **intervention**.



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

Unary QII

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



intervening on one feature at a time will not have any effect

images by Anupam Datta

Marginal QII

- Not all features are equally important within a set.

- *Marginal QII*: Influence of age and income over only income.

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

Need to aggregate Marginal QII across all sets

- But age is a part of many sets!

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

[Datta, Sen & Zick, 2016]

Aggregating influence across sets

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

"In voting systems with multiple agents with differing weights, voting power often does not directly correspond to the weights of the agents. For example, the US presidential election can roughly be modeled as a cooperative game where each state is an agent. The ***weight of a state is the number of electors in that state*** *(i.e., the number of votes it brings to the presidential candidate who wins that state). Although states like California and Texas have higher weight, swing states like Pennsylvania and Ohio tend to have higher power in determining the outcome of elections."*

[Datta, Sen & Zick, 2016]

QII summary

- A principled (and beautiful!) framework for determining the influence of a feature, or a set of features, on a decision
- Works for black-box models, with the assumption that the full set of inputs is available
- Accounts for correlations between features
- “Parametrizes” on what quantity we want to set (QII), how we intervene, how we aggregate the influence of a feature across sets
- Experiments in the paper: interesting results
- Also in the paper: a discussion of **transparency under differential privacy**

[Datta, Sen & Zick, 2016]

https://bit.ly/3DrCGIm

calculating SHAP values

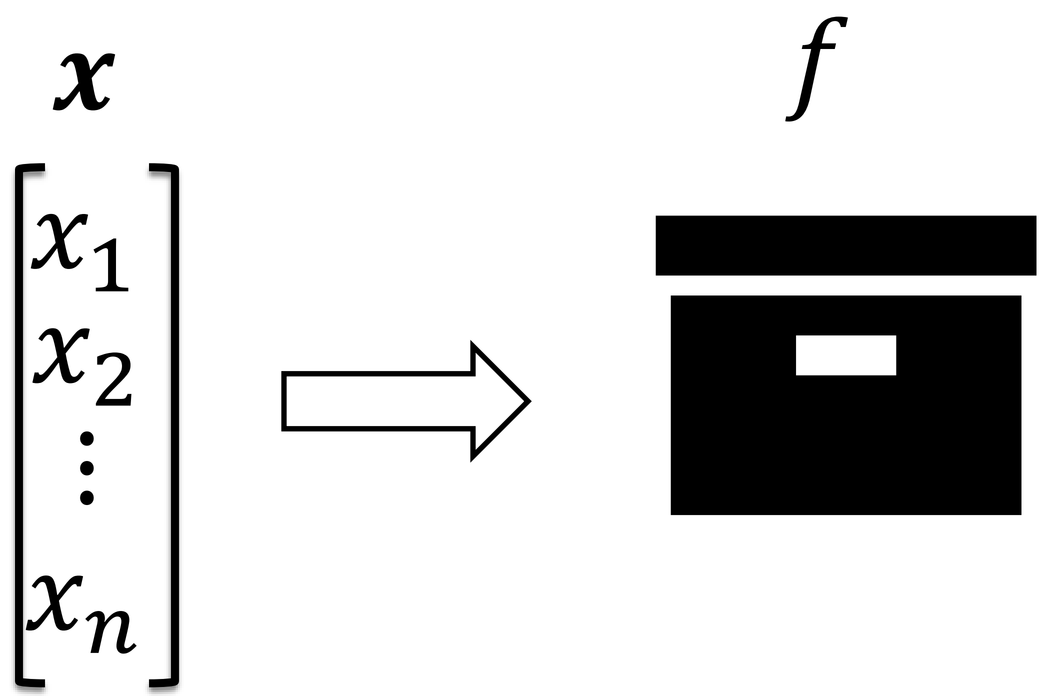
AI Prediction

$$\boldsymbol{x}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

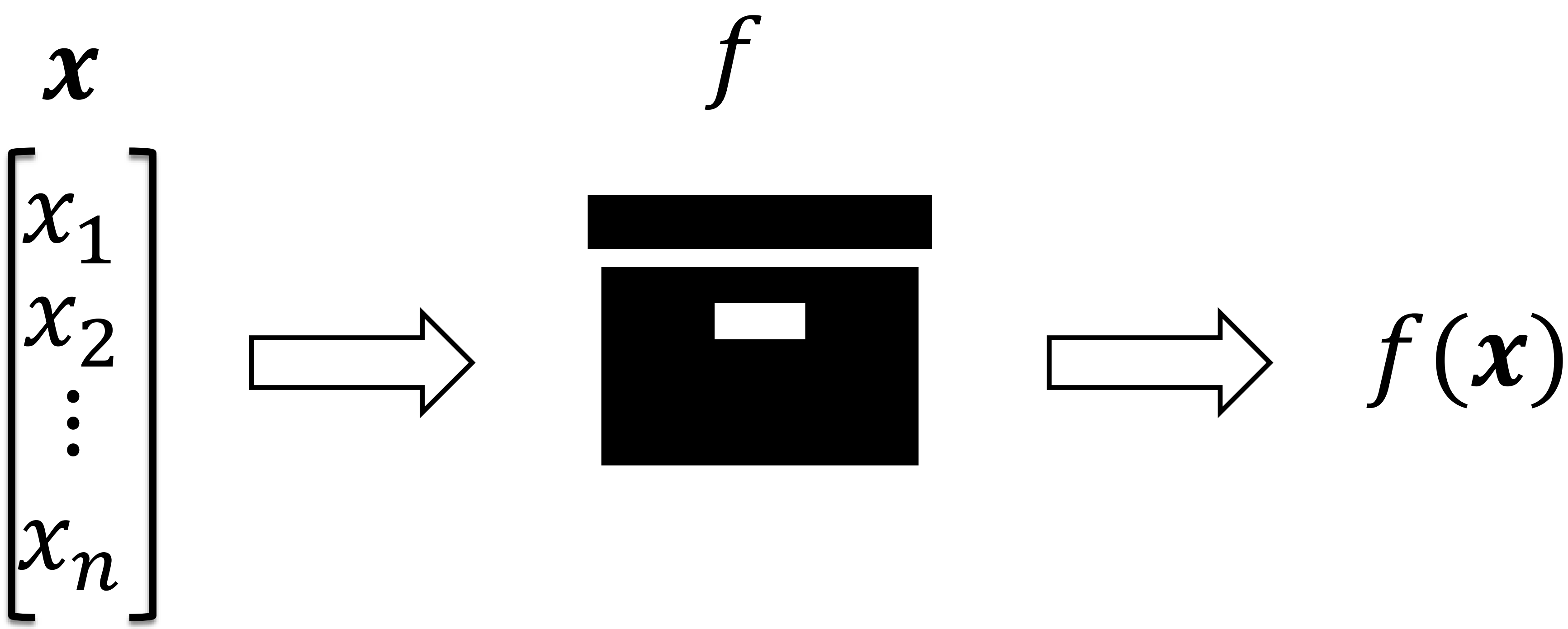
Slide Credit: R. Teal Witter

AI Prediction



Slide Credit: R. Teal Witter

AI Prediction



Slide Credit: R. Teal Witter

Example:

(biking enjoyment)

$\boldsymbol{x}$

Temperature

Wind speed

Chance of rain

Helicopters

Traffic

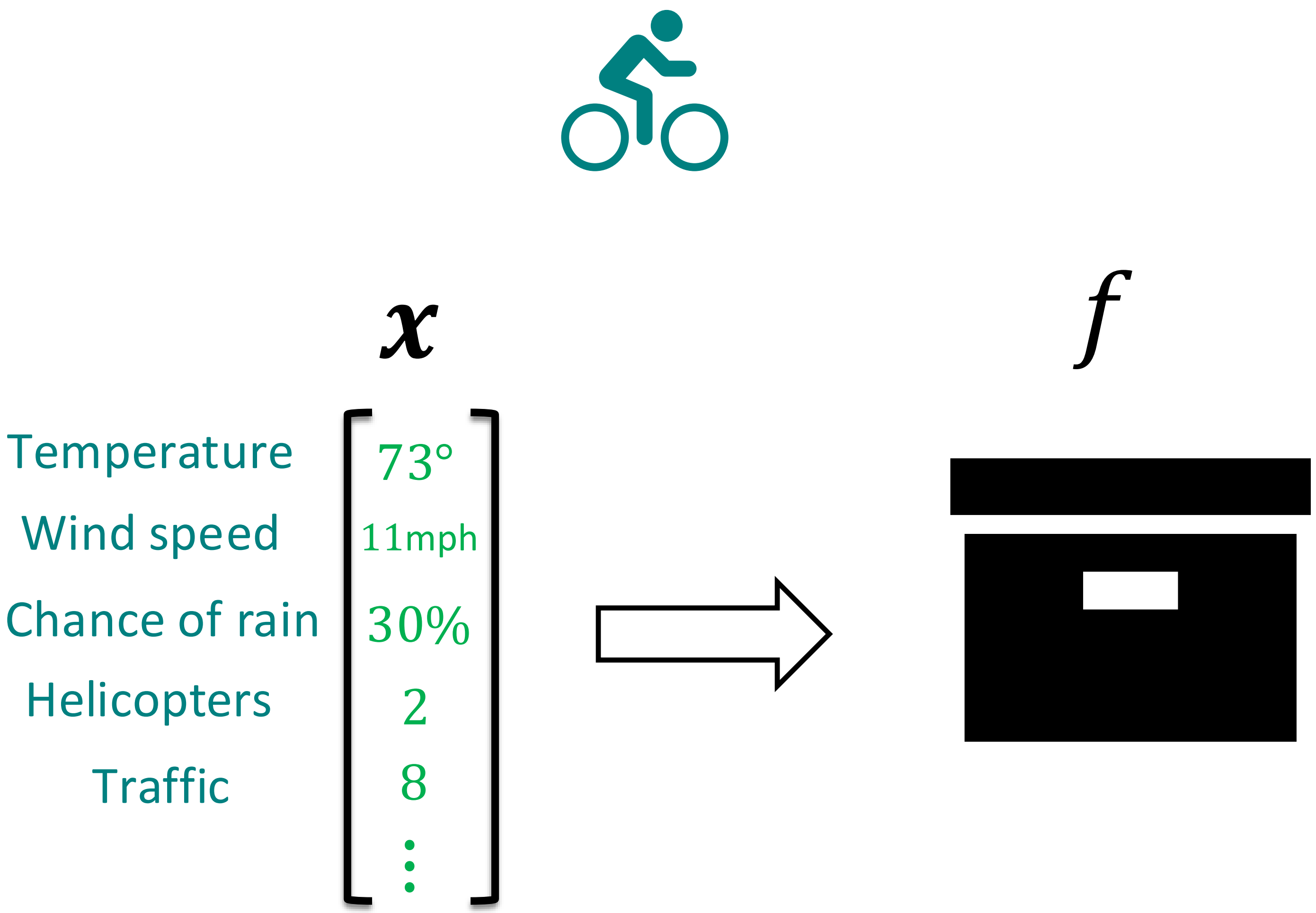
Slide Credit: R. Teal Witter

Example:

$\boldsymbol{x}$

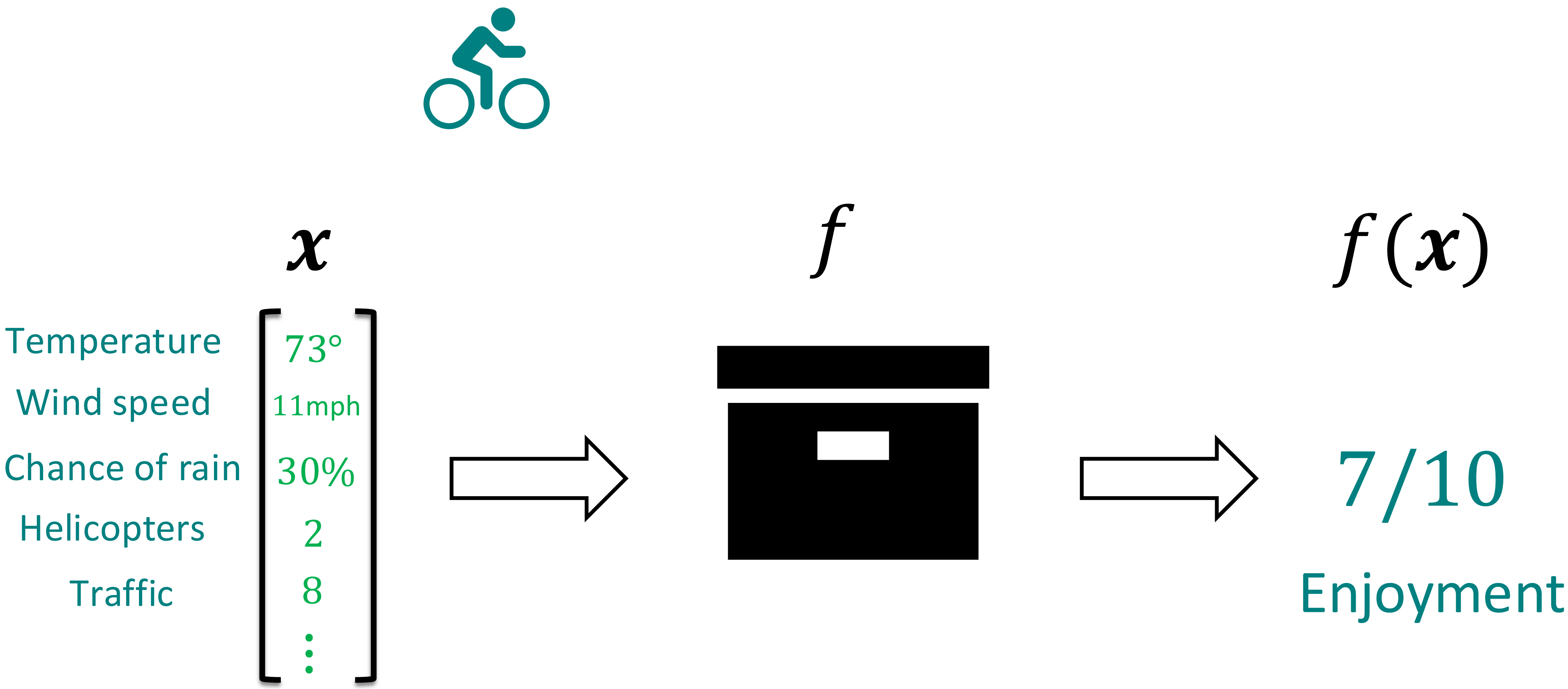
Temperature	73°
Wind speed	11mph
Chance of rain	30%
Helicopters	2
Traffic	8
	⋮

Example:



Slide Credit: R. Teal Witter

Example:



Slide Credit: R. Teal Witter

Explaining Predictions

Attribute the prediction to features relative to a baseline

"Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable."

Explaining Predictions

Attribute the prediction to features relative to a baseline

"Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable."

Attribution value!

Slide Credit: R. Teal Witter

Explaining Predictions

Attribute the prediction to features relative to a baseline

"Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable."

Temperature	73°
Wind speed	11mph
Chance of rain	30%
Helicopters	2
Traffic	8
	⋮

Explicand

Slide Credit: R. Teal Witter

Explaining Predictions

Attribute the prediction to features relative to a baseline

"Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable."

	Explicand	Baseline
Temperature	73°	89°
Wind speed	11mph	1mph
Chance of rain	30%	0%
Helicopters	2	5
Traffic	8	3
	⋮	⋮

Slide Credit: R. Teal Witter

Explaining Predictions

Attribute the prediction to features relative to a baseline

"Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable."

	Explicand	Baseline
Temperature	73°	89°
Wind speed	11mph	1mph
Chance of rain	30%	0%
Helicopters	2	5
Traffic	8	3
	⋮	⋮

					$f(x)$
89°	11mph	30%	5	3	5/10
89°	11mph	30%	5	8	4/10
73°	1mph	0%	5	3	6/10
73°	1mph	0%	5	8	8/10

Slide Credit: R. Teal Witter

Attribution Values

What is the effect of the feature in different settings?

Slide Credit: R. Teal Witter

Attribution Values

What is the effect of the feature in different settings?

Consider subsets $S \subseteq [n]$ and define $v(S) = f(\boldsymbol{x}^S)$ where

Slide Credit: R. Teal Witter

Attribution Values

What is the effect of the feature in different settings?

Consider subsets $S \subseteq [n]$ and define $v(S) = f(\boldsymbol{x}^S)$ where

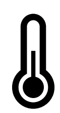
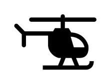
S						$f(\boldsymbol{x}^S)$
$\{2,3\}$	B	11mph	30%	B	B	$5/10$

Slide Credit: R. Teal Witter

Attribution Values

What is the effect of the feature in different settings?

Consider subsets $S \subseteq [n]$ and define $v(S) = f(\boldsymbol{x}^S)$ where

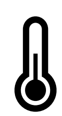
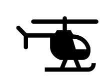
S						$f(\boldsymbol{x}^S)$
$\{2,3\}$	B	11mph	30%	B	B	$5/10$
$\{2,3,5\}$	B	11mph	30%	B	8	$4/10$

Slide Credit: R. Teal Witter

Attribution Values

What is the effect of the feature in different settings?

Consider subsets $S \subseteq [n]$ and define $v(S) = f(\boldsymbol{x}^S)$ where

S						$f(\boldsymbol{x}^S)$
$\{2,3\}$	$89°$	11mph	30%	5	3	$5/10$
$\{2,3,5\}$	$89°$	11mph	30%	5	8	$4/10$

Next: Define attribution value ϕ_i for every feature $i \in [n]$

Slide Credit: R. Teal Witter

Desirable Properties

Null Player: *If a feature never changes the prediction, then its attribution value is 0*

Slide Credit: R. Teal Witter

Desirable Properties

Null Player: *If a feature never changes the prediction, then its attribution value is 0*

Symmetry: *If two features always induce the same change, then their attribution values are the same*

Slide Credit: R. Teal Witter

Desirable Properties

Null Player: *If a feature never changes the prediction, then its attribution value is 0*

Symmetry: *If two features always induce the same change, then their attribution values are the same*

Additivity: *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*

Slide Credit: R. Teal Witter

Desirable Properties

Null Player: *If a feature never changes the prediction, then its attribution value is 0*

Symmetry: *If two features always induce the same change, then their attribution values are the same*

Additivity: *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*

Efficiency: *The attribution values sum to the difference between the predictions on the explicand and baseline*

Slide Credit: R. Teal Witter

Desirable Properties

Null Player: *If a feature never changes the prediction, then its attribution value is 0*

Symmetry: *If two features always induce the same change, then their attribution values are the same*

Additivity: *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*

Efficiency: *The attribution values sum to the difference between the predictions on the explicand and baseline*

$\Leftrightarrow$ Shapley values!

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v\colon 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \sum_{S \subseteq [n] \setminus \{i\}}$$

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{}$$

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Can be re-written as:

$$\phi_i = \underbrace{\frac{1}{n} \sum_{k \in [n-1]}}_{\text{Average over all sizes } k}$$

Average over all sizes k

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Can be re-written as:

$$\phi_i = \frac{1}{n} \sum_{k \in [n-1]} \frac{1}{\binom{n-1}{k}} \sum_{S \subseteq [n] \setminus \{i\} : |S| = k}$$

Average over sets of size k

Average over all sizes k

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Can be re-written as:

$$\phi_i = \frac{1}{n} \sum_{k \in [n-1]} \frac{1}{\binom{n-1}{k}} \sum_{S \subseteq [n] \setminus \{i\} : |S| = k} \boldsymbol{v(S \cup \{i\}) - v(S)}$$

Average over sets of size k

Average over all sizes k

Slide Credit: R. Teal Witter

Shapley Values for Feature Attribution

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Can be re-written as:

$$\phi_i = \frac{1}{n} \sum_{k \in [n-1]} \frac{1}{\binom{n-1}{k}} \sum_{S \subseteq [n] \setminus \{i\} : |S| = k} \boxed{v(S \cup \{i\}) - v(S)}$$

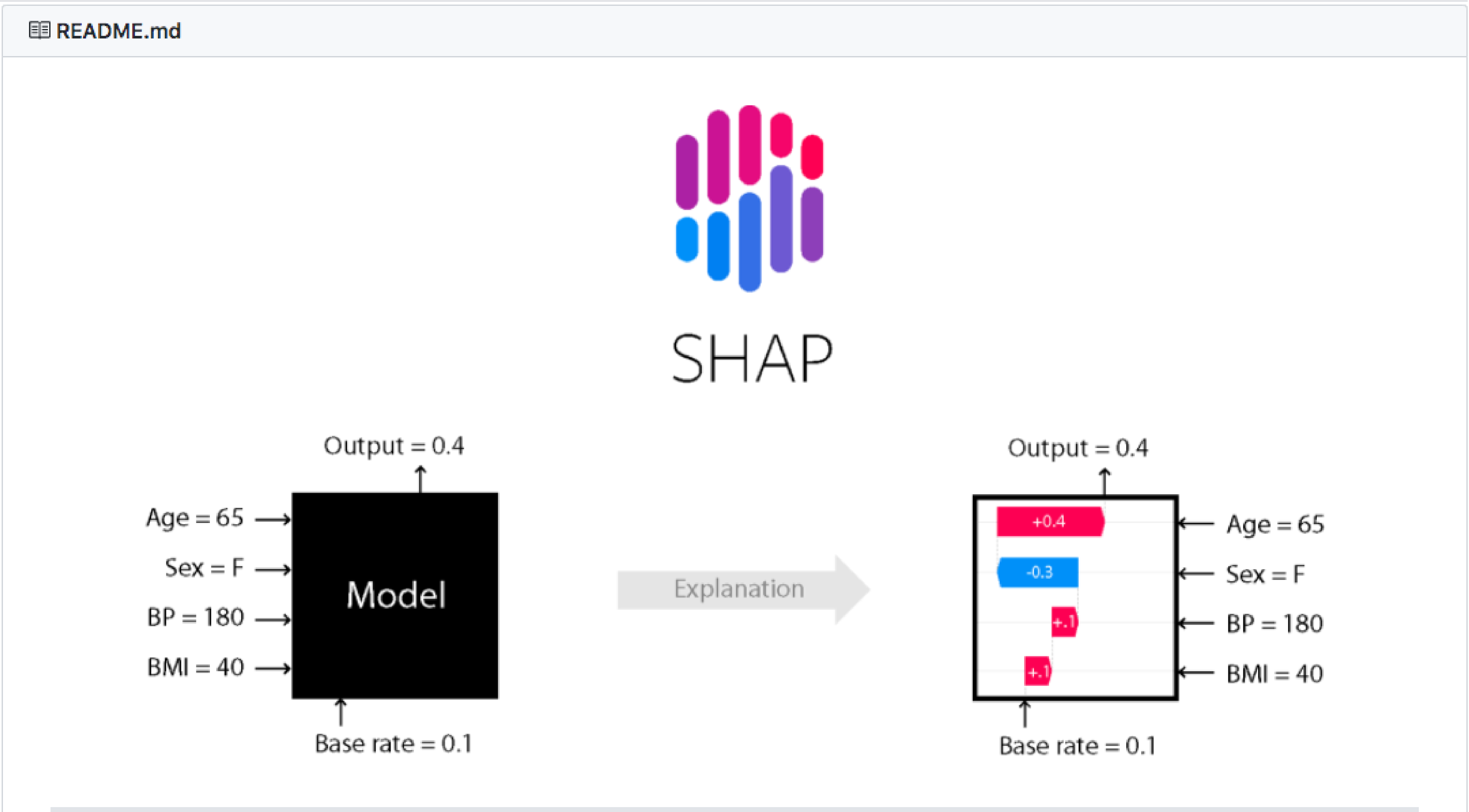
Actual value difference

Average over sets of size k

Average over all sizes k

Slide Credit: R. Teal Witter

Additive feature attribution methods



https://github.com/slundberg/shap

[Lundberg & Lee, 2017]

next lecture: a closer look at LIME, and more

r/ai