# Explainable Machine Learning

Joseph E. Gonzalez
Co-director of the RISE Lab
jegonzal@cs.berkeley.edu

# Two Papers Today

## "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 prob-

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

Widely cited early example of general exploitability.

## The Mythos of Model Interpretability

**Zachary C. Lipton** [1]

**Abstract**

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further explana-

no one has managed to set it in writing, or (ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept. In this paper, we seek to clarify both, suggesting that *interpretability* is not a monolithic concept, but in fact reflects several distinct ideas. We hope, through this critical analysis, to bring focus to the dialogue.

Here, we mainly consider supervised learning and not other

A good critique on the state of explainable AI research.

# Need for Explainability (The Problem)

- Don't trust black box models
  - confidence in the model
  - convince a user of a prediction
- Don't understand the data
  - Reveal relationships in data (science)
- Don't agree with the model
- Regulatory and legal reasons
  - GDPR — Right to an Explanation
  - US Equal Credit Opportunity Act. — Statement of Specific Reasons (for adverse actions)

# Classic Notion of "Interpretability"

- Model's form and parameters have meaning
  - Physical laws ($F = \boldsymbol{G}\ m^1\ m^2/d^2$), growth models ($p=e^{\boldsymbol{\alpha}t}$)
- Learning = Estimating parameters ➔ insight about the underlying phenomenon (and ability to make predictions)
- These models are often "simple" guided by "first principles"

# Classic “interpretable models”

- Linear models
- Decision trees (not random forrest’s)
- Bayesian models
- Nearest neighbor models

# Black Box (Less Interpretable) Models

- Deep Neural Networks
- Random Forests (ensembles in general …)
- Linear models with complex features

# Post-hoc Explainability

- Provide justification for a prediction after it is made
- May rely on training data as well as internal model calculations
- Like human explanations …
- Examples:
  - LIME, GradCam, RISE, Attentive Explanations, …

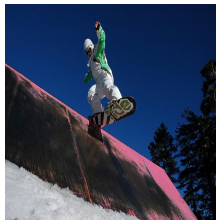
# Example Explanations

## LIME

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

**Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)**

## Attentive Explanations



A man on a snowboard is on a ramp.

Q: What is the person doing?
A: Snowboarding

Because... they are on a snowboard in snowboarding outfit.

A gang of biker police riding their bikes in formation down a street.

Q: Can these people arrest someone?
A: Yes

Because... they are Vancouver police.

## Grounding Visual Explanations

This bird is a **White Pelican** because this is a large white bird with a long orange beak and it is not a **Laysan Albatross** because it does not have a curved bill.

## FICO Score Reason Codes

**Your Credit Score Is: 705**

32: Balances on bankcard or revolving accounts too high compared to credit limits

16: The total of all balances on your open accounts is too high

85: You have too many inquiries on your credit report

13: Your most recently opened account is too new

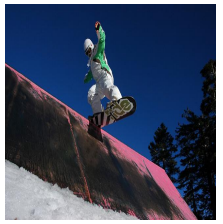
# Are these good/useful/helpful?

## LIME

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

**Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)**

## Attentive Explanations



A man on a snowboard is on a ramp.

Q: What is the person doing?
A: Snowboarding

Because... they are on a snowboard in snowboarding outfit.

A gang of biker police riding their bikes in formation down a street.

Q: Can these people arrest someone?
A: Yes

Because... they are Vancouver police.

## Grounding Visual Explanations

This bird is a **White Pelican** because this is a large white bird with a long orange beak and it is not a **Laysan Albatross** because it does not have a curved bill.

## FICO Score Reason Codes

**Your Credit Score Is: 705**

32: Balances on bankcard or revolving accounts too high compared to credit limits

16: The total of all balances on your open accounts is too high

85: You have too many inquiries on your credit report

13: Your most recently opened account is too new

# Metrics of Success?

- Can they persuade a user (user studies)
- Are the explanations consistent
- Are the explanations falsifiable
- Are the explanations teachable (improve learning)

# Systems Role in Explainability

- Maintaining Provenance
  - What code, data, people involved in developing and training models?
- Attesting/Verifying Provenance
  - Proving that model → decision follows the described provenance
- Providing mechanisms for users to falsify explanations
  - *"You will like 'Blue Planet' because we think you like 'BBC Documentaries about Nature'* ***[x]****."*

# Two Papers Today

## "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 prob-

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

Widely cited early example of general exploitability.

## The Mythos of Model Interpretability

**Zachary C. Lipton** [1]

**Abstract**

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further explana-

no one has managed to set it in writing, or (ii) the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that interpretability refers to more than one concept. In this paper, we seek to clarify both, suggesting that *interpretability* is not a monolithic concept, but in fact reflects several distinct ideas. We hope, through this critical analysis, to bring focus to the dialogue.

Here, we mainly consider supervised learning and not other

A good critique on the state of explainable AI research.