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

**Author: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin**

**Presenter: Gang Su**


UNIVERSITY OF GEORGIA

# Background

- Machine learning is at the core of many recent advances in science and technology.
- Unfortunately, if the users do not trust a model or a prediction, they will not use it. (Black Box)
- Determining trust in individual predictions is an important problem when the model is used for decision making.
- Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. Because real-world data is often significantly different.

It’s 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.

(2) Trusting a model, i.e. whether the user trusts a model to behave in reasonable ways if deployed.

# Contributions

- 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 for different datasets and different tasks.

# Explaining individual predictions

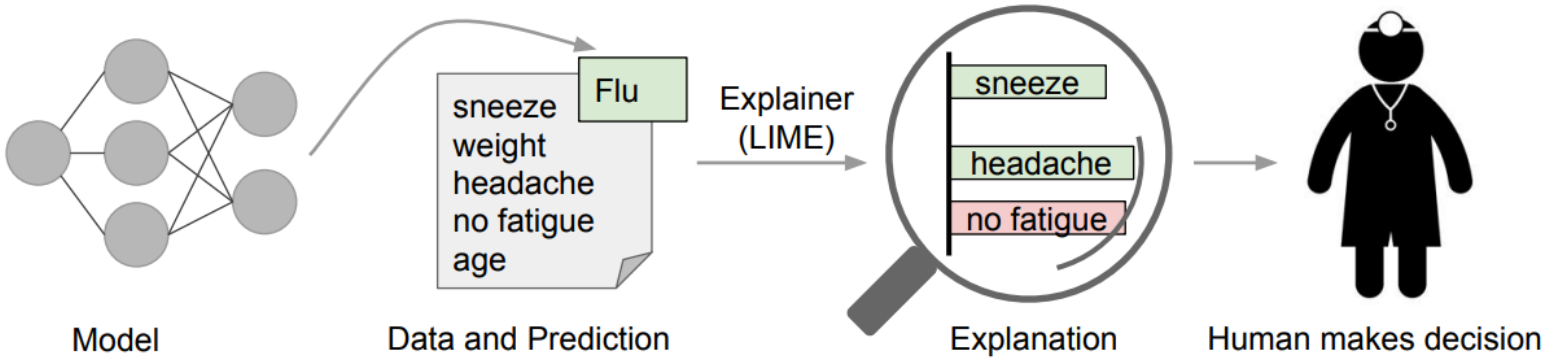



**Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.**

It's clear that a doctor is much better positioned to make a decision with the help of a model if intelligible explanations are provided. In this cases, an explanation is a small list of symptoms with relative weights – symptoms that either contribute to the prediction (in green) or are evidence against it (in red). Humans usually have prior knowledge about the application domain, which they can use to accept (trust) or reject a prediction if they understand the reasoning behind it.

# LIME (Local Interpretable Model-Agnostic Explanations)

- The overall goal of LIME is to identify an **interpretable** model over the interpretable representation that is **local faithful** to the classifier.

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

- We define an explanation as a model $g \in G$ , where G is a class of potentially interpretable models, such as linear model, decision trees, or falling rule lists, i.e. a model $g \in G$ can be readily presented to the user with visual or textual artifacts.

- $\pi_x(z)$ is a proximity measure between an instance z to x, so as to define locality around x.
- $\mathcal{L}(f, g, \pi_x)$ is a measure of how unfaithful g is in approximating f in the locality defined by $\pi_x$ .

# Sparse Linear Explanations

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

- This paper wants to minimize the locality-aware loss $\mathcal{L}(f, g, \pi_x)$ without making any assumptions about f, which also means **model-agnostic**. Thus, in order to learn the local behavior of f as the interpretable inputs vary, we approximate $\mathcal{L}(f, g, \pi_x)$ by drawing samples, weighted by $\pi_x$ . $\pi_x(z) = exp(-D(x, z)^2/\sigma^2)$

- We sample instances z around x uniformly at random.

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

- $\Omega(g)$ is a measure of complexity (as opposed to interpretability) of the explanation model G. For example, for decision trees complexity may be the depth of the tree, for linear models complexity may be the number of non-zero weights. $\Omega(g) = \infty \mathbb{1}[\|w_g\|_0 > K]$

**Algorithm 1** Sparse Linear Explanations using LIME

**Require:** Classifier $f$, Number of samples $N$
**Require:** Instance $x$, and its interpretable version $x'$
**Require:** Similarity kernel $\pi_x$, Length of explanation $K$

$\mathcal{Z} \leftarrow \{\}$
**for** $i \in \{1, 2, 3, ..., N\}$ **do**
  $z_i' \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z_i', f(z_i), \pi_x(z_i) \rangle$
**end for**
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with $z_i'$ as features, $f(z)$ as target
**return** $w$

- First selecting K features with Lasso (using the regularization) and then learning the weights via least squares.

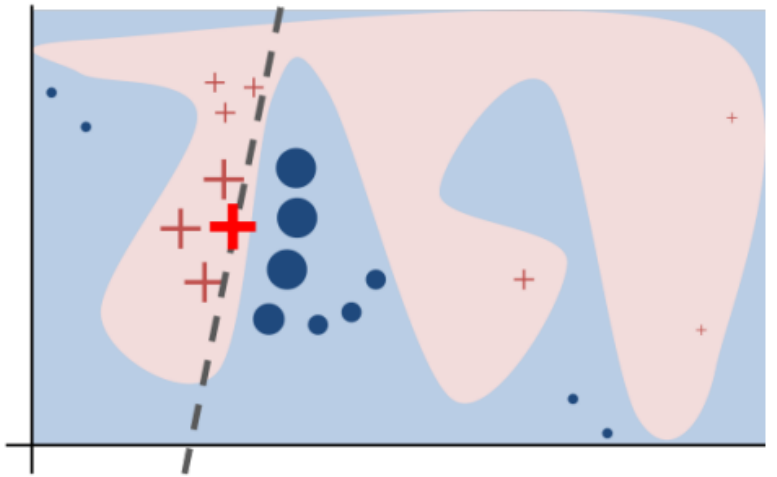


**Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function $f$ (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using $f$, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.**

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

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

# Example 1: Text classification with SVMs

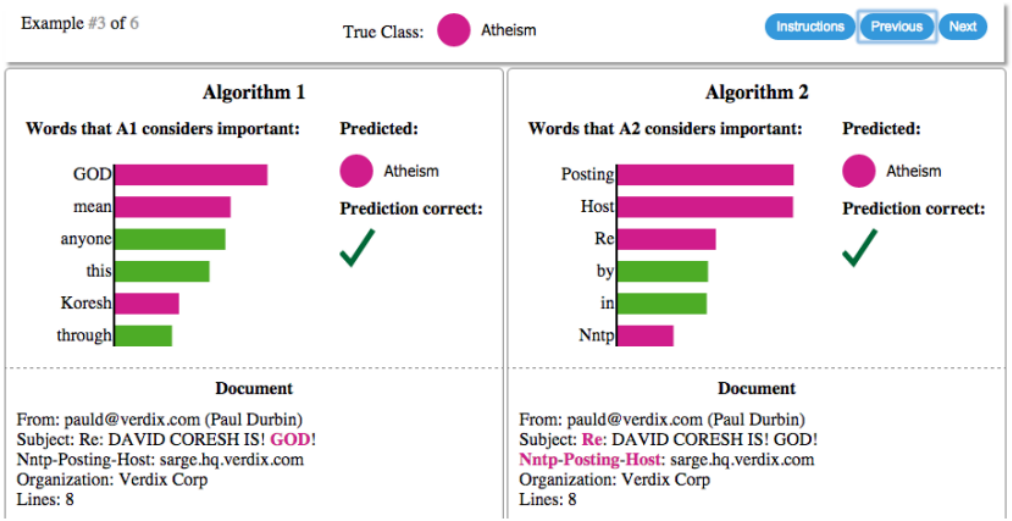
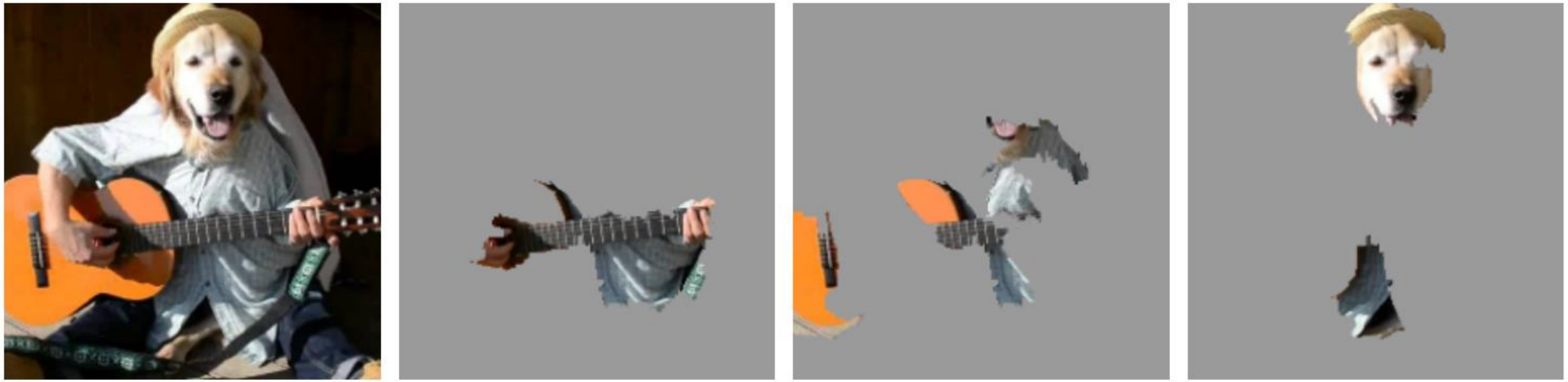



**Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").**

After getting such insights from explanations, it is clear that this dataset has serious issues, and this classifier cannot be trusted.
It is also clear what the problems are, and the steps that can be taken to fix these issues and train a more trustworthy classifier.

# Example 2: Deep networks for images



(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 neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)**

- When using sparse linear explanations for image classifiers, we just highlight the super-pixels with positive weight towards a specific class, as they give intuition as to why the model would think that class may be present.
- This kind of explanation enhances trust in the classifier, as it shows that it is not acting in an unreasonable manner.

# SP-LIME (Submodular Pick for Explaining Models)

Although an explanation of a single prediction provides some understanding into the reliability of the classifier to the user, it is not sufficient to evaluate and assess trust in the model as a whole.

This paper propose to give a global understanding of the model by explaining a set of individual instances. This approach is still model agnostic, and it complementary to computing summary statistics such as held-out accuracy.

The basic idea is to pick the fewest samples to efficiently cover the feature space. In this way, the reliability of the model can be roughly checked

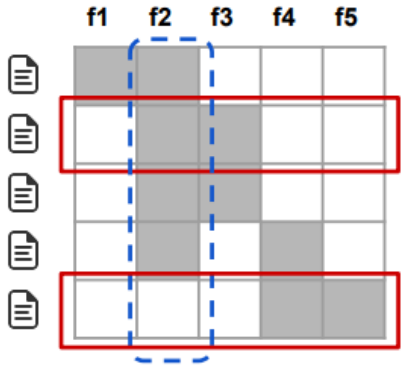



Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

# Conclusion

- This paper argued that trust is crucial for effective human interaction with machine learning systems, and that explaining individual predictions is important in assessing trust.
- This paper proposed LIME, a modular and extensible approach to faithfully explain the predictions of any model in an interpretable manner.
- This paper also introduced SP-LIME, a method to select representative and non-redundant predictions, providing a global view of the model to users.

# Thank you!

## Questions?