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

Authors: Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin

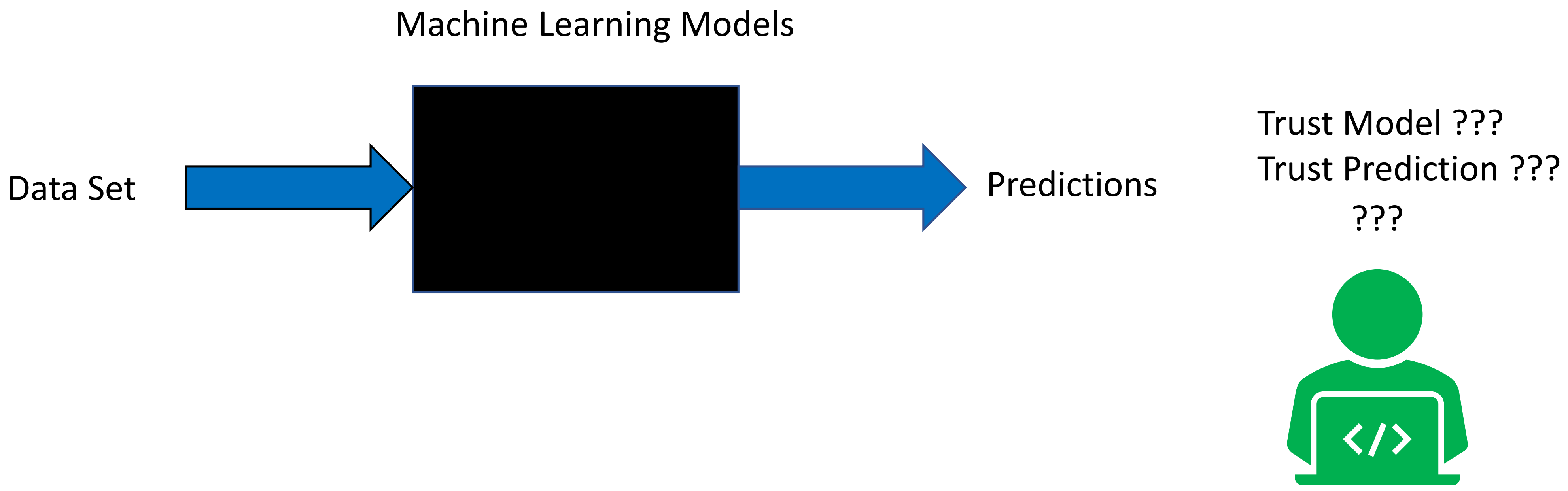
University of Washington

KDD 2016 San Francisco, CA, USA

Prepared by: Chandra Thapa (April 2019)

Outline

- Need of interpretability in classifiers.
- LIME
- Results



Machine Learning Models

Data Set

Predictions

Trust Model ???

Trust Prediction ???

???

Why we need interpretability in classifiers?

- For better models
    - To find biases in the data
- For legal and ethical provisions
    - GDPR
    - Establish Trust
- Interpretable results to human users

Classifiers

- Black-box
  - Neural Networks (DNN, RNN, CNN)
  - Random Forests
  - Support Vector Machines

More accurate not interpretable

- Interpretable
  - Decision Trees
  - Rule based models: If-then rules, lists of rules..

Simple but less accurate

- LIME – a novel explanation technique that explains the prediction of *any* classifier.

- LIME:
  - **L**ocally
  - **I**nterpretable
  - **M**odel-agnostic
  - **E**xplanations

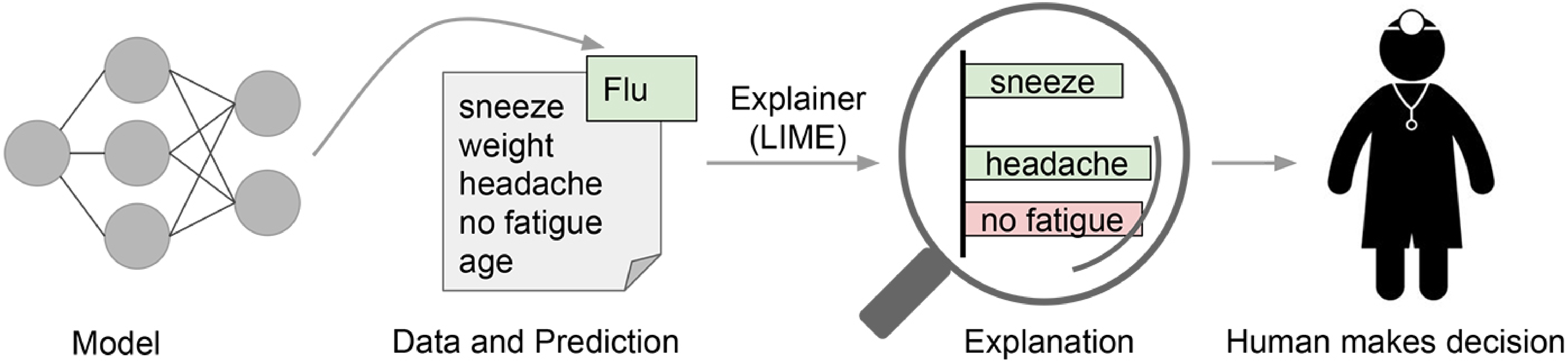


Fig: Explaining individual predictions to a human decision-maker

Explaining Global behavior

LIME explains a single prediction

local behavior for a single instance

Can't examine all explanations

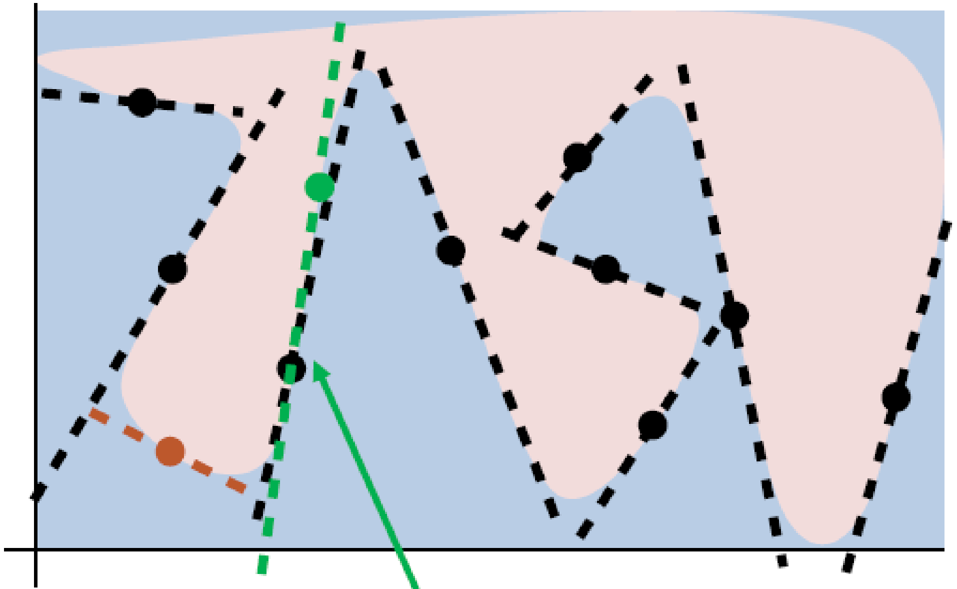
Instead pick *k* explanations to show to the user

Representative

Should summarize the model's global behavior

Diverse

Should not be redundant in their descriptions



Single explanation

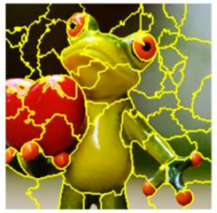
Source: http://oc.acm.org/docs/7-12-2017%20-%20Sameer%20Singh%20-%20Explaining%20Black-Box%20ML%20Predictions.pdf

Algorithm (LIME)

1. Permutate observation.
2. Complex model predict the outcome of all permuted observations.
3. Calculate distance between permutations and original observations.
4. Covert the distance to a similarity score.
5. Pick m features best describing the complex model outcome from the permuted data.
6. Fit a simple model to the permuted data, explaining the complex model outcome with the m features from the permute data weighted by its similarity to the original observation.
7. Feature weights from the simple model make explanations for the complex model local behaviour.

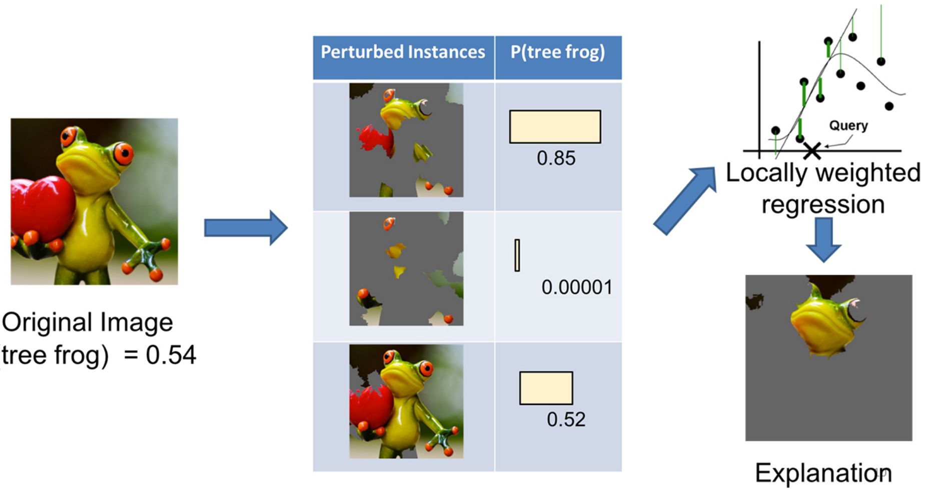
Example (how likely it is a tree frog?)

Original Image



Interpretable Components

1) Divide the image into interpretable components



Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

Locally weighted regression

Explanation

2) Generate a data set of perturbed instance by turning some of the interpretable components “off” (in this case grey).

3) Each perturbed instance – get the probability that a tree frog is in the image according to the model.

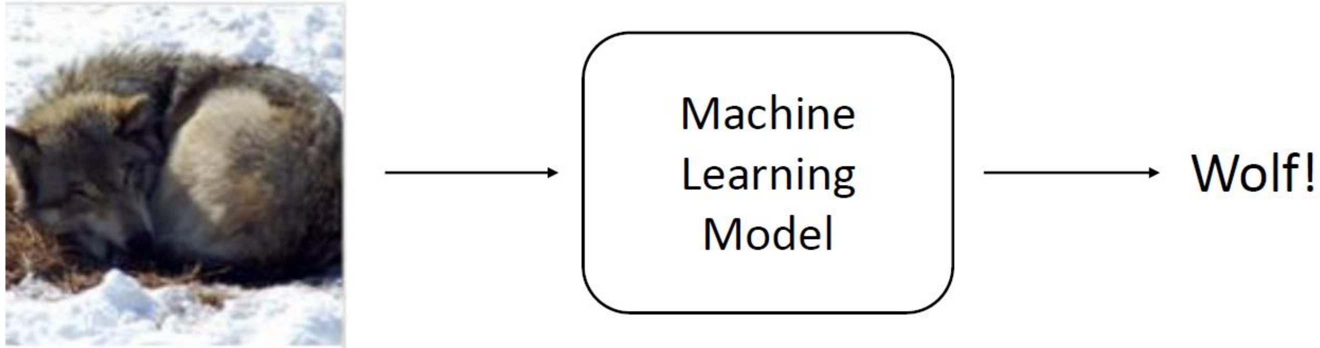
4) Learn a simple (linear) model on this data set, which is locally weighted, i.e., we care more about making mistakes in perturbed instances that are more similar to the original image.

5) We present the interpretable component with highest positive weights as an explanation, greying out everything else.

Source: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Classification: Wolf or a Husky?

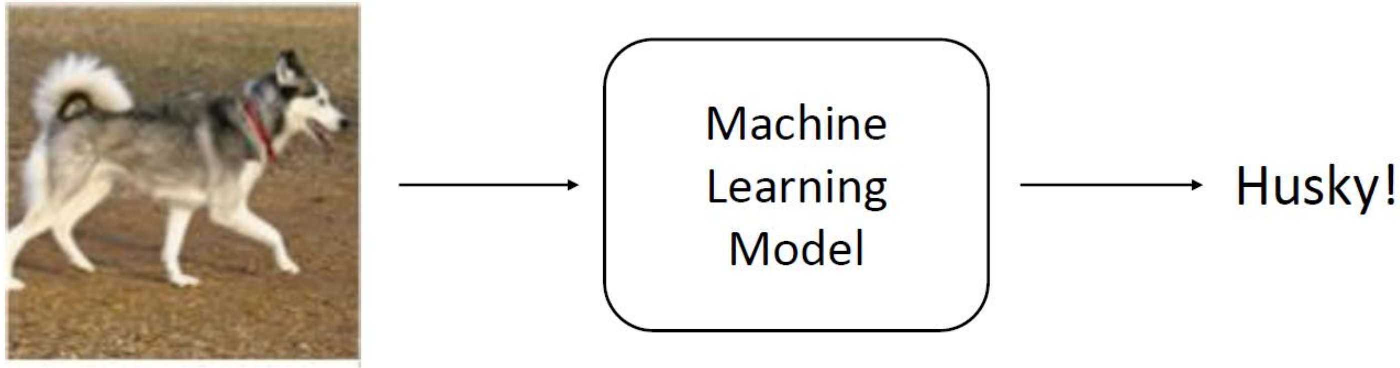
Adopt or not?



Machine Learning Model

Wolf!

Adopt or not?

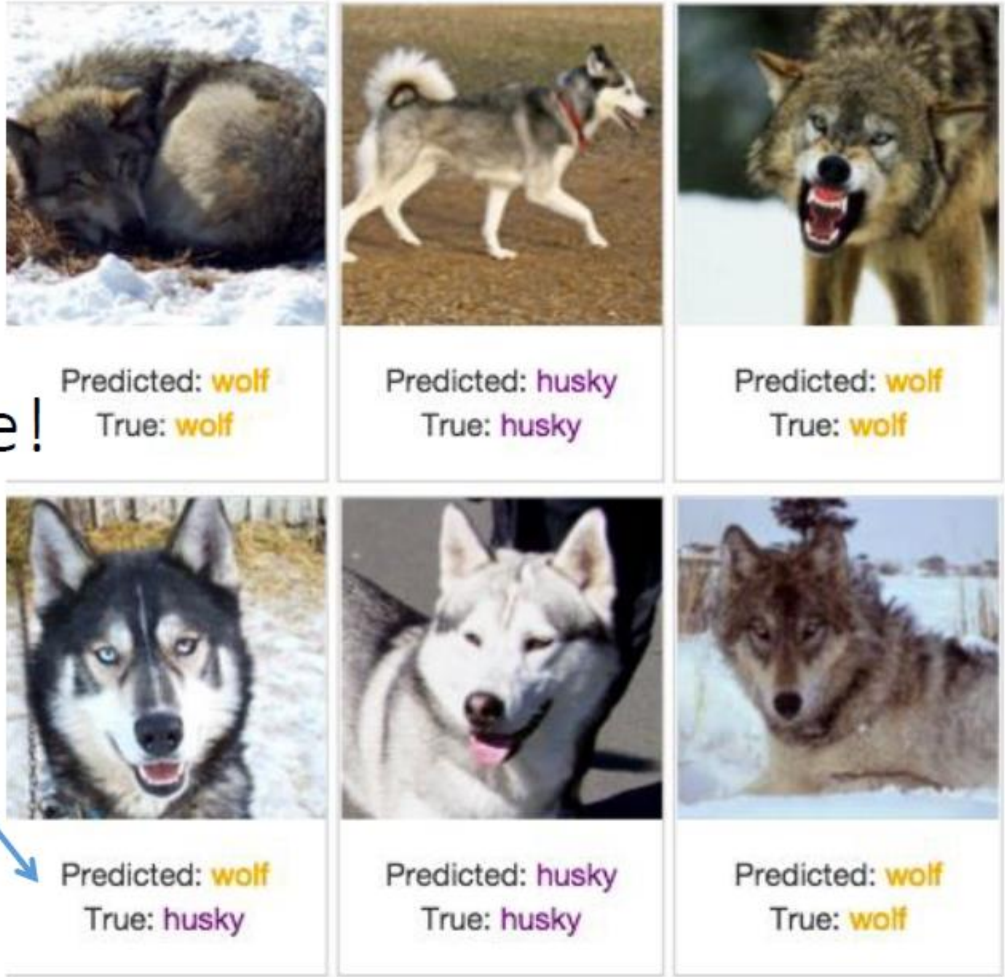


Machine Learning Model

Husky!

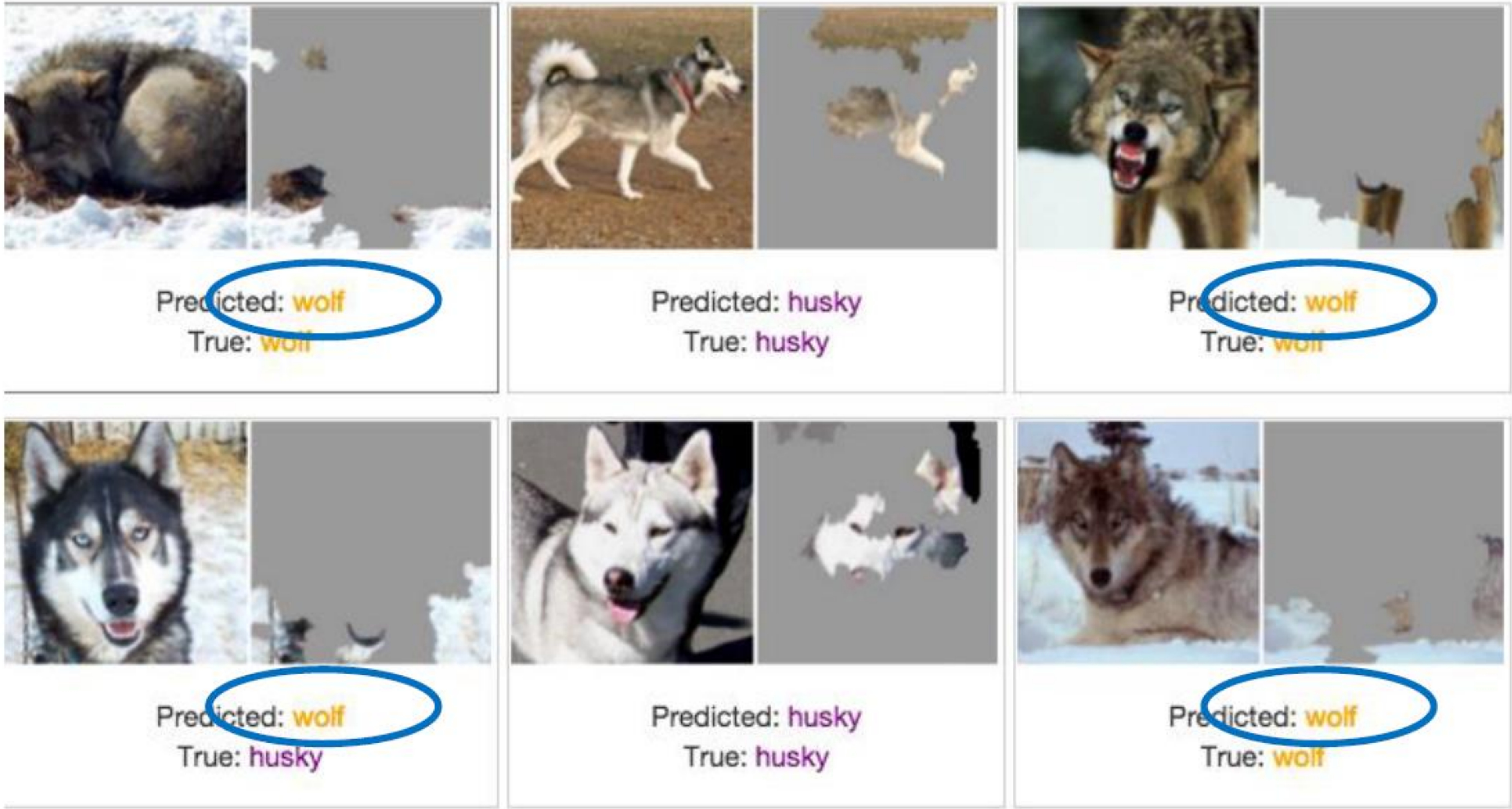
Source: http://oc.acm.org/docs/7-12-2017%20-%20Sameer%20Singh%20-%20Explaining%20Black-Box%20ML%20Predictions.pdf

Classification: Wolf or a Husky?



Only 1 mistake!

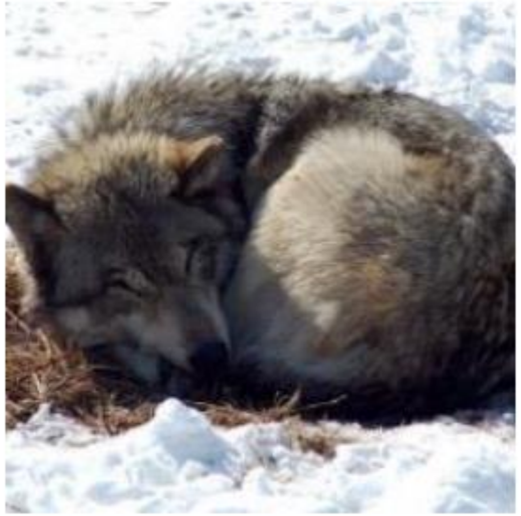
Source: http://oc.acm.org/docs/7-12-2017%20-%20Sameer%20Singh%20-%20Explaining%20Black-Box%20ML%20Predictions.pdf



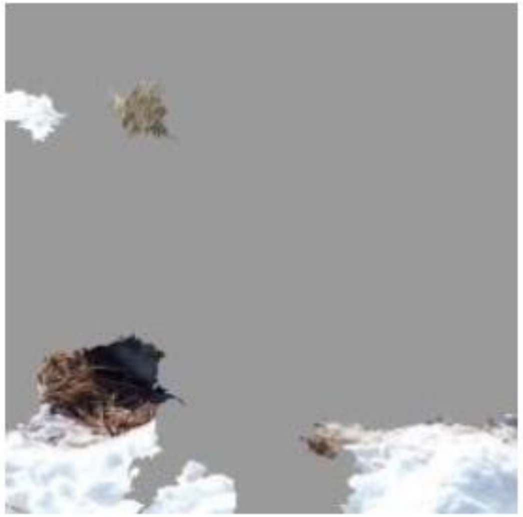
We’ve built a great snow detector…

Source: http://oc.acm.org/docs/7-12-2017%20-%20Sameer%20Singh%20-%20Explaining%20Black-Box%20ML%20Predictions.pdf

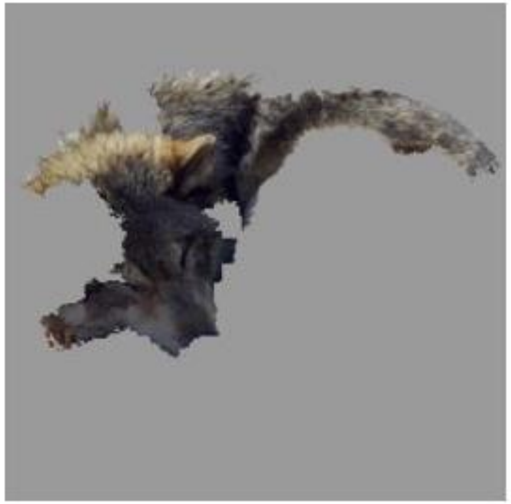
Comparing Classifiers



Original Image



“Bad” Classifier



“Good” Classifier

Source: http://oc.acm.org/docs/7-12-2017%20-%20Sameer%20Singh%20-%20Explaining%20Black-Box%20ML%20Predictions.pdf

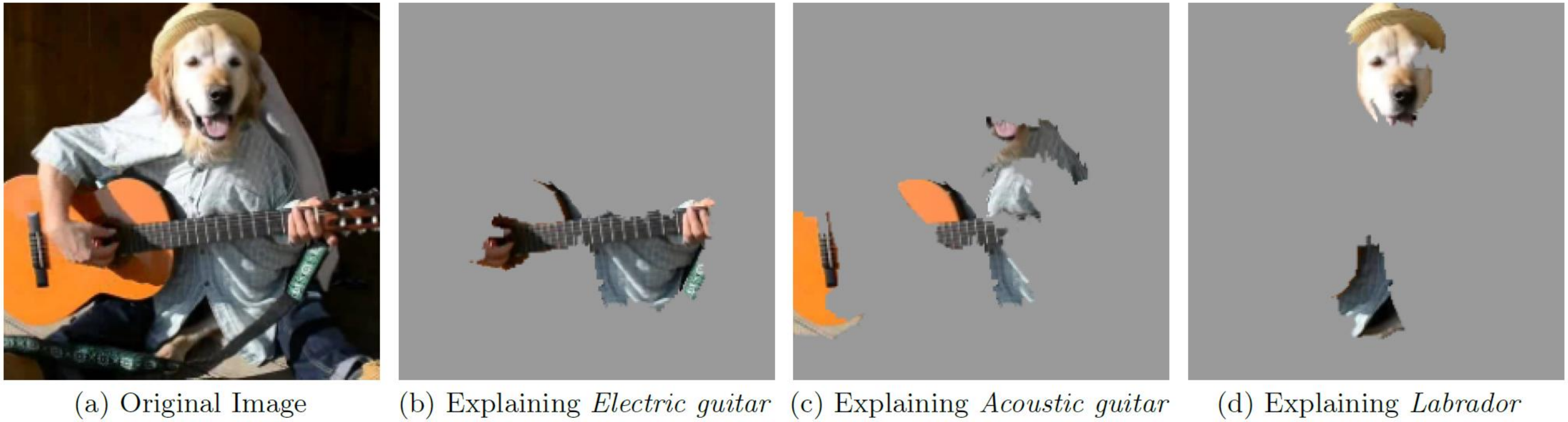
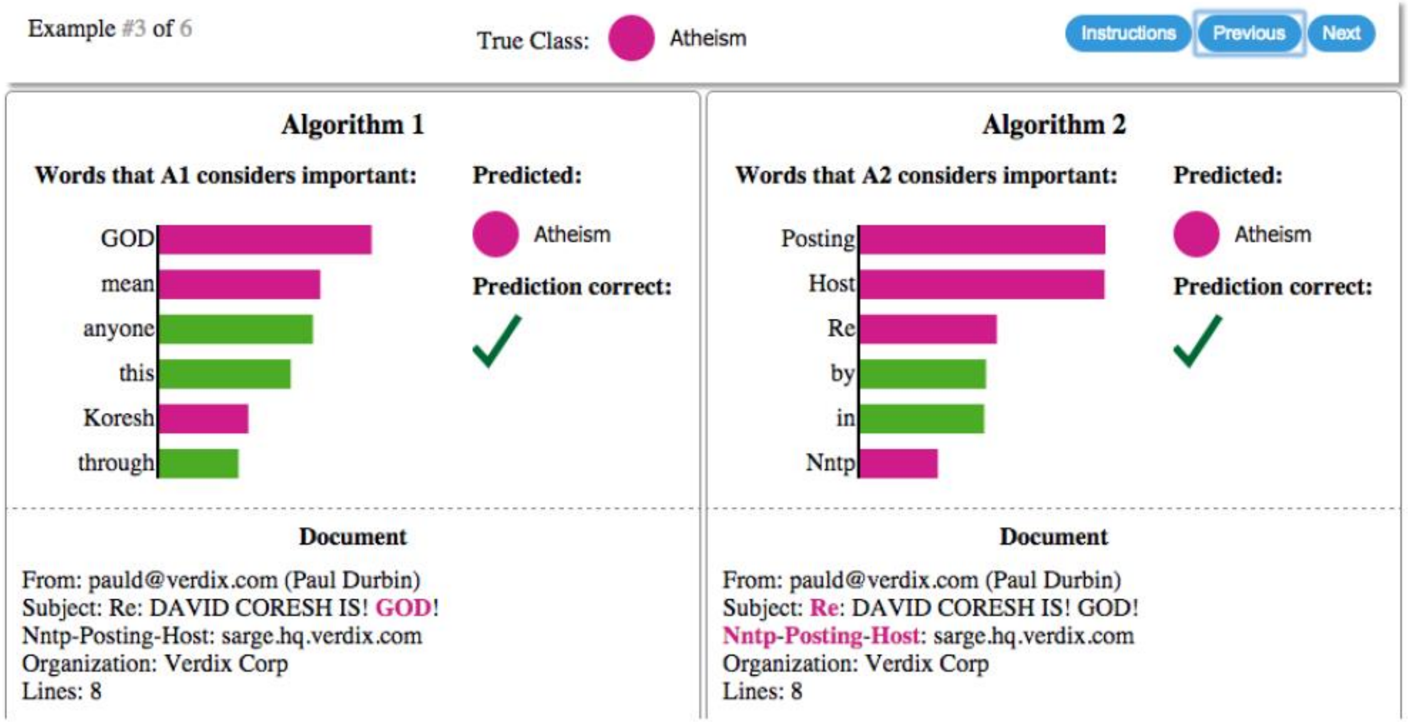

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Fig: 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)



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

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

Fig: 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").

Simulated user experiments

- presented simulated user experiments to evaluate the **utility of explanations in trust-related tasks**.
- addressed the following questions:

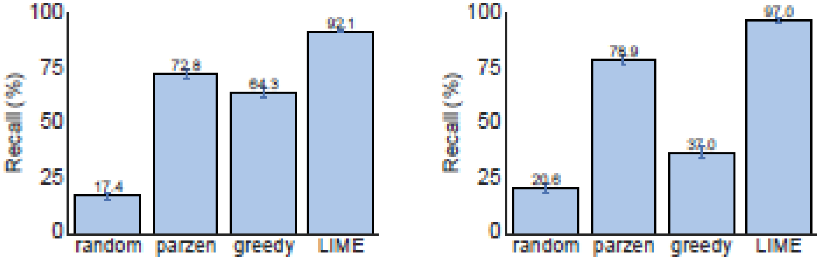
1. Are the **explanations faithful** to the model?
2. Can the explanations aid users in ascertaining **trust in predictions**?, and
3. Are the **explanations useful** for evaluating the model as a whole?

- Code and data for the experiments are available at https://github.com/marcotcr/lime-experiments.

Experimental Setup

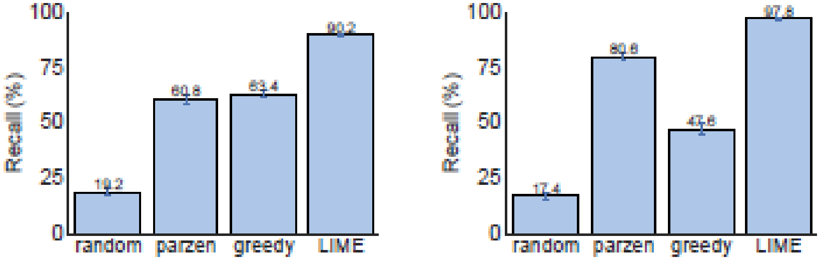
- Two datasets (books and DVDs), 2000 instances each
- Task is to classify product reviews as positive or negative
- Trained
  - Decision trees, logistic regression with L2 regularizations, nearest neighbours, support vector machines with RBF kernel – trained with all using bag of words as features.
  - Random forest (with 1000 trees) – trained with the average word2vec embedding.
- Training set (1600 instances) and test (400 instances)
- For explanation: used **LIME, Parzen, Greedy, Random**

Are explanations faithful to the model?



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.



(a) Sparse LR (b) Decision Tree

Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Should I trust this prediction? Can I trust this model?

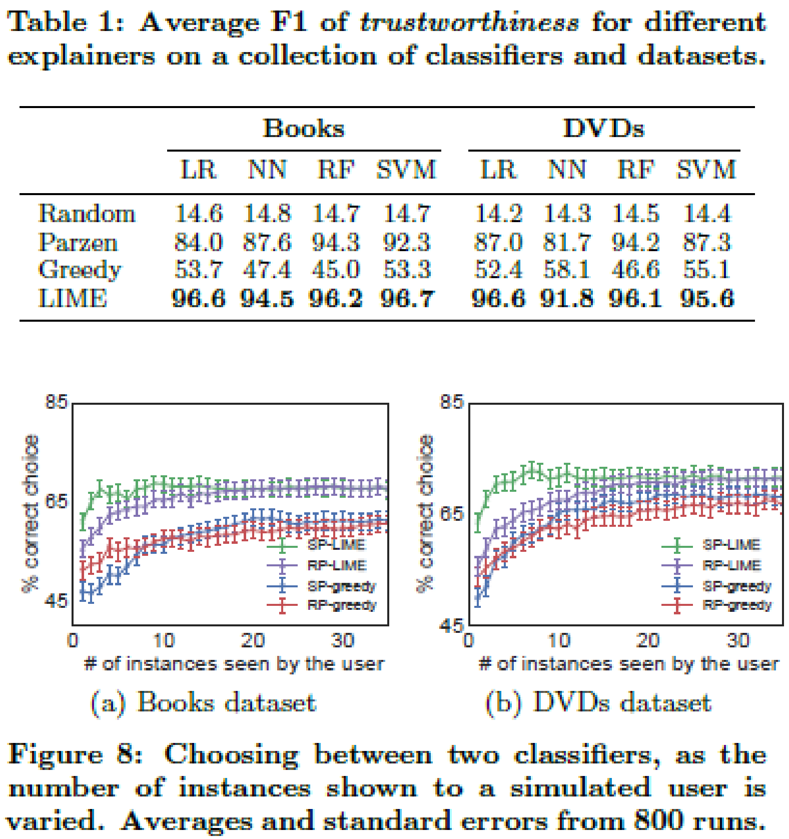


Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

(a) Books dataset

(b) DVDs dataset

Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

Can users select the best classifier?

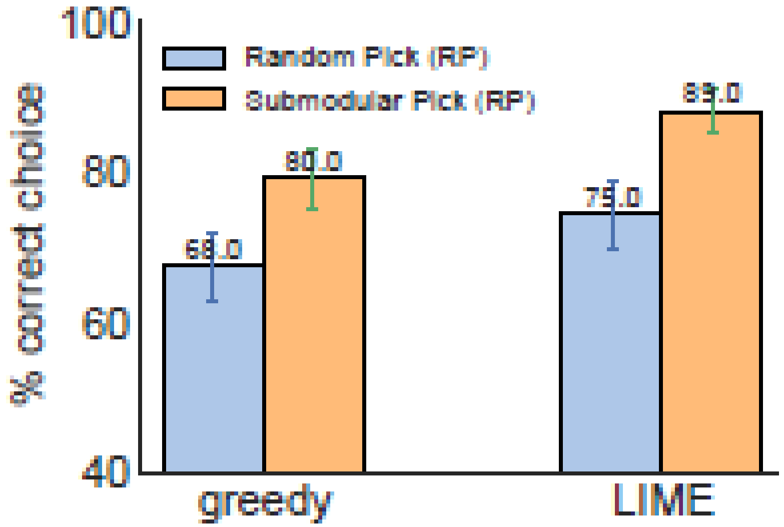


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

Thank You!

How to permute an observation

When it comes to permuting an observation, `lime` depends on the type of input data. Currently two types of inputs are supported: *tabular* and *text*

Tabular Data

When dealing with tabular data, the permutations are dependent on the training set. During the creation of the explainer the statistics for each variable are extracted and permutations are then sampled from the variable distributions. This means that permutations are in fact independent from the explained variable making the similarity computation even more important as this is the only thing establishing the locality of the analysis.

Text Data

When the outcome of text predictions are to be explained the permutations are performed by randomly removing words from the original observation.
Depending on whether the model uses word location or not, words occurring multiple times will be removed one-by-one or as a whole.