



Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**



Interpretability

Advanced Machine Learning for NLP
Jordan Boyd-Graber
NEED FOR INTERPRETABILITY

Slides adapted from Marco Tulio Ribeiro

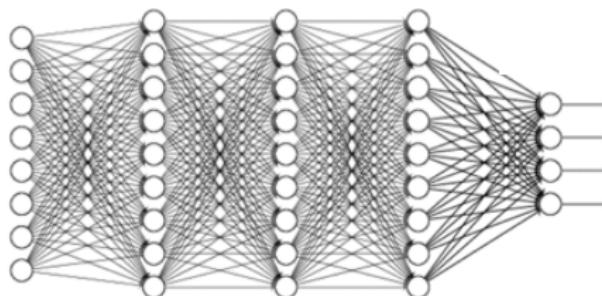
Classification

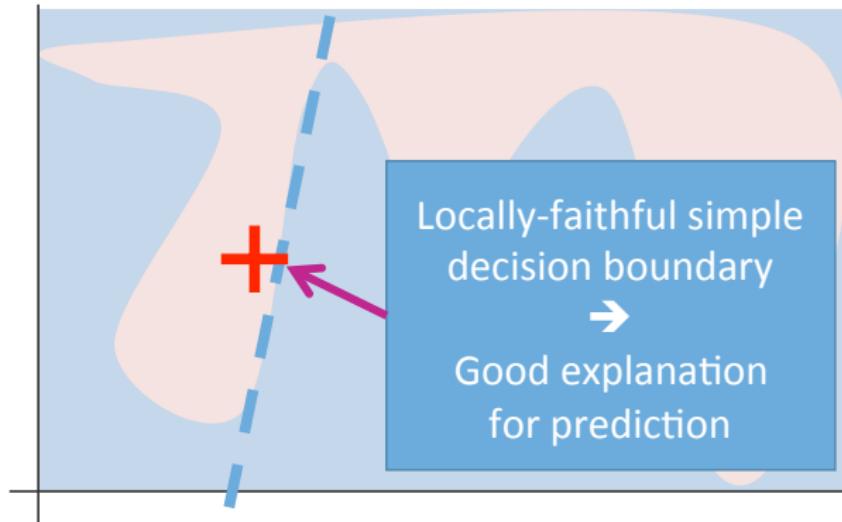
- But representation is often less important (means to end)
- We really care about end result
- And not doing simple things like decision trees / linear classifier

Classification

- But representation is often less important (means to end)
- We really care about end result
- And not doing simple things like decision trees / linear classifier
- That's why we're making complicated algorithms

Can explain
this mess ☺





Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier.
LIME: Local Interpretable Model-Agnostic Explanations

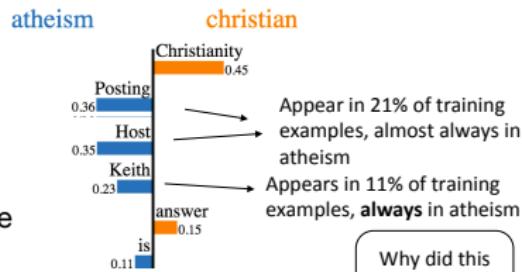
What's an Explanation

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note



Prediction probabilities



Why did this happen? How do I fix it?



What's an Explanation



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



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



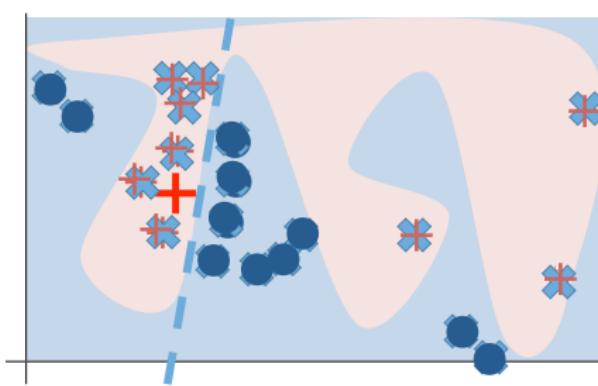
$P(\text{dog}) = 0.21$



What makes good Explanation?

- Interpretable: Humans can Understand
- Faithful: Describes Model
- Model Agnostic: Generalize to Many Models

Method



- Sample points around x_i
- Use complex model to predict labels for each sample
- Weigh samples according to distance to x_i
- Learn new simple model on weighted samples
- Use simple model to explain

Perturbing an Example

x (3 color channels / pixel)



Model

x' (contiguous superpixels)



Human

Perturbing an Example

x (embeddings)

0.5	0.3	1.3	4.4	1.1	...
-----	-----	-----	-----	-----	-----



Model

x' (words)

This is a horrible movie.



Human

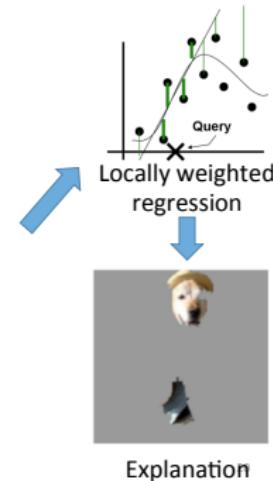
Image Example



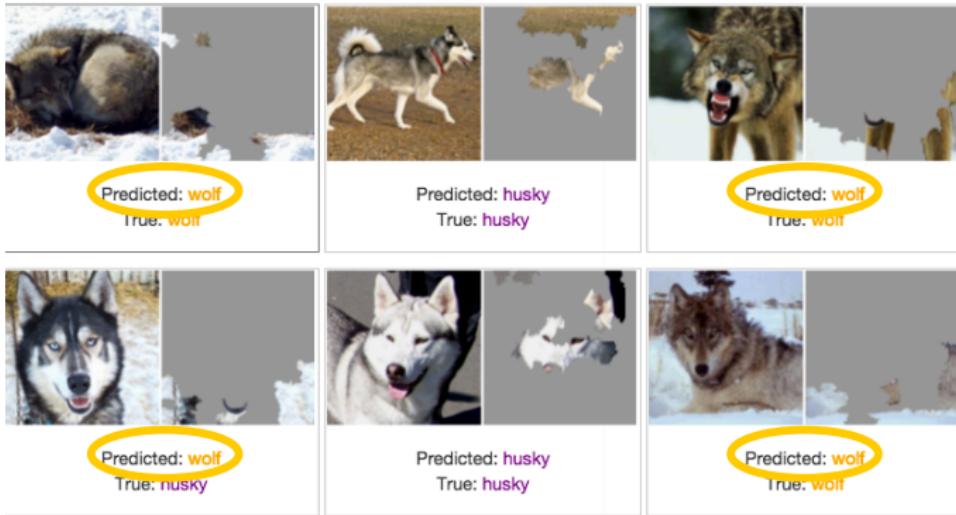
Original Image
 $P(\text{labrador}) = 0.21$



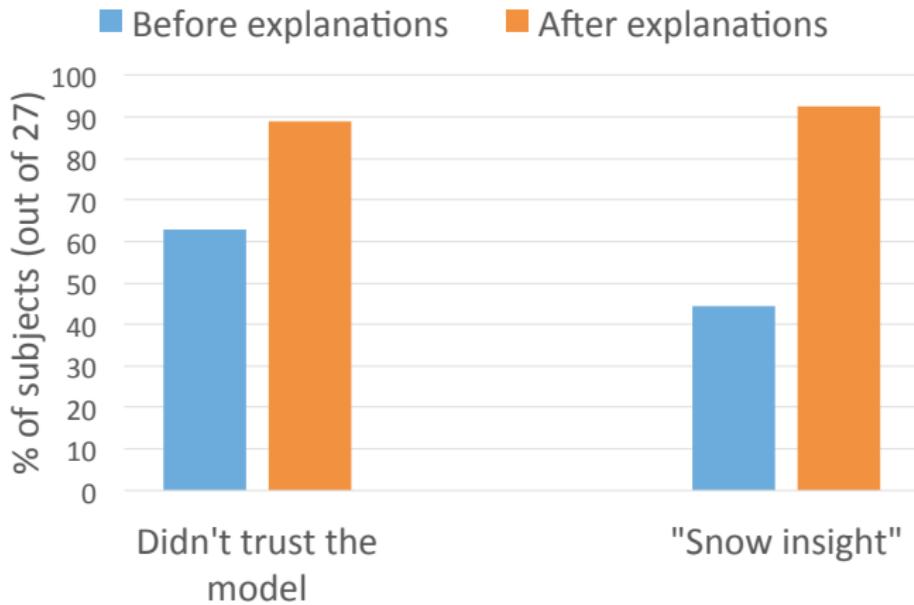
Perturbed Instances	$P(\text{Labrador})$
	0.92
	0.001
	0.34



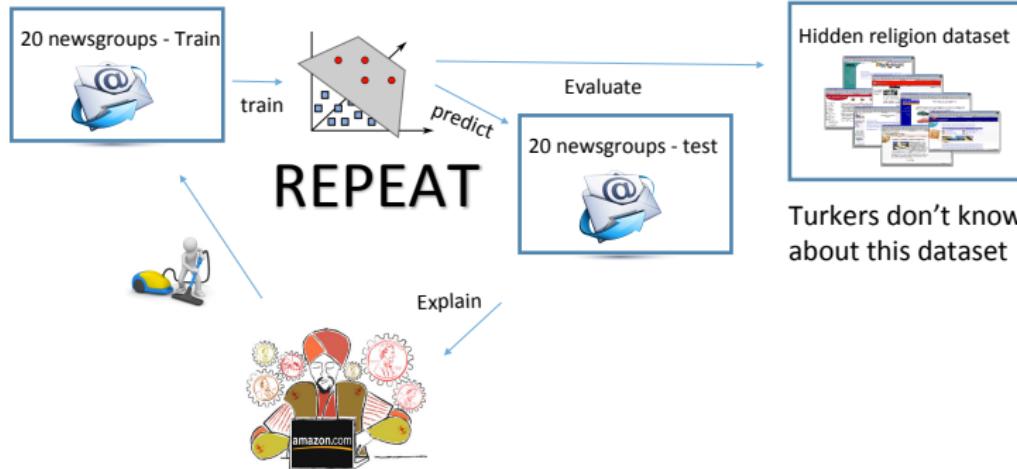
Is this a good Classifier?



Is this a good Classifier?



Improving ML Algorithms



Improving ML Algorithms

Example #5 of 10

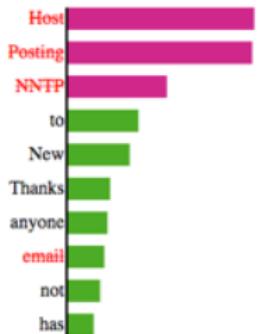
True Class:  Atheism

[Instructions](#)

[Previous](#)

[Next](#)

Words that the algorithm considers important.



Bar length indicates importance, and color indicates to which topic: Christianity (green) or Atheism (Pink).

Please click on the words (right next to the bars) that you think the algorithm is using incorrectly, because they are not important to distinguish between Atheism and Christianity. They should be red and crossed off after you click them.

Document

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where **to** obtain the DARWIN fish.
This is the same question I have and I have **not** seen an answer on the net. If **anyone has** a contact please post on the net or **email** me.

Thanks,

john chadwick
johnchad@triton.unm.edu
or

Improving ML Algorithms

