Interpretability

Advanced Machine Learning for NLP

Jordan Boyd-Graber

NEED FOR INTERPRETABILITY

Slides adapted from Marco Tulio Ribeiro

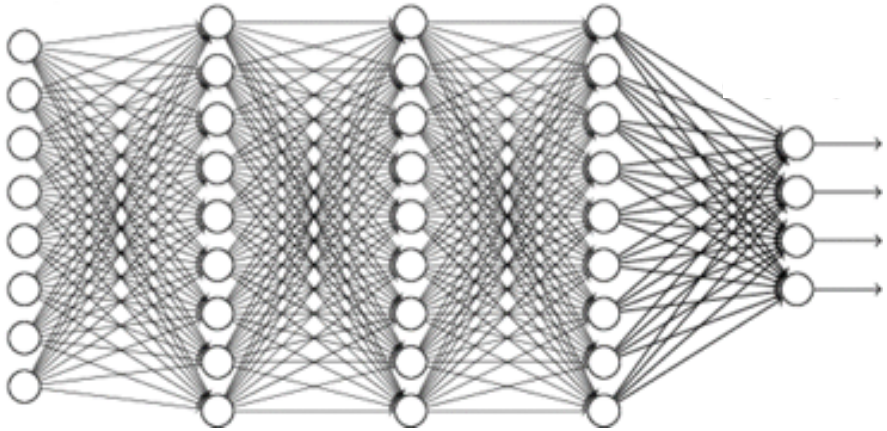
Classification

- 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

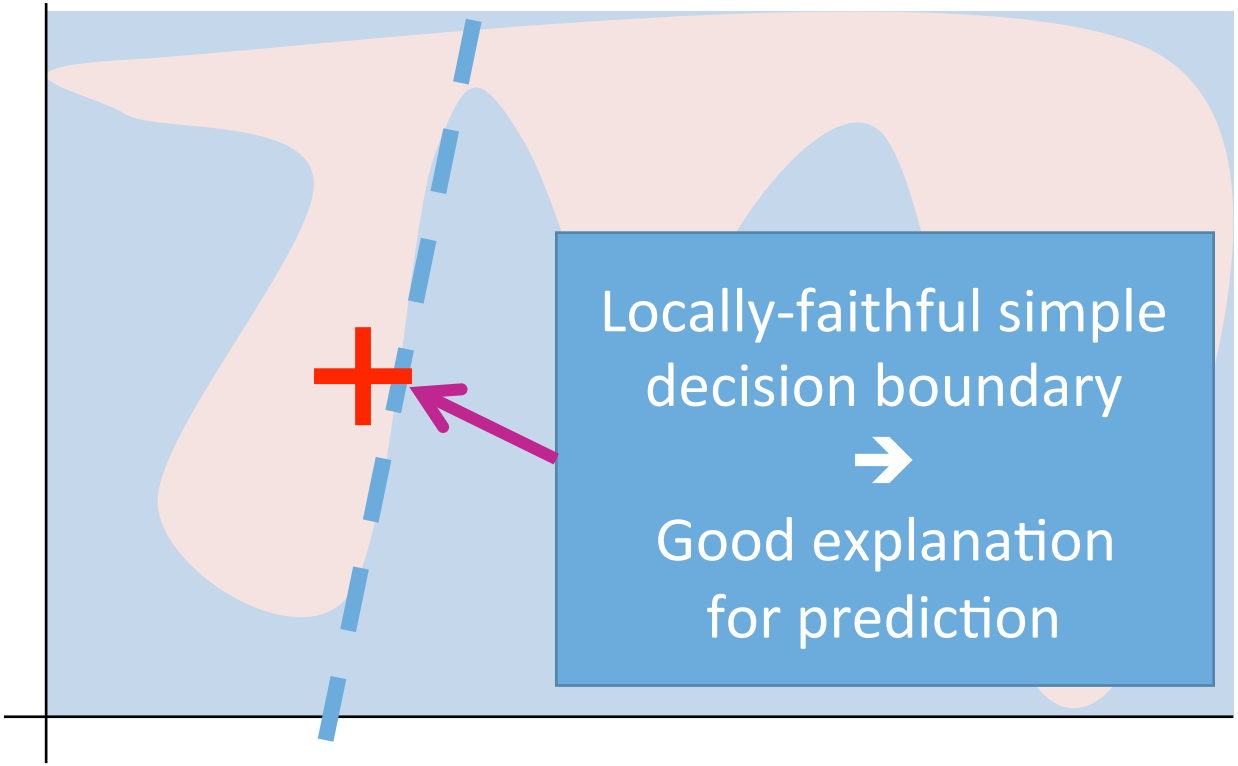
Classification

- 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 ☺



LIME



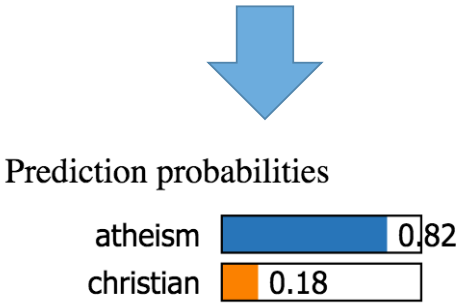
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?” Explaining the Predictions of Any Classifier.

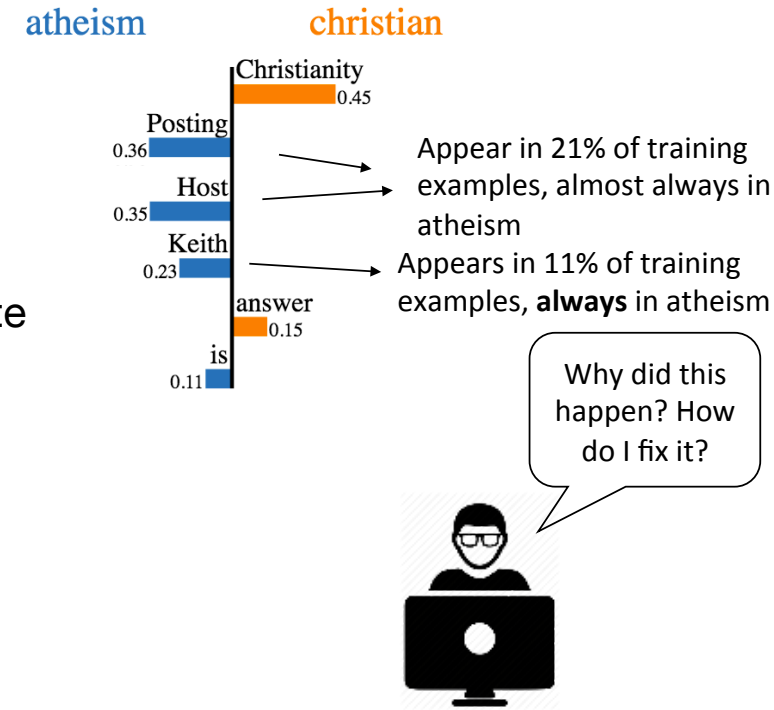
LIME: Local Interpretable Model-Agnostic Explanations

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





What's an Explanation

P( ) = 0.32

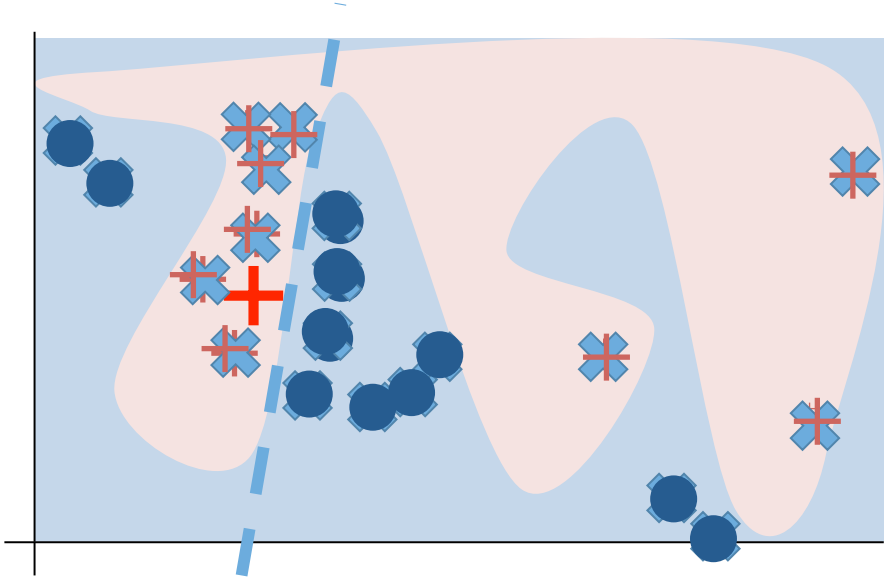
P( ) = 0.24

P( ) = 0.21

What makes good Explanation?

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

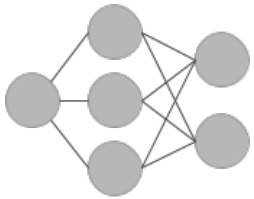
Method



- 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

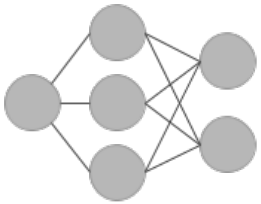
x' (contiguous superpixels)

Human

Perturbing an Example

x (embeddings)

0.5	0.3	1.3	4.4	1.1	...



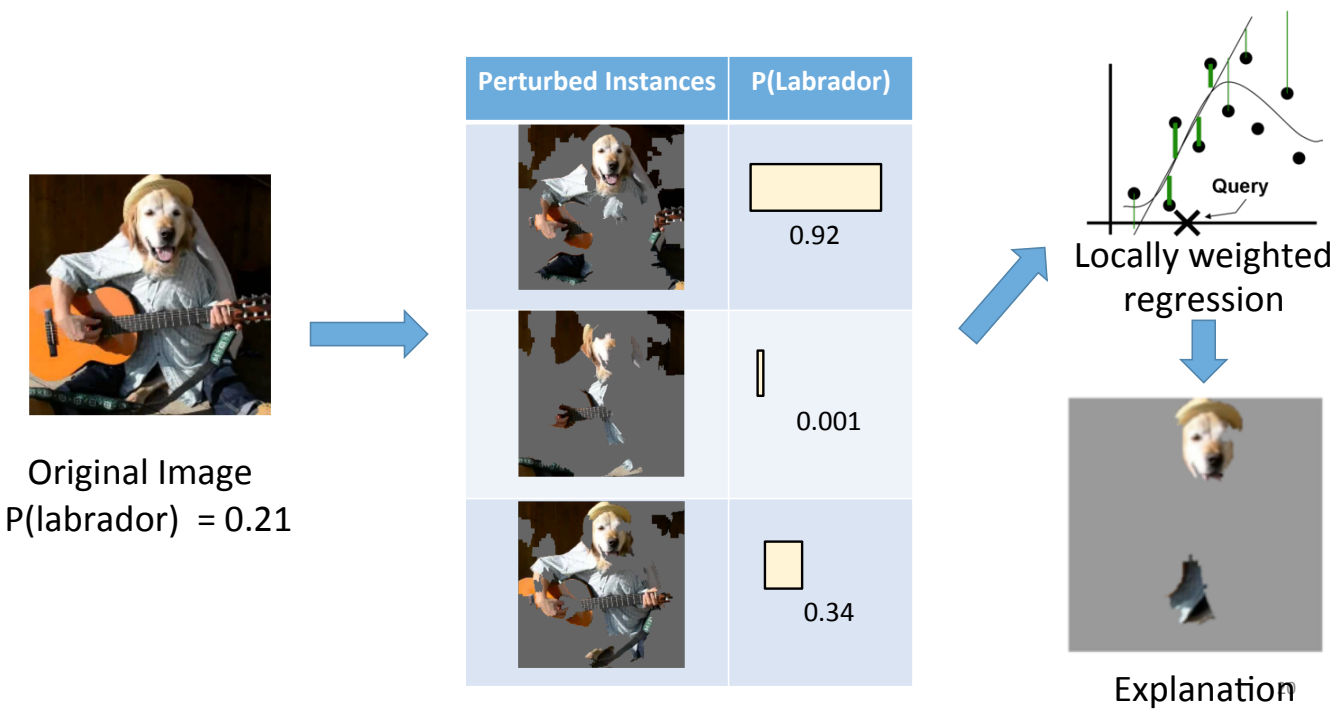
Model

x' (words)

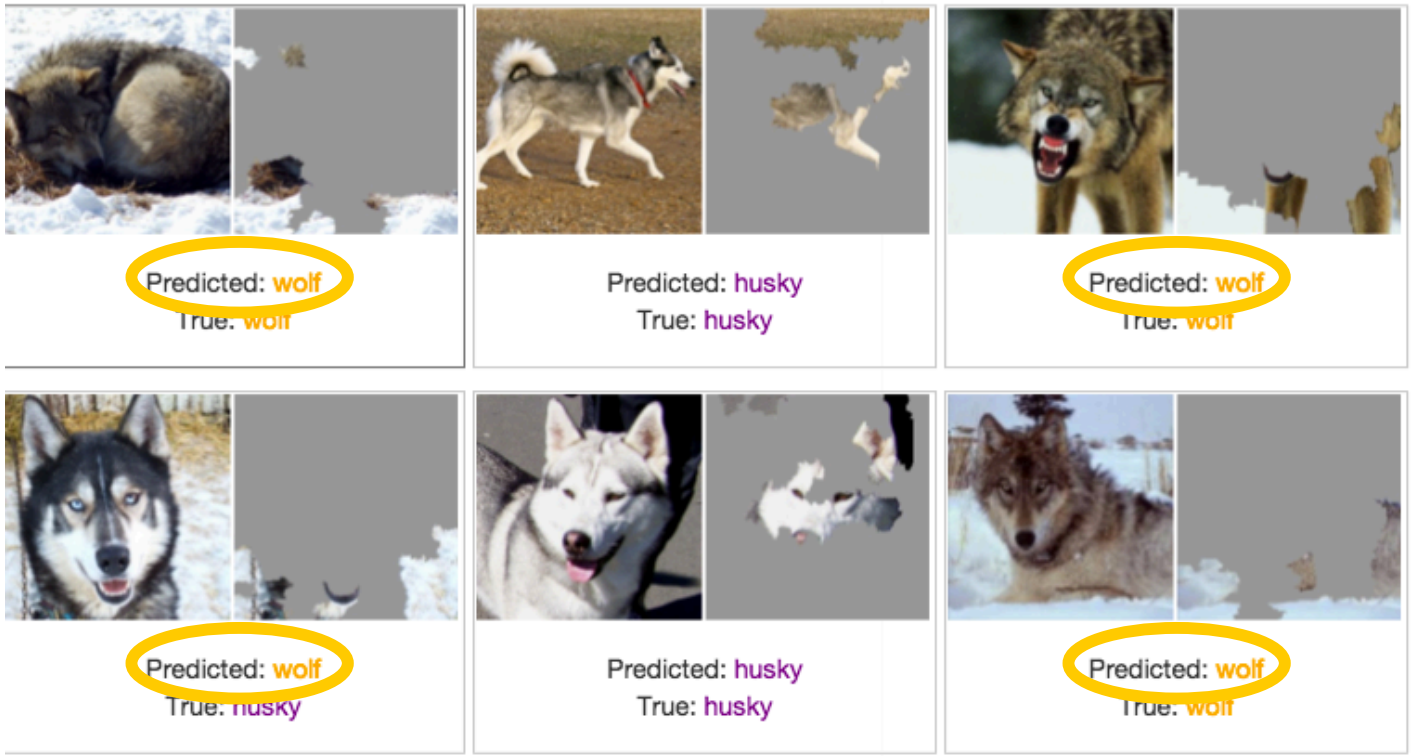
This is a horrible movie.

Human

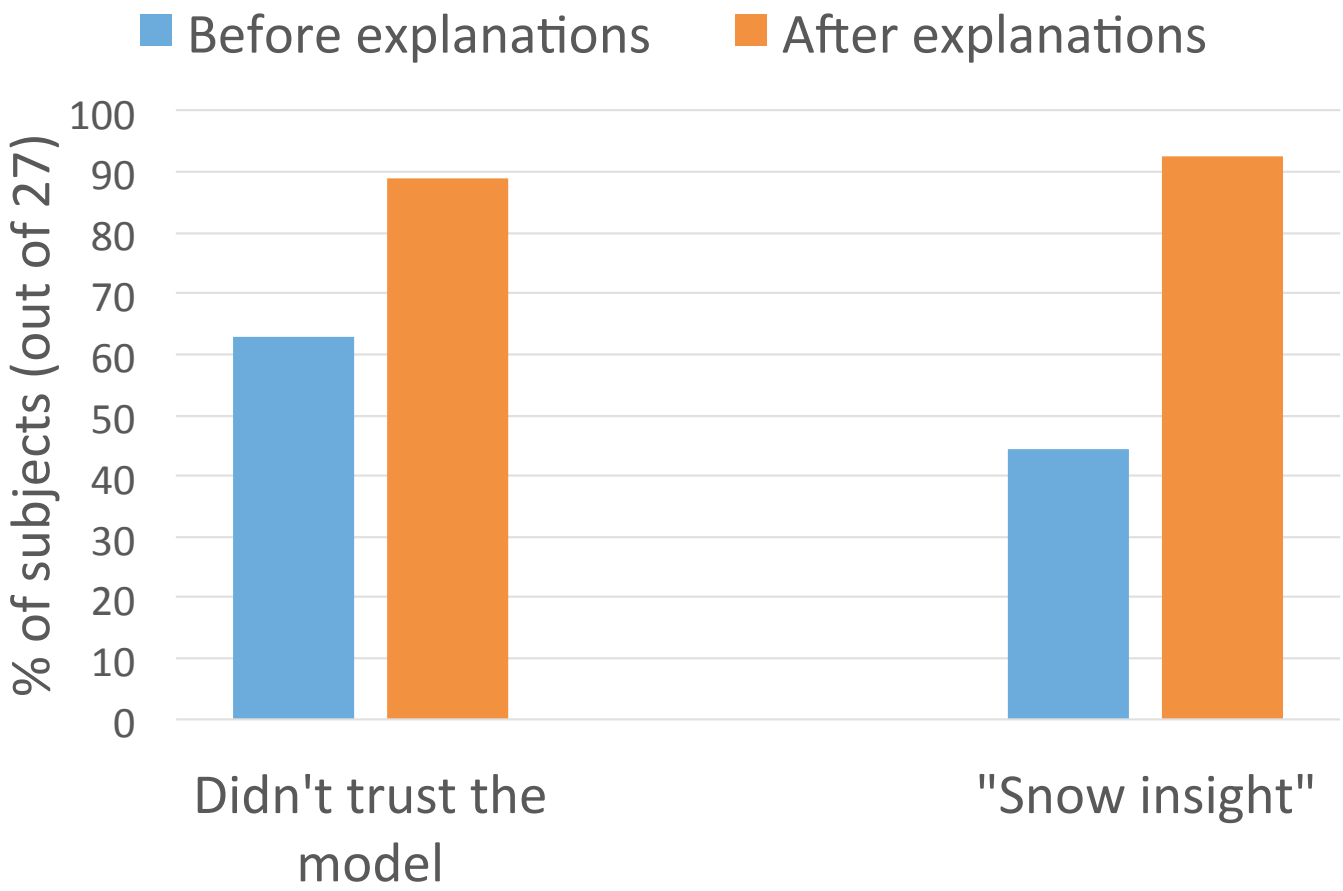
Image Example



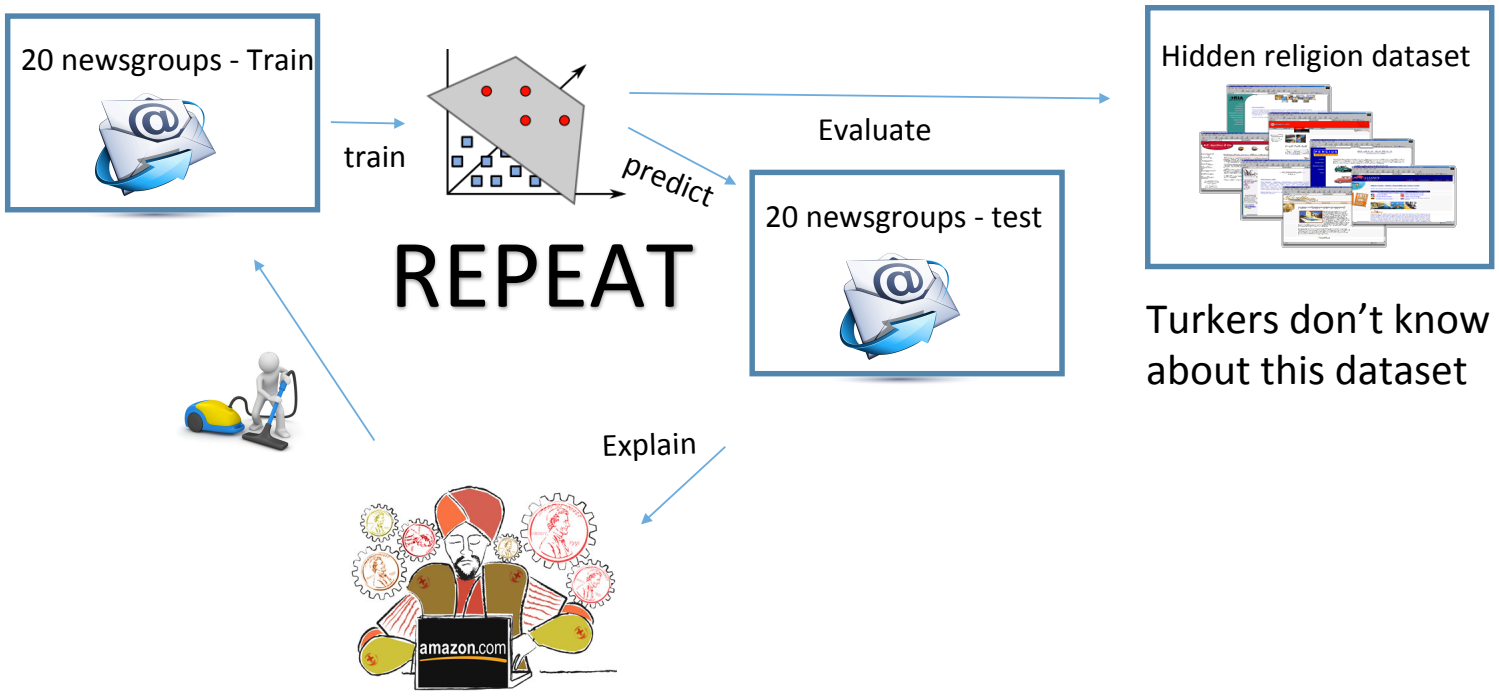
Is this a good Classifier?



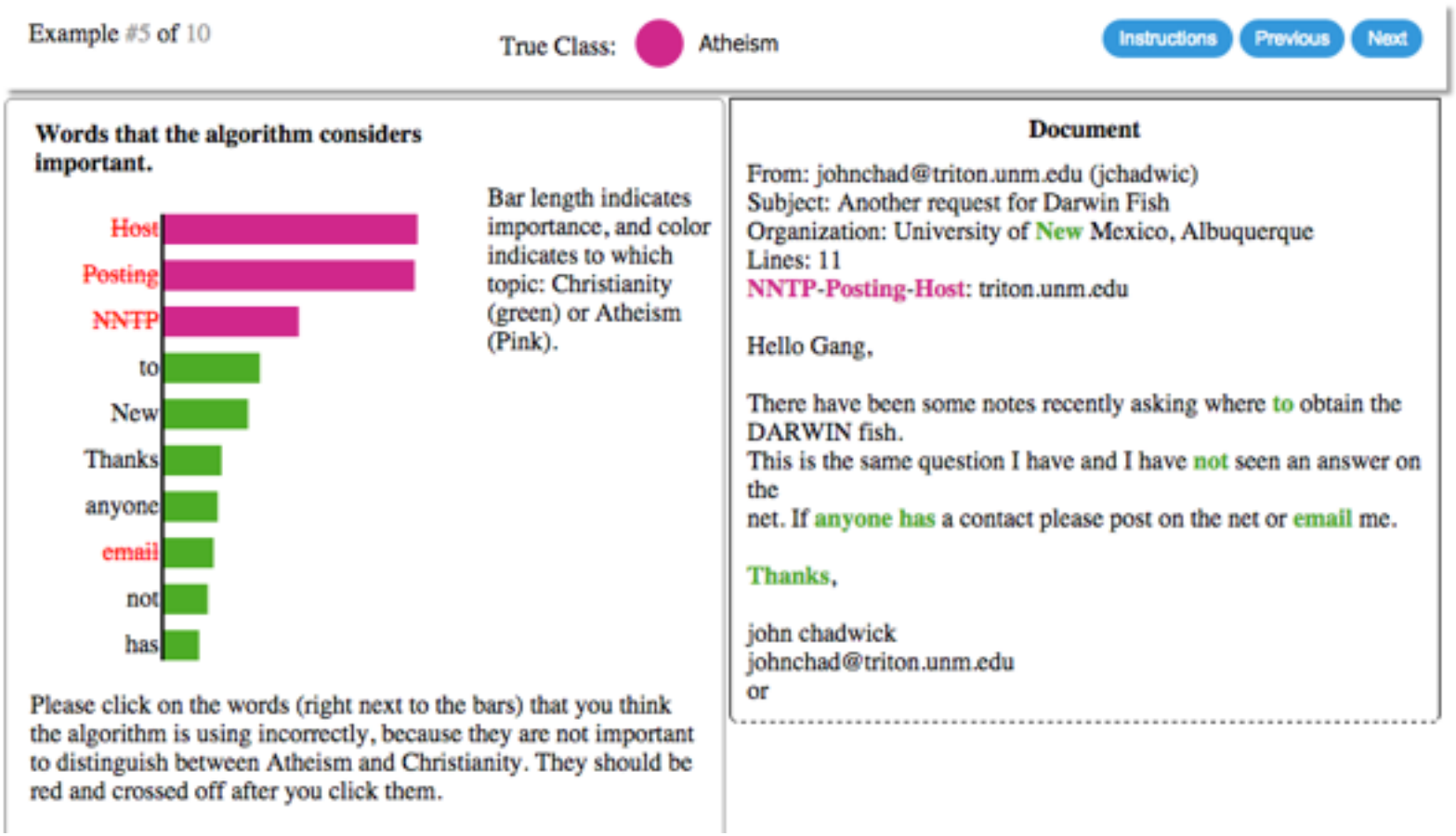
Is this a good Classifier?



Improving ML Algorithms



Improving ML Algorithms



Improving ML Algorithms