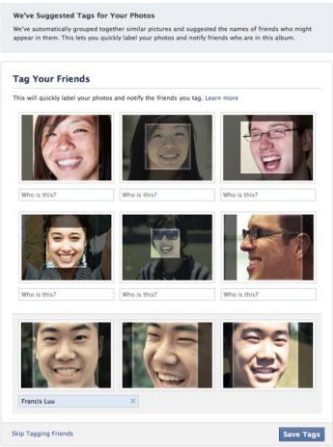
Explaining Black-Box Machine Learning Predictions

Sameer Singh

University of California, Irvine

work with Marco T. Ribeiro and Carlos Guestrin

Machine Learning is Everywhere…



Google Translate

NETFLIX

Classification: Wolf or a Husky?

Adopt or not?

Machine Learning Model

Wolf!

Classification: Wolf or a Husky?

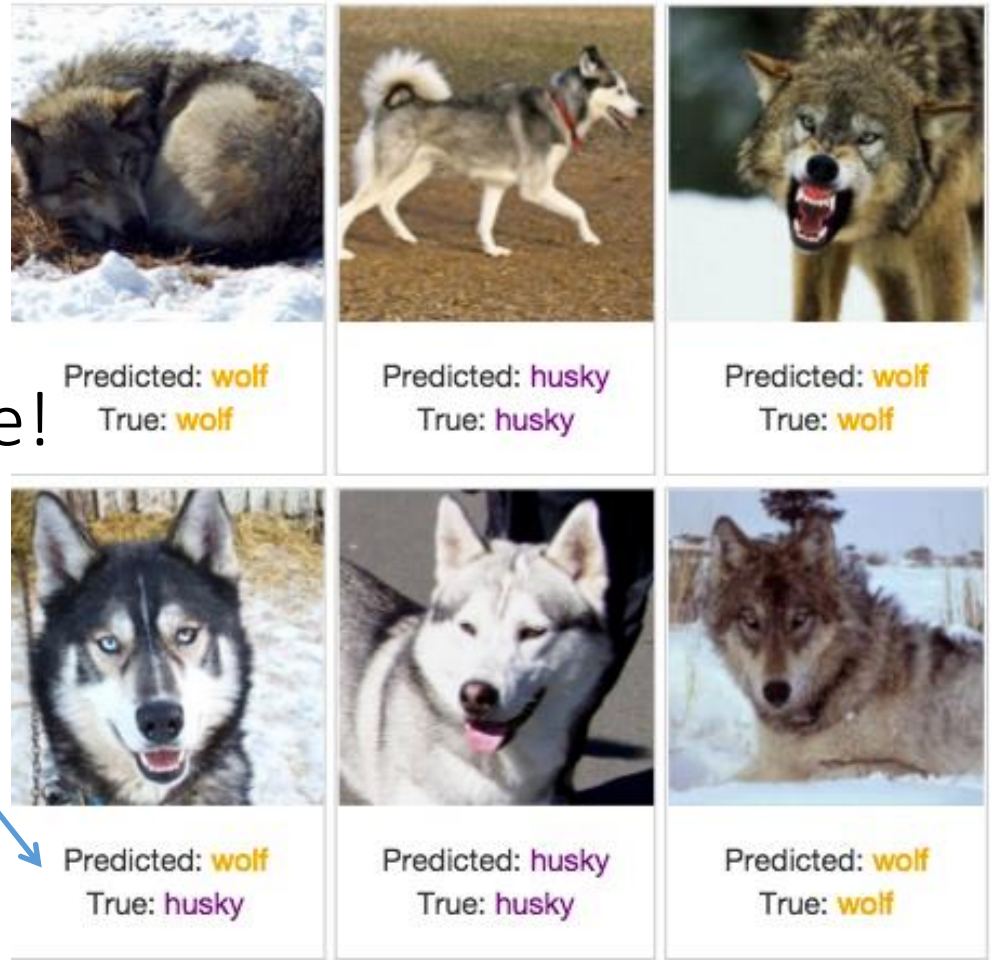
Adopt or not?

Machine Learning Model

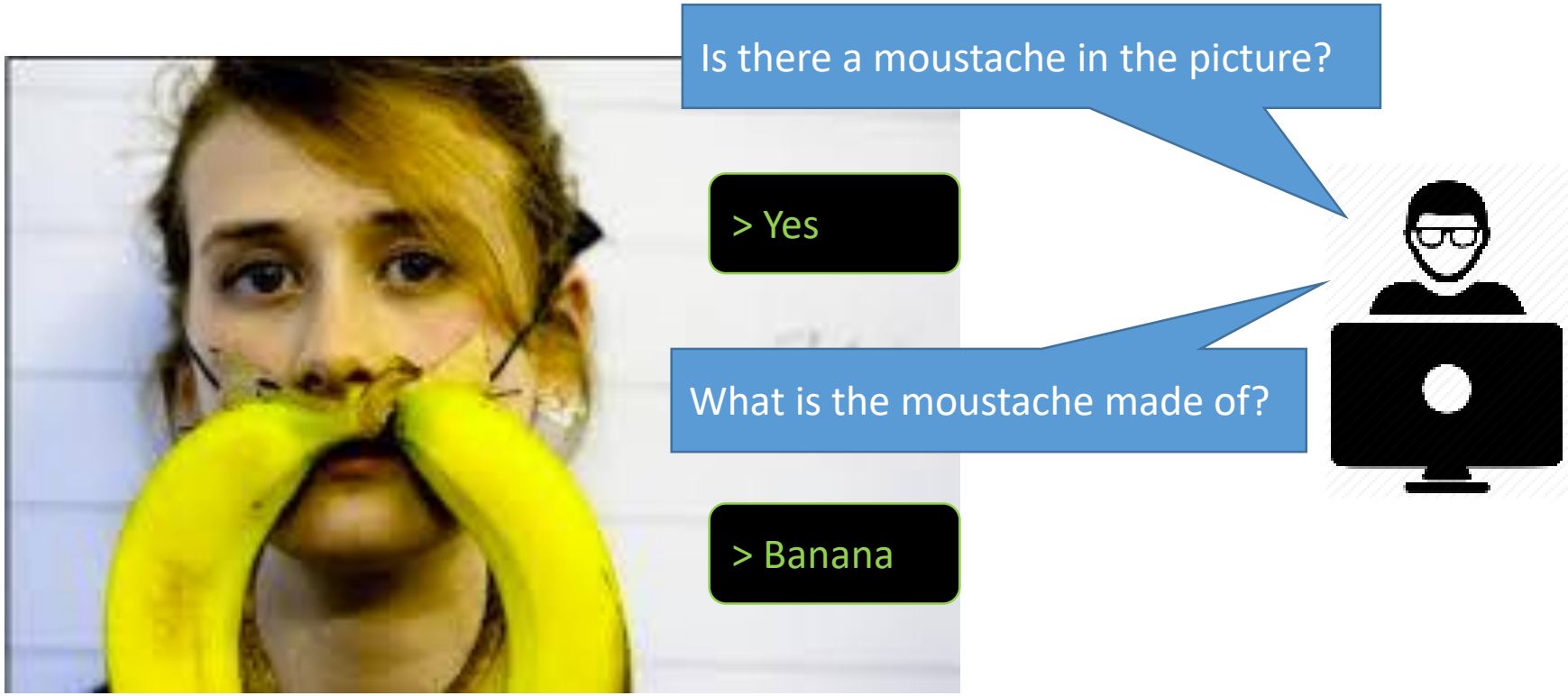
Husky!

Classification: Wolf or a Husky?

Only 1 mistake!



Visual Question Answering

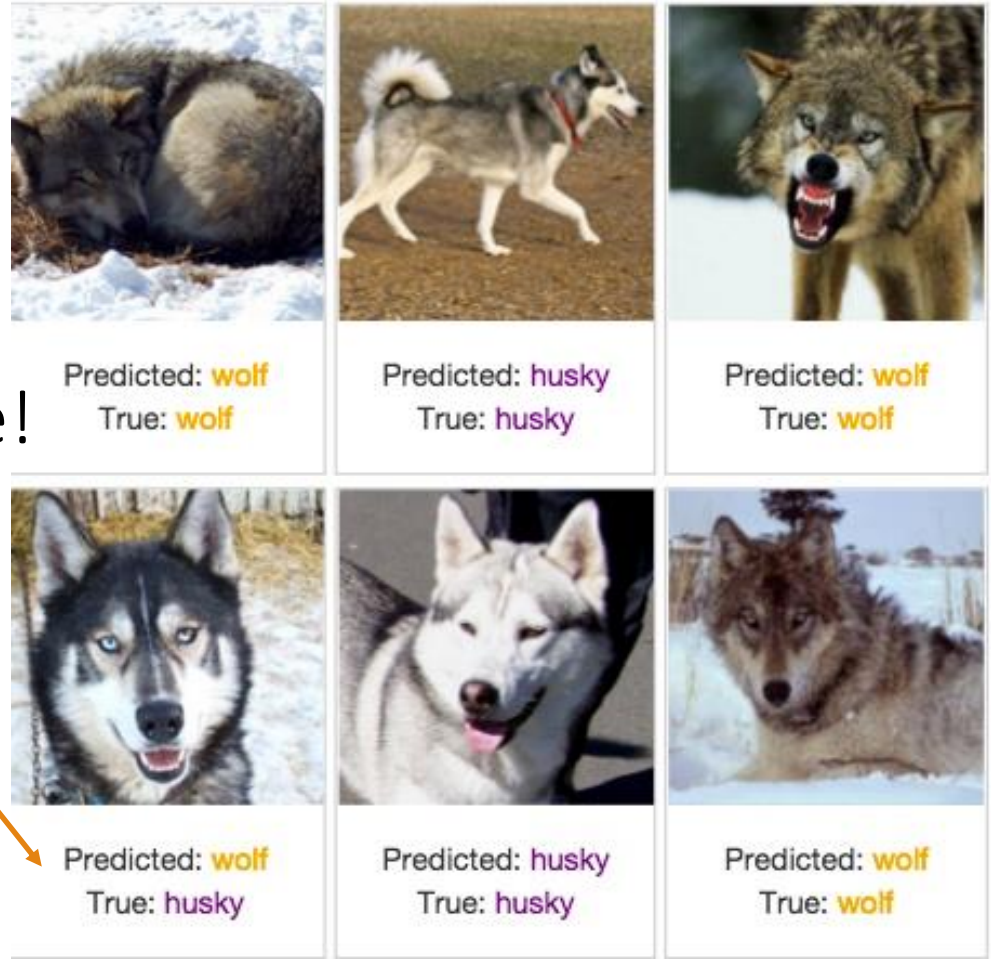


Essentially black-boxes!

Trust	Predict	Improve
How can we trust the predictions are correct?	How can we understand and predict the behavior?	How do we improve it to prevent potential mistakes?

Classification: Wolf or a Husky?

Only 1 mistake!



We've built a snow detector…

Slate

VIDEO SLATE IN MOTION. OCT. 14 2016 3:18 PM

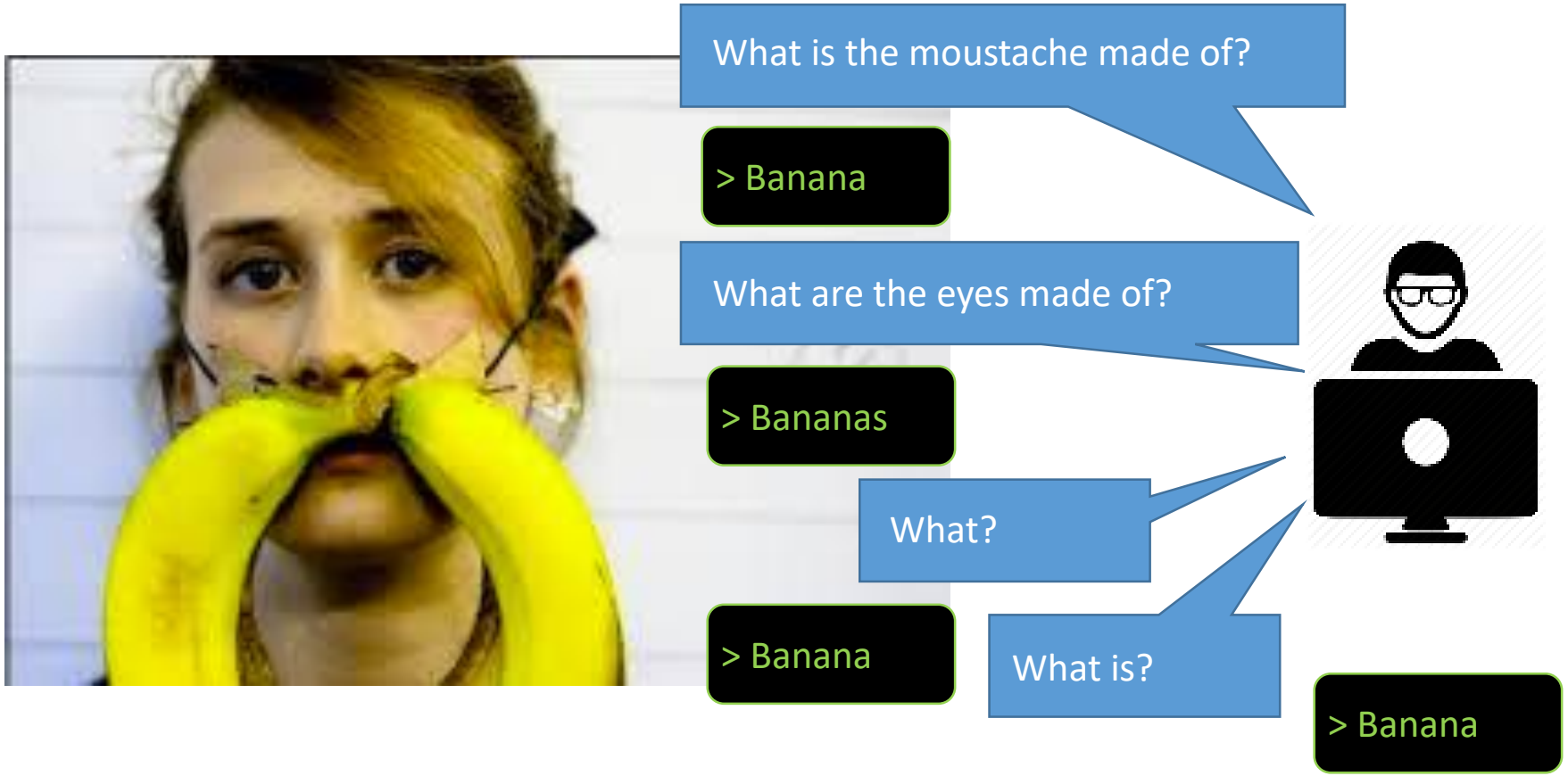
The Man Who Accidentally Adopted a Wolf Pup

It did not go well.

By A.J. McCarthy

10k 547 6

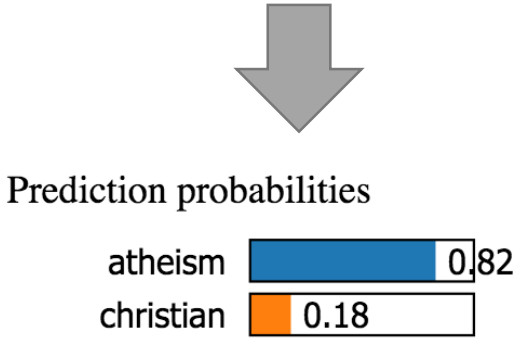
Visual Question Answering



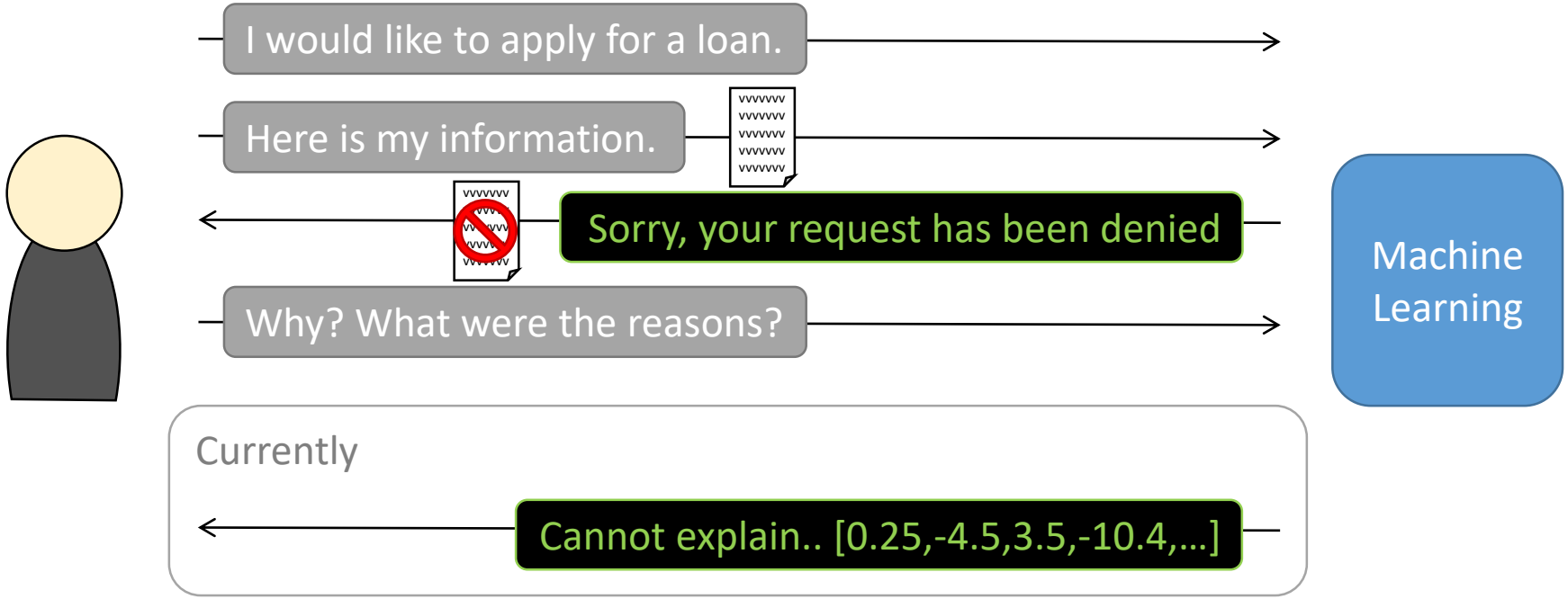
Text Classification

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

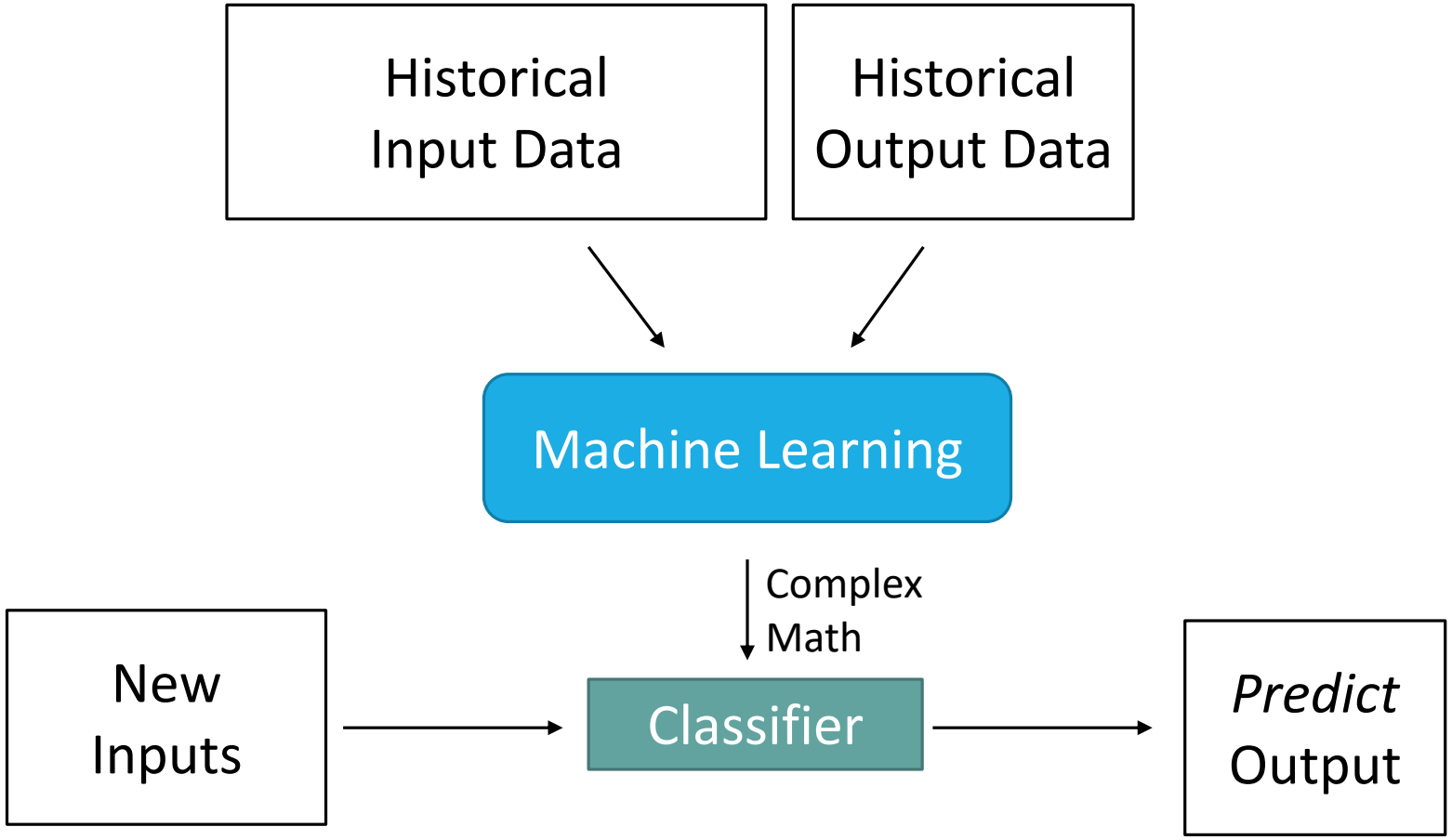


Loan Applications (in a Blackbox-ML World)

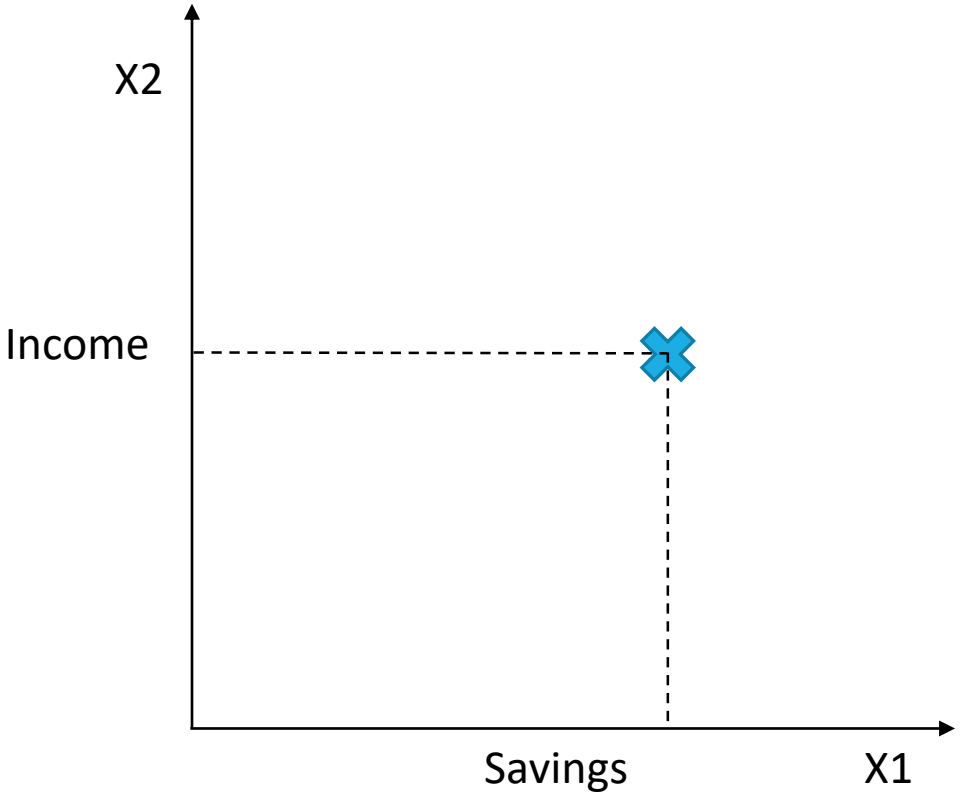


How did we get here?

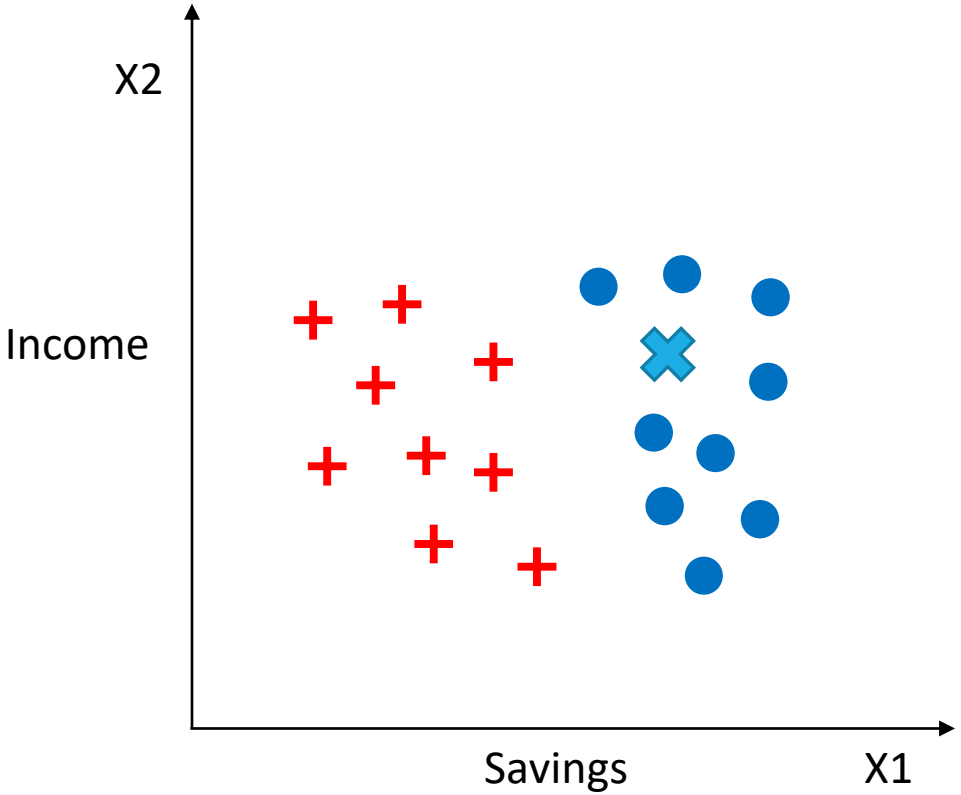
What is Machine Learning?



Should I give out a loan?



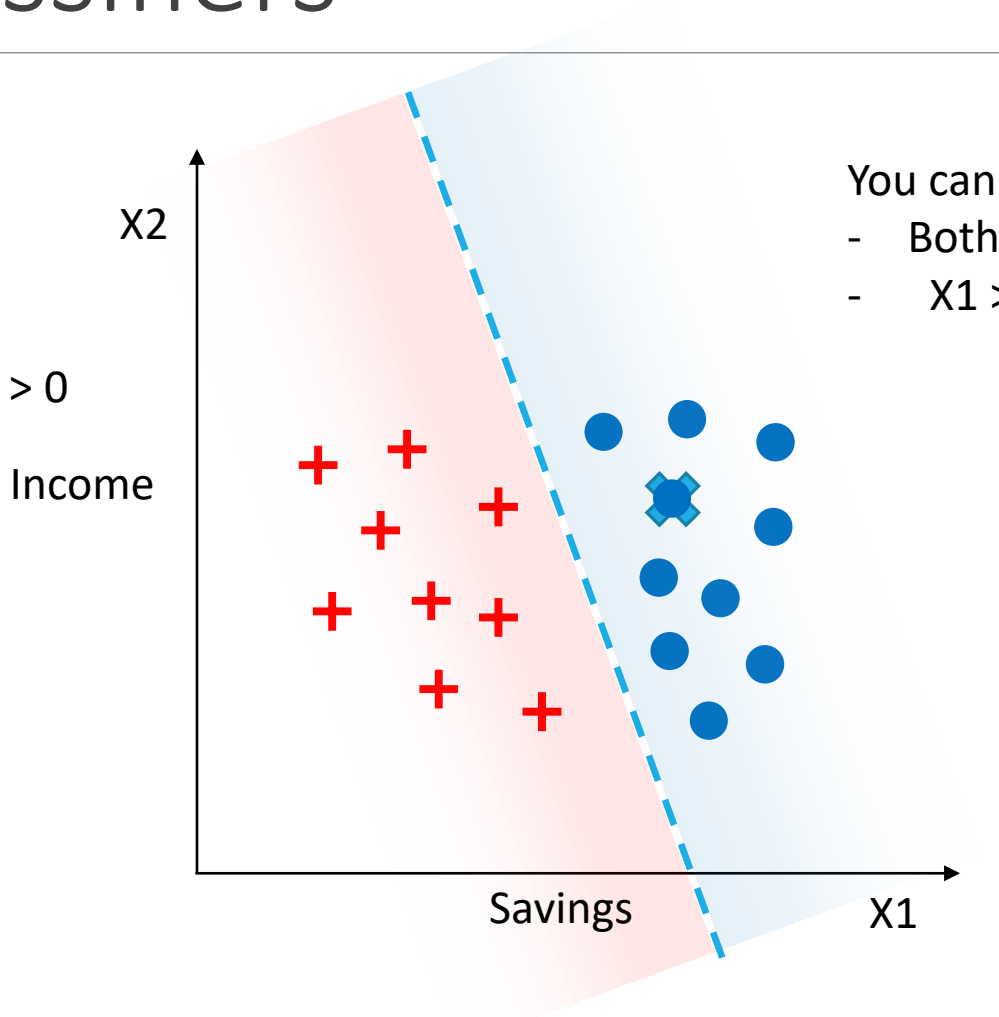
Get Historical Data



Linear Classifiers

if: 10X1 + X2 - 5 > 0

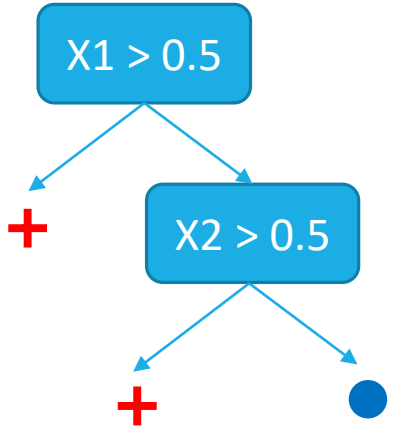
otherwise

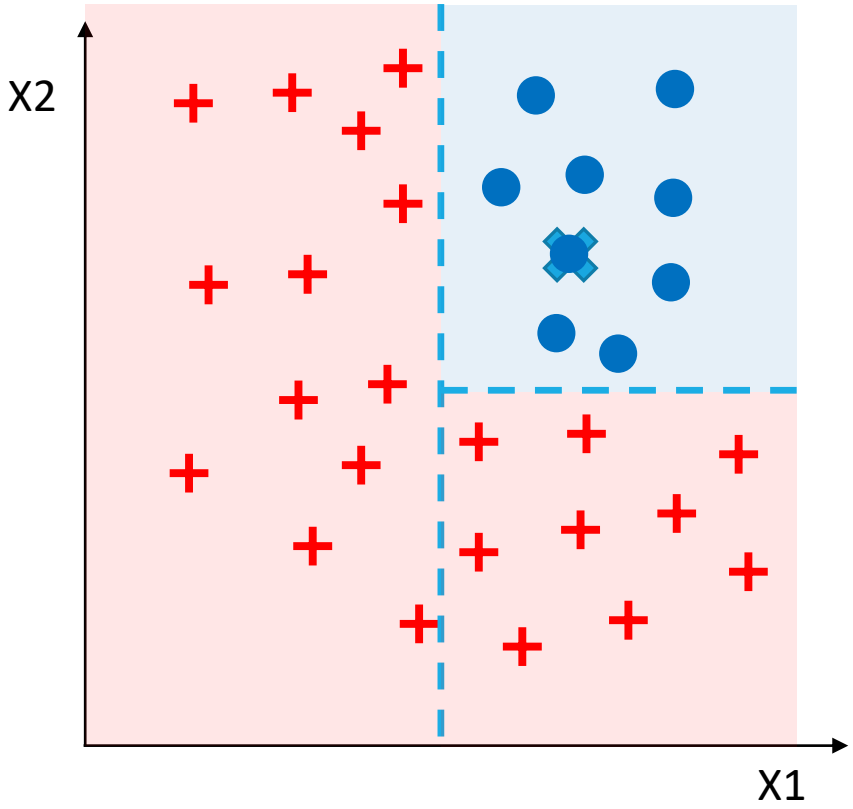


You can interpret it…

- Both have a positive effect
- X1 > X2

Decision trees





You can interpret it…

- X2 is irrelevant if X1<0.5
- Otherwise X2 is enough

Looking at the structure

Trust

How can we trust the predictions are correct?

Test whether the structure agrees with our intuitions.

Predict

How can we understand and predict the behavior?

Structure tells us exactly what will happen on any data.

Improve

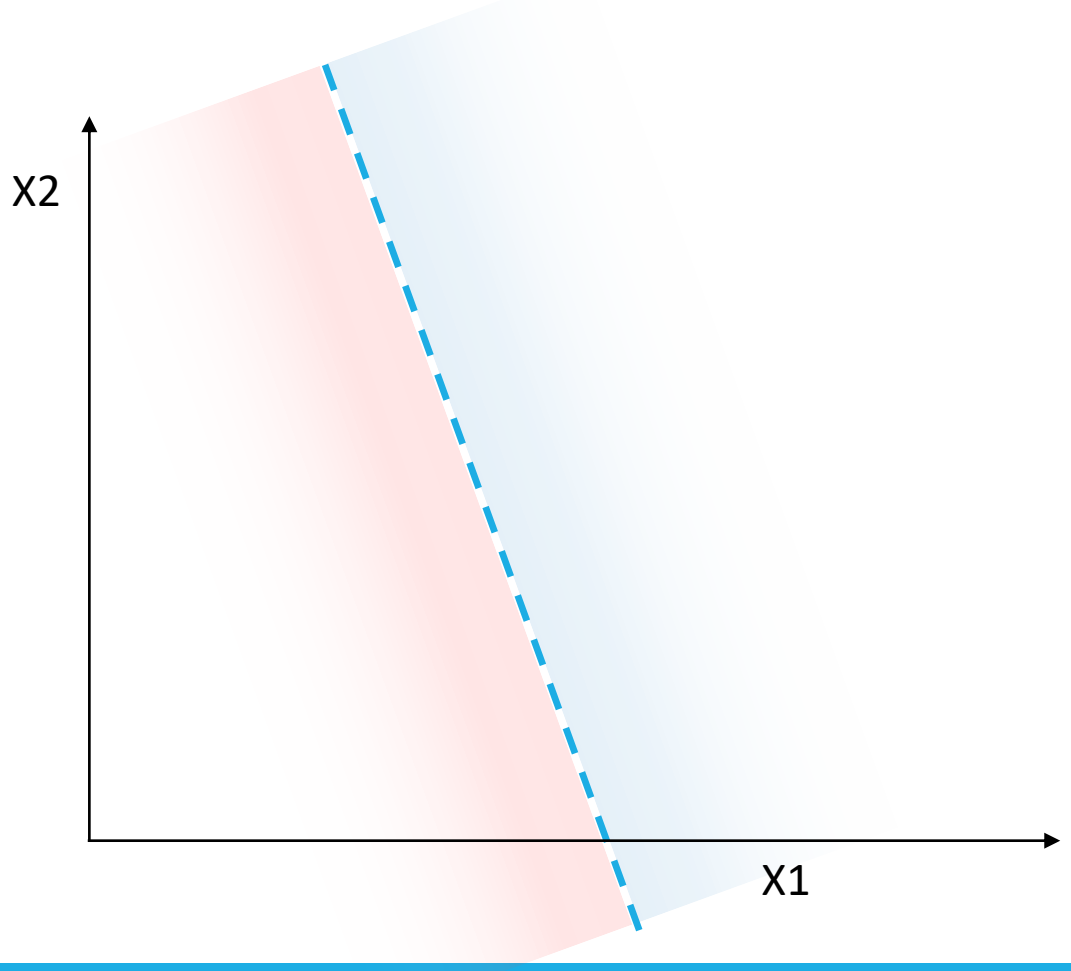
How do we improve it to prevent potential mistakes?

Structure tells you where the error is, thus how to fix it.

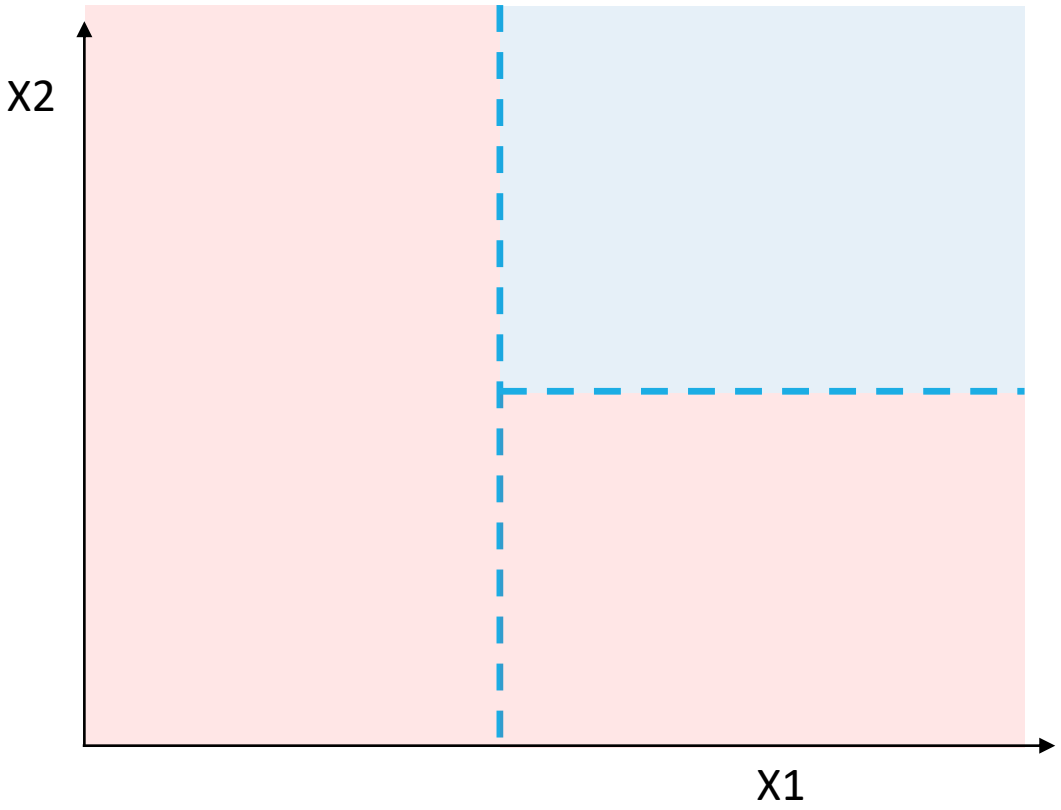
Arrival of Big Data

Big Data: Applications of ML

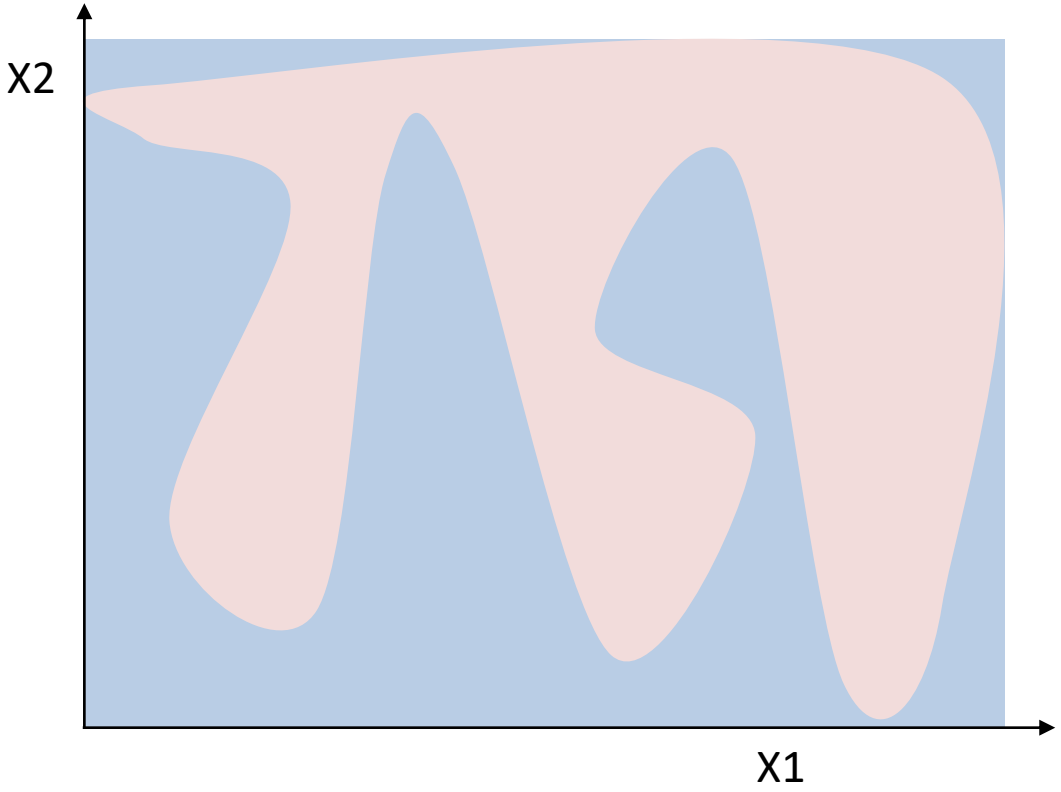
Big Data: More Complexity

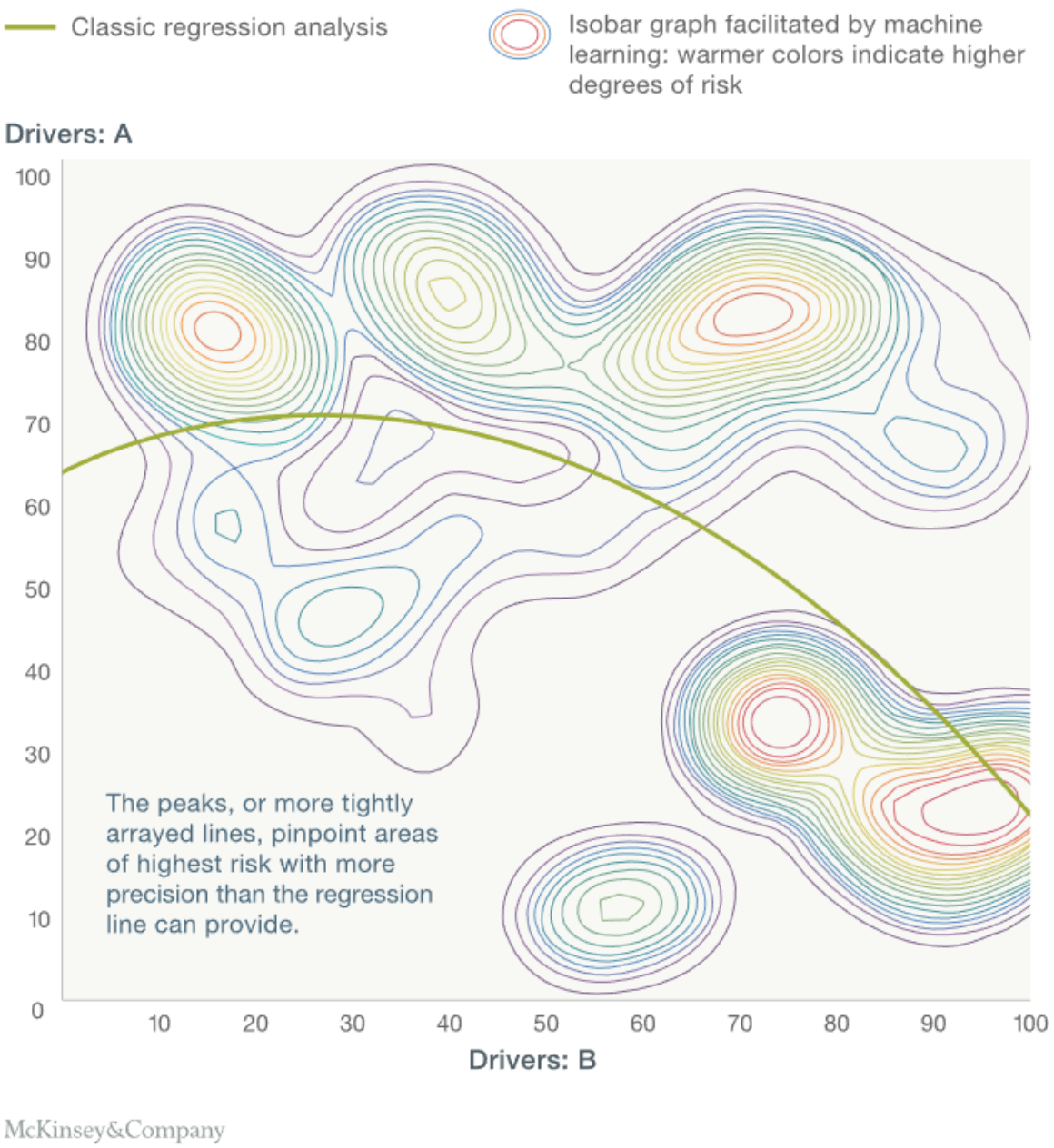


Big Data: More Complexity



Big Data: More Complexity





http://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning

Big Data: More Dimensions

Savings

Income

Credit scores

Past defaults

Loan Amount

Recent defaults

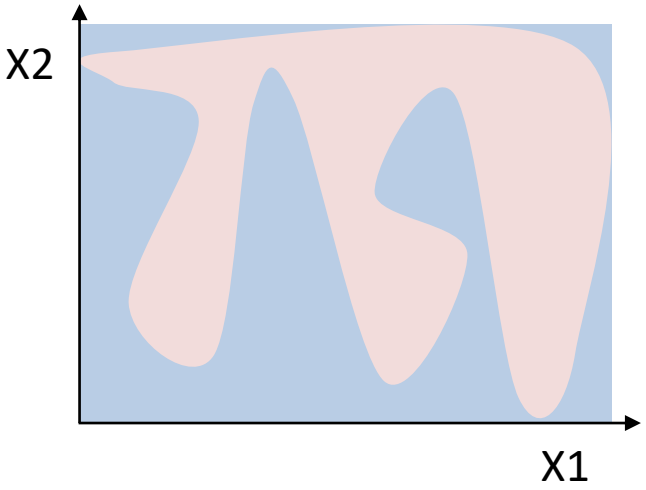
Profession

Age

Marital Status

This easily goes to hundreds

- Images: thousands
- Text: tens of thousands
- Video: millions
- … and so on



Savings

Income

Married

Loan Amount

Past defaults

Profession

Credit scores

Age

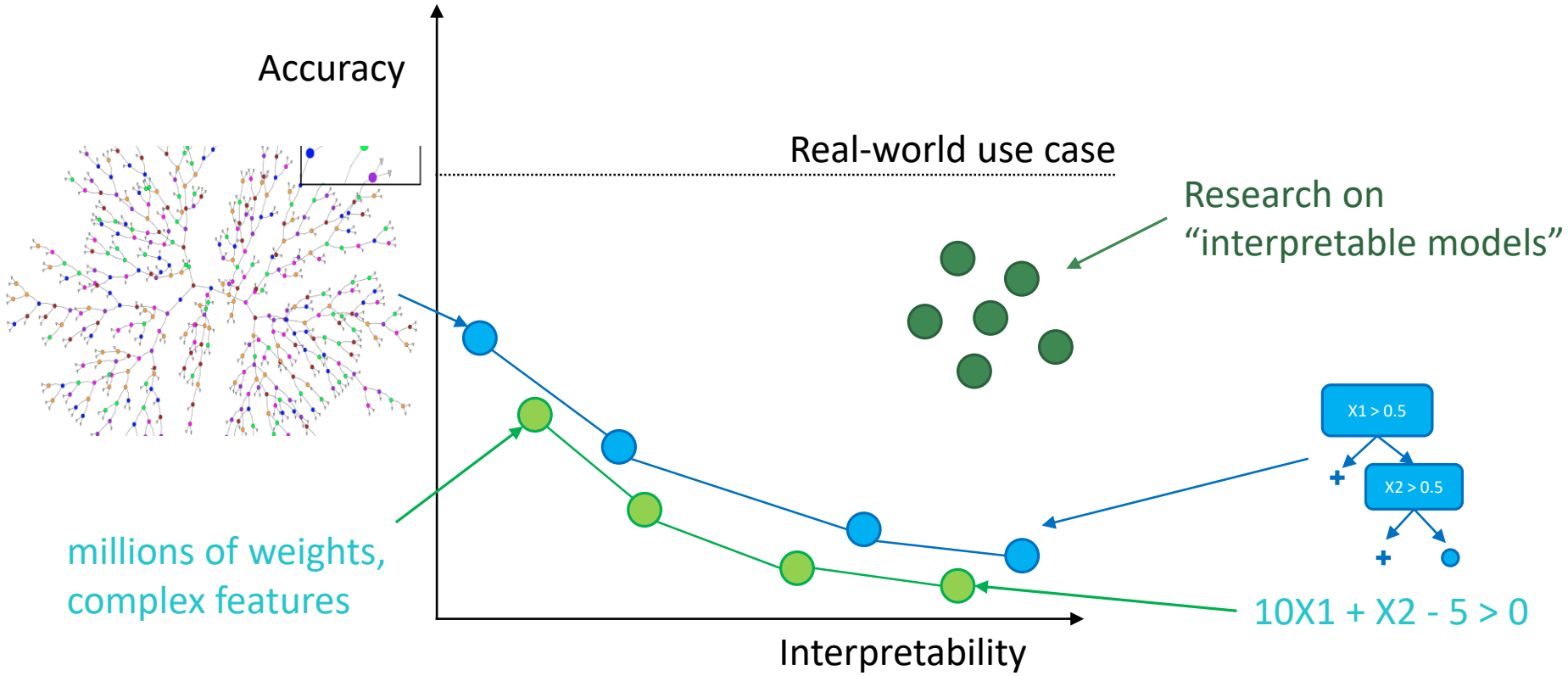
Recent defaults

…

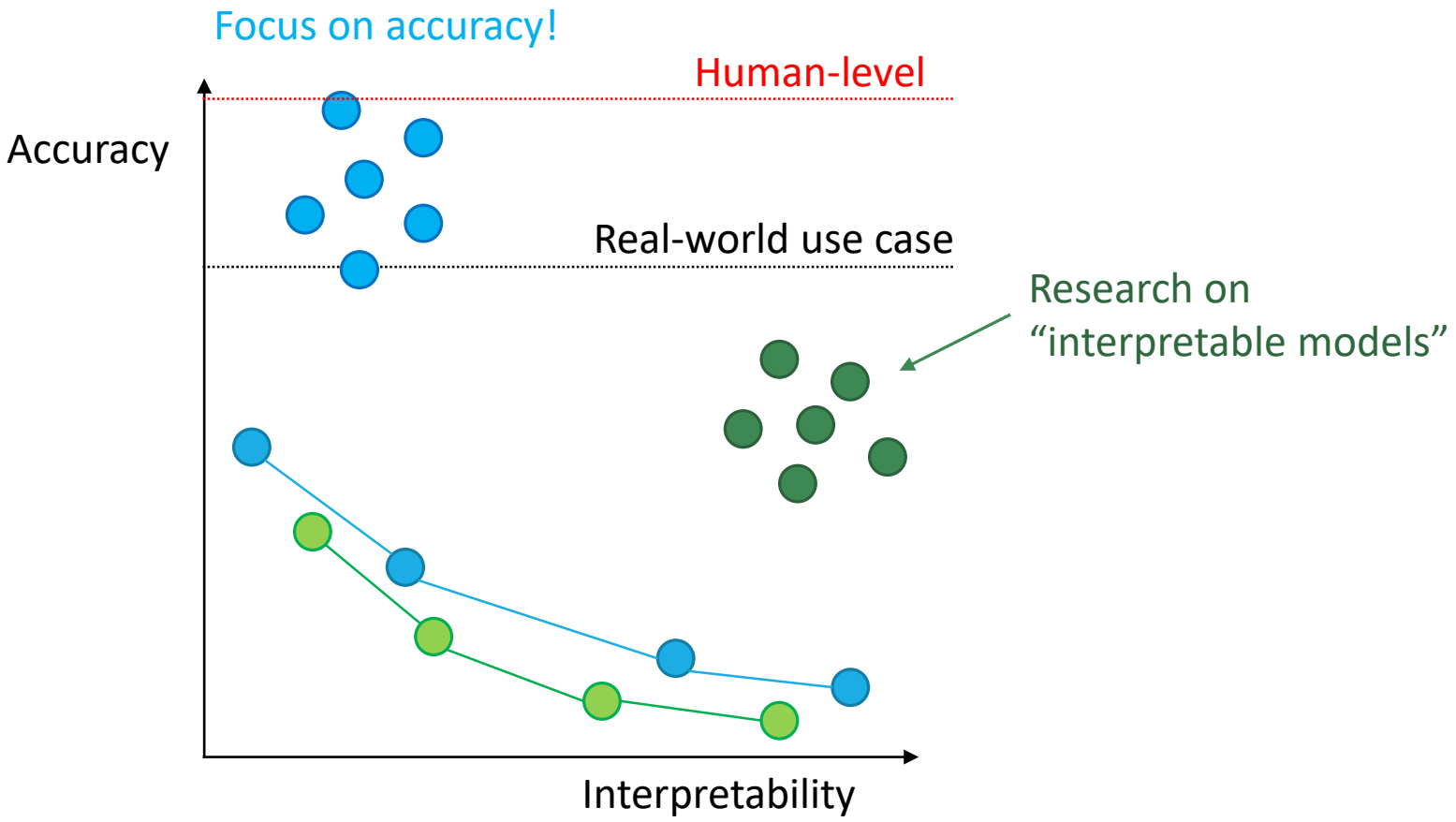
Complex Surfaces + Lots of dimensions

Black-boxes!

Accuracy vs Interpretability



Deep Learning



Looking at the structure

Trust

How can we trust the predictions are correct?

Test whether the structure agrees with our intuitions.

Predict

How can we understand and predict the behavior?

Structure tells us exactly what will happen on any data.

Improve

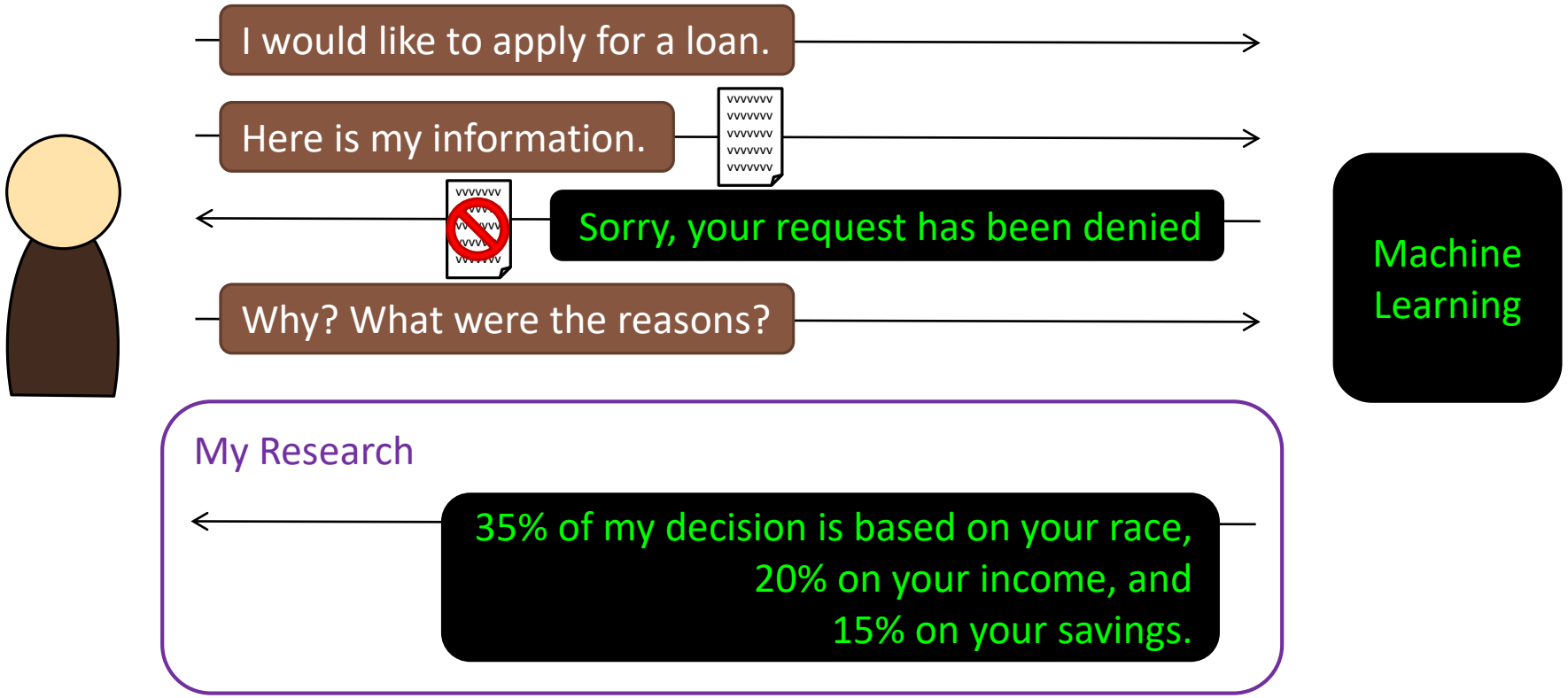
How do we improve it to prevent potential mistakes?

Structure tells you where the error is, thus how to fix it.

Explaining Predictions

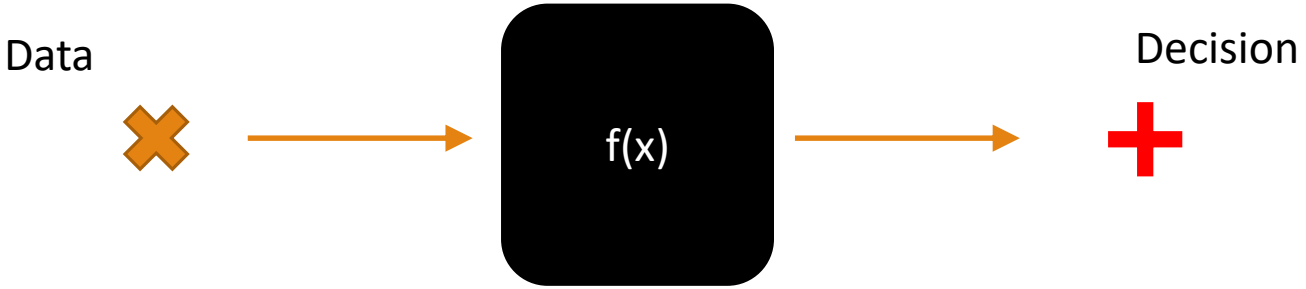
The LIME Algorithm

Applying for a Loan



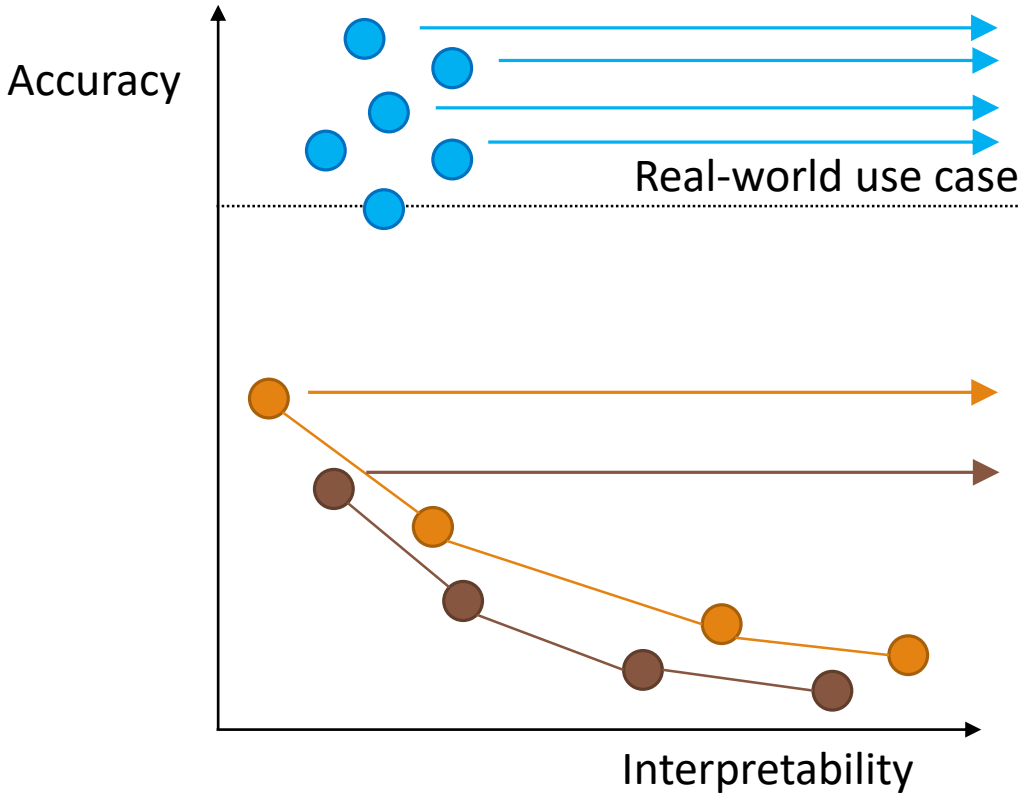
Being Model-Agnostic…

No assumptions about the internal structure…



Explain any existing, *or future*, model

LIME: Explain Any Classifier!

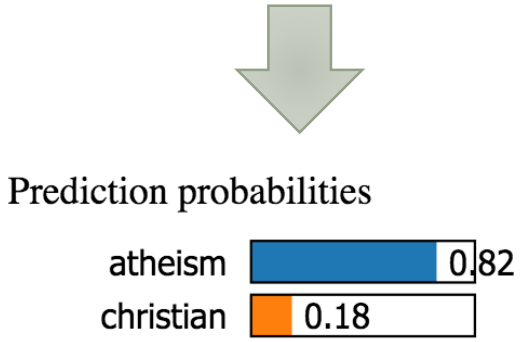


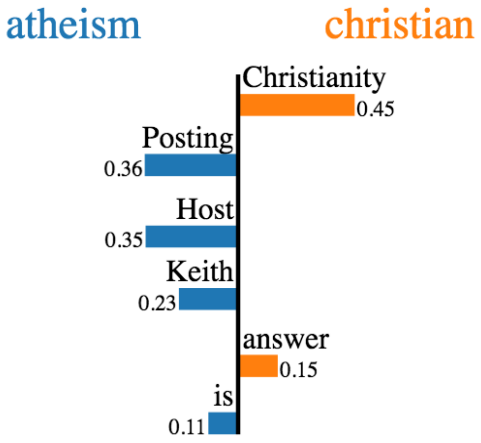
Make everything interpretable!

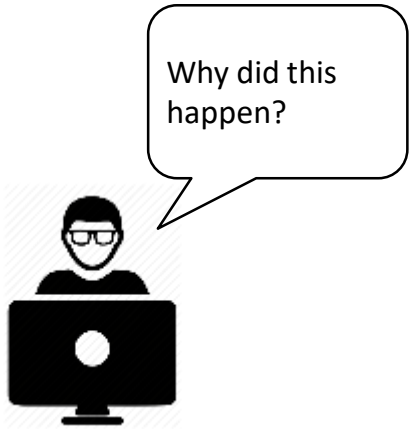
What is an “Explanation”?

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

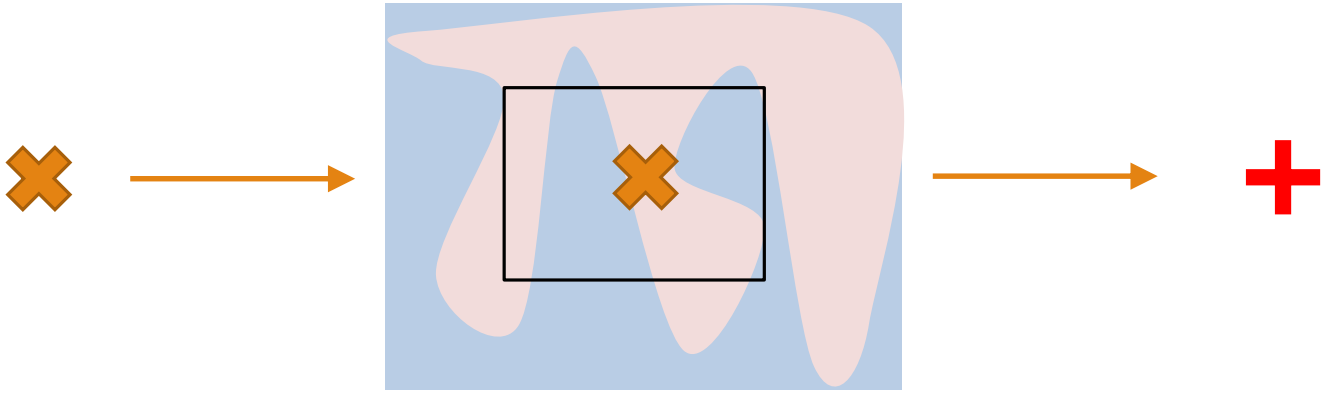






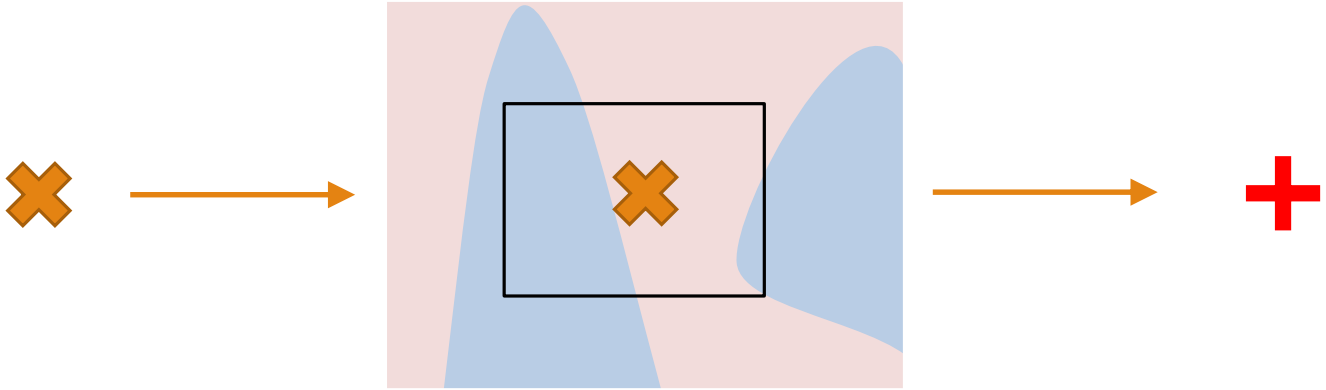
Being Model-Agnostic…

“Global” explanation is too complicated



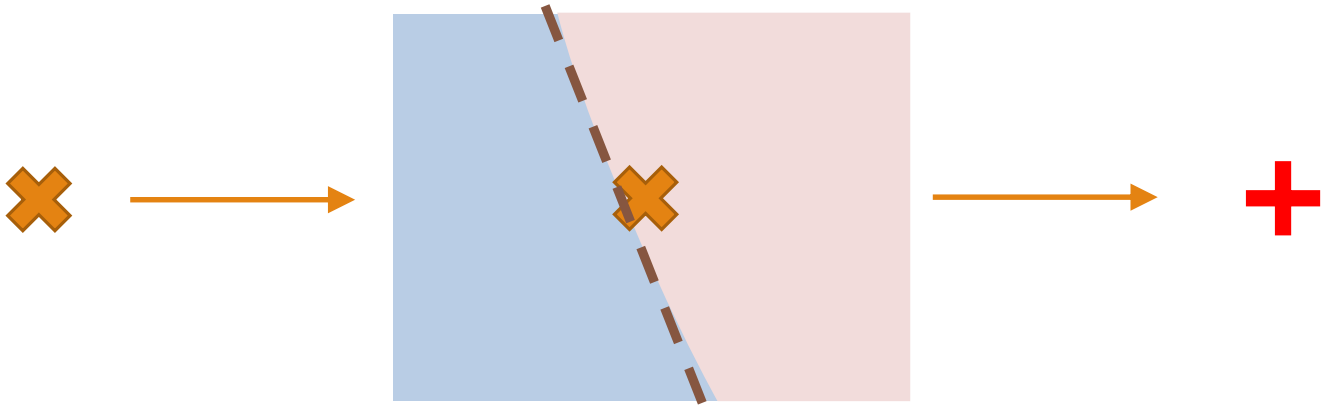
Being Model-Agnostic…

"Global" explanation is too complicated



Being Model-Agnostic…

"Global" explanation is too complicated

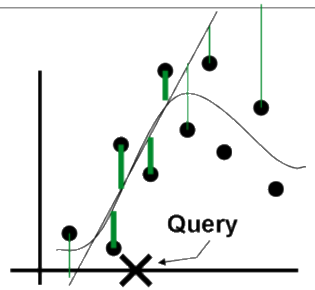


Explanation is an interpretable model, that is locally accurate

Example – Image Classification

Original Image
P(labrador) = 0.21

Perturbed Instances	P(Labrador)
	0.92
	0.001
	0.34

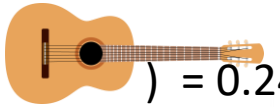


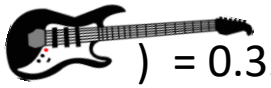
Locally weighted regression

Explanation

Google’s Object Detector

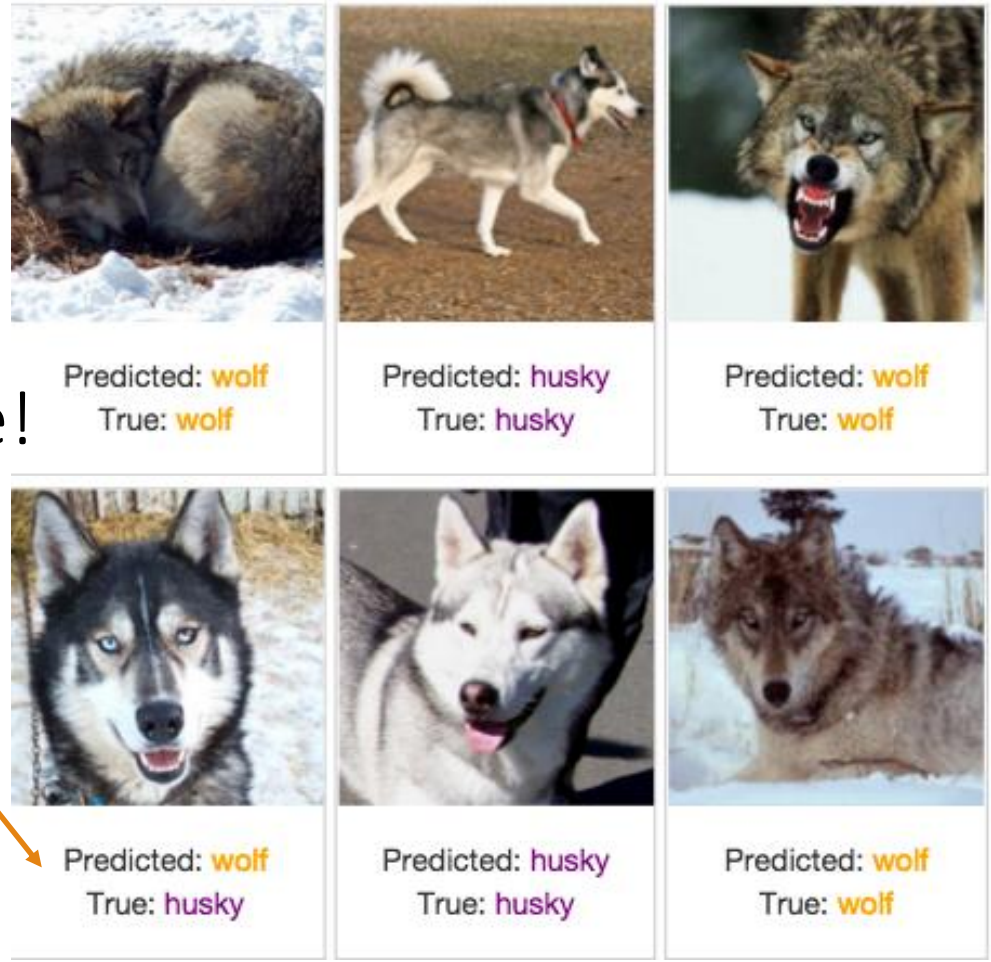
P() = 0.21

P() = 0.24

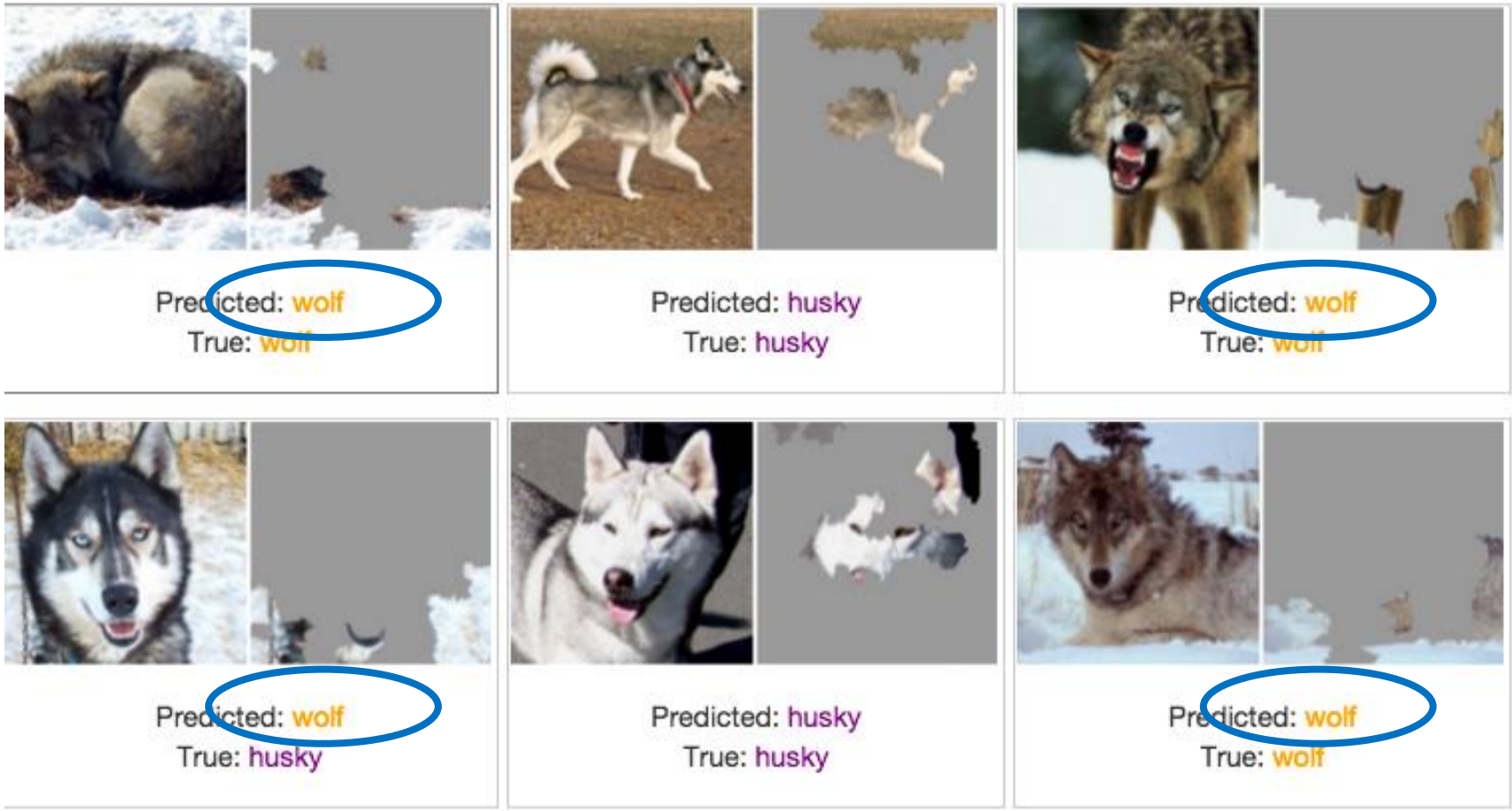
P() = 0.32

Classification: Wolf or a Husky?

Only 1 mistake!



Neural Network Explanations



We've built a great snow detector…

Visual QA

What is the mustache made of?	banana
How **many** bananas are in the picture?	2

Neural Machine Translation

English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar.

Neural Machine Translation

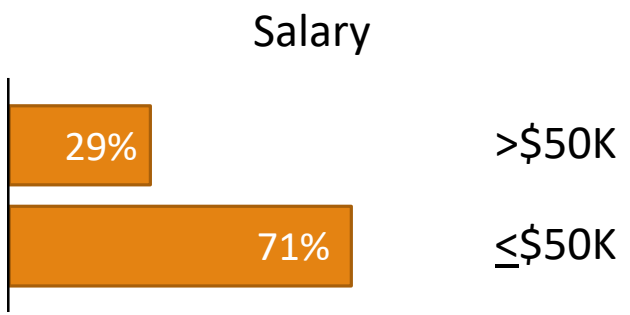
English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar.
This is the **problem** we must address	Este é o problema que temos que enfrentar.

Neural Machine Translation

English	Portuguese
This is the **question** we must address	Esta é a questão que temos que enfrentar.
This is the **problem** we must address	Este é o problema que temos que enfrentar.
This is what we must address	É isso que temos de enfrentar.

Salary Prediction

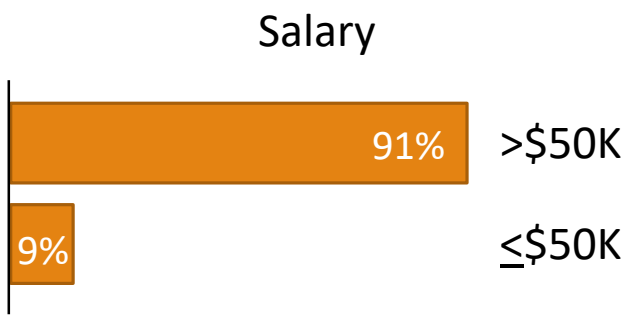
Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	$\leq$ High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	≤ 40
Country	United States



IF Education ≤ High School
Then Predict Salary ≤ 50K

Salary Prediction

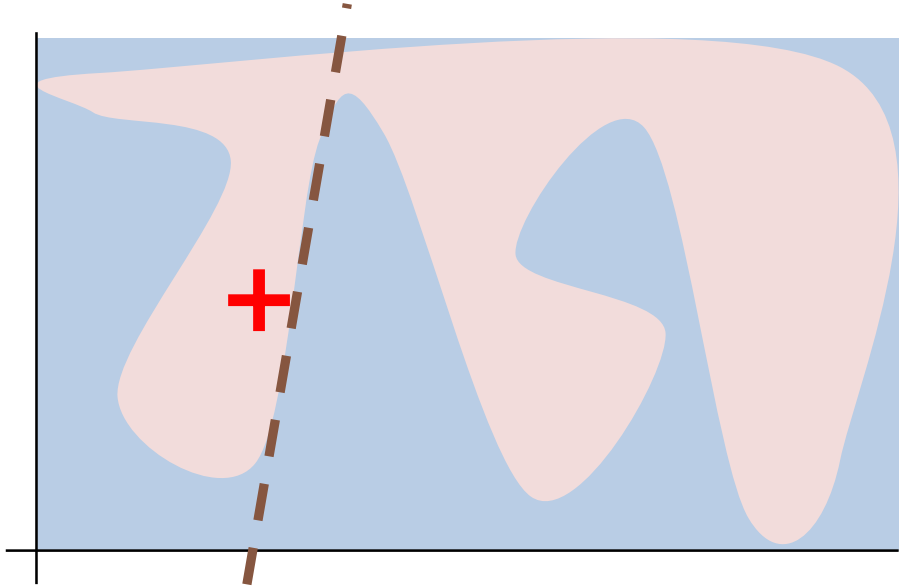
$28 < \text{Age} \leq 37$
Workclass = Private
Education = Doctorate
Marital Status = Married
Occupation = Professional
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = None
Hours per week > 45.00
Country = United-States



IF Married **and**
Education = Doctorate
Then Predict Salary > 50K

Data from *the US Census*

“Global” Behavior



What about explaining the rest of the model?

Explaining Global behavior

LIME explains a single prediction

local behavior for a single instance

Can’t examine all explanations

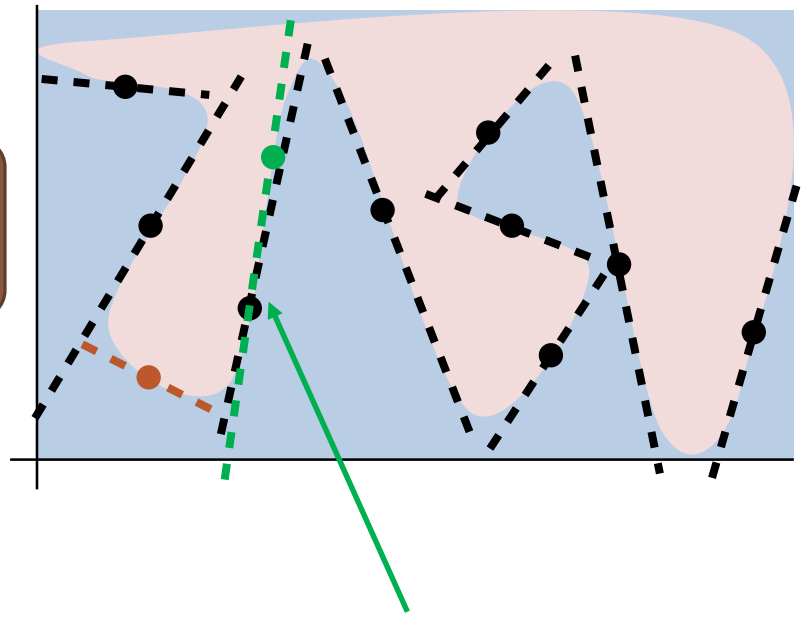
Instead pick *k* explanations to show to the user

Representative

Should summarize the model’s global behavior

Diverse

Should not be redundant in their descriptions



Are they useful?

Quantitative Evaluation

Understand what ML is doing

Compare different ML algorithms

Improve the existing model

Predict how ML will behave

Quantitative Evaluation

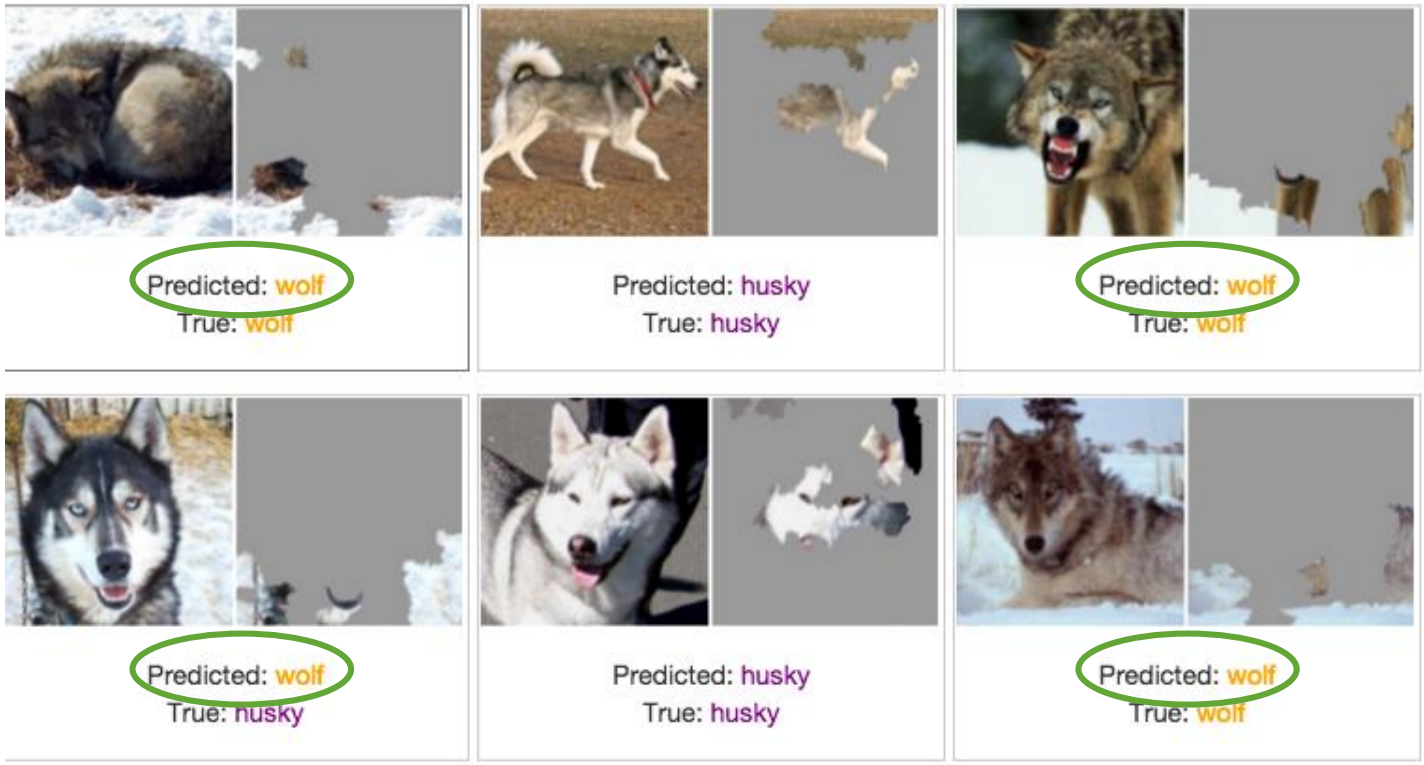
Understand what ML is doing

Compare different ML algorithms

Improve the existing model

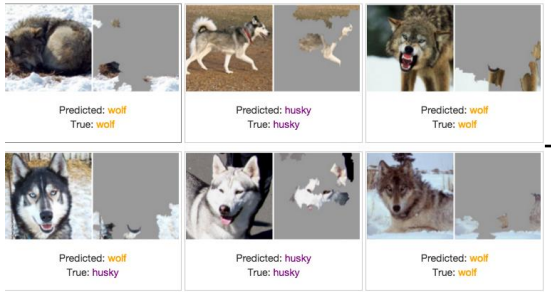
Predict how ML will behave

Understanding Behavior



We’ve built a great snow detector…

Understanding Behavior



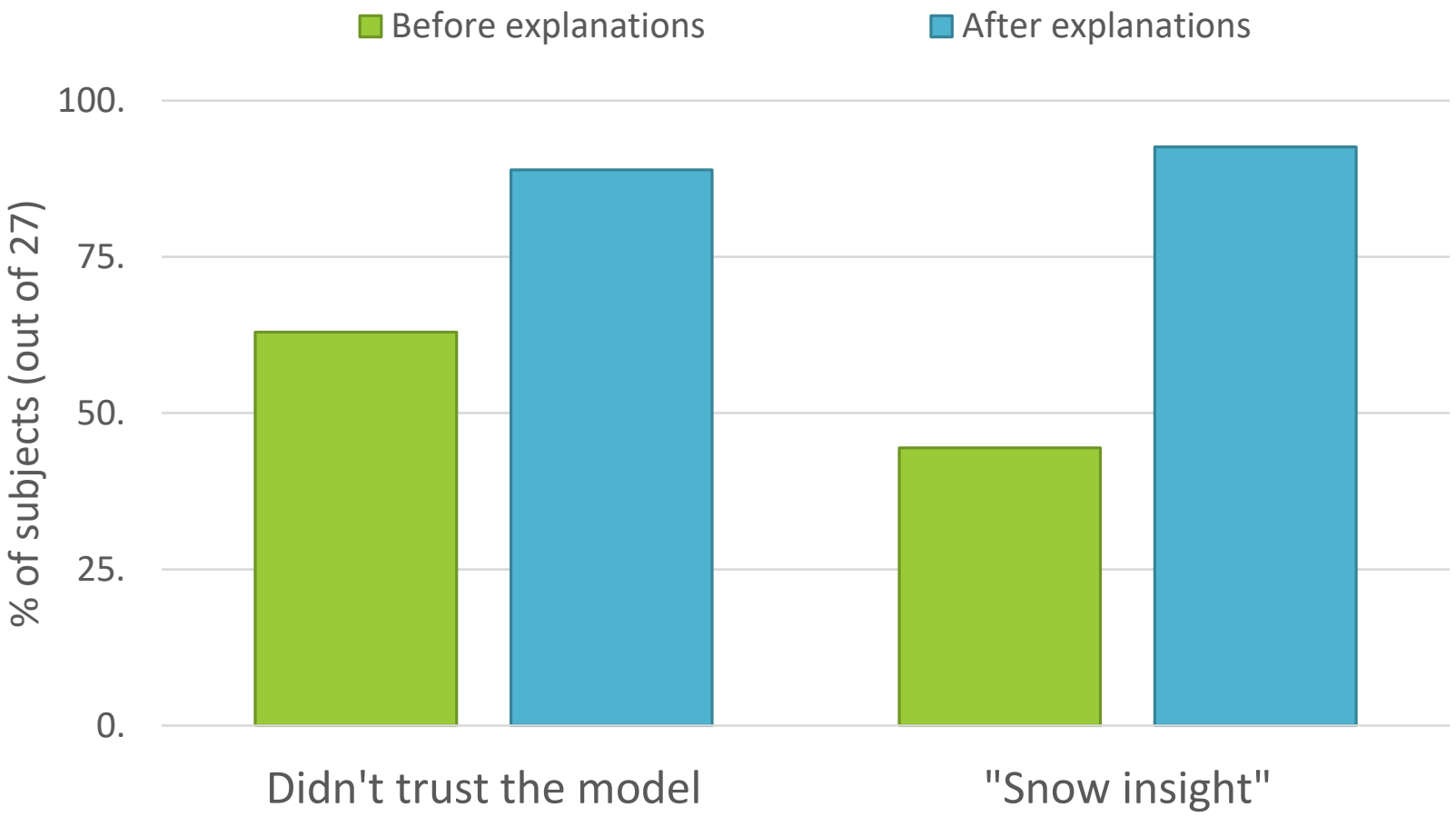
Question 1

Would you trust this model?

Question 2

What is the classifier is doing?

Did they notice it?



Quantitative Evaluation

Understand what ML is doing

Compare different ML algorithms

Improve the existing model

Predict how ML will behave

Comparing Classifiers

Classifier 1

Change the model
Different data
Different parameters
Different "features"
…

Classifier 2

Accuracy?

Look at Examples?

Deploy and Check?

"I have a gut feeling.."

Explanations?

Comparing Classifiers

Original Image

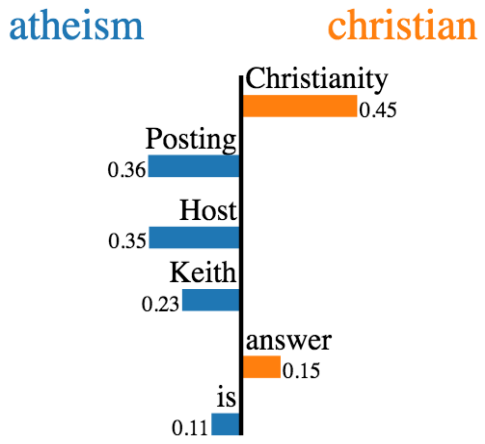
“Bad” Classifier

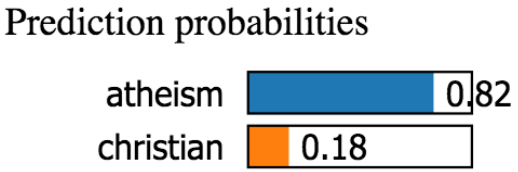
“Good” Classifier

Explanation for a bad classifier

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

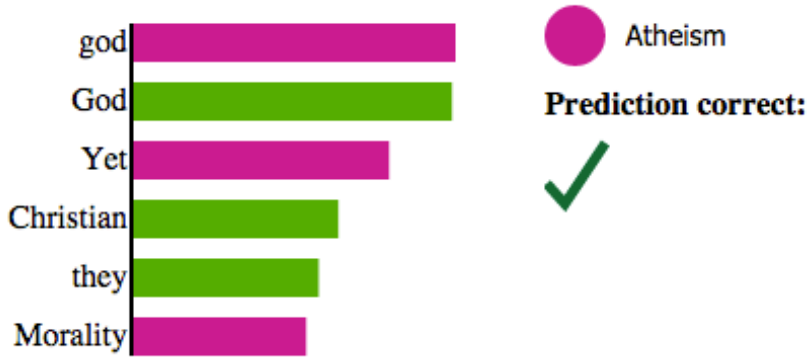
I think Christianity is the one true religion.
If you'd like to know more, send me a note





After looking at the explanation, we shouldn't trust the model!

“Good” Explanation

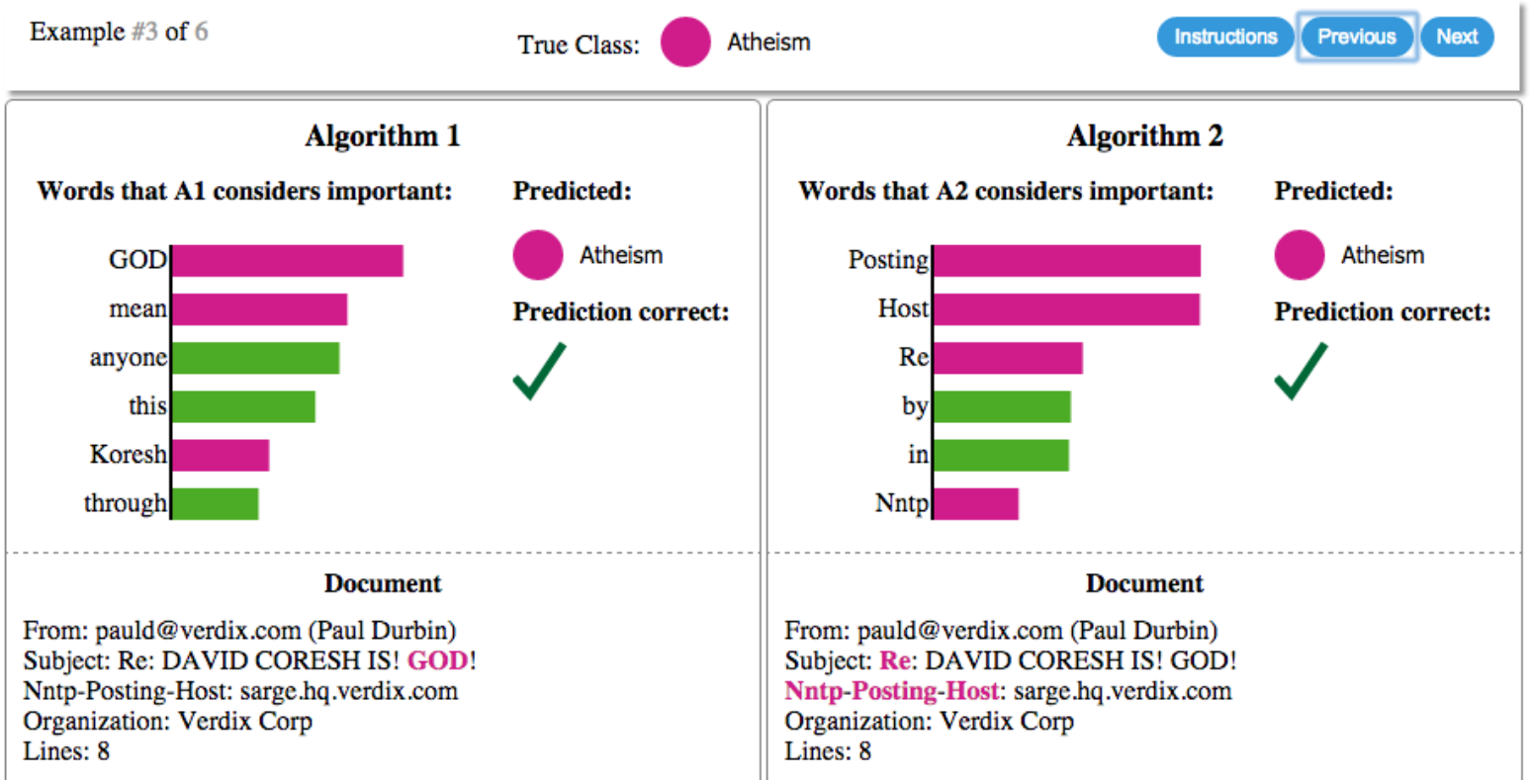


From: arromdee@jyusenkyou.cs.jhu.edu (Ken Arromdee)
Subject: Re: Christian Morality is
Organization: Johns Hopkins University CS Dept.
Lines: 24

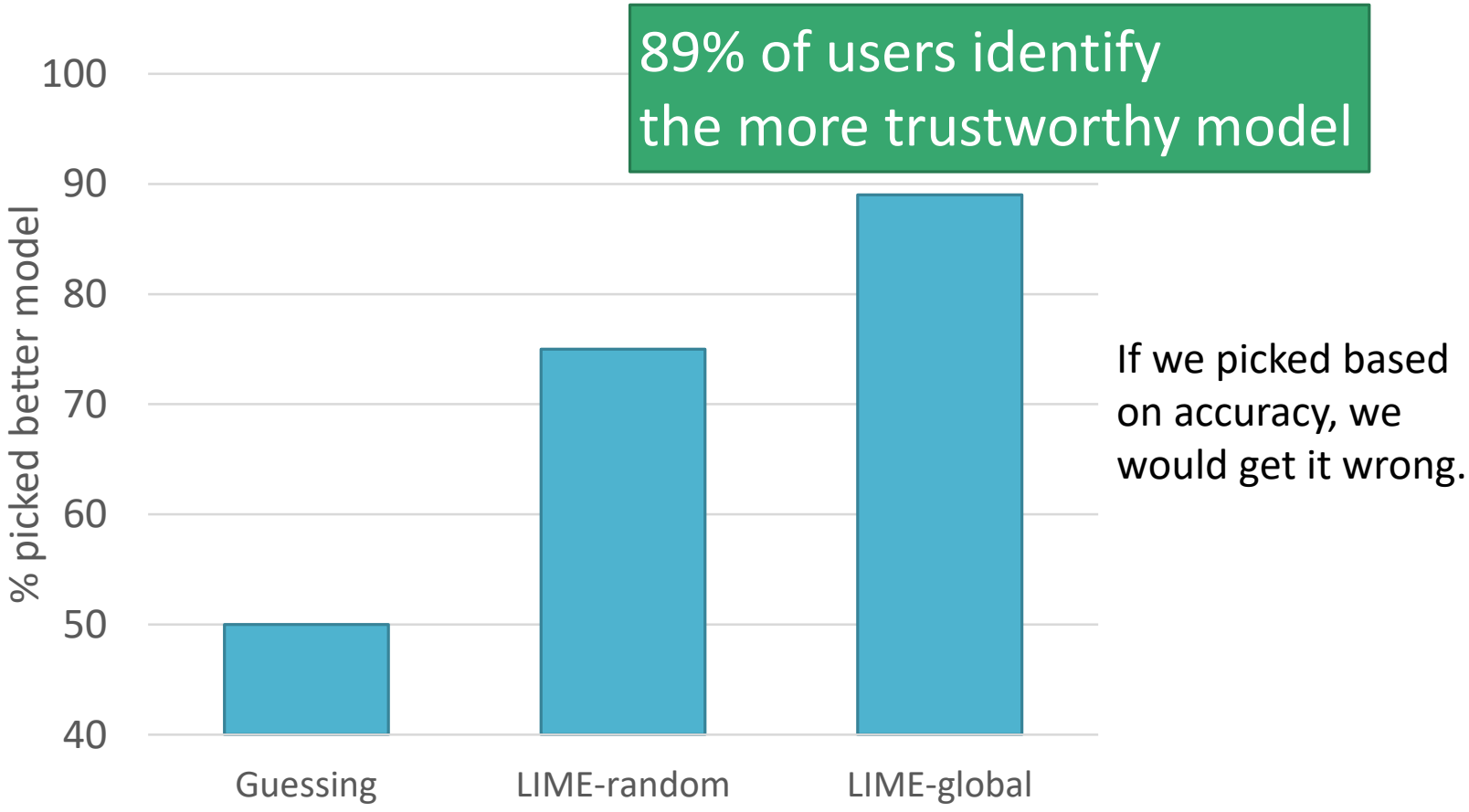
In article <4949@eastman.UUCP> dps@nasa.kodak.com writes:
>|> Yet I am still not a believer. Is god not concerned with my
>|> disposition? Why is it beneath him to provide me with the
>|> evidence I would require to believe? The evidence that my
>|> personality, given to me by this god, would find compelling?
>The fact is God could cause you to believe anything He wants you to.
>But think about it for a minute. Would you rather have someone love
>you because you made them love you, or because they wanted to
>love you.

It seems to be picking up on more reasonable things.. good!

UI for Comparing Classifiers



Comparing Models



Quantitative Evaluation

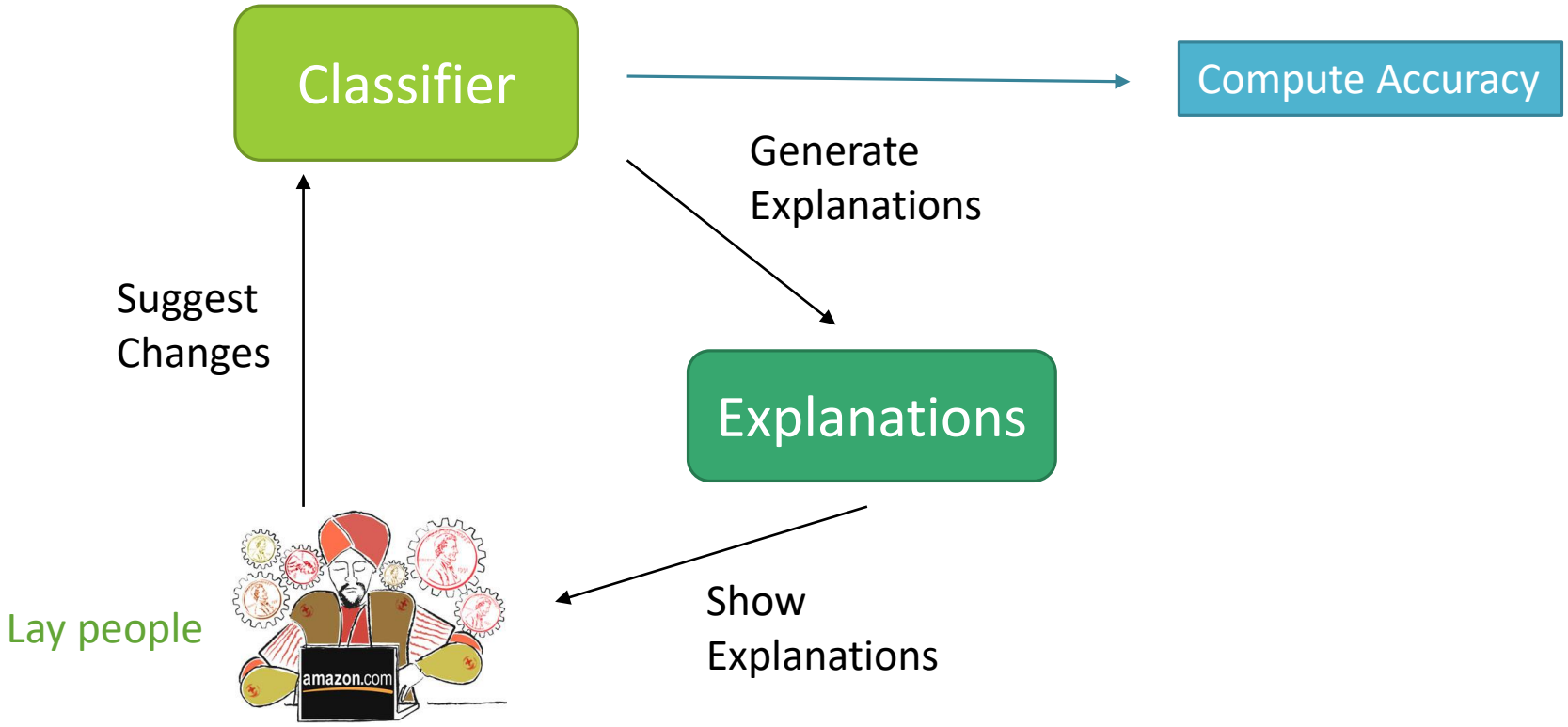
Understand what ML is doing

Compare different ML algorithms

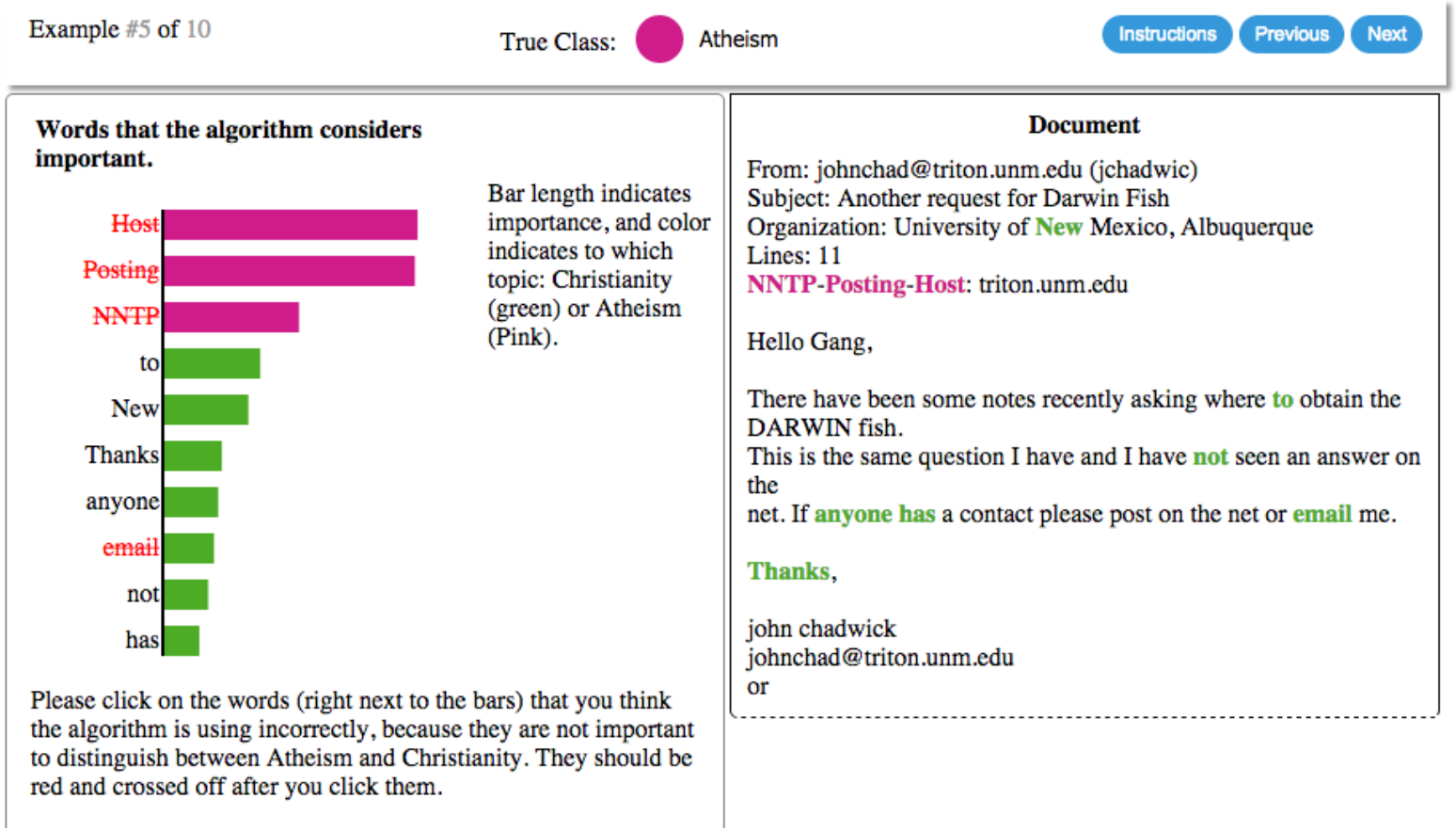
Improve the existing model

Predict how ML will behave

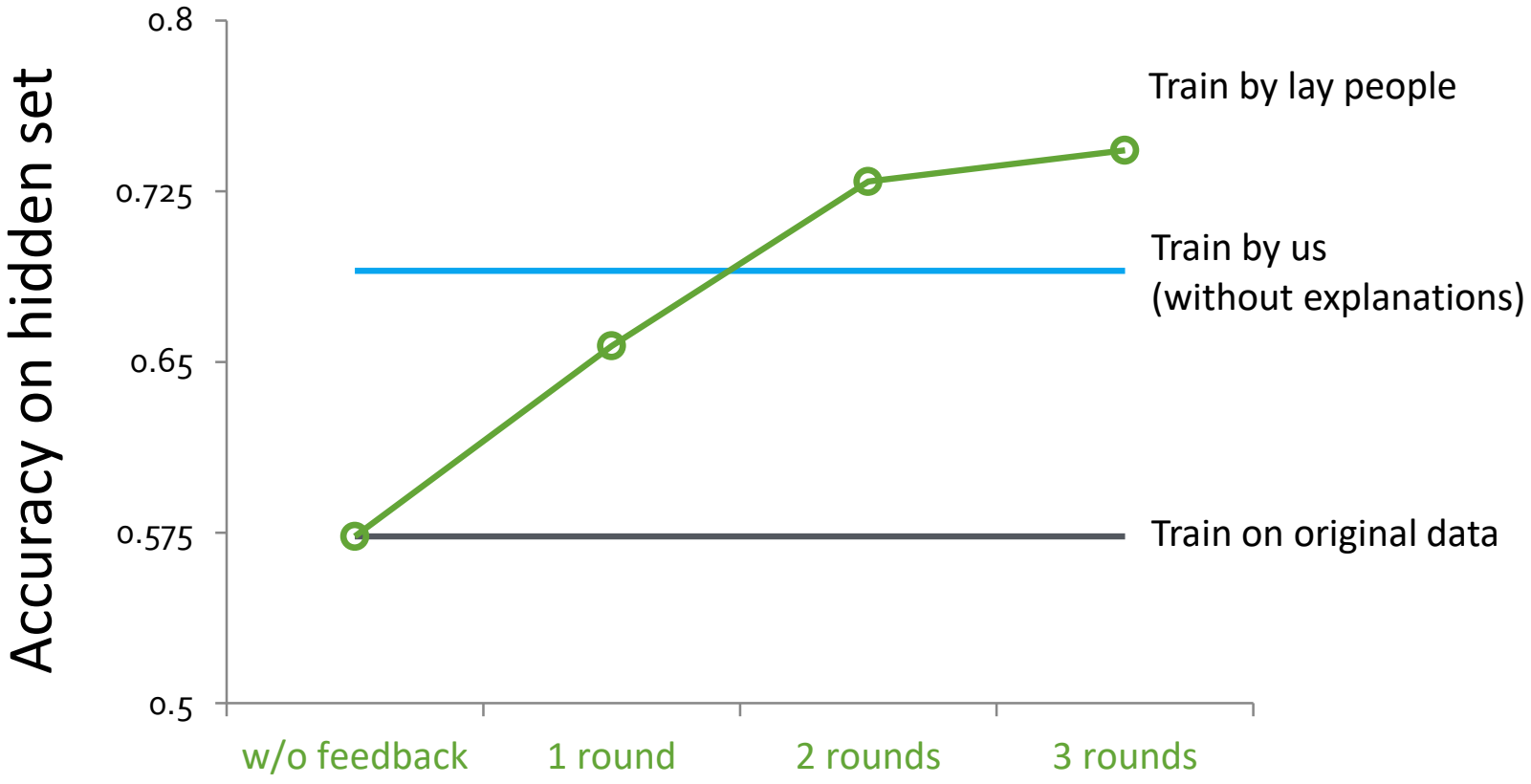
Improving Classifiers



UI for fixing bad classifiers



Fixing bad classifiers



Quantitative Evaluation

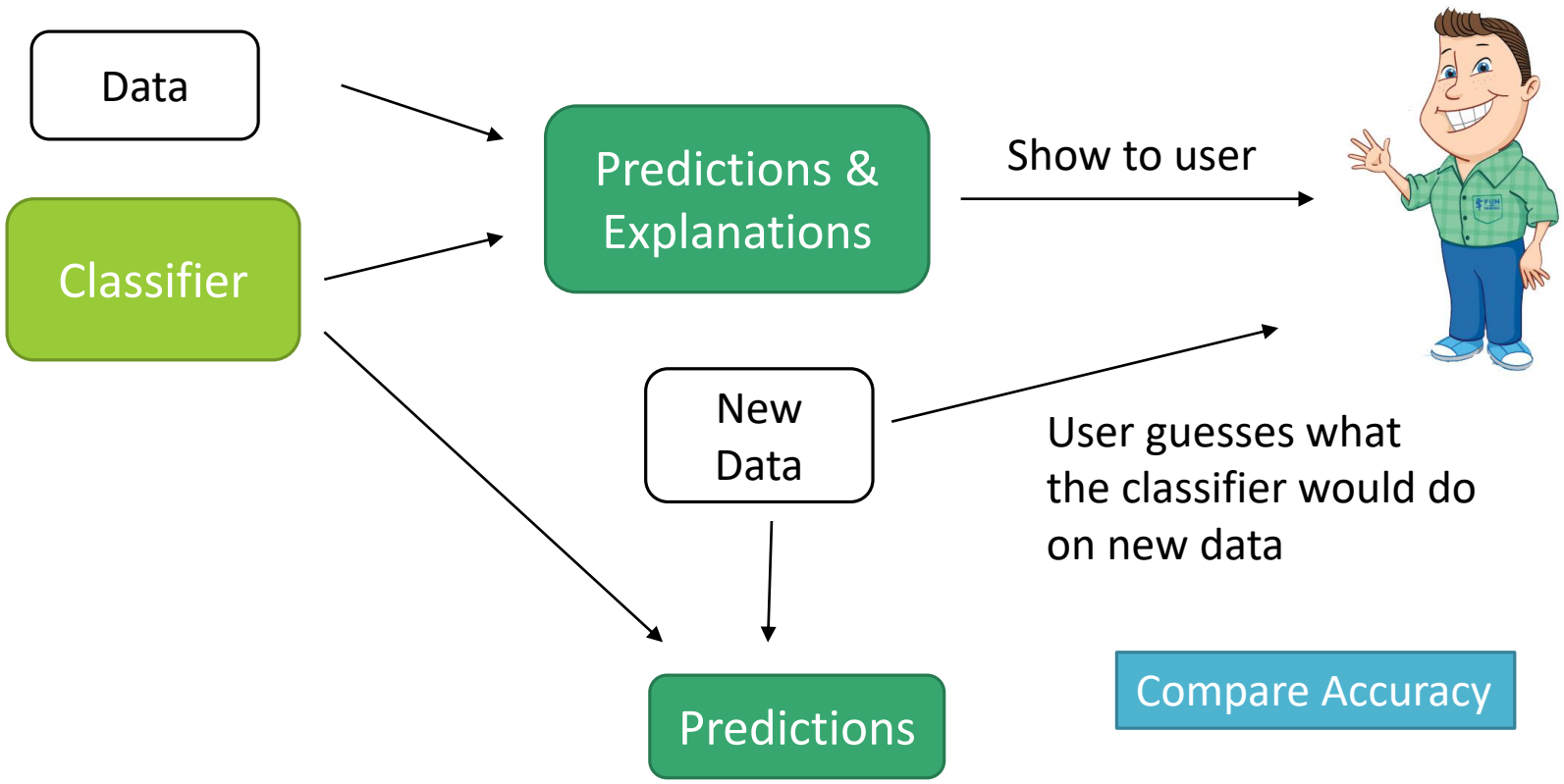
Understand what ML is doing

Compare different ML algorithms

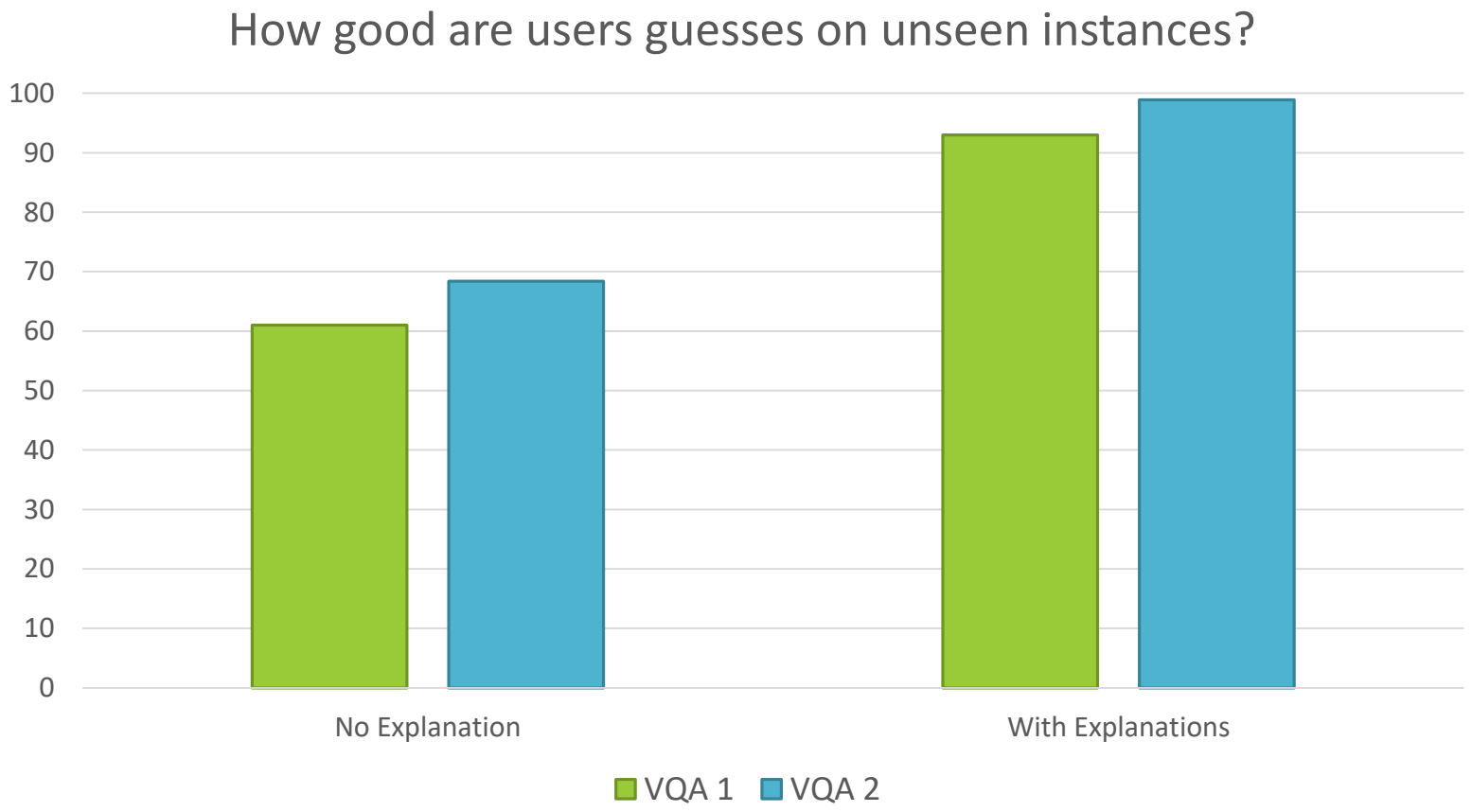
Improve the existing model

Predict how ML will behave

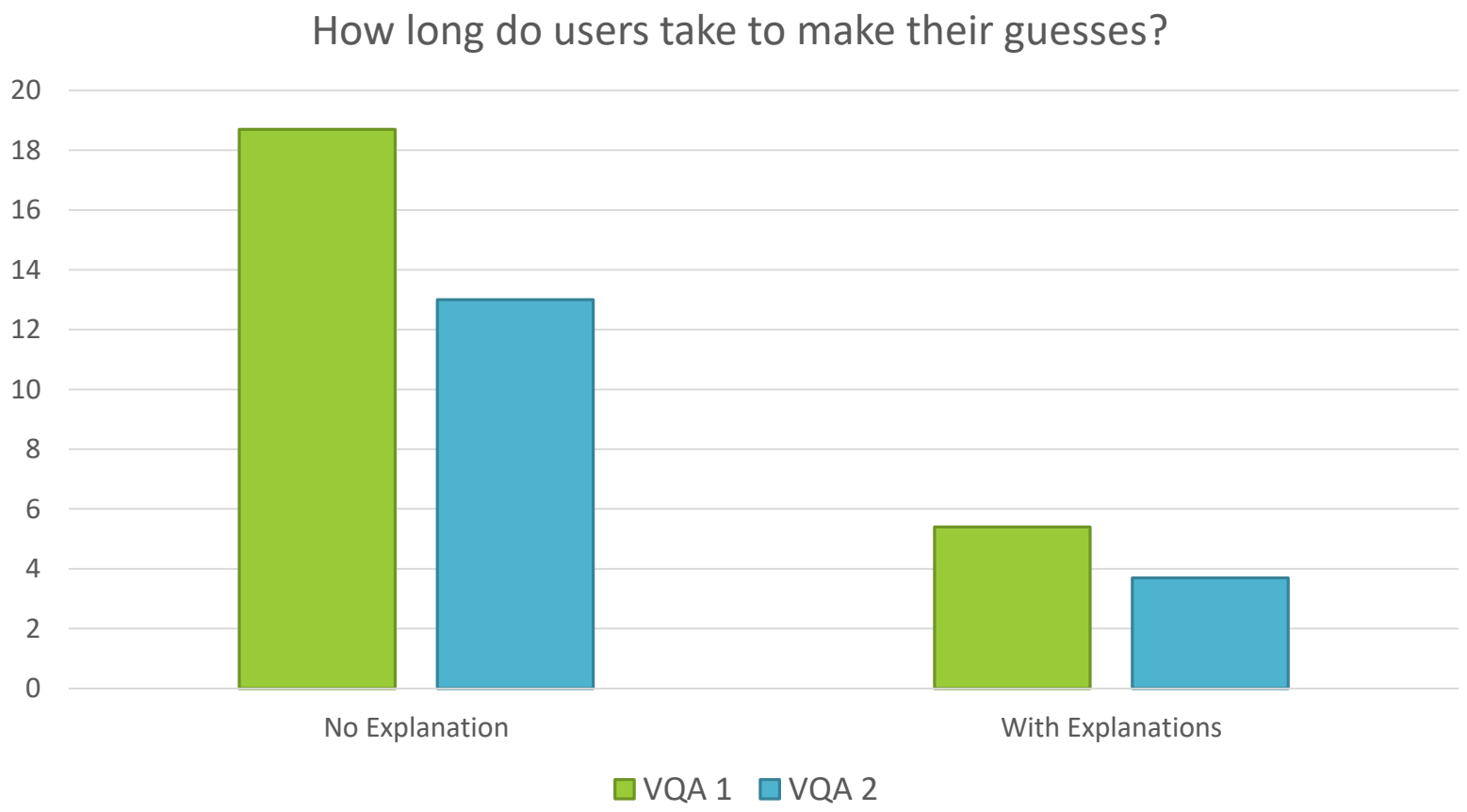
Predicting Behavior



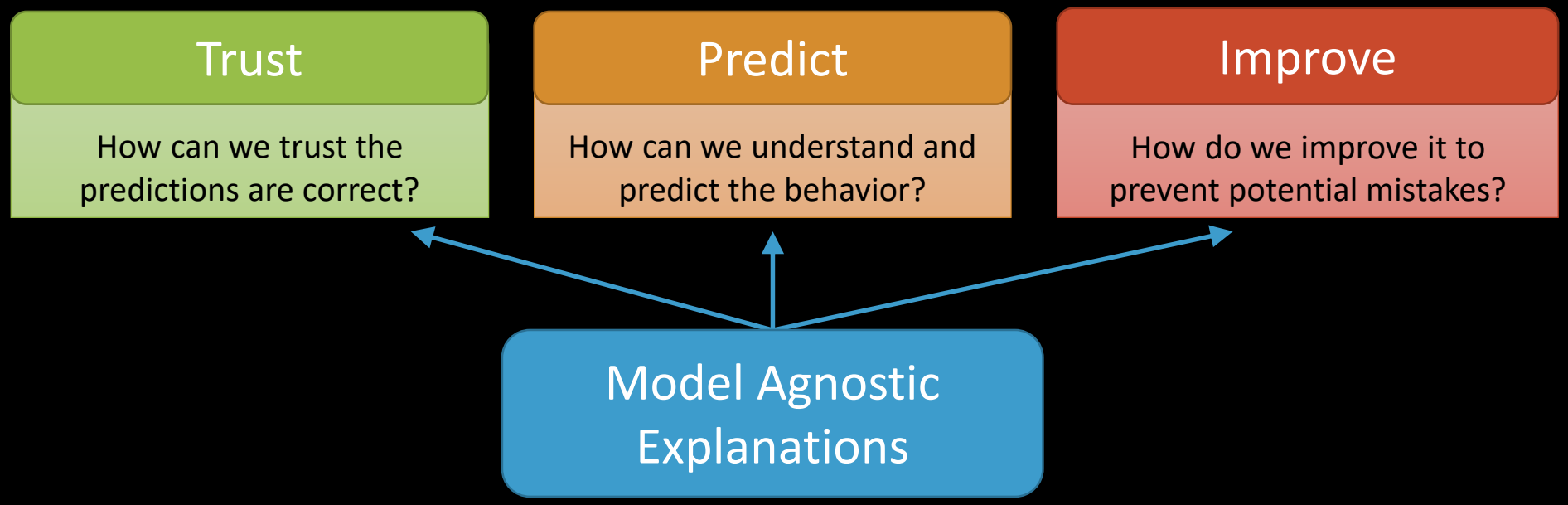
User Studies: *Precision*



User Studies: *Time*



Explanations are important!



Model Agnostic Explanations

"Why should I trust you?"
Explaining the predictions of any classifier
Ribeiro, **Singh**, Guestrin, KDD 2016

github.com/marcotcr/lime

Thanks!

sameer@uci.edu
sameersingh.org