

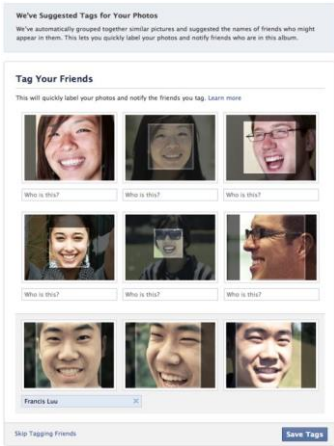
Explaining Black-Box Machine Learning Predictions

Sameer Singh

University of California, Irvine

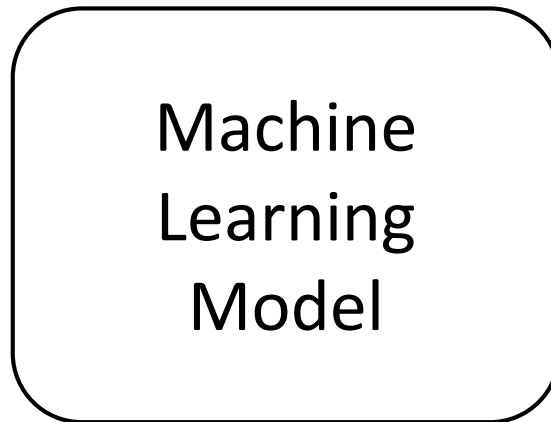
work with Marco T. Ribeiro and Carlos Guestrin

Machine Learning is Everywhere...



Classification: Wolf or a Husky?

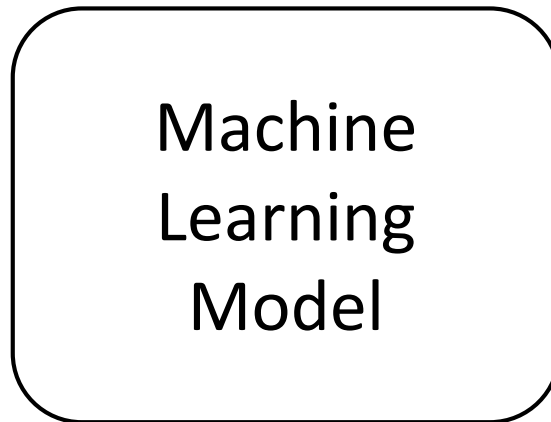
Adopt or not?



Wolf!

Classification: Wolf or a Husky?

Adopt or not?



Husky!

Classification: Wolf or a Husky?



Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

Only 1 mistake!



Predicted: **wolf**
True: **husky**



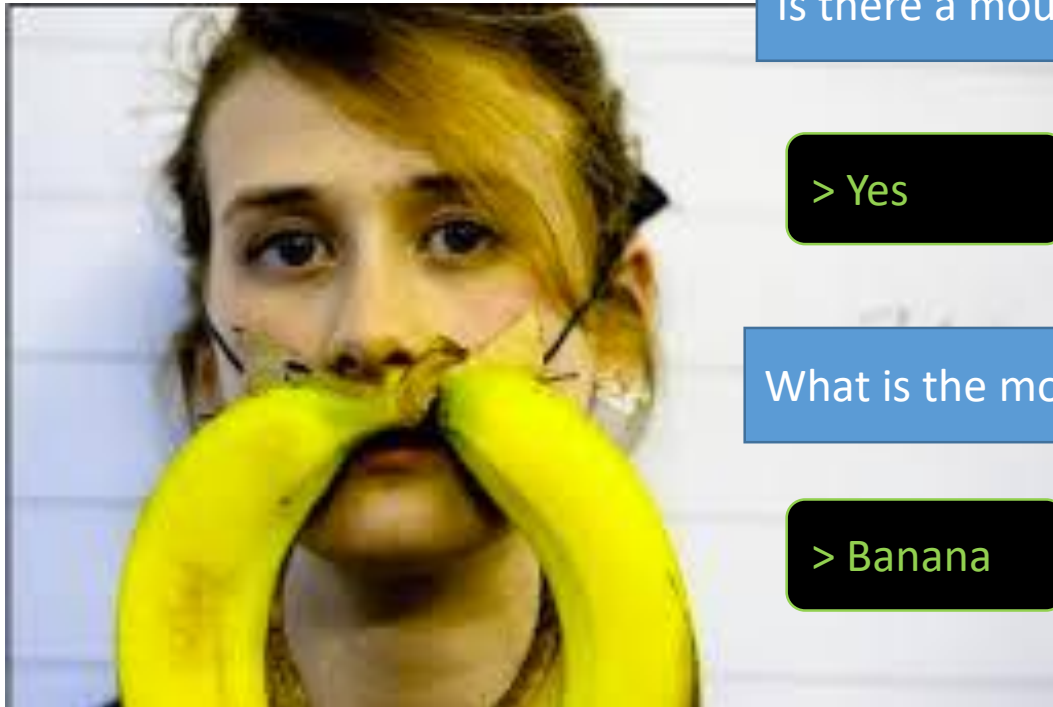
Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Visual Question Answering

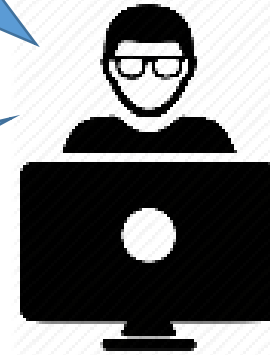


Is there a moustache in the picture?

> Yes

What is the moustache made of?

> Banana



Essentially black-boxes!

Trust

How can we trust the predictions are correct?

Predict

How can we understand and predict the behavior?

Improve

How do we improve it to prevent potential mistakes?

Classification: Wolf or a Husky?

Only 1 mistake!



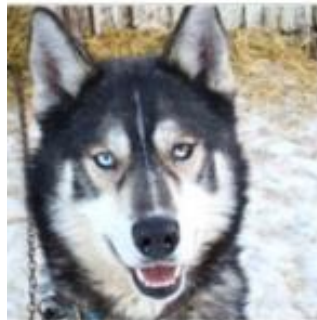
Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **husky**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

We've built a snow detector...

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

The Man Who Accidentally Adopted a Wolf Pup

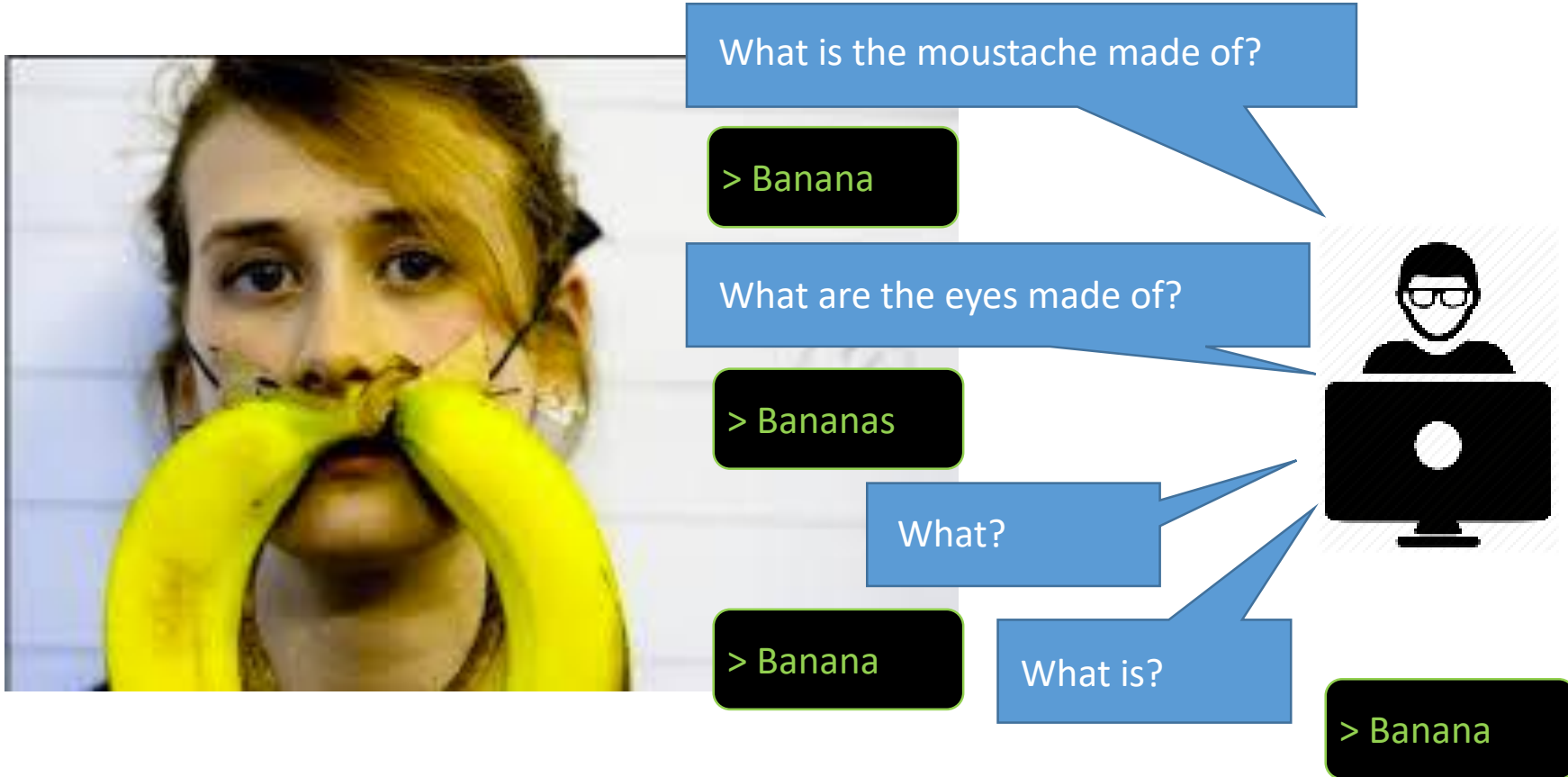
It did not go well.

By *A.J. McCarthy*

  
10k 547 6



Visual Question Answering



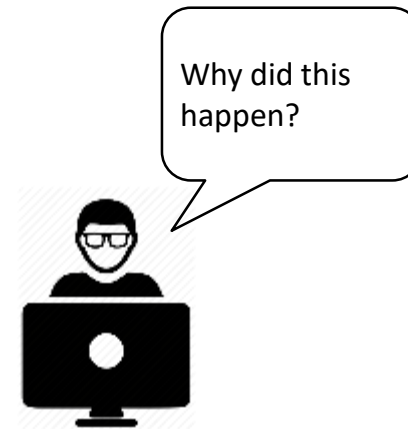
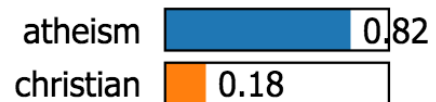
Text Classification

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

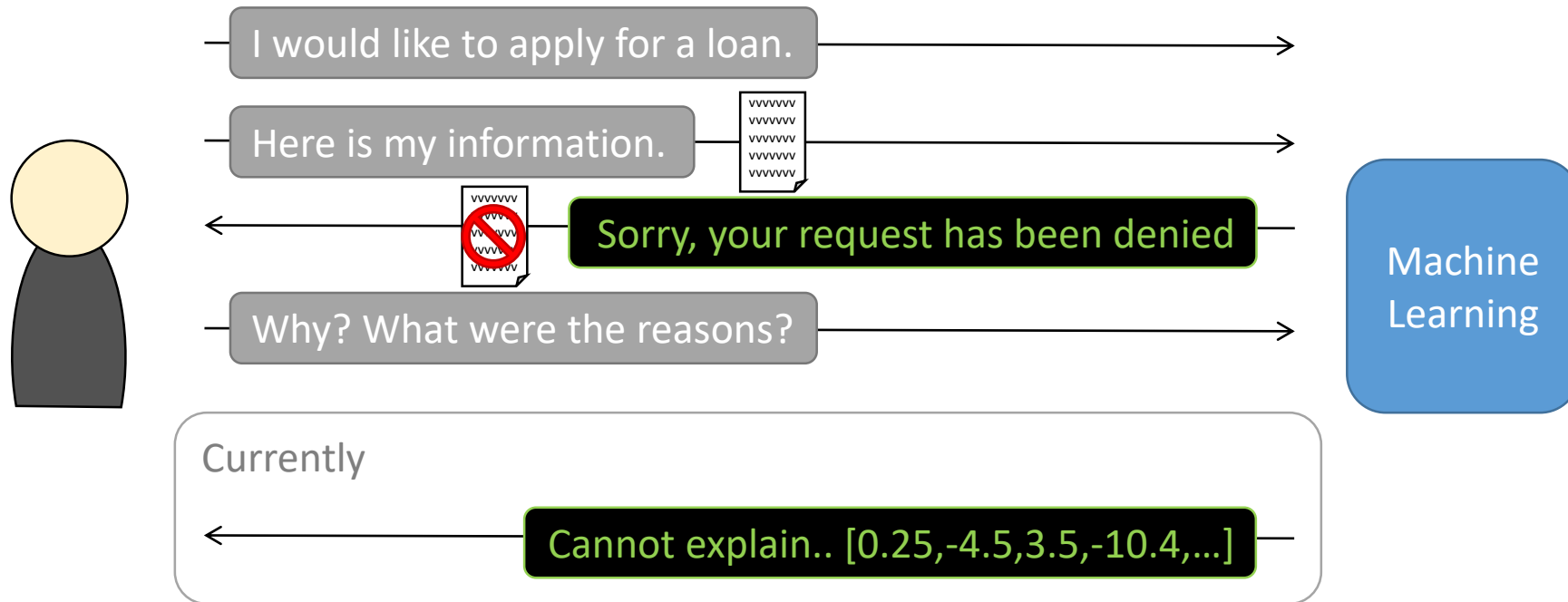
I think Christianity is the one true religion.
If you'd like to know more, send me a note



Prediction probabilities

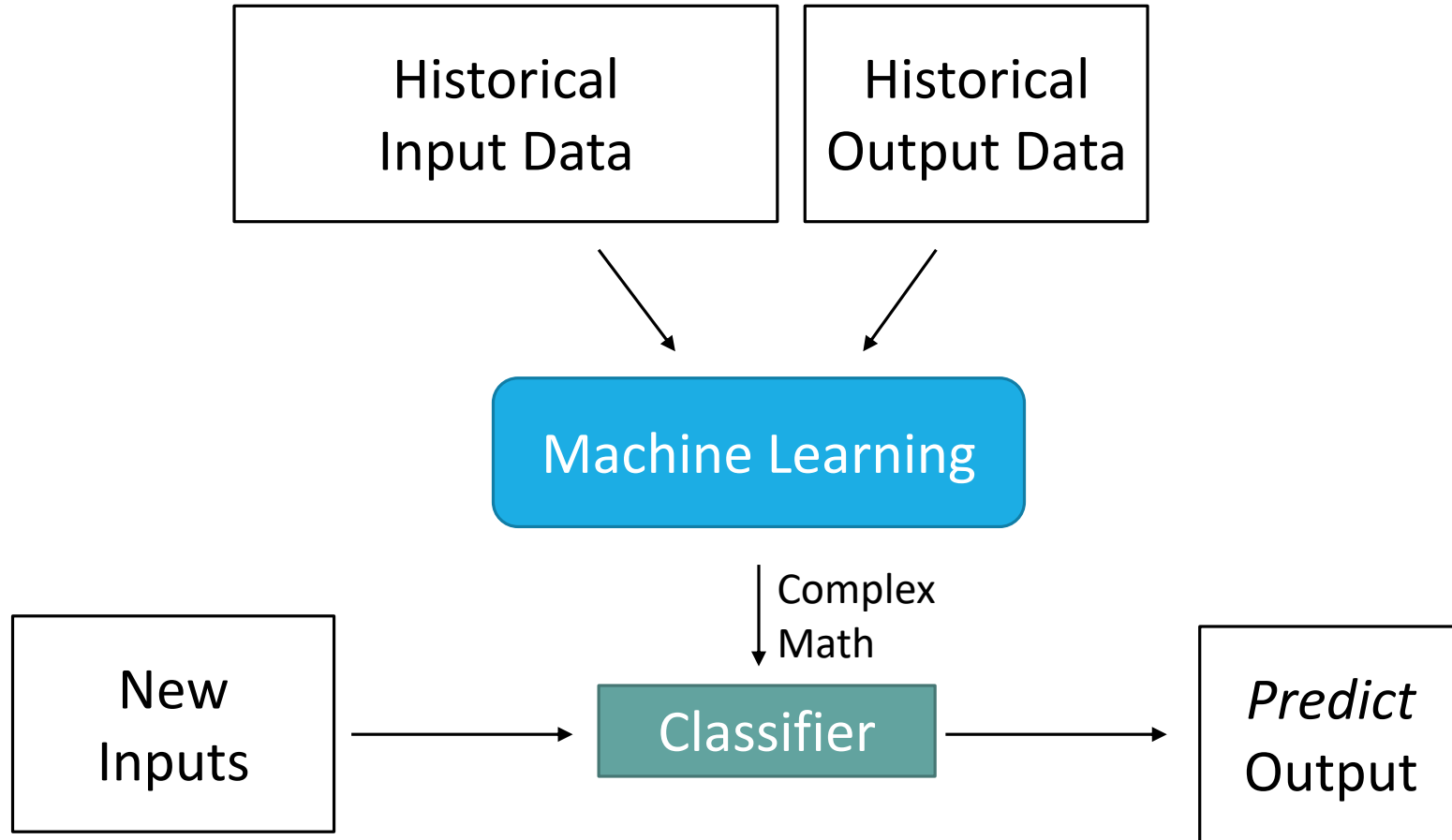


Loan Applications (in a Blackbox-ML World)

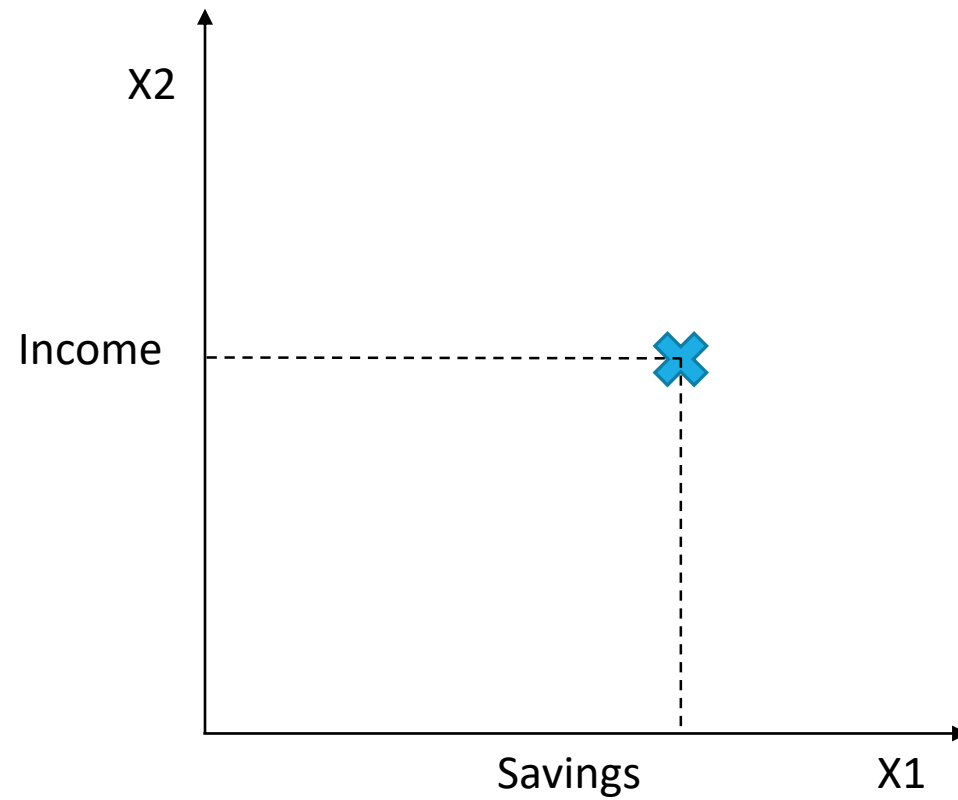


How did we get here?

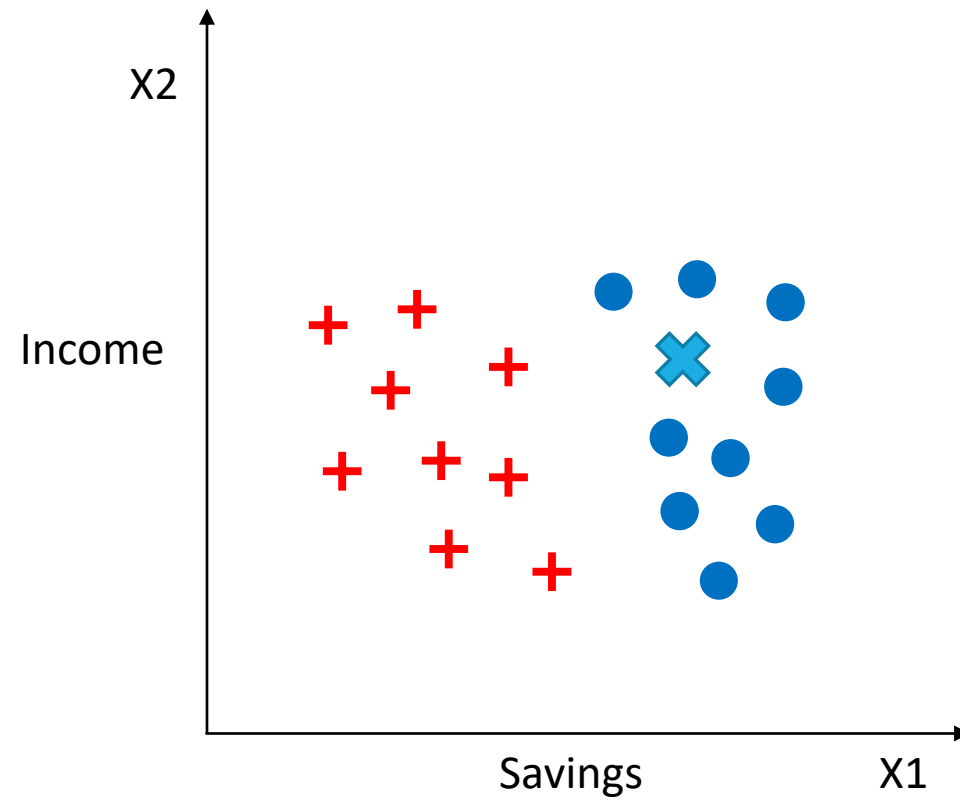
What is Machine Learning?



Should I give out a loan?



Get Historical Data

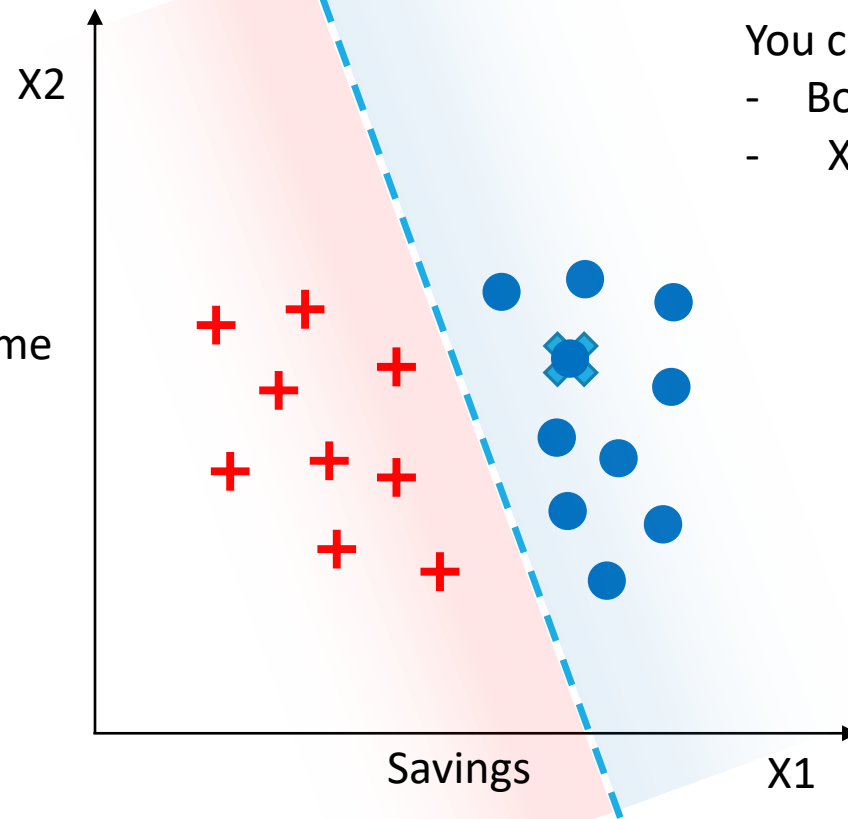


Linear Classifiers

if: $10X_1 + X_2 - 5 > 0$

●
otherwise
+

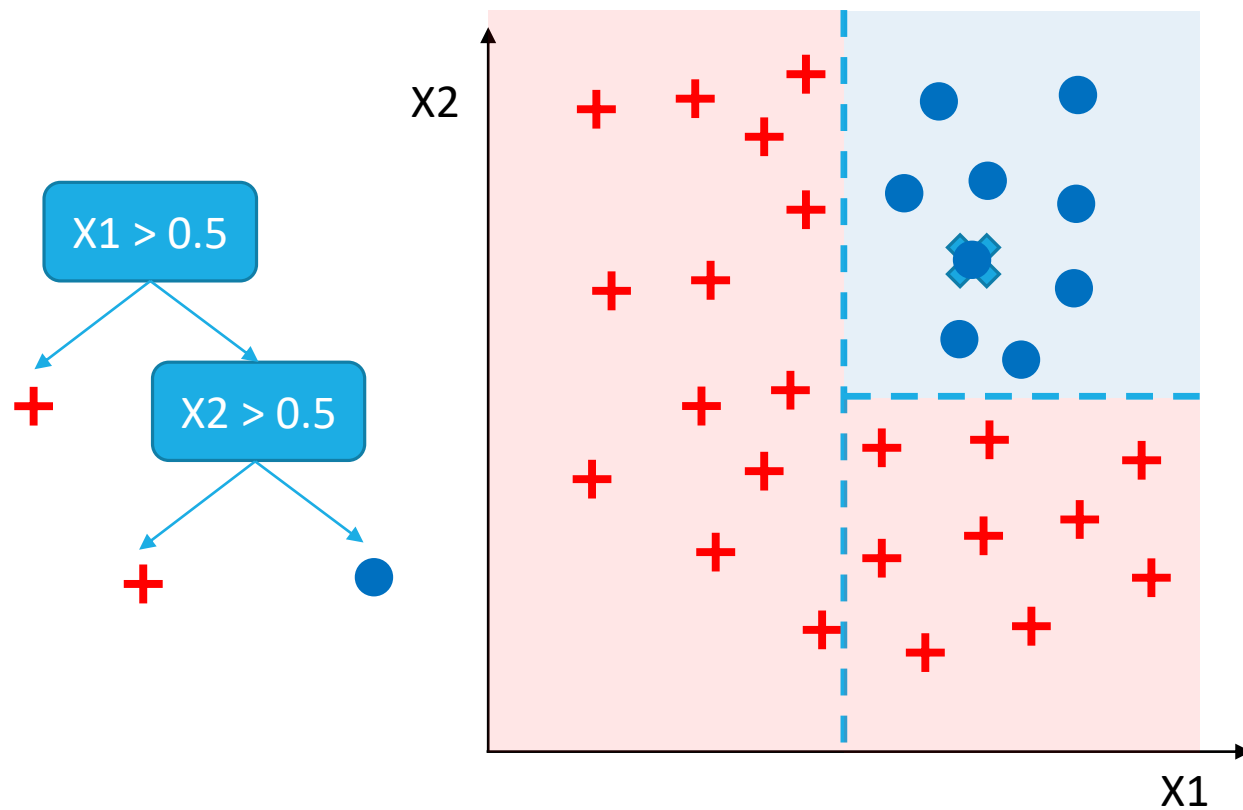
Income



You can interpret it...

- Both have a positive effect
- $X_1 > X_2$

Decision trees



You can interpret it...

- X_2 is irrelevant if $X_1 < 0.5$
- Otherwise X_2 is enough

Looking at the structure

Trust

How can we trust the predictions are correct?



Test whether the structure agrees with our intuitions.

Predict

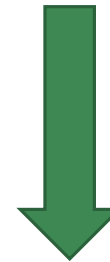
How can we understand and predict the behavior?



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.

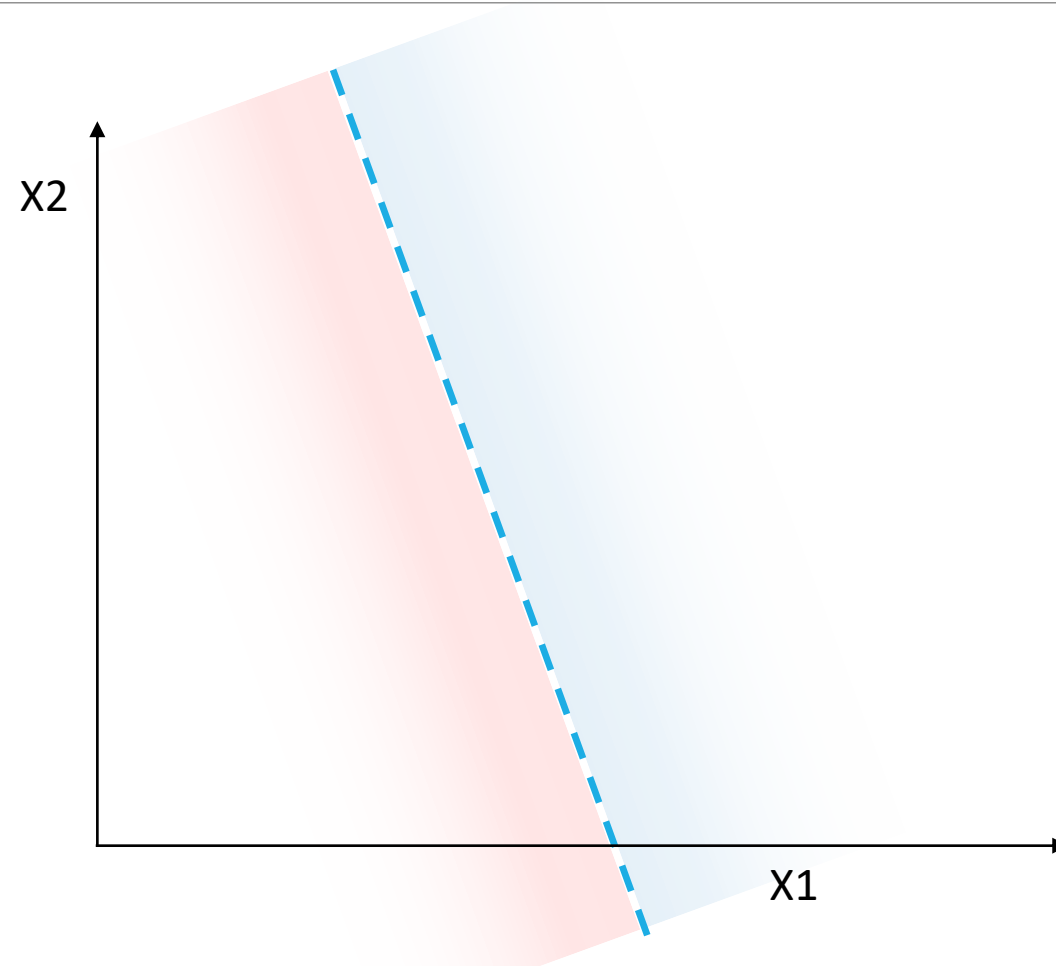
Arrival of Big Data

Big Data: Applications of ML

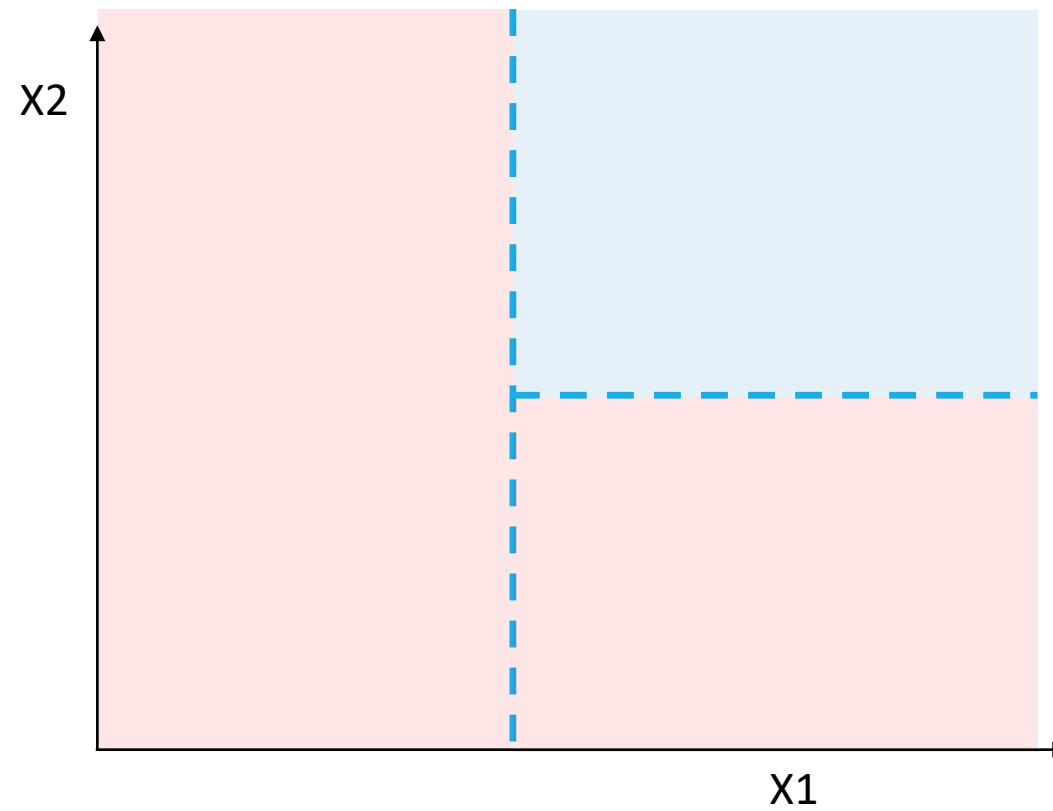
The collage features several key elements:

- Top Left:** A yellow speech bubble containing the text "Xconomy Forum Big Data Meets Big Biology" and the date "March 31, 2016".
- Top Center:** A network diagram with blue and green nodes and connecting lines.
- Top Right:** A word cloud with prominent words like "word", "language", "translation", "text", "dialogue", "ontology", "document", "compositional", "content", "automatic", and "natural".
- Middle Left:** A grey box with the text "big data, big opportunities: retail industry" and a row of four colorful icons (blue, green, orange, blue).
- Middle Right:** A stylized illustration of a building facade with a dark roof, white columns, and a blue sky with a yellow sun and clouds. The text "Big Data in BANKING" is displayed on the building's front.
- Bottom Left:** A blue-tinted image of a circuit board with the text "Digital Humanities" overlaid.
- Bottom Center:** Two more colorful icons (orange and blue).
- Background:** The word "medicine" is partially visible in a large, light-colored font.

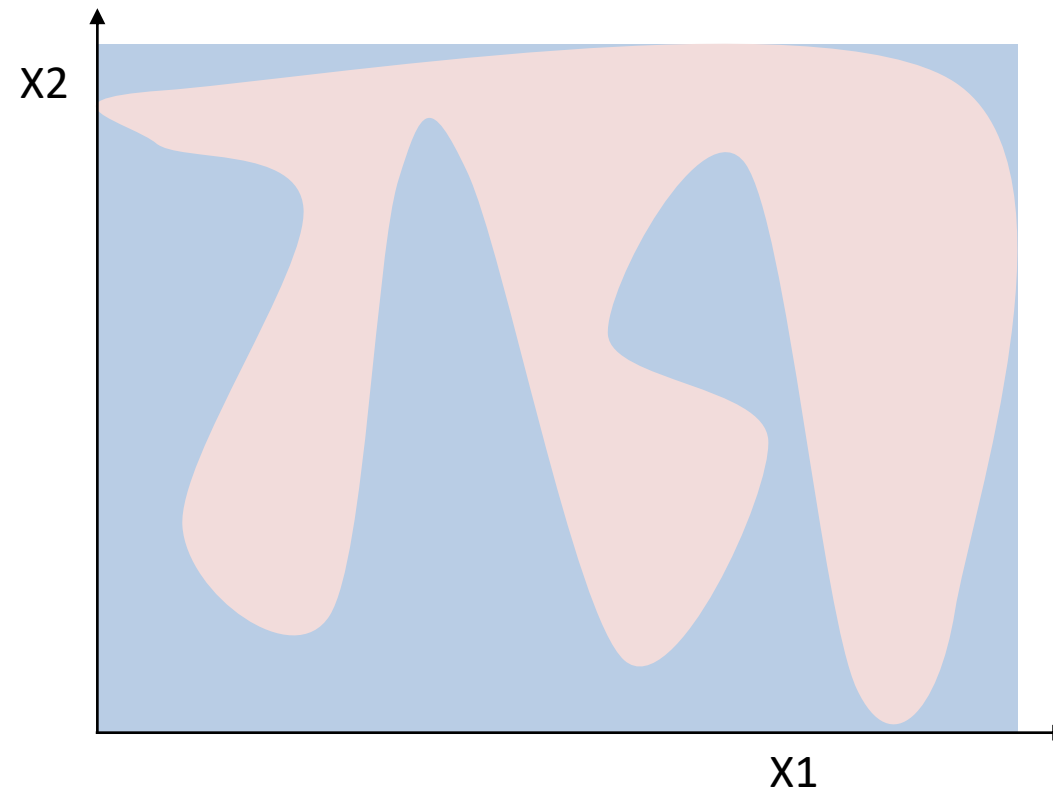
Big Data: More Complexity



Big Data: More Complexity



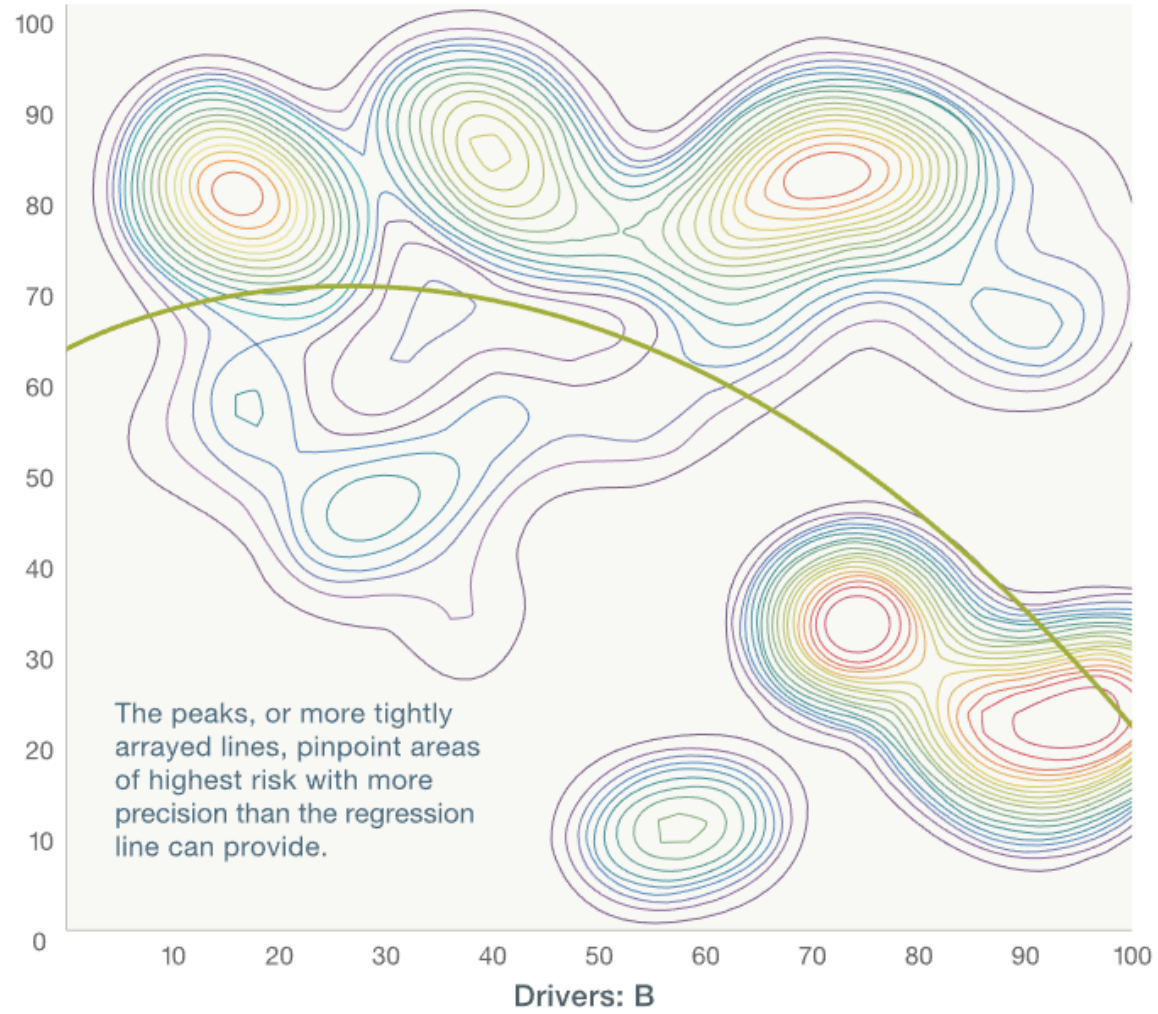
Big Data: More Complexity



— Classic regression analysis

○ Isobar graph facilitated by machine learning: warmer colors indicate higher degrees of risk

Drivers: A



McKinsey&Company

Big Data: More Dimensions

Savings

Income

Credit scores

Loan Amount

Past defaults

Recent defaults

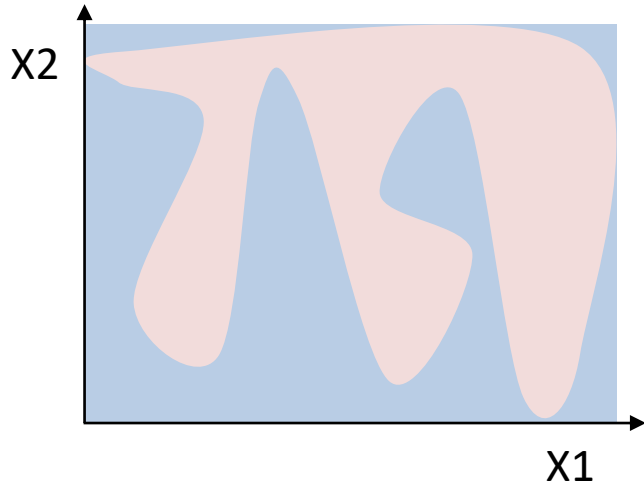
Profession

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
- ...

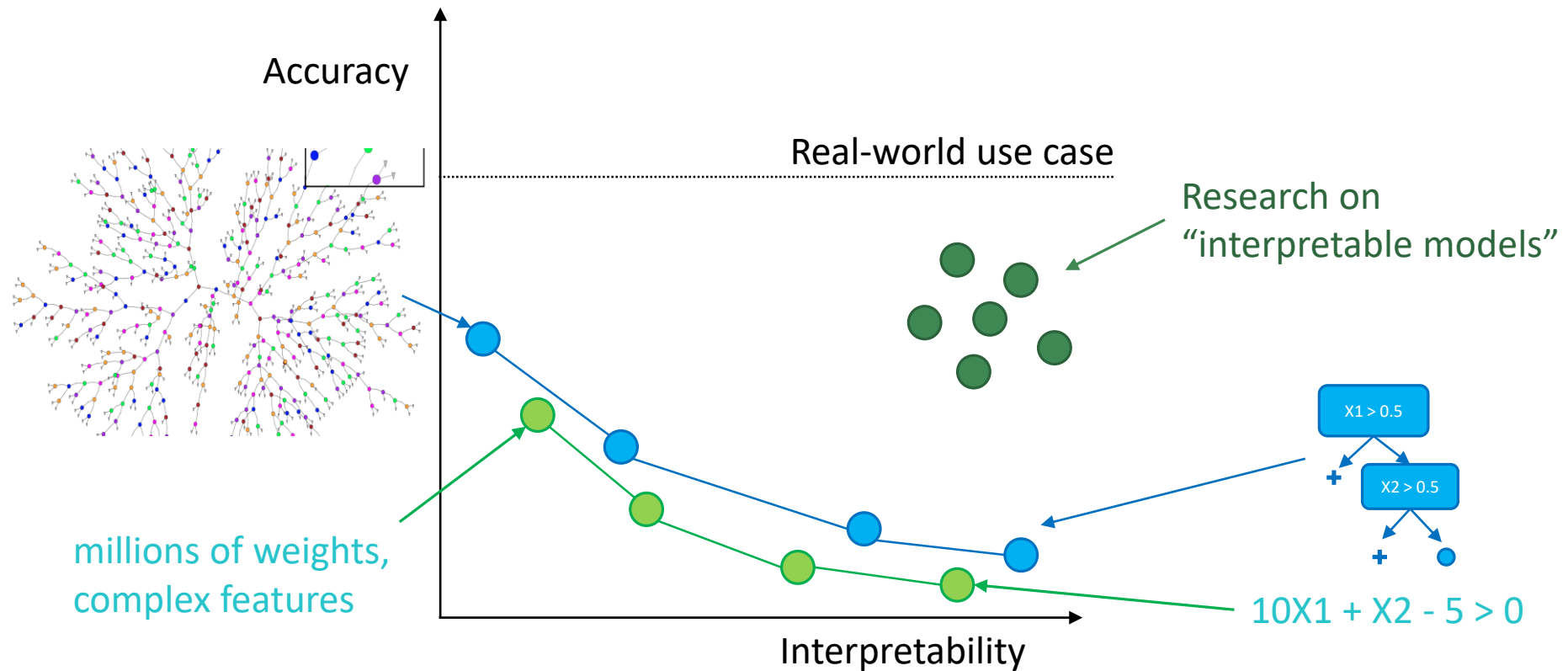
Complex Surfaces

+

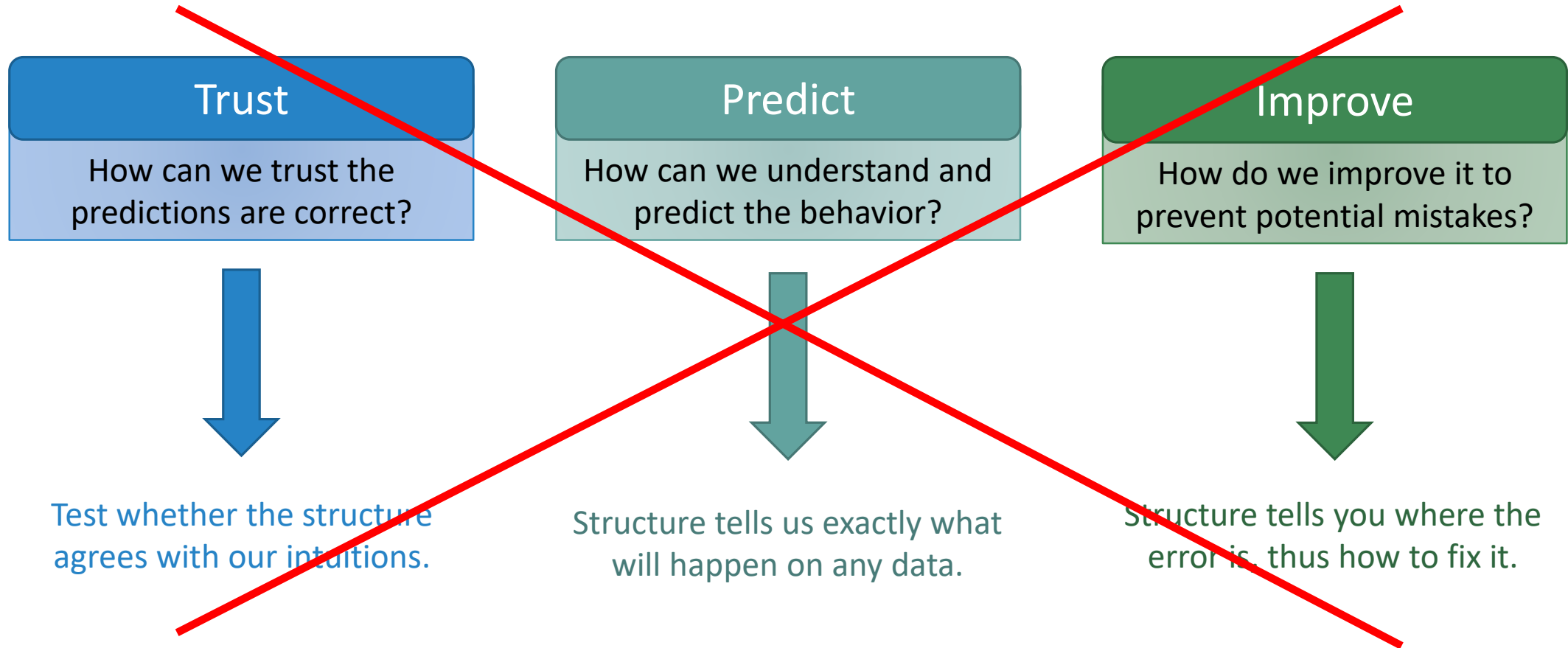
Lots of dimensions

Black-boxes!

Accuracy vs Interpretability



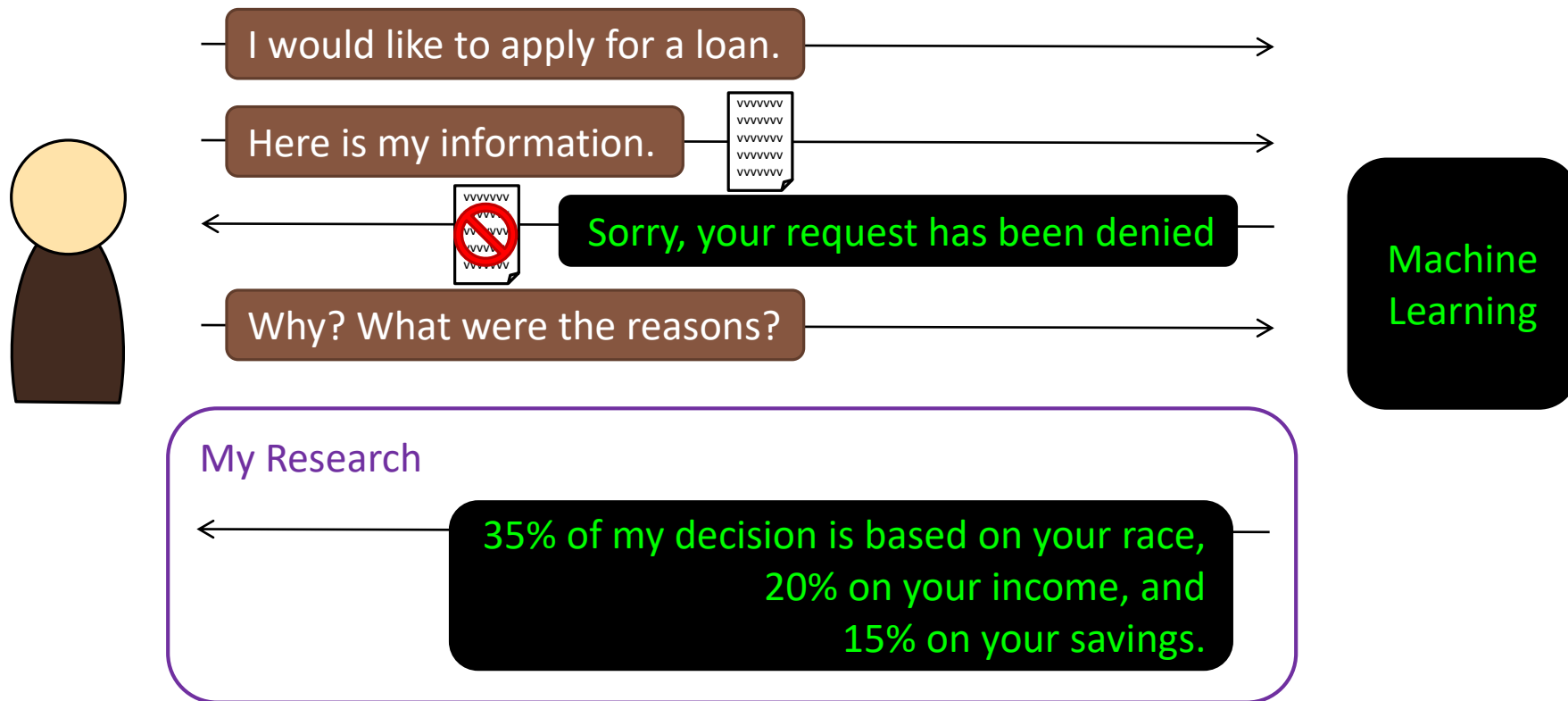
Looking at the structure



Explaining Predictions

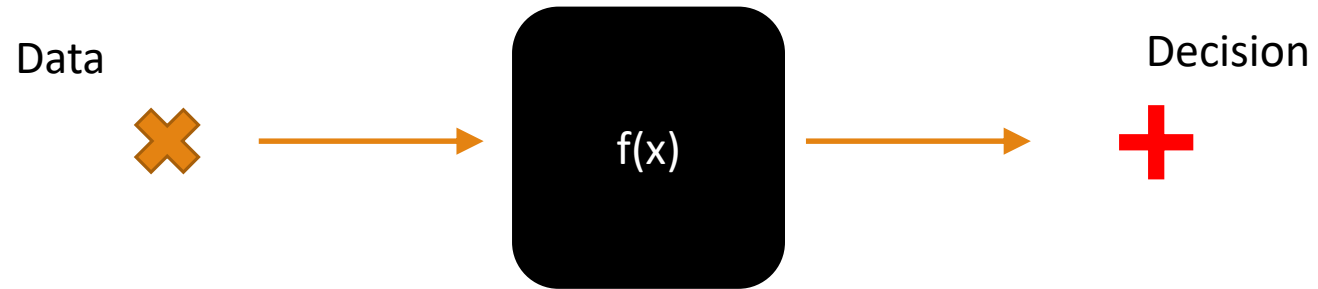
The LIME Algorithm

Applying for a Loan



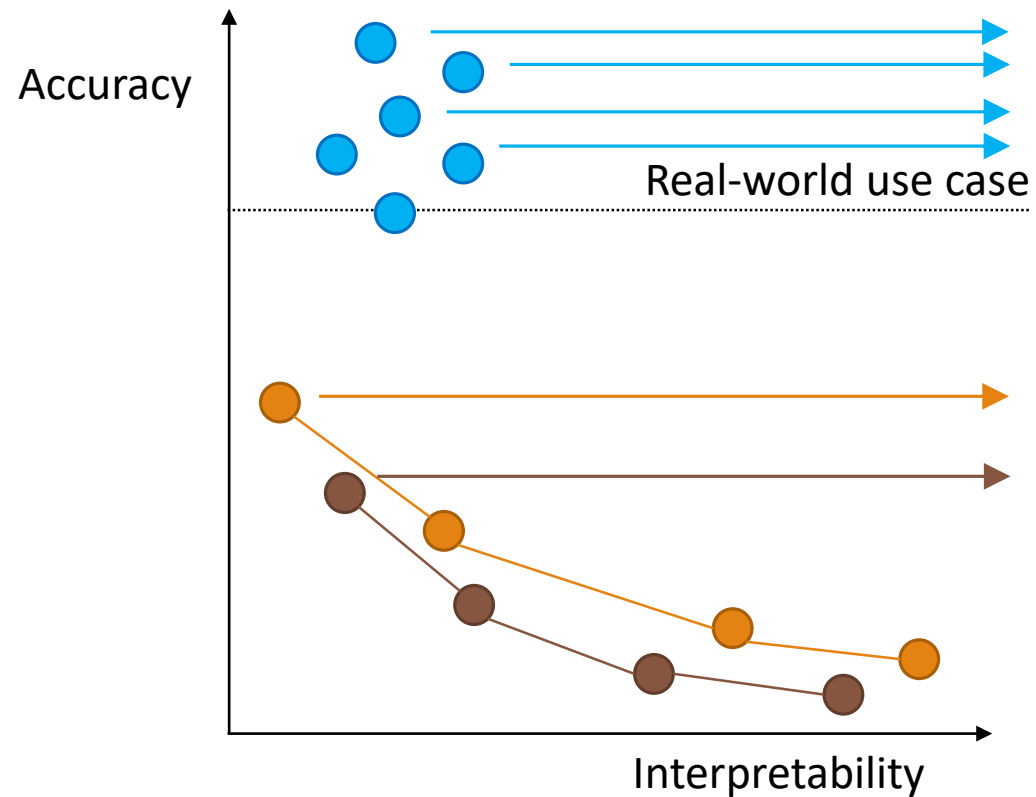
Being Model-Agnostic...

No assumptions about the internal structure...



Explain any existing, or *future*, model

LIME: Explain Any Classifier!



Make everything interpretable!

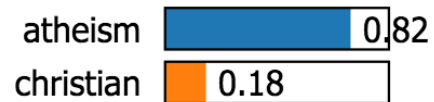
What is an “Explanation”?

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

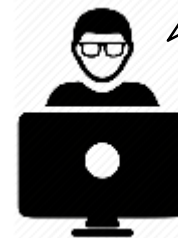
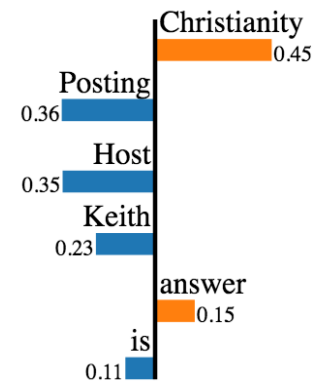


Prediction probabilities



atheism

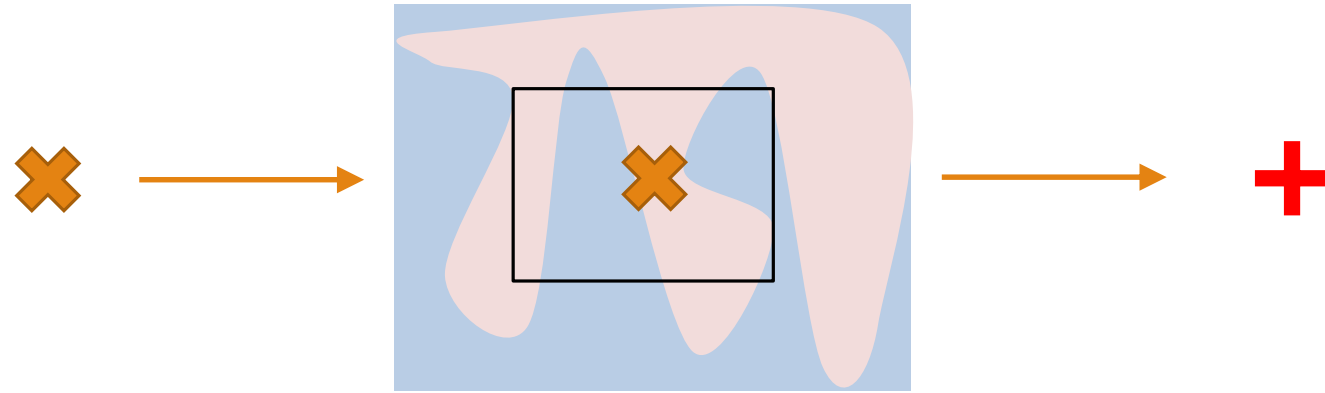
christian



Why did this happen?

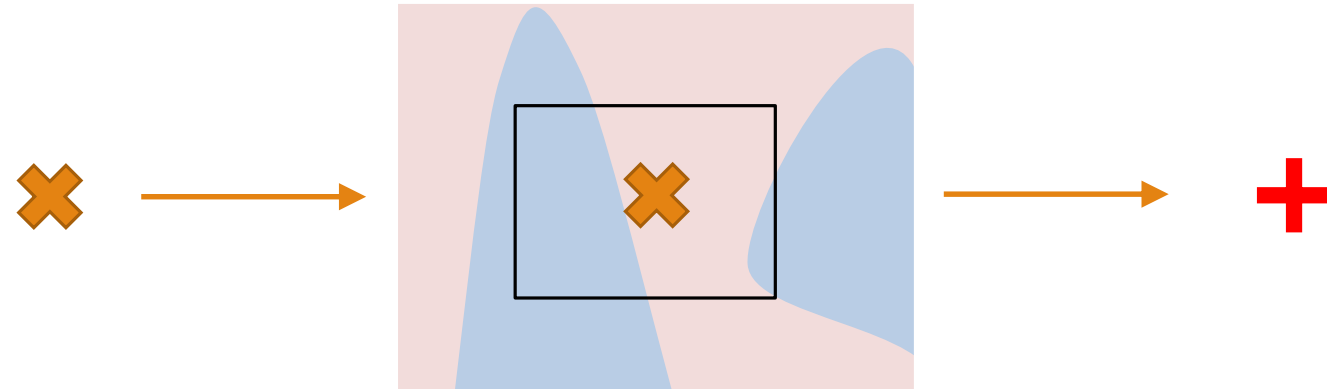
Being Model-Agnostic...

“Global” explanation is too complicated



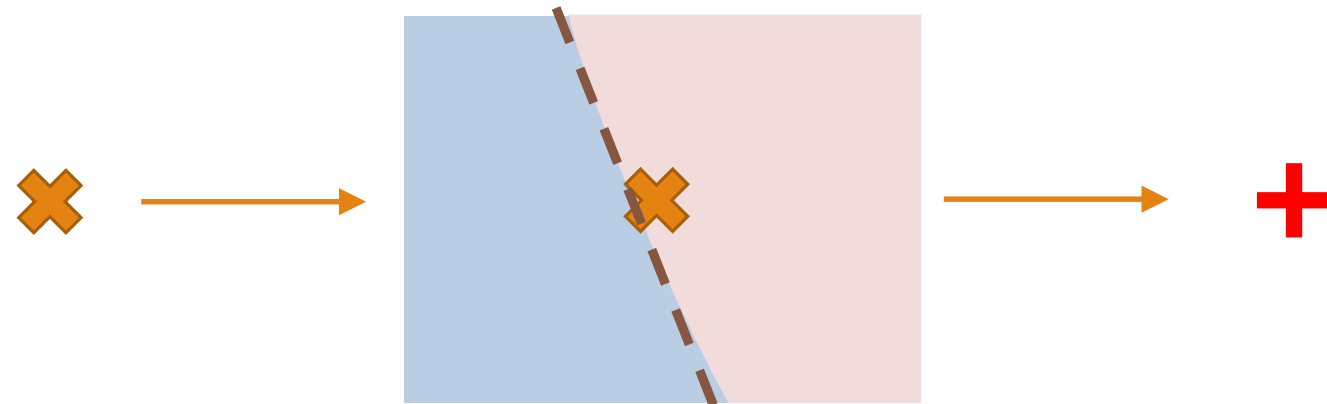
Being Model-Agnostic...

“Global” explanation is too complicated



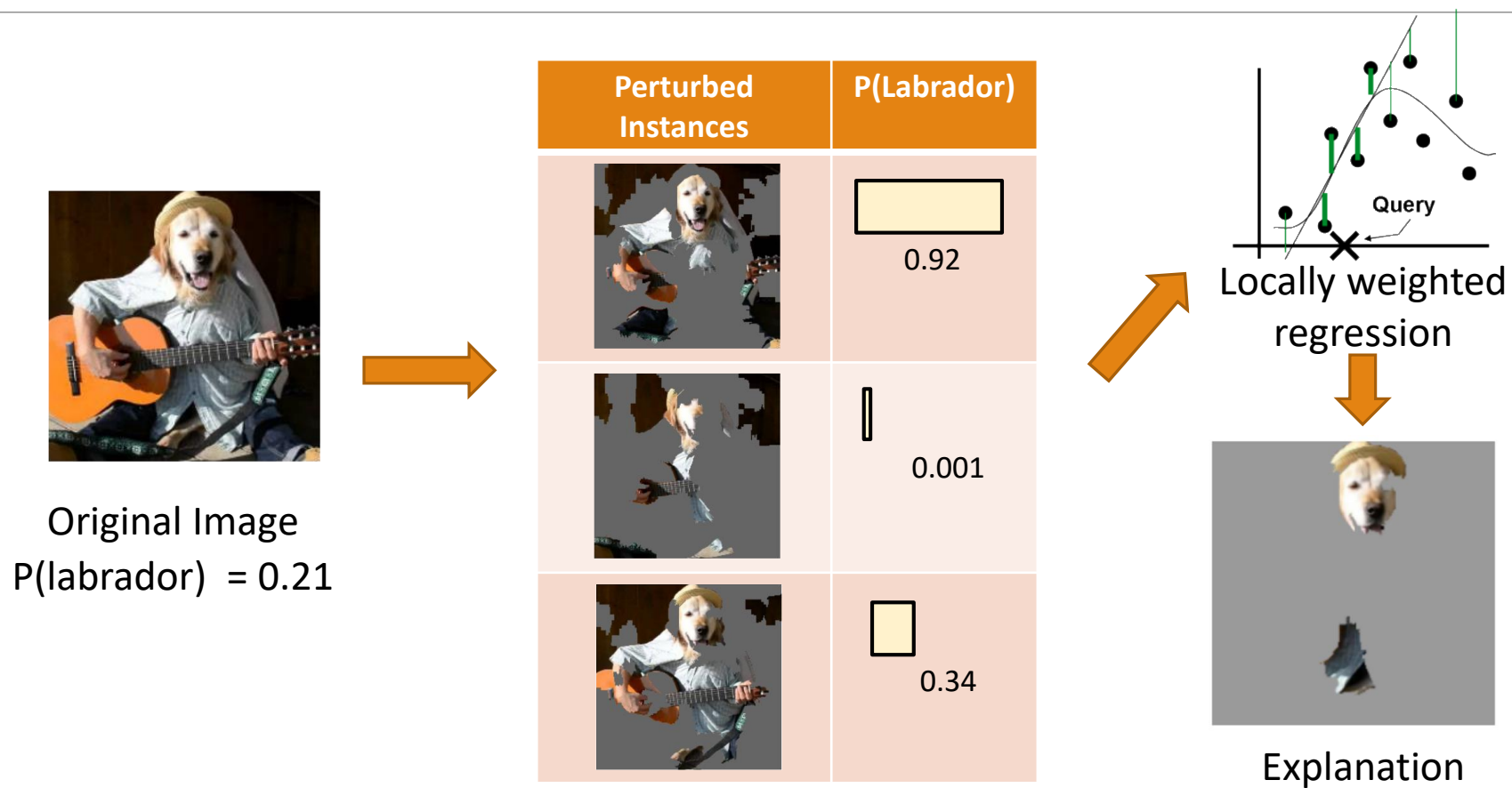
Being Model-Agnostic...

“Global” explanation is too complicated




Explanation is an interpretable model,
that is locally accurate

Example – Image Classification



Google's Object Detector



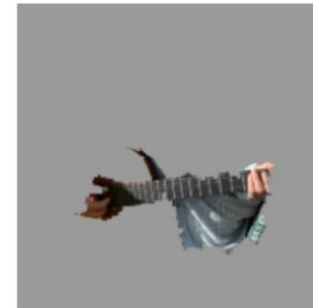
 $P(\text{dog}) = 0.21$



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



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



Classification: Wolf or a Husky?



Predicted: **wolf**
True: **wolf**

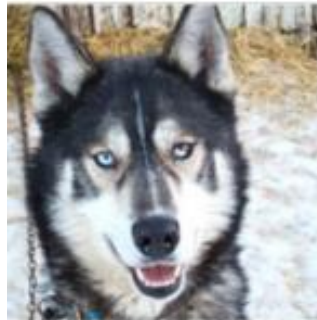


Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

Only 1 mistake!



Predicted: **wolf**
True: **husky**



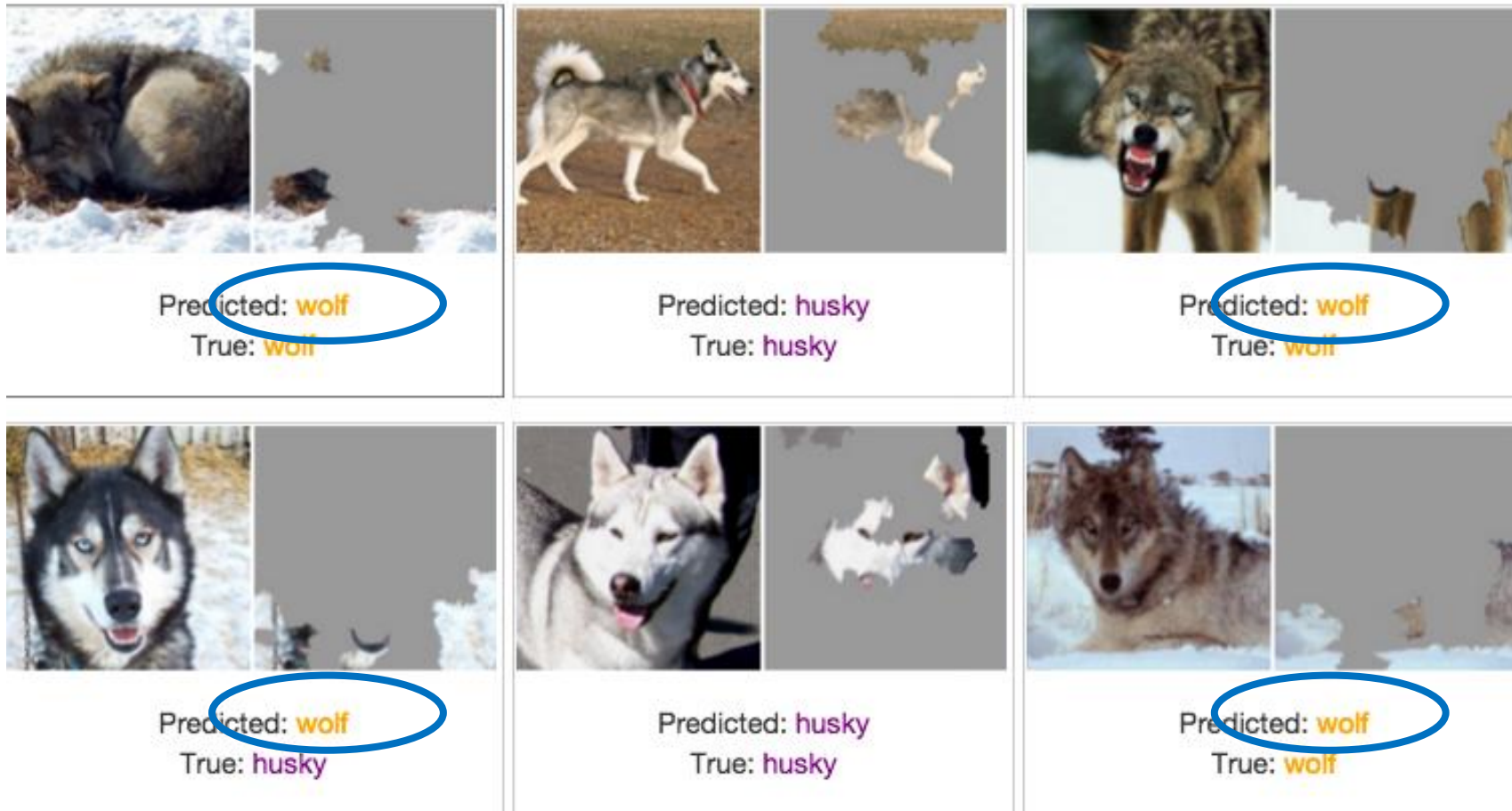
Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



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

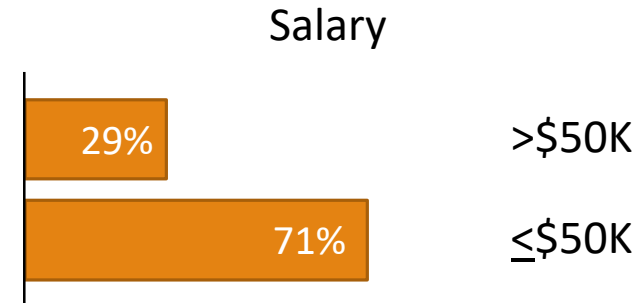
English	Portuguese
This is the question we must address	Esta é a questão que temos que enfrentar.
This is the problem we must address	Este é o problema que temos que enfrentar.

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

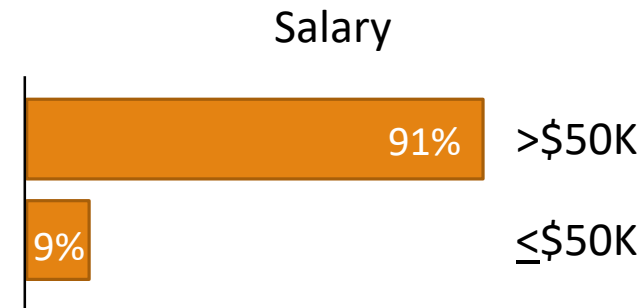
Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	\leq High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	≤ 40
Country	United States



**IF Education \leq High School
Then Predict Salary \leq 50K**

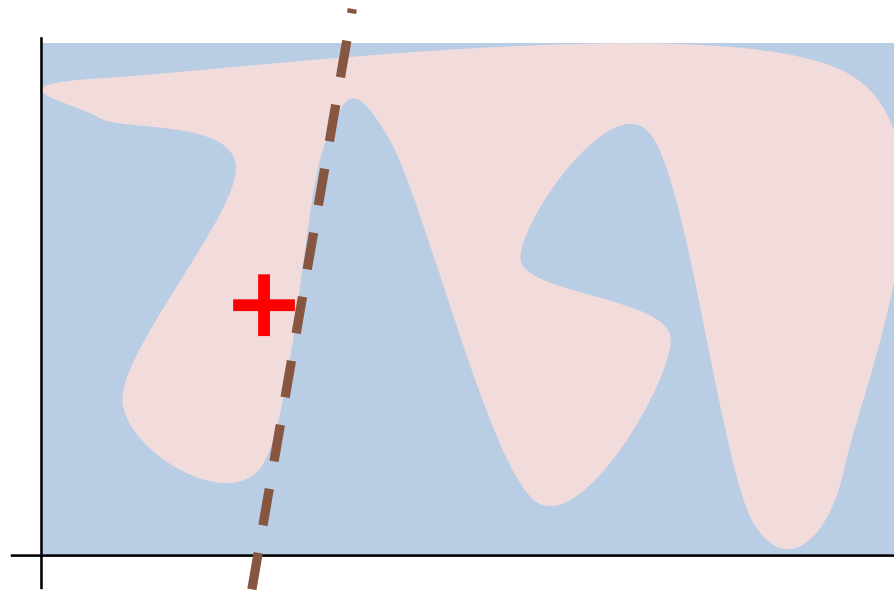
Salary Prediction

28 < Age ≤ 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**

“Global” Behavior



What about explaining the rest of the model?

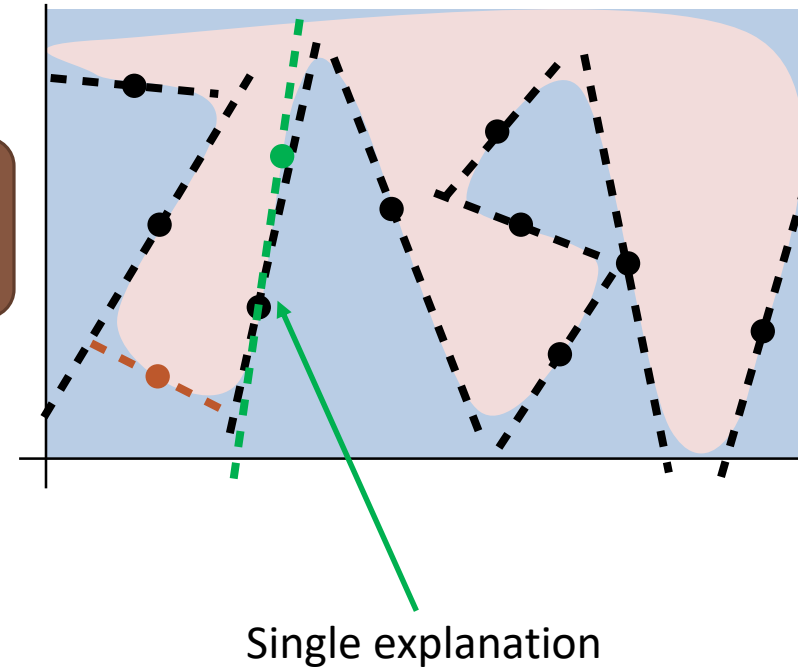
Explaining Global behavior

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

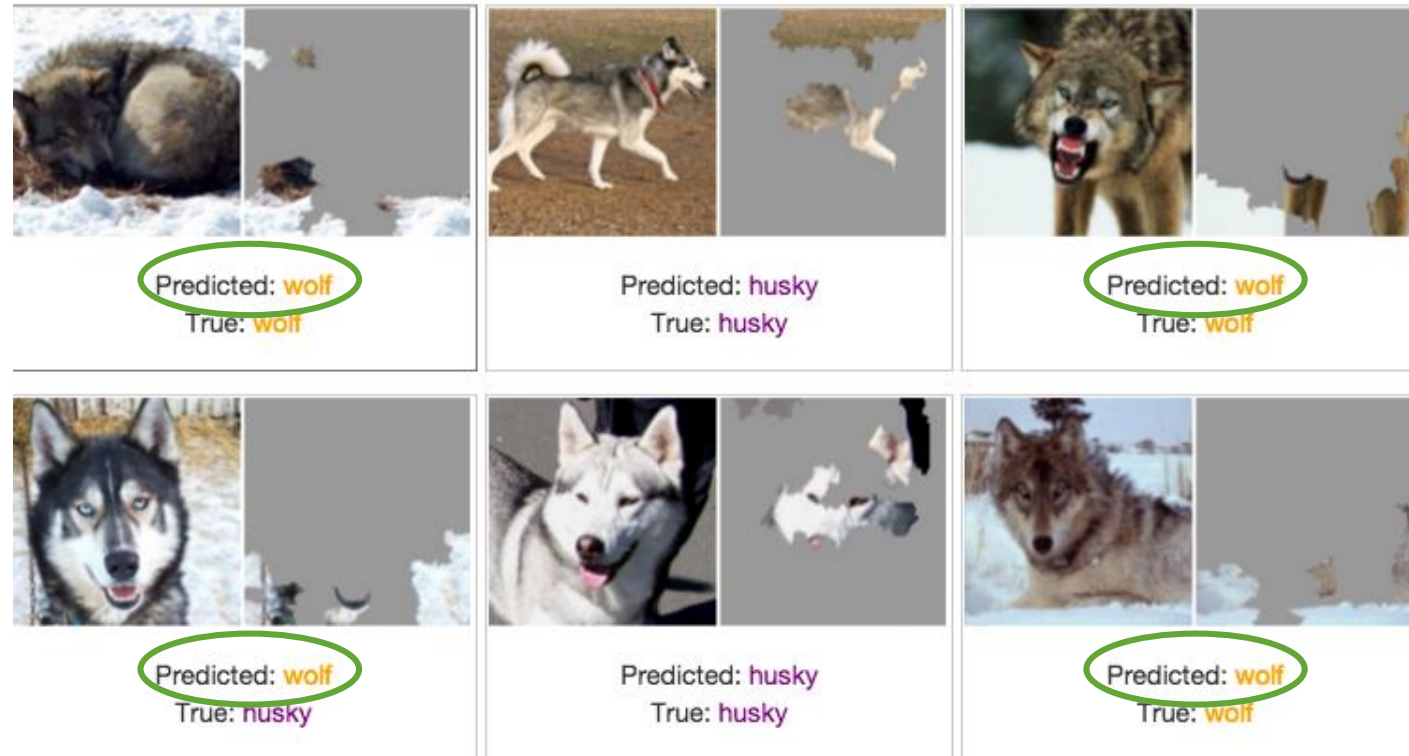
Understand what
ML is doing

Compare different
ML algorithms

Improve the
existing model

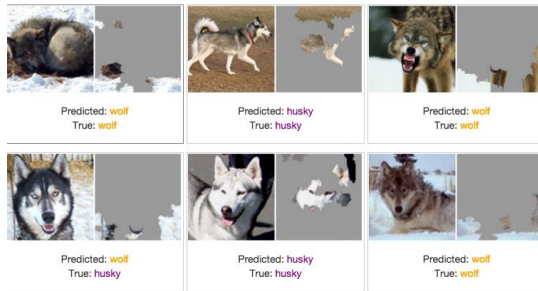
Predict how
ML will behave

Understanding Behavior



We've built a great snow detector...

Understanding Behavior



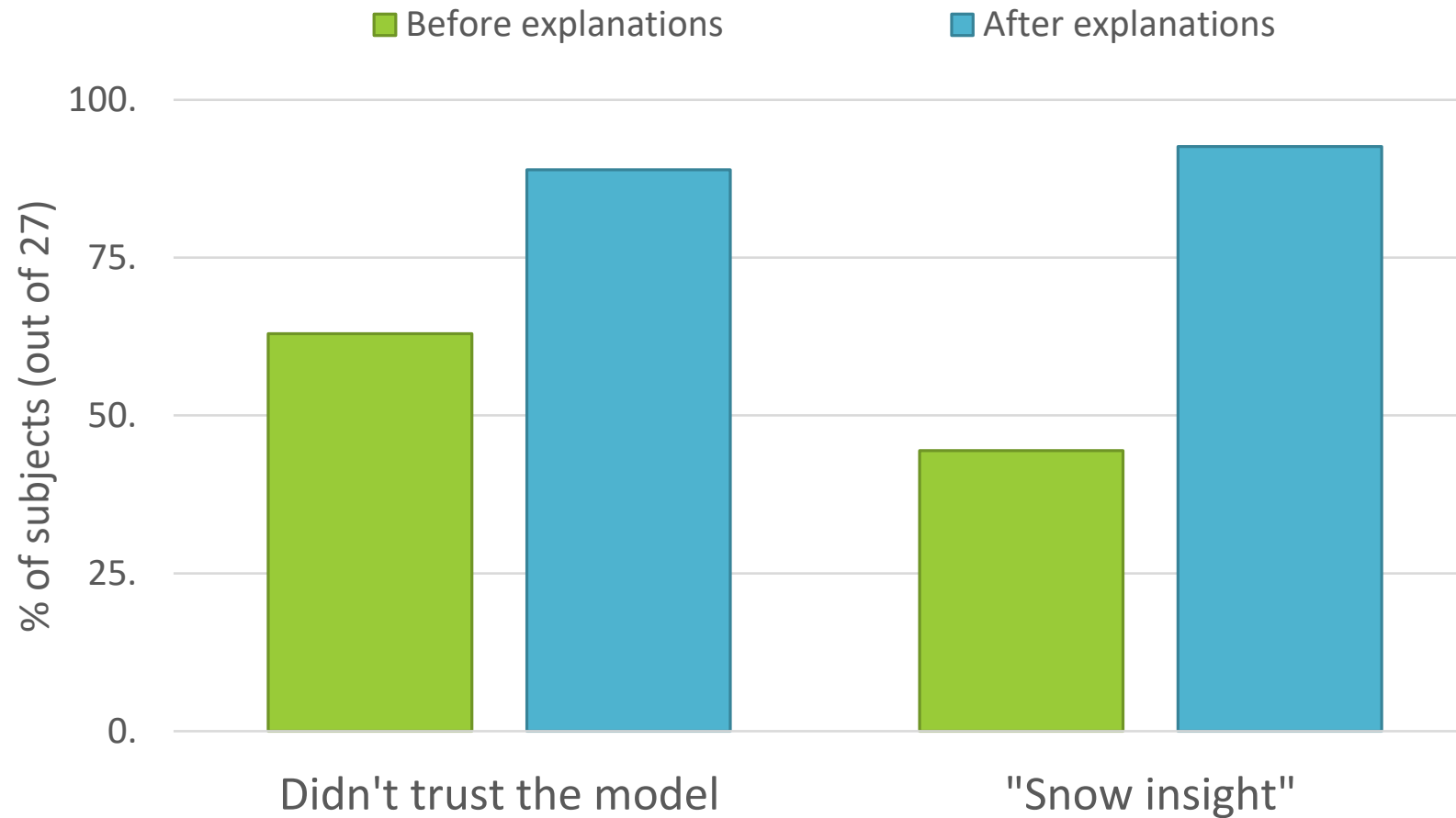
Question 1

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

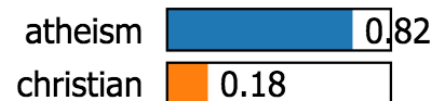
Explanation for a bad classifier

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

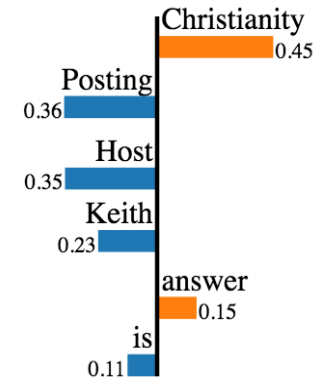
I think Christianity is the one true religion.
If you'd like to know more, send me a note



Prediction probabilities

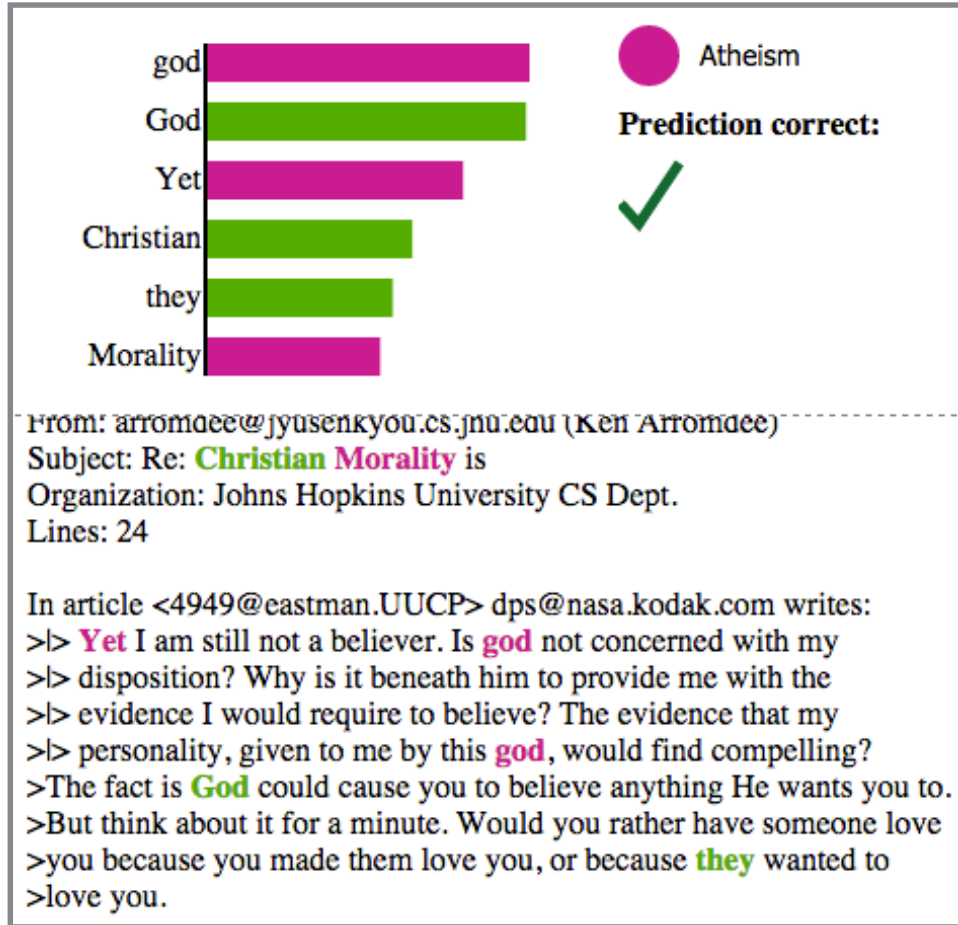


atheism christian



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


“Good” Explanation



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

UI for Comparing Classifiers

Example #3 of 6

True Class:  Atheism

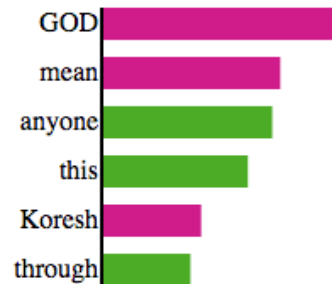
Instructions

Previous

Next

Algorithm 1

Words that A1 considers important:



Predicted:

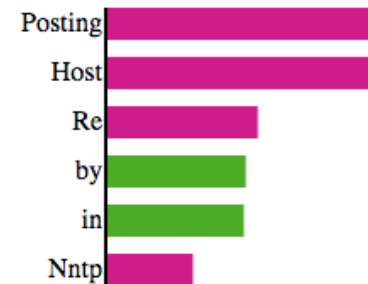
 Atheism

Prediction correct:



Algorithm 2

Words that A2 considers important:



Predicted:

 Atheism

Prediction correct:



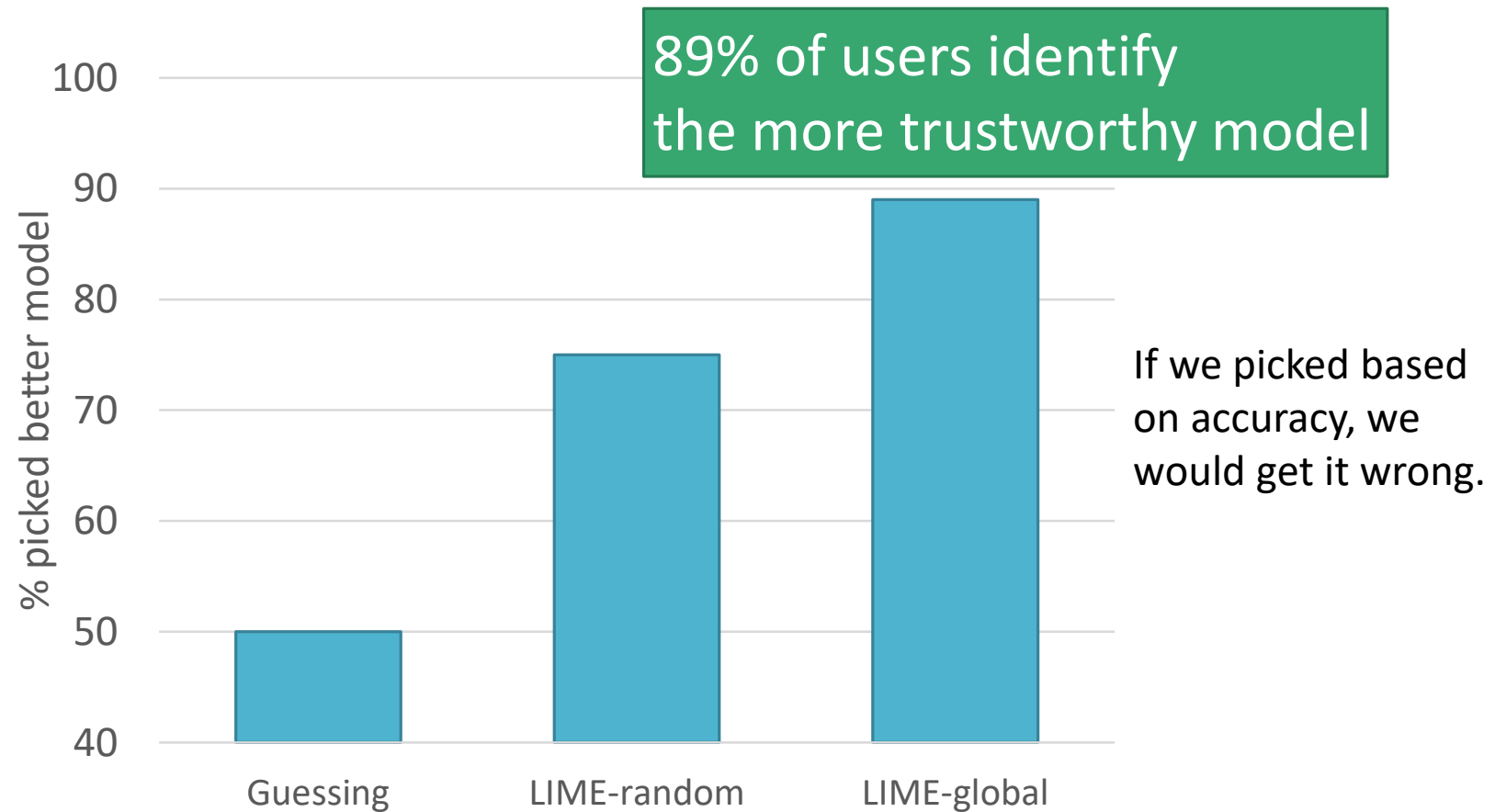
Document

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

Comparing Models



Quantitative Evaluation

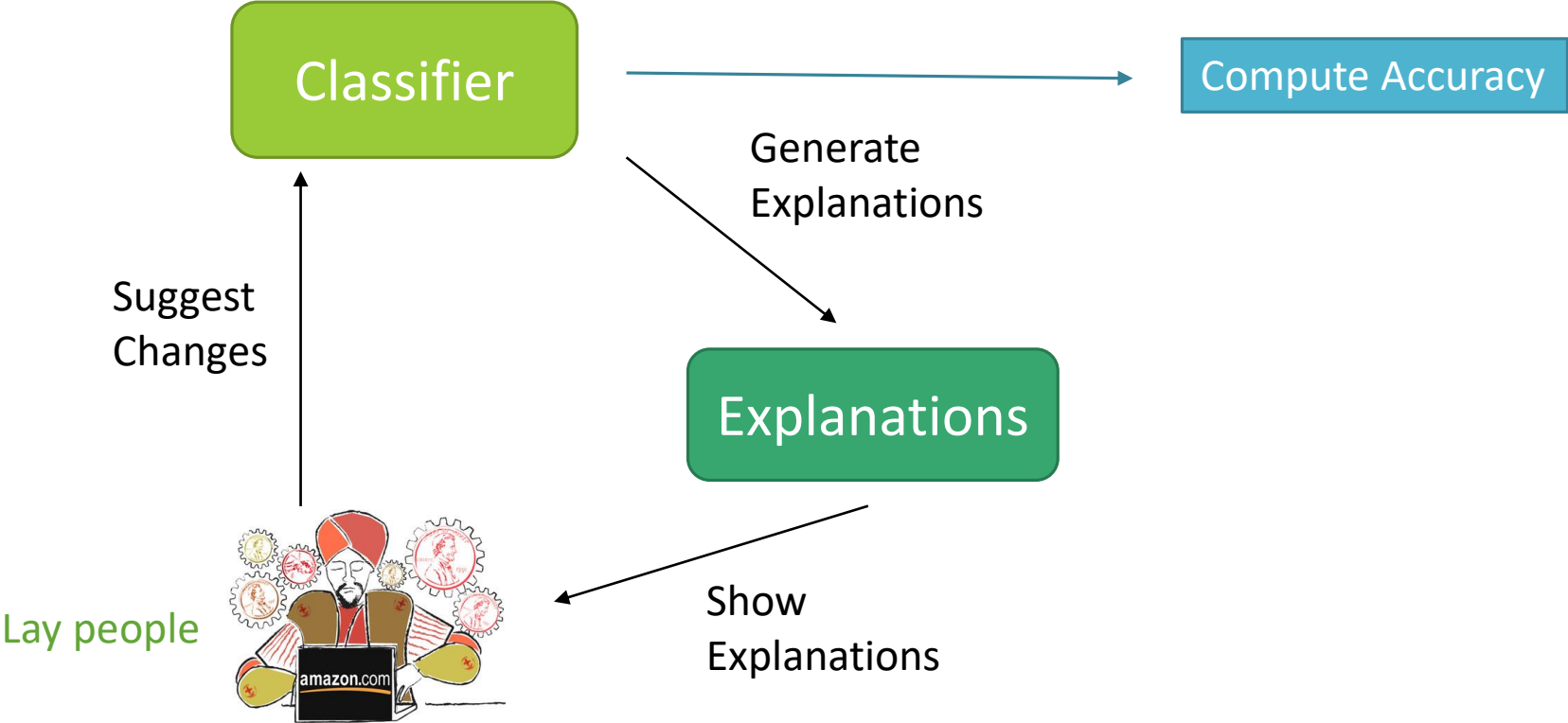
Understand what
ML is doing

Compare different
ML algorithms

Improve the
existing model

Predict how
ML will behave

Improving Classifiers



UI for fixing bad classifiers

Example #5 of 10

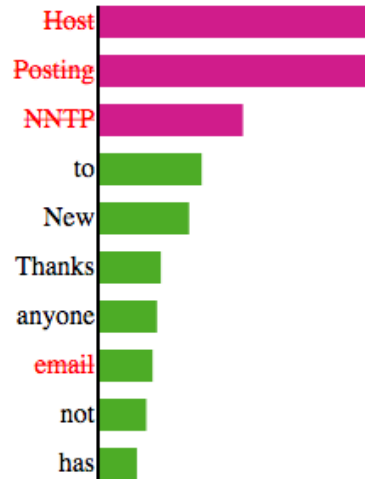
True Class: ● Atheism

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

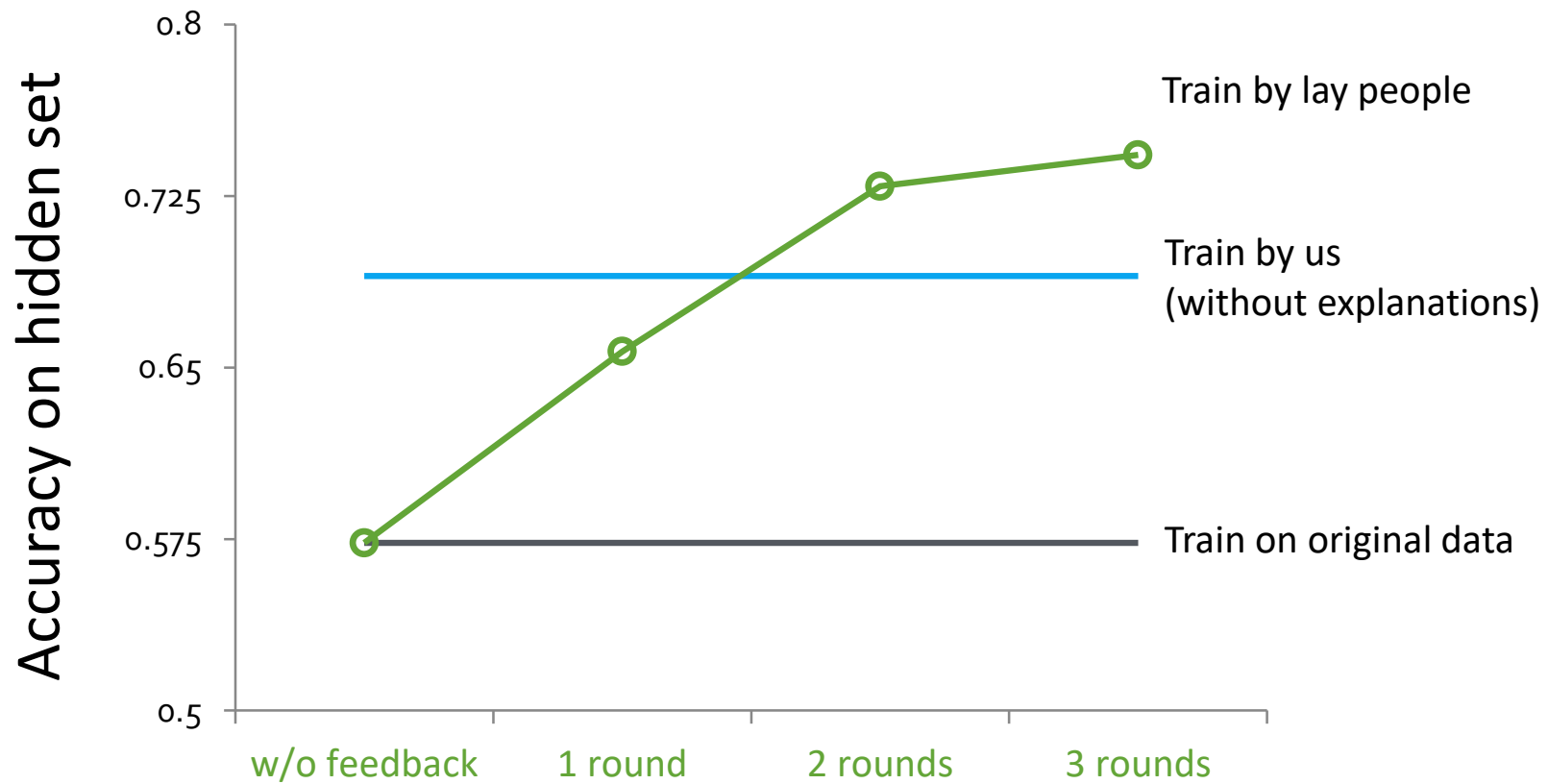
Hello Gang,

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

Fixing bad classifiers



Quantitative Evaluation

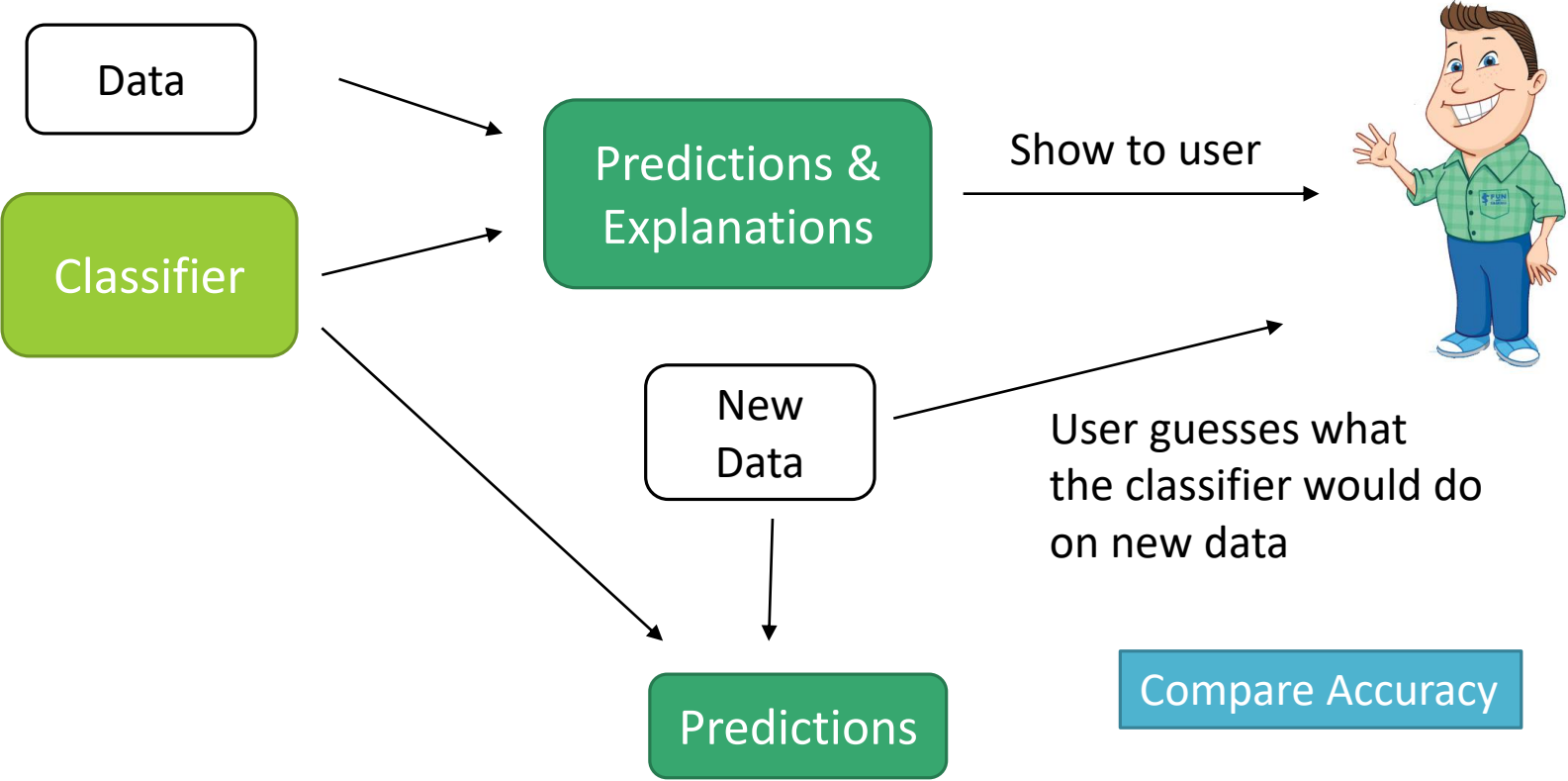
Understand what
ML is doing

Compare different
ML algorithms

Improve the
existing model

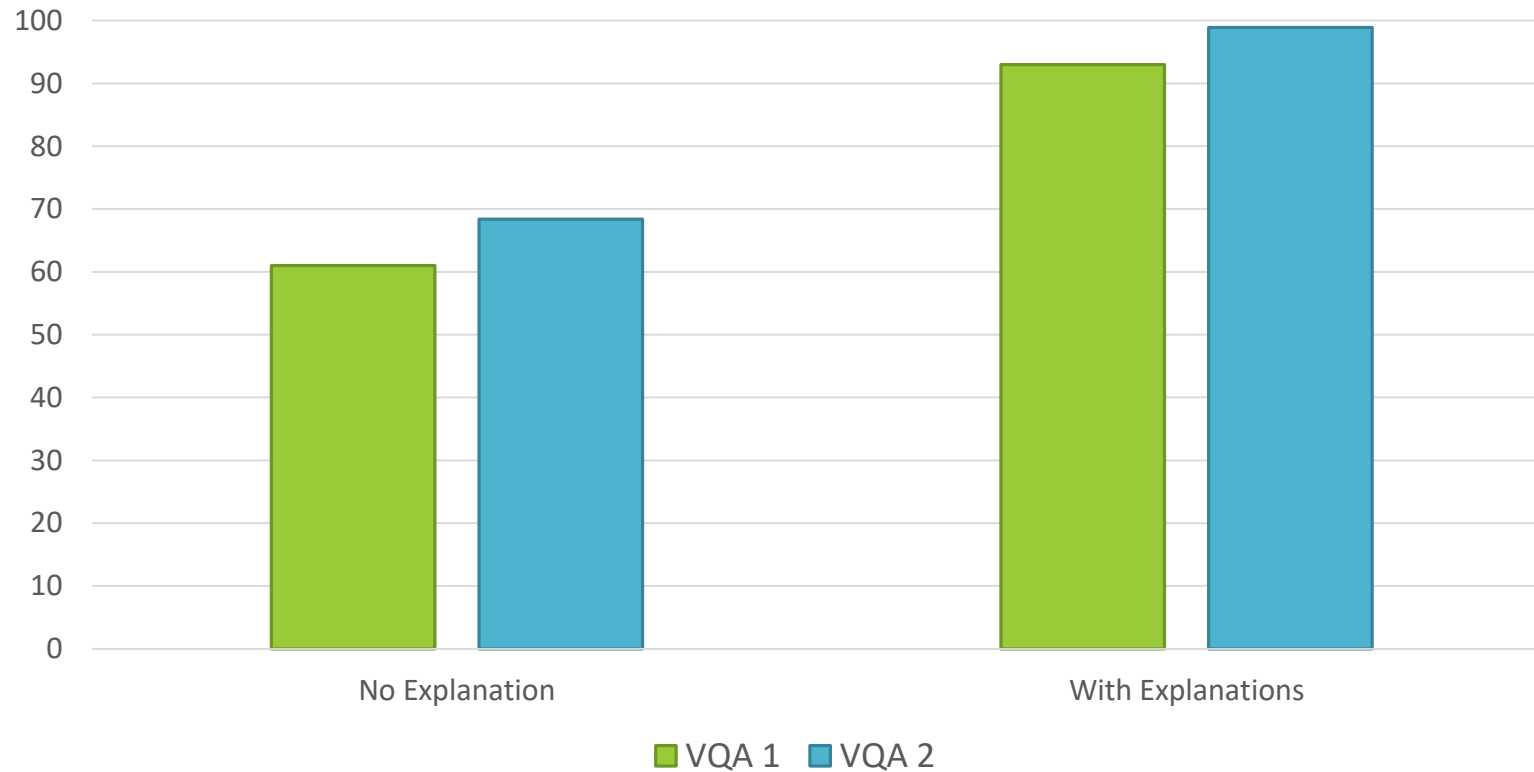
Predict how
ML will behave

Predicting Behavior



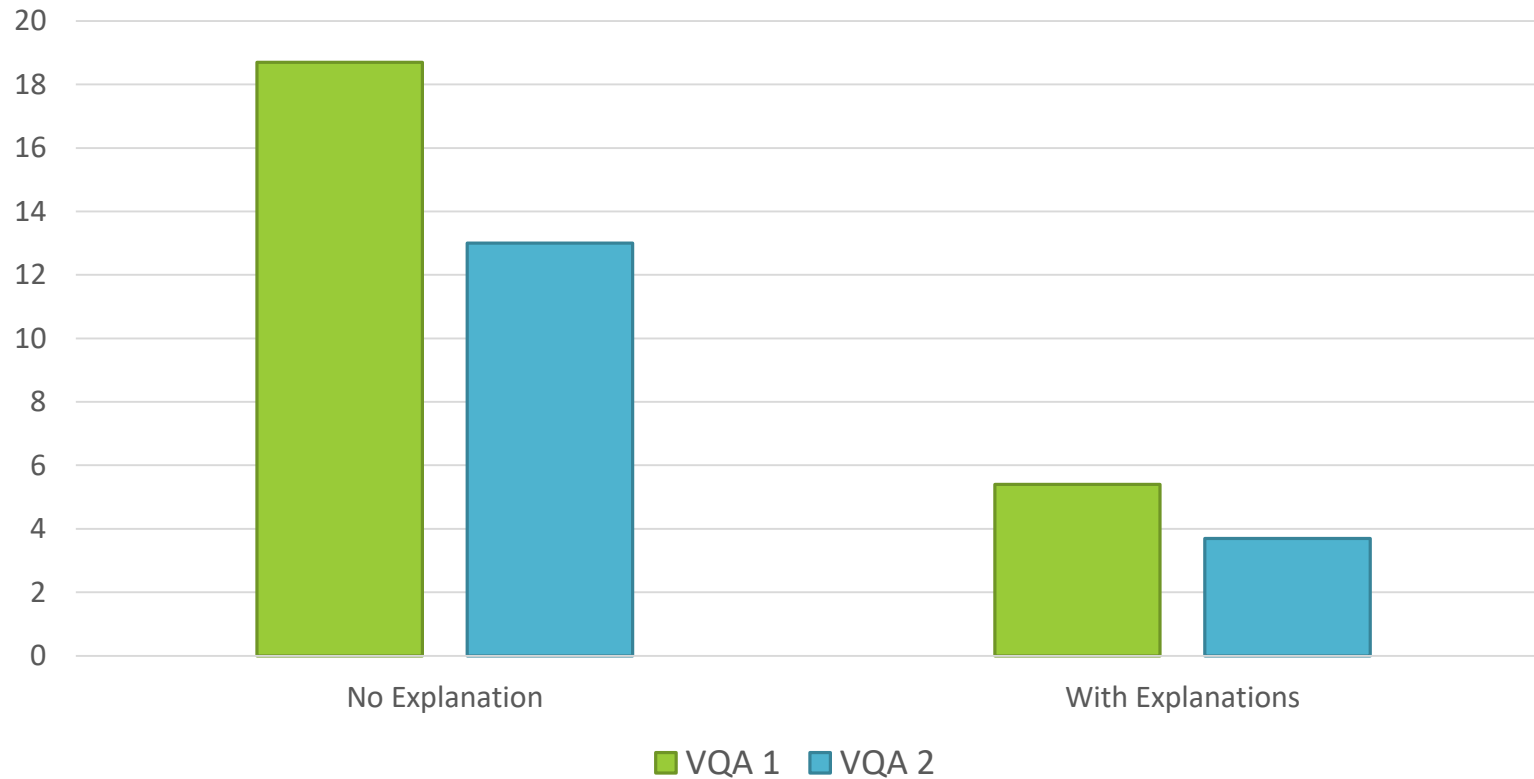
User Studies: *Precision*

How good are users guesses on unseen instances?

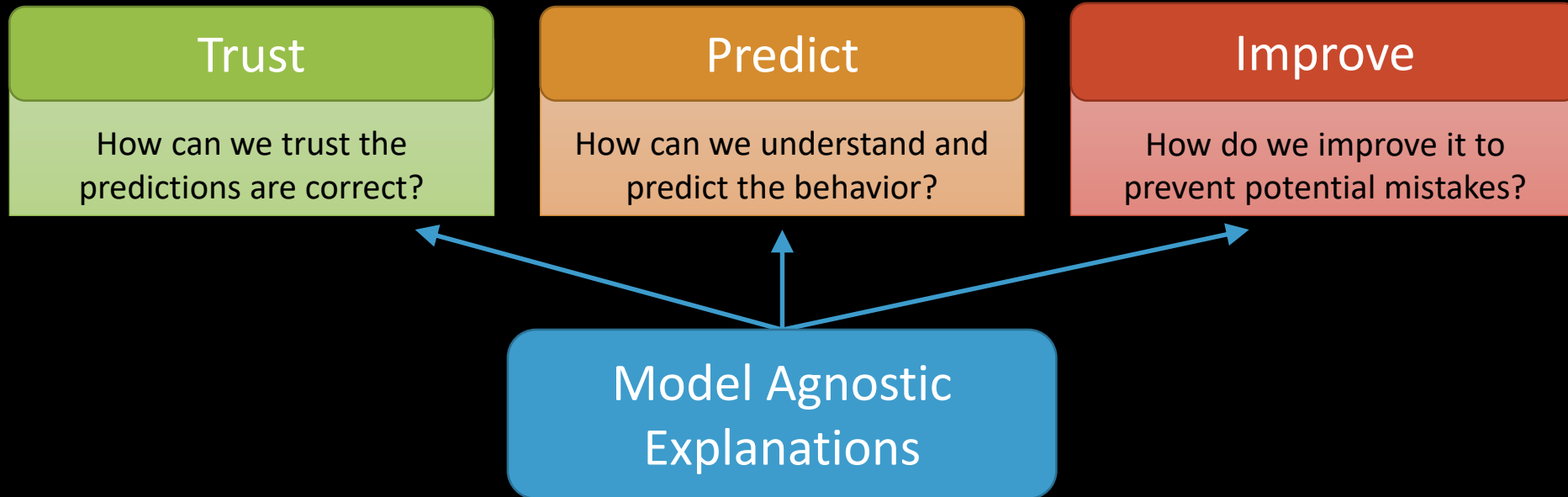


User Studies: *Time*

How long do users take to make their guesses?



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