Explaining Black-Box Machine Learning Predictions

Sameer Singh

University of California, Irvine

Machine Learning is Everywhere…

Google Translate

Classification: Wolf or a Husky?

Machine Learning Model

Wolf!

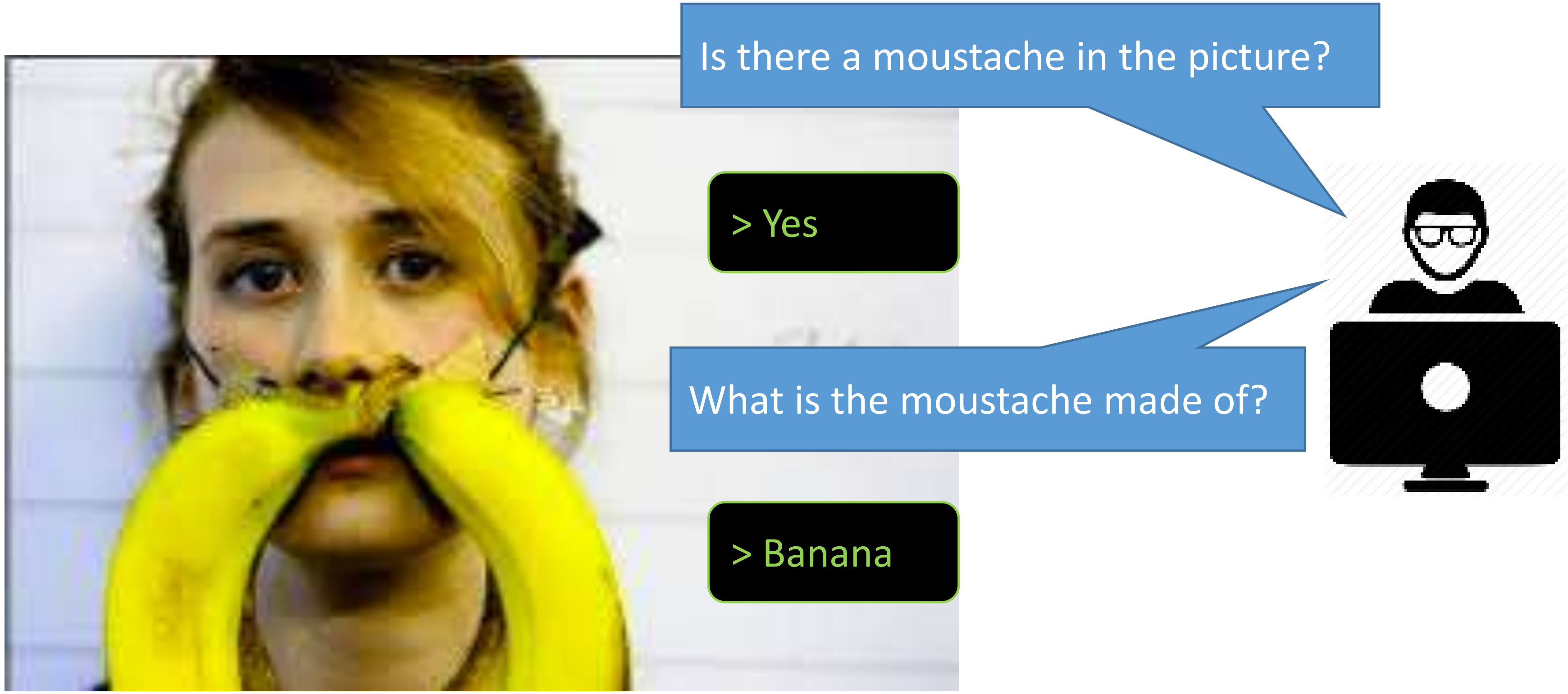
Classification: Wolf or a Husky?

→ Machine Learning Model → Husky!

Classification: Wolf or a Husky?

Only 1 mistake!

More Complex: Question Answering



Essentially black-boxes!

Trust

How can we trust the predictions are correct?

How do we know they are not breaking regulations?

How do we avoid "stupid mistakes"?

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!

We've built a snow detector…

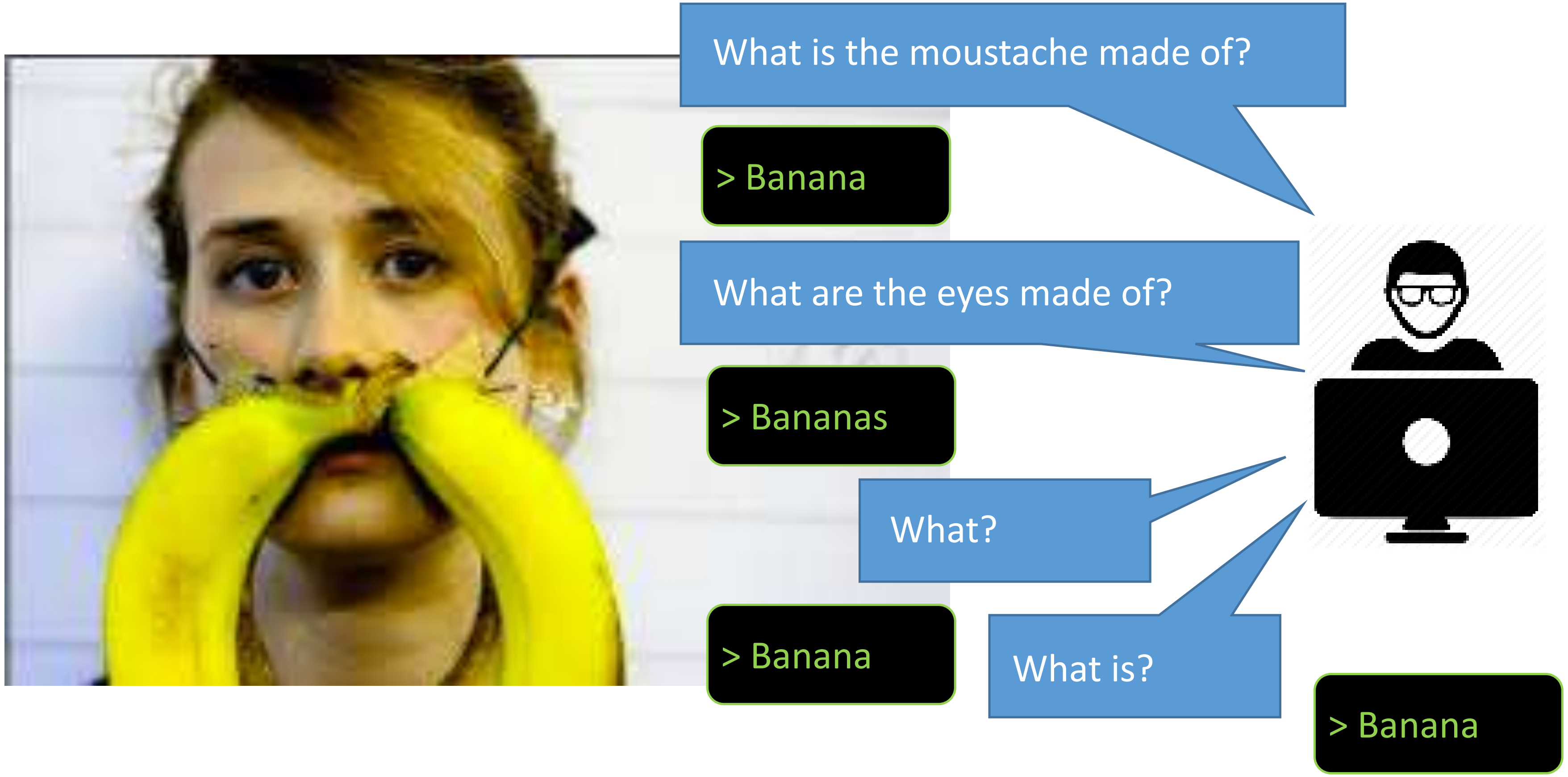
VIDEO SLATE IN MOTION. OCT. 14 2016 3:18 PM

The Man Who Accidentally Adopted a Wolf Pup

It did not go well.

By A.J. McCarthy

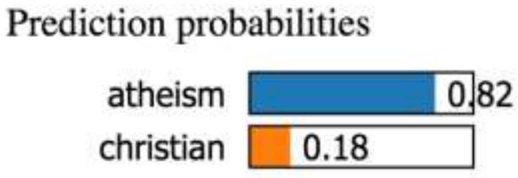
Visual Question Answering

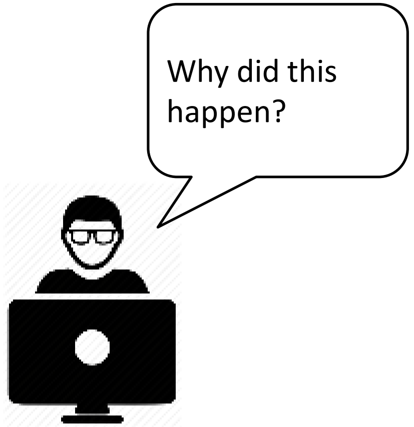


Text Classification

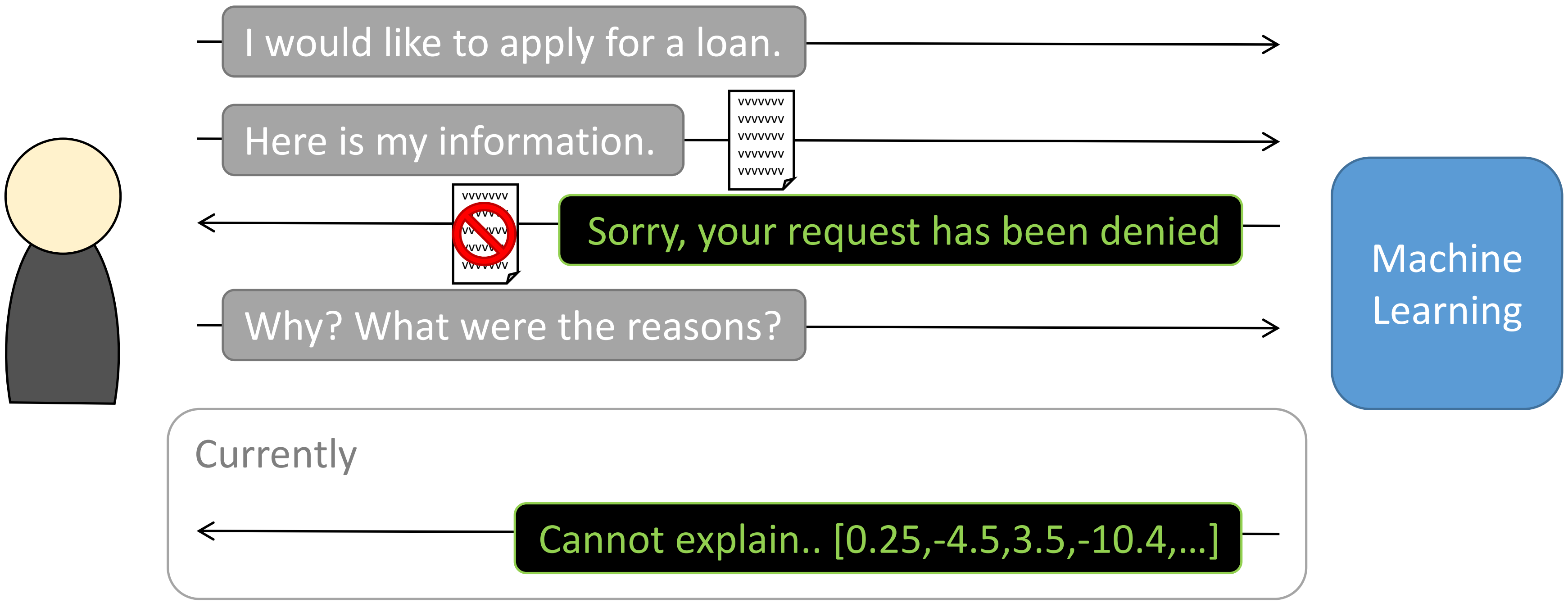
From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

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





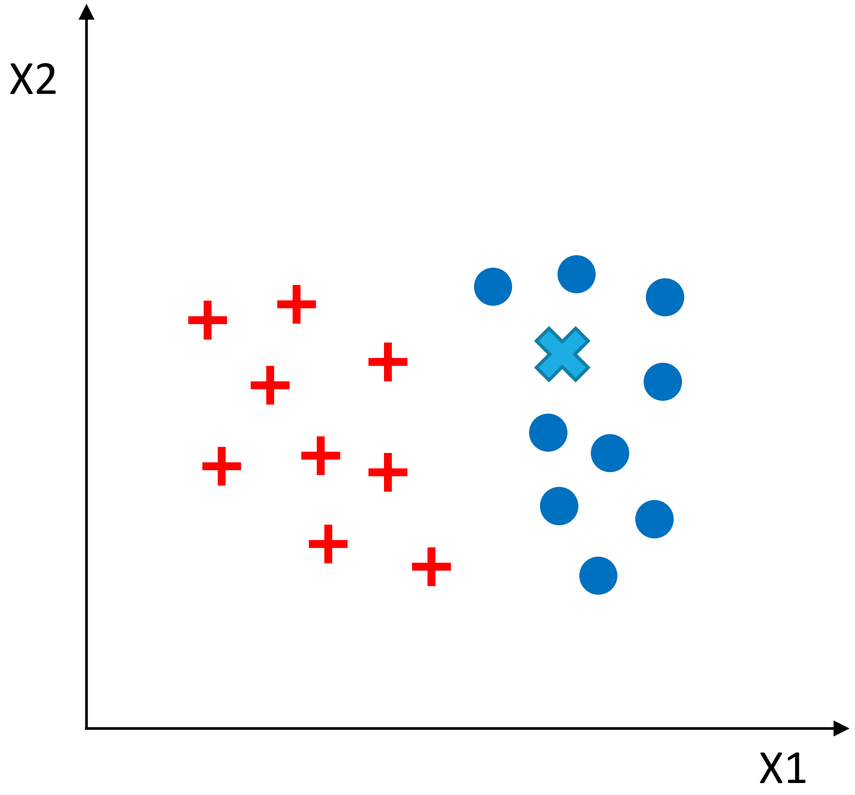
Applying for a Loan



How did we get here?

Big Data and Deep Learning

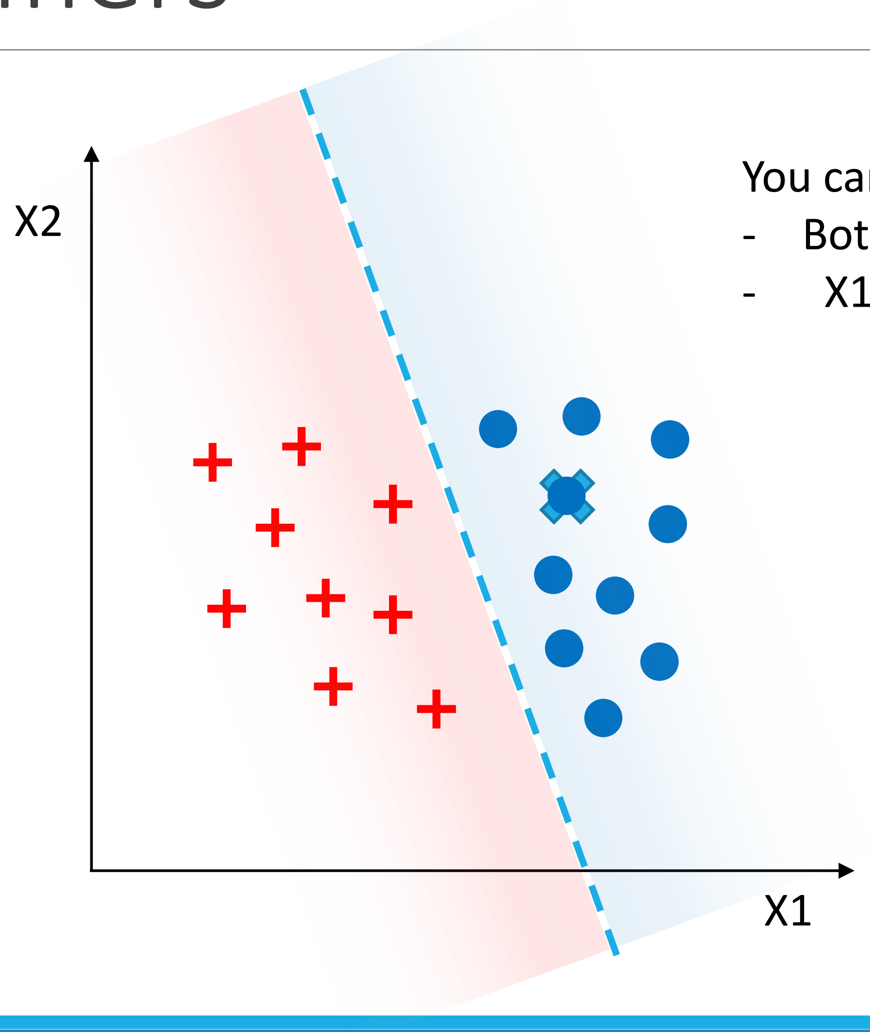
Simple Data



Linear Classifiers

if: $10X1 + X2 - 5 > 0$

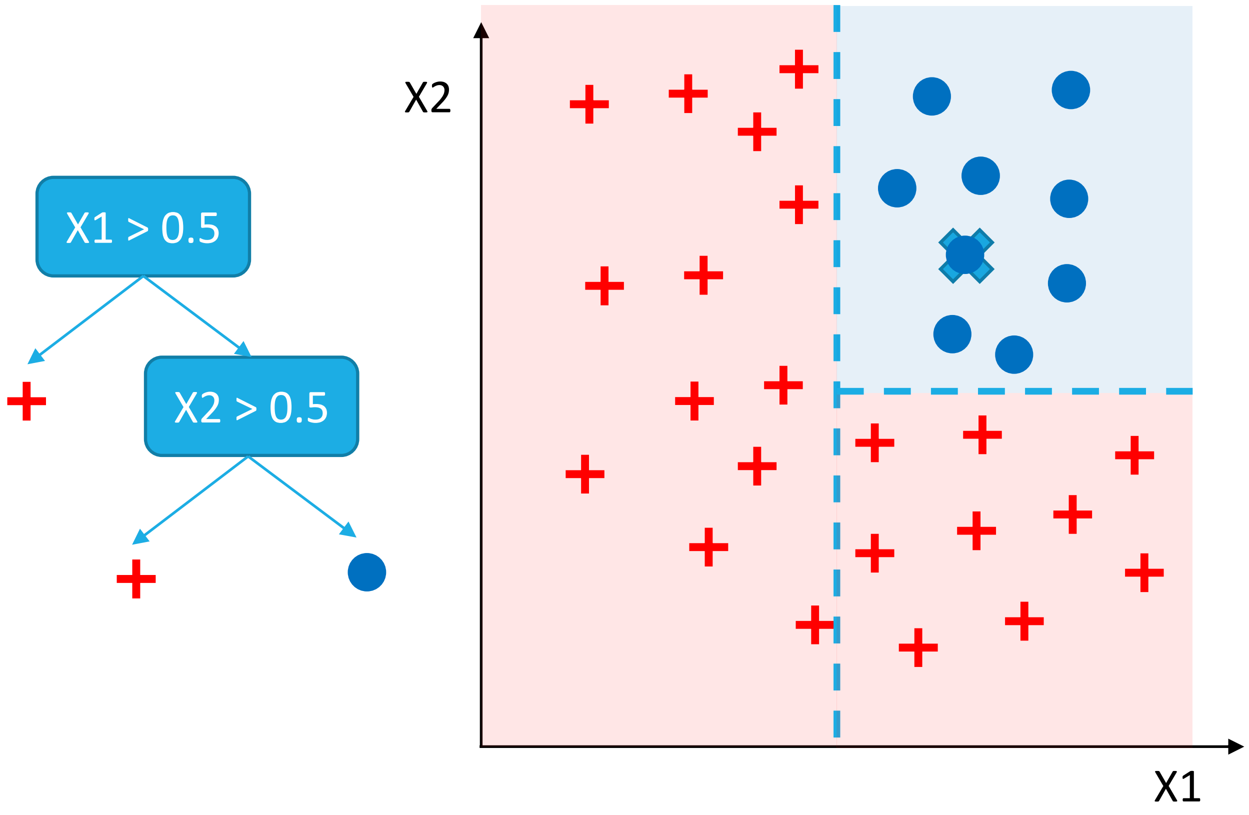
otherwise



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

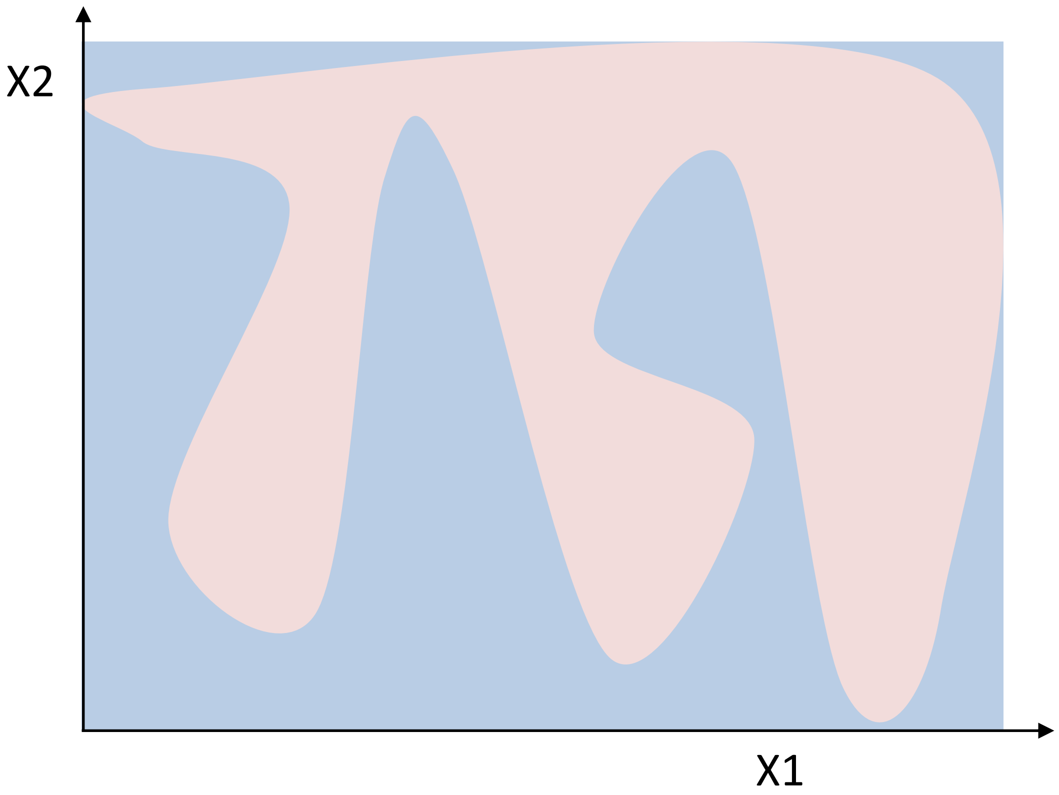
How do we improve it to prevent potential mistakes?

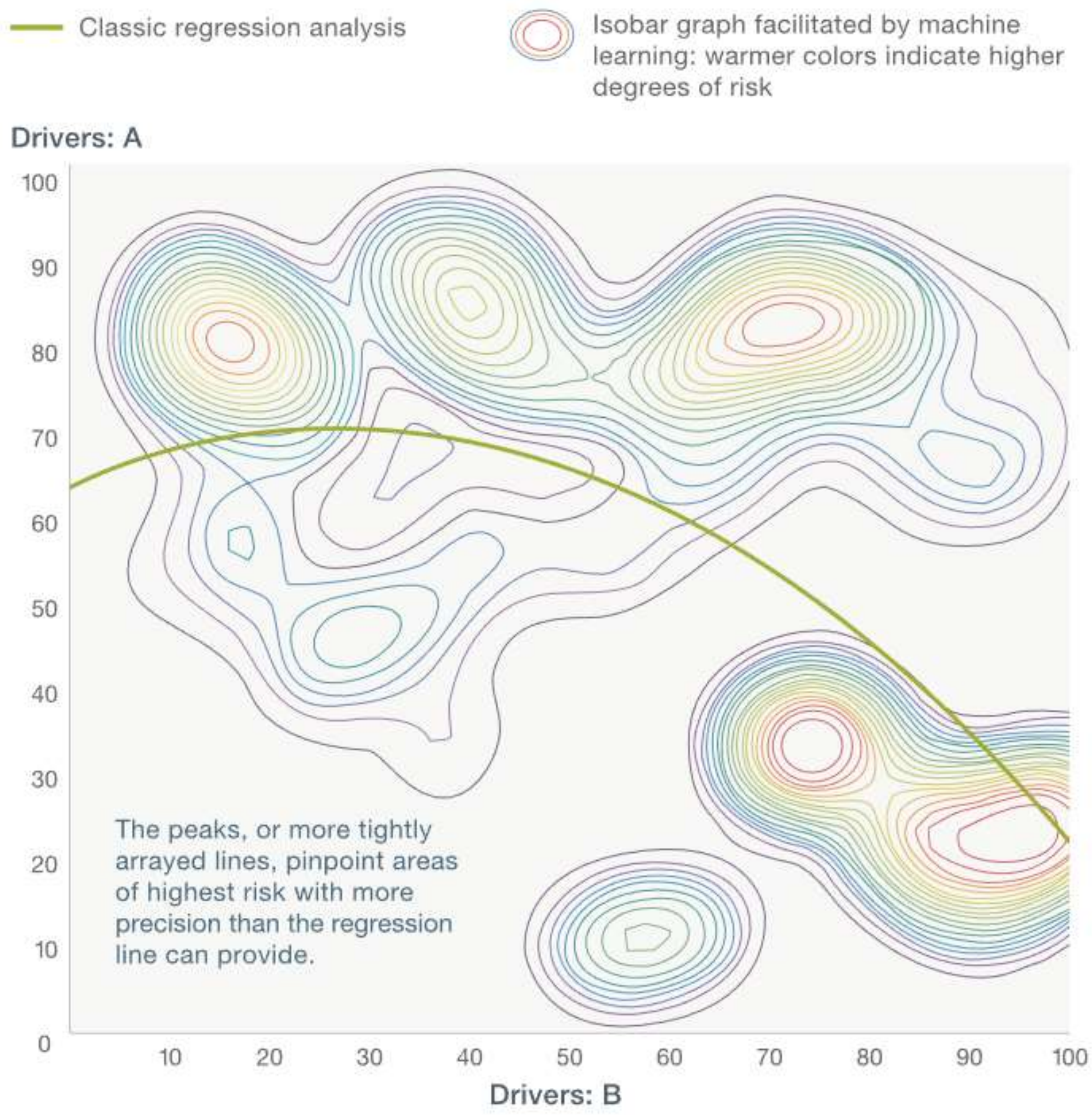
→ Structure tells you where the error is, thus how to fix it.

Arrival of Big Data

Big Data: Applications of ML

Big Data: More Complexity





McKinsey&Company

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

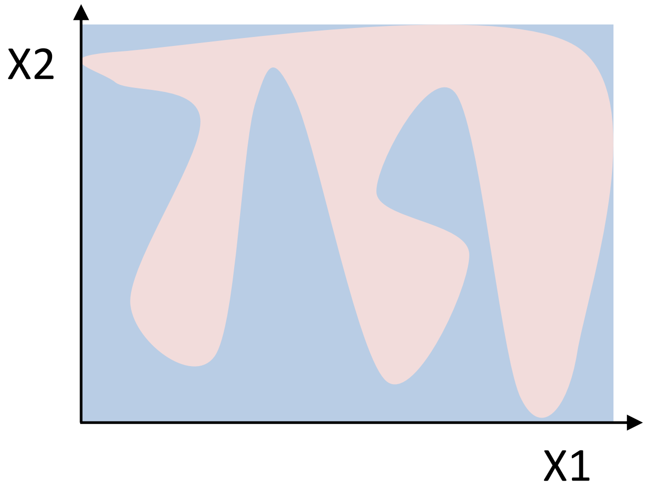
Recent defaults

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

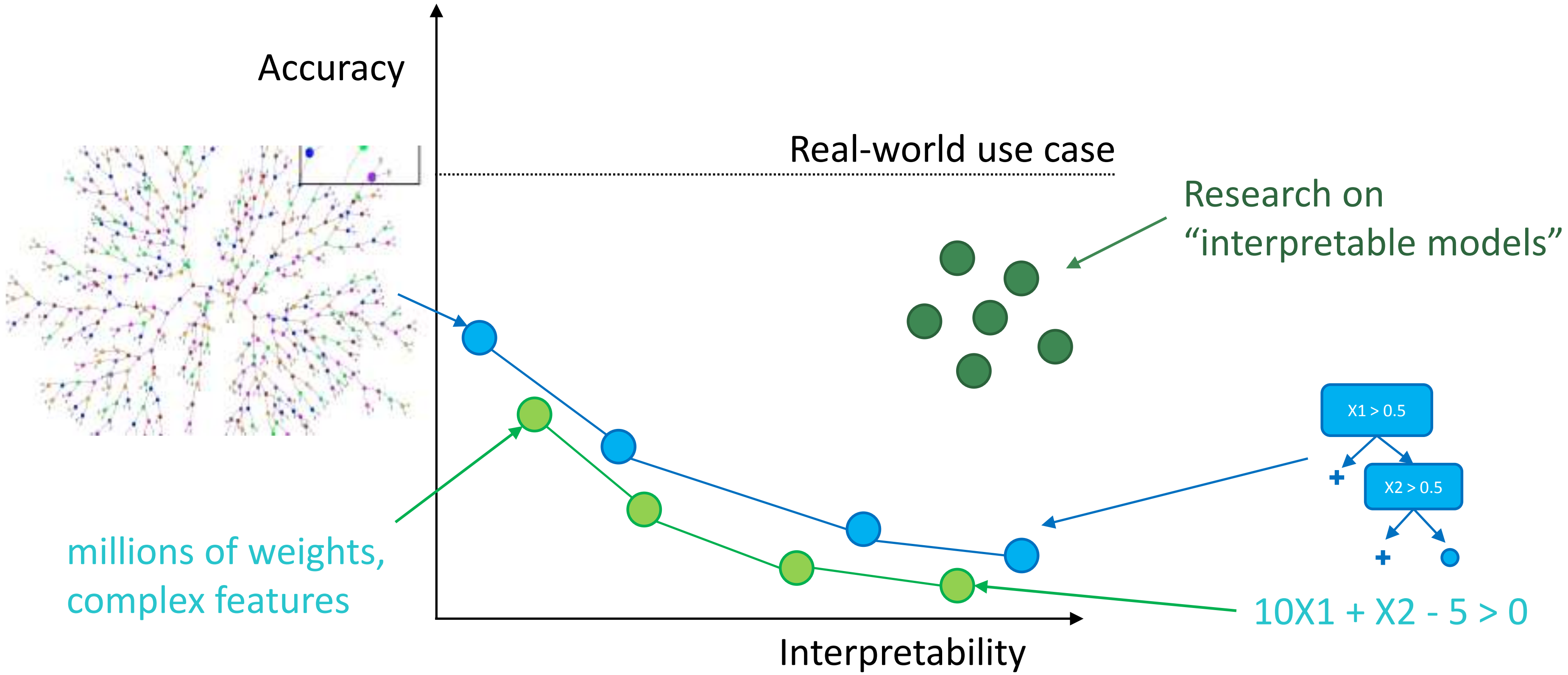
Recent defaults

…

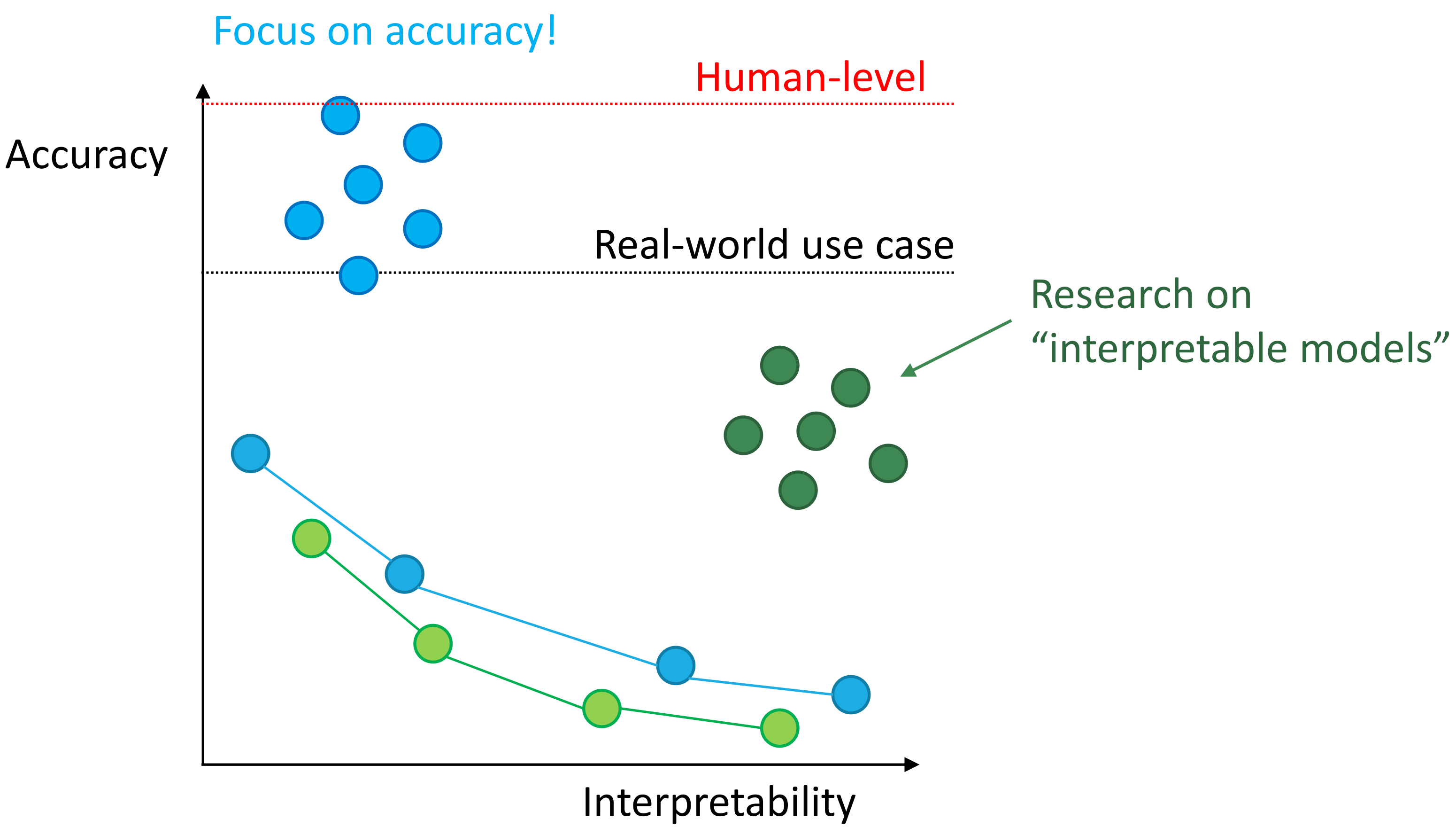
Complex Surfaces + Lots of dimensions

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?

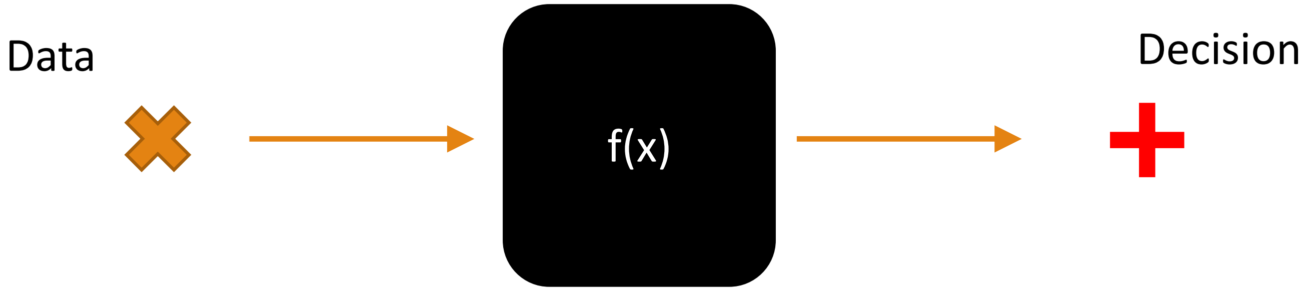
→ Structure tells you where the error is, thus how to fix it.

Explaining Predictions

The LIME Algorithm

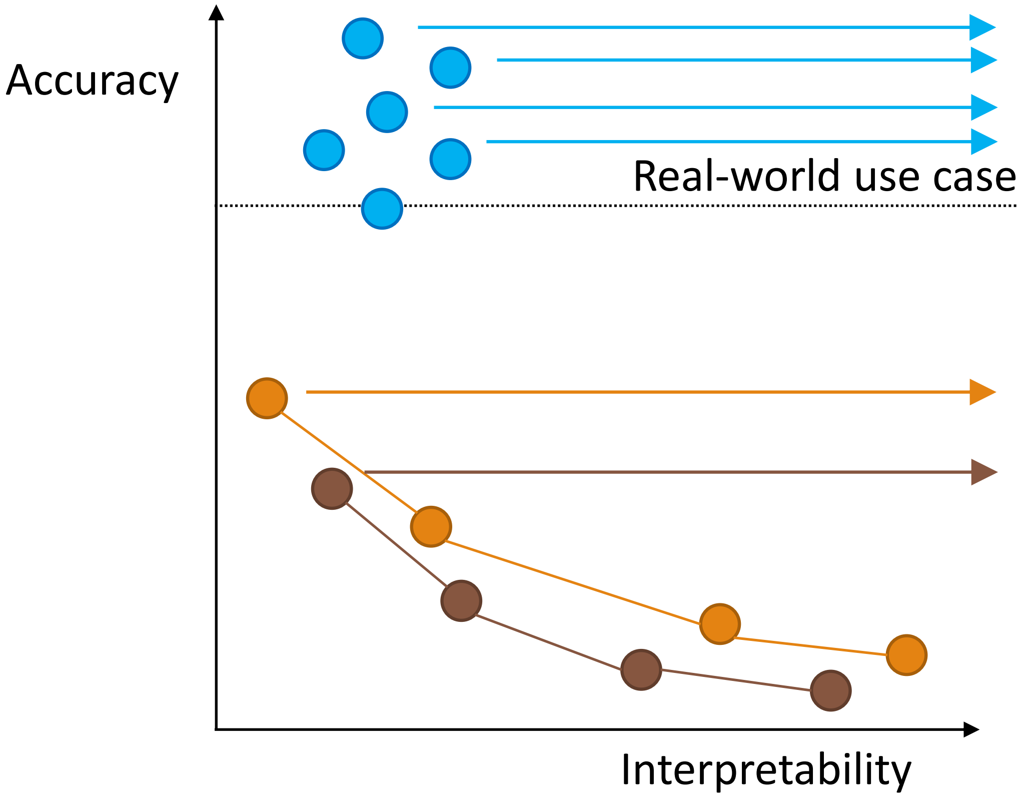
Being Model-Agnostic…

No assumptions about the internal structure…



Explain any existing, *or future*, model

LIME: Explain Any Classifier!

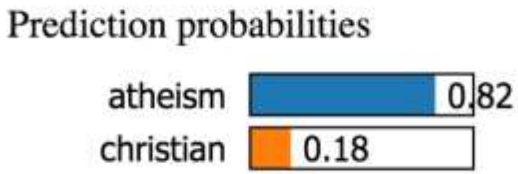


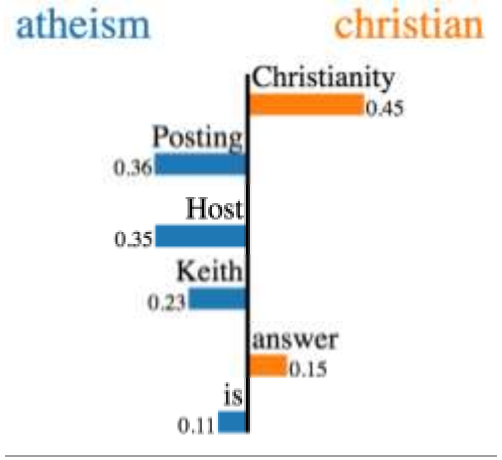
Make everything interpretable!

What an explanation looks like

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

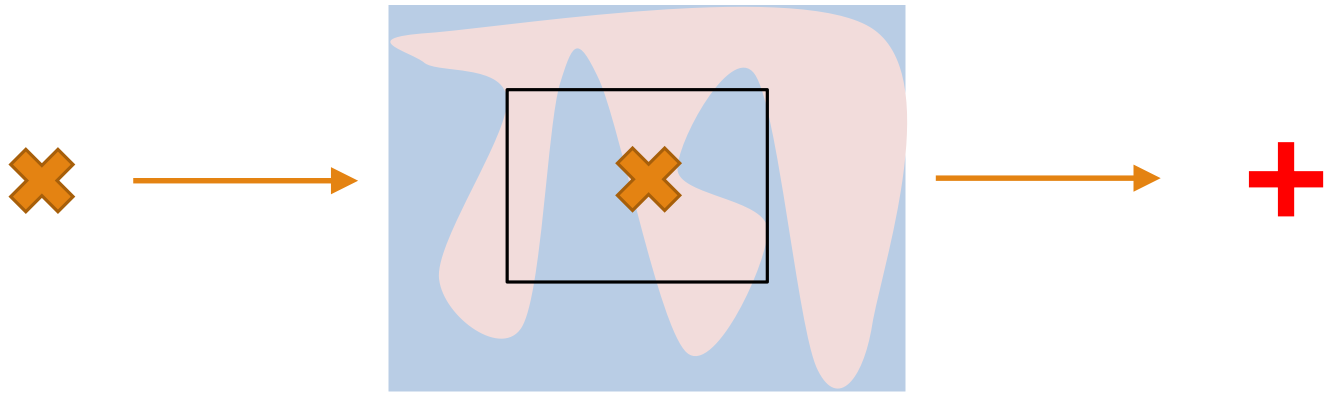
I think Christianity is the one true religion.
If you'd like to know more, send me a note





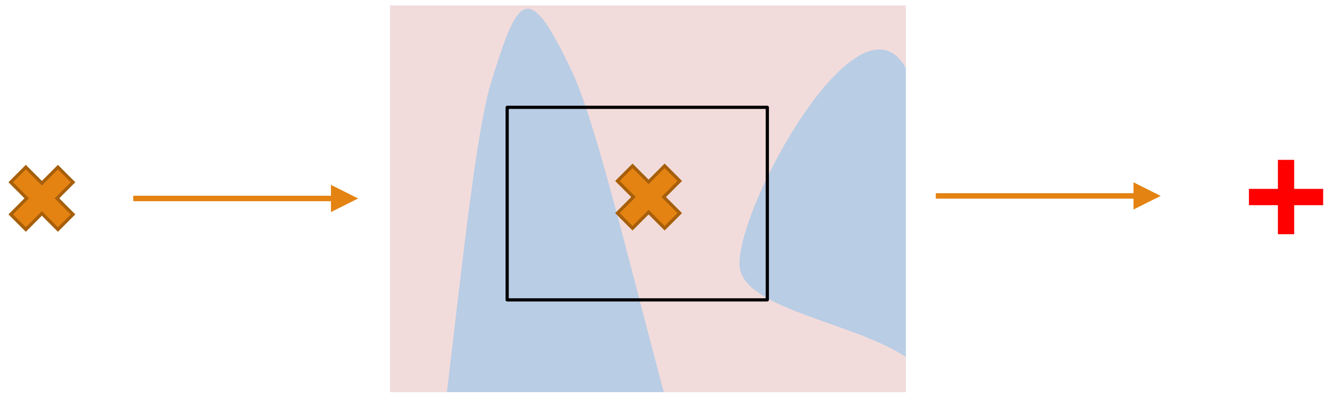
Being Model-Agnostic…

“Global” explanation is too complicated



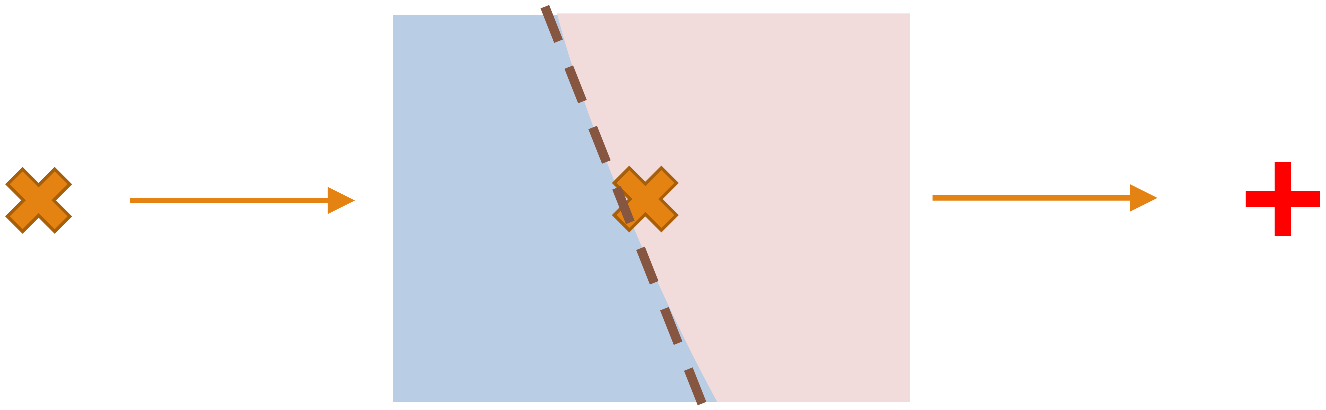
Being Model-Agnostic…

"Global" explanation is too complicated



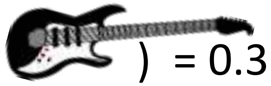
Being Model-Agnostic…

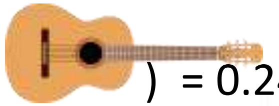
“Global” explanation is too complicated



Explanation is an interpretable model, that is locally accurate

Google's Object Detector

P() = 0.32

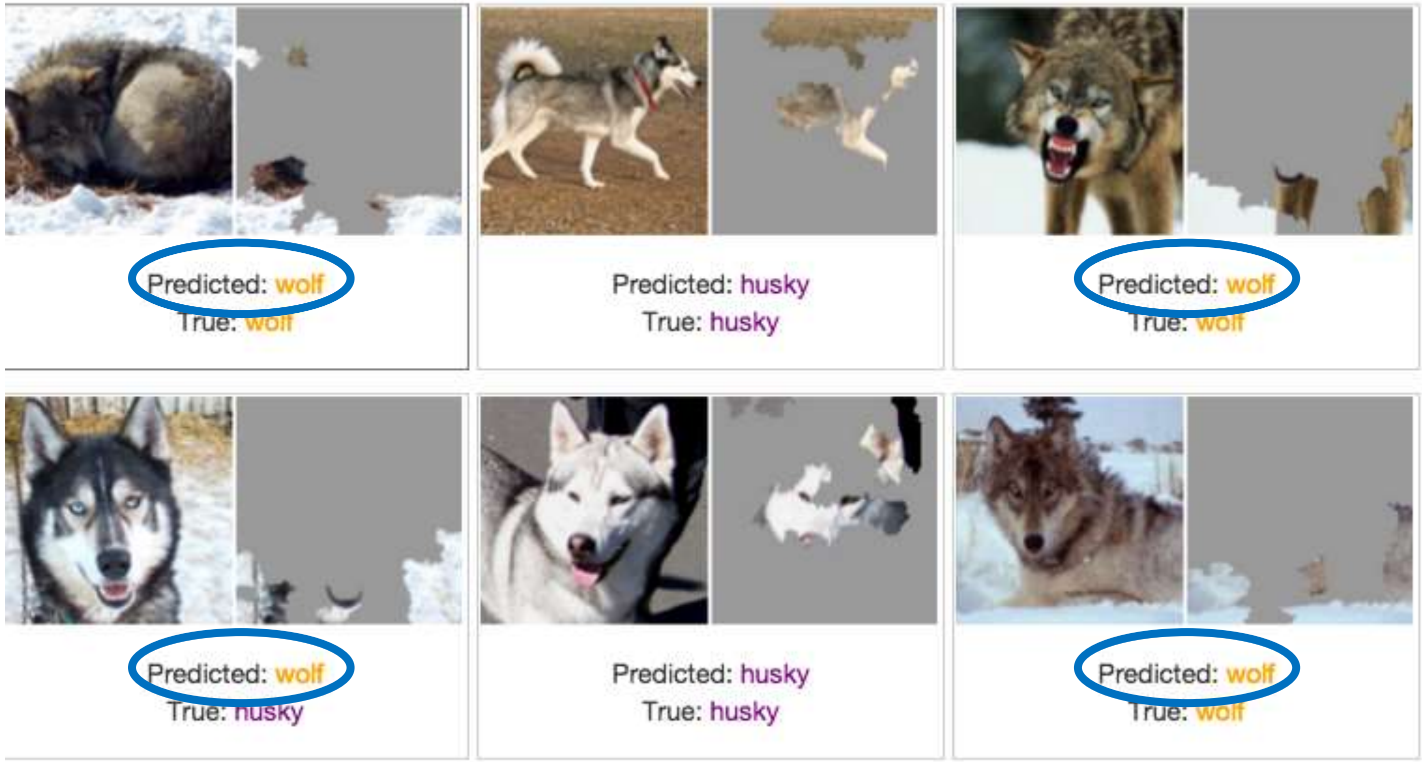
P() = 0.24

P() = 0.21

Classification: Wolf or a Husky?

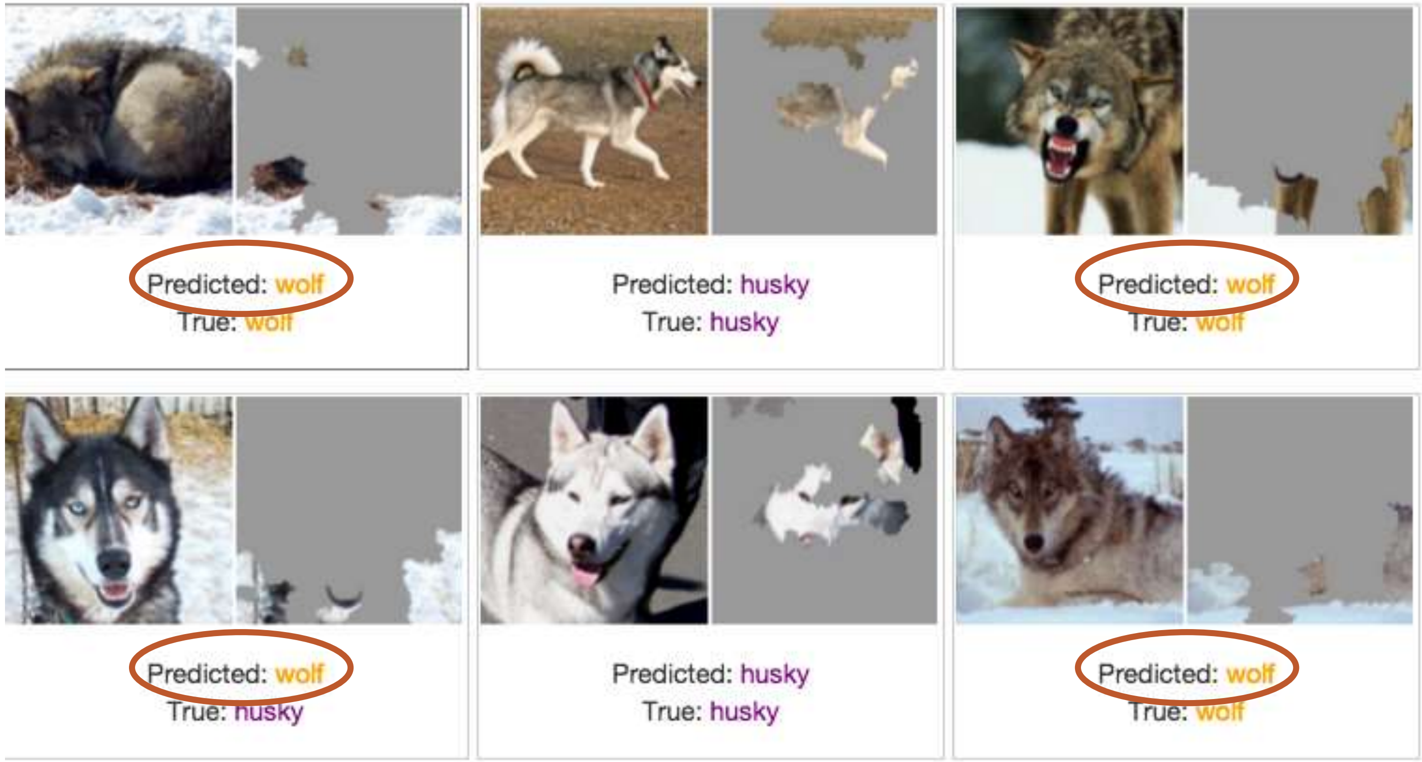
Only 1 mistake!

Neural Network Explanations



We've built a great snow detector…

Understanding Behavior



We’ve built a great snow detector…

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

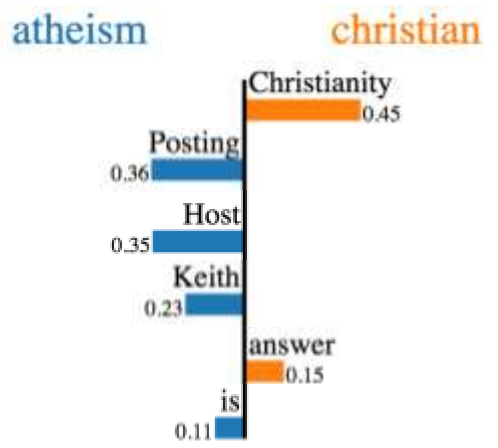
“Bad” Classifier

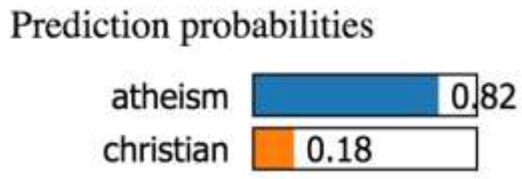
“Good” Classifier

Explanation for a bad classifier

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

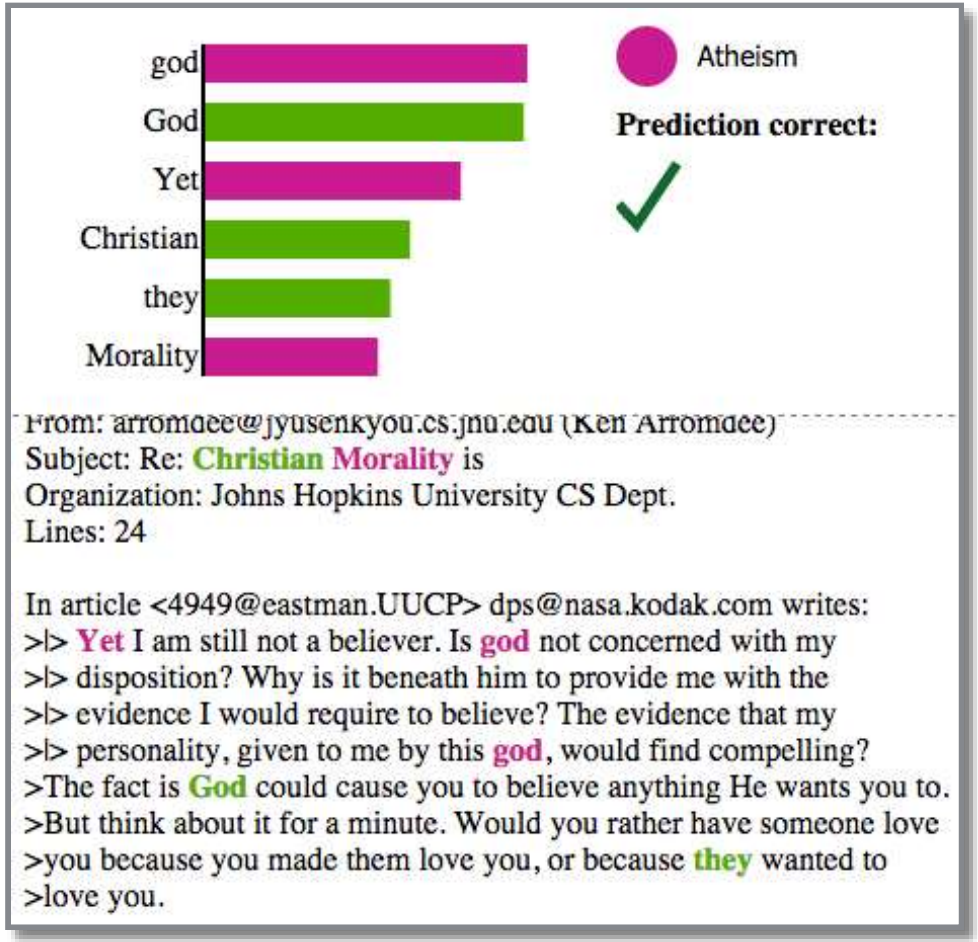
I think Christianity is the one true religion.
If you'd like to know more, send me a note





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

“Good” Explanation



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

Recent Work

Counter-examples and Counter-factuals

Understanding via Predicting

Users "understand" a model if they can predict its behavior on unseen instances

Precision

How accurate are the users guesses?
If the users guess wrong, they don't understand

Coverage

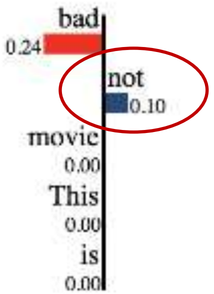
How often do the users make confident guesses?
It's okay not to be able to guess!

Precision is much more important than Coverage!

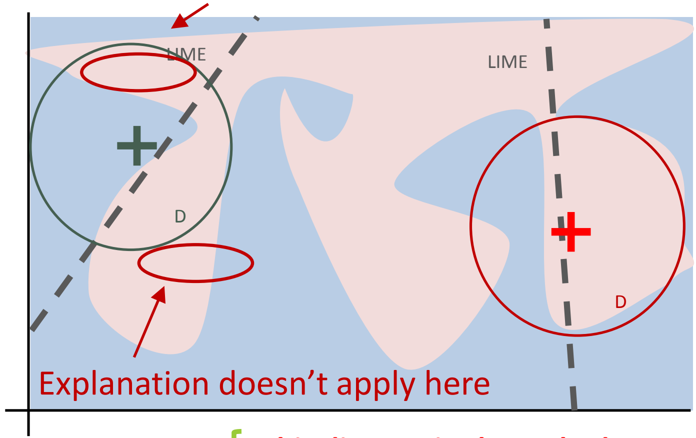
It's much better not to guess than to guess confidently, but be completely wrong!

Linear Explanations

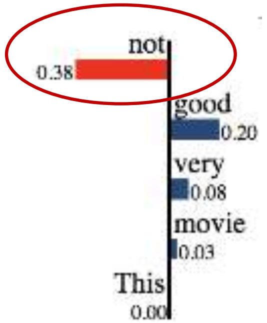
This movie is not bad.



Explanation is wrong in this region



This movie is not very good.



This explanation is a better approximation than the other one.

D

- **This director is always bad.**
- **This movie is not nice.**
- **This stuff is rather honest.**
- **This star is not bad.**
- ...

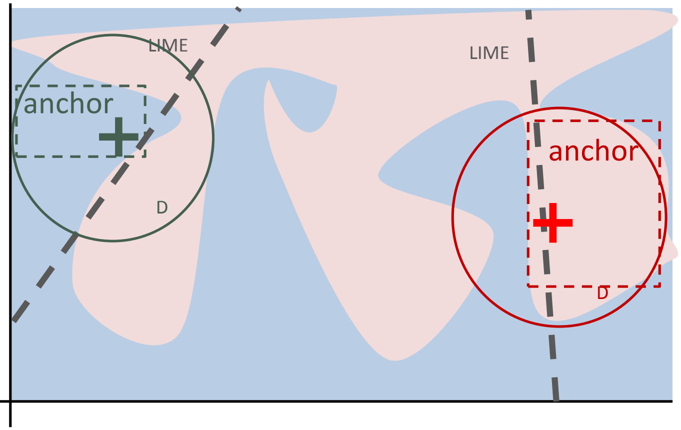
Problem 1: Where is the explanation good?

Problem 2: What is the coverage?

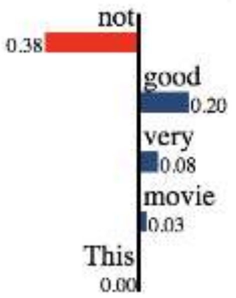
→ Users will make mistakes!

Anchors: Precise Counter-factuals

This movie is not bad.



This movie is not very good.



Anchor: ”**not bad**” → Positive

Anchor: ”**not good**” → Negative

D(.|A)

This audio is **not bad**.
This novel is **not bad**.
This footage is **not bad**.

An anchor is a **sufficient condition**

D(.|A)

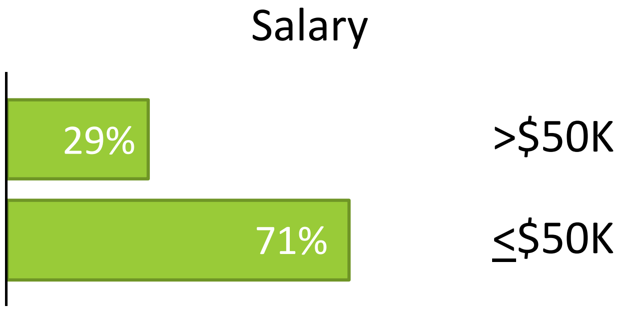
This poster is **not** ever **good**.
This picture is **not** rarely **good**.
This actor is **not** incredibly **good**.

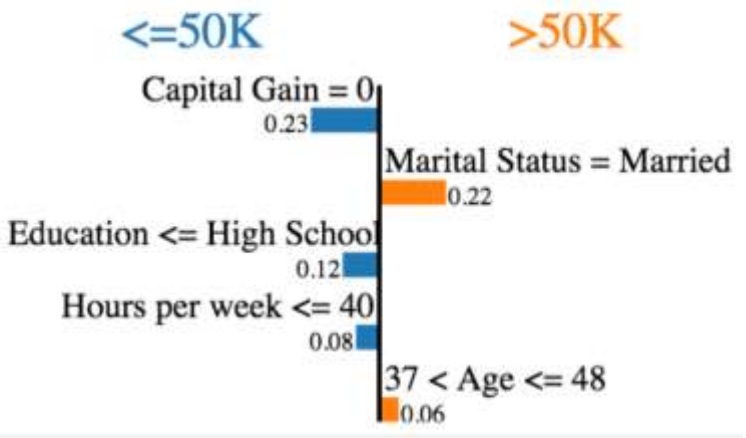
Clear (and adaptive) coverage

Probabilistic guarantee avoids human mistakes

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

Visual QA

What is the mustache made of?	banana

How **many** bananas are in the picture?	2

Encoder/Decoder LSTMs

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

Encoder/Decoder LSTMs

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.

Encoder/Decoder LSTMs

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.

What’s a Good Explanation?

We want to understand the models

Compact description

Lines, Decision Trees,
Simple Rules, etc.

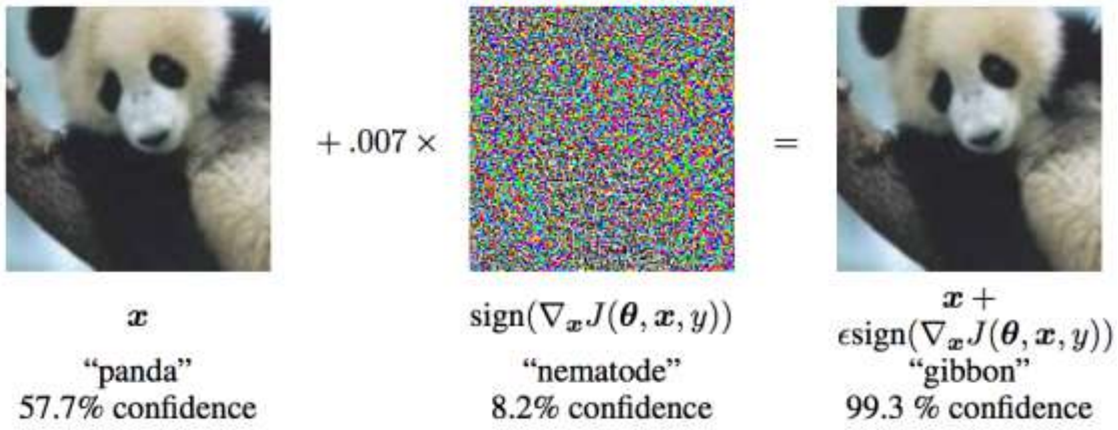
When we read them,
we imagine instances
where they apply, and
where they don’t

Directly show useful examples?

What examples describe the behavior?

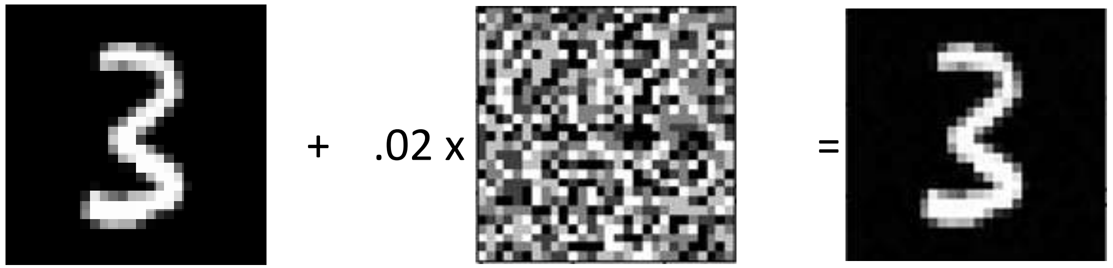
Closest **Counter-example**:
How can we change this example
to change the prediction?

Adversarial Examples



Goodfellow et al, "Explaining and Harnessing Adversarial Examples", ICLR 2015.

"inputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed input results in the model outputting an incorrect answer with high confidence"



original MNIST digit "3"   adversarial noise   adversary predicted as "2"

Adversarial Examples: Pros

$$x^* = \text{argmin}_{\tilde{x}} \|x - \tilde{x}\|_2 \text{ s.t. } f(x) \neq f(\tilde{x})$$

Advantages:

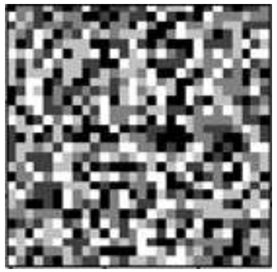
- Applicable to any gradient -based classifier
- Useful to evaluate the *robustness* of the model against adversaries
- Small perturbations often lead to imperceivable adversarial examples

Adversarial Examples: Cons

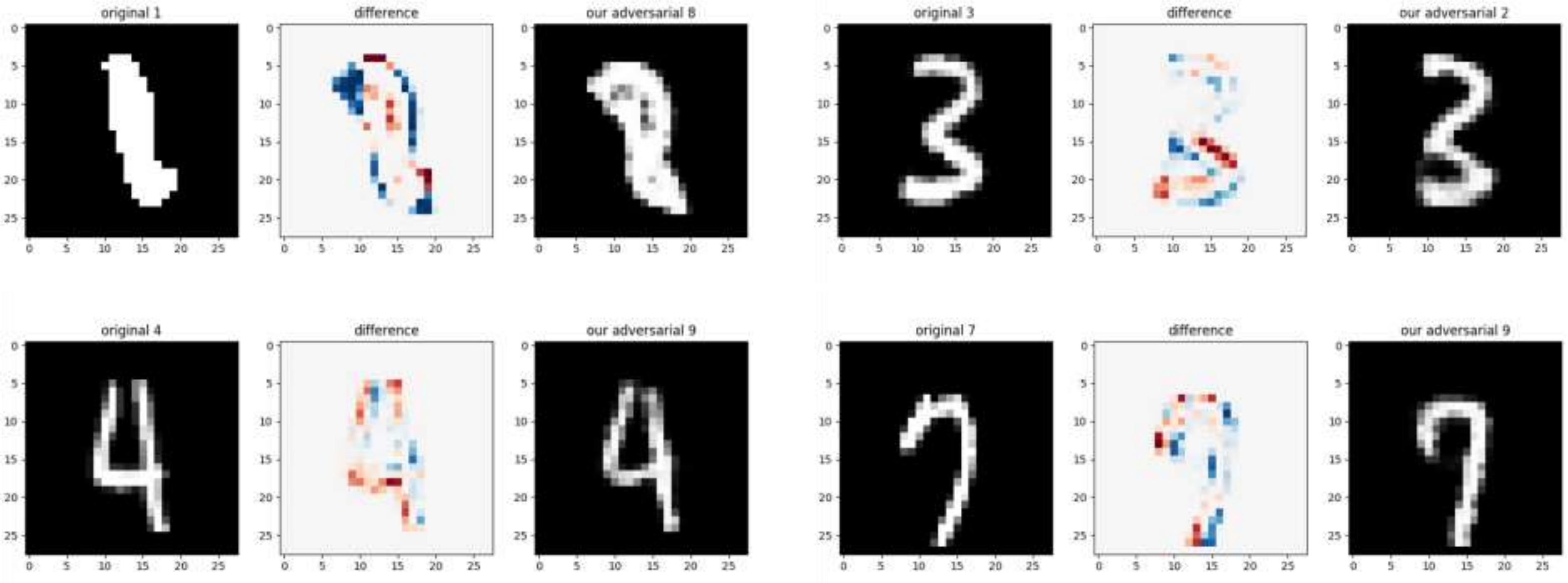
$$x^* = \text{argmin}_{\tilde{x}} \|x - \tilde{x}\|_2 \text{ s.t. } f(x) \neq f(\tilde{x})$$

Disadvantages:

- Examples are unnatural
  - may not look anything you would naturally see in the "wild"
- Distance is not always meaningful
  - E.g. color change or translation/rotation of an image
- Cannot be used for structured domains like text, code, etc.:
  - E.g. replacing/removing words results in sentences that are not grammatical
- Do not provide insights into why the sample is an adversary
  - How is the model working?
  - How to fix the model?



Example: MNIST Digits



Example: Church vs Tower

church→tower

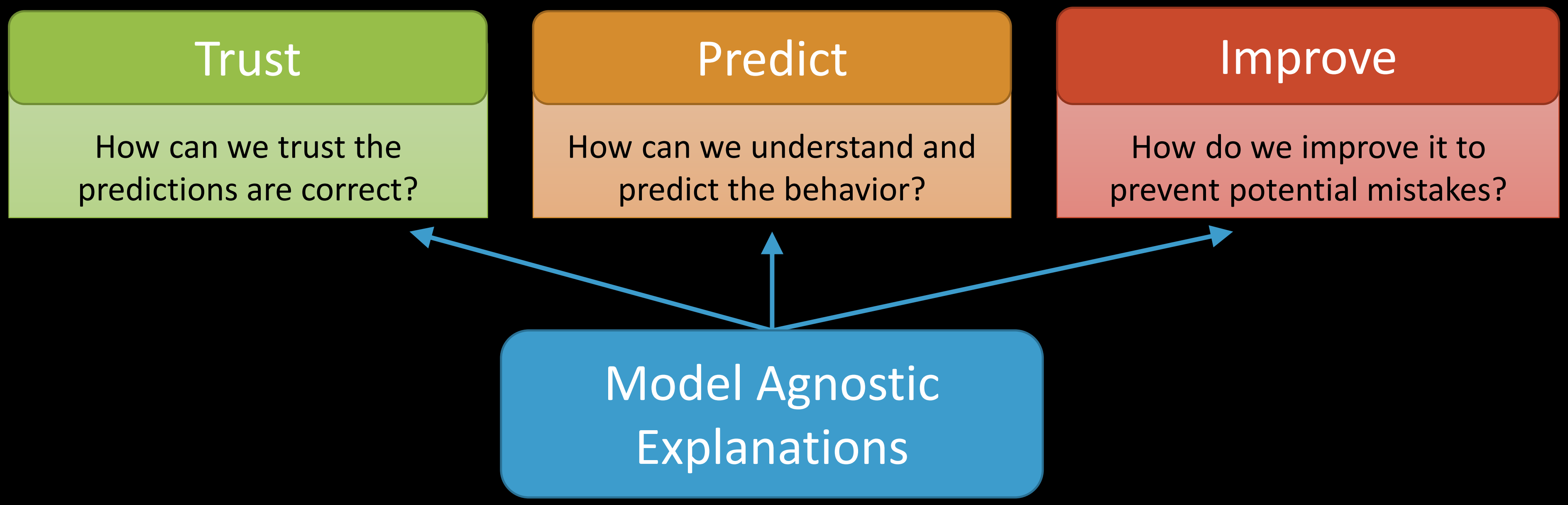
tower→church

Machine Translation

Debug Google Translate, remotely!

Source Sentence (English)	Generated Translation (German)
s : People sitting in a dim restaurant **eating** s′ : People sitting in a living room **eating** .	Leute, die in einem dim Restaurant **essen** sitzen. Leute, die in einem Wohnzimmeressen sitzen. *(People sitting in a living room)*
s : Elderly people **walking** down a city street . s′ : A man **walking** down a street playing	Ältere Menschen, die eine Stadtstraße **hinuntergehen**. Ein Mann, der eine Straße entlang spielt. *(A man playing along a street.)*

Explanations are important!



Model Agnostic Explanations

Work with Marco T. Ribeiro, Carlos Guestrin, Dheeru Dua, and Zhengli Zhao

Thanks!

sameer@uci.edu

sameersingh.org