"Why Should I Trust You?" Explaining the Predictions of Any Classifier

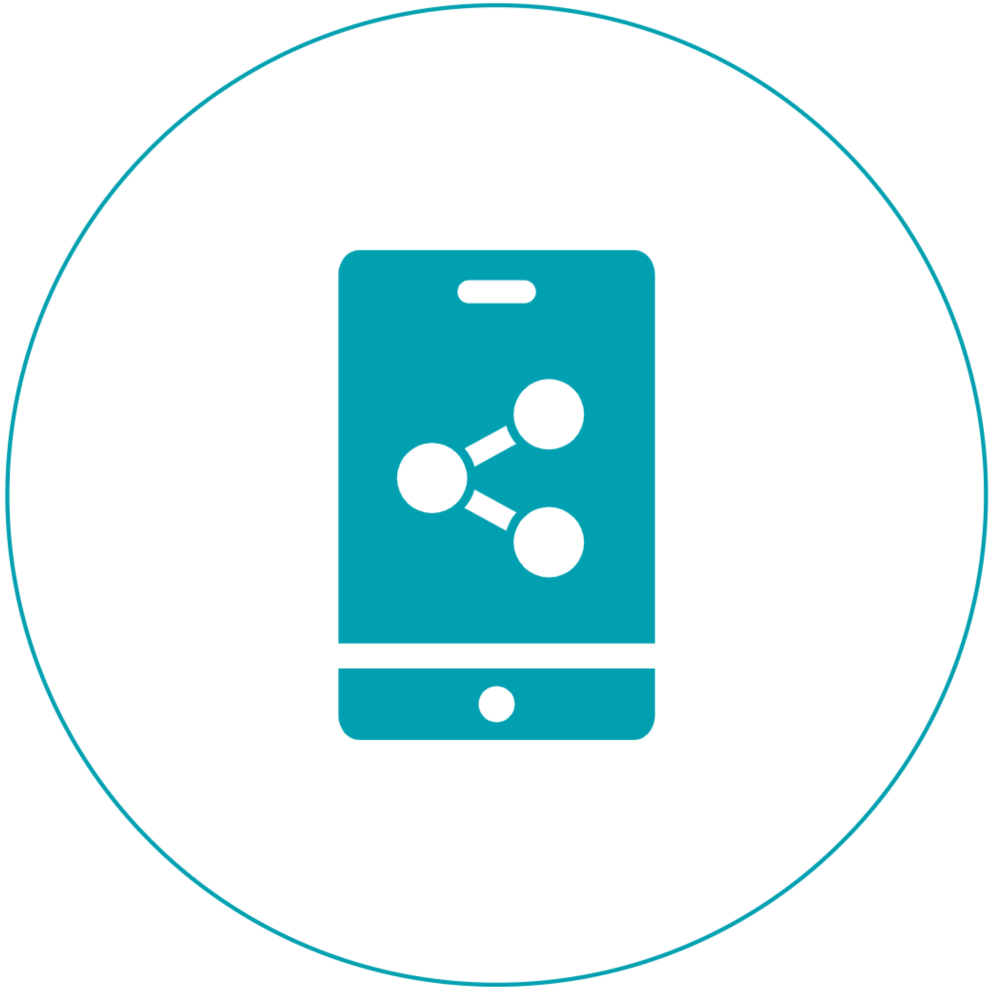
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Fengmin Wu

2019.10.16

Why is interpretability discussed today?

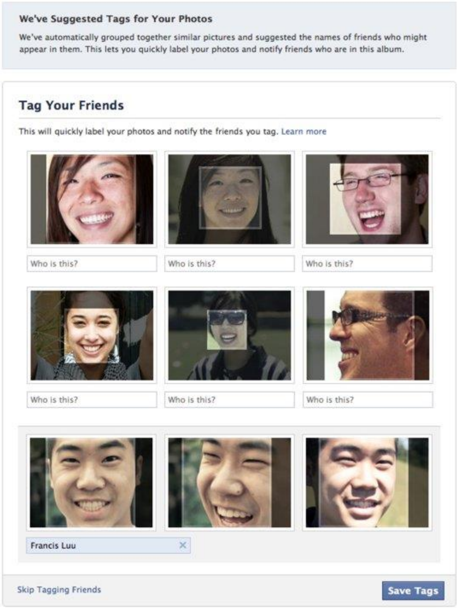
... since AI and machine learning is not really new ...

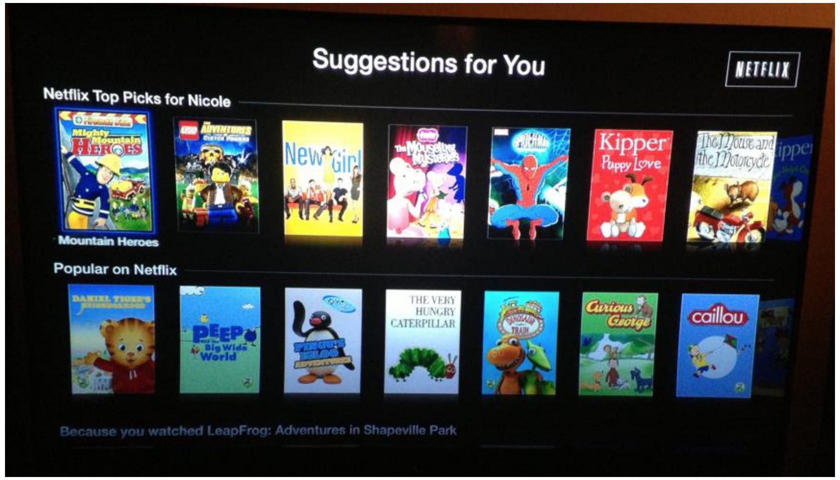


AI and machine learning algorithms become omnipresent.

- In (almost) every area of our lives
- Because algorithms tend to be more efficient and faster than humans

Machine Learning applications becoming pervasive…





Why is interpretability discussed today?

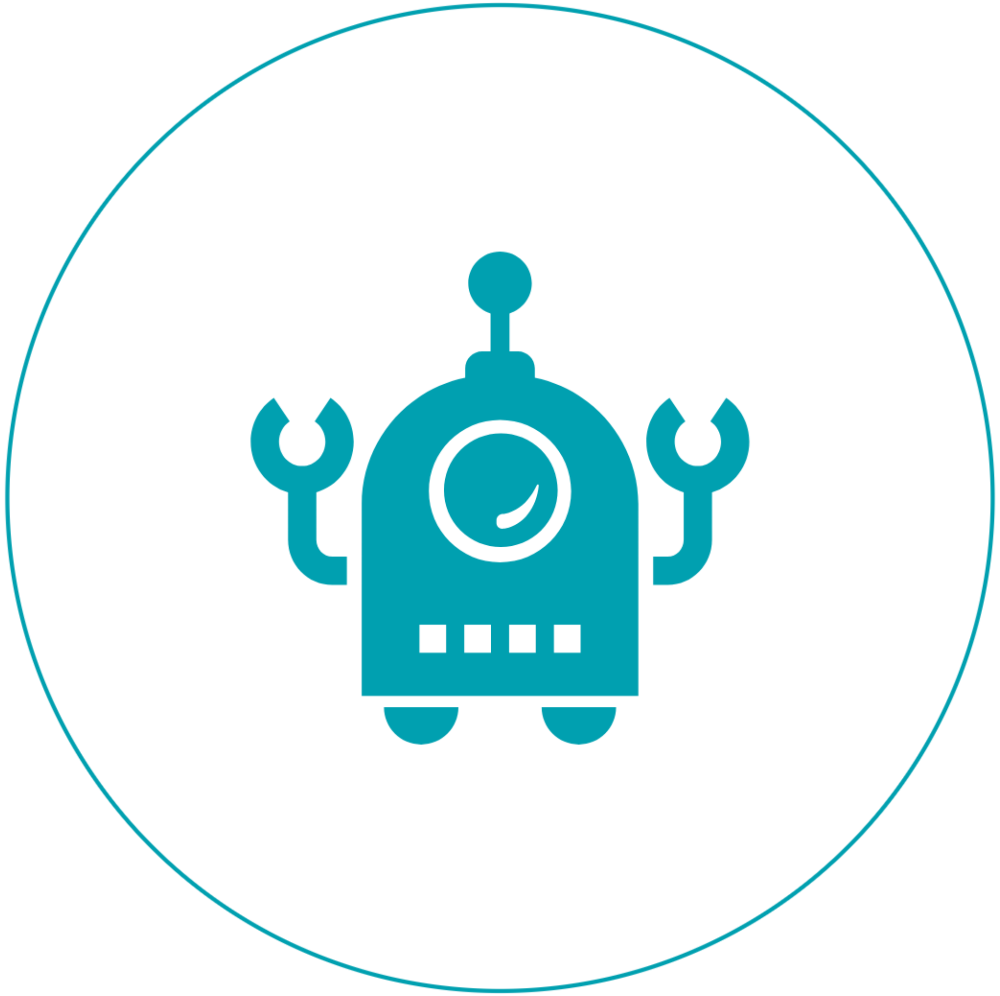
... since AI and machine learning is not really new ...

Data privacy and security become more important.

- Majority of our society is touched by it
- Talked about in the media
- Thus became an important political topic

Why is interpretability discussed today?

... since AI and machine learning is not really new ...

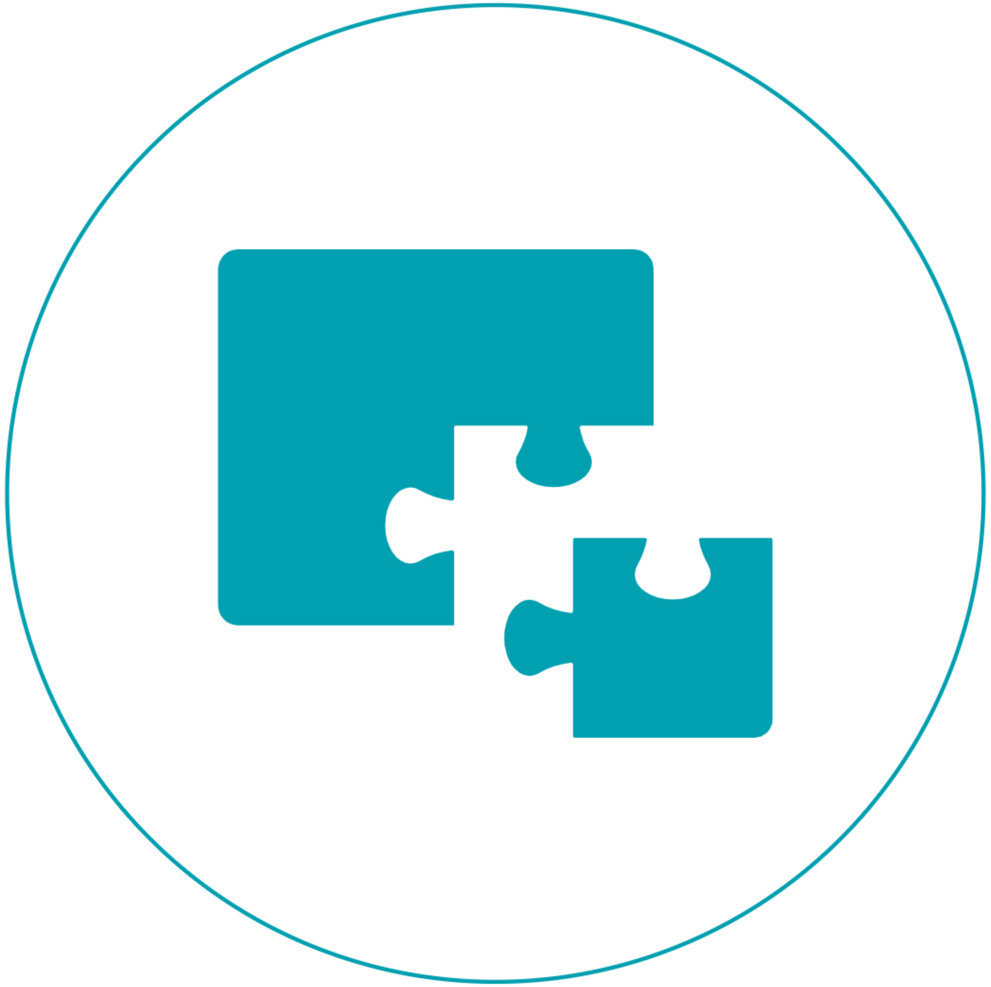


AI is hyped.

- The media reports about it a lot
- And tends to be critical about it

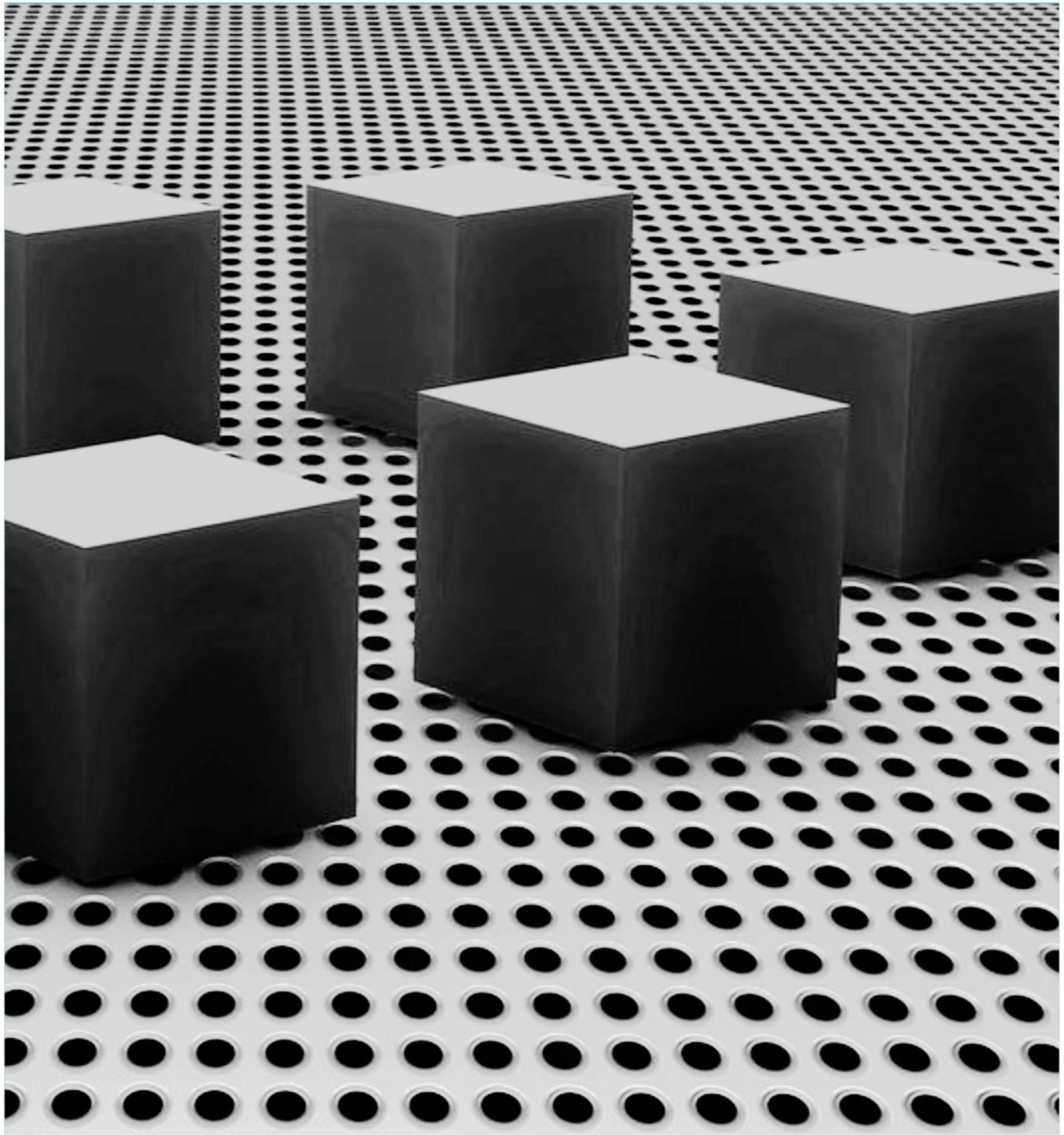
Why is interpretability discussed today?

... since AI and machine learning is not really new ...



AI is complex.

- Possibilities and dangers are hard to differentiate by most people
- “Brave New World”?

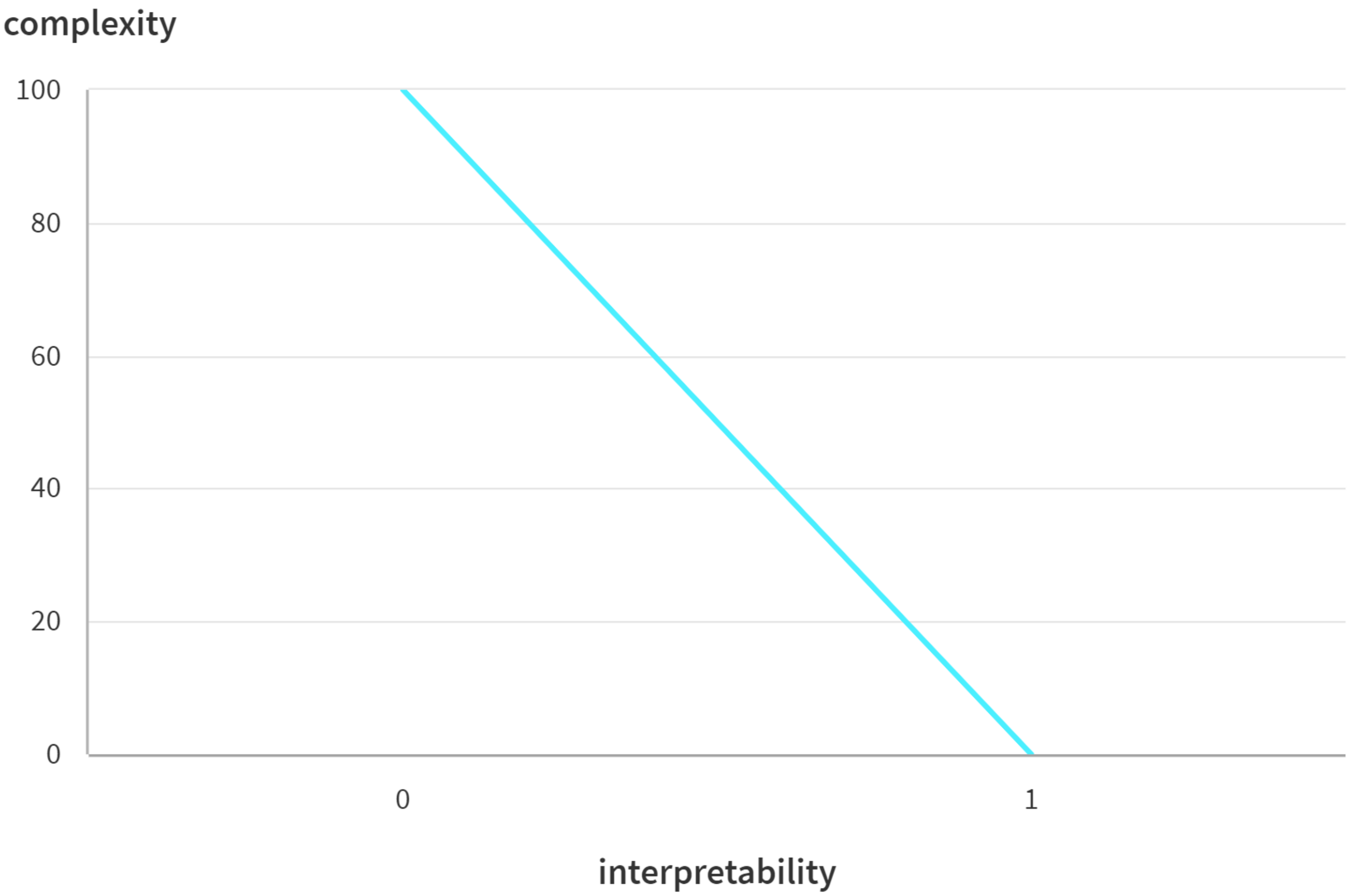


Why are ML models

“Black-Boxes”?

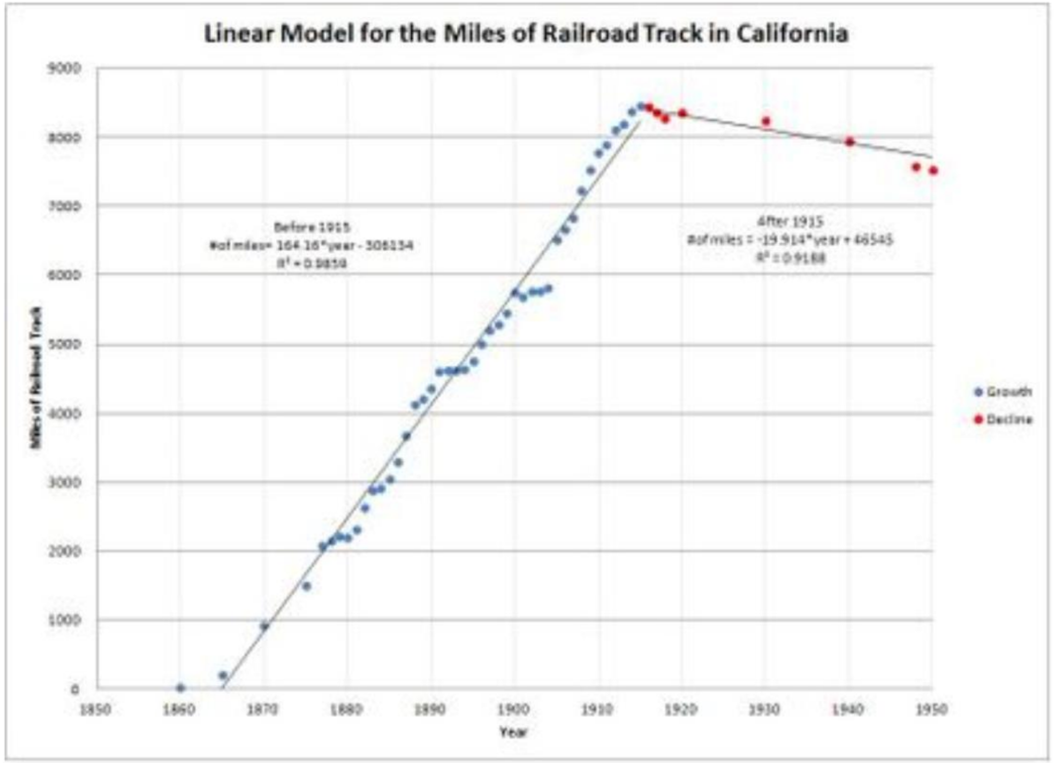
Trade-off between interpretability & complexity

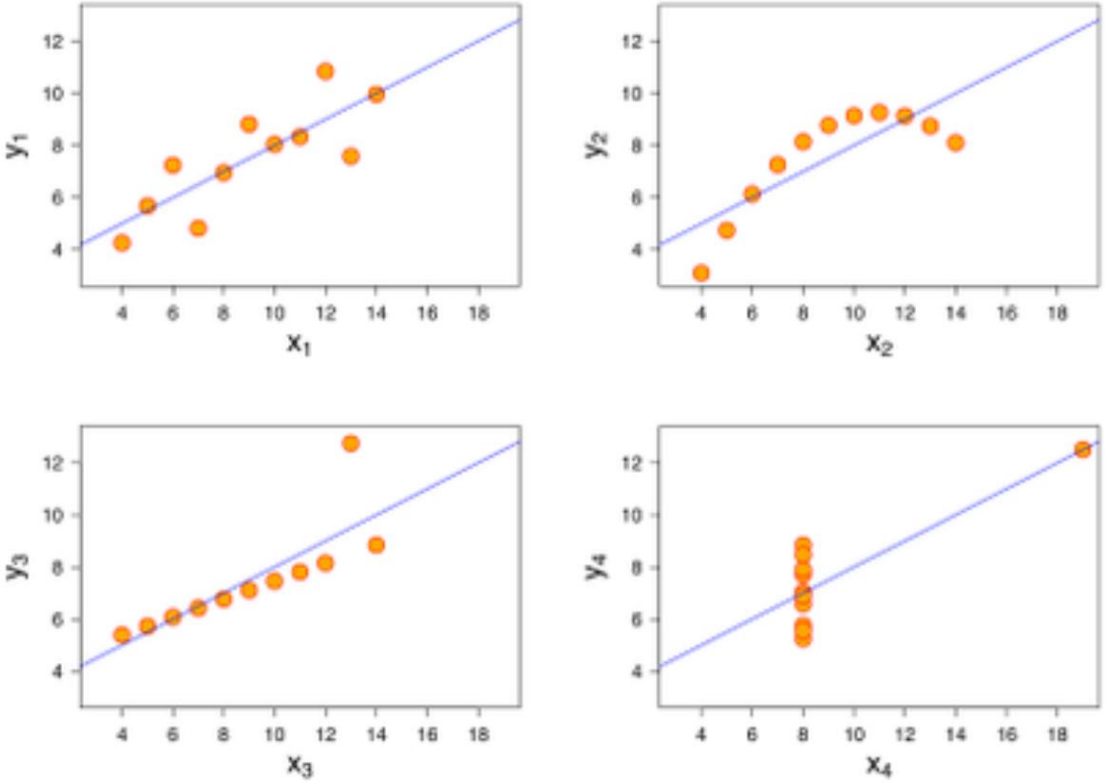
The more complex a model, the harder it tends to be to understand.



Trade-off between interpretability & complexity

Linear regression models are easy to understand: If ..., then ...

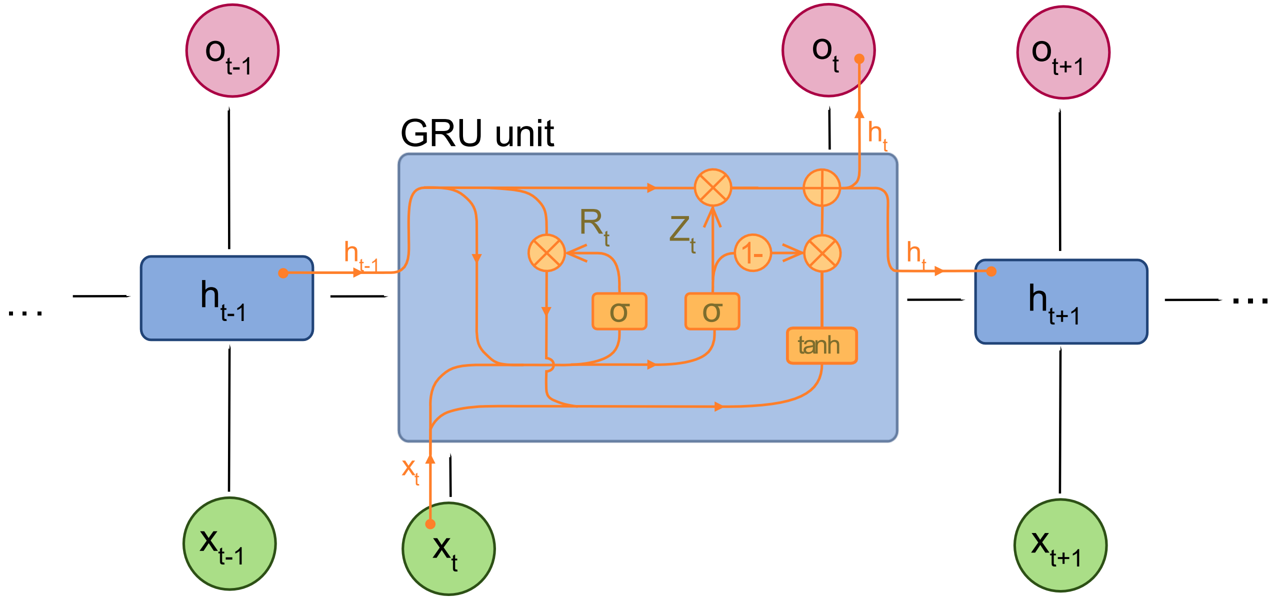


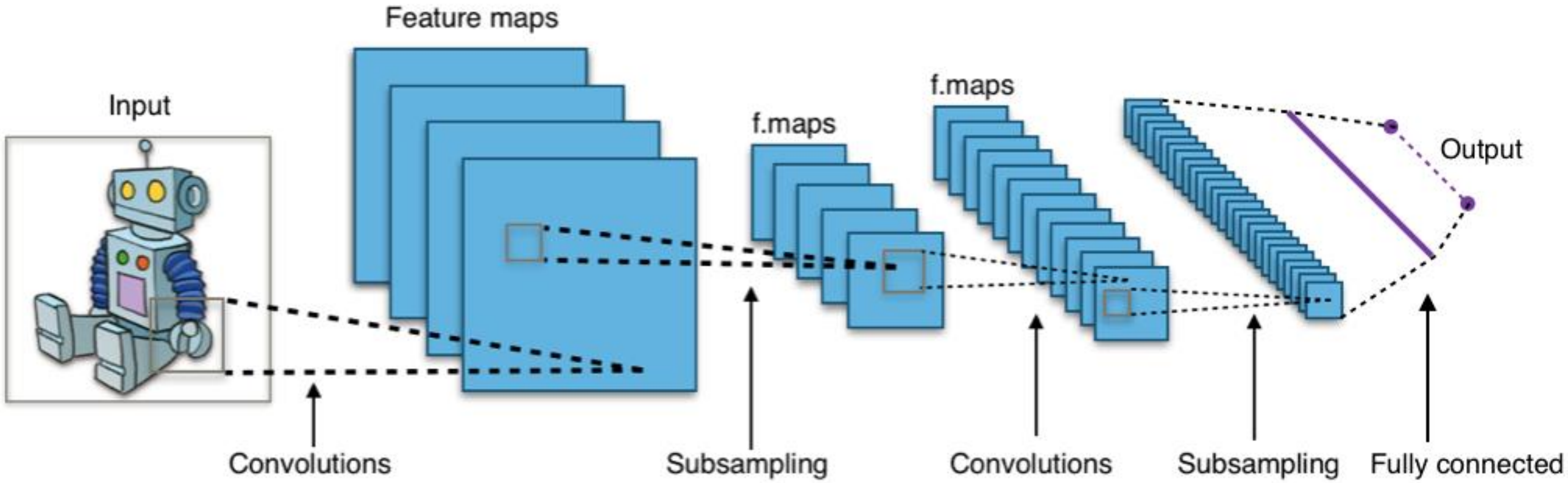


Source: Wikipedia

Trade-off between interpretability & complexity

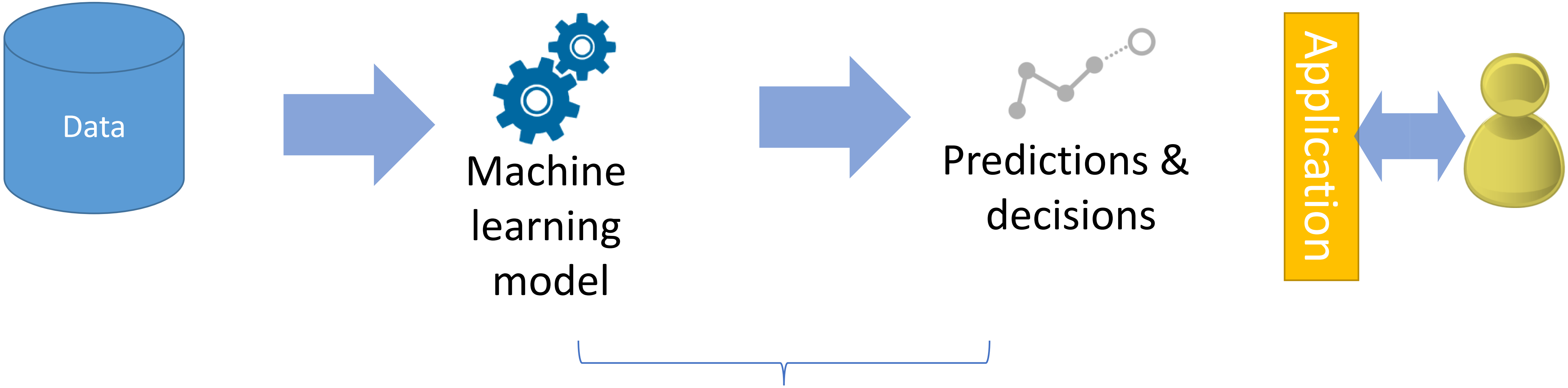
Complexity tends to make models perform better but harder to understand.





Can we **trust** our models?
Are they fit for the wild?

How to build an application with ML



Trust Challenge

- Is model really working?
- Convince myself and others?

Why should we improve our understanding of ML models?

... if technically it isn't necessary ...

Improving our models

Generalisability

“Sanity Check”

Prevent wrong conclusions & potentially adversarial attacks

Improving our models

When we understand our models better, we are better at detecting wrong conclusions.

Example 1: Image classification of wolves & Huskies

The model based its predictions on the snow in the background.

Example 2: Text classification of Christian & Atheist posts

Not all words that predictions were based on made sense.

Example 3: Image classification with Google's Inception Net

Knowing which areas of an image contributed to a decision helps us trust it.

Why should we improve our understanding of ML models?

... if technically it isn't necessary ...

Improving our models

Generalisability

“Sanity Check”

Prevent wrong conclusions & potentially adversarial attacks

Trust and transparency

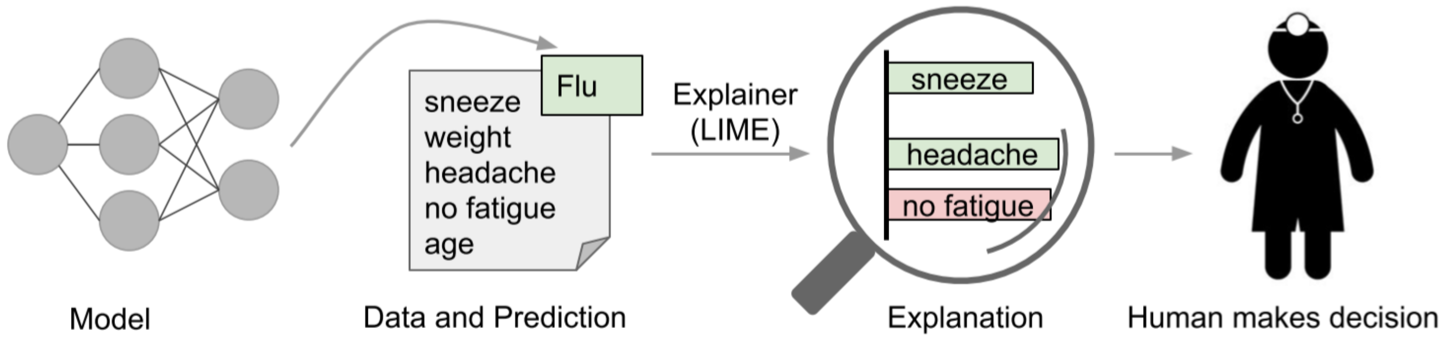
Can I trust my model's decisions?

Why does my model make the predictions it makes?

Trust and transparency

But trust is also important in others areas!

- Decision can be a question of life and death.
- Medical intervention needs to be based on a diagnosis.



Trust and transparency

But trust is also important in others areas!

- Decision can be a question of life and death.
- Medical intervention needs to be based on a diagnosis.

- In business, we can save time and money by improving our understanding of machine learning.

Why should we improve our understanding of ML models?

... if technically it isn't necessary ...

Improving our models

Generalisability

“Sanity Check”

Prevent wrong conclusions &

potentially adversarial attacks

Trust and transparency

Can I trust my model's decisions?

Why does my model make the predictions it makes?

Prevent Bias

Fairness

Identify and prevent bias

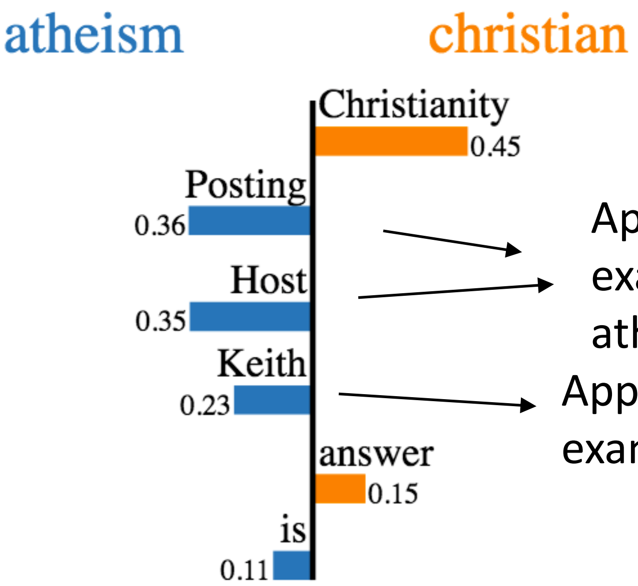
Explaining individual predictions:

Making any model interpretable

What an explanation looks like

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

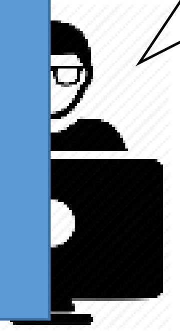
I think Christianity is the one true religion.
If you'd like to know more, send me a note



Appear in 21% of training examples, almost always in atheism

Appears in 11% of training examples, **always** in atheism

Why did this happen? How do I fix it?

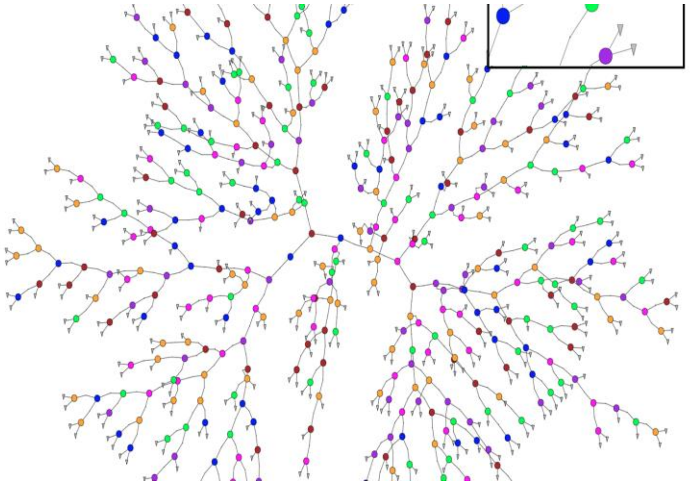


→ Will not generalize
→ Don't trust this model!

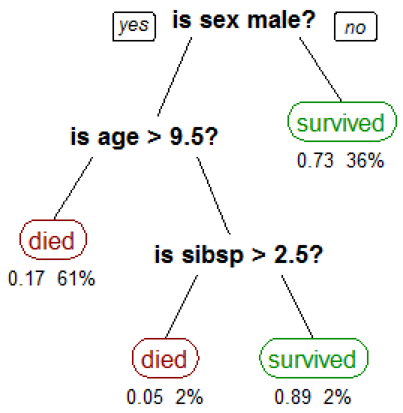
Three must-haves for a good explanation

Interpretable

- Humans can easily interpret reasoning



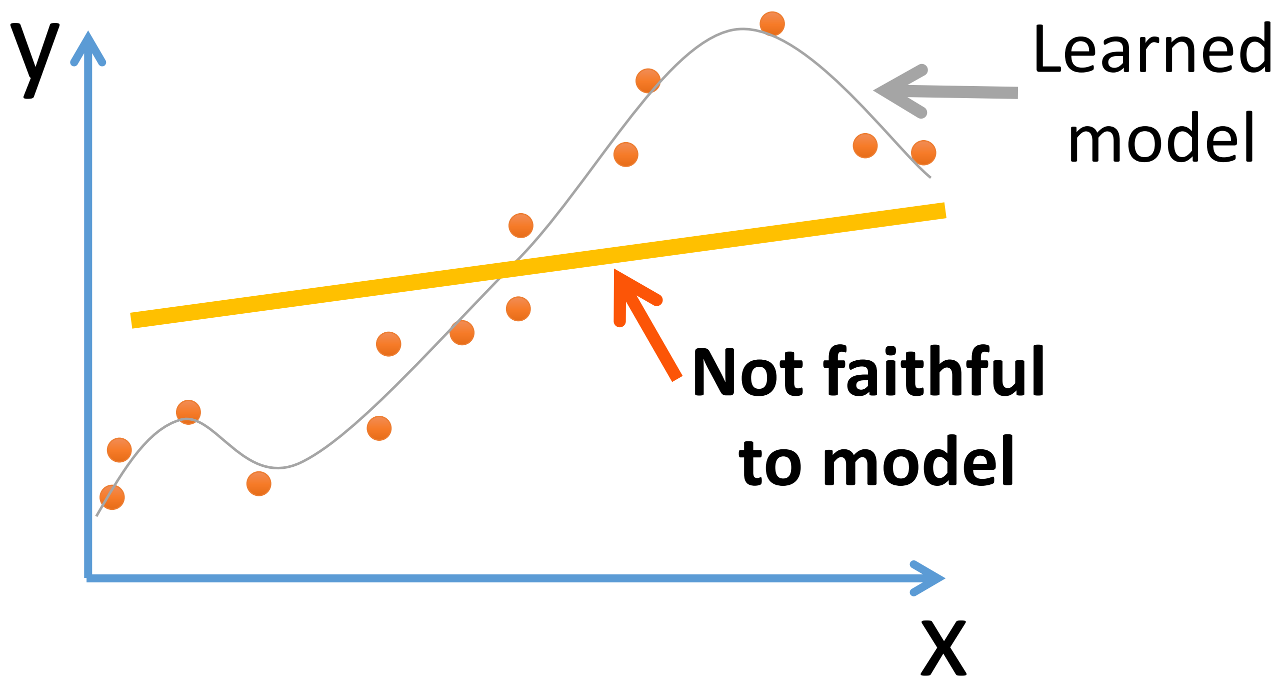
Definitely
not interpretable



Potentially
interpretable

Three must-haves for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves



Three must-haves for a good explanation

Interpretable	• Humans can easily interpret reasoning
Faithful	• Describes how this model actually behaves
Model agnostic	• Can be used for *any* ML model

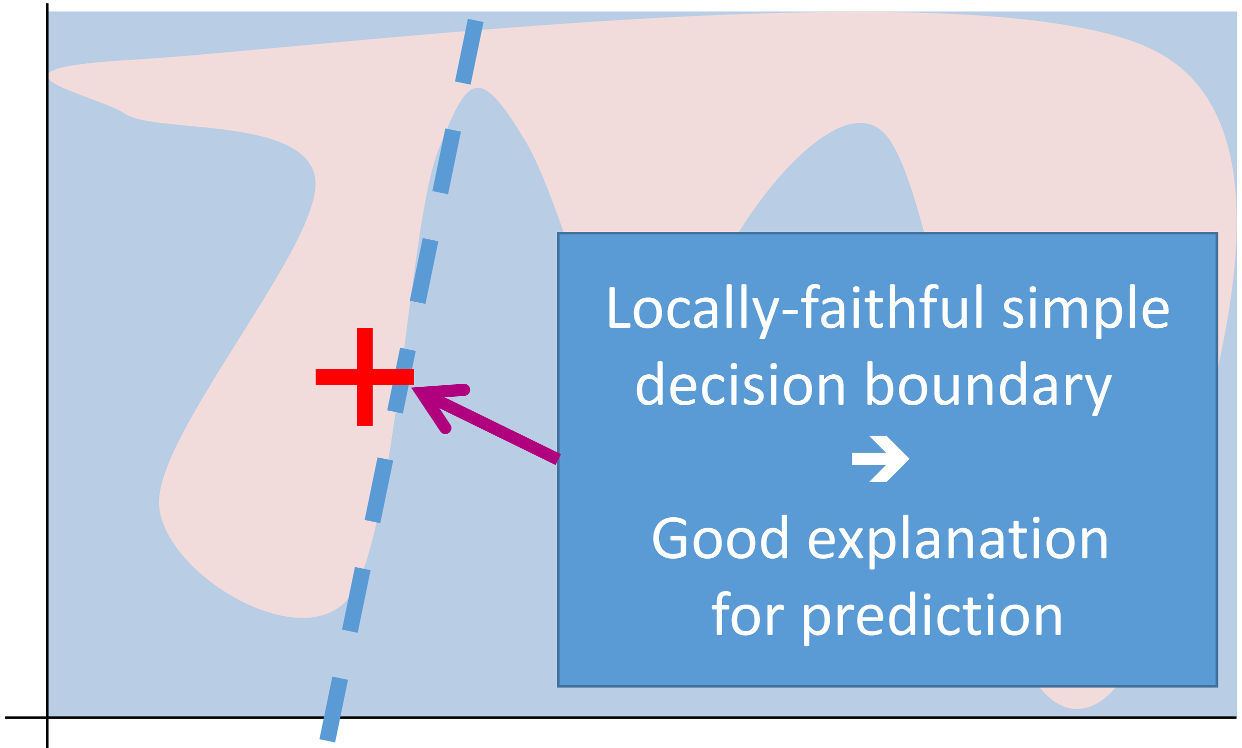
LIME: *Local Interpretable Model-Agnostic Explanations*

LIME – Key Ideas

1. Pick a model class interpretable by humans
   - Not globally faithful… ☹

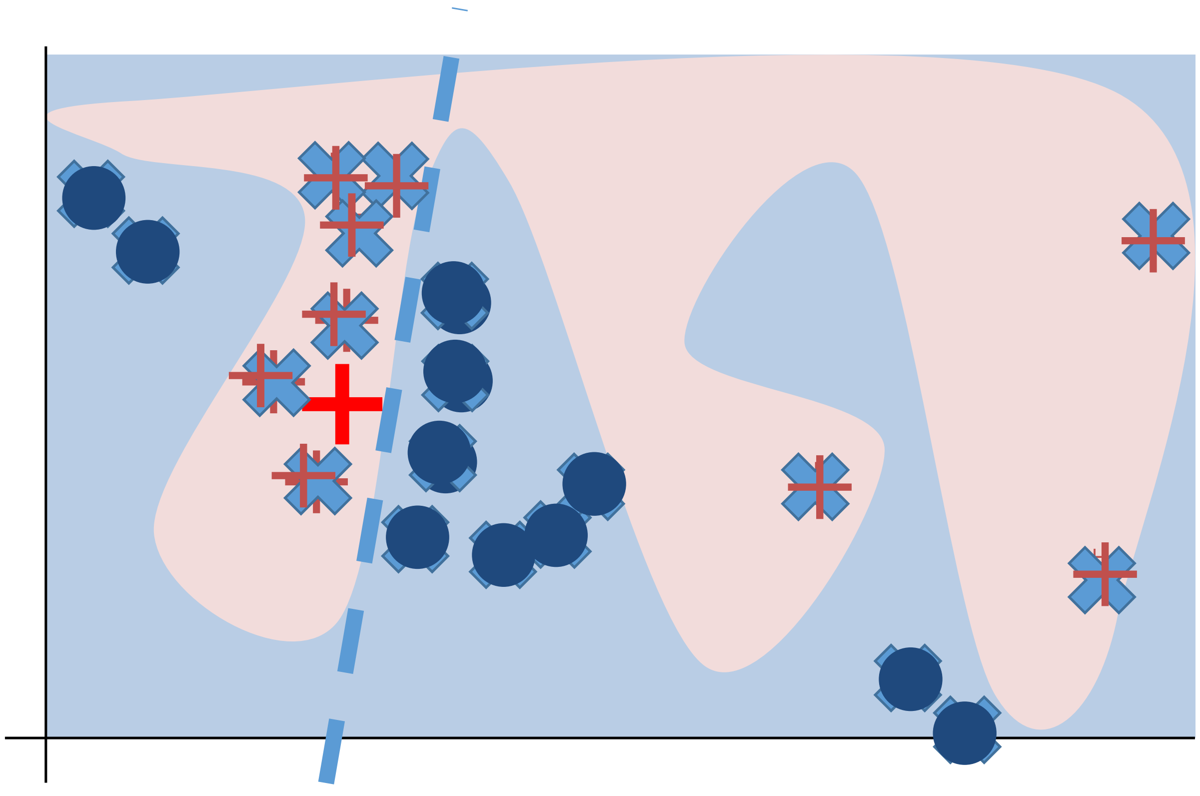
Line,
shallow decision tree,
sparse features, …

2. Locally approximate global (blackbox) model
   - Simple model globally bad, but locally good



Using LIME to explain a complex model's prediction for input x_i

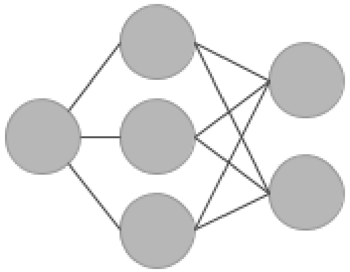
1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples
5. Use simple model to explain



Interpretable representations

x (embeddings)

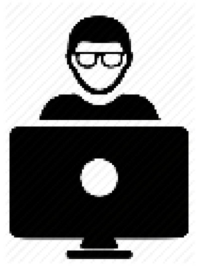
0.5	0.3	1.3	4.4	1.1	…



Model

x' (words)

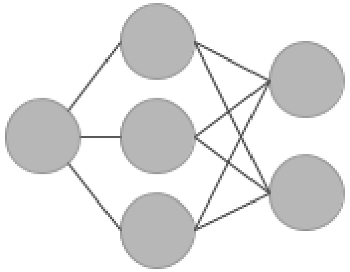
This is a horrible movie.



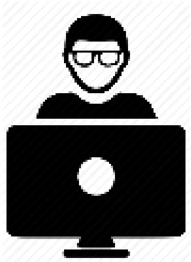
This is what we perturb, and this is what we use in the interpretable approximation

Interpretable representation: images

x (3 color channels / pixel)



x' (contiguous superpixels)



LIME

Explain model : $\boldsymbol{g} \in \boldsymbol{G}$
Model being explained *:* $\boldsymbol{f}$*:* $\mathbb{R}^{\boldsymbol{d}} \rightarrow \mathbb{R}$
A measure of complexity *:* $\boldsymbol{\Omega}(\boldsymbol{g})$ - depth of the tree, number of non-zero weights
Capturing the locality *:* $\boldsymbol{\pi}_{\boldsymbol{x}}$ - far away from x,low weight, or vice versa

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

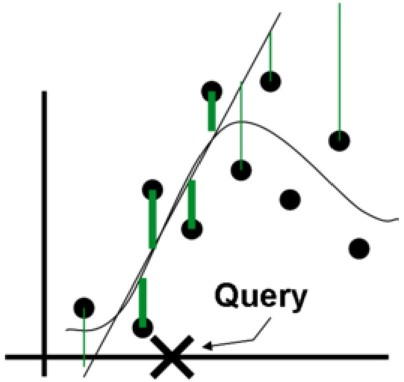
- A measure of how unfaithful $\boldsymbol{g}$ is in approximating $\boldsymbol{f}$ in the locality defined by $\boldsymbol{\pi}_{\boldsymbol{x}}$
- Get an explanation model $\boldsymbol{\xi}(\boldsymbol{x})$ by optimizing it.

Sampling example - images

Original Image

P(labrador) = 0.21

Perturbed Instances	P(Labrador)
	0.92
	0.001
	0.34



Locally weighted regression

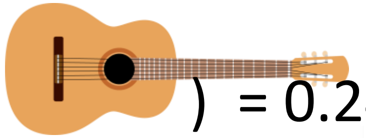
$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(z)\big(f(z) - g(z')\big)^2$$

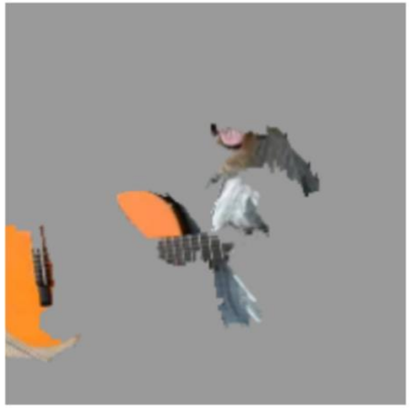
Explanation

Gaining insights from explanations

Explaining Google’s Inception NN

P() = 0.32

P() = 0.24



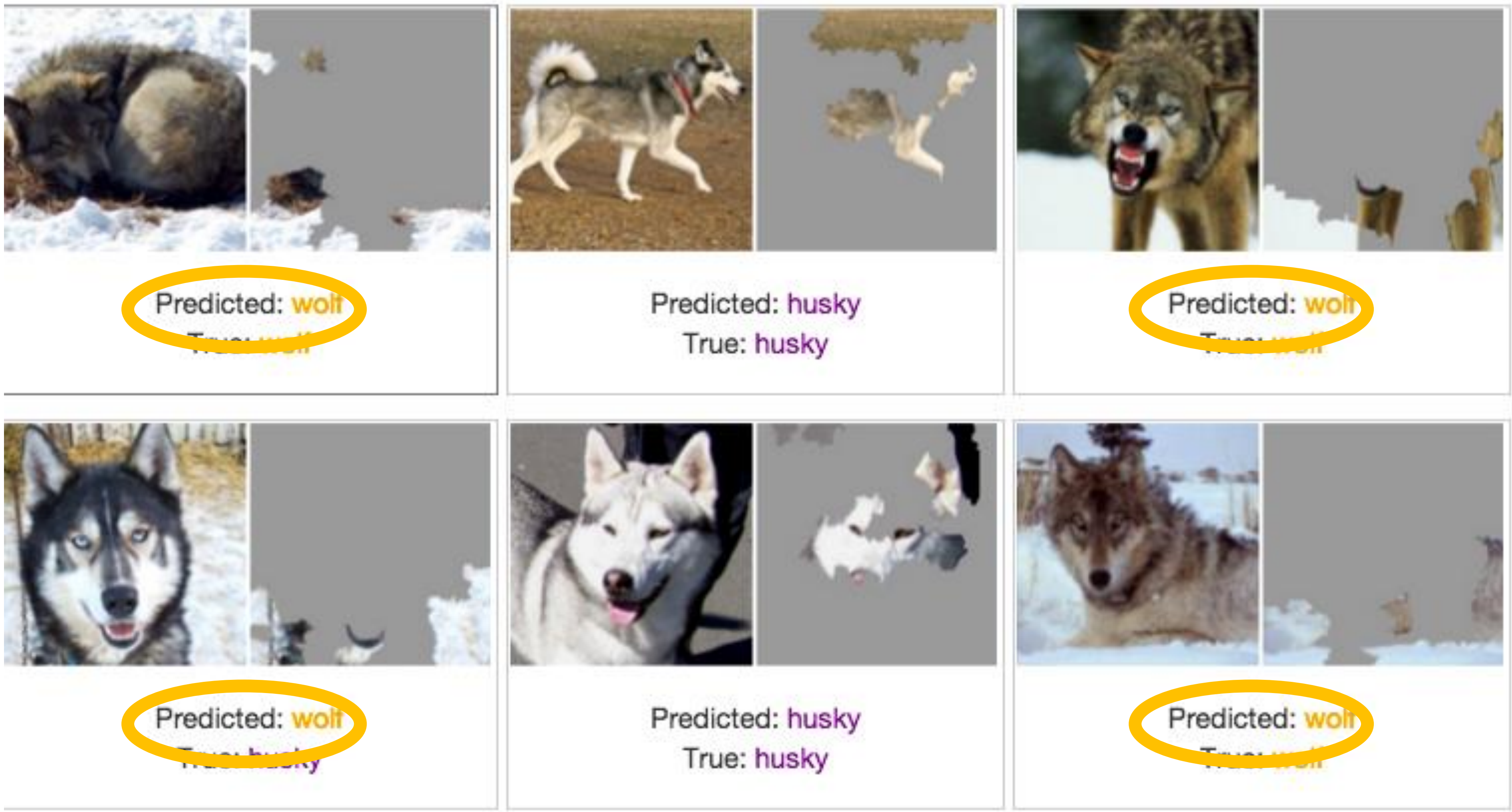
P() = 0.21

Train a neural network to predict wolf v. husky

Only 1 mistake!!!

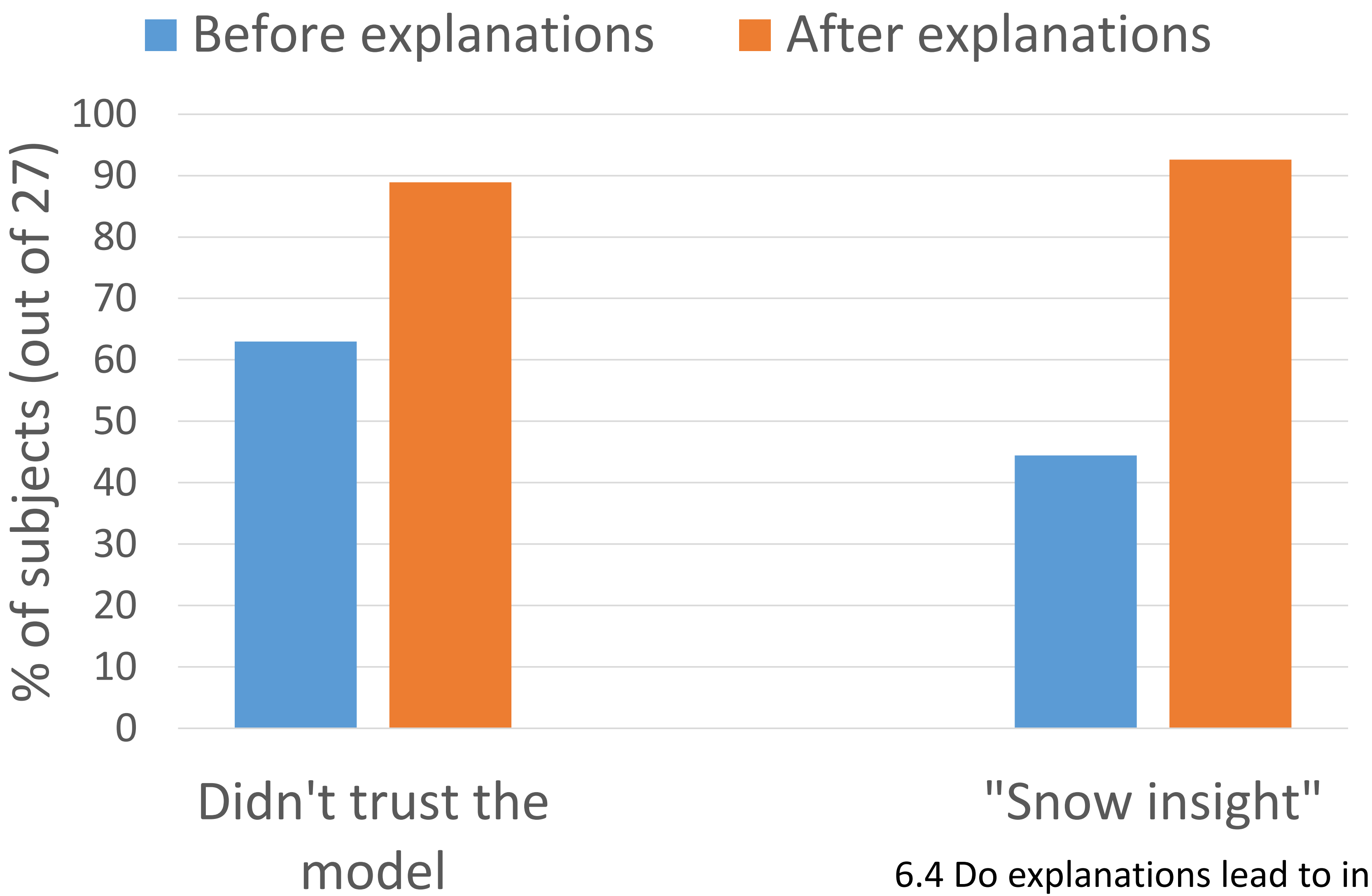
Do you trust this model?
How does it distinguish between huskies and wolves?

Explanations for neural network prediction



We’ve built a great snow detector… ☹

Did machine learning people notice it?



6.4 Do explanations lead to insights?

Beyond explaining predictions:

Explaining whole models

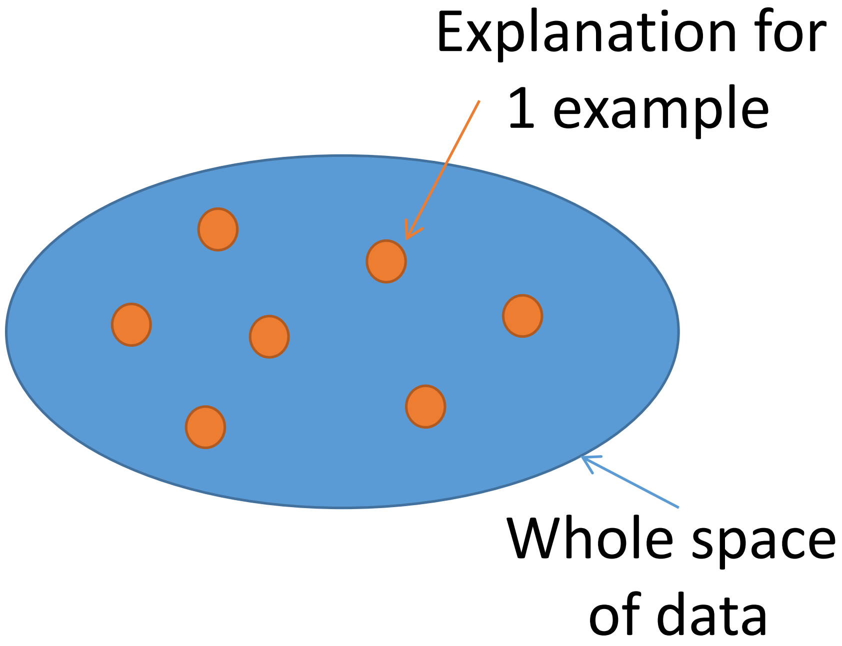
Explaining model behavior for whole space

Explaining 1 prediction describes local behavior of model

Pick k predictions to describe overall behavior

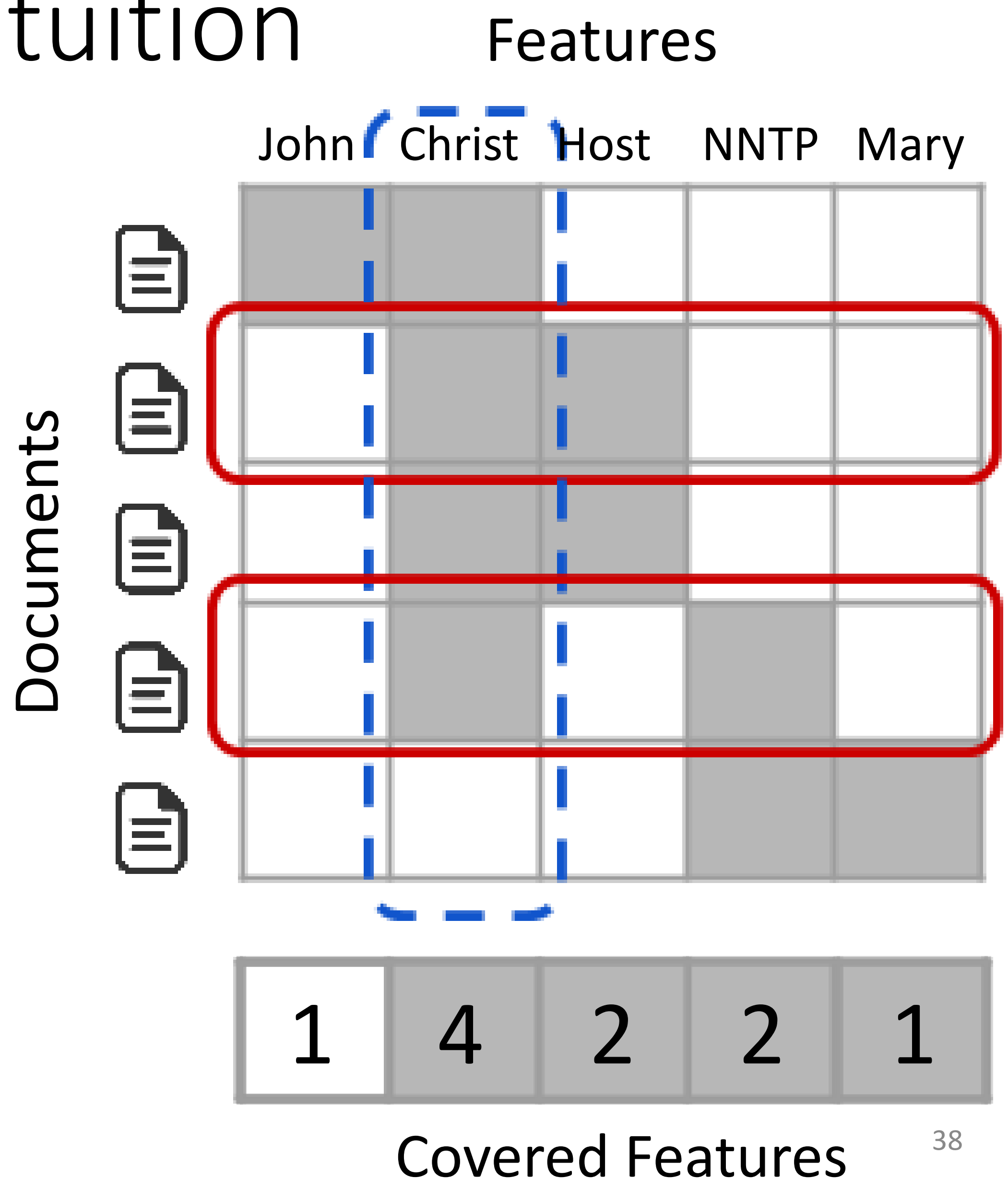
Indispensable to avoid redundancy in explanations

Submodular function optimization provides diverse set of explanations



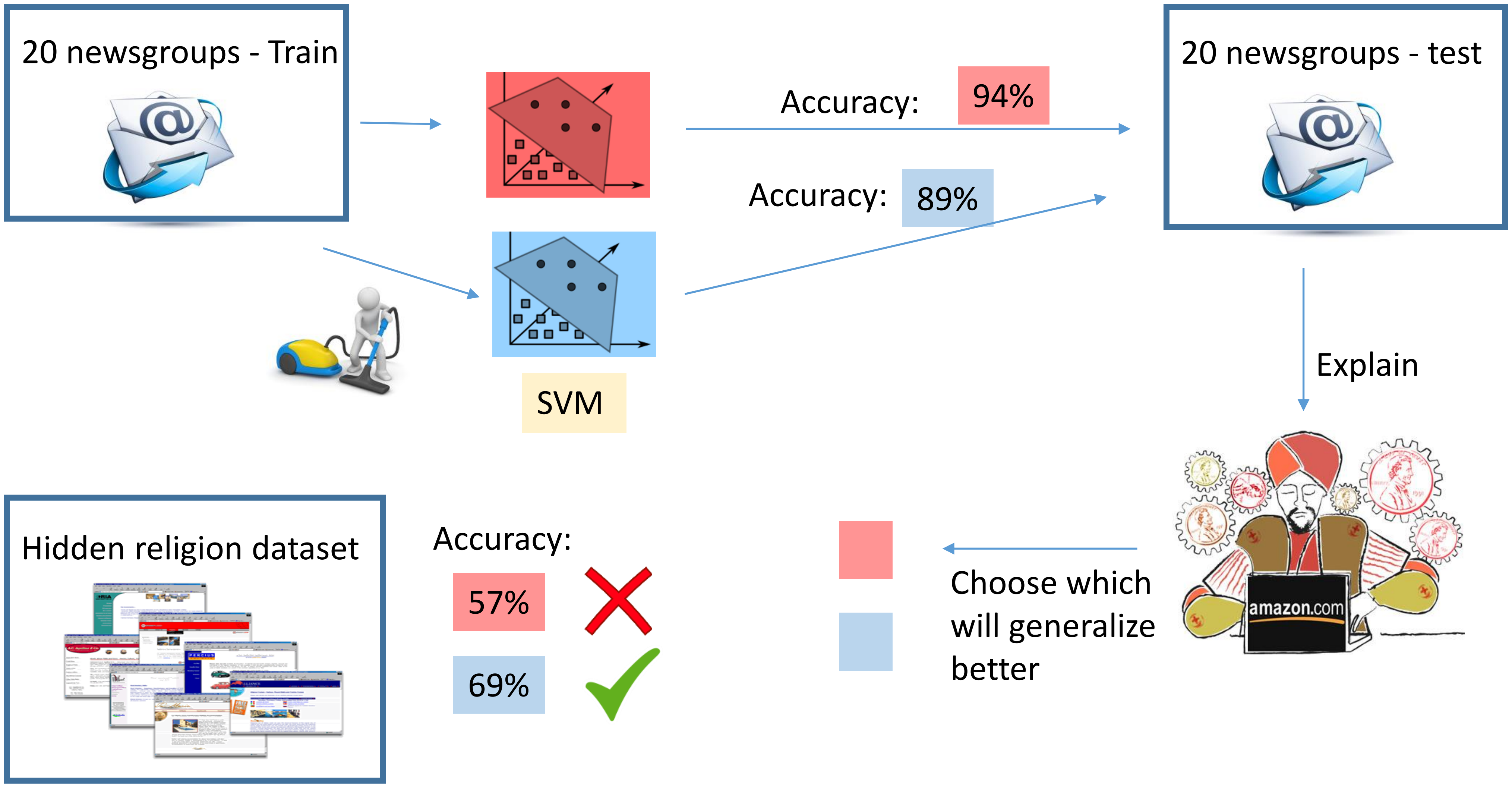
Submodular selection- intuition

1. Representative explanations
2. Avoid redundancy



Evaluating whole-model explanations

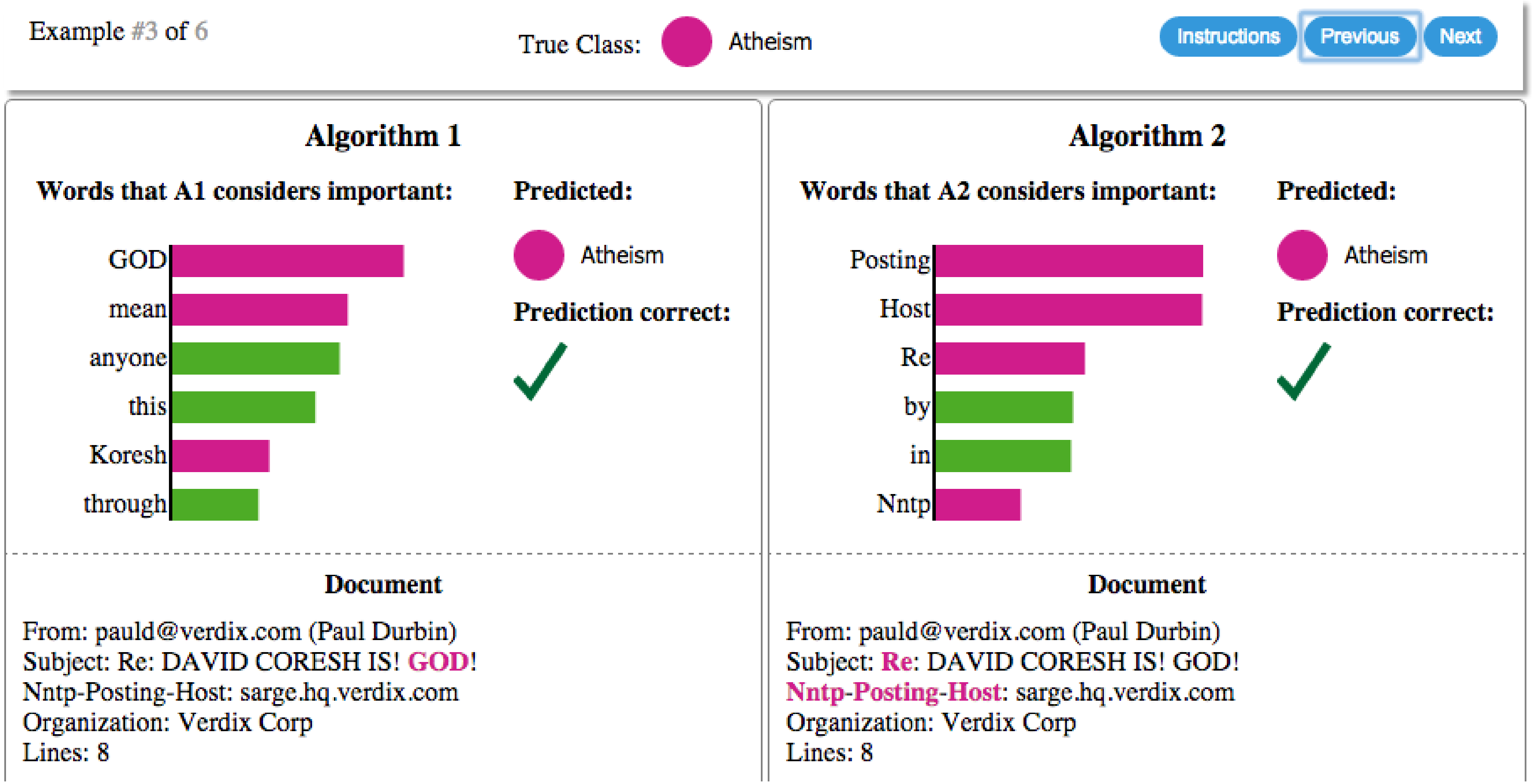
Choosing between competing models



6.2 Can users select the best classifier?

Choosing between competing models

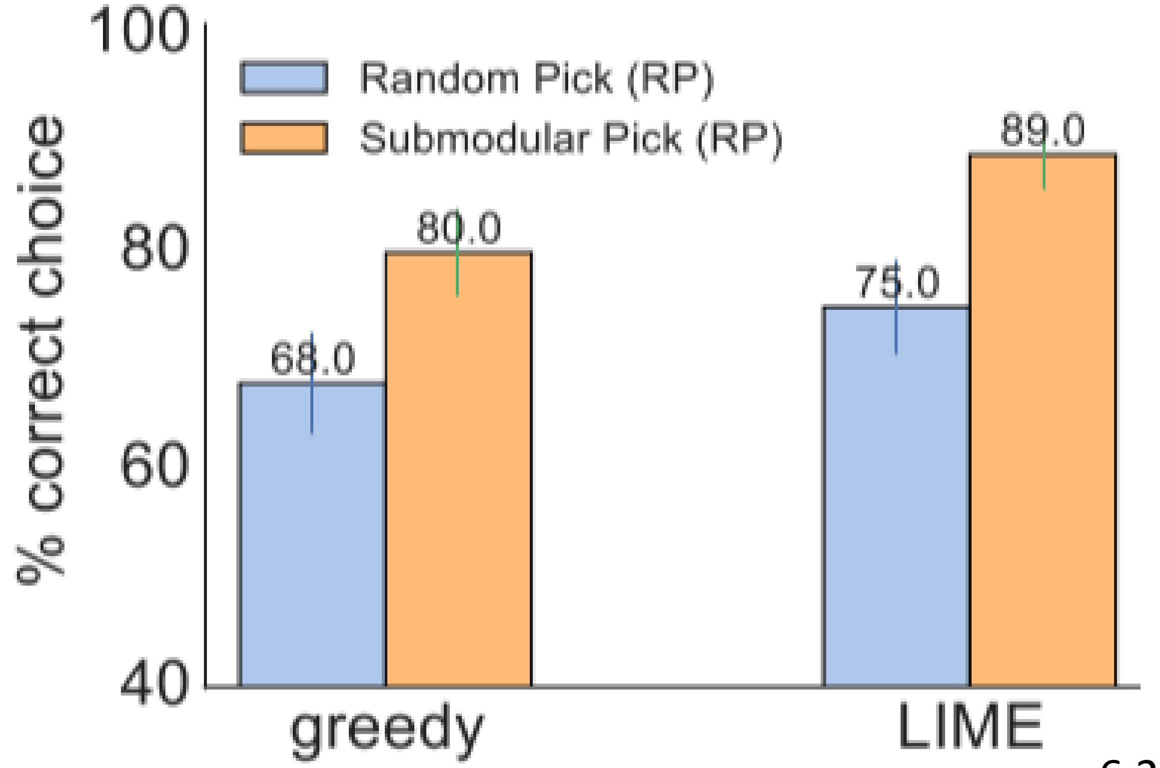
- Ask people on Mechanical Turk which model generalizes better



6.2 Can users select the best classifier?

Choosing between competing models

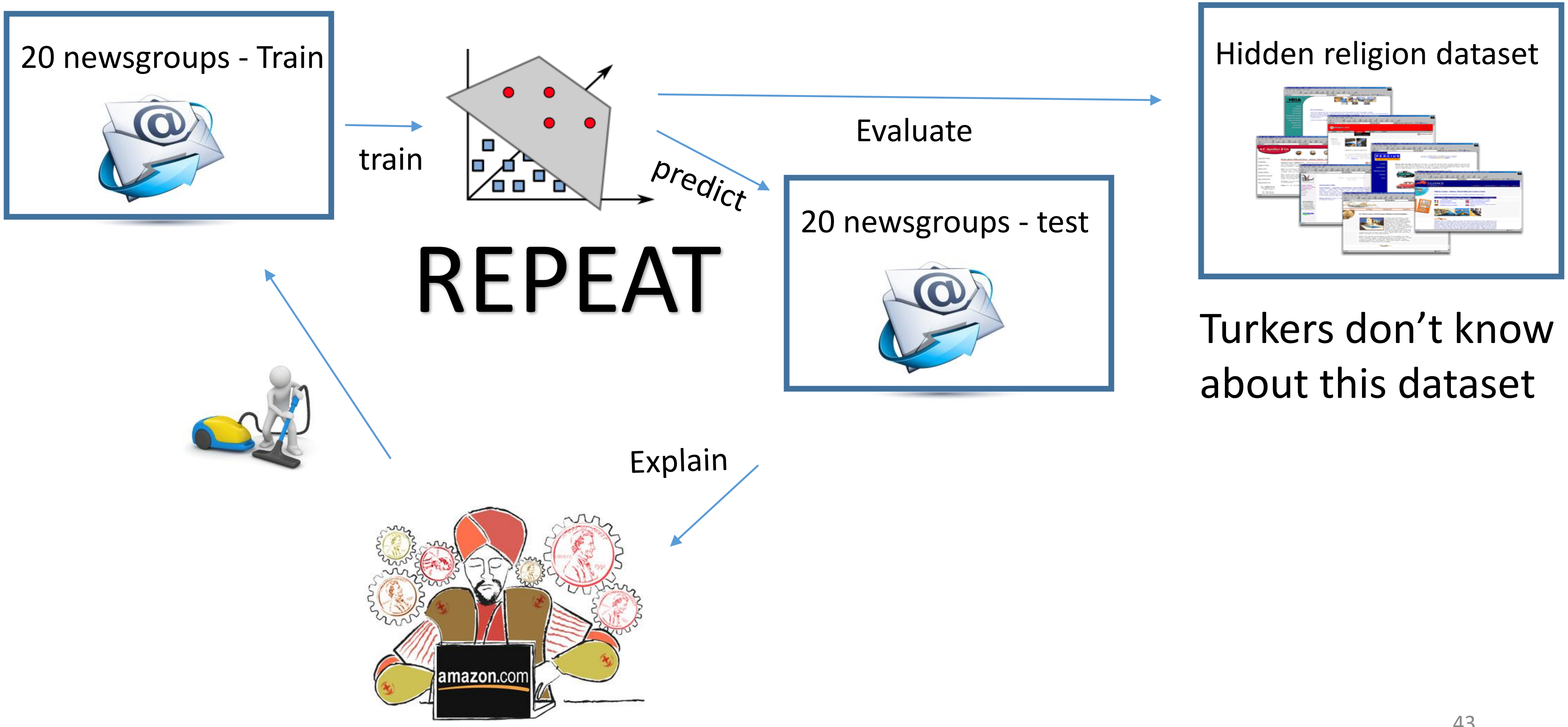
89% of Mechanical Turkers identify more trustworthy model



If we picked based on accuracy, we would get it wrong.

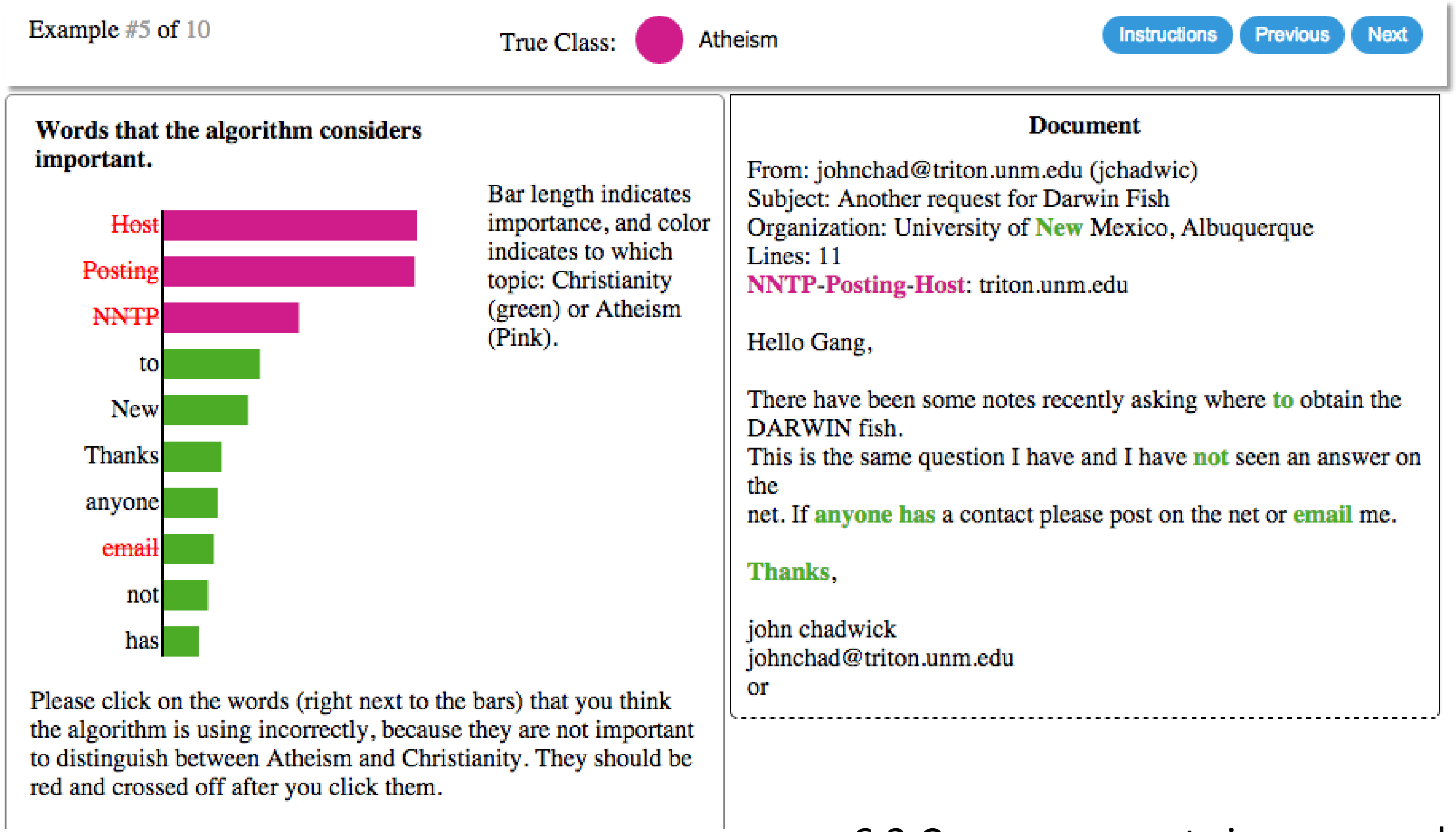
6.2 Can users select the best classifier?

Fixing bad classifiers



Fixing bad classifiers

- Turkers click on 'useless' words for the task in each round

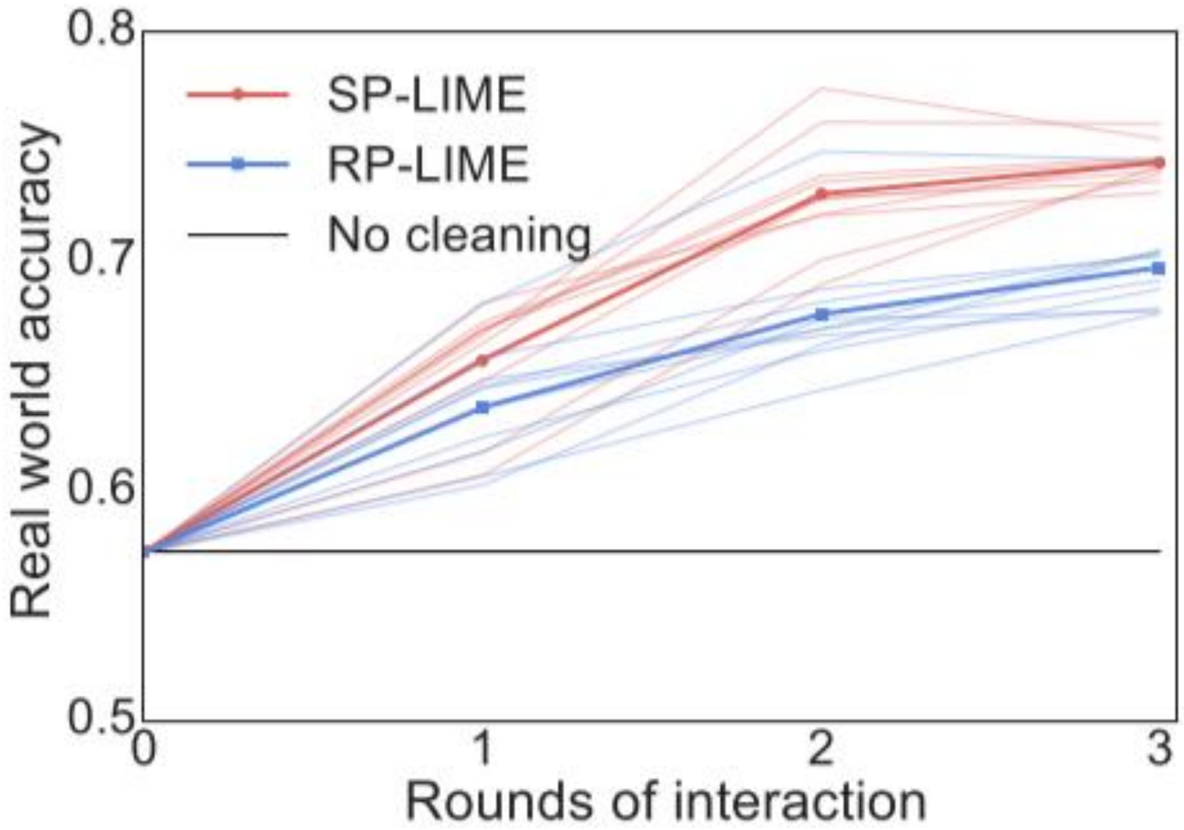


6.3 Can non-experts improve a classifier?

Fixing bad classifiers

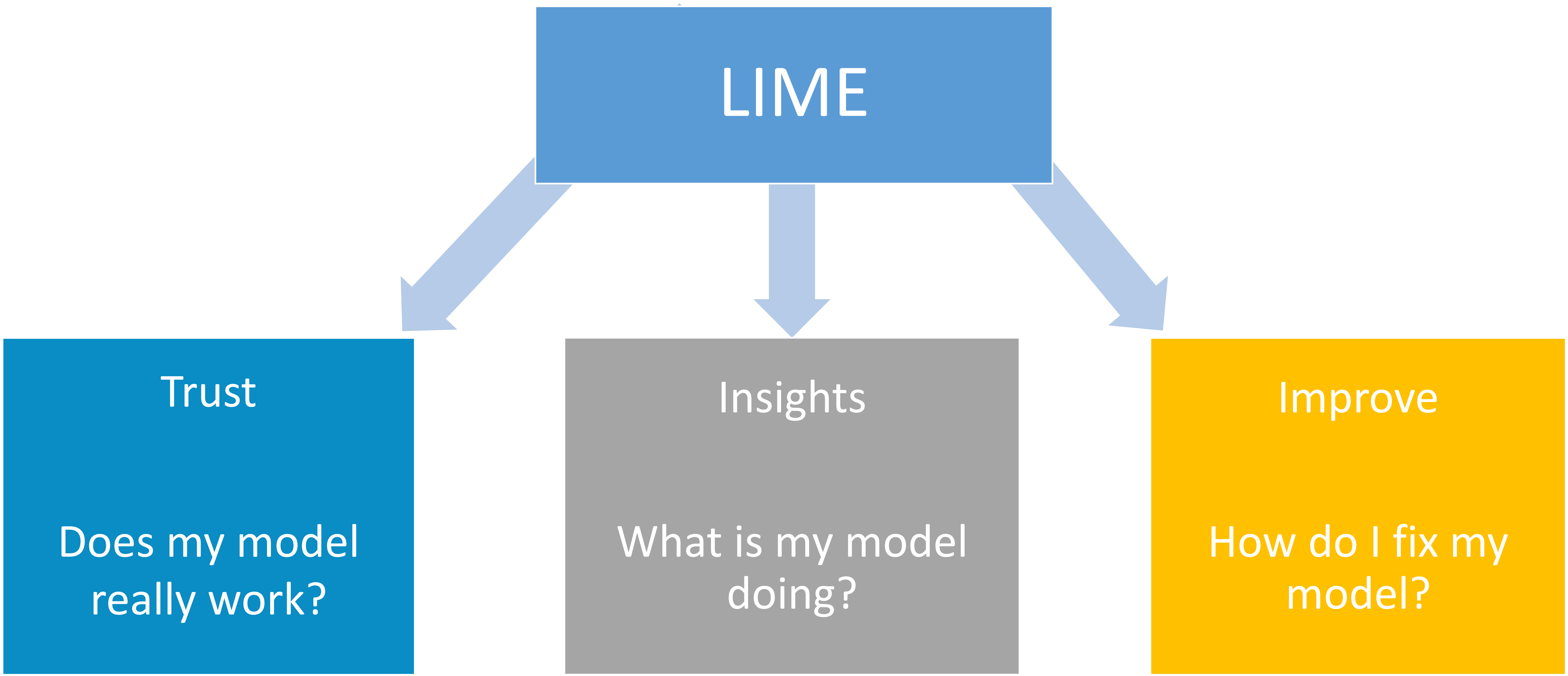
Mechanical Turkers can do 'feature engineering' really well!

Accuracy on hidden set



6.3 Can non-experts improve a classifier?

Conclusion



Open source project: https://github.com/marcotcr/lime