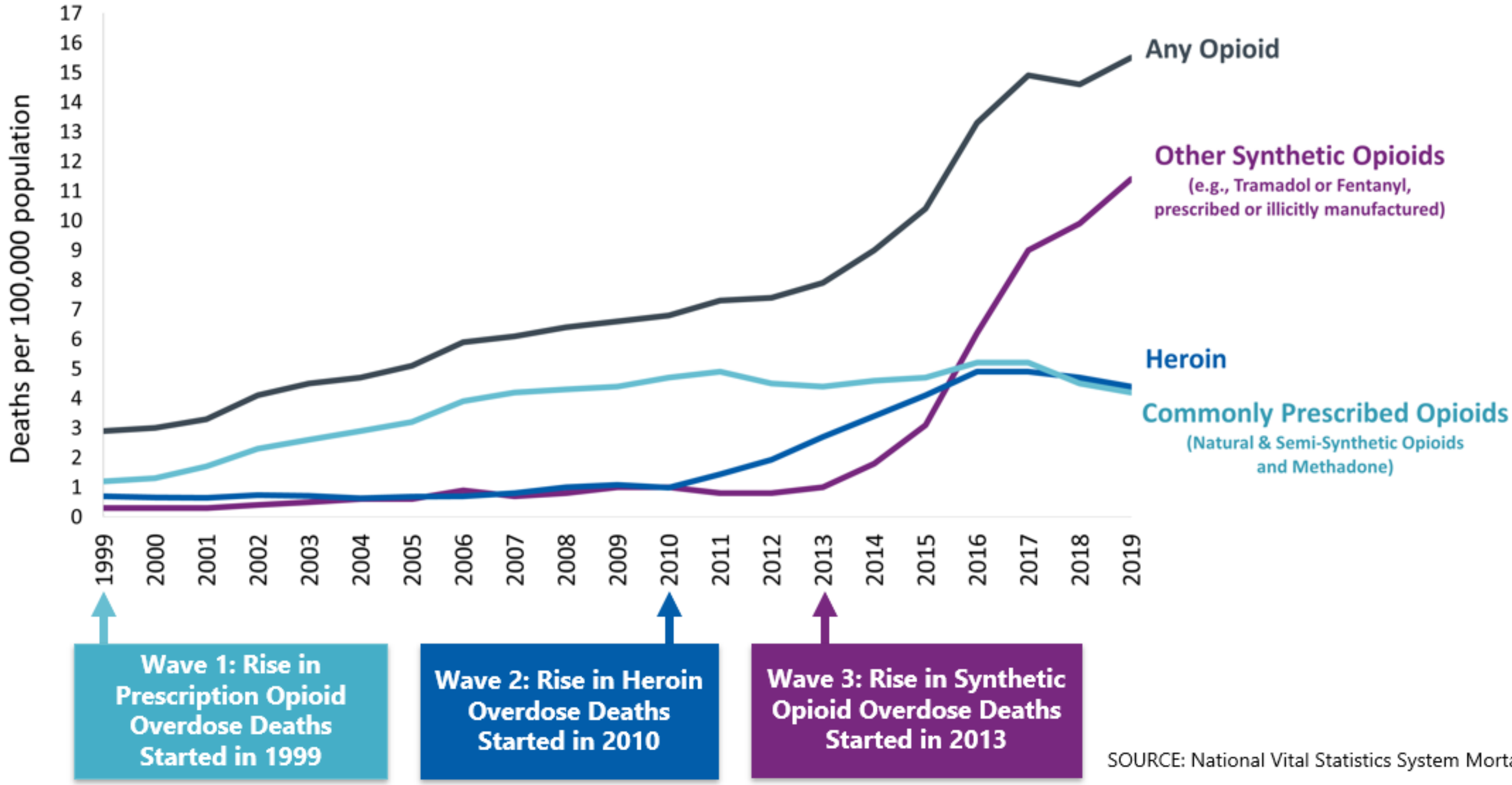
# *AI Ethics:* Explainability of Machine Learning

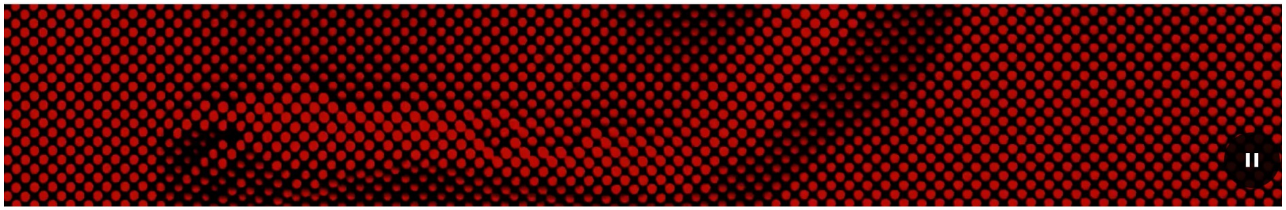
CS229: Machine Learning
Carlos Guestrin
Stanford University

# Three Waves of the Rise in Opioid Overdose Deaths




SOURCE: National Vital Statistics System Mortality File.

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE



VIDEO: SAM CANNON

MAIA SZALAVITZ BACKCHANNEL AUG 11, 2021 6:00 AM

# The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

**A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.**

**The AI Database →**

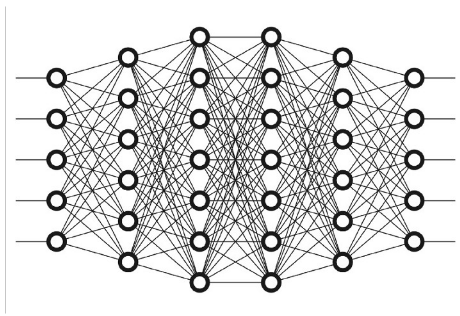
APPLICATION: ETHICS, PREDICTION, REGULATION

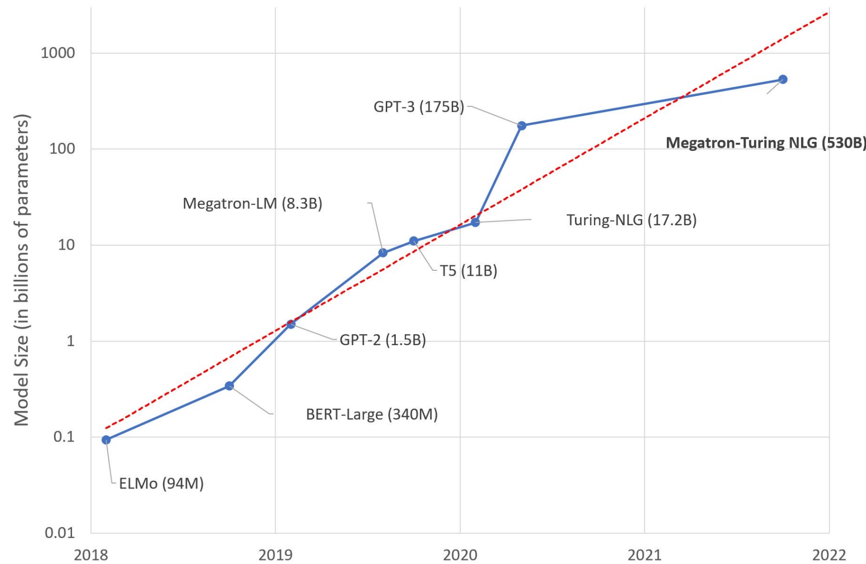
SECTOR: HEALTH CARE, PUBLIC SAFETY

**ONE EVENING IN** July of 2020, a woman named Kathryn went to the hospital in excruciating pain.

A 32-year-old psychology grad student in Michigan, Kathryn lived with endometriosis, an agonizing condition that causes uterine-like cells to abnormally develop in the wrong

# ML Models More and More Complex





When is a model ready to deploy?
*Hard to understand when models are working (for the right reasons) and not working!!*

# Isn't test accuracy enough?

# A User Study on Test Accuracy

“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

# Train a neural network to predict wolf v. husky

Husky

Wolf

Slate

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

# The Man Who Accidentally Adopted a Wolf Pup

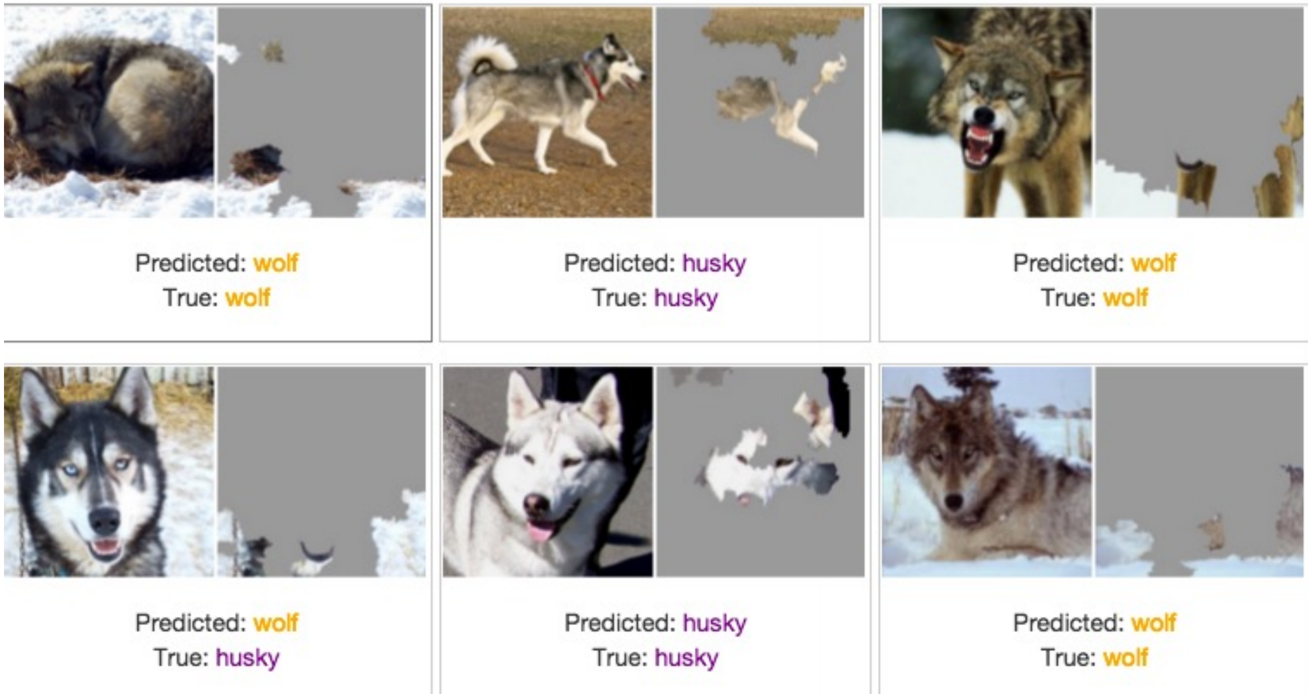
It did not go well.

By *A.J. McCarthy*

10k 547 6

# Train a neural network to predict wolf v. husky

# Explanations for neural network prediction



# Test accuracy may not capture critical issues

- Bad data
- Biases
- Poor performance in critical cases
- ...

# Examining Models

# Debugging is One Reason to Examine Models

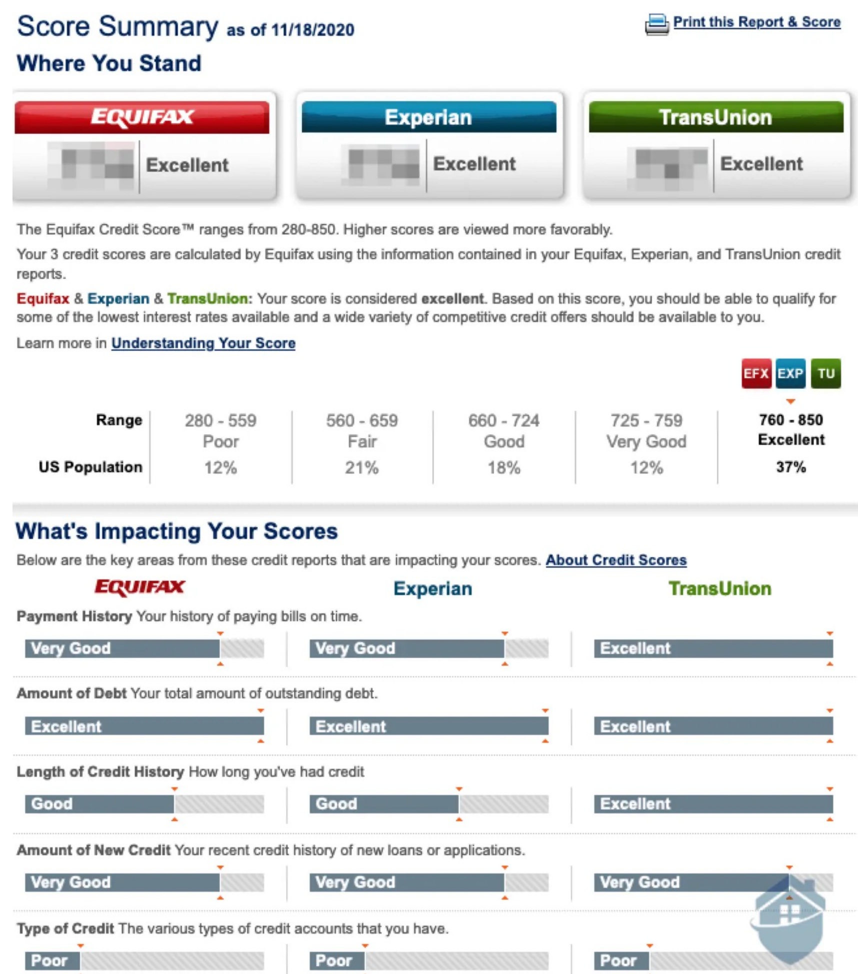
- Examining models:
  - Why a model makes particular predictions
  - What alternative predictions are possible
  - How robust/stable are predictions
  - What data supports predictions
- Examining models for debugging: discover bad, unexpected or unstable behavior
  - Typically not discovered by accuracy in train/test data

# Examining Models to Detect Algorithmic Bias

- Evaluate multiple fairness criteria
- Verify how/if decisions depend on sensitive features
- Discover what groups are privileged/disadvantaged by predictions

# Examine Models for Recourse

- In opioid overdose risk case, patient deemed risky had no way to discover why
  - Or how to fix bad data
- Understanding why could enable individuals to:
  - Address data issues
  - Change their actions to change outcomes

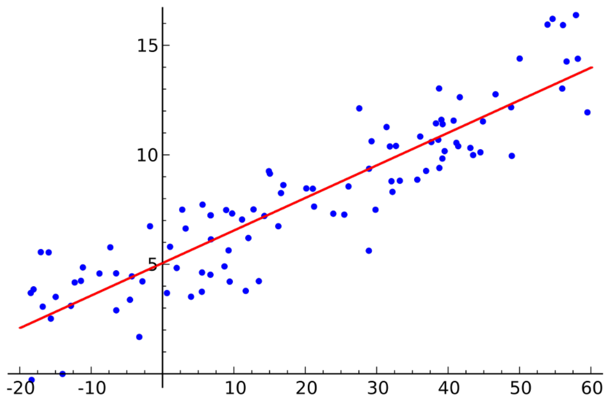


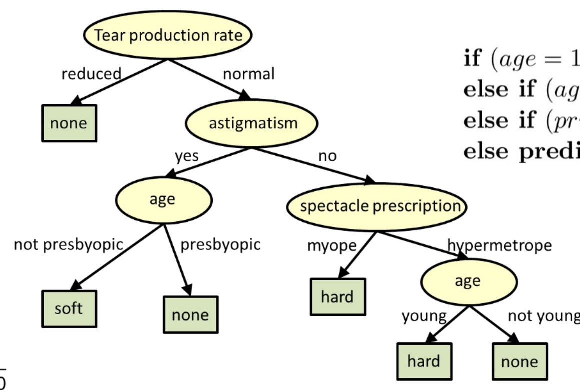
# Interpretable Models vs Post-hoc Explanations

# Interpretability in ML

*Giving humans a **mental model** of the machine's model behavior*

# Learning Interpretable Models (c.f., Lethan & Rudin 2015)

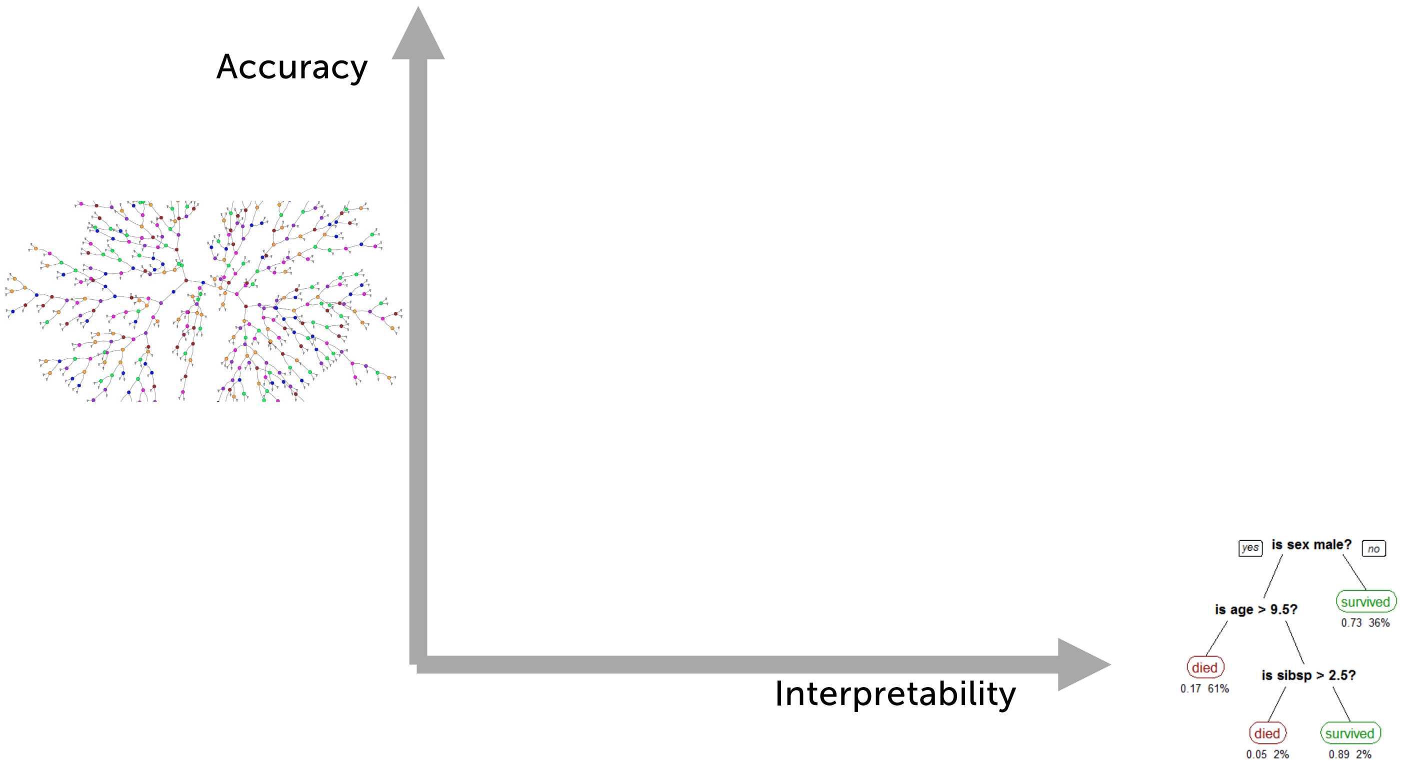



**if** $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** $yes$
**else if** $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** $yes$
**else if** $(priors > 3)$ **then predict** $yes$
**else predict** $no$

Image credit: Lakkaraju, Adebayo, Singh NeurIPS 2020 Tutorial

# Accuracy vs Interpretability



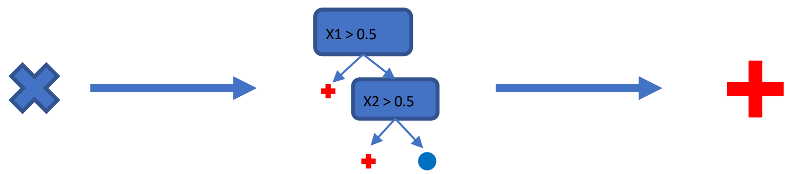
# Post-hoc Explanations

- Given a (huge, complex) model, provide human explanations for predictions

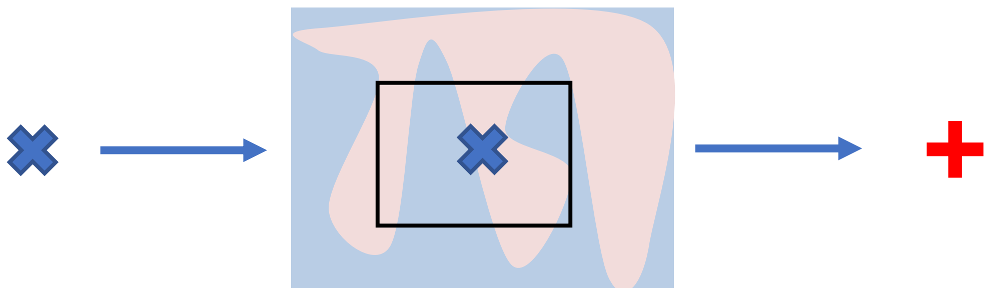
# LIME: Local, Interpretable Model-Agnostic Explanations

"Why should I trust you?": Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

**Model agnostic** → Ignore any internal structure

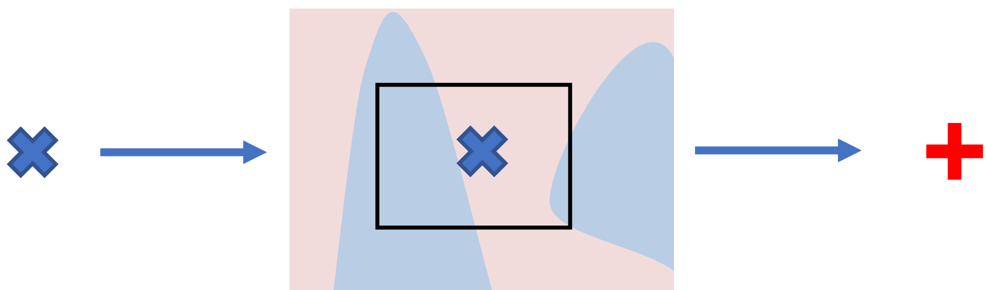


**Explaining predictions** Global decision may be very complicat



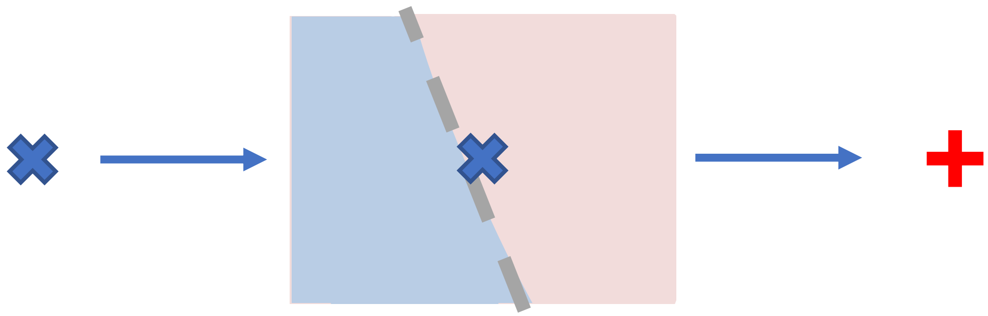
# Explaining predictions

## Locally, decision looks simpler…



## Explaining predictions

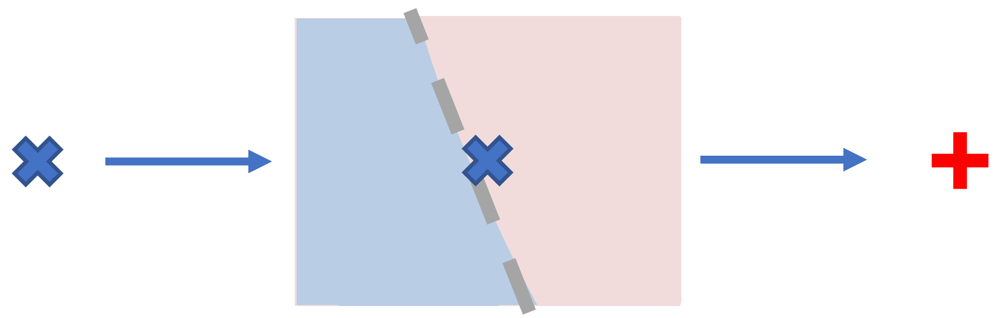
Very locally, decision looks linear



“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

Explaining predictions Very locally, decision looks linear

**LIME**: Learn locally sparse linear model around each prediction

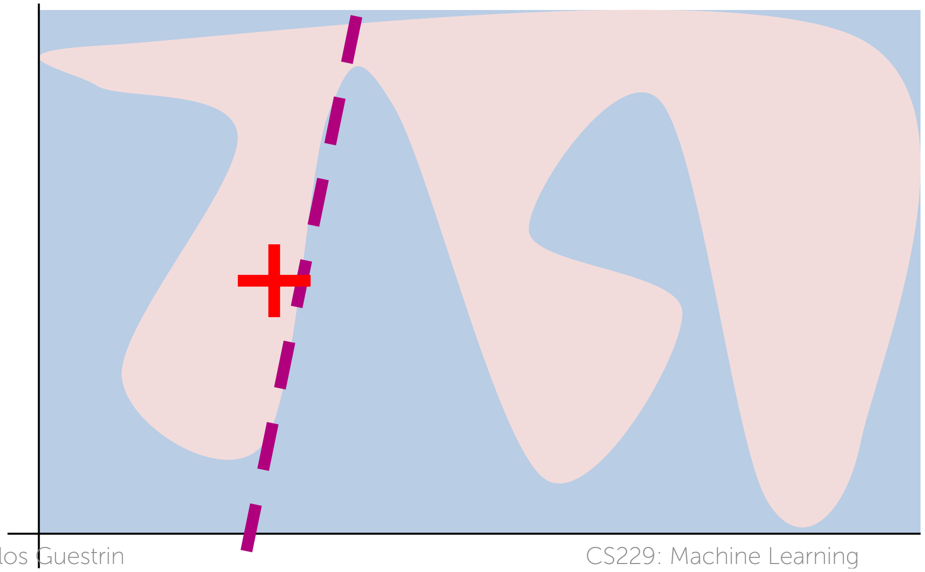


"Why should I trust you?": Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

# LIME – Key Ideas

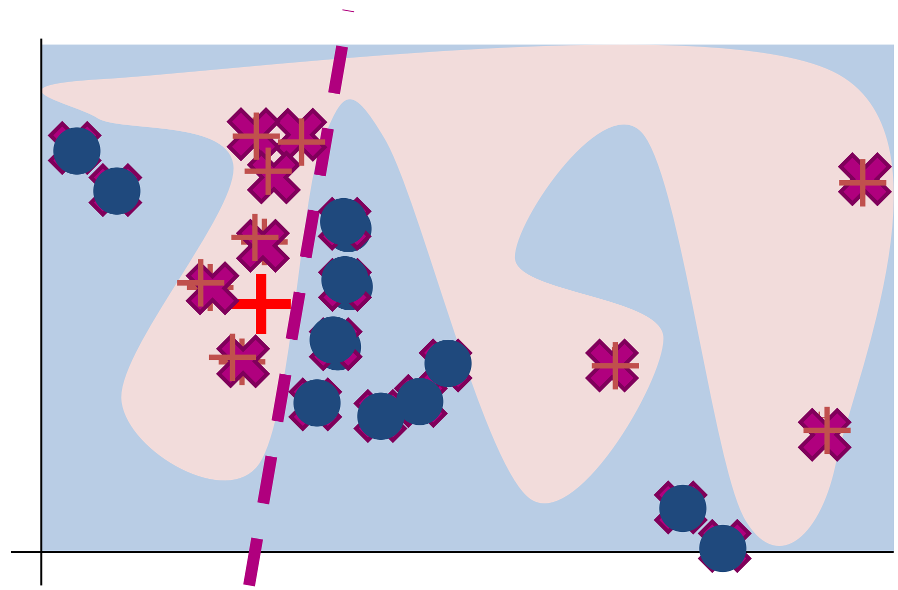
1. Pick a model class interpretable by humans

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



# Sparse linear Explanations

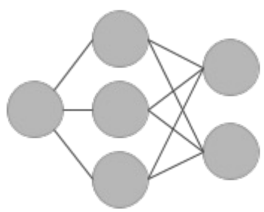
1. Sample points around $x_i$
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to $x_i$
4. Learn new simple model on weighted samples
5. Use simple model to explain



# Interpretable representations

x (embeddings)

| 0.5 | 0.3 | 1.3 | 4.4 | 1.1 | … |
|---|---|---|---|---|---|



Model

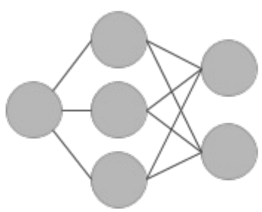
x' (words)

This is a horrible movie.

Human

# Interpretable representation: images

x (3 color channels / pixel)



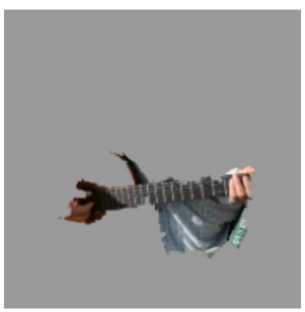
Model

x' (contiguous superpixels)

Human

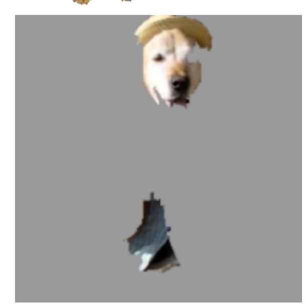
# Explaining prediction of Inception Neural Network

P() = 0.32



P() = 0.24

P() = 0.21



# Achieving target metric may not be enough

Atheism vs Christianity posts
(Newsgroups data, circa 1995)

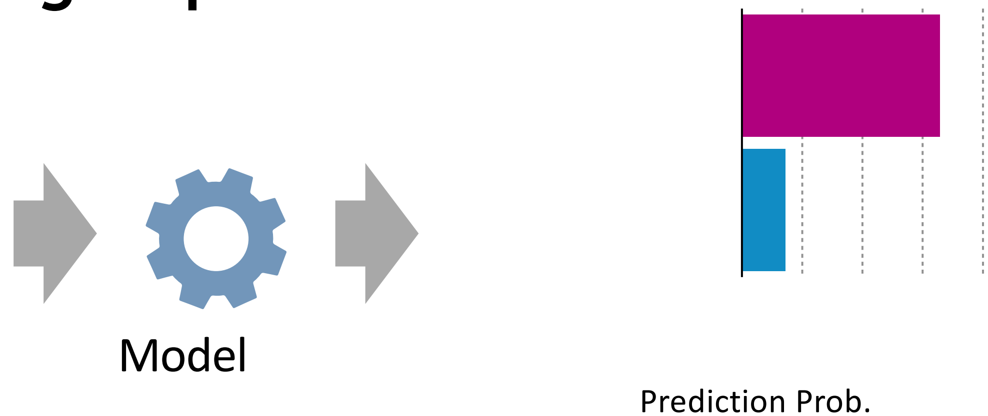
↓

94% accuracy!!!

# LIME applied to 20 newsgroups

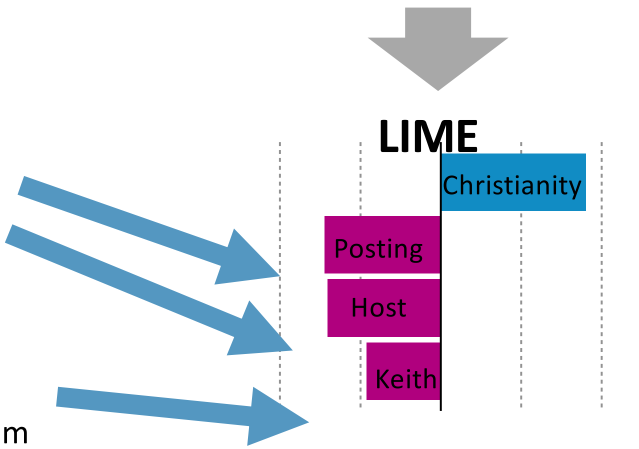
From: Keith Jones
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

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

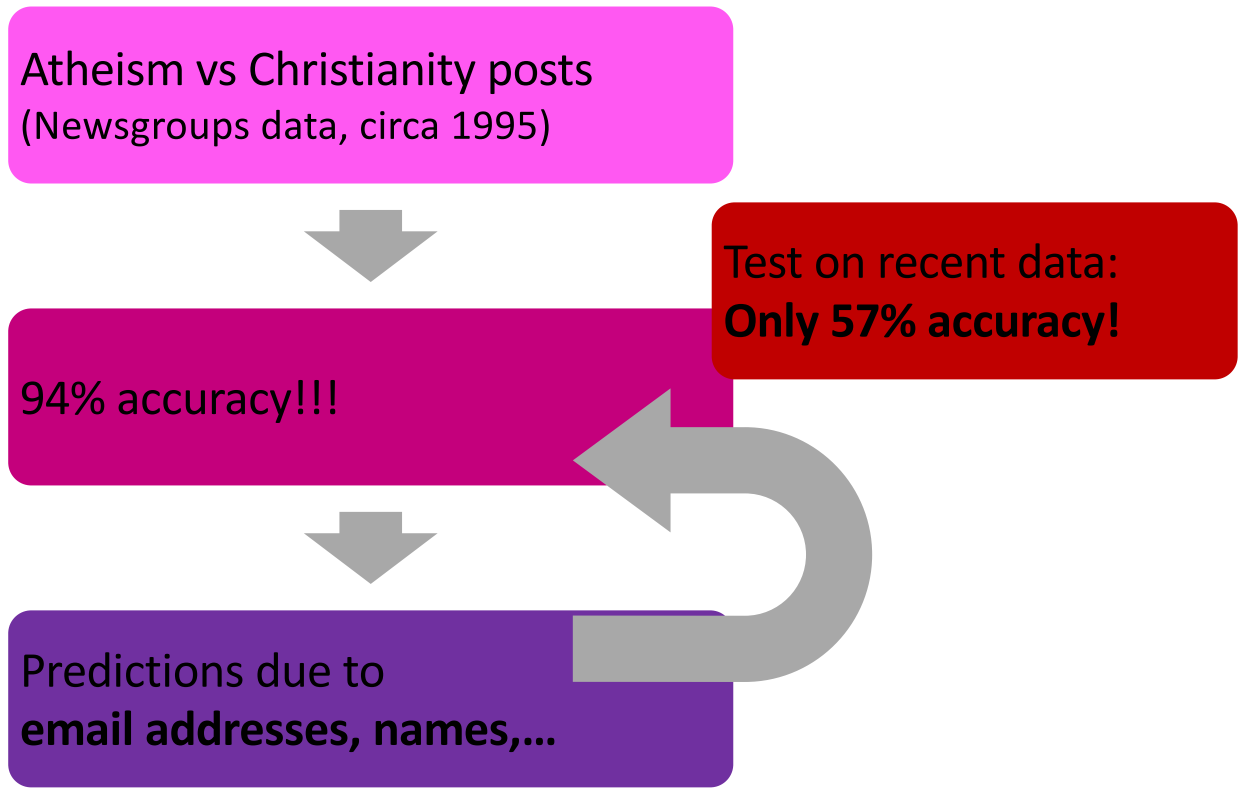



Appear in 21% of training examples, almost always in Atheism

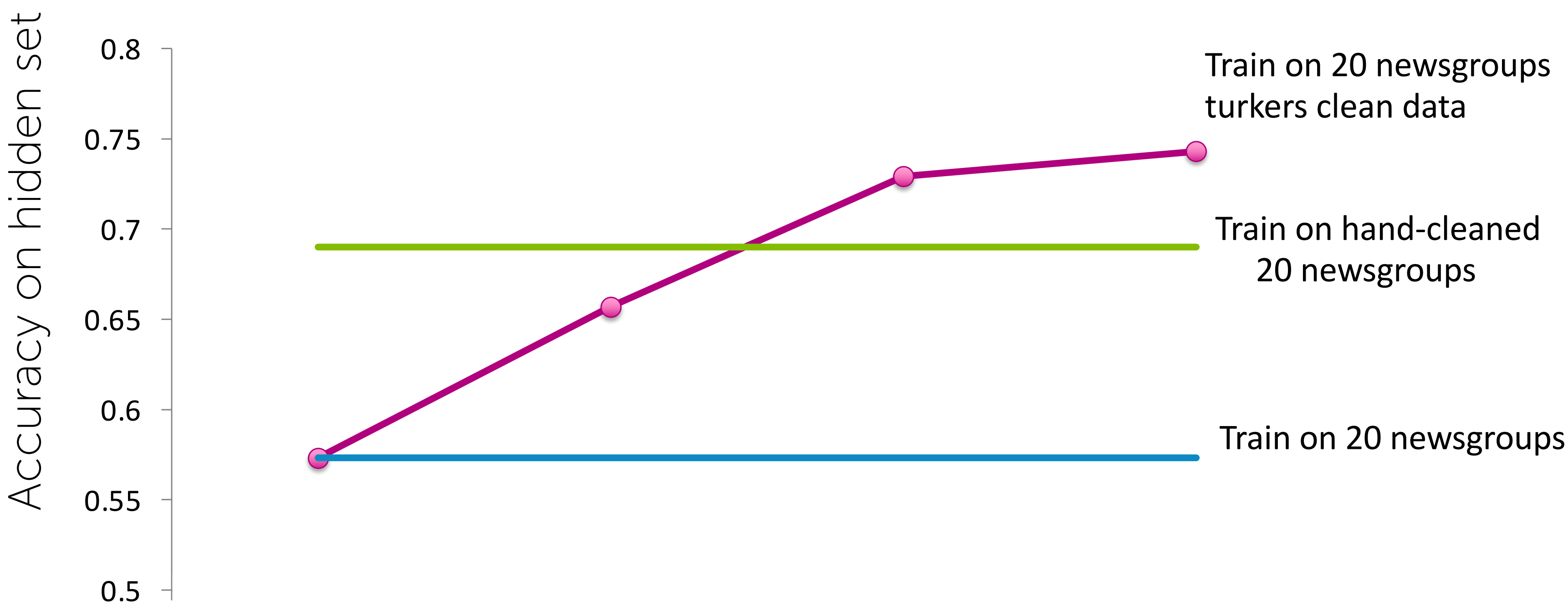
Appears in 11% of training examples, **always** in atheism

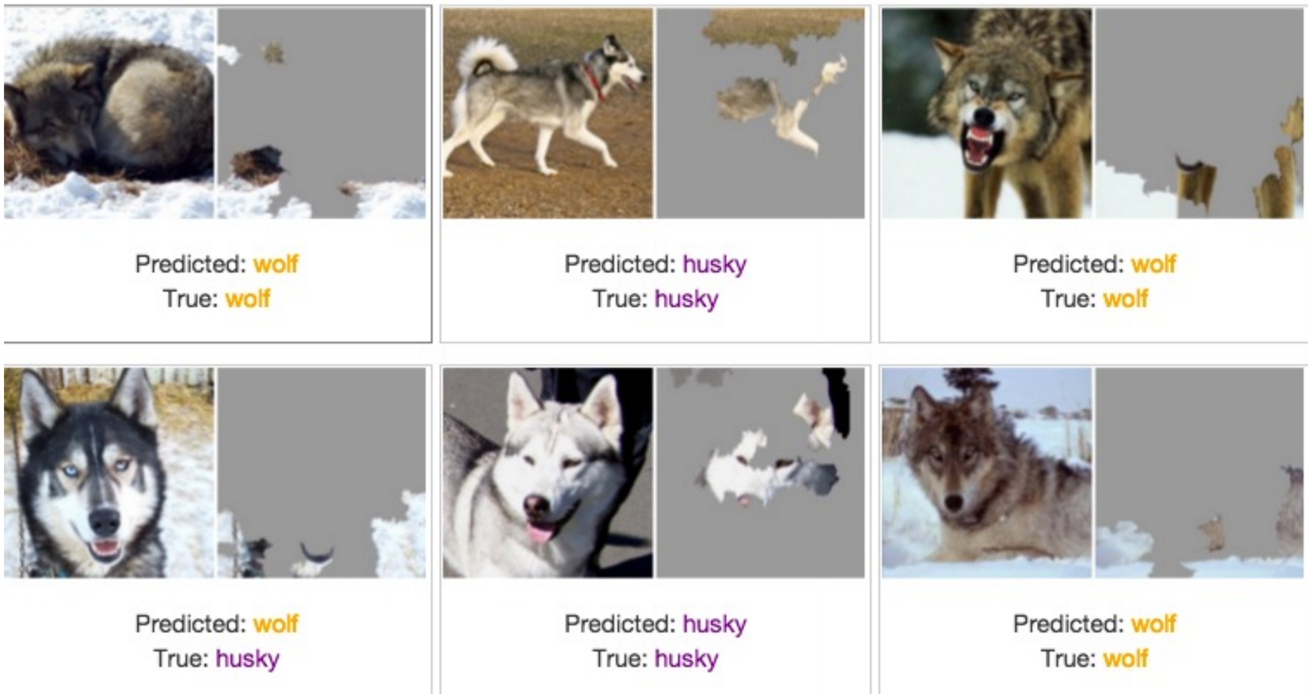


# Achieving target metric may not be enough

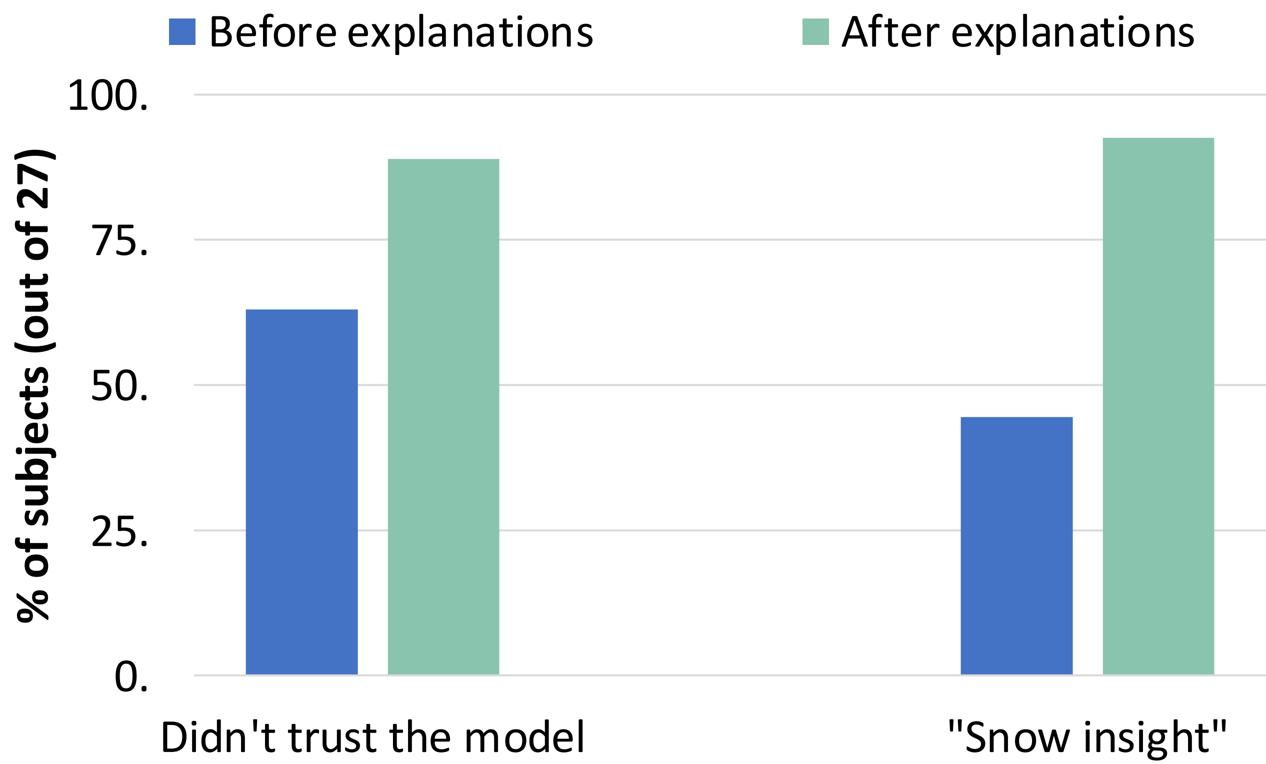


# Fixing bad classifiers



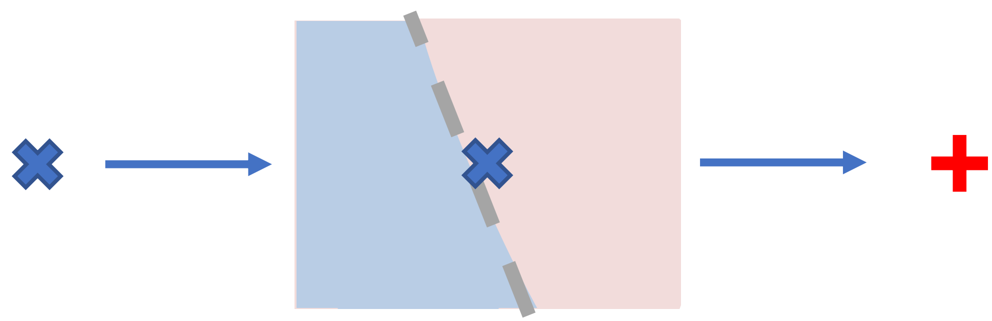

Predicted: wolf
True: wolf
Predicted: husky
True: husky
Predicted: wolf
True: wolf
Predicted: wolf
True: husky
Predicted: husky
True: husky
Predicted: wolf
True: wolf

# Did explanations help with wolf problem?



# More Examples

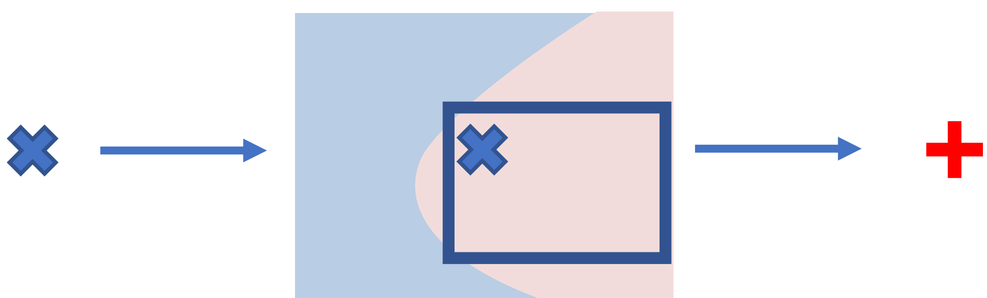
# **LIME:** Learn locally sparse linear model around each prediction



“Why should I trust you?”: Explaining the Predictions of Any Classifier. Ribeiro, Singh & G. KDD 16

## Anchors: Sufficient Conditions

Conditions under which classifier makes same prediction



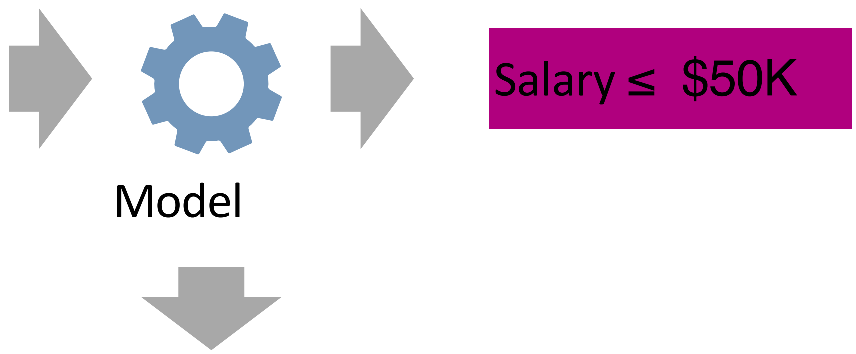
Anchors: High-Precision Model-Agnostic Explanations. Ribeiro, Singh & G. AAAI 18

# Salary Prediction

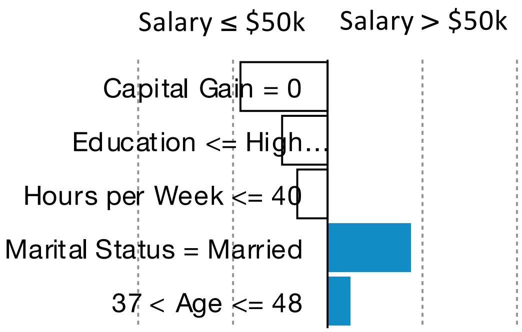
| Feature | Value |
| --- | --- |
| Age | $37 < \text{Age} \leq 48$ |
| Workclass | Private |
| Education | $\leq$ High School |
| Marital Status | Married |
| Occupation | Craft-repair |
| Relationship | Husband |
| Race | Black |
| Sex | Male |
| Capital Gain | 0 |
| Capital Loss | 0 |
| Hours per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

# Salary Prediction: LIME vs Anchors

| 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 | $\leq 40$ |
| Country | United States |




**LIME**




**Anchor**

**IF Education ≤ High School**
**Then**
**P(prediction = ≤ 50K) > 0.95**

# Anchors for Images: Classification

Prediction: Beagle

Anchor for Beagle

# Anchors for Visual Question Answering


| | |
|---|---|
| What is the mustache made of? | *Banana* |
| How many bananas are in the picture? | *2* |

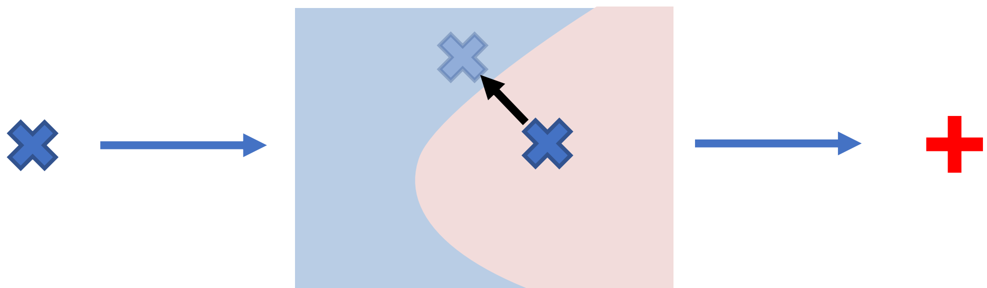
# Anchors for Visual Question Answering

| | |
|---|---|
| **What** is the mustache made of? | *Banana* |
| **What** is the ground made of? | *Banana* |
| **What** is the hair made of? | *Banana* |
| **What** is the picture of? | *Banana* |
| **What** was the head of the US? | *Banana* |

| | |
|---|---|
| How **many** bananas are in the picture? | *2* |
| How **many** are in the picture? | *2* |
| How **many** people in the picture? | *2* |
| Are there **many** animals in the picture? | *2* |
| How **many** is too many? | *2* |

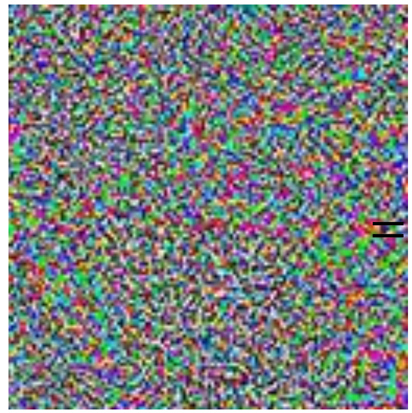
## Adversarial Bug Discovery

Find closest input with different prediction



# Oversensitivity in image classification

+ ε



=

"Panda"

"Gibbon"

Adversary not distinguishable by human
→ Unlikely to be a real-world issue (except for attacks)

Explaining and Harnessing Adversarial Examples. Goodfellow, Shlens & Szegedy 2015


What type of road sign is shown?

STOP

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.

How long is the Rhine?

1,230 km


What type of
road sign is shown?

Which type of
road sign is shown?

STOP

Do not enter

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe, at about 1,230 km.

How long is
the Rhine?

How long is
the Rhine??

1,230 km

More than
1,050,000

# Goal: Find semantically-equivalent adversarial examples

Semantically Equivalent Adversarial Rules for Debugging NLP Models. Ribeiro, Singh & G. ACL 18

**Semantically-equivalent**

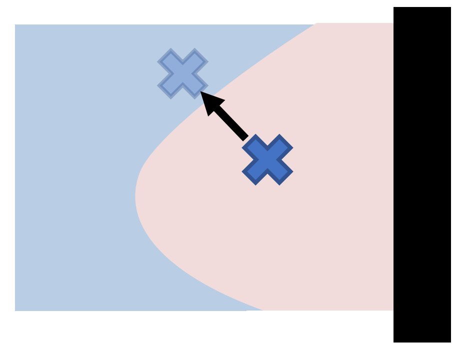
Use paraphrasing model [Lapata et al. 2017]

**+**

**Adversarial**

Changes correct model prediction

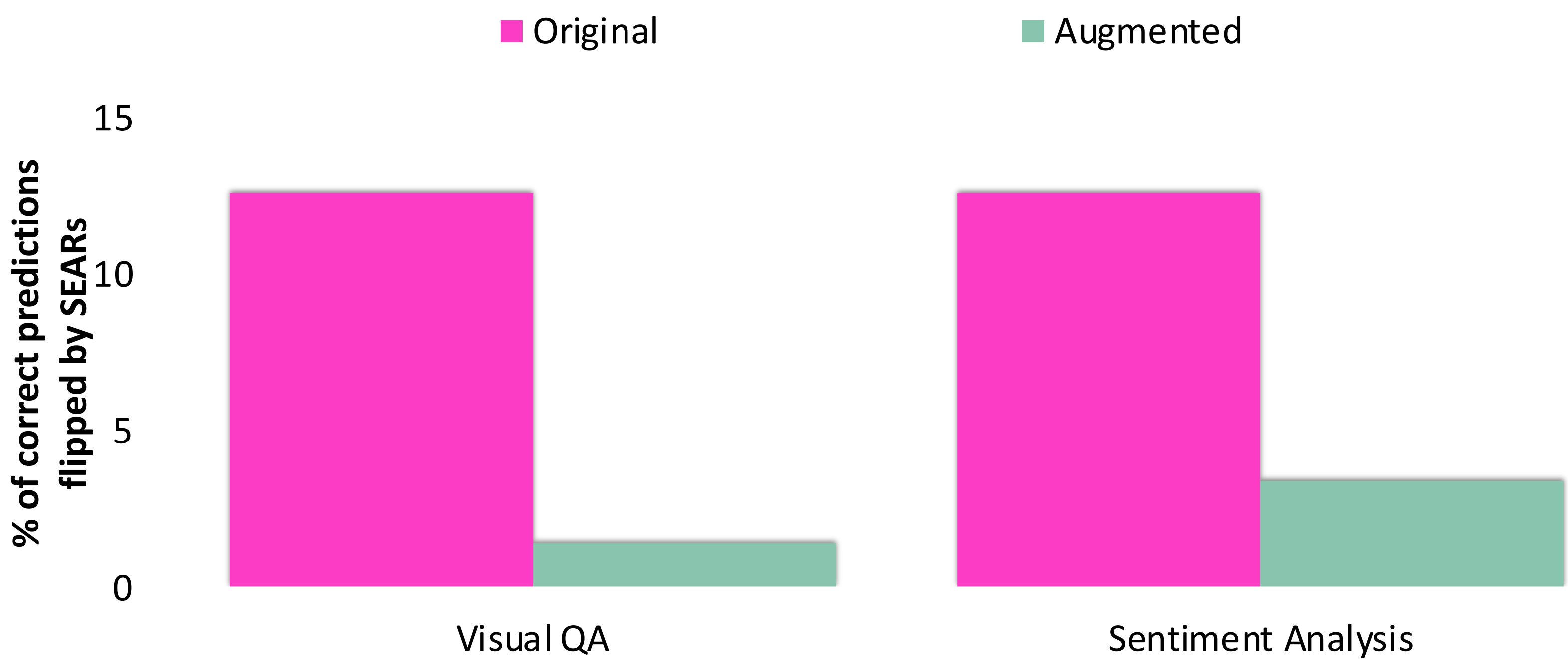
| | |
|---|---|
| What color is the tray? | *Pink* |
| What **colour** is the tray? | *Green* |
| **Which** color is the tray? | *Green* |
| What color is **it**? | *Green* |
| What color is ~~the~~ tray? | *Pink* |
| **How** color is ~~the~~ tray? | *Green* |



Semantically Equivalent Adversarial Rules for Debugging NLP Models. Ribeiro, Singh & G. ACL 18

# Closing the Loop with Simple Data Augmentation

## Augment by applying validated SEARs to training data



# Typical challenges with explainability methods

- Explanations to simplistic
- Not focused on information needs for task
- Unstable
- Not causal
- …