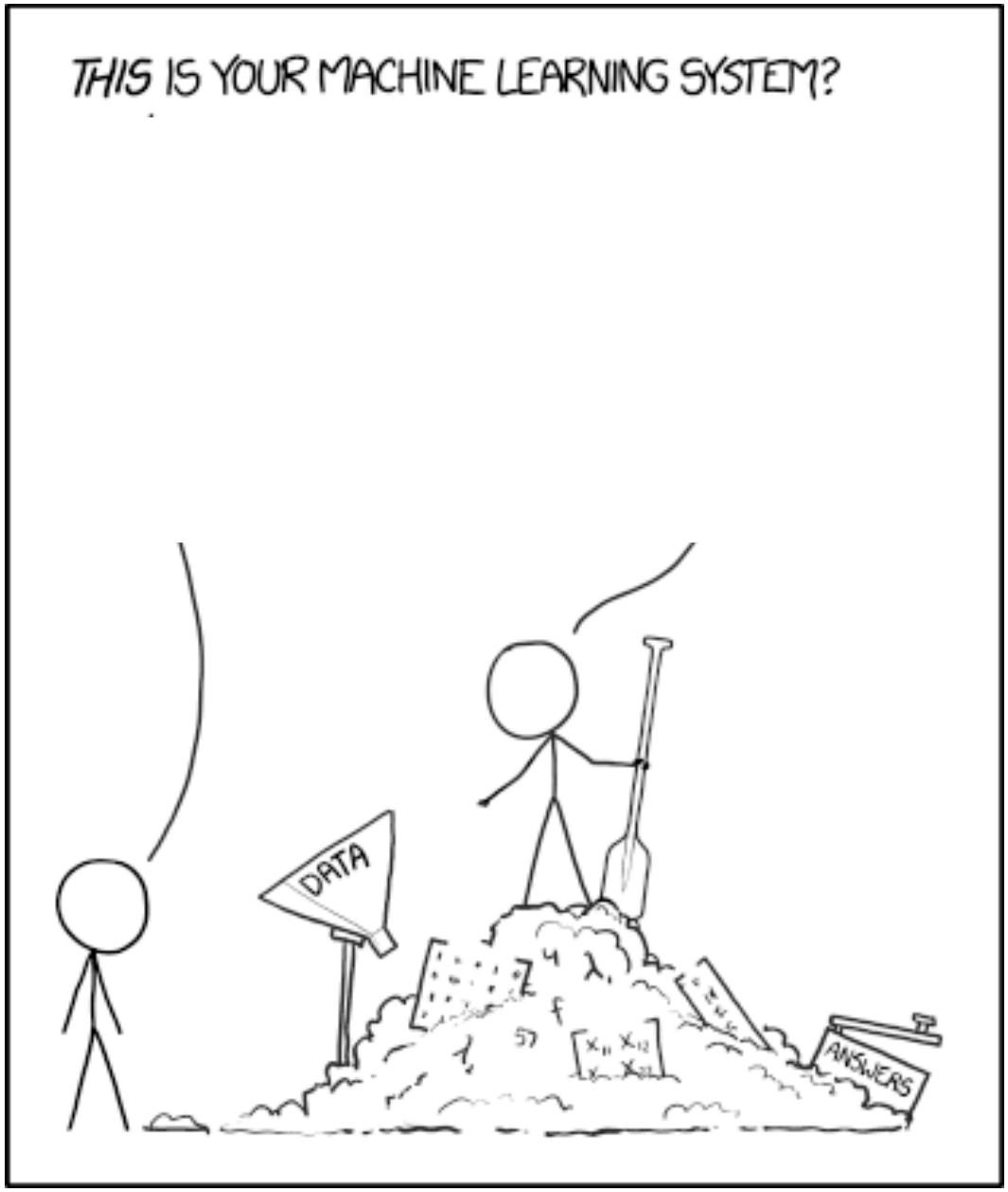
Interpretable Machine Learning

An overview-flavored talk containing a subset of work from the last year

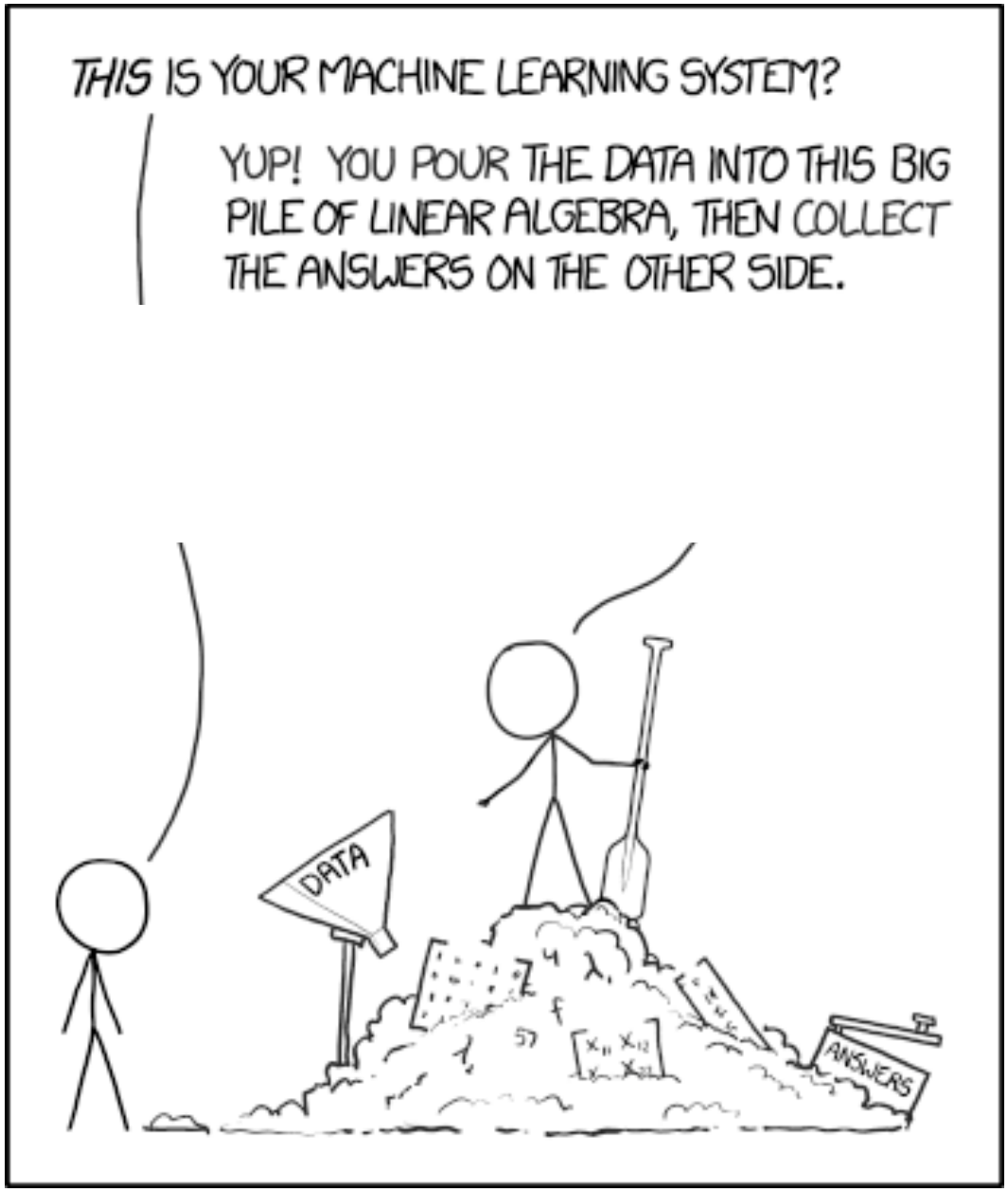
Been Kim

with a lot of awesome people inside and outside of Google:

Marten Wattenberg, Julius Adebayo
Justin Gilmer, Carrie Cai, Fernanda Viegas, Rory Sayres, Mortiz Hardt, Ian Goodfellow



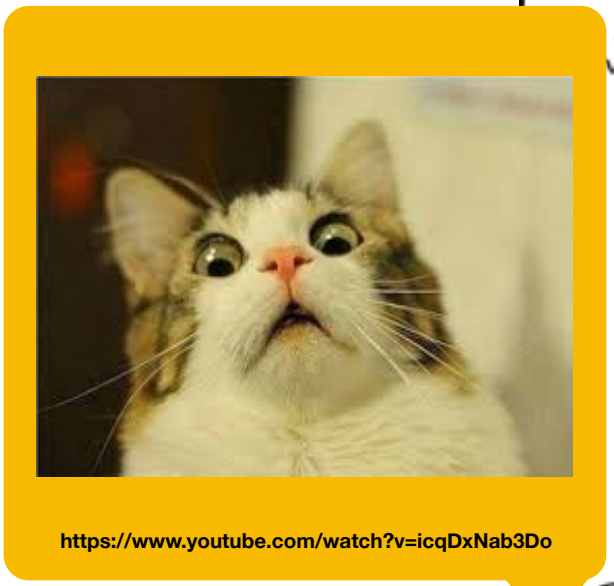
https://xkcd.com/



https://xkcd.com/

https://xkcd.com/

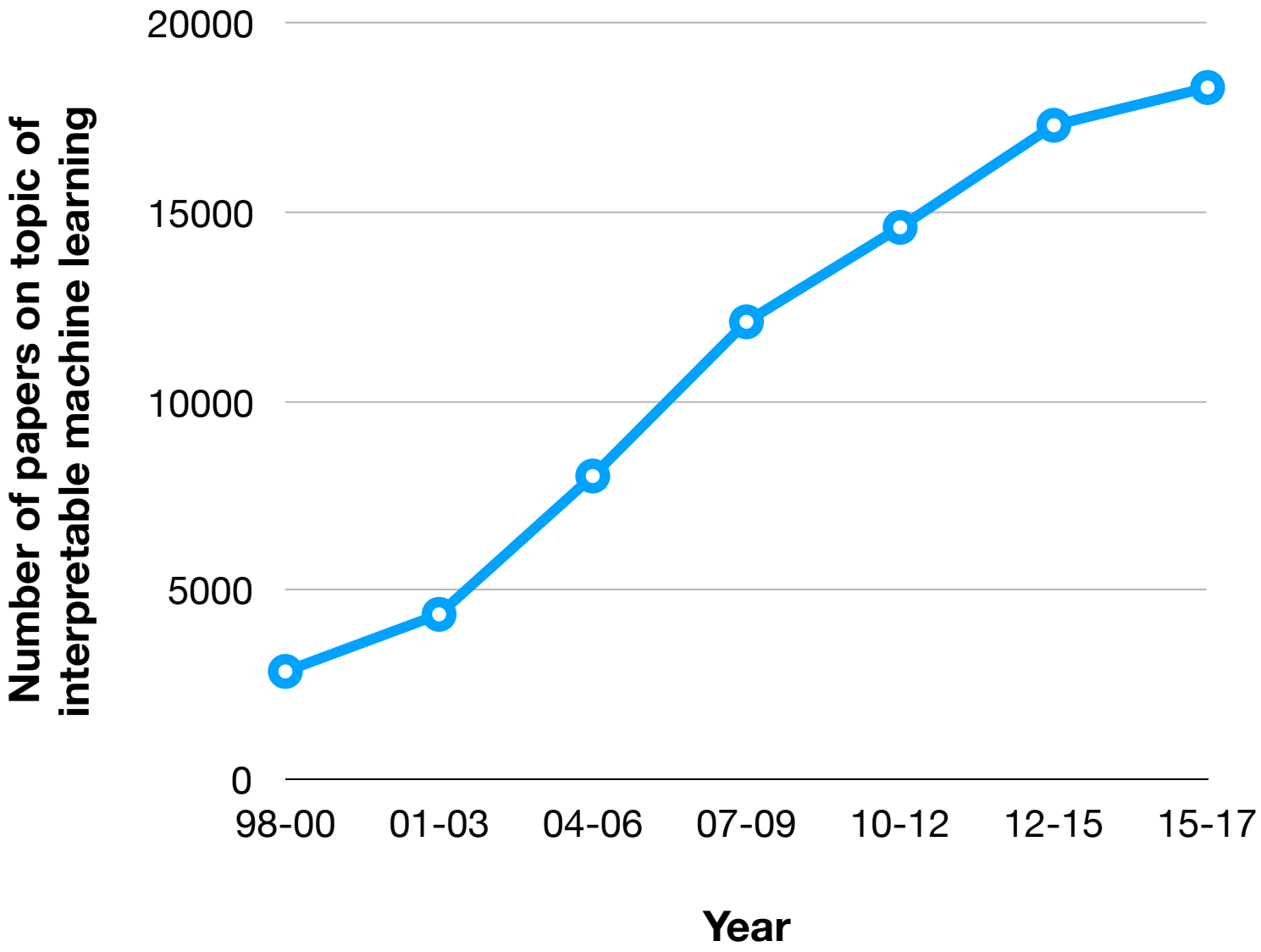
https://xkcd.com/



https://www.youtube.com/watch?v=icqDxNab3Do

https://xkcd.com/

ML community is responding



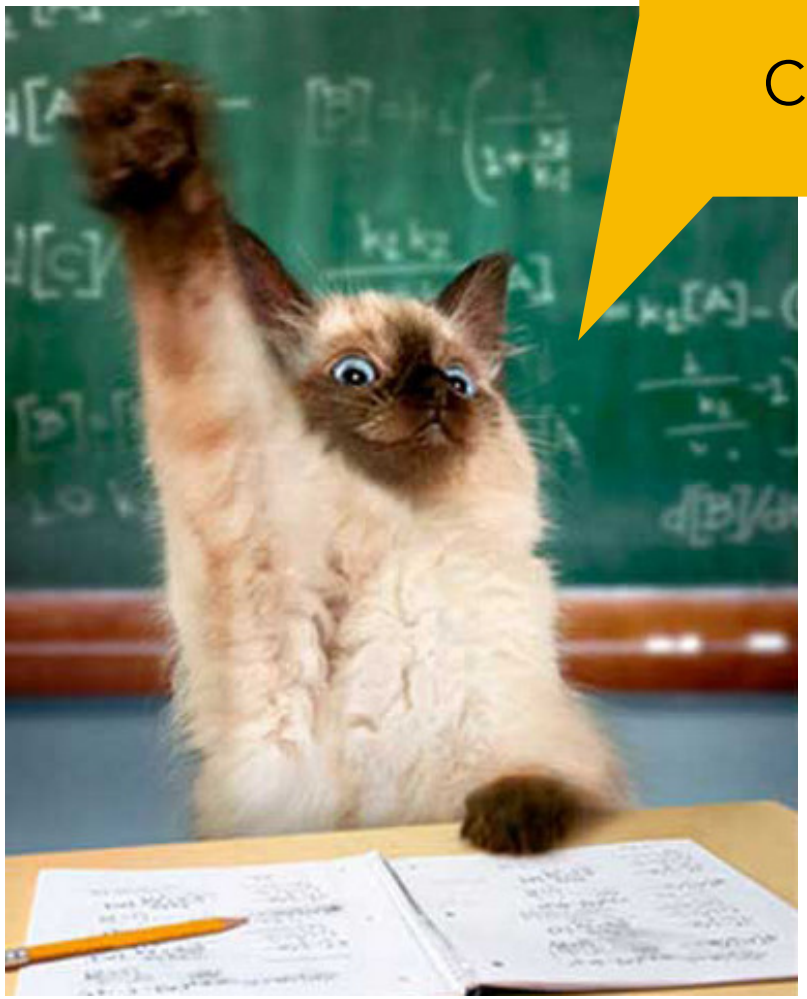
This is not a new problem. Why now?

Complexity and prevalence!

http://godeater.wikia.com

I heard you can just use decision trees…

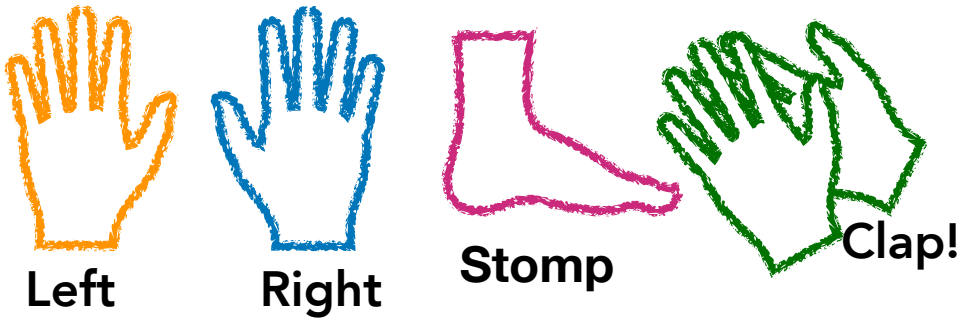
Can we go home now?



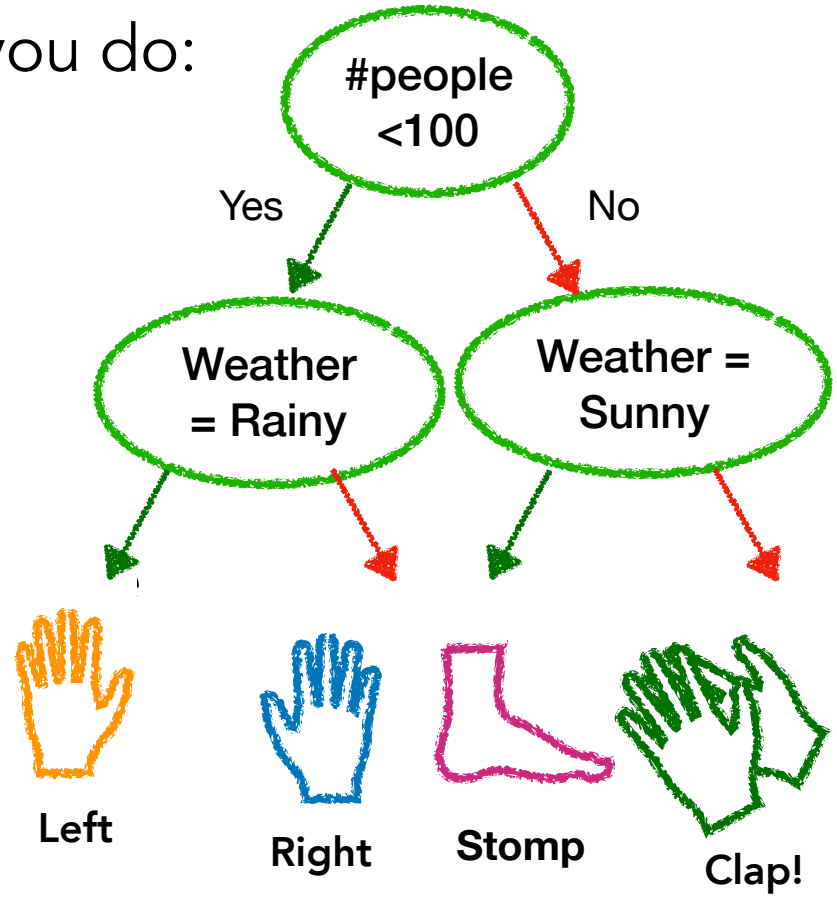
http://www.ogroup.com.au/raise-your-hand-when-you-should-and-why-you-should/

Experiment.

- I will show you a decision tree. Follow the right path given a data point, and you do:

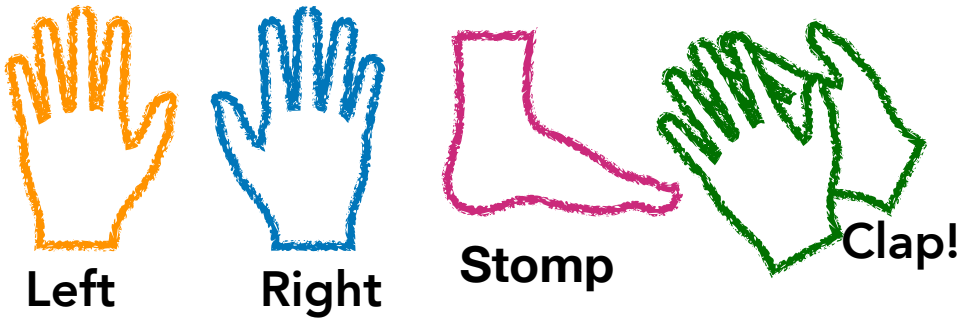


Data = [Sunny, 200]

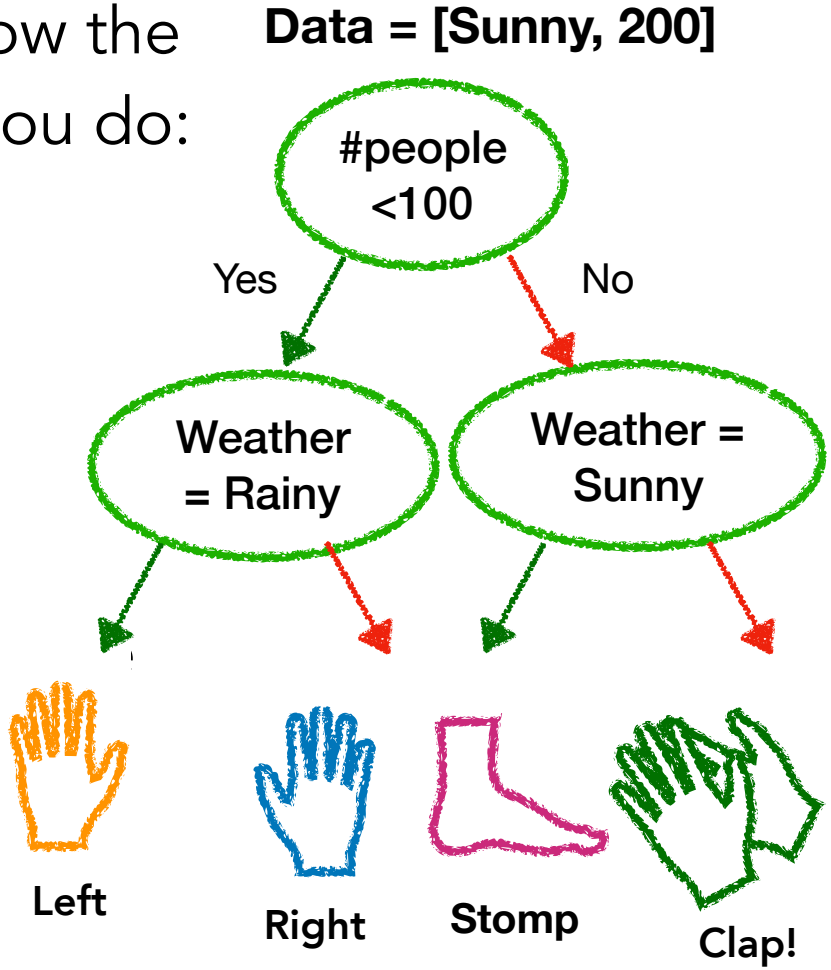


Experiment.

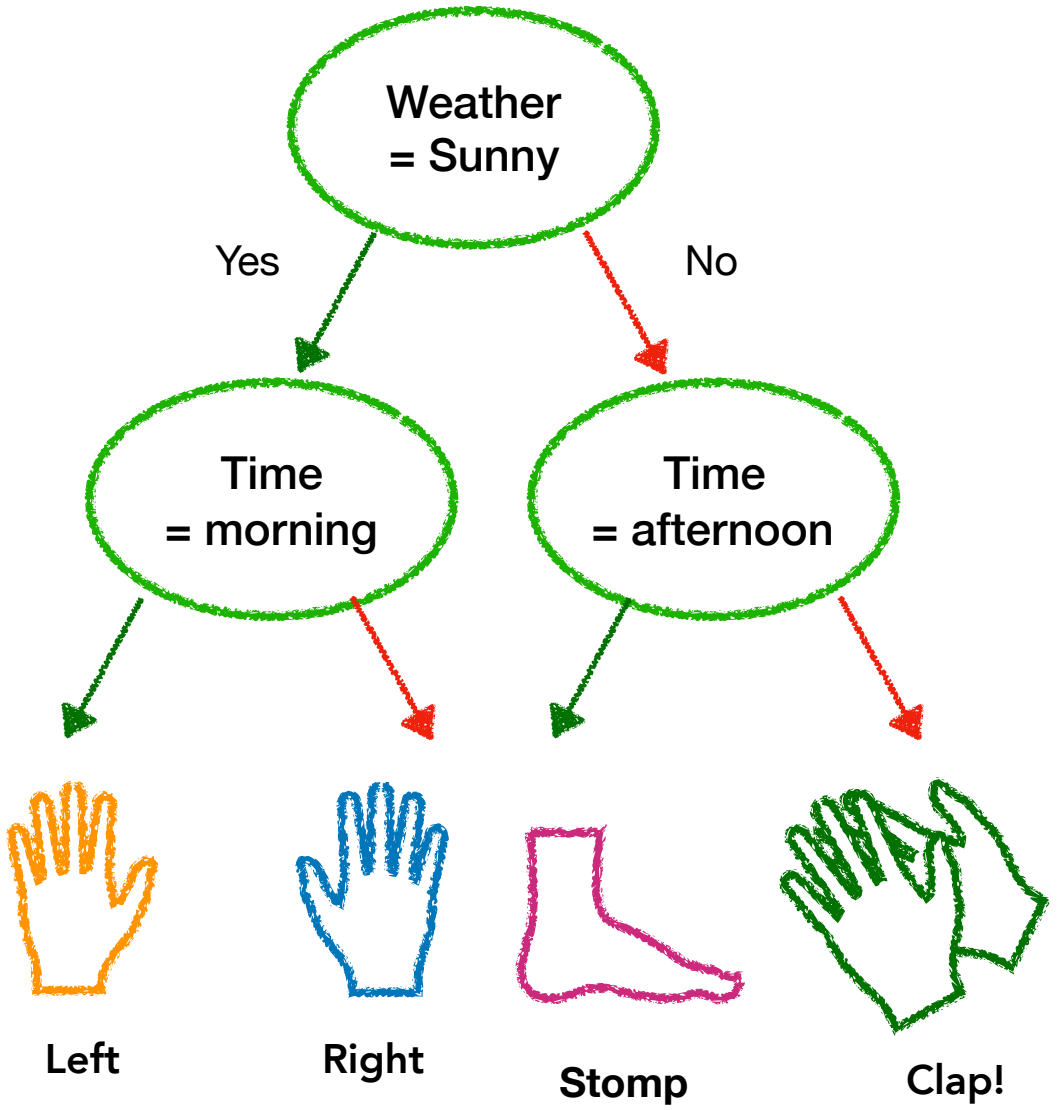
- I will show you a decision tree. Follow the right path given a data point, and you do:



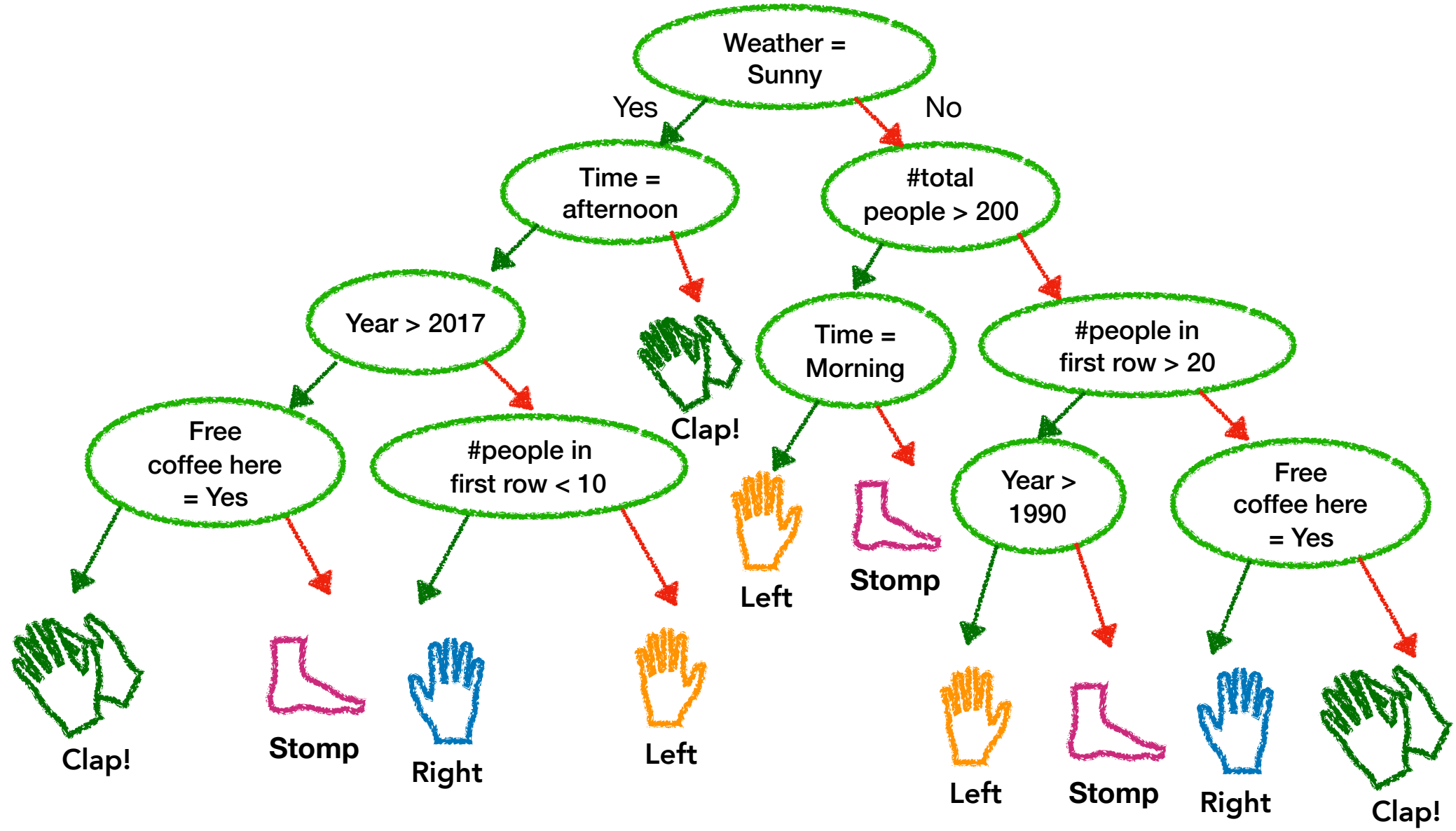
- As soon as you know the answer, do the action!



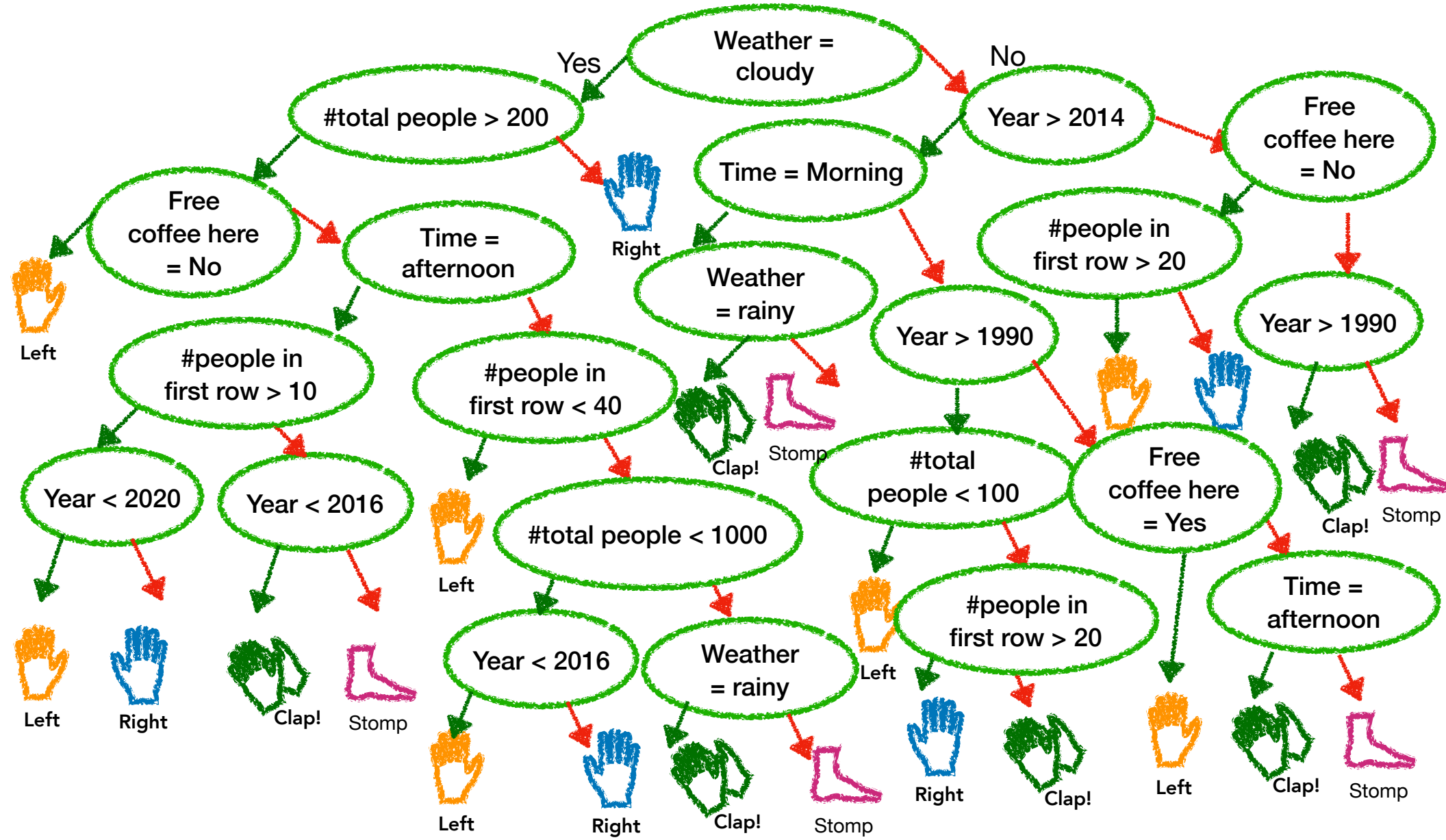
Sample decision tree #1



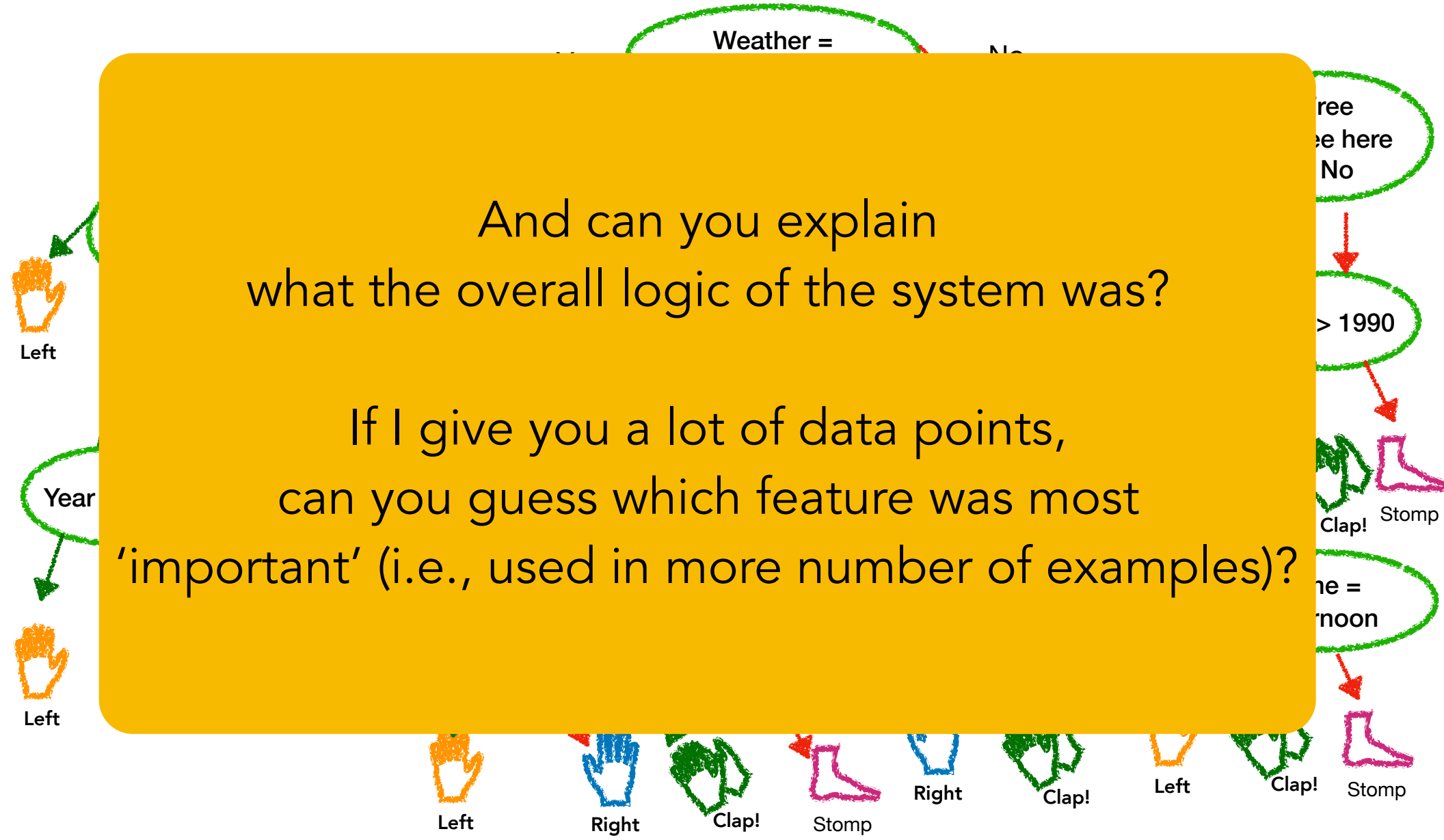
Sample decision tree #2



Sample decision tree #3



Sample decision tree #3



And can you explain
what the overall logic of the system was?

If I give you a lot of data points,
can you guess which feature was most
'important' (i.e., used in more number of examples)?

Common misunderstanding:
Decision trees and linear models are always interpretable.

Do we need a different model? How about rule lists?

If (sunny and hot)	then	go swim
Else if (sunny and cold)	then	go ski
Else	then	go work

Do we need a different model? How about rule lists?

If (sunny and hot)	then	go swim
Else if (sunny and cold)	then	go ski
Else if (wet and weekday)	then	go work
Else if (free coffee)	then	attend tutorial
Else if (cloudy and hot)	then	go swim
Else if (snowing)	then	go ski
Else if (New Rick and Morty)	then	watch TV
Else if (paper deadline)	then	go work
Else if (hungry)	then	go eat
Else if (tired)	then	watch TV
Else if (advisor might come)	then	go work
Else if (code running)	then	watch TV
Else	then	go work

Maybe rule sets are better?

IF (sunny and hot) OR (cloudy and hot) OR (sunny and thirsty and bored)
THEN go to beach
ELSE work

Maybe rule sets are better?

IF (sunny and hot) OR (cloudy and hot) OR (sunny and thirsty and bored) OR (bored and tired) OR (thirty and tired) OR (code running) OR (friends away and bored) OR (sunny and want to swim) OR (sunny and friends visiting) OR (need exercise) OR (want to build castles) OR (sunny and bored) OR (done with deadline and hot) OR (need vitamin D and sunny) OR (just feel like it)
THEN go to beach
ELSE work

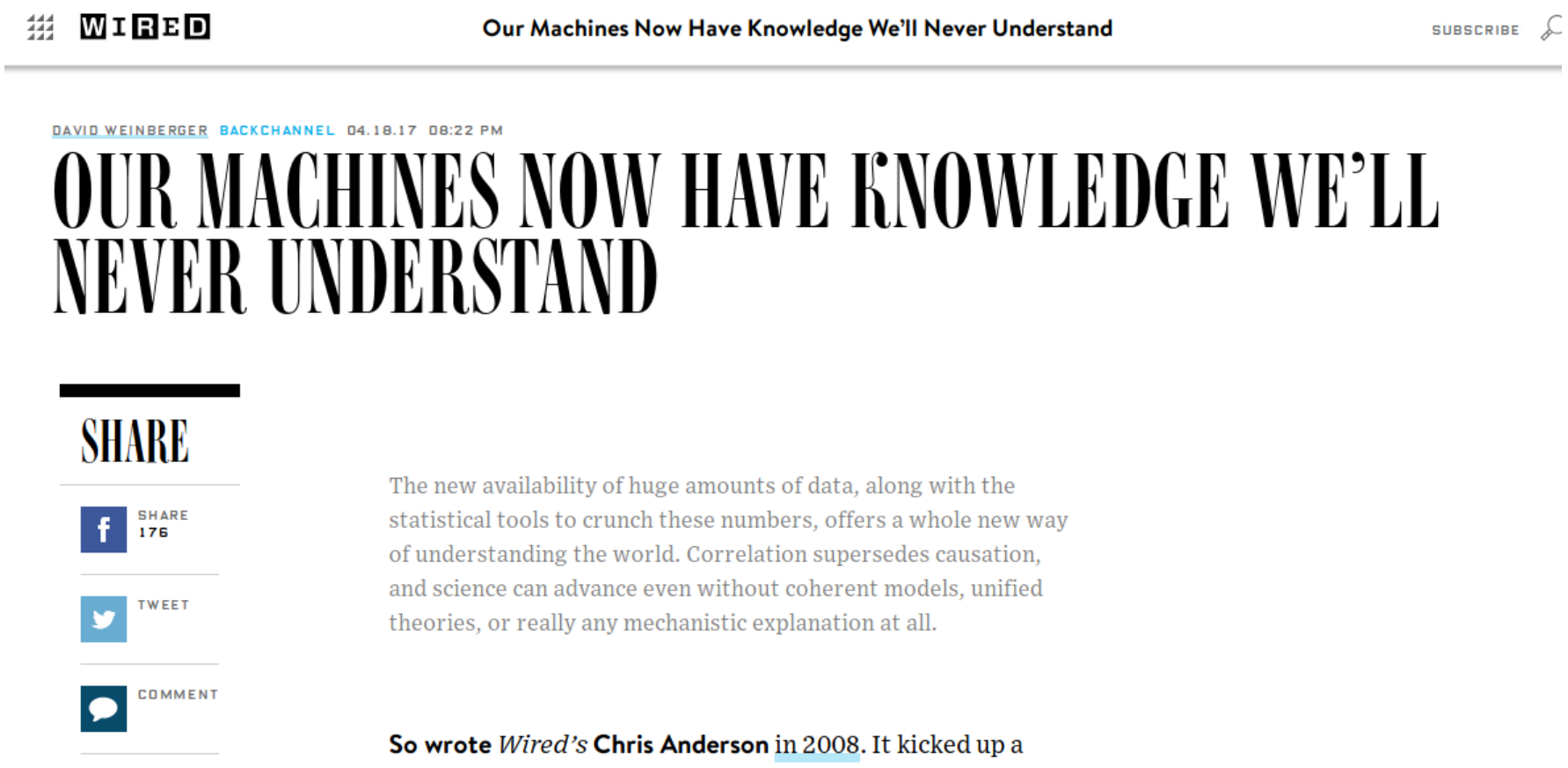
Are you saying decision trees, rule lists and rule sets don't work?!

Decision trees, rule lists or rule sets may work for your case!

The point here is that there is no one-size-fits-all method.

http://blog.xfree.hu/myblog.tvn?SID=&from=20&pid=&pev=2016&pho=02&pnap=&kat=1083&searchkey=&hol=&n=sarkadykati

Is interpretability possible at all?



WIRED Our Machines Now Have Knowledge We'll Never Understand SUBSCRIBE

DAVID WEINBERGER BACKCHANNEL 04.18.17 08:22 PM

OUR MACHINES NOW HAVE KNOWLEDGE WE'LL NEVER UNDERSTAND

SHARE

SHARE 176

TWEET

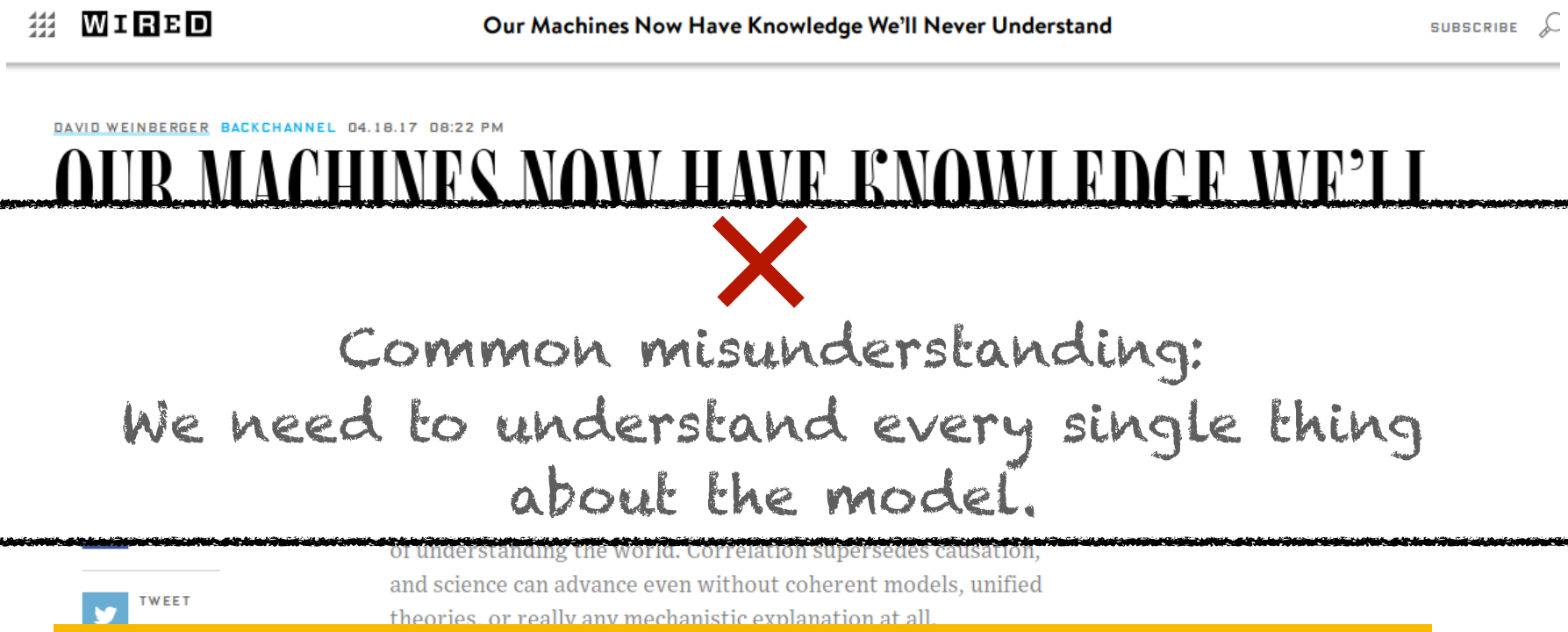
COMMENT

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

So wrote *Wired's* Chris Anderson in 2008. It kicked up a

https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/

Is interpretability possible at all?



Key Point:

Interpretability is NOT about understanding all bits and bytes of the model for all data points.

It is about knowing enough for your goals/downstream tasks.

http

My goal

interpretability

To use machine learning **responsibly**
we need to ensure that
1. our **values** are aligned
2. our **knowledge** is reflected

http://blogs.teradata.com/

Why interpretability?

Fundamental **underspecification** in the problem

Why interpretability?

example1: Safety

Fundamental **underspecification** in the problem

Why interpretability?

example1: Safety

example 2: Science

Fundamental **underspecification** in the problem

Why interpretability?

example1: Safety

example 2: Science

Fundamental **underspecification** in the problem

example3: mismatched objectives

Why interpretability?

example1: Safety

example 2: Science

Fundamental **underspecification** in the problem

Common misunderstanding:
More data or more clever algorithm will solve interpretability.

What is NOT underspecification?

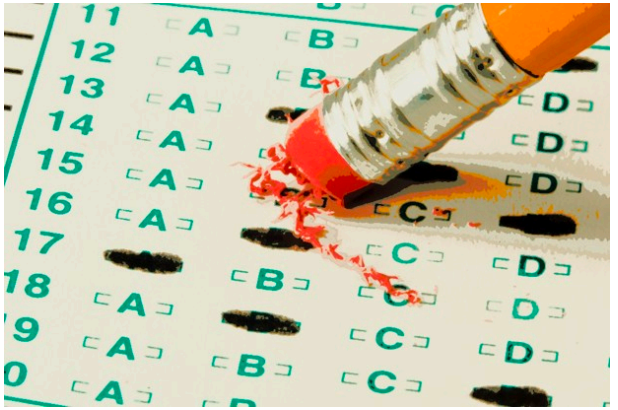
https://www.pinterest.com/dowd3128/type-o-negative/

When we may **not** want interpretability

- No significant consequences or when predictions are all you need.

- Sufficiently well-studied problem

- Prevent gaming the system - mismatched objectives.

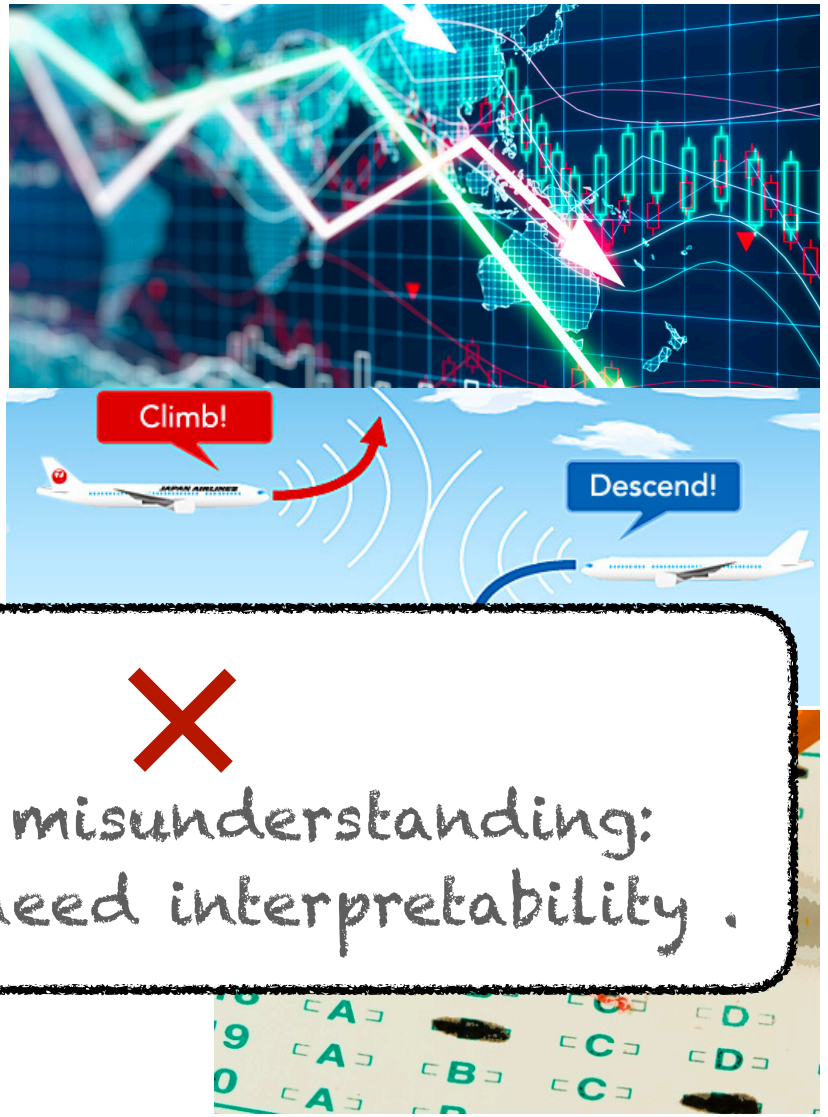


https://cdn.theatlantic.com/assets/media/img/mt/2015/04/shutterstock_11926084/lead_large.jpg
https://www.jal.com/assets/img/flight/safety/equipment/pic_tcas_001_en.jpg

http://www.cinemablend.com/pop/Netflix-Using-Amazon-Cloud-Explore-Artificial-Intelligence-Movie-Recommendations-62248.html

When we may **not** want interpretability

- No significant consequences or when predictions are all you need.
- Sufficiently well-studied problem
- Prevent gaming the mismatched objectives.



Common misunderstanding:
We always need interpretability.

https://cdn.theatlantic.com/assets/media/img/mt/2015/04/shutterstock_11926084/lead_large.jpg
https://www.jal.com/assets/img/flight/safety/equipment/pic_tcas_001_en.jpg

http://www.cinemablend.com/pop/Netflix-Using-Amazon-Cloud-Explore-Artificial-Intelligence-Movie-Recommendations-62248.html

Interpretability

?

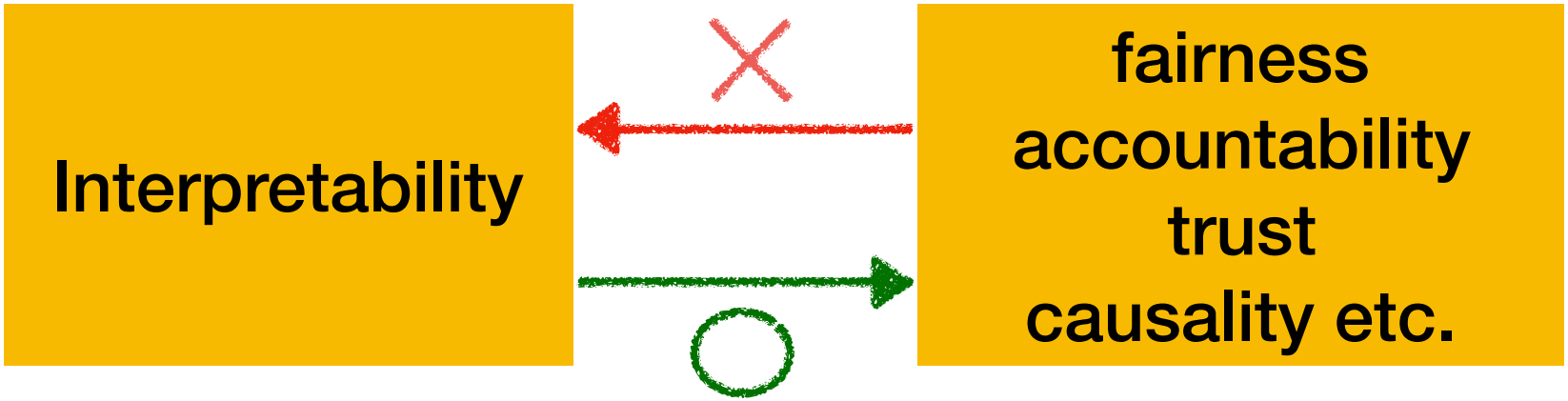
fairness
accountability
trust
causality etc.

Interpretability

fairness
accountability
trust
causality etc.

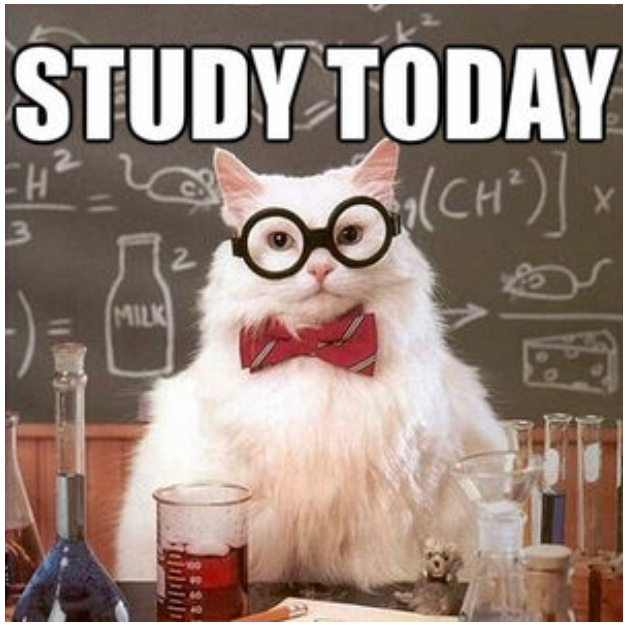
Common misunderstanding:
Trust, fairness and interpretability
are all the same thing.

Our cousins are not us



- Interpretability can help with them **when we cannot formalize these ideas**
- But once formalized, you may not need interpretability.

Let’s build some interpretability methods.



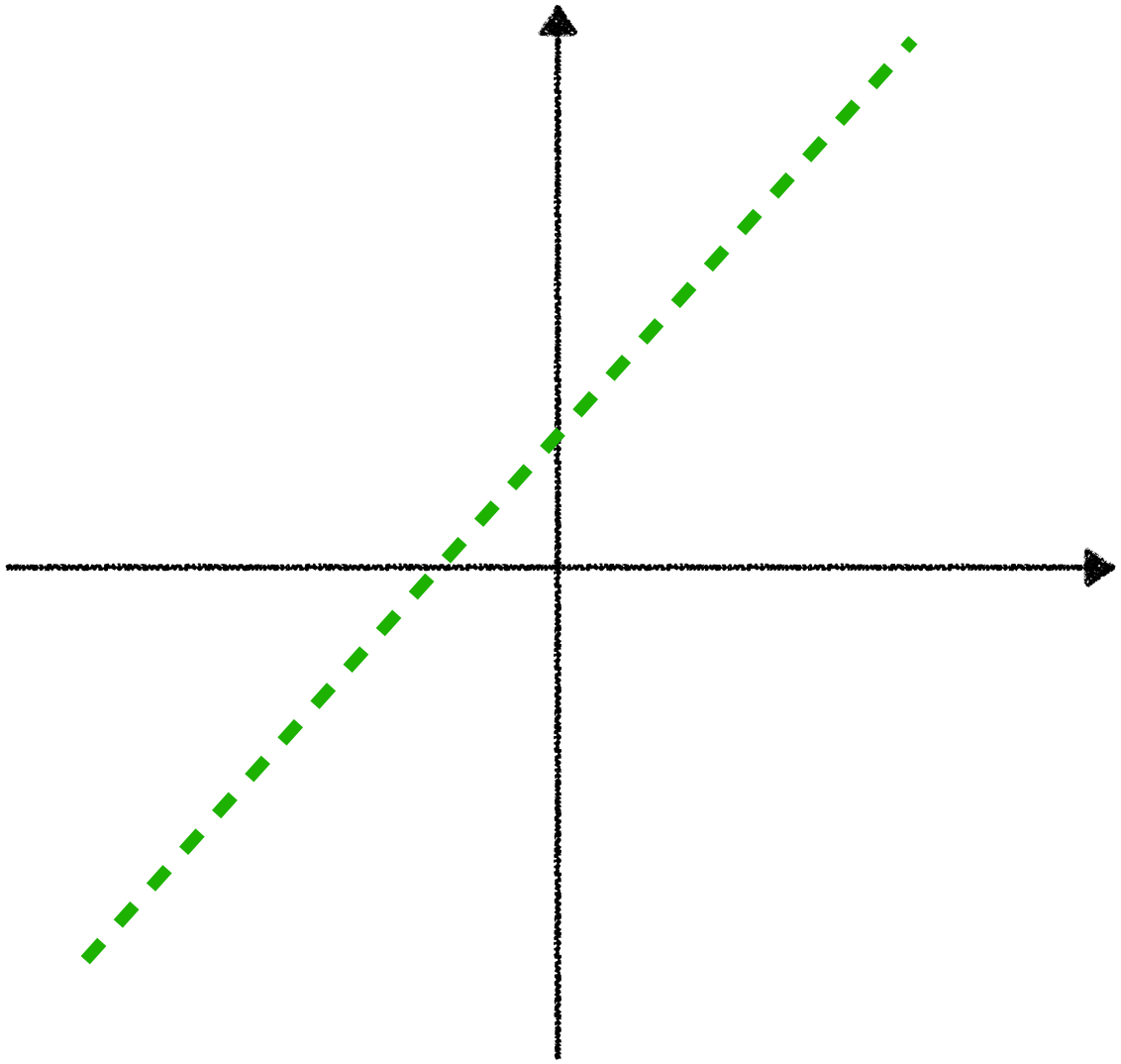
www.memecenter.com

ingredients for interpretability methods.

$$\underset{E}{\operatorname{argmax}}\ Q(E|\ ?)$$

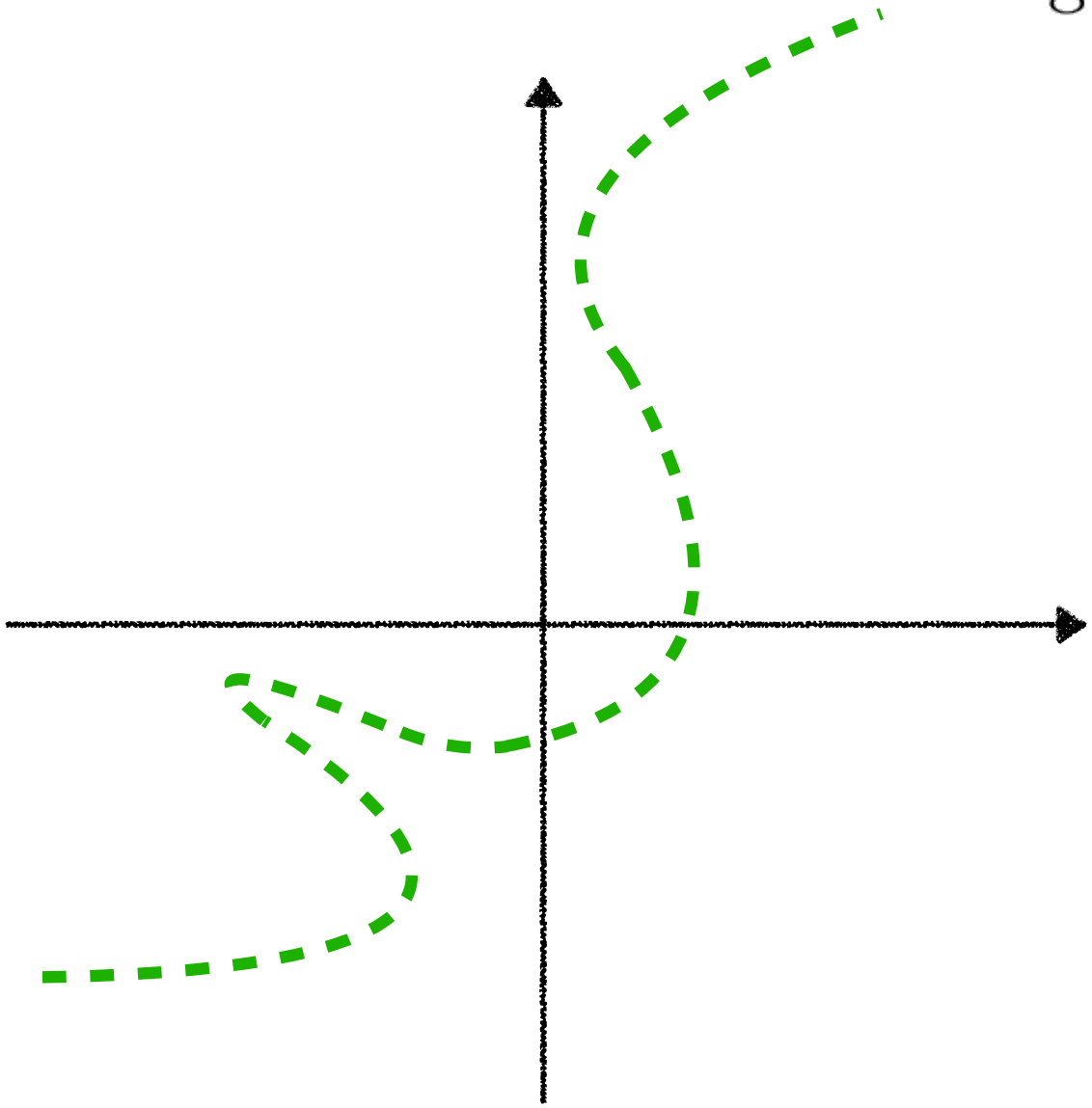
Some quality function

$$\underset{E}{\operatorname{argmax}}\ Q(E|M, \quad)$$



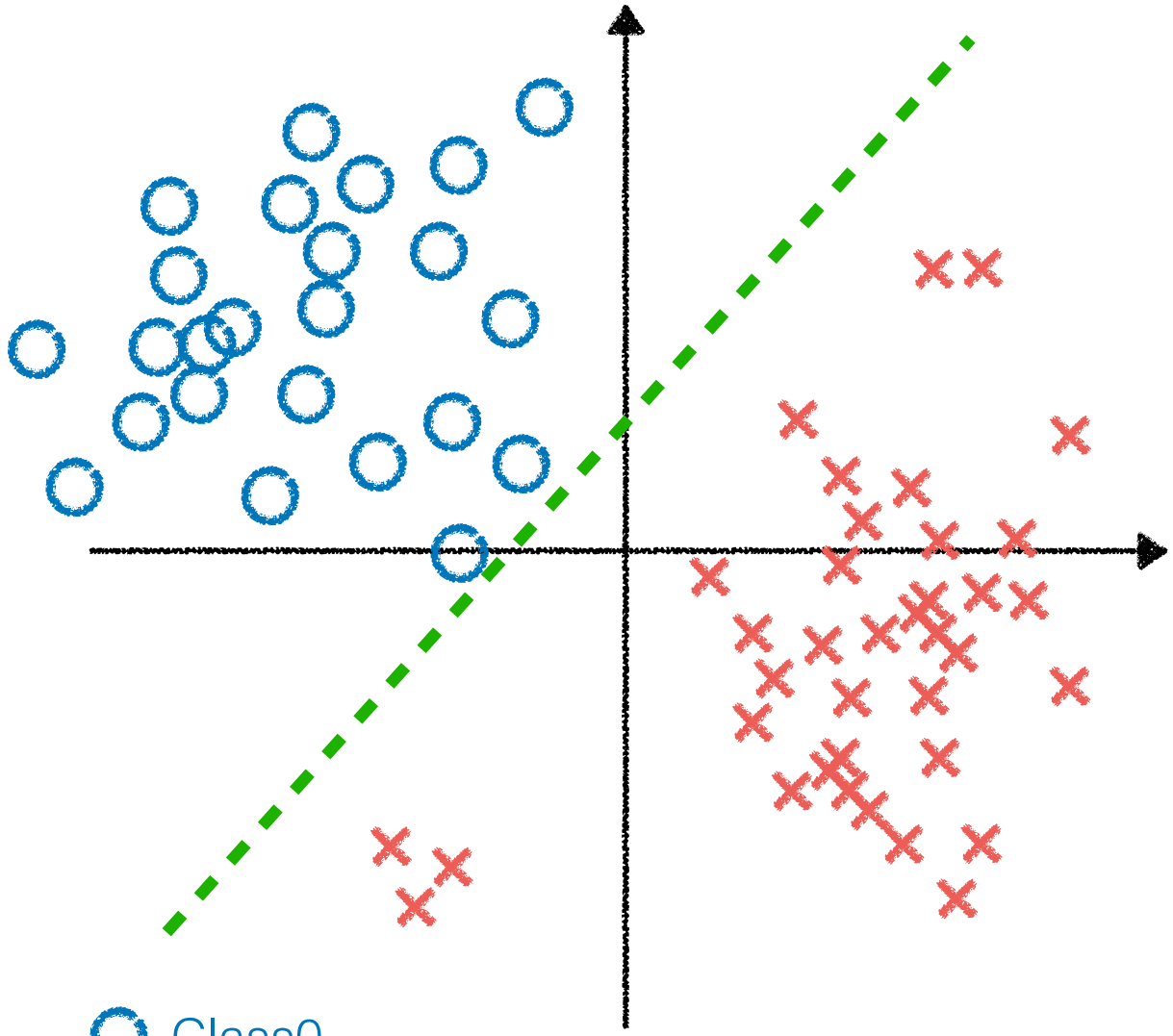
Model

$$\underset{E}{\operatorname{argmax}}\ Q(E|M, \quad)$$



Model

$$\underset{E}{\operatorname{argmax}}\ Q(E|M, \quad D, \quad)$$

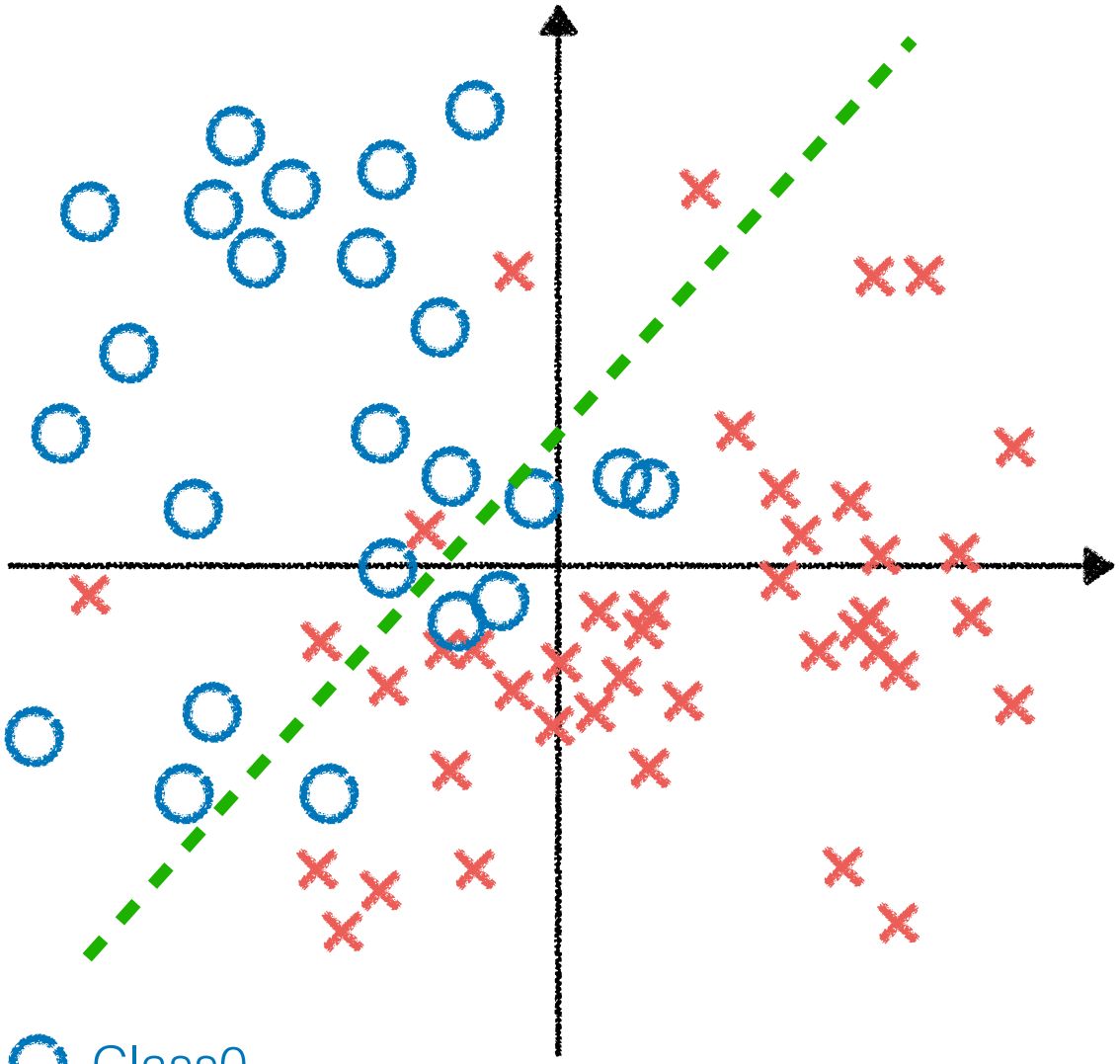


Data

Class0

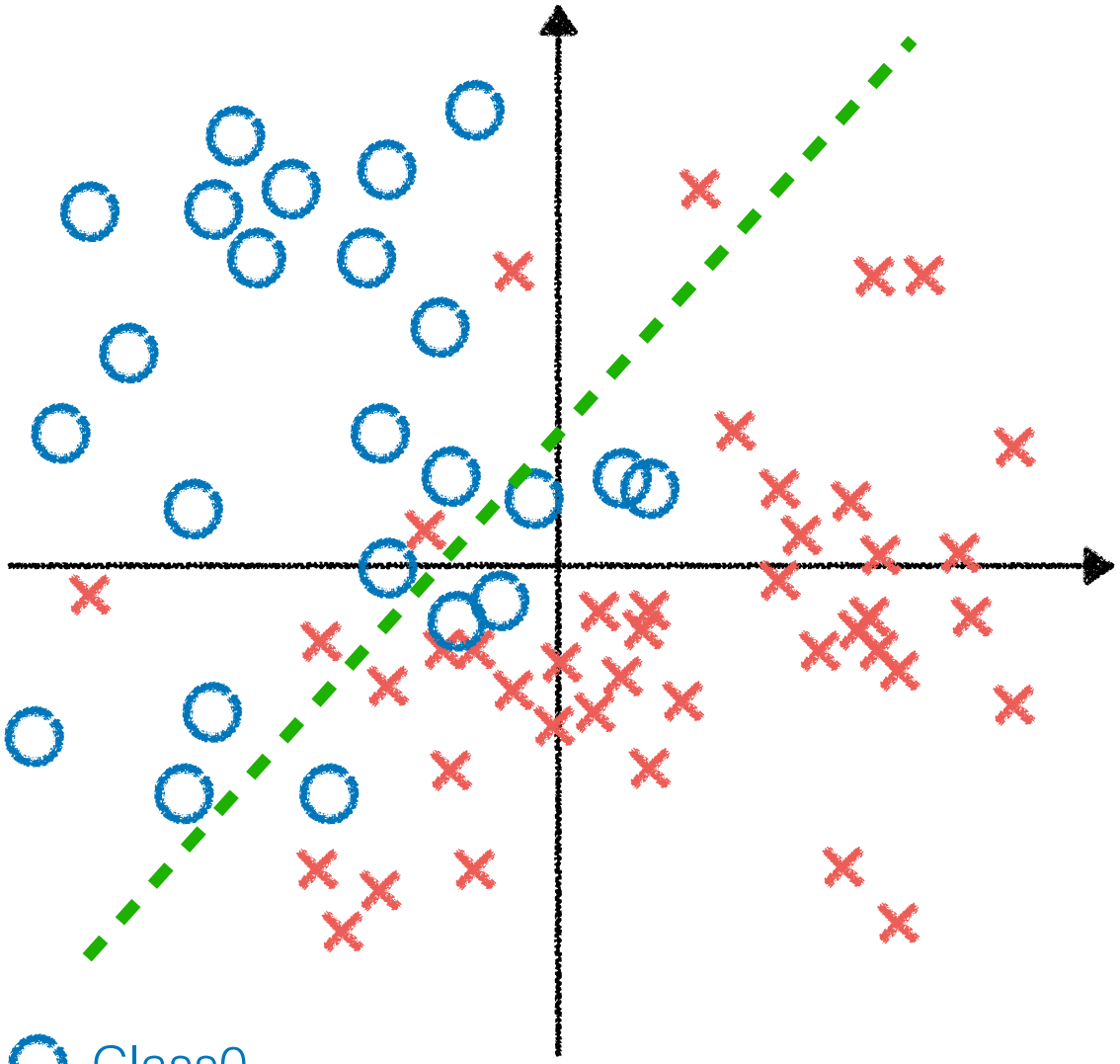
Class1

$$\underset{E}{\operatorname{argmax}}\ Q(E|M, \quad D, \quad)$$

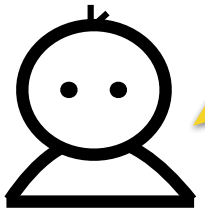


Data

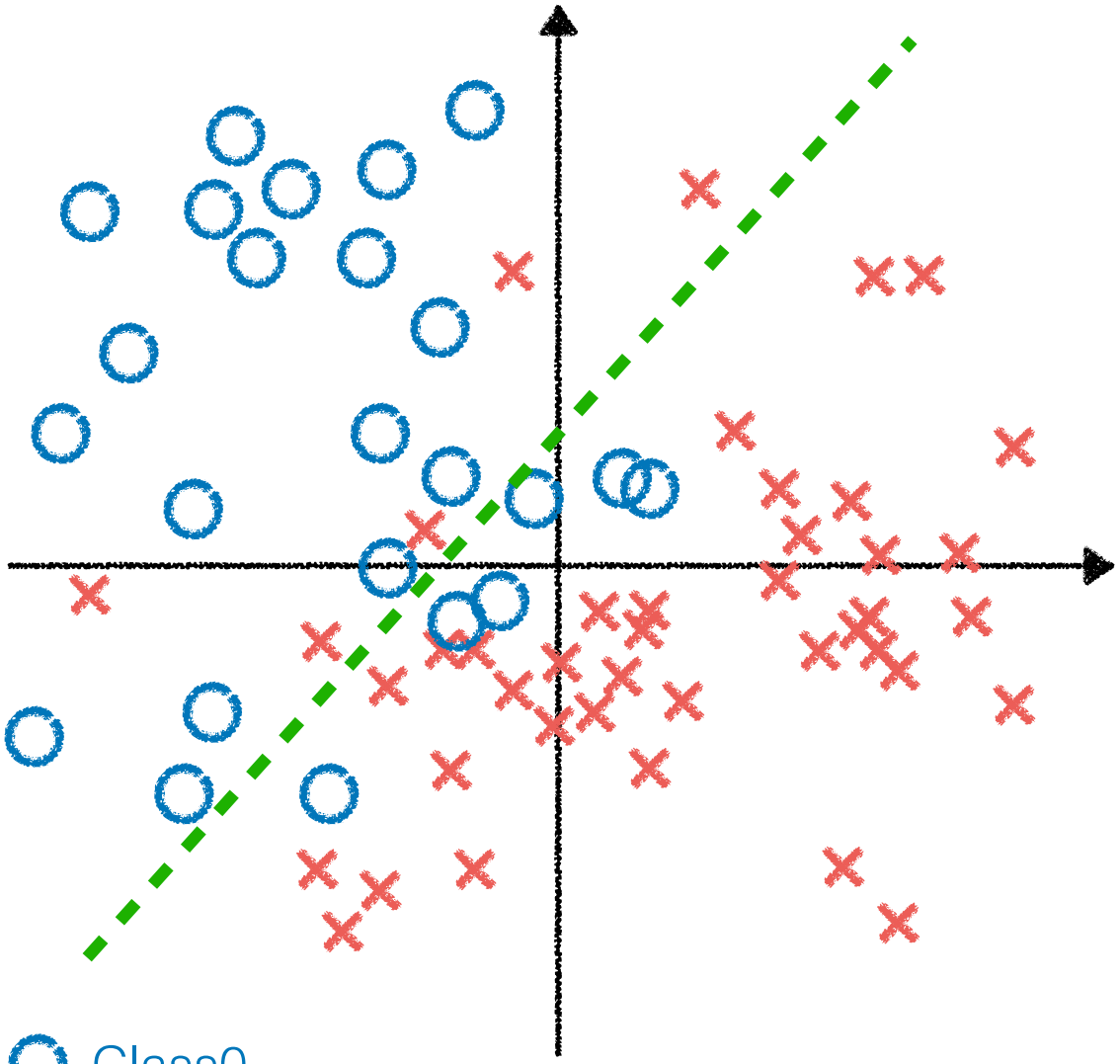
$$\underset{E}{\operatorname{argmax}}\, Q(E|M, H, D, \quad)$$



Human



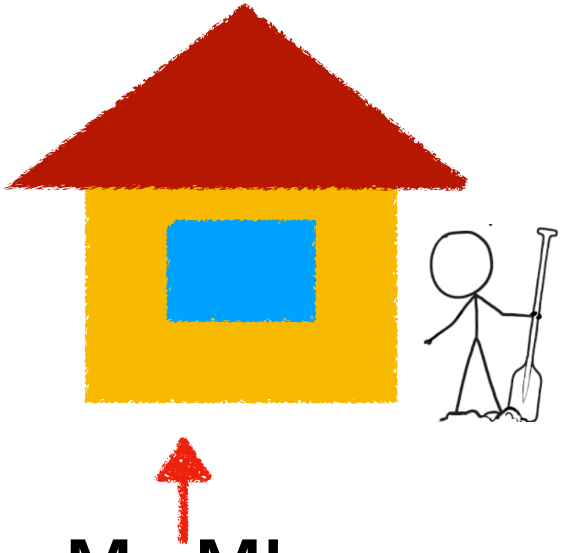
$$\underset{E}{\operatorname{argmax}}\, Q(E|M, H, D, T)$$



Task

- Local vs. global
- Simple explanations vs. more complex but more accurate explanations
- Low or high stake domains

Types of interpretability methods

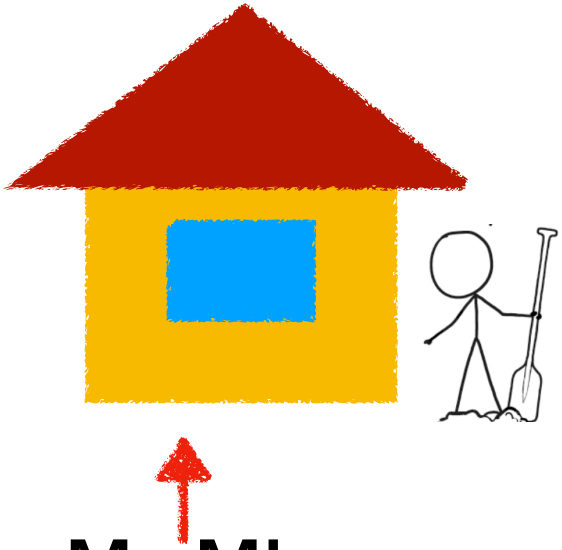


Post-training interpretability methods

Building inherently interpretable model

Explaining data

Types of interpretability methods



Post-training interpretability methods

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model}, \mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

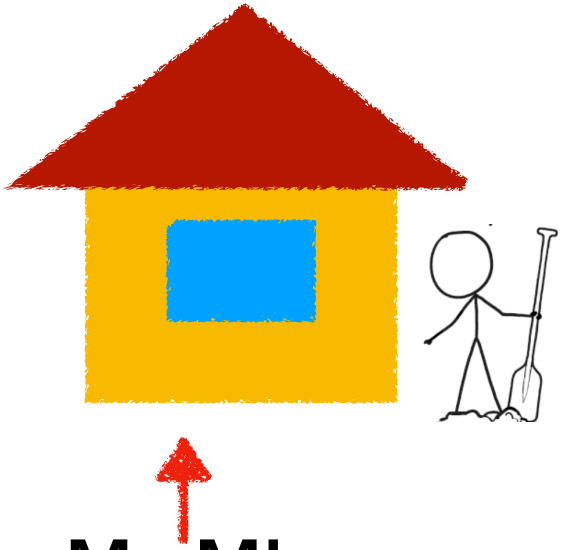
Building inherently interpretable model

$$\underset{E,M}{\text{argmax}}\ Q(\mathbf{Explanation}, \mathbf{Model}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Explaining data

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Types of interpretability methods



Post-training interpretability methods

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model}, \mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

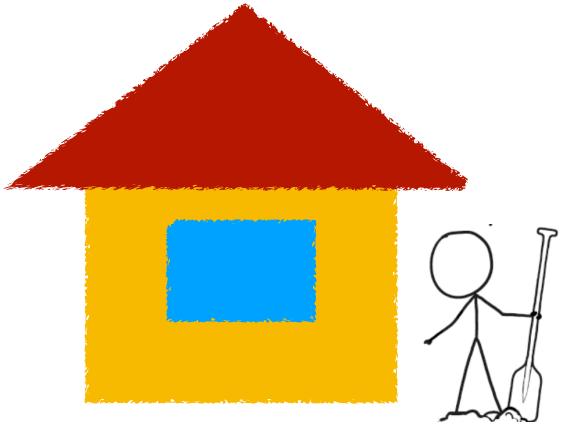
Building inherently interpretable model

$$\underset{E,M}{\text{argmax}}\ Q(\mathbf{Explanation}, \mathbf{Model}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Explaining data

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Types of interpretability methods



My ML

Post-training interpretability methods

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model}, \mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Building inherently interpretable model

rule, feature, example-based and many more.

$$\underset{E,M}{\text{argmax}}\ Q(\mathbf{Explanation}, \mathbf{Model}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Explaining data

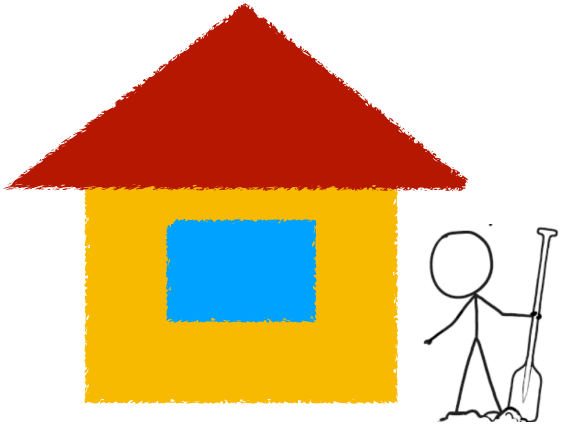
data visualization, exploratory data analysis

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Agenda

Post-training interpretability methods

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model}, \mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$



Building inherently interpretable model

rule, feature, example-based and many more.

$$\underset{E,M}{\text{argmax}}\ Q(\mathbf{Explanation}, \mathbf{Model}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

My ML

For more details
ICML tutorial 2017

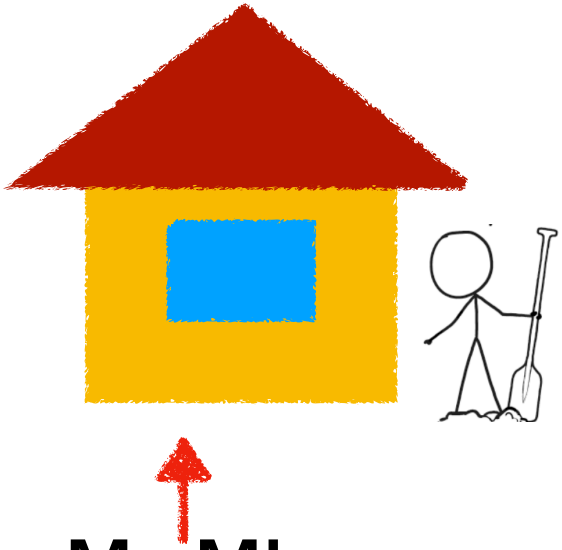
(on my website)

Explaining data

data visualization, exploratory data analysis

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Agenda



Post-training interpretability methods

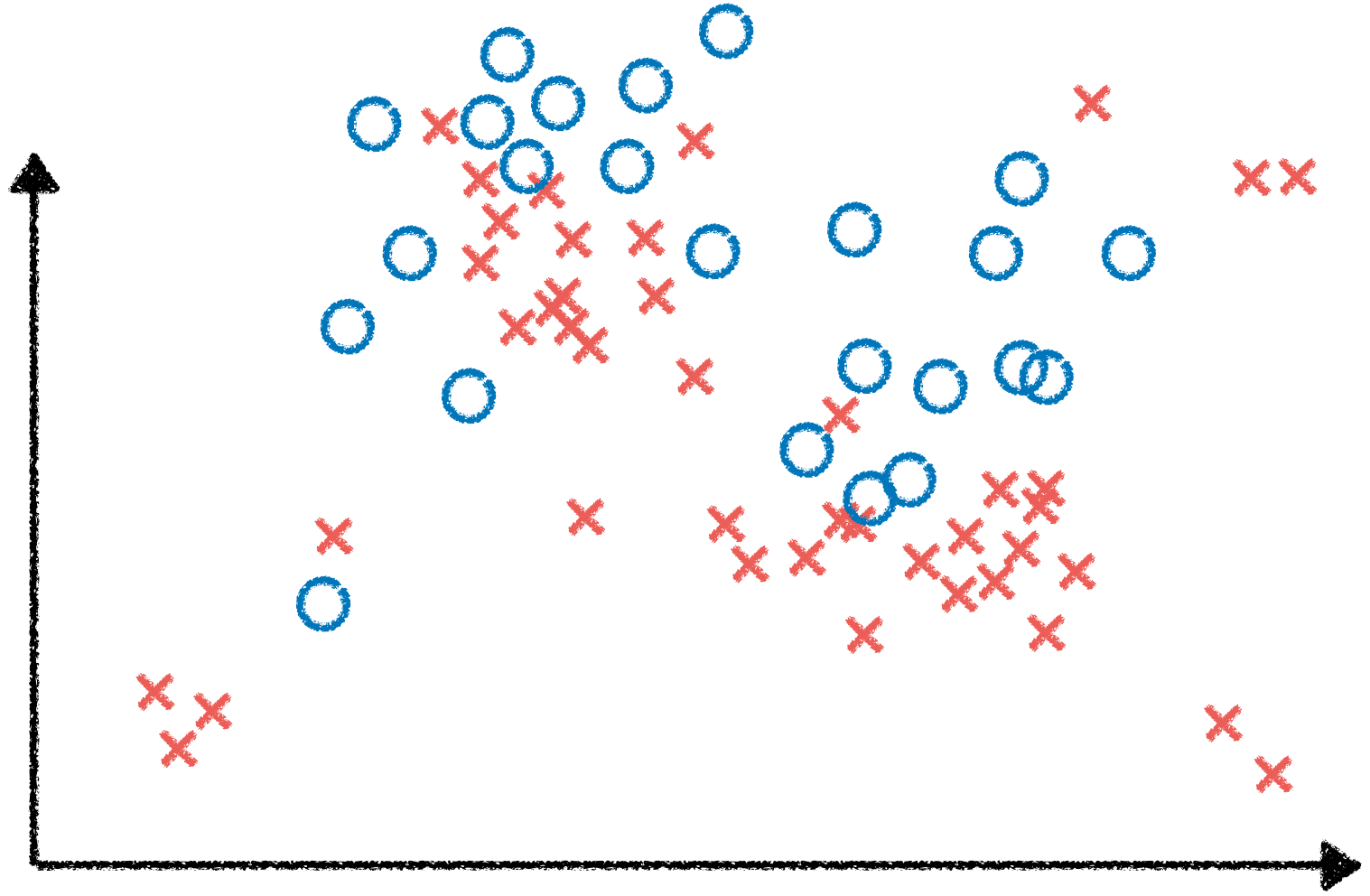
$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model}, \mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Building inherently interpretable model

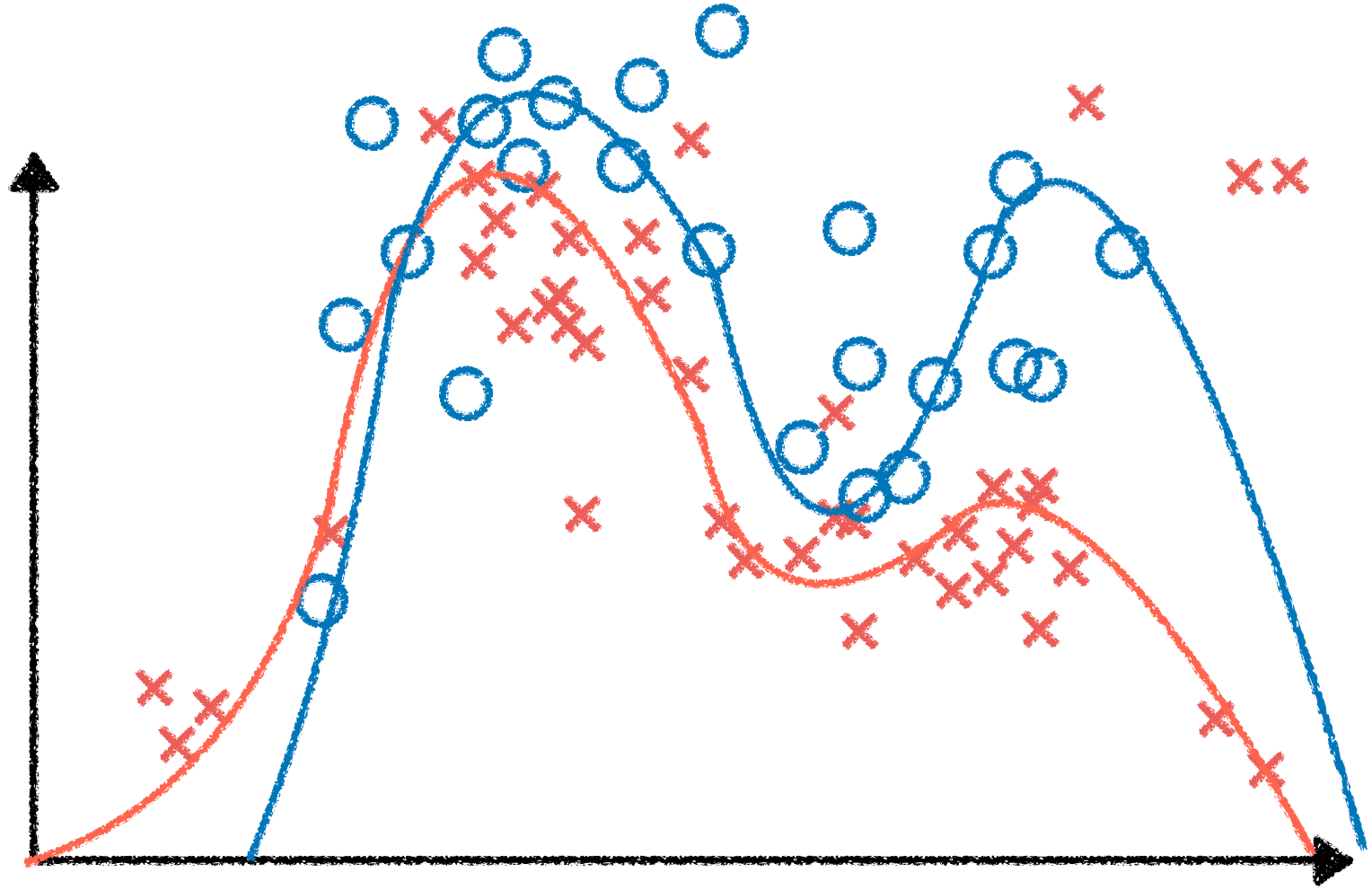
$$\underset{E,M}{\text{argmax}}\ Q(\mathbf{Explanation}, \mathbf{Model}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

Explaining data

$$\underset{E}{\text{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$


Class0
Class1

After building a model

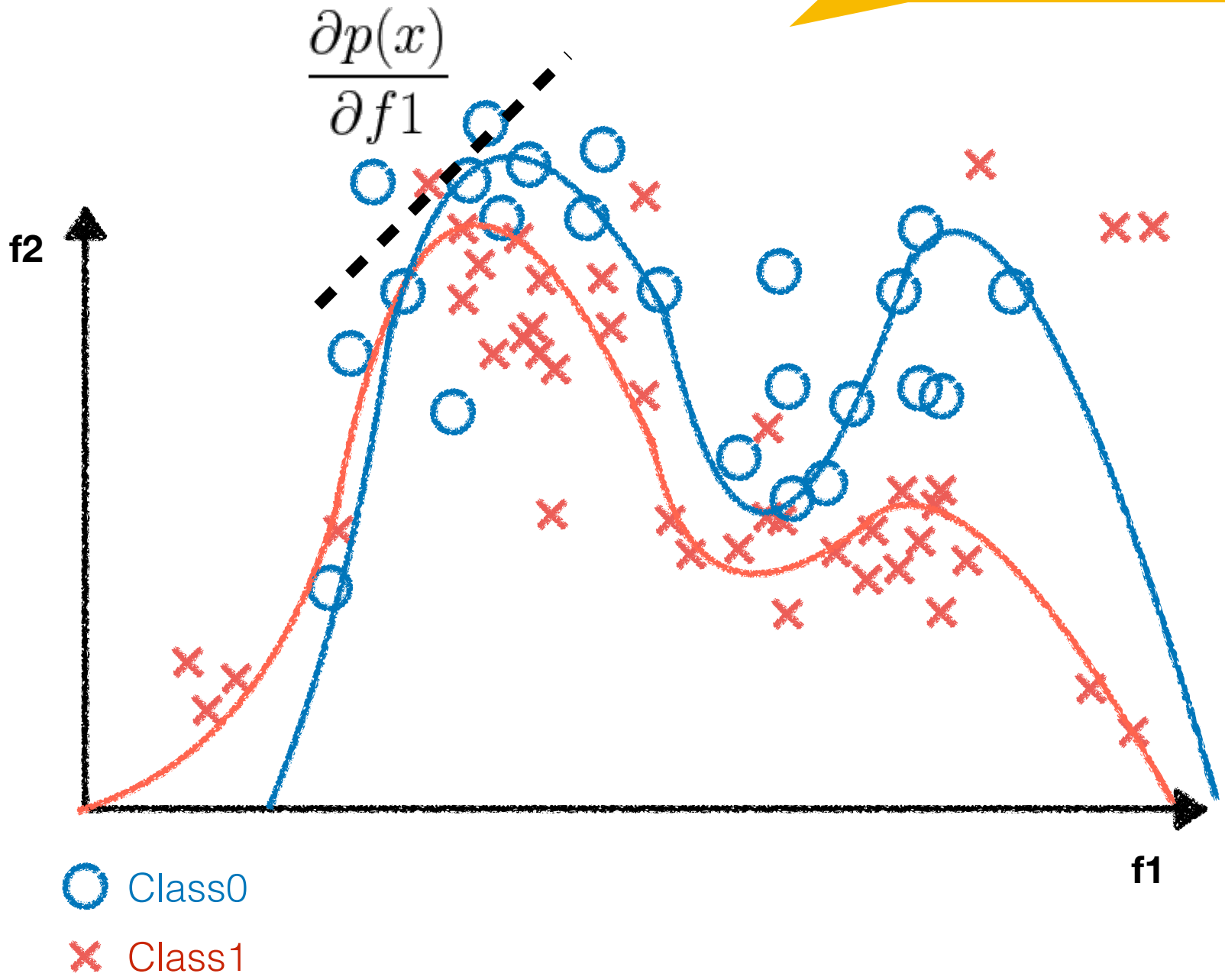


Class0

Class1

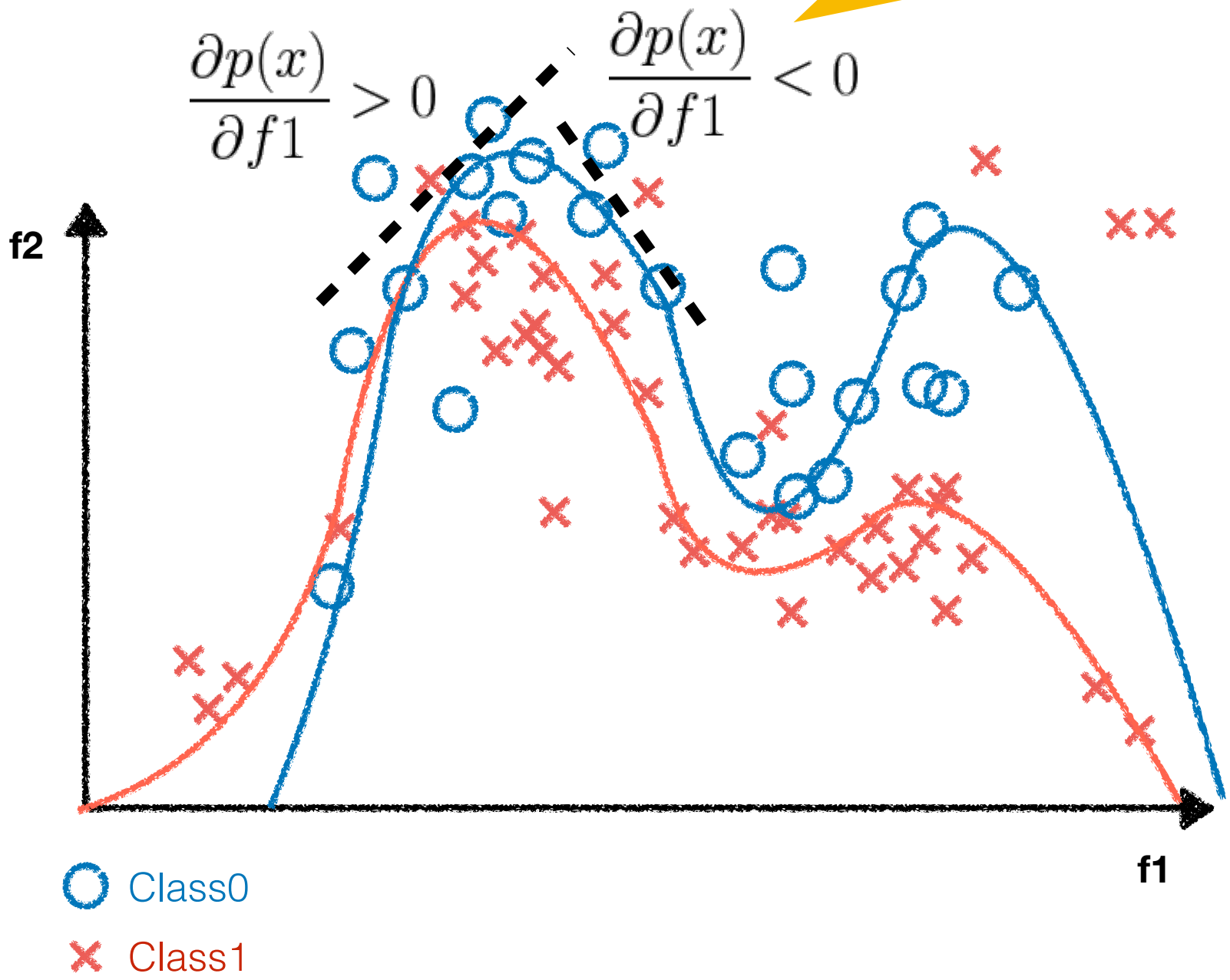
After building a model

Local explanations



After building a model

Local explanations



Problem:
Post-training explanation

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model}, \mathbf{Human}, \mathbf{Data}, \mathbf{Task})$$

A trained
machine learning model
(e.g., neural network)

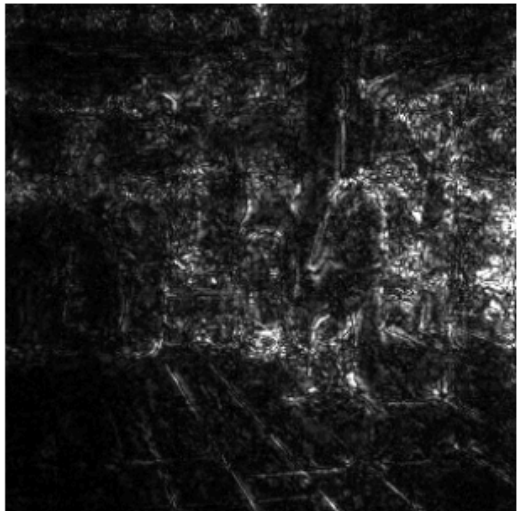
$p(z)$

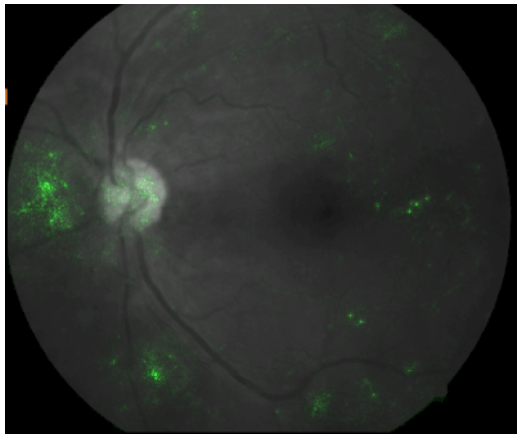
cash-machine-ness

Why was this a cash machine?

One of the most popular interpretability methods for images:

Saliency maps





picture credit: @sayres

Used for image classification and medical applications.

a logit → $\frac{\partial p(z)}{\partial x_{i,j}}$

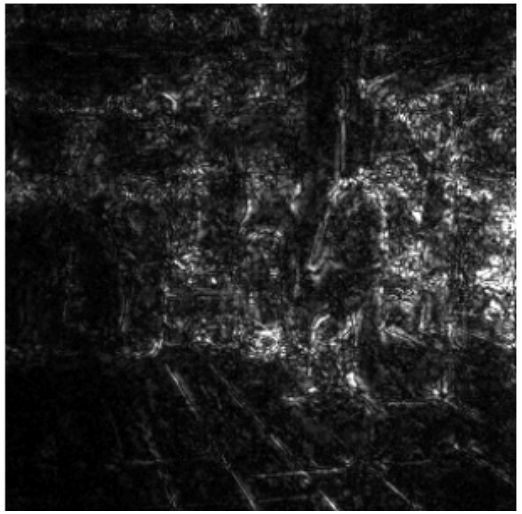
pixel i,j →

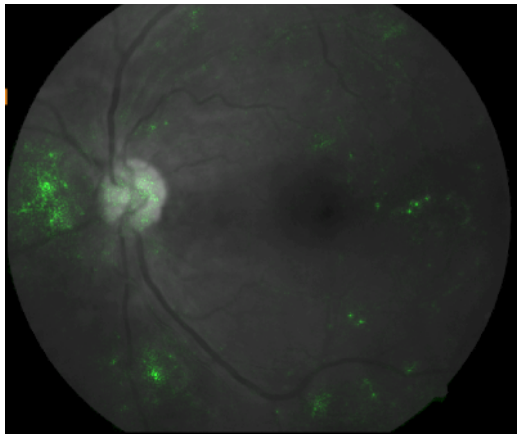
$$\underset{E}{\operatorname{argmax}}\ Q(E|M,H,D,T)$$

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg '17]
Integrated gradient [Sundararajan, Taly, Yan '17]

One of the most popular interpretability methods for images:

Saliency maps





picture credit: @sayres

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg '17]
Integrated gradient [Sundararajan, Taly, Yan '17]

Used for image classification and medical applications.

a logit → $\frac{\partial p(z)}{\partial x_{i,j}}$
pixel i,j →

$$\underset{E}{\operatorname{argmax}}\, Q(E|M, H, D, T)$$

NN

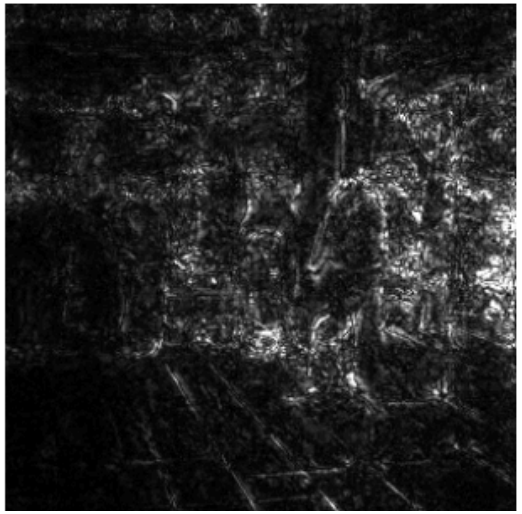
local undestanding

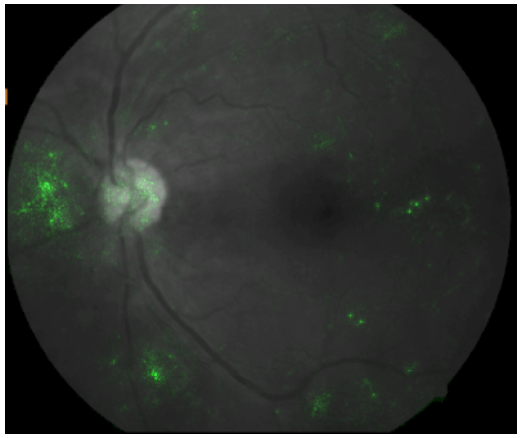
humans' subjective judgement

widely used for images

One of the most popular interpretability methods for images:

Saliency maps





Used for image classification and medical applications.

a logit → $\frac{\partial p(z)}{\partial x_{i,j}}$

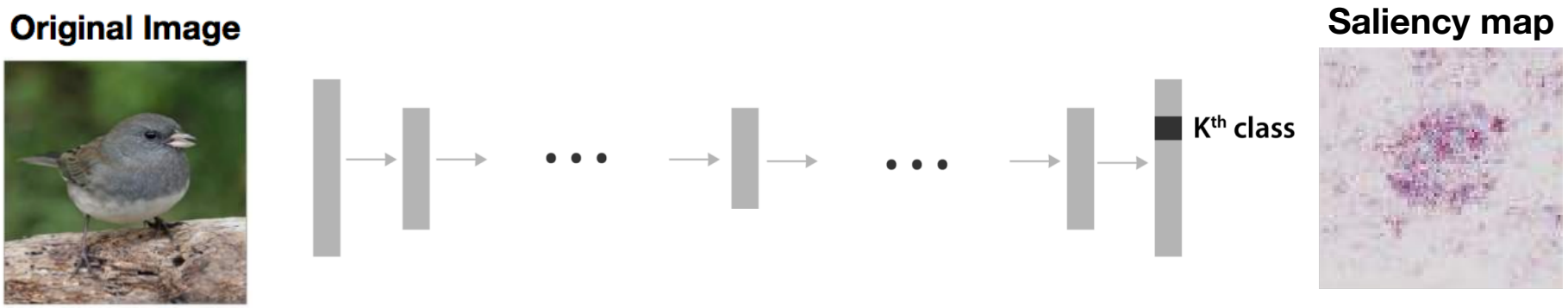
pixel i,j →

$$\underset{E}{\operatorname{argmax}}\ Q(E|M,H,D,T)$$

Sanity check:
If I change M a lot, will human perceive that E has changed a lot?

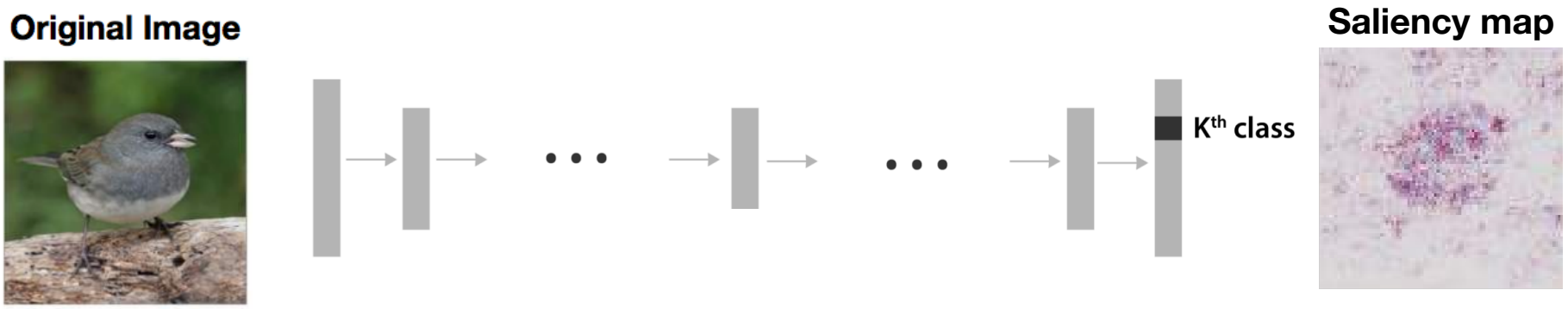
SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg '17]
Integrated gradient [Sundararajan, Taly, Yan '17]

Some confusing behaviors of saliency maps.

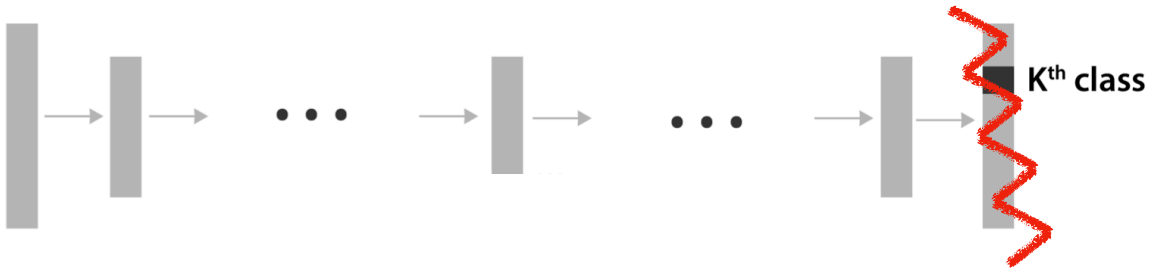


Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Some confusing behaviors of saliency maps.

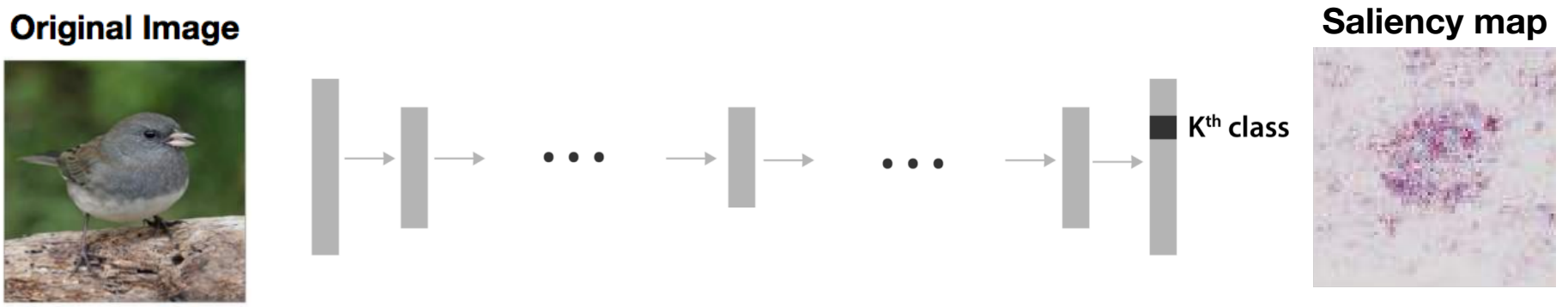


Randomized weights!
Network now makes garbage prediction.



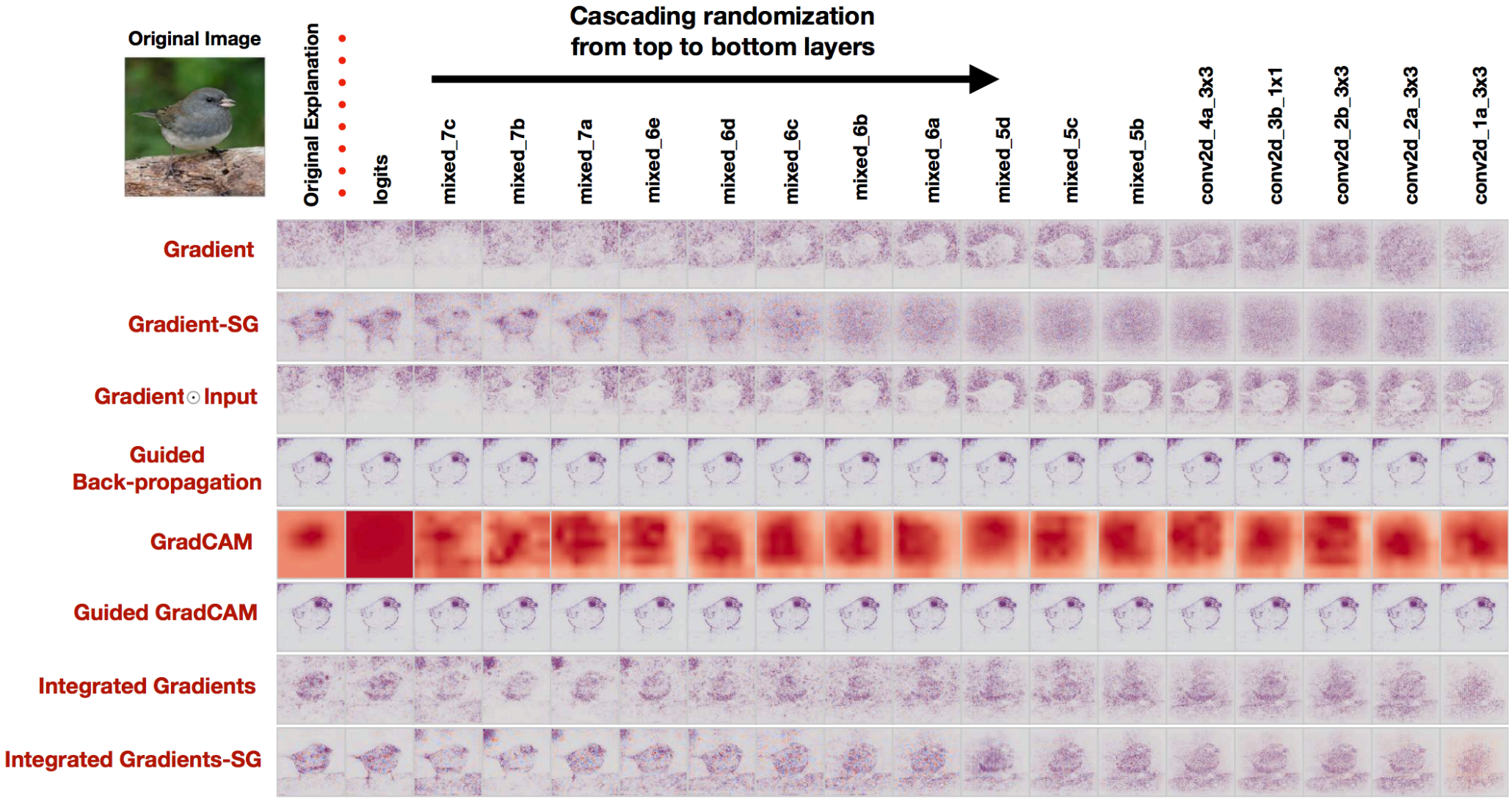
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Some confusing behaviors of saliency maps.



Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Some saliency maps look similar when we randomize the network.



Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

What can we learn from this?

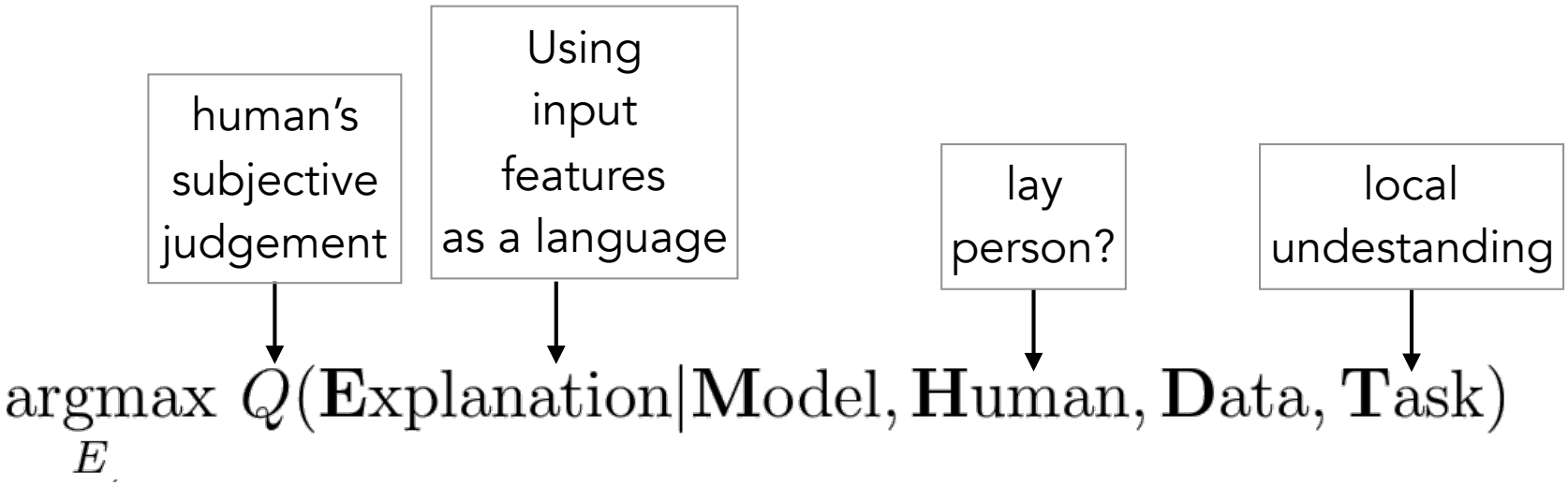
- Potential human confirmation bias: Just because it "makes sense" to humans, doesn't mean they reflect evidence for the prediction.

- Our discovery is consistent with other findings
  [Nie, Zhang, Patel '18] [Ulyanov, Vedaldi, Lempitsky '18]

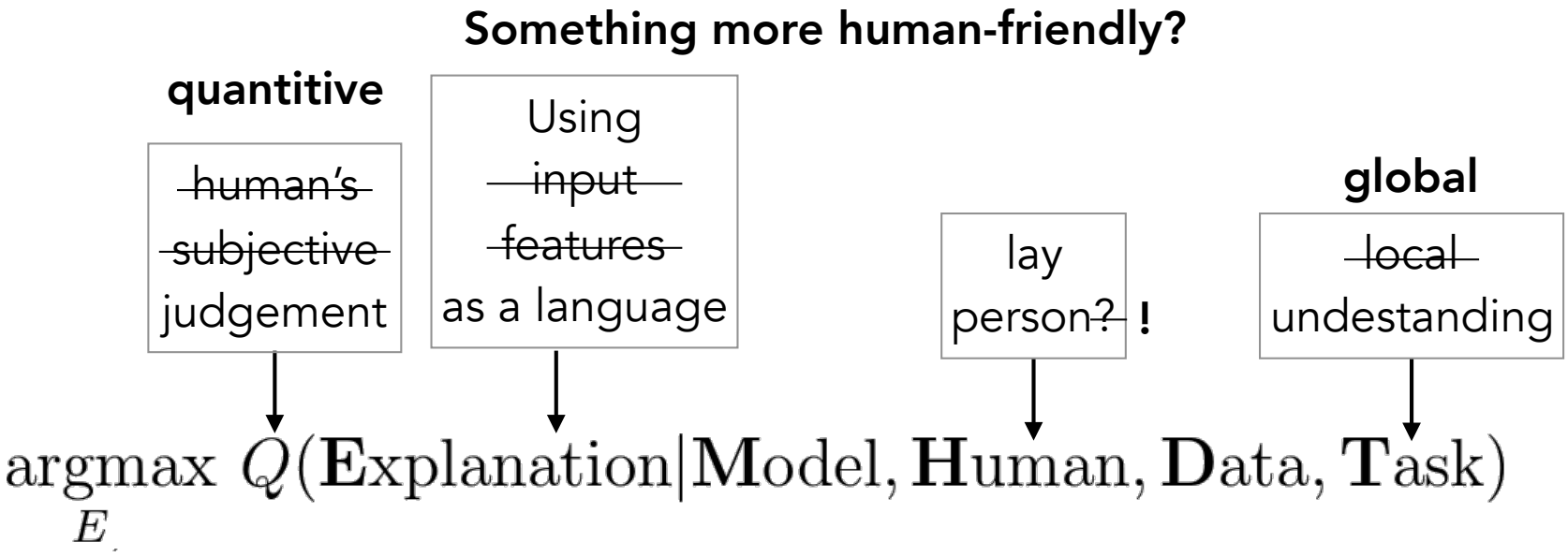
- Some of these methods have been shown to be useful in practice. Explaining predictions or features? More studies needed.

Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

What can we do better? Creating a wishlist.



What can we do better?
Creating a wishlist.

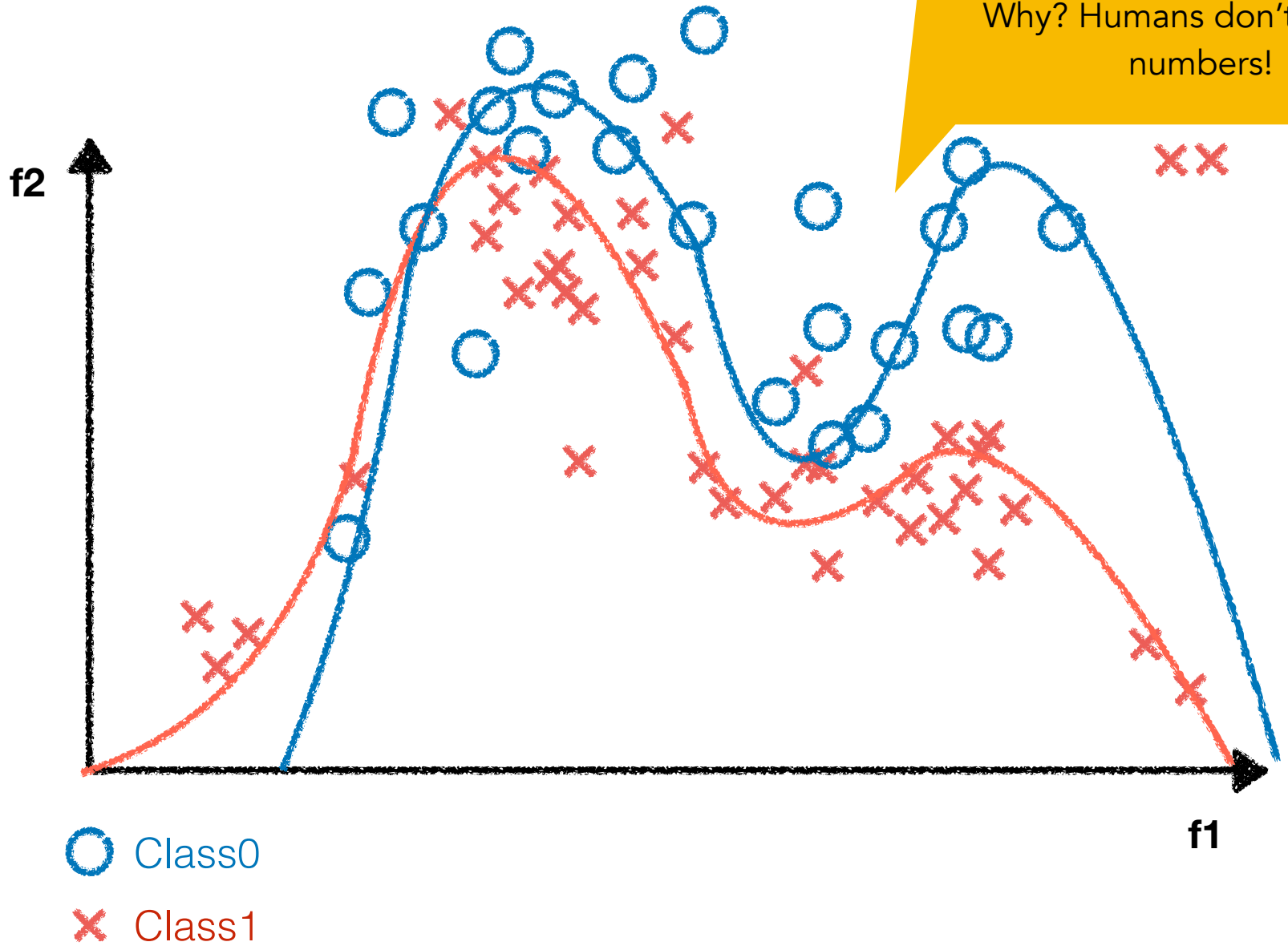


After building a model

$$\frac{\partial p(x)}{\partial \text{ concept}}$$

Concept-based:

Why? Humans don't think in numbers!

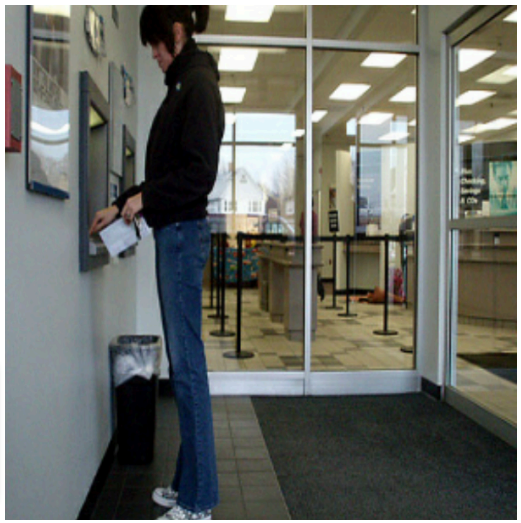


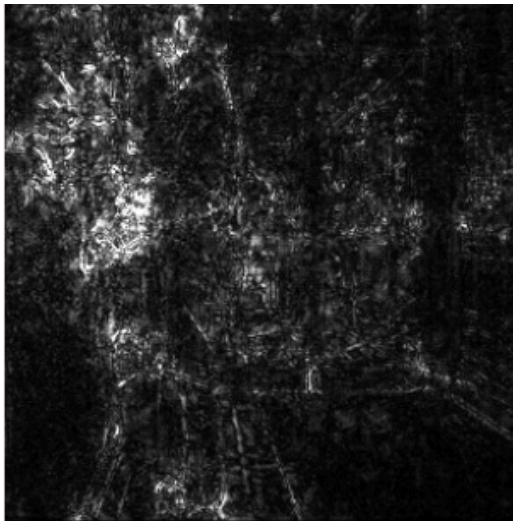
What we really want to ask…

prediction:
Cash machine

Why correct?
Why incorrect?

prediction:
Sliding door

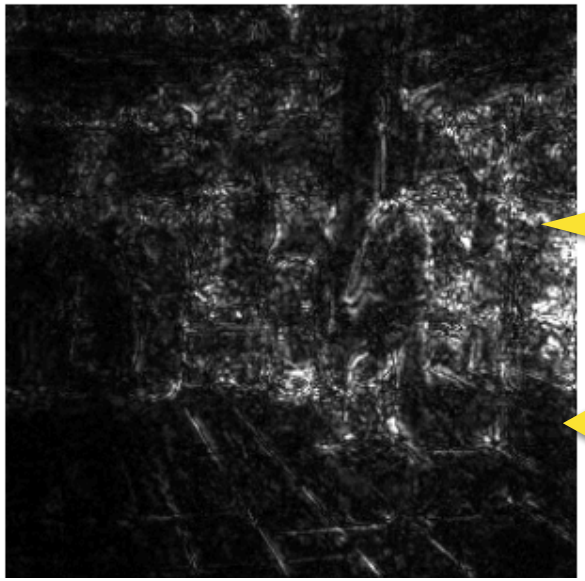


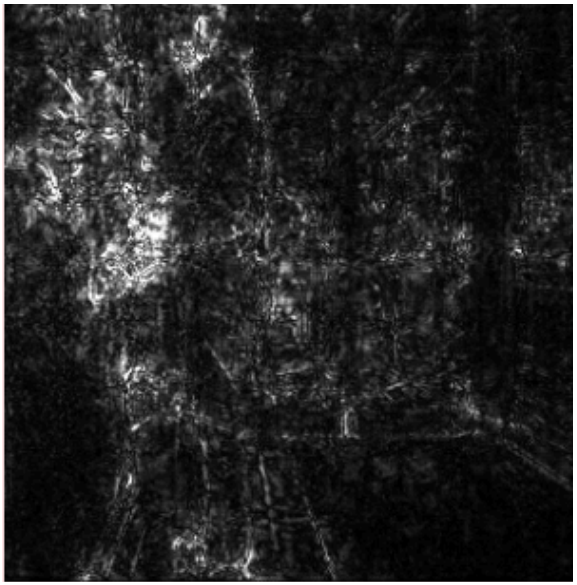


https://pair-code.github.io/saliency/

SmoothGrad [Smilkov, Thorat, K., Viégas, Wattenberg '17]

What we really want to ask…

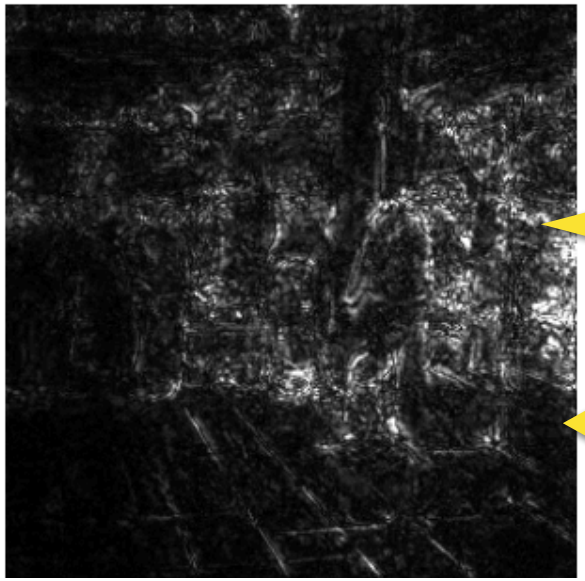


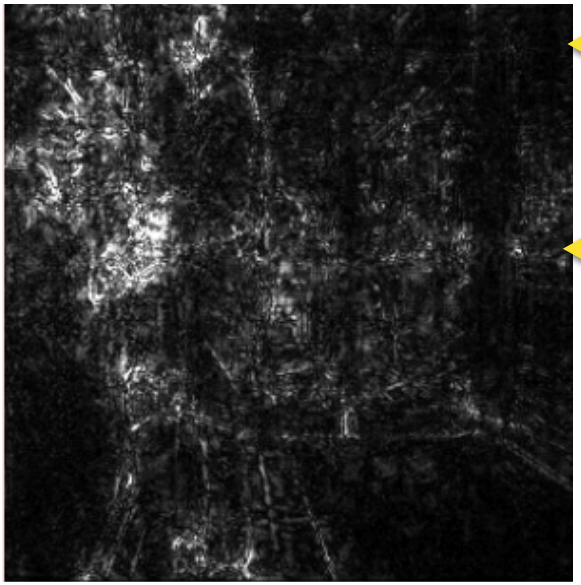


Were there more pixels on the cash machine than on the person?

Did the ‘human’ concept matter?
Did the ‘glasses’ or ‘paper’ matter?

What we really want to ask…





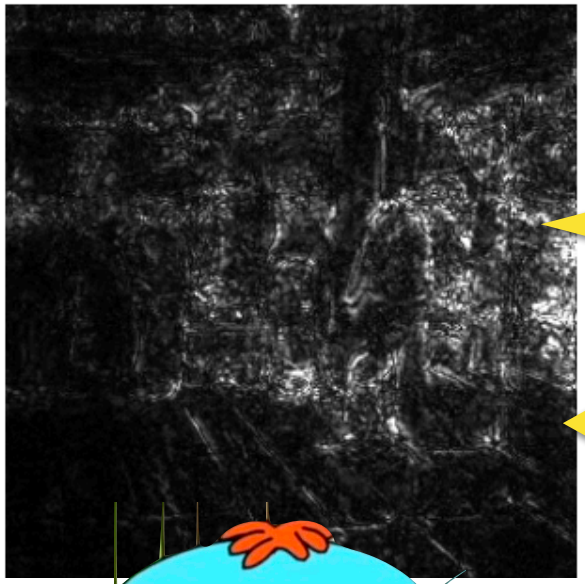
Were there more pixels on the cash machine than on the person?

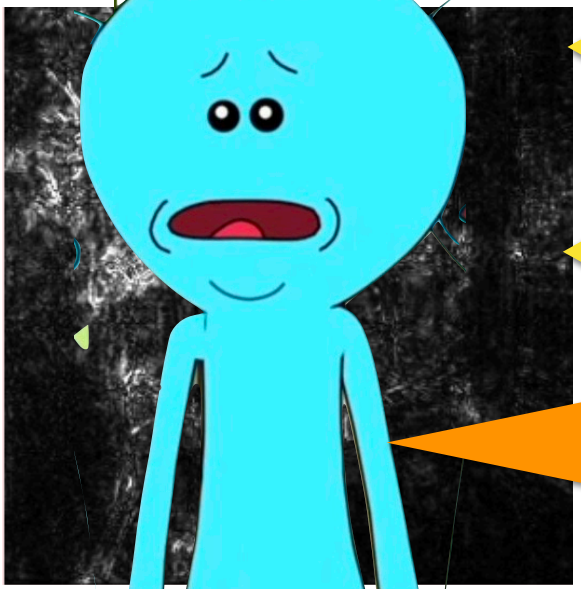
Did the ‘human’ concept matter? Did the ‘glasses’ or ‘paper’ matter?

Which concept mattered more?

Is this true for all other cash machine predictions?

What we really want to ask…





creative-analytics.corsairs.network/

Were there more pixels on the cash machine than on the person?

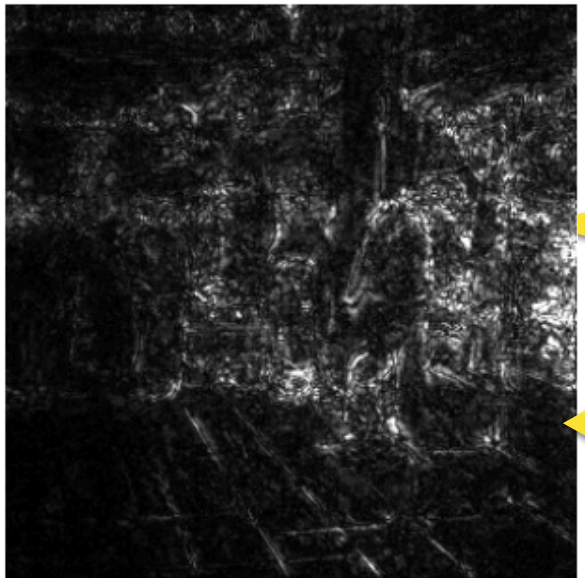
Did the ‘human’ concept matter?
Did the ‘glasses’ or ‘paper’ matter?

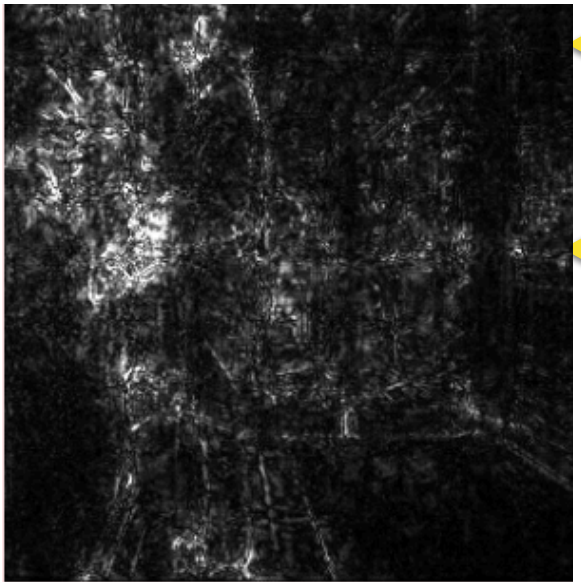
Which concept mattered more?

Is this true for all other cash machine predictions?

Oh no! I can’t express these concepts as pixels!!
They weren’t my input features either!

What we really want to ask…





Were there more pixels on the cash machine than on the person?

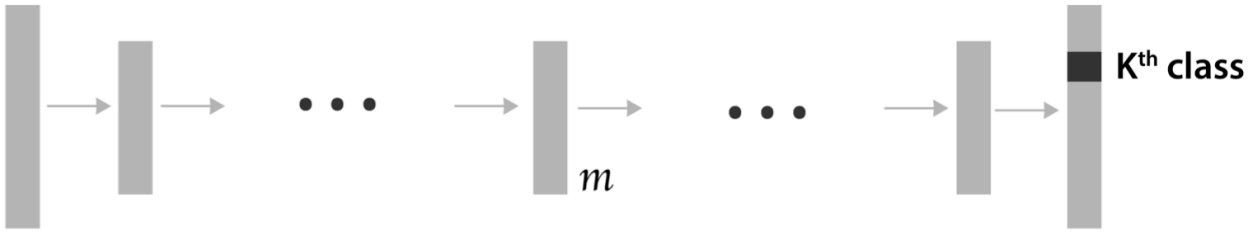
Did the ‘human’ concept matter? Did the ‘glasses’ or ‘paper’ matter?

Which concept mattered more?

Is this true for all other cash machine predictions?

Wouldn’t it be great if we can **quantitatively** measure how important *any* of these **user-chosen concepts** are?

Goal of TCAV:
Testing with Concept Activation Vectors



Quantitative explanation: how much a concept (e.g., gender, race) was important for a prediction in a trained model.

…even if the concept was not part of the training.

Goal of TCAV:
Testing with Concept Activation Vectors

Goal of TCAV:
Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network)

$p(z)$

Doctor-ness

Was gender concept important to this doctor image classifier?

Goal of TCAV:
Testing with Concept Activation Vectors

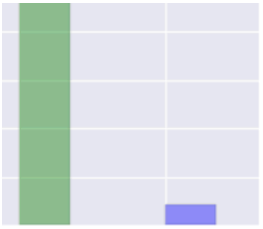
Was gender concept important to this doctor image classifier?

A trained machine learning model (e.g., neural network)

$p(z)$

Doctor-ness

TCAV score for **Doctor**



Goal of TCAV:
Testing with Concept Activation Vectors

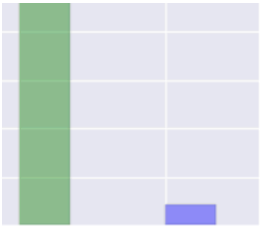
A trained machine learning model (e.g., neural network)

$p(z)$

Doctor-ness

Was gender concept important to this doctor image classifier?

TCAV score for **Doctor**



TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

Goal of TCAV:
Testing with Concept Activation Vectors

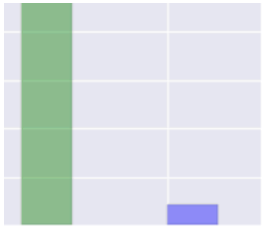
A trained
machine learning model
(e.g., neural network)

$p(z)$

zebra-ness

Was striped concept important to this zebra image classifier?

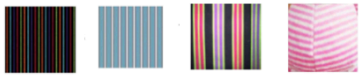
TCAV score for **Zebra**



striped **not striped**

TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

TCAV:
Testing with Concept Activation Vectors



A trained machine learning model (e.g., neural network)

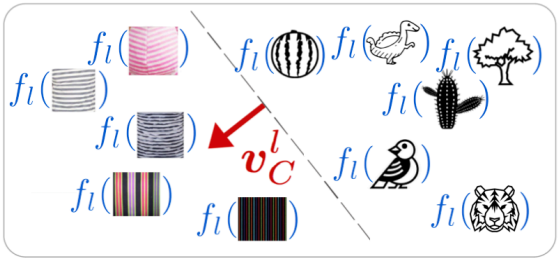
$p(z)$

zebra-ness

TCAV

Was striped concept important to this zebra image classifier?

1. Learning CAVs



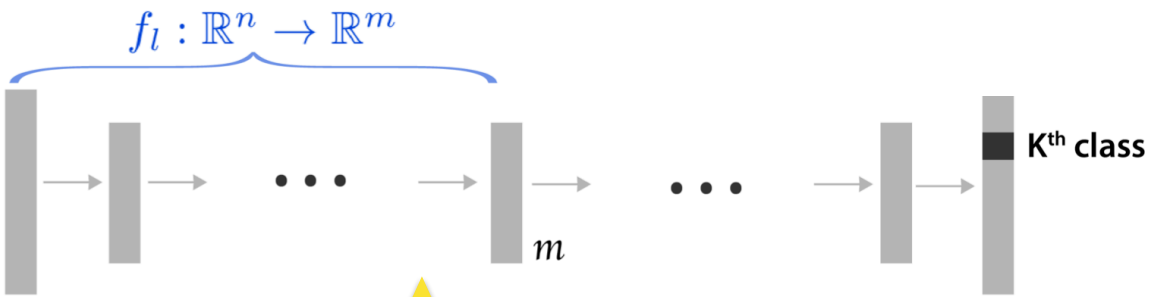
1. How to define concepts?

Defining concept activation vector (CAV)

Inputs:

Examples of concepts

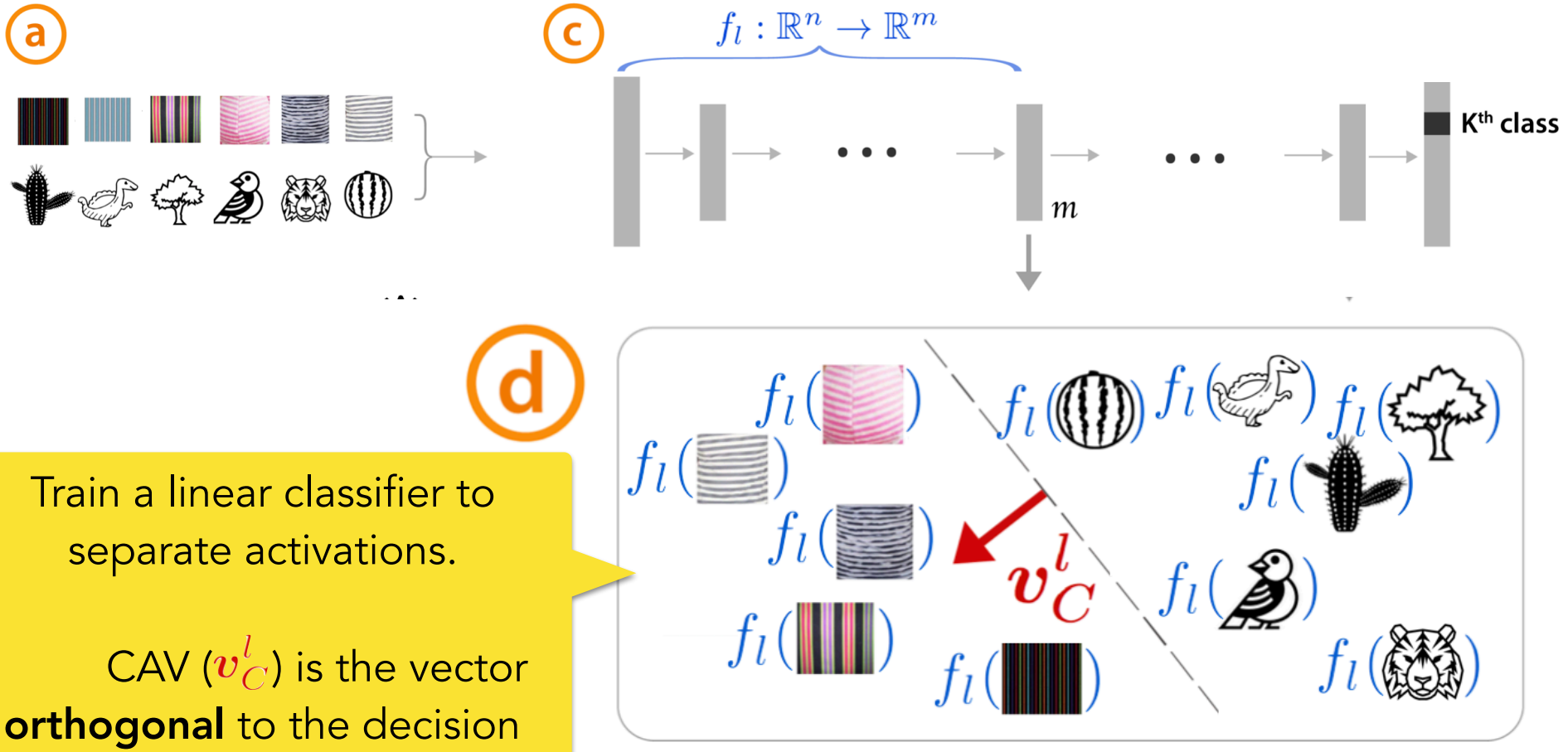
Random images



A trained network under investigation
and
Internal tensors

Defining concept activation vector (CAV)

Inputs:



Train a linear classifier to separate activations.

CAV ($\boldsymbol{v}_C^l$) is the vector **orthogonal** to the decision boundary.

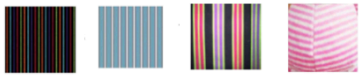
[Smilkov '17, Bolukbasi '16 , Schmidt '15]

TCAV:
Testing with Concept Activation Vectors

A trained machine learning model (e.g., neural network)

$p(z)$

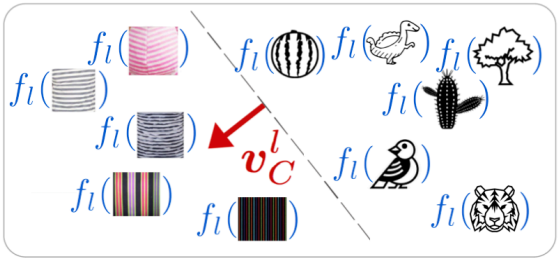
zebra-ness



Was striped concept important to this zebra image classifier?

TCAV

1. Learning CAVs



2. Getting TCAV score

$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

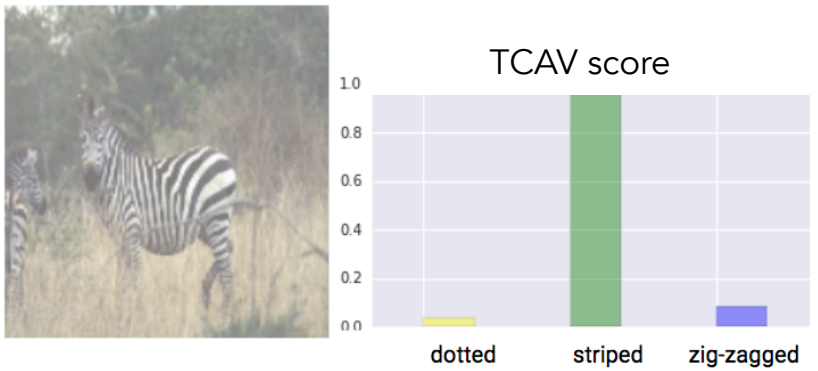
$S_{C,k,l}(\ )$

$\text{TCAV}_{Q_{C,k,l}}$

2. How are the CAVs useful to get explanations?

TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV



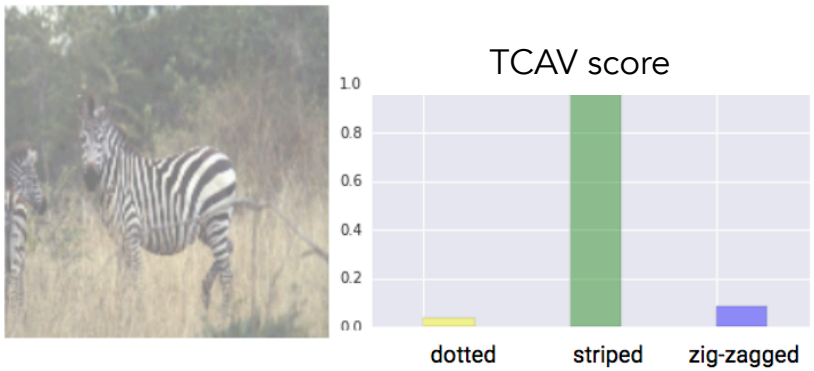
zebra-ness → $\frac{\partial p(z)}{\partial \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$

striped CAV →

Directional derivative with CAV

TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV



$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

$S_{C,k,l}($  $)$

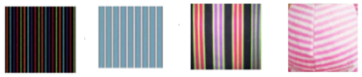
zebra-ness → $\dfrac{\partial p(z)}{\partial \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$

striped CAV →

$$\text{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

Directional derivative with CAV

TCAV:
Testing with Concept Activation Vectors



Was striped concept important to this zebra image classifier?

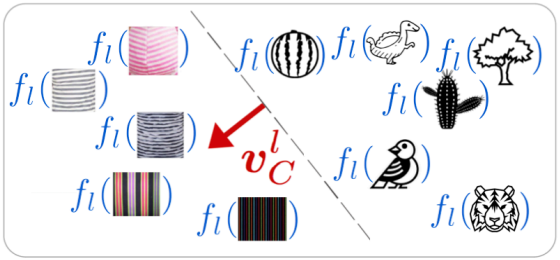
A trained machine learning model (e.g., neural network)

$p(z)$

zebra-ness

TCAV

1. Learning CAVs



2. Getting TCAV score

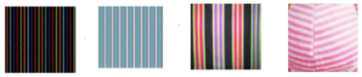
$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

$\text{TCAV}_{Q_{C,k,l}}$

TCAV:
Testing with Concept Activation Vectors



Was striped concept important to this zebra image classifier?

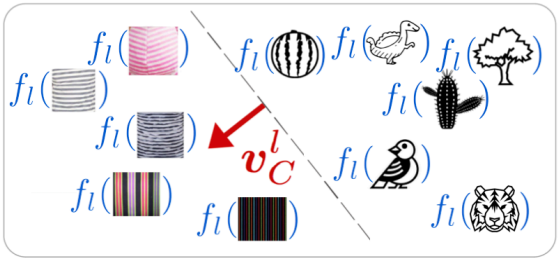
A trained machine learning model (e.g., neural network)

$p(z)$

zebra-ness

TCAV

1. Learning CAVs



2. Getting TCAV score

$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

$\text{TCAVQ}_{C,k,l}$

3. CAV validation

Qualitative

Quantitative

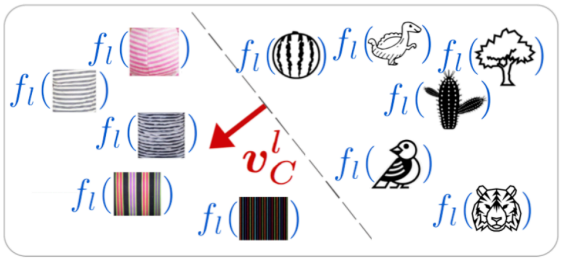
Quantitative validation:

Guarding against spurious CAV

Did my CAVs returned high sensitivity by chance?

Quantitative validation:

Guarding against spurious CAV



Learn many stripes CAVs using different sets of random images

Quantitative validation:

Guarding against spurious CAV

Zebra

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

Quantitative validation:

Guarding against spurious CAV

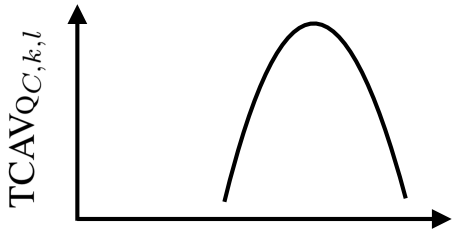
Zebra

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

$\longrightarrow \text{TCAVQ}_{C,k,l}$:

$\longrightarrow \text{TCAVQ}_{C,k,l}$:



Quantitative validation:

Guarding against spurious CAV

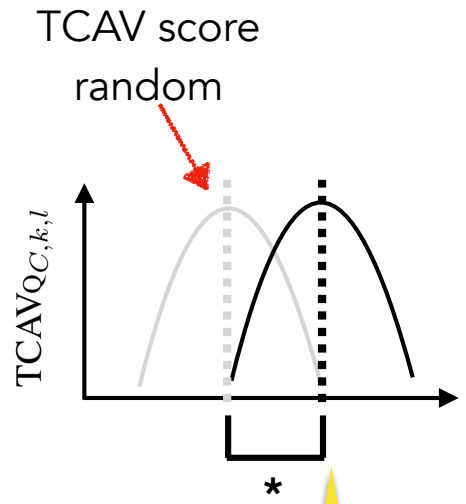
Zebra

$\rightarrow \text{TCAVQ}_{C,k,l}$:

$\rightarrow \text{TCAVQ}_{C,k,l}$:

$\rightarrow \text{TCAVQ}_{C,k,l}$:

$\rightarrow \text{TCAVQ}_{C,k,l}$:



Check the distribution of $\text{TCAVQ}_{C,k,l}$ is statistically different from random using t-test

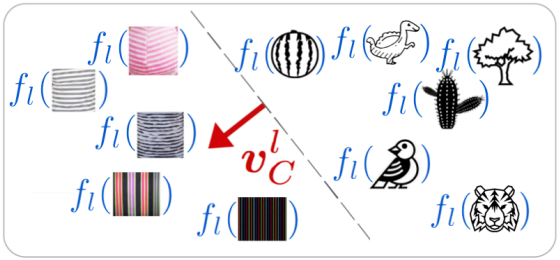
Recap TCAV:
Testing with Concept Activation Vectors

TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

Even if your training data wasn't tagged with the concept

Even if your input feature did not include the concept

1. Learning CAVs



2. Getting TCAV score

$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

$S_{C,k,l}(\ )$

$\rightarrow \text{TCAVQ}_{C,k,l}$

3. CAV validation

Qualitative

Quantitative

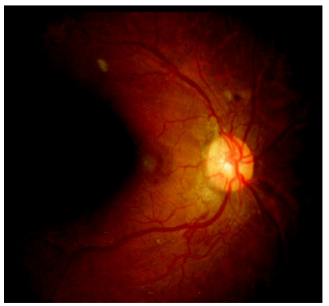
Results

1. Sanity check experiment

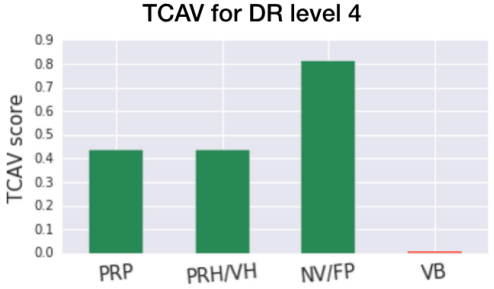
cab image cab image with caption

2. Biases in Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy



DR level 4 Retina



Results

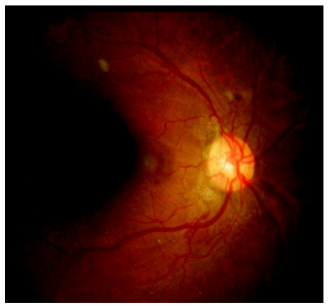
1. Sanity check experiment

cab image cab image with caption

2. Biases from Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy

DR level 4 Retina



TCAV for DR level 4

Sanity check experiment

If we know the ground truth
(important concepts),
will TCAV match?

Sanity check experiment setup

An image

+

Potentially noisy Caption

Sanity check experiment setup

image concept

caption concept

An image

+

Potentially noisy Caption

models can use either image or caption concept for classification.

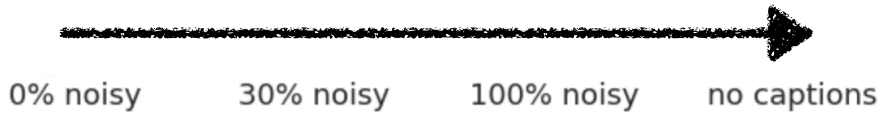
Sanity check experiment setup

image concept

caption concept

An image
+
Potentially noisy Caption

models can use either image or caption concept for classification.



Caption noise level in training set

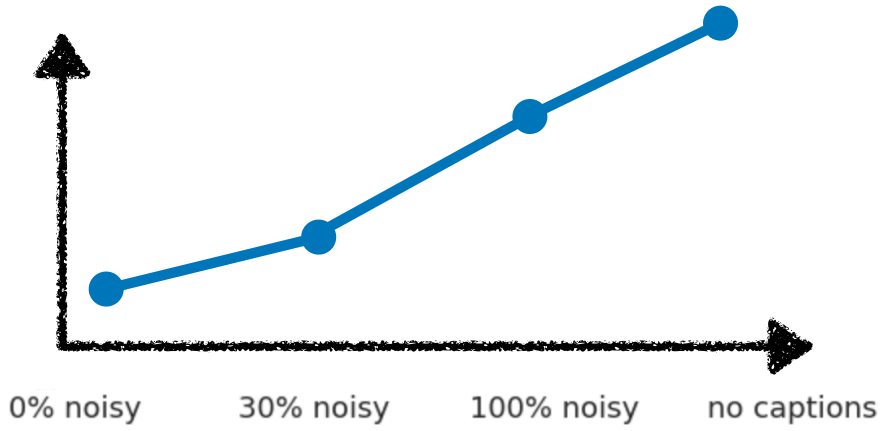
Sanity check experiment setup

image concept

caption concept

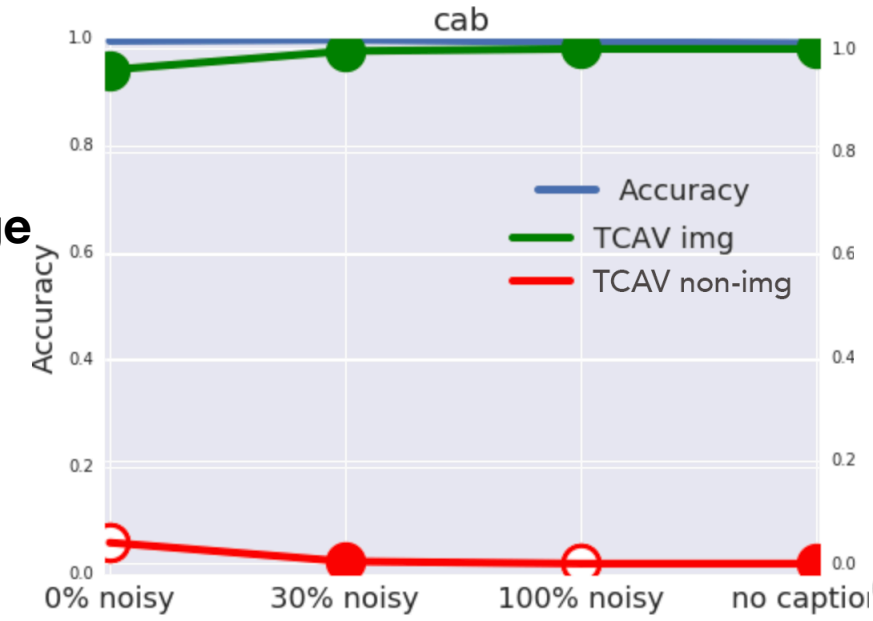
models can use either image or caption concept for classification.

Test accuracy with no caption image = Importance of image concept

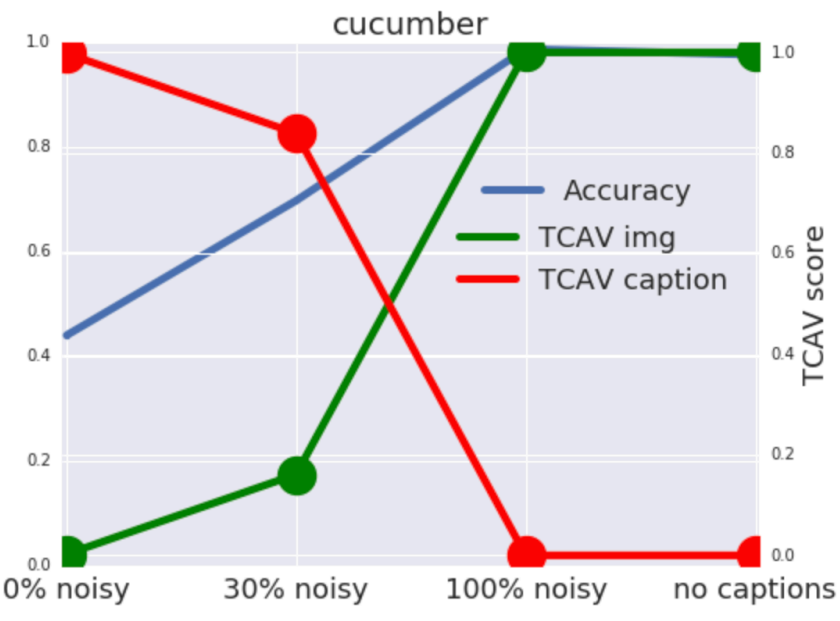


Sanity check experiment

Test accuracy with no caption image



Caption noise level in training set

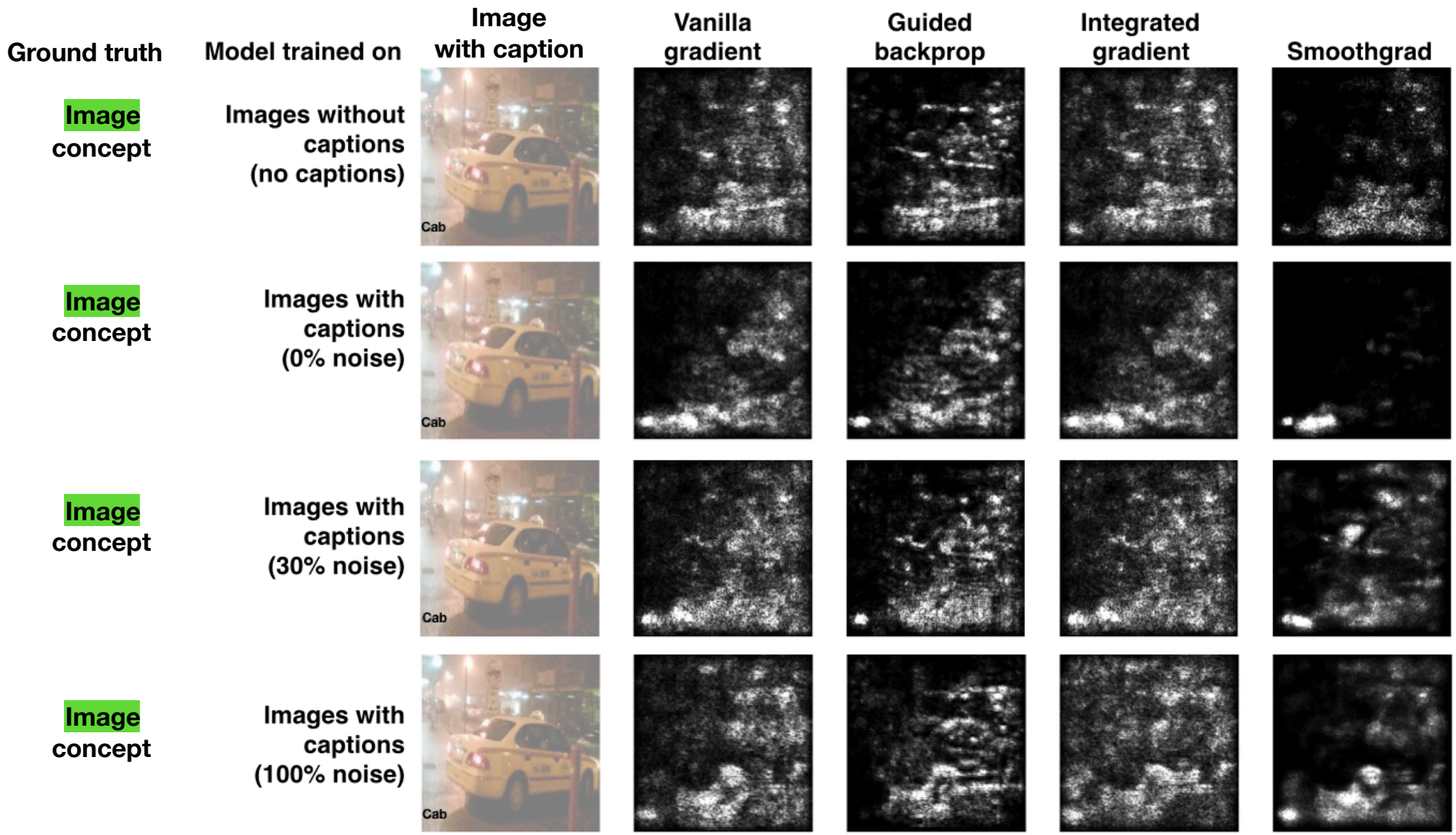


Caption noise level in training set

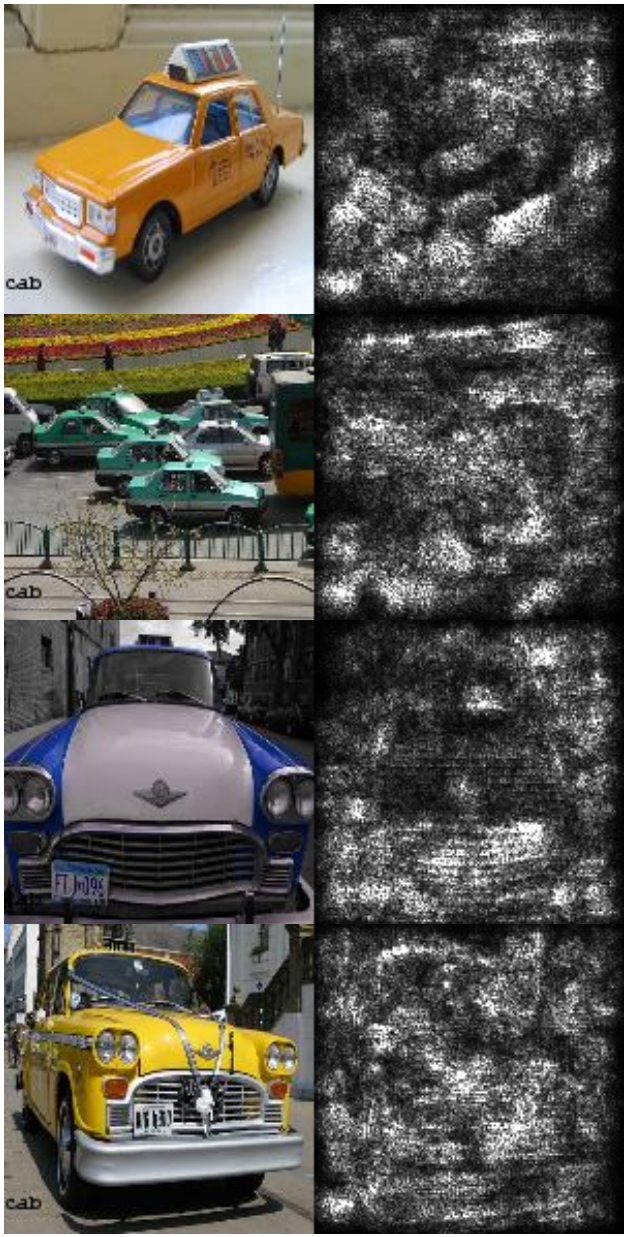
Cool, cool.

Can saliency maps do this too?

Can saliency maps communicate the same information?



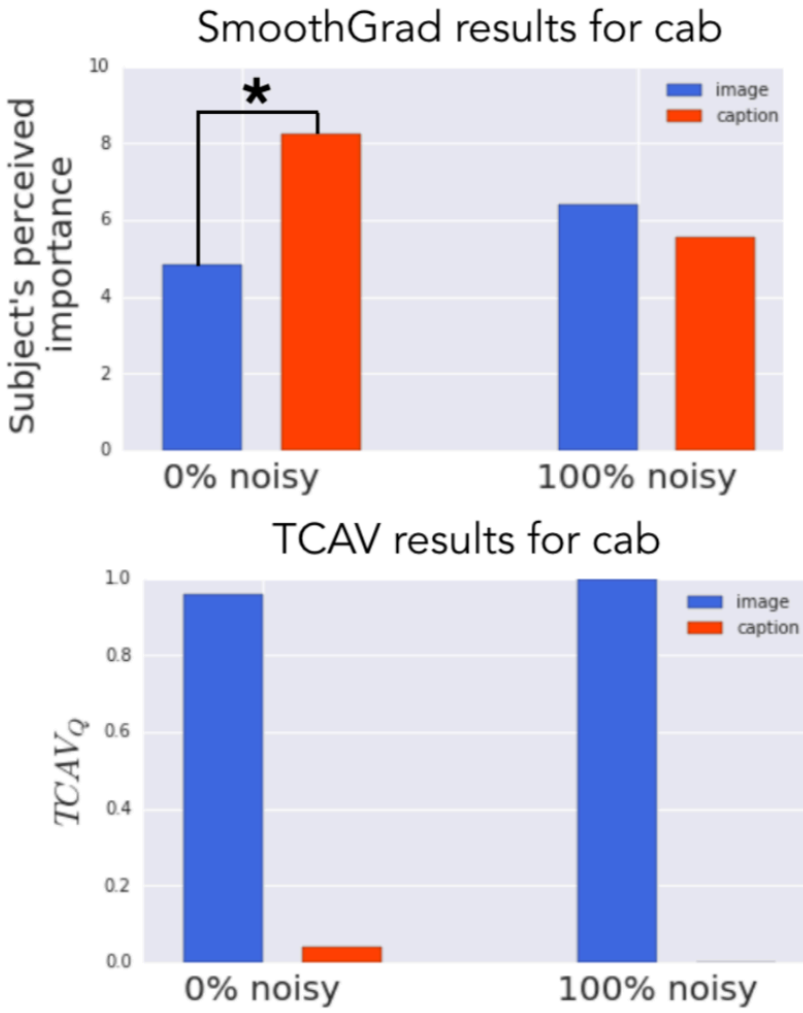
Human subject experiment: Can saliency maps communicate the same information?



- 50 turkers are
  - asked to judge importance of image vs. caption given saliency maps.
  - asked to indicate their confidence
  - shown 3 classes (cab, zebra, cucumber) x 2 saliency maps for one model

Human subject experiment: Can saliency maps communicate the same information?

- Random chance: 50%
- Human performance with saliency map: 52%
- Humans can't agree: more than 50% no significant consensus



Human subject experiment: Can saliency maps communicate the same information?

- Random chance: 50%
- Human performance with saliency map: 52%
- Humans can't agree: more than 50% no significant consensus
- Humans are **very** confident even when they are wrong.

Results

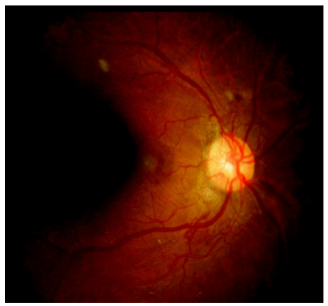
1. Sanity check experiment

cab image cab image with caption

2. Biases from Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy

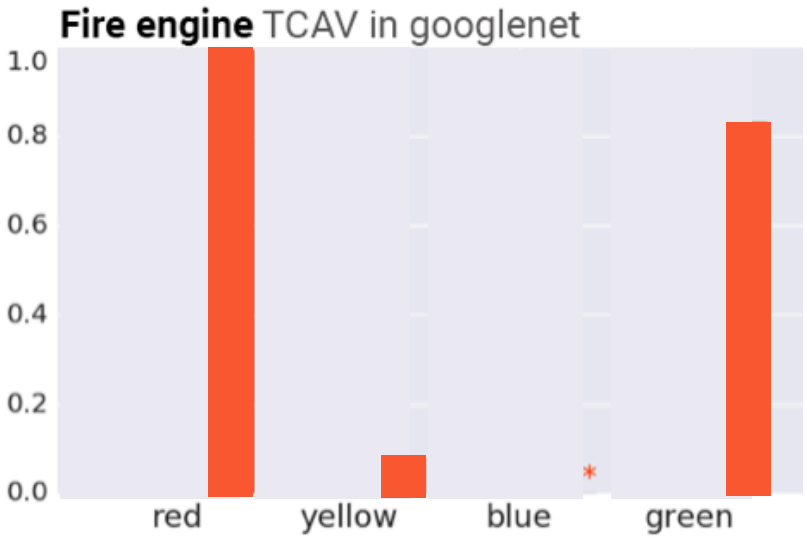
DR level 4 Retina

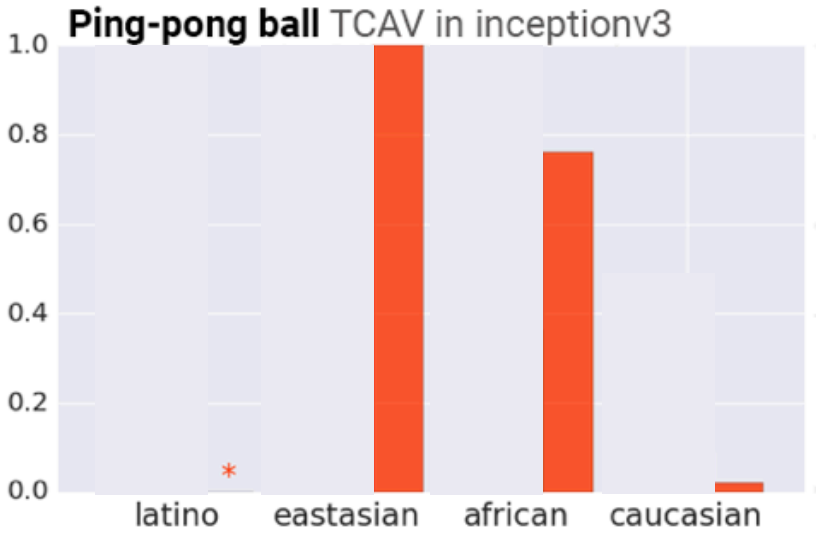


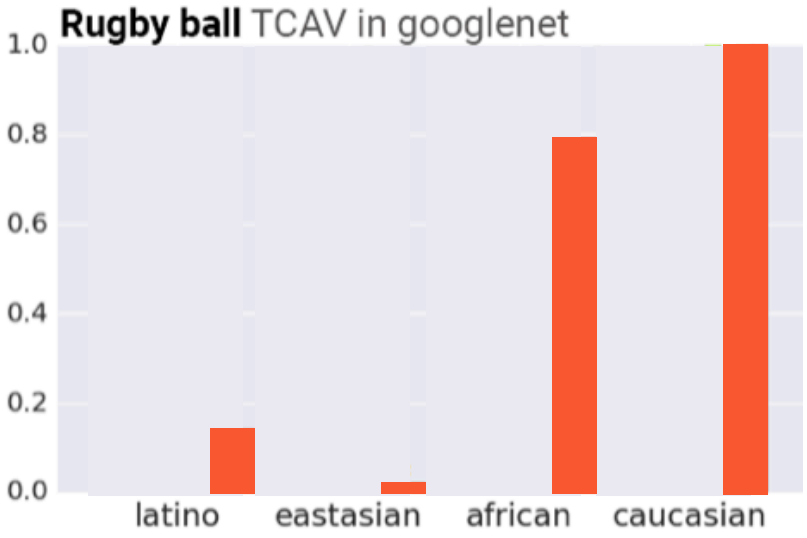
TCAV for DR level 4

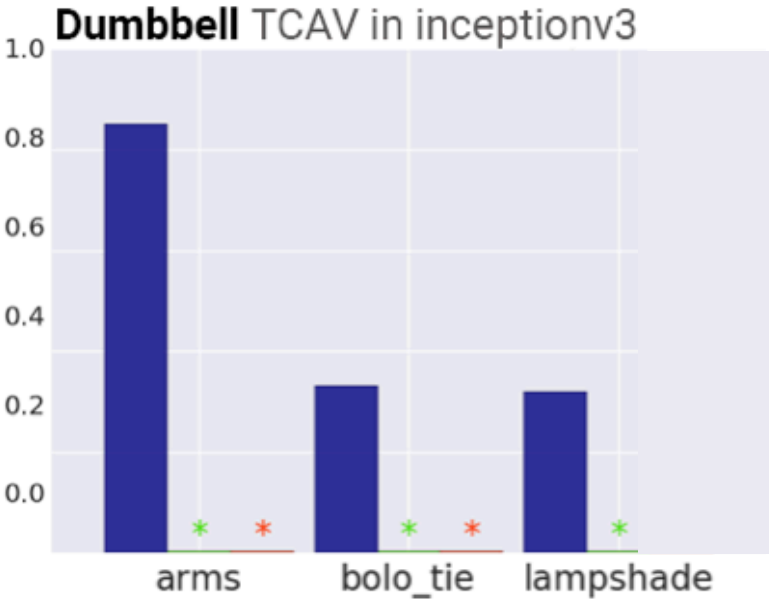
TCAV in

Two widely used image prediction models





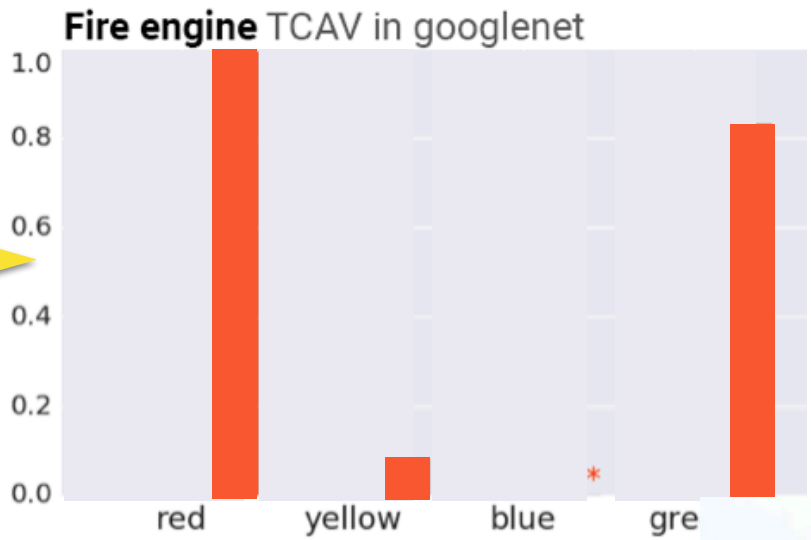


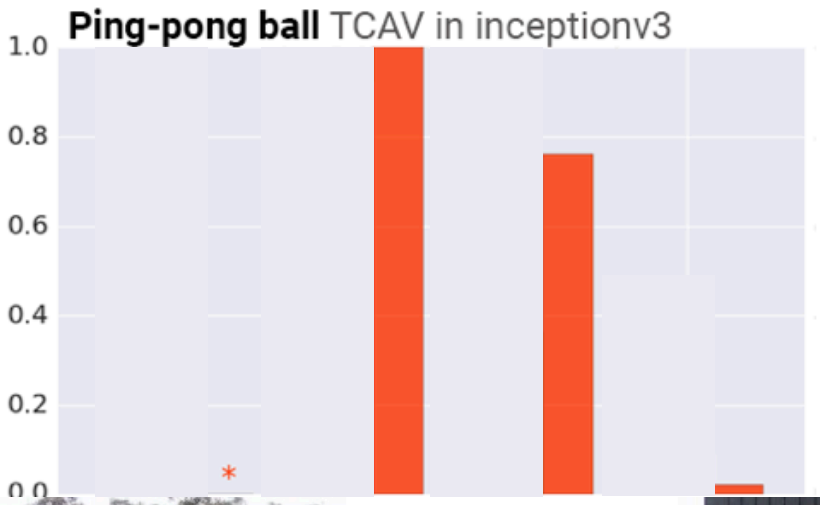


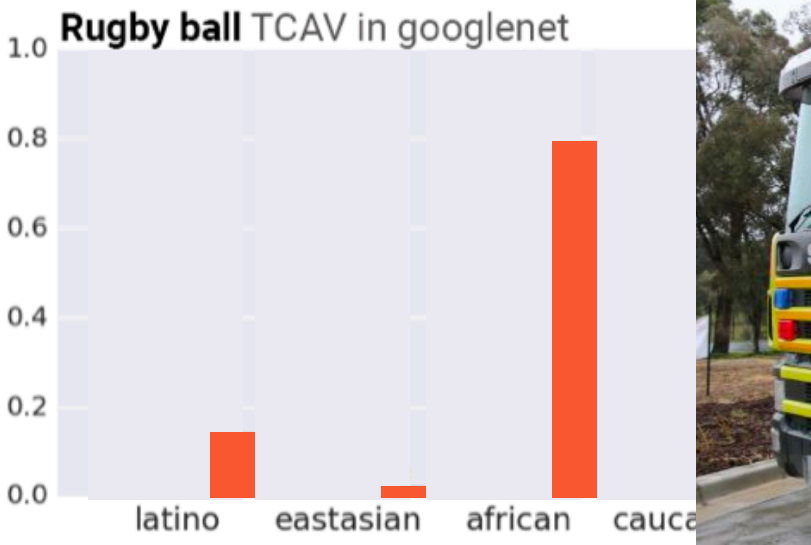
TCAV in

Two widely used image prediction models

Geographical bias!



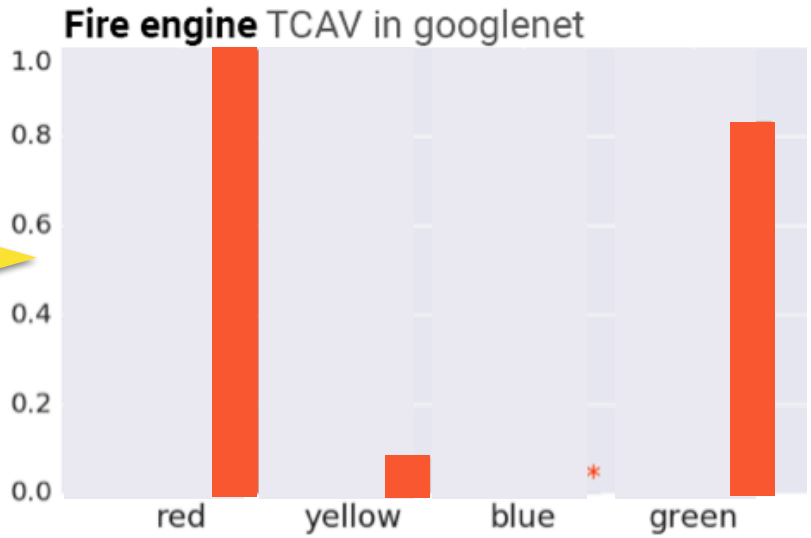


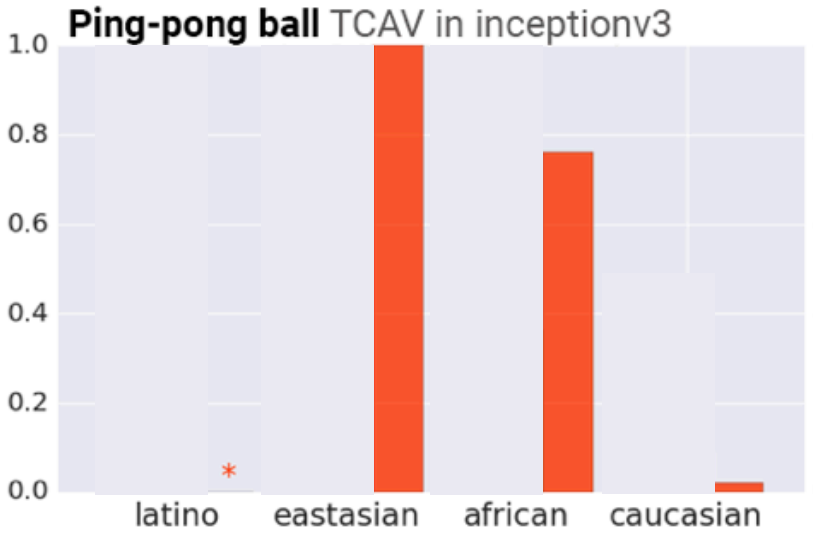


TCAV in

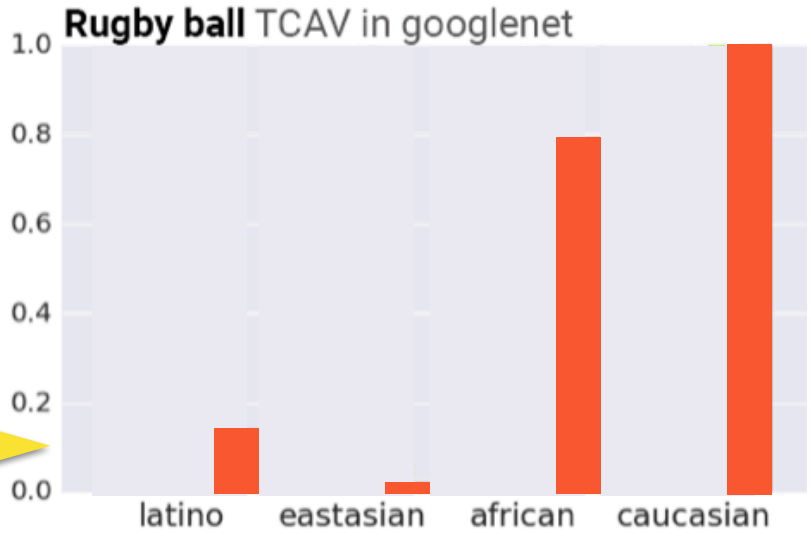
Two widely used image prediction models

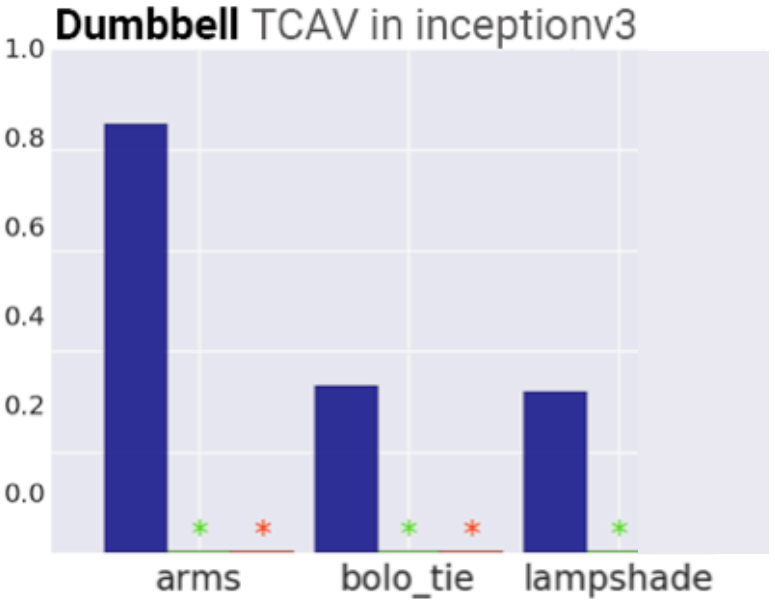
Geographical bias?





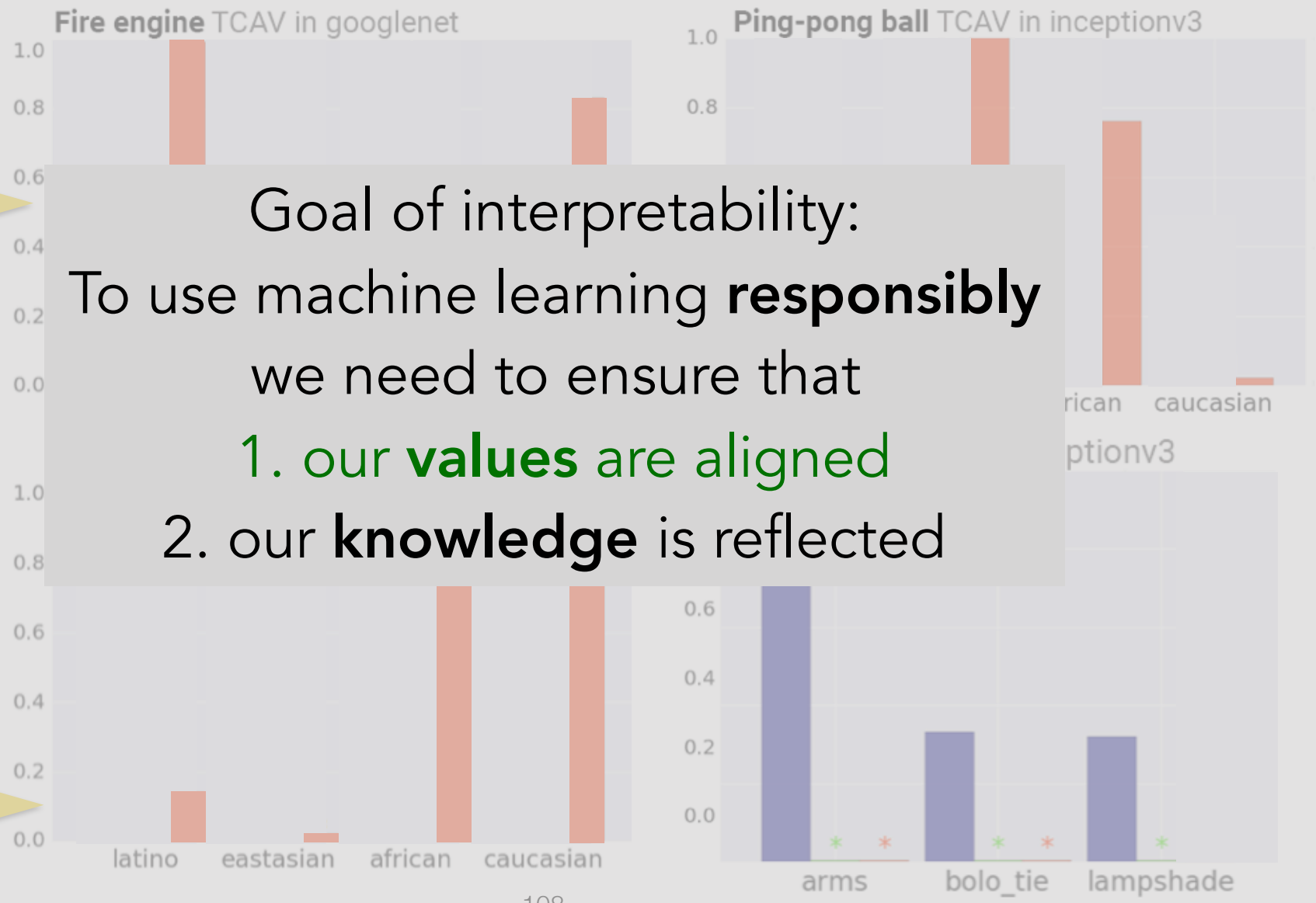
Quantitative confirmation to previously qualitative findings [Stock & Cisse, 2017]





TCAV in

Two widely used image prediction models



Geographical bias?

Quantitative confirmation to previously qualitative findings [Stock & Cisse, 2017]

Goal of interpretability:

To use machine learning **responsibly**

we need to ensure that

1. our **values** are aligned
2. our **knowledge** is reflected

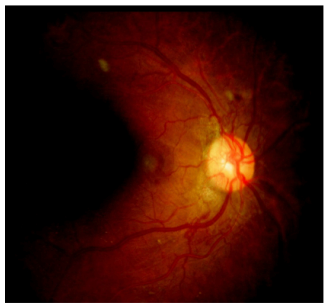
Results

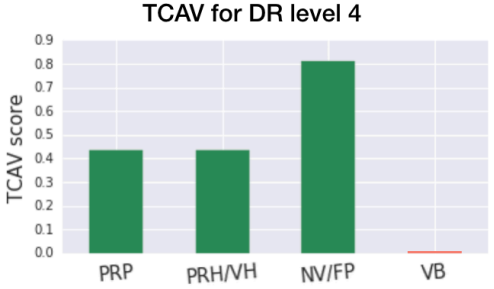
1. Sanity check experiment

cab image cab image with caption

2. Biases Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy





Diabetic Retinopathy

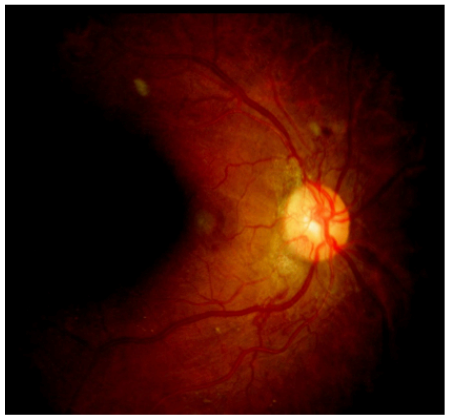
- Treatable but sight-threatening conditions
- Have model to with accurate prediction of DR (85%) [Krause et al., 2017]

Concepts the **ML model** uses

Vs

Diagnostic Concepts **human** doctors use

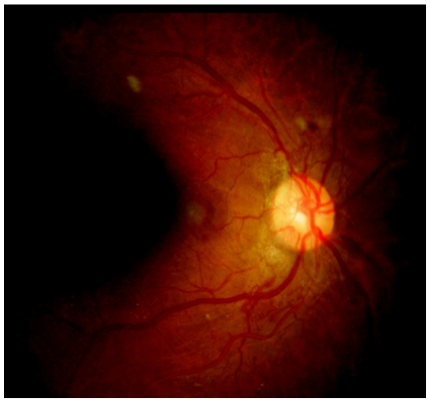
DR level 4 Retina



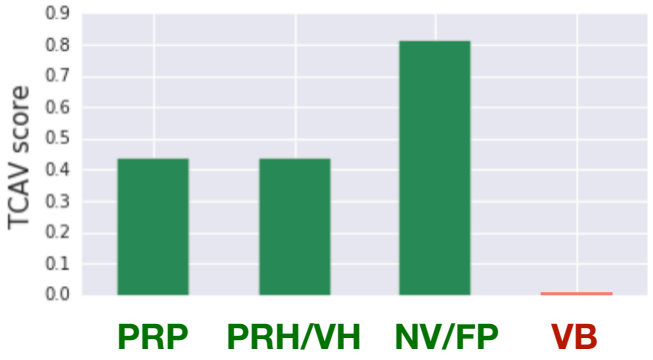
TCAV for Diabetic Retinopathy

Prediction class	Prediction accuracy
DR level 4	High

Example



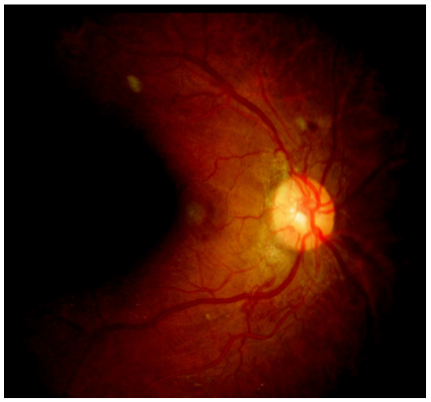
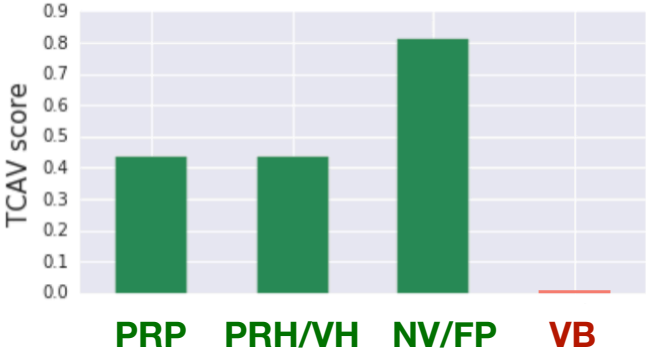
TCAV scores



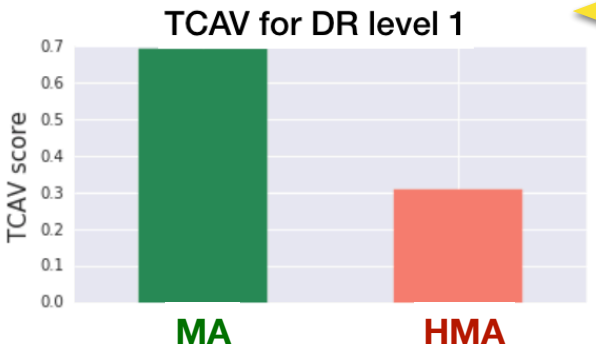
TCAV shows the model is **consistent** with doctor's knowledge when model is **accurate**

Green: domain expert's label on concepts belong to the level
Red: domain expert's label on concepts does not belong to the level

TCAV for Diabetic Retinopathy

Prediction class	Prediction accuracy	Example	TCAV scores
DR level 4	High		

TCAV shows the model is **consistent** with doctor's knowledge when model is **accurate**

DR level 1	Med		

TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate

Green: domain expert's label on concepts belong to the level
Red: domain expert's label on concepts does not belong to the level

TCAV for Diabetic Retinopathy

Prediction class	Prediction accuracy	Example
DR level 4	Hi	
DR level 1	Low	

Level 1 was often confused to level 2.

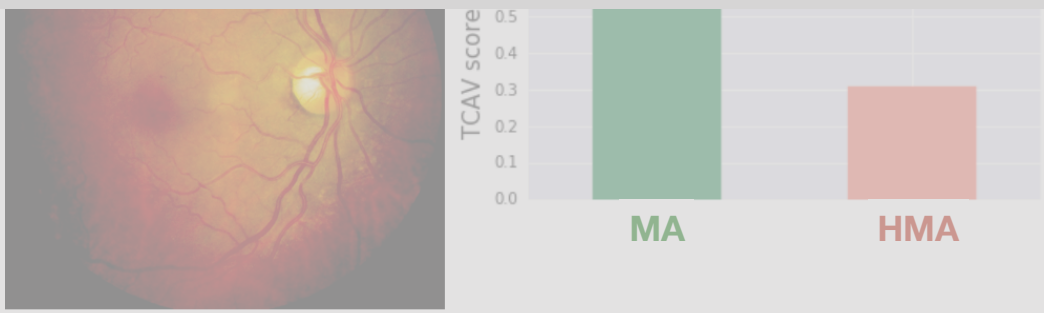
Goal of interpretability:

To use machine learning **responsibly**

we need to ensure that

1. our **values** are aligned

2. our **knowledge** is reflected



CAV shows the
del is **inconsistent**
with doctor's
knowledge for classes
when model is less
accurate

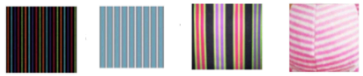
Green: domain expert's label on concepts belong to the level
Red: domain expert's label on concepts does not belong to the level

Summary:

code: github.com/tensorflow/tcav

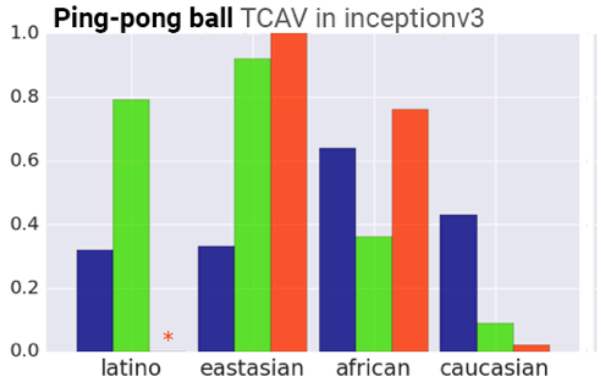
Testing with Concept Activation Vectors

Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres

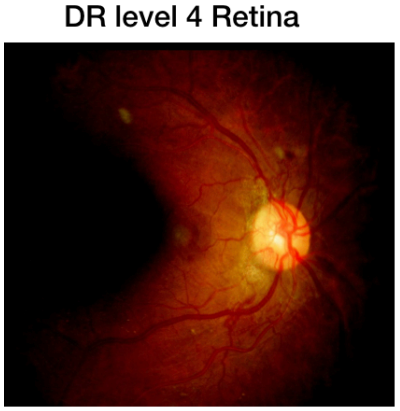


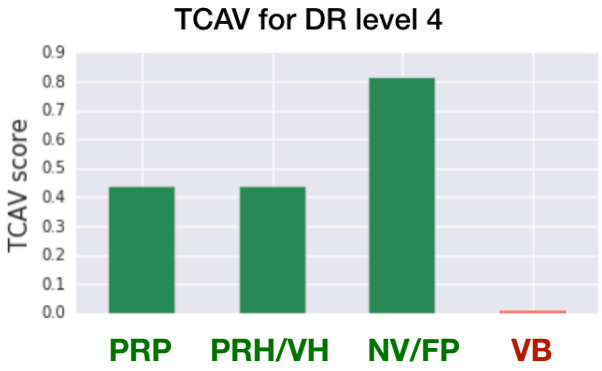
stripes concept (score: 0.9) was important to **zebra** class for this trained network. 

TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

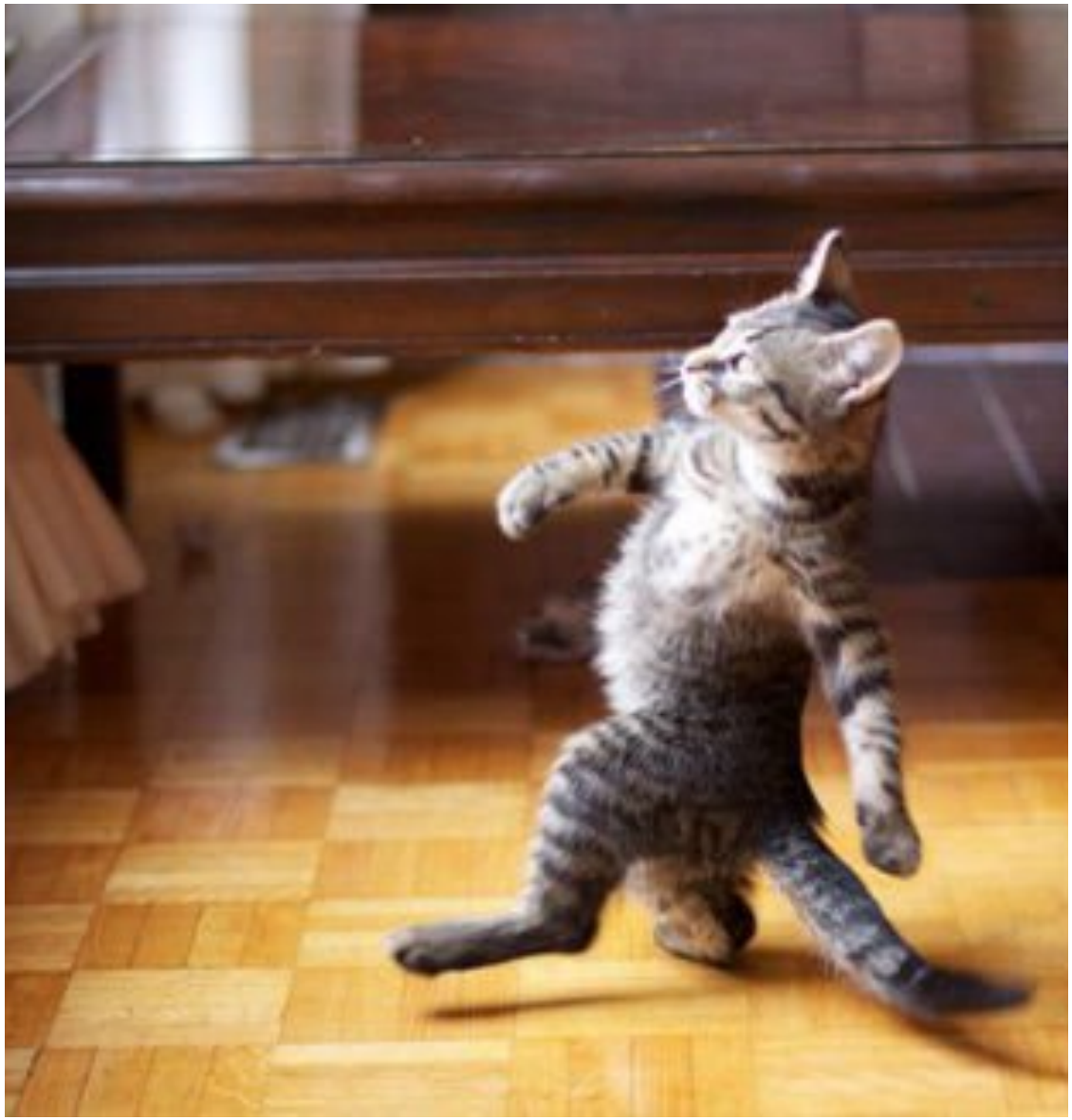


Our values





Our knowledge



https://imgflip.com