Google AI

Interpretability: the myth, questions, and some answers.

Been Kim

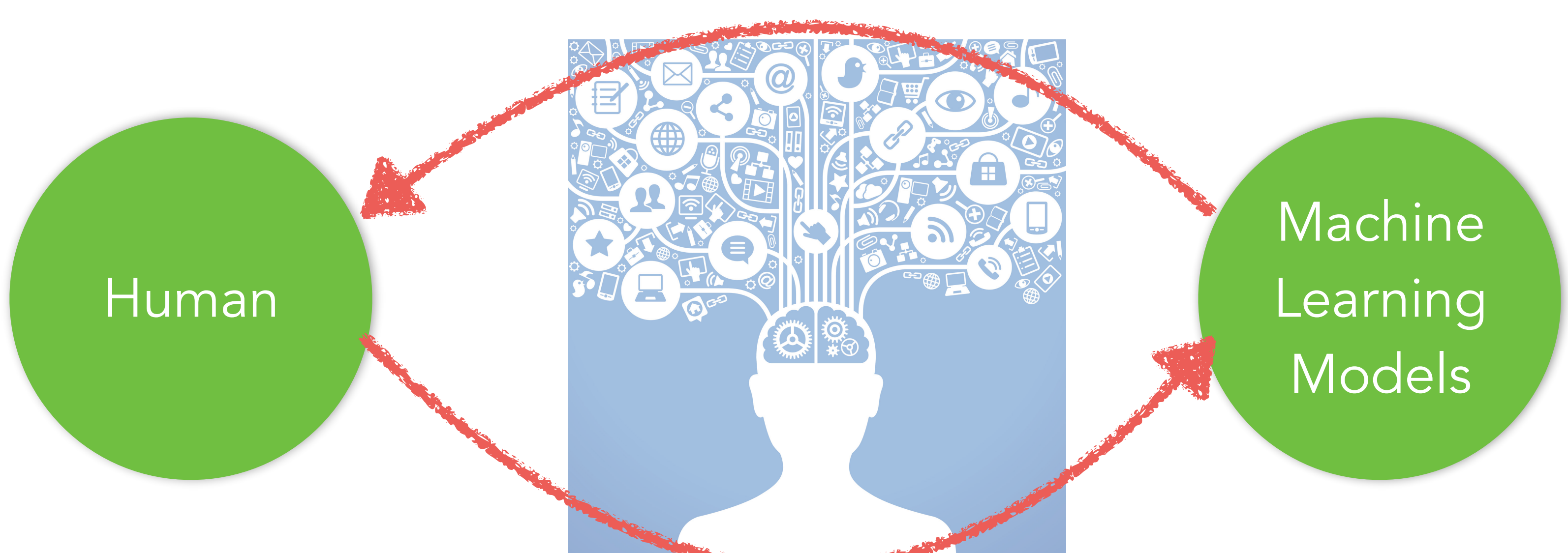
Presenting a subset of work
with a lot of awesome people inside and outside of Google:

Martin Wattenberg, Finale Doshi-Velez, Julius Adebayo, Heinrich Jiang, Maya Gupta, Ike Lage, Andrew Ross, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres, Ian Goodfellow, Mortiz Hardt, Sam Gershman, Menaka Narayanan, Emily Chen, Jeffrey He

My goal

interpretability

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



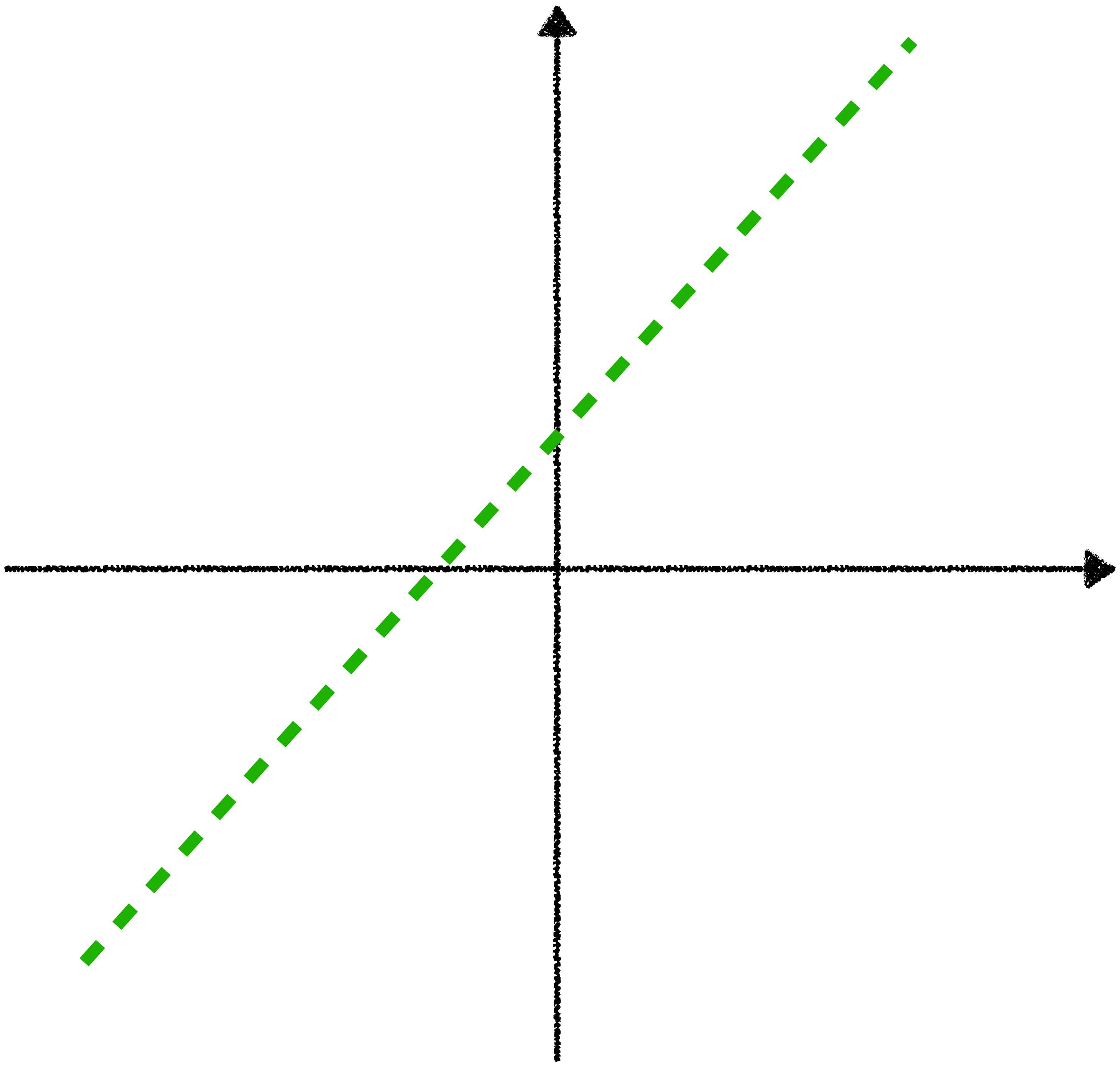
http://blogs.teradata.com/

ingredients for interpretability methods.

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

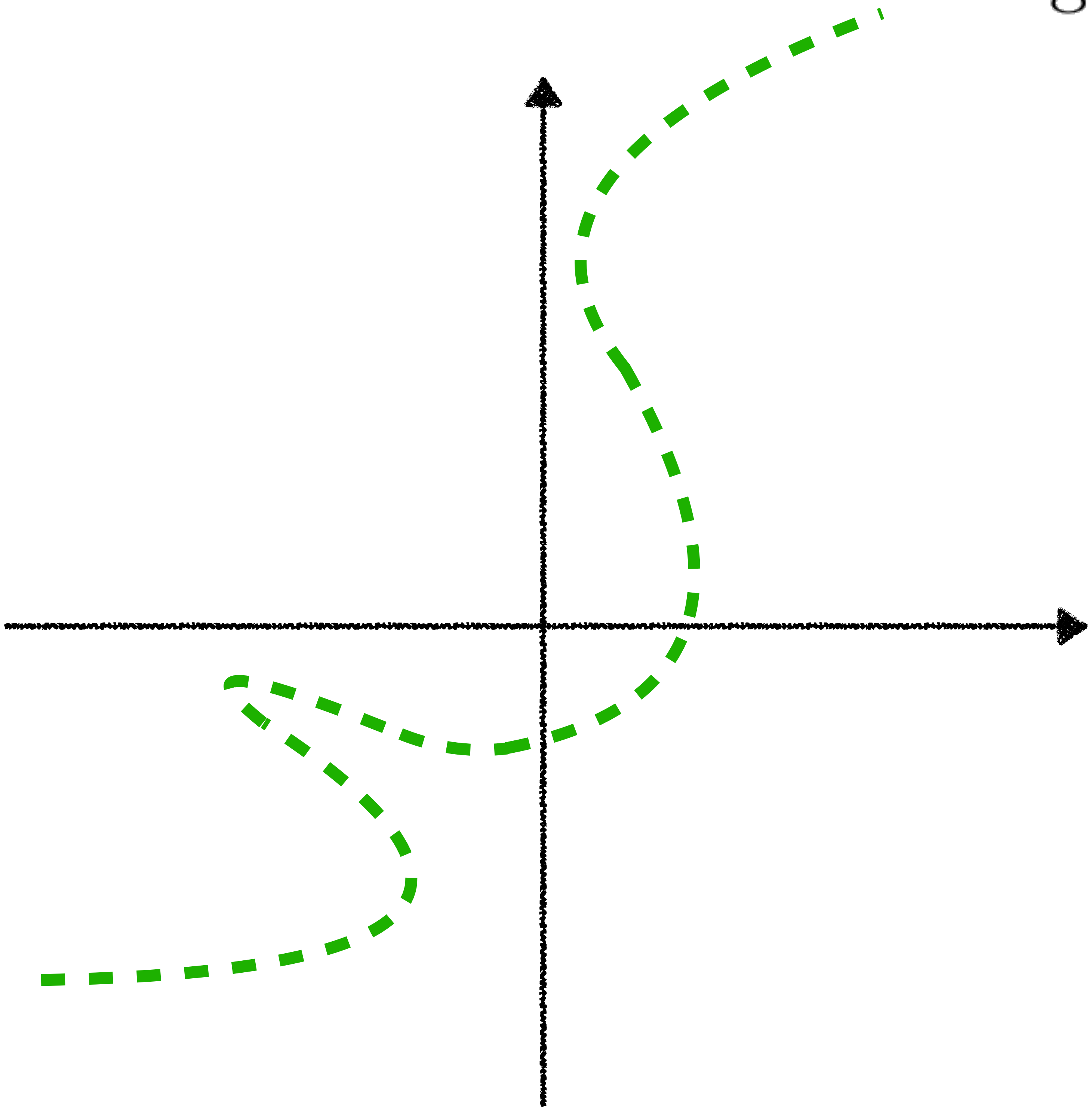
Some quality function

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



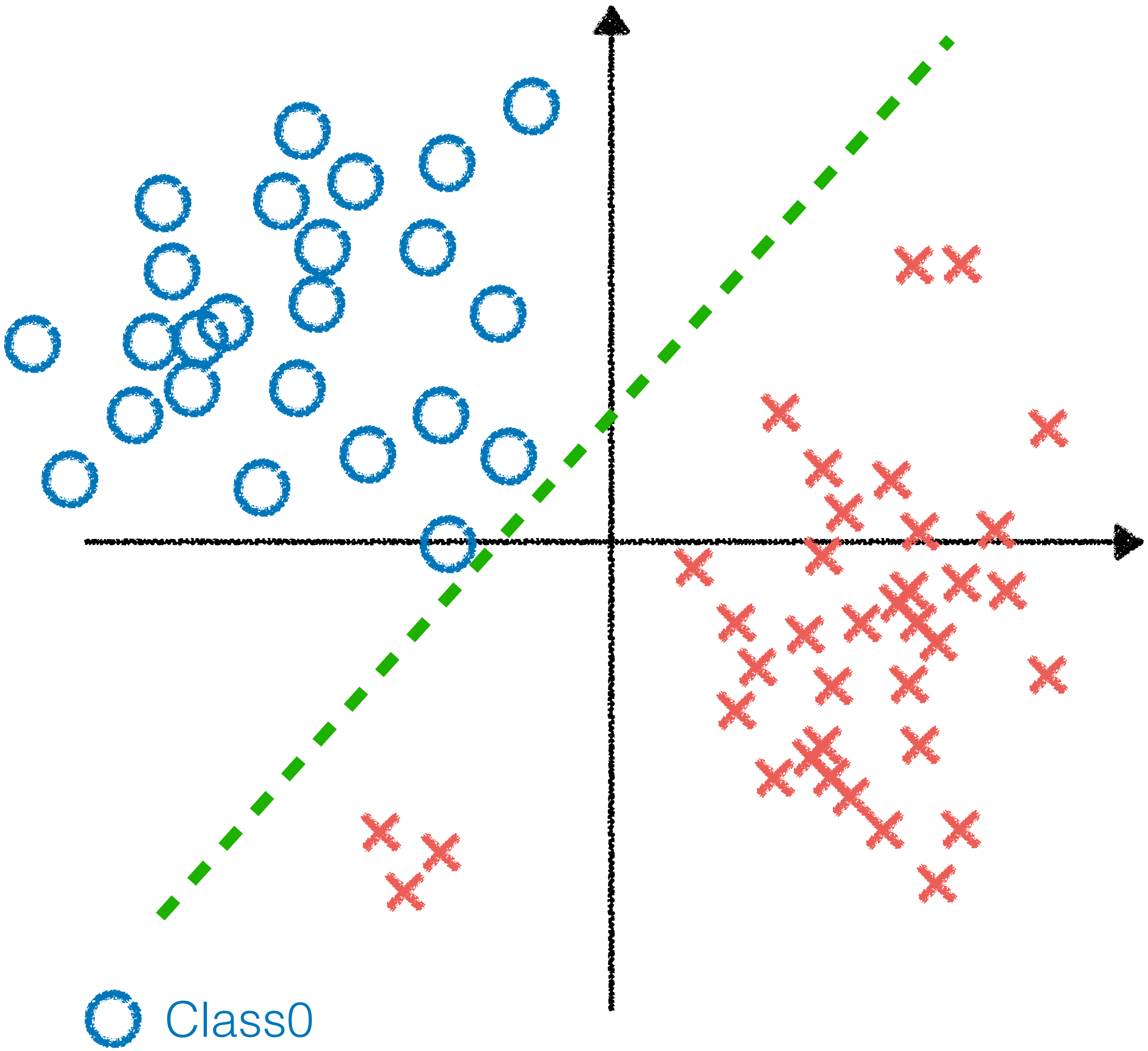
Model

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



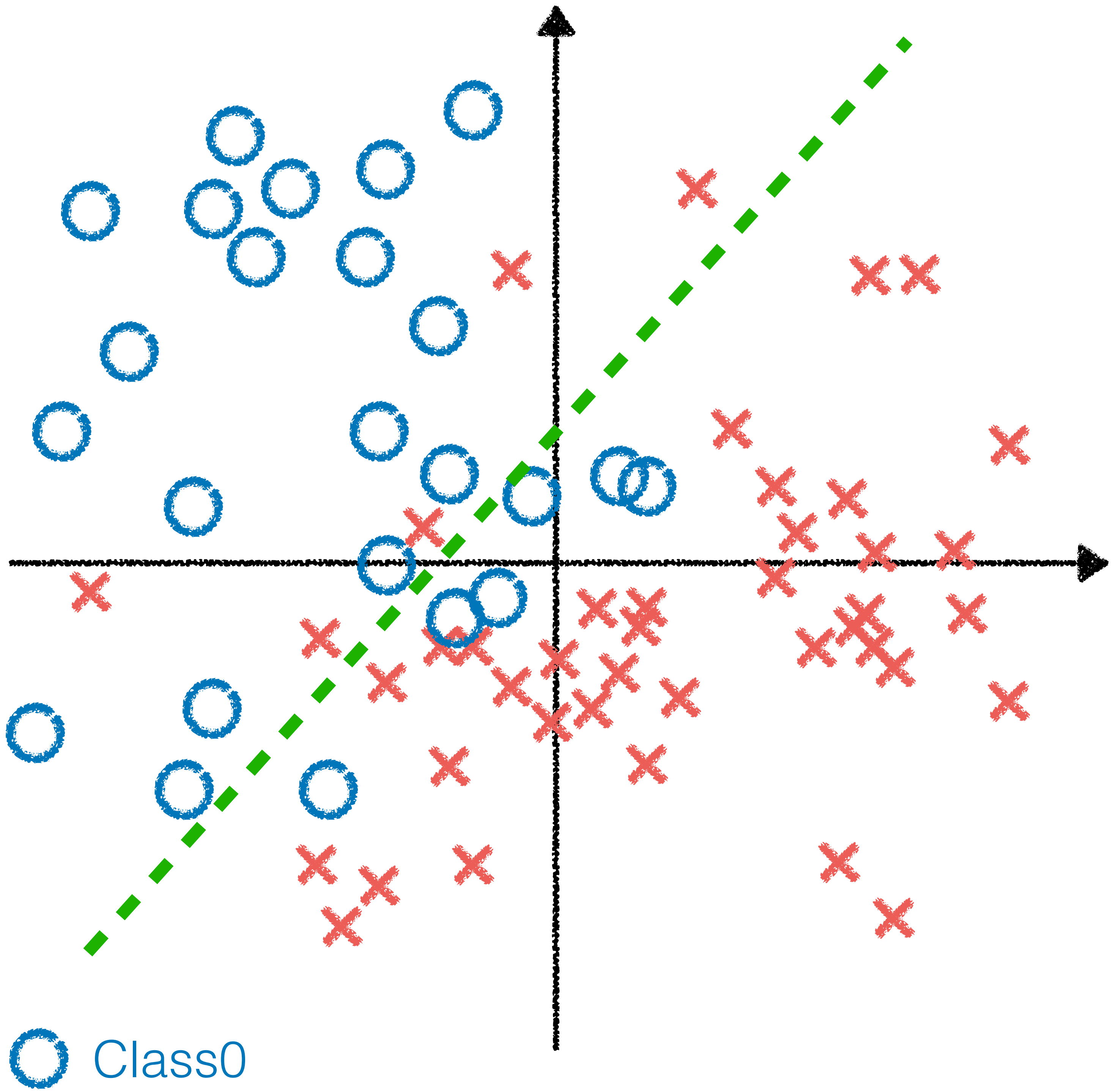
Model

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



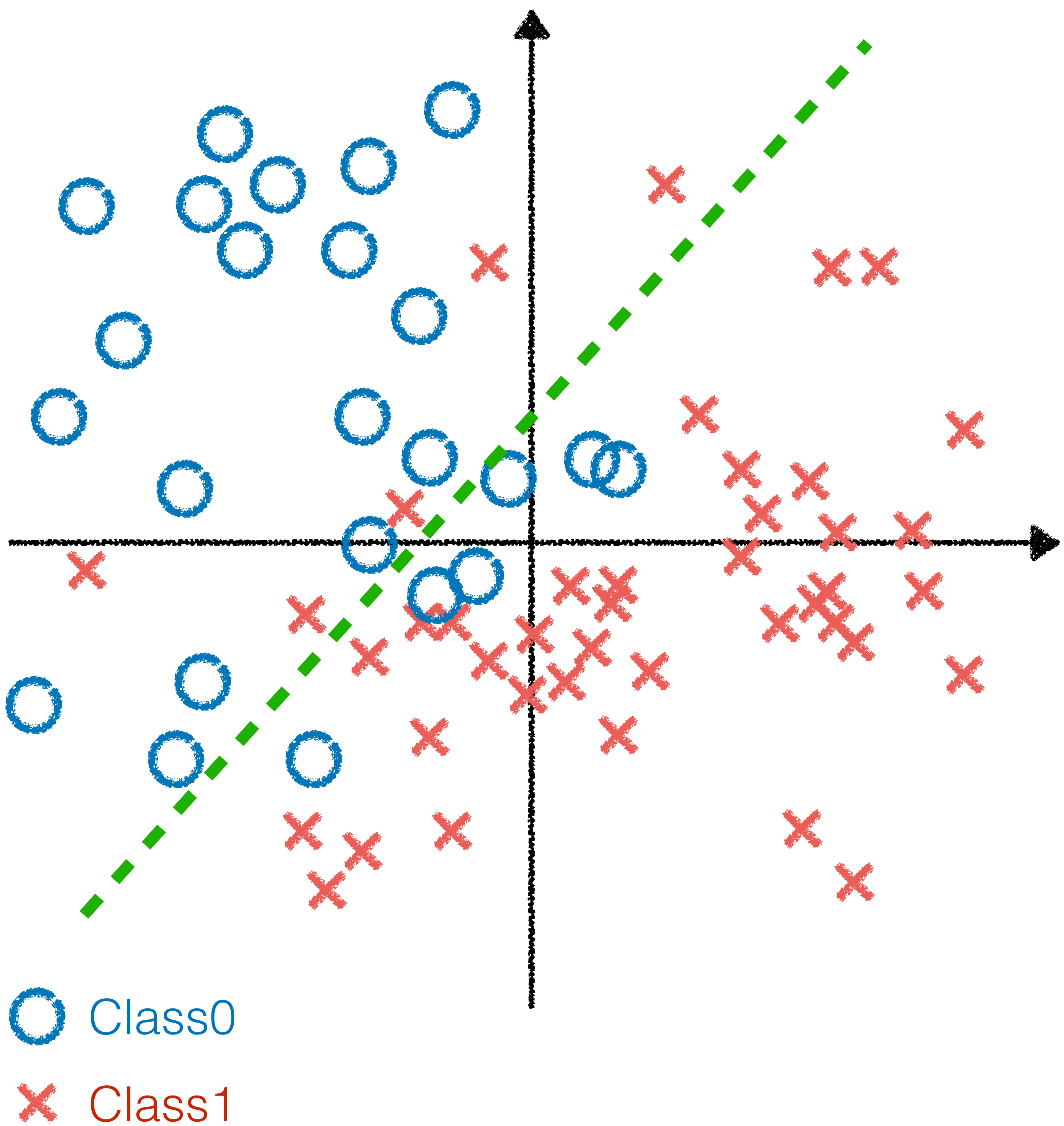
Data

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

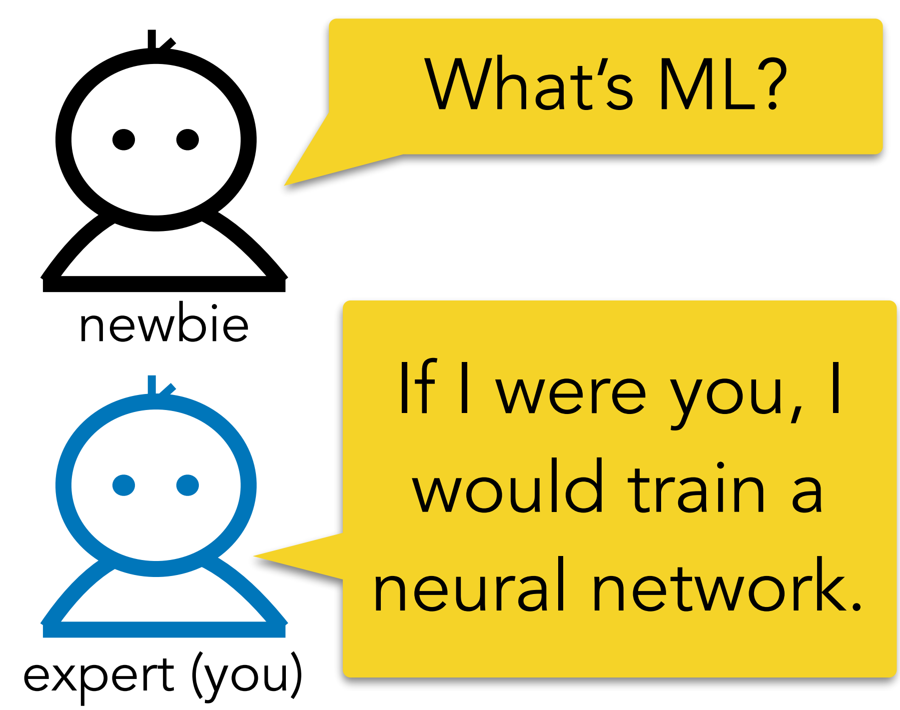


Data

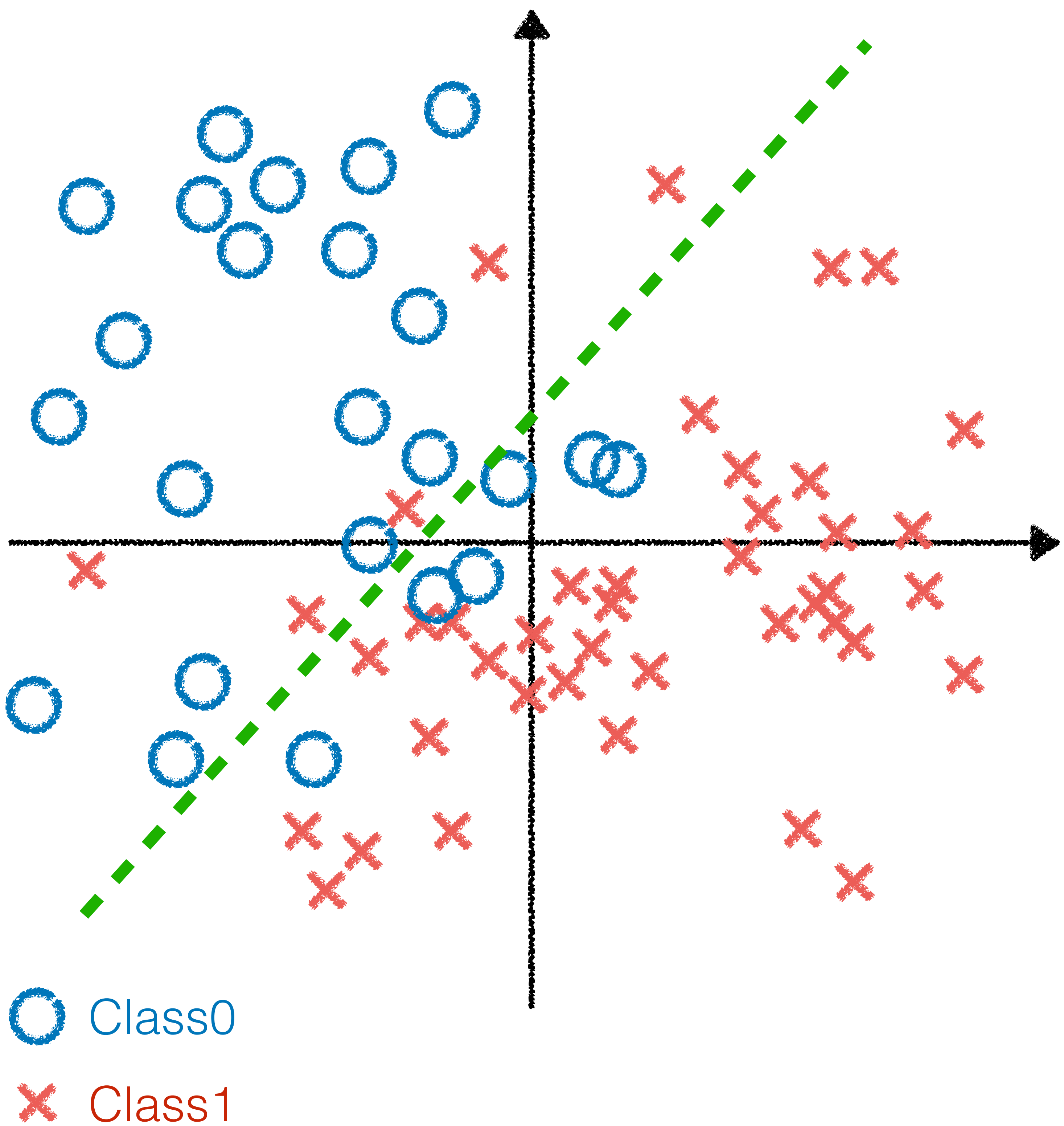
$$\underset{E}{\operatorname{argmax}}\ Q(E|M, \underline{H}, D, \quad)$$



Human



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



Task

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

Agenda

Post-training explanations

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Building inherently interpretable models

$$\underset{E,M}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}, \mathbf{M}\text{odel}|\mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Agenda

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Agenda

1. Revisit some existing methods: Sanity check questions

2. Make explanations that work for lay people.

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.

Agenda

1. Revisit some existing methods: Sanity check questions

2. Make explanations that work for lay people.

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\mathrm{xplanation}|\mathbf{M}\mathrm{odel}, \mathbf{H}\mathrm{uman}, \mathbf{D}\mathrm{ata}, \mathbf{T}\mathrm{ask})$$

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.

Problem:
Post-training explanation

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

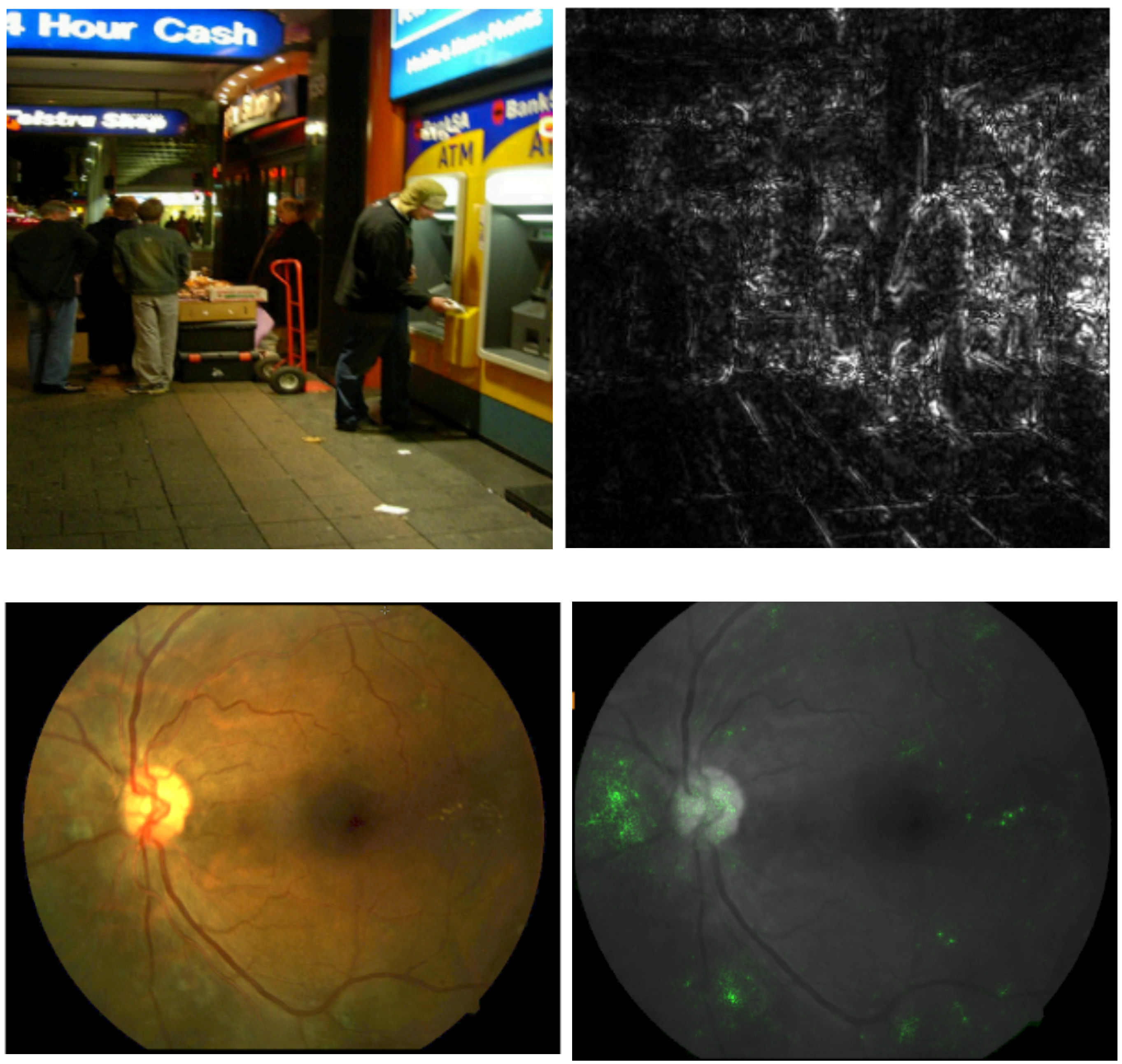
→ A trained machine learning model (e.g., neural network) → $p(z)$

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 $\rightarrow$ $\dfrac{\partial p(z)}{\partial x_{i,j}}$
pixel i,j $\rightarrow$

NN

local undestanding

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

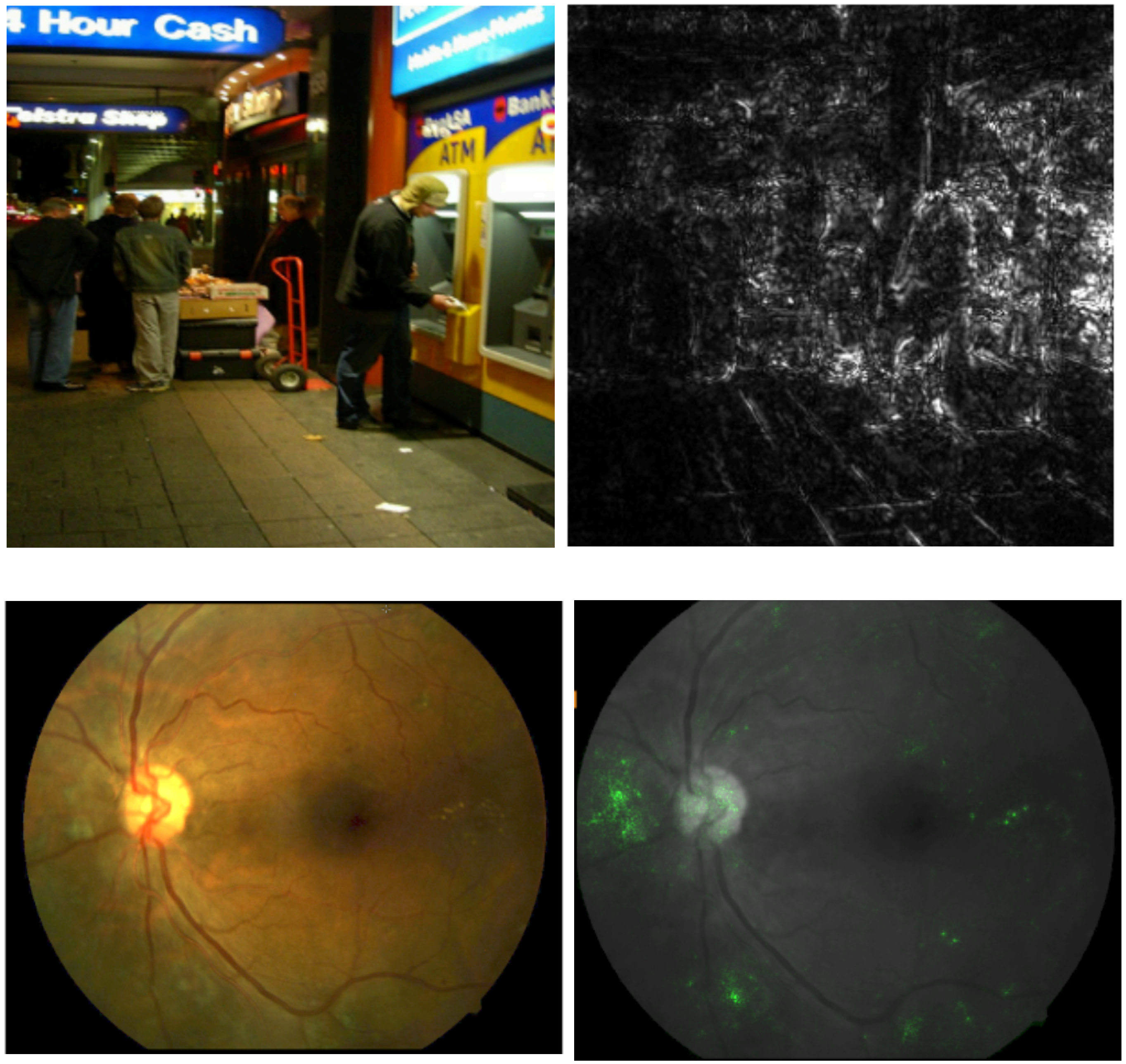
humans' subjective judgement

widely used for images

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

One of the most popular interpretability methods for images:

Saliency maps



Used for image classification and medical applications.

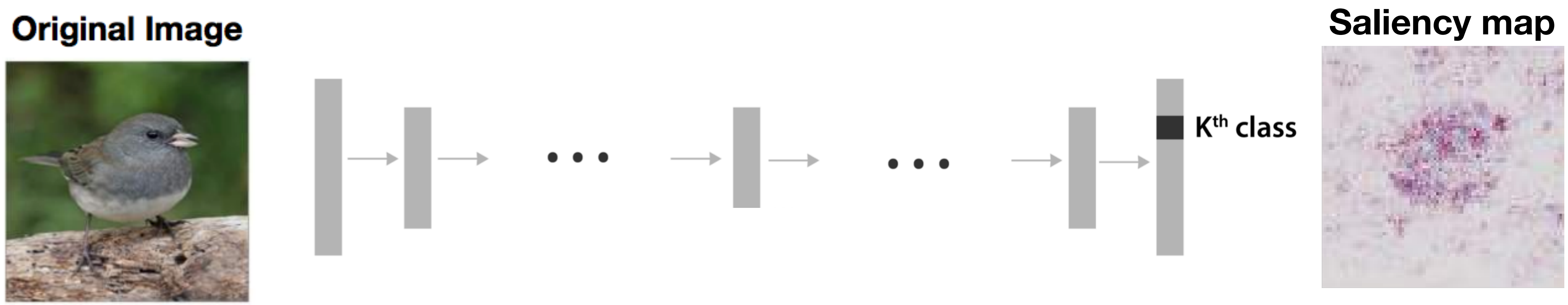
$$\begin{array}{lc} \text{a logit} \rightarrow & \dfrac{\partial p(z)}{\partial x_{i,j}} \\ \text{pixel i,j} \rightarrow & \end{array}$$

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

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.

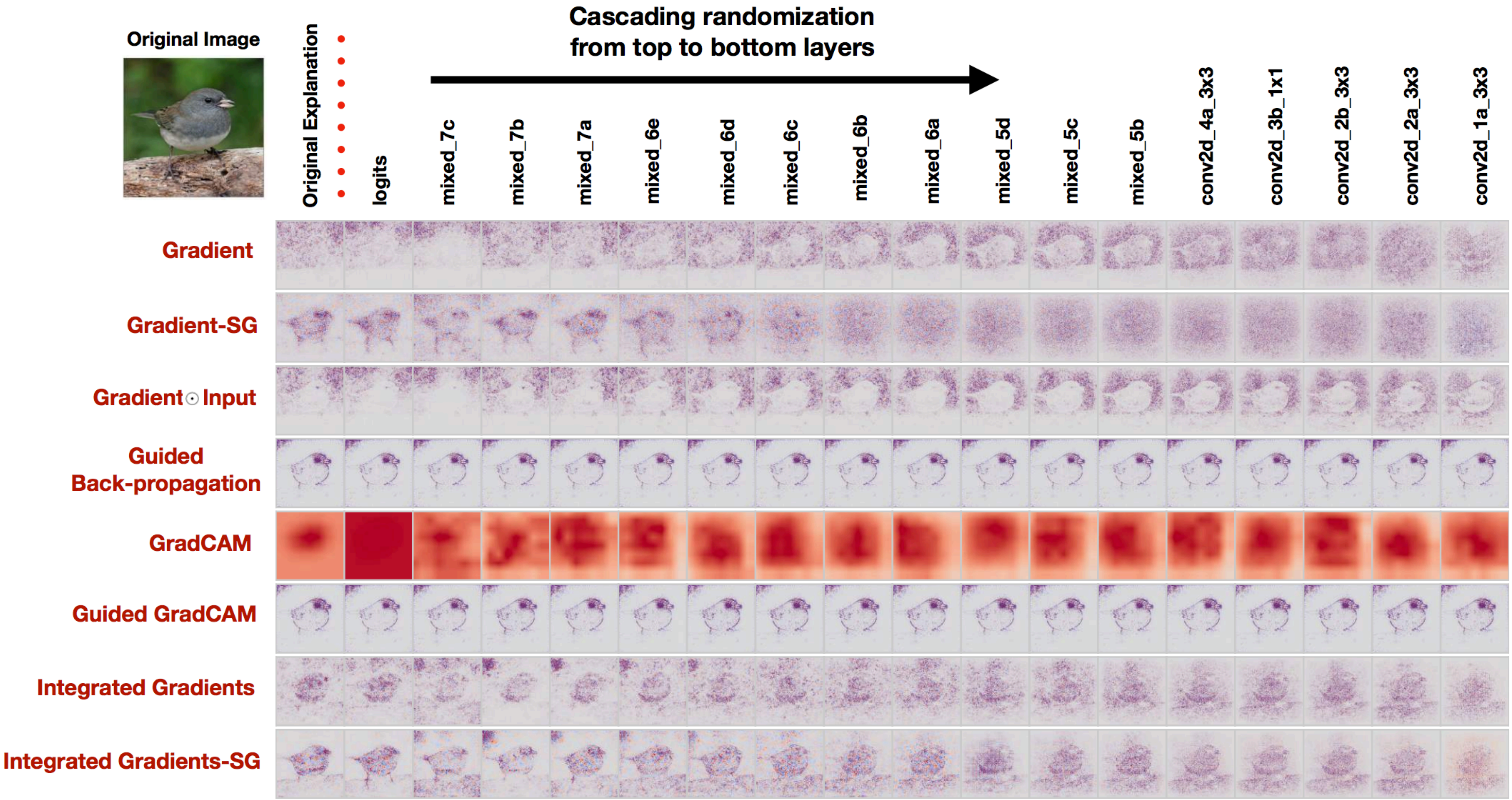


!!!!!???!?

Randomized weights!
Network now makes garbage prediction.

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

Some saliency maps look similar when we randomize the network.



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

What can we do better? Creating a wishlist.

human's subjective judgement

Using input features as a language

lay person?

local undestanding

$$\underset{E}{\operatorname{argmax}}\; Q(\mathbf{E}\mathrm{xplanation}|\mathbf{M}\mathrm{odel}, \mathbf{H}\mathrm{uman}, \mathbf{D}\mathrm{ata}, \mathbf{T}\mathrm{ask})$$

What can we do better?
Creating a wishlist.

Something more human-friendly?

quantitive

~~human's~~
~~subjective~~
judgement

Using
~~input~~
~~features~~
as a language

lay
person~~?~~ **!**

global

~~local~~
undestanding

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\mathrm{xplanation}|\mathbf{M}\mathrm{odel}, \mathbf{H}\mathrm{uman}, \mathbf{D}\mathrm{ata}, \mathbf{T}\mathrm{ask})$$

Agenda

1. Revisit some existing methods:
Sanity check questions

2. Make explanations
that work for lay people.

$$\underset{E}{\operatorname{argmax}}\; Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

3. Understand how humans
understand explanations

4. Make explanations to detect
trustworthy predictions.

TCAV [ICML'18]
Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres

Problem:
Post-training explanation

$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

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

$p(z)$

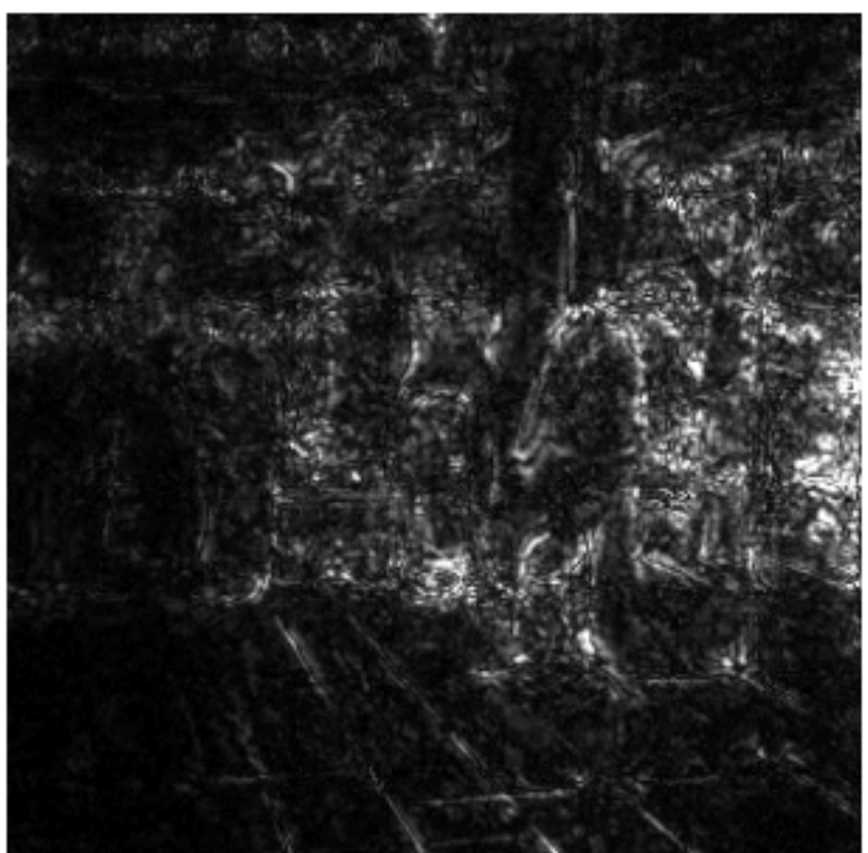
cash-machine-ness

Why was this a cash machine?

TCAV [ICML'18]
Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres

Common solution: Saliency map

prediction:
Cash machine



Let's use this to help us think about what we really want to ask.

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

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

What we really want to ask…

prediction:
Cash machine

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

Did the 'human' concept matter?
Did the 'wheels' concept matter?

Which concept mattered more?

Is this true for all other cash machine predictions?

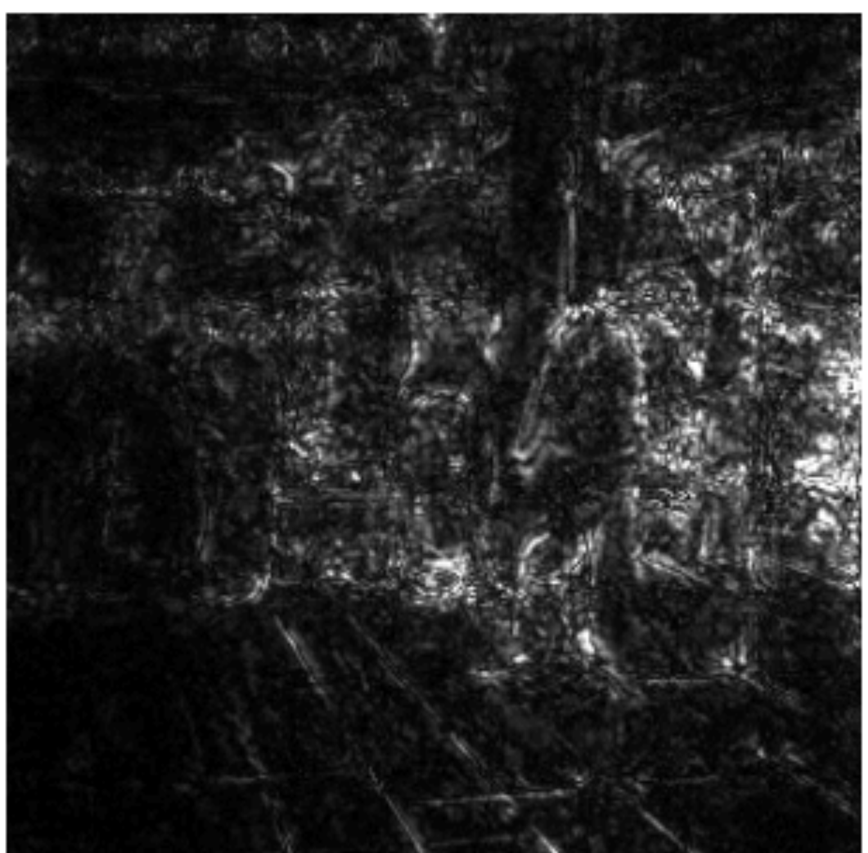
Oh no! I can't express these concepts as pixels!!
They weren't my input features either!

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

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

What we really want to ask…

prediction:
Cash machine



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

Did the 'human' concept matter?
Did the 'wheels' concept matter?

Which concept mattered more?

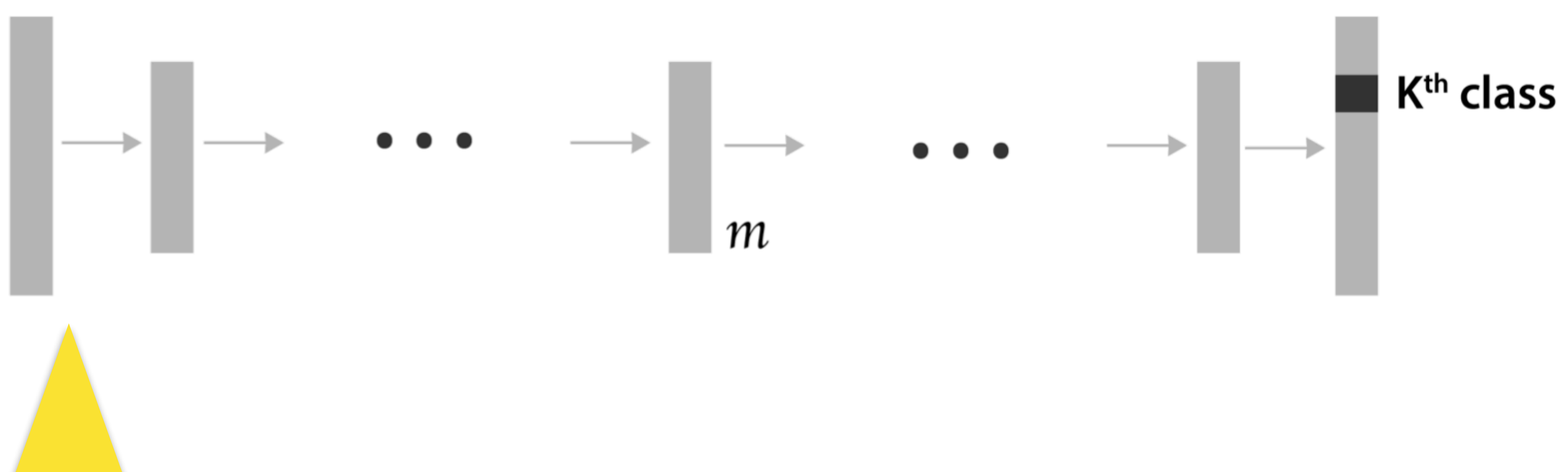
Is this true for all other cash machine predictions?

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

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

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

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.

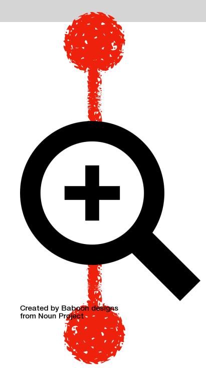
Goal of TCAV:
Testing with Concept Activation Vectors

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

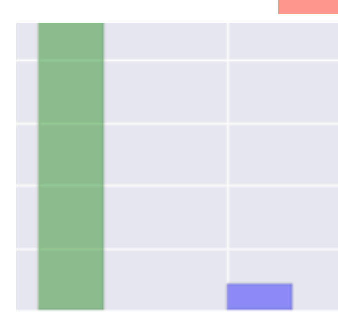
$p(z)$

Doctor-ness

Was gender concept important to this doctor image classifier?



TCAV score for **Doctor**



not women **women**

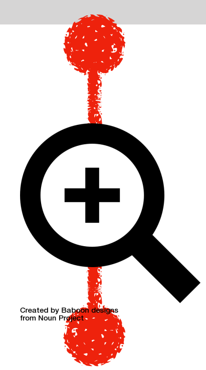
TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

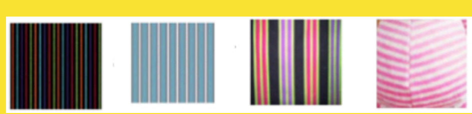
Goal of TCAV:
Testing with Concept Activation Vectors

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

$p(z)$

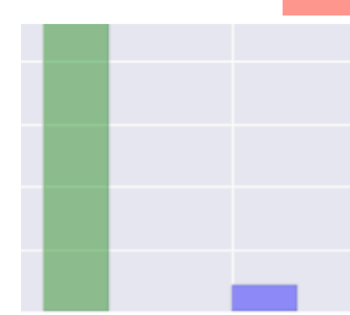
zebra-ness





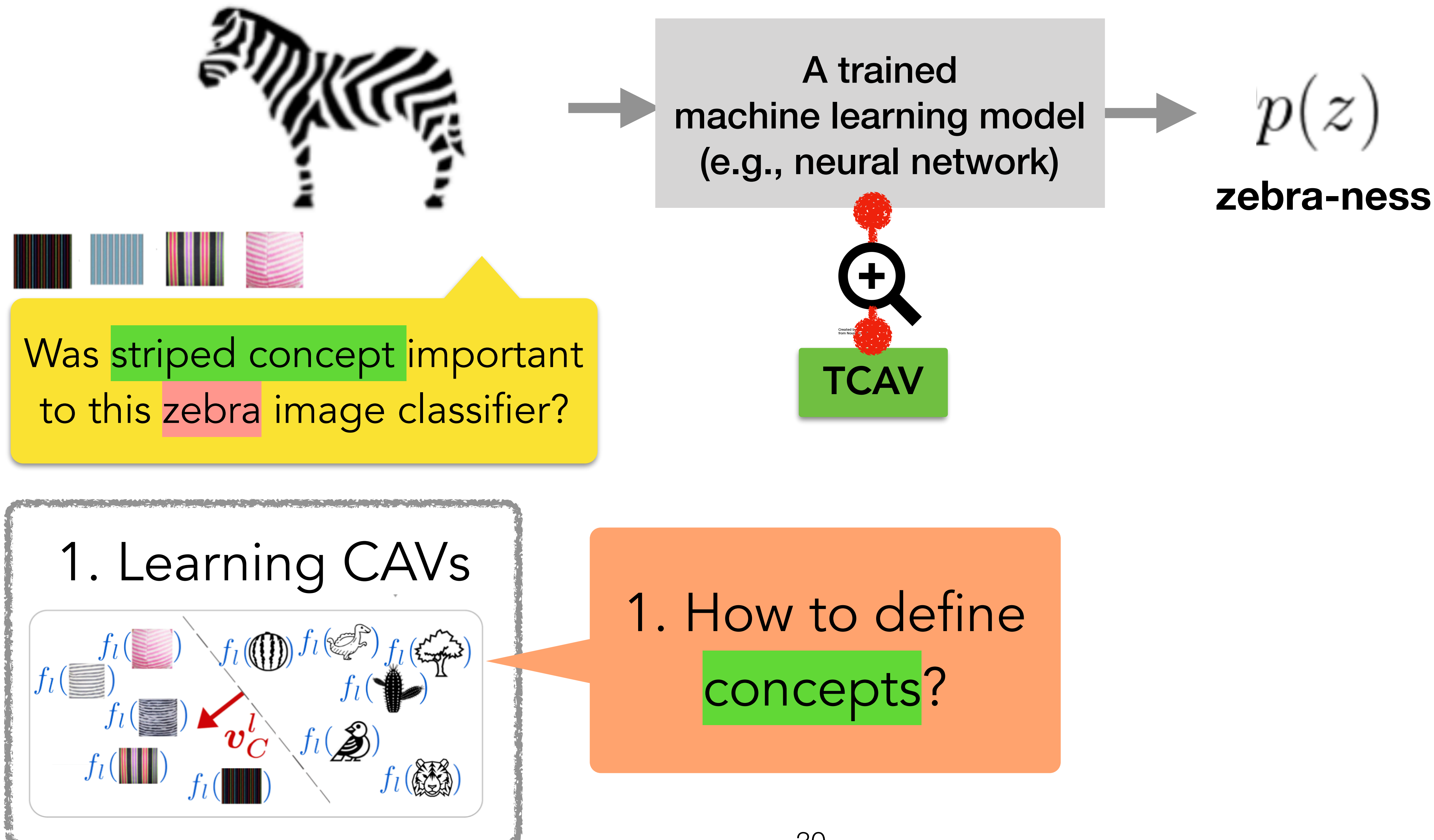
Was striped concept important
to this zebra image classifier?

TCAV score for **Zebra**



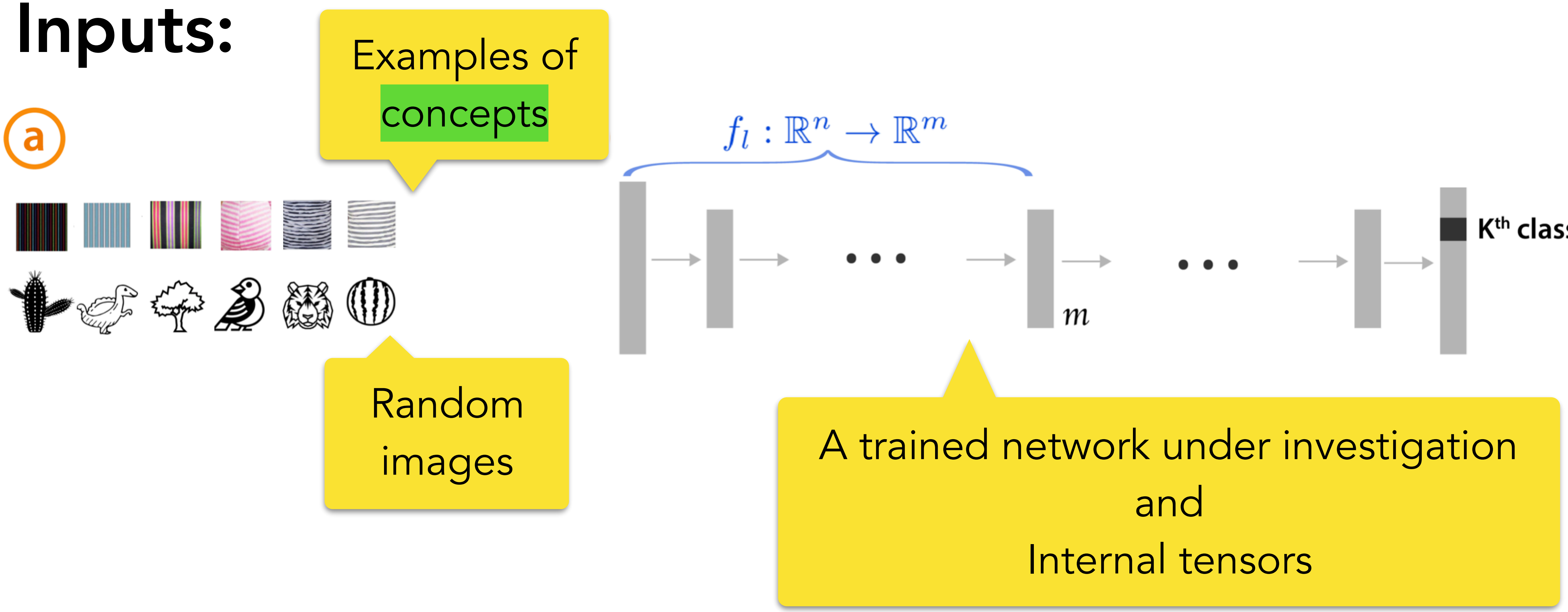
TCAV provides
quantitative importance of
a concept **if and only if** your
network learned about it.

TCAV:
Testing with Concept Activation Vectors



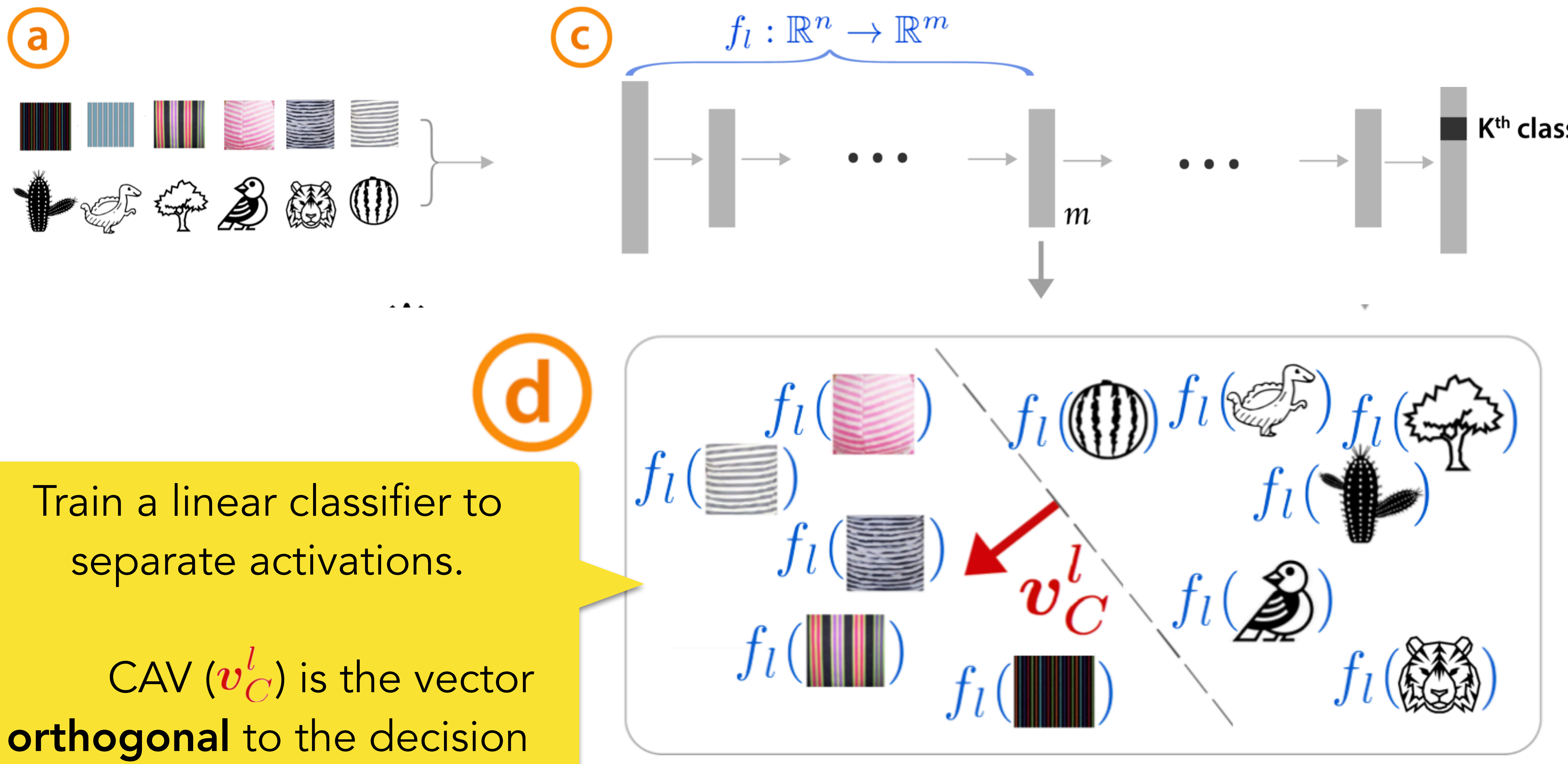
Defining concept activation vector (CAV)

Inputs:



Defining concept activation vector (CAV)

Inputs:

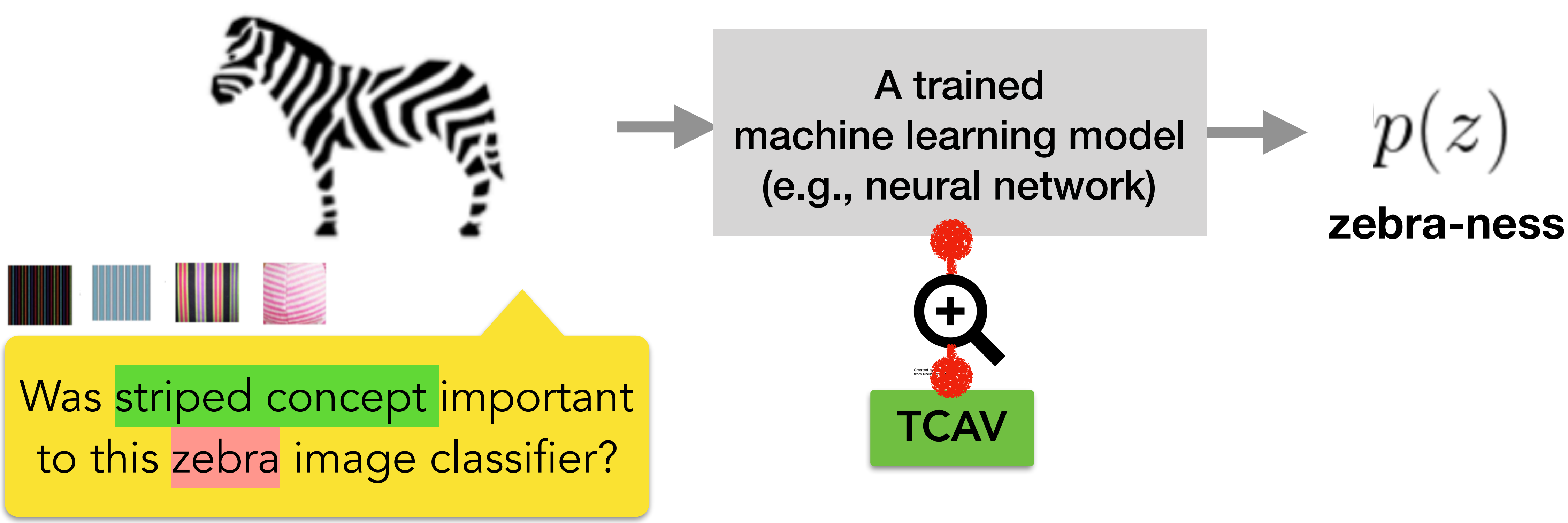


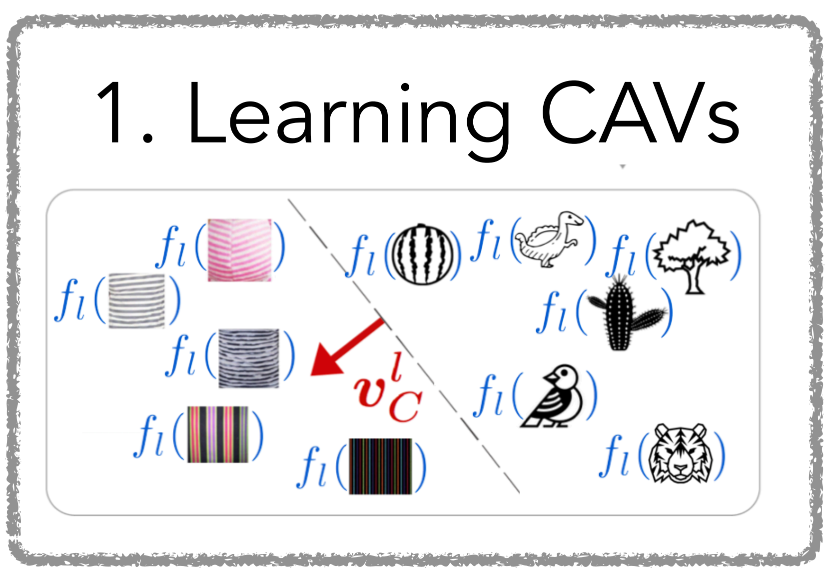
Train a linear classifier to separate activations.

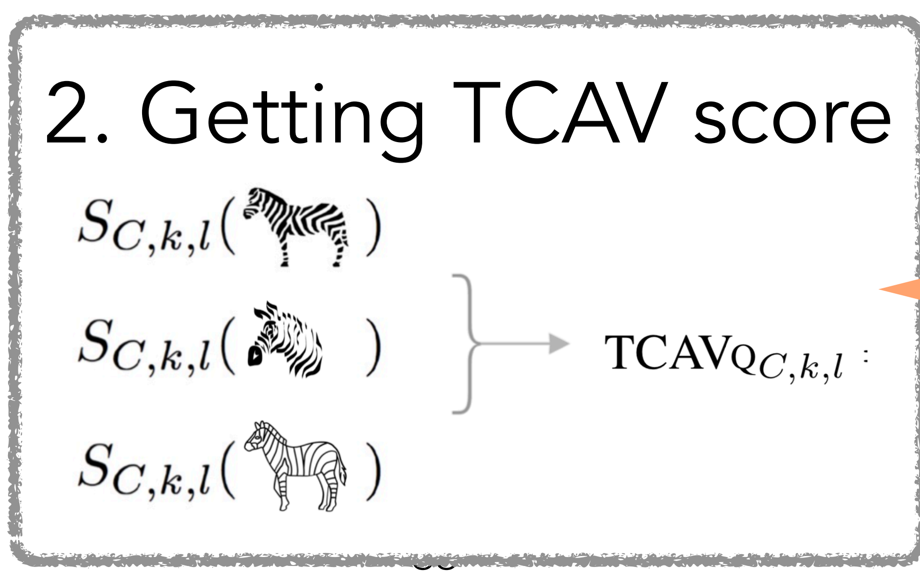
CAV (v_C^l) is the vector **orthogonal** to the decision boundary.

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

TCAV:
Testing with Concept Activation Vectors



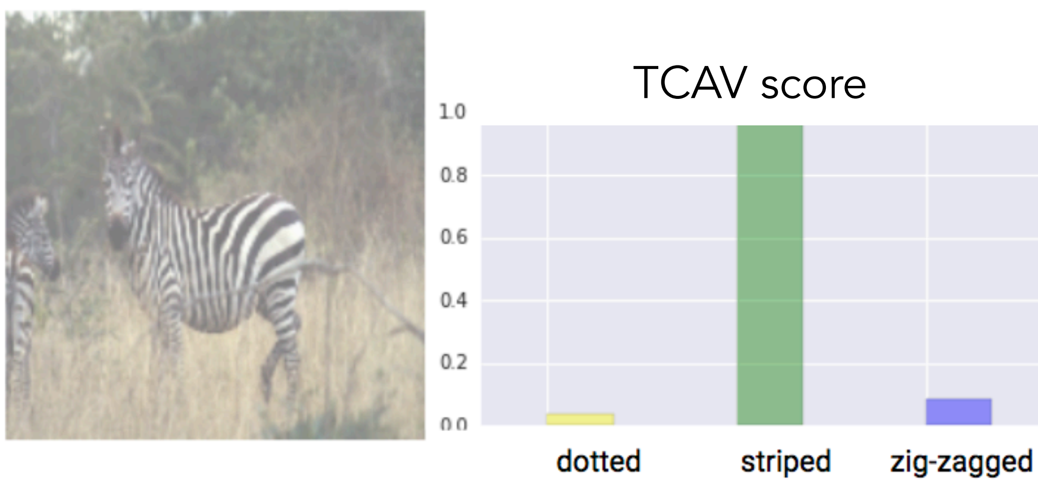




2. How are the CAVs useful to get explanations?

TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV



zebra-ness → striped CAV →

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

Directional derivative with CAV

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

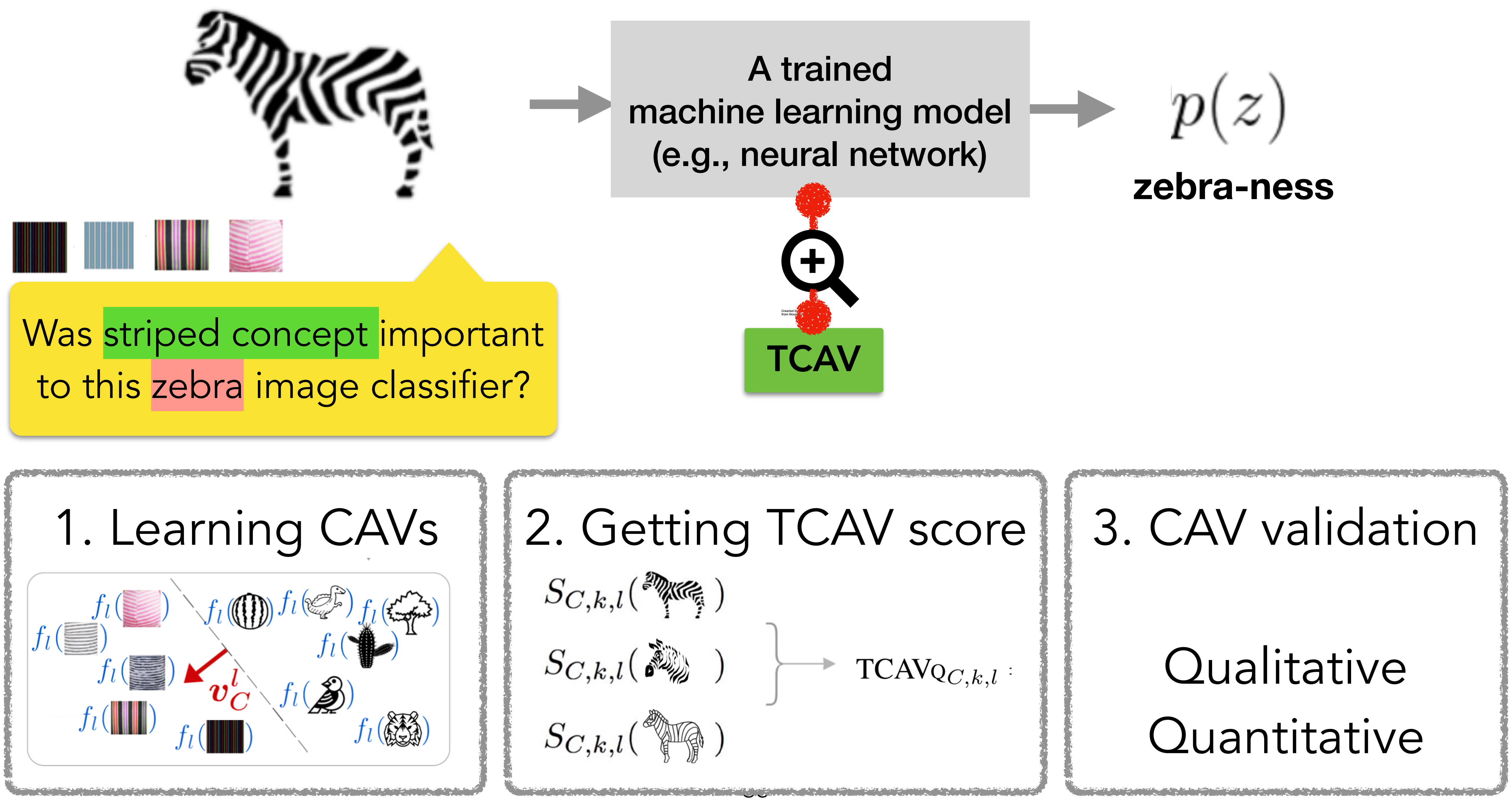
$S_{C,k,l}($  $)$

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

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

$$\boxed{\mathrm{TCAV}_{\mathrm{Q}_{C,k,l}} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}}$$

TCAV:
Testing with Concept Activation Vectors



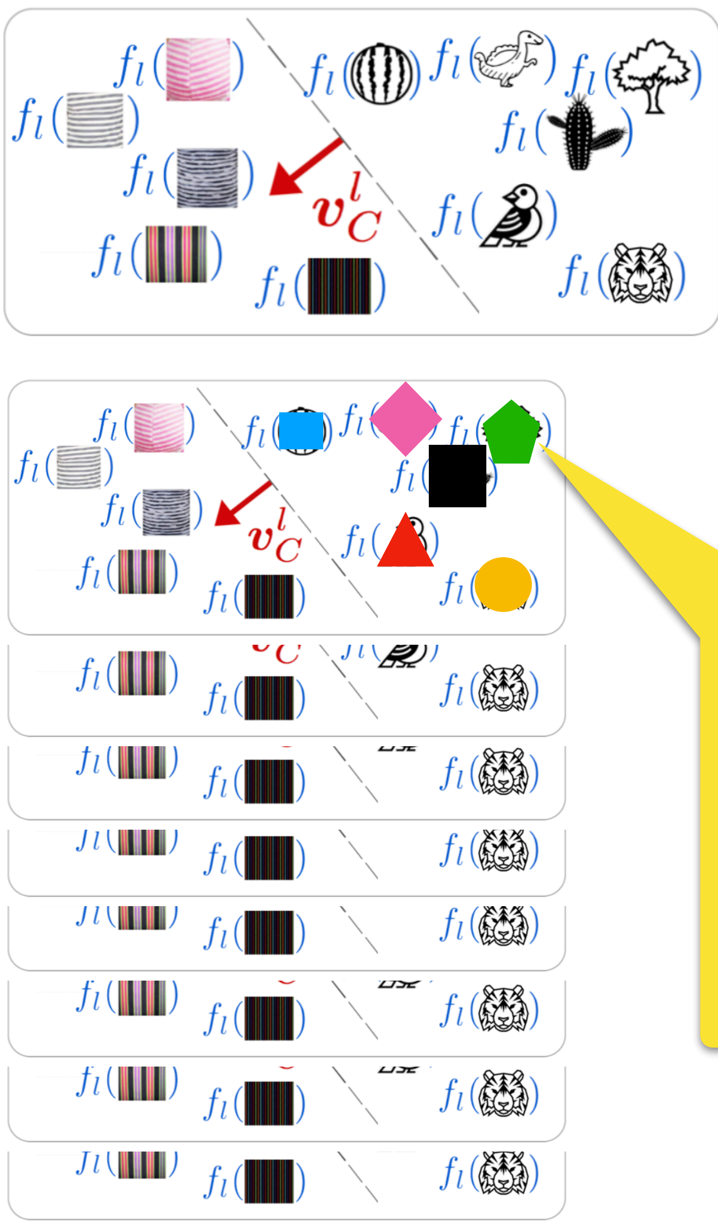
Quantitative validation:

Guarding against spurious CAV

Did my CAVs returned high sensitivity by chance?

Quantitative validation:

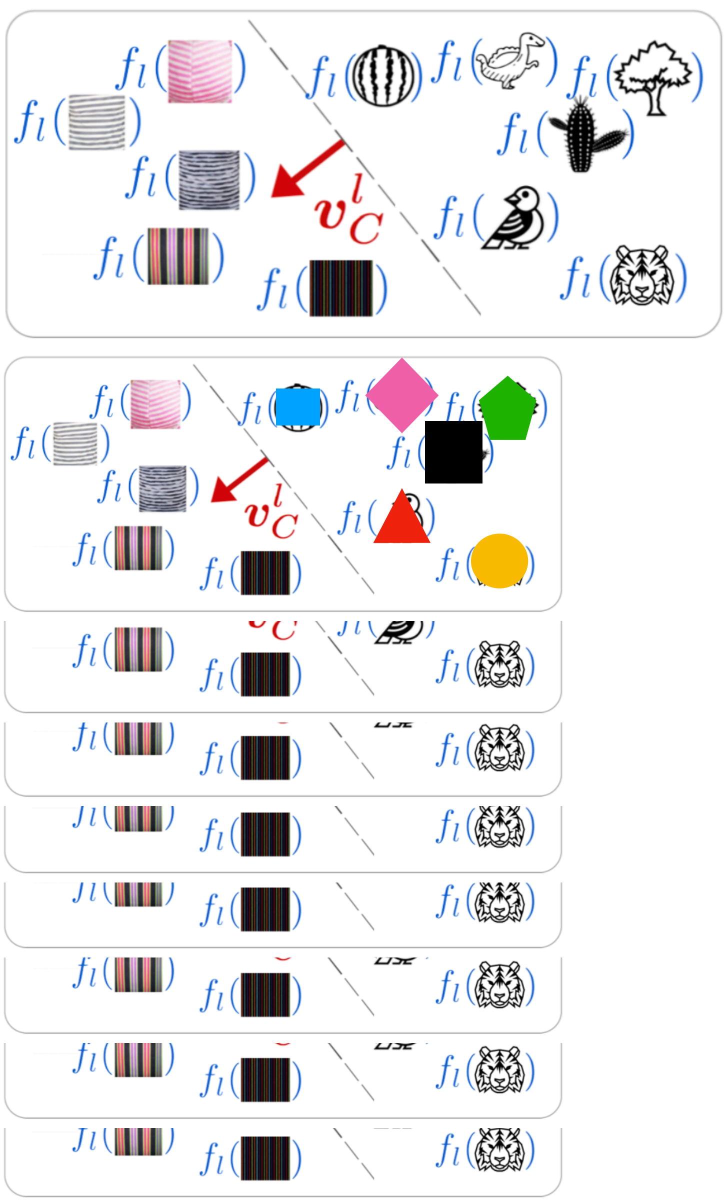
Guarding against spurious CAV



Learn many stripes CAVs using different sets of random images

Quantitative validation:

Guarding against spurious CAV



Zebra

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

⋮

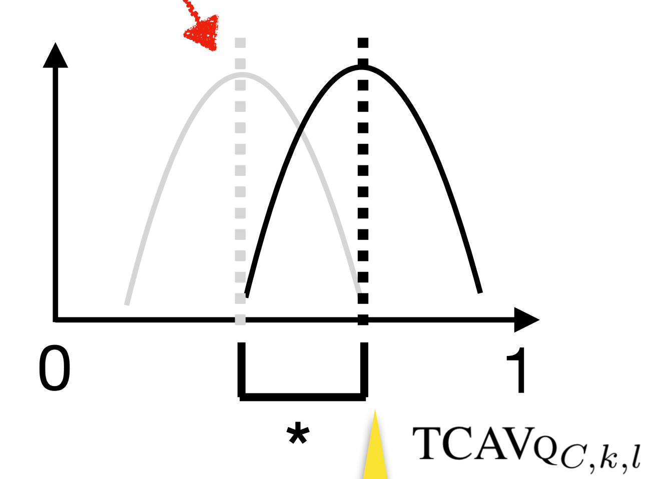
$\longrightarrow \mathrm{TCAVQ}_{C,k,l}$:

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

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

⋮

TCAV score random



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

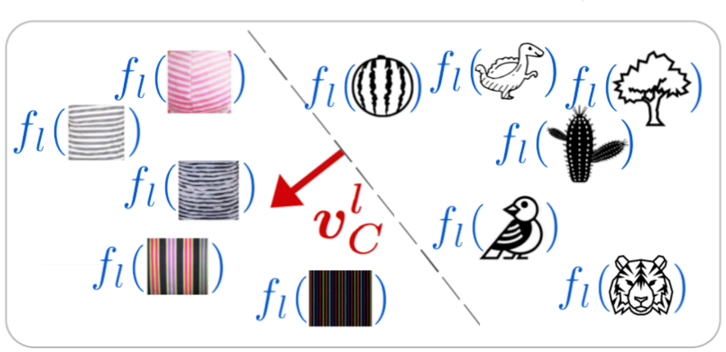
Recap TCAV:
Testing with Concept Activation Vectors

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

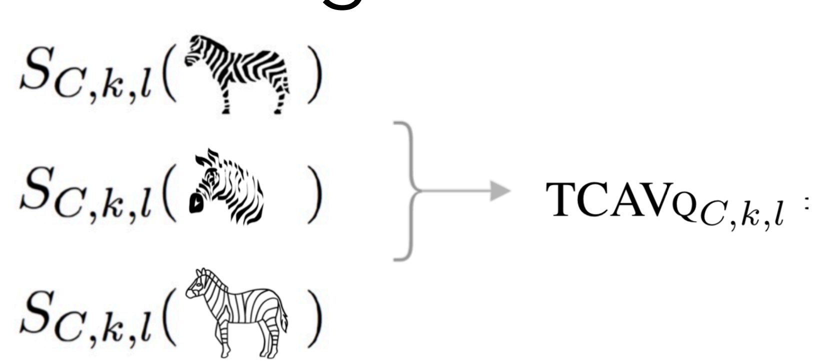
Even if your training data wasn't tagged with the concept

Even if your input feature did not include the concept

1. Learning CAVs



2. Getting TCAV score



3. CAV validation

Qualitative

Quantitative

Results

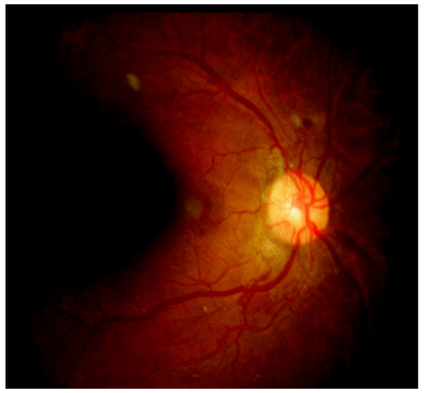
1. Sanity check experiment

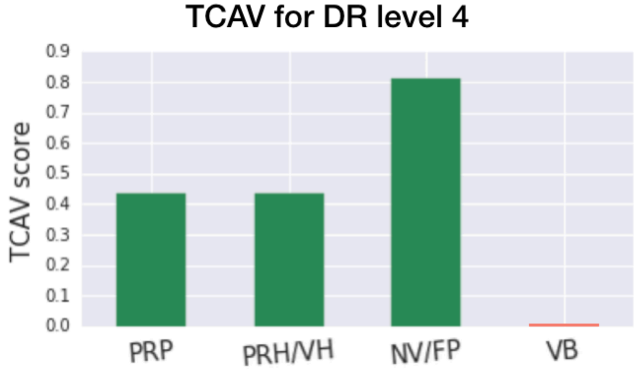
cab image   cab image with caption

2. Biases from Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy

DR level 4 Retina





Sanity check experiment

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

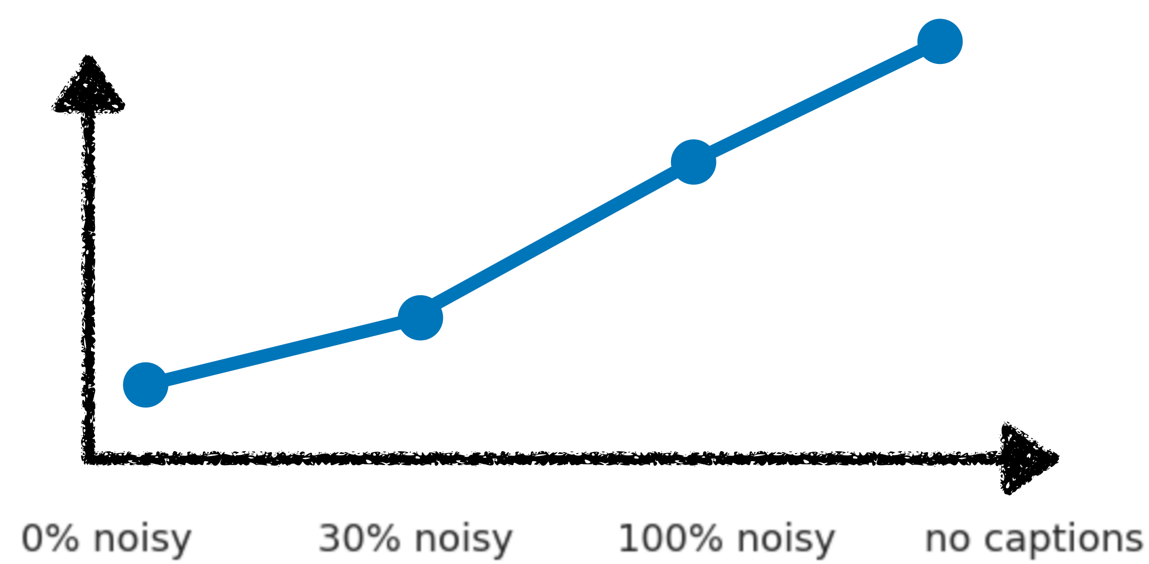
Sanity check experiment setup

caption concept

image concept

models can use either image or caption concept for classification.

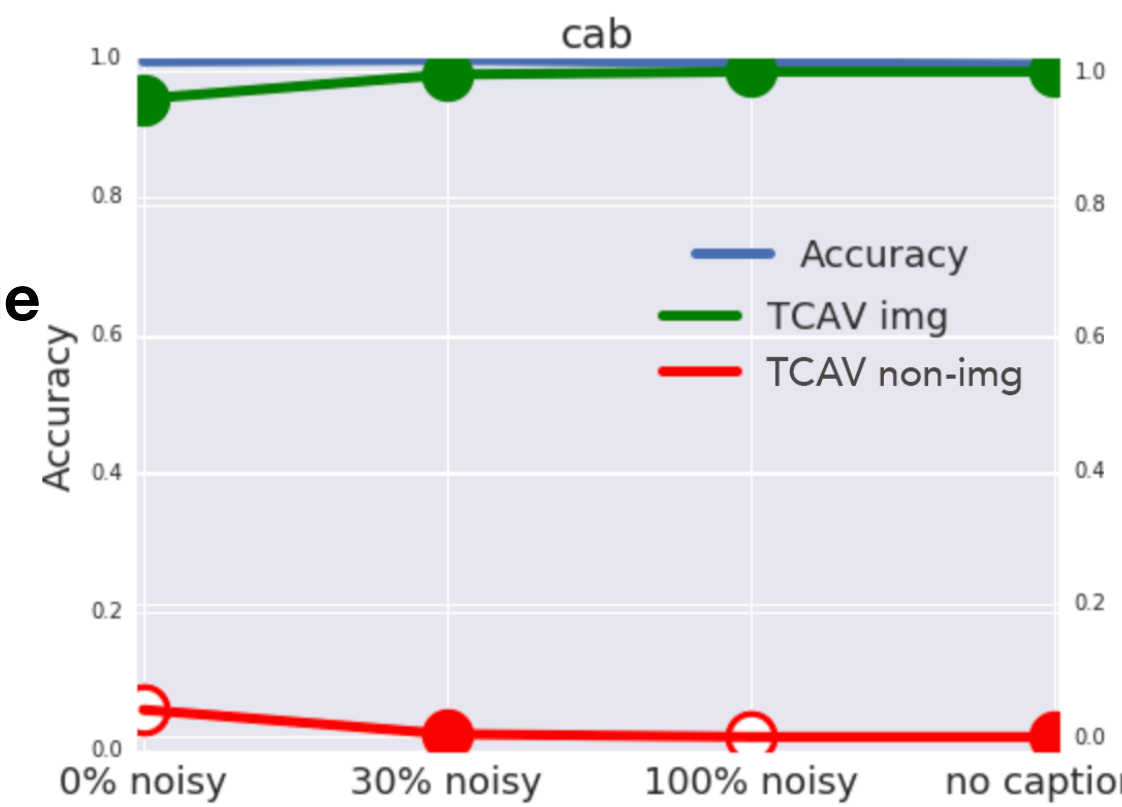
Test accuracy with no caption image = Importance of image concept



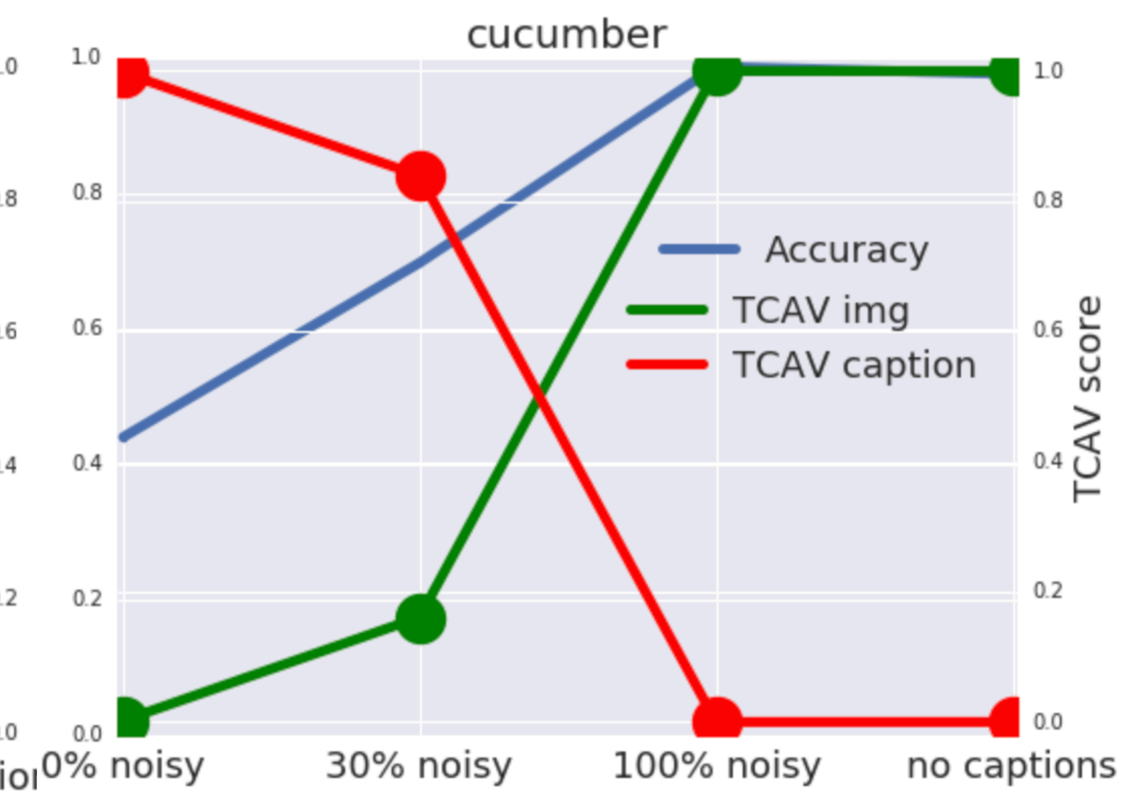
Caption noise level in training set of each model

Sanity check experiment

Test accuracy with no caption image



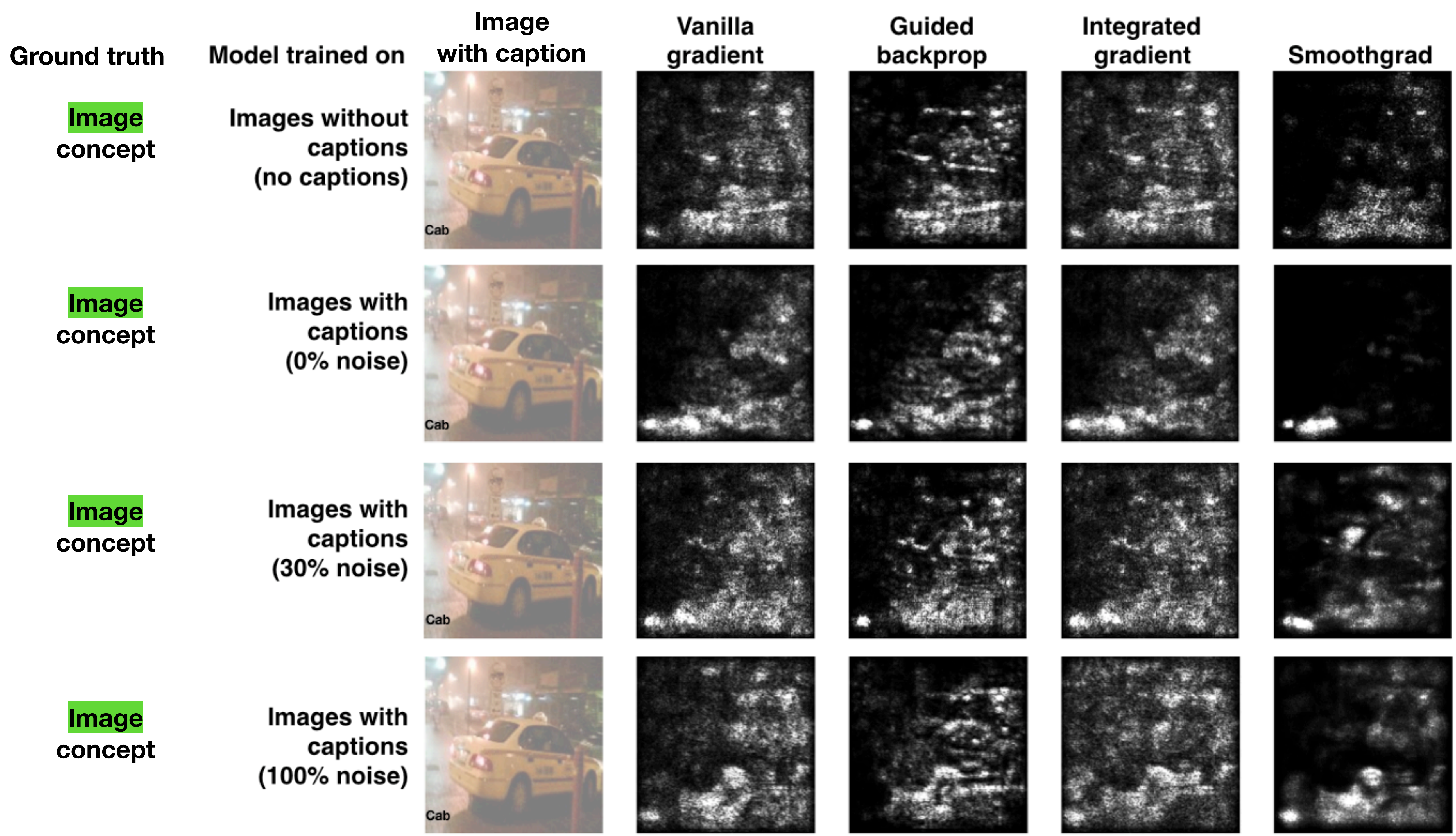
Caption noise level in training set



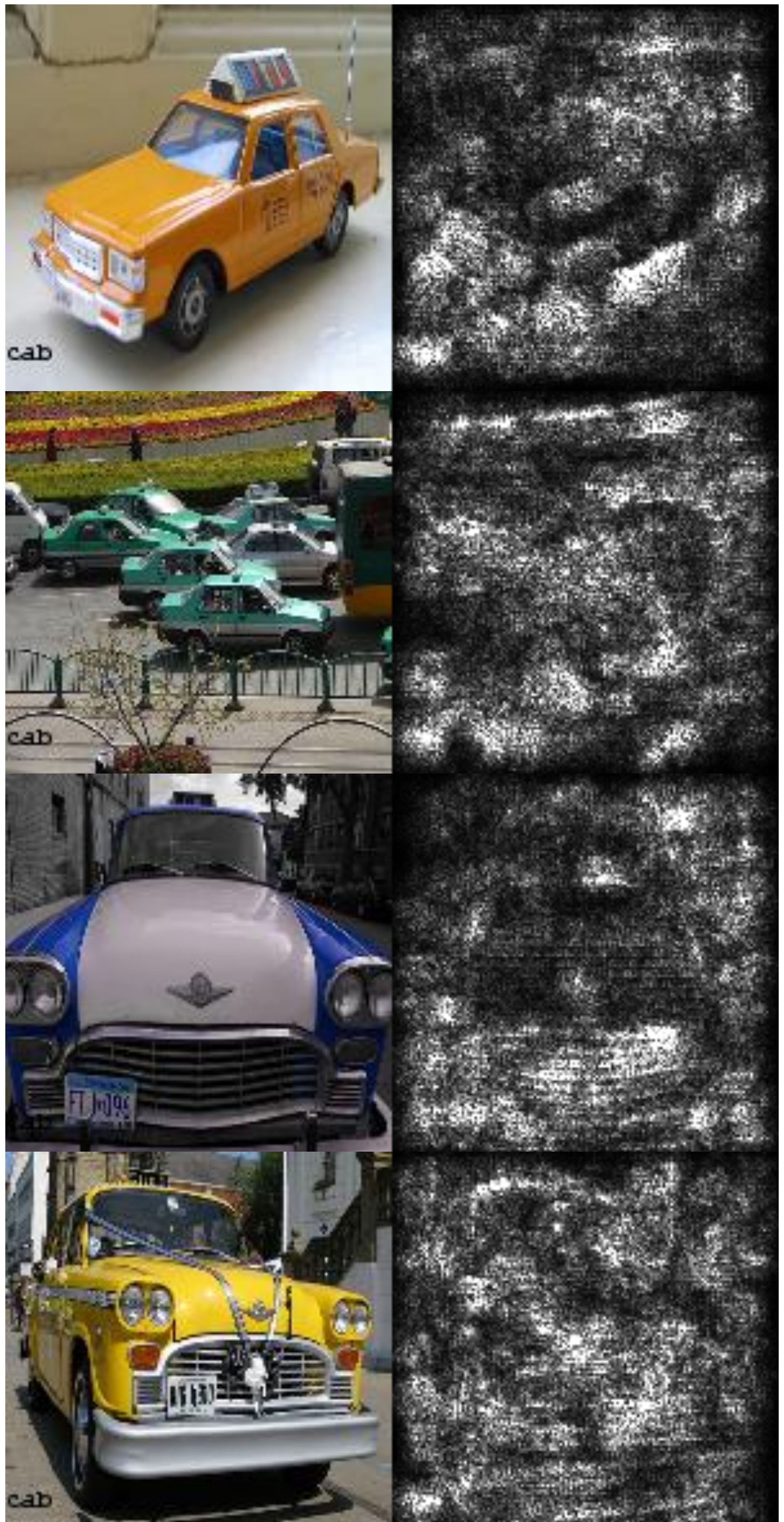
Caption noise level in training set

Cool, cool.
Can saliency maps do this too?

Can saliency maps communicate the same information?



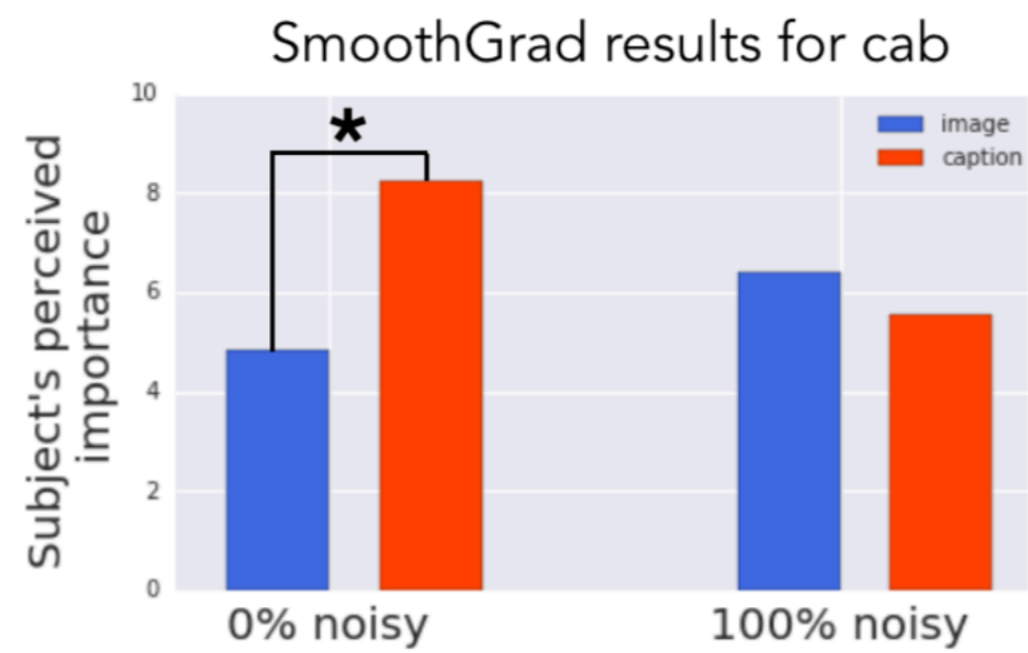
Human subject experiment: Can saliency maps communicate the same information?

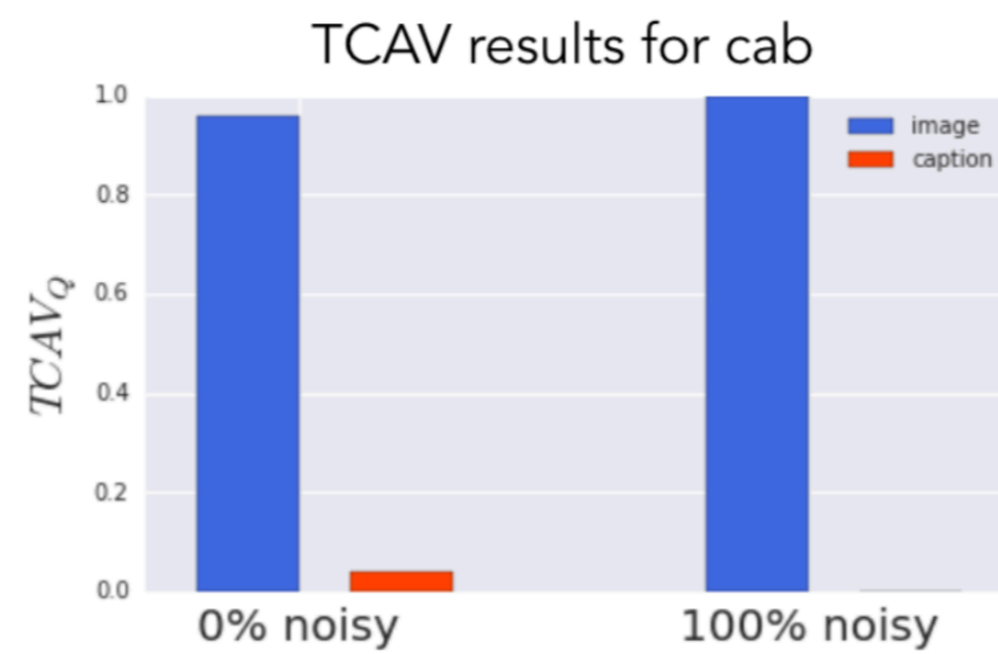


- 50 turkers are
  - asked to judge importance of image vs. caption given saliency maps.
  - asked to indicate their confidence
  - shown 3 classes (cab, zebra, cucumber) x 2 saliency maps for one model

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

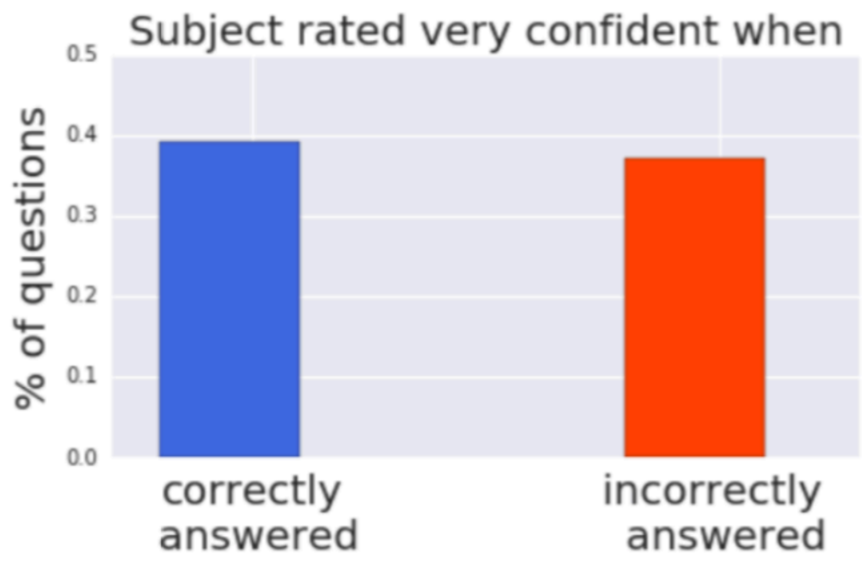
- Random chance: 50%
- Human performance with saliency map: 52%
- Humans can't agree: more than 50% no significant consensus





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.

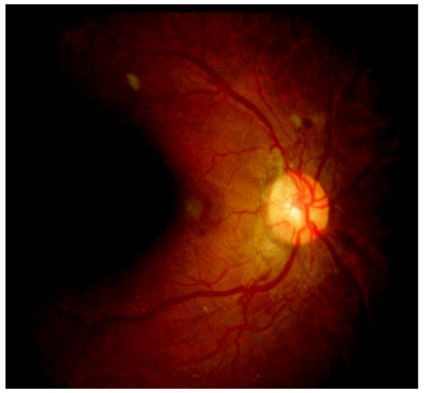


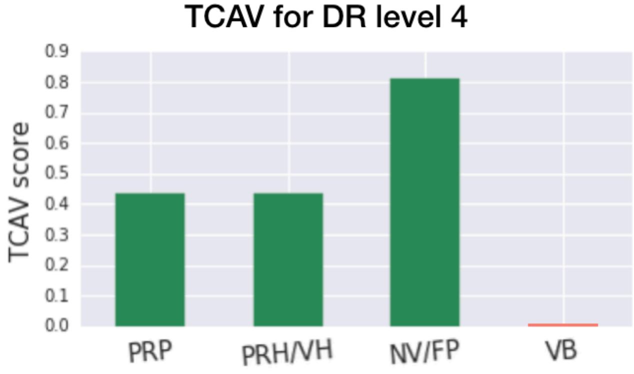
Results

cab image cab image with caption

1. Sanity check experiment
2. Biases from Inception V3 and GoogleNet
3. Domain expert confirmation from Diabetic Retinopathy

DR level 4 Retina

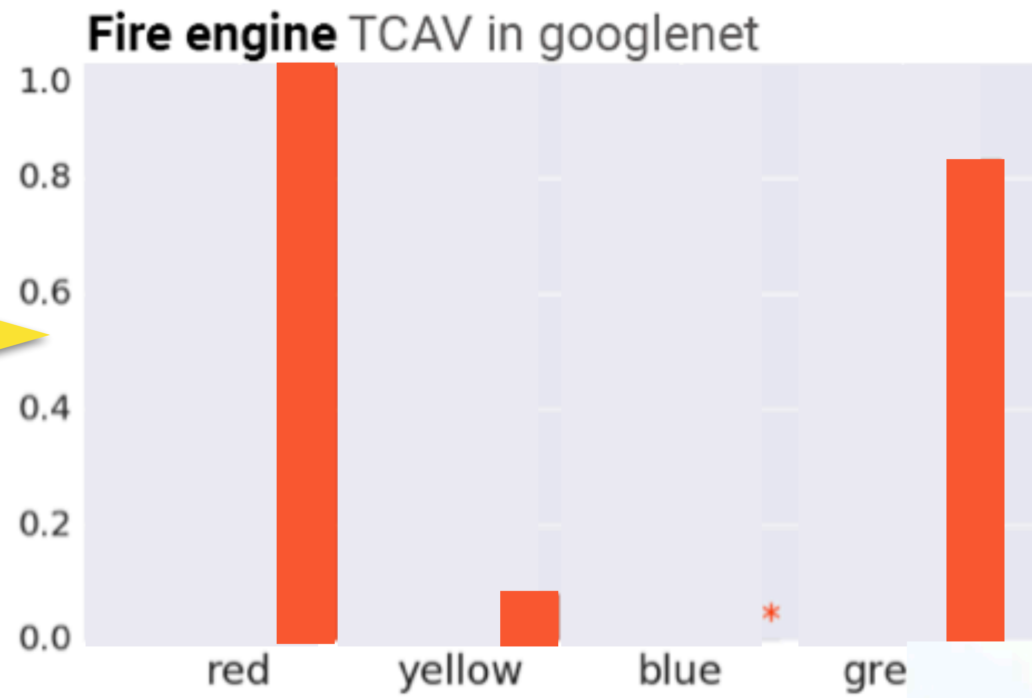


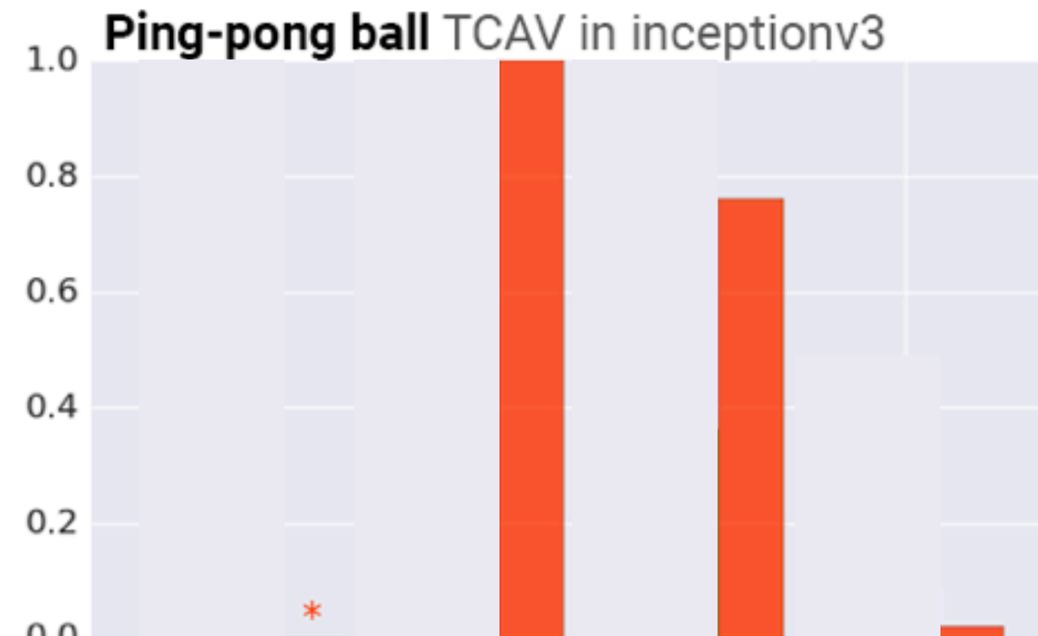


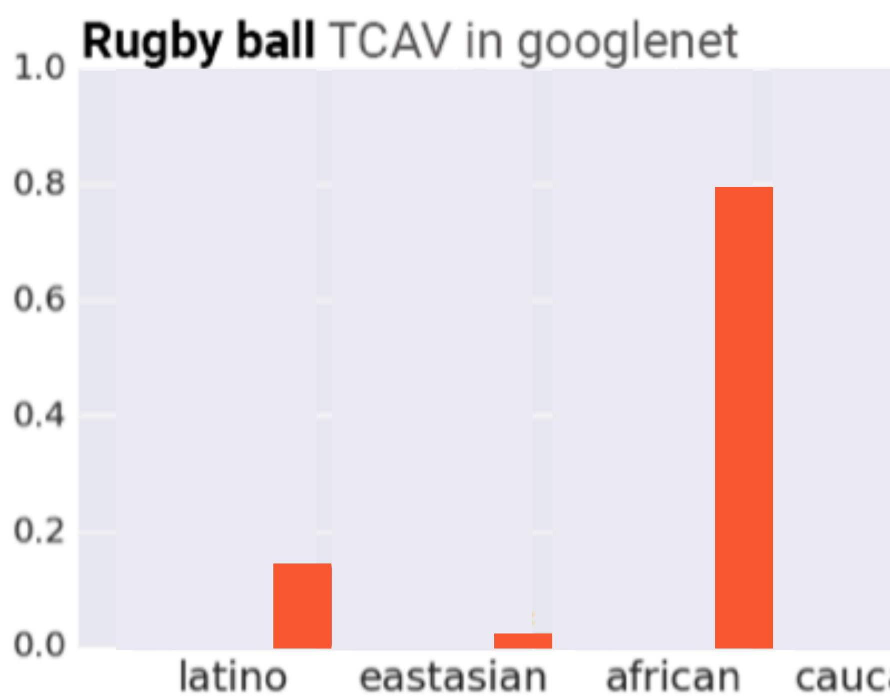
TCAV in

Two widely used image prediction models

Geographical bias!



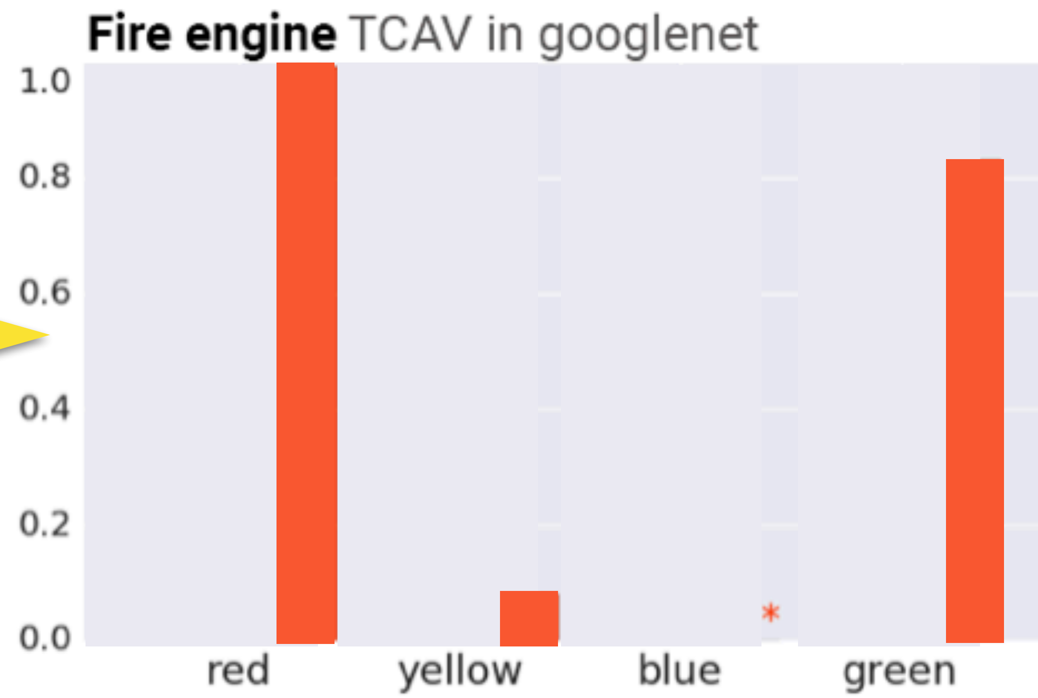


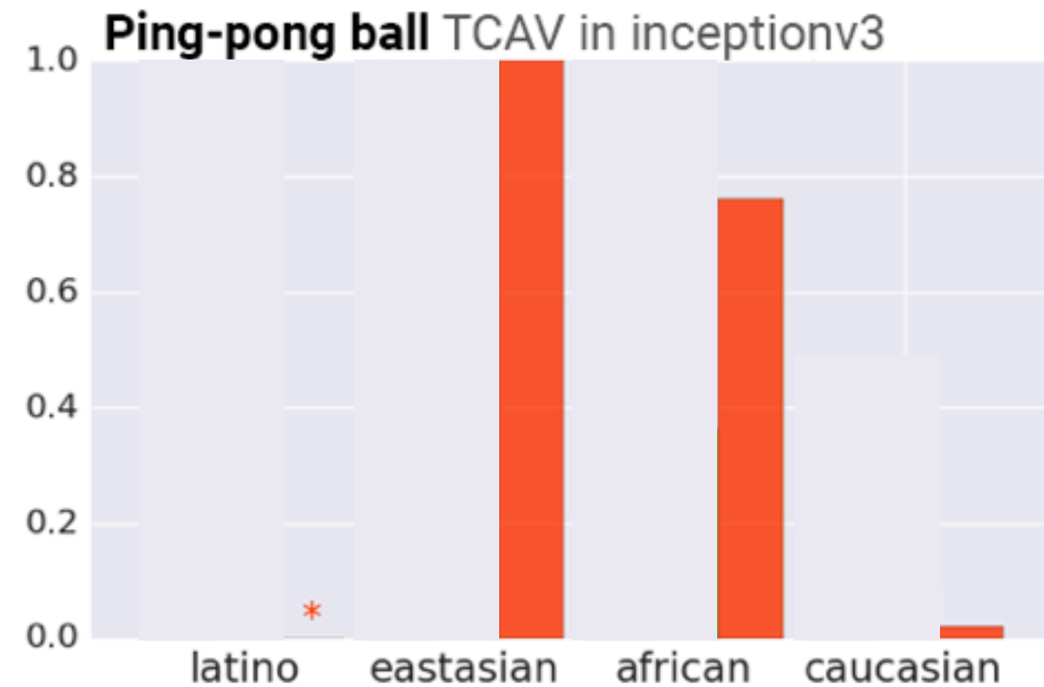


TCAV in

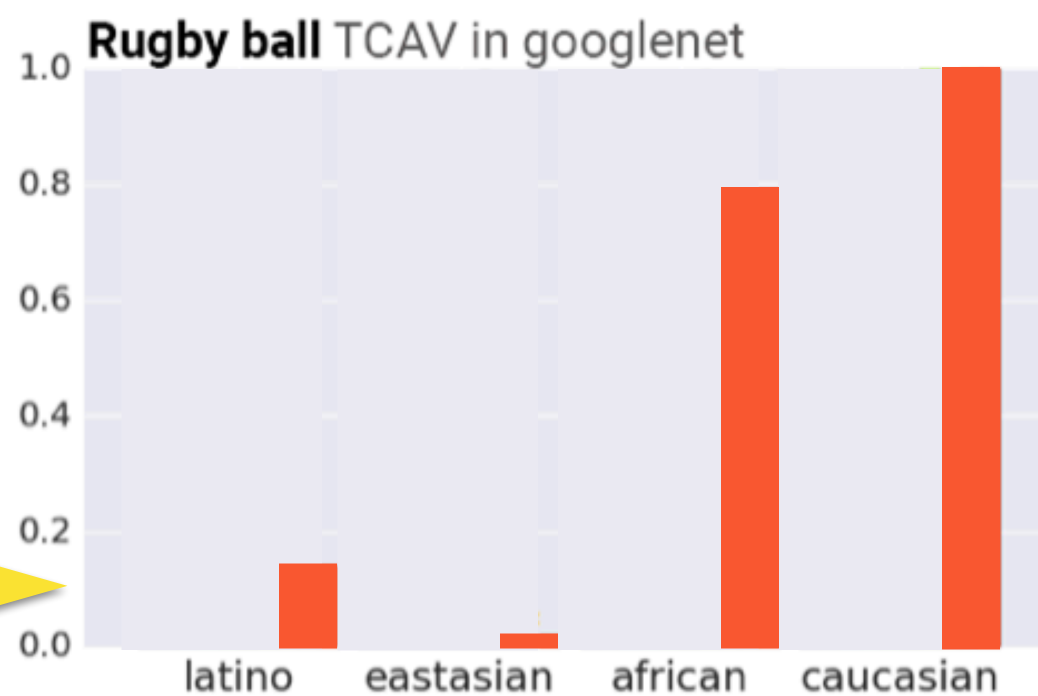
Two widely used image prediction models

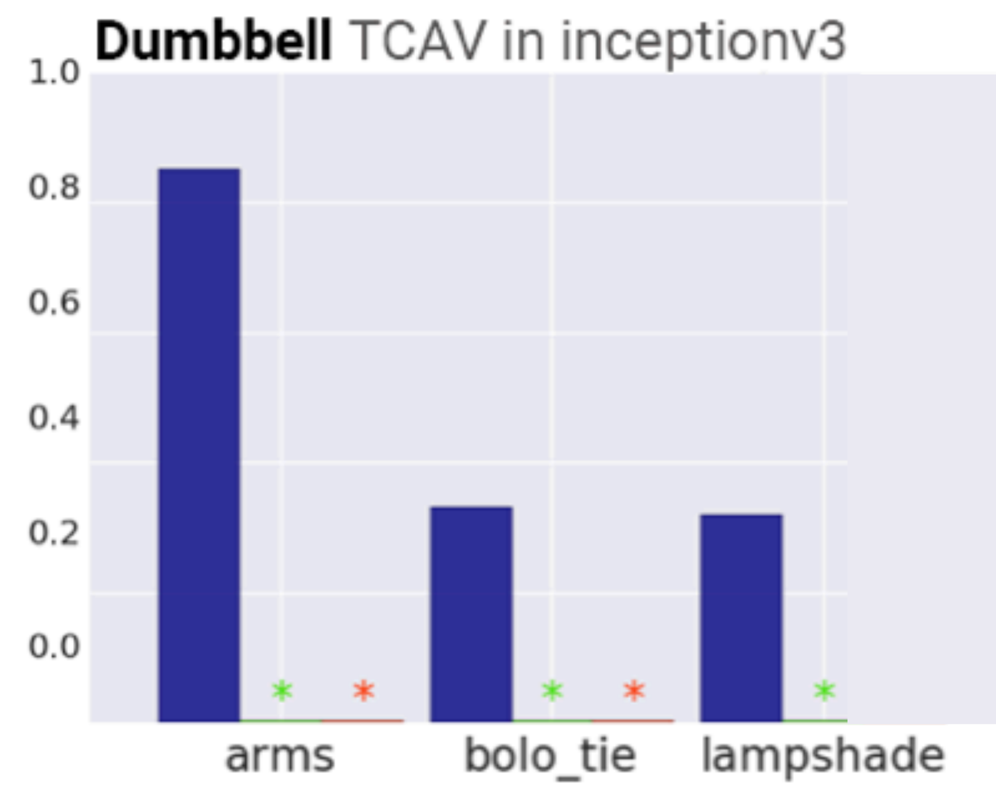
Geographical bias?





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

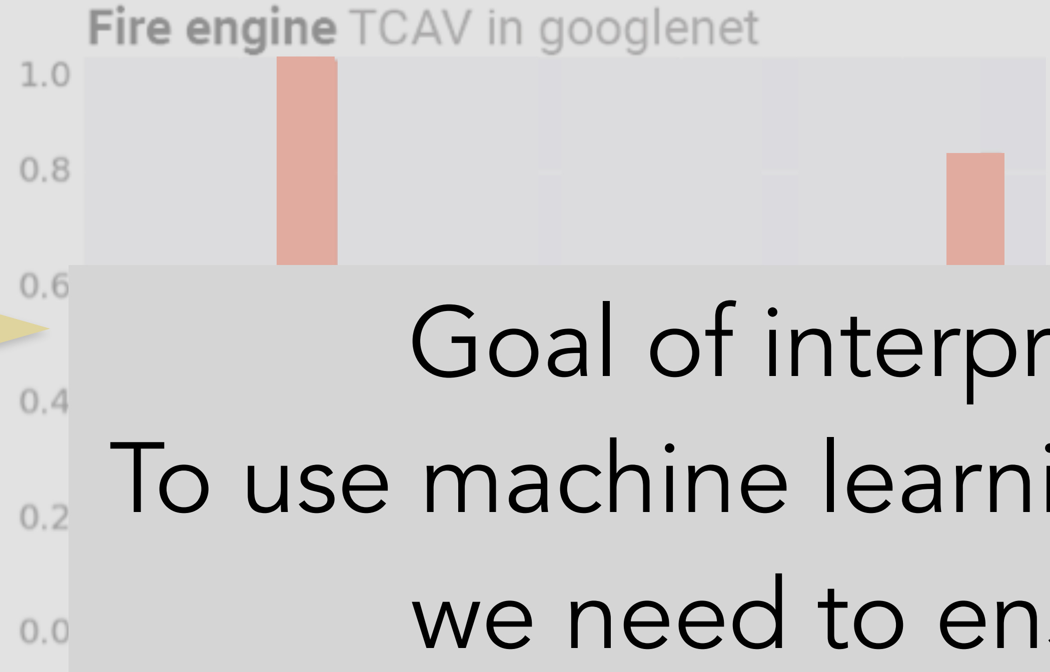




TCAV in

Two widely used image prediction models

Geographical bias?



Goal of interpretability:

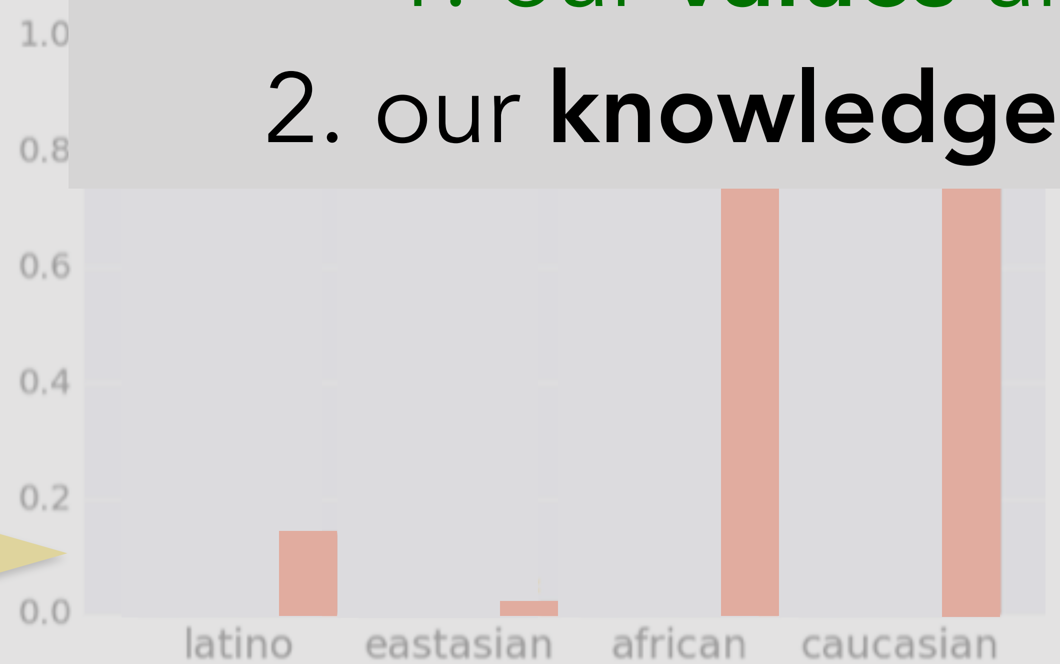
To use machine learning **responsibly**

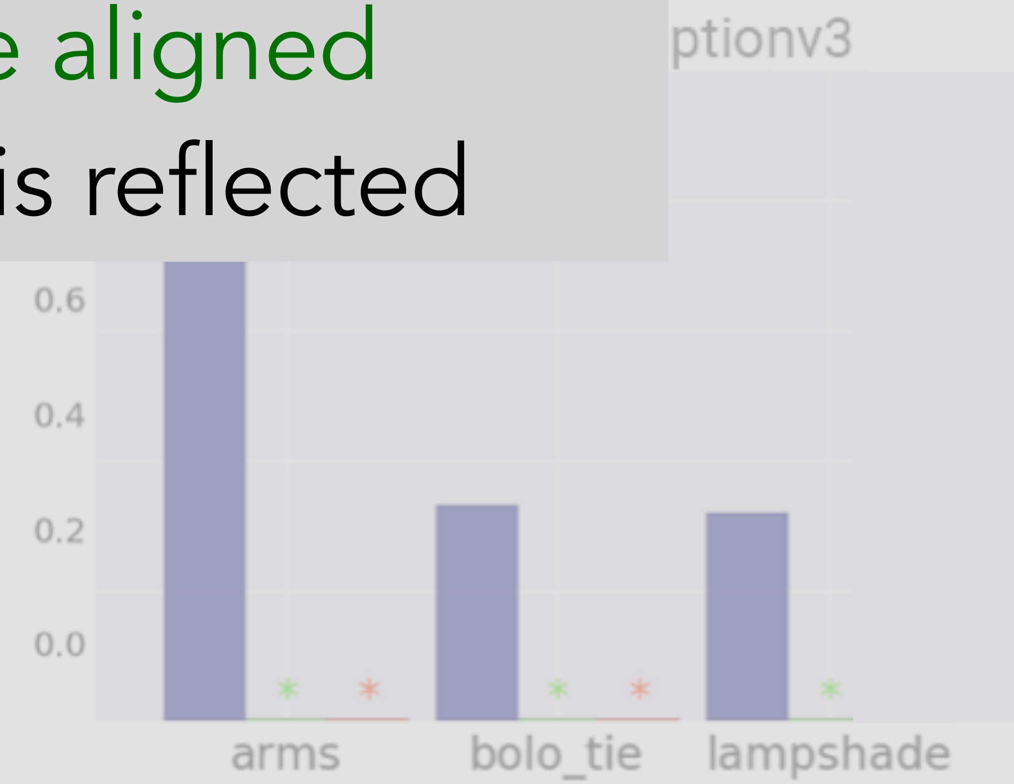
we need to ensure that

1. our **values** are aligned

2. our **knowledge** is reflected

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



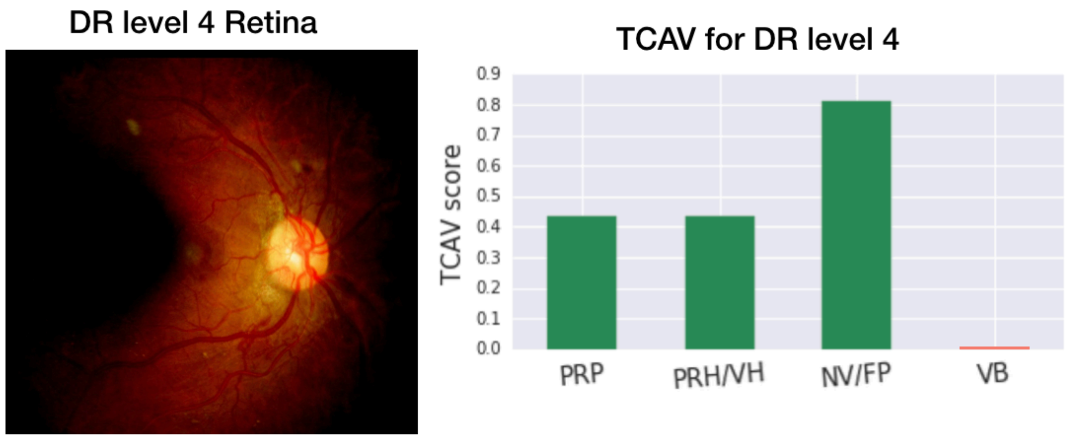


Results

1. Sanity check experiment

2. Biases Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy



Diabetic Retinopathy

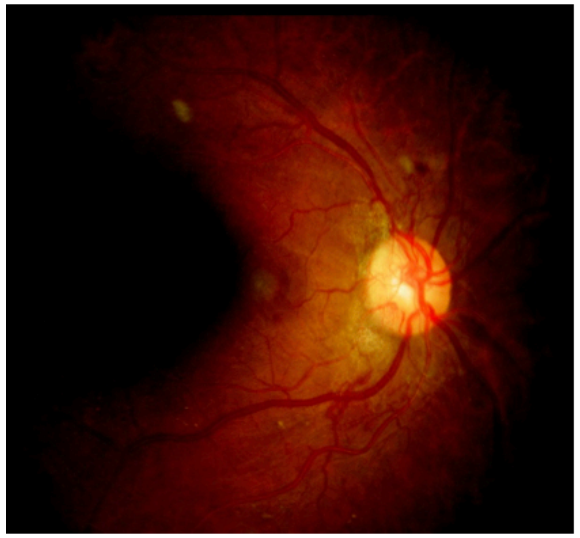
- Treatable but sight-threatening conditions
- Have model to with accurate prediction of DR (85%) [Krause et al., 2017]

Concepts the **ML model** uses

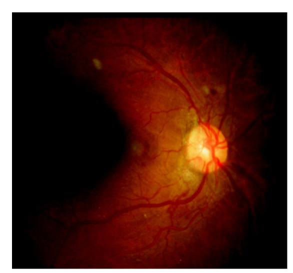
Vs

Diagnostic Concepts **human** doctors use

DR level 4 Retina



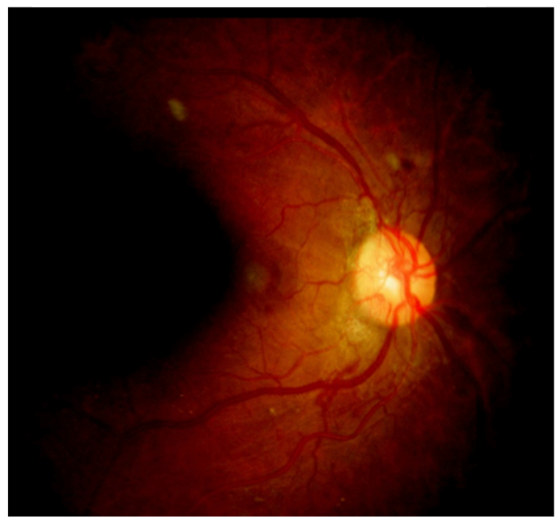
Collect human doctor's knowledge

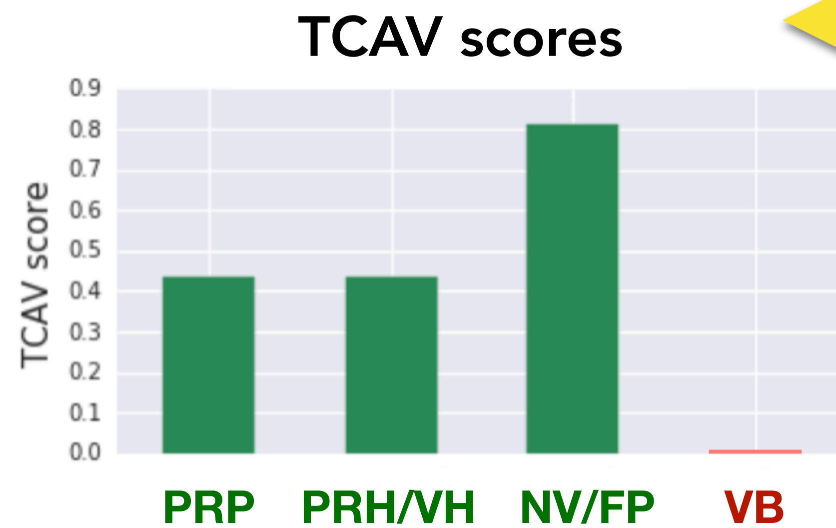
	Concepts belong to this level	Concepts do not belong to this level
DR level 4 	**PRP** **PRH/VH** **NV/FP**	**VB**
DR level 1 	**MA**	**HMA**

TCAV for Diabetic Retinopathy

Prediction class | **Prediction accuracy** | **Example**

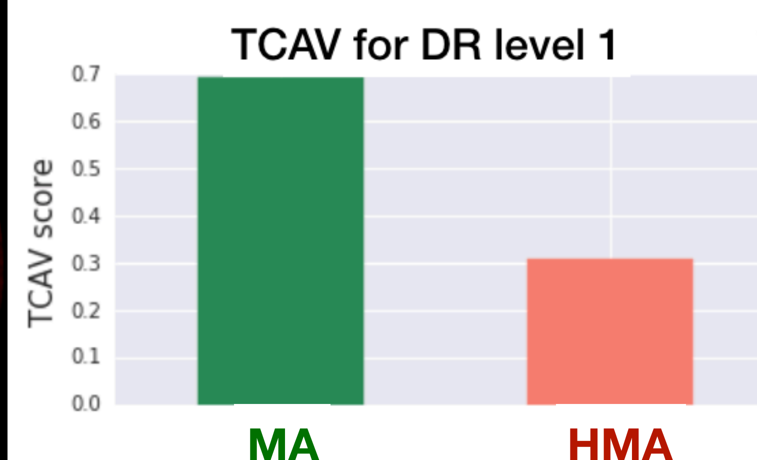
DR level 4 | **High**





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

DR level 1 | **Med**

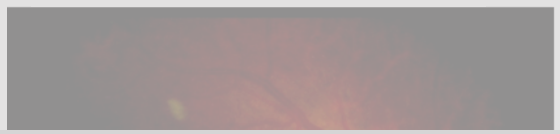


TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate

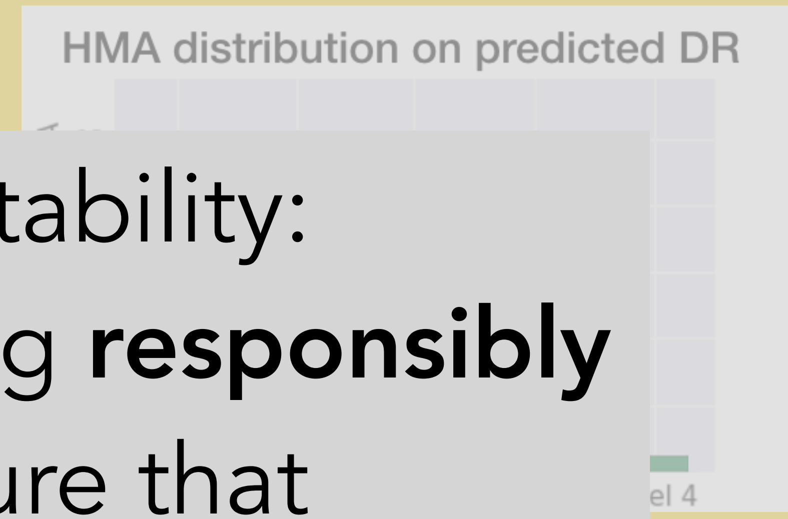
Green: domain expert's label on concepts belong to the level
Red: domain expert's label on concepts does not belong to the level

TCAV for Diabetic Retinopathy

Prediction class | Prediction accuracy | Example



Level 1 was often confused to level 2.



DR level 4 Hi

Goal of interpretability:

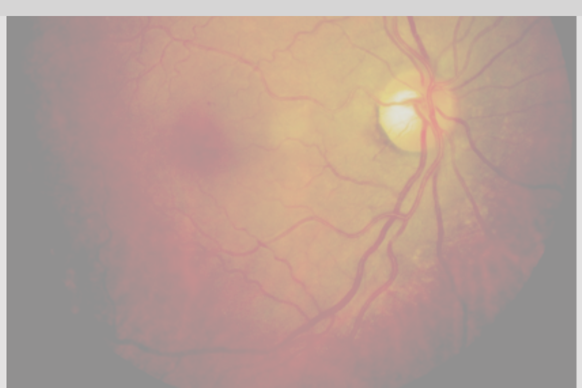
To use machine learning **responsibly**

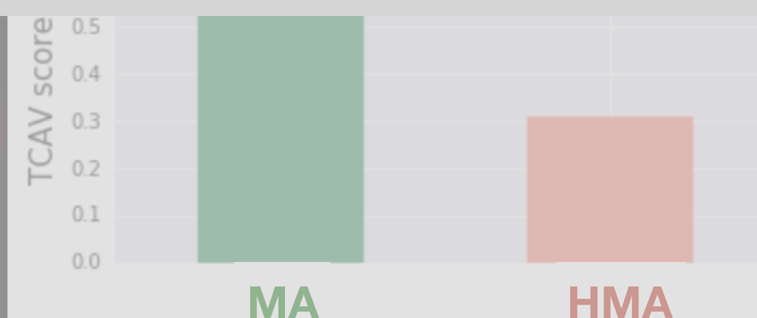
we need to ensure that

1. our **values** are aligned

2. our **knowledge** is reflected

DR level 1 Low





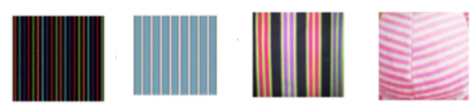
TCAV shows the del is **inconsistent** with doctor's knowledge for classes when model is less accurate

Green: domain expert's label on concepts belong to the level

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

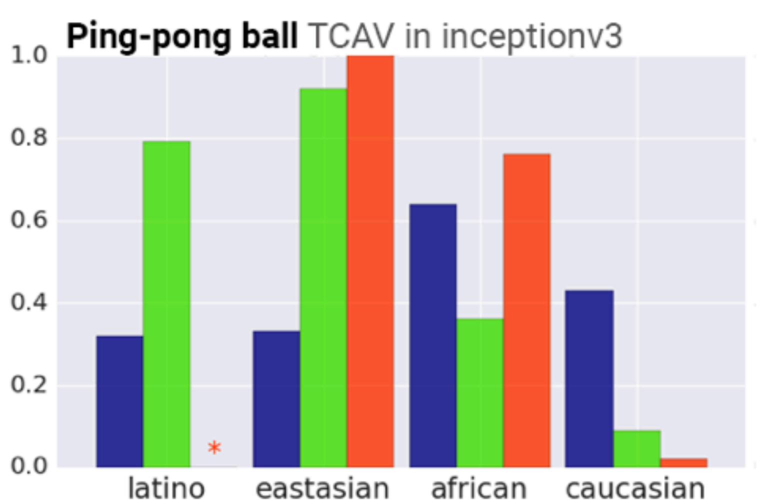
Summary:
Testing with Concept Activation Vectors

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



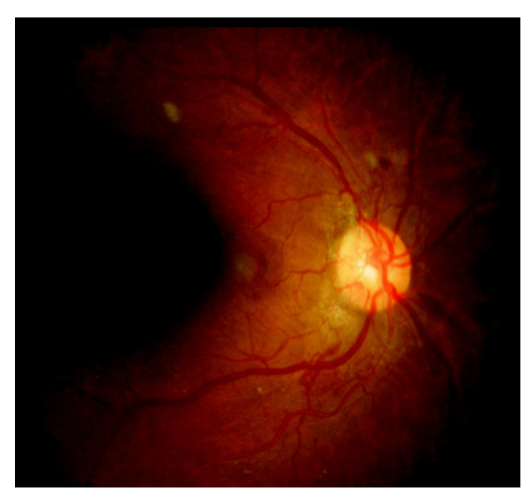
stripes concept (score: 0.9) was important to **zebra** class for this trained network. 

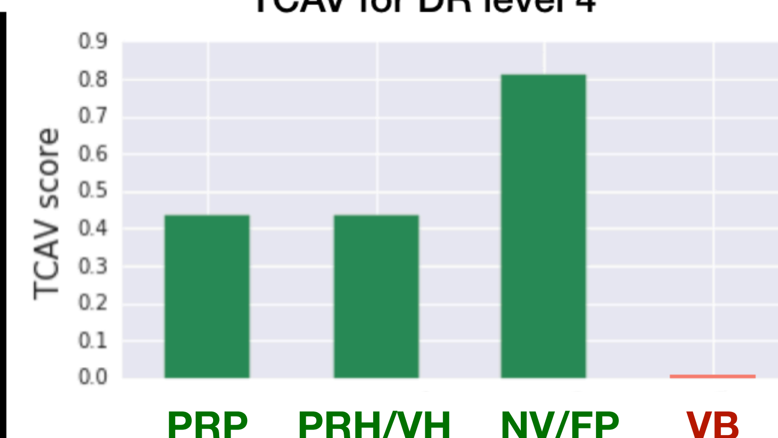
TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.



Our values

DR level 4 Retina





Our knowledge

Agenda

1. Revisit some existing methods: Sanity check questions

2. Make explanations that work for lay people.

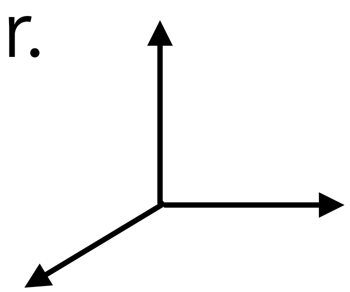
$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.

What makes explanations hard or easy for humans?

- How do humans' understanding changes as we vary factors in explanations? $\underset{E}{\operatorname{argmax}}\ Q(E|M, H, D, T)$
- Among many explanations, we choose a rule-set.
- Among many factors, we choose a subset based on what prior literatures assumed to matter.

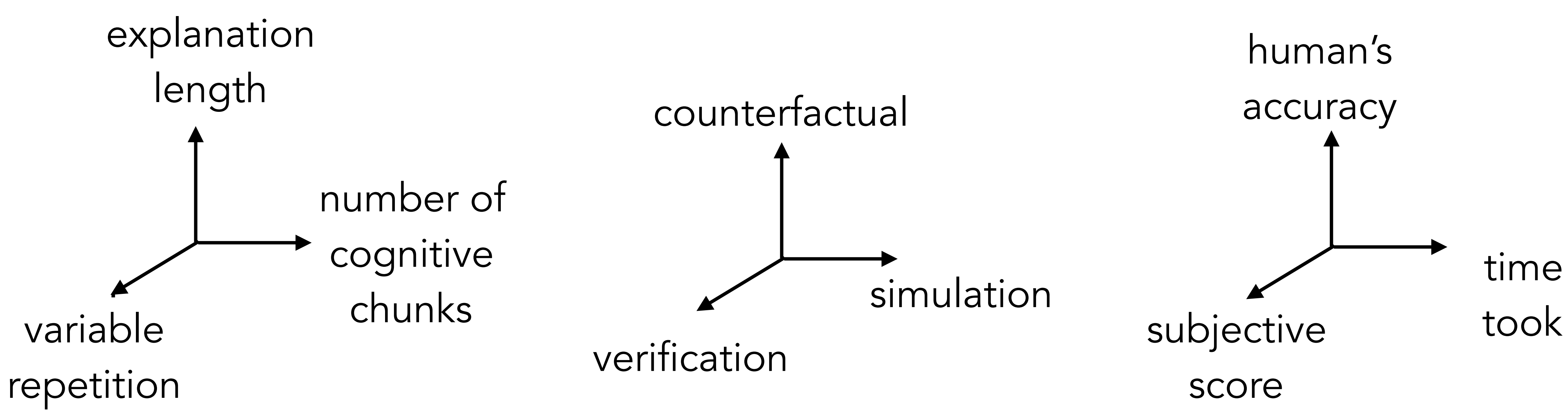


factors

How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation
joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017

What makes explanations hard or easy for humans?

- How do humans' understanding changes as we vary factors in explanations? $\underset{E}{\operatorname{argmax}}\ Q(E|M,H,D,T)$



we vary these factors

humans do these tasks

measures of interpretability

How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation
joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017

Controlling for prior knowledge.

The alien's preferences:

checking the news and coughing → windy
snowing or humid and weekend → spices or vegetables and grains
embarrassed and grouchy or raining → dairy or vegetables
snowing or windy and energetic → candy or dairy and fruit
grouchy or weekend and windy → spices or grains and fruit

Using a made-up 'alien world' to control for prior knowledge.

Is the alien happy with his meal?

○ Yes ○ No

Observations: Saturday, coughing, checking the news

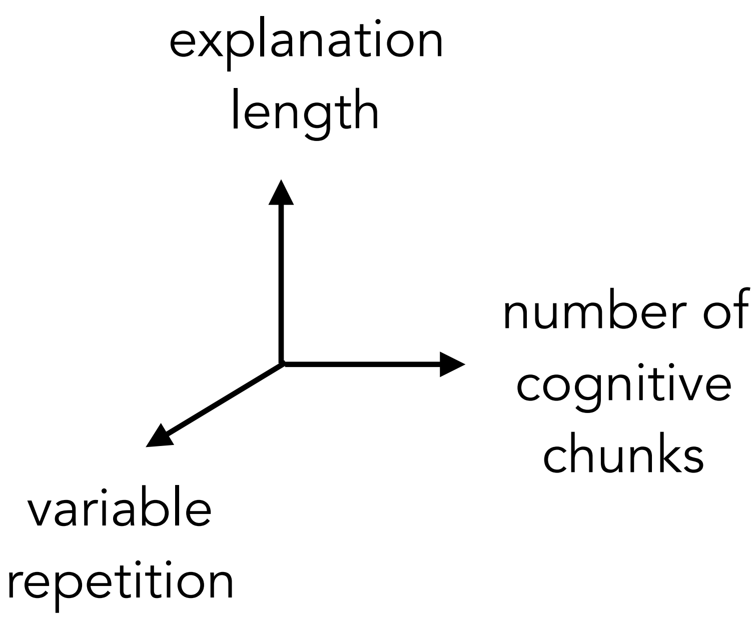
Recommendation: bagel, rice, strawberry

Ingredients:

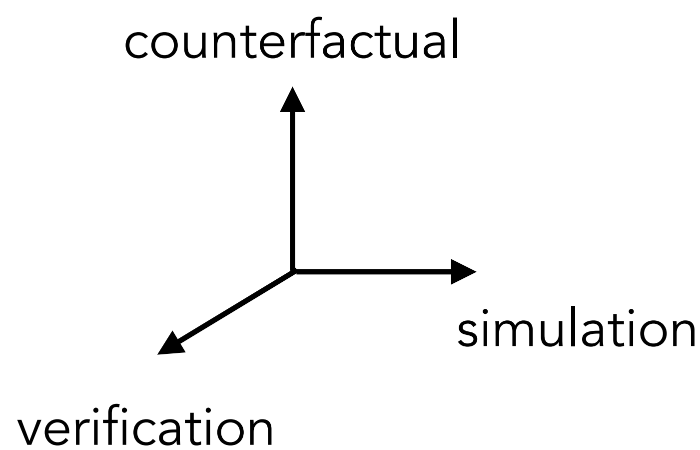
- **Vegetables:** okra, carrot, spinach
- **Spices:** turmeric, thyme, cinnamon
- **Dairy:** milk, butter, yogurt
- **Fruit:** mango, strawberry, guava
- **Candy:** chocolate, taffy, caramel
- **Grains:** bagel, rice, pasta

How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation
joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017

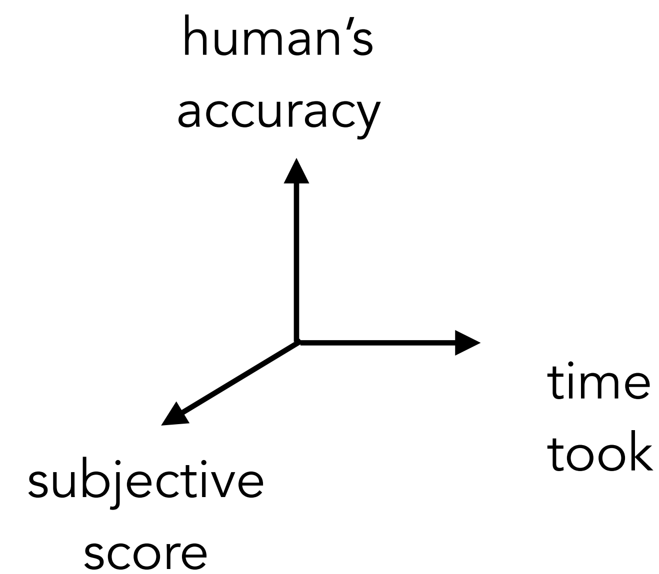
(a small subset of) results



we vary
these factors



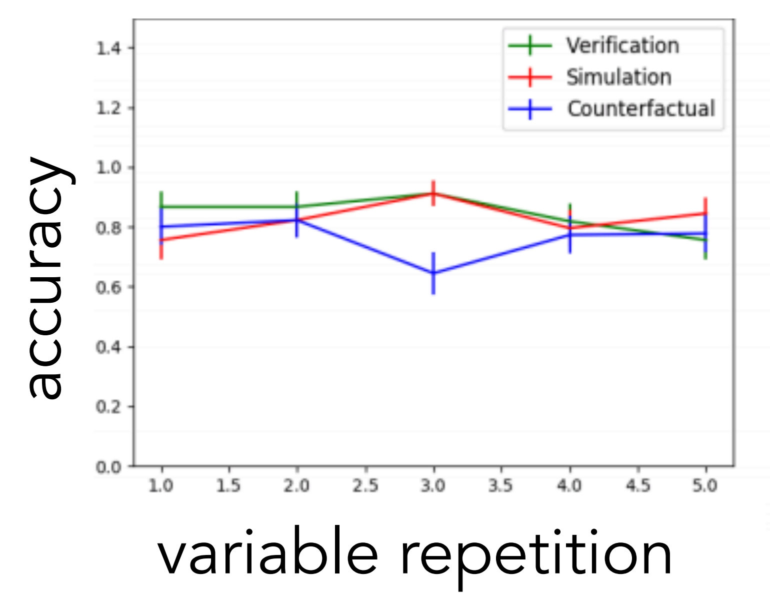
humans do
these tasks



measures of
interpretability

Variable repetition mattered less for accuracy than other factors.

*all repeated variables are needed for task completion



How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation
joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017

Agenda

1. Revisit some existing methods: Sanity check questions

2. Make explanations that work for lay people.

$$\operatorname*{argmax}_{E} Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

3. Understand how humans understand explanations

4. Make explanations to detect trustworthy predictions.

To trust or not to trust a classifier
joint work with Jiang and Gupta [NeurIPS 2018]

Ultimate goal is to use ML more responsibly.

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

Not using the classifier when it's suspicious.

↕

Improve confidence measure coming from a classifier

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

Simply confidence scores

Problem:
precision of
"definitely trustworthy (correct)" predictions

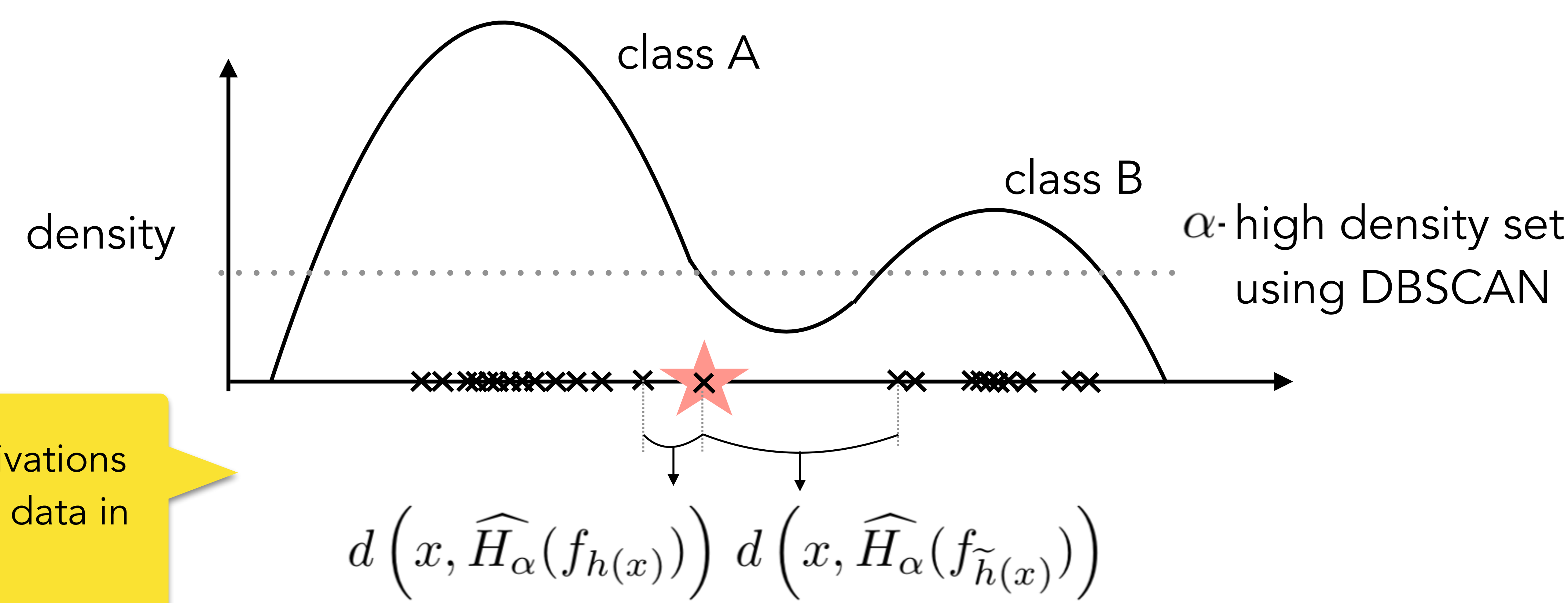
"definitely suspicious (incorrect)" predictions

To trust or not to trust a classifier
joint work with Jiang and Gupta [NeurIPS 2018]

Trust score: a super simple method

★ was predicted as class A. Can we trust this?

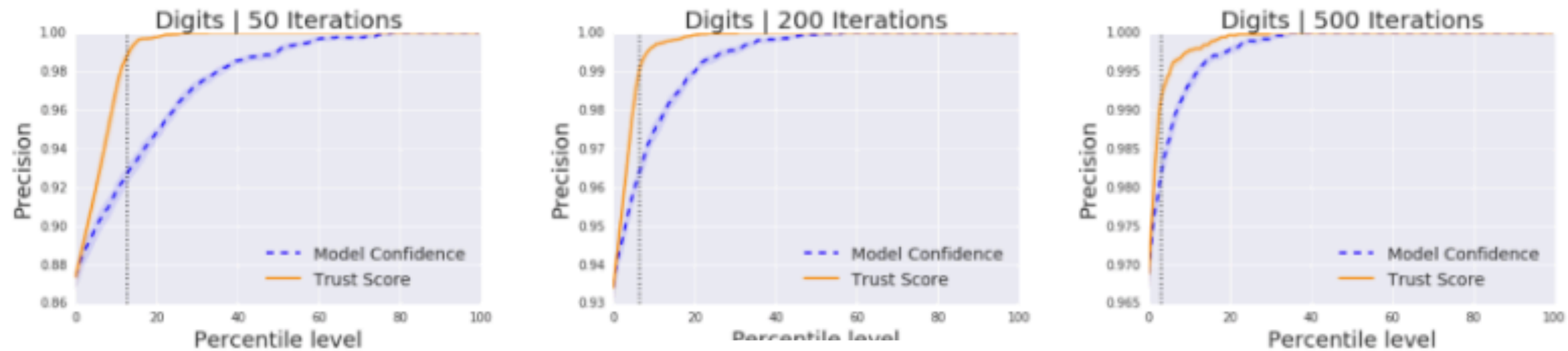
$$\text{Trust score} := \frac{d\left(x, \widehat{H_\alpha}(f_{\widetilde{h}(x)})\right)}{d\left(x, \widehat{H_\alpha}(f_{h(x)})\right)}$$



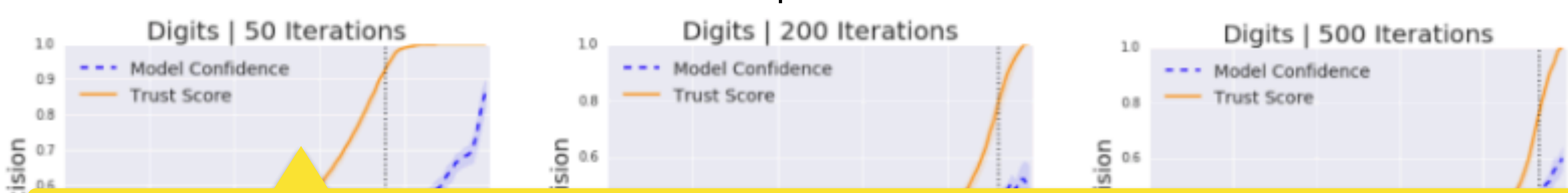
We can use activations instead of input data in NN!

results:
We can detect trustworthy and suspicious predictions with high precision.

Detect trustworthy (correct)



Detect suspicious (incorrect)



Theoretical results: why does this work?

The trust score reveals the signal from a Bayes optimal classifier (with high probability).

To trust or not to trust a classifier
joint work with Jiang and Gupta [NeurIPS 2018]

Summary, future work

1. Revisit some existing methods:
Sanity check questions
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt
NIPS 2018

2. Make explanations that work for lay people.
TCAV: Testing with concept activation vectors
Joint work with Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres
ICML 2018

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

3. Understand how humans understand explanations
How do Humans Understand Explanations from Machine Learning Systems?
An Evaluation of the Human-Interpretability of Explanation
joint work with Narayanan, Chen, He, Gershman, and Doshi-Velez 2017

4. Make explanations to detect trustworthy predictions.
To trust or not to trust a classifier
joint work with Jiang and Gupta
NIPS 2018

Understanding superhuman performance networks
Understanding models under production @ Google
Detect 'different types of mistakes' that a model makes.
…lots of others.