Google AI

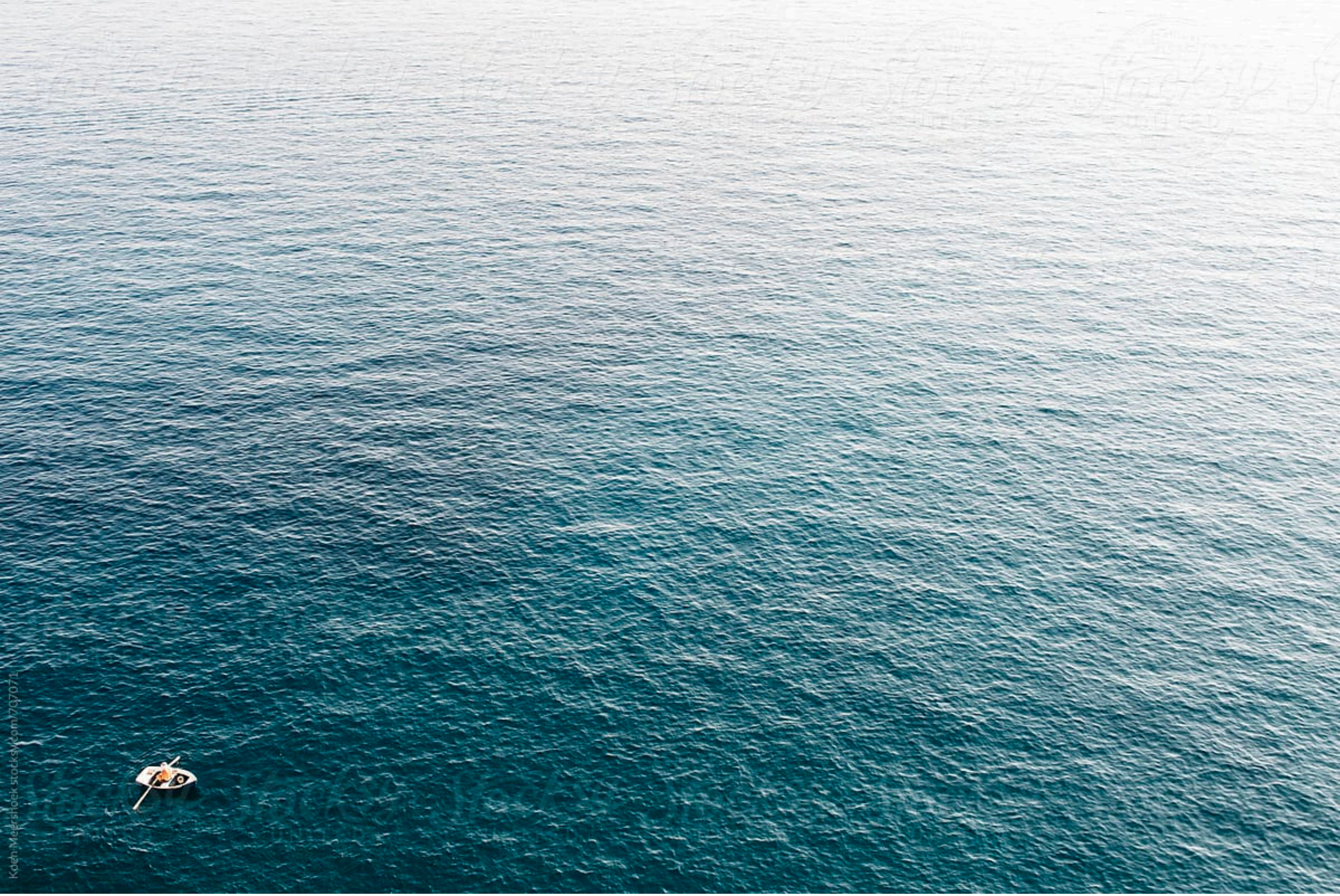
Interpretability: what now?

Been Kim

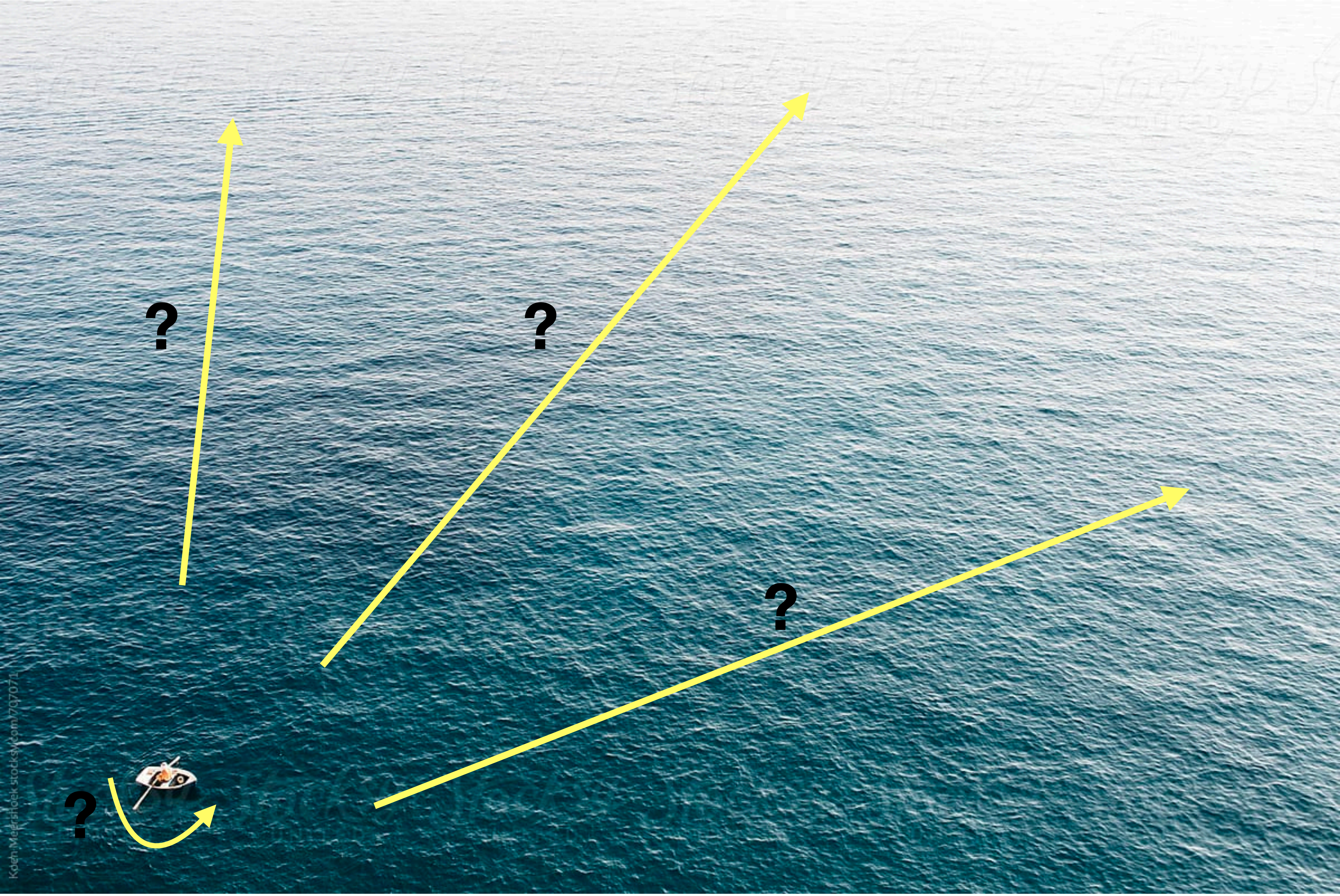
Presenting work with a lot of awesome people inside and outside of Google

Julius Adebayo, Sherry Yang, Justin Gilmer, Martin Wattenberg, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres, Ian Goodfellow, Mortiz Hardt, Michael Muelly

Sea of interpretability

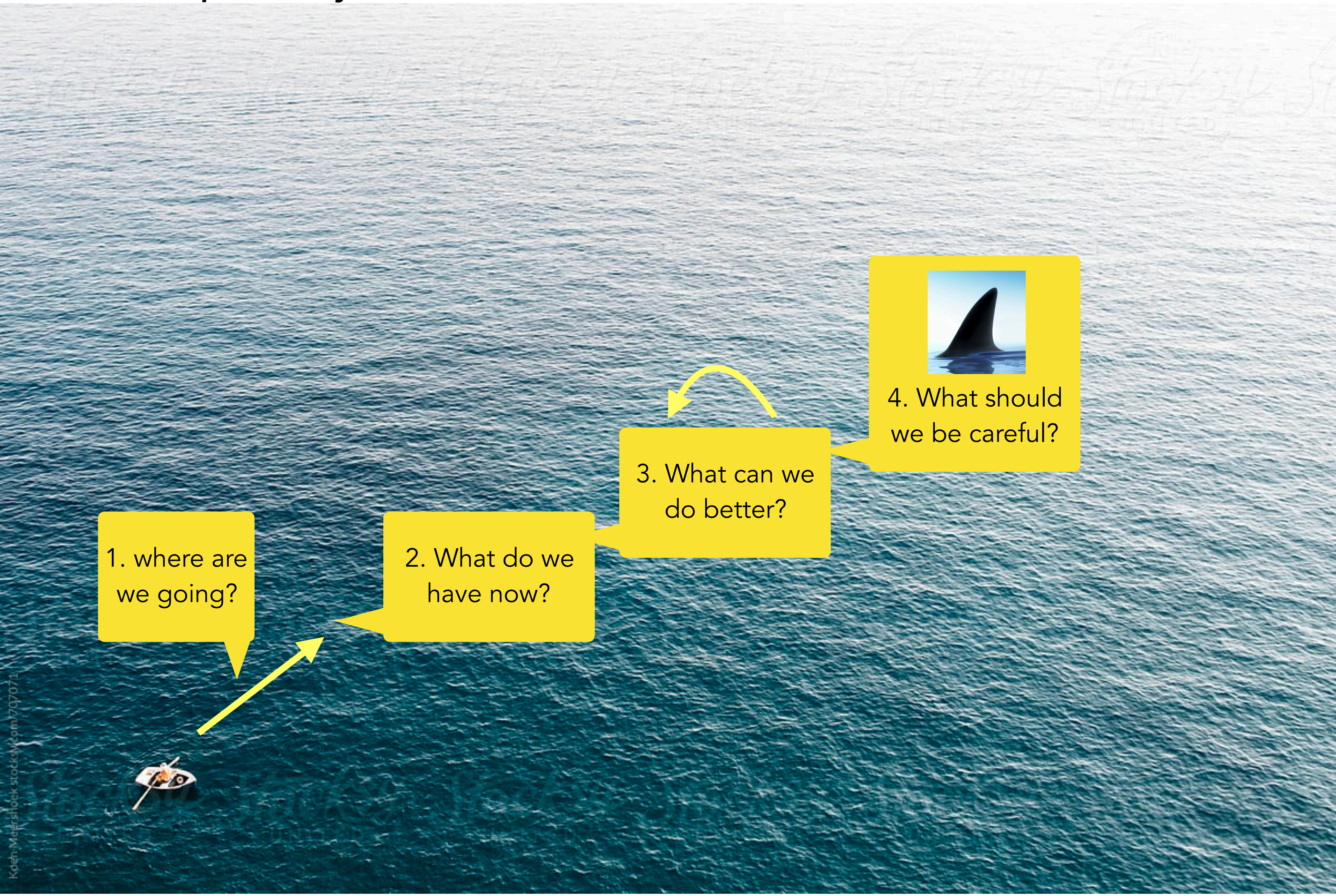


Sea of interpretability



www.stocksy.com

Sea of interpretability



Sea of interpretability

My goal

interpretability

To use machine learning **responsibly**
we need to ensure that

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

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.

NON-goals

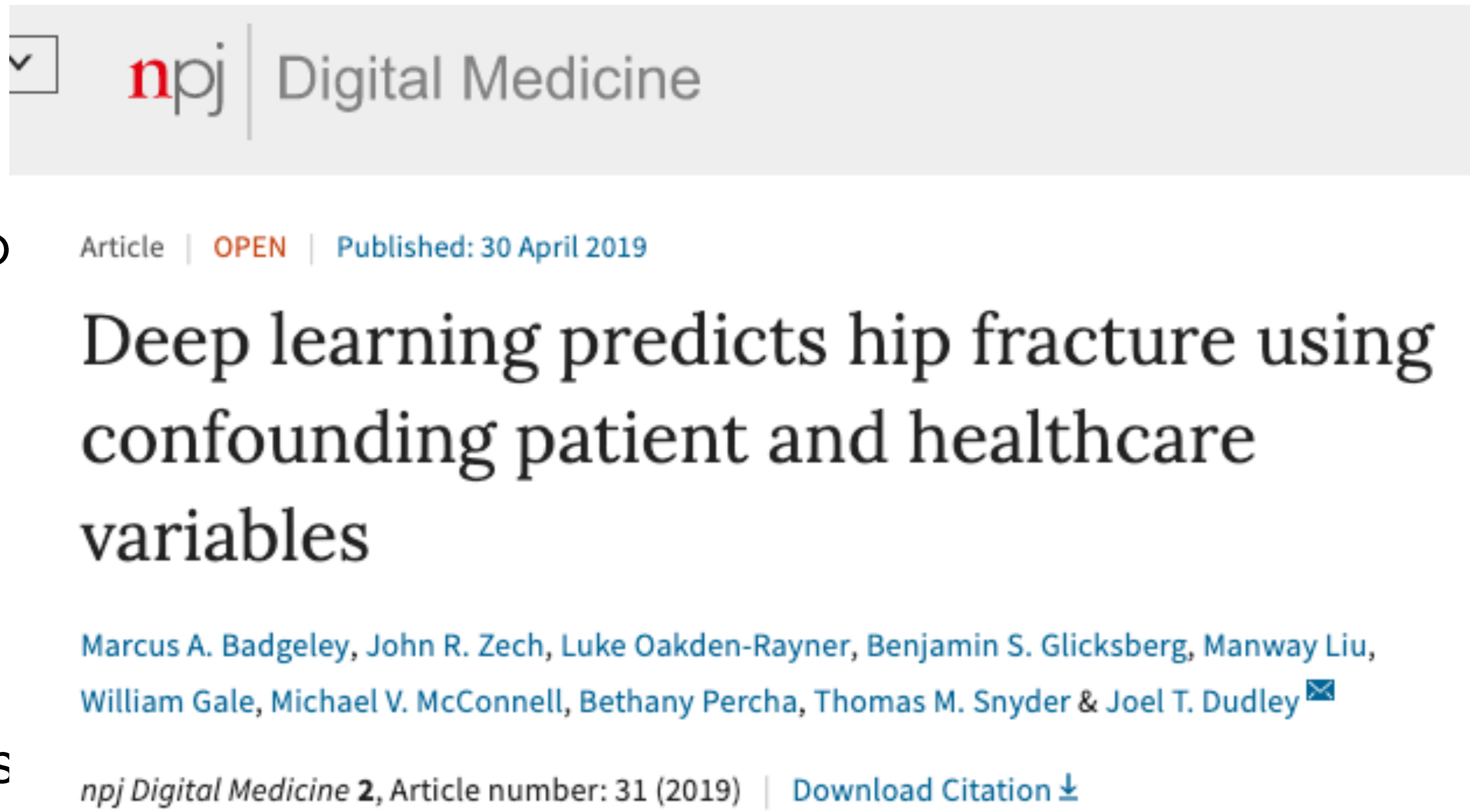
Interpretability is NOT…

- about making ALL models interpretable.
- about understanding EVERY SINGLE BIT about the model
- against developing highly complex models.
- only about gaining user trust or fairness

NON-goals

Interpretab

- about
- about
- agains



npj | Digital Medicine

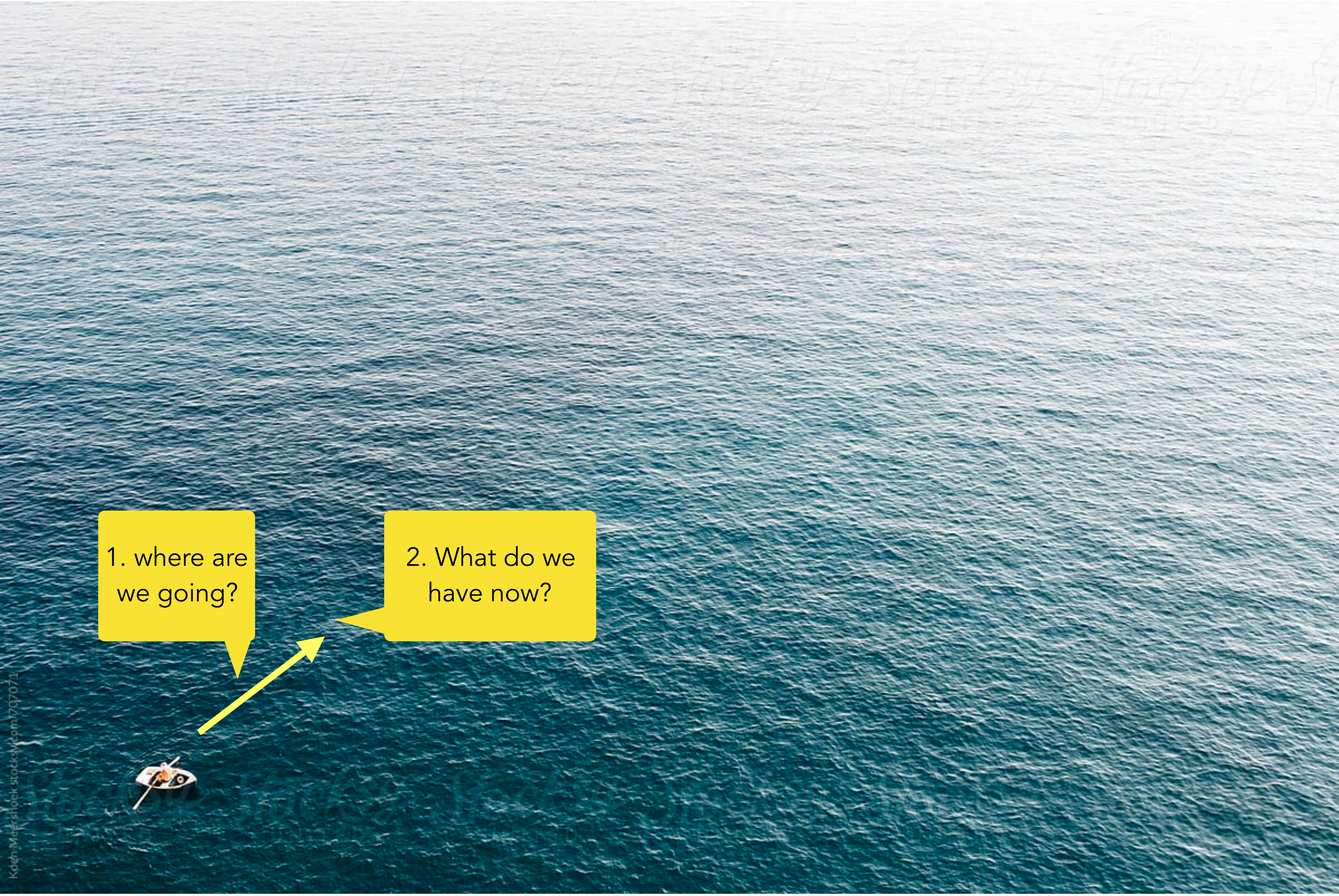
Article | OPEN | Published: 30 April 2019

Deep learning predicts hip fracture using confounding patient and healthcare variables

Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder & Joel T. Dudley

npj Digital Medicine **2**, Article number: 31 (2019) | Download Citation

- only about gaining user trust or fairness


1. where are
we going?
2. What do we
have now?
Koen Meershoek stocksy.com/707071

Investigating post-training interpretability methods.

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

$p(z)$

Junco Bird-ness

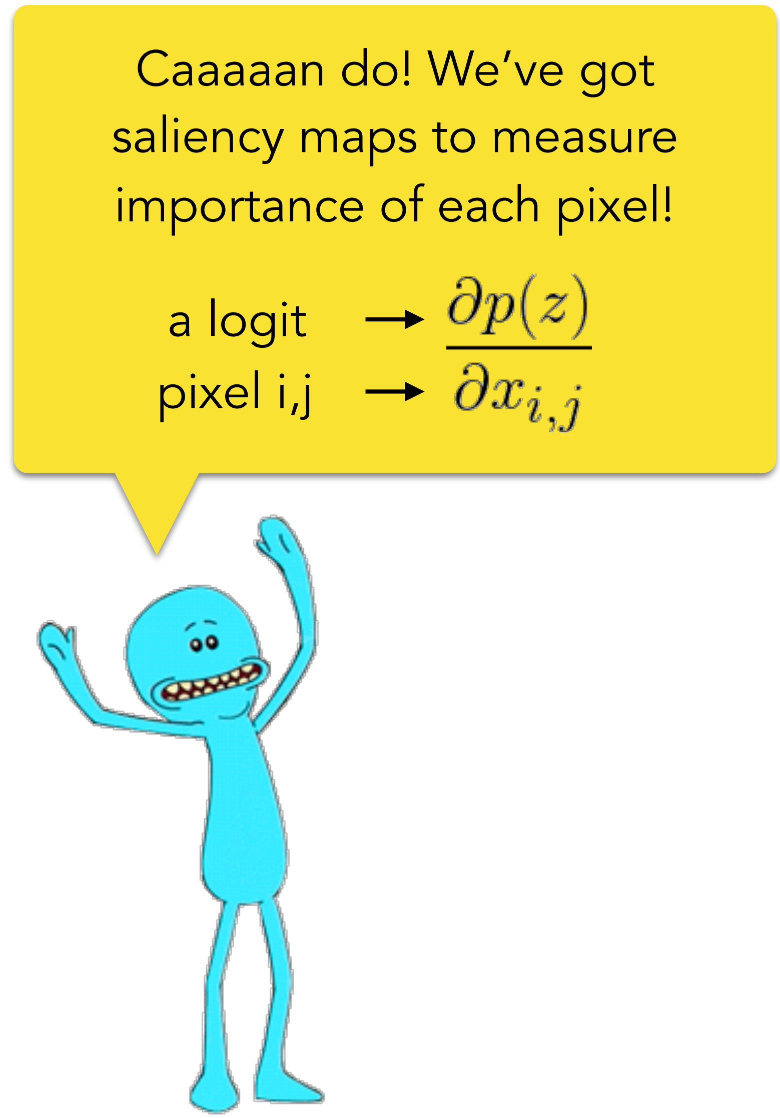
Given a fixed model, find the **evidence** of **prediction**.

Why was this a Junco bird?

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

One of the most popular interpretability methods for images:

Saliency maps



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

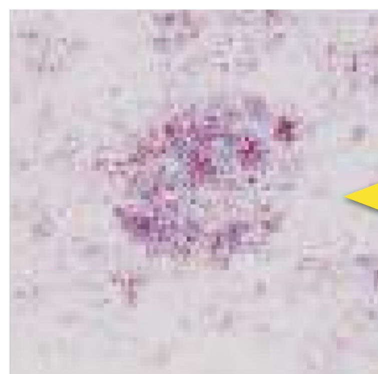
One of the most popular interpretability methods for images:

Saliency maps

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

$$p(z)$$

Junco Bird-ness



The promise:
these pixels are the
evidence of
prediction.

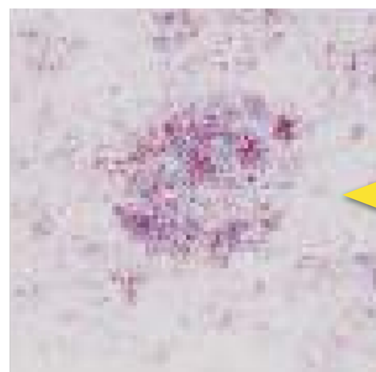
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Sanity check question.

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

$p(z)$

Junco Bird-ness



The promise:
these pixels are the
evidence of
prediction.

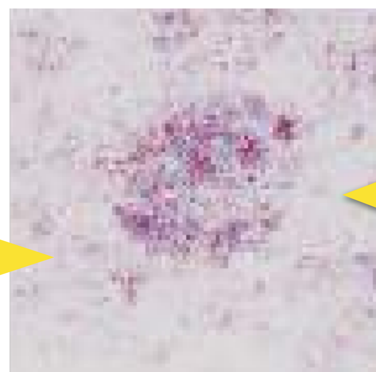
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Sanity check question.

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

$p(z)$

Junco Bird-ness



The promise: these pixels are the **evidence** of **prediction.**

If so, when **prediction** changes, the explanation should change.

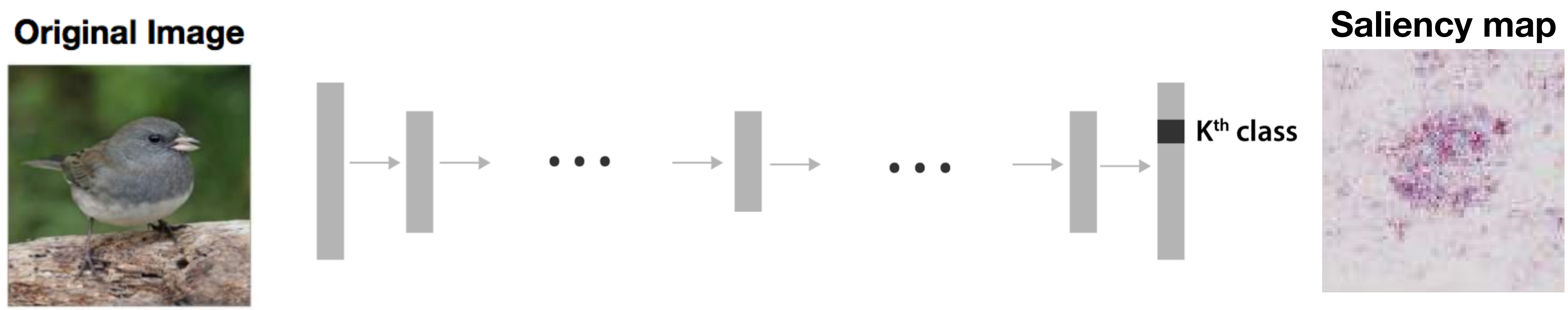
Extreme case:
If **prediction** is random,
the **explanation** should
REALLY change.

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

Some confusing behaviors of saliency maps.

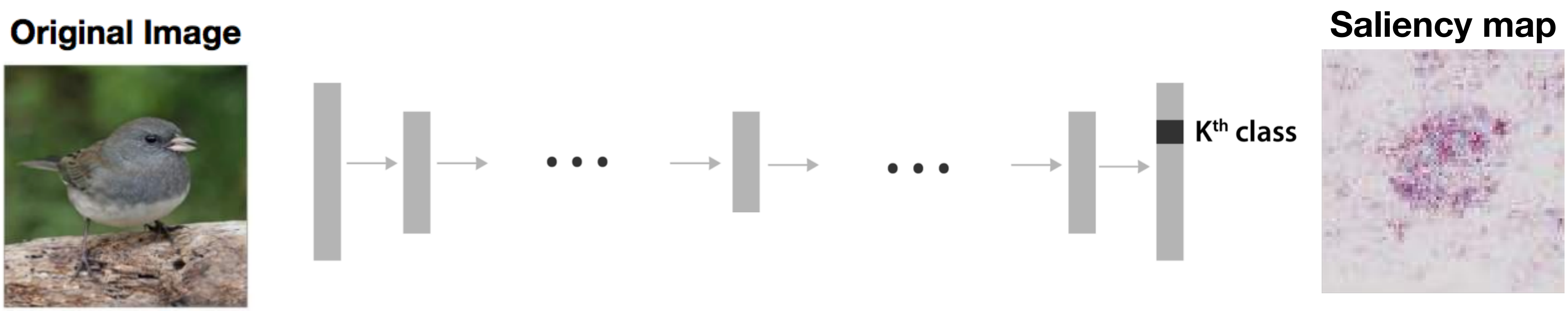
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Some confusing behaviors of saliency maps.

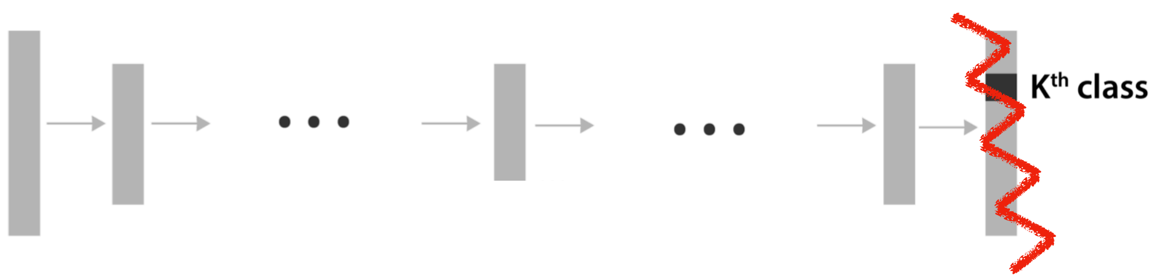


Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NeurIPS 18]

Some confusing behaviors of saliency maps.

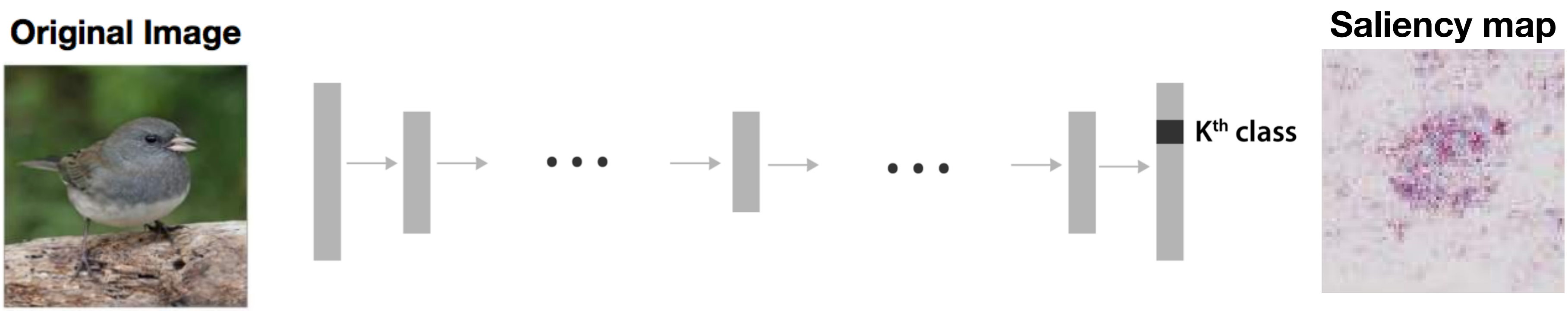


Randomized weights!
Network now makes garbage prediction.



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

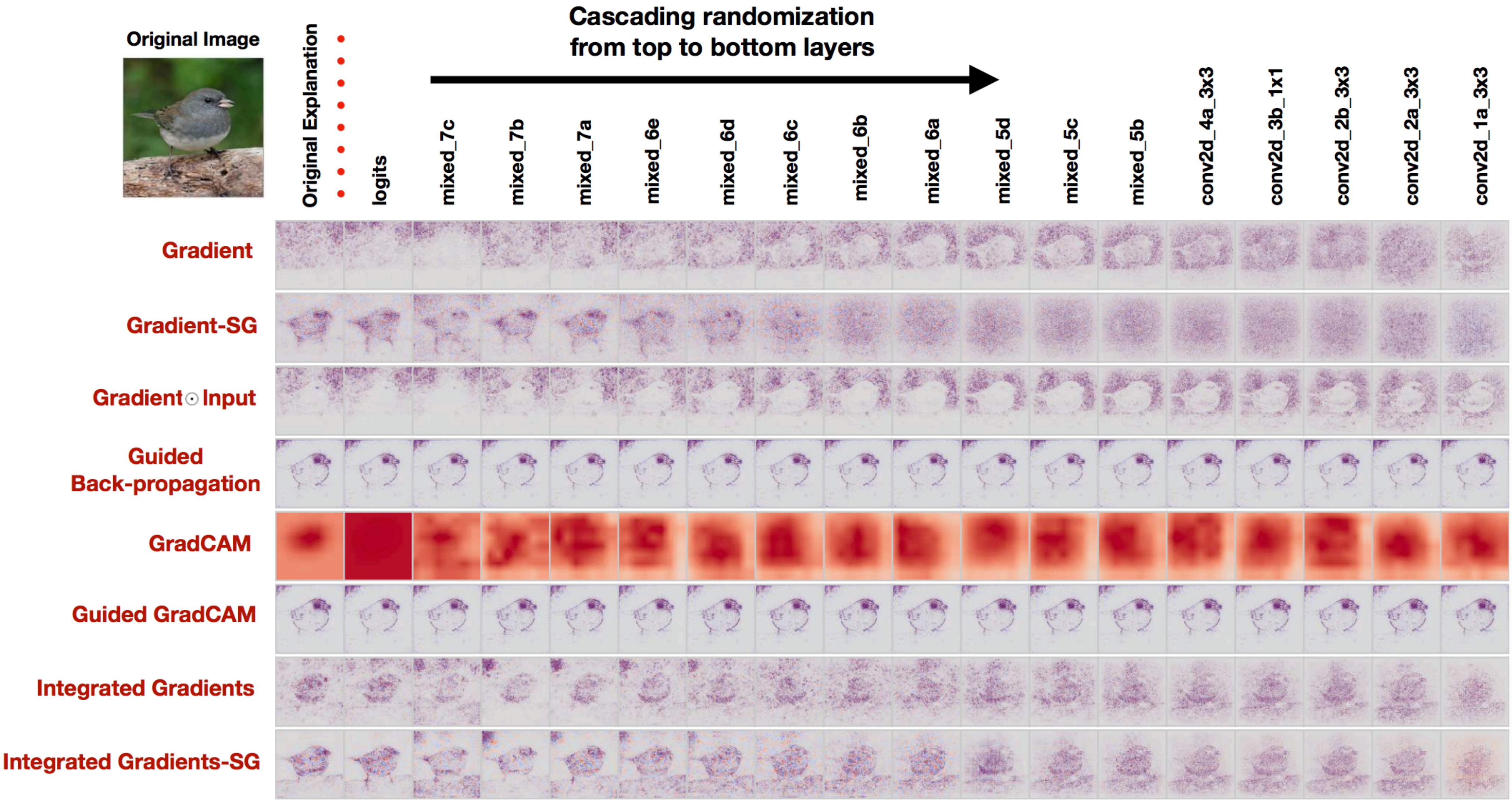
Some confusing behaviors of saliency maps.

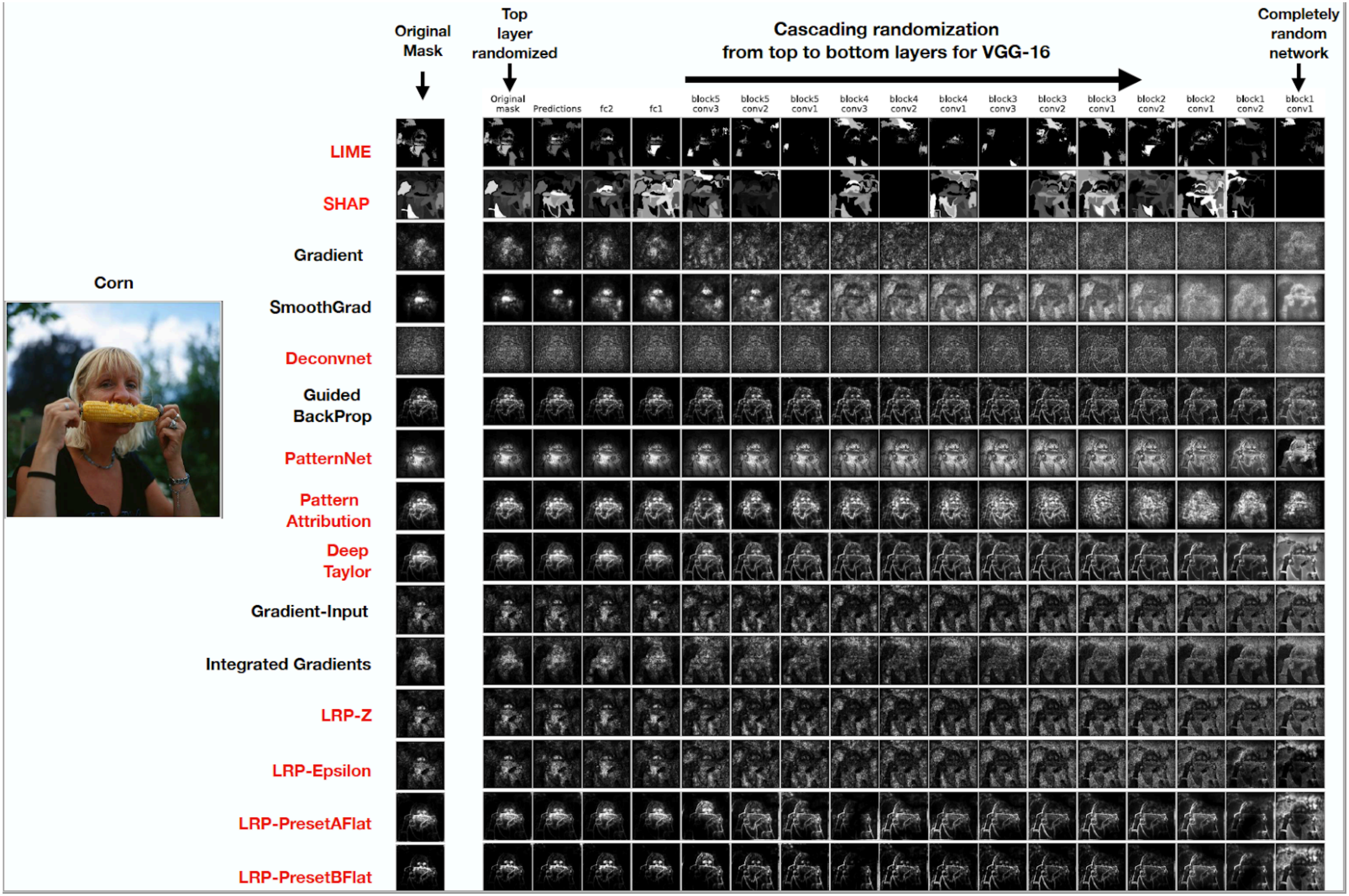


Randomized weights!
Network now makes garbage prediction.

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

Sanity check1:
When prediction changes, do explanations change?
No!

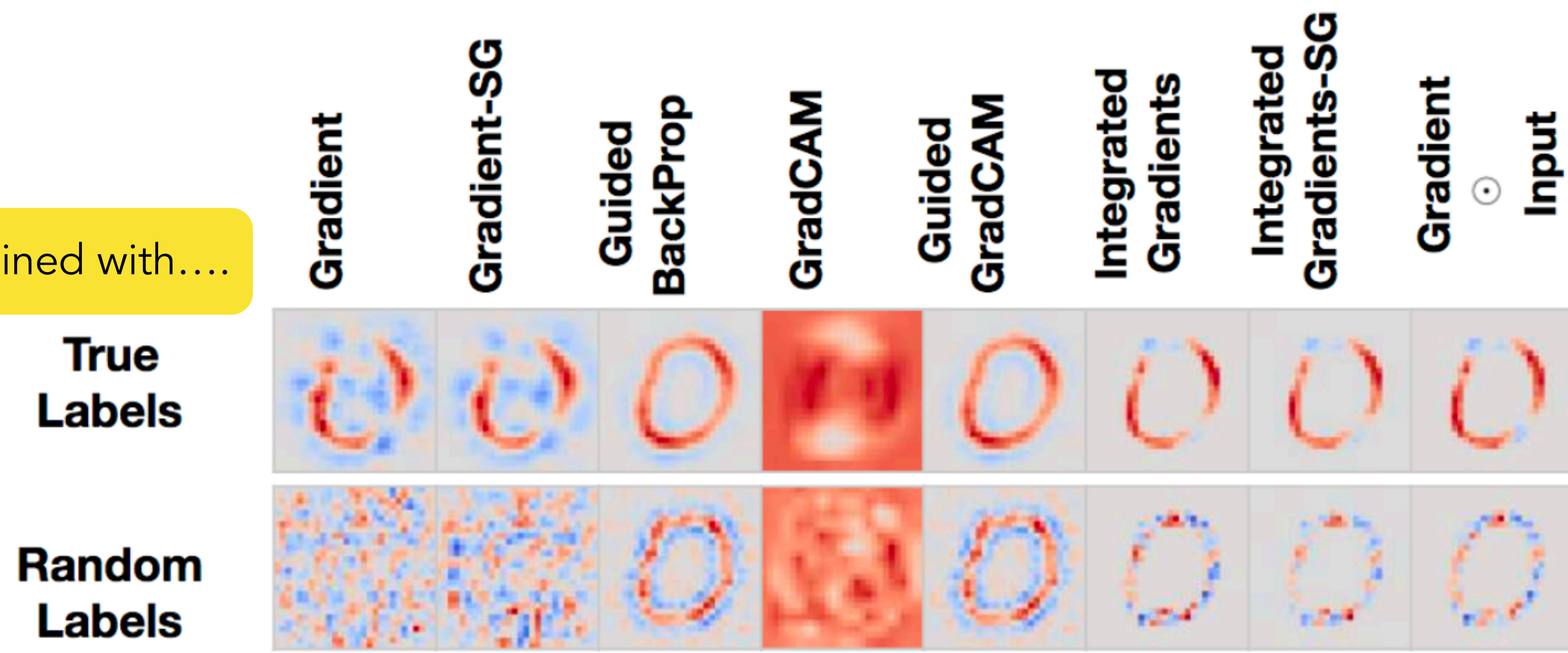




Original Mask
Top layer randomized
Cascading randomization from top to bottom layers for VGG-16
Completely random network
Original mask
Predictions
fc2
fc1
block5 conv3
block5 conv2
block5 conv1
block4 conv3
block4 conv2
block4 conv1
block3 conv3
block3 conv2
block3 conv1
block2 conv2
block2 conv1
block1 conv2
block1 conv1
Corn
LIME
SHAP
Gradient
SmoothGrad
Deconvnet
Guided BackProp
PatternNet
Pattern Attribution
Deep Taylor
Gradient-Input
Integrated Gradients
LRP-Z
LRP-Epsilon
LRP-PresetAFlat
LRP-PresetBFlat

Sanity check2:
Networks trained with true and random labels, Do explanations deliver different messages?
No!

Networks trained with….



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

What can we learn from this?

- **Confirmation bias**: Just because it "makes sense" to humans, doesn't mean it reflects the evidence for prediction.

- Others who independently reached the same conclusions: [Nie, Zhang, Patel '18] [Ulyanov, Vedaldi, Lempitsky '18]

- Some of these methods have been shown to be useful for humans. Why? More studies needed.

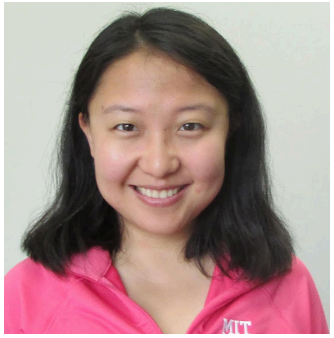
- Recent work by Gupta and Arora 19' suggests a simple fix

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

This was a low bar test.

Can we put interpretability methods on a harder test?

Benchmarking interpretability methods (BIM)

work with Sherry Yang 

Benchmarking interpretability methods (BIM)

Forest

A thing

work with Sherry Yang

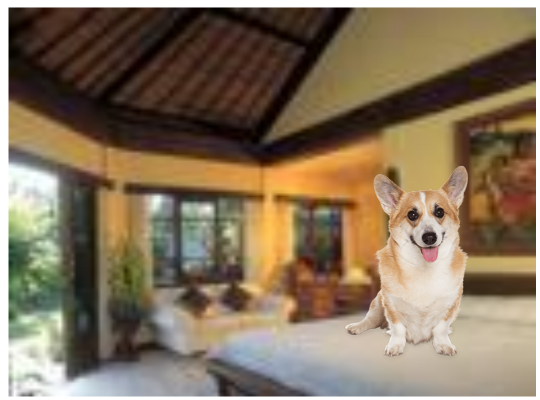
Benchmarking interpretability methods (BIM)

Forest

A thing

Forest

Bedroom



Kitchen

work with Sherry Yang

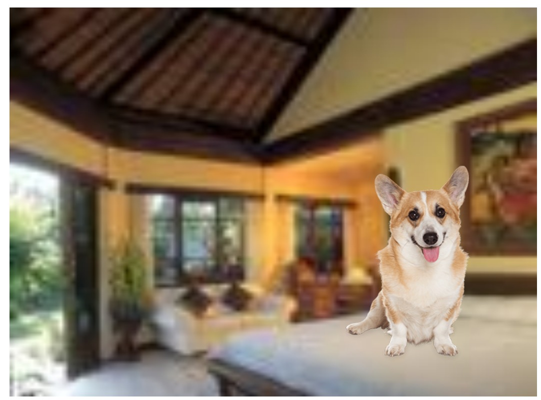
Benchmarking interpretability methods (BIM)

Forest

A thing

Forest

Bedroom



Kitchen

 is NOT important for predicting scene classes.

↓

 should NOT Be part of explanation

work with Sherry Yang

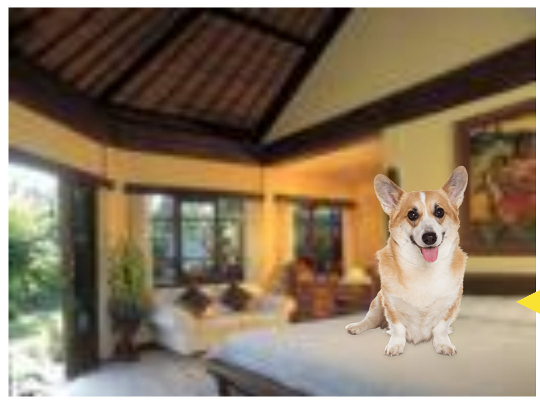
Benchmarking interpretability methods (BIM)

Forest

A thing

Forest

Bedroom



Kitchen

 is NOT important for predicting scene classes.

↓

 should NOT Be part of explanation

We can also make  more important to some classes by controlling when it appears.

↓

 should be more important explanation in some classes than others.

github.com/google-research-datasets/bim

Three metrics for measuring false positives

		Model's truth	
		important	Not important
Interp. methods estimates	important	TP	FP
	not important	FN	TN

Our Focus

github.com/google-research-datasets/bim

Three metrics for measuring false positives

		Model's truth	
		important	Not important
Interp. methods estimates	important	TP	FP
	not important	FN	TN

Our Focus

Suggested metrics

- Model contrast score (MCS)
- Input dependence rate (IDR)
- Input independence rate (IIR)

github.com/google-research-datasets/bim

Three metrics for measuring false positives

		Model's truth	
		important	Not important
Interp. methods estimates	important	TP	FP
	not important	FN	TN

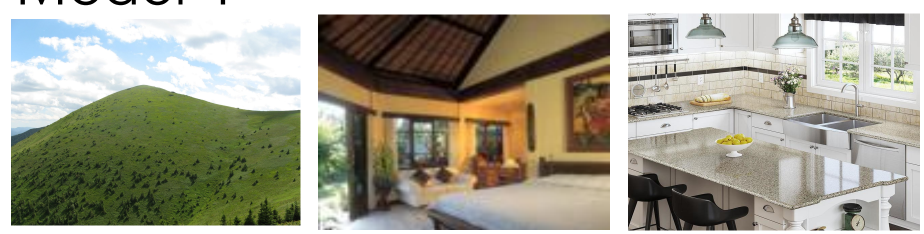
Our Focus

Suggested metrics

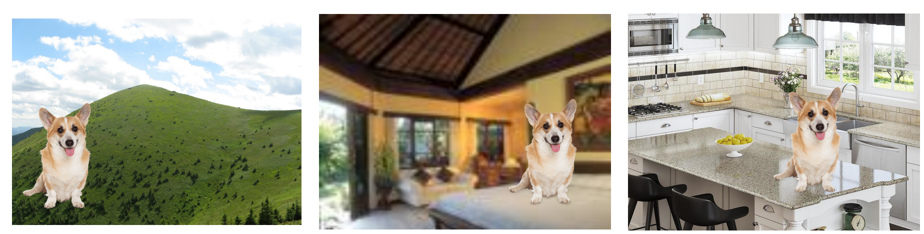
- Model contrast score (MCS)
- Input dependence rate (IDR)
- Input independence rate (IIR)

Two models trained to classify scenes.

Model 1



Model 2



github.com/google-research-datasets/bim

Three metrics for measuring false positives

		Model's truth	
		important	Not important
Interp. methods estimates	important	TP	FP
	not important	FN	TN

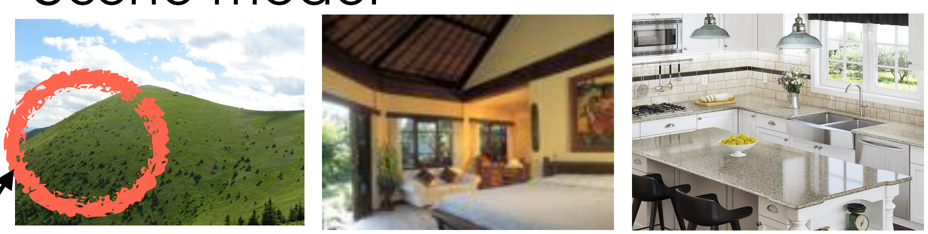
Our Focus

Suggested metrics

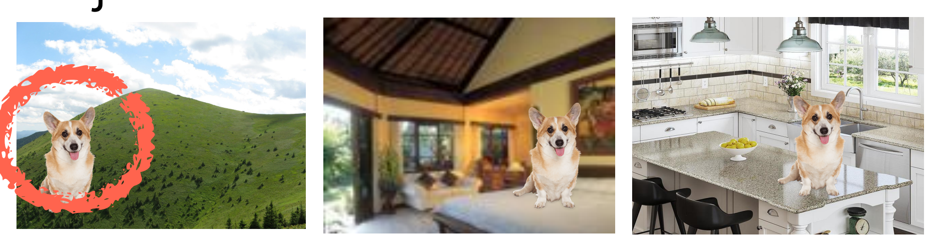
- Model contrast score (MCS)
- Input dependence rate (IDR)
- Input independence rate (IIR)

Two models trained to classify

Scene model



Object model



We expect big contrast on where the object is.

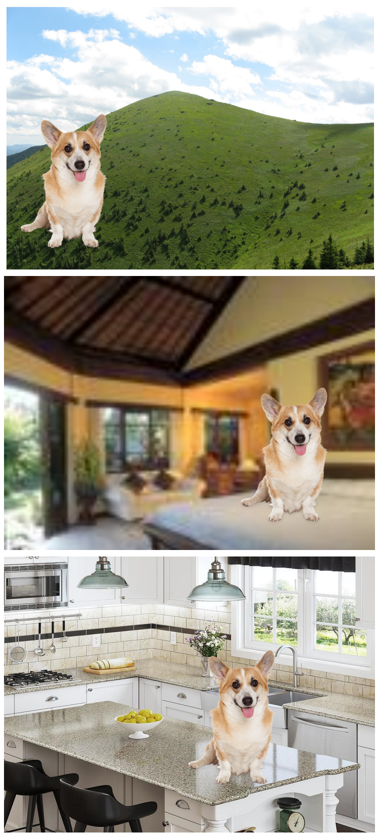
github.com/google-research-datasets/bim

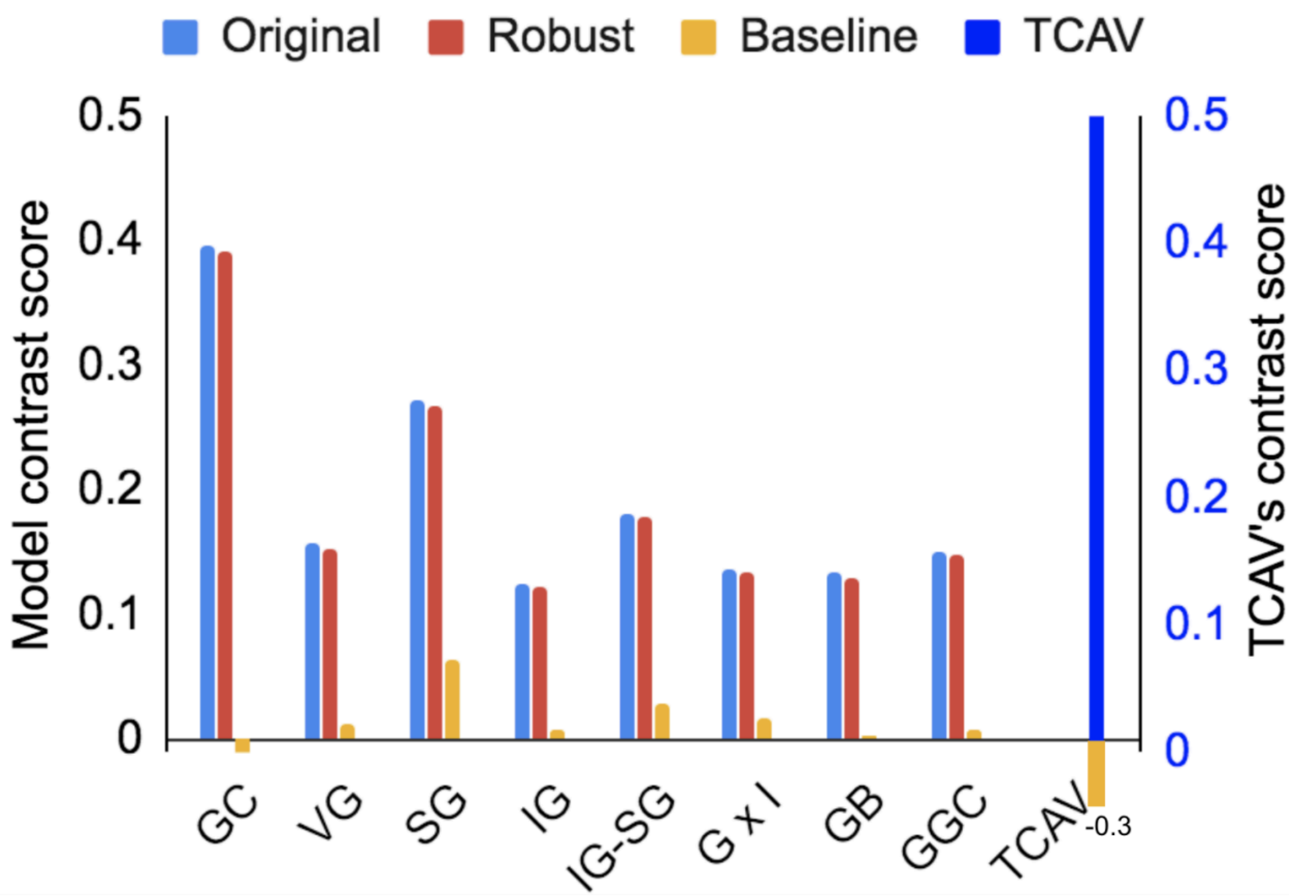
Model Contrast Score (MCS)

Forest

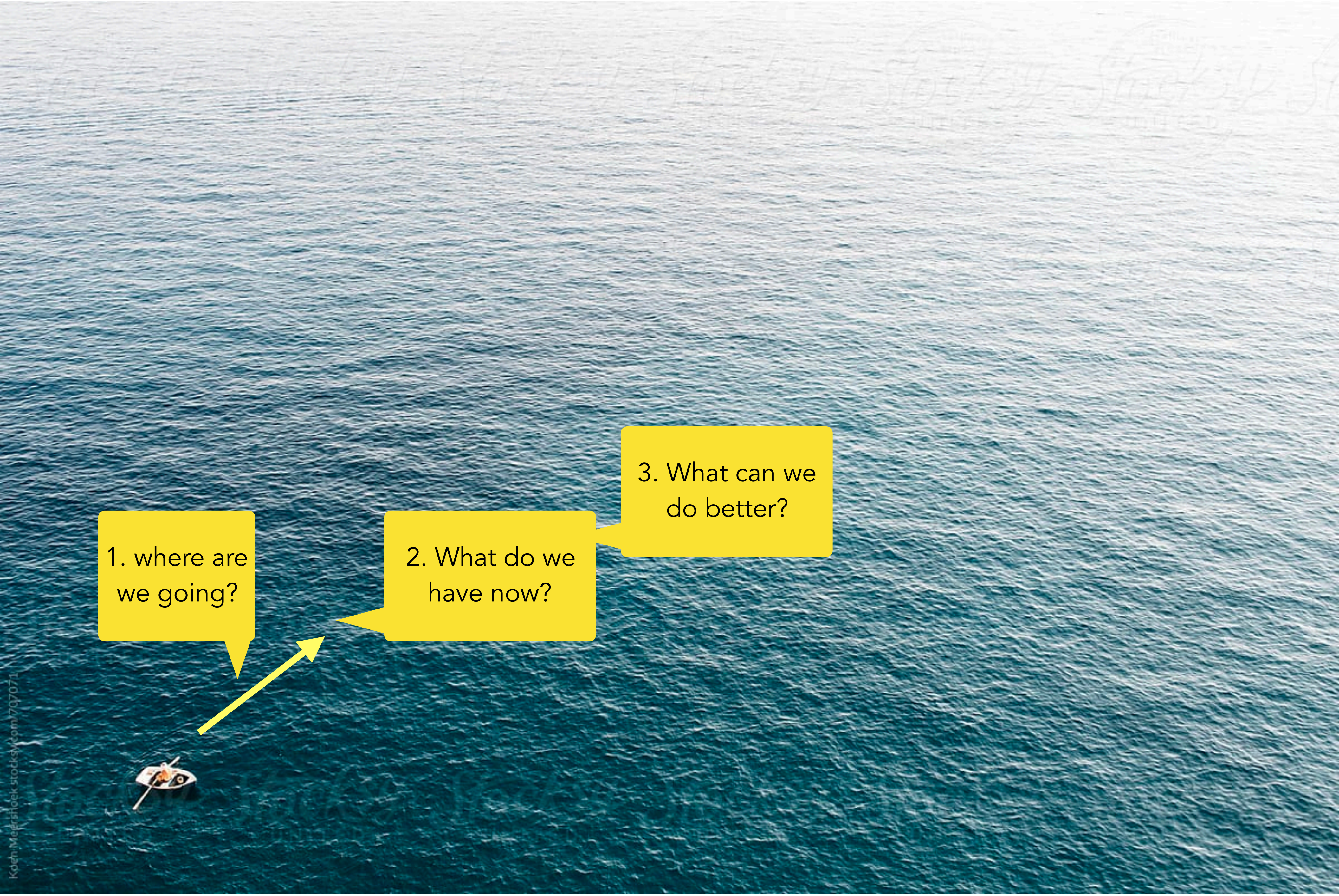
Bedroom

Kitchen





github.com/google-research-datasets/bim


1. where are we going?
2. What do we have now?
3. What can we do better?

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)

$p(z)$

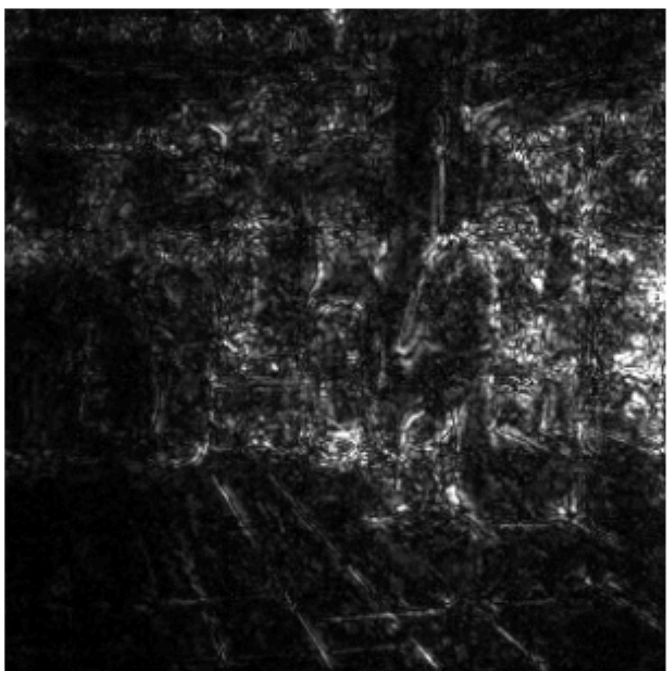
cash-machine-ness

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

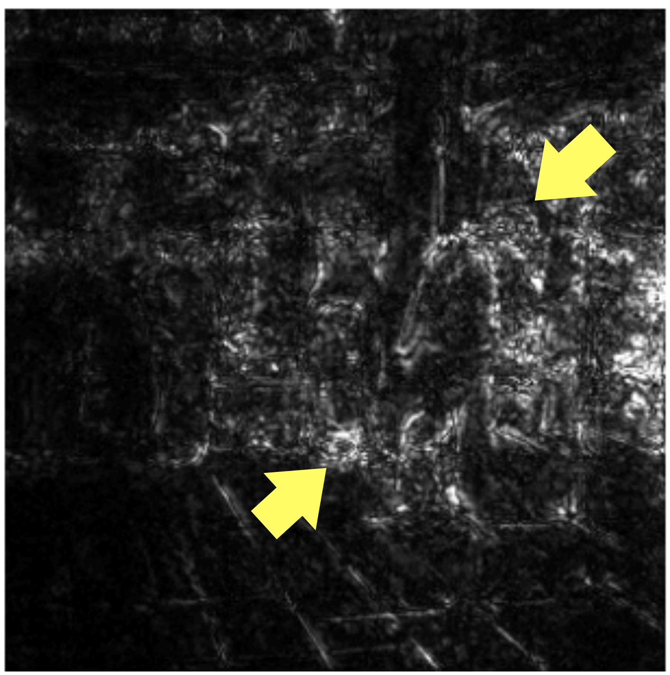


Let's use this to help us think about what what we really want to ask.

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?

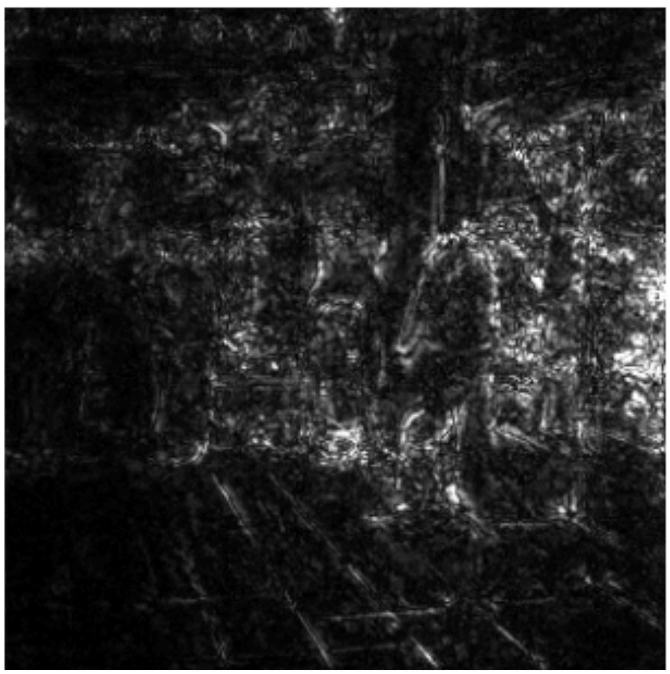
Did the 'human' concept matter?
Did the 'wheels' concept matter?

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?

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?

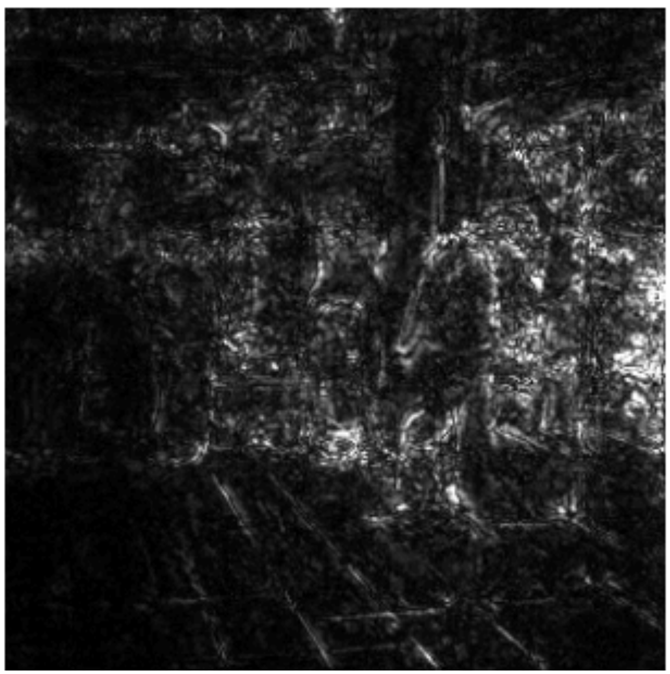
Oh no! I can’t express these concepts as pixels!!
They weren’t my input features either!

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?

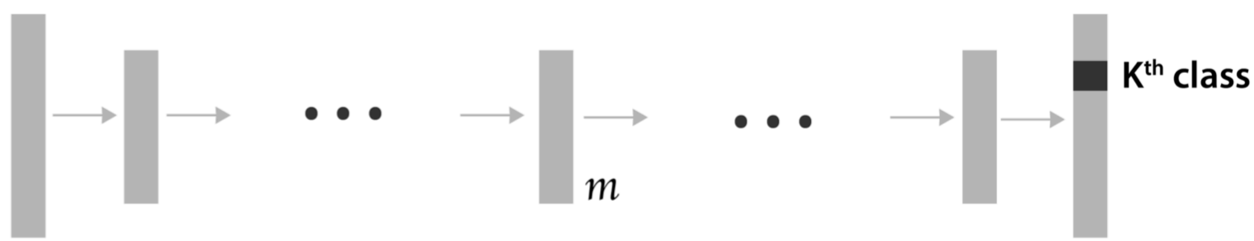
Wouldn't it be great if we can **quantitatively** measure how important *any* of these **user-chosen concepts** are?

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)

$p(z)$

zebra-ness

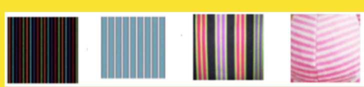
Was striped concept important
to this zebra image classifier?

Goal of TCAV:
Testing with Concept Activation Vectors

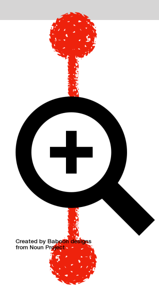
A trained machine learning model (e.g., neural network)

$p(z)$

zebra-ness



Was striped concept important to this zebra image classifier?



TCAV score for **Zebra**

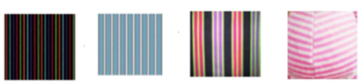
TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

TCAV:
Testing with Concept Activation Vectors

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

$p(z)$

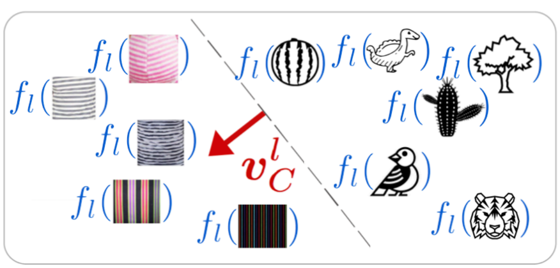
zebra-ness



Was striped concept important
to this zebra image classifier?

TCAV

1. Learning CAVs



1. How to define
concepts?

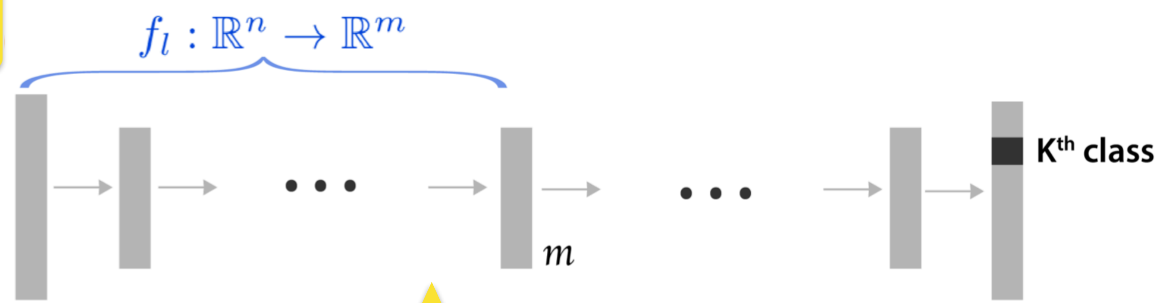
Defining concept activation vector (CAV)

Inputs:

Examples of concepts

a

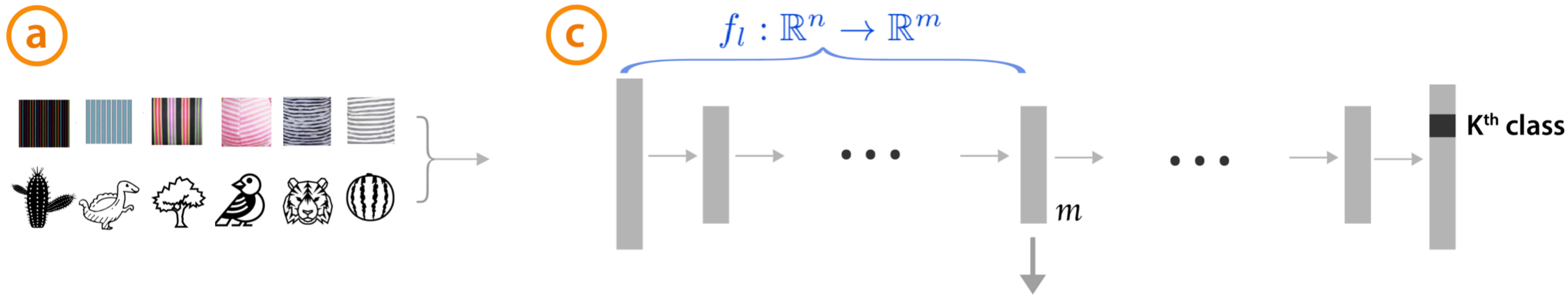
Random images

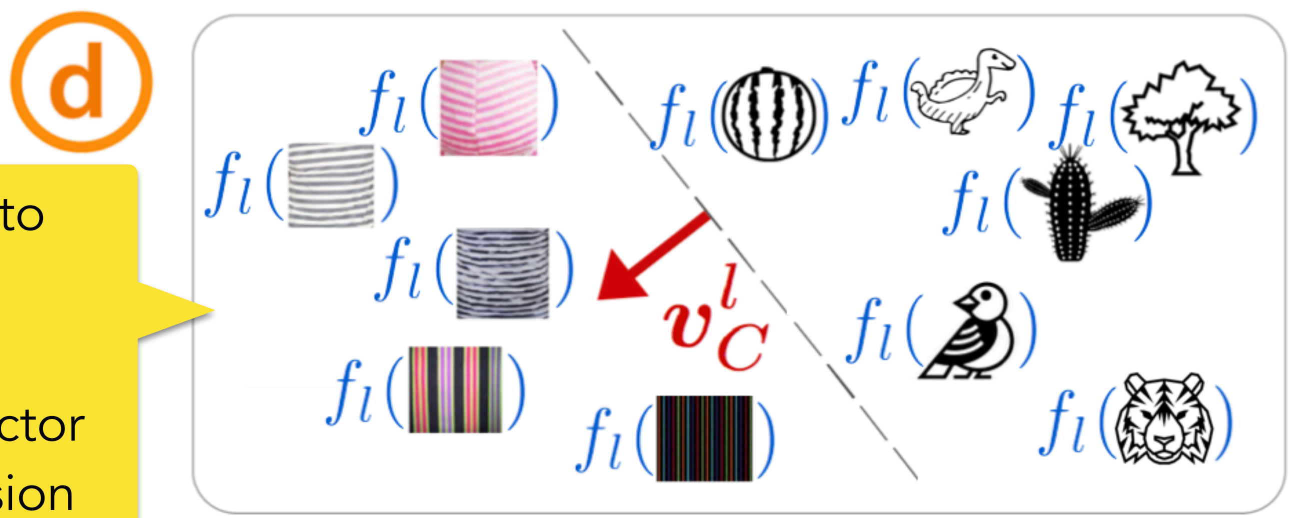


A trained network under investigation
and
Internal tensors

Defining concept activation vector (CAV)

Inputs:



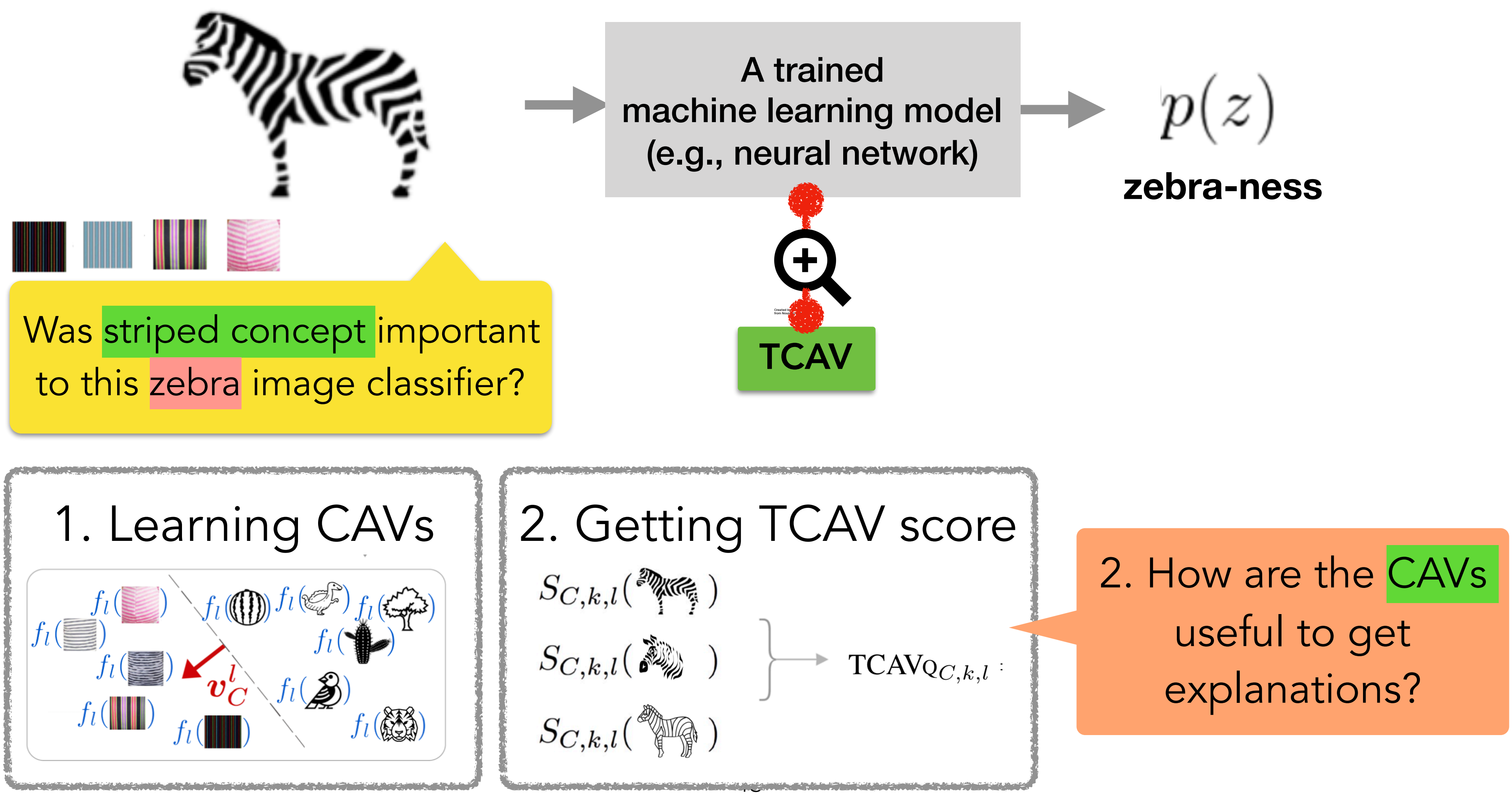


Train a linear classifier to separate activations.

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

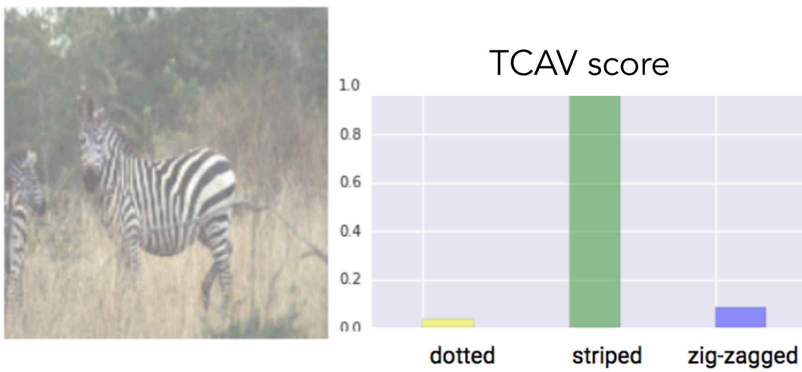
[Smilkov '17, Bolukbasi '16 , Schmidt '15]

TCAV:
Testing with Concept Activation Vectors



TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV



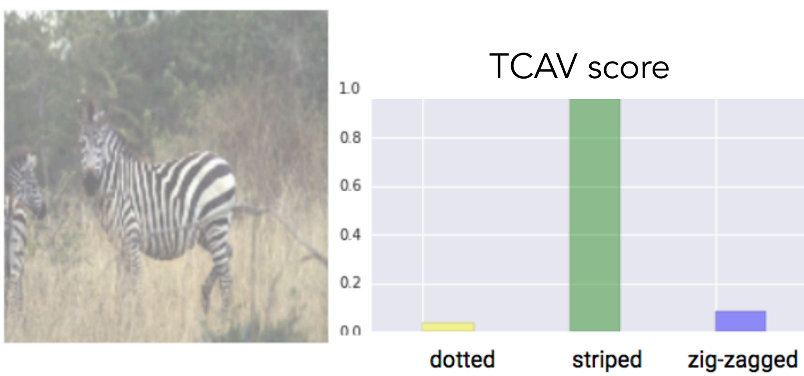
zebra-ness →
striped CAV →

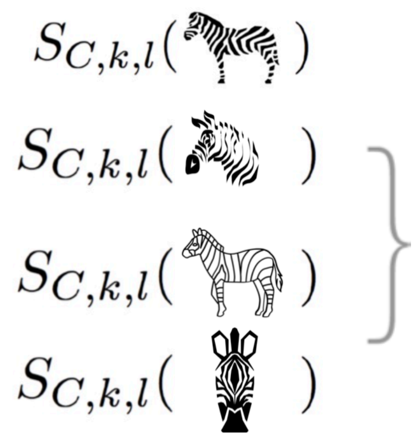
$$\frac{\partial p(z)}{\partial\, \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$$

Directional derivative with CAV

TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV





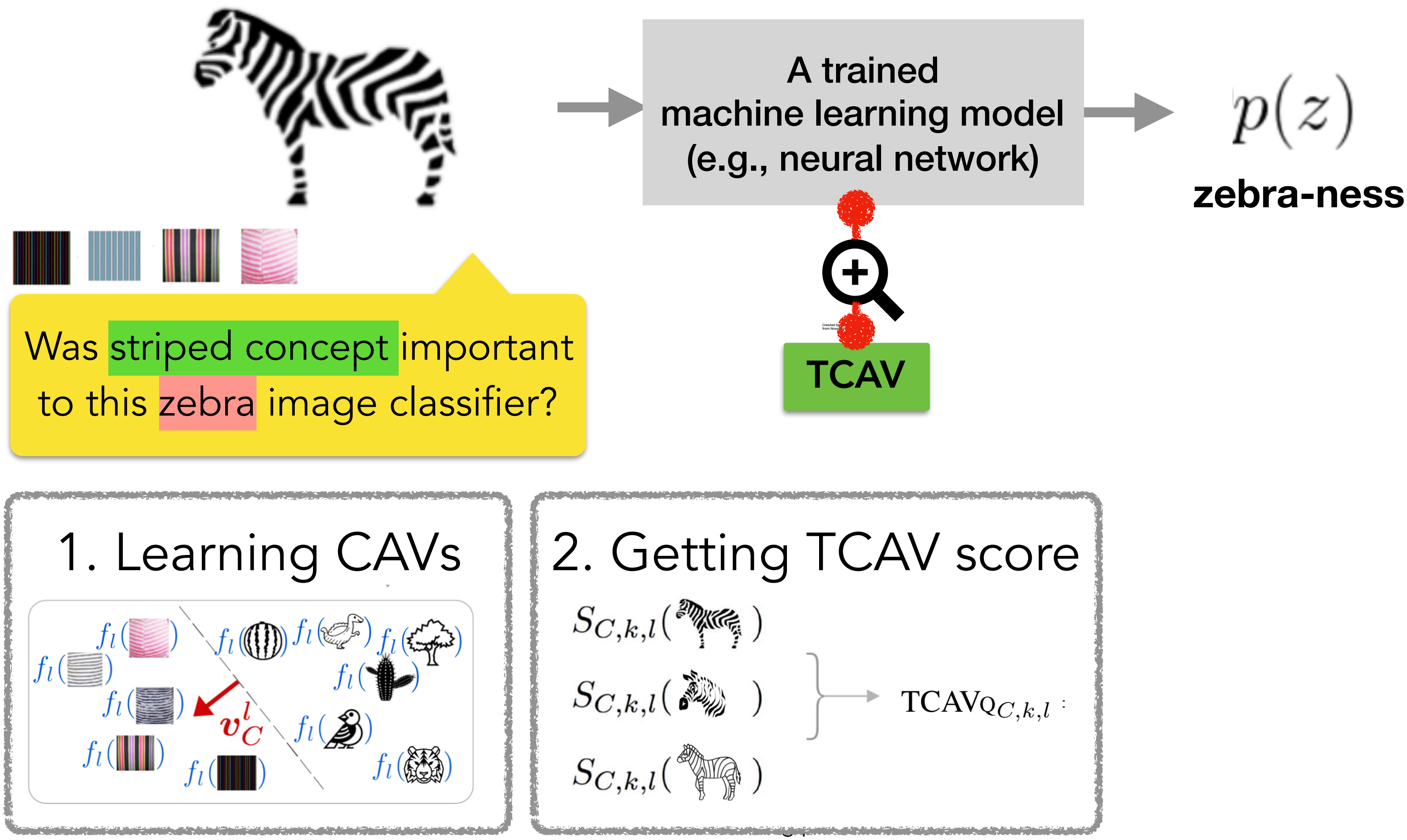
zebra-ness →
striped CAV →

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

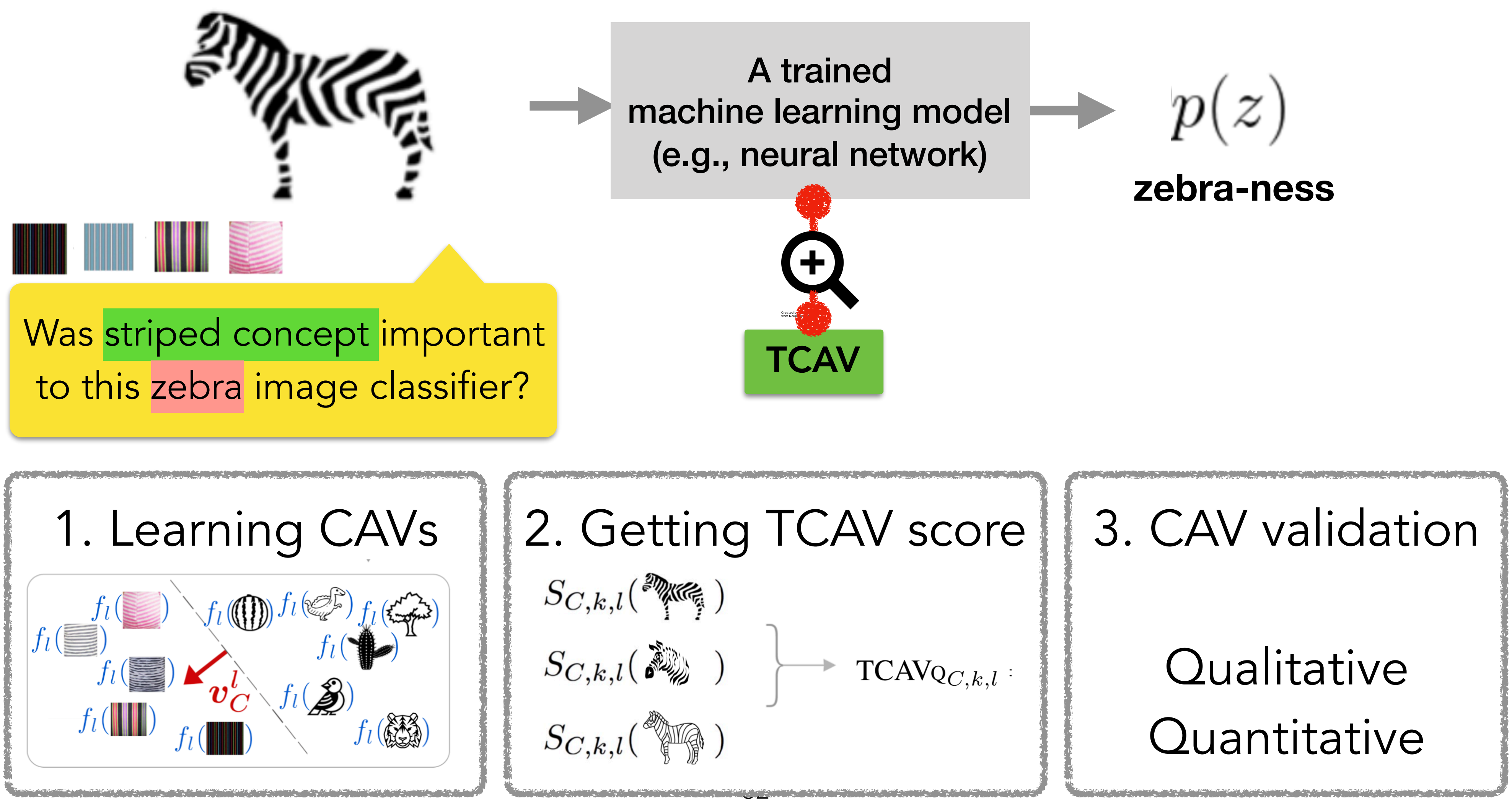
$$\boxed{\mathrm{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}}$$

Directional derivative with CAV

TCAV:
Testing with Concept Activation Vectors



TCAV:
Testing with Concept Activation Vectors



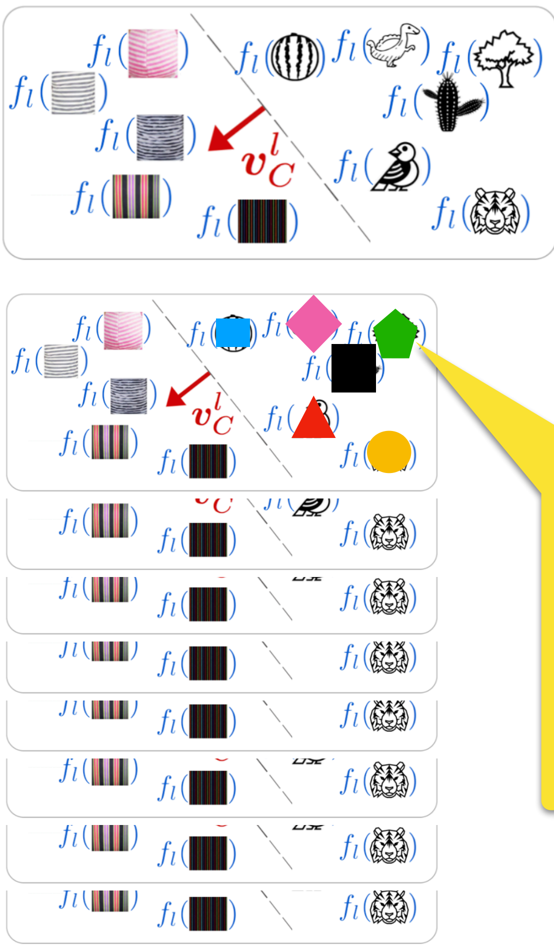
Quantitative validation:

Guarding against spurious CAV

Did my CAVs returned high sensitivity by chance?

Quantitative validation:

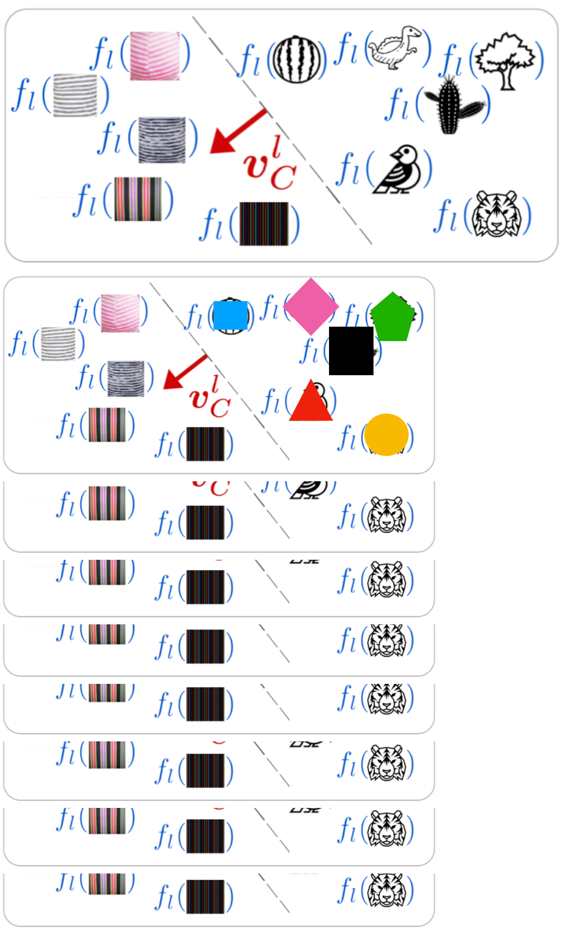
Guarding against spurious CAV



Learn many stripes CAVs using different sets of random images

Quantitative validation:

Guarding against spurious CAV



Zebra

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

⋮

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

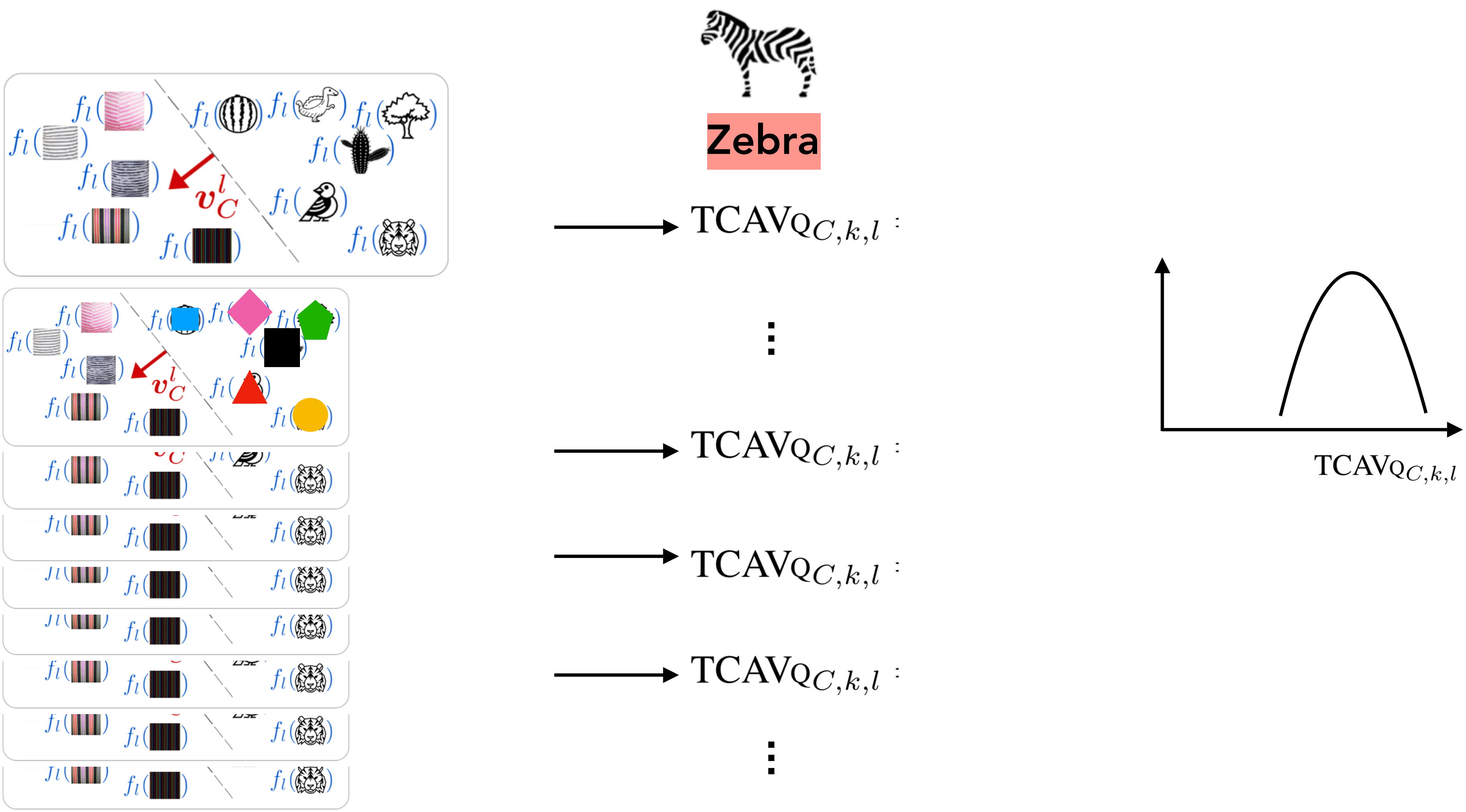
$\longrightarrow$ $\mathrm{TCAVQ}_{C,k,l}$:

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

⋮

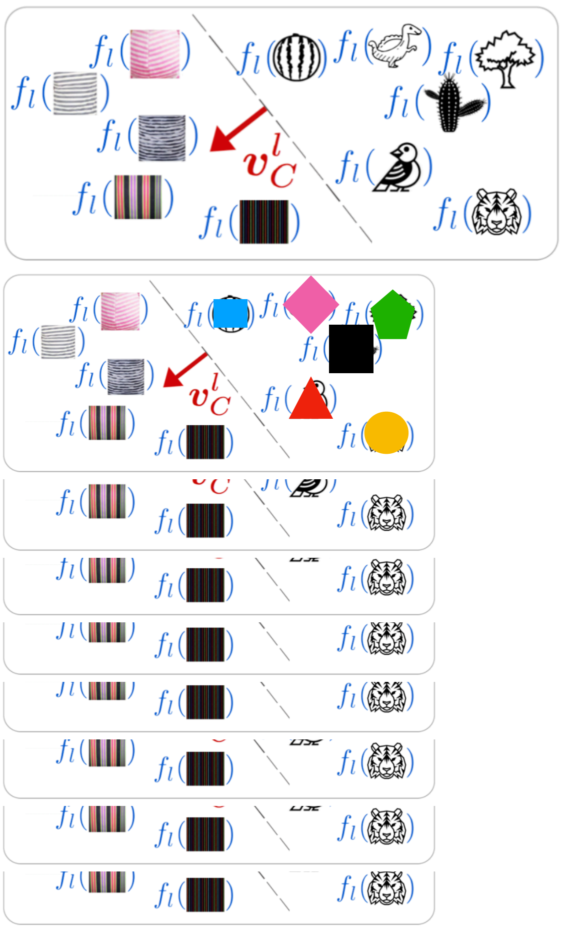
Quantitative validation:

Guarding against spurious CAV



Quantitative validation:

Guarding against spurious CAV



Zebra

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

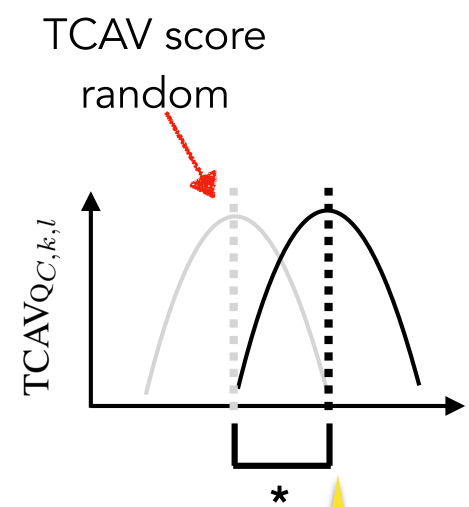
⋮

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

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

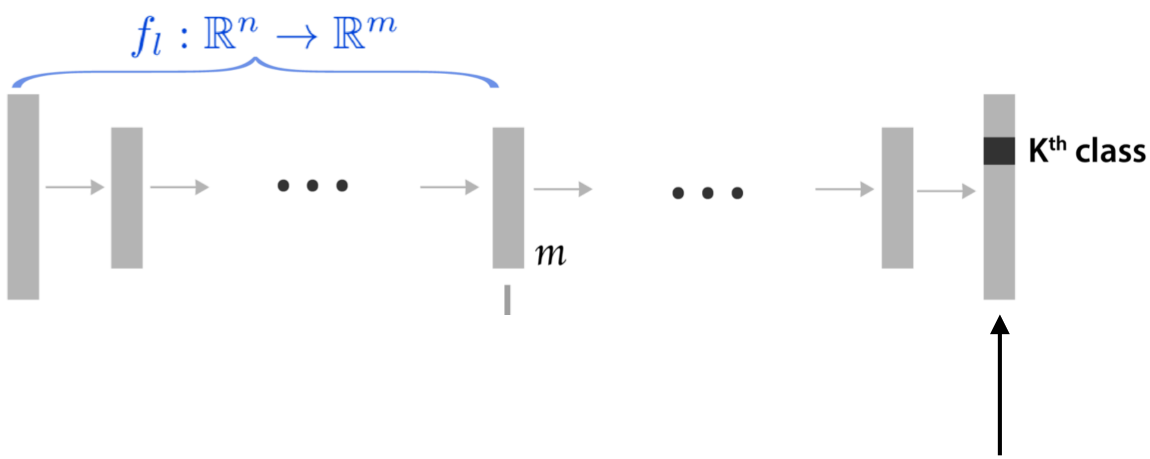
$\longrightarrow \mathrm{TCAVQ}_{C,k,l}$:

⋮



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

How to choose a layer



Start from the top layer (closest to prediction).

Go down a layer if the current layer doesn't pass the statistical testing.

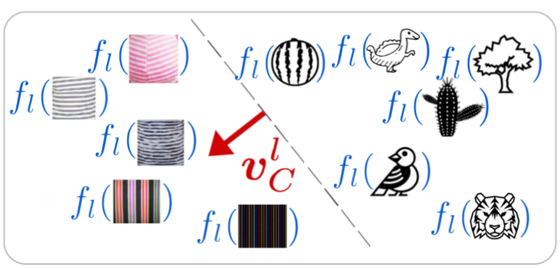
Recap TCAV:
Testing with Concept Activation Vectors

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

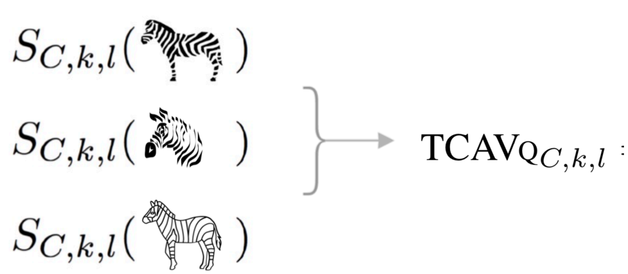
Even if your training data wasn't tagged with the concept

Even if your input feature did not include the concept

1. Learning CAVs



2. Getting TCAV score



3. CAV validation

Qualitative

Quantitative

Results

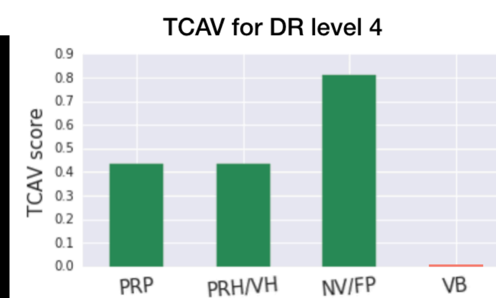
1. Sanity check experiment

cab image cab image with caption

2. Biases in Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy

DR level 4 Retina



Results

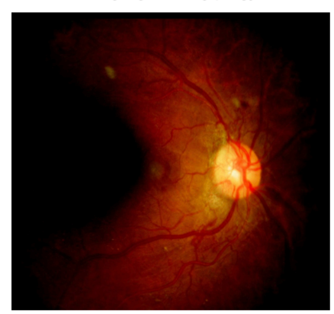
1. Sanity check experiment

cab image | cab image with caption

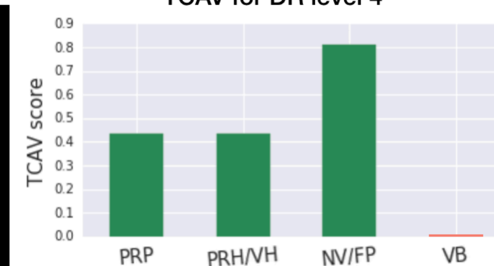
2. Biases from Inception V3 and GoogleNet

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

models can use either image or caption concept for classification.

caption
concept

An image
+
Potentially noisy Caption

0% noisy | 30% noisy | 100% noisy | no captions

Four models trained with
different caption noise levels

Sanity check experiment setup

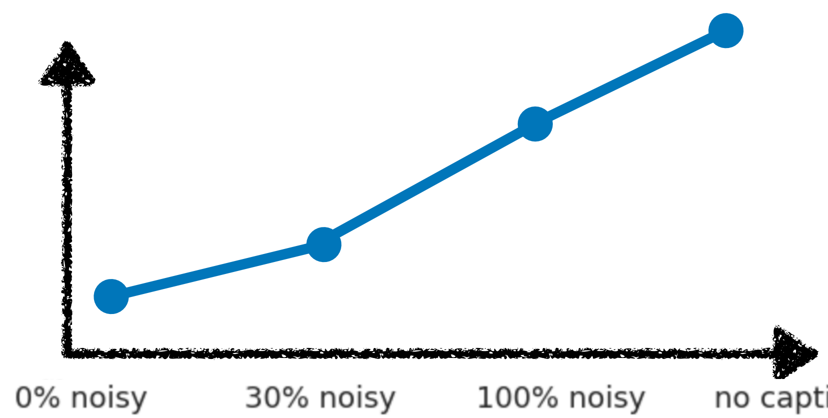
image concept

caption concept

models can use either image or caption concept for classification.

Test models with no caption image.

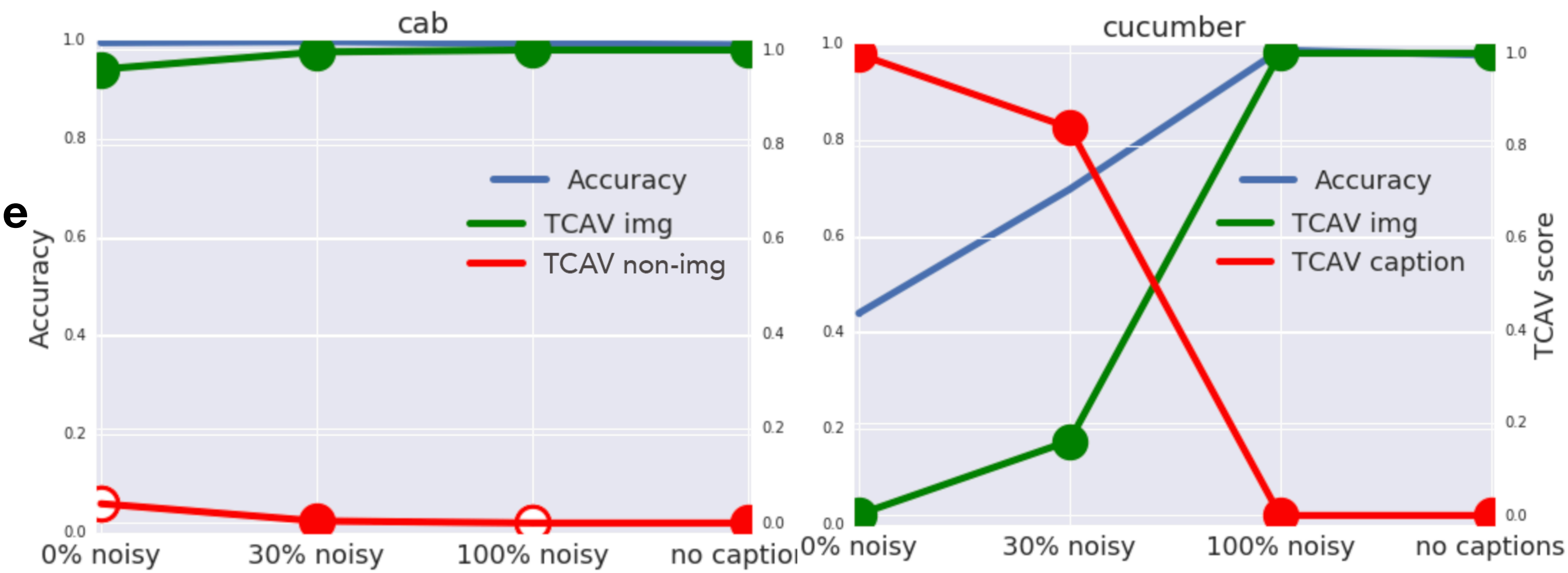
Test accuracy = Importance of image concept



Four models trained with different caption noise levels

Sanity check experiment

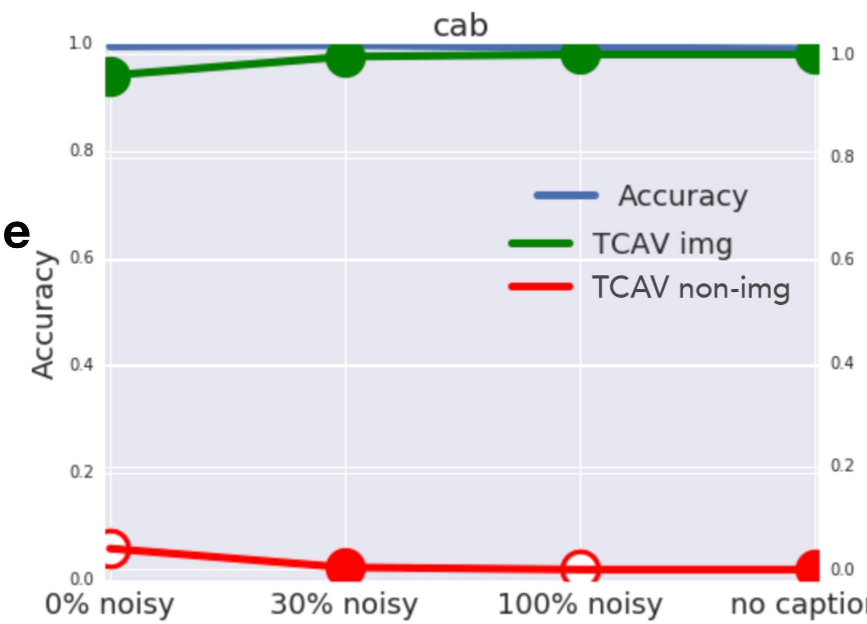
Test accuracy with no caption image



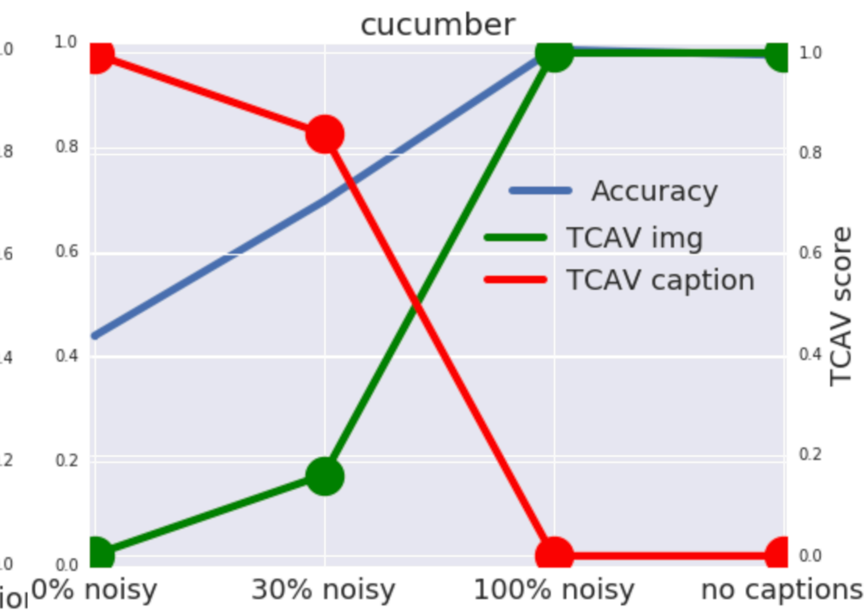
Caption noise level in training set **Caption noise level in training set**

Sanity check experiment

Test accuracy with no caption image



Caption noise level in training set



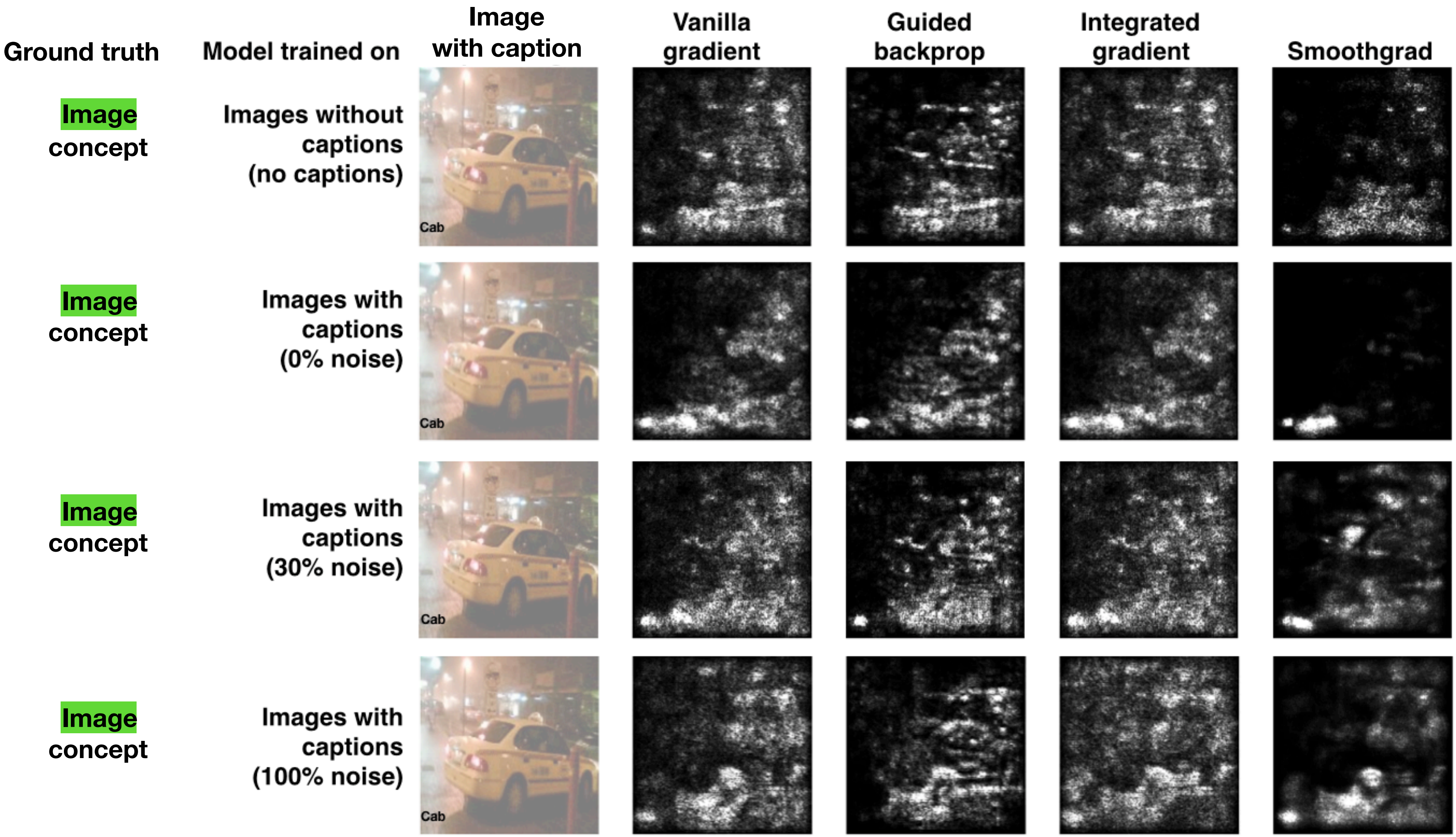
Caption noise level in training set

Cab class never cared about the caption!

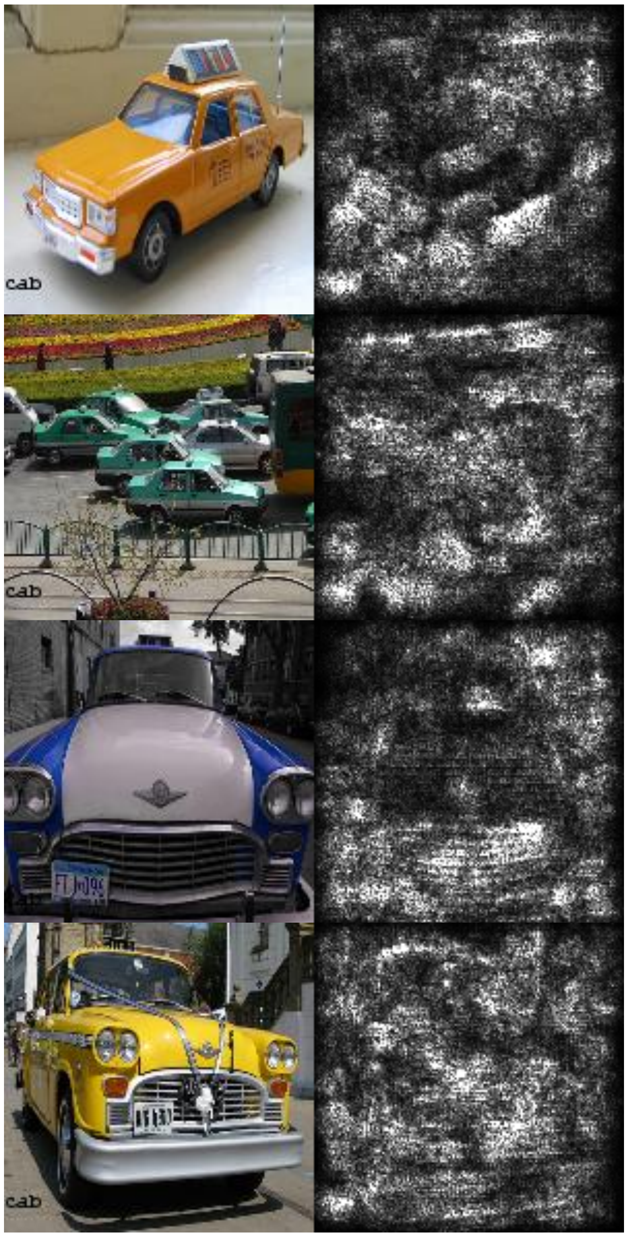
Cool, cool.

Can saliency maps do this too?

Can saliency maps communicate the same information?



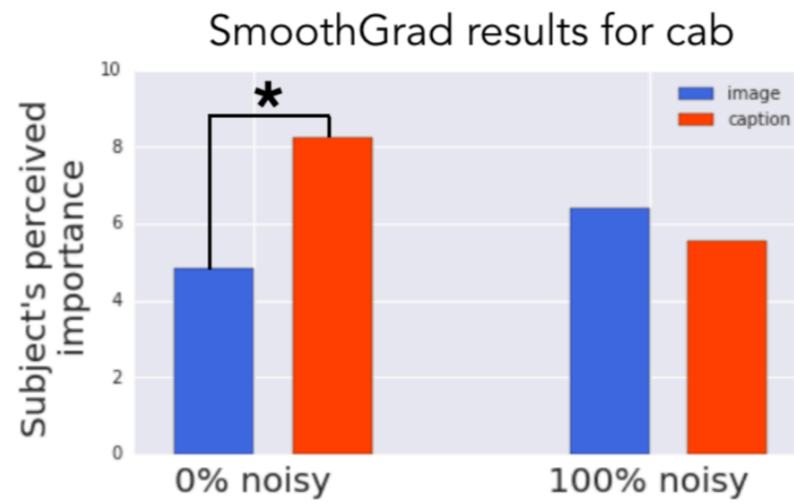
Human subject experiment: Can saliency maps communicate the same information?

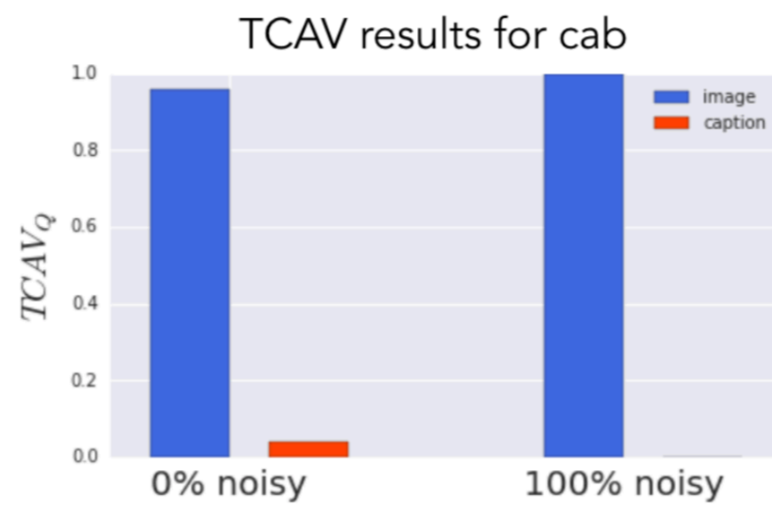


- 50 turkers are
  - asked to judge importance of image vs. caption given saliency maps.
  - asked to indicate their confidence
  - shown 3 classes (cab, zebra, cucumber) x 2 saliency maps for one model

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?

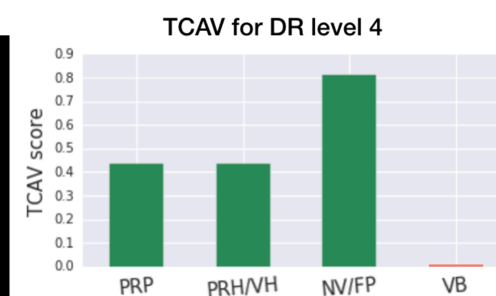
- Random chance: 50%
- Human performance with saliency map: 52%
- Humans can't agree: more than 50% no significant consensus
- Humans are **very** confident even when they are wrong.

Results

cab image cab image with caption

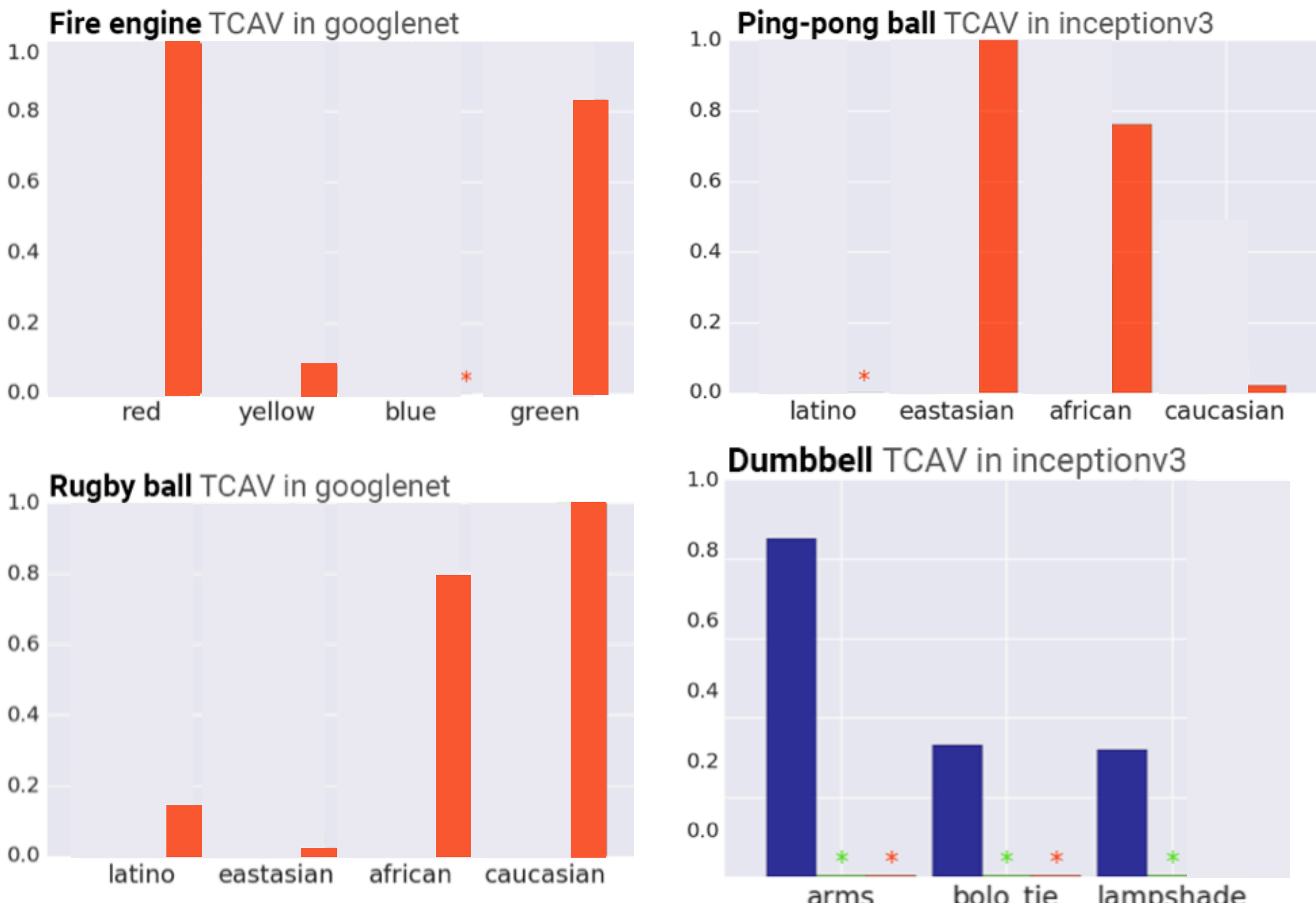
1. Sanity check experiment
2. Biases from Inception V3 and GoogleNet
3. Domain expert confirmation from Diabetic Retinopathy

DR level 4 Retina



TCAV in

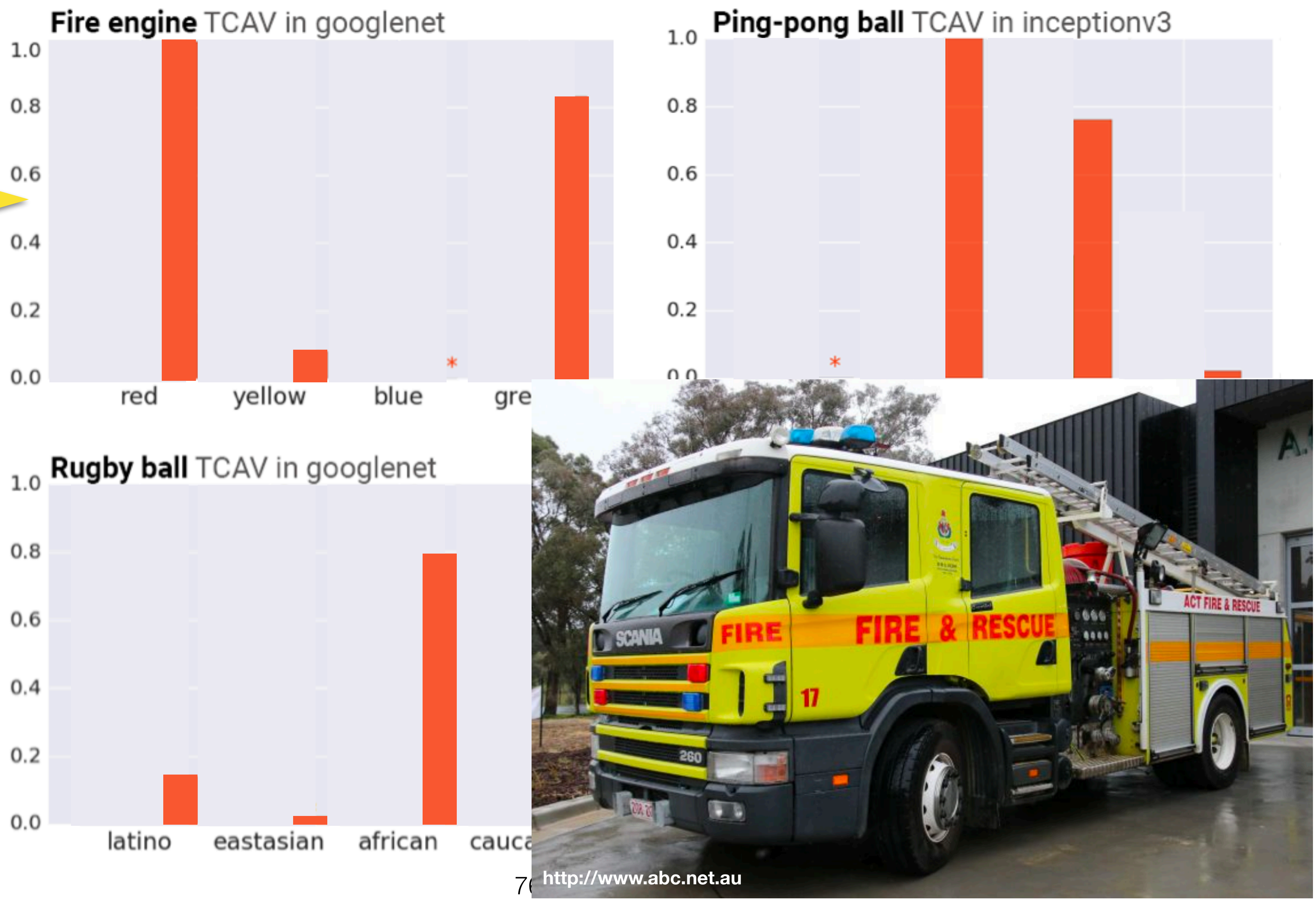
Two widely used image prediction models



TCAV in

Two widely used image prediction models

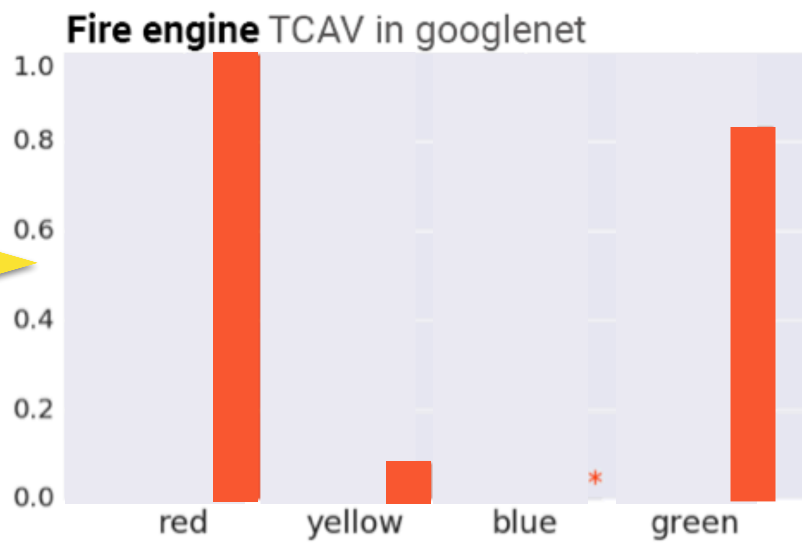
Geographical bias!

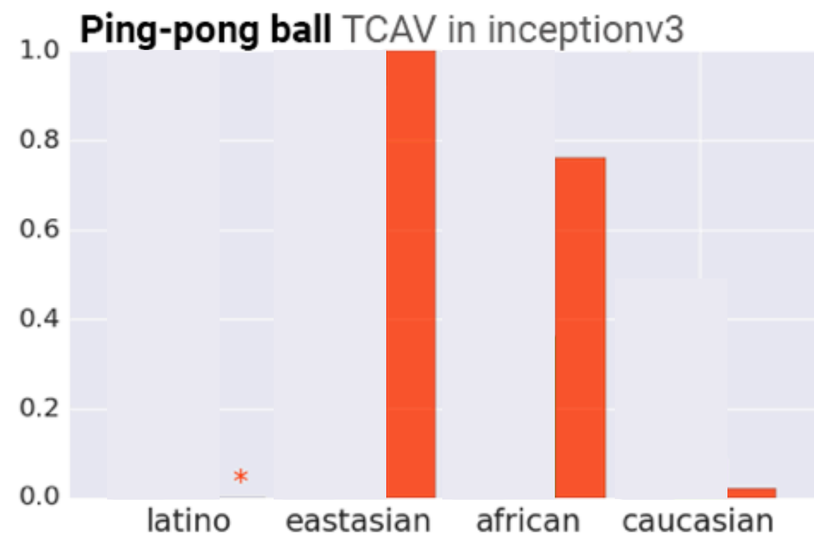


TCAV in

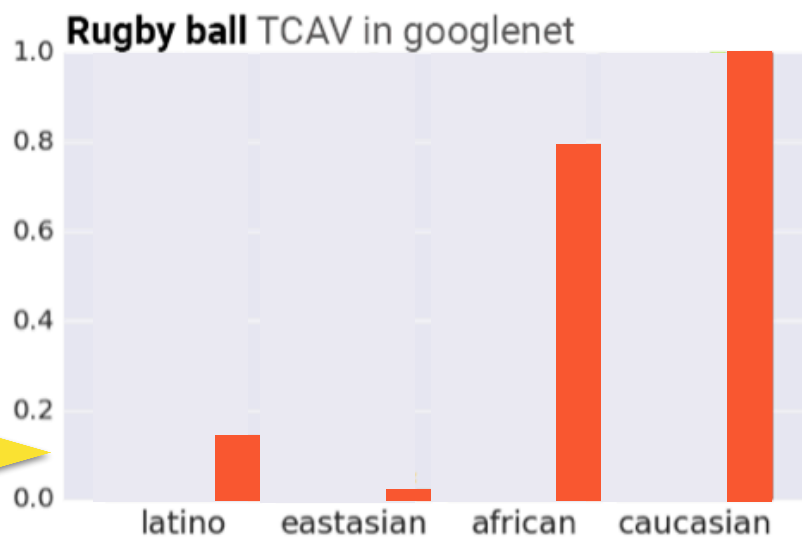
Two widely used image prediction models

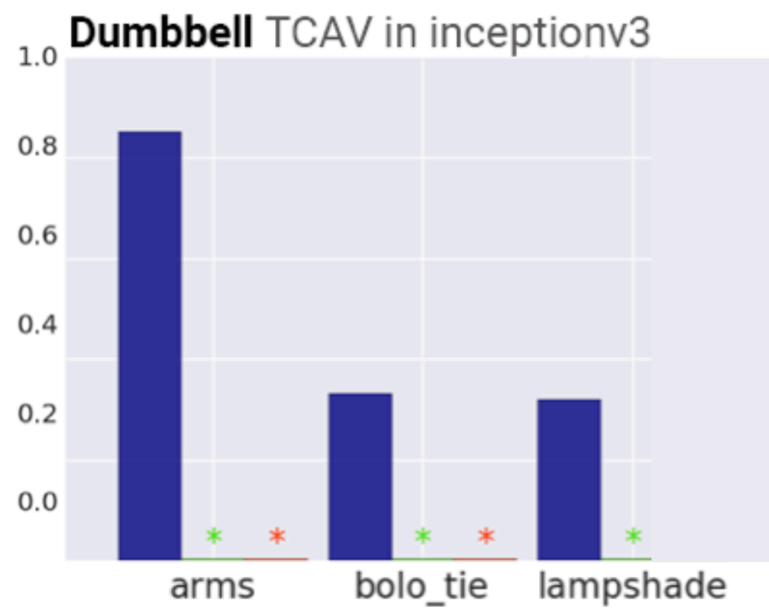
Geographical bias?





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

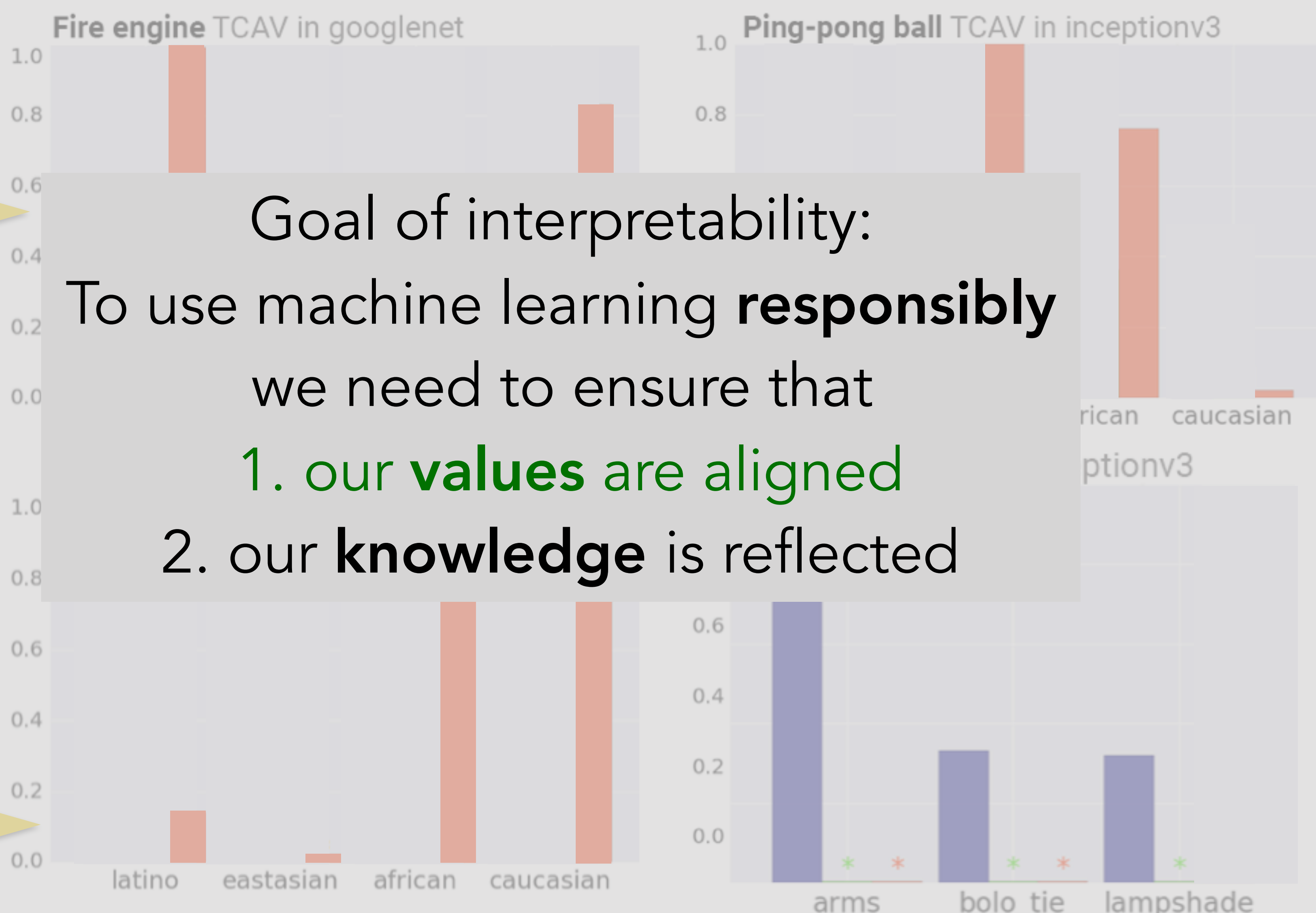




TCAV in

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

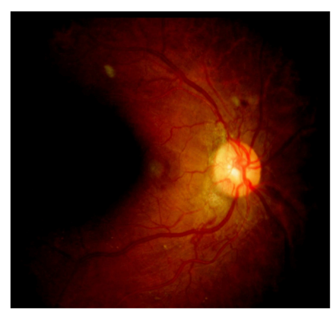
Results

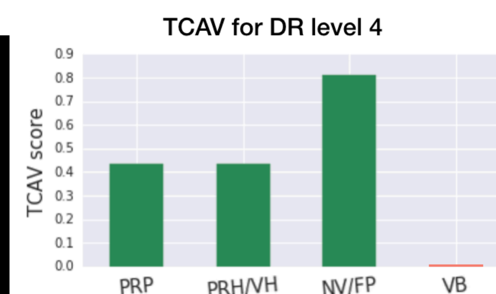
1. Sanity check experiment

cab image cab image with caption

2. Biases Inception V3 and GoogleNet

3. Domain expert confirmation from Diabetic Retinopathy

DR level 4 Retina




Diabetic Retinopathy

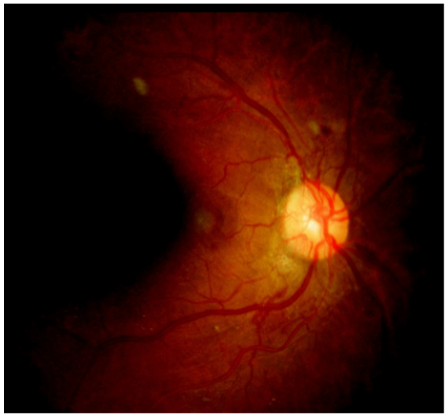
- Treatable but sight-threatening conditions
- Have model to with accurate prediction of DR (85%) [Krause et al., 2017]

Concepts the **ML model** uses

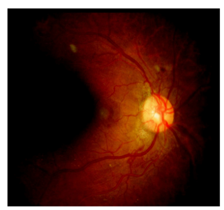
Vs

Diagnostic Concepts **human** doctors use

DR level 4 Retina



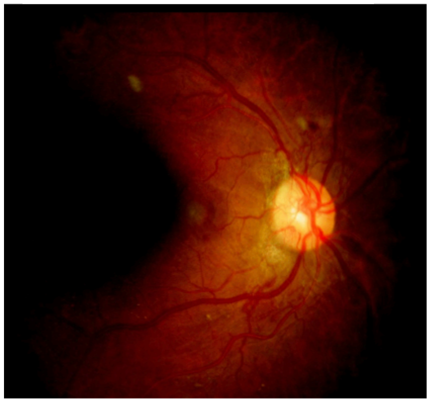
Collect human doctor’s knowledge

	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**

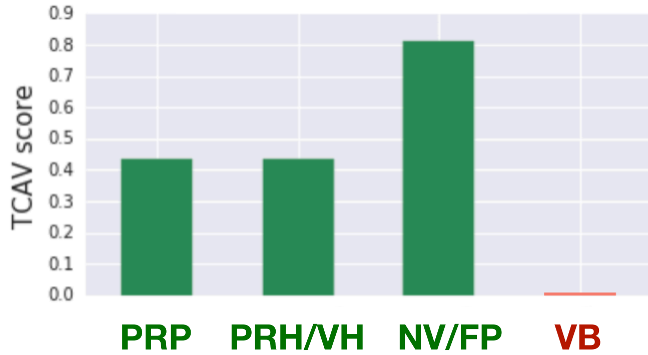
TCAV for Diabetic Retinopathy

Prediction class	Prediction accuracy
DR level 4	High

Example



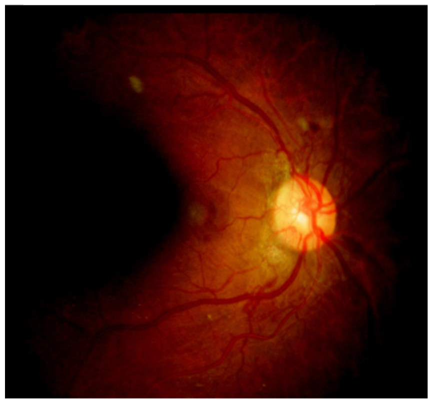
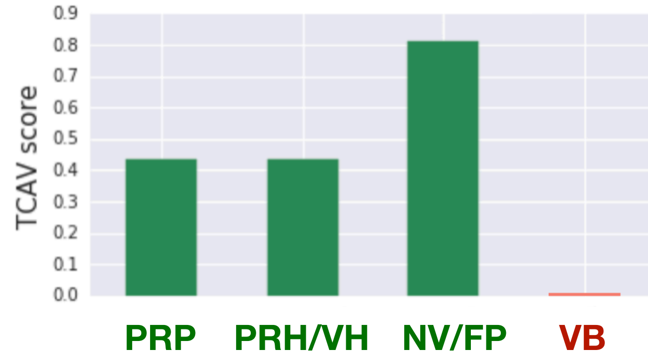
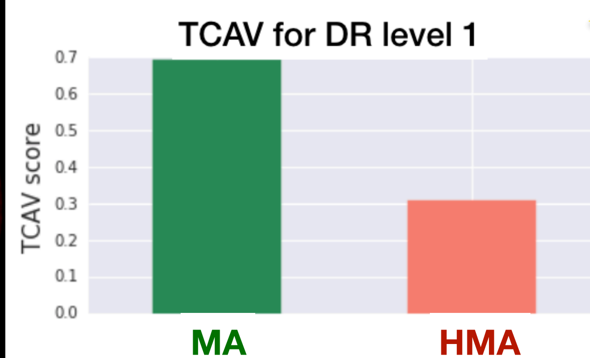
TCAV scores



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		
DR level 1	Med		

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

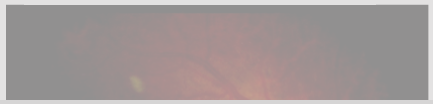
TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate

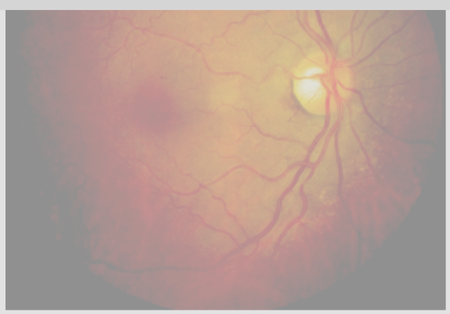
Green: domain expert's label on concepts belong to the level

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	





Goal of interpretability:

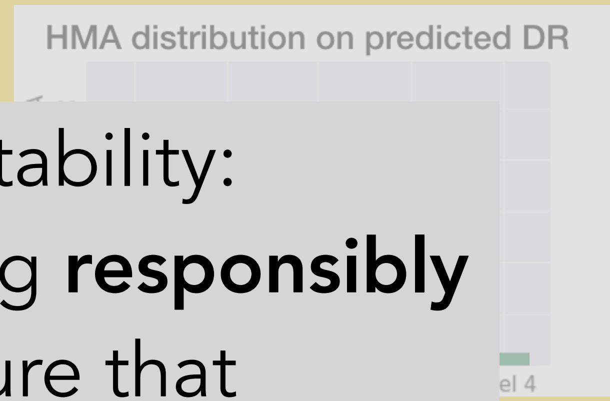
To use machine learning **responsibly**

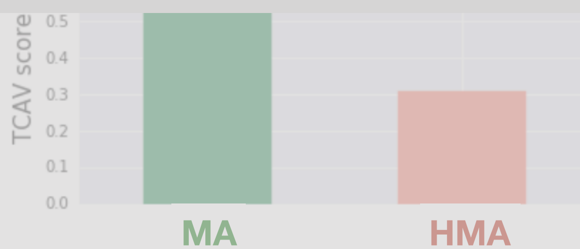
we need to ensure that

1. our **values** are aligned

2. our **knowledge** is reflected

Level 1 was often confused to level 2.





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

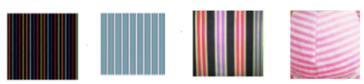
Green: domain expert's label on concepts belong to the level

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

github.com/tensorflow/tcav

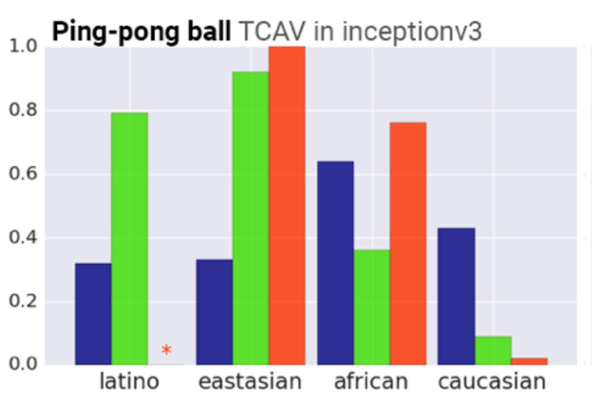
Summary:
Testing with Concept Activation Vectors

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



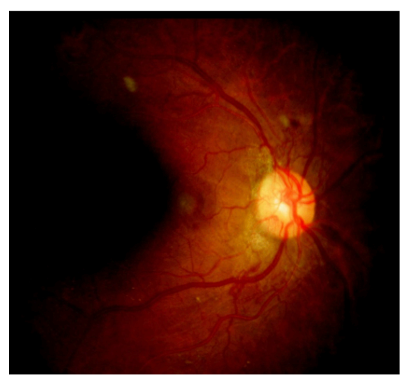
stripes concept (score: 0.9) was important to **zebra** class for this trained network.

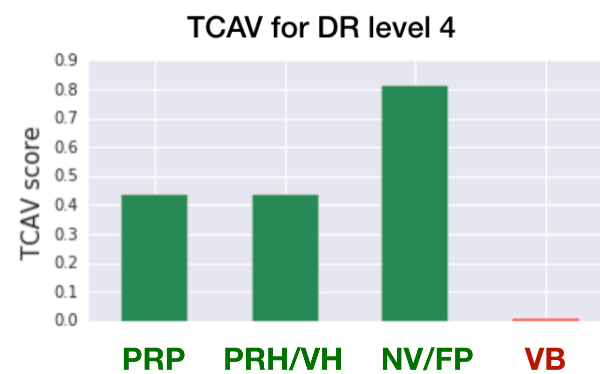
TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.



Our values

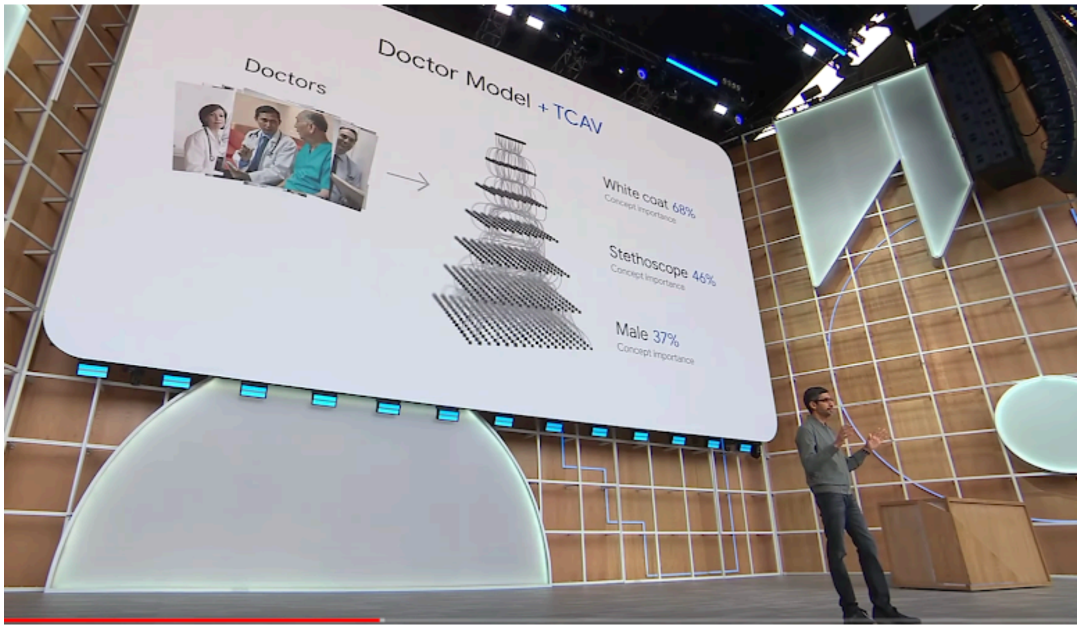
DR level 4 Retina





Our knowledge

Responses from outside of academia



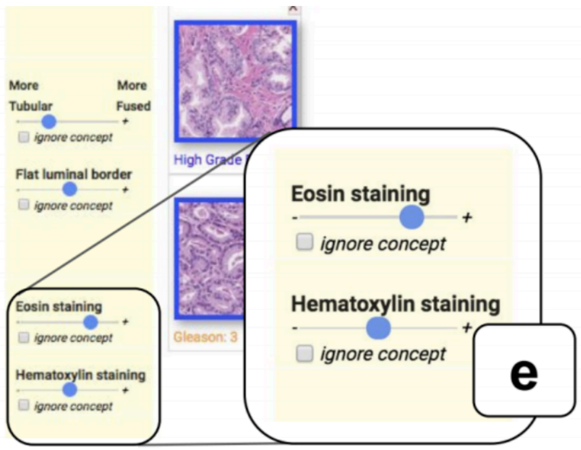
Sundar (CEO of Google) explaining how TCAV works in his keynote at Google I/O 2019

UNESCO

UNESCO NetExplo award 2019

Selected as one of ten "cutting-edge digital innovations with the potential of profound and lasting impact."

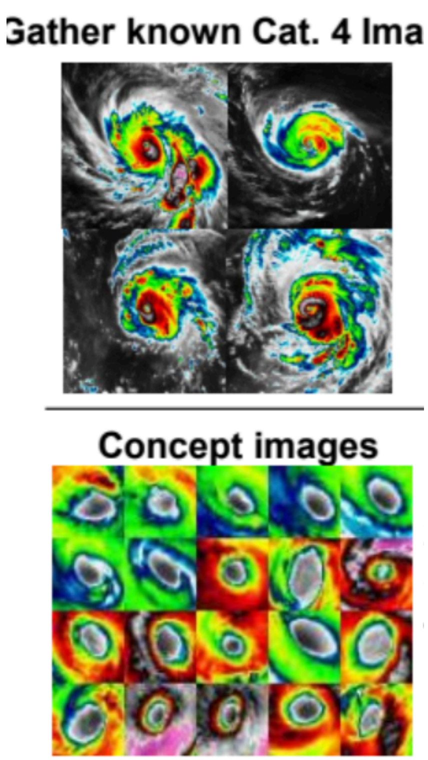
Responses from inside of academia



Using CAVs to help doctors find more diagnostically relevant images

"Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making" Work by Carrie J. Cai et al.

CHI conference, best paper honorable mention

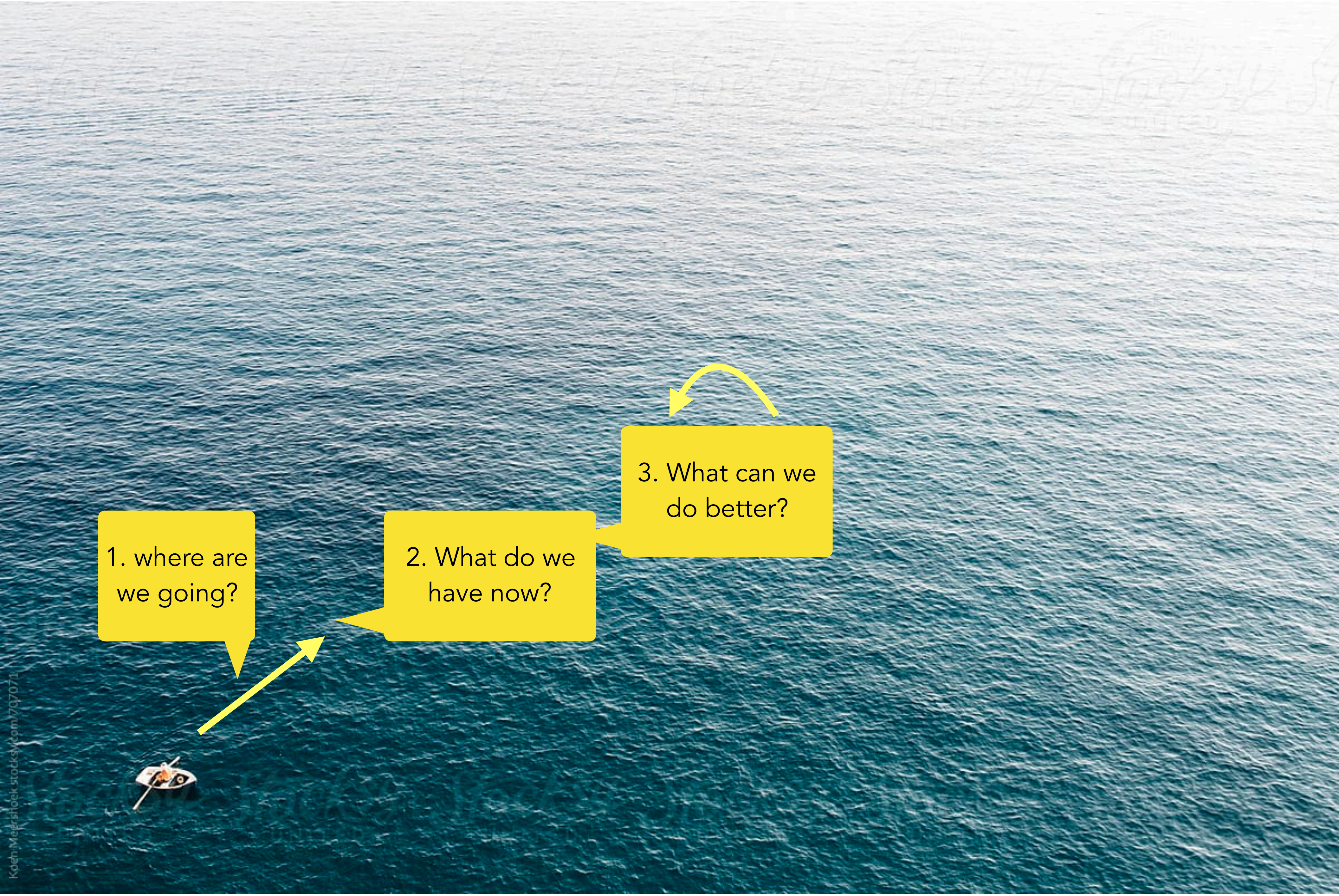


TCAV for storm prediction models

"Interpretable AI for deep-learning based meteorological applications" Work by Eric Wendoloski,

Extending TCAV to regression models

"Regression Concept Vectors for Bidirectional Explanations in Histopathology" Work by Mara Graziani et al.

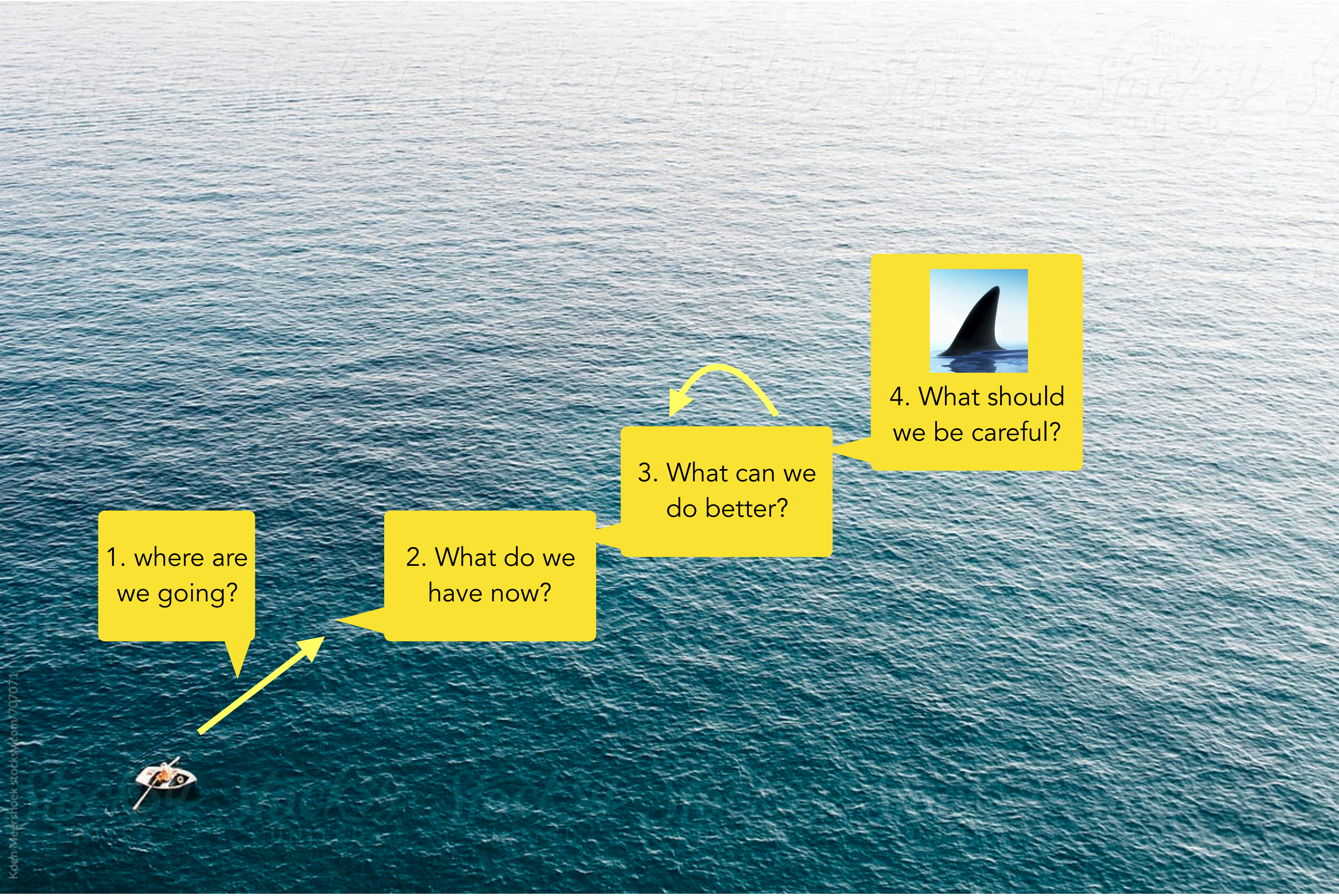

1. where are we going?
2. What do we have now?
3. What can we do better?
Koen Meershoek stocksy.com/707071

Limitations of TCAV

- Concept has to ‘expressible’ using examples (e.g., “love” concept might be hard).

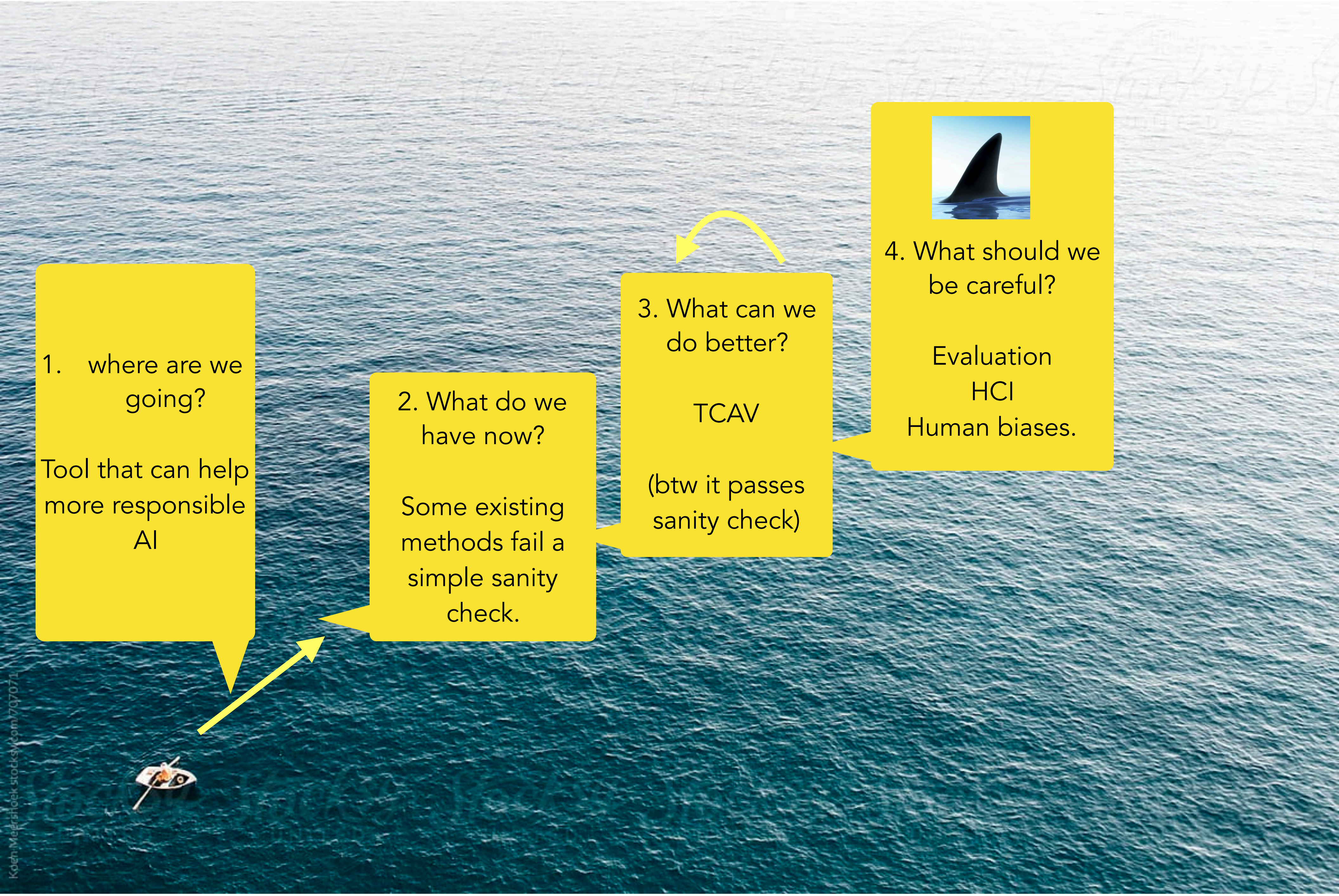
- User needs to know which concepts they want to test, and have examples for it. Follow-up work to automatically discover concepts for images (submitted), but many more directions are possible.

- Explanations provided by TCAV are not-causal - Follow-up work on causal TCAV (submitted)


1. where are we going?
2. What do we have now?
3. What can we do better?
4. What should we be careful?

Things to keep in mind during our journey.

- Proper evaluations
  - Sanity check and ground-truth-based evaluations
  - Test with humans!
- Remember that humans are biased and irrational.
- Designing the right interaction - HCI.
- Try to criticize - think about what wasn't talked about in this talk but should have!
- Keep checking if we are going to the right direction!


1. where are we going?
Tool that can help more responsible AI
2. What do we have now?
Some existing methods fail a simple sanity check.
3. What can we do better?
TCAV
(btw it passes sanity check)
4. What should we be careful?
Evaluation
HCI
Human biases.