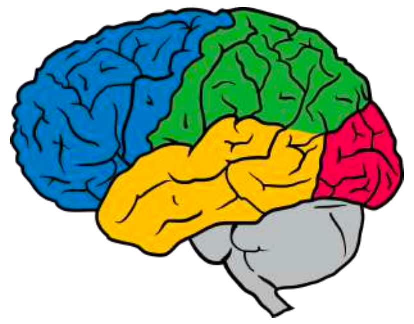
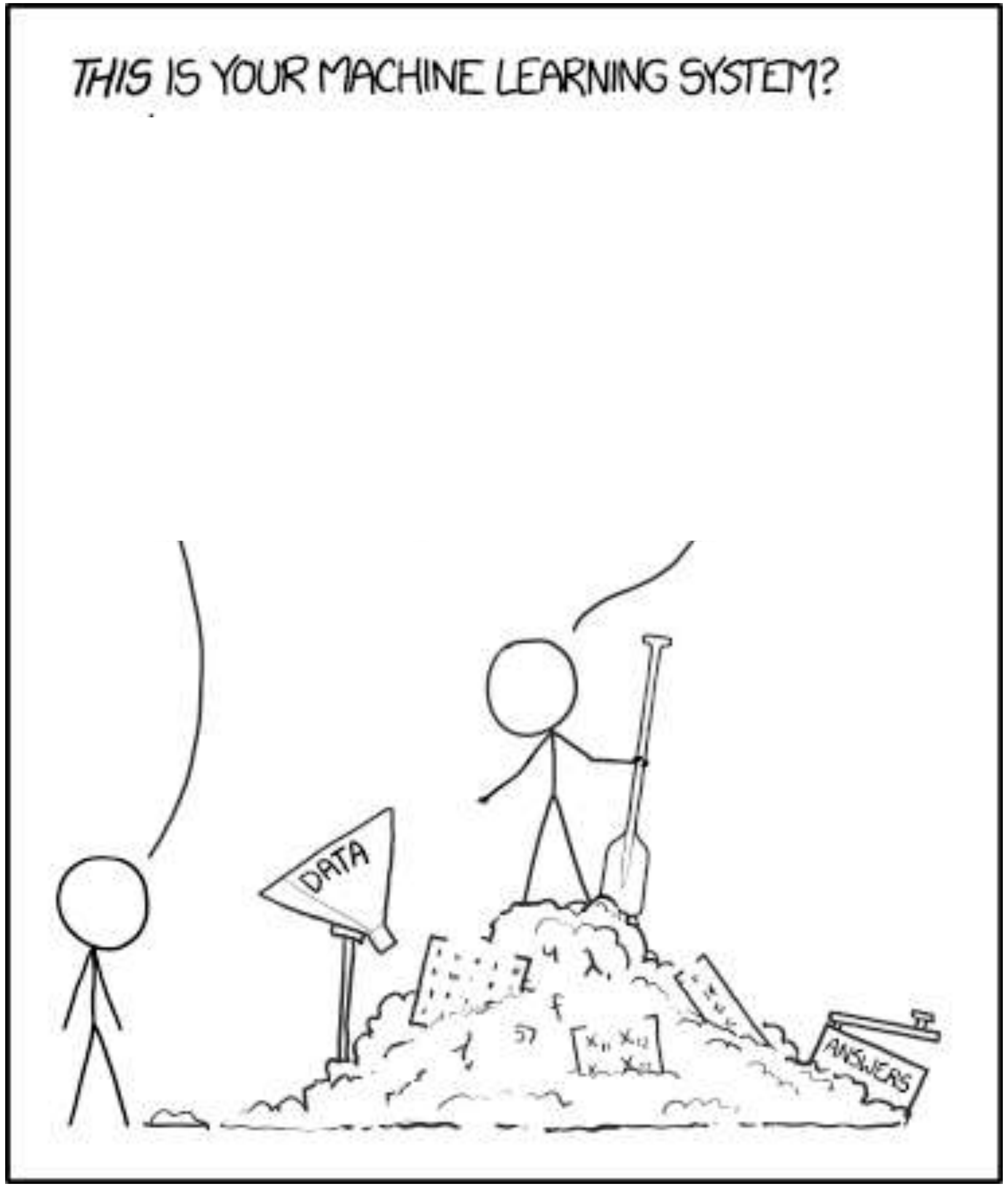


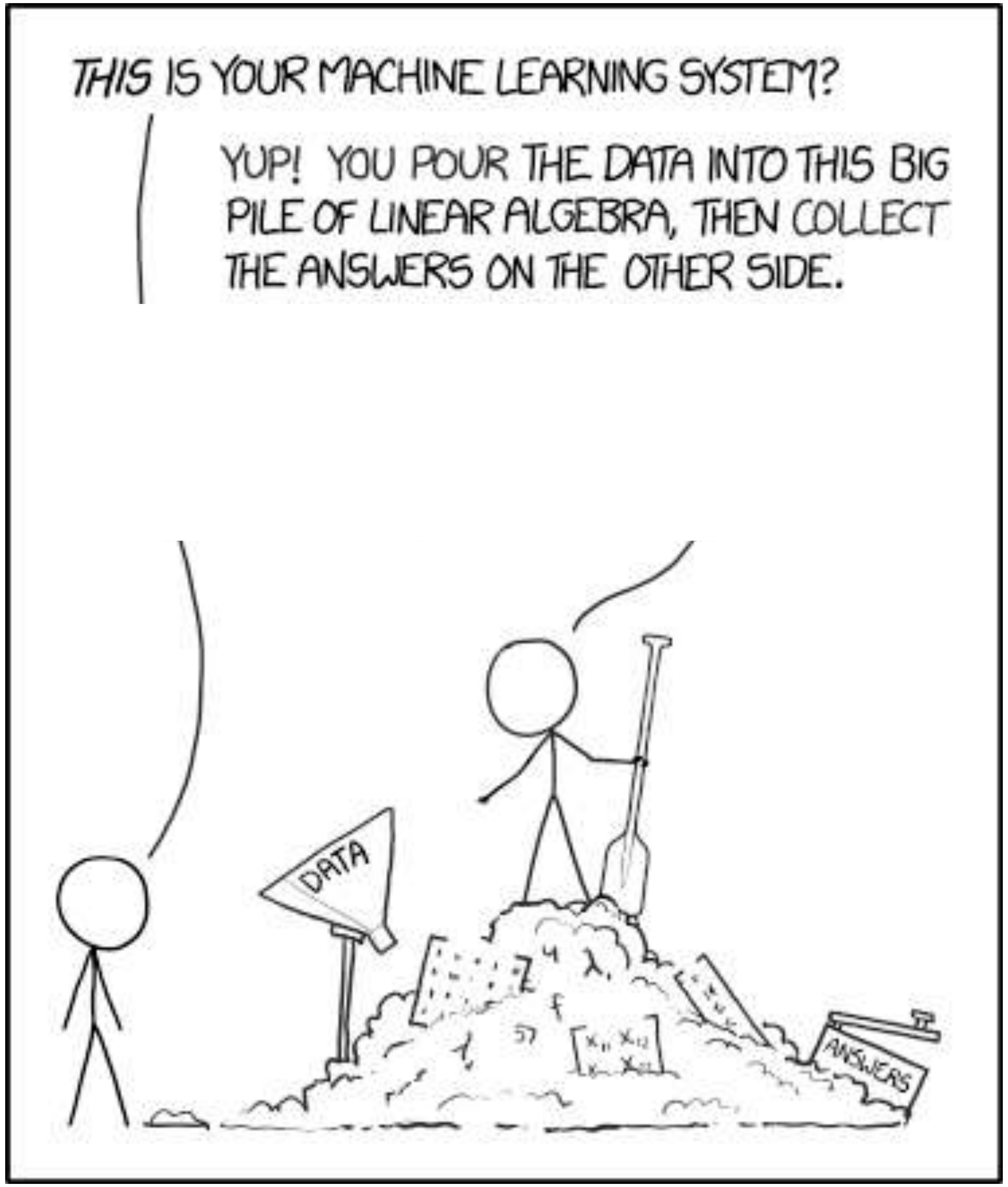
MLSS 2021
Interpretability

why, what and how to

Been Kim



https://xkcd.com/



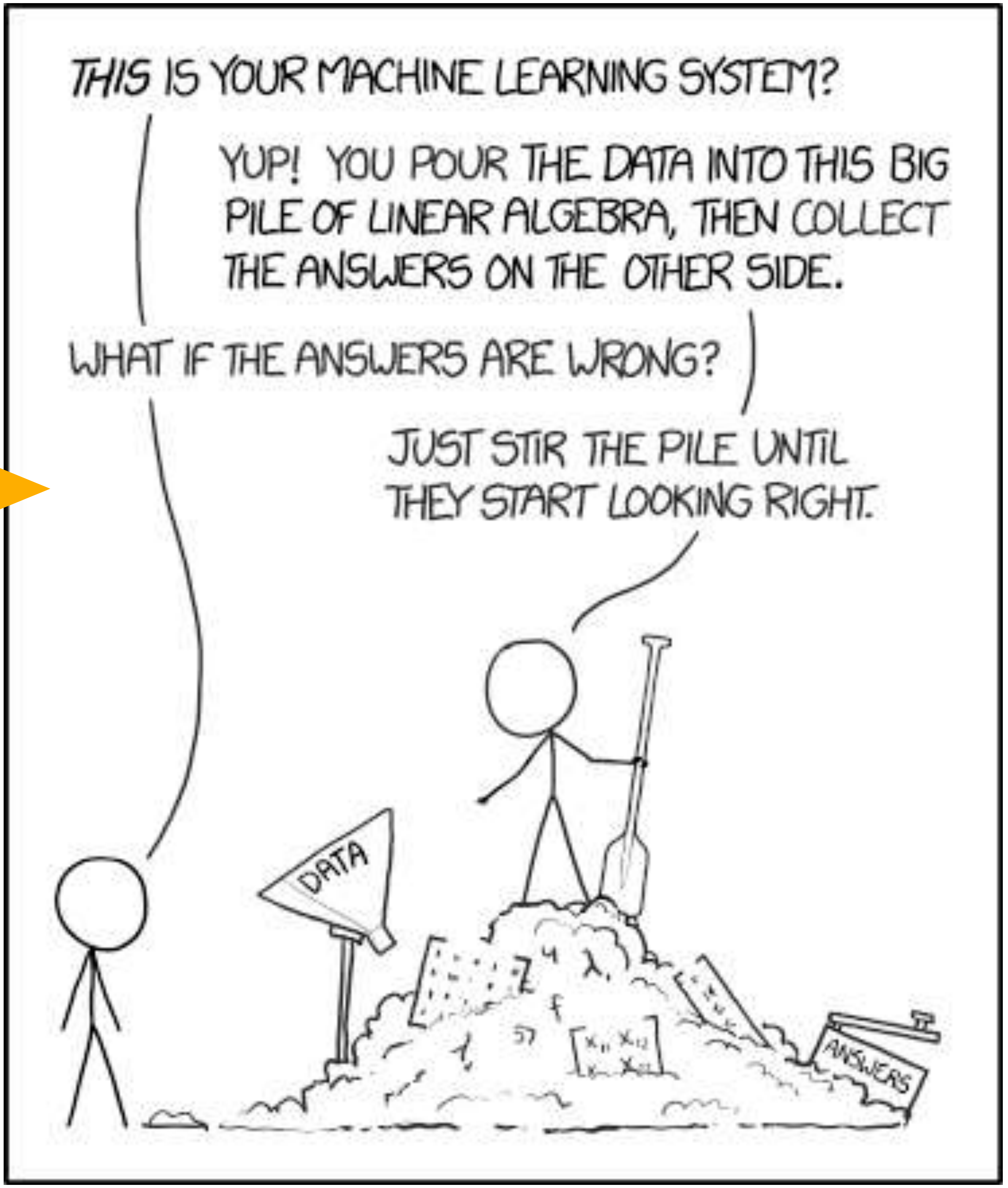
https://xkcd.com/

https://xkcd.com/

https://xkcd.com/

Oh no.

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

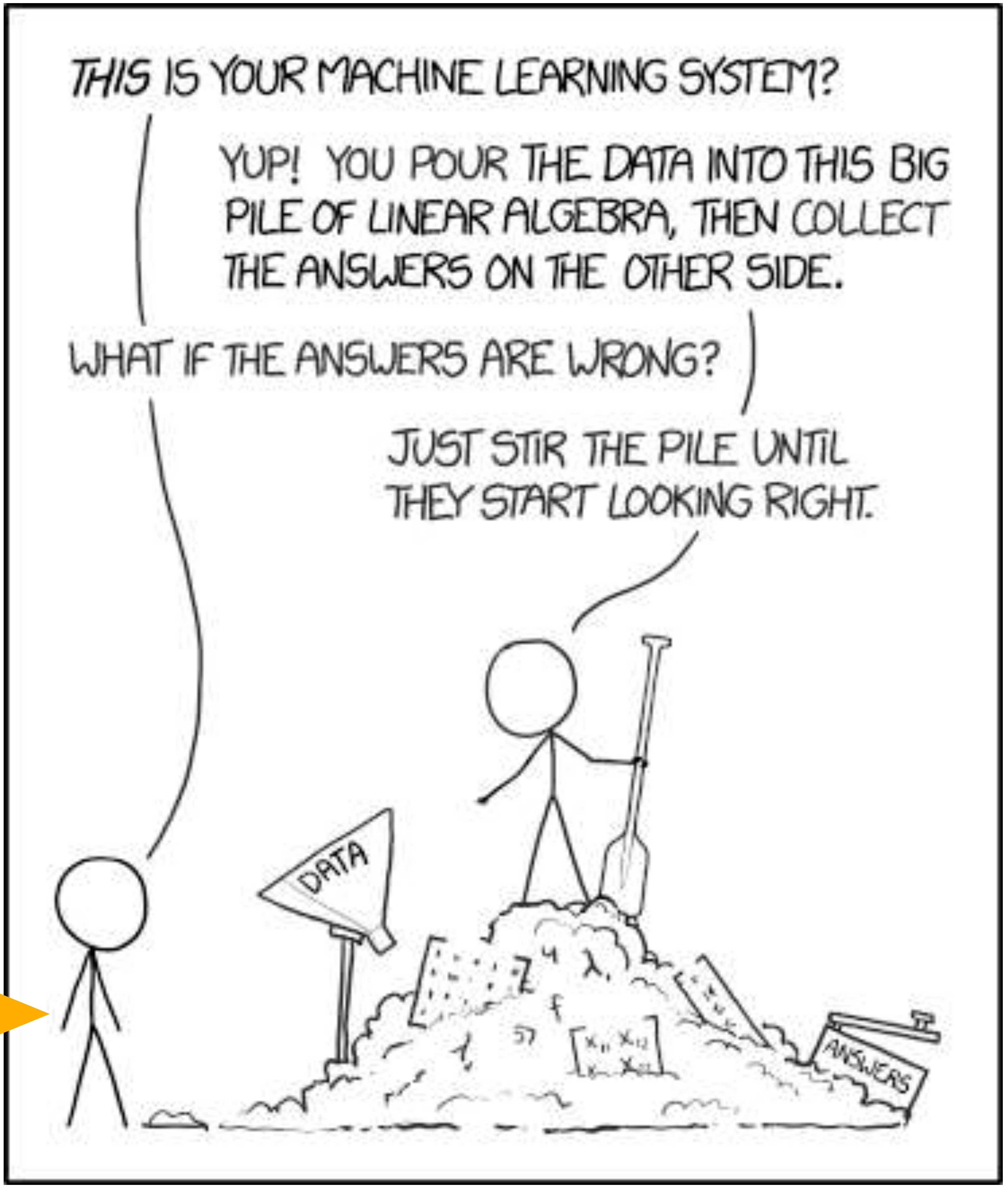


https://xkcd.com/

For this reason, this tutorial won't be about list and lists of methods.

Focusing on more methods is not what we need.

Instead, it'll be about more Important things.



https://xkcd.com/

Agenda

- What and why
- !Caution!: Things to be careful when using and developing interpretability methods
- Evaluate: How to evaluate interpretability methods
- Methods: 3 types of methods and examples

Agenda

- What and why
- !Caution!: Things to be careful when using and developing interpretability methods
- Evaluate: How to evaluate interpretability methods
- Methods: 3 types of methods and examples

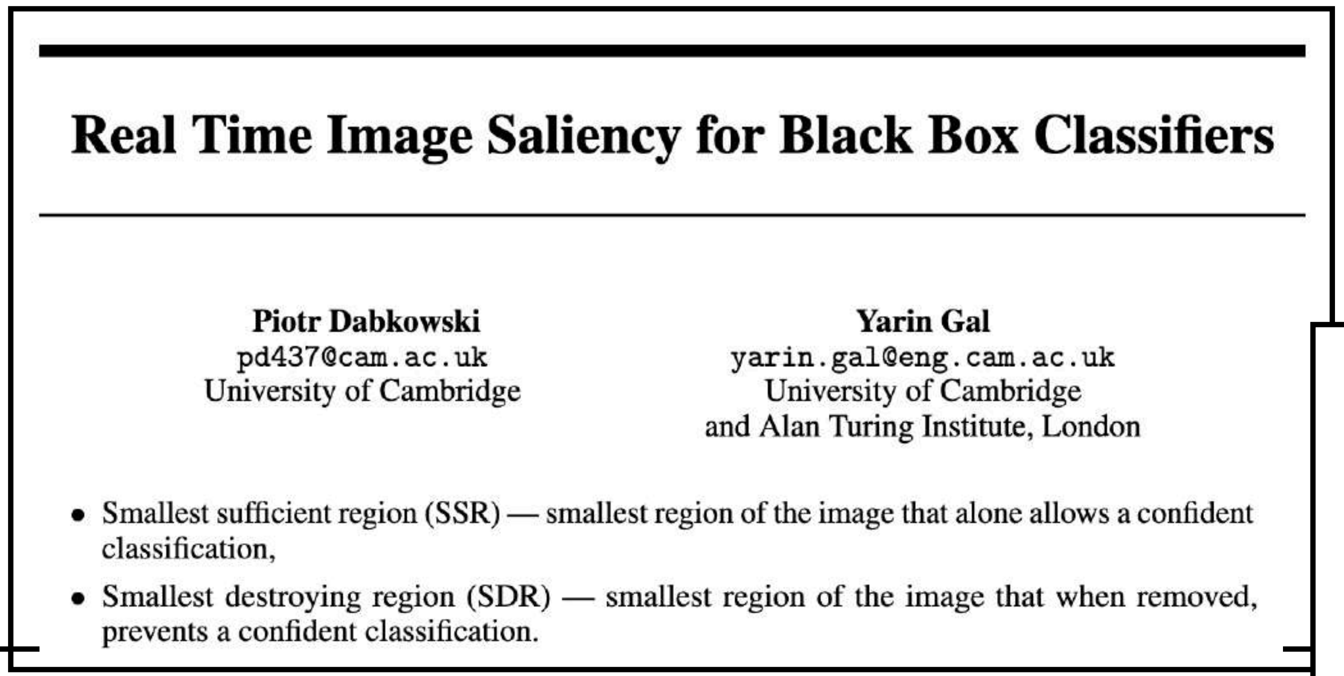
What do we mean by interpretability?

- In a dictionary (Merriam-Webster):
  - “to explain or to present in understandable terms”
- In ML (among many)
  - “ability to explain or to present in understandable terms to a human”[Doshi-Velez, K. 16]
  - “Interpretability is the degree to which a human can understand the cause of a decision.” [Miller 17]
- In cognitive science (among many)
  - “explanations are... the currency in which we exchanged beliefs” [Lombrozo 06]

Sure, but how do we make a working definition for my paper?

Operationalizing interpretability

- Define your desiderata - clearly specify what your definition is, and what you are optimizing.
- Do proper quantitative and qualitative evaluation with your end-task in mind -'users like the explanation' says nothing (more on this later)

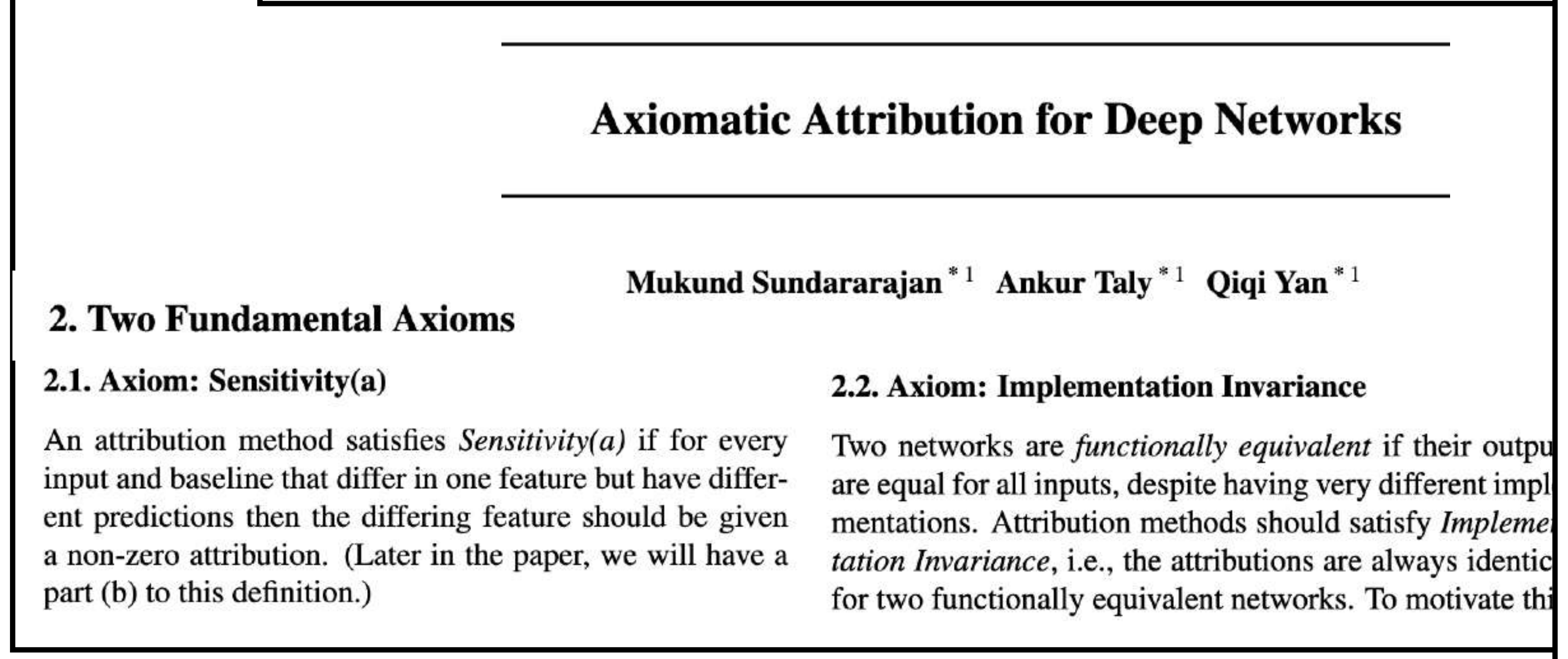


Real Time Image Saliency for Black Box Classifiers

Piotr Dabkowski
`pd437@cam.ac.uk`
University of Cambridge

Yarin Gal
`yarin.gal@eng.cam.ac.uk`
University of Cambridge
and Alan Turing Institute, London

- Smallest sufficient region (SSR) — smallest region of the image that alone allows a confident classification,
- Smallest destroying region (SDR) — smallest region of the image that when removed, prevents a confident classification.



Axiomatic Attribution for Deep Networks

Mukund Sundararajan [*1] **Ankur Taly** [*1] **Qiqi Yan** [*1]

2. Two Fundamental Axioms

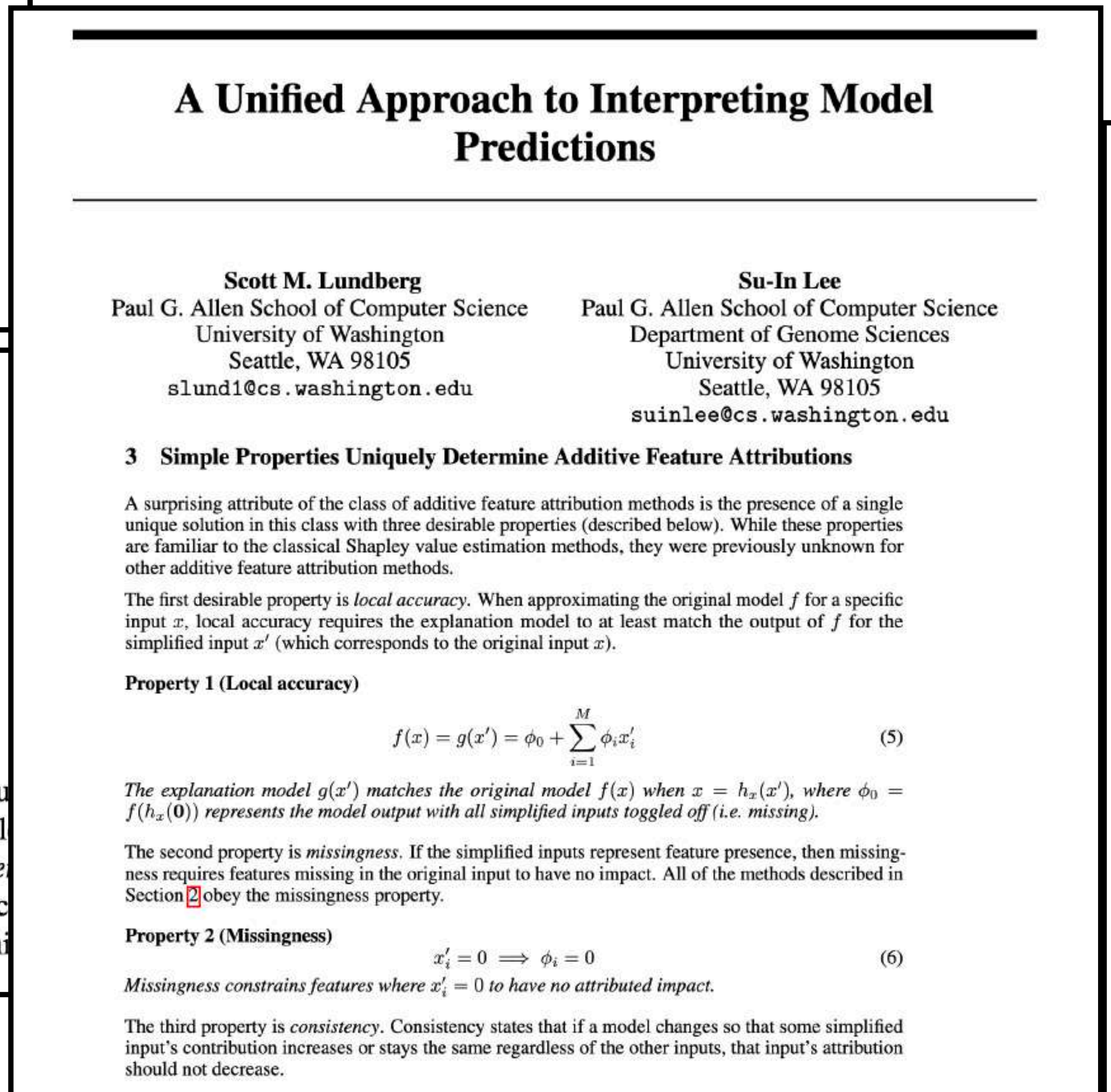
2.1. Axiom: Sensitivity(a)

An attribution method satisfies *Sensitivity(a)* if for every input and baseline that differ in one feature but have different predictions then the differing feature should be given a non-zero attribution. (Later in the paper, we will have a part (b) to this definition.)

2.2. Axiom: Implementation Invariance

Two networks are *functionally equivalent* if their outpu are equal for all inputs, despite having very different impl mentations. Attribution methods should satisfy *Impleme tation Invariance*, i.e., the attributions are always identic for two functionally equivalent networks. To motivate thi

---



A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
`slund1@cs.washington.edu`

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
`suinlee@cs.washington.edu`

3 Simple Properties Uniquely Determine Additive Feature Attributions

A surprising attribute of the class of additive feature attribution methods is the presence of a single unique solution in this class with three desirable properties (described below). While these properties are familiar to the classical Shapley value estimation methods, they were previously unknown for other additive feature attribution methods.

The first desirable property is *local accuracy*. When approximating the original model f for a specific input x, local accuracy requires the explanation model to at least match the output of f for the simplified input x' (which corresponds to the original input x).

Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{5}$$

The explanation model $g(x')$ matches the original model $f(x)$ when $x = h_x(x')$, where $\phi_0 = f(h_x(\mathbf{0}))$ represents the model output with all simplified inputs toggled off (i.e. missing).

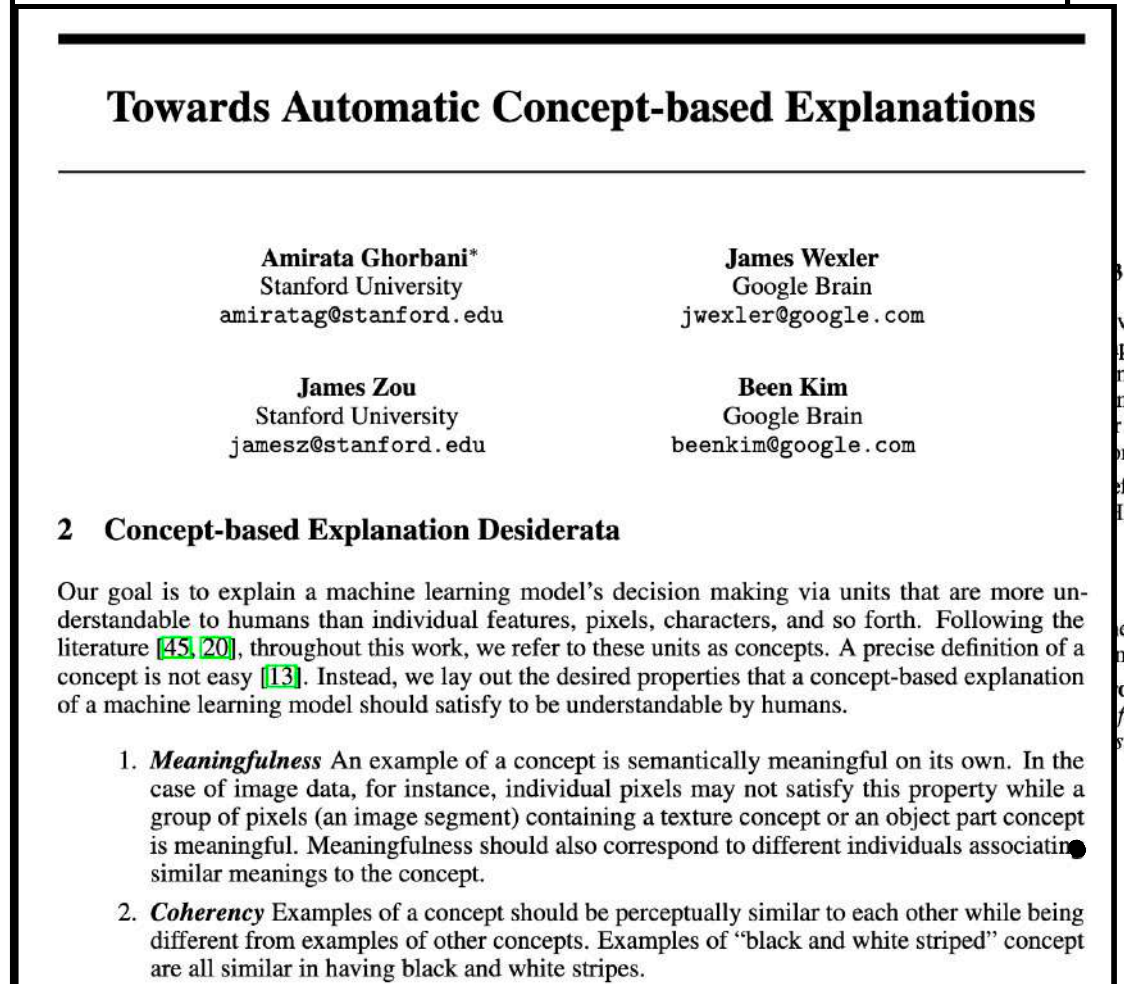
The second property is *missingness*. If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact. All of the methods described in Section 2 obey the missingness property.

Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \tag{6}$$

Missingness constrains features where $x'_i = 0$ to have no attributed impact.

The third property is *consistency*. Consistency states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.



Towards Automatic Concept-based Explanations

Amirata Ghorbani[*]
Stanford University
`amiratag@stanford.edu`

James Wexler
Google Brain
`jwexler@google.com`

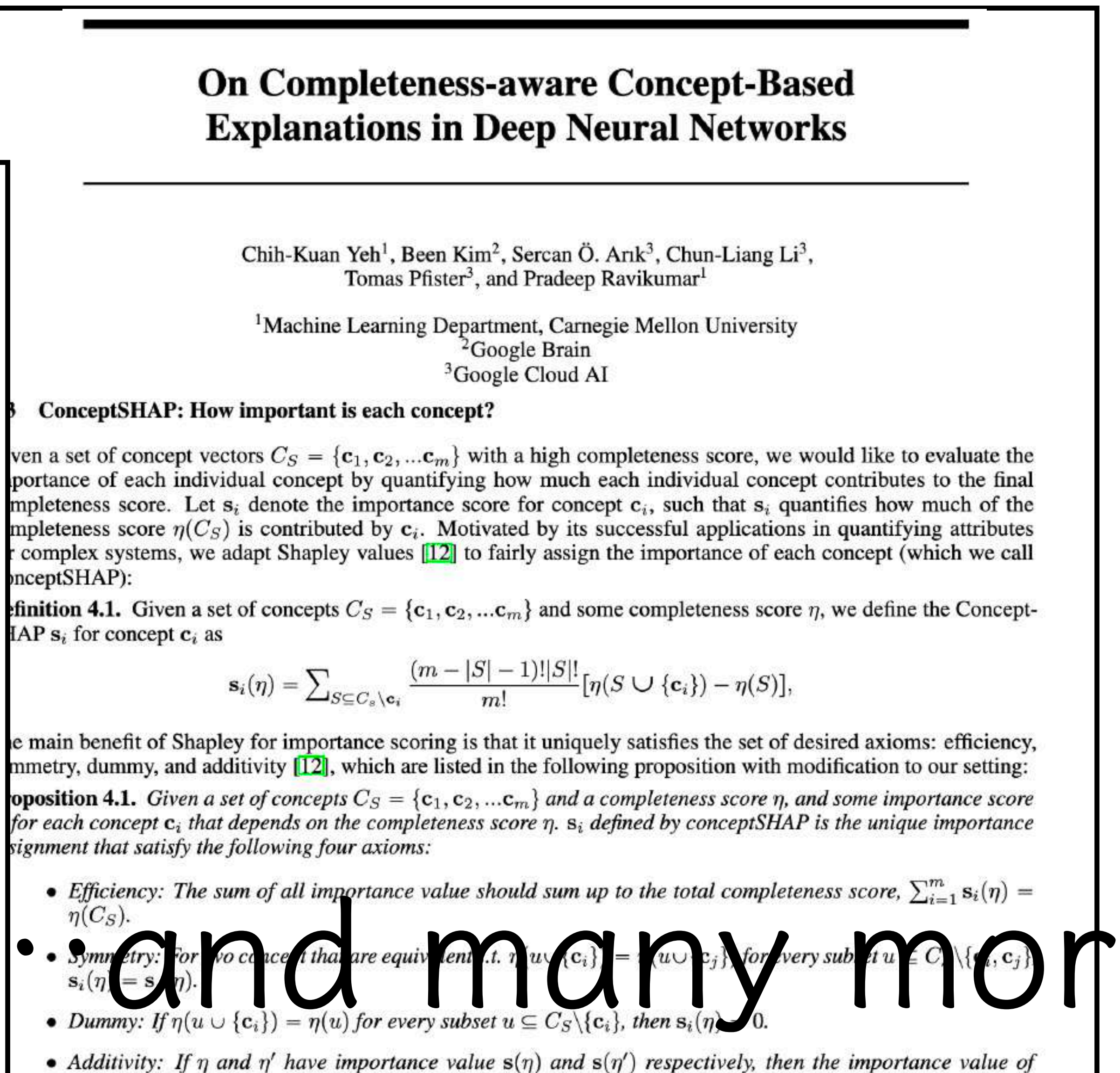
James Zou
Stanford University
`jamesz@stanford.edu`

Been Kim
Google Brain
`beenkim@google.com`

2 Concept-based Explanation Desiderata

Our goal is to explain a machine learning model's decision making via units that are more understandable to humans than individual features, pixels, characters, and so forth. Following the literature [45, 20], throughout this work, we refer to these units as concepts. A precise definition of a concept is not easy [13]. Instead, we lay out the desired properties that a concept-based explanation of a machine learning model should satisfy to be understandable by humans.

1. ***Meaningfulness*** An example of a concept is semantically meaningful on its own. In the case of image data, for instance, individual pixels may not satisfy this property while a group of pixels (an image segment) containing a texture concept or an object part concept is meaningful. Meaningfulness should also correspond to different individuals associating similar meanings to the concept.
2. ***Coherency*** Examples of a concept should be perceptually similar to each other while being different from examples of other concepts. Examples of "black and white striped" concept are all similar in having black and white stripes.



On Completeness-aware Concept-Based Explanations in Deep Neural Networks

Chih-Kuan Yeh[1], Been Kim[2], Sercan Ö. Arık[3], Chun-Liang Li[3],
Tomas Pfister[3], and Pradeep Ravikumar[1]

[1]Machine Learning Department, Carnegie Mellon University
[2]Google Brain
[3]Google Cloud AI

ConceptSHAP: How important is each concept?

ven a set of concept vectors $C_S = \{\mathbf{c}_1, \mathbf{c}_2, ...\mathbf{c}_m\}$ with a high completeness score, we would like to evaluate the portance of each individual concept by quantifying how much each individual concept contributes to the final mpleteness score. Let $\mathbf{s}_i$ denote the importance score for concept $\mathbf{c}_i$, such that $\mathbf{s}_i$ quantifies how much of the mpleteness score $\eta(C_S)$ is contributed by $\mathbf{c}_i$. Motivated by its successful applications in quantifying attributes complex systems, we adapt Shapley values [12] to fairly assign the importance of each concept (which we call nceptSHAP):

finition 4.1. Given a set of concepts $C_S = \{\mathbf{c}_1, \mathbf{c}_2, ...\mathbf{c}_m\}$ and some completeness score η, we define the Concept-AP $\mathbf{s}_i$ for concept $\mathbf{c}_i$ as

$$\mathbf{s}_i(\eta) = \sum_{S \subseteq C_s \setminus \mathbf{c}_i} \frac{(m - |S| - 1)!|S|!}{m!}[\eta(S \cup \{\mathbf{c}_i\}) - \eta(S)],$$

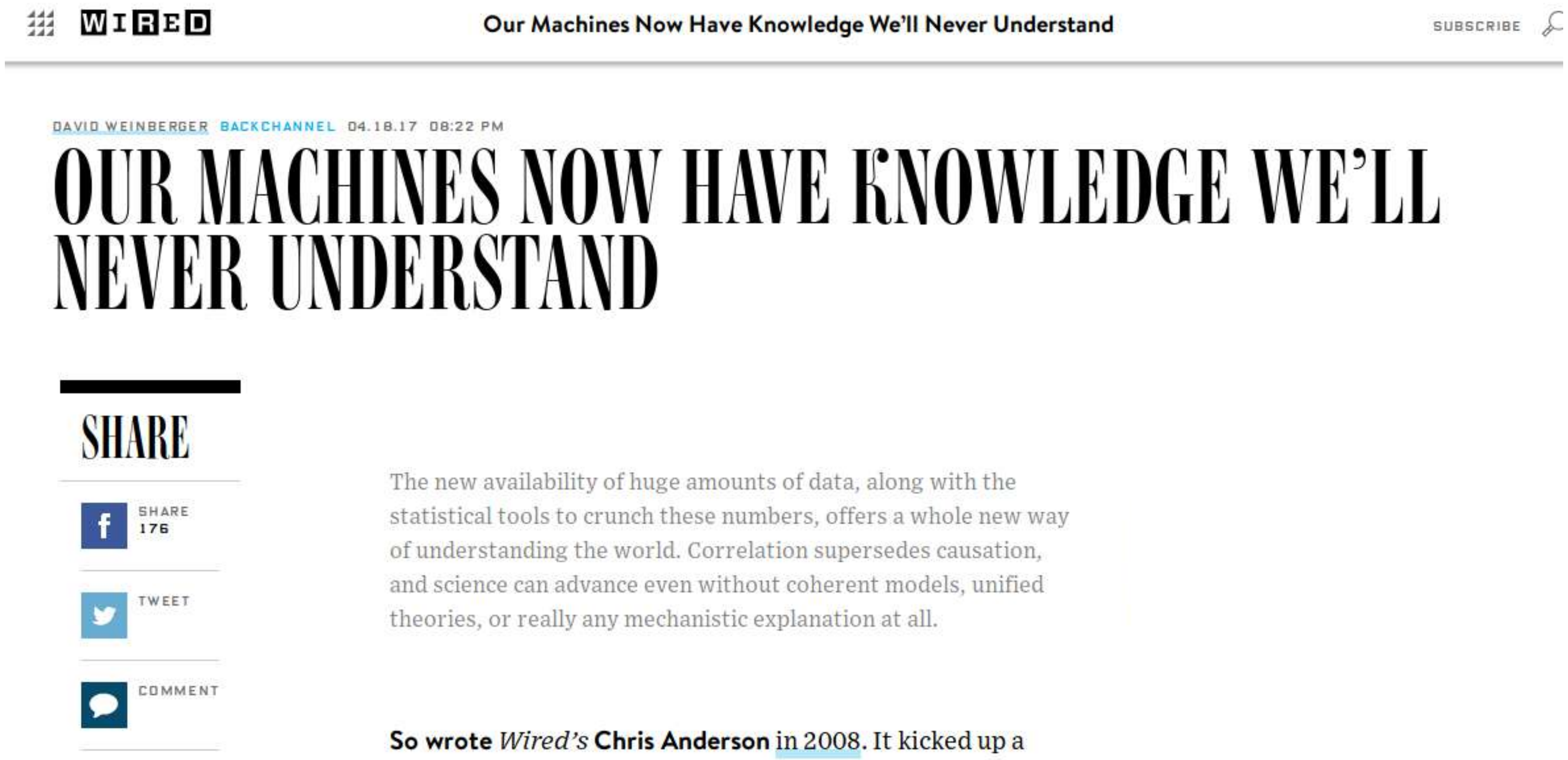
e main benefit of Shapley for importance scoring is that it uniquely satisfies the set of desired axioms: efficiency, mmetry, dummy, and additivity [12], which are listed in the following proposition with modification to our setting:

oposition 4.1. *Given a set of concepts $C_S = \{\mathbf{c}_1, \mathbf{c}_2, ...\mathbf{c}_m\}$ and a completeness score η, and some importance score for each concept $\mathbf{c}_i$ that depends on the completeness score η. $\mathbf{s}_i$ defined by conceptSHAP is the unique importance signment that satisfy the following four axioms:*

- *Efficiency: The sum of all importance value should sum up to the total completeness score, $\sum_{i=1}^{m} \mathbf{s}_i(\eta) = \eta(C_S)$.*
- *Symmetry: For* [illegible] *that are equiv*[illegible] *for every sub*[illegible]
- *Dummy: If $\eta(u \cup \{\mathbf{c}_i\}) = \eta(u)$ for every subset $u \subseteq C_S \setminus \{\mathbf{c}_i\}$, then $\mathbf{s}_i(\eta)$* [illegible] *0.*
- *Additivity: If η and η' have importance value $\mathbf{s}(\eta)$ and $\mathbf{s}(\eta')$ respectively, then the importance value of*

…and many more!

Is interpretability possible at all?



Our Machines Now Have Knowledge We'll Never Understand

DAVID WEINBERGER BACKCHANNEL 04.18.17 08:22 PM

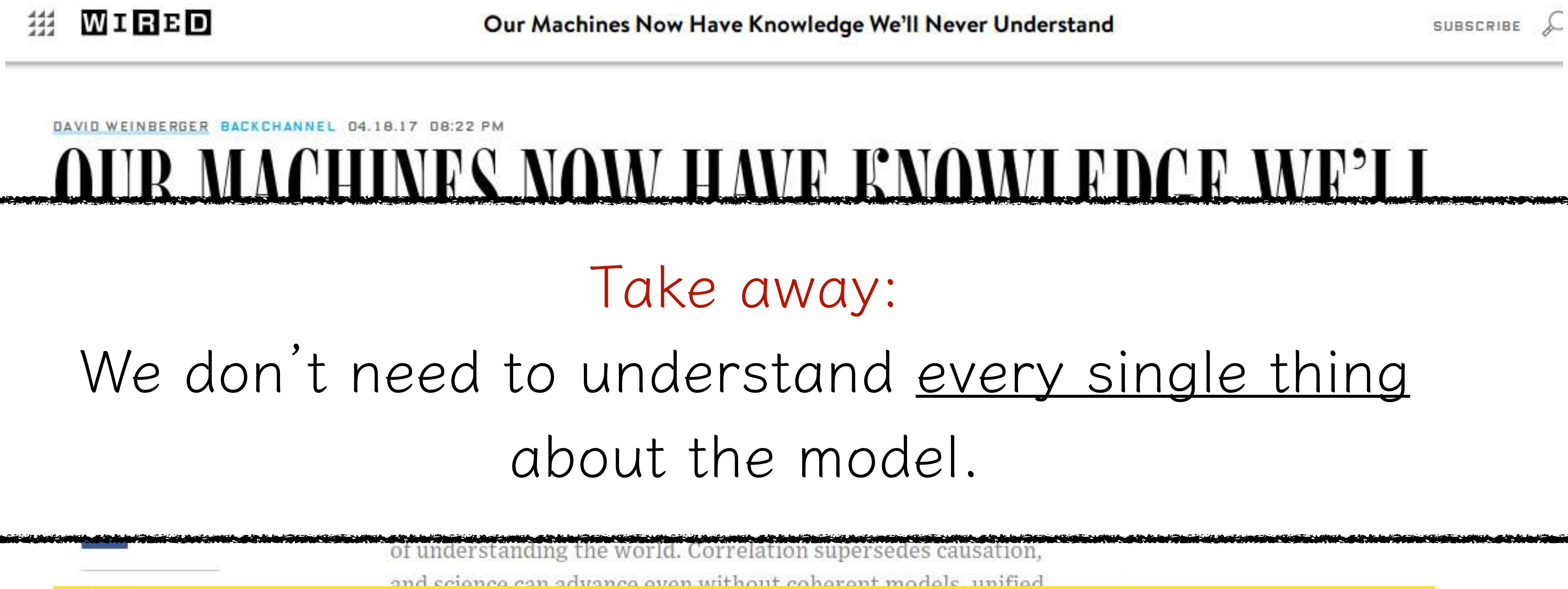
OUR MACHINES NOW HAVE KNOWLEDGE WE'LL NEVER UNDERSTAND

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?



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

of understanding the world. Correlation supersedes causation,

and science can advance even without coherent models, unified

Take away:

We don't need to understand <u>every single thing</u> about the model.

Key Point:

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

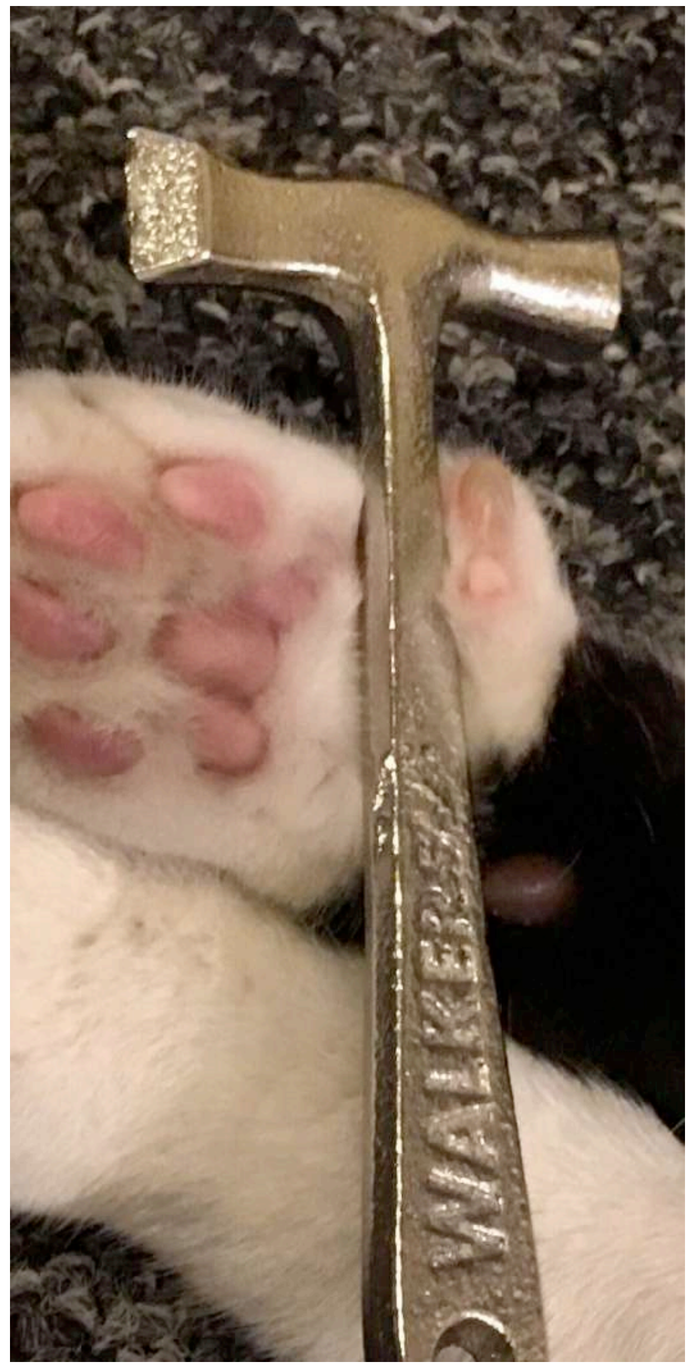
It is about knowing enough for your goals/downstream tasks.

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

How much is enough?

- What does it mean “the system is fair enough”?

- This hammer isn’t perfect, but it’s “good enough”



reddit.com

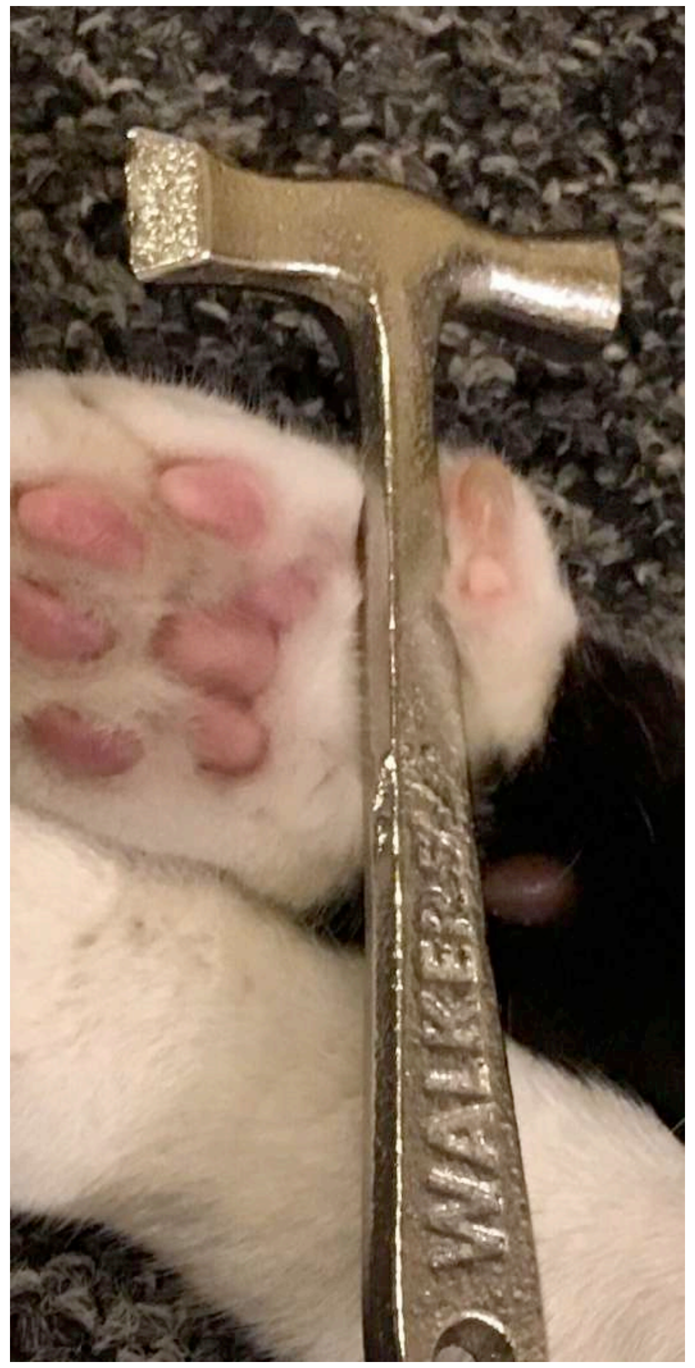
How much is enough?

- What does it mean “the system is fair enough”?

➡[for what we are trying to do]

- This hammer isn’t perfect, but it’s “good enough”

➡[for what we are trying to do]



reddit.com

How much is enough?

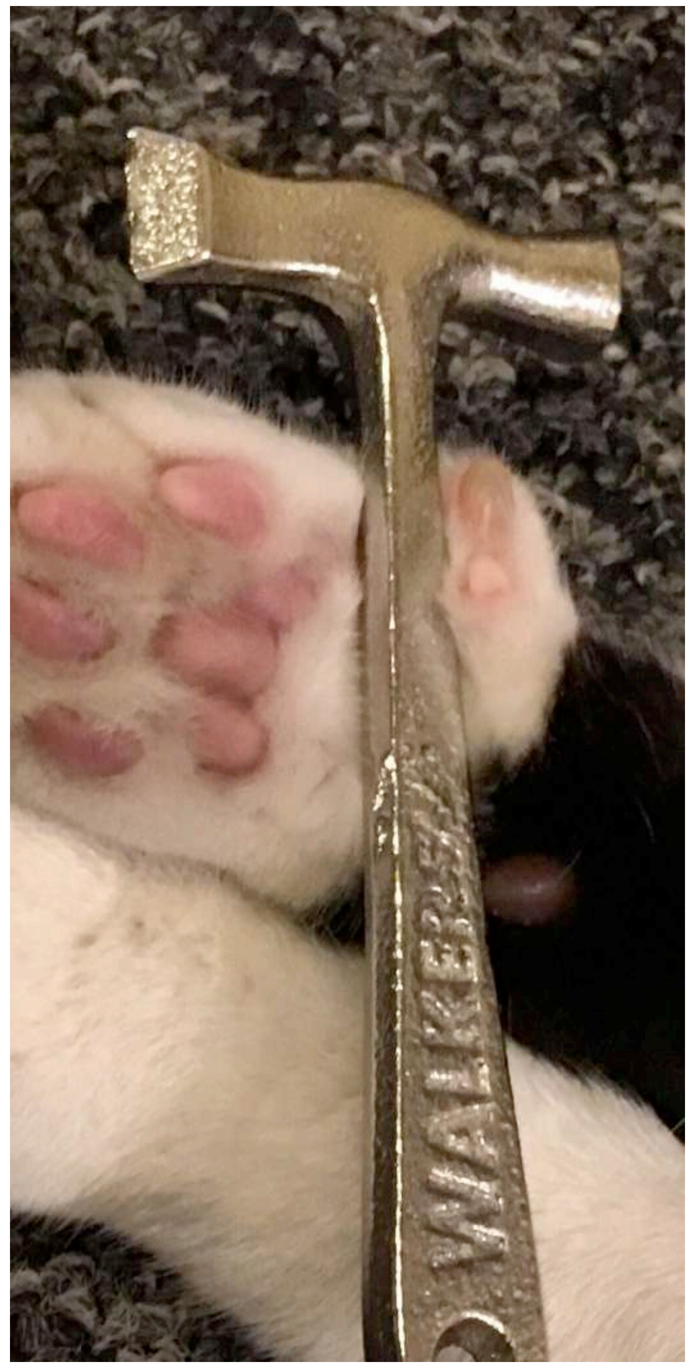
- What does it mean "the system is fair enough"?

➡[for what we are trying to do]

- This hammer isn't perfect, but it's "good enough"

➡[for what we are trying to do]

I'm better off having this tool
for [my goal].



reddit.com

What is the goal?

- End-task metric!
- Everyone's goals are different, but mine is generally:
  - Tools to help people use ML more effectively and responsibly such that

    1. our values are respected
    2. human knowledge is reflected when appropriate

What is the goal?

- End-task metric!

- Everyone's goals are different, but mine is generally:

  everyone

  - Tools to help people use ML more effectively and responsibly such that

    1. our values are respected

    2. human knowledge is reflected when appropriate

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

Interpretability is NOT…

- *about making ALL mode*
- *about understanding EV*
- *against developing highl*
- only *about gaining user*

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 ⤓

Take away:

Helping people to distrust the model is often more important than helping to trust it.

Interpretability is not a new problem. Why now?

- Prevalence: It's everywhere, and used to make potentially life changing decisions.
- Complexity: layers and layers of models of models

When do you need interpretability?

Fundamental underspecification in the problem

Humans often don't know *exactly* what they want.

When do you need interpretability?

example1: Safety

example 2: Science

Fundamental underspecification in the problem

example3: mismatched objectives

When do you need interpretability?

example1: Safety

example 2: Science

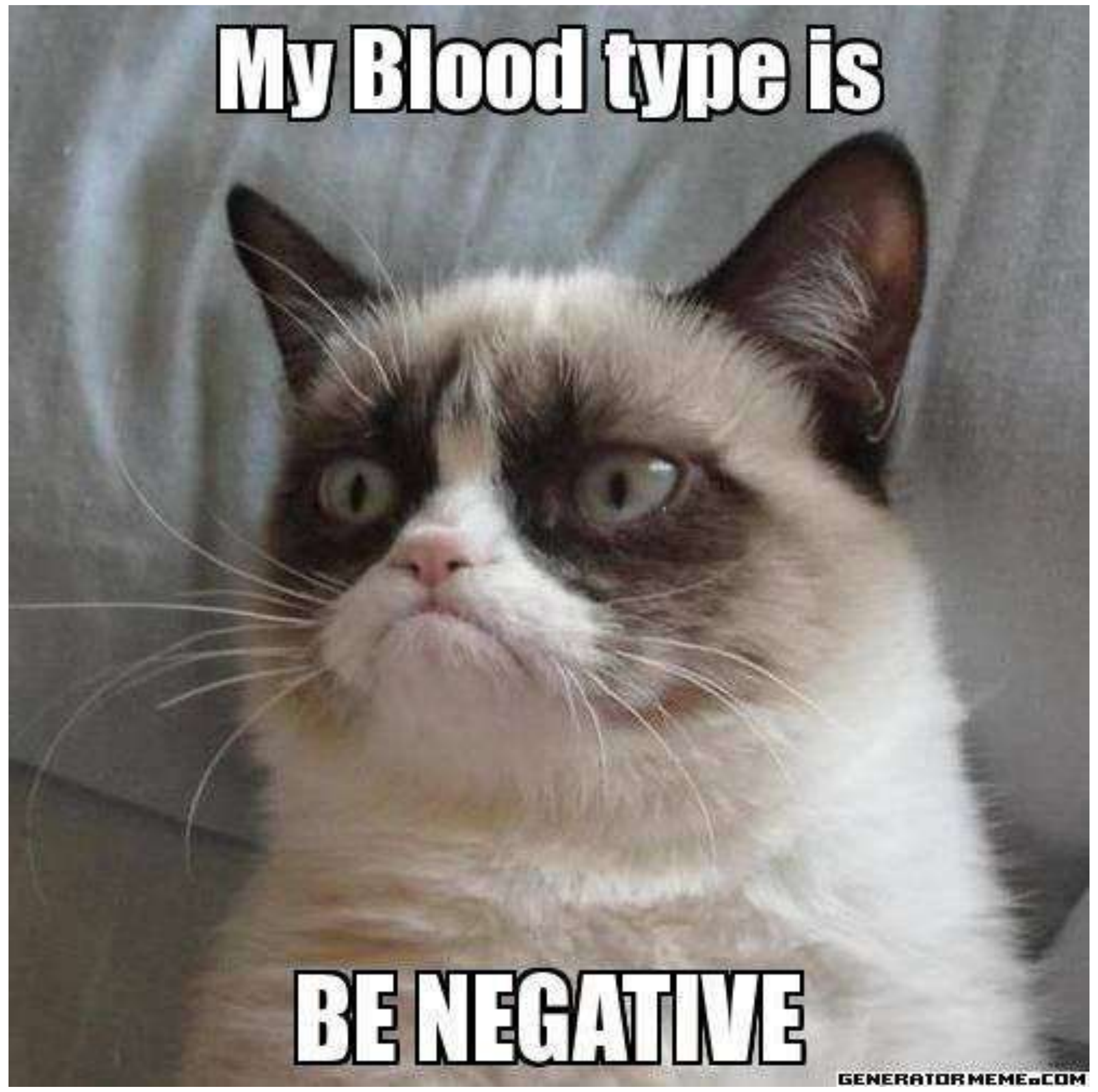
Fundamental underspecification in the problem

example3: mismatched objectives

Take away:

More data or more clever algorithm will not solve interpretability.

Wait, then what is NOT underspecification?



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

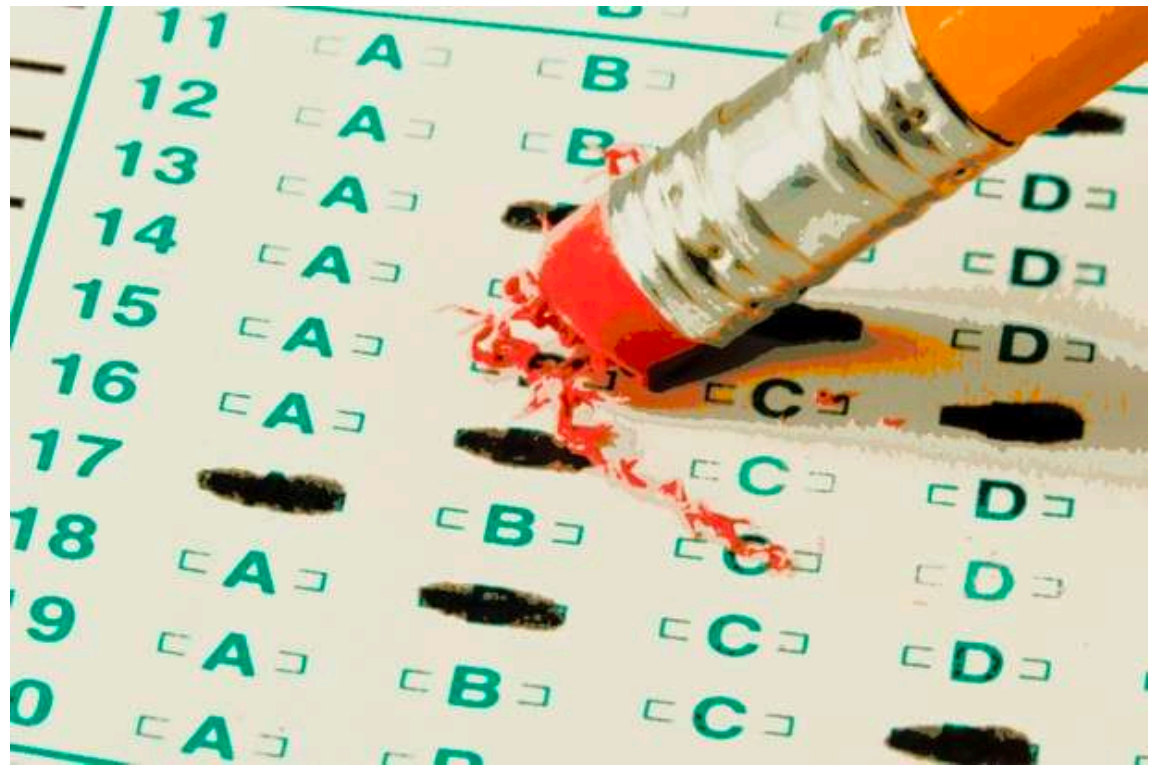
When we may not need/want interpretability

- No significant consequences. Prediction is what *everyone cares*.

- Sufficiently well-studied problem with *abundance of empirical evidence*

Instructs the optimal avoidance direction based on radio waves received from the other aircraft

- People might game the system (*example of* mismatched objectives)

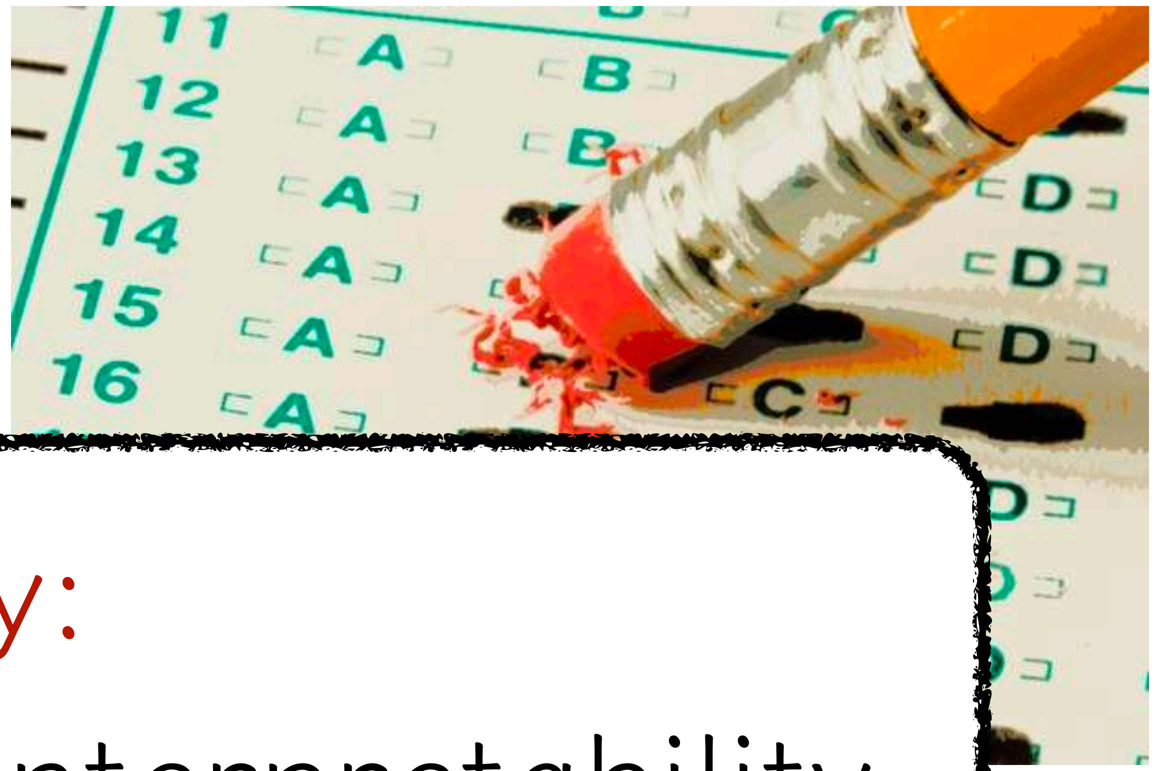


When we may not need/want interpretability

- No significant consequences. Prediction is what everyone cares.

- Sufficiently well-studied problem with abundance of empirical evidence

- People might game the system (example of mismatched objectives)



Take away:

We don't always need interpretability.

But certainly, there will be performance trade-off, right?

- "It is a myth that there is necessarily a trade-off between accuracy and interpretability." [Rudin 19]
- Carefully building structure in the model (e.g., architecture, prior, loss function) has long been done to increase performance with or without interpretability in mind.

Take away:

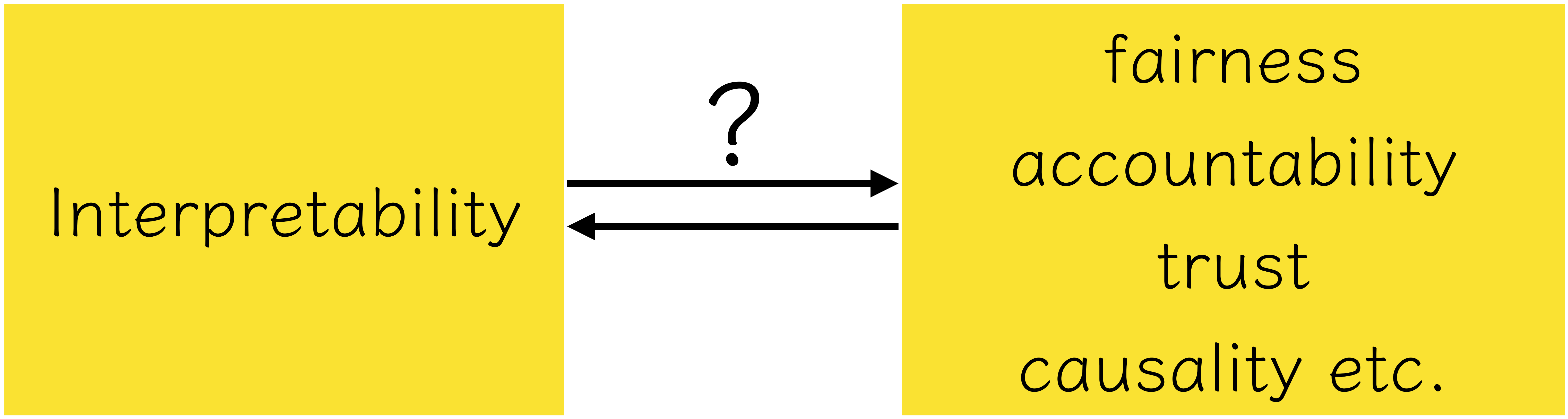
Interpretability and performance trade-off often don't exists.

True that.

Here are a small subset of vast amount of evidence by many researchers.

[1] Finale Doshi-Velez, Byron Wallace, and Ryan Adams. Graph-sparse lda: a topic model with structured sparsity. Association for the Advancement of Artificial Intelligence, 2015.
[2] Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander Van Esbroeck. Monotonic calibrated interpolated look-up tables. Journal of Machine Learning Research, 2016
[3] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1675–1684. ACM, 2016
[4] Been Kim, Julie Shah, and Finale Doshi-Velez. Mind the gap: A generative approach to interpretable feature selection and extraction. In Advances in Neural Information Processing Systems, 2015b
[5] Lou Y, Caruana R, Gehrke J, Hooker G. Accurate Intelligible Models with Pairwise Interactions. In: Proceedings of 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). ACM; 2013.
[6] Rudin C, Passonneau R, Radeva A, Dutta H, Ierome S, Isaac D. A Process for Predicting Manhole Events In Manhattan. Machine Learning. 2010;80:1–31.
[7] Rudin C, Ustun B. Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice. Interfaces. 2018;48:399–486. Special Issue: 2017 Daniel H. Wagner Prize for Excellence in Operations Research Practice September-October 2018.
[8] Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T. An Interpretable Model with Globally Consistent Explanations for Credit Risk. In: Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy; 2018.

What about our cousins?



Interpretability

?

fairness

accountability

trust

causality etc.

What about our cousins?

Interpretability

fairness
accountability
trust
causality etc.

Take away:

Trust, fairness and interpretability are not the same thing.

What about our cousins?

Interpretability

fairness
accountability
trust
causality etc.

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

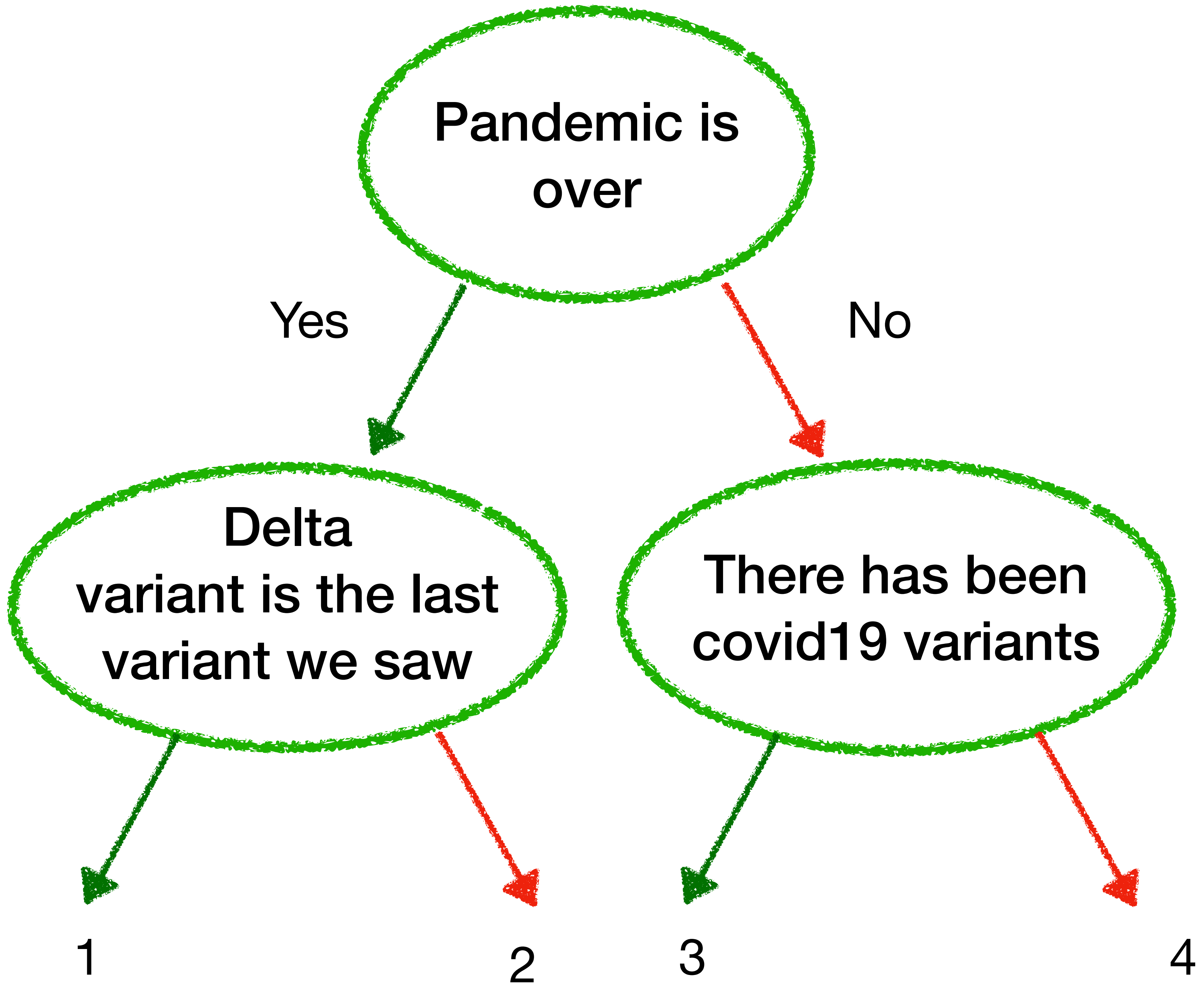
Agenda

- What and why
- !Caution!: Things to be careful when using and developing interpretability methods
- Evaluate: How to evaluate interpretability methods
- Methods: 3 types of methods and examples

I heard you can just use decision trees.

That's all we need, right?

Sample decision tree #1



Pandemic is over

Yes

No

Delta variant is the last variant we saw

There has been covid19 variants

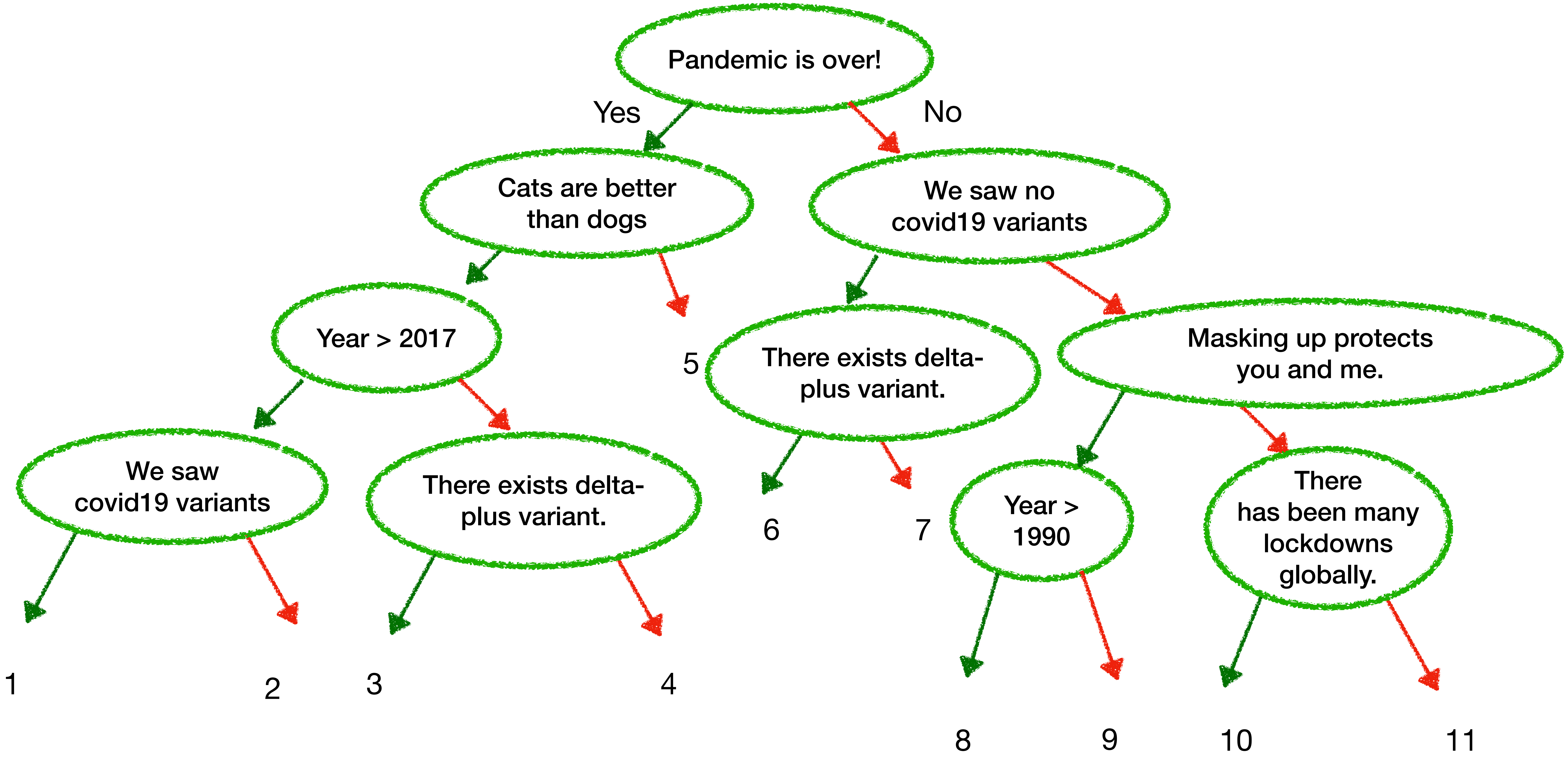
1

2

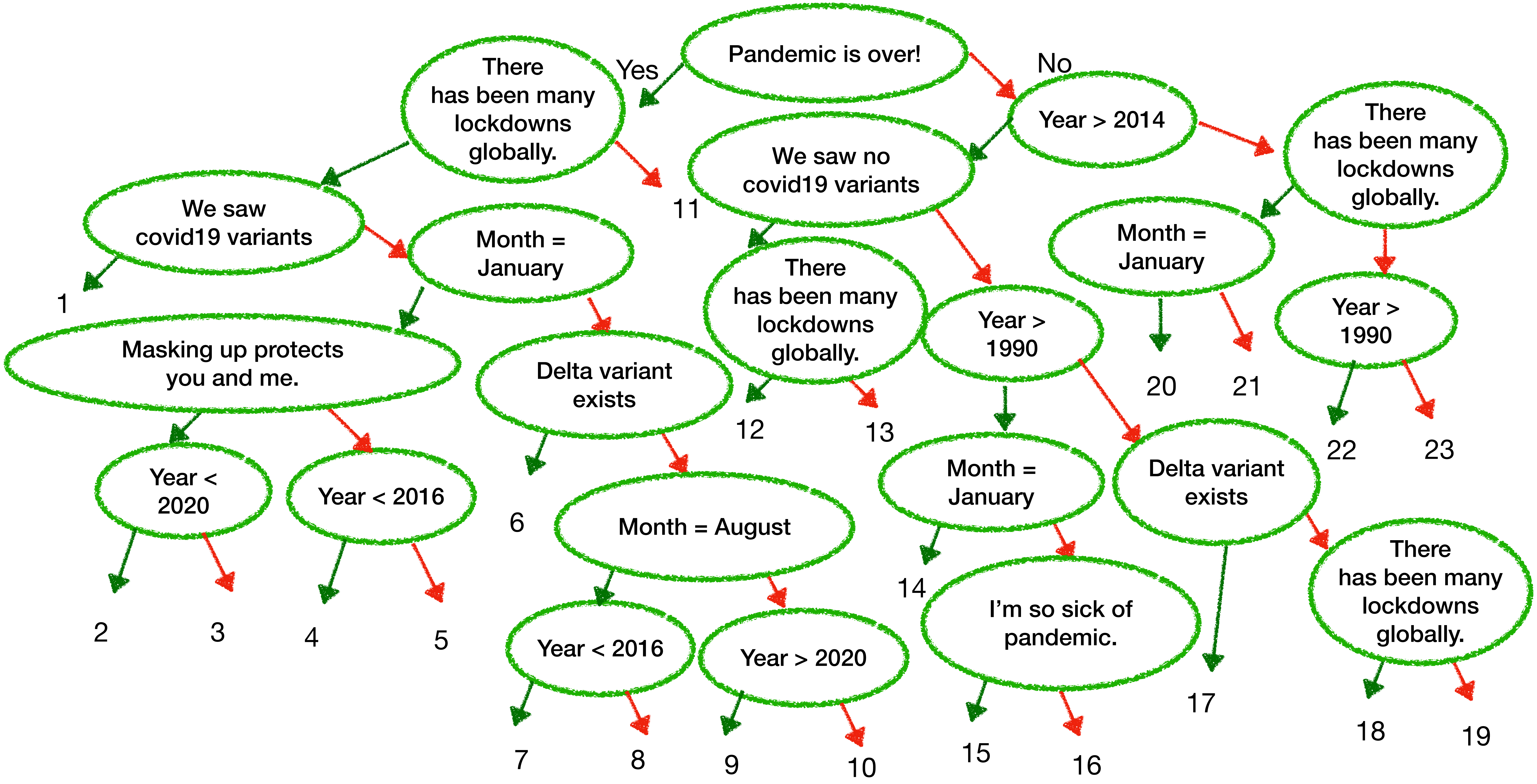
3

4

Sample decision tree #2



Sample decision tree #3



Pandemic is over!

Yes

No

There has been many lockdowns globally.

Year > 2014

There has been many lockdowns globally.

We saw no covid19 variants

We saw covid19 variants

Month = January

11

Month = January

There has been many lockdowns globally.

Year > 1990

1

Masking up protects you and me.

Delta variant exists

Year > 1990

20

21

12

13

22

23

Year < 2020

Year < 2016

Month = January

Delta variant exists

6

Month = August

14

I'm so sick of pandemic.

There has been many lockdowns globally.

2

3

4

5

Year < 2016

Year > 2020

17

18

19

7

8

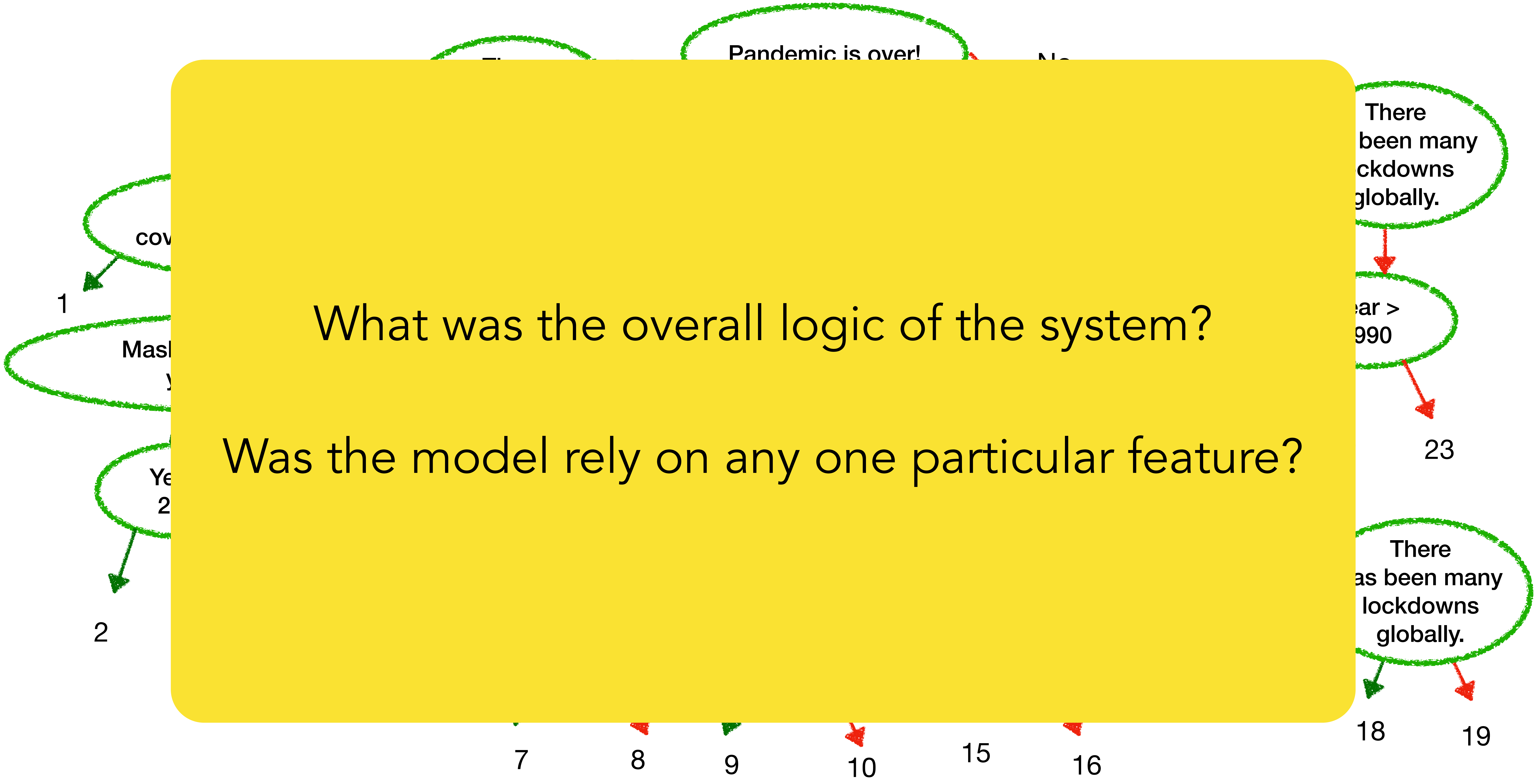
9

10

15

16

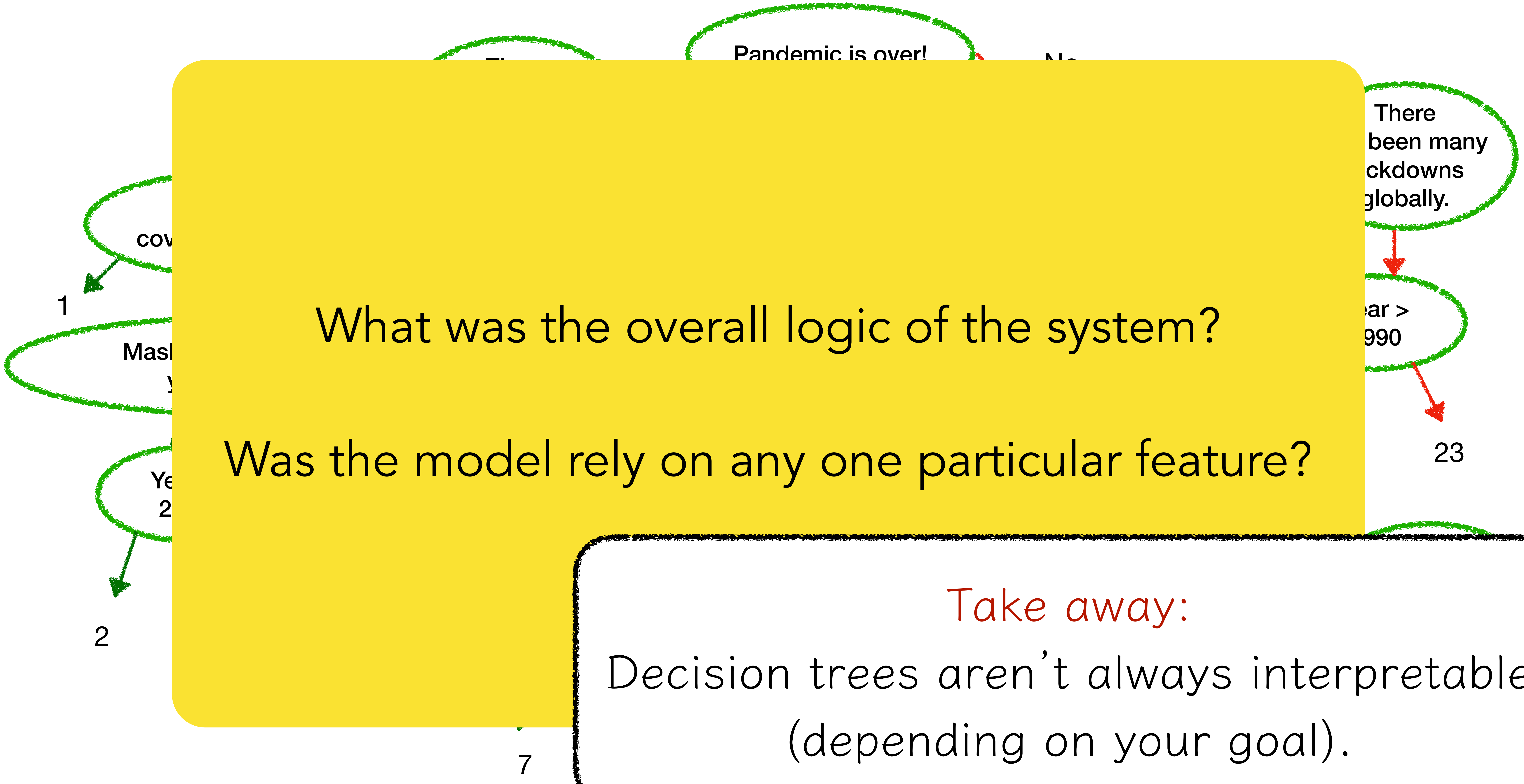
Sample decision tree #3



What was the overall logic of the system?

Was the model rely on any one particular feature?

Sample decision tree #3



What was the overall logic of the system?

Was the model rely on any one particular feature?

Take away:

Decision trees aren't always interpretable (depending on your goal).

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

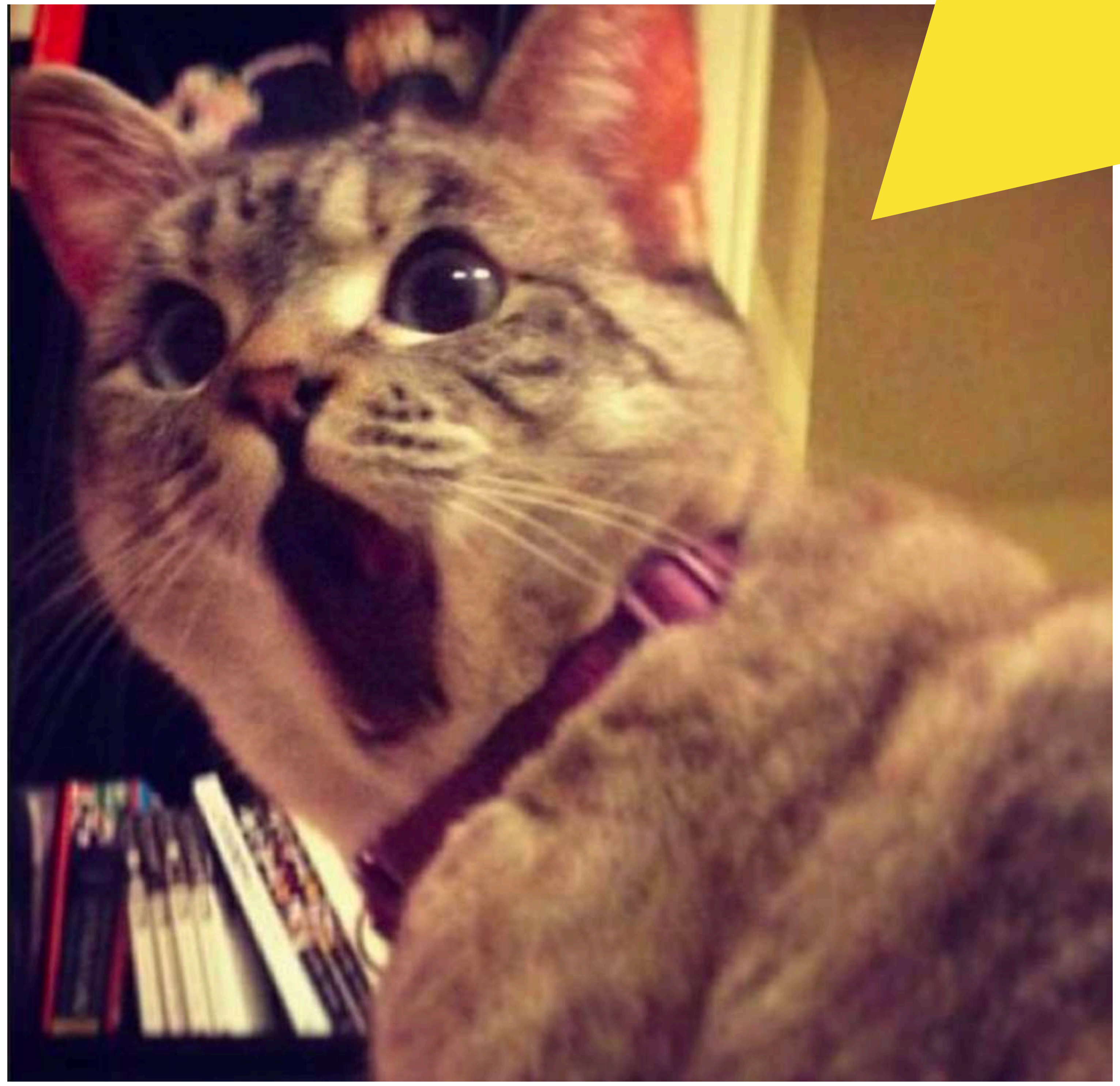
Maybe rule sets are better?

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?!



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

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

The point here is that there is <u>no one-size-fits-all</u> method.

Linear classifiers are interpretable, right?

So why not just fit locally linear functions everywhere?

Linear models are not always interpretable

- Can human interpret a linear model with many features, *each* with a floating number (normalized): *e.g.*, feature 1 weighted 0.1,⋯feature 134 weighted 0.05, feature 201 weighted 0.8..
  - "Probability distortion is that people generally do not look at the value of probability uniformly between 0 and 1. Lower probability is said to be over-weighted while medium to high probability is under-weighted" - Kahneman

Linear models are not always interpretable

- Can human interpret a linear model with many features, *each* with a floating number (normalized): *e.g.*, feature 1 weighted 0.1,⋯feature 134 weighted 0.05, feature 201 weighted 0.8..
  - "Probability distortion is that people generally do not look at the value of probability uniformly between 0 and 1. Lower probability is said to be over-weighted while medium to high probability is under-weighted" - Kahneman

Take away:
Using linear model isn't always the answer.

Causality should be the one and only methods for interpretability, right?

Pursuing causality is great, but it's not always simple

- It is one of the areas of huge importance, no doubt about that!
- But (currently) it often comes with a lot of assumptions (e.g., no hidden confounders) that starts to matter for high dimensional real-world applications.
- "Without causality, explanation is meaningless" -> I'd rather have useful, well-validated explanation than nothing at all for high stake applications.

A very intuitive (and funny) tutorial on causal inference!

https://matheusfacure.github.io/

Causal Inference for the Brave and True

Search this book...

Causal Inference for The Brave and True

PART I - THE YANG

01 - Introduction To Causality

02 - Randomised Experiments

03 - Stats Review: The Most Dangerous Equation

04 - Graphical Causal Models

05 - The Unreasonable Effectiveness of Linear Regression

06 - Grouped and Dummy Regression

07 - Beyond Confounders

08 - Instrumental Variables

Causal Inference for The Brave and True

A light-hearted yet rigorous approach to learning impact estimation and sensitivity analysis. Everything in Python and with

So once we have an explanation, that IS how the model thinks, right?

Some explanation methods fails a simple sanity check.

Input image

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

prediction

$p(z)$

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]

Some explanation methods fails a simple sanity check.

Input image

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

prediction

$p(z)$

Junco Bird-ness

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

Why was this a Junco bird?

One definition of explanation:

Tell me how **sensitive** the prediction is when we slightly **change** each input feature (pixel).

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

Input image

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

prediction

$p(z)$

Junco Bird-ness

In jargon: take derivative of the prediction wrt each pixel.

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

In English: take one pixel in the image, and imagine changing it by a little. See how much prediction changes. Do this for all pixels.

One definition of explanation:

Tell me how **sensitive** the prediction is when we slightly **change** each input feature (pixel).

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

One of the most popular interpretability methods for images:

Saliency maps

Input image

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

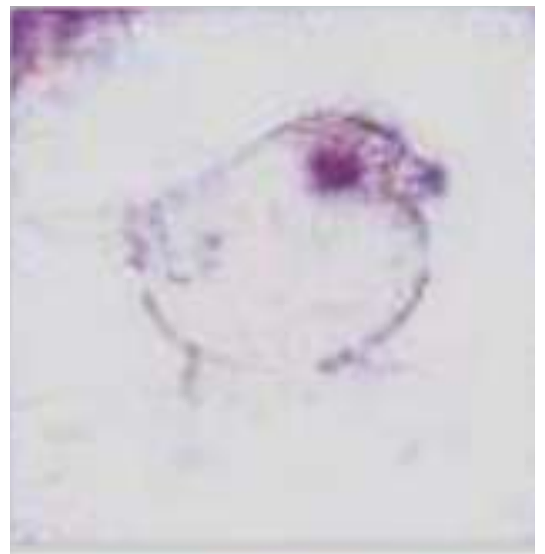
prediction

$p(z)$

Junco Bird-ness

Popular method #1

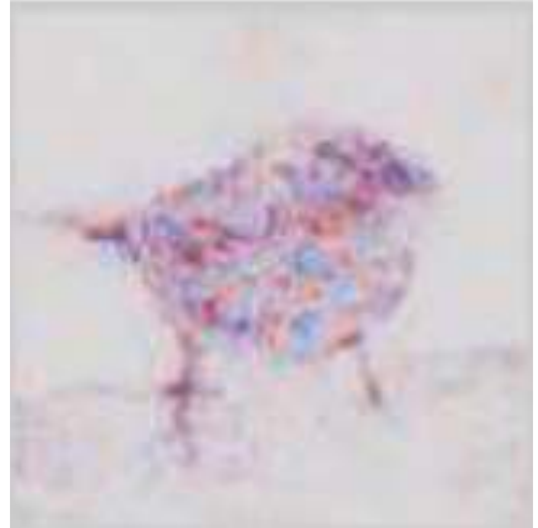
Popular method #2



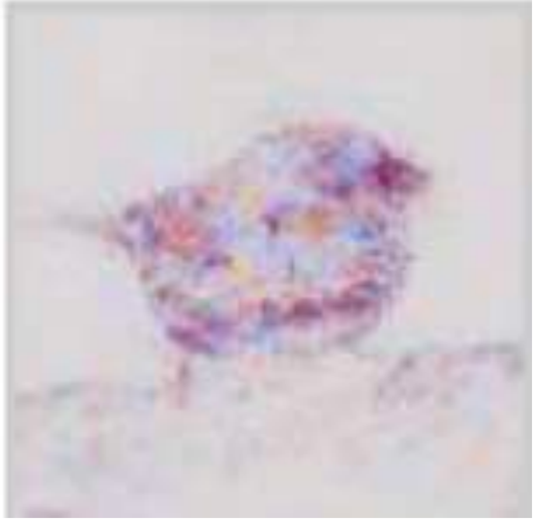
My work from 2018 #1

My work from 2018 #2

Popular method #3



Popular method #4



A sanity check question:

Input image

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

prediction

$p(z)$

Junco Bird-ness

So these pixels are the **evidence** of **prediction.**

g(prediction) = explanation

When **prediction** changes, the explanations will probably change.

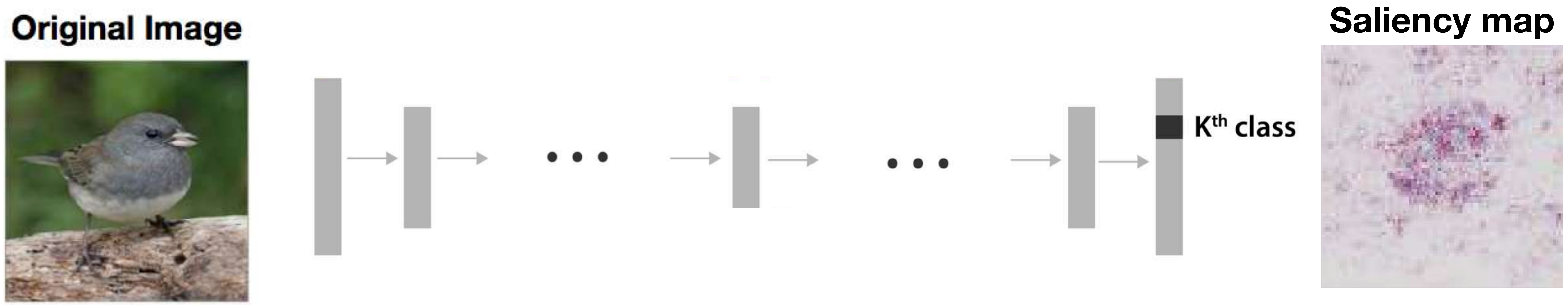
g(prediction') = explanation'

When **prediction** is random, the explanations really should change!

g(random) != explanation ?

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

A sanity check results



Randomized weights!
Network now makes garbage prediction.

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

A sanity check results

Input image

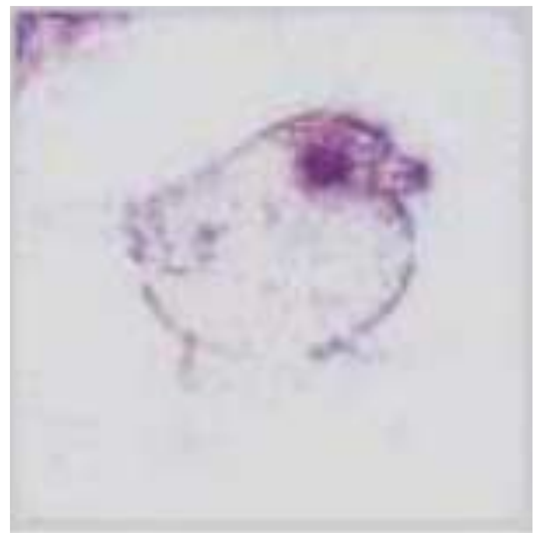
A trained
machine learning model
(e.g., neural network)

prediction

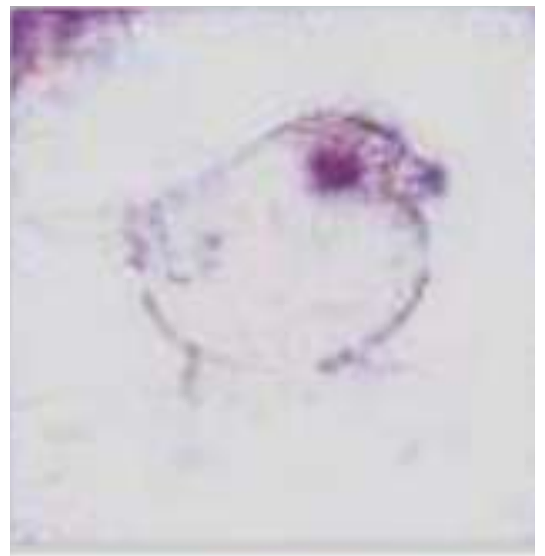
$$p(z)$$

Junco Bird-ness

Popular method #1



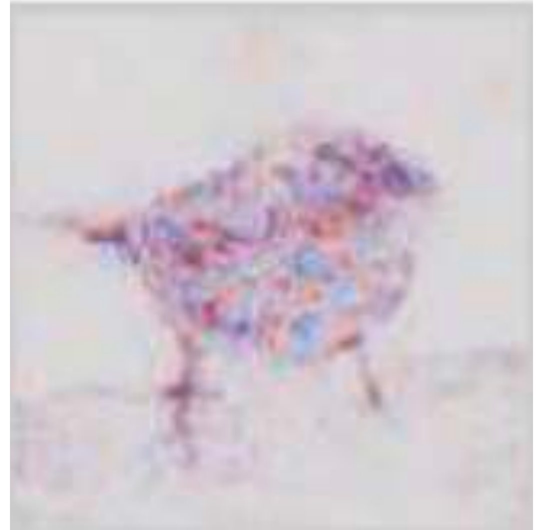
Popular method #2



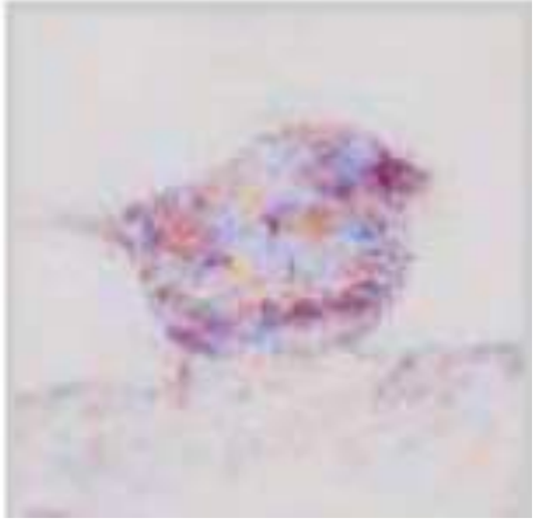
My work from 2018 #1

My work from 2018 #2

Popular method #3

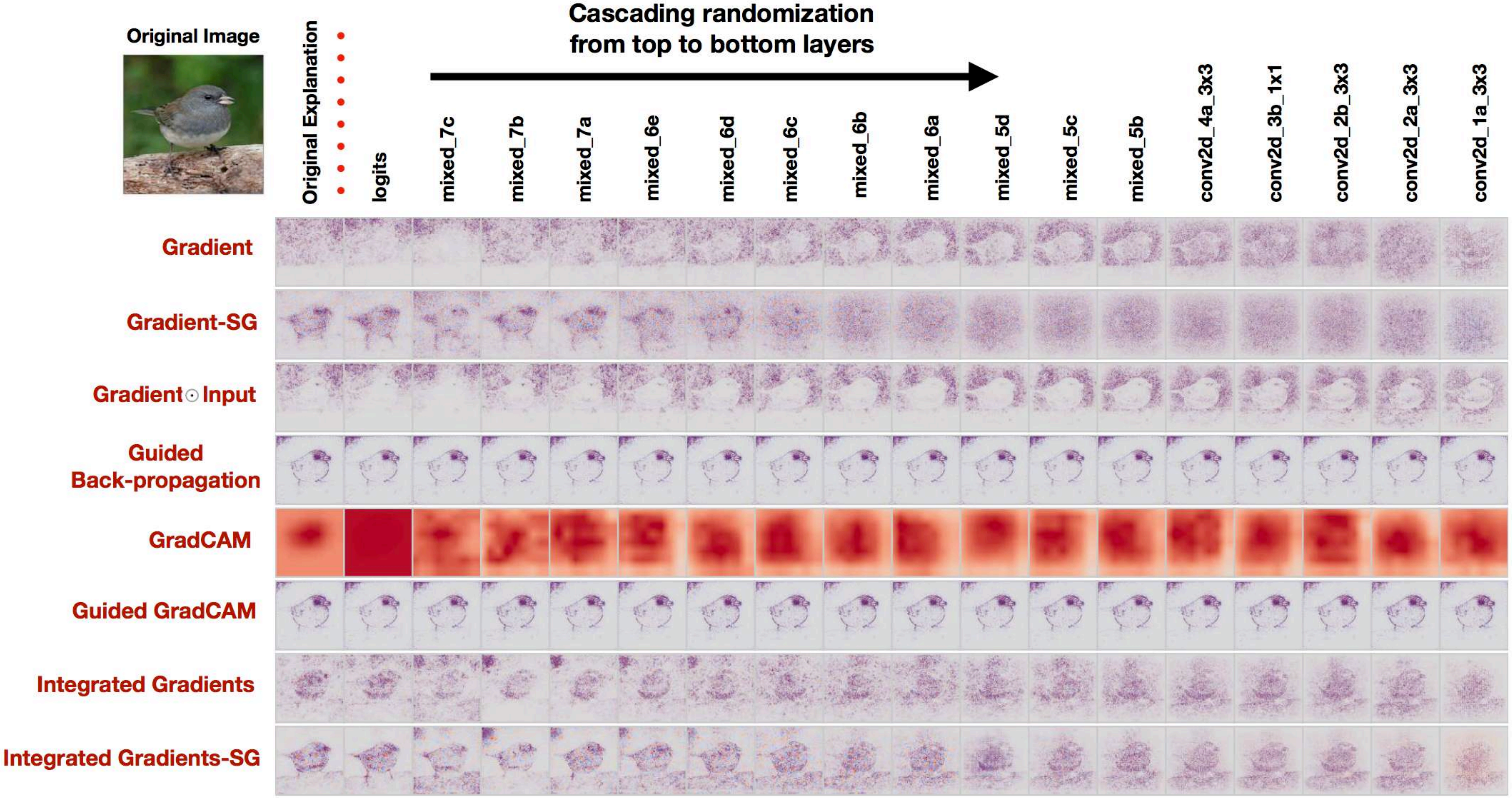


Popular method #4



Sanity check 1: Explanations from random vs trained network

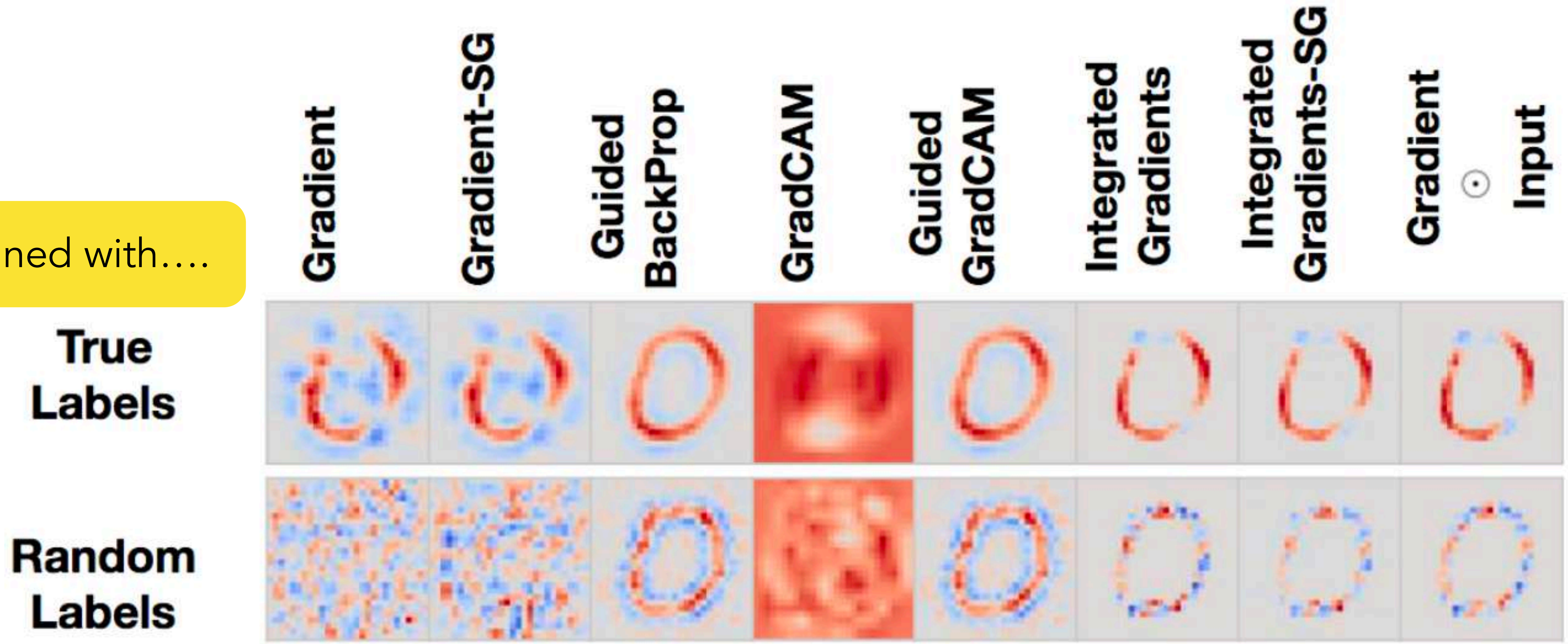
Most of methods produce quantitatively and qualitative similar results



Sanity check 2: Network trained with random vs true labels

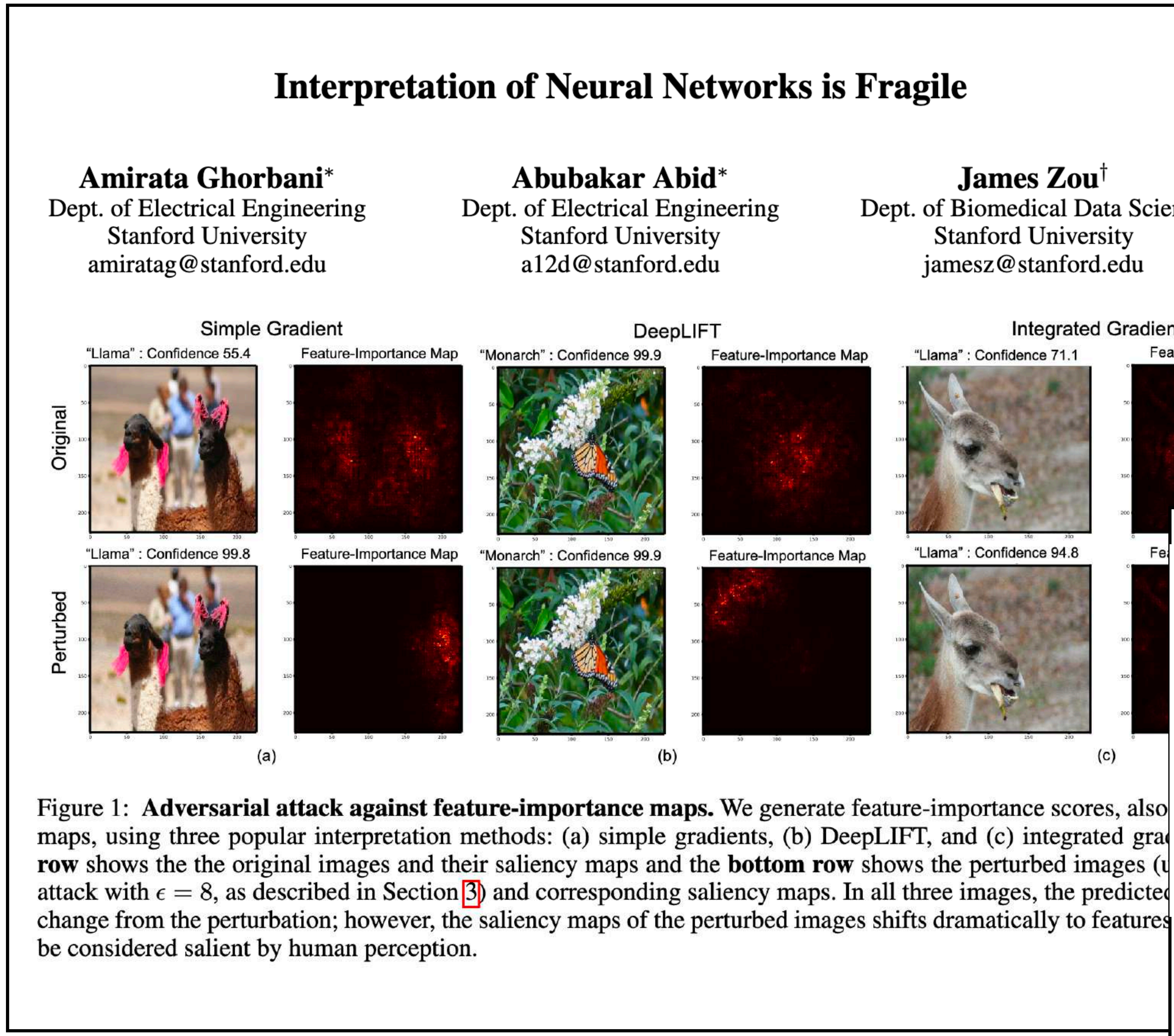
Most of methods produce quantitatively and qualitative similar results

Networks trained with….



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

Explanations can be (easily) attacked!



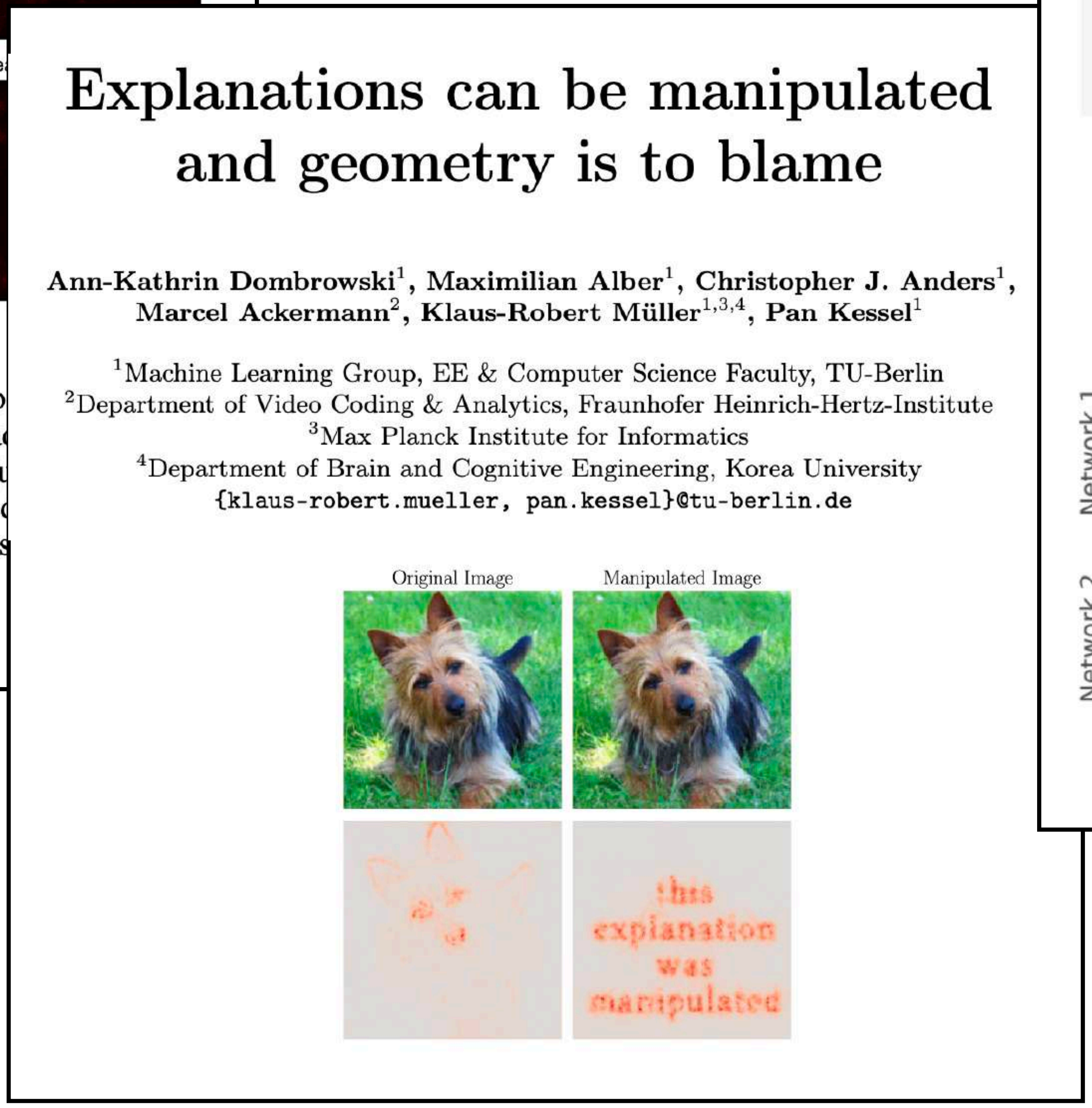
Interpretation of Neural Networks is Fragile

Amirata Ghorbani*
Dept. of Electrical Engineering
Stanford University
amiratag@stanford.edu

Abubakar Abid*
Dept. of Electrical Engineering
Stanford University
a12d@stanford.edu

James Zou†
Dept. of Biomedical Data Science
Stanford University
jamesz@stanford.edu

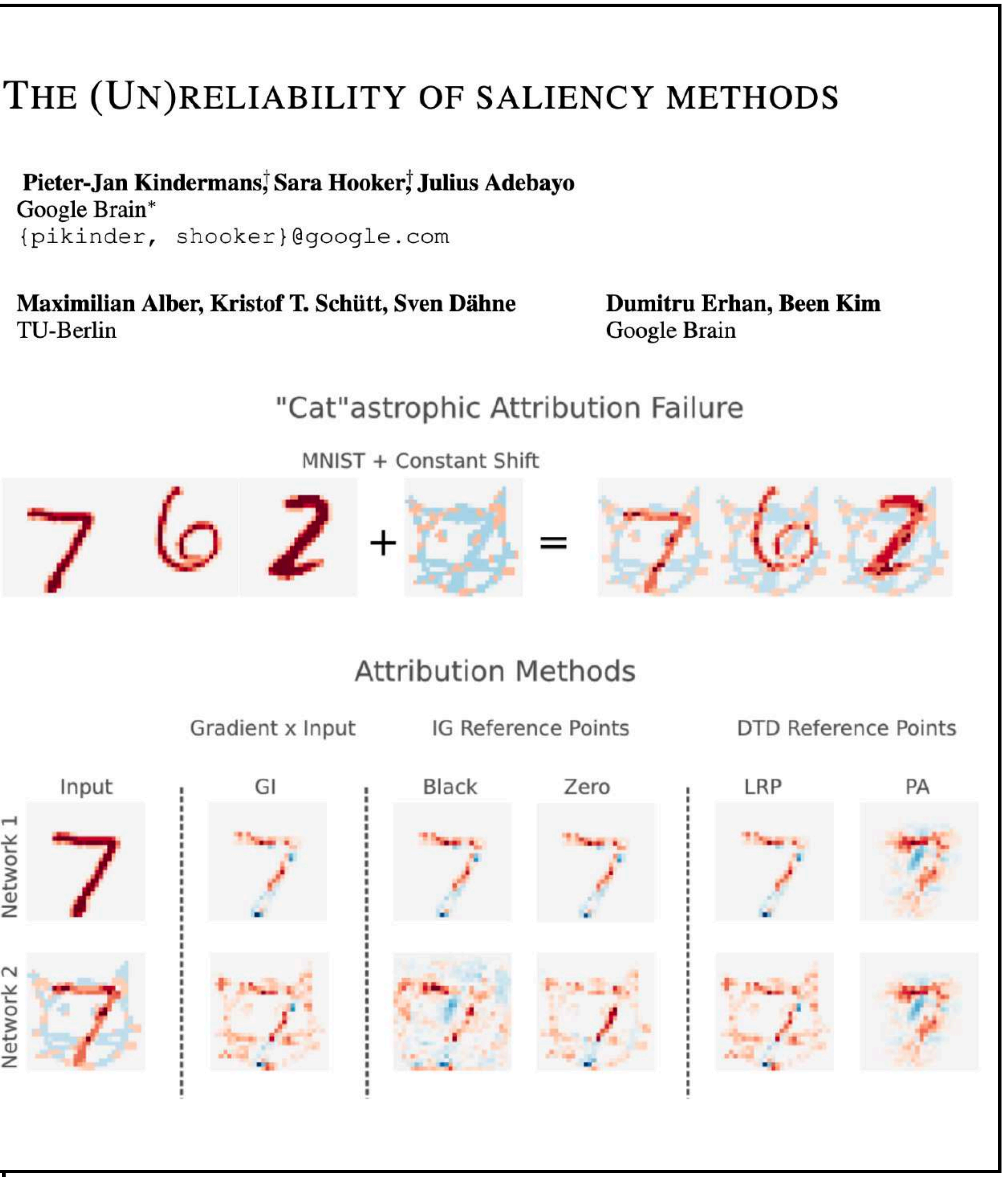
Figure 1: **Adversarial attack against feature-importance maps.** We generate feature-importance scores, also
maps, using three popular interpretation methods: (a) simple gradients, (b) DeepLIFT, and (c) integrated gra
row shows the the original images and their saliency maps and the **bottom row** shows the perturbed images (u
attack with $\epsilon = 8$, as described in Section 3) and corresponding saliency maps. In all three images, the predicte
change from the perturbation; however, the saliency maps of the perturbed images shifts dramatically to features
be considered salient by human perception.



Explanations can be manipulated and geometry is to blame

Ann-Kathrin Dombrowski[1], **Maximilian Alber**[1], **Christopher J. Anders**[1],
Marcel Ackermann[2], **Klaus-Robert Müller**[1,3,4], **Pan Kessel**[1]

[1]Machine Learning Group, EE & Computer Science Faculty, TU-Berlin
[2]Department of Video Coding & Analytics, Fraunhofer Heinrich-Hertz-Institute
[3]Max Planck Institute for Informatics
[4]Department of Brain and Cognitive Engineering, Korea University
{klaus-robert.mueller, pan.kessel}@tu-berlin.de



THE (UN)RELIABILITY OF SALIENCY METHODS

Pieter-Jan Kindermans‡, **Sara Hooker**‡, **Julius Adebayo**
Google Brain*
{pikinder, shooker}@google.com

Maximilian Alber, Kristof T. Schütt, Sven Dähne
TU-Berlin

Dumitru Erhan, Been Kim
Google Brain

Approximated explanations can be and will be wrong sometimes

- A number of methods "approximates" models behavior in some way. This means, there will be errors.
- Sometimes it's just plain wrong (e.g., not robust to distributional shifts)

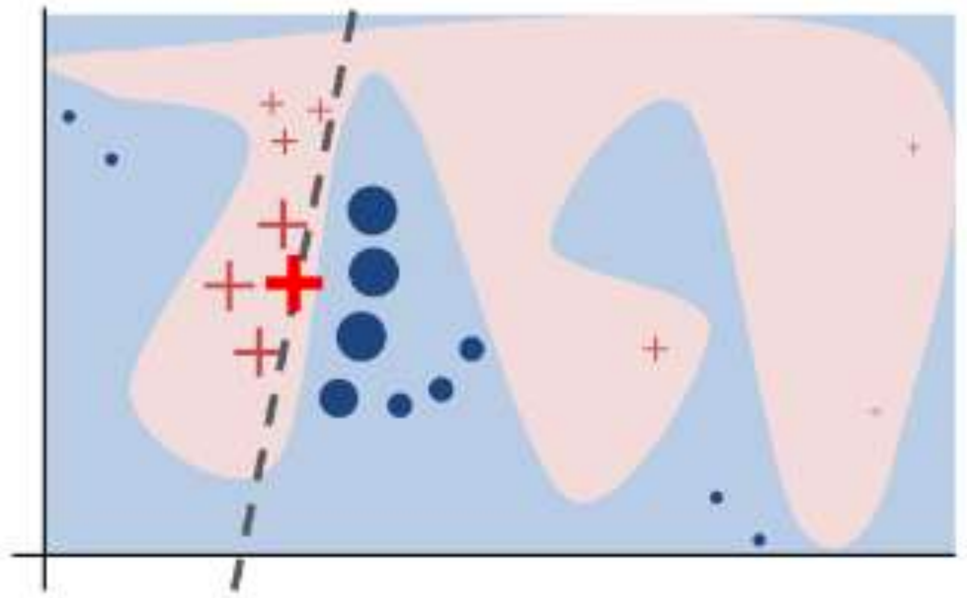
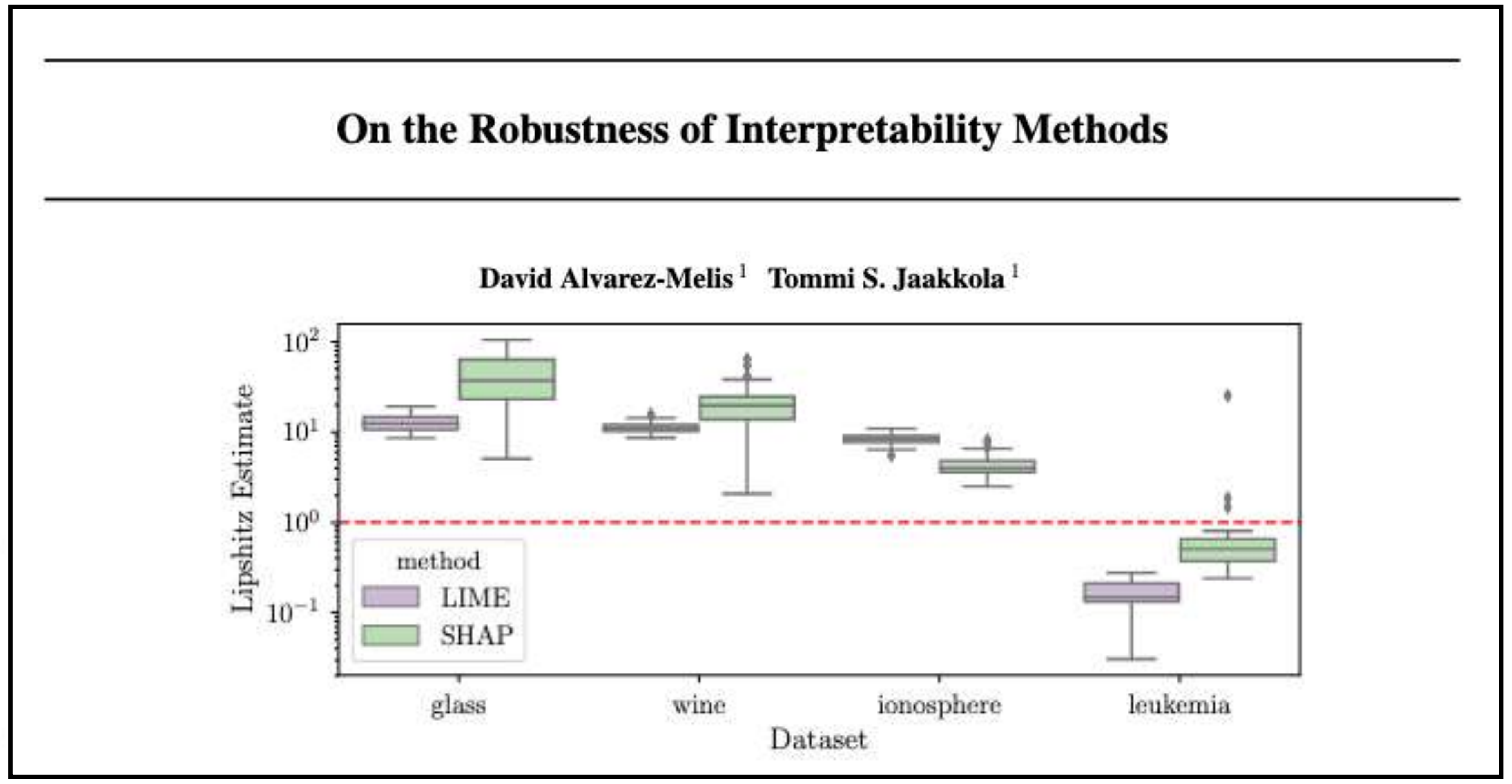
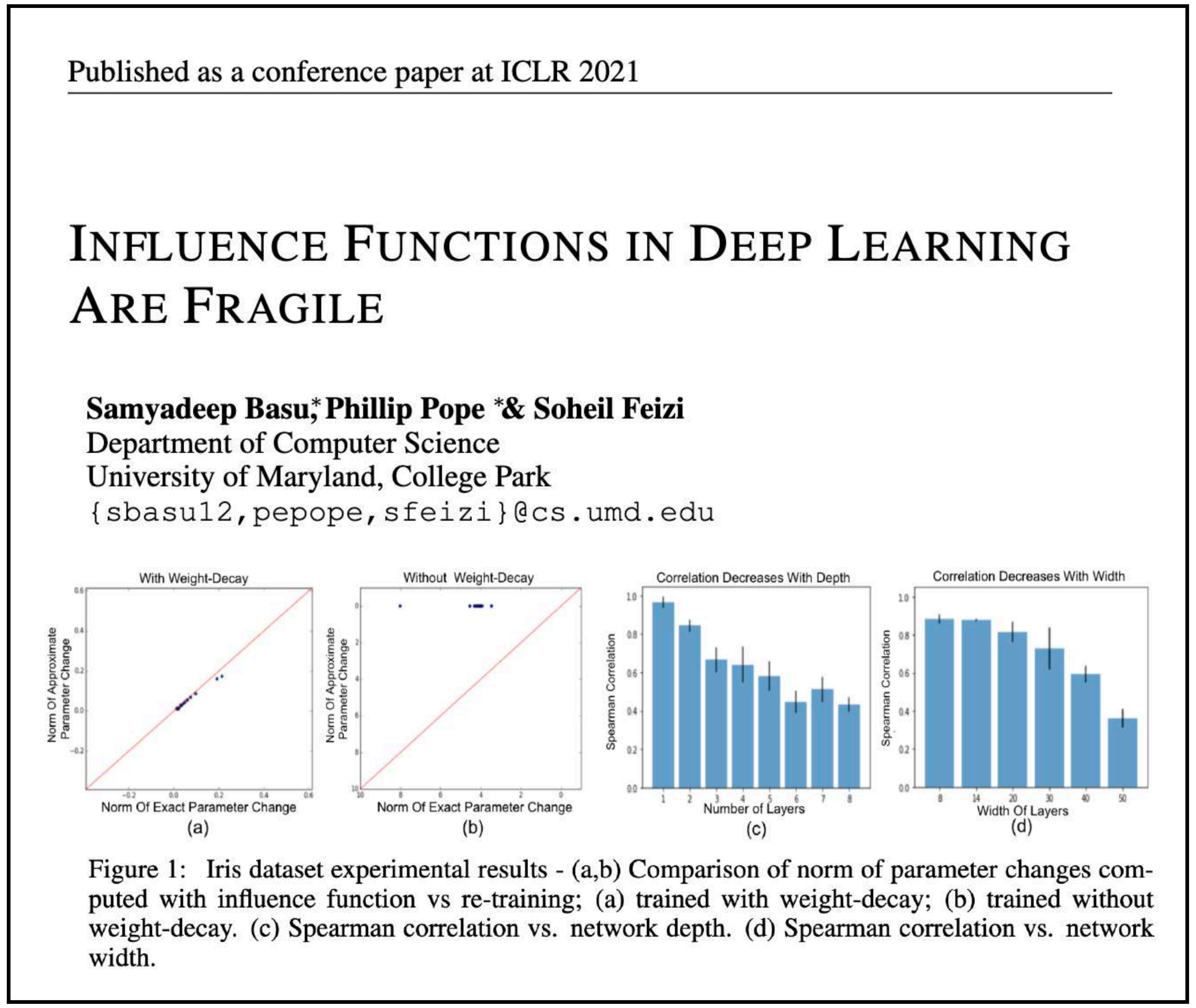


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.



On the Robustness of Interpretability Methods

David Alvarez-Melis [1] **Tommi S. Jaakkola** [1]



Published as a conference paper at ICLR 2021

INFLUENCE FUNCTIONS IN DEEP LEARNING ARE FRAGILE

Samyadeep Basu*, Phillip Pope *& Soheil Feizi
Department of Computer Science
University of Maryland, College Park
`{sbasu12,pepope,sfeizi}@cs.umd.edu`

Figure 1: Iris dataset experimental results - (a,b) Comparison of norm of parameter changes computed with influence function vs re-training; (a) trained with weight-decay; (b) trained without weight-decay. (c) Spearman correlation vs. network depth. (d) Spearman correlation vs. network width.

Approximated explanations can be and will be wrong sometimes

- A number of methods "approximates" models behavior in some way. This means, there will be errors.
- Sometimes it's just plain wrong (e.g., not robust to distributional shifts)

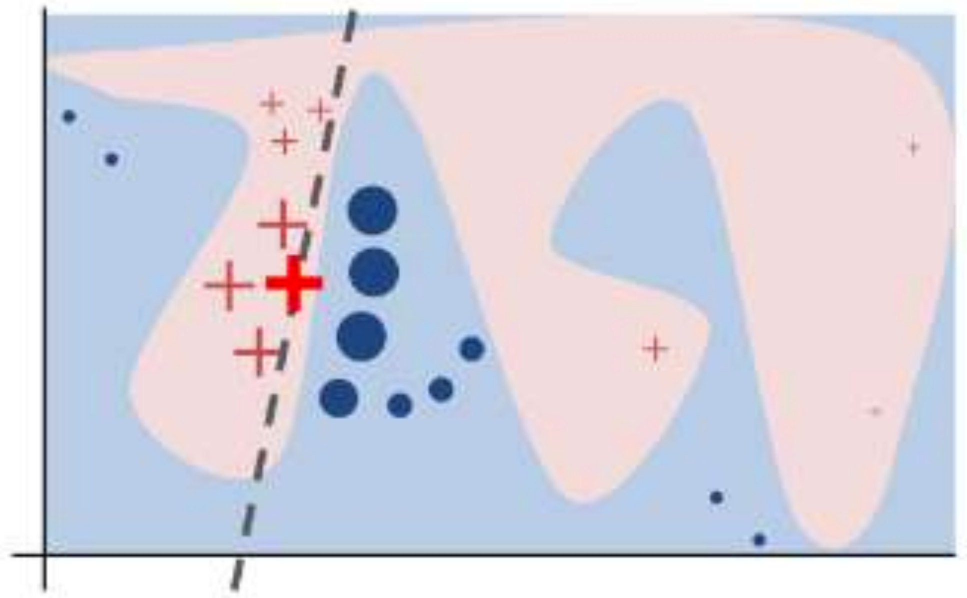
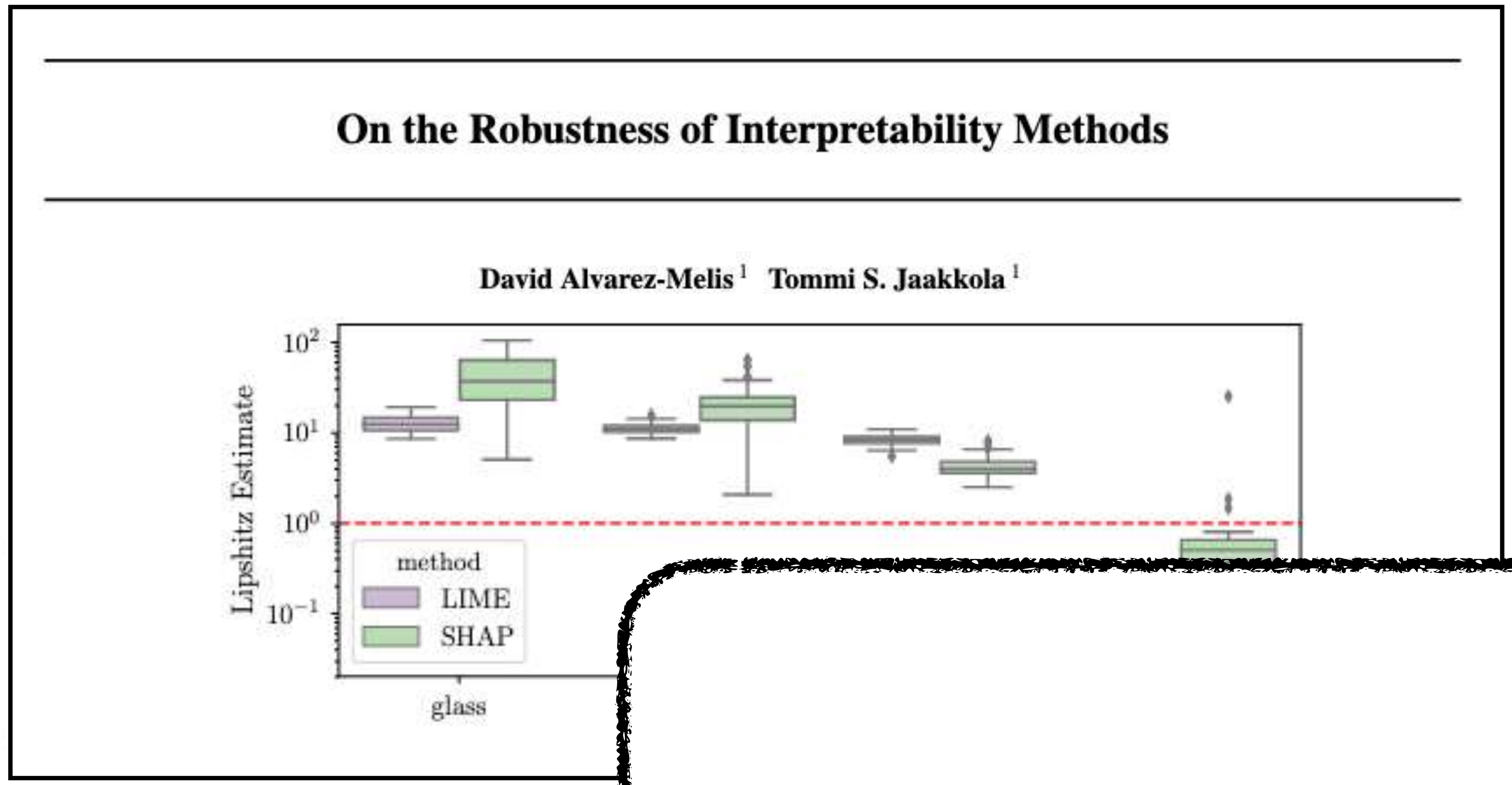
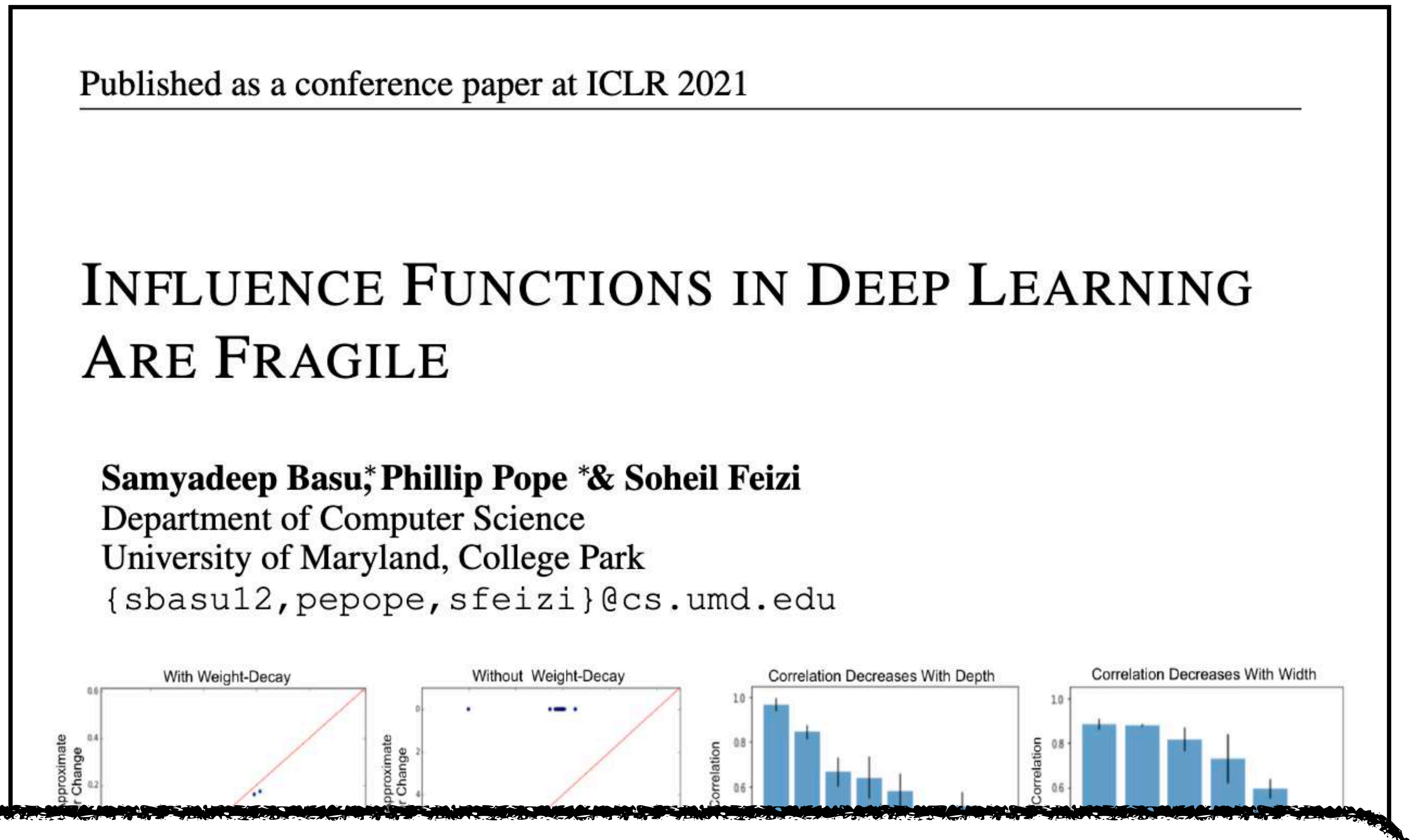


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.



On the Robustness of Interpretability Methods

David Alvarez-Melis [1] Tommi S. Jaakkola [1]



Published as a conference paper at ICLR 2021

INFLUENCE FUNCTIONS IN DEEP LEARNING ARE FRAGILE

Samyadeep Basu*, Phillip Pope *& Soheil Feizi
Department of Computer Science
University of Maryland, College Park
`{sbasu12,pepope,sfeizi}@cs.umd.edu`

Take away:

Always skeptical about the explanations you get.

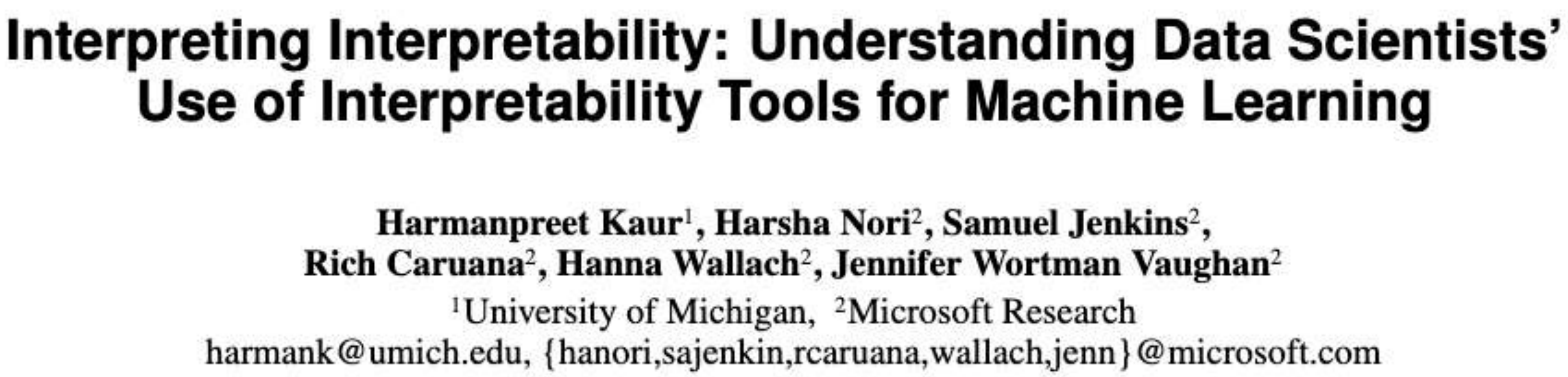
We already know all the learned parameters of the function of the neural network. This is an open book and transparent system.

[This has actually been said]

Oh, come on.

I’m an ML person.
Human experiments are only for HCI folks, right?

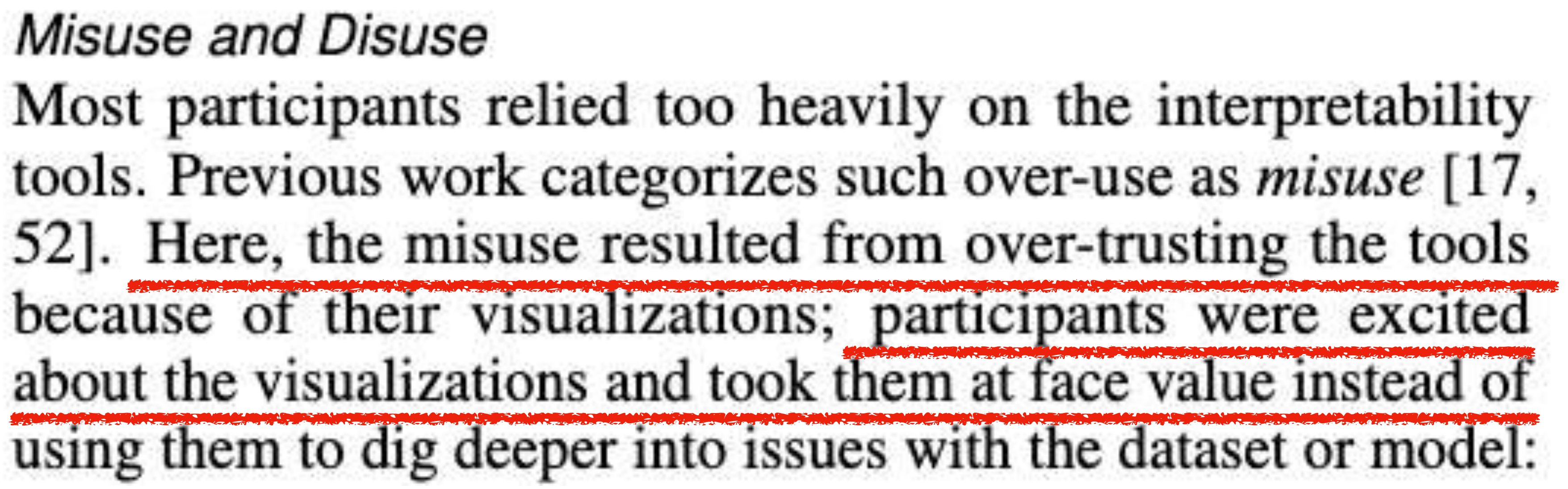
"How" *explanations are presented* is *as* important *as* the *explanations* themselves. Knowing how that impacts users is *even* more important



Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur[1], **Harsha Nori**[2], **Samuel Jenkins**[2],
Rich Caruana[2], **Hanna Wallach**[2], **Jennifer Wortman Vaughan**[2]
[1]University of Michigan, [2]Microsoft Research
harmank@umich.edu, {hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

"Our results indicate that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualizations output by these tools."



Misuse and Disuse

Most participants relied too heavily on the interpretability tools. Previous work categorizes such over-use as *misuse* [17, 52]. Here, the misuse resulted from over-trusting the tools because of their visualizations; participants were excited about the visualizations and took them at face value instead of using them to dig deeper into issues with the dataset or model:

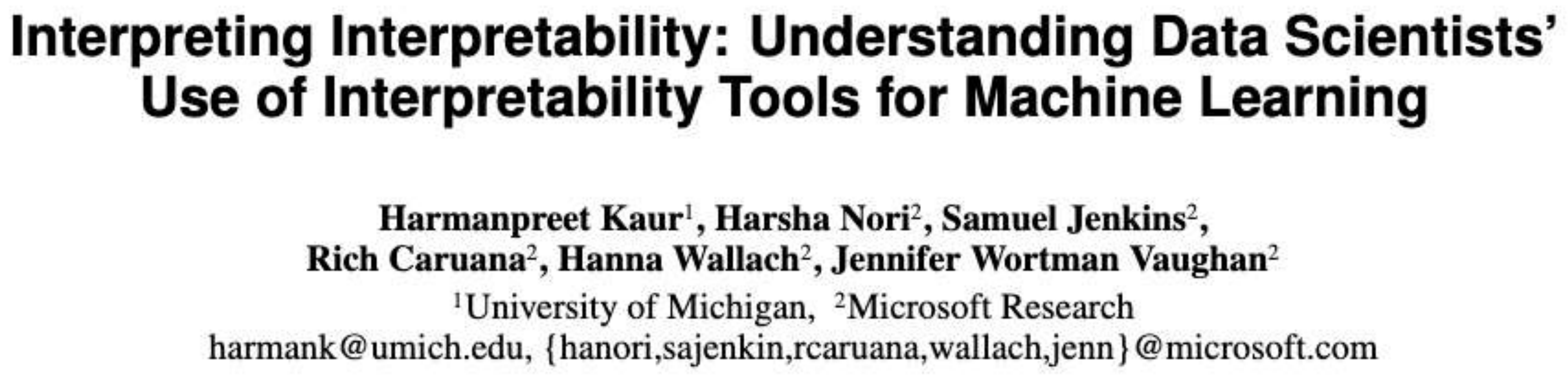
[Submitted on 22 Oct 2020]

Towards falsifiable interpretability research

Matthew L. Leavitt, Ari Morcos

"illustrative power of visualization is a double-edged sword: an evocative graphic can elicit a strong feeling of comprehension regardless of whether the graphic faithfully represents the phenomenon it is attempting to depict."

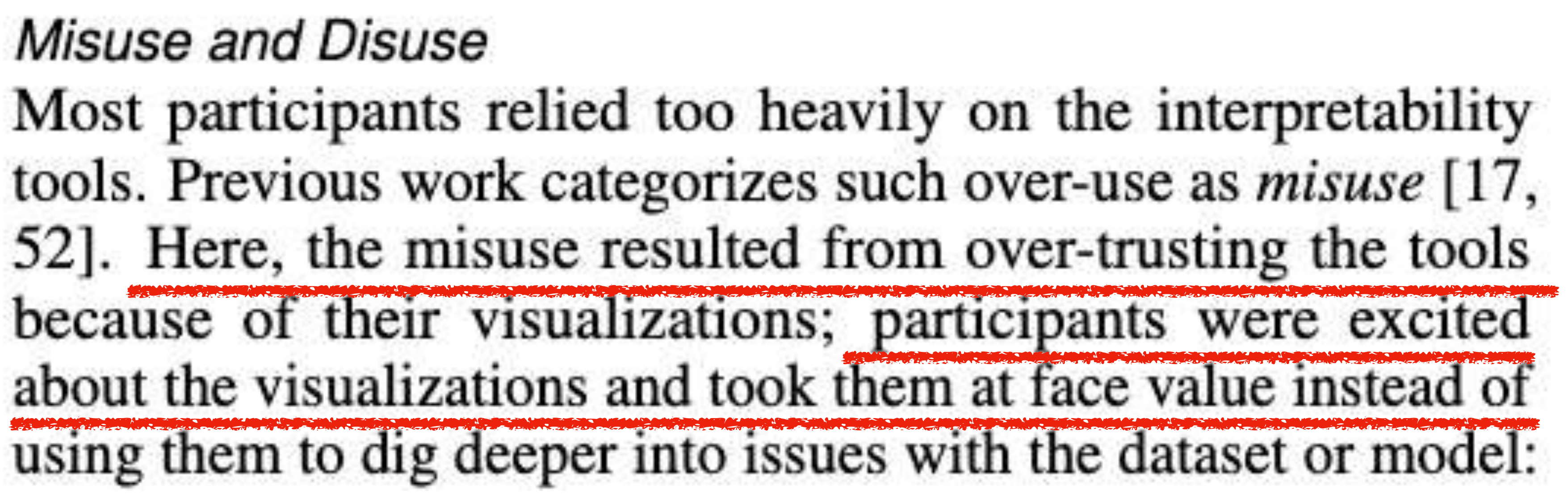
"How" *explanations are presented* is *as* important as the *explanations* themselves. Knowing how that impacts users is *even* more important



Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur[1], Harsha Nori[2], Samuel Jenkins[2],
Rich Caruana[2], Hanna Wallach[2], Jennifer Wortman Vaughan[2]
[1]University of Michigan, [2]Microsoft Research
harmank@umich.edu, {hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

"Our results indicate that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualizations output by these tools."



Misuse and Disuse

Most participants relied too heavily on the interpretability tools. Previous work categorizes such over-use as *misuse* [17, 52]. Here, the misuse resulted from over-trusting the tools because of their visualizations; participants were excited about the visualizations and took them at face value instead of using them to dig deeper into issues with the dataset or model:

[Submitted on 22 Oct 2020]

Towards falsifiable interpretability research

Matthew L. Leavitt, Ari Morcos

"illustrative power of visualization is a double-edged sword: an *evocative graphic* can elicit a strong feeling of comprehension regardless of whether the graphic faithfully represents the phenomenon it is *attempting* to depict."

Take away:

Human factors is tricky but important.

Agenda

- What and why
- !Caution!: Things to be careful when using and developing interpretability methods
- Evaluate: How to evaluate interpretability methods
- Methods: 3 types of methods and examples

Evaluation - yes you can.

- Testing with no humans, proxy task
- Testing with humans, proxy task
- Testing with humans and real task

no humans, proxy task

Using ground truth dataset and Sanity check

- Idea: Test the obvious

1. Test hypothesis that should be true by craft a ground-truth dataset
2. Test hypothesis that should be true using results on real dataset
3. Do sanity check: often testing hypothesis that should NOT be true. a.k.a. as crazy questions.

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

Forest

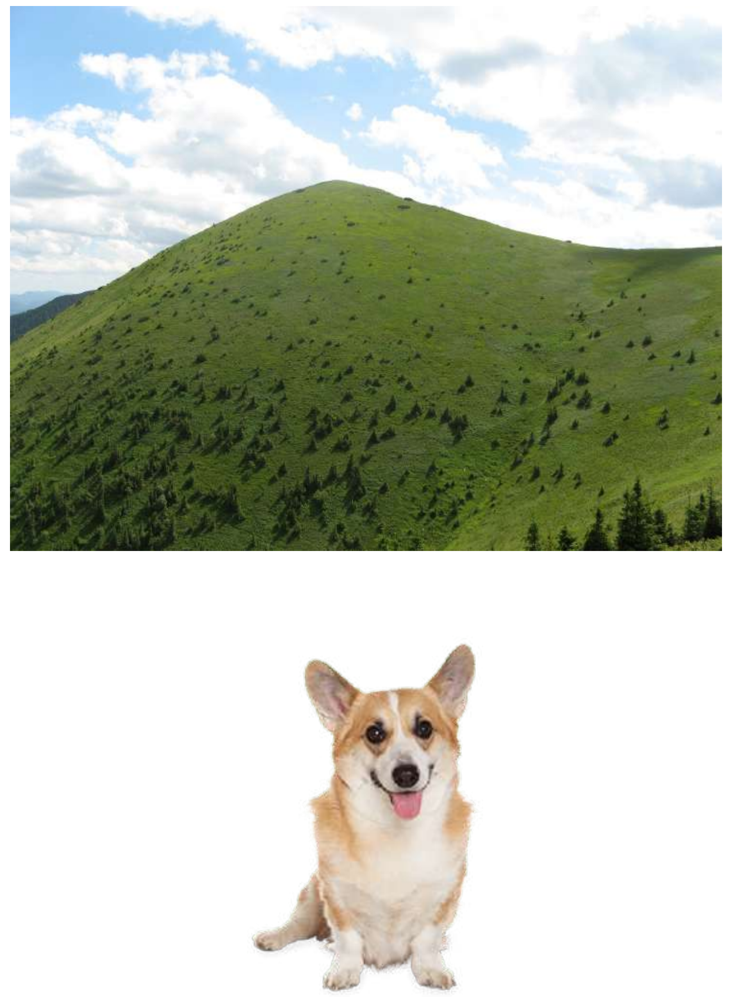
A thing

Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset



Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

Forest

A thing

Forest

 is NOT important for predicting scene classes.

↓

 should NOT Be part of explanation

Bedroom

Kitchen

Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

Forest

A thing

Forest

Bedroom

Kitchen

🐕 is NOT important for predicting scene classes.
↓
🐕 should NOT
Be part of explanation

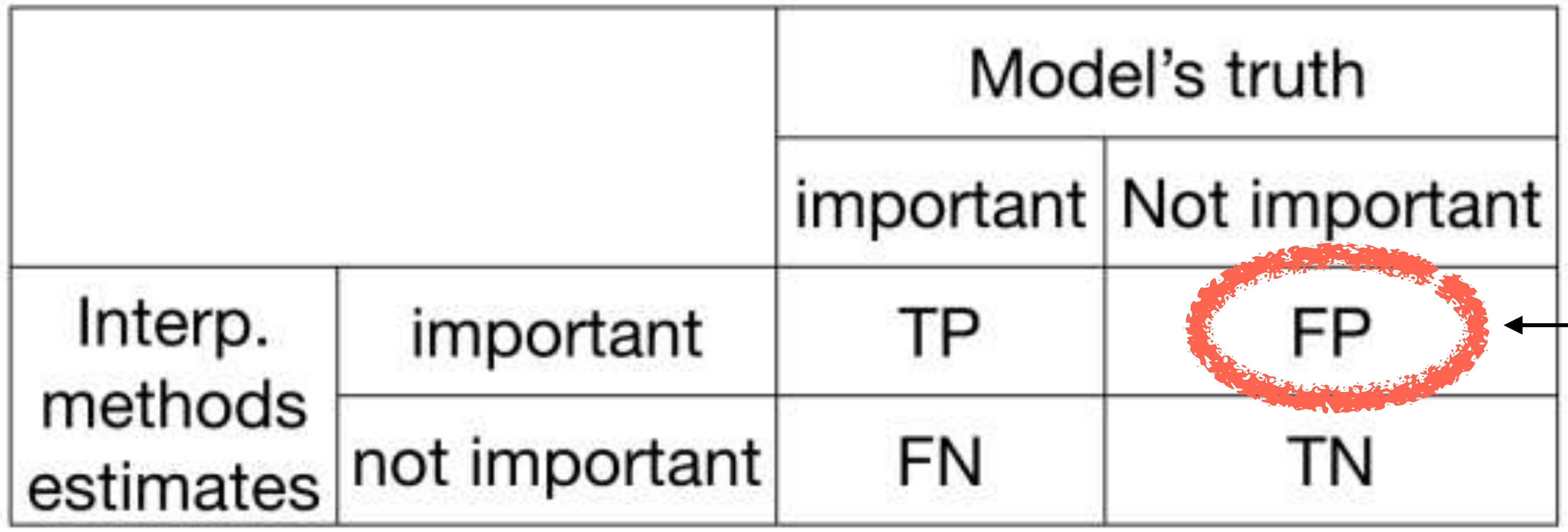
We can also make
🐕 more important
to some classes by
controlling when it appears.
↓
🐕 should be more important
explanation in some classes
than others.

Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset



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

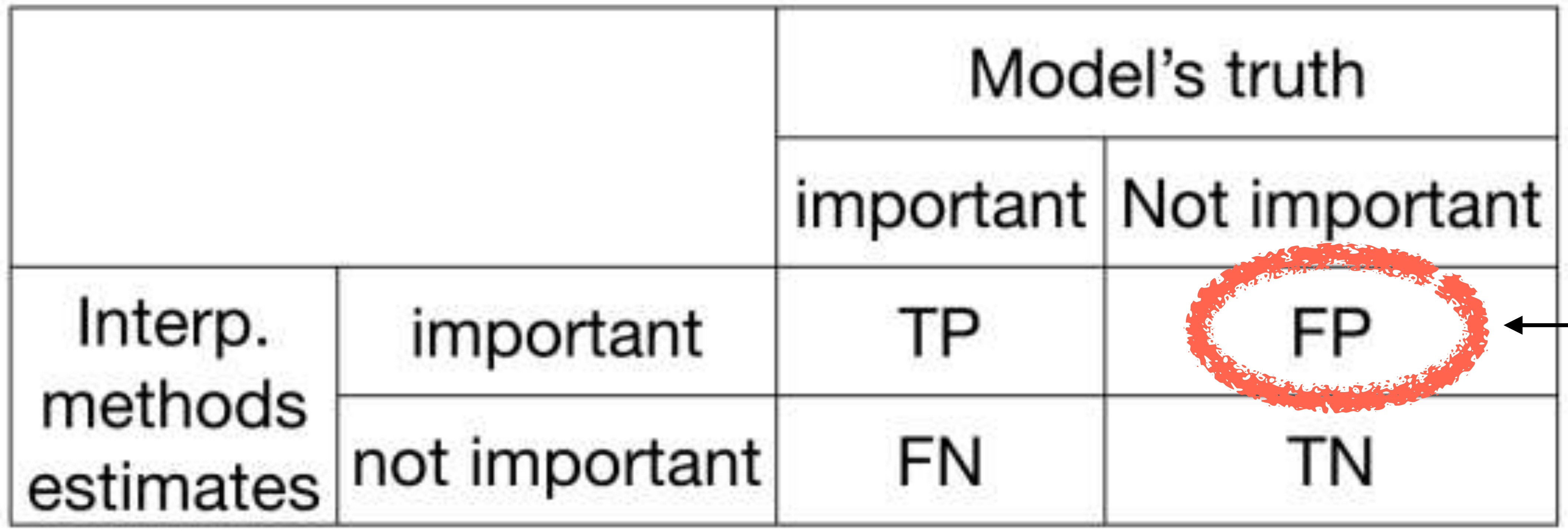
Our Focus: False positives

Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset



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

Our Focus: False positives

Suggested metrics

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

Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

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

← Our Focus: False positives

Suggested metrics

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

→

Two models trained to classify scenes.

Model 1

Model 2

Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

		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

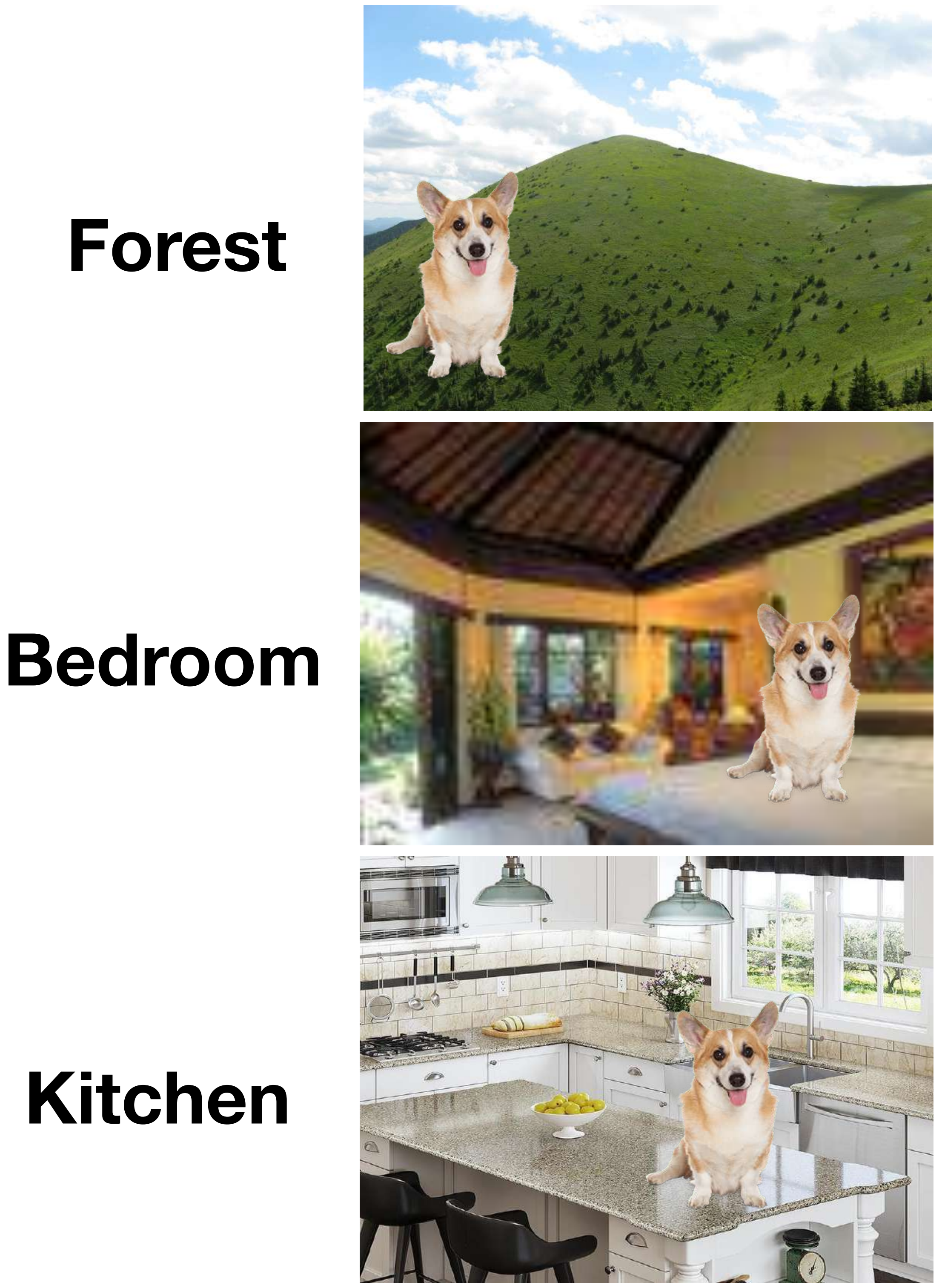
We expect big contrast on where the object is.

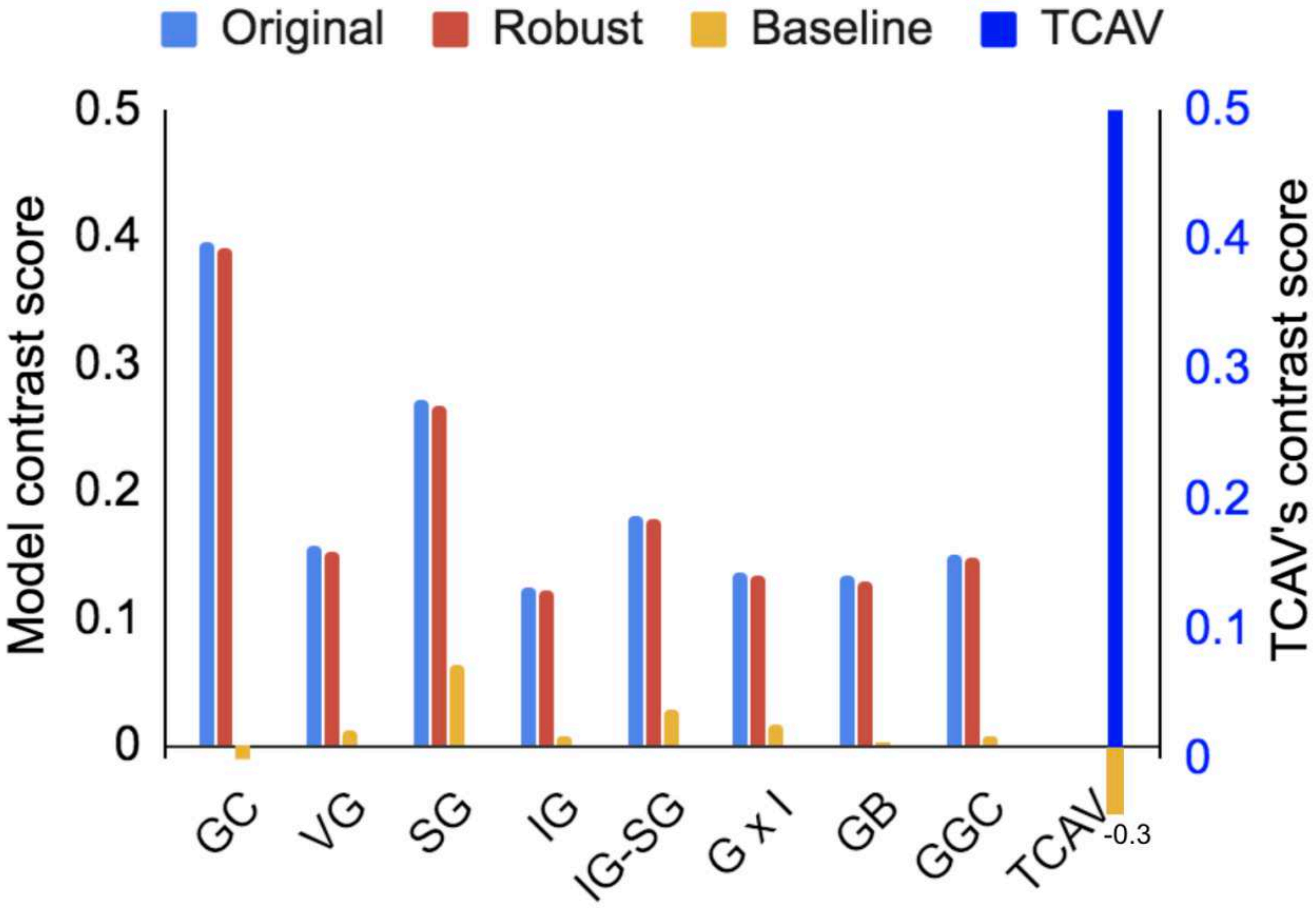
Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset



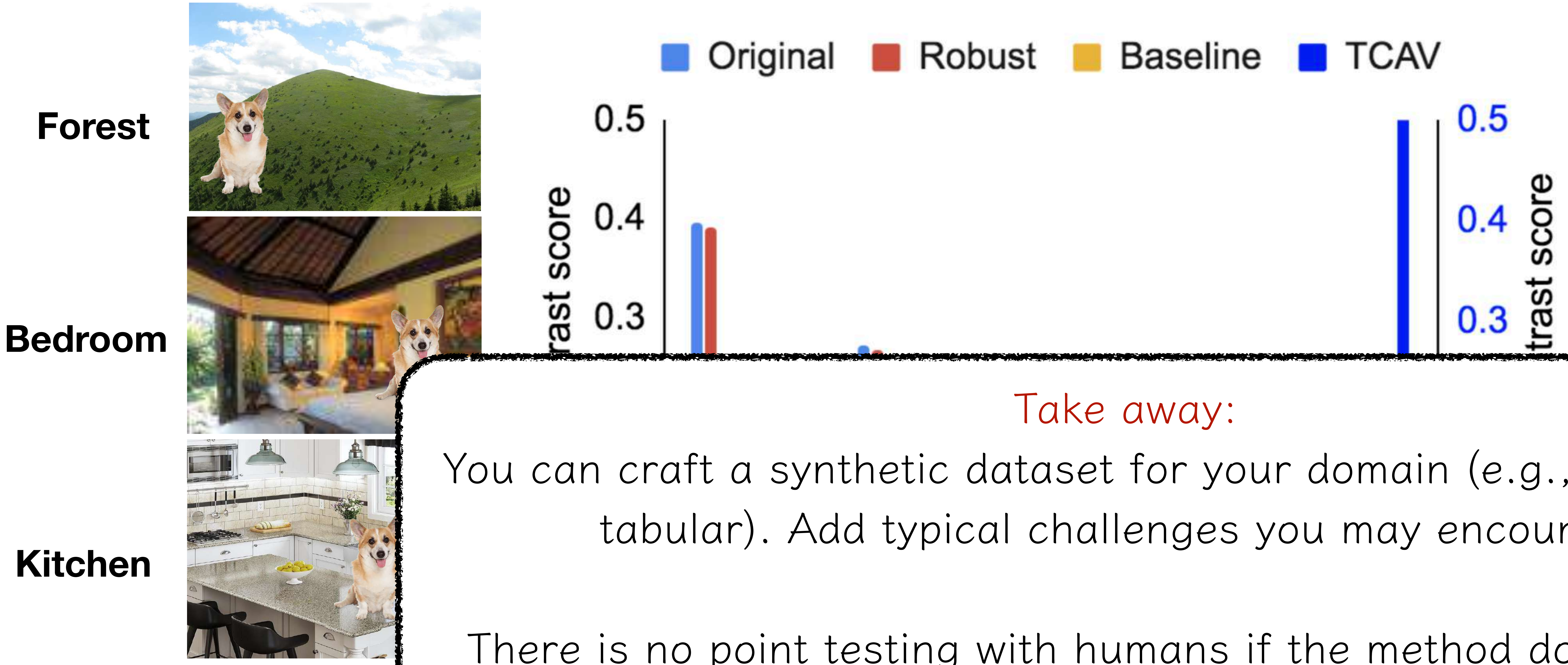


Benchmarking interpretability methods (BIM)

github.com/google-research-datasets/bim

no humans, proxy task

1: Test hypothesis that should be true by craft a <u>ground-truth dataset</u>



Take away:

You can craft a synthetic dataset for your domain (e.g., sequential, tabular). Add typical challenges you may encounter.

There is no point testing with humans if the method doesn't pass these tests.

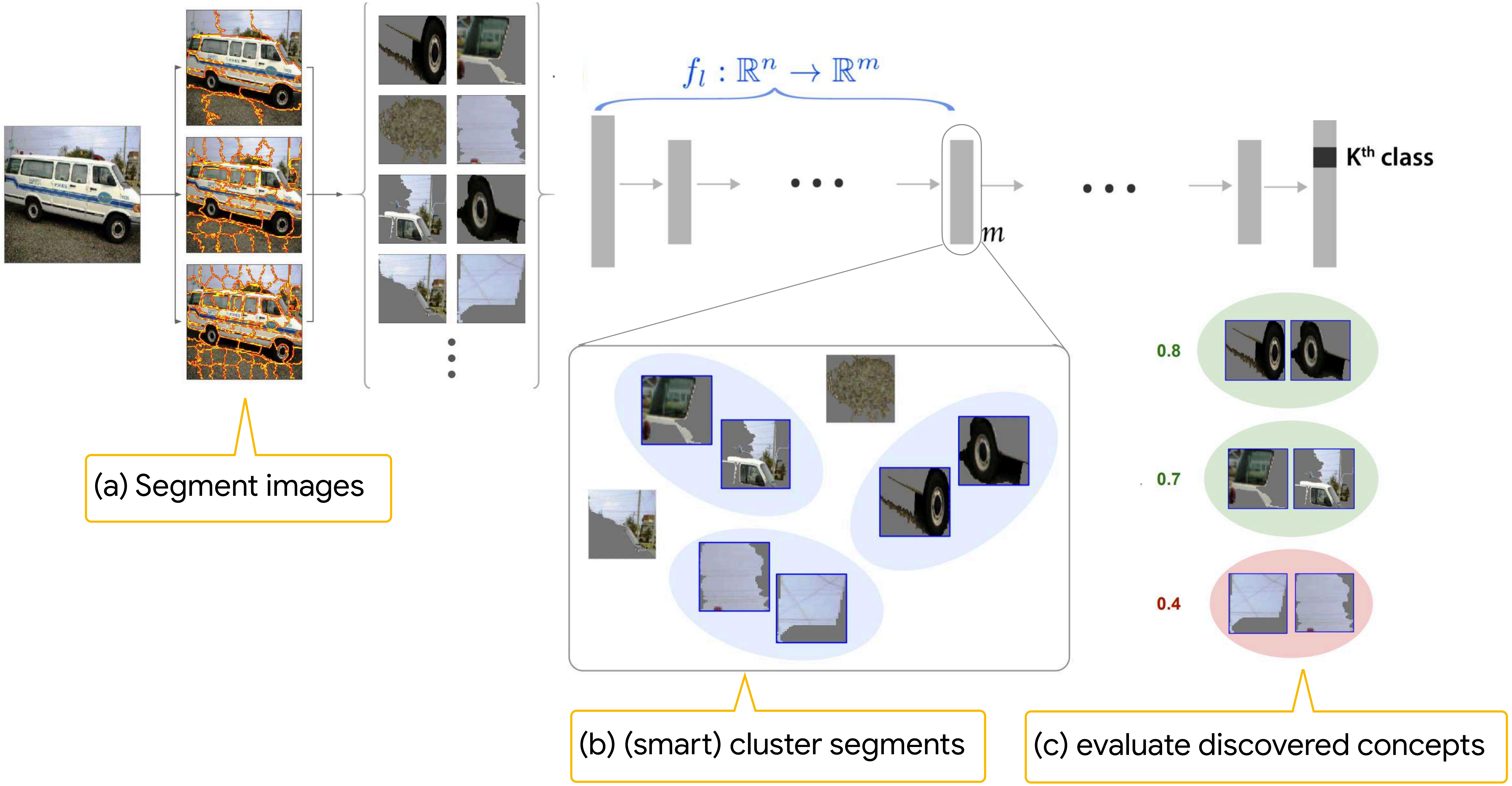
Benchmarking interpretability methods (

github.com/google-research-datasets/bim

no humans, proxy task

2: Test hypothesis that should be true using results on real dataset

Automatic Concept-based Explanations (ACE)
[Ghorbani et al. NeurIPS 19]



(a) Segment images

(b) (smart) cluster segments

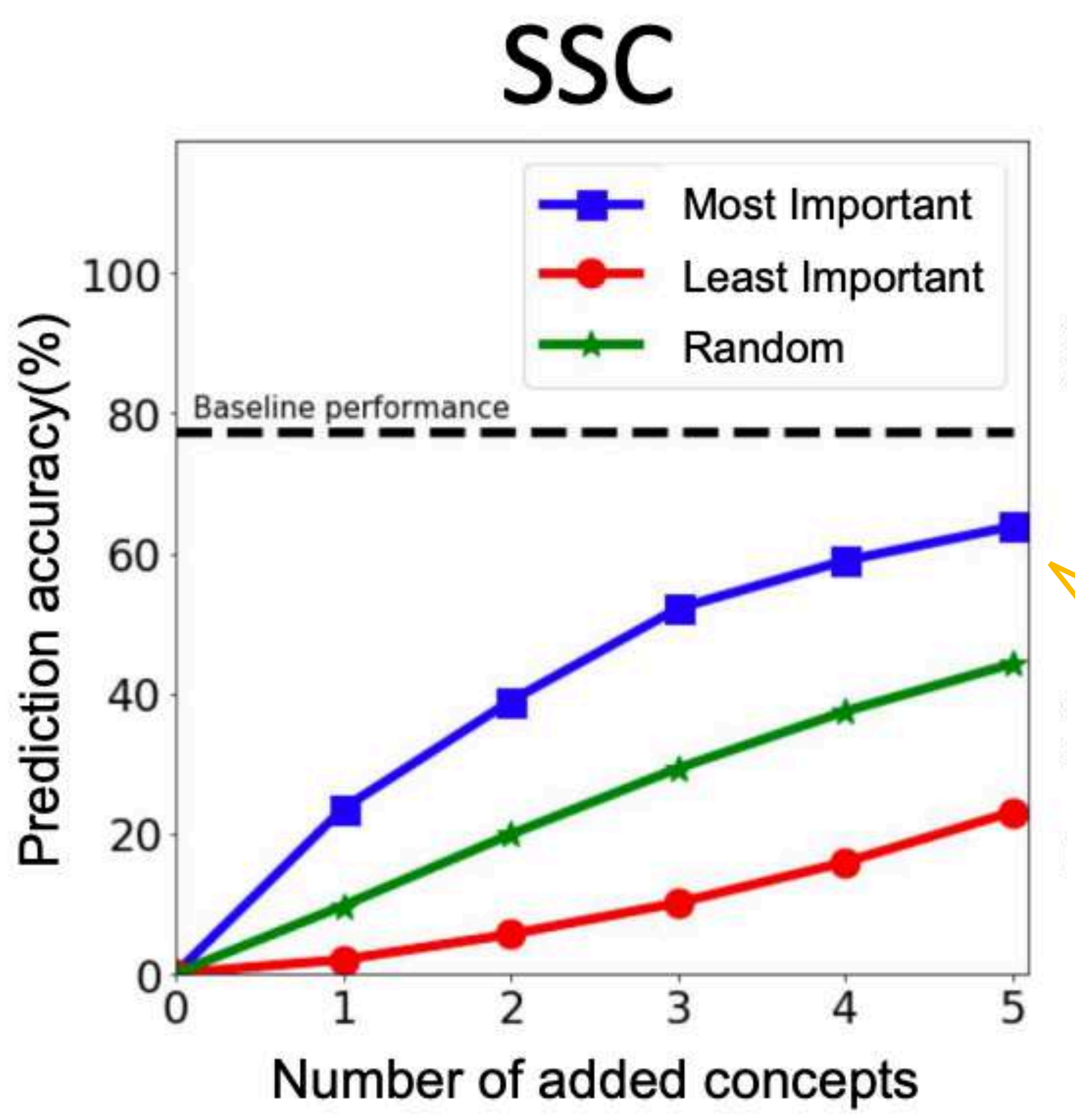
(c) evaluate discovered concepts

no humans, proxy task

2: Test hypothesis that should be true using results on real dataset

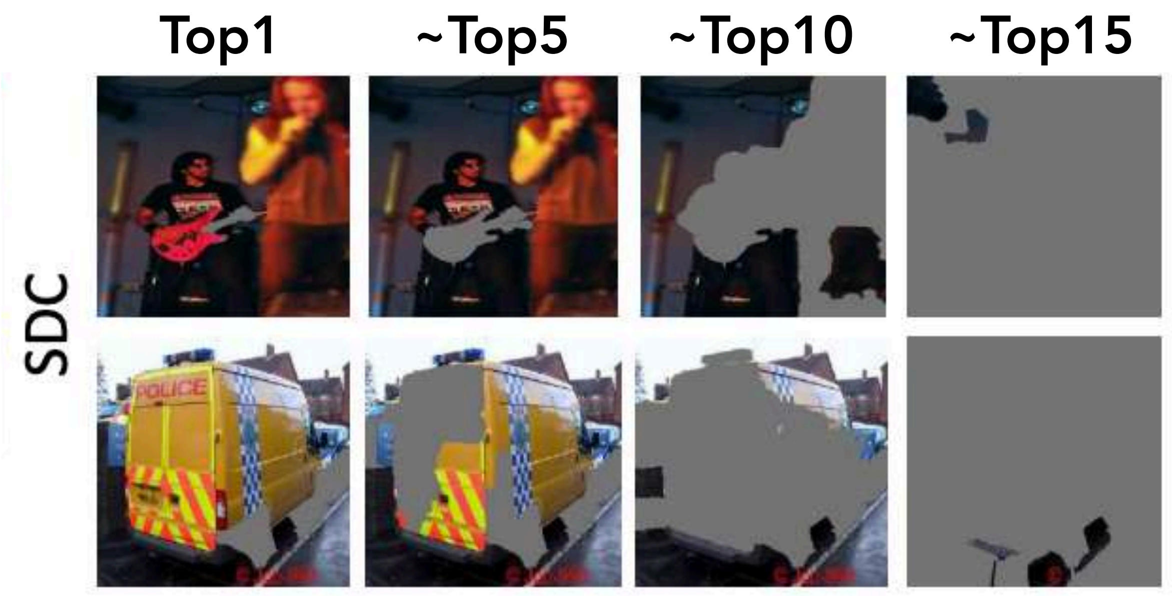
Automatic Concept-based Explanations (ACE)
[Ghorbani et al. NeurIPS 19]

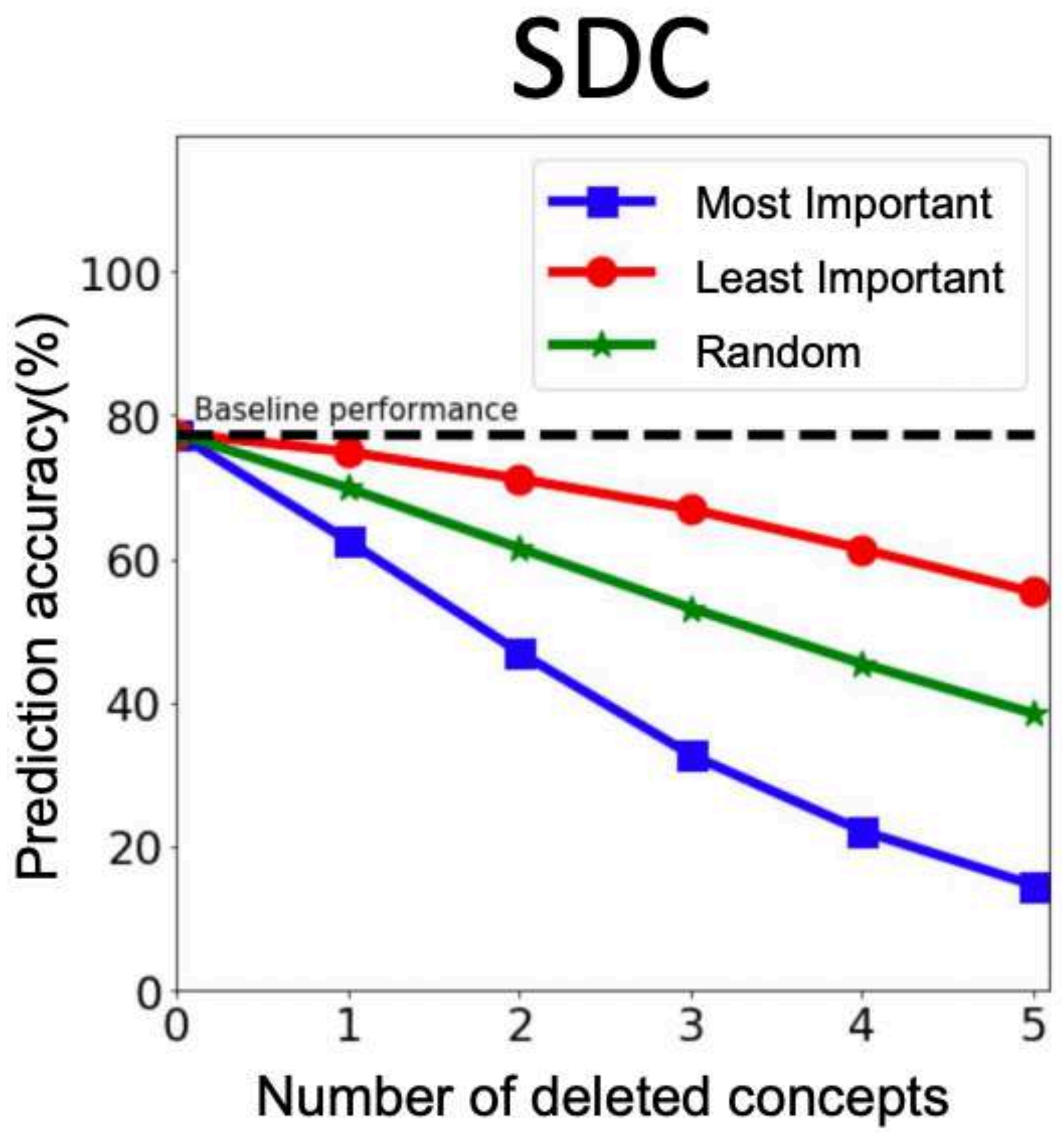
Adding top-rated patches



Adding the top5 discovered concepts alone achieves 70% of the original accuracy

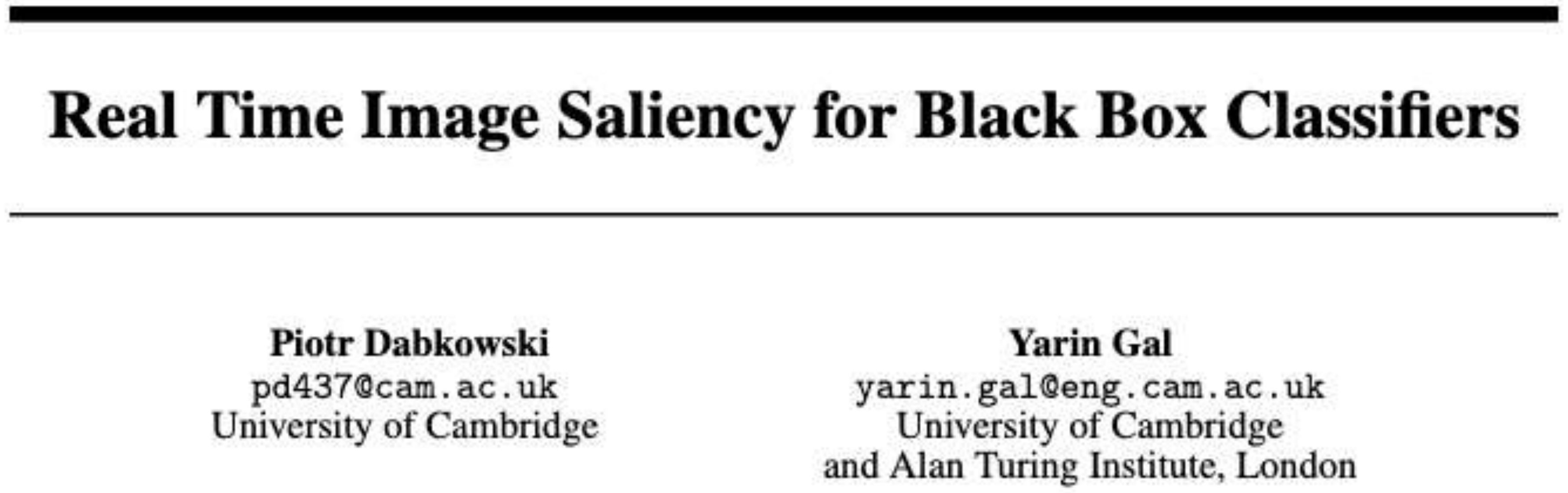
Deleting top-rated patches





no humans, proxy task

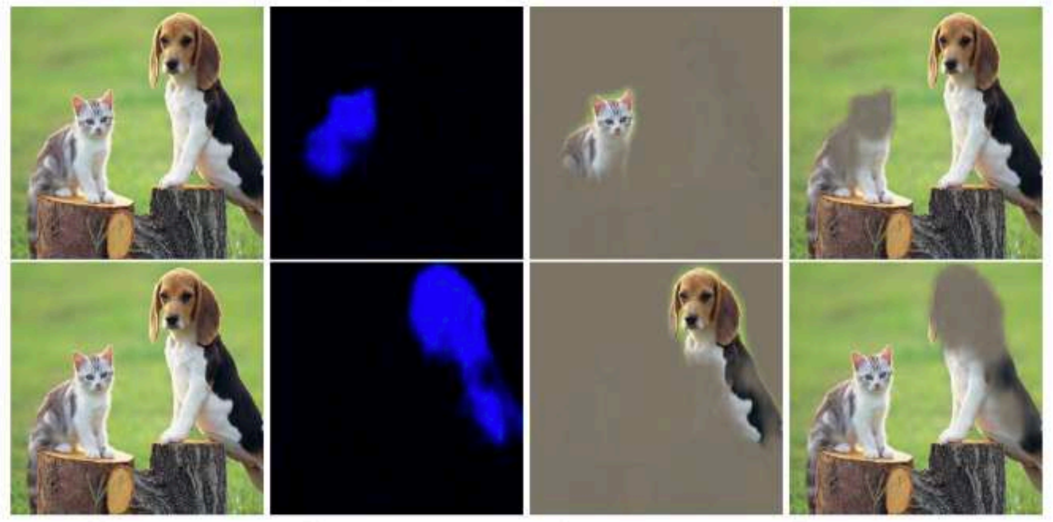
2: Test hypothesis that should be true using results on real dataset



Real Time Image Saliency for Black Box Classifiers

Piotr Dabkowski
pd437@cam.ac.uk
University of Cambridge

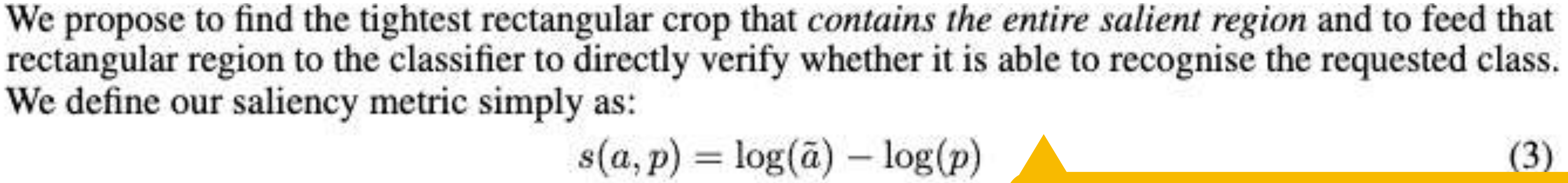
Yarin Gal
yarin.gal@eng.cam.ac.uk
University of Cambridge
and Alan Turing Institute, London



(a) Input Image (b) Generated saliency map (c) Image multiplied by the mask (d) Image multiplied by inverted mask

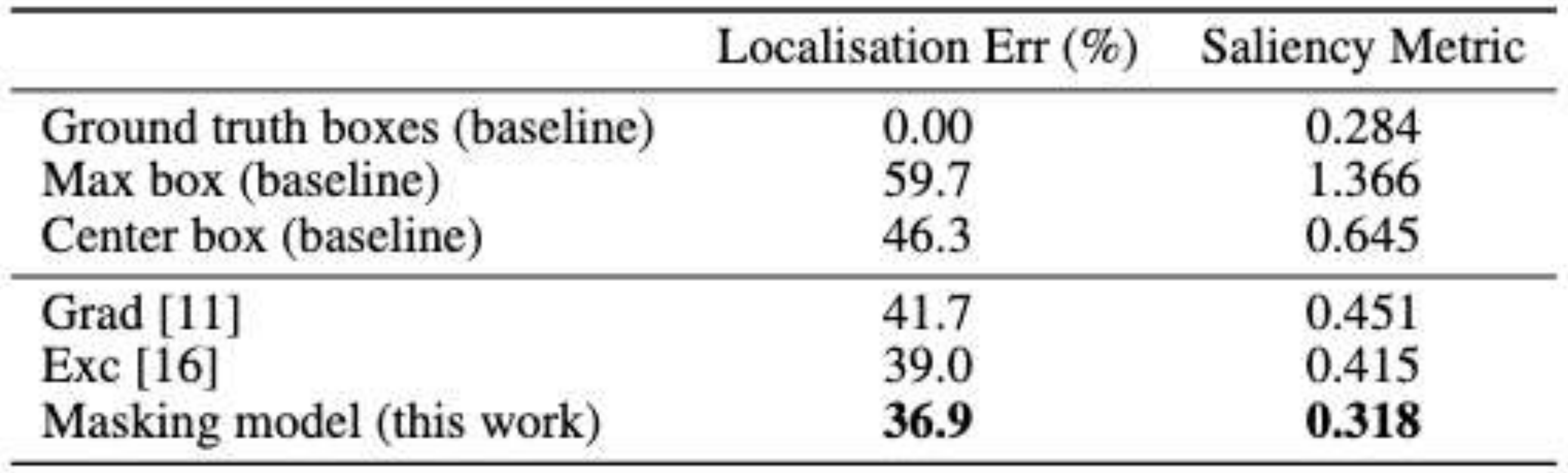
Figure 1: An example of explanations produced by our model. The top row shows the explanation for the "Egyptian cat" while the bottom row shows the explanation for the "Beagle". Note that produced explanations can precisely both highlight and remove the selected object from the image.

smallest sufficient region (SSR) - smallest region of the image that alone allows a confident classification.



We propose to find the tightest rectangular crop that *contains the entire salient region* and to feed that rectangular region to the classifier to directly verify whether it is able to recognise the requested class. We define our saliency metric simply as:

$$s(a, p) = \log(\tilde{a}) - \log(p) \tag{3}$$



	Localisation Err (%)	Saliency Metric
Ground truth boxes (baseline)	0.00	0.284
Max box (baseline)	59.7	1.366
Center box (baseline)	46.3	0.645
Grad [11]	41.7	0.451
Exc [16]	39.0	0.415
Masking model (this work)	**36.9**	**0.318**

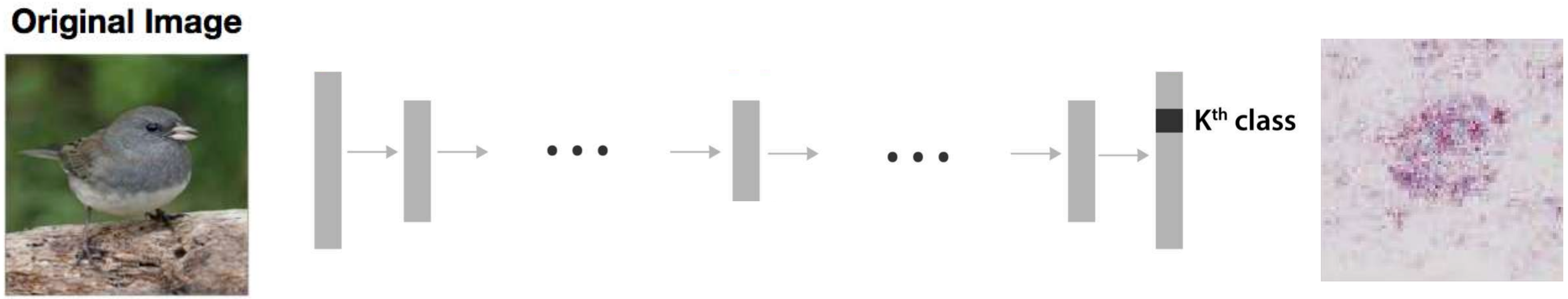
Lower is the better:

Relative amount of information between p and a (concentration of information in the cropped region)

no humans, proxy task

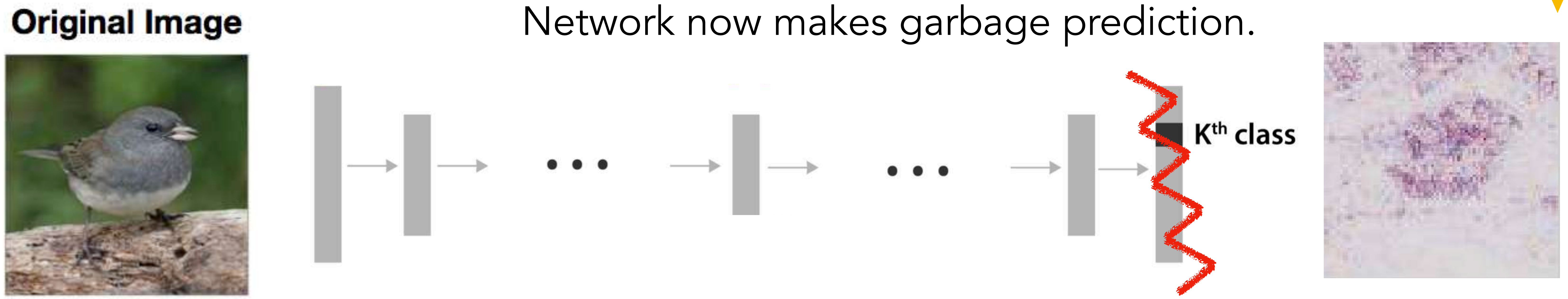
3: Do sanity check: often testing hypothesis that <u>should NOT be true</u>.

a.k.a. ask crazy questions.



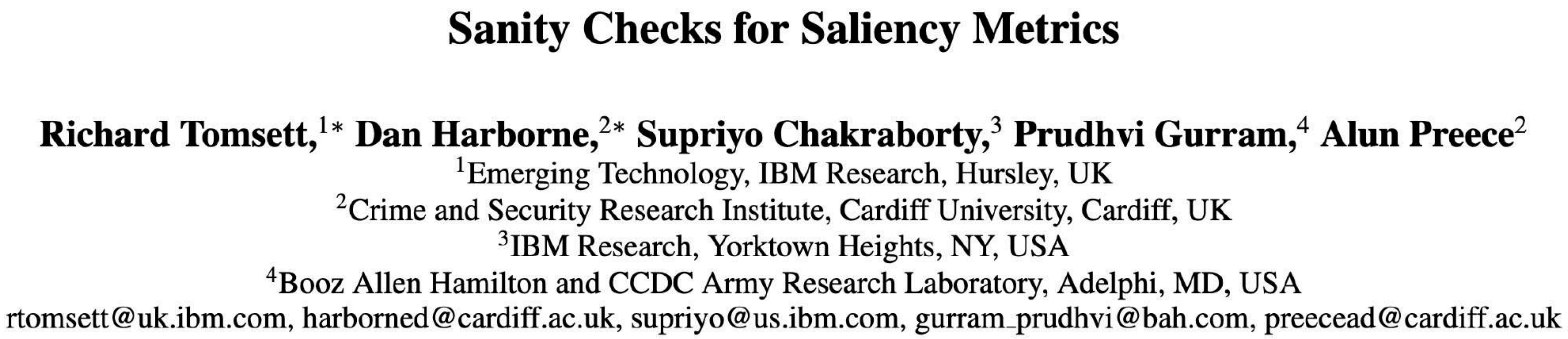
Randomized weights!
Network now makes garbage prediction.

!!!!!???!?



no humans, proxy task

3: Do sanity check: often testing hypothesis that <u>should NOT be true</u>.
a.k.a. ask crazy questions.

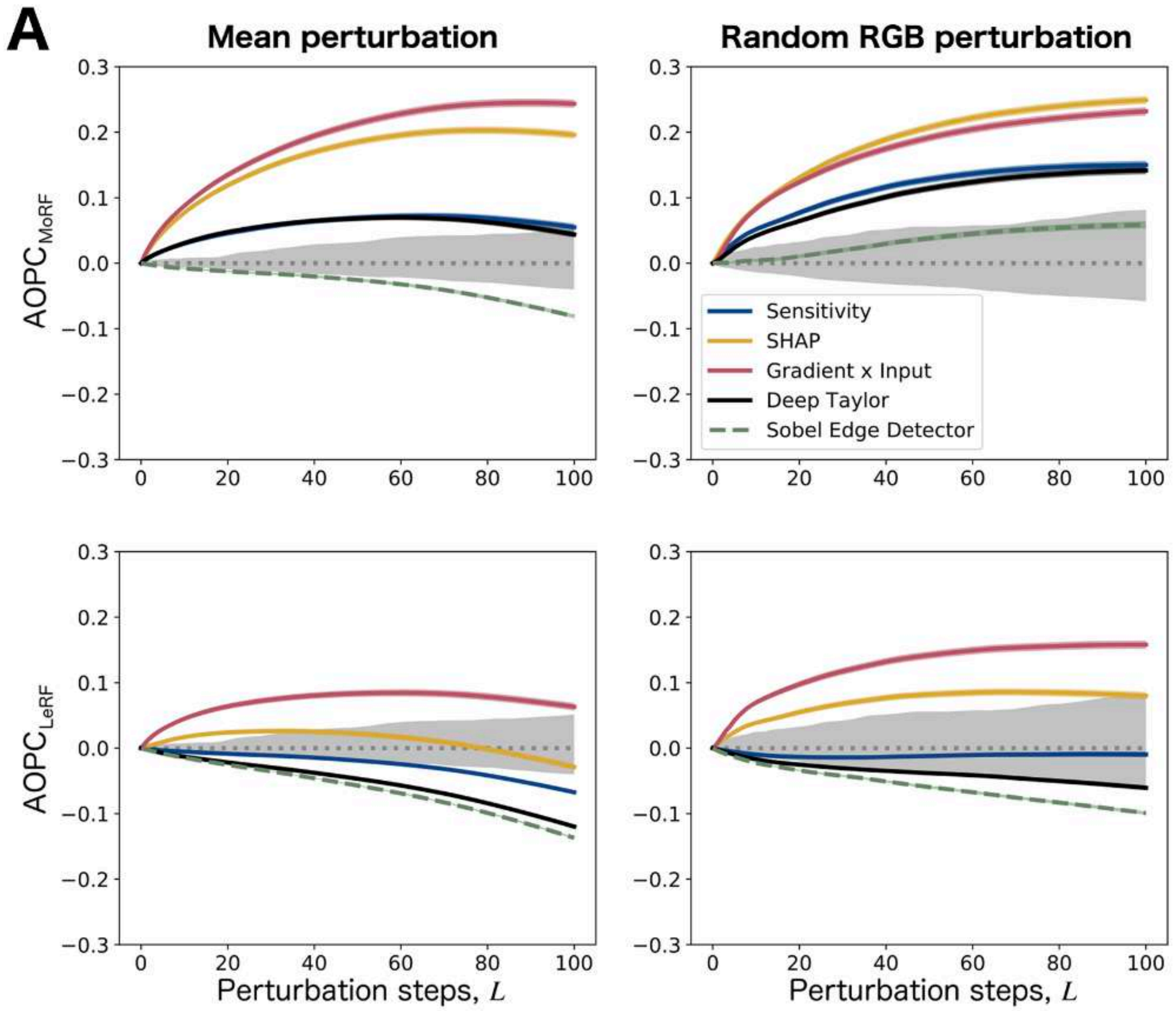


Sanity Checks for Saliency Metrics

Richard Tomsett,[1]* Dan Harborne,[2]* Supriyo Chakraborty,[3] Prudhvi Gurram,[4] Alun Preece[2]
[1]Emerging Technology, IBM Research, Hursley, UK
[2]Crime and Security Research Institute, Cardiff University, Cardiff, UK
[3]IBM Research, Yorktown Heights, NY, USA
[4]Booz Allen Hamilton and CCDC Army Research Laboratory, Adelphi, MD, USA
rtomsett@uk.ibm.com, harborned@cardiff.ac.uk, supriyo@us.ibm.com, gurram_prudhvi@bah.com, preecead@cardiff.ac.uk

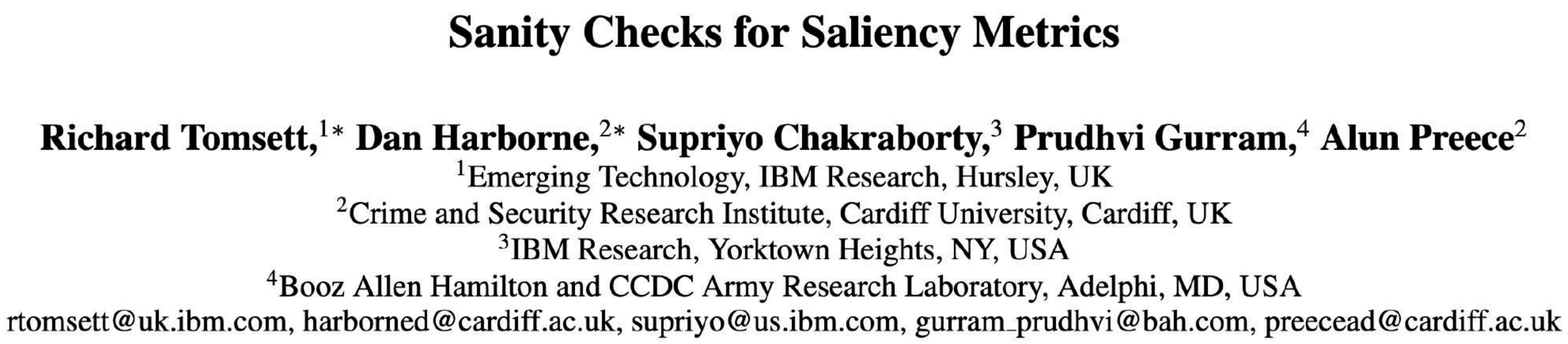
"Our results show that saliency metrics can be statistically unreliable and inconsistent, indicating that comparative rankings between saliency methods generated using such metrics can be untrustworthy.

1. Global saliency metrics had high variance
2. Saliency metrics were sensitive to the specifics of their implementation
3. Saliency maps from different saliency methods were ranked inconsistently image-by-image
4. The internal consistency of different metrics that all attempt to measure fidelity was low



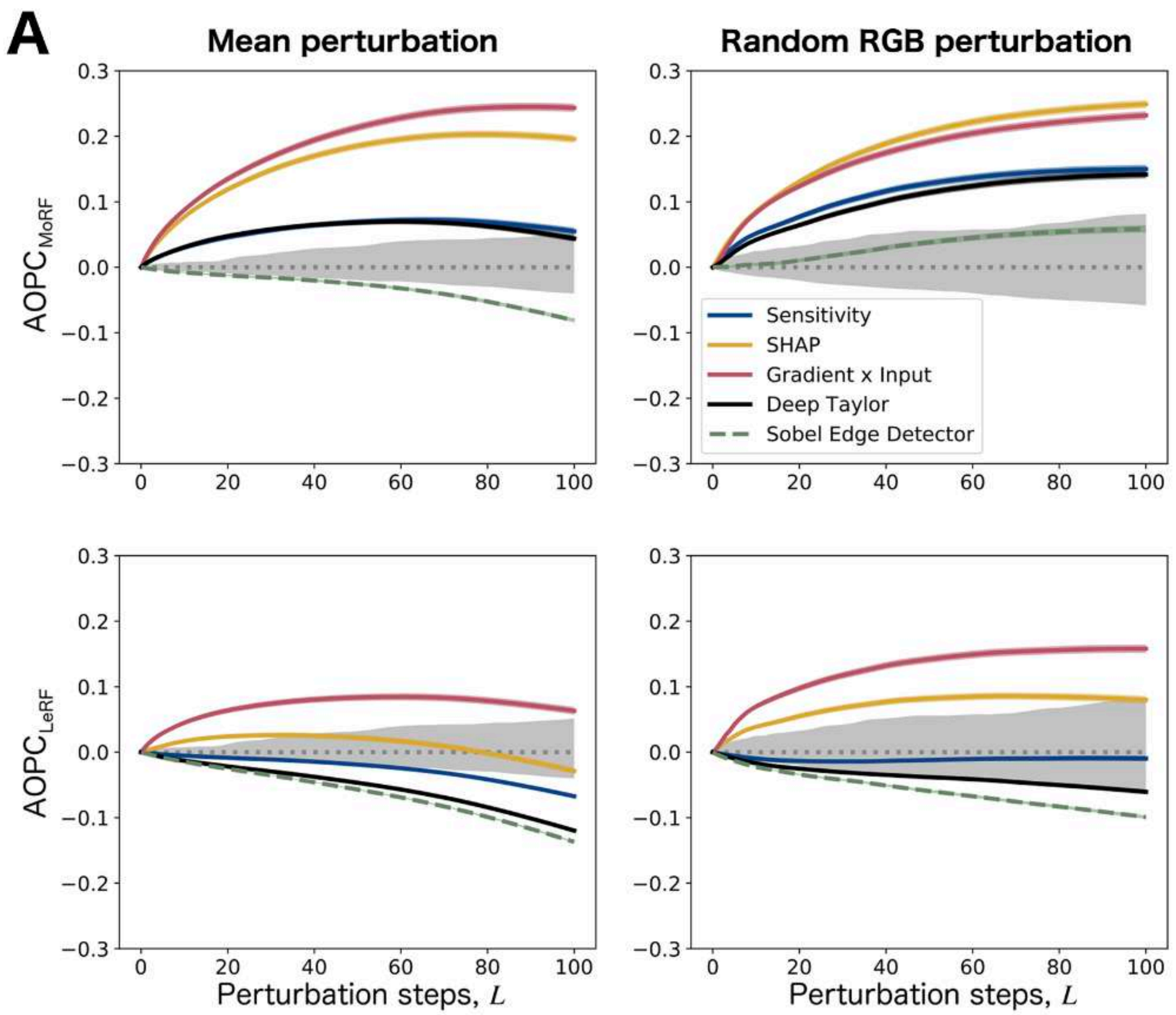
no humans, proxy task

3: Do sanity check: often testing hypothesis that should NOT be true. a.k.a. ask crazy questions.



Sanity Checks for Saliency Metrics

Richard Tomsett,[1]* **Dan Harborne,**[2]* **Supriyo Chakraborty,**[3] **Prudhvi Gurram,**[4] **Alun Preece**[2]
[1]Emerging Technology, IBM Research, Hursley, UK
[2]Crime and Security Research Institute, Cardiff University, Cardiff, UK
[3]IBM Research, Yorktown Heights, NY, USA
[4]Booz Allen Hamilton and CCDC Army Research Laboratory, Adelphi, MD, USA
rtomsett@uk.ibm.com, harborned@cardiff.ac.uk, supriyo@us.ibm.com, gurram_prudhvi@bah.com, preecead@cardiff.ac.uk



"Our results show that saliency metrics can be statistically unreliable and inconsistent, indicating that comparative rankings between saliency methods generated using such metrics can be untrustworthy.

1. Global saliency metrics had high variance
2. Saliency metrics were sensitive to the specifics of their implementation
3. Saliency maps from different saliency methods were ranked inconsistently image-by-image
4. The internal consistency of different metrics that all attempt to measure fidelity was low

Take away:

Being skeptical can be healthy and productive.

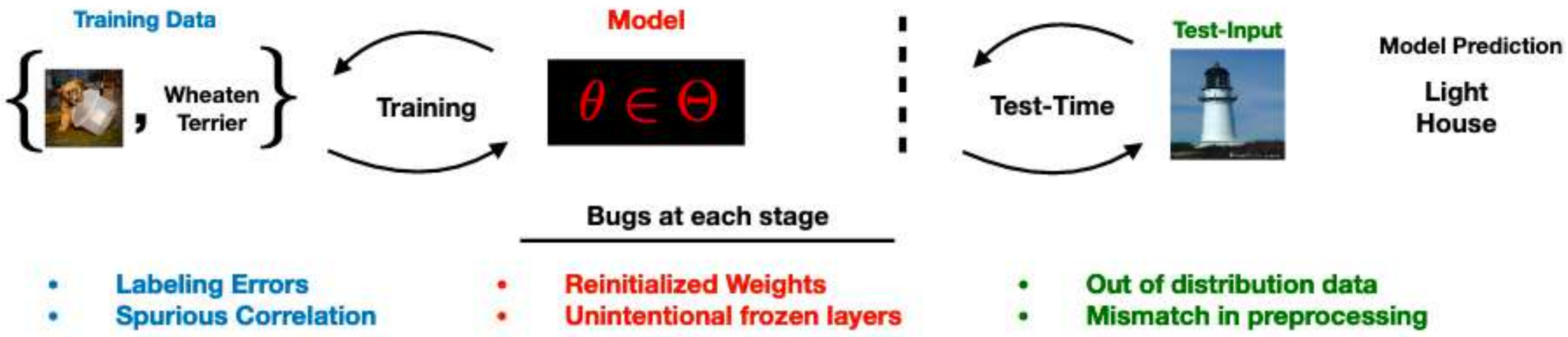
Testing with humans, proxy task

- In *a* proxy task that maintains the *essence* of the final task (but likely ground truth is known)

- with humans who may not be your *idea* users (*e.g.*, doctors) but still *can* help *evaluating*

humans, proxy task

Testing methods with users and concrete end-tasks



- **Task for subjects**: You work at a start-up selling animal classification ML model. Here are the images, predictions and attribution maps. (We gave users prediction labels as it is unrealistic not to).
- **Questions:** Would you recommend this model? Why? [because the wrong/correct label/explanation]? All in Likert scale.

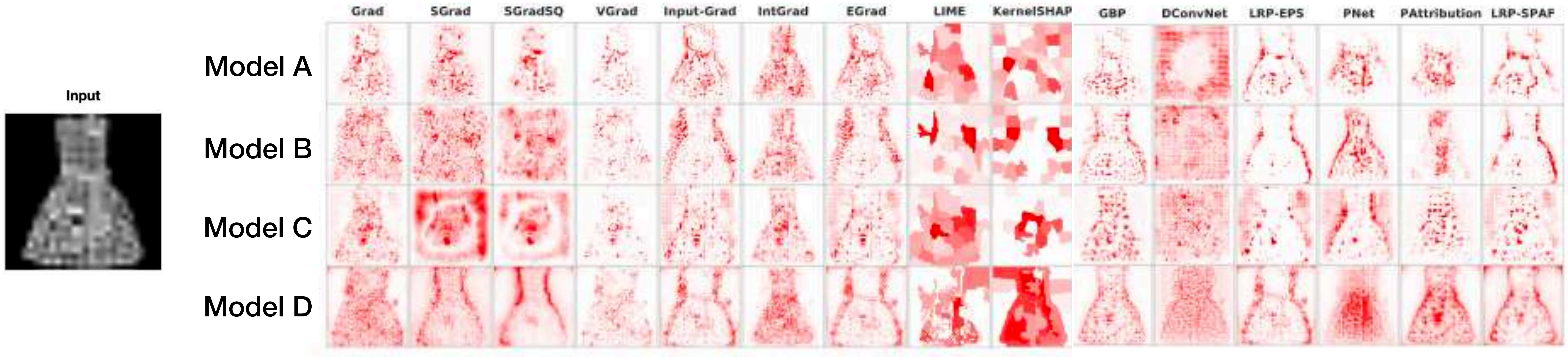
[Adebayo, Muelly, Liccardi, K. Neurips 2020]

humans, proxy task

Can these methods tell us about

Out of distribution?

- Out of distribution data

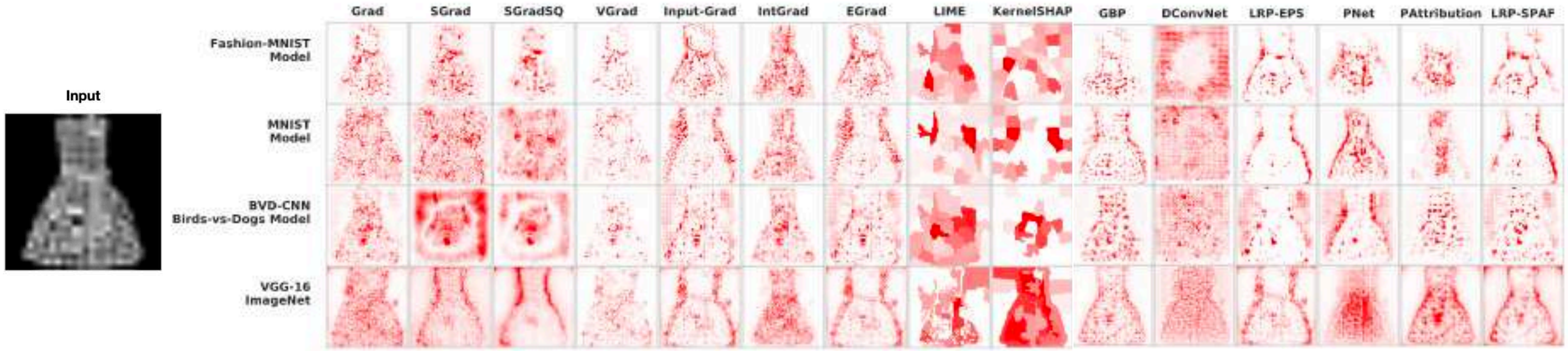


humans, proxy task

Can these methods tell us about

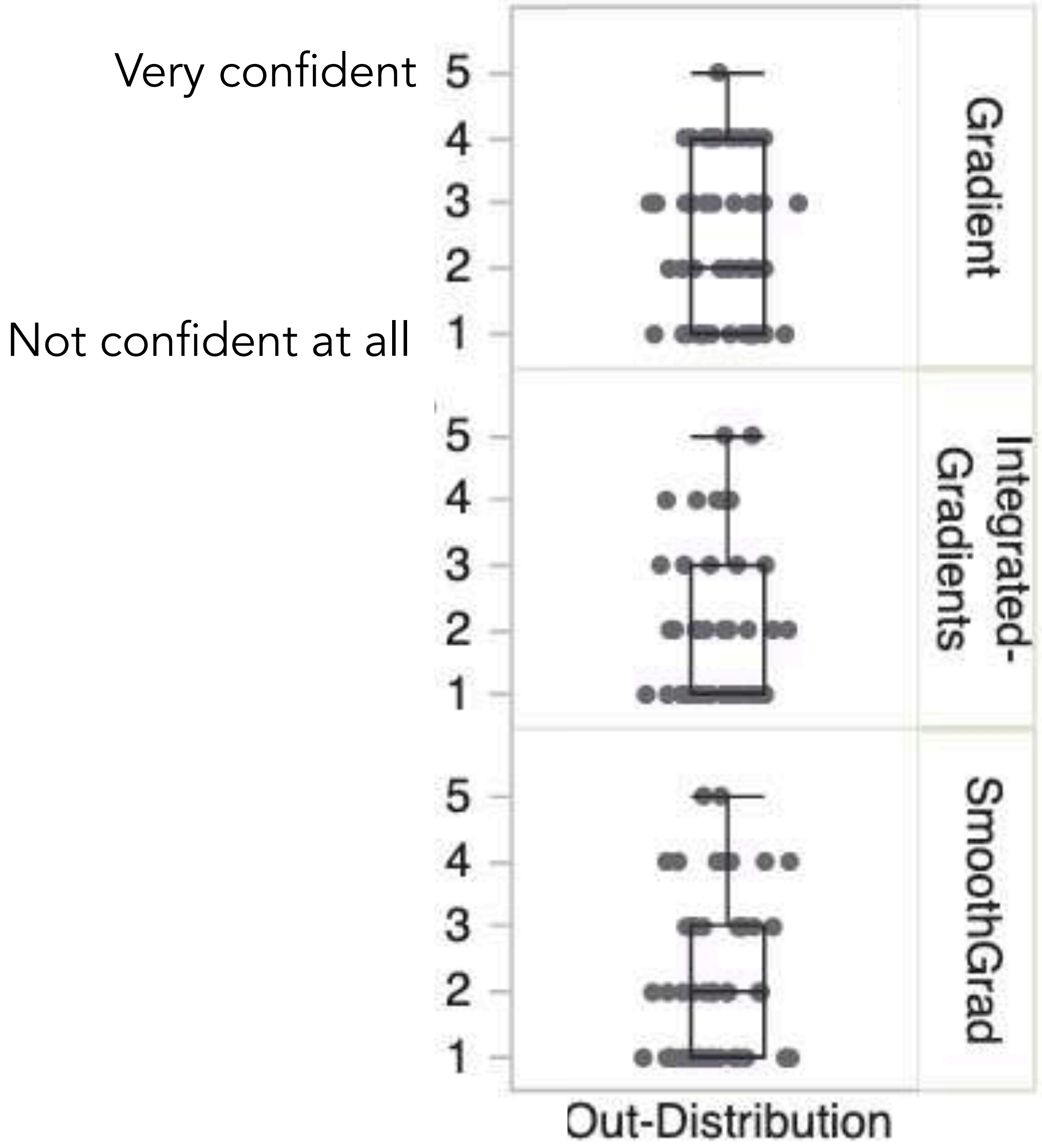
Out of distribution? probably not.

- Out of distribution data

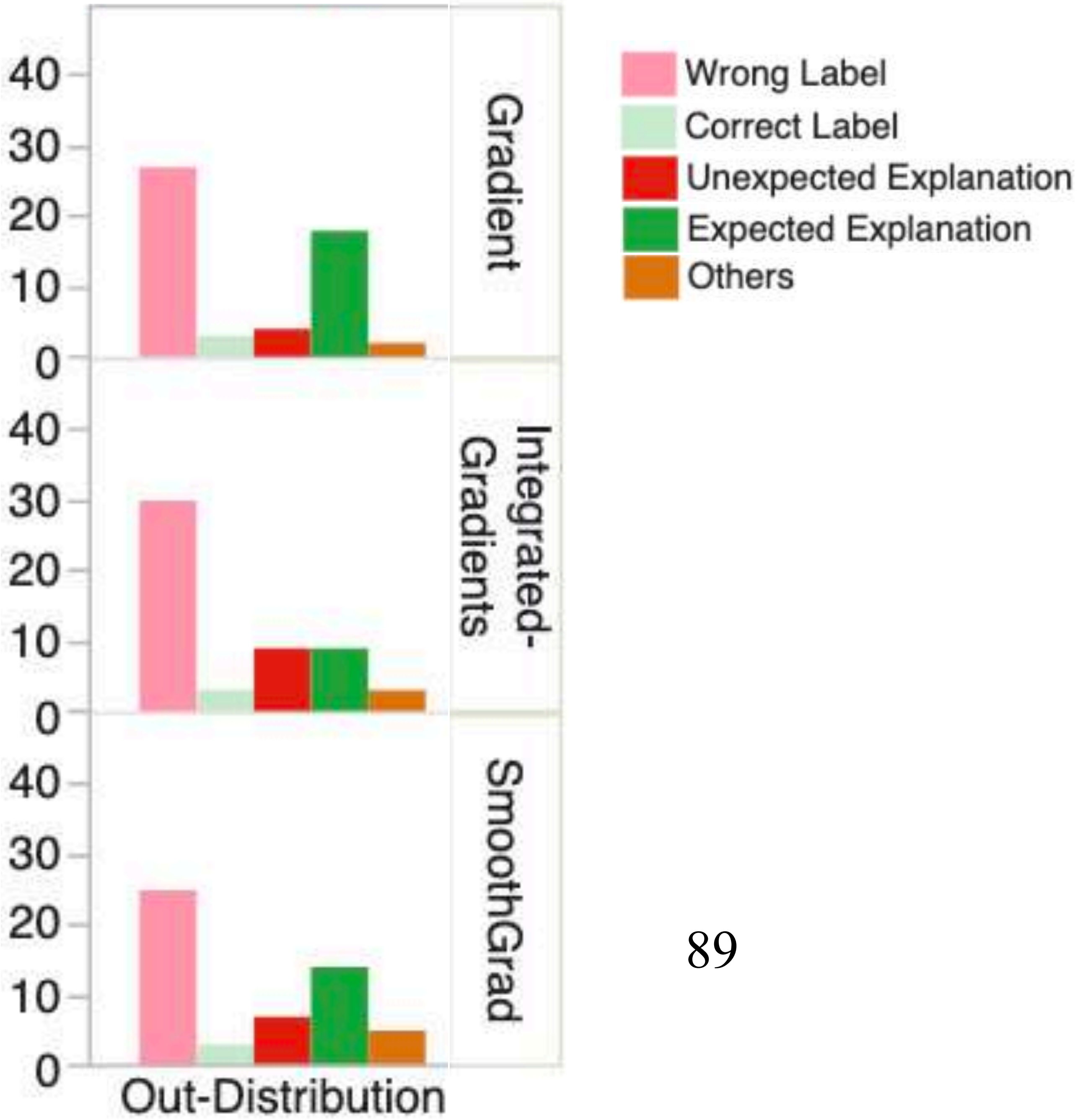


Subjects are uncertain, mostly because of wrong label, but some **expected** explanations.

How confident are you to deploy this model?



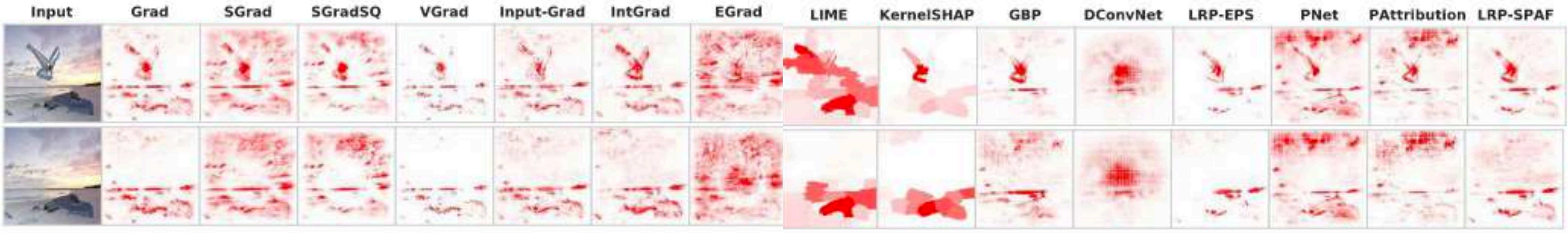
% why



humans, proxy task

Can these methods tell us about

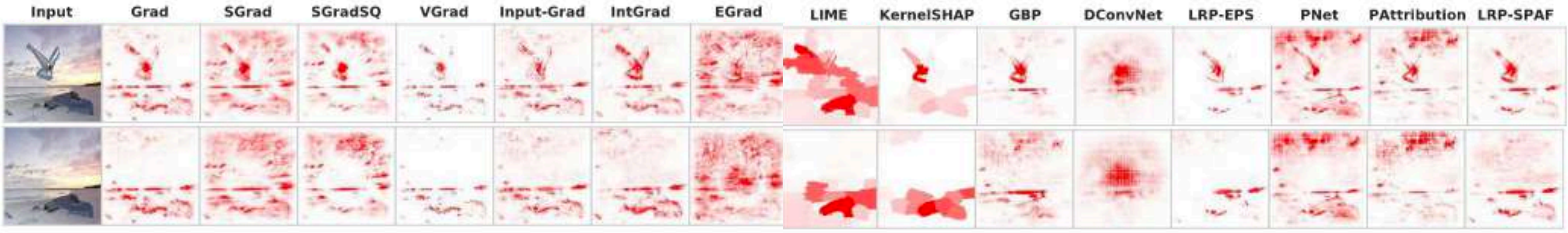
Spurious correlation?

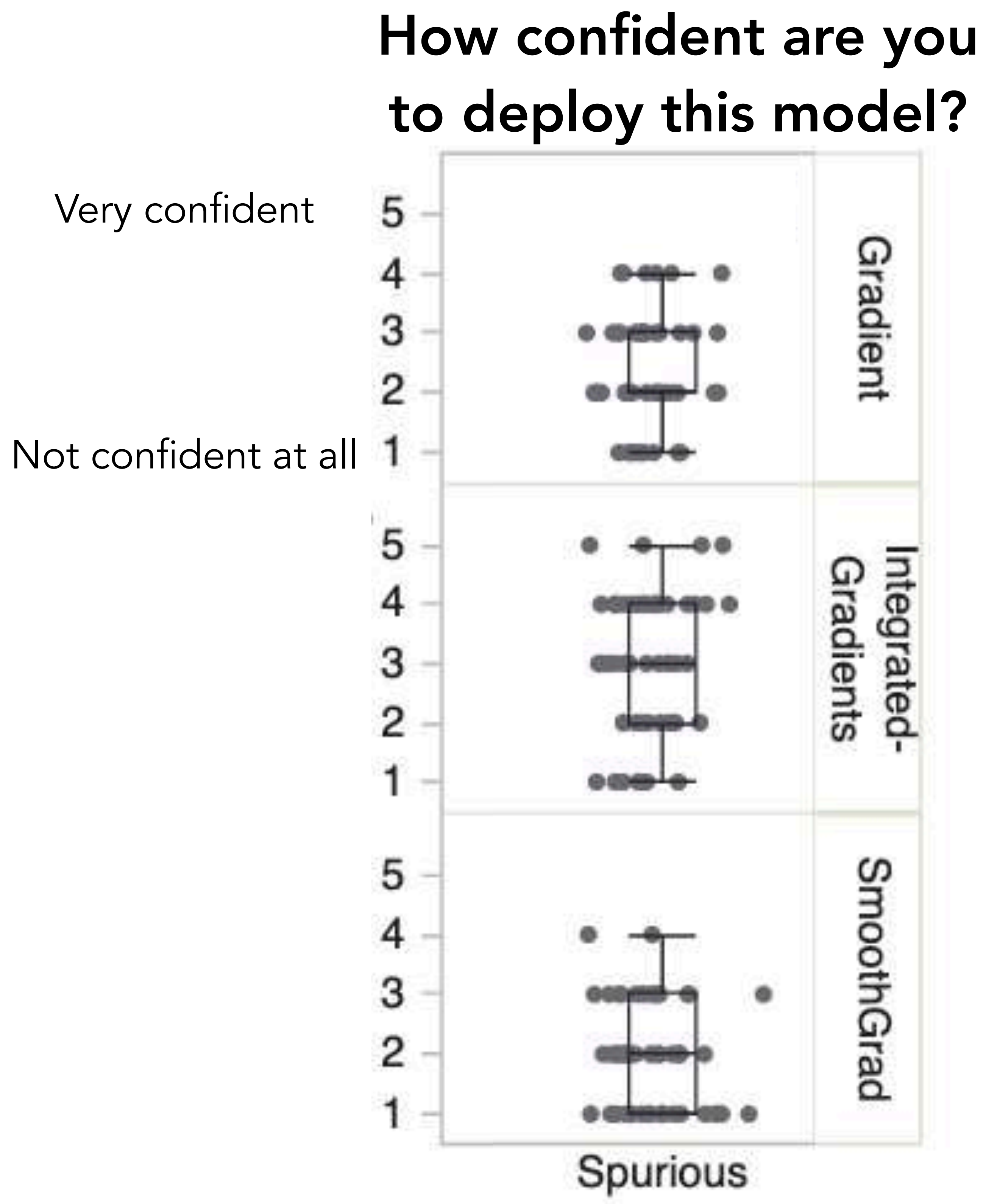


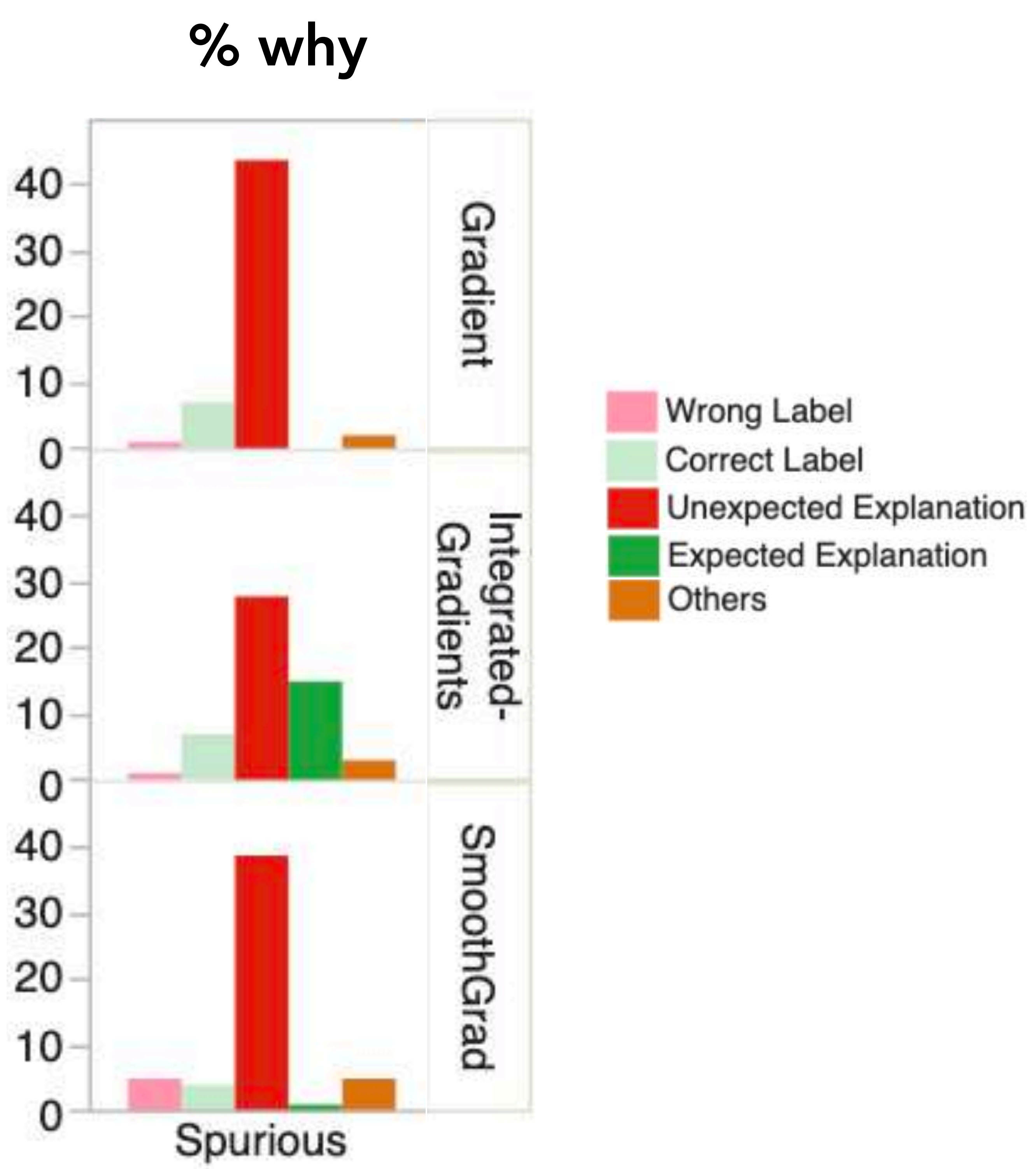
humans, proxy task

Can these methods tell us about

Spurious correlation? maybe!



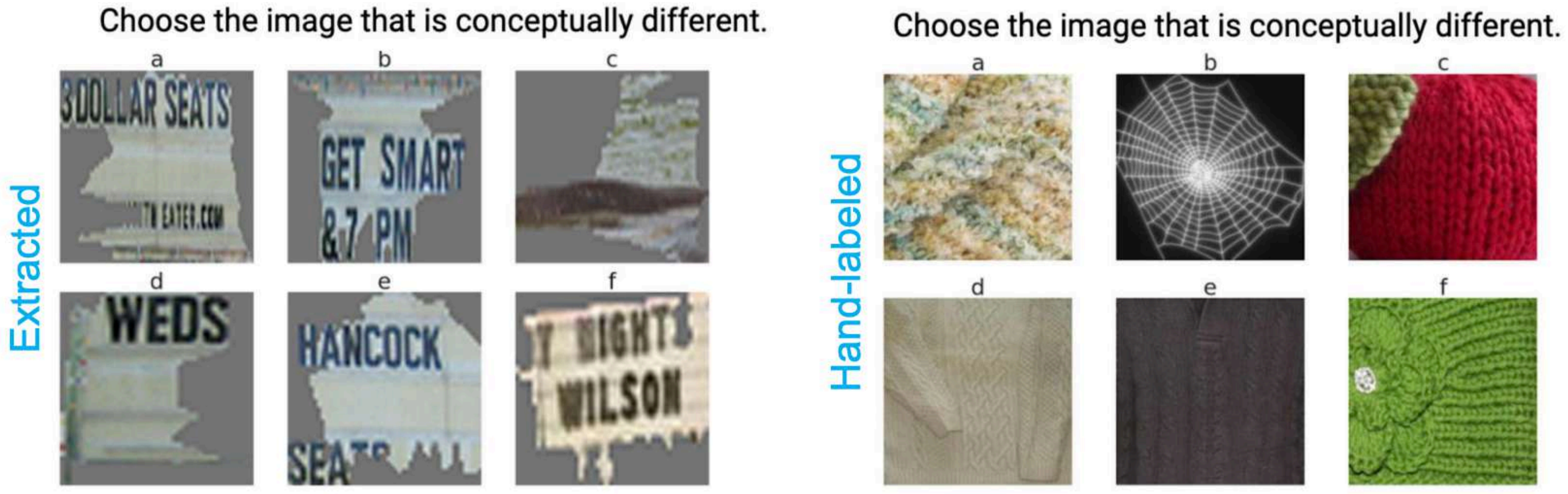




Subjects are uncertain, mostly because of **unexpected** explanations!

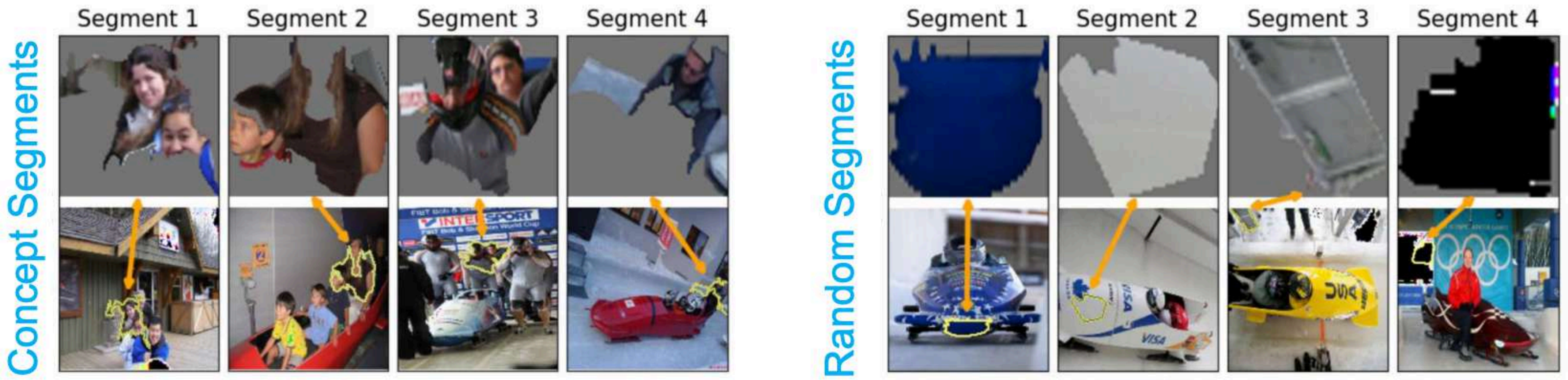
humans, proxy task

Example: Evaluating discovered concepts with subjects



Experiment 1: Identifyig intruder concept

Look at the following two groups of segments. In each group, you should look at the top row. Each image in the top row is a zoomed-in version of another image shown on the bottom row. Now the question is that which of the groups seems more meaningful to you.



Which groups of images is more meaningful to you? ○ right ○ left

If possible please describe the chosen row in one word. Your answer

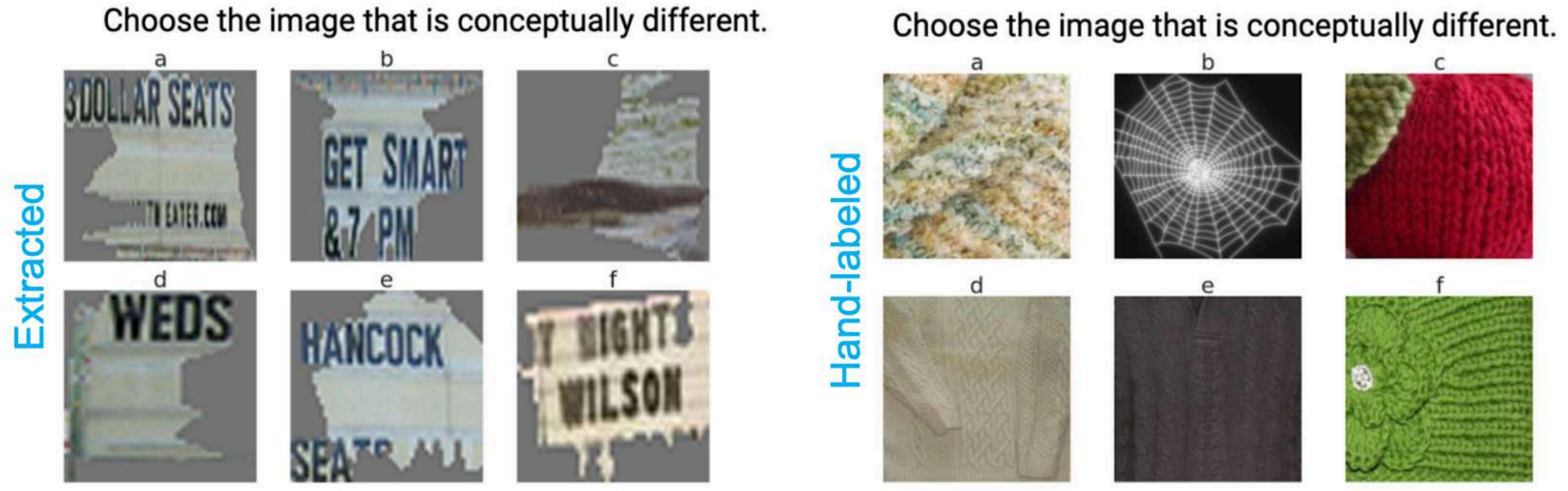
Experiment 2: Identifying the meaning of concept

- Exp1: Intruder test
  - Task: Identify an odd one out
  - Discovered concepts: 99%, similar to hand-labeled dataset, 97%
- Exp2: Meaning test
  - Task: Select between discovered concepts vs random segments and name them.
  - Correctly chosen 95% of time
  - 56% used the same name and 77% named the same or top two terms (e.g., human, face)

Automatic Concept-based Explanations (ACE)
[Ghorbani et al. NeurIPS 19]

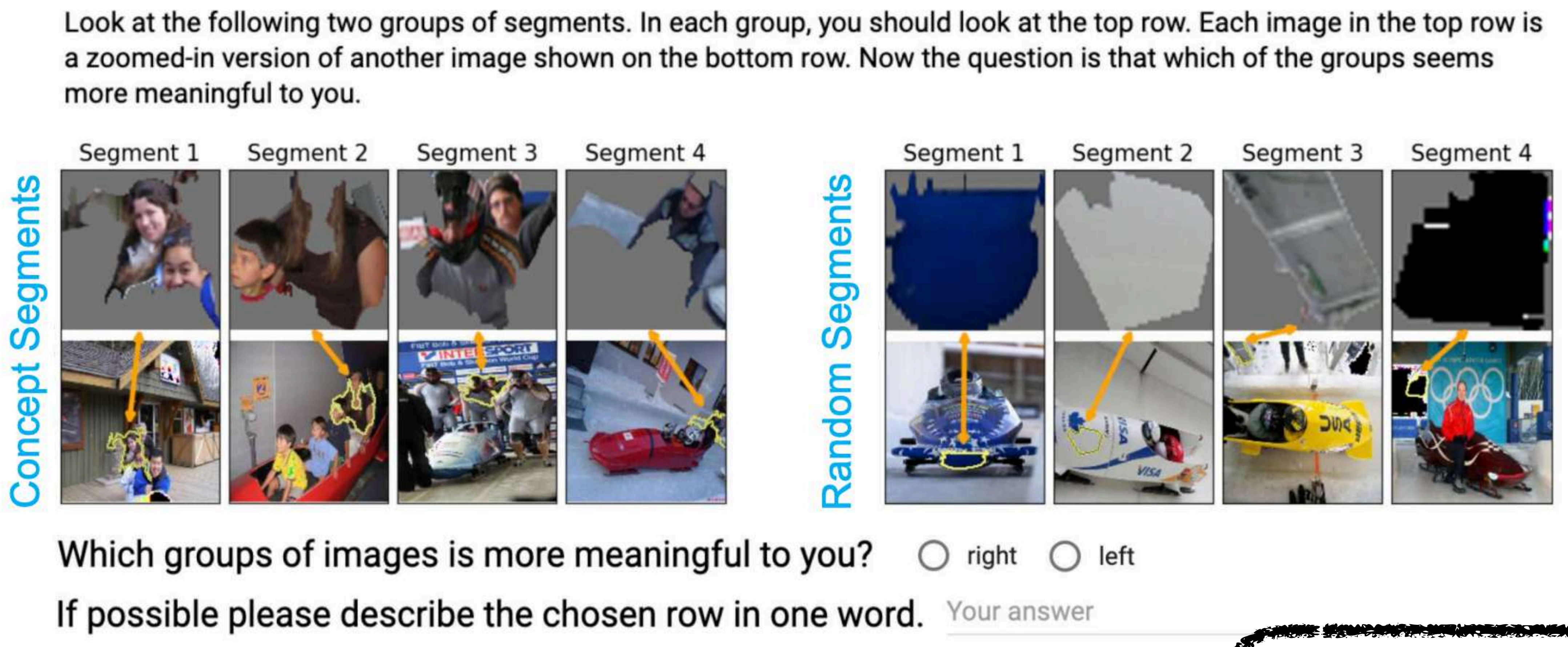
humans, proxy task

Example: Evaluating discovered concepts with subjects


Choose the image that is conceptually different.

Choose the image that is conceptually different.

Experiment 1: Identifyig intruder concept


Look at the following two groups of segments. In each group, you should look at the top row. Each image in the top row is a zoomed-in version of another image shown on the bottom row. Now the question is that which of the groups seems more meaningful to you.

Which groups of images is more meaningful to you? ○ right ○ left

If possible please describe the chosen row in one word. Your answer

Experiment 2: Identifying the meaning of co

- Exp1: Intruder test
  - Task: Identify an odd one out
  - Discovered concepts: 99%, similar to hand-labeled dataset, 97%
- Exp2: Meaning test
  - Task: Select between discovered concepts vs random segments and name them.
  - Correctly chosen 95% of time
  - 56% used the same name and 77% named the same or top two terms

Take away:

Proxy task can be an effective way to evaluate a method (often) before running real experts on real tasks.

With humans on real tasks

“Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making”

[Cai et al. 19 CHI]

“In two evaluations with pathologists, we found that these refinement tools increased the diagnostic utility of images found and increased user trust in the algorithm. The tools were preferred over a traditional interface, without a loss in diagnostic accuracy.”

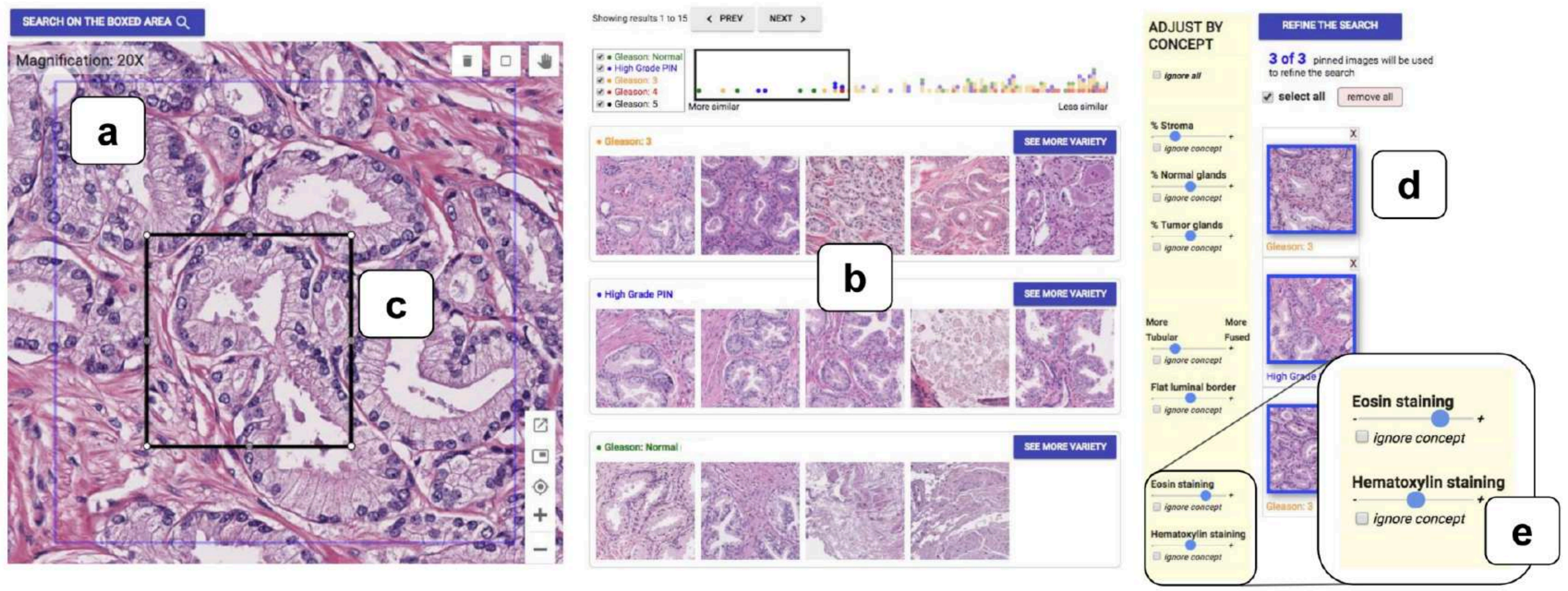


Figure 2: Key components of SMILY: a) the query image (e.g. pathology tissue possibly containing cancer), b) the search results (images from previously diagnosed cases), c) refine-by-region tool: users crop a region to emphasize its importance, d) refine-by-example tool: for clinical concepts that can’t be pinpointed to a specific region (e.g. visual patterns), users can pin examples from search results to emphasize that concept, e) refine-by-concept tool: users increase or decrease the presence of clinical concepts by sliding sliders.

“Explainable machine-learning predictions for the prevention of hypoxaemia during surgery”

[Lundberg et al. 18 Nature biomedical engineering]

“The system, which was trained on minute-by-minute data from the electronic medical records of over 50,000 surgeries, improved the performance of anesthesiologists by providing interpretable hypoxaemia risks and contributing factors.”

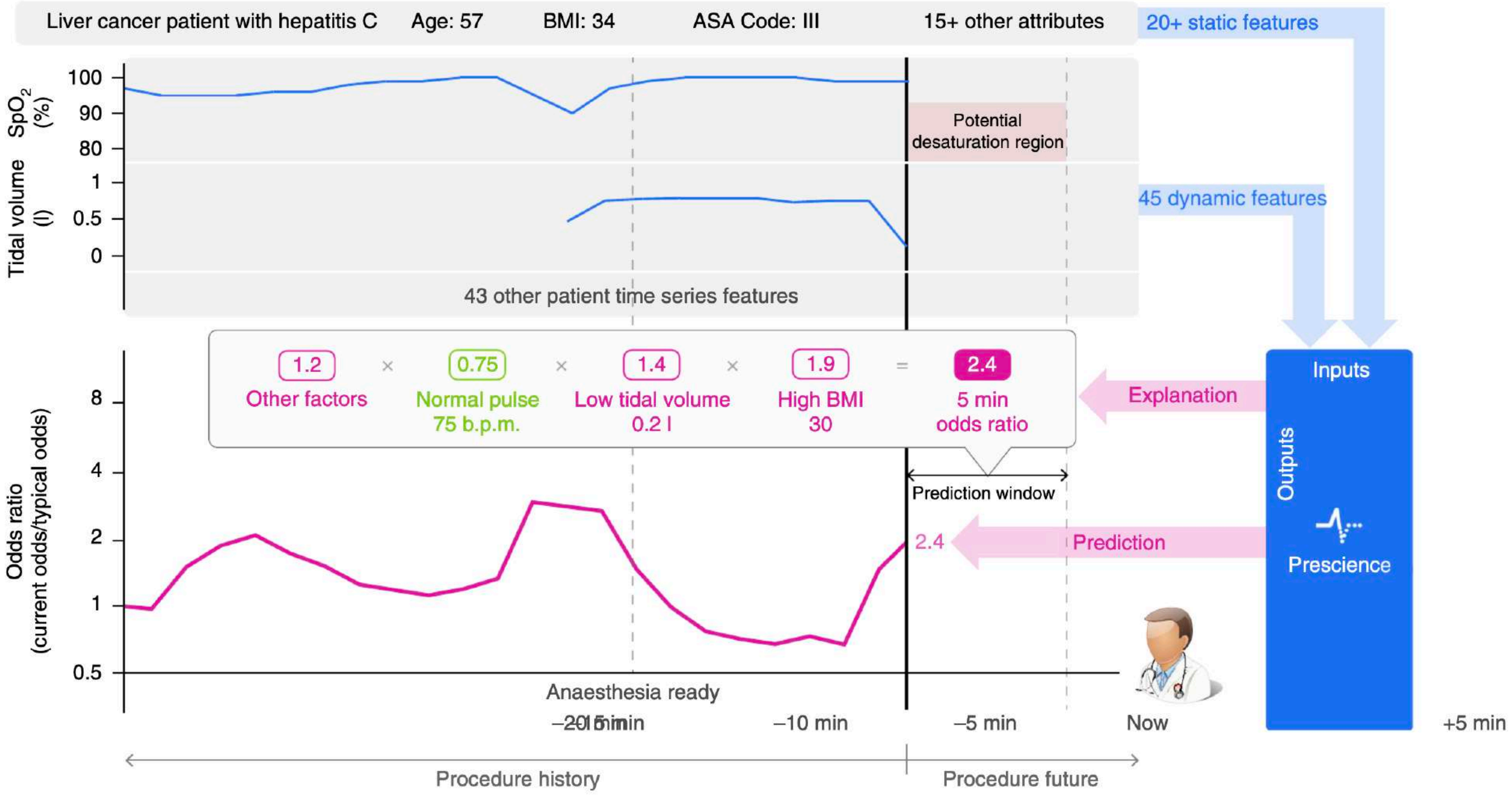


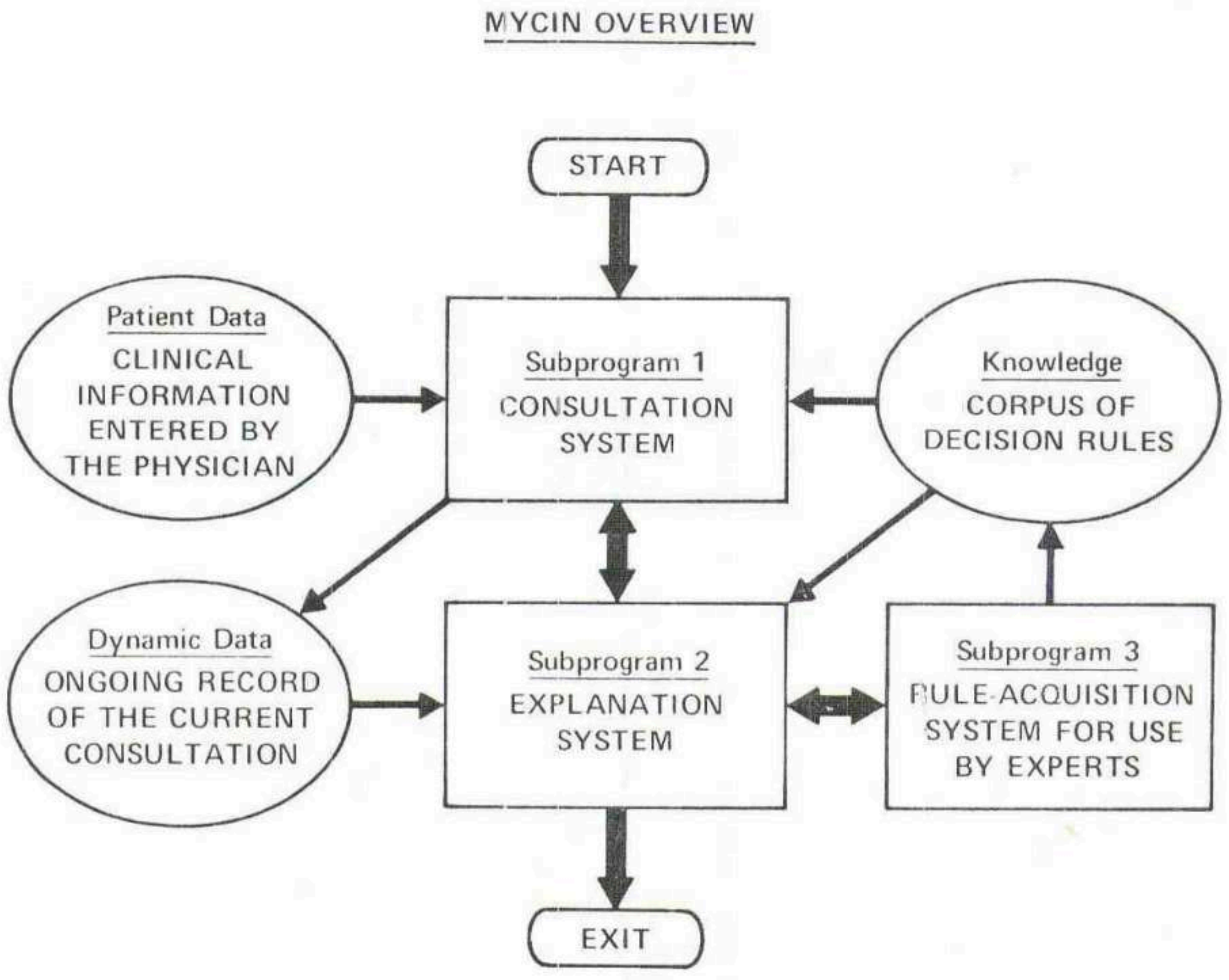
Fig. 1 | Prescience integrates many data sources into a single risk, which is explained through a succinct visual summary. A wide variety of data sources were used to build a predictive model of hypoxaemia events. An explanation (overlaid) is then built for each prediction. Pink features have values that increased risk, whereas green features decreased hypoxaemia risk. The combination of the impacts of all features is the predicted Prescience risk; in this case, the odds are 2.4× higher than normal. Each feature impact value represents the change in risk when the value of that feature is known versus unknown. Qualitative terms such as ‘low’ or ‘high’ are based on the distribution of a feature value in our dataset.

Agenda

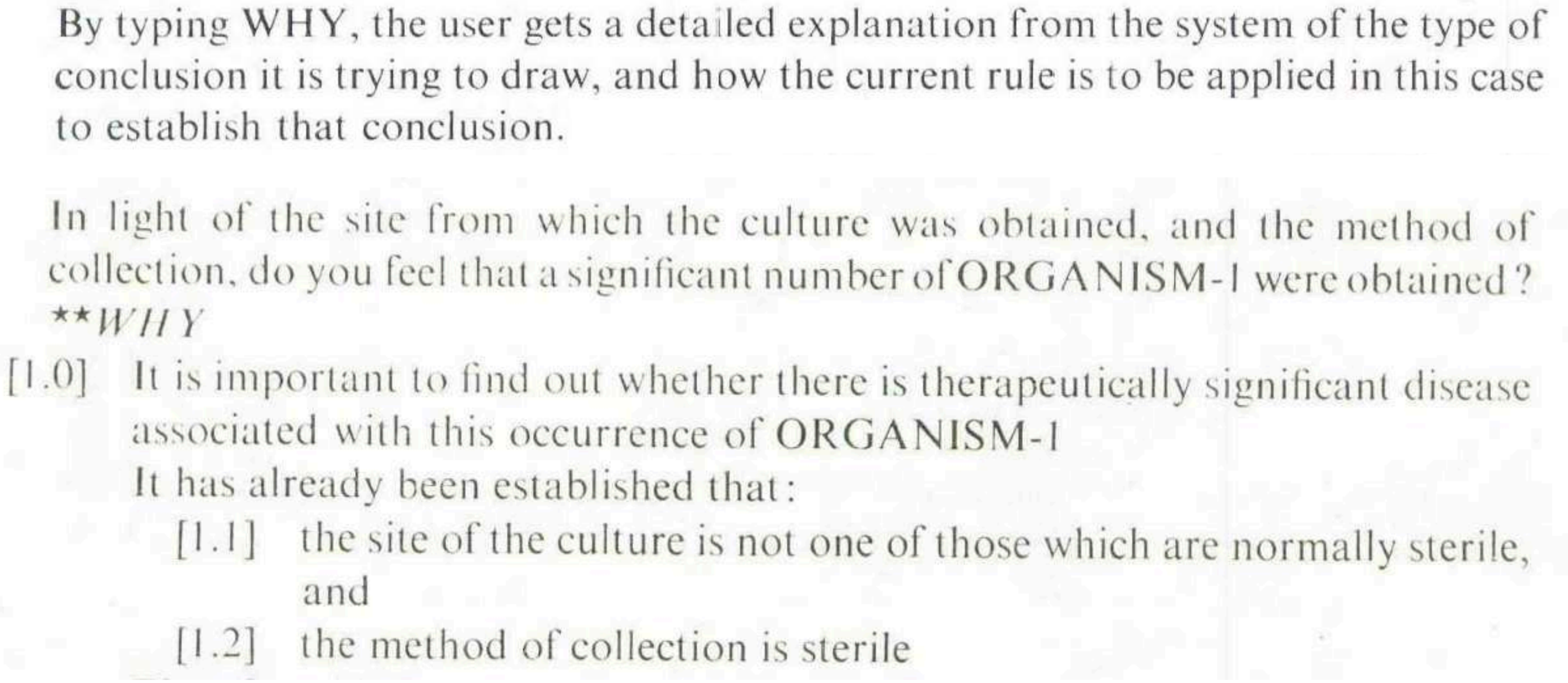
- What and why
- !Caution!: Things to be careful when using and developing interpretability methods
- Evaluate: How to evaluate interpretability methods
- Methods: 3 types of methods and examples

Again, it's not all about deep learning.
Interpretability isn't a new problem.

- Lots of pre-deep learning work going all the way back to 70's.



[Shortliffe et al. 1975]



By typing WHY, the user gets a detailed explanation from the system of the type of conclusion it is trying to draw, and how the current rule is to be applied in this case to establish that conclusion.

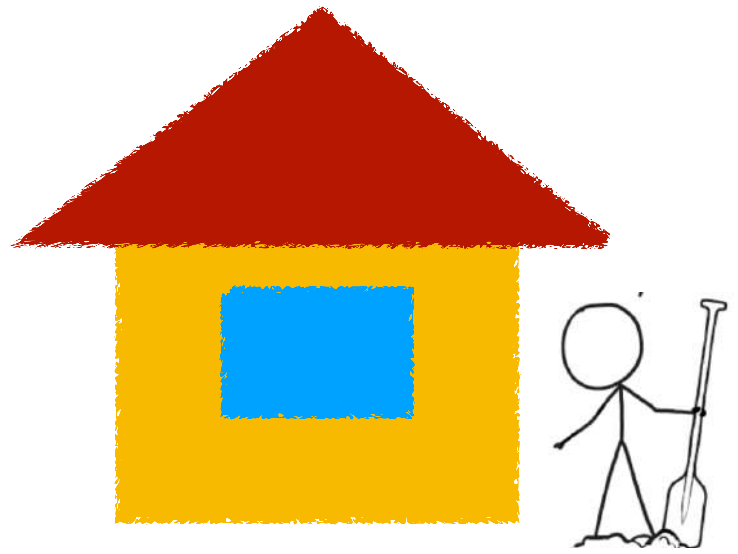
In light of the site from which the culture was obtained, and the method of collection, do you feel that a significant number of ORGANISM-1 were obtained?
*****WHY***
[1.0] It is important to find out whether there is therapeutically significant disease associated with this occurrence of ORGANISM-1
It has already been established that:
[1.1] the site of the culture is not one of those which are normally sterile, and
[1.2] the method of collection is sterile

Good reference

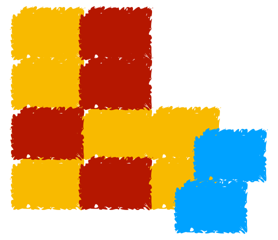
Comprehensible Classification Models – a position paper

Alex A. Freitas
School of Computing
University of Kent
Canterbury, CT2 7NF, UK

Types of interpretability methods



My ML



Explaining data

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

Building inherently interpretable model

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

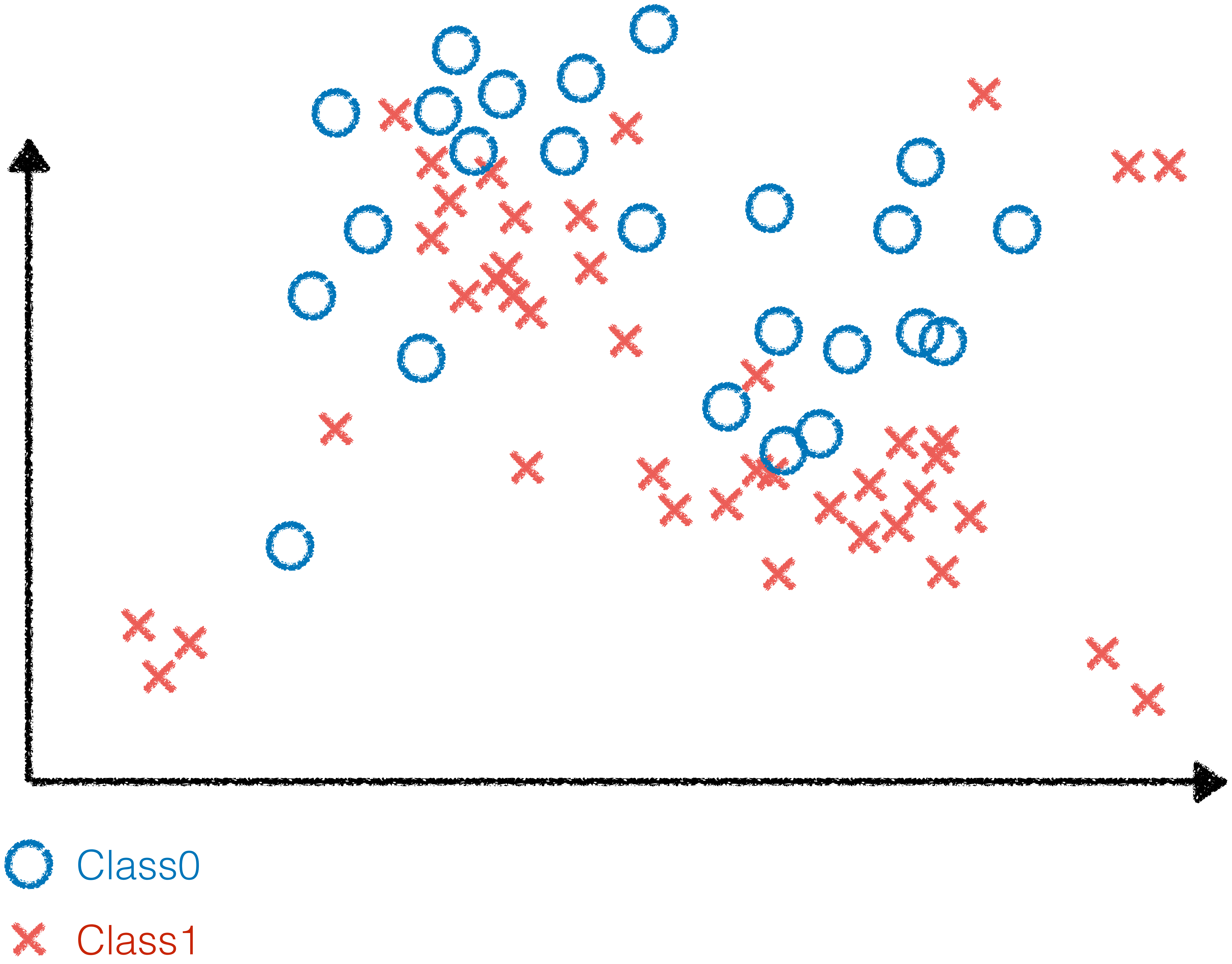
Post-training interpretability methods

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

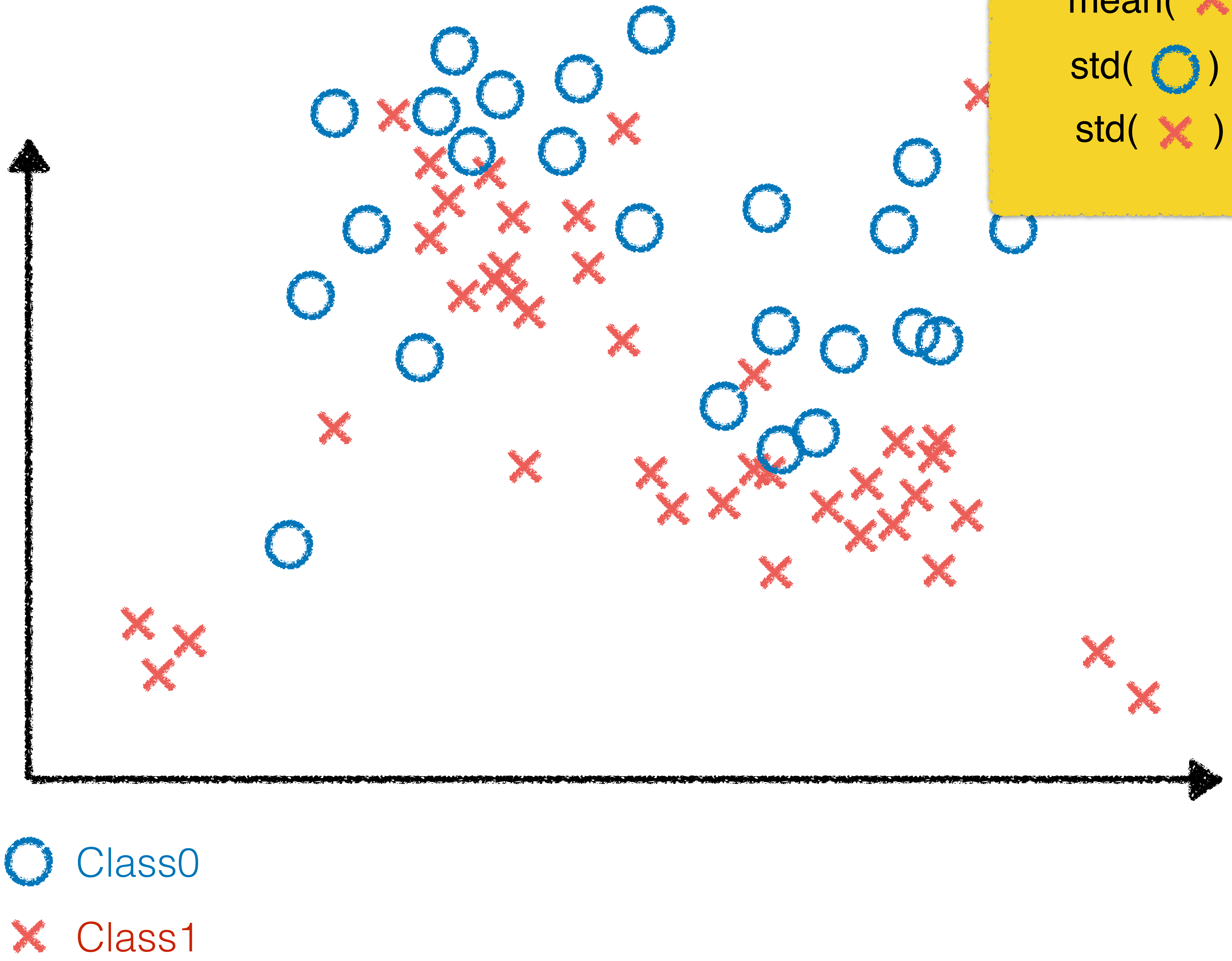
Interpreting data (not just the model) is important.

- Exploratory data analysis:
    - “an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. [It] is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. "

https://en.wikipedia.org/wiki/Exploratory_data_analysis


Class0
Class1

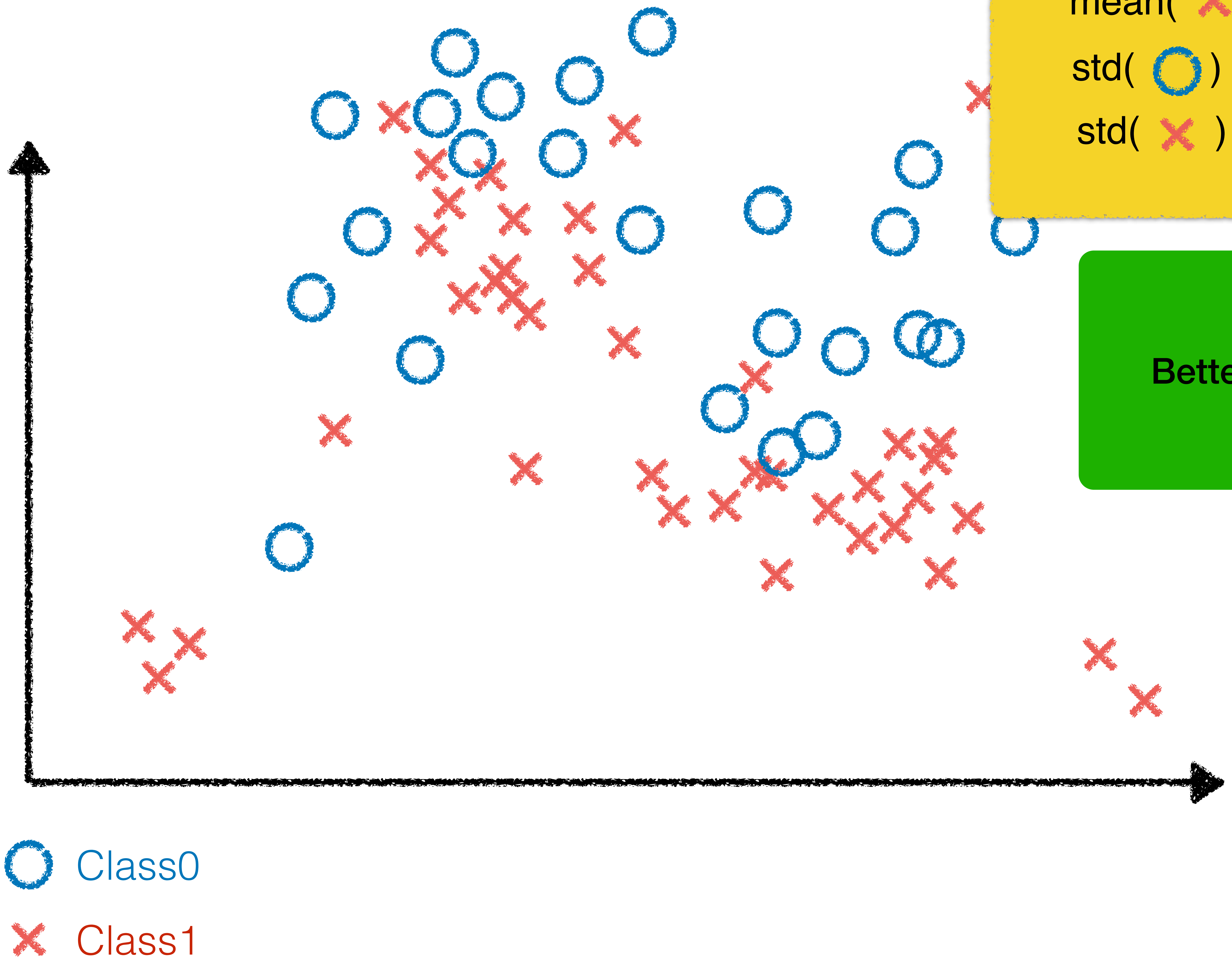
Before building any model



Descriptive statistics

mean(○)
mean(×)
std(○)
std(×)

Before building any model

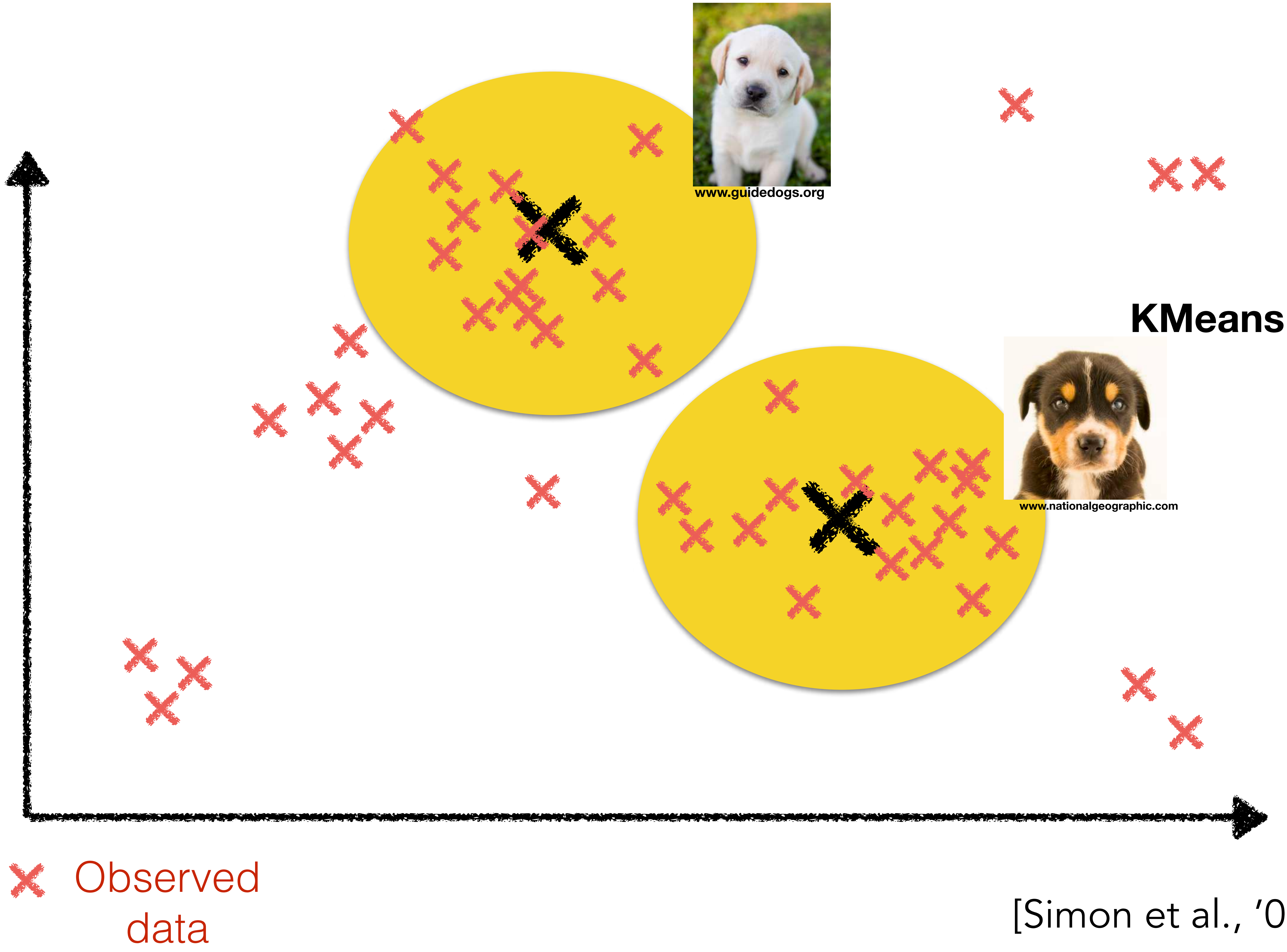


Descriptive statistics

mean(○)
mean(×)
std(○)
std(×)

Better way?

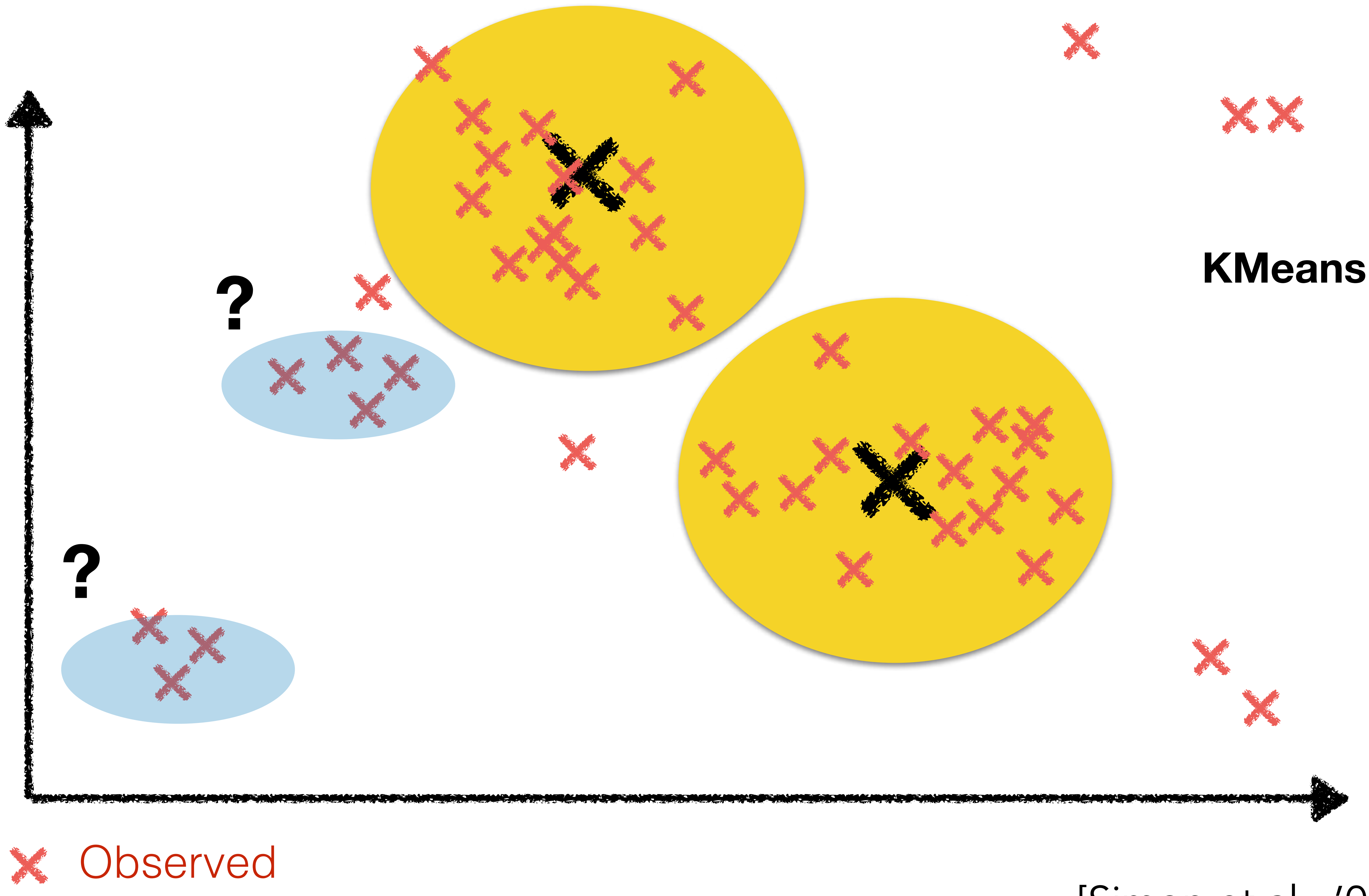
Exploratory data analysis



KMeans, KNN

[Simon et al., ’07]
[Lin and Bilmes, ’11]

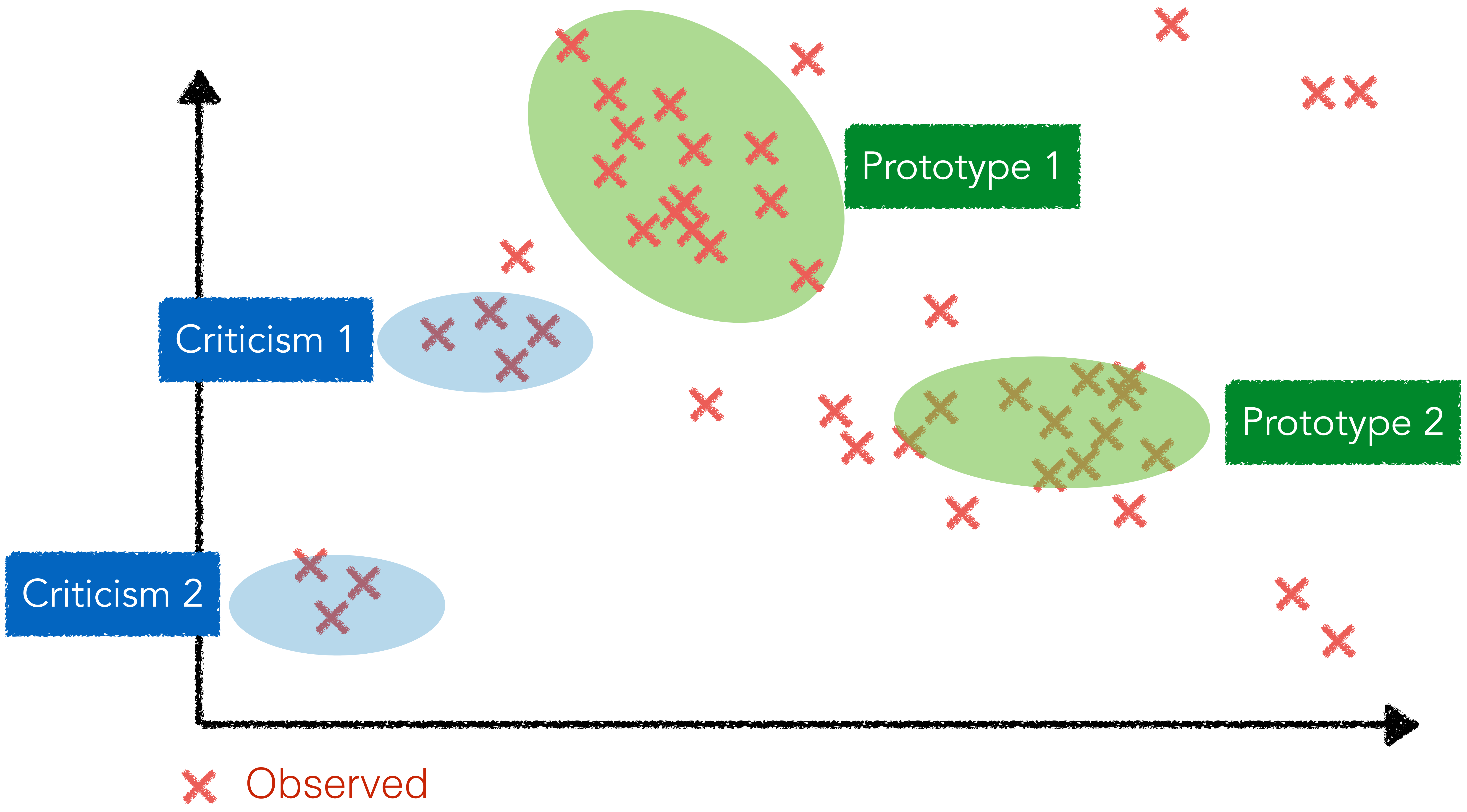
Exploratory data analysis



KMeans, KNN

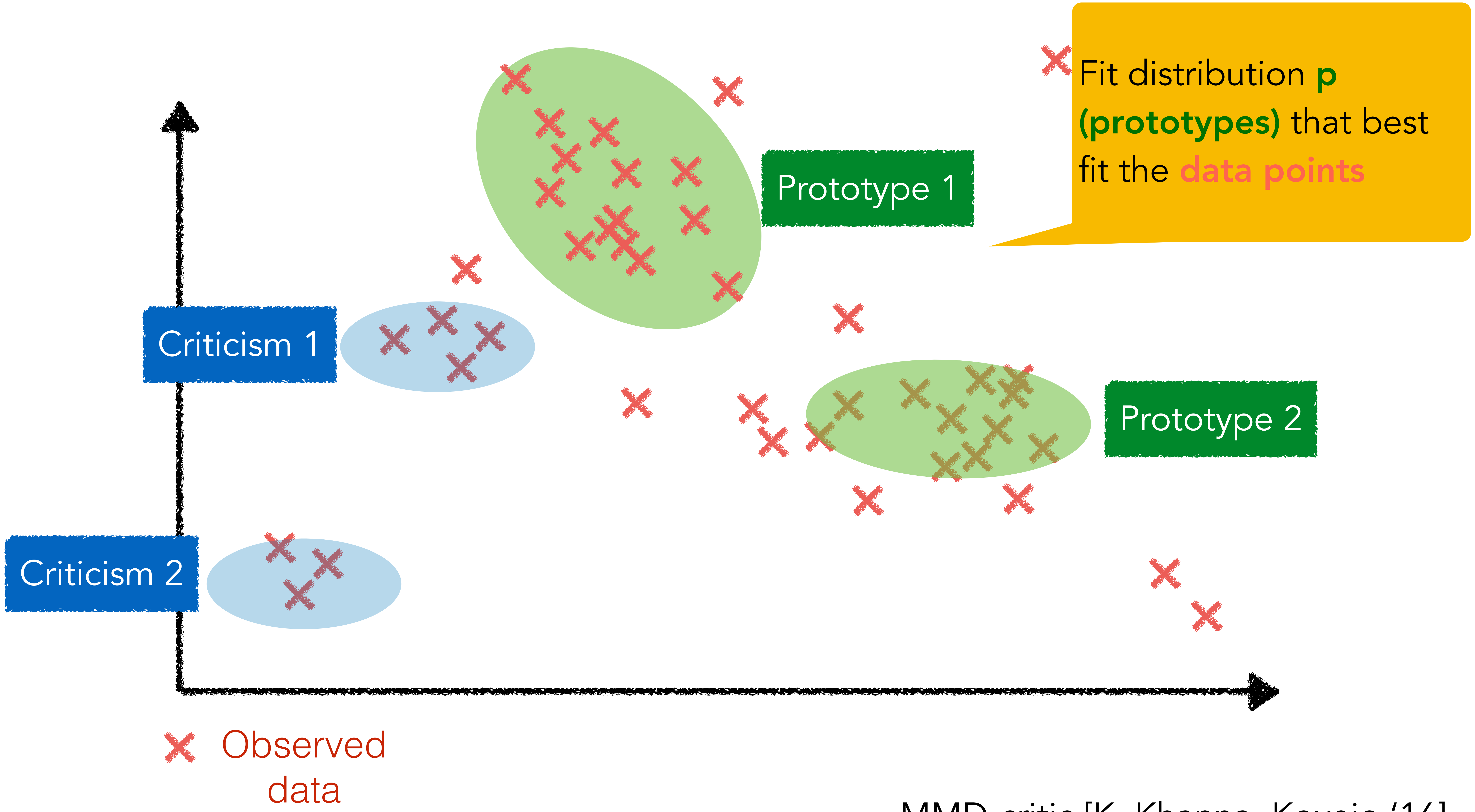
[Simon et al., '07]
[Lin and Bilmes, '11]

Exploratory data analysis



MMD-critic [K. Khanna, Koyejo '16]

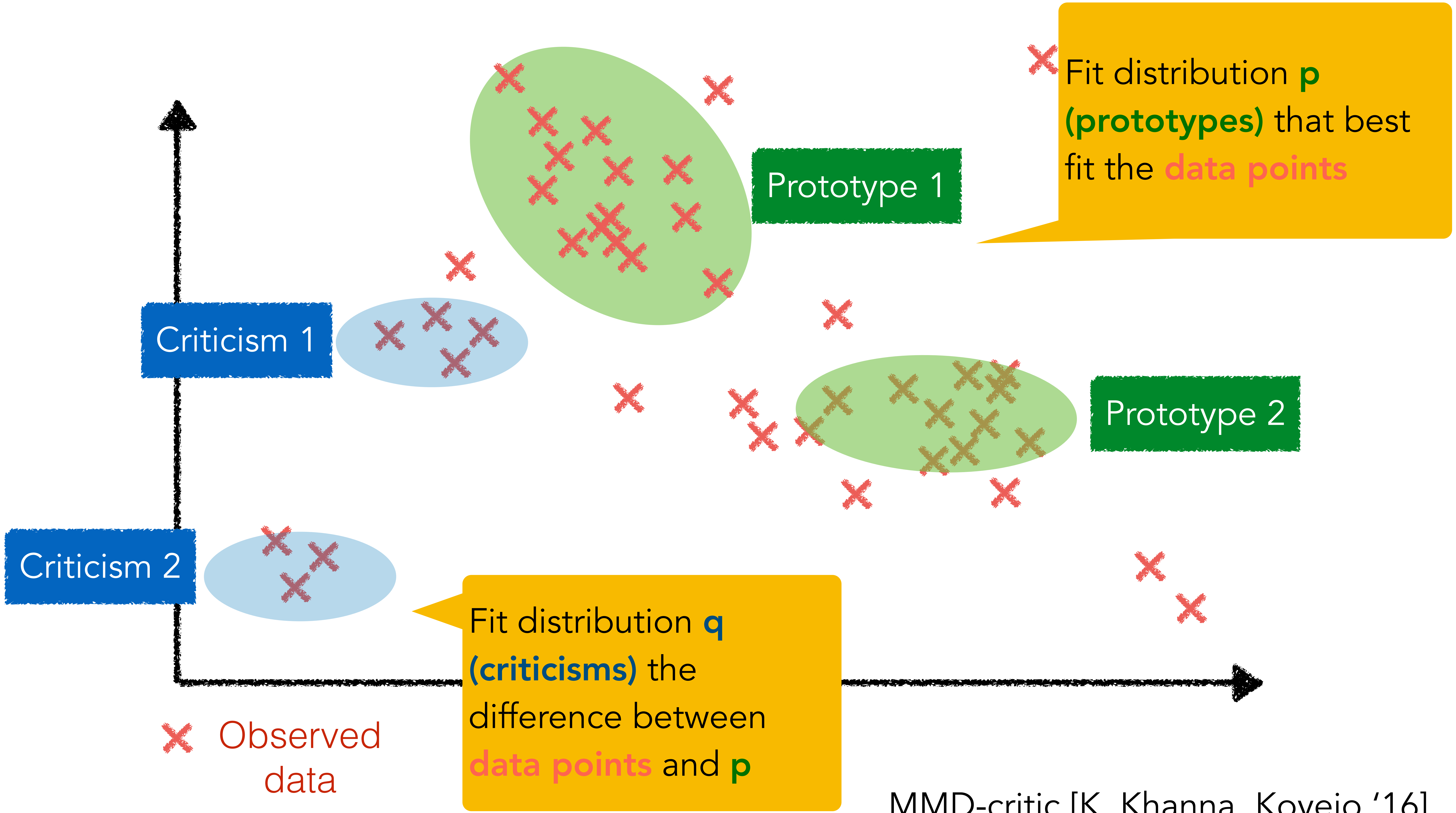
Exploratory data analysis



Fit distribution **p (prototypes)** that best fit the **data points**

MMD-critic [K. Khanna, Koyejo '16]

Exploratory data analysis

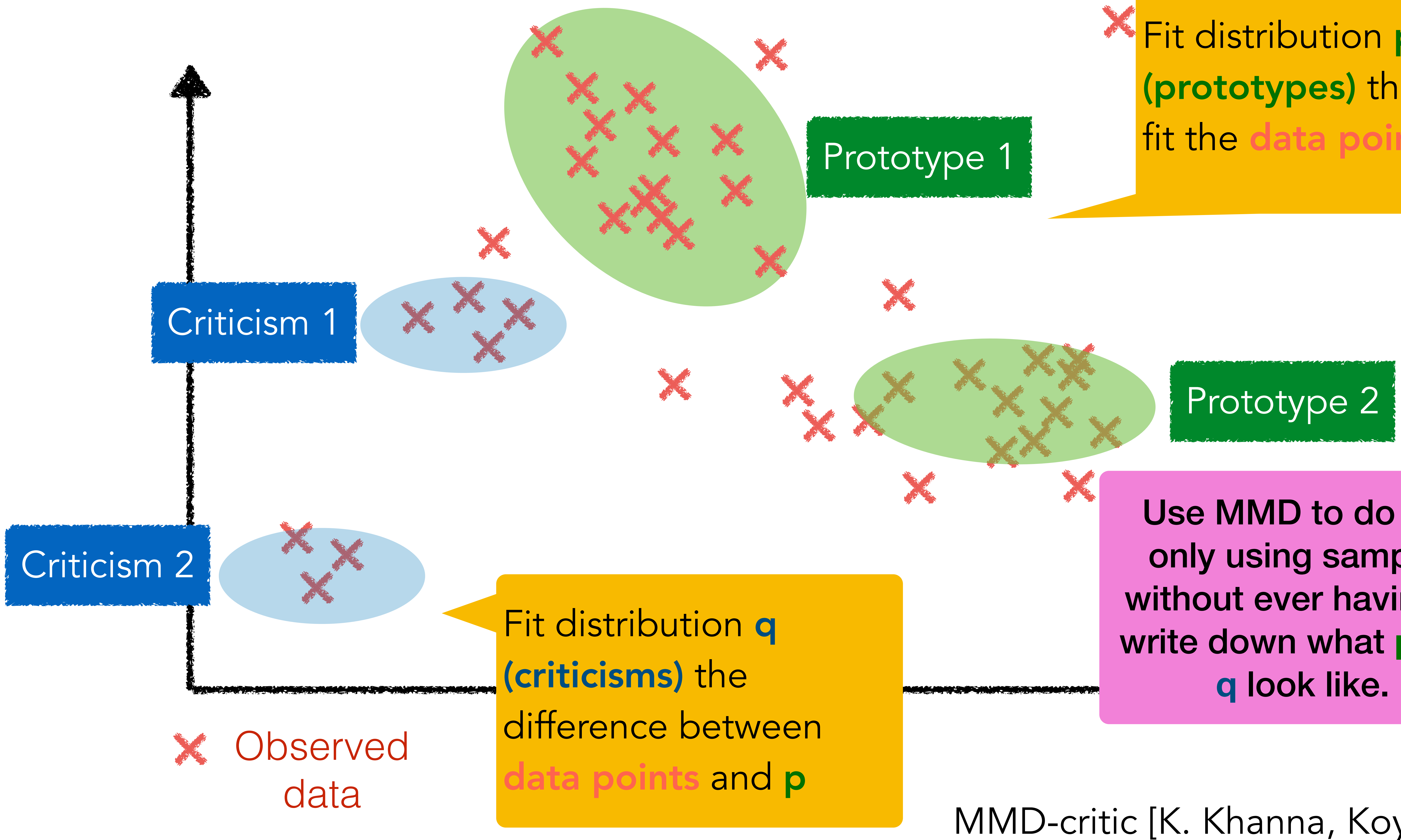


Fit distribution **p (prototypes)** that best fit the **data points**

Fit distribution **q (criticisms)** the difference between **data points** and **p**

MMD-critic [K. Khanna, Koyejo '16]

Exploratory data analysis



Fit distribution **p** **(prototypes)** that best fit the **data points**

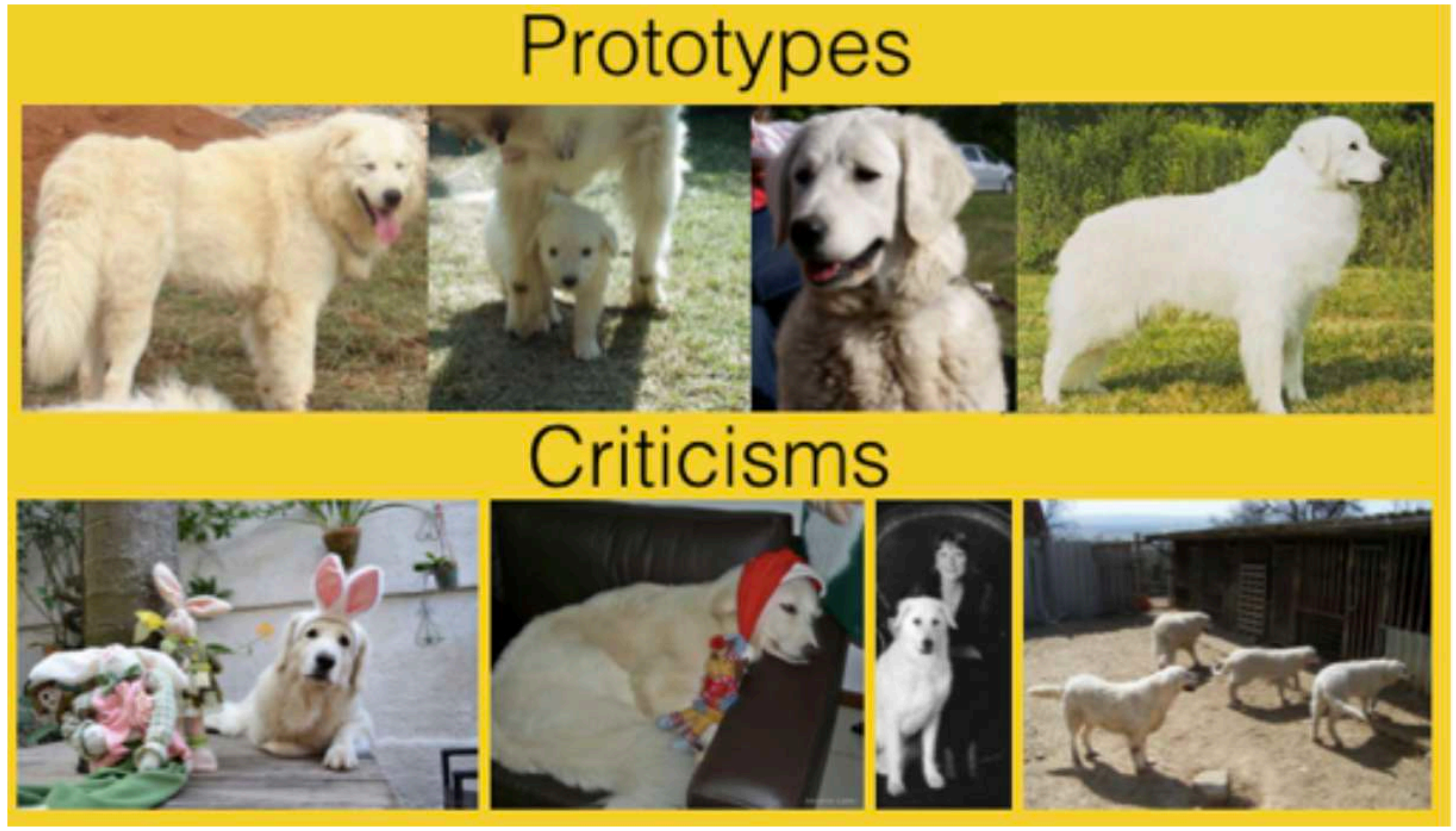
Fit distribution **q** **(criticisms)** the difference between **data points** and **p**

Use MMD to do this only using samples without ever having to write down what p and q look like.

MMD-critic [K. Khanna, Koyejo '16]

Before building any model

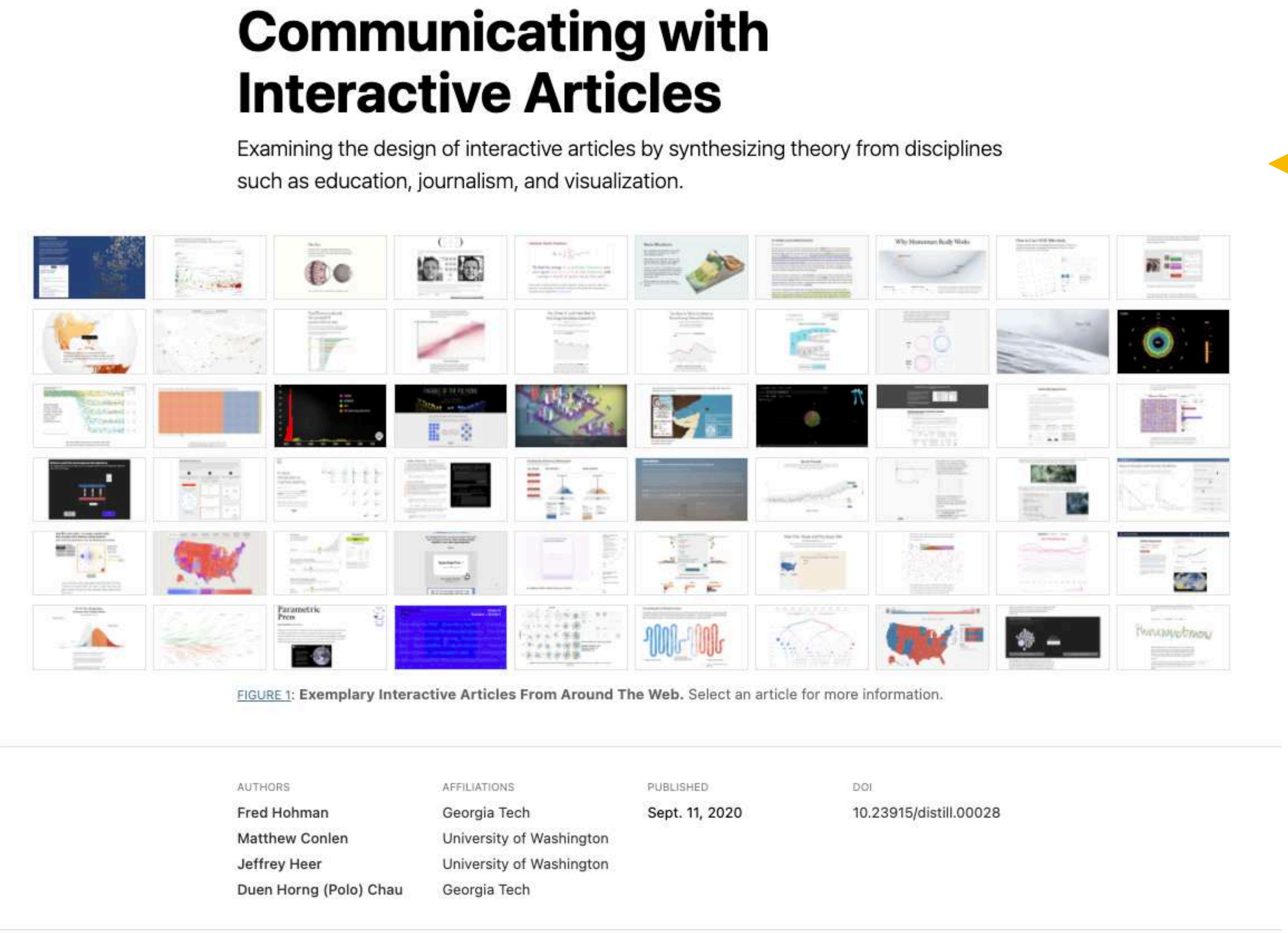
Exploratory data analysis



MMD-critic [K. Khanna, Koyejo '16]

Before building any model

Exploratory data analysis



Communicating with Interactive Articles

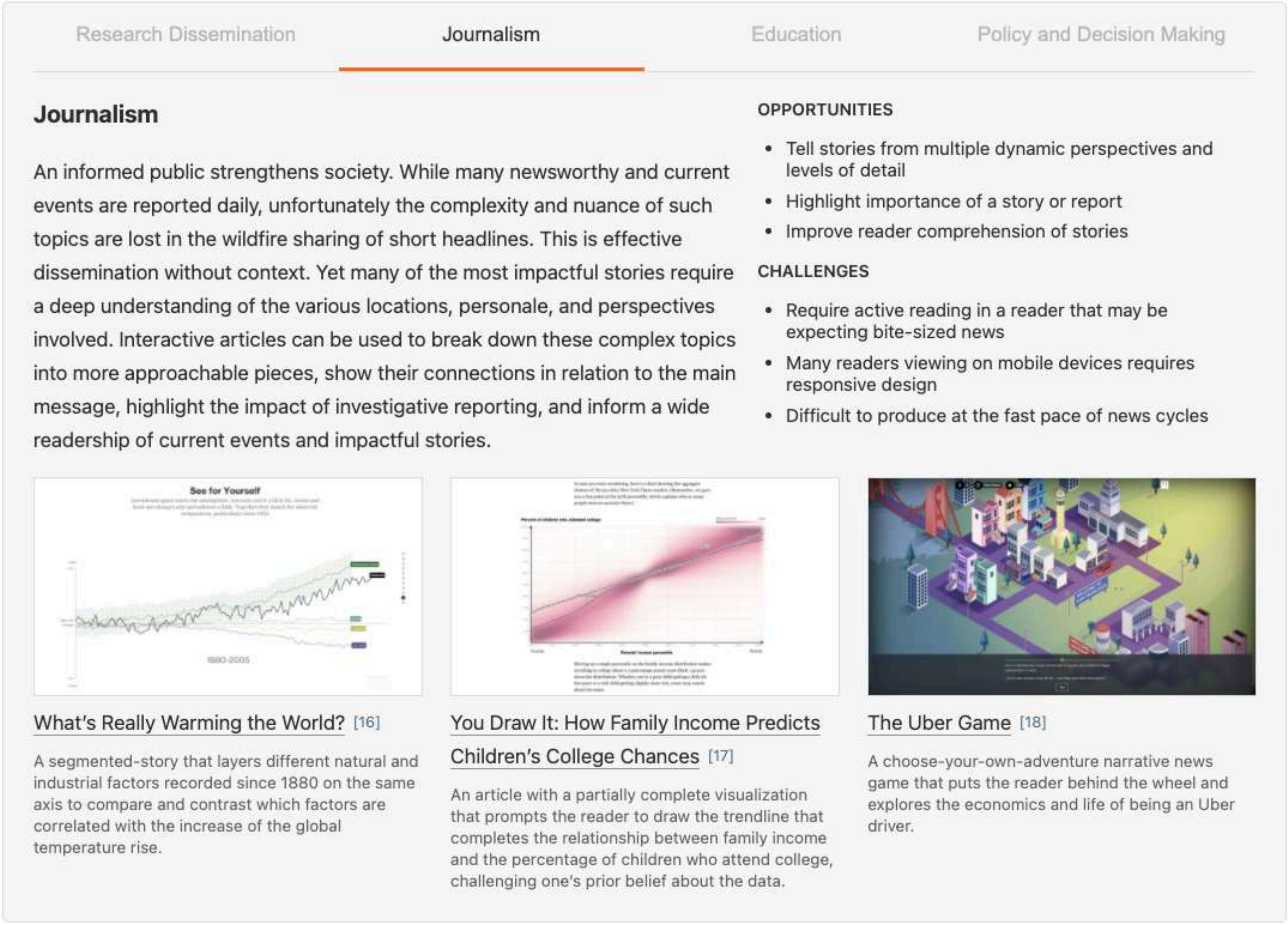
Examining the design of interactive articles by synthesizing theory from disciplines such as education, journalism, and visualization.

FIGURE 1: **Exemplary Interactive Articles From Around The Web.** Select an article for more information.

AUTHORS	AFFILIATIONS	PUBLISHED	DOI
Fred Hohman	Georgia Tech	Sept. 11, 2020	10.23915/distill.00028
Matthew Conlen	University of Washington		
Jeffrey Heer	University of Washington		
Duen Horng (Polo) Chau	Georgia Tech		

Great overview of many exploratory data analysis, highly recommend.

https://distill.pub/2020/communicating-with-interactive-articles/



Research Dissemination | Journalism | Education | Policy and Decision Making

Journalism

An informed public strengthens society. While many newsworthy and current events are reported daily, unfortunately the complexity and nuance of such topics are lost in the wildfire sharing of short headlines. This is effective dissemination without context. Yet many of the most impactful stories require a deep understanding of the various locations, personale, and perspectives involved. Interactive articles can be used to break down these complex topics into more approachable pieces, show their connections in relation to the main message, highlight the impact of investigative reporting, and inform a wide readership of current events and impactful stories.

OPPORTUNITIES

- Tell stories from multiple dynamic perspectives and levels of detail
- Highlight importance of a story or report
- Improve reader comprehension of stories

CHALLENGES

- Require active reading in a reader that may be expecting bite-sized news
- Many readers viewing on mobile devices requires responsive design
- Difficult to produce at the fast pace of news cycles

What's Really Warming the World? [16]

A segmented-story that layers different natural and industrial factors recorded since 1880 on the same axis to compare and contrast which factors are correlated with the increase of the global temperature rise.

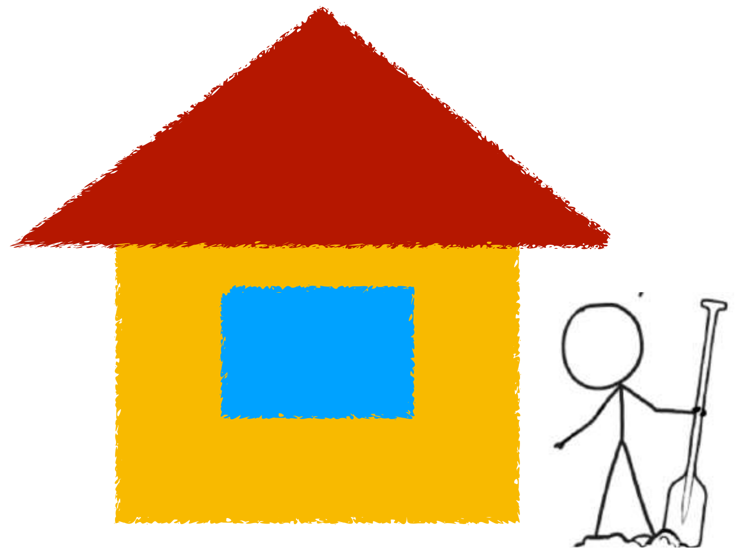
You Draw It: How Family Income Predicts Children's College Chances [17]

An article with a partially complete visualization that prompts the reader to draw the trendline that completes the relationship between family income and the percentage of children who attend college, challenging one's prior belief about the data.

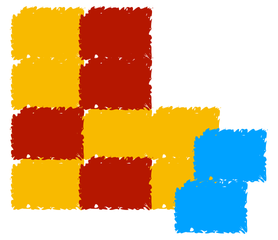
The Uber Game [18]

A choose-your-own-adventure narrative news game that puts the reader behind the wheel and explores the economics and life of being an Uber driver.

Types of interpretability methods



My ML



Explaining data

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

Building inherently interpretable model

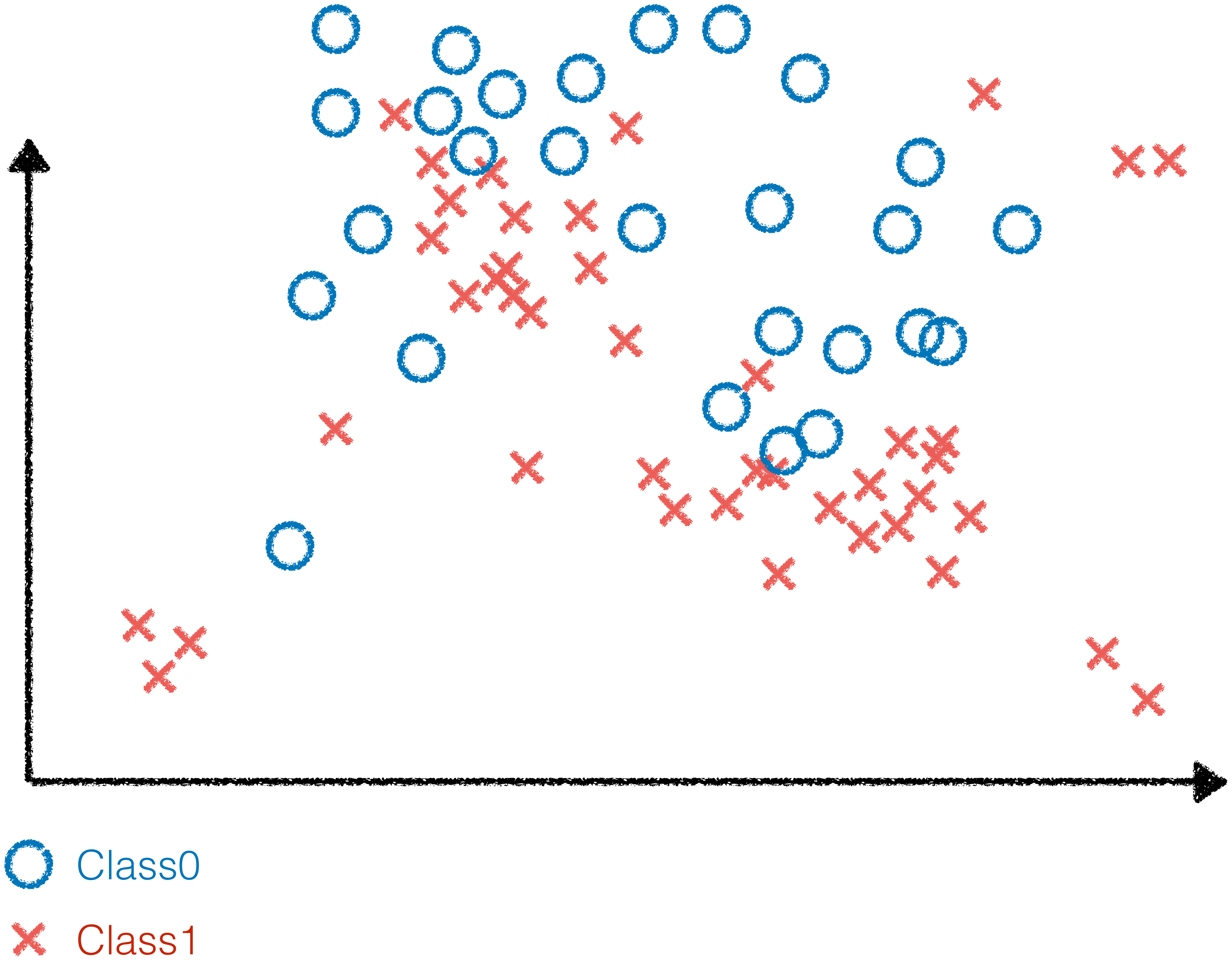
$$\underset{E,M}{\text{argmax}}\ Q(\mathbf{E}\text{xplanation}, \mathbf{M}\text{odel}|\mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Model == explanation

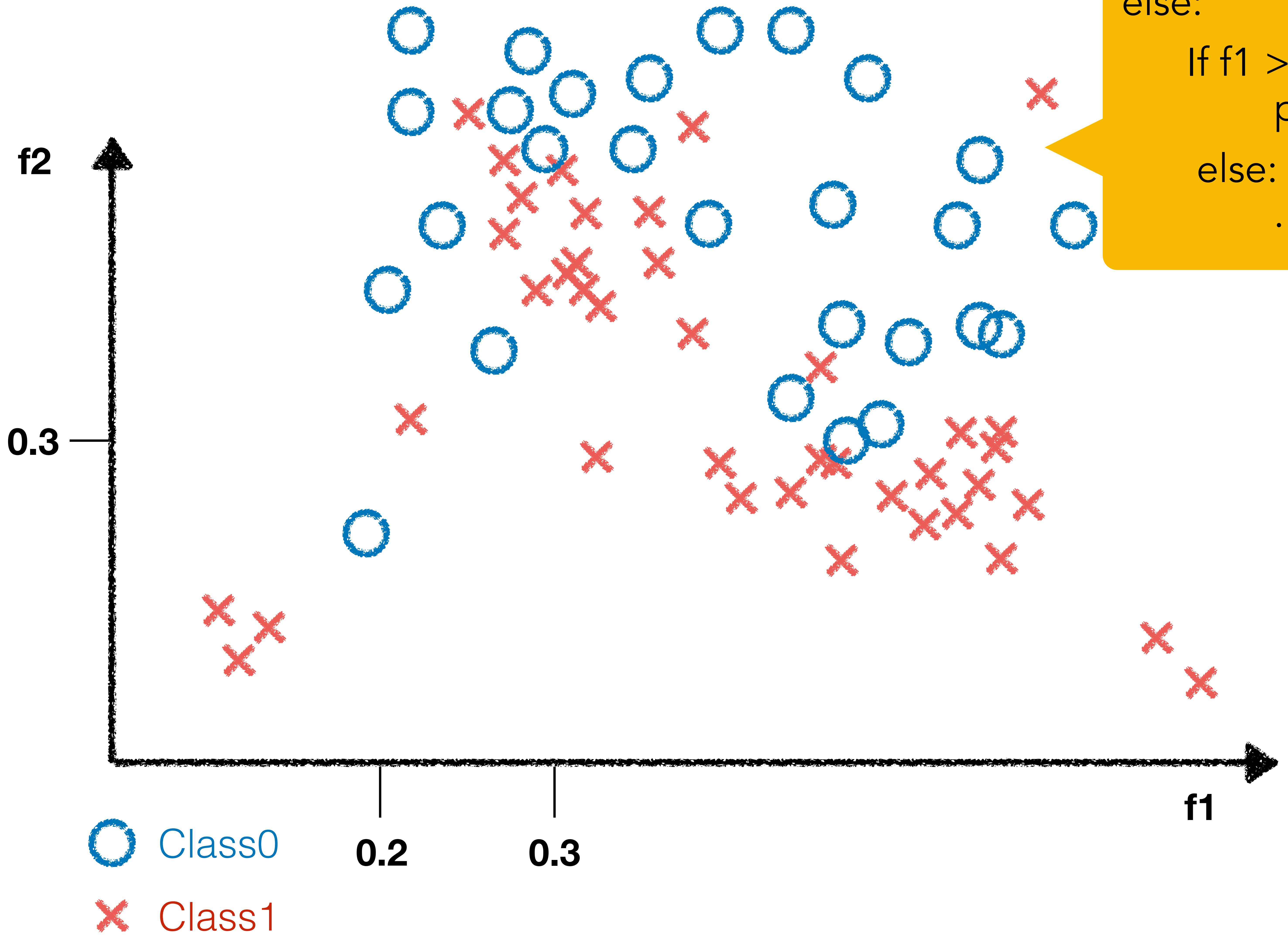
No approximation needed

Post-training interpretability methods

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


Class0
Class1

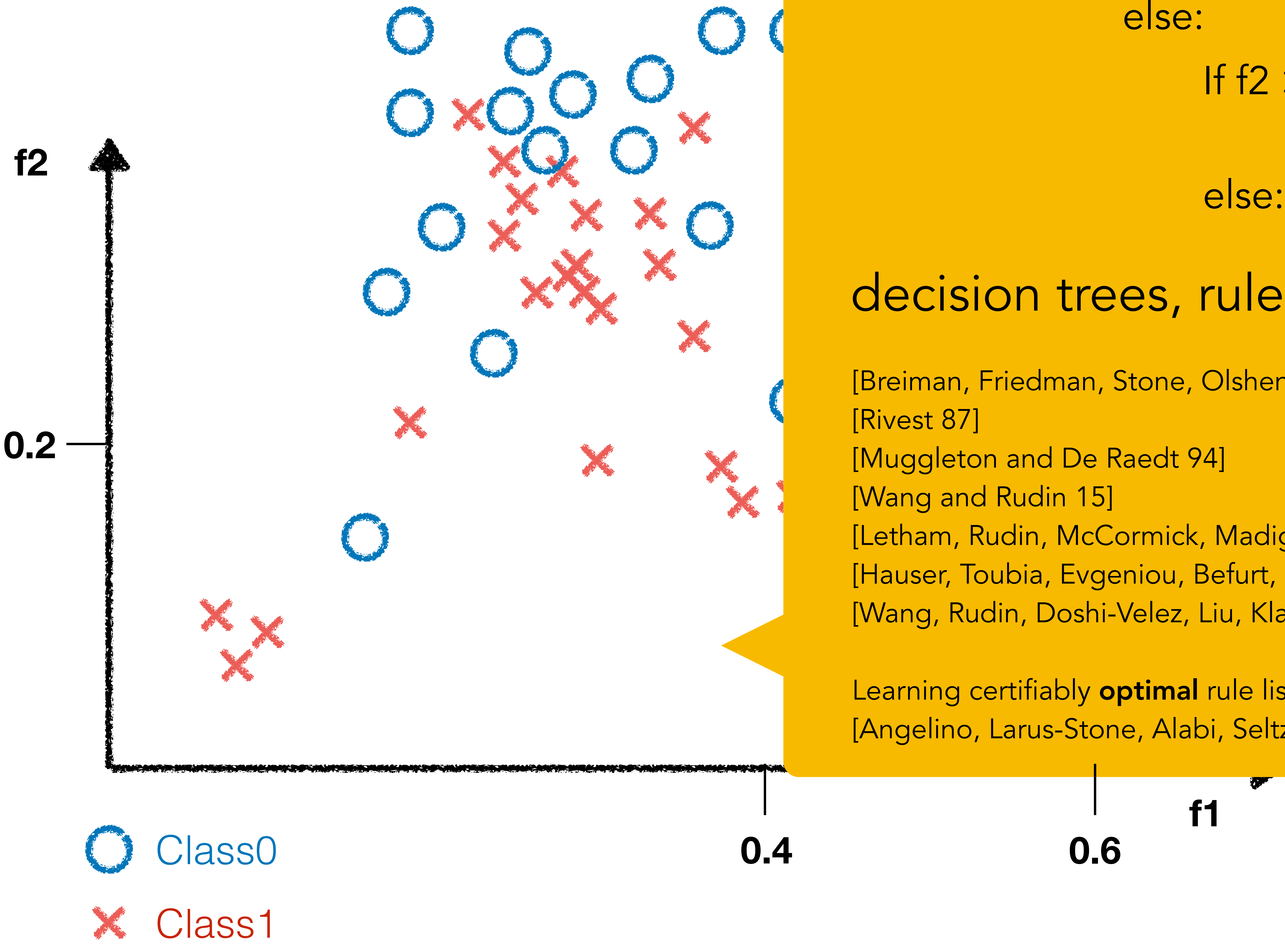
Building a new model



Rule based

```
If f2 < 0.3:
    predict ✗
else:
    If f1 > 0.2 and f1 < 0.3:
        predict ○
    else:
        ...
```

Building a new model



Rule based

```
If f1 < 0.1:
    predict ×
else:
    If f2 > 0.4 and f2 < 0.6:
        predict ○
    else:
        ...
```

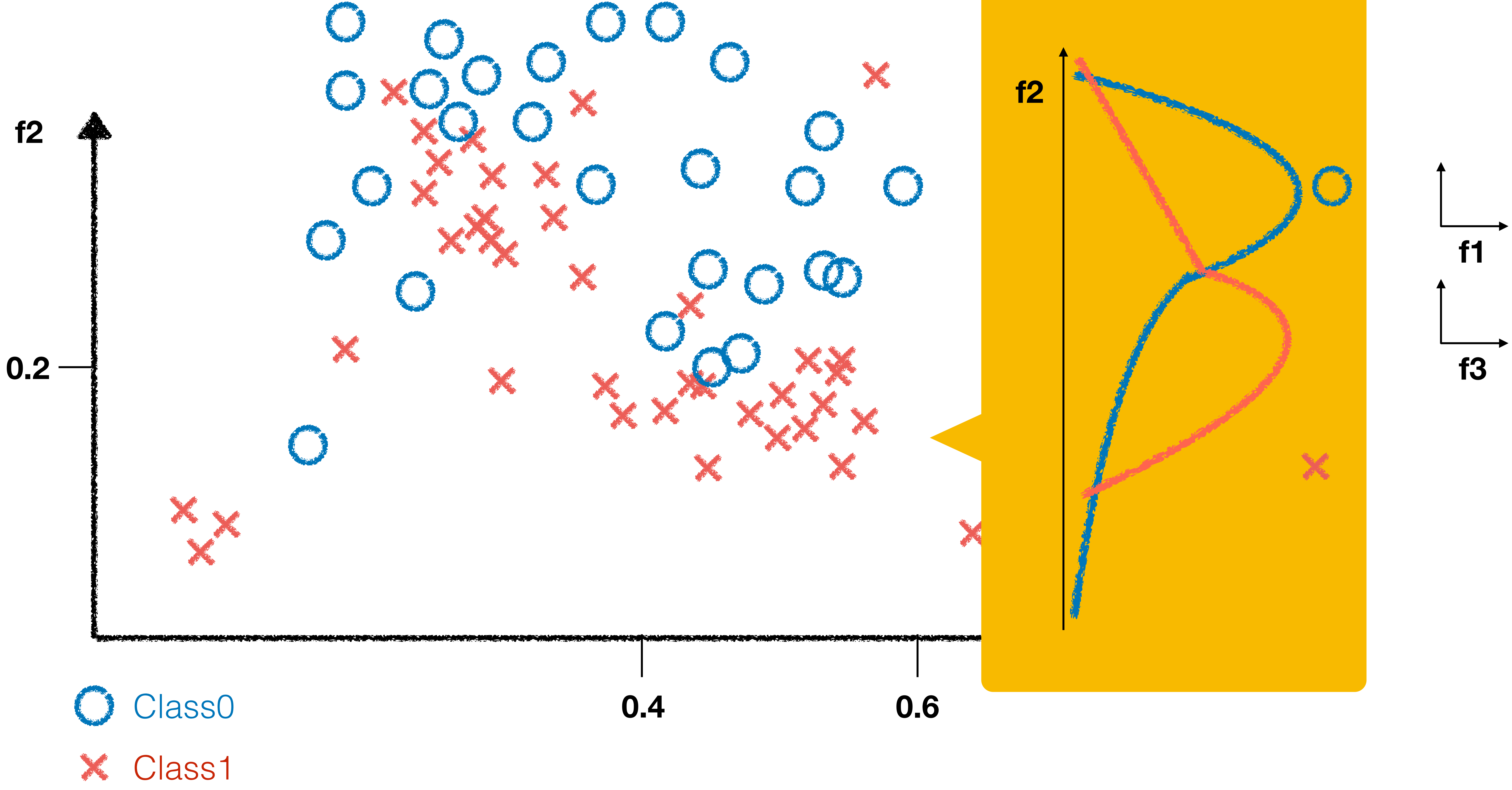
decision trees, rule lists, rule sets

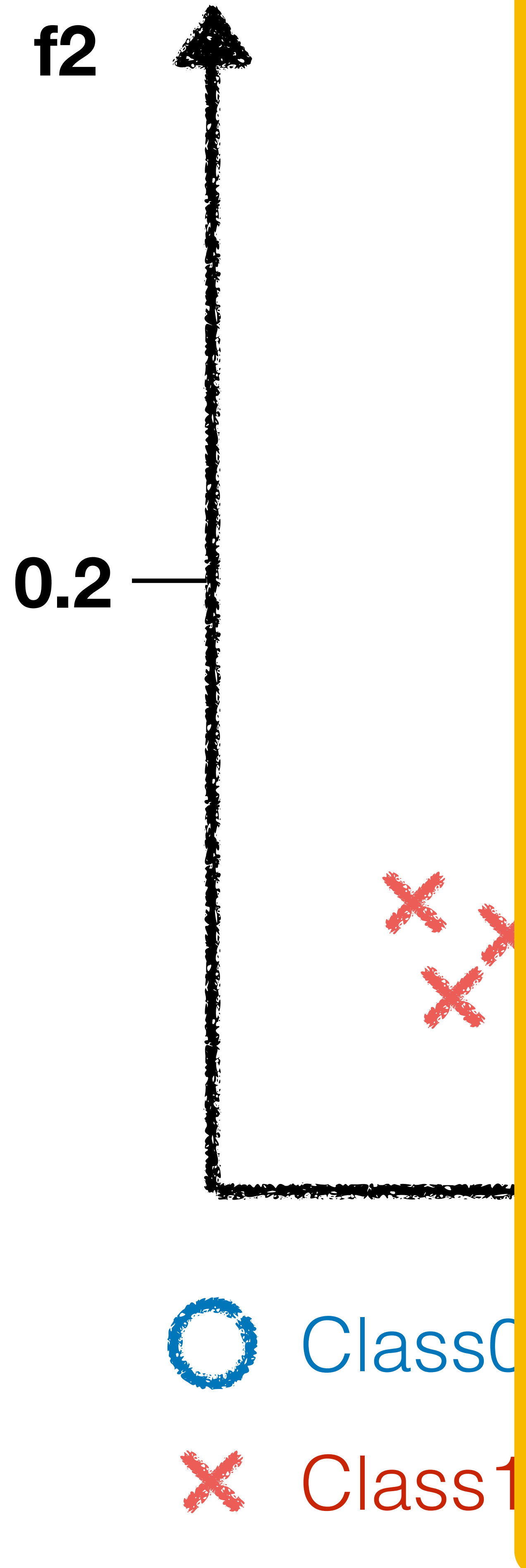
[Breiman, Friedman, Stone, Olshen 84]
[Rivest 87]
[Muggleton and De Raedt 94]
[Wang and Rudin 15]
[Letham, Rudin, McCormick, Madigan '15]
[Hauser, Toubia, Evgeniou, Befurt, Dzyabura 10]
[Wang, Rudin, Doshi-Velez, Liu, Klampfl, MacNeille 17]

Learning certifiably **optimal** rule list
[Angelino, Larus-Stone, Alabi, Seltzer, Rudin '18]

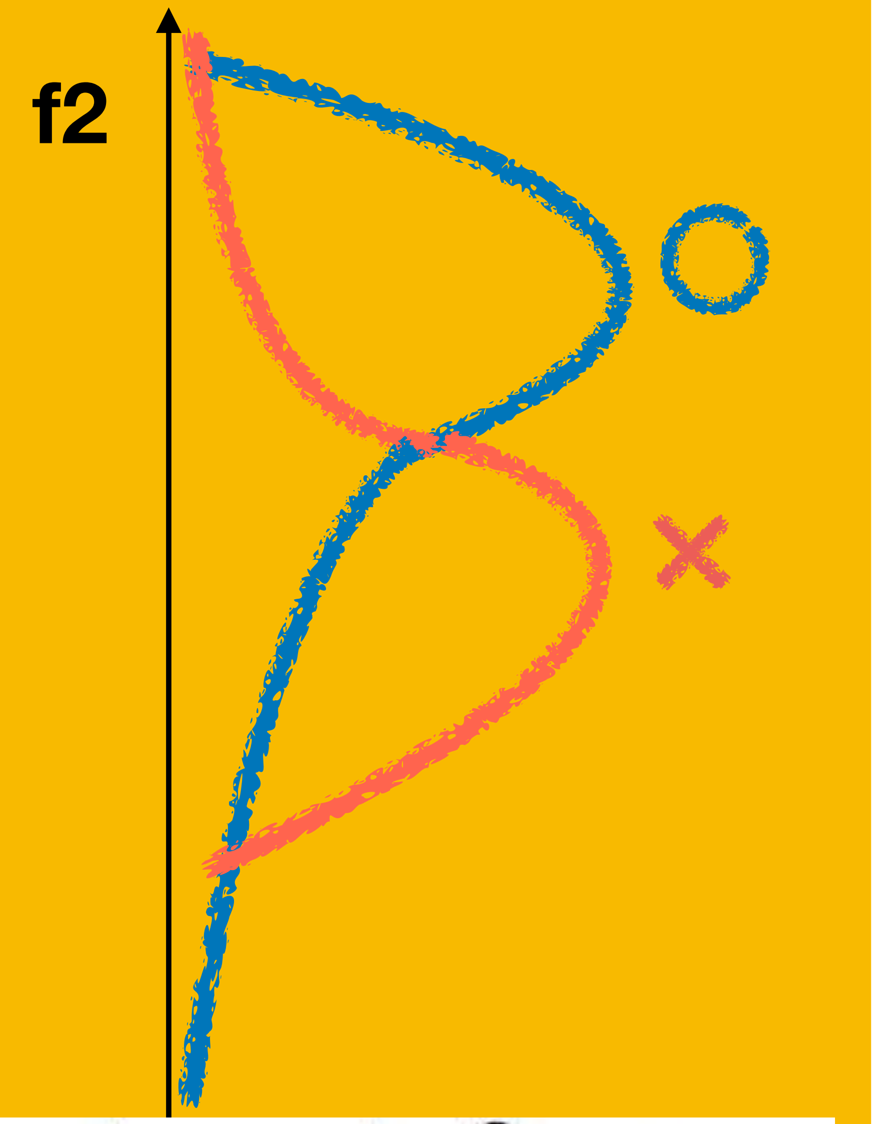
Building a new model

Fit a simpler function for each feature





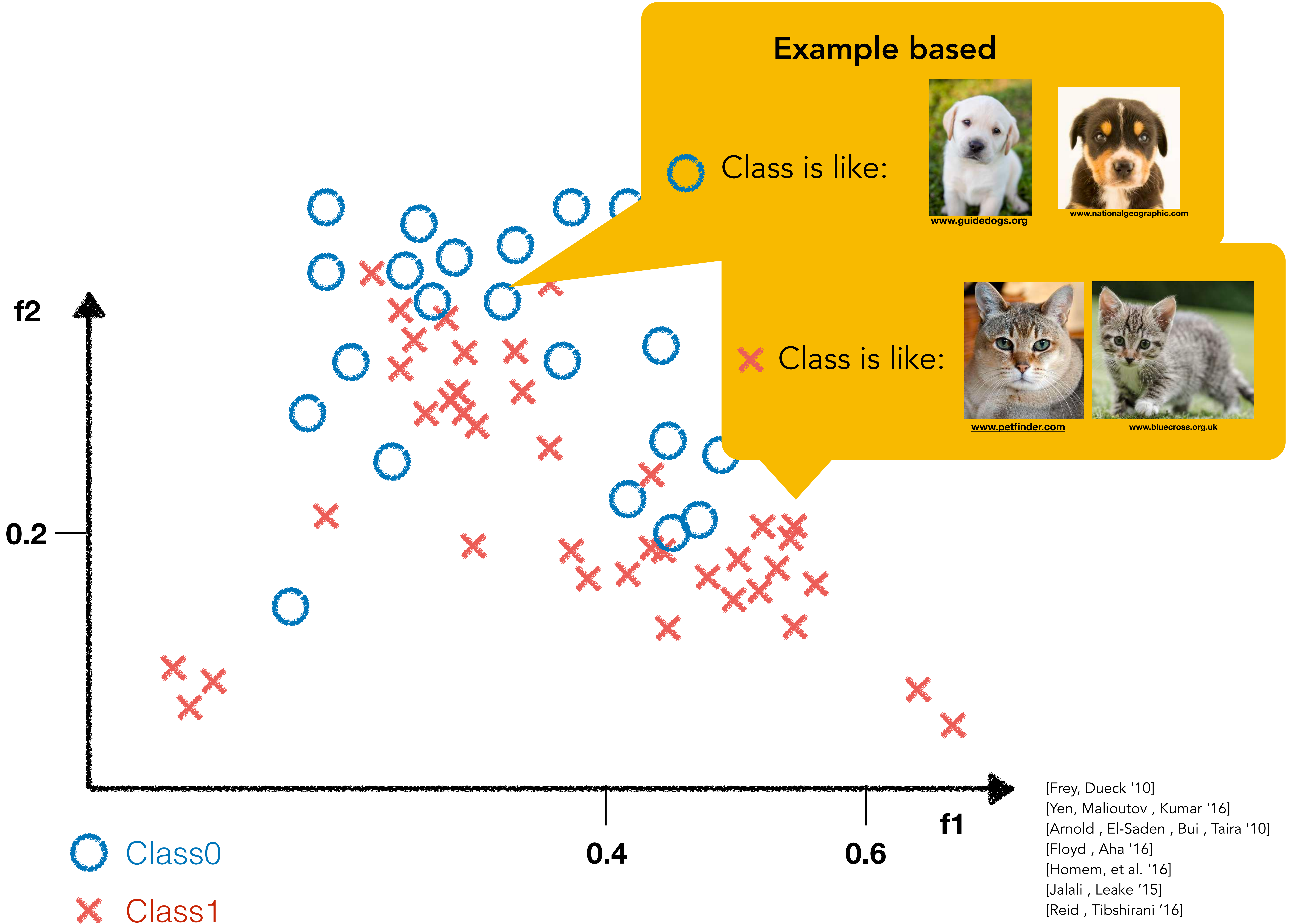
Fit a simpler function



Linear model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
generalized linear model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$
generalized additive model	$g(y) = f_1(x_1) + ... + f_n(x_n)$
generalized additive2 model	$g(E[y]) = \sum f_i(x_i) + \sum f_{ij}(x_i, x_j);$

[Lou et al. '12]

Table edited from [Gehrke et al. '12]



Example based

Class is like:

www.guidedogs.org

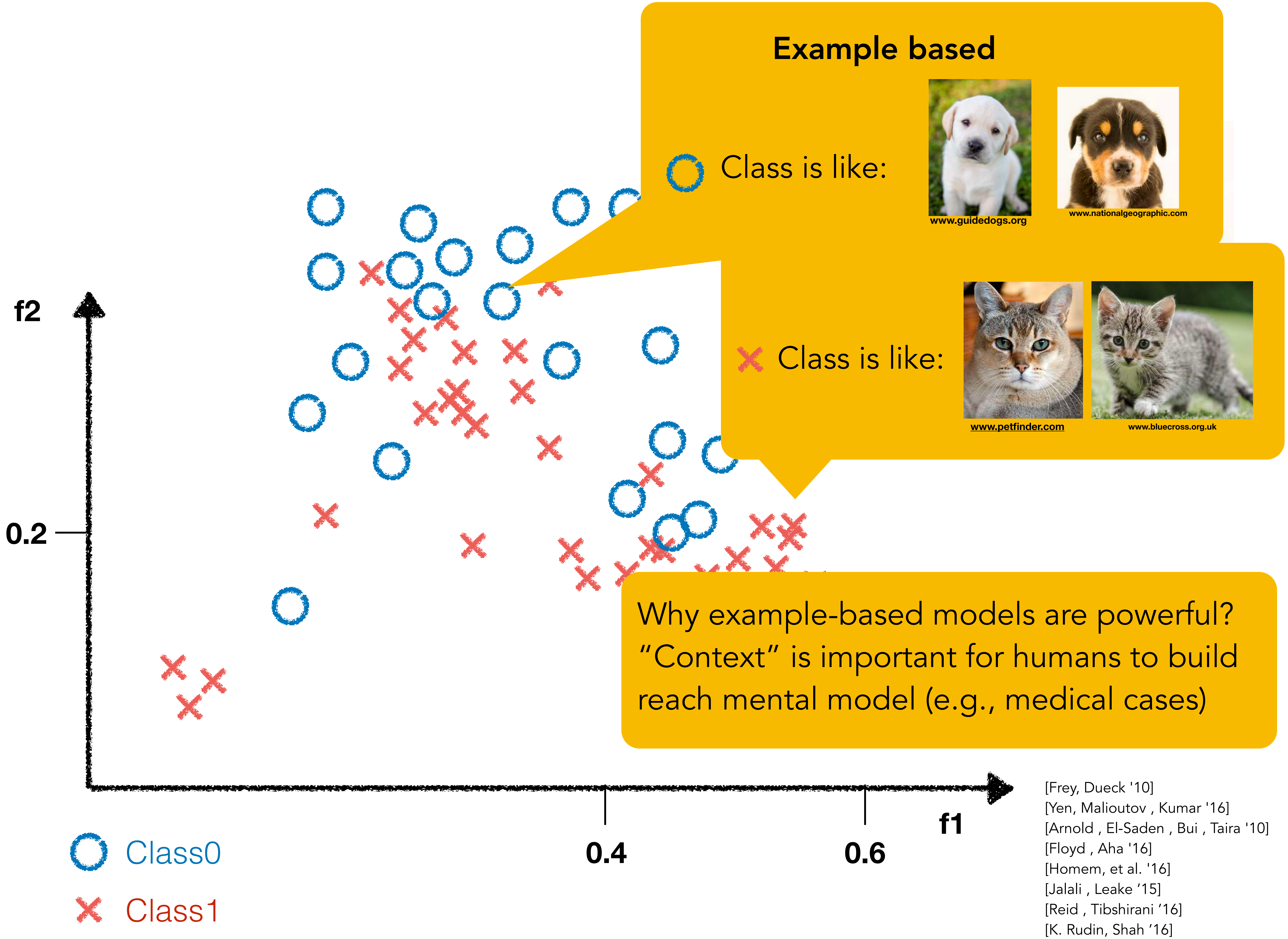
www.nationalgeographic.com

Class is like:

www.petfinder.com

www.bluecross.org.uk

[Frey, Dueck '10]
[Yen, Malioutov , Kumar '16]
[Arnold , El-Saden , Bui , Taira '10]
[Floyd , Aha '16]
[Homem, et al. '16]
[Jalali , Leake '15]
[Reid , Tibshirani '16]
[K. Rudin, Shah '16]
[Koh, Liang '17]



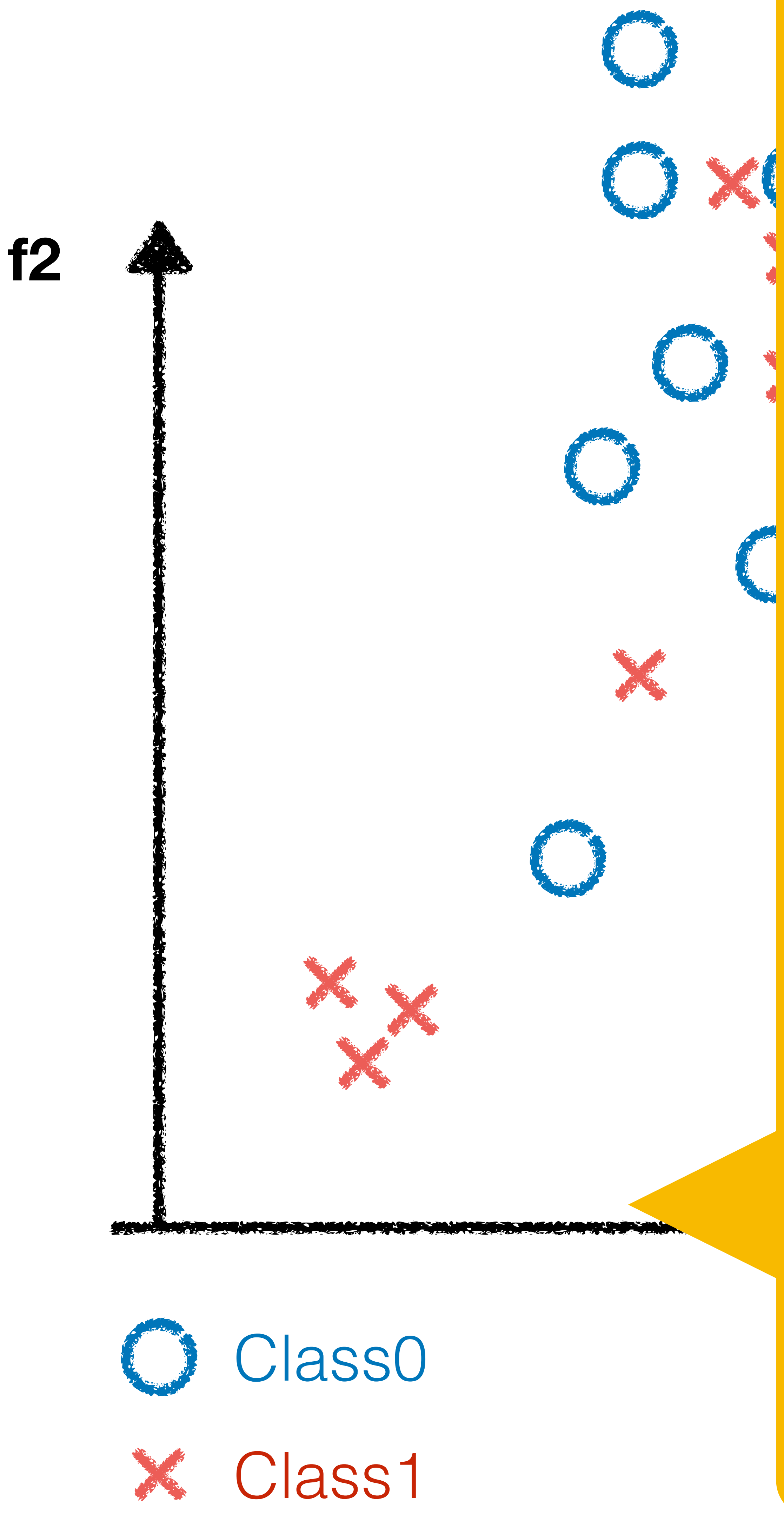
Example based

○ Class is like:

✕ Class is like:

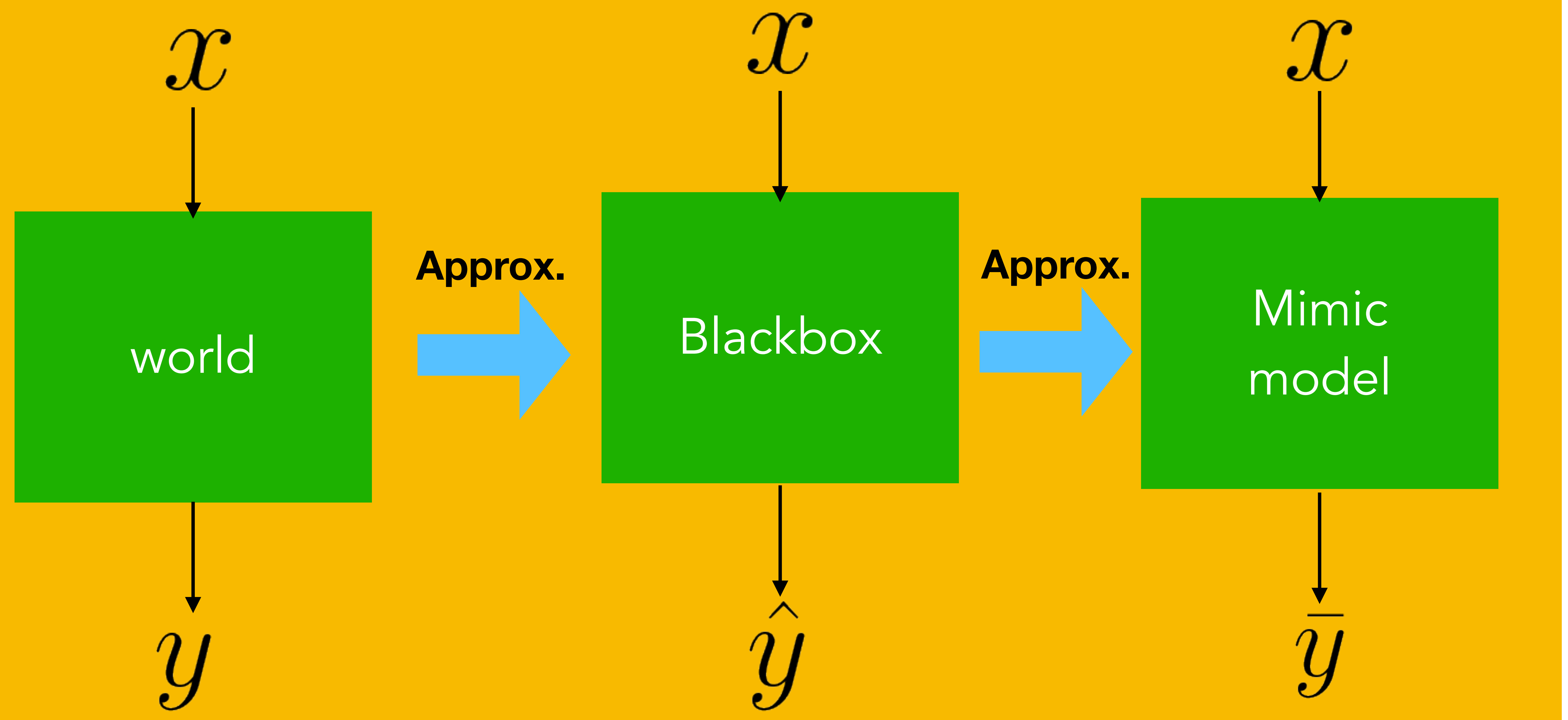
Why example-based models are powerful?
“Context” is important for humans to build reach mental model (e.g., medical cases)

[Frey, Dueck '10]
[Yen, Malioutov , Kumar '16]
[Arnold , El-Saden , Bui , Taira '10]
[Floyd , Aha '16]
[Homem, et al. '16]
[Jalali , Leake '15]
[Reid , Tibshirani '16]
[K. Rudin, Shah '16]
[Koh, Liang '17]



Mimic models, model distillation

Building a simpler model that walks,
talks, barks like the complex model.



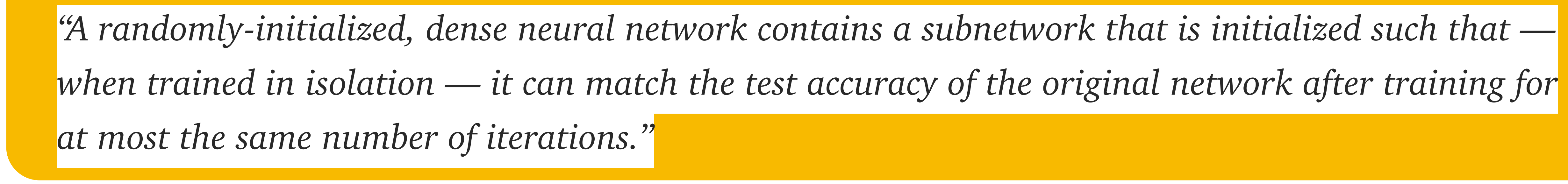
[Bucila et al. '06]

Do Deep Nets Really Need to be Deep? [Ba et al. '14]

Distilling the Knowledge in a Neural Network [Hinton et al. '15] [Frosst '17]

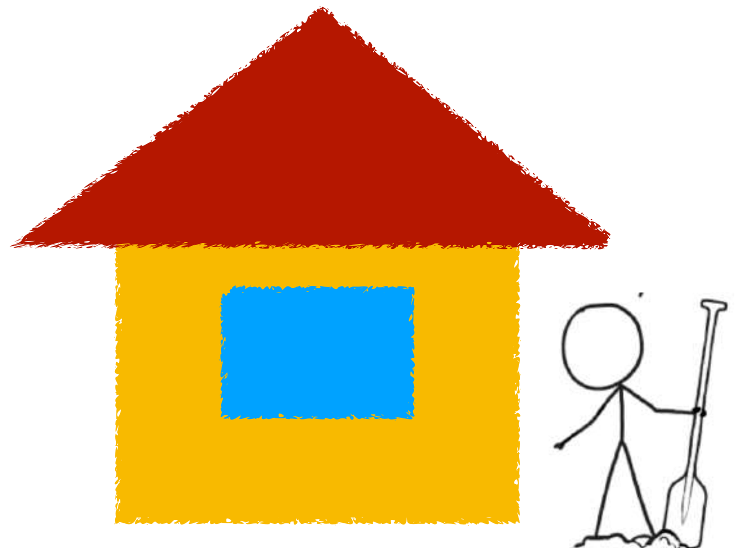
The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

Jonathan Frankle, Michael Carbin

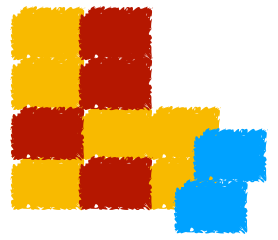


"A randomly-initialized, dense neural network contains a subnetwork that is initialized such that — when trained in isolation — it can match the test accuracy of the original network after training for at most the same number of iterations."

Types of interpretability methods



My ML



Explaining data

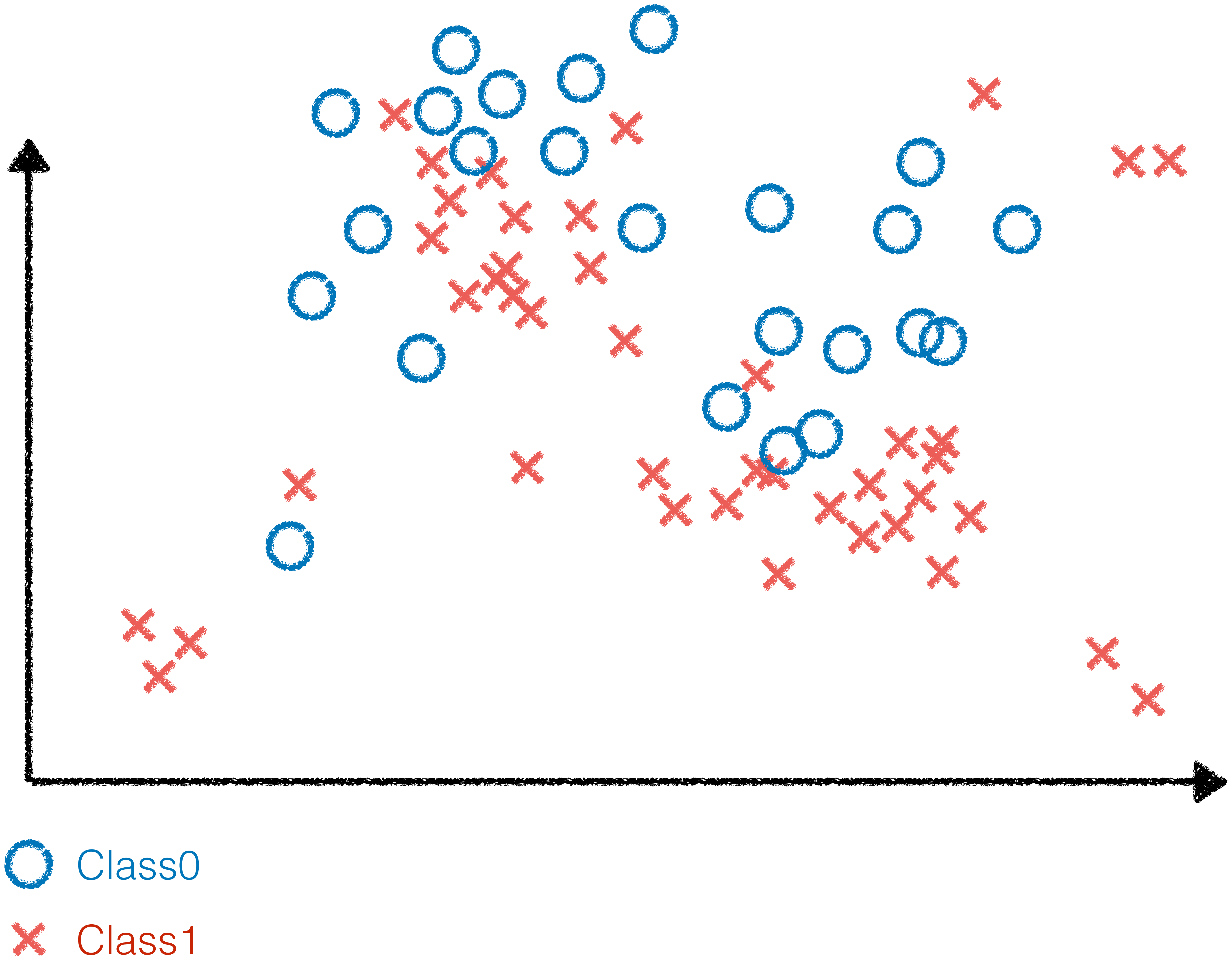
$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Building inherently interpretable model

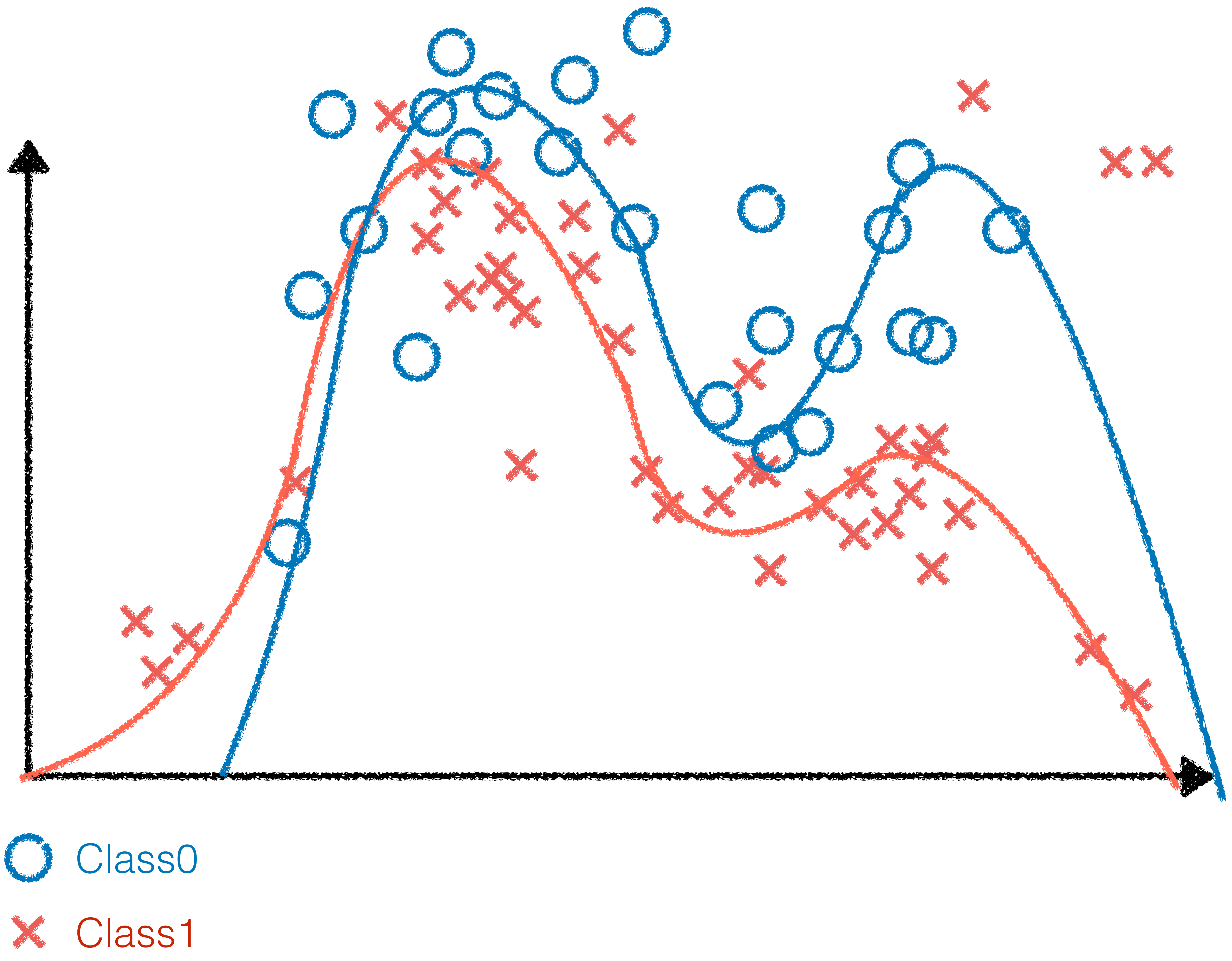
$$\underset{E,M}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}, \mathbf{M}\text{odel}|\mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Post-training interpretability methods

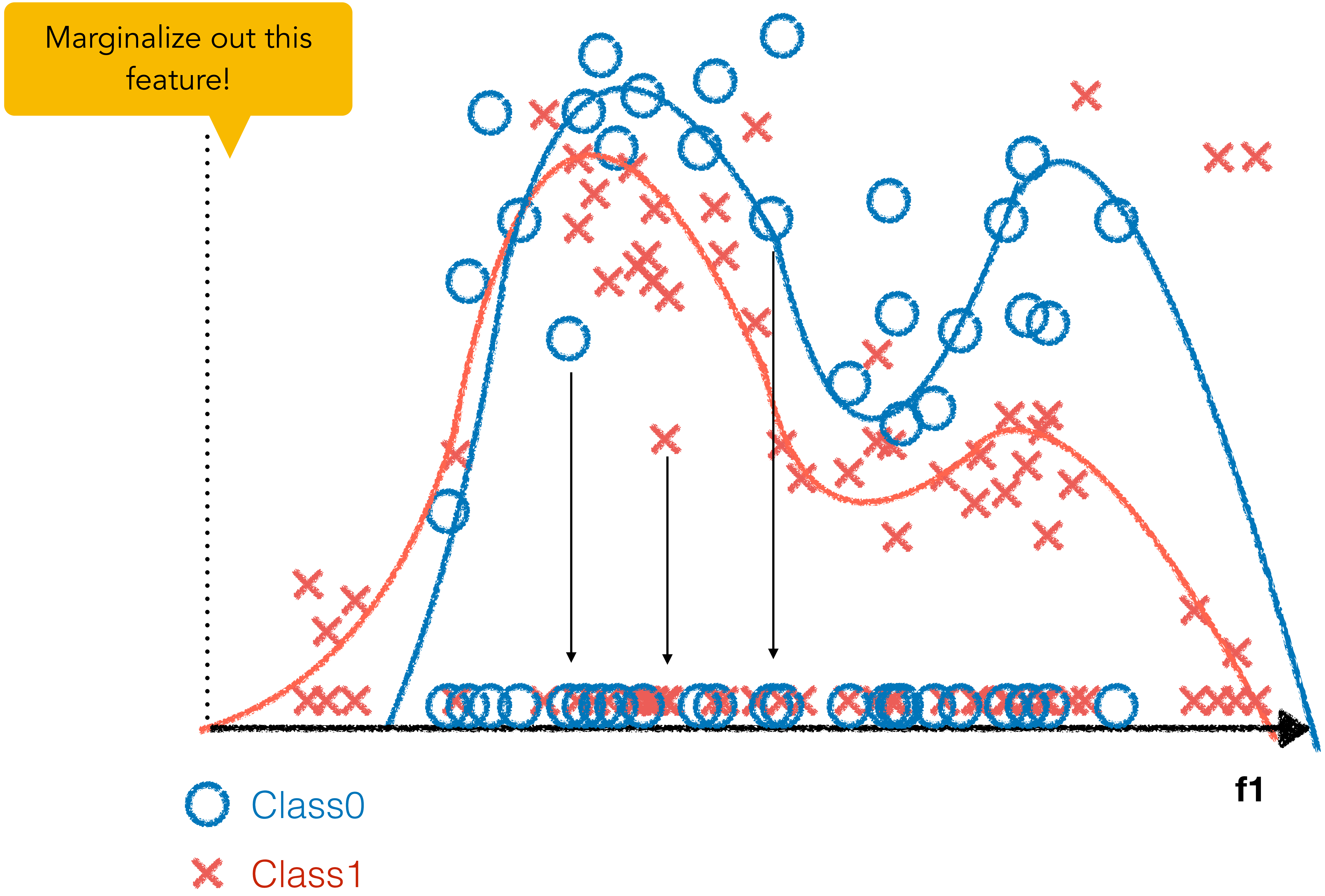
$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}|\mathbf{M}\text{odel}, \mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$


Class0
Class1

After building a model



After building a model

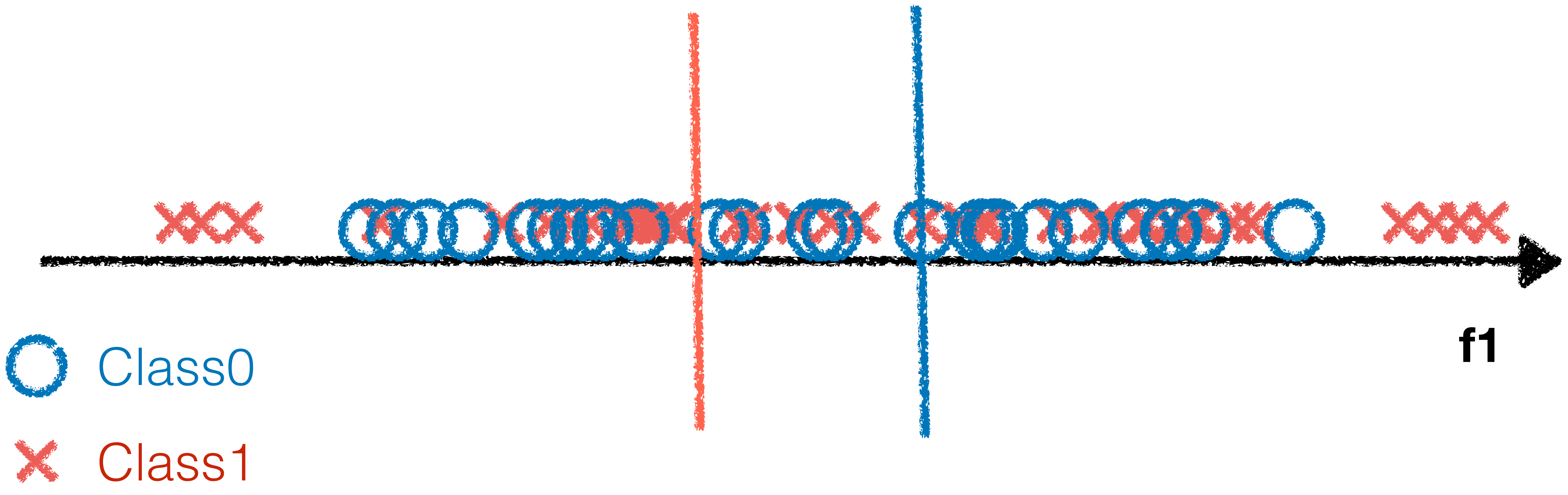

Marginalize out this feature!

○ Class0

× Class1

After building a model

1. **Ablation test: train without that feature/data points and see the impact**



After building a model

Definition of importance: "how would the model's predictions change if a training input were modified?"

1. **Ablation test: train without that feature/data points and see the impact**

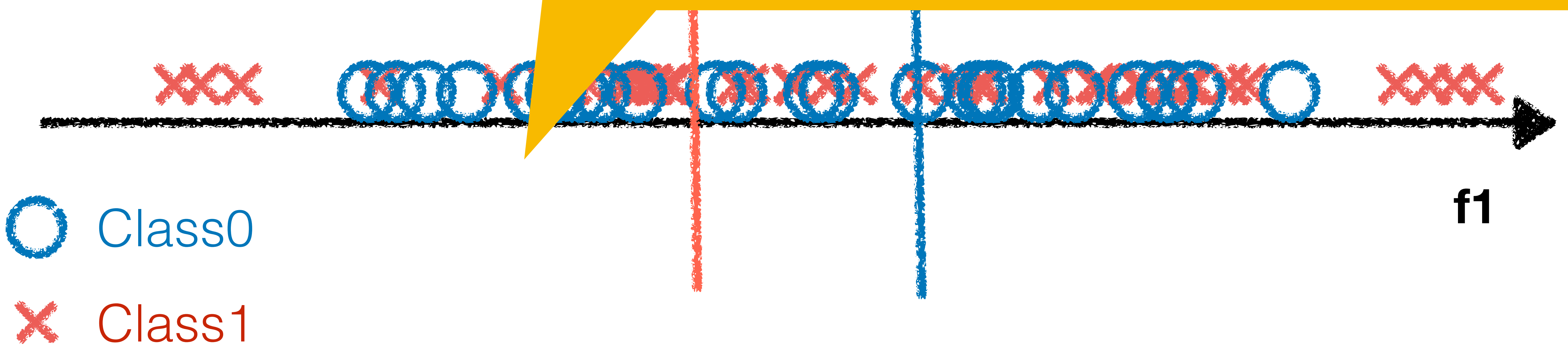
Smarter ablation Influence functions [Koh et al.'17]

To classify this image:

Model found these images most helpful

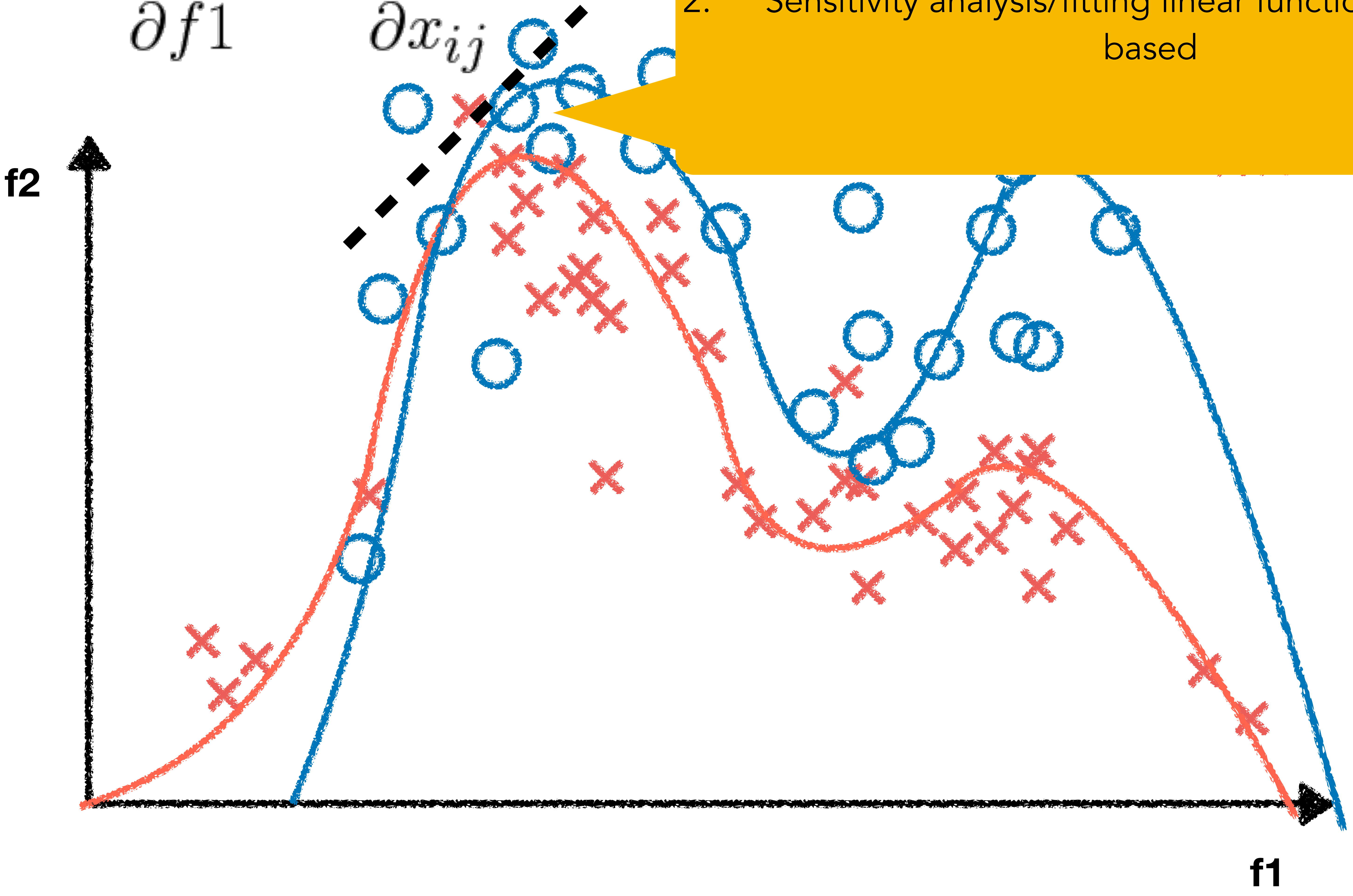
SVM

Inception



After building a model

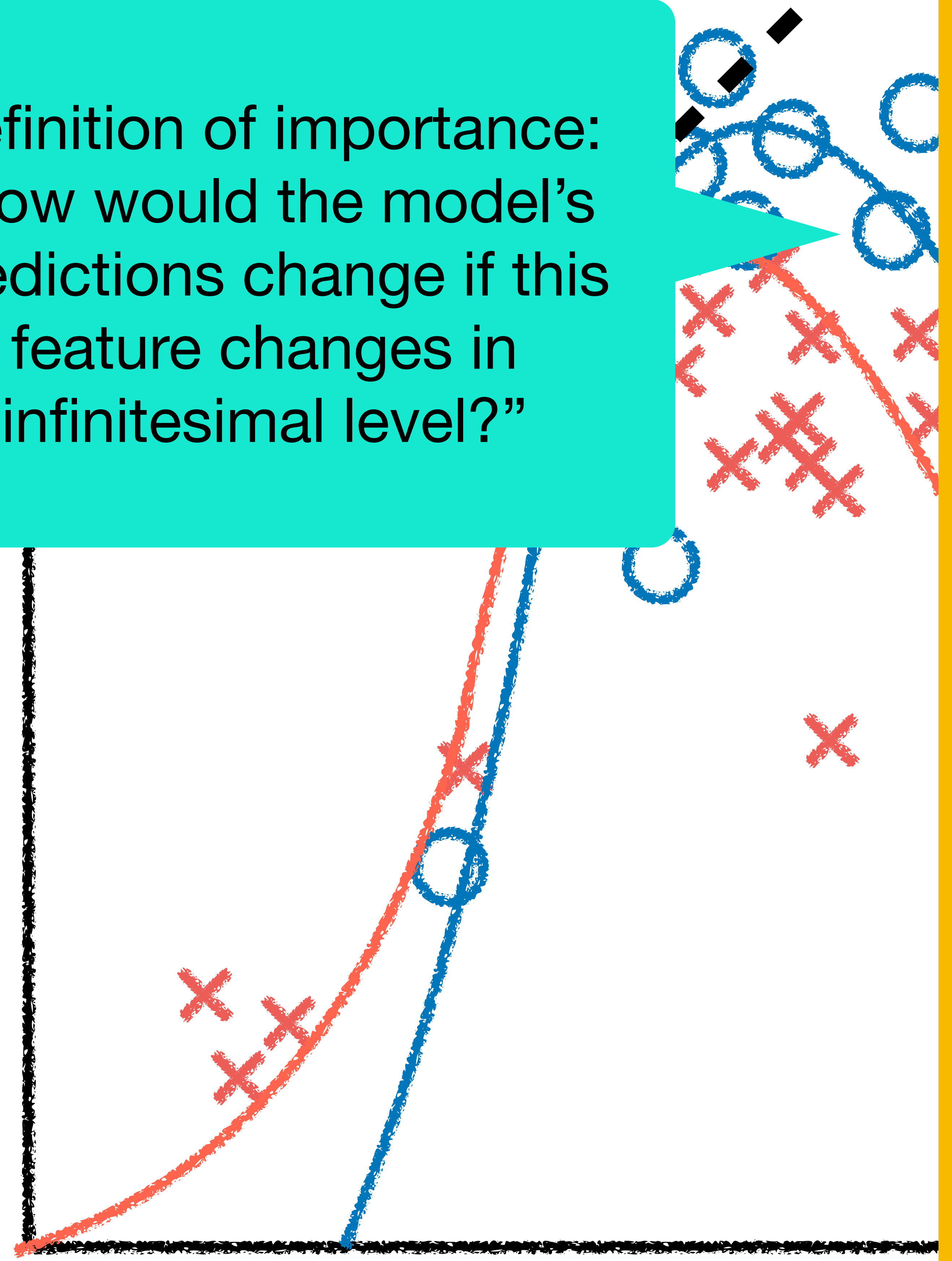
$$\frac{\partial p(x)}{\partial f1} = \frac{\partial p(x)}{\partial x_{ij}}$$



1. Ablation test: train without that feature/data and see the impact
2. Sensitivity analysis/fitting linear function/gradient-based

After building a model

$$\frac{\partial p(x)}{\ } = \frac{\partial p(x)}{\ }$$



Definition of importance: "how would the model's predictions change if this feature changes in infinitesimal level?"

1. Ablation test: train without that feature/data and see the impact
2. **2.** Sensitivity analysis/fitting linear function/gradient-based

Sensitivity analysis on model [Ribeiro et al. '16]

Want local explanation

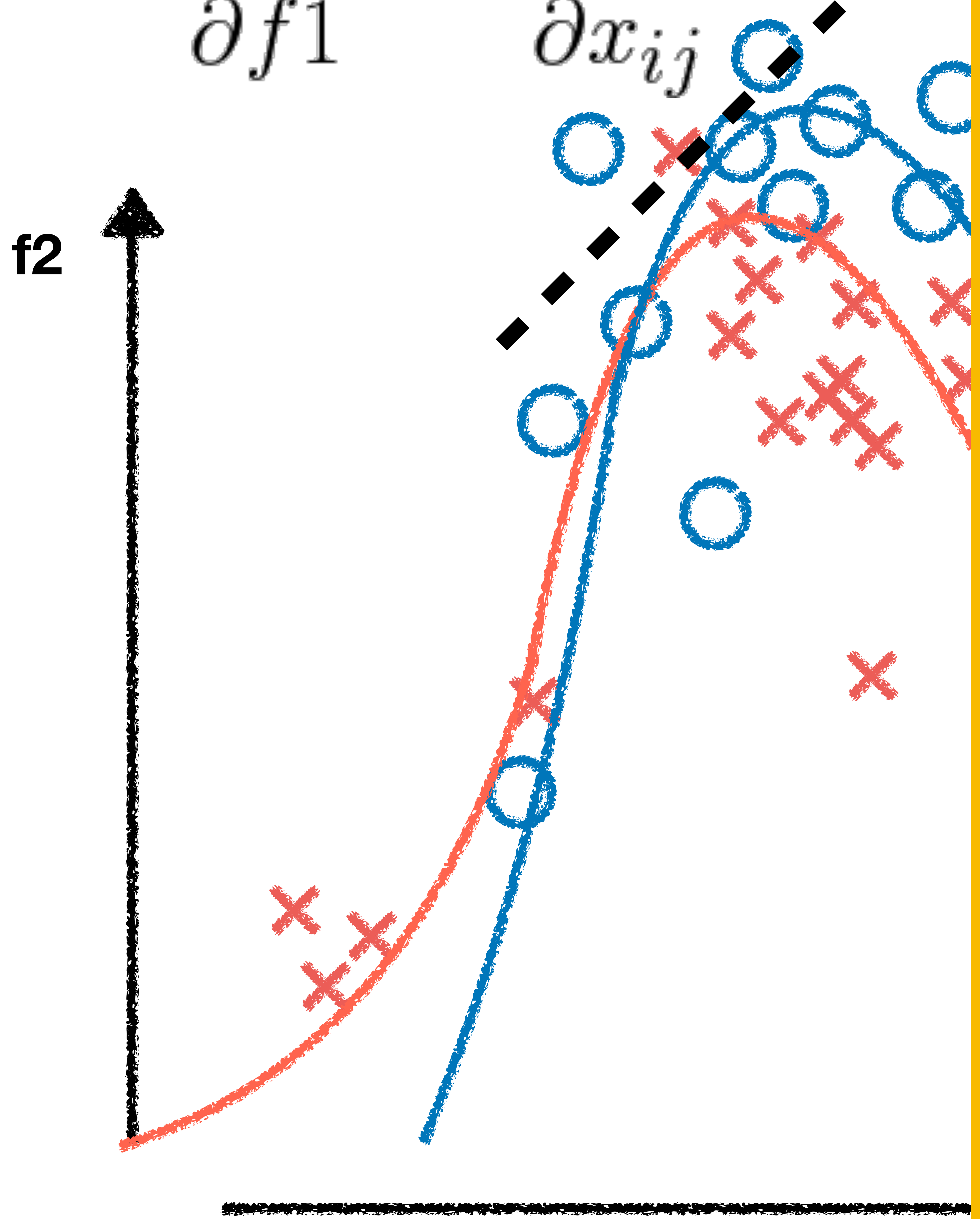
of the + data point

Locally fitted linear function

Many sensitivity analysis literature
[Ribeiro et al. '16] [Simonyan et al., '13] [Li et al., '16]
[Datta et al. '16] [Adler et al., '16] [Bach '15]

After building a model

$$\frac{\partial p(x)}{\partial f1} = \frac{\partial p(x)}{\partial x_{ij}}$$



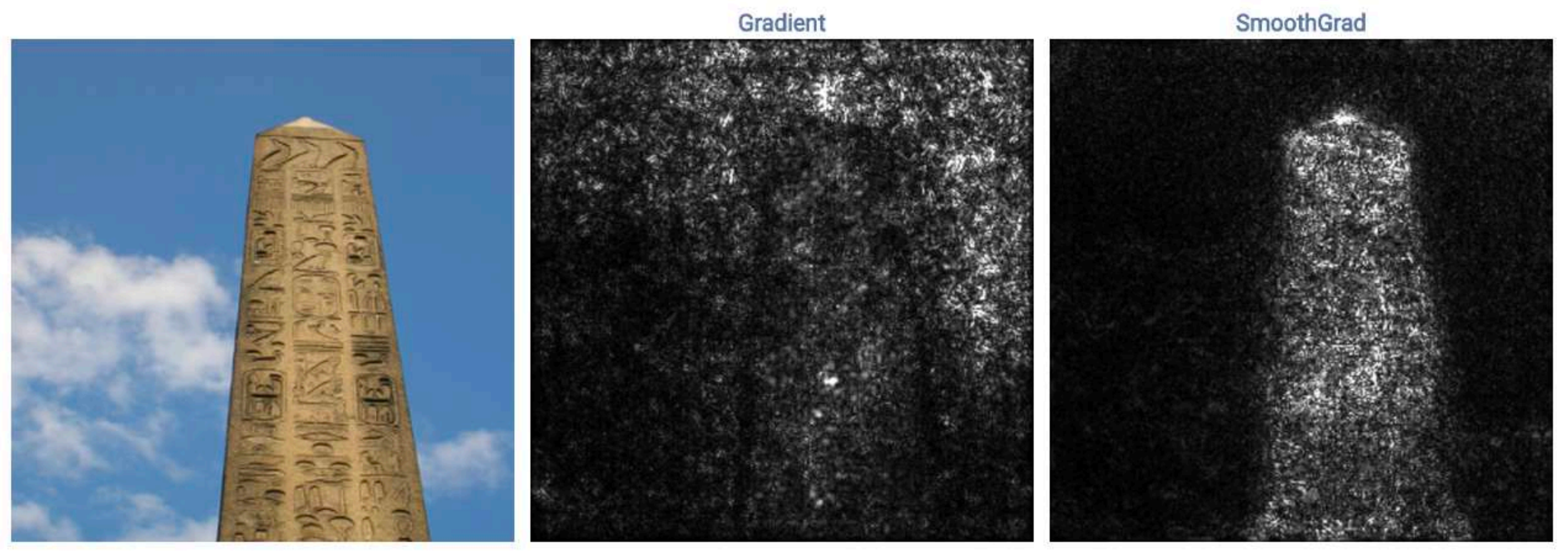
1. Ablation test: train without that feature/data and see the impact
2. Sensitivity analysis/fitting linear function/ **gradient-based**

Integrated gradients [Sundararajan et al. 17]


Top label: starfish

Score: 0.999992

SmoothGrad [Smilkov et al. 17]

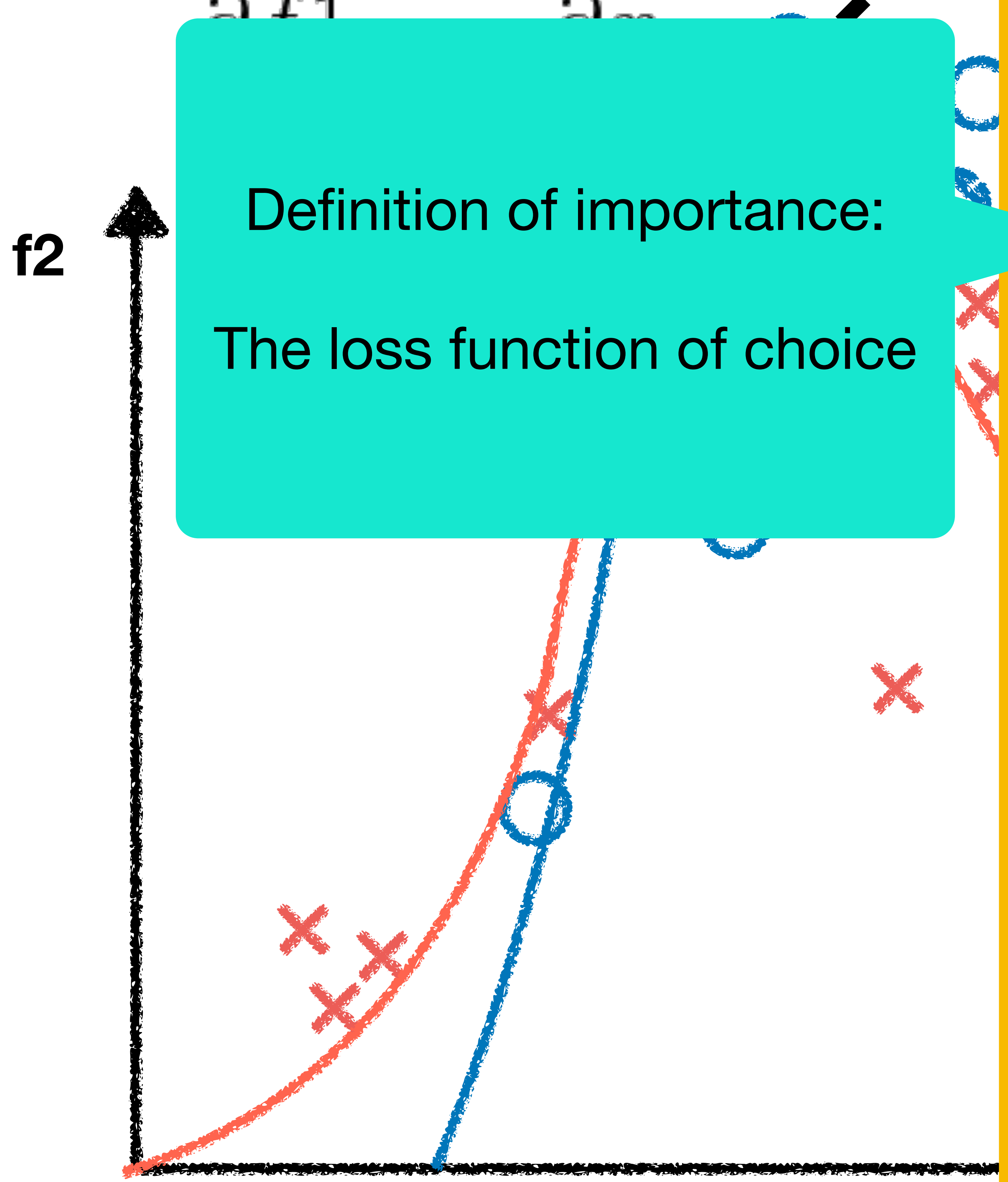


[Zeiler et al. '13] [Selvaraju et al. 16]

[Erhan 2009] [Springenberg, '14] [Shrikumar '17] and many more..

After building a model

$$\frac{\partial p(x)}{\partial f1} = \frac{\partial p(x)}{\partial x}$$

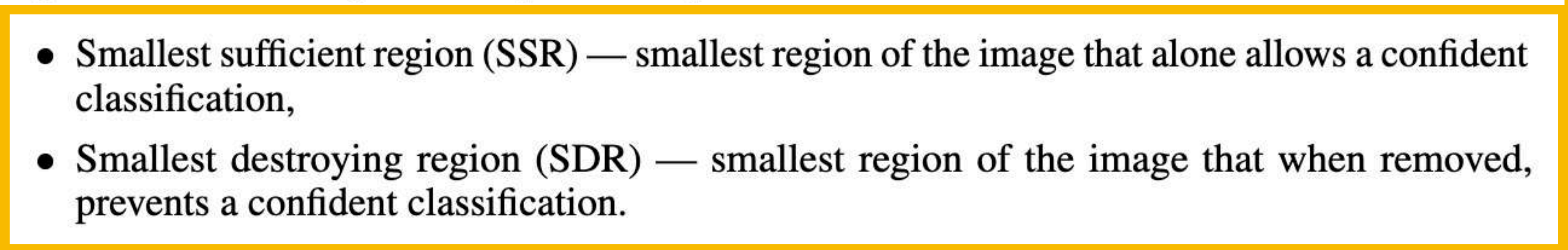


Definition of importance:

The loss function of choice

1. Ablation test: train without that feature/data and see the impact
2. Sensitivity analysis/fitting linear function/ gradient-based
3. Optimization based methods

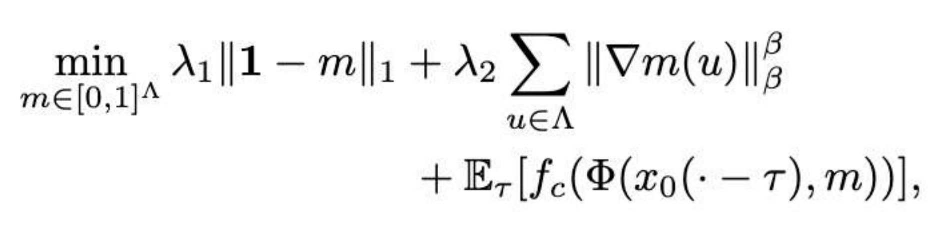
Figure 2: From left to right: the input image; smallest sufficient region (SSR); smallest destroying region (SDR). Regions were found using the mask optimisation procedure from [3].



- Smallest sufficient region (SSR) — smallest region of the image that alone allows a confident classification,
- Smallest destroying region (SDR) — smallest region of the image that when removed, prevents a confident classification.

[Dabkowski et al. 17]

Figure 1. An example of a mask learned (right) by blurring an image (middle) to suppress the softmax probability of its target class (left: original image; softmax scores above images).

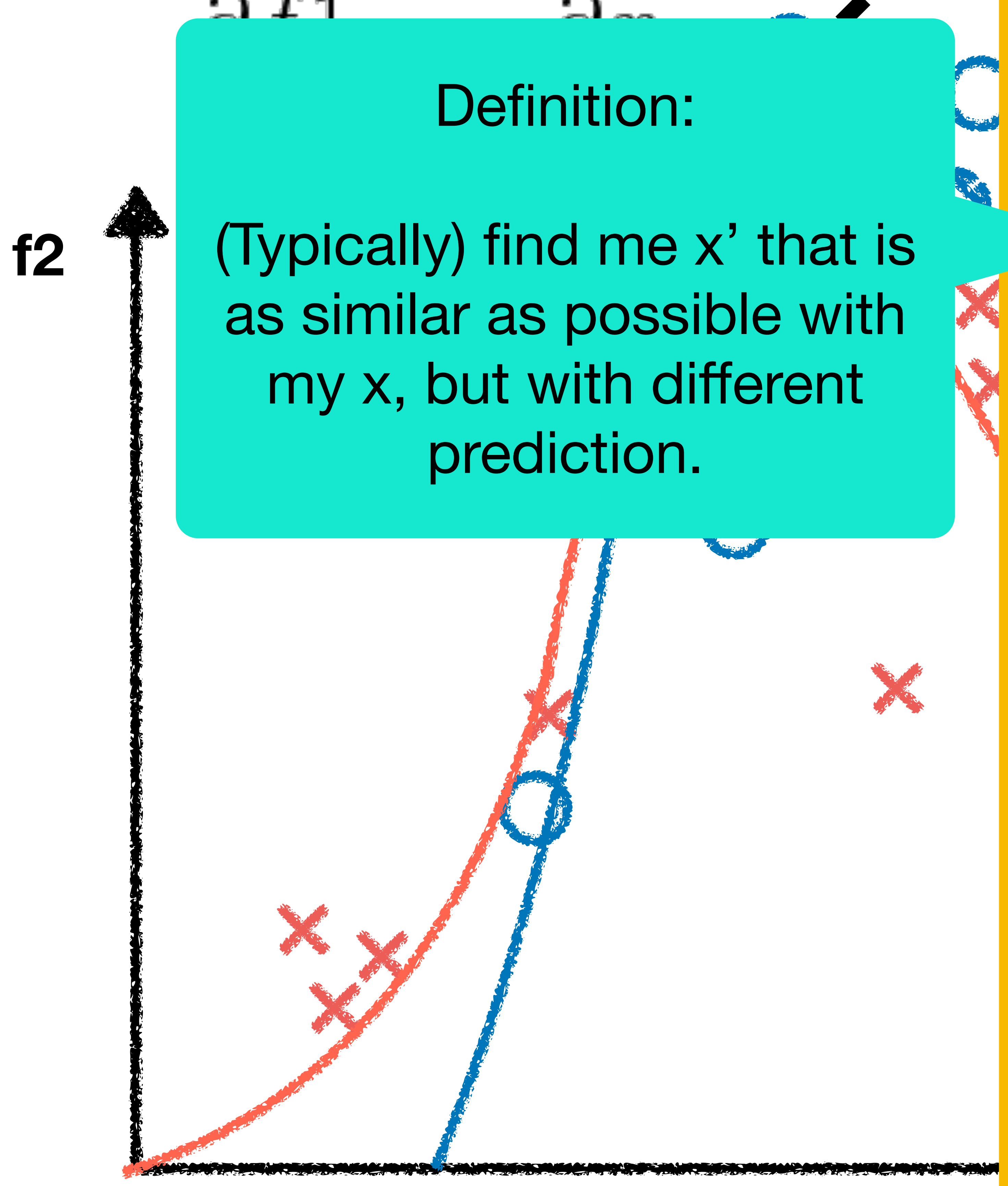


$$\min_{m \in [0,1]^{\Lambda}} \lambda_1 \|\mathbf{1} - m\|_1 + \lambda_2 \sum_{u \in \Lambda} \|\nabla m(u)\|_\beta^\beta + \mathbb{E}_\tau[f_c(\Phi(x_0(\cdot - \tau), m))],$$

[Fong et al. 17, 18, 20]

After building a model

$$\frac{\partial p(x)}{\partial f1} = \frac{\partial p(x)}{\partial x}$$



Definition:

(Typically) find me x’ that is as similar as possible with my x, but with different prediction.

1. Ablation test: train without that feature/data and see the impact
2. **Sensitivity analysis/fitting linear function/ gradient-based**
3. Optimization based methods
4. Counterfactual explanations

$$\arg\min_{x'}\max_{\lambda} \lambda(f_w(x') - y')^2 + d(x_i, x')$$

[Watcher et al. 18]

$$d(x_i, x') = \sum_{k \in F} \frac{|x_{i,k} - x'_k|}{\mathrm{MAD}_k}$$

$$\widetilde{\mathcal{L}_{\mathcal{M}}}(y, \bar{x}) = \mathbb{1}\left[y = \arg\max_{y'} \mathcal{M}(y' \mid \bar{x})\right] \cdot \widetilde{\mathcal{M}}(y \mid \bar{x}).$$

Extend Watcher’s framework to non-differentiable (e.g., trees) [Lucic et al. 21]

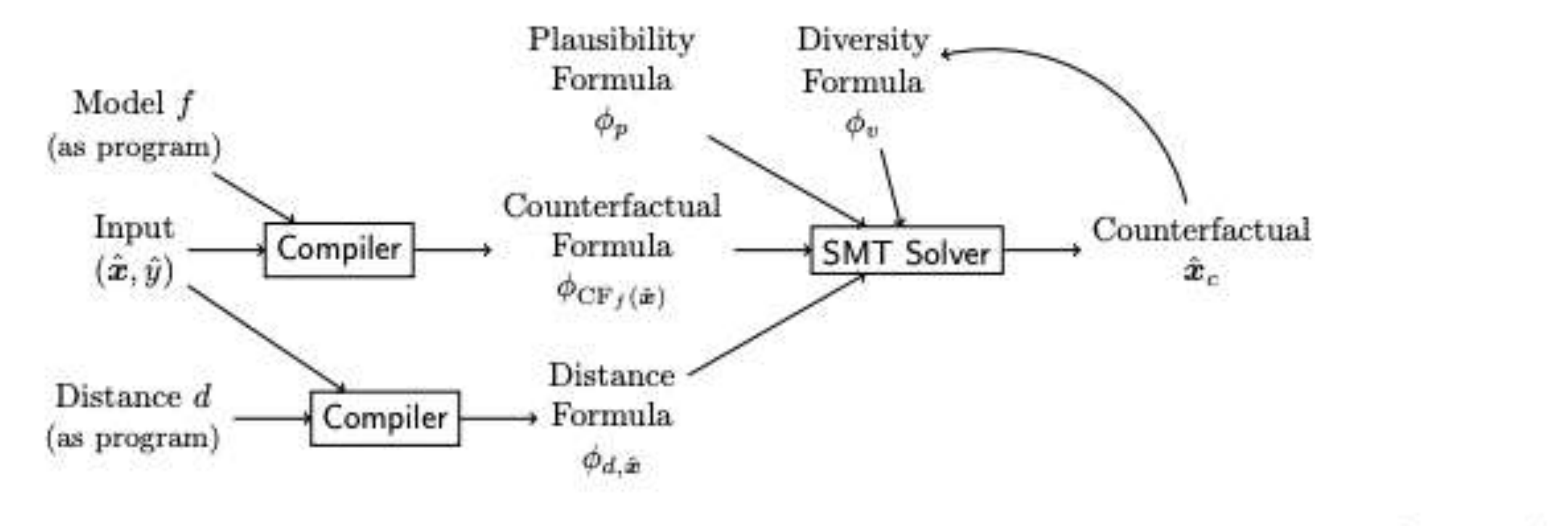
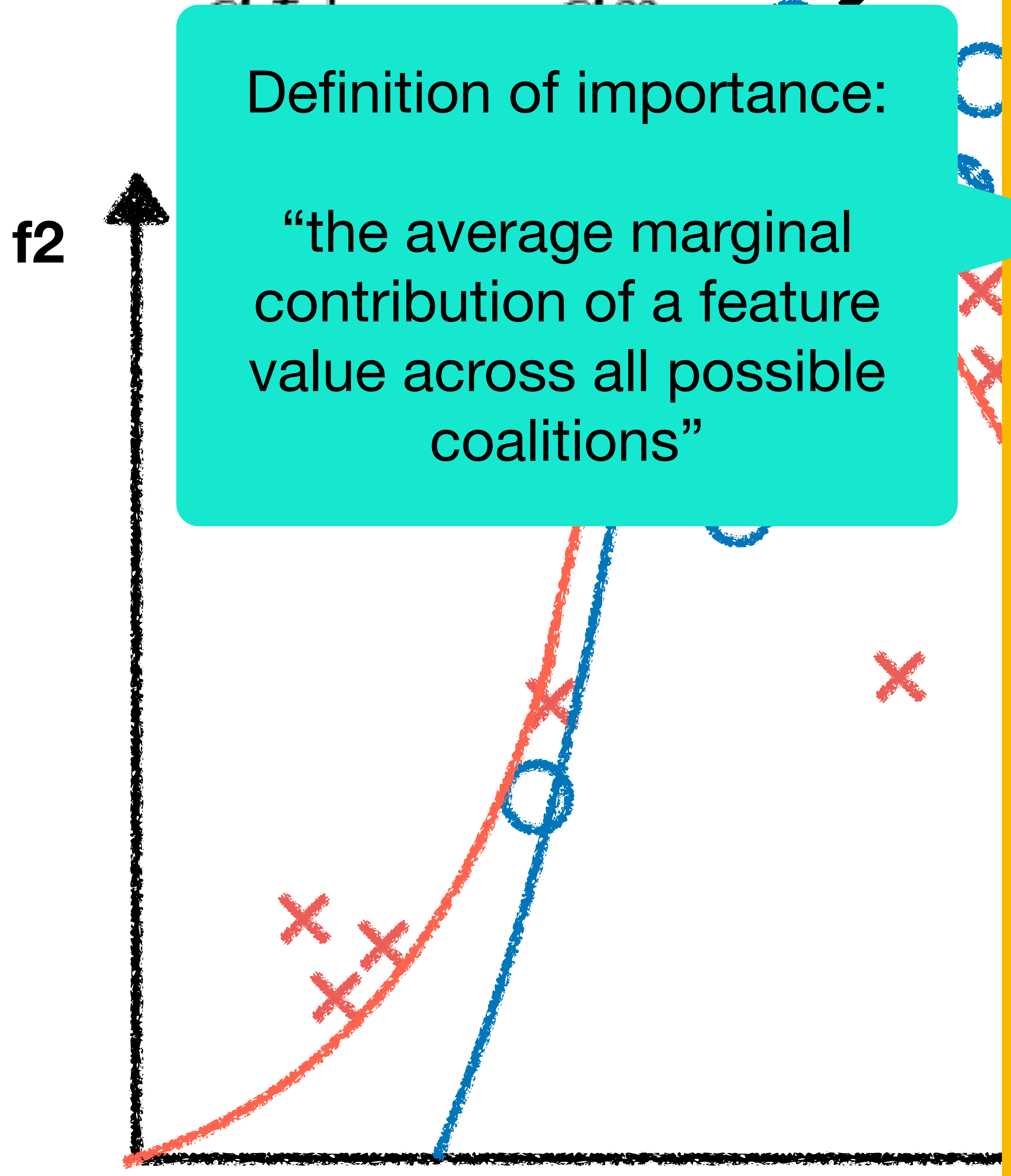


Figure 1: Architecture Overview for Model-Agnostic Counterfactual Explanations (MACE) [Karimi 20]

After building a model

$$\frac{\partial p(x)}{\partial f1} = \frac{\partial p(x)}{\partial x}$$



Definition of importance:

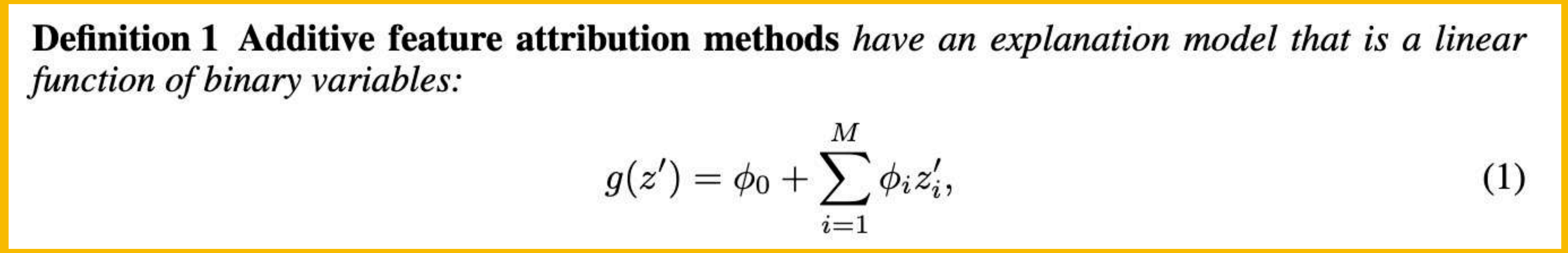
"the average marginal contribution of a feature value across all possible coalitions"

1. Ablation test: train without that feature/data and see the impact
2. Sensitivity analysis/fitting linear function/ gradient-based
3. Optimization based methods
4. Counterfactual explanations
5. Game theoretic approach

Shapley [Shapley 53]
how important is each player to the overall cooperation ("gain")
<-> The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

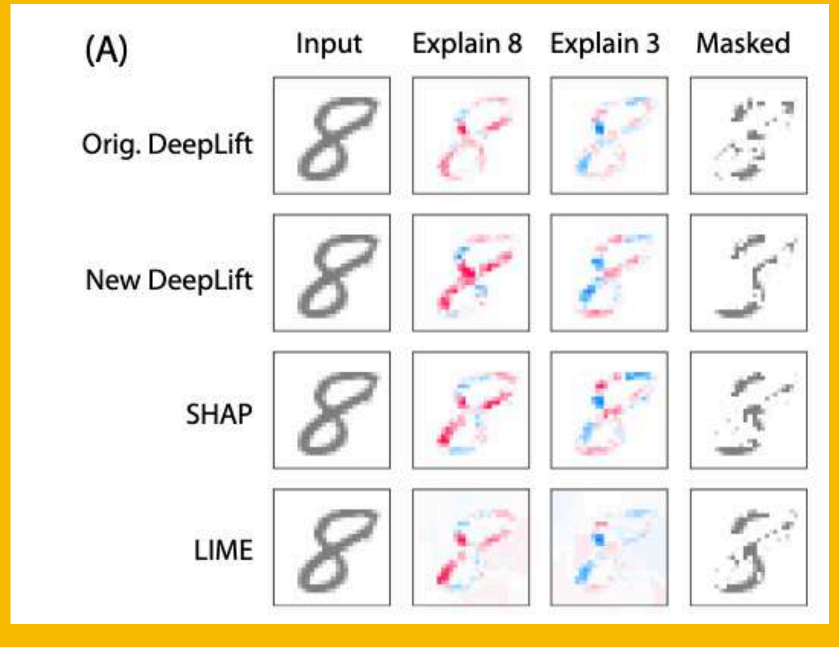
gain = prediction value for x - E[all prediction values]

SHAP [Lundberg et al. 17]



Definition 1 Additive feature attribution methods *have an explanation model that is a linear function of binary variables:*

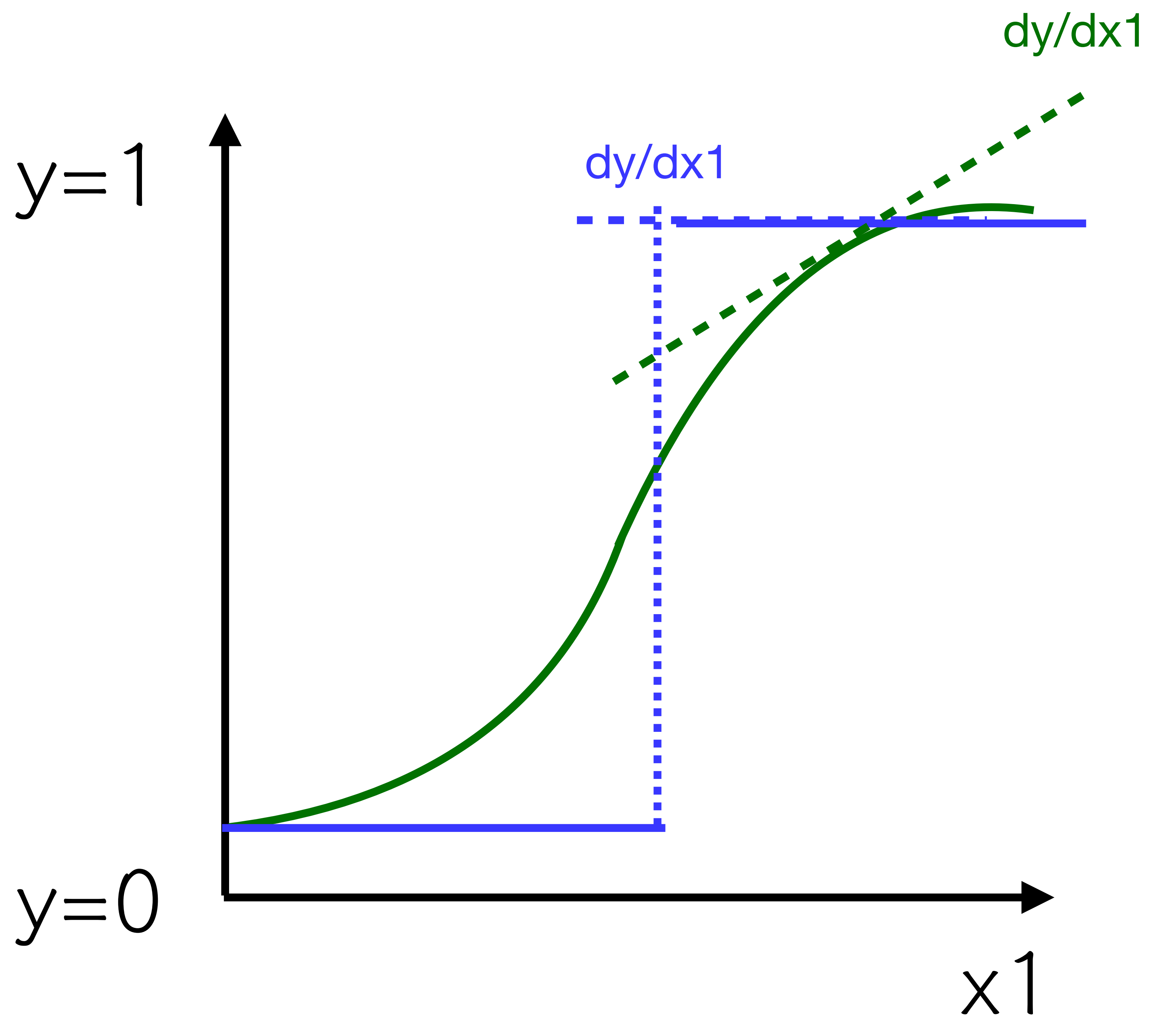
$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$



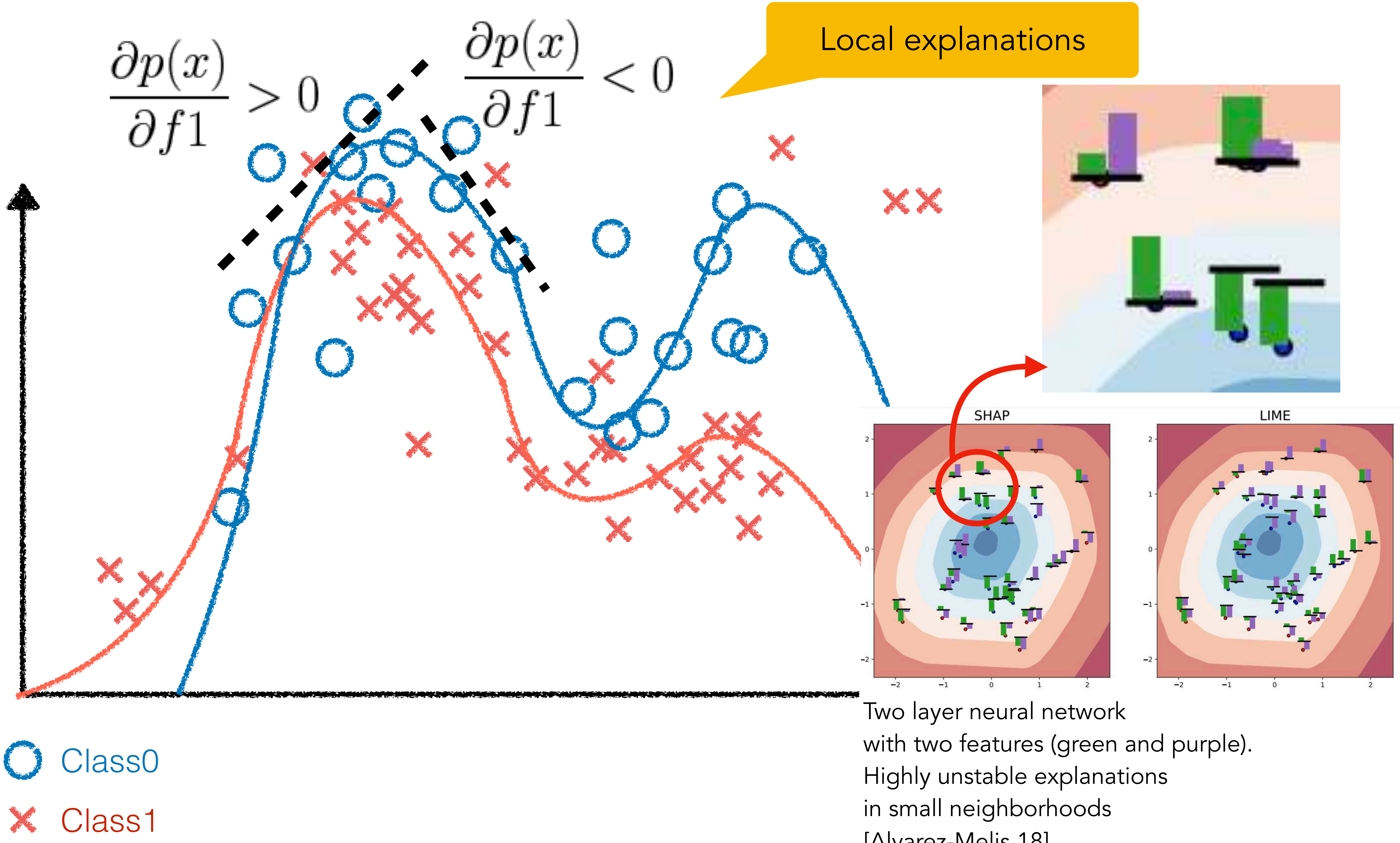
Aren't all these different definition of importance get you the similar answer, typically? Which one should I choose?

0-th order or 1-st order derivatives could lead to very different intuition

- Which feature is dominant (0-th order derivative)
  - feature x1 is important distinction between class y=1 and y=0 for both blue curve and green curve.
- Which feature is sensitive (1-th order derivative)
  - feature x1 is important distinction between class y=1 and y=0 for green curve (dy/dx > 0), but not for blue curve (dy/dx1=0).
- Neither represents causal relationship (of course)
- What you think you want may not be what you need! -> Test with the end-task.



That's all good. What could go wrong?

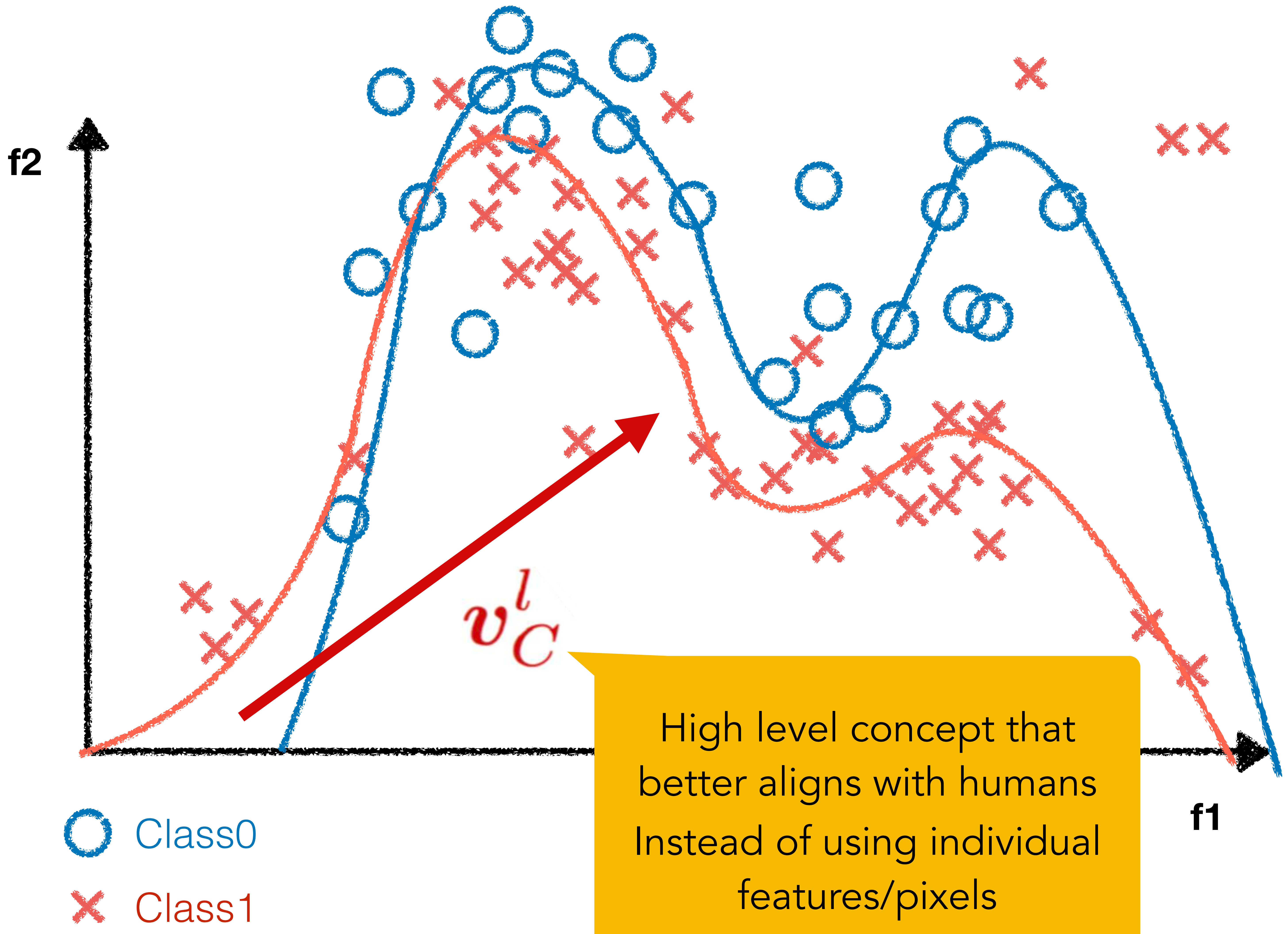


$$\frac{\partial p(x)}{\partial f1} > 0 \qquad \frac{\partial p(x)}{\partial f1} < 0$$

Local explanations

Two layer neural network
with two features (green and purple).
Highly unstable explanations
in small neighborhoods
[Alvarez-Melis 18]

After building a model



High level concept that better aligns with humans Instead of using individual features/pixels

Problem:
Post-training explanation

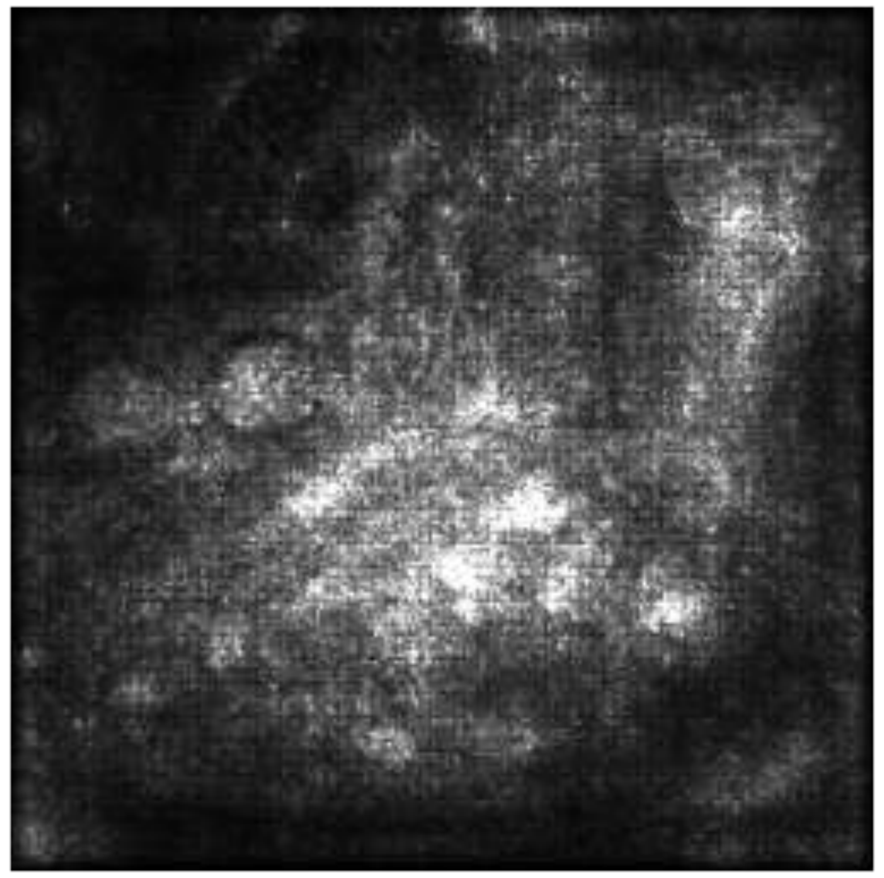
$$\underset{E}{\operatorname{argmax}}\ Q(\mathbf{Explanation}|\mathbf{Model},\mathbf{Human},\mathbf{Data},\mathbf{Task})$$

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

$p(z)$

popularity

Why was this a popular pizza?



VS

🍅 was important 0.8
🧀 was important 0.3
🍷 was important 0.1

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

Problem:
Post-training explanation

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

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

$p(z)$

popularity

Why was this a popular pizza?

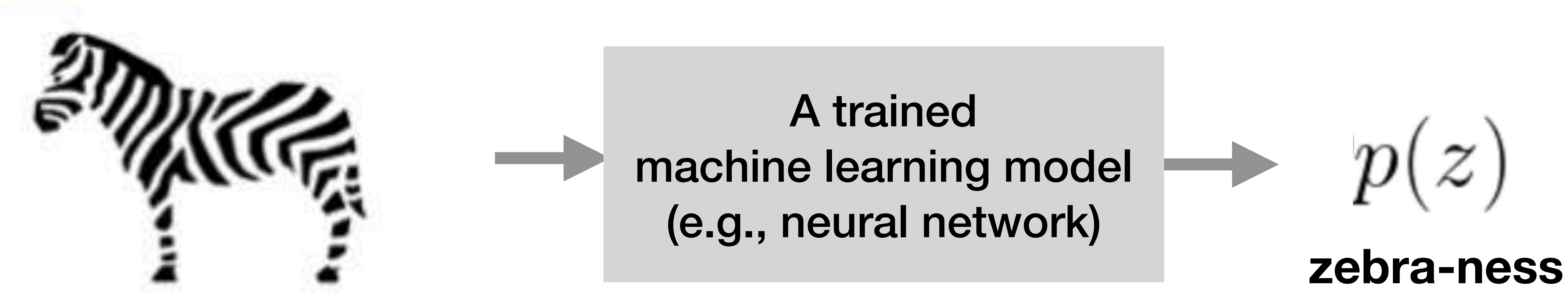
Quantitative explanation:
how much a concept
(e.g., gender, race)
was important for a prediction in a
trained model.

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

🍅 was important 0.8
🧀 was important 0.3
🍷 was important 0.1

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

TCAV: Testing with Concept Activation Vectors



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

$p(z)$

zebra-ness

How important was the striped concept
to this zebra image classifier?

[K., Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres ICML 2018]

TCAV: Testing with Concept Activation Vectors

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

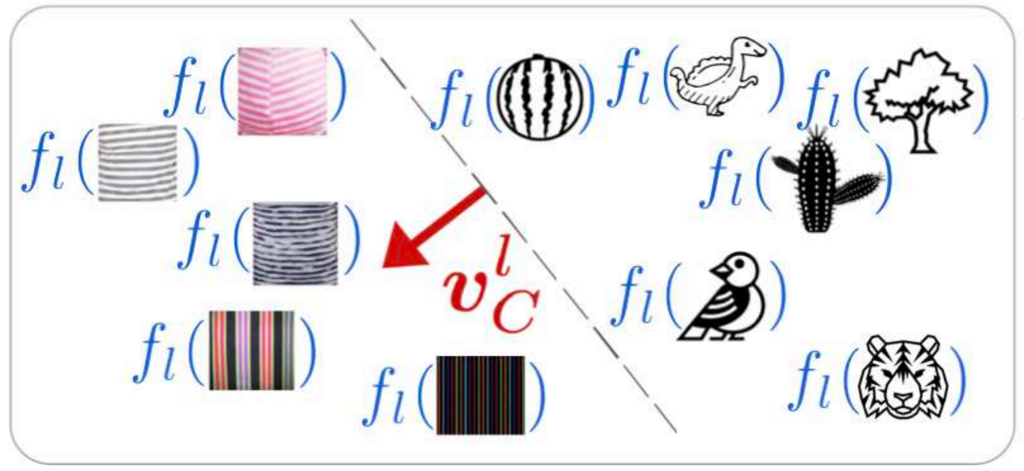
$p(z)$

zebra-ness

How important was the striped concept
to this zebra image classifier?

TCAV

1. Learning CAVs



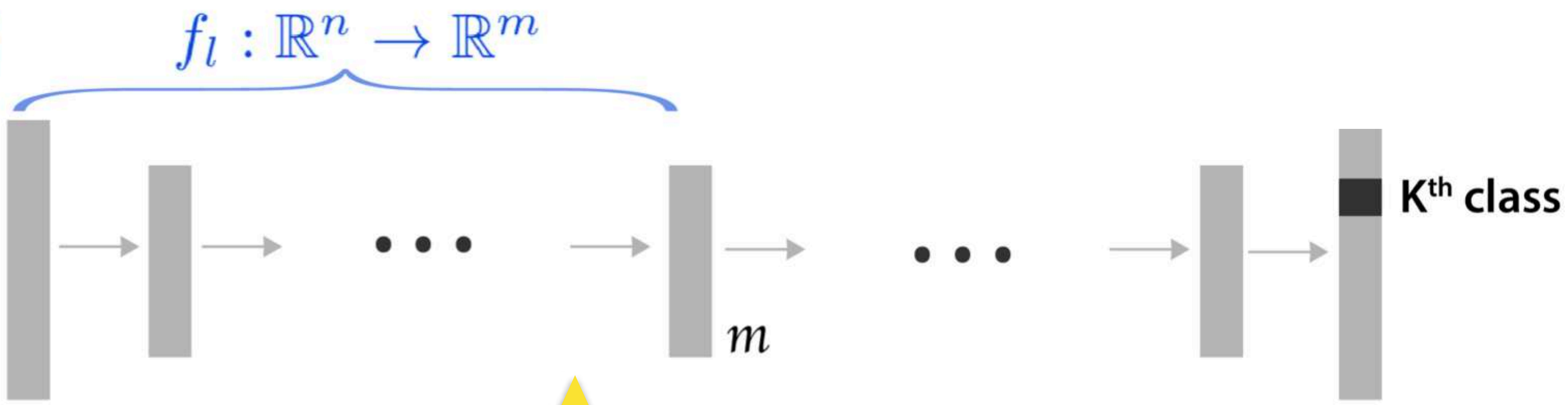
1. How to define concepts?

Defining concept activation vector (CAV)

Inputs:

Examples of concepts

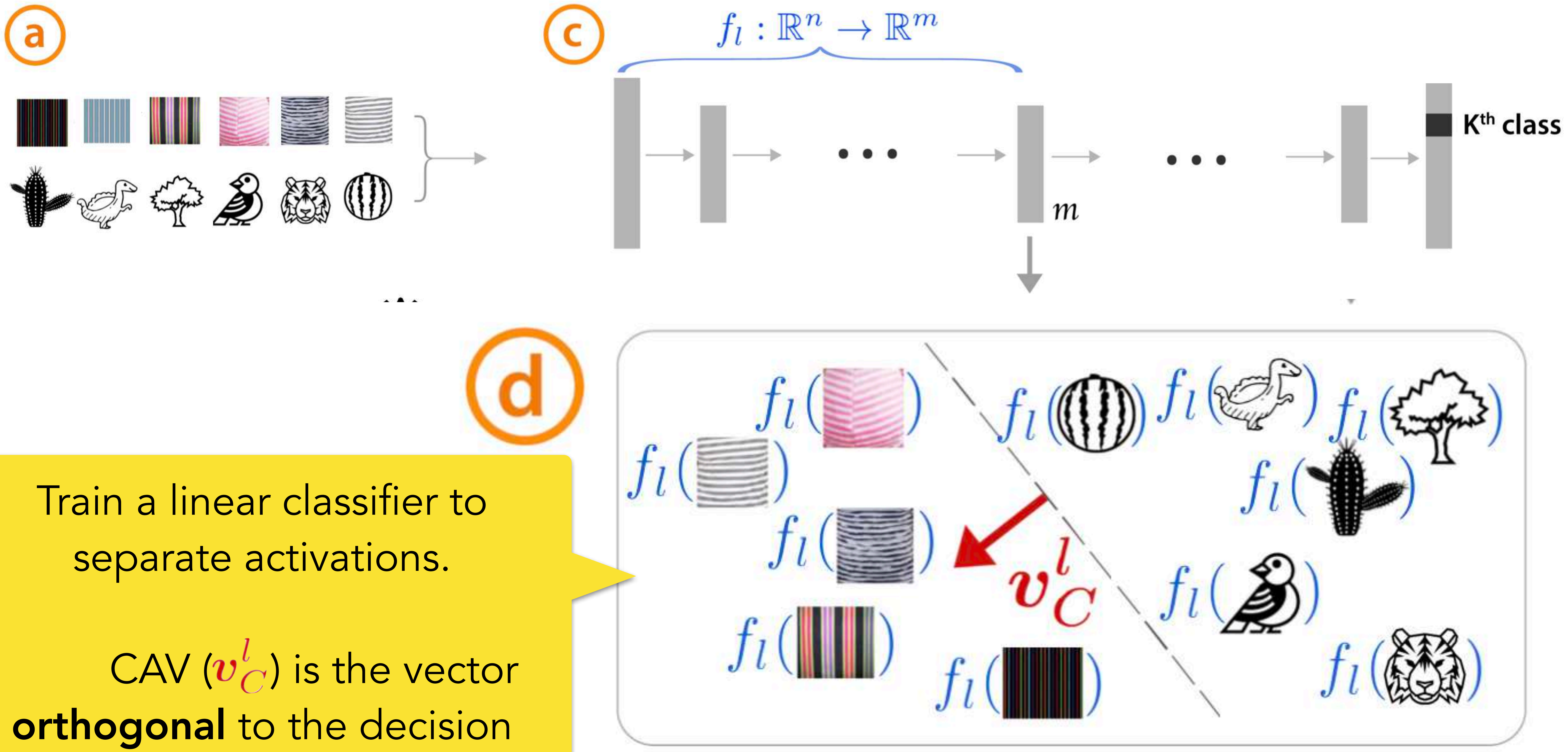
Random images



A trained network under investigation
and
Internal tensors

Defining concept activation vector (CAV)

Inputs:

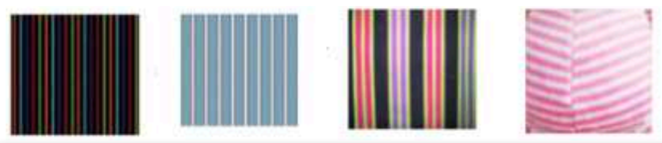


Train a linear classifier to separate activations.

CAV (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)

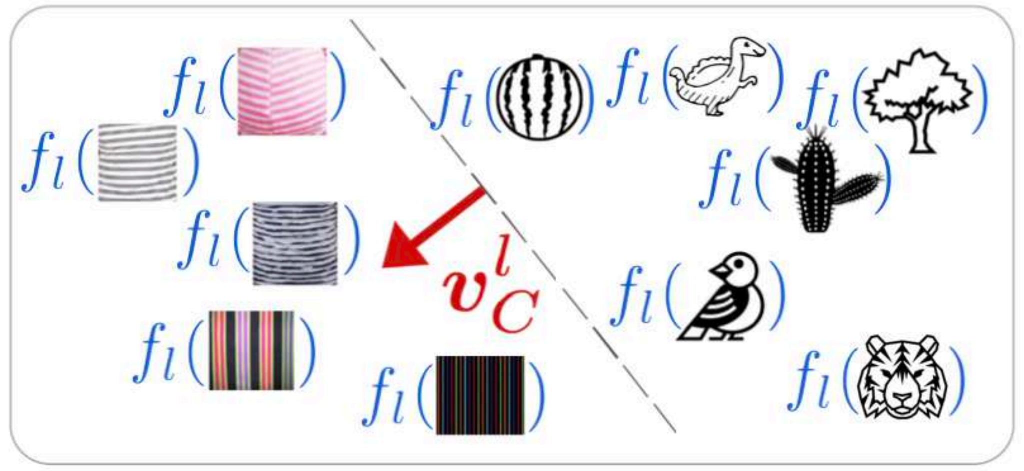
$p(z)$

zebra-ness

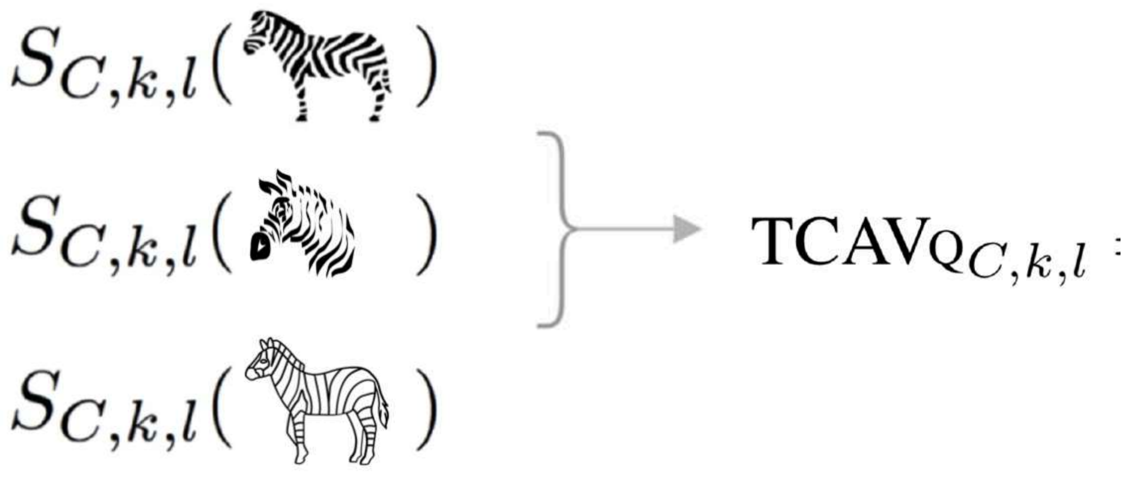
TCAV

How important was the striped concept to this zebra image classifier?

1. Learning CAVs



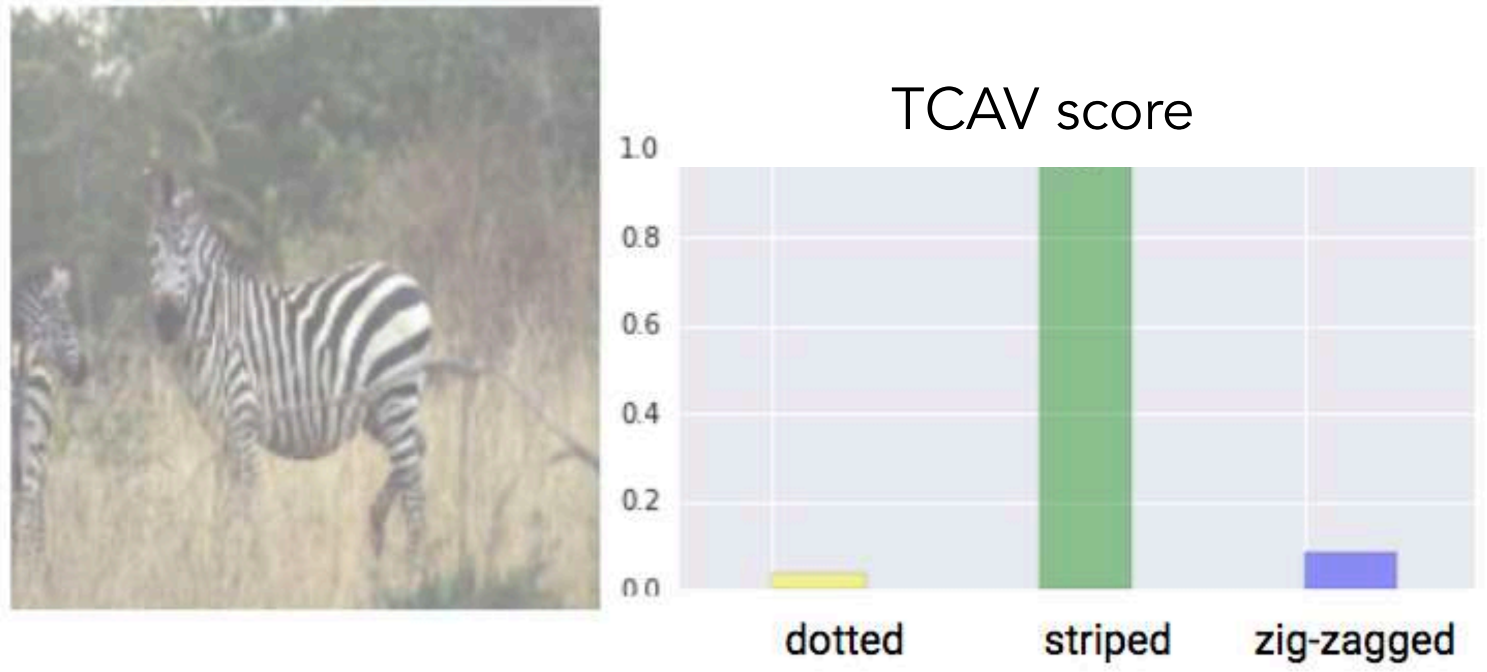
2. Getting TCAV score



2. How are the CAVs useful to get explanations?

TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV



zebra-ness →

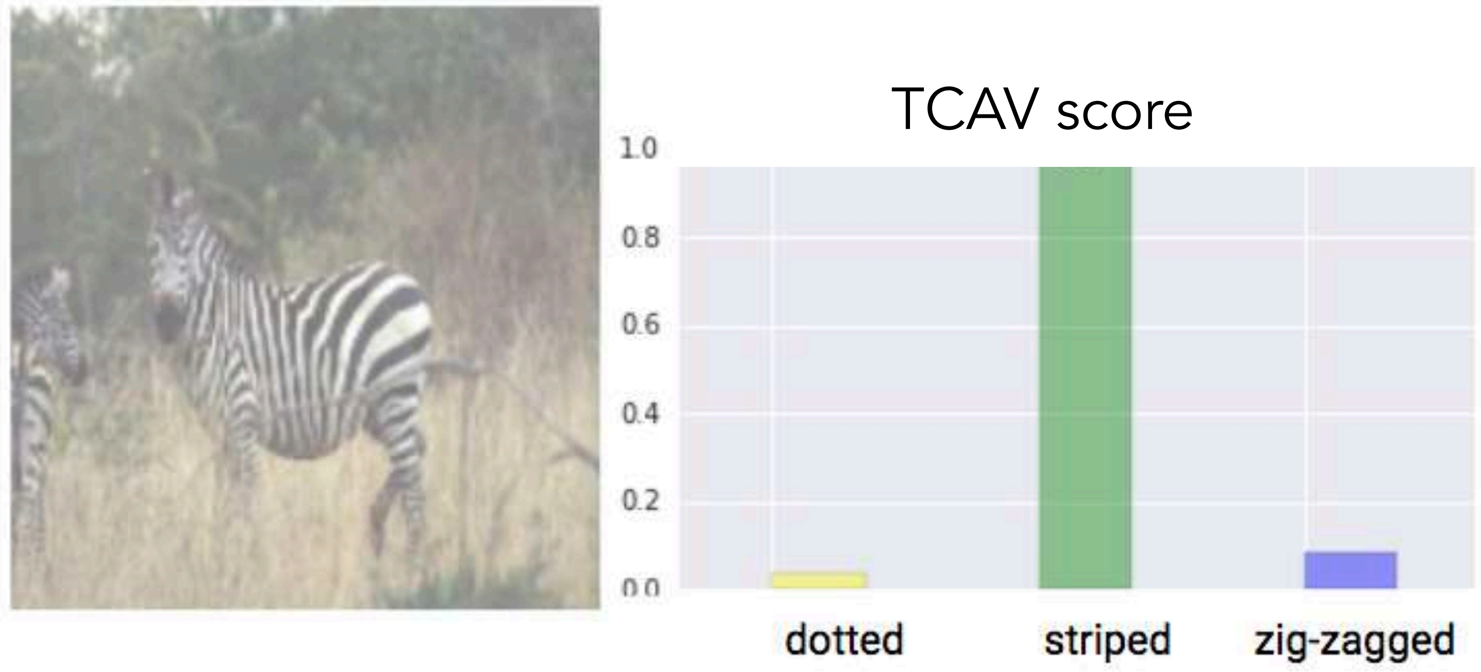
striped CAV →

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

Directional derivative with CAV

TCAV core idea:
Derivative with CAV to get prediction sensitivity

TCAV



zebra-ness →

striped CAV →

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

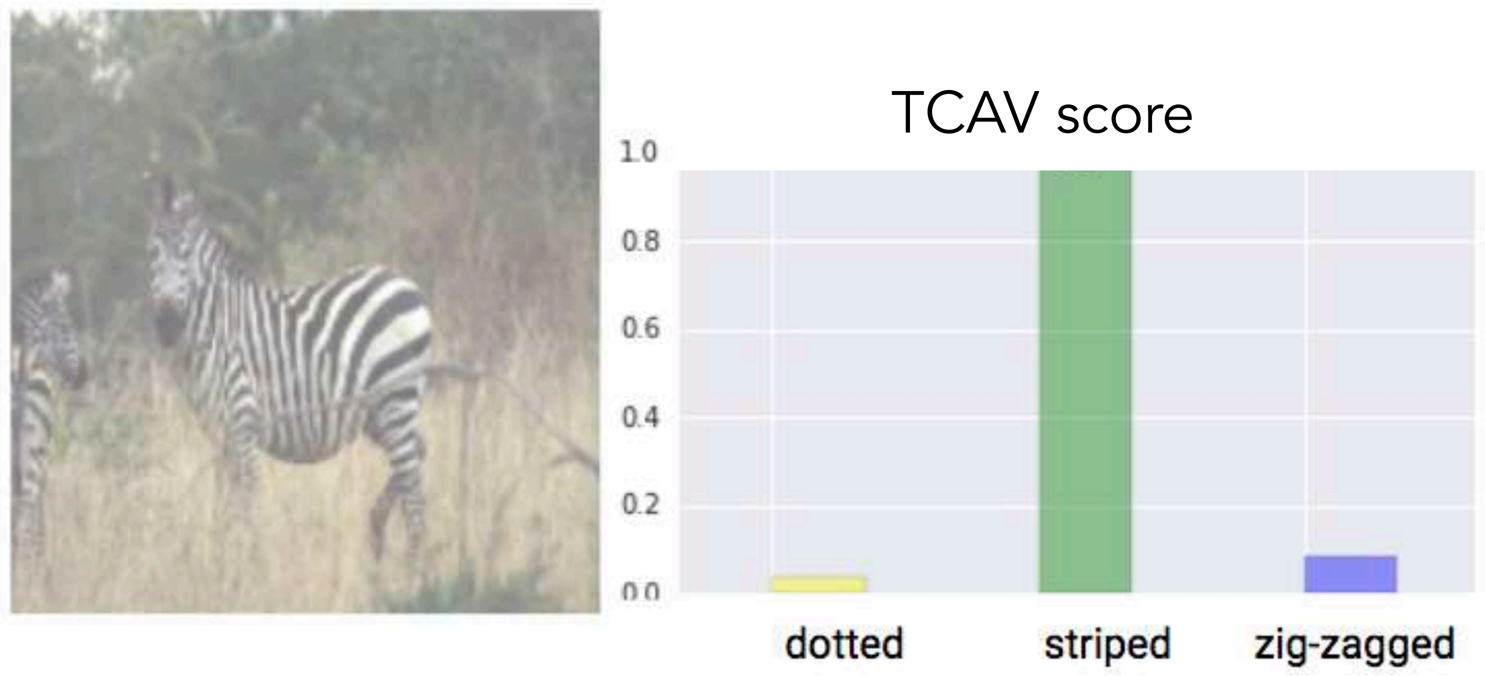
Directional derivative with CAV

One definition of explanation:

Tell me how **sensitive** the prediction is when we slightly **change** each concept.

TCAV core idea:
Derivative with CAV to get prediction sensitivity

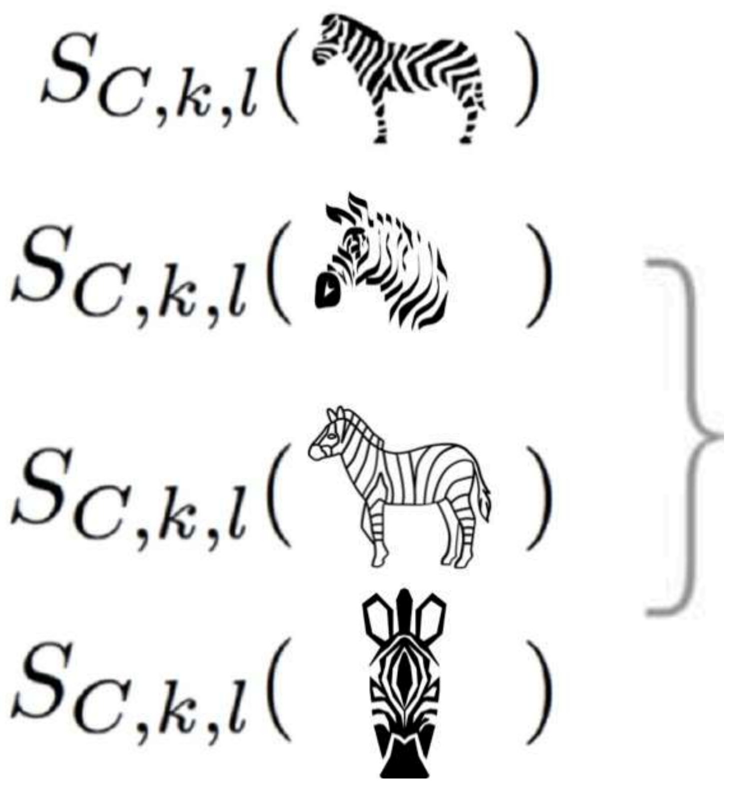
TCAV



zebra-ness →
striped CAV →

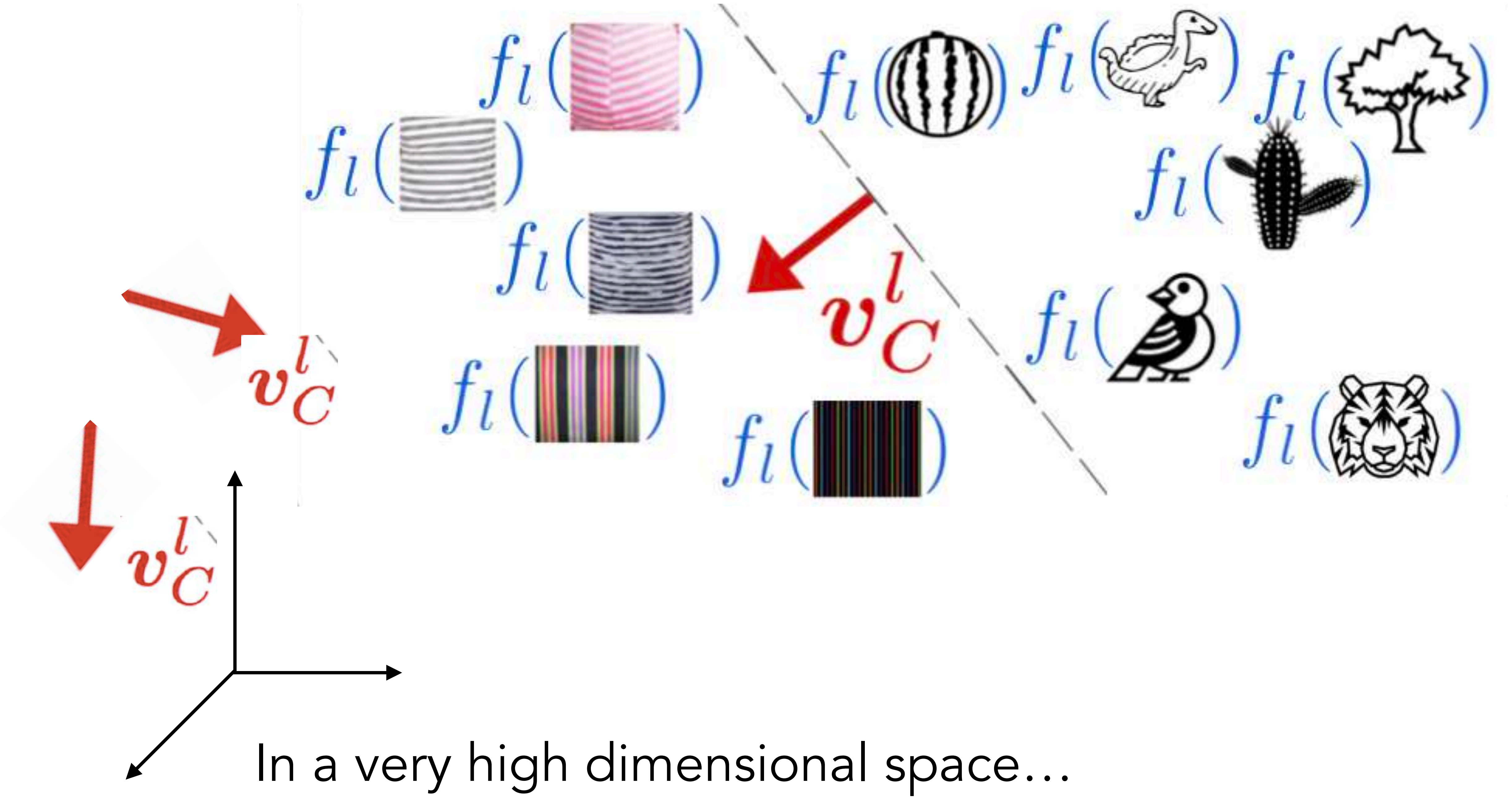
$$\frac{\partial p(z)}{\partial \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$$

Directional derivative with CAV



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

Is this CAV legit?



In a very high dimensional space…
funky things can happen.

Quantitative validation:

Guarding against spurious CAV

Zebra

→ $\mathrm{TCAVQ}_{C,k,l}$

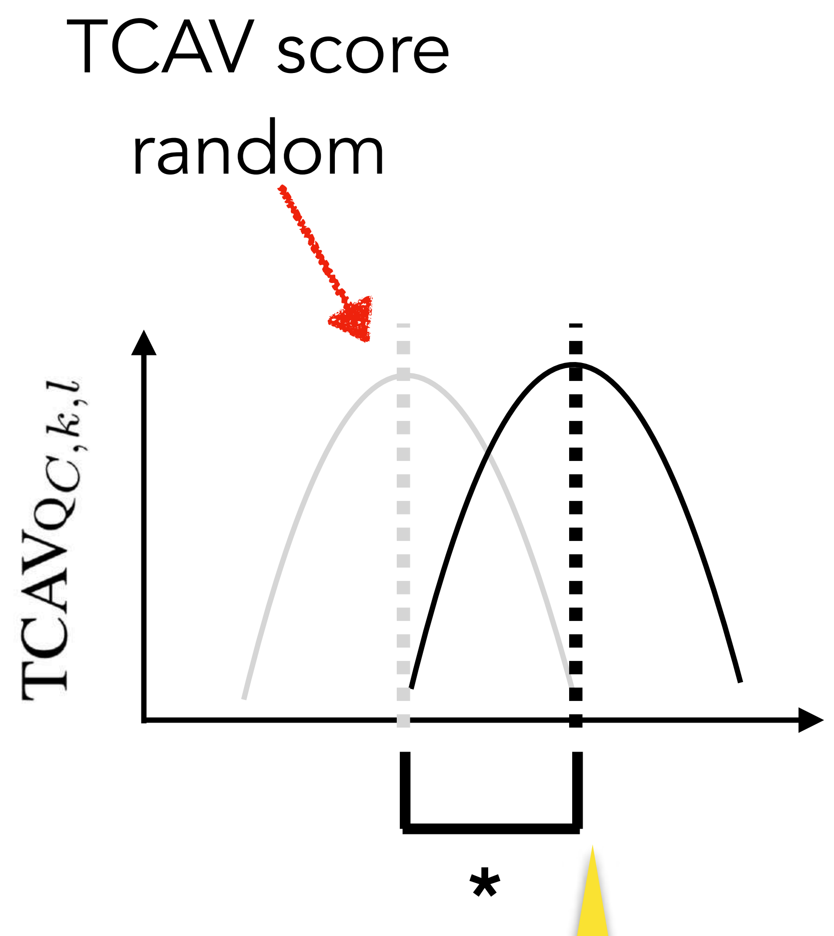
⋮

→ $\mathrm{TCAVQ}_{C,k,l}$

→ $\mathrm{TCAVQ}_{C,k,l}$

→ $\mathrm{TCAVQ}_{C,k,l}$

⋮

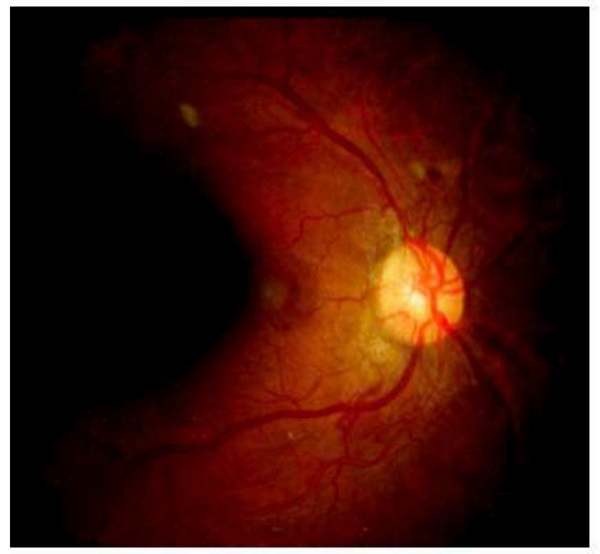


Check the distribution of $\mathrm{TCAVQ}_{C,k,l}$ is statistically different from random using t-test

Results

cab image
cab image with caption

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


DR level 4 Retina


TCAV for DR level 4

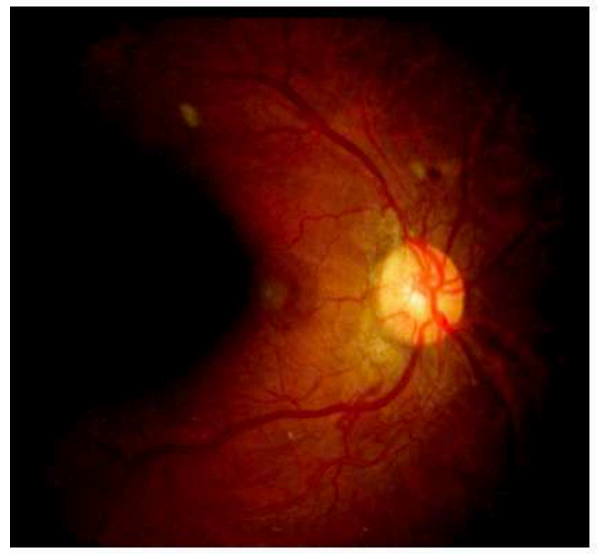
Real!

Results

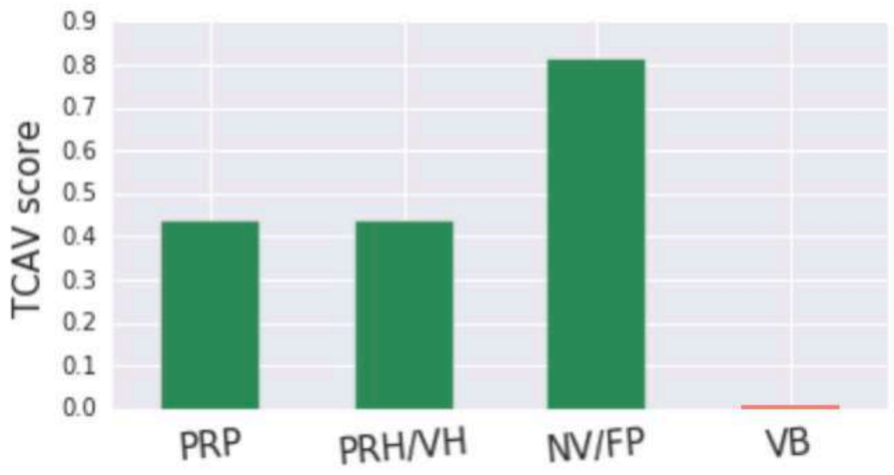
cab image

cab image with caption

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



DR level 4 Retina



TCAV for DR level 4

Global and Local Interpretability for Cardiac MRI Classification

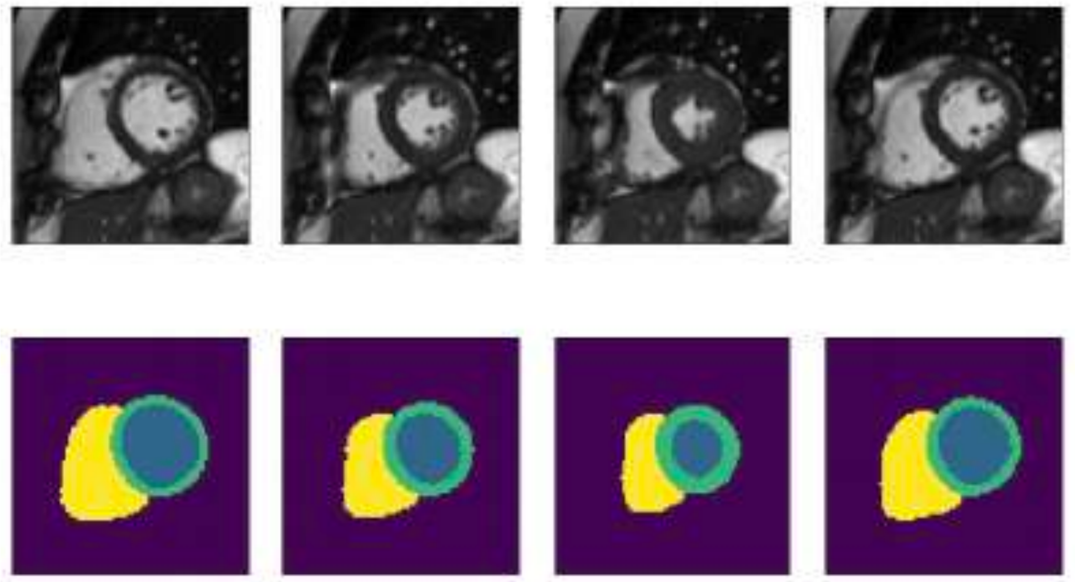
James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink,
Andrew P. King, and Julia A. Schnabel

School of Biomedical Engineering & Imaging Sciences, King's College London, UK
`james.clough@kcl.ac.uk`

CAV	Description	$\nabla \tilde{y} \cdot \mathbf{v}_c > 0$	$\langle \nabla \tilde{y} \cdot \mathbf{v}_c \rangle$
Low EF	Ejection Fraction	78.2%	0.0417
Low PER	Peak Ejection Rate	88.8%	0.0770
Low PFR	Peak Filling Rate	99.6%	0.1560
Low APFR	Atrial Peak Filling Rate	58.2%	0.0048
High LVT	Variance of LV wall thickening	63.4%	0.0156

Table 1: The sensitivity of the classifier to clinical biomarkers of poor cardiac health. A biomarker with no relevance would have $\nabla_{\mathbf{z}} \tilde{y} \cdot \mathbf{v}_c = 0$ on average.

Interpreting a jointly trained VAE+classification model.



Global and Local Interpretability for Cardiac MRI Classification

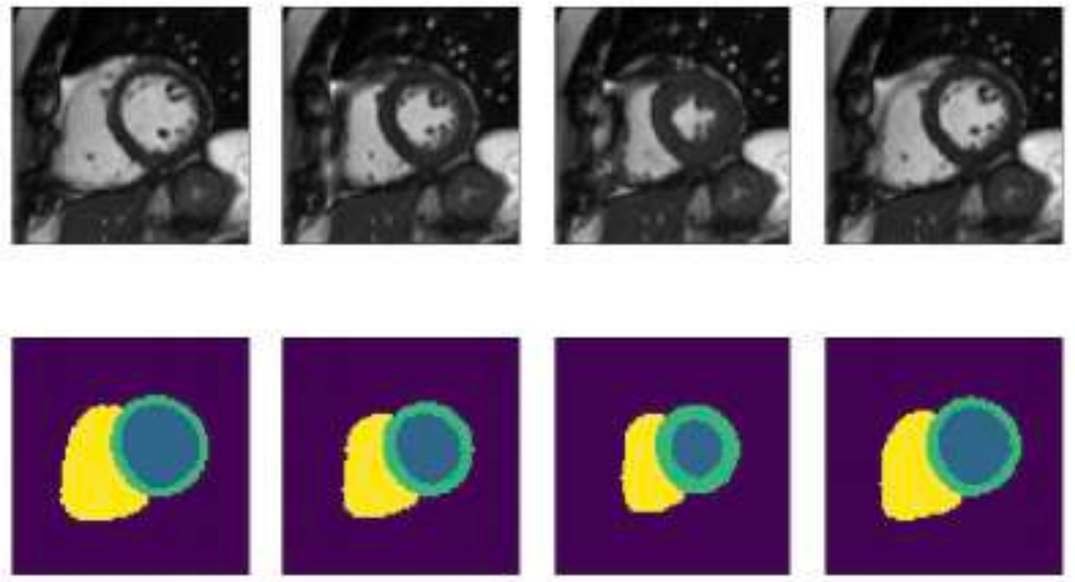
James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink,
Andrew P. King, and Julia A. Schnabel

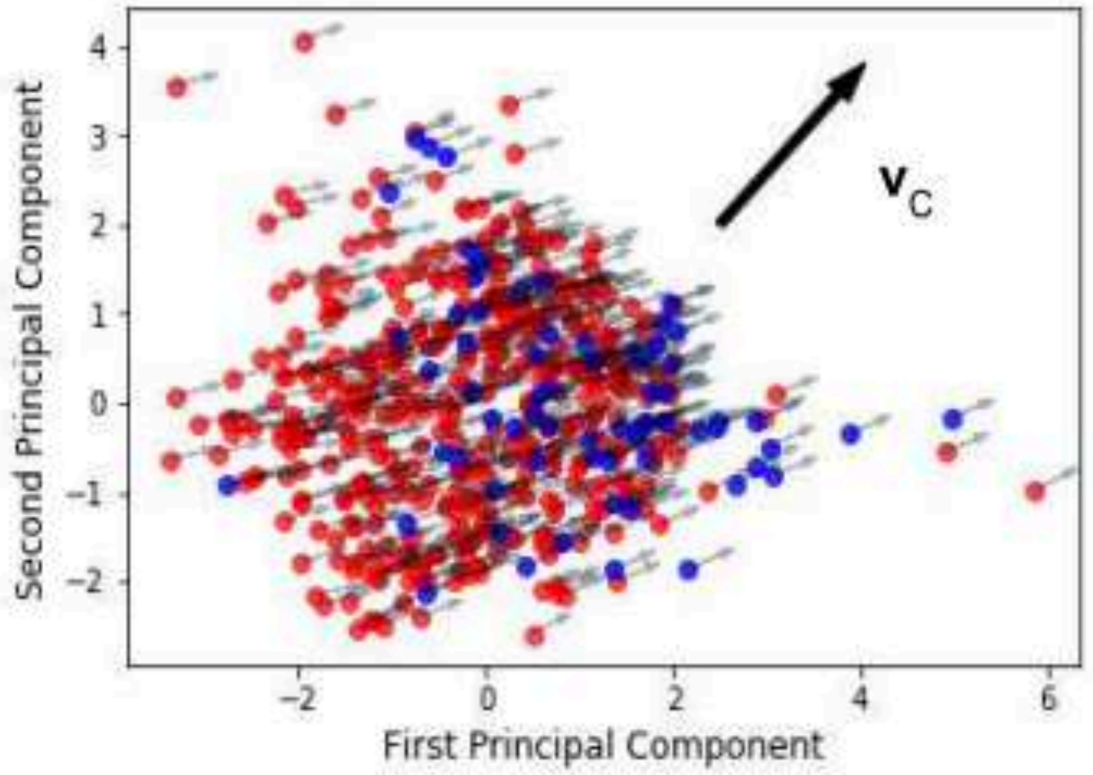
School of Biomedical Engineering & Imaging Sciences, King's College London, UK
james.clough@kcl.ac.uk

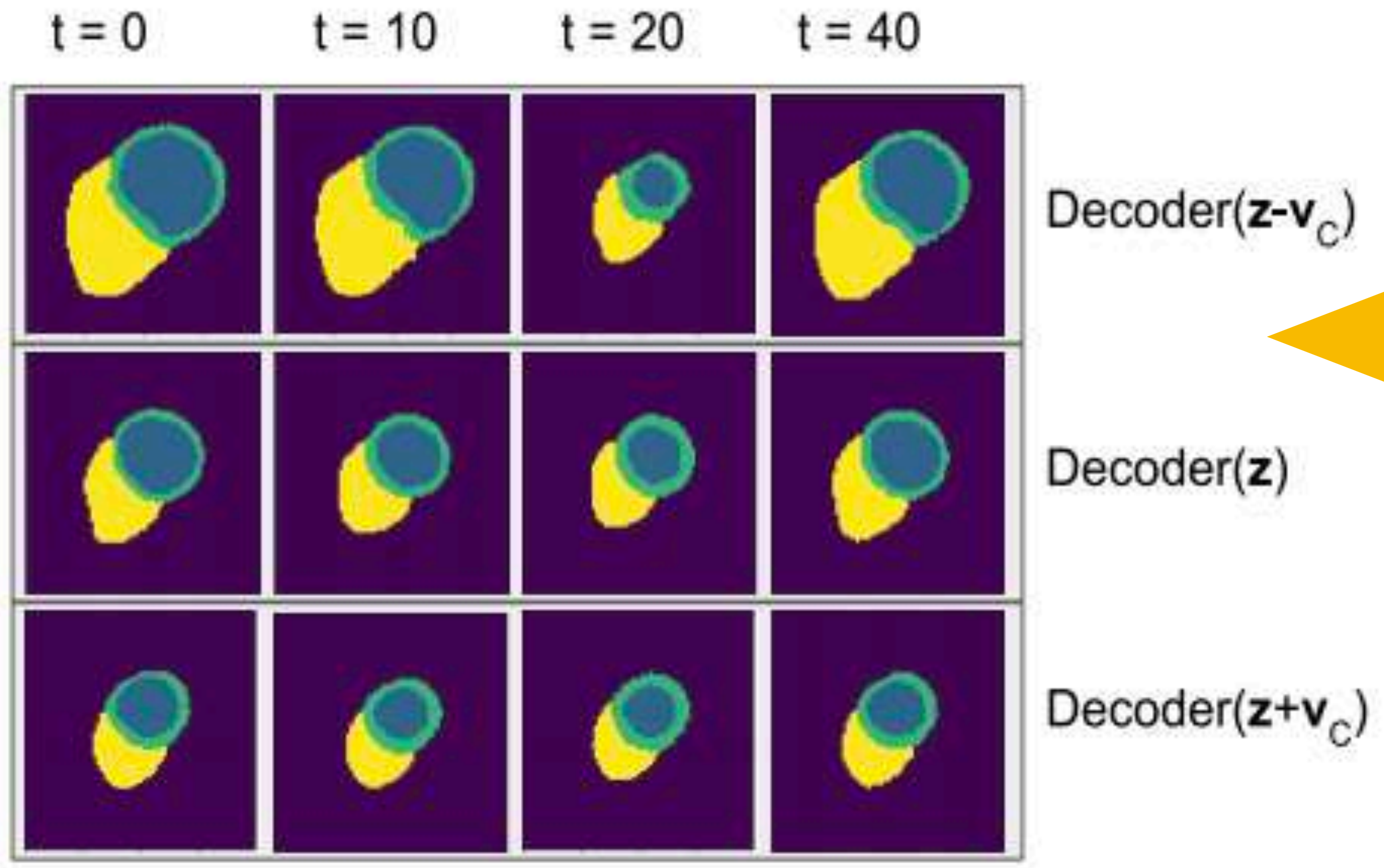
CAV	Description	$\nabla\tilde{y} \cdot \mathbf{v}_c > 0$	$\langle\nabla\tilde{y} \cdot \mathbf{v}_c\rangle$
Low EF	Ejection Fraction	78.2%	0.0417
Low PER	Peak Ejection Rate	88.8%	0.0770
Low PFR	Peak Filling Rate	99.6%	0.1560
Low APFR	Atrial Peak Filling Rate	58.2%	0.0048
High LVT	Variance of LV wall thickening	63.4%	0.0156

Table 1: The sensitivity of the classifier to clinical biomarkers of poor cardiac health. A biomarker with no relevance would have $\nabla_{\mathbf{z}}\tilde{y} \cdot \mathbf{v}_c = 0$ on average.

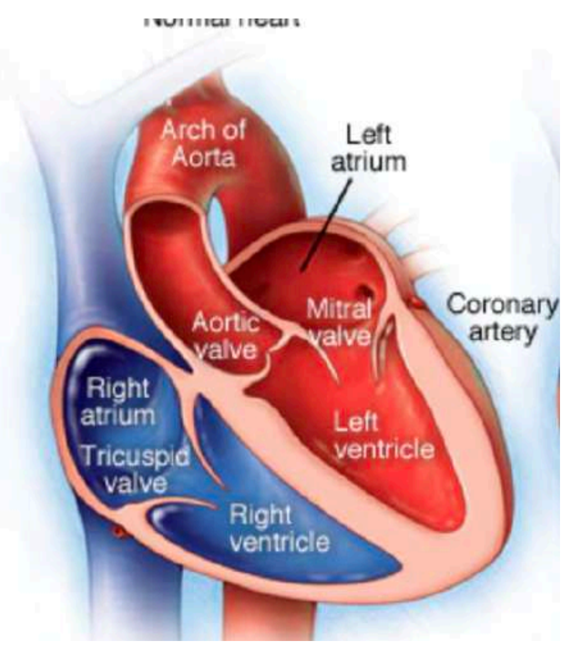
Interpreting a jointly trained VAE+classification model.







Can generate images with more/less
LV (left ventricle) concept

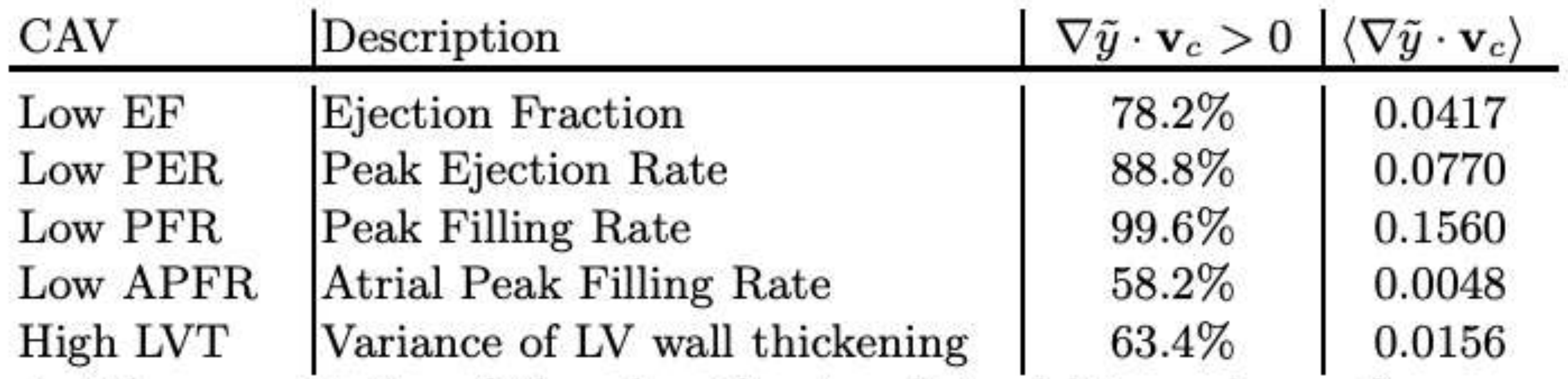


Global and Local Interpretability for Cardiac MRI Classification

James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink,
Andrew P. King, and Julia A. Schnabel

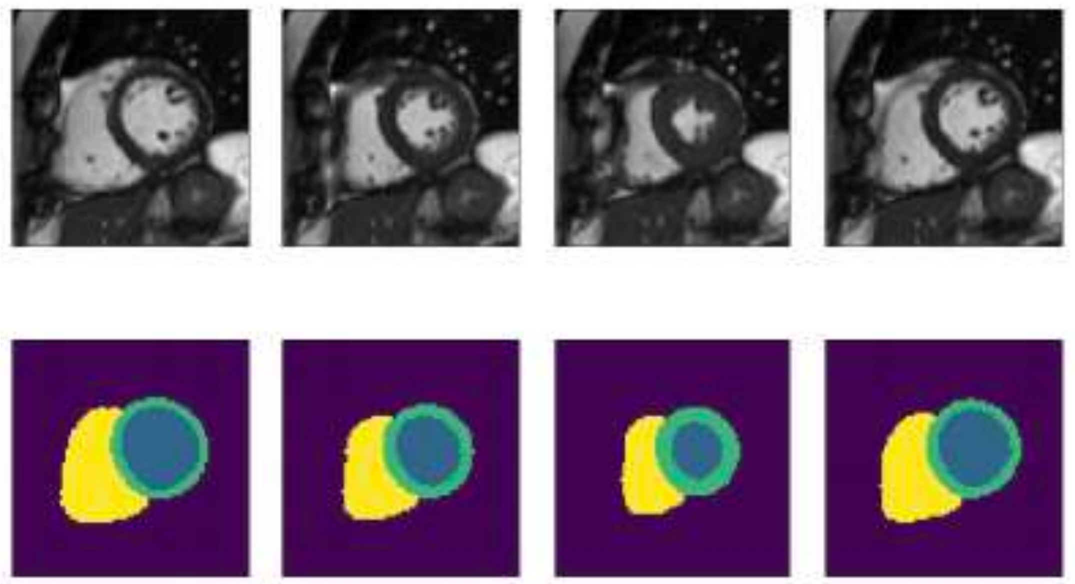
School of Biomedical Engineering & Imaging Sciences, King's College London, UK
`james.clough@kcl.ac.uk`

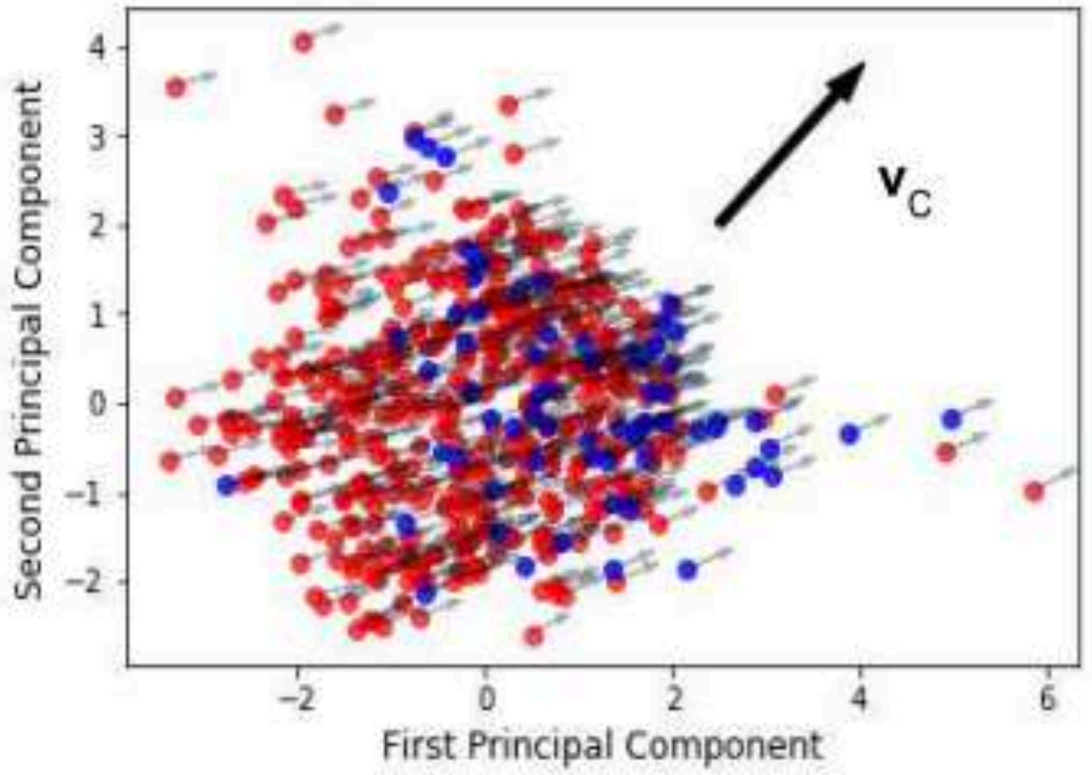
Interpreting a jointly trained VAE+classification model.

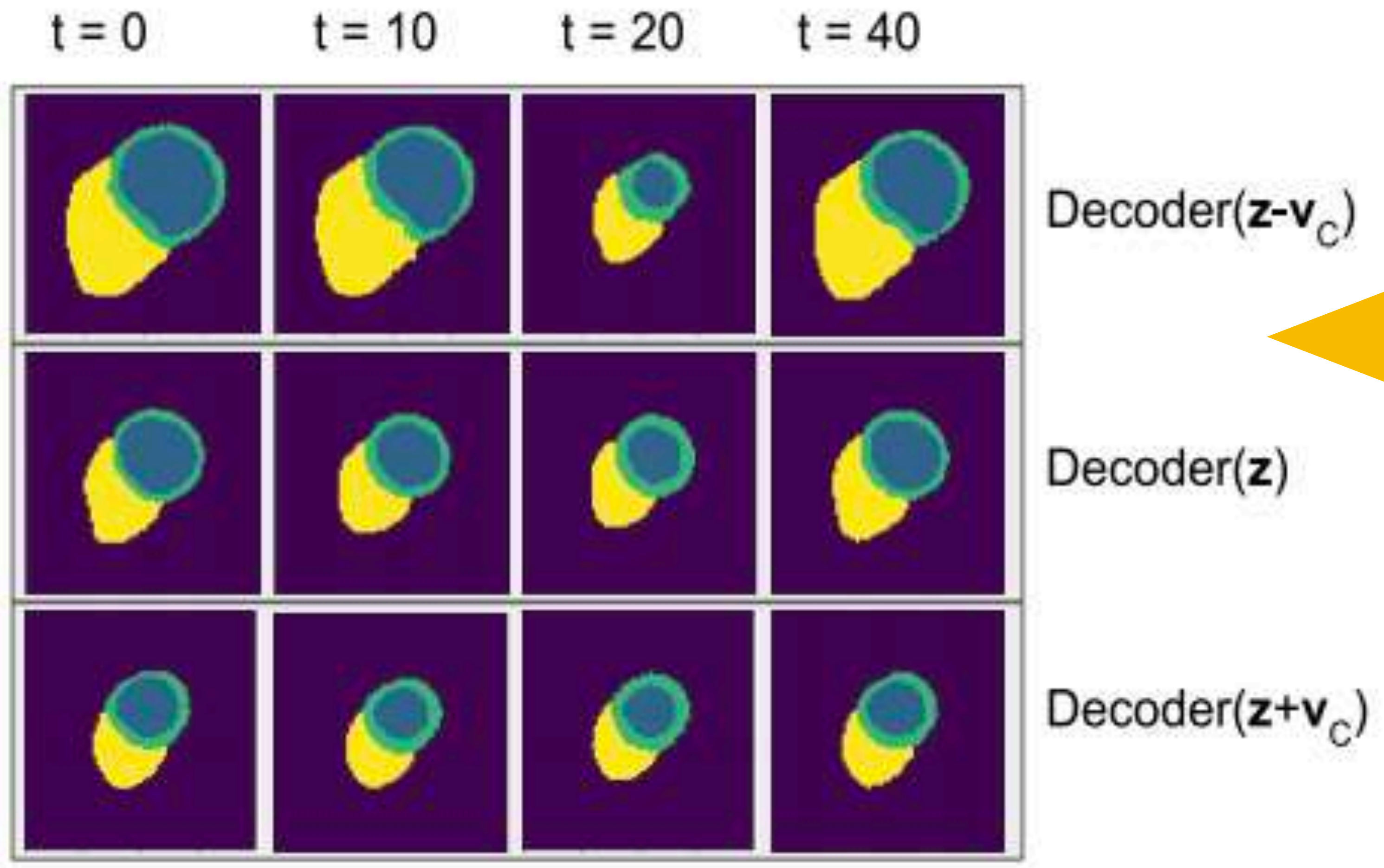


CAV	Description	$\nabla\tilde{y}\cdot\mathbf{v}_c > 0$	$\langle\nabla\tilde{y}\cdot\mathbf{v}_c\rangle$
Low EF	Ejection Fraction	78.2%	0.0417
Low PER	Peak Ejection Rate	88.8%	0.0770
Low PFR	Peak Filling Rate	99.6%	0.1560
Low APFR	Atrial Peak Filling Rate	58.2%	0.0048
High LVT	Variance of LV wall thickening	63.4%	0.0156

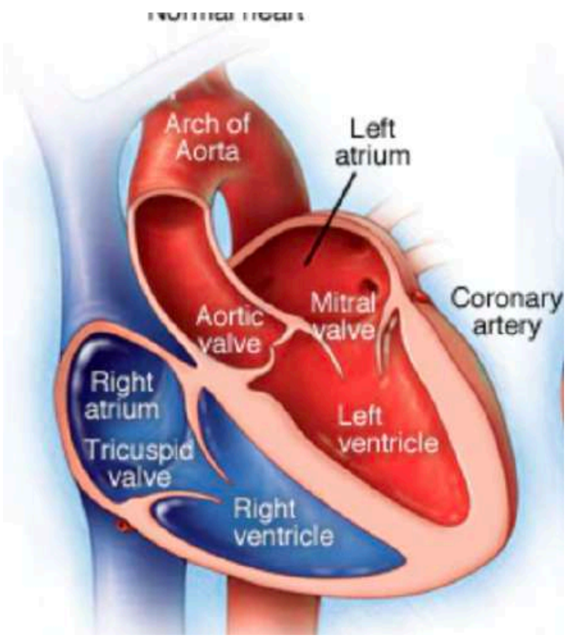
Table 1: The sensitivity of the classifier to clinical biomarkers of poor cardiac health. A biomarker with no relevance would have $\nabla_{\mathbf{z}}\tilde{y}\cdot\mathbf{v}_c = 0$ on average.

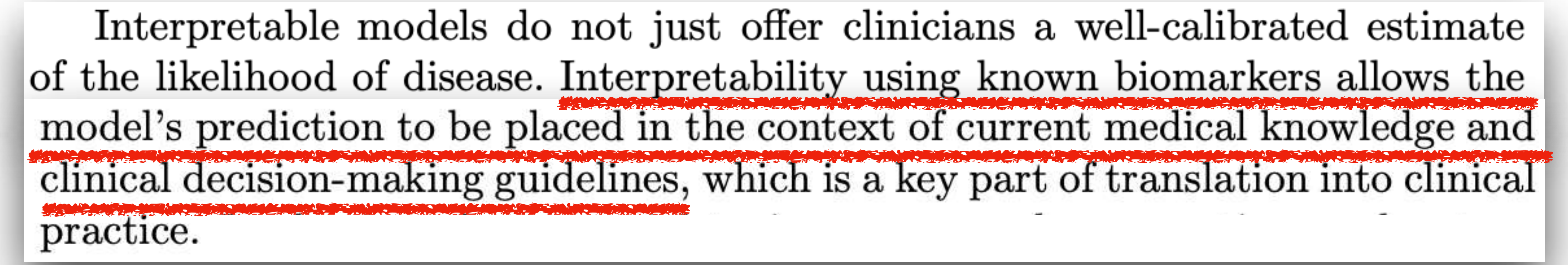




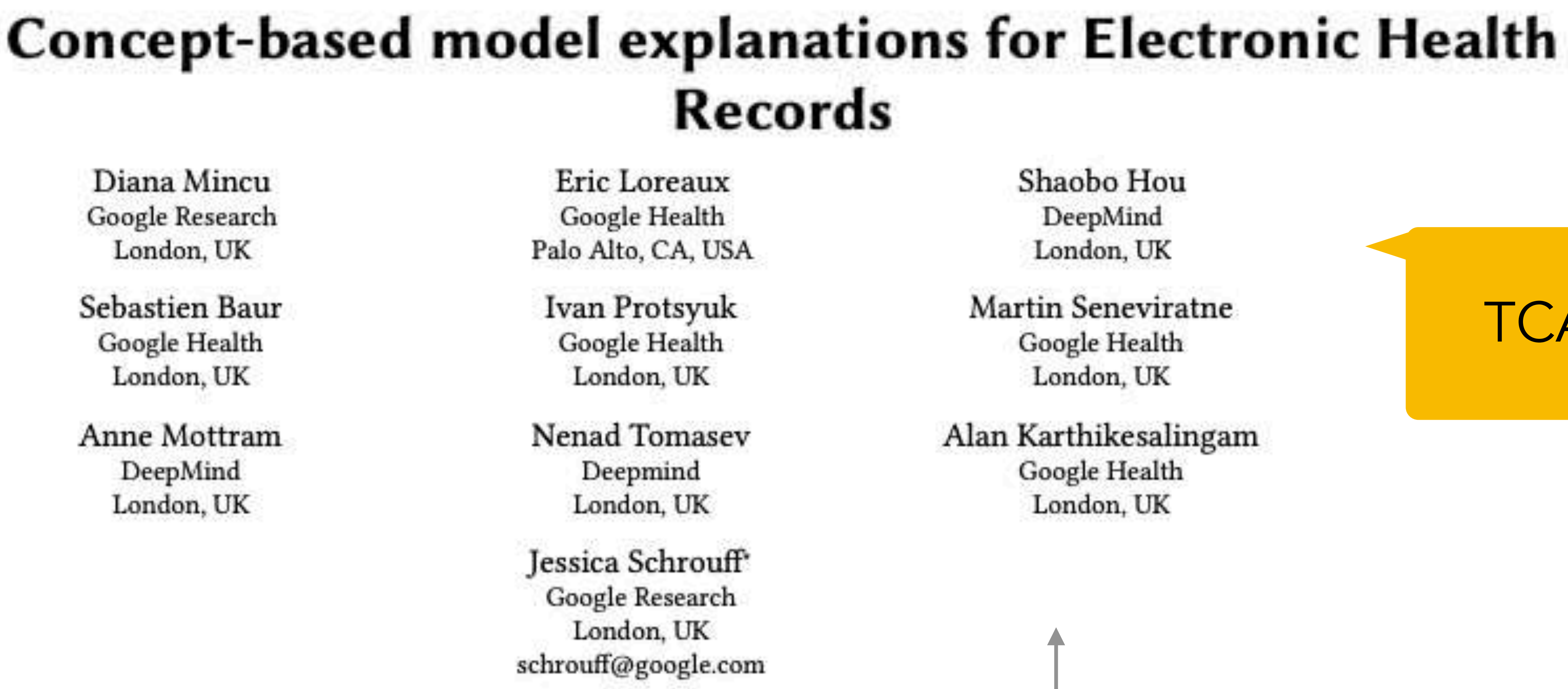


Can generate images with more/less
LV (left ventricle) concept





Interpretable models do not just offer clinicians a well-calibrated estimate of the likelihood of disease. Interpretability using known biomarkers allows the model's prediction to be placed in the context of current medical knowledge and clinical decision-making guidelines, which is a key part of translation into clinical practice.



Concept-based model explanations for Electronic Health Records

Diana Mincu
Google Research
London, UK

Eric Loreaux
Google Health
Palo Alto, CA, USA

Shaobo Hou
DeepMind
London, UK

Sebastien Baur
Google Health
London, UK

Ivan Protsyuk
Google Health
London, UK

Martin Seneviratne
Google Health
London, UK

Anne Mottram
DeepMind
London, UK

Nenad Tomasev
Deepmind
London, UK

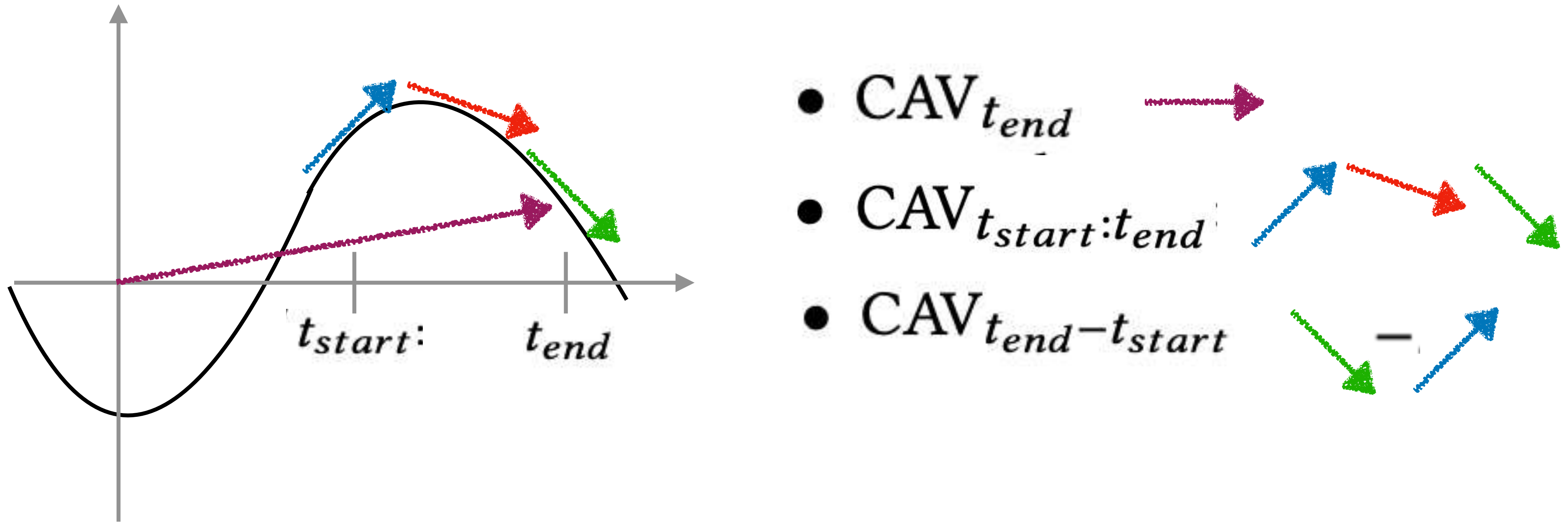
Alan Karthikesalingam
Google Health
London, UK

Jessica Schrouff*
Google Research
London, UK
schrouff@google.com

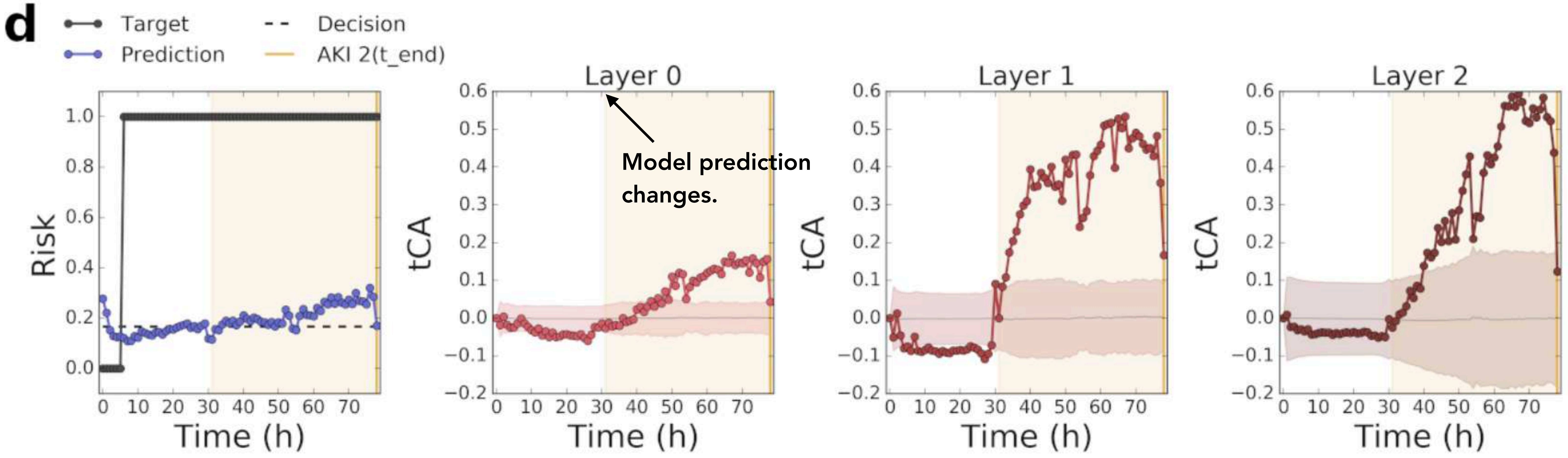
TCAV for RNNs

Presence of a concept in one data point

$$\mathrm{tCA}_C(\vec{x}_t) = \frac{\vec{a}_t^T}{||\vec{a}_t||_2}\vec{v}_C$$



- $\mathrm{CAV}_{t_{end}}$
- $\mathrm{CAV}_{t_{start}:t_{end}}$
- $\mathrm{CAV}_{t_{end}-t_{start}}$



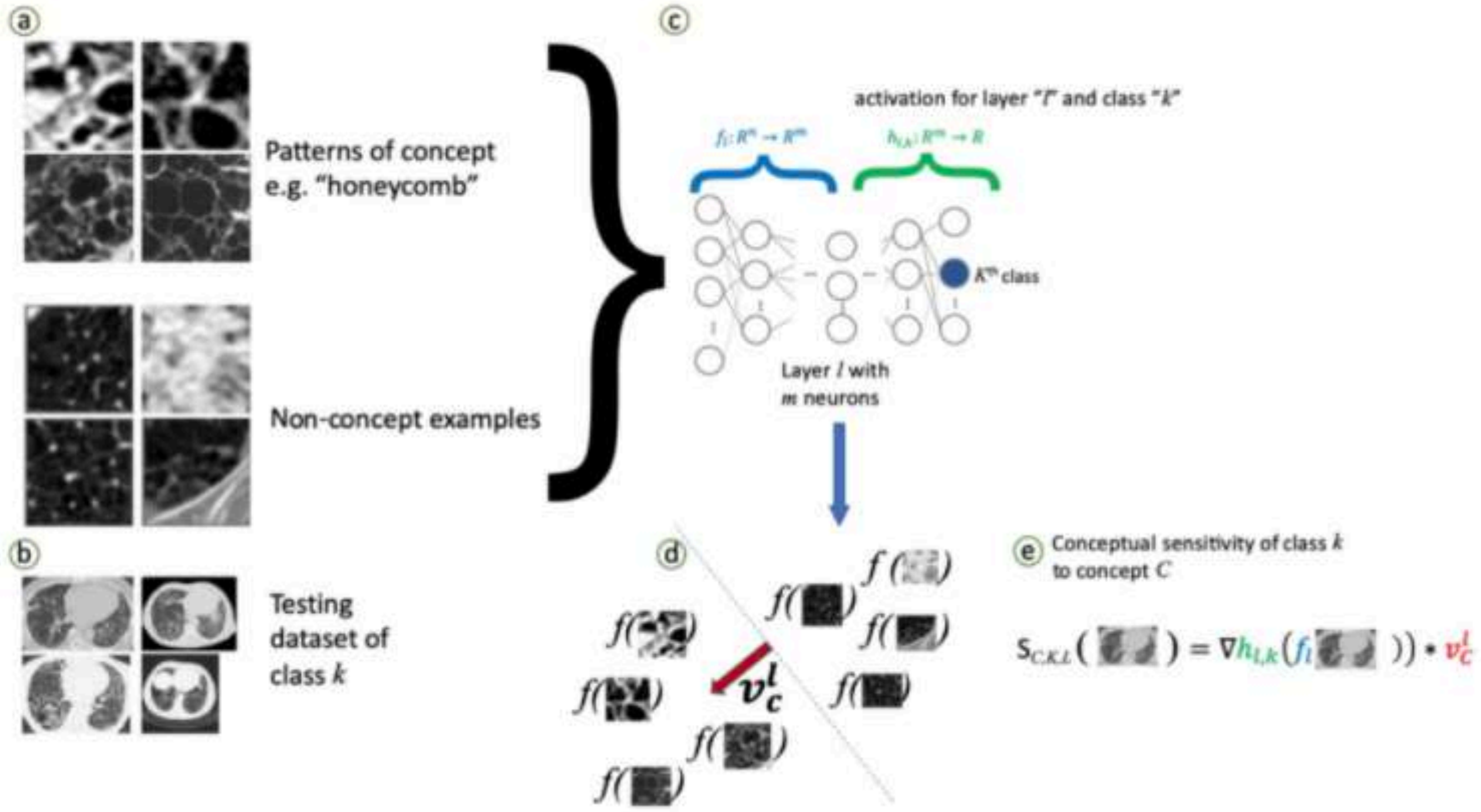
Model prediction changes.

github.com/google/ehr-predictions/tree/master/tcav-for-ehr

Radiology:Artificial Intelligence

On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, Roland Wiest



UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics

Authors Authors and affiliations

Hugo Yeche, Justin Harrison, Tess Berthier

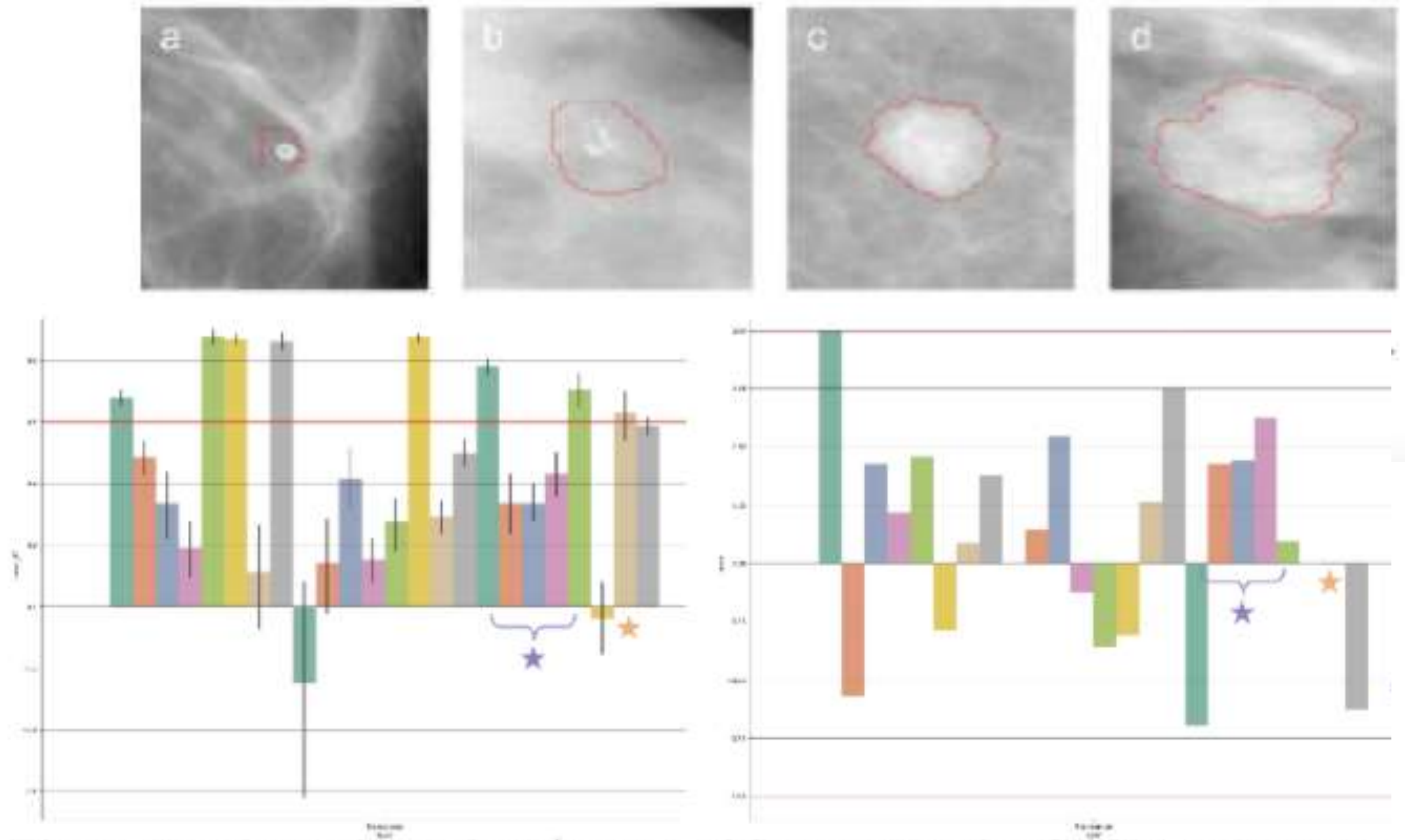
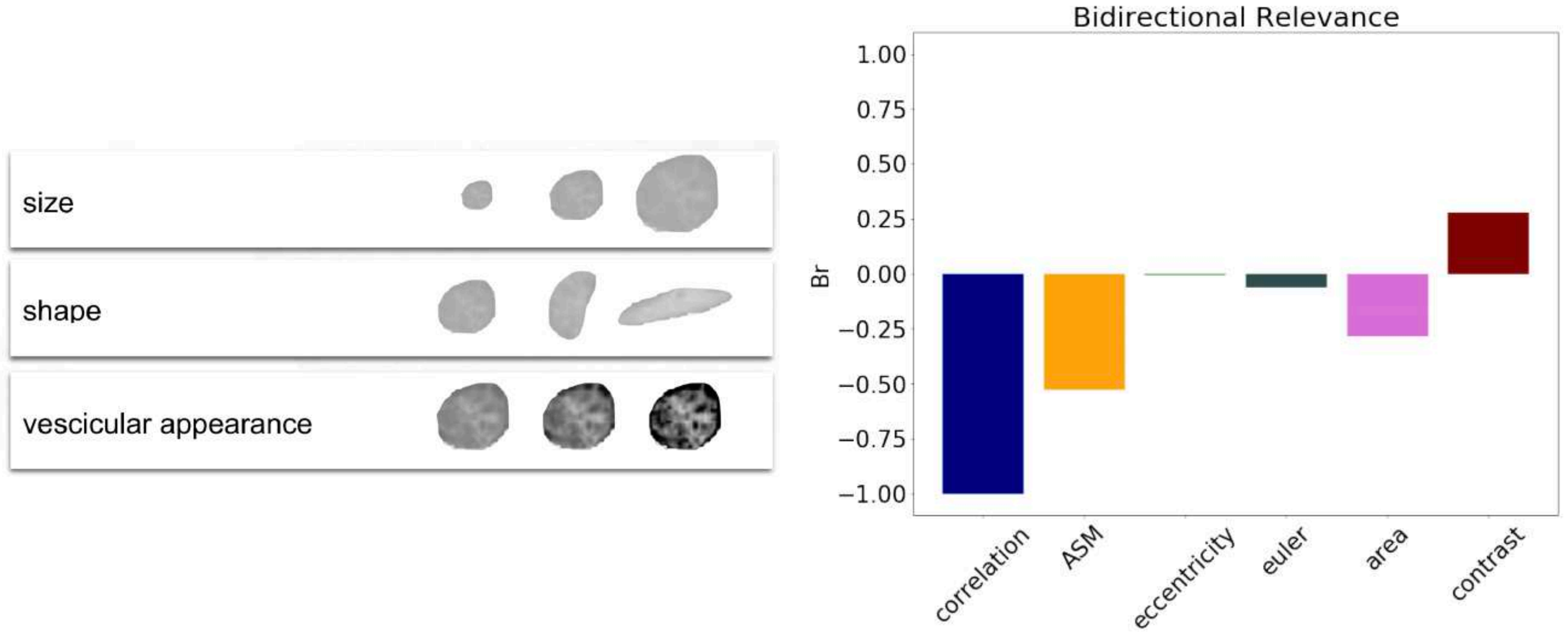


Fig. 4. Results at layer fire6/concat of SqueezeNet for GLCM radiomics,
sponding R^2 scores (left) and Br scores (right) for calcification prediction.

[Submitted on 9 Apr 2019]

Regression Concept Vectors for Bidirectional Explanations in Histopathology

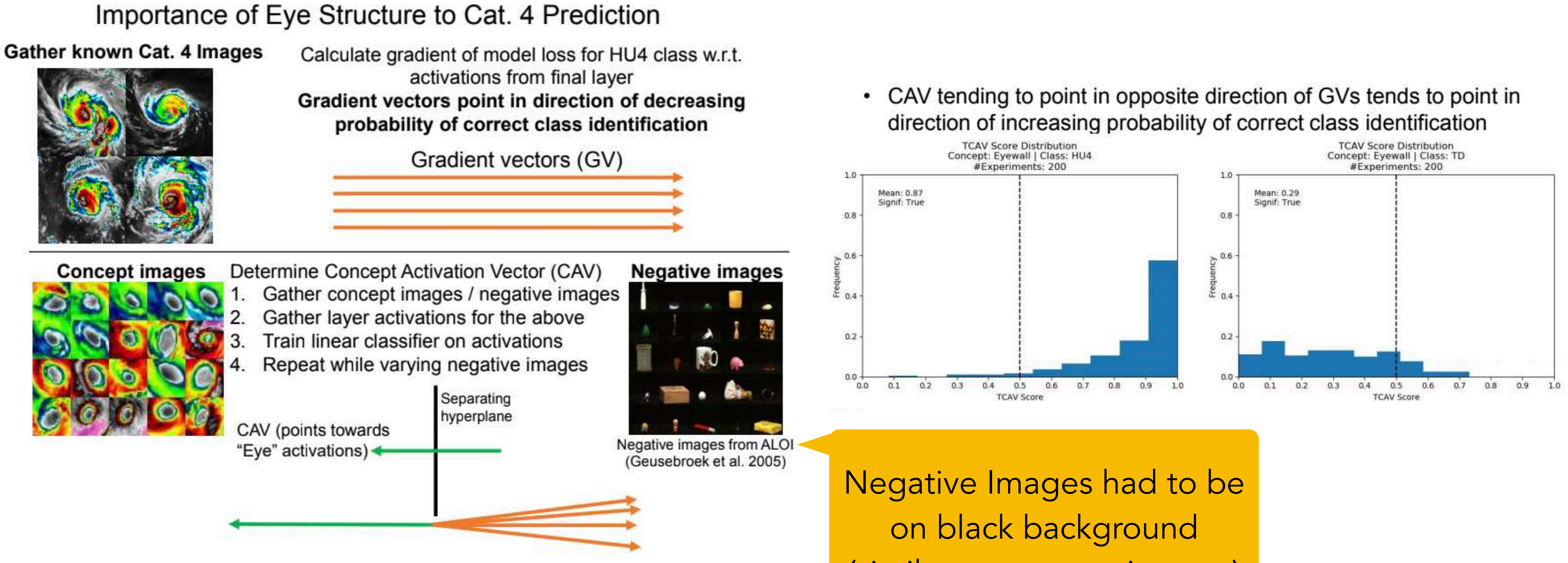
Mara Graziani, Vincent Andrearczyk, Henning Müller



Interpretable AI for Deep-Learning-Based Meteorological Applications

Eric B. Wendoloski, Ingrid C. Guch

The Aerospace Corporation



Importance of Eye Structure to Cat. 4 Prediction

Gather known Cat. 4 Images

Calculate gradient of model loss for HU4 class w.r.t. activations from final layer

Gradient vectors point in direction of decreasing probability of correct class identification

Gradient vectors (GV)

Concept images

Determine Concept Activation Vector (CAV)

1. Gather concept images / negative images
2. Gather layer activations for the above
3. Train linear classifier on activations
4. Repeat while varying negative images

Negative images

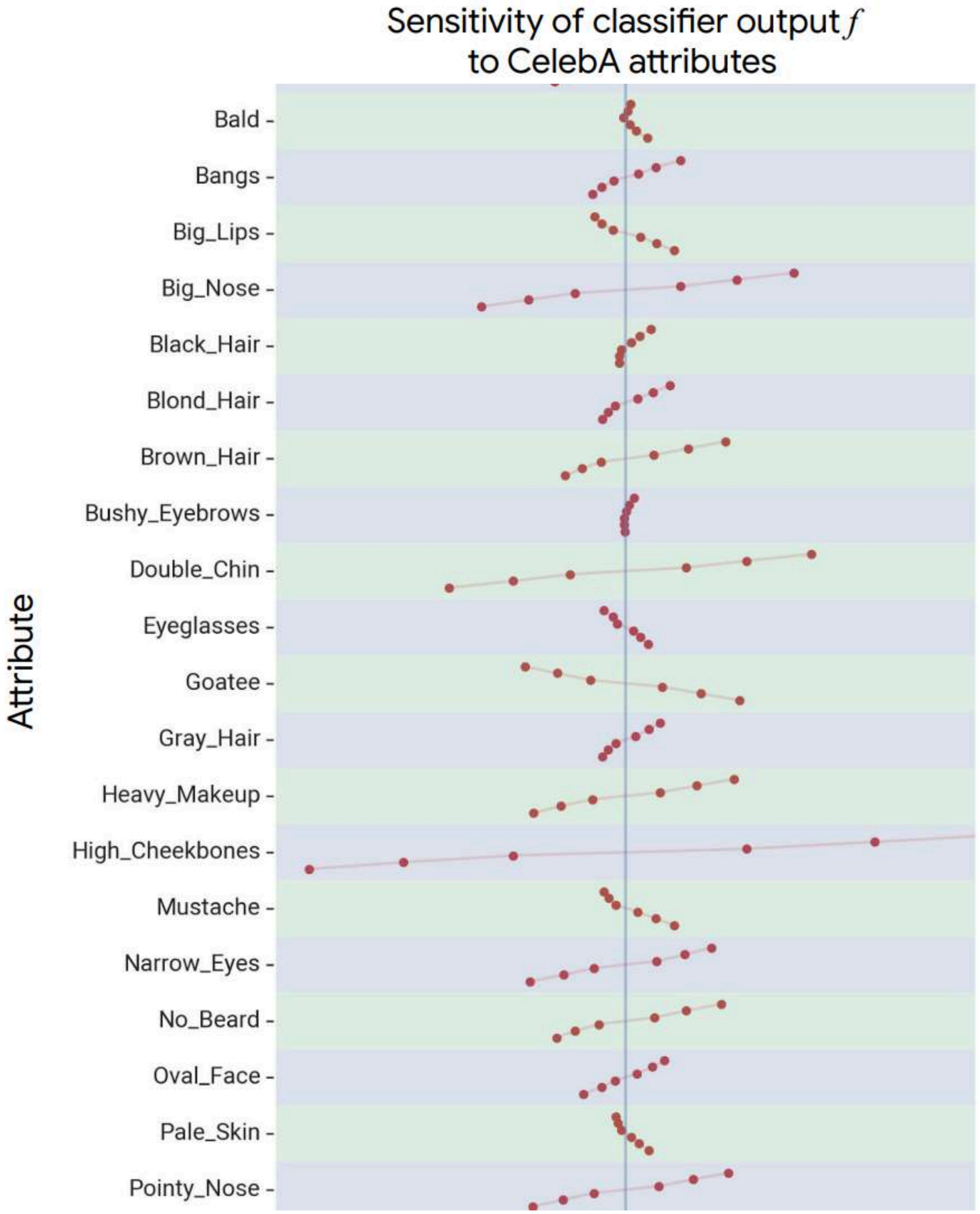
Negative images from ALOI (Geusebroek et al. 2005)

- CAV tending to point in opposite direction of GVs tends to point in direction of increasing probability of correct class identification

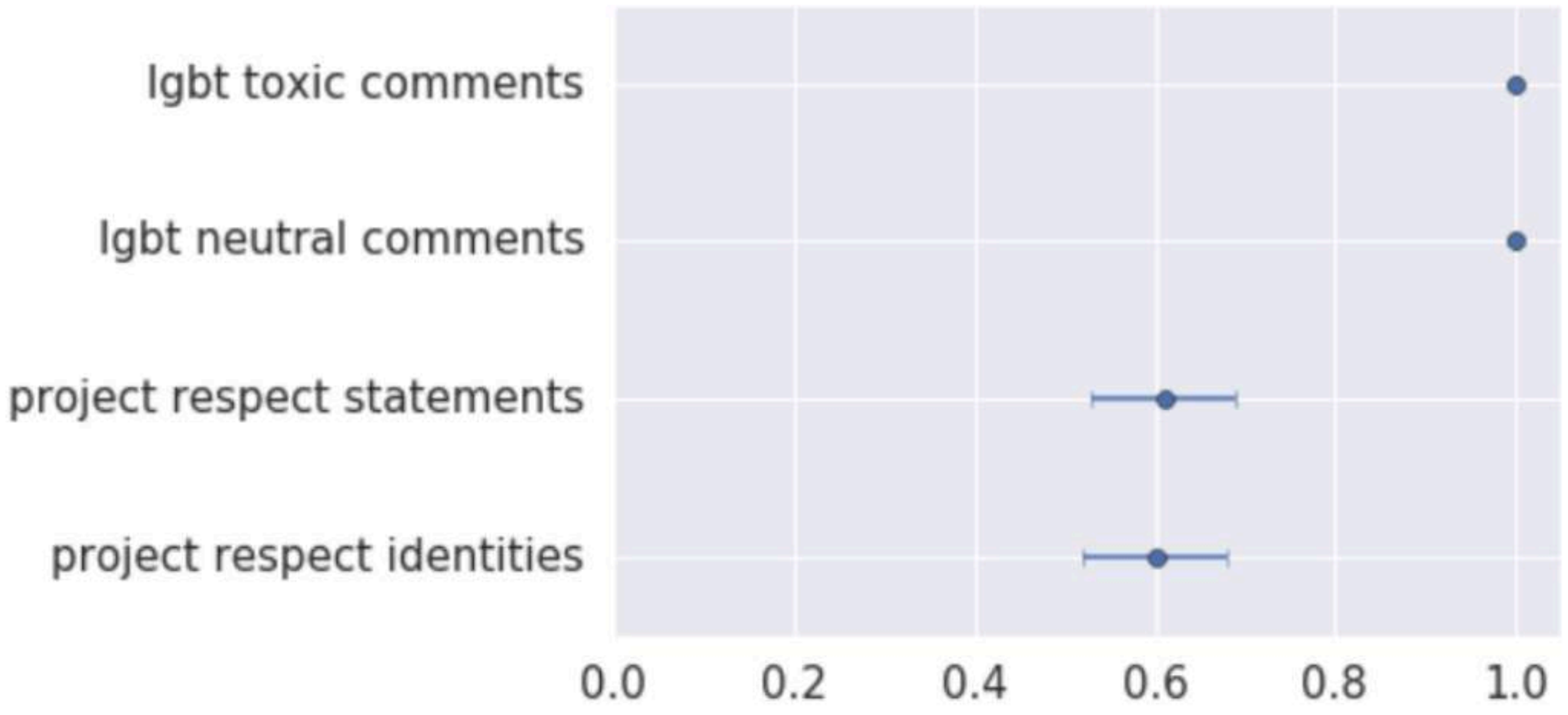
Negative Images had to be on black background (similar to concept images)

Follow up work on ML fairness @Google with non-image data

Language model

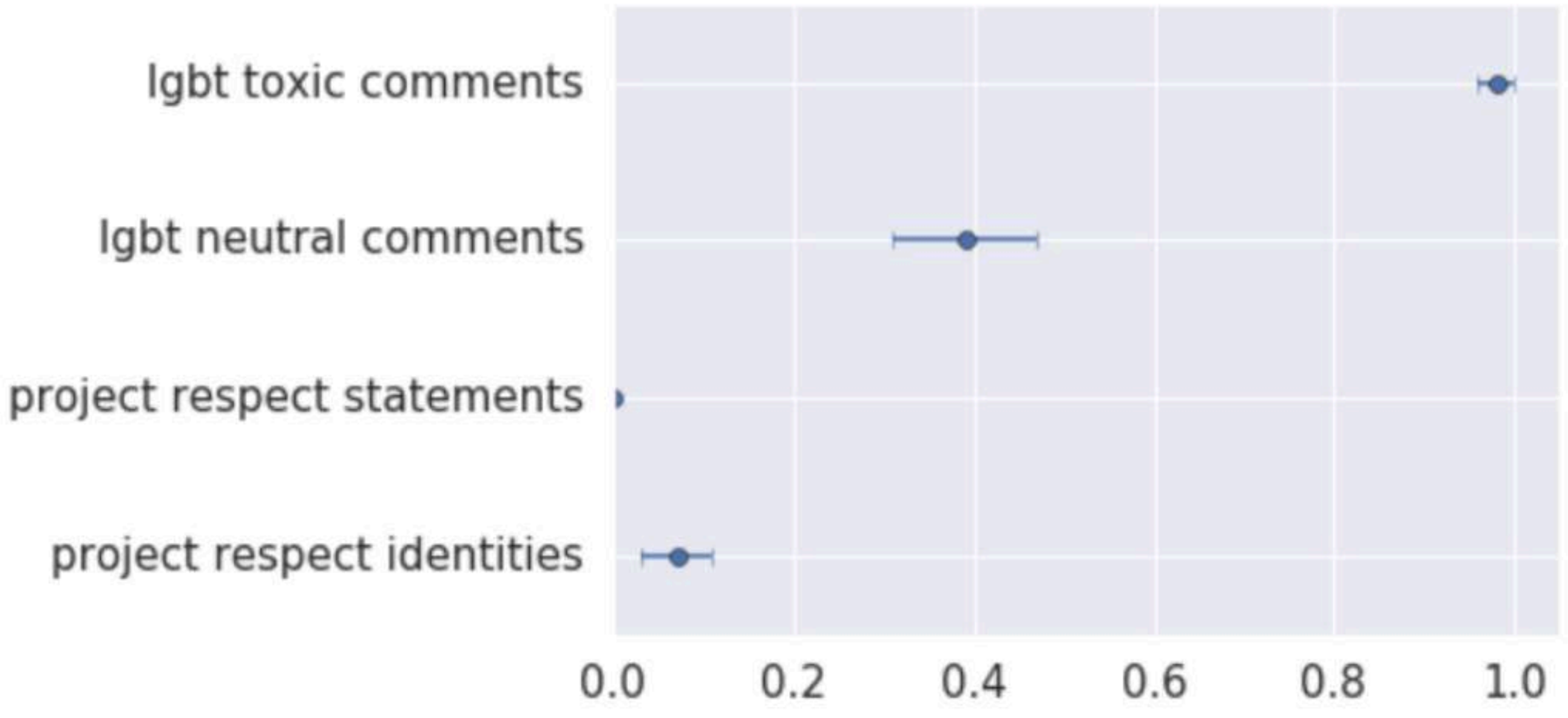


[Denton et al. 19]



(a) TOXICITY@1

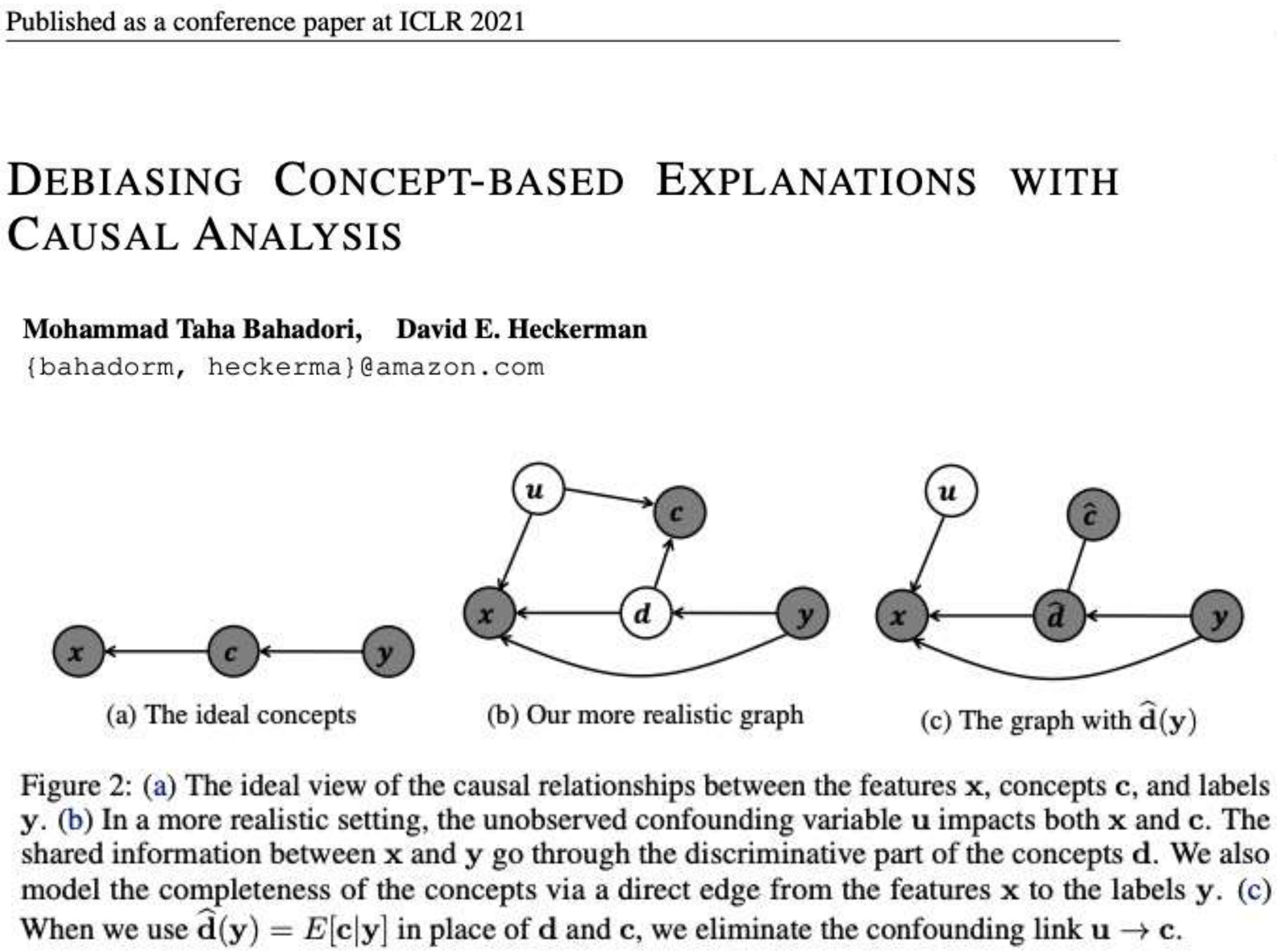
↓ Unbiasing



(b) TOXICITY@6

[Hutchinson et al. 19]

What if concepts are confounded/overlap?



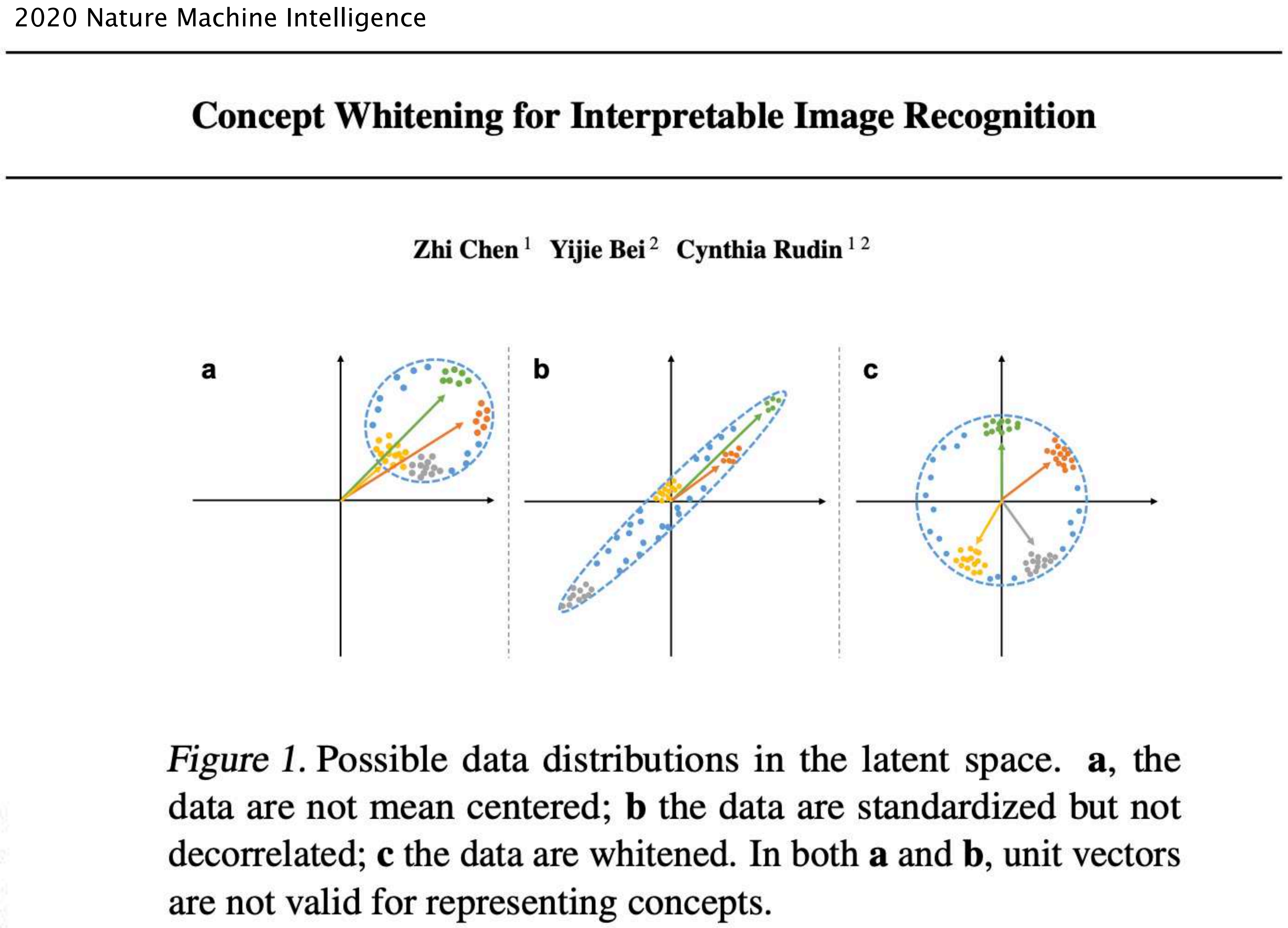
Published as a conference paper at ICLR 2021

DEBIASING CONCEPT-BASED EXPLANATIONS WITH CAUSAL ANALYSIS

Mohammad Taha Bahadori, David E. Heckerman
{bahadorm, heckerma}@amazon.com

(a) The ideal concepts (b) Our more realistic graph (c) The graph with $\widehat{\mathbf{d}}(\mathbf{y})$

Figure 2: (a) The ideal view of the causal relationships between the features $\mathbf{x}$, concepts $\mathbf{c}$, and labels $\mathbf{y}$. (b) In a more realistic setting, the unobserved confounding variable $\mathbf{u}$ impacts both $\mathbf{x}$ and $\mathbf{c}$. The shared information between $\mathbf{x}$ and $\mathbf{y}$ go through the discriminative part of the concepts $\mathbf{d}$. We also model the completeness of the concepts via a direct edge from the features $\mathbf{x}$ to the labels $\mathbf{y}$. (c) When we use $\widehat{\mathbf{d}}(\mathbf{y}) = E[\mathbf{c}|\mathbf{y}]$ in place of $\mathbf{d}$ and $\mathbf{c}$, we eliminate the confounding link $\mathbf{u} \rightarrow \mathbf{c}$.



2020 Nature Machine Intelligence

Concept Whitening for Interpretable Image Recognition

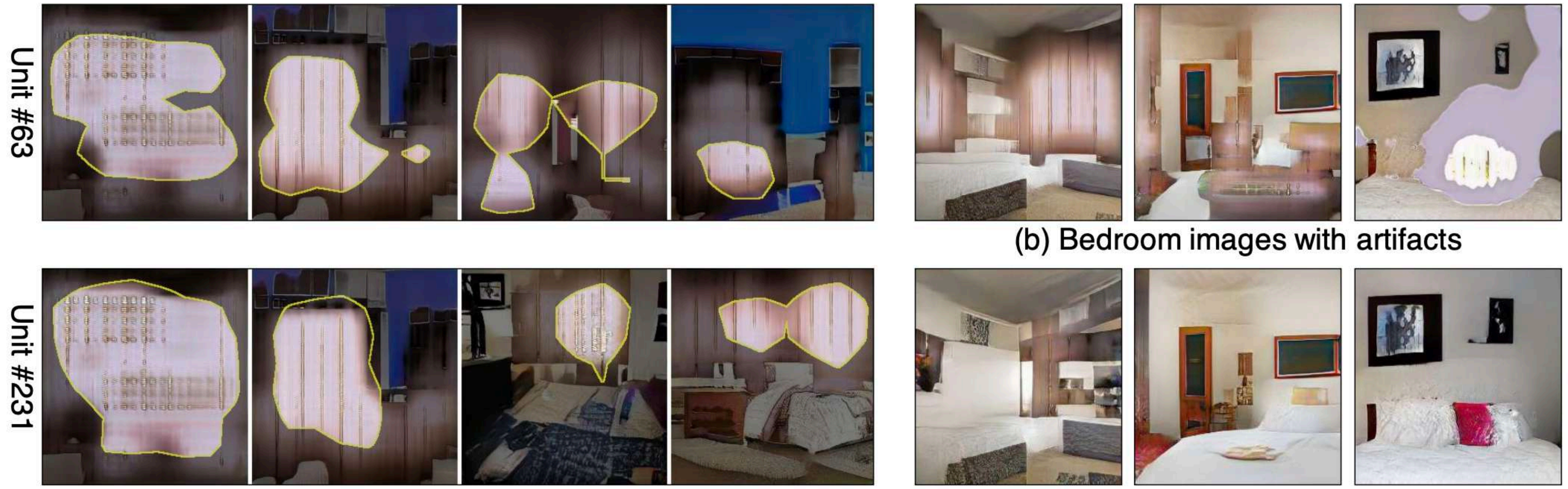
Zhi Chen [1] **Yijie Bei** [2] **Cynthia Rudin** [1 2]

Figure 1. Possible data distributions in the latent space. **a**, the data are not mean centered; **b** the data are standardized but not decorrelated; **c** the data are whitened. In both **a** and **b**, unit vectors are not valid for representing concepts.

Debugging GAN with concepts

GAN DISSECTION: VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS

David Bau[1,2], **Jun-Yan Zhu**[1], **Hendrik Strobelt**[2,3], **Bolei Zhou**[4], **Joshua B. Tenenbaum**[1], **William T. Freeman**[1], **Antonio Torralba**[1,2]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab,
[3]IBM Research, [4]The Chinese University of Hong Kong

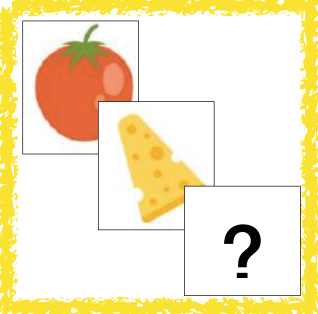


(b) Bedroom images with artifacts

(a) Example artifact-causing units

(c) Ablating "artifact" units improves results

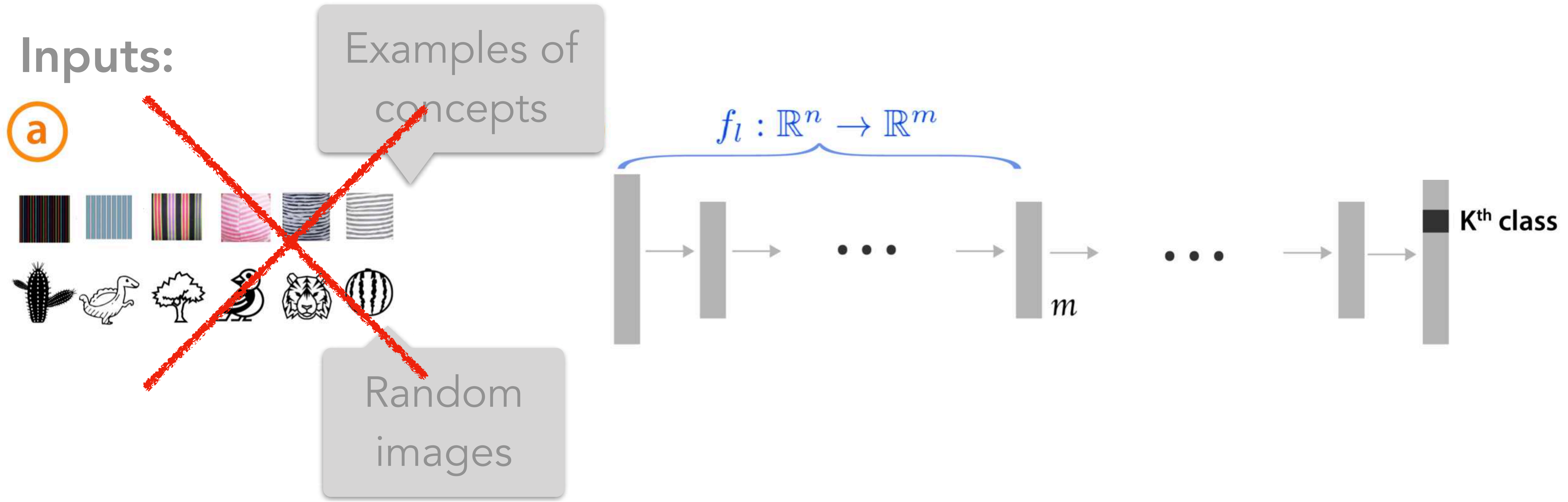
Figure 8: (a) We show two example units that are responsible for visual artifacts in GAN results. There are 20 units in total. By ablating these units, we can fix the artifacts in (b) and significantly improve the visual quality as shown in (c).



Automatically learning CAVs

[Ghorbani et al. NeurIPS 19]

Segment training images into patches, cluster them to discover new concepts (and rigorously validate them).



Inputs:

ⓐ

Examples of concepts

Random images

$f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$

m

Kth class

Hand concept in dumbbell class

Human concept in Jinrikisha class

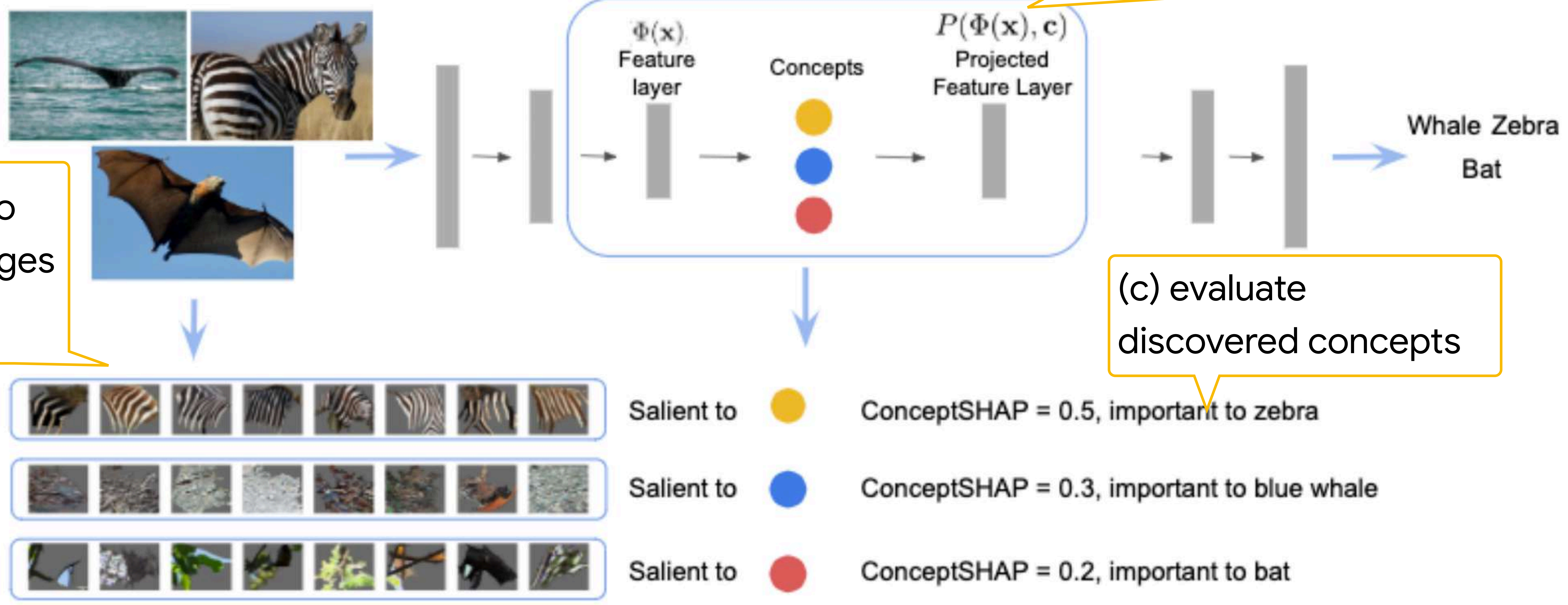
Poles concept in carousel class

Discovering “complete” set of concepts

[Yeh et al. Neurips 20]

Decompose activations into concept vectors that span the activation space.

(b) Decompose feature vectors into concept vectors



(a) Use ACE to segment images and cluster

(c) evaluate discovered concepts

Completeness:

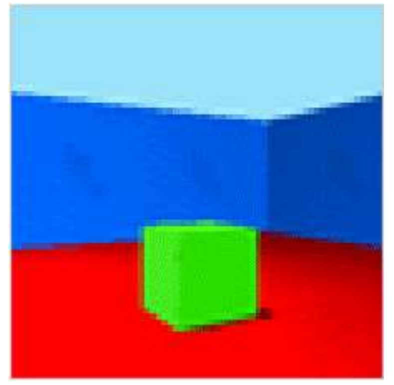
The relative prediction accuracy if I only had this concept.

$$\frac{\mathbb{E}_{\mathbf{x},y\sim V}[\mathbb{1}[y=\arg\max_{y'} P(y'|\mathbb{E}[z_{1:T}],h)]]-R}{\mathbb{E}_{\mathbf{x},y\sim V}[\mathbb{1}[y=\arg\max_{y'} P(y'|\mathbf{x}_{1:T},f)]]-R},$$

Why this metric?
Under simple assumptions, this metric is equivalent to top k PCA vectors.

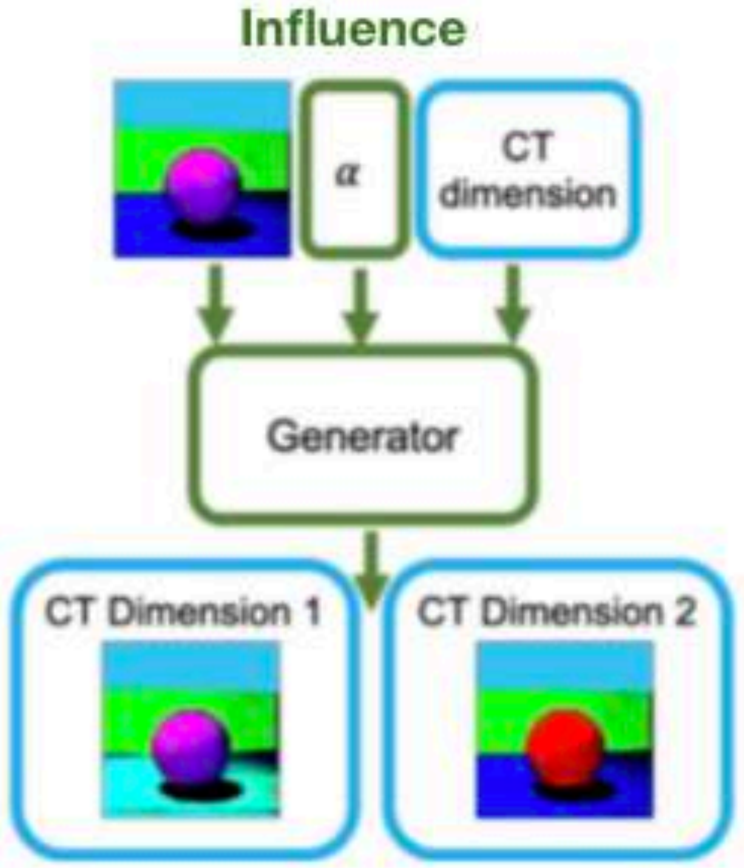
DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

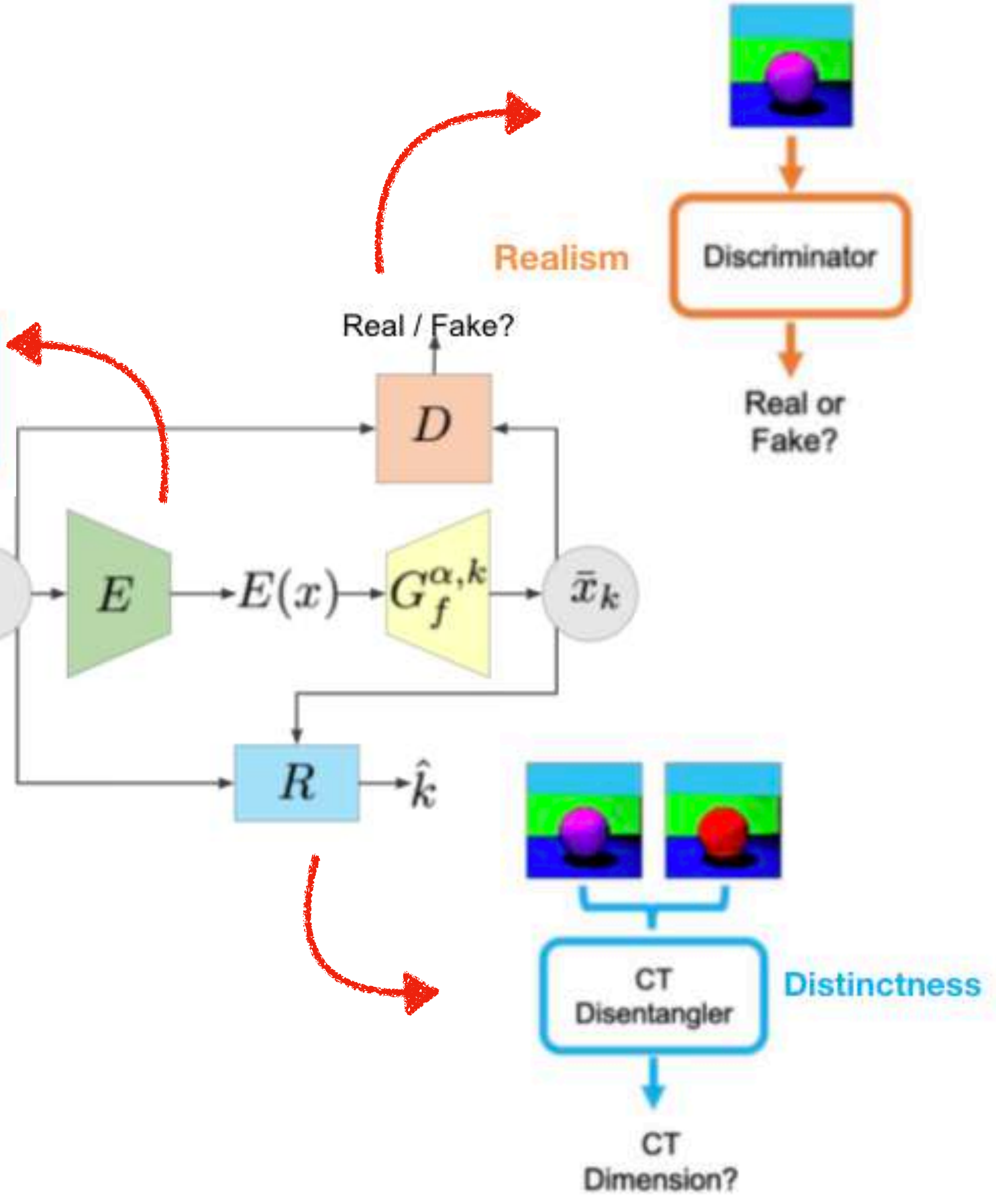
[Ghandeharioun et al. 2021]

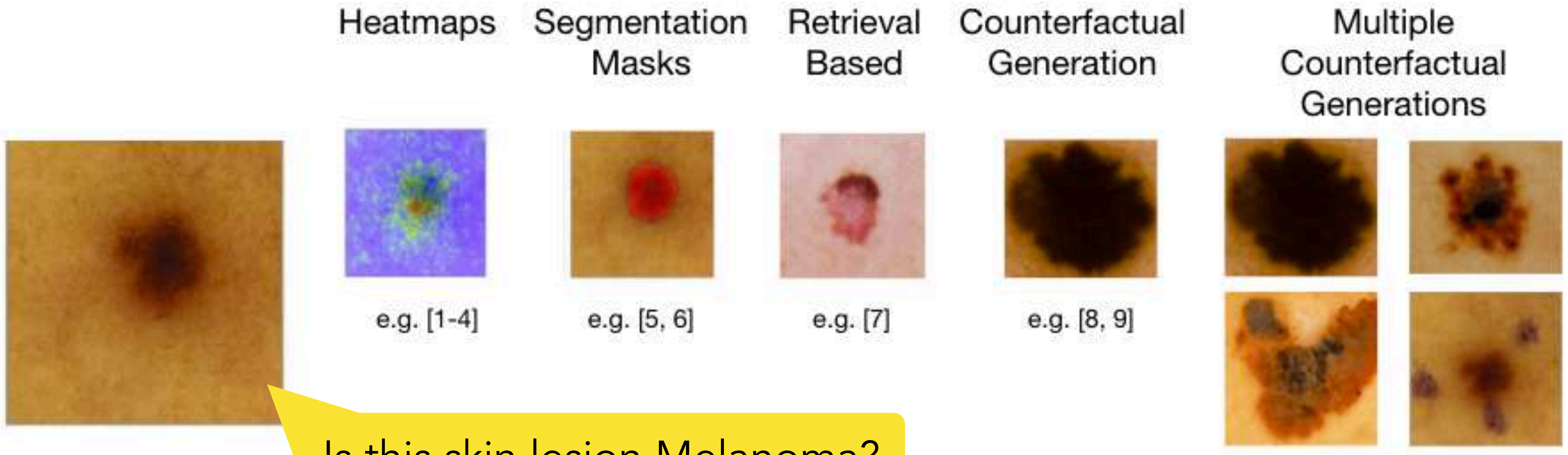


[1]

Joint train a generative model to produce multiple counter factual concepts







Query: Benign

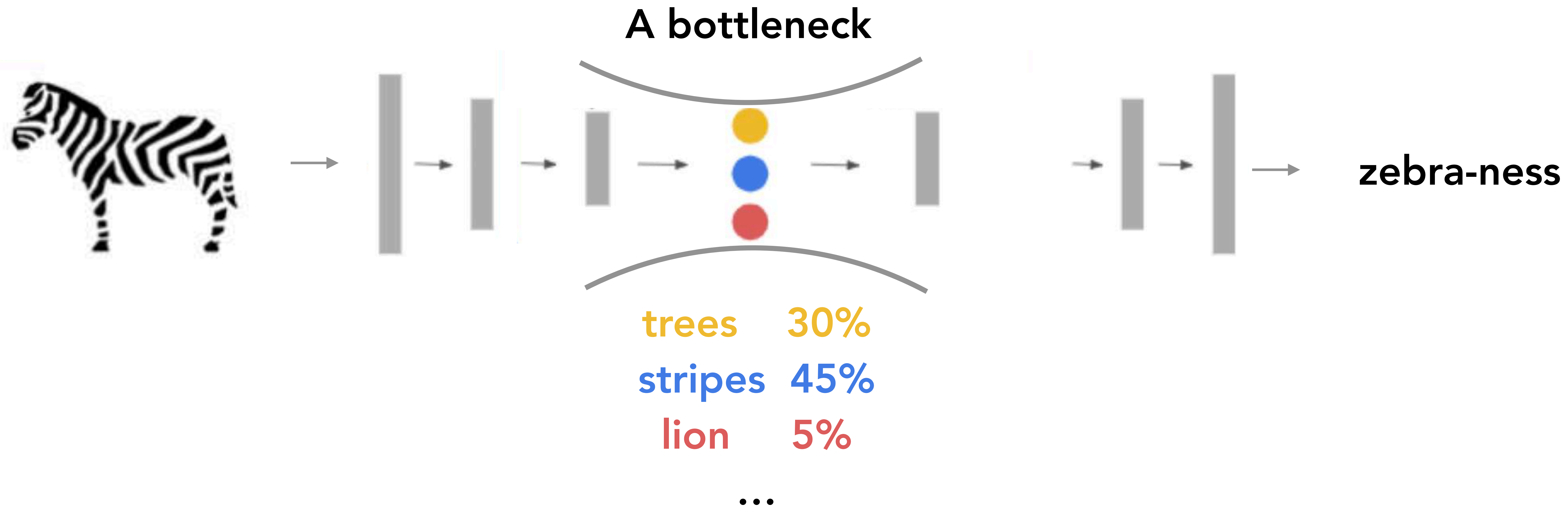
Is this skin lesion Melanoma?

Slide credit: Asma Ghandeharioun

[1] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. ICML, 2018.

Concept bottleneck models

[Goh et al., ICML 20]



A bottleneck

zebra-ness

trees 30%

stripes 45%

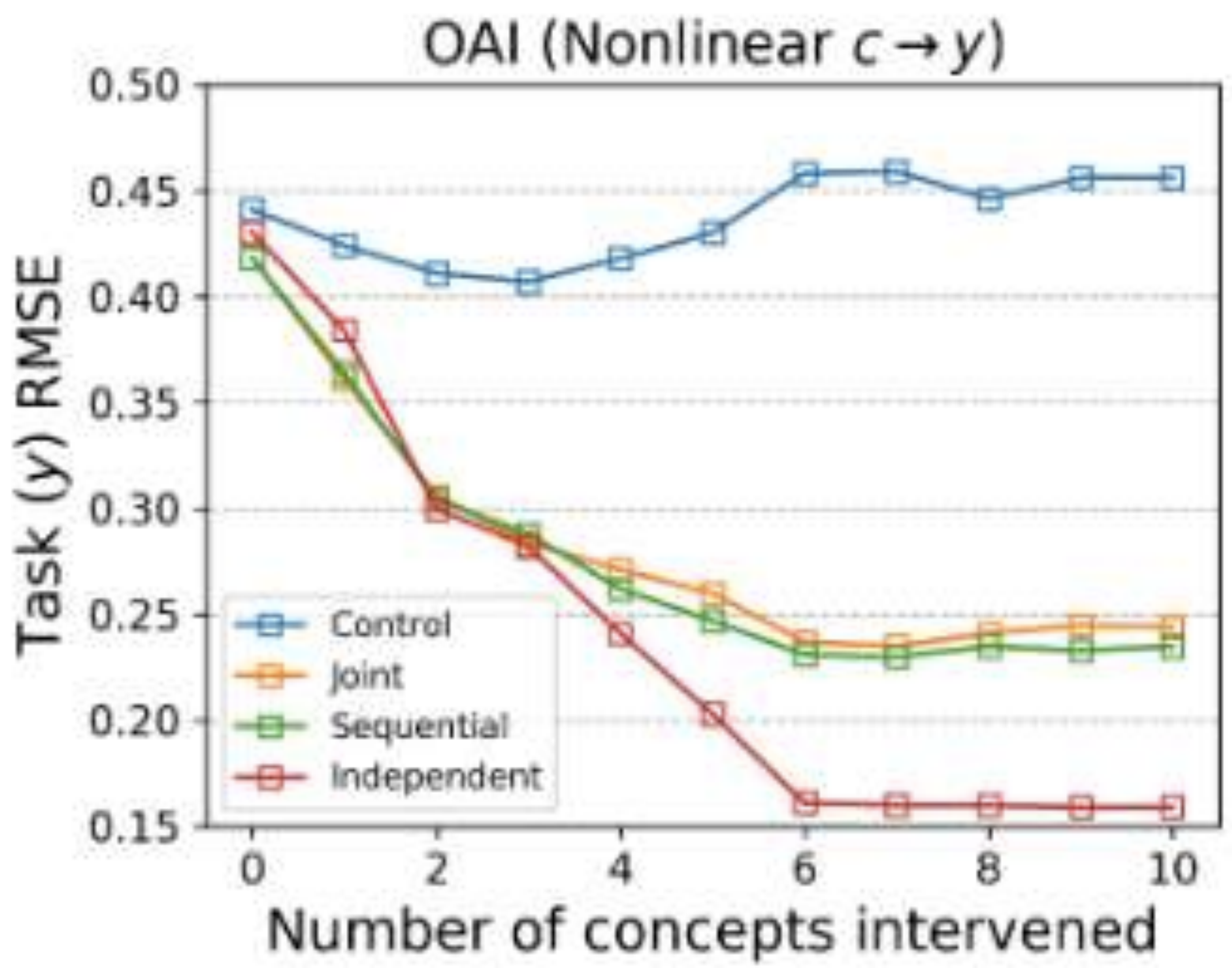
lion 5%

...

Train a model with concepts as neurons in the middle.

Bonus: we can **interact** and **control** the model

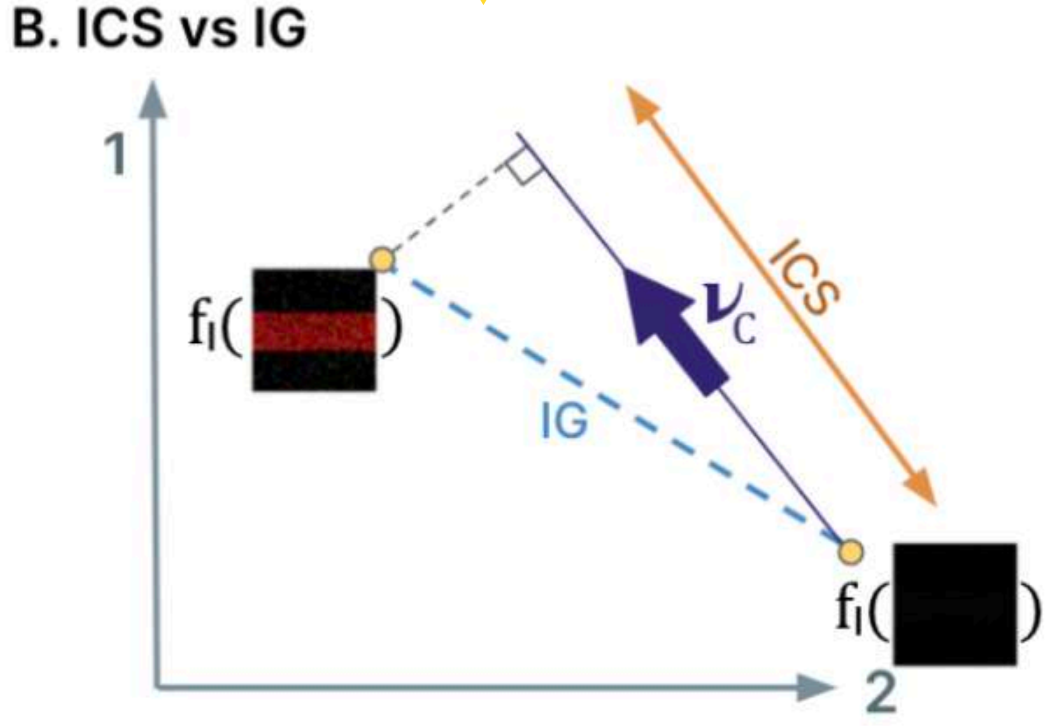
Based on my expertise, symptom X should not contribute to the diagnosis.

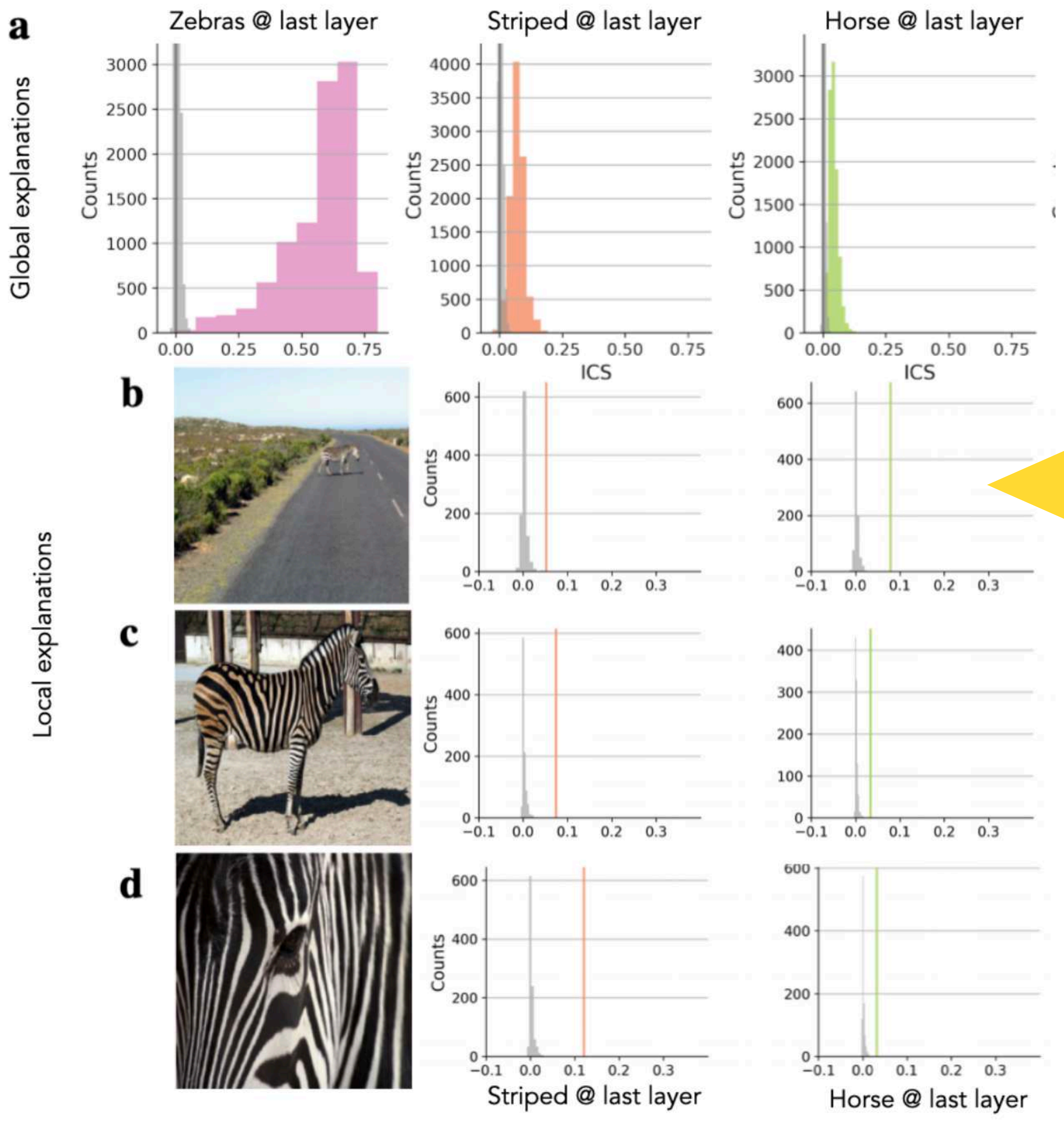


ICS: Combine TCAV + IG to provide both global and local explanations

[Schrouff et al. 21]

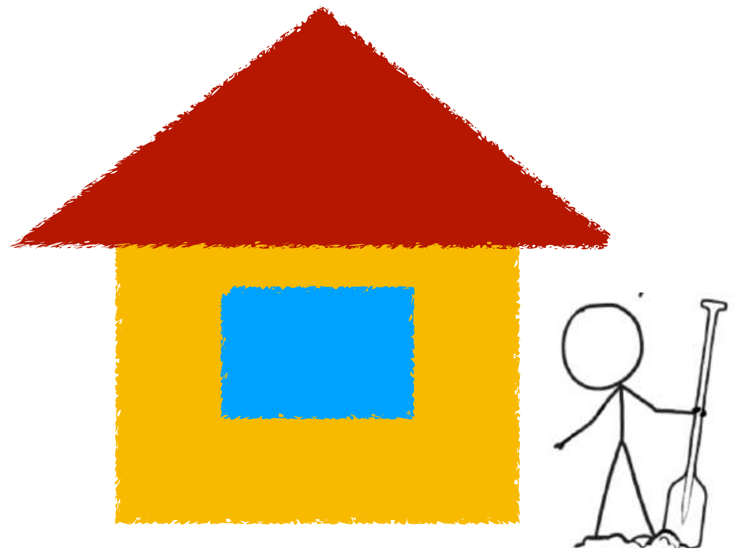
Integrate on the path of a CAV
<->
A projection of path integration



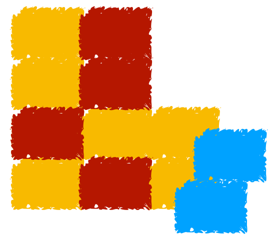


Zebras look like horses from far a way

Types of interpretability methods



My ML



Explaining data

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

Building inherently interpretable model

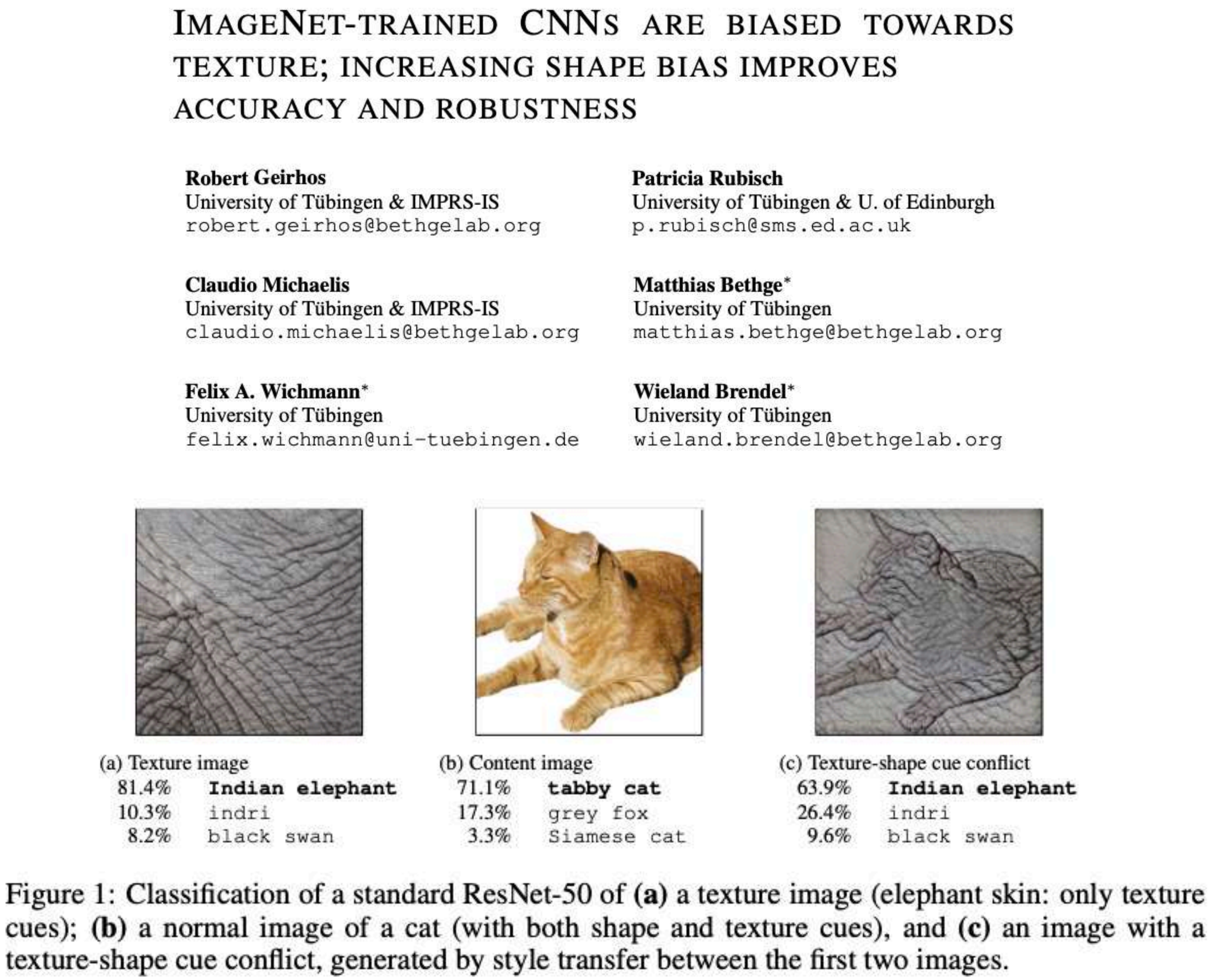
$$\underset{E,M}{\operatorname{argmax}}\ Q(\mathbf{E}\text{xplanation}, \mathbf{M}\text{odel}|\mathbf{H}\text{uman}, \mathbf{D}\text{ata}, \mathbf{T}\text{ask})$$

Post-training interpretability methods

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

Things that aren't covered but important - science

- Science of it - studying models as a scientific object



IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

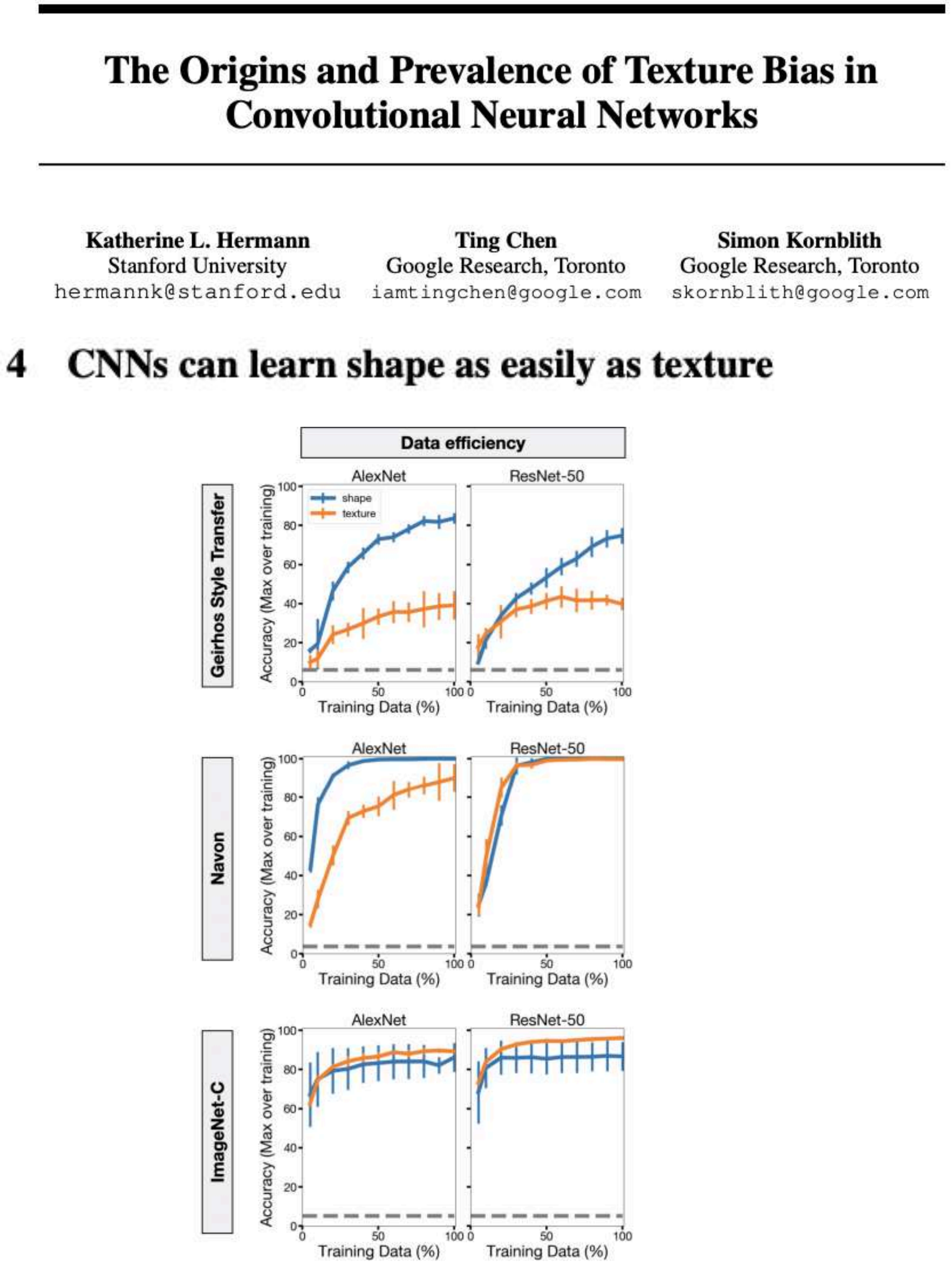
Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.



The Origins and Prevalence of Texture Bias in Convolutional Neural Networks

Katherine L. Hermann
Stanford University
hermannk@stanford.edu

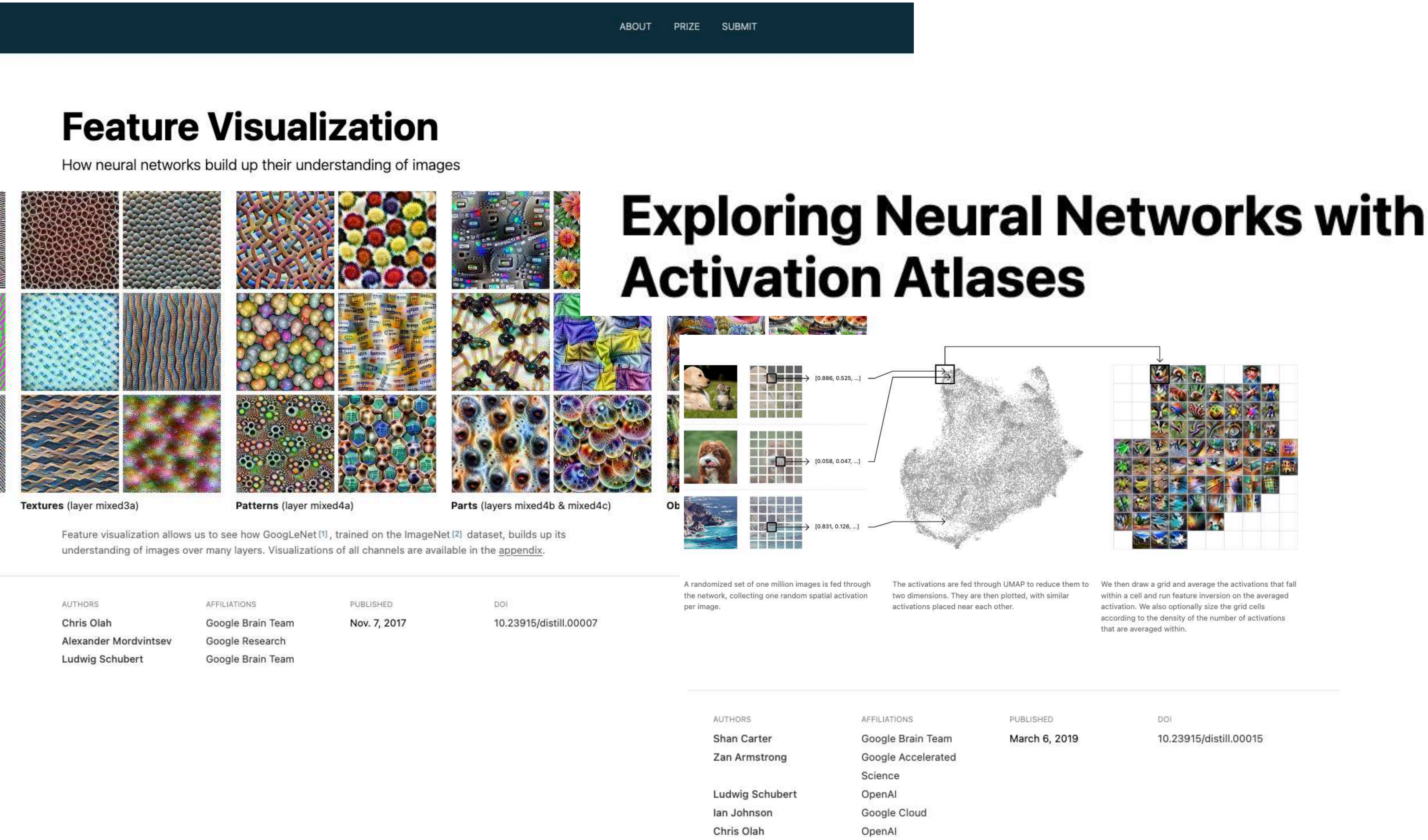
Ting Chen
Google Research, Toronto
iamtingchen@google.com

Simon Kornblith
Google Research, Toronto
skornblith@google.com

4 CNNs can learn shape as easily as texture

Things that aren't covered but important - science

- Science of it - studying models as a scientific object



Feature Visualization

How neural networks build up their understanding of images

Feature visualization allows us to see how GoogLeNet [1], trained on the ImageNet [2] dataset, builds up its understanding of images over many layers. Visualizations of all channels are available in the appendix.

AUTHORS	AFFILIATIONS	PUBLISHED	DOI
Chris Olah	Google Brain Team	Nov. 7, 2017	10.23915/distill.00007
Alexander Mordvintsev	Google Research		
Ludwig Schubert	Google Brain Team		

Exploring Neural Networks with Activation Atlases

A randomized set of one million images is fed through the network, collecting one random spatial activation per image.

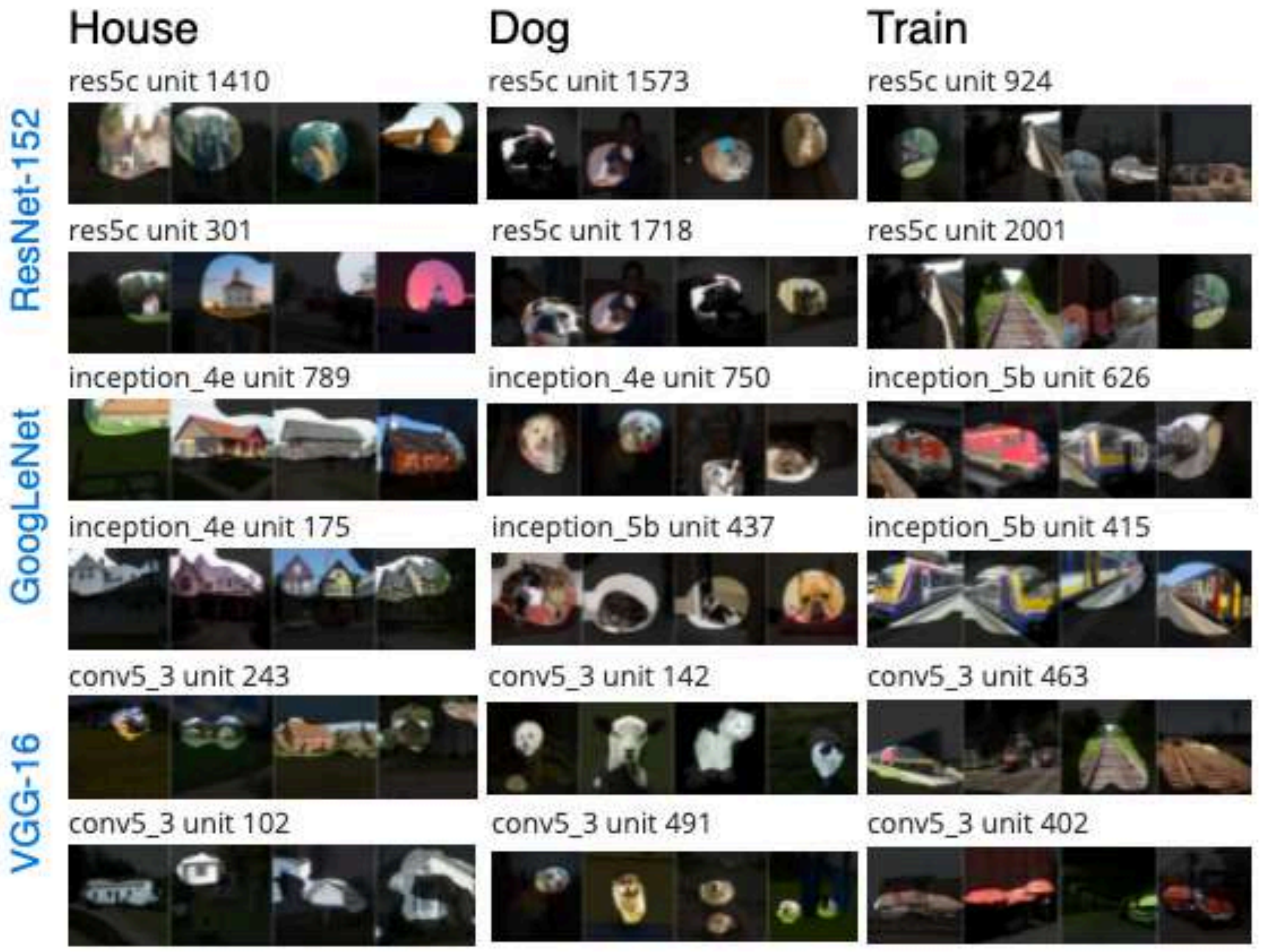
The activations are fed through UMAP to reduce them to two dimensions. They are then plotted, with similar activations placed near each other.

We then draw a grid and average the activations that fall within a cell and run feature inversion on the averaged activation. We also optionally size the grid cells according to the density of the number of activations that are averaged within.

AUTHORS	AFFILIATIONS	PUBLISHED	DOI
Shan Carter	Google Brain Team	March 6, 2019	10.23915/distill.00015
Zan Armstrong	Google Accelerated Science		
Ludwig Schubert	OpenAI		
Ian Johnson	Google Cloud		
Chris Olah	OpenAI		

Network Dissection: Quantifying Interpretability of Deep Visual Representations

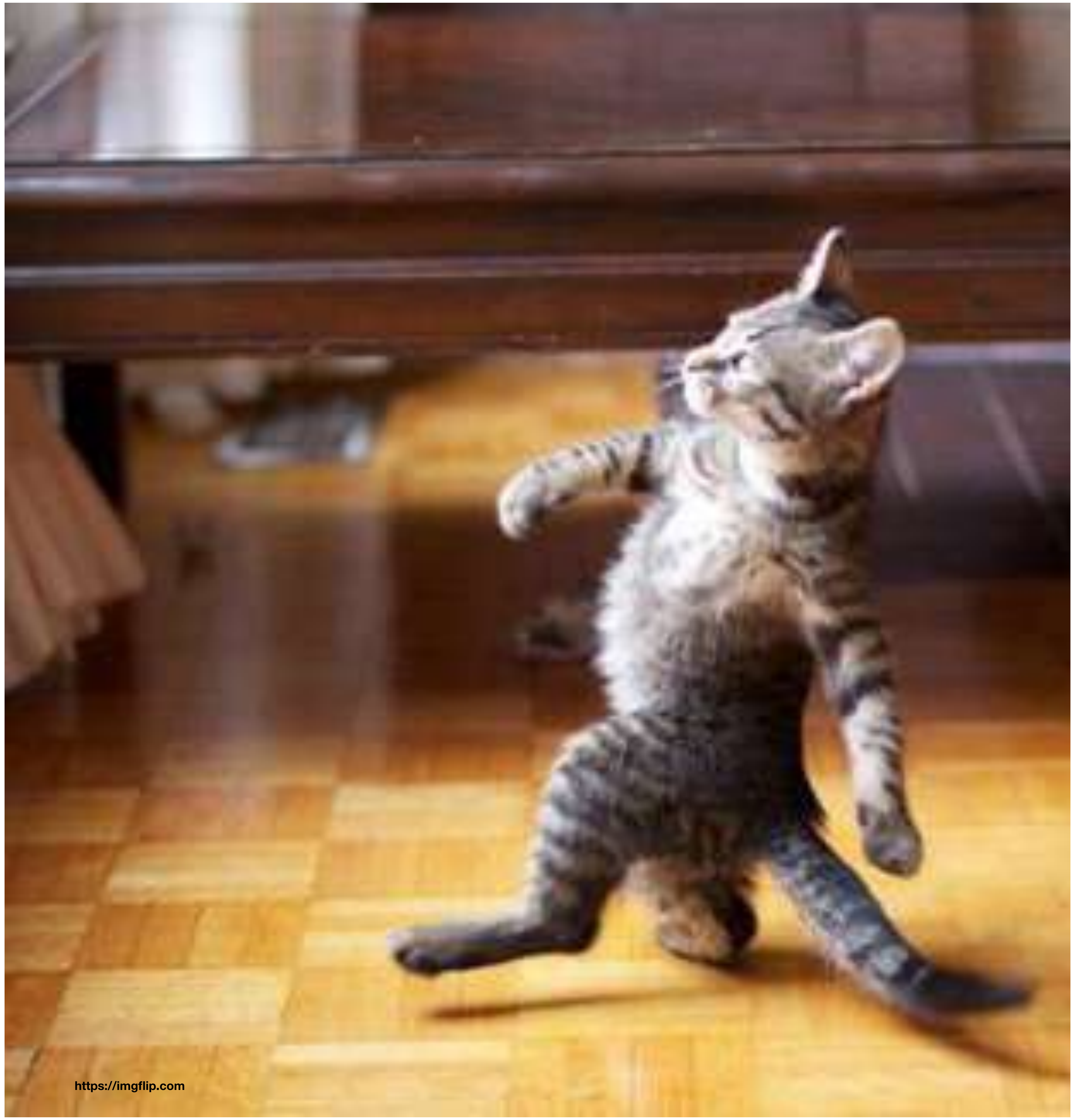
David Bau*, Bolei Zhou*, Aditya Khosla, Aude Oliva, and Antonio Torralba
CSAIL, MIT



That's a wrap!

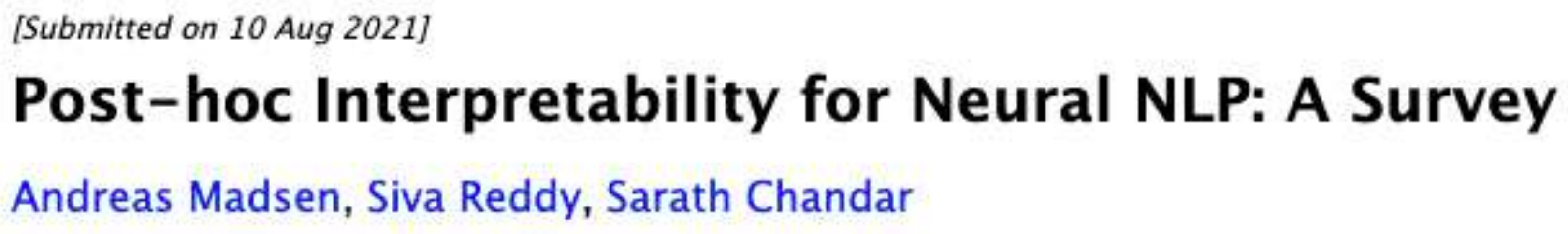
- What and why
- !Caution!: Things to be careful when using and developing interpretability methods
- Evaluate: How to evaluate interpretability methods
- Methods: 3 types of methods and examples

cheeseburger.com


https://imgflip.com

Backups

Other domains



[Submitted on 10 Aug 2021]

Post-hoc Interpretability for Neural NLP: A Survey

Andreas Madsen, Siva Reddy, Sarath Chandar

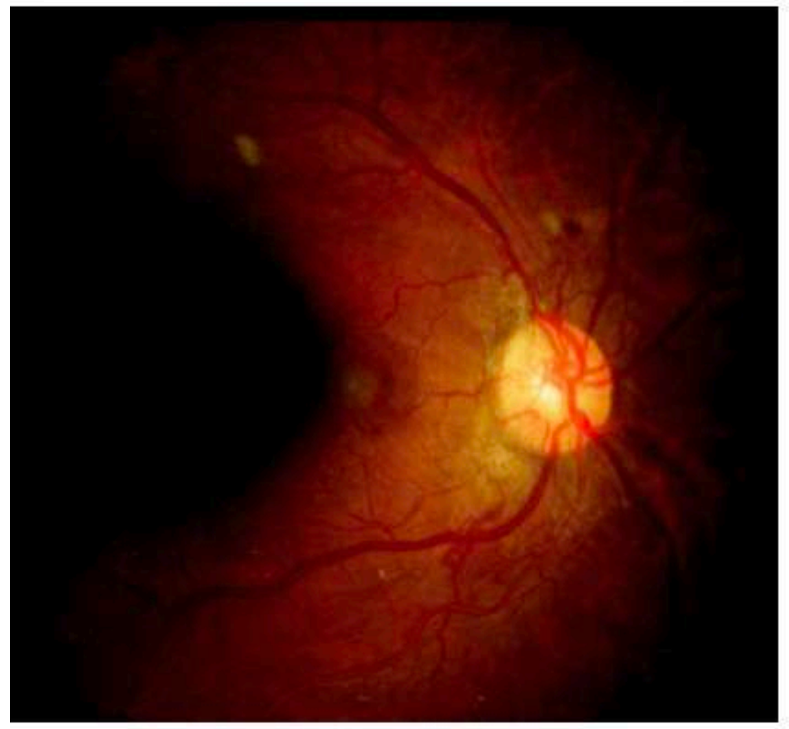
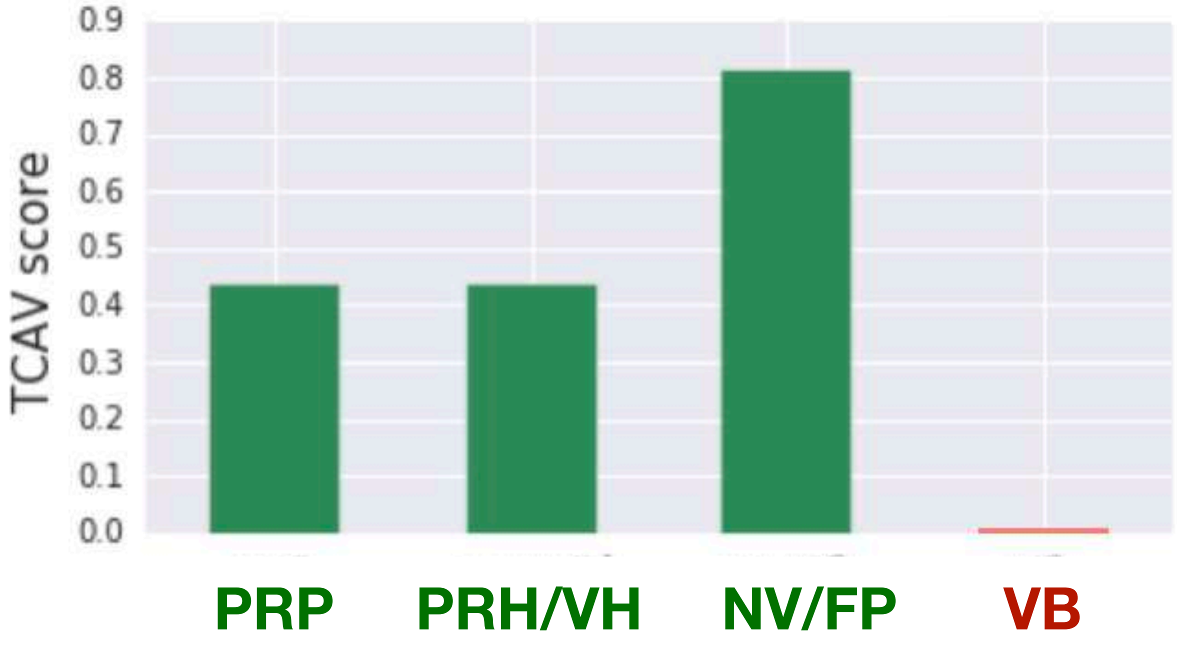
Radiology: Artificial Intelligence

On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, Roland Wiest

Ok, great but…

What if I don’t have concepts?

Prediction class	Prediction accuracy	Example	TCAV scores
DR level 4	High		

I don’t have these!

Can we automate this?

Automatically learning CAVs

[Ghorbani et al. NeurIPS 19]

Amirata Ghorbani

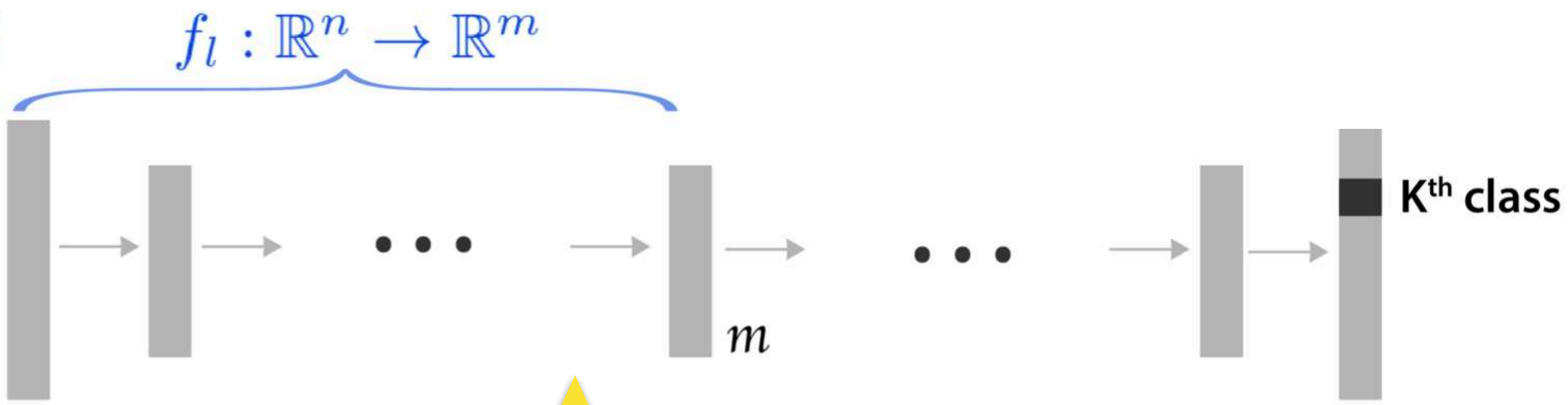
Automatically learning CAVs

[Ghorbani et al. NeurIPS 19]

Inputs:

Examples of concepts

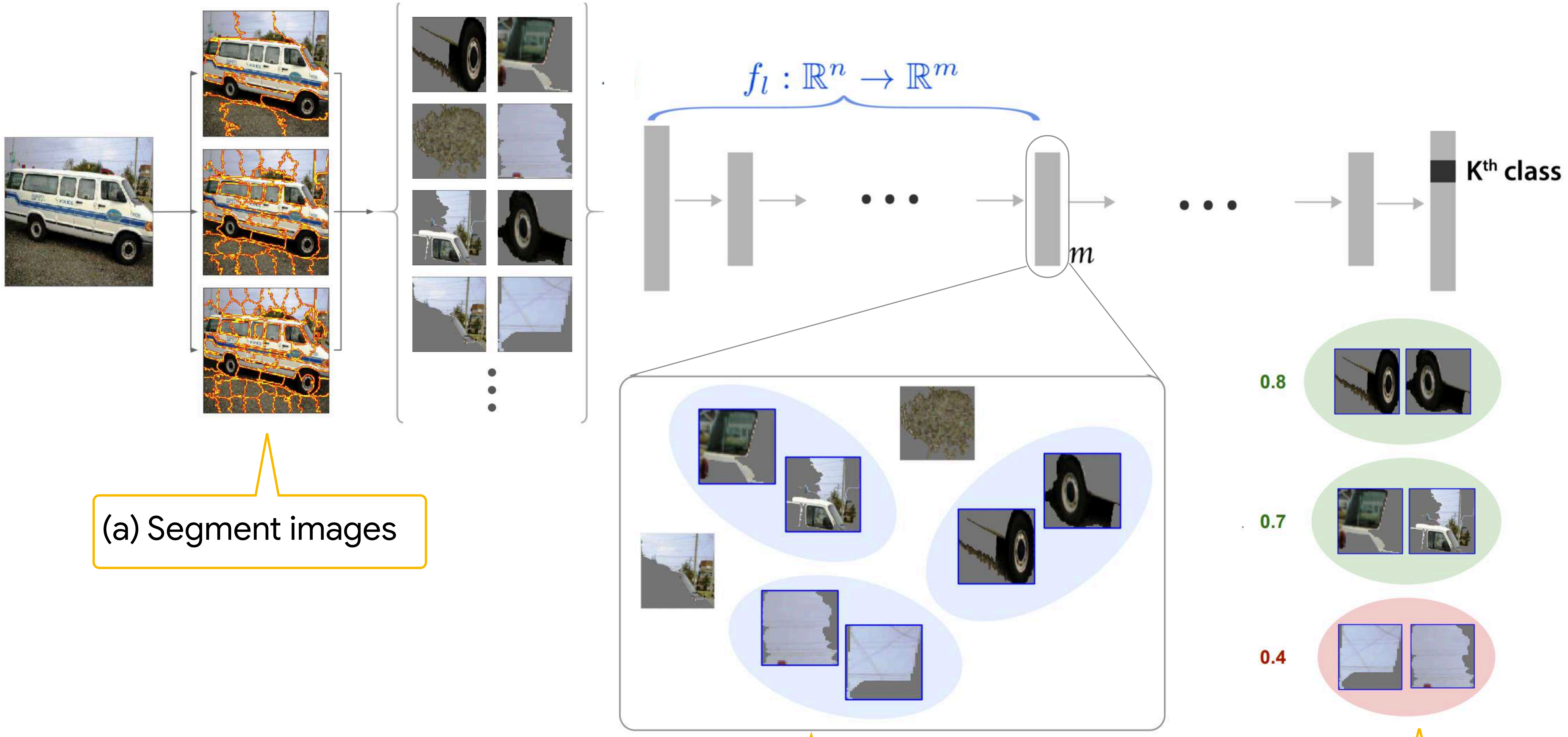
Random images



A trained network under investigation
and
Internal tensors

Automatic Concept-based Explanations (ACE)

[Ghorbani et al. NeurIPS 19]

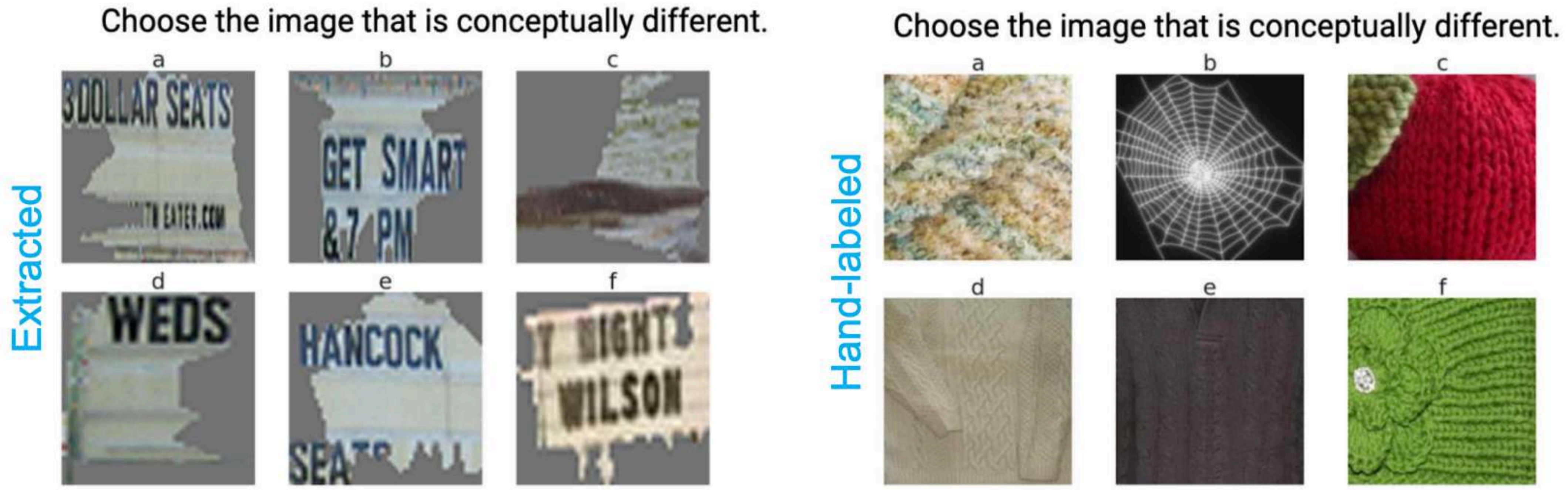


(a) Segment images

(b) (smart) cluster segments

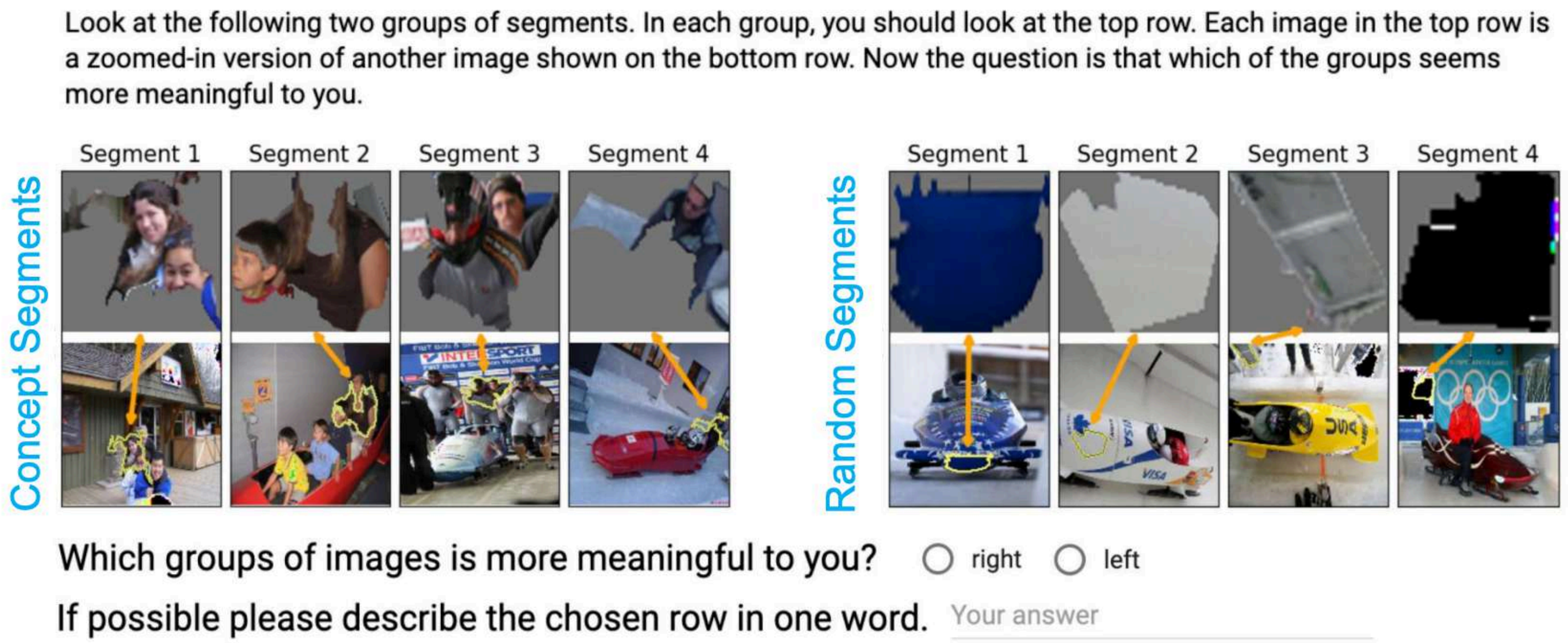
(c) evaluate discovered concepts

Validating with human experiments: Intruder and meaning test


Choose the image that is conceptually different.

Choose the image that is conceptually different.

Experiment 1: Identifyig intruder concept


Look at the following two groups of segments. In each group, you should look at the top row. Each image in the top row is a zoomed-in version of another image shown on the bottom row. Now the question is that which of the groups seems more meaningful to you.

Which groups of images is more meaningful to you? ○ right ○ left

If possible please describe the chosen row in one word. Your answer

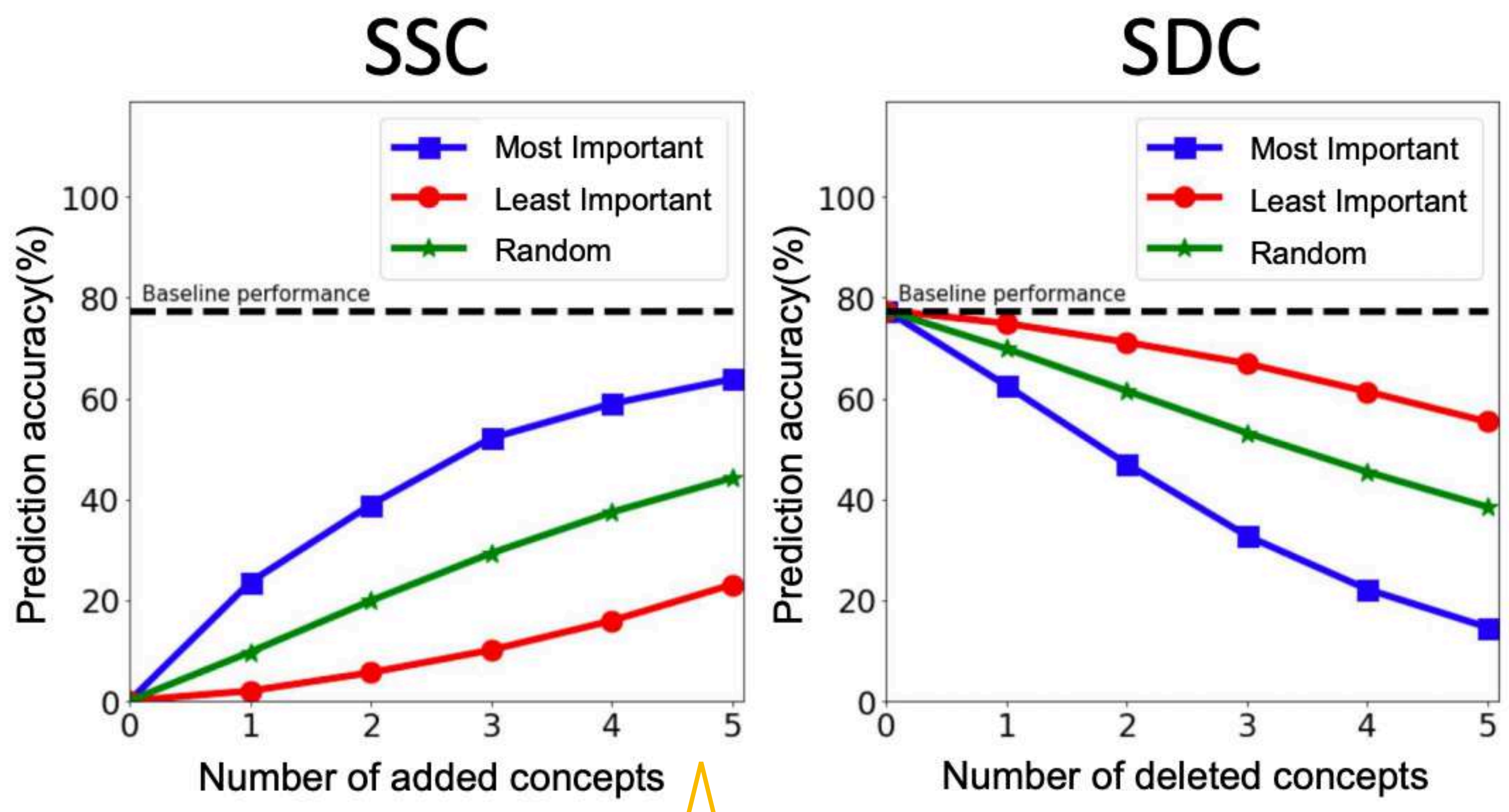
Experiment 2: Identifying the meaning of concept

- Exp1: Intruder test
  - Task: Identify an odd one out
  - Discovered concepts: 99%, similar to hand-labeled dataset, 97%
- Exp2: Meaning test
  - Task: Select between discovered concepts vs random segments and name them.
  - Correctly chosen 95% of time
  - 56% used the same name and 77% named the same or top two terms (e.g., human, face)

Validating importance: Addition and deletion test

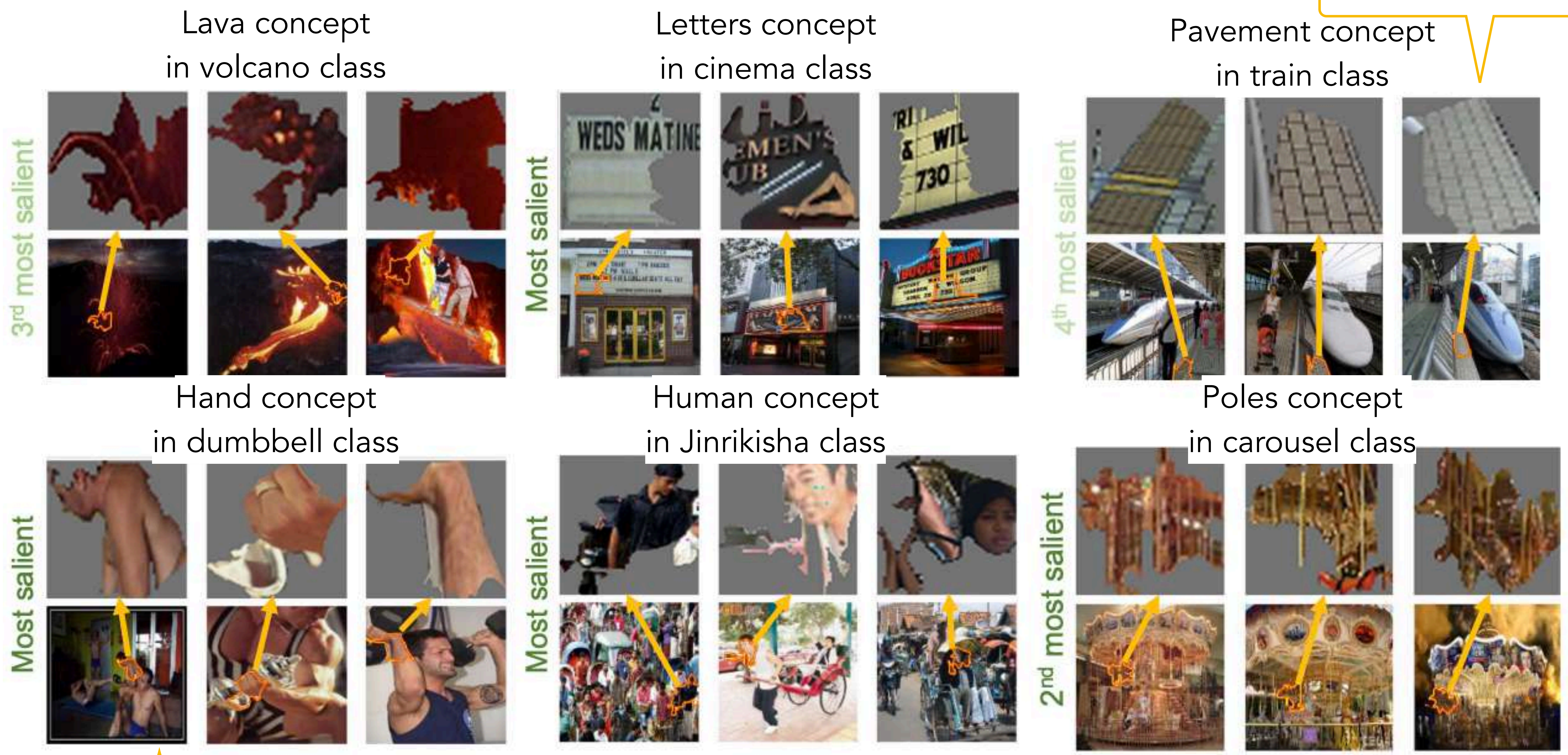
Adding top-rated

Deleting top-rated



Adding the top5 discovered concepts achieves 80% of the original accuracy

Qualitative results:
Surprises and non-surprises

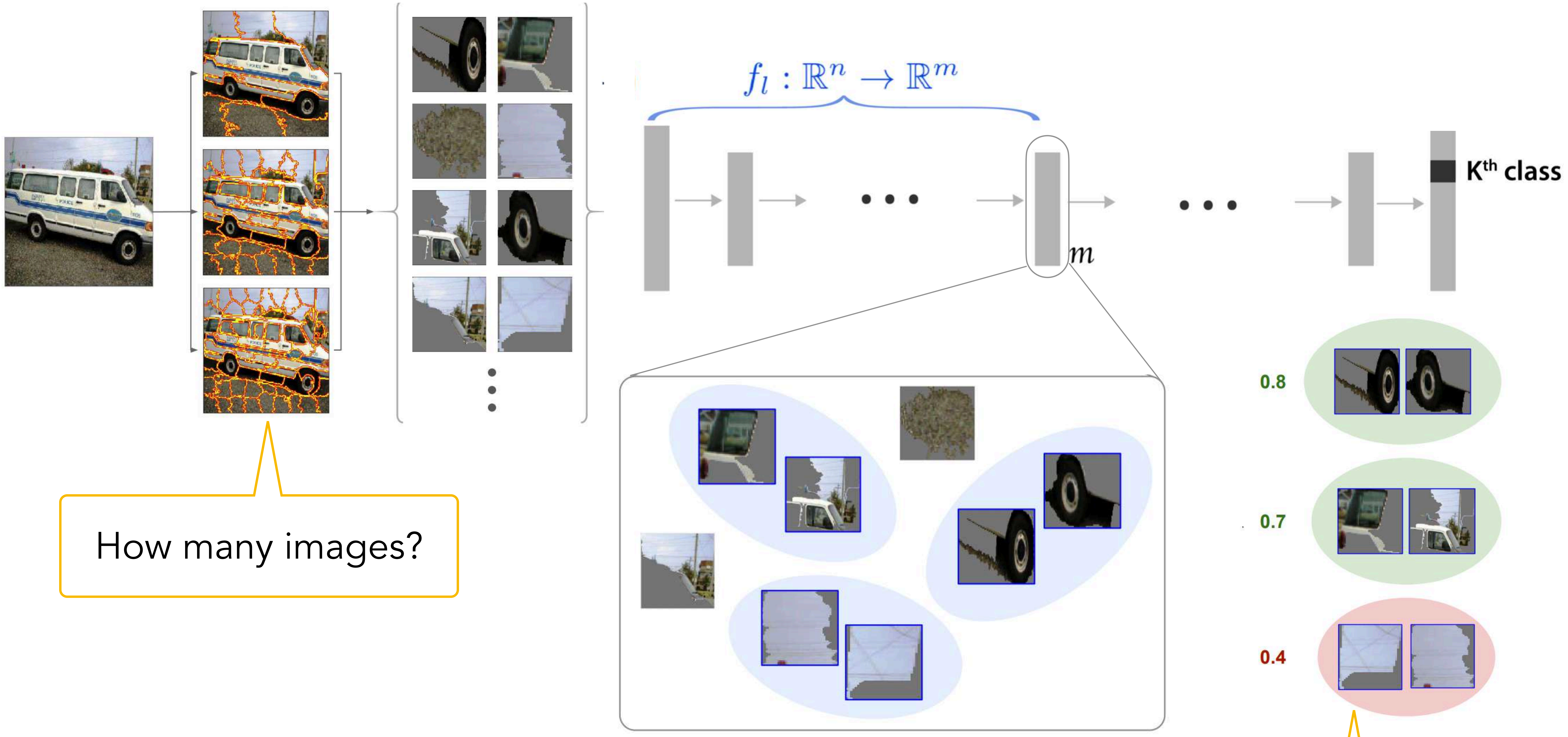


This may not work in Korea?

Hands are not dumbbells...

Ok, great but…

When do you stop?



How many images?

Are these concepts “enough”?

Discovering “complete” set of concepts

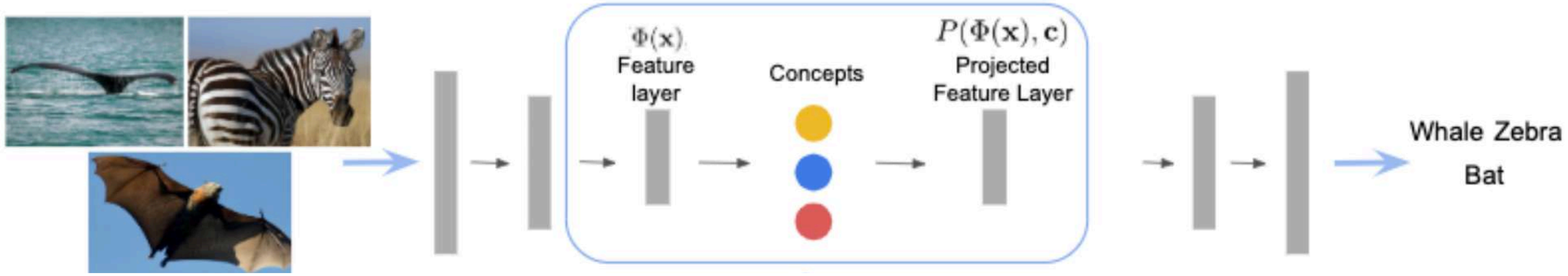
[Yeh, Arik, Ravikumar, Pfister, K. Neurips 20]

Chih-Kuan Yeh

Discovering “complete” set of concepts

[Yeh, Arik, Ravikumar, Pfister, K. Neurips 20]

Decompose activations into concept vectors that span the activation space.

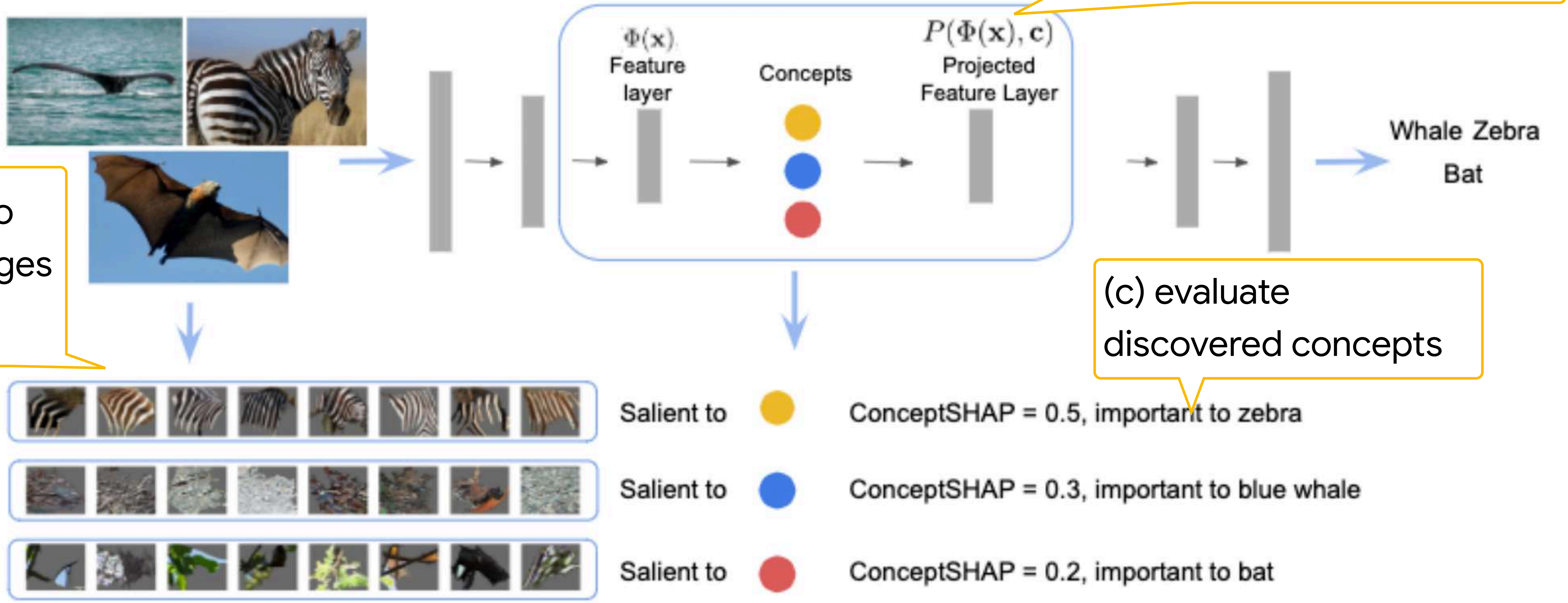


Discovering "complete" set of concepts

[Yeh, Arik, Ravikumar, Pfister, K. Neurips 20]

Decompose activations into concept vectors that span the ac

(b) Decompose feature vectors into concept vectors



(a) Use ACE to segment images and cluster

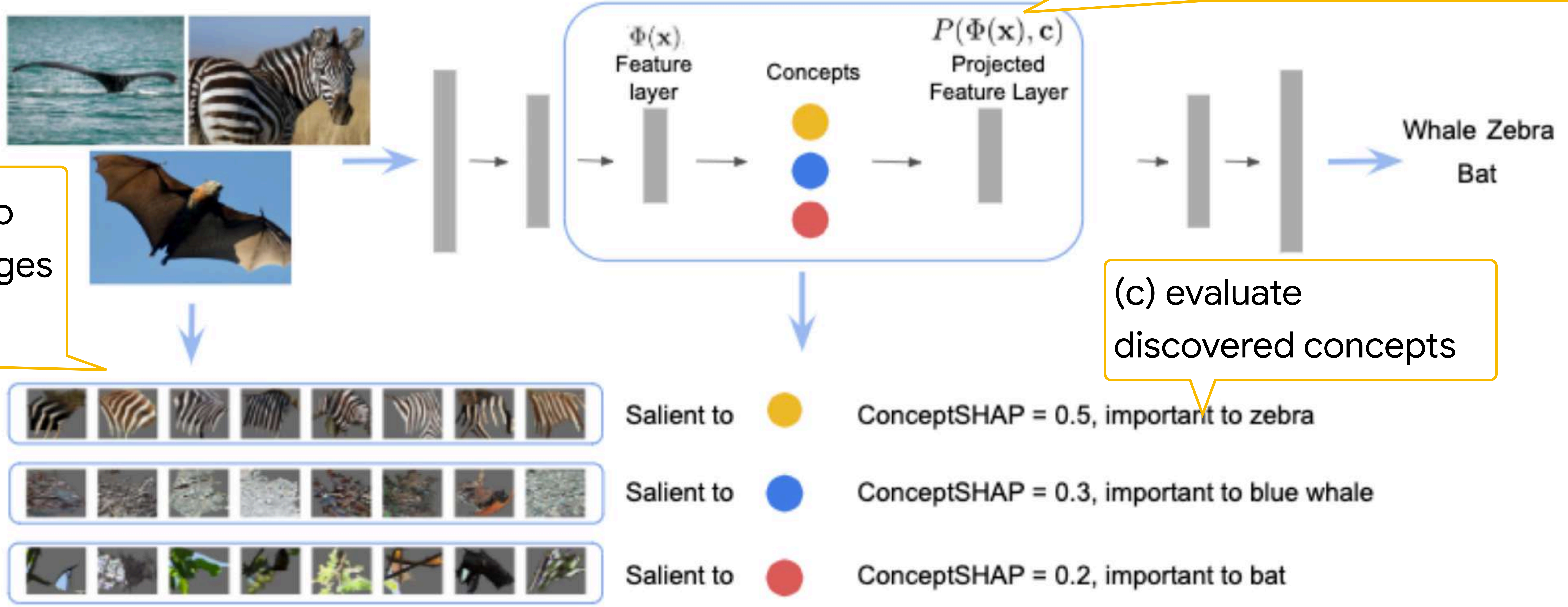
(c) evaluate discovered concepts

Discovering “complete” set of concepts

[Yeh, Arik, Ravikumar, Pfister, K. Neurips 20]

Decompose activations into concept vectors that span the ac

(b) Decompose feature vectors into concept vectors



(a) Use ACE to segment images and cluster

(c) evaluate discovered concepts

Completeness:

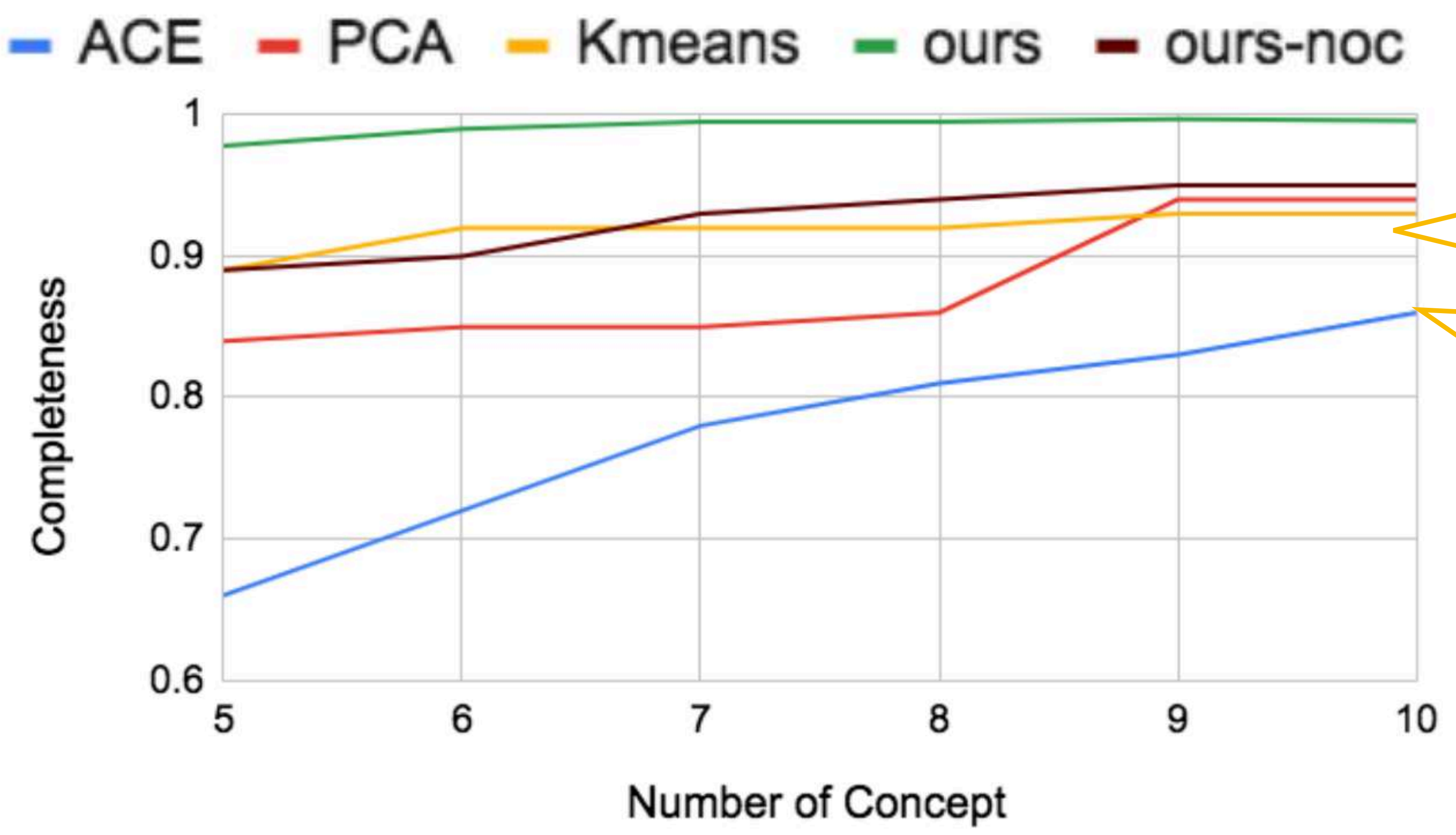
The relative prediction accuracy if I only had this concept.

$$\frac{\mathbb{E}_{\mathbf{x},y\sim V}[\mathbb{1}[y=\arg\max_{y'} P(y'|\mathbb{E}[z_{1:T}],h)]]-R}{\mathbb{E}_{\mathbf{x},y\sim V}[\mathbb{1}[y=\arg\max_{y'} P(y'|\mathbf{x}_{1:T},f)]]-R},$$

Why this metric?
Under simple assumptions, this metric is equivalent to top k PCA vectors.

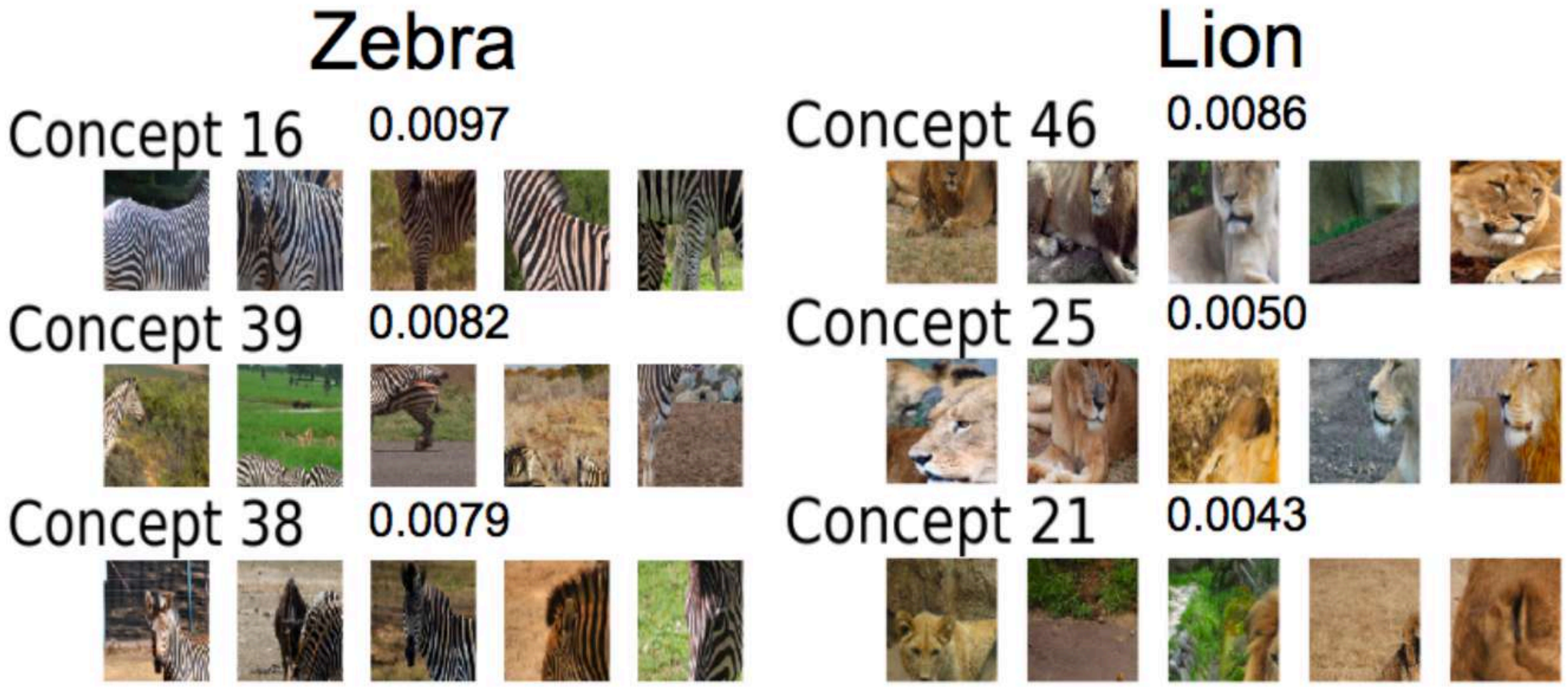
Discovering "complete" set of concepts

[Yeh, Arik, Ravikumar, Pfister, K. Neurips 20]



Kmeans are often surprisingly good.

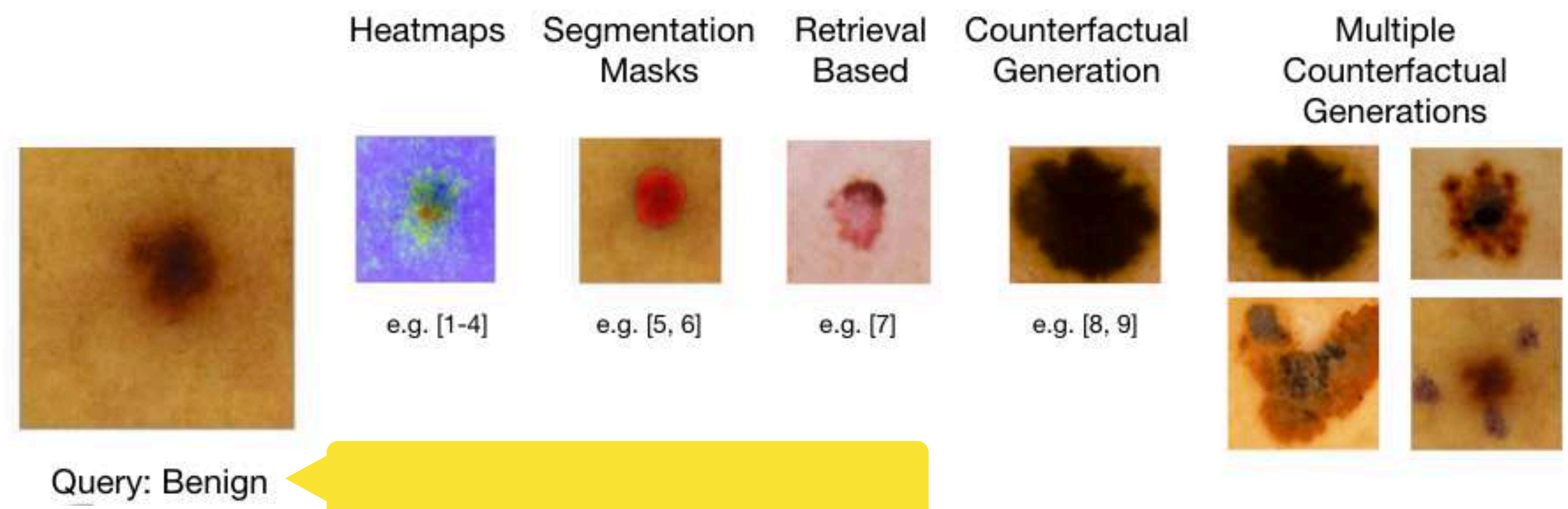
ACE might take a while to converge.



Ok, great but…

Haven’t you heard about generative models?

Instead of looking for concepts **within training set,** can we just use **generate** concepts using generative models?



Is this skin lesion Melanoma?

Slide credit: Asma Ghandeharioun

DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

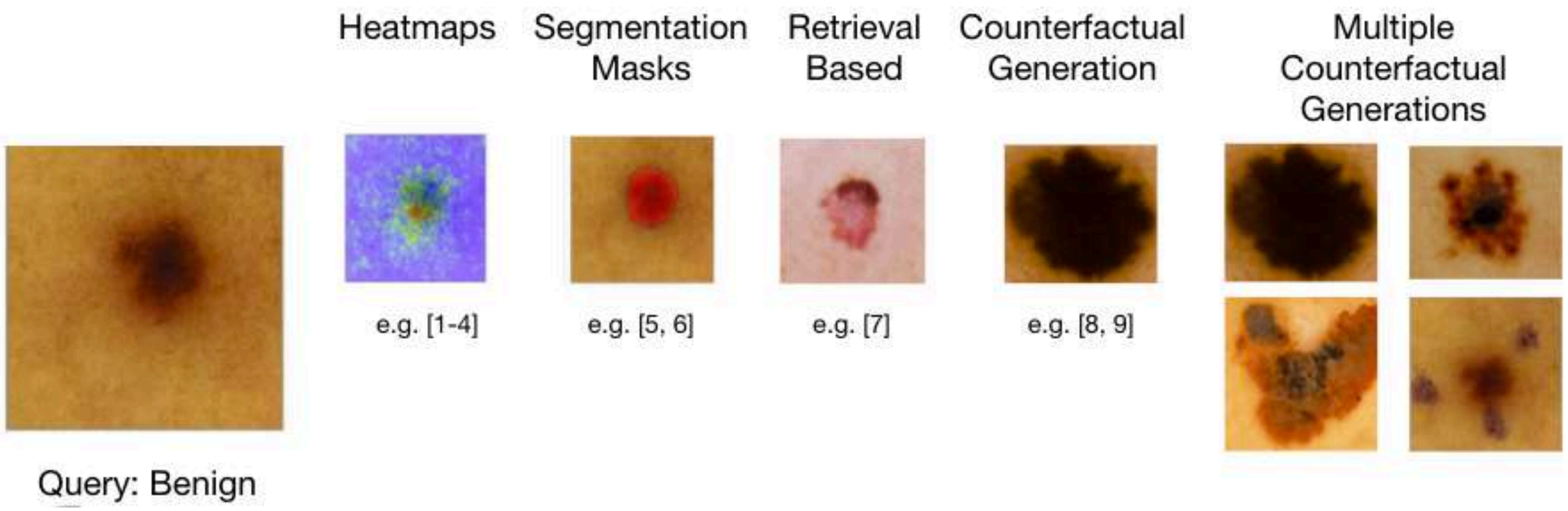
[Ghandeharioun, K., Li, Jou, Eoff, Picard, 2021]

Asma Ghandeharioun

Slide credit: Asma Ghandeharioun

DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

[Ghandeharioun, K., Li, Jou, Eoff, Picard, 2021]



Is this skin lesion Melanoma?

[1] Erhan et al. “Visualizing higher-layer features of a deep network”. University of Montreal, 2009.

[2] Smilkov et al. “SmoothGrad: Removing noise by adding noise”. ICMLW 2017.

[3] Sundararajan et al. “Axiomatic attribution for deep networks”. ICML 2017.

[4] Lundberg et al. “A unified approach to interpreting model predictions”. NeurIPS 2017.

[5] Ghorbani et al. “Towards automatic concept-based explanations”. NeurIPS 2019.

[6] Santamaria-Pang, et al. “Towards Emergent Language Symbolic Semantic Segmentation and Model Interpretability”. MICCAI 2020.

[7] Silva et al. “Interpretability-guided content-based medical image retrieval”. MICCAI 2020.

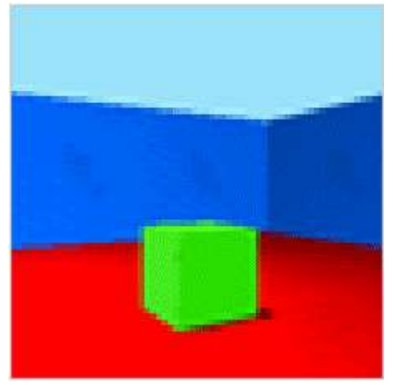
[8] Samangouei et al.“ExplainGAN: Model explanation via decision boundary crossing transformations”. ECCV 2018.

[9] Singla et al. “Explanation by progressive exaggeration”. ICLR 2020.

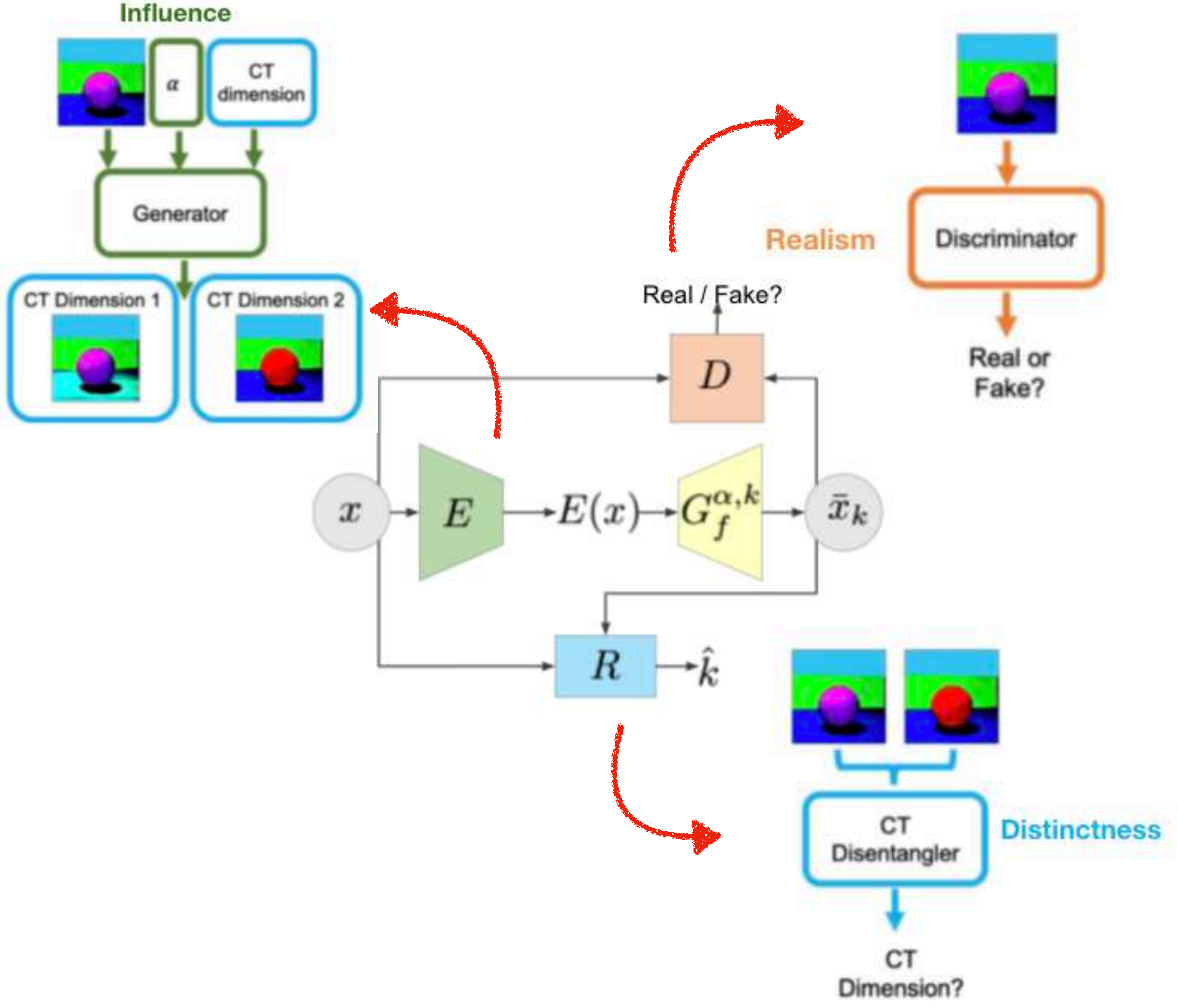
Slide credit: Asma Ghandeharioun

DISSECT: Disentangled Simultaneous Explanations via Concept Traversals

[Ghandeharioun, K., Li, Jou, Eoff, Picard, 2021]


[1]

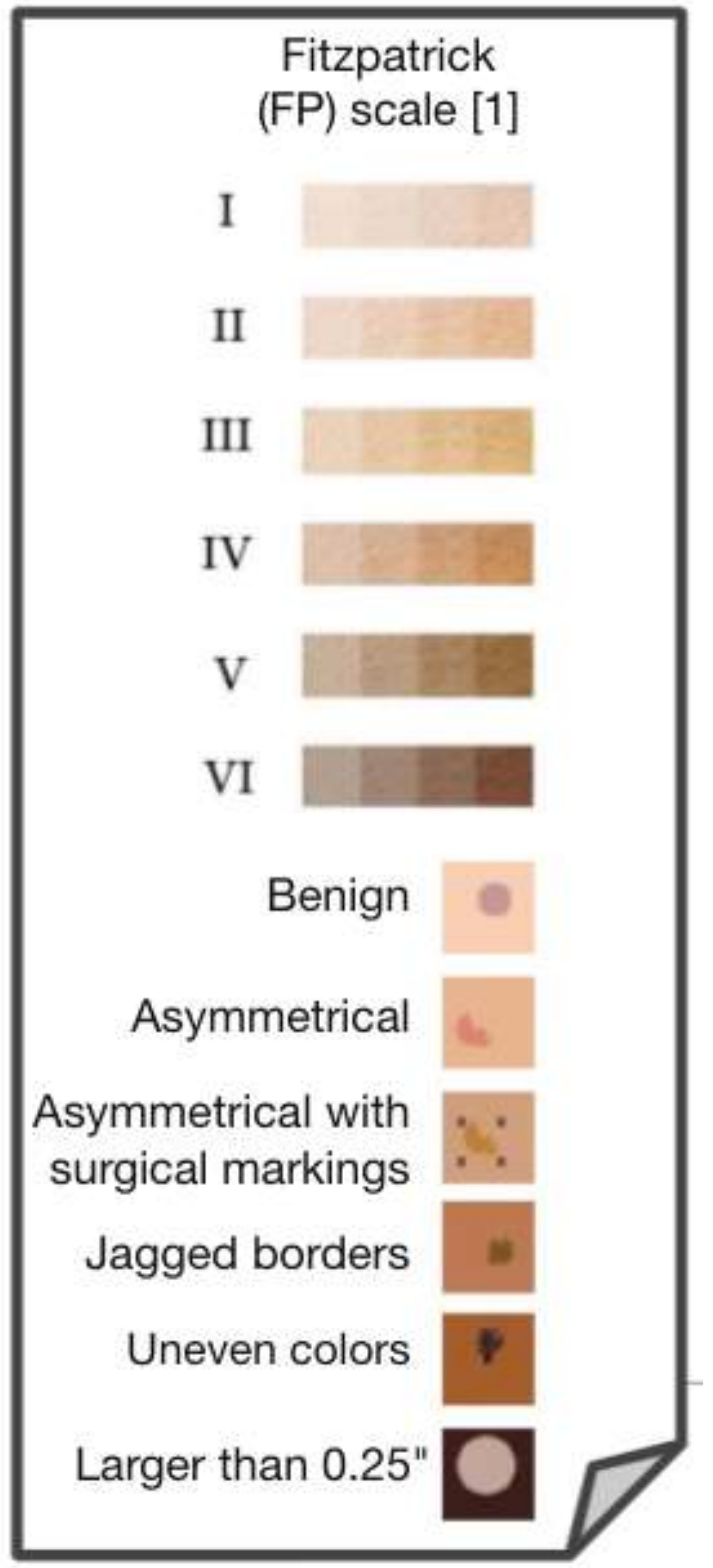
- Desiderata
  - Influential (to classifier's decision)
  - Distinct concept traversals
  - Stable generation
  - High substitutability (can replace real data)
  - High realism (in data manifold)

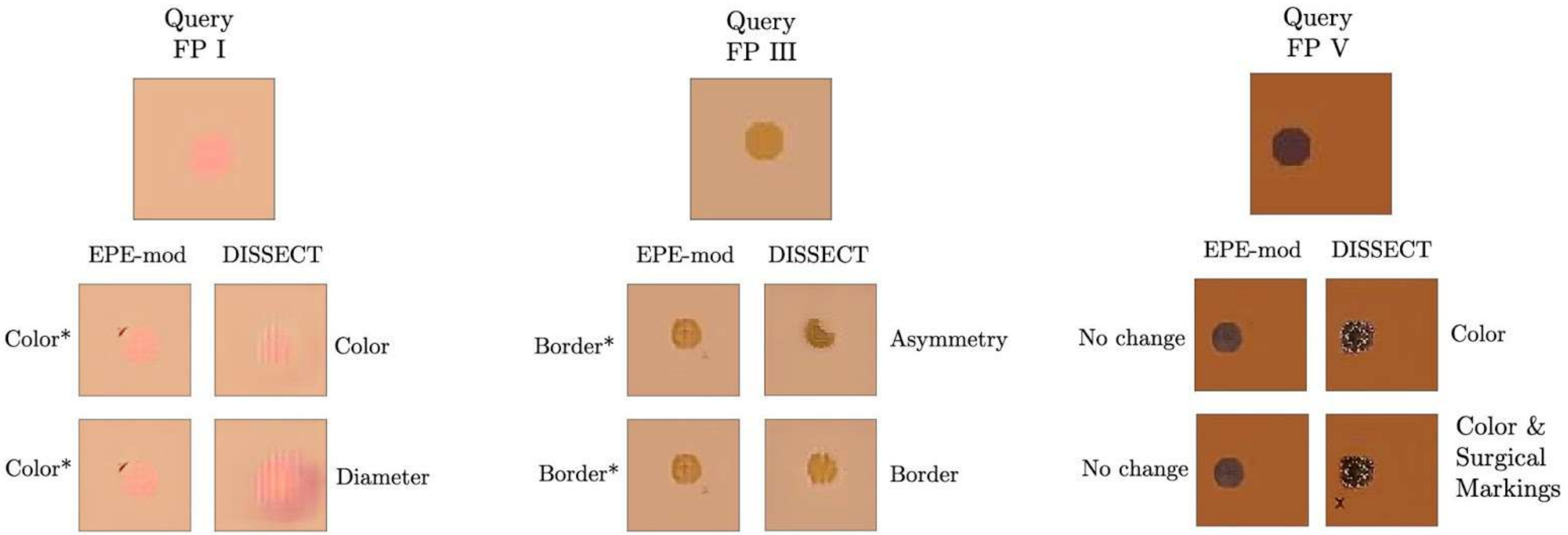


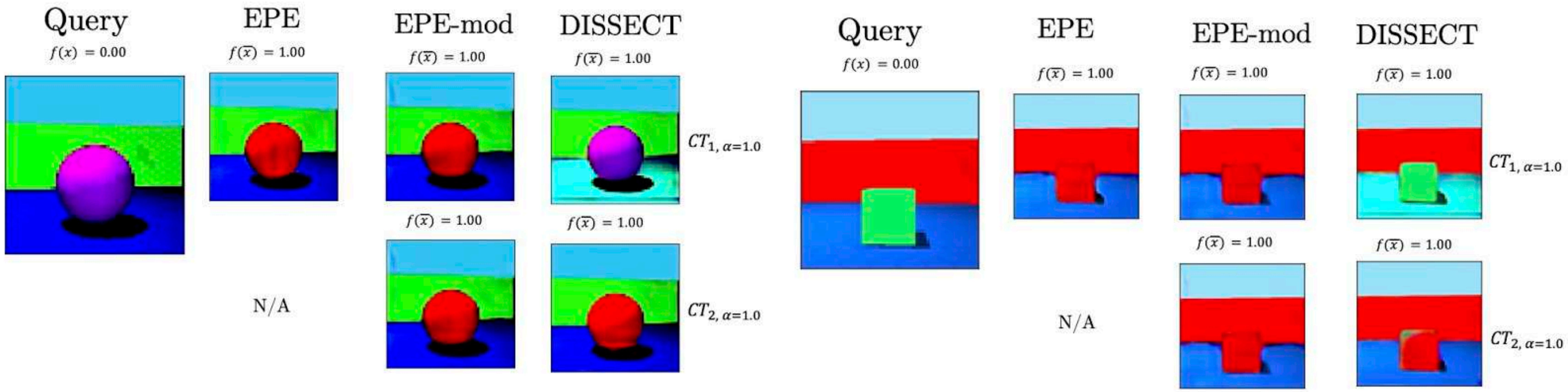
Slide credit: Asma Ghandeharioun

[1] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. ICML, 2018.

Concept traversals in dermatology and 3D shapes dataset





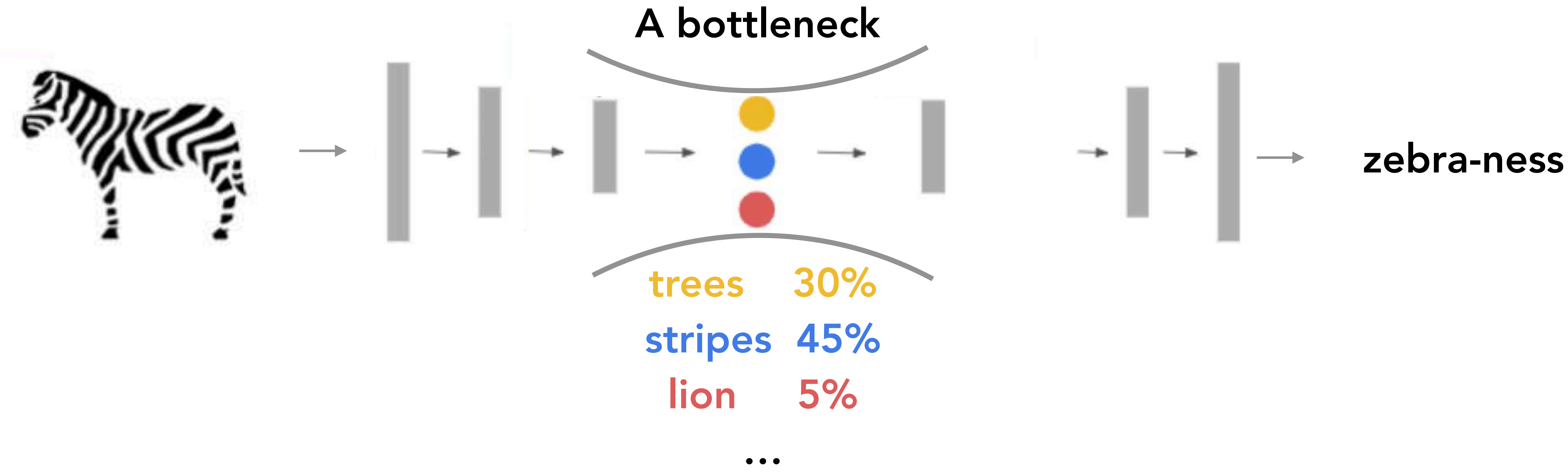


Slide credit: Asma Ghandeharioun

[1] Fitzpatrick, Thomas B. The validity and practicality of sun-reactive skin types I through VI. Archives of dermatology 124.6 (1988): 869-871.

Ok, great but…

Can we flip this around and build a new model?



A bottleneck

zebra-ness

trees 30%

stripes 45%

lion 5%

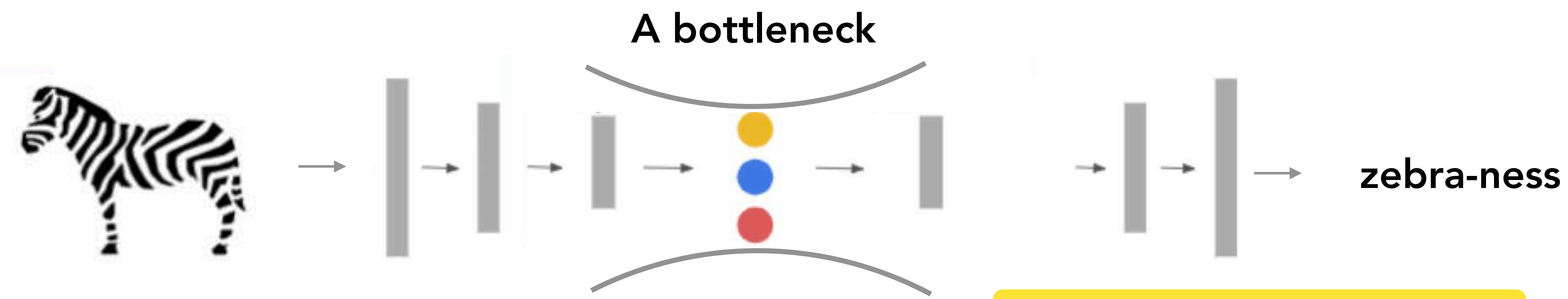
…

Concept bottleneck models

[Goh et al., ICML 20]

Concept bottleneck models

[Goh et al., ICML 20]



A bottleneck

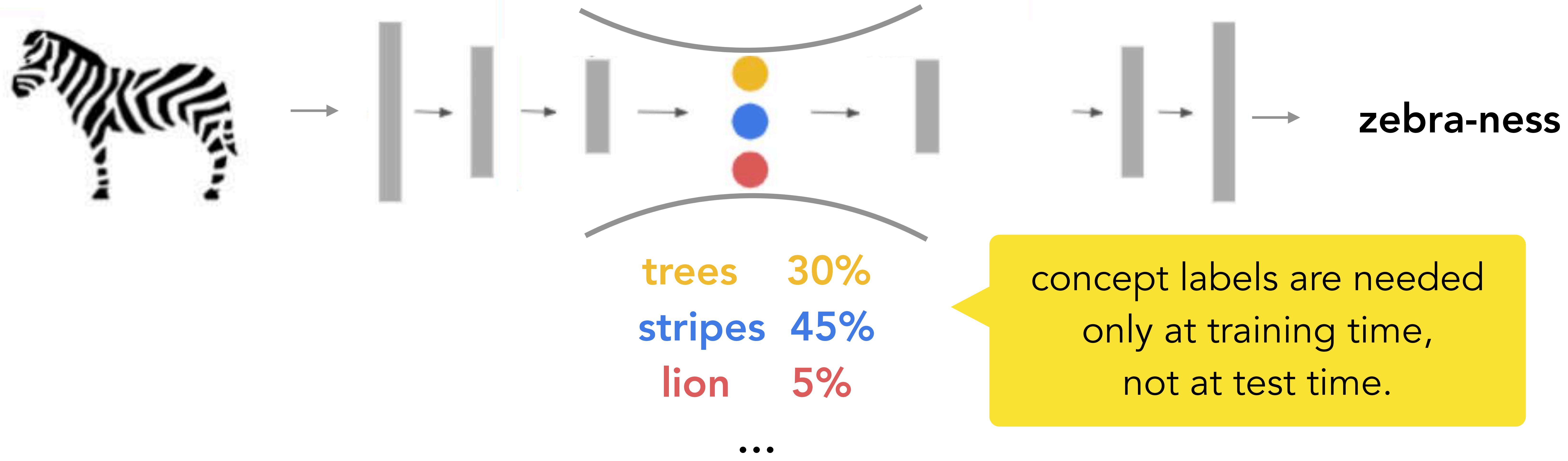
zebra-ness

trees 30%
stripes 45%
lion 5%
...

concept labels are needed
only at training time,
not at test time.

Concept bottleneck models

[Goh et al., ICML 20]



concept labels are needed
only at training time,
not at test time.

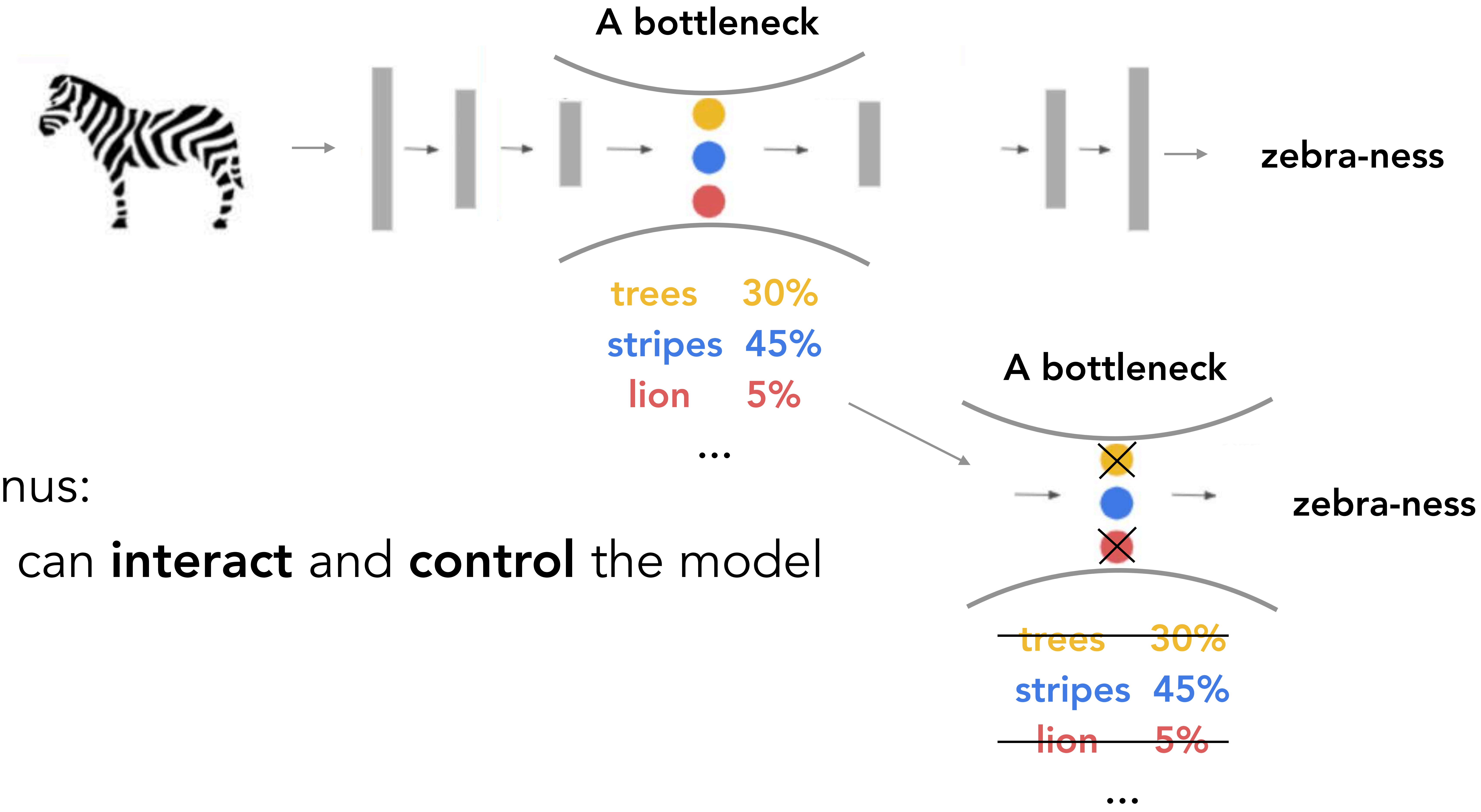
First thing to check: is the performance impacted? - No.

concept bottleneck models

MODEL	y RMSE (OAI)	y ERROR (CUB)
INDEPENDENT	0.435± 0.024	0.240±0.012
SEQUENTIAL	0.418± 0.004	0.243±0.006
JOINT	0.418± 0.004	0.199±0.006
STANDARD	0.441± 0.006	0.175±0.008
NO BOTTLENECK	0.443 ± 0.008	0.173±0.003

Concept bottleneck models

[Goh et al., ICML 20]



A bottleneck

zebra-ness

trees 30%

stripes 45%

lion 5%

...

A bottleneck

zebra-ness

~~trees 30%~~

stripes 45%

~~lion 5%~~

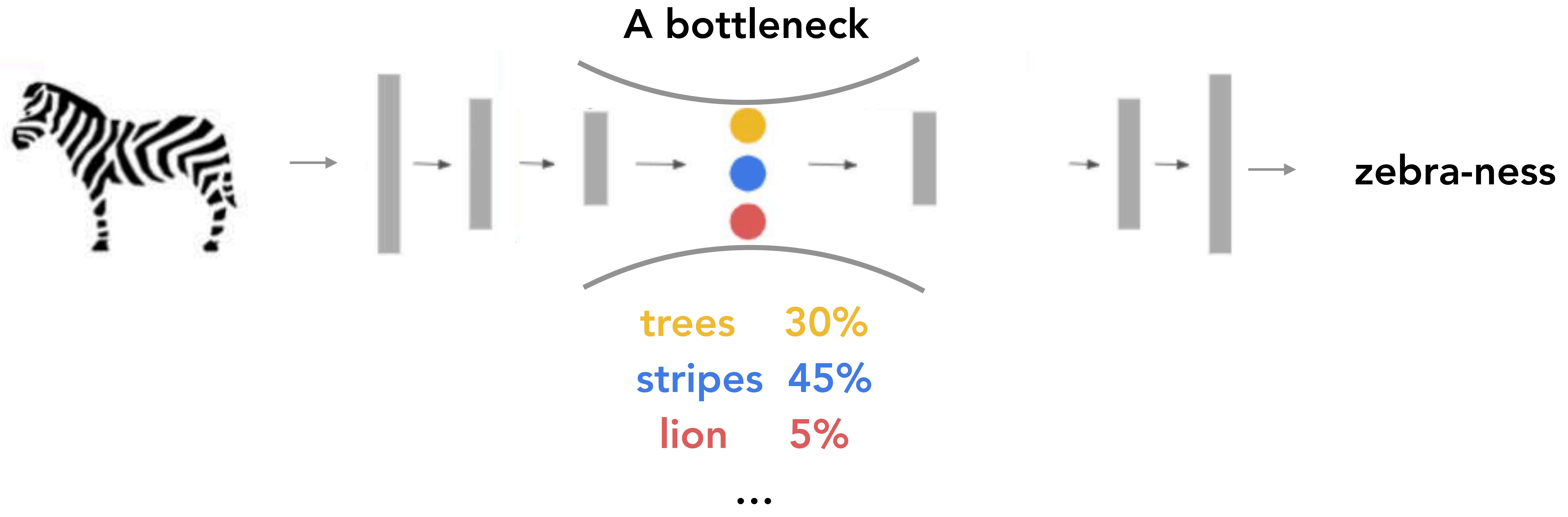
...

Bonus:

we can **interact** and **control** the model

Concept bottleneck models

[Goh et al., ICML 20]



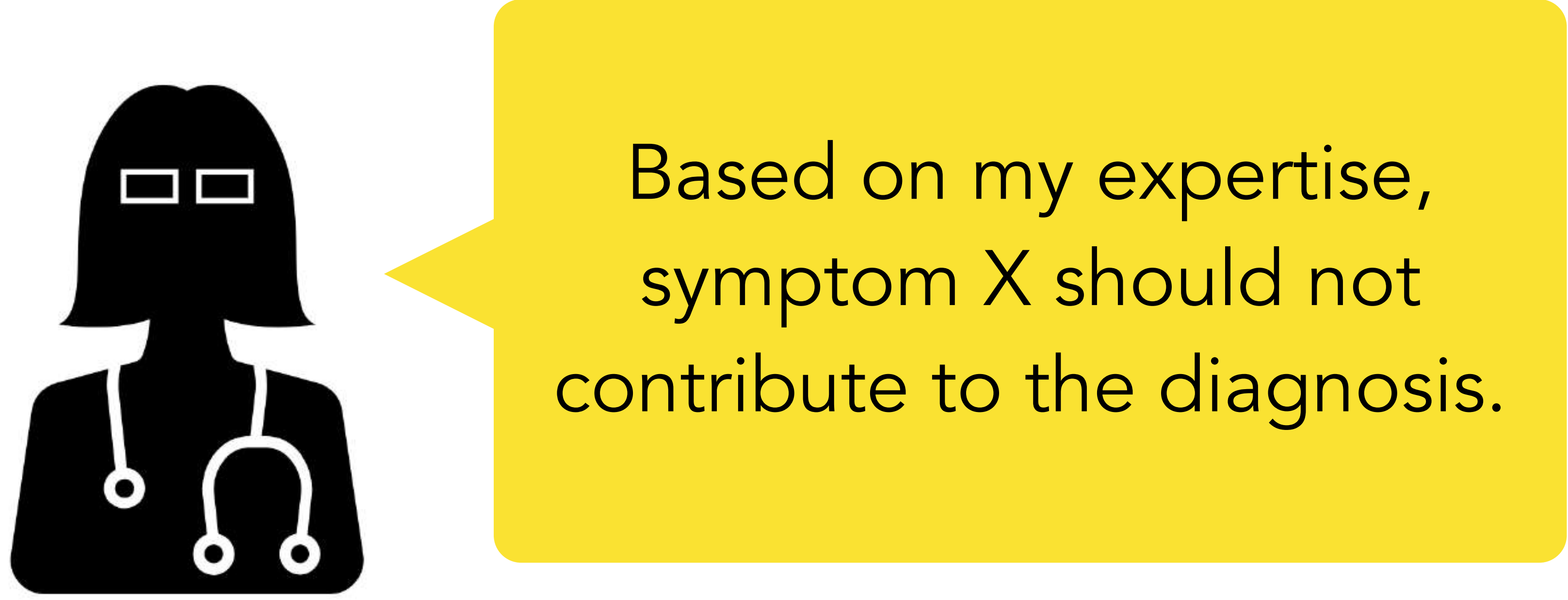
A bottleneck

zebra-ness

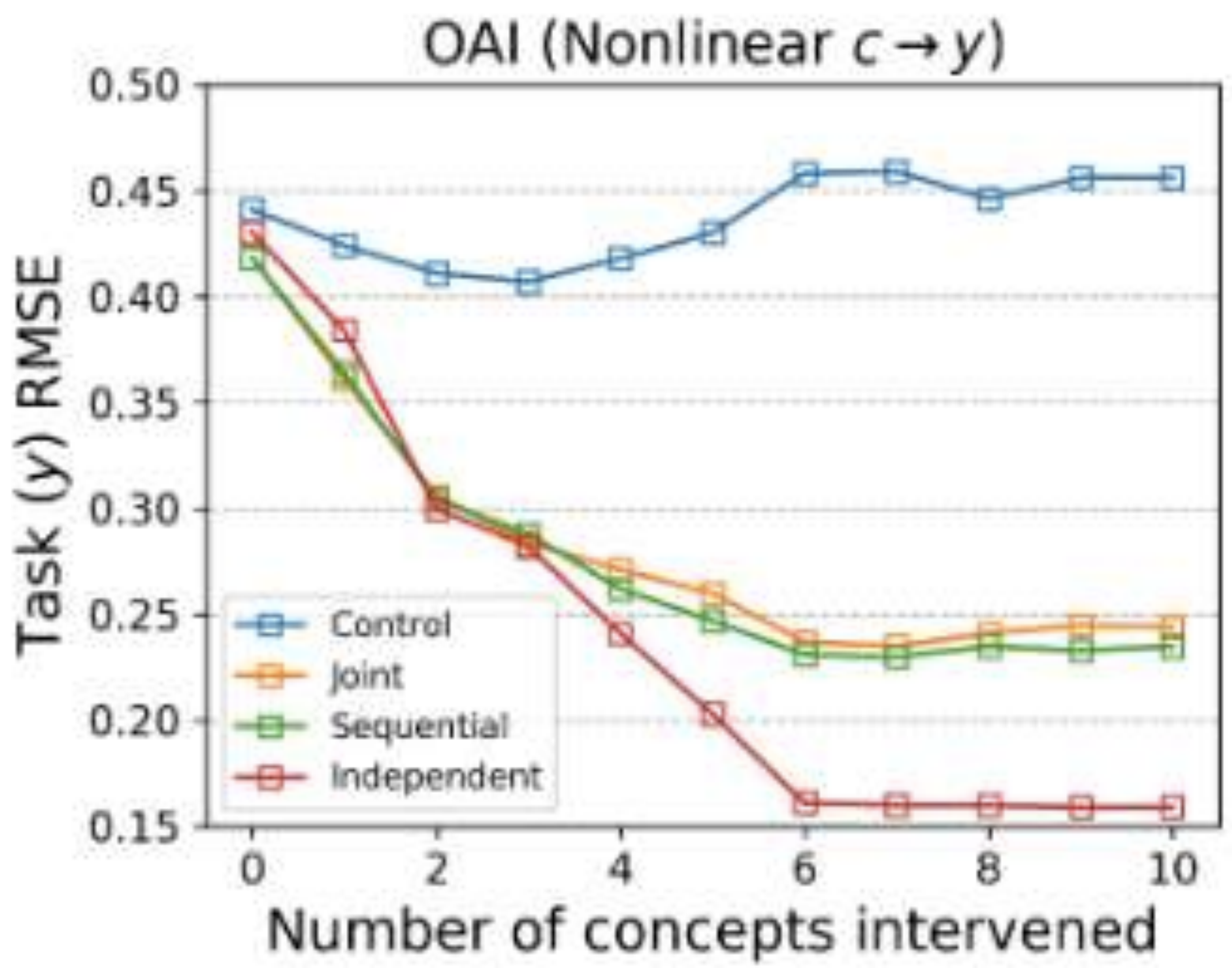
trees 30%
stripes 45%
lion 5%
...

Bonus:

we can **interact** and **control** the model

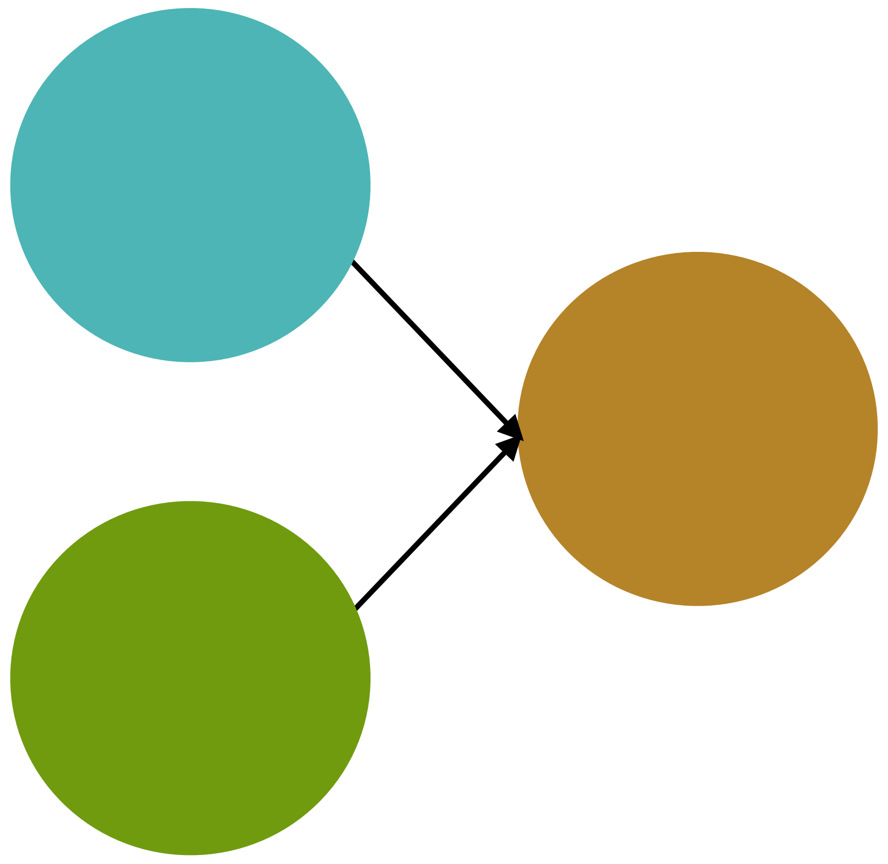


Based on my expertise, symptom X should not contribute to the diagnosis.



Ok, great but…

what about causality?



Based on my expertise, symptom X should not contribute to the diagnosis.

CaCE: Causal TCAV score.

[Goyal et al., 20]


Yash Goyal

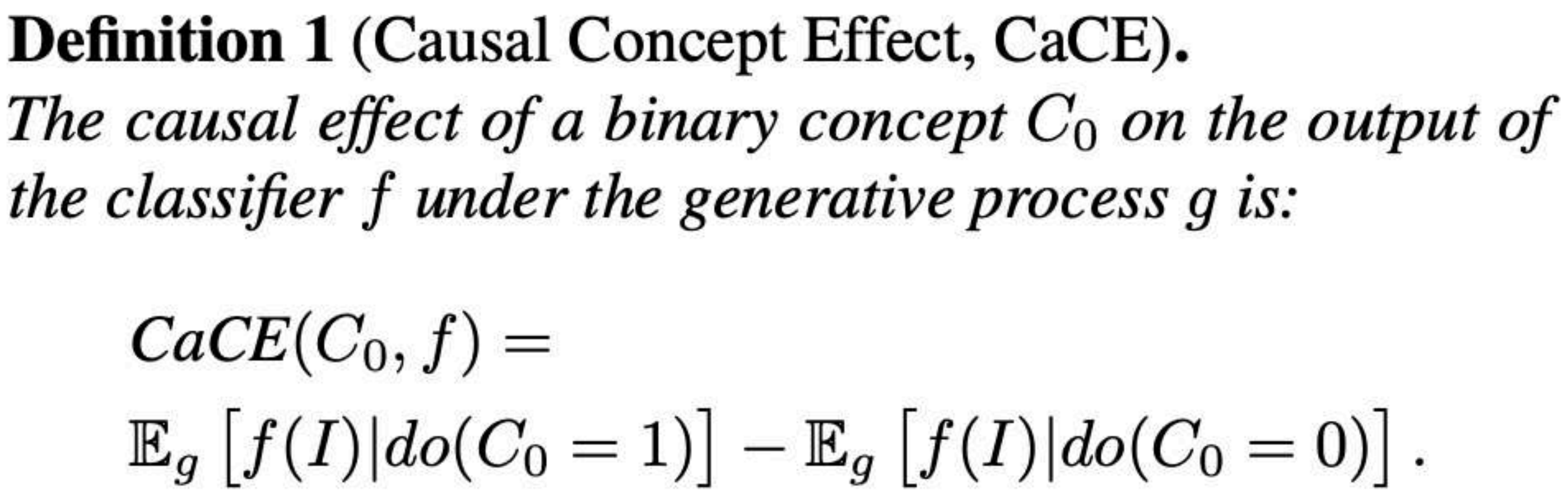

Amir Feder


Uri Salit

CaCE: Causal TCAV score.

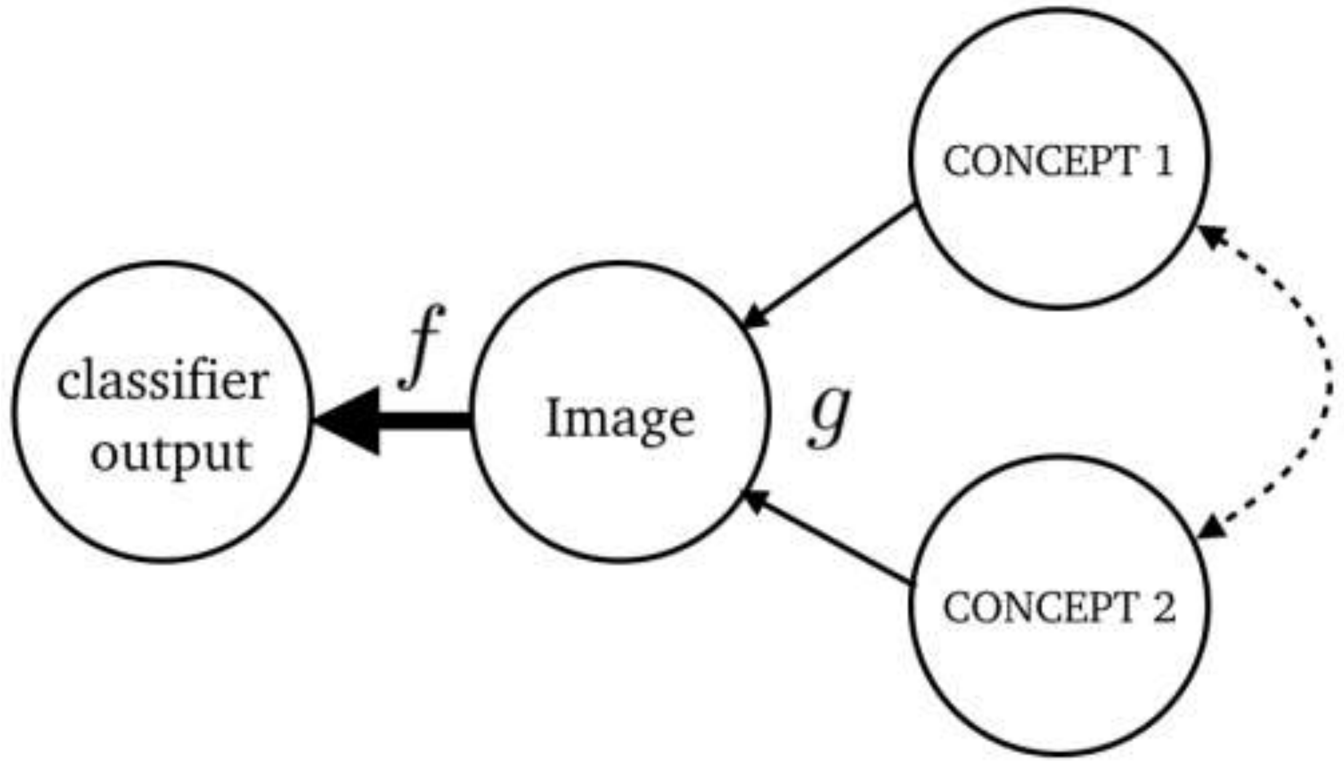
[Goyal et al., 20]

- Basic idea: Do operations on concepts using a generative model to produce $do(C_0 = 1)$ and $do(C_0 = 0)$ Calculate ATE.



Definition 1 (Causal Concept Effect, CaCE).
The causal effect of a binary concept C_0 on the output of the classifier f under the generative process g is:

$$CaCE(C_0, f) = \mathbb{E}_g\left[f(I)|do(C_0 = 1)\right] - \mathbb{E}_g\left[f(I)|do(C_0 = 0)\right].$$



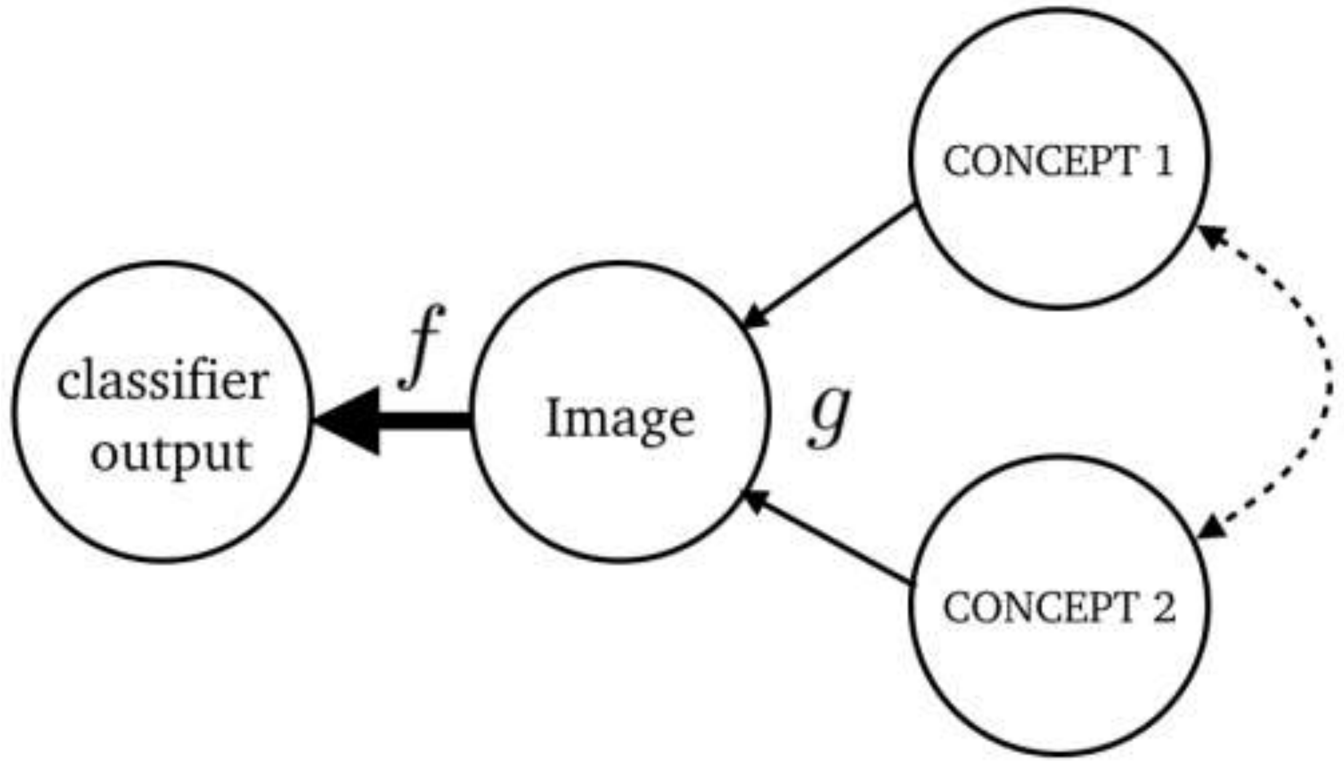
CaCE: Causal TCAV score.

[Goyal et al., 20]

- Basic idea: Do operations on concepts using a generative model to produce $do(C_0 = 1)$ and $do(C_0 = 0)$. Calculate ATE.

Definition 1 (Causal Concept Effect, CaCE).
The causal effect of a binary concept C_0 on the output of the classifier f under the generative process g is:

$$CaCE(C_0, f) =$$
$$\mathbb{E}_g\left[f(I)|do(C_0 = 1)\right] - \mathbb{E}_g\left[f(I)|do(C_0 = 0)\right].$$



e.g., gender classifier f, ATE with glasses concept?

$\mathbb{E}_g [f($  $)|do(C_0 = 1)] - \mathbb{E}_g [f($  $)|do(C_0 = 0)]$

p(woman) = 0.88

glasses

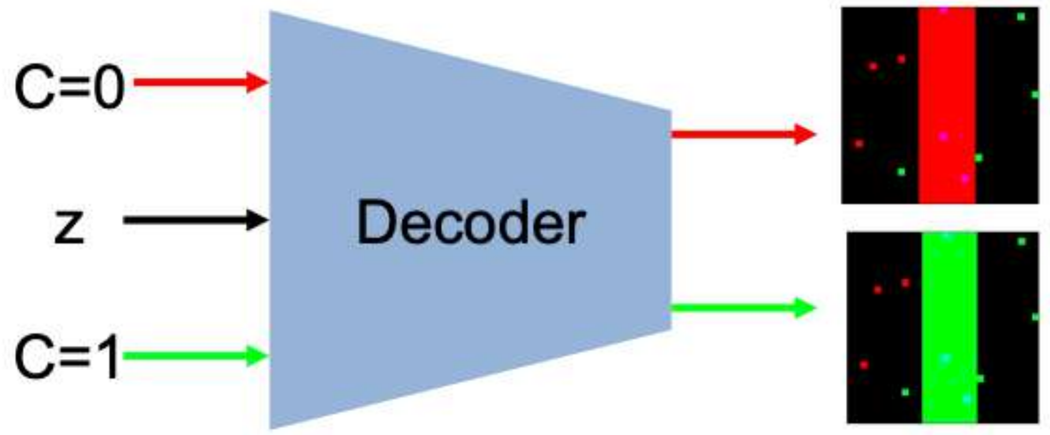
p(woman) = 0.4

no glasses

CaCE: Causal TCAV score.

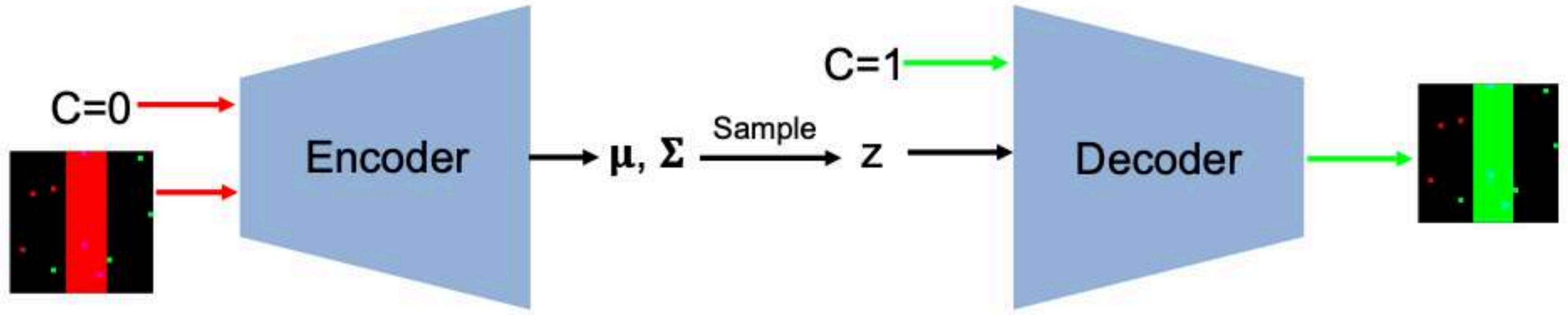
[Goyal et al., 20]

- Can we train a generative model ‘good enough’ to make this work?



Dec-CaCE

Use sampled z
testing general distributions



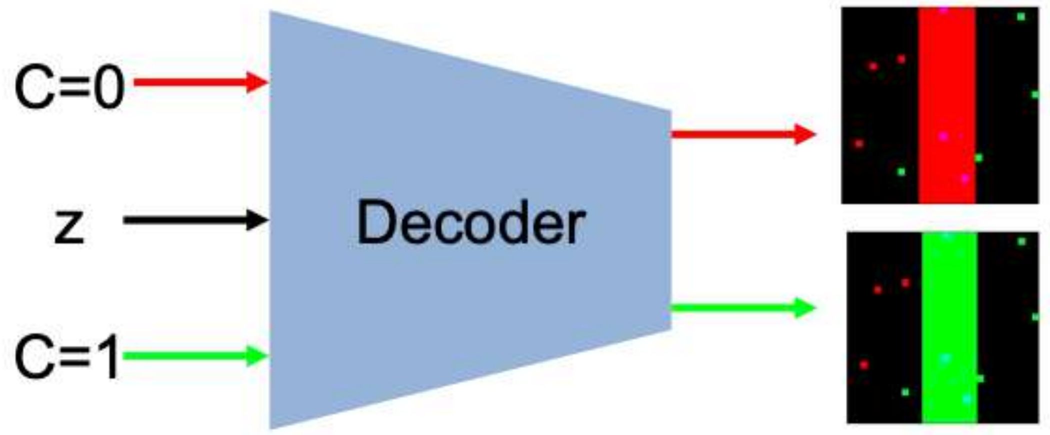
EncDec-CaCE

Use a particular instance
testing particular population

CaCE: Causal TCAV score.

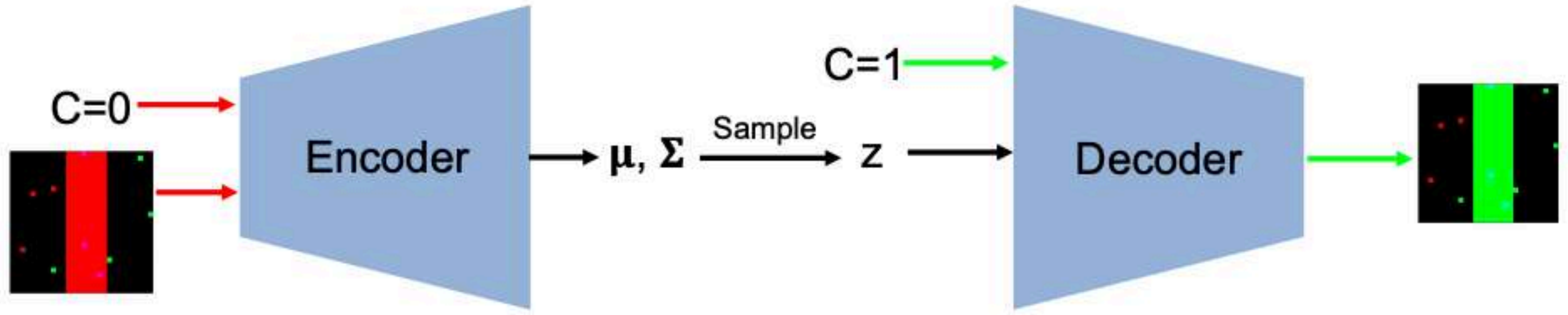
[Goyal et al., 20]

- Can we train a generative model 'good enough' to make this work?



Dec-CaCE

Use sampled z
testing general distributions



EncDec-CaCE

Use a particular instance
testing particular population

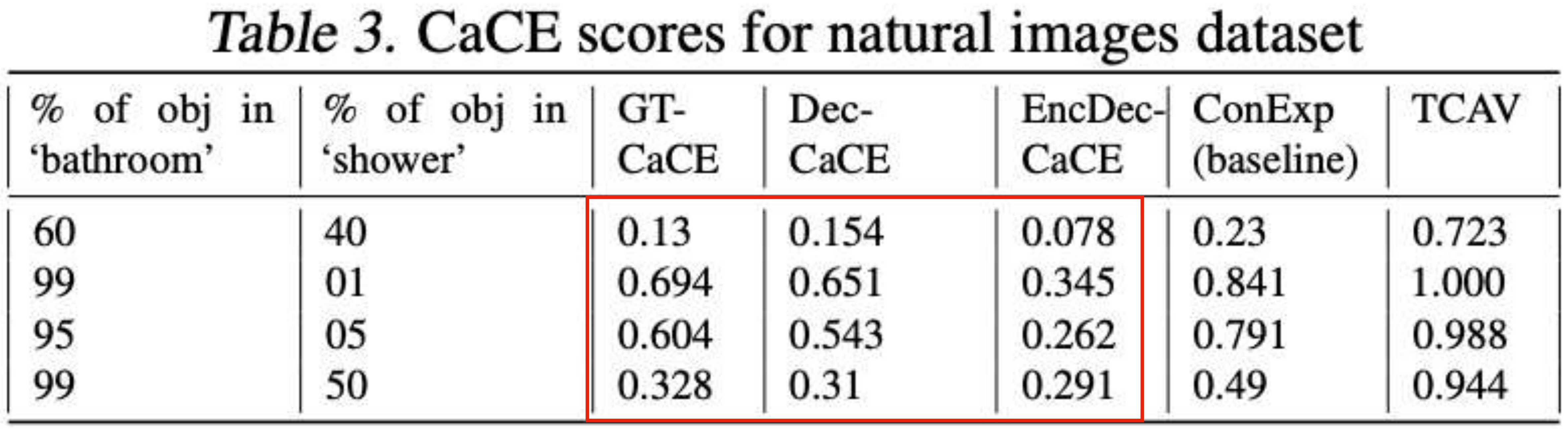
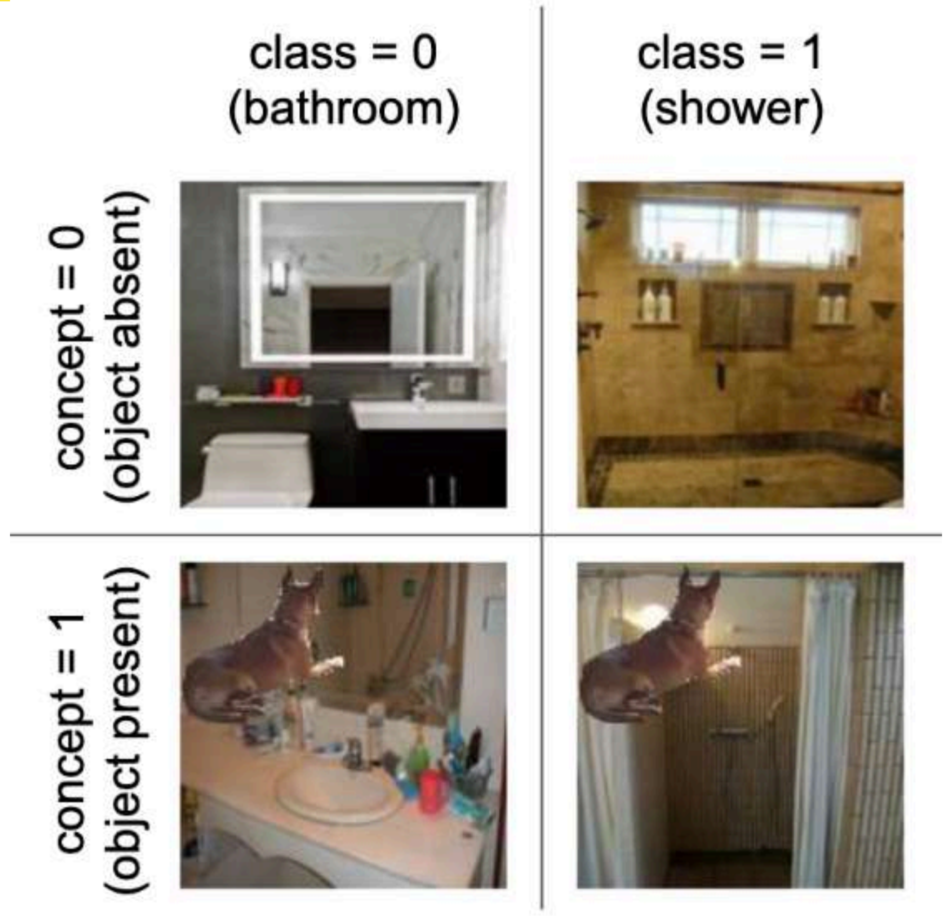


Table 3. CaCE scores for natural images dataset

% of obj in 'bathroom'	% of obj in 'shower'	GT-CaCE	Dec-CaCE	EncDec-CaCE	ConExp (baseline)	TCAV
60	40	0.13	0.154	0.078	0.23	0.723
99	01	0.694	0.651	0.345	0.841	1.000
95	05	0.604	0.543	0.262	0.791	0.988
99	50	0.328	0.31	0.291	0.49	0.944

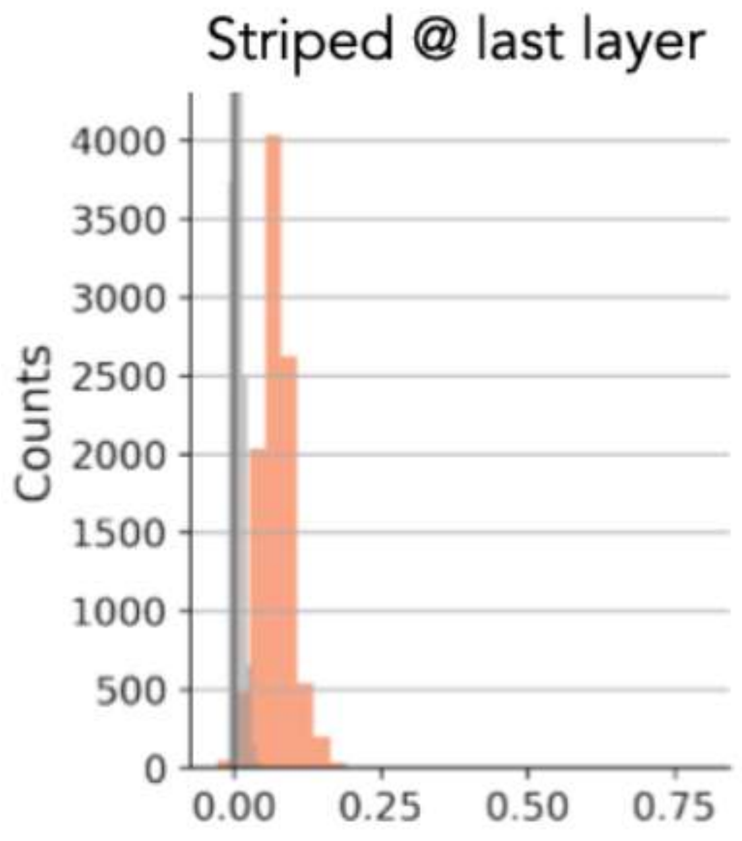


Natural
images
dataset

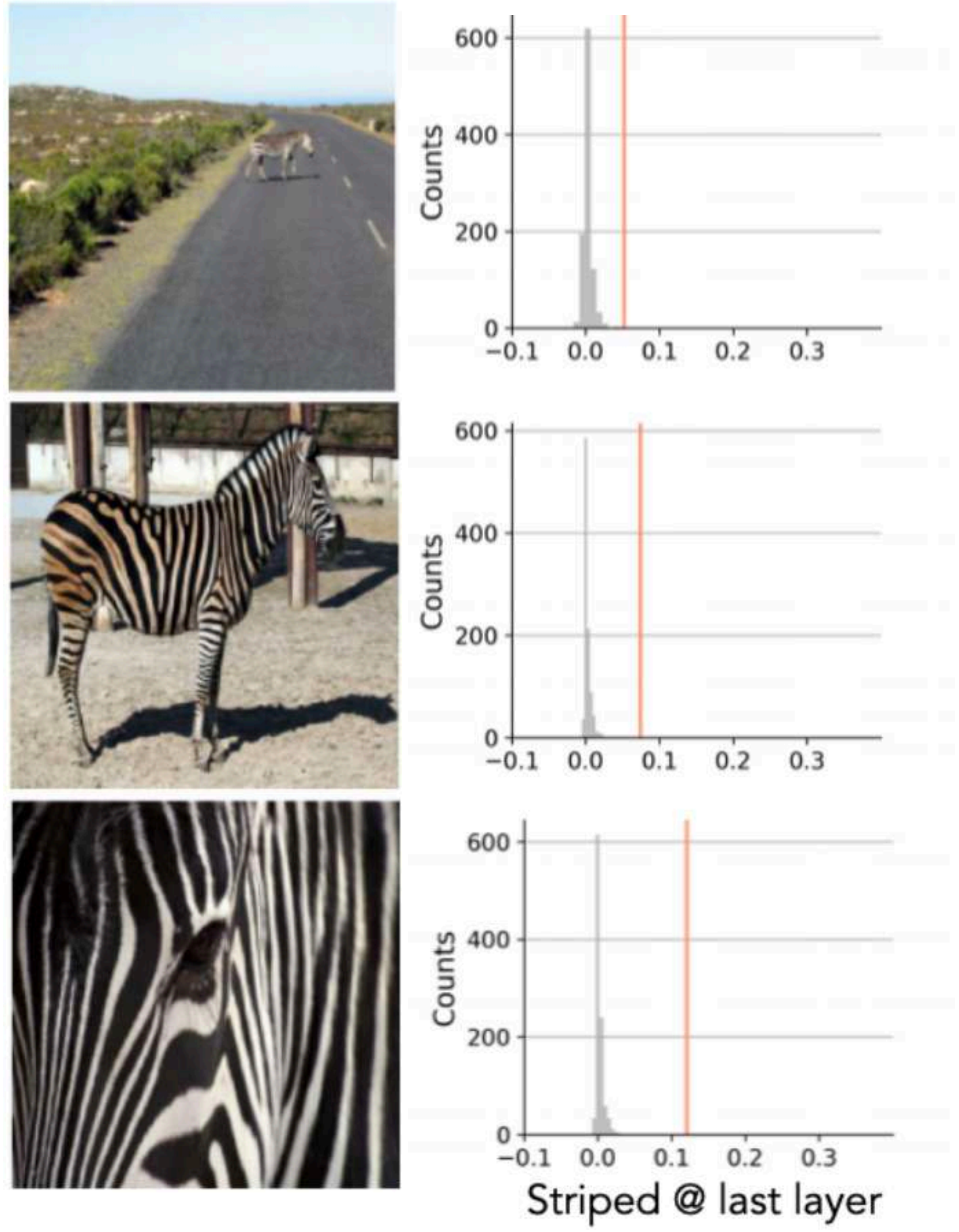
But

These are all global explanations only.. what about local? [ongoing work]

All the zebras



My zebras!

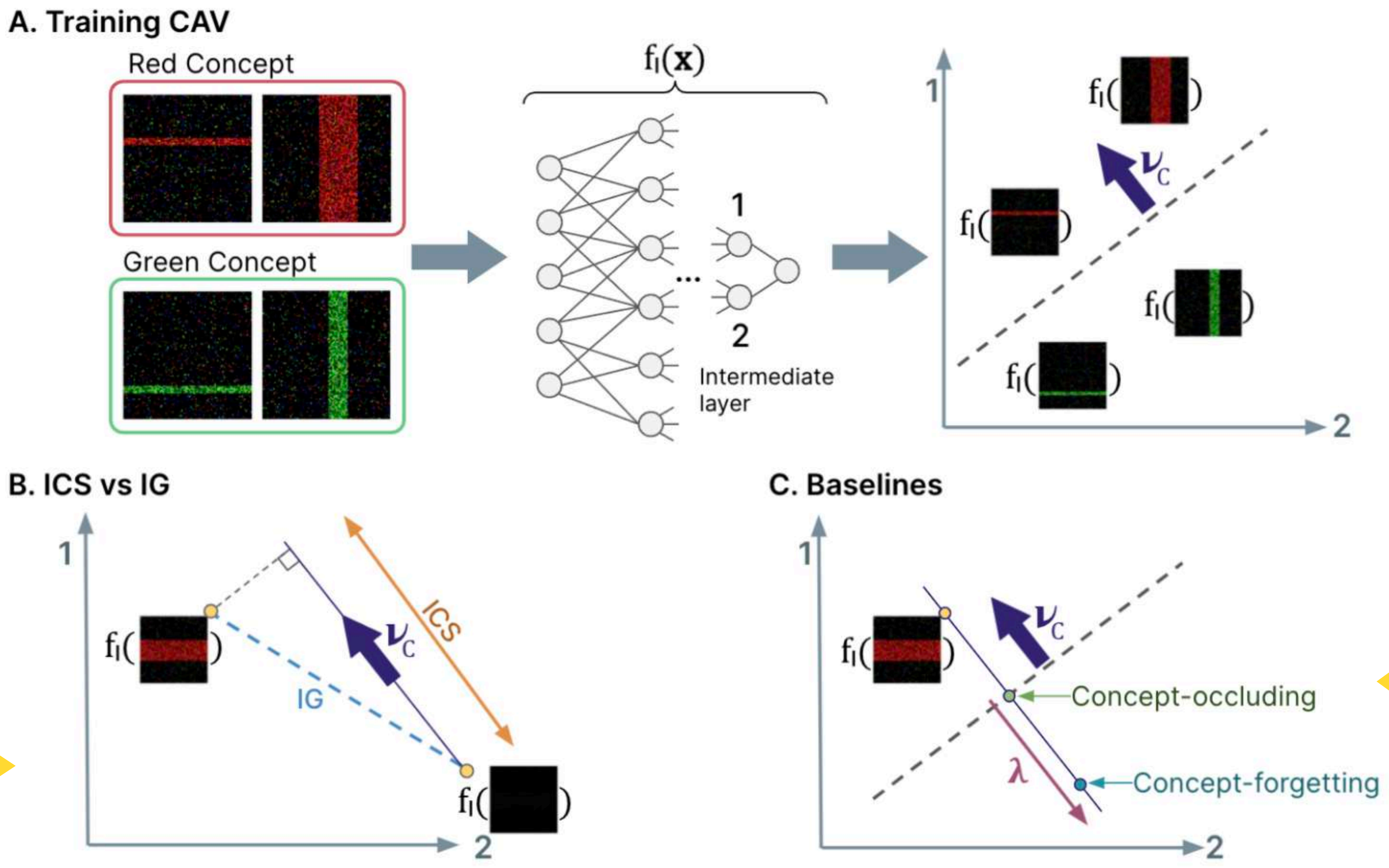


Combine TCAV + IG to provide both global and local explanations

Jessica Schrouff

Sebastien Baur

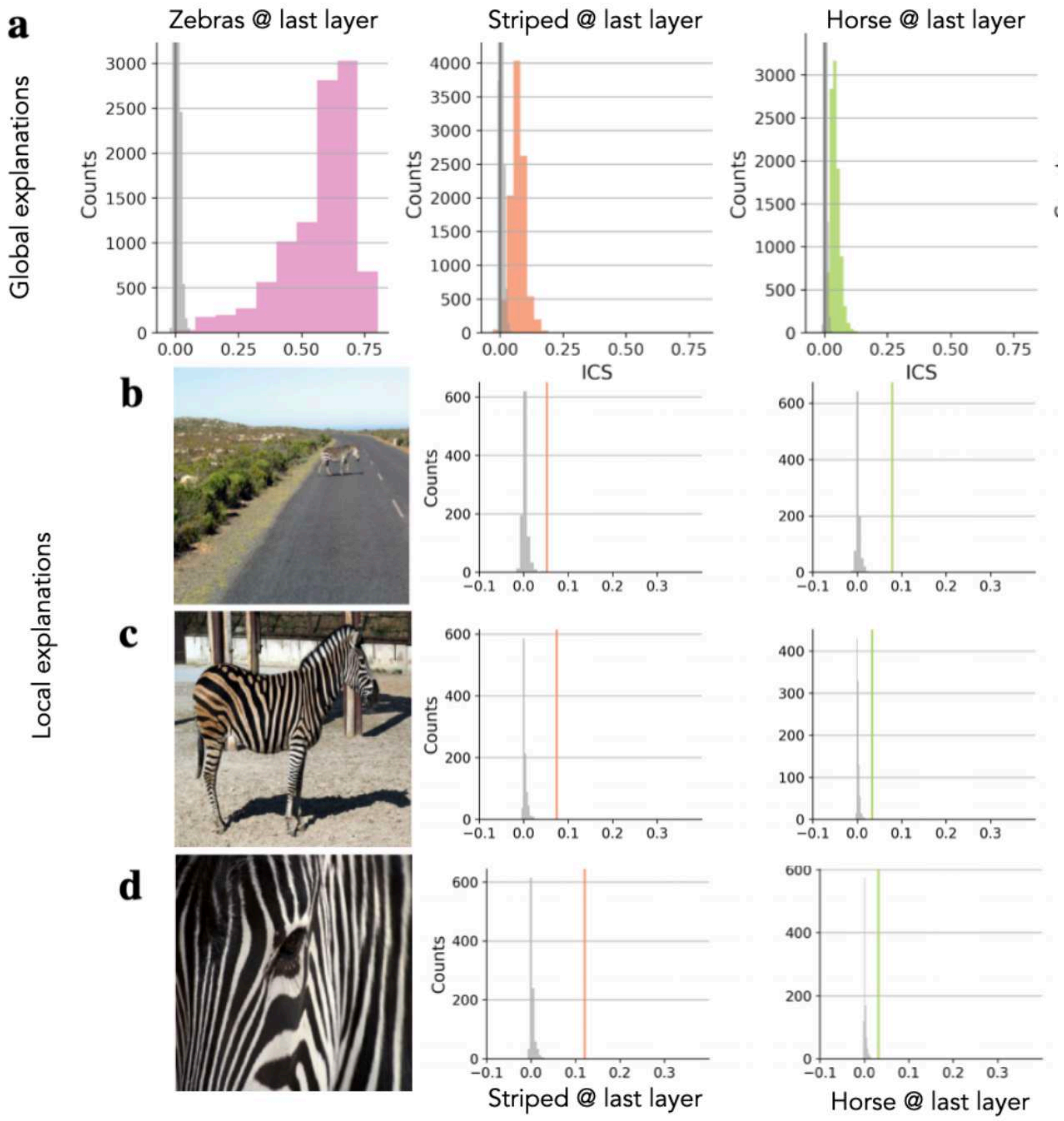
Combine TCAV + IG to provide both global and local explanations

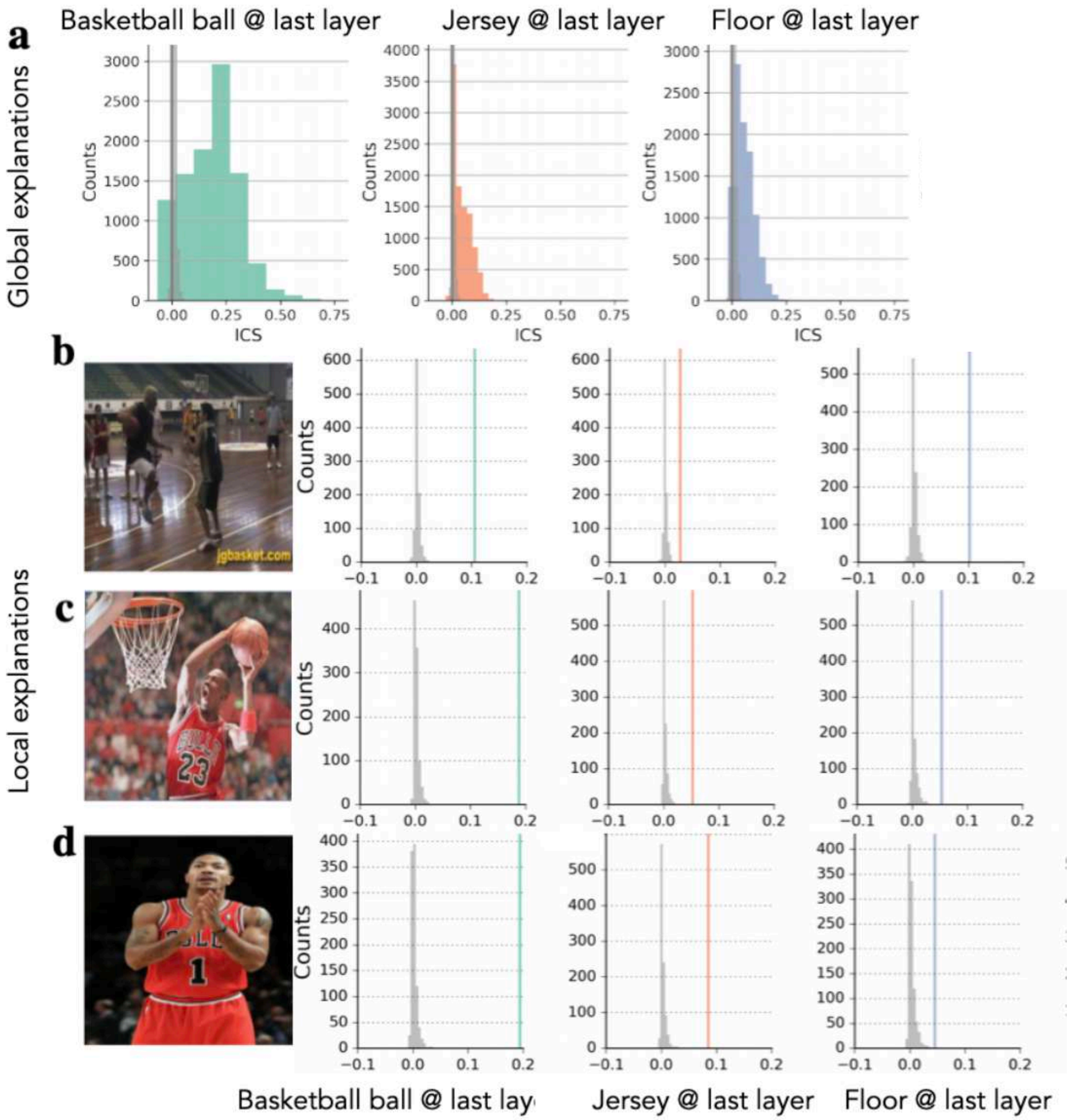


Integrate on the path of a CAV
<->
A projection of path integration

New baselines for concepts

Combine TCAV + IG to provide both global and local explanations

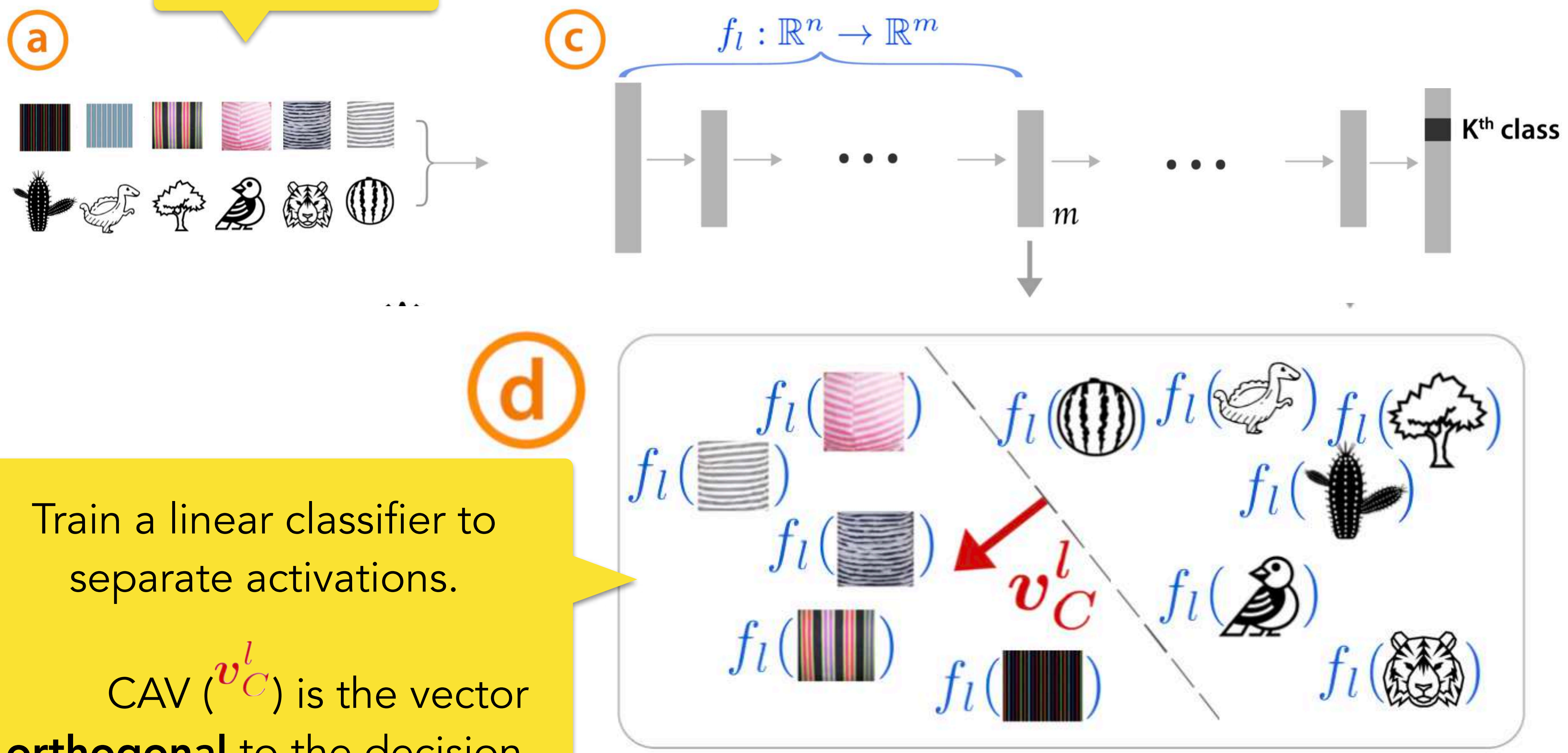




no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

Examples of concepts



Train a linear classifier to separate activations.

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

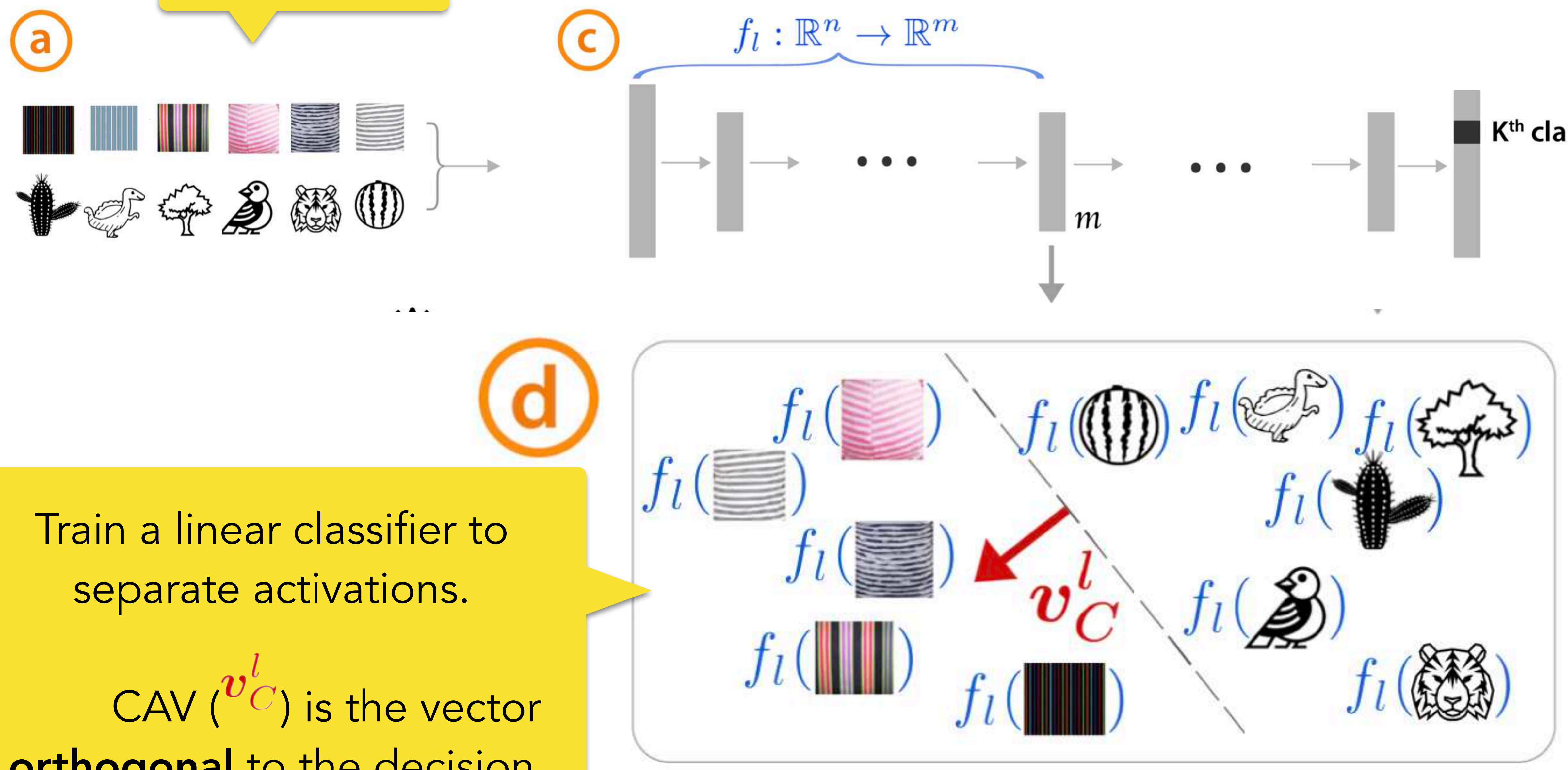
How important was the striped concept to this zebra image classifier?

[K., Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres ICML 2018]

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

Examples of concepts



Train a linear classifier to separate activations.

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

zebra-ness →

striped CAV →

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

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

How important was the striped concept to this zebra image classifier?

[K., Wattenberg, Gilmer, Cai, Wexler, Viegas, Sayres ICML 2018]

no humans, proxy task

1: Test hypothesis that should be true by craft a <u>ground-truth dataset</u>

An image

+

Potentially noisy Caption

no humans, proxy task

1: Test hypothesis that should be true by craft a <u>ground-truth dataset</u>

caption
concept

An image
+
Potentially noisy Caption

image
concept

models can use either
image or caption
concept for classification.

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

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

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

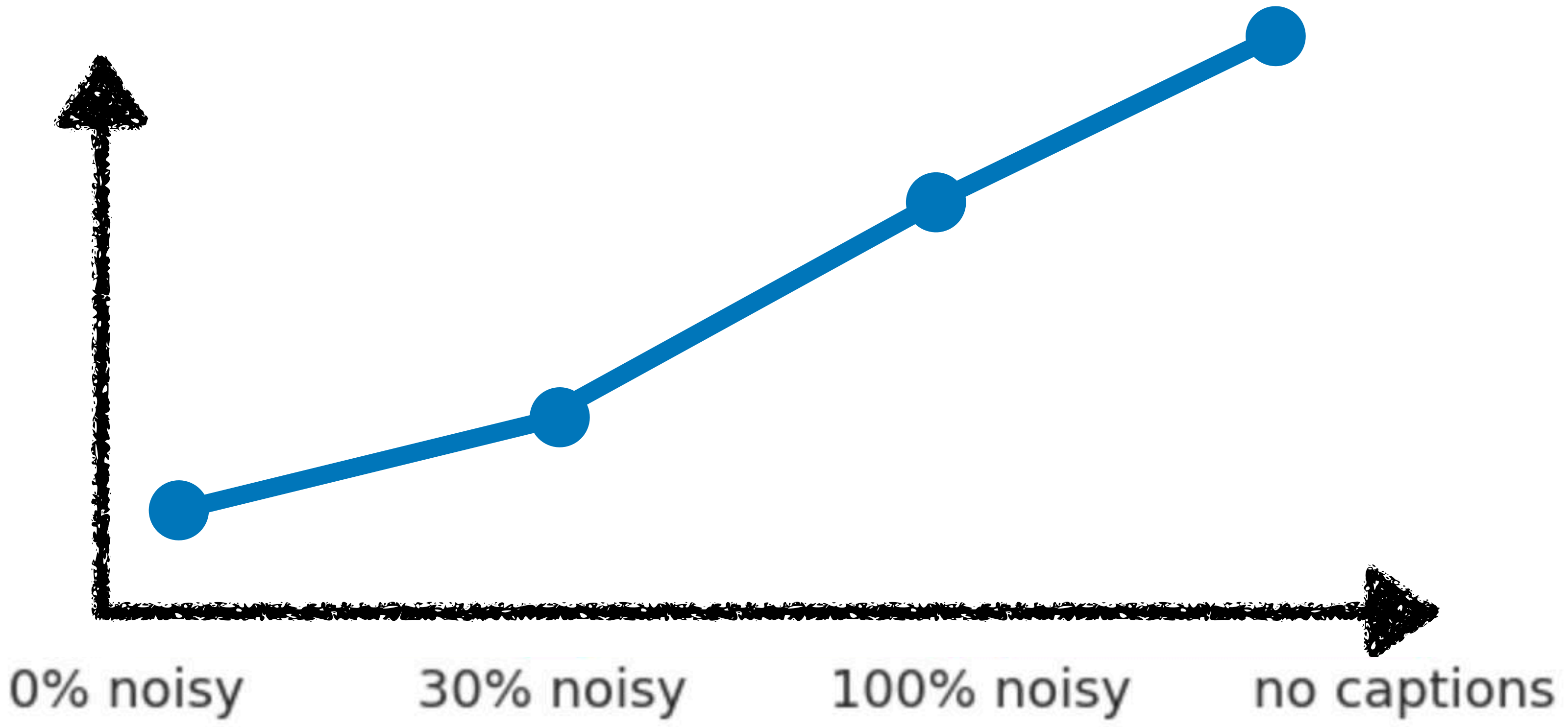
caption concept

image concept

models can use either image or caption concept for classification.

Test models with no caption image.

Test accuracy = Importance of image concept

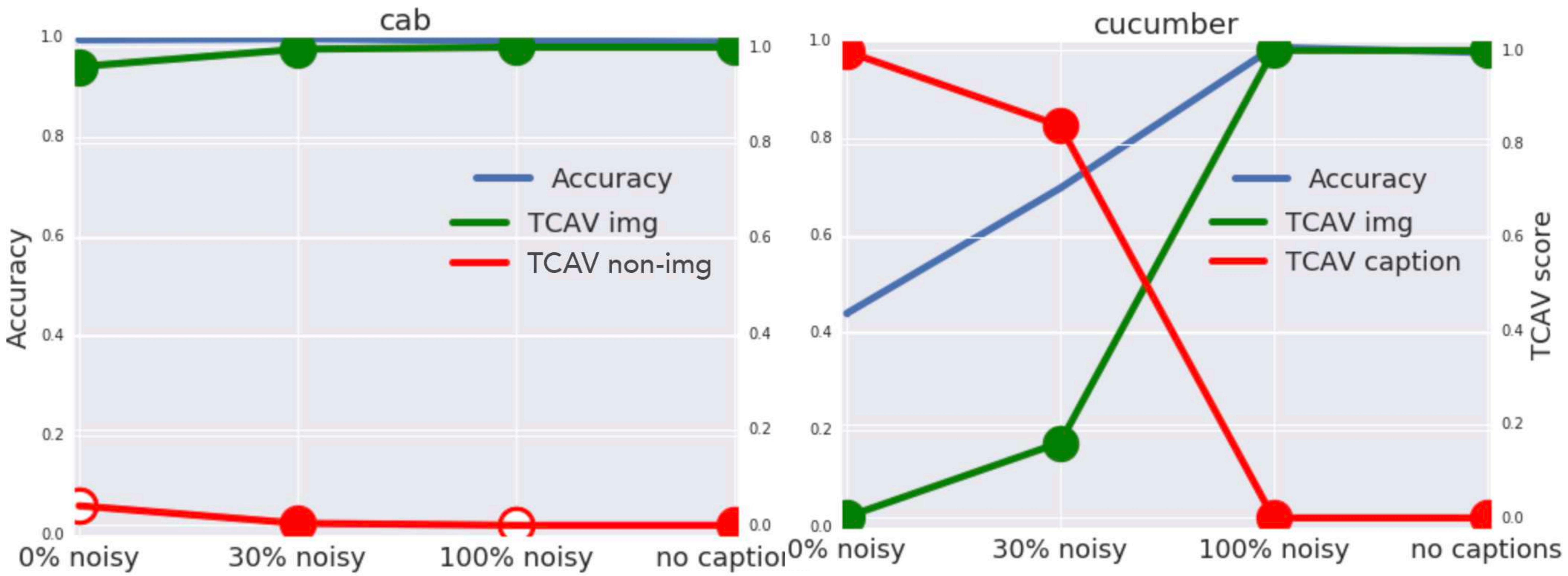


Four models trained with different caption noise levels

no humans, proxy task

1: Test hypothesis that should be true by craft a ground-truth dataset

Test accuracy
with
no caption image



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