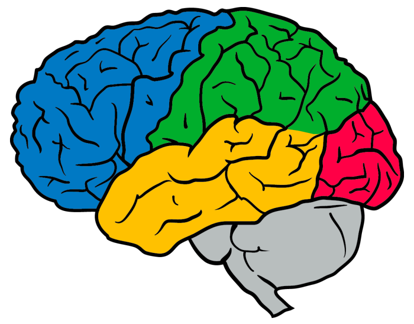


Interpretability for philosophical and skeptical minds

Been Kim

What is the best explanation? From philosophers

- Many attempts to come up with **a single model of explanation**
    - Deductive Nomological (1942, Hempel) Statistical relevance (1971, Salmon), Causal Mechanical (1984, Salmon), Unificationist (1974, Freidman, 1989, Kitcher) with the hope that there exists ONE OPTIMAL model for explanations.
- Then pragmatic theories (1980, van Fraassen) came out.

> The discussion of explanation went wrong at the very beginning when explanation was conceived of as a relation like description: a relation between a theory and a fact. Really, it is a three-term relation between theory, fact, and context. No wonder that no single relation between theory and fact ever managed to fit more than a few examples! Being an explanation is essentially relative for an explanation is an *answer*… it is evaluated vis-à-vis a question, which is a request for information. But exactly… what is requested differs from context to context. (1980: 156)

- The importance of "context" is something that ML community also came to (generally) agree.
- I doubt we will over find a single best model of explanation without context.

Carl Gustav Hempel

Wesley C. Salmon

Bas van Fraassen

Reference: https://plato.stanford.edu/entries/scientific-explanation/#Bib

What is the best explanation?
Illuminating example

- Structural explanation by Prof. Sally Haslanger
- The Invisible Foot (Okin 1989, Cudd 2006): Lisa and Larry, equally intelligent and talented at work, both capable of taking care of a child. But they live in a society where there is wage gap between men and women. They don't have means to pay for childcare. Lisa decides to quit her job.

Prof. Haslanger

- What's the "best" explanations for "why did Lisa quit her job?"
  - Why did **Lisa** quit her job and Larry?
  - Why did Lisa **quit** instead of going part time?
  - The society that unconsciously shaped her preference? "I'm not as good as Larry".
  - The society that created the bias and wage gap?
  - ...

What is the best explanation?
Illuminating example

- Structural explanation by Prof. Sally Haslanger
- The Invisible Foot (Okin 1989, Cudd 2006): Lisa and Larry, equally intelligent and talented at work, both capable of taking care of a child. But they live in a society where there is wage gap between men and women. They don't have means to pay for childcare. Lisa decides to quit her job.

Prof. Haslanger

- What's the "best" explanations for "why did Lisa quit her job?"
  - Why did **Lisa** quit her job and Larry?
  - Why did Lisa **quit** instead of going part time?
  - The society that unconsciously shaped her preference? "I'm not as good as Lar
  - The society that created the bias and wage gap?
  - ...

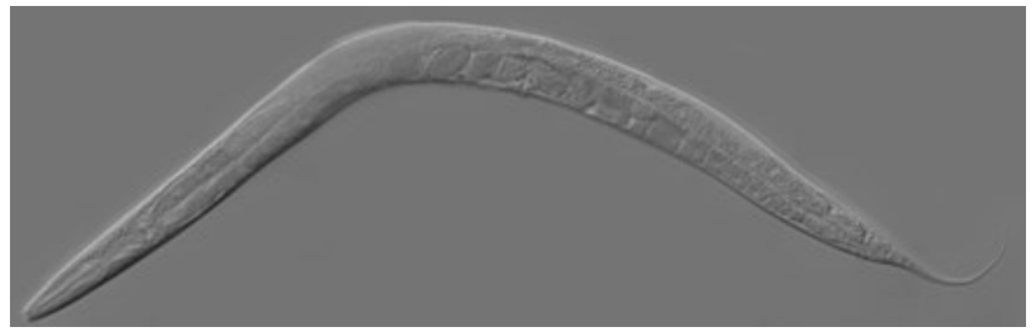
What's the "best" explanations for "why was this predicted as a dog?"

- This and that pixel?
- Other training data and their delicate interaction during training process?
- The choice of architecture or optimizer?
- How the pictures are taken and when?
- The human history of domesticating wolves into dogs...

Wait, why are we talking about philosophy?

- Giving "explanations" isn't a new problem. It's century-old one.
- The complexity of "how/what/when" to explain: it's always more complicated than we think.
- We should not take "good" explanation on its face value: we need to be skeptical (as we will see more soon).

Trying to understand something new isn't new. Neuroscience?



- Understanding human brain: came a long way, but not enough.
  - "**We still don't understand** a worm (Caenorhabditis elegans) with 302 neurons. Humans have **86 billion** of them." - Koch, Allen institute for brain science.
  - "Let's say we could actually record from 1 million neurons in a brain while it's operating. You'd get a lot of data, but **what would we look for**? That is what we have to get some idea of." - Prof. Roland
  - "Throughout, Understanding the Brain reads like a compendium of things we still don't know. **We don't know how many neurons** are in the human brain. [..] **We don't know how alcohol relieves anxiety**, or how dopamine signaling is impaired in schizophrenia[...]" - article

BRAINS! —

Understanding the Brain is a catalog of all we don't know about the brain

Updated version of *Creating Mind* mostly tells us what we don't understand.

DIANA GITIG - 11/10/2018, 7:00 AM

Will It Ever Be Possible to Understand the Human Brain?

Despite technical breakthroughs like Elon Musk's Neuralink, scientists still have no reliable model of how the brain actually works

Brian Bergstein Aug 21, 2019 · 15 min read ★

Oh bummer... Are you still giving this talk?

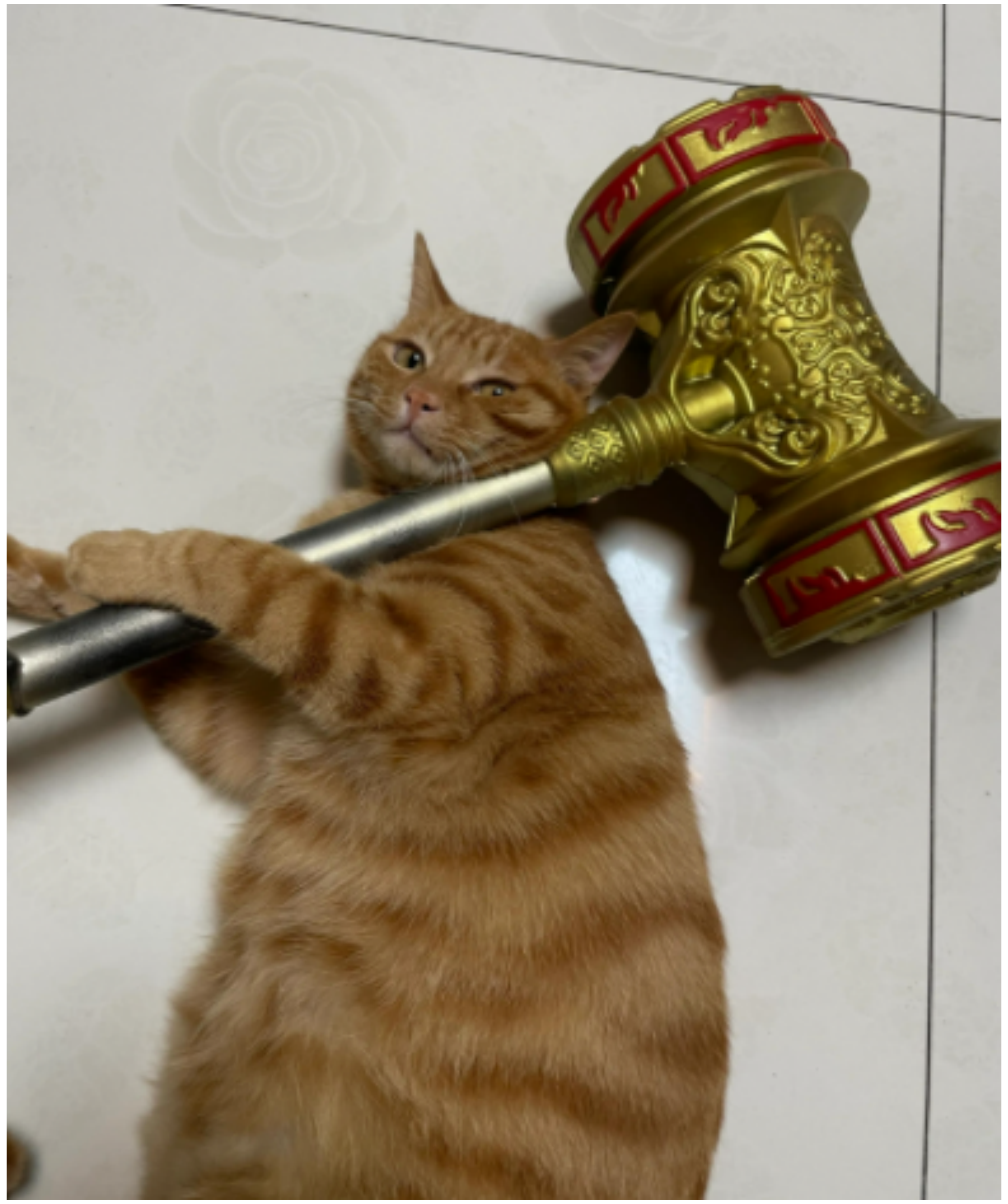
- Yes, neuroscience feels like my future in 40 years... "We still don't understand..."
  - But, I'm still optimistic. Because, while we still don't understand human brain, without a doubt **studying human brain helped the world,** because for example, 1) we have ways to help people via psychological treatments 2) we can sometimes cure seizure (e.g., epilepsy surgery) and the list goes on.
  - The point is: the goal of interpretability is similar. it's not about understanding **everything** all the time. It's about understanding **enough** so that they are **useful.**

What's enough?

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

  [for what I am trying to do = context]"

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



inf.news

What's enough in medicine?

- For example:
  - "Solve" medicine (?)
  - Help doctors to be more effective, efficient, and precise.
  - Use less resources, help more patients.
  - ...
  - ...
  - At minimum, do no harm.

Low bar

What's enough in medicine?

- For example:
  - "Solve" medicine (?)
  - Help doctors to be more effective, efficient, and precise.
  - Use less resources, help more patients.
  - …
  - …
  - At minimum, do no harm.

Low bar

Investigating post-training interpretability methods.

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]

Investigating post-training interpretability methods.

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

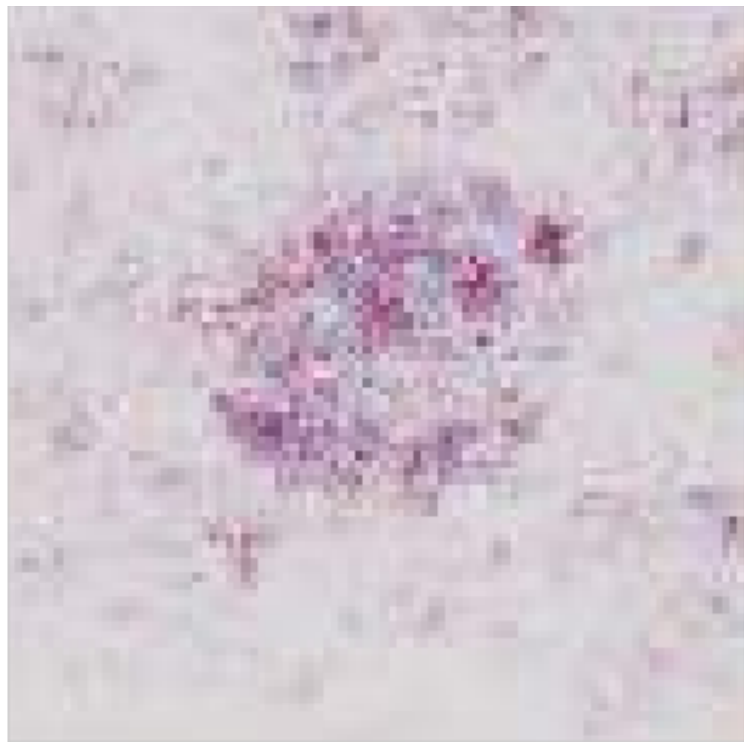
Input image

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.

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

a logit → numerator

pixel i,j → denominator

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

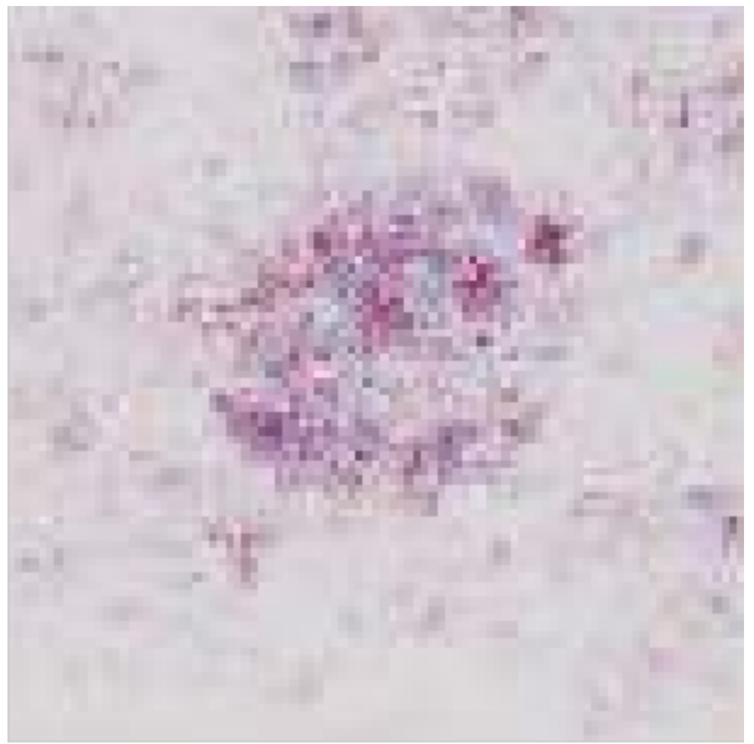
Input image

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

prediction

$p(z)$

Junco Bird-ness



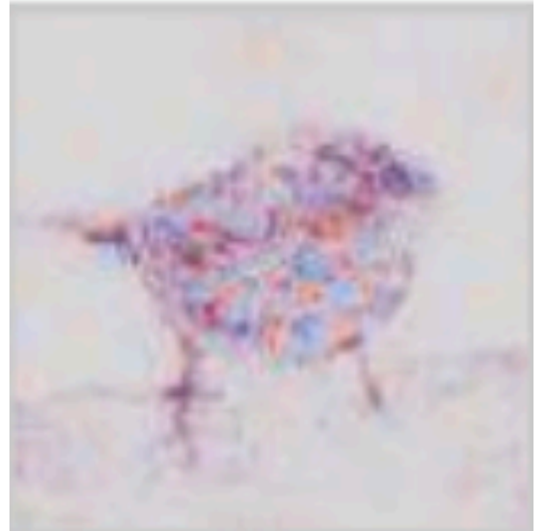
Popular method #1

Popular method #2

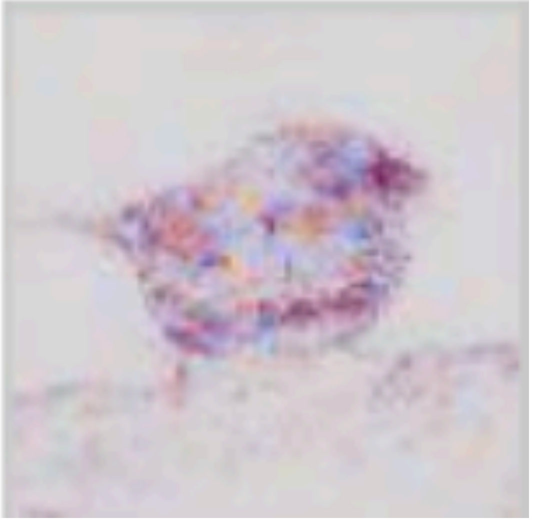
My work from 2018 #1

My work from 2018 #2

Popular method #3



Popular method #4



Sanity check question

Input image

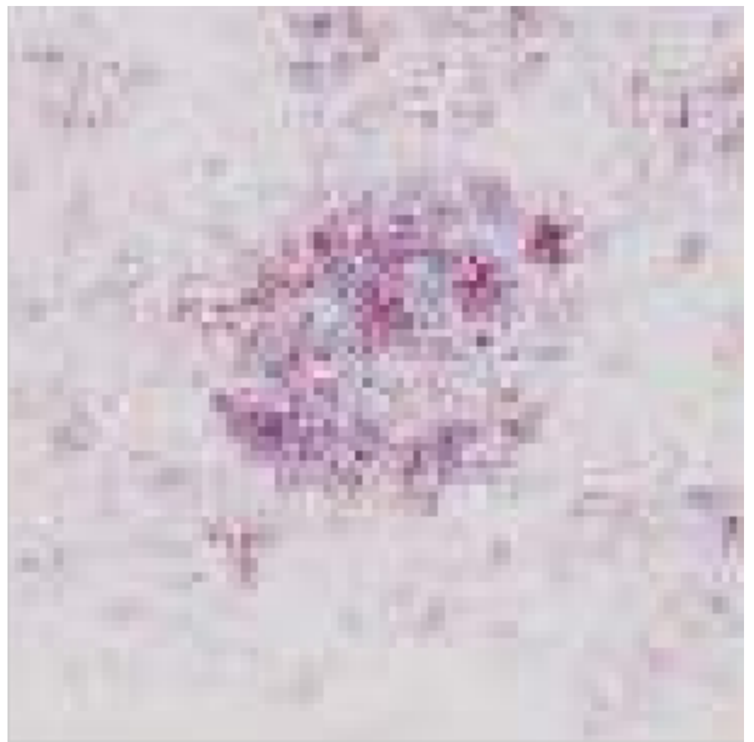
A trained
machine learning model
(e.g., neural network)

prediction

$p(z)$

Junco Bird-ness

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



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

Sanity check question

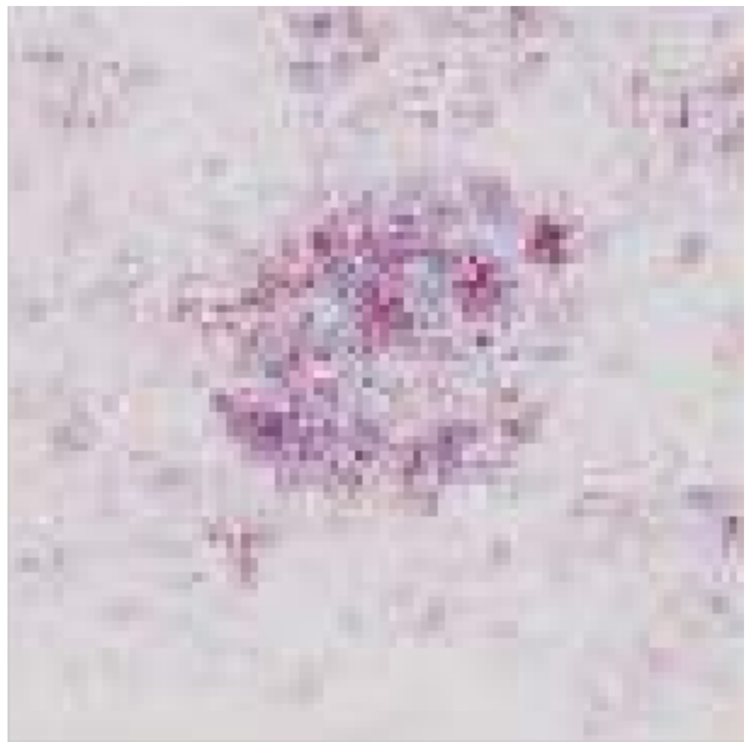
Input image

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

prediction

$p(z)$

Junco Bird-ness



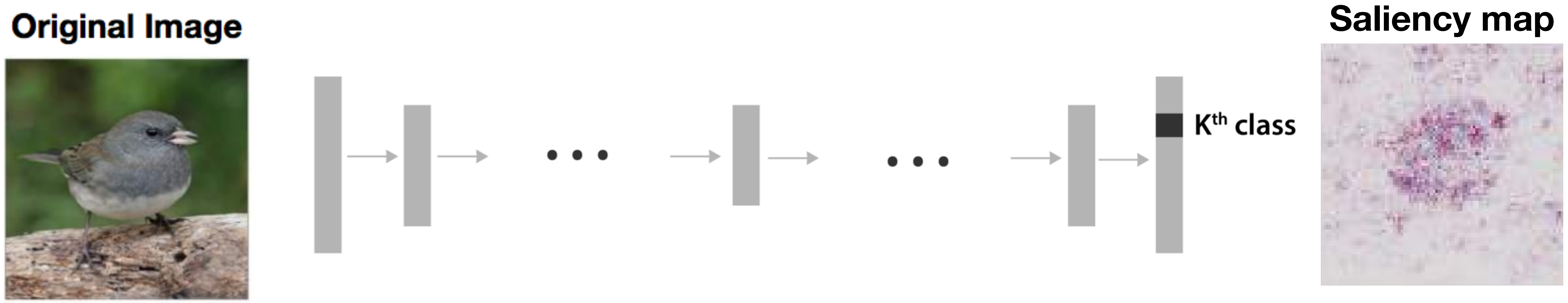
So these pixels are the **evidence** of **prediction.**

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

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

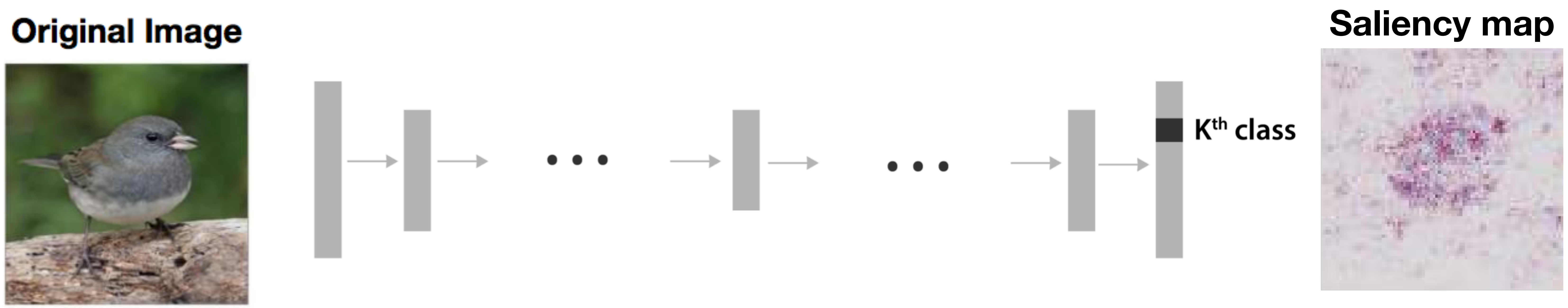
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Some confusing behaviors of saliency maps.

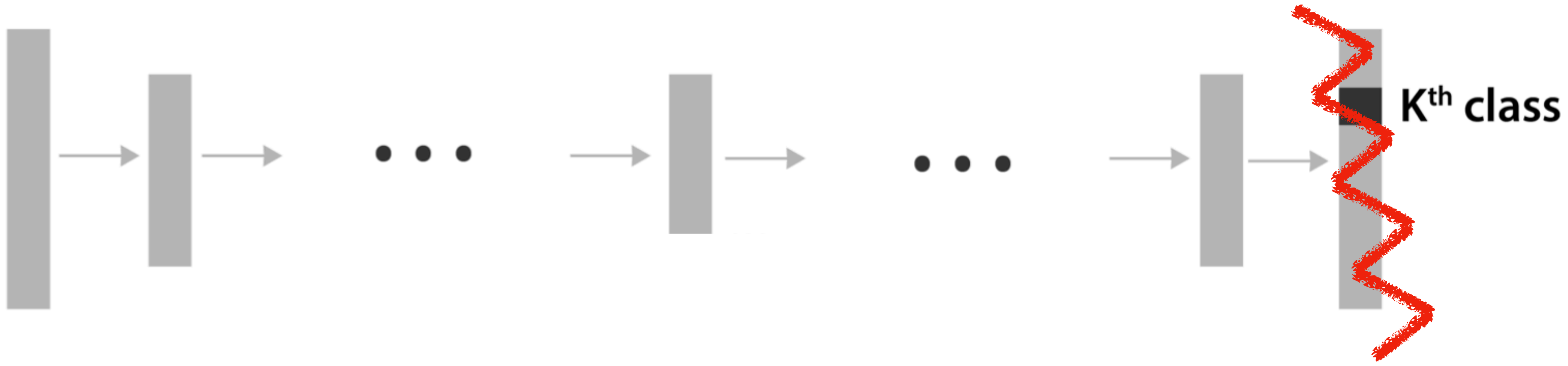


Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NeurIPS 18]

Some confusing behaviors of saliency maps.

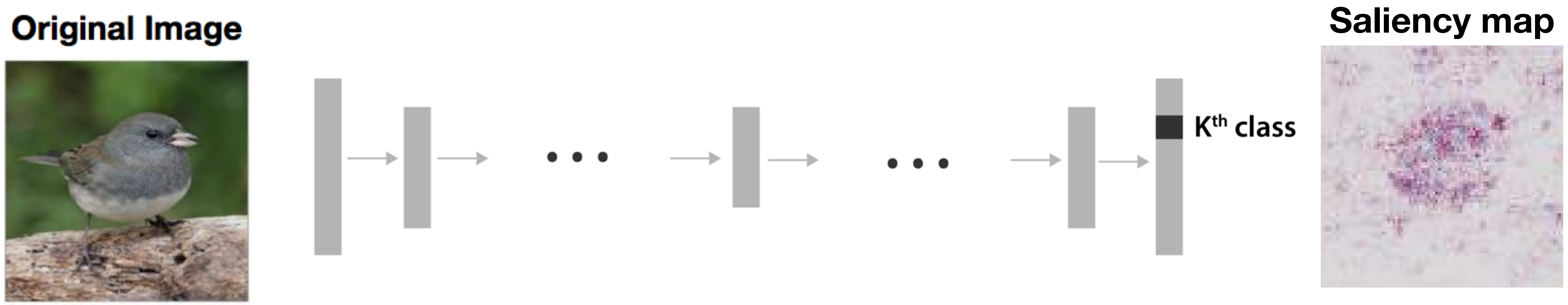


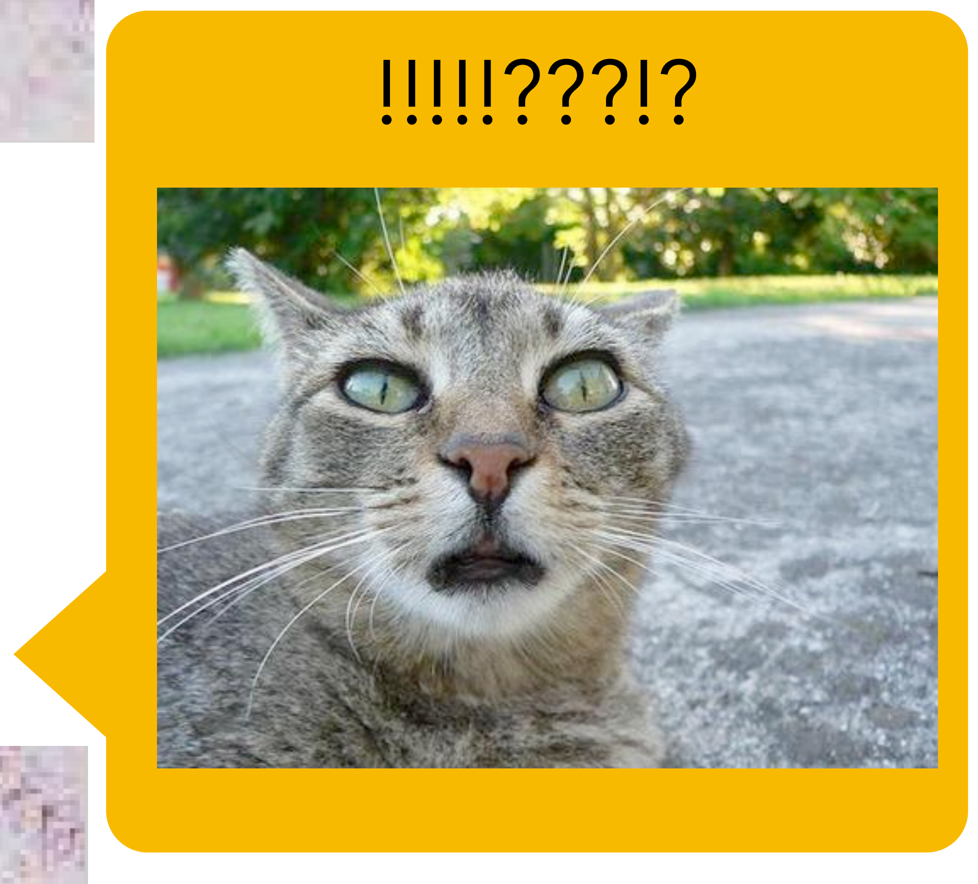
Randomized weights!
Network now makes garbage prediction.



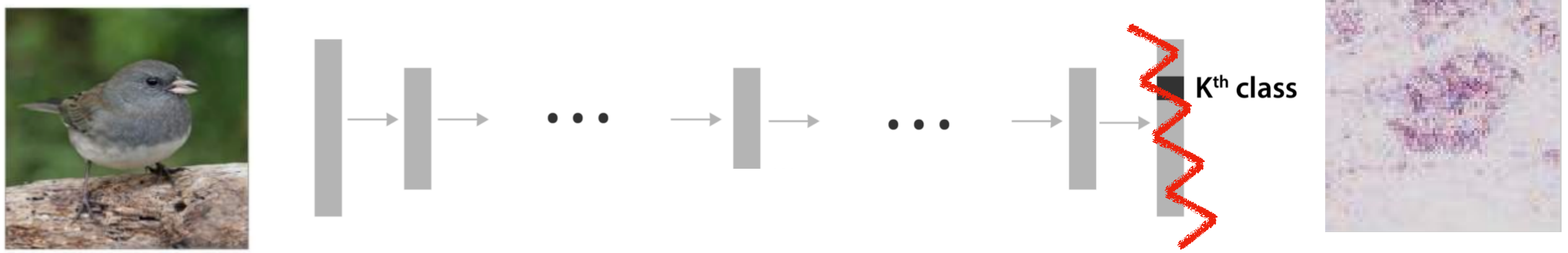
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

Some confusing behaviors of saliency maps.





Randomized weights!
Network now makes garbage prediction.



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

One of the most popular interpretability methods for images:

Saliency maps

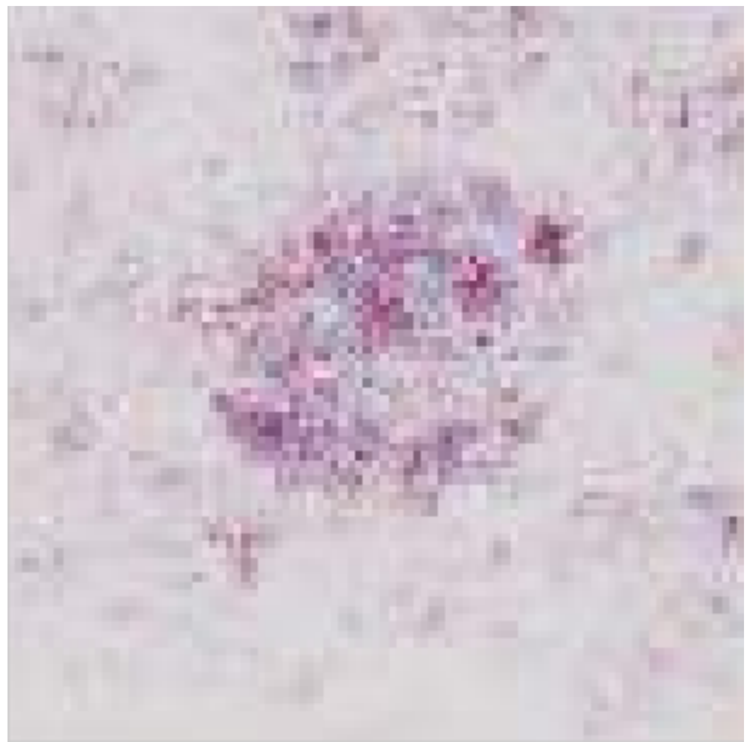
Input image

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

prediction

$p(z)$

Junco Bird-ness



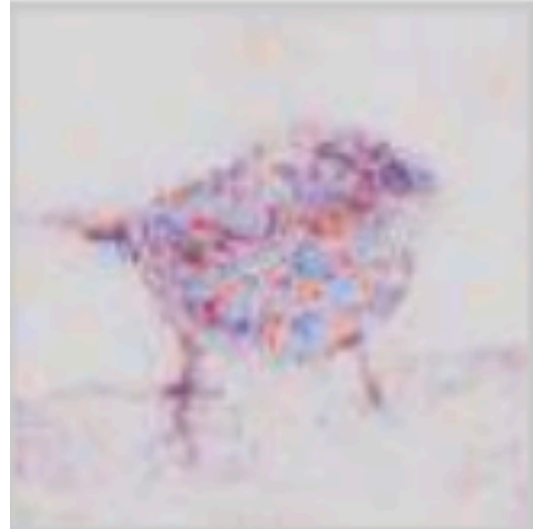
Popular method #1

Popular method #2

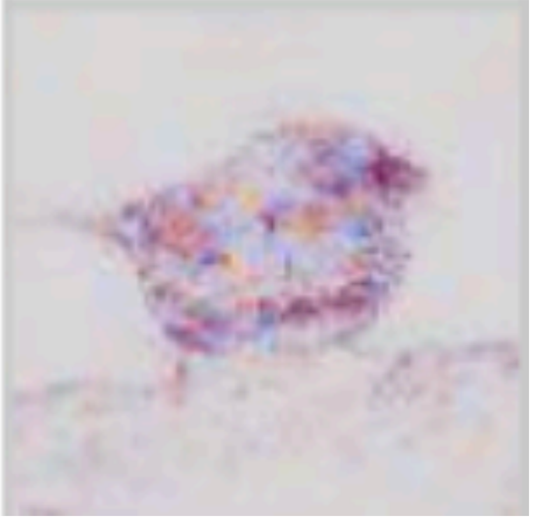
My work from 2018 #1

My work from 2018 #2

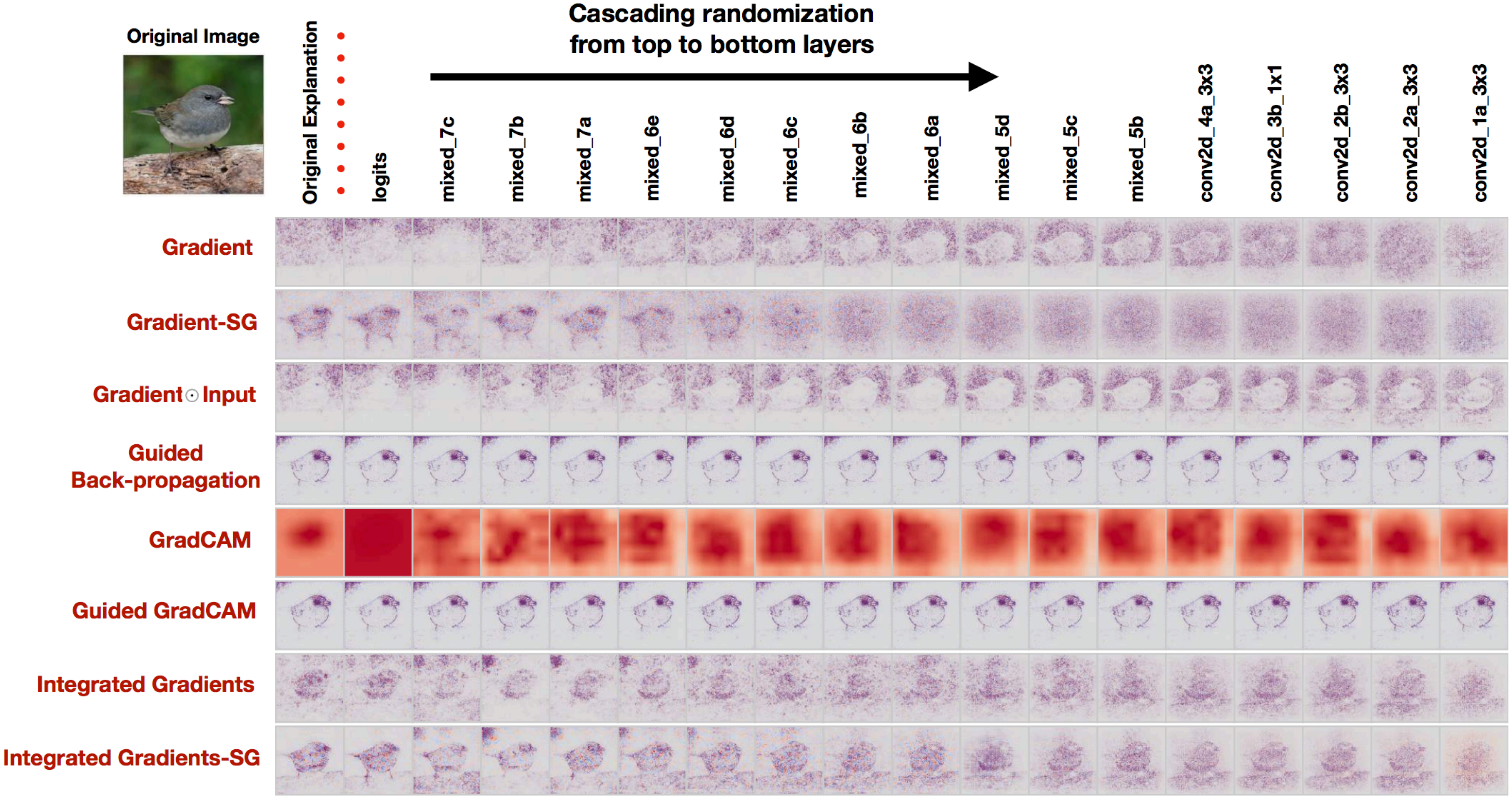
Popular method #3



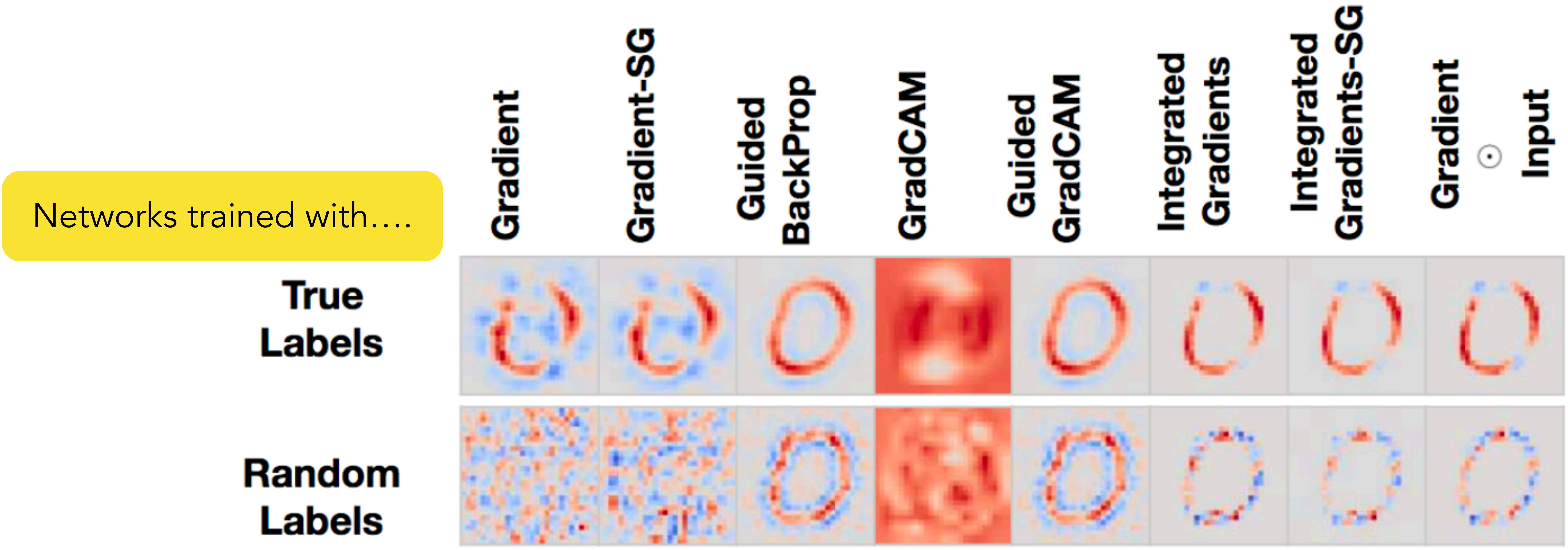
Popular method #4



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



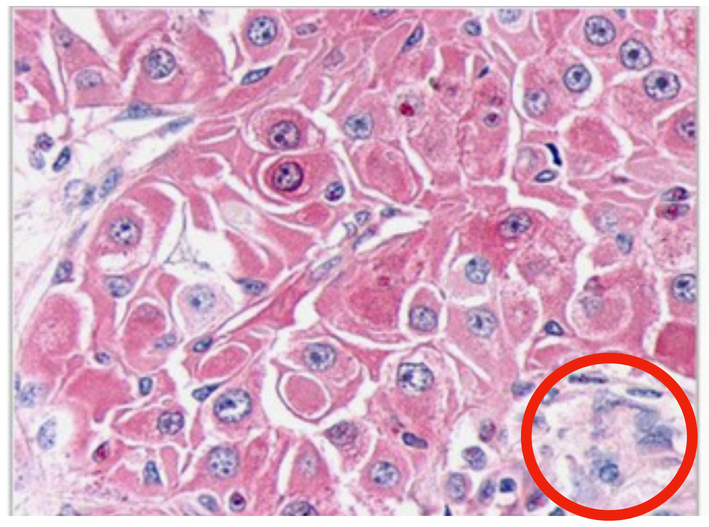
Sanity check2:
Networks trained with random labels,
Do explanations deliver different messages?
No!

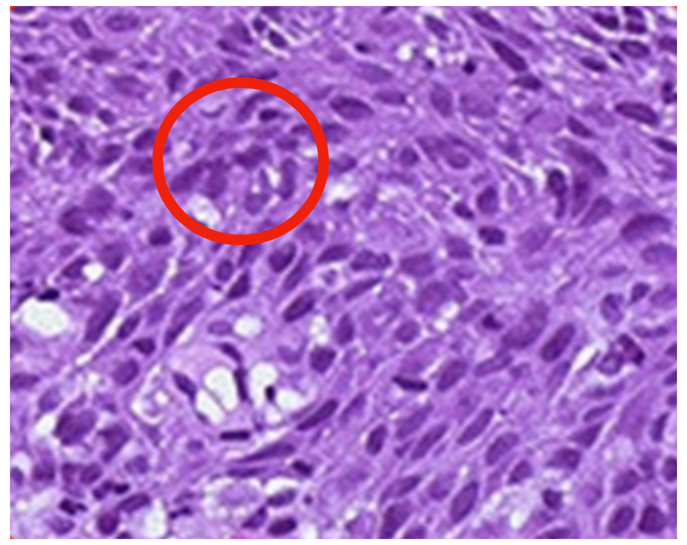


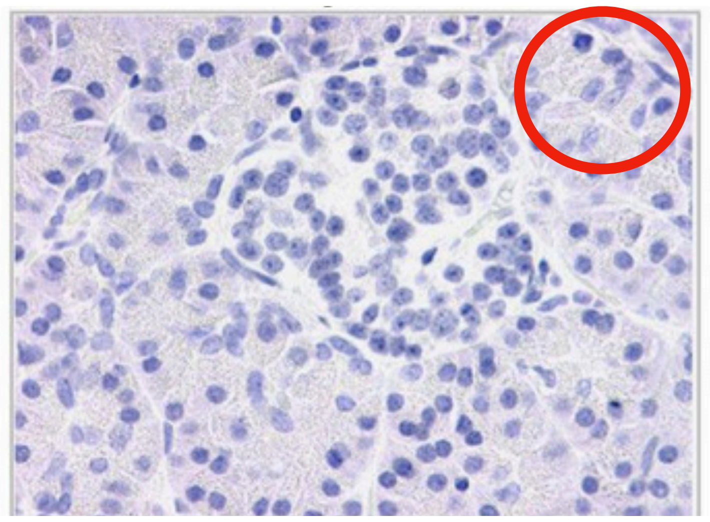
Sanity Checks for Saliency Maps
Joint work with Adebayo, Gilmer, Goodfellow, Hardt, [NIPS 18]

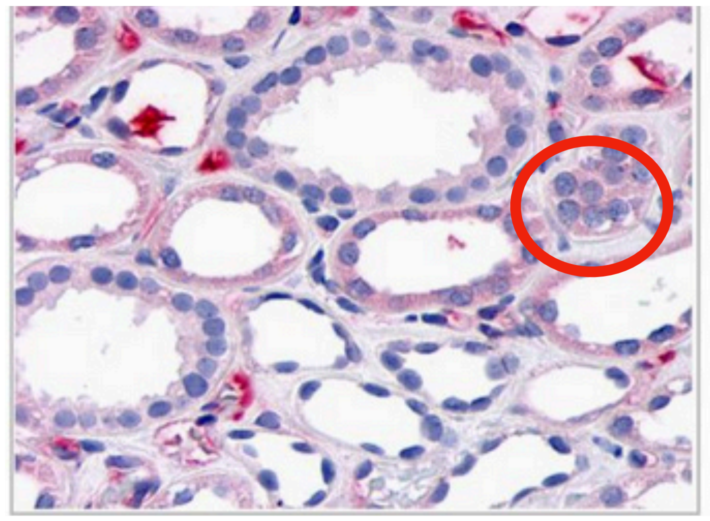
Wait, what's so bad about this?

- What's this obsession about prediction? Maybe it's showing "features" that could have been 'used' in prediction. That's still relevant.


Your kidney

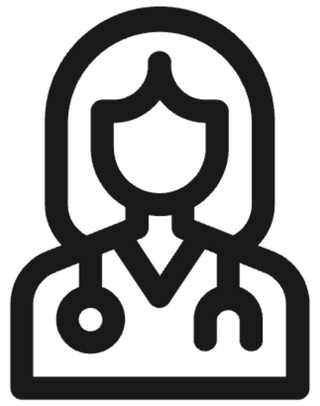

Your lung


Your pancreas


Your colon

Explanations:
"Dotty" feature **used** to classify cancer.

Oh it's all cancer.

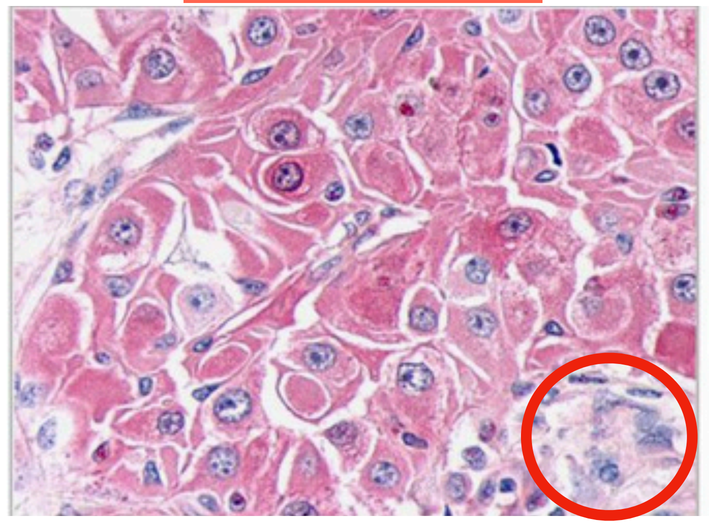


www.biomedcentral.com

Wait, what's so bad about this?

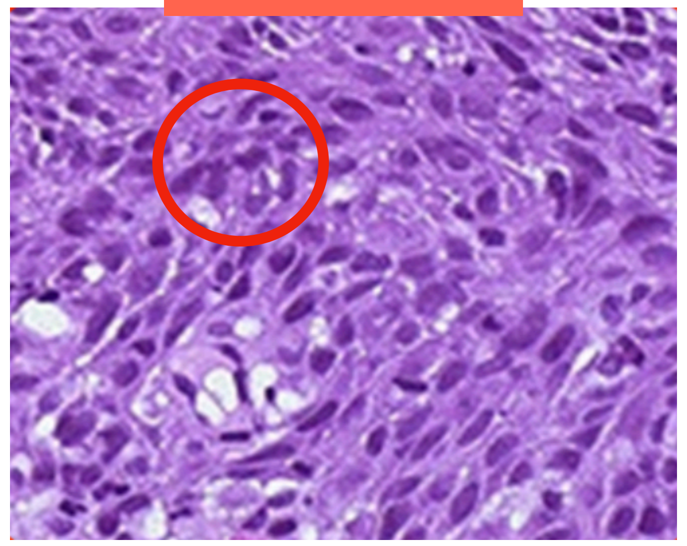
- What's this obsession about prediction? Maybe it's showing "features" that could have been 'used' in prediction. That's still relevant.

Cancer



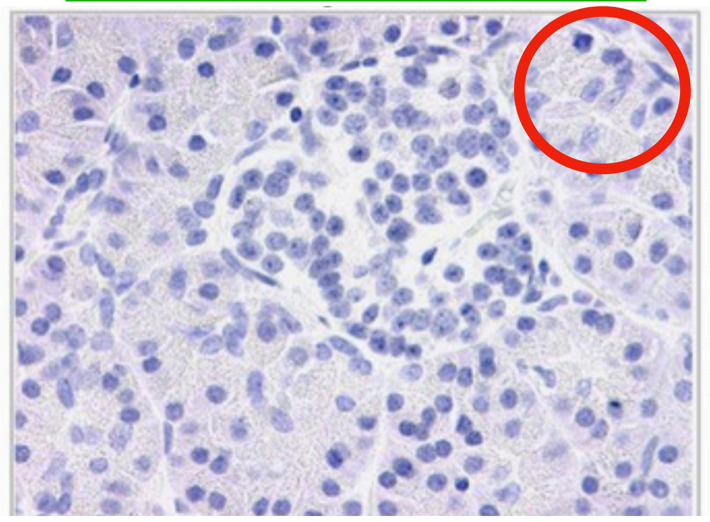
Your kidney

Cancer



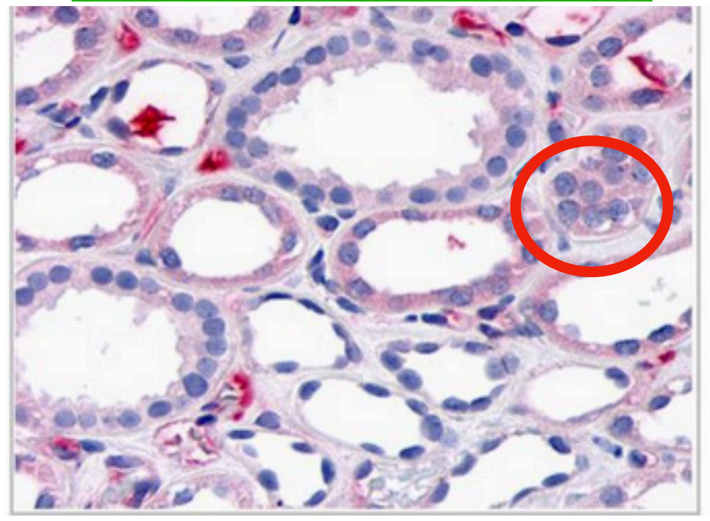
Your lung

Not cancer



Your pancreas

Not cancer



Your colon

Explanations:
"Dotty" feature **used** to classify cancer.

www.biomedcentral.com

Many skeptics followed! But still long way to go.

Statistically Consistent Saliency Estimation

Emre Barut, Shunyan Luo

25 Sep 2019 (modified: 23 Dec 2019) ICLR 2020 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisio

TL;DR: We propose a statistical framework and a theoretically consistent procedure for saliency estimation.

Abstract: The use of deep learning for a wide range of data problems has increased the need for understanding and diagnosing these m become an essential tool for data analysts. Although numerous model interpretation methods have been proposed in recent years, most theoretical guarantees. In this work, we propose a statistical framework for saliency estimation for black box computer vision models. We consistent and passes the saliency checks of Adebayo et al. (2018). Our method requires solving a linear program, whose solution can be analysis, we establish an upper bound on the number of model evaluations needed to recover the region of importance with high probab gradients that is shown to be more efficient than the commonly used random perturbation schemes. Validity of the new method is demo

5 EXAMPLES

In this section, we demonstrate the robustness and validity of our procedure by two numerical experiments. In Section 5.1, we perform sanity checks as laid out by Adebayo et al. (2018b), and show that the LEG-TV estimator fails to detect objects when the weights of the neural network are chosen

Official Blind Review #1

ICLR 2020 Conference Paper2432 AnonReviewer1

23 Oct 2019 (modified: 22 Nov 2019) ICLR 2020 Conference Paper2432 Official Review Readers: Everyone

In Adebayo et al., NIPS'18 (and very related efforts), there are presented a set of sanity checks to be applied to explanation methods to ensure their predictions are relate to the class and model being predicted. Could you provide any indication on whether the proposed method passes these checks?

Some issues

ICLR 2019 Conference Paper294 AnonReviewer1

05 Nov 2018 (modified: 30 Nov 2018) ICLR 2019 Conference Paper294 Official Review Readers: Everyone Show Revisions

3. Recently several papers pointed out some significant issues in Guided BP,

Xie et al. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. ICML 2018
Adebayo et al. Sanity Checks for Saliency Maps. NIPS 2018
Kindermans et al. The (Un)reliability of saliency methods. NIPS workshop 2017

Official Blind Review #2

ICLR 2020 Conference Paper709 AnonReviewer2

The authors referred to Sanity Checks for Saliency Maps (Adebayo et al) without using it for their results, it would be nice to add it to the experiments.

[Submitted on 27 May 2019 (v1), last revised 7 Jun 2019 (this version, v2)]

A Simple Saliency Method That Passes the Sanity Checks

Arushi Gupta, Sanjeev Arora

There is great interest in "saliency methods" (also called "attribution methods"), which give "explanations" for a deep credit-assignment via the gradient of the output with respect to input. Recently Adebayo et al. [arXiv:1810.03292] qu whether the scores shift/vanish when layers of the trained net are randomized, or when the net is retrained using ran We propose a simple fix to existing saliency methods that helps them pass sanity checks, which we call "competition using a simple competition among them to identify and remove less relevant pixels from the map. The simplest varia only the input and gradient. Some theoretical justification is provided for it (especially for ReLU networks) and its per

[Submitted on 29 Nov 2019]

Sanity Checks for Saliency Metrics

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, Alun Preece

Saliency maps are a popular approach to creating post-hoc explanations of image classifier outputs. These methods produce estimates of the relevance of each pixel to the classific as a saliency map that highlights important pixels. Despite a proliferation of such methods, little effort has been made to quantify how good these saliency maps are at capturing th classifier output (i.e. their "fidelity"). We therefore investigate existing metrics for evaluating the fidelity of saliency methods (i.e. saliency metrics). We find that there is little consiste are calculated, and show that such inconsistencies can have a significant effect on the measured fidelity. Further, we apply measures of reliability developed in the psychometric test

[Submitted on 4 Oct 2019 (v1), last revised 5 Dec 2019 (this version, v3)]

Can I Trust the Explainer? Verifying Post-hoc Explanatory Methods

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, Phil Blunsom

For AI systems to garner widespread public acceptance, we must develop methods capable of explaining the decisions of black-box models such as neural networks. In methods. First, we show that two prevalent perspectives on explanations --- feature-additivity and feature-selection --- lead to fundamentally different instance-wise perspectives are currently being directly compared, despite their distinct explanation goals. The second issue is that current post-hoc explainers are either validated und linear regression, or on models trained on syntactic datasets), or, when applied to real-world neural networks, explainers are commonly validated under the assumption neural networks often rely on unreasonable correlations, even when producing correct decisions. We introduce a verification framework for explanatory methods under t based on a non-trivial neural network architecture trained on a real-world task, and for which we are able to provide guarantees on its inner workings. We validate the e

[Submitted on 25 Feb 2021]

Do Input Gradients Highlight Discriminative Features?

Harshay Shah, Prateek Jain, Praneeth Netrapalli

Interpretability methods that seek to explain instance-specific model predictions [Simonyan et al. 2014, Smilko the premise that the magnitude of input-gradient -- gradient of the loss with respect to input -- highlights dis relevant for prediction over non-discriminative features that are irrelevant for prediction. In this work, we intro

[Submitted on 16 Jun 2020 (v1), last revised 3 Mar 2021 (this version, v2)]

Rethinking the Role of Gradient-Based Attribution Methods for Model Interpretability

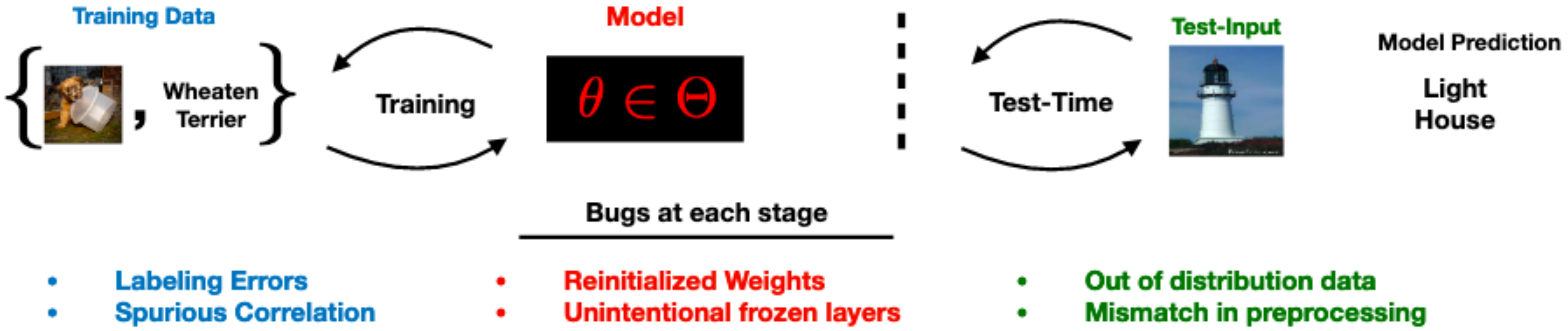
Suraj Srinivas, Francois Fleuret

Current methods for the interpretability of discriminative deep neural networks commonly rely on the model's input-gradients, i.e., the gradients of the output logits w.r.t. the inputs. The common assumption is that these input-gradients contain information regarding $p_\theta(y \mid x)$, the model's discriminative

But how do some of these methods still helpful for some end-tasks?

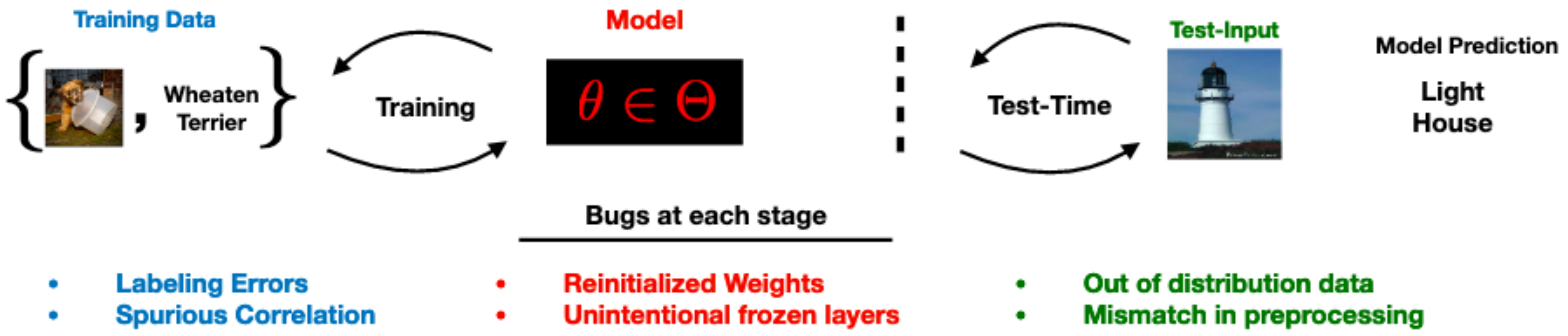
…

What are those tasks?



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

Testing methods with users and concrete end-tasks



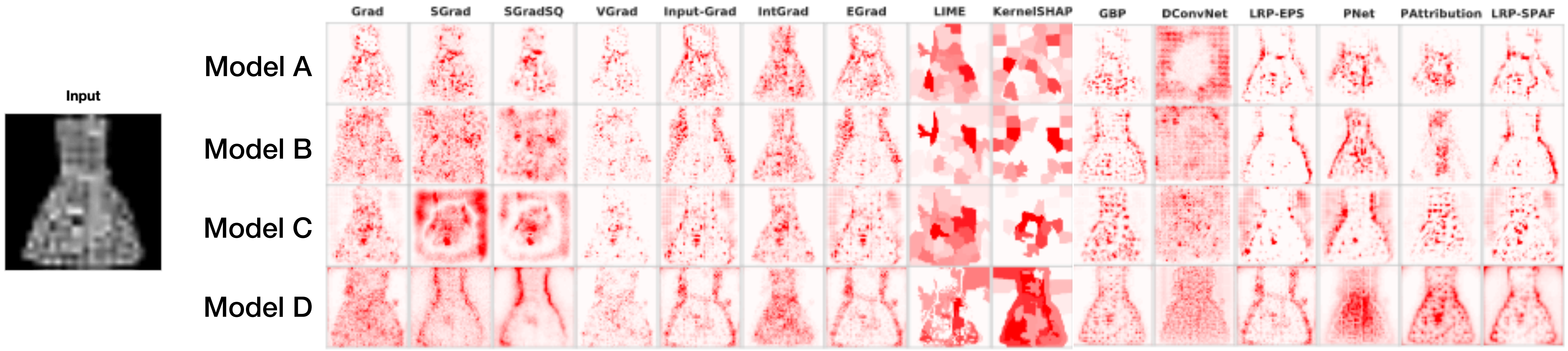
- **Task for subjects**: You work at a start-up selling animal classification ML model. Here are the images, predictions and attribution maps. (We gave users prediction labels as it is unrealistic not to).

- **Questions:** Would you recommend this model? Why? [because the wrong/correct label/explanation]? All in Likert scale.

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

Can these methods tell us about

Out of distribution?

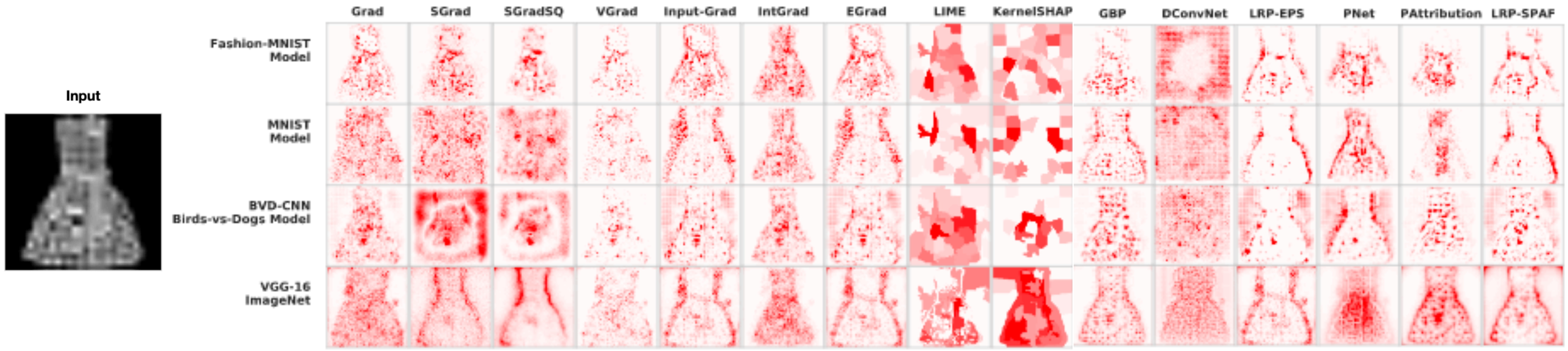


- Out of distribution data

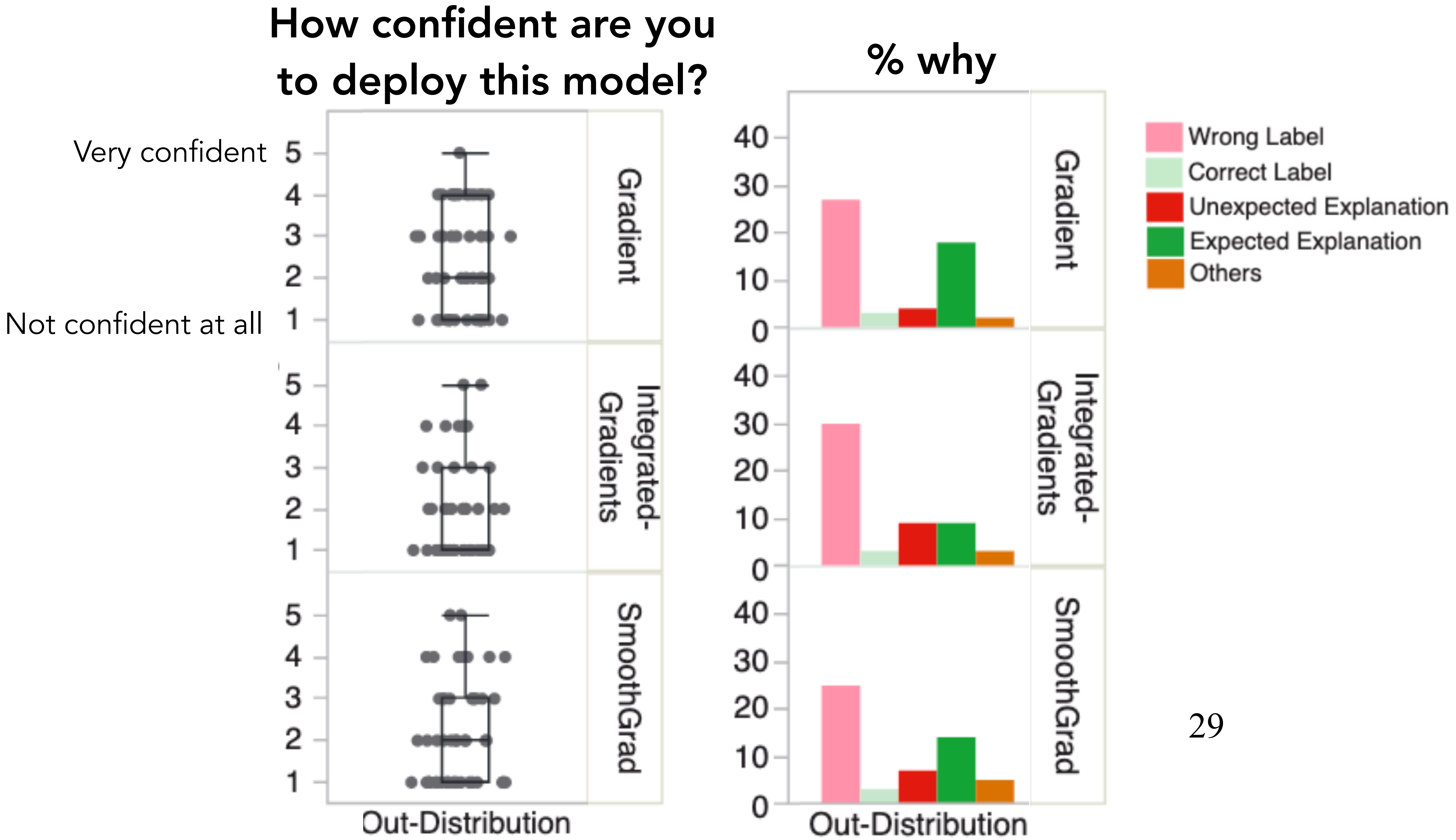
Can these methods tell us about

Out of distribution? probably not.

- Out of distribution data

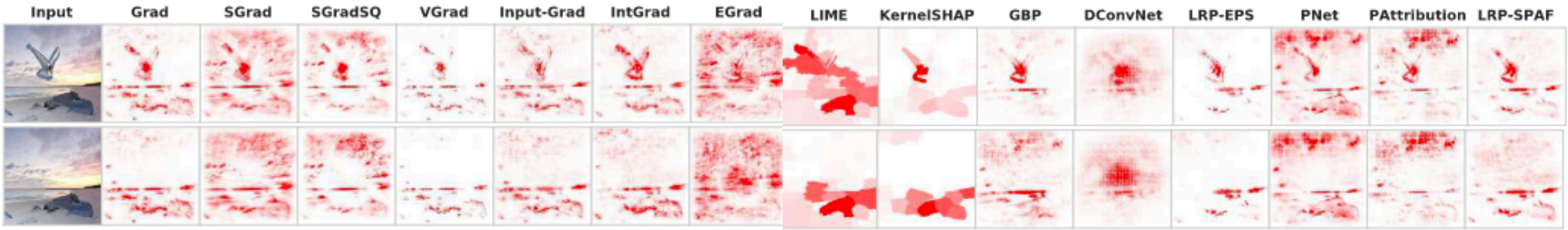


Subjects are uncertain, mostly because of wrong label, but some **expected** explanations.



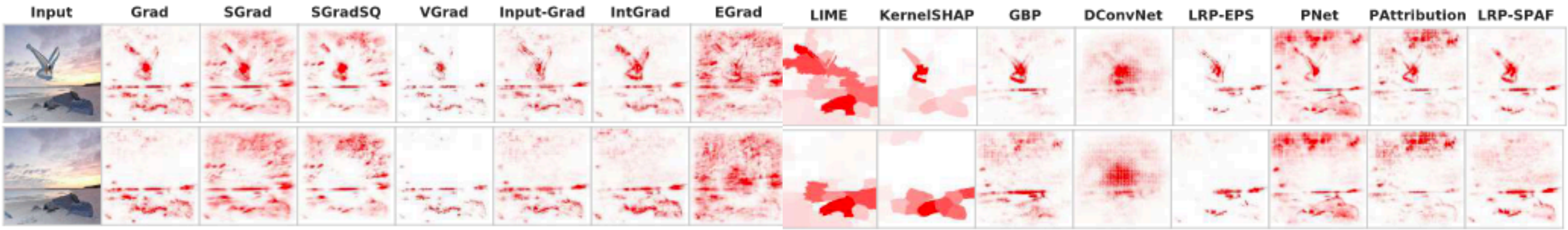
Can these methods tell us about

Spurious correlation?

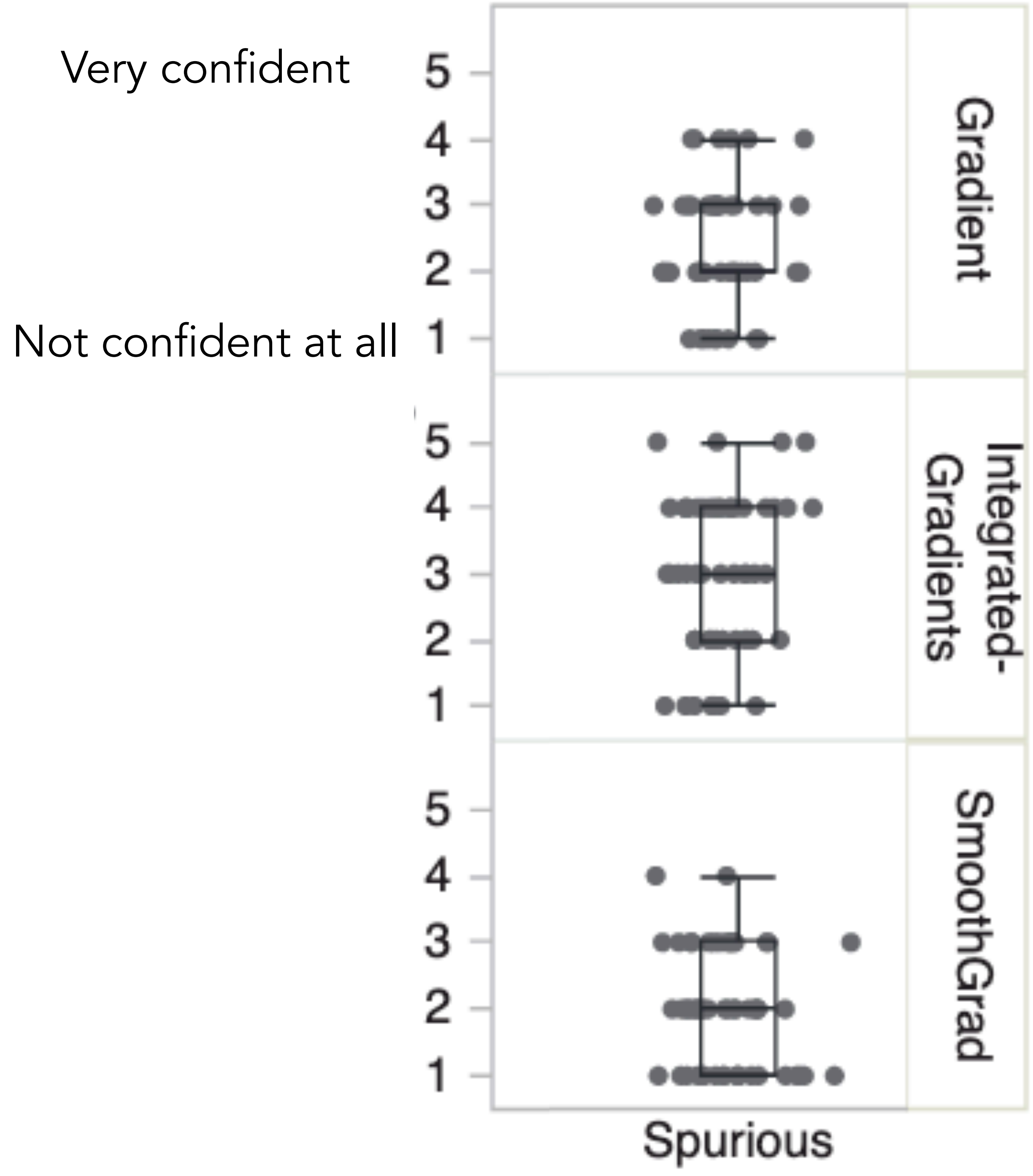


Can these methods tell us about

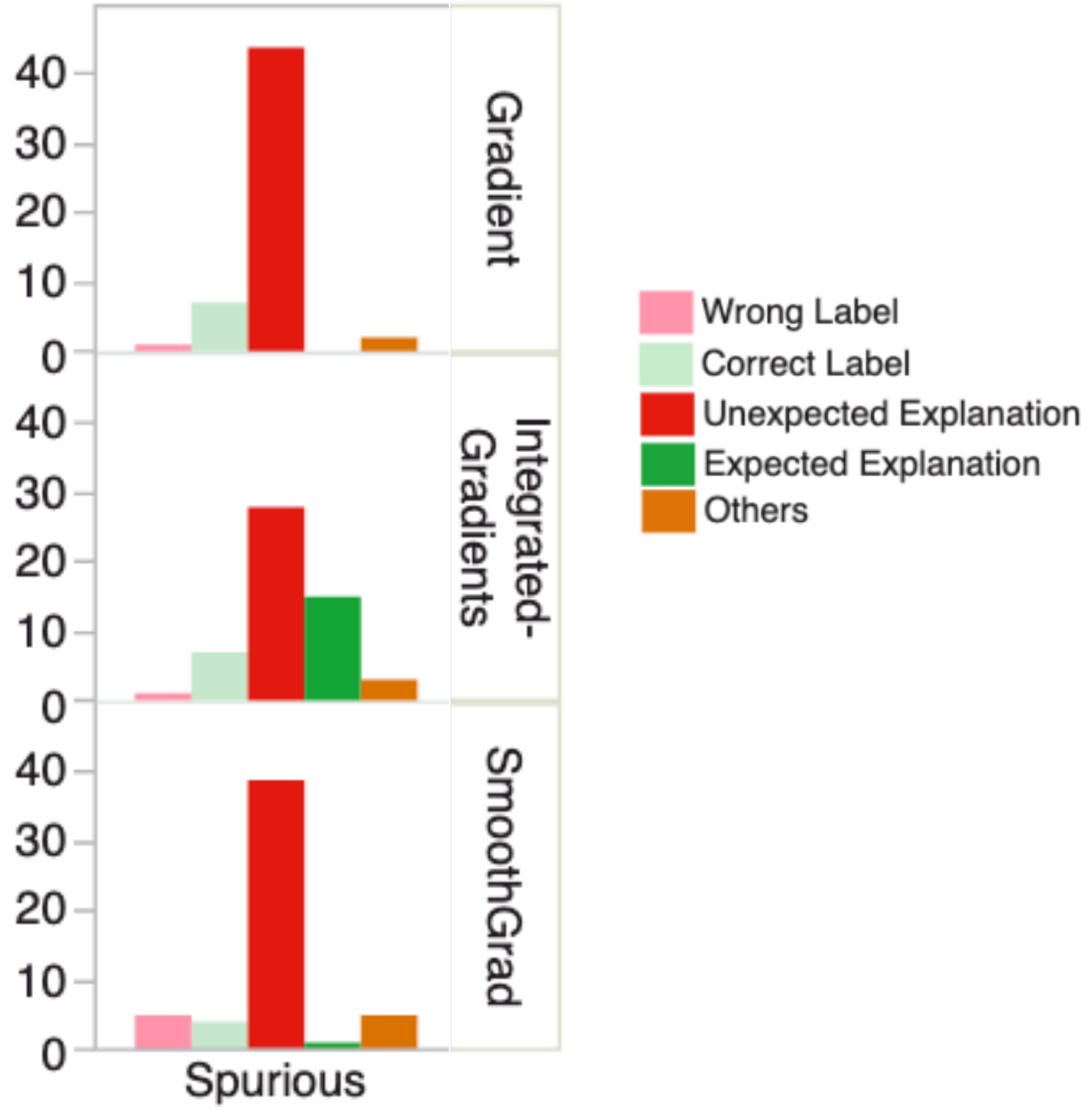
Spurious correlation? maybe!



How confident are you to deploy this model?

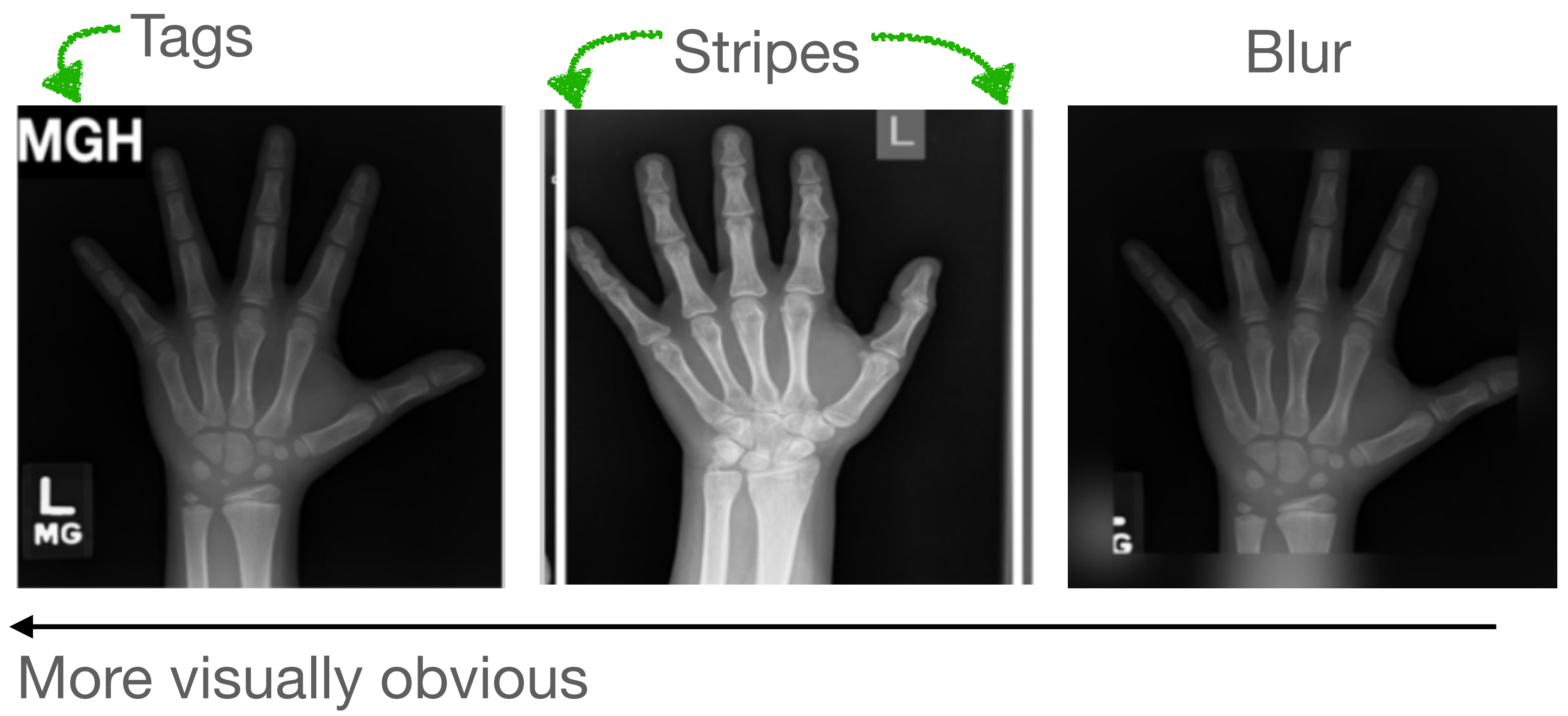


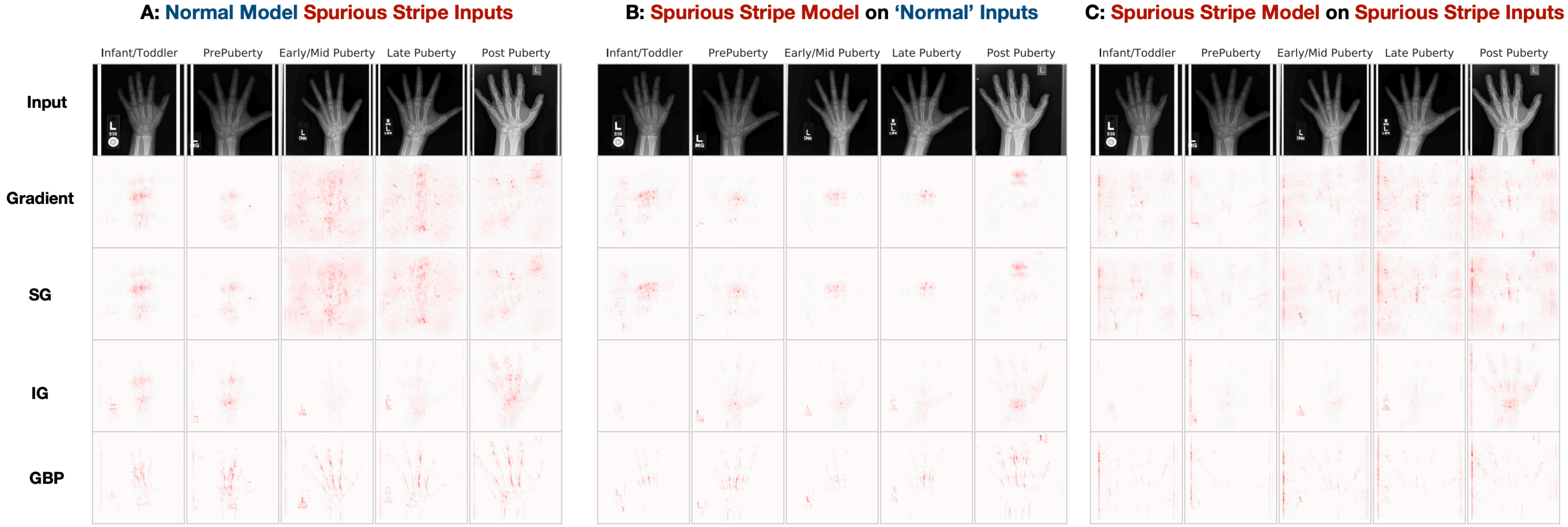
% why



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

[Ongoing work]
What kind of spurious correlation can we hope to capture?
TL;DR: Not many.





[Adebayo, Muelly, K. In submission]

Take away

- Please be skeptical! Think of explanations as your (potentially incompetent) colleague. Maybe they are helpful, but maybe not.

- Explanations are complex in nature (we've known this for quite a few centuries); they are powerful, but we need to be careful how we use them.

- Many explanations can give plausible explanations, but we need to be careful (e.g., even explanations from an inherently interpretable model could be misleading in distributional shift)

- Test, test and test.