Visualizing Model Behavior

Jorge Poco, @jpocom

Fundação Getulio Vargas

FGV EMAp

Material base on:

- Slides from Julius Adebayo (“Sanity Checks for ‘Saliency’ Maps”)
- Slides from Julius Adebayo & Hima Lakkaraju (“Visualizing Model Behavior”)

Introduction

Recent ML Systems achieve superhuman

AlphaGo beats Go human champ

Computer out-plays humans in "doom"

Deep Net outperforms humans in image classification

IMAGENET

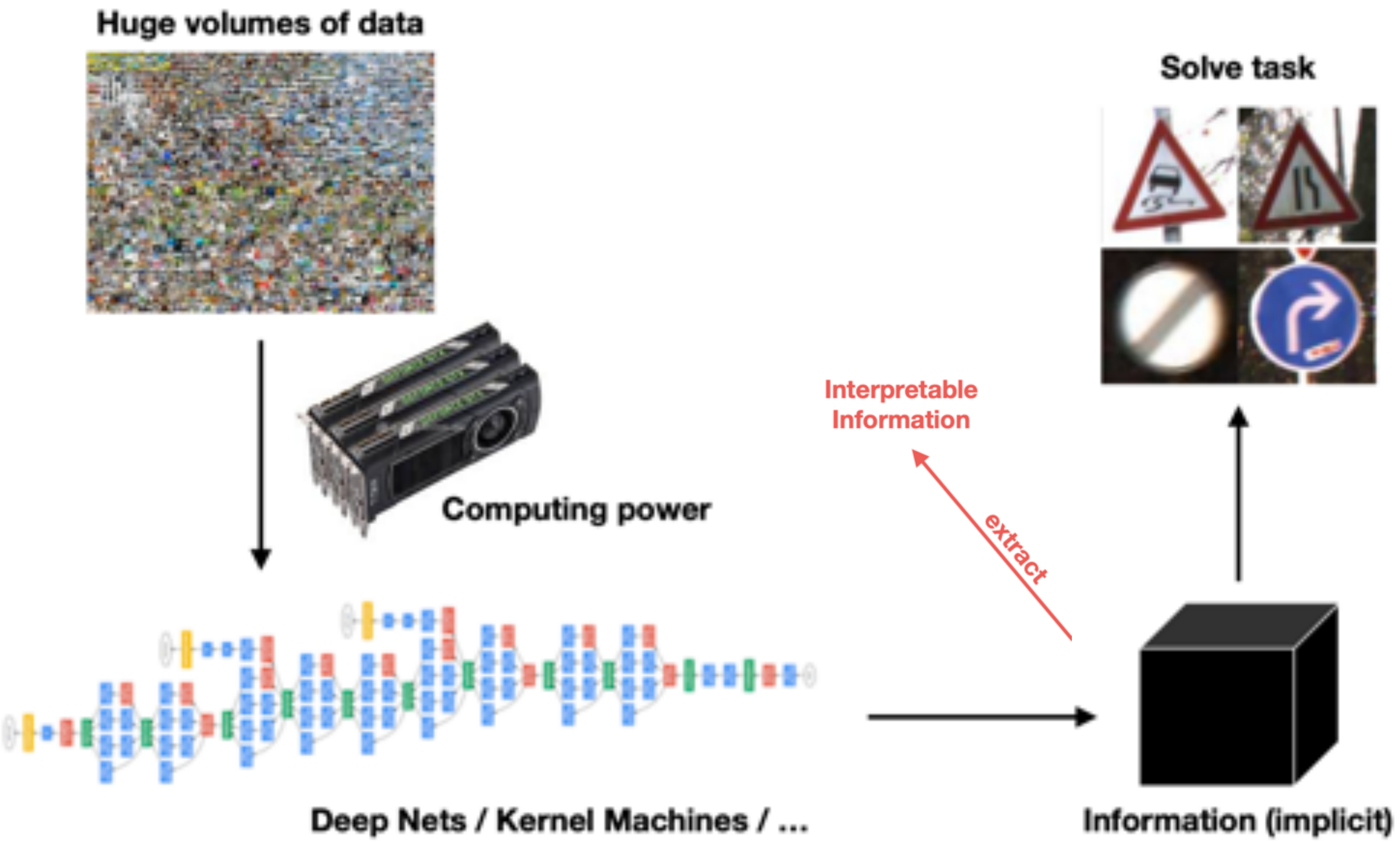
Autonomous search-and-rescue drones outperform humans

IBM's Watson destroys humans in jeopardy

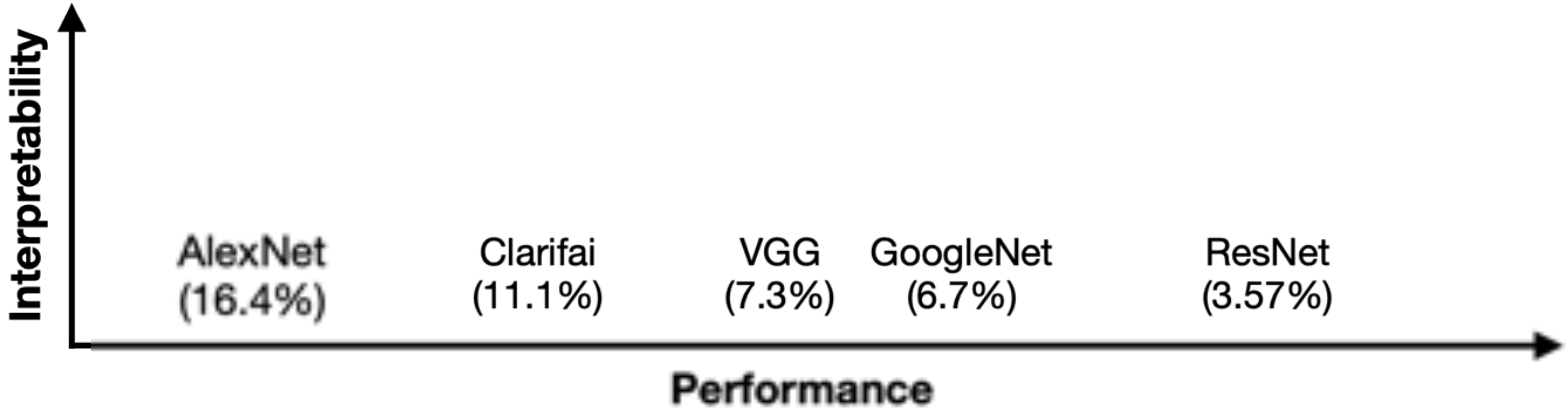
DeepStack beats professional poker players

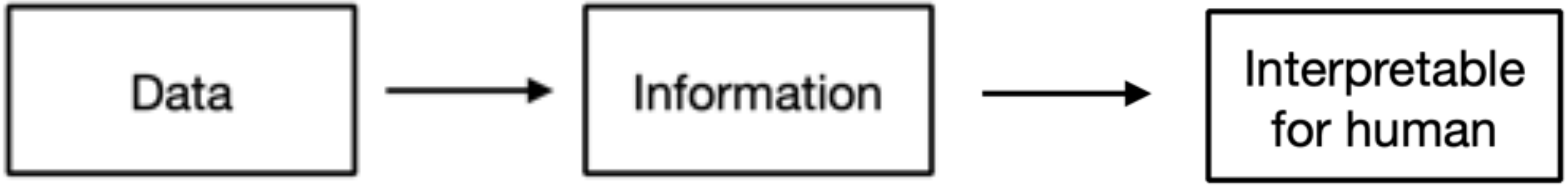
Deep Net beats human at recognizing traffic signs

From Data to Information



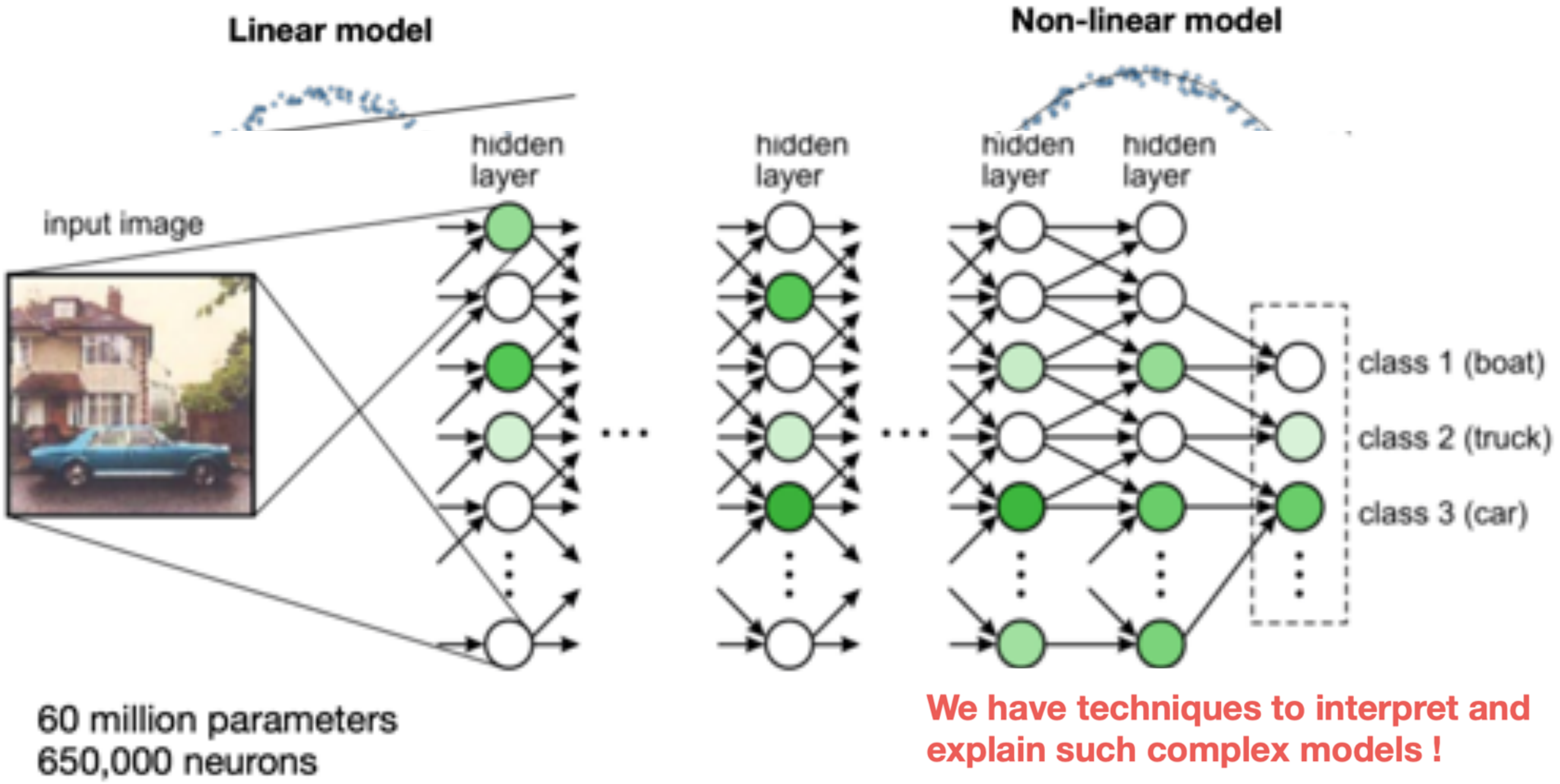
From Data to Information





Crucial in many applications

Interpretable vs. Powerful Models ?



We have techniques to interpret and explain such complex models !

Interpretable vs. Powerful Models ?

Ante-hoc interpretability:

Choose a model that is readily interpretable and train it.

Example:

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} \overbrace{g_i(x_i)}^{\text{contribution of } i\text{th variable}}$$

Is the model expressive enough to predict the data?

Post-hoc interpretability:

Choose a model that works well in practice, and develop a special technique to interpret it.

Example:

$$f(\boldsymbol{x}) = \mathrm{DeepNet}(\boldsymbol{x})$$

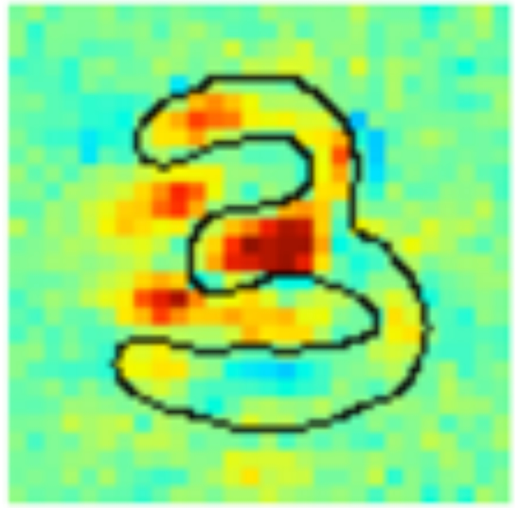
How to determine the contribution each input variable?

Dimensions of Interpretability

Different dimensions of "interpretability"

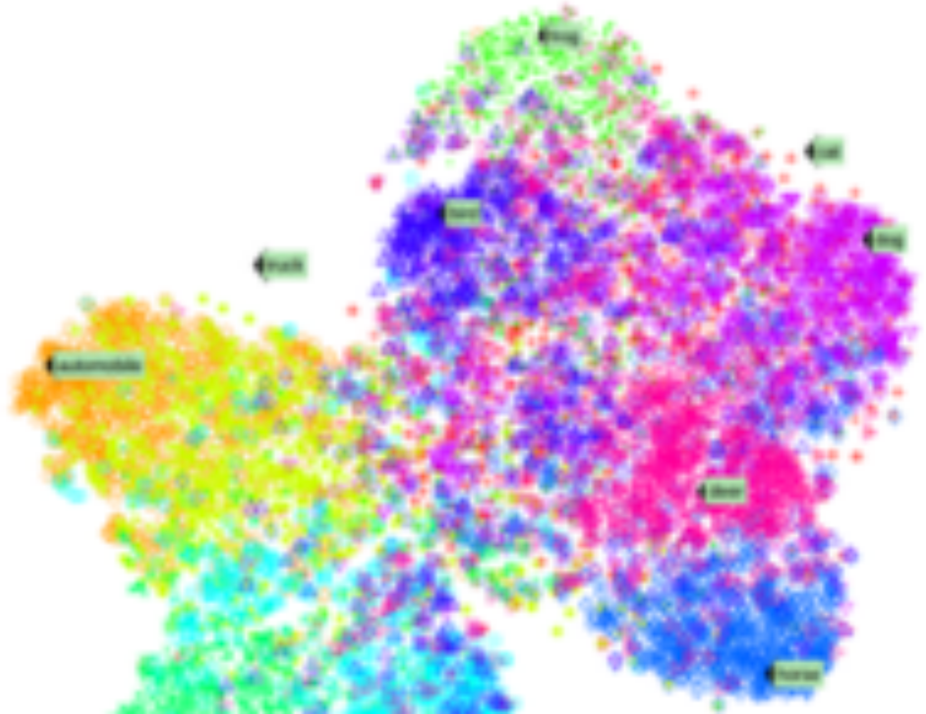
prediction

"Explain why a certain pattern x has been classified in a certain way f(x)."



model

"What would a pattern belonging to a certain category typically look like according to the model."



data

"Which dimensions of the data are most relevant for the task."

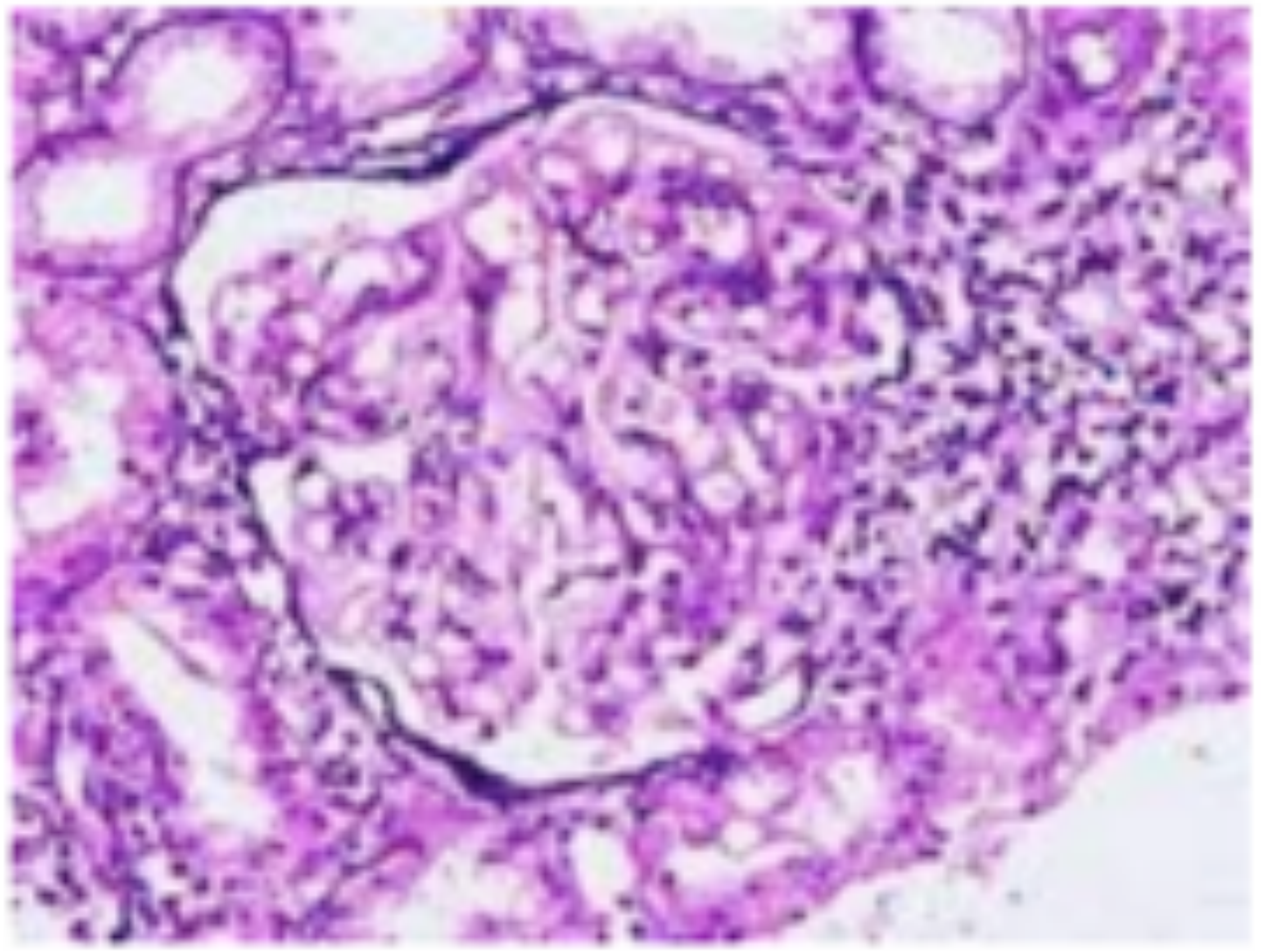
Why Interpretability ?

1) Verify that classifier works as expected

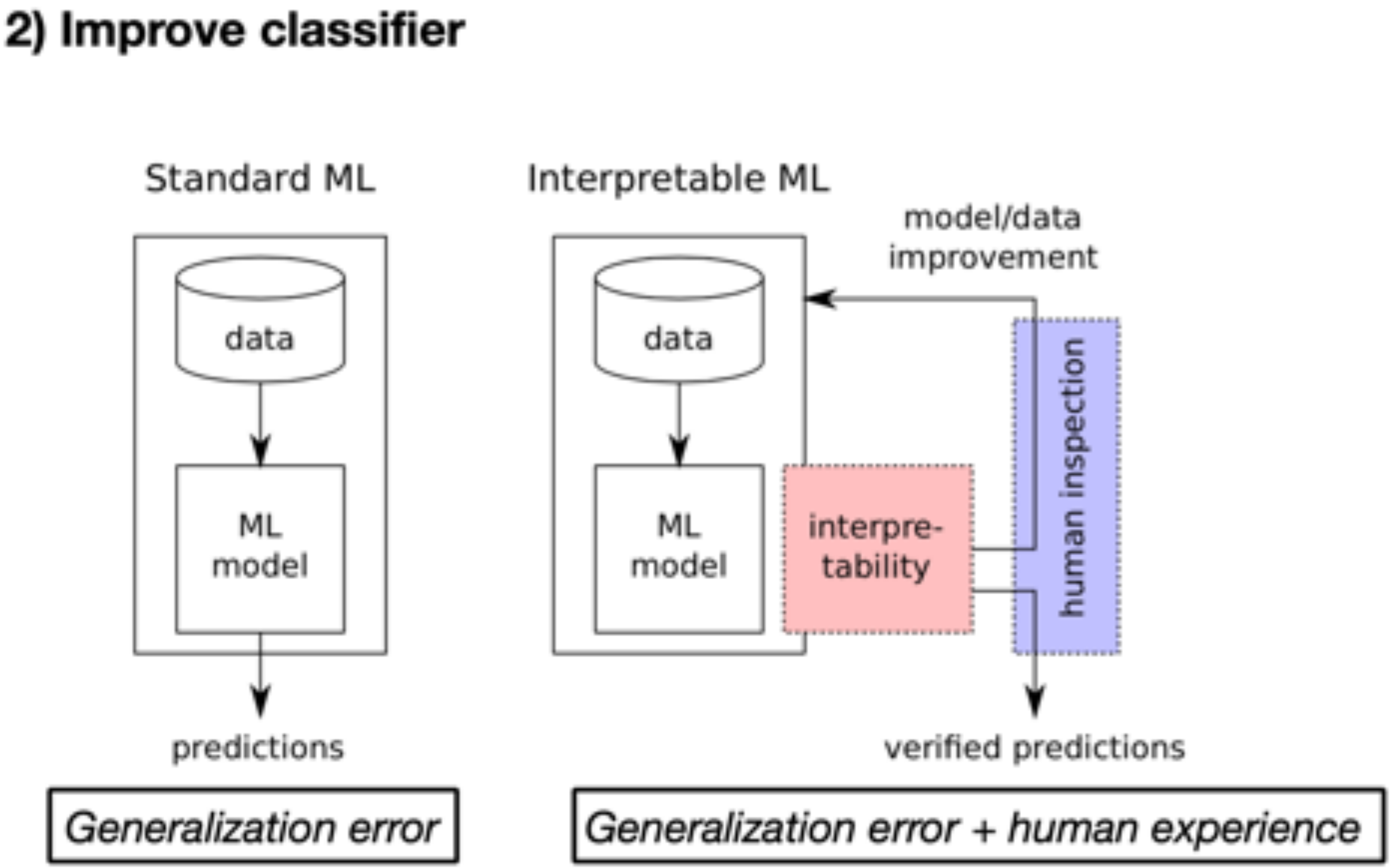
Wrong decisions can be costly
and dangerous

*"Autonomous car crashes,
because it wrongly recognizes …"*

*"AI medical diagnosis system
misclassifies patient's disease …"*



Why Interpretability ?

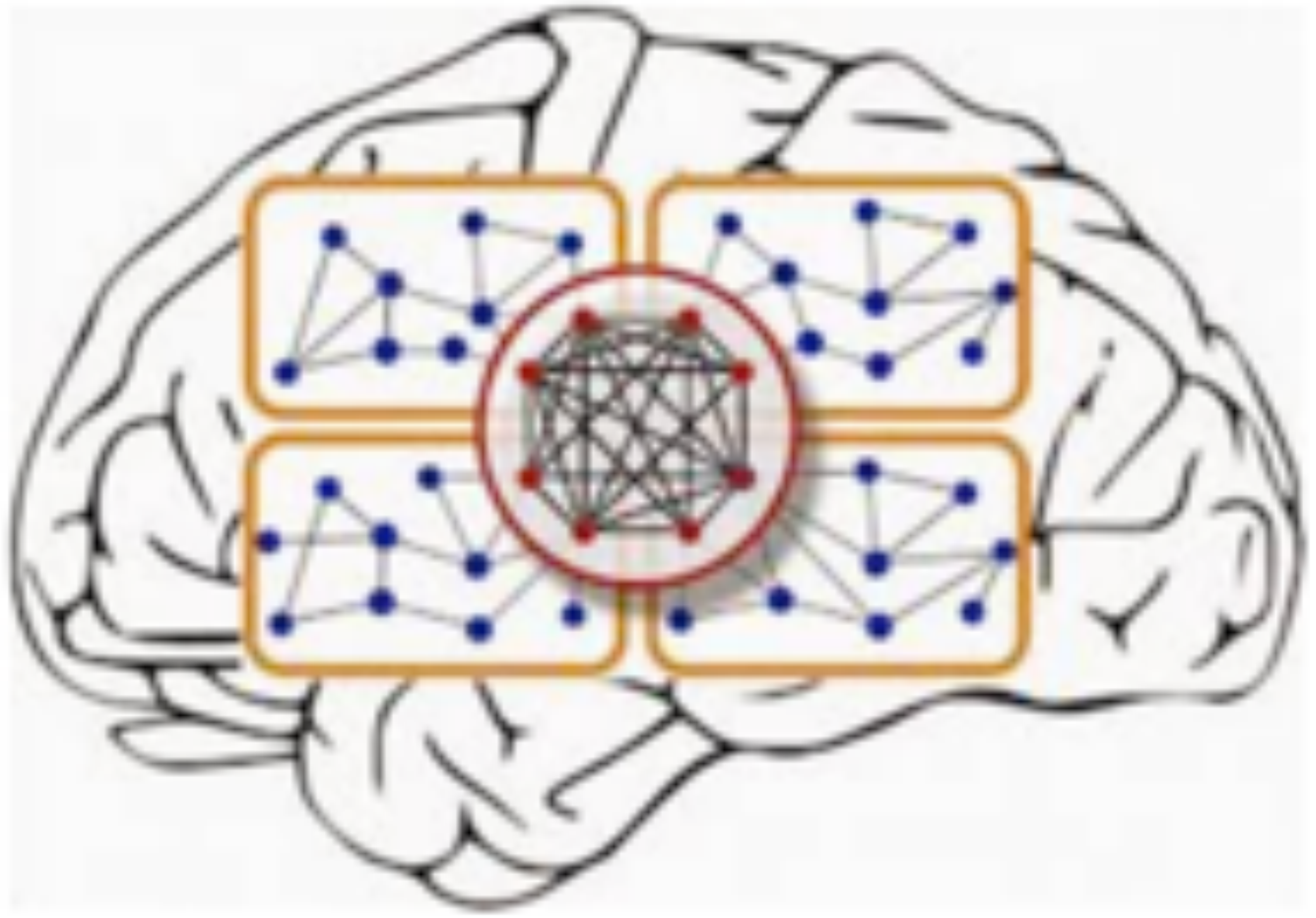


Why Interpretability ?

3) Learn from the learning machine

"It's not a human move. I've never seen a human play this move." (Fan Hui)

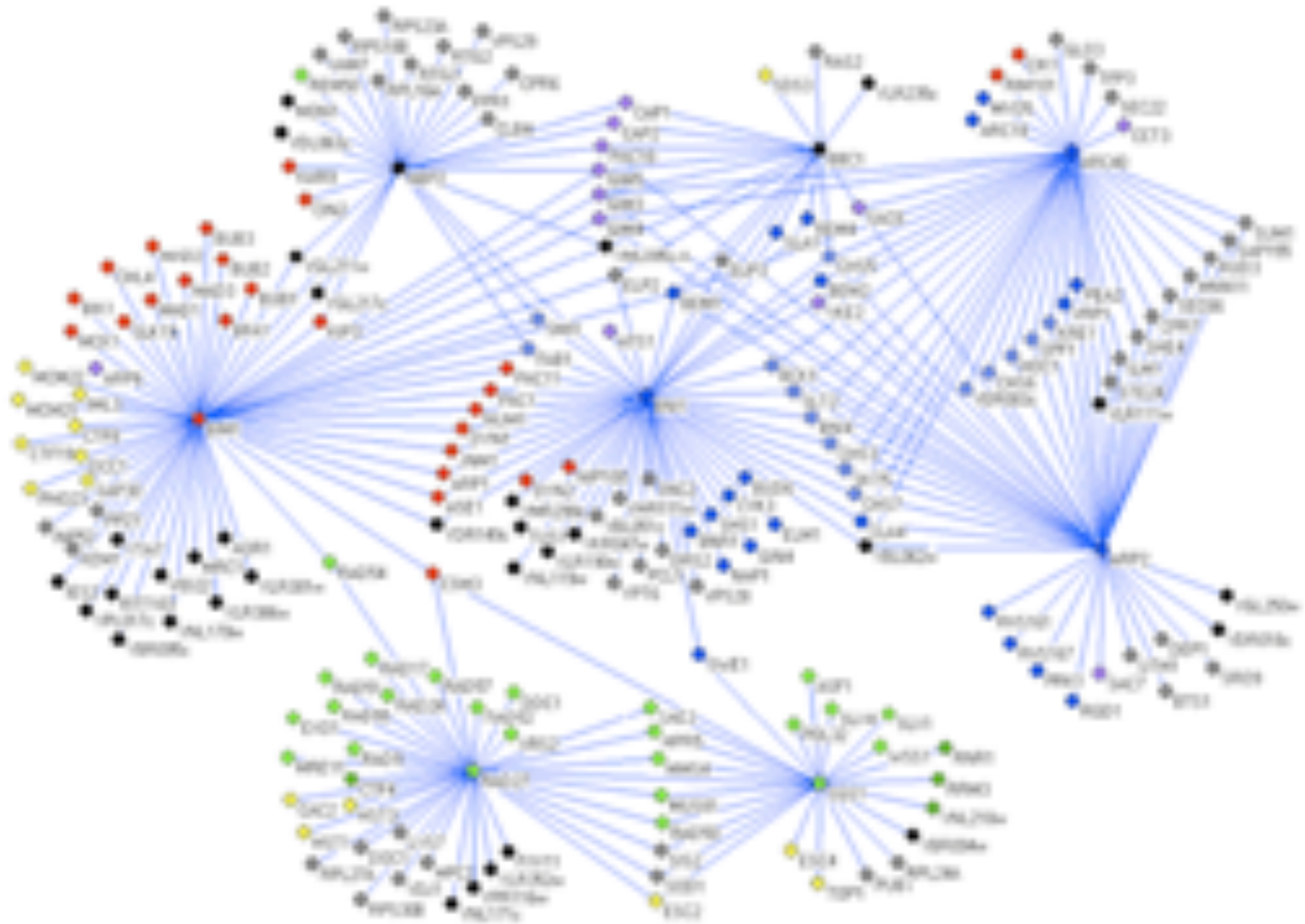
Old promise:
"Learn about the human brain."

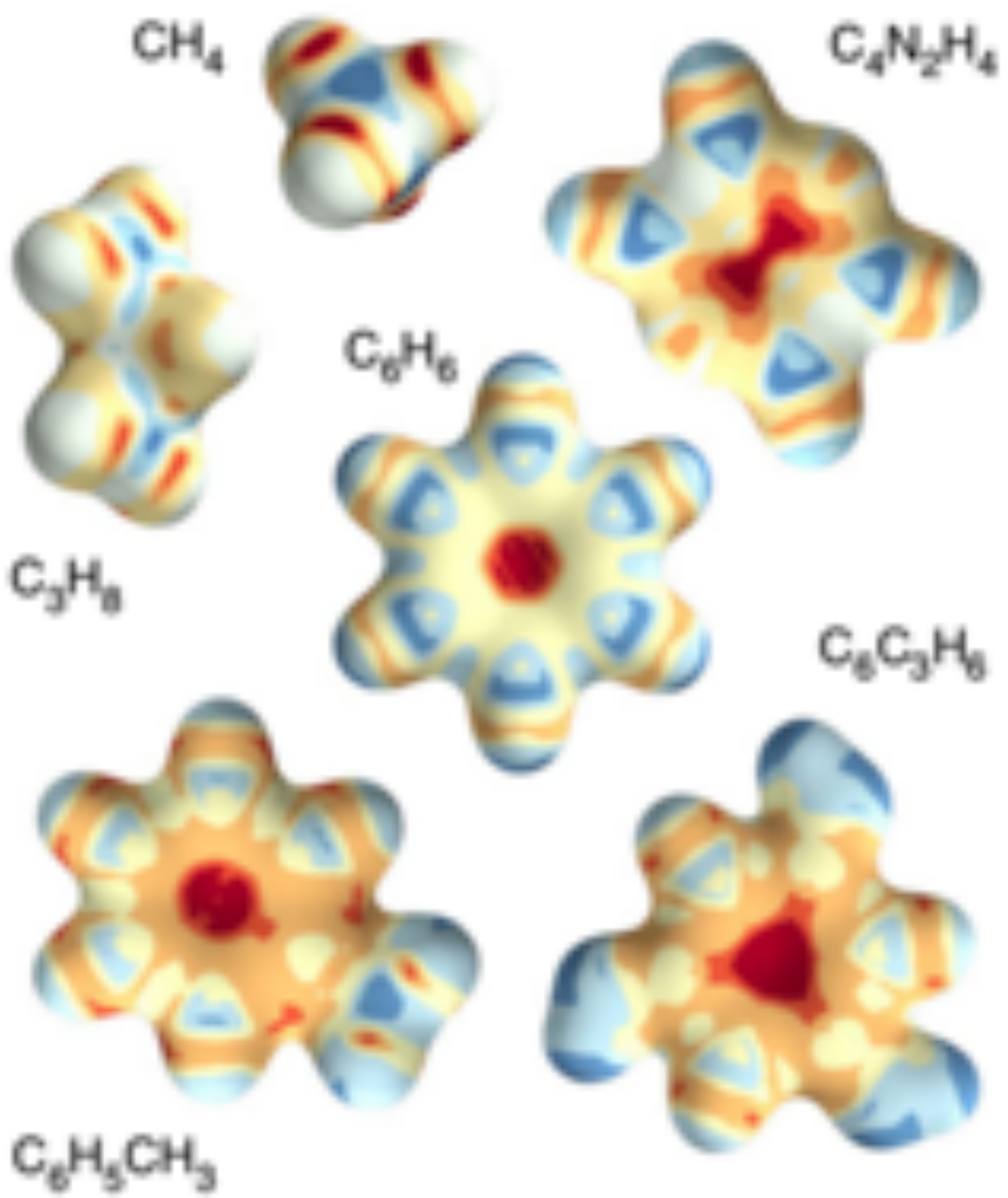


Why Interpretability ?

4) Interpretability in the sciences

Learn about the physical / biological / chemical mechanisms.
(e.g. find genes linked to cancer, identify binding sites …)





Why Interpretability ?

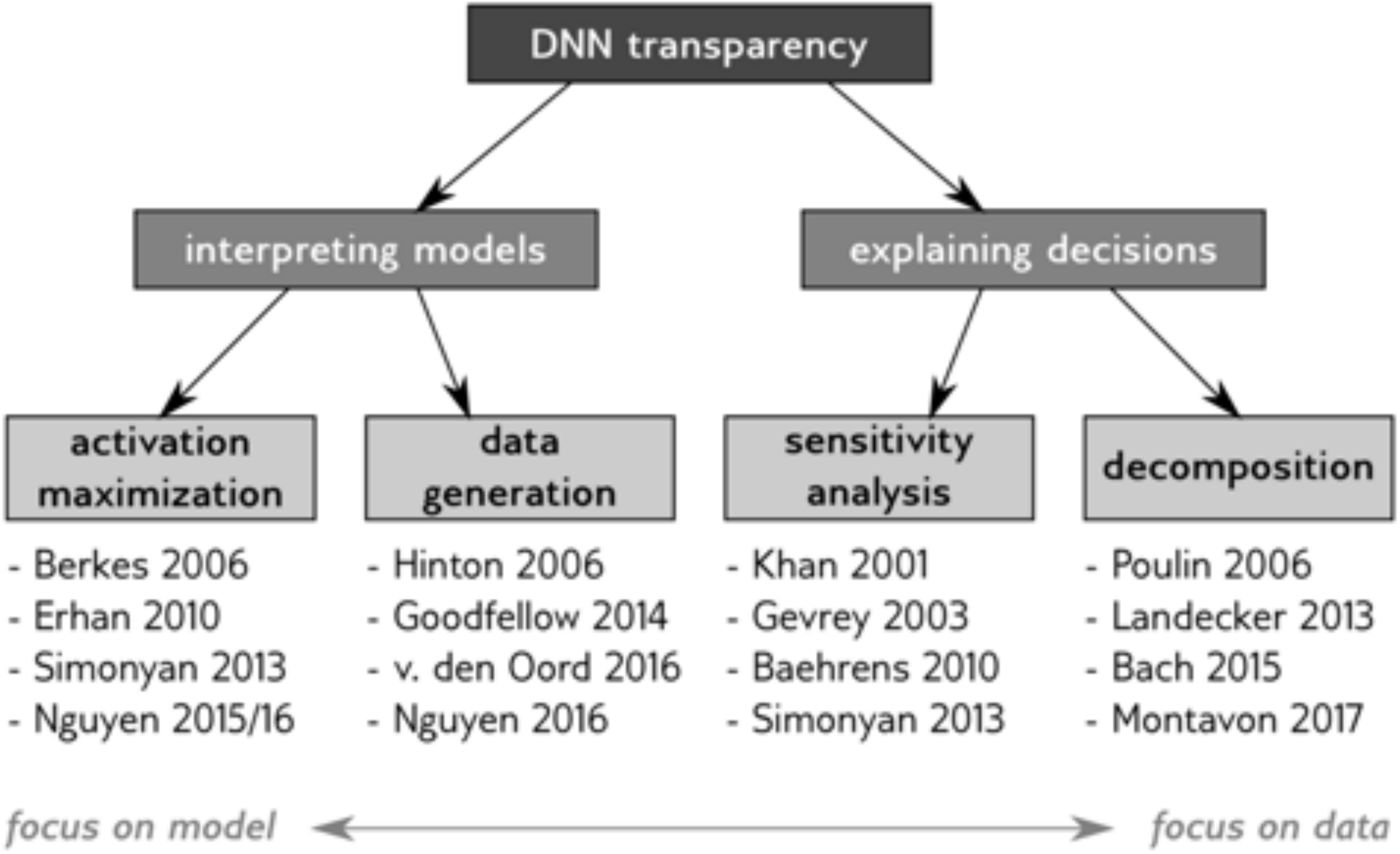
5) Compliance to legislation

European Union's new General Data Protection Regulation → "right to explanation"

Retain human decision in order to assign responsibility.

"With interpretability we can ensure that ML models work in compliance to proposed legislation."

Techniques of Interpretation



Techniques of Interpretation

Interpreting models
(ensemble)

- *find prototypical example of a category*
- *find pattern maximizing activity of a neuron*

→ **better understand internal representation**

Explaining decisions
(individual)

- *"why" does the model arrive at this particular prediction*
- *verify that model behaves as expected*

→ **crucial for many practical applications**

Techniques of Interpretation

In medical context

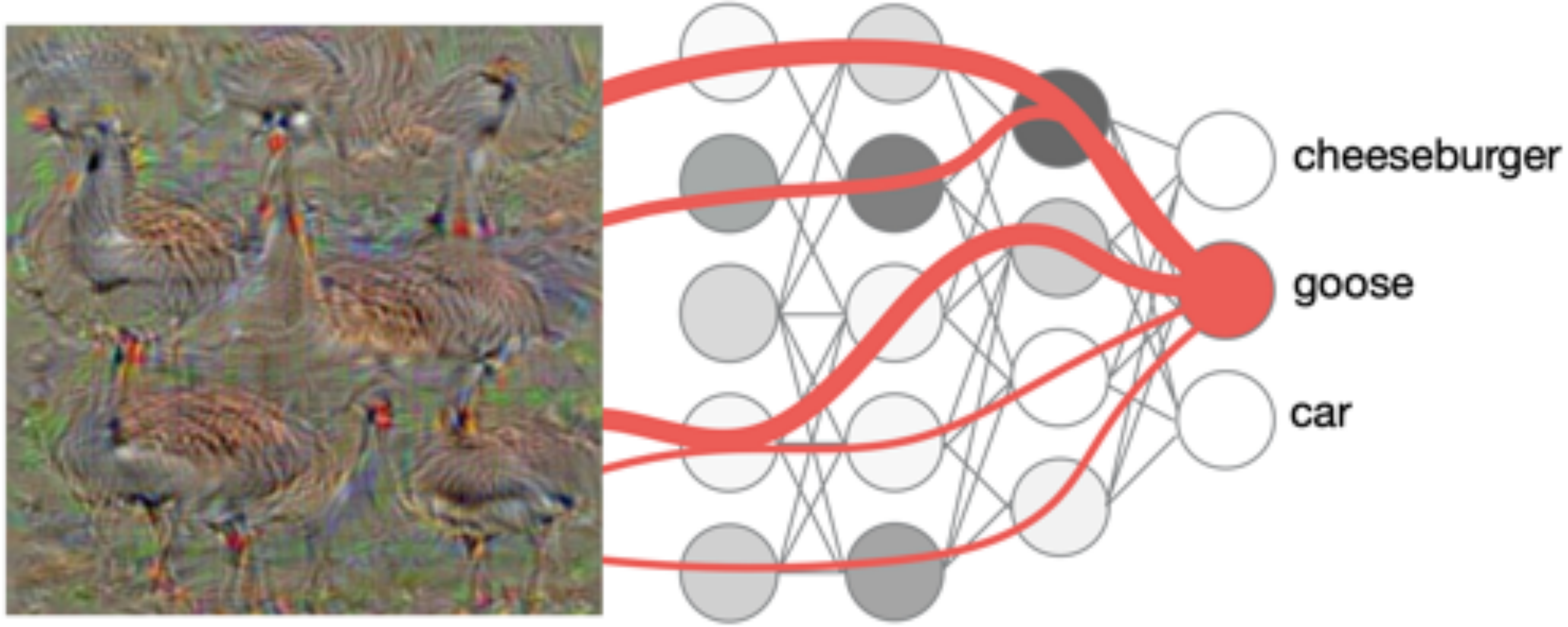
- Population view (ensemble)
  - Which symptoms are most common for the disease
  - Which drugs are most helpful for patients
- Patient's view (individual)
  - Which particular symptoms does the patient have
  - Which drugs does he need to take in order to recover

Both aspects can be important depending on who you are (FDA, doctor, patient).

Techniques of Interpretation

Interpreting models

- *find prototypical example of a category*
- *find pattern maximizing activity of a neuron*



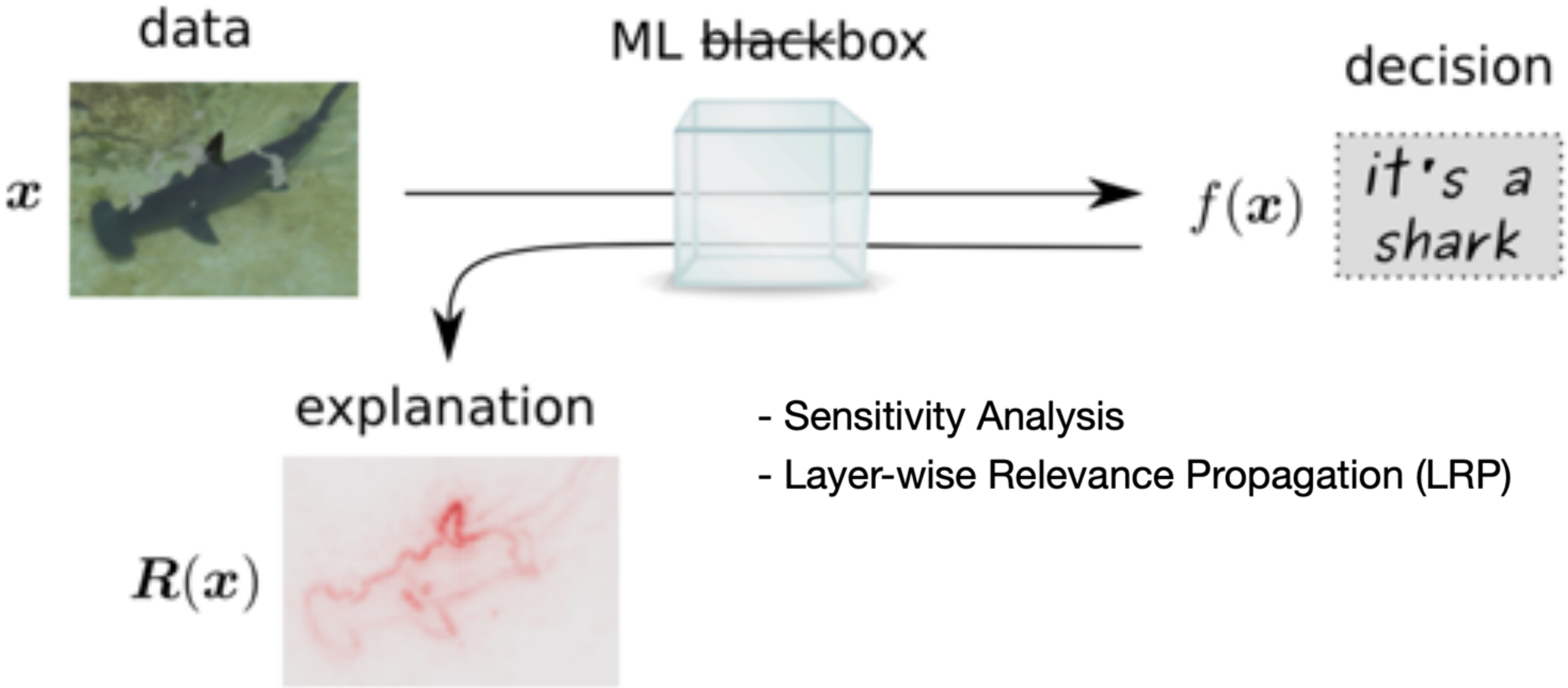
simple regularizer
(Simonyan et al. 2013)

$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$

Techniques of Interpretation

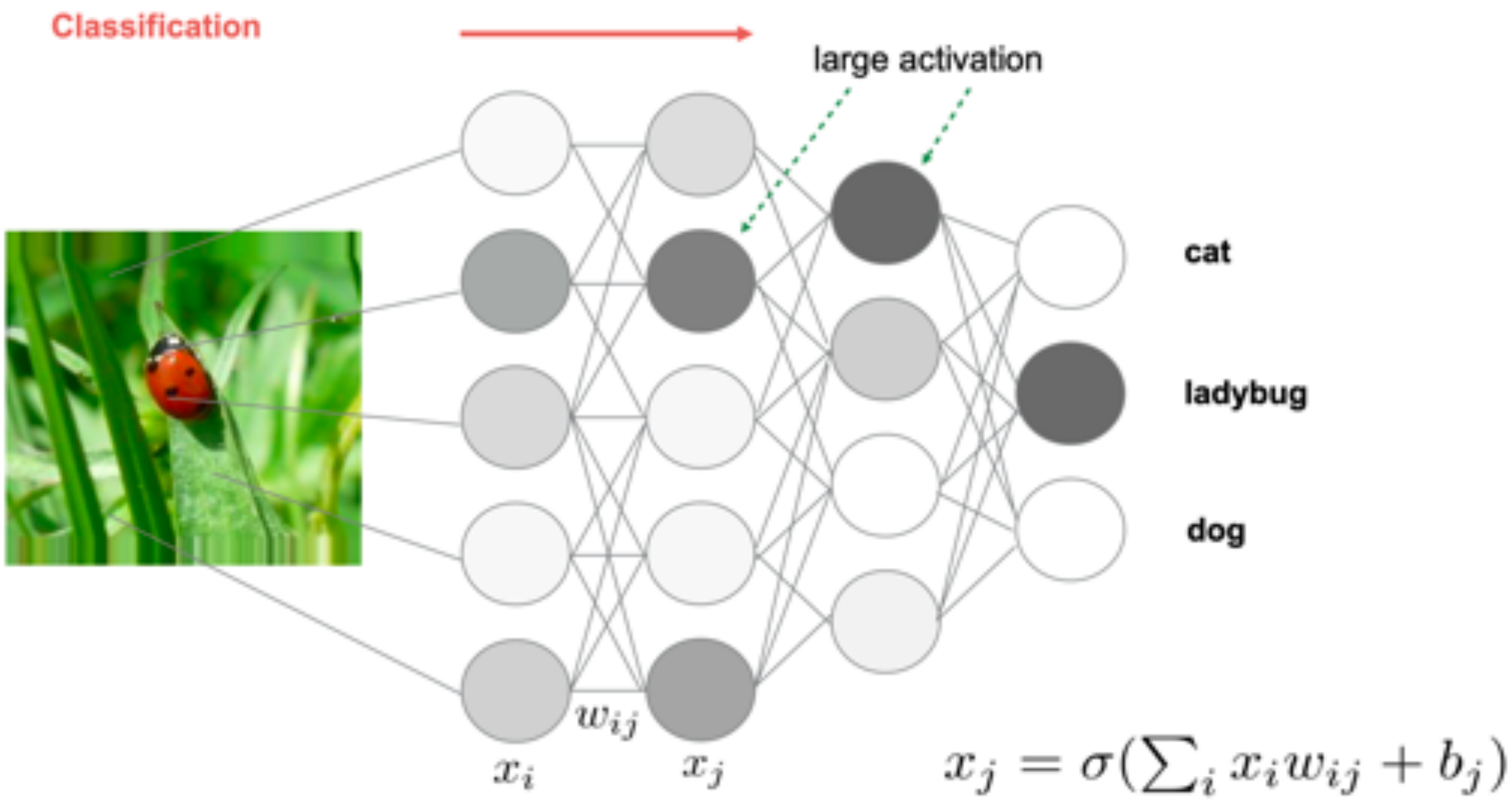
Explaining decisions

- *"why" does the model arrive at a certain prediction*
- *verify that model behaves as expected*

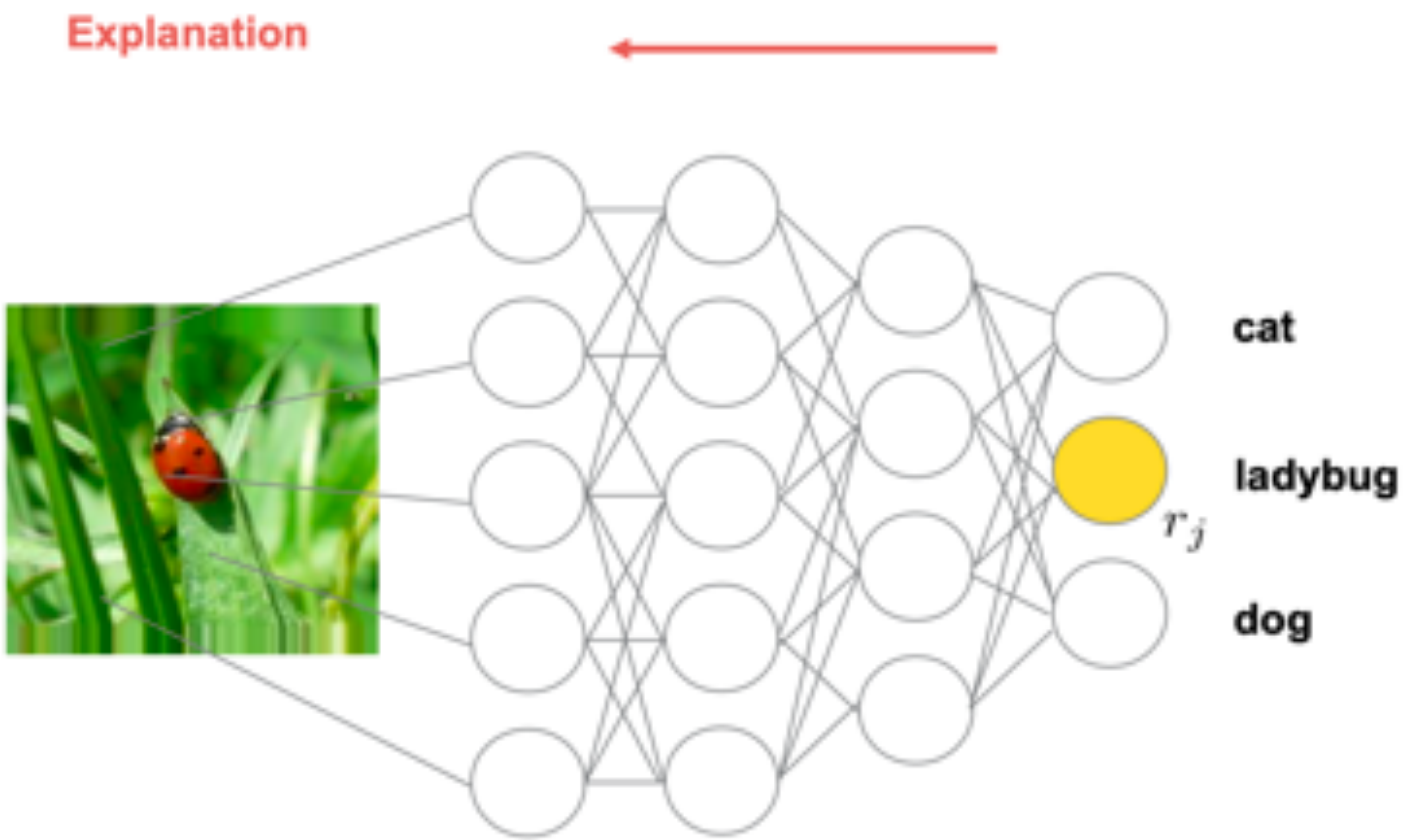


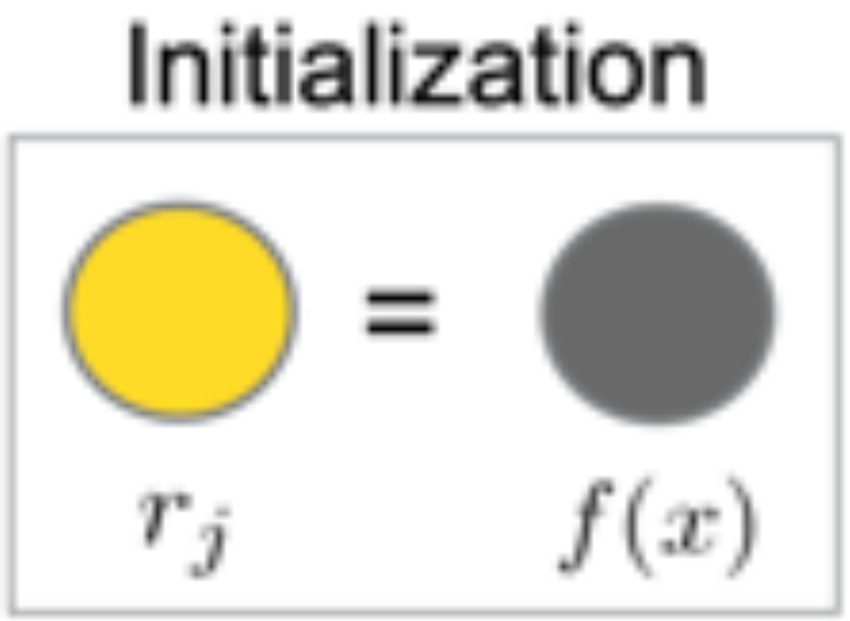
- Sensitivity Analysis
- Layer-wise Relevance Propagation (LRP)

Explaining Neural Network Predictions

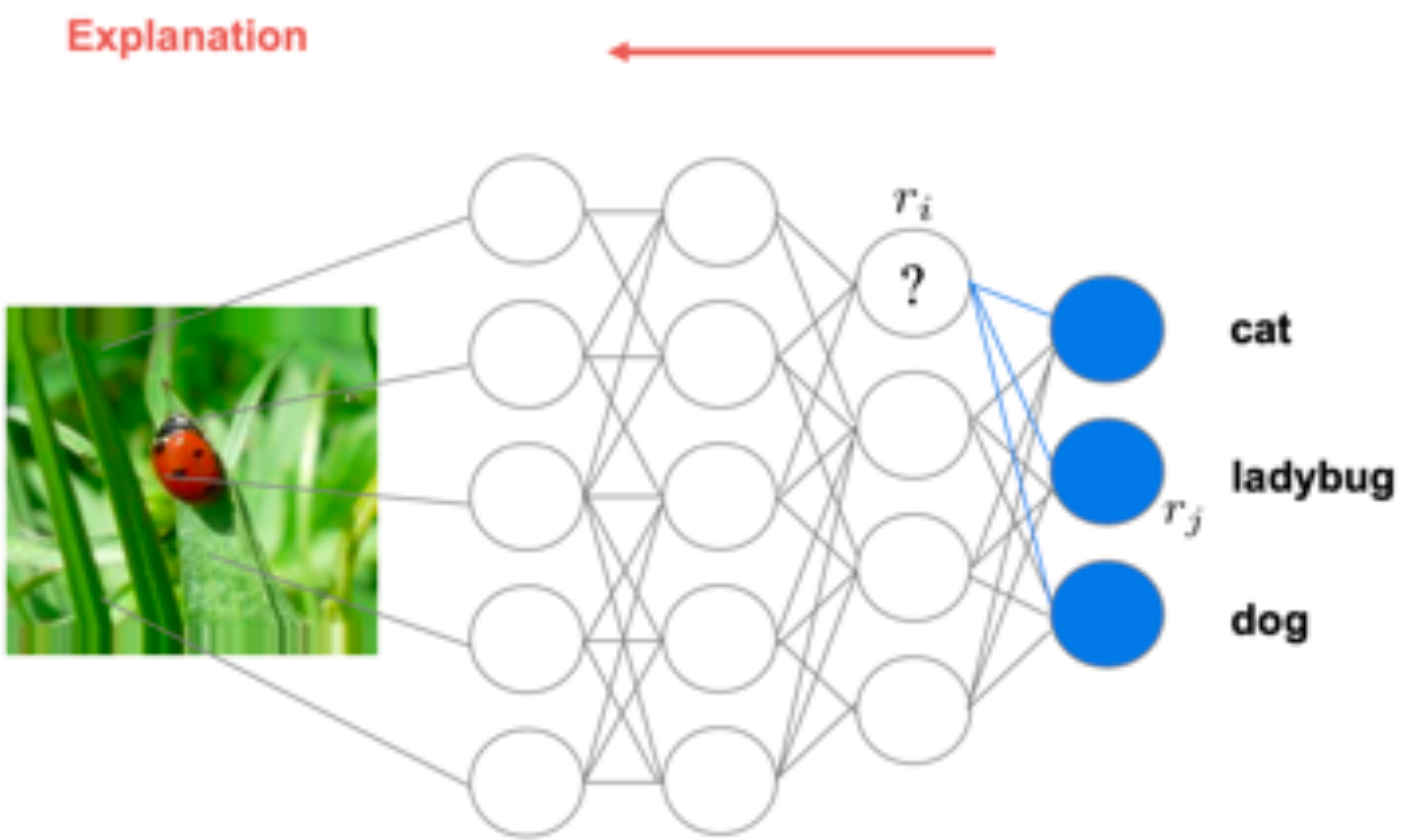


Explaining Neural Network Predictions

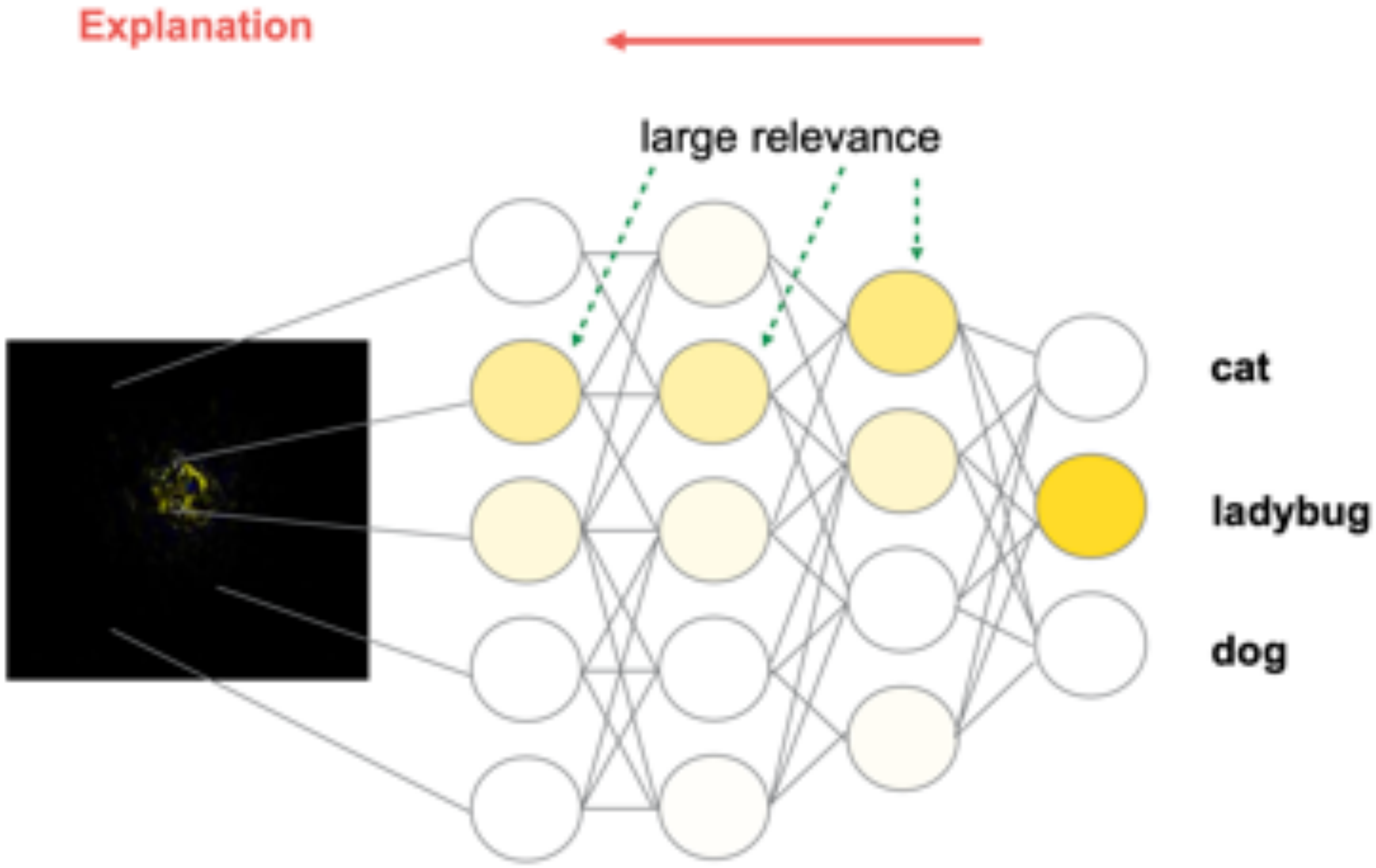


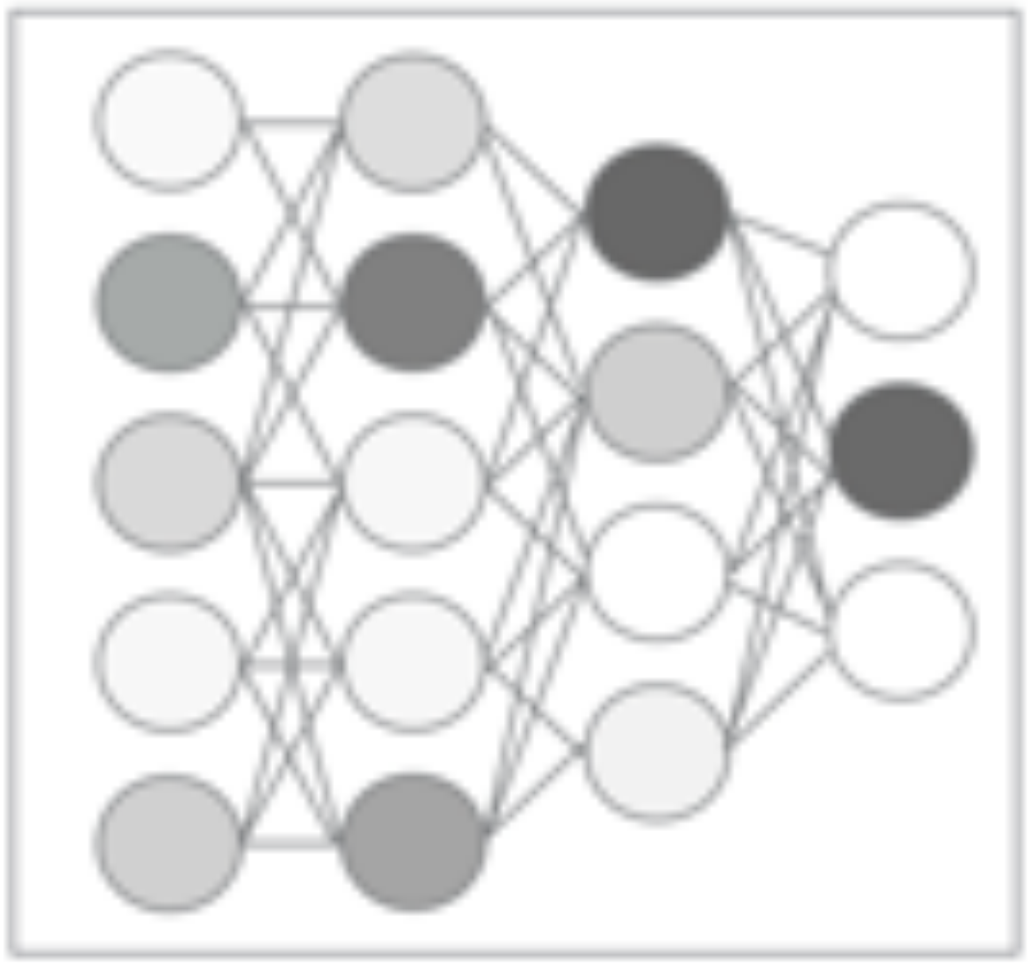


Explaining Neural Network Predictions

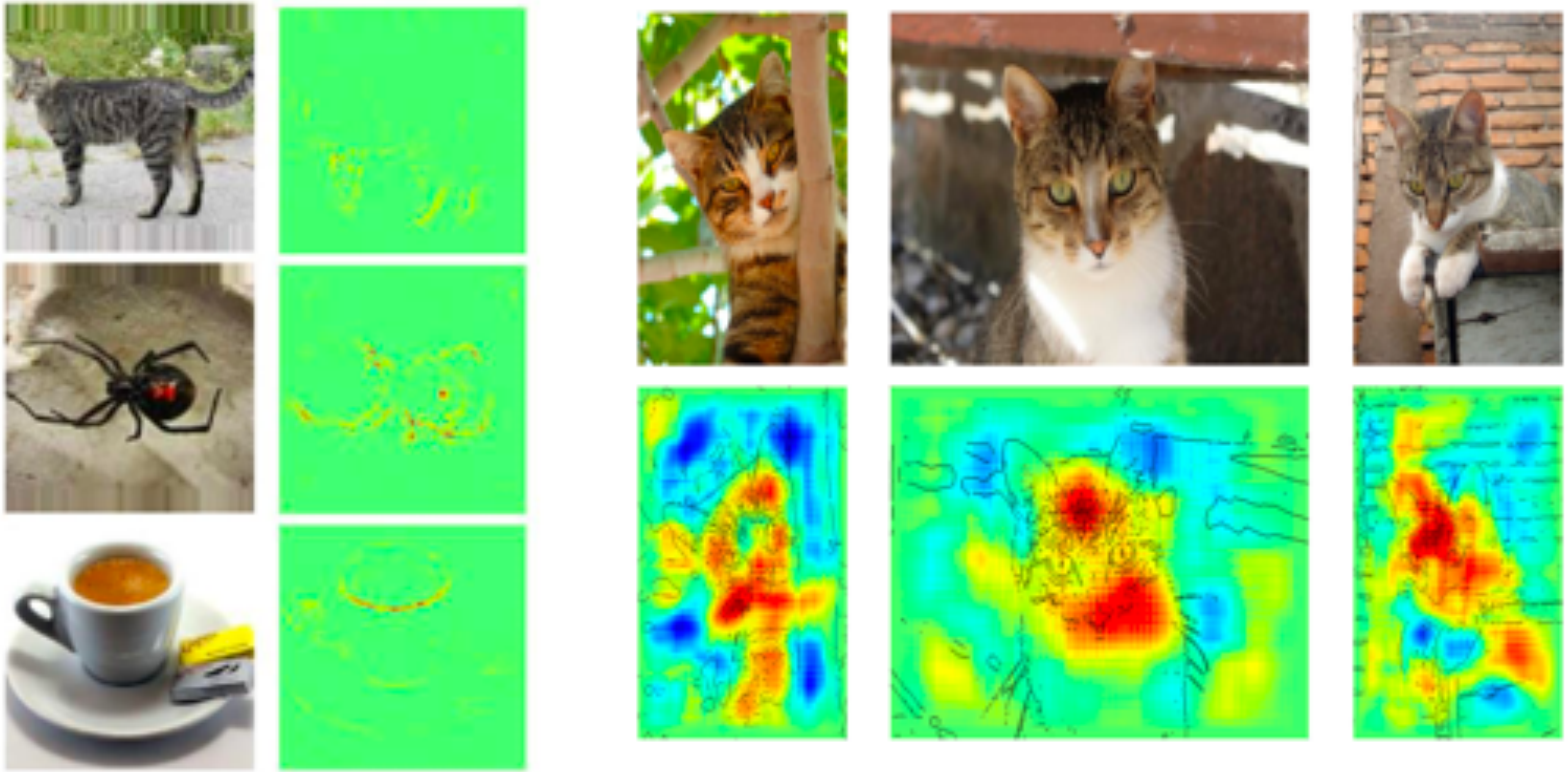


Explaining Neural Network Predictions





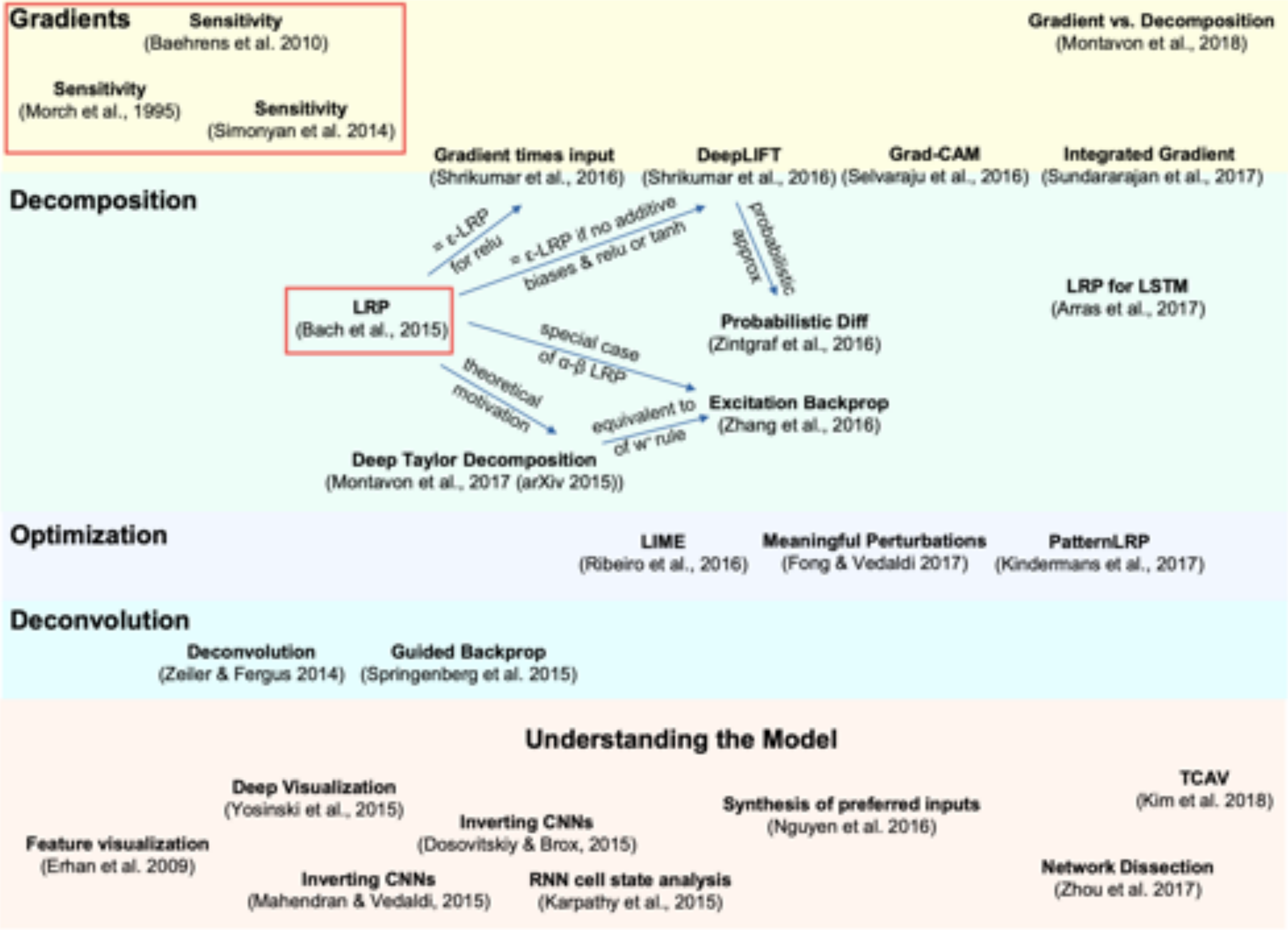
Explaining Predictions Pixel-wise



Neural networks

Kernel methods

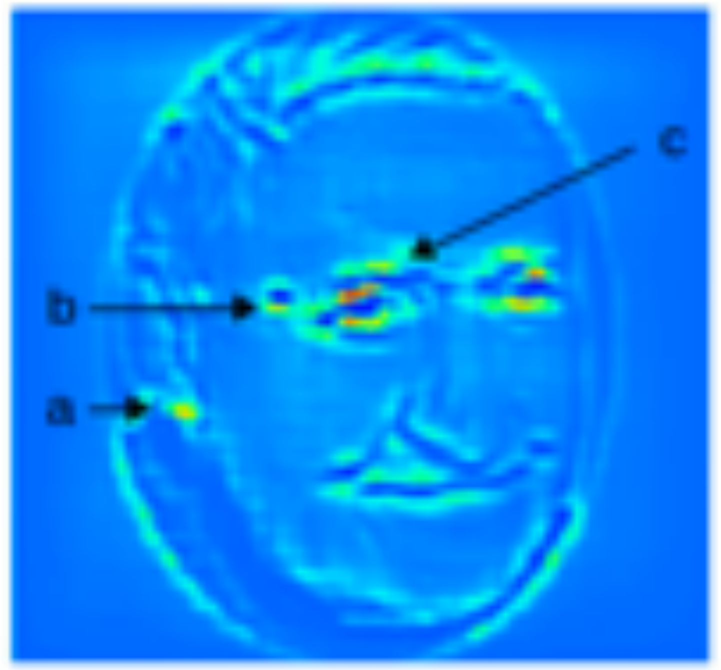
Historical remarks on Explaining Predictors



Applying Explanation in Vision and Text

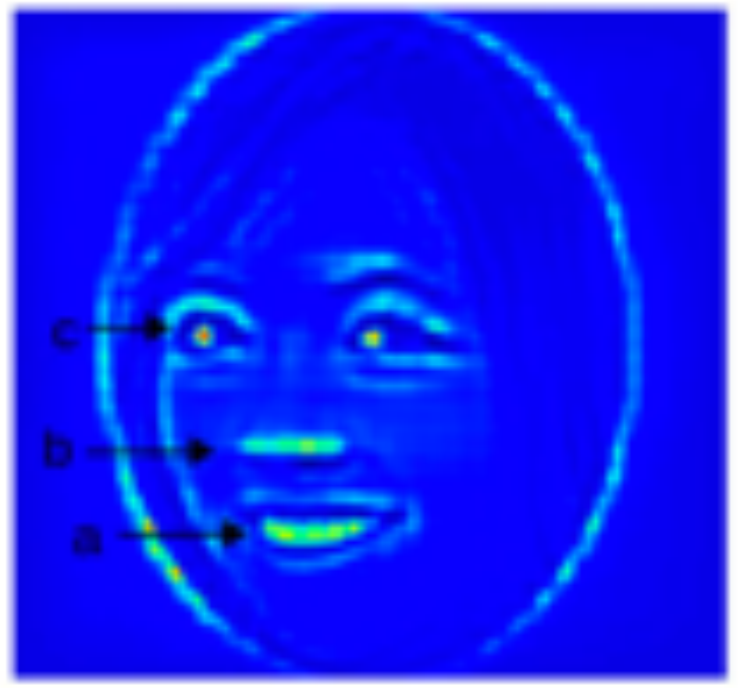
Application: Faces

What makes
you look old ?

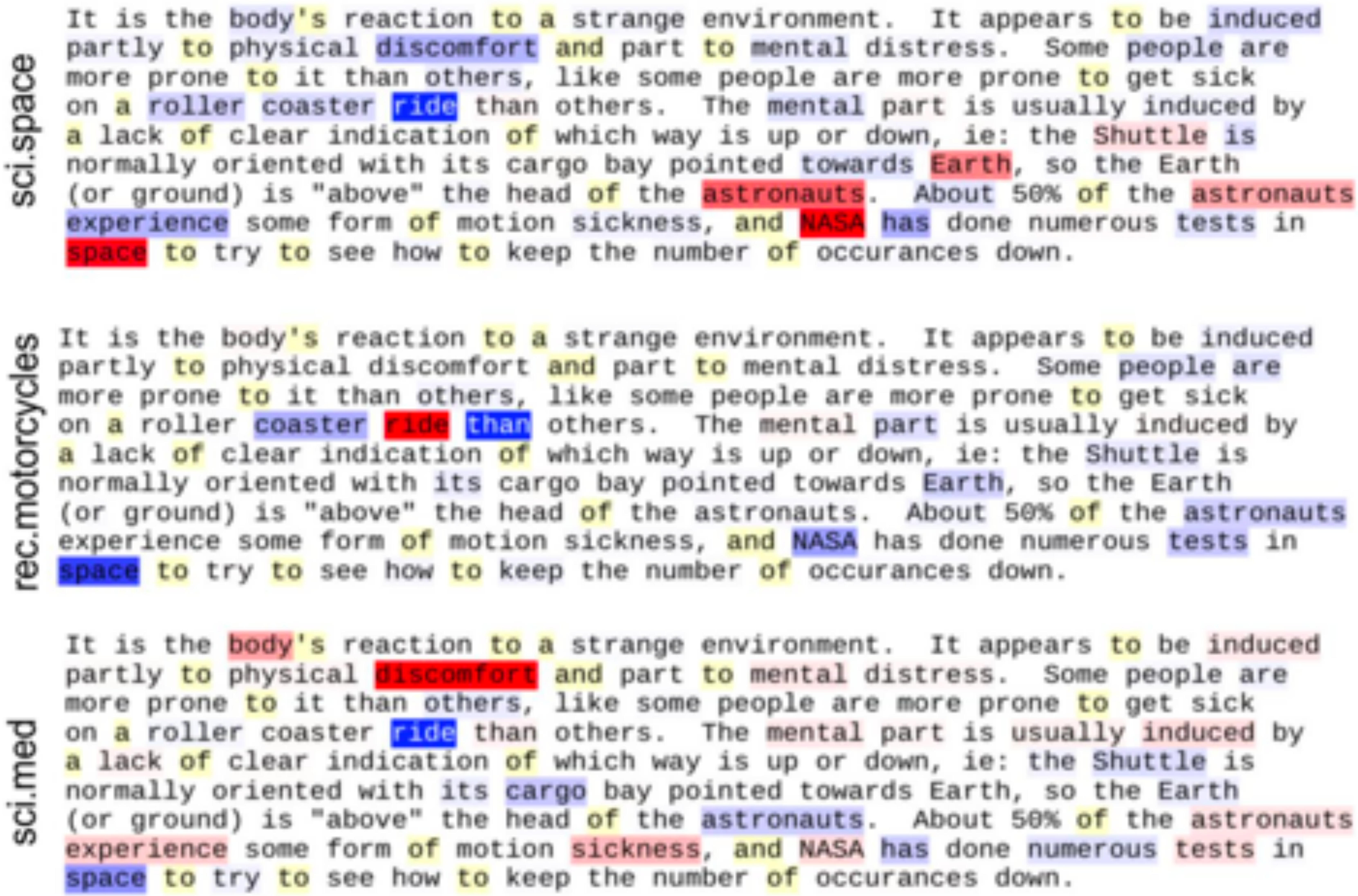


What makes
you look sad ?

What makes
you look attractive ?



Application: Document Classification



Sanity Checks for ‘Saliency’ Maps

Motivation

- Developer/Researcher: Model Debugging.
- Safety concerns.
- Ethical concerns.
- Trust: Satiate ‘societal’ need for reasoning to trust an automated system learned from data.

Goals: Model Debugging

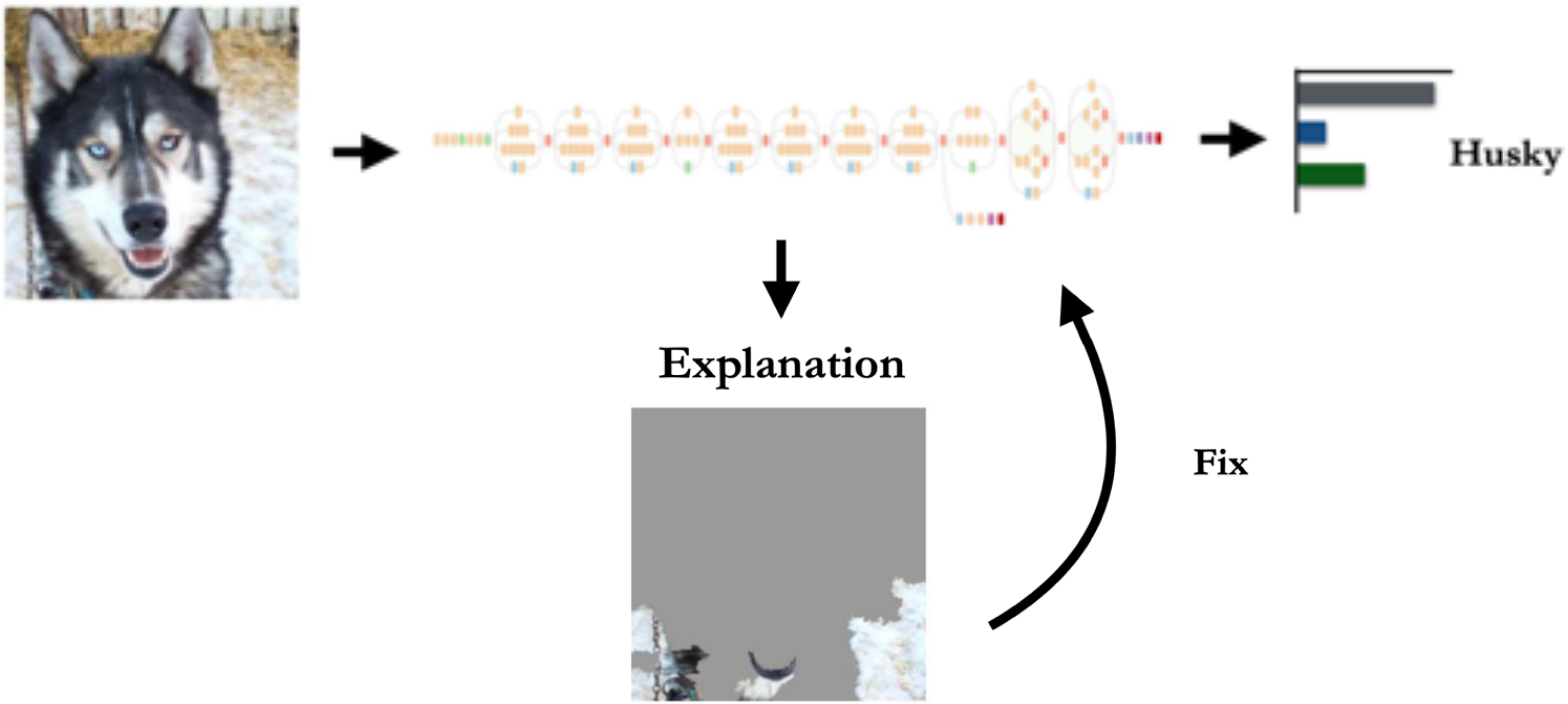
Model Debugging: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.

(a) Husky classified as wolf (b) Explanation

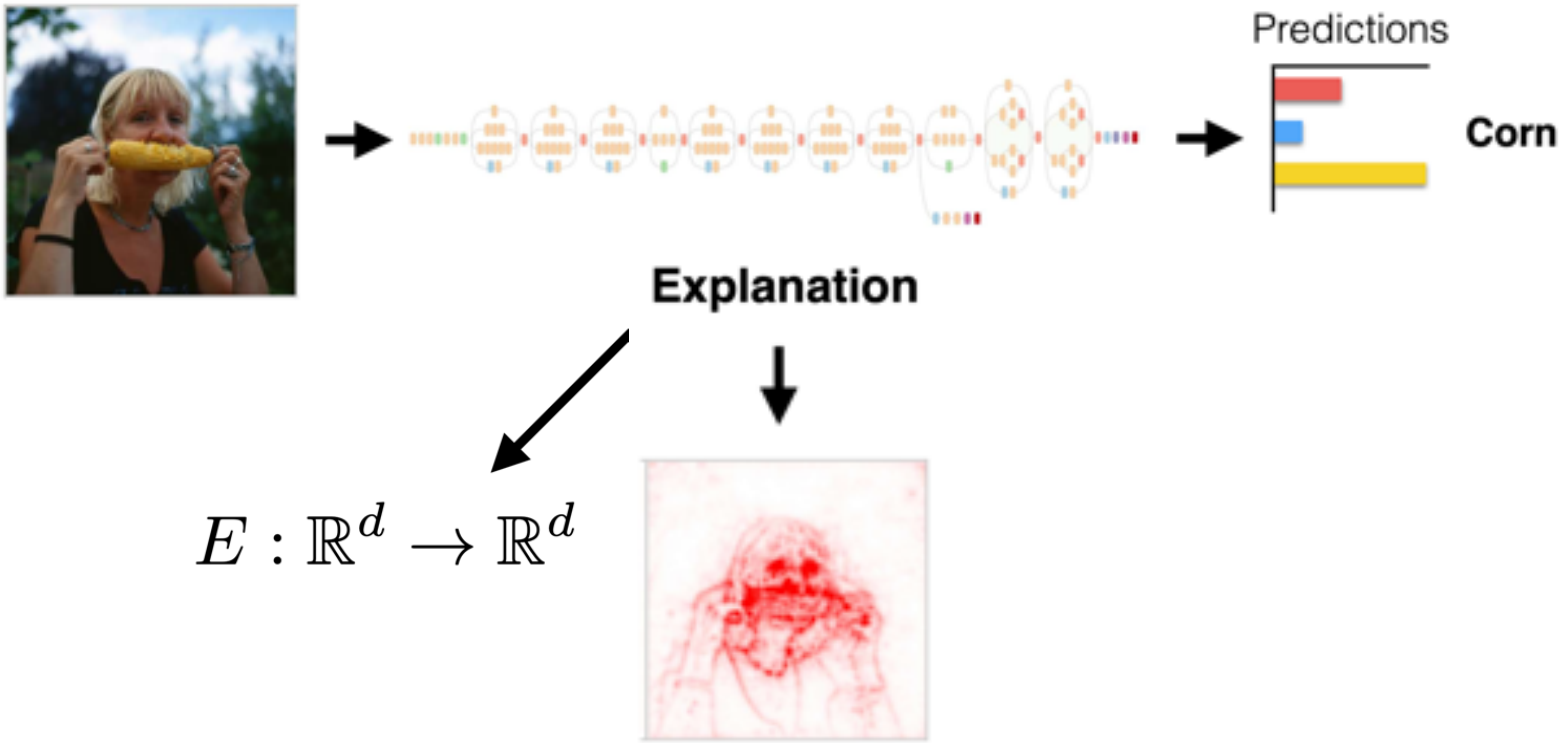
[Ribeiro+ 2016]

Promise of Explanations

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.



Saliency/Attribution Maps



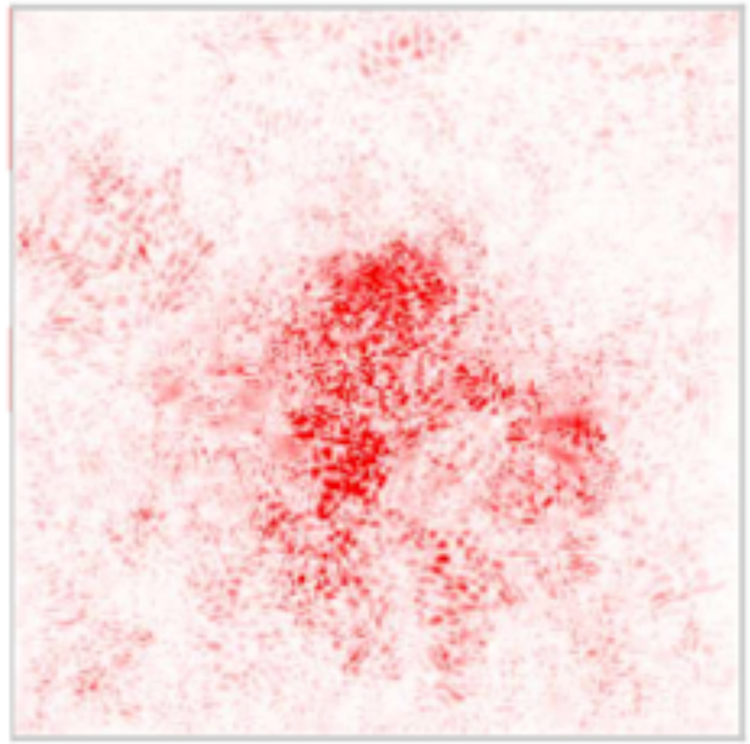
Attribution maps provide 'relevance' scores for each dimension of the input.

How to compute attribution?

Attribution

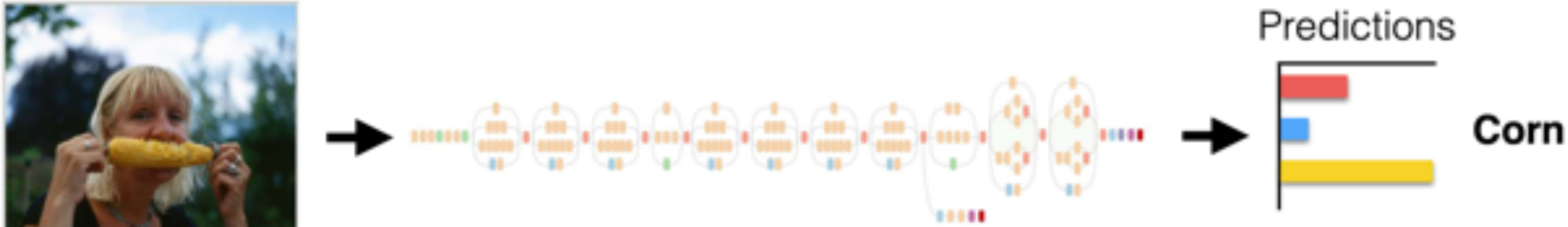
$$E_{grad}(x) = \frac{\partial S_i}{\partial x}$$

Gradient

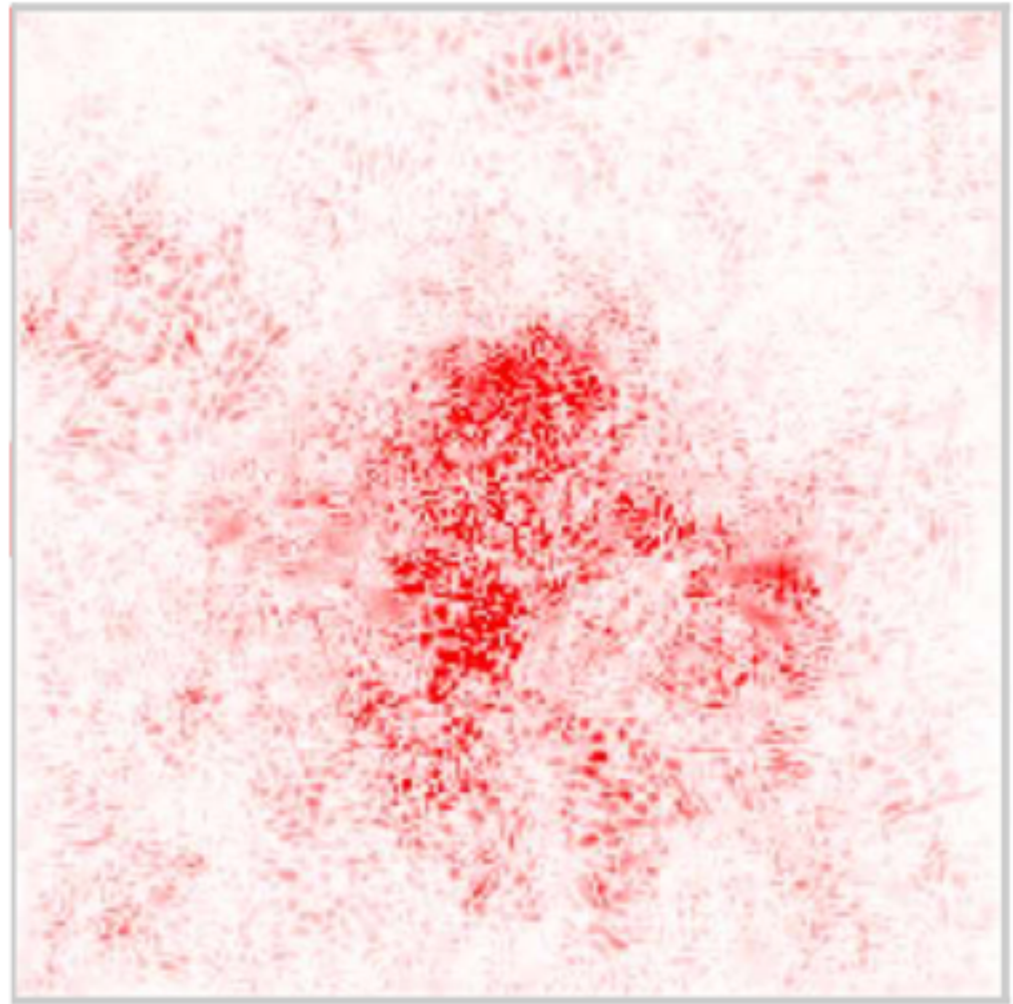


[SVZ'13]

Some Issues with the Gradient



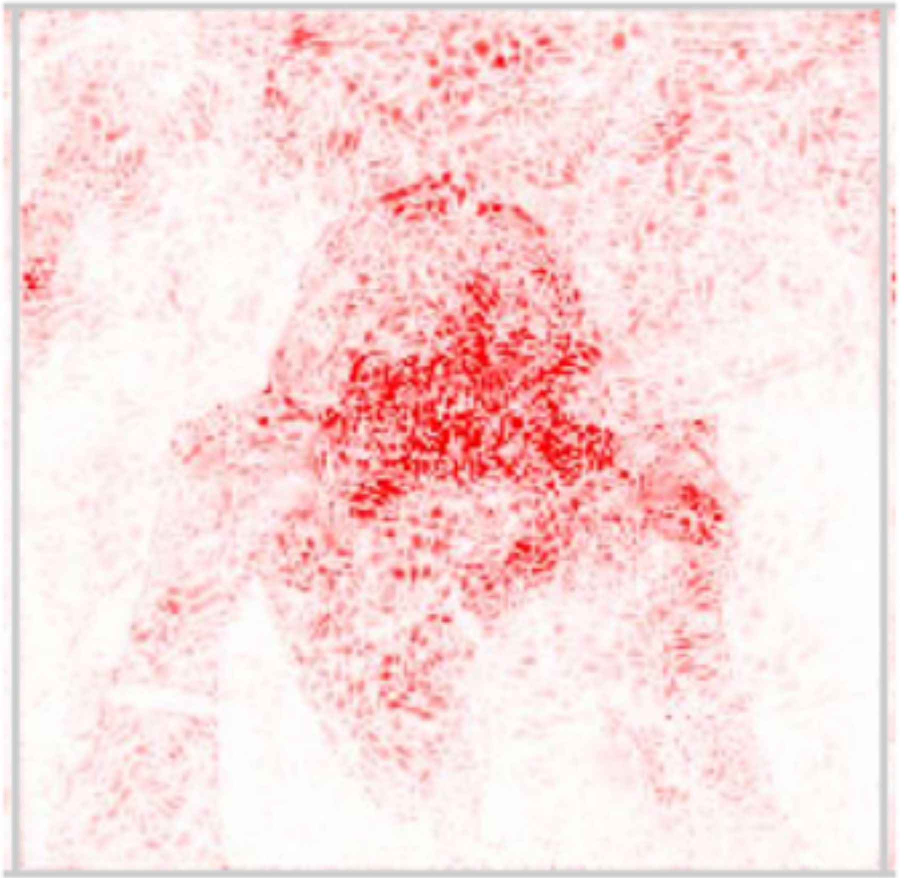
Gradient



'Visually noisy', and can violate sensitivity w.r.t. a baseline input [Sundararajan et. al., Shrikumar et. al., and Smilkov et. al.]

Integrated Gradients

Integrated Gradients



[STY'17]

$$E_{\mathrm{IG}}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x} + \alpha(x - \bar{x})}{\partial x} d\alpha$$

Sum of 'interior' gradients.

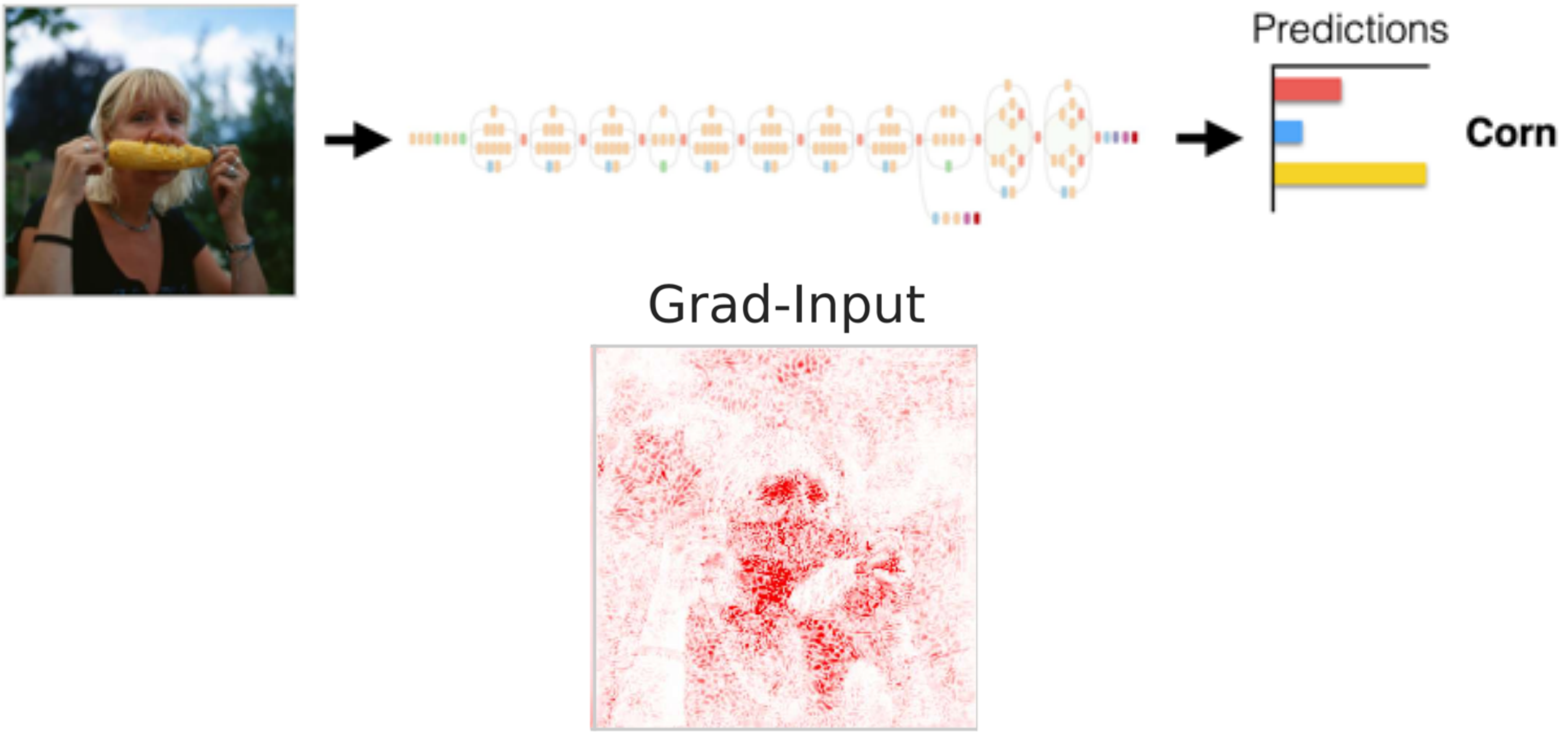
SmoothGrad

$$E_{\text{sg}}(x) = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i),$$

[STKVW'17]

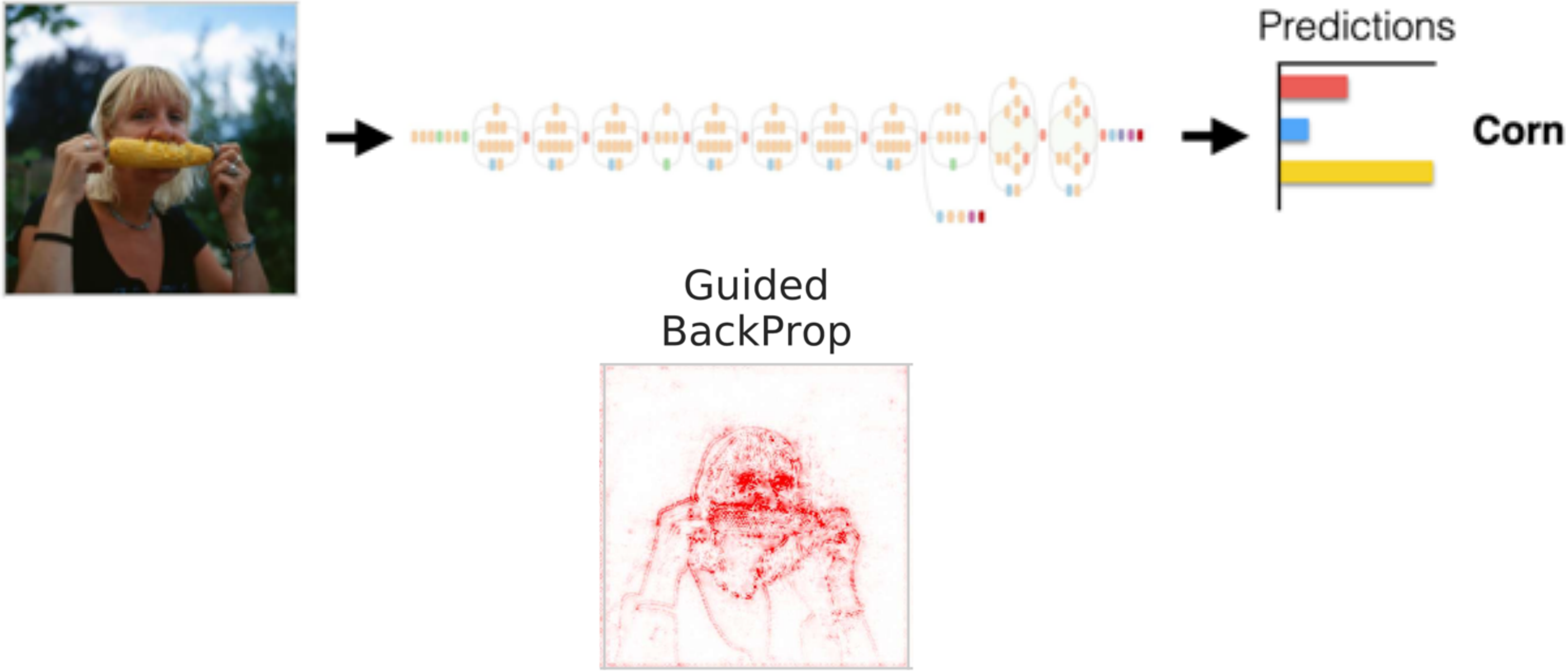
Average attribution of 'noisy' inputs.

Gradient-Input



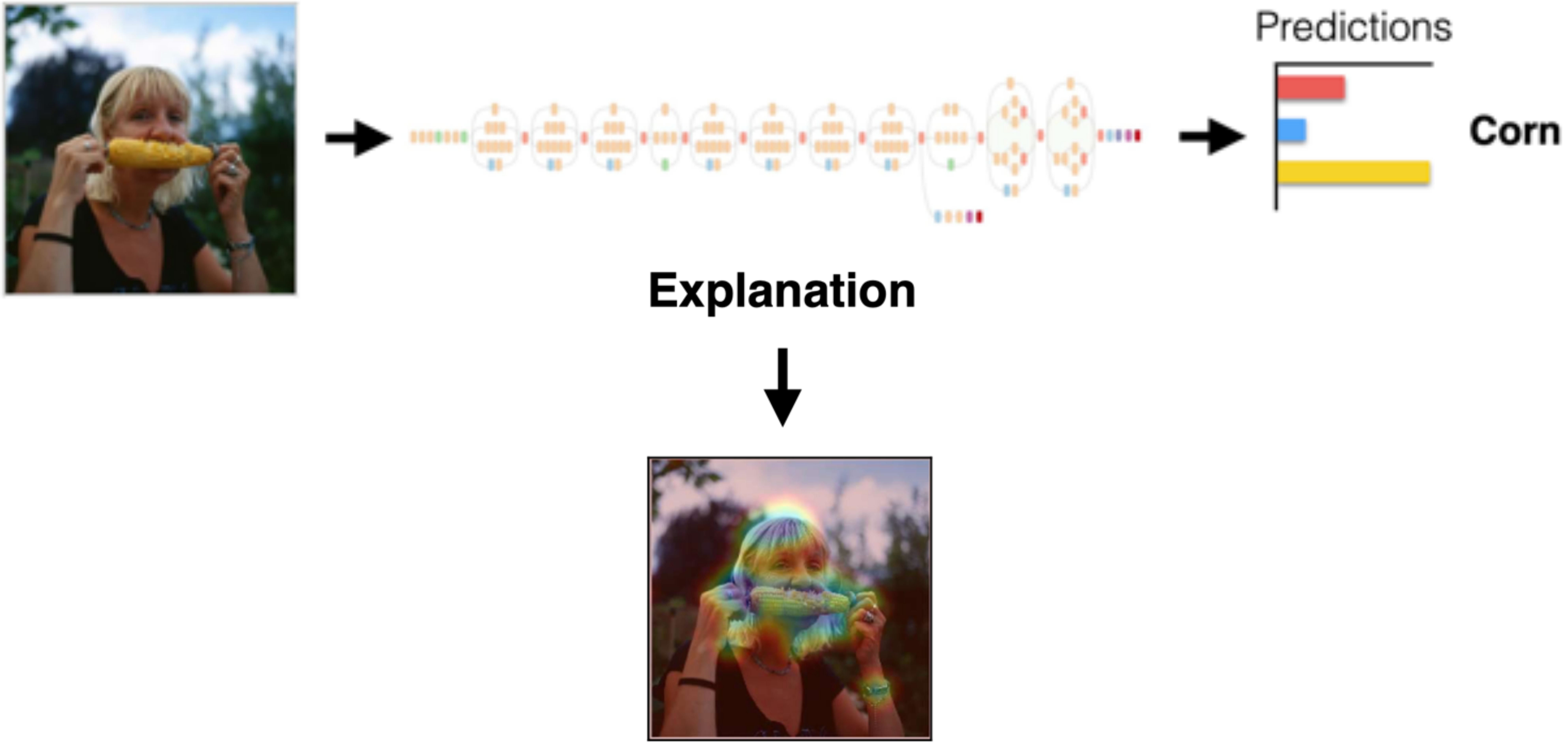
Element-wise product of gradient and input.

Guided BackProp



Zero out ‘negative’ gradients and ‘activations’ while back-propagating.

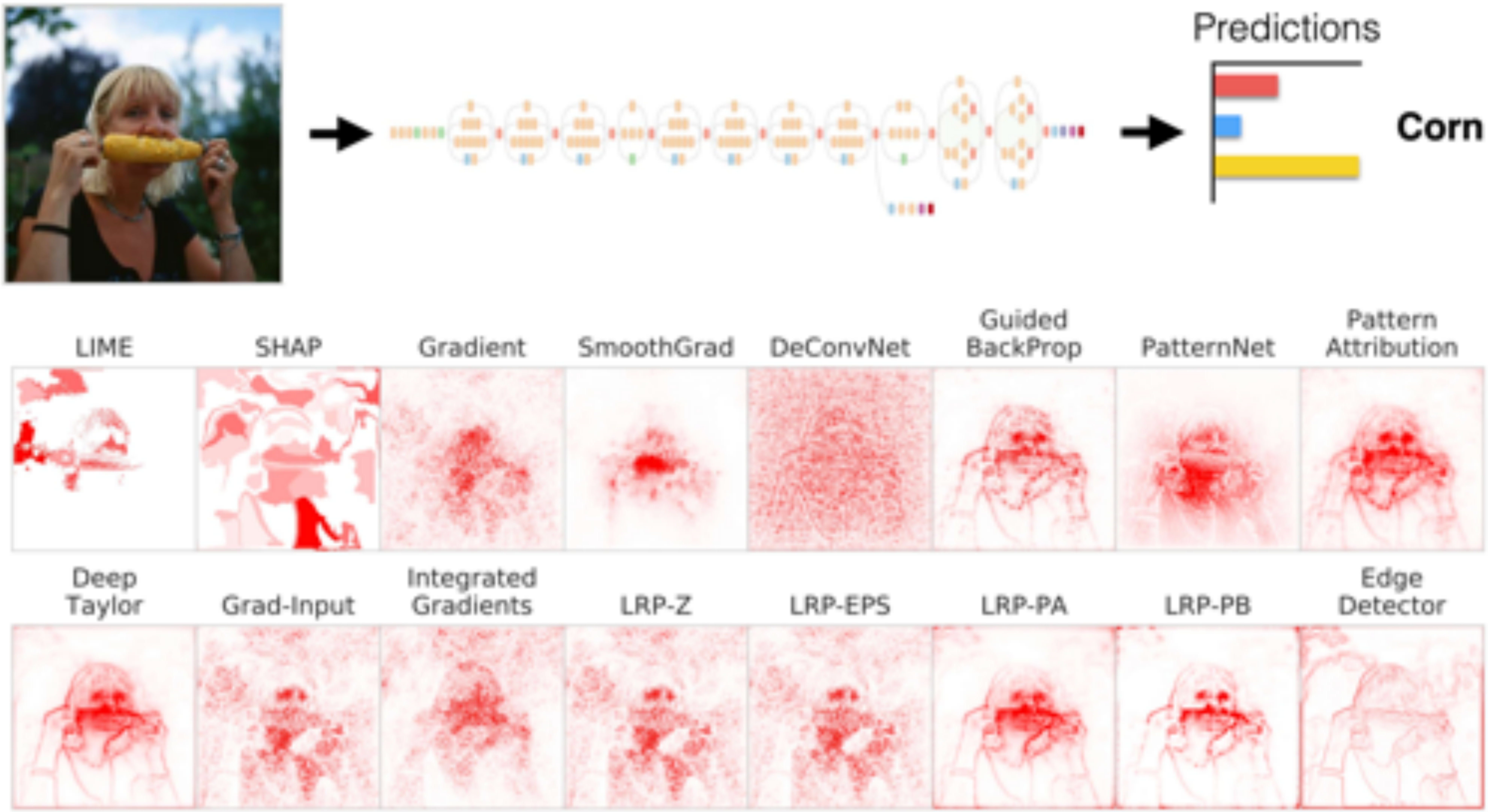
Other Learned Kinds



[FV'17]

Formulate an explanation as through learned patch removal.

The Selection Conundrum



The Selection Conundrum

For a particular **task** and **model**, how should a developer/researcher select **which method to use**?

Desirable Properties

Sensitivity to the parameters of a **model** to be explained.

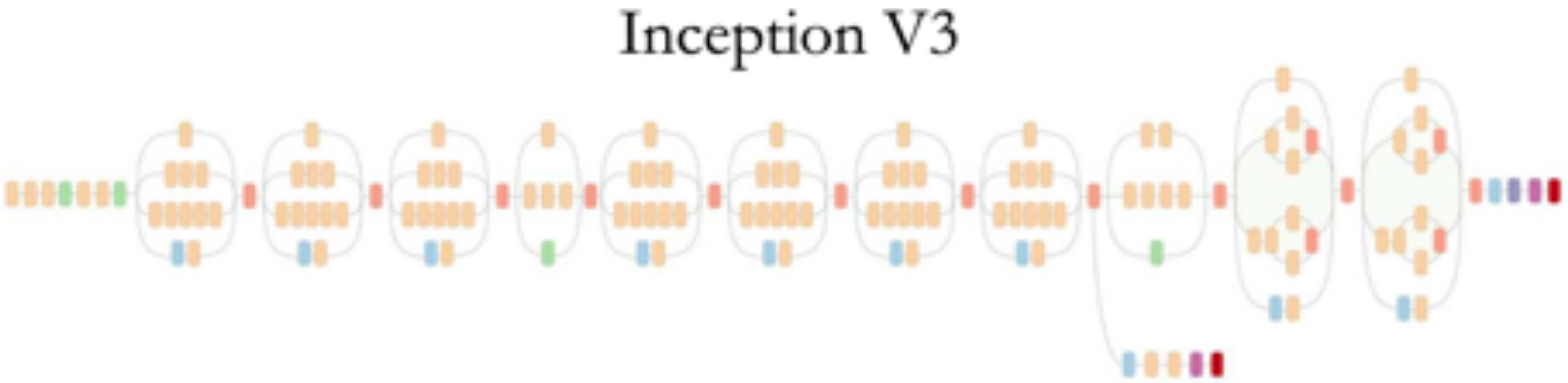
Depend on the labeling of the **data**, i.e., reflect the relationship between inputs and outputs.

Sanity Checks

- **Model parameter randomization test:** randomize (re-initialize) the parameters of a model and now compare attribution maps for a trained model to those derived from a randomized model.

- **Data randomization test**: compare attribution maps for a model trained with correct labels to those derived from a model trained with random labels.

Model Parameter Randomization

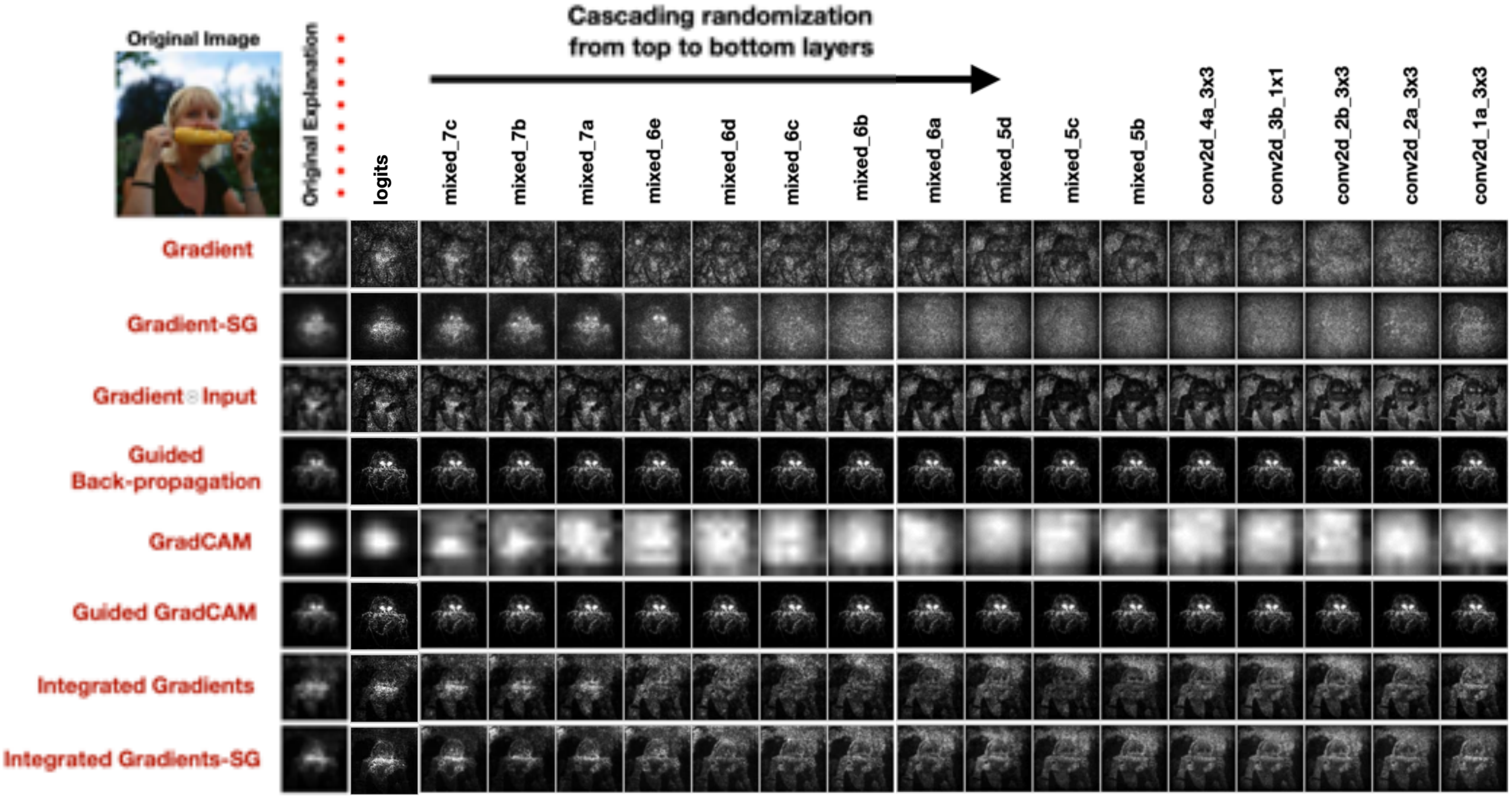


Cascading randomization from top to bottom layers.

Independent layer randomization.

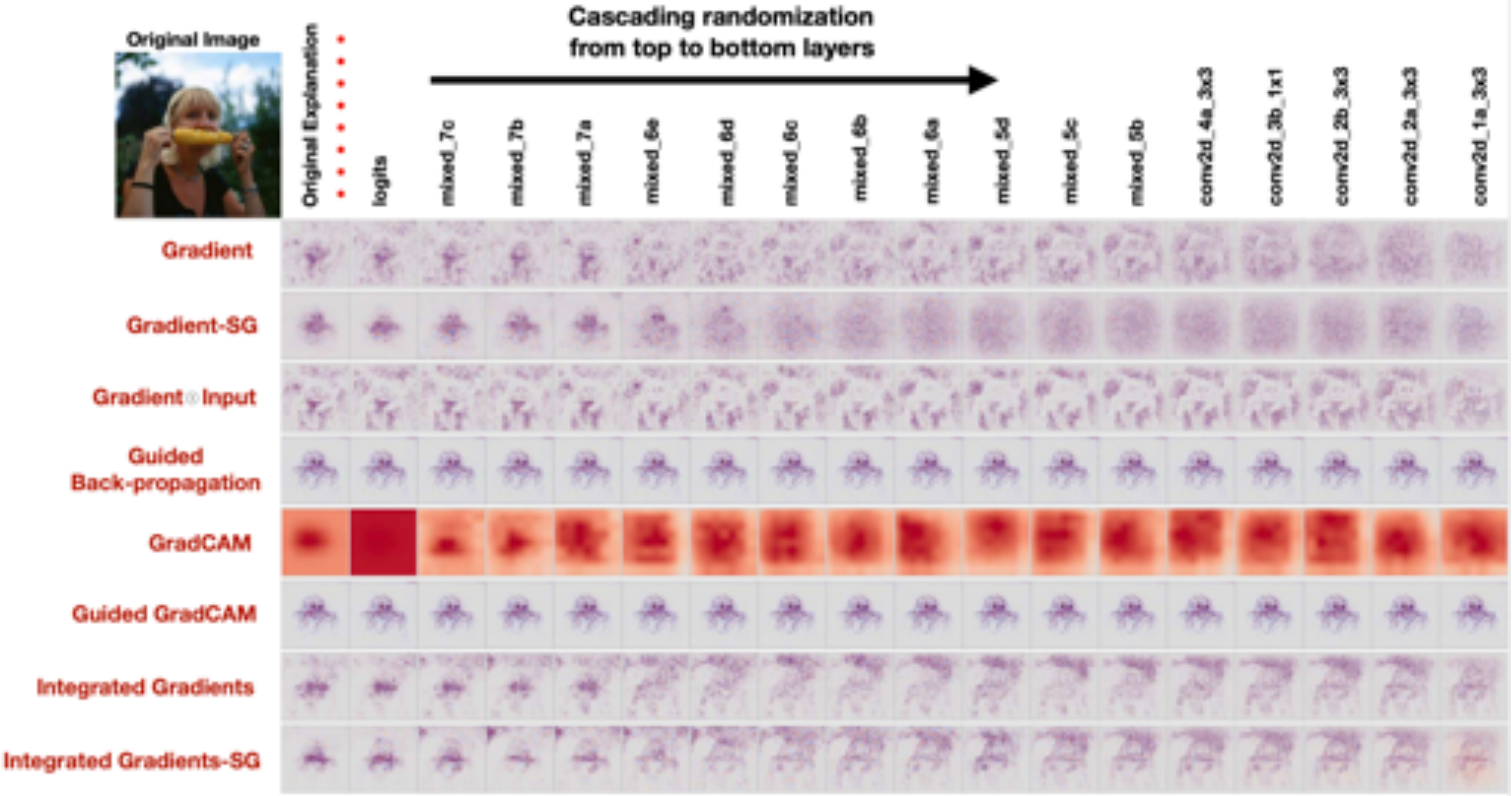
Model Parameter Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



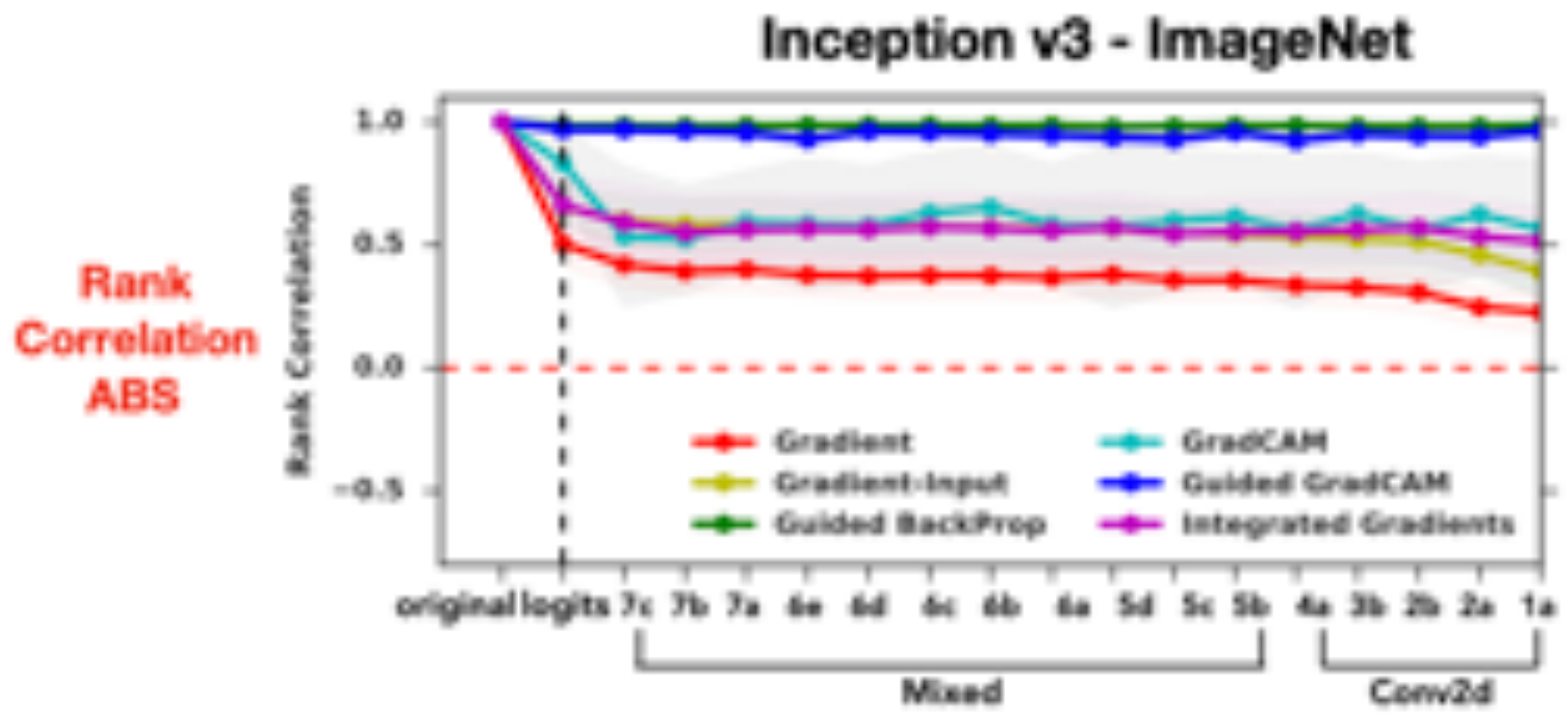
Model Parameter Randomization

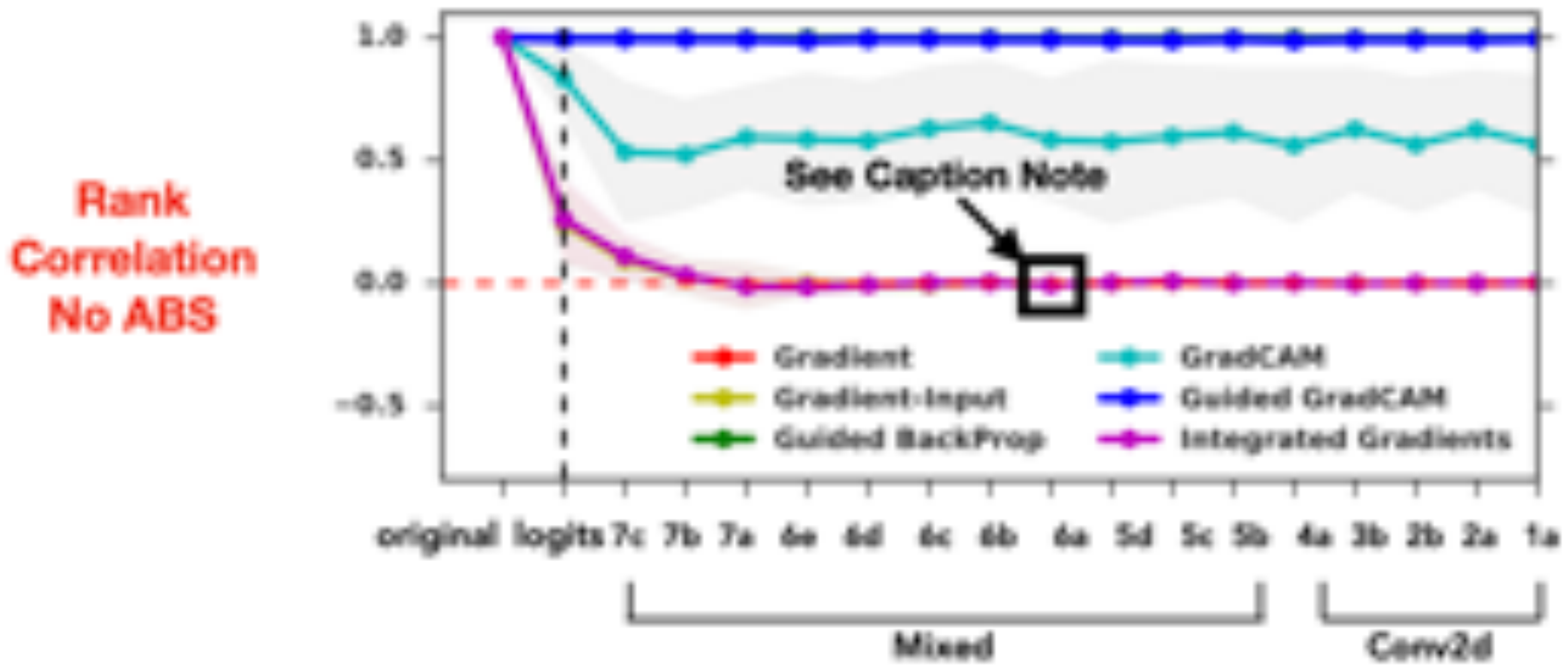
Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



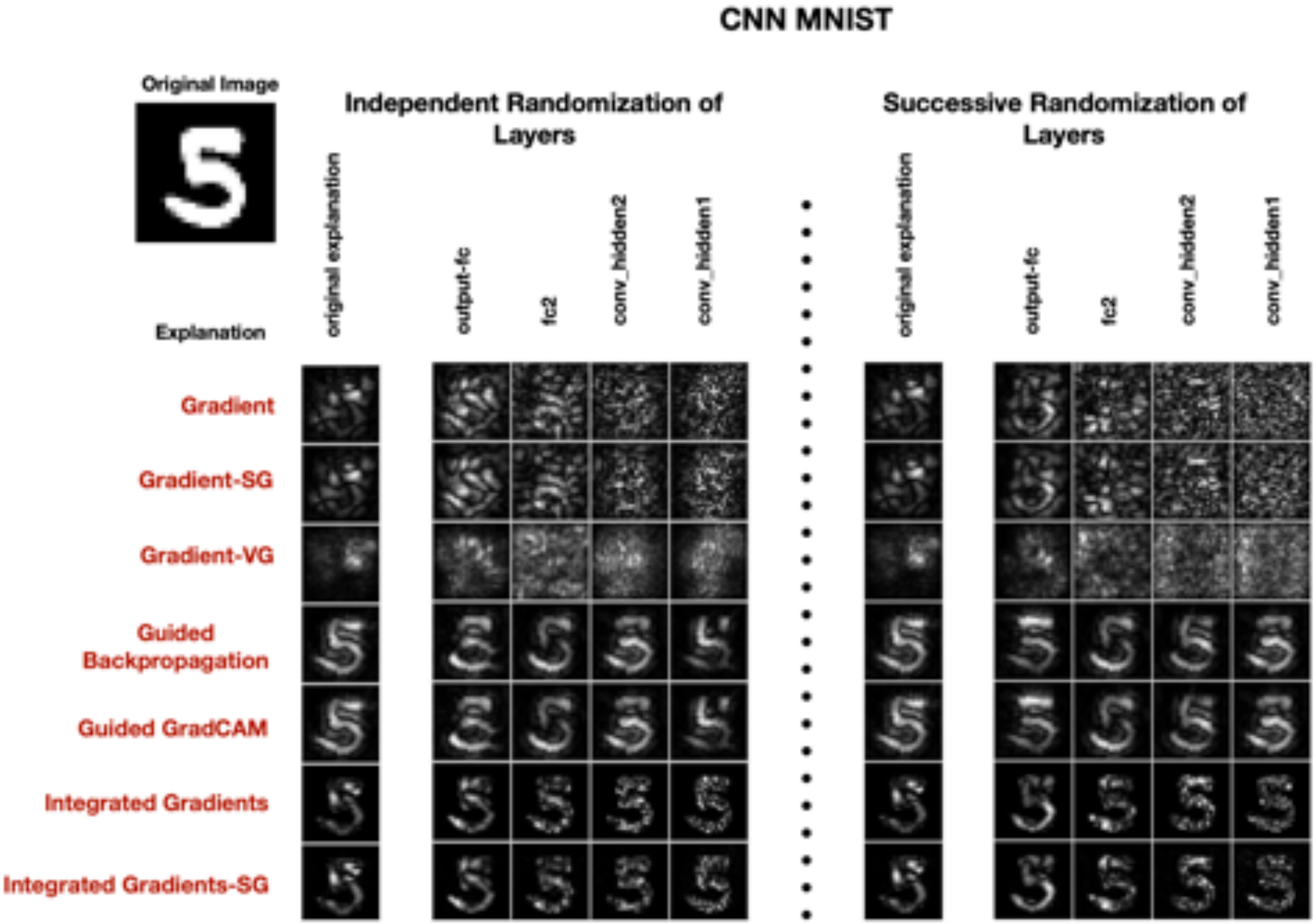
Metrics

- Rank correlation of attribution from model with trained weights to those derived from partially randomized models.
- Attribution sign changes. Roughly similar regions are, however, still attributed.

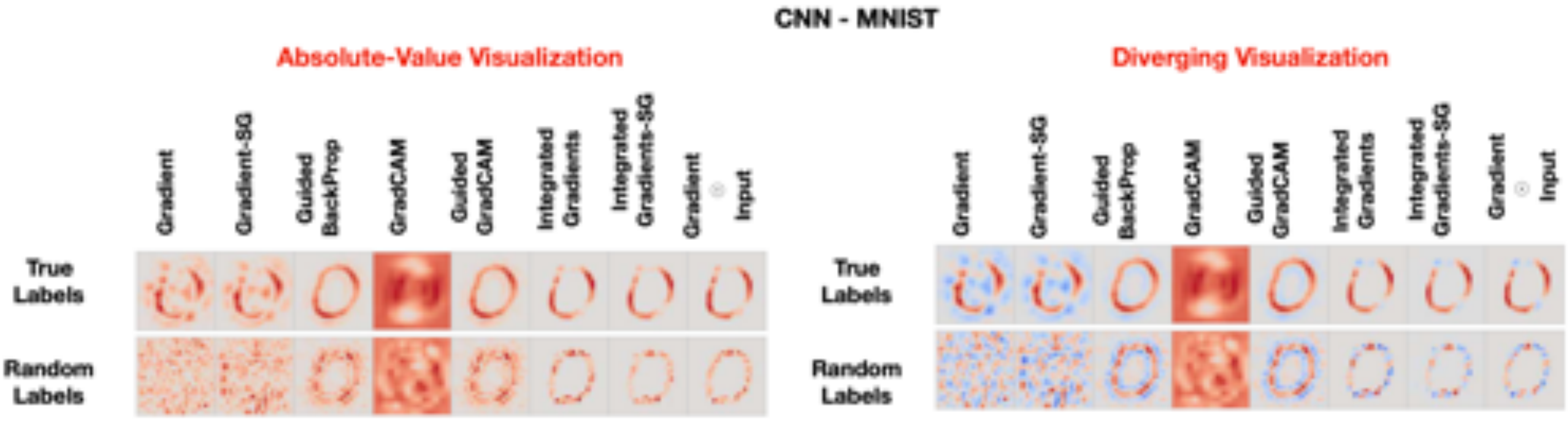




Model Parameter Randomization



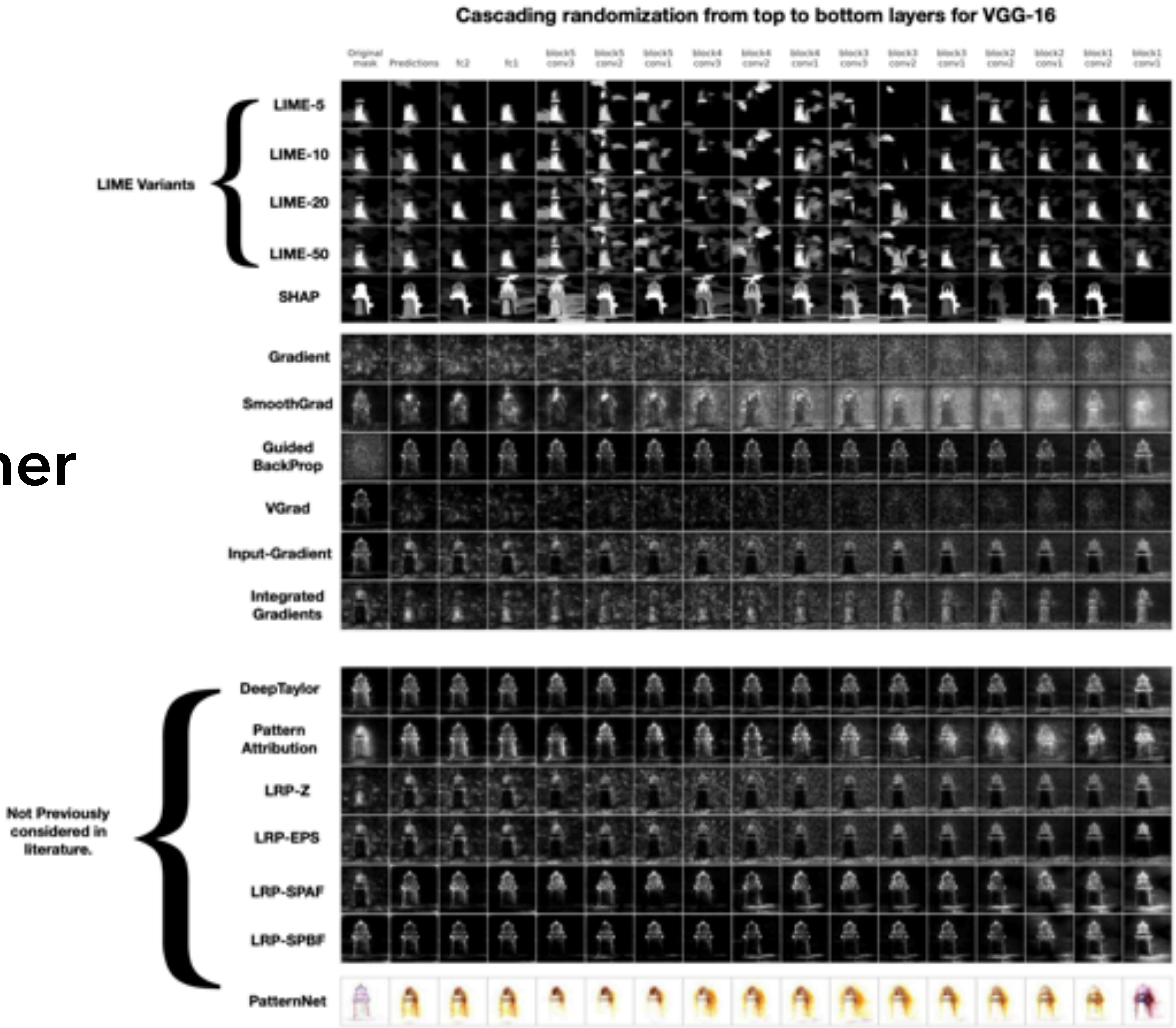
Data Randomization



Summary

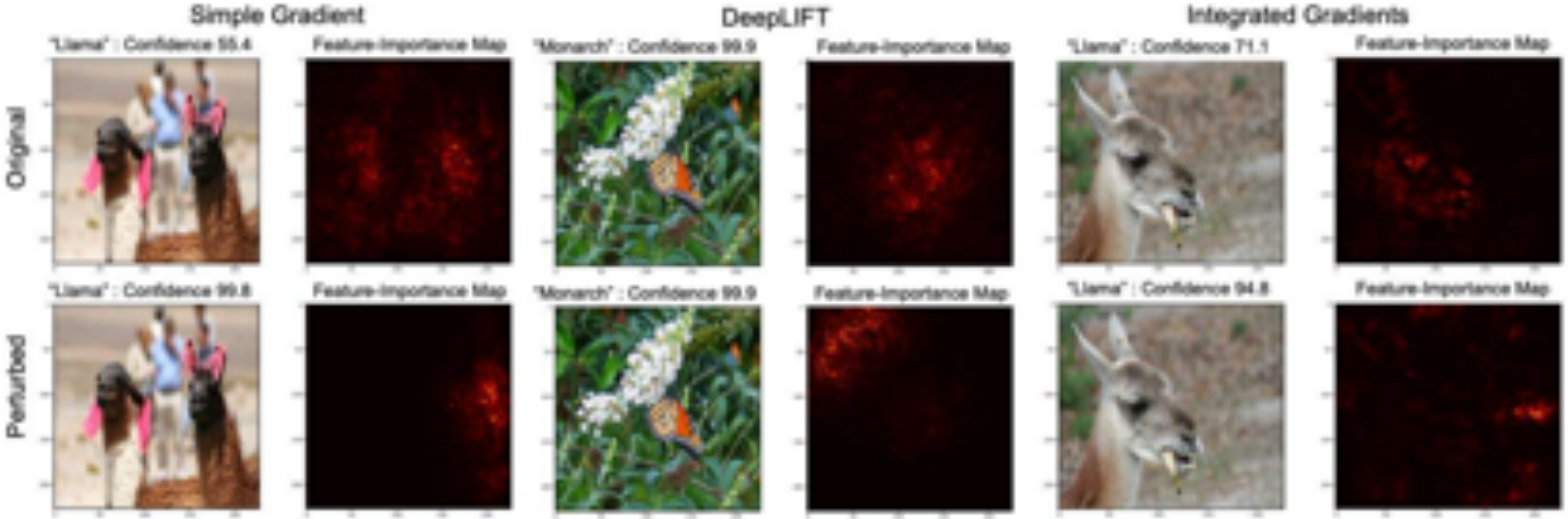
- Focused on gradient-based methods mostly.
- Sanity checks don’t tell if a method is good, just if it is invariant.
- Sole visual inspection can be deceiving.

What about other methods?



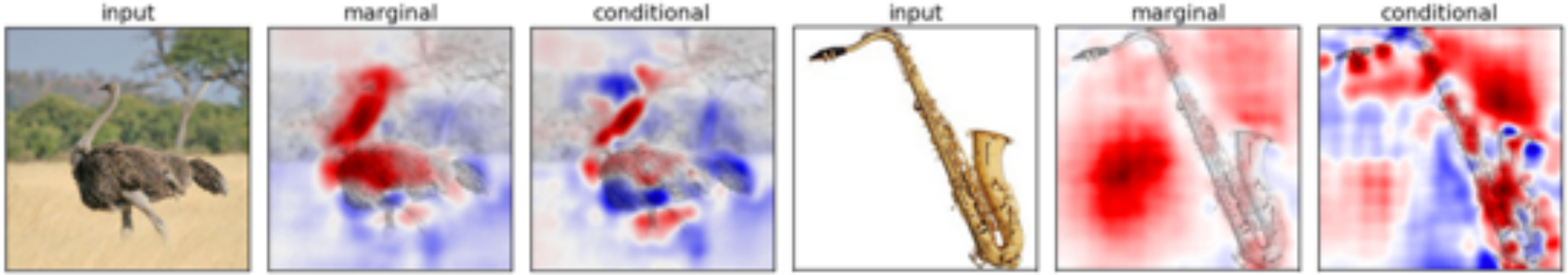
Attacks

‘Adversarial’ attack on explanations by Ghorbani et. al.



Visualizing Deep Neural Network Decisions: Prediction Difference Analysis

Marginal vs Conditional Sampling

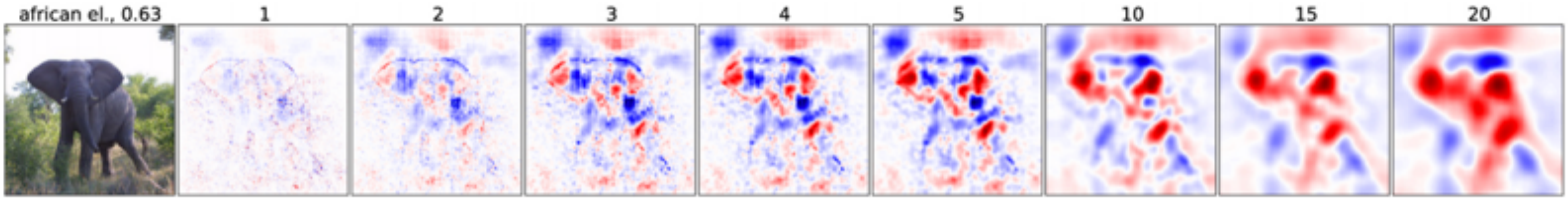


- Marginal Sampling → pixels that can be easily predicted using neighborhood are important

- Conditional Sampling → more specific and fine grained results

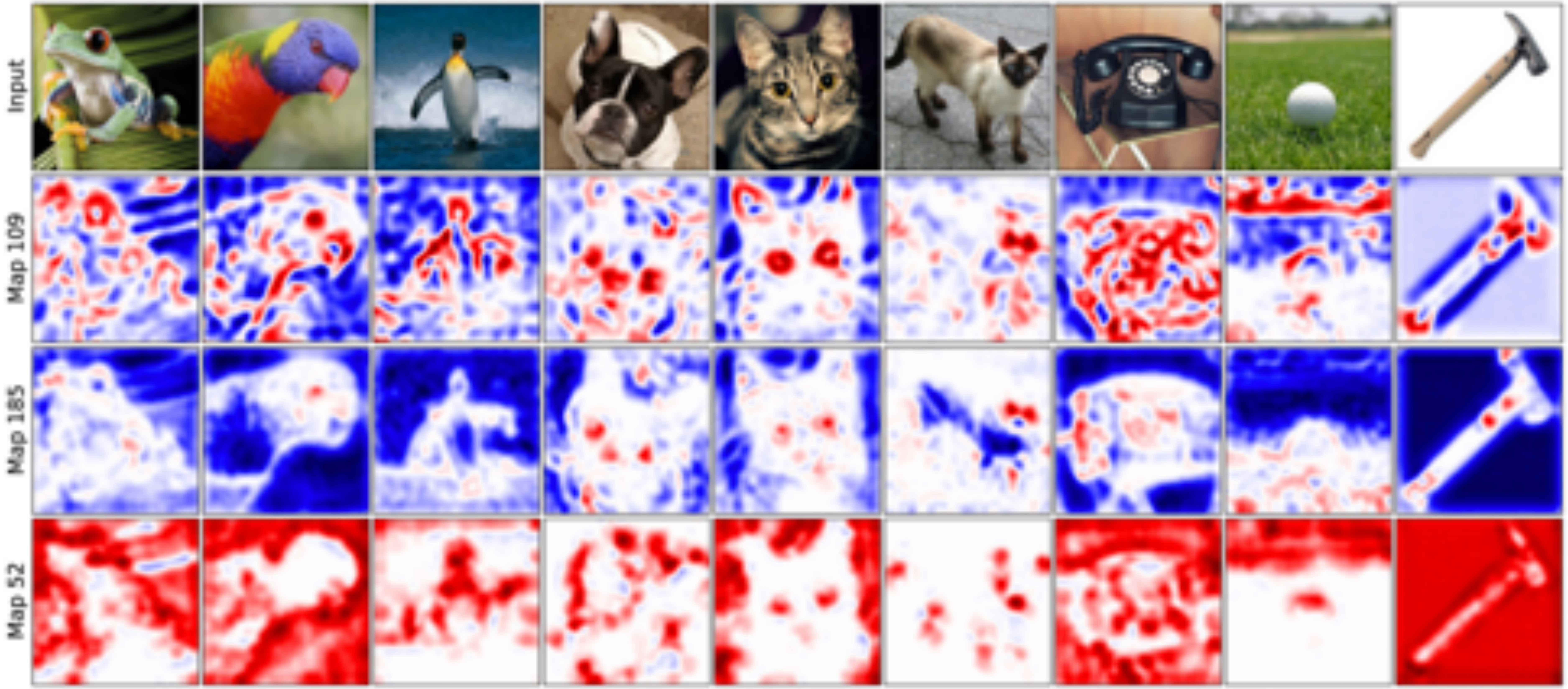
Multivariate Analysis : Window Sizes

- AlexNet, $l = k + 4$, varying k



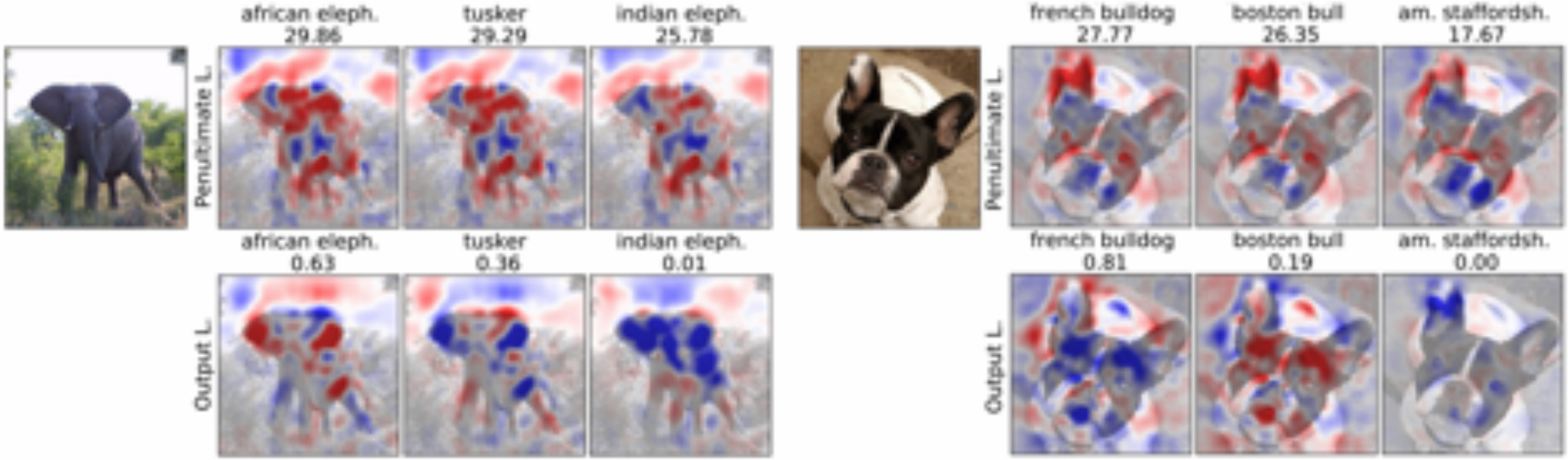
- Increasing window size → more easily interpretable, smooth until image gets blurry

Visualization of Hidden Layers



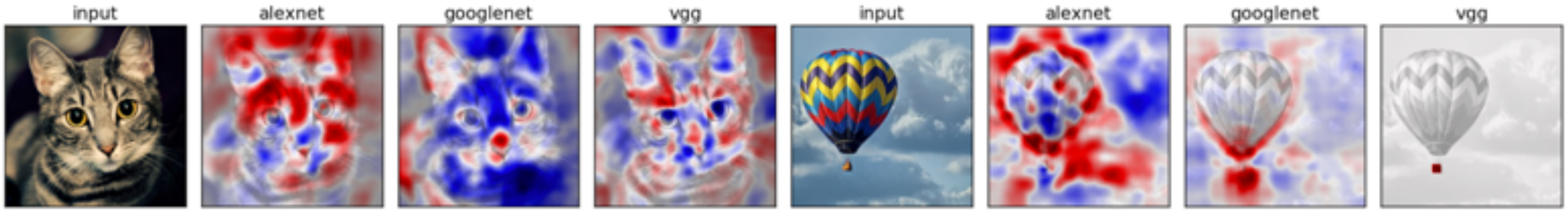
- Visualize 3 different feature maps react to multiple images
  - Middle of the network -- GoogLeNet

Penultimate vs Output Layers



- Visualizations in penultimate layer look similar if classes are similar
- In the final layer, values of nodes are all interdependen

Comparing Neural Architectures



- AlexNet is looking at more contextual info
  - E.g., sky in balloon image

- VGG: last image
  - Basket differentiates between balloon and parachute

Visualizing Model Behavior

Jorge Poco, @jpocom
Fundação Getulio Vargas
FGV EMAp

Questions?