Tutorials

Interpretable Deep Learning: Towards Understanding & Explaining DNNs

Part 1: Introduction

Wojciech Samek, Grégoire Montavon, Klaus-Robert Müller

From ML Successes to Applications

Deep Net outperforms humans in image classification

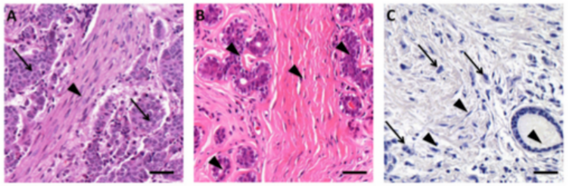
IMAGENET

AlphaGo beats Go human champ

Visual Reasoning

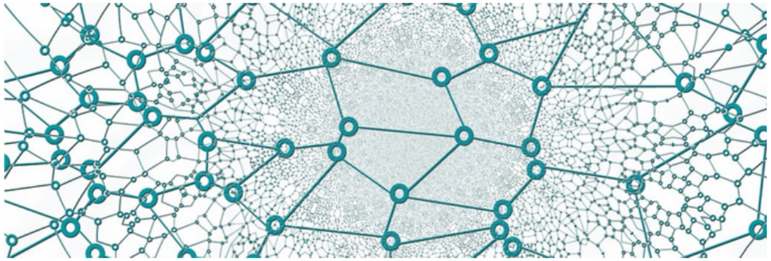
What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

Medical Diagnosis



Autonomous Driving

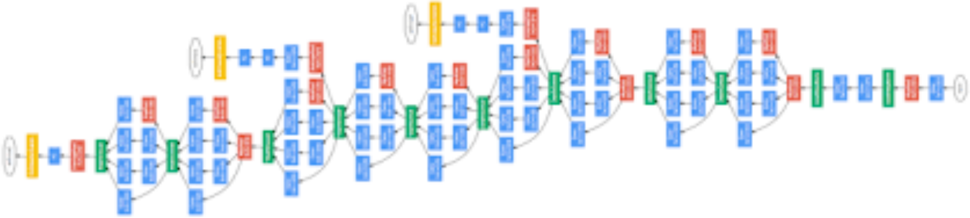
Networks (smart grids, etc.)



Black Box Models

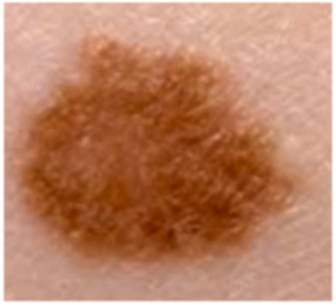
Huge volumes of data

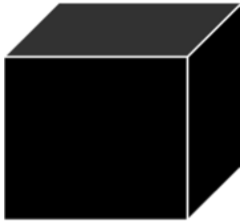
Computing power



Deep Neural Network

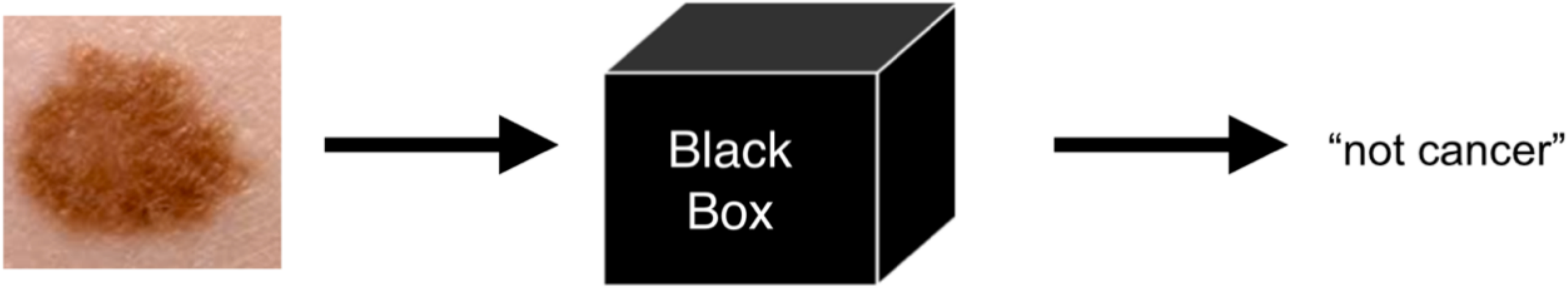
Solve task





Information (implicit)

Black Box Models



Is minimizing the error a guarantee for the model to work well in practice?

Why interpretability ?

Why Interpretability ?

We need interpretability in order to:

verify system

legal aspects

understand weaknesses

learn new things from data

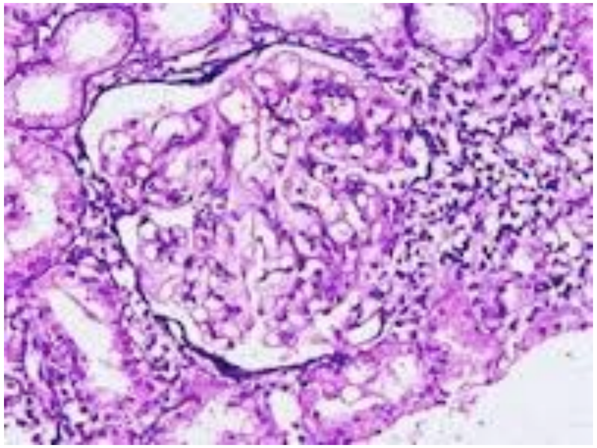
Why Interpretability ?

1) Verify that classifier works as expected

Wrong decisions can be costly
and dangerous

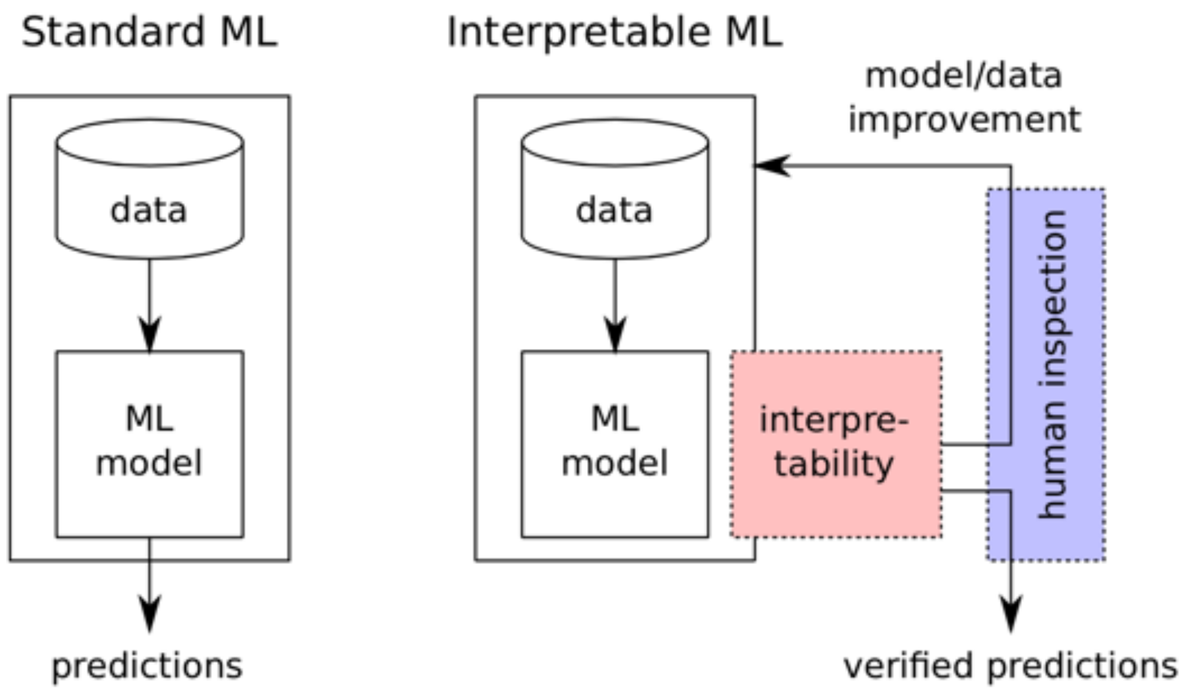
“Autonomous car crashes, because it wrongly recognizes …”

“AI medical diagnosis system misclassifies patient’s disease …”



Why Interpretability ?

2) Understand weaknesses & improve classifier



Generalization error

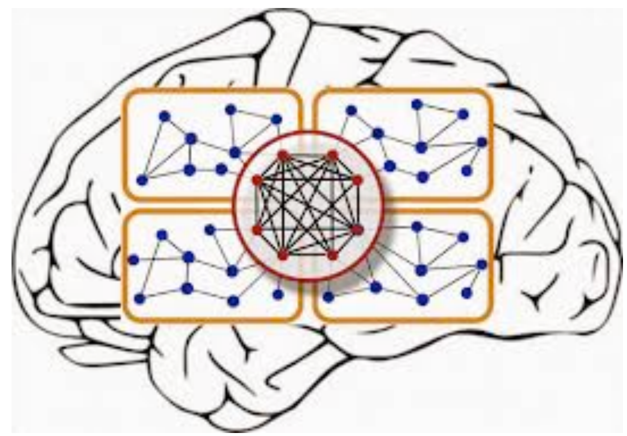
Generalization error + human experience

Why Interpretability ?

3) Learn new things from the learning machine

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

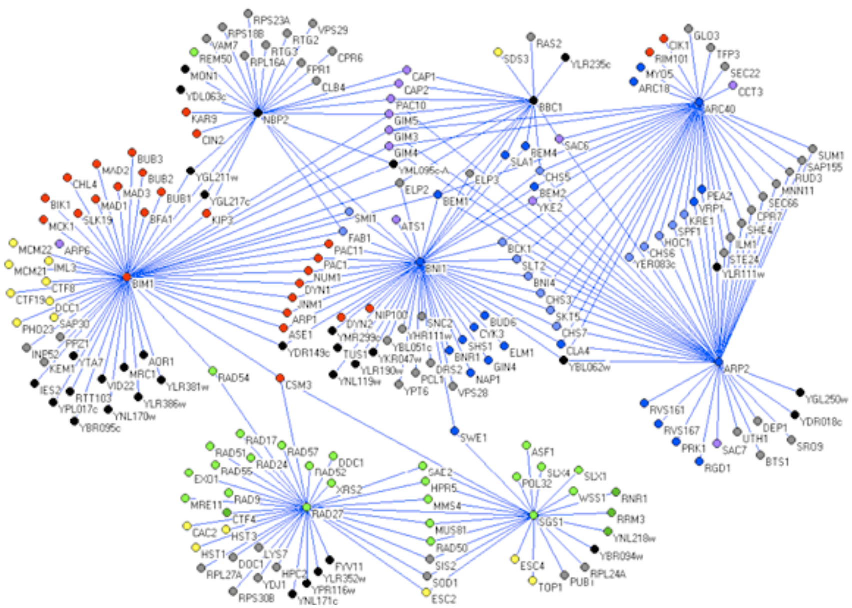
Old promise:
"Learn about the human brain."

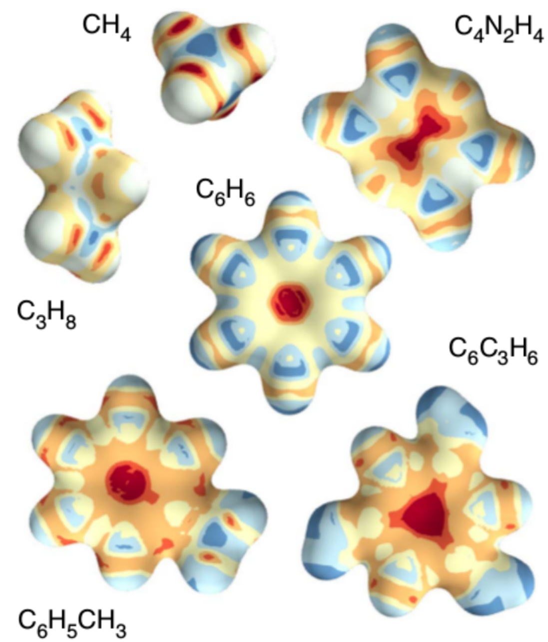


Why Interpretability ?

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.

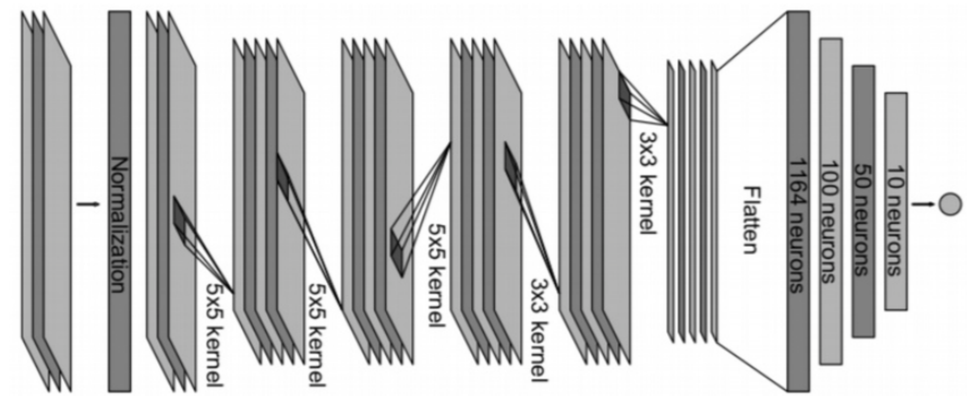
"With interpretability we can ensure that ML models work in compliance to proposed legislation."

Example: Autonomous Driving

Bojarski et al. 2017 “Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car”

PilotNet

Input:

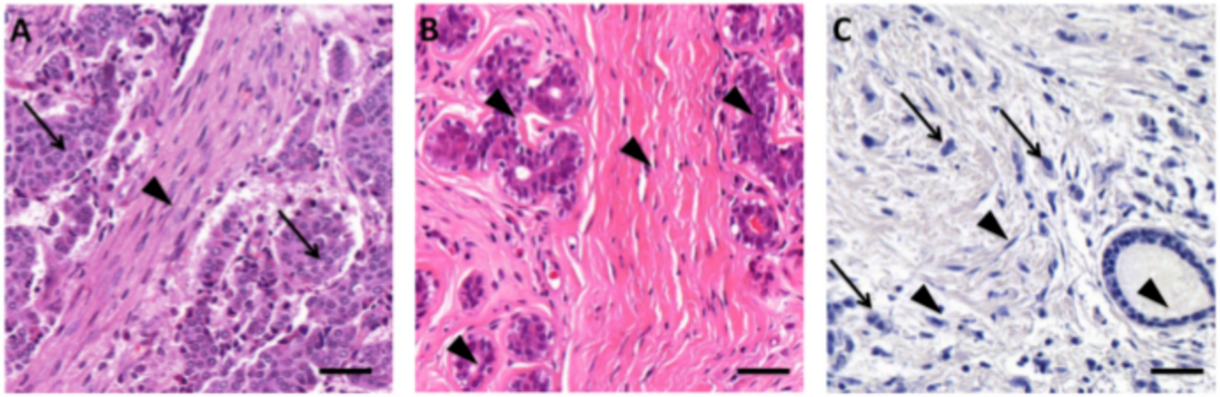


Decision

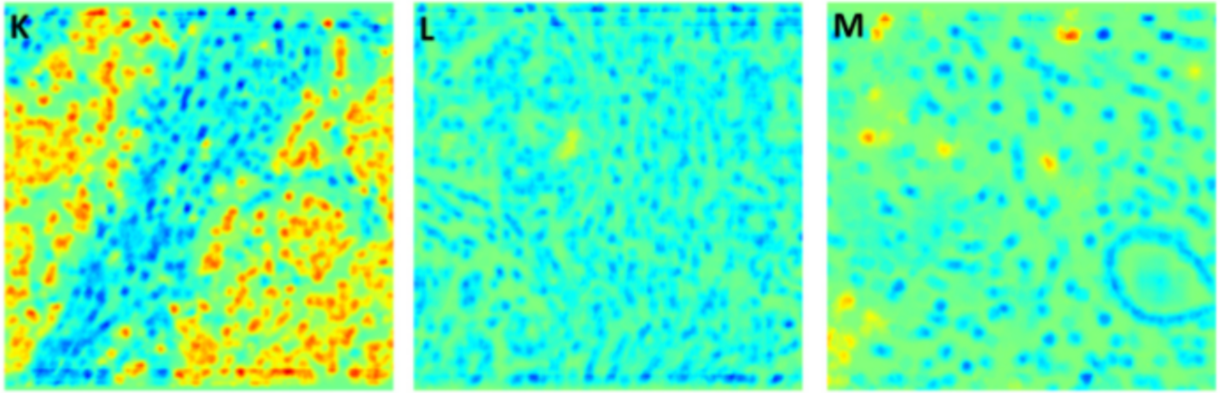
Explanation:

Example: Medical Diagnosis

Binder et al. 2018 “Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles”



A: Invasive breast cancer, H&E stain; B: Normal mammary glands and fibrous tissue, H&E stain; C: Diffuse carcinoma infiltrate in fibrous tissue, Hematoxylin stain



Example: Quantum Chemistry

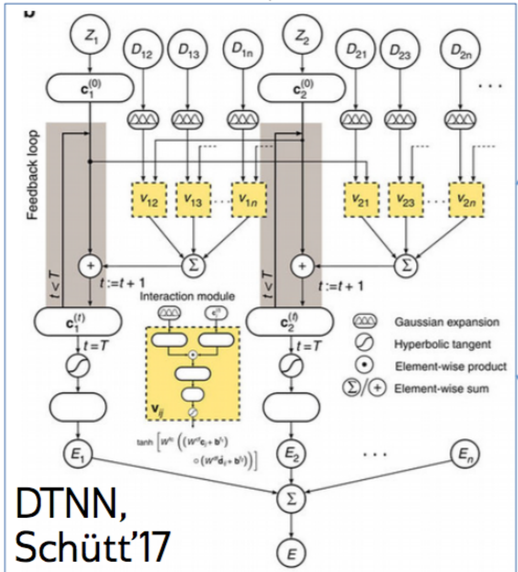
Schütt et al. 2017: Quantum-Chemical Insights from Deep Tensor Neural Networks

molecular structure (e.g. atoms positions)

DFT calculation of the stationary Schrödinger Equation

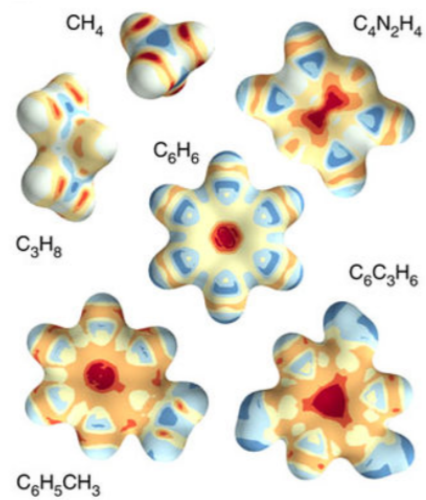
$$H\Phi = E\Phi$$

PBE0, Pedrew’86



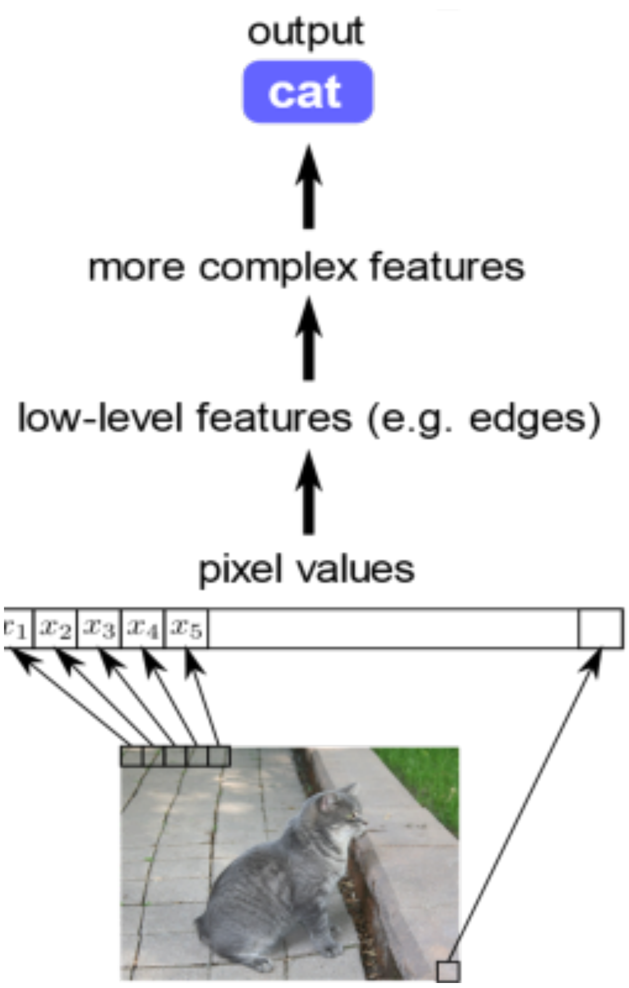
DTNN, Schütt’17

interpretable insight

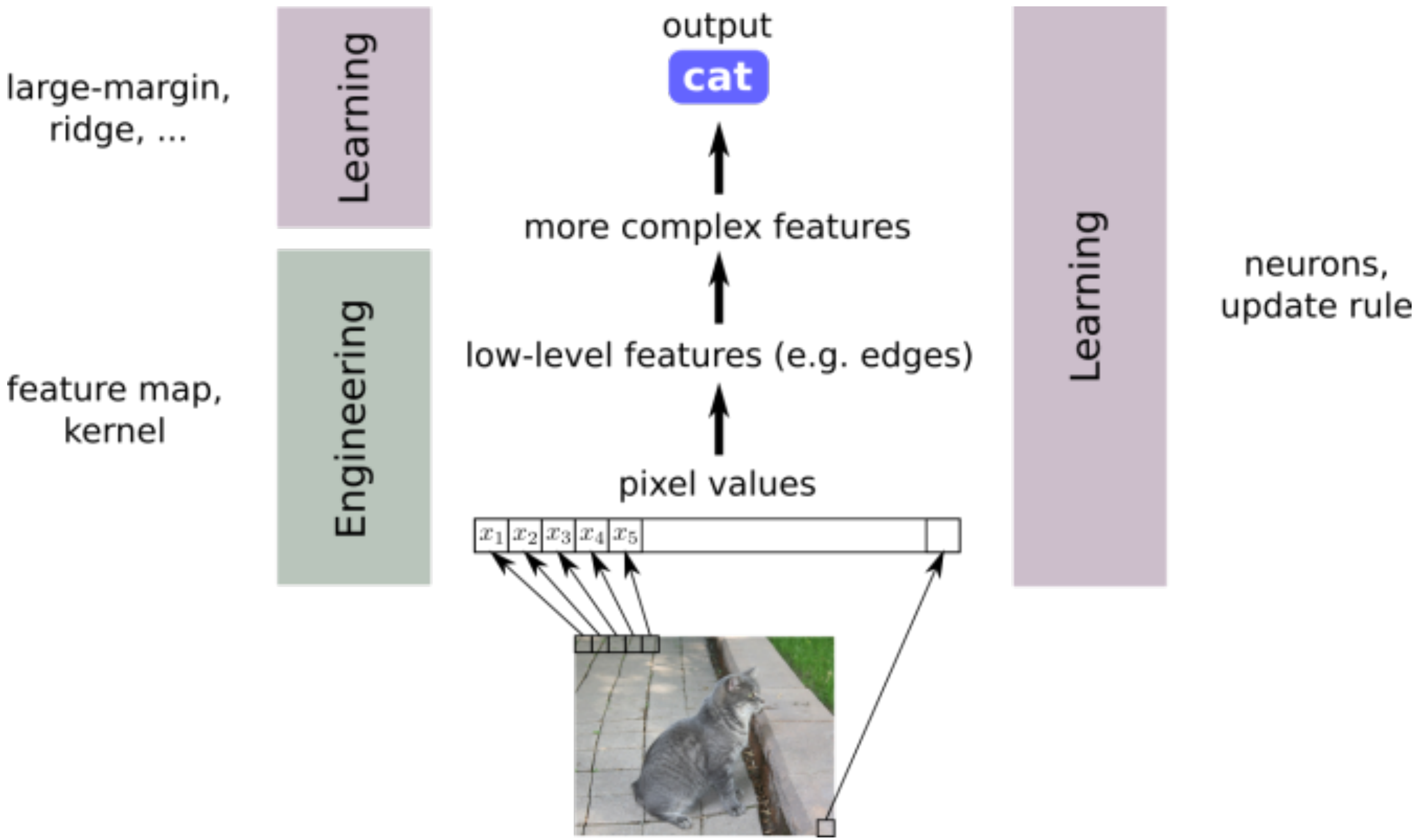


molecular electronic properties (e.g. atomization energy)

From Input to Abstractions

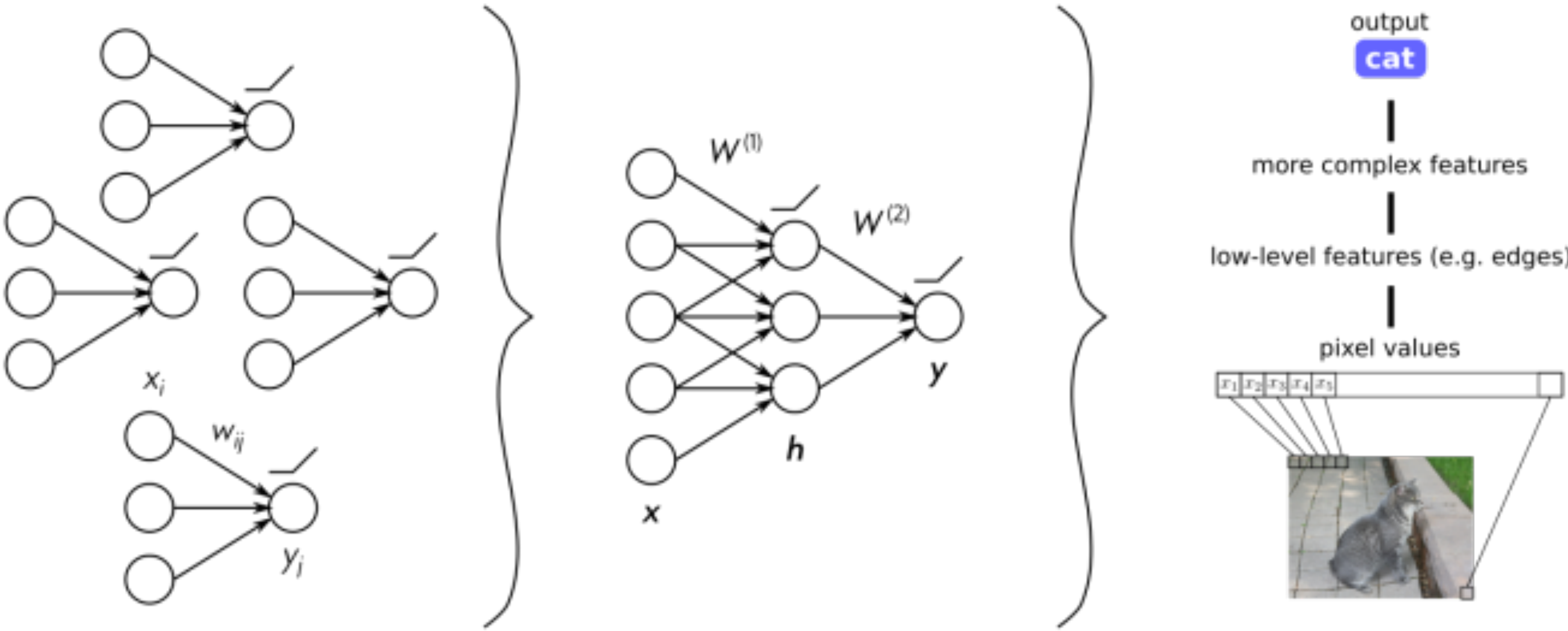


Learning Hierarchical Representations



Learning Hierarchical Representations

- Multiple neurons with similar structure, but with different weight parameters.
- Compose them into a deep layered architecture.

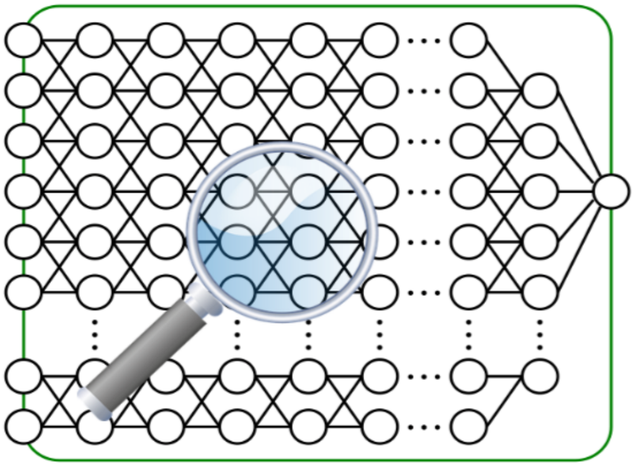


$$y_j = \max(0, \textstyle\sum_i x_i w_{ij})$$

$$\boldsymbol{h} = \max(0, W^{(1)}\boldsymbol{x})$$

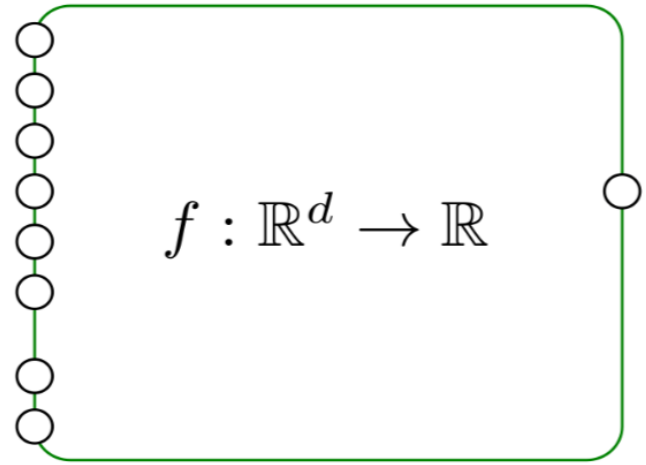
Dimensions of Interpretability

mechanistic understanding



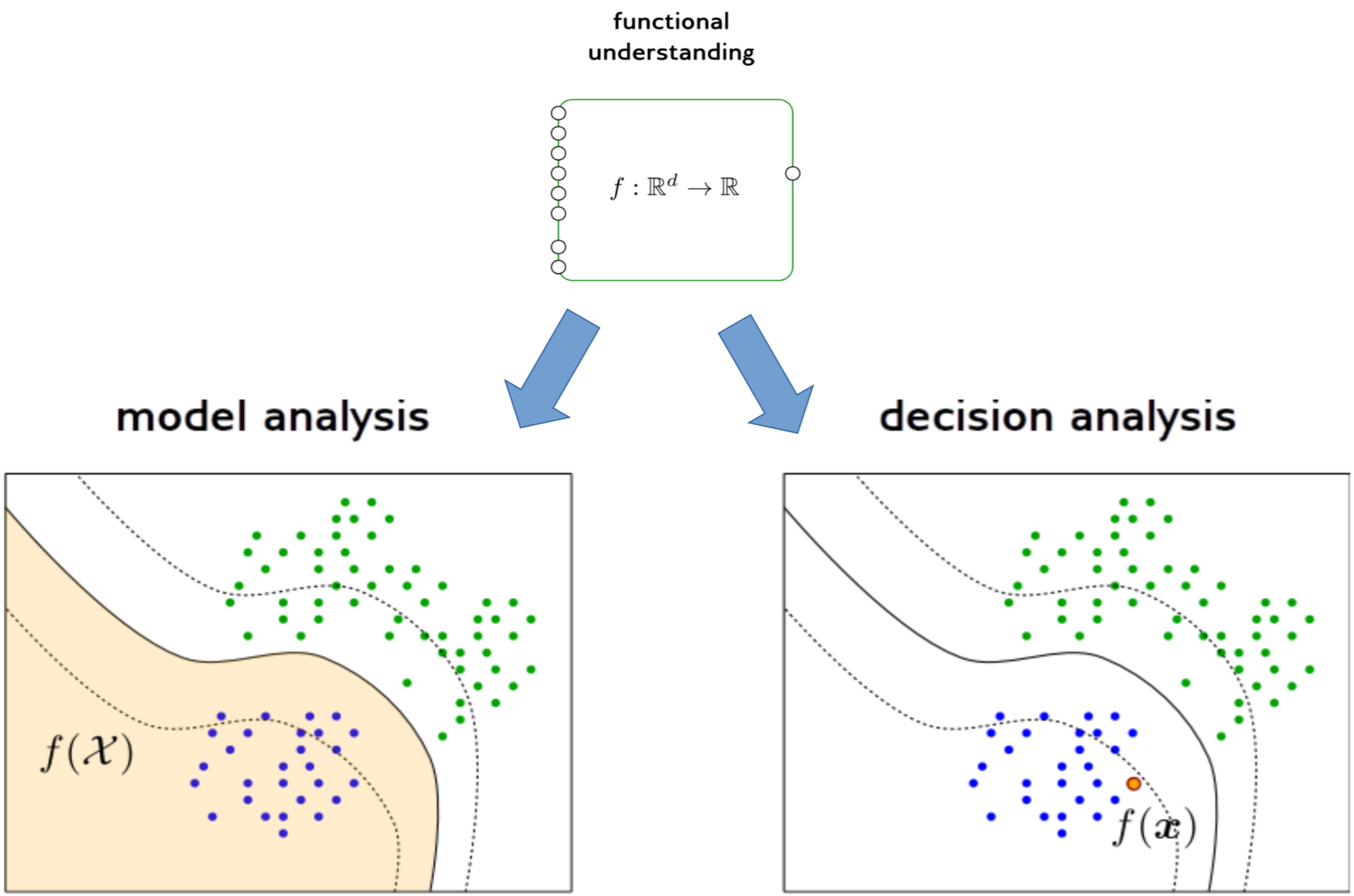
Understanding what mechanism the network uses to solve a problem or implement a function.

functional understanding



Understanding how the network relates the input to the output variables.

Dimensions of Interpretability



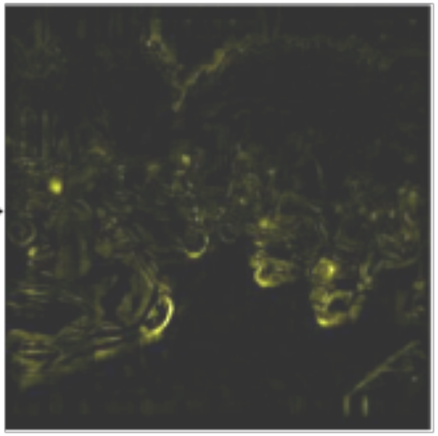
Dimensions of Interpretability

Model Analysis
"what does something predicted as a scooter typically look like."

model's prototypical scooter

Decision Analysis
"why a given image is classified as a scooter"

some image with scooter



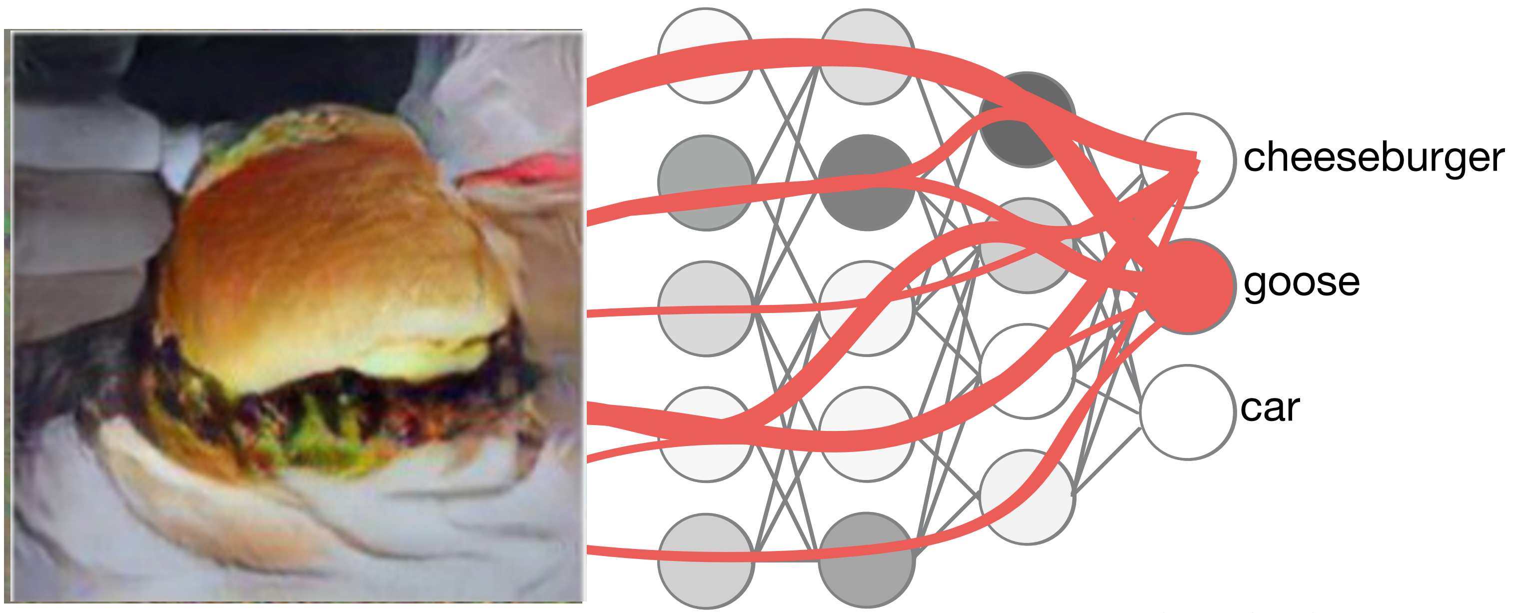
why it is classified as scooter

Model analysis

Interpreting the Model

Activation Maximization

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

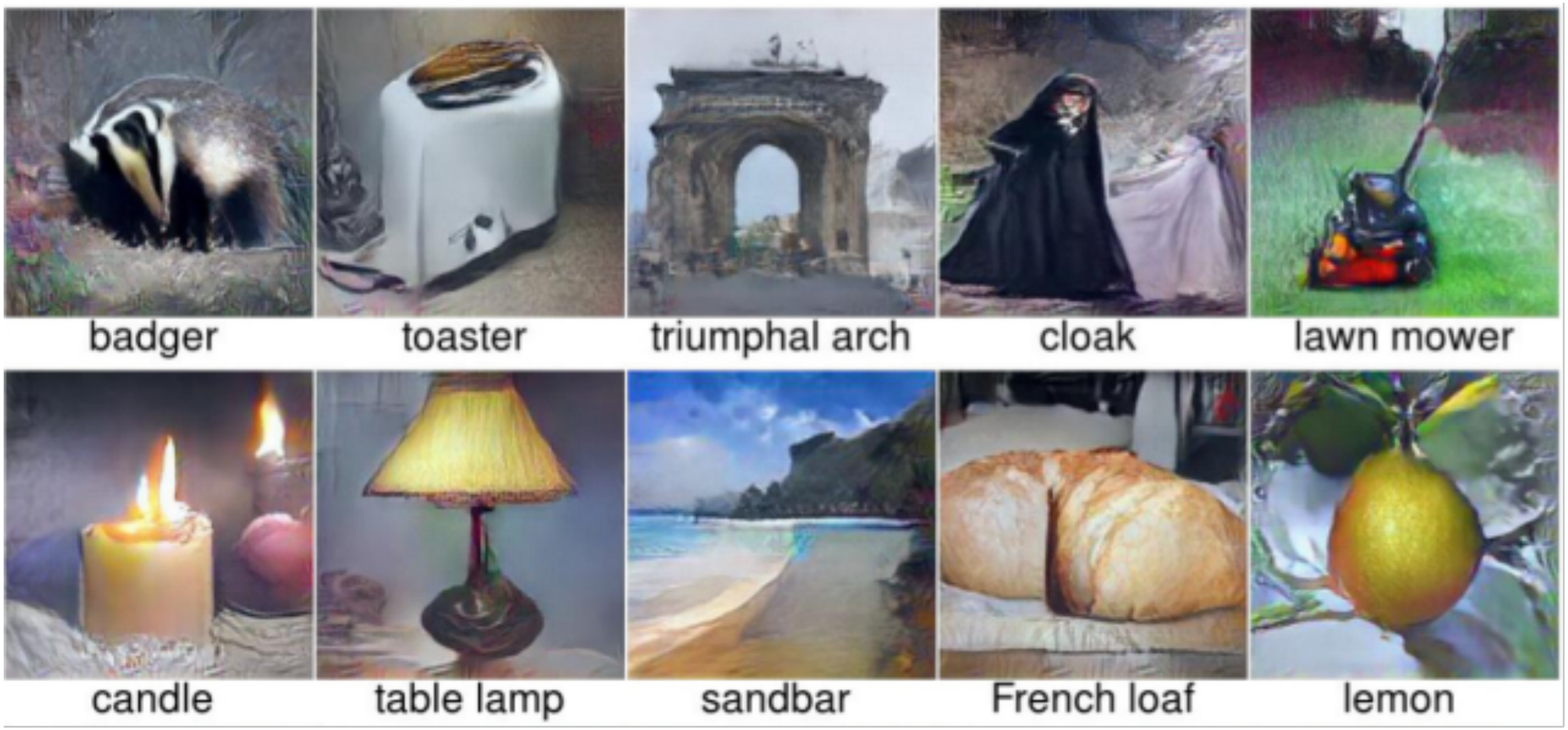


simple regularizer (Simonyan et al 2013)

complex regularizer (Nguyen et al 2016)

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

Interpreting the Model



Nguyen'16: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.

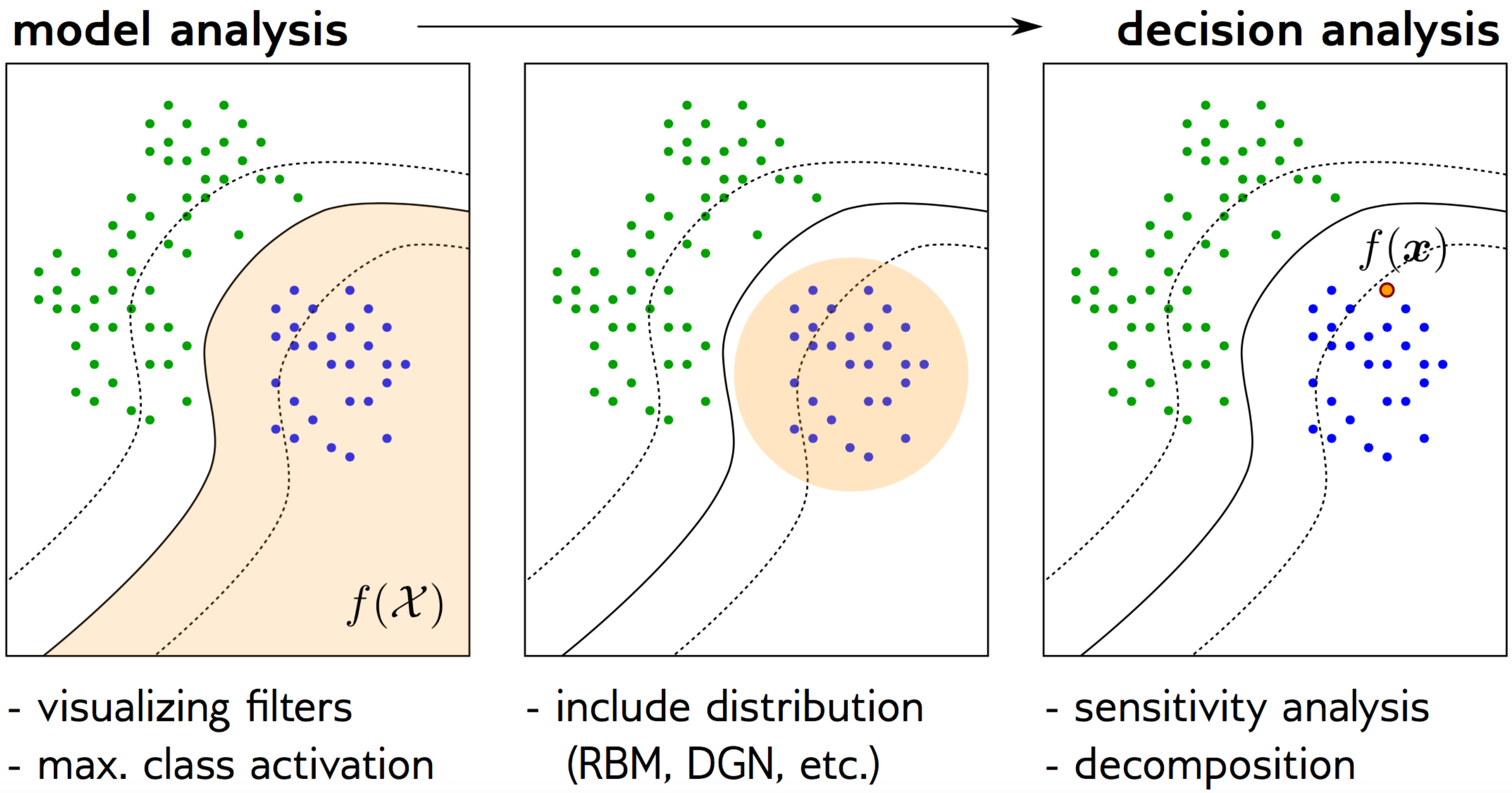
Limitations of Global Interpretations

Question: Below are some images of motorbikes. What would be the best prototype to interpret the class "motorbike"?

Observations:

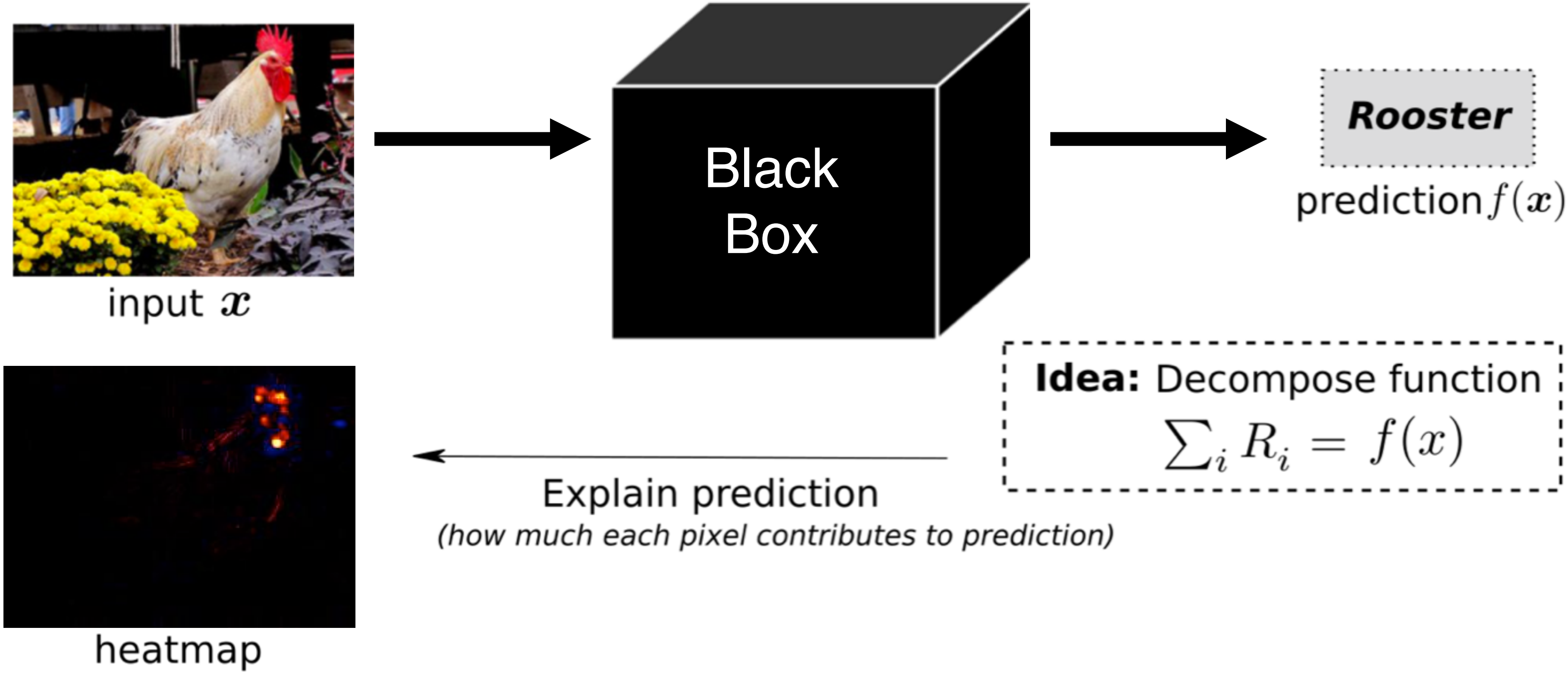
- Summarizing a concept or category like "motorbike" into a single image can be difficult (e.g. different views or colors).
- A good interpretation would grow as large as the diversity of the concept to interpret.

Making Deep Neural Nets Transparent



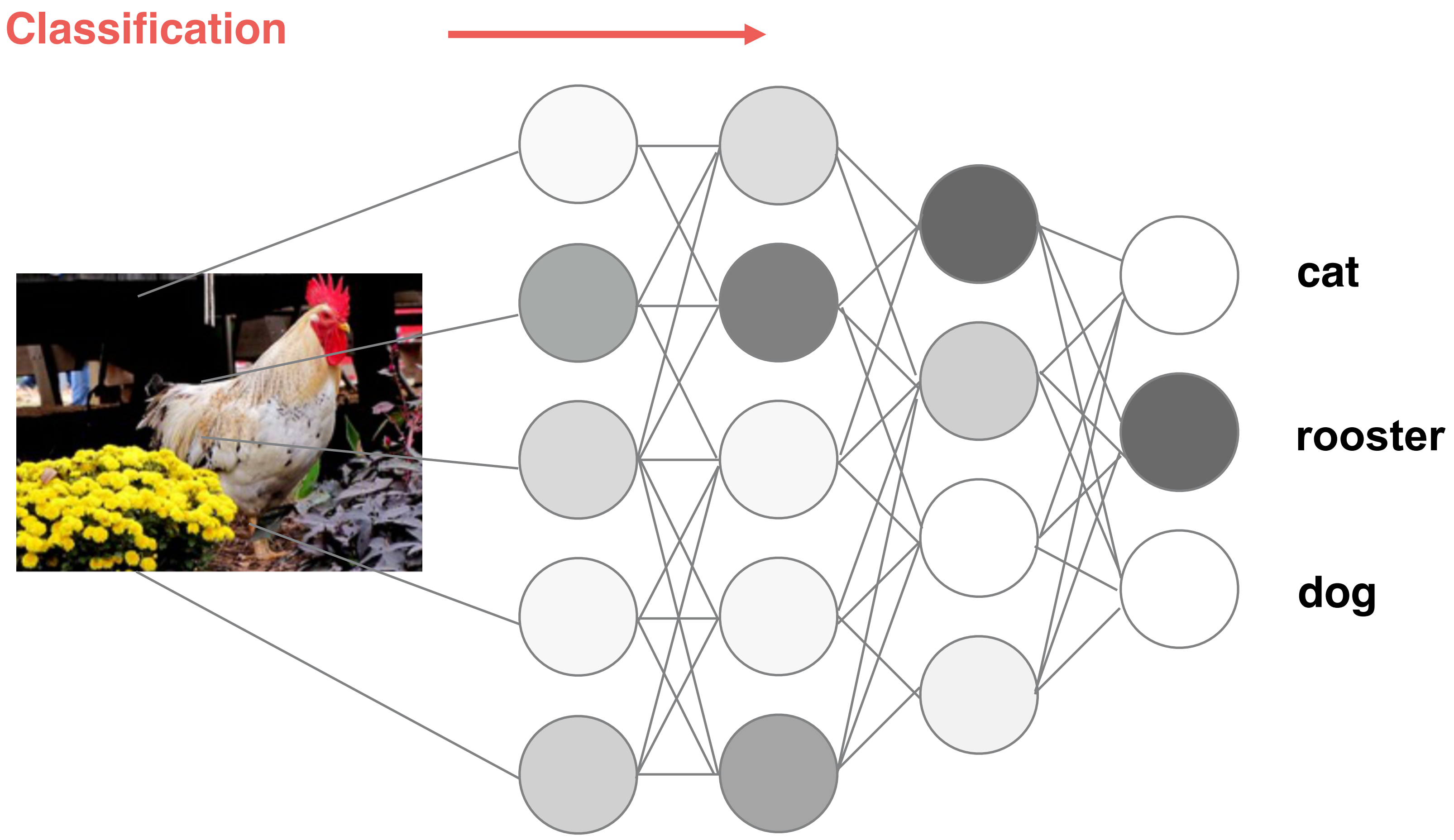
Decision analysis

Decision Analysis: LRP

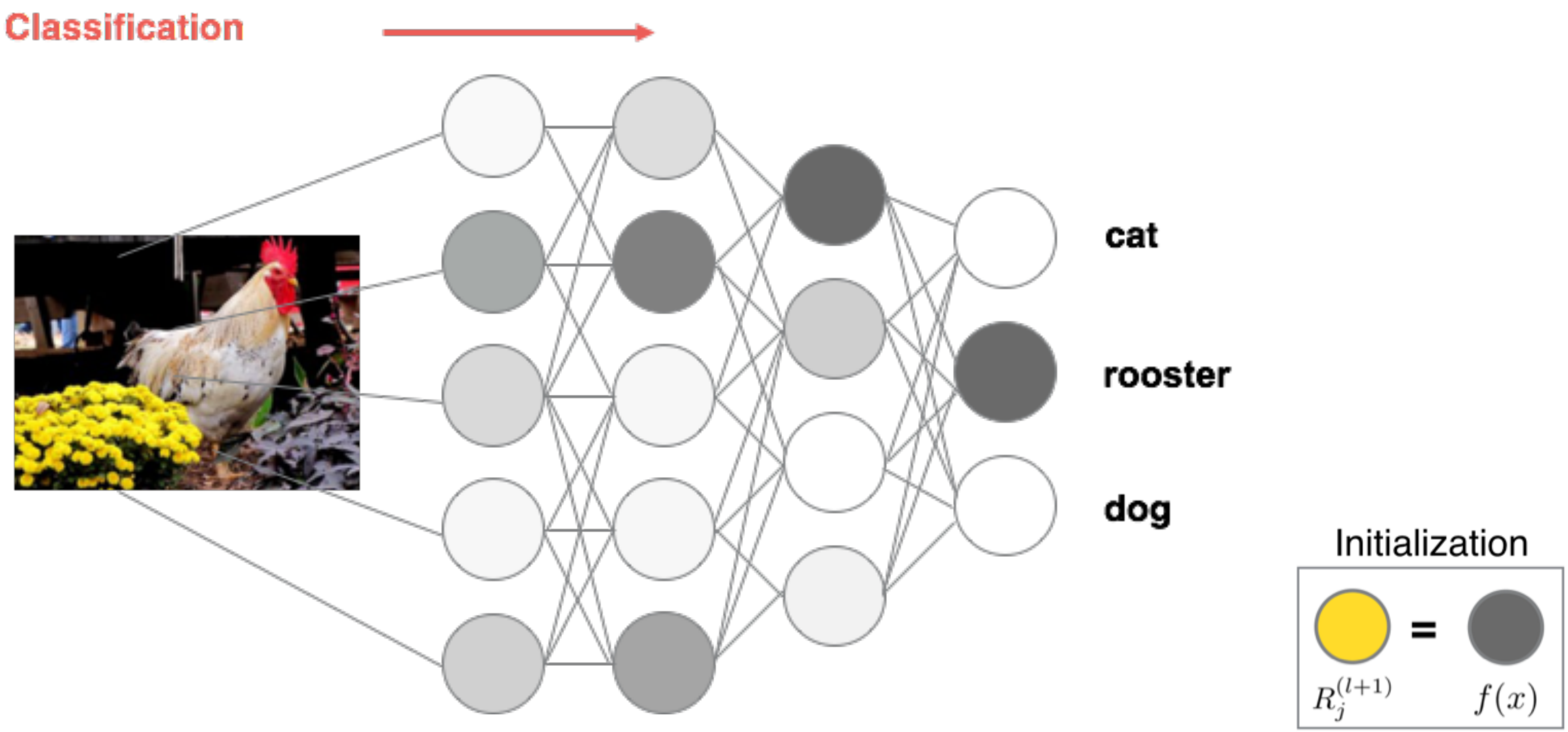


Layer-wise Relevance Propagation (LRP)
(Bach et al., PLOS ONE, 2015)

Decision Analysis: LRP



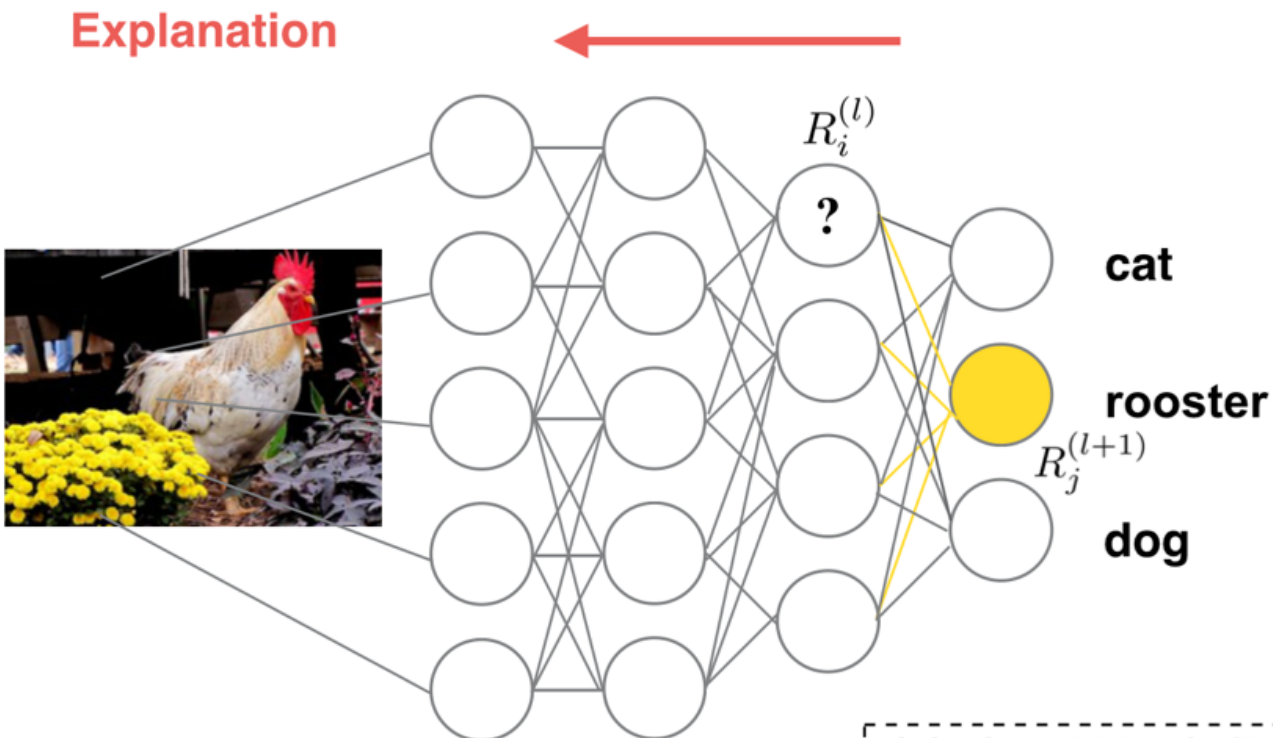
Decision Analysis: LRP



What makes this image a “rooster image” ?

Idea: Redistribute the evidence for class rooster back to image space.

Decision Analysis: LRP



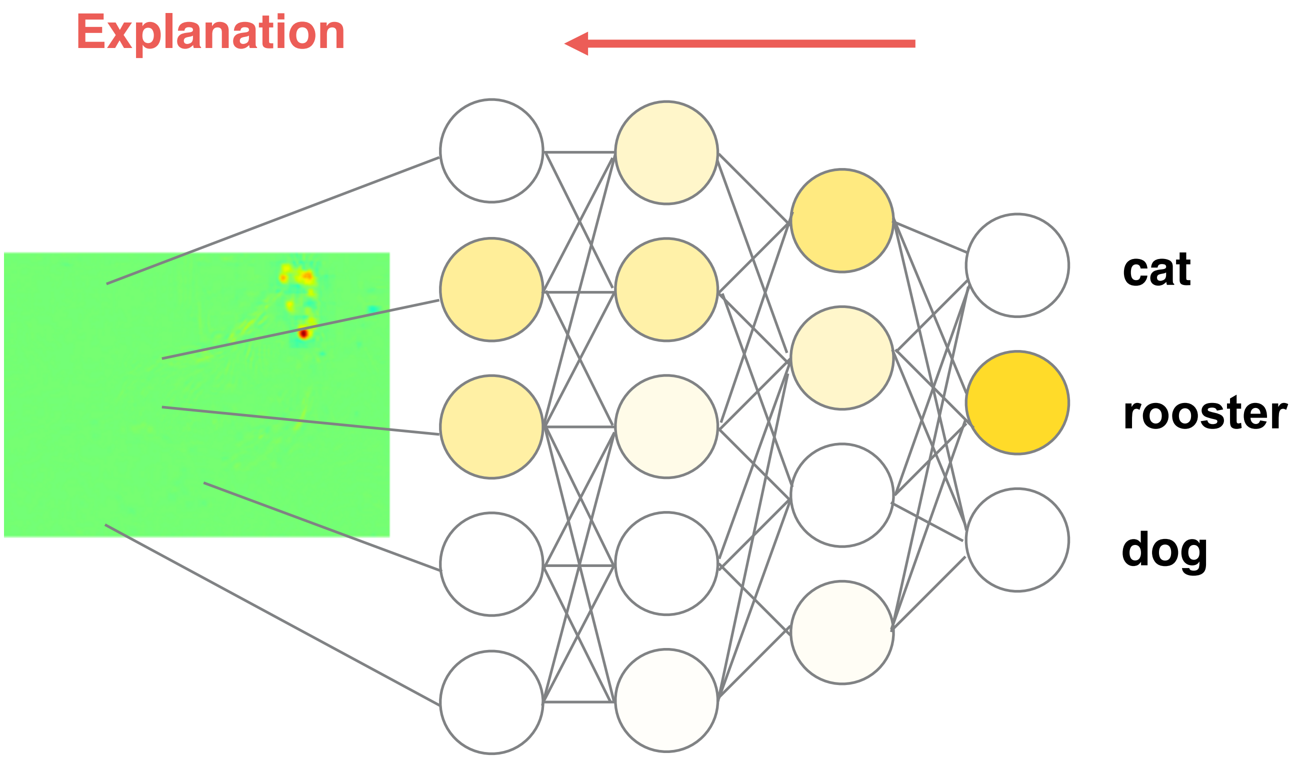
alpha-beta LRP rule (Bach et al. 2015)

$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{(x_i \cdot w_{ij})^+}{\sum_{i'}(x_{i'} \cdot w_{i'j})^+} + \beta \cdot \frac{(x_i \cdot w_{ij})^-}{\sum_{i'}(x_{i'} \cdot w_{i'j})^-}\right) R_j^{(l+1)}$$

where $\alpha + \beta = 1$

Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)

Decision Analysis: LRP

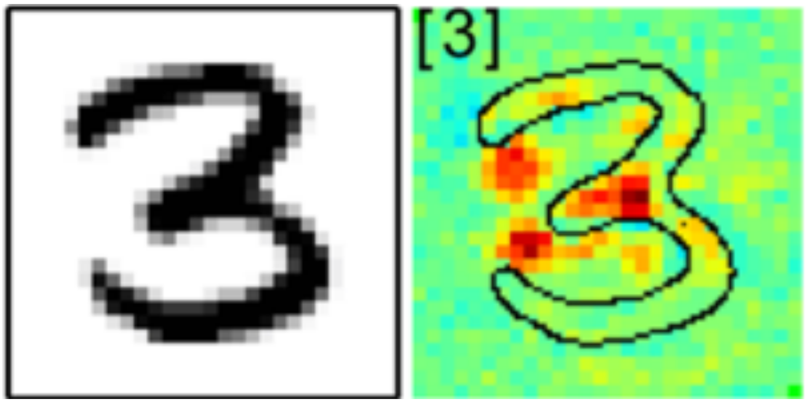


Layer-wise relevance conservation

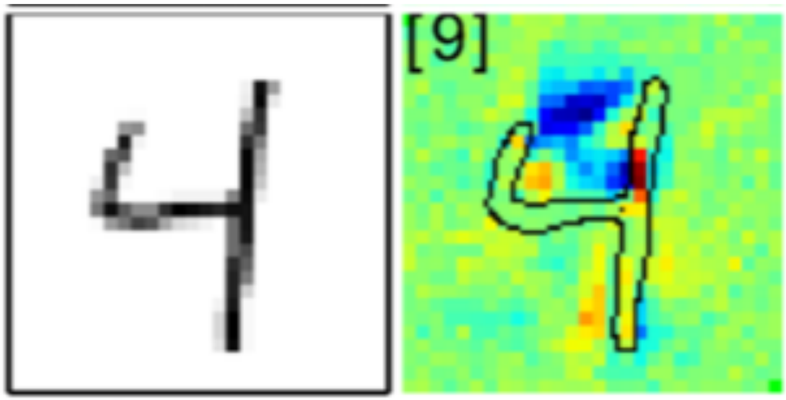
$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \ldots = f(x)$$

Decision Analysis: LRP

Heatmap of prediction “3”



Heatmap of prediction “9”



Other Explanation Methods