Explaining Machine Learning Decisions

Grégoire Montavon, TU Berlin

Joint work with: Wojciech Samek, Klaus-Robert Müller, Sebastian Lapuschkin, Alexander Binder

18/09/2018 Intl. Workshop ML & AI, Telecom ParisTech

From ML Successes to Applications

Deep Net outperforms humans in image classification

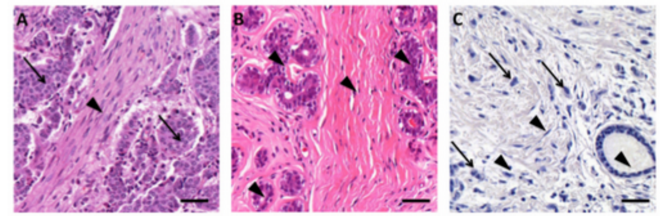
IMAGENET

AlphaGo beats Go human champ

Visual Reasoning

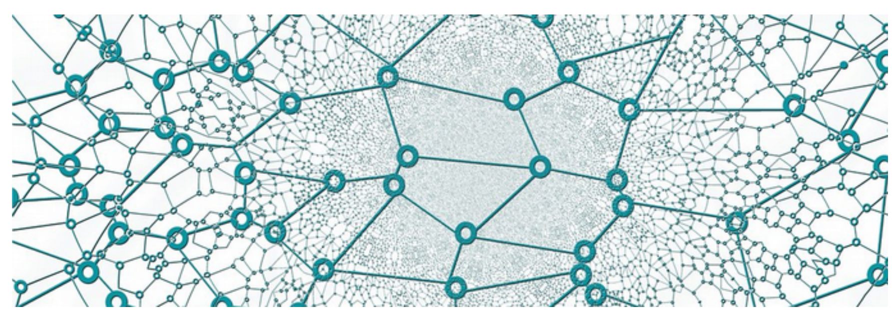
What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

Medical Diagnosis



Autonomous Driving

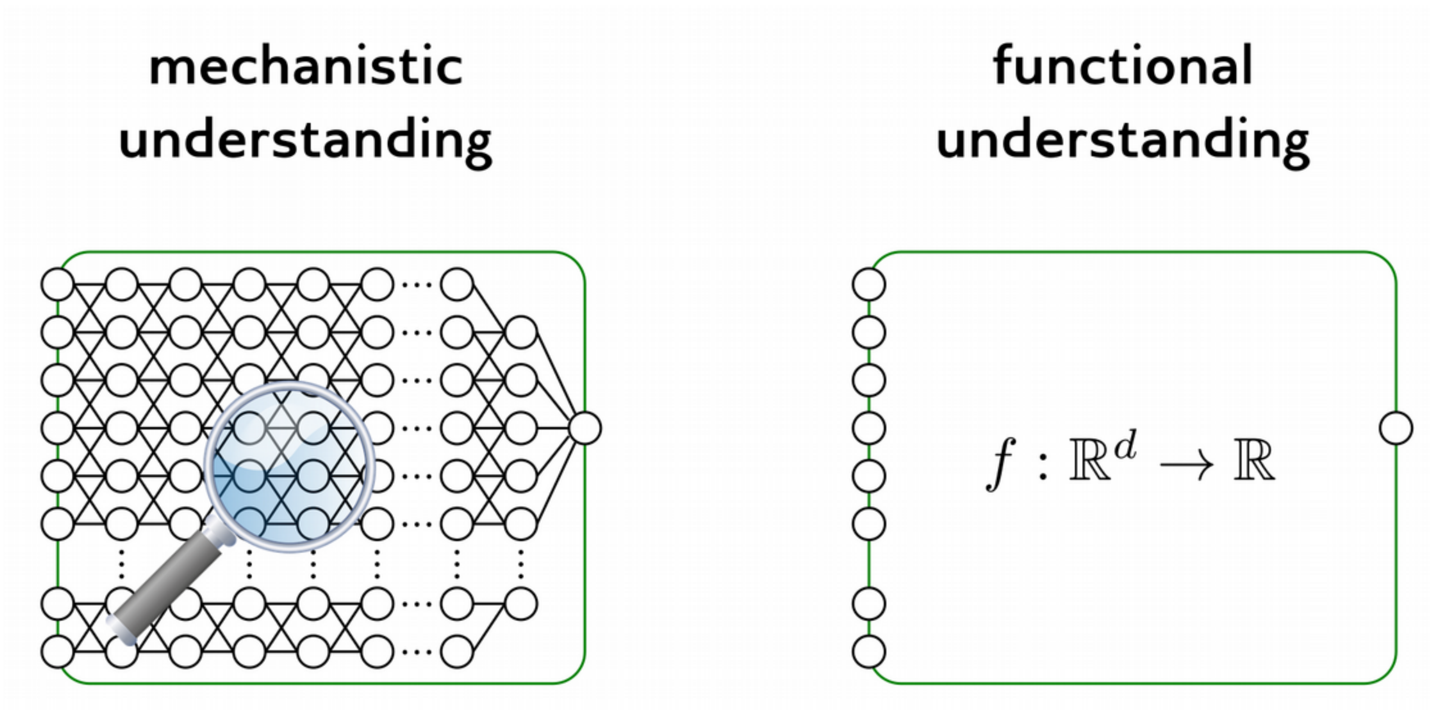
Networks (smart grids, etc.)



Can we interpret
what a ML model has
learned?

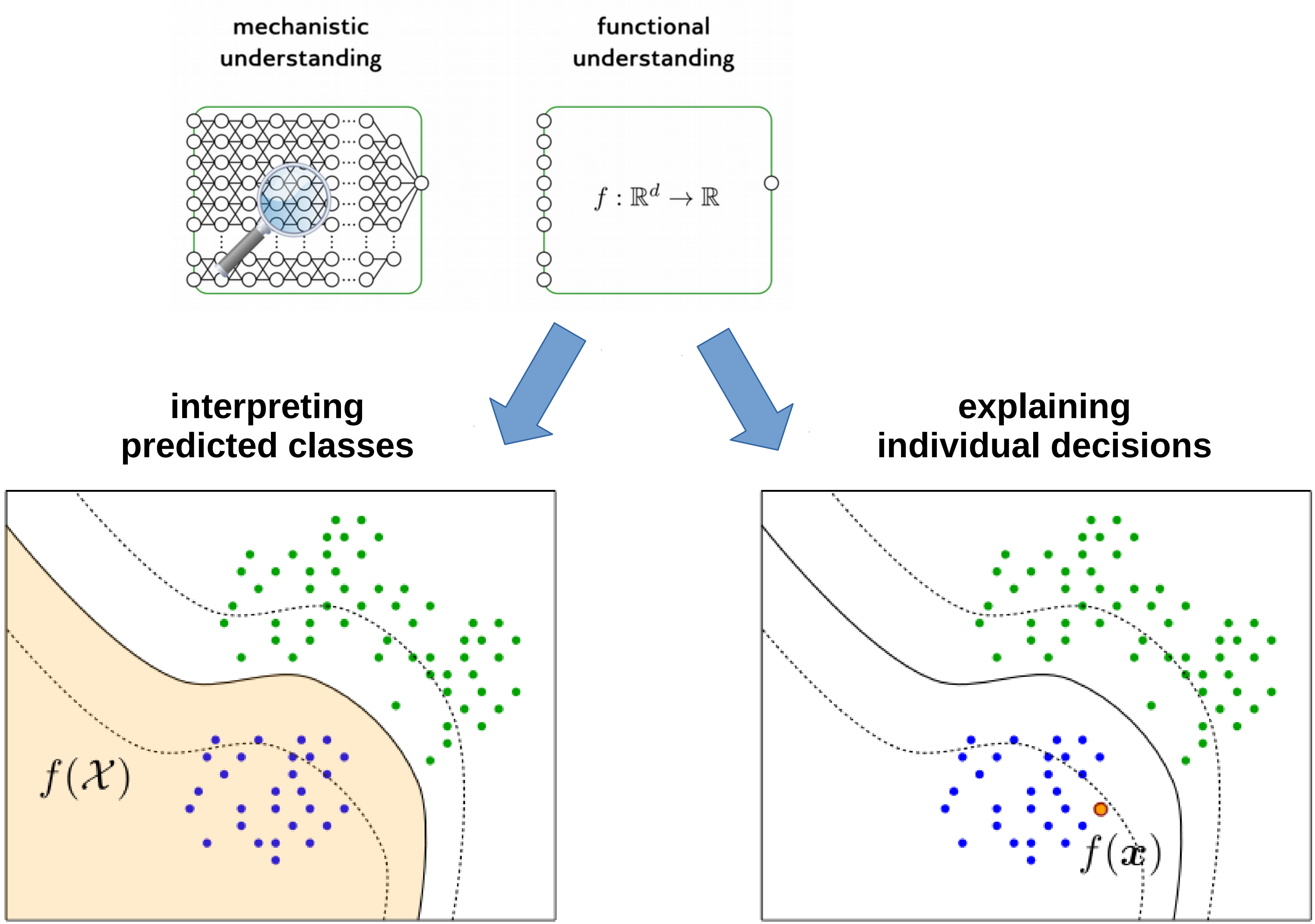
First, we need to define what we want from interpretable ML.

Understanding Deep Nets: Two Views



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

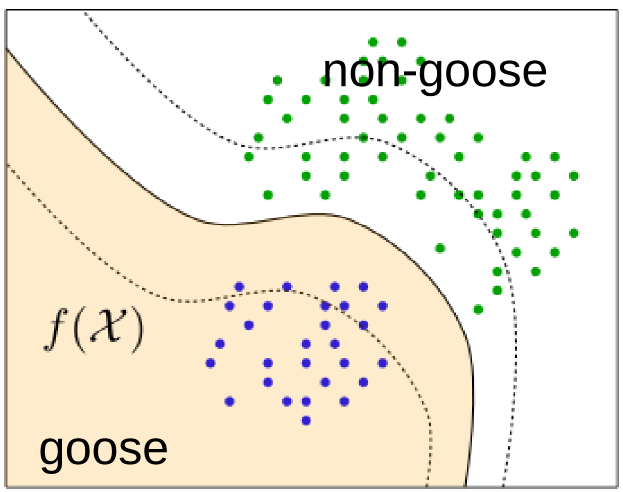
Understanding how the network relates the input to the output variables.



mechanistic understanding
functional understanding
$f : \mathbb{R}^d \to \mathbb{R}$
interpreting predicted classes
explaining individual decisions
$f(\mathcal{X})$
$f(\boldsymbol{x})$

Interpreting Predicted Classes

Example: *"How does a goose typically look like according to the neural network?"*

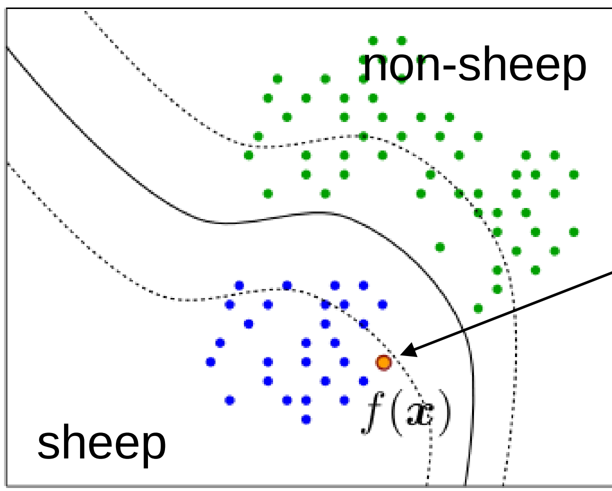


$$\arg\max_x f(x) + \text{reg.}$$

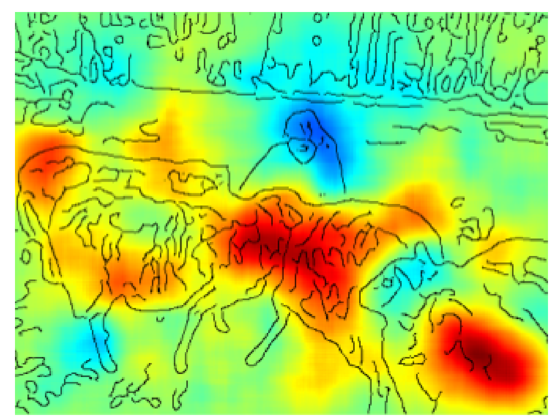
Image from **Symonian'13**

Explaining Individual Decisions

Example: *"Why is a given image classified as a sheep?"*



$$\boldsymbol{R}(\boldsymbol{x}, f)$$



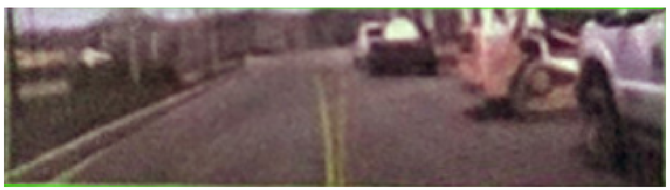
Images from **Lapuschkin'16**

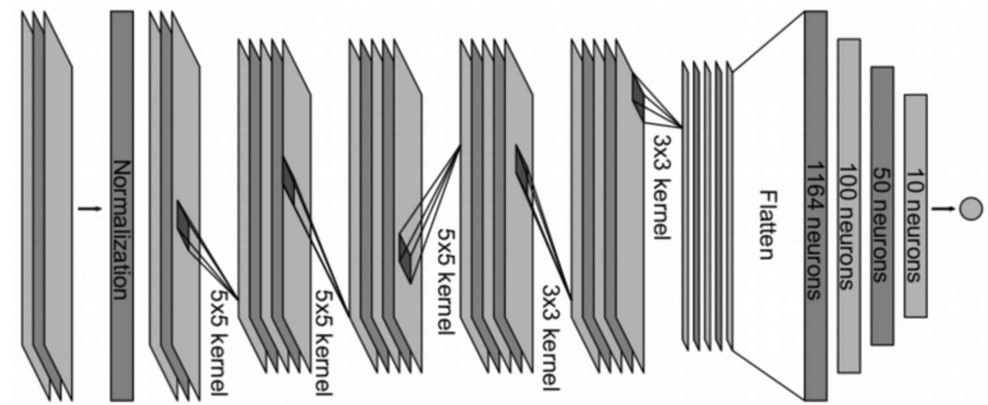
Example: Autonomous Driving [Bojarski'17]

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

PilotNet

Input:





Decision

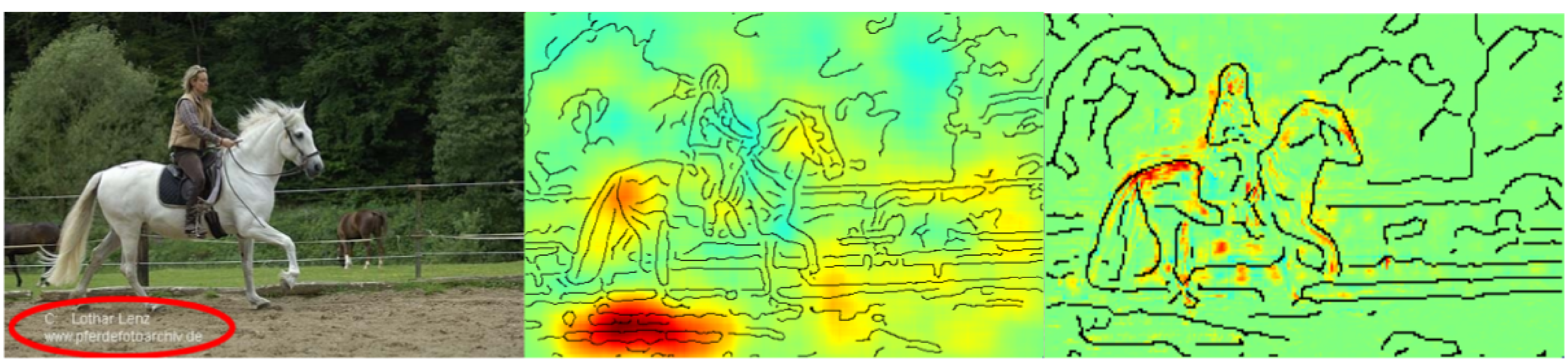
Explanation:

Example: Pascal VOC Classification [Lapuschkin'16]

Comparing Performance on Pascal VOC 2007
(Fisher Vector Classifier vs. DeepNet pretrained on ImageNet)

	aeroplane	**bicycle**	**bird**	**boat**	**bottle**	**bus**	**car**
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	**chair**	**cow**	**diningtable**	**dog**	**horse**	**motorbike**
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

Fisher classifier (pretrained) deep net



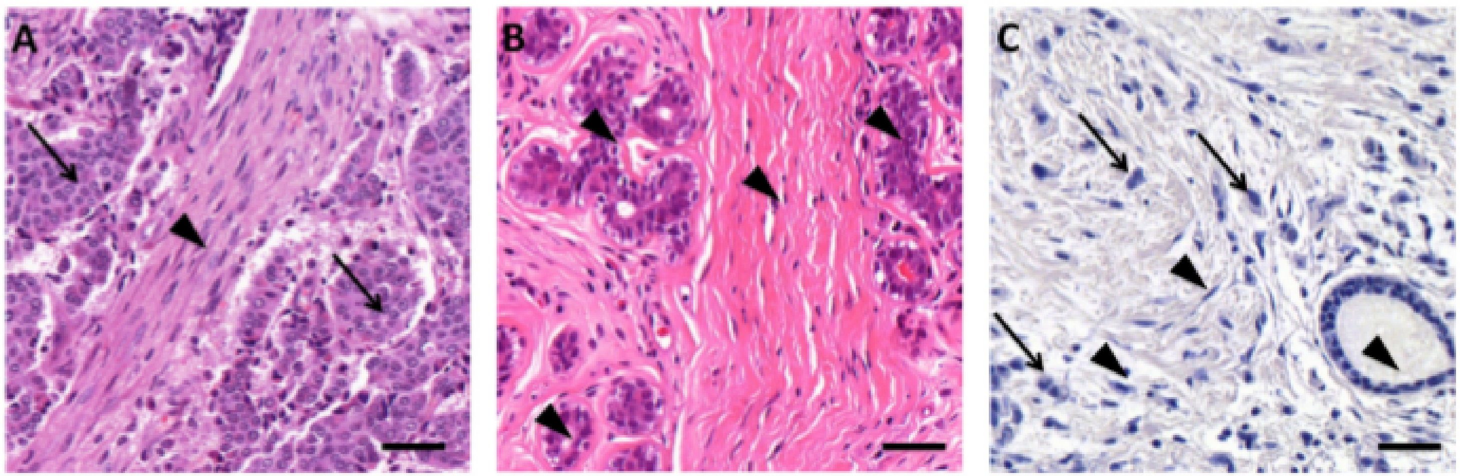
Lapuschkin et al. 2016. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks

Example: Pascal VOC Classification [Lapuschkin'16]

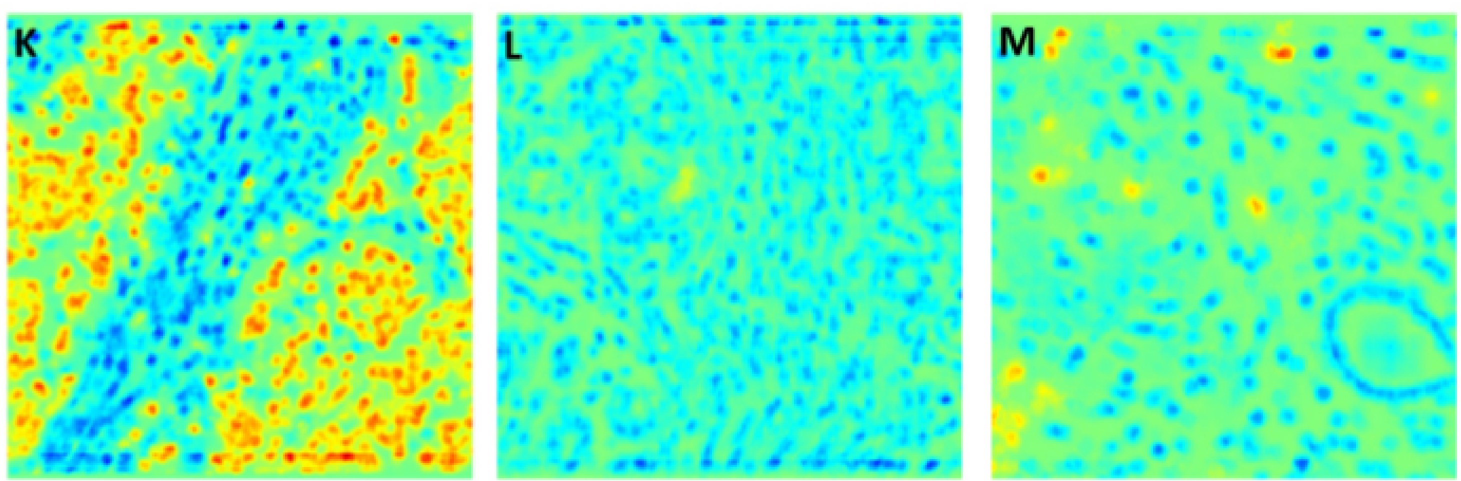
Lapuschkin et al. 2016. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks

Example: Medical Diagnosis [Binder'18]

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'17]

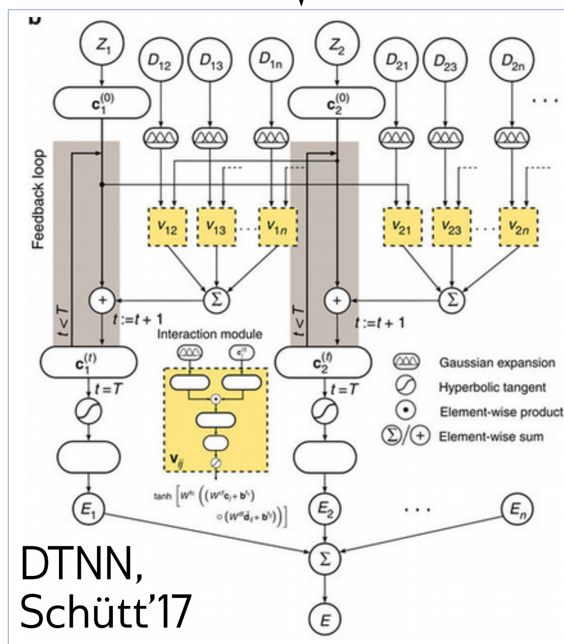
Schütt et al. 2017: Quantum-Chemical Insights from Deep Tensor Neural Networks

molecular structure (e.g. atoms positions)

DFT calculation of the stationary Schrödinger Equation

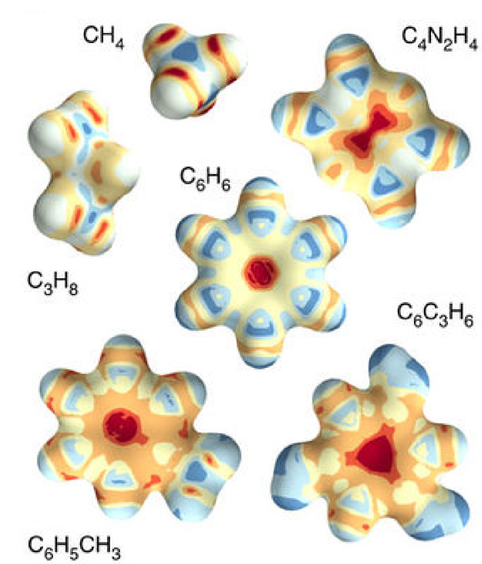
$$H\Phi = E\Phi$$

PBE0, Pedrew'86



DTNN, Schütt'17

interpretable insight



molecular electronic properties (e.g. atomization energy)

Examples of Explanation Methods

Baehrens'10
Gradient

Sundarajan'17
Int Grad

Zintgraf'17
Pred Diff

Ribeiro'16
LIME

Haufe'15
Pattern

Zurada'94
Gradient

Symonian'13
Gradient

Zeiler'14
Occlusions

Fong'17
M Perturb

Kindermans'17
PatternNet

Poulin'06
Additive

Lundberg'17
Shapley

Bazen'13
Taylor

Montavon'17
Deep Taylor

Shrikumar'17
DeepLIFT

Zeiler'14
Deconv

Landecker'13
Contrib Prop

Bach'15
LRP

Zhang'16
Excitation BP

Caruana'15
Fitted Additive

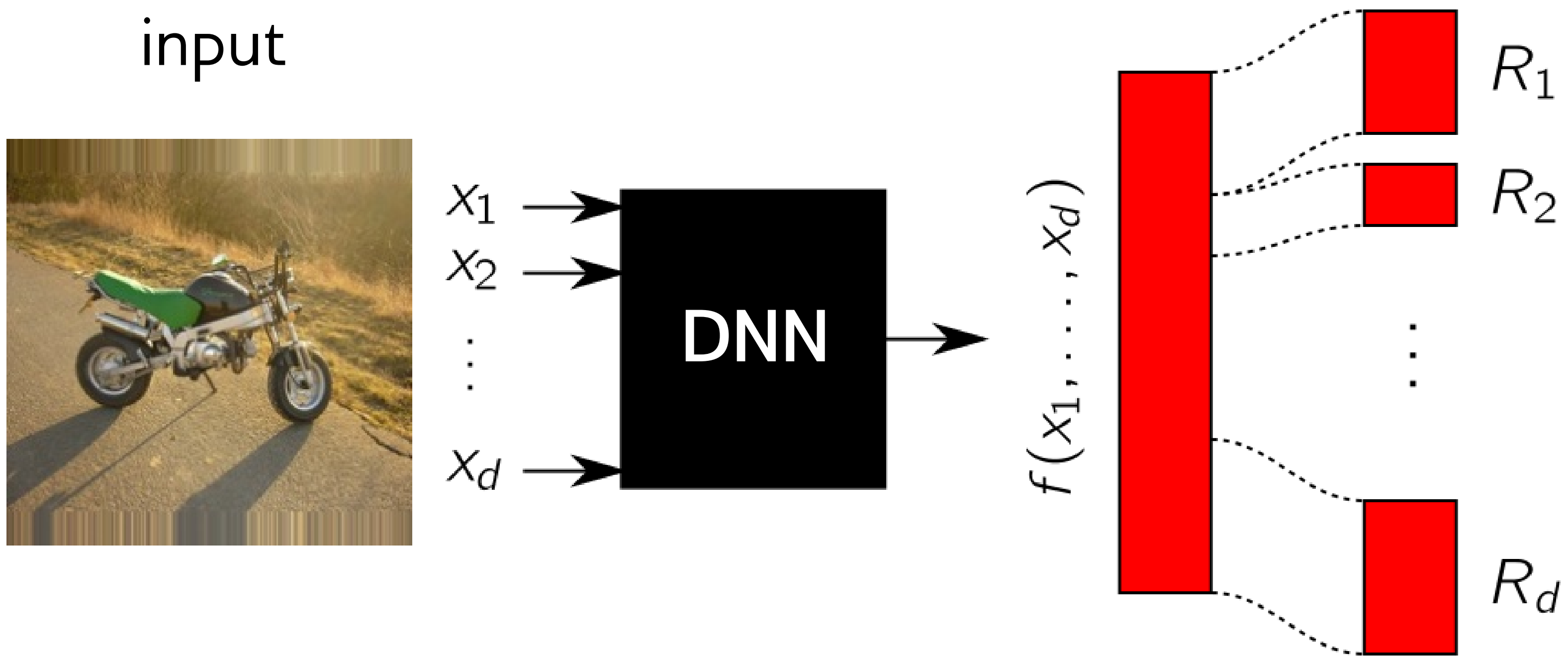
Springenberg'14
Guided BP

Zhou'16
GAP

Selvaraju'17
Grad-CAM

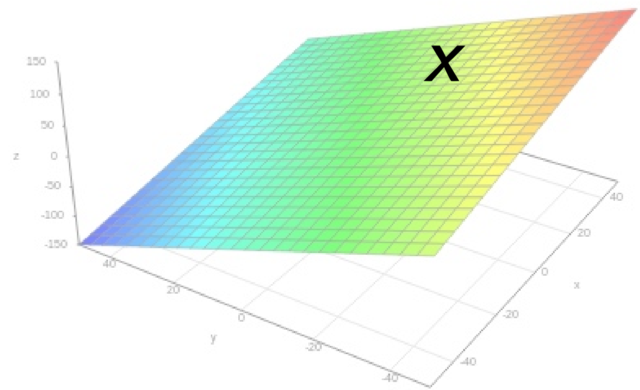
Explaining by Decomposing

Importance of a variable is the share of the function score that can be attributed to it.



Decomposition property: $f(x_1, \ldots, x_d) = \sum_{i=1}^{d} R_i$

Explaining Linear Models



$$f(x_1, \dots, x_d) = \sum_{i=1}^{d} w_i x_i + b$$

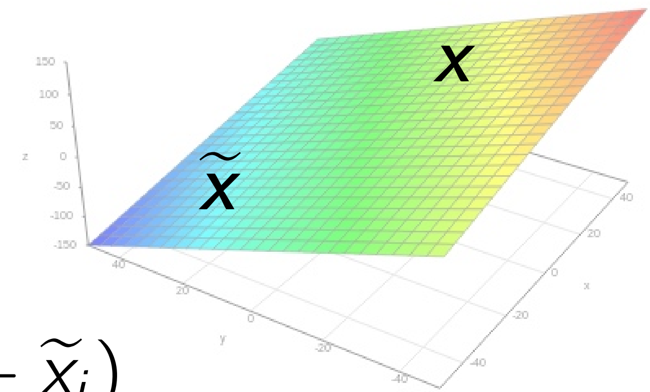
A simple method:

$$f(x_1, \dots, x_d) = \sum_{i=1}^{d} \underbrace{w_i x_i}_{R_i} + b$$

$$f(x_1, \dots, x_d) \approx \sum_{i=1}^{d} R_i$$

Explaining Linear Models

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{d} w_i x_i + b$$



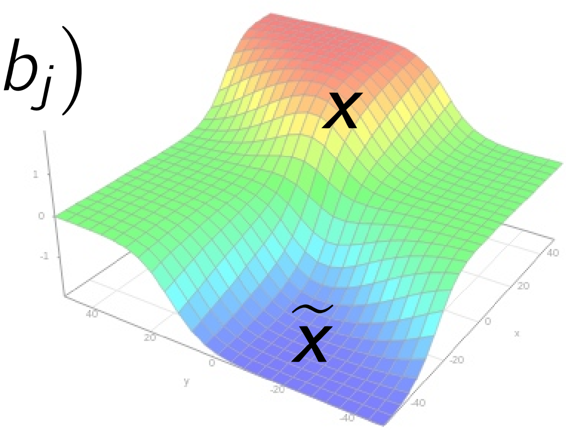
Taylor decomposition approach:

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{d} \underbrace{\frac{\partial f}{\partial x_i}\Big|_{\widetilde{\boldsymbol{x}}} \cdot (x_i - \widetilde{x}_i)}_{R_i = w_i \cdot (x_i - \widetilde{x}_i)}$$

Insight: explanation depends on the root point.

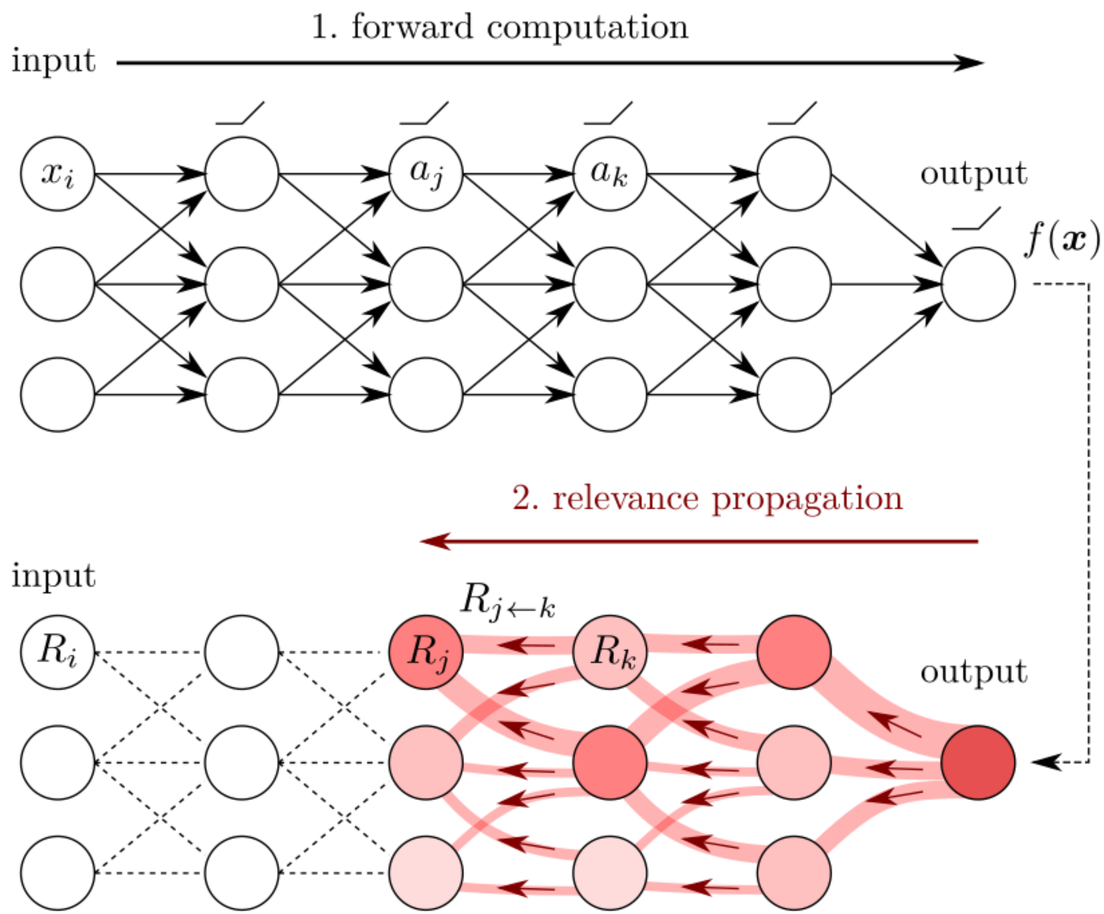
Explaining Nonlinear Models

$$f(x_1,\dots,x_d) = \sum_j \rho\Big(\sum_{i=1}^d w_{ij}x_i + b_j\Big)$$



$$f(x_1,\dots,x_d) = \sum_{i=1}^d \frac{\partial f}{\partial x_i}\Big|_{\widetilde{\boldsymbol{x}}} \cdot (x_i - \widetilde{x}_i) + \underbrace{o(\boldsymbol{x}\boldsymbol{x}^\top)}_{\text{second-order terms are hard to interpret and can be very large}}$$

Explaining Nonlinear Models by Propagation



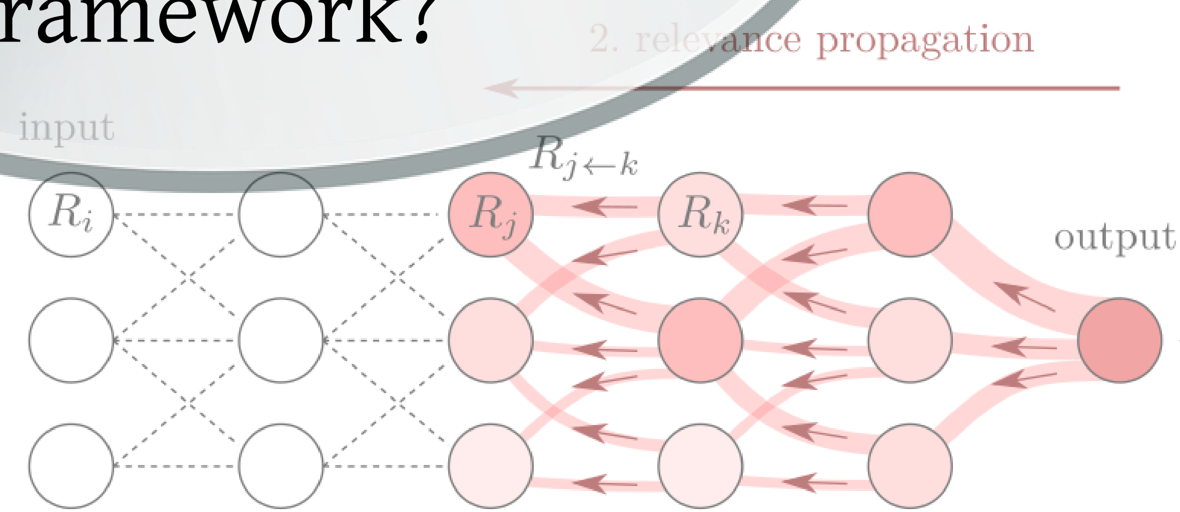
$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

Layer-Wise Relevance Propagation (LRP) [Bach'15]

Explaining Nonlinear Models by Propagation

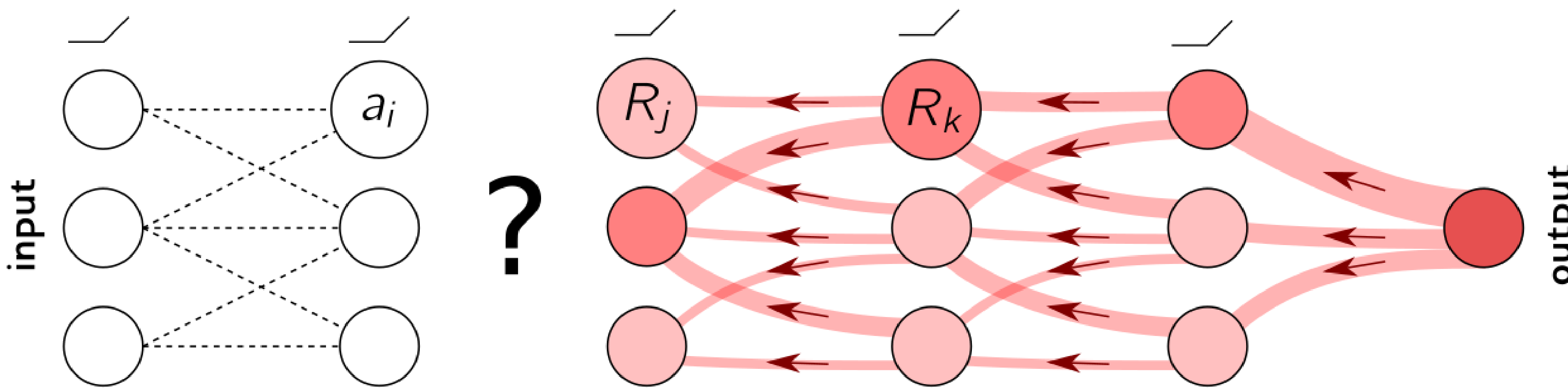
Is there an underlying mathematical framework?

$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$



Deep Taylor Decomposition (DTD) [Montavon'17]

Question: Suppose that we have propagated LRP scores ("relevance") until a given layer. How should it be propagated one layer further?



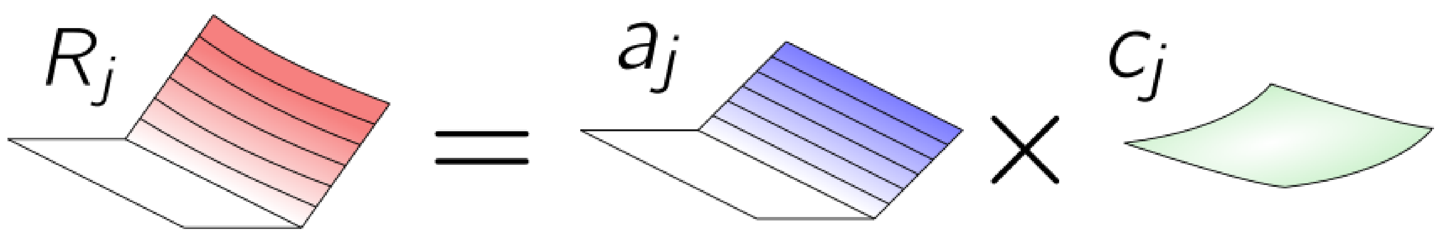
$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

Key idea: Let's use Taylor expansions for this.

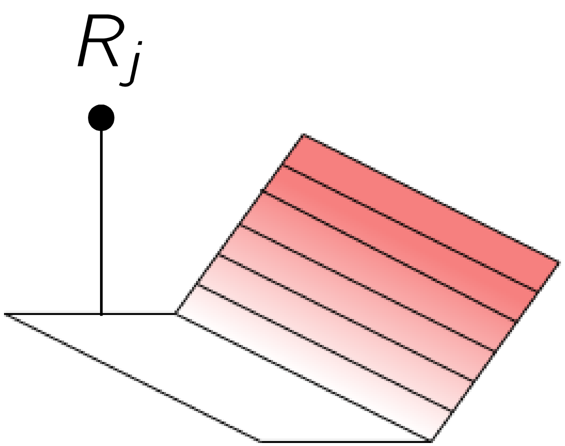
DTD Step 1: The Structure of Relevance

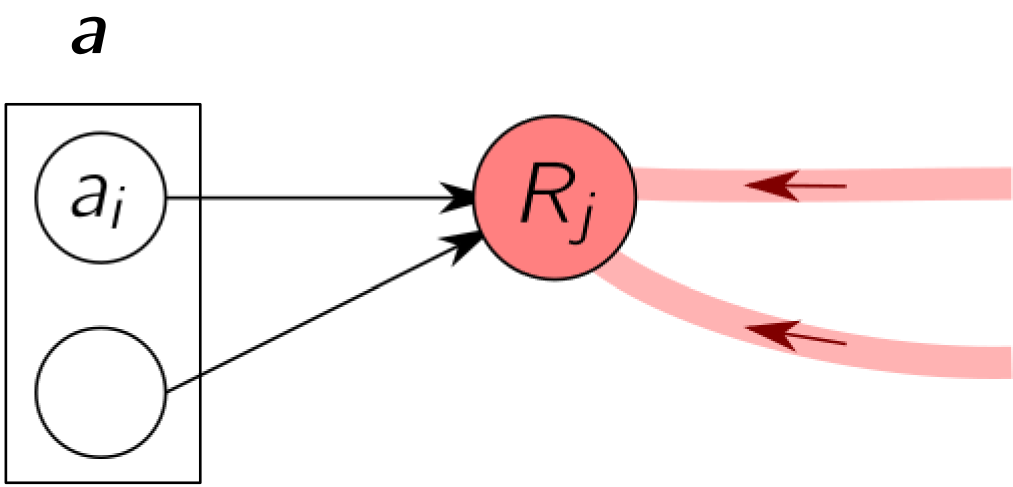
$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

Observation: Relevance at each layer is a product of the activation and an approximately constant term.



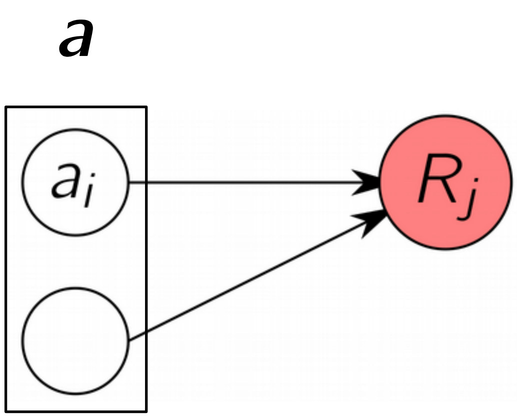
DTD Step 1: The Stucture of Relevance

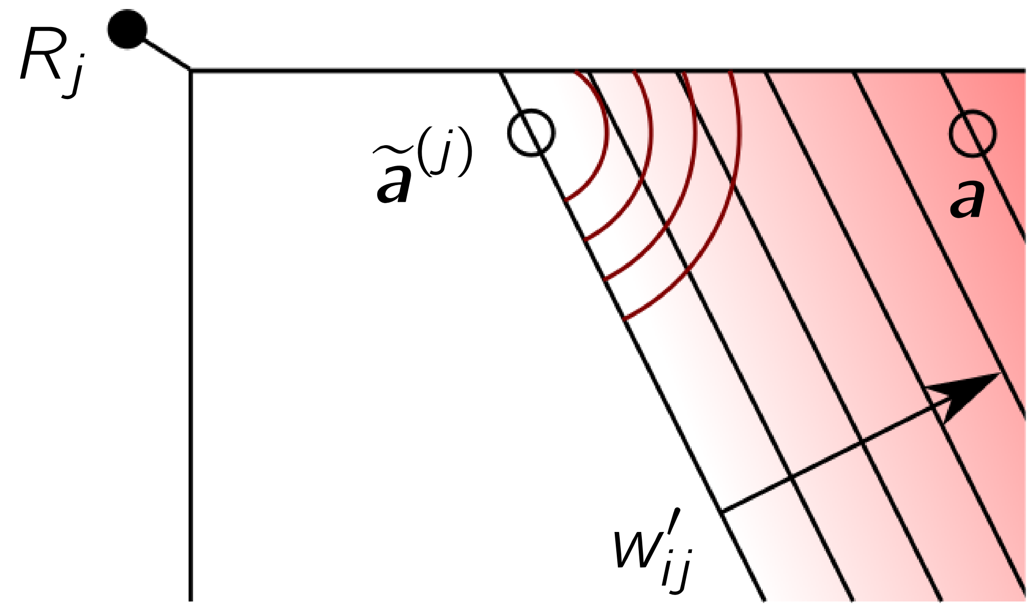




$$
\begin{aligned}
R_j(\boldsymbol{a}) &= \max\left(0, \textstyle\sum_i a_i w_{ij} + b_j\right) \cdot c_j \\
&= \max\left(0, \textstyle\sum_i a_i \underbrace{w_{ij} c_j}_{w'_{ij}} + \underbrace{b_j c_j}_{b'_j}\right)
\end{aligned}
$$

DTD Step 2: Taylor Expansion





$$R_j(\boldsymbol{a}) = \sum_i \frac{\partial R_j}{\partial a_i}\bigg|_{\widetilde{\boldsymbol{a}}^{(j)}} \cdot (a_i - \widetilde{a}_i^{(j)})$$

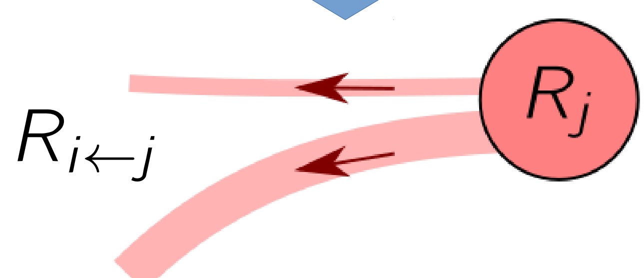
DTD Step 2: Taylor Expansion

Taylor expansion at root point:

$$R_j(\boldsymbol{a}) = \sum_i \frac{\partial R_j}{\partial a_i}\bigg|_{\widetilde{\boldsymbol{a}}^{(j)}} \cdot (a_i - \widetilde{a}_i^{(j)})$$

$$\frac{(a_i - \widetilde{a}_i^{(j)})w_{ij}}{\sum_i (a_i - \widetilde{a}_i^{(j)})w_{ij}} R_j$$

Relevance can now be backward propagated



DTD Step 3: Choosing the Root Point

$$R_{i \leftarrow j} = \frac{(a_i - \widetilde{a}_i^{(j)})w_{ij}}{\sum_i (a_i - \widetilde{a}_i^{(j)})w_{ij}} R_j$$

(Deep Taylor generic)

Choice of root point		$\widetilde{\boldsymbol{a}}^{(j)} \in \mathcal{D}$
1. nearest root	$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{w}_j$	
2. rescaled excitations	$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{a} \odot \mathbf{1}_{\boldsymbol{w}_j \succ 0}$	✔

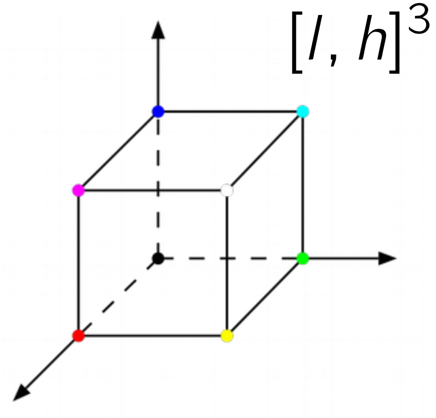
$$R_{i \leftarrow j} = \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

(same as LRP-$\alpha_1\beta_0$)

DTD: Choosing the Root Point

$$R_{i\leftarrow j} = \frac{(x_i - \widetilde{x}_i^{(j)})w_{ij}}{\sum_i (x_i - \widetilde{x}_i^{(j)})w_{ij}} R_j \qquad \text{(Deep Taylor generic)}$$

Pixels domain



Choice of root point

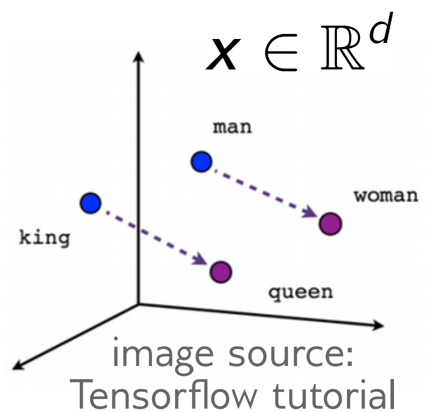
$$(\boldsymbol{x} - \widetilde{\boldsymbol{x}}^{(j)}) = t \cdot (\boldsymbol{x} - \boldsymbol{l} \odot 1_{\boldsymbol{w}_j \succ 0} - \boldsymbol{h} \odot 1_{\boldsymbol{w}_j \prec 0})$$

$$R_{i\leftarrow j} = \frac{x_{ij}w_{ij} - l_i w_{ij}^{+} - h_i w_{ij}^{-}}{\sum_i x_{ij}w_{ij} - l_i w_{ij}^{+} - h_i w_{ij}^{-}} R_j$$

DTD: Choosing the Root Point

$$R_{i \leftarrow j} = \frac{(x_i - \widetilde{x}_i^{(j)}) w_{ij}}{\sum_i (x_i - \widetilde{x}_i^{(j)}) w_{ij}} R_j$$ (Deep Taylor generic)

Embedding:



Choice of root point

$$(\boldsymbol{x} - \boldsymbol{x}^{(j)}) = t \cdot \boldsymbol{w}_j$$

$$R_{i \leftarrow j} = \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$$

DTD: Application to Pooling Layers

A sum-pooling layer over positive activations is equivalent to a ReLU layer with weights 1.

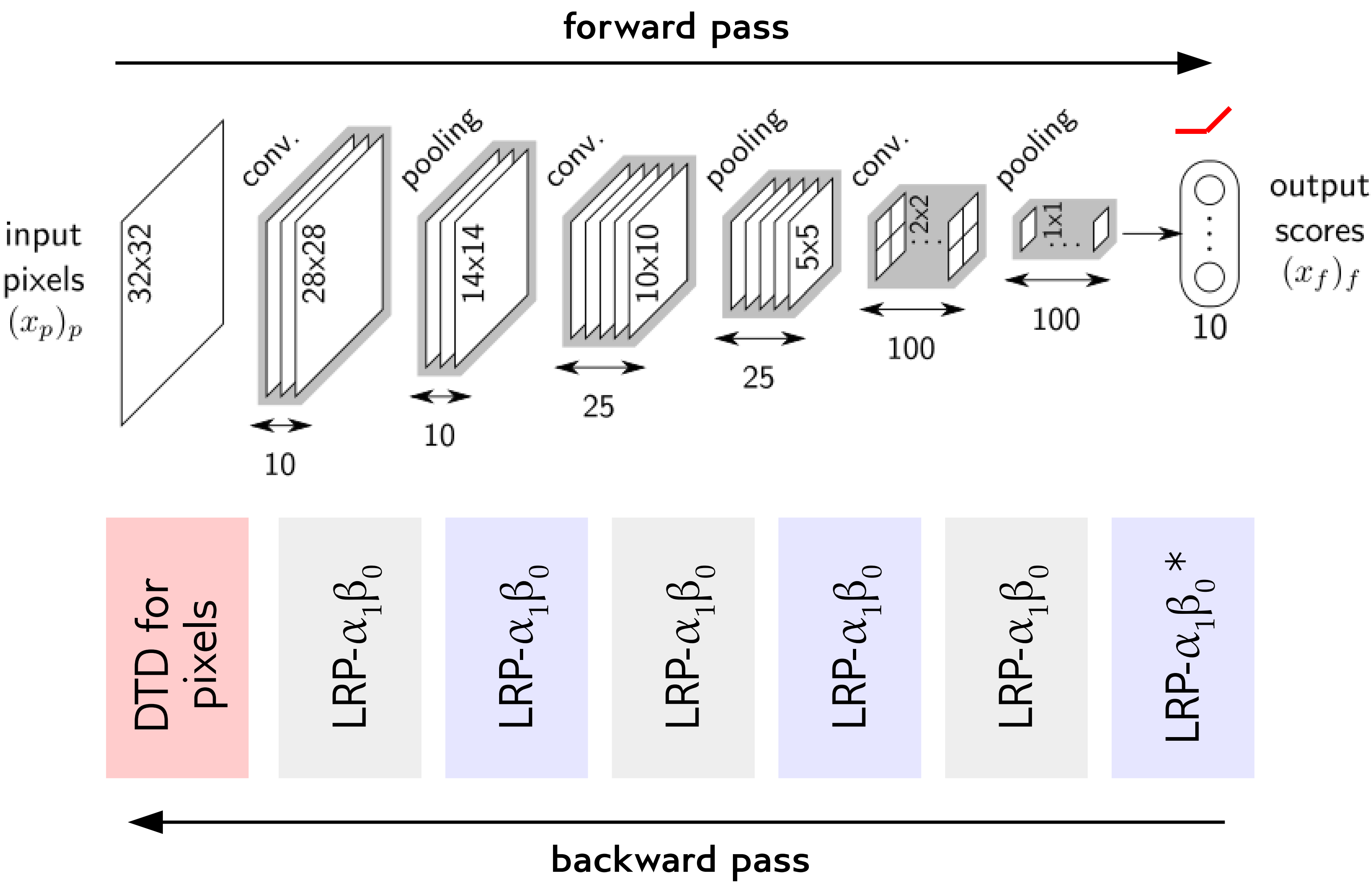
$$a_j = \big(\textstyle\sum_i a_i\big) = \max\big(0, \textstyle\sum_i a_i 1_{ij} + 0_j\big)$$

A p-norm pooling layer can be approximated as a sum-pooling layer multiplied by a ratio of norms that we treat as constant [Montavon'17].

$$a_j = \big(\textstyle\sum_i a_i\big) \cdot \frac{\|(a_i)_i\|_p}{\|(a_i)_i\|_1}$$

→ Treat pooling layers as ReLU detection layers

Deep Taylor Decomposition on ConvNets



* For top-layers, other rules may improve selectivity

Implementing Propagation Rules

Example: LRP-$\alpha_1\beta_0$:

$$\boxed{R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j}$$

Sequence of element-wise computations	Sequence of vector computations
$z_j \to \sum_i a_i w_{ij}^+$	$\boldsymbol{z} \to W_+^\top \cdot \boldsymbol{a}$
$s_j \to R_j / z_j$	$\boldsymbol{s} \to \boldsymbol{R} \oslash \boldsymbol{z}$
$c_i \to \sum_j w_{ij}^+ s_j$	$\boldsymbol{c} \to W_+ \cdot \boldsymbol{s}$
$R_i \to a_i c_i$	$\boldsymbol{R} \to \boldsymbol{a} \odot \boldsymbol{c}$

Implementing Propagation Rules

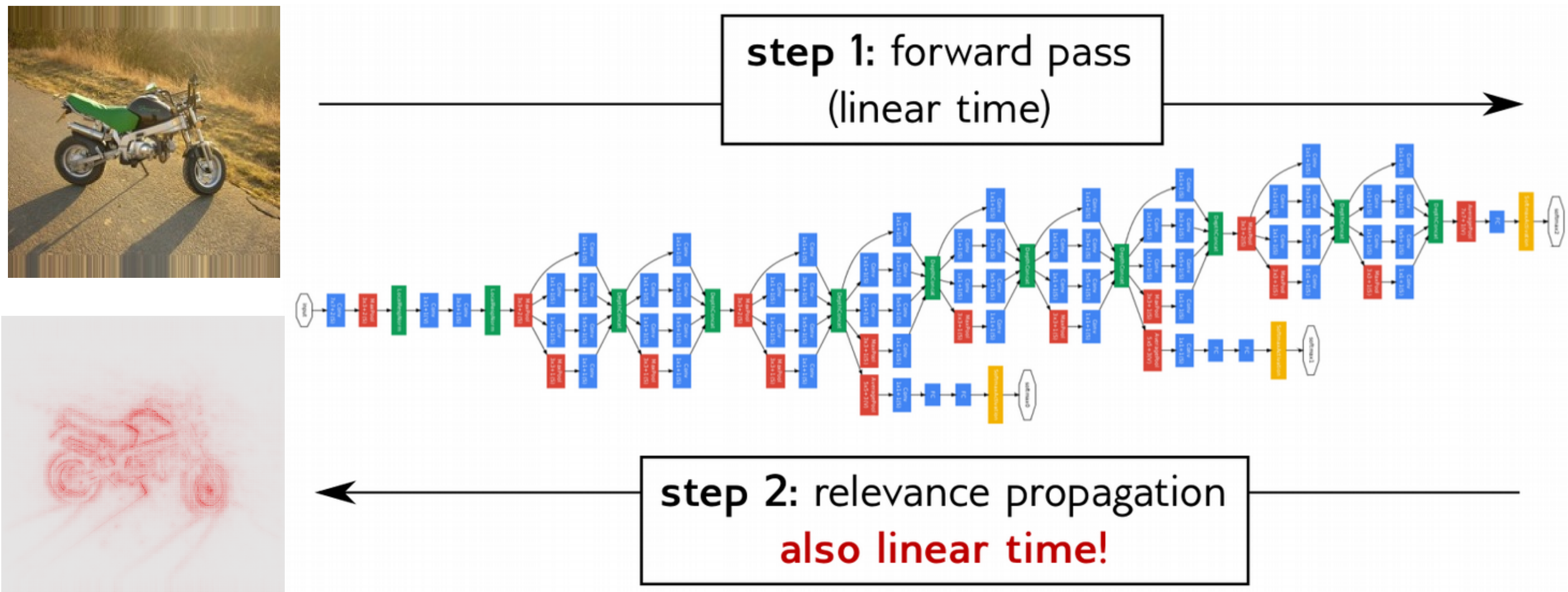
Example: LRP-$\alpha_1\beta_0$:

$$\boxed{R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j}$$

Code that reuses forward and gradient computations:

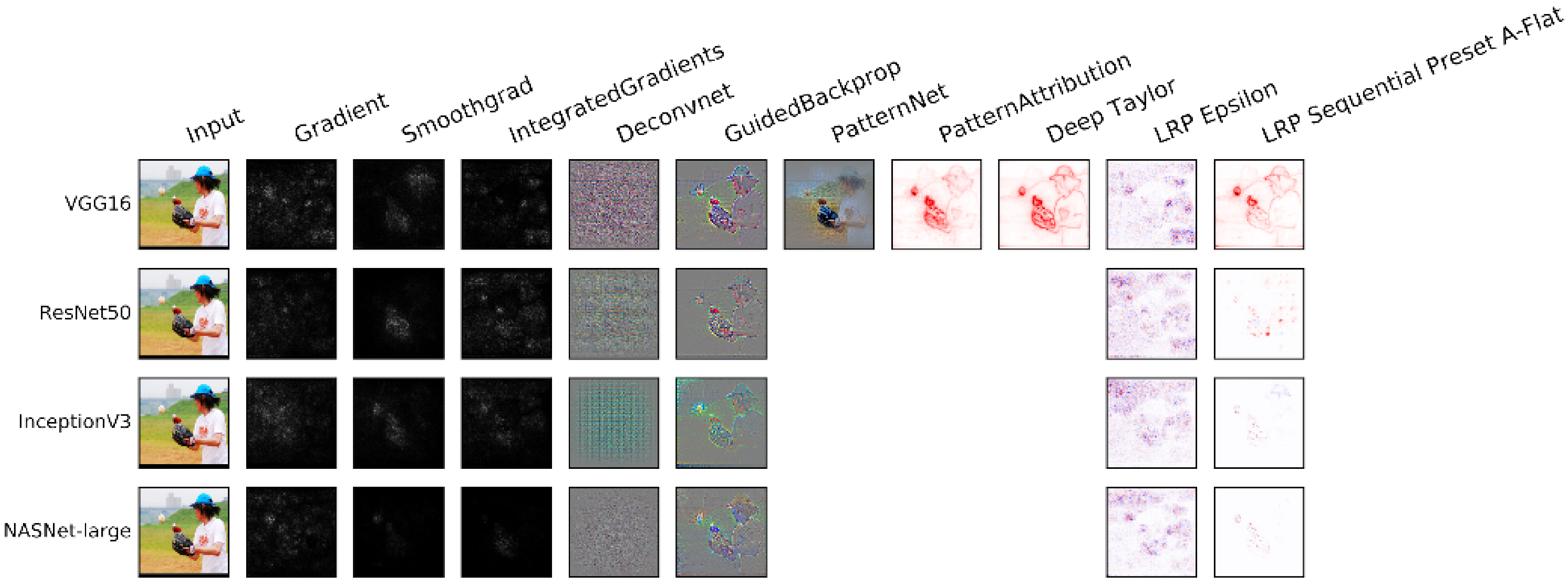
```
def lrp(layer,a,R):

    clone = layer.clone()
    clone.W = maximum(0,layer.W)
    clone.B = 0

    z = clone.forward(a)
    s = R / z
    c = clone.backward(s)

    return a * c
```

How Deep Taylor / LRP Scales



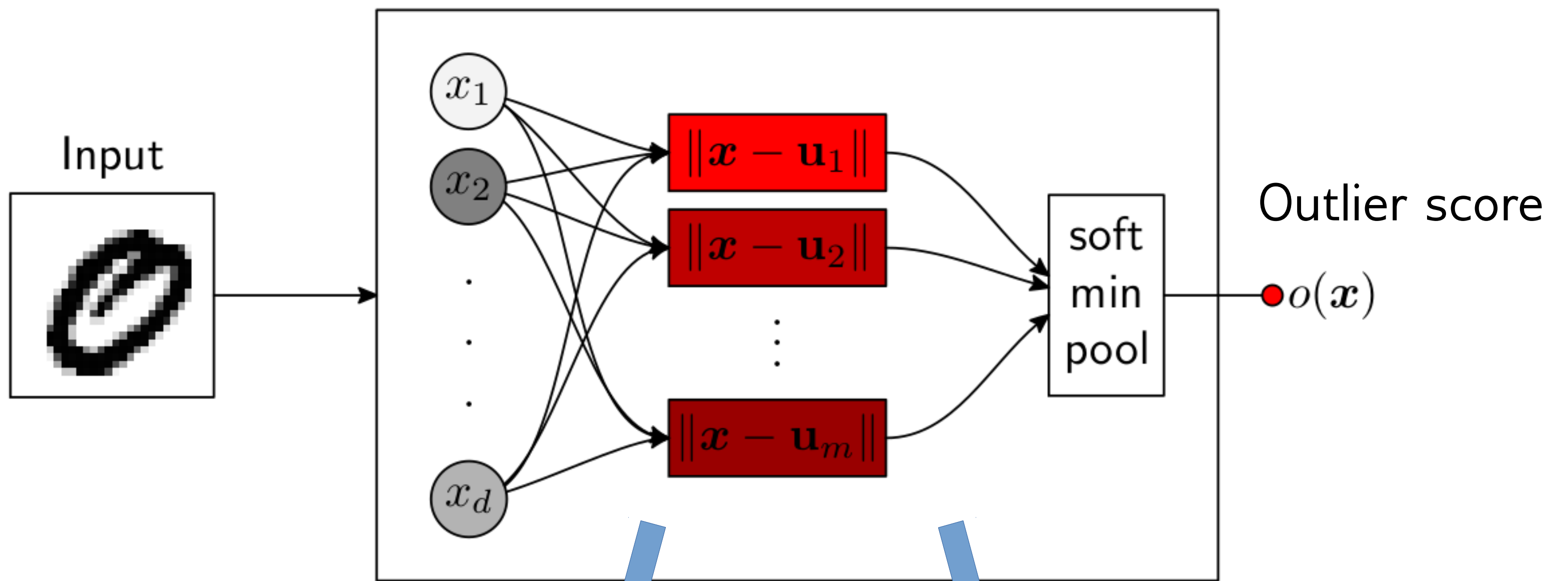
Implementation on Large-Scale Models [Alber'18]

https://github.com/albermax/innvestigate



DTD for Kernel Models [Kauffmann'18]

1. Build a neural network equivalent of the One-Class SVM:



2. Computes its deep Taylor decomposition

$$R_i = \sum_j \frac{(x_i - u_{ij})^2}{\|\boldsymbol{x} - \boldsymbol{u}_j\|_2^2}(R_j - D_j^+)$$

Gaussian/Laplace Kernel

$$R_j = (a_j + \varepsilon_j) \cdot \frac{\exp(-a_j)}{\sum_j \exp(-a_j)}$$

Student Kernel

$$R_j = a_j \cdot \mathbb{H}[(h_{j'}/h_j)_{j'}]$$

DTD: Choosing the Root Point (Revisited)

$$R_{i \leftarrow j} = \frac{(a_i - \widetilde{a}_i^{(j)}) w_{ij}}{\sum_i (a_i - \widetilde{a}_i^{(j)}) w_{ij}} R_j$$

(Deep Taylor generic)

Choice of root point		$\widetilde{\boldsymbol{a}}^{(j)} \in \mathcal{D}$
1. nearest root	$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{w}_j$	
2. rescaled excitations	$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{a} \odot \mathbf{1}_{\boldsymbol{w}_j \succ 0}$	✔
3. rescaled activation	$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{a}$	✔

2. LRP-$\alpha_1\beta_0$

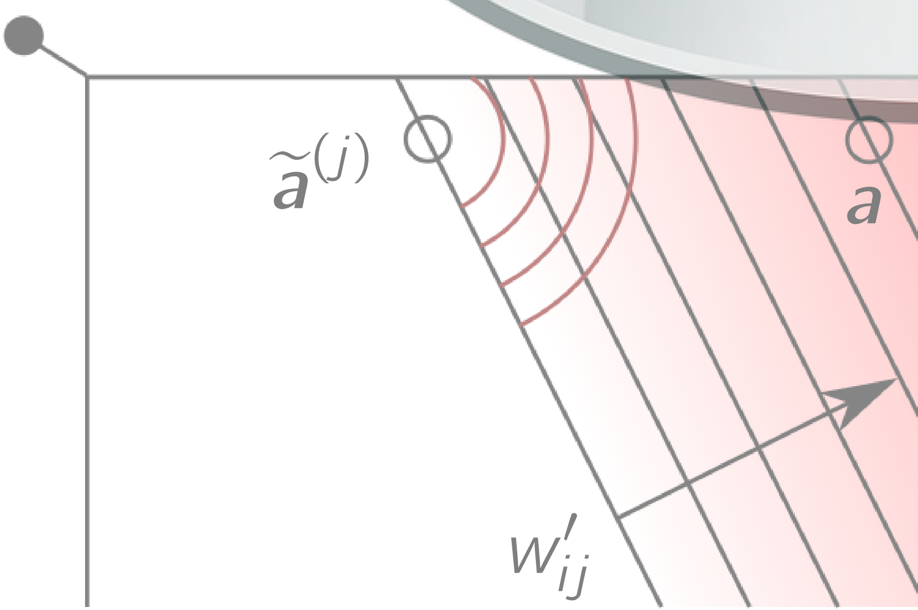
$$R_{i \leftarrow j} = \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

3. Another rule

$$R_{i \leftarrow j} = \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j$$

Selecting the Explanation Technique

How to select the best root points?



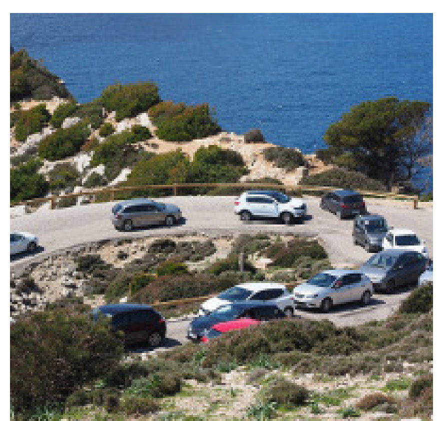
$$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{a} \odot \mathbf{1}_{\boldsymbol{w}_j \succ 0}$$

$$\widetilde{\boldsymbol{a}}^{(j)} = \boldsymbol{a} - t \cdot \boldsymbol{a}$$

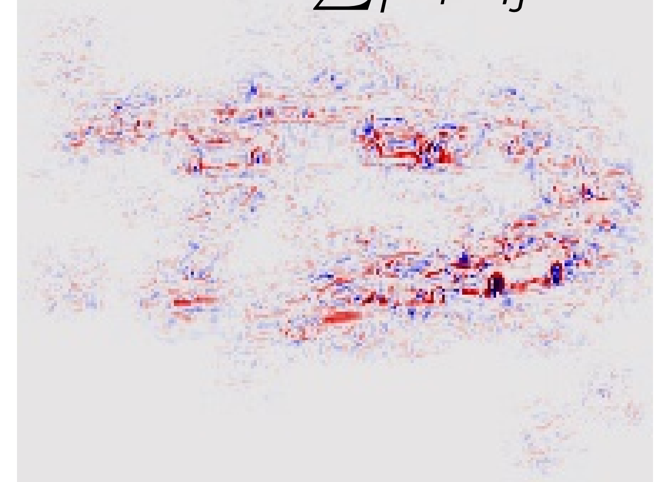
Selecting the Explanation Technique

Which rule leads to the best explanation?

$$R_{i \leftarrow j} = \frac{a_i w_{ij}^{+}}{\sum_i a_i w_{ij}^{+}} R_j$$



$$R_{i \leftarrow j} = \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j$$



Selecting the Explanation Technique

What axioms should
an explanation satisfy?

$$\left\{ \begin{array}{rl} \text{Continuity:} & \big(\boldsymbol{x} \approx \boldsymbol{x}'\big) \wedge \big(f(\boldsymbol{x}) \approx f(\boldsymbol{x}')\big) \Rightarrow \boldsymbol{R}(\boldsymbol{x}) \approx \boldsymbol{R}(\boldsymbol{x}') \\ \text{Conservation:} & \big(\sum_p R_p(\boldsymbol{x}) = f(\boldsymbol{x})\big) \wedge \big(\sum_p |R_p(\boldsymbol{x})| < A \cdot |f(\boldsymbol{x})|\big). \end{array} \right\}$$

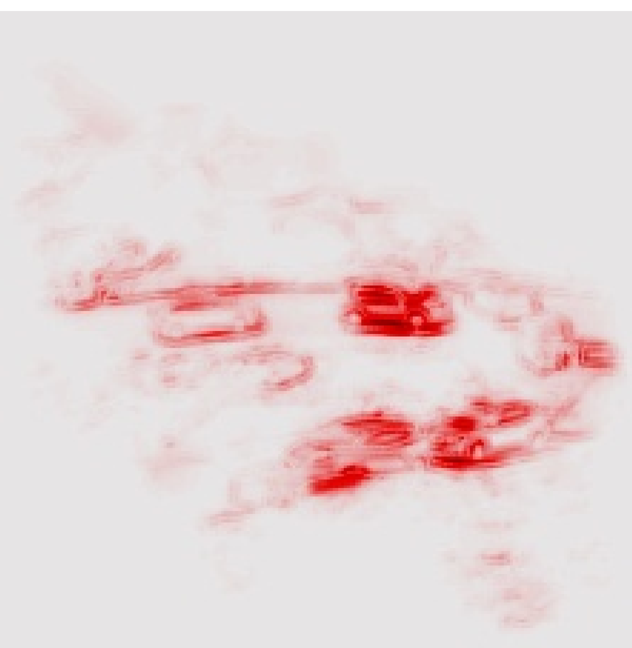
Selection based on Axioms

$$\left\{ \text{Conservation: } \left(\sum_p R_p(\boldsymbol{x}) = f(\boldsymbol{x})\right) \wedge \left(\sum_p |R_p(\boldsymbol{x})| < A \cdot |f(\boldsymbol{x})|\right). \right\}$$

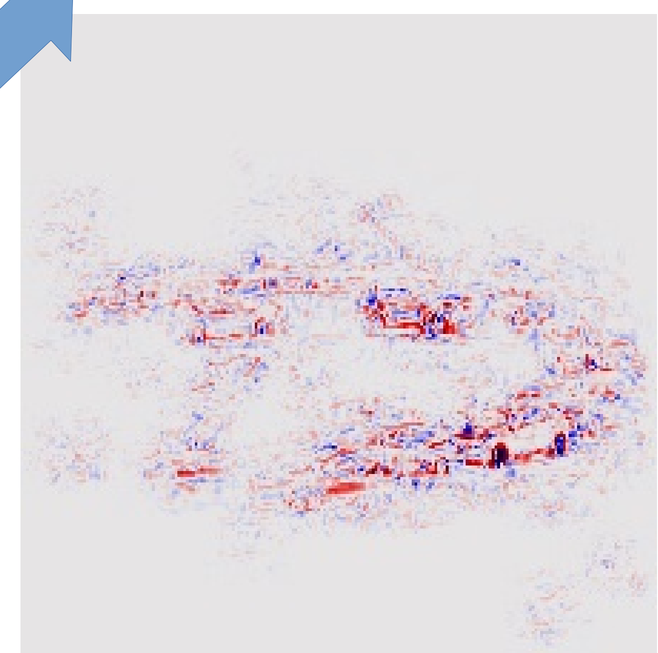
LRP-$\alpha_1\beta_0$

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j$$



division by zero → scores explode.



Selection based on Axioms

$$\left\{ \begin{array}{r} \text{Conservation:} \;\; \left(\sum_p R_p(\boldsymbol{x}) = f(\boldsymbol{x})\right) \wedge \left(\sum_p |R_p(\boldsymbol{x})| < A \cdot |f(\boldsymbol{x})|\right). \\ \text{Continuity:} \;\; (\boldsymbol{x} \approx \boldsymbol{x}') \wedge \big(f(\boldsymbol{x}) \approx f(\boldsymbol{x}')\big) \Rightarrow \boldsymbol{R}(\boldsymbol{x}) \approx \boldsymbol{R}(\boldsymbol{x}') \end{array} \right\}$$

LRP-$\alpha_1\beta_0$

$$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_i a_i w_{ij}} R_j$$

$$\Downarrow \quad R_i = a_i \delta_i$$

$$\Downarrow \quad R_i = a_i \delta_i$$

$$\delta_i = \sum_j w_{ij}^+ \frac{(\sum_i a_i w_{ij} + b_j)^+}{\sum_i a_i w_{ij}^+} \delta_j$$

$$\delta_i = \sum_j w_{ij} \underbrace{\frac{(\sum_i a_i w_{ij} + b_j)^+}{\sum_i a_i w_{ij}}}_{\text{discontinuous step function for } b_j = 0} \delta_j$$

Explainable ML: Challenges

Underlying mathematical framework

Human perception

Validating Explanations

Similarity to ground truth

Perturbation analysis [Samek'17]

Axioms of an explanation

Explainable ML: Opportunities

Detecting unexpected ML behavior

Finding weaknesses of a dataset

Using Explanations

Human interaction

Designing better ML algorithms?

Extracting new domain knowledge

Check our webpage

with interactive demos, software, tutorials, ...

and our tutorial paper:

Montavon, G., Samek, W., Müller, K.-R. Methods for Interpreting and Understanding Deep Neural Networks, Digital Signal Processing, 2018

References

- **[Alber'18]** Alber, M. Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., Kindermans. P.-J. iNNvestigate neural networks. CoRR abs/1808.04260, 2018
- **[Bach'15]** Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W. On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. PLOS ONE 10 (7), 2015
- **[Binder'18]** Binder, A., Bockmayr, M., ..., Müller, K.-R., Klauschen, F. Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles CoRR abs/1805.11178, 2018
- **[Bojarski'17]** Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L. D., Muller, U. Explaining how a deep neural network trained with end-to-end learning steers a car. CoRR abs/1704.07911, 2017
- **[Kauffmann'18]** Kauffmann, J. Müller, K.-R., Montavon, G., Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models CoRR abs/1805.06230, 2018
- **[Lapuschkin'16]** Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., Samek, W. Analyzing classifiers: Fisher vectors and deep neural networks. CVPR 2016
- **[Montavon'17]** Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition 65, 211–222, 2017
- **[Samek'17]** Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 2017
- **[Schütt'17]** Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K.-R., Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks, Nature Communications 8, 13890, 2017
- **[Symonian'13]** Symonian, K. Vedaldi, A. Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ArXiv 2013