

Tutorial on Interpretable Machine Learning



Wojciech Samek
(Fraunhofer HHI)



Klaus-Robert Müller
(TU Berlin)

9:00 - 9:30 Part 1a: Introduction

9:30 - 10:30 Part 1b: Making Deep Neural Networks Transparent

10:30 - 11:00 Coffee Break

11:00 - 11:30 Part 2a: Making Deep Neural Networks Transparent

11:30 - 12:30 Part 2b: Applications & Discussion

Before we start

We thank our collaborators !

Grégoire Montavon (TU Berlin)

Alexander Binder (SUTD)

Sebastian Lapuschkin (Fraunhofer HHI)

Leila Arras (Fraunhofer HHI)

...

<http://interpretable-ml.org/>

Please ask questions at any time !

**NIPS'17 Workshop “Interpreting, Explaining and Visualizing
Deep Learning - Now what ?”**

Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

Part 1a: Introduction

Recent ML Systems achieve superhuman Performance

AlphaGo beats Go human champ



Deep Net outperforms humans in image classification



Autonomous search-and-rescue drones outperform humans



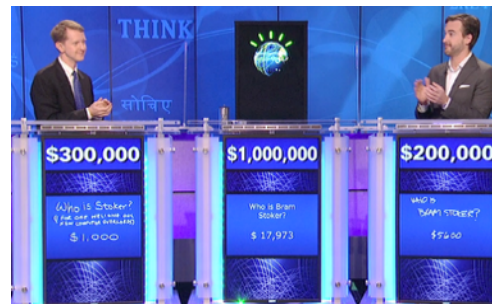
DeepStack beats professional poker players



Computer out-plays humans in "doom"



IBM's Watson destroys humans in jeopardy



Deep Net beats human at recognizing traffic signs



From Data to Information

Huge volumes of data

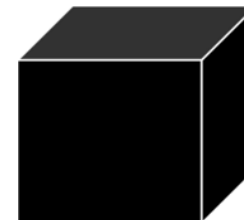


Computing power



Deep Nets / Kernel Machines / ...

Solve task



Information (implicit)

From Data to Information

Huge volumes of data



Computing power



Deep Nets / Kernel Machines / ...

Interpretable
Information

extract

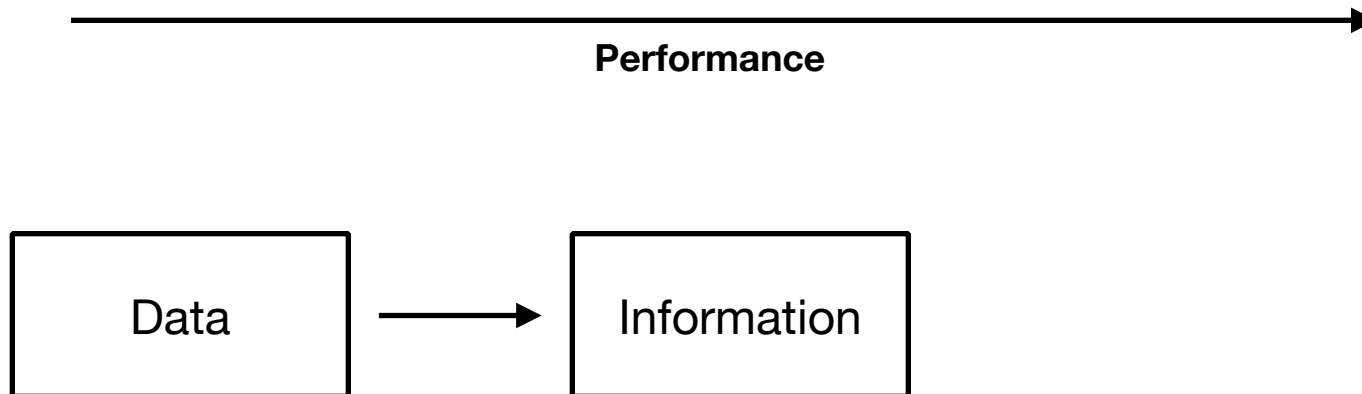
Solve task



Information (implicit)

From Data to Information

IMGENET

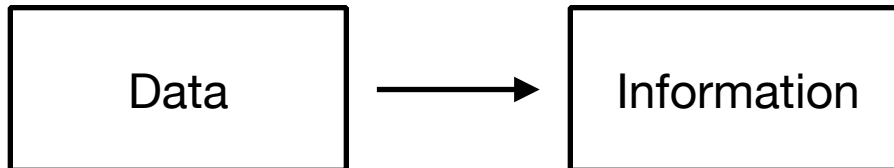


From Data to Information

IMGENET

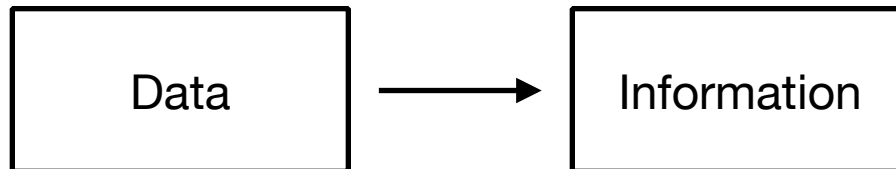
AlexNet
(16.4%)

Performance



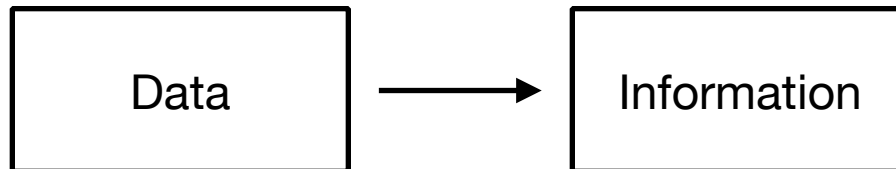
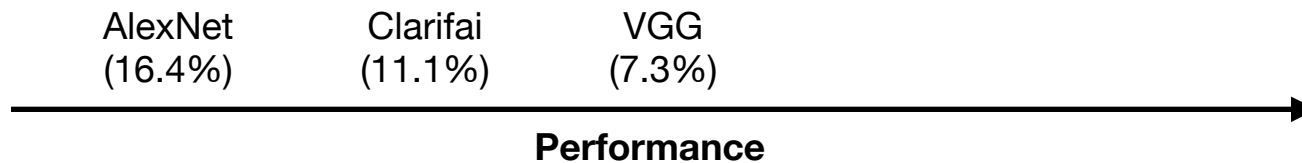
From Data to Information

IMGENET



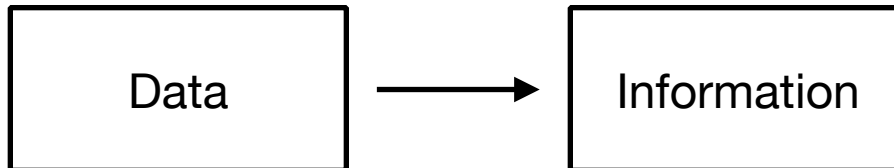
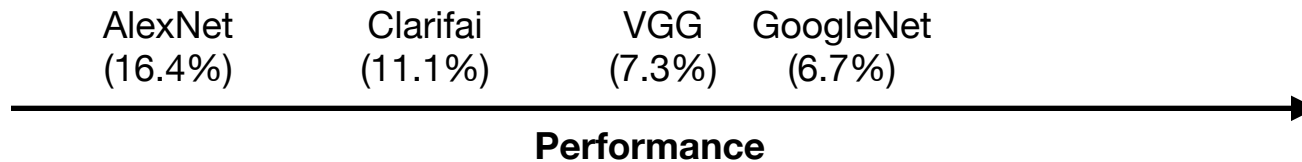
From Data to Information

IMGENET



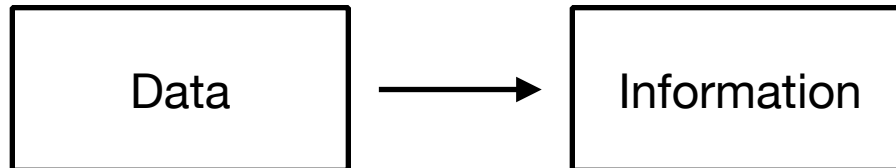
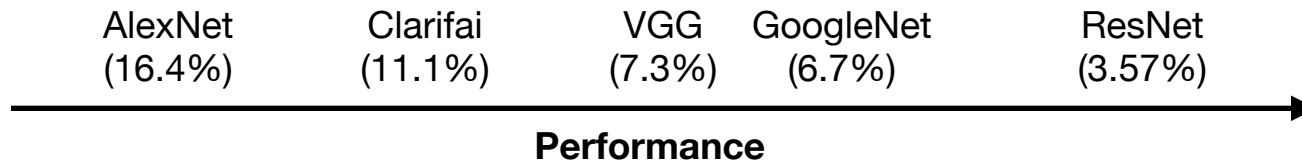
From Data to Information

IMGENET



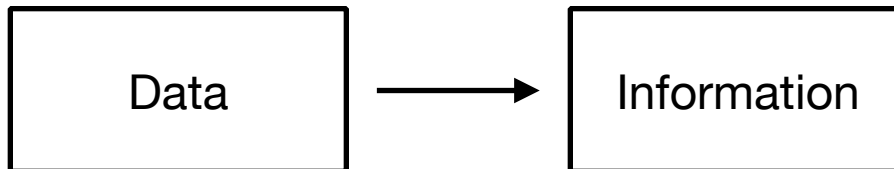
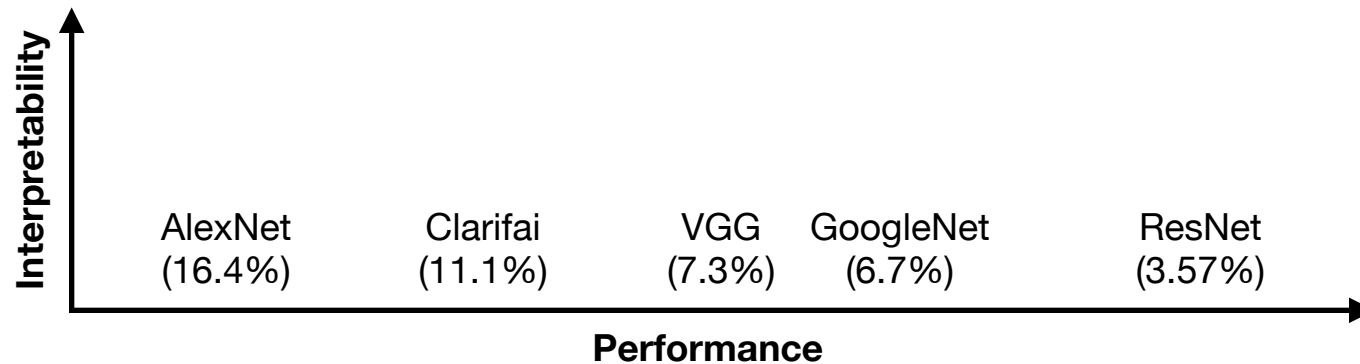
From Data to Information

IMGENET



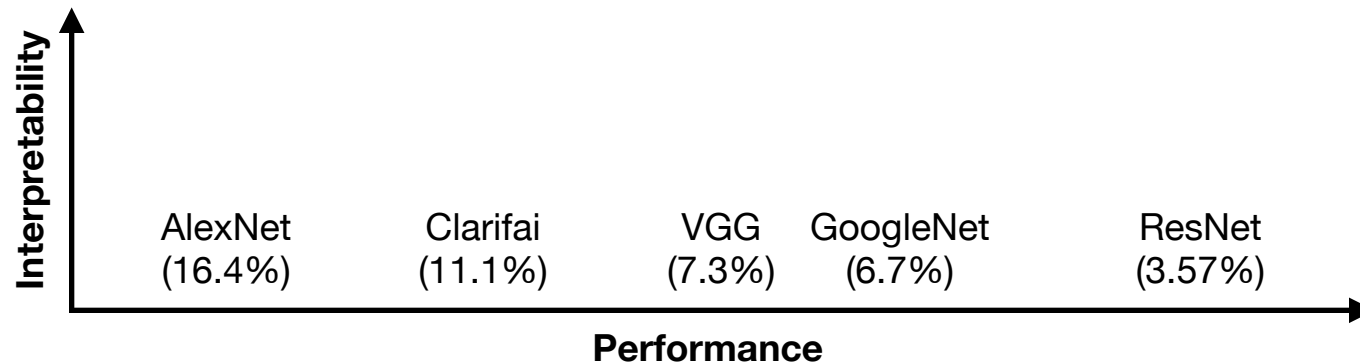
From Data to Information

IM  GENET



From Data to Information

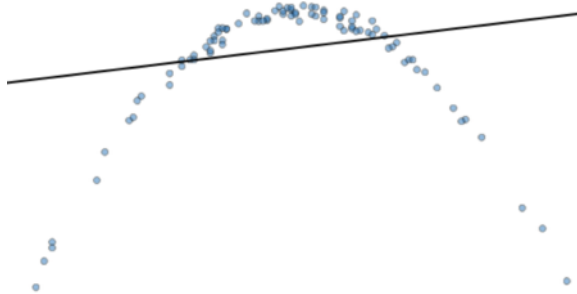
IMAGENET



**Crucial in many applications
(industry, sciences ...)**

Interpretable vs. Powerful Models ?

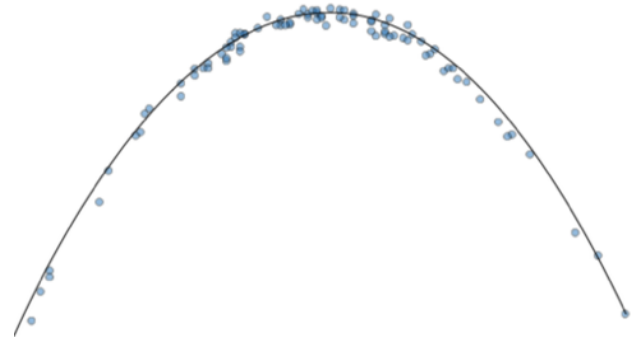
Linear model



Poor fit, but easily interpretable
“global explanation”

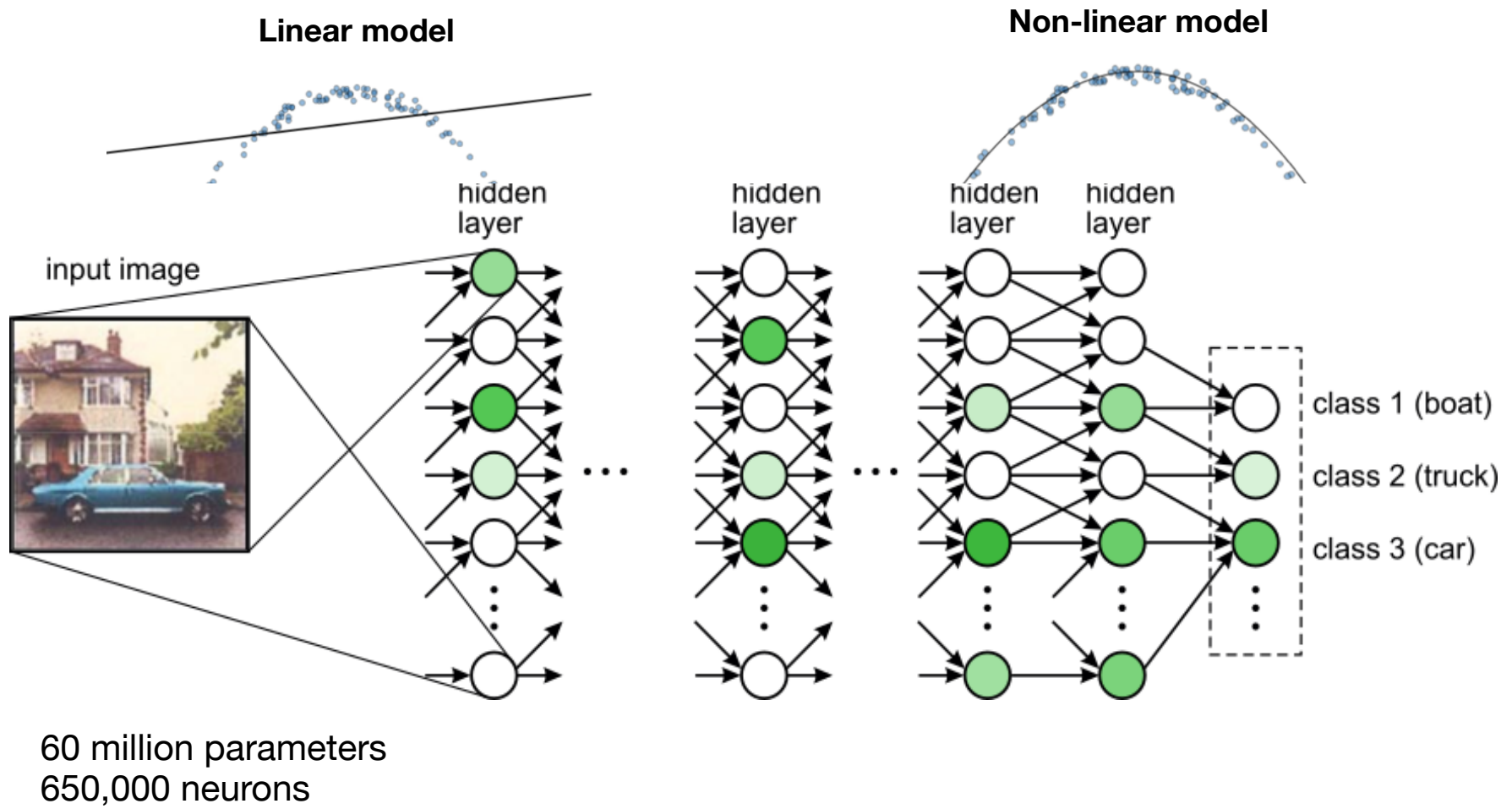
vs.

Non-linear model

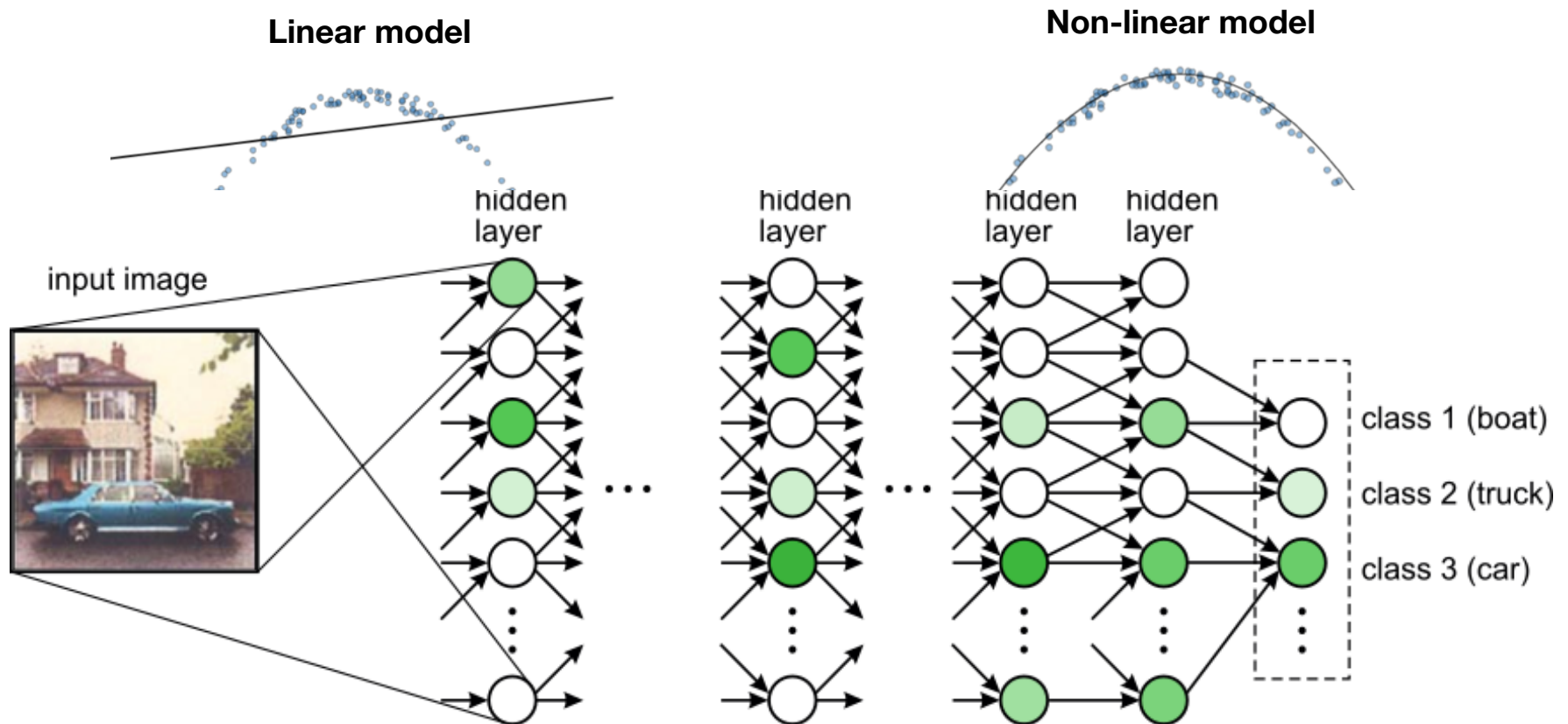


Can be very complex
“individual explanation”

Interpretable vs. Powerful Models ?



Interpretable vs. Powerful Models ?



60 million parameters
650,000 neurons

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:

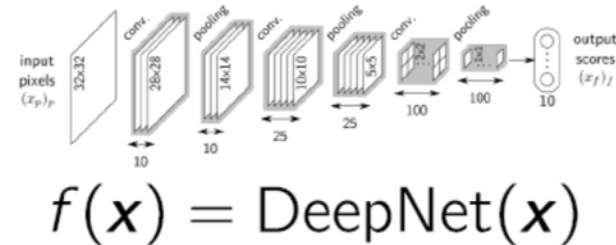
$$f(\mathbf{x}) = \sum_{i=1}^d \underbrace{g_i(x_i)}_{\text{contribution of } i\text{th variable}}$$

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.

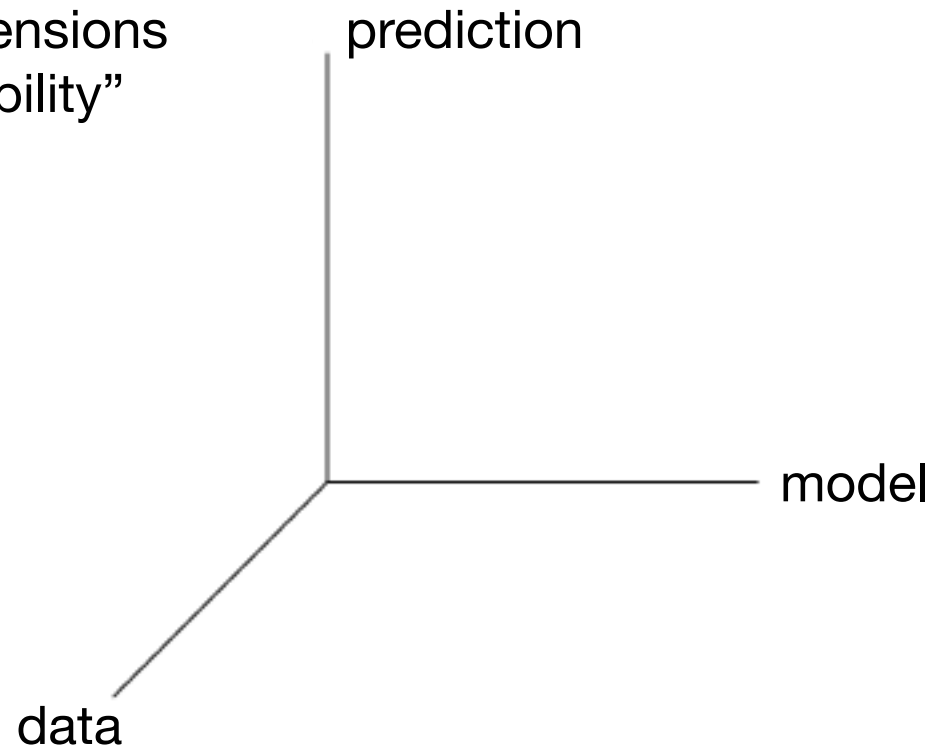
Example:



How to determine the contribution each input variable?

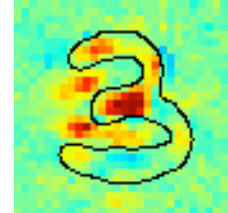
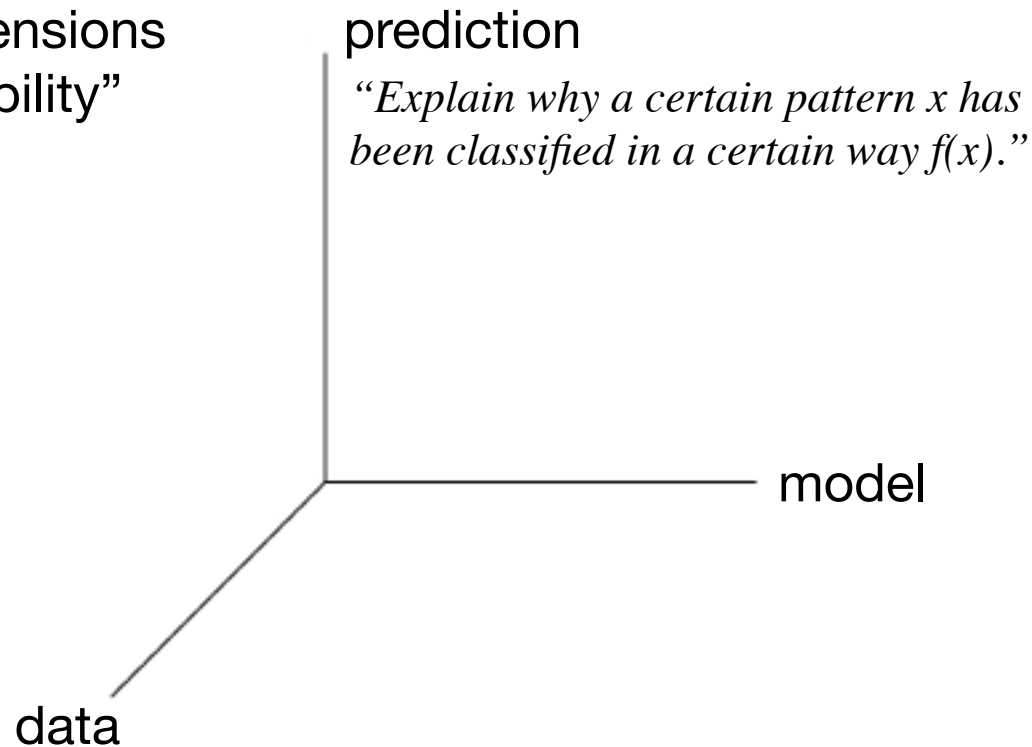
Dimensions of Interpretability

Different dimensions
of “interpretability”



Dimensions of Interpretability

Different dimensions
of “interpretability”

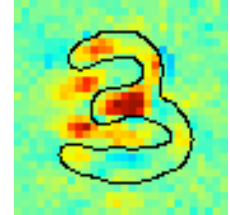


Dimensions of Interpretability

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.”



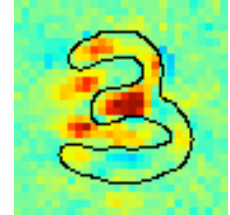
data

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.”

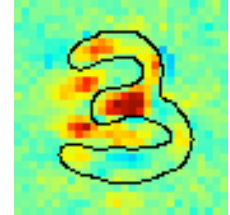


Dimensions of Interpretability

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 ?

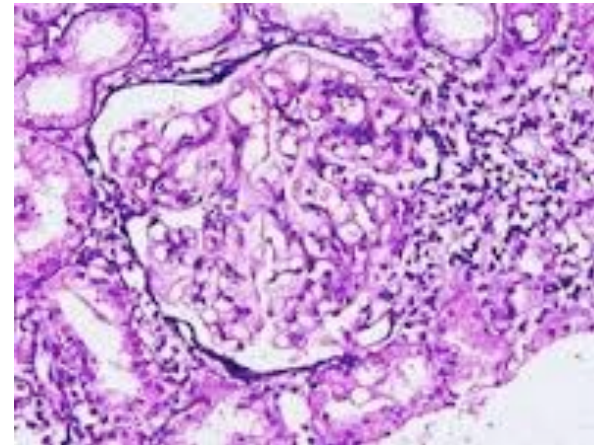
1) Verify that classifier works as expected

Wrong decisions can be costly and dangerous

“Autonomous car crashes, because it wrongly recognizes ...”

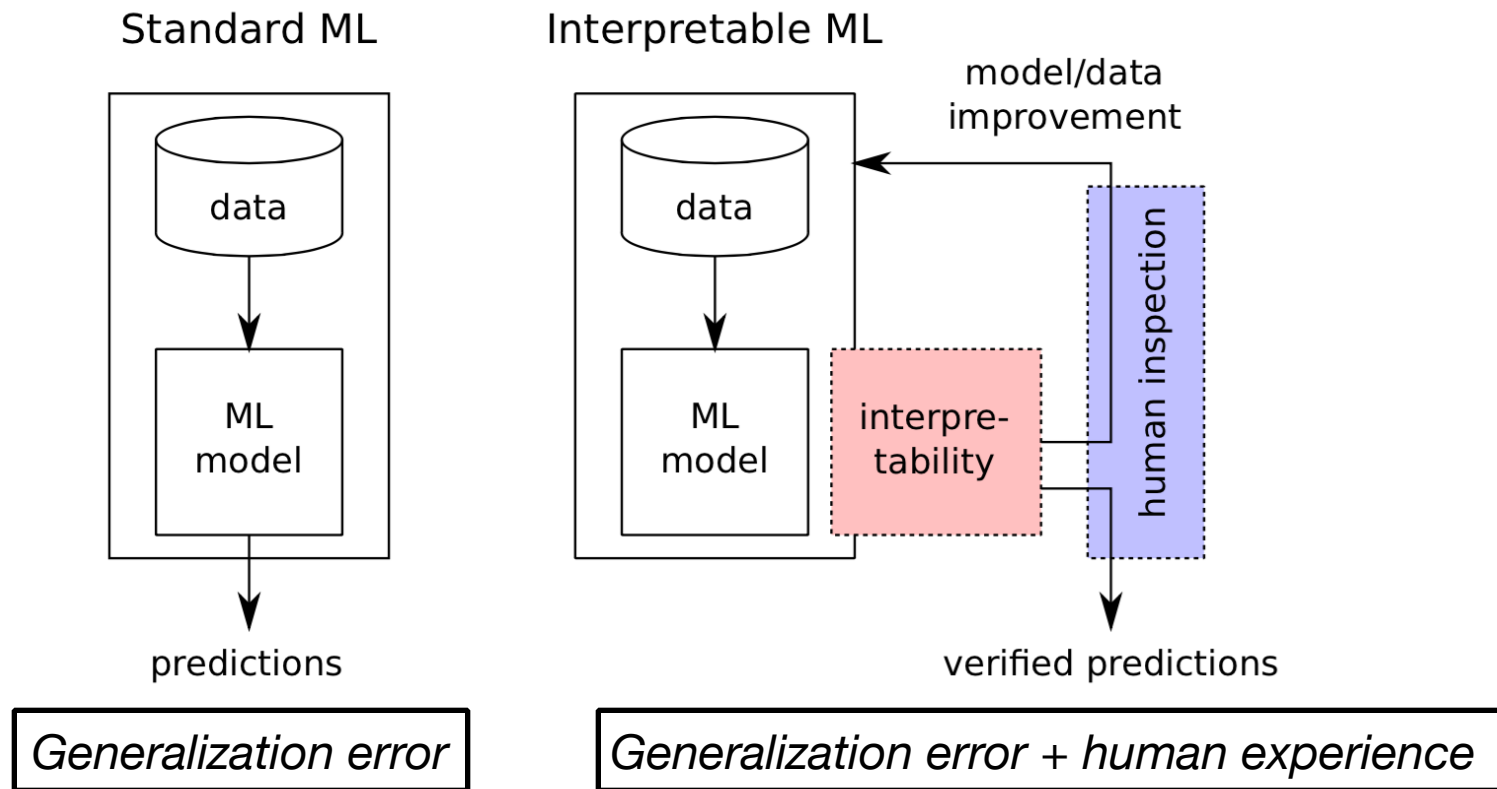


“AI medical diagnosis system misclassifies patient’s disease ...”



Why Interpretability ?

2) Improve classifier



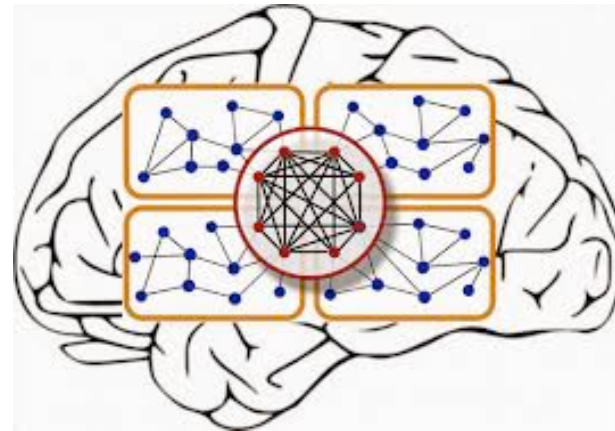
Why Interpretability ?

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 ?

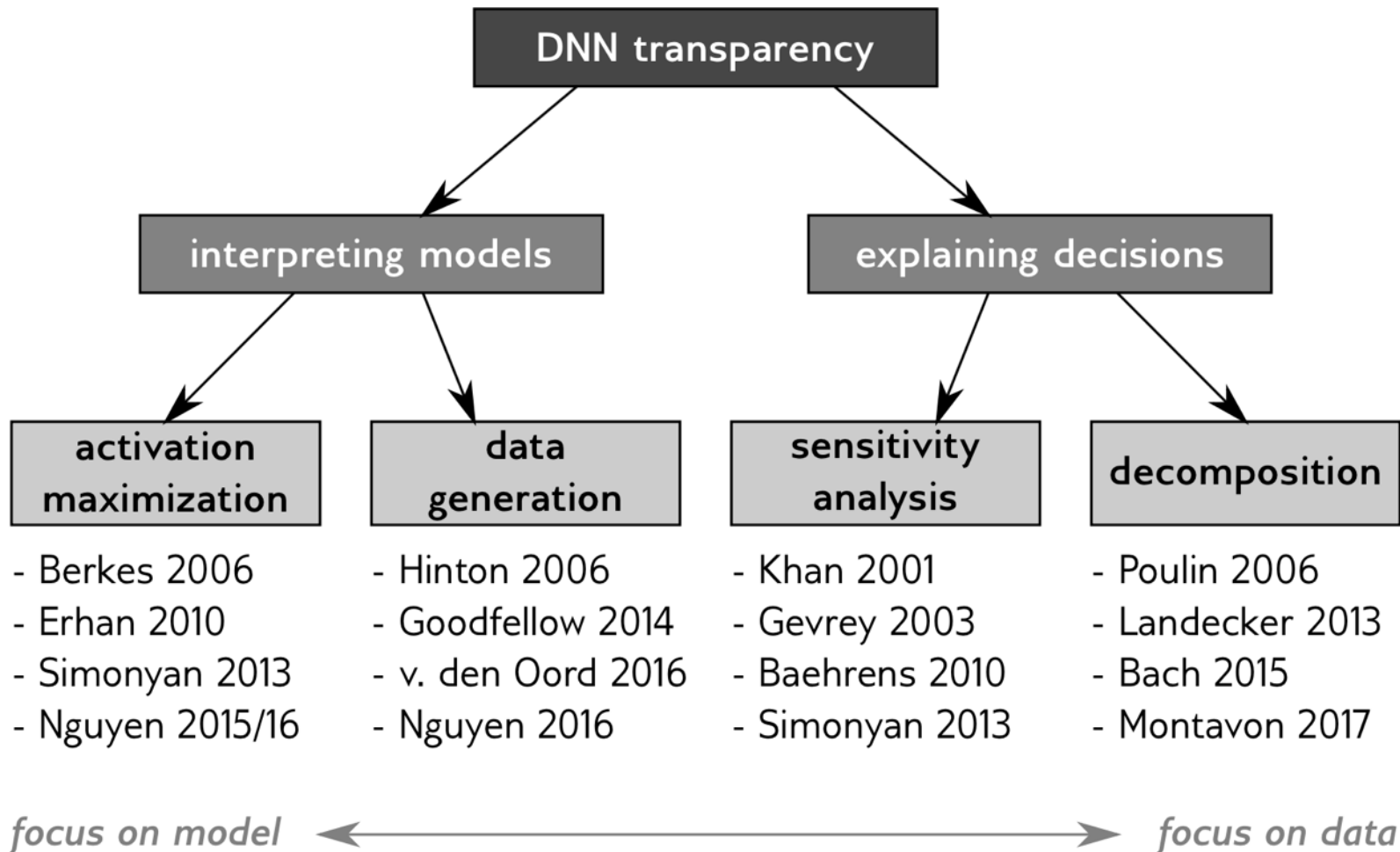
5) Compliance to legislation

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)



**better understand
internal representation**

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

Explaining decisions (individual)



**crucial for many
practical applications**

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

Techniques of Interpretation

In medical context

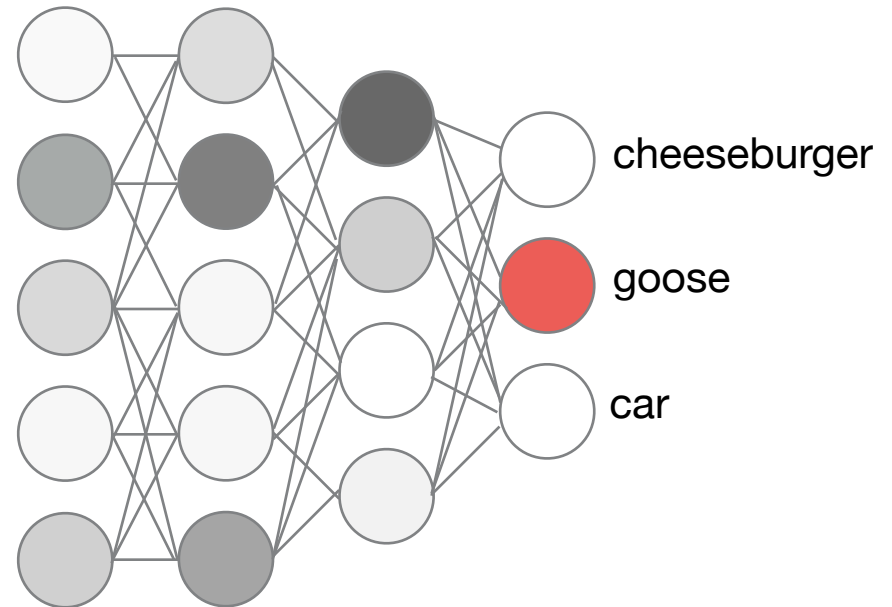
- Population view (ensemble)
 - Which symptoms are most common for the disease
 - Which drugs are most helpful for patients
- Patient's view (individual)
 - Which particular symptoms does the patient have
 - Which drugs does he need to take in order to recover

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*

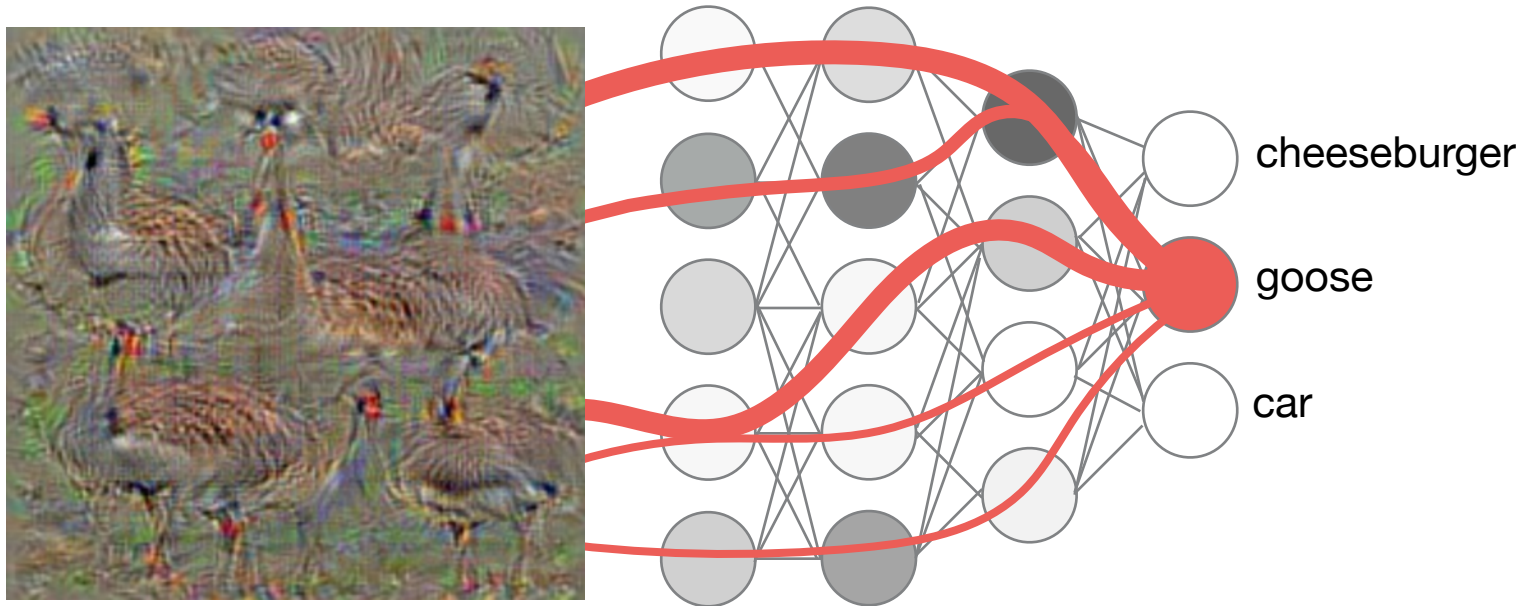


$$\max_{x \in \mathcal{X}} p_{\theta}(\omega_c | x) + \lambda \Omega(x)$$

Techniques of Interpretation

Interpreting models

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



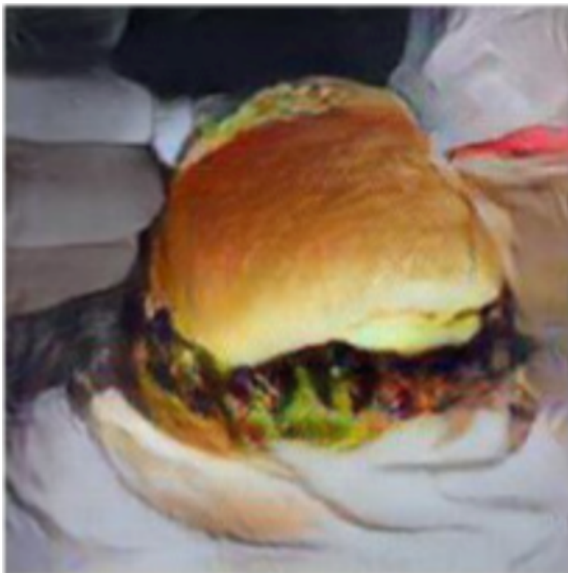
simple regularizer
(Simonyan et al. 2013)

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

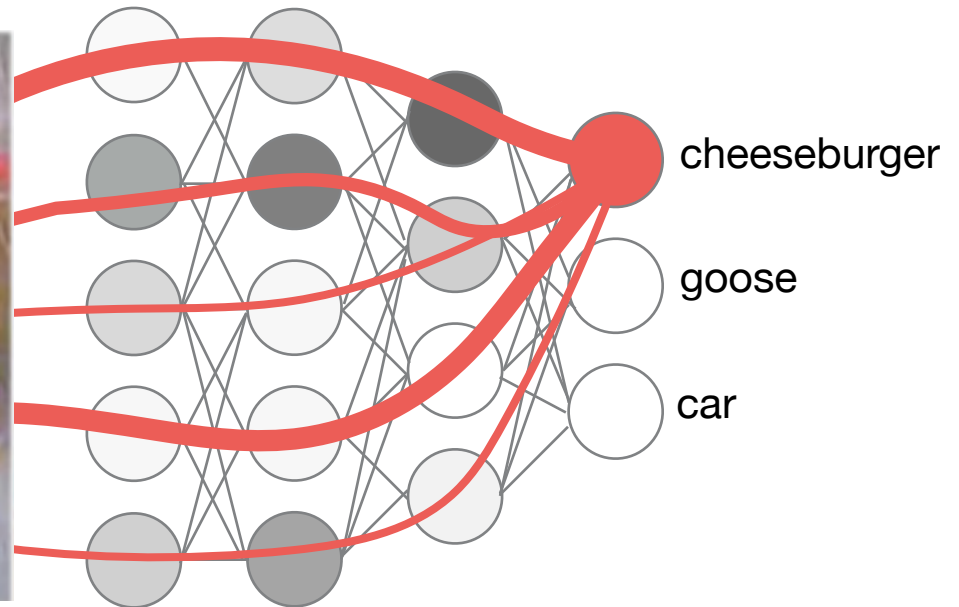
Techniques of Interpretation

Interpreting models

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



**complex regularizer
(Nguyen et al. 2016)**

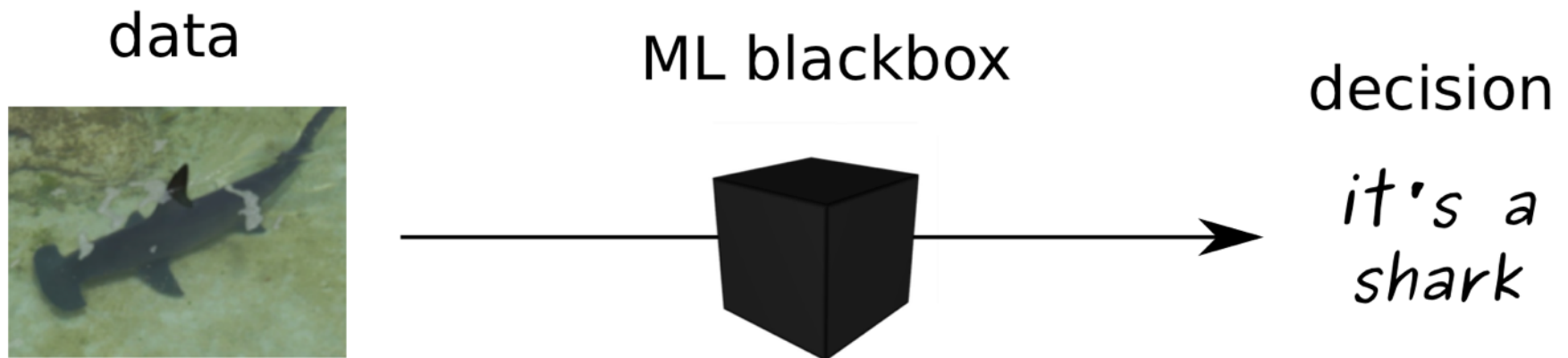


$$\max_{x \in \mathcal{X}} p_{\theta}(\omega_c | x) + \lambda \Omega(x)$$

Techniques of Interpretation

Explaining decisions

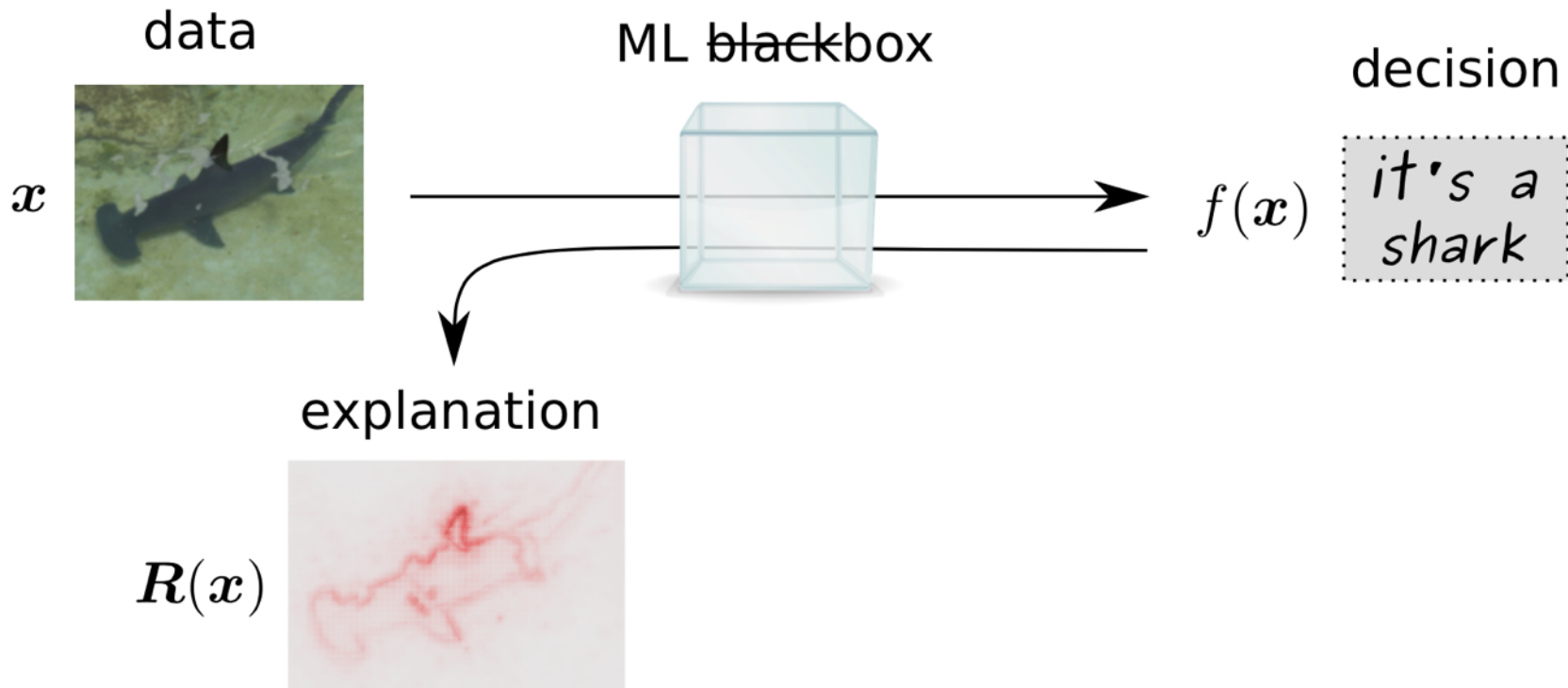
- “why” does the model arrive at a certain prediction
- verify that model behaves as expected



Techniques of Interpretation

Explaining decisions

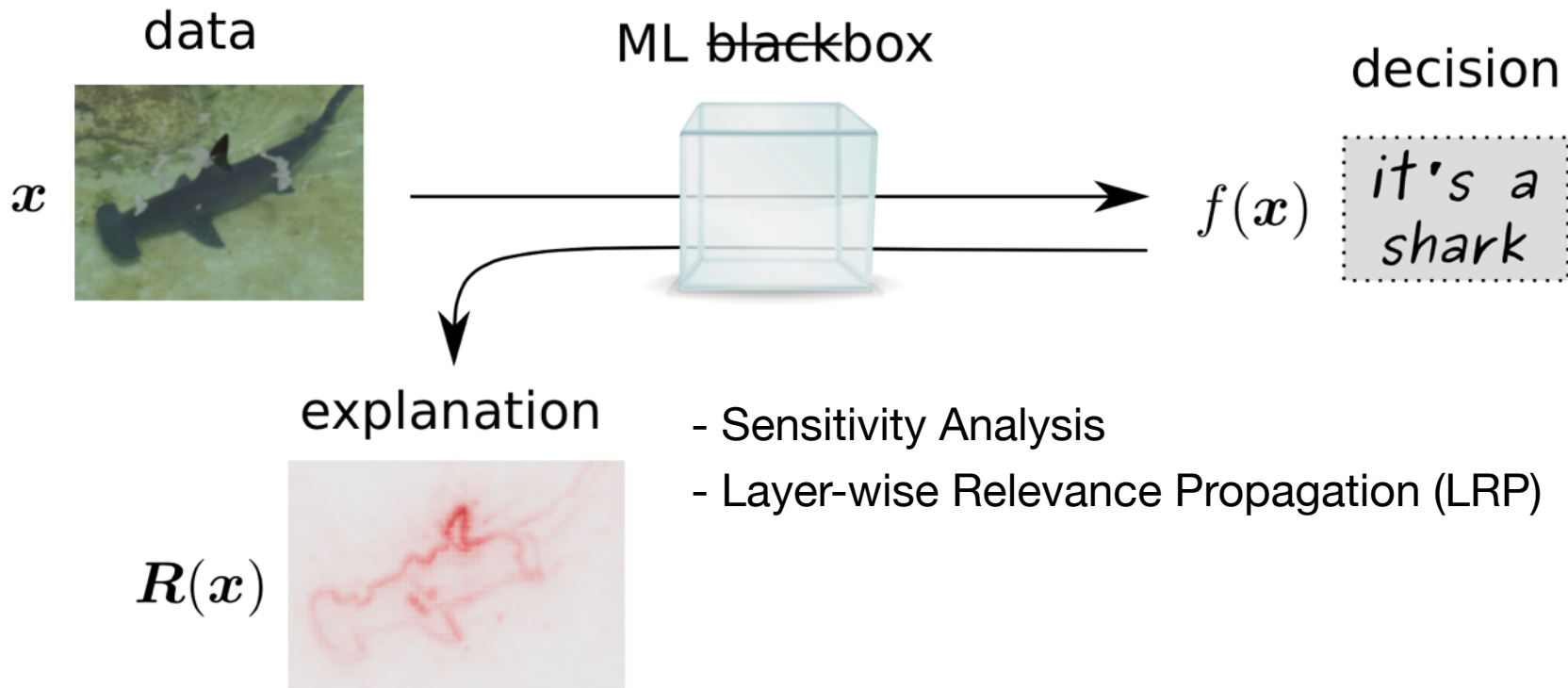
- “why” does the model arrive at a certain prediction
- verify that model behaves as expected



Techniques of Interpretation

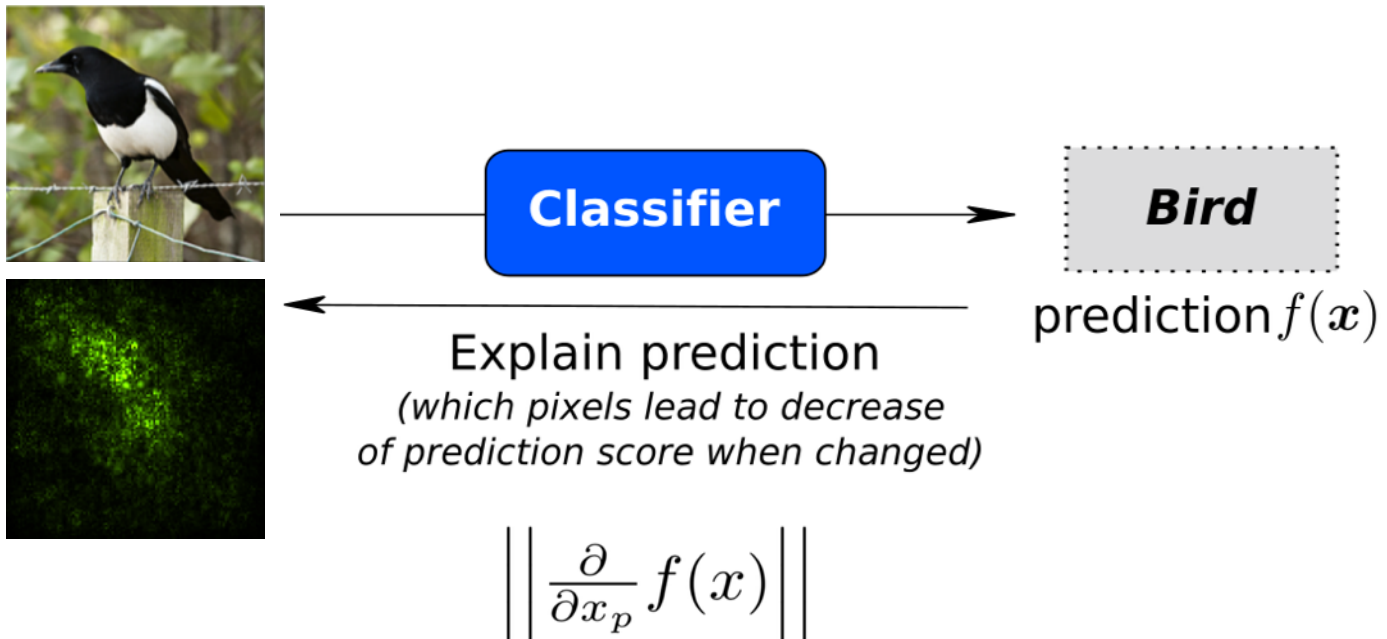
Explaining decisions

- “why” does the model arrive at a certain prediction
- verify that model behaves as expected



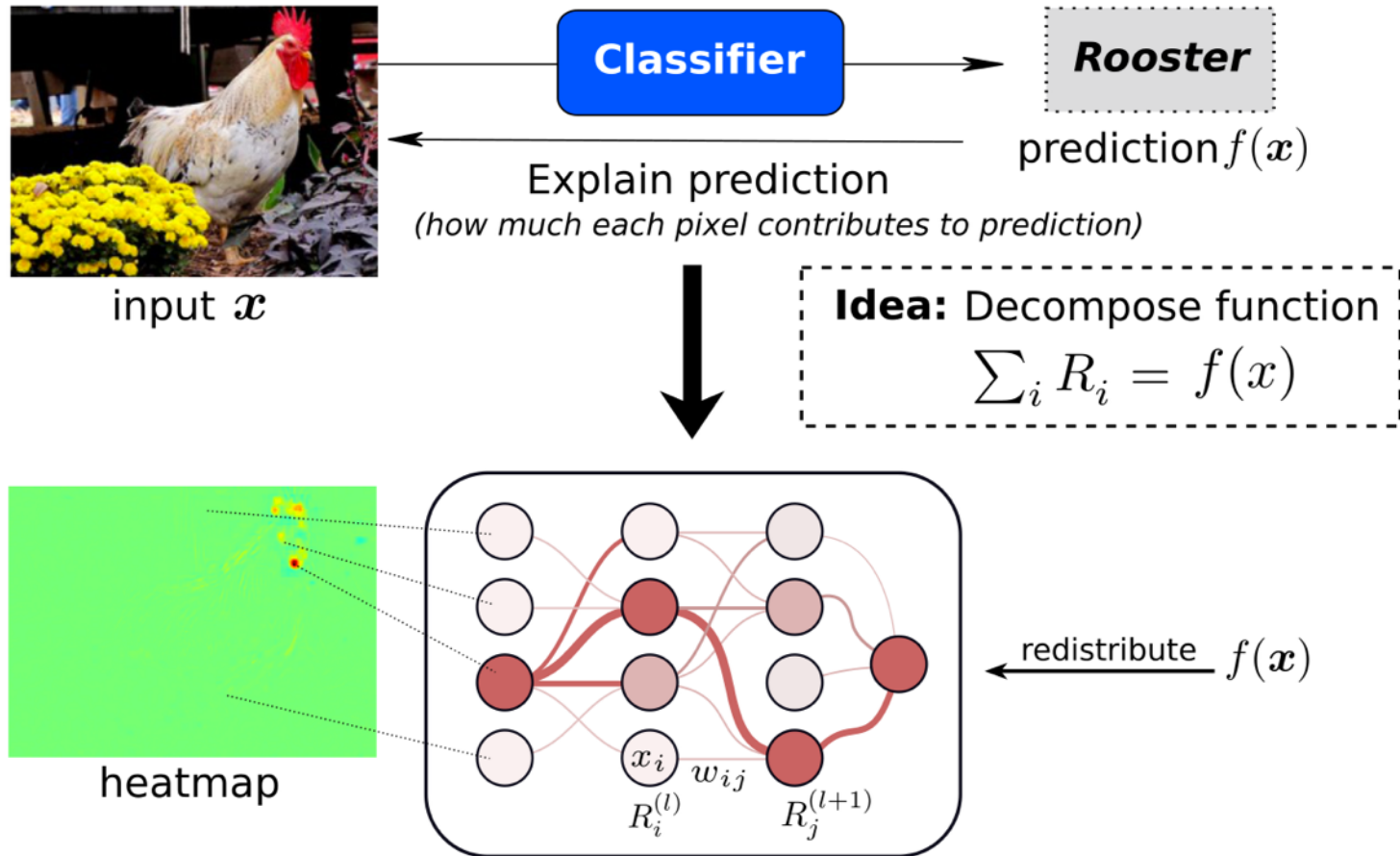
Techniques of Interpretation

Sensitivity Analysis
(Simonyan et al. 2014)



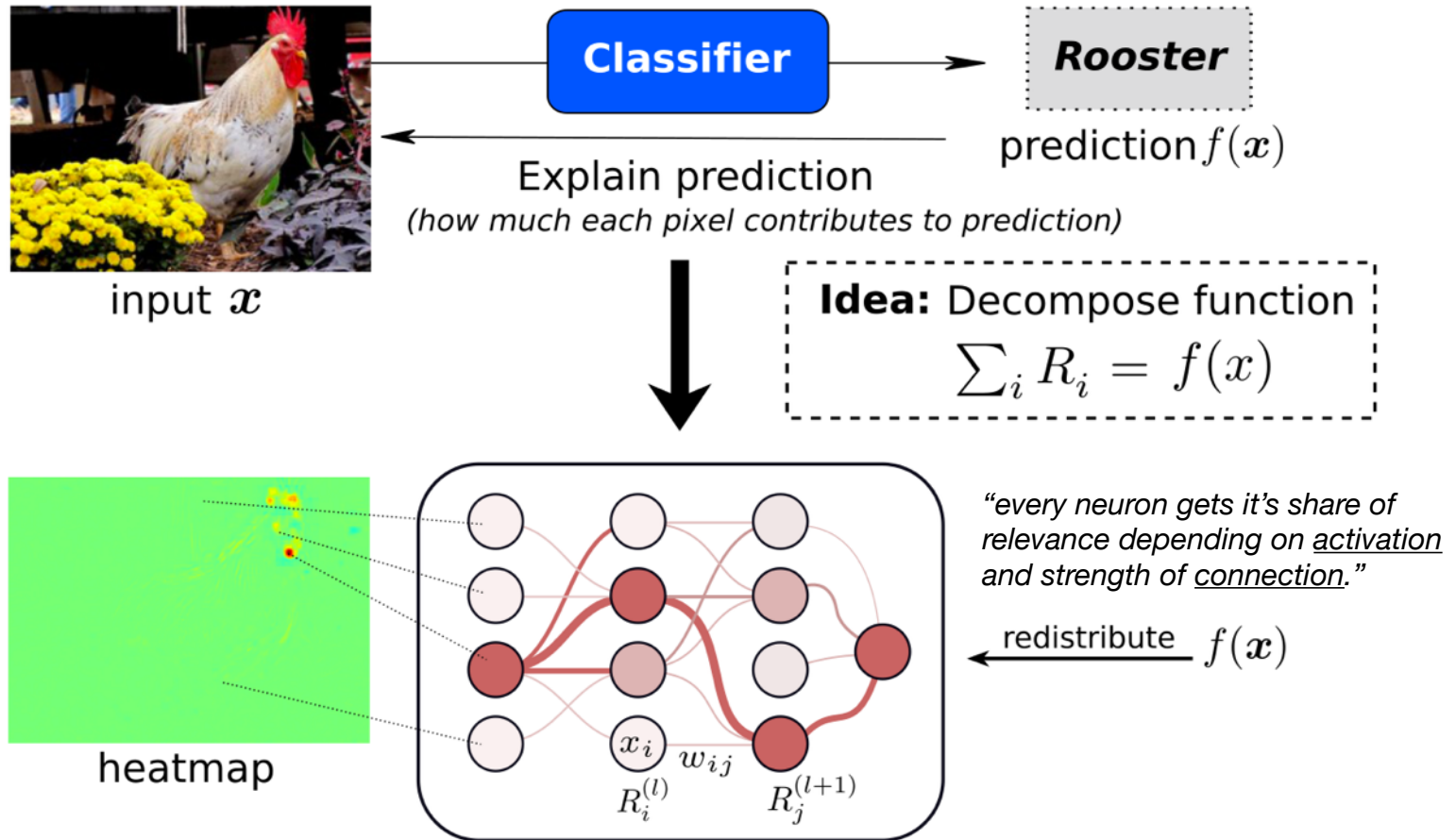
Techniques of Interpretation

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



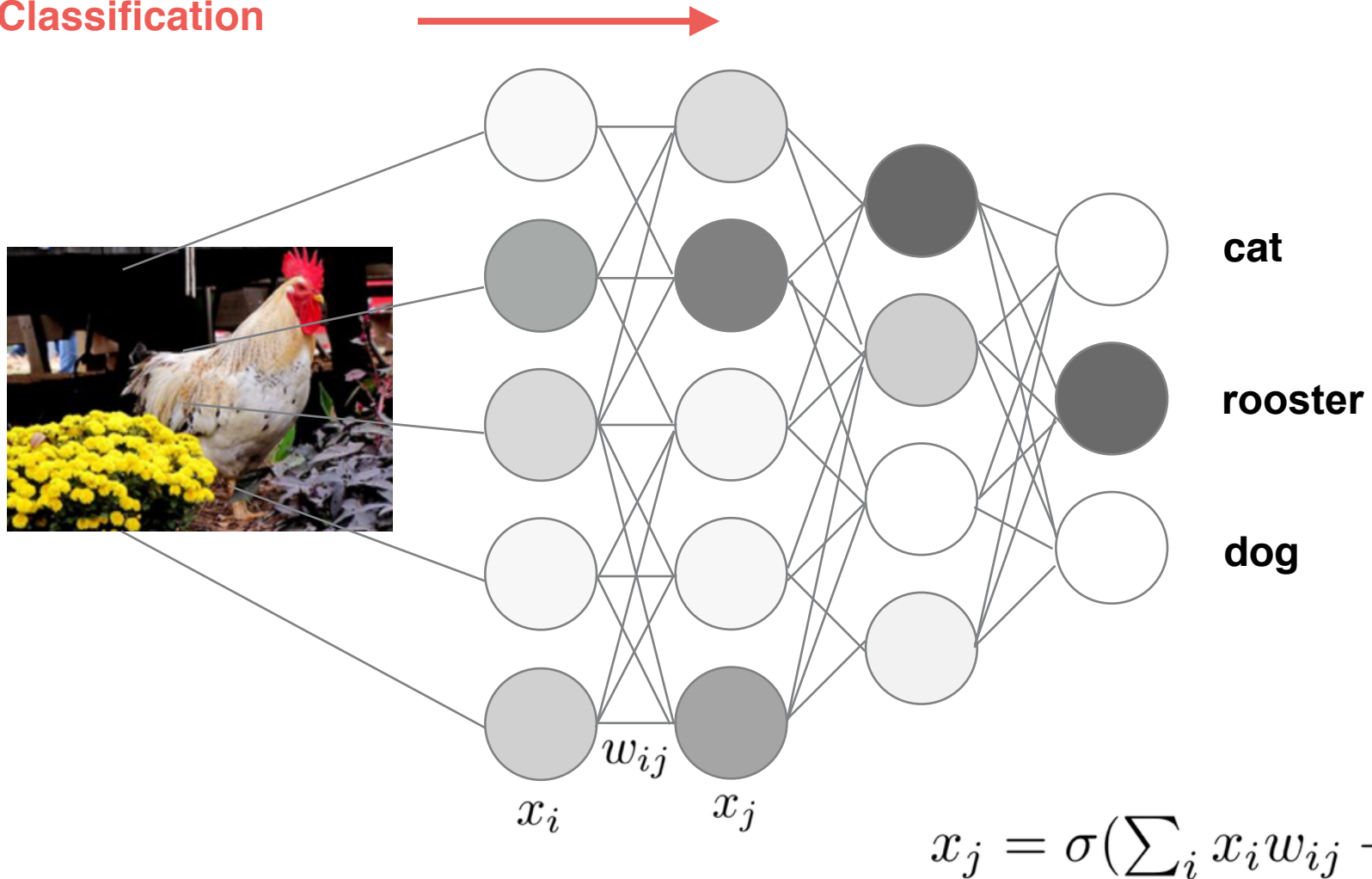
Techniques of Interpretation

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



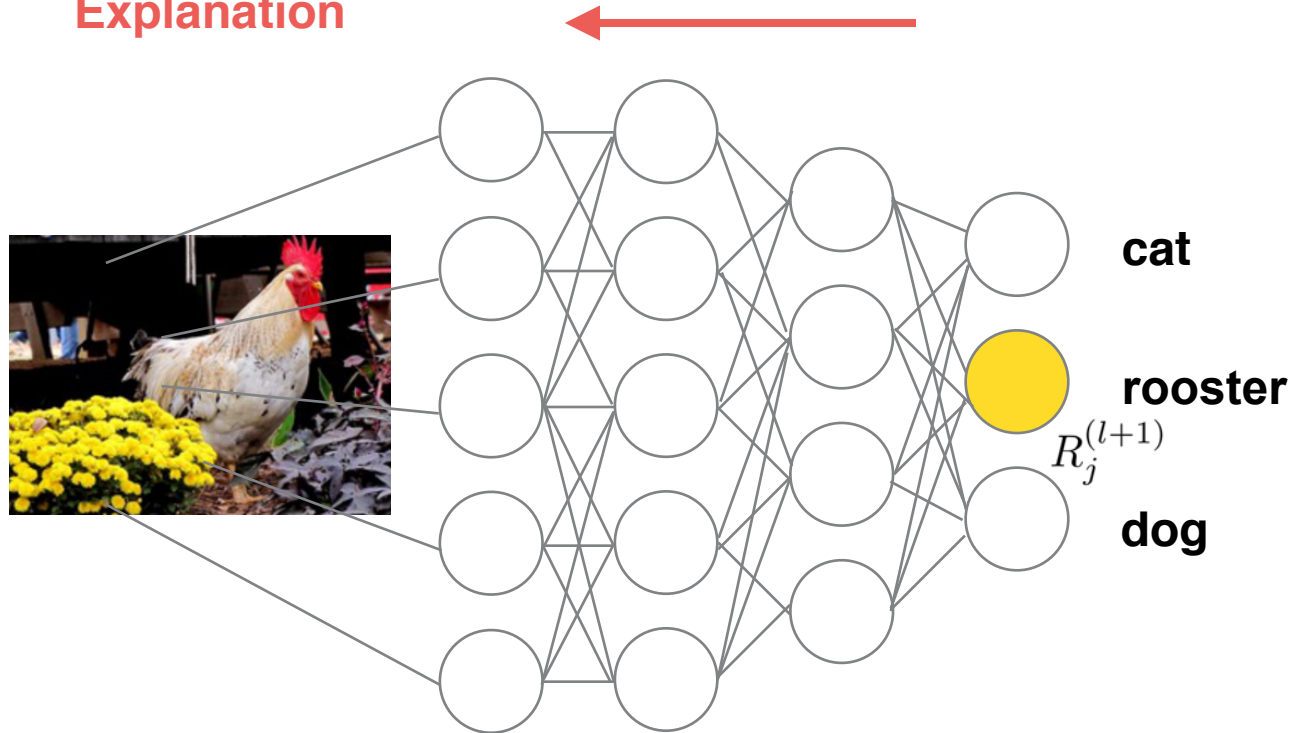
Techniques of Interpretation

Classification



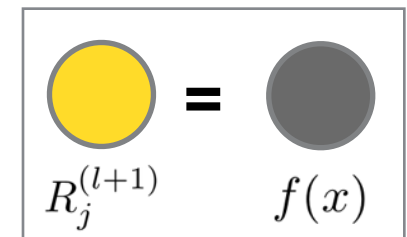
Techniques of Interpretation

Explanation



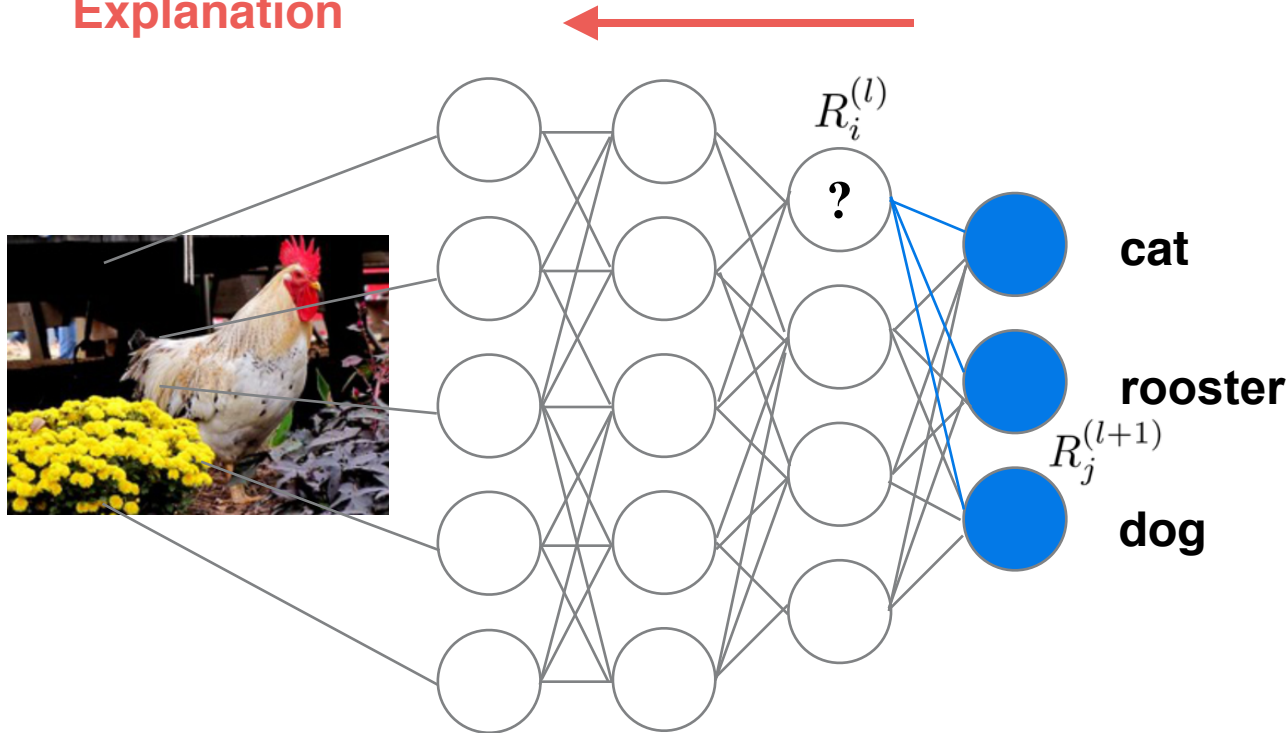
Idea: Backpropagate “relevances”

Initialization



Techniques of Interpretation

Explanation



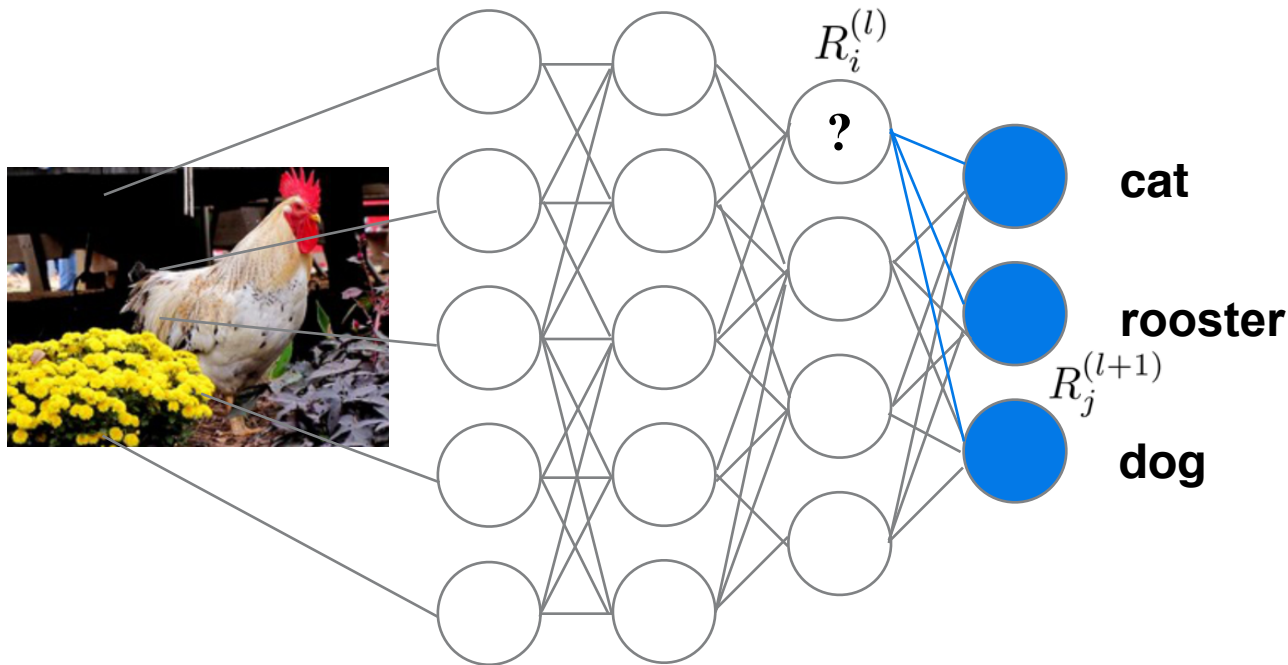
Simple LRP rule (Bach et al. 2015)

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

Every neuron gets its "share" of the redistributed relevance

Techniques of Interpretation

Explanation



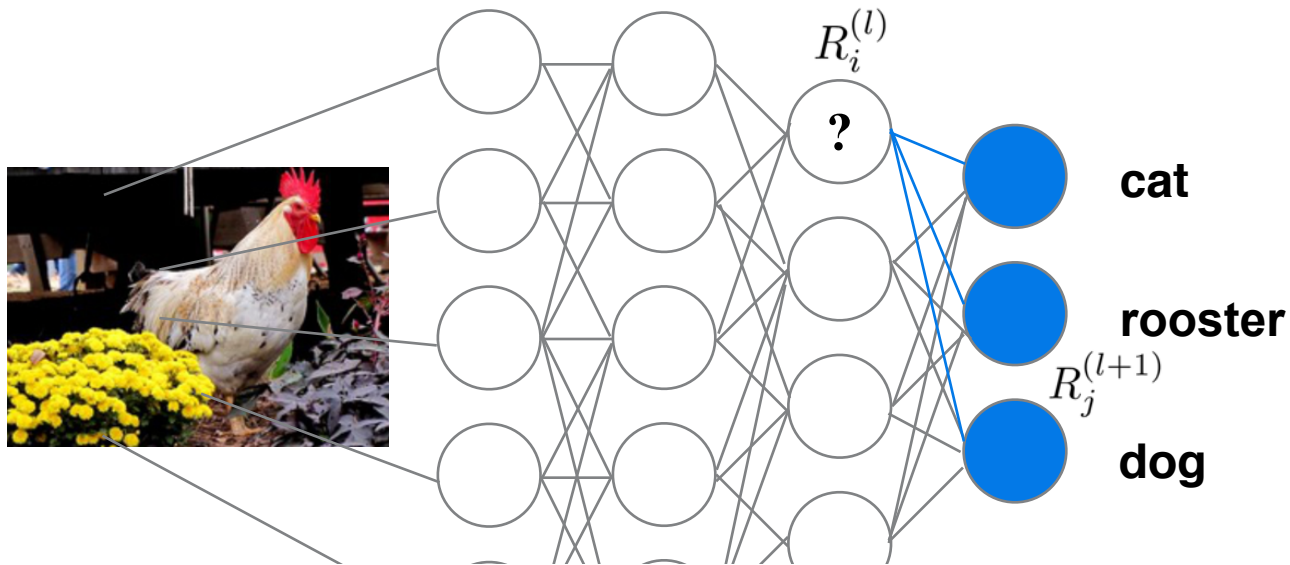
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

Techniques of Interpretation

Explanation



special case
 $\alpha = 1, \beta = 0$

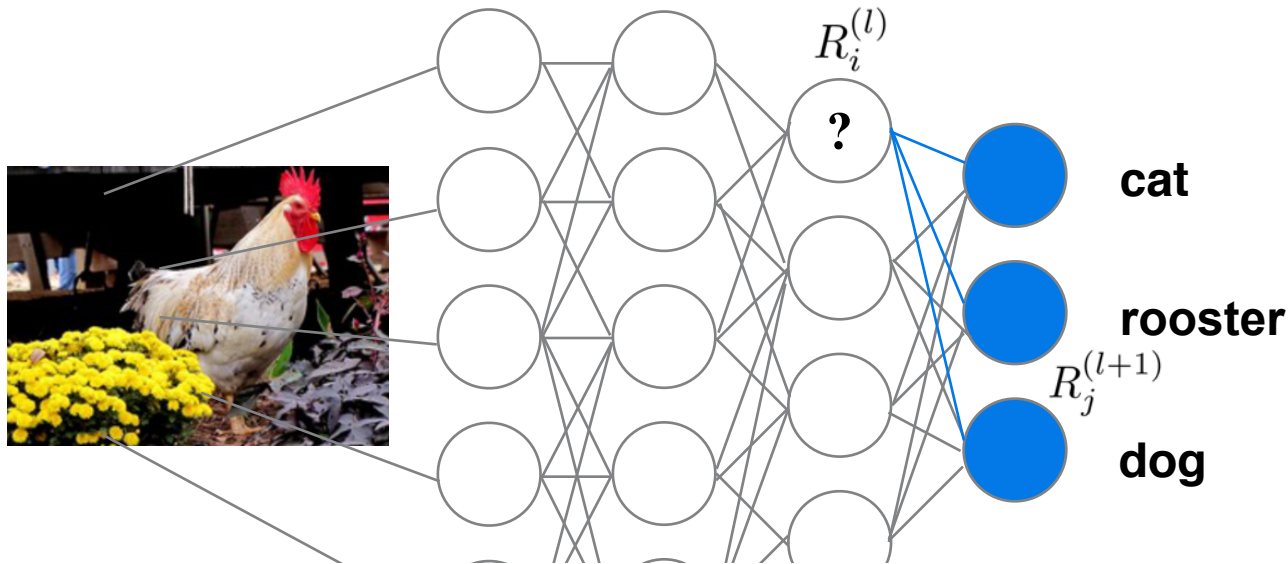
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

Techniques of Interpretation

Explanation



special case
 $\alpha = 1, \beta = 0$

Equivalent to redistribution rule proposed in
Excitation Backprop (Zhang et al., 2016)

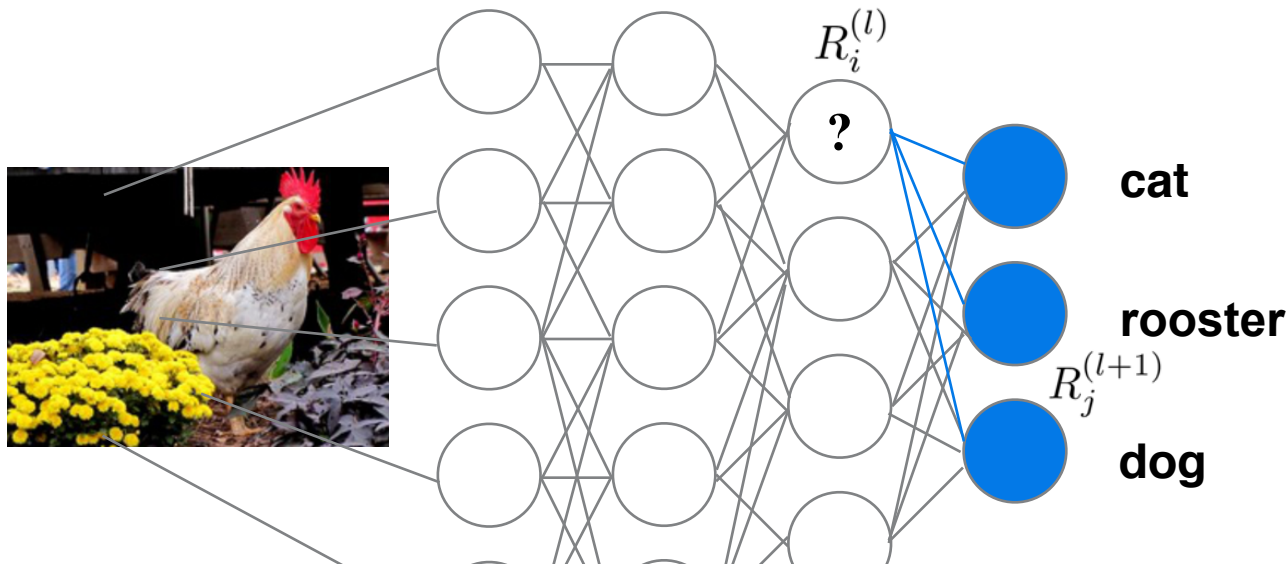
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

Techniques of Interpretation

Explanation



special case
 $\alpha = 1, \beta = 0$

Equivalent to redistribution rule proposed in
 Excitation Backprop (Zhang et al., 2016)

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

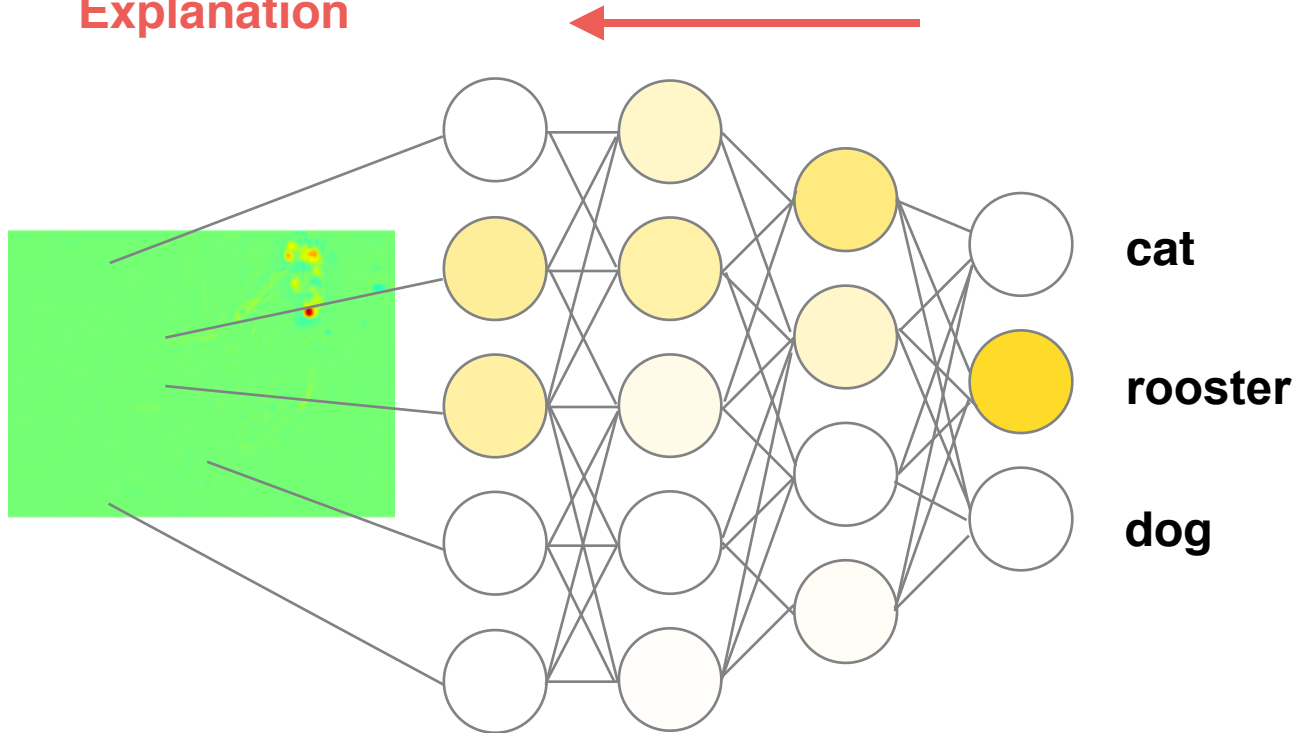
alpha-beta LRP rule (Bach et al. 2015)

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

where $\alpha + \beta = 1$

Techniques of Interpretation

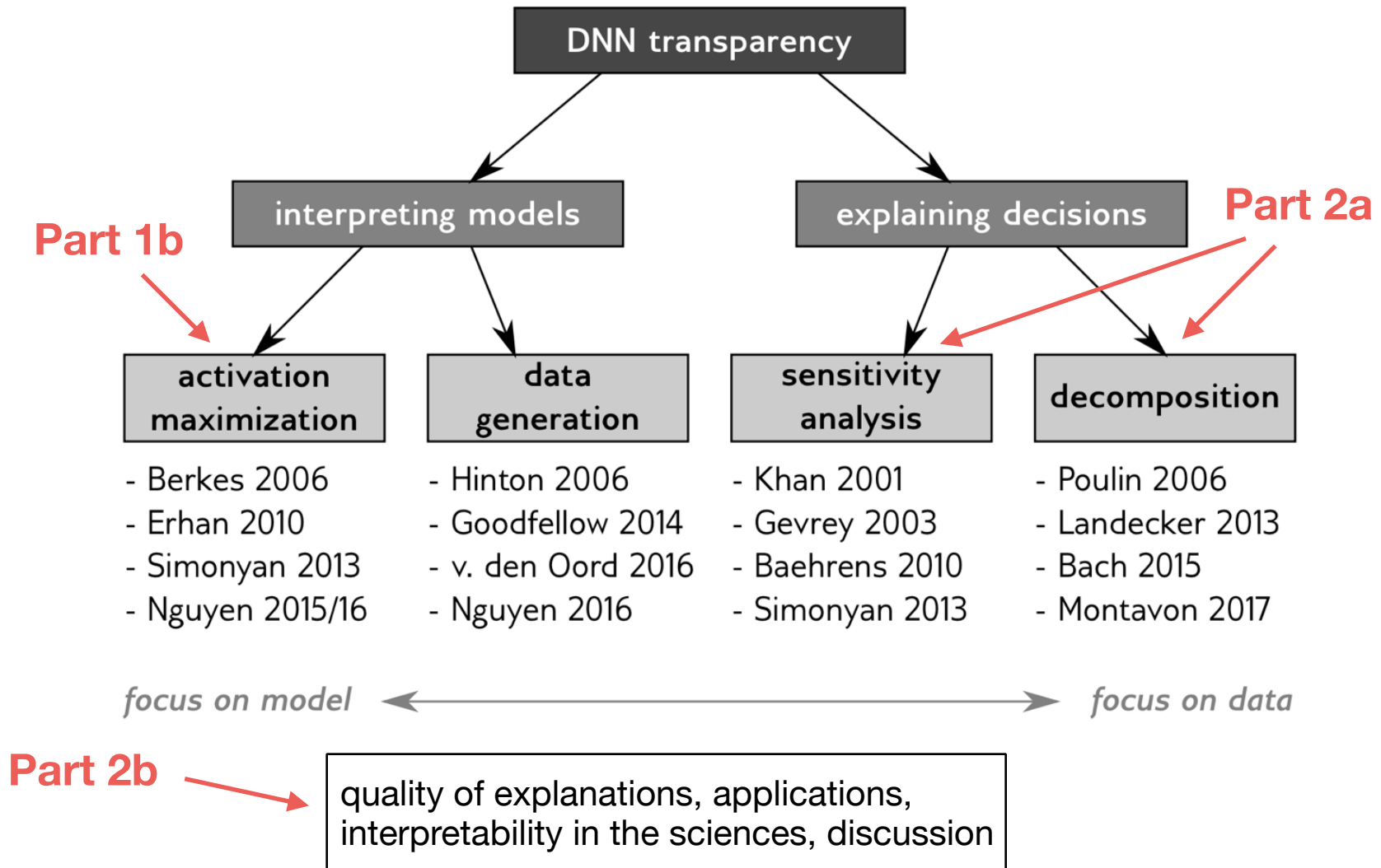
Explanation



Layer-wise relevance conservation

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

More to come



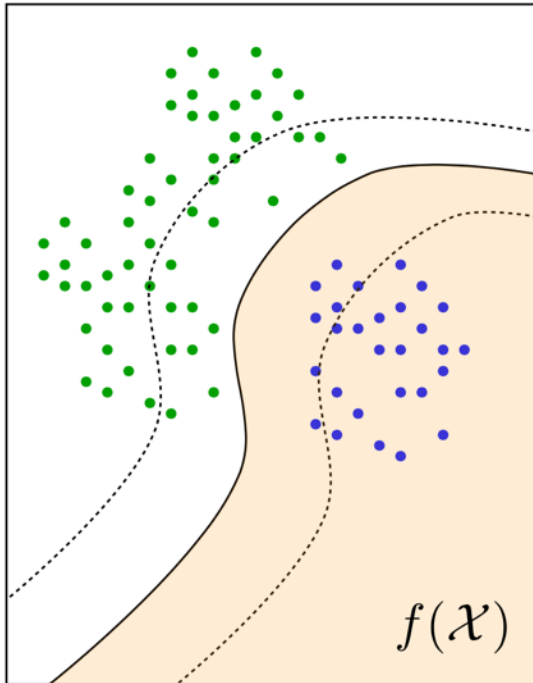
Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

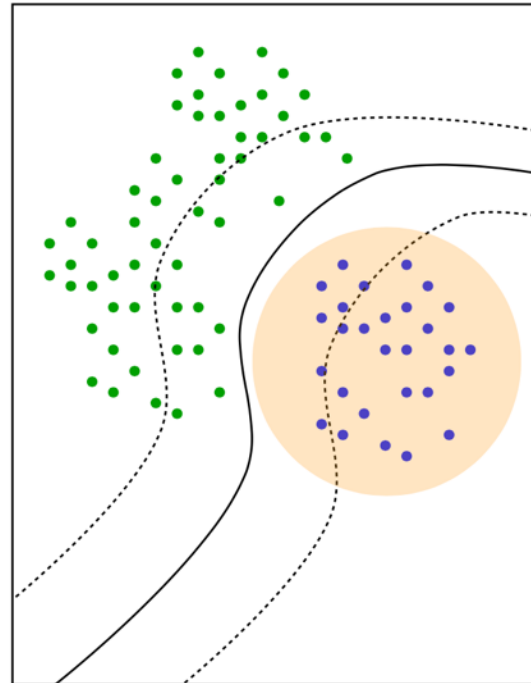
Part 1b: Making Deep Neural Networks Transparent

Making Deep Neural Nets Transparent

model analysis

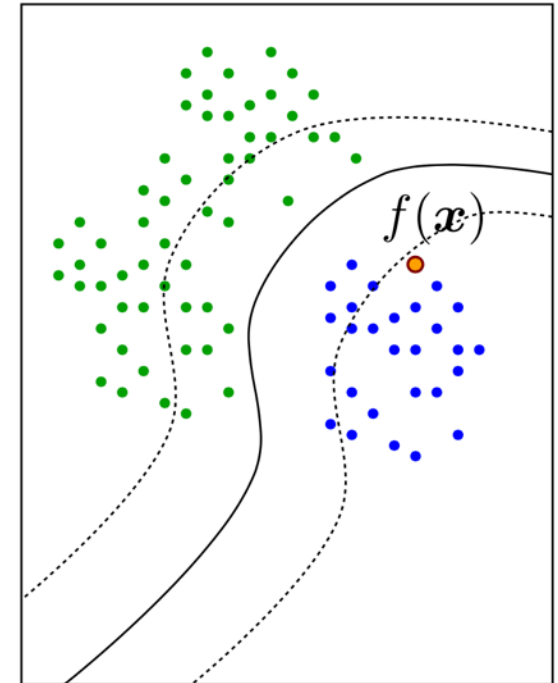


- visualizing filters
- max. class activation



- include distribution (RBM, DGN, etc.)

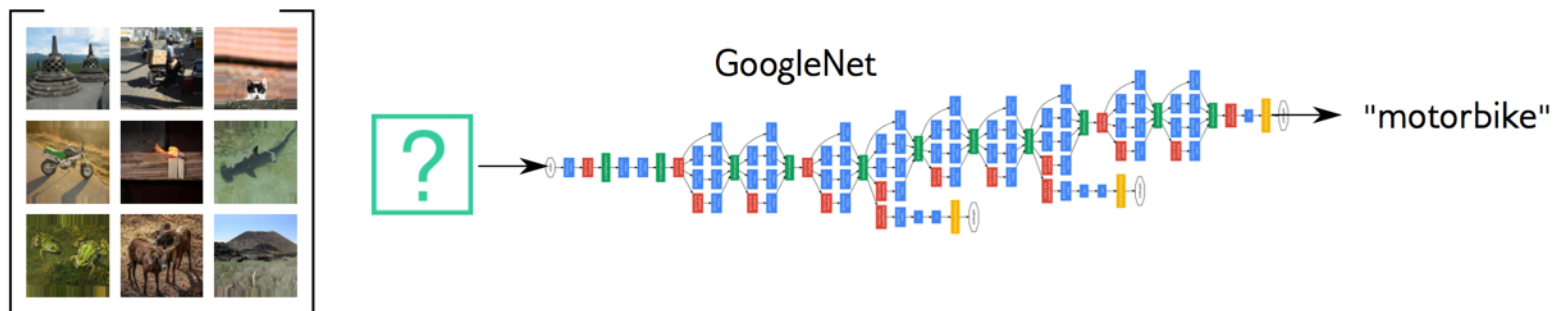
decision analysis



- sensitivity analysis
- decomposition

Interpreting Classes and Outputs

Image classification:



Question: How does a “motorbike” typically look like?

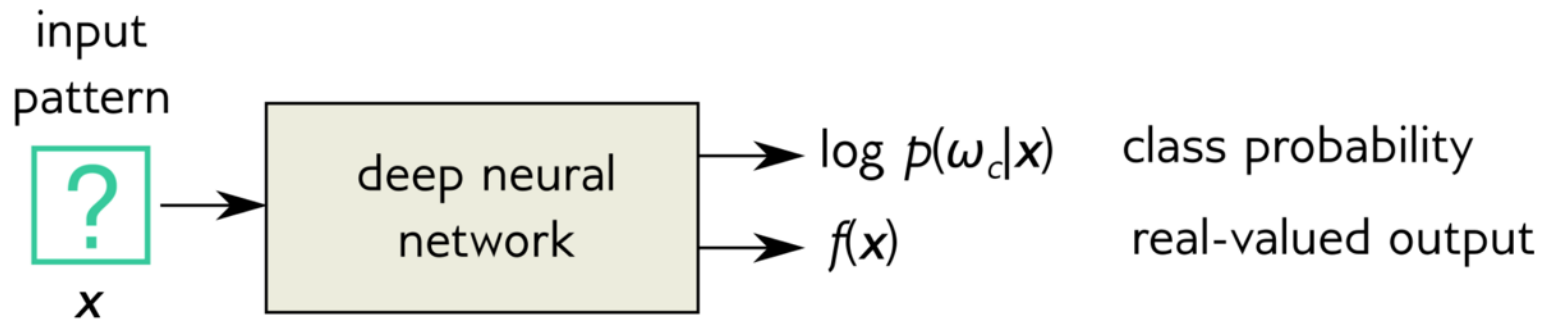
Quantum chemical calculations:



Question: How to interpret “ α high” in terms of molecular geometry?

The Activation Maximization (AM) Method

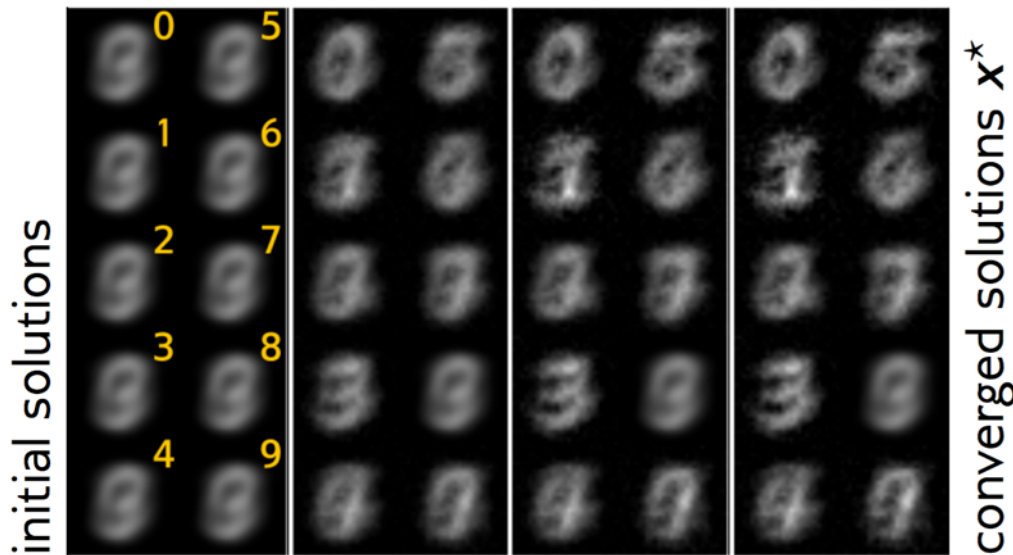
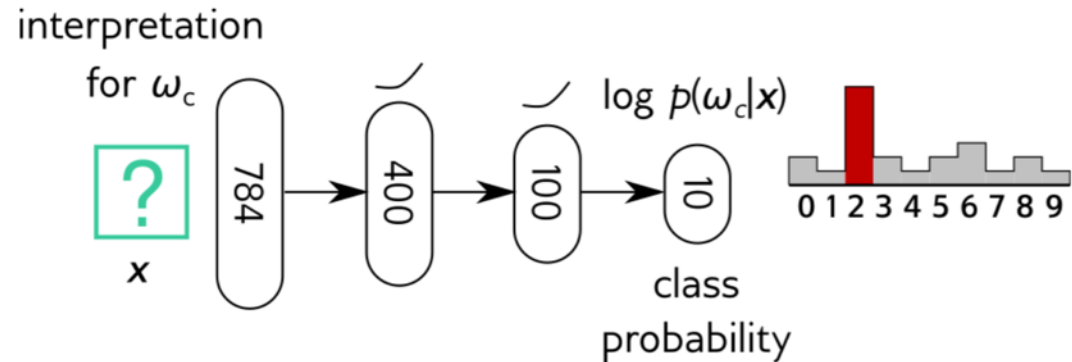
Let us interpret a concept predicted by a deep neural net (e.g. a class, or a real-valued quantity):



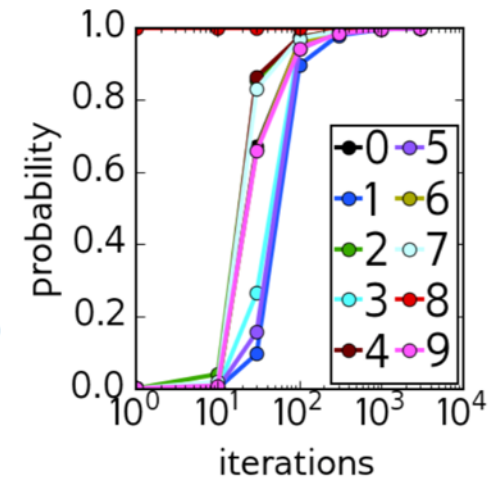
Examples:

- ▶ Creating a class prototype: $\max_{\mathbf{x} \in \mathcal{X}} \log p(\omega_c | \mathbf{x})$.
- ▶ Synthesizing an extreme case: $\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$.

Interpreting a Handwritten Digits Classifier



→ → optimizing $\max_x p(\omega_c|x)$ → →



Interpreting a DNN Image Classifier

goose



ostrich



Images from **Simonyan et al. 2013** “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”

Observations:

- ▶ AM builds typical patterns for these classes (e.g. beaks, legs).
- ▶ Unrelated background objects are not present in the image.

Improving Activation Maximization

Activation-maximization produces class-related patterns, but they are not resembling true data points. This can lower the quality of the interpretation for the predicted class ω_c .

Idea:

- ▶ Force the interpretation $\mathbf{x}^*$ to match the data more closely.

This can be achieved by redefining the optimization problem:

Find the input pattern that maximizes class probability.



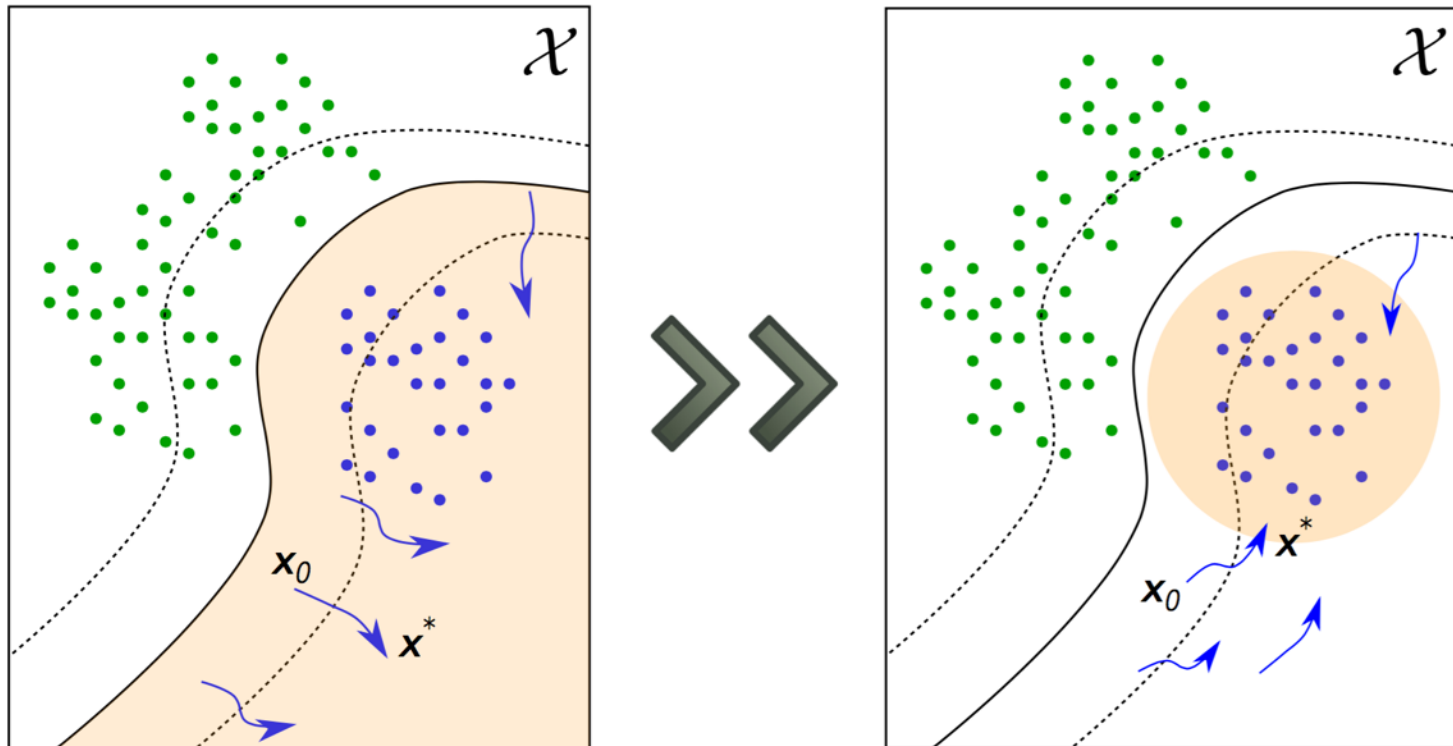
Find the most likely input pattern for a given class.

Improving Activation Maximization

Find the input pattern that maximizes class probability.



Find the most likely input pattern for a given class.



Improving Activation Maximization

Find the input pattern that maximizes class probability.

→

Find the most likely input pattern for a given class.

Nguyen et al. 2016 introduced several enhancements for activation maximization:

- ▶ Multiplying the objective by an expert $p(\mathbf{x})$:

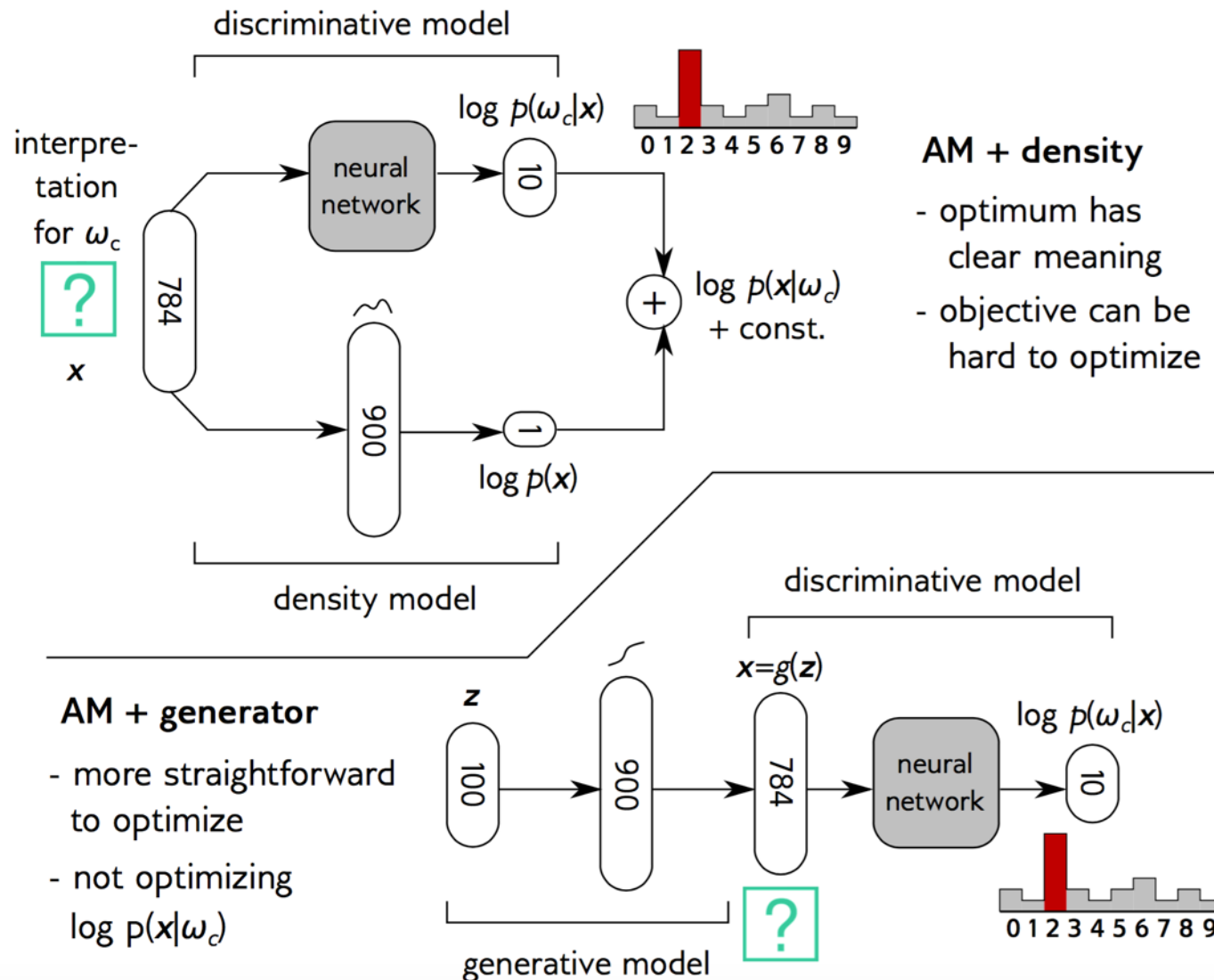
$$p(\mathbf{x}|\omega_c) \propto \underbrace{p(\omega_c|\mathbf{x})}_{\text{old}} \cdot p(\mathbf{x})$$

- ▶ Optimization in code space:

$$\max_{\mathbf{z} \in \mathcal{Z}} p(\omega_c | \underbrace{g(\mathbf{z})}_{\mathbf{x}}) + \lambda \|\mathbf{z}\|^2 \quad \mathbf{x}^* = g(\mathbf{z}^*)$$

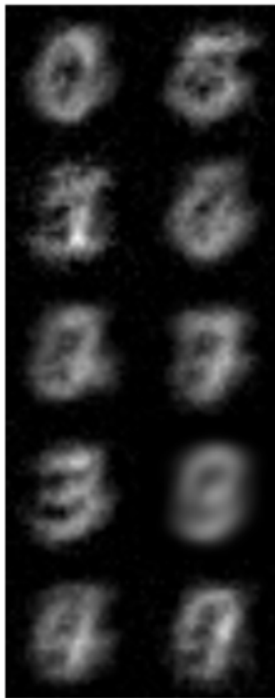
These two techniques require an unsupervised model of the data, either a density model $p(\mathbf{x})$ or a generator $g(\mathbf{z})$.

Improving Activation Maximization

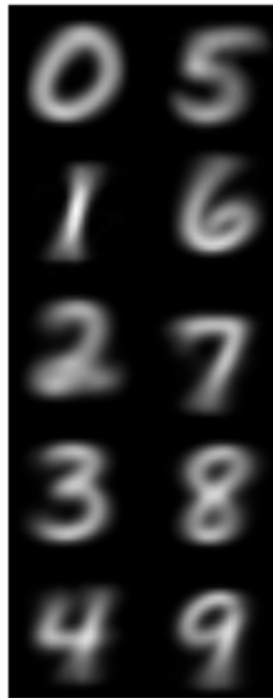


Comparison of Activation Maximization Variants

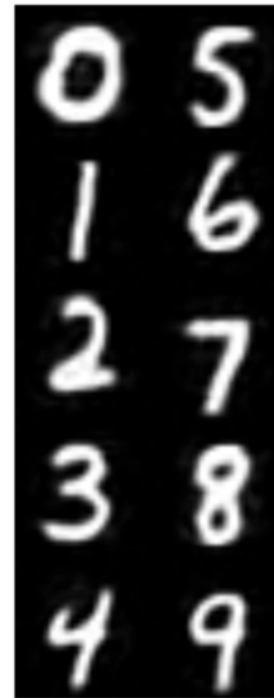
simple AM
(initialized
to mean)



simple AM
(init. to
class
means)



AM-density
(init. to
class
means)



AM-gen
(init. to
class
means)



Observation: Connecting to the data leads to sharper prototypes.

Enhanced AM on Natural Images

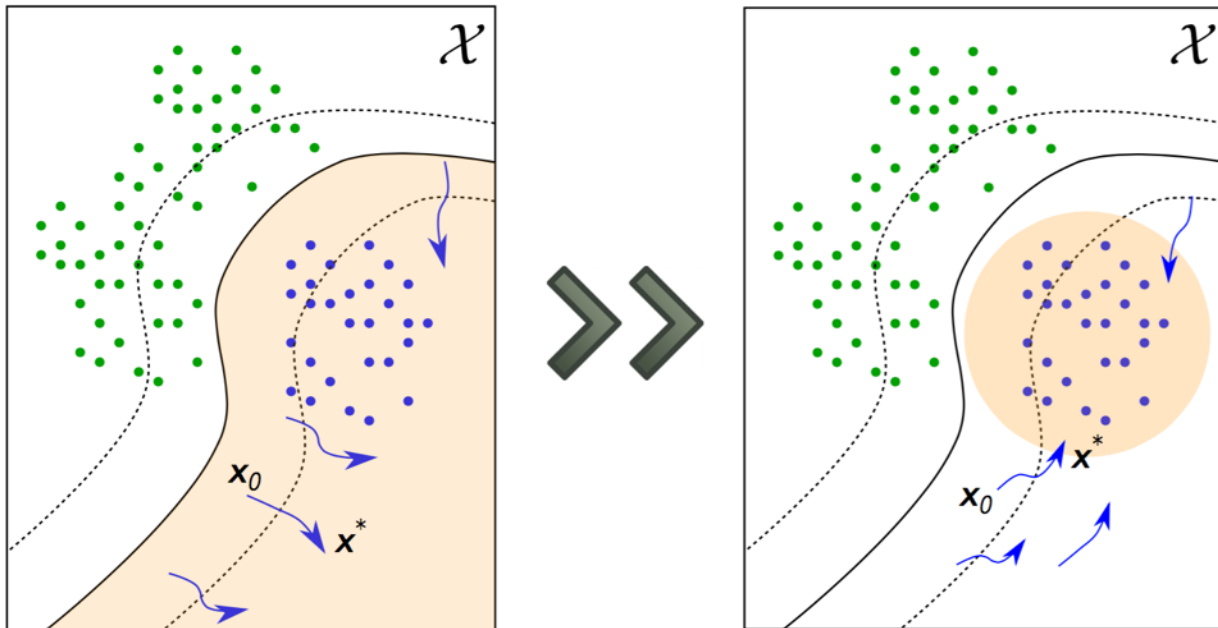
Images from **Nguyen et al. 2016**. “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”



Observation: Connecting AM to the data distribution leads to more realistic and more interpretable images.

Summary: Interpreting models

- ▶ Deep neural networks can be interpreted by finding input patterns that maximize a certain output quantity (e.g. class probability).
- ▶ Connecting to the data (e.g. by adding a generative or density model) improves the interpretability of the solution.



Limitations of Global Interpretations

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

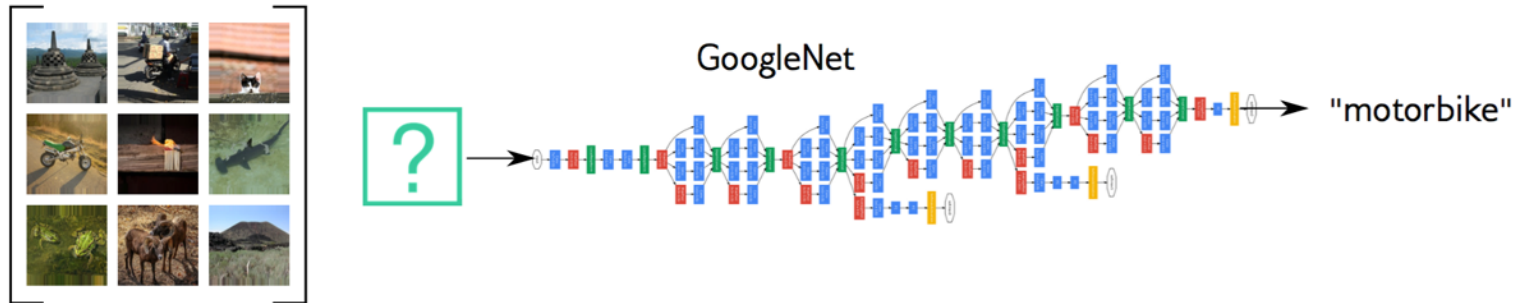


Observations:

- ▶ 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.

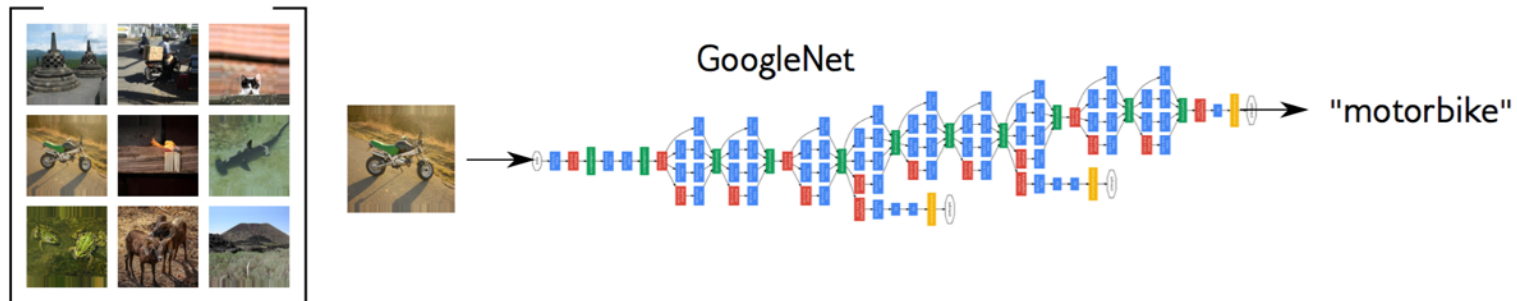
From Prototypes to Individual Explanations

Finding a prototype:



Question: How does a “motorbike” typically look like?

Individual explanation:



Question: Why is *this* example classified as a motorbike?

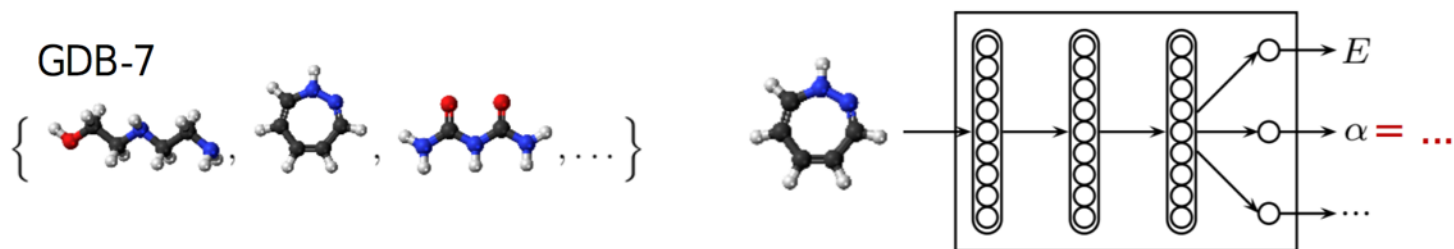
From Prototypes to Individual Explanations

Finding a prototype:



Question: How to interpret “ α high” in terms of molecular geometry?

Individual explanation:

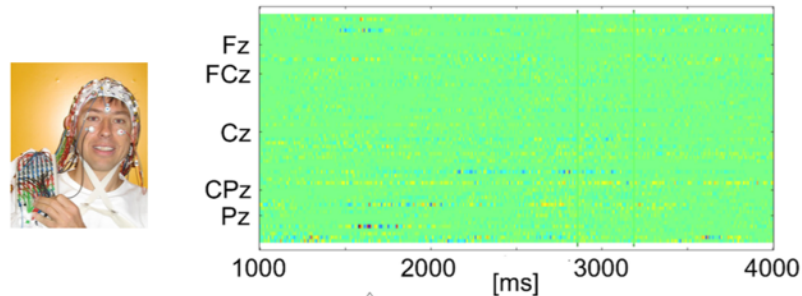


Question: Why α has a certain value for *this* molecule?

From Prototypes to Individual Explanations

Other examples where individual explanations are preferable to global interpretations:

- ▶ **Brain-computer interfaces:** Analyze input data for a *given* user at a *given* time in a *given* environment.

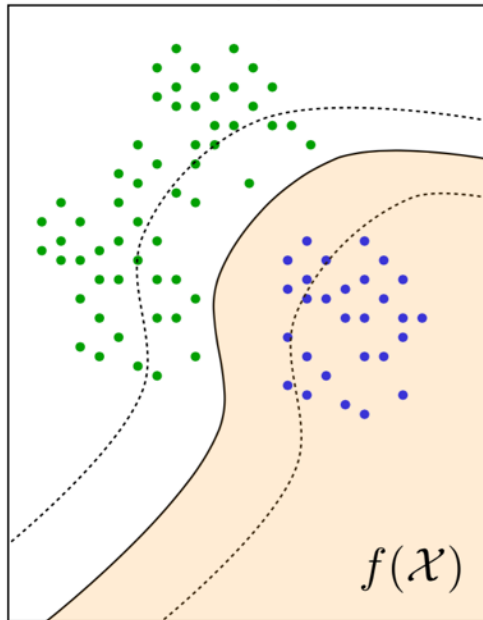


- ▶ **Personalized medicine:** Extracting the relevant information about a medical condition for a *given* patient at a *given* time.

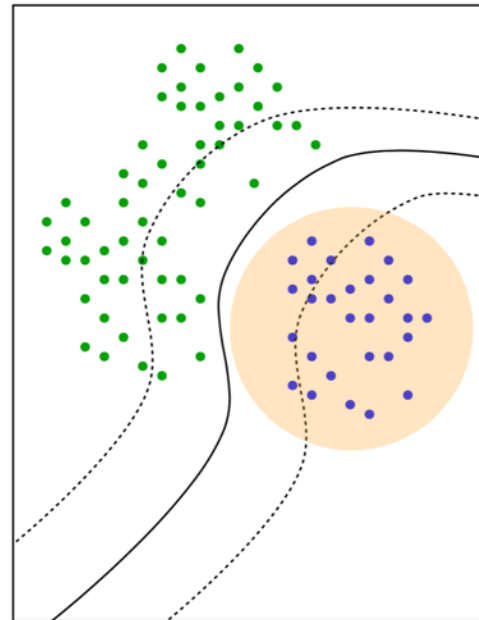
Each case is unique and needs its own explanation.

From Prototypes to Individual Explanations

model analysis

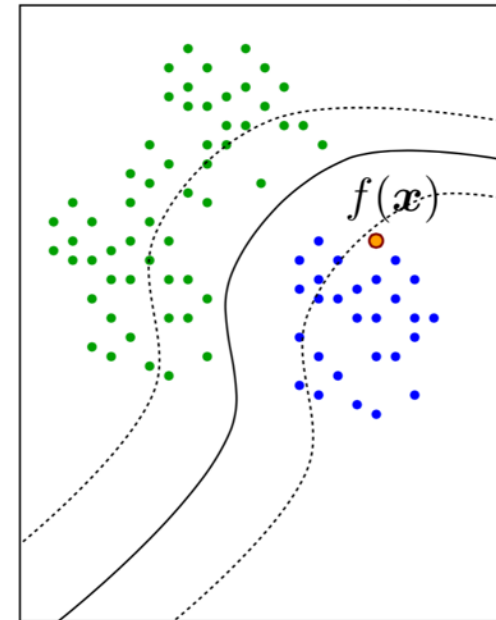


- visualizing filters
- max. class activation



- include distribution (RBM, DGN, etc.)

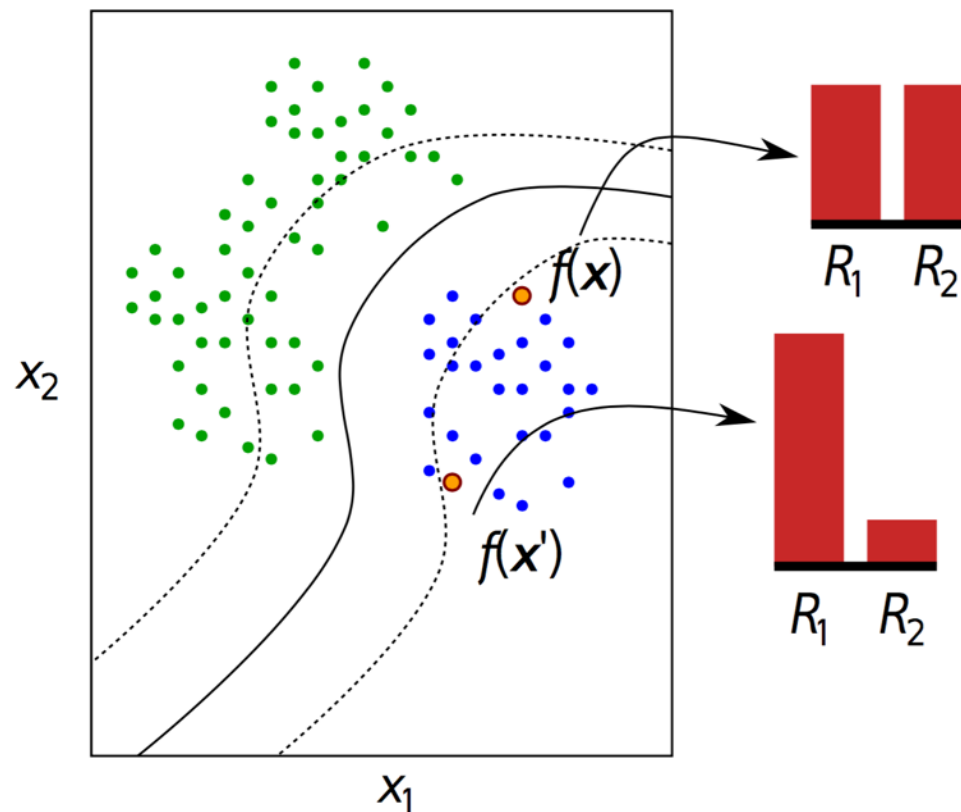
decision analysis



- sensitivity analysis
- decomposition

From Prototypes to Individual Explanations

Goal: Determine the relevance of each input variable for a given decision $f(x_1, x_2, \dots, x_d)$, by assigning to these variables *relevance scores* $R_1, R_2, \dots, R_d$.



Basic Technique: Sensitivity Analysis

Consider a function f , a data point $\mathbf{x} = (x_1, \dots, x_d)$, and the prediction

$$f(x_1, \dots, x_d).$$

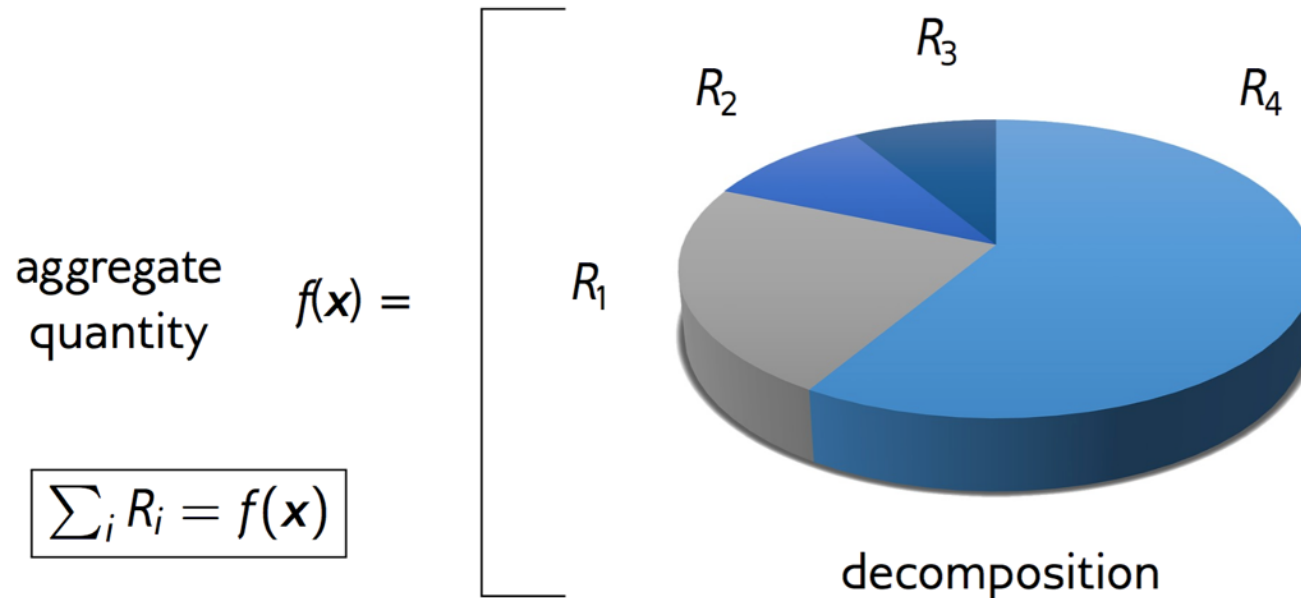
Sensitivity analysis measures the local variation of the function along each input dimension

$$R_i = \left(\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{x}} \right)^2$$

Remarks:

- ▶ Easy to implement (we only need access to the gradient of the decision function).
- ▶ But does it really explain the prediction?

Explaining by Decomposing



Examples:

- ▶ Economic activity (e.g. petroleum, cars, medicaments, ...)
- ▶ Energy production (e.g. coal, nuclear, hydraulic, ...)
- ▶ Evidence for object in an image (e.g. pixel 1, pixel 2, pixel 3, ...)
- ▶ Evidence for meaning in a text (e.g. word 1, word 2, word 3, ...)

What Does Sensitivity Analysis Decompose?

Sensitivity analysis

$$R_i = \left(\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}=\mathbf{x}} \right)^2$$

is a decomposition of the gradient norm $\|\nabla_{\mathbf{x}} f\|^2$.

Proof: $\sum_i R_i = \|\nabla_{\mathbf{x}} f\|^2$



**Sensitivity analysis explains
a *variation* of the function,
not the function value itself.**

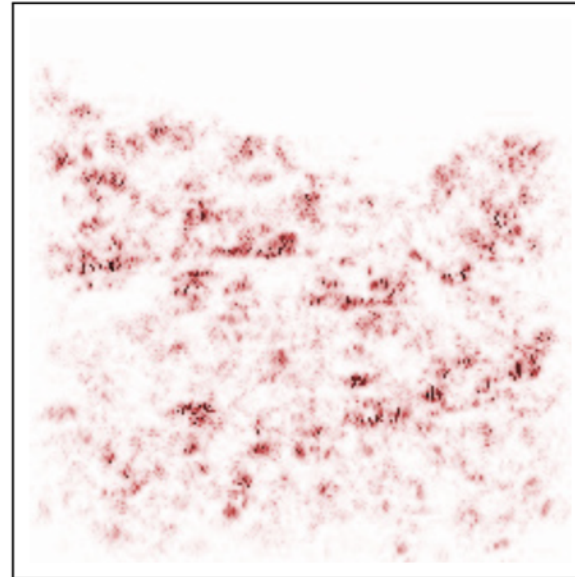
What Does Sensitivity Analysis Decompose?

Example: Sensitivity for class “car”

input image



sensitivity



- ▶ Relevant pixels are found both on cars and on the background.
- ▶ Explains what *reduces/increases* the evidence for cars rather what *is* the evidence for cars.

Decomposing the Correct Quantity

slope decomposition

$$\sum_i R_i = \|\nabla_{\mathbf{x}} f\|^2$$

value decomposition

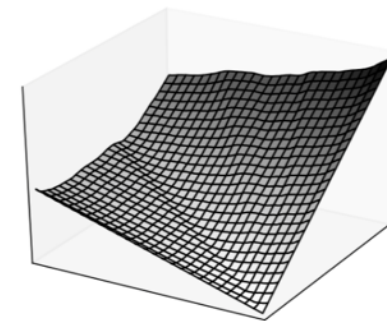
$$\sum_i R_i = f(\mathbf{x})$$

Candidate: Taylor decomposition

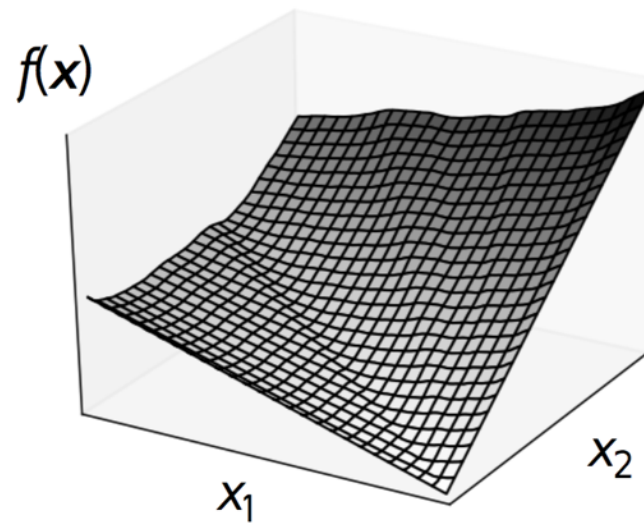
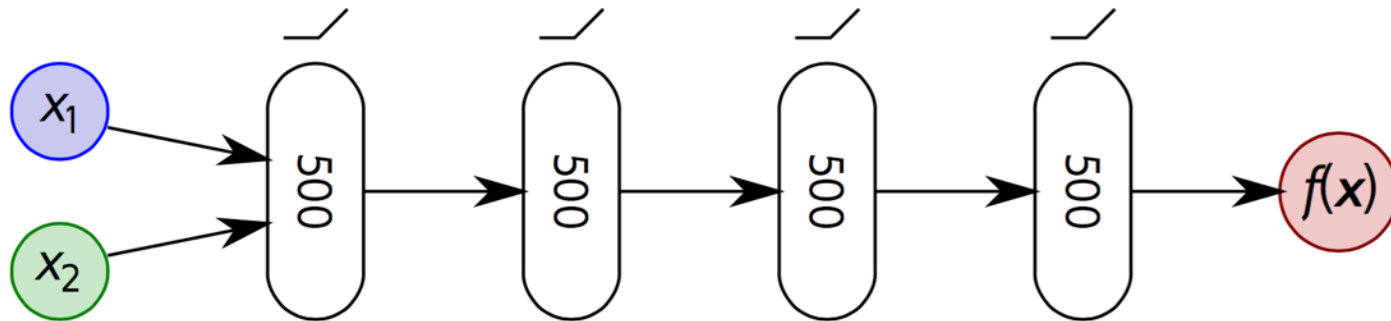
$$f(\mathbf{x}) = \underbrace{f(\tilde{\mathbf{x}})}_0 + \sum_{i=1}^d \underbrace{\frac{\partial f}{\partial x_i} \Big|_{\mathbf{x}=\tilde{\mathbf{x}}}}_{R_i} (x_i - \tilde{x}_i) + \underbrace{O(\mathbf{x}\mathbf{x}^T)}_0$$

- ▶ Achievable for linear models and deep ReLU networks without biases, by choosing:

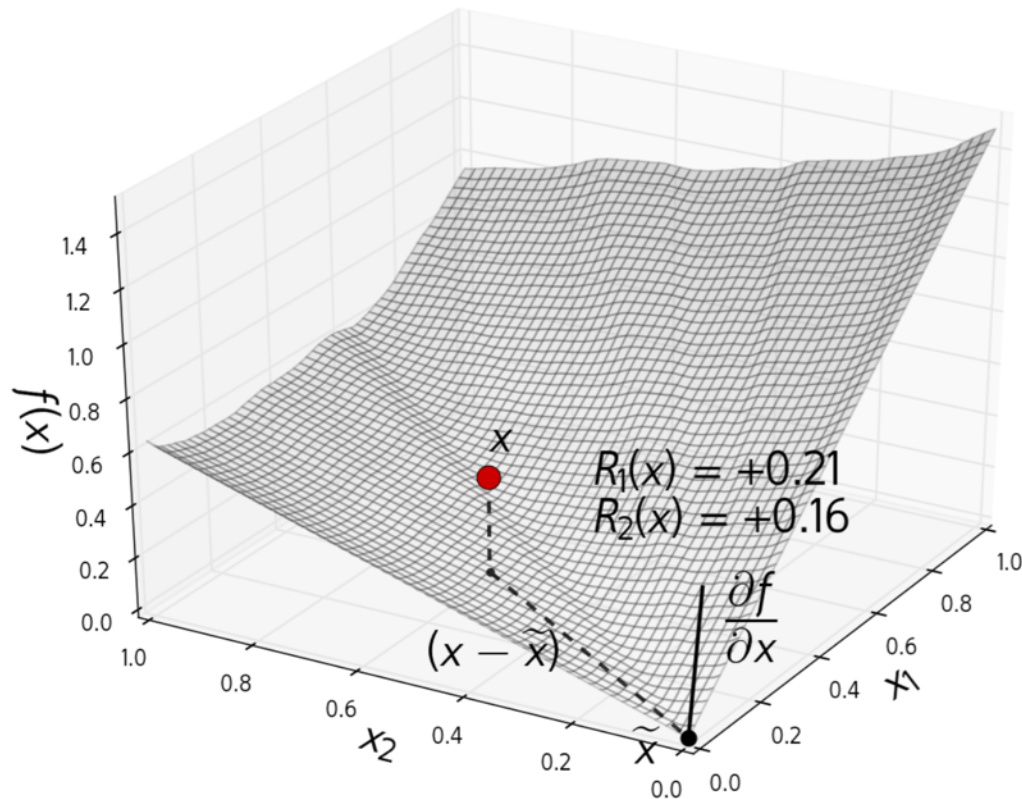
$$\tilde{\mathbf{x}} = \lim_{\varepsilon \rightarrow 0} \varepsilon \cdot \mathbf{x} \approx \mathbf{0}.$$



Experiment on a Randomly Initialized DNN

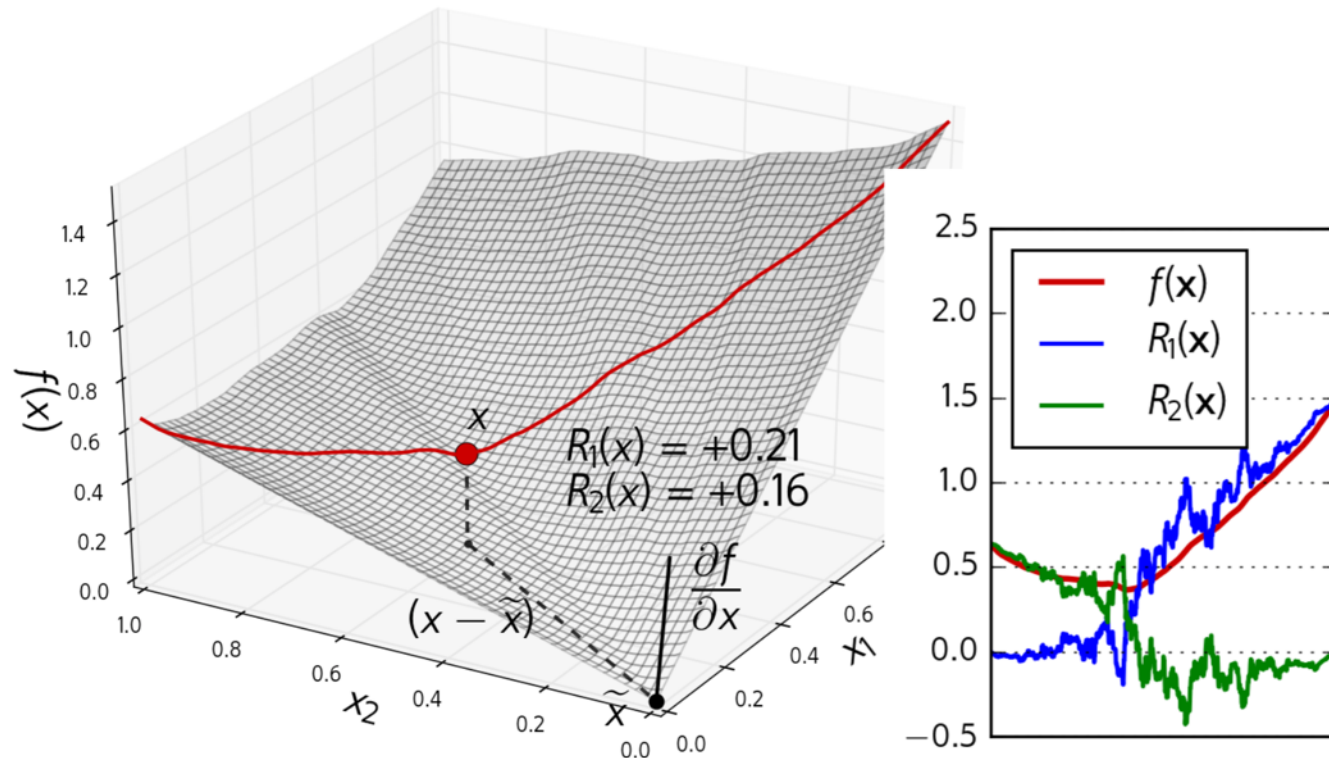


Decomposing the Output of the DNN



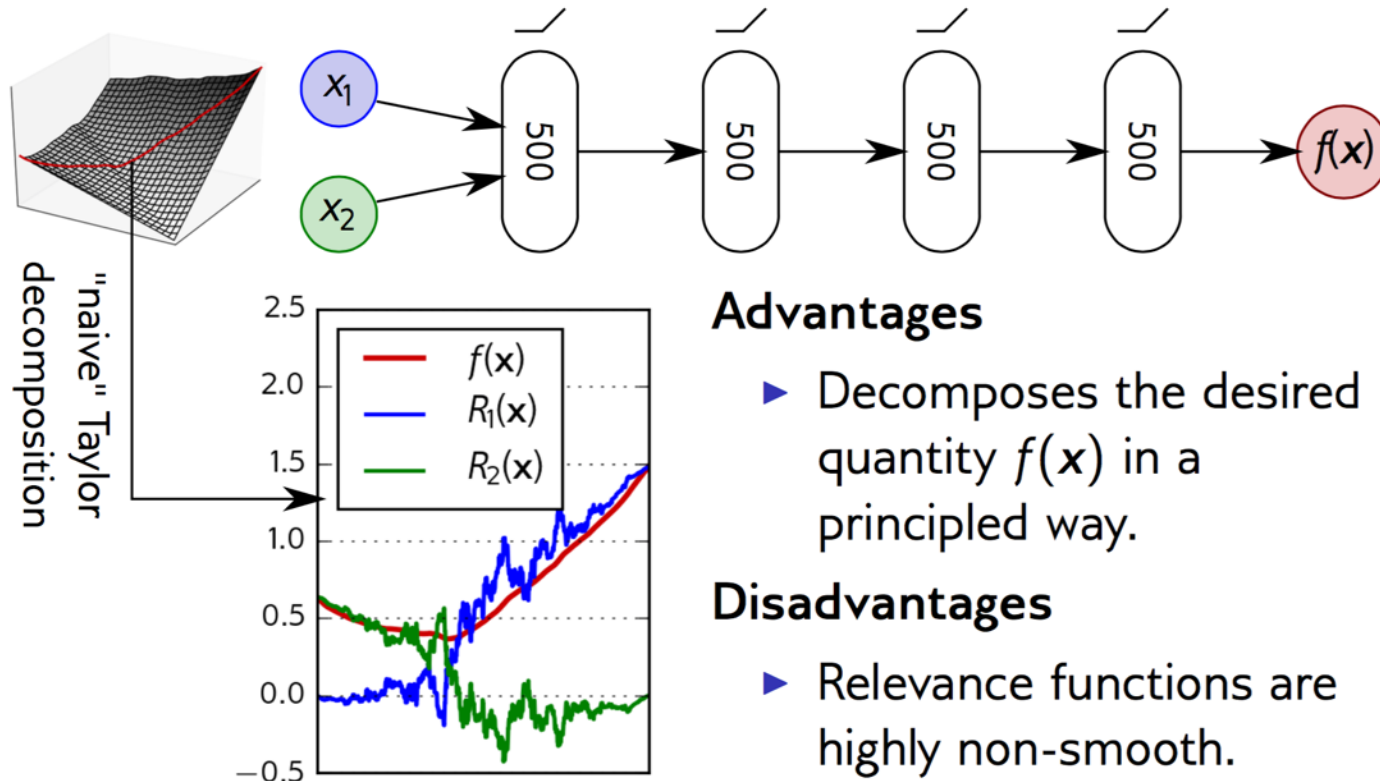
$$R_i = \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}=\tilde{\mathbf{x}}} \cdot (x_i - \tilde{x}_i)$$

Decomposing the Output of the DNN



$$R_i = \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{x}=\tilde{\mathbf{x}}} \cdot (x_i - \tilde{x}_i) \Rightarrow \text{"Naive" Taylor decomposition}$$

Decomposing the Output of the DNN



Advantages

- ▶ Decomposes the desired quantity $f(x)$ in a principled way.

Disadvantages

- ▶ Relevance functions are highly non-smooth.
- ▶ Relevance scores are sometimes negative.
- ▶ Inflexible w.r.t. the model.

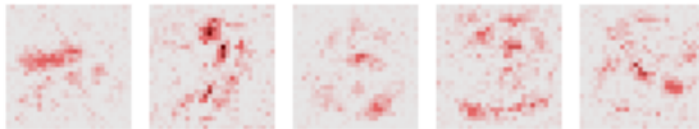
Experiment on Handwritten Digits

Data to classify:

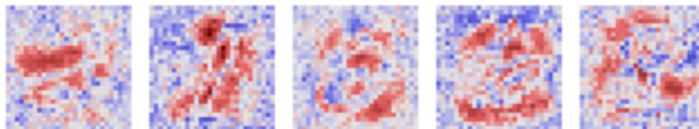


3-layer MLP:

Sensitivity analysis



Naive Taylor ($\tilde{x} = 0$)

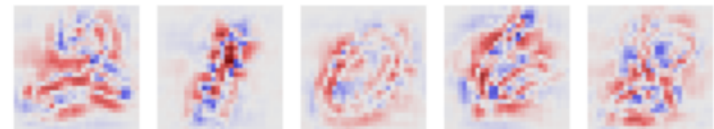


6-layer CNN:

Sensitivity analysis



Naive Taylor ($\tilde{x} = 0$)



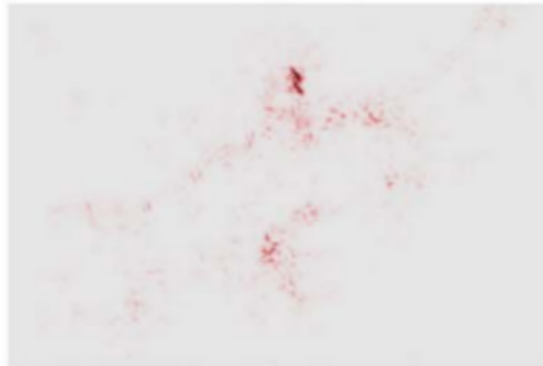
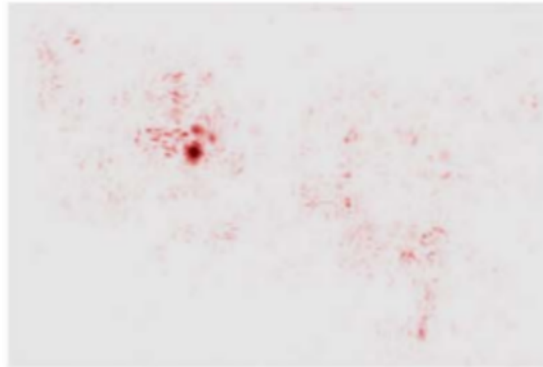
Observation: Both analyses produce noisy explanations of the MLP and CNN predictions.

Experiment on BVLC CaffeNet

Input images

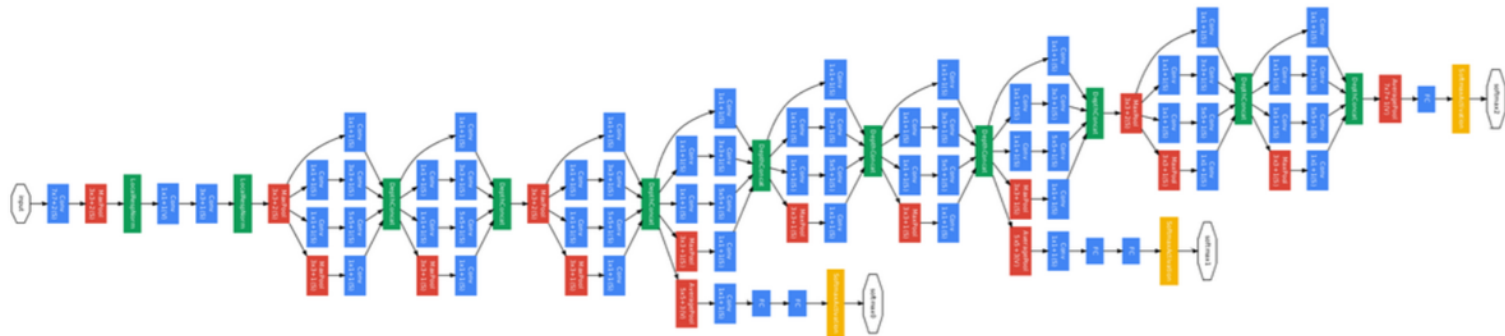


Sensitivity analysis



Observation: Explanations are noisy and (over/under)represent certain regions of the image.

Explaining DNN Predictions



- ▶ Standard methods (sensitivity analysis, naive Taylor decomposition) are subject to gradient noise and do not work well on deep neural networks.

DNN predictions need more advanced explanation methods.

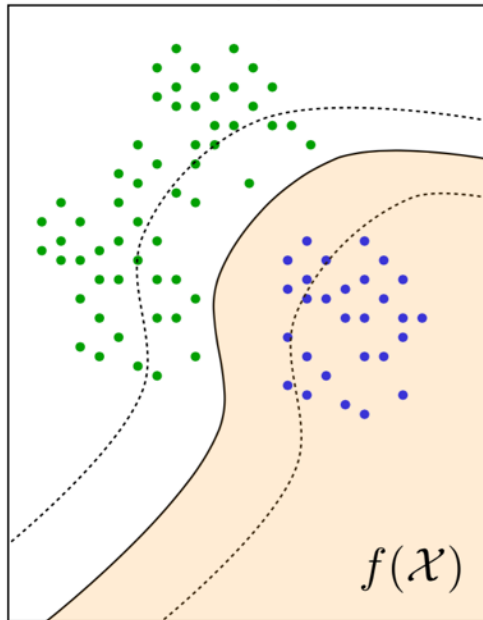
Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

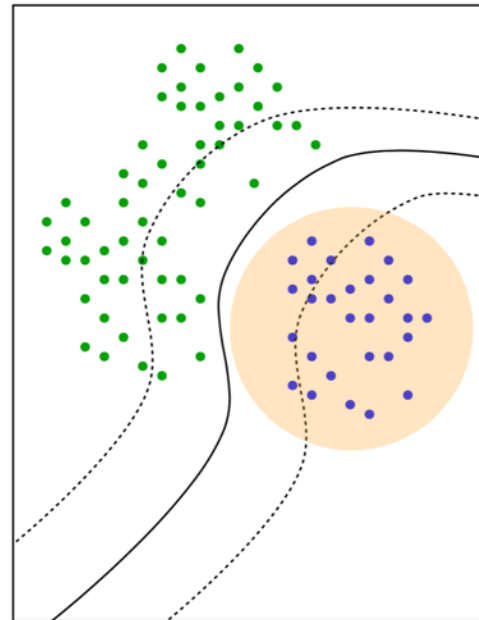
Part 2a: Making Deep Neural Networks Transparent

From Prototypes to Individual Explanations

model analysis

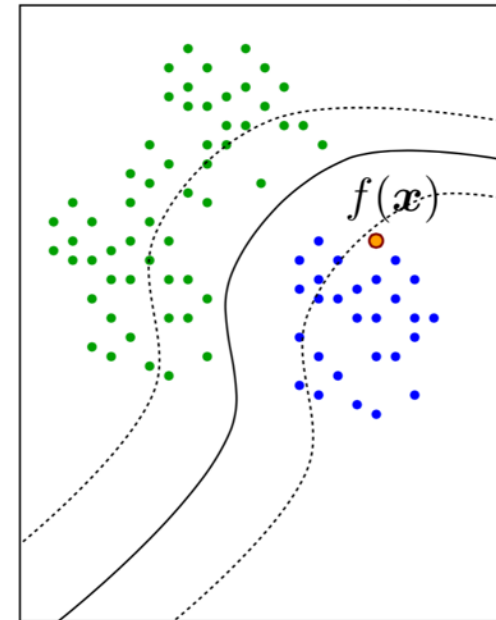


- visualizing filters
- max. class activation



- include distribution (RBM, DGN, etc.)

decision analysis



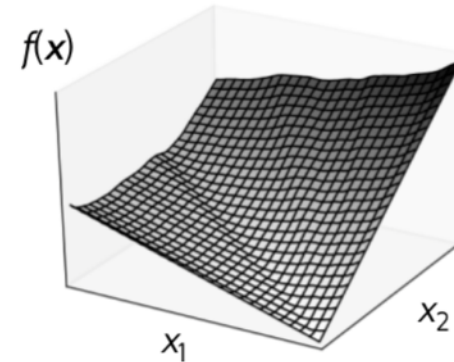
- sensitivity analysis
- decomposition

Why Simple Taylor doesn't work?

Two Reasons:

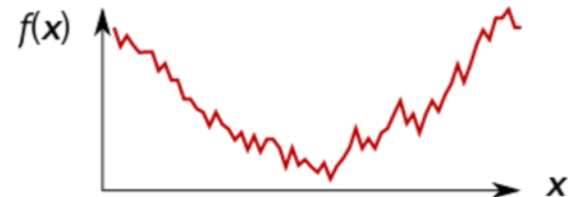
1

Root point is hard to find or too far $\rightarrow$ includes too much information (incl. negative evidence)



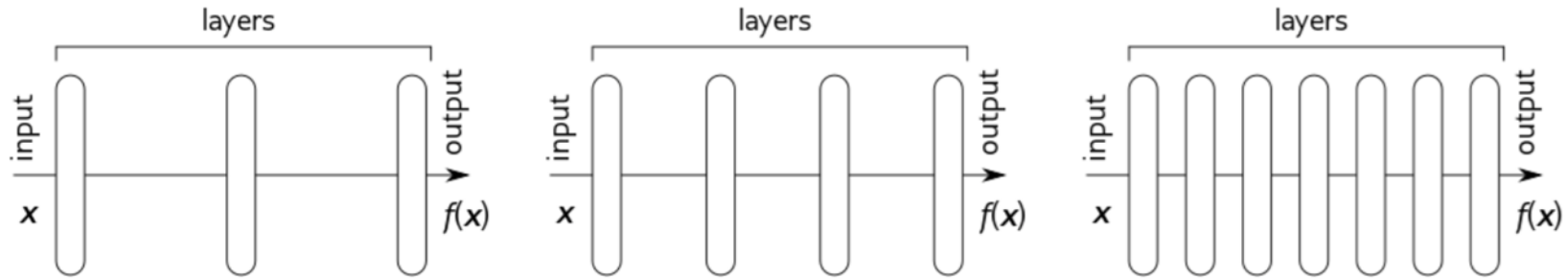
2

Gradient shattering problem $\rightarrow$ gradient of deep nets has low informative value

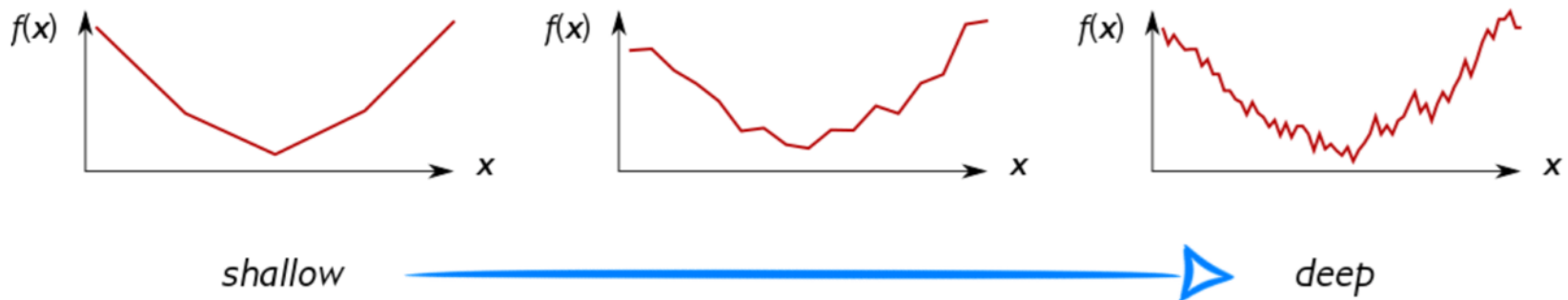


Gradient Shattering

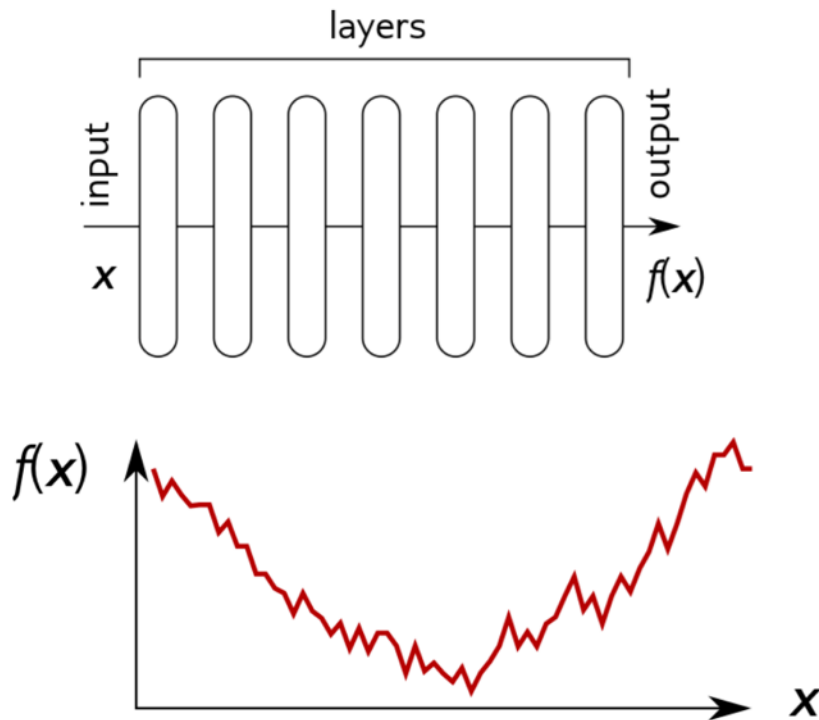
Structure's view



Function's view (cartoon)



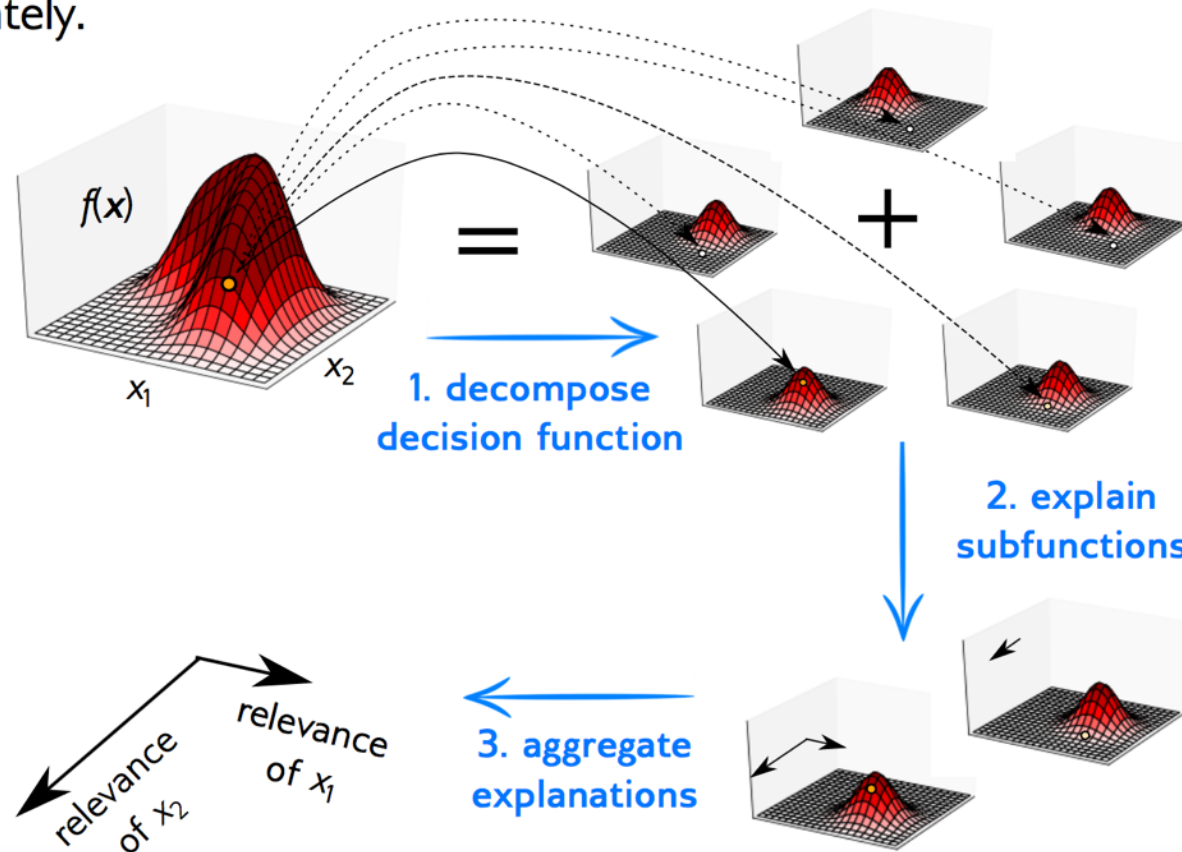
Sensitivity and Simple Taylor



Gradient-based methods usually do not work for explaining deep nets.

Deep Taylor Decomposition

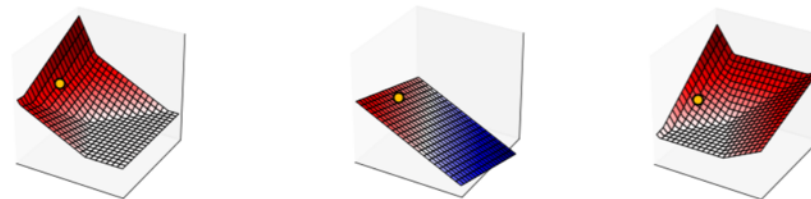
Key Idea: If a decision is too complex to explain, break the decision function into sub-functions, and explain each sub-decision separately.



Deep Taylor Decomposition

Taylor
decomposition
(TD)

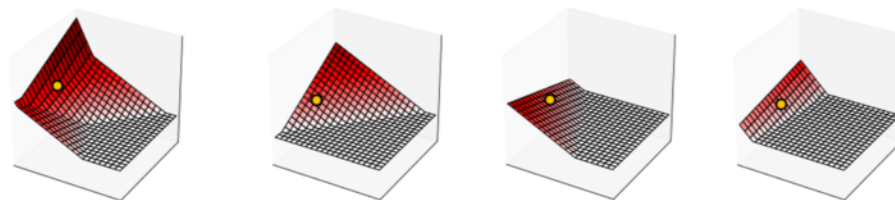
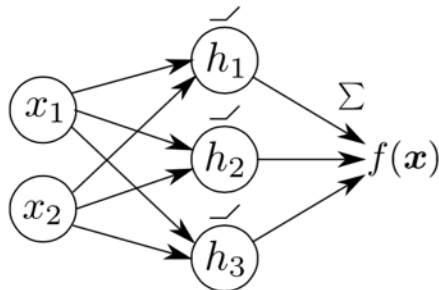
$$f(\mathbf{x}), \nabla f, \dots$$



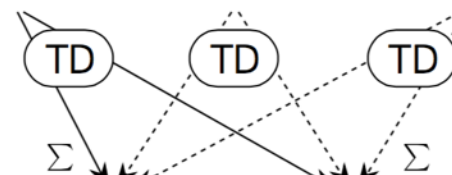
$$f(\mathbf{x}) = \nabla f|_{\mathbf{x}=\tilde{\mathbf{x}}}^\top \cdot (\mathbf{x} - \tilde{\mathbf{x}}) + \varepsilon$$

$$f(\mathbf{x}) = R_1 + R_2 + \varepsilon$$

deep Taylor
decomposition
(DTD)

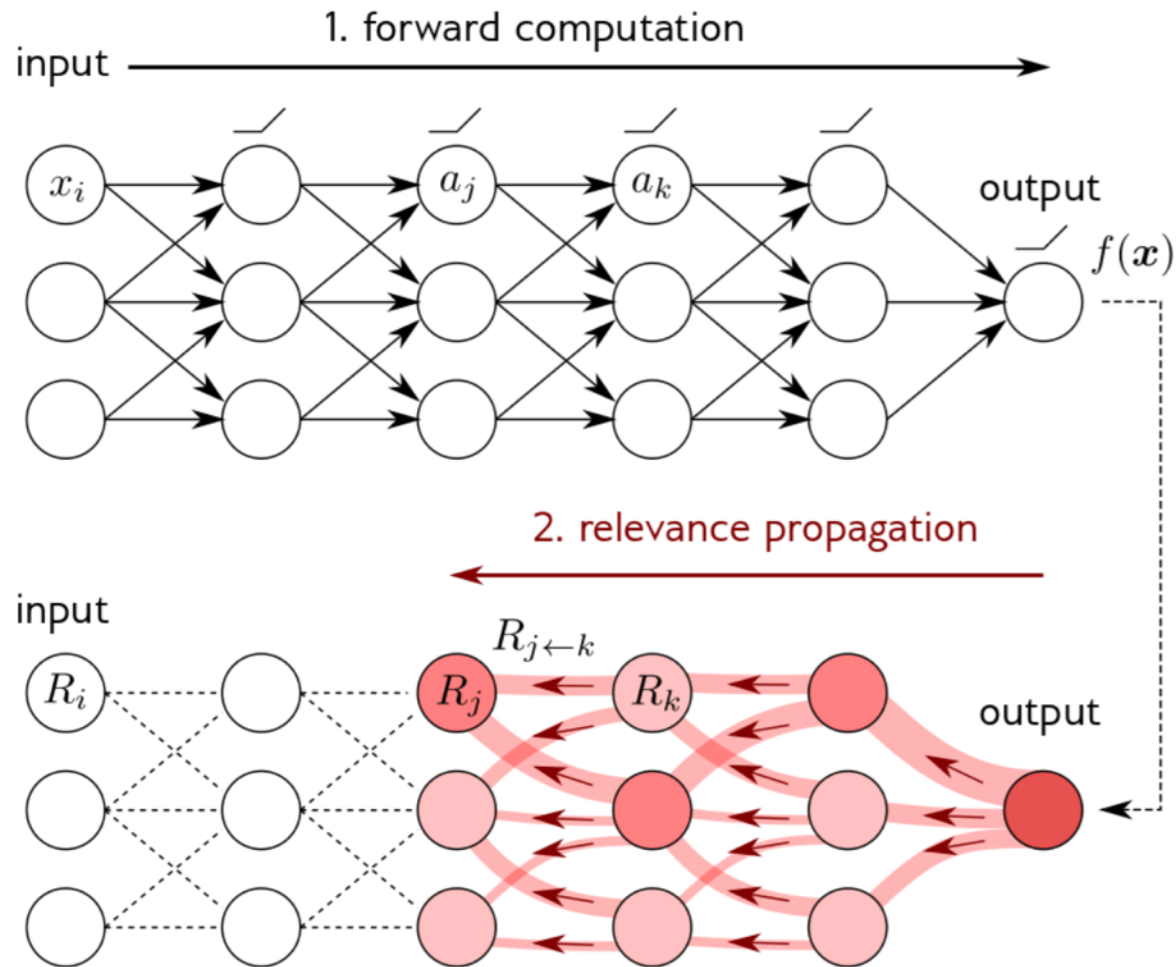


$$f(\mathbf{x}) = h_1 + h_2 + h_3$$

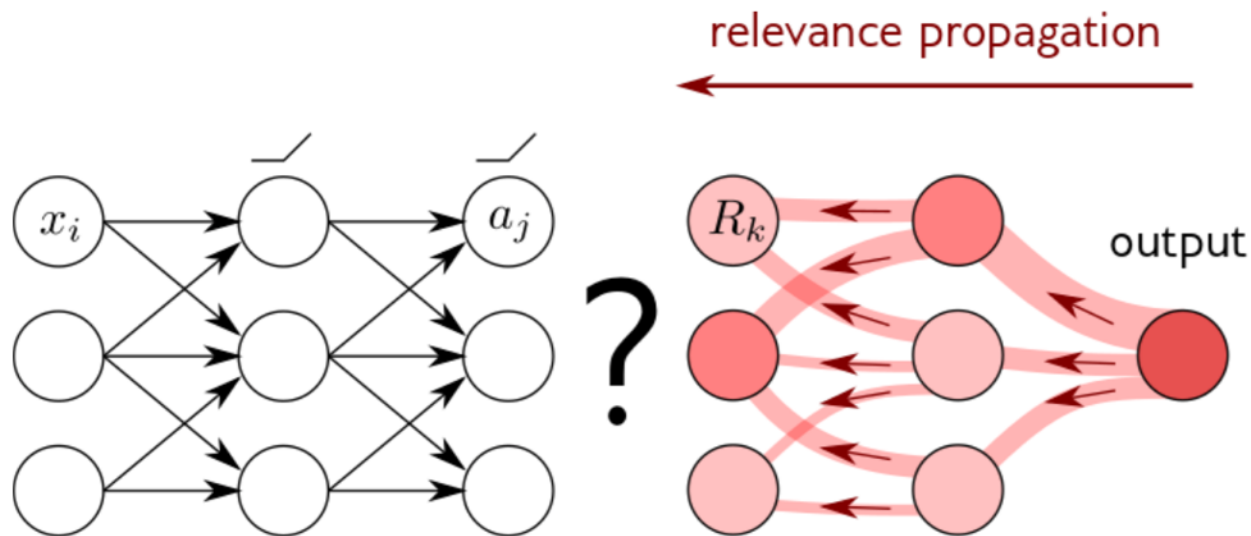


$$f(\mathbf{x}) = R_1 + R_2$$

Deep Taylor Decomposition



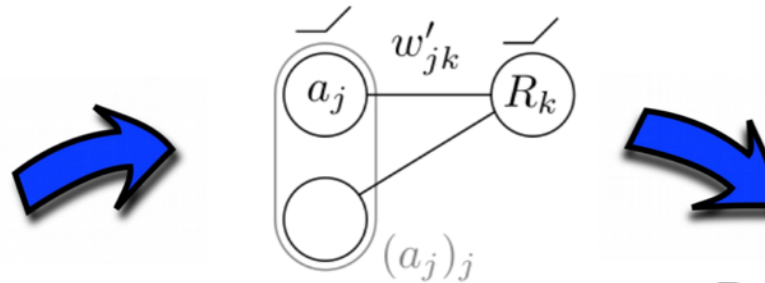
Deep Taylor Decomposition



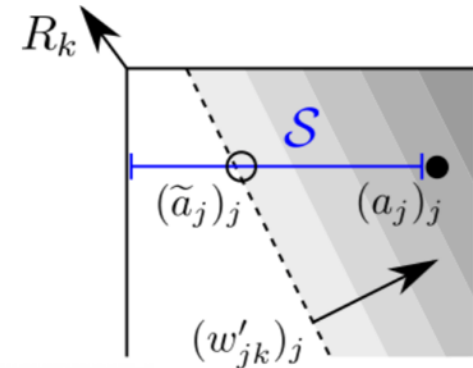
Can we express R_k as a simple function of $(a_j)_j$?

Can we do a Taylor decomposition of $R_k((a_j)_j)$?

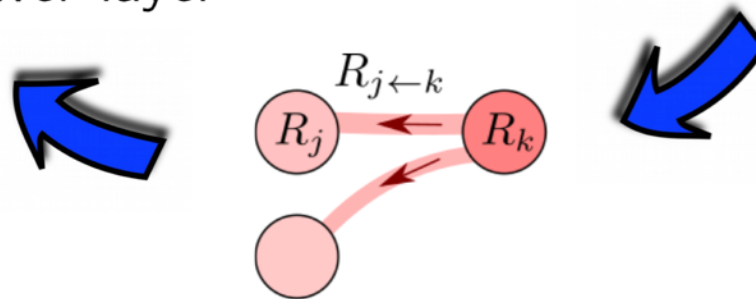
Deep Taylor Decomposition



Observe that $R_k \approx a_k \cdot \text{const.}$



Move to the lower-layer



Deep Taylor Decomposition / LRP

$$R_j = \sum_k \left(\alpha \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} - \beta \frac{a_j w_{jk}^-}{\sum_j a_j w_{jk}^-} \right) R_k$$

intuition
[Bach'15]



Relevance should be redistributed to the lower-layer neurons $(a_j)_j$ in proportion to their excitatory effect on a_k . “Counter-relevance” should be redistributed to the lower-layer neurons $(a_j)_j$ in proportion to their inhibitory effect on a_k .

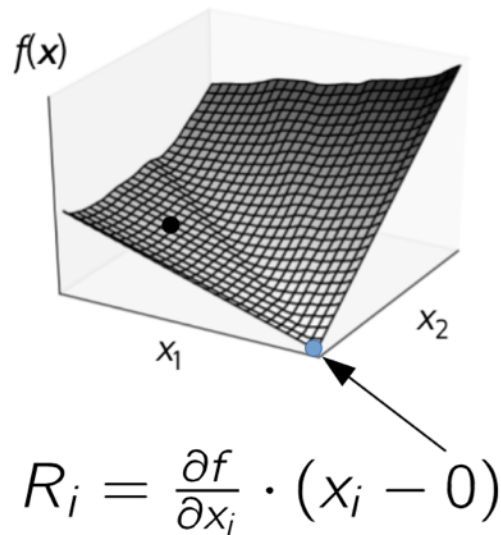


analysis
[Montavon'17]

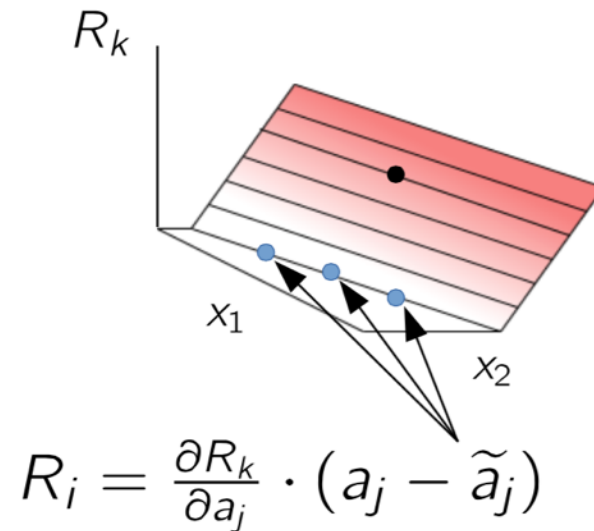
For the specific case $\alpha = 1$, the whole LRP procedure can be seen as a *deep Taylor decomposition* of the neural network function.

Deep Taylor Decomposition / LRP

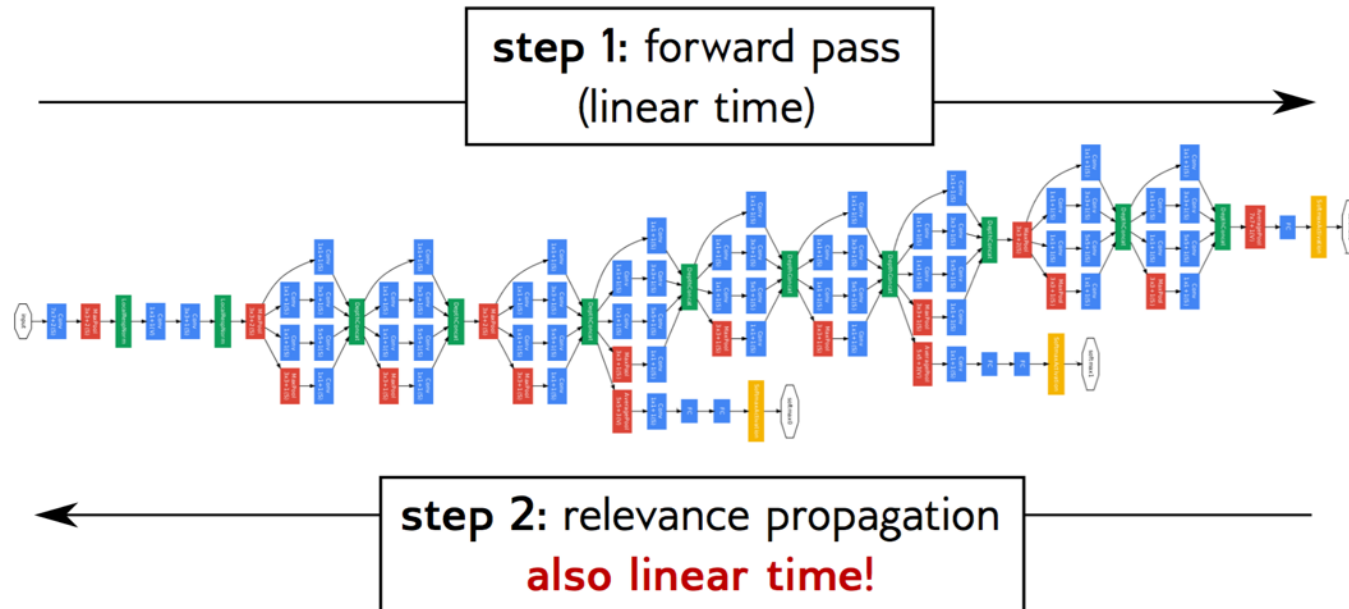
Simple Taylor



Deep Taylor



Deep Taylor Decomposition / LRP



Propagation rule:

$$R_i = \sum_j q_{ij} R_j \quad \sum_i q_{ij} = 1$$

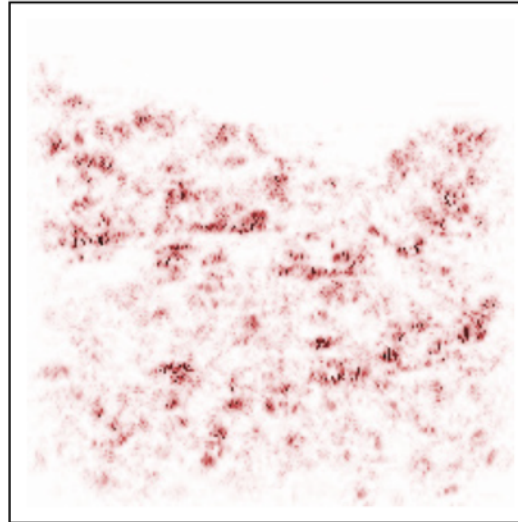
Various rules are available for pixel layers, intermediate layers, or special layers.

Deep Taylor Decomposition / LRP

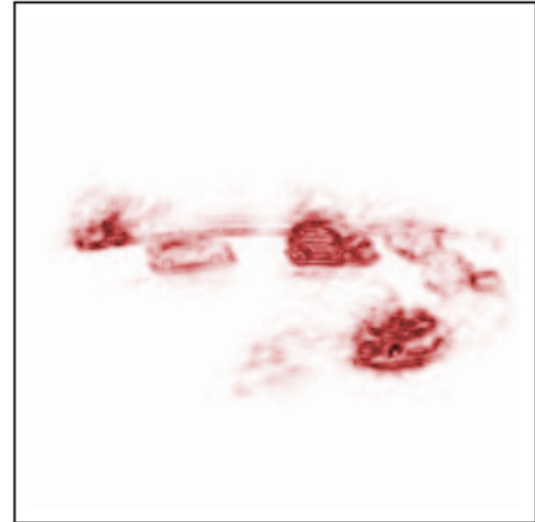
Image



Sensitivity Analysis



Deep Taylor LRP



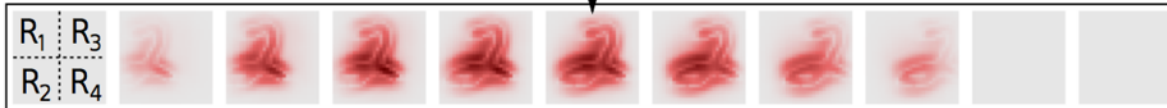
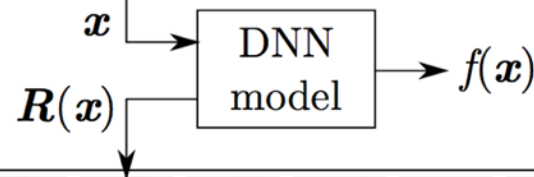
Observation: Only deep Taylor LRP focuses on cars.

Deep Taylor Decomposition / LRP

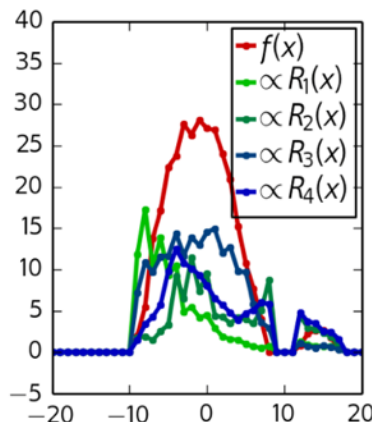
input sequence



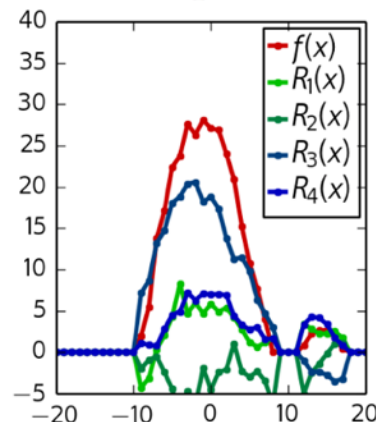
explanation with
relevance propagation



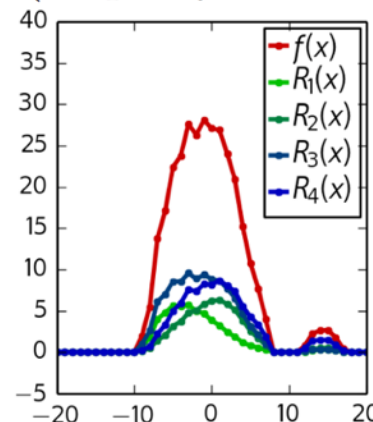
sensitivity analysis



simple Taylor
decomposition



relevance propagation
(deep Taylor LRP)

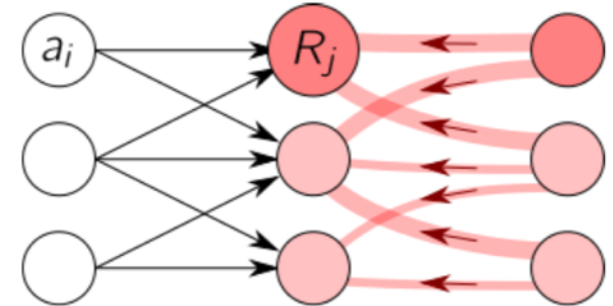


Two similar images
should have similar
explanations.

Technical Details

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

$$R_j = a_j c_j$$



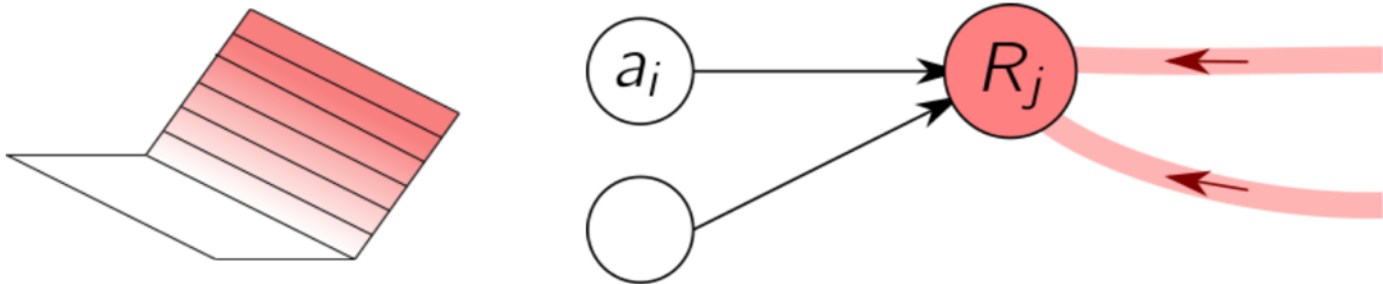
$$R_i = a_i \sum_j w_{ij}^+ \frac{\max(0, \sum_i a_i w_{ij})}{\sum_i a_i w_{ij}^+} c_j$$

$$R_i = a_i c_j$$

Relevance has product structure at all layers.

1

Build the Relevance Neuron



$$R_j = a_j c_j$$

$$= \max(0, \sum_i a_i w_{ij}) \cdot c_j$$

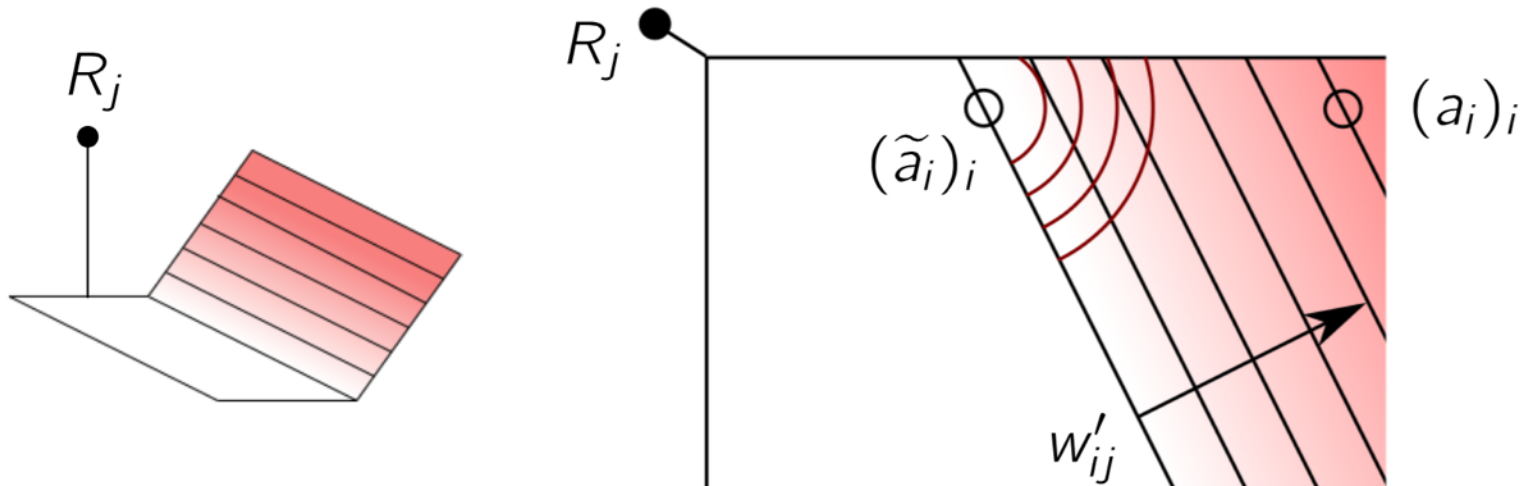
$$= \max(0, \sum_i a_i w'_{ij})$$

$$w'_{ij} = w_{ij} c_j$$

2

Expand the Relevance Neuron

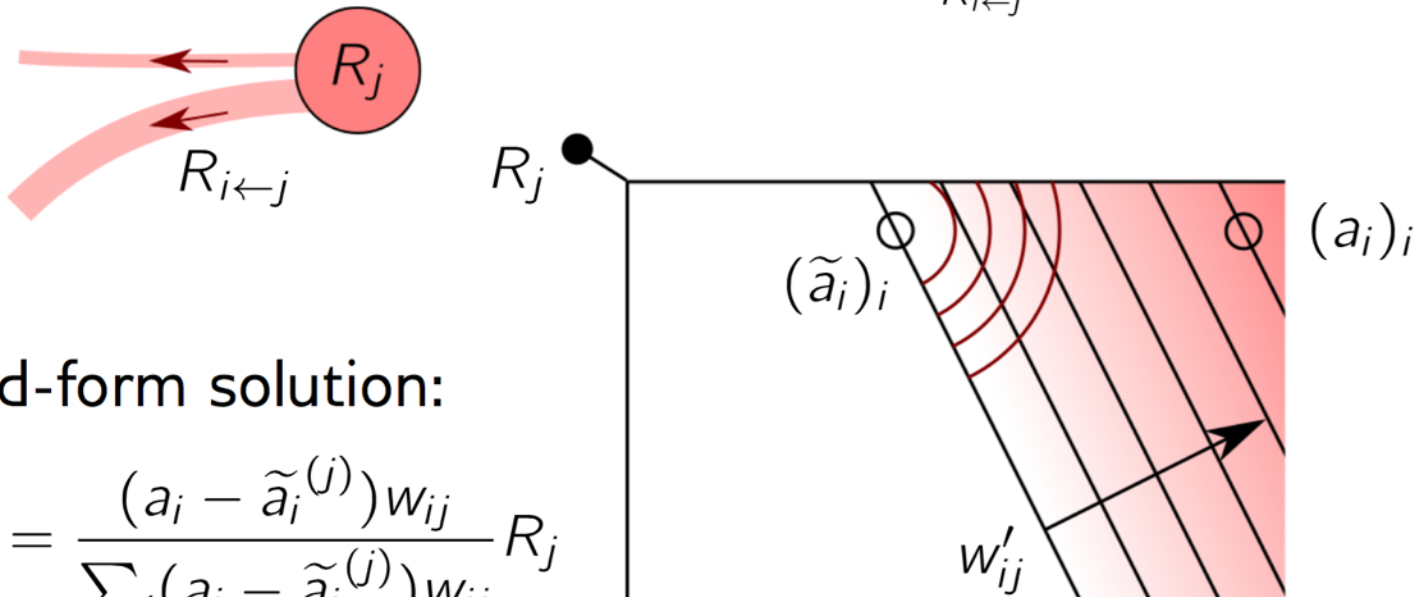
$$R_j((a_i)_i) = R_j((\tilde{a}_i)_i) + \underbrace{\sum_i \frac{\partial R_j}{\partial a_i} \Big|_{(\tilde{a}_i)_i}}_{R_{i \leftarrow j}} \cdot (a_i - \tilde{a}_i) + \varepsilon$$



3

Decompose Relevance

$$R_j((a_i)_i) = R_j((\tilde{a}_i)_i) + \underbrace{\sum_i \frac{\partial R_j}{\partial a_i} \Big|_{(\tilde{a}_i)_i}}_{R_{i \leftarrow j}} \cdot (a_i - \tilde{a}_i) + \varepsilon$$



Closed-form solution:

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

Technical Details

Closed-form solution:

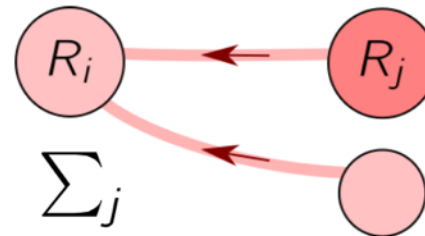
$$R_{i \leftarrow j} = \frac{(a_i - \tilde{a}_i^{(j)}) w_{ij}}{\sum_i (a_i - \tilde{a}_i^{(j)}) w_{ij}} R_j$$

LRP rule [Bach15, Zhang16]

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

4

Pooling relevance
over all outgoing
neurons



5

Search root point along
specific direction:

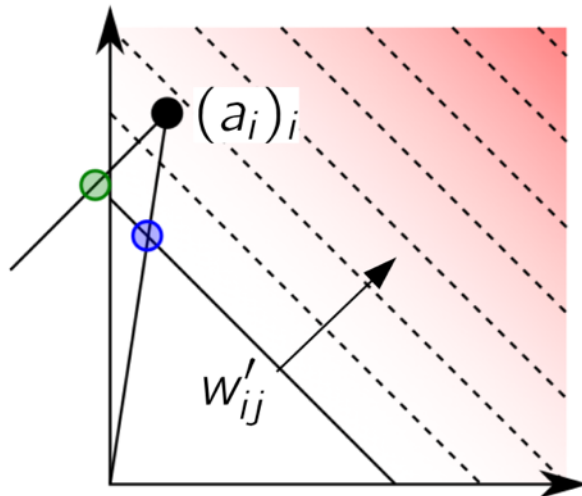
$$(a_i - \tilde{a}_i^{(j)}) \propto a_i 1_{w'_{ij} > 0}$$

**But what does
that mean?**

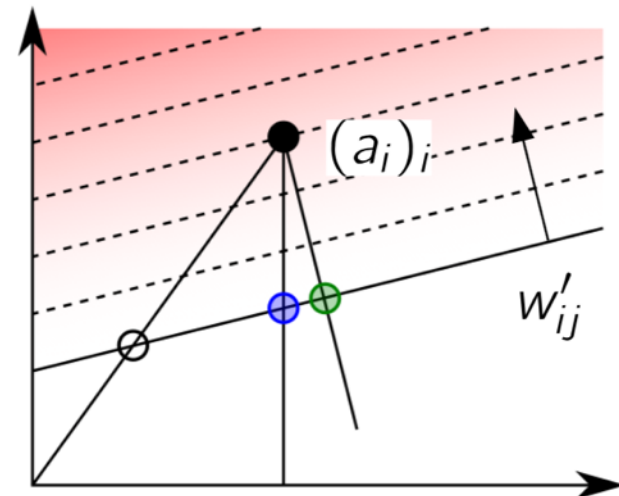
18/22

Technical Details

case 1



case 2



- $(a_i - \tilde{a}_i^{(j)}) \propto w'_{ij}$
- $(a_i - \tilde{a}_i^{(j)}) \propto a_i$
- $(a_i - \tilde{a}_i^{(j)}) \propto a_i 1_{w'_{ij} > 0}$

nearest root

origin (like simple Taylor)

root along positive
activations (LRP rule)

19/22

Technical Details

Input domain	Rule
ReLU activations ($a_j \geq 0$)	$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$
Pixel intensities ($x_i \in [l_i, h_i]$, $l_i \leq 0 \leq h_i$)	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$
Real values ($x_i \in \mathbb{R}$)	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$

Deep Taylor LRP rules [Montavon'17]

More refined rules can also be constructed to match the input data distribution [Kindermans'17]

20

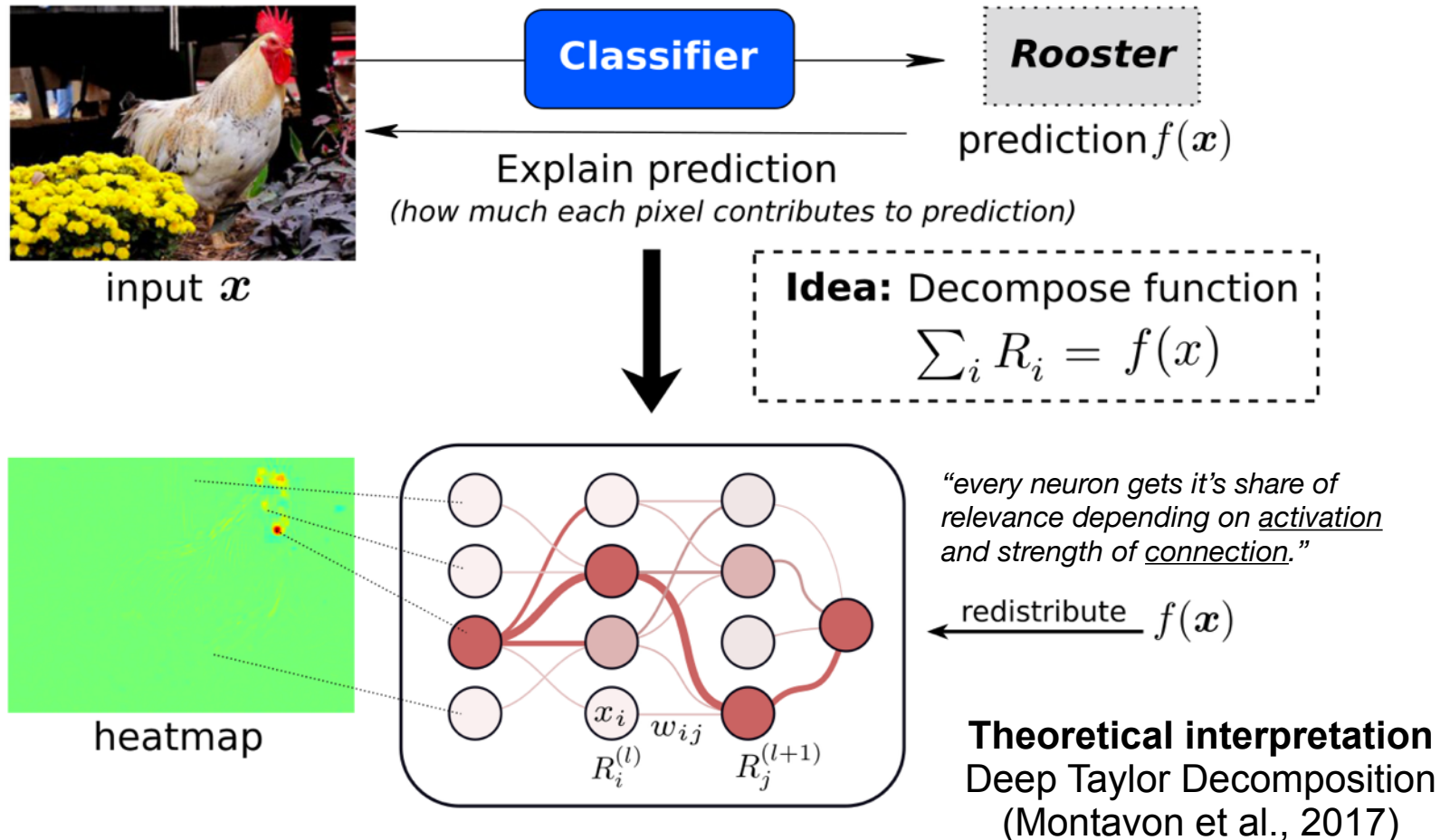
Tutorial on Interpretable Machine Learning

W. Samek & K.-R. Müller

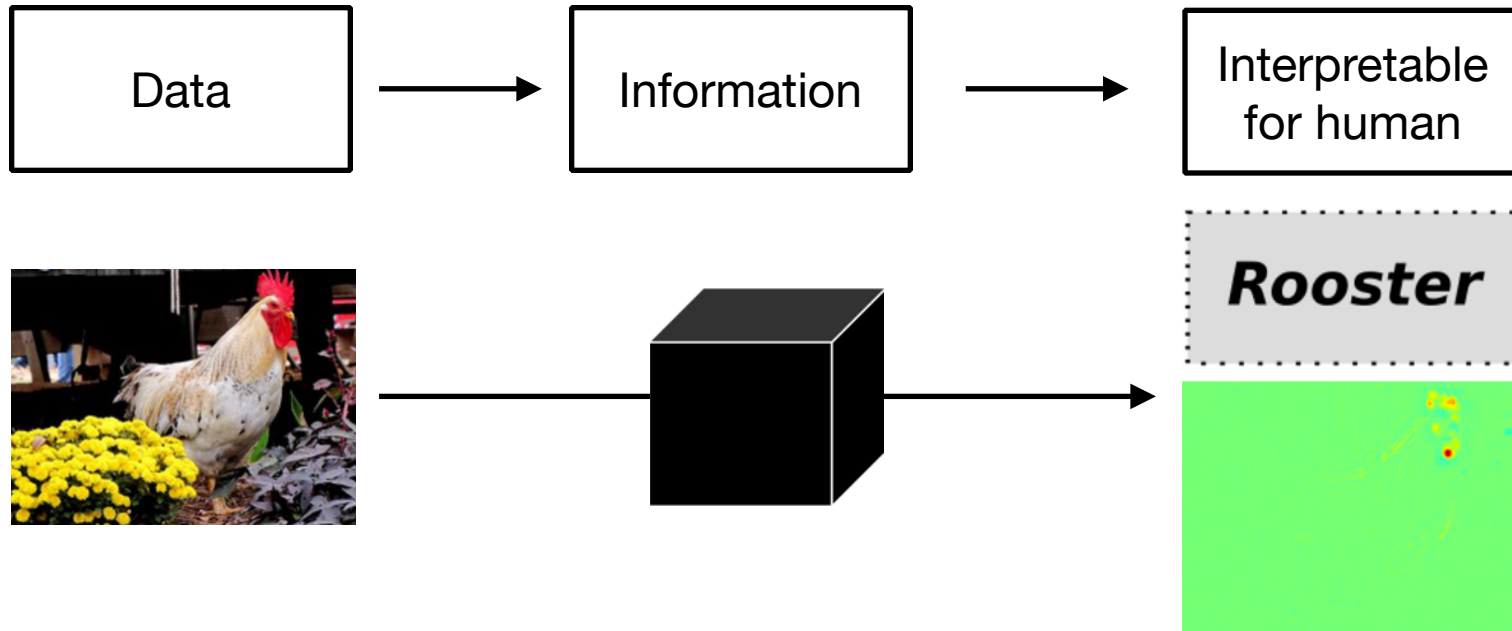
Part 2b: Applications & Discussion

Recap: Layer-wise Relevance Propagation (LRP)

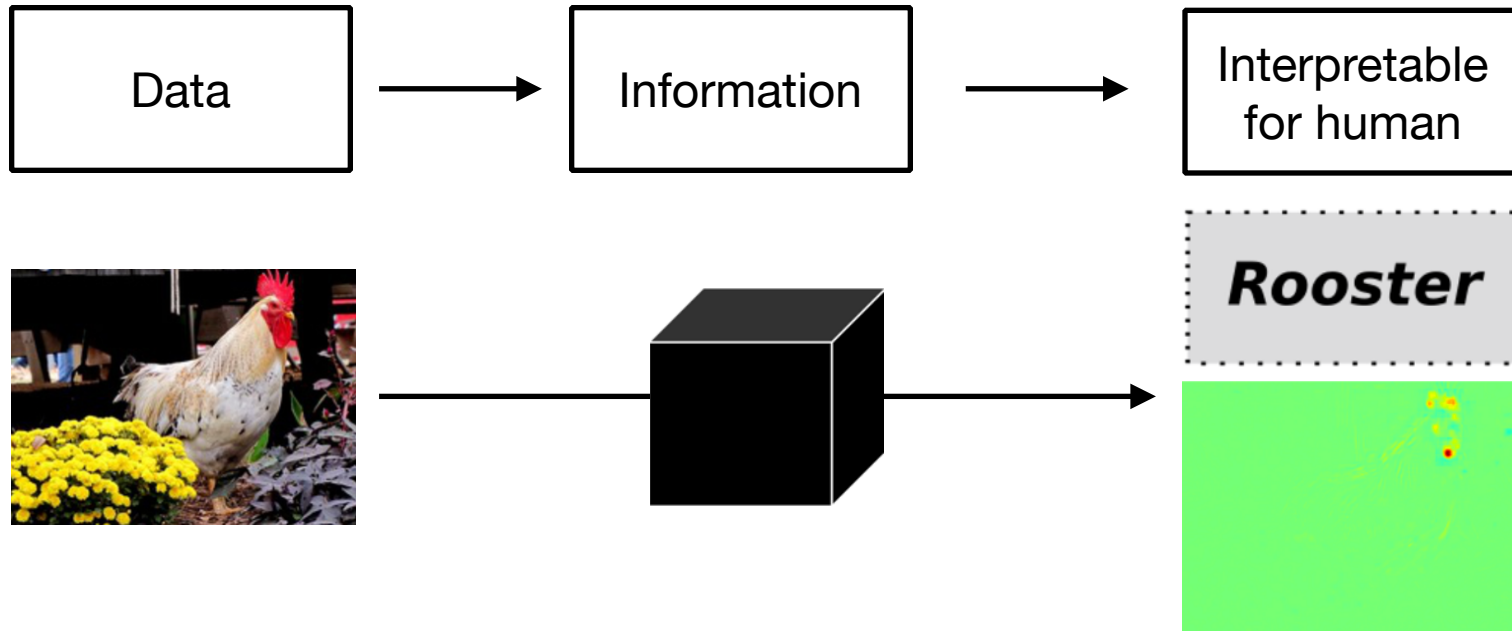
Layer-wise Relevance Propagation (LRP)
(Bach et al. 2015)



Explanations and now ?

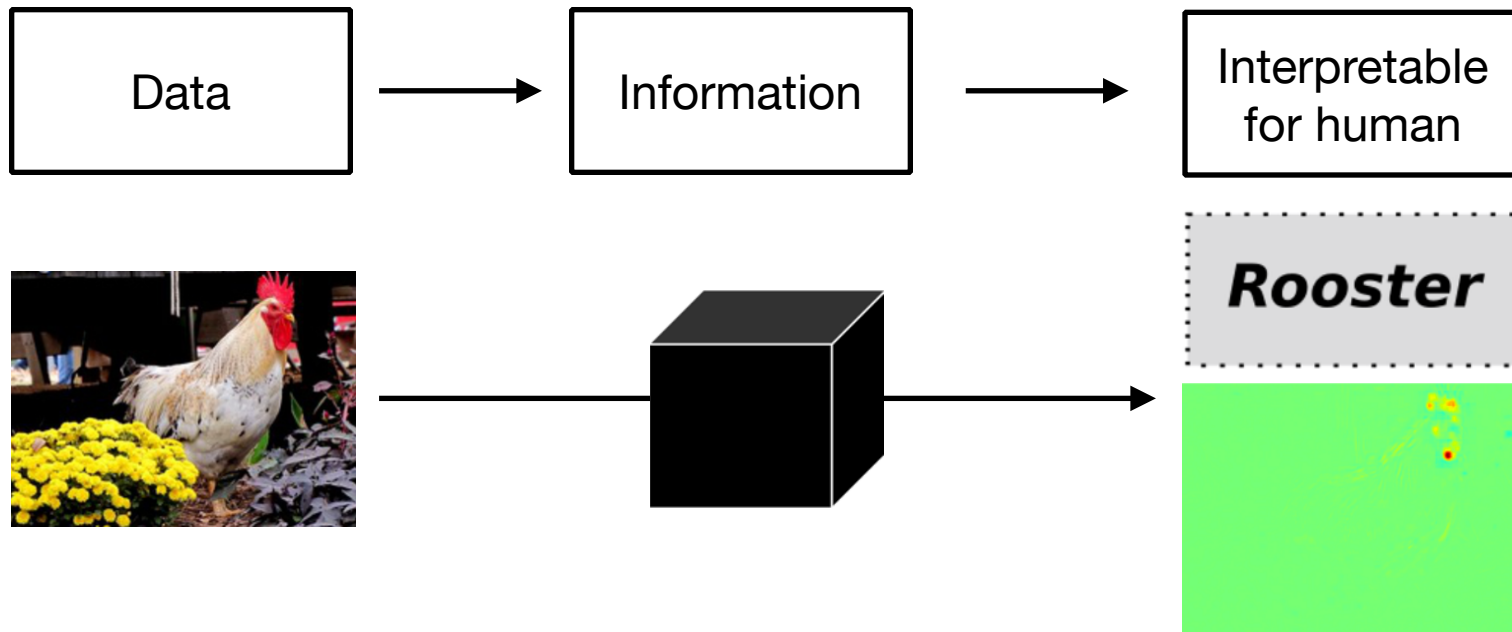


Explanations and now ?



How good is the explanation ?

Explanations and now ?



How good is the explanation ?

What can we do with it ?

Explanations and now ?

How good is the explanation ?

- Objective measure of quality
- Compare explanation methods

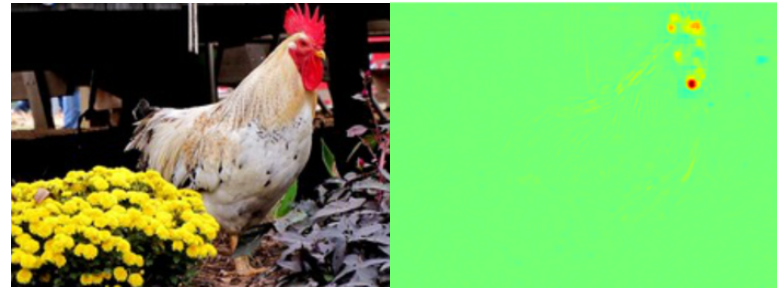
What can we do with it ?

- Compare classifiers
- Detect biases and flaws
- Quantify use of context
- Novel representation
- ...

Measuring Quality of Explanations

Heatmap depends on

- classifier
- explanation method

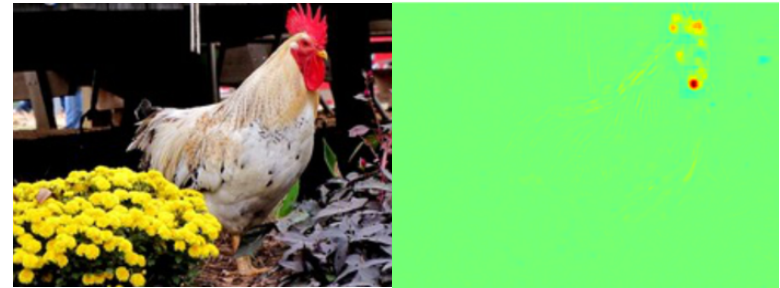


If we want to compare classifiers or explanations methods, we need an *objective* measure of heatmap quality.

Measuring Quality of Explanations

Heatmap depends on

- classifier
- explanation method



If we want to compare classifiers or explanations methods, we need an *objective* measure of heatmap quality.

Algorithm (Pixel Flipping)

Sort pixel scores

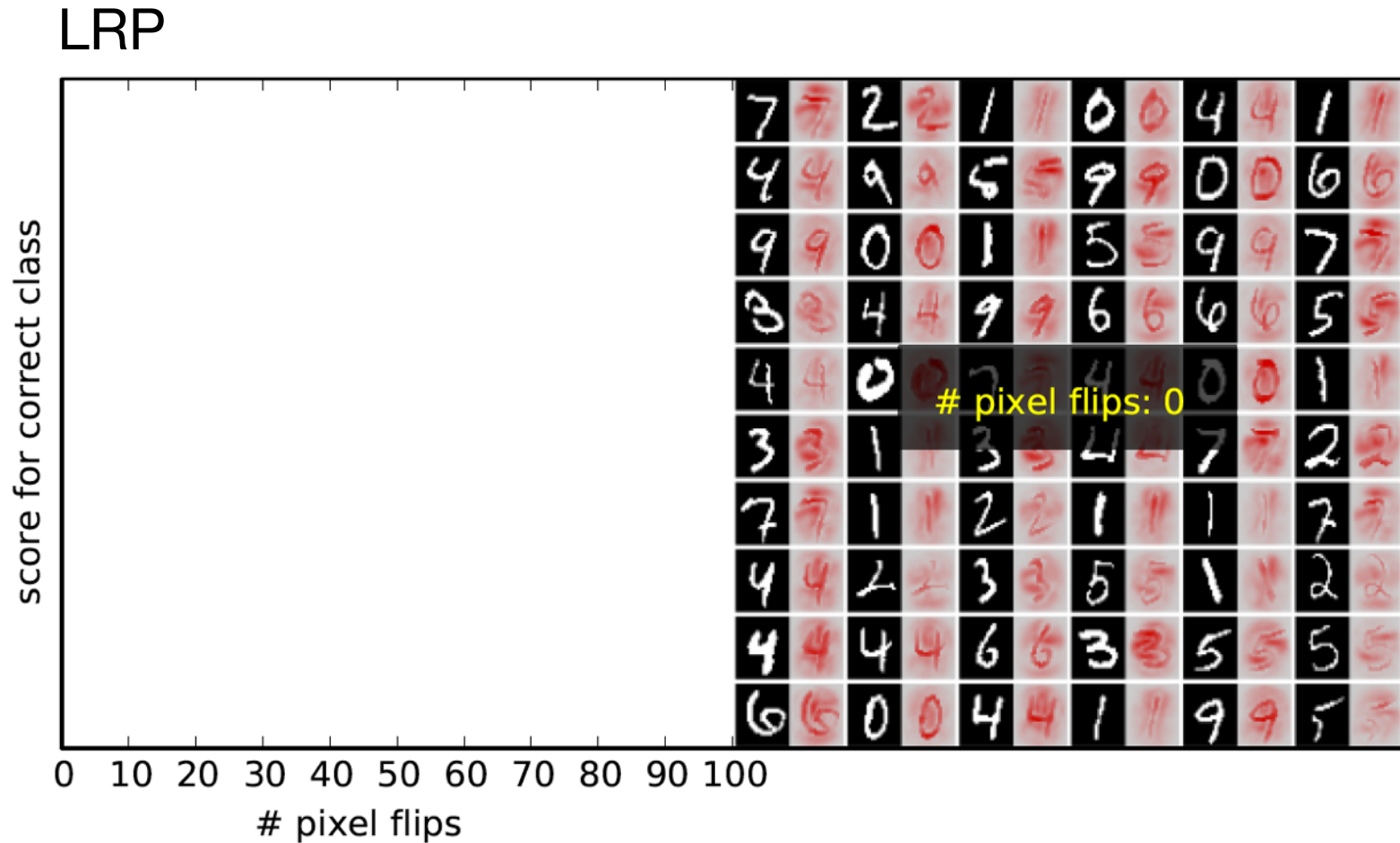
Iterate

 flip pixels

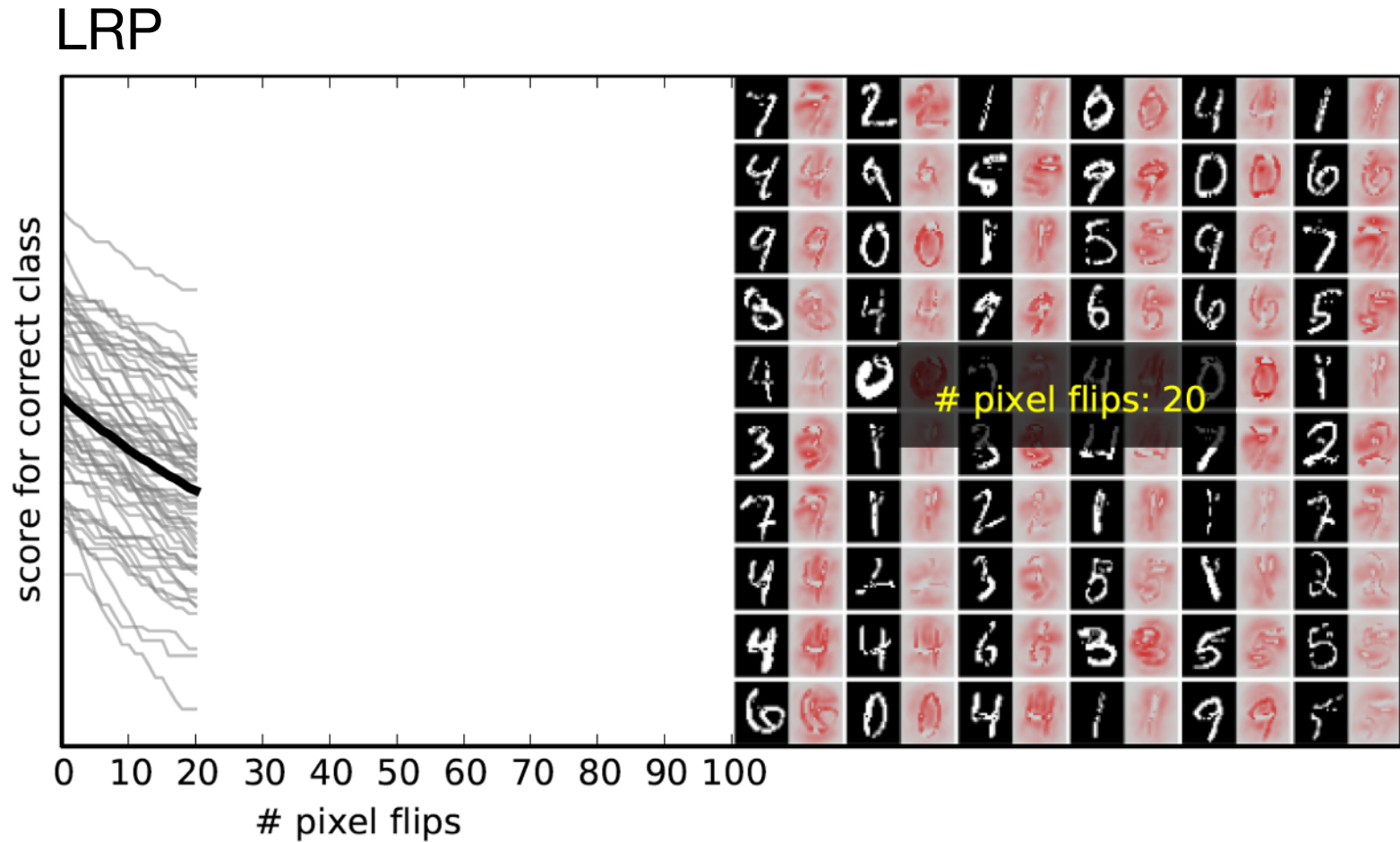
 evaluate $f(x)$

Measure decrease of $f(x)$

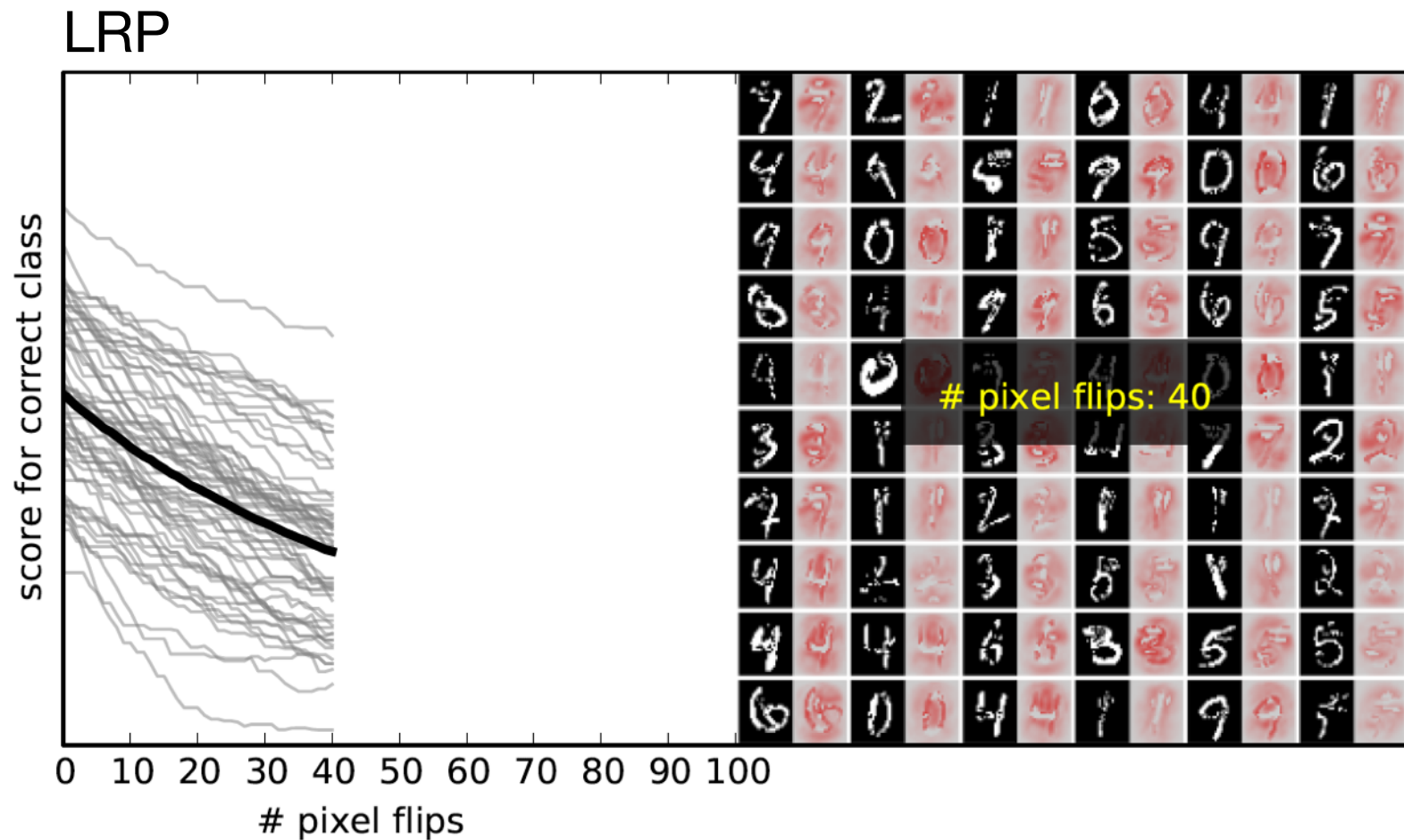
Compare Explanation Methods



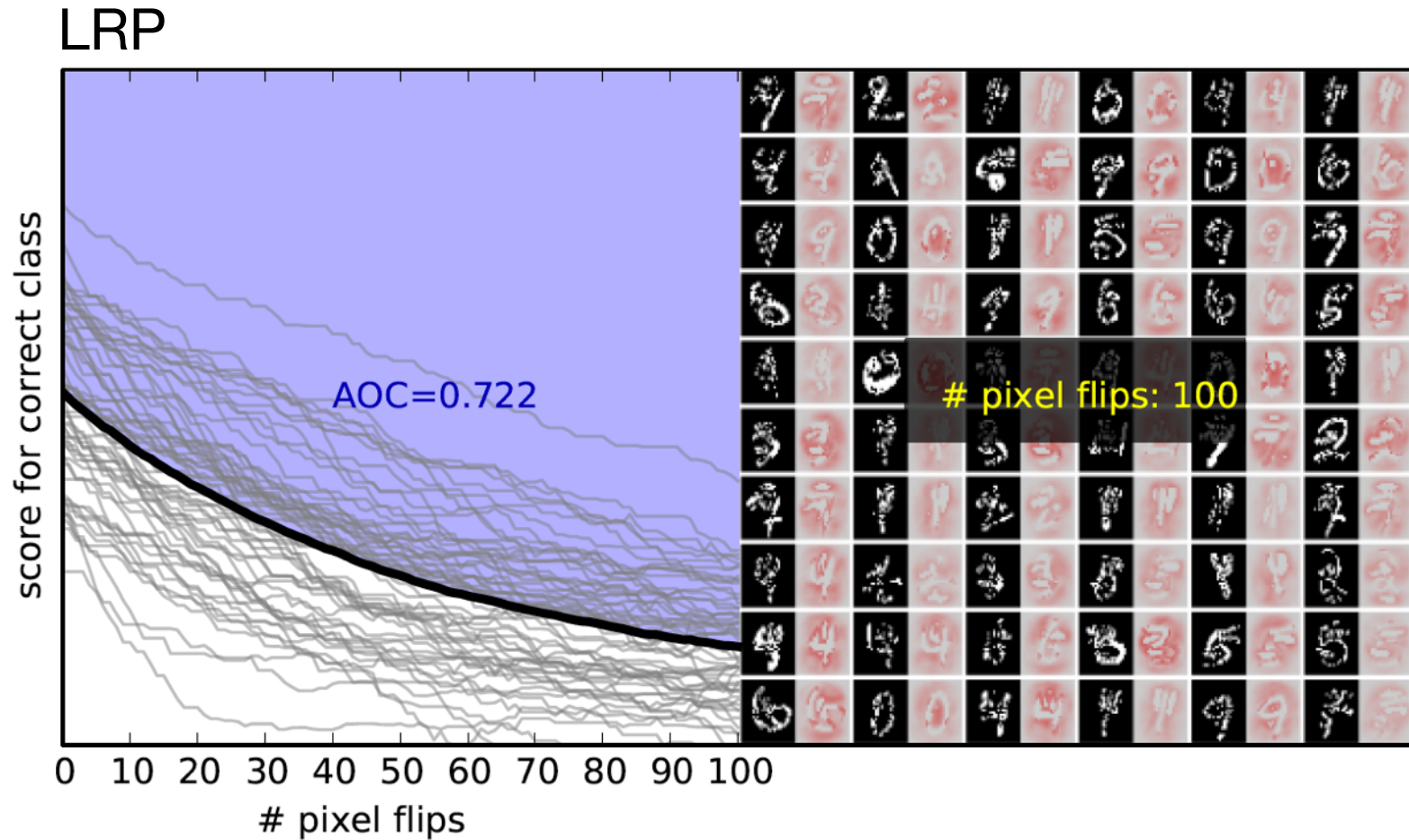
Compare Explanation Methods



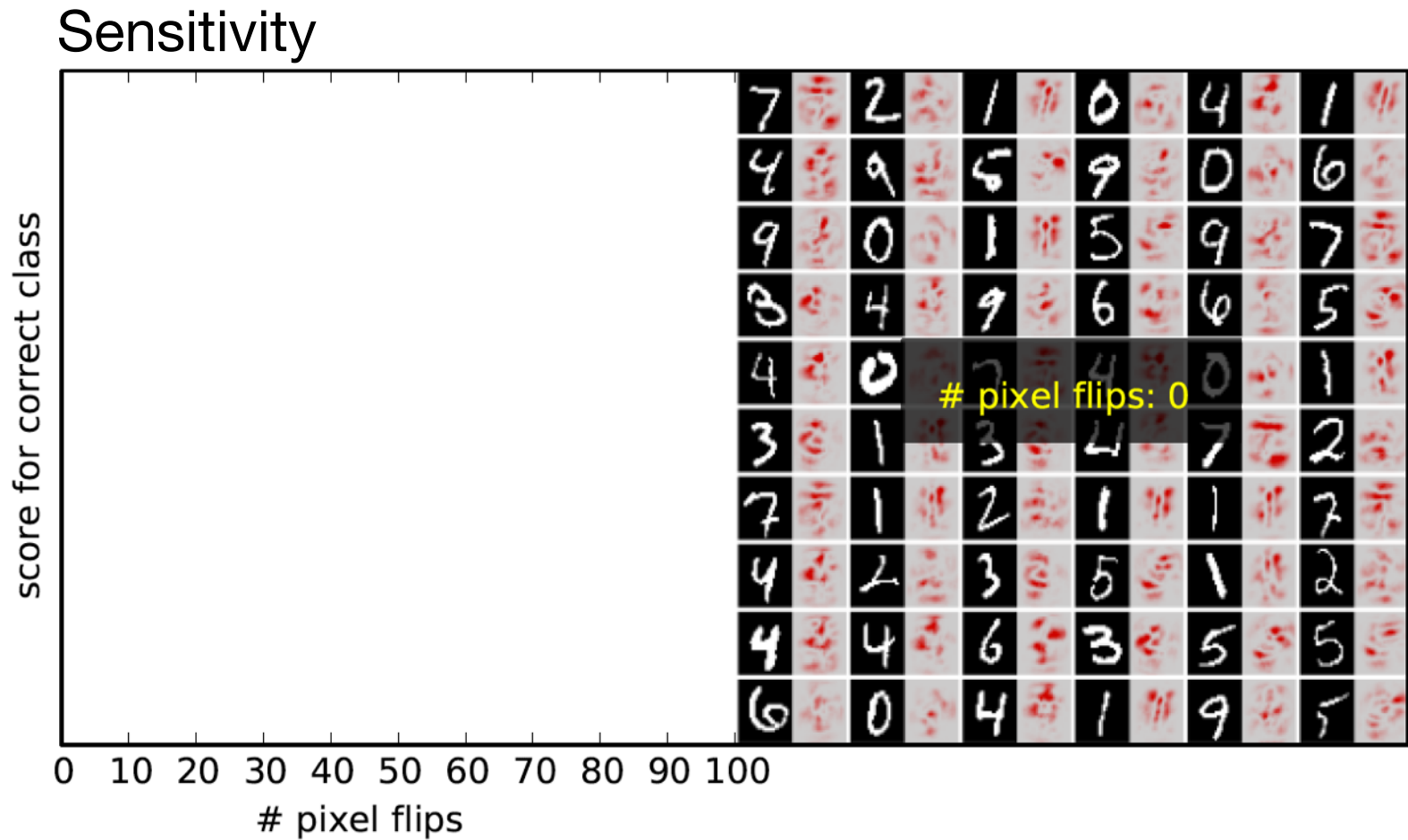
Compare Explanation Methods



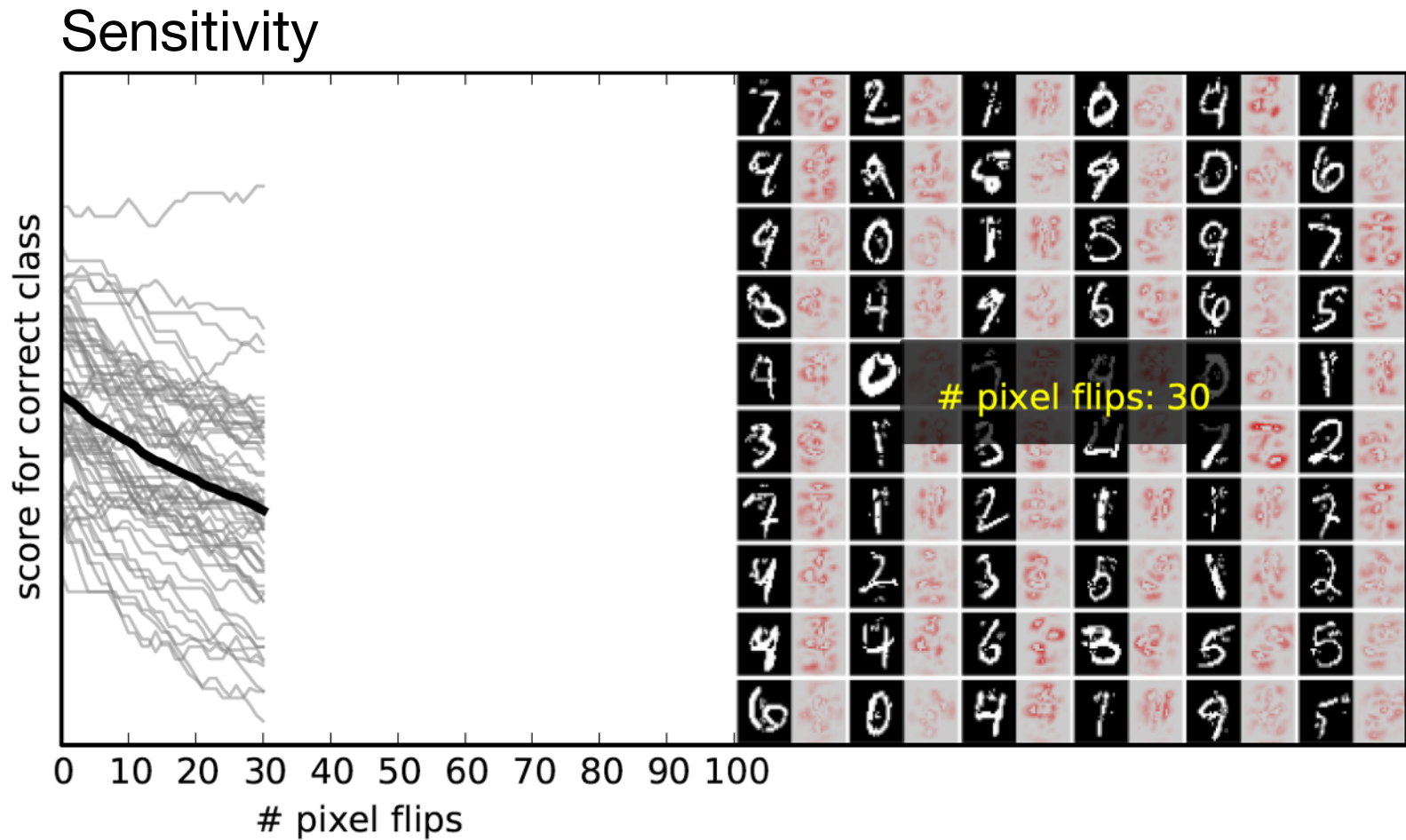
Compare Explanation Methods



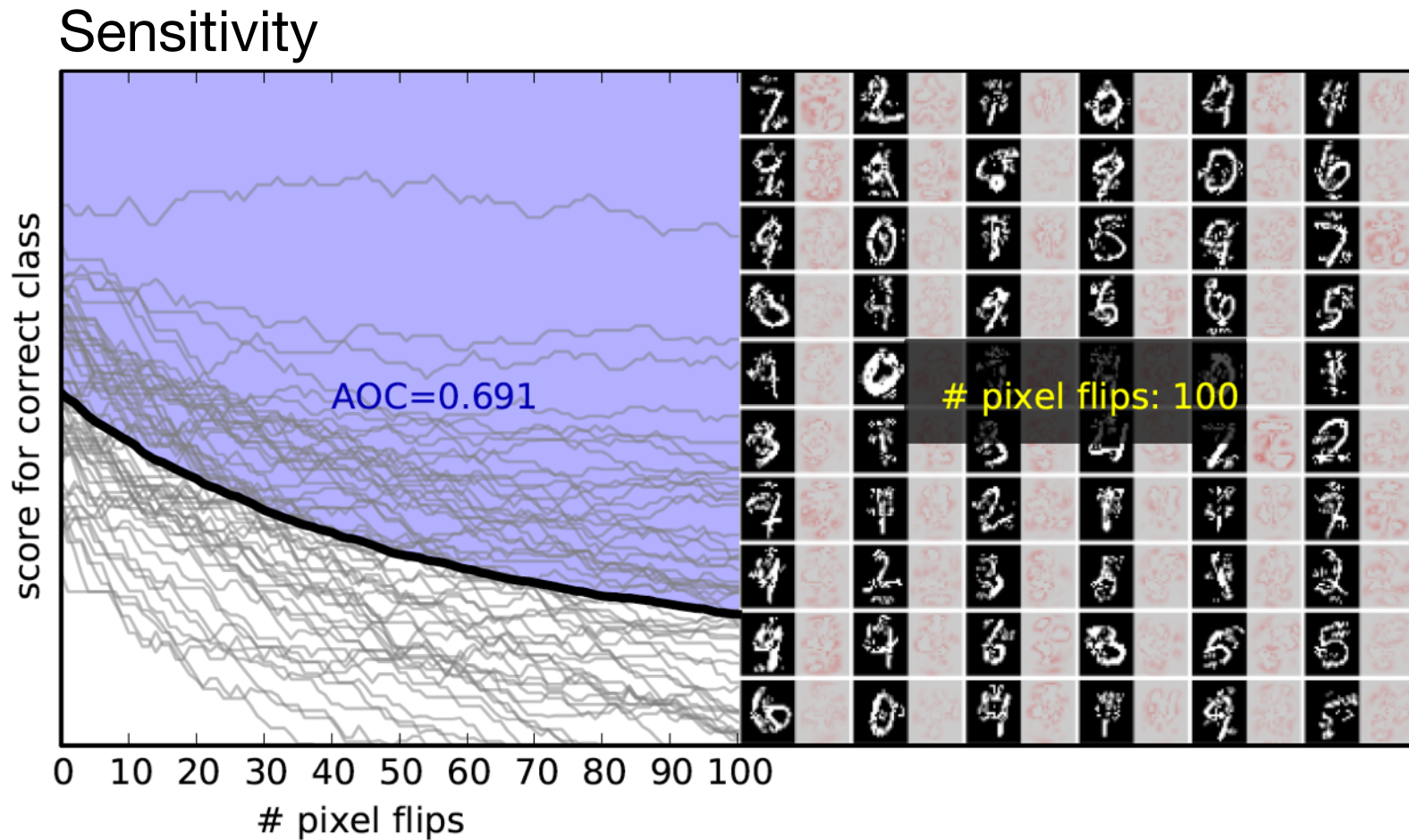
Compare Explanation Methods



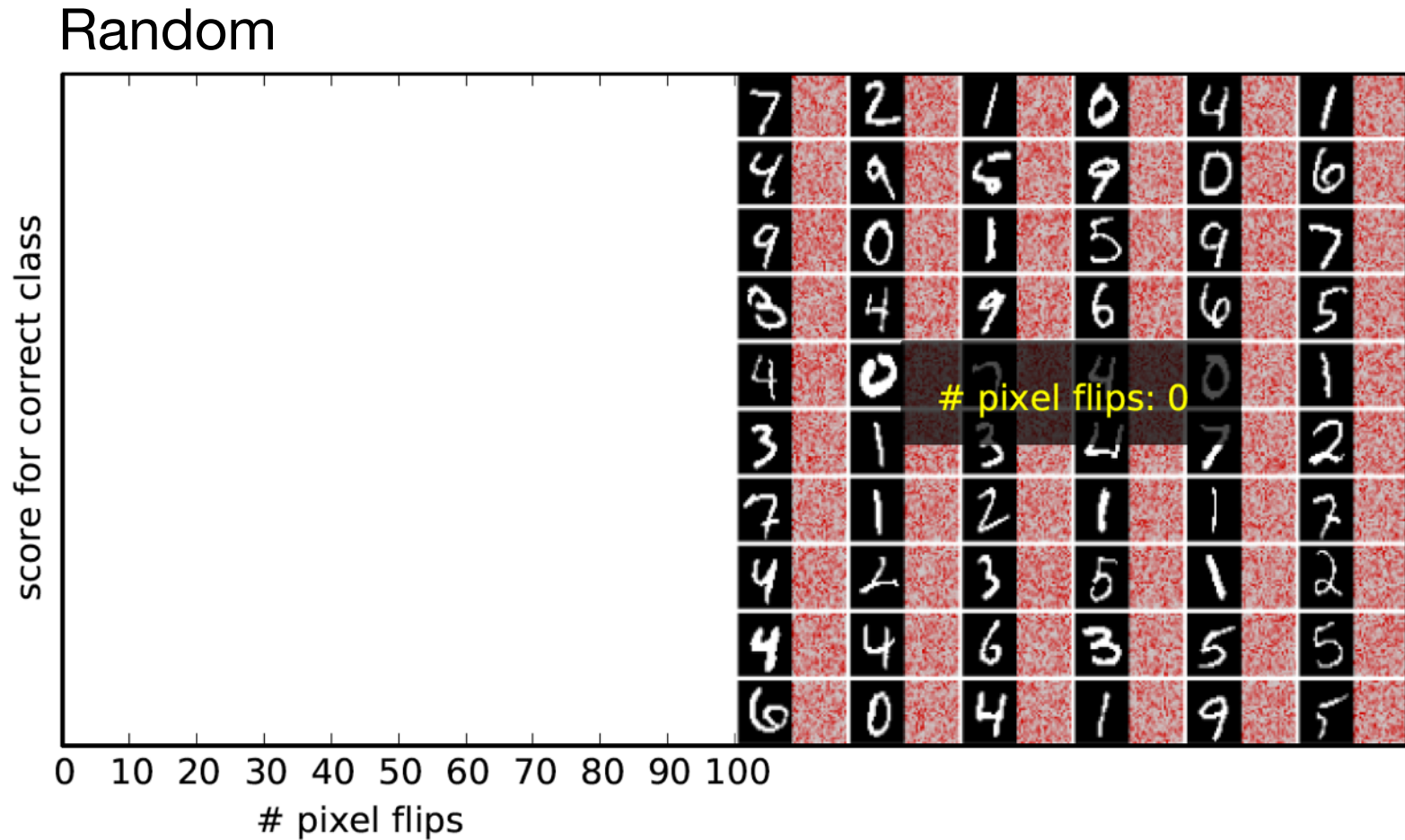
Compare Explanation Methods



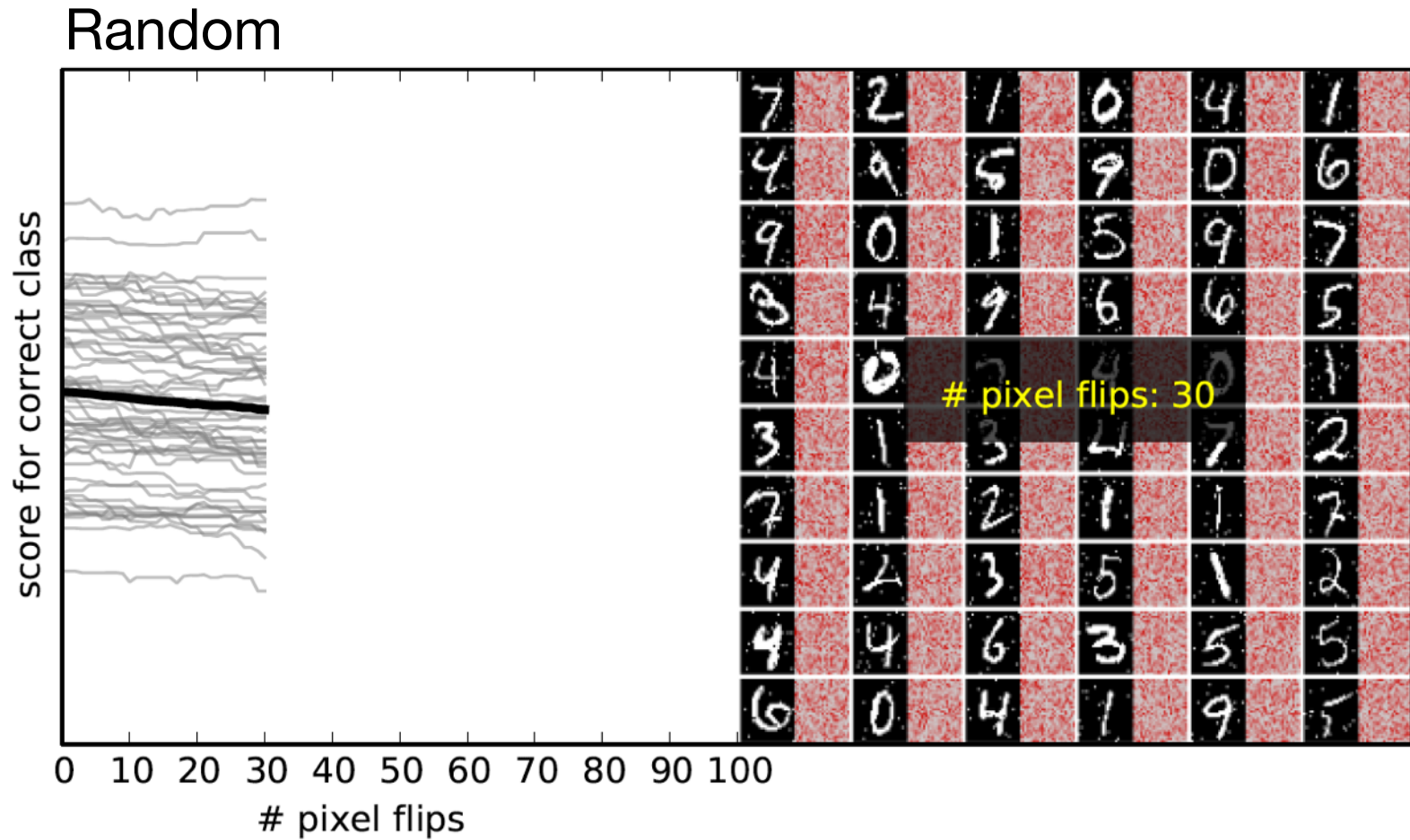
Compare Explanation Methods



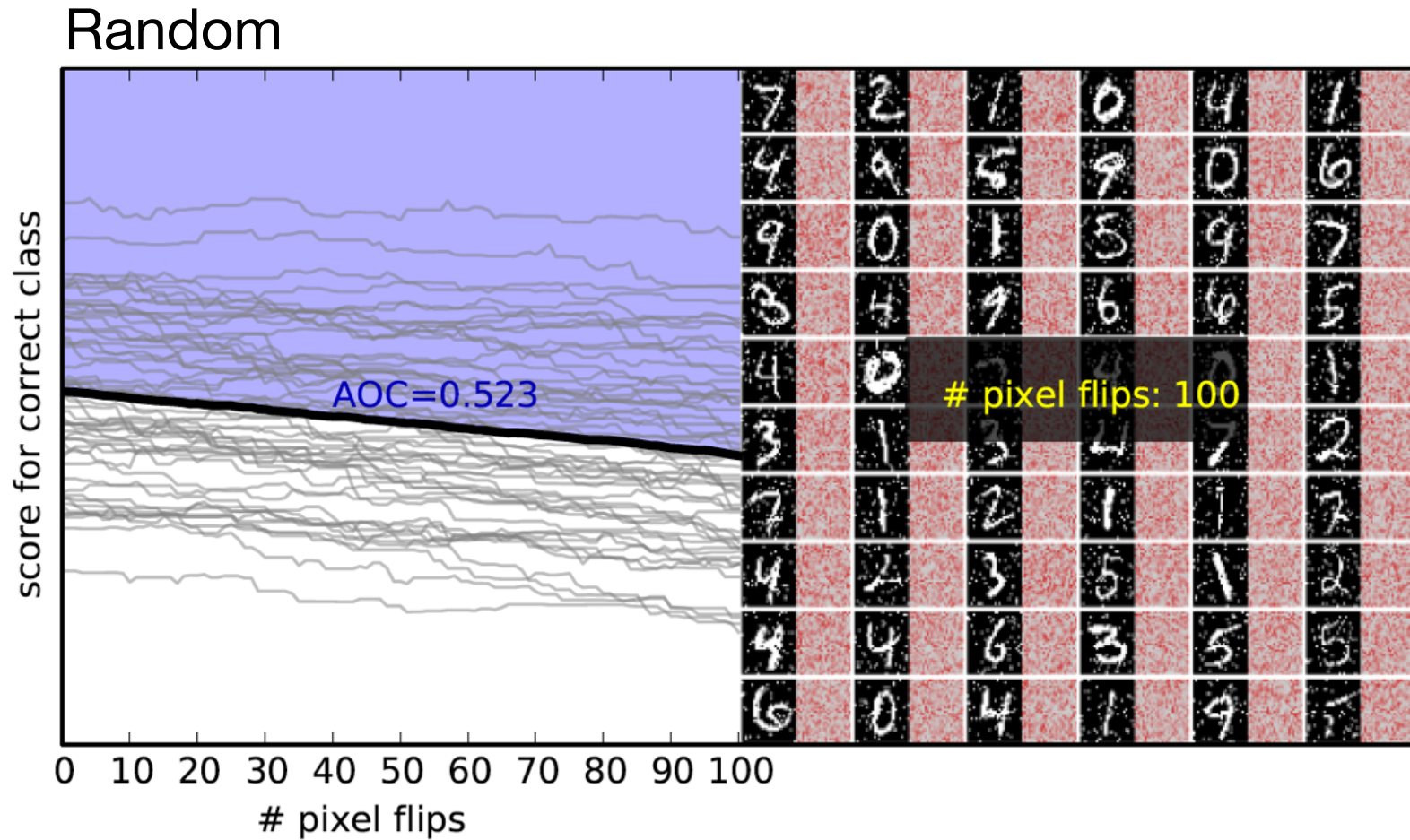
Compare Explanation Methods



Compare Explanation Methods



Compare Explanation Methods



Compare Explanation Methods

LRP:	0.722
Sensitivity:	0.691
Random:	0.523

LRP produces quantitatively better heatmaps than sensitivity analysis and random.

Compare Explanation Methods

LRP: **0.722**

Sensitivity: 0.691

Random: 0.523

LRP produces quantitatively better heatmaps than sensitivity analysis and random.

What about more complex datasets ?

SUN397



397 scene categories
(108,754 images in total)

ILSVRC2012



1000 categories
(1.2 million training images)

MIT Places



205 scene categories
(2.5 millions of images)

Compare Explanation Methods

Sensitivity Analysis
(Simonyan et al. 2014)

Deconvolution Method
(Zeiler & Fergus 2014)

LRP Algorithm
(Bach et al. 2015)



(Samek et al. 2016)

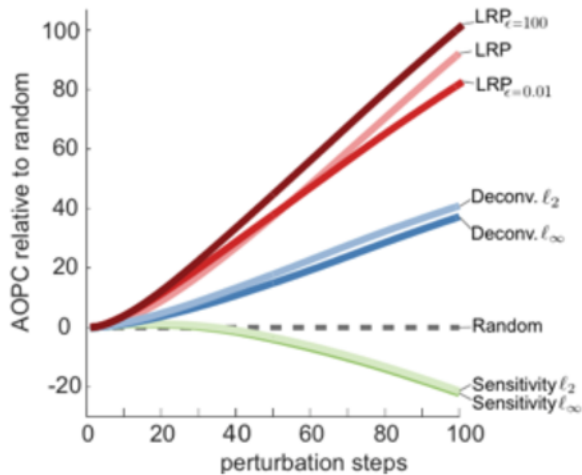
Compare Explanation Methods

Red: LRP method

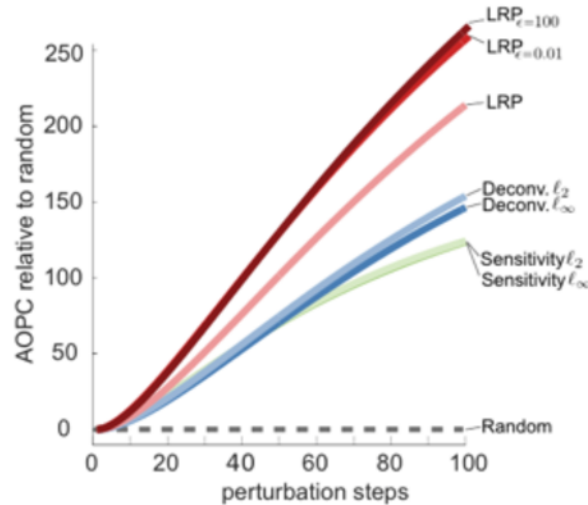
Blue: Deconvolution method (Zeiler & Fergus, 2014)

Green: Sensitivity method (Simonyan et al., 2014)

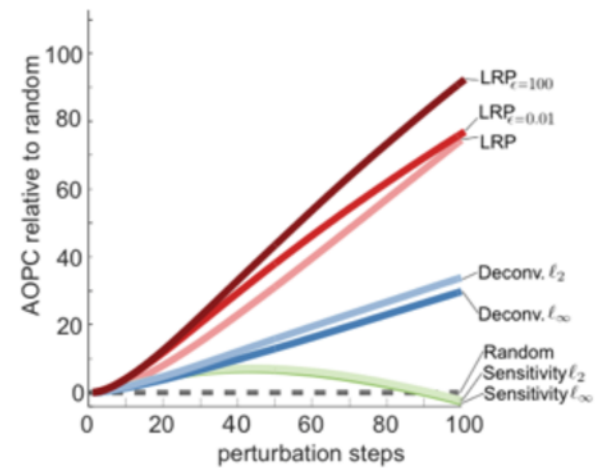
SUN397



ILSVRC2012



MIT Places



(Samek et al. 2016)

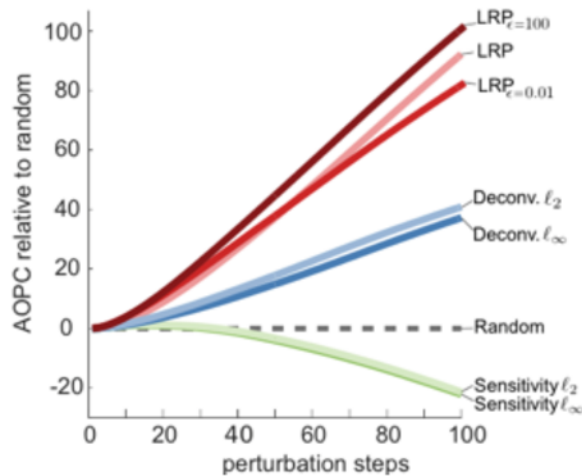
Compare Explanation Methods

Red: LRP method

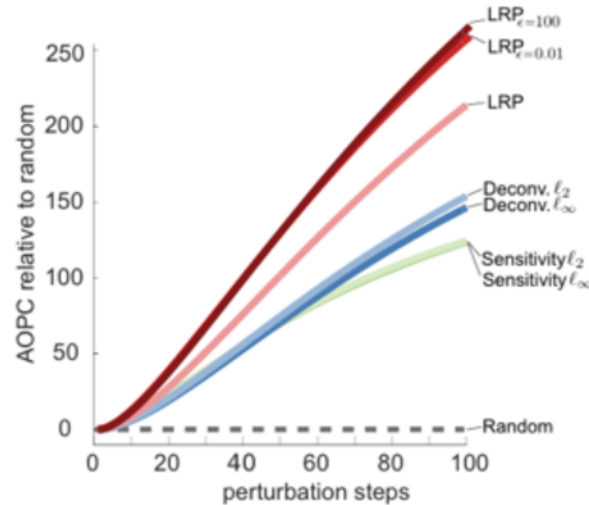
Blue: Deconvolution method (Zeiler & Fergus, 2014)

Green: Sensitivity method (Simonyan et al., 2014)

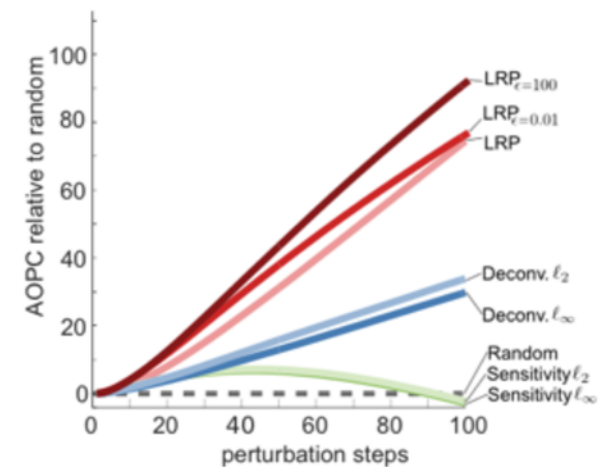
SUN397



ILSVRC2012



MIT Places



LRP produces quantitatively better heatmaps.

(Samek et al. 2016)

Compare Explanation Methods

Same idea can be applied for other domains (e.g. text document classification)

“Pixel flipping”
=
“Word deleting”

Compare Explanation Methods

Same idea can be applied for other domains (e.g. text document classification)

“Pixel flipping”
=
“Word deleting”

Text classified as “sci.med” —> LRP identifies most relevant words.

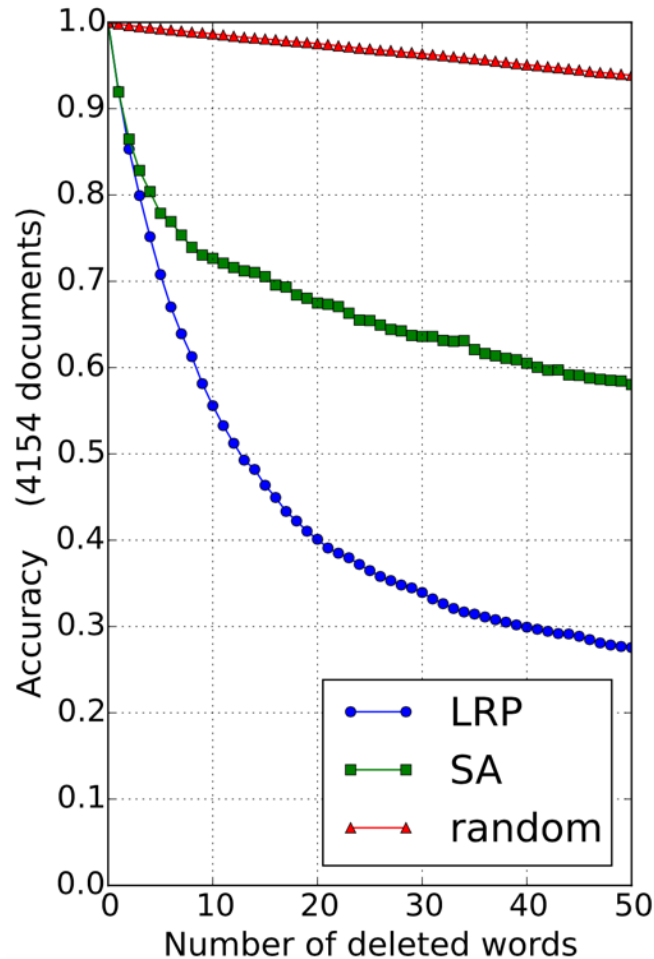
Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

sci.med (4.1)
>And what is the motion sickness
>that some astronauts occasionally experience?

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

(Arras et al. 2016)

Compare Explanation Methods



Deleting relevant words leads to a quick decrease of classifier accuracy.

The decrease is much steeper for LRP than for random word deletion and deletion according to sensitivity.

LRP better identifies relevant words.

(Arras et al. 2016)

Explanations and now ?

How good is the explanation ?

- Objective measure of quality
- Compare explanation methods

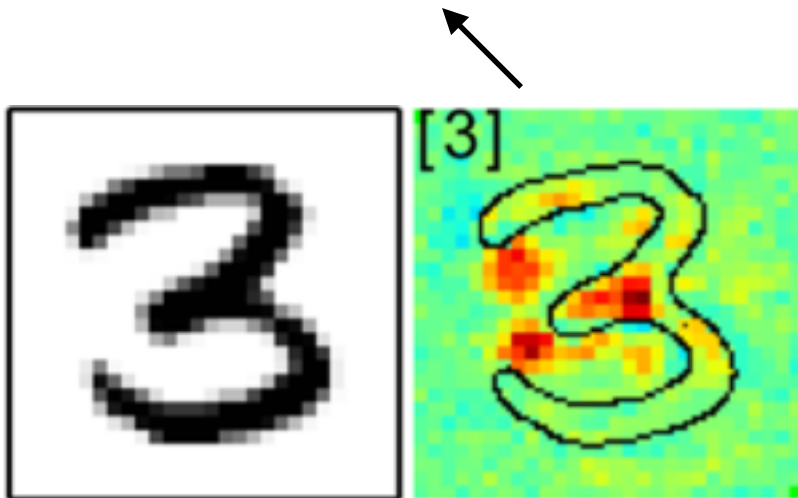
What can we do with it ?

- Compare classifiers
- Detect biases and flaws
- Quantify use of context
- Novel representation
- Application in the sciences

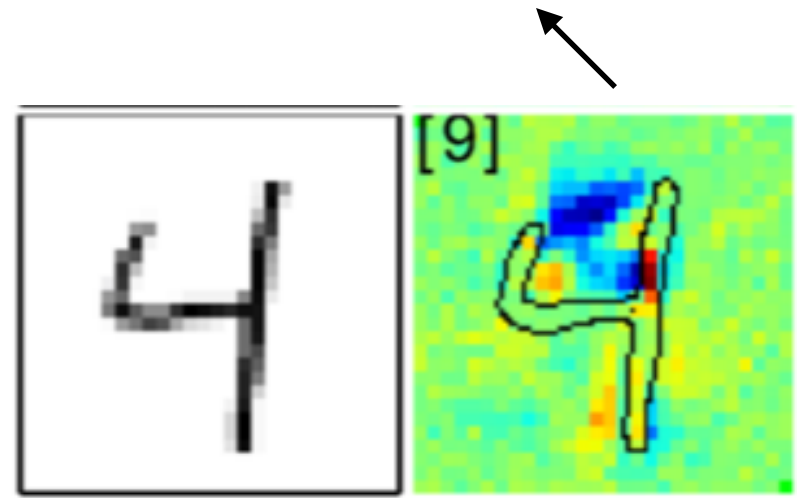
...

Opening the Black-Box

what speaks for / against
classification as “3”



what speaks for / against
classification as “9”



[number]: explanation target class

red color: evidence for prediction

blue color: evidence against prediction

Application: Compare Classifiers

20 Newsgroups data set

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Test set performance

word2vec / CNN model: 80.19%

Application: Compare Classifiers

20 Newsgroups data set

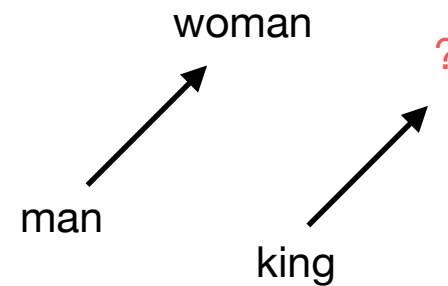
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

man king woman

$$\begin{array}{ccc} \downarrow & \downarrow & \downarrow \\ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix} & \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} & \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} \end{array}$$

Test set performance

word2vec / CNN model: 80.19%



Application: Compare Classifiers

20 Newsgroups data set

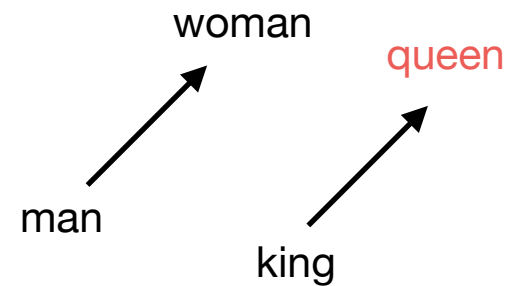
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

man king woman

$$\begin{array}{ccc} \downarrow & \downarrow & \downarrow \\ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix} & \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} & \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} \end{array}$$

Test set performance

word2vec / CNN model: 80.19%



Application: Compare Classifiers

20 Newsgroups data set

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

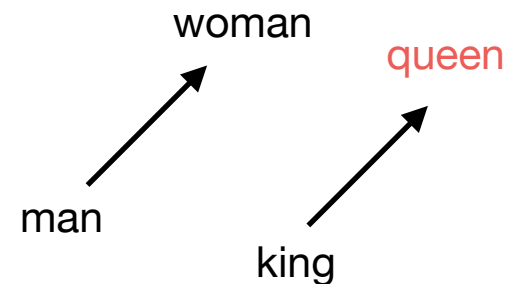
man king woman

$$\begin{array}{ccc} \downarrow & \downarrow & \downarrow \\ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix} & \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} & \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} \end{array}$$

Test set performance

word2vec / CNN model: 80.19%

BoW/SVM model: 80.10%



Application: Compare Classifiers

20 Newsgroups data set

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

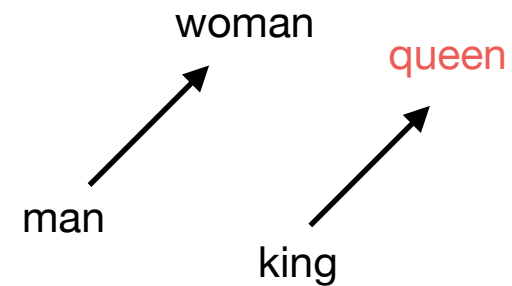
man king woman

$$\begin{array}{ccc} \downarrow & \downarrow & \downarrow \\ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{pmatrix} & \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} & \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{pmatrix} \end{array}$$

Test set performance

word2vec / CNN model: 80.19%

BoW/SVM model: 80.10%



same performance → same strategy ?

Application: Compare Classifiers

word2vec /
CNN model

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

sci.med (4.1)
>And what is the motion sickness
>that some astronauts occasionally experience?

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

Yes, weightlessness does feel like falling. It may feel strange at first, but the body does adjust. The feeling is not too different from that of sky diving.

sci.med (-0.6)
>And what is the motion sickness
>that some astronauts occasionally experience?

It is the body's reaction to a strange environment. It appears to be induced partly to physical discomfort and part to mental distress. Some people are more prone to it than others, like some people are more prone to get sick on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in space to try to see how to keep the number of occurrences down.

(Arras et al. 2016)

Application: Compare Classifiers

word2vec / CNN model

sci.med

symptoms (7.3), treatments (6.6), medication (6.4), osteopathy (6.3), ulcers (6.2), sciatica (6.0), hypertension (6.0), herb (5.6), doctor (5.4), physician (5.1), Therapy (5.1), antibiotics (5.1), Asthma (5.0), renal (5.0), medicines (4.9), caffeine (4.9), infection (4.9), gastrointestinal (4.8), therapy (4.8), homeopathic (4.7), medicine (4.7), allergic (4.7), dosages (4.7), esophagitis (4.7), inflammation (4.6), arrhythmias (4.6), cancer (4.6), disease (4.6), migraine (4.6), patients (4.5).

BoW/SVM model

sci.med

cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn (0.5), circumcision (0.5), syndrome (0.5), migraine (0.5), antibiotic (0.5), water (0.5), blood (0.5), fat (0.4), weight (0.4).

Words with maximum relevance

(Arras et al. 2016)

Application: Compare Classifiers

word2vec / CNN model

sci.med

symptoms (7.3), treatments (6.6), medication (6.4), osteopathy (6.3), ulcers (6.2), sciatica (6.0), hypertension (6.0), herb (5.6), doctor (5.4), physician (5.1), Therapy (5.1), antibiotics (5.1), Asthma (5.0), renal (5.0), medicines (4.9), caffeine (4.9), infection (4.9), gastrointestinal (4.8), therapy (4.8), homeopathic (4.7), medicine (4.7), allergic (4.7), dosages (4.7), esophagitis (4.7), inflammation (4.6), arrhythmias (4.6), cancer (4.6), disease (4.6), migraine (4.6), patients (4.5).

BoW/SVM model

sci.med

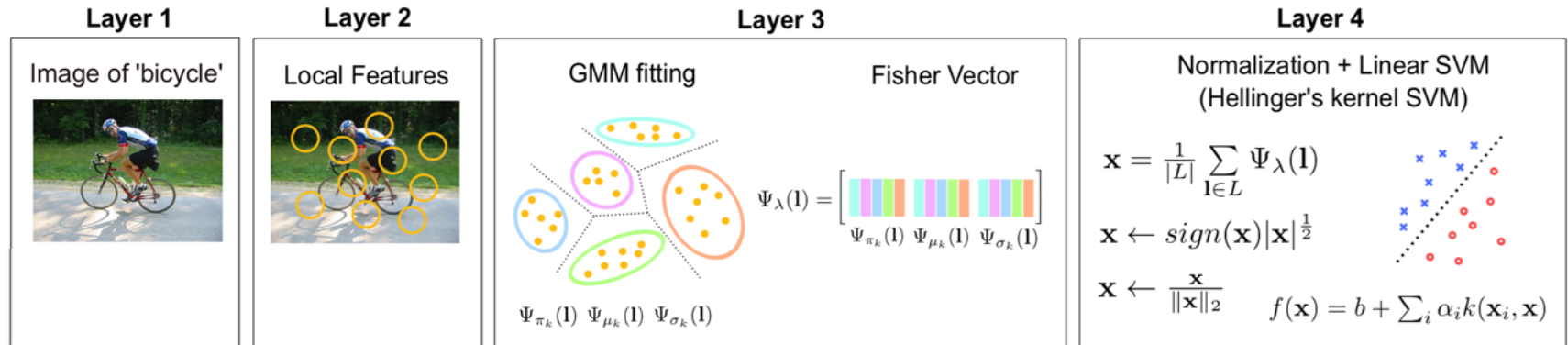
cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn (0.5), circumcision (0.5), syndrome (0.5), migraine (0.5), antibiotic (0.5), water (0.5), blood (0.5), fat (0.4), weight (0.4).

Words with maximum relevance

(Arras et al. 2016)

Application: Compare Classifiers

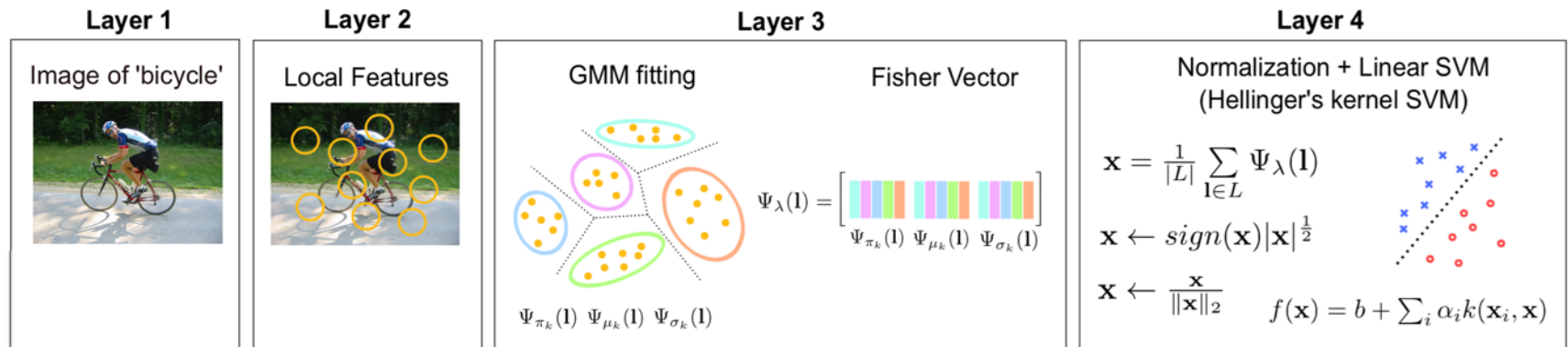
Fisher Vector / SVM Classifier



(Lapuschkin et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier

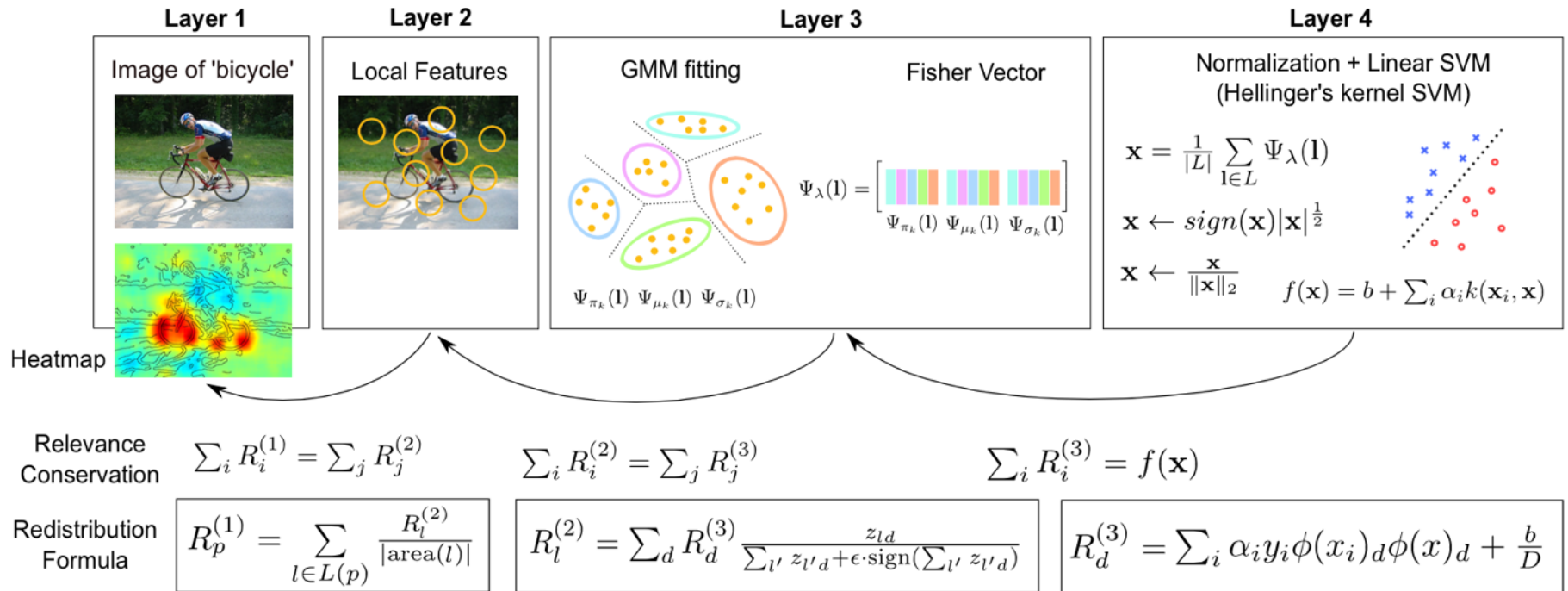


LRP general method for non-linear classifiers

(Lapuschkin et al. 2016)

Application: Compare Classifiers

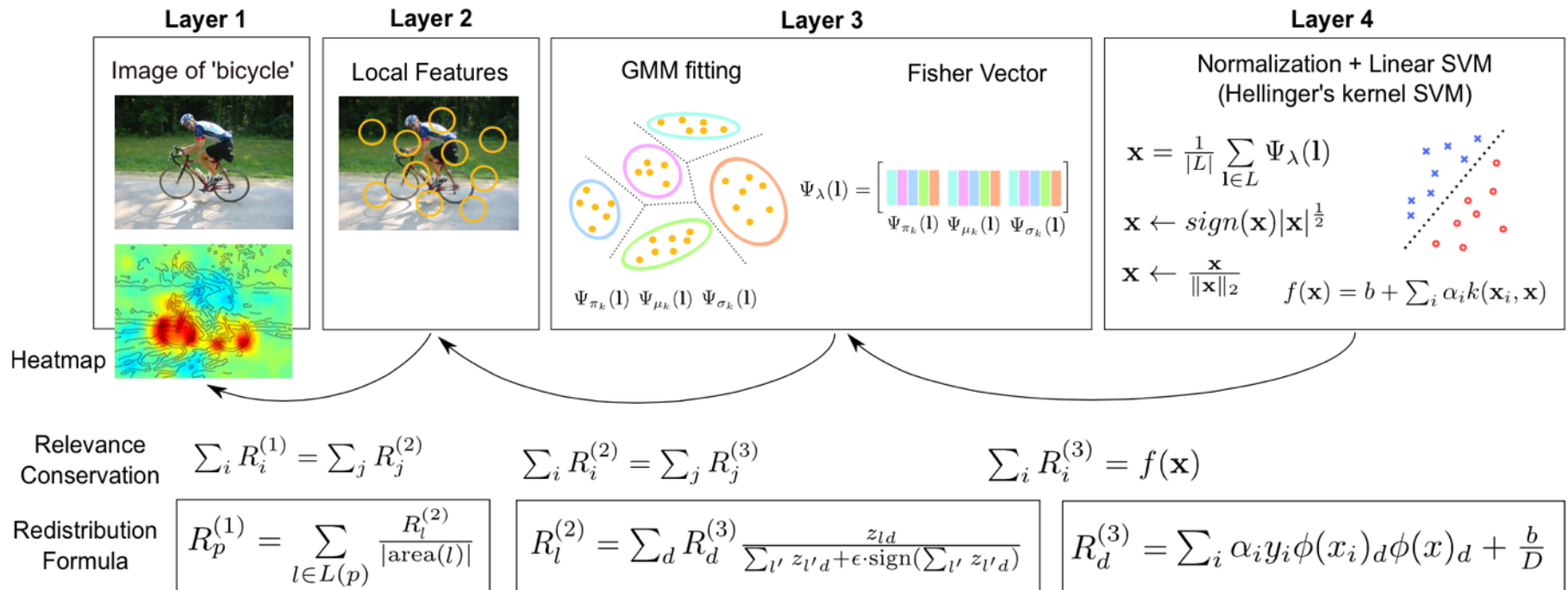
Fisher Vector / SVM Classifier



(Lapuschkin et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier



Deep Neural Network

- BVLC reference model + fine tuning.

Dataset

- PASCAL VOC 2007

(Lapuschkin et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier



Aeroplanes



Bicycles



Birds



Boats



Bottles



Buses



Cars



Cats



Chairs



Cows



Dining tables



Dogs



Horses



Motorbikes



People



Potted plants



Sheep



Sofas



Trains



TV/Monitors

Heatmap

Relevance

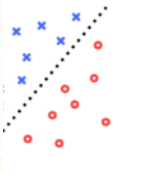
Consistency

Redistribution

Form

Linear SVM

SVM



$$\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

$$(x)_d + \frac{b}{D}$$

2007

(Lapuschkin et al. 2016)

Application: Compare Classifiers

Test error for various classes:

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

(Lapuschkin et al. 2016)

Application: Compare Classifiers

Test error for various classes:

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

same performance —> same strategy ?

(Lapuschkin et al. 2016)

Application: Compare Classifiers

Test error for various classes:

	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%

Image



same performance —> same strategy ?

(Lapuschkin et al. 2016)

Application: Compare Classifiers

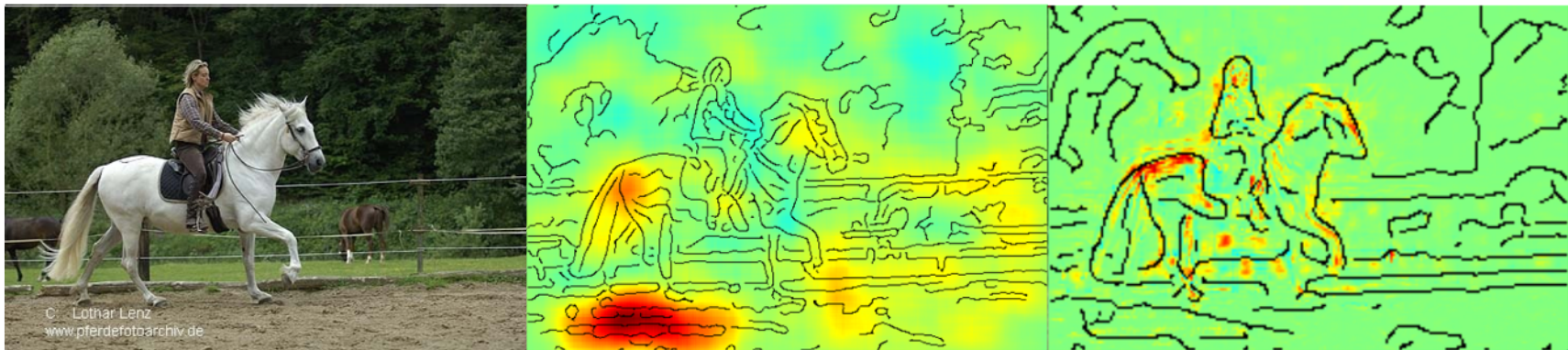
Test error for various classes:

	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%

Image

FV

DNN



same performance → same strategy ?

(Lapuschkin et al. 2016)

Application: Compare Classifiers

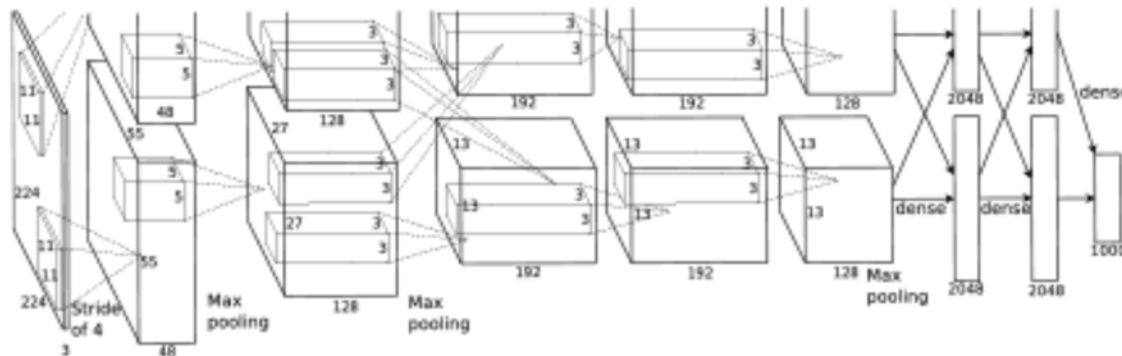
'horse' images in PASCAL VOC 2007



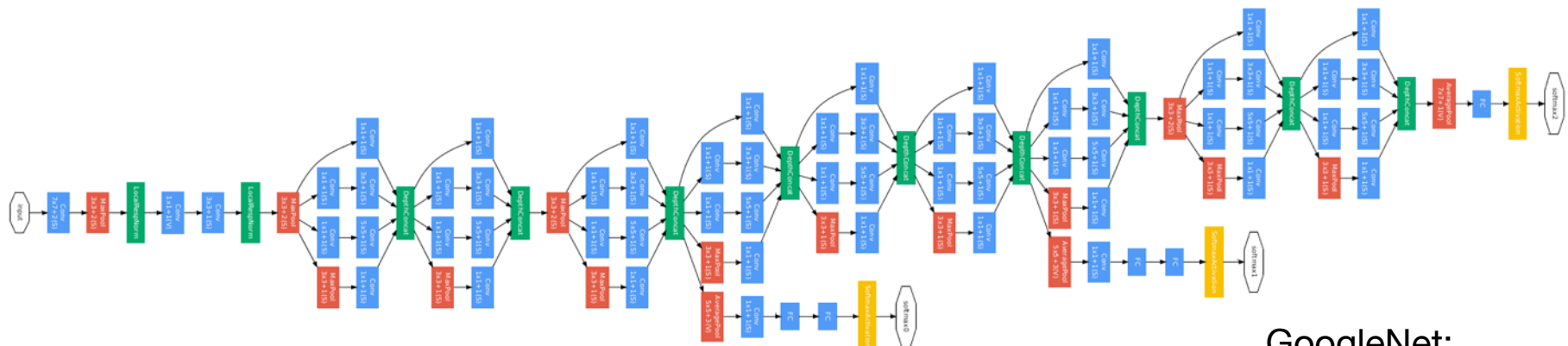
C: Lothar Lenz
www.pferdefotoarchiv.de



Application: Compare Classifiers



BVLC:
 - 8 Layers
 - ILSRCV: 16.4%



GoogleNet:
 - 22 Layers
 - ILSRCV: 6.7%
 - Inception layers

Application: Compare Classifiers



GoogLeNet focuses on faces of animal.
—> suppresses background noise

(Binder et al. 2016)

Application: Measure Context Use



how important
is context ?



how important
is context ?

classifier



Application: Measure Context Use



how important
is context ?



how important
is context ?

classifier

**LRP decomposition allows
meaningful pooling over bbox !**

Application: Measure Context Use



how important
is context ?



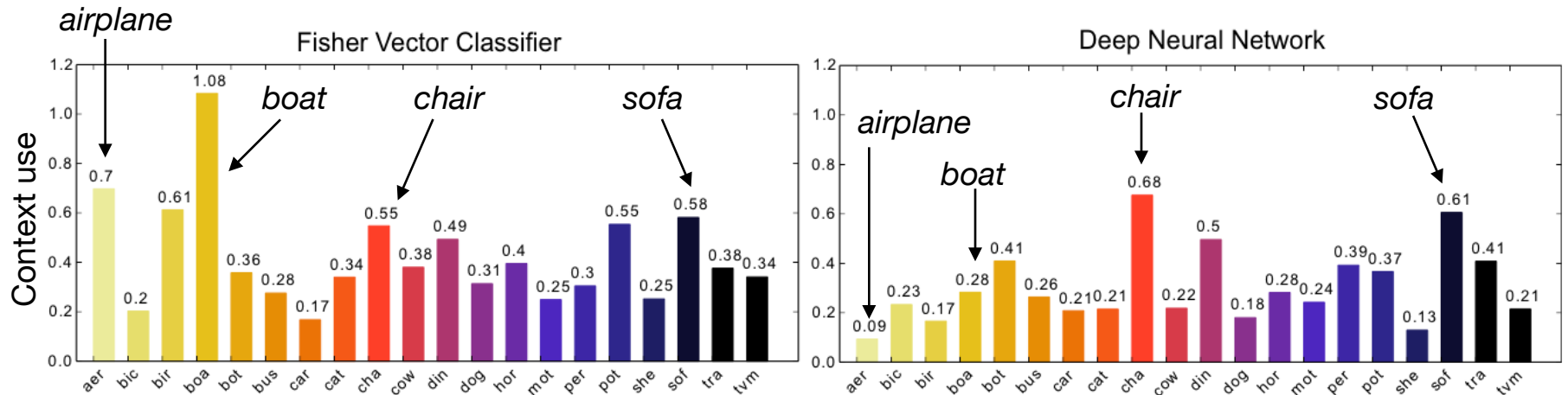
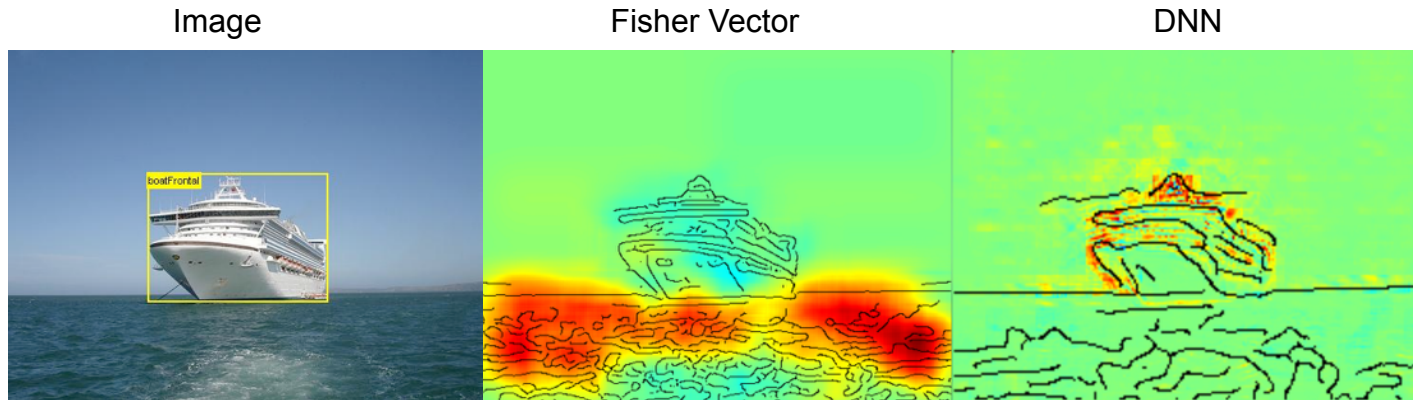
how important
is context ?

classifier

**LRP decomposition allows
meaningful pooling over bbox !**

$$\text{importance of context} = \frac{\text{relevance outside bbox}}{\text{relevance inside bbox}}$$

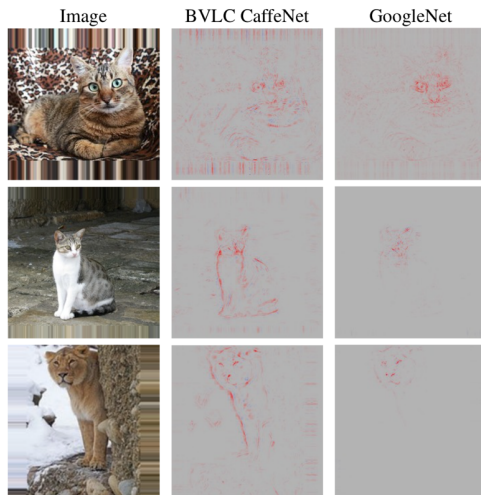
Application: Measure Context Use



Large values indicate importance of context

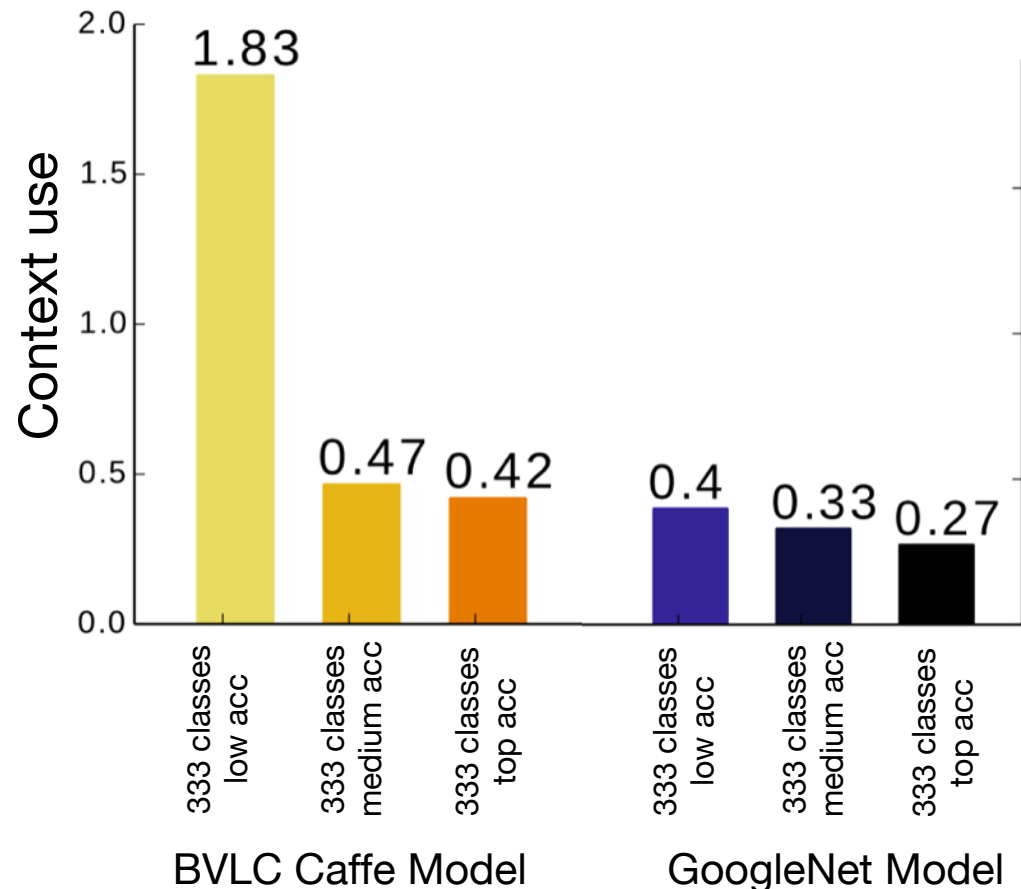
(Lapuschkin et al. 2016)

Application: Measure Context Use



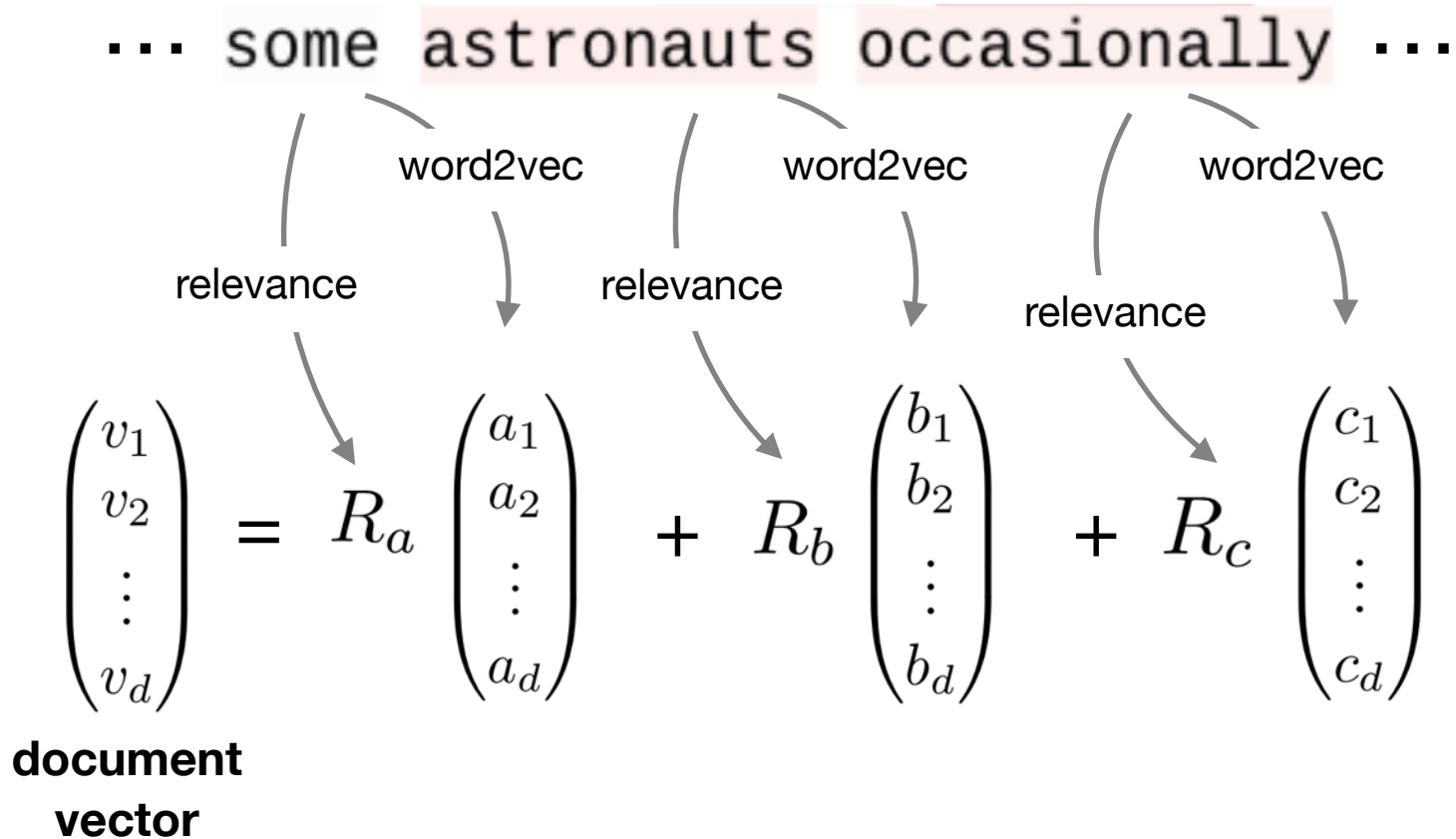
GoogleNet uses less context than BVLC model.

Context use anti-correlated with performance.



(Lapuschkin et al. 2016)

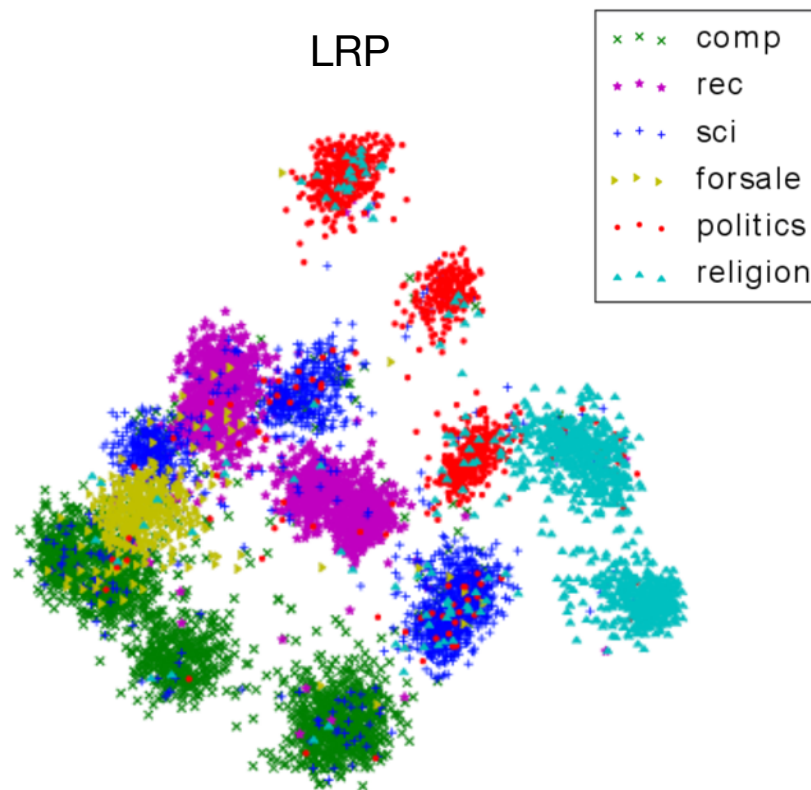
Application: Novel Representation



(Arras et al. 2016)

Application: Novel Representation

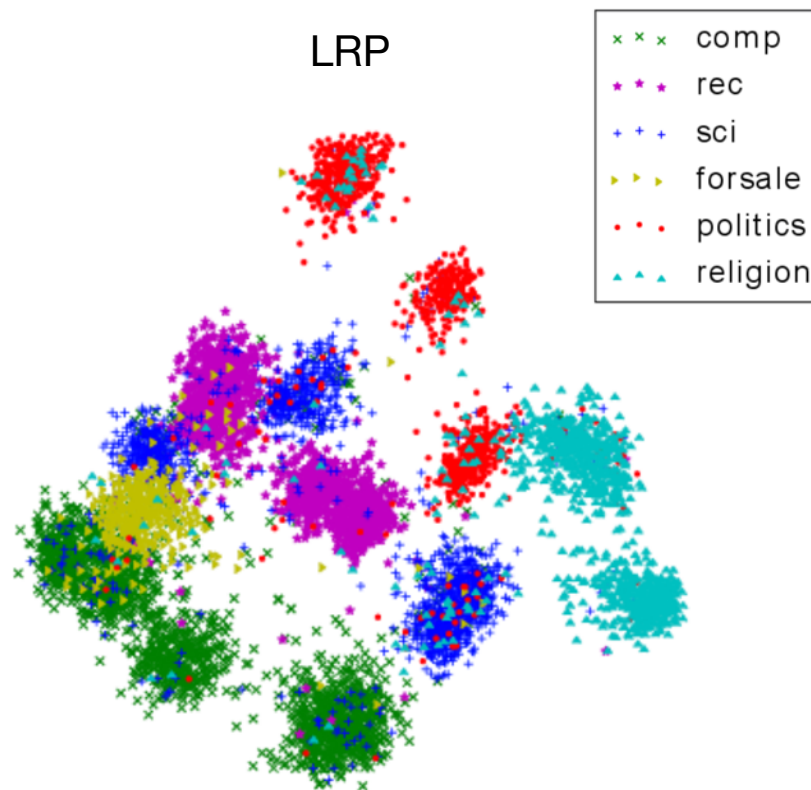
2D PCA projection of
document vectors



(Arras et al. 2016)

Application: Novel Representation

2D PCA projection of
document vectors



uniform



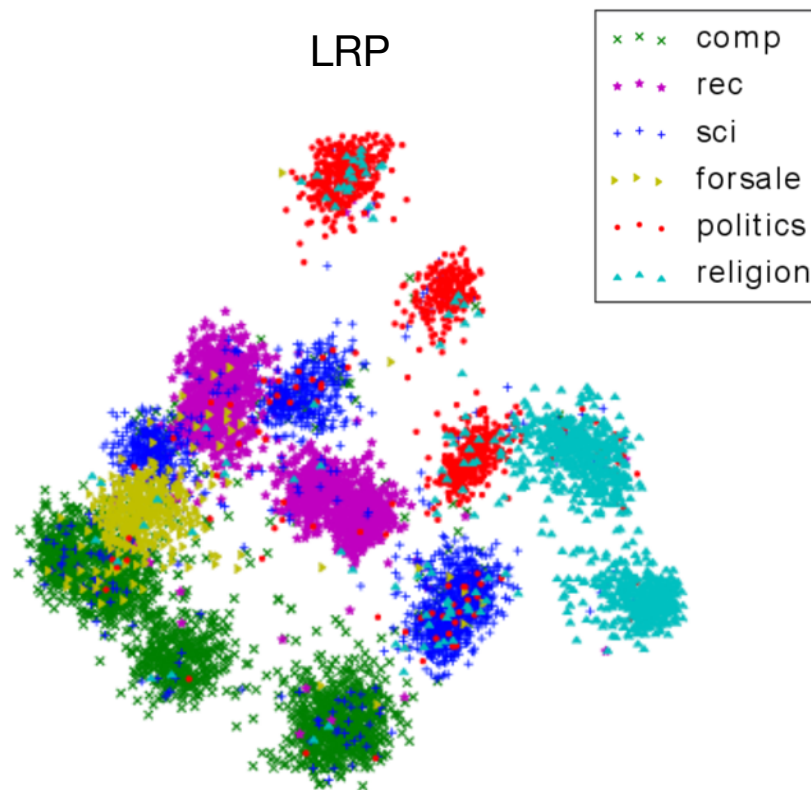
TFIDF



(Arras et al. 2016)

Application: Novel Representation

2D PCA projection of
document vectors



uniform



TFIDF

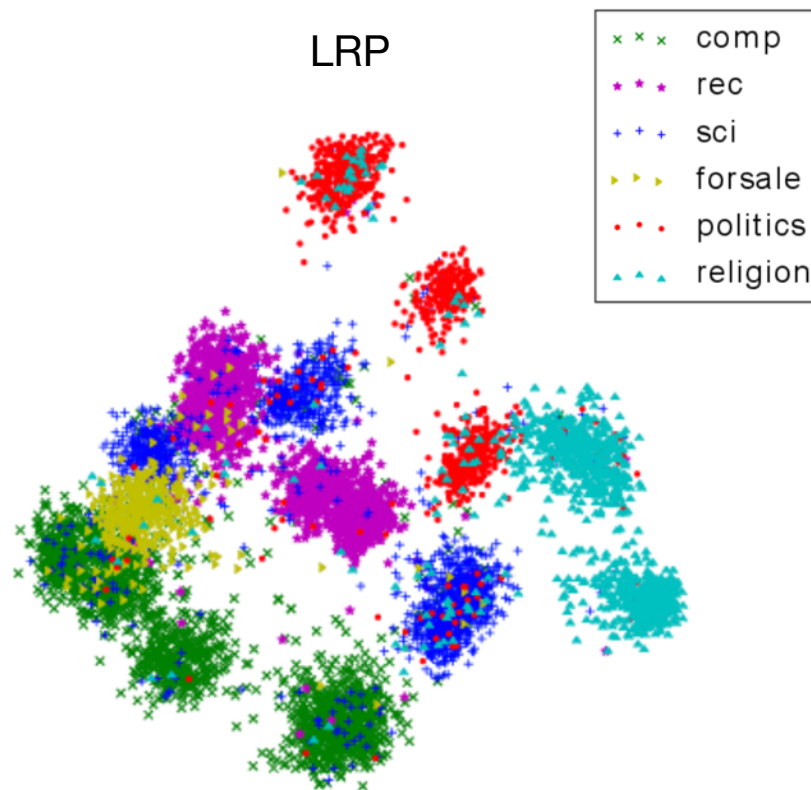


Document vector
computation is unsupervised.

(Arras et al. 2016)

Application: Novel Representation

2D PCA projection of document vectors



uniform



TFIDF



Document vector computation is unsupervised.

KNN-Performance on document vectors
(Explanatory Power Index)

LRP: 0.8076

TFIDF: 0.6816

uniform: 0.6208 (Arras et al. 2016)

Application: Sentiment Analysis

How to handle multiplicative interactions ?

$$z_j = z_g \cdot z_s$$

$$R_g = 0 \quad R_s = R_j$$

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

--	--	1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .

Positive sentiment

++	++	19.	a worthy entry into a very difficult genre .
		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Application: Sentiment Analysis

How to handle multiplicative interactions ?

$$z_j = z_g \cdot z_s$$

$$R_g = 0 \quad R_s = R_j$$

gate neuron indirectly affect relevance distribution in forward pass

Negative sentiment

--	--	1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .

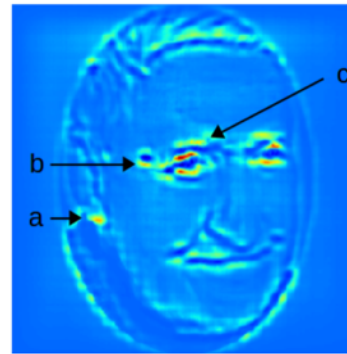
Positive sentiment

++	++	19.	a worthy entry into a very difficult genre .
		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

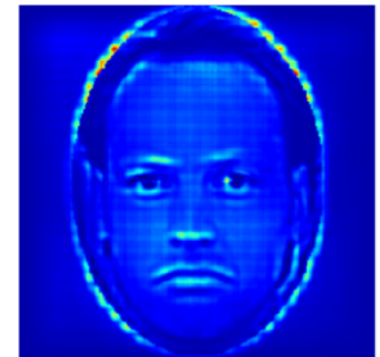
(Arras et al., 2017)

Application: Face Analysis

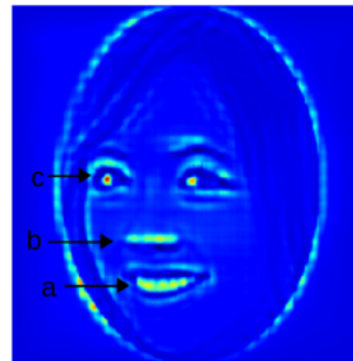
Identifying
age-related
features



Emotions



Attractivity



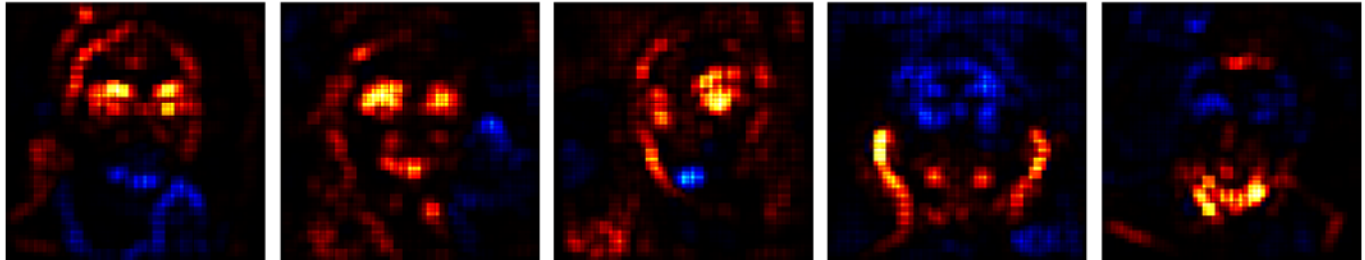
(Arbabzadah et al. 2016)

Application: Face Analysis

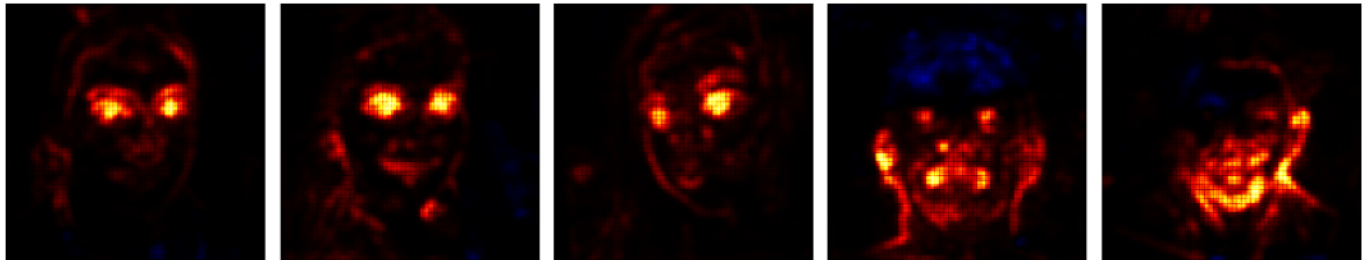
Gender
Classification



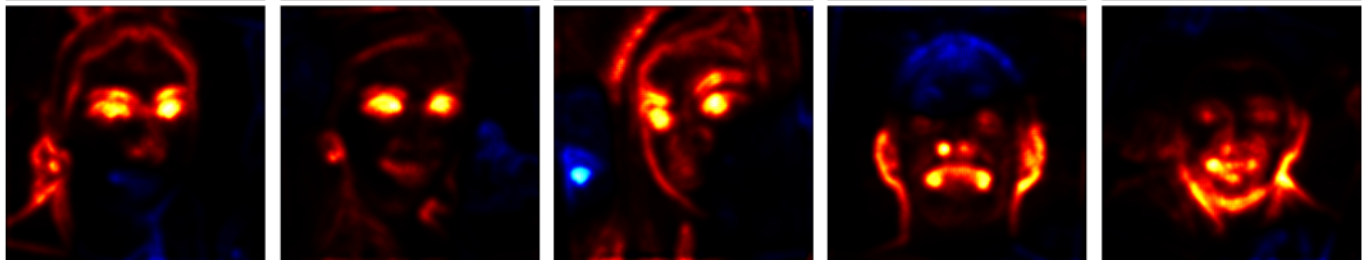
CaffeNet



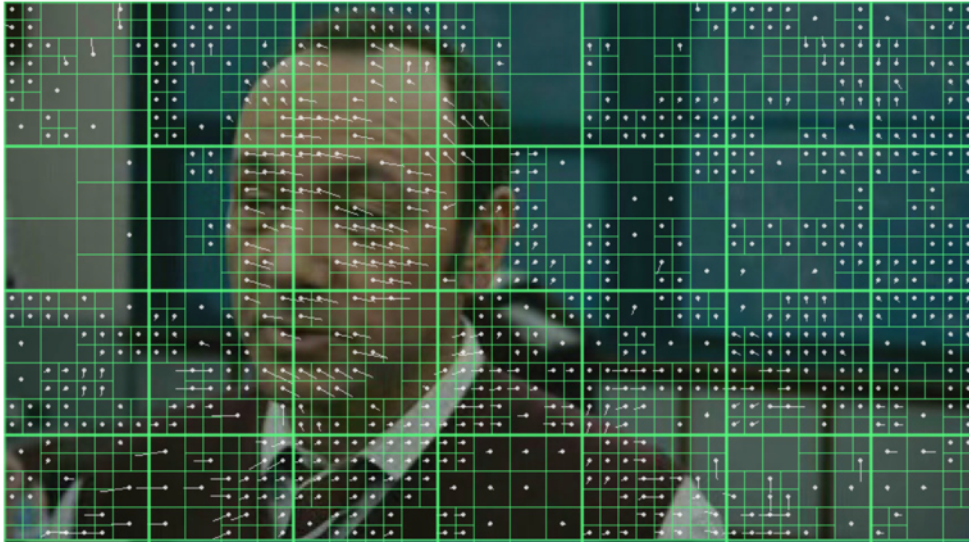
GoogleNet



VGG16



Application: Video Analysis

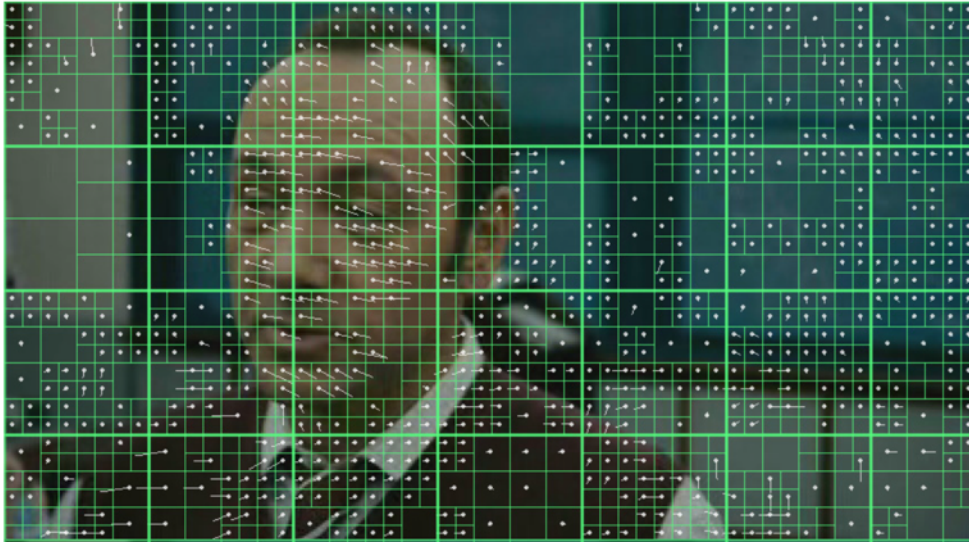


Motion vectors can be extracted from the compressed video
-> allows very efficient analysis



(Srinivasan et al. 2017)

Application: Video Analysis



Motion vectors can be extracted from the compressed video
-> allows very efficient analysis



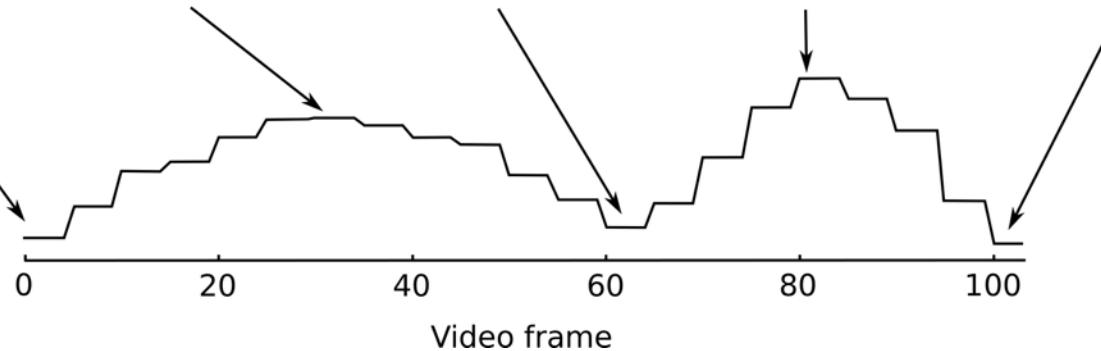
(Srinivasan et al. 2017)

Application: Video Analysis

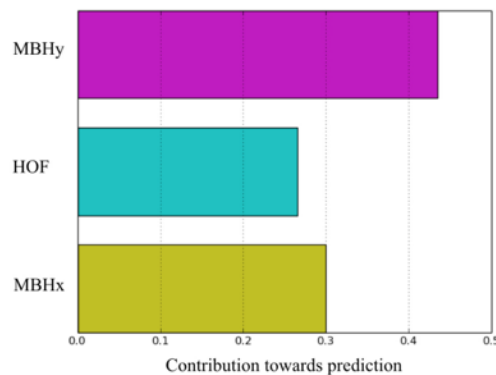
Explaining prediction: "sit-up"



LRP relevances
per frame



Which features are most relevant ?



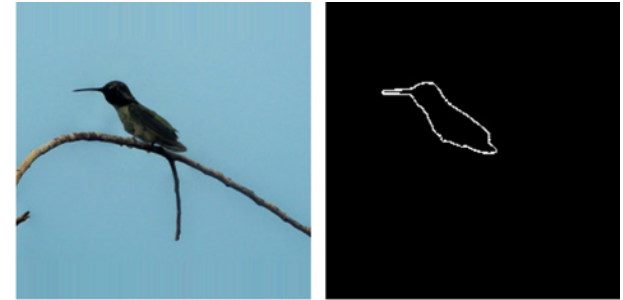
(Srinivasan et al. 2017)

Application: Semantic Boundary Detection

Pixel-wise labelling are very costly

Idea: Use image-label labels

+ apply explanation methods

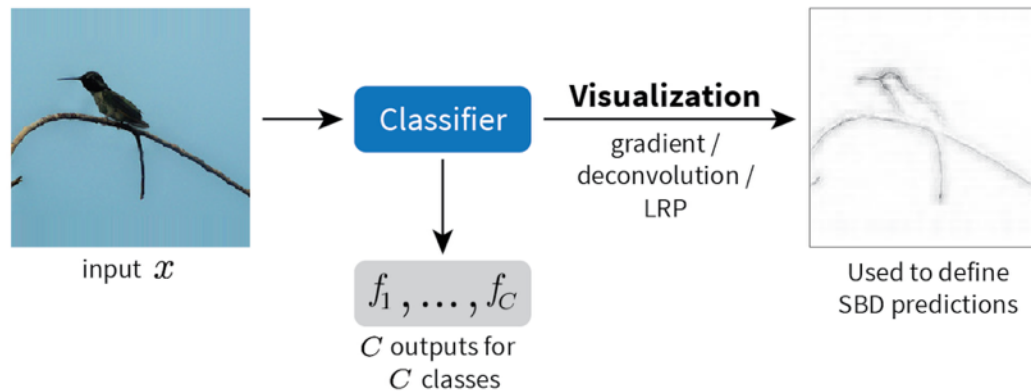
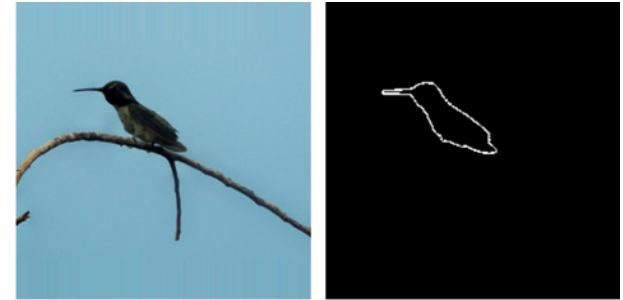


(Koh et al. 2017)

Application: Semantic Boundary Detection

Pixel-wise labelling are very costly

Idea: Use image-label labels
+ apply explanation methods

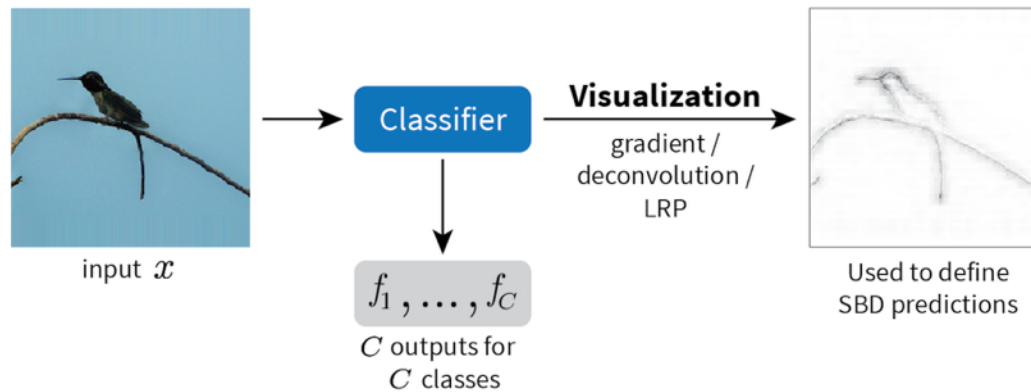


(Koh et al. 2017)

Application: Semantic Boundary Detection

Pixel-wise labelling are very costly

Idea: Use image-label labels
+ apply explanation methods



training phase: Method:	image-level labels						pixel-level labels	
	Gradient	Deconv	$\beta = 0$	$\beta = -1$	$\epsilon = 1$	$\epsilon = 0.01$	InvDet [1]	HFL [5]
AP	22.5	25.0	28.4	27.3	31.4	31.2	19.9	54.6
MF	31.0	33.3	35.1	34.1	38.0	38.1	28.0	62.5

(Koh et al. 2017)

Application: Interpretability in the Sciences

Brain-Computer Interfacing



How brain works subject-dependent
—> individual explanations



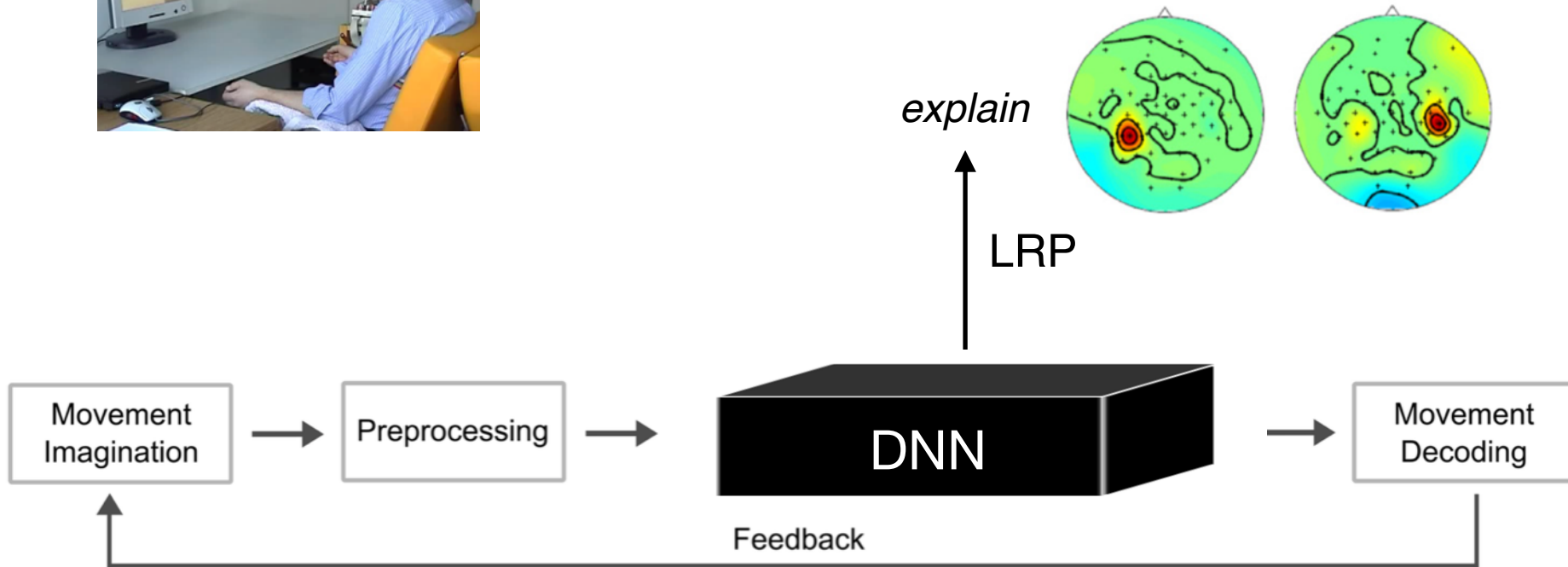
(Sturm et al. 2016)

Application: Interpretability in the Sciences

Brain-Computer Interfacing



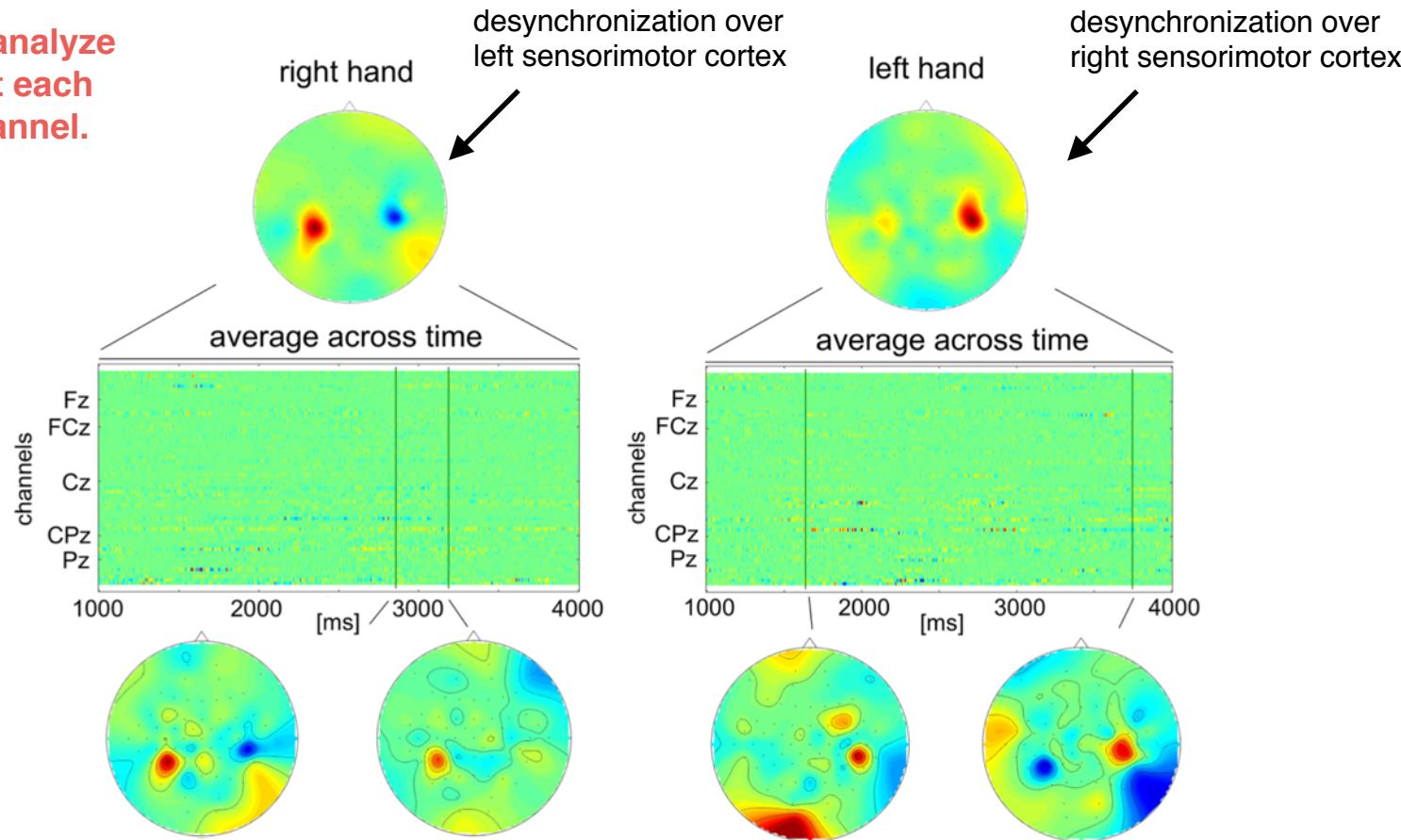
How brain works subject-dependent
—> individual explanations



(Sturm et al. 2016)

Application: Interpretability in the Sciences

With LRP we can analyze relevant activity at each time point and channel.

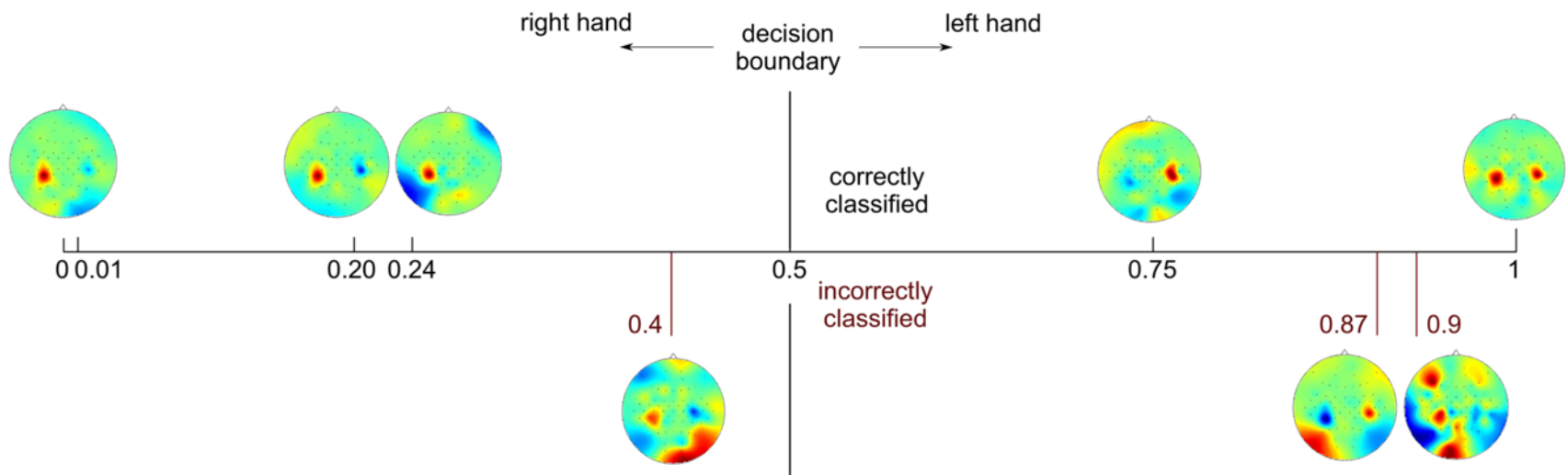


Allows to spatially & temporally identify important activity in EEG data.

(Sturm et al. 2016)

Application: Interpretability in the Sciences

With LRP we can analyze what made a trial being misclassified.

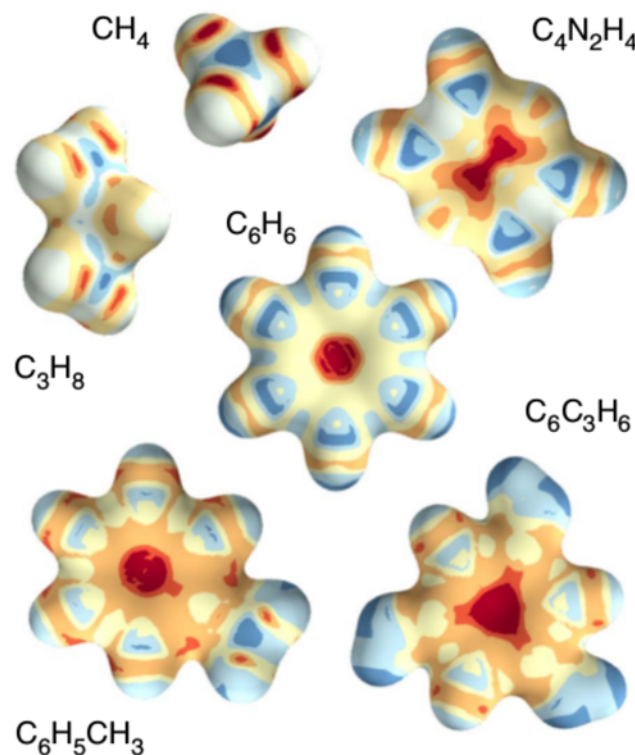
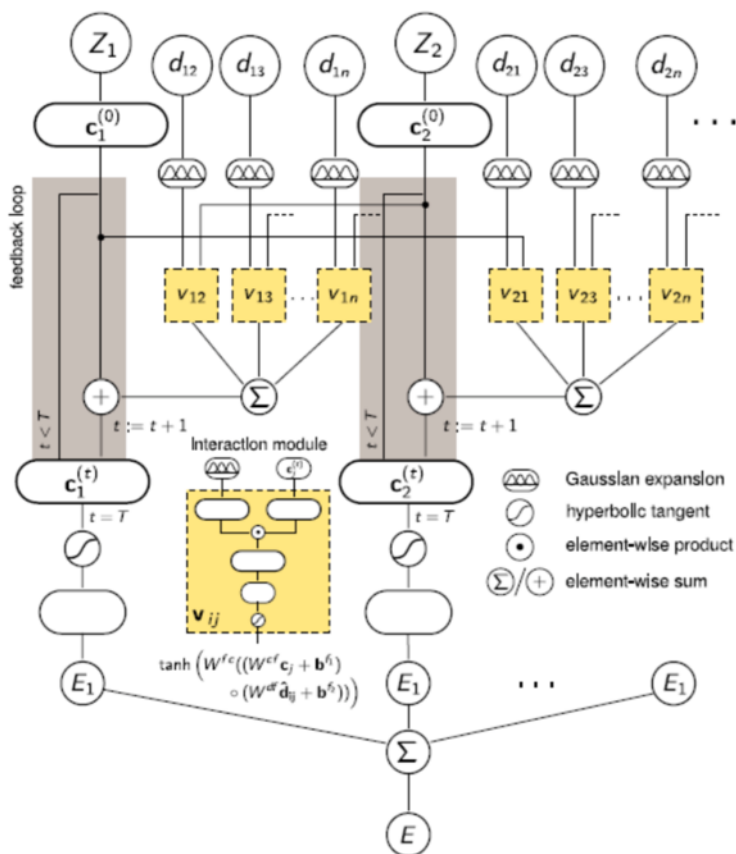


(Sturm et al. 2016)

Application: Interpretability in the Sciences

Deep Tensor Network predicts molecular properties with state-of-the-art accuracy.

Effect of energy of test charge
—> interpretable for human expert



(Schütt et al. 2017)

Summary

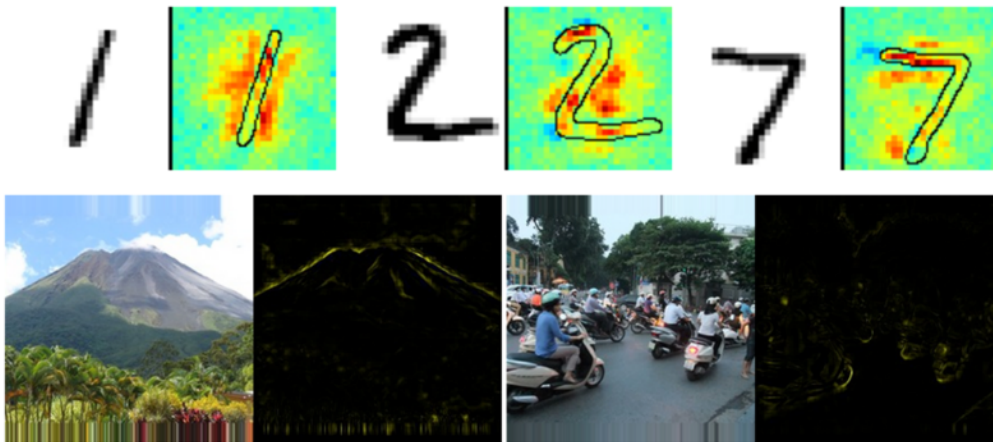
- In many problems interpretability as important as prediction.
- Explaining individual predictions is key.
- We have powerful, mathematically well-founded methods (LRP / deep Taylor) to explain individual predictions.
- How can we use interpretability for improving models (focus of NIPS'17 Workshop “Interpretability - Now what ?”)
- Many interesting applications with interpretable deep nets
—> more to come soon !

Thank you for your attention

Visit:

<http://www.heatmapping.org>

- ▶ Tutorials
- ▶ Software
- ▶ Online Demos



For more information, check out our tutorial paper:

Montavon et al. “Methods for Interpreting and Understanding Deep Neural Networks”

<https://arxiv.org/abs/1706.07979>

References

- F Arbabzadeh, G Montavon, KR Müller, W Samek. Identifying Individual Facial Expressions by Deconstructing a Neural Network. *Pattern Recognition - 38th German Conference, GCPR 2016*, Lecture Notes in Computer Science, 9796:344-54, Springer International Publishing, 2016.
- L Arras, F Horn, G Montavon, KR Müller, W Samek. Explaining Predictions of Non-Linear Classifiers in NLP. *Workshop on Representation Learning for NLP*, Association for Computational Linguistics, 1-7, 2016.
- L Arras, F Horn, G Montavon, KR Müller, W Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *PLOS ONE*, 12(8):e0181142, 2017.
- L Arras, G Montavon, K-R Müller, W Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. *EMNLP'17 Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)*, 1-10, 2017.
- S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015.
- S Bach, A Binder, KR Müller, W Samek. Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth. *IEEE International Conference on Image Processing (ICIP)*, 2271-75, 2016.
- A Binder, G Montavon, S Lapuschkin, KR Müller, W Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *Artificial Neural Networks and Machine Learning – ICANN 2016*, Part II, Lecture Notes in Computer Science, Springer-Verlag, 9887:63-71, 2016.
- A Binder, W Samek, G Montavon, S Bach, KR Müller. Analyzing and Validating Neural Networks Predictions. *ICML Workshop on Visualization for Deep Learning*, 2016
- J Y Koh, W Samek, KR Müller, A Binder. "Object Boundary Detection and Classification with Image-level Labels" *Pattern Recognition - 39th German Conference, GCPR 2017*, Lecture Notes in Computer Science, Springer International Publishing, 2017.

References

- S Lapuschkin, A Binder, G Montavon, KR Müller, Wojciech Samek. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912-20, 2016.
- S Lapuschkin, A Binder, KR Müller, W Samek. Understanding and Comparing Deep Neural Networks for Age and Gender Classification. *ICCV'17 Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, 2017.
- S Lapuschkin, A Binder, G Montavon, KR Müller, W Samek. The Layer-wise Relevance Propagation Toolbox for Artificial Neural Networks. *Journal of Machine Learning Research*, 17(114):1-5, 2016.
- G Montavon, S Bach, A Binder, W Samek, KR Müller. DeepTaylor Decomposition of Neural Networks. *ICML Workshop on Visualization for Deep Learning*, 2016.
- G Montavon, S Bach, A Binder, W Samek, KR Müller. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222, 2017.
- G Montavon, W Samek, KR Müller. Methods for Interpreting and Understanding Deep Neural Networks. *arXiv:1706.07979*, 2017.
- W Samek, A Binder, G Montavon, S Lapuschkin, KR Müller. Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2017.
- W Samek, G Montavon, A Binder, S Lapuschkin, KR Müller. Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation. *NIPS'16 Workshop on Interpretable ML for Complex Systems*, 1-5, 2016
- V Srinivasan, S Lapuschkin, C Hellge, KR Müller, W Samek. Interpretable human action recognition in compressed domain. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.
- I Sturm, S Lapuschkin, W Samek, KR Müller. Interpretable Deep Neural Networks for Single-Trial EEG Classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.