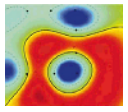
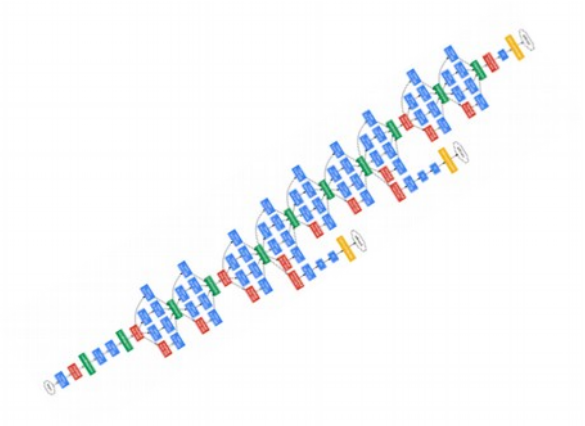
SoSe 2018: Deep Neural Networks

Lecture 6: Explanations

Machine Learning Group
Technische Universität Berlin



Deep Neural Networks

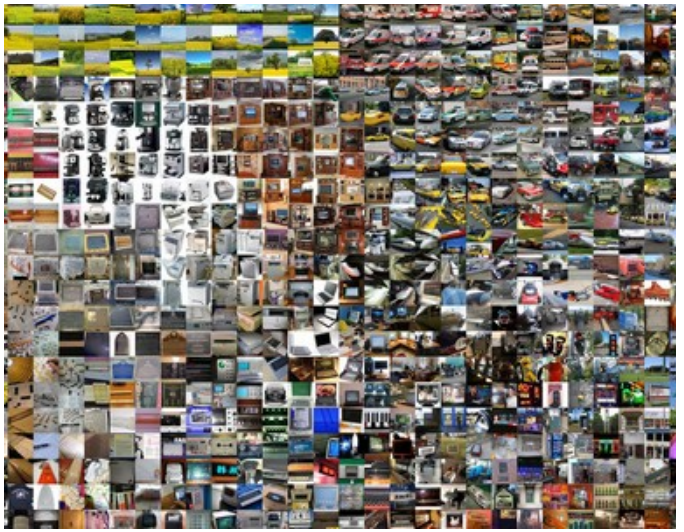


neural nets enable

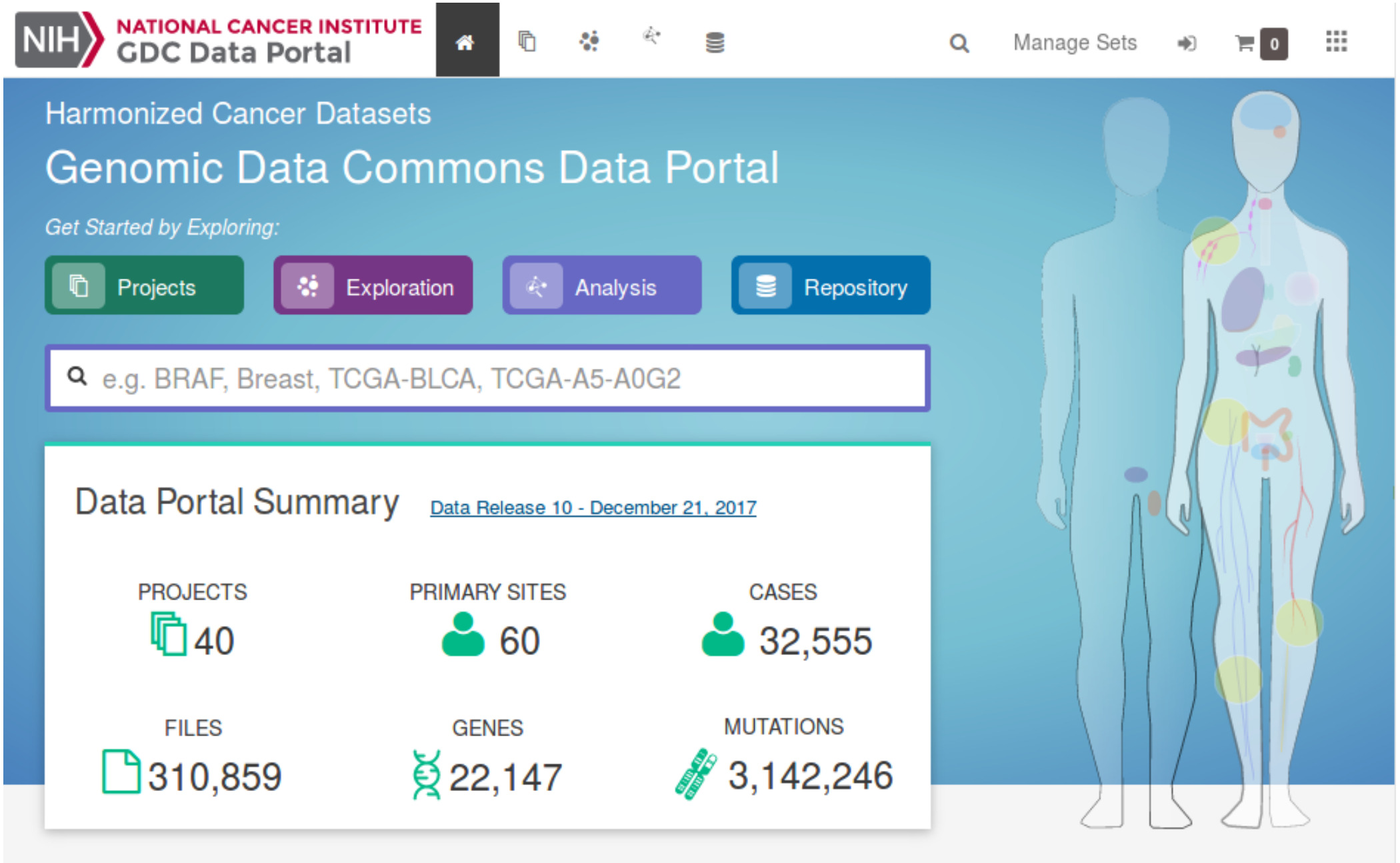
big models

GPUs + fast algorithms

enable **training** these models



What about the training data?


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Manage Sets
0
Harmonized Cancer Datasets
Genomic Data Commons Data Portal
Get Started by Exploring:
Projects
Exploration
Analysis
Repository
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2
Data Portal Summary
Data Release 10 - December 21, 2017
PROJECTS
40
PRIMARY SITES
60
CASES
32,555
FILES
310,859
GENES
22,147
MUTATIONS
3,142,246

“Big Data” and “Big Confusion”

Observation: Many confounding factors in practice.

different qualities of measurement

different subjects

technology evolves over time

time of the day

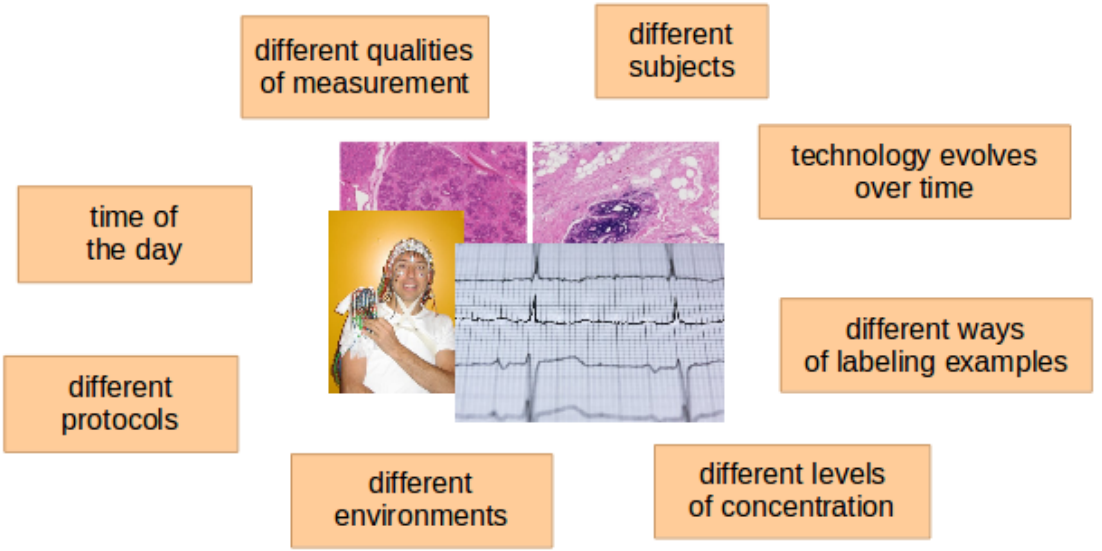
different ways of labeling examples

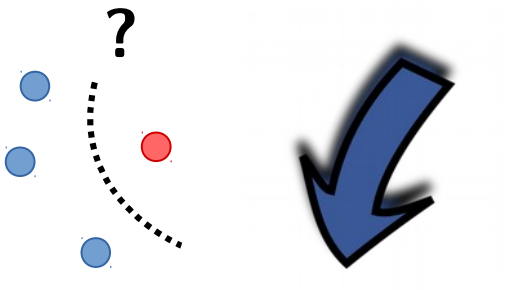
different protocols

different environments

different levels of concentration

Avoiding “Big Confusion”



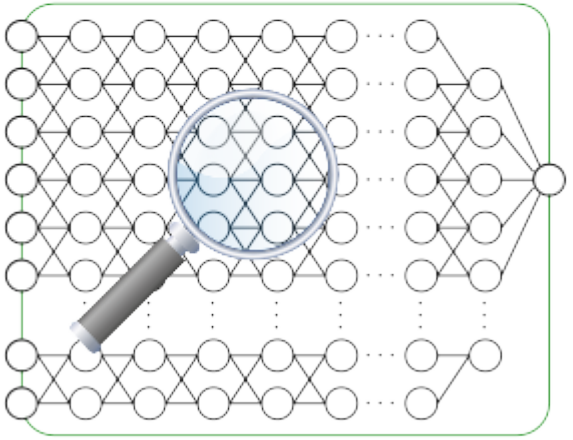


Give up on big data, use a controlled setting, and make careful use of “little” data.

Inspect what the big data model has learned.

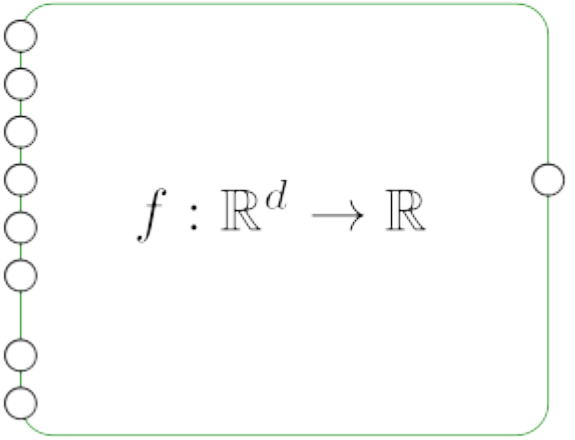
Understanding Deep Nets: Two Views

mechanistic understanding



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

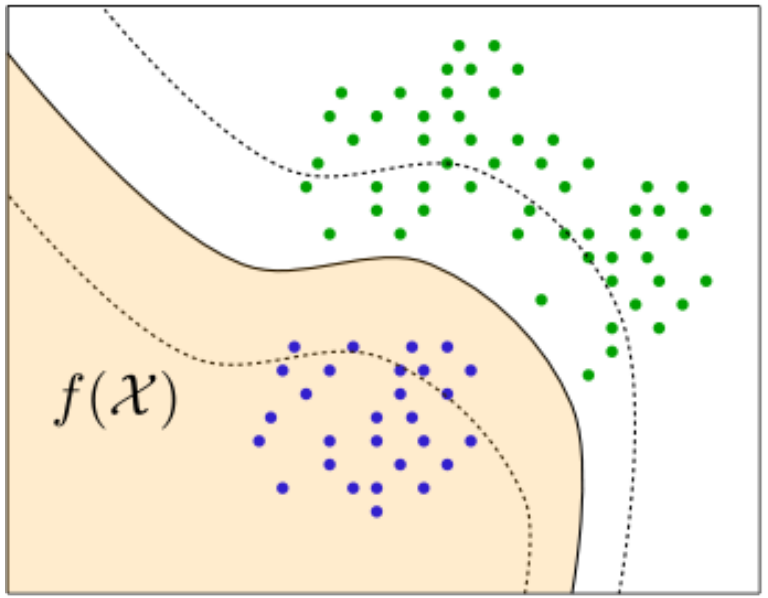
functional understanding



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

Understanding Deep Nets: Two Problems

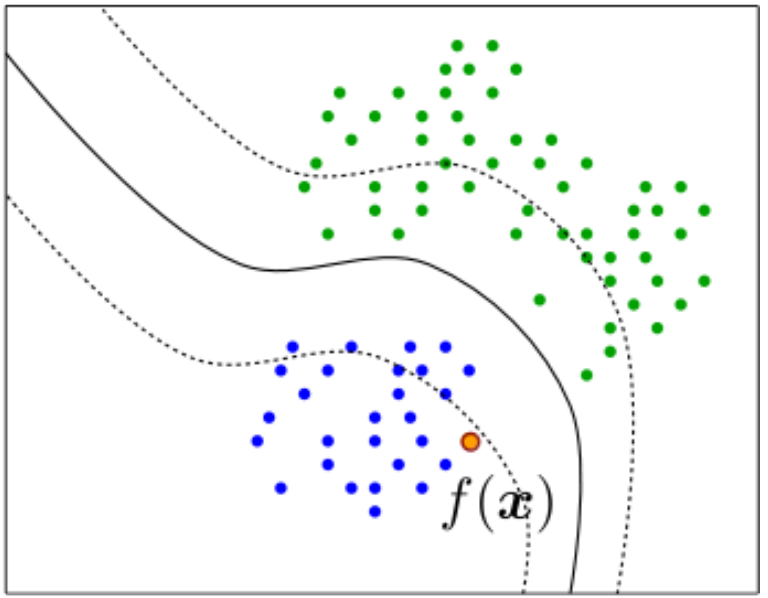
model analysis



possible approach

- build prototypes of "typical" examples of a certain class.

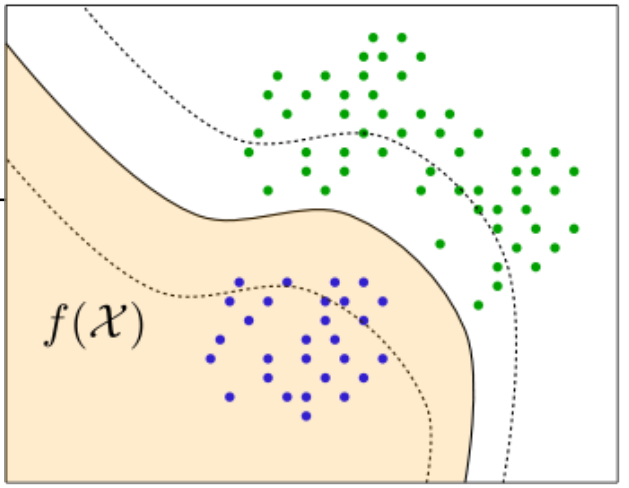
decision analysis



possible approach

- identify which input variables contribute to the prediction.

Model Analysis



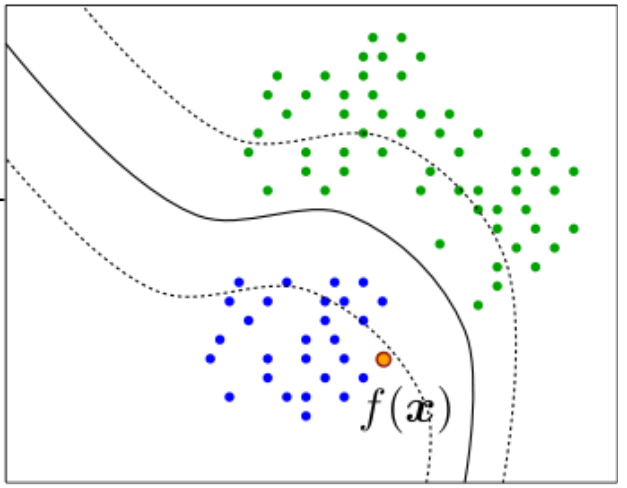
Example: Understanding how a deep net models the ImageNet class "Junco" by:

Approach 1: Looking which image in the dataset maximally activates that class.

Approach 2: Synthesizing an input image with maximum activation for that class.

Images from **Nguyen'16**

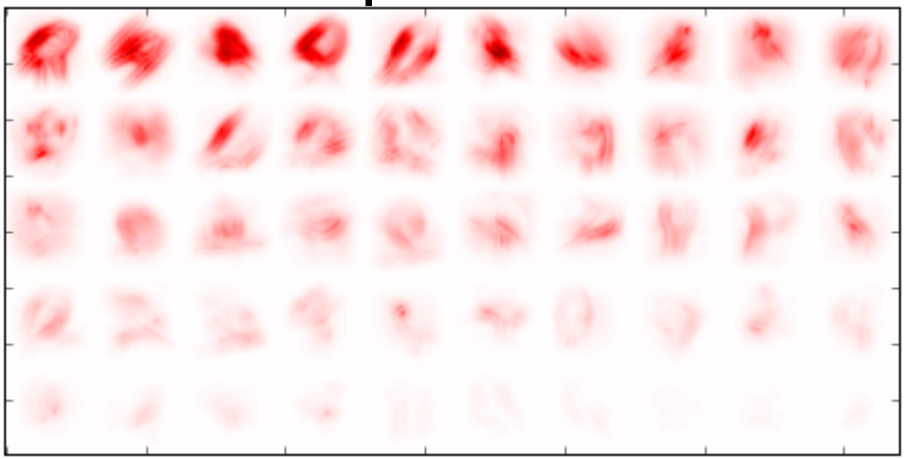
Decision Analysis



Example: Understanding what makes given images belong to their CIFAR-10 image class.

CIFAR-10 Images of birds

Pixel-wise explanation

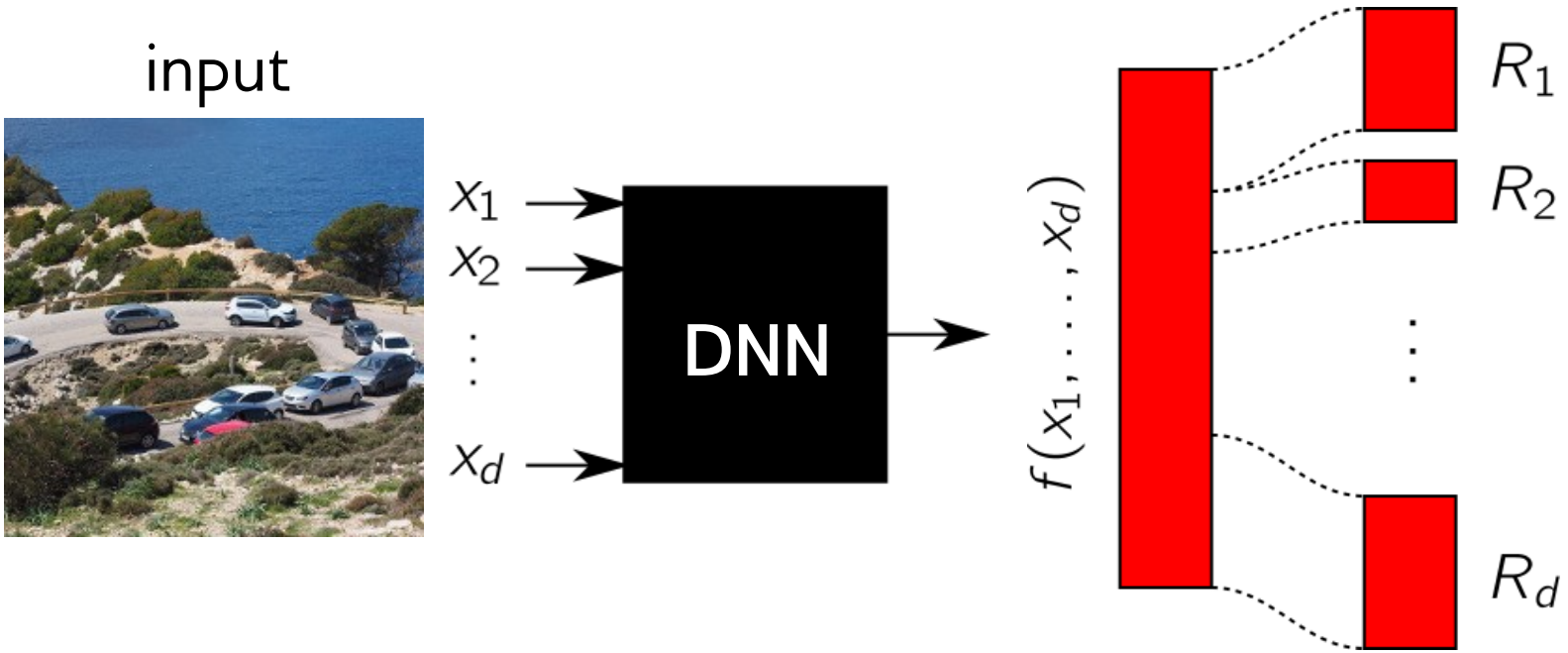


Observation: each image has a different explanation.

Question: how to find which pixels are relevant for a given image?

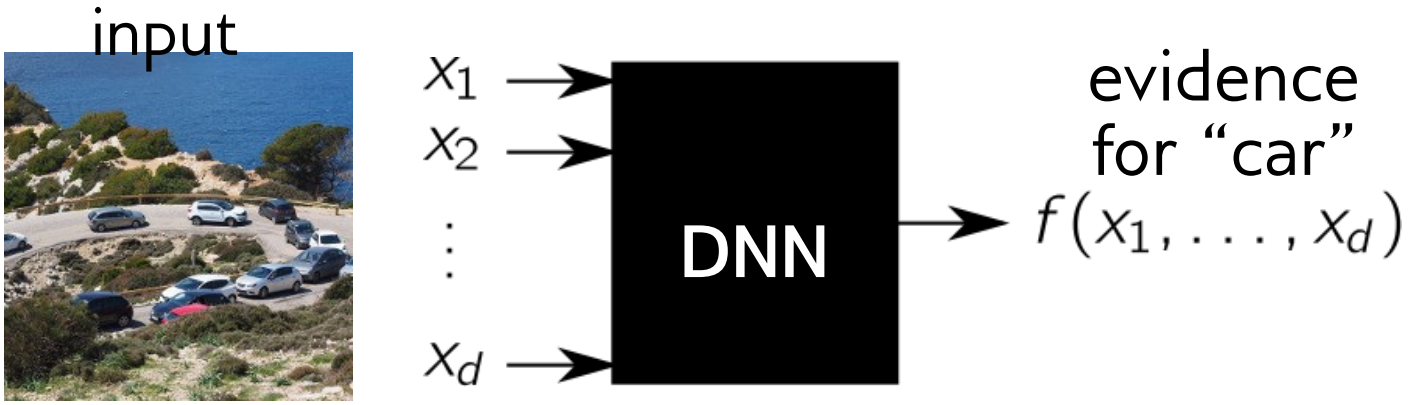
Explaining by Decomposing

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



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

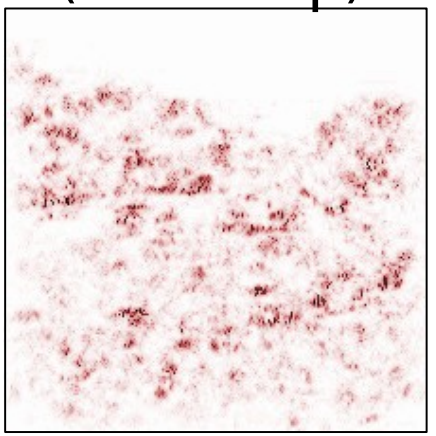
Sensitivity Analysis



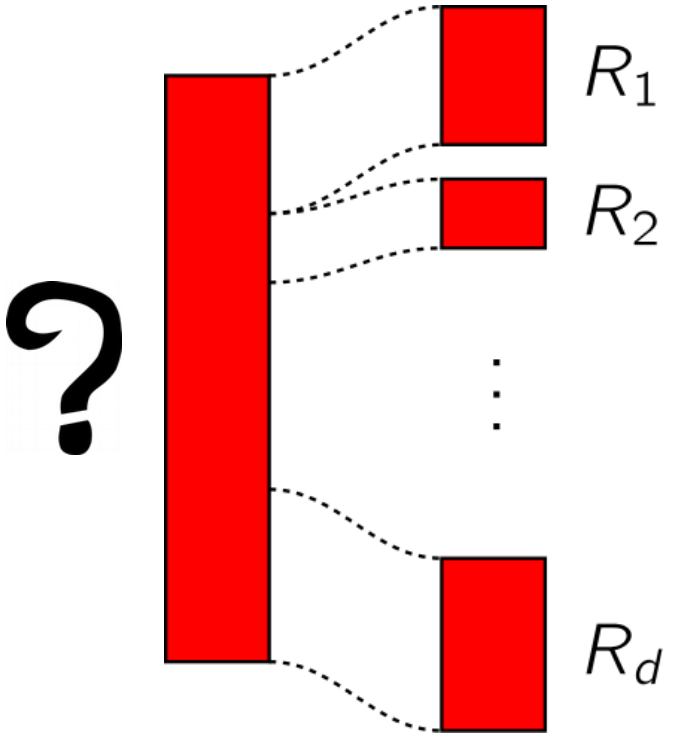
explanation for “car”
(heatmap):

We compute for each pixel:

$$R_i = \left(\frac{\partial f}{\partial x_i}\right)^2$$



Sensitivity Analysis and Decomposition



Question: If sensitivity analysis computes a decomposition of something: Then, *what* does it decompose?

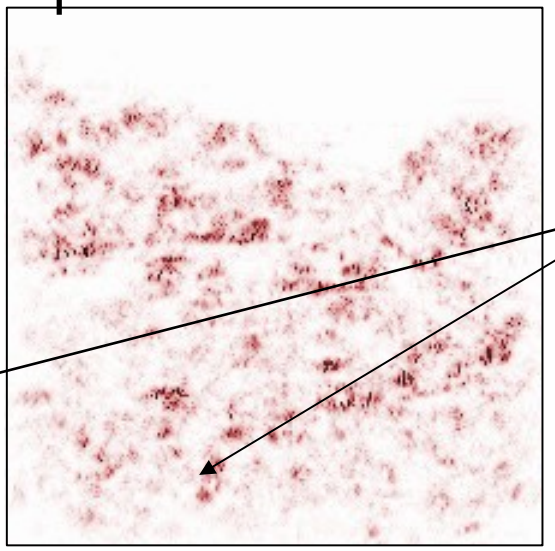
$$R_i = \left(\frac{\partial f}{\partial x_i}\right)^2 \quad \Rightarrow \quad \sum_{i=1}^{d} R_i = \|\nabla f(x_1, \ldots, x_d)\|^2$$

Sensitivity Analysis and Decomposition

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

input

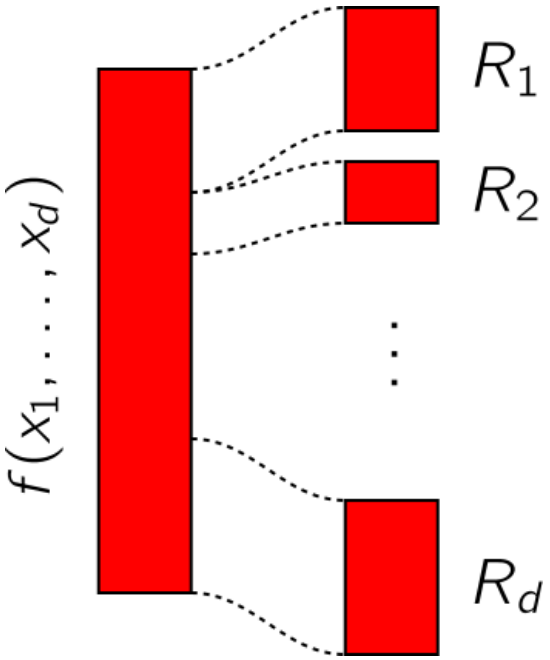
explanation for “car”



variation = *transforming trees into cars.*

$$\sum_{i=1}^{d} R_i = \|\nabla f(x_1, \ldots, x_d)\|^2$$

Simple Taylor Decomposition



$$f(\boldsymbol{x}) = f(\widetilde{\boldsymbol{x}}) + \langle \nabla f \big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}}, \boldsymbol{x} - \widetilde{\boldsymbol{x}} \rangle + \varepsilon$$

write it as a sum and identify first-order terms

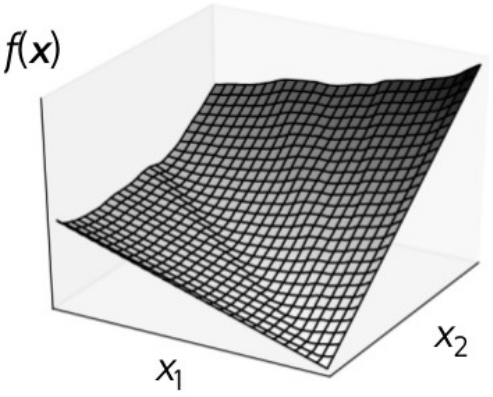
$$\sum_i \underbrace{\frac{\partial f}{\partial x_i}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} (x_i - \widetilde{x}_i)}_{R_i}$$

$\widetilde{x}$ is the root point (similar to the data point but without the evidence)

Simple Taylor Decomposition

Special case: unbiased Deep ReLU nets

$$f(tx) = tf(x) \quad \Rightarrow \quad f(0) = 0$$



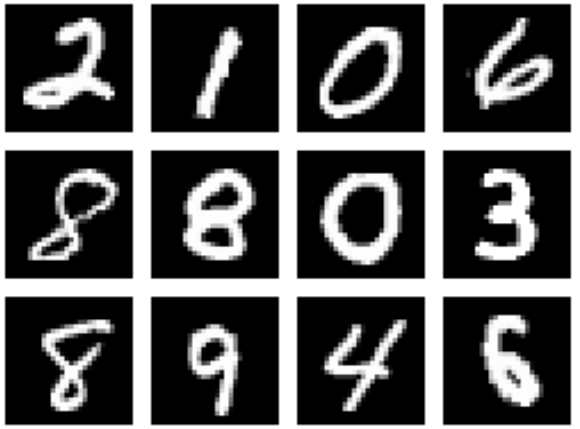
choose:

$$\widetilde{x} = \lim_{t \to 0} tx$$

then:

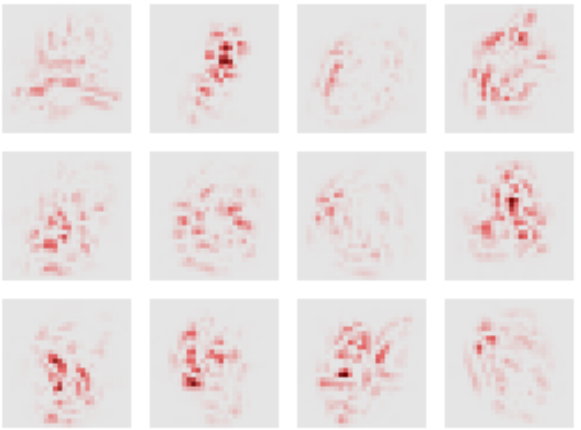
$$R_i = \frac{\partial f}{\partial x_i}\Big|_{x} \cdot x_i$$

Sensitivity vs. Simple Taylor Decomposition



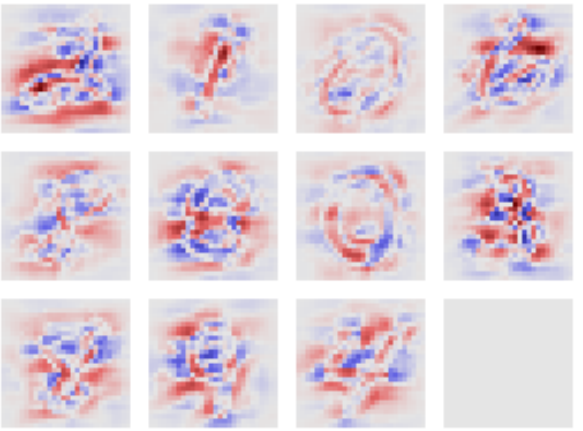
sensitivity analysis

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

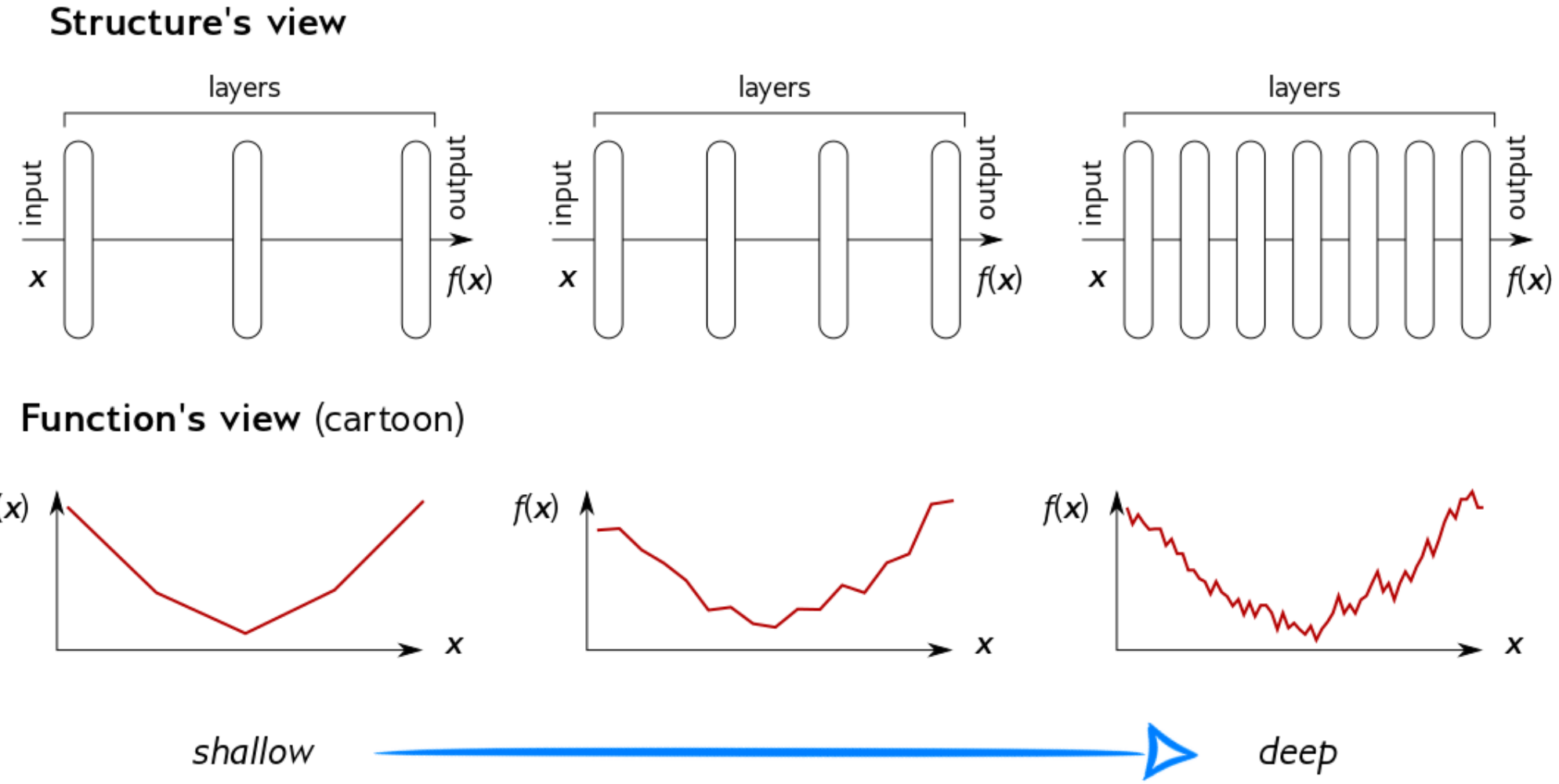


simple Taylor decomposition

$$R_i = \frac{\partial f}{\partial x_i}\Big|_x \cdot x_i$$

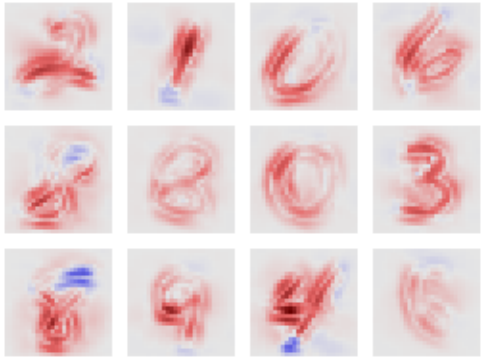


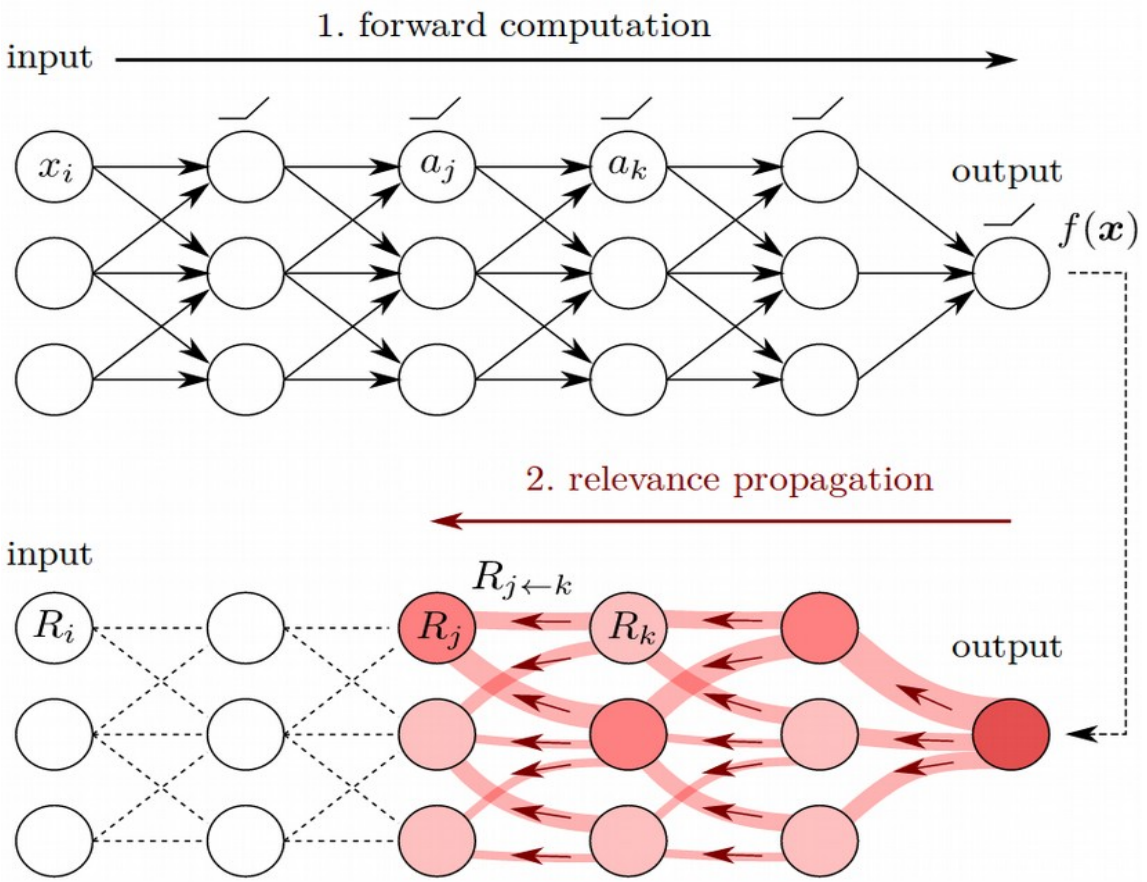
Gradient Shattering [Montufar'14, Balduzzi'17]



Explanation methods should not rely on the gradient.

Beyond Gradient-Based Techniques

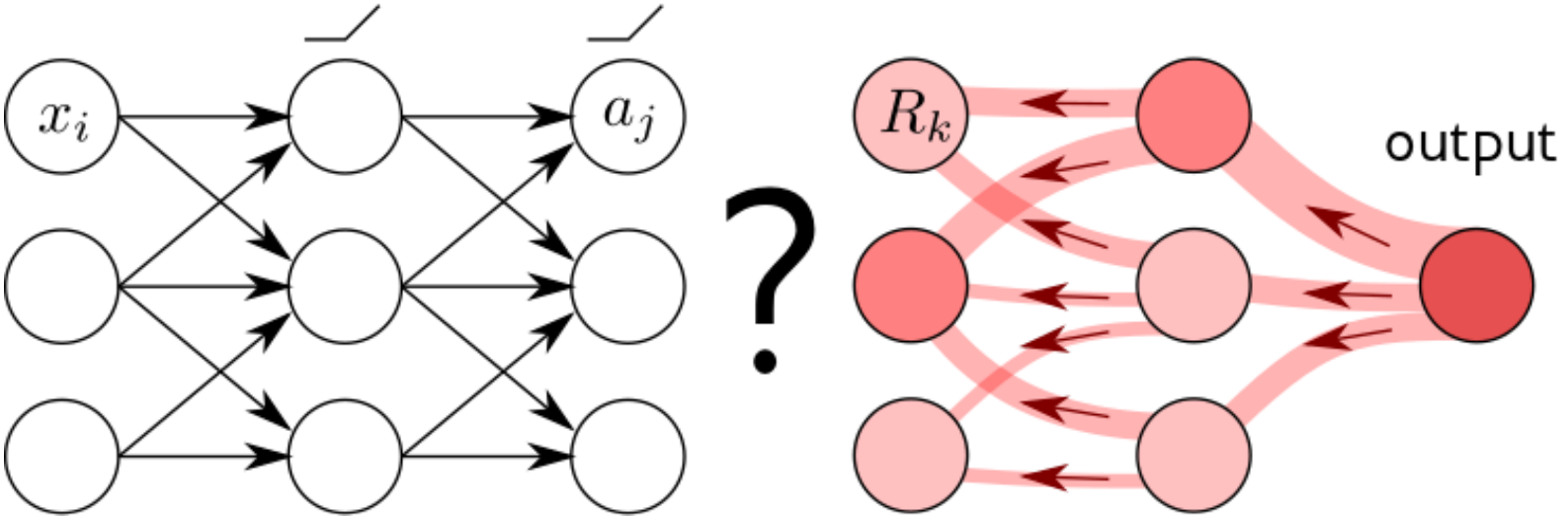




Bach'15 Layer-wise relevance propagation (LRP)

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

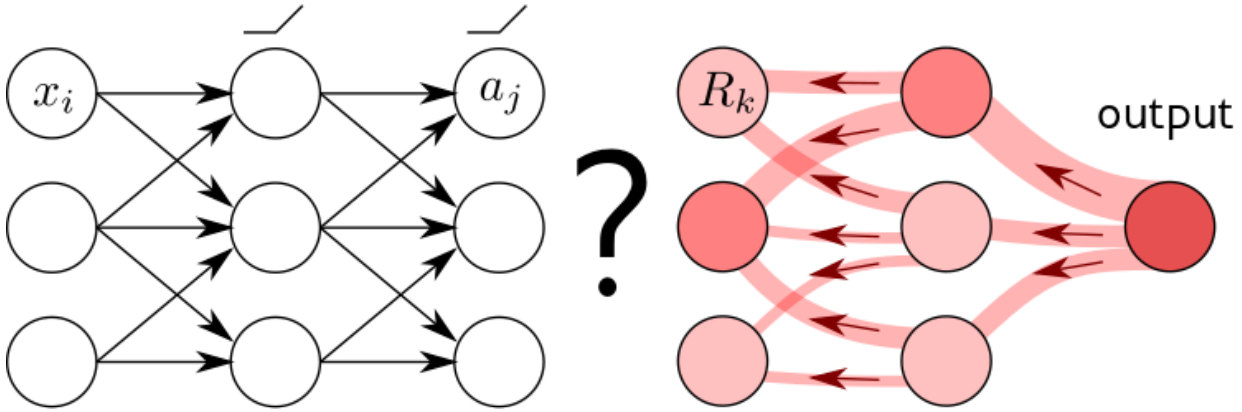
LRP: The Basic Question



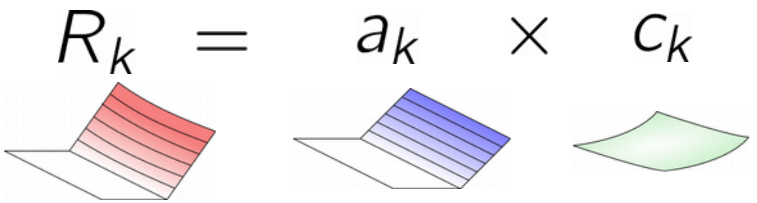
Question:
Suppose that we have propagated the relevance until a given layer.
How should the relevance be propagated to the previous layer?

Idea:
Perform a Taylor decomposition of the function $R_k((a_j)_j)$

Relevance as a Function

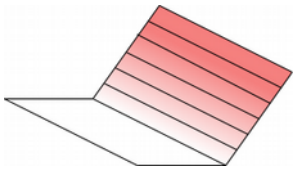


1 Assume relevance is (approximately) a multiple of the neuron activation and a constant.



2 We build a relevance function

$$R_k((a_j)_j) = \max(0, \sum_j a_j w_{jk} + b_k) \cdot \text{ const.}$$
$$= \max(0, \sum_j a_j w'_{jk} + b'_k)$$

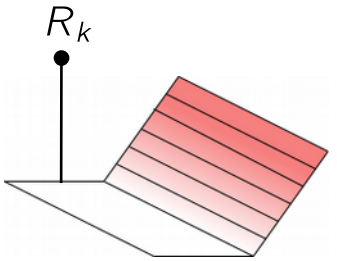


and we use Taylor decomposition to decompose relevance on the layer before.

Decomposing on the Lower-Layer

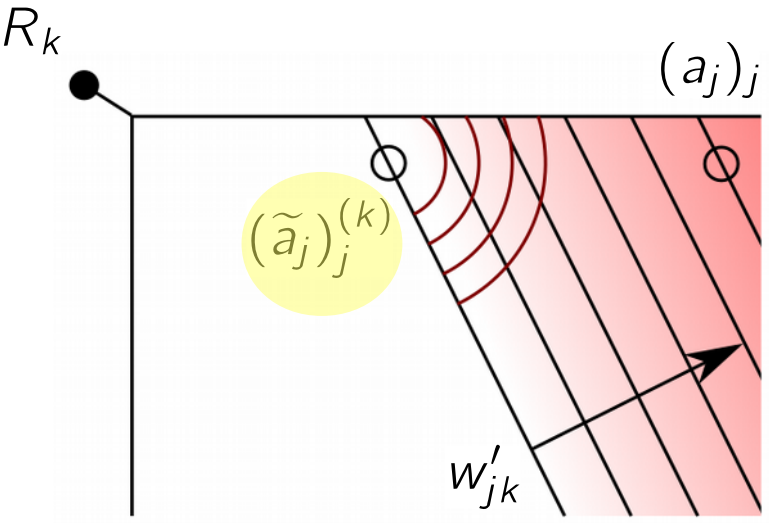
Relevance function:

$$R_k((a_j)_j) = \max(0, \sum_j a_j w'_{jk} + b'_k)$$



Taylor expansion:

$$R_k((a_j)_j) = \sum_j \underbrace{\frac{\partial R_k}{\partial a_j}\Big|_{(\widetilde{a}_j)_j^{(k)}} \cdot (a_j - \widetilde{a}_j^{(k)})}_{R_{j \leftarrow k}}$$



Propagation rule:

$$R_{j \leftarrow k} = \frac{(a_j - \widetilde{a}_j^{(k)}) w_{jk}}{\sum_j (a_j - \widetilde{a}_j^{(k)}) w_{jk}} R_k$$

Building Rules for Specific Domains

Generic propagation rule derived from Taylor decomposition:

$$R_{j \leftarrow k} = \frac{(a_j - \widetilde{a}_j^{(k)}) w_{jk}}{\sum_j (a_j - \widetilde{a}_j^{(k)}) w_{jk}} R_k$$

Practical propagation rules with root points selected to belong to input domain:

Input domain	Root point	Rule
ReLU activations ($a_j \geq 0$)	$(a_j - \widetilde{a}_j)^{(k)} \propto a_j \cdot 1_{w_{jk}>0}$	$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$
Pixel intensities ($x_i \in [l_i, h_i]$, $l_i \leq 0 \leq h_i$)	$(x_i - \widetilde{x}_i)^{(j)} \propto x_i - l_i \cdot 1_{w_{jk}>0} - h_i \cdot 1_{w_{jk}<0}$	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$
Real values ($x_i \in \mathbb{R}$)	$(x_i - \widetilde{x}_i)^{(j)} \propto w_{ij}$	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$

LRP and Taylor Decomposition

Bach'15 Layer-wise relevance propagation (LRP)

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

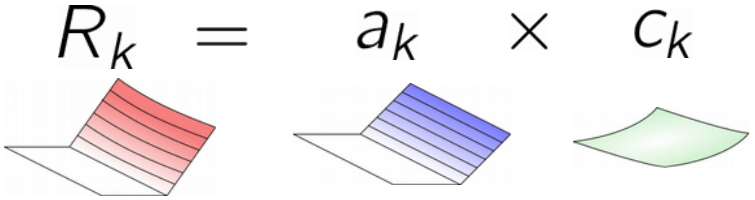
equivalent for α=1 and β=0

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 ($x_i \in \mathbb{R}$)	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$

specialized rules for input domain

Verifying the Product Structure

1. Assume relevance is (approximately) a multiple of the neuron activation and a constant.



Question: Was it true?

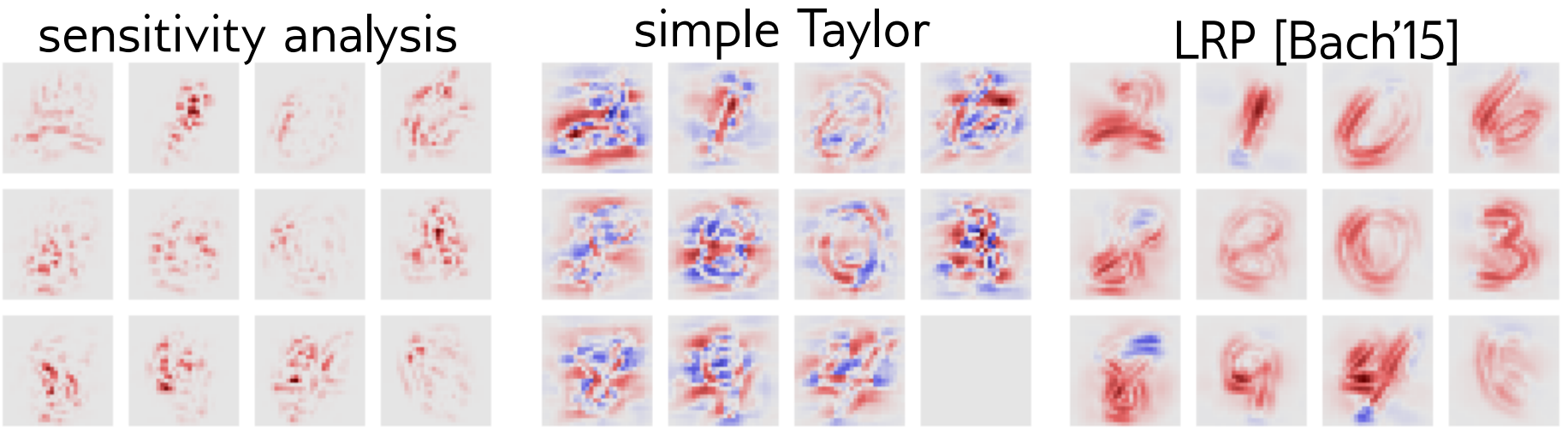
Inductive reasoning:
Assume that it holds for R_k, then show that it also holds for R_j

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

$$R_j \approx a_j \cdot \sum_k w_{jk}^+ \frac{\left(\sum_j a_j w_{jk} + b_j\right)^+}{\sum_j a_j w_{jk}^+} \cdot \text{ const}$$

$$R_j \approx a_j \cdot \text{ const}$$

Comparing Explanation Methods

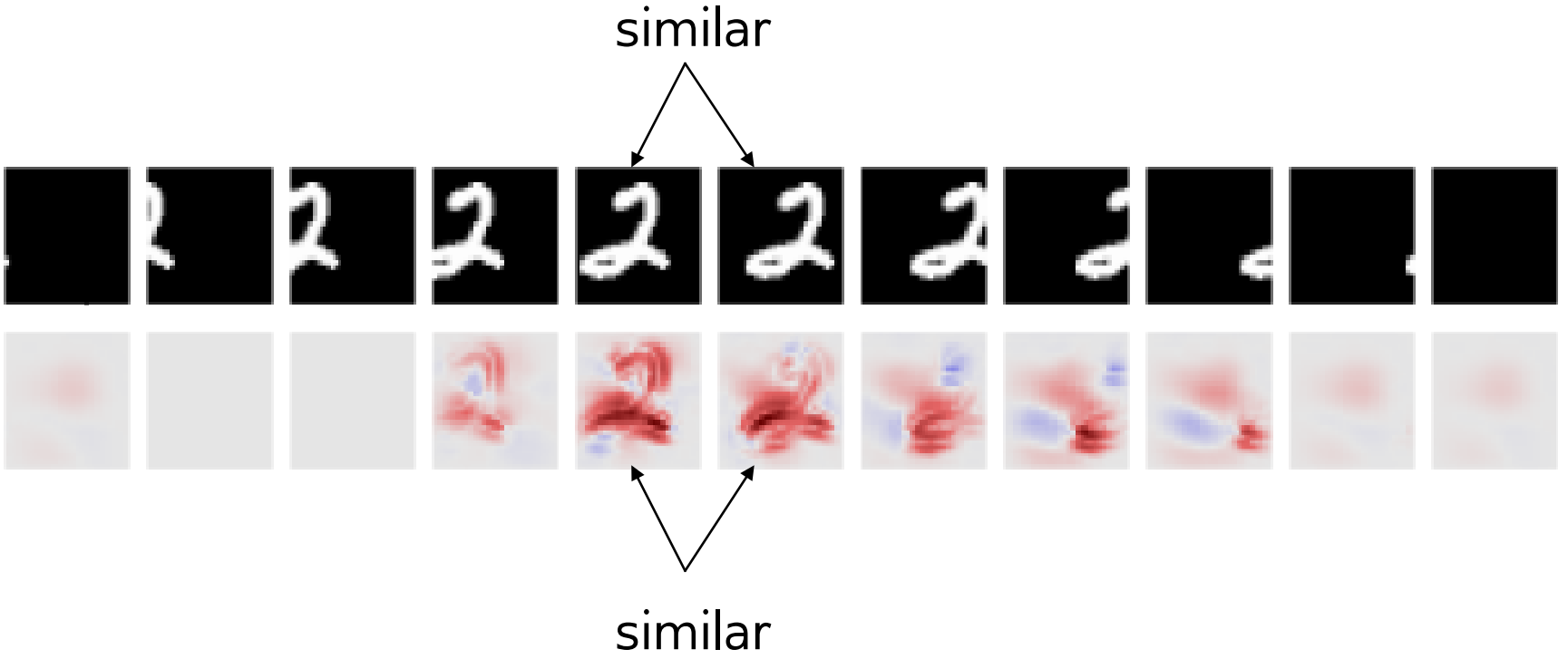


Visual intuition suggests that LRP produces better explanations.

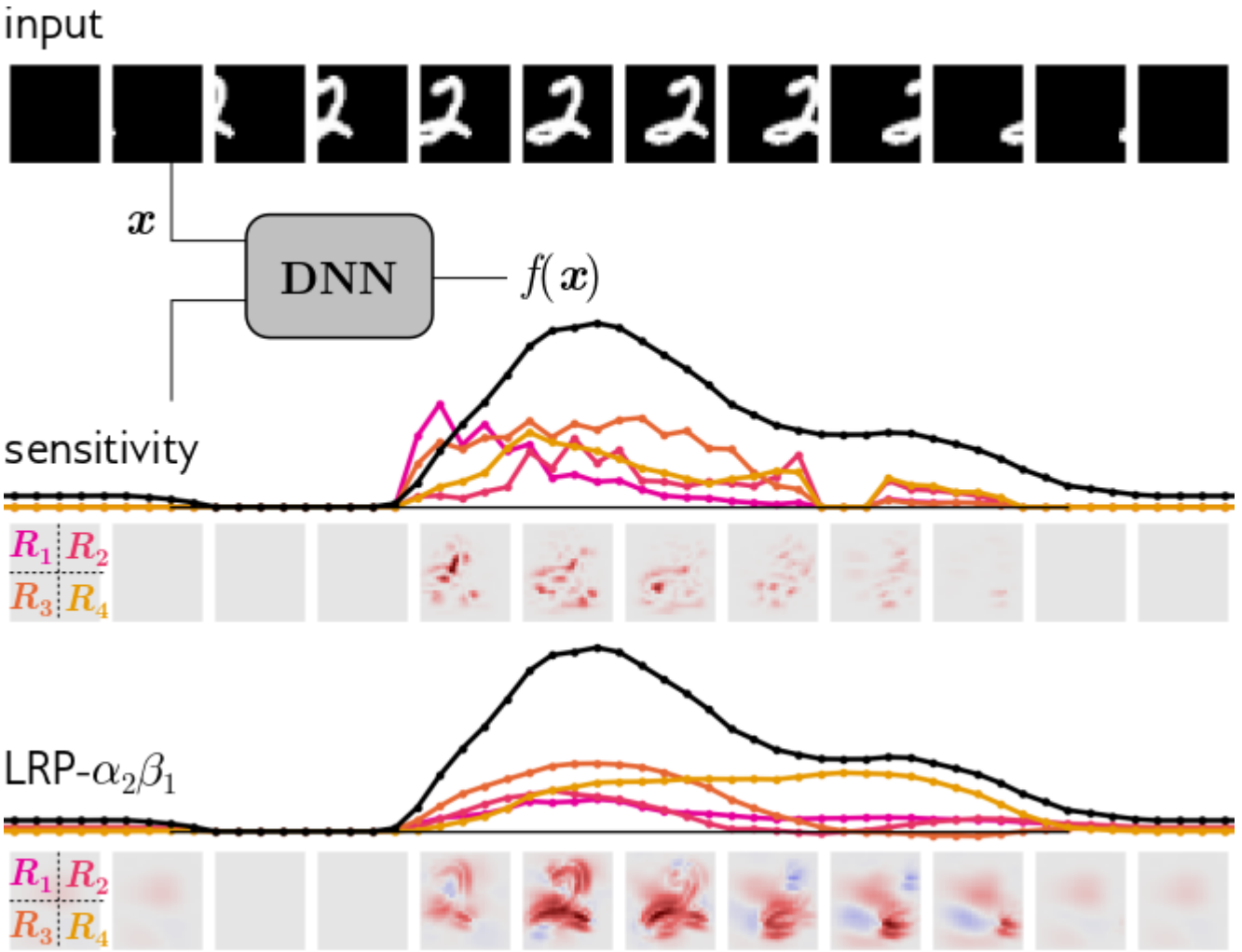
But can we *measure* heatmap quality?

Explanation Continuity

Idea: If two examples are almost the same (and their prediction too), then, the explanations should also be almost the same.



Explanation Continuity



Follow a path on the data manifold.

LRP explanations are more continuous

LRP-α1β0 and Continuity

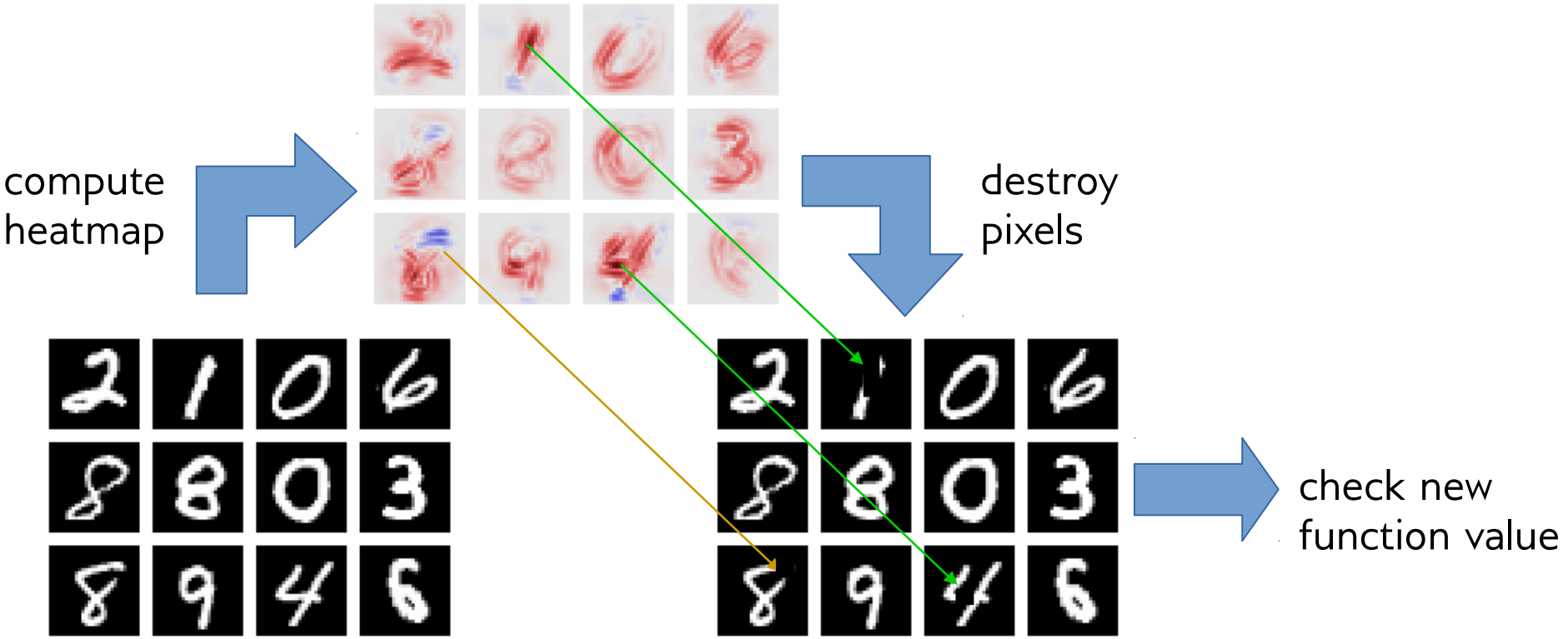
LRP view 1 (redistribution): $$R_i = \sum_j \frac{a_i w_{ij}^+}{\sum_i a_i w_{ij}^+} R_j$$

LRP view 2 (constant terms): $$c_i = \sum_j w_{ij}^+ \frac{(\sum_i a_i w_{ij} + b_j)^+}{\sum_i a_i w_{ij}^+} c_j$$

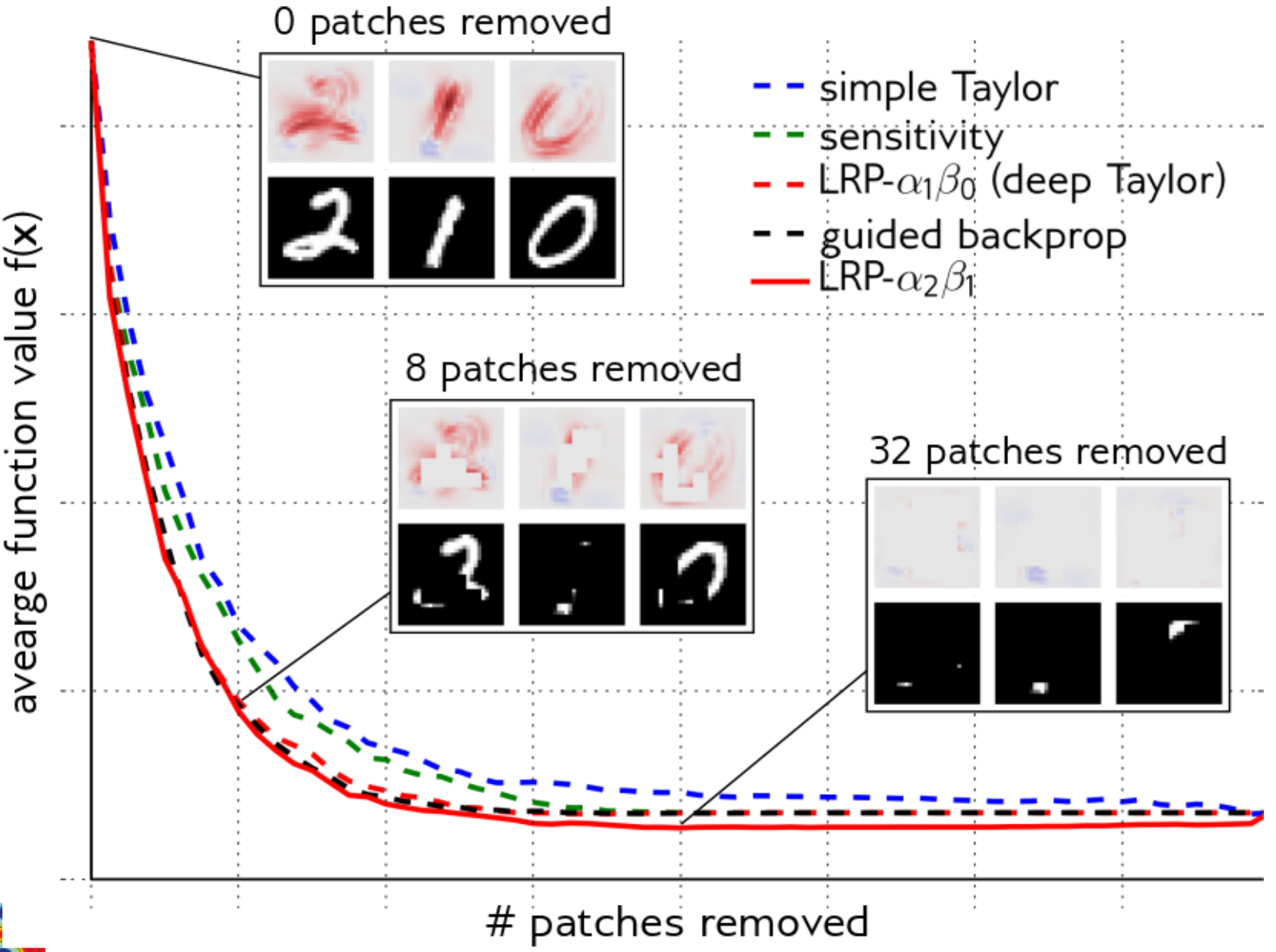
gradient propagation: $$\delta_i = \sum_j w_{ij} 1_{\sum_i a_i w_{ij} + b_j > 0} \delta_j$$

Explanation Selectivity

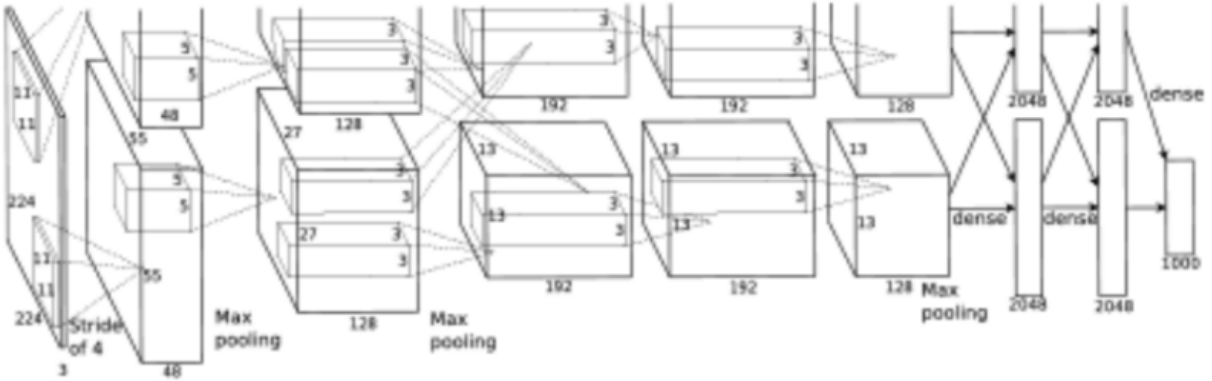
Idea: Removing input variables with high relevance should make the function value drop.



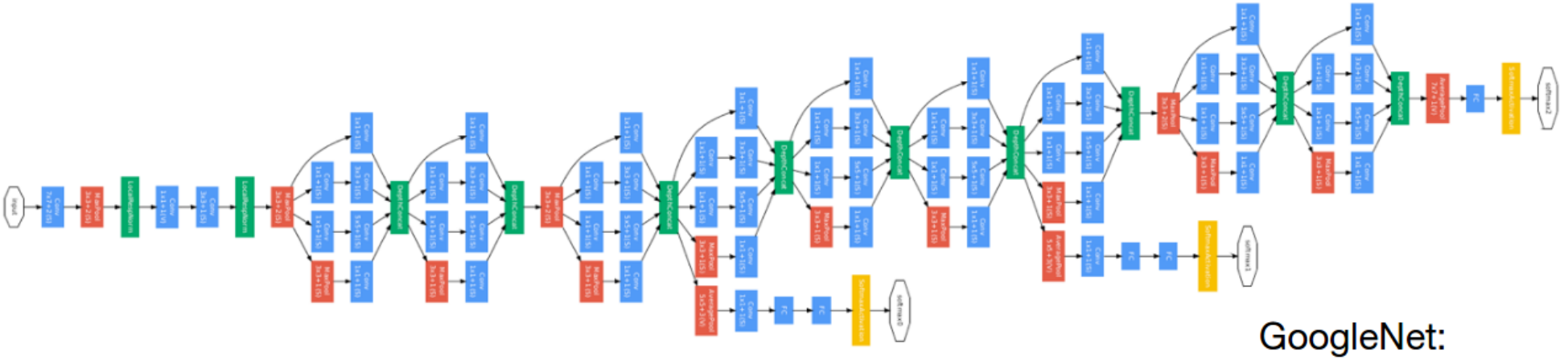
Explanation Selectivity



LRP for Comparing Classifiers [Binder'16]



BVLC:
- 8 Layers
- ILSRCV: 16.4%



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

LRP for Comparing Classifiers [Binder'16]

Googlenet focuses on the face of animals (suppresses background noise)

LRP for Comparing Classifiers [Arras'17]

Table 1. Test set performance of the ML models for 20-class document classification.

ML Model	Test Accuracy (%)
BoW/SVM ($V = 70631$ words)	80.10
CNN1 ($H = 1$, $F = 600$)	79.79
CNN2 ($H = 2$, $F = 800$)	**80.19**
CNN3 ($H = 3$, $F = 600$)	79.75

SVM/BoW classifier

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

CNN/word2vec classifier

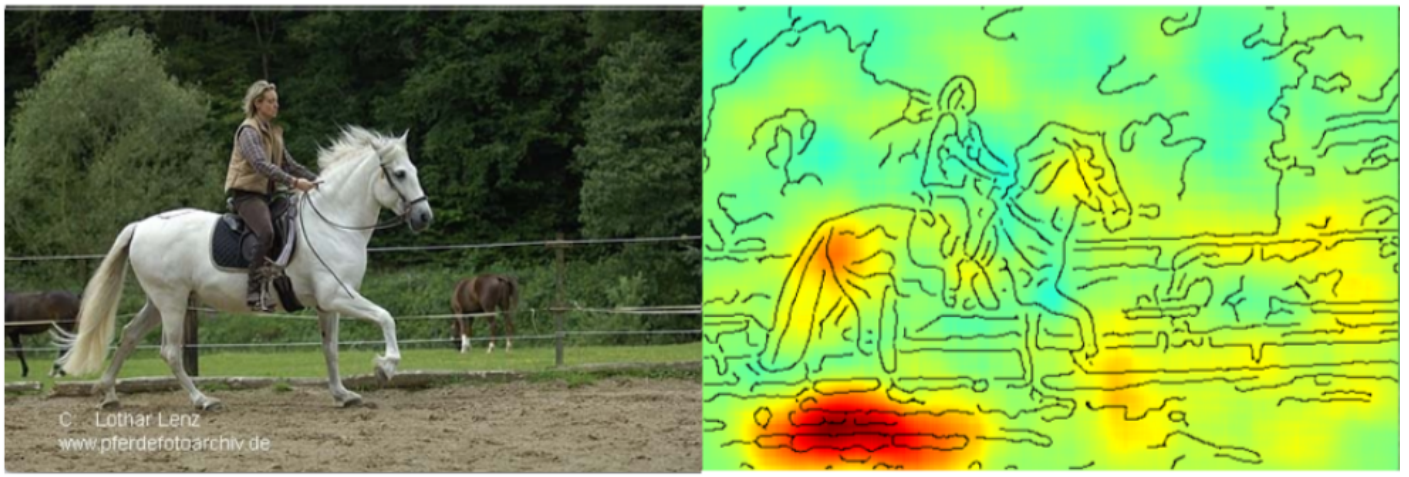
on a roller coaster ride than others. The mental part is usually induced by a lack of clear indication of which way is up or down, ie: the Shuttle is normally oriented with its cargo bay pointed towards Earth, so the Earth (or ground) is "above" the head of the astronauts. About 50% of the astronauts experience some form of motion sickness, and NASA has done numerous tests in

LRP for Comparing Classifiers [Lapushkin'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%

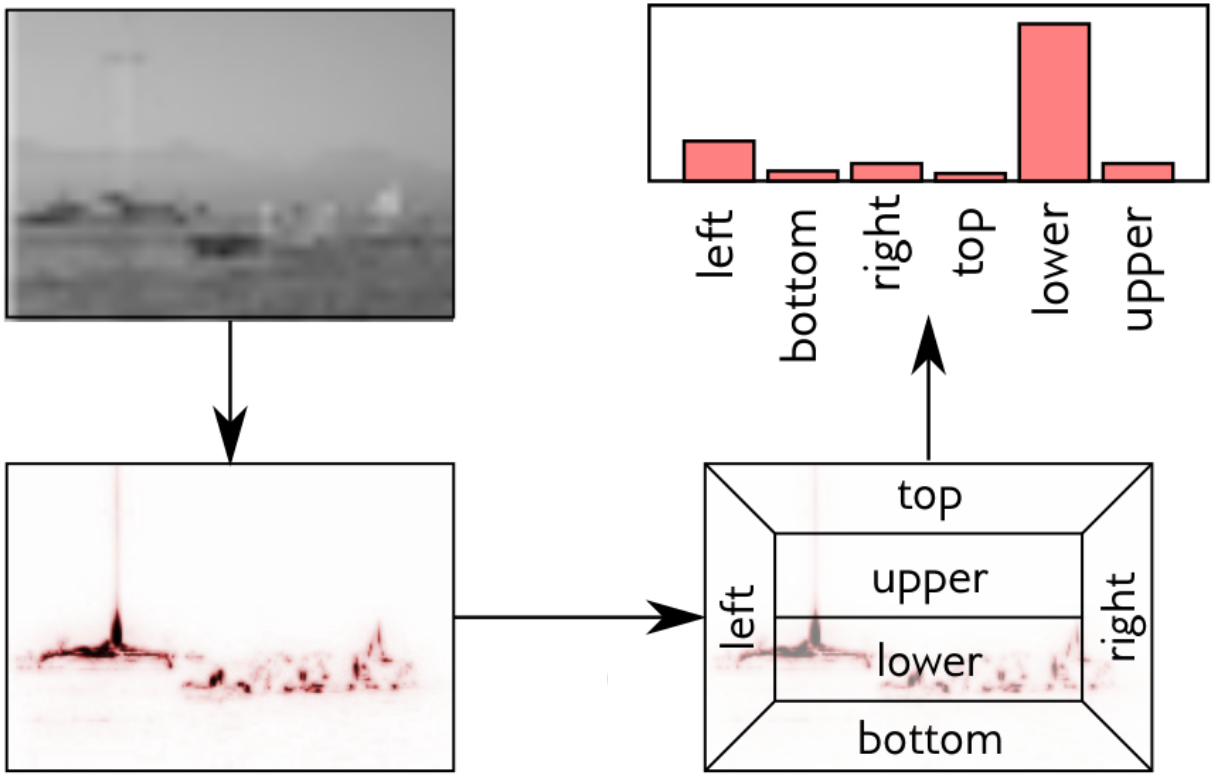
Example of a horse image next to the LRP analysis of the Fisher's decision:



LRP for Comparing Classifiers [Lapushkin’16]

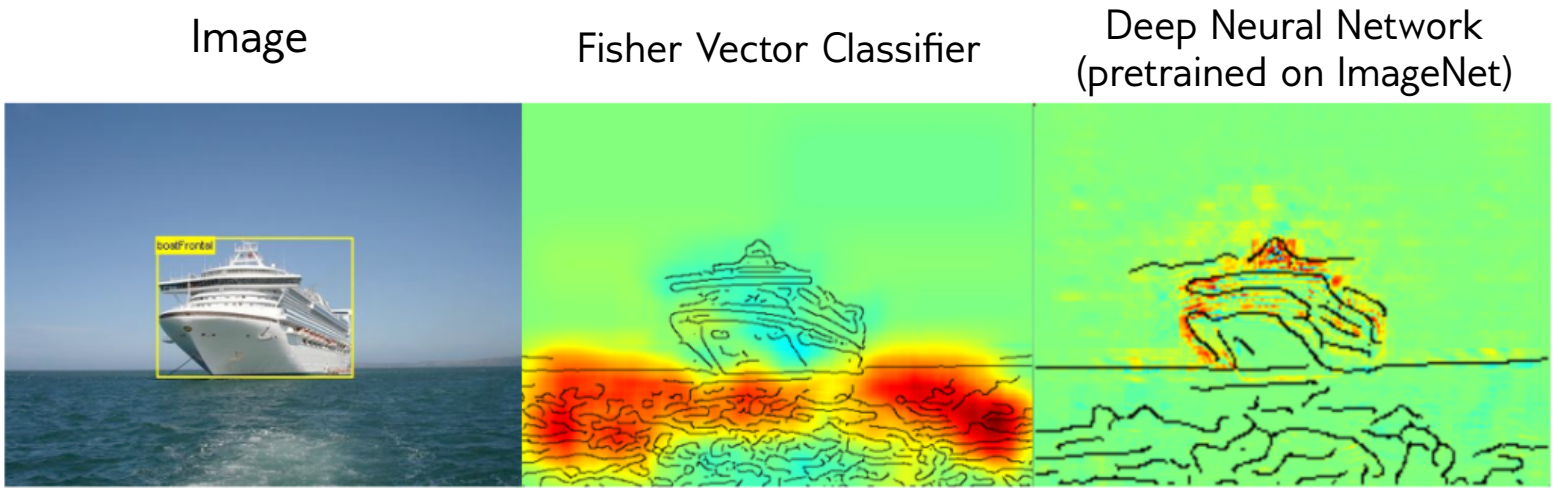
LRP and Region-Based Analysis

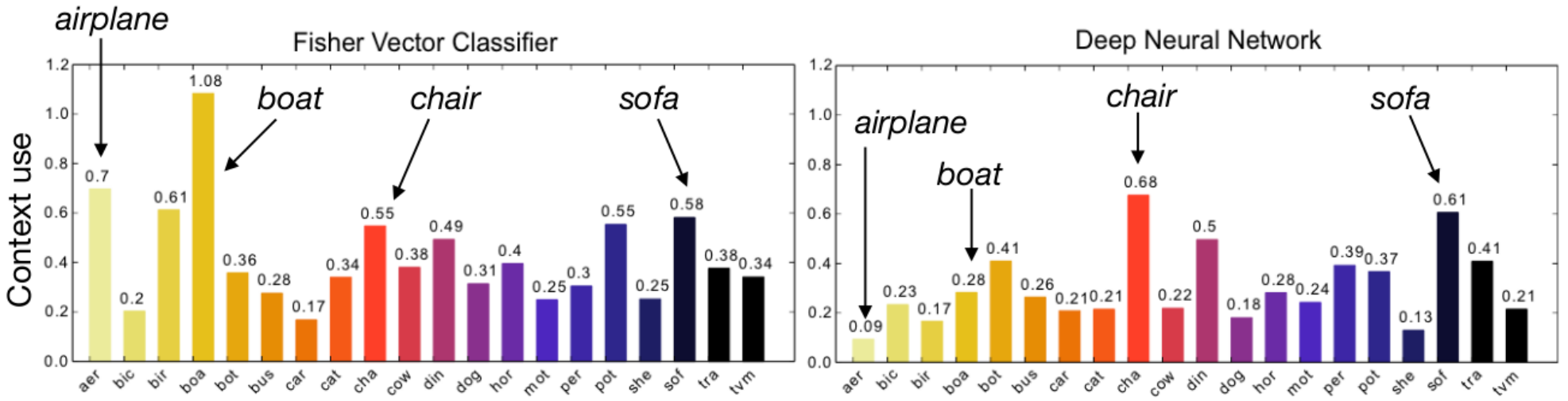
$$f(\boldsymbol{x}) = \sum_{\mathcal{I}} R_{\mathcal{I}}(\boldsymbol{x})$$



$$f(\boldsymbol{x}) = \sum_{i} R_i(\boldsymbol{x}) \qquad R_{\mathcal{I}}(\boldsymbol{x}) = \sum_{i \in \mathcal{I}} R_i(\boldsymbol{x})$$

Comparing Reliance of Classifiers on Context





[Lapuschkin'16]

Interpretable ML for the Complex World

Can we use interpretable ML to make sense of complex systems for which we have little understanding or intuition?

Winning strategies in board games

Deep Blue, Alpha Go

Quantum chemistry

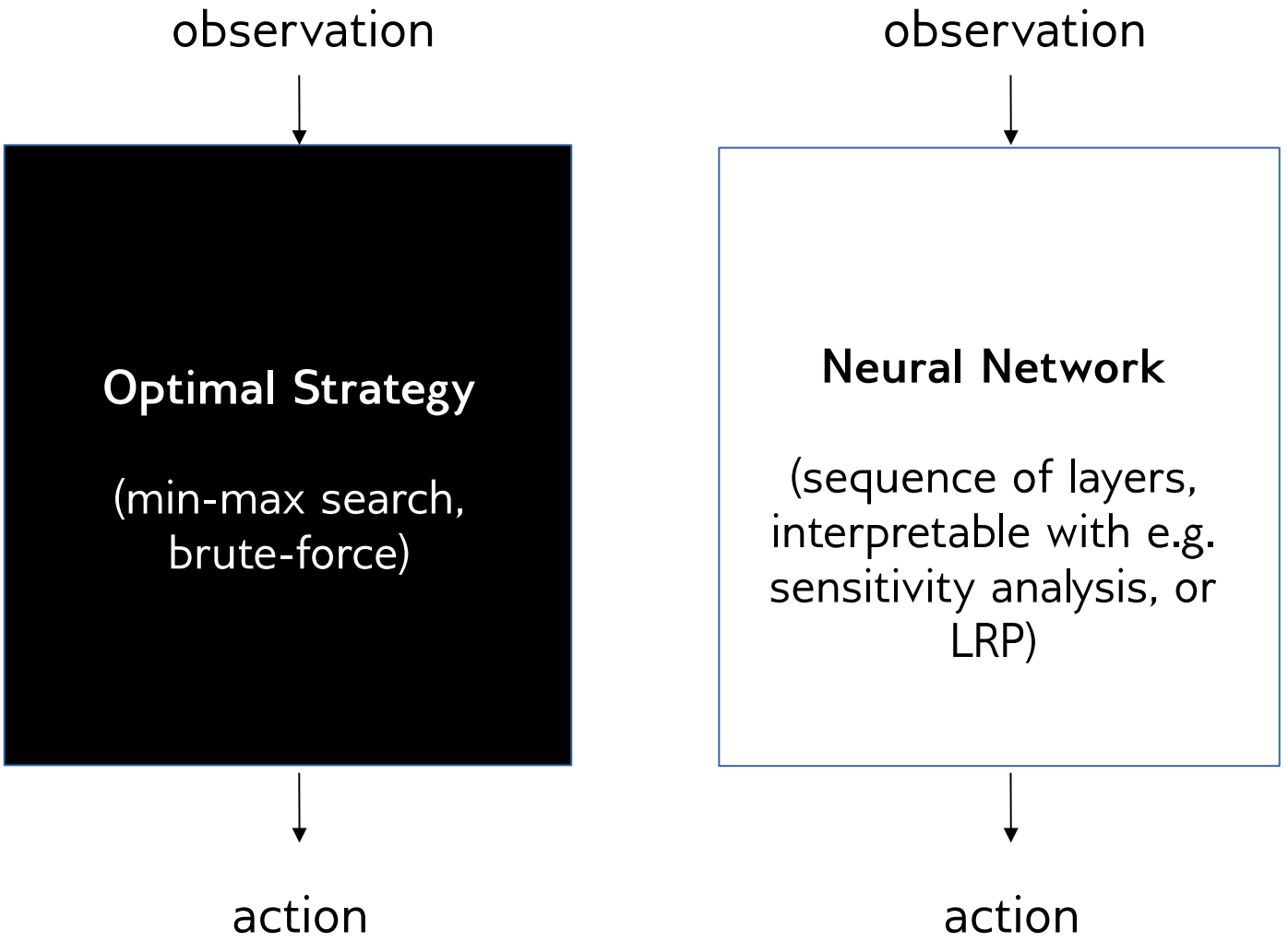
$$i\hbar\frac{\partial}{\partial t}\Psi = H\Psi$$

DFT-based solvers, numerical simulations

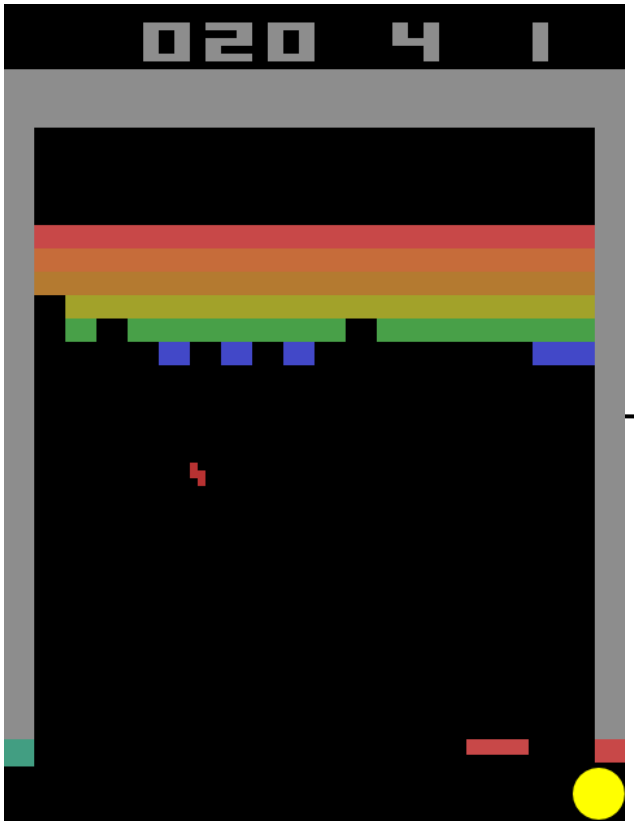
Neurosciences, biology

Data, experiments

Approximating Optimal Strategies with ML

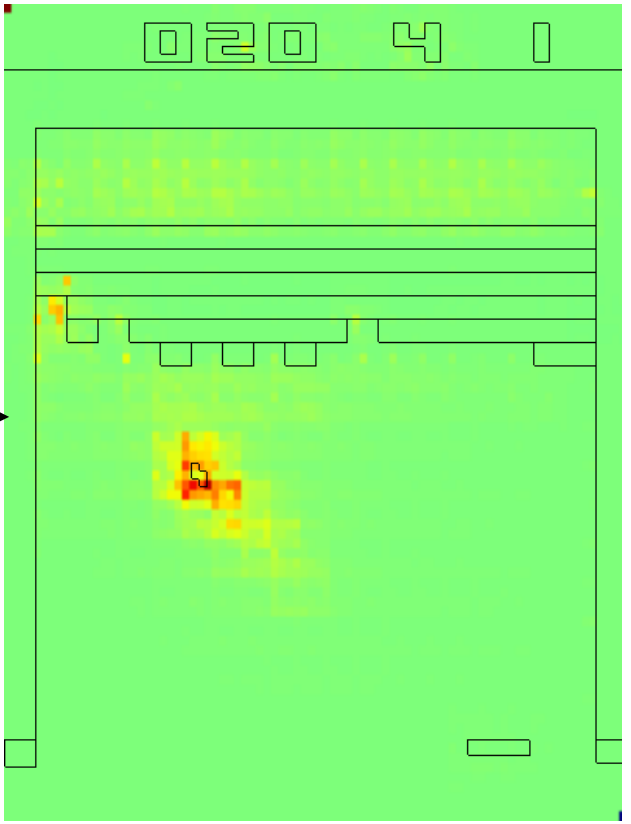


Interpreting a DNN Playing “Breakout”



Input image

action
(left/right)



LRP heatmap

Insight: DNN player keeps track of the ball to position its cursor.

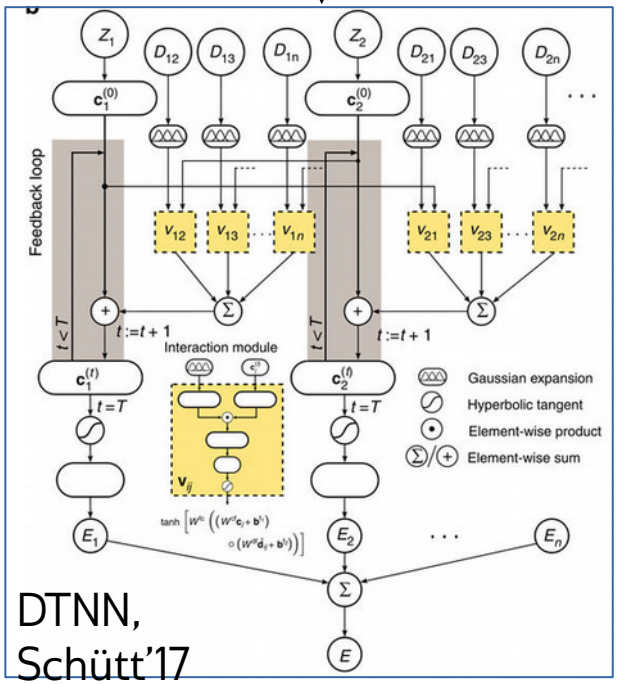
Approximating the Schrödinger Equation

molecular structure (e.g. atoms positions)

DFT calculation of the stationary Schrödinger Equation

$$H\Phi = E\Phi$$

PBE0, Pedrew'86



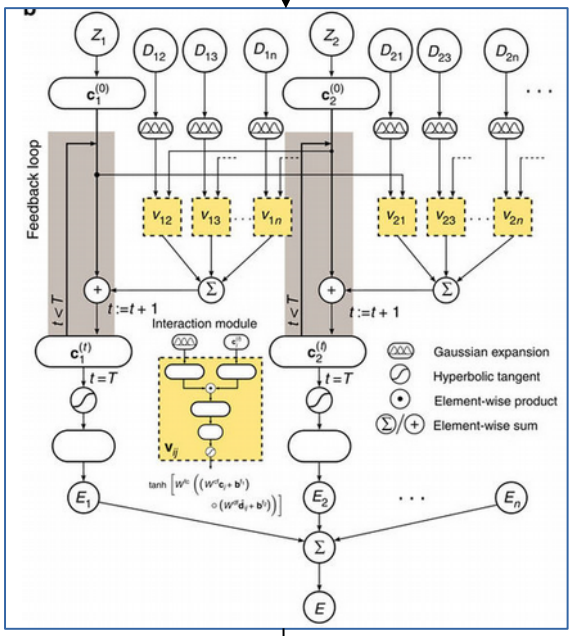
DTNN, Schütt'17

molecular electronic properties (e.g. atomization energy)

Interpreting the Schrödinger Equation

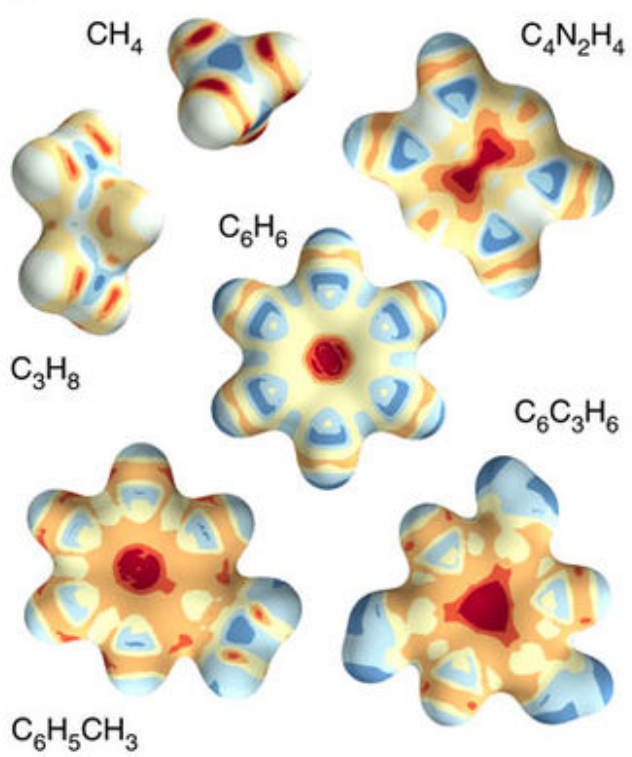
Schütt'17 Quantum-Chemical Insights from Deep Tensor Neural Networks

molecular structure



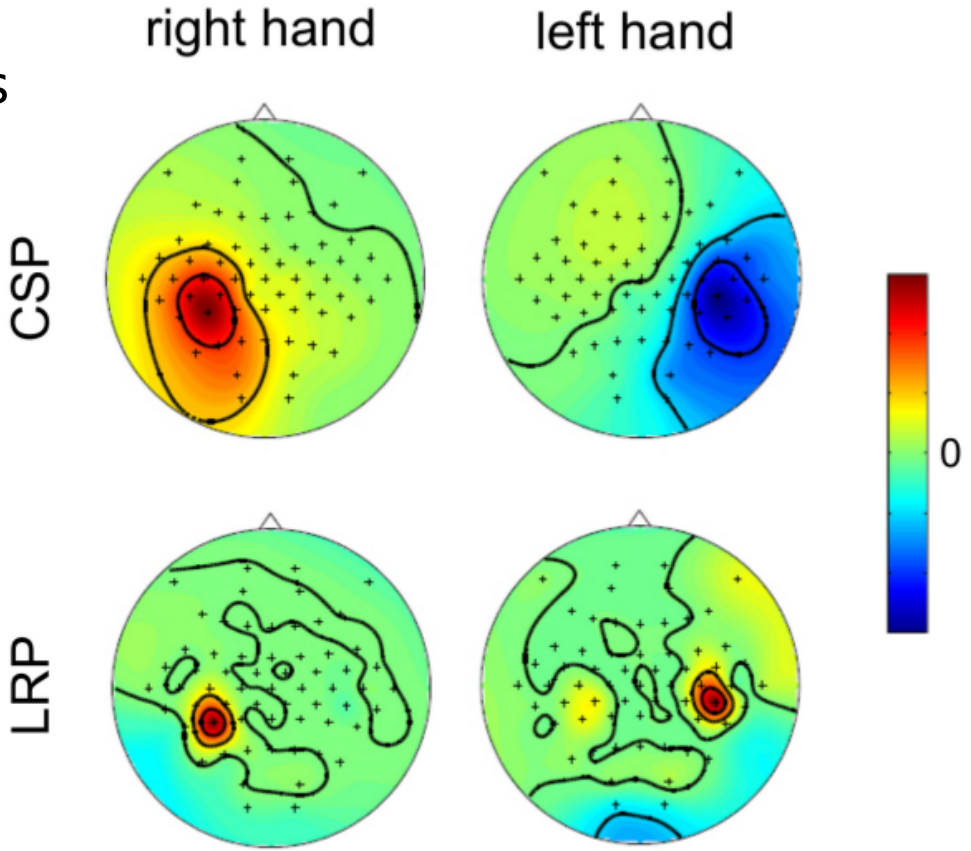
atomization energy

Example: Perturbation analysis of the neural network for various molecules.

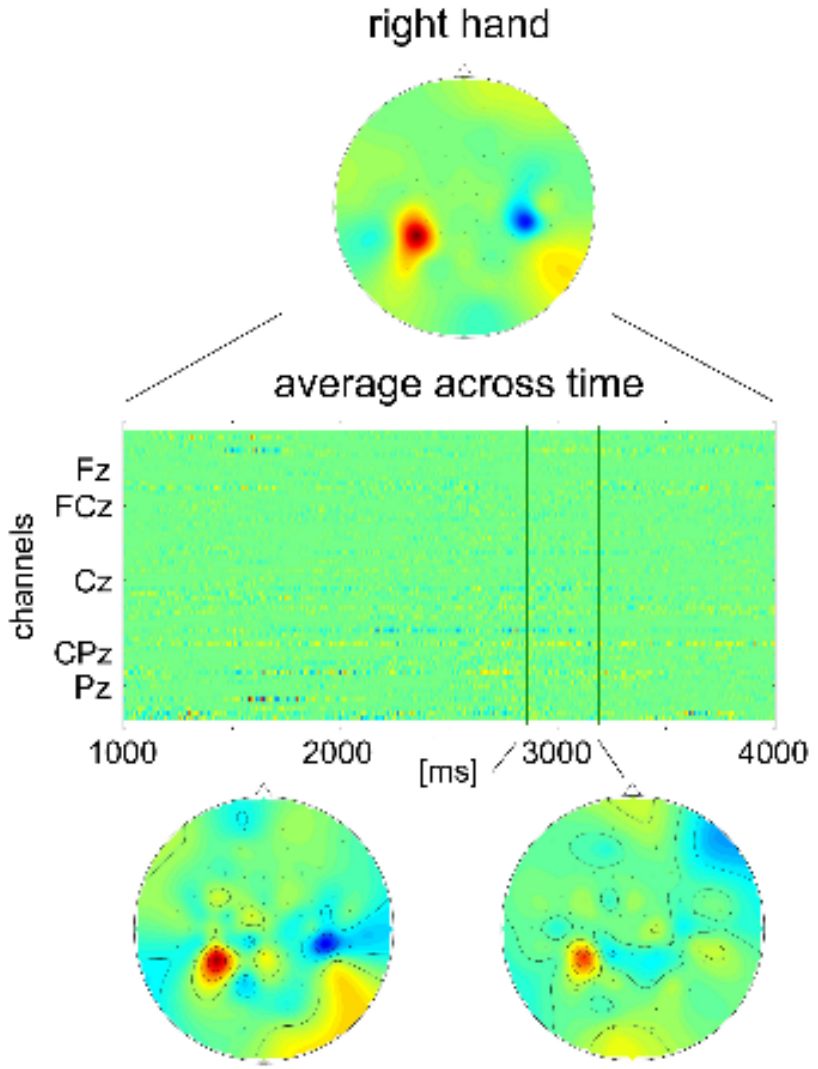


Interpreting Human “Left/Right” Decisions

Sturm’16 Understanding in terms of EEG data why a person thinks “left hand” or “right hand”



EEG Data and Pooling Explanations



$$f(\boldsymbol{x}) = \sum_{\mathcal{I}} R_{\mathcal{I}}(\boldsymbol{x})$$

Temporal pooling

$$f(\boldsymbol{x}) = \sum_{i} R_i(\boldsymbol{x})$$

Relevance is first redistributed on channels and time. It can then be pooled either in time, or in space.

Image from **Sturm'16**

Summary

- Big data can sometimes have adverse effect on learned models (e.g. base its decision on spurious features).
- The standard approach to remediate this problem is to concentrate on a small dataset of “clean” data points, and make efficient use of this finite amount of data.
- Another approach is to explain the predictions of the classifier trained on big data to spot potential biases/weaknesses.
- Graph-propagation approaches are more robust than methods relying directly on the function and its gradient.
- Explanation methods can also be used to learn something new about the problem (e.g. in the sciences).