

Machine Learning and AI for the sciences – Towards Understanding



Berliner Zentrum für
MASCHINELLES LERNEN



BERLIN BIG
DATA CENTER

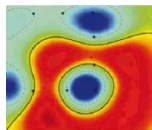


mpii max planck institut
informatik

Klaus-Robert Müller !!et al.!!

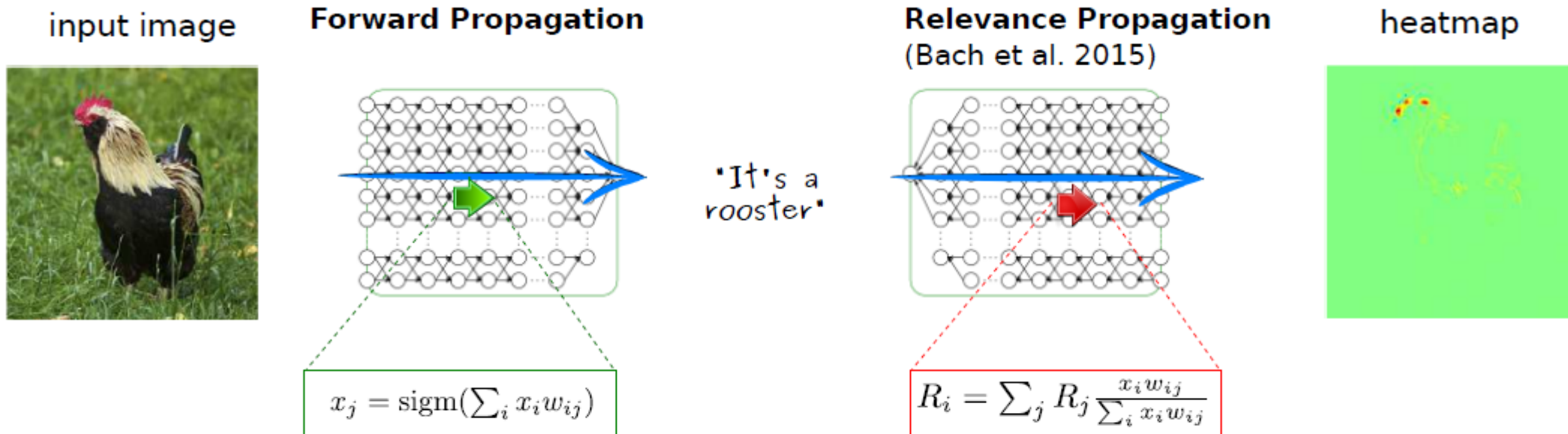
Outline

- understanding single decisions of nonlinear learners
- Layer-wise Relevance Propagation (LRP)
- Applications in Neuroscience, Medicine and Physics



Towards Explaining:
Machine Learning = black box?

Explaining single Predictions Pixel-wise



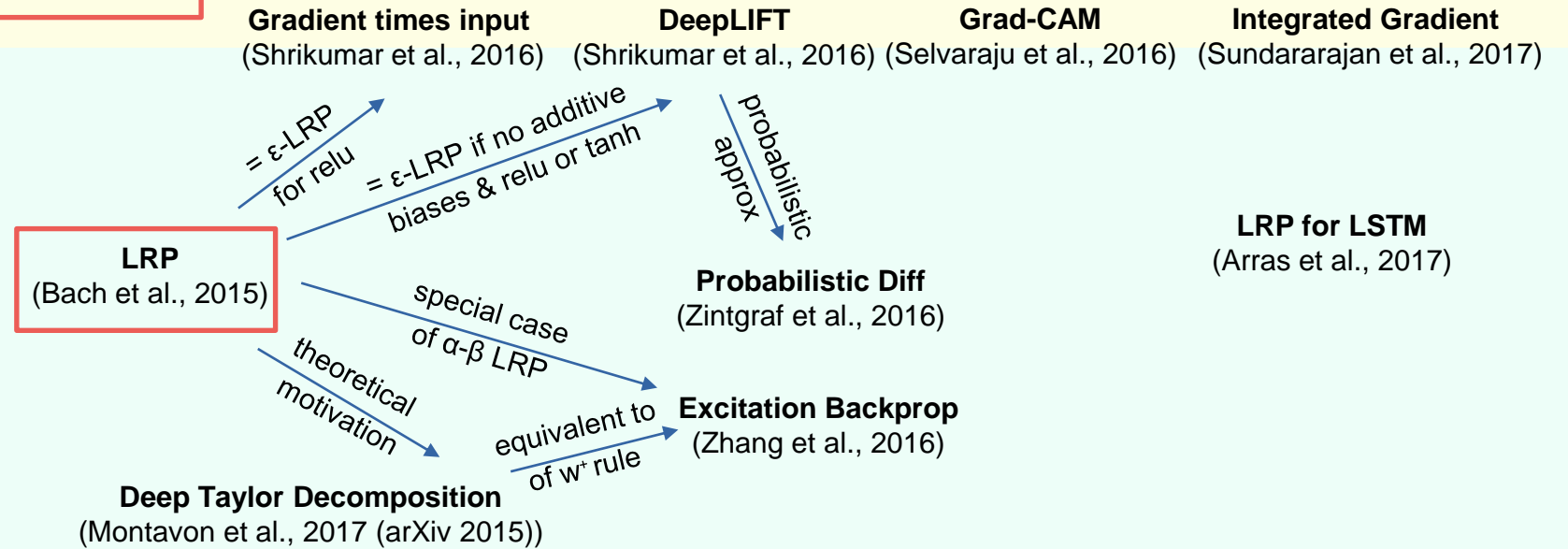
Goodbye Blackbox ML!

Historical remarks on Explaining Predictors

Gradients
 Sensitivity (Baehrens et al. 2010)
 Sensitivity (Morch et al., 1995)
 Sensitivity (Simonyan et al. 2014)

Gradient vs. Decomposition
 (Montavon et al., 2018)

Decomposition



Optimization

LIME (Ribeiro et al., 2016) **Meaningful Perturbations** (Fong & Vedaldi 2017) **PatternLRP** (Kindermans et al., 2017)

Deconvolution

Deconvolution (Zeiler & Fergus 2014) **Guided Backprop** (Springenberg et al. 2015)

Understanding the Model

Feature visualization (Erhan et al. 2009) **Deep Visualization** (Yosinski et al., 2015) **Inverting CNNs** (Mahendran & Vedaldi, 2015) **Inverting CNNs** (Dosovitskiy & Brox, 2015) **RNN cell state analysis** (Karpathy et al., 2015) **Synthesis of preferred inputs** (Nguyen et al. 2016) **TCAV** (Kim et al. 2018) **Network Dissection** (Zhou et al. 2017)

Explaining Neural Network Predictions

- Layer-wise relevance Propagation (LRP, **Bach et al 15**) first method to **explain** nonlinear classifiers
- based on generic **theory** (related to Taylor decomposition – deep Taylor decomposition **M et al 17**)
 - applicable to any NN with monotonous activation, BoW models, Fisher Vectors, SVMs etc.

Explanation: “Which pixels contribute how much to the classification” (**Bach et al 2015**)
(what makes this image to be classified as a car)

$$f(x) = \sum_p h_p$$

Sensitivity / Saliency: “Which pixels lead to increase/decrease of prediction score when changed”
(what makes this image to be classified more/less as a car) (Baehrens et al 10, **Simonyan et al 14**)

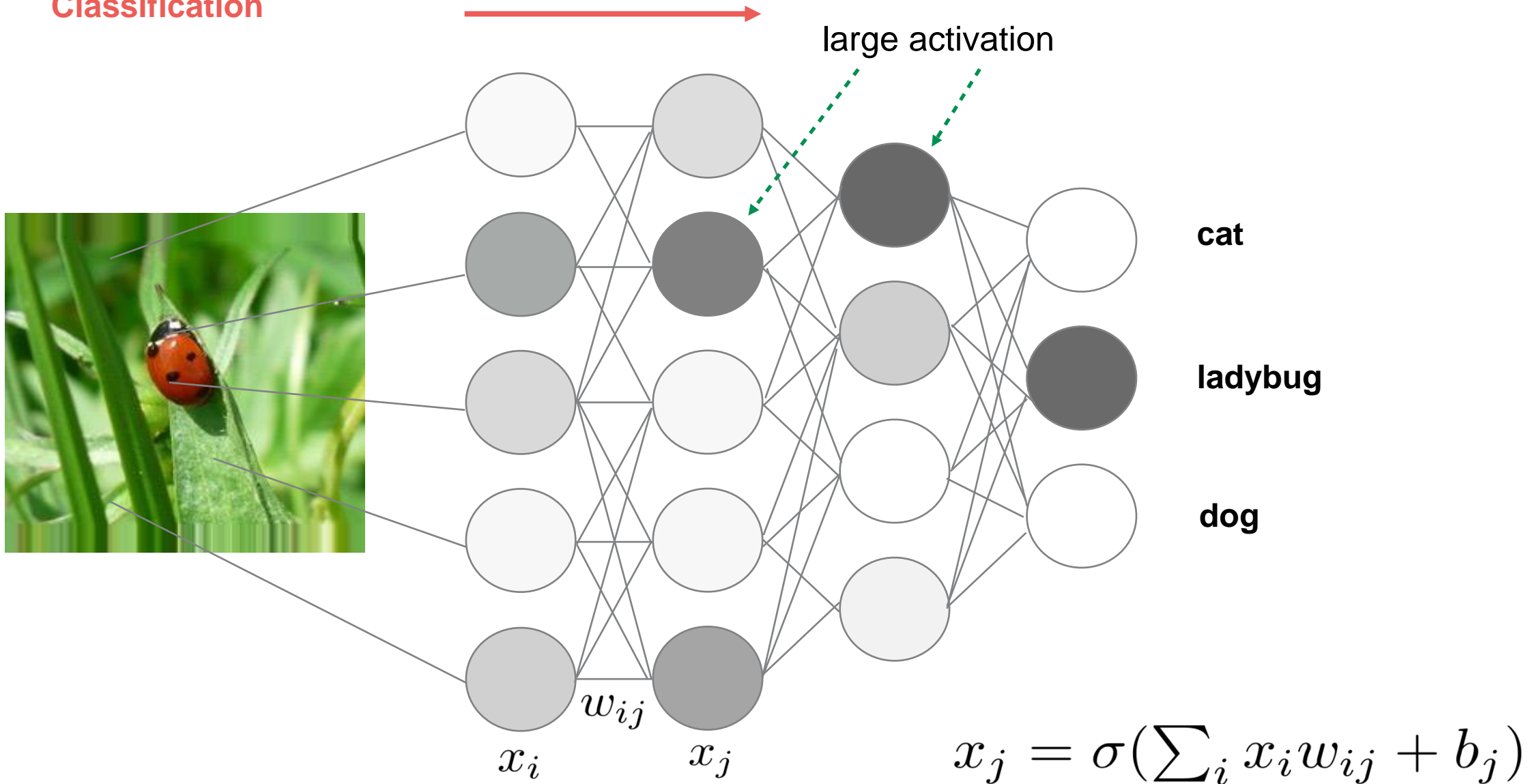
$$h_p = \left\| \left\| \frac{\partial}{\partial x_p} f(x) \right\| \right\|_{\infty}$$

Deconvolution: “Matching input pattern for the classified object in the image” (**Zeiler & Fergus 2014**)
(relation to $f(x)$ not specified)

Each method solves a **different** problem!!!

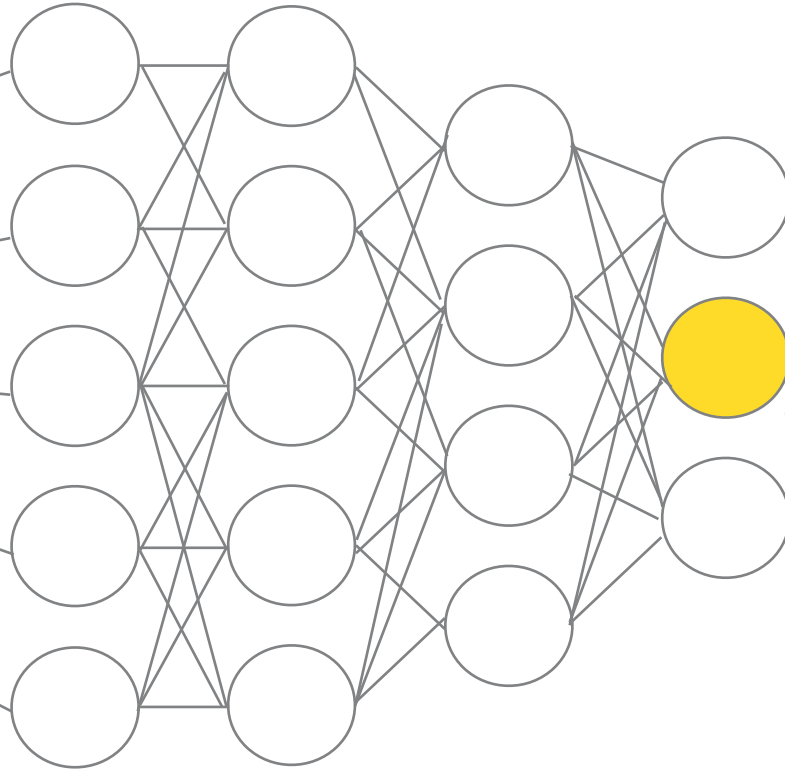
Explaining Neural Network Predictions

Classification



Explaining Neural Network Predictions

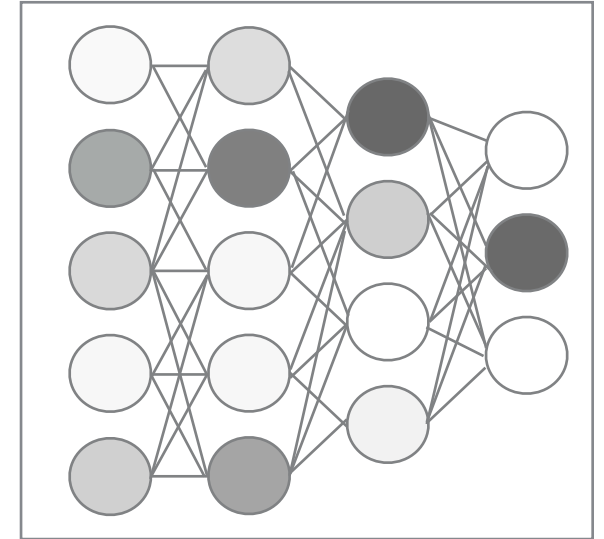
Explanation



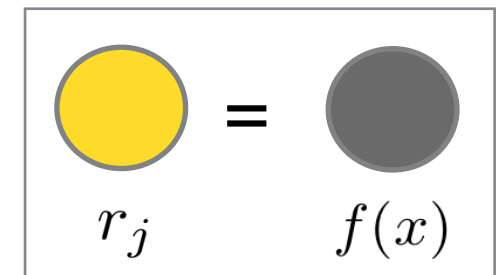
cat

ladybug

dog

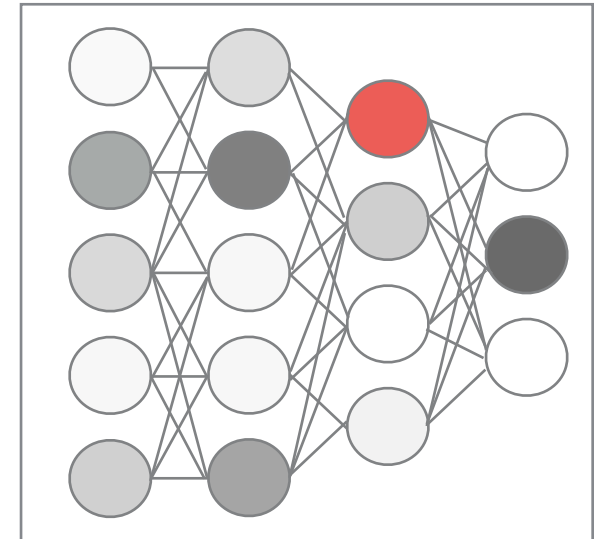
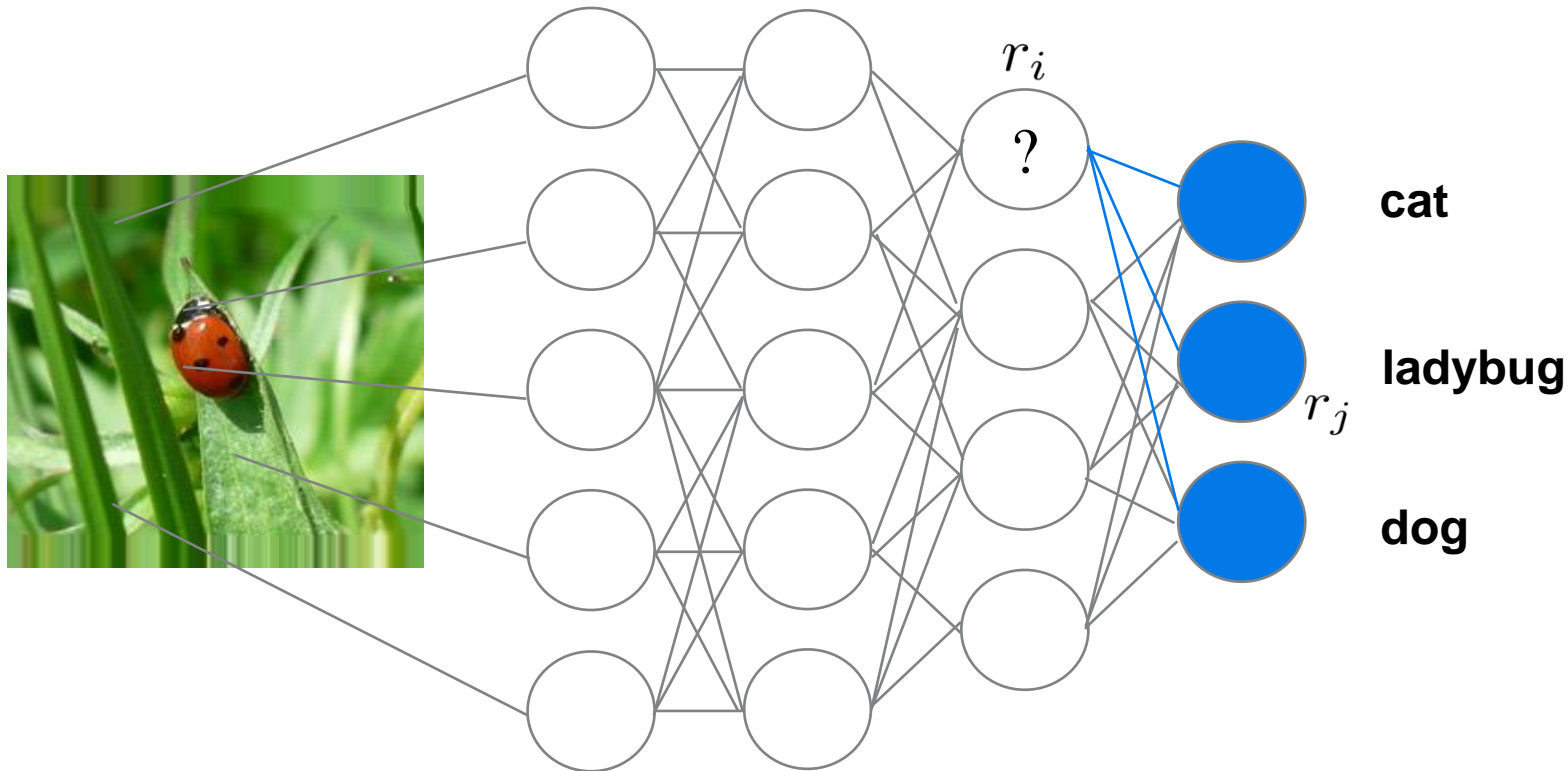


Initialization



Explaining Neural Network Predictions

Explanation



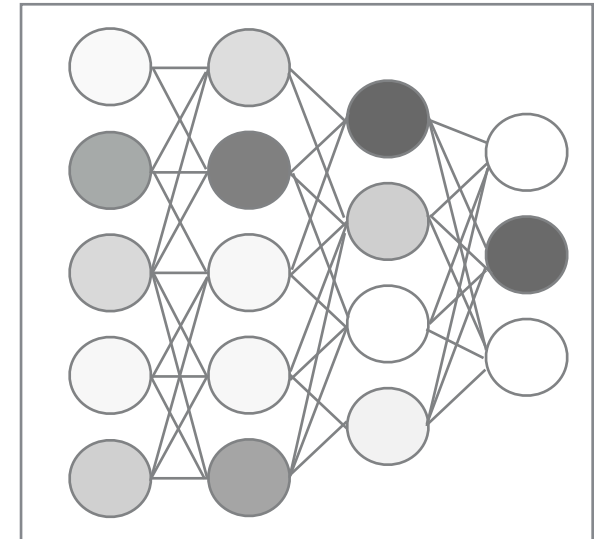
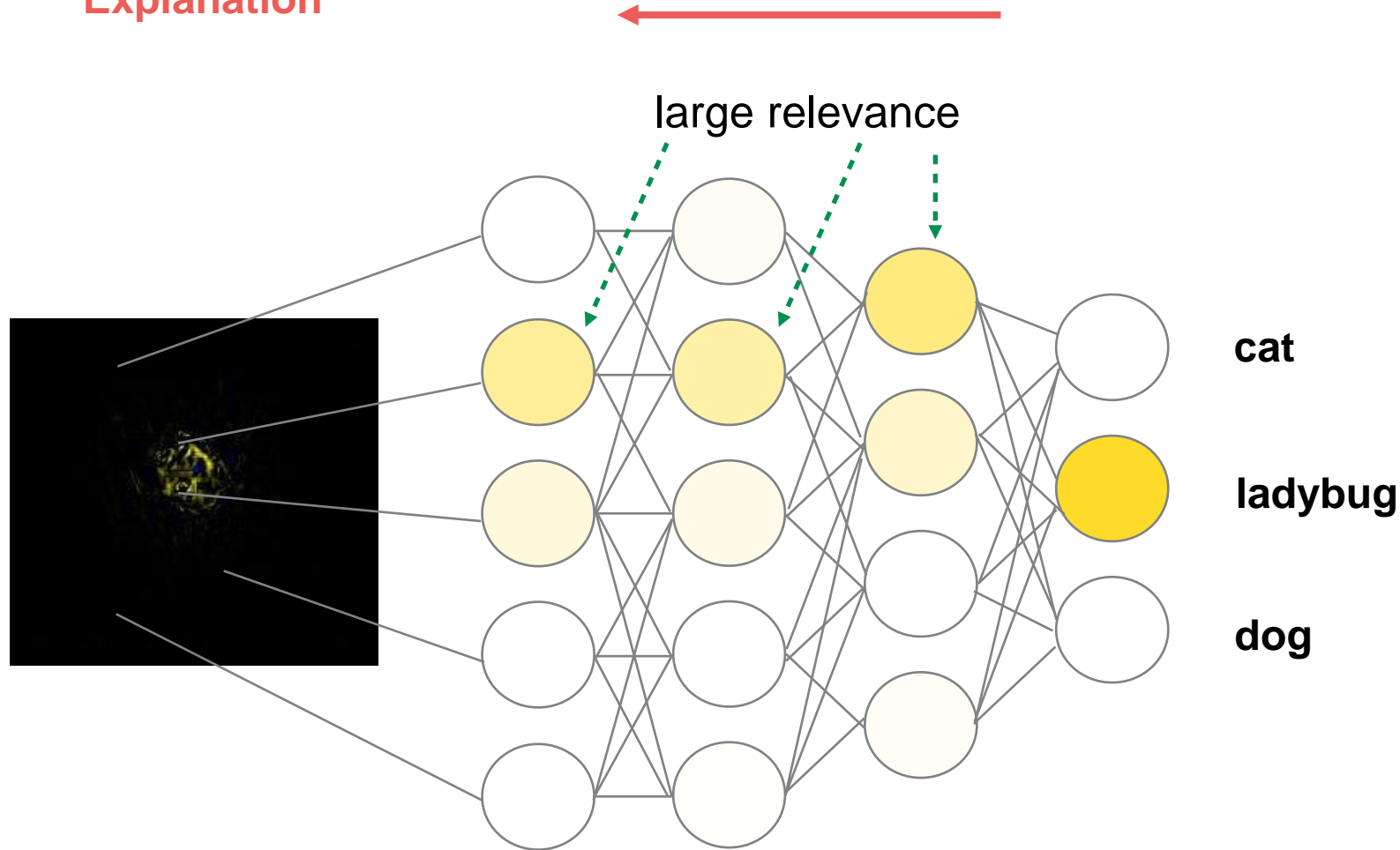
Theoretical interpretation
Deep Taylor Decomposition

$$r_i = x_i \sum_j \frac{w_{ij} r_j}{\sum_i x_i w_{ij}} = x_i C_i$$

r_i depends on the activations **and** the weights

Explaining Neural Network Predictions

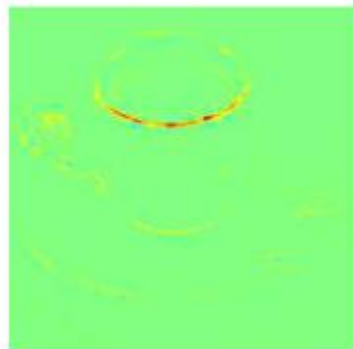
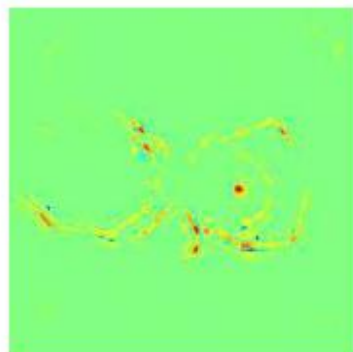
Explanation



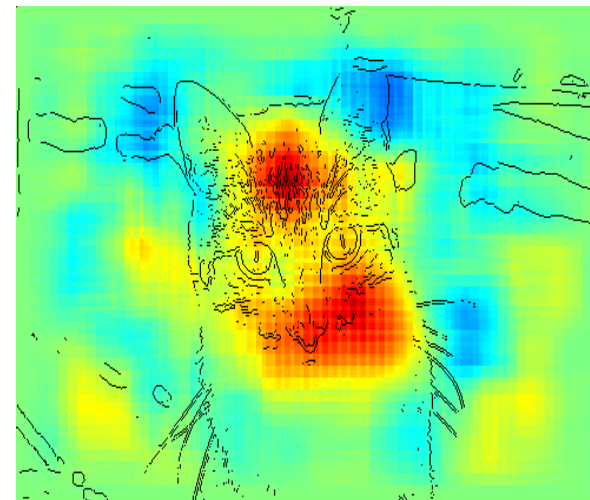
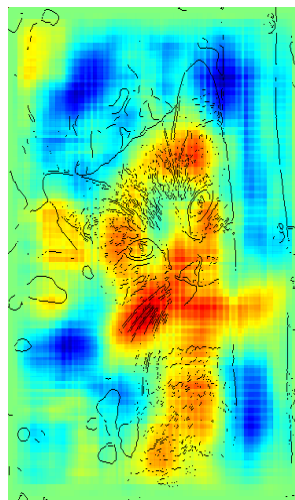
Relevance Conservation Property

$$\sum_p r_p = \dots = \sum_i r_i = \sum_j r_j = \dots = f(x)$$

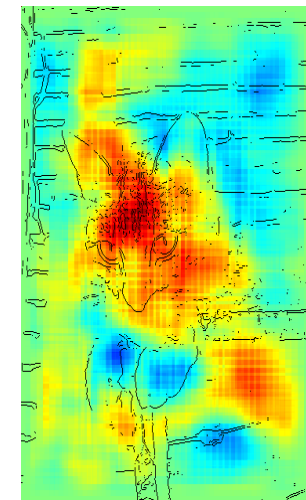
Explaining Predictions Pixel-wise



Neural networks



Kernel methods



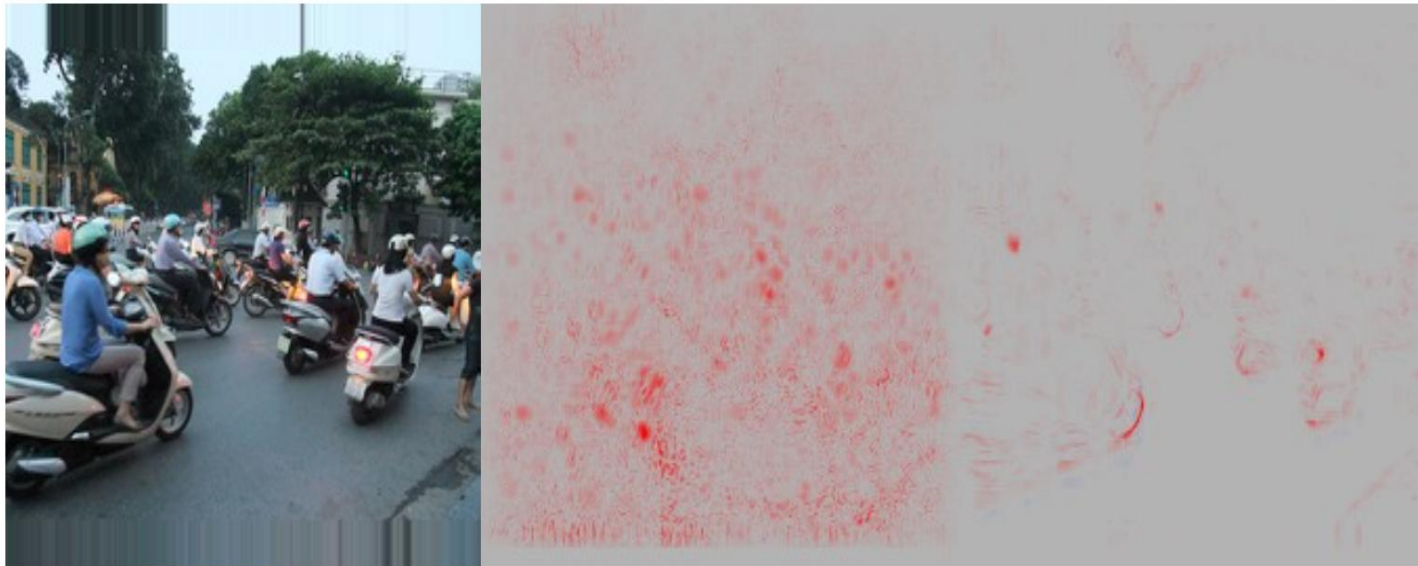
Some Digestion on Explaining

Sensitivity analysis is often not the question that you would like to ask!

Image

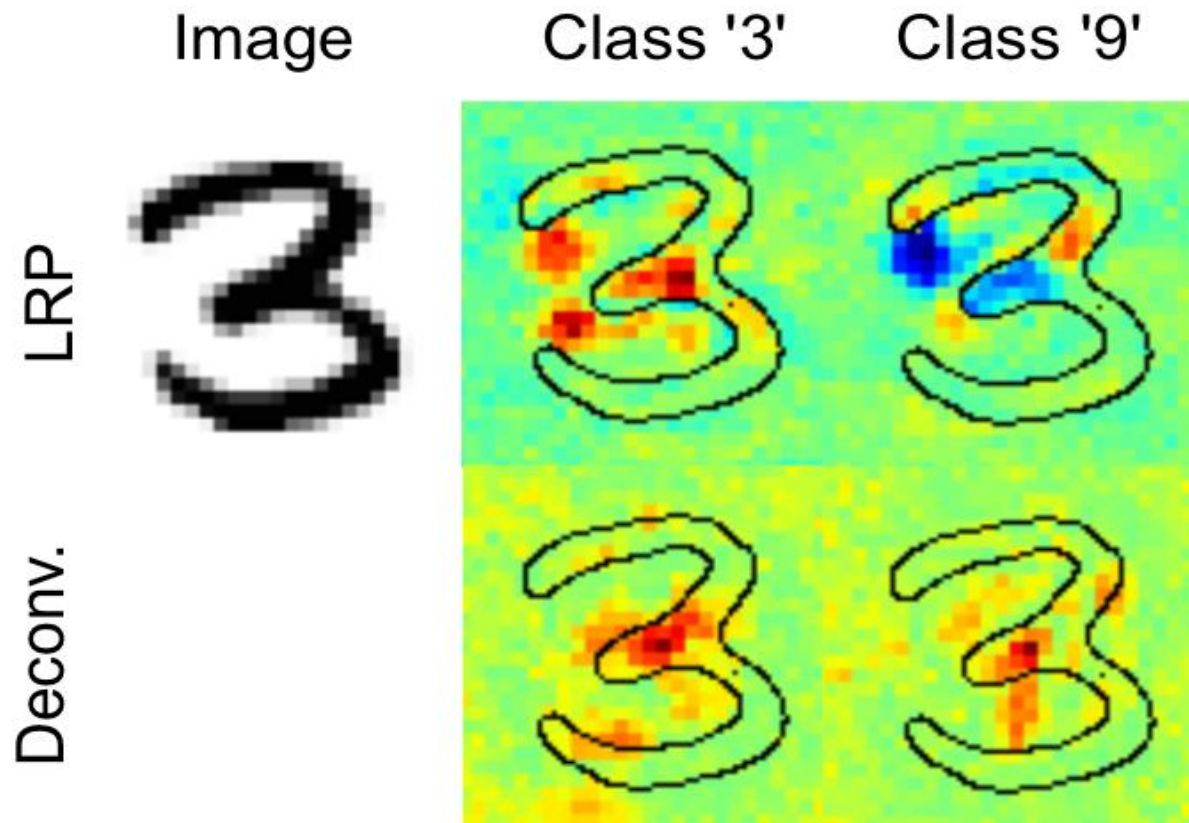
Sensitivity ℓ_2

LRP



Advantages of LRP over both Sensitivity and Deconvolution

Positive and Negative Evidence: LRP distinguishes between positive evidence, supporting the classification decision, and negative evidence, speaking against the prediction



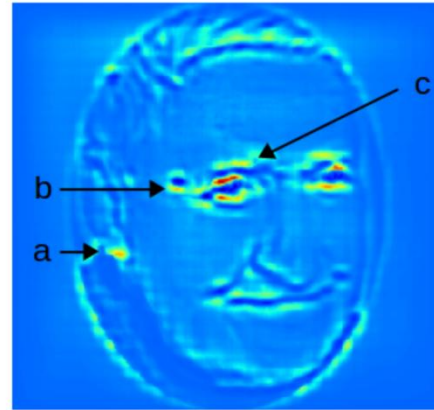
LRP indicates what speaks for class '3' and speaks against class '9'

The sign of Sensitivity and Deconvolution does not have this interpretation.
-> taking norm gives unsigned visualizations

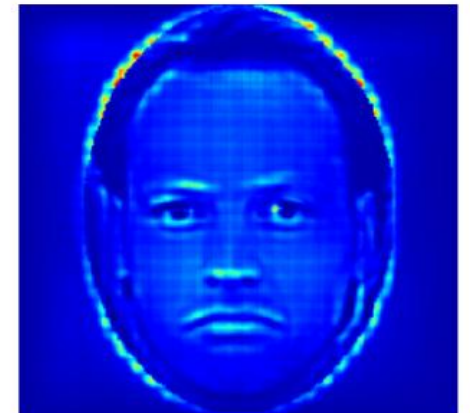
Applying Explanation in Vision and Text

Application: Faces

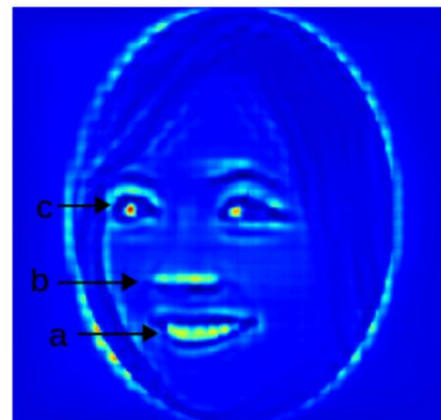
What makes
you look old ?



What makes
you look sad ?



What makes
you look attractive ?



Application: Document Classification

sci.space

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.

rec.motorcycles

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.

sci.med

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.

Explaining LSTMs

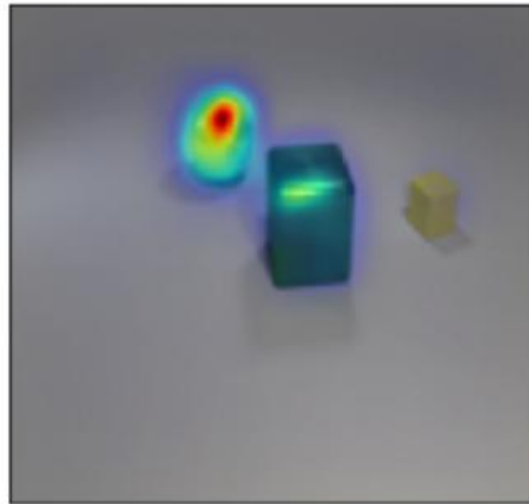
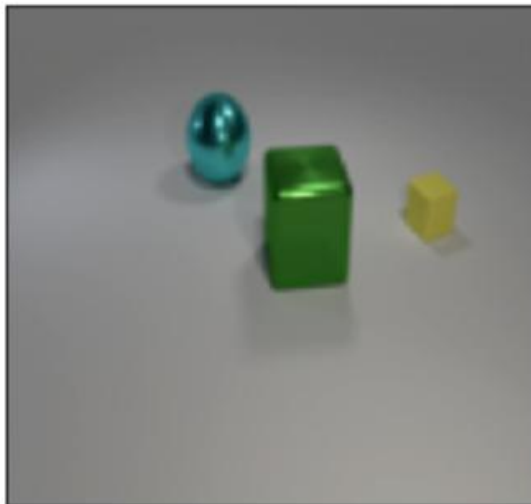
Second example: Visual question answering on the CLEVR dataset.

Question

LRP

there is a metallic cube ; are there any large cyan metallic objects behind it ?

there is a metallic cube ; are there any large cyan metallic objects behind it ?

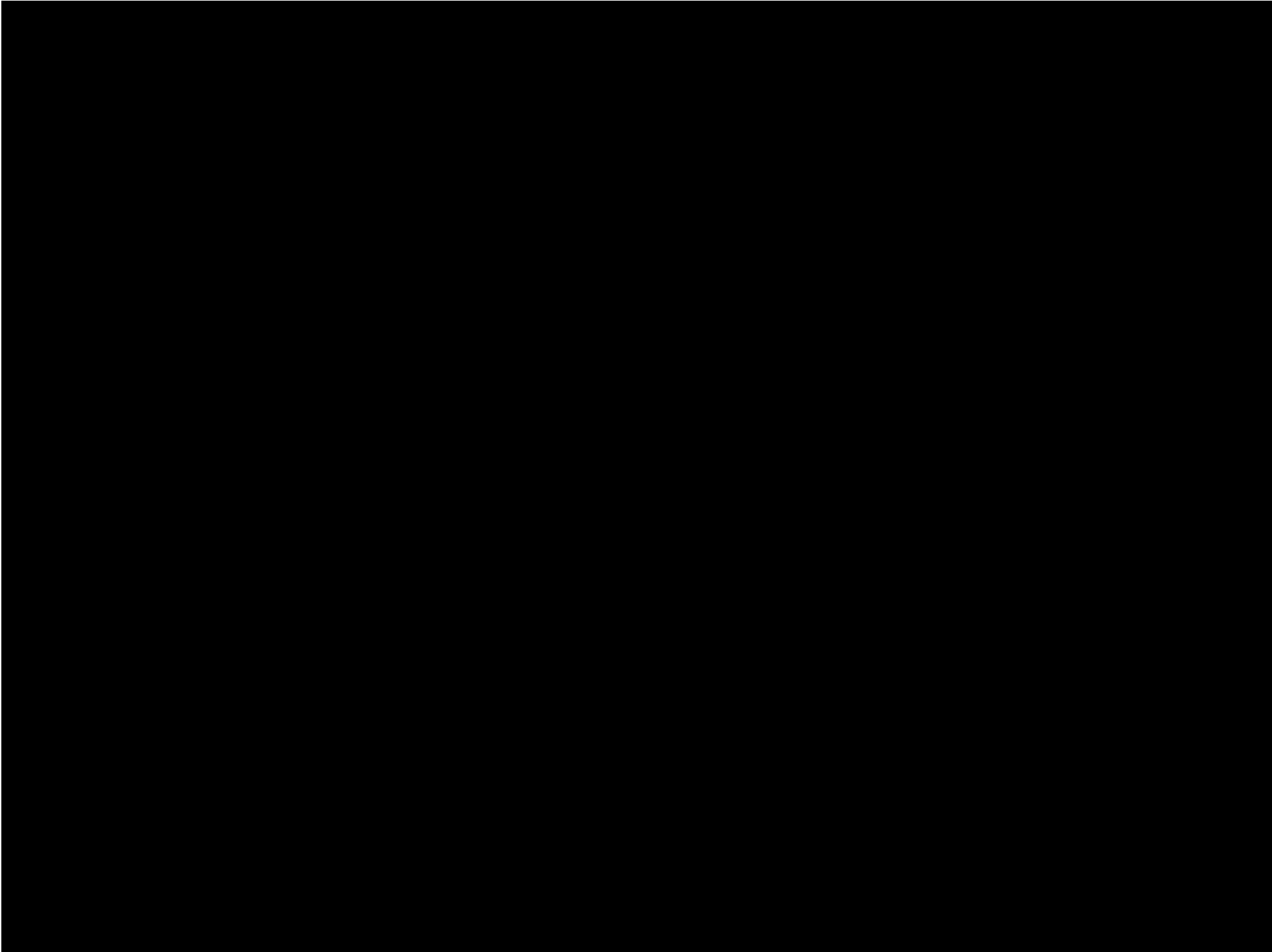


—> model understands the question and correctly identifies the object of interest

(Arras et al., in prep)

Understanding learning models for complex gaming scenarios

Analysing Breakout: LRP vs. Sensitivity



Machine Learning in the Sciences

Machine Learning in Neuroscience

Brain Computer Interfacing: ‚Brain Pong‘

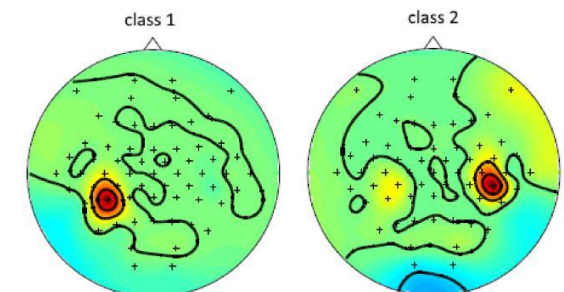


Berlin Brain Computer Interface

- ML reduces patient training from 300h -> 5min

Applications

- help/hope for patients (ALS, stroke...)
- neuroscience
- neurotechnology (video coding, gaming, monitoring driving)



Leitmotiv: ›let the machines learn‹

ML4 Quantum Chemistry

Machine Learning in Chemistry, Physics and Materials

Matthias Rupp, Anatole von Lilienfeld,
Alexandre Tkatchenko, Klaus-Robert Müller

[Rupp et al. Phys Rev Lett 2012, Snyder et al. Phys Rev Lett
2012, Hansen et al. JCTC 2013 and JPCL 2015]

Machine Learning for chemical compound space

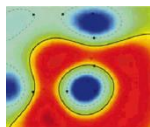
Ansatz:

$$\{Z_I, \mathbf{R}_I\} \xrightarrow{\text{ML}} E$$

instead of

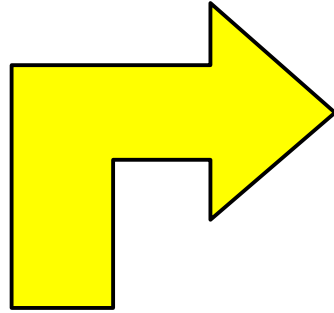
$$\hat{H}(\{Z_I, \mathbf{R}_I\}) \xrightarrow{\Psi} E$$

$$\hat{H}\Psi = E\Psi$$

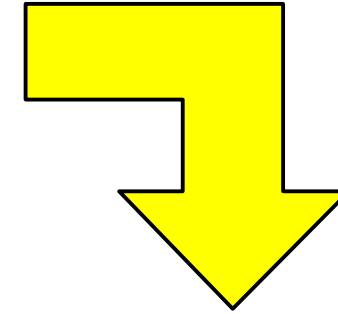


[from von Lilienfeld]

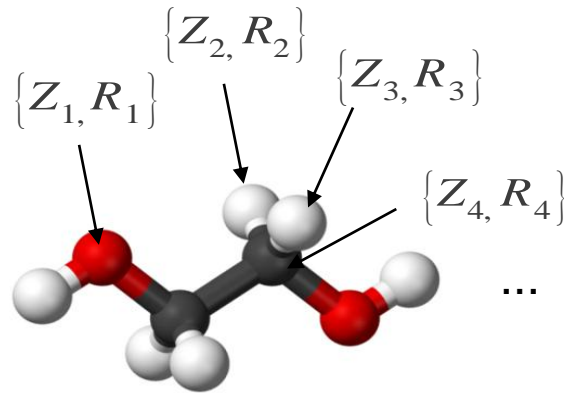
Coulomb representation of molecules



$$M_{ii} = Z_i^{2.4}$$
$$M_{ij} = \frac{Z_i Z_j}{\|R_i - R_j\|}$$

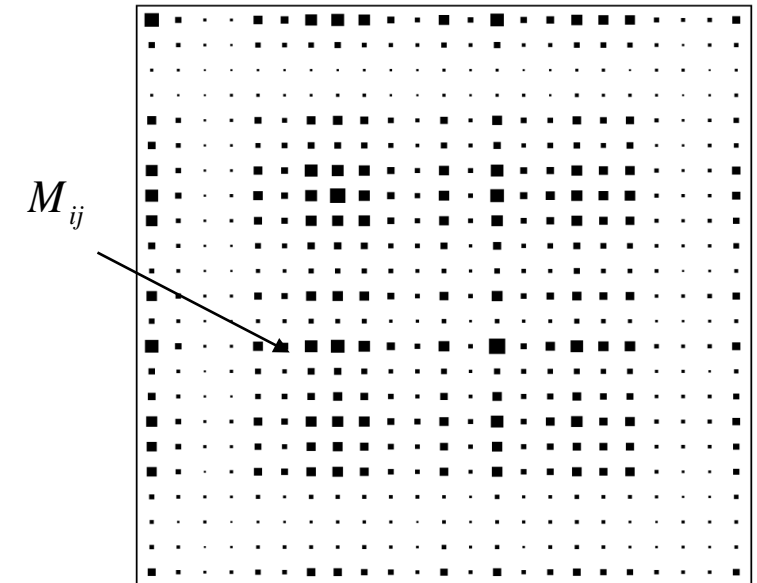


$$M \in {}^{23 \times 23}$$



+ phantom atoms

$$\{0, R_{21}\} \quad \{0, R_{22}\} \quad \{0, R_{23}\}$$



Coulomb Matrix (Rupp, Müller et al 2012, PRL)

$$d(\mathbf{M}, \mathbf{M}') = \sqrt{\sum_{IJ} |M_{IJ} - M'_{IJ}|^2}$$

Kernel ridge regression

Distances between \mathbf{M} define Gaussian kernel matrix \mathbf{K}

$$k(\mathbf{M}, \mathbf{M}') = \exp\left(-\frac{d(\mathbf{M}, \mathbf{M}')^2}{2\sigma^2}\right)$$

Predict energy as sum over weighted Gaussians

$$E^{est}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i) + b$$

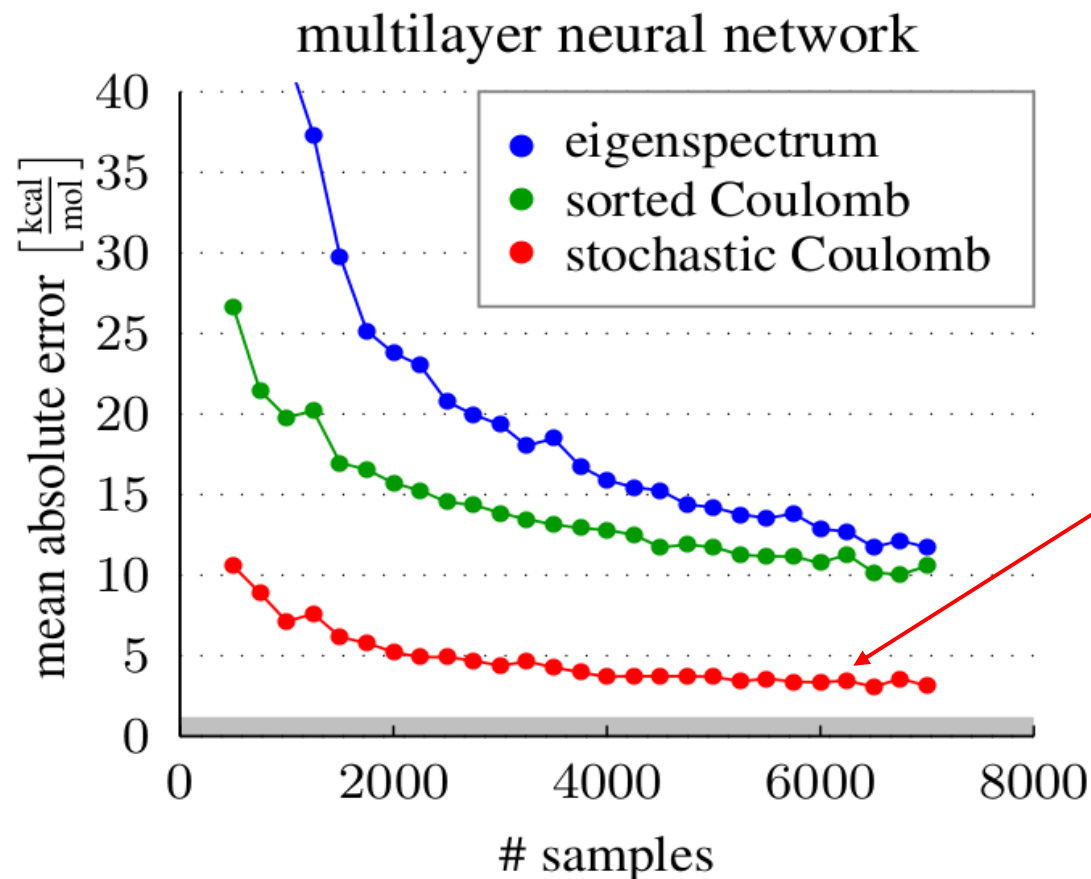
using weights that minimize error in training set

$$\min_{\alpha} \sum_i (E^{est}(\mathbf{M}_i) - E_i^{ref})^2 + \lambda \sum_i \alpha_i^2$$
$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$

Exact solution

As many parameters as molecules + 2 global parameters, characteristic length-scale or kT of system (σ), and noise-level (λ)

Predicting Energy of small molecules: Results



March 2012

Rupp et al., PRL

9.99 kcal/mol

(kernels + eigenspectrum)

December 2012

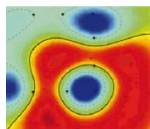
Montavon et al., NIPS

3.51 kcal/mol

(Neural nets + Coulomb sets)

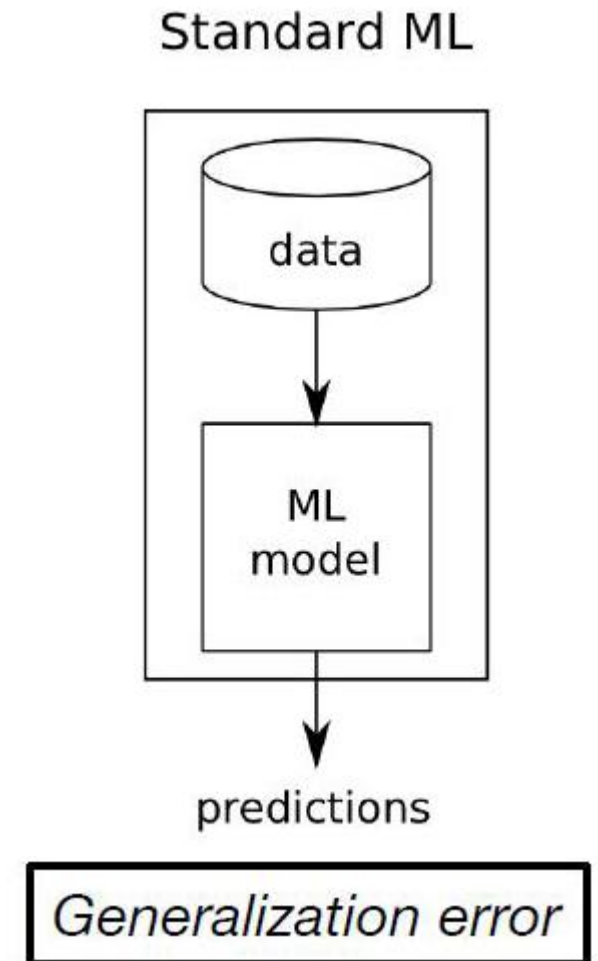
2015 Hansen et al 1.3kcal/mol at
10 million times faster than the
state of the art

Prediction considered chemically
accurate when MAE is below **1**
kcal/mol



Dataset available at <http://quantum-machine.org>

Is the Generalization Error all we need?



Application: Comparing Classifiers (Lapuschkin et al CVPR 2016)

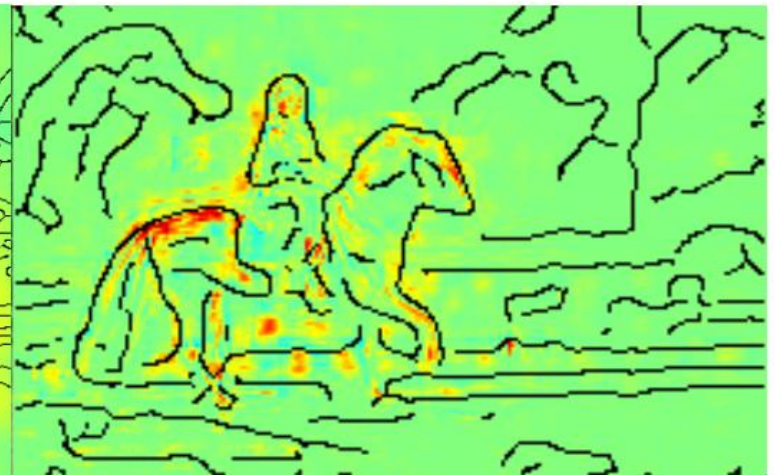
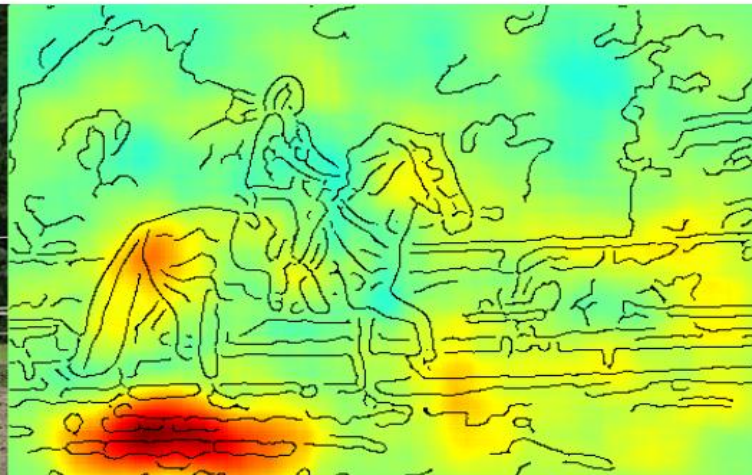
Test error for various classes:

Fisher	aeroplane	bicycle	bird	boat	bottle	bus	car
	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%
Fisher	cat	chair	cow	diningtable	dog	horse	motorbike
	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%
Fisher	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
	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



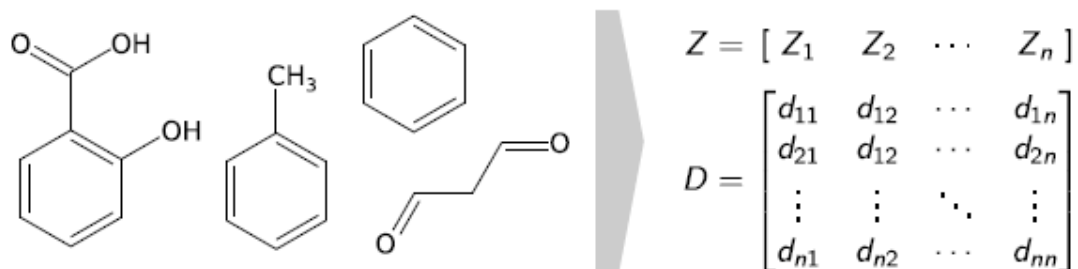
Learning Atomistic Representations with Deep Tensor Neural Networks

Kristof Schütt, Farhad Arbabzadah,
Stefan Chmiela, Alexandre Tkatchenko,
Klaus-Robert Müller

[Schütt et al. Nature Communications 2017, Chmiela et al
Science Advances 2017, Brockherde et al Nat. Comm. 2017]

Deep Tensor Neural Network (DTNN) for representing molecules

Input: Atomic numbers and interatomic distances



Embedding of based on atom types

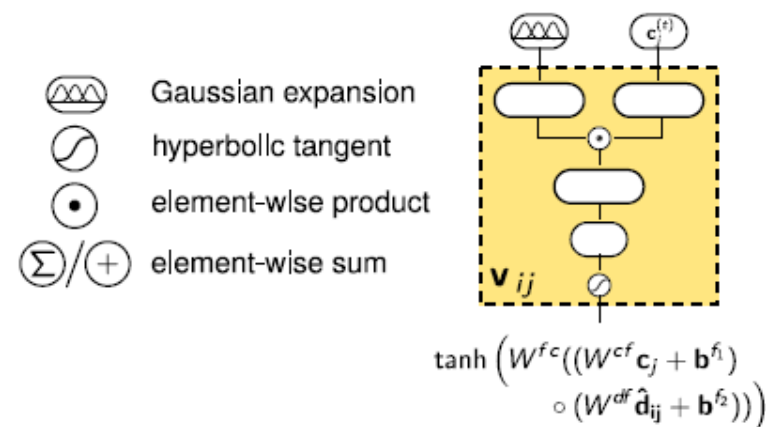
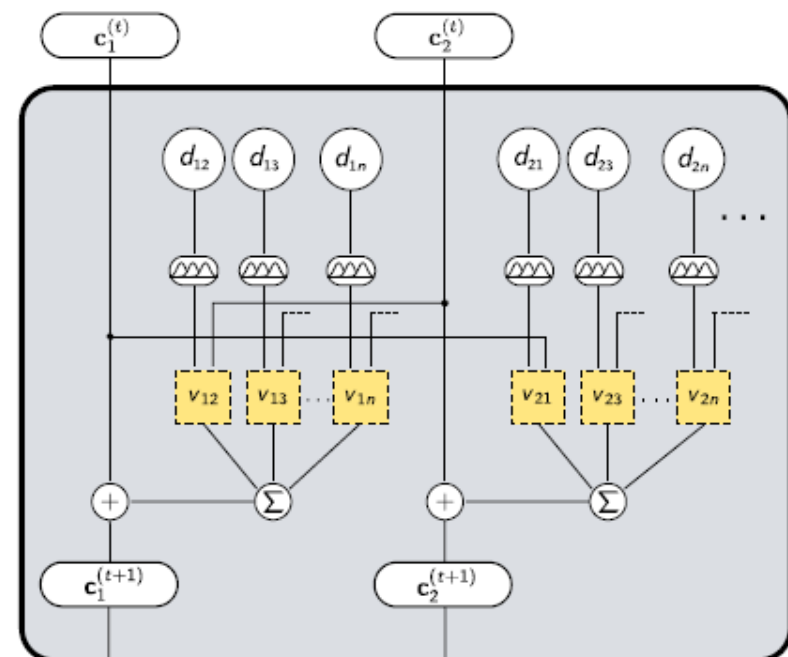
$$\mathbf{x}_i^{(0)} = \mathbf{x}_{Z_i} \in \mathbb{R}^d$$

Add interaction with environment using $t = 1 \dots T$ sequential refinements $\mathbf{v}_i^{(t)}$

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \mathbf{v}_i^{(t)} \left(\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_{n_{\text{atoms}}}^{(t)}, d_{i1}, \dots, d_{in_{\text{atoms}}} \right)$$

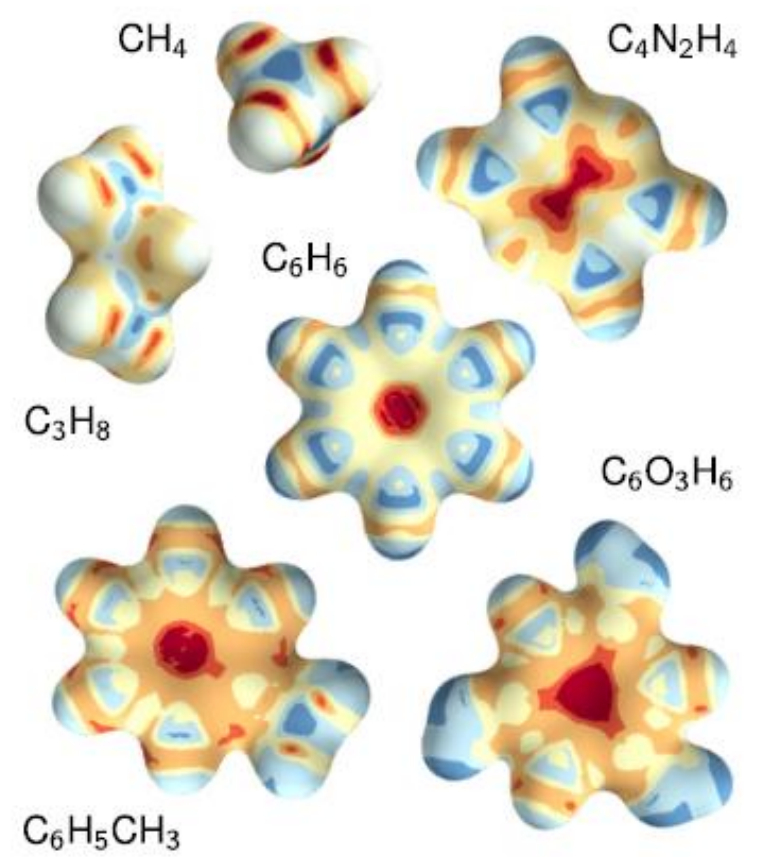
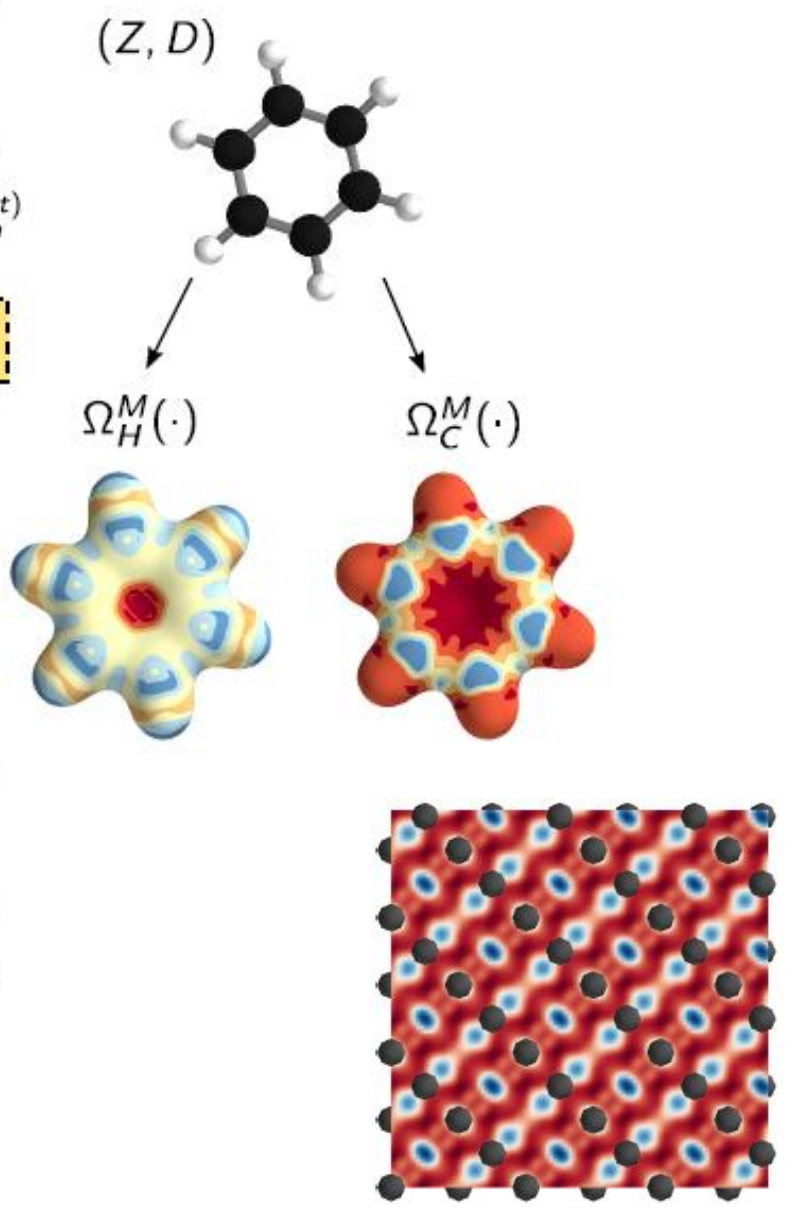
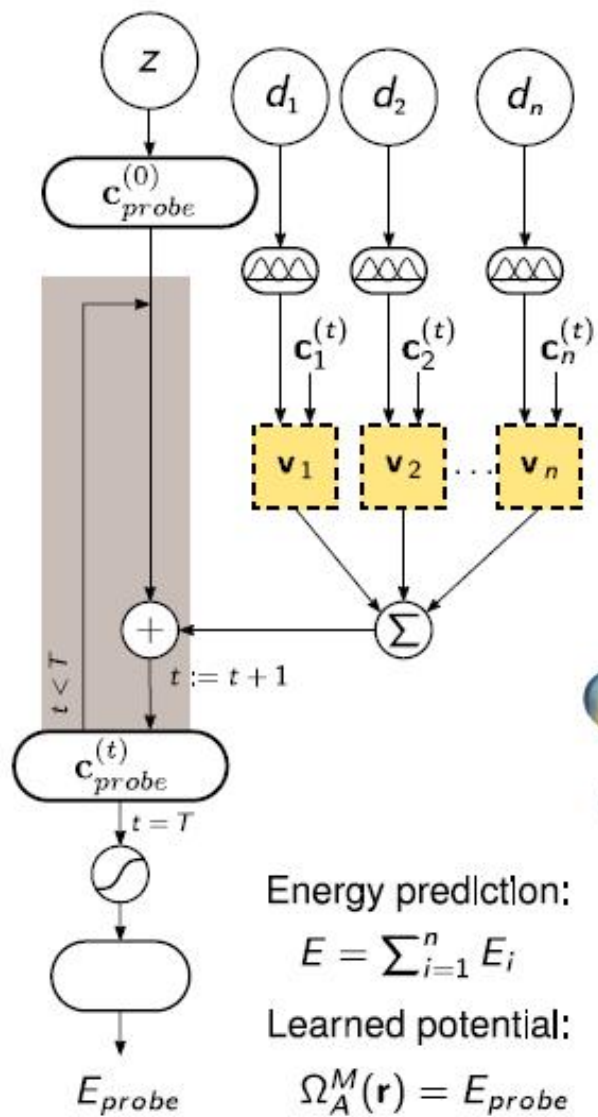
Prediction via atom-wise contributions:

$$\hat{E} = \sum_{i=1}^{n_{\text{atoms}}} f_{\text{out}}(\mathbf{x}_i^{(T)})$$

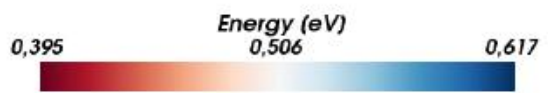


Gaining insights for Physics

Toward Quantum Chemical Insights: supervised



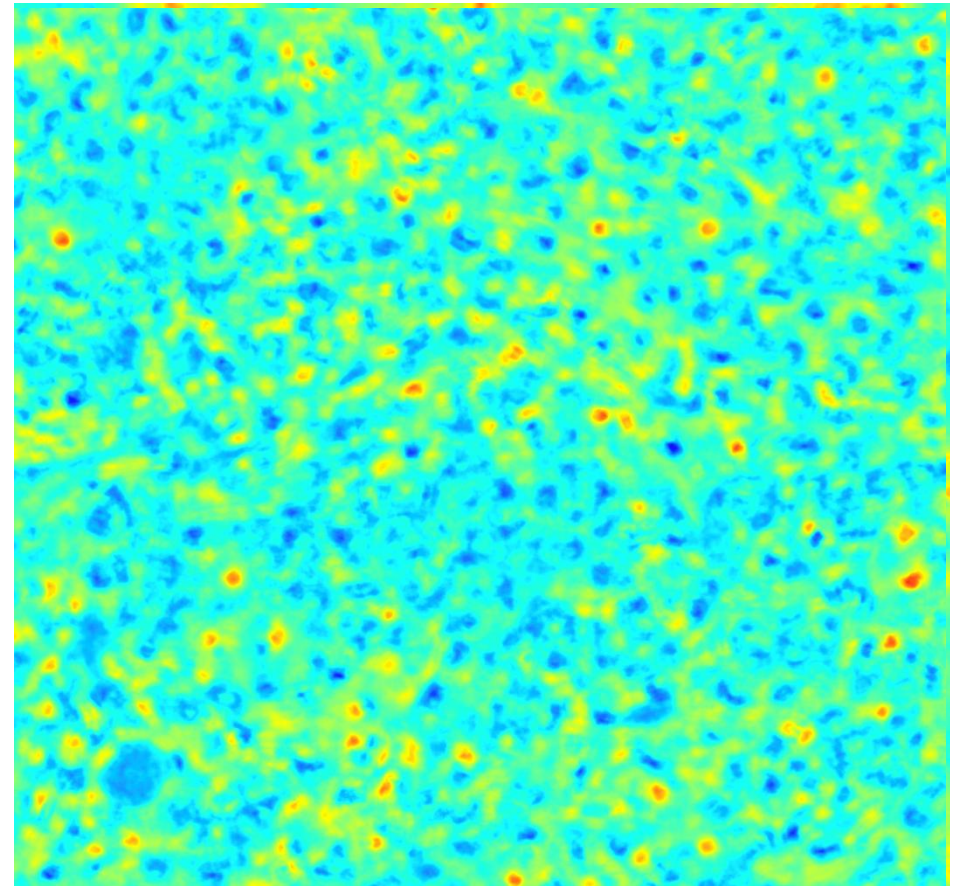
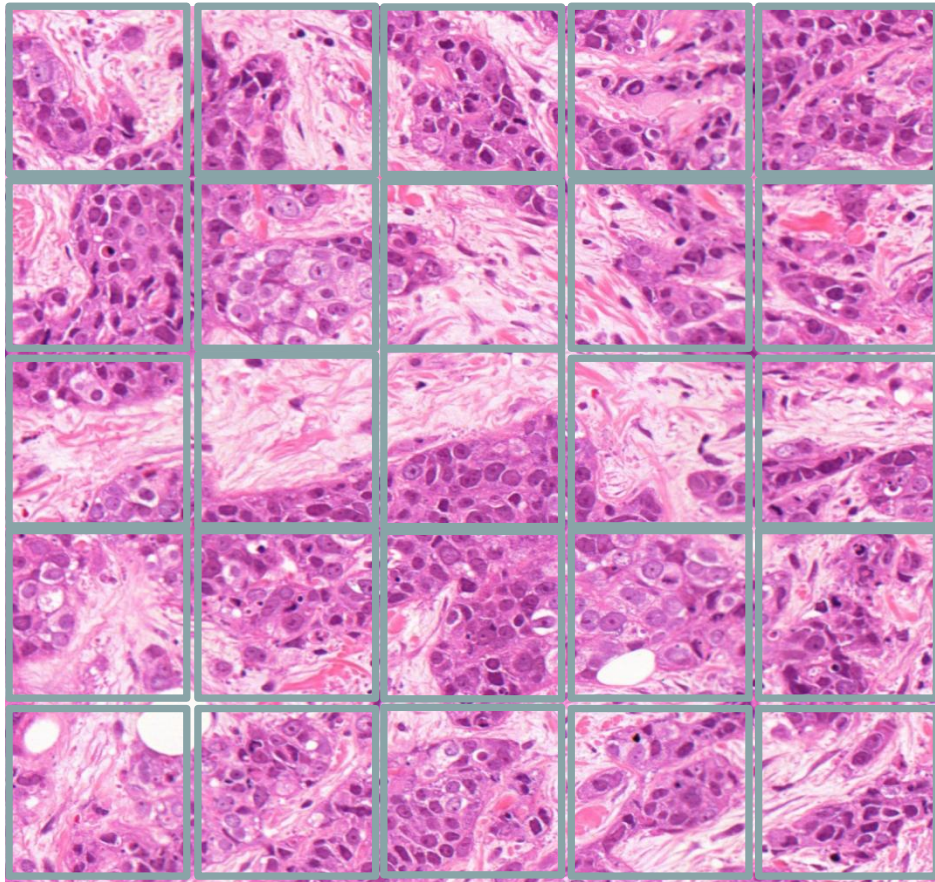
[Schütt et al. Nat Comm. 2017,
 Schütt et al JCP 2018]



Machine Learning for morpho-molecular Integration

Alexander Binder^{1,6}, Michael Bockmayr^{2,10}, Miriam Hägele¹, Stephan Wienert², Daniel Heim², Katharina Hellweg³, Albrecht Stenzinger⁴, Laura Parlow², Jan Budczies², Benjamin Goepfert⁴, Denise Treue², Manato Kotani⁵, Masaru Ishii⁵, Manfred Dietel², Andreas Hocke³, Carsten Denkert^{2,7}, Klaus-Robert Müller^{1,8,9,*} and Frederick Klauschen^{2,7,*}

Interpretable ML



Bach et al., PLoS1 2015
Klauschen et al., US Patent #9558550
Binder et al., *in revision*

Machine learning based integration of morphological and molecular tumor profiles

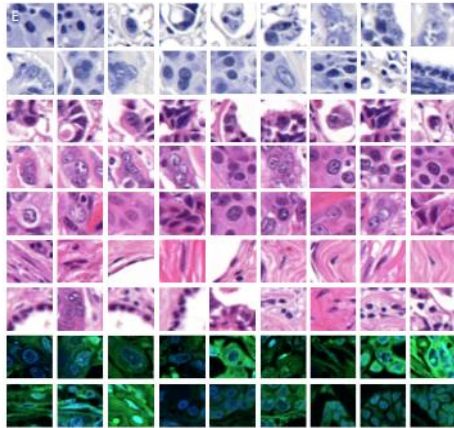
MICROSCOPIC AND MOLECULAR DATA

TRAINING

PREDICTION

INTEGRATION/ INTERPRETATION

histo-morphological features



div. Image modalities: brightfield, confocal

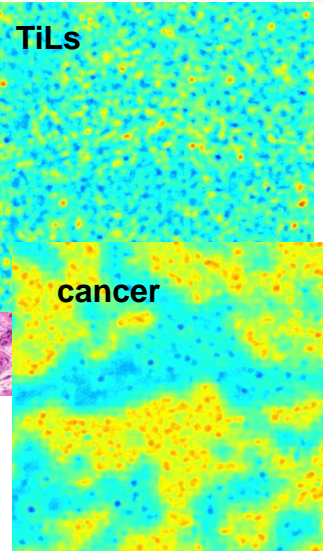
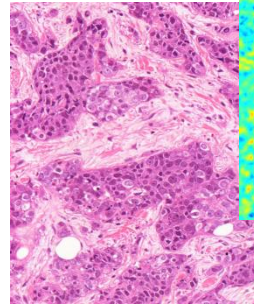
in-house data base

cell types

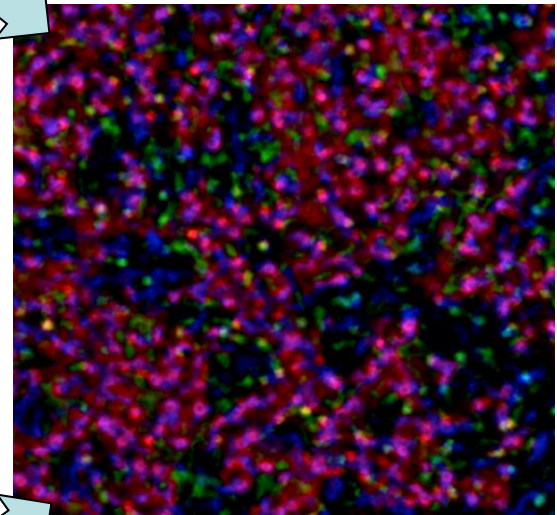
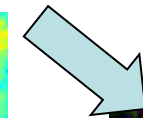
cancer

lymphocytes

stroma

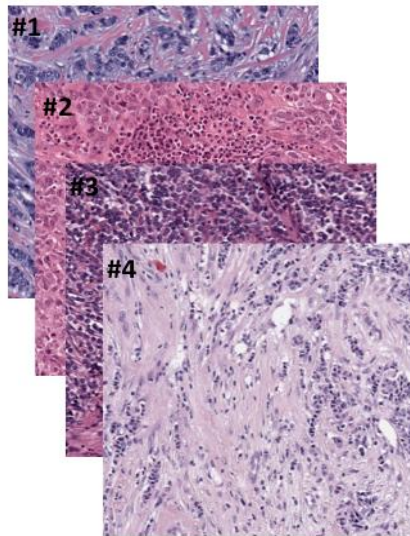


red: carcinoma,
green: TiLs,
blue: molecular property



„computational
fluorescence
microscopy“

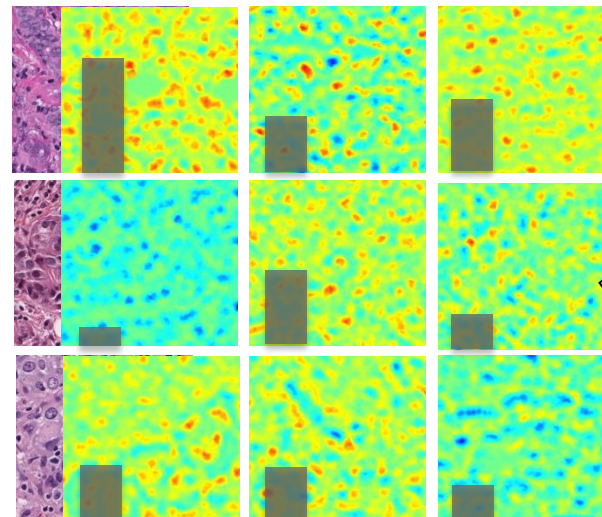
molecular features



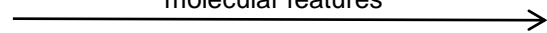
TCGA data base

molecular profiles

Run	49624671	40261700	51380533	60125348
Run_2	190224746	202220437	206107867	208243491
Run_4	490185471	418474483	486118242	491821100
Run_5	110224746	101076187	109472222	121108173
Run_Catenin	596029913	587967633	605510243	620601009
Run_7	44468229	418791197	440002208	44862116
Run_8	483171767	438840249	481120764	438499843
Run_9	432228023	420406513	444410128	442227871
Run_p175	21031542	21224494	2178986	238424894
c_4b	342334	40021429	48861061	492124314
c_4b_1	425644807	416461824	44486827	4212616027
c_4b_2	98724048	117641098	121712211	121648876
c_4b_3	43138274	410268601	400176784	424618786
C_Ant_p133	48293204	418848424	44322771	423841278
Catenin_7_p1660138	41306118	416442022	44467122	47006118
Catenin_1	7390208	45815871	4602287	40112007
CD30	43844222	411076024	4444677	44884209
CD31	42181708	414614401	439982127	458780766
CD36	44248474	410772228	421148984	440021616
CD41	41888184	410480588	421024247	417392992
CD43	401746364	410010089	442171054	421087107
CD43_p145	410891806	410946624	41748182	428032481
CD43_p178	42684708	410176144	42769184	41418711
CD43_p184	411864185	411861924	422019084	421048278
CD47	401746364	410010089	442171054	421087107
Claudin_7	421734951	418444882	418822114	418822114
Claudin_11	401746364	410010089	442171054	421087107
Cyclin_D1	424001238	421710388	444444441	421140749
Cyclin_E1	74444444	417444442	417444425	421141789
Cyclin_E2	44444411	444444441	444444441	417087008
DNA	44444444	44444444	44444444	44444444
DNA2	44444444	44444444	44444444	44444444
E_Cadherin	44444444	44444444	44444444	44444444
HIF2	44444444	44444444	44444444	44444444
HIF3	44444444	44444444	44444444	44444444
MAPK1	44444444	44444444	44444444	44444444
MAPK3	44444444	44444444	44444444	44444444
MAPK8	44444444	44444444	44444444	44444444
MAPK9	44444444	44444444	44444444	44444444
MDM2	44444444	44444444	44444444	44444444
MDM4	44444444	44444444	44444444	44444444
MDM5	44444444	44444444	44444444	44444444
MDM6	44444444	44444444	44444444	44444444
MDM7	44444444	44444444	44444444	44444444
MDM8	44444444	44444444	44444444	44444444
MDM9	44444444	44444444	44444444	44444444
MDM10	44444444	44444444	44444444	44444444
MDM11	44444444	44444444	44444444	44444444
MDM12	44444444	44444444	44444444	44444444
MDM13	44444444	44444444	44444444	44444444
MDM14	44444444	44444444	44444444	44444444
MDM15	44444444	44444444	44444444	44444444
MDM16	44444444	44444444	44444444	44444444
MDM17	44444444	44444444	44444444	44444444
MDM18	44444444	44444444	44444444	44444444
MDM19	44444444	44444444	44444444	44444444
MDM20	44444444	44444444	44444444	44444444
MDM21	44444444	44444444	44444444	44444444
MDM22	44444444	44444444	44444444	44444444
MDM23	44444444	44444444	44444444	44444444
MDM24	44444444	44444444	44444444	44444444
MDM25	44444444	44444444	44444444	44444444
MDM26	44444444	44444444	44444444	44444444
MDM27	44444444	44444444	44444444	44444444
MDM28	44444444	44444444	44444444	44444444
MDM29	44444444	44444444	44444444	44444444
MDM30	44444444	44444444	44444444	44444444
MDM31	44444444	44444444	44444444	44444444
MDM32	44444444	44444444	44444444	44444444
MDM33	44444444	44444444	44444444	44444444
MDM34	44444444	44444444	44444444	44444444
MDM35	44444444	44444444	44444444	44444444
MDM36	44444444	44444444	44444444	44444444
MDM37	44444444	44444444	44444444	44444444
MDM38	44444444	44444444	44444444	44444444
MDM39	44444444	44444444	44444444	44444444
MDM40	44444444	44444444	44444444	44444444
MDM41	44444444	44444444	44444444	44444444
MDM42	44444444	44444444	44444444	44444444
MDM43	44444444	44444444	44444444	44444444
MDM44	44444444	44444444	44444444	44444444
MDM45	44444444	44444444	44444444	44444444
MDM46	44444444	44444444	44444444	44444444
MDM47	44444444	44444444	44444444	44444444
MDM48	44444444	44444444	44444444	44444444
MDM49	44444444	44444444	44444444	44444444
MDM50	44444444	44444444	44444444	44444444
MDM51	44444444	44444444	44444444	44444444
MDM52	44444444	44444444	44444444	44444444
MDM53	44444444	44444444	44444444	44444444
MDM54	44444444	44444444	44444444	44444444
MDM55	44444444	44444444	44444444	44444444
MDM56	44444444	44444444	44444444	44444444
MDM57	44444444	44444444	44444444	44444444
MDM58	44444444	44444444	44444444	44444444
MDM59	44444444	44444444	44444444	44444444
MDM60	44444444	44444444	44444444	44444444
MDM61	44444444	44444444	44444444	44444444
MDM62	44444444	44444444	44444444	44444444
MDM63	44444444	44444444	44444444	44444444
MDM64	44444444	44444444	44444444	44444444
MDM65	44444444	44444444	44444444	44444444
MDM66	44444444	44444444	44444444	44444444
MDM67	44444444	44444444	44444444	44444444
MDM68	44444444	44444444	44444444	44444444
MDM69	44444444	44444444	44444444	44444444
MDM70	44444444	44444444	44444444	44444444
MDM71	44444444	44444444	44444444	44444444
MDM72	44444444	44444444	44444444	44444444
MDM73	44444444	44444444	44444444	44444444
MDM74	44444444	44444444	44444444	44444444
MDM75	44444444	44444444	44444444	44444444
MDM76	44444444	44444444	44444444	44444444
MDM77	44444444	44444444	44444444	44444444
MDM78	44444444	44444444	44444444	44444444
MDM79	44444444	44444444	44444444	44444444
MDM80	44444444	44444444	44444444	44444444
MDM81	44444444	44444444	44444444	44444444
MDM82	44444444	44444444	44444444	44444444
MDM83	44444444	44444444	44444444	44444444
MDM84	44444444	44444444	44444444	44444444
MDM85	44444444	44444444	44444444	44444444
MDM86	44444444	44444444	44444444	44444444
MDM87	44444444	44444444	44444444	44444444
MDM88	44444444	44444444	44444444	44444444
MDM89	44444444	44444444	44444444	44444444
MDM90	44444444	44444444	44444444	44444444
MDM91	44444444	44444444	44444444	44444444
MDM92	44444444	44444444	44444444	44444444
MDM93	44444444	44444444	44444444	44444444
MDM94	44444444	44444444	44444444	44444444
MDM95	44444444	44444444	44444444	44444444
MDM96	44444444	44444444	44444444	44444444
MDM97	44444444	44444444	44444444	44444444
MDM98	44444444	44444444	44444444	44444444
MDM99	44444444	44444444	44444444	44444444
MDM100	44444444	44444444	44444444	44444444

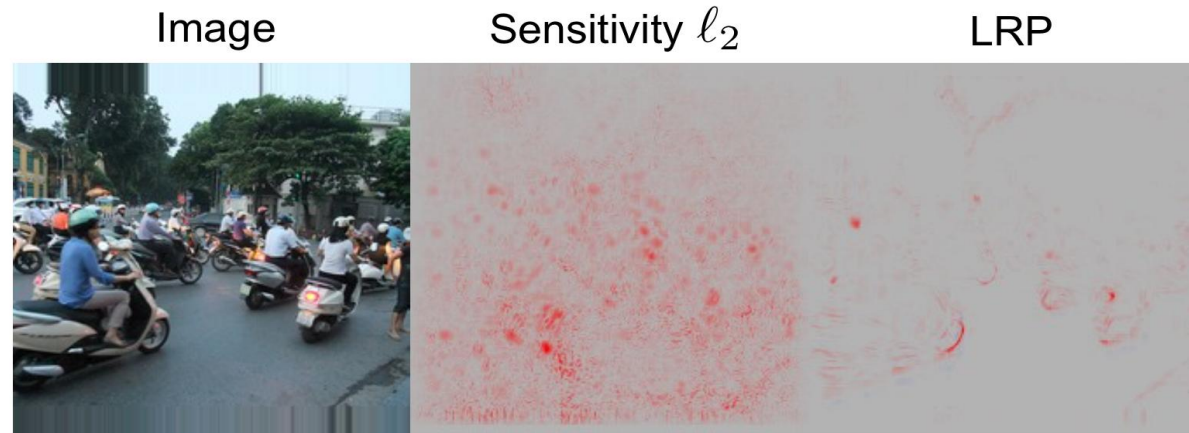


molecular features



Take Home messages

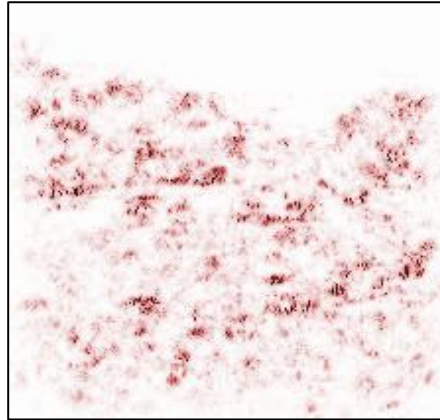
Sensitivity analysis is not the question that you would like to ask!



Sensitivity analysis:



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



Problem: sensitivity analysis does not highlight cars

Observation:

$$\sum_{i=1}^d \left(\frac{\partial f}{\partial x_i} \right)^2 = \|\nabla_{\mathbf{x}} f\|^2$$

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

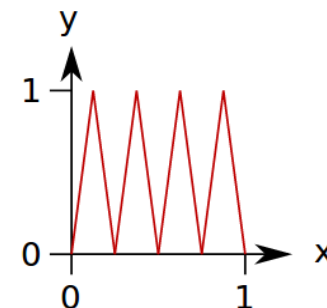
Explanation for simple models does not necessary work for deep models

What works for simple models doesn't work for deep models.



gradient-based methods

vulnerable to shattered gradients



Our LRP method is robust to this.

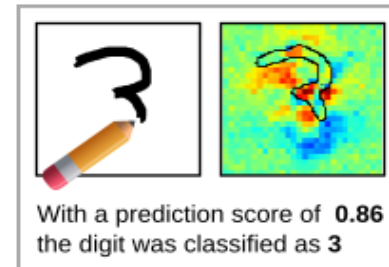
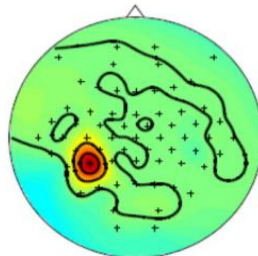
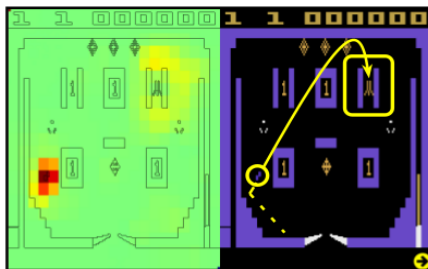
Layer-Wise Relevance Propagation



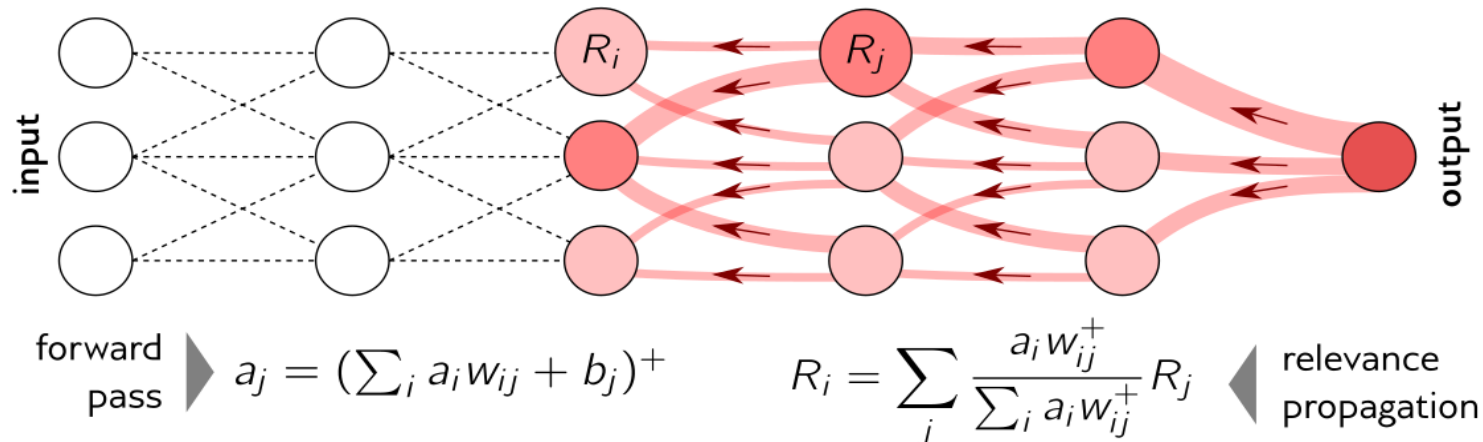
LRP Explanation Framework

e people are more prone to g
The mental part is usually
y is up or down, ie: the Shu
ointed towards Earth, so the
astronauts. About 50% of t
s, and NASA has done numerou

(software, tutorials, demos,
insights, applications)



LRP works 4 all: deep models, LSTMs, kernel methods ...



A Clarification on LRP

LRP \neq Gradient \times Input

... except for special cases. LRP was developed among others because gradient-based methods aren't satisfying.

When comparing with LRP, please use appropriate LRP parameters (Like when comparing different ML techniques).

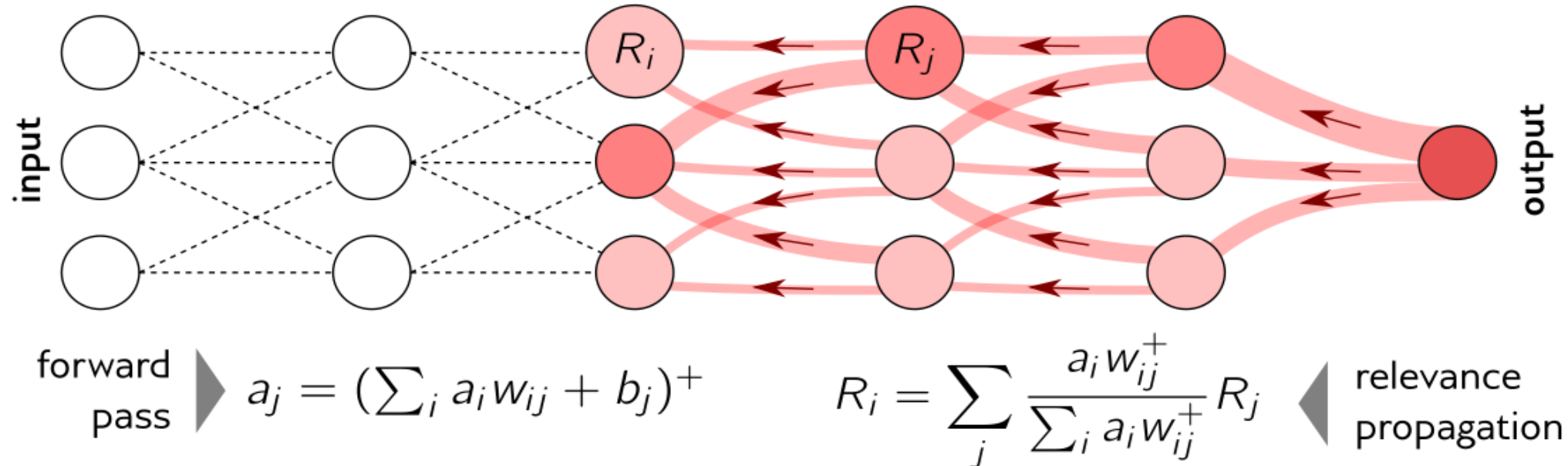
Good news: No need to reimplement LRP, check our software at www.heatmapping.org.

Layer-Wise Relevance Propagation

Robustly and reliably explains complex state-of-the-art deep neural networks.

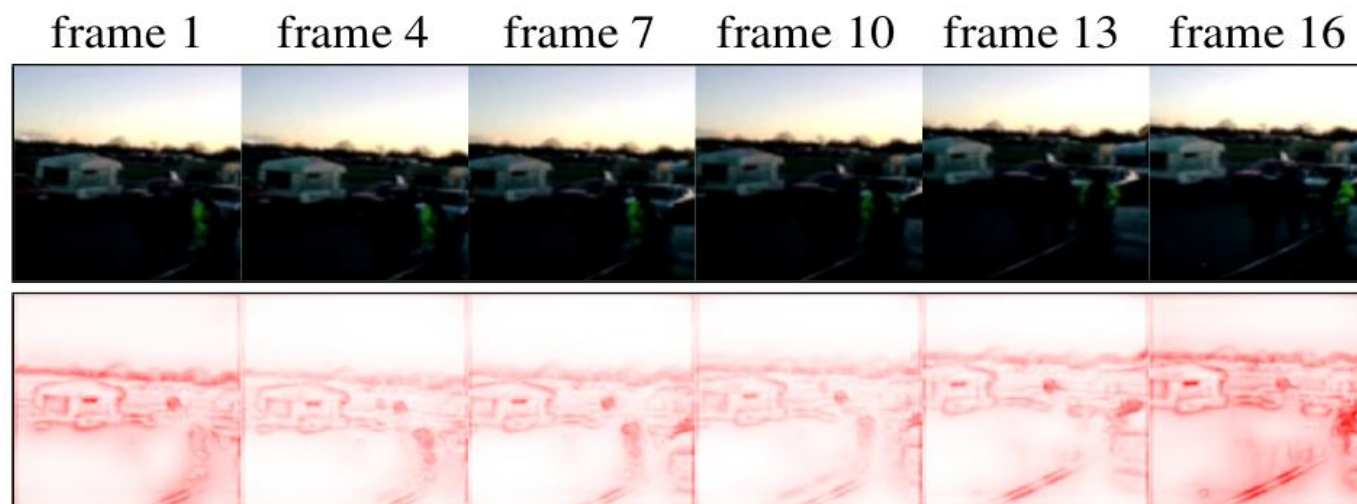
Applicable to general deep networks, but also (kernel) SVMs, LSTMs, Bag-of-words classifiers.

Rules can be engineered to enforce desirable properties or derived from a theoretical principle (deep Taylor decomposition).



Explanations can be evaluated:
Pixel flipping (model agnostic)
And beyond LRP and DTD

Explanation helps to improve models



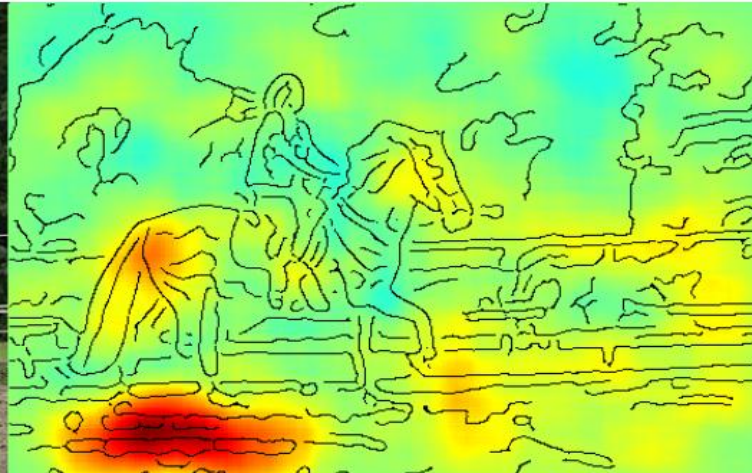
Explaining ML, Now What?

Explanation helps to find flaws in models

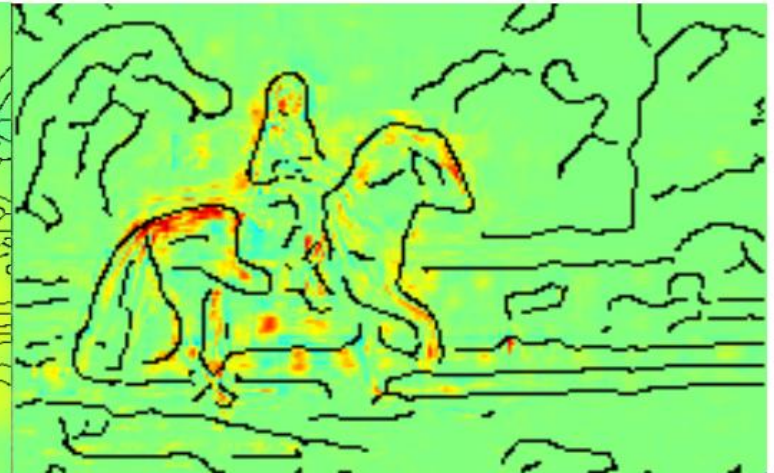
Image



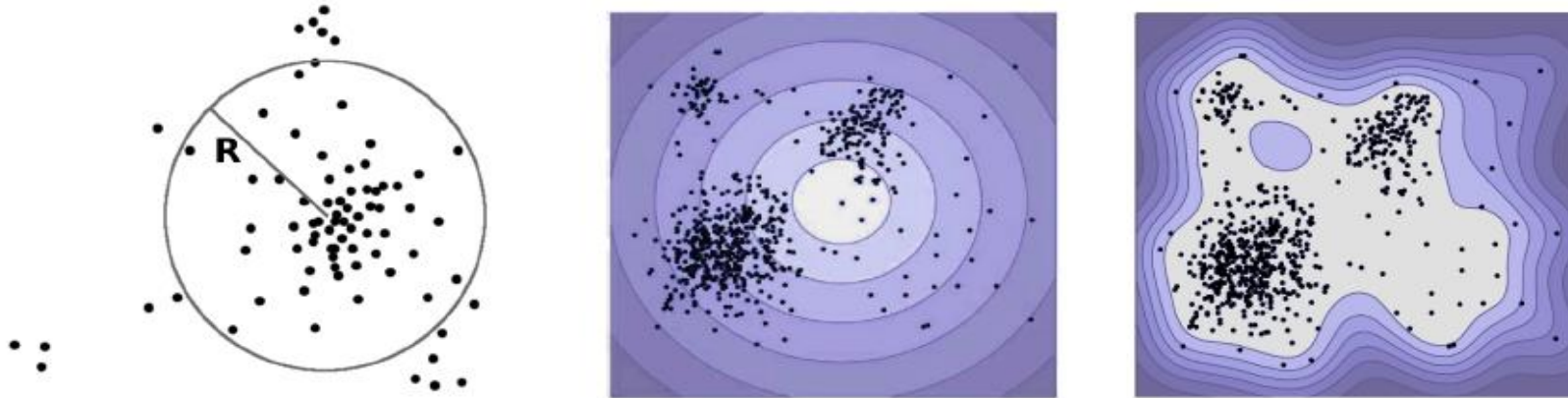
FV



DNN



Support Vector Data description



Support Vector Data Description (SVDD)

- Compute minimal enclosing sphere with center \mathbf{c} and radius R
- Anomaly score as the distance to center \mathbf{c} , that is $f(\mathbf{x}) = \|\phi(\mathbf{x}) - \mathbf{c}\|$
- Accept data point \mathbf{x} if $f(\mathbf{x}) \leq R$ and ...
... reject \mathbf{x} if $f(\mathbf{x}) > R$

Explaining one-class

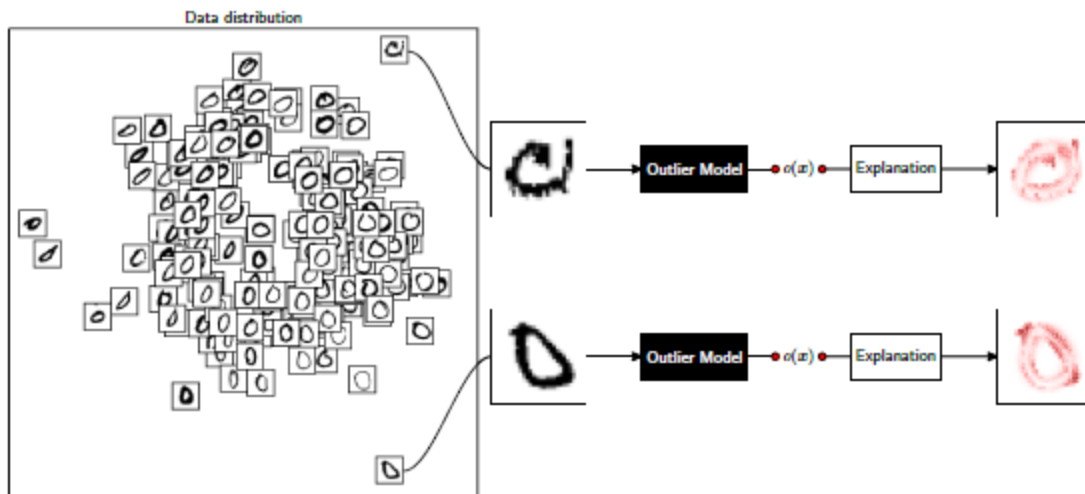


Figure 1: Illustration of the outlier detection and explanation setting. *Left:* Data is generated from an unknown distribution, we are for example interested in potential outliers; *Middle:* Unsupervised machine learning techniques estimate the data generating distribution and assign an outlier score $o(x)$ to unlikely data points; *Right:* Our explanation method assigns a relevance score to every input variable that reflects the contribution of input variable x_i to the model decision. We apply dithering to all heatmaps for printing reliability.

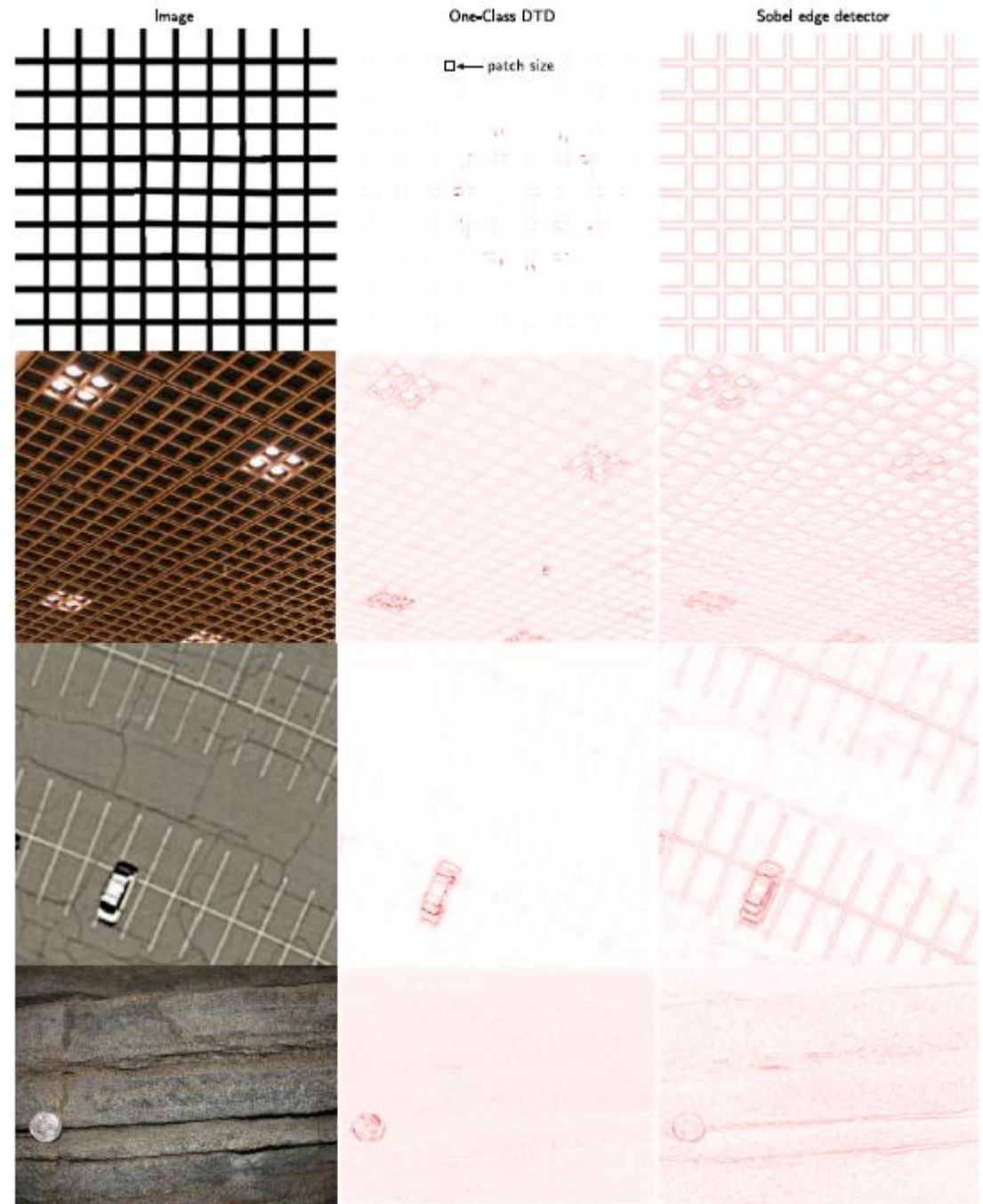


Figure 5: A One-Class SVM is trained on small 7×7 patches of the very image itself. Parameter $\nu = 0.1$ is set to allow at most 10% outliers. Images from a texture data set [11] (row one, two and four) and PatternNet [61]; top image is altered by us. For every image, we show *Left:* input image; *Middle:* decomposition of one-class SVM; *Right:* Sobel filter for reference. All images were resized to 256 pixels width.

Getting **new** Insights in the Sciences

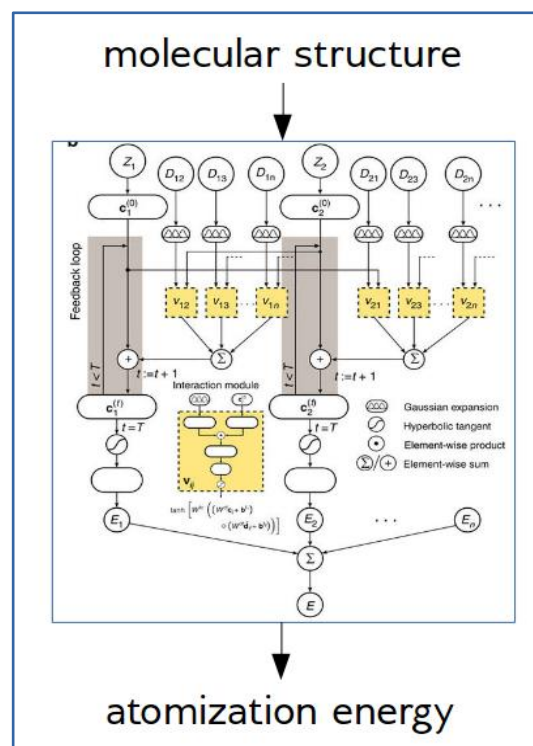
Example: Understanding physical systems at the quantum level.

time-independent Schrödinger Equation

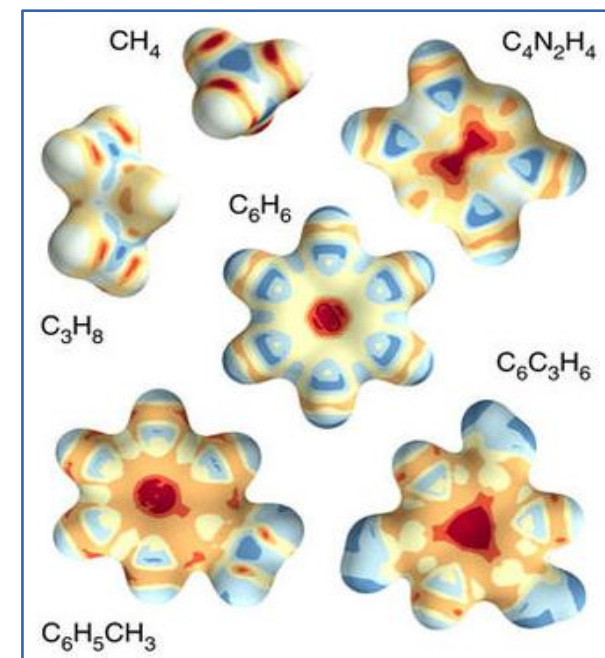
$$\hat{H}\Psi = E\Psi$$

↑ Hamiltonian ↑ energy

equation describing general physical systems



DNN approximation for organic molecules



Interpretation of the trained DNN model

[Schütt et al. Nat Comm. 2017, Schütt et al JCP 2018, Chmiela et al. Sci. Adv. 2017, Chmiela et al Nat Comms 2018...]

Semi-final Conclusion

- explaining & interpreting nonlinear models is essential
- orthogonal to improving DNNs and other models
- need for opening the blackbox ...
- understanding nonlinear models is essential for Sciences & AI
- new **theory**: LRP is based on deep taylor expansion
- when looking at XAI techniques: compare the right thing!
- XAI and WHO & ITU, Regulations etc.
- Note: **even the most complex DL models are explainable nowadays**

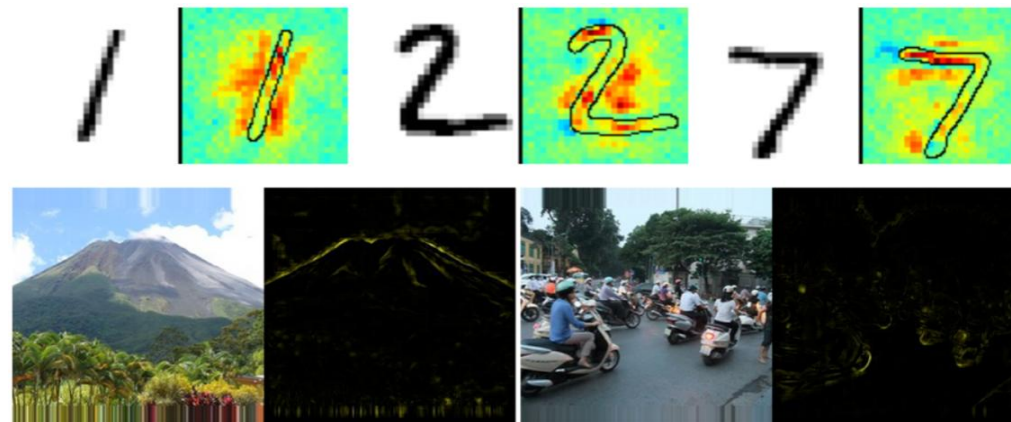
www.heatmapping.org

Thank you for your attention

Visit:

<http://www.heatmapping.org>

- ▶ Tutorials
- ▶ Software
- ▶ Online Demos



Tutorial Paper

Montavon et al., “Methods for interpreting and understanding deep neural networks”, Digital Signal Processing, 73:1-5, 2018

Keras Explanation Toolbox

<https://github.com/albermax/investigate>

State-of-the-Art
Survey

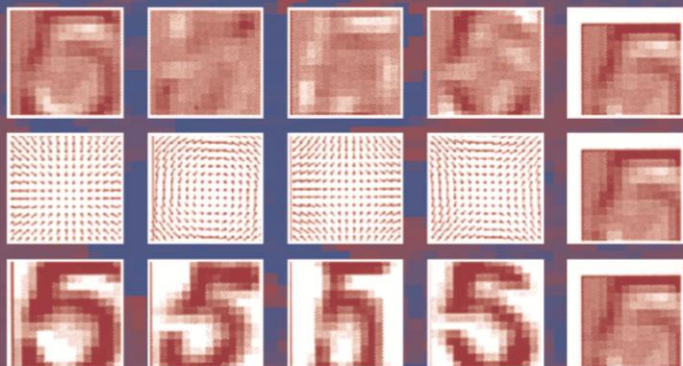
Grégoire Montavon
Genevieve B. Orr
Klaus-Robert Müller (Eds.)

LNCS 7700

Neural Networks: Tricks of the Trade

Second Edition

RELOADED



 Springer



Toward Brain-Computer Interfacing

edited by
Guido Dornhege, José del R. Millán,
Thilo Hinterberger, Dennis J. McFarland,
and Klaus-Robert Müller

foreword by Terrence J. Sejnowski

The background features a light blue world map with binary code (0s and 1s) overlaid on it. The text 'BBDCC' is prominently displayed in the center.

BBDCC

BERLIN **BIG**
DATA CENTER



Further Reading I

- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.
- Bach, S., Binder, A., Montavon, G., Müller, K.-R. & Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11, 1803-1831.
- Brockherde, F., Vogt, L., Li, L., Tuckerman, M., Burke, K., Müller, K. R., By-passing the Kohn-Sham Equations with machine learning, *Nature Communications*, 8:872 (2017)
- Blum, L. C., & Reymond, J. L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131(25), 8732-8733.
- Braun, M. L., Buhmann, J. M., & Müller, K. R. (2008). On relevant dimensions in kernel feature spaces. *The Journal of Machine Learning Research*, 9, 1875-1908
- Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., & Müller, K. R. (2017). Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5), e1603015.
- Chmiela, S., Sauceda, HE, Müller, K. R., Tkatchenko, A. (2018). Towards exact molecular dynamics simulations with machine-learned force fields. *Nat Commun*. 2018 Sep 24;9(1):3887
- Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, A.O., Tkatchenko, A., and Müller, K.-R. "Assessment and validation of machine learning methods for predicting molecular atomization energies." *Journal of Chemical Theory and Computation* 9, no. 8 (2013): 3404-3419.
- Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K. R., & Tkatchenko, A. (2015). Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space, *J. Phys. Chem. Lett.* 6, 2326–2331.

Further Reading II

- Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12(2), 181-201.
- Montavon, G., Braun, M. L., & Müller, K. R. (2011). Kernel analysis of deep networks. *The Journal of Machine Learning Research*, 12, 2563-2581.
- Montavon, Grégoire, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V. Lilienfeld, and Klaus-Robert Müller. "Learning invariant representations of molecules for atomization energy prediction." In *Advances in Neural Information Processing Systems*, pp. 440-448. 2012.
- Montavon, G., Braun, M., Krueger, T., & Müller, K. R. (2013). Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *IEEE Signal Processing Magazine*, 30(4), 62-74.
- Montavon, G., Orr, G. & Müller, K. R. (2012). *Neural Networks: Tricks of the Trade*, Springer LNCS 7700. Berlin Heidelberg.
- Montavon, Grégoire, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Machine learning of molecular electronic properties in chemical compound space." *New Journal of Physics* 15, no. 9 (2013): 095003.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.R., 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65, pp.211-222.
- Montavon, G., Samek, W. and Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1-15.
- Snyder, J. C., Rupp, M., Hansen, K., Müller, K. R., & Burke, K. Finding density functionals with machine learning. *Physical review letters*, 108(25), 253002. 2012.

Further Reading III

- Pozun, Z. D., Hansen, K., Sheppard, D., Rupp, M., Müller, K. R., & Henkelman, G., Optimizing transition states via kernel-based machine learning. *The Journal of chemical physics*, 136(17), 174101. 2012 .
- K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties *Phys. Rev. B* 89, 205118 (2014)
- K.T. Schütt, F. Arbabzadah, S. Chmiela, K.R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nature Communications* 8, 13890 (2017)
- Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. *Machine learning*, 42(3), 287-320.
- Rupp, M., Tkatchenko, A., Müller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5), 058301.
- Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5), 1299-1319.
- Smola, A. J., Schölkopf, B., & Müller, K. R. (1998). The connection between regularization operators and support vector kernels. *Neural networks*, 11(4), 637-649.
- Schölkopf, B., Mika, S., Burges, C. J., Knirsch, P., Müller, K. R., Rätsch, G., & Smola, A. J. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, 10(5), 1000-1017.
- Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., & Müller, K. R. (2002). A new discriminative kernel from probabilistic models. *Neural Computation*, 14(10), 2397-2414.
- Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., & Müller, K. R. (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9), 799-807.