Machine Learning and AI for the sciences – Towards Understanding

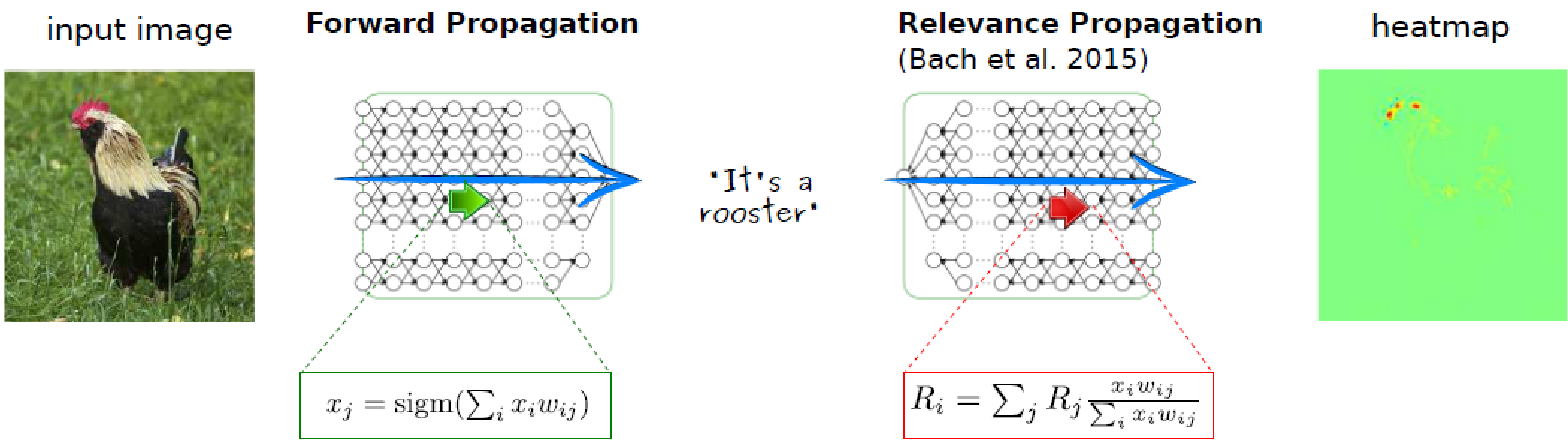
Klaus-Robert Müller !!et al.!!

Outline

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

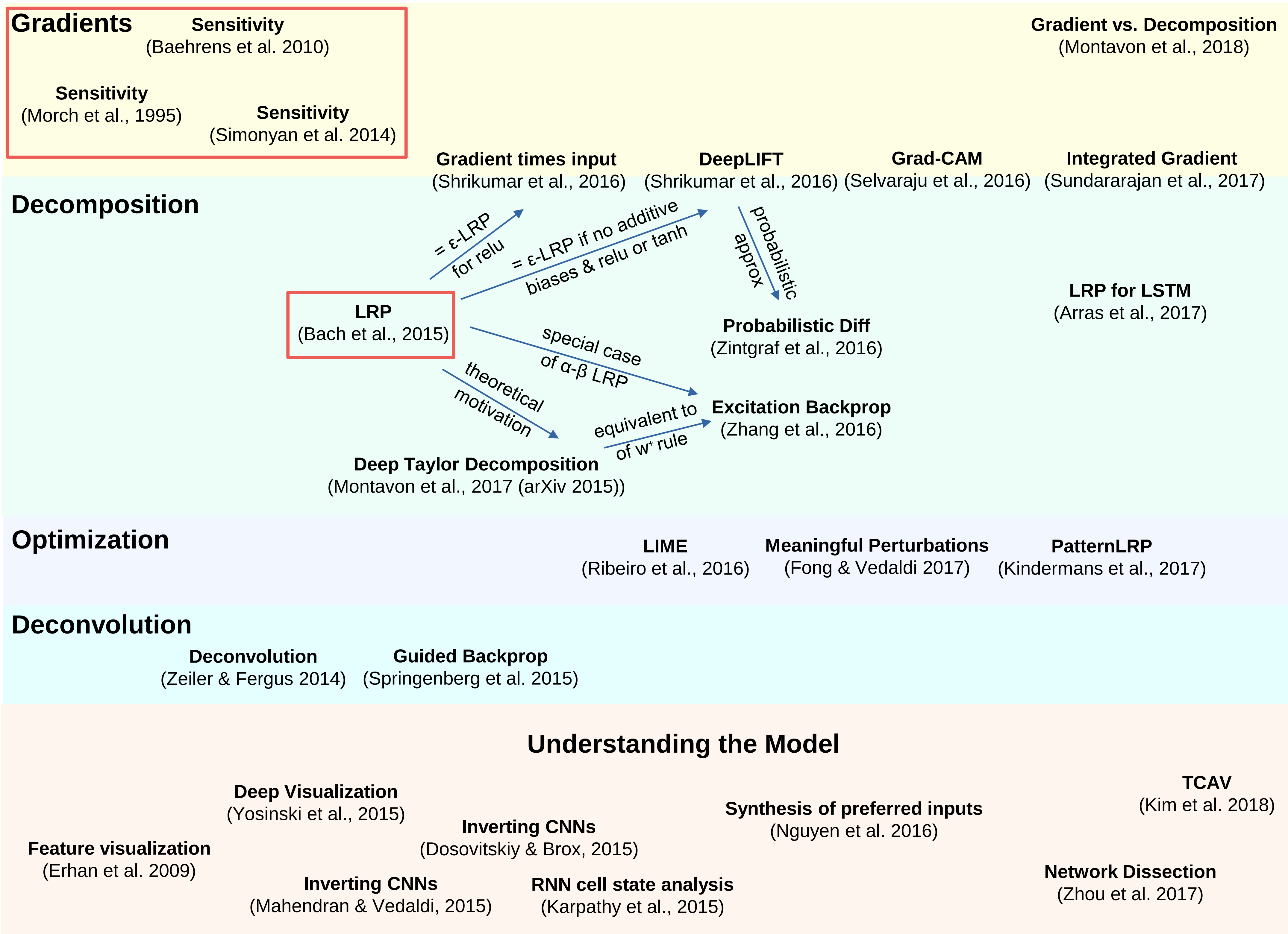
Towards Explaining:
Machine Learning = black box?

Explaining single Predictions Pixel-wise



Goodbye Blackbox ML!

Historical remarks on Explaining Predictors



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**)*

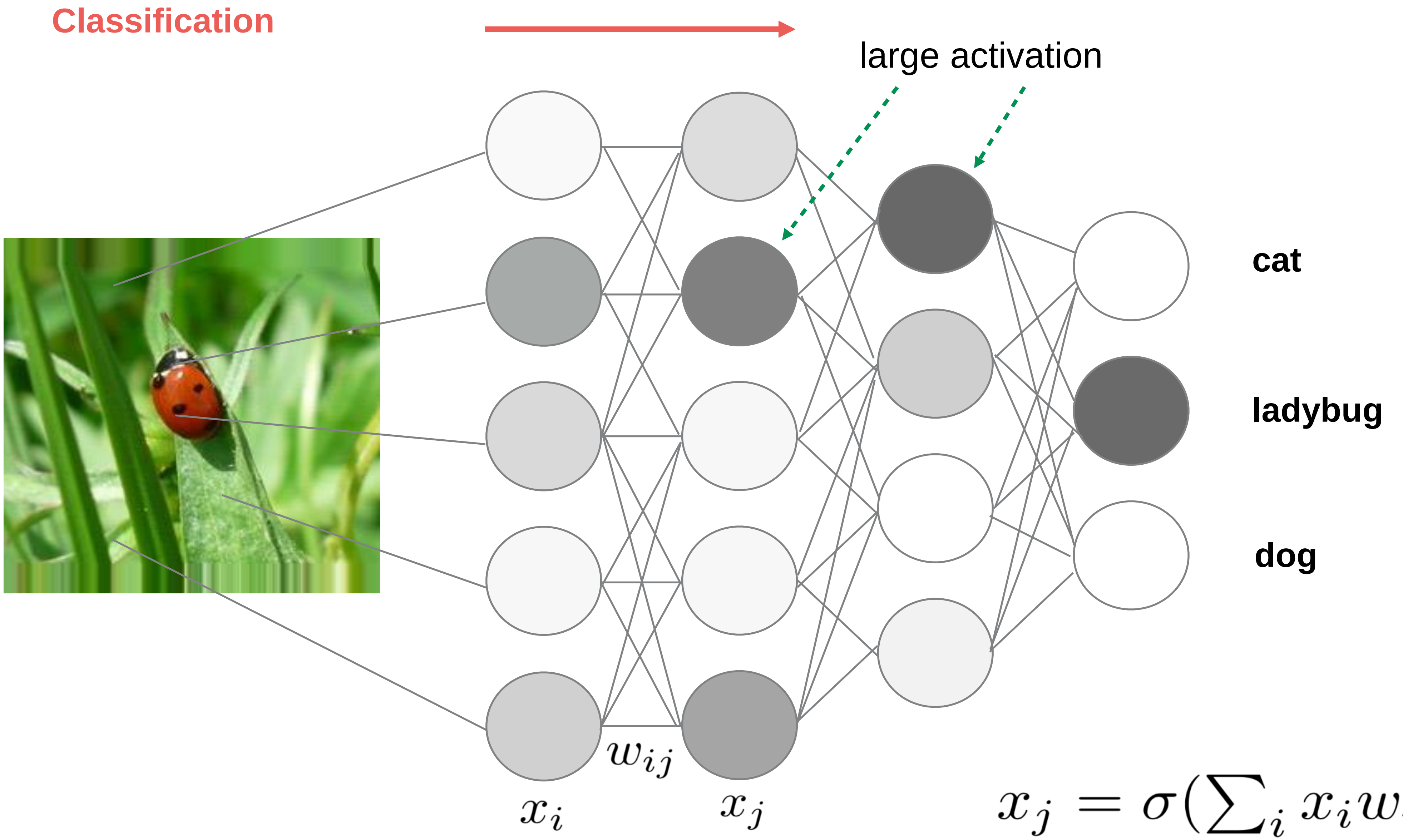
$$h_p = \left|\left| \frac{\partial}{\partial x_p} f(x) \right|\right|_\infty$$

Deconvolution: "Matching input pattern for the classified object in the image" *(**Zeiler & Fergus 2014**)*

(relation to f(x) not specified)

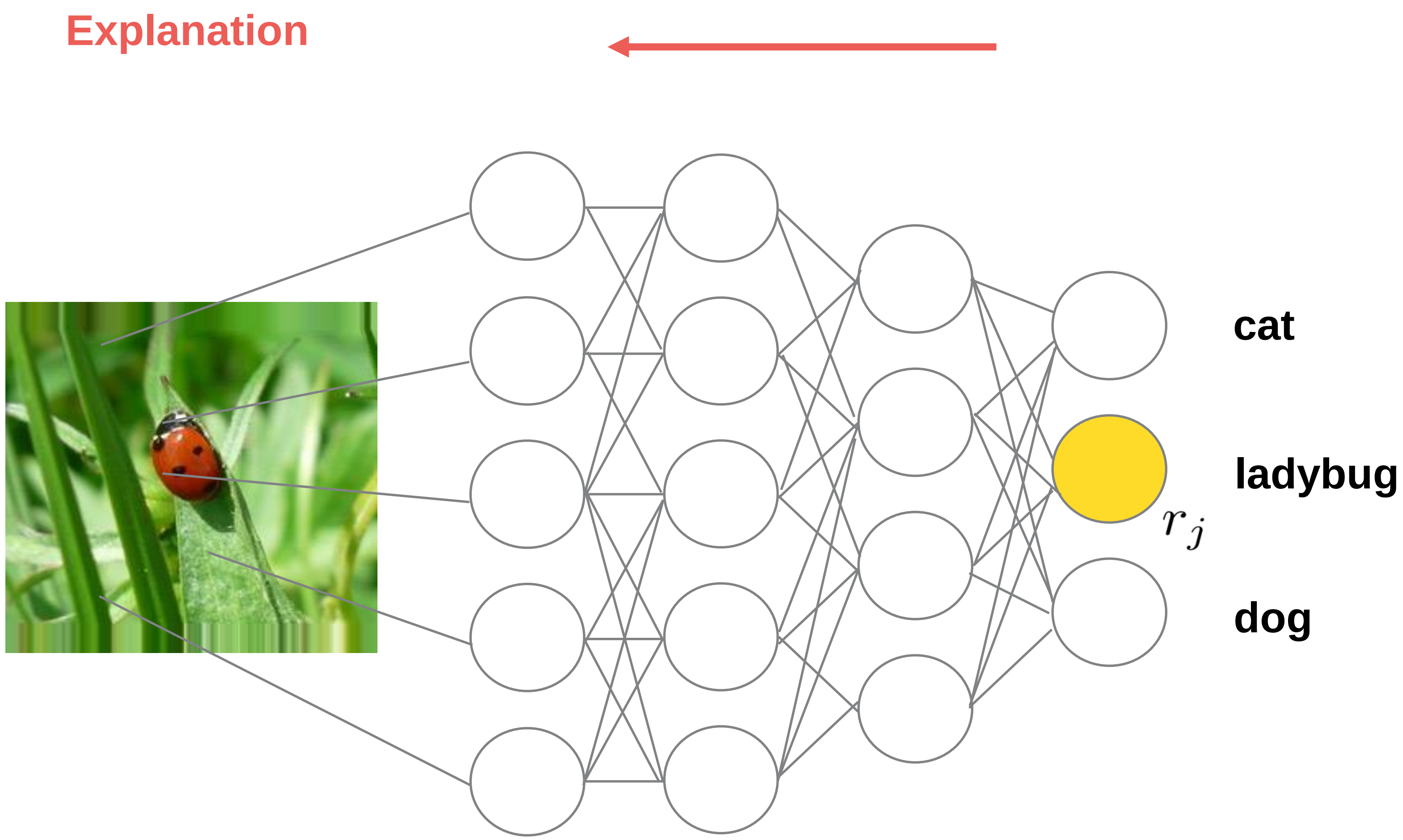
Each method solves a **different** problem!!!

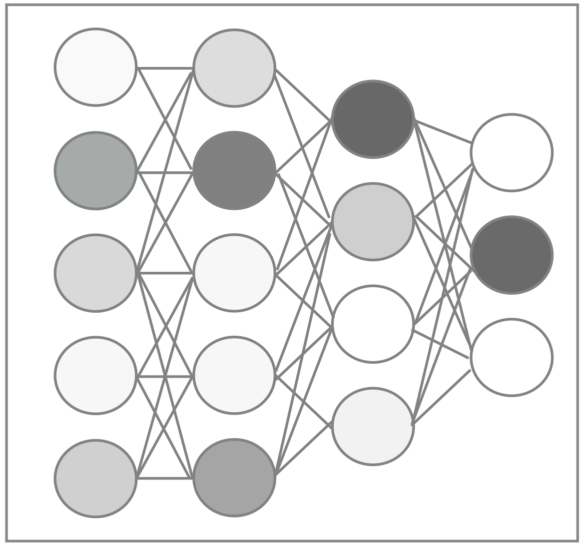
Explaining Neural Network Predictions



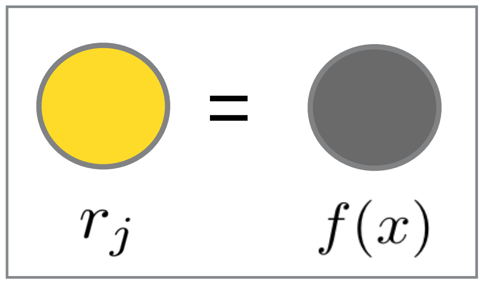
$$x_j = \sigma(\sum_i x_i w_{ij} + b_j)$$

Explaining Neural Network Predictions

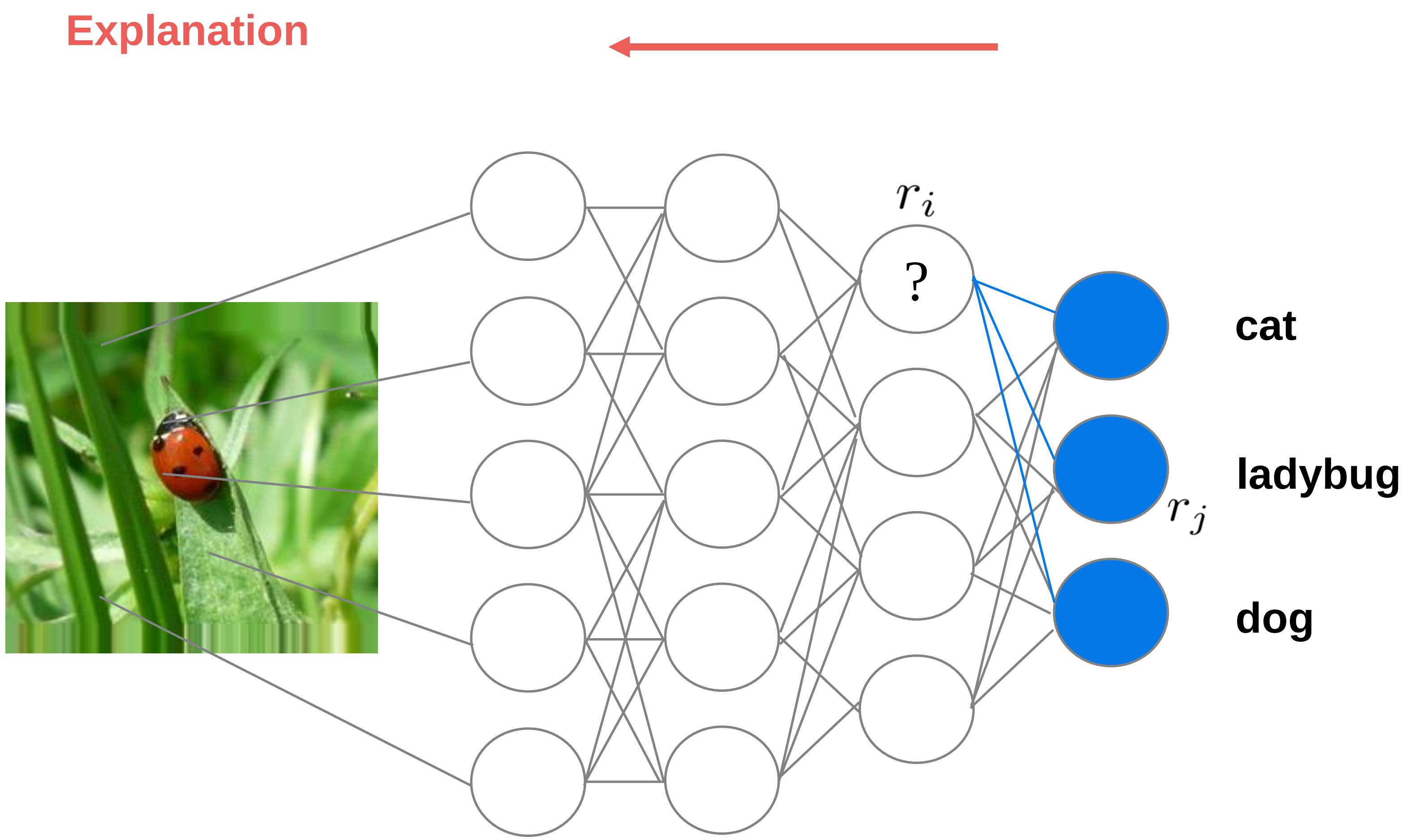


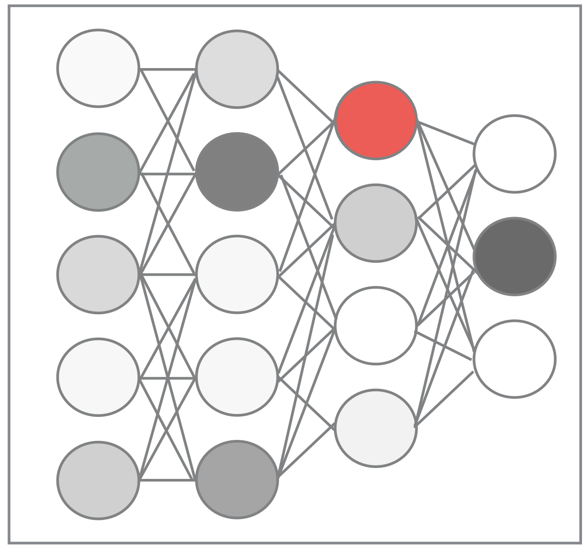


Initialization



Explaining Neural Network Predictions



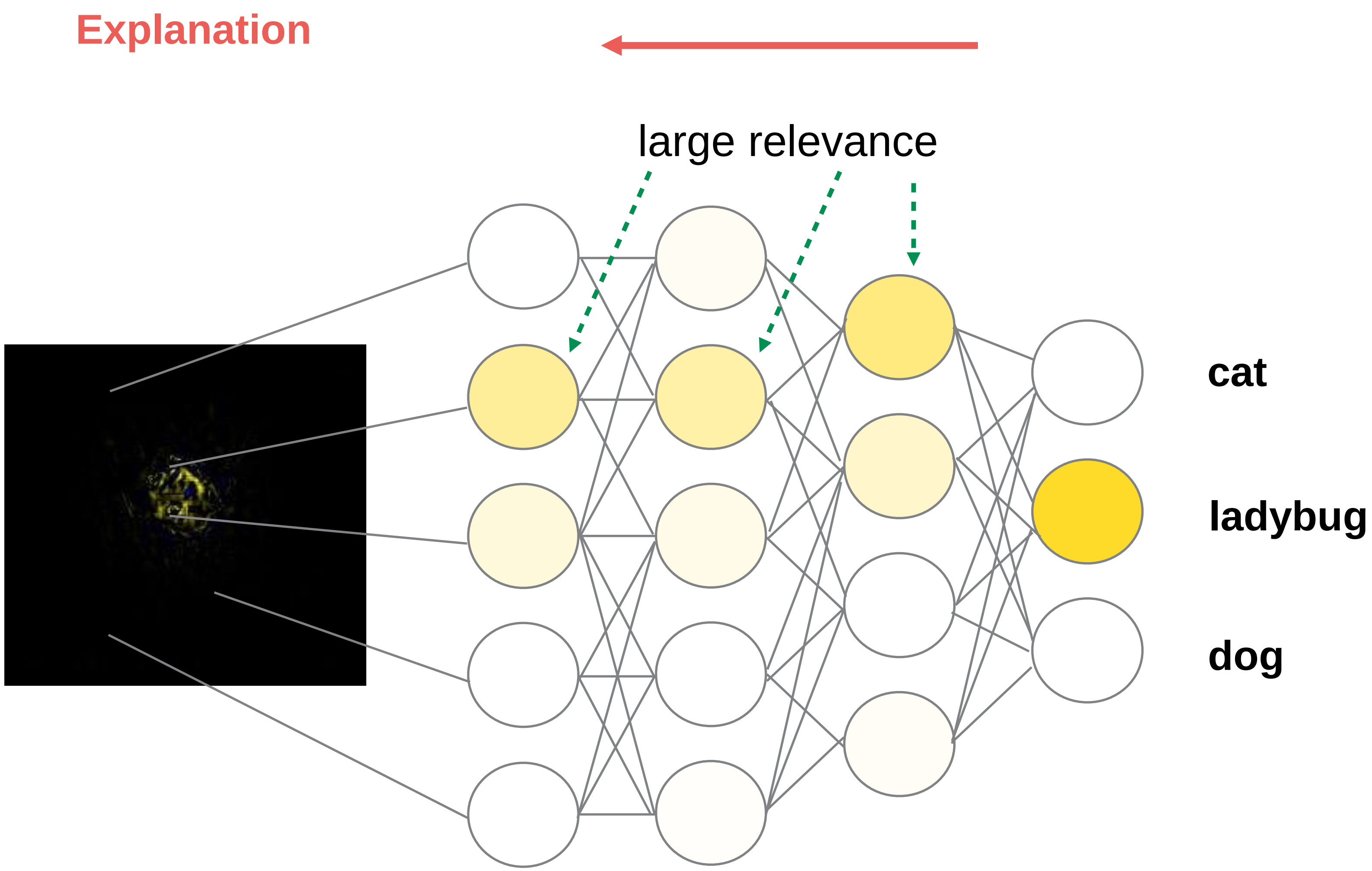


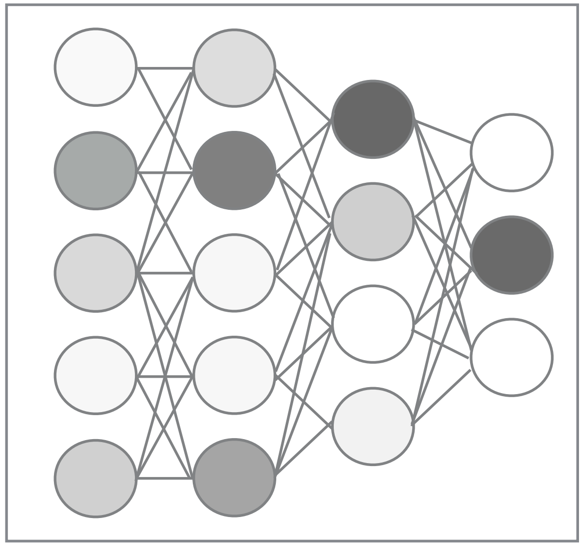
$$r_i = x_i \sum_j \frac{w_{ij} r_j}{\sum_i x_i w_{ij}} = x_i c_i$$

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

Theoretical interpretation
Deep Taylor Decomposition

Explaining Neural Network Predictions

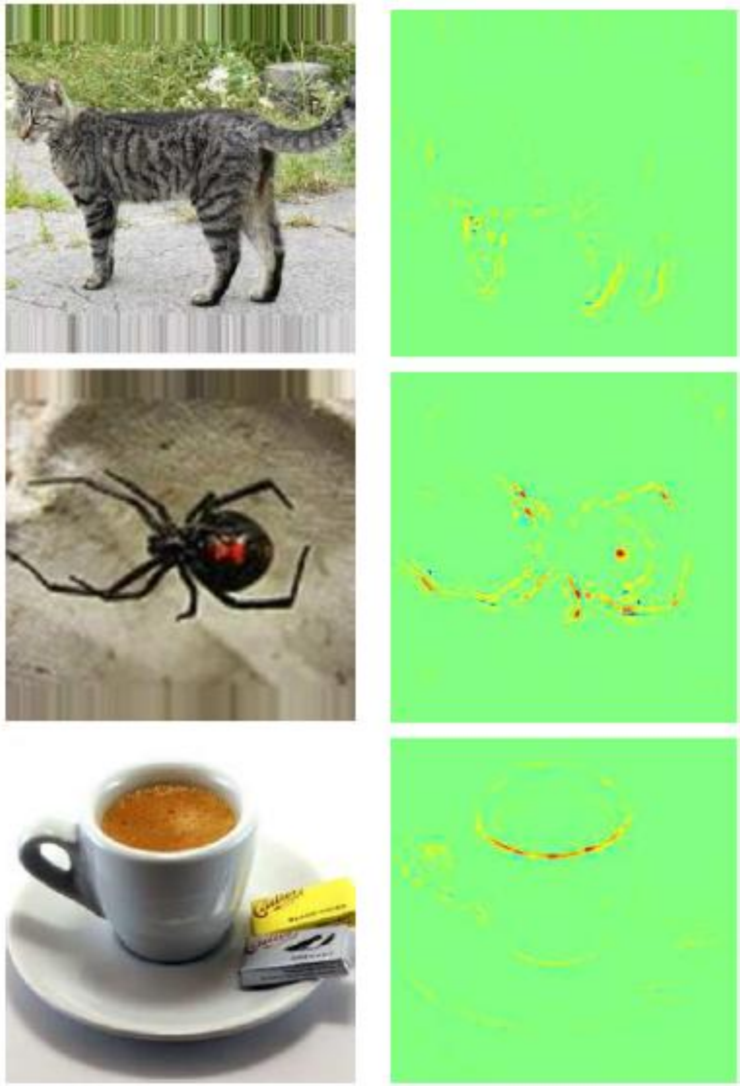




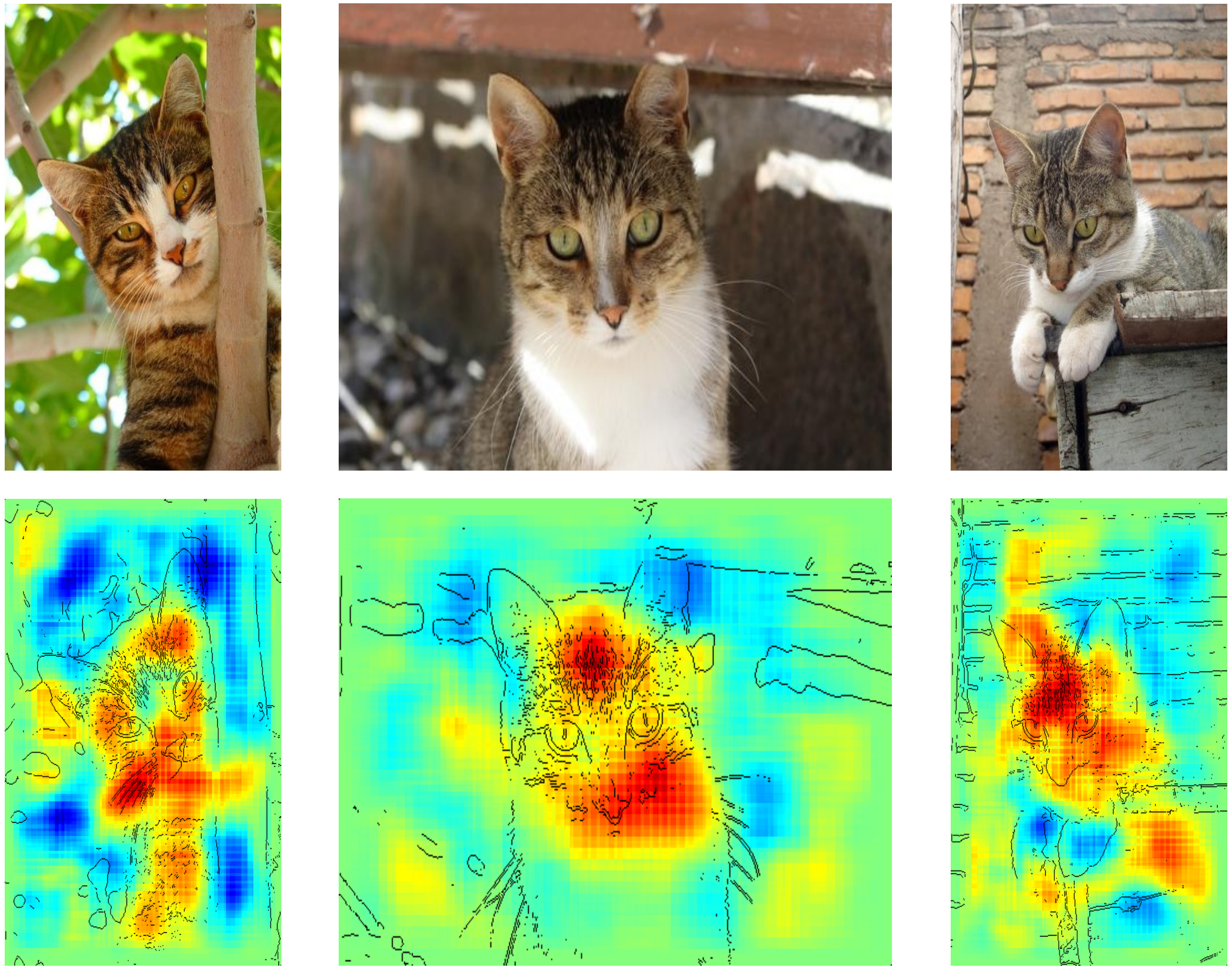
Relevance Conservation Property

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

Explaining Predictions Pixel-wise



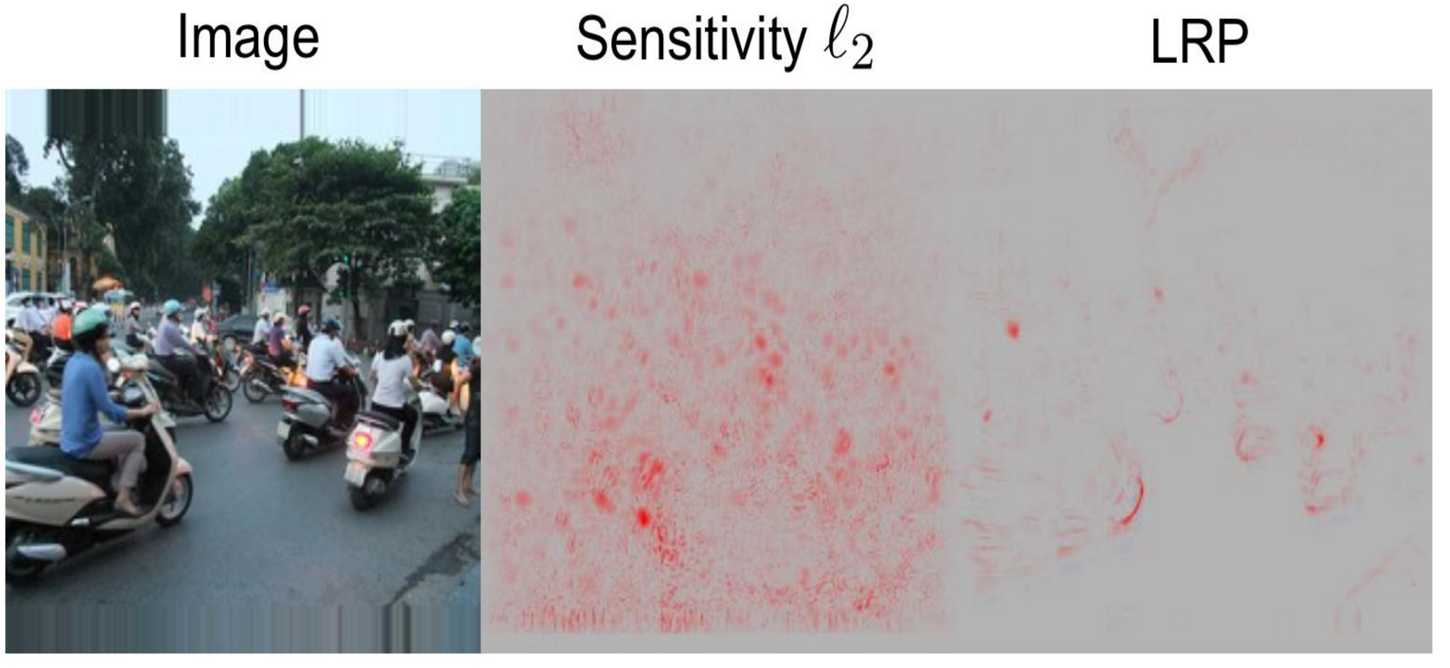
Neural networks



Kernel methods

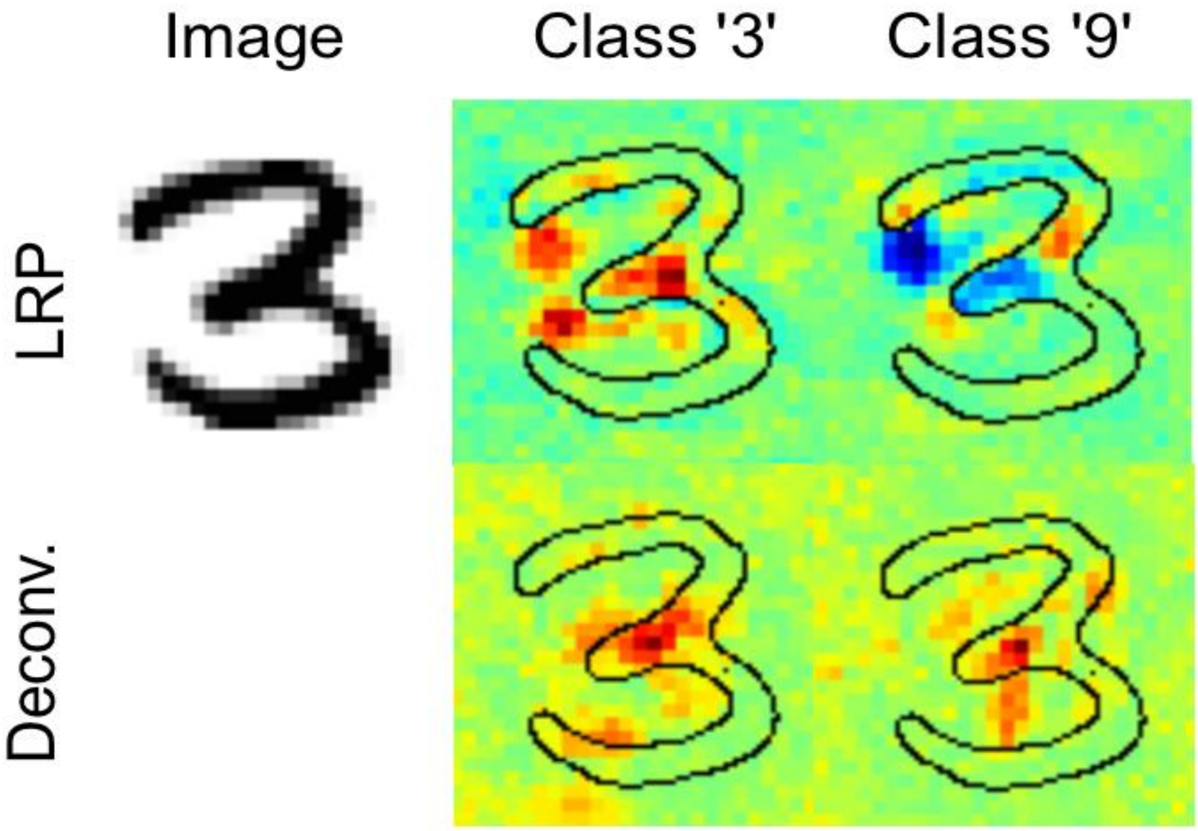
Some Digestion on Explaining

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



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



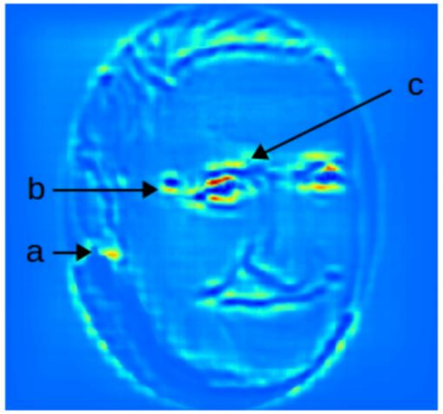
LRP indicates what speaks for class ‘3’ and speaks against class ‘9’

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

Applying Explanation in Vision and Text

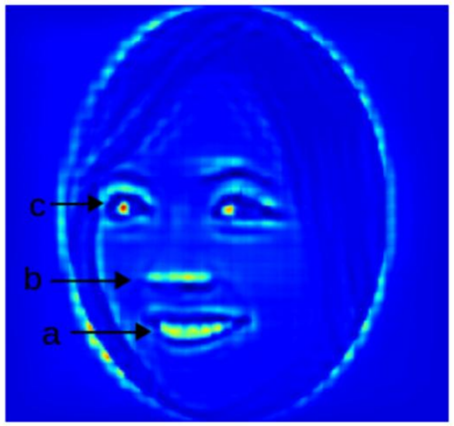
Application: Faces

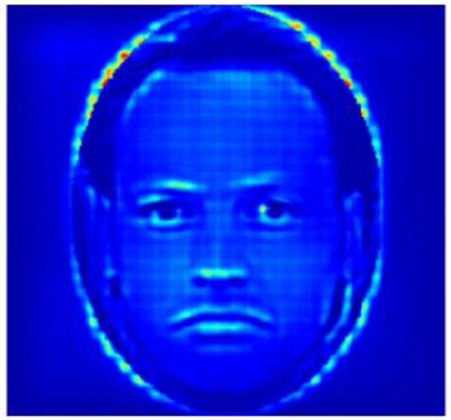
What makes
you look old ?



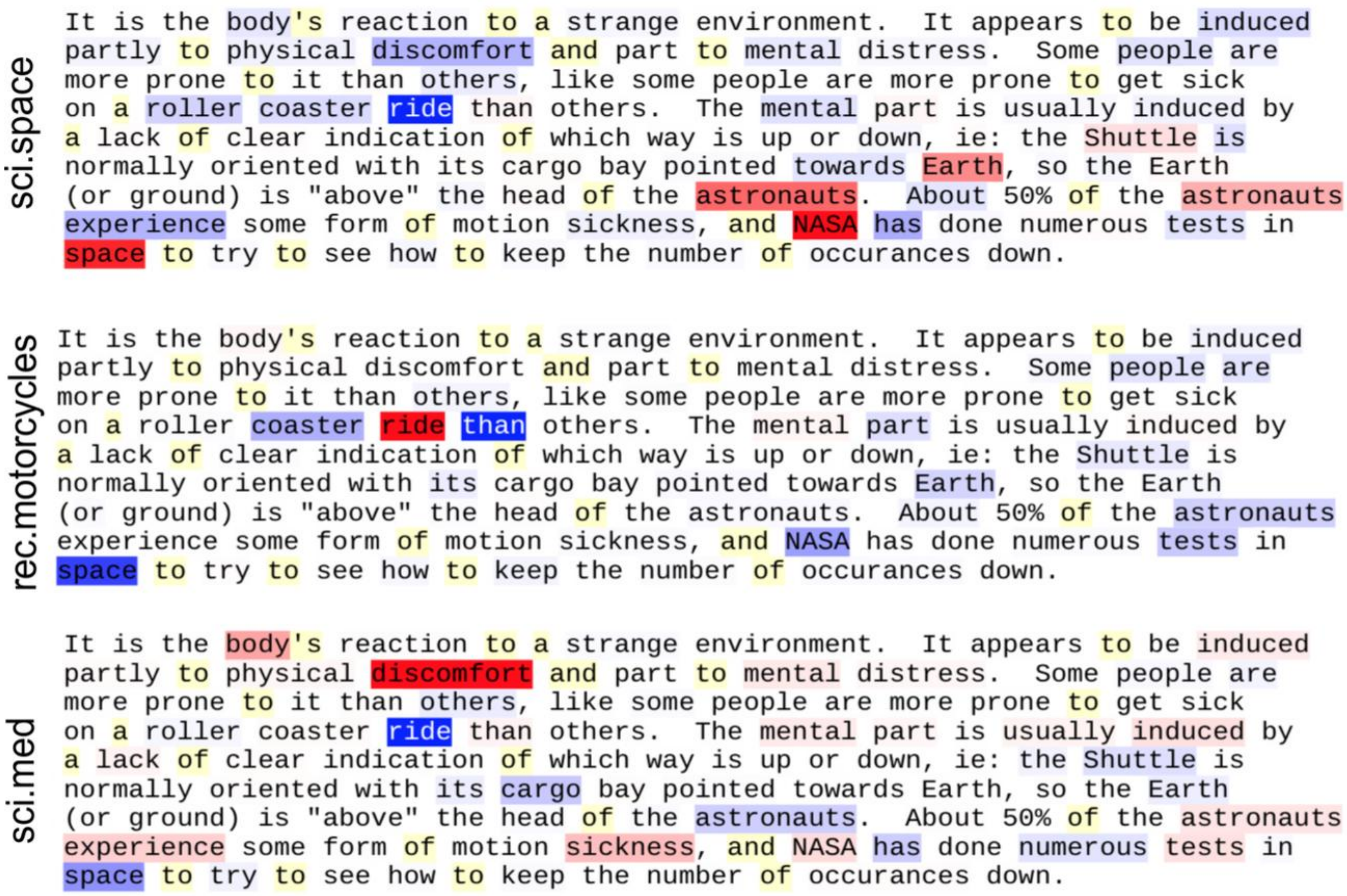
What makes
you look sad ?

What makes
you look attractive ?





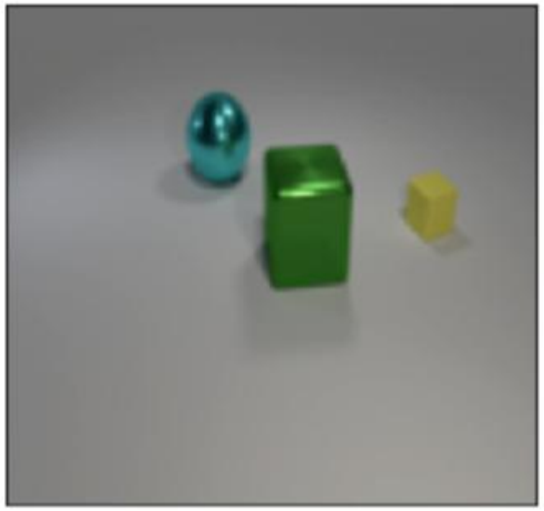
Application: Document Classification

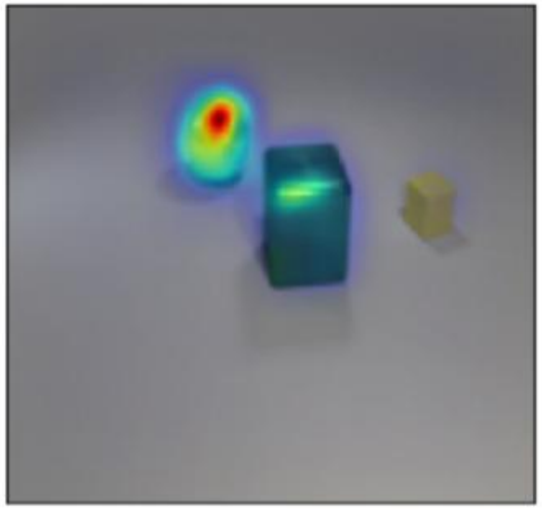


Explaining LSTMs

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 ?



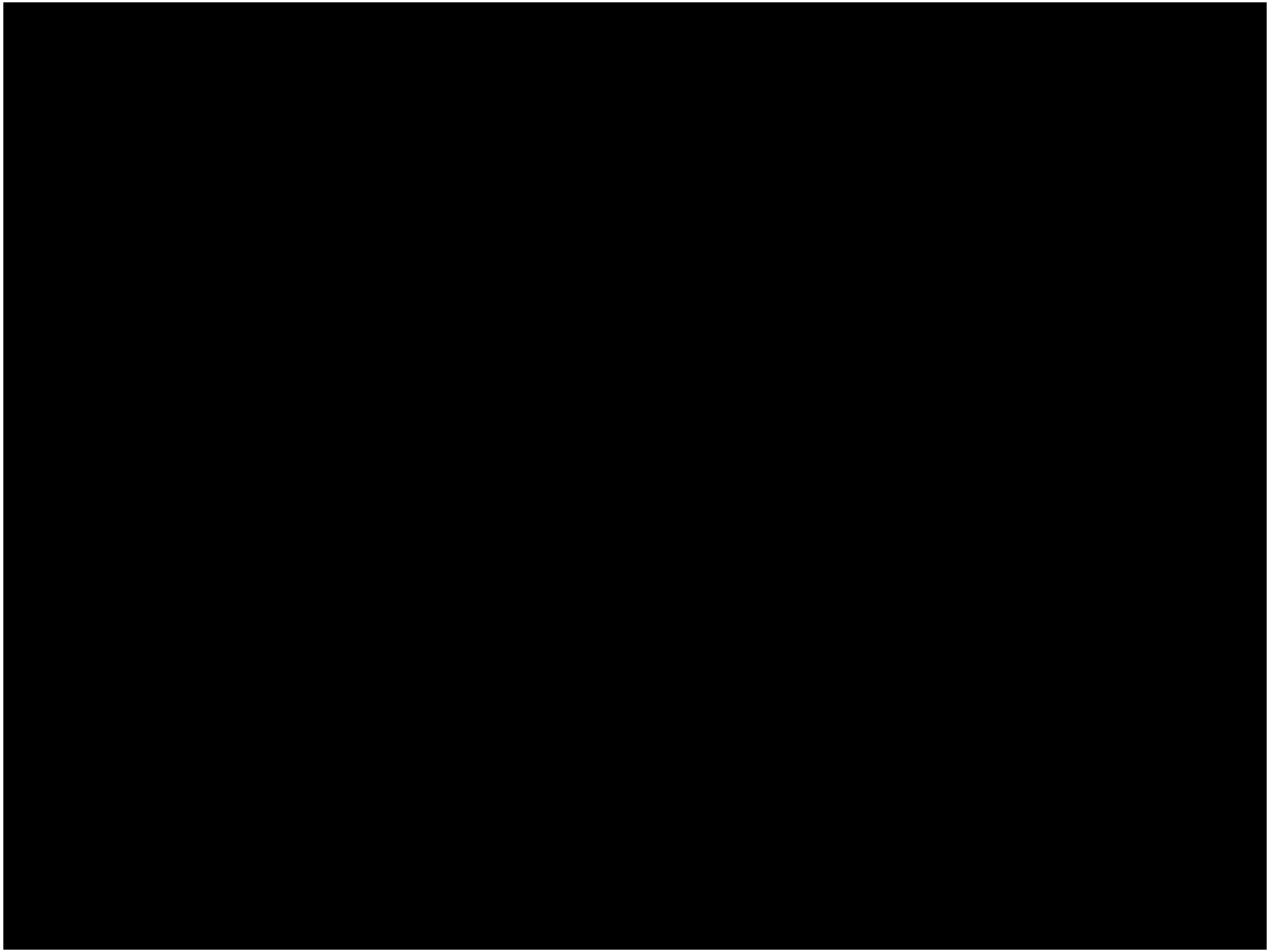


—> model understands the question and correctly identifies the object of interest

(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

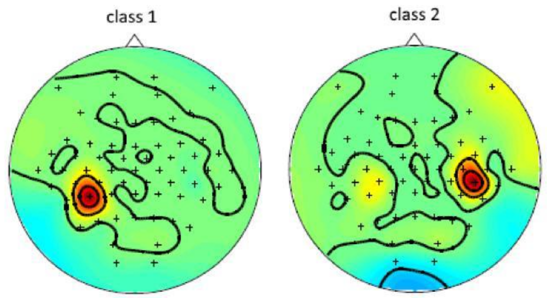
Brain Computer Interfacing: ‚Brain Pong‘

Berlin Brain Computer Ínterface

- 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:

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

instead of

$$\hat{H}(\{Z_I, \mathbf{R}_I\}) \quad \overset{\Psi}{\longmapsto} \quad E$$

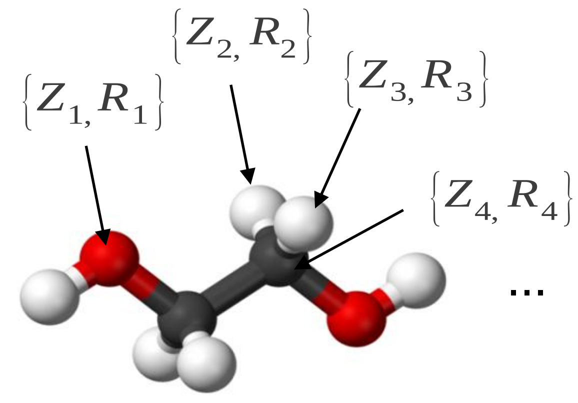
$$\hat{H}\Psi = E\Psi$$

[from von Lilienfeld]

Coulomb representation of molecules

$$M_{ii} = Z_i^{2.4}$$

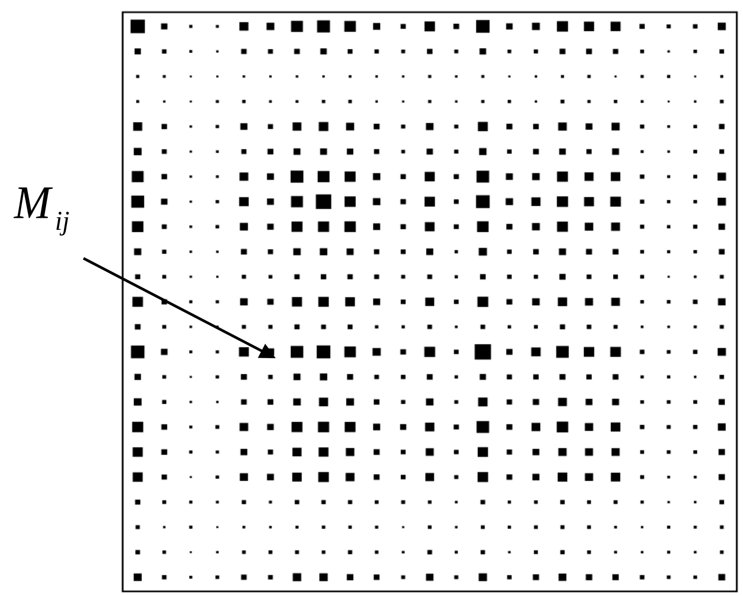
$$M_{ij} = \frac{Z_i Z_j}{\left\| R_i - R_j \right\|}$$



+ phantom atoms

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

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



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 **M** define Gaussian kernel matrix **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

$$\min_{\alpha} \quad \sum_i \left(E^{est}(\mathbf{M}_i) - E_i^{ref}\right)^2 + \lambda \sum_i \alpha_i^2$$

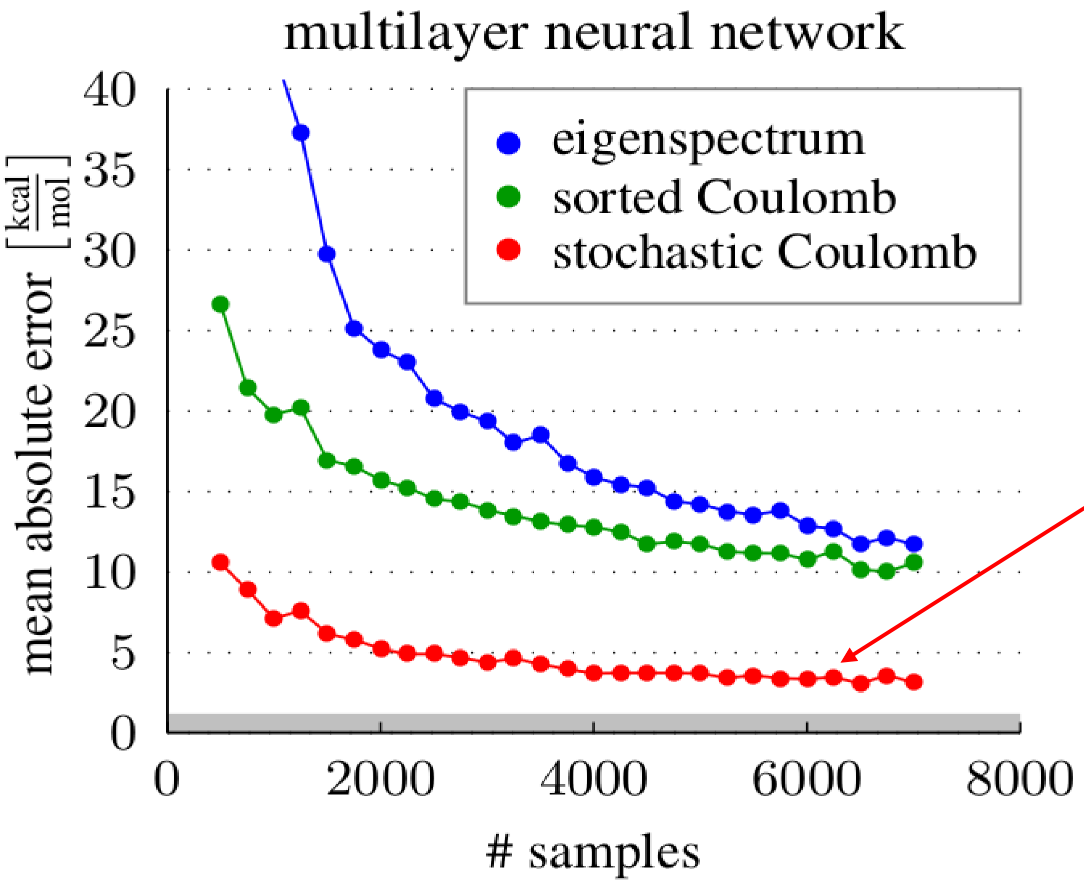
$$\alpha = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{E}^{ref}$$

Exact solution

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

[from von Lilienfeld]

Predicting Energy of small molecules: Results



March 2012
Rupp et al., PRL
9.99 kcal/mol
(kernels + eigenspectrum)

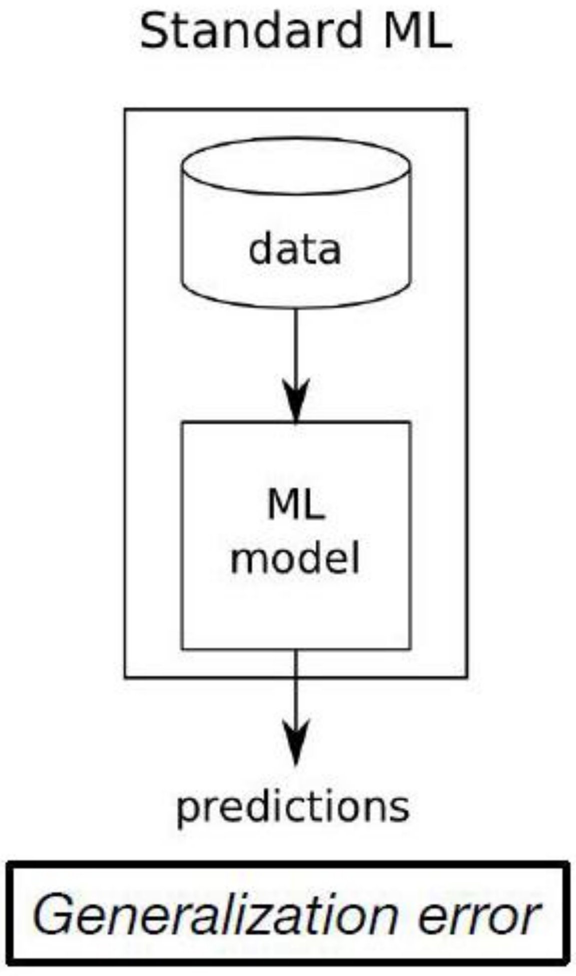
December 2012
Montavon et al., NIPS
3.51 kcal/mol
(Neural nets + Coulomb sets)

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

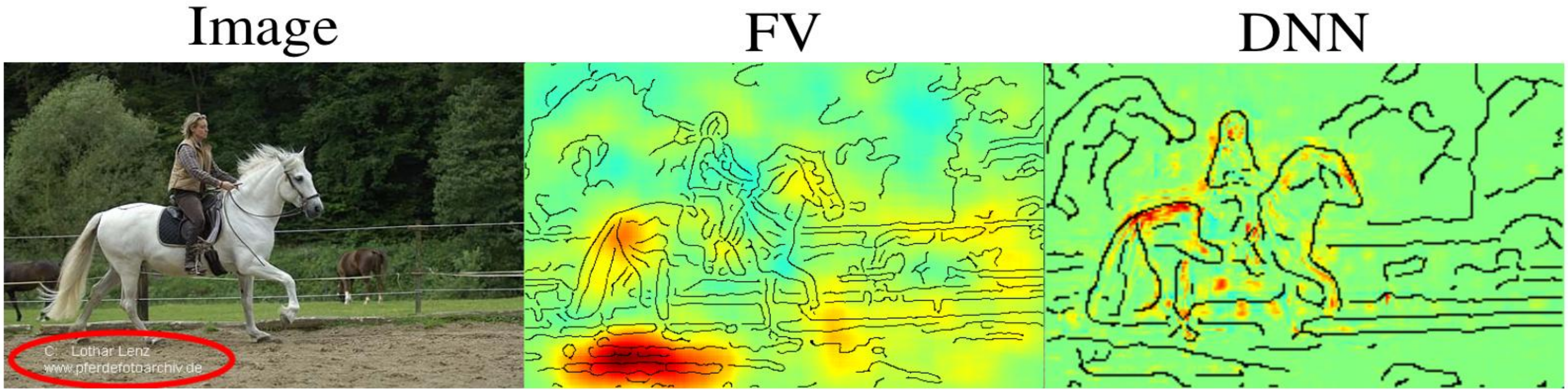
Is the Generalization Error all we need?



Application: Comparing Classifiers (Lapuschkin et al CVPR 2016)

Test error for various classes:

	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%



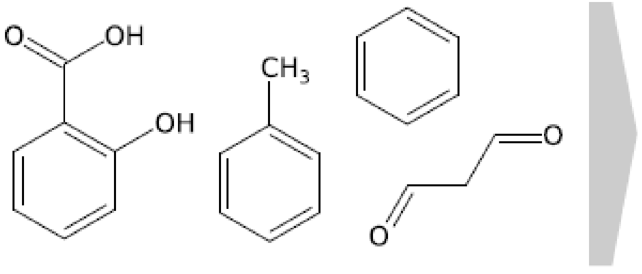
Learning Atomistic Representations with Deep Tensor Neural Networks

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



$$Z = [\, Z_1 \quad Z_2 \quad \cdots \quad Z_n \,]$$

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{12} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Embedding of based on atom types

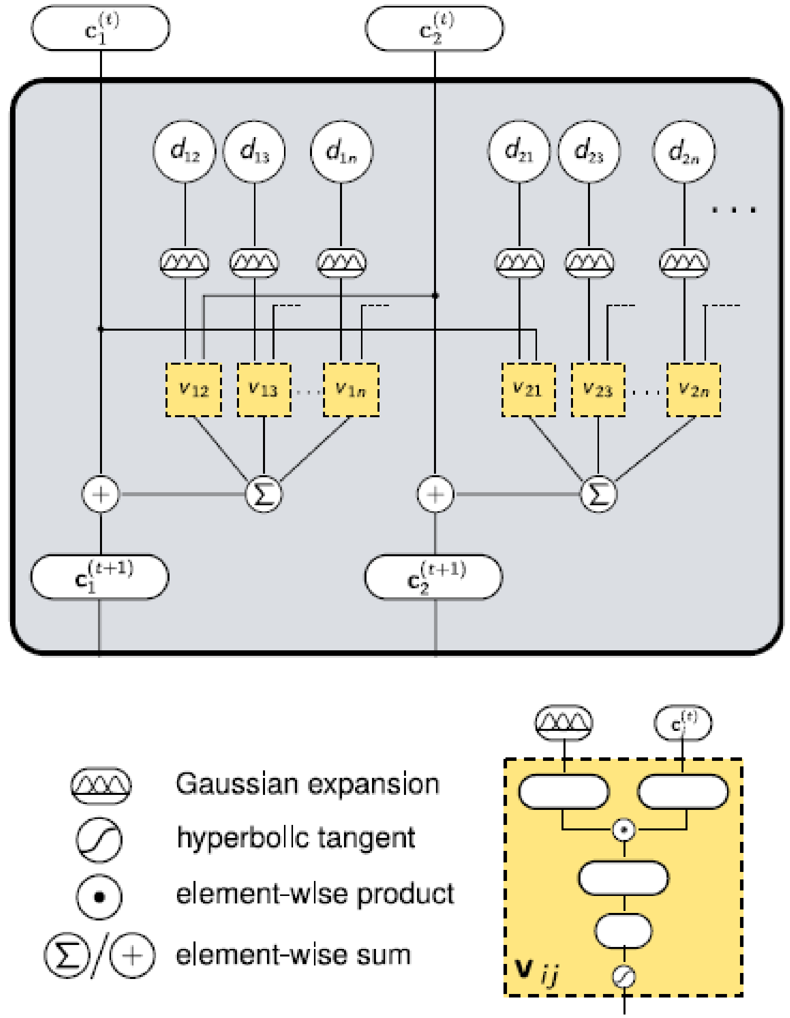
$$\mathbf{x}_i^{(0)} = \mathbf{x}_{Z_i} \in \mathbb{R}^d$$

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

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

Prediction via atom-wise contributions:

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

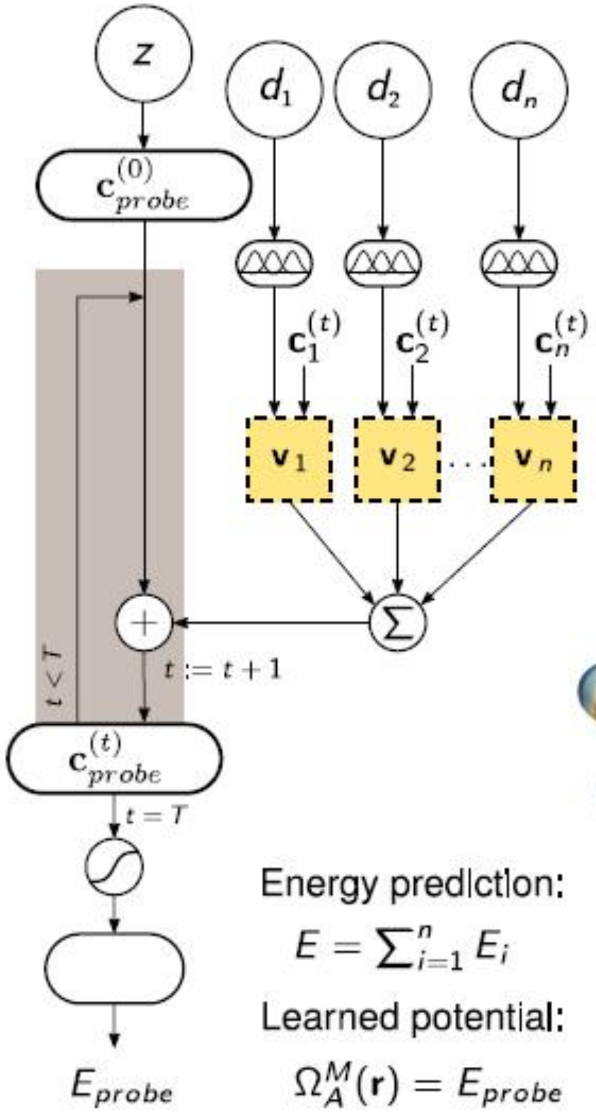


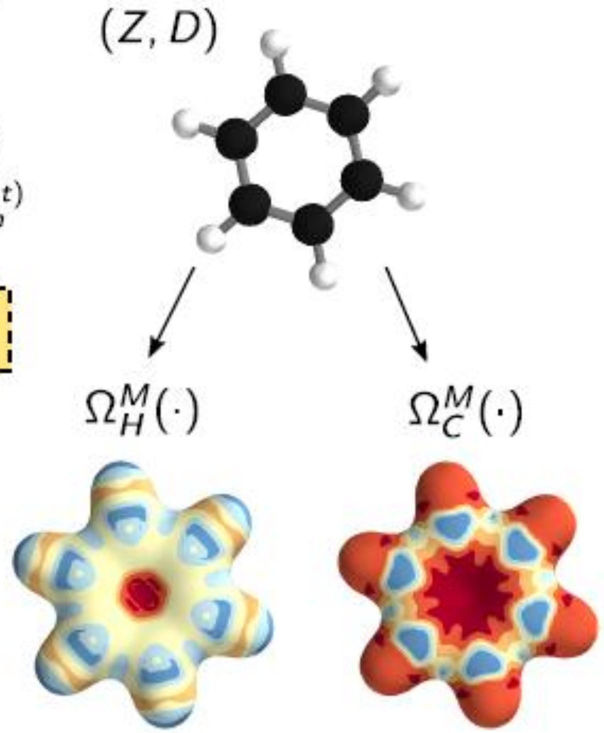
$$\tanh\left(W^{fc}((W^{cf}\mathbf{c}_j + \mathbf{b}^{f_1}) \circ (W^{df}\hat{\mathbf{d}}_{ij} + \mathbf{b}^{f_2}))\right)$$

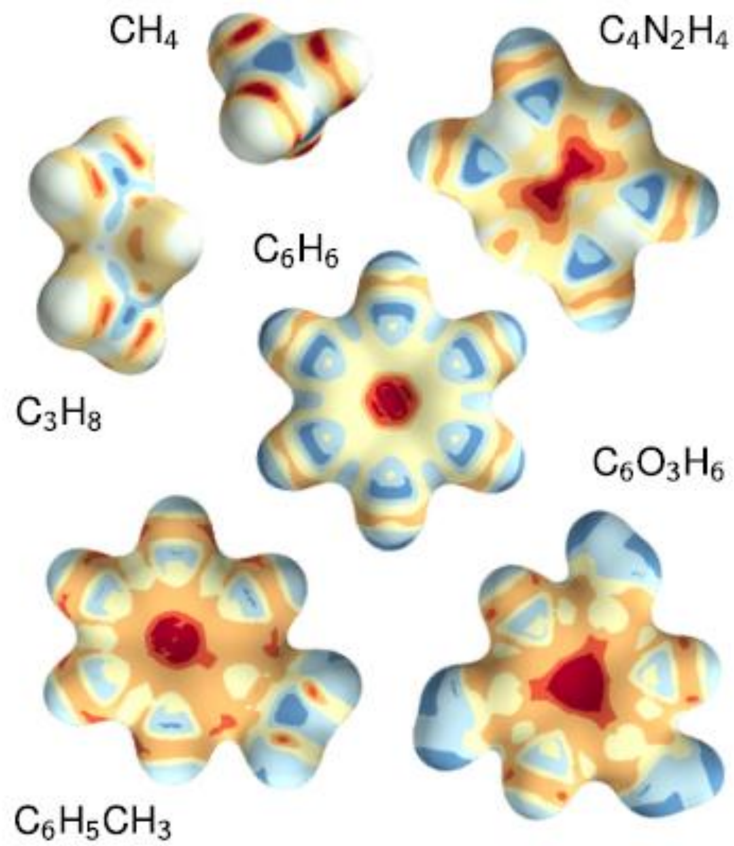
Schütt, Arbabzadah, Chmiela, Müller, Tkatchenko, Nature Communications **8**, *13890 (2017)*

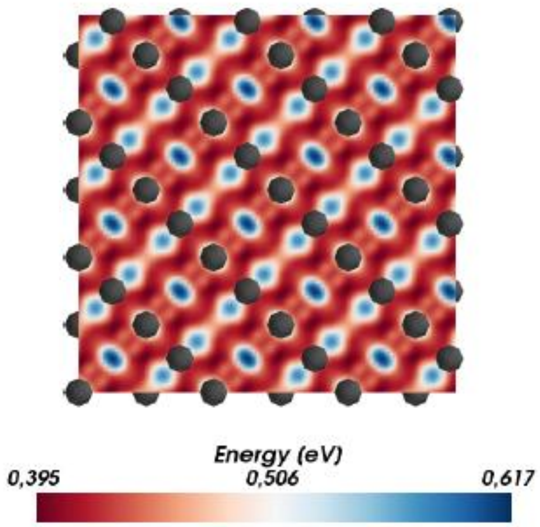
Gaining insights for Physics

Toward Quantum Chemical Insights: supervised







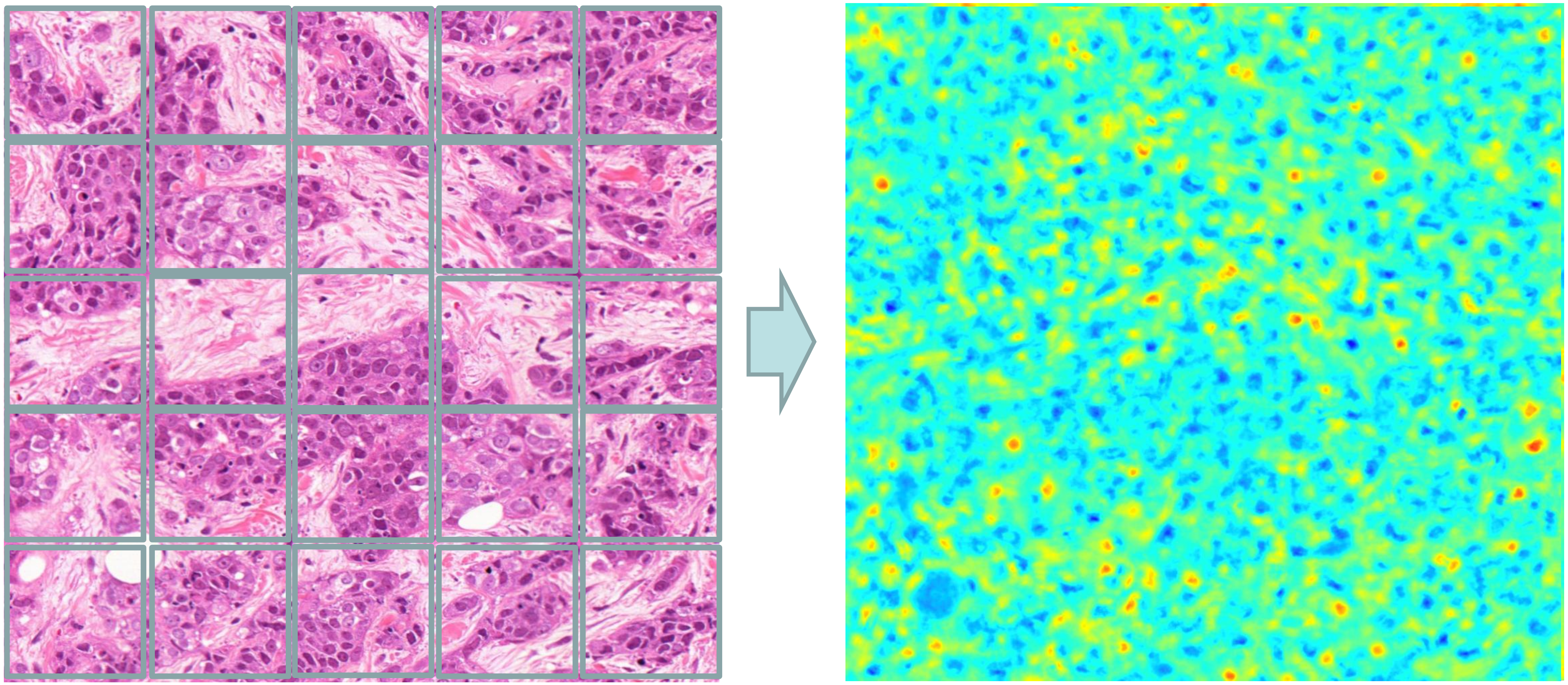


[Schütt et al. Nat Comm. 2017, Schütt et al JCP 2018]

Machine Learning for morpho-molecular Integration

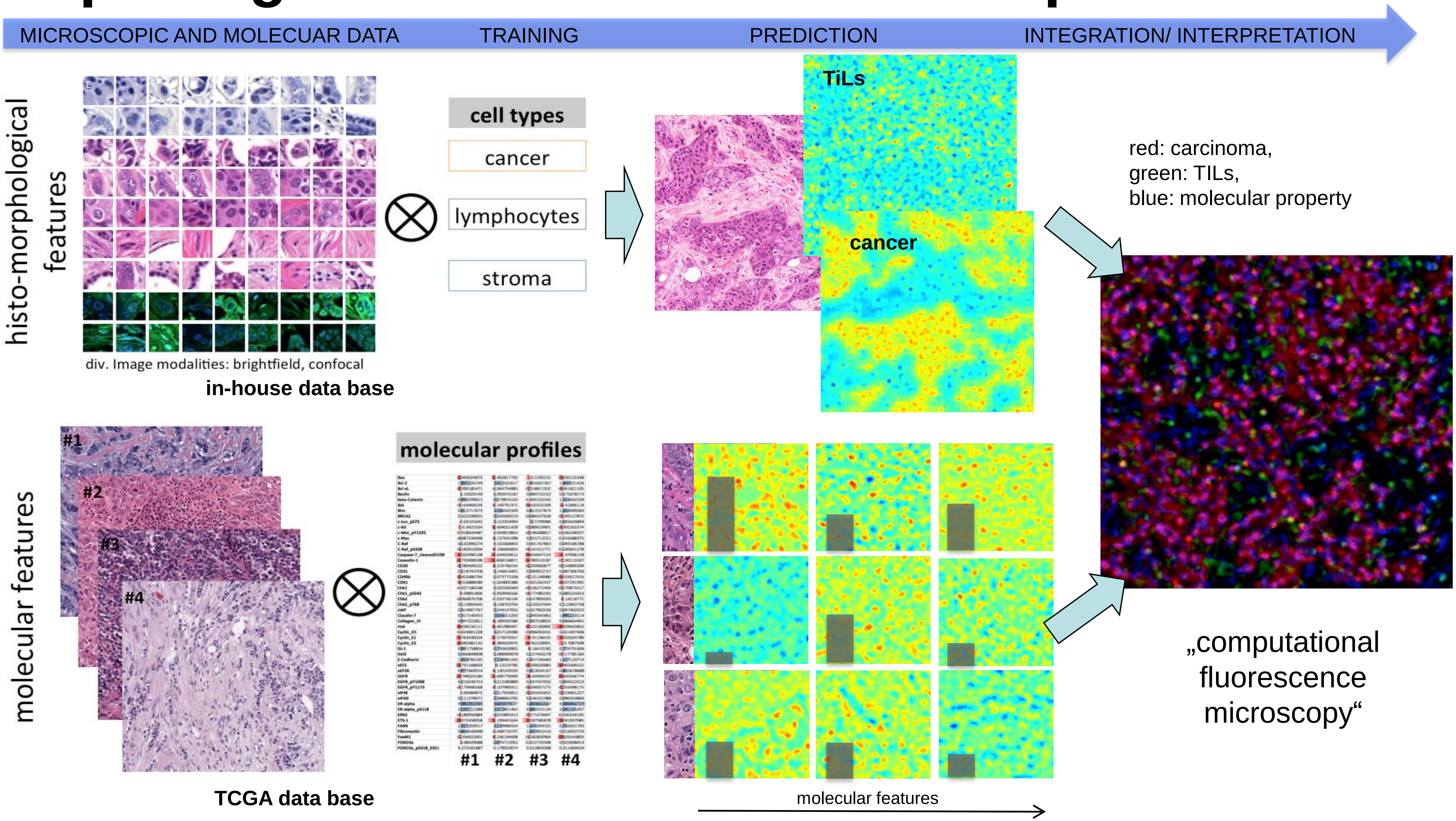
Alexander Binder[1,6], Michael Bockmayr[2,10], Miriam Hägele[1], Stephan Wienert[2], Daniel Heim[2], Katharina Hellweg[3], Albrecht Stenzinger[4], Laura Parlow[2], Jan Budczies[2], Benjamin Goeppert[4], Denise Treue[2], Manato Kotani[5], Masaru Ishii[5], Manfred Dietel[2], Andreas Hocke[3], Carsten Denkert[2,7], Klaus-Robert Müller[1,8,9,*] and Frederick Klauschen[2,7,*]

Interpretable ML



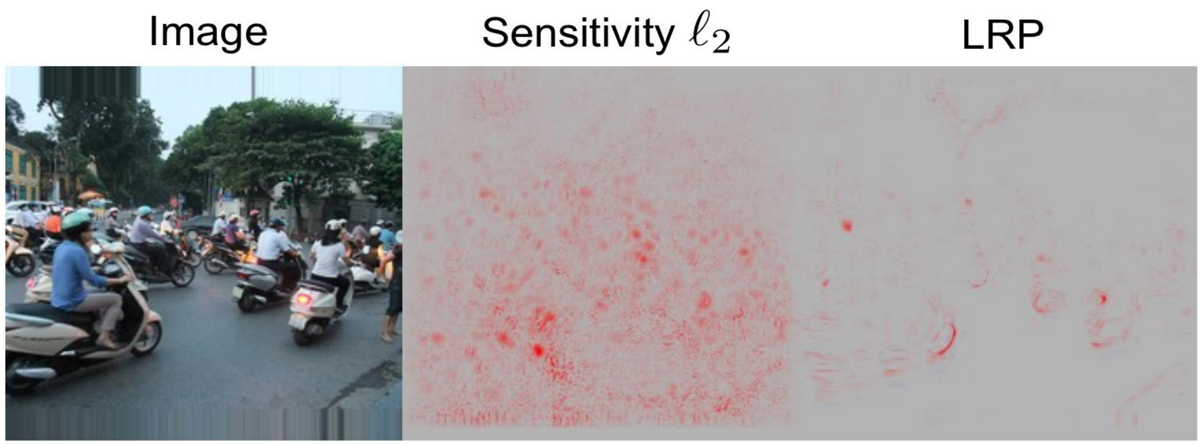
Bach et al., PLoS1 2015
Klauschen et al., US Patent #9558550
Binder et al., *in revision*

Machine learning based integration of morphological and molecular tumor profiles



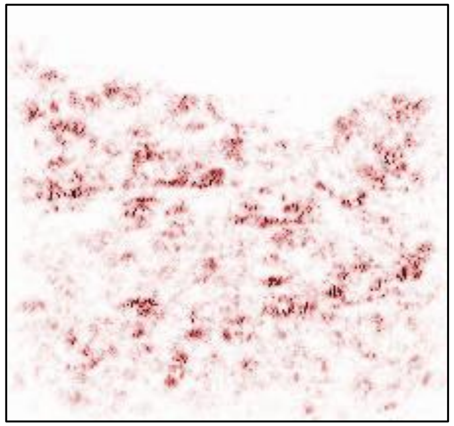
Take Home messages

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

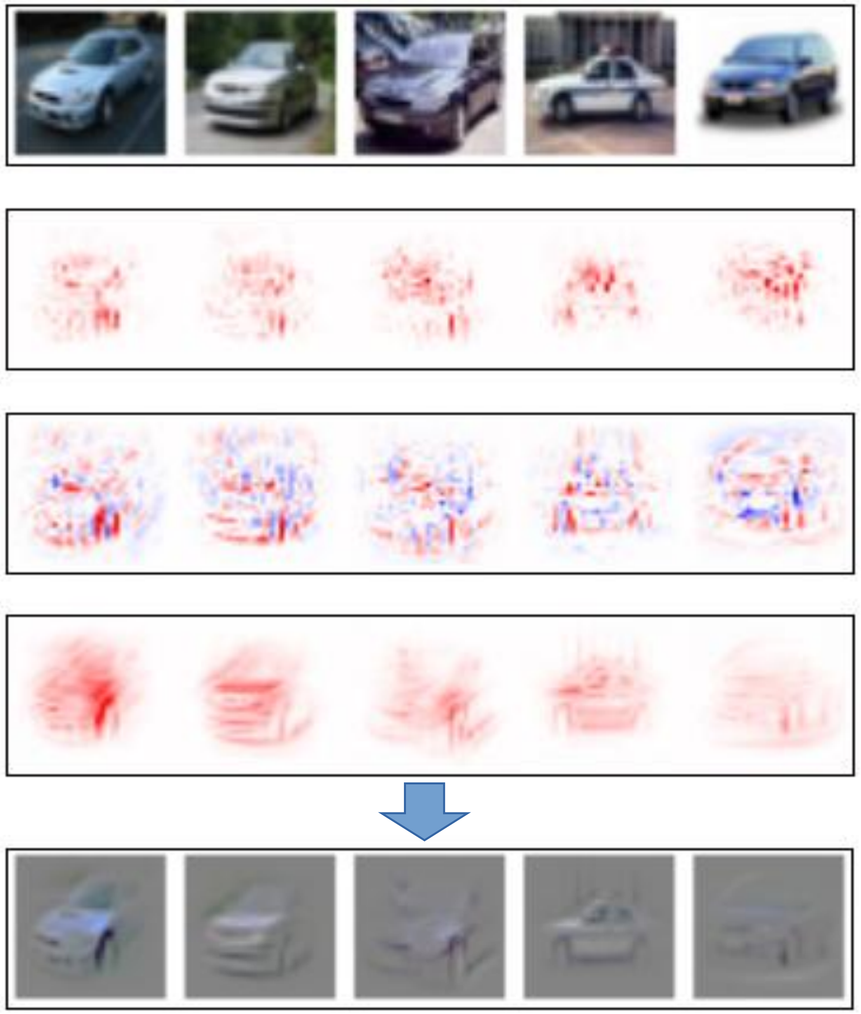
Observation:

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

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

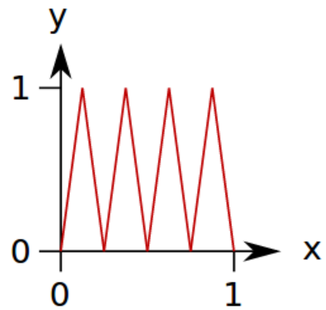
Explanation for simple models does not necessary work for deep models

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



gradient-based methods

vulnerable to shattered gradients

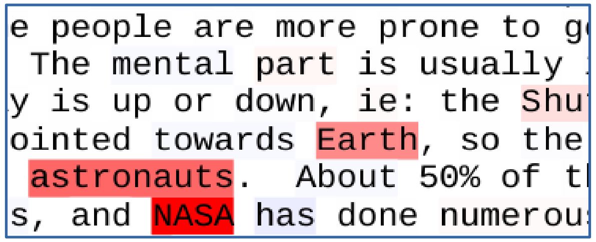


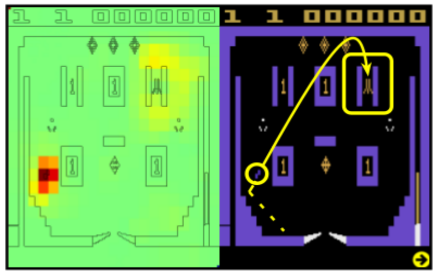
Our LRP method is robust to this.

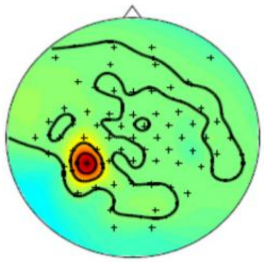
Layer-Wise Relevance Propagation

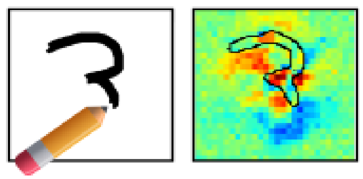
LRP Explanation Framework

(software, tutorials, demos, insights, applications)



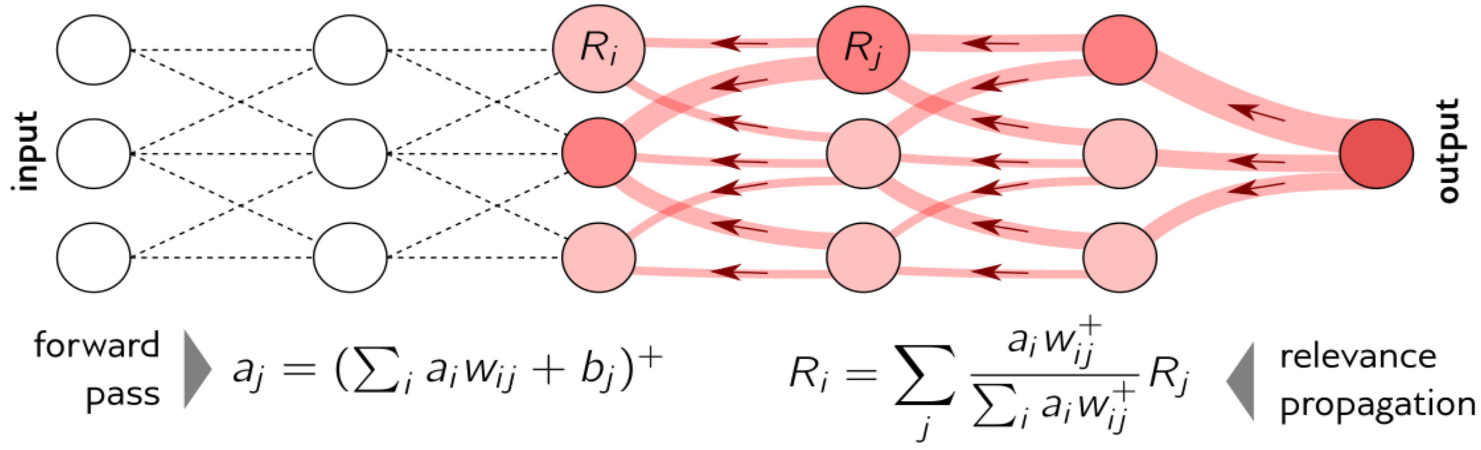






With a prediction score of **0.86**
the digit was classified as **3**

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



A Clarification on LRP

$$\text{LRP} \neq \text{Gradient} \times \text{Input}$$

… 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).

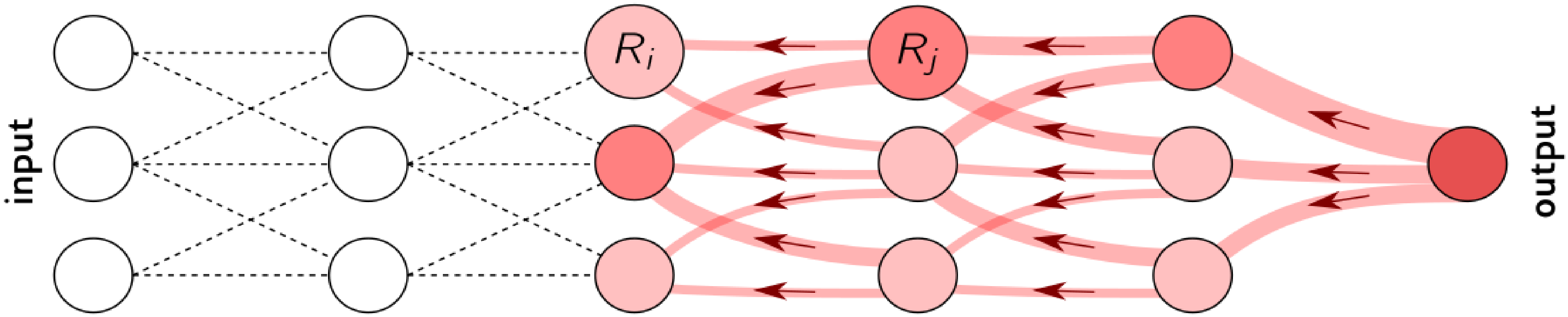
Good news: No need to reimplement LRP, check our software at www.heatmapping.org.

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).



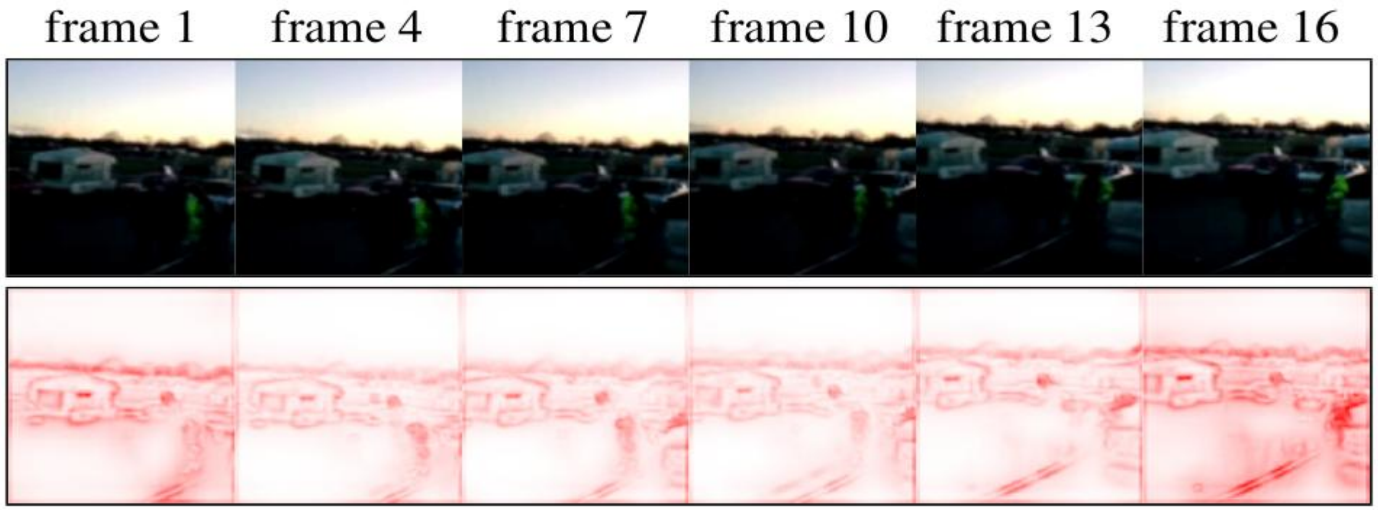
forward pass $a_j = (\sum_i a_i w_{ij} + b_j)^+$

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

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

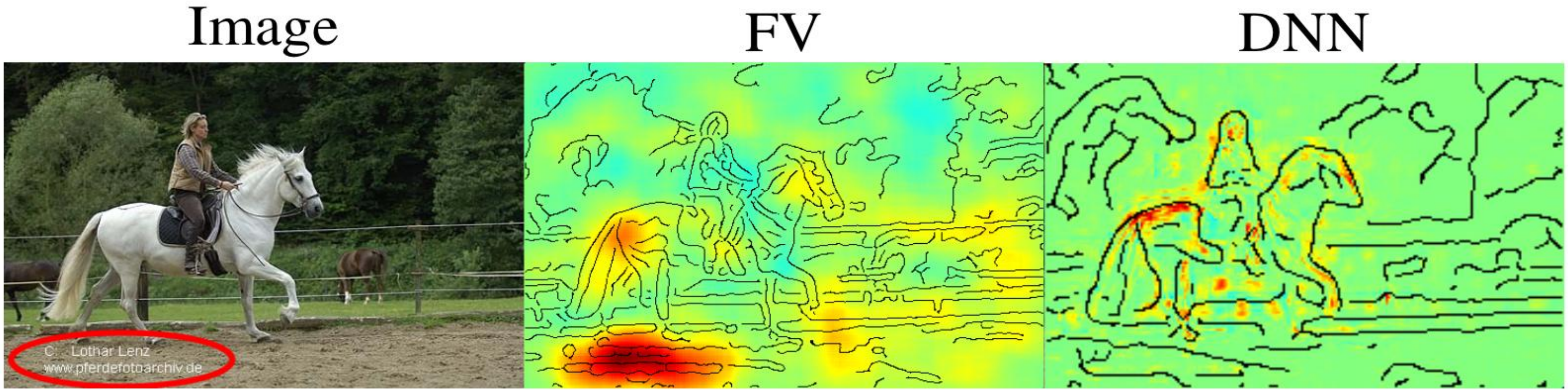
[Samek et al. IEEE TNNLS 2017]

Explanation helps to improve models



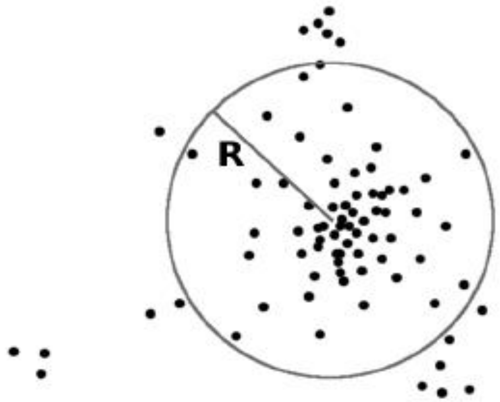
Explaining ML, Now What?

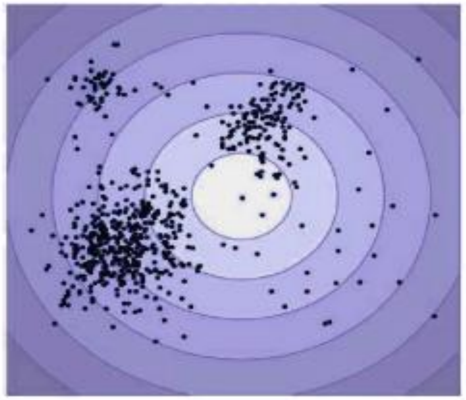
Explanation helps to find flaws in models

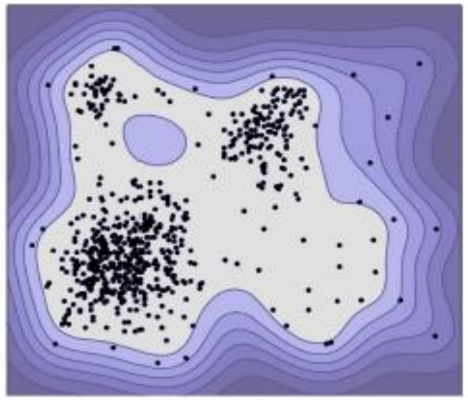


[Lapuschkin et al CVPR 2016]

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

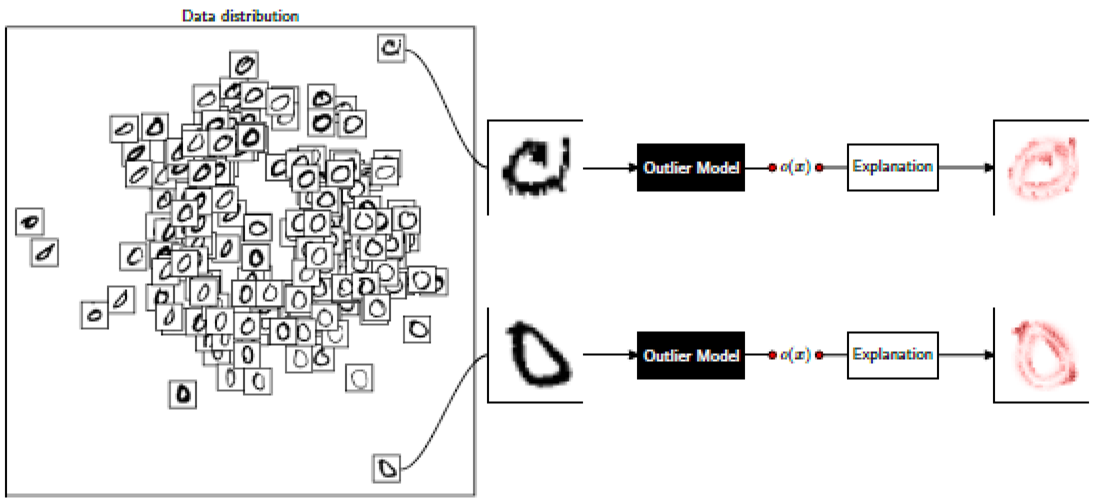


Figure 1: Illustration of the outlier detection and explanation setting. *Left:* Data is generated from an unknown distribution, we are for example interested in potential outliers; *Middle:* Unsupervised machine learning techniques estimate the data generating distribution and assign an outlier score $o(x)$ 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.

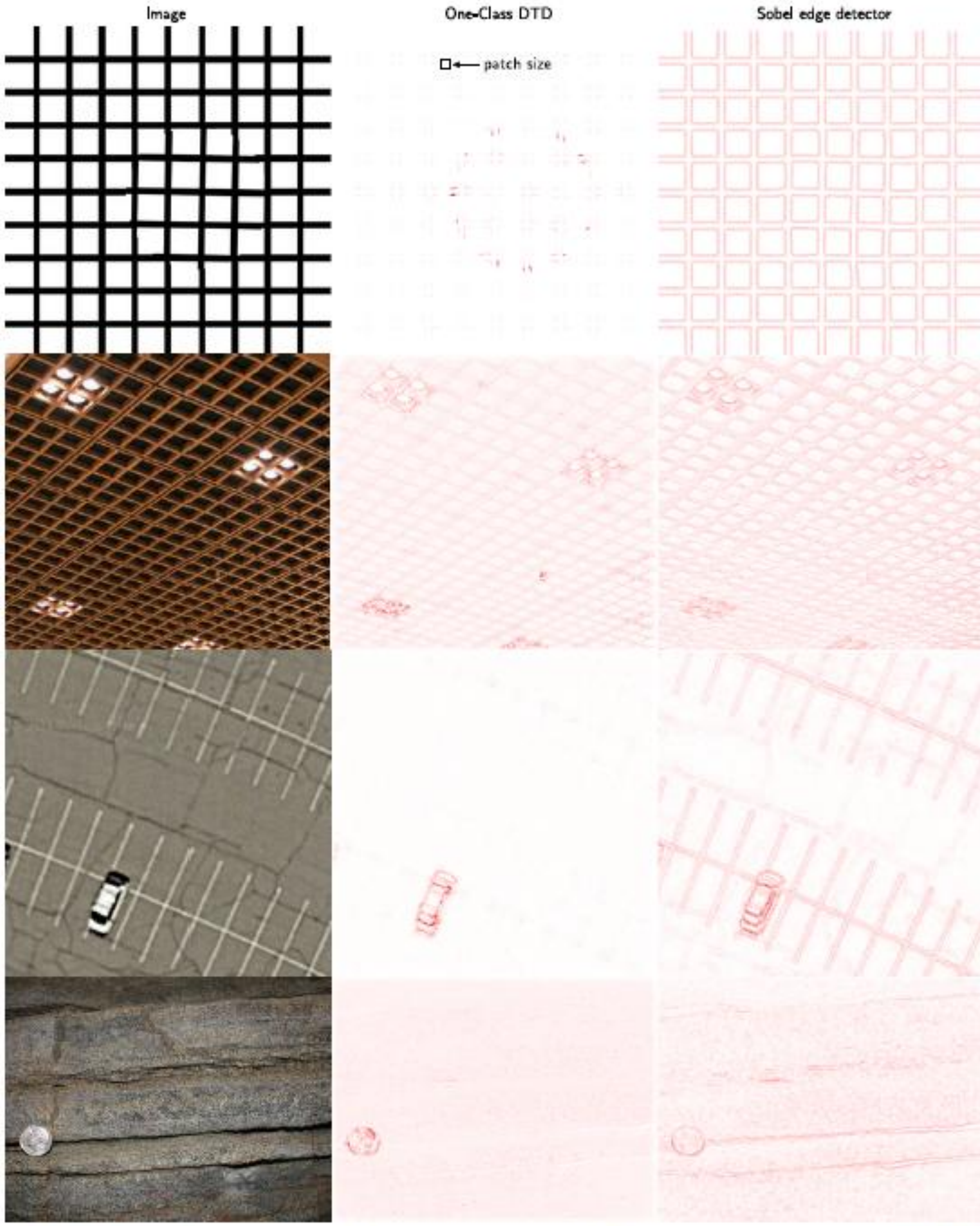


Figure 5: A One-Class SVM is trained on small 7×7 patches of the very image itself. Parameter $\nu = 0.1$ is set to allow at most 10% outliers. Images from a texture data set [11] (row one, two and four) and PatternNet [61]; top image is altered by us. For every image, we show *Left:* input image; *Middle*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

[Kaufmann, Müller, Montavon 2018]

Getting new Insights in the Sciences

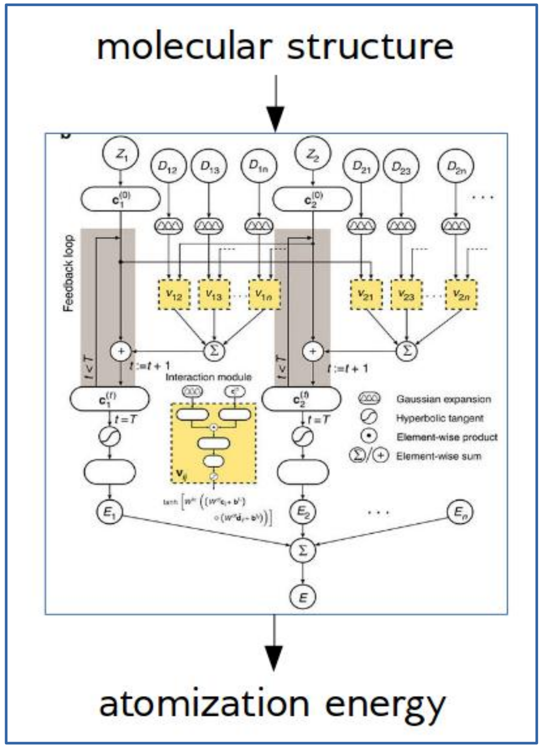
Example: Understanding physical systems at the quantum level.

time-independent Schrödinger Equation

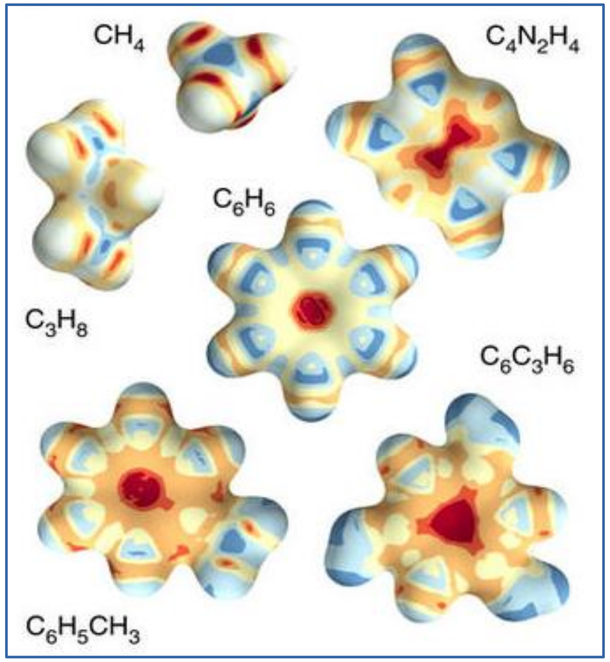
$$\hat{H}\Psi = E\Psi$$

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**

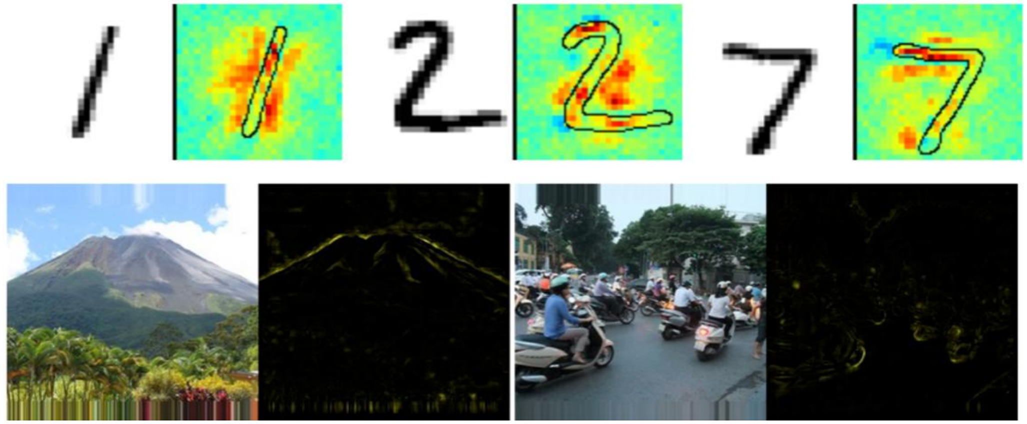
www.heatmapping.org

Thank you for your attention

Visit:

`http://www.heatmapping.org`

- Tutorials
- Software
- Online Demos



Tutorial Paper

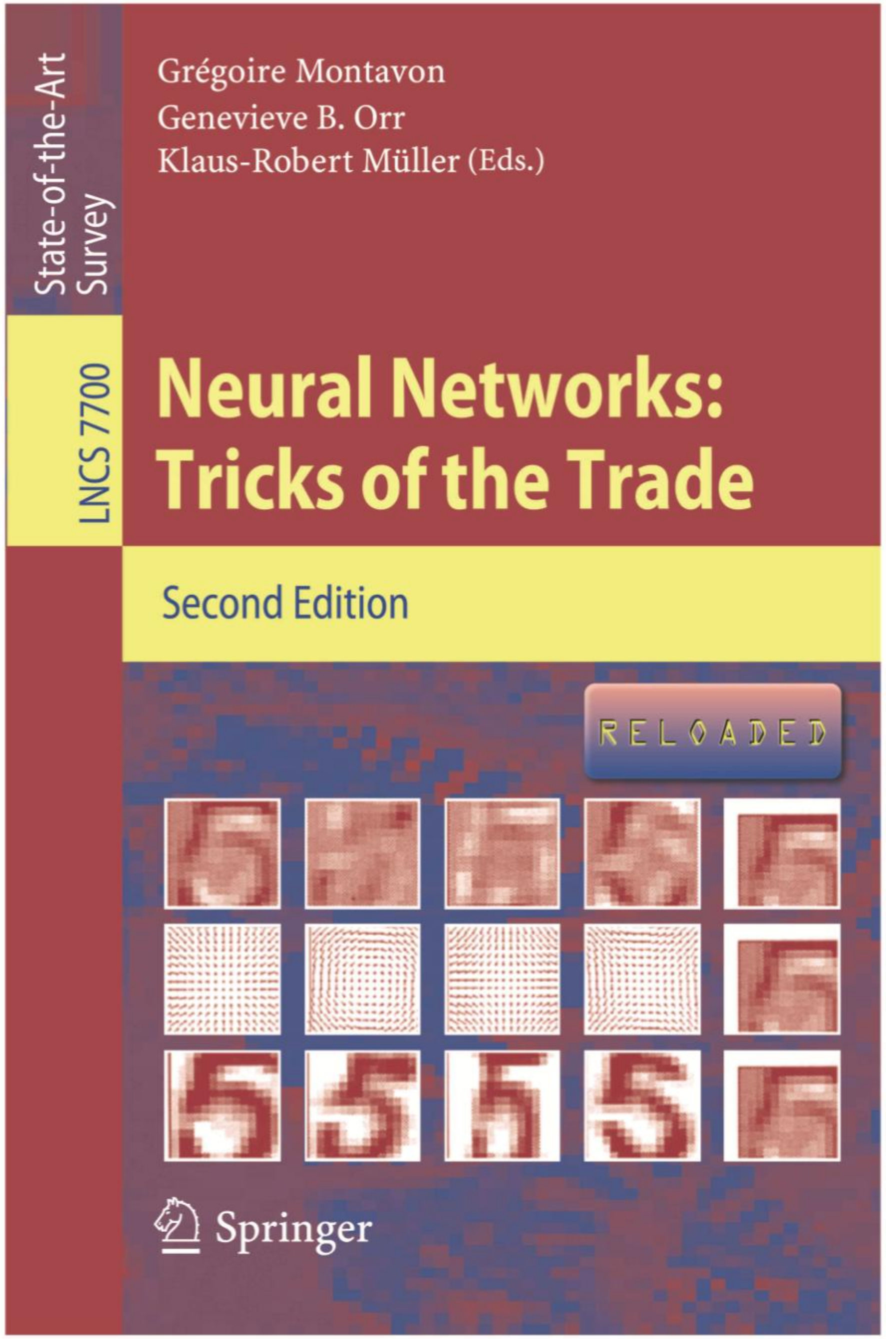
Montavon et al., "Methods for interpreting and understanding deep neural networks", Digital Signal Processing, 73:1-5, 2018

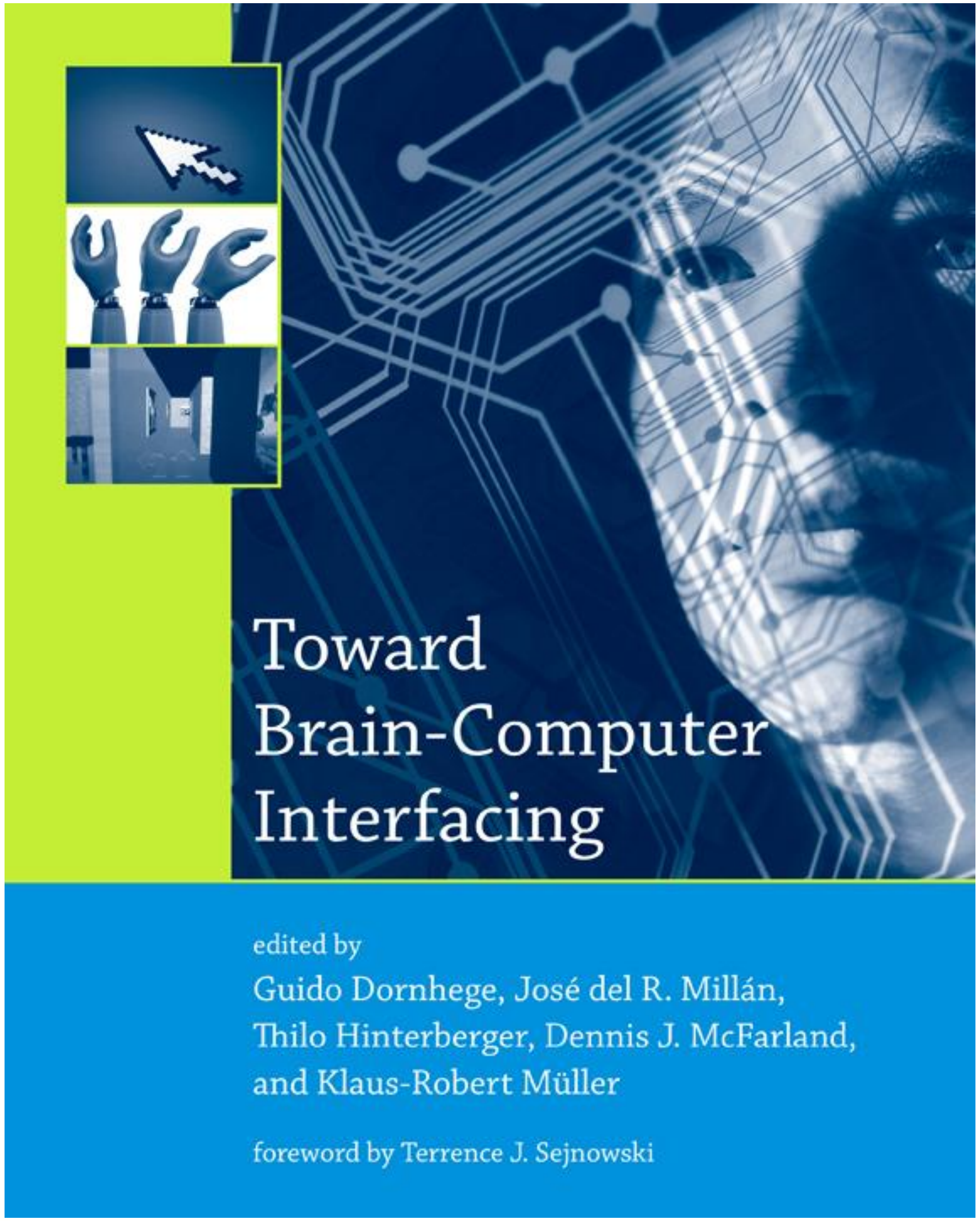
Keras Explanation Toolbox

https://github.com/albermax/innvestigate

Fraunhofer Heinrich Hertz Institute

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN


State-of-the-Art Survey
LNCS 7700
Grégoire Montavon
Genevieve B. Orr
Klaus-Robert Müller (Eds.)
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

BBDC

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., & Muller, 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, KR 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.

.