# Understanding ML Models



Berliner Zentrum für
MASCHINELLES LERNEN

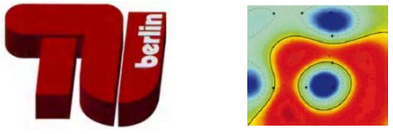
BBDC
BERLIN BIG
DATA CENTER

mpii max planck institut informatik

**Klaus-Robert Müller !!et al.!!**

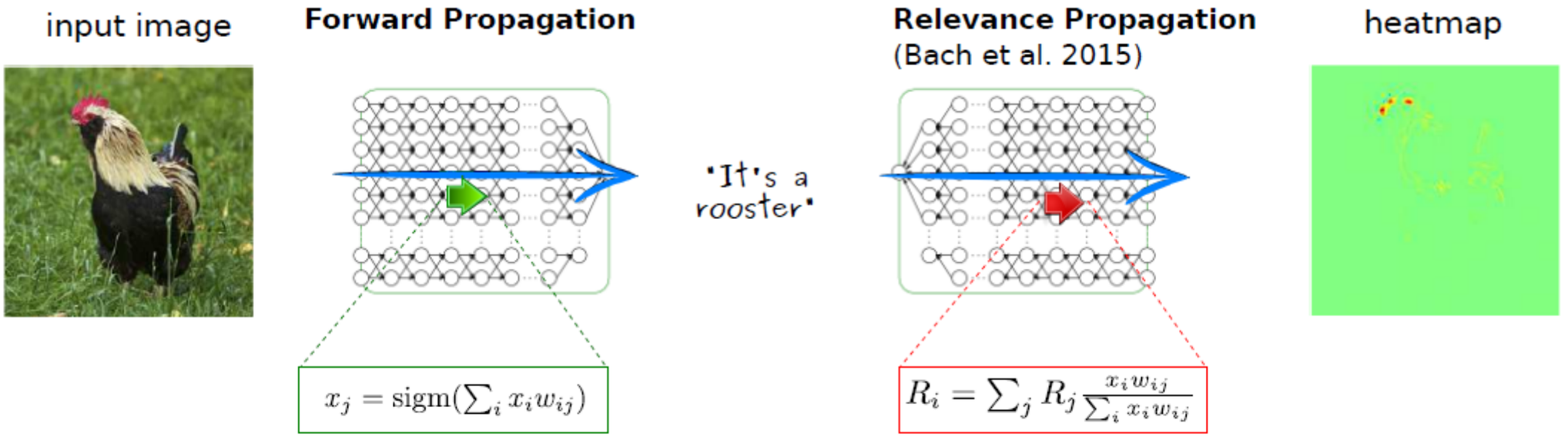
# Outline

- understanding single decisions of nonlinear learners
- Layer-wise Relevance Propagation (LRP)
- Applications in Physics, Chemistry and Medicine: **towards insights**



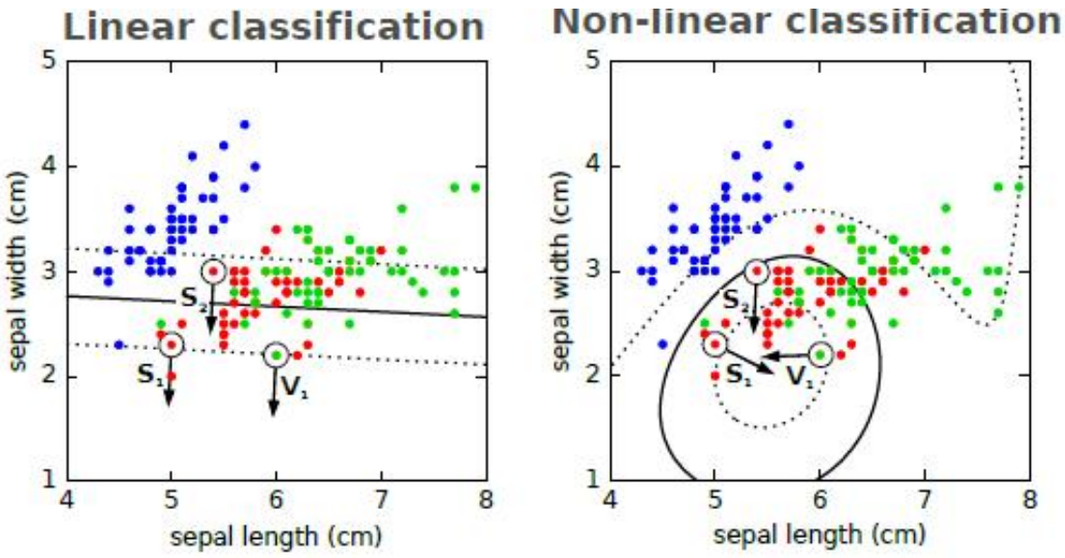
# Towards Explaining:
# Machine Learning = black box?

# Explaining single Predictions Pixel-wise




**Explaining single decisions is difficult!**

# Explaining nonlinear decisions is difficult




Iris setosa (red)


Iris virginica (green)


Iris versicolor (blue)

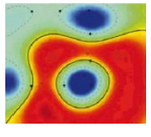

**Explaining individual classification decisions**

Linear classification

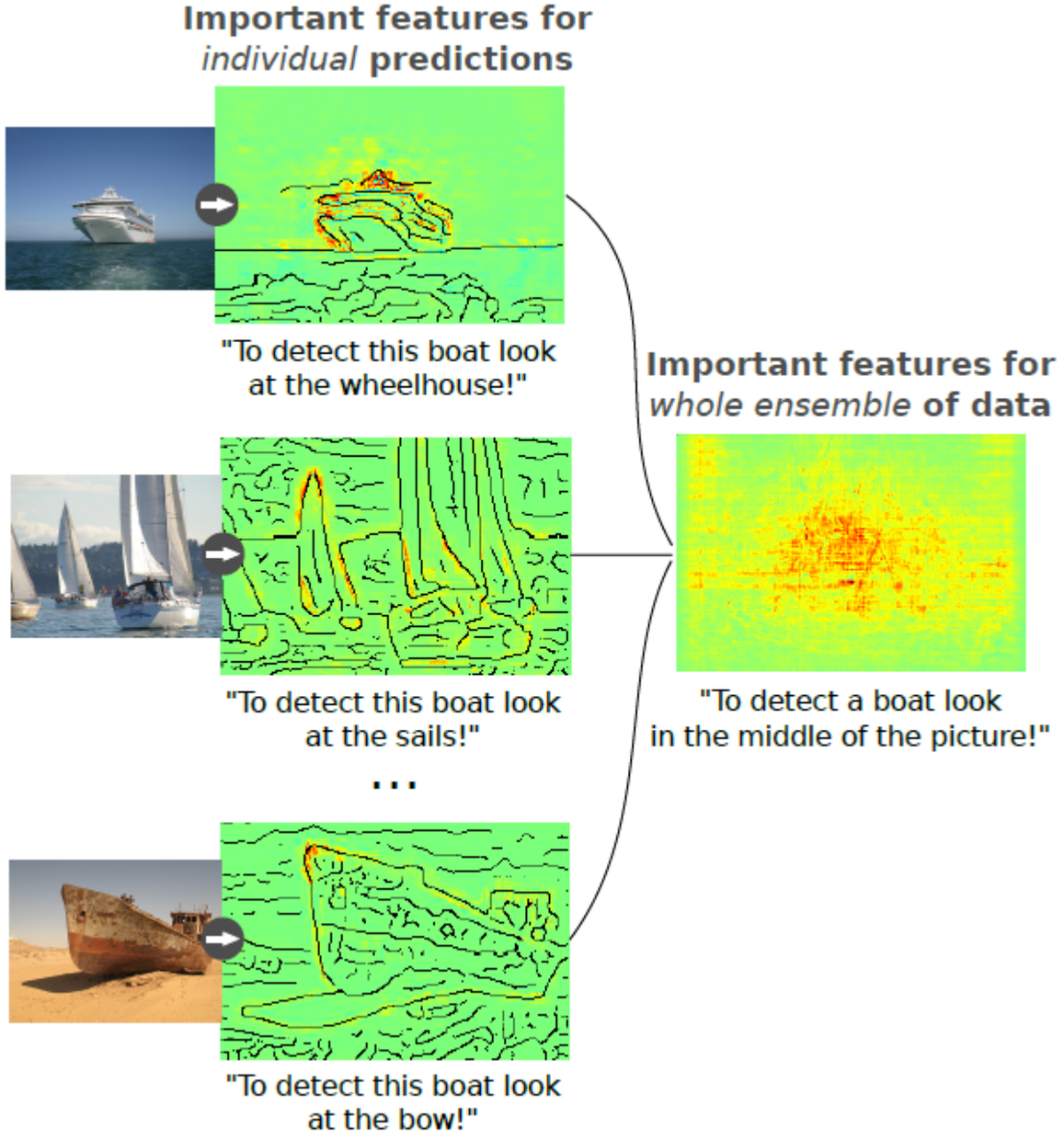
$S_1$: sepal width
$S_2$: sepal width
$V_1$: sepal width

Non-linear classification

$S_1$: sepal width & length
$S_2$: sepal width
$V_1$: sepal length



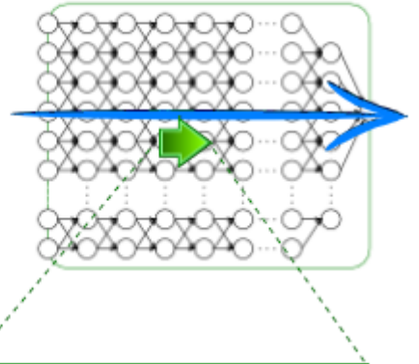
# Explaining single decisions is difficult



# Explaining single Predictions Pixel-wise

input image

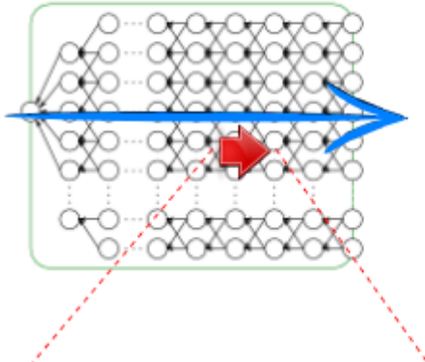
**Forward Propagation**



$$x_j = \text{sigm}(\sum_i x_i w_{ij})$$

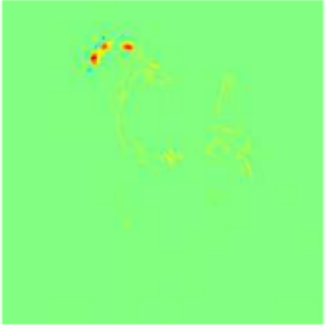
"It's a rooster"

**Relevance Propagation**
(Bach et al. 2015)



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

heatmap



**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)

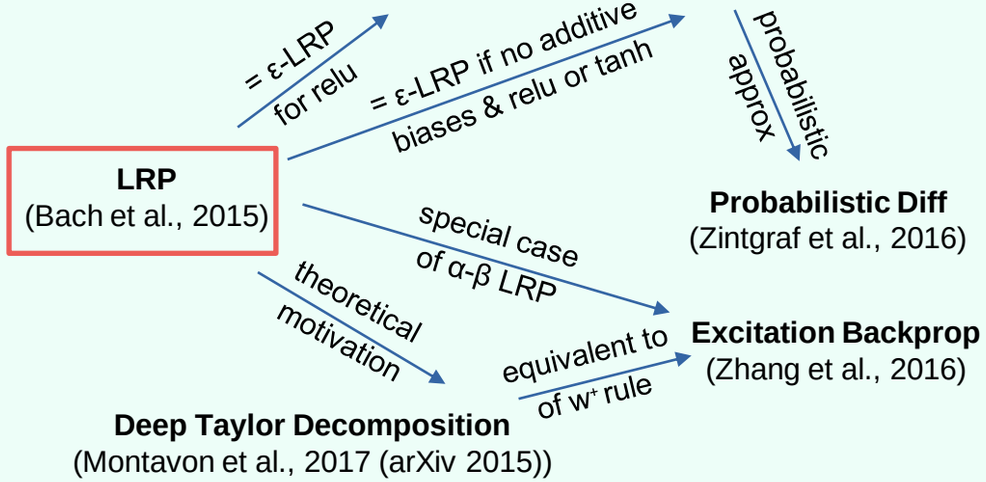
**Gradient times input**
(Shrikumar et al., 2016)

**DeepLIFT**
(Shrikumar et al., 2016)

**Grad-CAM**
(Selvaraju et al., 2016)

**Integrated Gradient**
(Sundararajan et al., 2017)

## Decomposition




**LRP for LSTM**
(Arras et al., 2017)

## 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**
(Dosovitskiy & Brox, 2015)

**Synthesis of preferred inputs**
(Nguyen et al. 2016)

**TCAV**
(Kim et al. 2018)

**Inverting CNNs**
(Mahendran & Vedaldi, 2015)

**RNN cell state analysis**
(Karpathy et al., 2015)

**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)*

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

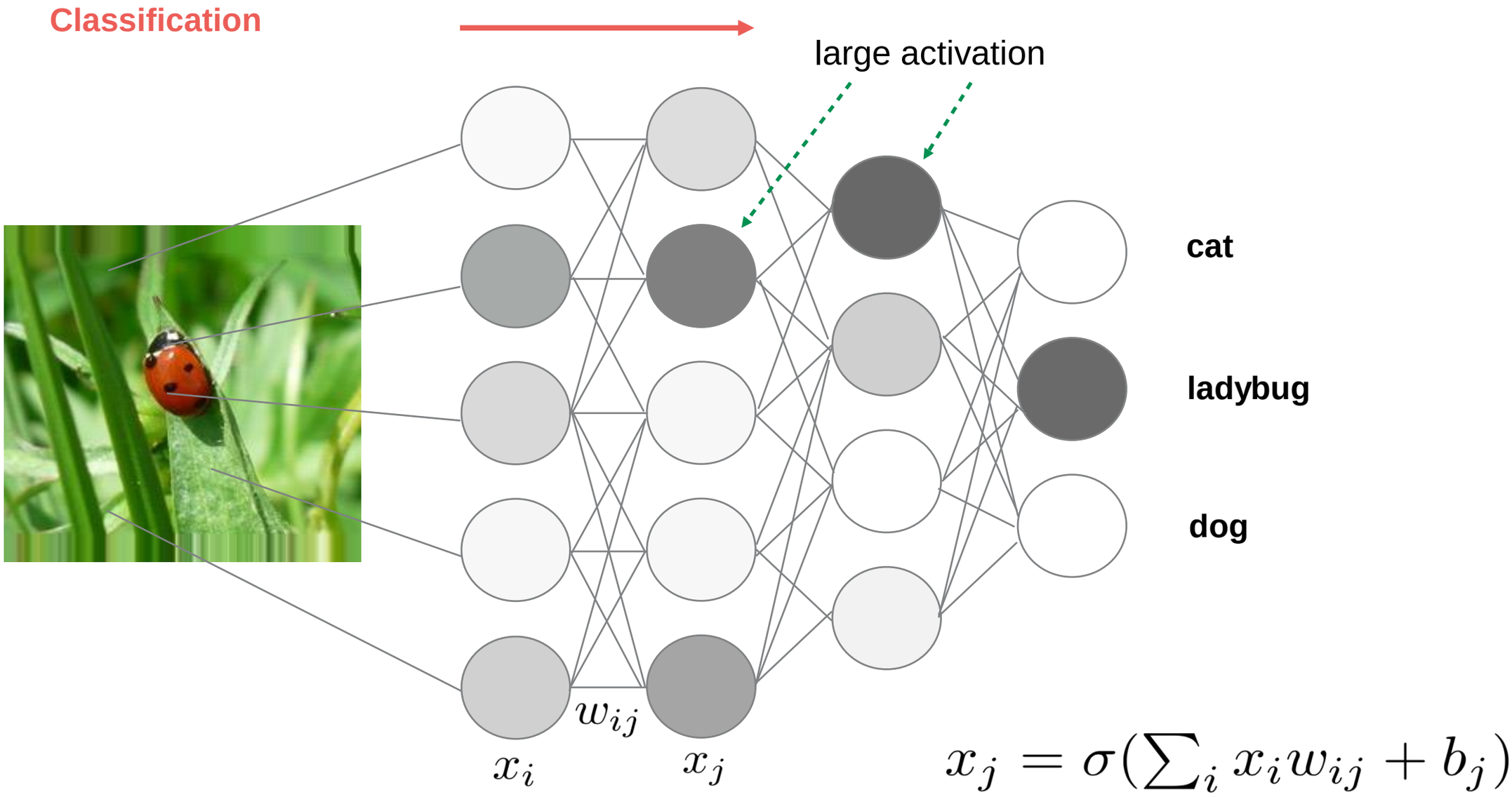
**Sensitivity / Saliency**: "*Which pixels lead to increase/decrease of prediction score when changed" (what makes this image to be classified more/less as a car) (Baehrens et al 10, **Simonyan et al 14**)*

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

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

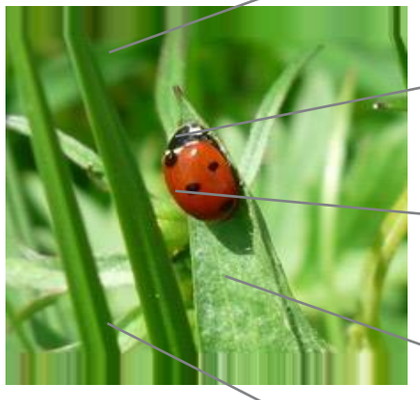
Each method solves a **different** problem!!!

# Explaining Neural Network Predictions

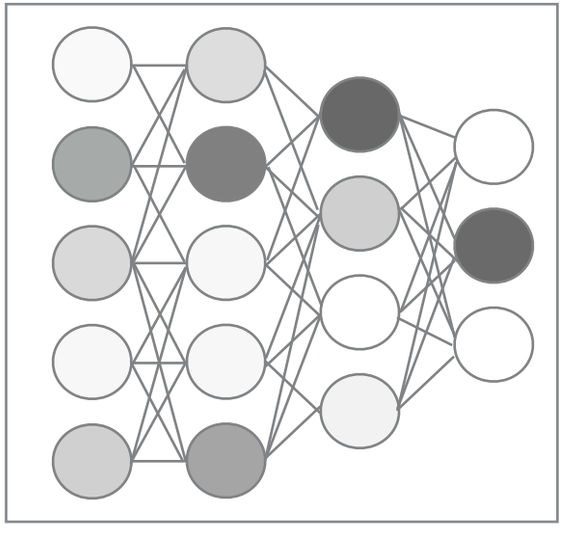


# Explaining Neural Network Predictions

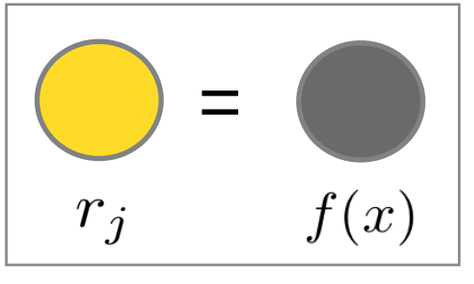
**Explanation**





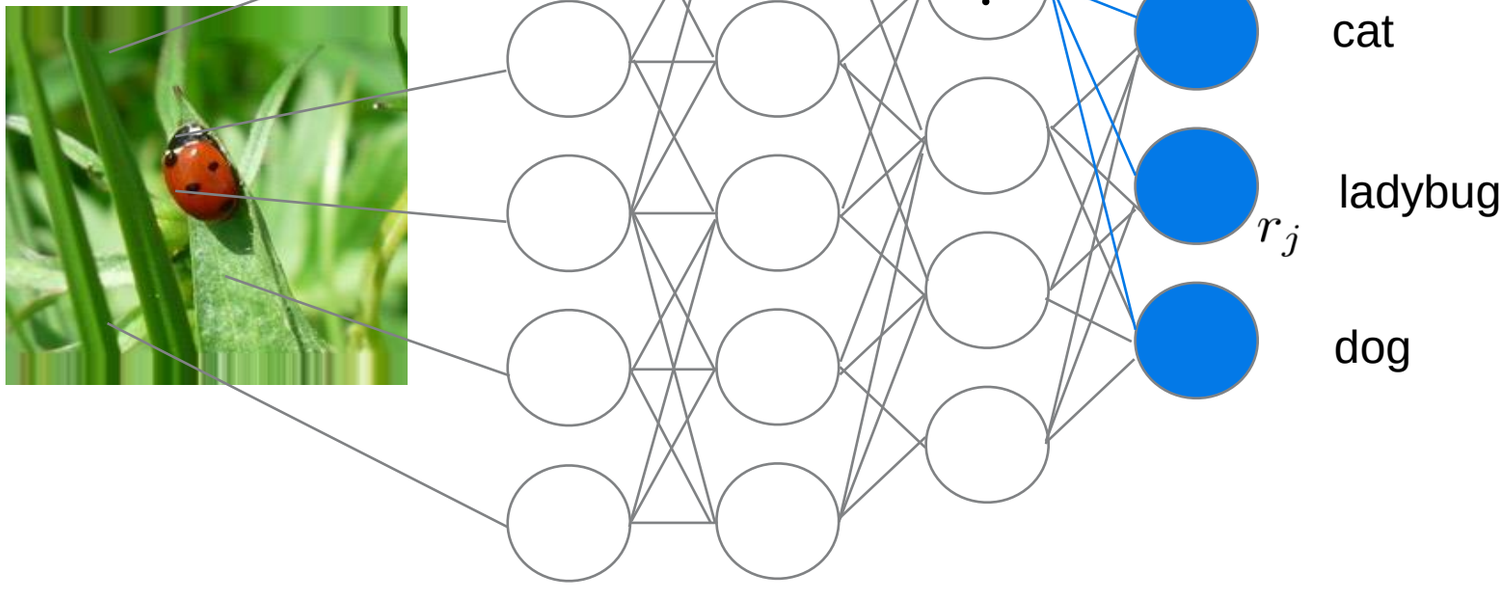



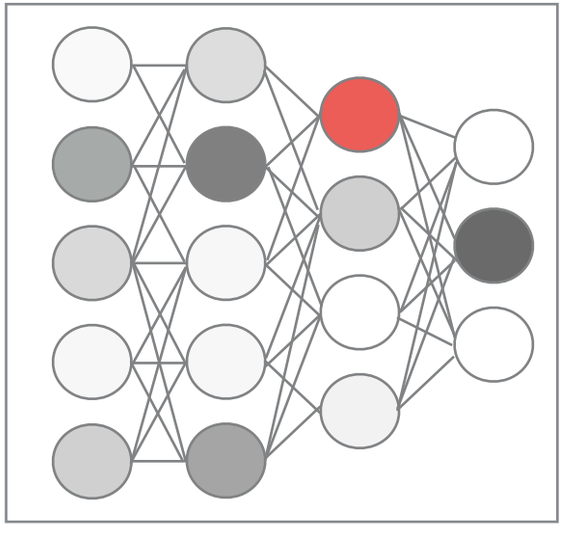
Initialization



# Explaining Neural Network Predictions

Explanation



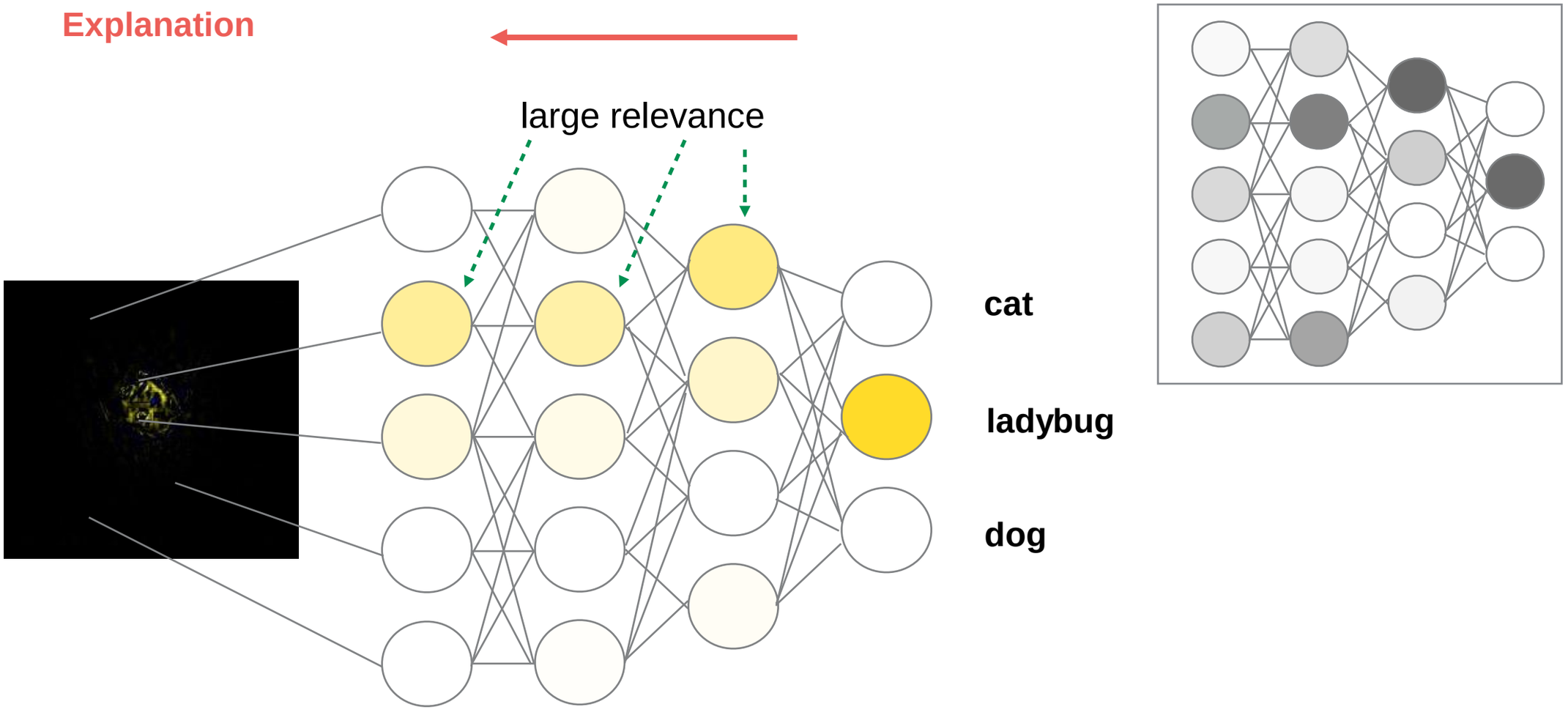



$$r_i = x_i \sum_j \frac{w_{ij} r_j}{\sum_i x_i w_{ij}} = x_i c_i$$

**Theoretical interpretation**
Deep Taylor Decomposition

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

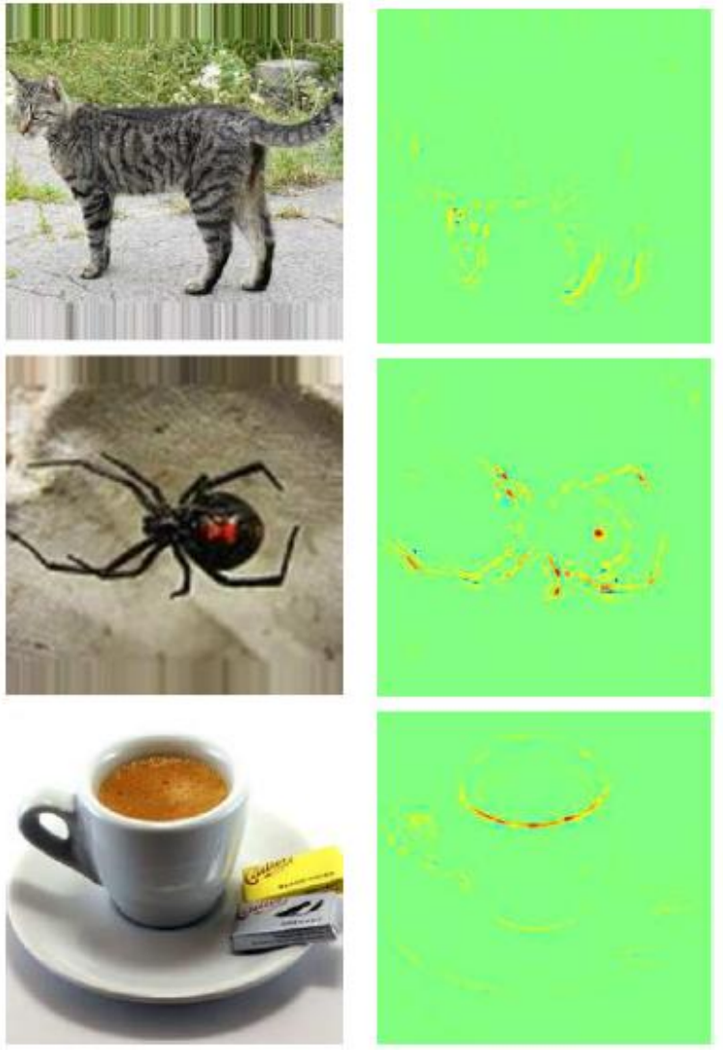
# Explaining Neural Network Predictions



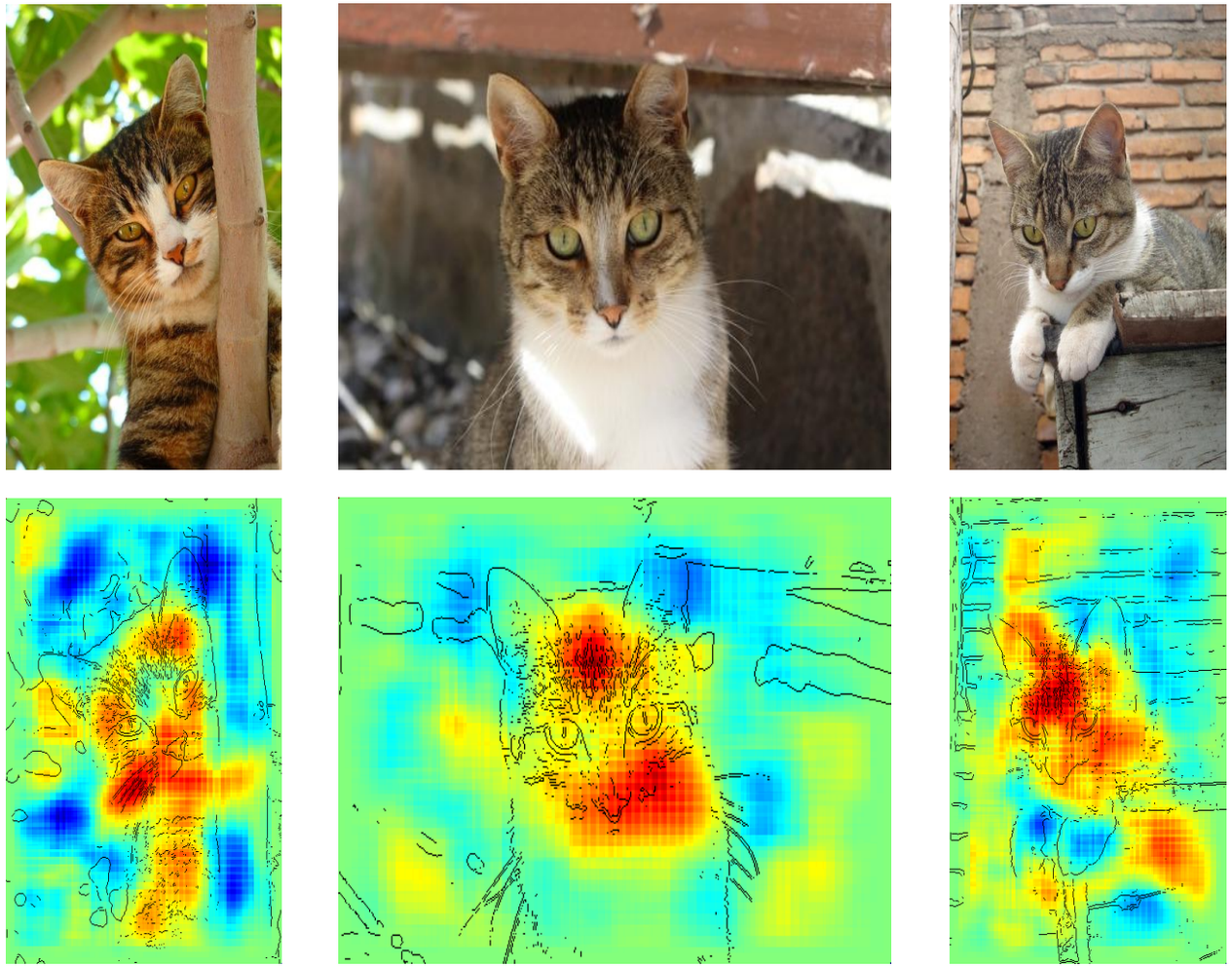

Relevance Conservation Property

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

# Explaining Predictions Pixel-wise



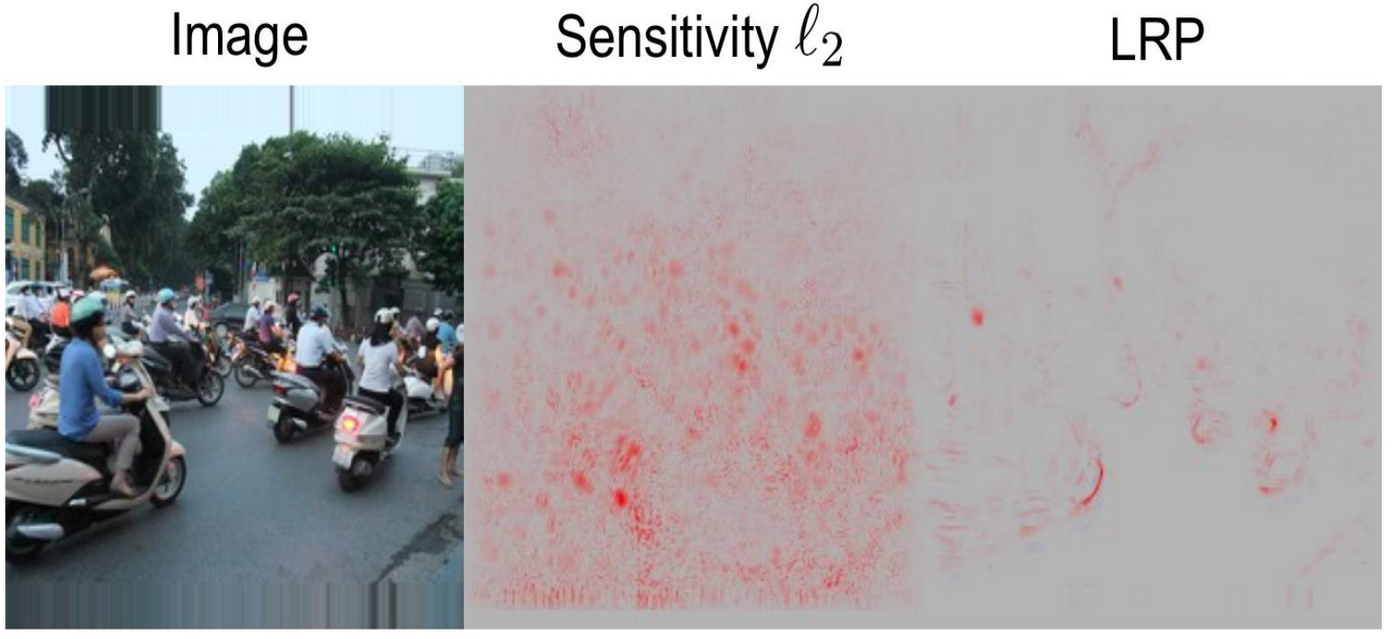
Neural networks



Kernel methods

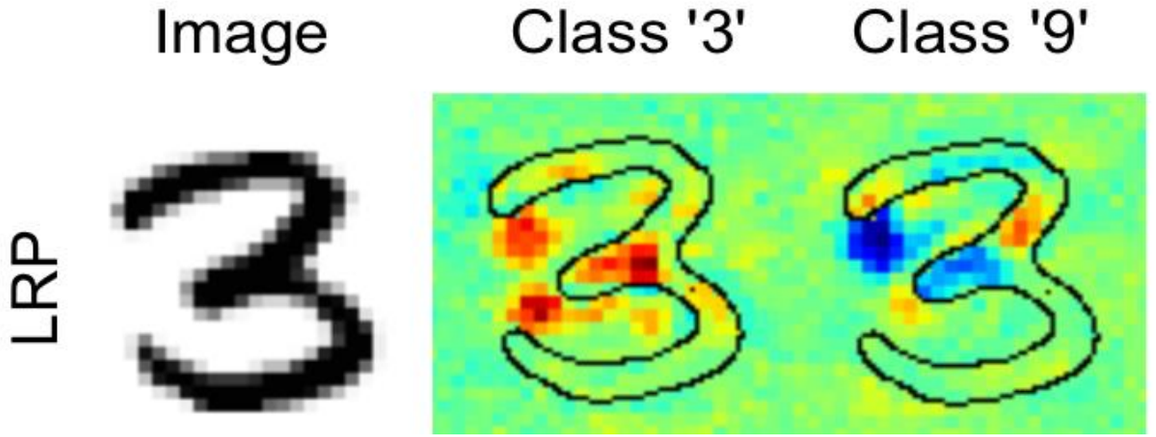
# Some Digestion on Explaining

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



# LRP can ‘say’ positive and negative things

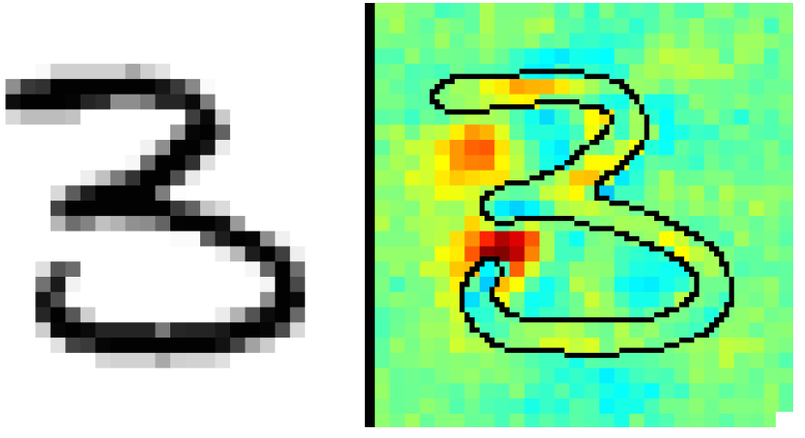
Positive and Negative Evidence: LRP distinguishes between positive evidence, supporting the classification decision, and negative evidence, speaking against the prediction




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

Play Video

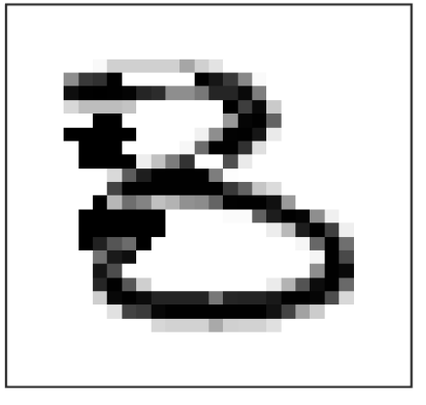
# Measuring the Quality of Explanation (Samek et al 2017)

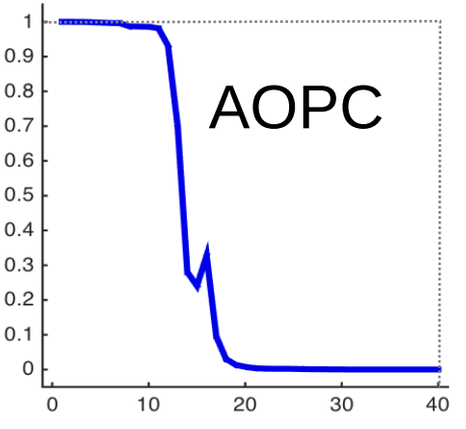


Is this a good explanation ?

Algorithm

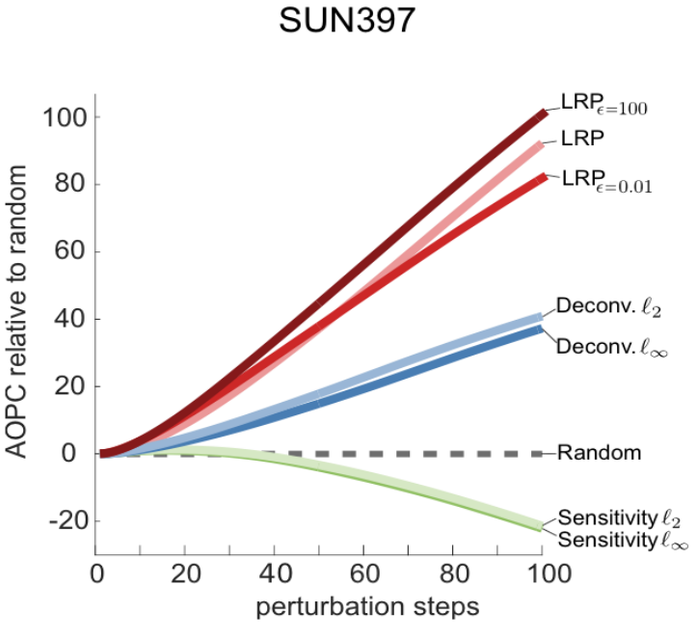
```
Sort pixel scores
Iterate
    flip pixels
    evaluate f(x)
Measure decrease of f(x)
```

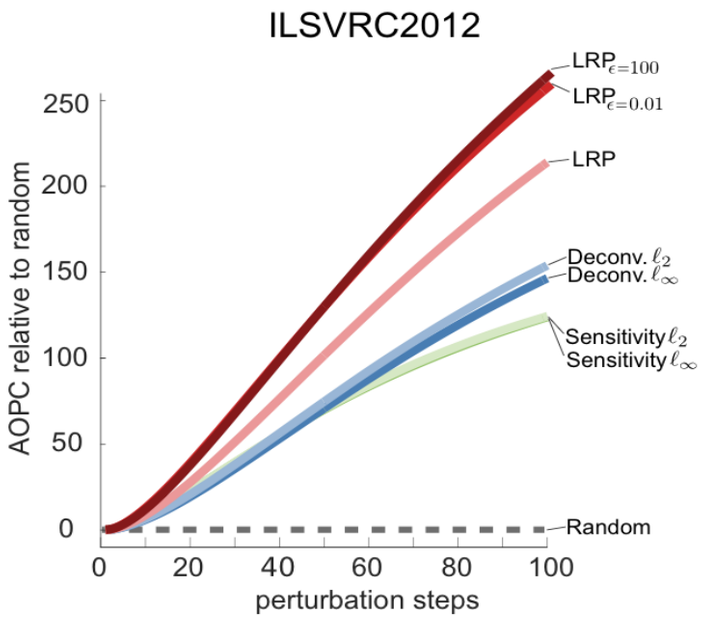


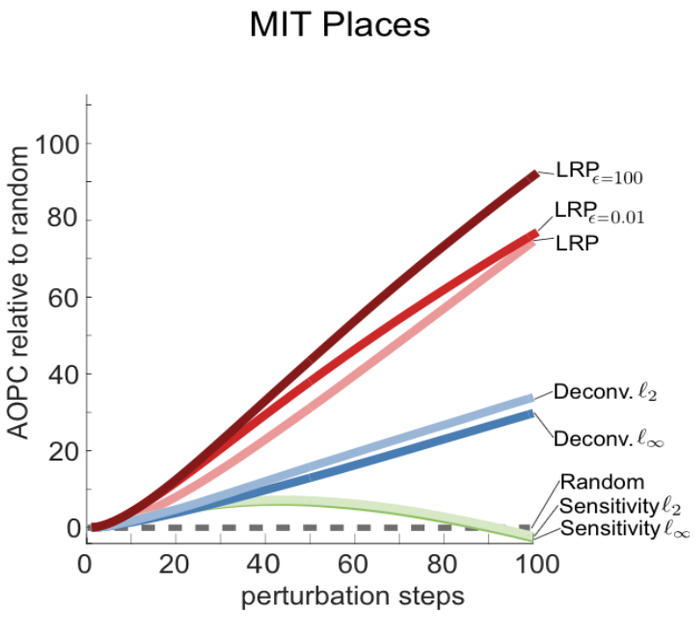



[Samek et al IEEE TNNLS 2017]

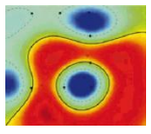
# Measuring the Quality of Explanation



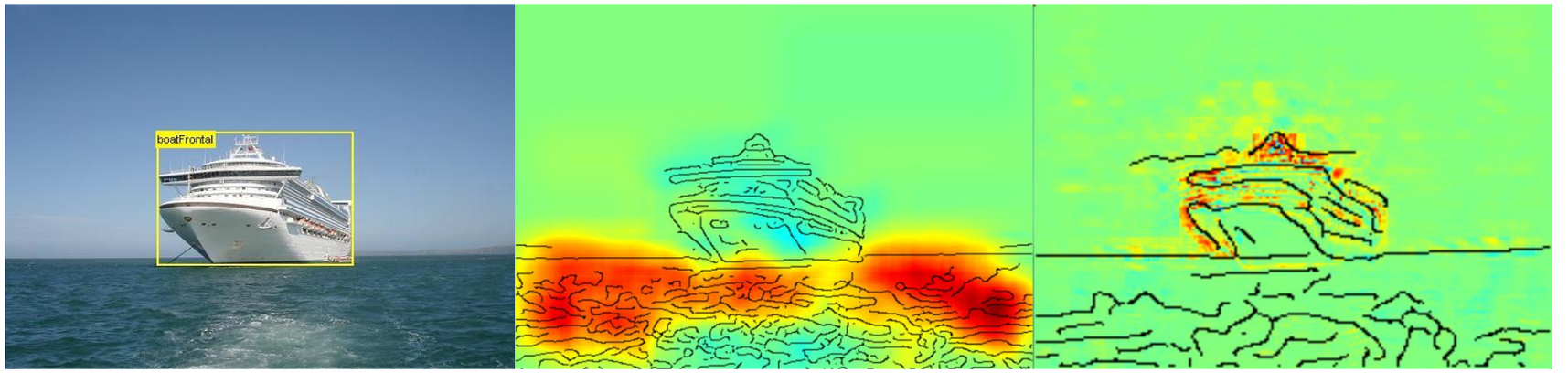



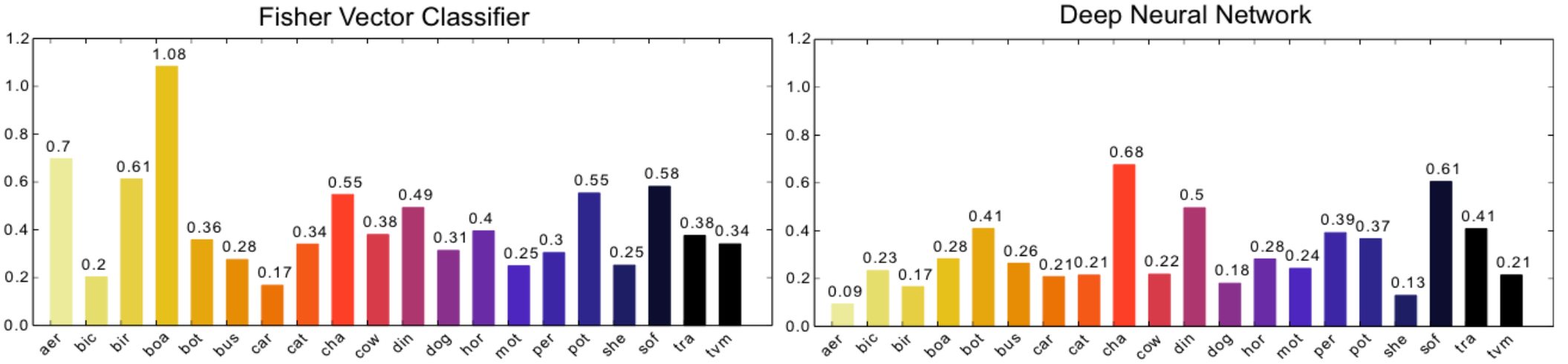




LRP outperforms Sensitivity and Deconvolution on all three datasets.

# Application: Comparing Classifiers



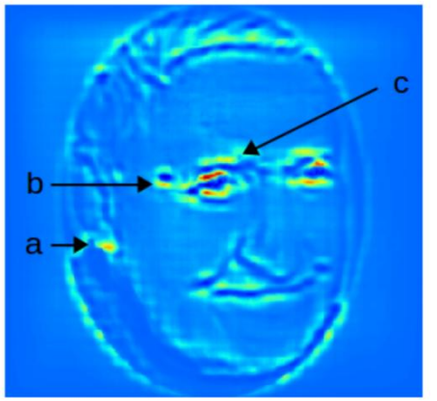

**Large values indicate importance of context**

[Lapuschkin et al CVPR 2016]

# Applying Explanation in Vision and Text

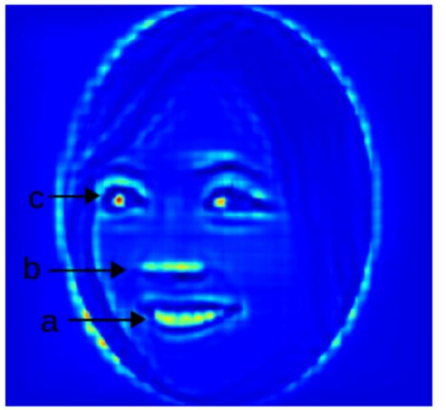
# Application: Faces

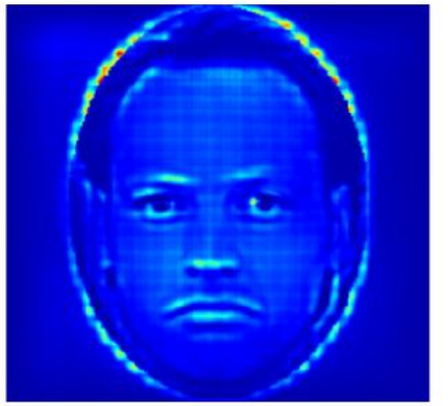
What makes
you look old ?



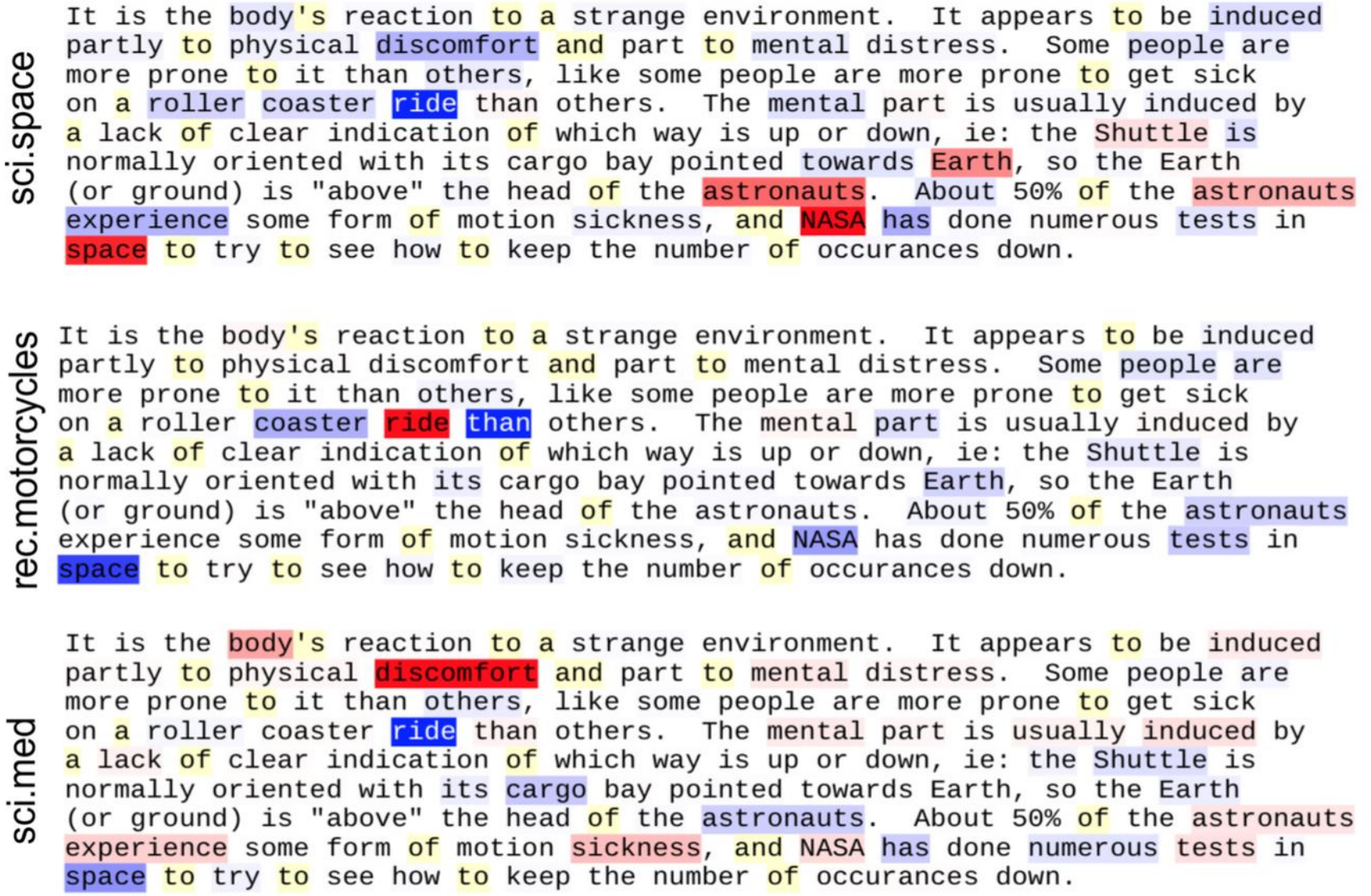

What makes
you look sad ?

What makes
you look attractive ?



# Application: Document Classification

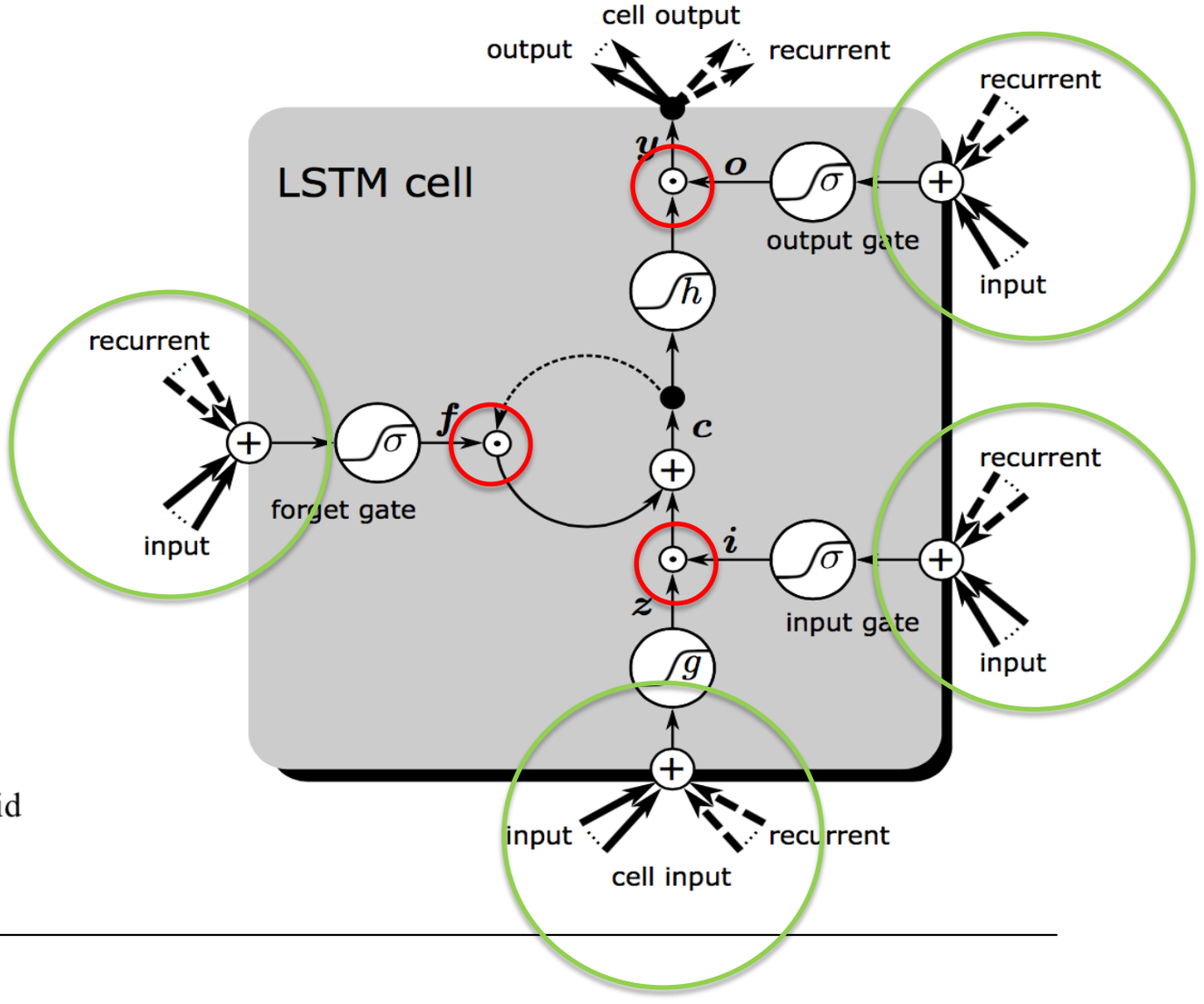


# LRP for LSTMs

- 2 types of operations:
  - Dot Product
  - Multiplicative

- Dot Product:
  - $R_l = \frac{z_{ij}}{z_j + \varepsilon * sign(z_j)} * R_{l+1}$
  - $z_{ij} = x_i w_{ij}$
  - $z_j = \sum_i x_i w_{ij}$

- Multiplicative:
  - Occurrence:
    - output of gate * source (e.g. cell state)
  - $R_{gate} = 0$
  - $R_{source} = R_{l+1}$
  - Gate already accounted for in forward-pass (decides which information to keep)

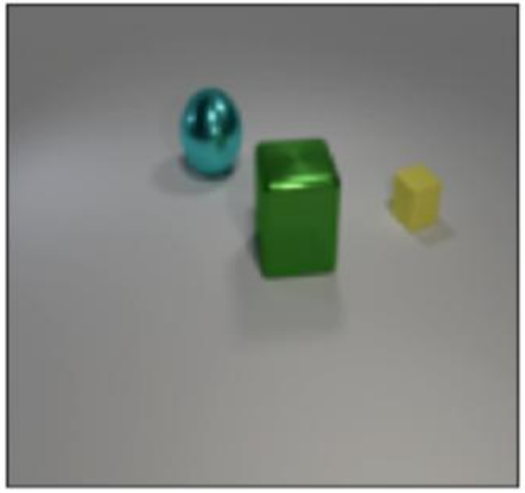


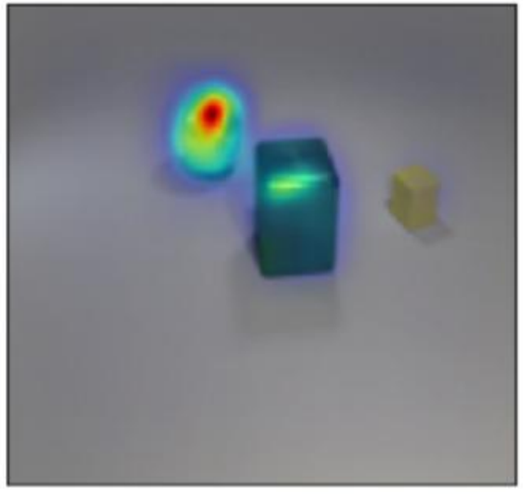

(Arras et al., 2017)

# Explaining LSTMs

**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 ? |

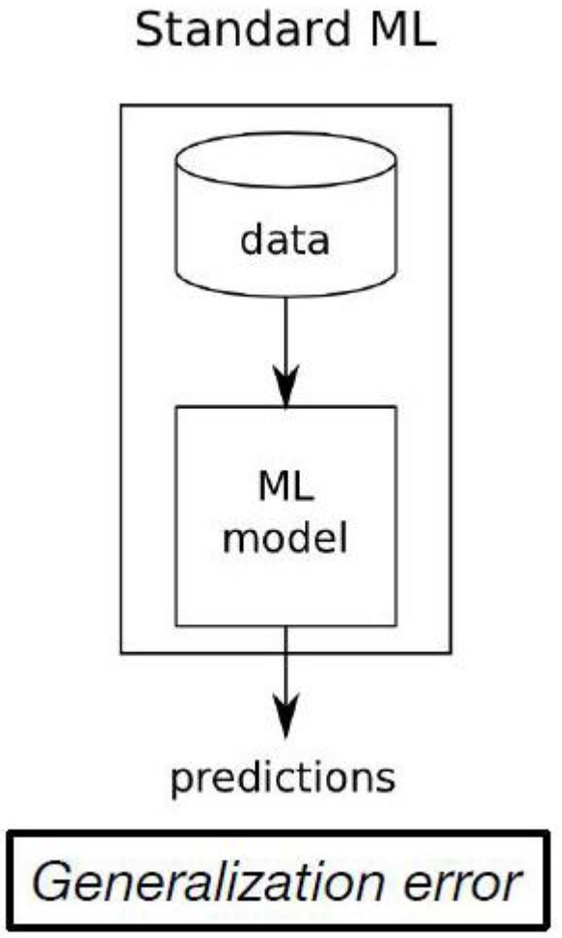




—> model understands the question and correctly identifies the object of interest

(Arras et al., in Press)

# Is the Generalization Error all we need?

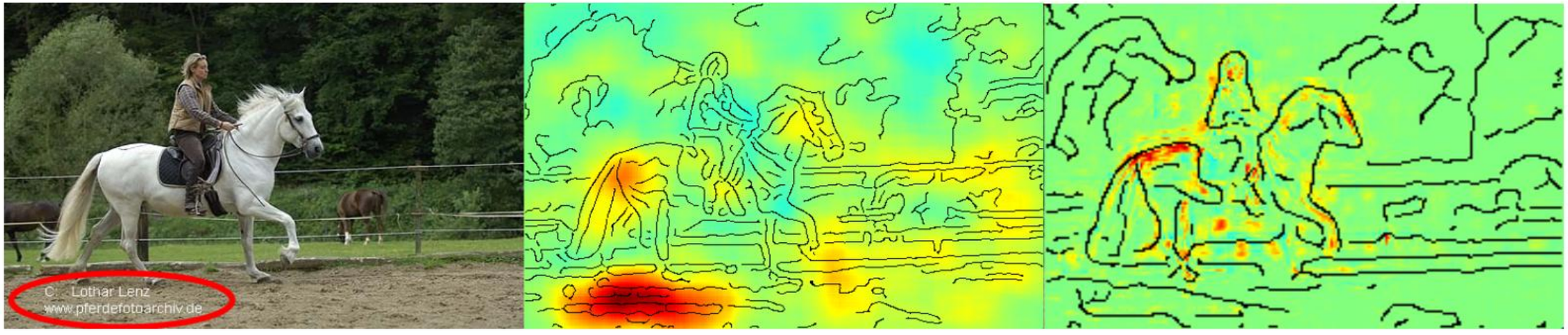


# Application: Comparing Classifiers (Lapuschkin et al CVPR 2016)

Test error for various classes:

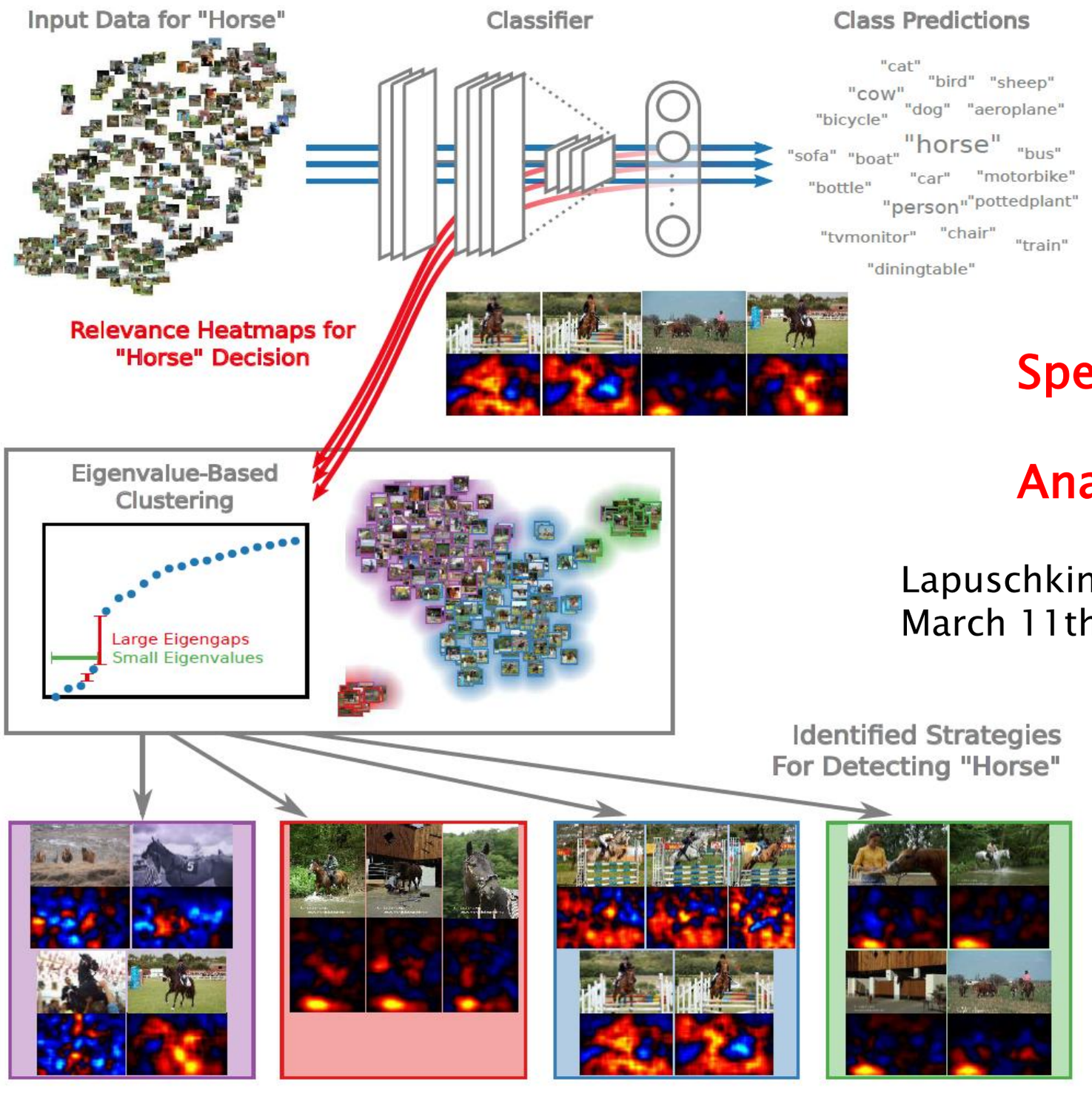
| | aeroplane | bicycle | bird | boat | bottle | bus | car |
|---|---|---|---|---|---|---|---|
| **Fisher** | 79.08% | 66.44% | 45.90% | 70.88% | 27.64% | 69.67% | 80.96% |
| **DeepNet** | 88.08% | 79.69% | 80.77% | 77.20% | 35.48% | 72.71% | 86.30% |
| | **cat** | **chair** | **cow** | **diningtable** | **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% |

Image FV DNN




Xarif

# Explaining problem solving strategies in scale




# Spectral Relevance Analysis (SpRAy)

Lapuschkin et al. Nat Comms, March 11th 2019

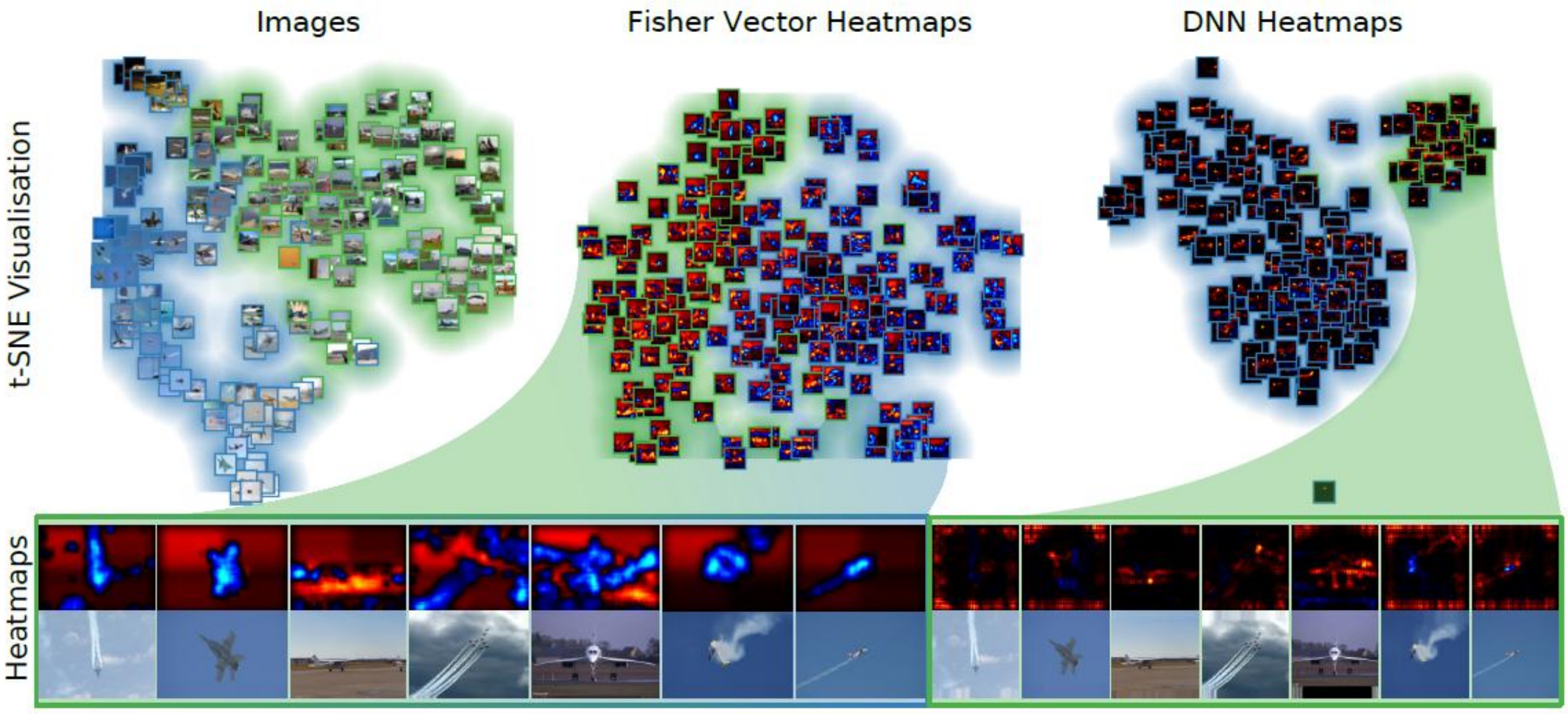



**Figure 28:** Cluster label assignments for class "aeroplane" via SC for input images, FV model relevance maps and DNN relevance maps. Embedding coordinates in $\mathbb{R}^2$ for visualization have been computed on pair-wise distances derived from the weighted affinity matrix $W$ used for SC. The samples at the bottom right (square images) show DNN relevance maps and images with strong reaction of the DNN models to the border padding. FV relevance maps for the same images are shown to the left. Enlarged relevance maps and images are shown without preprocessing.

# 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:*

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

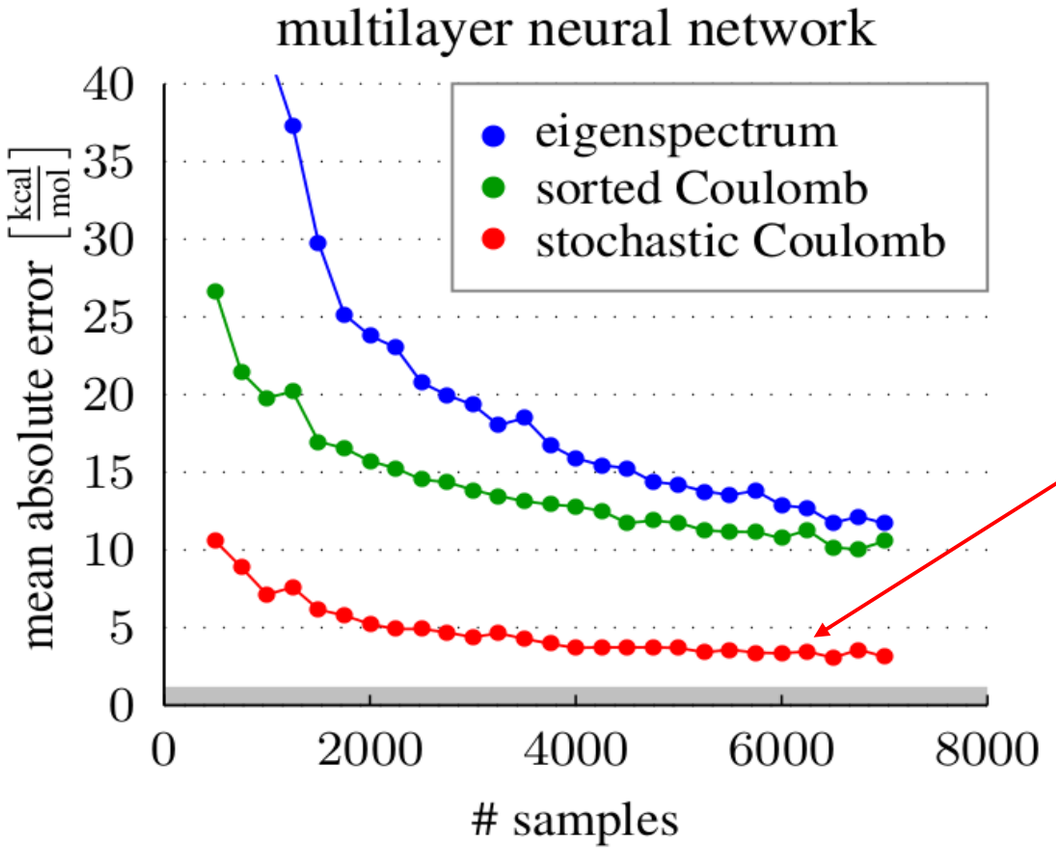
instead of

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

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

[from von Lilienfeld]

# Predicting Energy of small molecules: Results



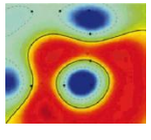

March 2012
Rupp et al., PRL
**9.99 kcal/mol**
(kernels + eigenspectrum)

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

**Now: 0.3 kcal/mol with DTNN and SchNet**

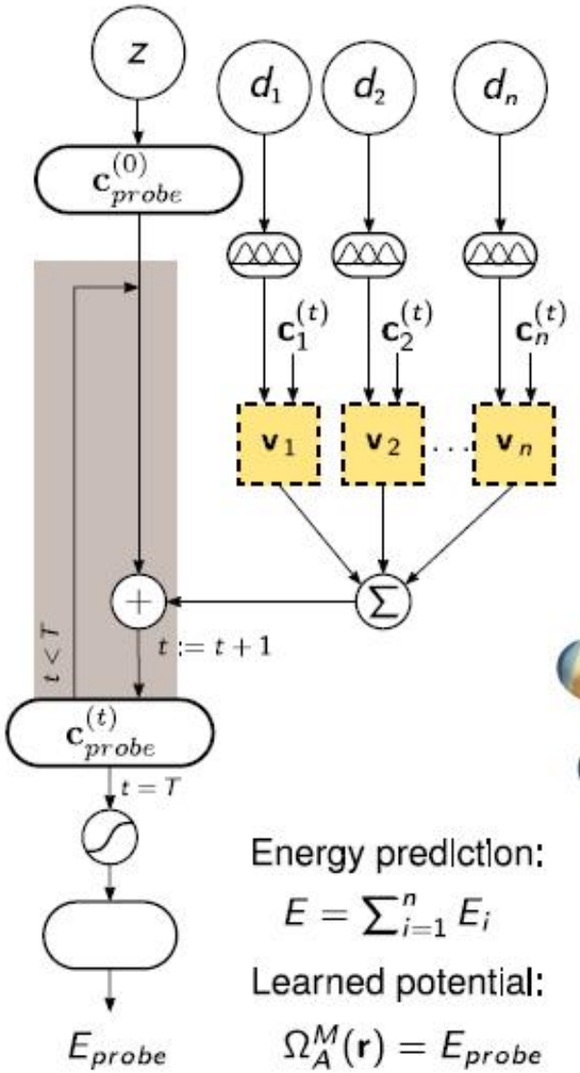
Prediction considered chemically accurate when MAE is below **1 kcal/mol**

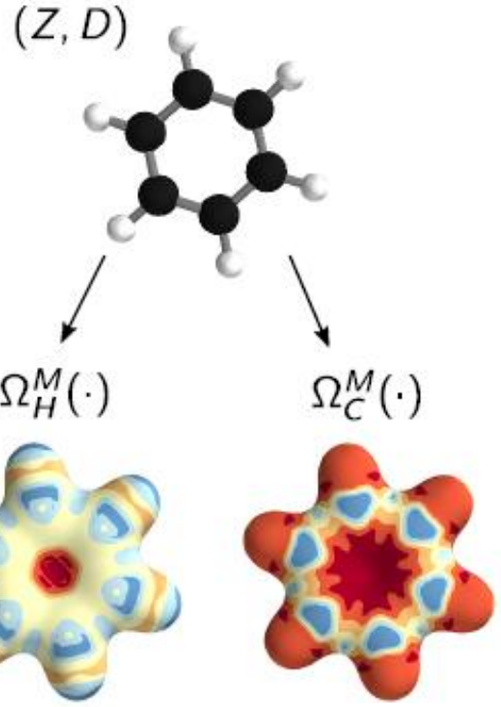



Dataset available at http://quantum-machine.org

# Gaining insights for Physics

# Toward Quantum Chemical Insight



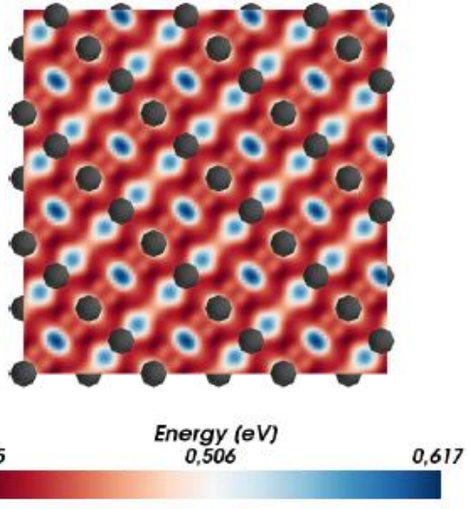




Energy prediction:

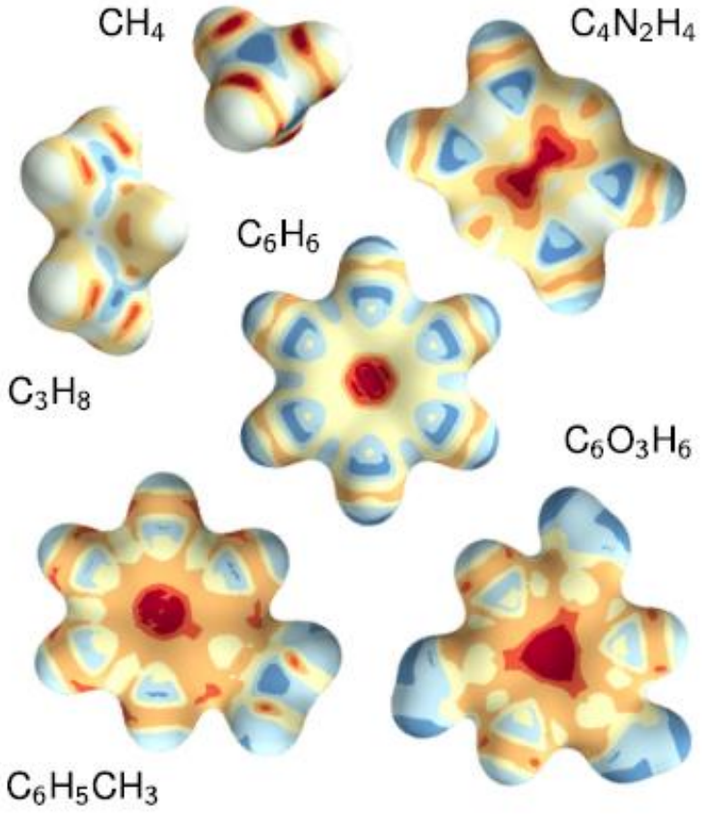
$E = \sum_{i=1}^{n} E_i$

Learned potential:

$\Omega^M_A(\mathbf{r}) = E_{probe}$



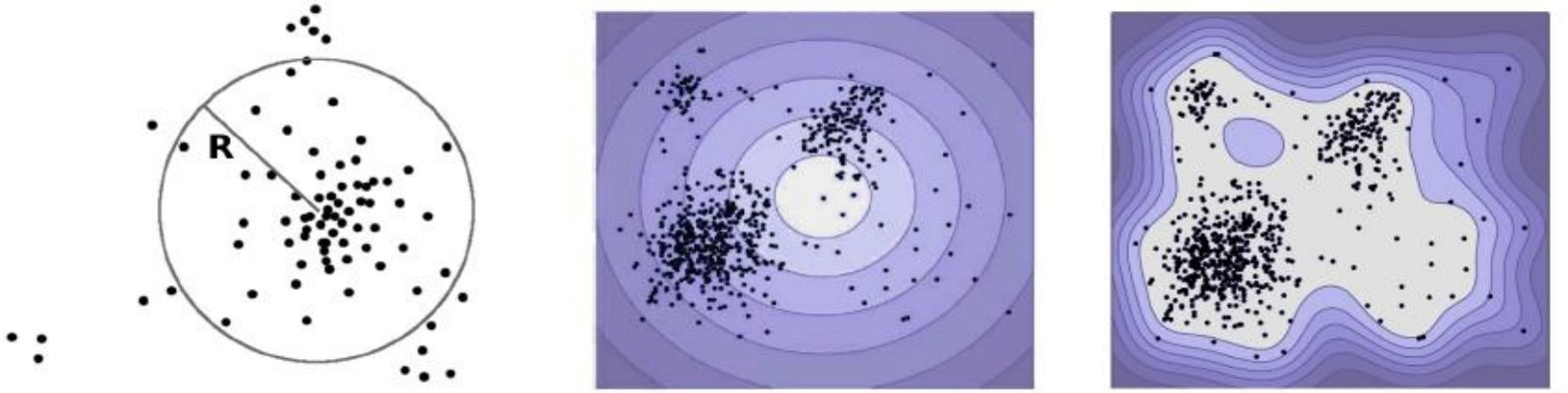

**0.3 kcal/mol with DTNN and SchNet**




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

# XAI for unsupervised learning

# 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

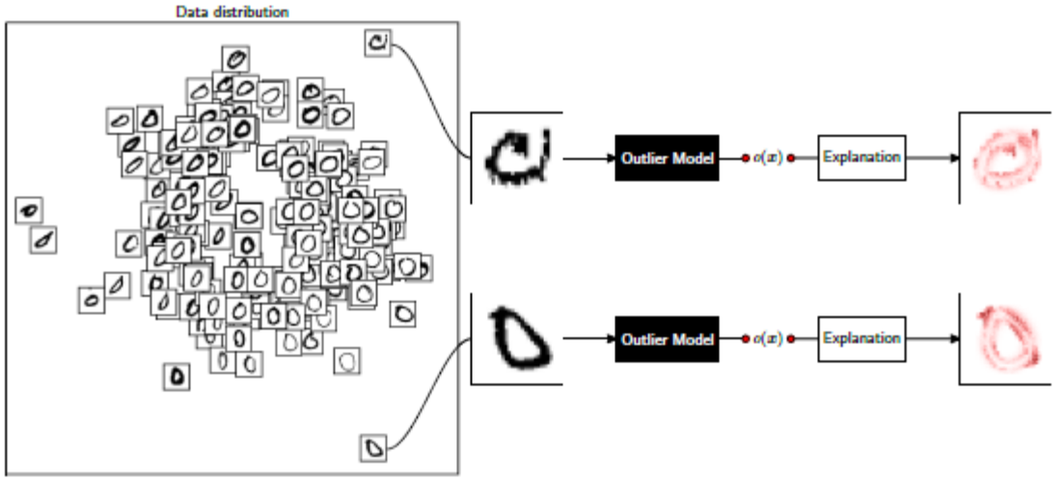



Figure 1: Illustration of the outlier detection and explanation setting. *Left:* Data is generated from an unknown distribution, we are for example interested in potential outliers; *Middle:* Unsupervised machine learning techniques estimate the data generating distribution and assign an outlier score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

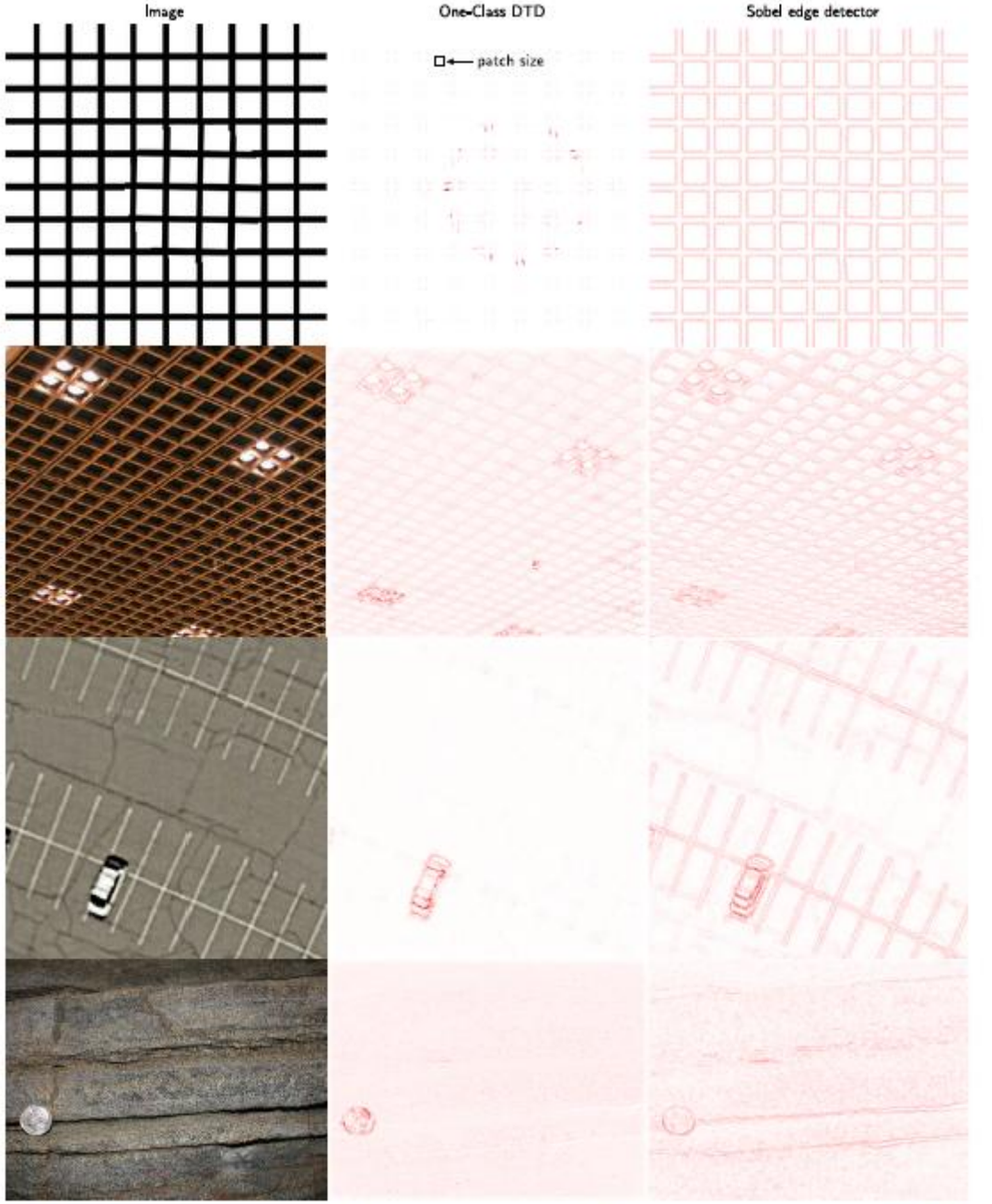


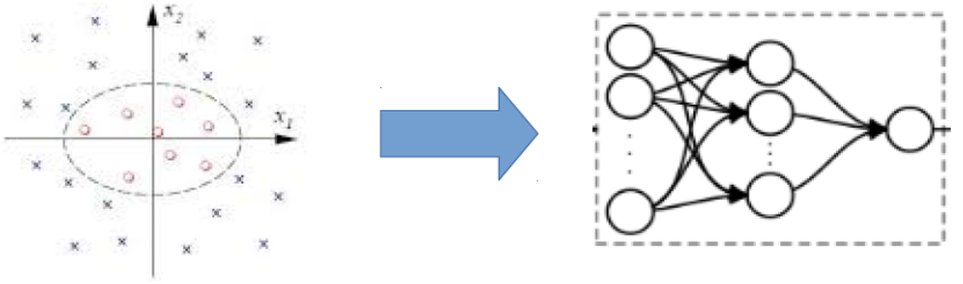

Figure 5: A One-Class SVM is trained on small $7 \times 7$ patches of the very image itself. Parameter $\nu = 0.1$ is set to allow at most 10% outliers. Images from a texture data set [11] (row one, two and four) and PatternNet [61]; top image is altered by us. For every image, we show *Left:* input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

[Kaufmann, Müller, Montavon 2018, 2019]

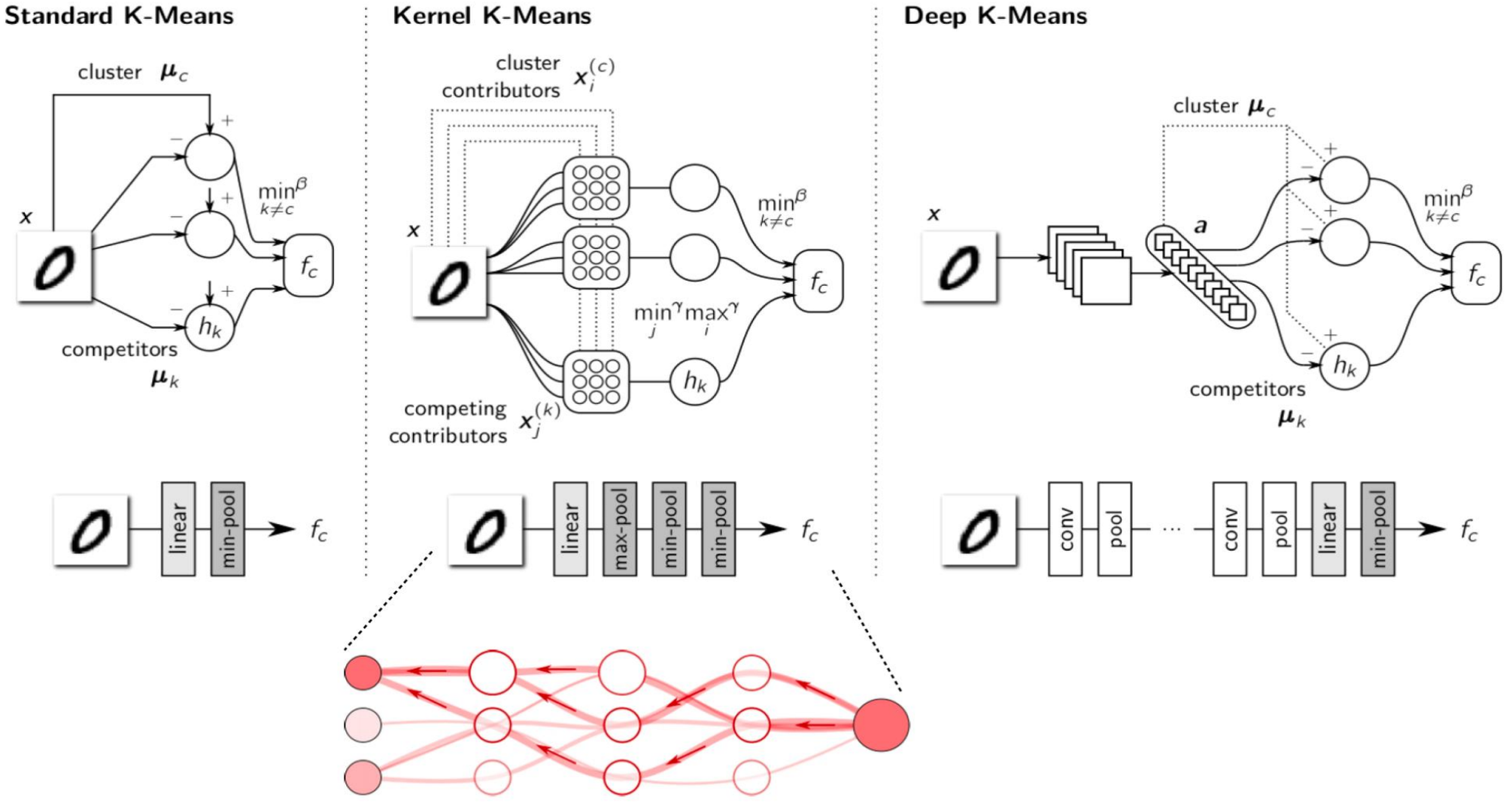
# Interpretable Clustering

# NEON (Neuralization-Propagation)

**NEON's idea:** When the ML model is not a neural network (e.g. a kernel machine), convert it into a neural network first ('neuralize' it).



# Neuralizing K-means




J Kauffmann, M Esders, G Montavon, W Samek, KR Müller. From Clustering to Cluster Explanations via Neural Networks, arXiv:1906.07633, 2019

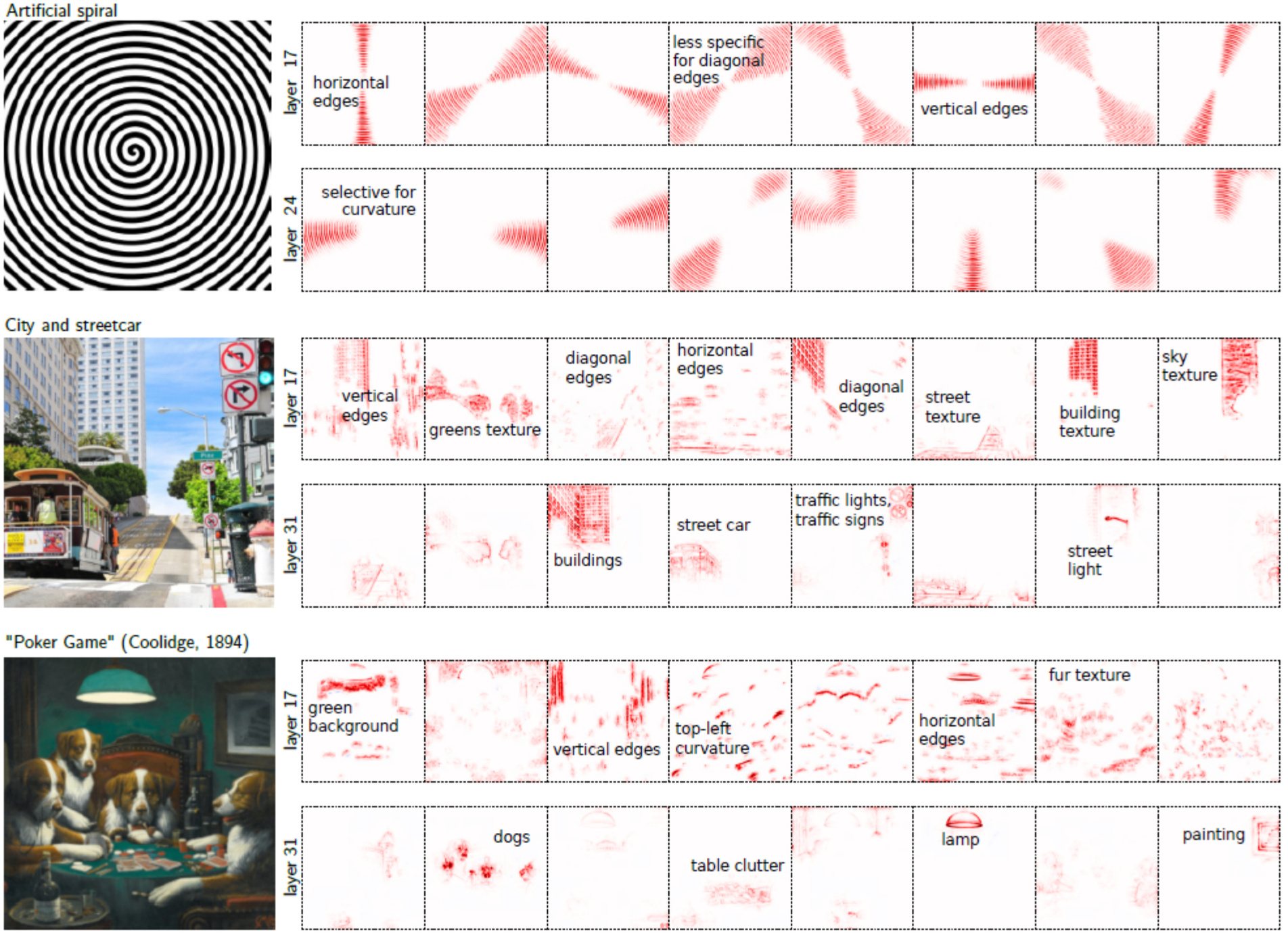



Fig. 7. NEON analysis of images represented at different layers of a deep neural network (pretrained VGG16). K-means clustering with $K = 8$ is performed at these two layers. Each column shows the pixel-contributions for one of these clusters.

## 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
- tool for gaining **insight**

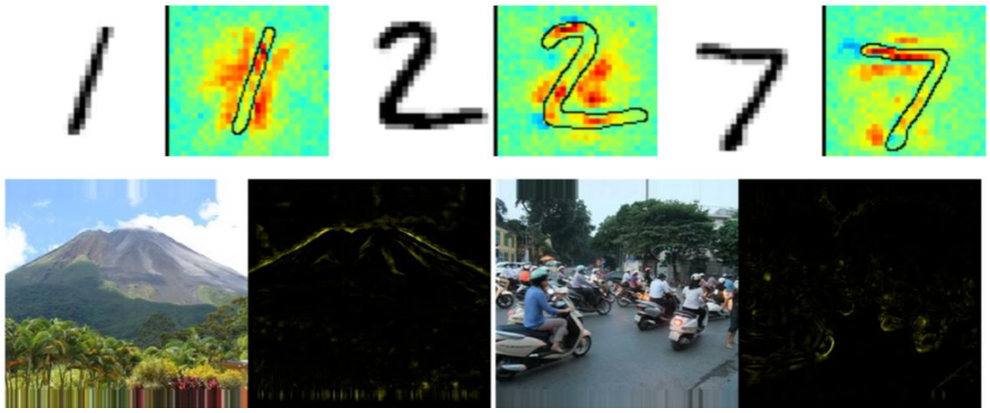
**www.heatmapping.org**

# 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

**New Book:** Samek, Montavon, Vedaldi, Hansen, Müller (eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNAI 11700, Springer (2019) (**coming up in 1 month**)

**Keras Explanation Toolbox**
**https://github.com/albermax/innvestigate**

Fraunhofer Heinrich Hertz Institute

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

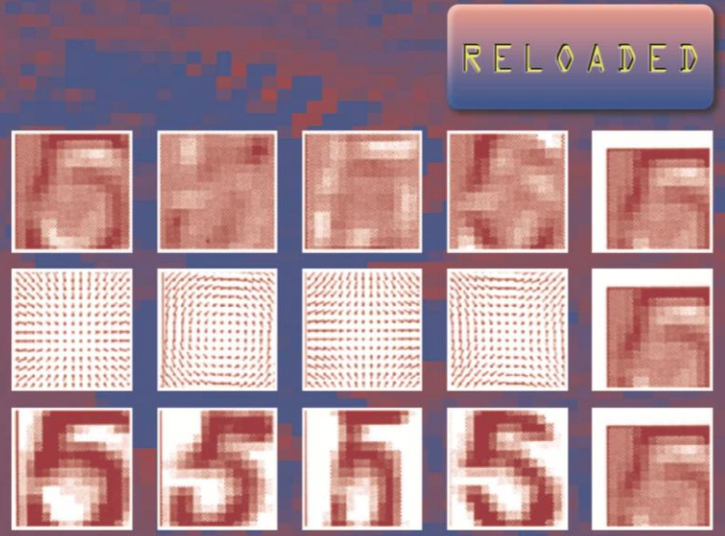
State-of-the-Art Survey

LNCS 7700

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

# Neural Networks: Tricks of the Trade

Second Edition



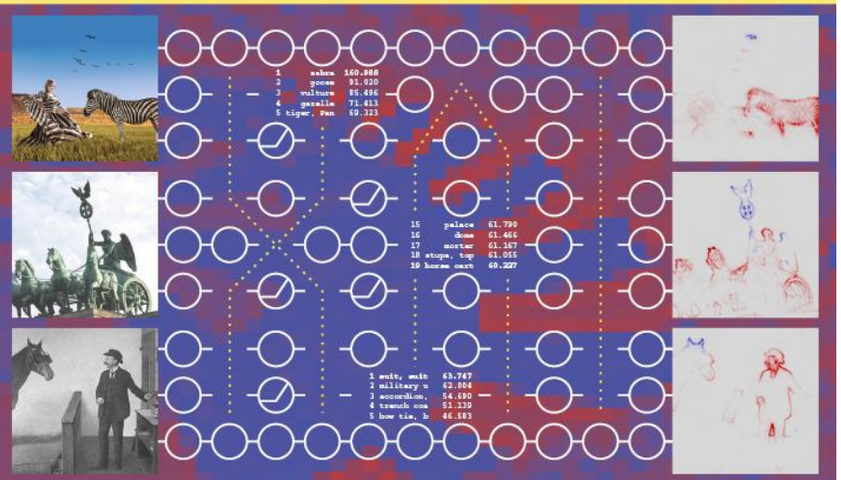

Springer

State-of-the-Art Survey

LNAI 11700

Wojciech Samek · Grégoire Montavon · Andrea Vedaldi · Lars Kai Hansen · Klaus-Robert Müller (Eds.)

# Explainable AI: Interpreting, Explaining and Visualizing Deep Learning



Springer

# Further Reading I


Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.R., Dähne, S. and Kindermans, P.J., 2019. iNNvestigate neural networks!. Journal of Machine Learning Research, 20(93), pp.1-8.

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, e0130140 (7).

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. The Journal of Machine Learning Research, 11, 1803-1831.

Binder et al. Machine Learning for morpho-molecular Integration, *arXiv:1805.11178* (2018)

Blankertz, B., Curio, G. and Müller, K.R., 2002. Classifying single trial EEG: Towards brain computer interfacing. In Advances in neural information processing systems (pp. 157-164).

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.R. and Curio, G., 2007. The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. NeuroImage, 37(2), pp.539-550.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. and Muller, K.R., 2007. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal processing magazine, 25(1), pp.41-56.

Blankertz, B., Lemm, S., Treder, M., Haufe, S. and Müller, K.R., 2011. Single-trial analysis and classification of ERP components—a tutorial. NeuroImage, 56(2), pp.814-825.

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.

Brockherde, F., Vogt, L., Li, L., Tuckerman, M.E., Burke, K. and Müller, K.R., 2017. Bypassing the Kohn-Sham equations with machine learning. Nature communications, 8(1), p.872.

# Further Reading II


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, H.E., Müller, K.R. and Tkatchenko, A., 2018. Towards exact molecular dynamics simulations with machine-learned force fields. Nature communications, 9(1), p.3887.

Dornhege, G., Millan, J.D.R., Hinterberger, T., McFarland, D.J. and Müller, K.R. eds., 2007. Toward brain-computer interfacing. MIT press.

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.

Horst, F., Lapuschkin, S., Samek, W., Müller, K.R. and Schöllhorn, W.I., 2019. Explaining the unique nature of individual gait patterns with deep learning. Scientific reports, 9(1), p.2391.

Kauffmann, J., Müller, K.R. and Montavon, G., 2018. Towards explaining anomalies: A deep taylor decomposition of one-class models. arXiv preprint arXiv:1805.06230.

# Further Reading III


Klauschen, F., Müller, K.R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., de Maria, S., Badve, S. and Michiels, S., 2018, October. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Seminars in cancer biology (Vol. 52, pp. 151-157).

Lemm, S., Blankertz, B., Dickhaus, T. and Müller, K.R., 2011. Introduction to machine learning for brain imaging. Neuroimage, 56(2), pp.387-399.

Lapuschkin, 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) 2912-2920 (2016).

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), p.1096.

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.

Muller, K.R., Anderson, C.W. and Birch, G.E., 2003. Linear and nonlinear methods for brain-computer interfaces. IEEE transactions on neural systems and rehabilitation engineering, 11(2), pp.165-169.

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., Orr, G. & Müller, K. R. (2012). Neural Networks: Tricks of the Trade, Springer LNCS 7700. Berlin Heidelberg.

# Further Reading IV


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., Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222 (2017)

Montavon, G., Samek, W., & Müller, K. R., Methods for interpreting and understanding deep neural networks, Digital Signal Processing, 73:1-5, (2018).

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.

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)

K.T. Schütt, H.E. Sauceda, , P.J. Kindermans, , A. Tkatchenko and K.R. Müller, SchNet–A deep learning architecture for molecules and materials. The Journal of Chemical Physics, 148(24), p.241722. (2018)

Samek, W., Binder, A., Montavon, G., Lapuschkin, S. and Müller, K.R., Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11), pp.2660-2673 (2017)

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNAI 11700, Springer (2019)

Sturm, I., Lapuschkin, S., Samek, W. and Müller, K.R., 2016. Interpretable deep neural networks for single-trial EEG classification. Journal of neuroscience methods, 274, pp.141-145.