Tutorial on Methods for Interpreting and Understanding Deep Neural Networks

W. Samek, G. Montavon, K.-R. Müller

Part 3: Applications & Discussion

Recap: Layer-wise Relevance Propagation (LRP)

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

input $\boldsymbol{x}$

Classifier

Rooster

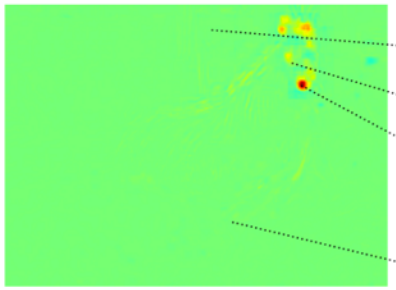
prediction $f(\boldsymbol{x})$

Explain prediction

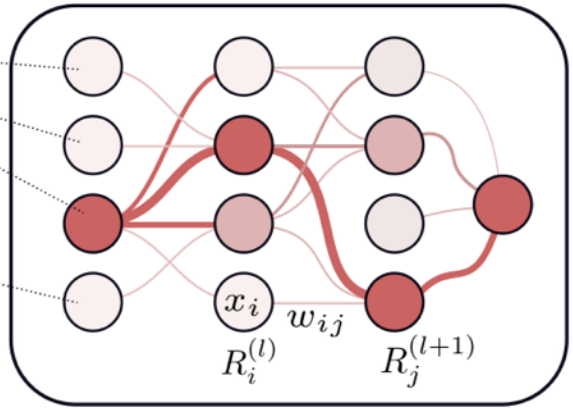
(how much each pixel contributes to prediction)

Idea: Decompose function

$$\sum_i R_i = f(x)$$



heatmap



"every neuron gets it's share of relevance depending on activation and strength of connection."

redistribute $f(\boldsymbol{x})$

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

Recap: Layer-wise Relevance Propagation (LRP)

Classification

cat

rooster

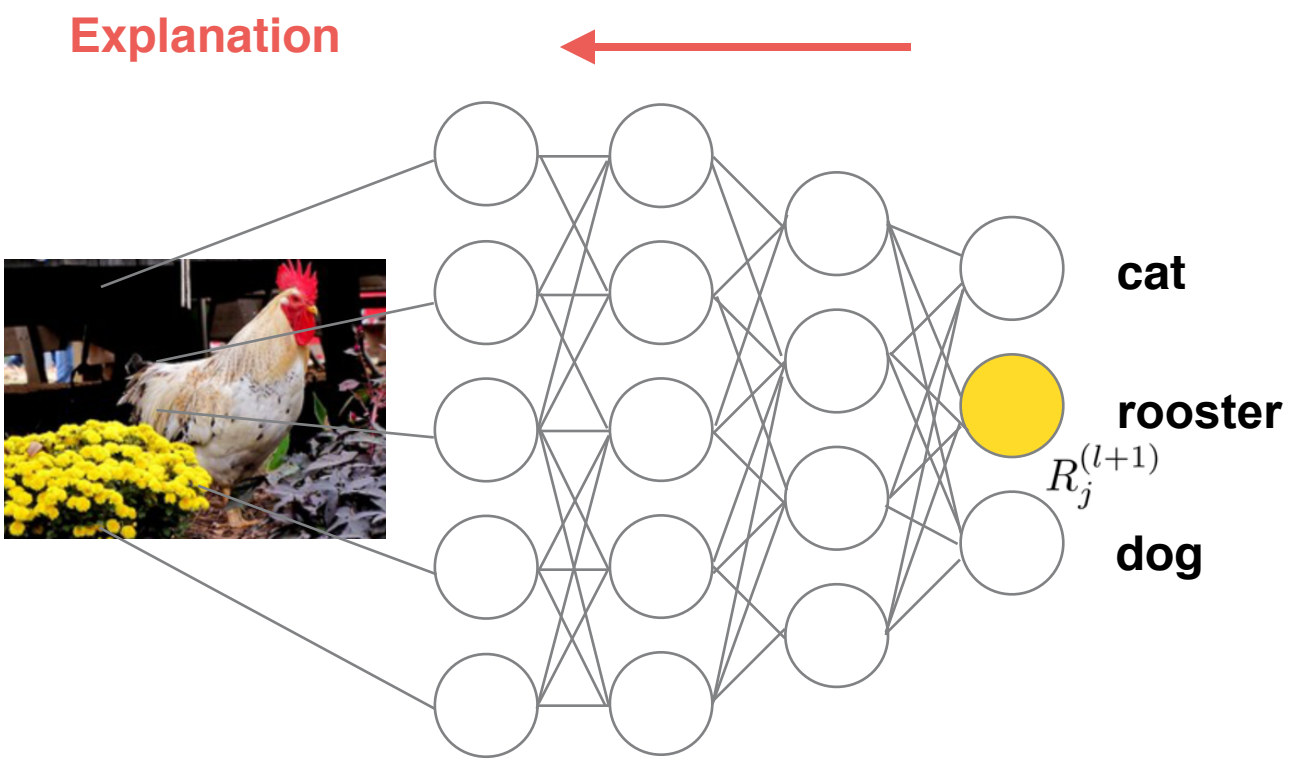
dog

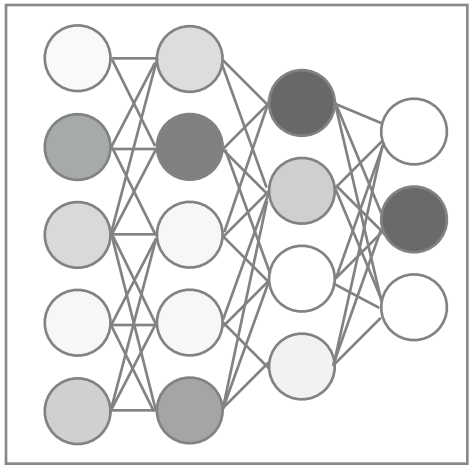
w_{ij}

x_i x_j

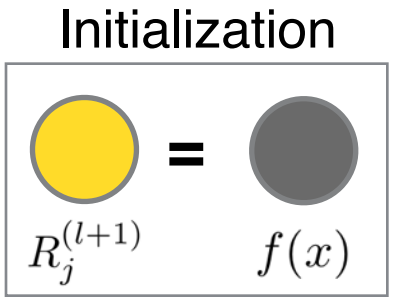
$$x_j = \sigma(\sum_i x_i w_{ij} + b_j)$$

Recap: Layer-wise Relevance Propagation (LRP)

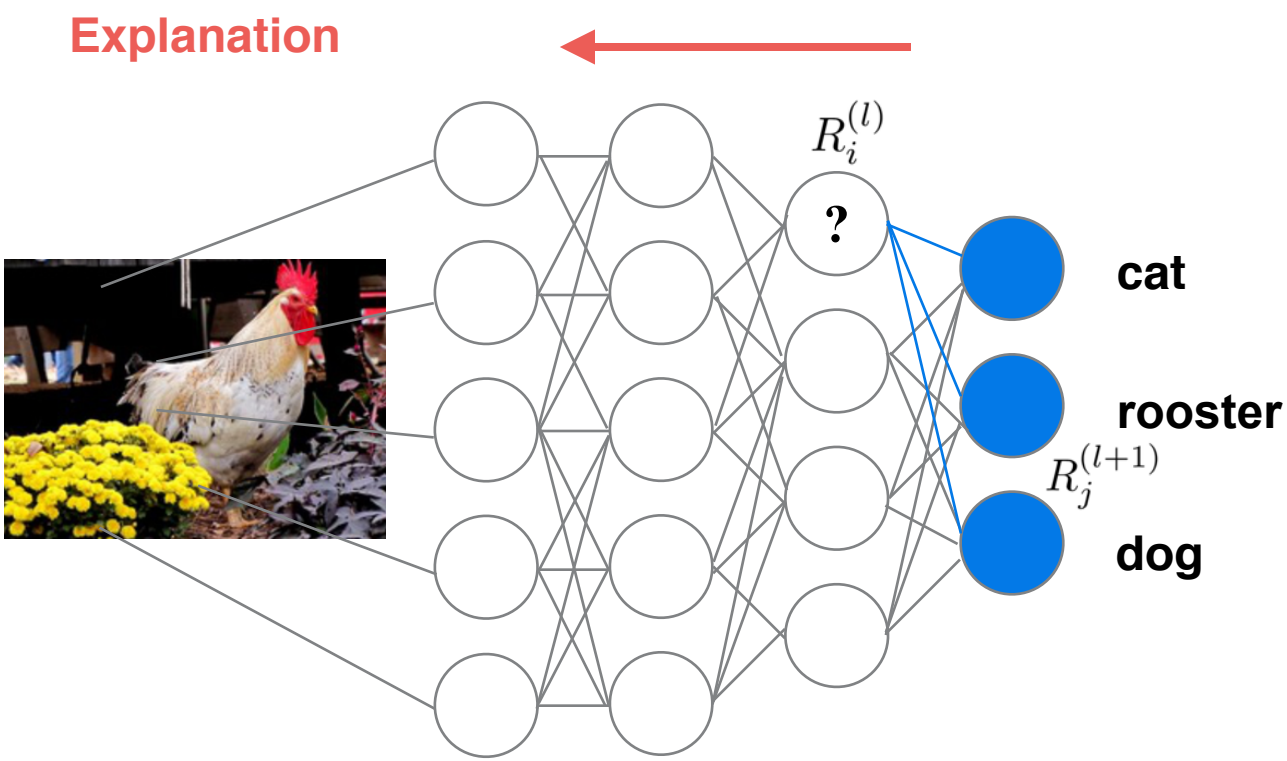


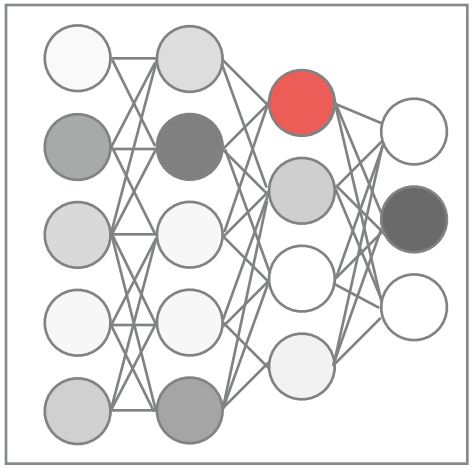


Idea: Backpropagate “relevances”



Recap: Layer-wise Relevance Propagation (LRP)



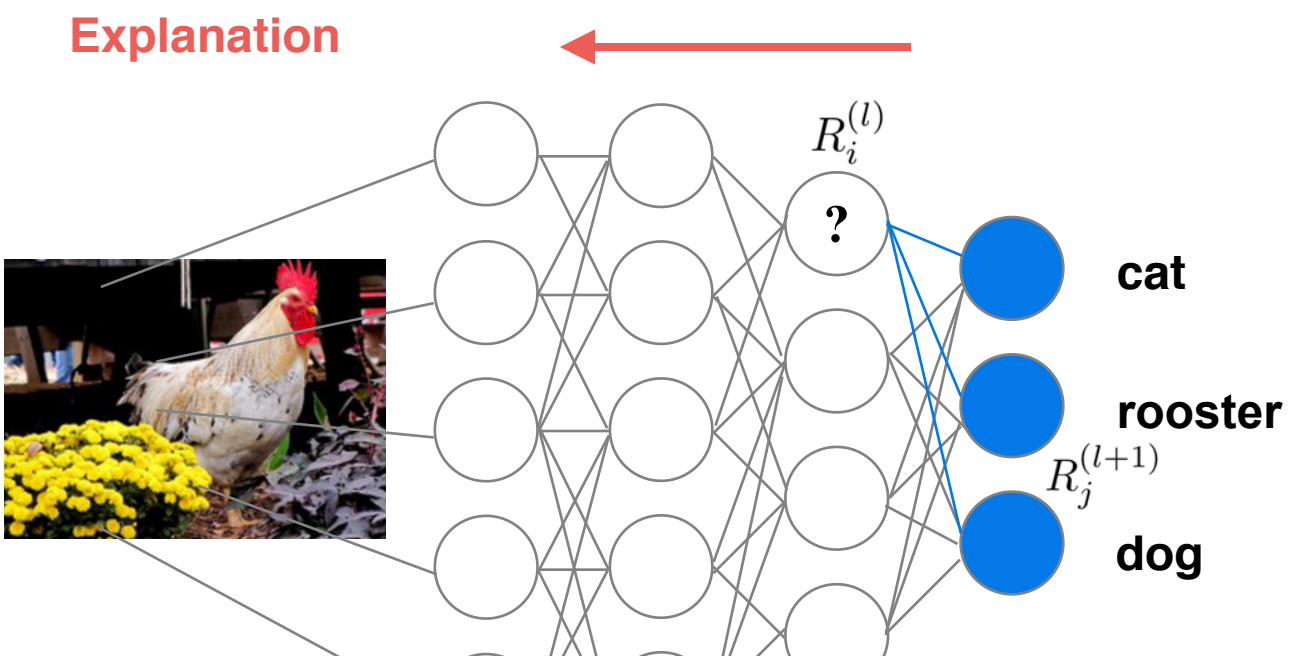


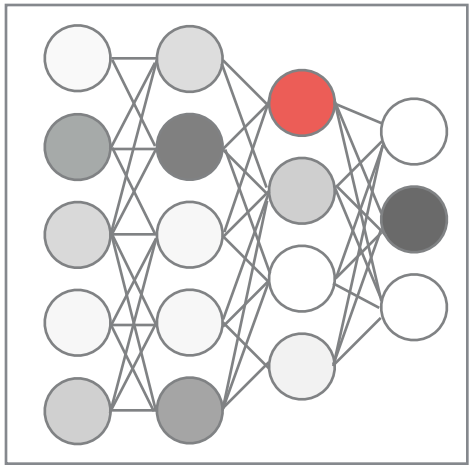
Simple LRP rule (Bach et al. 2015)

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

Every neuron gets its "share" of the redistributed relevance

Recap: Layer-wise Relevance Propagation (LRP)





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

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

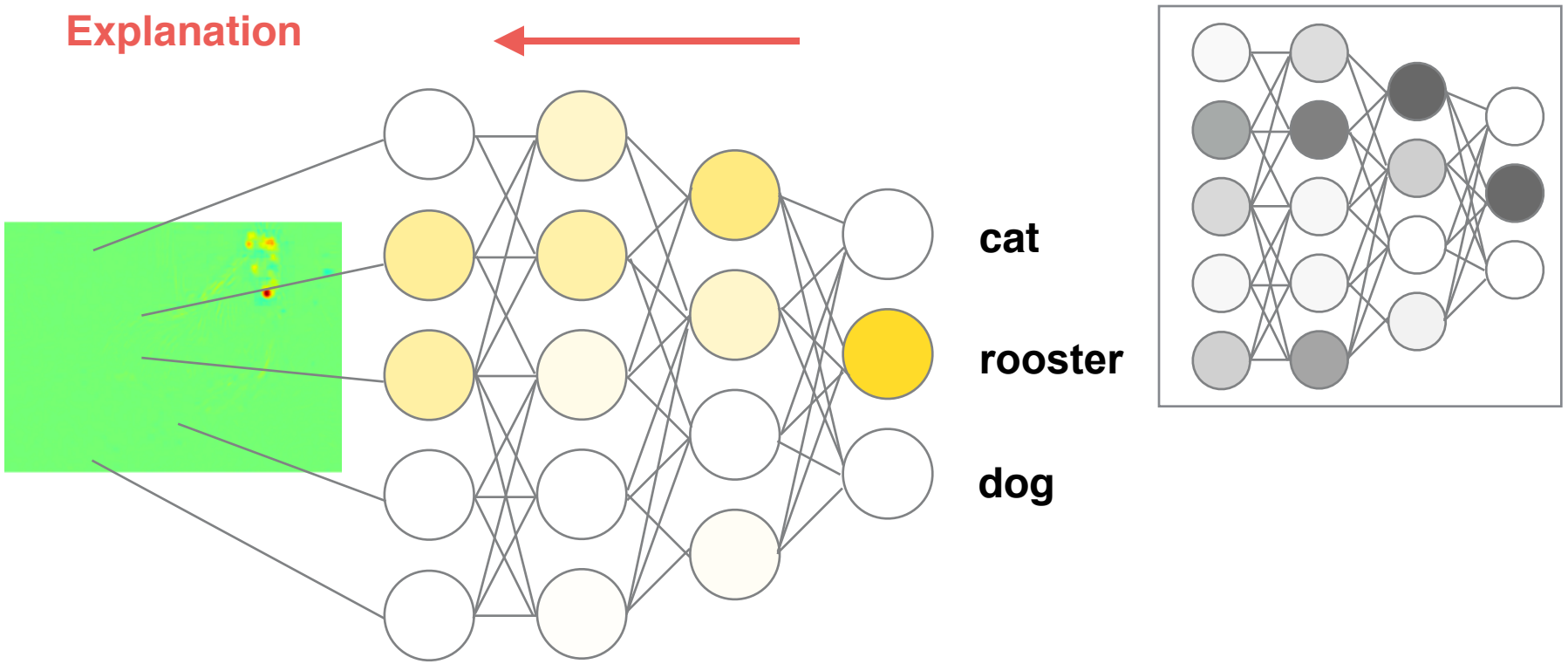
Theoretical interpretation
Deep Taylor Decomposition
(Montavon et al., 2017)

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

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

where $\alpha + \beta = 1$

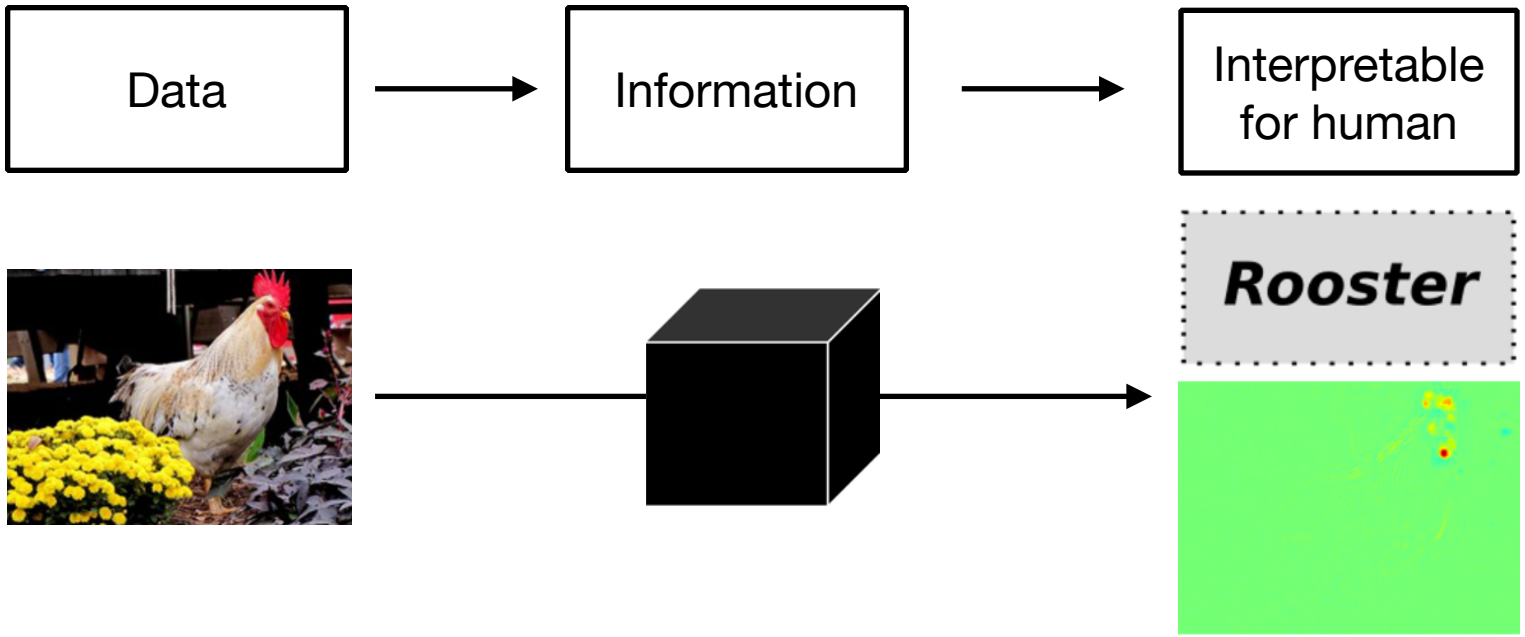
Recap: Layer-wise Relevance Propagation (LRP)



Layer-wise relevance conservation

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

Explanations and now ?



How good is the explanation ?

What can we do with it ?

Explanations and now ?

How good is the explanation ?

- Objective measure of quality
- Compare explanation methods

What can we do with it ?

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

...

Measuring Quality of Explanations

Heatmap quality depends on

classifier (better performance -> better heatmap ?)

- no guarantee that “meaningful” for humans

explanation method

- need objective measure to compare explanation methods

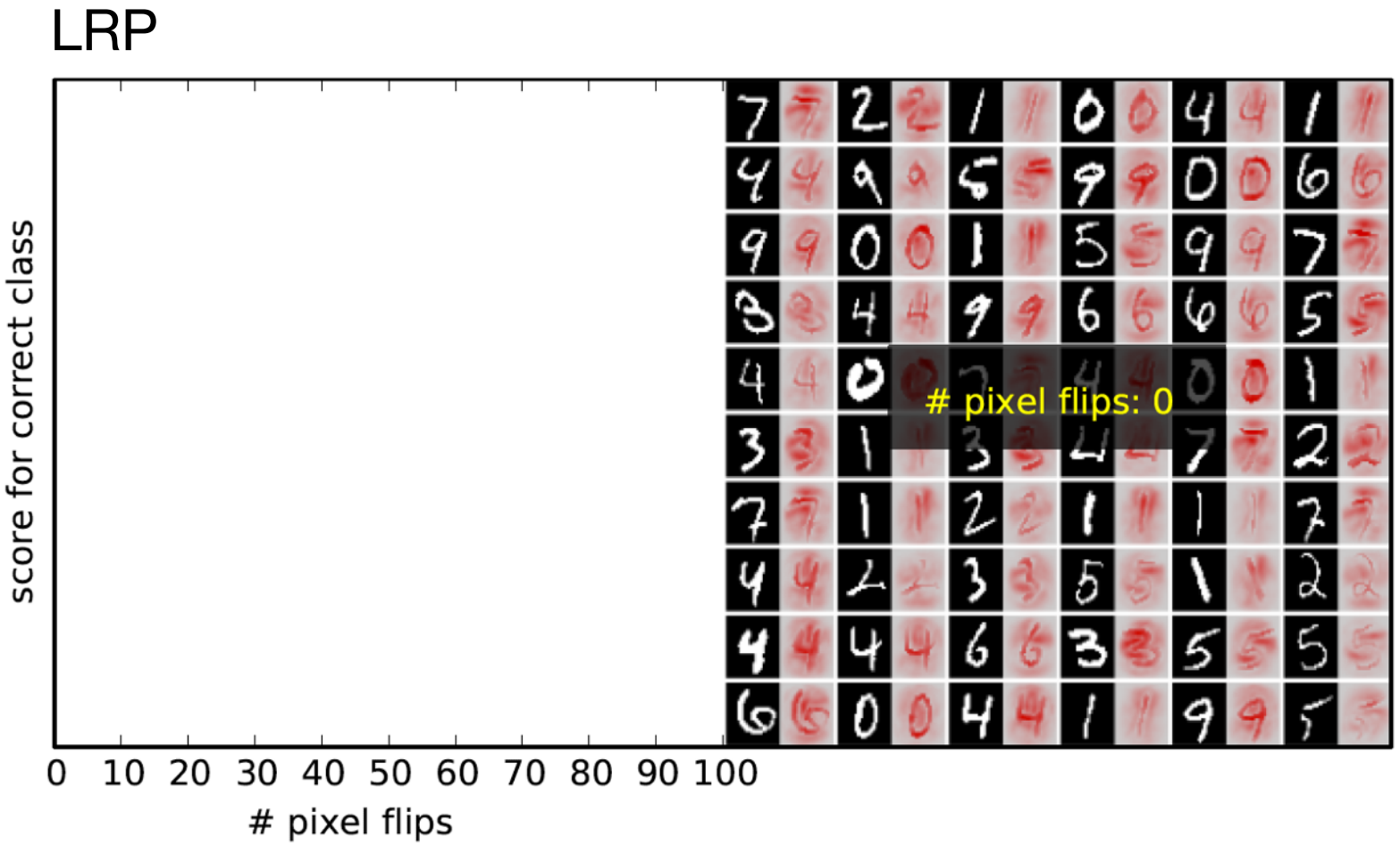
“Pixel flipping”

- good heatmap assigns high relevance to truly relevant pixels.
- destroying these relevant pixels will strongly decrease classification score.
- by measuring this decrease we can assess the quality of heatmaps.

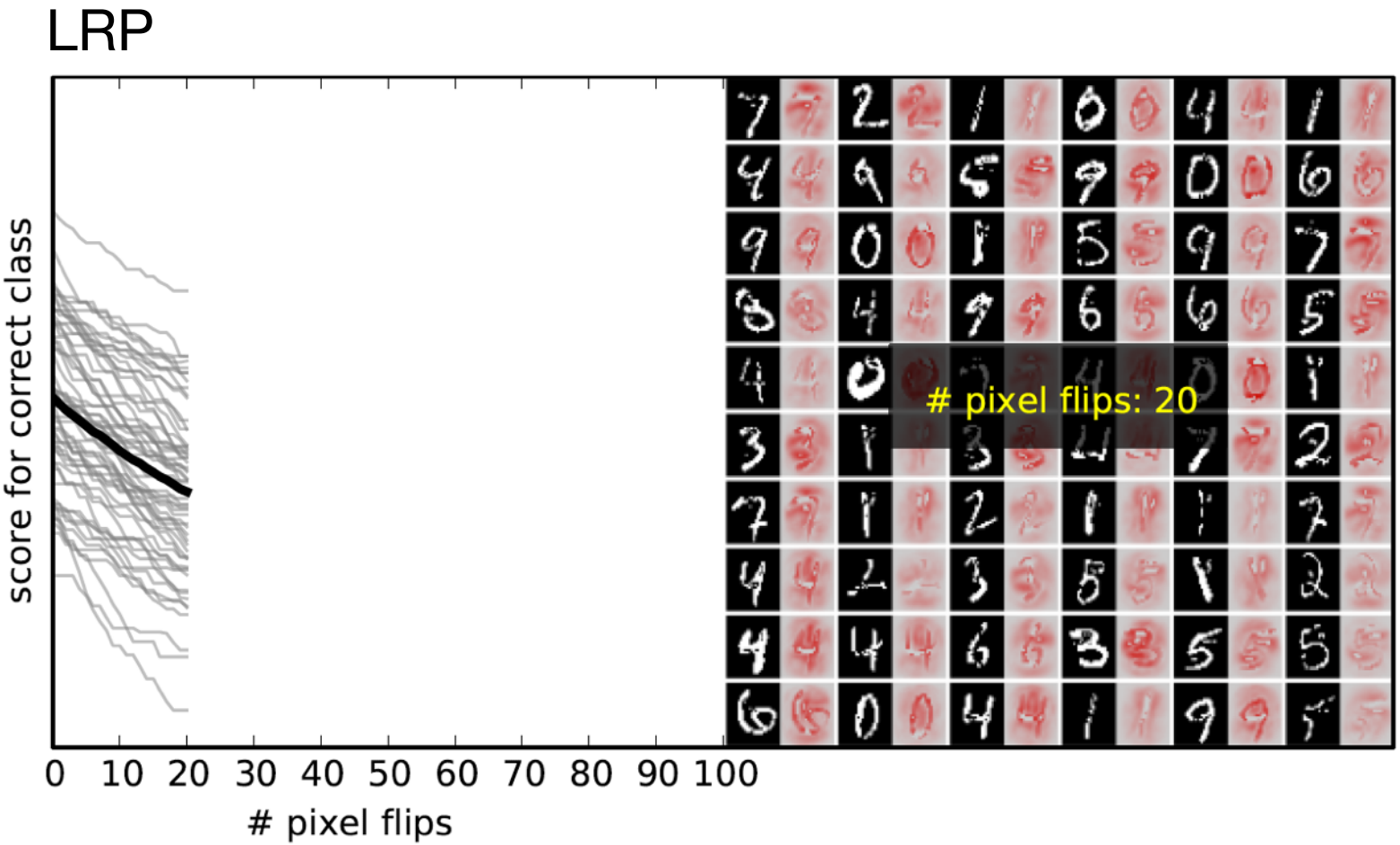
Algorithm (Pixel Flipping)

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

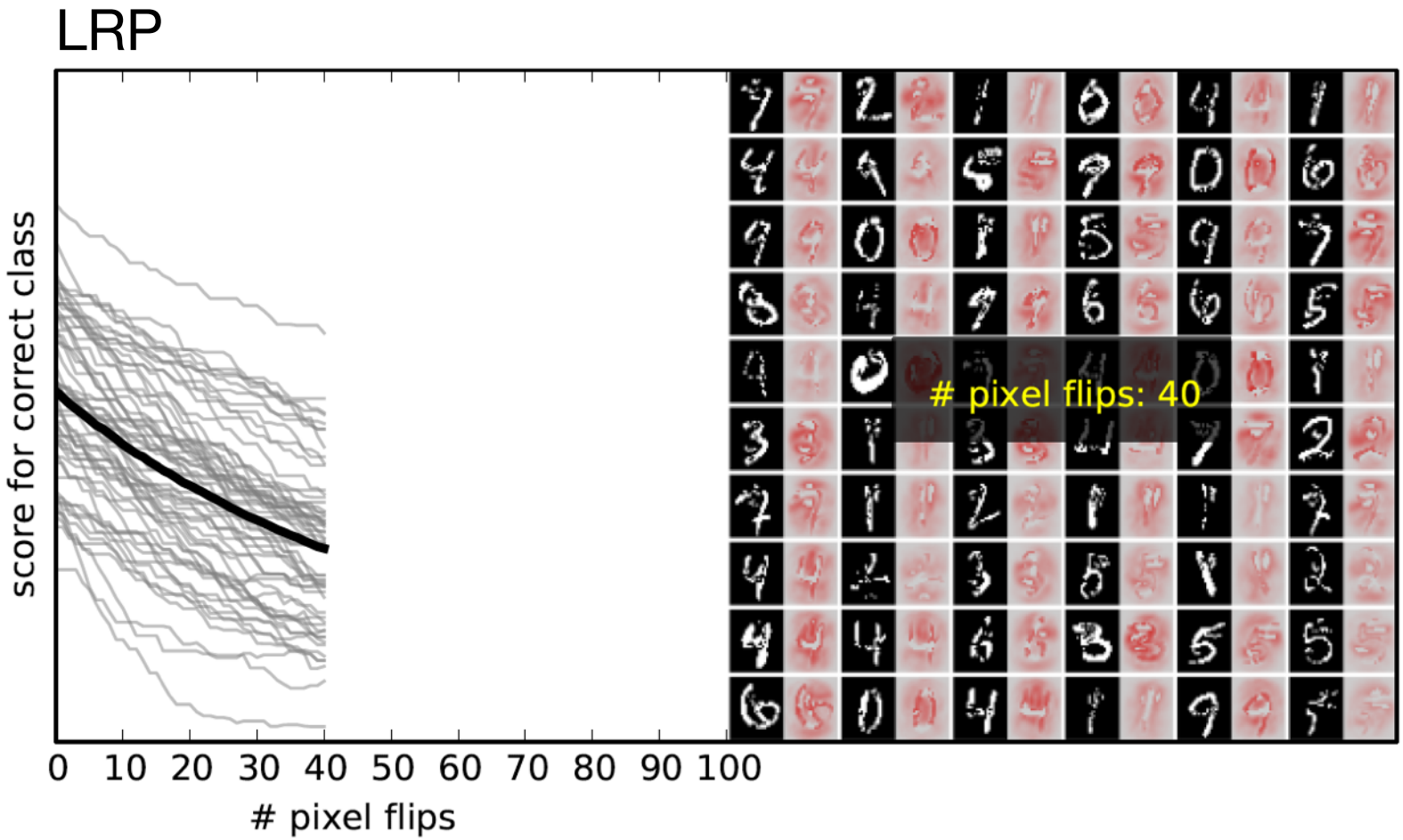
Compare Explanation Methods



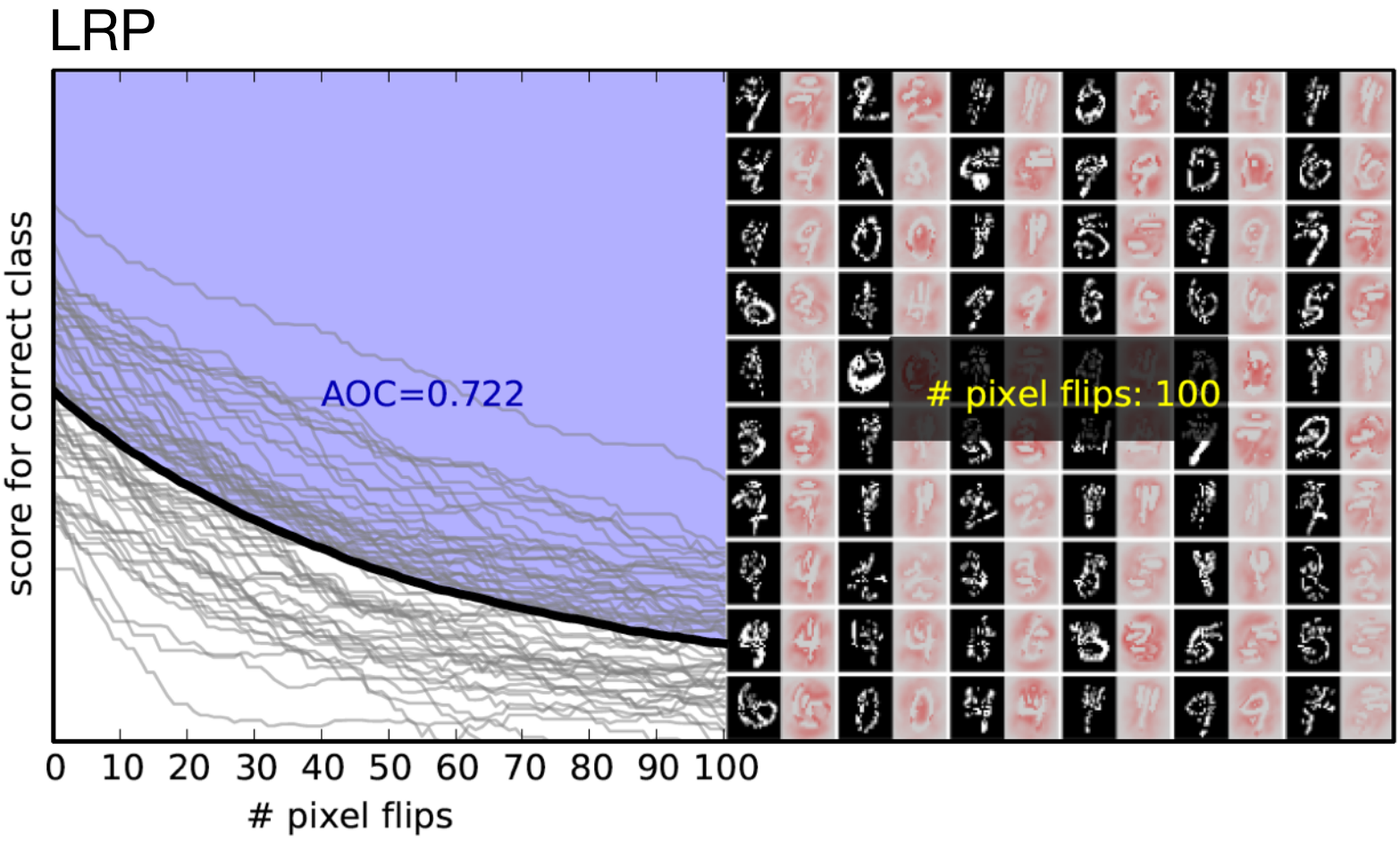
Compare Explanation Methods



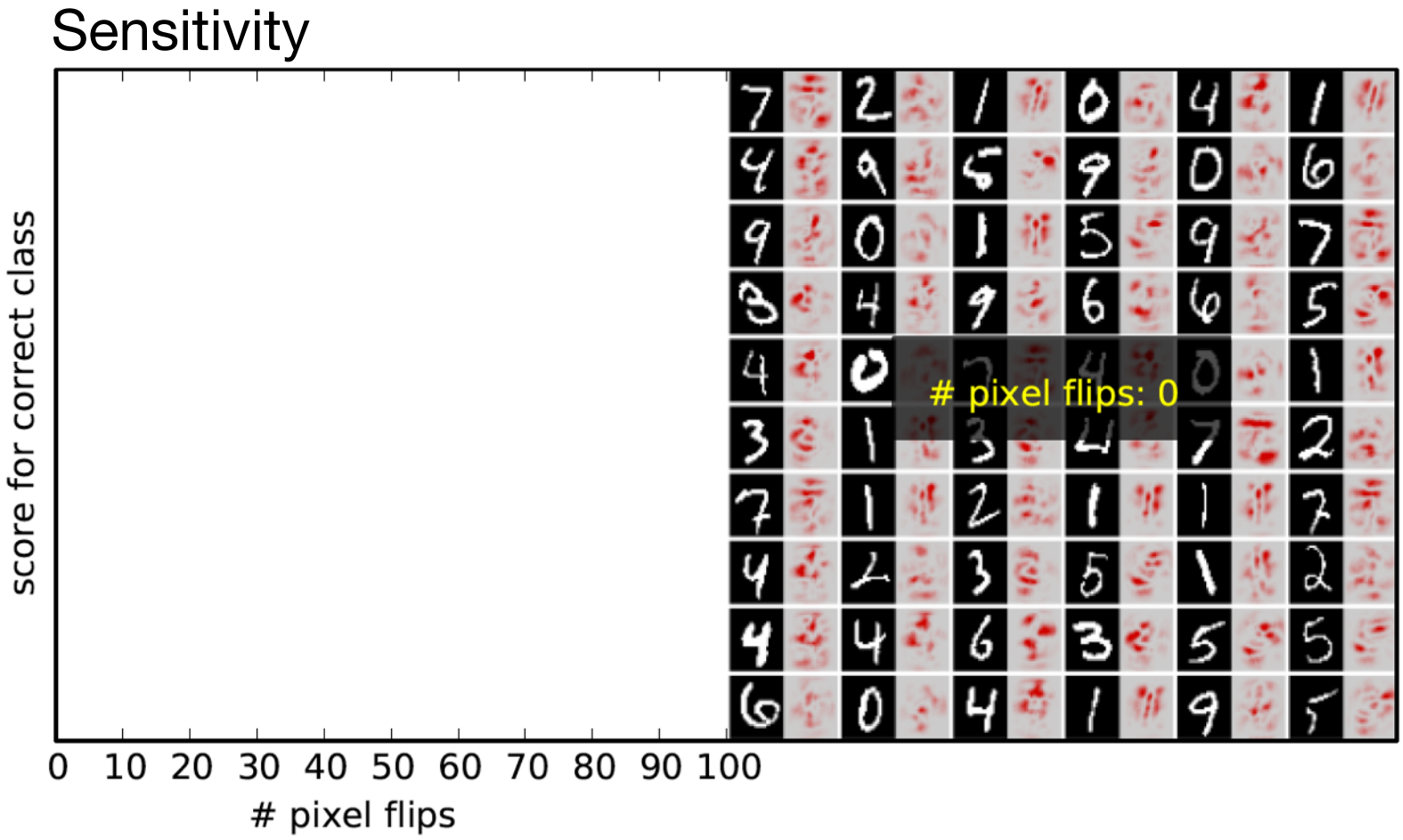
Compare Explanation Methods



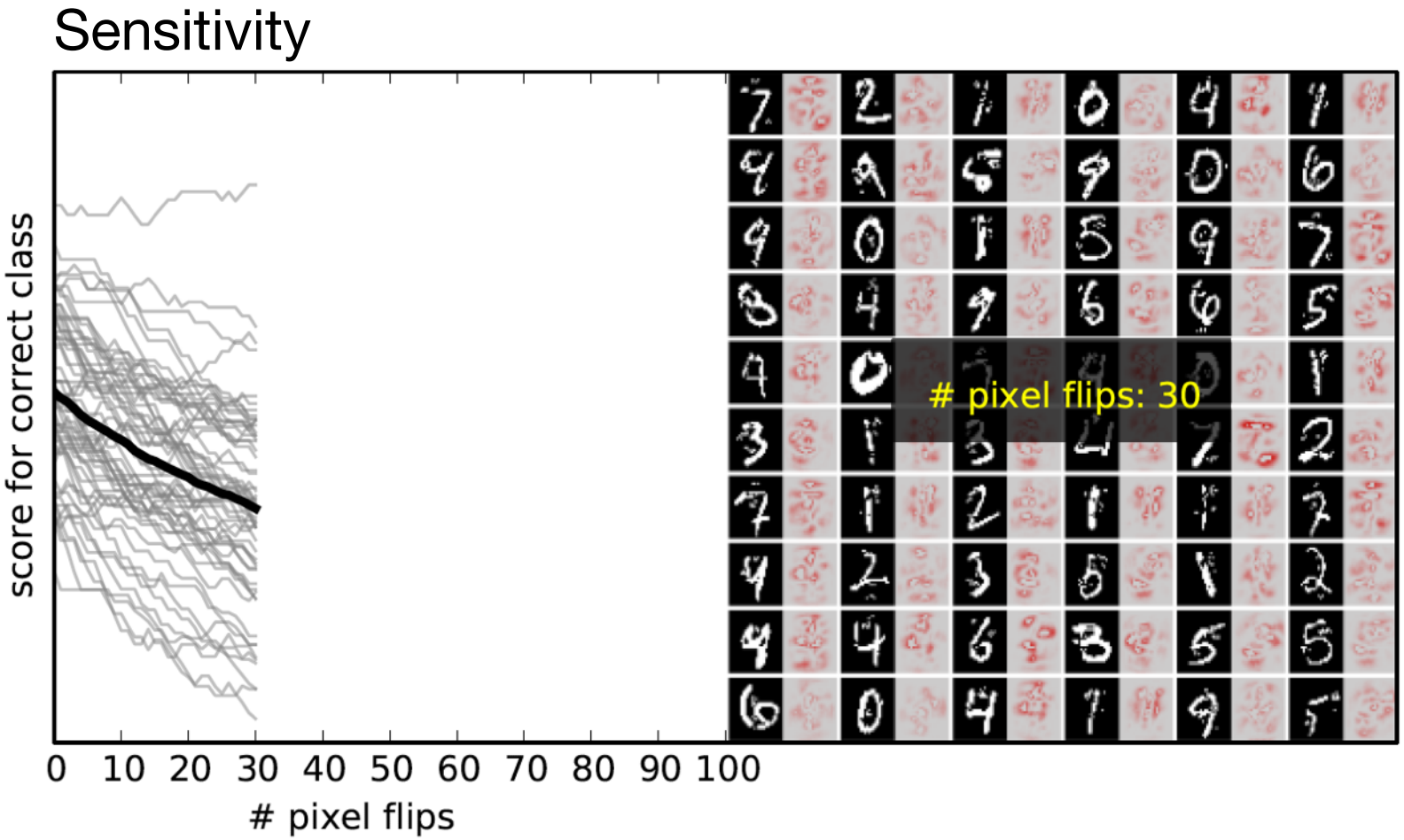
Compare Explanation Methods



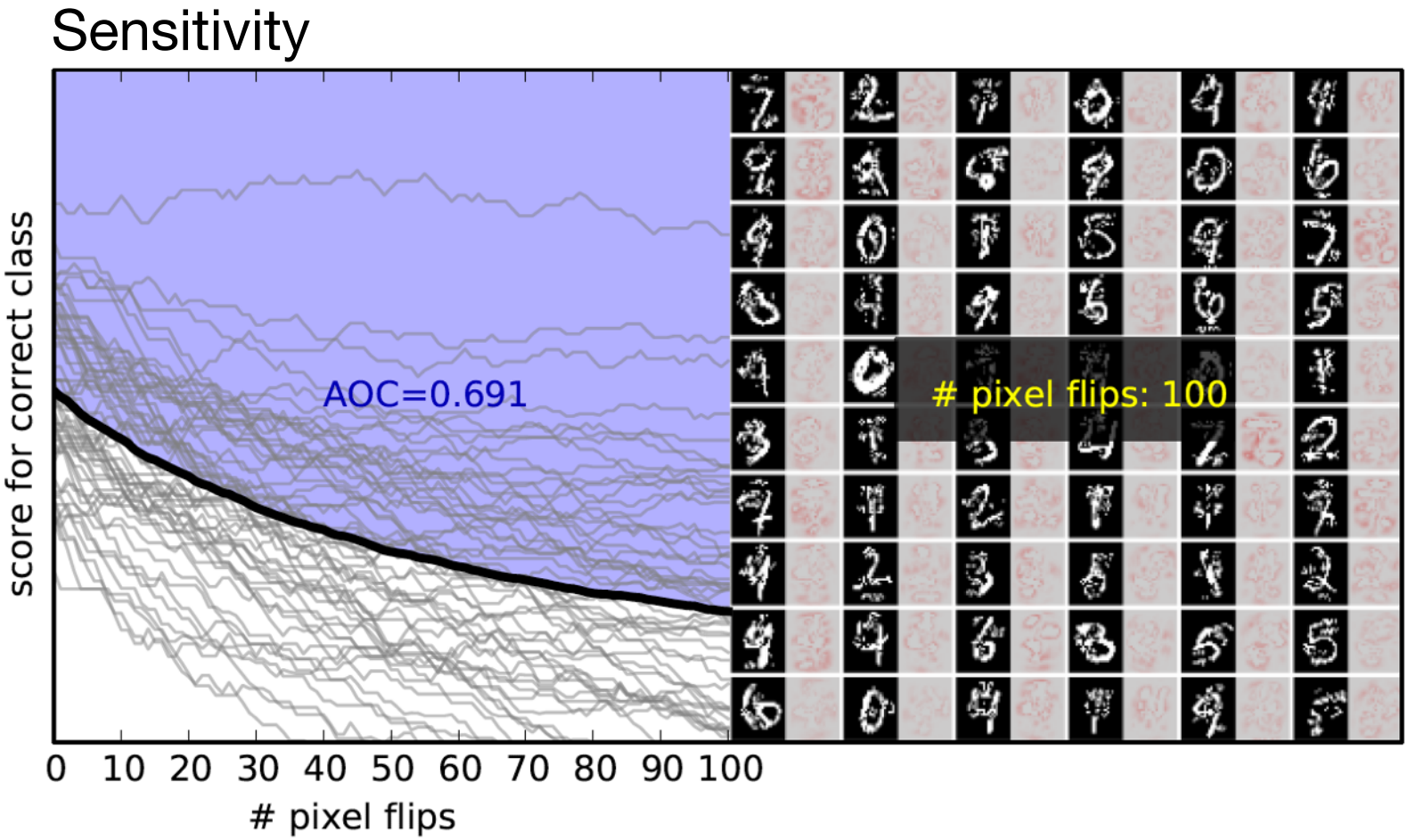
Compare Explanation Methods



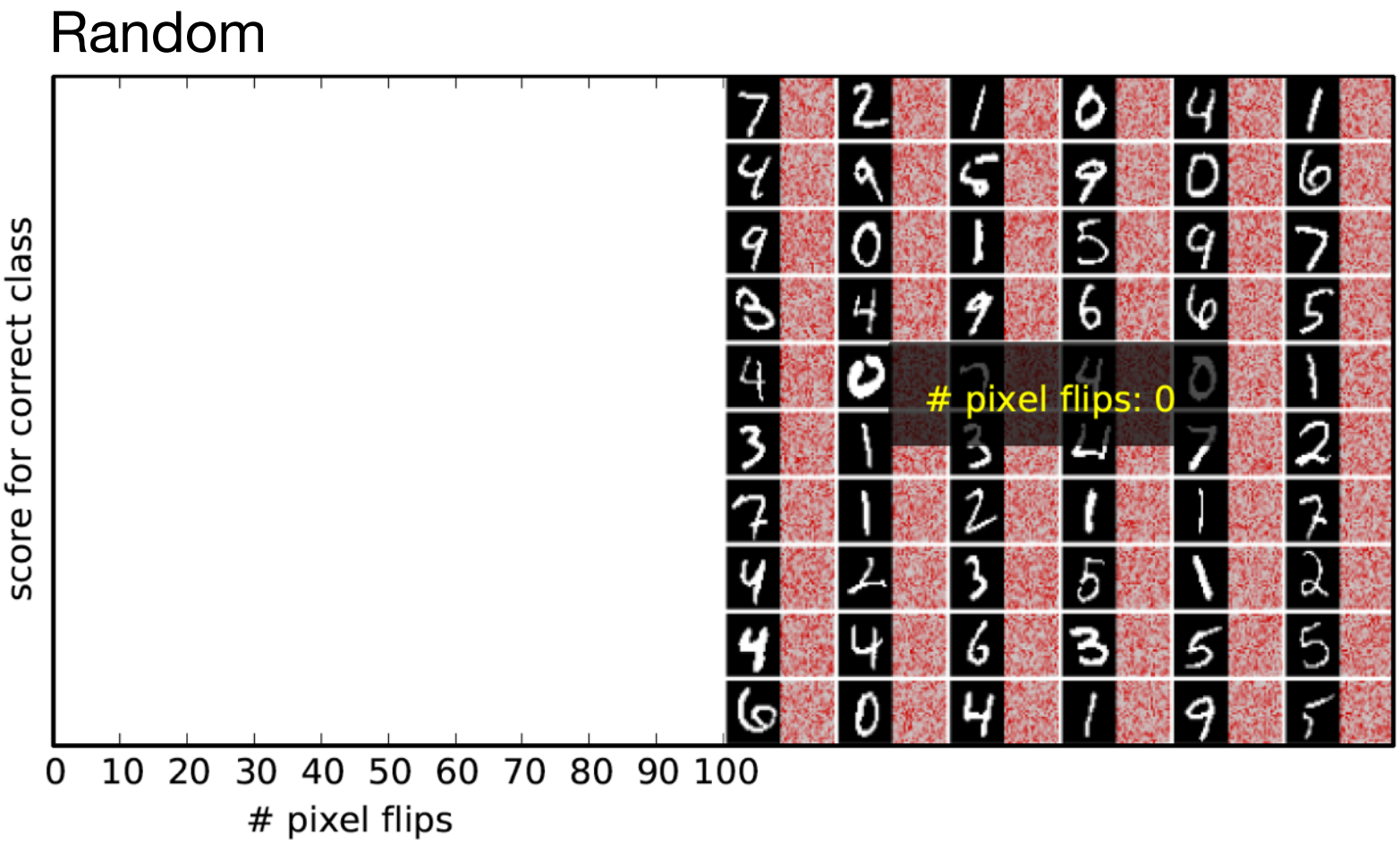
Compare Explanation Methods



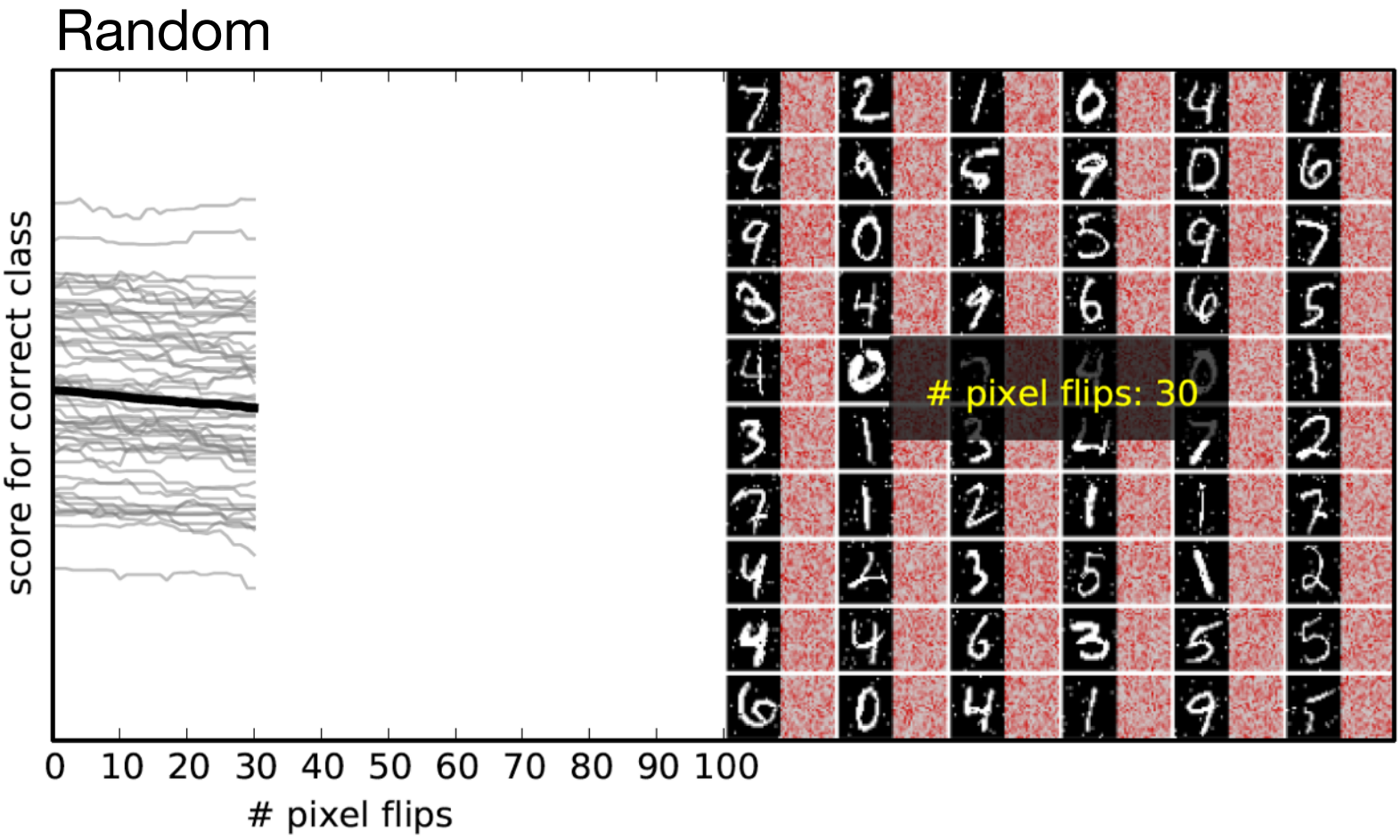
Compare Explanation Methods



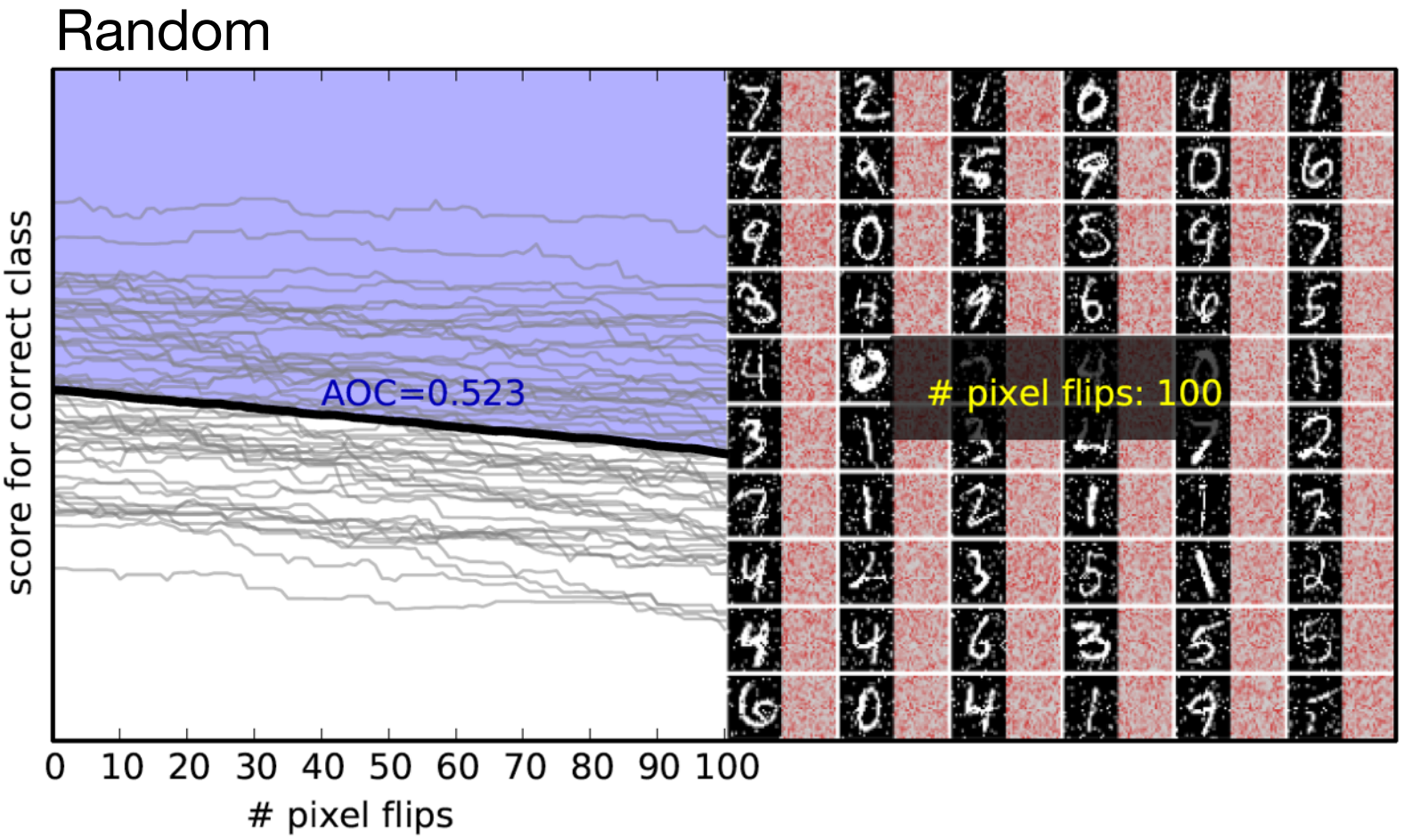
Compare Explanation Methods



Compare Explanation Methods



Compare Explanation Methods



Compare Explanation Methods

LRP:	**0.722**
Sensitivity:	0.691
Random:	0.523

LRP produces quantitatively better heatmaps than sensitivity analysis and random.

What about more complex datasets ?

SUN397

397 scene categories
(108,754 images in total)

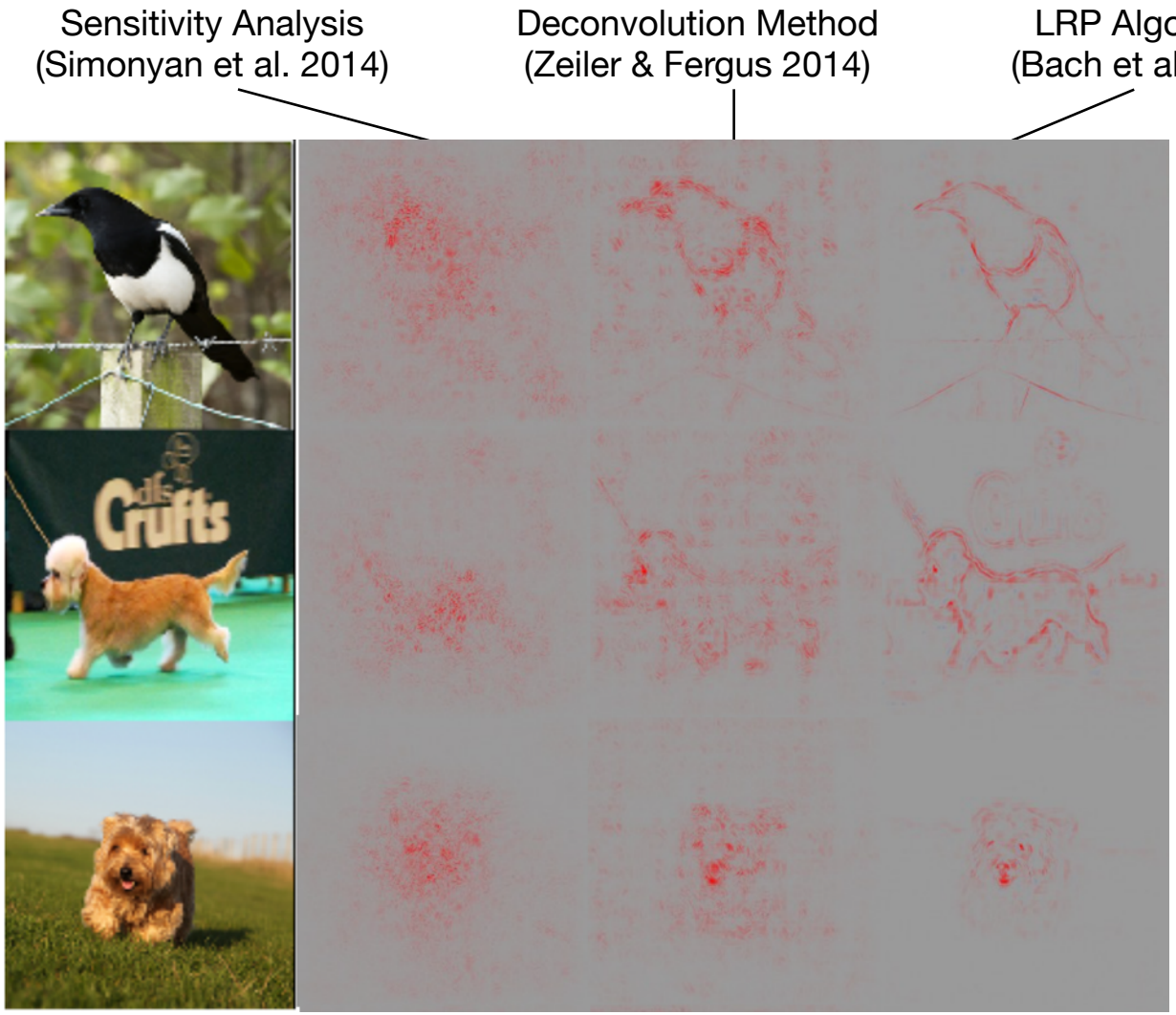
ILSVRC2012

1000 categories
(1.2 million training images)

MIT Places

205 scene categories
(2.5 millions of images)

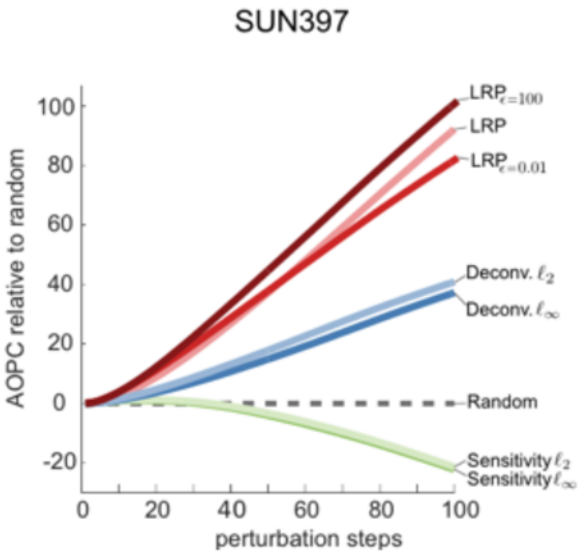
Compare Explanation Methods

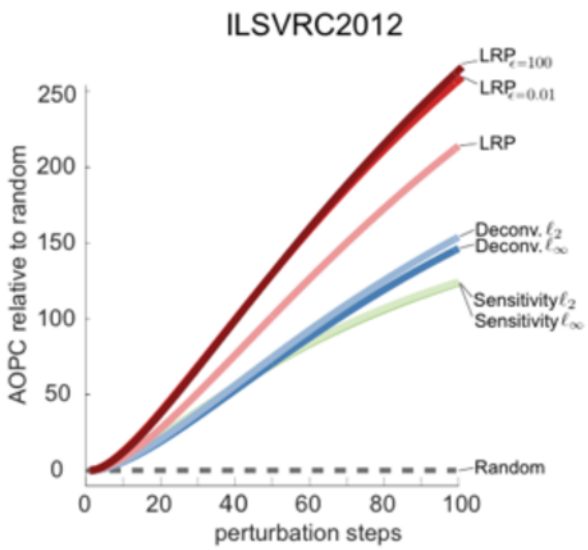


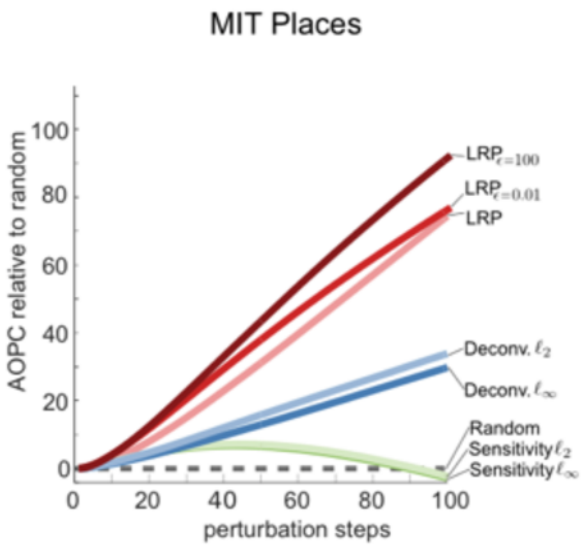
(Samek et al. 2016)

Compare Explanation Methods

Red: LRP method
Blue: Deconvolution method (Zeiler & Fergus, 2014)
Green: Sensitivity method (Simonyan et al., 2014)







LRP produces quantitatively better heatmaps.

(Samek et al. 2016)

Compare Explanation Methods

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

"Pixel flipping"
=
"Word deleting"

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

sci.med (4.1)

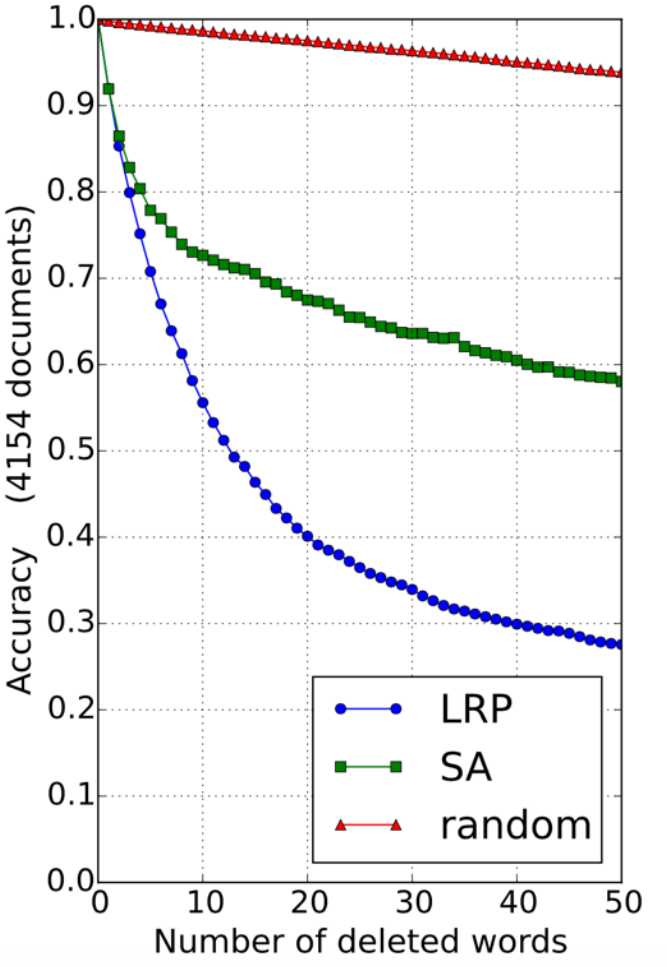
```
Yes, weightlessness does feel like falling.  It may feel strange at first,
but the body does adjust.  The feeling is not too different from that
of sky diving.

>And what is the motion sickness
>that some astronauts occasionally experience?

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

(Arras et al. 2016)

Compare Explanation Methods



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

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

LRP better identifies relevant words.

(Arras et al. 2016)

Explanations and now ?

How good is the explanation ?

- Objective measure of quality
- Compare explanation methods

What can we do with it ?

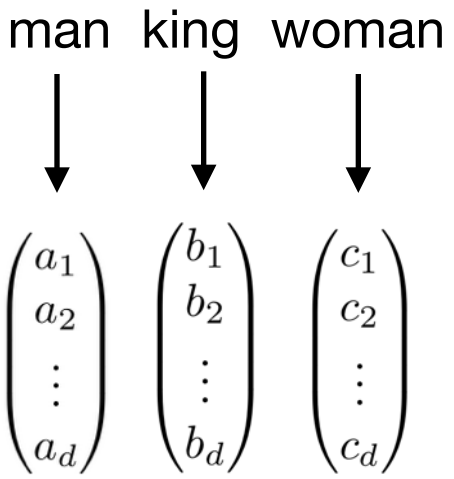
- Compare classifiers
- Detect biases and flaws
- Quantify use of context
- Novel representation
- Application in the sciences

...

Application: Compare Classifiers

20 Newsgroups data set

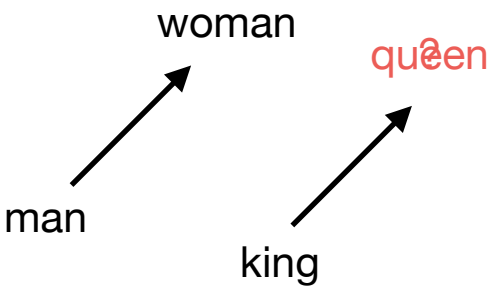
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism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



Test set performance

word2vec / CNN model: 80.19%

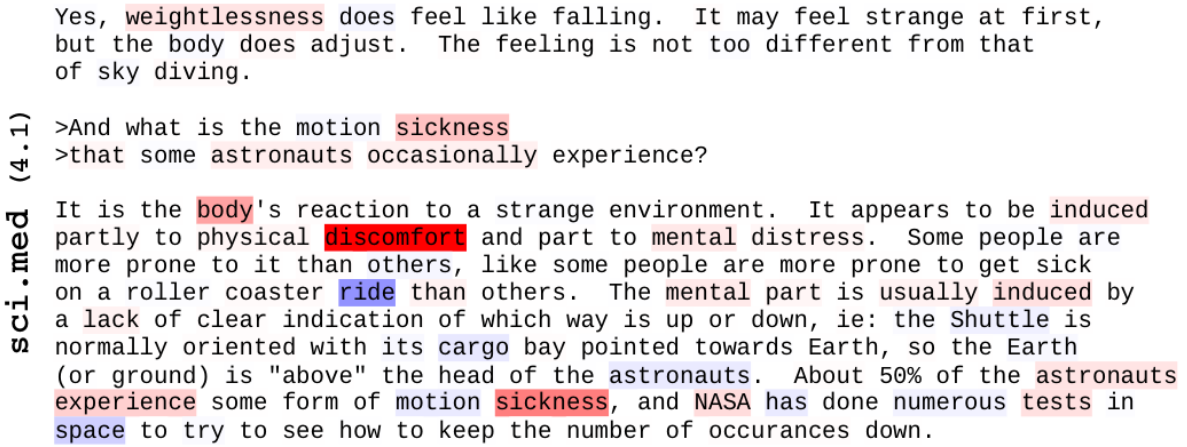
BoW/SVM model: 80.10%



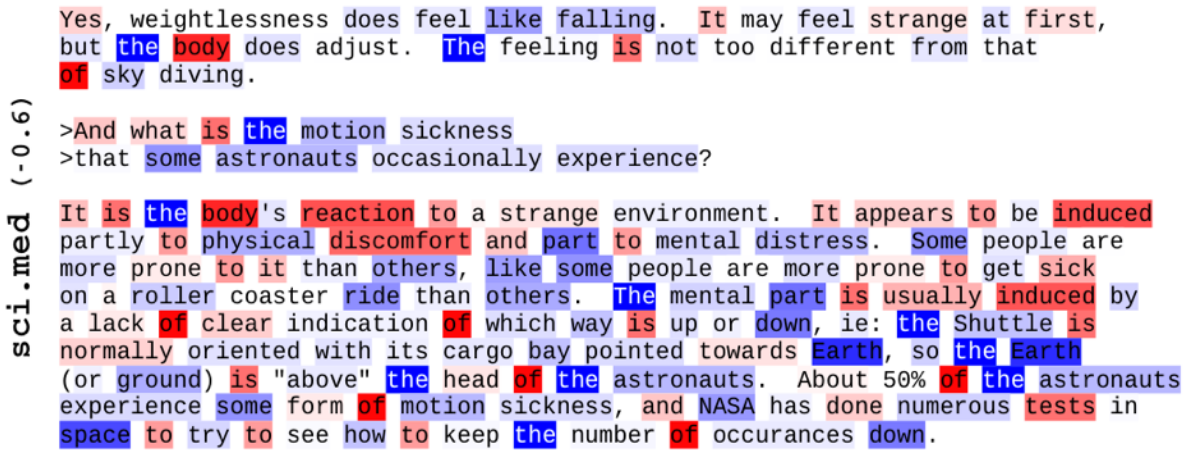
same performance —> same strategy ?

Application: Compare Classifiers

word2vec / CNN model



BoW/SVM model



(Arras et al. 2016)

Application: Compare Classifiers

word2vec / CNN model

sci.med

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

BoW/SVM model

sci.med

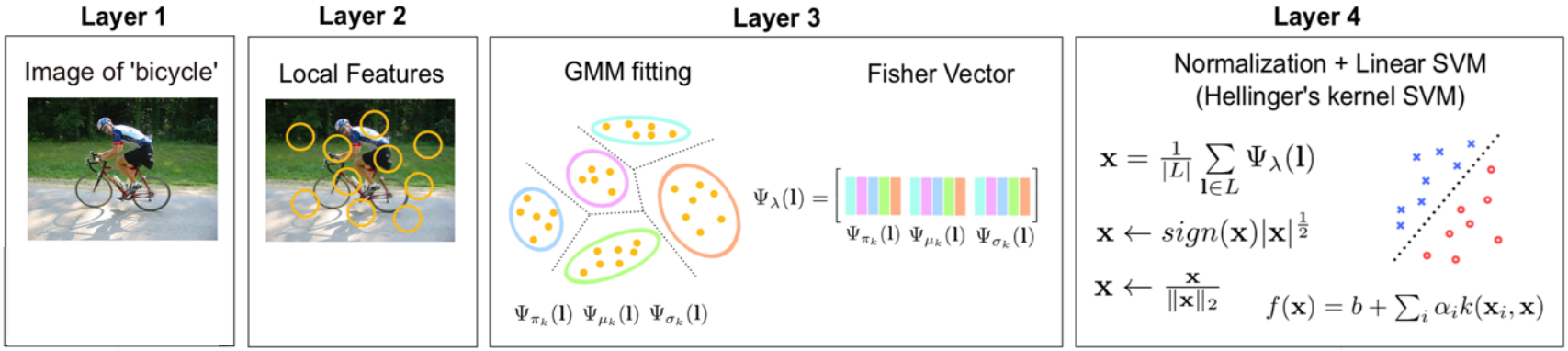
cancer (1.4), photography (1.0), doctor (1.0), msg (0.9), disease (0.9), medical (0.8), sleep (0.8), radiologist (0.7), eye (0.7), treatment (0.7), prozac (0.7), vitamin (0.7), epilepsy (0.7), health (0.6), yeast (0.6), skin (0.6), pain (0.5), liver (0.5), physician (0.5), she (0.5), needles (0.5), dn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

Words with maximum relevance

(Arras et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier

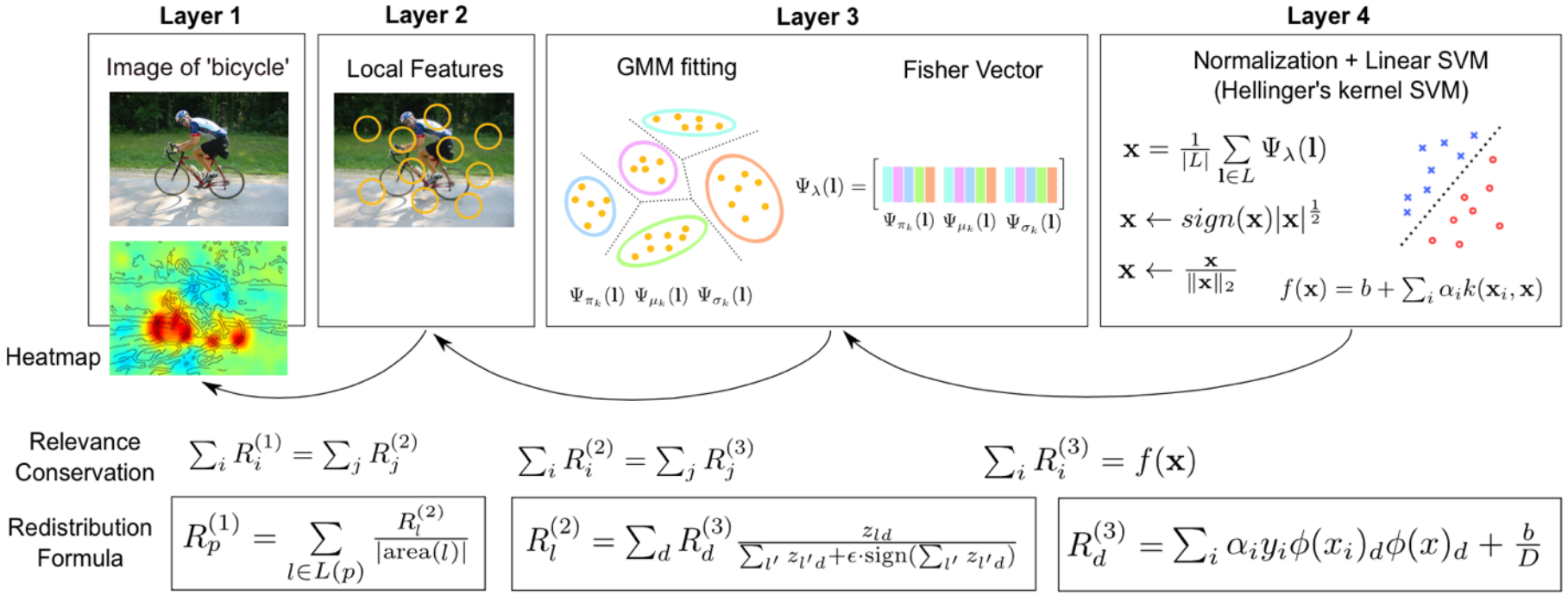


LRP general method for non-linear classifiers

(Lapuschkin et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier



Deep Neural Network

- BVLC reference model + fine tuning.

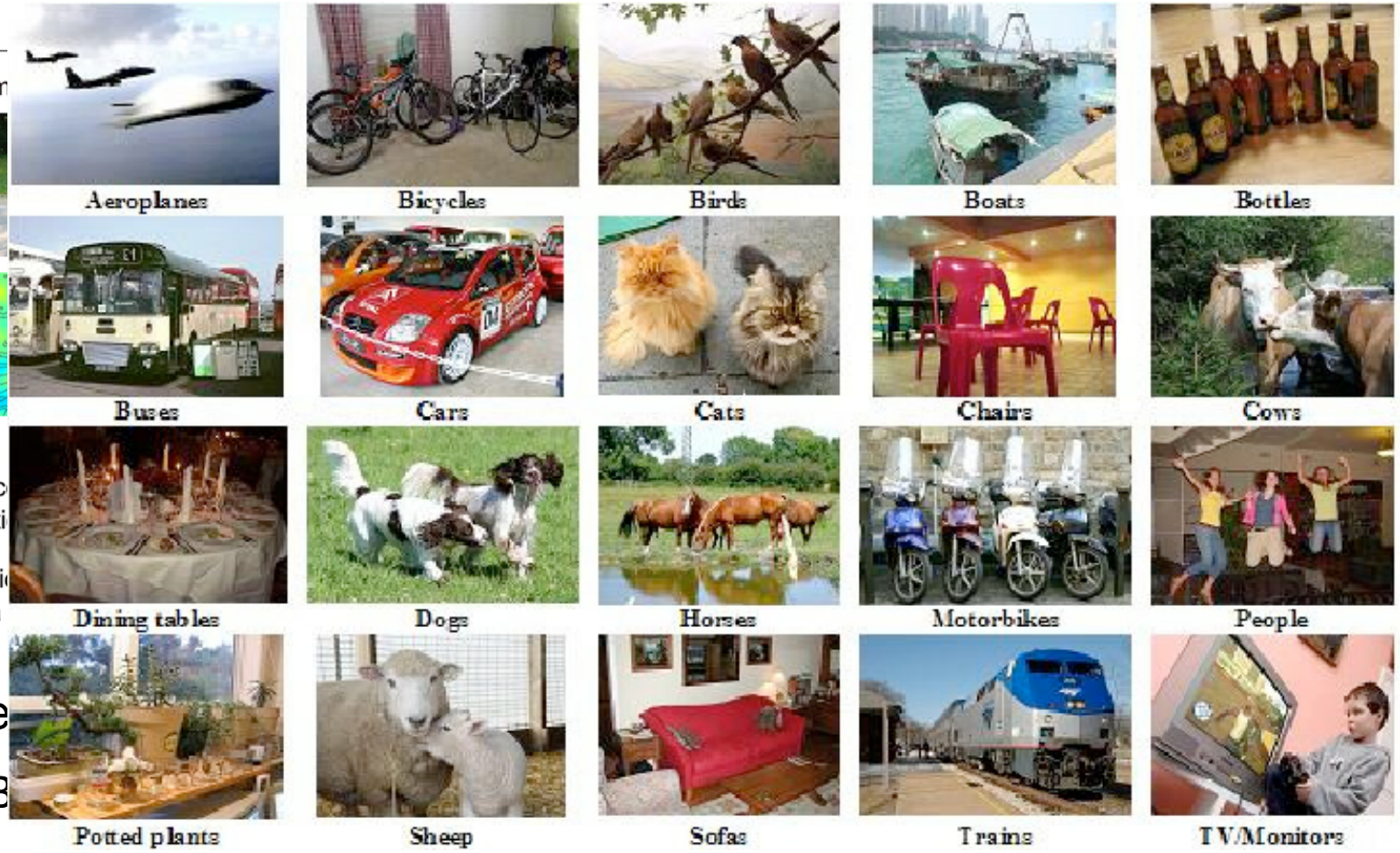
Dataset

- PASCAL VOC 2007

(Lapuschkin et al. 2016)

Application: Compare Classifiers

Fisher Vector / SVM Classifier

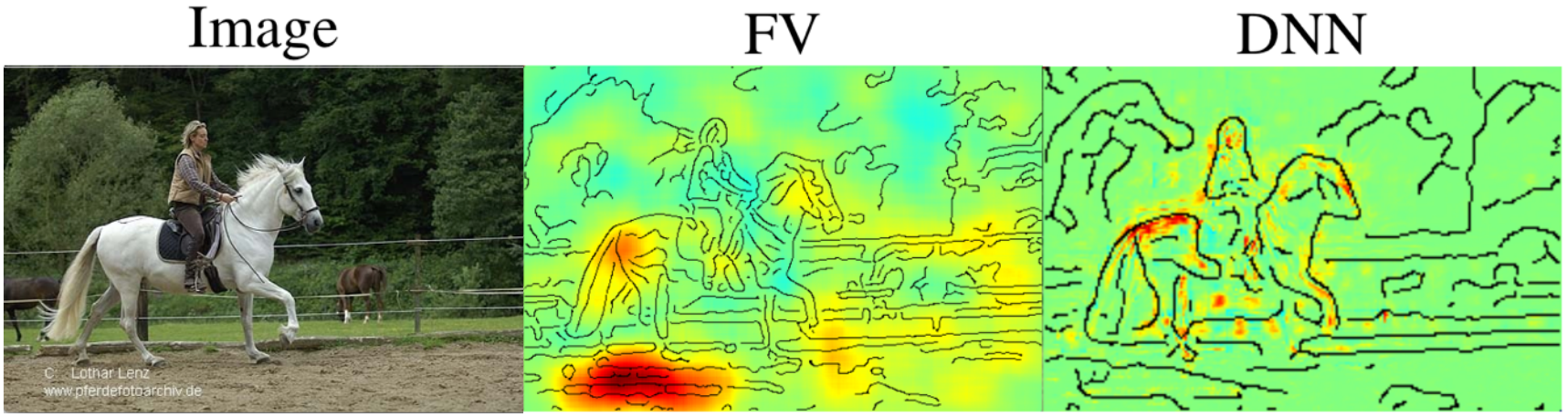


(Lapuschkin et al. 2016)

Application: Compare Classifiers

Test error for various classes:

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

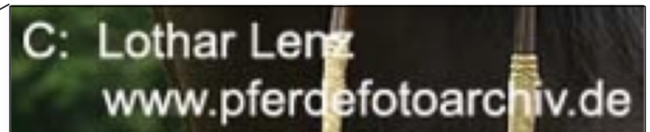


same performance —> same strategy ?

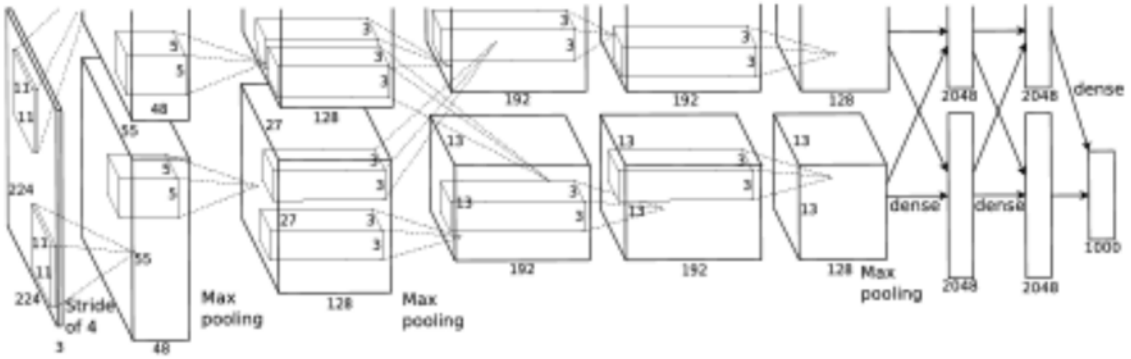
(Lapuschkin et al. 2016)

Application: Compare Classifiers

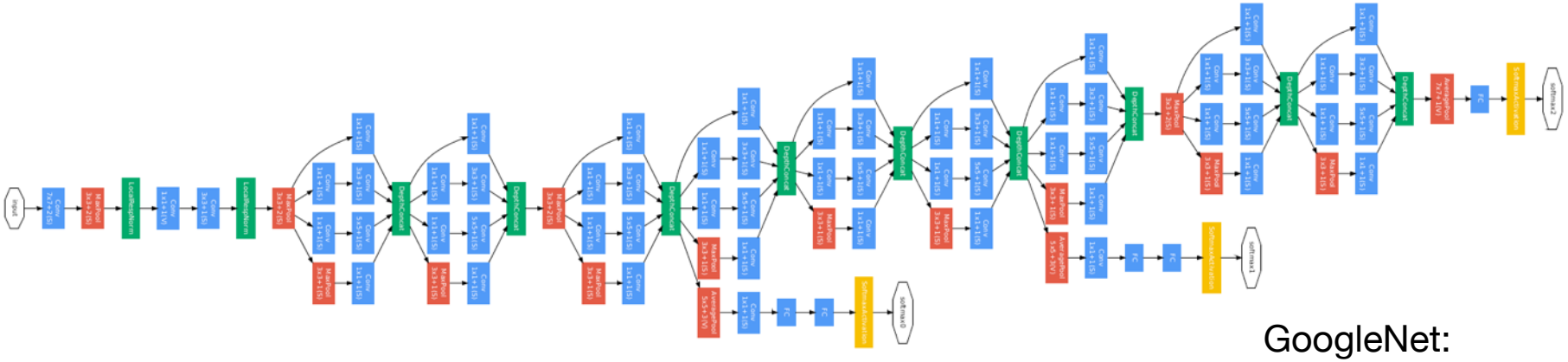
‘horse’ images in PASCAL VOC 2007



Application: Compare Classifiers



BVLC:
- 8 Layers
- ILSRCV: 16.4%



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

Application: Compare Classifiers

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

(Binder et al. 2016)

Application: Measure Context Use

how important is context ?

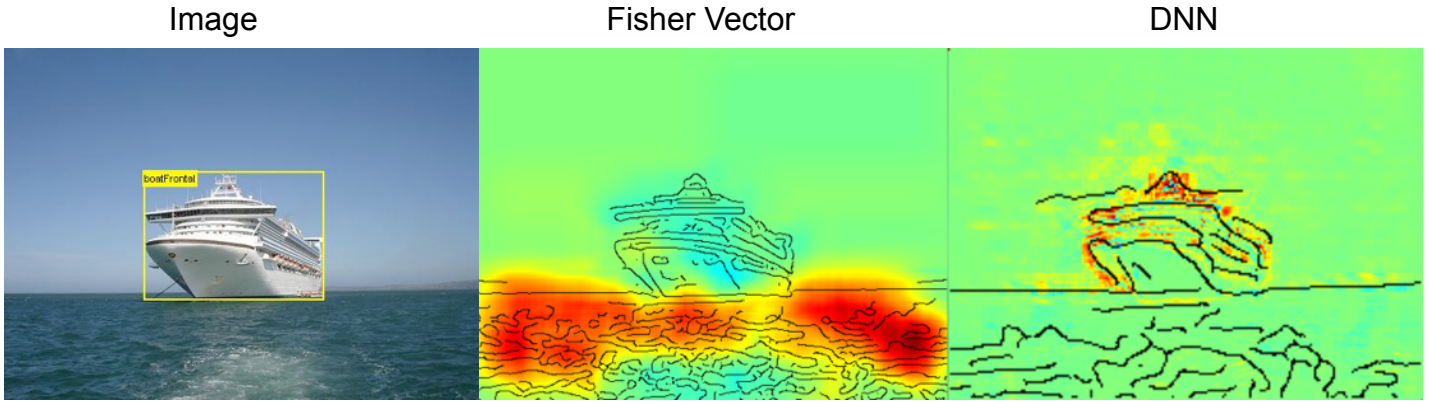
how important is context ?

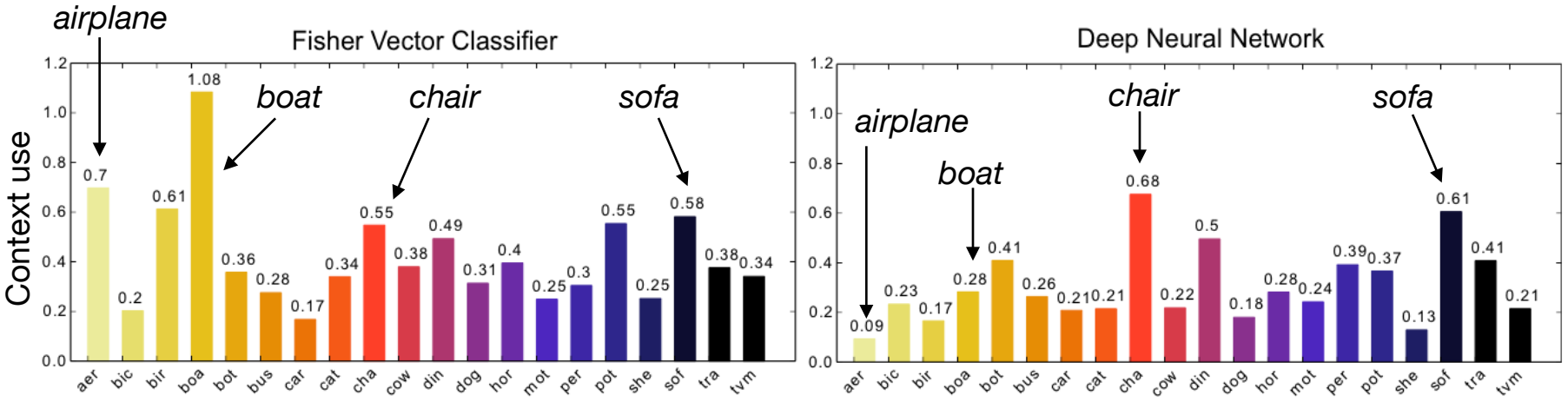
classifier

LRP decomposition allows meaningful pooling over bbox !

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

Application: Measure Context Use





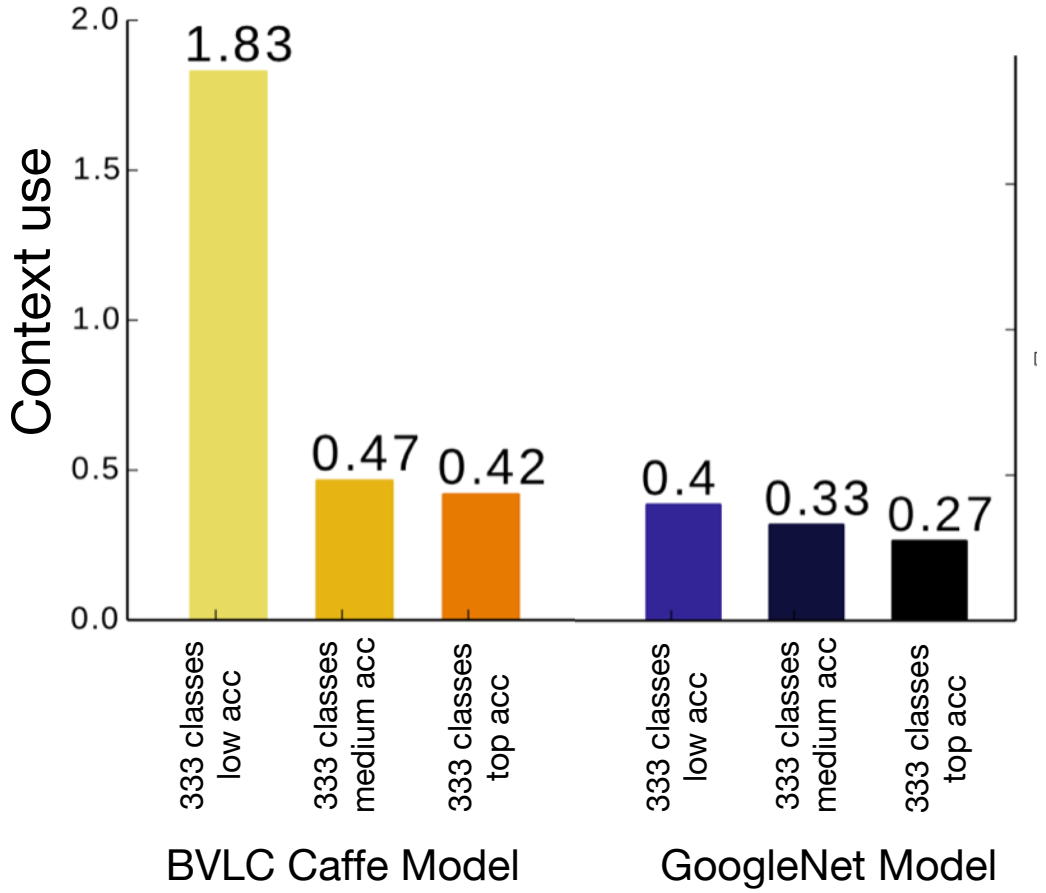
Large values indicate importance of context

(Lapuschkin et al. 2016)

Application: Measure Context Use

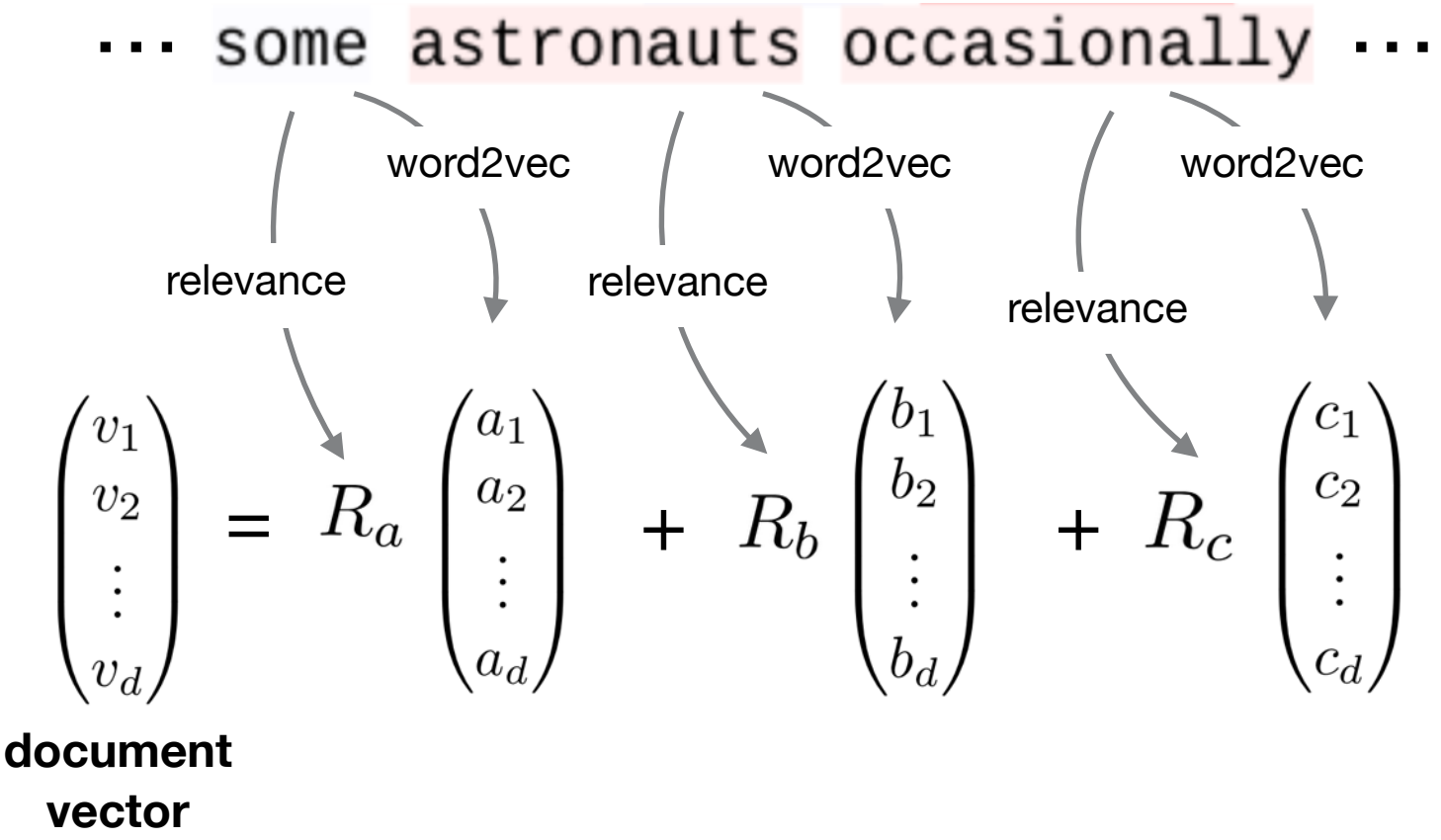
GoogleNet uses less context than BVLC model.

Context use anti-correlated with performance.



(Lapuschkin et al. 2016)

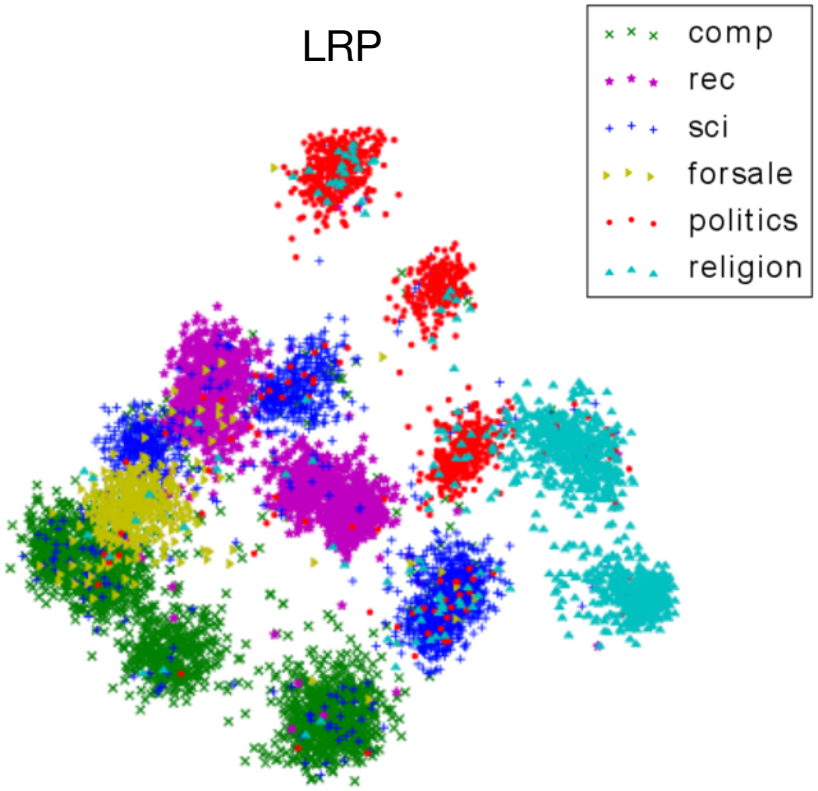
Application: Novel Representation



(Arras et al. 2016)

Application: Novel Representation

2D PCA projection of document vectors



Document vector computation is unsupervised.

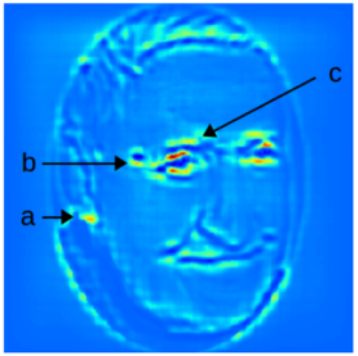
KNN-Performance on document vectors (Explanatory Power Index)

LRP: 0.8076
TFIDF: 0.6816
uniform: 0.6208

(Arras et al. 2016)

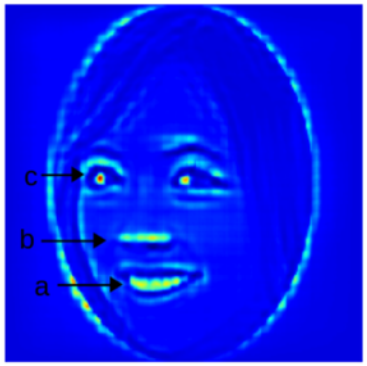
Application: Face Analysis

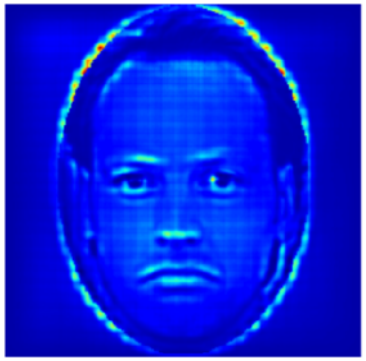
Identifying age-related features



Emotions

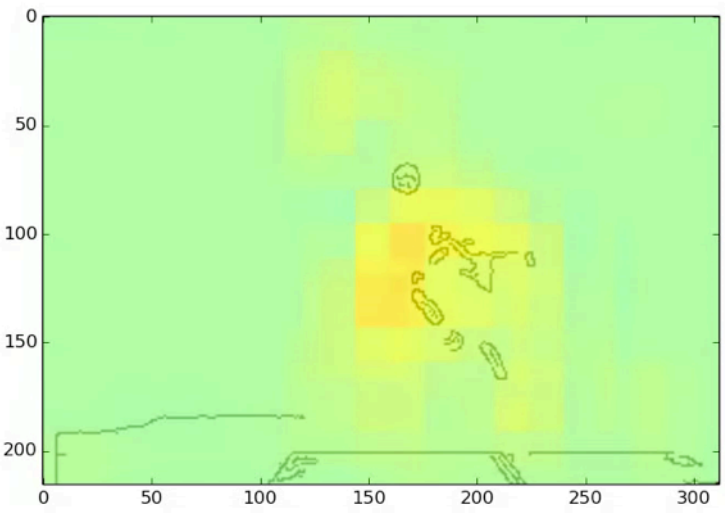
Attractivity





(Arbabzadah et al. 2016)

Application: Video Analysis



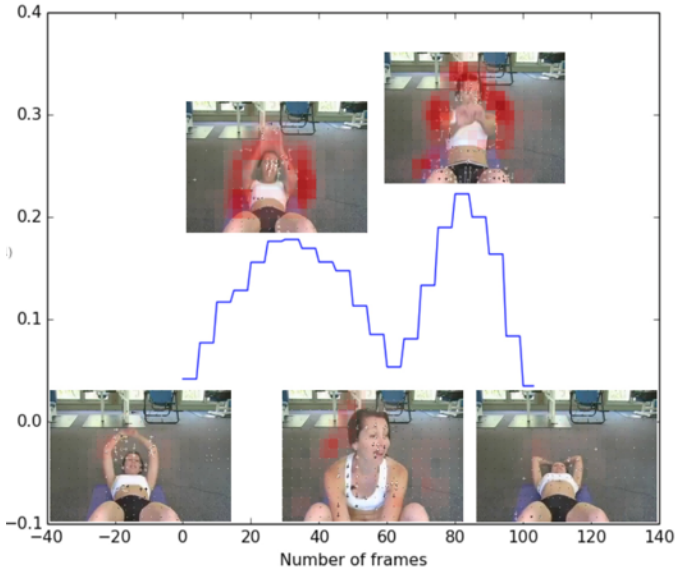
MVs assigned by encoder
from RD-perspective
—> not necessarily real motion

Session: IVMSP-P7: Image/Video Analysis & Applications
Location: Churchill: Poster Area D
Time: **Tuesday, March 7, 13:30 - 15:30**

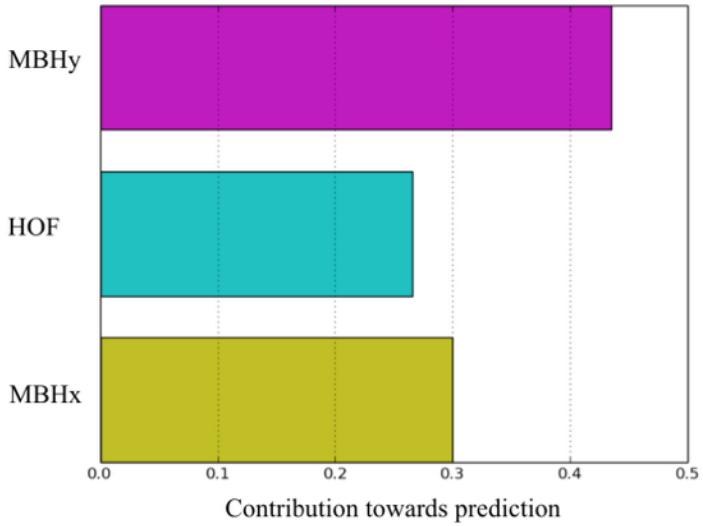
(Srinivasan et al. 2016)

Application: Video Analysis

When and where is the action ?



Which features are most relevant ?



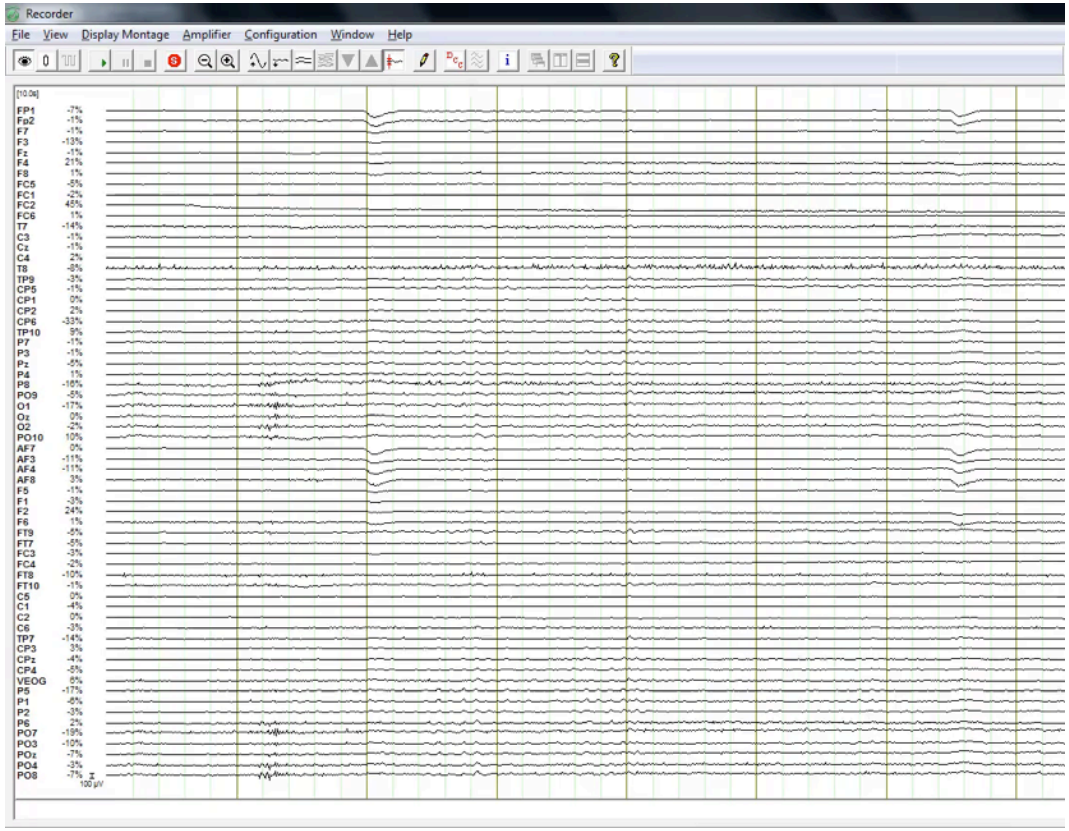
class “chew”

Session: IVMSP-P7: Image/Video Analysis & Applications
Location: Churchill: Poster Area D
Time: **Tuesday, March 7, 13:30 - 15:30**

(Srinivasan et al. 2016)

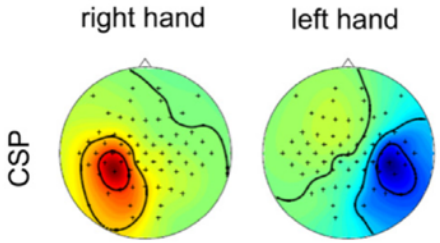
Application: Interpretability in the Sciences

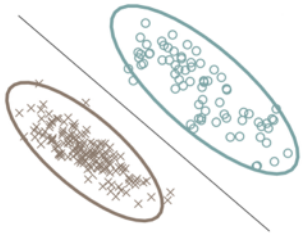
Brain-Computer Interfacing



Application: Interpretability in the Sciences

Brain-Computer Interfacing





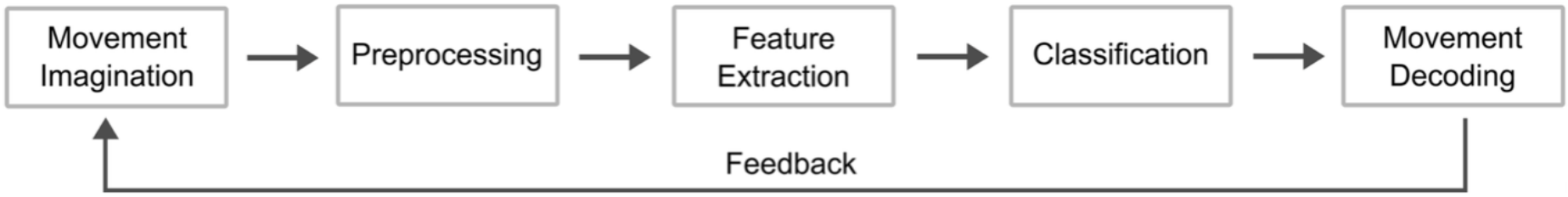
interpret

interpret

(linear) spatial filtering

variance computation

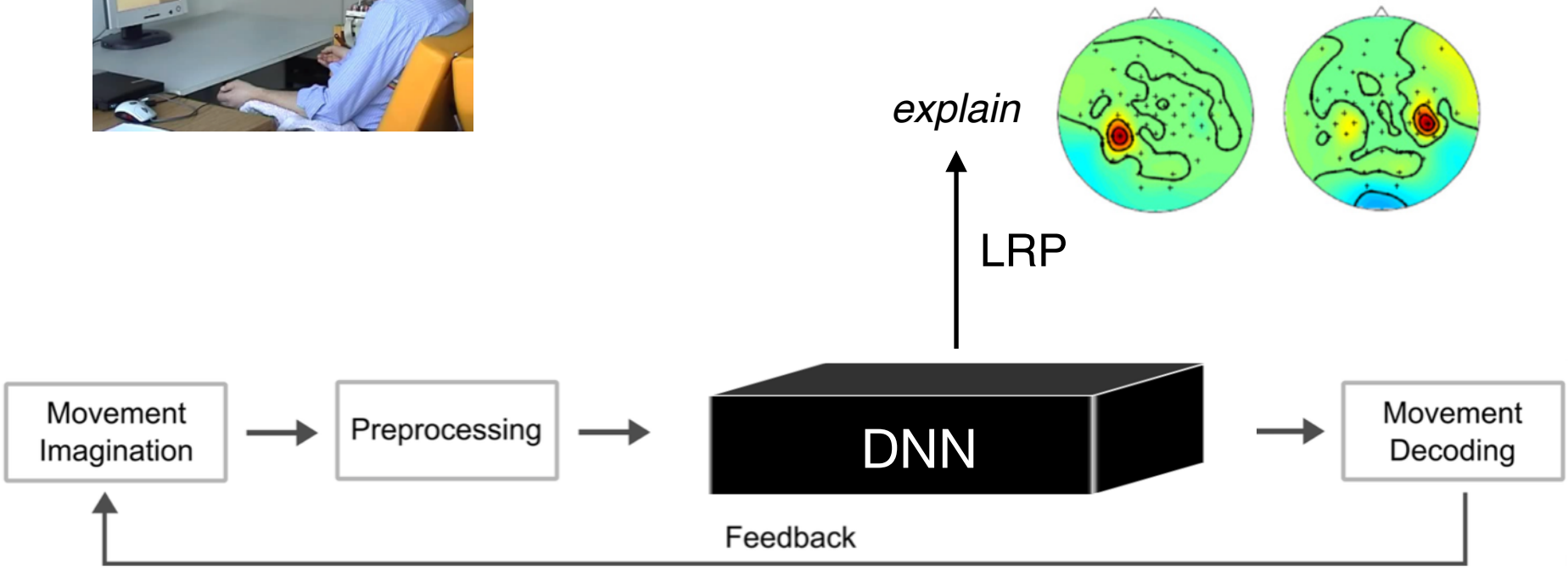
(linear) classifier (LDA)



Application: Interpretability in the Sciences

Brain-Computer Interfacing

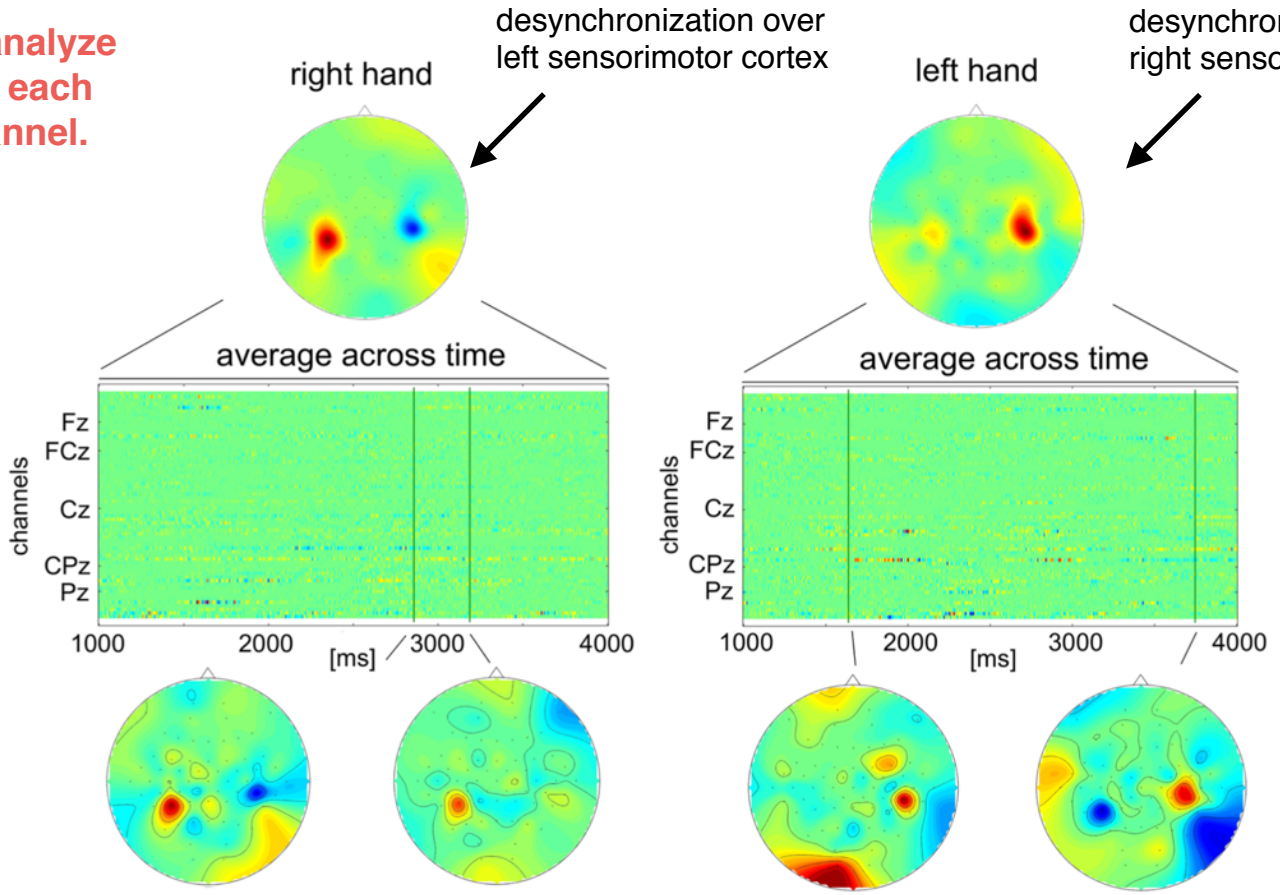
How brain works subject-dependent —> individual explanations



(Sturm et al. 2016)

Application: Interpretability in the Sciences

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

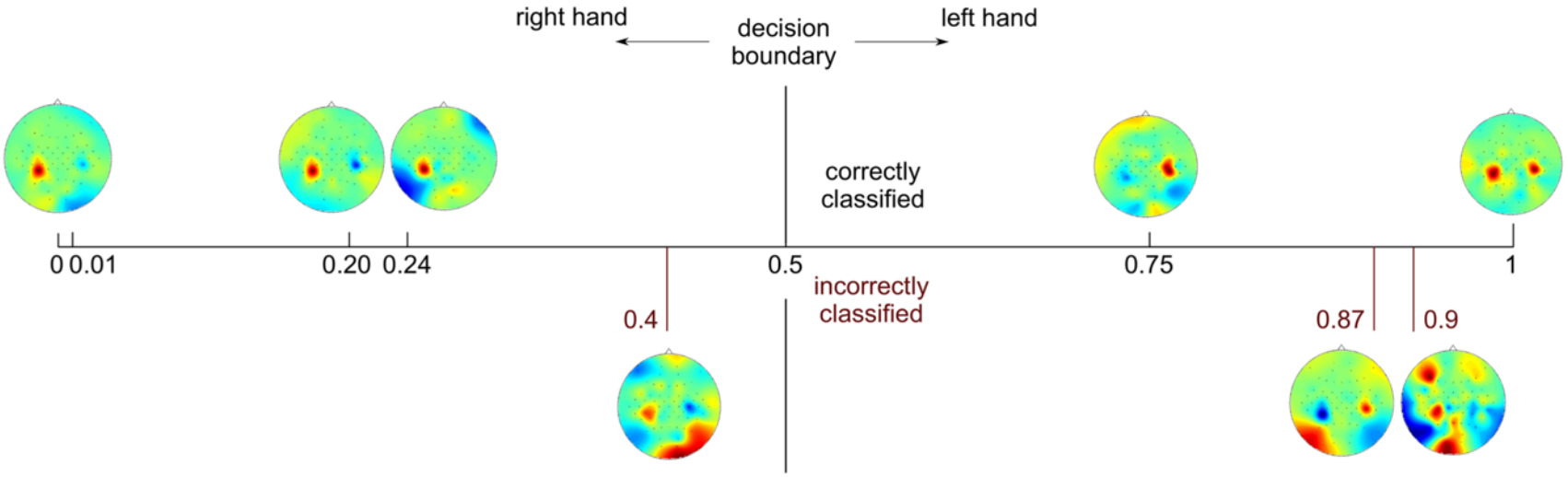


Allows to <u>spatially</u> & <u>temporally</u> identify important activity in EEG data.

(Sturm et al. 2016)

Application: Interpretability in the Sciences

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



(Sturm et al. 2016)

Application: Interpretability in the Sciences

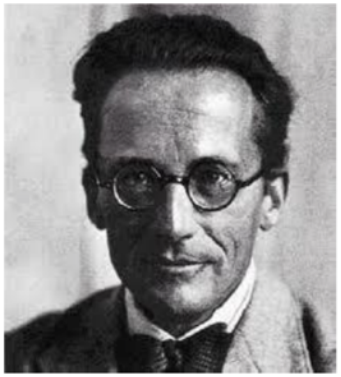
Ansatz:

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

instead of

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

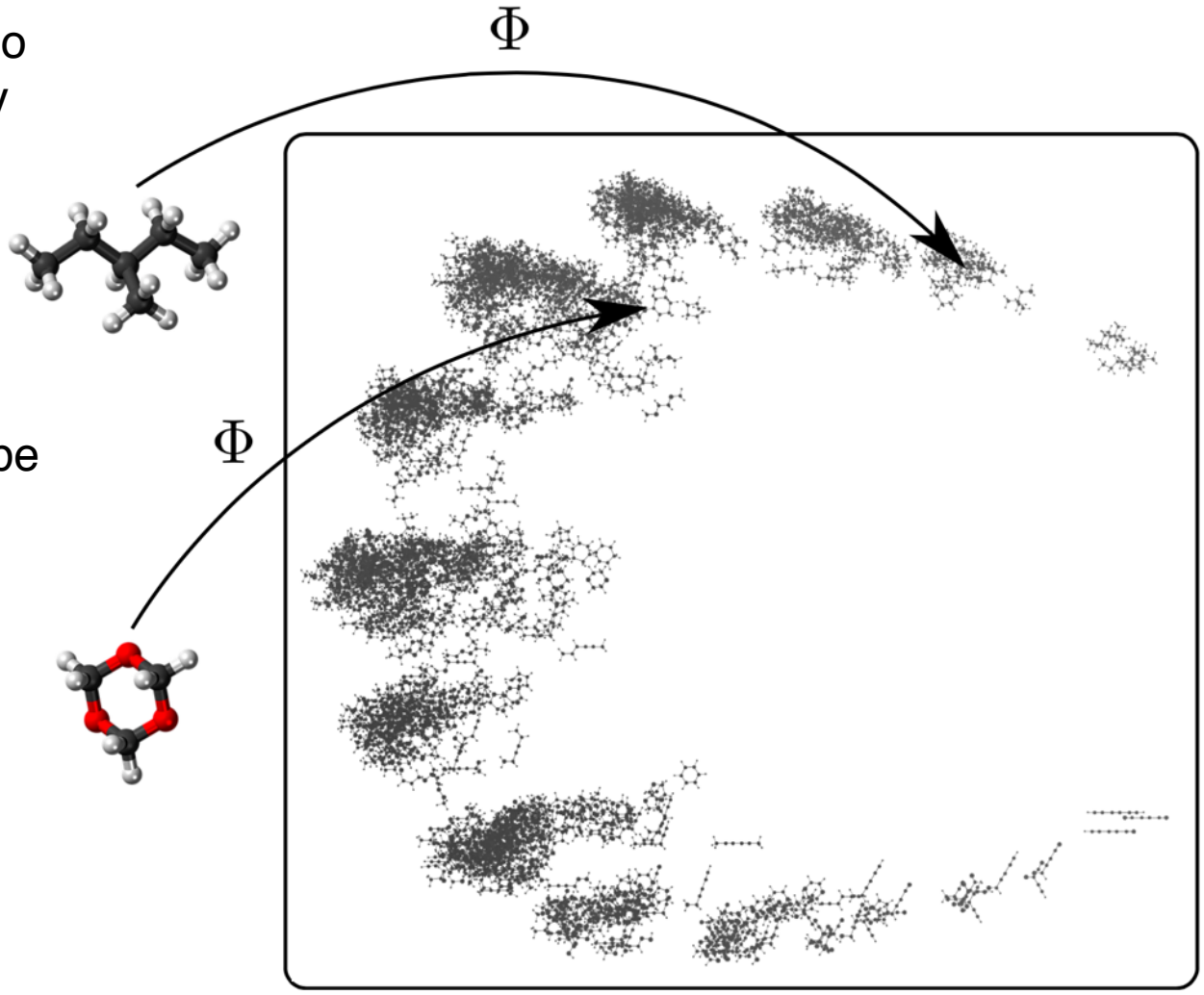
$$\hat{H}\Psi = E\Psi$$



Application: Interpretability in the Sciences

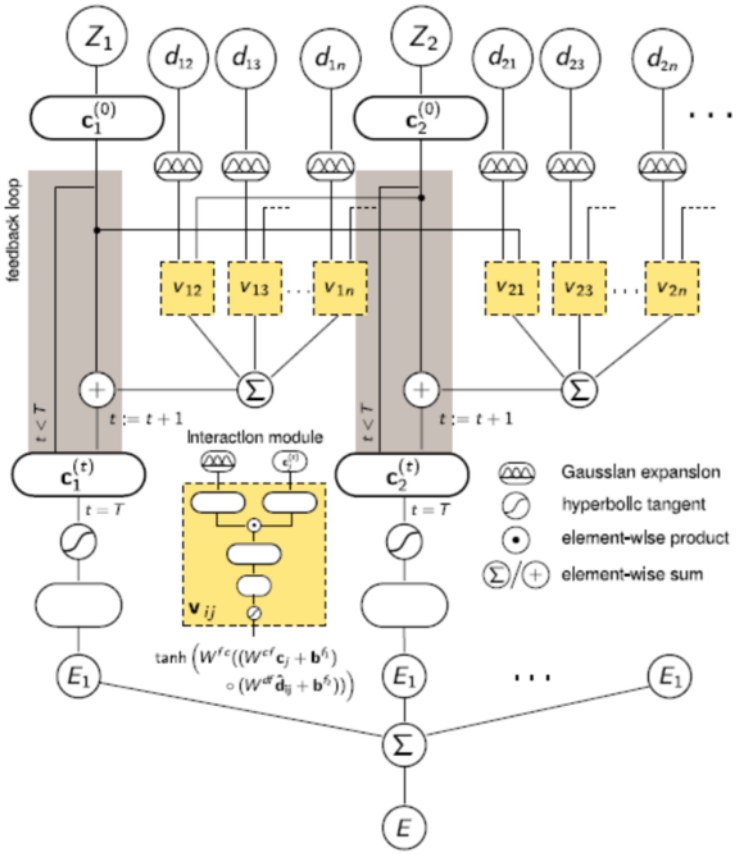
Basic simulations do not exploit similarity structure between molecules.

Our approach can be seed up analysis significantly.

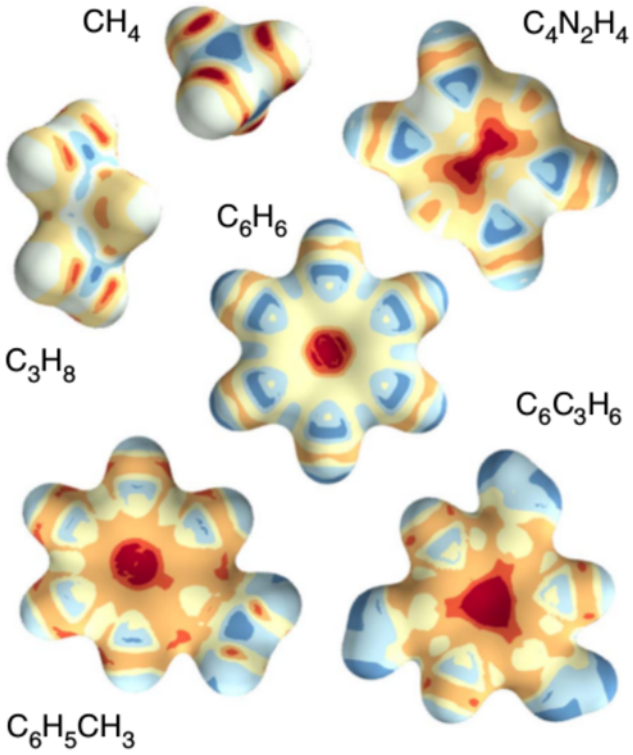


Application: Interpretability in the Sciences

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



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



(Schütt et al. 2017)

Discussion

Individual Explanations vs. Feature Selection

"explain prediction of individual data sample"

What makes these images belong to category "boat"

"average explanations"

"What are the most important features of boat images"
or
"how does a typical boat image look like"

Medical context: Patient's view vs. Population view

Discussion

Post-hoc Interpretability vs. Interpretability as Model Prior

"Train best model you can get and explain it afterwards"

Suboptimal or biased results if prior does not fit the data.

"Include interpretability aspect into the model from the beginning"

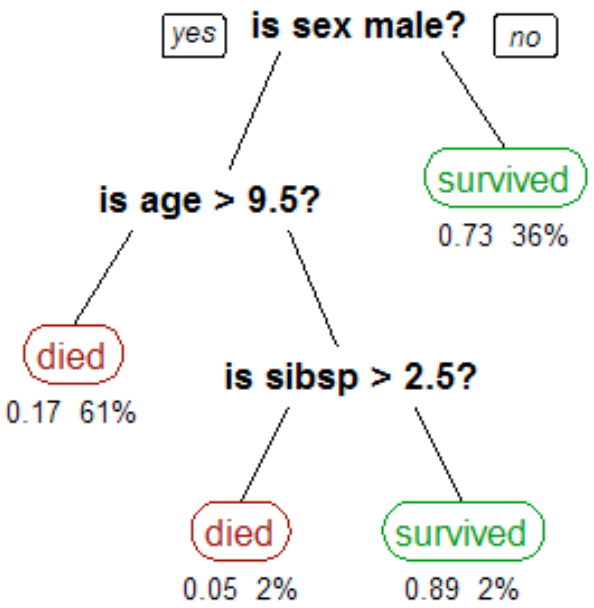
Examples: Linearity, sparsity, hierarchy, disentangled representations …

Discussion

Post-hoc Interpretability vs. Interpretability as Model Prior

Suboptimal or biased results if prior does not fit the data.

"Train best mc
and explain it



retability aspect
from the beginning"

ty, sparsity, hierarchy,
sentations …

Decision trees are highly interpretable.

Discussion

Post-hoc Interpretability vs. Interpretability as Model Prior

Suboptimal or biased results if prior does not fit the data.

"Train best model you can get

"Include interpretability aspect

(a) Rotation

(b) Width

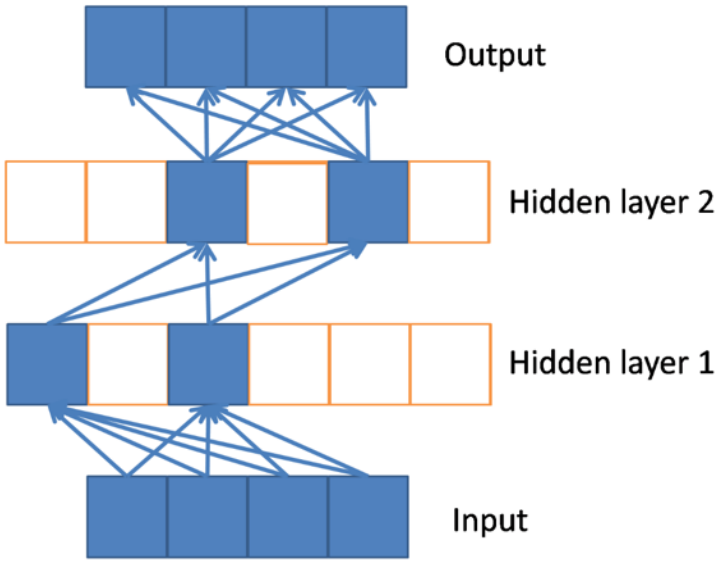
Disentangled representations are highly interpretable, but it may be hard to assign meaning to factors (often mixed).

Discussion

Post-hoc Interpretability vs. Interpretability as Model Prior

Suboptimal or biased results if prior does not fit the data.

"Train best mc … pretability aspect and explain it … from the beginning"



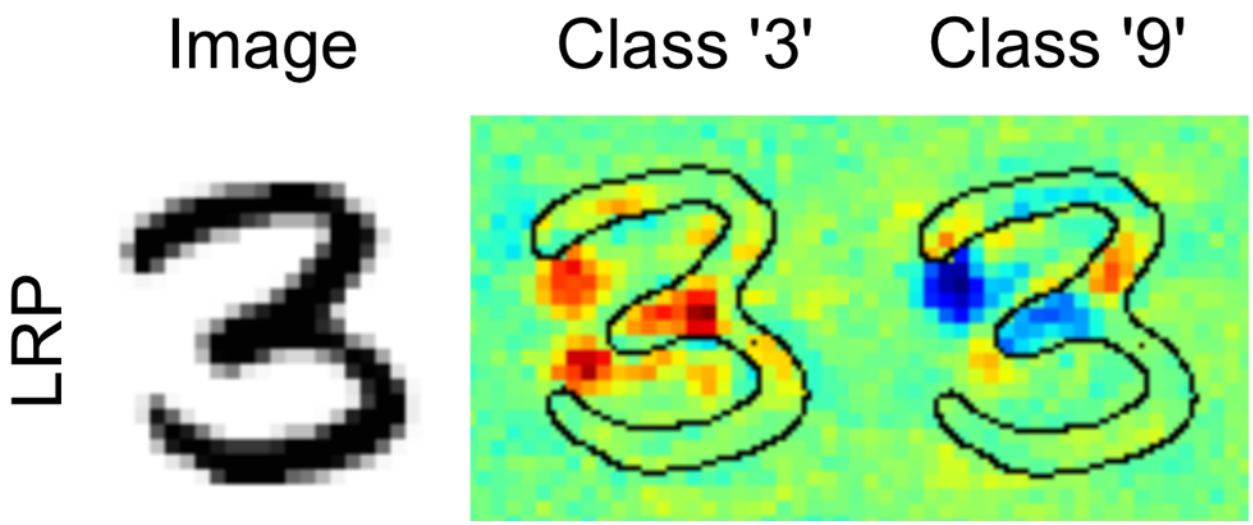
…ity, sparsity, hierarchy, …esentations …

Sparse features can lie in non-interpretable domain (e.g. hidden layer activations)

Different aspects of interpretability

What about negative evidence ?

Distinguish between positive evidence supporting a prediction (red) and negative evidence speaking against it (blue).



LRP naturally distinguishes between positive and negative evidence.
-> Sensitivity analysis and deconvolution approach do not

Different aspects of interpretability

Explanation relative to what ?

- what makes this car a car (i.e. distinguishes it from other categories)
- what distinguishes this car from another car
- what changes to the image would make it belong less / more to category ‘car’

Summary

- In many problems interpretability as important as prediction.

- Explaining individual predictions is key.

- We have powerful, mathematically well-founded methods (LRP / deep Taylor) to explain individual predictions.

- More research needed on how to compare and evaluation all the different aspects of interpretability.

- Many interesting applications with interpretable deep nets
  —> more to come soon !

Take Home Message

Don’t be afraid of complex models.

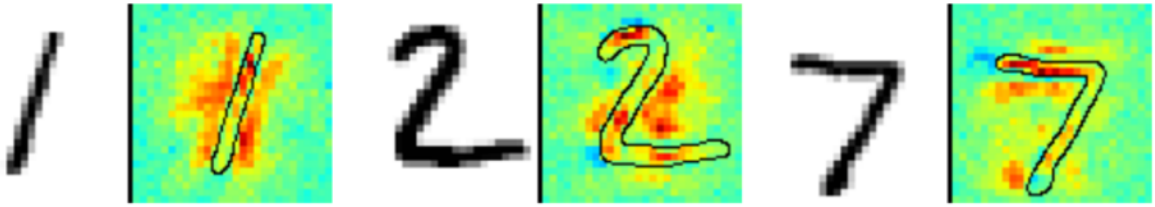
There are powerful, mathematically well-founded tools to explain *individual* predictions.

Thank you for your attention

Visit:

`http://www.heatmapping.org`

- Tutorials
- Software
- Online Demos



References

F Arbabzadeh, G Montavon, KR Müller, W Samek. Identifying Individual Facial Expressions by Deconstructing a Neural Network. *Pattern Recognition - 38th German Conference, GCPR 2016*, Lecture Notes in Computer Science, 9796:344-54, Springer International Publishing, 2016.

L Arras, F Horn, G Montavon, KR Müller, W Samek. Explaining Predictions of Non-Linear Classifiers in NLP. *Workshop on Representation Learning for NLP,* Association for Computational Linguistics, 1-7, 2016.

L Arras, F Horn, G Montavon, KR Müller, W Samek. "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. *arXiv:1612.07843*, 2016.

S Bach, A Binder, G Montavon, F Klauschen, KR Müller, W Samek. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015.

A Binder, G Montavon, S Lapuschkin, KR Müller, W Samek. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *Artificial Neural Networks and Machine Learning – ICANN 2016*, Part II, Lecture Notes in Computer Science, Springer-Verlag, 9887:63-71, 2016.

A Binder, W Samek, G Montavon, S Bach, KR Müller. Analyzing and Validating Neural Networks Predictions. *ICML Workshop on Visualization for Deep Learning*, 2016

S Lapuschkin, A Binder, G Montavon, KR Müller, Wojciech Samek. Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912-20, 2016

References

G Montavon, S Bach, A Binder, W Samek, KR Müller. DeepTaylor Decomposition of Neural Networks. *ICML Workshop on Visualization for Deep Learning*, 2016.

G Montavon, S Bach, A Binder, W Samek, KR Müller. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222, 2017.

A Nguyen, A Dosovitskiy, J Yosinski, T Brox, J Clune. Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks. *Conference on Neural Information Processing Systems (NIPS)*, 3387–3395, 2016.

W Samek, A Binder, G Montavon, S Lapuschkin, KR Müller. Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

K Schütt, F Arbabzadah, S Chmiela, KR Müller, A Tkatchenko. Quantum-chemical insights from deep tensor neural networks. Nature Communications, 13890, 2017

K Simonyan, A Vedaldi, A Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034*, 2013.

V Srinivasan, S Lapuschkin, C Hellge, KR Müller, W Samek. Interpretable human action recognition in compressed domain. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

I Sturm, S Lapuschkin, W Samek, KR Müller. Interpretable Deep Neural Networks for Single-Trial EEG Classification. *Journal of Neuroscience Methods*, 274:141–145, 2016.