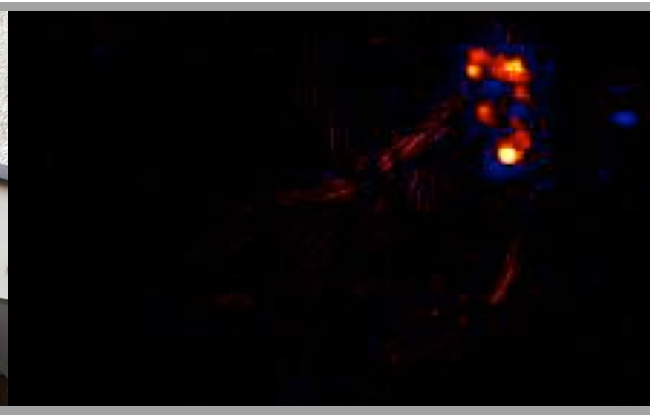
Fraunhofer
Heinrich Hertz Institute

Opening the Black Box: Making Deep Learning Interpretable & Transparent

Dr. Wojciech Samek, ML Group
Fraunhofer Heinrich Hertz Institute



Applied Machine Learning Days

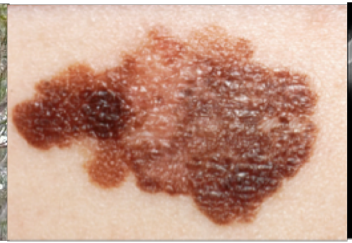
29th January, Lausanne, Switzerland

Black Box AI with "Superhuman" Performance

Game GO

Traffic Sign Recognition

Skin cancer detection



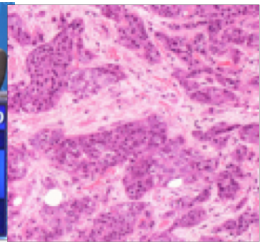
Lung cancer detection

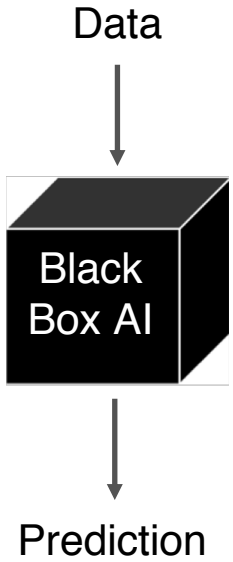
Poker

Computer games

Jeopardy

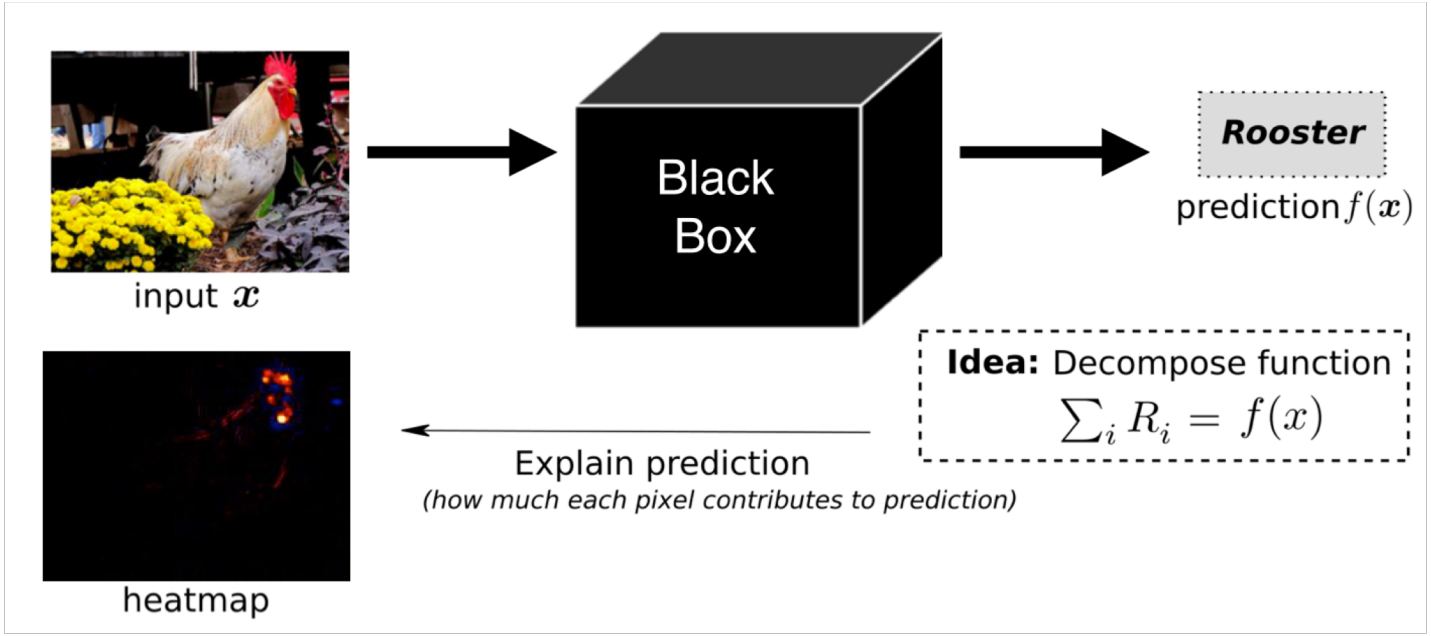
Histopathology





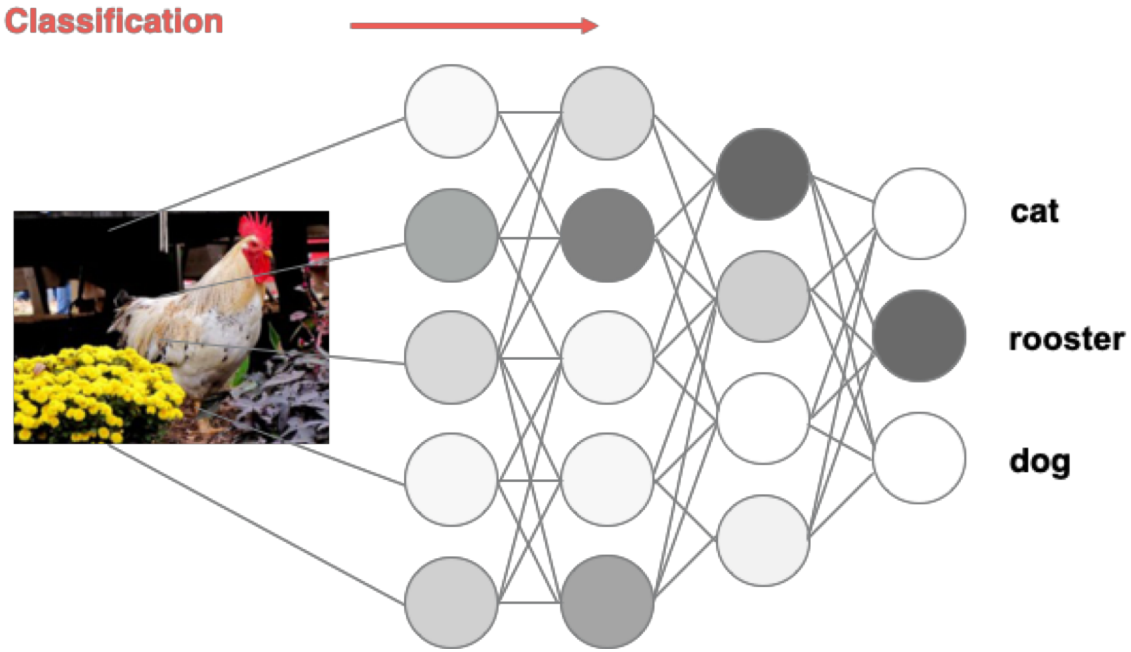
Black Box AI with "Superhuman" Performance

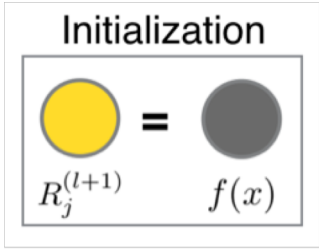
Opening the black box



Layer-wise Relevance Propagation (Bach et al. 2015)
is a general approach to explain predictions of AI.

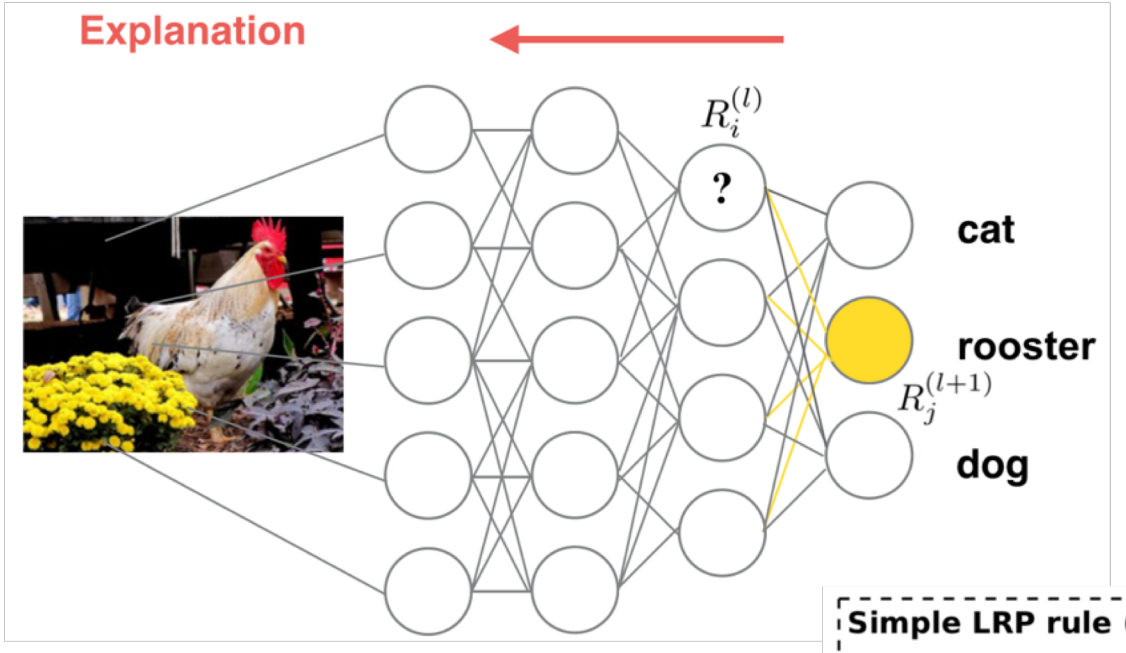
Opening the black box





Idea: Redistribute the evidence for class rooster back to image space.

Opening the black box



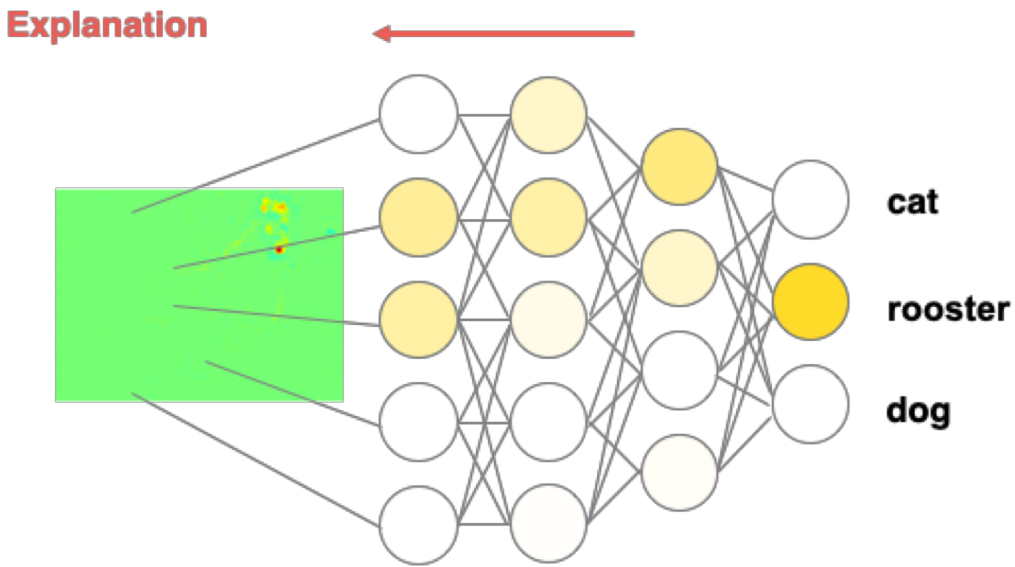
Simple LRP rule (Bach et al. 2015)

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

Every neuron gets its "share" of the redistributed relevance

Mathematical Interpretation: Deep Taylor Decomposition

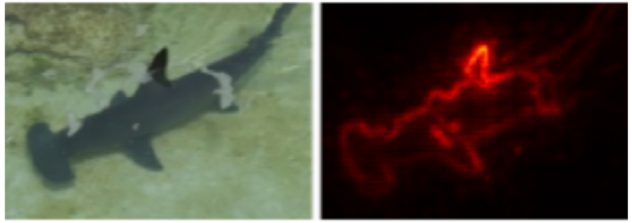
Opening the black box



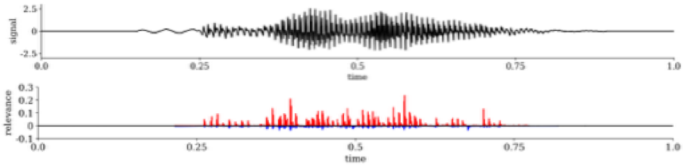
Layer-wise relevance conservation

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

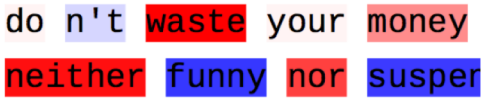
General Images (Bach' 15, Lapuschkin'16)



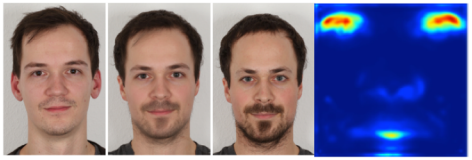
Speech (Becker'18)



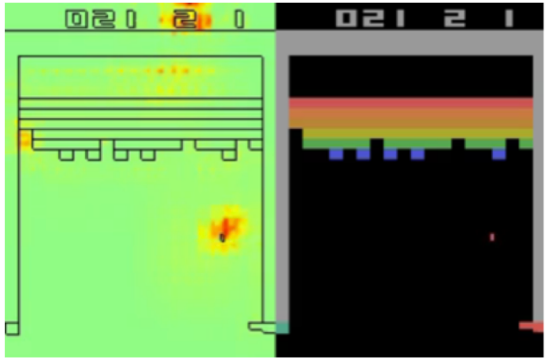
Text Analysis (Arras'16 &17)



Morphing (Seibold'18)

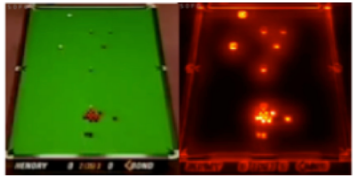


Games (Lapuschkin'19)

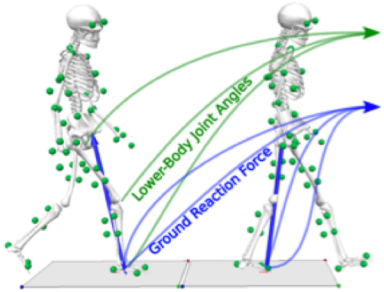


VQA (Arras'18)

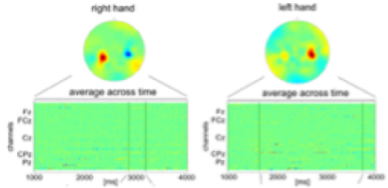
Video (Anders'18)



Gait Patterns (Horst'19)

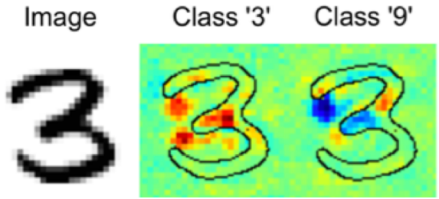


EEG (Sturm'16)

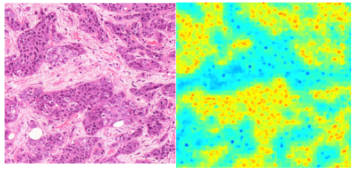


Faces (Lapuschkin'17)

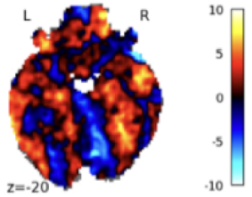
Digits (Bach' 15)



Histopathology (Binder'18)



fMRI (Thomas'18)



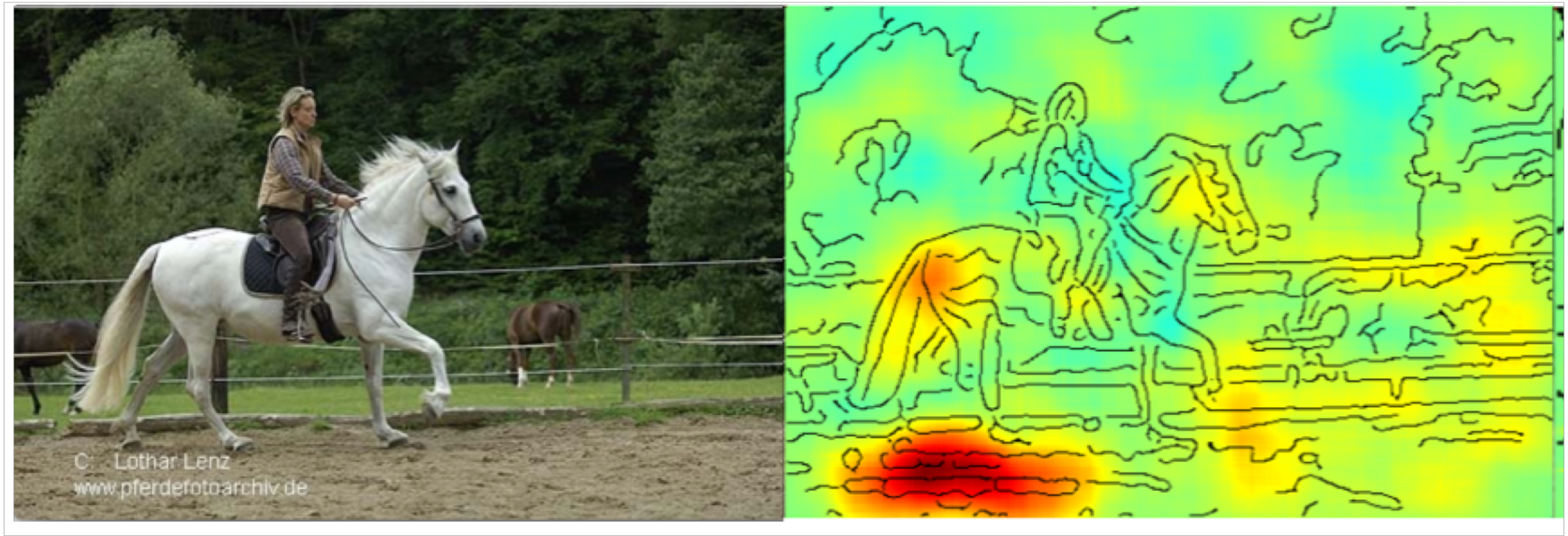
Examples from Our Research

Pascal VOC Challenges 2005-2012

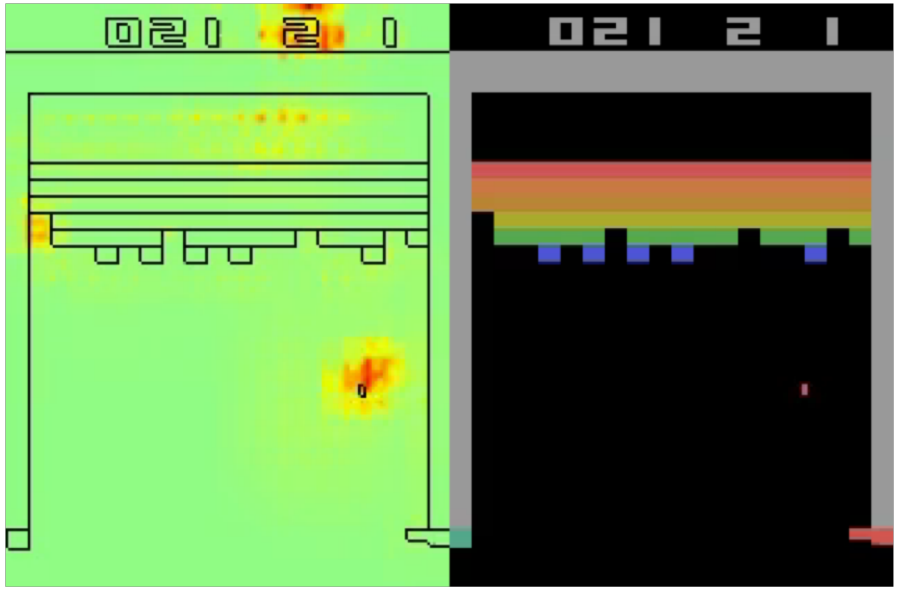
Best results for classes:

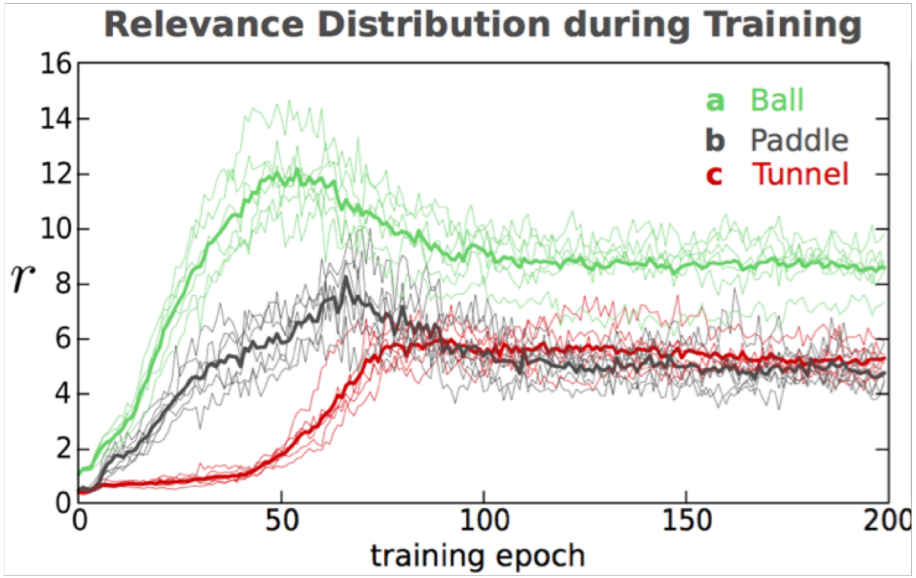
- Person
- Train
- Horse

Pascal VOC Challenges 2005-2012

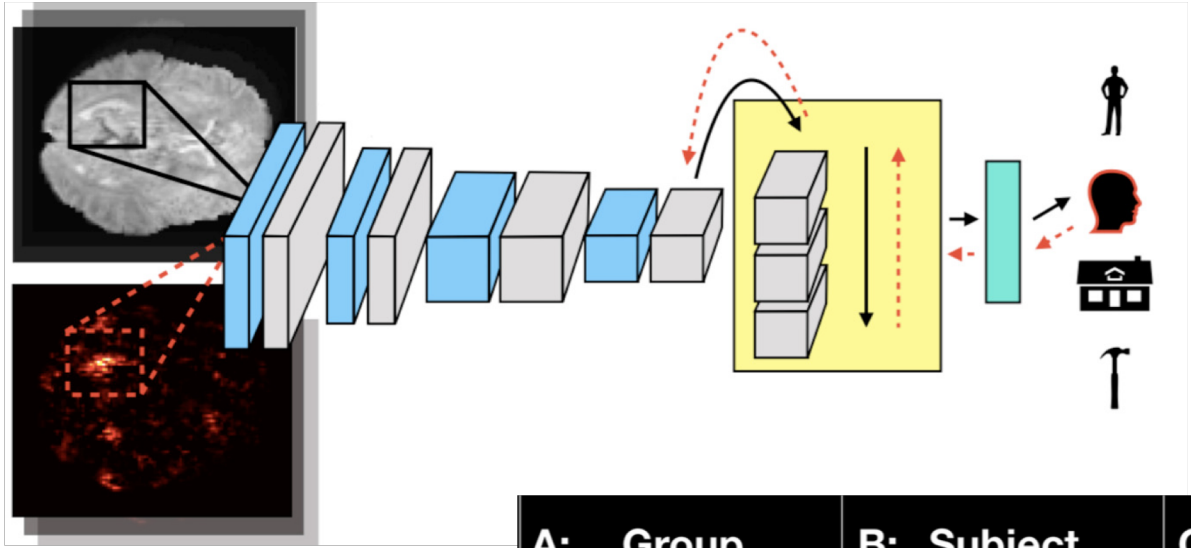


Understanding Machines Playing Games

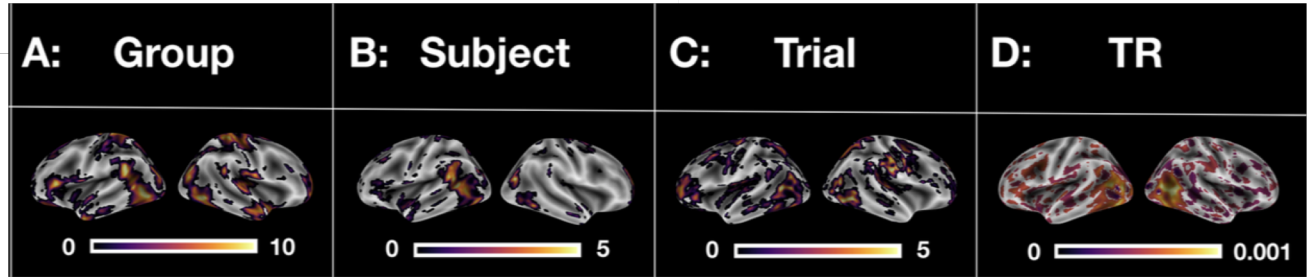




Application to Health: fMRI Decoding



Our approach:
- Recurrent neural networks (CNN + LSTM) for whole-brain analysis
- LRP allows to interpret the results

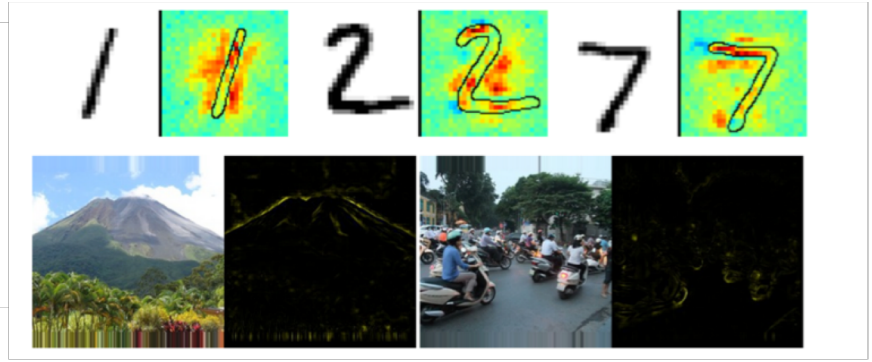


Thank you for your attention

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`