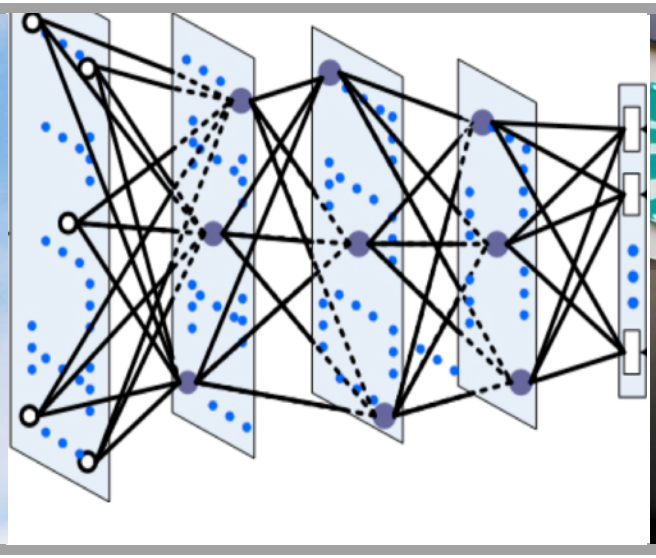
Fraunhofer
Heinrich Hertz Institute

Neuronale Netzwerke beim Denken beobachten

Fraunhofer HHI, Machine Learning Group

Dr. Wojciech Samek



From Lab to Real Applications

Impressive performance

Deep Net outperforms humans in image classification

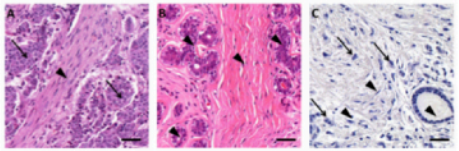
IMAGENET

AlphaGo beats Go human champ

Dermatologist-level classification of skin cancer

Real applications

Medical Diagnosis



Autonomous Driving

Smart devices, 5G, IoT etc.

From Lab to Real Applications

Impressive performance

Deep Net outperforms humans in image classification

IMAGE

AlphaGo beats human champ

Dermatologist-level classification of skin cancer

Explainability

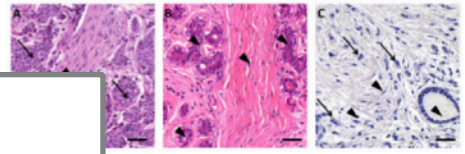
Reliability

Energy Efficiency

Privacy Preservation

Real applications

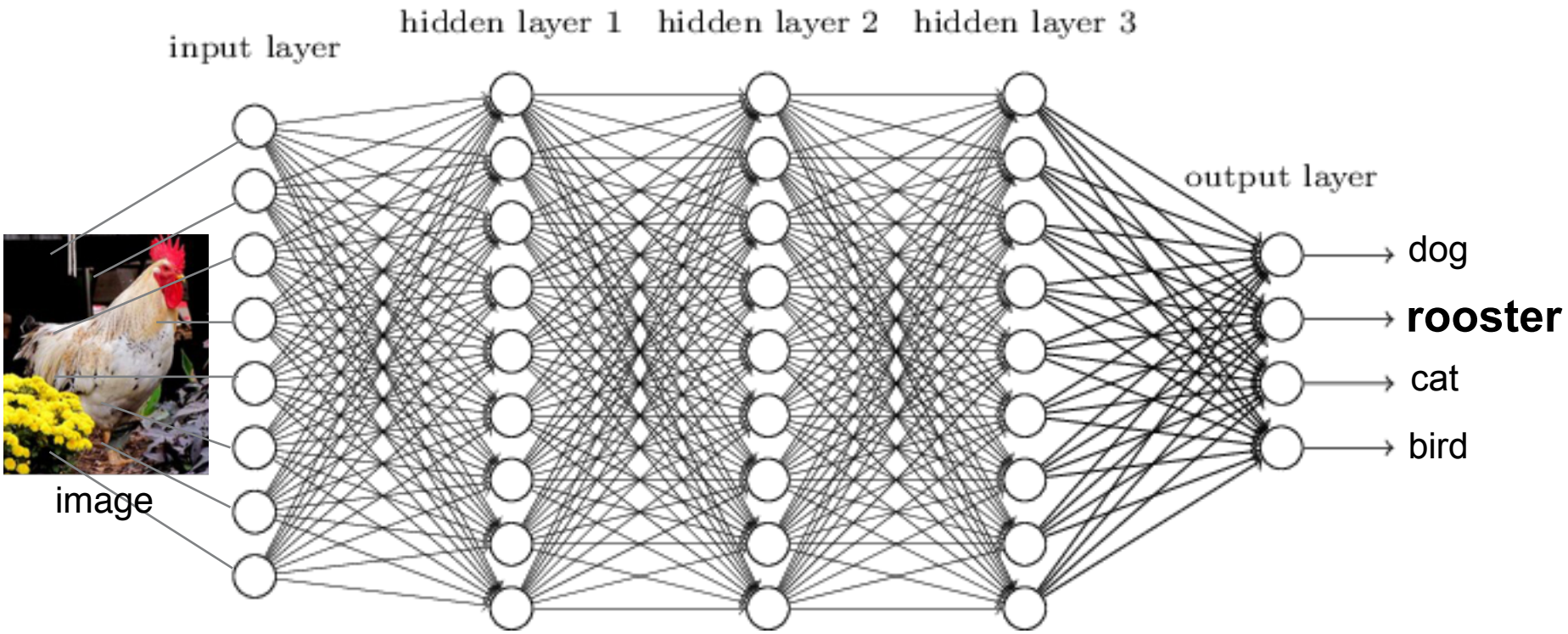
Medical Diagnosis



Autonomous Driving

Smart devices, 5G, IoT etc.

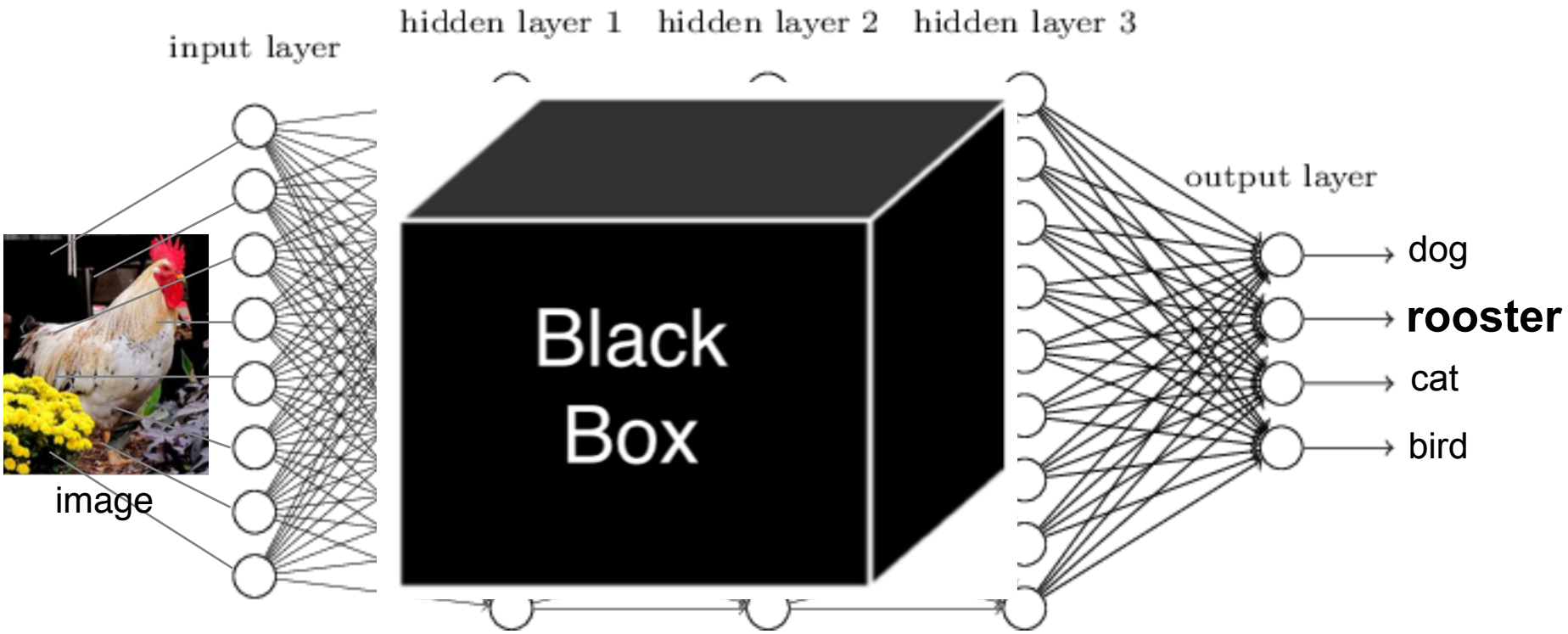
Deep Neural Networks



e.g. 20 of layers,
150 Mio. parameters

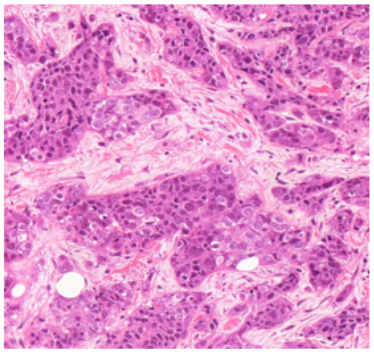
$$x_j = \sigma(\sum_i x_i w_{ij} + b_j)$$

Deep Neural Networks

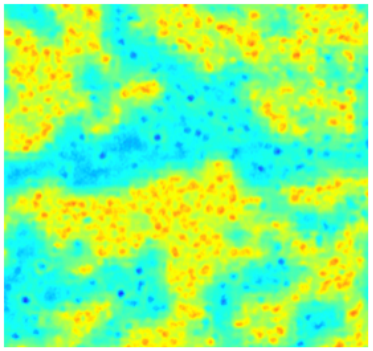


Why is the image classified as rooster?

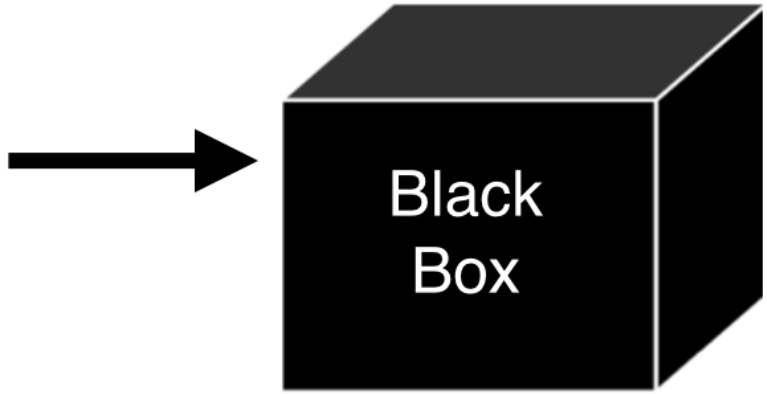
Opening the Black Box



input image



heatmap



Idea: Decompose function

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

Explain prediction

(how much each pixel contributes to prediction)

trust & verification

improve system

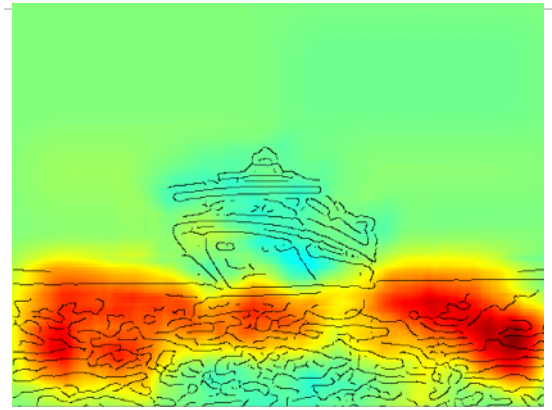
legal aspects

new insights

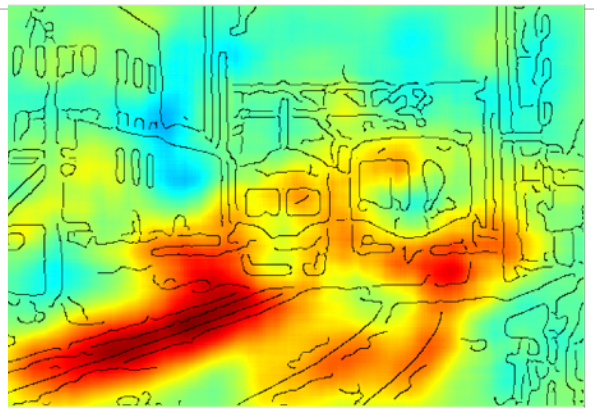
nature COMMUNICATIONS

Unmasking Clever Hans Predictors

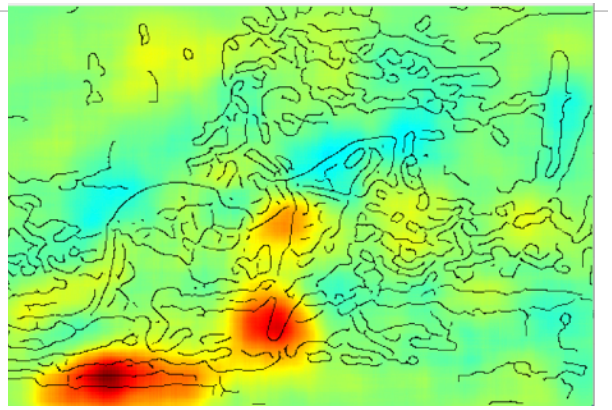
Prediction: "Boat"



Prediction: "Train"

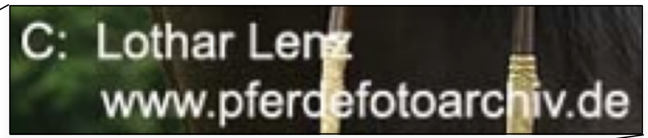


Prediction: "Horse"

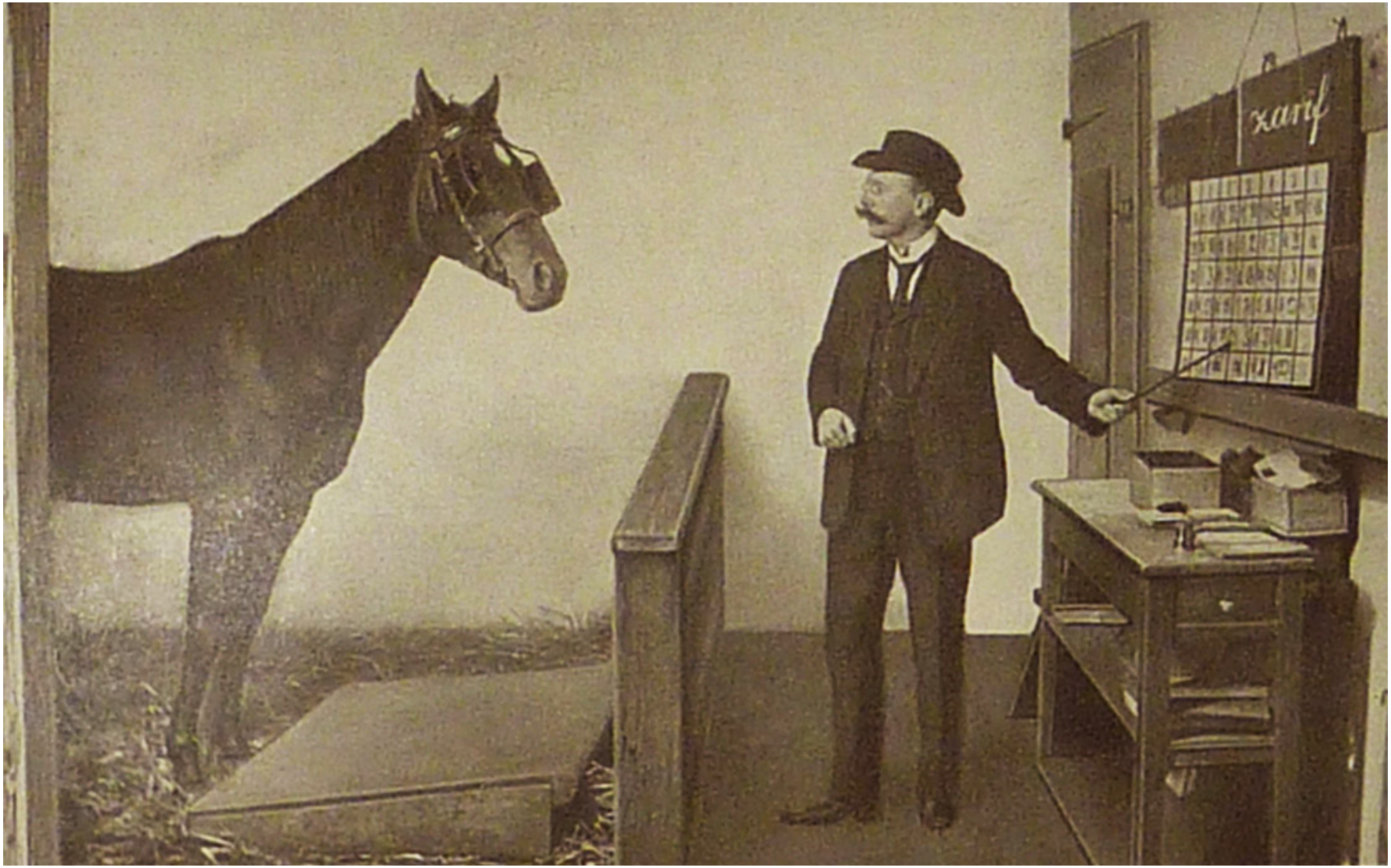


Unmasking Clever Hans Predictors

‘horse’ images in PASCAL VOC 2007

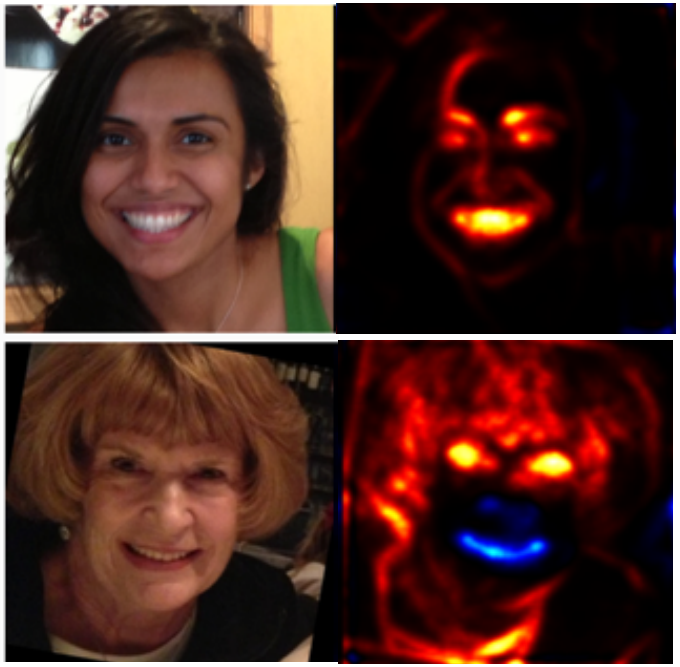


Unmasking Clever Hans Predictors



Identifying Biases

Neural Network-Based Age prediction



Predictions

25-32 years old

Laughing **speaks for** age 25-32

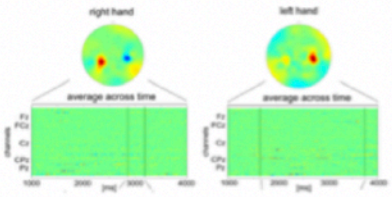
60+ years old

Laughing **speaks against** age 60+

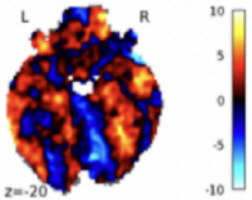
(Lapuschkin et al. 2017)

Applications in Health

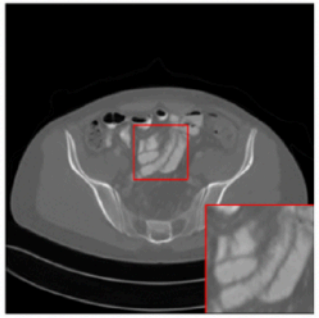
EEG (Sturm'16)



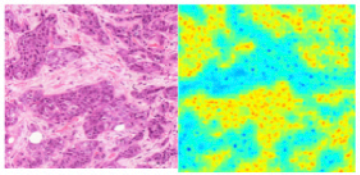
fMRI (Thomas'18)



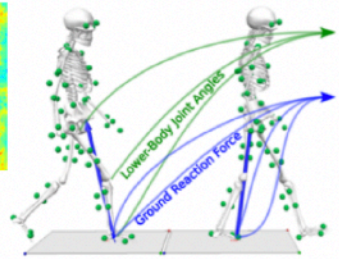
Limited Angle Tomography (März'19)



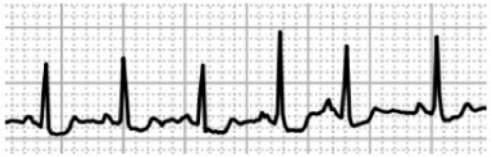
Histopathology (Hägele'19)



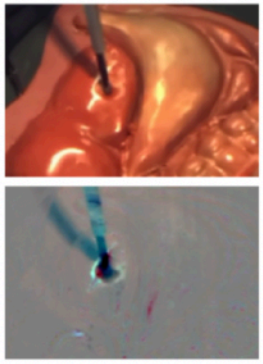
Gait Analysis (Horst'19)



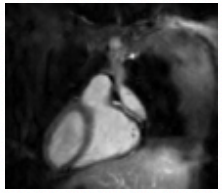
ECG (Strodthoff'18)



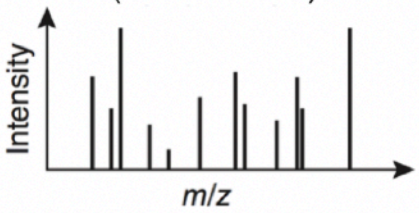
Robotic Surgery (Marban'17)



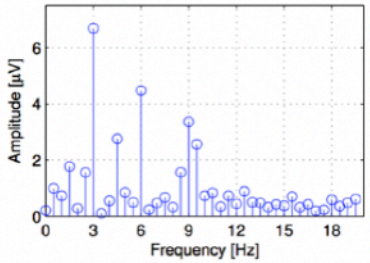
3D MRT (Ma'18)



Proteomics (Strodthoff'19)



SSVEP (Bosse'17)



Neutrophile Analysis (Wagner'19)

Robust & Trustworthy AI

Relying on the test error is often not enough.

Interpretability helps to
- understand the model
- identify biases and flaws in the data
- compare different training procedures

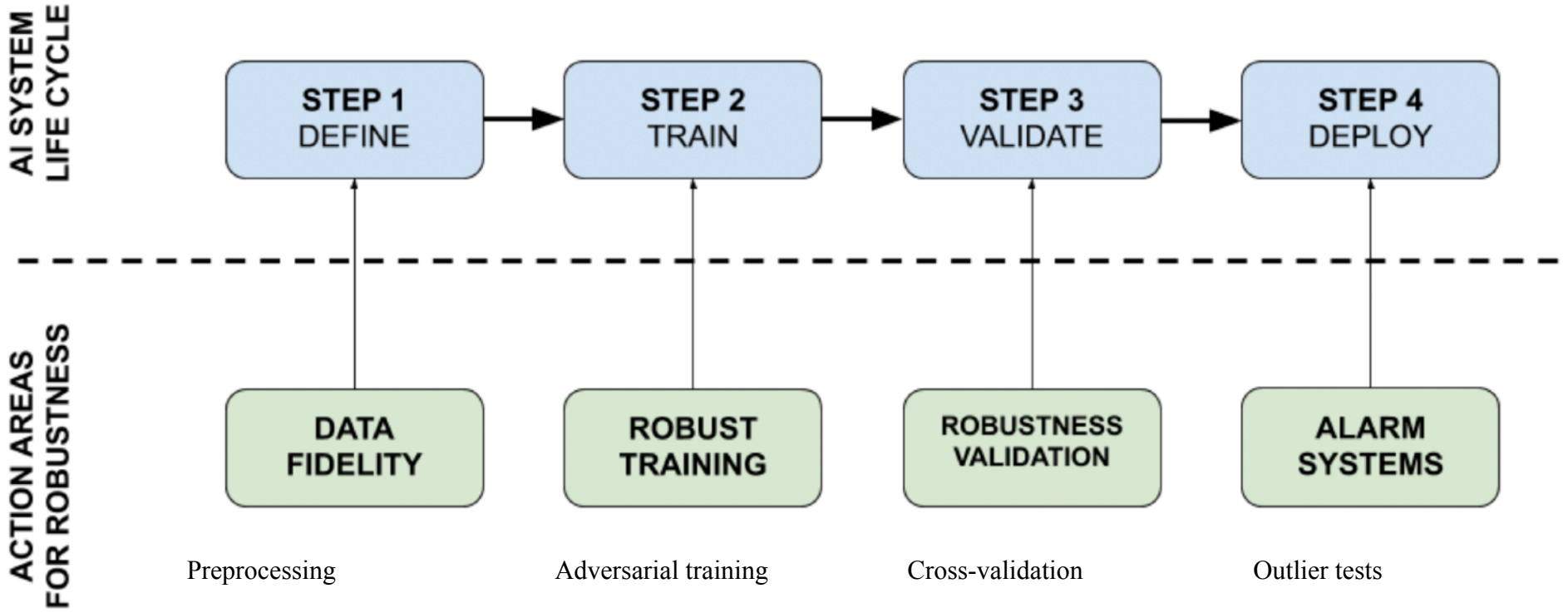
But is it enough to ensure robust and safe AI ?

ITU/WHO Focus Group on Artificial Intelligence for Health

More information about the group:

`https://www.itu.int/en/ITU-T/focusgroups/ai4h`

Robust & Trustworthy AI



Thank you for your attention

Questions ???

Contact Information:

Wojciech Samek
Machine Learning Group
Fraunhofer HHI
Einsteinufer 37, 10587 Berlin, Germany

Phone: +49 30 31002-417
Mail: wojciech.samek@hhi.fraunhofer.de
Web: http://iphome.hhi.de/samek