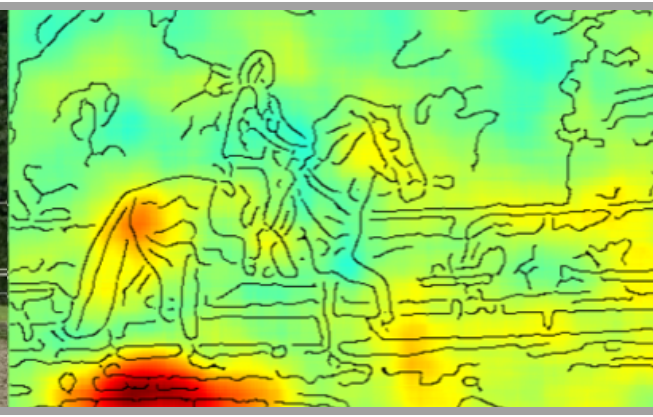
Fraunhofer
Heinrich Hertz Institute

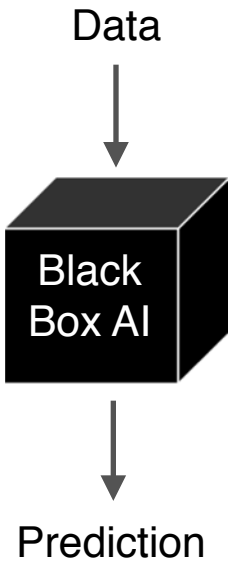
How to make the Black Box of Neural Networks Transparent – The Path towards Explainable AI

Wojciech Samek

ML Group, Fraunhofer HHI



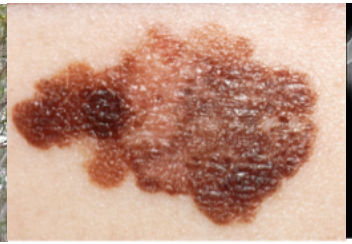
Black Box AIs Achieve “Superhuman” Performances



Game GO

Traffic Sign Recognition

Skin cancer detection



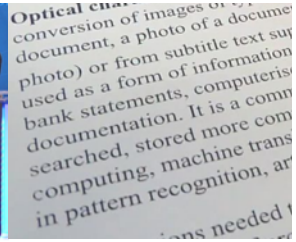
Lung cancer detection

Poker

Computer games

Jeopardy

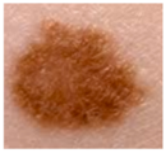
OCR

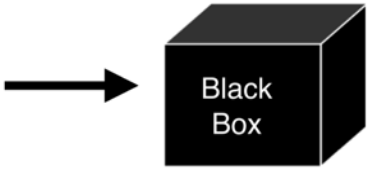


Disadvantages of a Black Box System

TRUST

Can we base important decisions on it?

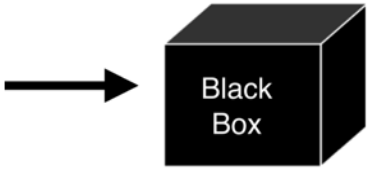




"everything ok, no therapy needed"

SAFETY

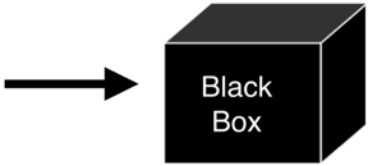
Can we be sure that it resists manipulation?



"don't stop, you have right of way"

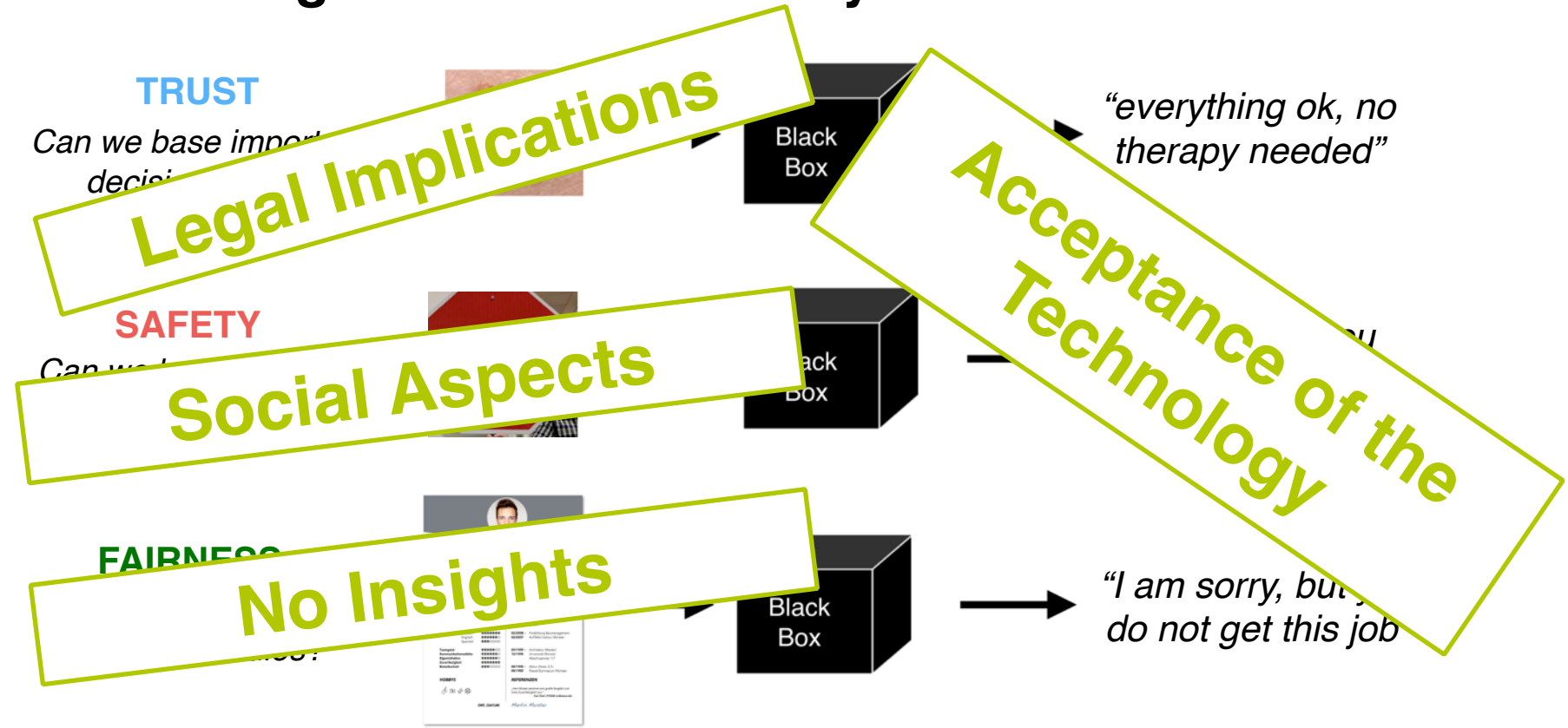
FAIRNESS

Can we be sure that it follows the rules?



"I am sorry, but you do not get this job"

Disadvantages of a Black Box System

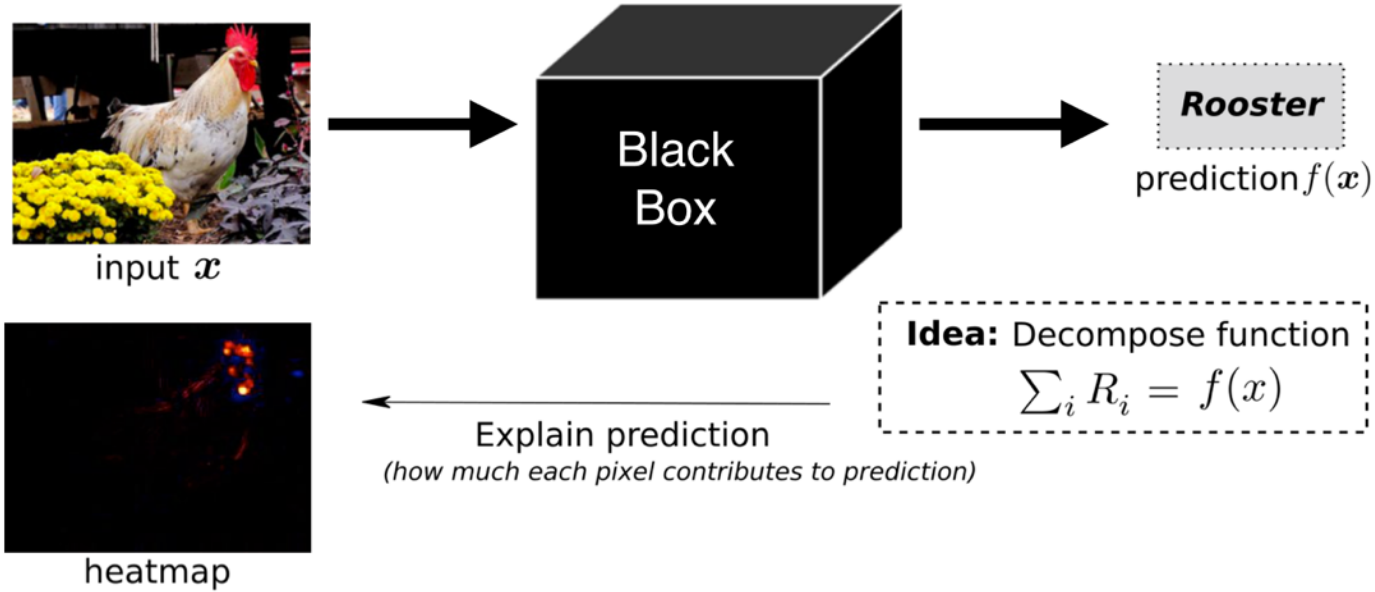


“Right to Explanation”

Recital 71 (Profiling)

“such processing [...] should include [...] the right to obtain an explanation of the decision reached after such assessment and to challenge the decision”

The Path Towards Explainable AI



Layer-wise Relevance Propagation is a general approach to explain predictions of AI.
Mathematical Interpretation: Deep Taylor Decomposition of the neural network.

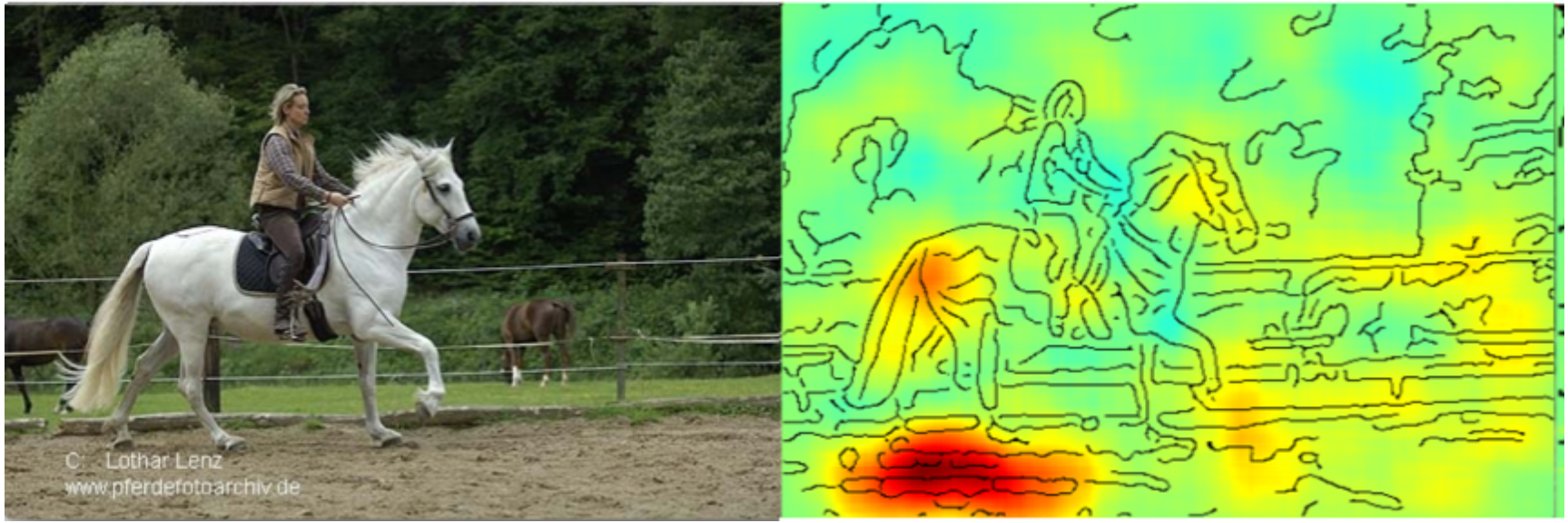
(Bach et al., PLOS ONE, 2015)

Two Examples from Our Research

Pascal VOC Challenges 2005-2012

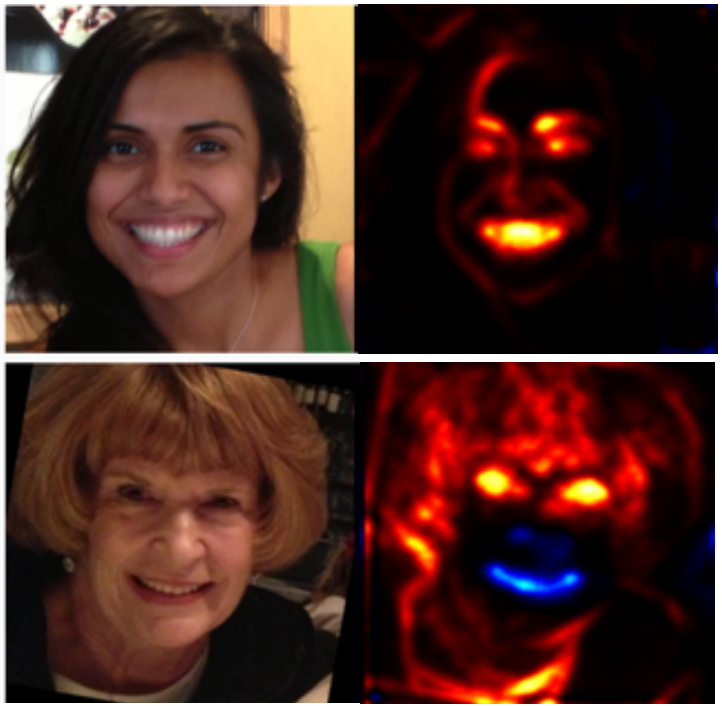
	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor
INRIA_Flat	*74.8*	*62.5*	*51.2*	*69.4*	*29.2*	*60.4*	*76.3*	*57.6*	*53.1*	*41.1*	*54.0*	*42.8*	*76.5*	*62.3*	*84.5*	*35.3*	*41.3*	*50.1*	*77.6*	*49.3*
INRIA_Genetic	**77.5**	**63.6**	**56.1**	**71.9**	**33.1**	**60.6**	**78.0**	**58.8**	**53.5**	**42.6**	**54.9**	**45.8**	**77.5**	**64.0**	**85.9**	**36.3**	**44.7**	*50.6*	**79.2**	**53.2**
INRIA_Larlus	62.6	54.0	32.8	47.5	17.8	46.4	69.6	44.2	44.6	26.0	38.1	34.0	66.0	55.1	77.2	13.1	29.1	36.7	62.7	43.3
MPI_BOW	58.9	46.0	31.3	59.0	16.9	40.5	67.2	40.2	44.3	28.3	31.9	34.4	63.6	53.5	75.7	22.3	26.6	35.4	60.6	40.6
PRIPUVA	48.6	20.9	21.3	17.2	6.4	14.2	45.0	31.4	27.4	12.3	14.3	23.7	30.1	13.3	62.0	10.0	12.4	13.3	26.7	26.2
QMUL_HSLS	70.6	54.8	35.7	64.5	27.8	51.1	71.4	54.0	46.6	36.6	34.4	39.9	71.5	55.4	80.6	15.8	35.8	41.5	73.1	45.5
QMUL_LSPCH	71.6	55.0	41.1	65.5	27.2	51.1	72.2	*55.1*	47.4	35.9	37.4	41.5	71.5	57.9	80.8	15.6	33.3	41.9	*76.5*	45.9
TKK	71.4	51.7	48.5	63.4	27.3	49.9	70.1	51.2	51.7	32.3	46.3	41.5	72.6	*60.2*	82.2	31.7	30.1	39.2	71.1	41.0
ToshCam_rdf	59.9	36.8	29.9	40.0	23.6	33.3	60.2	33.0	41.0	17.8	33.2	33.7	63.9	53.1	77.9	29.0	27.3	31.2	50.1	37.6
ToshCam_svm	54.0	27.1	30.3	35.6	17.0	22.3	58.0	34.6	38.0	19.0	27.5	32.4	48.0	40.7	78.1	23.4	21.8	28.0	45.5	31.8
Tsinghua	62.9	42.4	33.9	49.7	23.7	40.7	62.0	35.2	42.7	21.0	38.9	34.7	65.0	48.1	76.9	16.9	30.8	32.8	58.9	33.1
UVA_Bigrams	61.2	33.2	29.4	45.0	16.5	37.6	54.6	31.3	39.9	17.2	31.4	30.6	61.6	42.4	74.6	14.5	20.9	23.5	49.9	30.0
UVA_FuseAll	67.1	48.1	43.3	58.1	19.9	46.3	61.8	41.9	48.4	27.8	41.9	38.5	69.8	51.4	79.4	32.5	31.9	36.0	66.2	40.3
UVA_MCIP	66.5	47.9	41.0	58.0	16.8	44.0	61.2	40.5	48.5	27.8	41.7	37.1	66.4	50.1	78.6	31.2	32.3	31.9	66.6	40.3
UVA_SFS	66.3	49.7	43.5	60.7	18.8	44.9	64.8	41.9	46.8	24.9	42.3	33.9	71.5	53.4	80.4	29.7	31.2	31.8	67.4	43.5
UVA_WGT	59.7	33.7	34.9	44.5	22.2	32.9	55.9	36.3	36.8	20.6	25.2	34.7	65.1	40.1	74.2	26.4	26.9	25.1	50.7	29.7
XRCE	*72.3*	*57.5*	*53.2*	*68.9*	*28.5*	*57.5*	*75.4*	50.3	*52.2*	*39.0*	*46.8*	*45.3*	*75.7*	58.5	*84.0*	*32.6*	*39.7*	**50.9**	75.1	*49.5*

Pascal VOC Challenges 2005-2012



(Lapuschkin et al., IEEE CVPR, 2016)

Age & Gender Prediction (2017)



Predictions

25-32 years old

“Laughing speaks for prediction 25-23 years”

60+ years old

“Laughing speaks against prediction 60+ years”

(Lapuschkin et al., IEEE ICCVW, 2017)

Questions ???

All our papers available on:

http://iphome.hhi.de/samek

Acknowledgement

Klaus-Robert Müller (TUB)
Grégoire Montavon (TUB)
Sebastian Lapuschkin (HHI)
Leila Arras (HHI)
Alexander Binder (SUDT)
…