

LV 706.046 Summer Term 2019

Monday, March, 18

## From measuring Usability to measuring Causability: Methods of Explainable AI

Assoc.Prof. Dr. Andreas Holzinger

Holzinger-Group, HCI-KDD, Institute for Medical Informatics/Statistics

Medical University Graz, Austria

&

Institute of interactive Systems & Data Science

Graz University of Technology, Austria

[a.holzinger@tugraz.at](mailto:a.holzinger@tugraz.at)

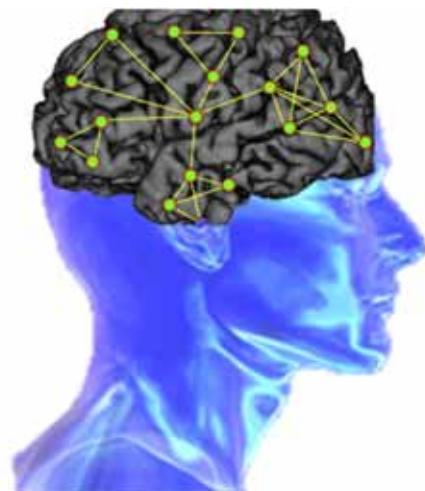
<https://hci-kdd.org/intelligent-user-interfaces-2019>



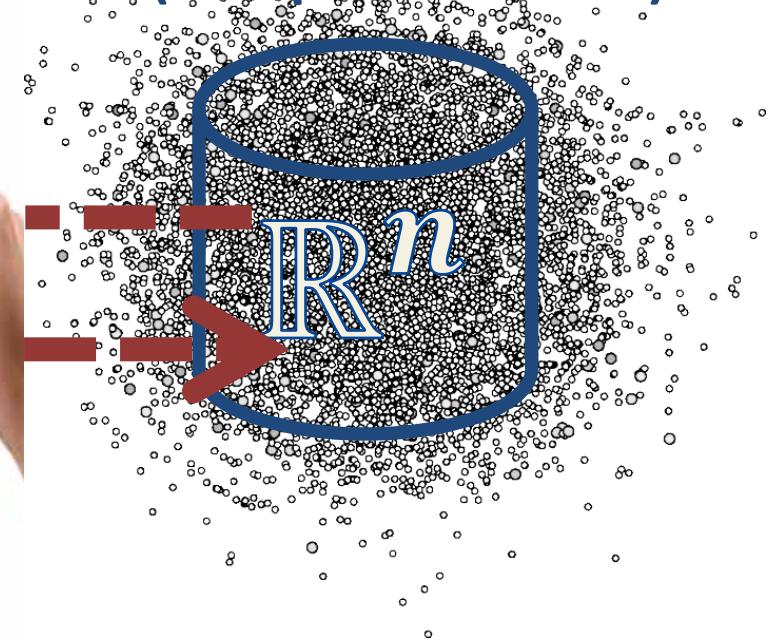
# Our Goal in this AK: design, develop & test a System Causability Scale

- **Causability := a property of a person (Human)**
- **Explainability := a property of a system (Computer)**

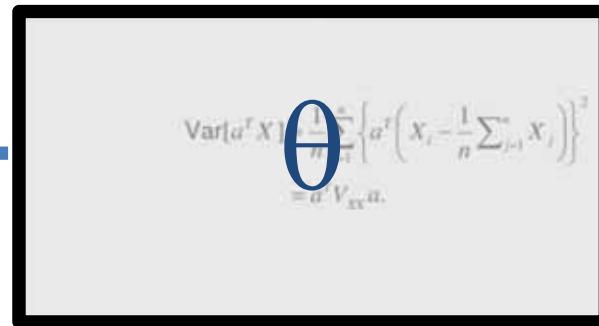
Human intelligence  
(Cognitive Science)



Machine intelligence  
(Computer Science)



*Why did the algorithm do that?  
Can I trust these results?  
How can I correct an error?*

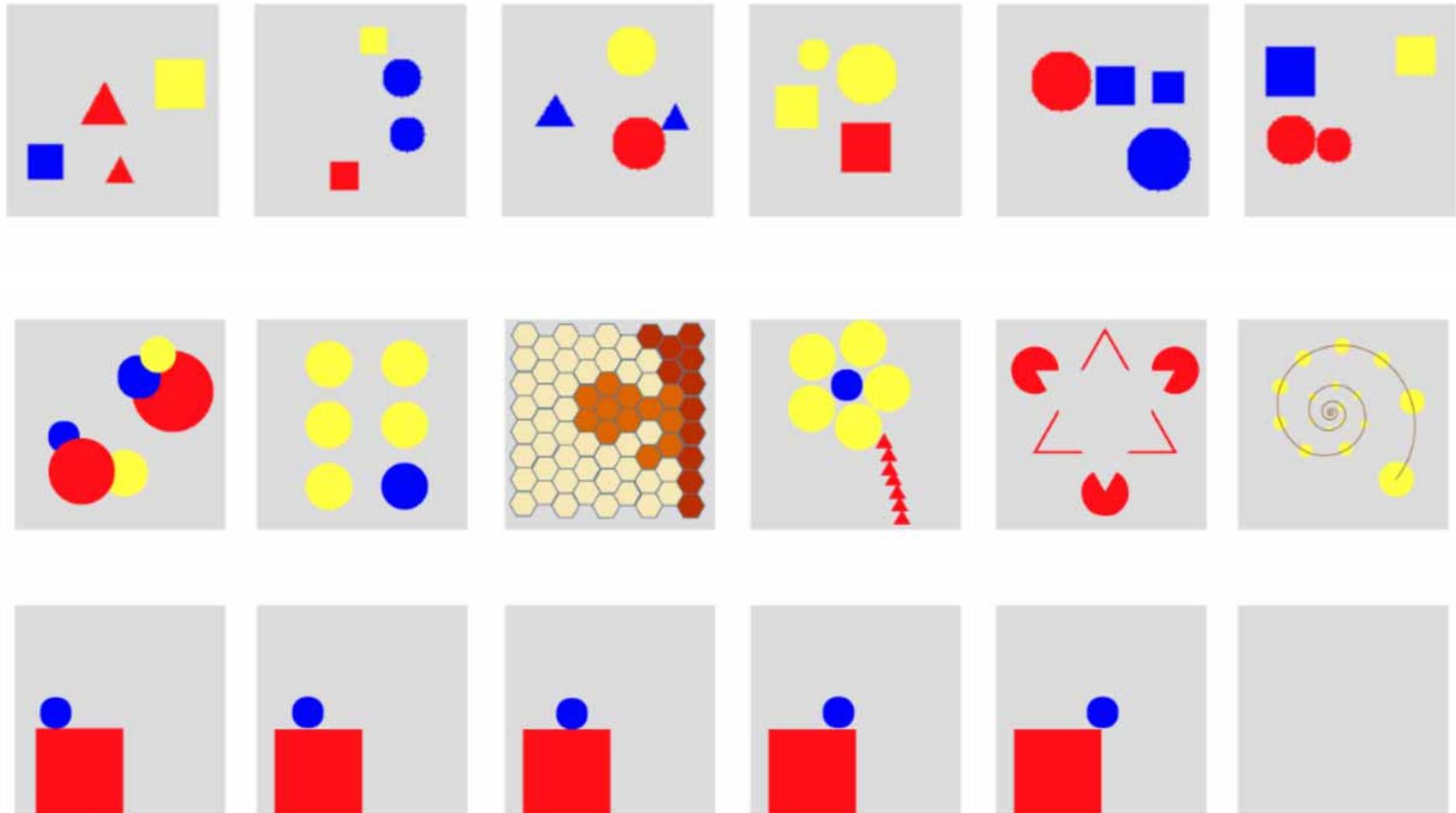


Input data

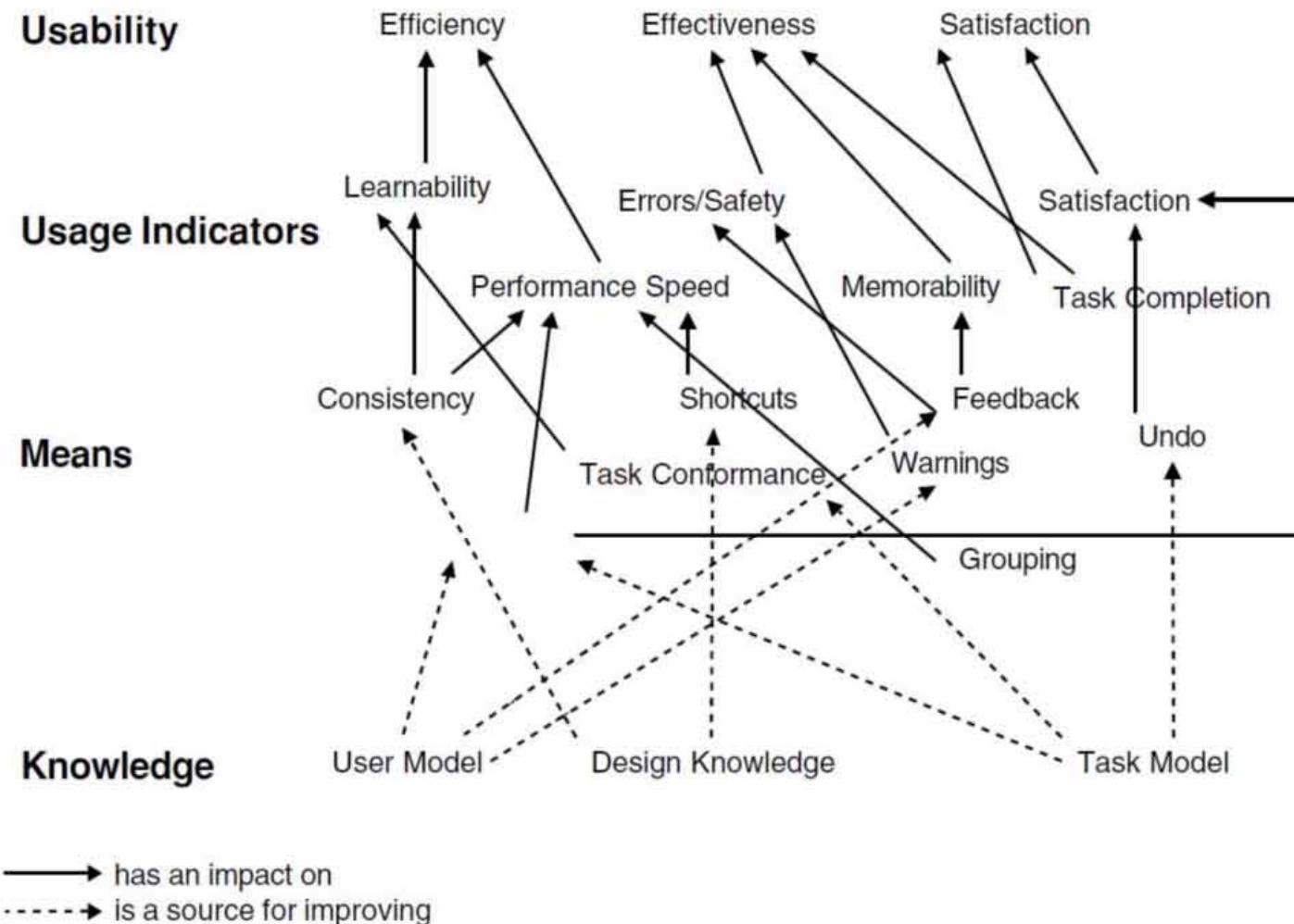
### A possible solution



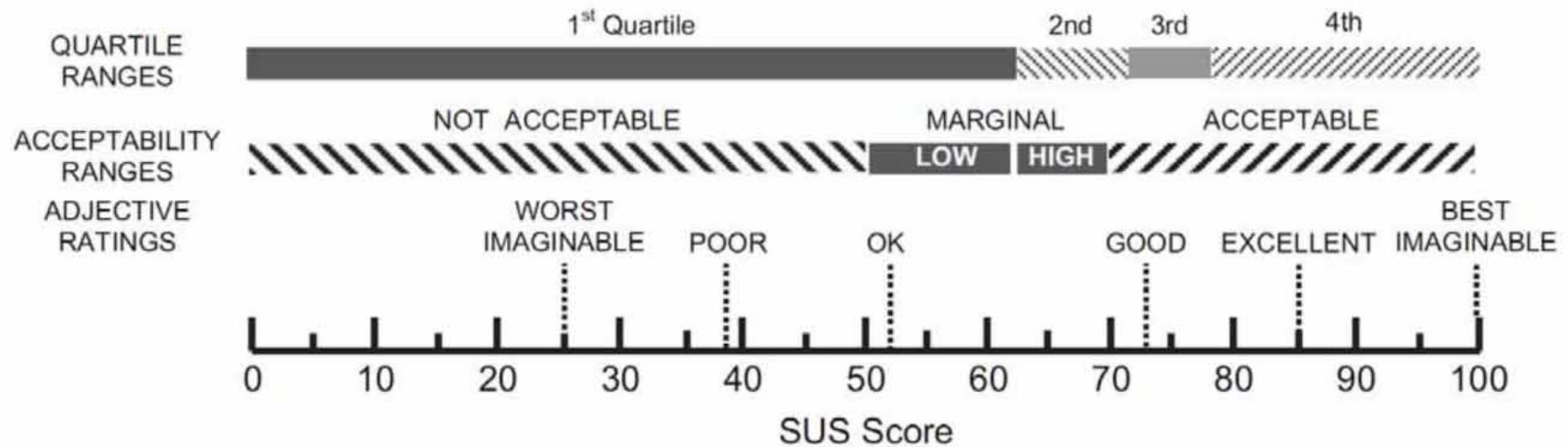
*The domain expert can understand why ...  
The domain expert can learn and correct errors ...  
The domain expert can re-enact on demand ...*



- <https://www.tensorflow.org/tutorials>



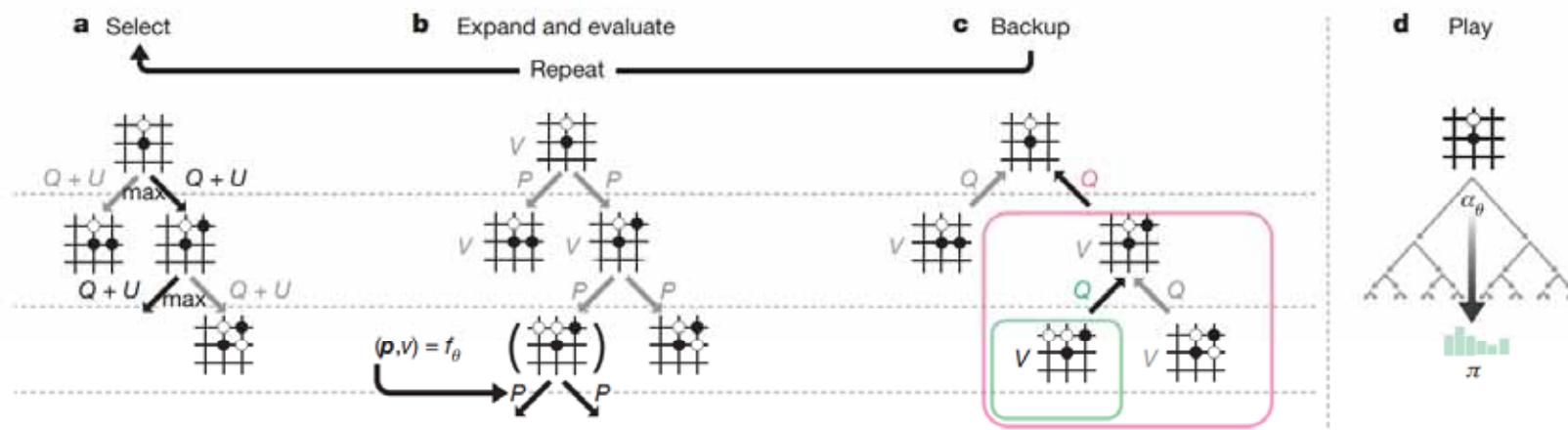
Veer, G. C. v. d. & Welie, M. v. (2004) DUTCH: Designing for Users and Tasks from Concepts to Handles. In: Diaper, D. & Stanton, N. (Eds.) *The Handbook of Task Analysis for Human-Computer Interaction*. Mahwah (New Jersey), Lawrence Erlbaum, 155-173.



Bangor, A., Kortum, P. T. & Miller, J. T. (2008) An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24, 6, 574-594.

The System Usability Scale Standard Version		Strongly Disagree	Strongly Agree			
		1	2	3	4	5
1	I think that I would like to use this system frequently.		0	0	0	0
2	I found the system unnecessarily complex.		0	0	0	0
3	I thought the system was easy to use.		0	0	0	0
4	I think that I would need the support of a technical person to be able to use this system.		0	0	0	0
5	I found the various functions in this system were well integrated.		0	0	0	0
6	I thought there was too much inconsistency in this system.		0	0	0	0
7	I would imagine that most people would learn to use this system very quickly.		0	0	0	0
8	I found the system very awkward to use.		0	0	0	0
9	I felt very confident using the system.		0	0	0	0
10	I needed to learn a lot of things before I could get going with this system.		0	0	0	0

# Explainability



**Figure 2 | MCTS in AlphaGo Zero.** **a**, Each simulation traverses the tree by selecting the edge with maximum action value  $Q$ , plus an upper confidence bound  $U$  that depends on a stored prior probability  $P$  and visit count  $N$  for that edge (which is incremented once traversed). **b**, The leaf node is expanded and the associated position  $s$  is evaluated by the neural network  $(P(s, \cdot), V(s)) = f_\theta(s)$ ; the vector of  $P$  values are stored in

the outgoing edges from  $s$ . **c**, Action value  $Q$  is updated to track the mean of all evaluations  $V$  in the subtree below that action. **d**, Once the search is complete, search probabilities  $\pi$  are returned, proportional to  $N^{1/\tau}$ , where  $N$  is the visit count of each move from the root state and  $\tau$  is a parameter controlling temperature.

19 OCTOBER 2017 | VOL 550 | NATURE | 355

$$(p, v) = f_\theta(s) \text{ and } l = (z - v)^2 - \pi^T \log p + c \|\theta\|^2$$

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel & Demis Hassabis 2017. Mastering the game of go without human knowledge. Nature, 550, (7676), 354-359, doi:doi:10.1038/nature24270.



David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, (7587), 484-489, doi:10.1038/nature16961.

Message Predictor 1.0.5.28868

Move message to folder... Only show predictions that just changed → OFF Search Stanley Clear

**Folders**

Original order	Subject	Predicted topic	Prediction confidence
9287	Re: Playoff Predictions	Hockey	99%
9294	Re: Schedule...	Baseball	60% <span style="color: green;">▲</span>
9306	Paul Kuryla and Canadian Worm	Hockey	99%
9308	Re: My Predictions For 1993	Baseball	64% <span style="color: green;">▲</span>
9312	Re: NHL Team Captains	Baseball	64% <span style="color: green;">▲</span>
9316	Re: ugliest swing	Baseball	63% <span style="color: green;">▲</span>
9319	Re: Octopus in Detroit?	Hockey	67% <span style="color: red;">▼</span>
9339	Sparky Anderson Gets win #2000, Tigers beat A's	Baseball	99%
9347	Re: Goalie masks	Baseball	53%
9362	Re: Young Catchers	Baseball	82% <span style="color: green;">▲</span>
9371	Re: Winning Streaks	Baseball	53%
9379	Royals	Baseball	64% <span style="color: green;">▲</span>
9390	Phillies Mailing List?	Baseball	65% <span style="color: green;">▲</span>
9410	Reds snap 5-game losing streak: RedReport 4-18	Baseball	98%
9423	Re: Juggling Dodgers	Baseball	57% <span style="color: green;">▲</span>
9424	Re: Candlestick Park experience (long)	Baseball	99%
9433	Re: Notes on Jays vs. Indians Series	Baseball	53%
9434	Re: When did Dodgers move from NY to LA?	Baseball	53%
9439	Playoff pool	Hockey	96%
9441	Re: Hockey and the Hispanic community	Hockey	99%
9449	Re: Yooi-isms	Baseball	53%

**Messages in the 'Unknown' folder**

**A** Unknown (1,180 messages) **B** Baseball 8/8 correct predictions

**C** Harold Zazula <DLMQC@CUNYVM.BITNET  
 >I was watching the Detroit-Minnesota game and thought I saw an  
 >octopus on the ice after Ysebaert scored a goal at two. What gives?  
 >(is there some custom to throw octopuses on ice in Detroit?)  
 It is a long standing good luck Redwing's tradition to throw an octopus  
 on the ice during a **Stanley** Cup game. They say it dates back to '52  
 at the Olympia when the Wings became the 1st team (I think) to sweep  
 the cup in 8 games. A lot harder to throw one from Joe Louis seats  
 than from the old Olympia balcony, though.

**D** Why Hockey?  
 Part 1: Important words  
 This message has important words about **Hockey** and **Baseball**  
**baseball** **hockey**  
**stanley** **tiger**

The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.

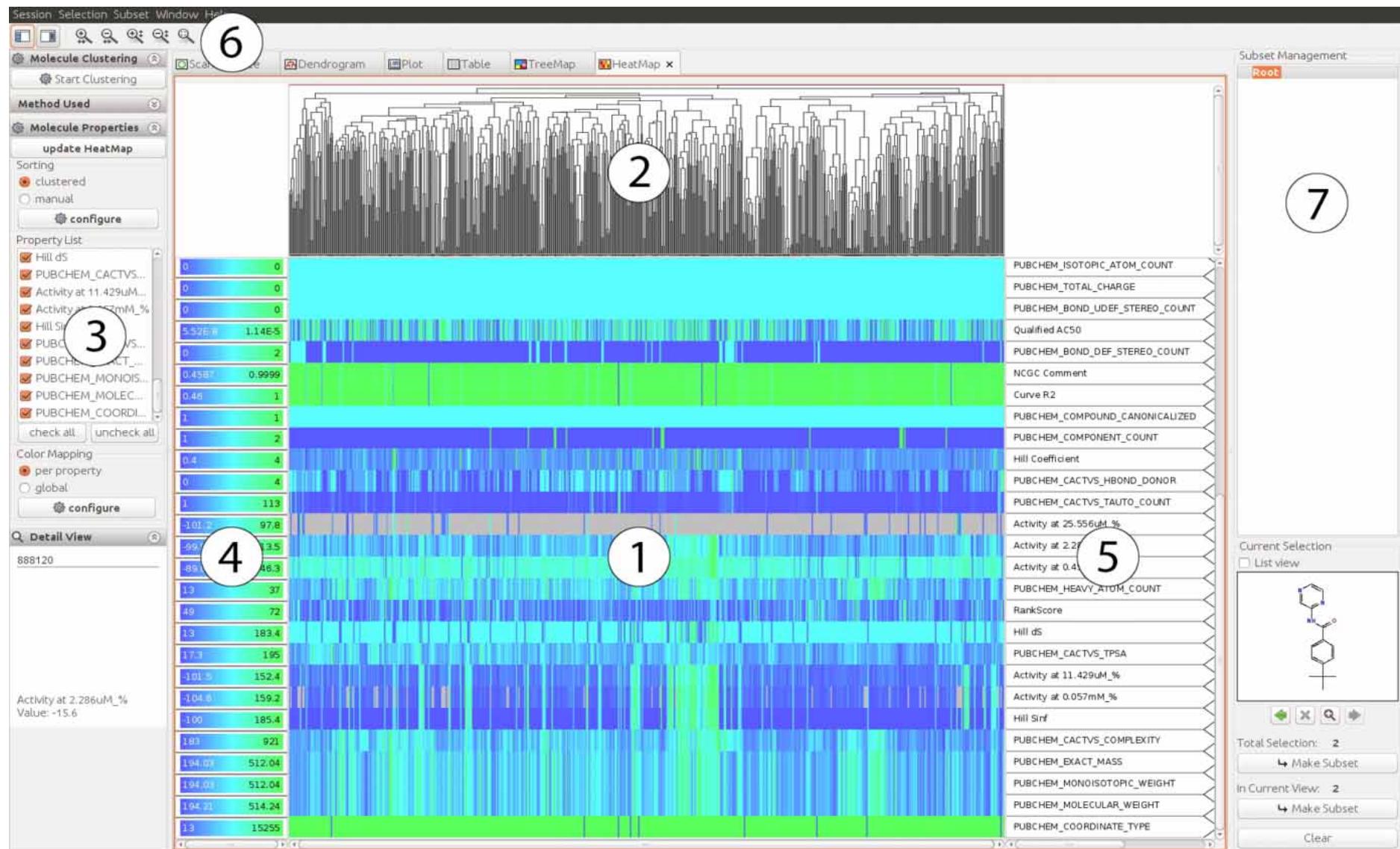
**E** Prediction totals  
 Hockey 278 ▼  
 Baseball 917 ▲

**F** Messages containing "Stanley"  
 Baseball  
 Hockey  
 Unknown

**Important words**  
 These are all of the words the computer used to make its prediction.  
 baseball bill canadian dave daniel hockey player players prime stanley stats tiger time

Add a new word or phrase  
 Remove word  
 Undo importance adjustment

Todd Kulesza, Margaret Burnett, Weng-Keen Wong & Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI 2015), 2015 Atlanta. ACM, 126-137, doi:10.1145/2678025.2701399.



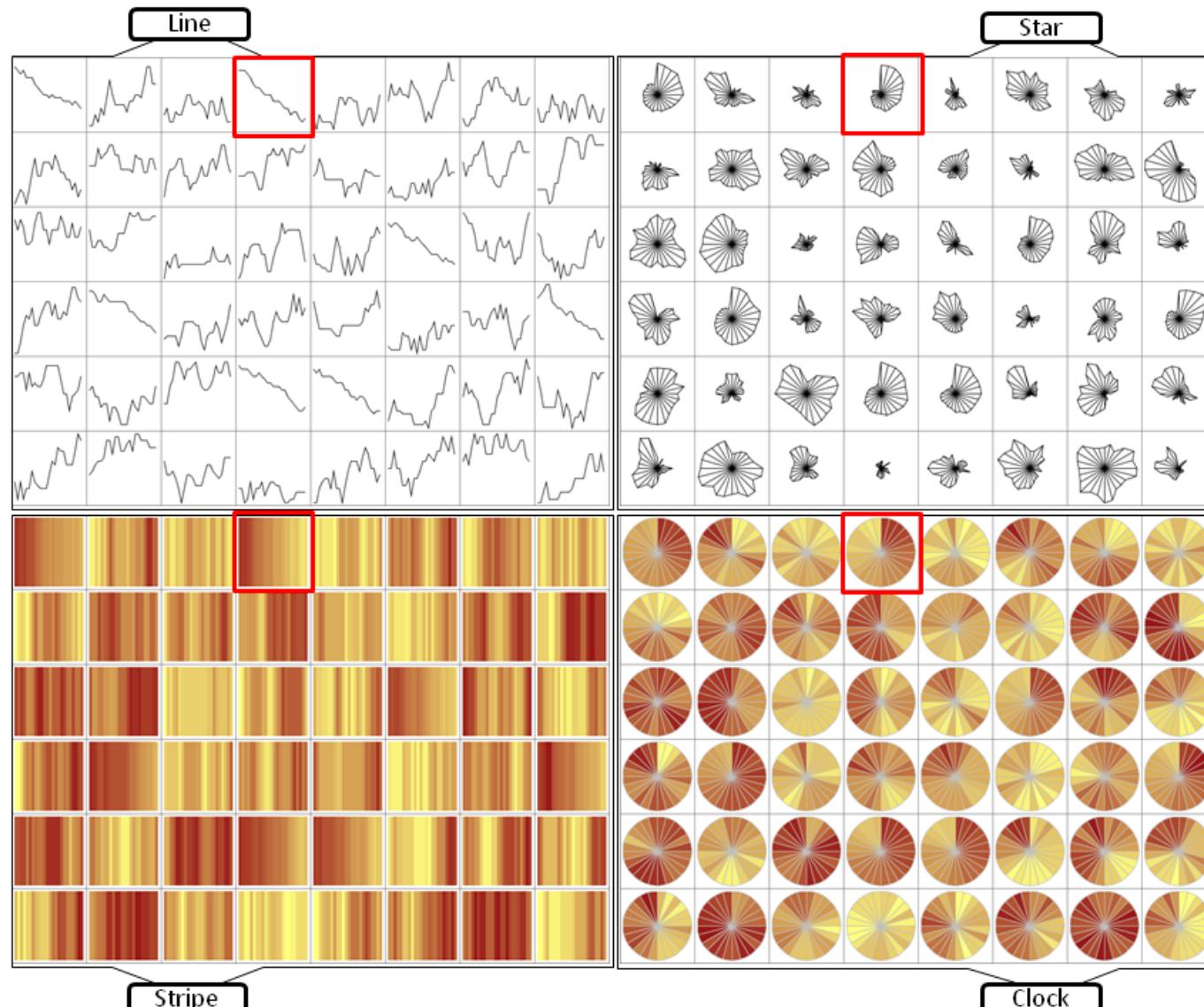
Werner Sturm, Till Schaefer, Tobias Schreck, Andreas Holzinger & Torsten Ullrich. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps In: Borgo, Rita & Turkay, Cagatay, eds. EG UK Computer Graphics & Visual Computing CGVC 2015, 2015 University College London (UCL). Euro Graphics (EG), 77-84, doi:10.2312/cgvc.20151247.

- Interpretability as a novel kind for supporting teaching, learning and knowledge discovery,
- Particularly in abstract fields (informatics)
- Compliance to European Law “the right of explanation”
- Check for bias in machine learning results
- Fostering trust, acceptance, making clear the reliability

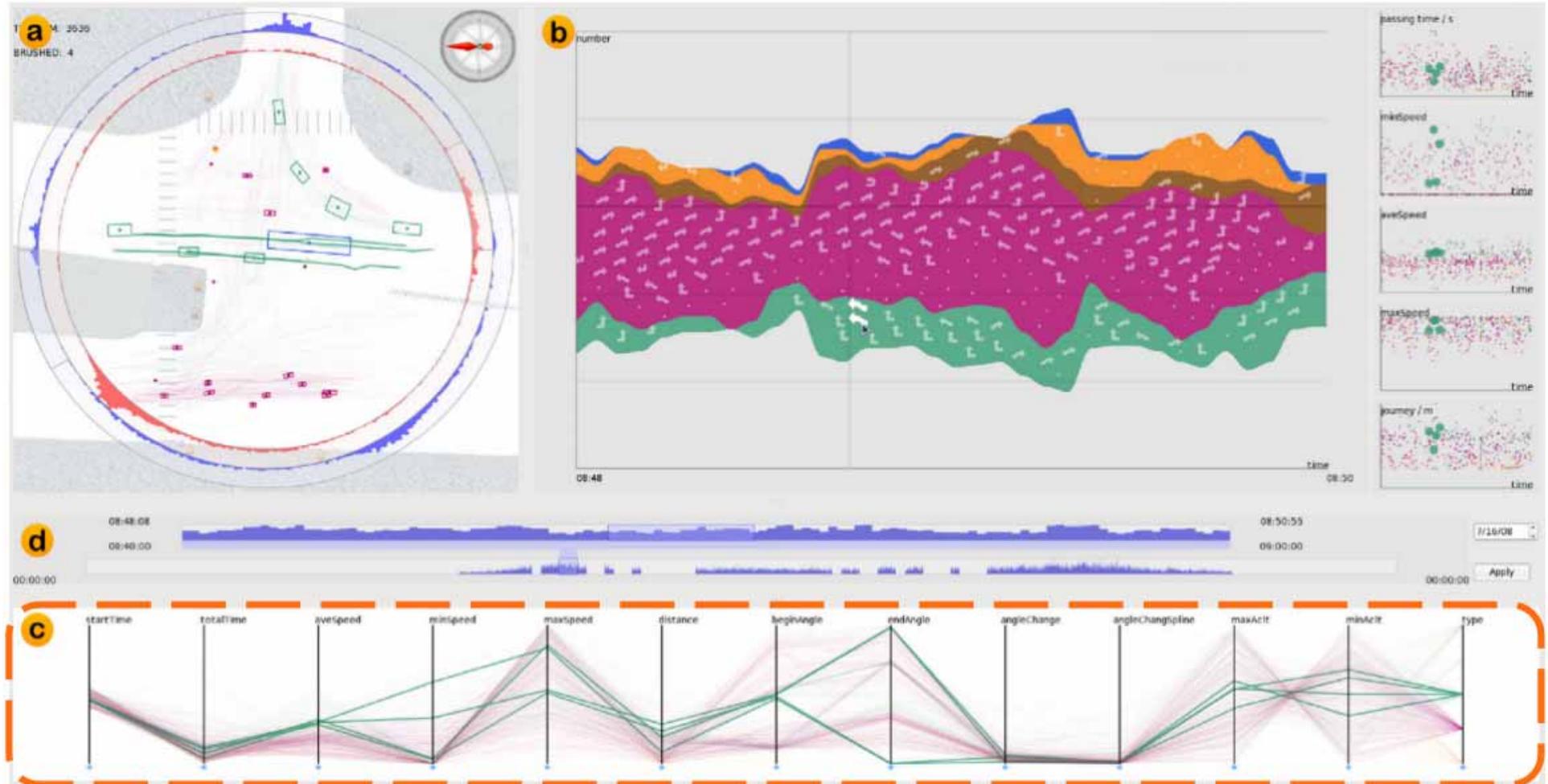
Andreas Holzinger 2018. Explainable AI (ex-AI). Informatik-Spektrum, 41, (2), 138-143,  
doi:10.1007/s00287-018-1102-5.

1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.953	0.894	0.620	0.699	0.629	0.546	0.540	1.000	0.526	1.000	0.522	0.483	0.471	1.000	0.522	0.576	0.658									
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.722	0.638	1.000	0.785	0.743	0.792	0.801	0.875	0.712	1.000	0.444	0.947	0.431	1.000	0.793	1.000	0.635									
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.658	0.633	0.569	0.561	0.589	0.640	0.659	0.845	0.932	0.512	0.575	0.941	1.000	0.991	1.000	0.892										
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.932	0.639	0.575	0.544	0.501	0.489	0.470	0.454	0.576	0.581	0.707	0.992	1.000	1.000	1.000	1.000										
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.711	0.644	0.569	0.541	0.461	0.430	0.425	0.381	0.364	0.437	0.562	0.509	0.528	0.678	1.000	0.991	1.000									
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.680	0.594	0.579	0.513	0.490	0.429	0.405	0.425	0.381	0.401	0.387	0.367	0.484	0.428	0.483	0.659	0.936	1.000								
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.761	0.677	0.610	0.565	0.511	0.498	0.457	0.416	0.396	0.388	0.369	0.355	0.359	0.468	0.392	0.380	0.487	0.4	0.500	0.744	1.000					
1.000	1.000	1.000	1.000	0.861	0.640	0.579	0.560	0.542	0.476	0.470	0.441	0.405	0.389	0.392	0.396	0.436	0.355	0.327	0.394	0.407	0.390	0.383	0.400	0.376	0.676	0.676	0.437						
1.000	1.000	1.000	0.827	0.646	0.579	0.556	0.545	0.489	0.505	0.489	0.478	0.411	0.387	0.404	0.401	0.391	0.452	0.352	0.350	0.350	0.350	0.350	0.350	0.354	0.318	0.462	0.491	0.426					
0.909	1.000	0.860	0.675	0.598	0.528	0.535	0.500	0.497	0.517	0.468	0.520	0.623	0.619	0.507	0.472	0.385	0.310	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350						
1.000	0.989	0.693	0.561	0.546	0.523	0.532	0.452	0.441	0.461	0.649	0.659	0.695	0.686	0.632	0.620	0.610	0.616	0.655	0.358	0.295	0.310	0.336	0.363	0.418	0.458								
0.969	0.849	0.606	0.530	0.521	0.494	0.437	0.396	0.421	0.626	0.698	0.741	0.737	0.730	0.729	0.720	0.718	0.680	0.565	0.506	0.435	0.358	0.311	0.299	0.313	0.402	0.488							
1.000	1.000	0.590	0.509	0.486	0.445	0.411	0.372	0.569	0.675	0.732	0.740	0.756	0.750	0.743	0.741	0.730	0.729	0.720	0.710	0.650	0.622	0.573	0.467	0.405	0.286	0.274	0.358	0.419	0.445				
1.000	0.924	0.554	0.517	0.450	0.416	0.449	0.378	0.585	0.700	0.727	0.730	0.737	0.730	0.720	0.714	0.700	0.653	0.626	0.500	0.502	0.431	0.338	0.279	0.295	0.330	0.446							
1.000	1.000	0.557	0.517	0.457	0.396	0.390	0.4	0.635	0.658	0.670	0.679	0.751	0.757	0.792	0.764	0.714	0.694	0.642	0.597	0.542	0.419	0.341	0.289	0.291	0.326	0.380							
1.000	1.000	0.556	0.4	0.42	0.385	0.52	0.623	0.63	0.670	0.711	0.748	0.771	0.775	0.772	0.724	0.594	0.4	0.434	0.378	0.354	0.414	0.307	0.282	0.278	0.402								
0.763	1.000	0.617	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.646	0.687	0.718	0.724	0.748	0.717	0.559	0.45	0.40	0.560	0.494	0.483	0.499	0.472	0.273	0.234	0.279	0.306						
1.000	1.000	0.750	0.6	0.6	0.6	0.6	0.6	0.644	0.328	0.490	0.550	0.623	0.593	0.55	0.521	0.65	0.68	0.6	0.7	0.74	0.71	0.69	0.62	0.519	0.500	0.566	0.521	0.286	0.249	0.234	0.258		
0.754	0.830	1.000	0.471	0.435	0.326	0.327	0.489	0.474	0.421	0.388	0.4	0.34	0.5	0	0.56	0.64	0.601	0.594	0.627	0.590	0.613	0.585	0.529	0.438	0.328	0.487	0.200						
0.929	0.672	0.503	0.654	0.388	0.335	0.306	0.475	0.416	0.375	0.46	0.4	0	0.74	0	0.559	0.616	0.550	0.649	0.686	0.658	0.667	0.587	0.564	0.486	0.416	0.546	0.263						
1.000	0.758	0.639	0.726	0.931	0.330	0.299	0.398	0.54	0.5	0	0.21	0.67	0.646	0.644	0.517	0.605	0.517	0.546	0.616	0.714	0.683	0.609	0.578	0.563	0.478	0.314	0.252						
1.000	0.790	0.907	0.701	0.897	0.382	0.296	0.358	0	0.63	0.28	0.674	0.683	0.666	0.605	0.526	0.620	0.527	0.514	0.616	0.666	0.670	0.628	0.549	0.512	0.262	0.321	0.254						
0.760	0.587	0.639	0.557	0.681	0.593	0.397	0.340	0.575	0.574	0.647	0.691	0.666	0.620	0.506	0.614	0.550	0.532	0.487	0.589	0.610	0.616	0.504	0.482	0.310	0.271	0.237							
0.577	0.599	0.443	0.561	0.657	0.363	0.914	0.626	0.482	0.553	0.631	0.678	0.722	0.561	0.523	0.639	0.634	0.510	0.481	0.558	0.533	0.597	0.570	0.509	0.342	0.263	0.243							
0.639	0.615	0.748	0.639	0.911	0.796	0.647	0.614	0.529	0.553	0.588	0.651	0.644	0.585	0.433	0.606	0.588	0.467	0.313	0.363	0.349	0.415	0.578	0.512	0.305	0.274	0.256							
0.569	0.661	0.486	0.605	0.448	0.494	0.705	0.730	0.579	0.532	0.526	0.623	0.518	0.387	0.310	0.338	0.466	0.378	0.559	0.479	0.444	0.430	0.494	0.465	0.232	0.248	0.237							
0.493	0.522	0.508	0.553	0.458	0.457	0.435	0.742	0.636	0.434	0.553	0.578	0.369	0.394	0.502	0.539	0.532	0.555	0.601	0.582	0.548	0.498	0.328	0.237	0.242	0.252	0.273							
0.891	0.817	0.441	0.445	0.473	0.452	0.720	0.423	0.700	0.492	0.525	0.509	0.463	0.614	0.466	0.477	0.603	0.615	0.509	0.517	0.563	0.405	0.224	0.258	0.234	0.211	0.228							
0.543	0.548	0.598	0.433	0.386	0.627	0.482	0.345	0.835	0.751	0.581	0.502	0.482	0.610	0.531	0.524	0.615	0.625	0.562	0.481	0.566	0.306	0.266	0.407	0.366	0.243	0.252							
0.762	0.720	0.506	0.496	0.495	0.698	0.396	0.627	0.555	0.317	0.491	0.294	0.382	0.393	0.572	0.449	0.405	0.407	0.357	0.567	0.518	0.243	0.255	0.465	0.415	0.323	0.248							
0.472	0.437	0.618	0.547	0.500	0.439	0.580	0.579	0.474	0.406	0.320	0.302	0.233	0.262	0.387	0.622	0.556	0.499	0.580	0.558	0.378	0.214	0.364	0.502	0.413	0.311	0.269							
0.461	0.503	0.513	0.432	0.537	0.537	0.467	0.530	0.387	0.504	0.353	0.362	0.456	0.222	0.241	0.342	0.510	0.622	0.454	0.441	0.285	0.218	0.545	0.502	0.445	0.508	0.623							
0.529	0.464	0.455	0.824	0.476	0.411	0.498	0.405	0.408	0.400	0.382	0.387	0.482	0.422	0.210	0.242	0.281	0.309	0.295	0.241	0.213	0.549	0.569	0.522	0.500	0.493	0.529							
0.383	0.458	0.482	0.370	0.384	0.361	0.400	0.391	0.320	0.319	0.425	0.377	0.433	0.528	0.497	0.285	0.247	0.198	0.226	0.410	0.570	0.597	0.576	0.588	0.531	0.493	0.546							
0.459	0.476	0.391	0.431	0.563	0.321	0.364	0.382	0.365	0.368	0.405	0.287	0.263	0.509	0.606	0.569	0.509	0.554	0.551	0.591	0.622	0.647	0.612	0.648	0.594	0.537	0.546							

What is interpretable  
for humans?



<https://www.vis.uni-konstanz.de/en/members/fuchs/>



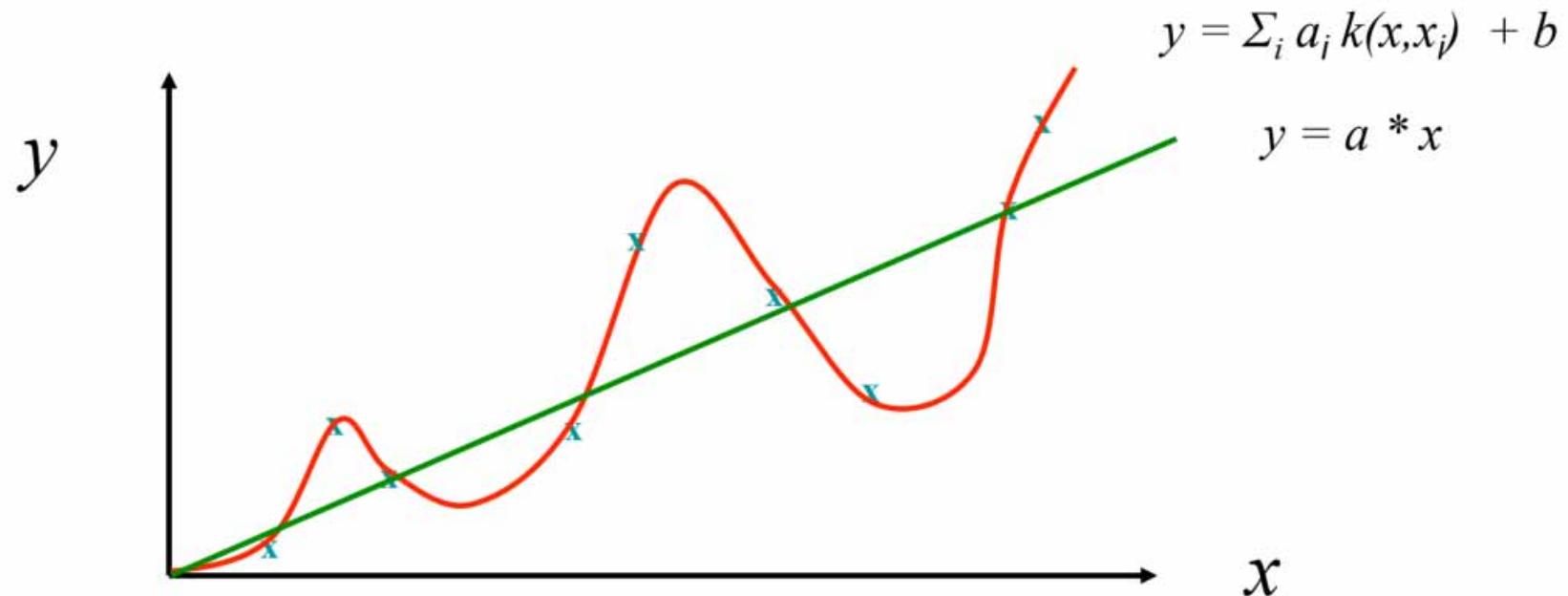
# Methods of Explainable AI

- 1) Gradients
- 2) Sensitivity Analysis
- 3) Decomposition Relevance Propagation  
(Pixel-RP, Layer-RP, Deep Taylor Decomposition, ...)
- 4) Optimization (Local-IME – model agnostic,  
BETA transparent approximation, ...)
- 5) Deconvolution and Guided Backpropagation
- 6) Model Understanding
  - Feature visualization, Inverting CNN
  - Qualitative Testing with Concept Activation Vectors TCAV
  - Network Dissection

Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course at Graz University of Technology  
<https://hci-kdd.org/explainable-ai-causability-2019> (course given since 2016)

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises:  $A=B$ ,  $B=C$ , therefore  $A=C$
- **Inductive reasoning** = makes broad generalizations from specific observations
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions
- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = drawing conclusions from empirical data (observations, measurements)
  - Causal inference = drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
    - Causal inference is an example of causal reasoning.



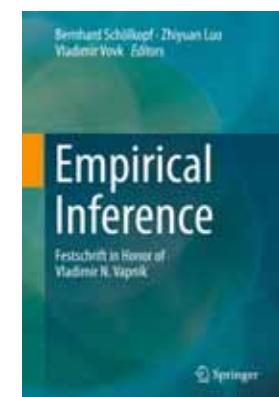
Gottfried W. Leibniz (1646-1716)

Hermann Weyl (1885-1955)

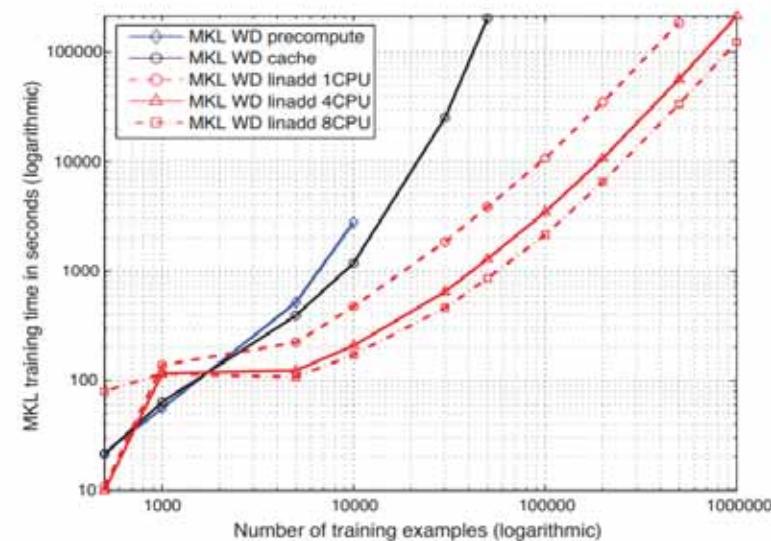
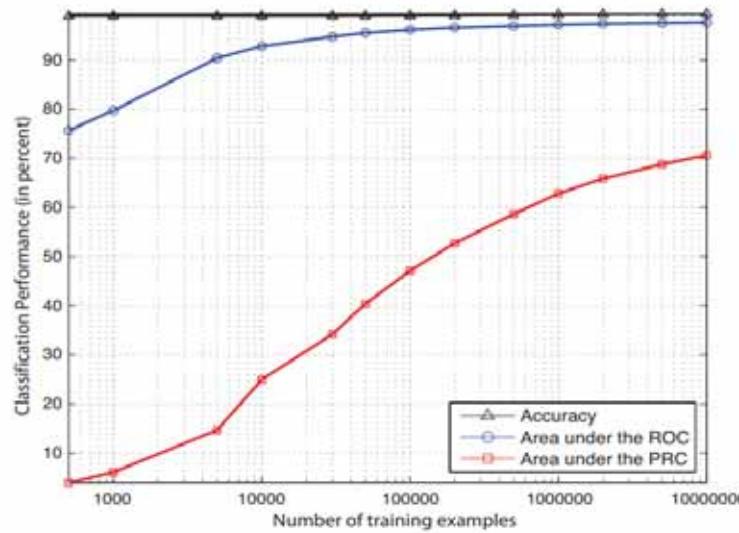
Vladimir Vapnik (1936-)

Alexey Chervonenkis (1938-2014)

Gregory Chaitin (1947-)

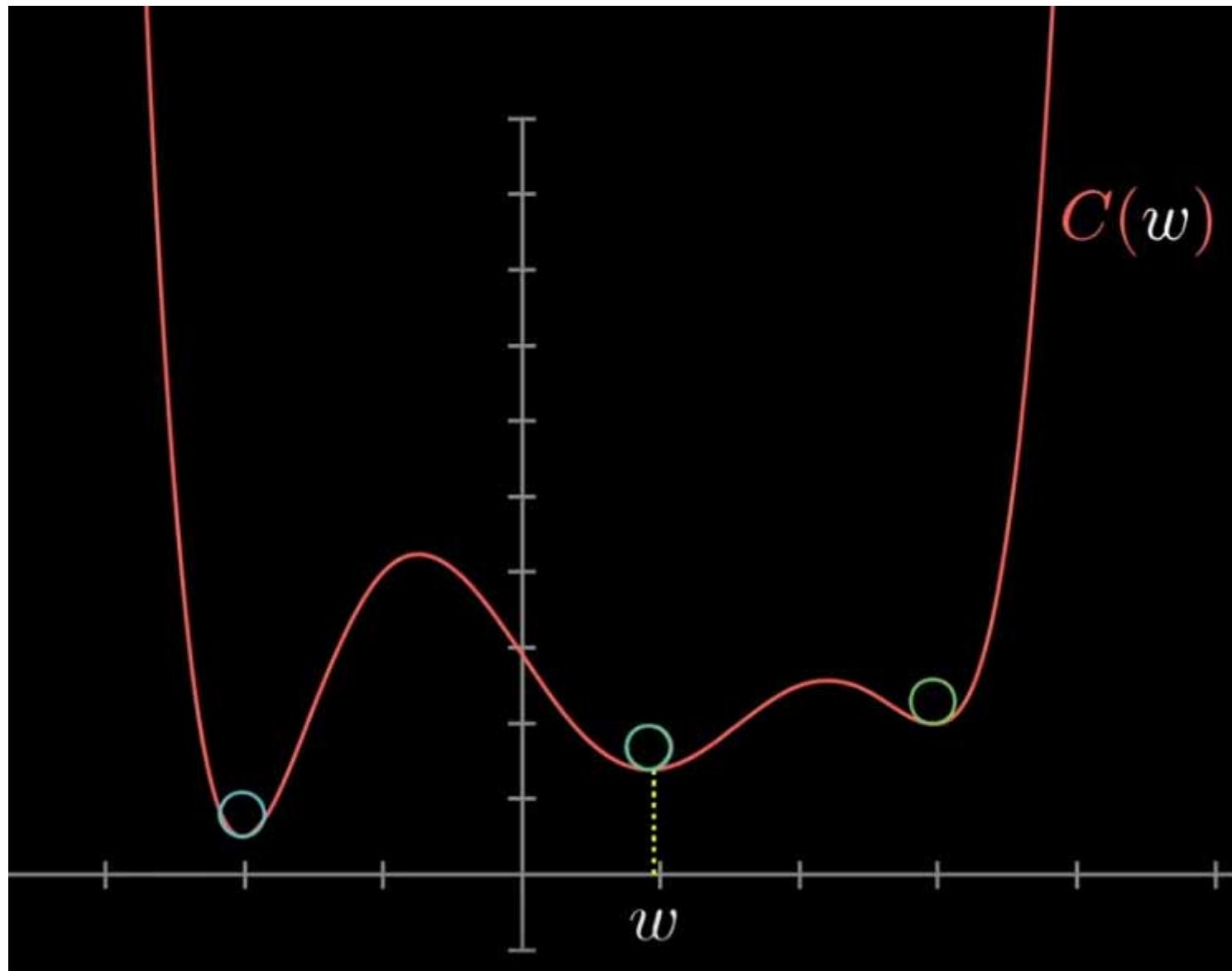


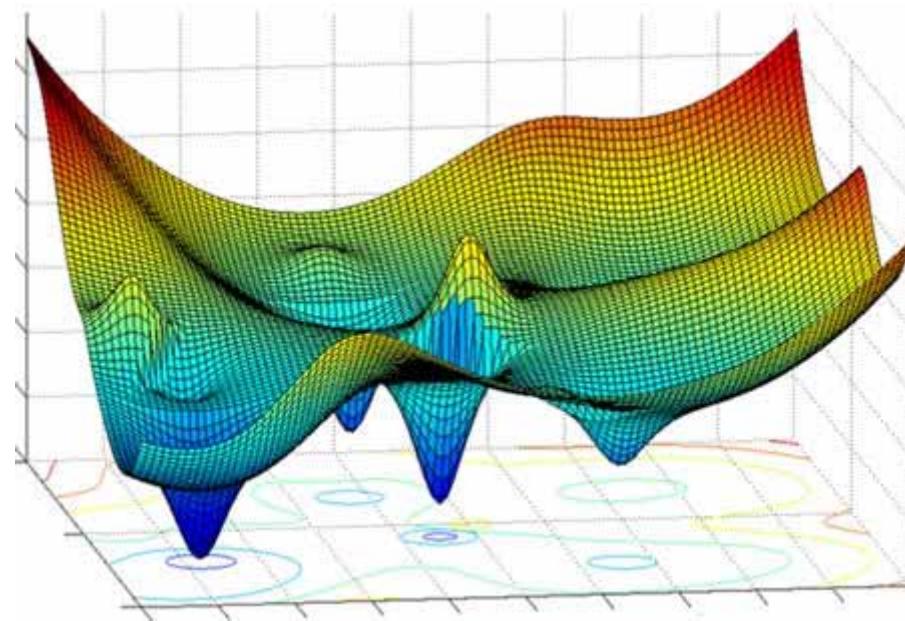
- High dimensionality (curse of dim., many factors contribute)
- Complexity (real-world is non-linear, non-stationary, non-IID \*)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - \*) = Def.: a sequence or collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent

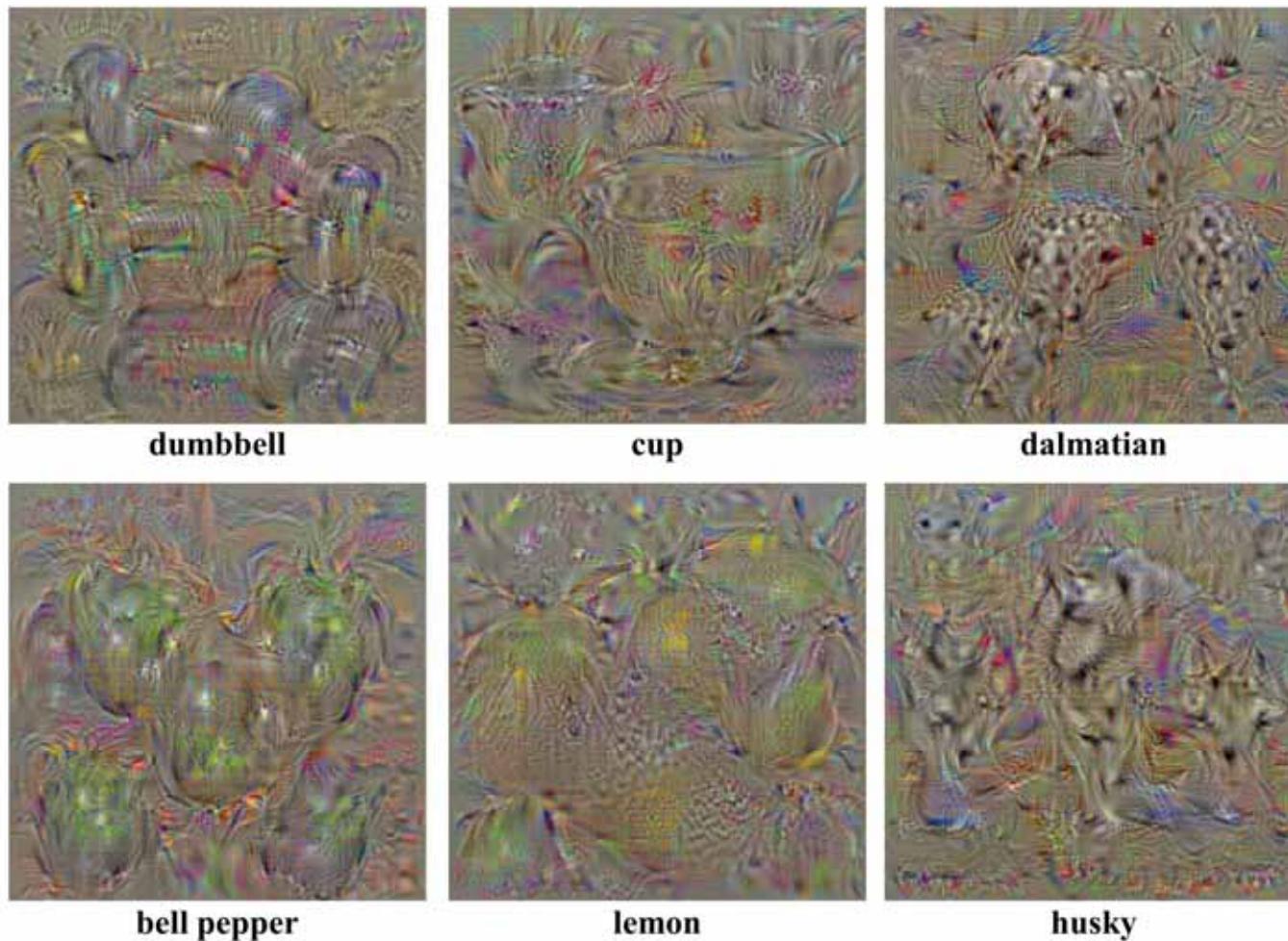


Sören Sonnenburg, Gunnar Rätsch, Christin Schaefer & Bernhard Schölkopf 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

<https://lrpserver.hhi.fraunhofer.de/handwriting-classification>







Karen Simonyan, Andrea Vedaldi & Andrew Zisserman 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

For  $\text{var}_f(x_0) = k(x_0, x_0) - k_*^T(K + \Sigma)^{-1}k_*$  the derivative is given by<sup>3</sup>

$$\nabla \text{var}_f(x)|_{x=x_0} = \frac{\partial \text{var}_f}{\partial x_{0,j}} = \left( \frac{\partial}{\partial x_{0,j}} k(x_0, x_0) \right) - 2 * k_*^T(K + \Sigma)^{-1} \frac{\partial}{\partial x_{0,j}} k_* \quad \text{for } j \in \{1, \dots, d\}.$$

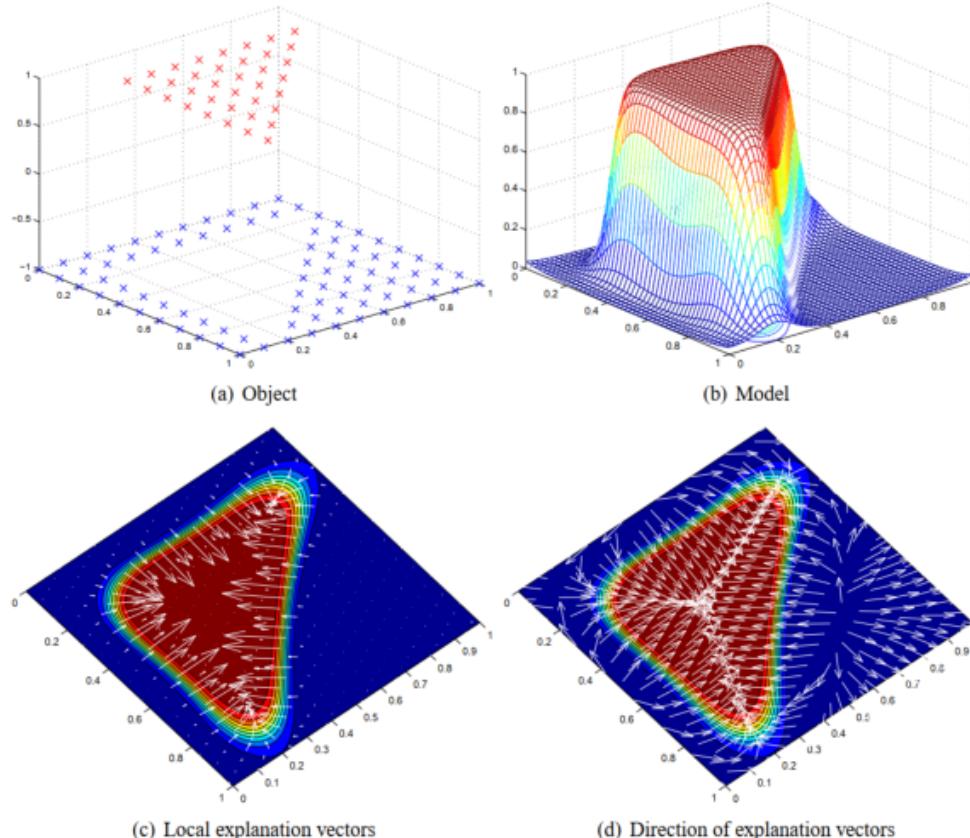
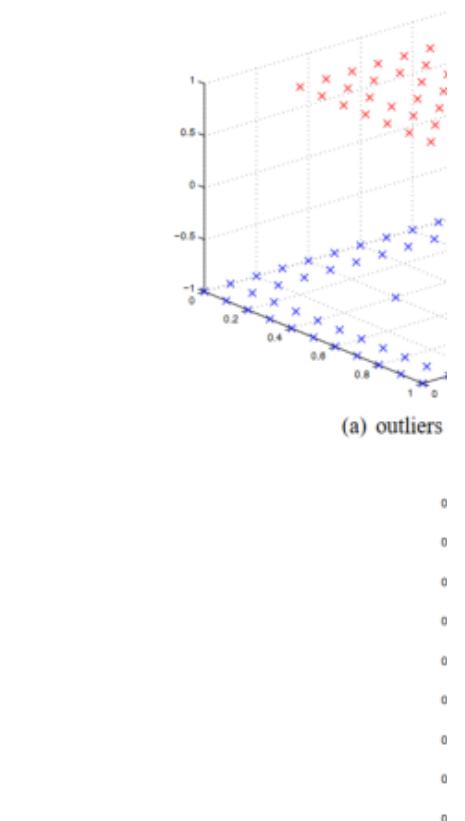


Figure 1: Explaining simple object classification with Gaussian Processes

Panel (a) of Figure 1 shows the training data of a simple object classification task and panel (b) shows the model learned using GPC.<sup>4</sup> The data is labeled  $-1$  for the blue points and  $+1$  for the red points. As illustrated in panel (b) the model is a probability function for the positive class which gives every data point a probability of being in this class. Panel (c) shows the probability gradient of the model together with the local gradient explanation vectors. Along the hypotenuse and at the corners of the triangle explanations from both features interact towards the triangle class while along

David Baehrens, Timor  
How to explain individual

Holzinger Group hci-kdd.org



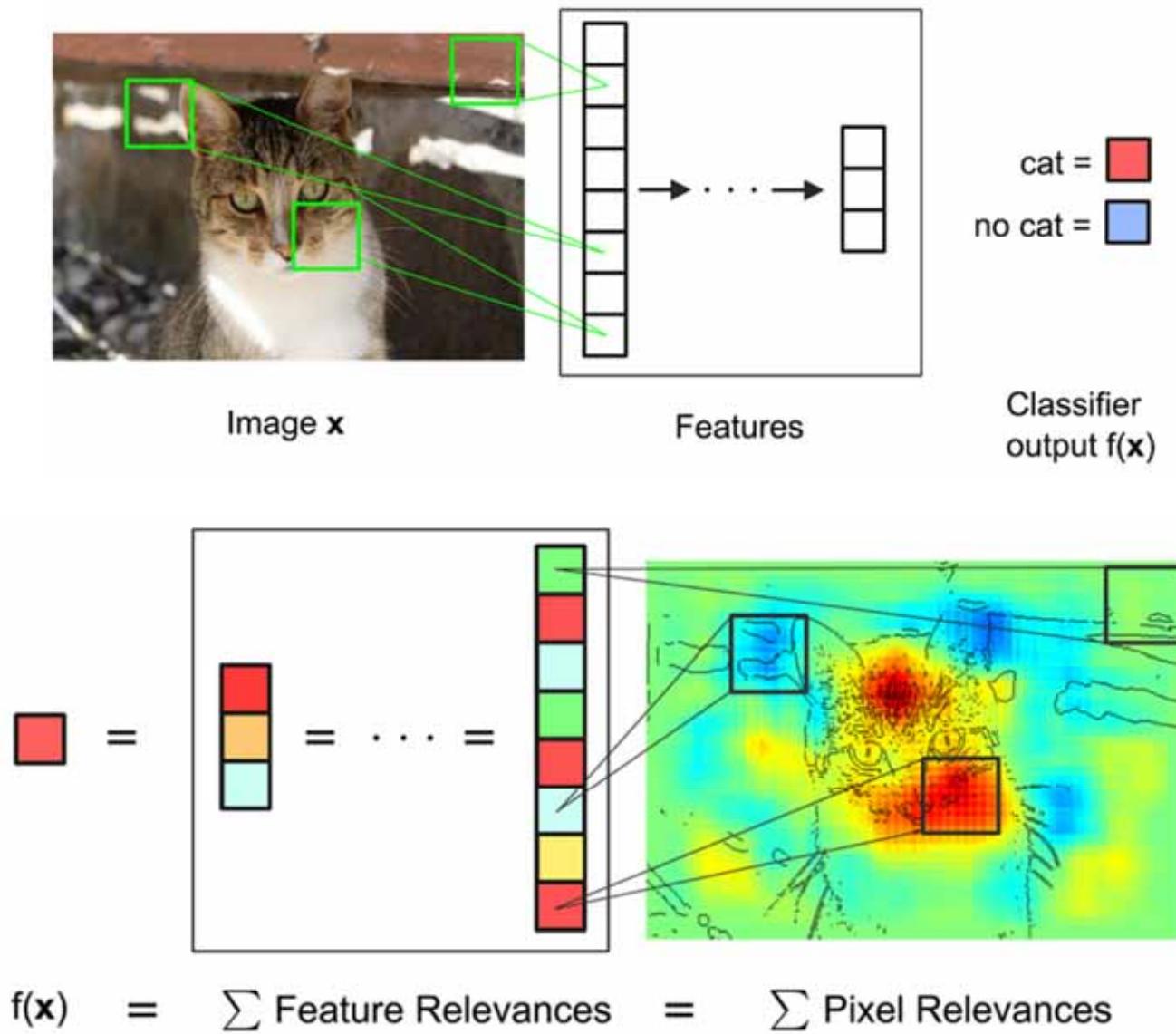
(a) outliers

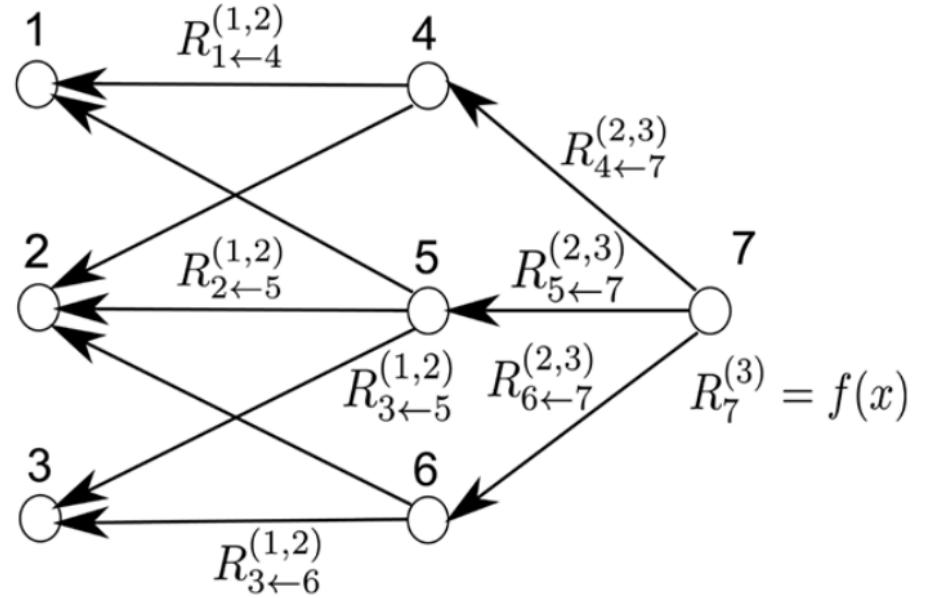
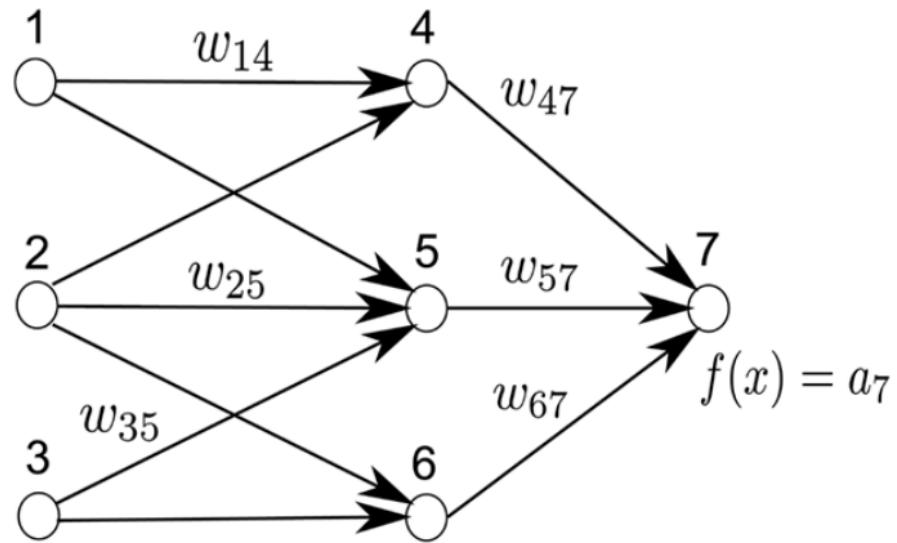
Figure 12

Robert Mueller 2010.  
11, (6), 1803-1831.

046 AK HCI Summer 2019 L01

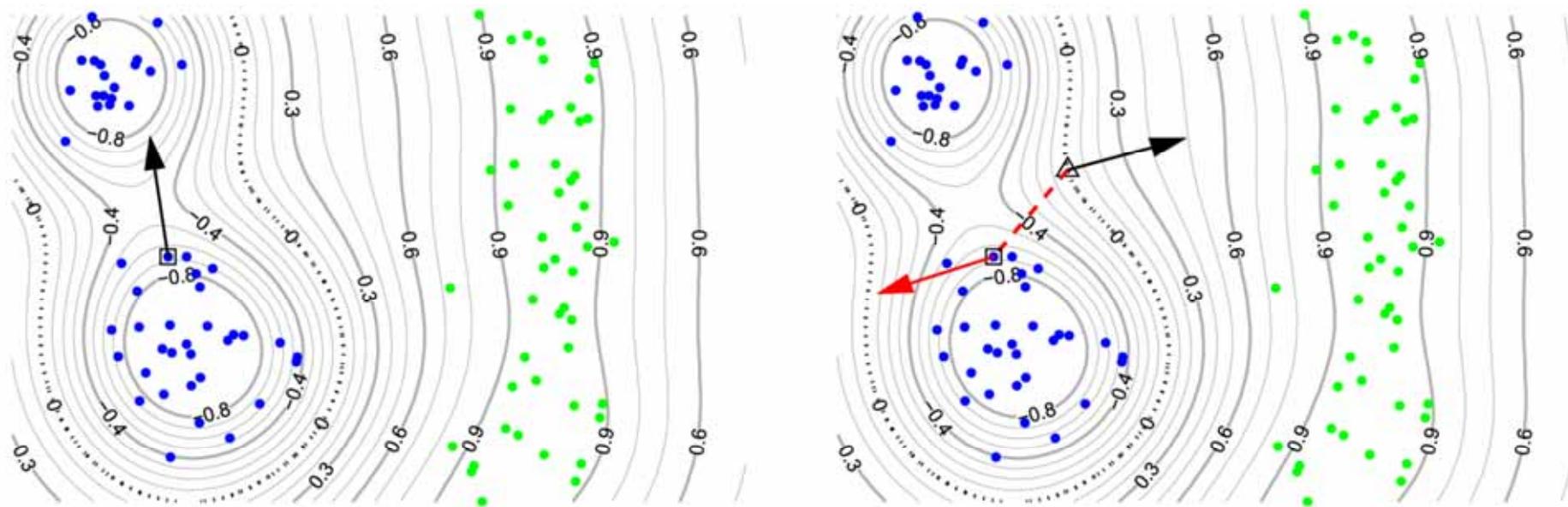
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.





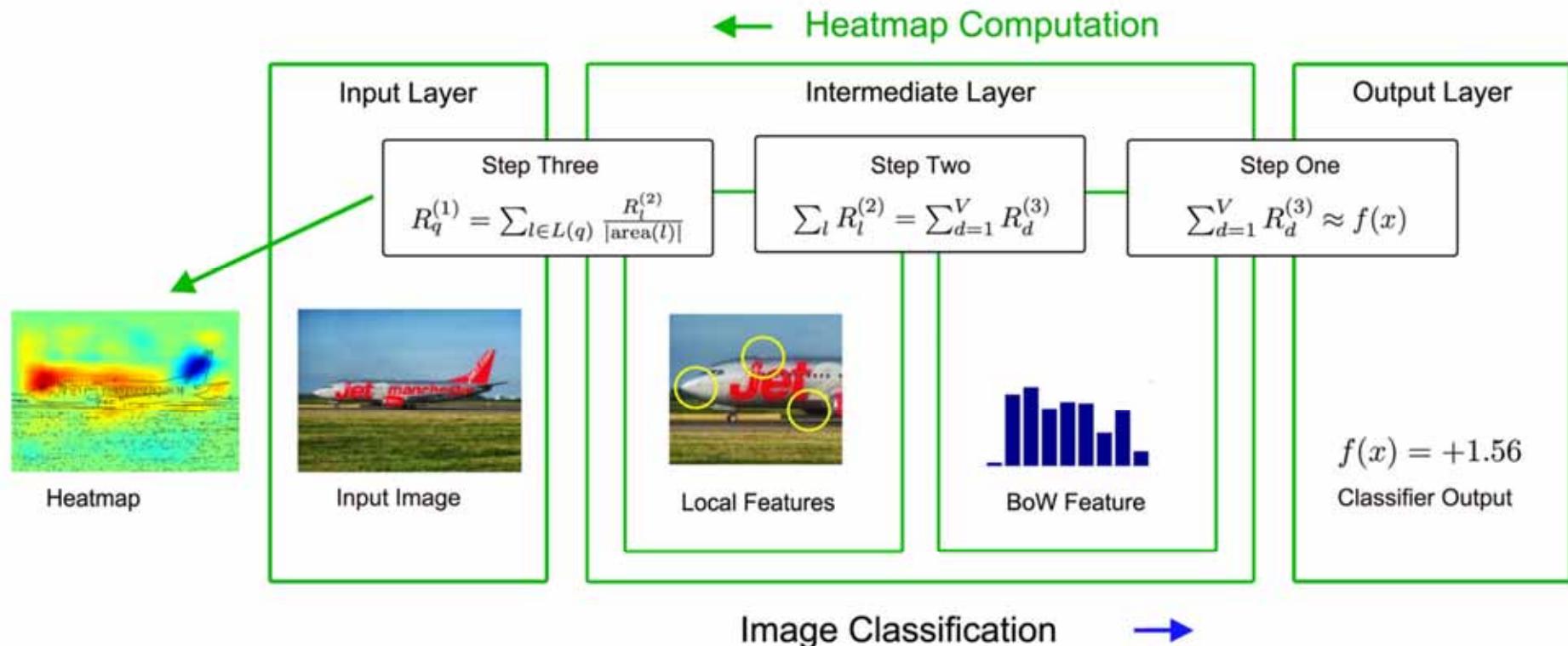
$$f(x) = \dots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \dots = \sum_d R_d^{(1)}$$

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



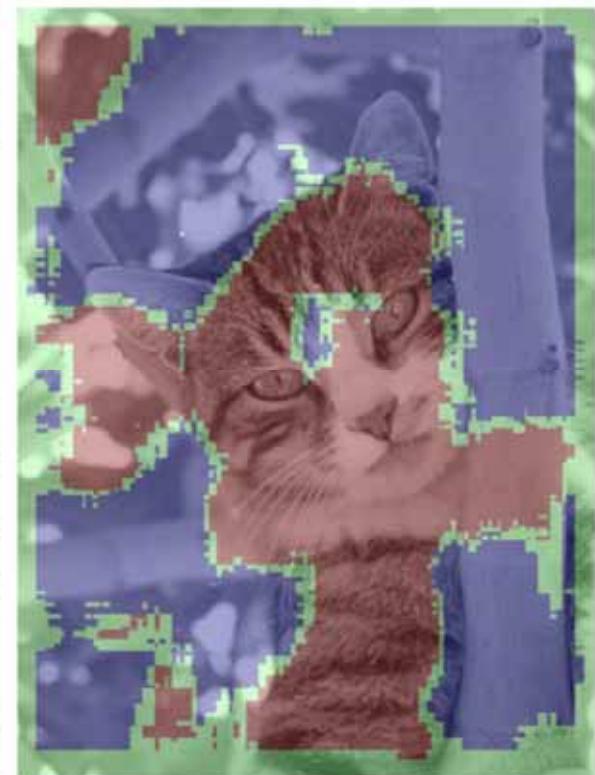
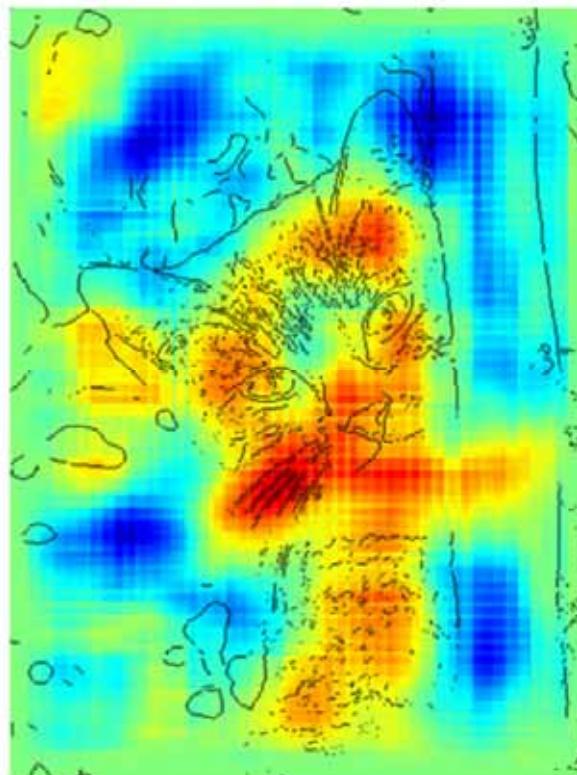
**Fig 3. An exemplary real-valued prediction function for classification with the dashed black line being the decision boundary which separates the blue from the green dots.** The blue dots are labeled negatively, the green dots are labeled positively. Left: Local gradient of the classification function at the prediction point. Right: Taylor approximation relative to a root point on the decision boundary. This figure depicts the intuition that a gradient at a prediction point  $x$ —here indicated by a square—does not necessarily point to a close point on the decision boundary. Instead it may point to a local optimum or to a far away point on the decision boundary. In this example the explanation vector from the local gradient at the prediction point  $x$  has a too large contribution in an irrelevant direction. The closest neighbors of the other class can be found at a very different angle. Thus, the local gradient at the prediction point  $x$  may not be a good explanation for the contributions of single dimensions to the function value  $f(x)$ . Local gradients at the prediction point in the left image and the Taylor root point in the right image are indicated by black arrows. The nearest root point  $x_0$  is shown as a triangle on the decision boundary. The red arrow in the right image visualizes the approximation of  $f(x)$  by Taylor expansion around the nearest root point  $x_0$ . The approximation is given as a vector representing the dimension-wise product between  $Df(x_0)$  (the black arrow in the right panel) and  $x - x_0$  (the dashed red line in the right panel) which is equivalent to the diagonal of the outer product between  $Df(x_0)$  and  $x - x_0$ .

doi:10.1371/journal.pone.0130140.g003



**Fig 4. Local and global predictions for input images are obtained by following a series of steps through the classification- and pixel-wise decomposition pipelines.** Each step taken towards the final pixel-wise decomposition has a complementing analogue within the Bag of Words classification pipeline. The calculations used during the pixel-wise decomposition process make use of information extracted by those corresponding analogues. Airplane image in the graphic by Pixabay user tpsdave.

doi:10.1371/journal.pone.0130140.g004



**Definition 1.** A heatmap  $R(x)$  is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

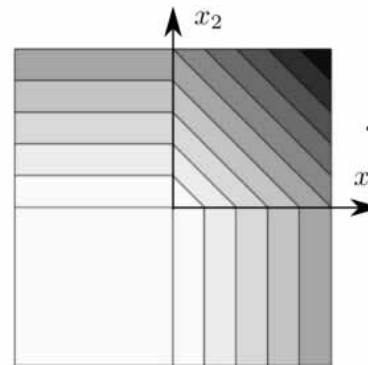
$$\forall x: f(x) = \sum_p R_p(x).$$

**Definition 2.** A heatmap  $R(x)$  is *positive* if all values forming the heatmap are greater or equal to zero, that is:

$$\forall x, p: R_p(x) \geq 0$$

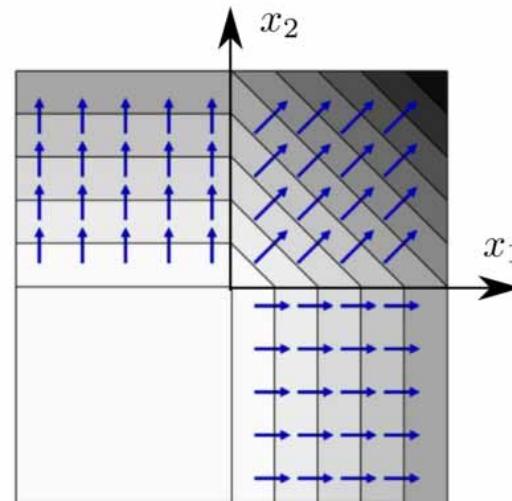
**Definition 3.** A heatmap  $R(x)$  is *consistent* if it is conservative *and* positive. That is, it is consistent if it complies with [Definitions 1 and 2](#).

Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.



function to analyze:

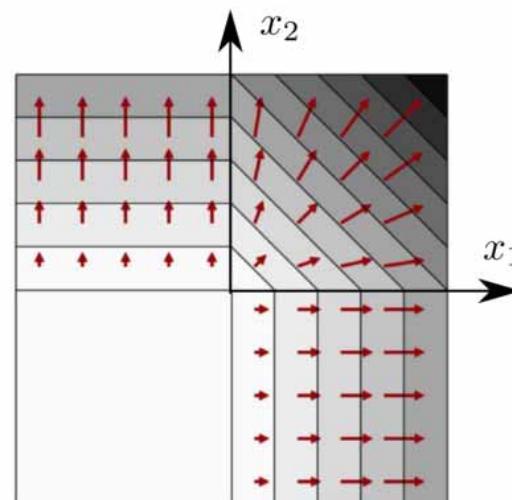
$$f(\mathbf{x}) = \max(0, x_1) + \max(0, x_2)$$



sensitivity analysis:

$$(\partial f / \partial x_1)^2 = 1_{x_1 > 0}$$

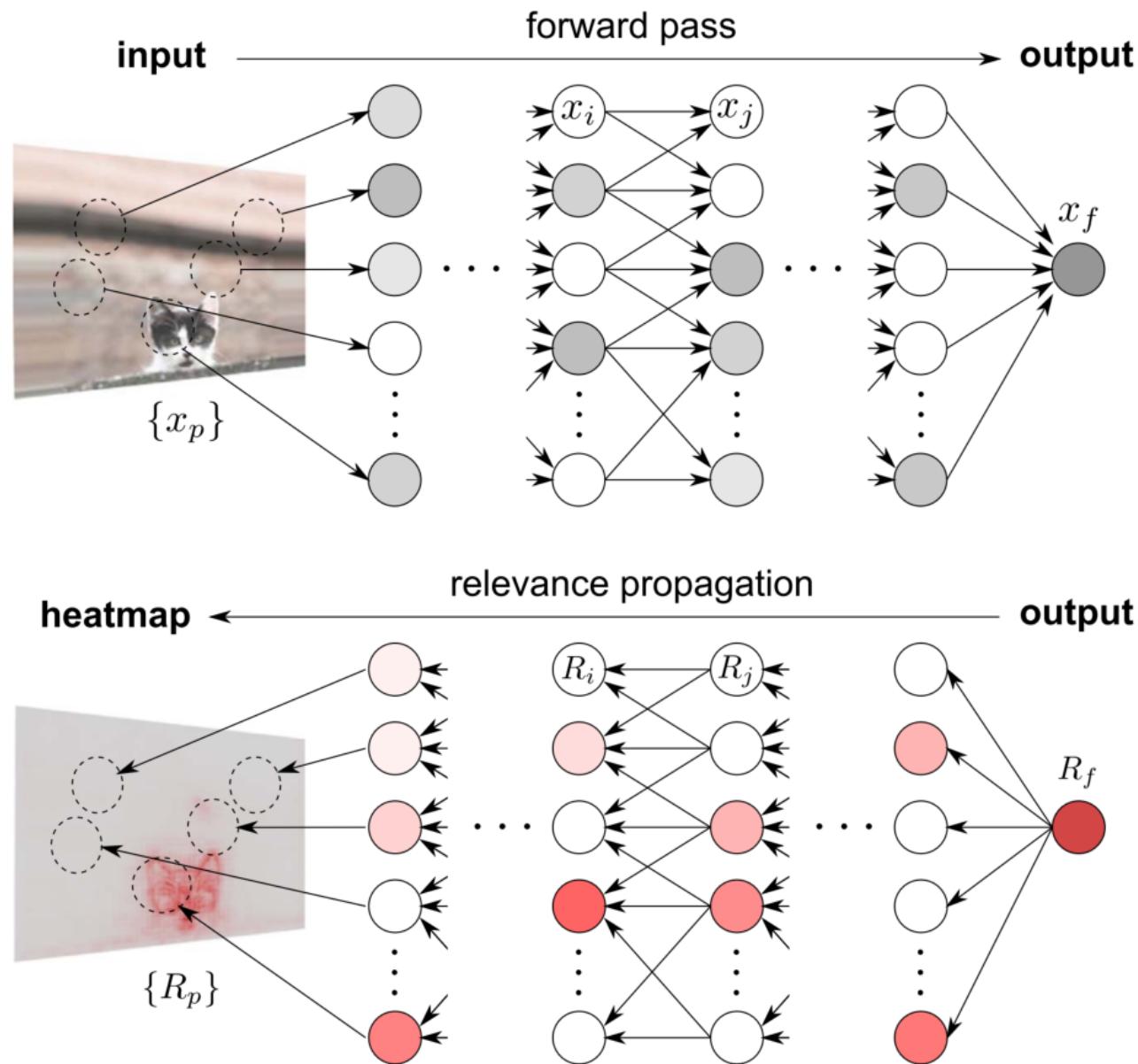
$$(\partial f / \partial x_2)^2 = 1_{x_2 > 0}$$

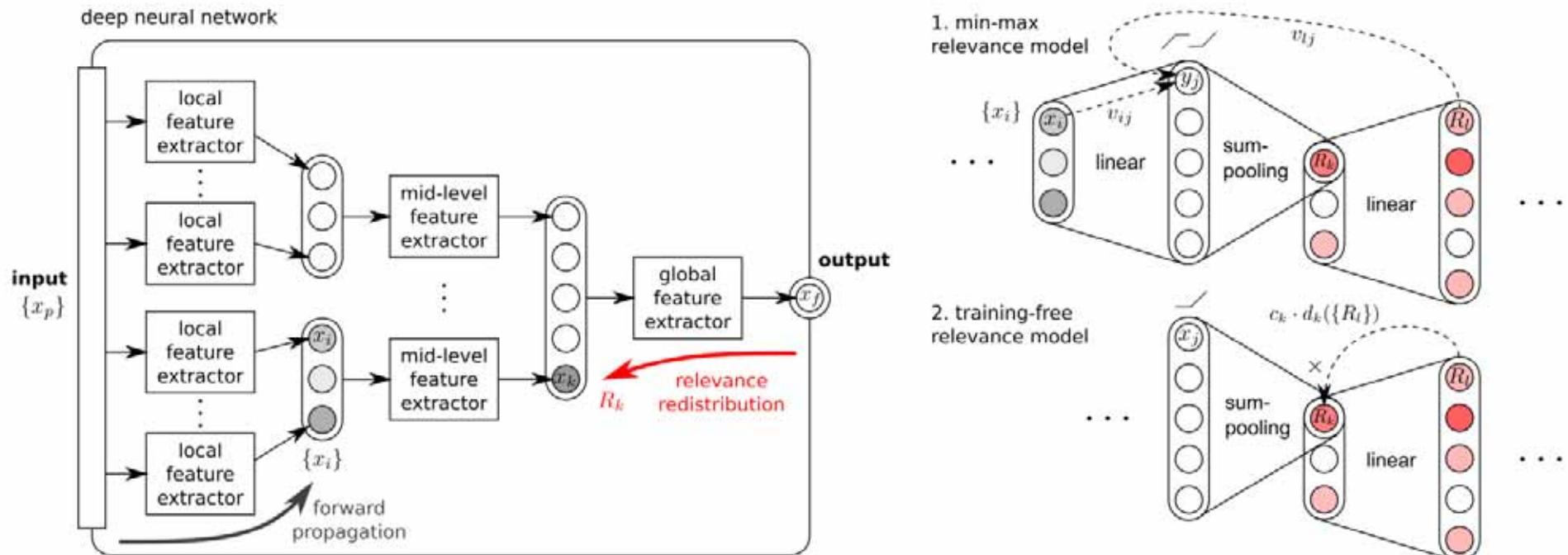


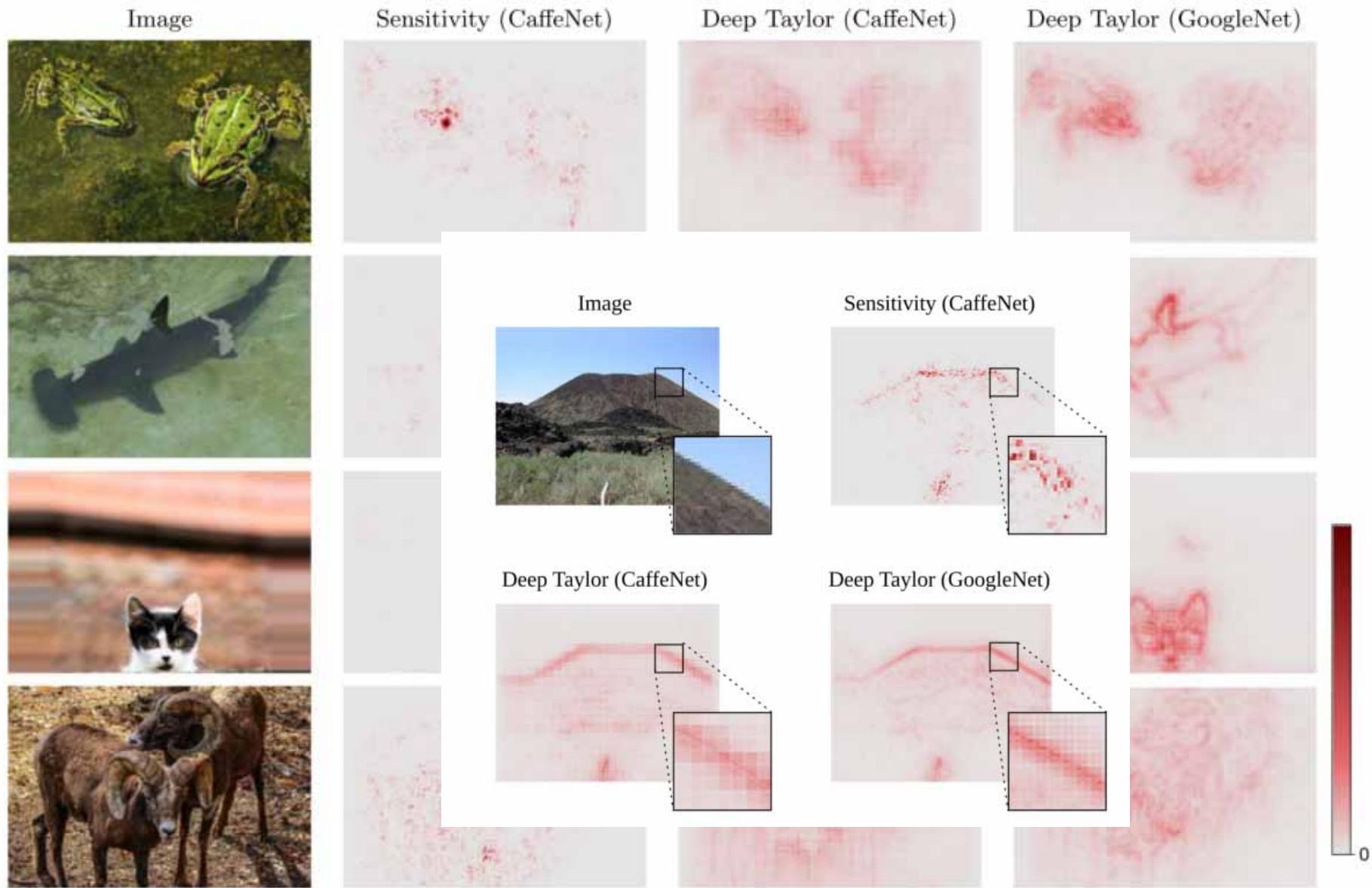
decomposition:

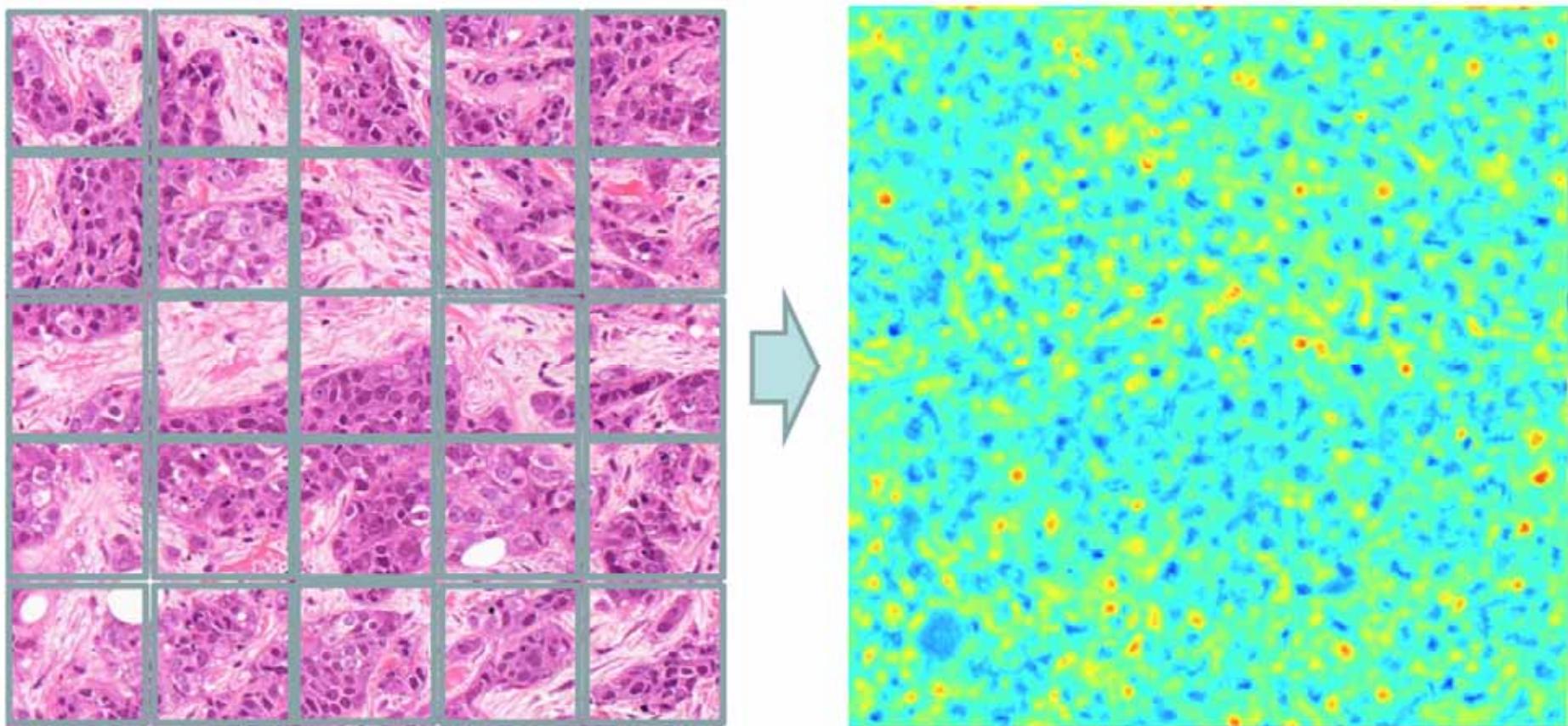
$$R_1(\mathbf{x}) = \max(0, x_1)$$

$$R_2(\mathbf{x}) = \max(0, x_2)$$

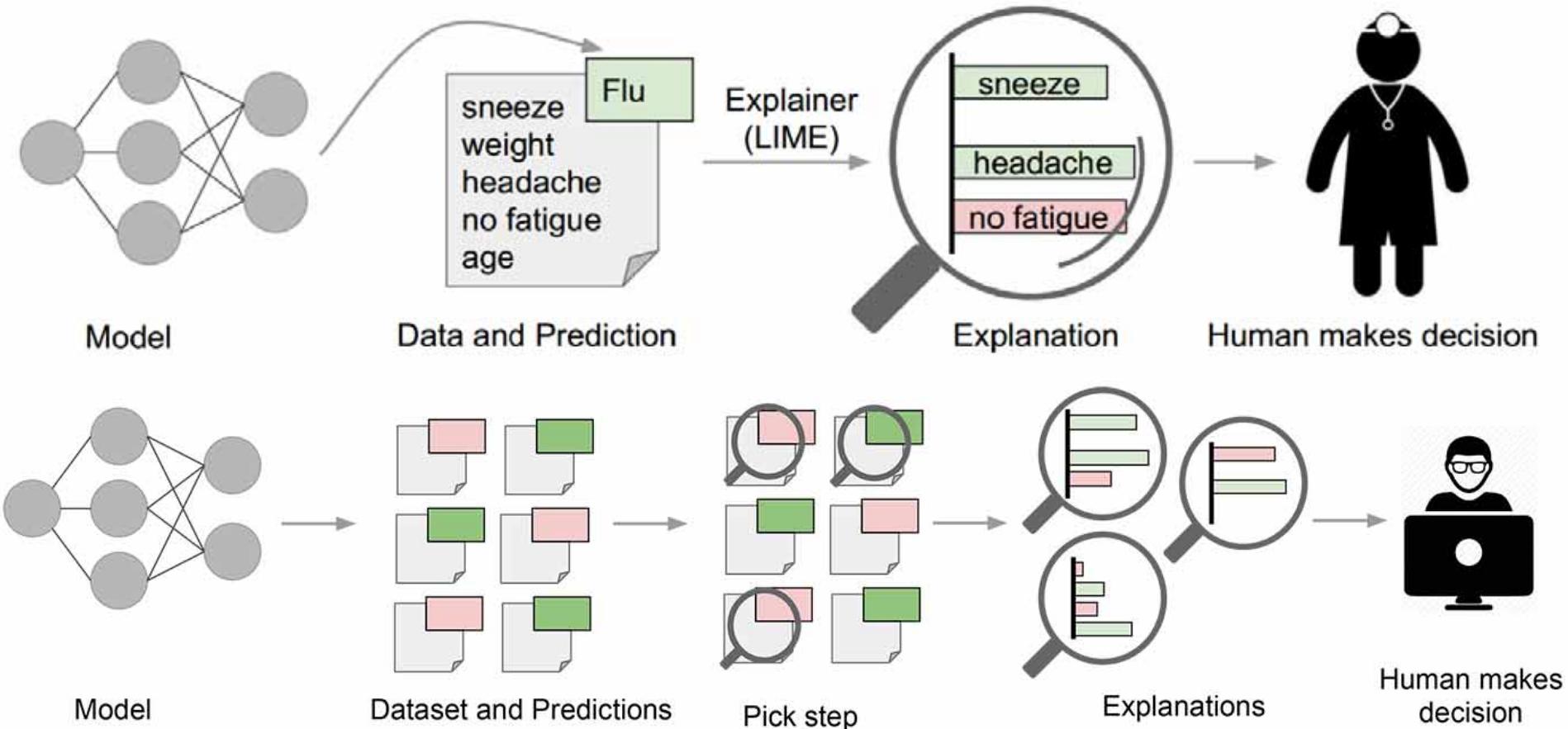








Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek  
2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one,  
10, (7), e0130140, doi:10.1371/journal.pone.0130140.



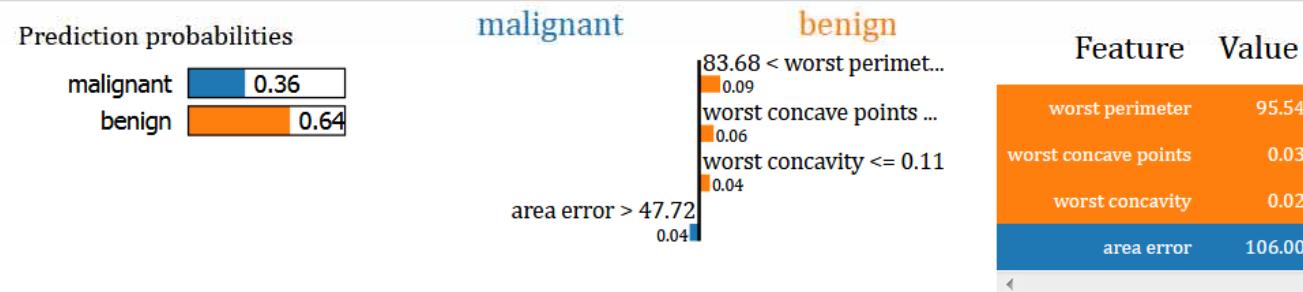
Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

```
In [12]: explainer = lime.lime_tabular.LimeTabularExplainer(X_train, feature_names=breast.feature_names, class_names=breast.targe
```

Here we will take a sample from the test set (in this case the sample at index 76) and create an explainer instance for this sample. This will let us see why the algorithm made its prediction visually.

```
In [18]: # For this demonstration, let's take the same sample each time, in this case sample index 86
i = 76
# For a random sample uncomment out the following line
# i = np.random.randint(0, X_test.shape[0])

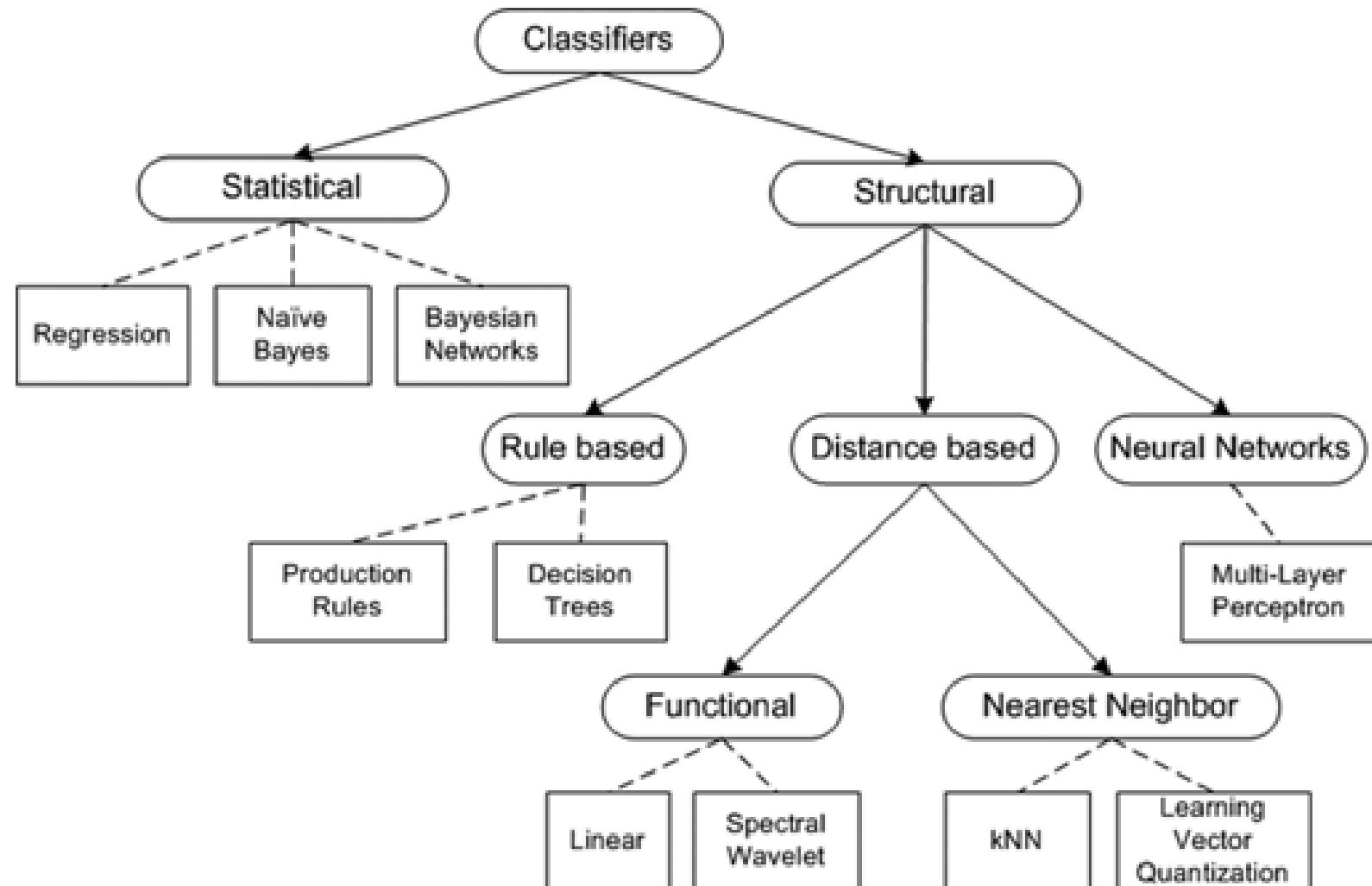
exp = explainer.explain_instance(X_test[i], random_forest.predict_proba, num_features=4)
exp.show_in_notebook(show_table=True, show_all=False)
```



As you can see, the random forest algorithm has predicted with a probability of 0.64 that the sample at index 76 in the test set is malignant.

When using the explainer, we set the `num_features` parameter to 4, meaning the explainer shows the top 4 features that contributed to the prediction probabilities.

We chose 76 as it was a borderline decision. For example sample 86 is much more clear (this will we will set the `num_features` parameter to include all features so that we see each feature's contribution to the probability):



<https://stats.stackexchange.com/questions/271247/machine-learning-statistical-vs-structural-classifiers>

If Age <50 and Male =Yes:

If Past-Depression =Yes and Insomnia =No and Melancholy =No, then Healthy

If Past-Depression =Yes and Insomnia =Yes and Melancholy =Yes and Tiredness =Yes, then Depression

If Age  $\geq$  50 and Male =No:

If Family-Depression =Yes and Insomnia =No and Melancholy =Yes and Tiredness =Yes, then Depression

If Family-Depression =No and Insomnia =No and Melancholy =No and Tiredness =No, then Healthy

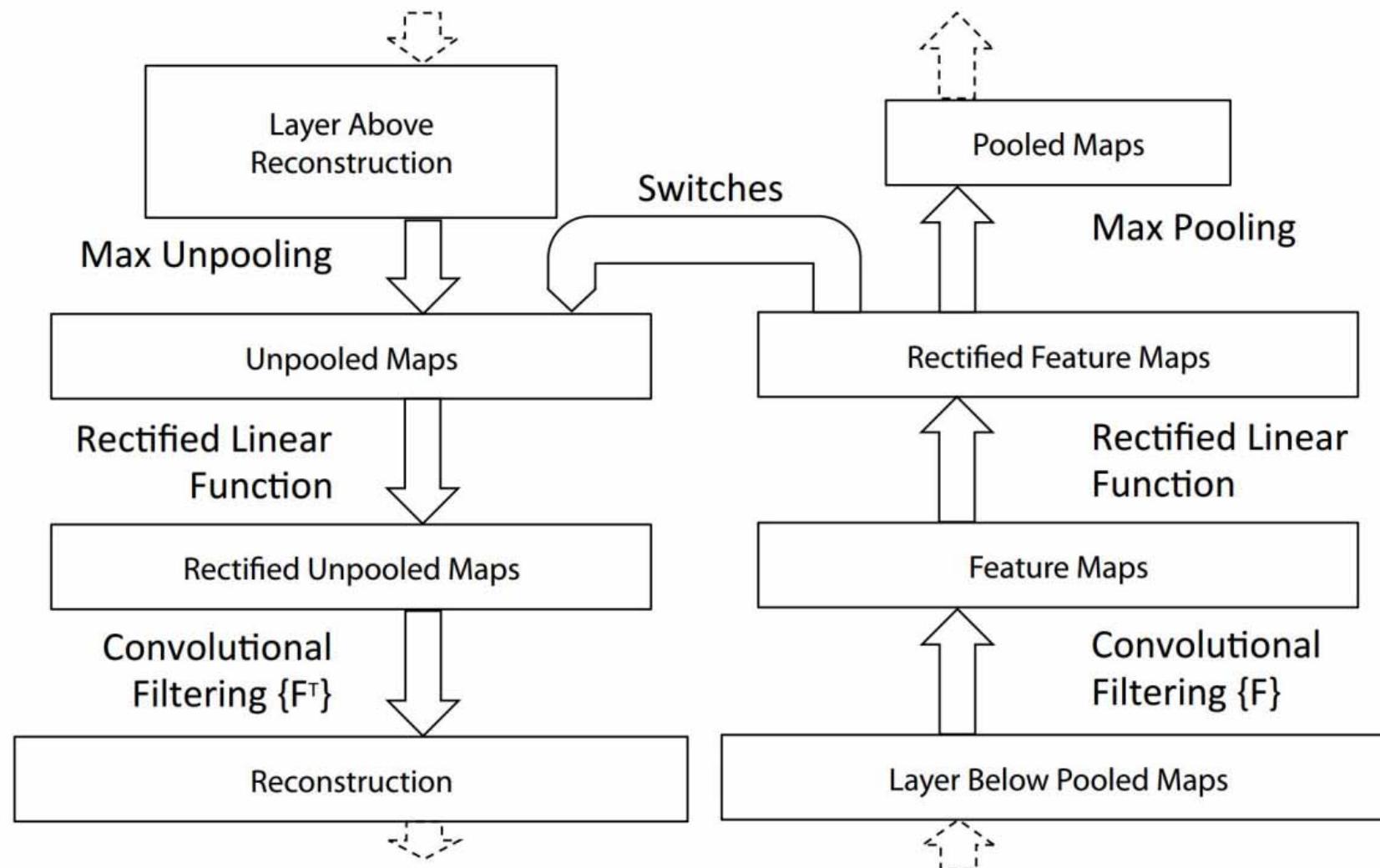
Default:

If Past-Depression =Yes and Tiredness =No and Exercise =No and Insomnia =Yes, then Depression

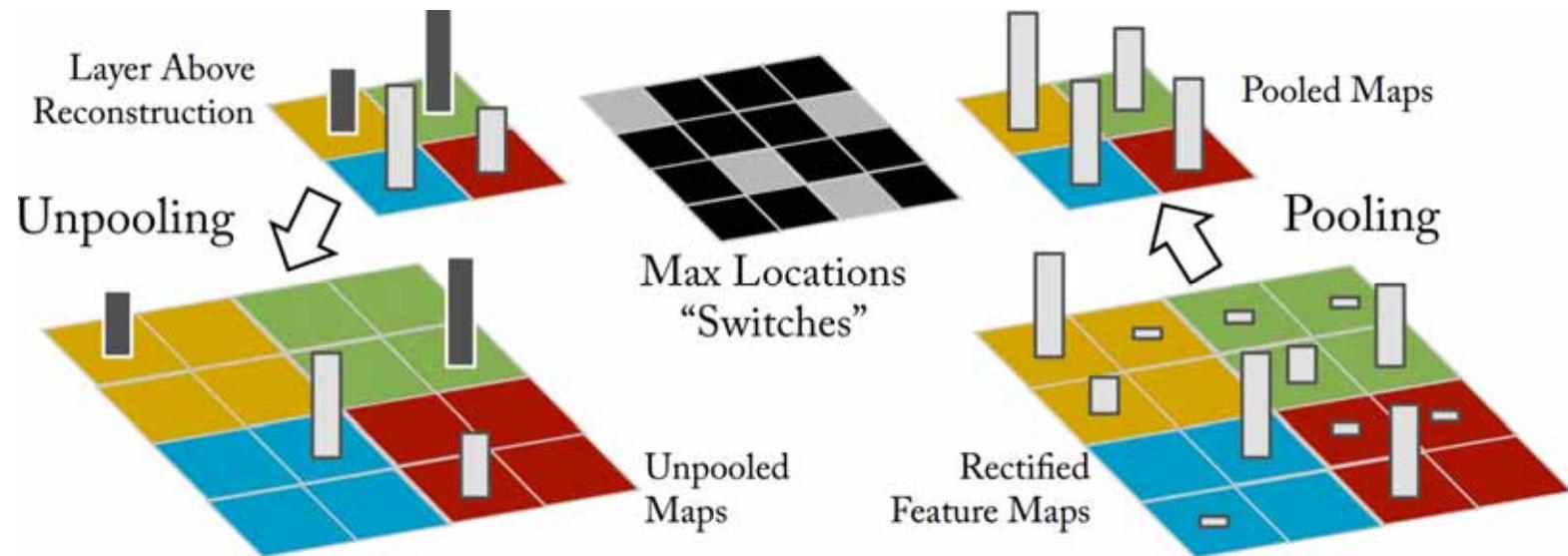
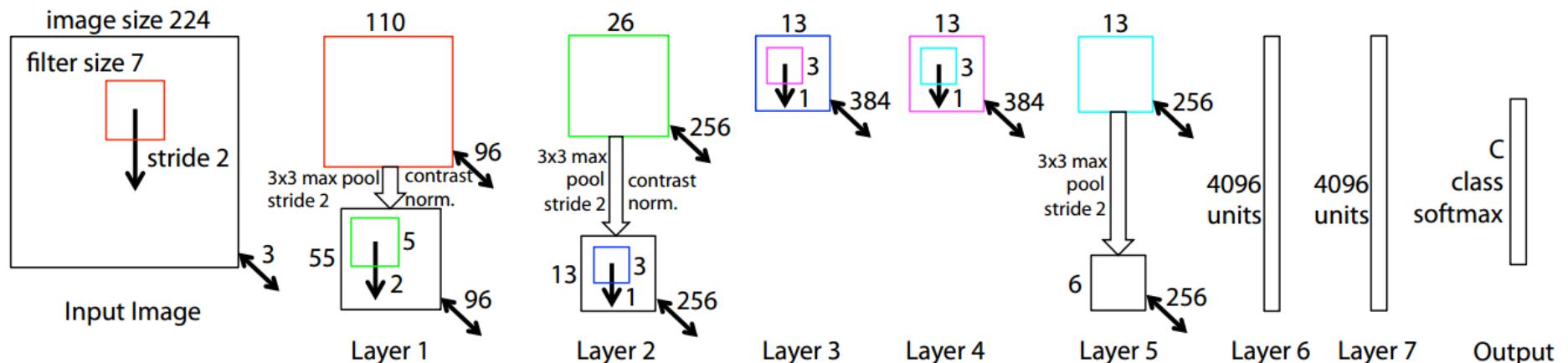
If Past-Depression =No and Weight-Gain =Yes and Tiredness =Yes and Melancholy =Yes, then Depression

If Family-Depression =Yes and Insomnia =Yes and Melancholy =Yes and Tiredness =Yes, then Depression

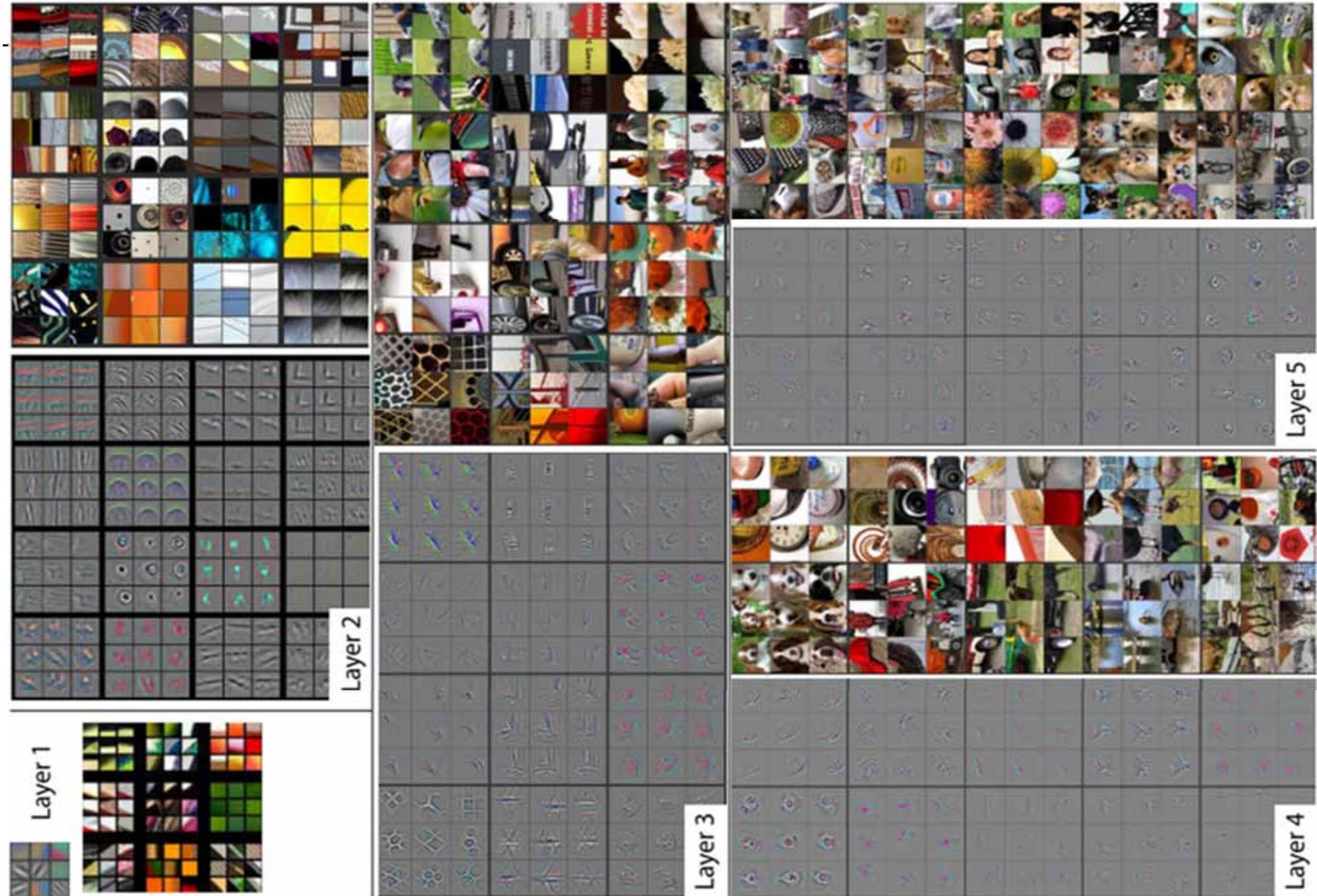
Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.



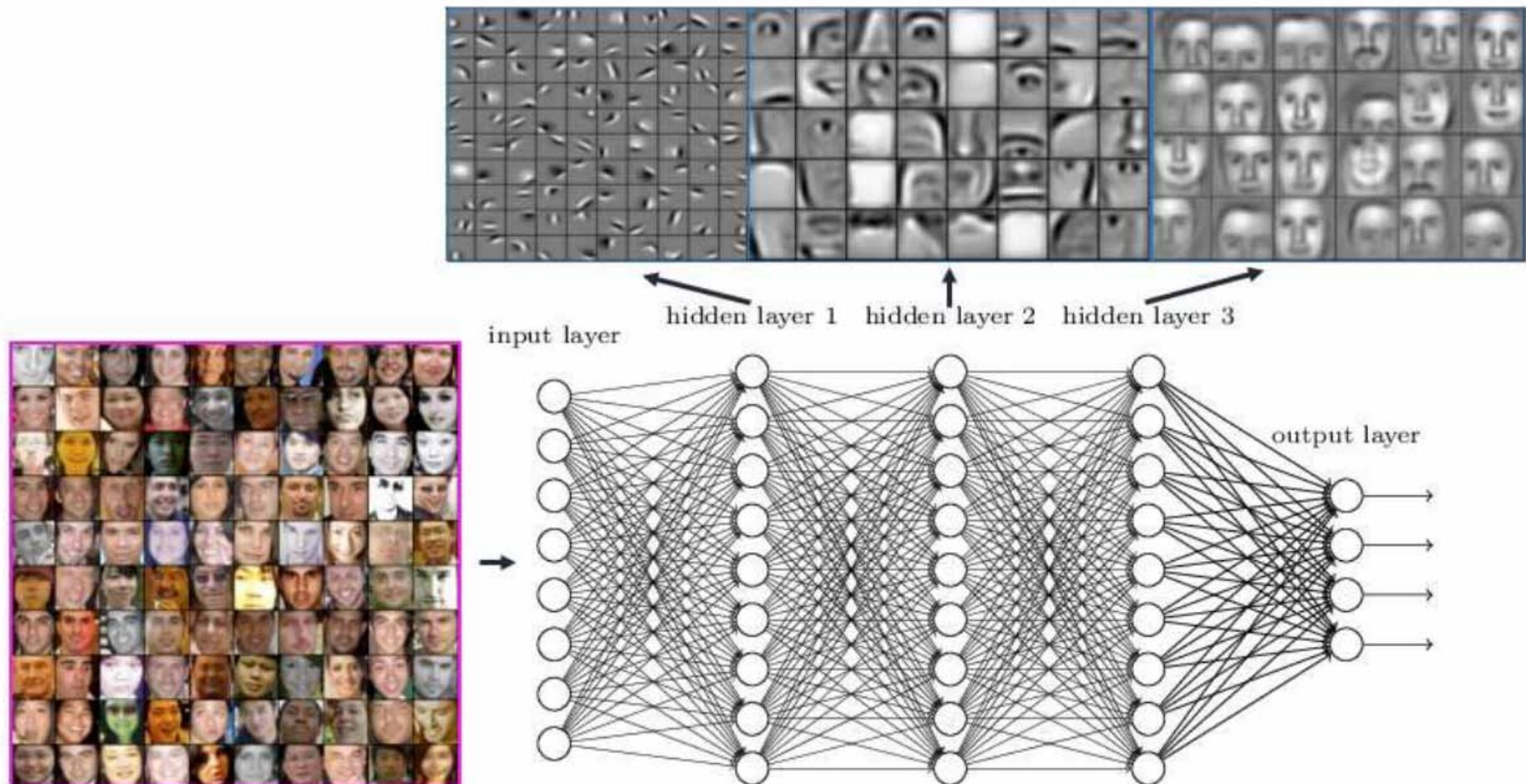
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.



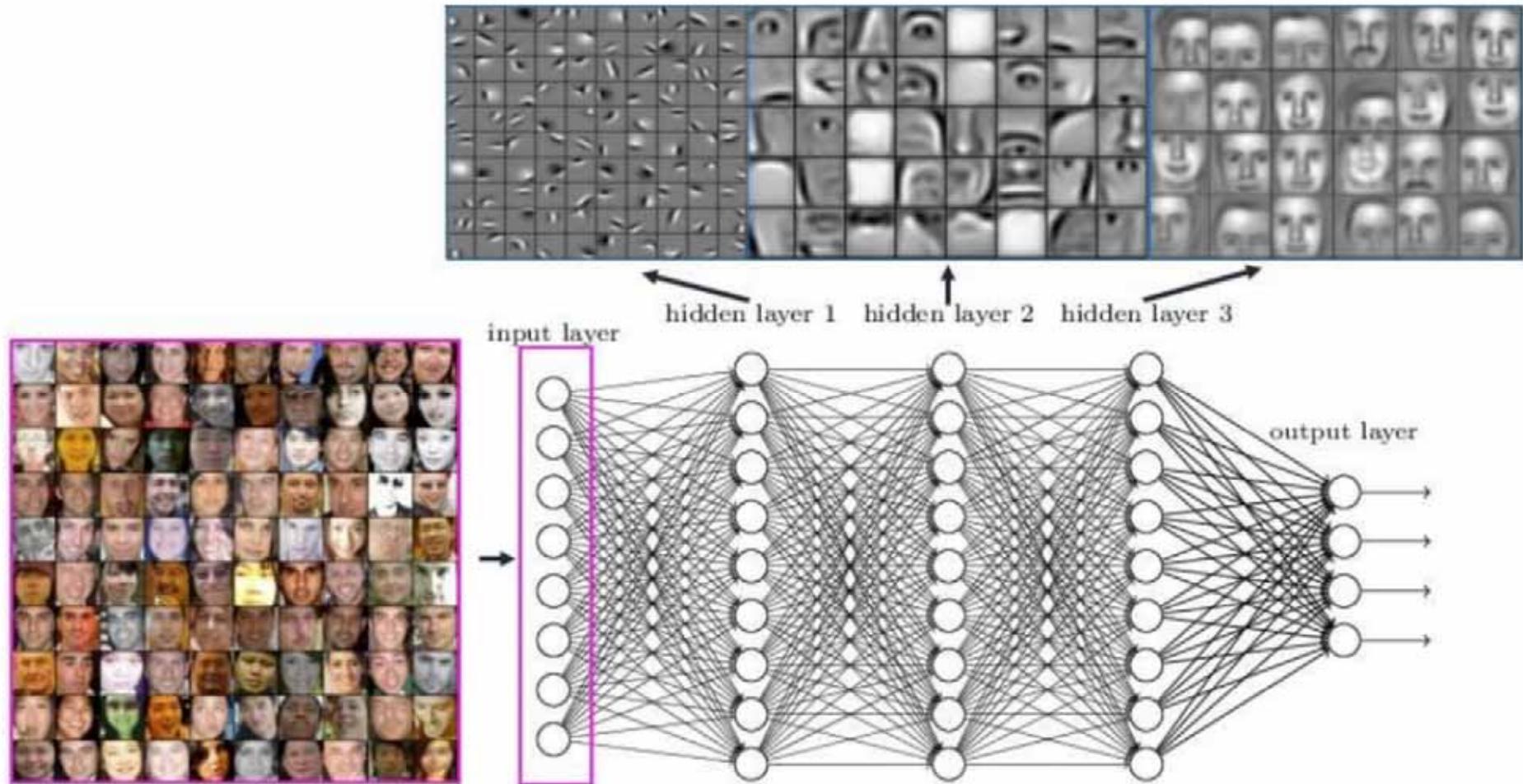
Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1\_53.

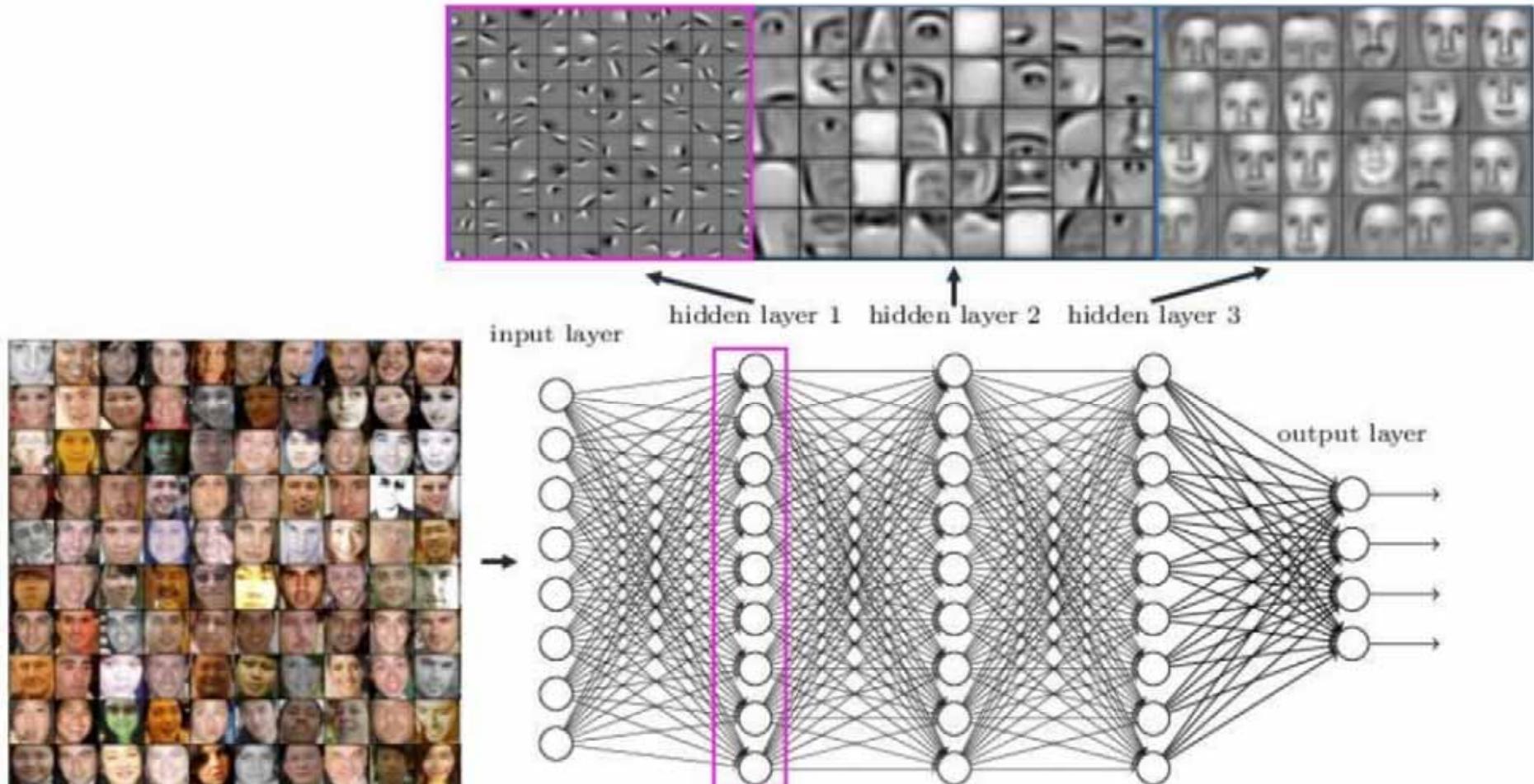


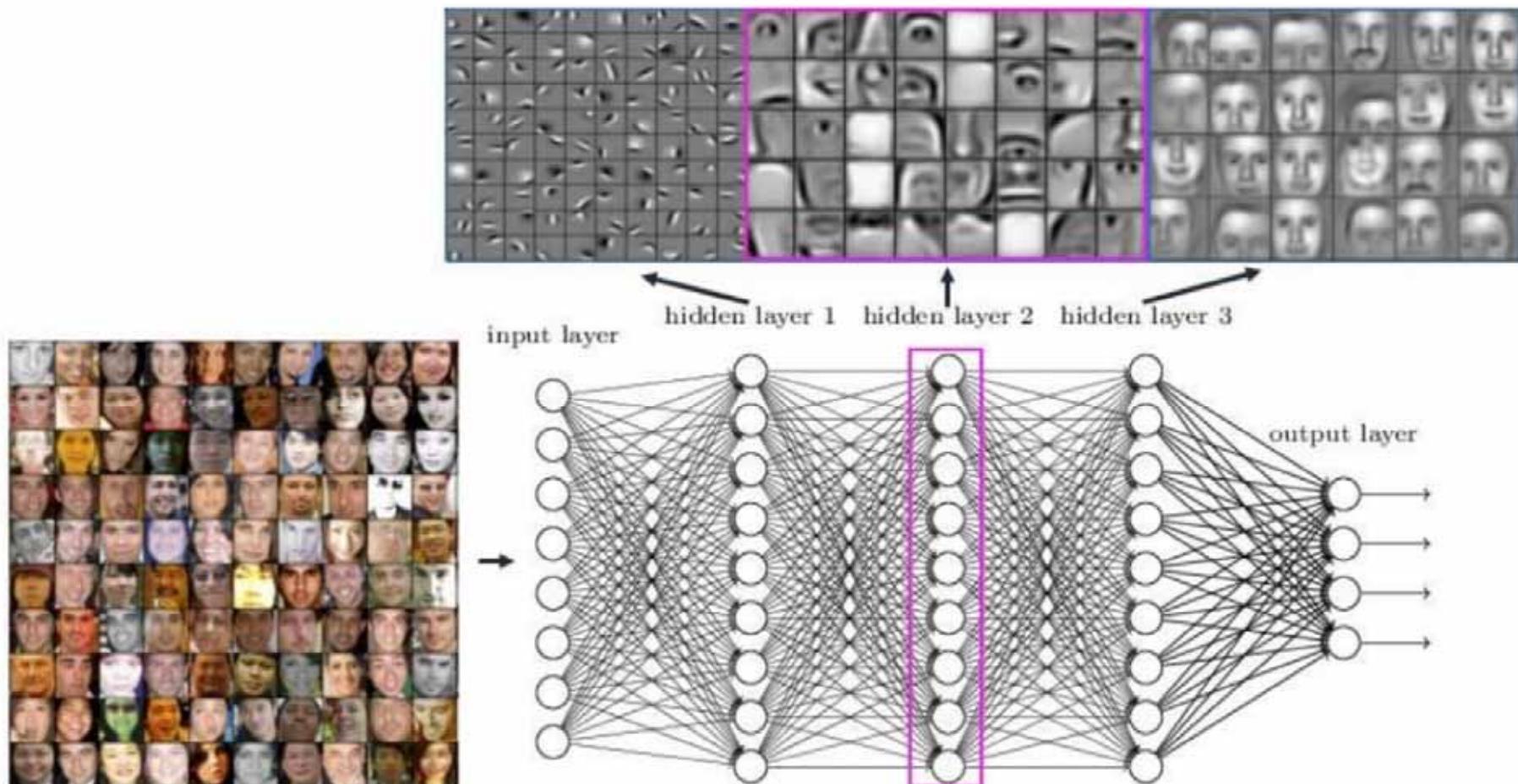
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.  
Holzinger Group hci-kdd.org

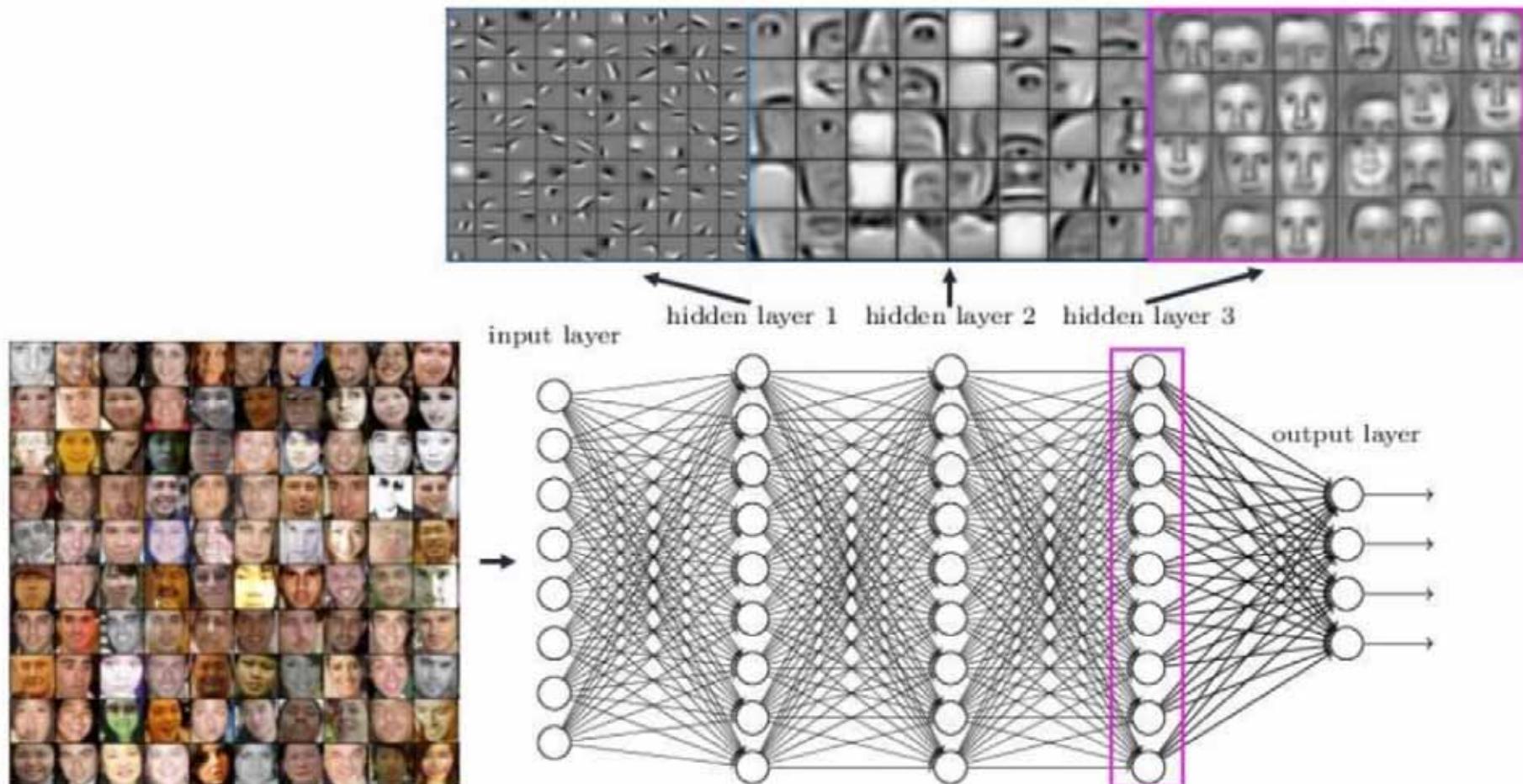


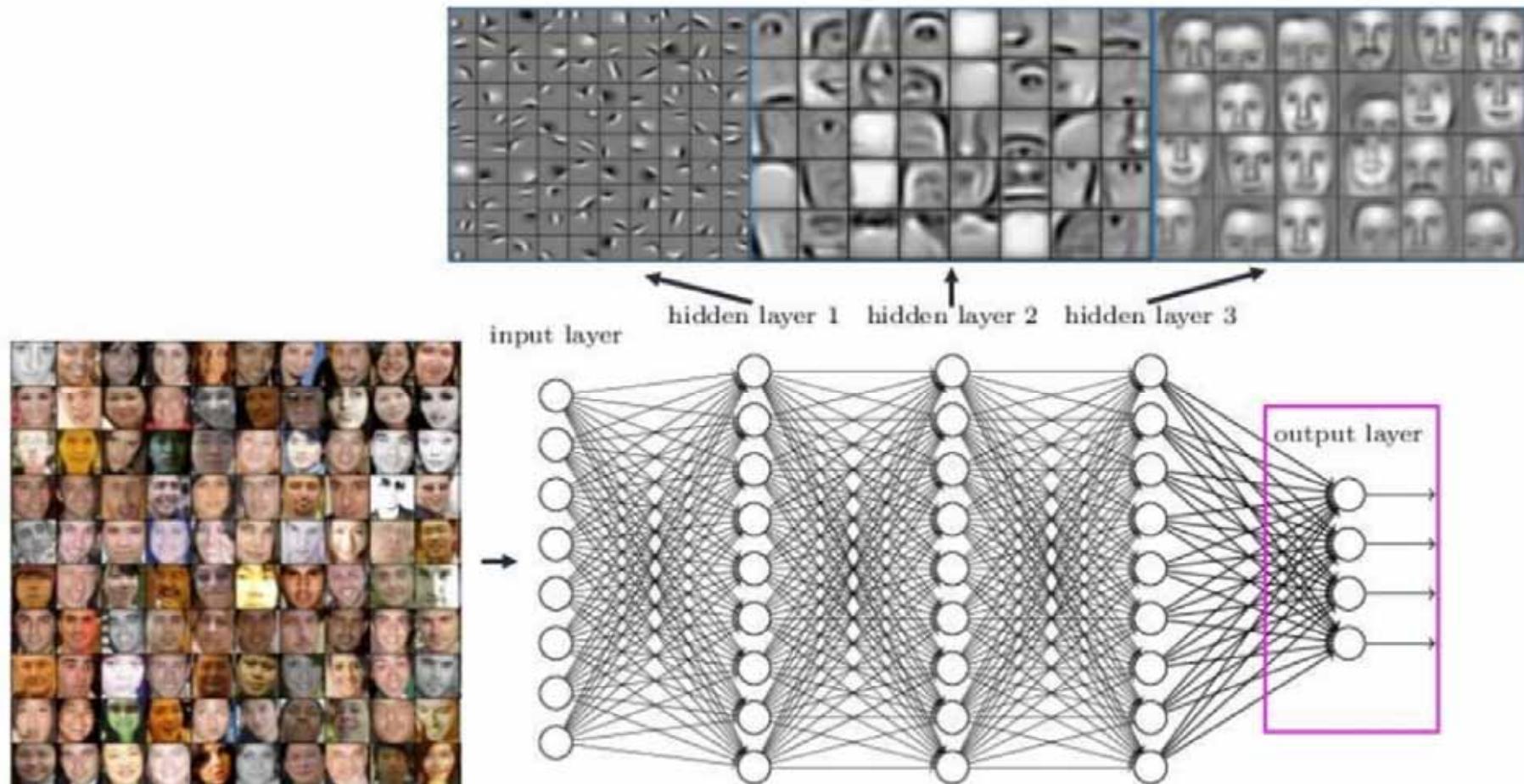
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

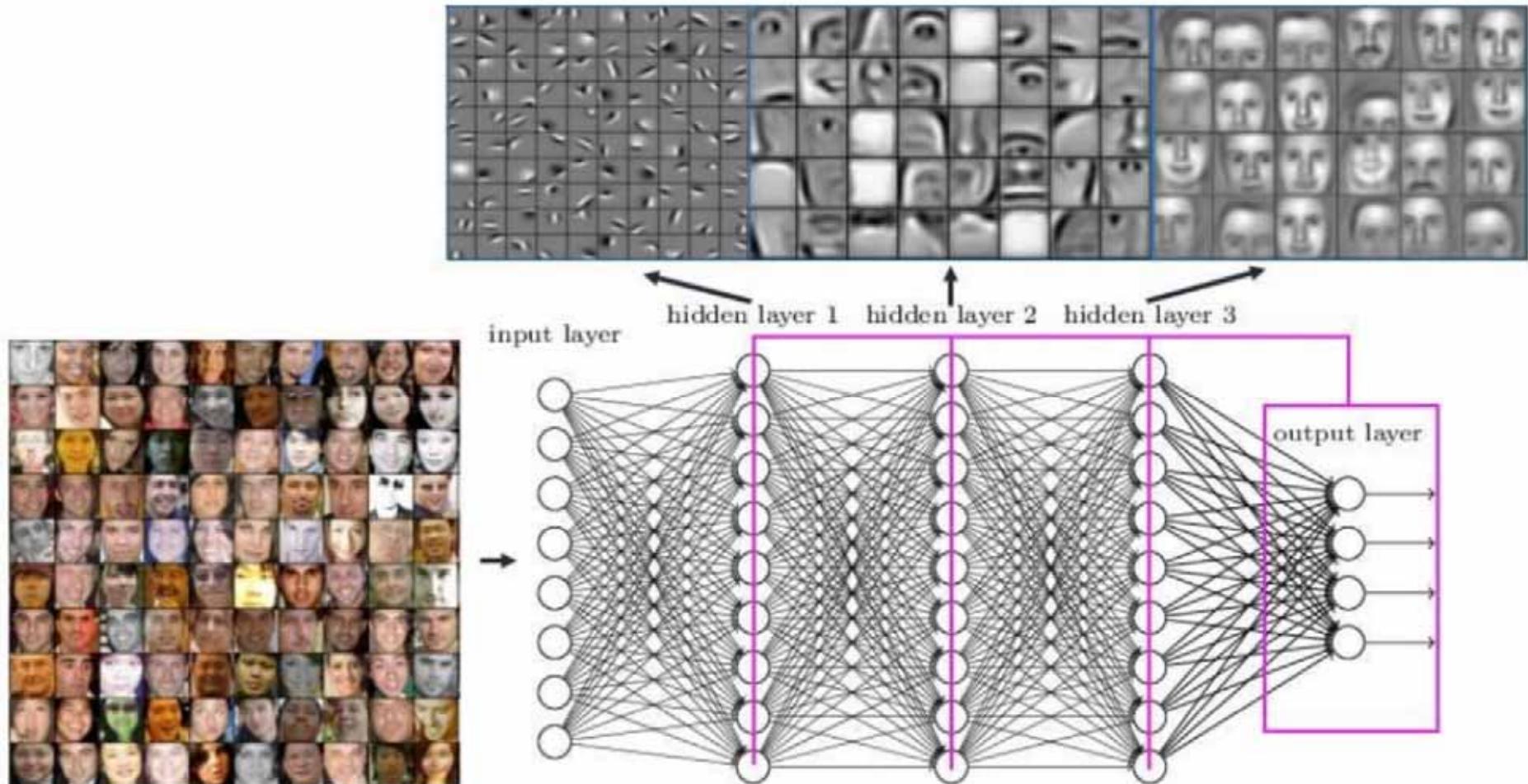




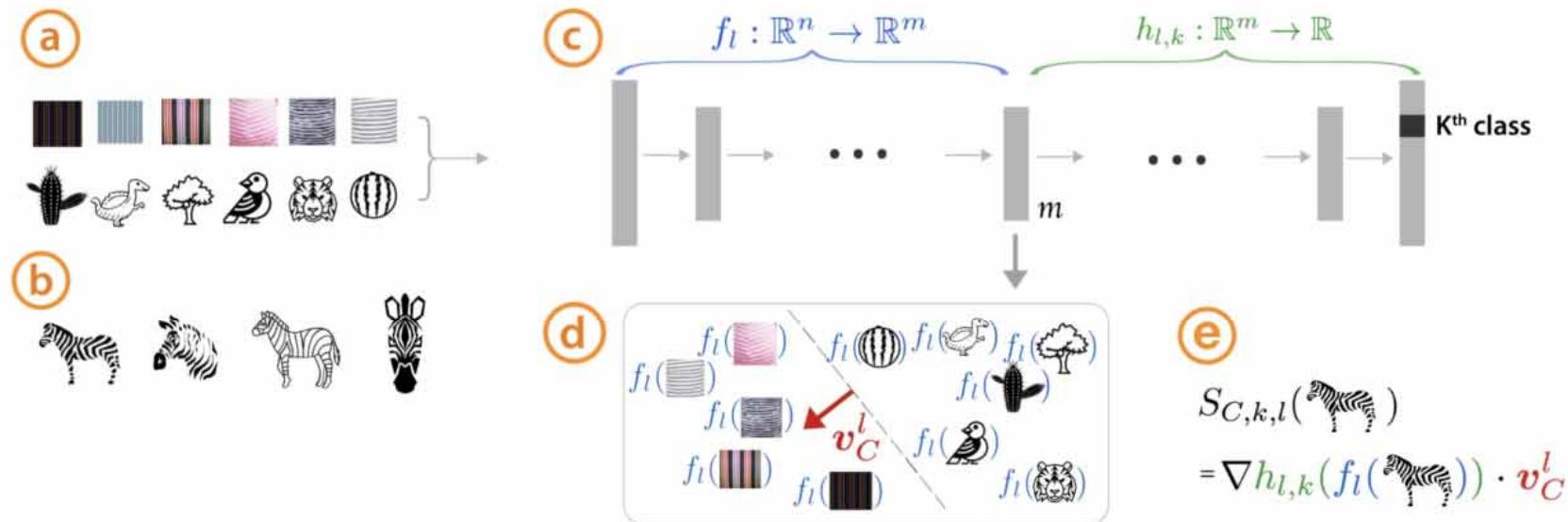








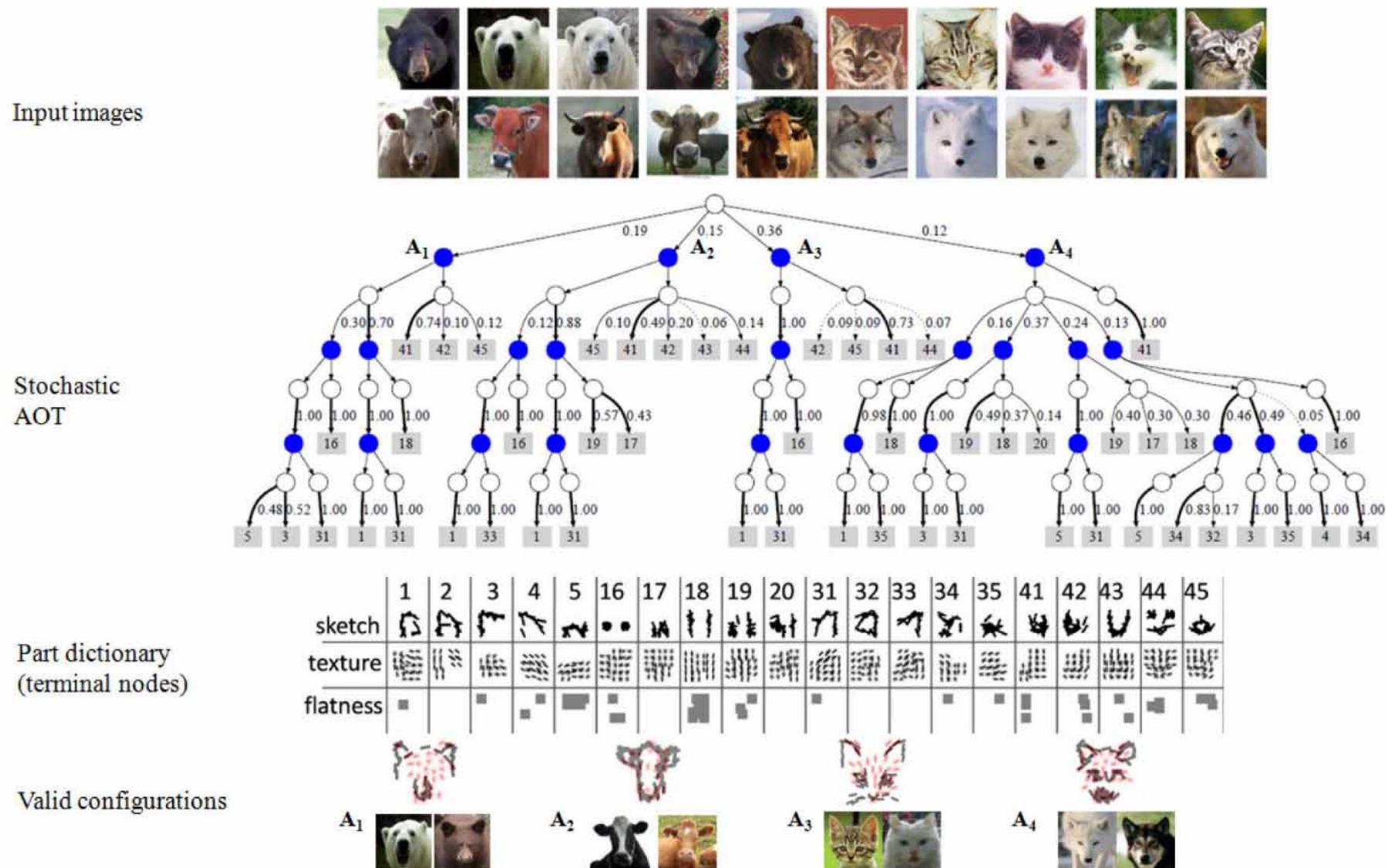
### Testing with Concept Activation Vectors (TCAV)



**Figure 1. Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., ‘striped’), and random examples ④, labeled training-data examples for the studied class (zebras) ⑤, and a trained network ③, TCAV can quantify the model’s sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept’s examples and examples in any layer ④. The CAV is the vector orthogonal to the classification boundary ( $v_C^l$ , red arrow). For the class of interest (zebras), TCAV uses the directional derivative  $S_{C,k,l}(x)$  to quantify conceptual sensitivity ⑥.

<https://github.com/tensorflow/tcav>

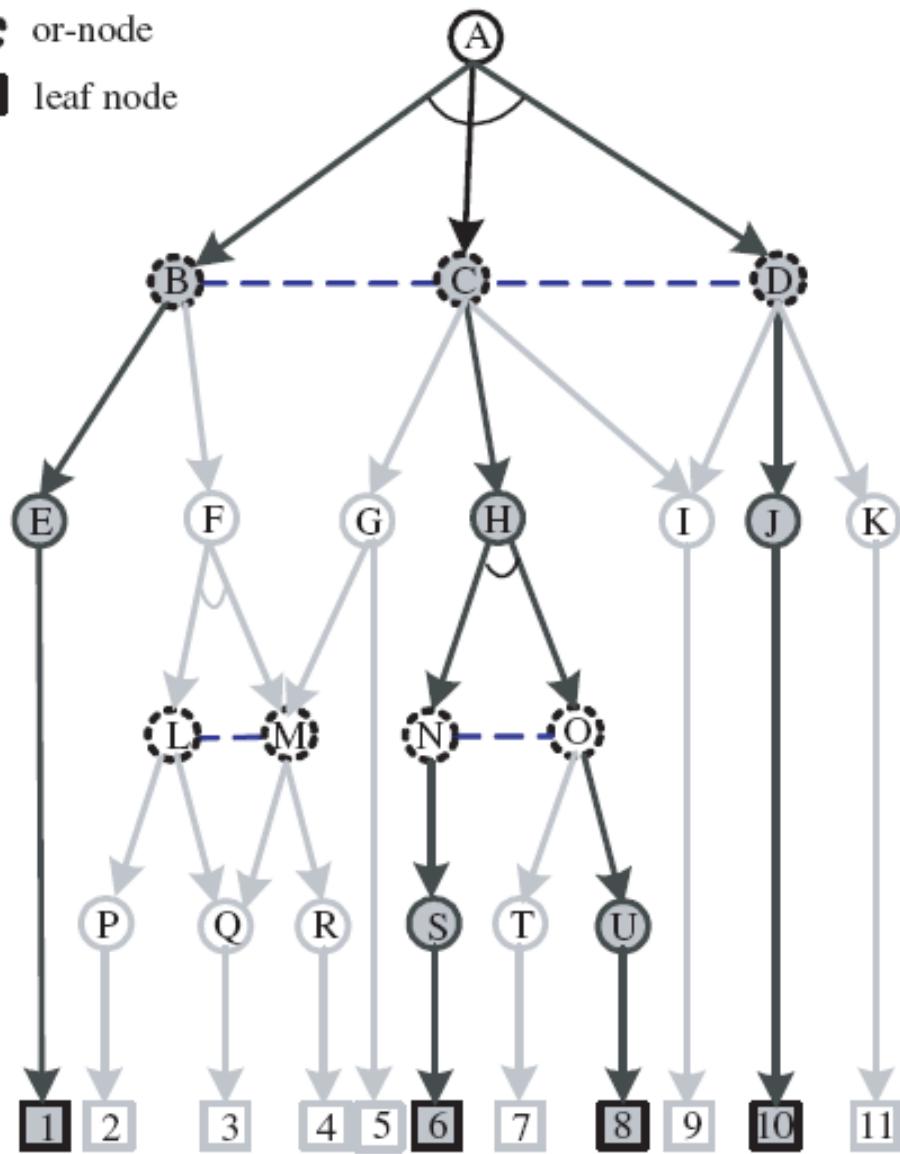
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.



	1	2	3	4	5	16	17	18	19	20	31	32	33	34	35	41	42	43	44	45
sketch																				
texture																				
flatness																				

Zhangzhang Si & Song-Chun Zhu 2013. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 35, (9), 2189-2205, doi:10.1109/TPAMI.2013.35.

- and-node
- or-node
- leaf node



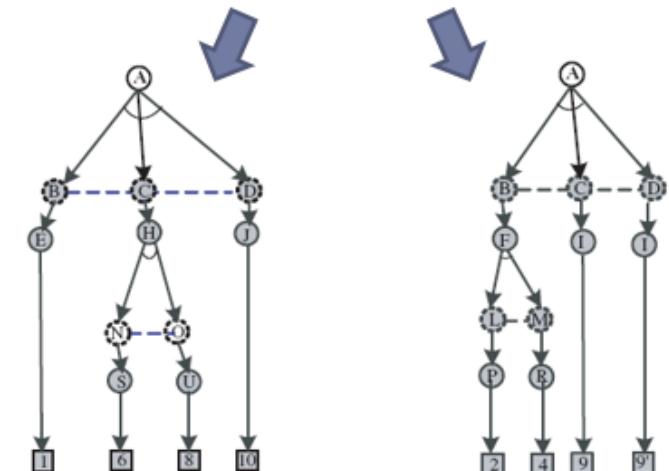
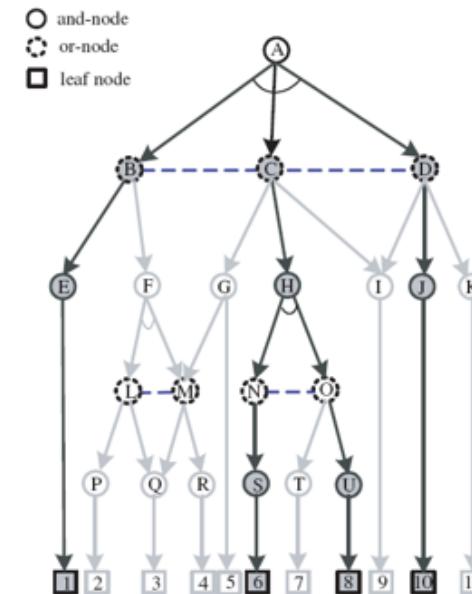
- Algorithm for this framework
  - Top-down/bottom-up computation
- Generalization of small sample
  - Use Monte Carlos simulation to synthesis more configurations
- Fill semantic gap

Images credit to Zhaoyin Jia (2009)

- ▶ Terminal (leaf) node:  $T(pg)$
- ▶ And-Or node:  $V^{or}(pg), V^{and}(pg)$
- ▶ Set of links:  $E(pg)$
- ▶ Switch variable at Or-node:  $w(t)$
- ▶ Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\begin{aligned} \xi(pg) = & \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_t(\alpha(t)) \\ & + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij}) \end{aligned}$$

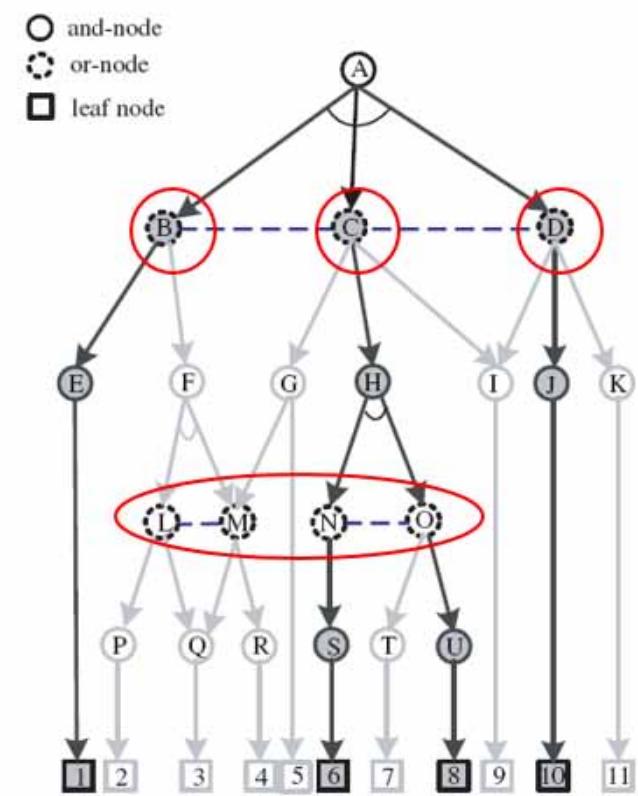


- ▶ Terminal (leaf) node:  $T(pg)$
- ▶ And-Or node:  $V^{or}(pg), V^{and}(pg)$
- ▶ Set of links:  $E(pg)$
- ▶ Switch variable at Or-node:  $w(t)$
- ▶ Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\begin{aligned} \xi(pg) &= \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_t(\alpha(t)) \\ &+ \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij}) \end{aligned}$$

SCFG: weigh the frequency at the children of or-nodes

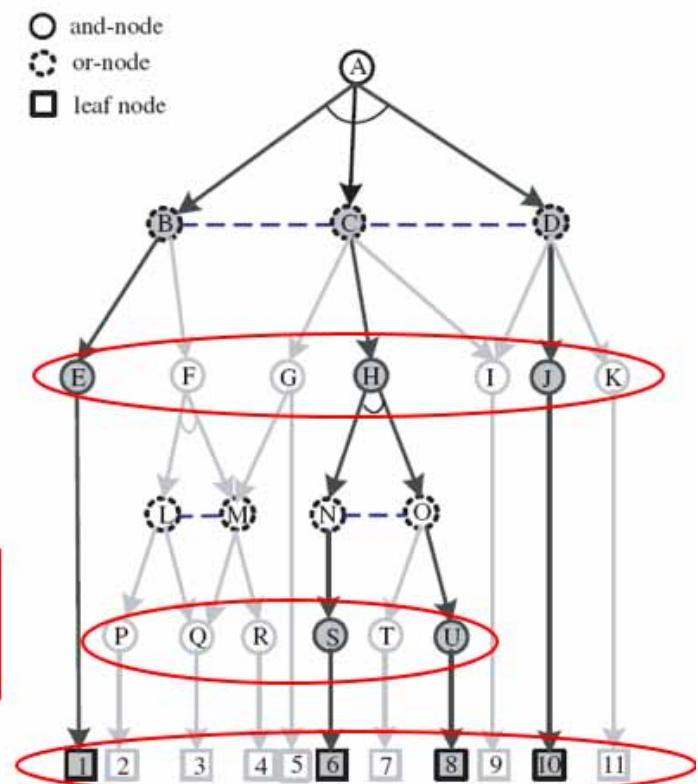


- ▶ Terminal (leaf) node:  $T(pg)$
- ▶ And-Or node:  $V^{or}(pg), V^{and}(pg)$
- ▶ Set of links:  $E(pg)$
- ▶ Switch variable at Or-node:  $w(t)$
- ▶ Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\begin{aligned} \xi(pg) = & \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \boxed{\sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_t(\alpha(t))} \\ & + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij}) \end{aligned}$$

Weigh the local compatibility of primitives (geometric and appearance)

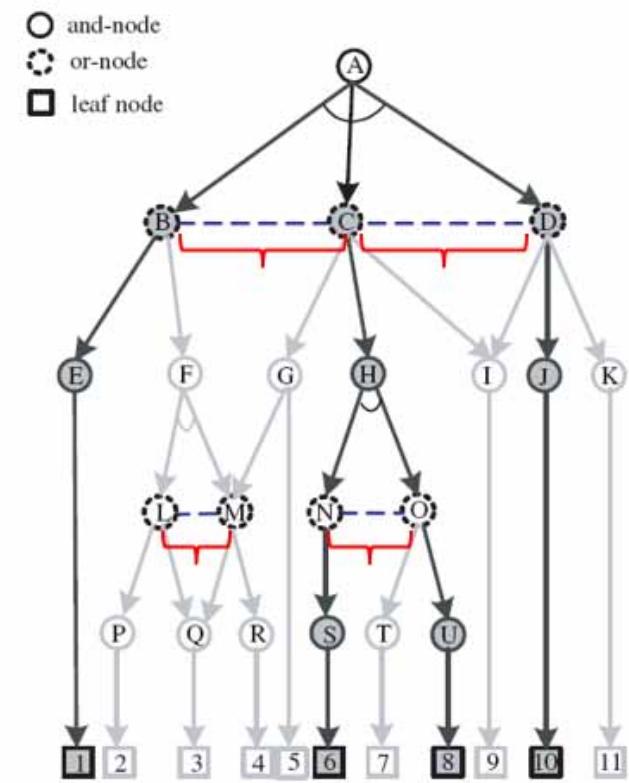


- ▶ Terminal (leaf) node:  $T(pg)$
- ▶ And-Or node:  $V^{or}(pg), V^{and}(pg)$
- ▶ Set of links:  $E(pg)$
- ▶ Switch variable at Or-node:  $w(t)$
- ▶ Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\begin{aligned} \xi(pg) = & \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_t(\alpha(t)) \\ & + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij}) \end{aligned}$$

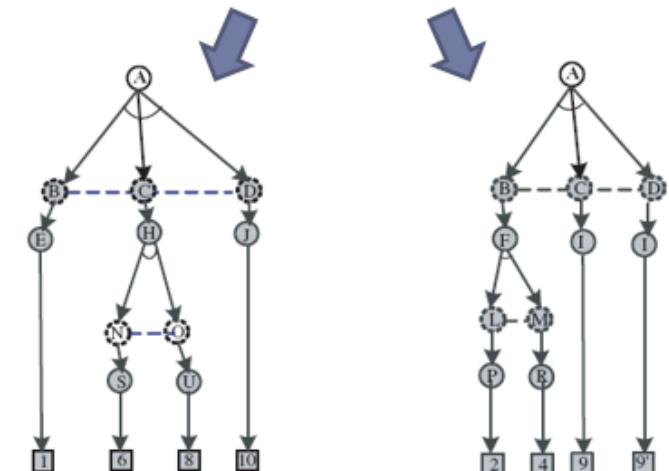
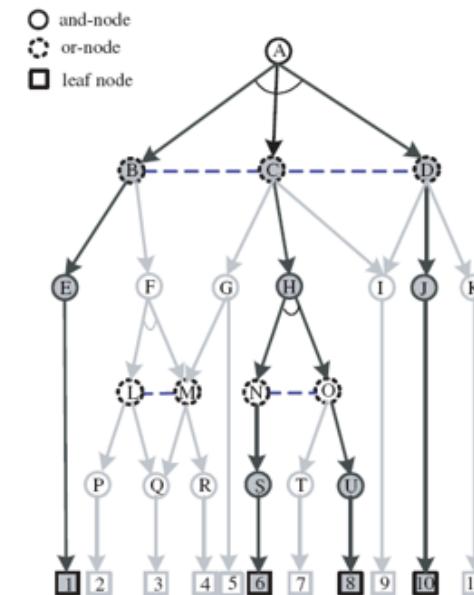
Spatial and appearance between primitives (parts or objects)

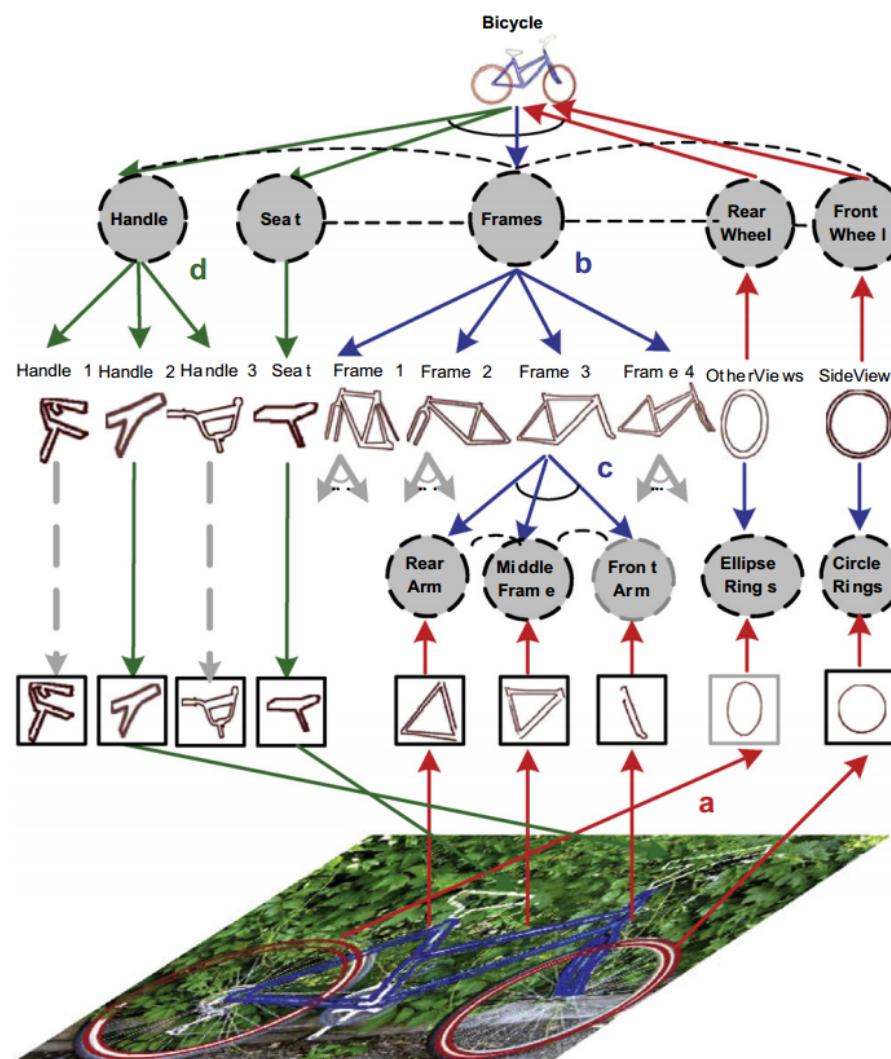


- ▶ Terminal (leaf) node:  $T(pg)$
- ▶ And-Or node:  $V^{or}(pg), V^{and}(pg)$
- ▶ Set of links:  $E(pg)$
- ▶ Switch variable at Or-node:  $w(t)$
- ▶ Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\begin{aligned} \xi(pg) = & \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_t(\alpha(t)) \\ & + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij}) \end{aligned}$$





**Input:** an input image  $I$ , and a set of constructed And-Or graphs of compositional object categories.

**Output:** a parsing graph  $pg_s$  of the scene that consists of the parsing graphs of detected objects.

- Repeat the following steps

- 1 Schedule the next node  $A$  to visit from the candidate parts.

- 2 Call Bottom-up( $A$ ) to update  $A$ 's **open** list.

- i Detect terminal instances of  $A$  from the image.

- ii Bind non-terminal instances of  $A$  from its children's **open** or **closed** lists

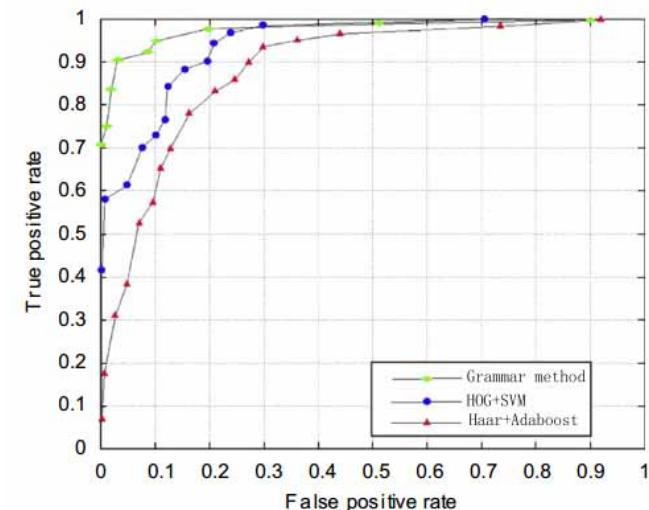
- 3 Call Top-down( $A$ ) to update  $A$ 's **open** or **closed** lists.

- i Accept hypotheses from  $A$ 's **open** list to its **closed** list.

- ii Remove (or disassemble) hypotheses from  $A$ 's **closed** list.

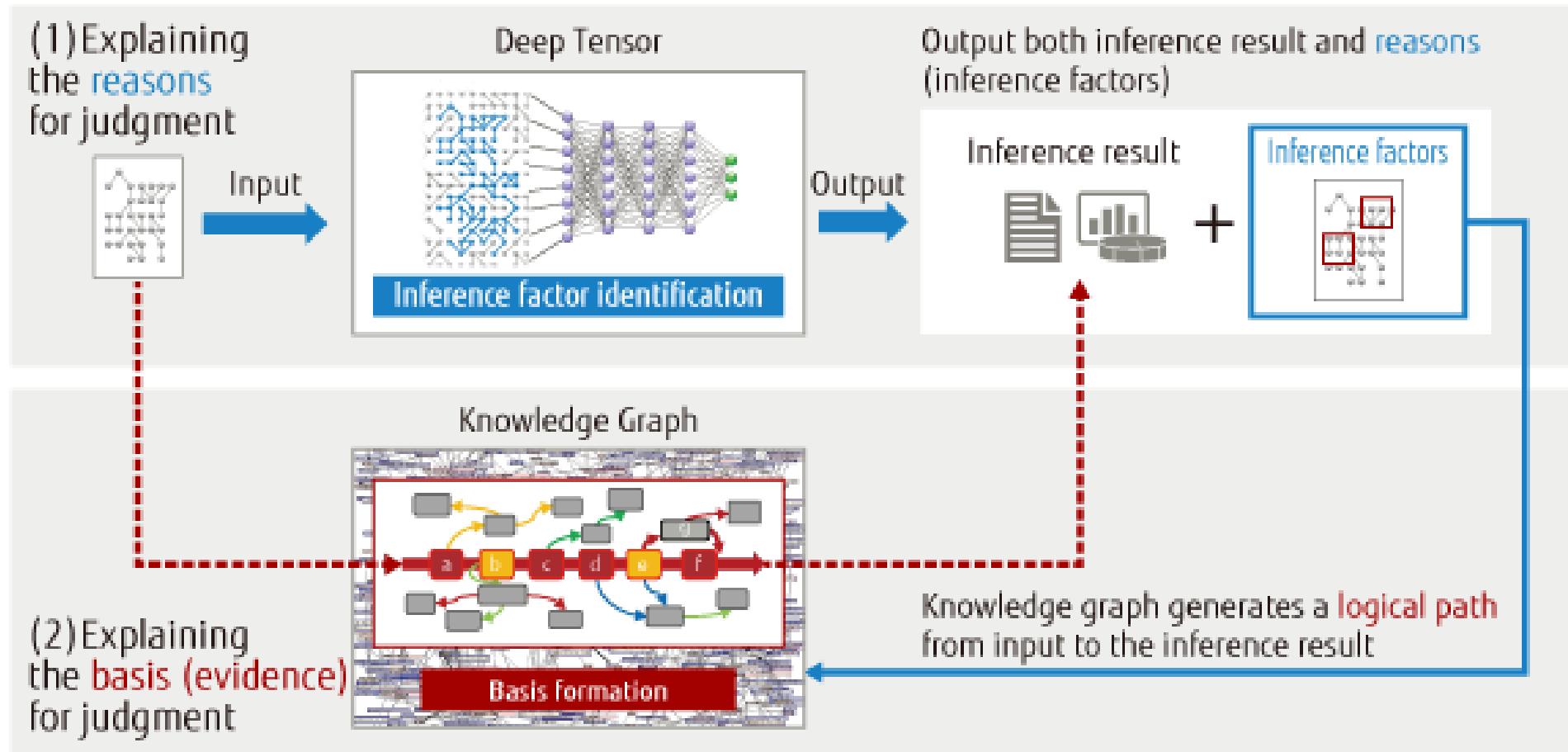
- iii Update the **open** lists for particles that overlap with node  $A$ .

- Until the particles in **open** list with weights higher than the empirical threshold are exhausted. Output all parsing graphs whose root nodes are reached.



Liang Lin, Tianfu Wu, Jake Porway & Zijian Xu 2009. A stochastic graph grammar for compositional object representation and recognition. Pattern Recognition, 42, (7), 1297-1307, doi:10.1016/j.patcog.2008.10.033.

# Future Work



## Explainable AI with Deep Tensor and Knowledge Graph

[http://www.fujitsu.com/jp/Images/artificial-intelligence-en\\_tcm102-3781779.png](http://www.fujitsu.com/jp/Images/artificial-intelligence-en_tcm102-3781779.png)

- What is a good explanation?
- (obviously if the other did understand it)
- Experiments needed!
- What is explainable/understandable/intelligible?
- When is it enough (Sättigungsgrad – you don't need more explanations – enough is enough)
- But how much is it ...



<https://www.newyorker.com/cartoon/a19697>

Noel C.F. Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei & Aleksandra Mojsilović 2018. Teaching Meaningful Explanations. arXiv:1805.11648.

iv:1805.11648v1 [cs.AI] 29 May 2018

# Teaching Meaningful Explanations

Noel C. F. Codella,\* Michael Hind,\* Karthikeyan Natesan Ramamurthy,\*  
Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei,  
Aleksandra Mojsilović

\* These authors contributed equally.

IBM Research

Yorktown Heights, NY 10598

{nccodell,hindm,knatesa,mcam,adhuran,krvarshn,dwei,aleksand}@us.ibm.com

## Abstract

The adoption of machine learning in high-stakes applications such as healthcare and law has lagged in part because predictions are not accompanied by explanations comprehensible to the domain user, who often holds ultimate responsibility for decisions and outcomes. In this paper, we propose an approach to generate such explanations in which training data is augmented to include, in addition to features and labels, explanations elicited from domain users. A joint model is then learned to produce both labels and explanations from the input features. This simple idea ensures that explanations are tailored to the complexity expectations and domain knowledge of the consumer. Evaluation spans multiple modeling techniques on a simple game dataset, an image dataset, and a chemical odor dataset, showing that our approach is generalizable across domains and algorithms. Results demonstrate that meaningful explanations can be reliably taught to machine learning algorithms, and in some cases, improve modeling accuracy.

## 1 Introduction

New regulations call for automated decision making systems to provide “meaningful information” on the logic used to reach conclusions [1–4]. Selbst and Powles interpret the concept of “meaningful information” as information that should be understandable to the audience (potentially individuals