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

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

Important Definition:

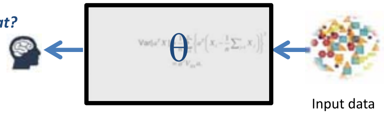
- Causability := a property of a person (Human)
- Explainability := a property of a system (Computer)

Human intelligence (Cognitive Science)

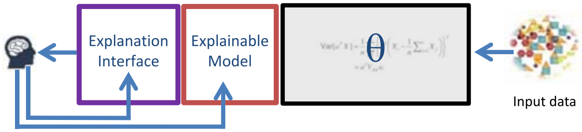
Machine intelligence (Computer Science)

The challenge

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

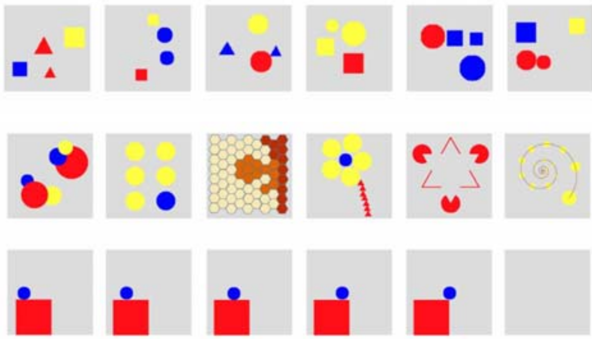


A possible solution



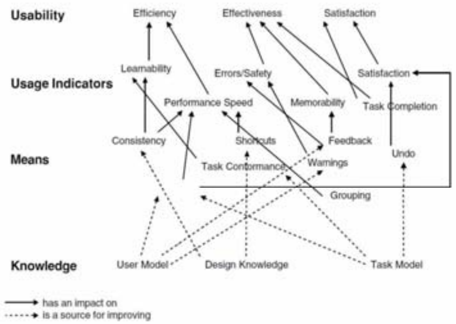
The domain expert can understand why …
The domain expert can learn and correct errors …
The domain expert can re-enact on demand …

Our Playground



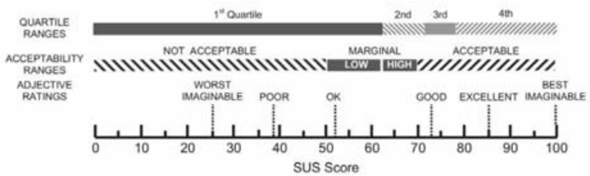
First Homework for you: make yourself familiar …

- 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.*

System Usability Scale

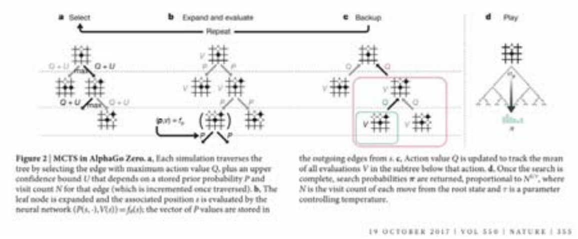


Bangor, A., Kortum, P. T. & Miller, J. T. (2008) An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24, 6, 574-594.*

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

Explainability

Mastering the game of Go without human knowledge



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

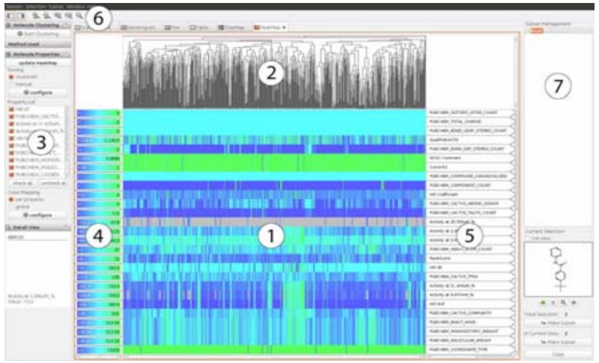
David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel & Demis Hassabis 2017. Mastering the game of go without 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.

Example for an Explanation Interface

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.

Example for an Explanation Interface - open work ☺

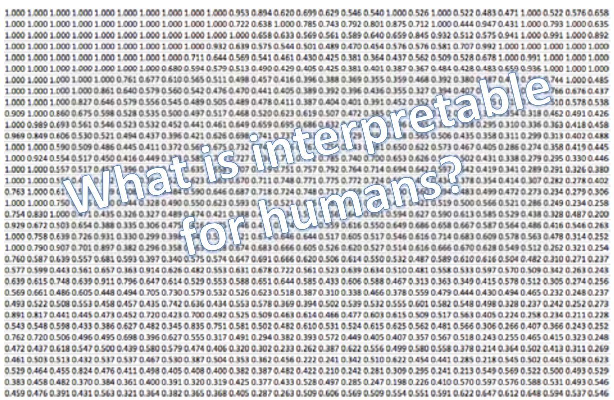


Werner Sturm, Till Schaefer, Tobias Schreck, Andeas Holzinger & Torsten Ullrich. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps In: Borgo, Rita & Turkay, Cagatay, eds. EG UK Computer Graphics & Visual Computing CGVC 2015, 2015 University College London (UCL). Euro Graphics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

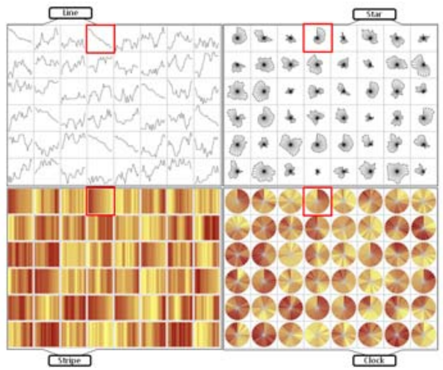
Research and Development in Interpretability …

- 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.

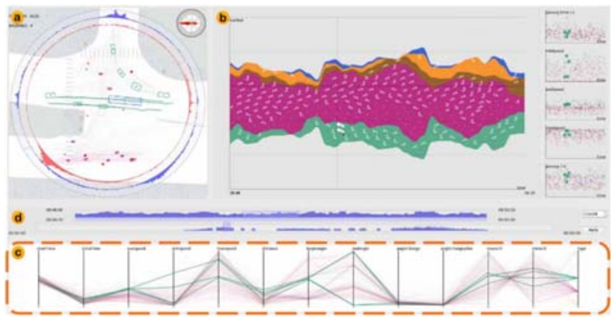


What is understandable, interpretable, intelligible?



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

Explainable AI is a huge challenge for visualization



Methods of Explainable AI

Methods of ex-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)

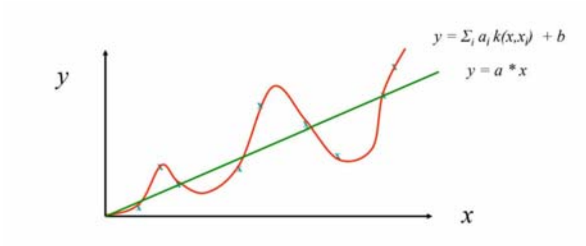
Remember: Reasoning = "Sensemaking"

- **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.

Important Definition: Ground truth

- := 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.

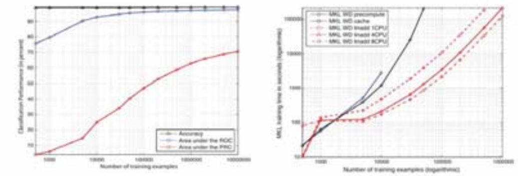
Empirical Inference Example



Gottfried W. Leibniz (1646-1716)
Hermann Weyl (1885-1955)
Vladimir Vapnik (1936-)
Alexey Chervonenkis (1938-2014)
Gregory Chaitin (1947-)

Remember: hard inference problems

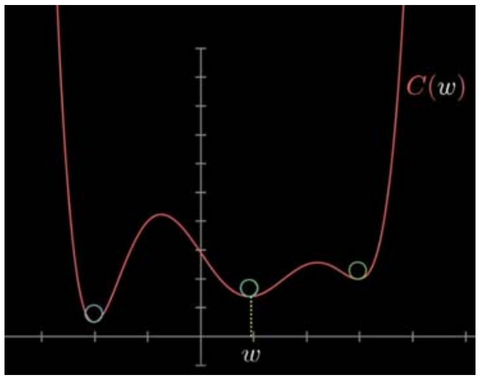
- High dimensionality (curse of dim., many factors contribute)
- Complexity (real-world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - *) = Def.: a sequence or collection of random variables is independent and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



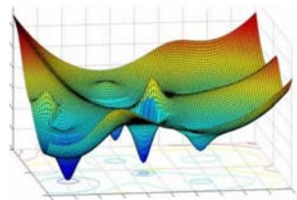
Sören Sonnenburg, Gunnar Rätsch, Christin Schaefer & Bernhard Schölkopf 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

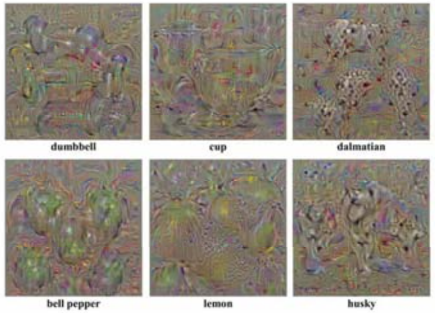
https://lrpserver.hhi.fraunhofer.de/handwriting-classification

Gradients



Gradients





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

Gradients

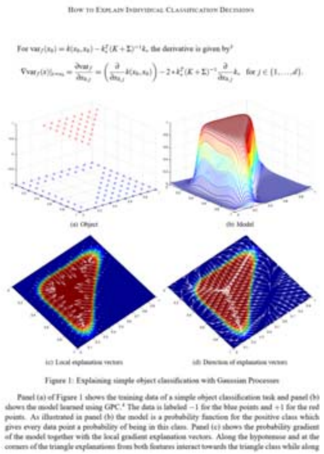
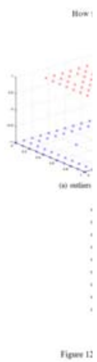


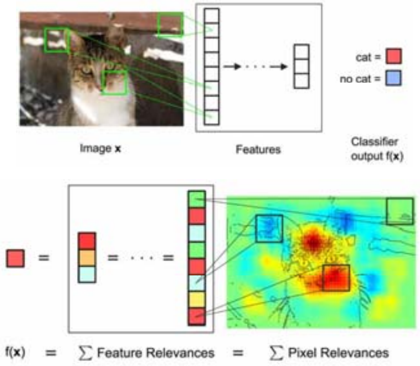
Figure 1: Explaining simple object classification with Gaussian Processes



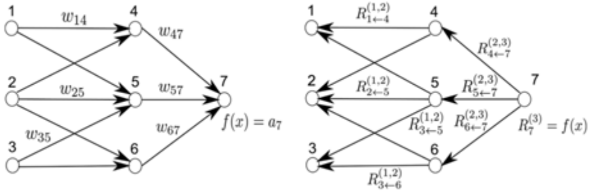
David Baehrens, Timor … How to explain individ… Robert Mueller 2010. … 11, (6), 1803-1831.

LRP Layer-Wise Relevance Propagation

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.



A NN-classifier during prediction time



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

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.

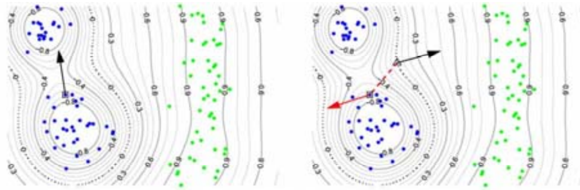


Fig 3. An exemplary real-valued prediction function for classification with the dashed black line being the decision boundary which separates the blue from the green dots. The blue dots are labeled negatively, the green dots are labeled positively. Left: Local gradient of the classification function 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

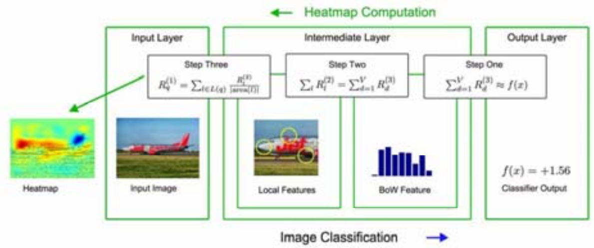
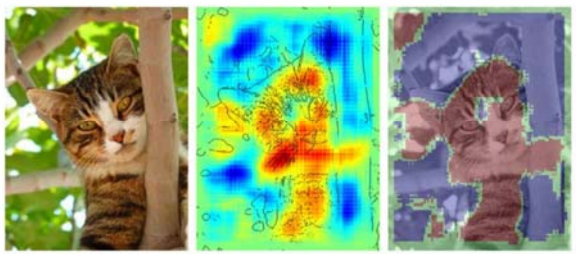


Fig 4. Local and global predictions for input images are obtained by following a series of steps through the classification- and pixel-wise decomposition pipelines. Each step taken towards the final pixel-wise decomposition has a complementing analogue within the Bag of Words classification pipeline. 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 heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

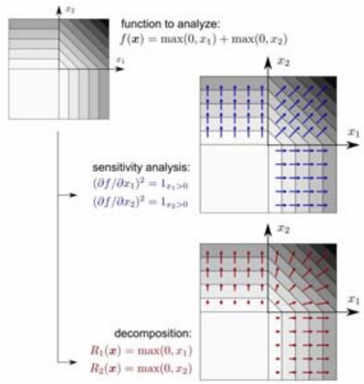
$$\forall\, \boldsymbol{x}: f(\boldsymbol{x}) = \sum_p R_p(\boldsymbol{x}).$$

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

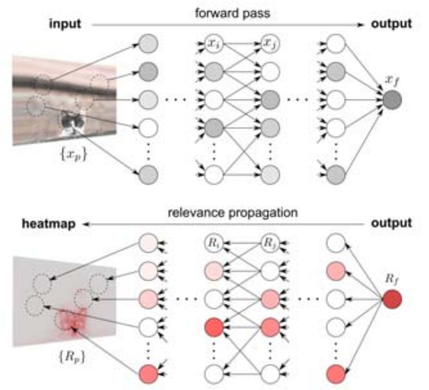
$$\forall\, \boldsymbol{x}, p: R_p(\boldsymbol{x}) \geq 0$$

Definition 3. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *consistent* if it is conservative *and* positive. That is, it is consistent if it complies with Definitions 1 and 2.

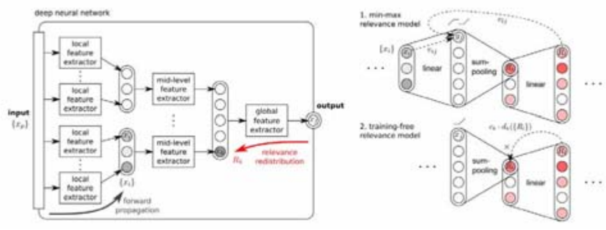
Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.



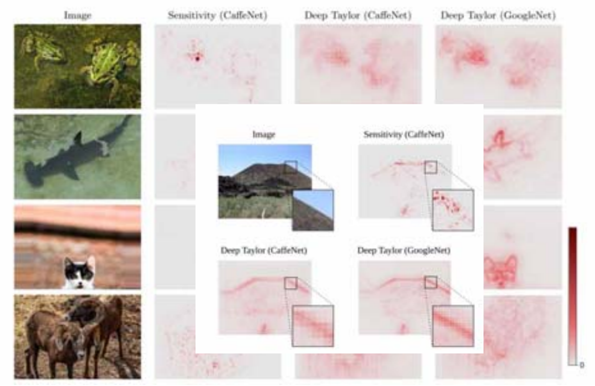
Relevance propagation



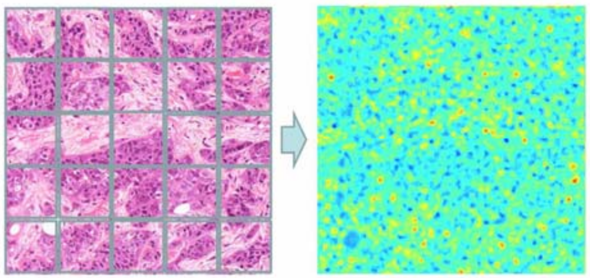
Relevance Redistribution



Example 1

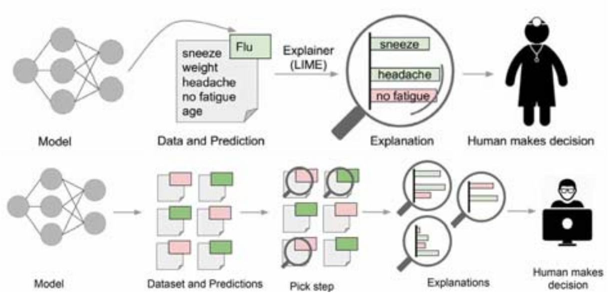


Example 2 Histopathology



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.

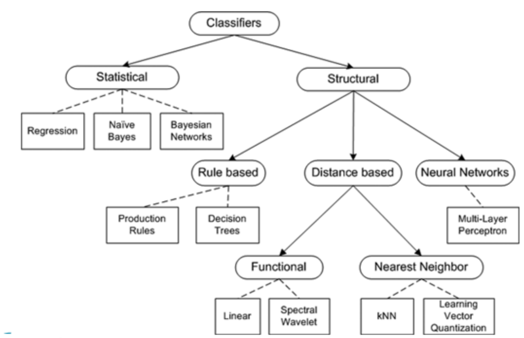
LIME – Local Interpretable Model Agnostic Explanations



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.

Example LIME – Model Agnostic Explanation

Remember: there are myriads of classifiers …



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

Black Box Explanations through Transparent Approximations

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:

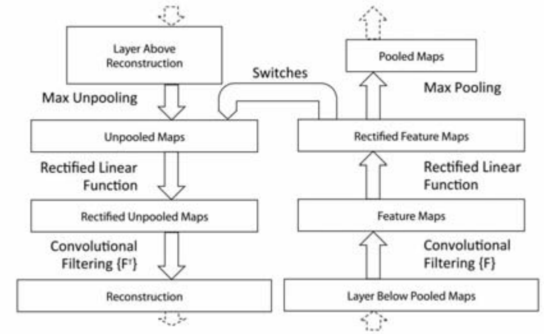
If Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression

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

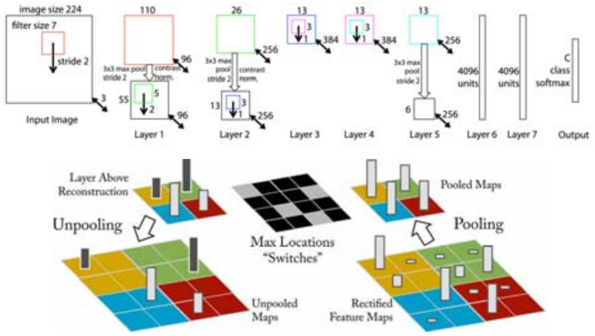
Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

Example: Interpretable Deep Learning Model

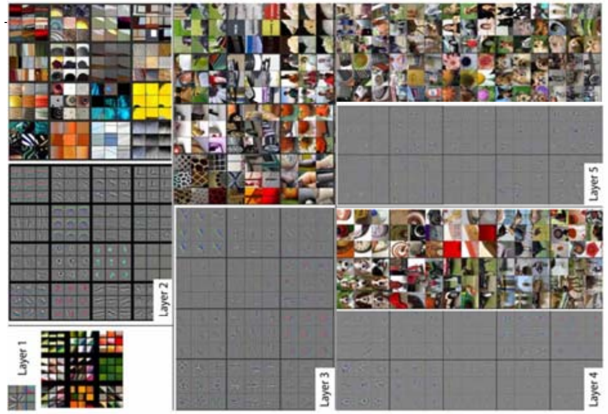


Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

Visualizing a Conv Net with a De-Conv Net

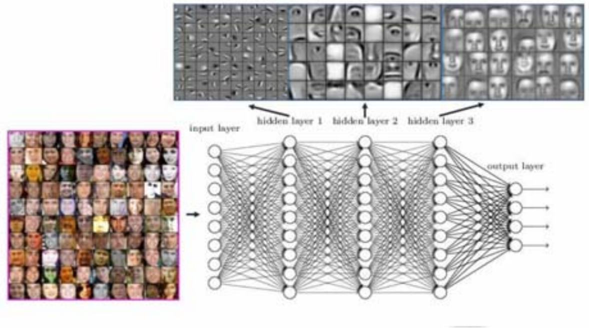


Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



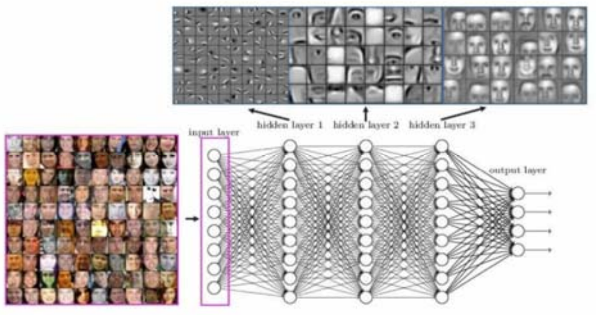
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

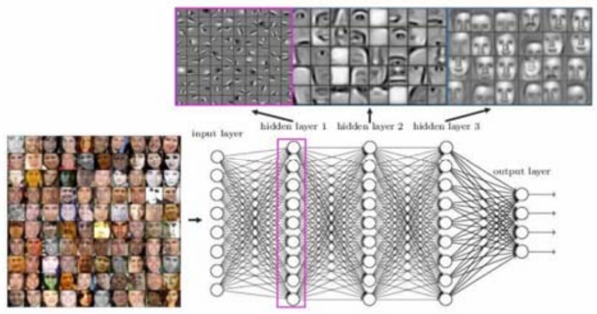
The world is compositional (Yann LeCun)

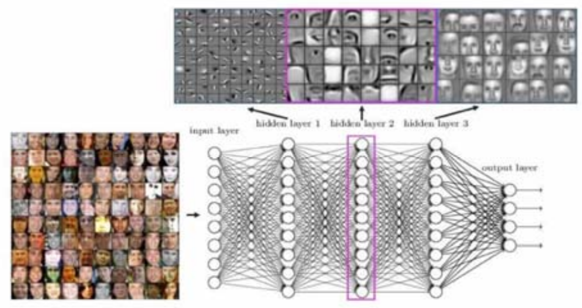


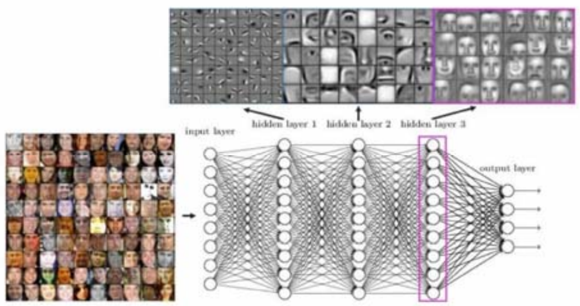
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

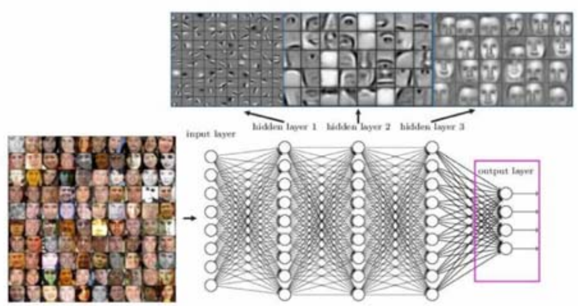
The world is compositional (Yann LeCun)

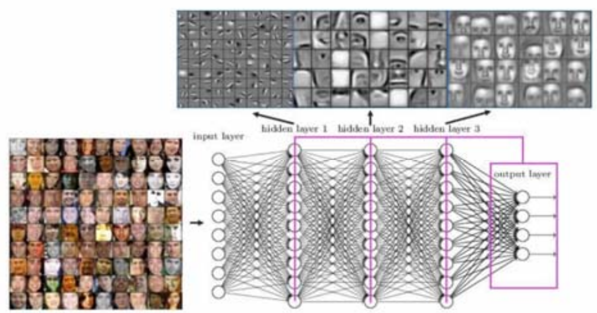












TCAV Testing with Concept Activation Vectors

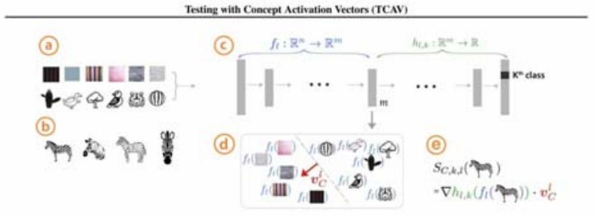
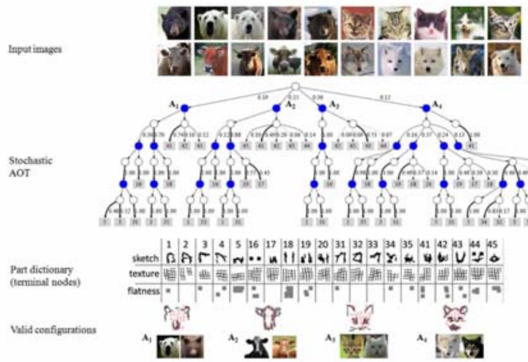


Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples (a), labeled training-data examples for the studied class (zebras) (b), and a trained network (c), TCAV can quantify the model's sensitivity to the concept 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 (d). 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 (e).

https://github.com/tensorflow/tcav

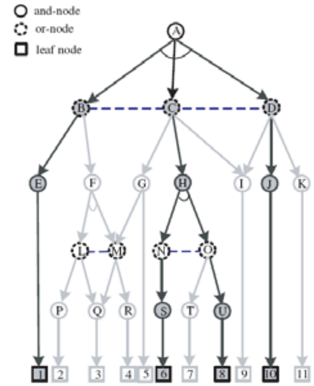
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Learning interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.

Stochastic AND-OR Templates for visual objects



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.

Framework for vision: AND-OR Graphs



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

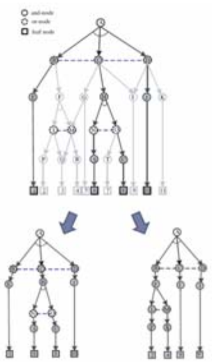
Images credit to Zhaoyin Jia (2009)

Stochastic Model on AND-OR graph: 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: $\alpha(t)$

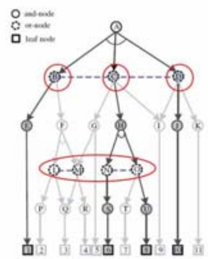
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

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



Stochastic Model on AND-OR graph: 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: $\alpha(t)$



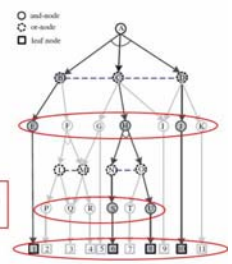
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

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

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

Stochastic Model on AND-OR graph: 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: $\alpha(t)$



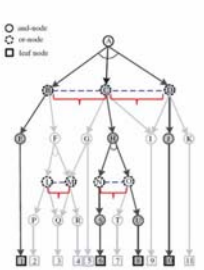
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

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

Weigh the local compatibility of primitives (geometric and appearance)

Stochastic Model on AND-OR graph: 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: $\alpha(t)$



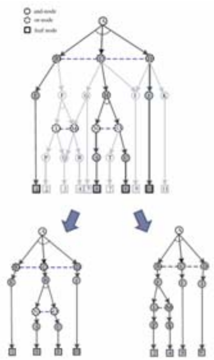
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

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

Spatial and appearance between primitives (parts or objects)

Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

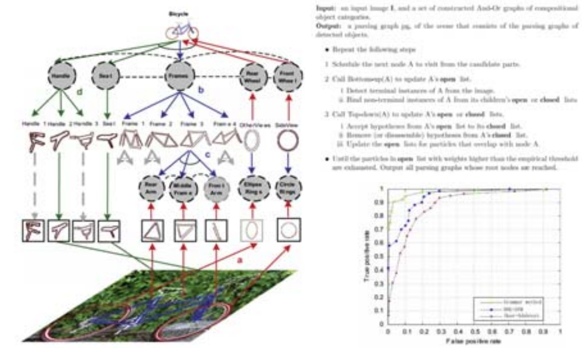
- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$



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

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

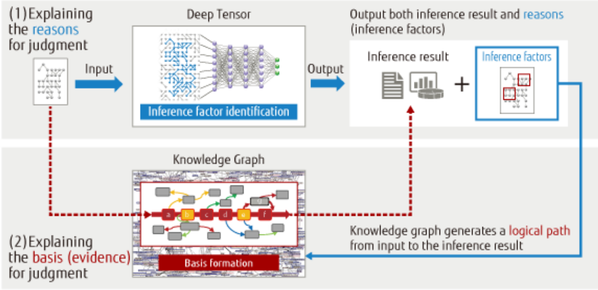
Stochastic graph grammar/comp. object representation



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

Combination of Deep Learning with Ontologies



Explainable AI with Deep Tensor and Knowledge Graph

http://www.fujitsu.com/jp/Images/artificial-intelligence-en_tcm102-3781779.png

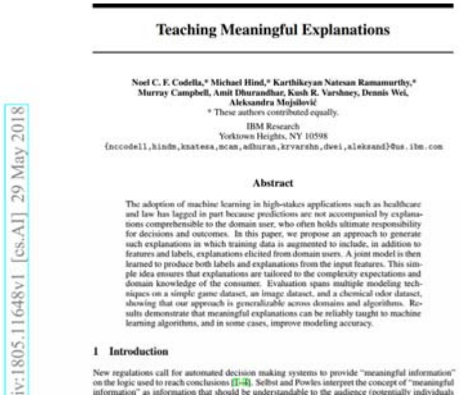
Seemingly trivial questions … ?

- 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 …

Explanations in Artificial Intelligence will be necessary

IBM is doing it now: teaching meaningful explanations

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



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