Assoc.Prof. Dr. Andreas Holzinger

185.A83 Machine Learning for Health Informatics
2020S, VU, 2.0 h, 3.0 ECTS
Andreas Holzinger, Marcus Bloice, Florian Endel, Anna Saranti
Lecture 06 - Week 19

Selected Methods of explainable AI

Contact: andreas.holzinger AT tuwien.ac.at

https://human-centered.ai/machine-learning-for-health-informatics-class-2020

LV 185.A83 Machine Learning for Health Informatics (Class of 2020)

Andreas HOLZINGER, Marcus BLOICE, Florian ENDEL, Anna SARANTI

Study Code: 066 936 Master program Medical Informatics

https://tiss.tuwien.ac.at/curriculum/public/curriculum.xhtml?dswid=9468&dsrid=253&key=56089&semester=NEXT

Semester hours: 2.0 h; ECTS-Credits: 3.0; Type: VU Lecture and Exercises with Python

ECTS-Breakdown (sum=75 h, corresponds with 3 ECTS, where 1 ECTS = 25 h workload):

Presence during lecture	8 * 3 h	24 h
Preparation before and after lecture	8 * 1 h	08 h
Preparation of assignments and presentation	28 h + 2 h	30 h
Written exam including preparation	1 h + 12 h	13 h
TOTAL students' workload		**75 h**

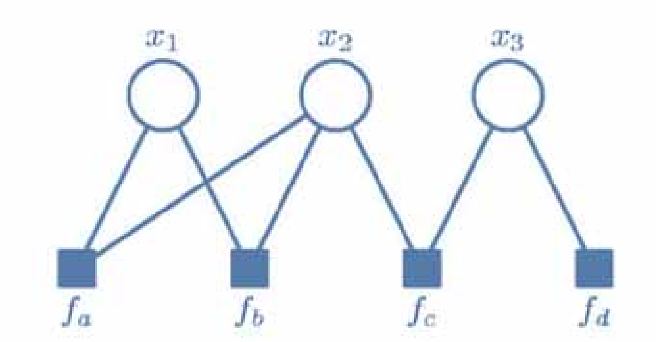
Class URL: https://human-centered.ai/machine-learning-for-health-informatics-class-2020

Class Schedule for 2020 (subject to change: please check class URL for any changes):

Nr	*Week*	*Topic*
01	12	Introduction and overview: From health informatics to ethical responsible medical AI
02	13	Data for machine learning, Probabilistic information and entropy: On data quality, data integration, data augmentation, information theory
03	14	Tutorial T01 and Python assignment A01 (Data augmentation) Tutor: Marcus BLOICE
		Happy Easter
04	17	Probabilistic graphical models: From knowledge representation to graph model learning
05	18	Tutorial T02 and Python assignment A02 (Probabilistic programming) Tutor: Florian ENDEL
06	19	Selected methods of explainable AI: LIME, BETA, LRP, Deep Taylor Decomposition, PDA, TCAV etc.
07	20	Tutorial T03 and Python assignment A03 (LRP) Tutor: Anna SARANTI
		Finalization of assignments and exam preparation
08	24	Course finalization and grading (detailed information will be given in due course)

- 00 Reflection – follow-up from last lecture
- 01 Explainability, Interpretability, Causability
- 02 Is xAI new ?
- 03 Examples for Ante-hoc models (explainable models, interpretable machine learning)
- 04 Examples for Post-hoc models (making the "black-box" model interpretable)
- 04a LIME, 04b BETA, 04c LRP, 04d Taylor, 04e Prediction Difference Analysis, 04f TCAV

00 Reflection



1

$$p(\boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\theta}) \left[\prod_{i=1}^{N} p(\mathbf{x}_i | \boldsymbol{\theta}) \right]$$

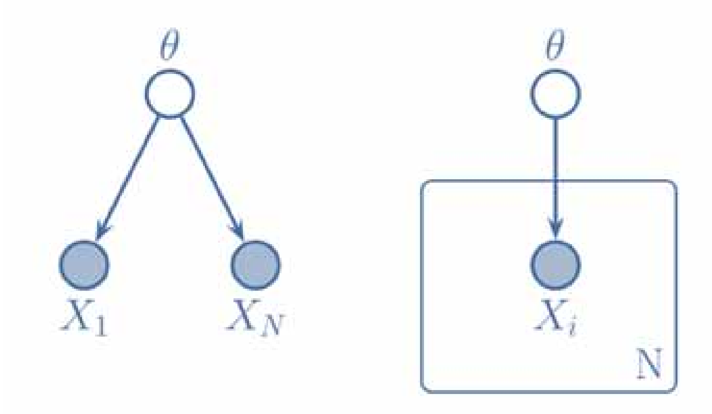
2

Deduction

Induction

Abduction

4



3

01 Explainability, Interpretability, Causability, …

- Explainability = motivated by the opaqueness of so called “black-box” ML approaches
- it is the ability to provide an explanation on **why** a machine decision has been reached (e.g. why is it a cat what the deep network recognized).
- Note: Finding an appropriate explanation is difficult, because this needs understanding the context
- and providing a description of causality and consequences of a given fact.
- German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz

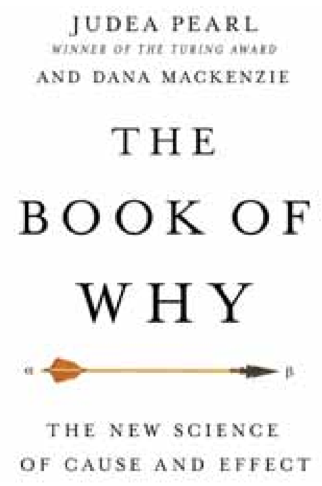
- Interpretability := ability to explain or to provide the meaning in understandable terms to a human
- Understandability (intelligibility) := characteristic of a model to make a human understand its function – how the model works (without any need for explaining its internal structure).
- Comprehensibility := ability of a learning algorithm to represent its learned knowledge in a human understandable fashion entities.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila & Francisco Herrera 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115, doi:10.1016/j.inffus.2019.12.012.

http://bayes.cs.ucla.edu/LECTURE/lecture_sec1.htm

Judea Pearl & Dana Mackenzie 2018. The book of why, New York, Basic Books

http://bayes.cs.ucla.edu/WHY



https://www.youtube.com/watch?v=pEBI0vF45ic

Causality: The art and science of cause and effect

Judea Pearl 2000. Causality: Models, Reasoning, and Inference, Cambridge: Cambridge University Press.

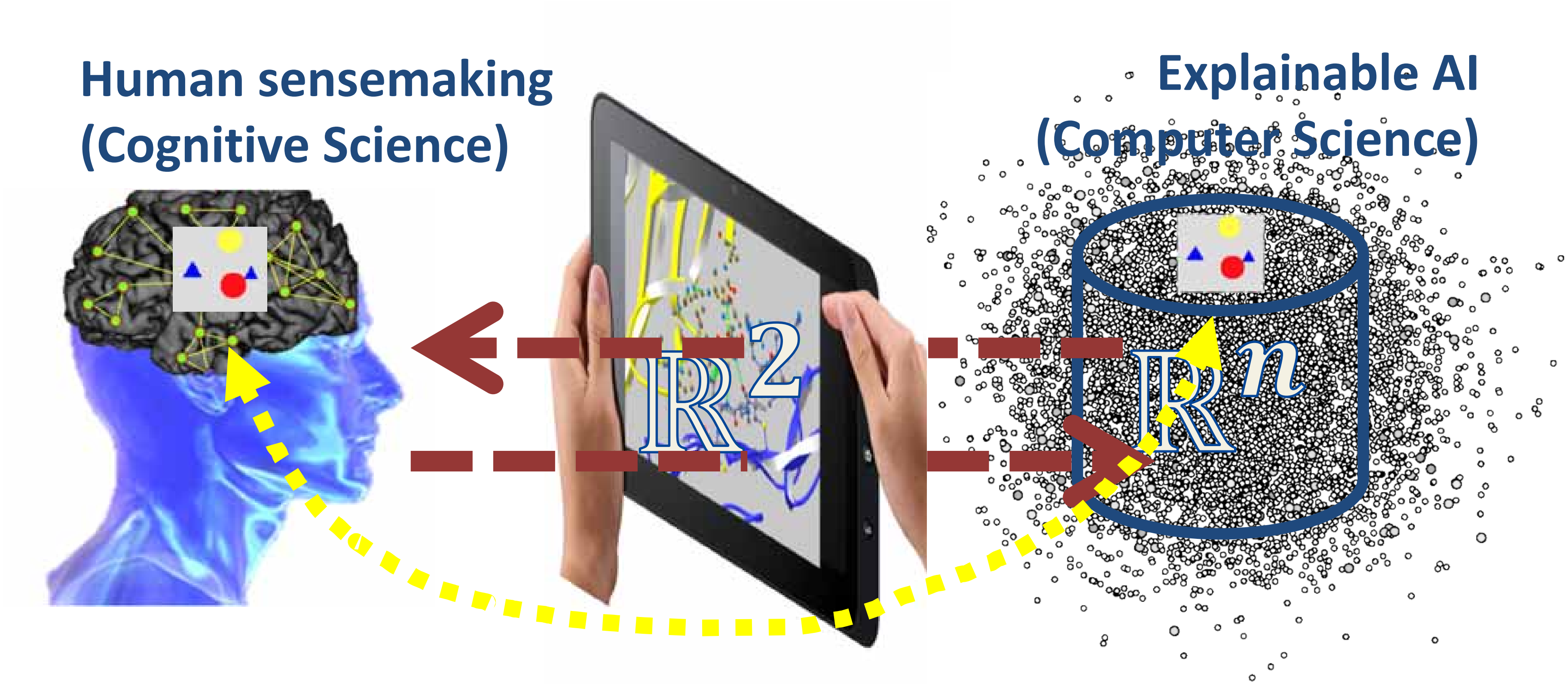
Causability: Mapping machine explanations with human understanding

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, (4), doi:10.1002/widm.1312.

- Explainability = in a technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation. **It does not refer to an explicit human model.**
- Causability = as the extent to which an explanation of a statement to a human expert achieves a specified level of **causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.**

Holzinger, A., Carrington, A. & Müller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Ed. Kristian Kersting, TU Darmstadt, 34, (2), doi:10.1007/s13218-020-00636-z.

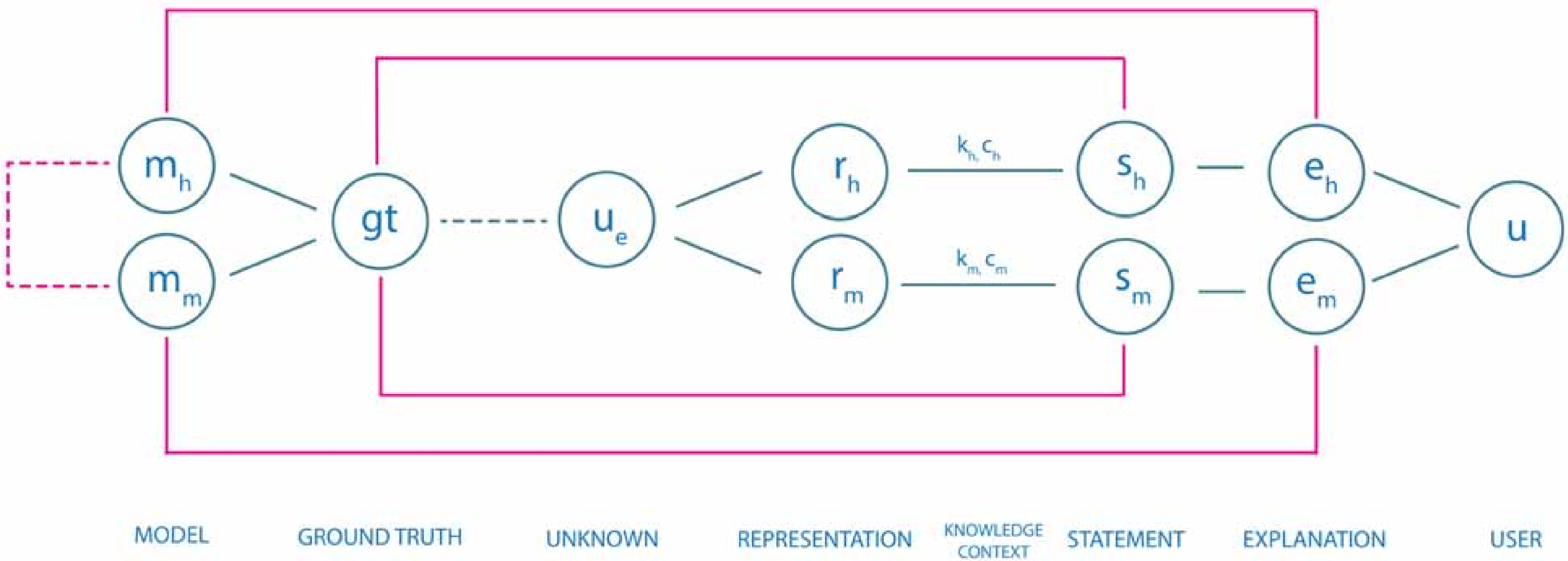
- **Explainability := a property of a system (Computer)**
- **Causability := a property of a person (Human)**



Andreas Holzinger. On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: Helfert, Markus, Fancalanci, Chiara & Filipe, Joaquim, eds. DATA 2012, International Conference on Data Technologies and Applications, 2012 Rome, Italy. INSTICC, 5-16.

Measuring the quality of Explanations: The Systems Causability Scale

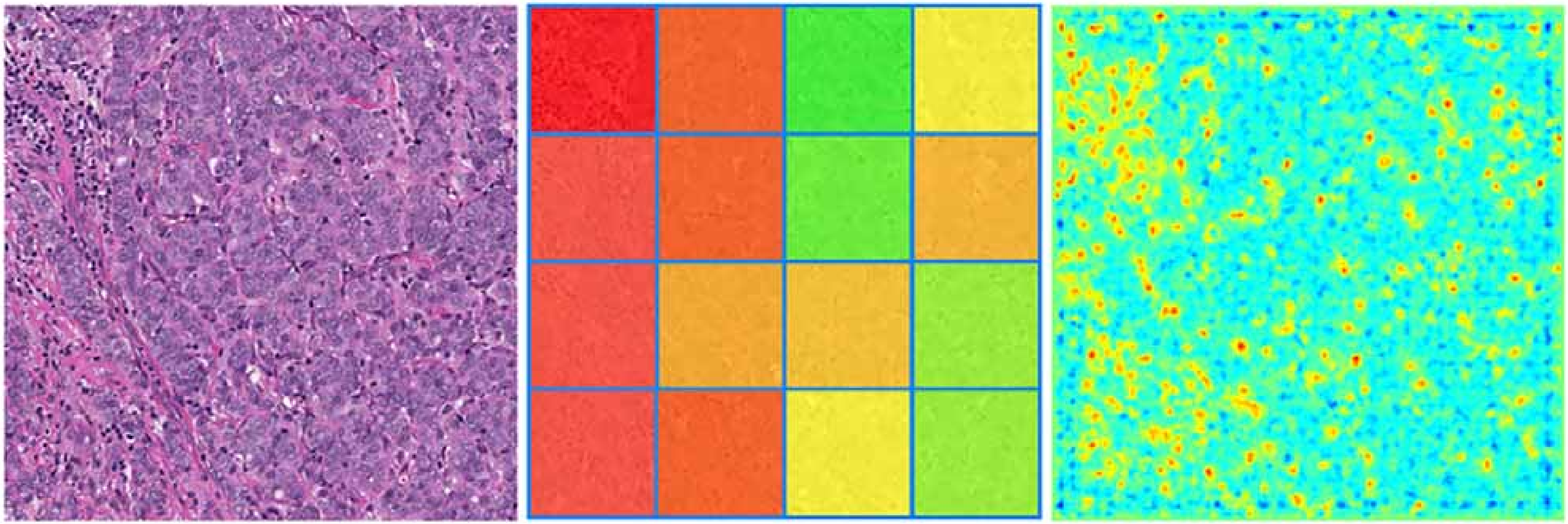
Andreas Holzinger, Andre Carrington & Heimo Müller 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt, 34, (2), doi:10.1007/s13218-020-00636-z



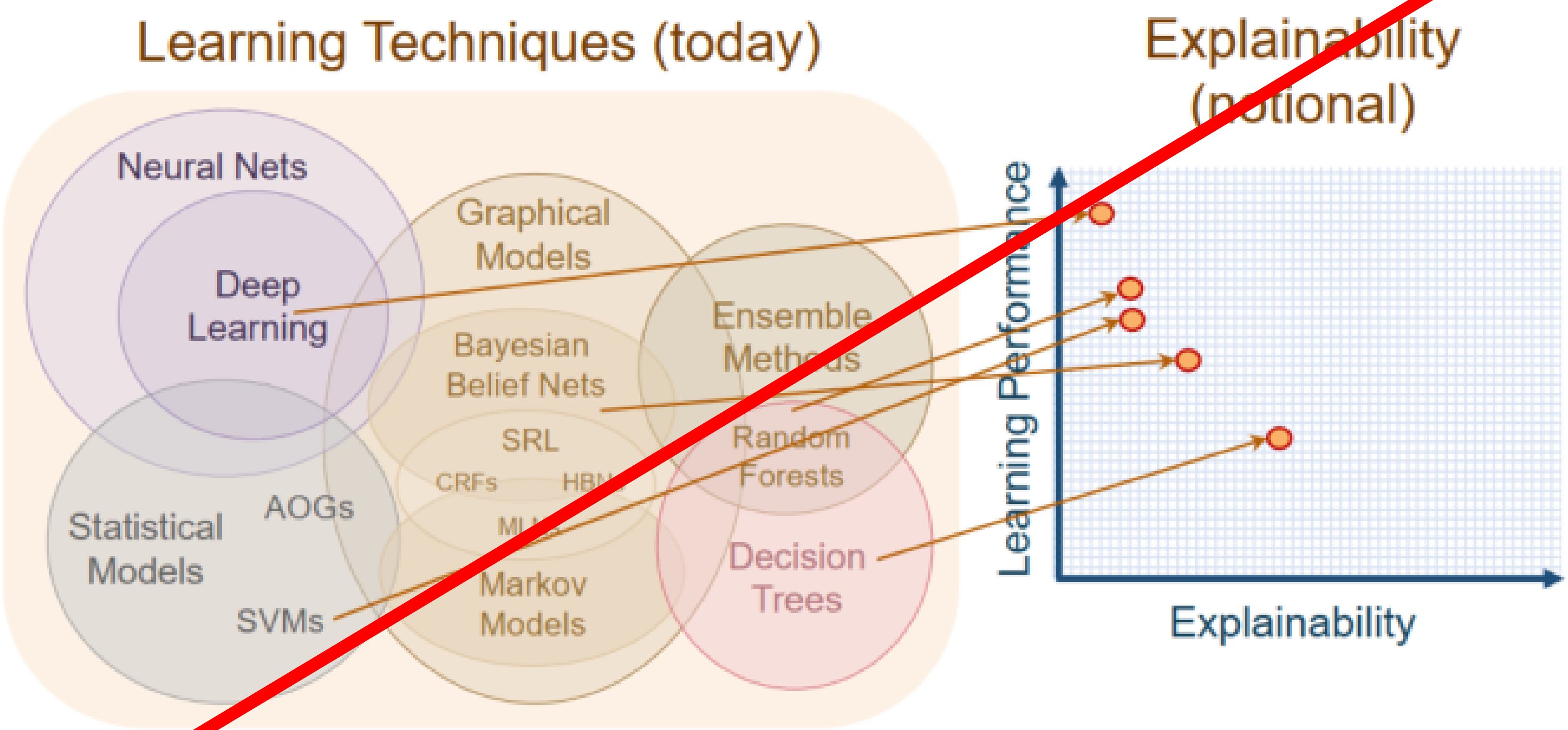
Andreas Holzinger, Andre Carrington & Heimo Müller 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt, 34, (2), doi:10.1007/s13218-020-00636-z.

- 1) ground truth is not always well defined, especially when making a medical diagnosis;
- 2) although human (scientific) models are often based on understanding causal mechanisms, today's successful machine models or algorithms are typically based on correlation or related concepts of similarity and distance!

what - to whom - how



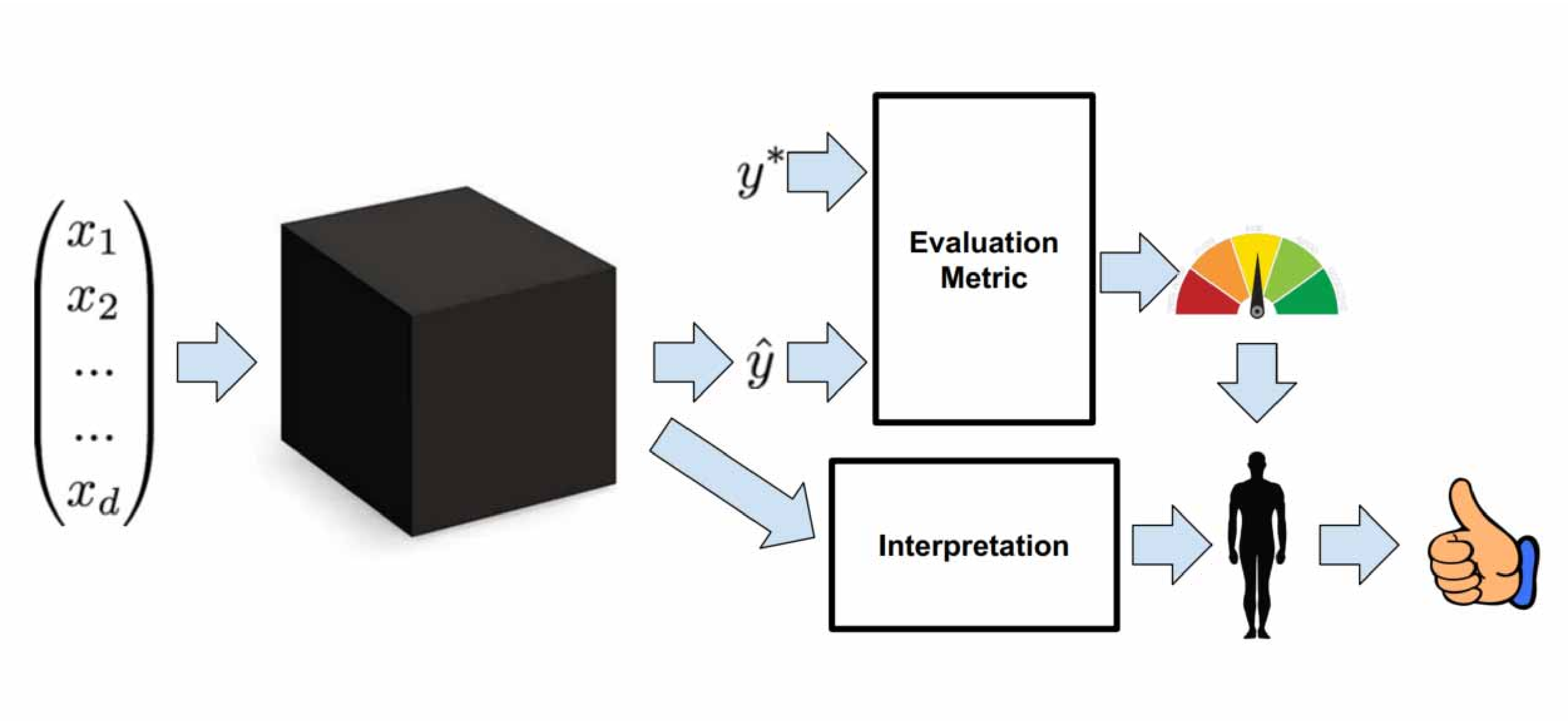
Frederick Klauschen, K.-R. Müller, Alexander Binder, Michael Bockmayr, M. Hägele, P. Seegerer, Stephan Wienert, Giancarlo Pruneri, S. De Maria & S. Badve 2018. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Seminars in cancer biology, 52, 151-157, doi:10.1016/j.semcancer.2018.07.001.



https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

This is far too naïve: Explainability (better: interpretability !) does not correlate with performance !!

How is Explainability/ Interpretability contrasted to performance ?



Zachary C. Lipton 2018. The mythos of model interpretability. ACM Queue, 16, (3), 31-57, doi:10.1145/3236386.3241340

- Explanation is a <u>reasoning process</u>
- Open questions:
  - What is a good explanation?
  - When is it enough (degree of saturation)?
  - Context dependent (Emergency vs. research)
  - How can we measure the degree of comprehensibility of a given explanation -> (System Causability Scale, SCS)
  - (obviously the explanation was good when it has been understood by the human)
  - What can the system learn from the human?
  - What can the human learn from the system?
  - Measuring explanation effectiveness!

Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland & Patrick Vinck 2018. Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology, 31, (4), 611-627.

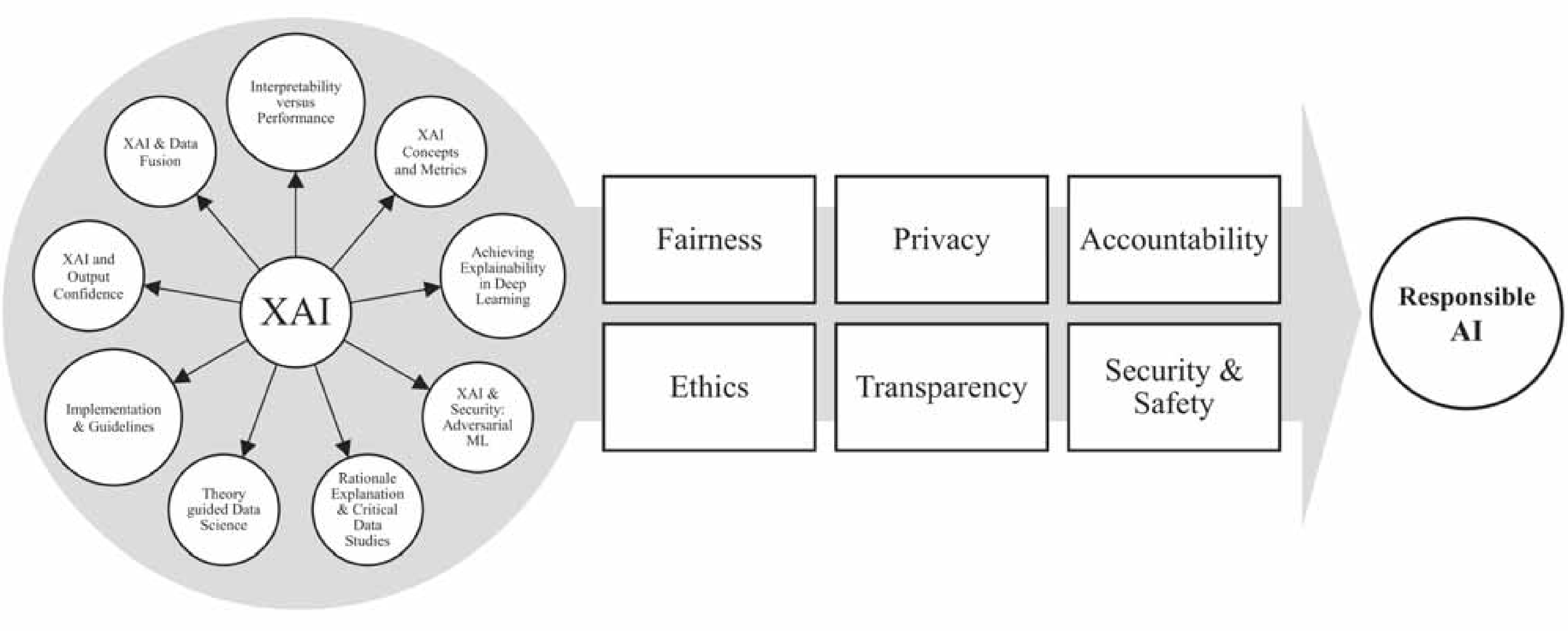
- **Causality** - inferring causal relationships from pure observational data has been extensively studied (Pearl, 2009), however it relies strongly on prior knowledge
- **Transferability** – humans have a much higher capacity to generalize, and can transfer learned skills to completely new situations; compare this with e.g. susceptibility of CNNs to adversarial data (please remember that we rarely have iid data in real world
- **Informativeness** - for example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision
- **Fairness and Ethical decision making** – interpretations for the purpose of assessing whether decisions produced automatically by algorithms conform to ethical standards
- **Trust AI** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?

Zachary C. Lipton 2016. The mythos of model interpretability. arXiv:1606.03490.

- **Interpretable Glass-Box Models,** the model itself is already interpretable, e.g.
  - Regression
  - Naïve Bayes
  - Random Forests
  - Decision Trees/Graphs
  - …
- **Interpreting Black-Box Models** (the model is not interpretable and needs a post-hoc interpretability method, e.g.:
  - Decomposition
  - LIME/BETA
  - LRP
  - …

- Rule-Based Models (e.g. decision trees):
  - Easy to interpret, the rules provide clear explanations
  - Can learn even from little data sets
  - Problems with high-dimensional data, with noise, and with images (ambiguity)
- Neuro-Symbolic Models (e.g. CNN):
  - Not easy or even impossible to interpret
  - Needs a lot of top-quality training data
  - Can well generalize even from high-dimensional data, with noise and good for images

So, is xAI a good name for it?

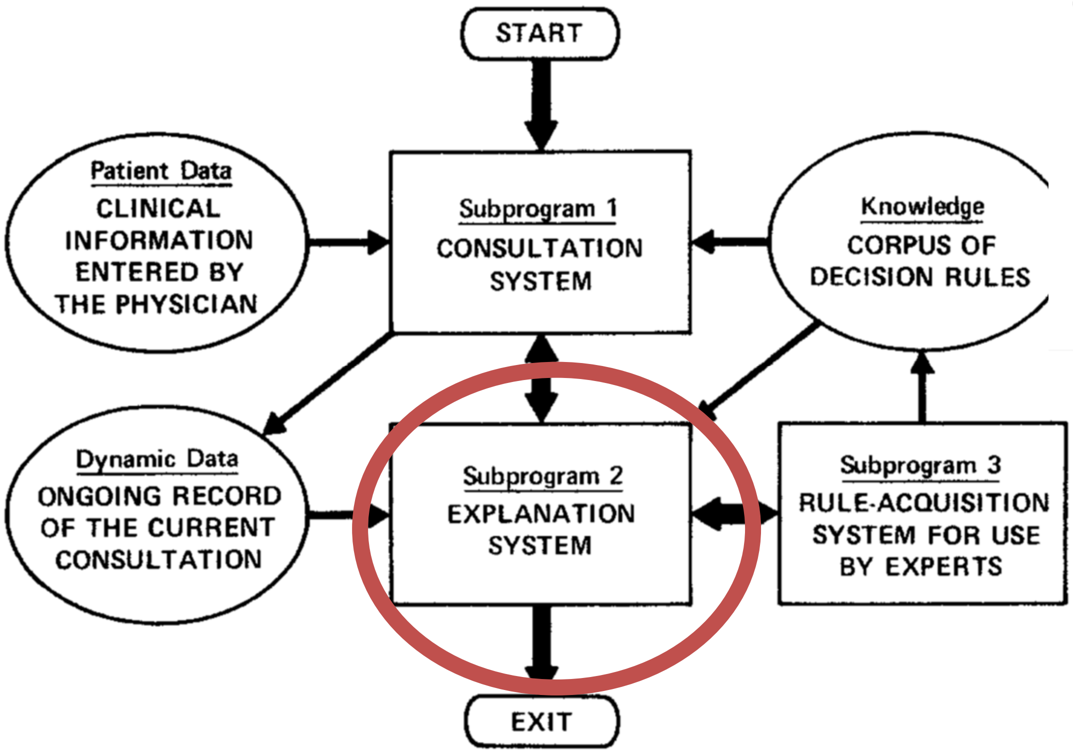


Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila & Francisco Herrera 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115, doi:10.1016/j.inffus.2019.12.012.

02 Is xAI new?

David Gunning & David W. Aha 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine,* 40, (2), 44-58.

- Explainability was the most requested feature of early medical decision support systems!



COMPUTERS AND BIOMEDICAL RESEARCH 8, 303–320 (1975)

Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System*

EDWARD H. SHORTLIFFE,† RANDALL DAVIS, STANTON G. AXLINE, BRUCE G. BUCHANAN, C. CORDELL GREEN, AND STANLEY N. COHEN

Stanford University, Stanford, California 94305

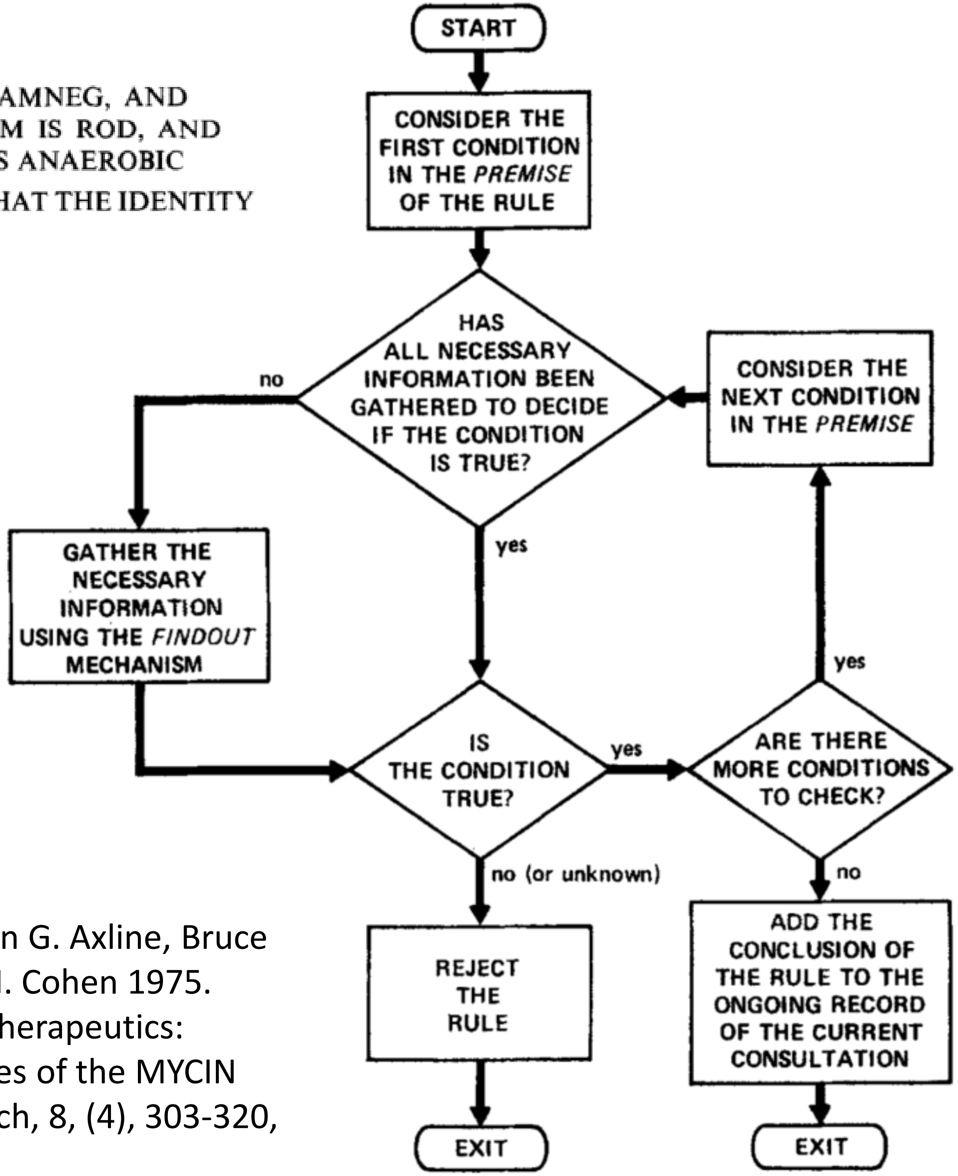
Received June 24, 1974

Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green & Stanley N. Cohen 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and biomedical research*, 8, (4), 303-320, doi:10.1016/0010-4809(75)90009-9.

How did MYCIN enable explainability ?

IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) THE AEROBICITY OF THE ORGANISM IS ANAEROBIC

THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE IDENTITY OF THE ORGANISM IS BACTEROIDES



Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green & Stanley N. Cohen 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Computers and biomedical research, 8, (4), 303-320, doi:10.1016/0010-4809(75)90009-9.

How did MYCIN support WHY questions ?

IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) THE AEROBICITY OF THE ORGANISM IS ANAEROBIC
THEN: THERE IS SUGGESTIVE EVIDENCE (.6) THAT THE IDENTITY OF THE ORGANISM IS BACTEROIDES

** *WHAT IS THE IDENTITY OF ORGANISM-1?*
THE IDENTITY OF ORGANISM-1 IS E. COLI (.30) PROTEUS (.26) PSEUDOMONAS (.24) KLEBSIELLA (.21)
** *HOW DID YOU DECIDE THAT ORGANISM-1 MIGHT BE A* PSEUDOMONAS?
I USED RULE085 TO CONCLUDE THAT THE IDENTITY OF ORGANISM-1 IS PSEUDOMONAS. THIS GAVE A CUMULATIVE CERTAINTY FACTOR OF .24.

IF: 1) THE STAIN OF THE ORGANISM IS GRAMNEG, AND
2) THE MORPHOLOGY OF THE ORGANISM IS ROD, AND
3) THE PATIENT IS A COMPROMISED HOST
THEN: THERE IS SUGGESTIVE EVIDENCE (.4) THAT THE IDENTITY OF THE ORGANISM IS PSEUDOMONAS

** *DO YOU EVER PRESCRIBE CARBENICILLIN FOR PSEUDOMONAS INFECTIONS?*

RELEVANT RULES: RULE064 RULE119 RULE120
WHICH ONES DO YOU WISH TO SEE?
** *RULE064*
IF: 1) IT IS KNOWN WITH CERTAINTY THAT THE IDENTITY OF THE ORGANISM IS PSEUDOMONAS, AND
2) THE DRUG UNDER CONSIDERATION IS GENTAMICIN
THEN: RECORD THE FOLLOWING AS A MORE APPROPRIATE THERAPY: GENTAMICIN-AND-CARBENICILLIN

Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green & Stanley N. Cohen 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Computers and biomedical research, 8, (4), 303-320, doi:10.1016/0010-4809(75)90009-9.

How was the interactivity for the doctors ensured ?

558 SHORTLIFFE ET AL.

THE QUESTION-ANSWERING MODULE

** Is either chloramphenicol or kanamycin okay for treatment of salmonella infections?

YES.

RULE083

IF:

THE IDENTITY OF THE ORGANISM IS SALMONELLA

THEN:

I RECOMMEND (IN ORDER OF PREFERENCE):

1–CHLORAMPHENICOL
2–AMPICILLIN
3–CEPHALOTHIN

** How do you know if an organism is a bacteroides?

RULE030

IF:

THE GRAMSTAIN OF THE ORGANISM IS GRAMNEG, AND THE MORPHOLOGY OF THE ORGANISM IS ROD, AND THE AEROBICITY OF THE ORGANISM IS ANAEROBIC

THEN:

CONCLUDE THAT THE IDENTITY OF THE ORGANISM IS BACTEROIDES (MODIFIER: THE CERTAINTY TALLY FOR THE PREMISE TIMES .6)

Edward H Shortliffe, Stanton G Axline, Bruce G Buchanan, Thomas C Merigan & Stanley N Cohen 1973. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Computers and Biomedical Research, 6, (6), 544-560.

- The evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical applications.
- Specific deficiencies that must be overcome include the program's inability
- to reason, its inability to construct differential diagnoses spanning multiple areas, its occasional attribution of findings to improper causes, and its inability to explain its "thinking“ …

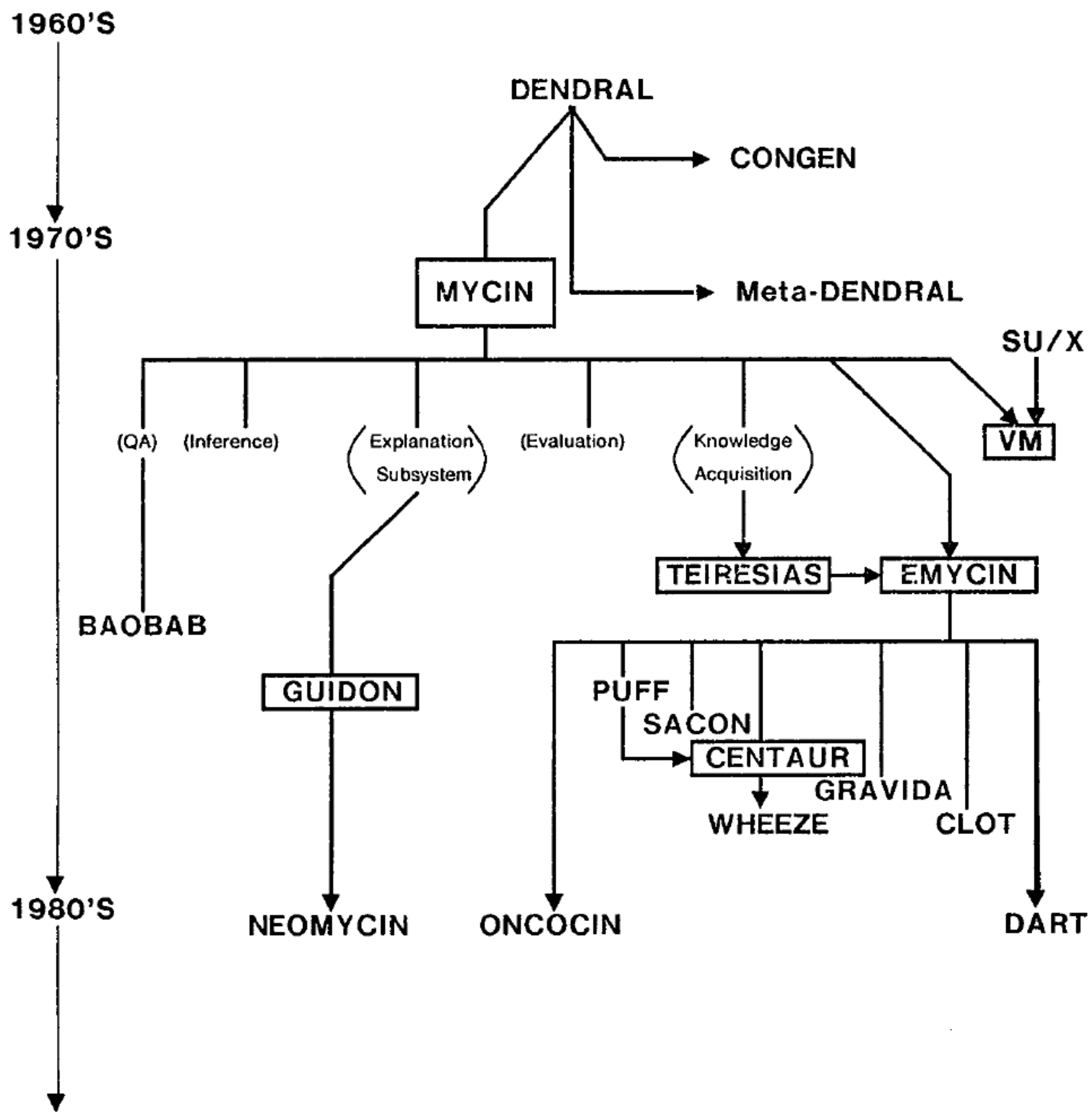
Randolph A Miller, Harry E Pople Jr & Jack D Myers 1982. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307, (8), 468-476.

Digression: History of DSS = History of AI

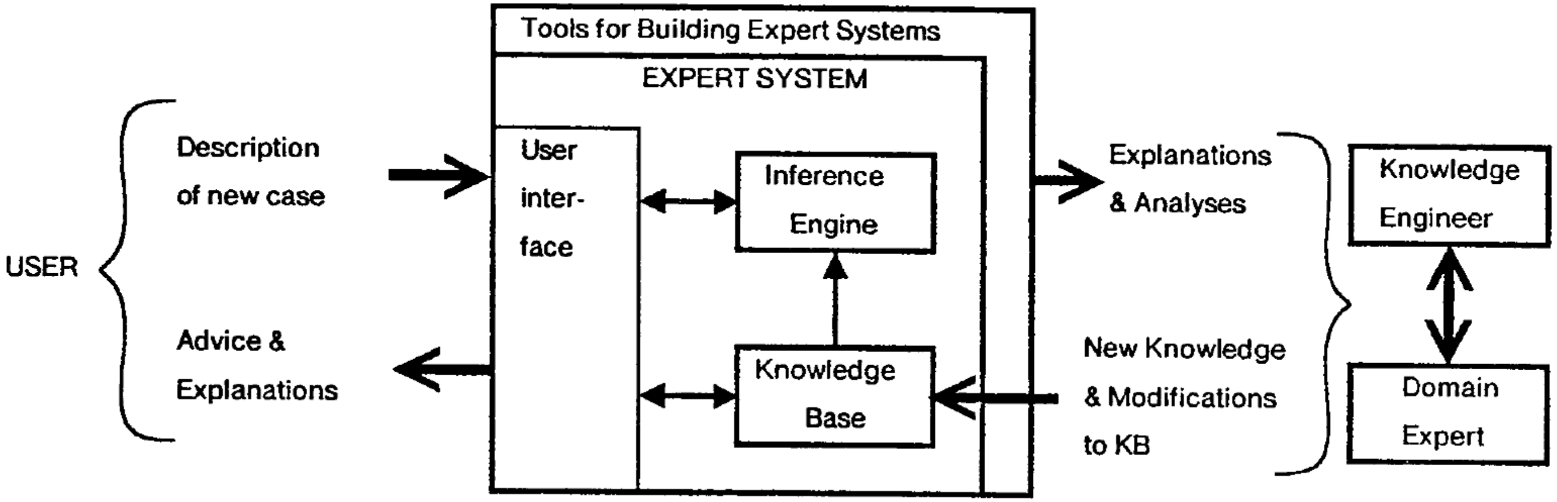
- **1943** Warren S. McCulloch & Walter Pitts: A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biology,* 5, (4), 115-133, doi:10.1007/BF02459570.
- **1950** Alan M. Turing: Computing machinery and intelligence. Mind, 59, (236), 433-460, doi:10.1093/mind/LIX.236.433
- **1959** John McCarthy: Programs with common sense. Mechanization of thought processes (Advice Taker)
- **1975** Ted Shortliffe & Bruce Buchanan: A model of inexact reasoning in medicine. Mathematical biosciences, 23, (3-4), 351-379, doi:10.1016/0025-5564(75)90047-4.
- **1978** Bellman, R. Can Computers Think? Automation of Thinking, problem solving, decision-making …

- **1986** David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams 1986. Learning representations by back-propagating errors. *Nature,* 323, (6088), 533-536, doi:10.1038/323533a0.
- **1988** Judea Pearl: Embracing causality in default reasoning. Artificial Intelligence, 35, (2), 259-271, doi:10.1016/0004-3702(88)90015-X.
- **1997** Deep Blue beats Geri Kasparov
- **2009** Successful autonomous driving
- **2011** IBM Watson in Jeopardy
- **2016** 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.

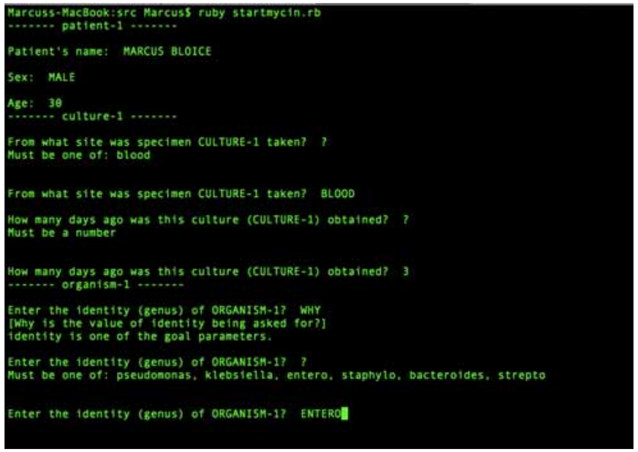
- 1960+ Medical Informatics (Classic AI Hype)
  - Focus on data acquisition, storage, accounting, Expert Systems
  - The term was first used in 1968 and the first course was set up 1978 !
- 1985+ Health Telematics (AI winter)
  - Health care networks, Telemedicine, CPOE-Systems, …
- 1995+ Web Era (AI is "forgotten")
  - Web based applications, Services, EPR, distributed systems, …
- 2005+ Success statistical learning (AI renaissance)
  - Pervasive, ubiquitous Computing, Internet of things, …
- 2010+ Data Era – Big Data (super for AI)
  - Massive increase of data – data integration, mapping, …
- 2020+ Explanation Era – (towards explainable AI)
  - Re-traceability, replicability, reenactment, explainability, interpretability, sensemaking, disentangling the underlying concepts, **causality,** causability, human-AI interfaces, ethical responsible machine learning, trust-AI…



Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*



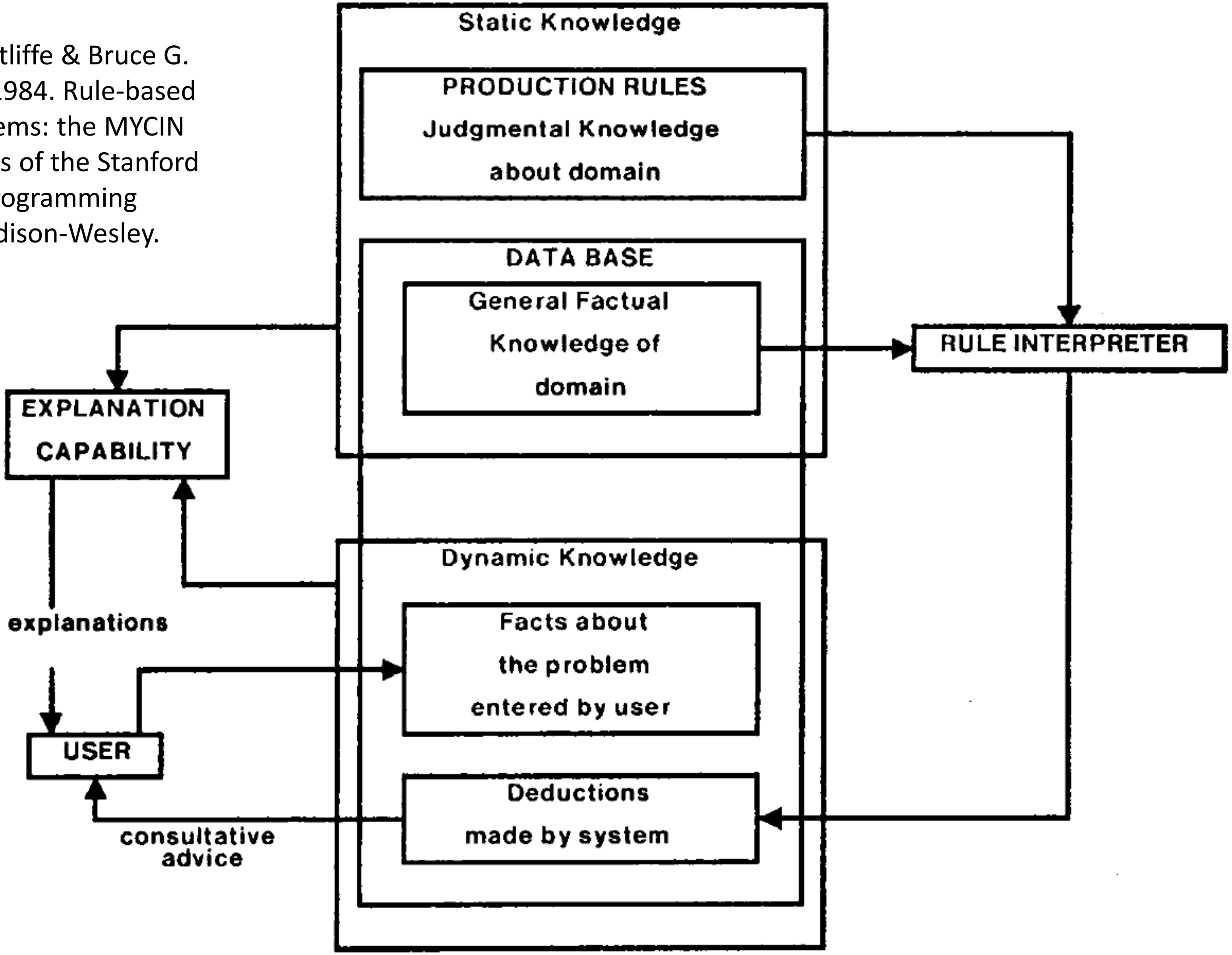
Ted Shortliffe & Randy Davis 1975. Some considerations for the implementation of knowledge-based expert systems ACM SIGART Bulletin, (55), 9-12.



Find an emulation and a Jupyter notebook here:
http://user.medunigraz.at/marcus.bloice/seminars/dss/g3/g3.htm

What was static knowledge versus dynamic knowledge ?

Ted H. Shortliffe & Bruce G. Buchanan 1984. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project, Addison-Wesley.



- The information available in medicine is often imperfect – imprecise - uncertain.
- **Human experts** can cope with deficiencies.
- Classical logic permits only **exact reasoning**:
- IF A is true THEN A is non-false and
  IF B is false THEN B is non-true
- Most real-world problems do not provide this exact information, mostly it is inexact, incomplete, uncertain and/or **un-measurable!**

- To every rule and every entry a certainty factor (CF) is assigned, which is between 0 und 1
- Two measures are derived:
- MB: measure of belief
- MD: measure of disbelief
- Certainty factor – CF of an element is calculated by:

$$CF[h] = MB[h] - MD[h]$$

- CF is positive, if more evidence is given for a hypothesis, otherwise CF is negative
- CF[h] = +1 -> h is 100 % true
- CF[h] = –1 -> h is 100% false

h_1 = The identity of ORGANISM-1 is streptococcus
h_2 = PATIENT-1 is febrile
h_3 = The name of PATIENT-1 is John Jones

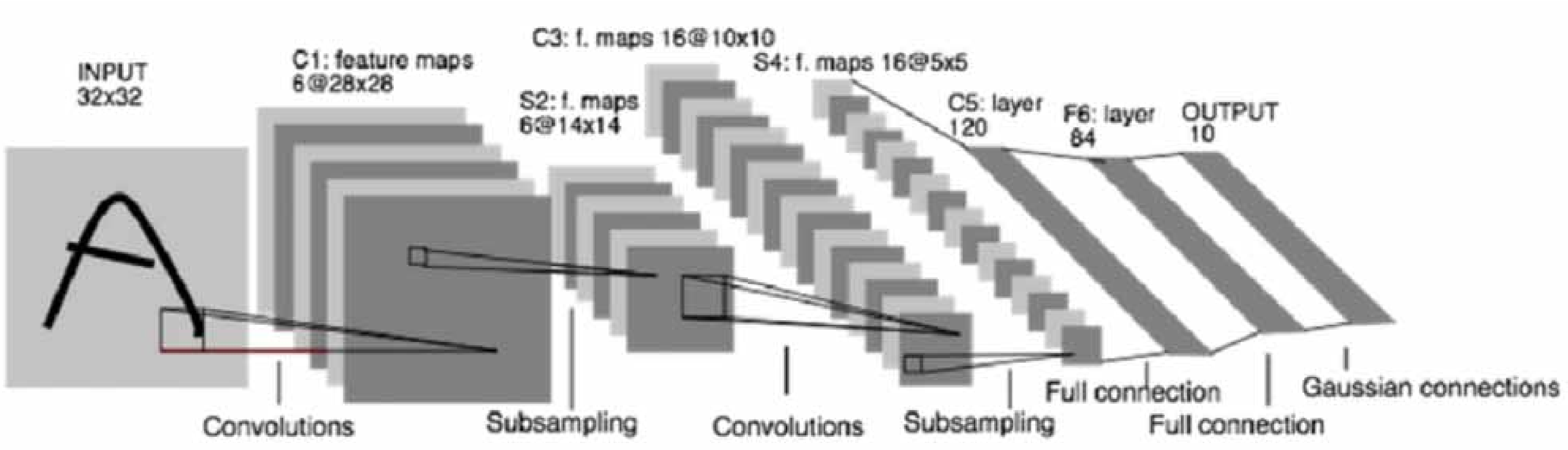
$CF[h_1,E] = .8$: There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus

$CF[h_2,E] = -.3$: There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile

$CF[h_3,E] = +1$: It is definite (1) that the name of PATIENT-1 is John Jones

Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*


YouTube DE
real nurse triage
Real Triage Nurse
Jon Bell
Subscribe 35
61,434 views
Add to
Share
More
77
12



Yann LeCun, Leon Bottou, Yoshua Bengio & Patrick Haffner 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86, (11), 2278-2324.

Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In: Pereira, Fernando, Burges, Christopher J.C., Bottou, Leon & Weinberger, Kilian Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. NIPS, 1097-1105.

- Won the 2012 ImageNet LSVRC. 60 Million parameters, 832M MAC ops

4M	FULL CONNECT	4Mflop
16M	FULL 4096/ReLU	16M
37M	FULL 4096/ReLU	37M
	MAX POOLING	
442K	CONV 3x3/ReLU 256fm	74M
1.3M	CONV 3x3ReLU 384fm	224M
884K	CONV 3x3/ReLU 384fm	149M
	MAX POOLING 2x2sub	
	LOCAL CONTRAST NORM	
307K	CONV 11x11/ReLU 256fm	223M
	MAX POOL 2x2sub	
	LOCAL CONTRAST NORM	
35K	CONV 11x11/ReLU 96fm	105M

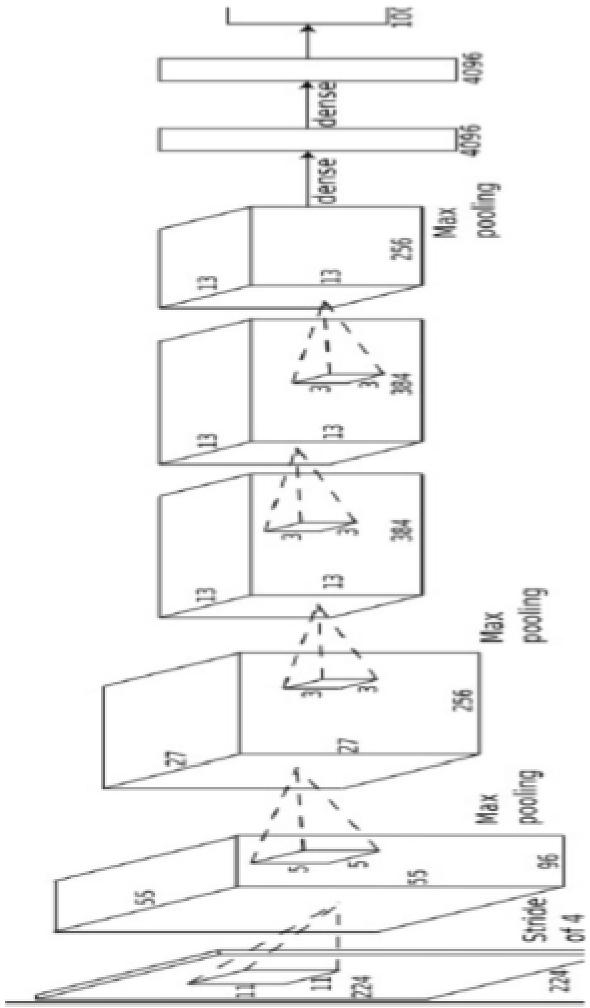
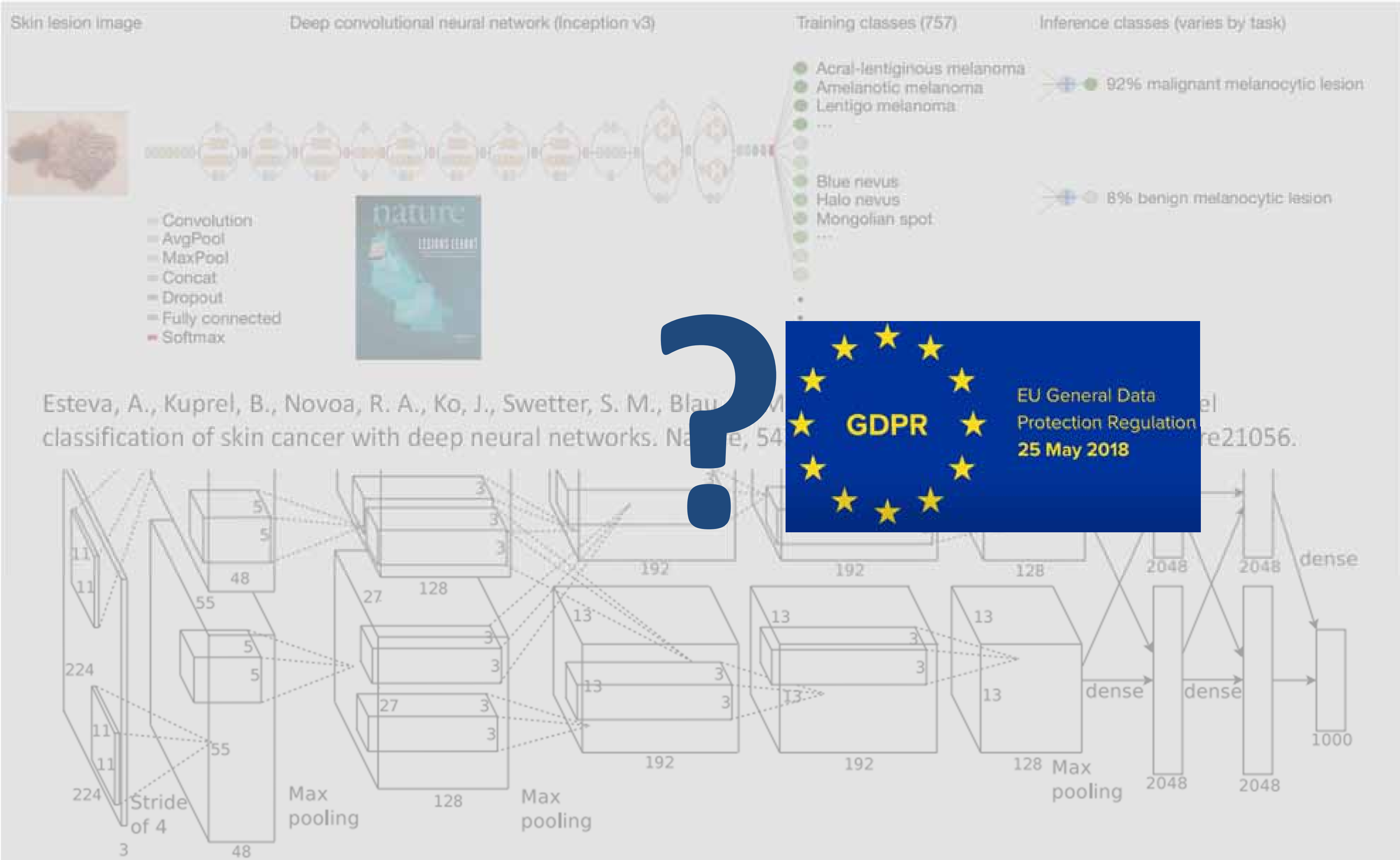


Image credit to Yann LeCun, ICML 2013 Deep Learning Tutorial



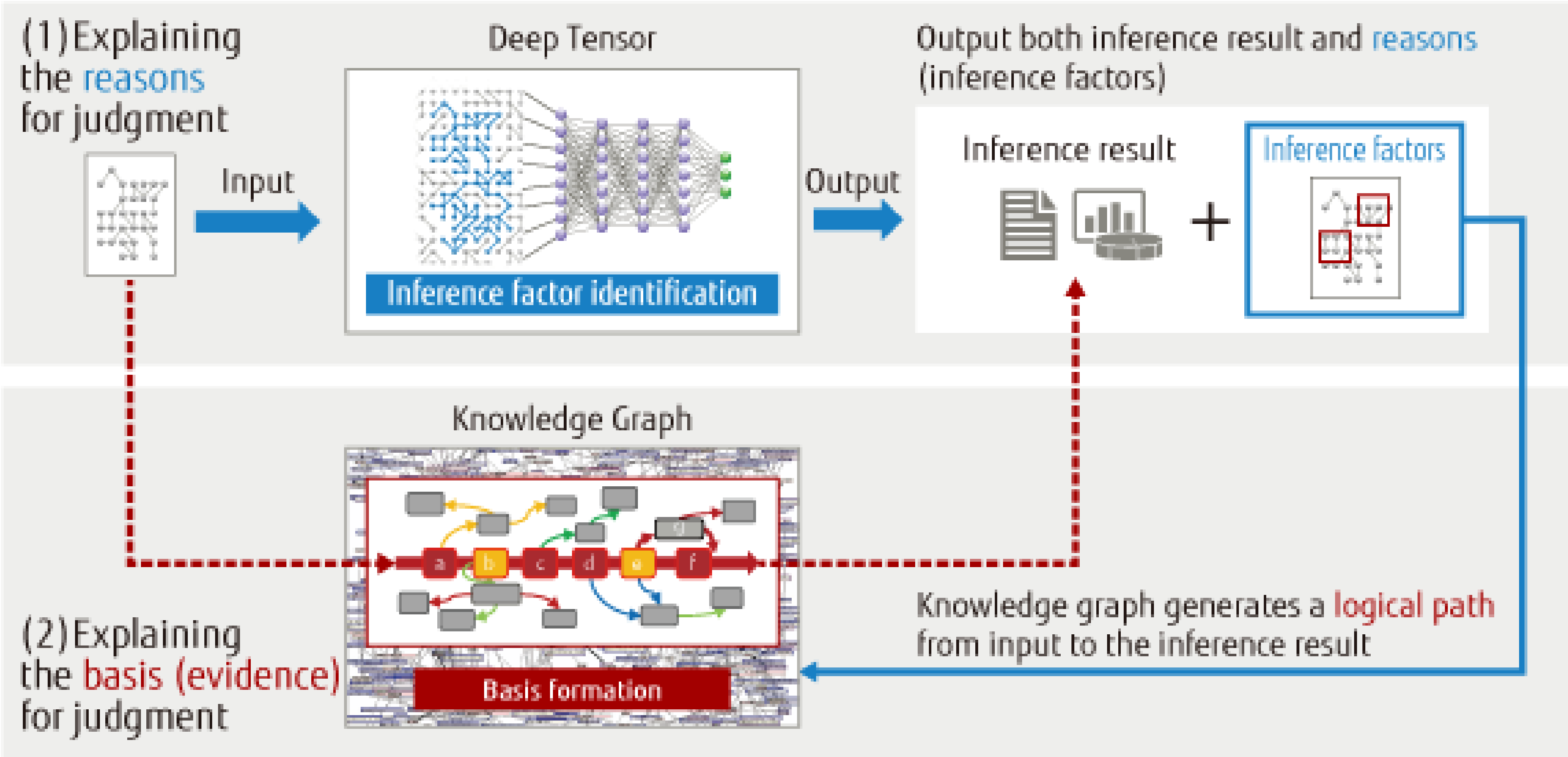
Skin lesion image
Deep convolutional neural network (Inception v3)
Training classes (757)
Inference classes (varies by task)
Acral-lentiginous melanoma
Amelanotic melanoma
Lentigo melanoma
Blue nevus
Halo nevus
Mongolian spot
92% malignant melanocytic lesion
8% benign melanocytic lesion
Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax
Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M.,
classification of skin cancer with deep neural networks.
Stride of 4
Max pooling
dense
GDPR
EU General Data Protection Regulation
25 May 2018

- Success in deep learning *) resulted in “deep problems” (e.g. complex and exploding gradients)
- *) Note: “DL” methods are representation learning methods with multiple layers of representations (see LeCun, Bengio & Hinton (2015), Nature 521, 7553)
- Problem in our society: “Secret algorithms” make important decisions about individuals
- **Black box Type 1** = too complicated for a human to understand
- **Black box Type 2** = proprietary = “secret algorithm”

Cynthia Rudin, Caroline Wang & Beau Coker 2018. The age of secrecy and unfairness in recidivism prediction. arXiv:1811.00731.

- **Post-Hoc** (latin) = after- this (event), i.e. such approaches provide an explanation for a specific solution of a "black-box" approach, e.g. LIME, BETA, LRP, …
- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis & Douglas B. Kell 2017. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923*.



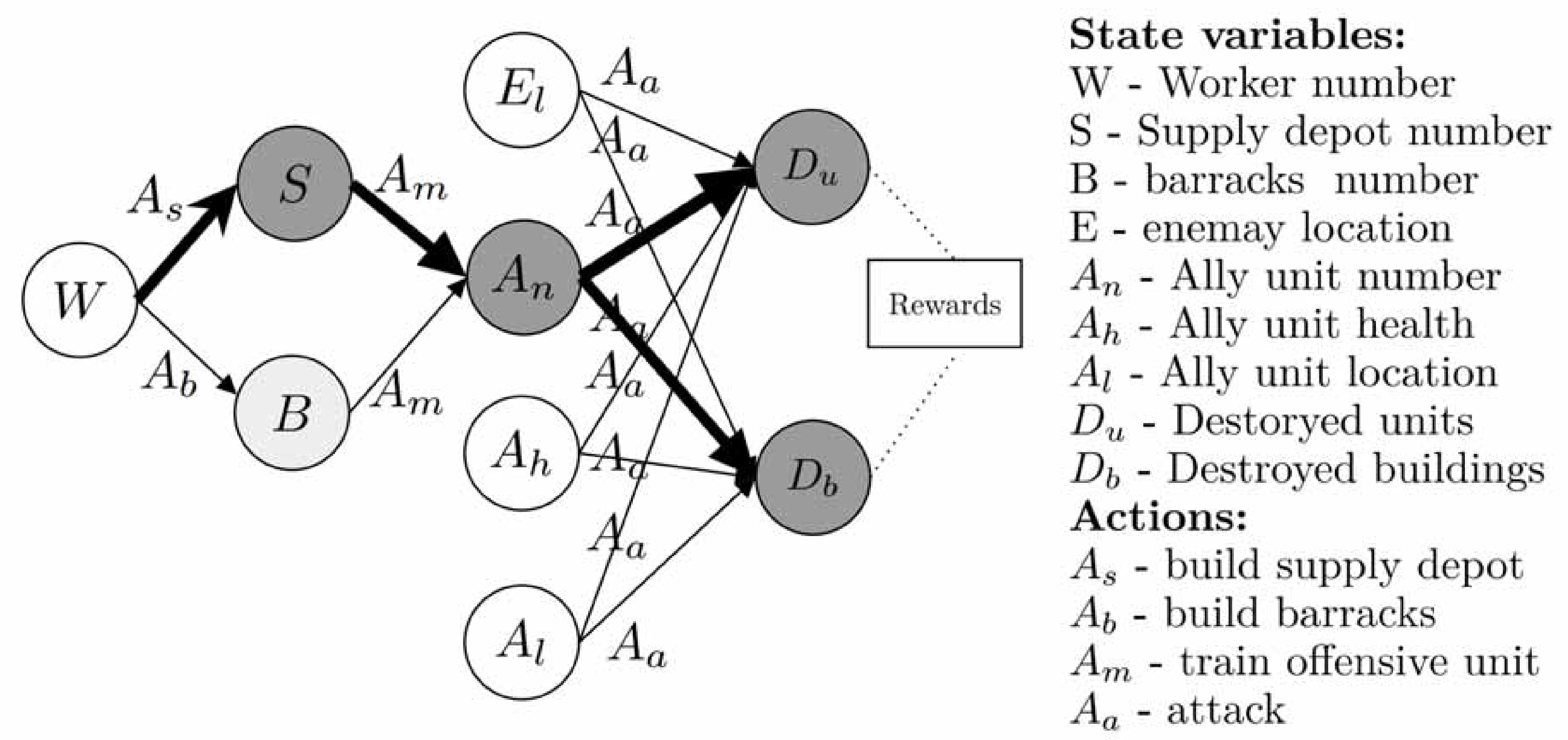
Explainable AI with Deep Tensor and Knowledge Graph

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

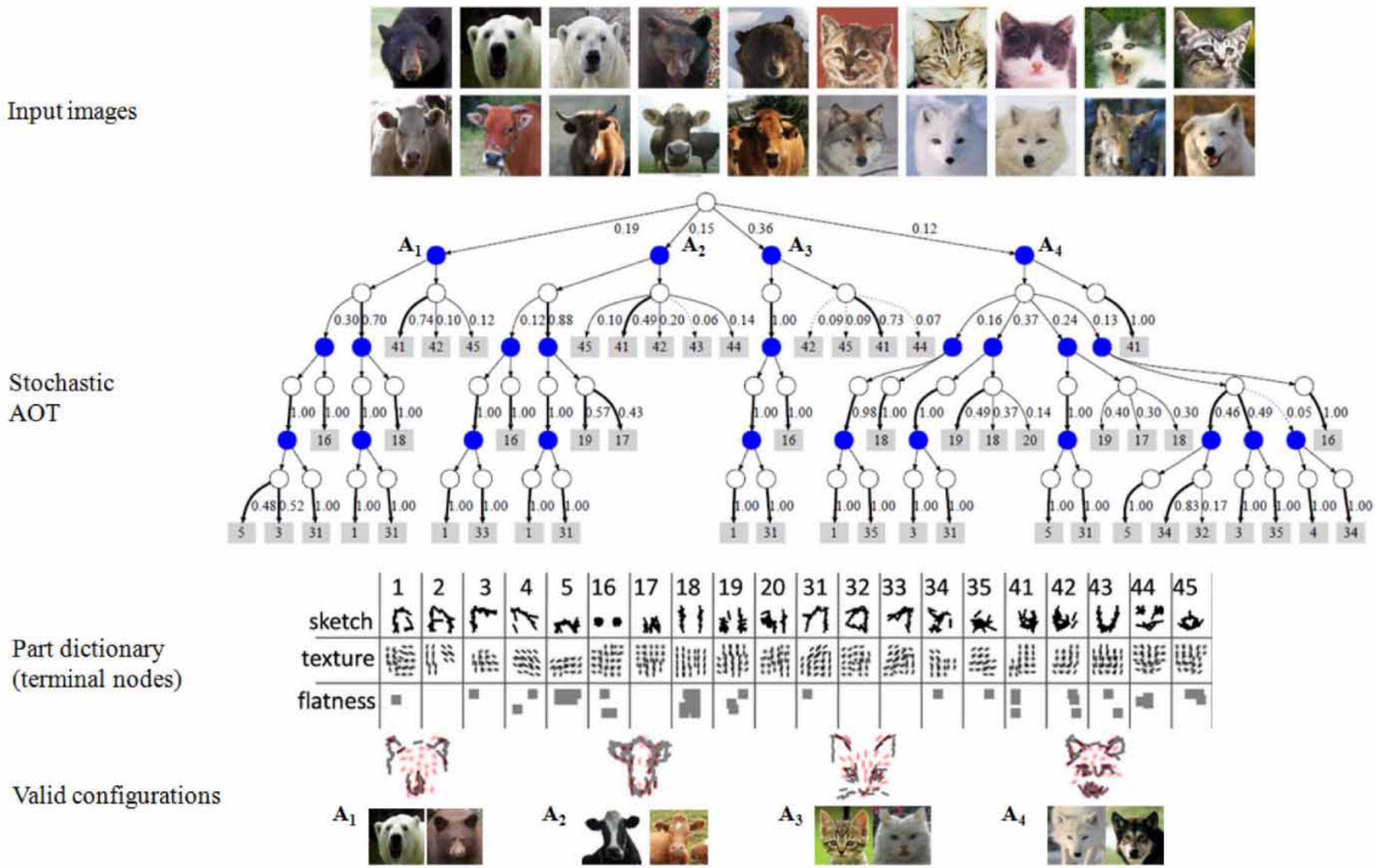
03 Examples for Ante Hoc Models (interpretable Machine Learning)

- **Post-Hoc** (latin) = after- this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, … (see module 5)
- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.
- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems; typical methods decision trees, linear regression, Random Forests, ...

Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis & Douglas B. Kell 2017. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923*.



Prashan Madumal, Tim Miller, Liz Sonenberg & Frank Vetere 2019. Explainable Reinforcement Learning Through a Causal Lens. arXiv:1905.10958

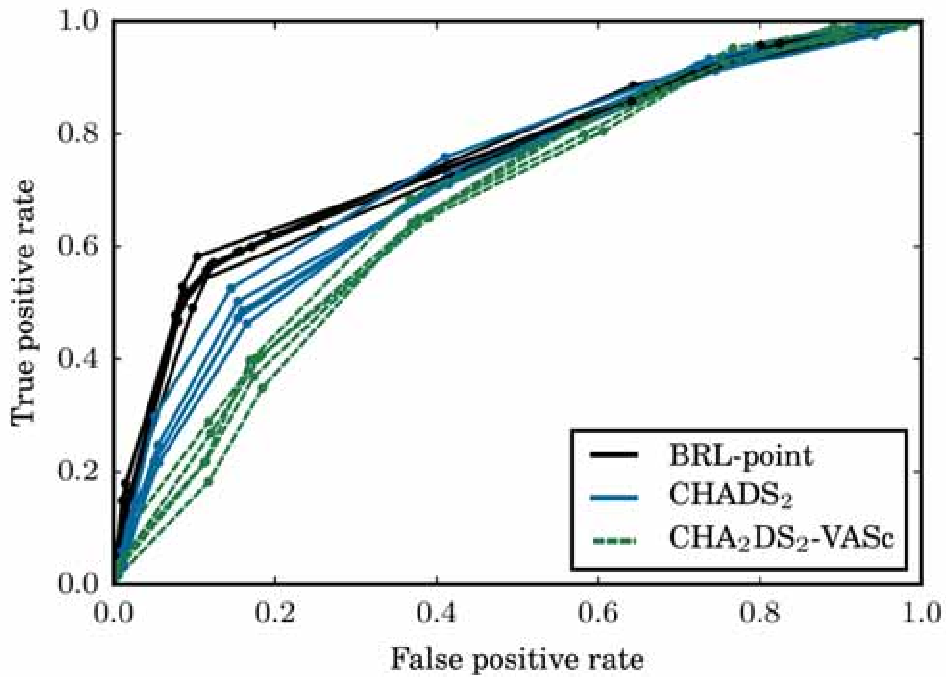


Zhangzhang Si & Song-Chun Zhu 2013. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 35, (9), 2189-2205, doi:10.1109/TPAMI.2013.35.

Example: Bayesian Rule Lists

if hemiplegia **and** age > 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age > 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

	BRL	C5.0	CART	ℓ_1-LR	SVM	RF	BCART
Mean accuracy	1.00	0.94	0.90	0.98	0.99	0.99	0.71
Standard deviation	0.00	0.01	0.04	0.01	0.01	0.01	0.04



Benjamin Letham, Cynthia Rudin, Tyler H McCormick & David Madigan 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics, 9, (3), 1350-1371, doi:10.1214/15-AOAS848.

04 Examples for Post Hoc Models (e.g. LIME, BETA, LRP)

- **Post-Hoc** (latin) = after- this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, …
- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.
- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems (see module 3); typical methods decision trees, linear regression, and Random Forests.

Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis & Douglas B. Kell 2017. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923*.

Caveat – Post hoc explanation can be misleading !

Cynthia Rudin 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, (5), 206-215, doi:10.1038/s42256-019-0048-x.

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

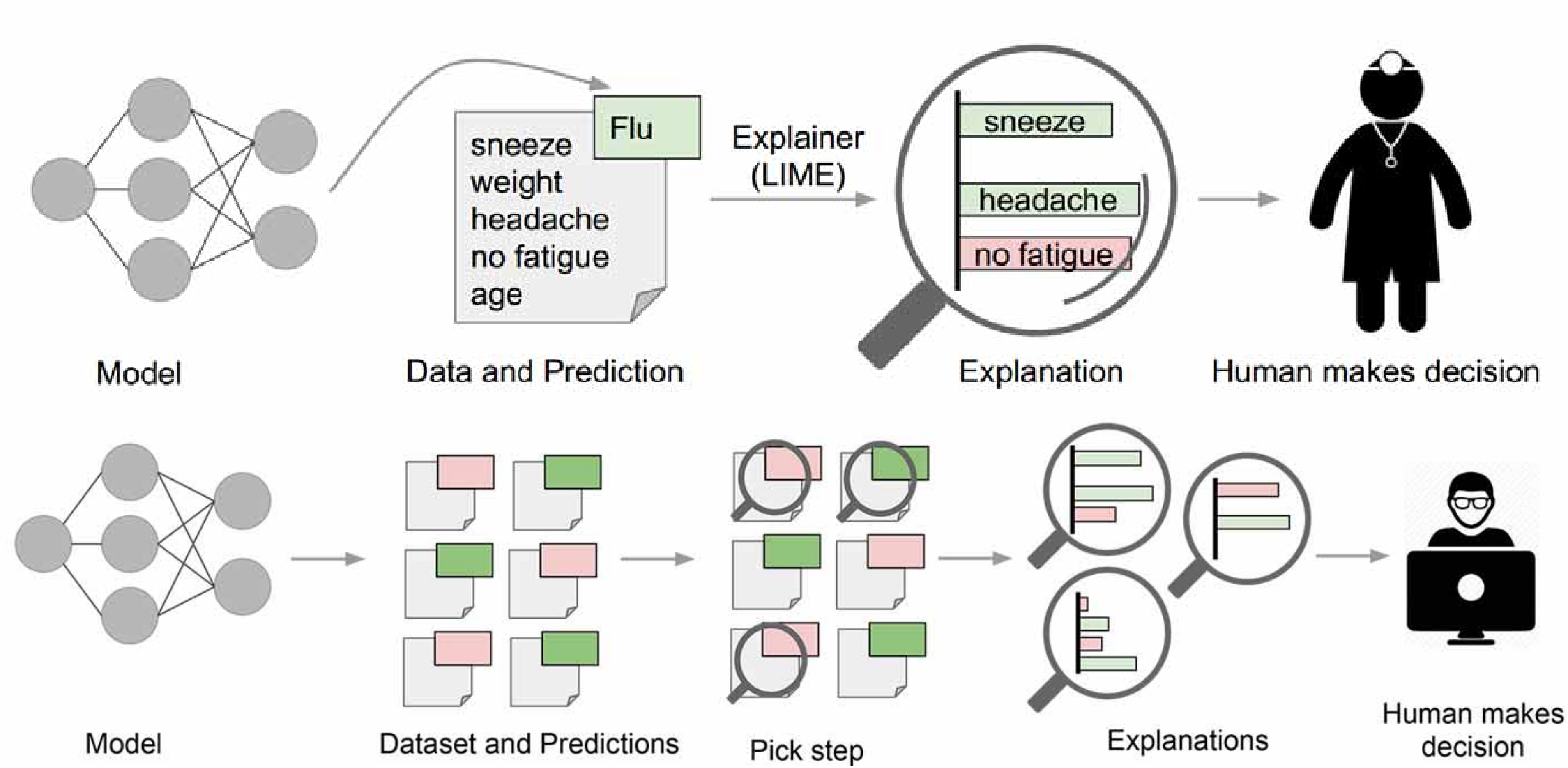
Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

- 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://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

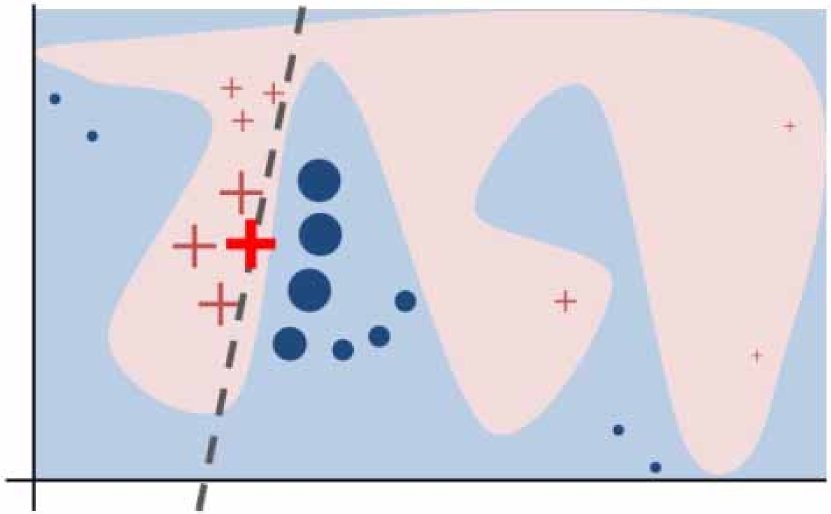
04a LIME – Local Interpretable Model Agnostic Explanations

What is the general principle of LIME ?



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.

- Explanation := local linear approximation of the model's behaviour. While the model may be very complex globally, it is easier to approximate it around the vicinity of a particular instance.



Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with z'_i as features, $f(z)$ as target
return w

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. https://github.com/marcotcr/lime

$$\xi(x) = \underset{g \in G}{\arg min} \; L(f, g, \pi_x) + \Omega(g)$$

Fidelity score (for local fidelity)

Complexity score (for interpretability)

$\pi_x(z)$ Distance metric (in feature space!)

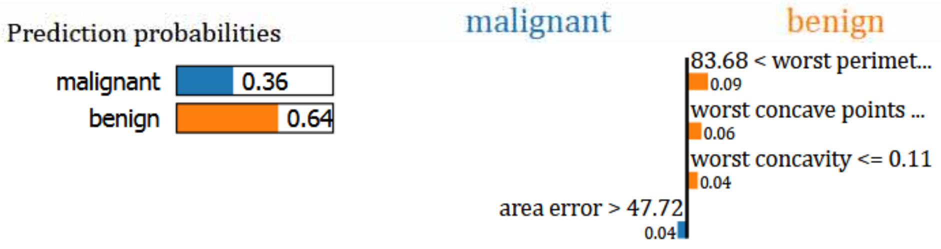
How does a typical LIME example look like ?

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)
```

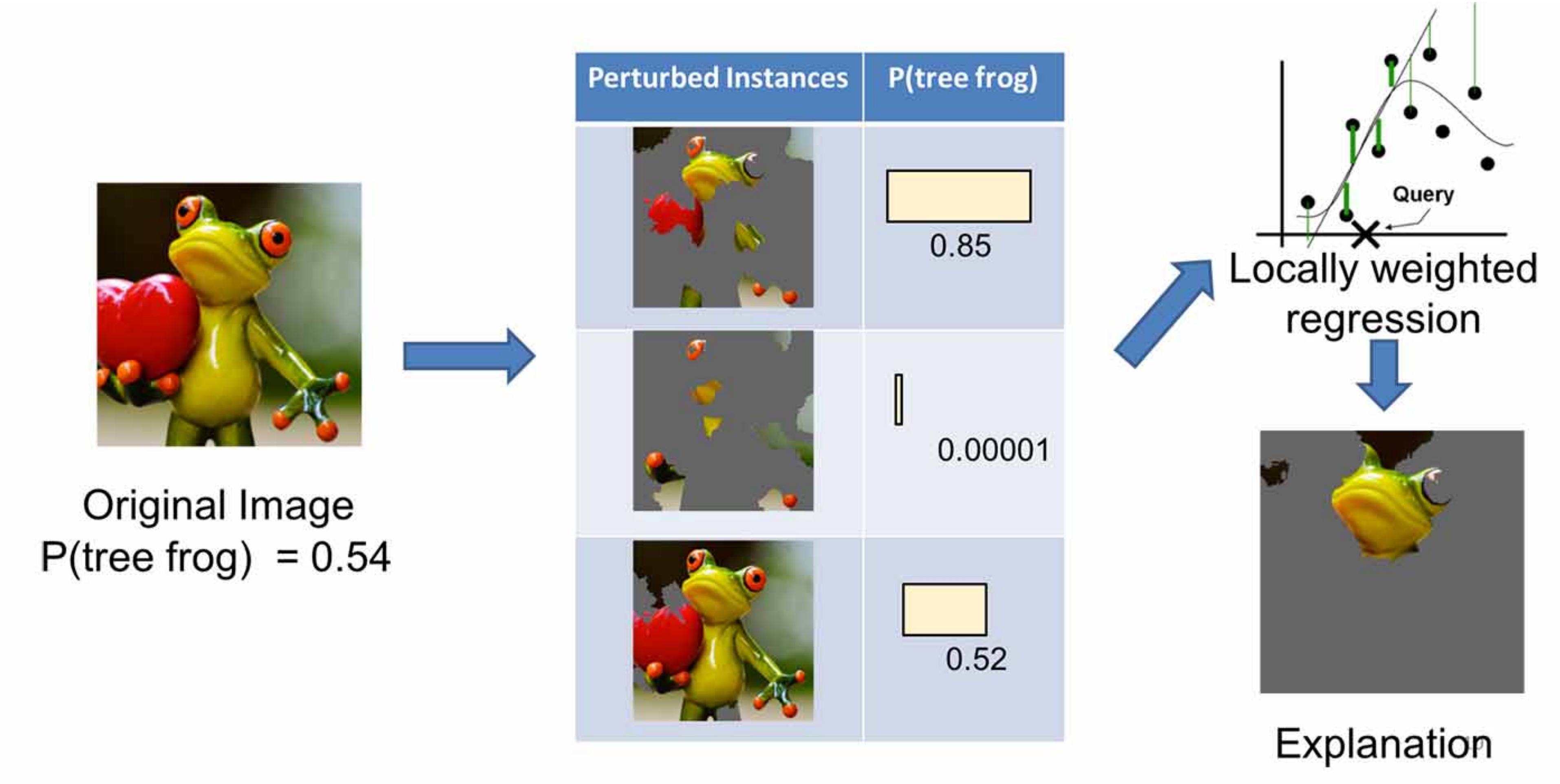


Feature	Value
worst perimeter	95.54
worst concave points	0.03
worst concavity	0.02
area error	106.00

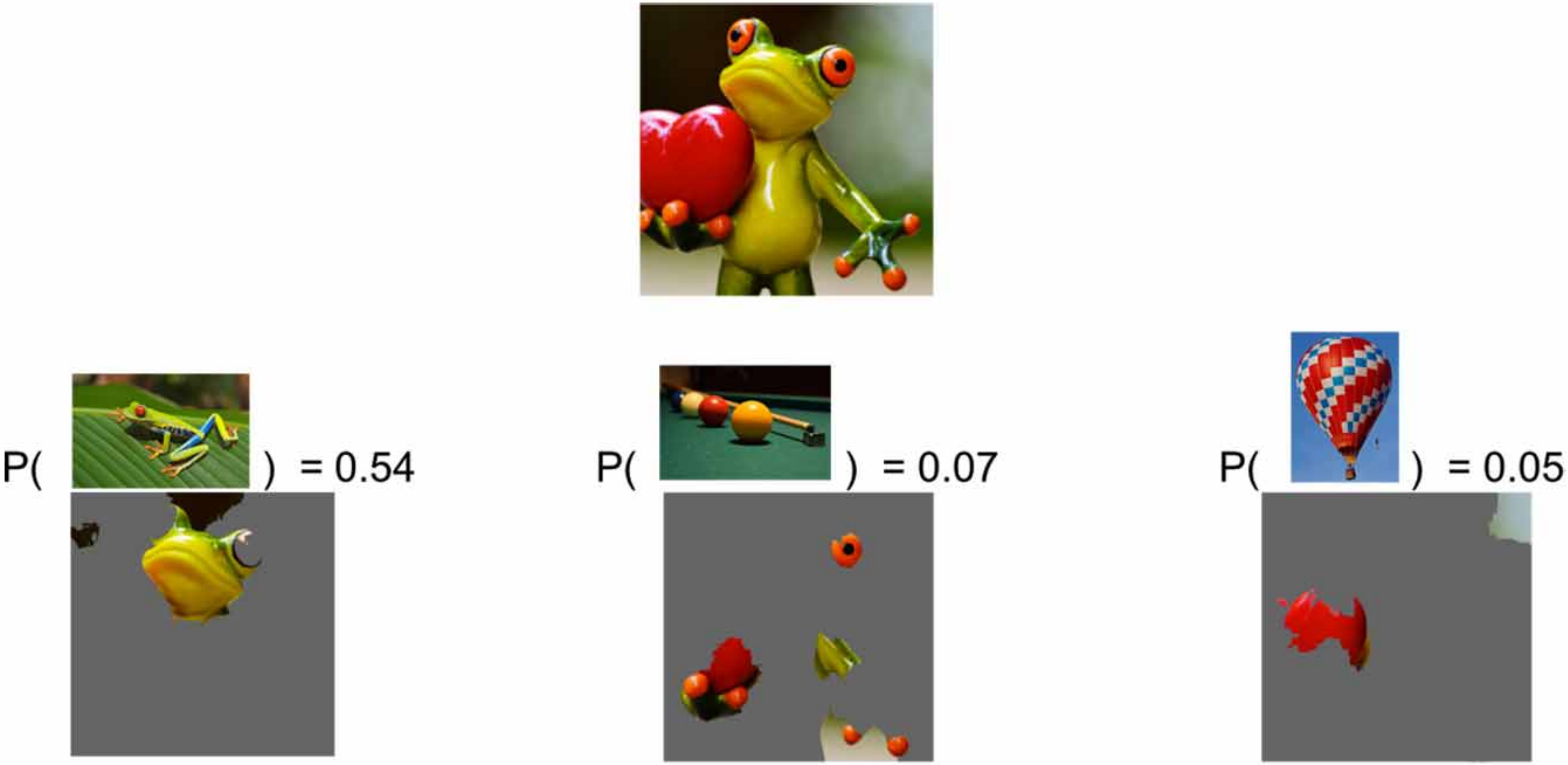
As you can see, the random forest algorithm has predicted with a probability of 0.64 that the sample at index 76 in the test set is malignant.

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):



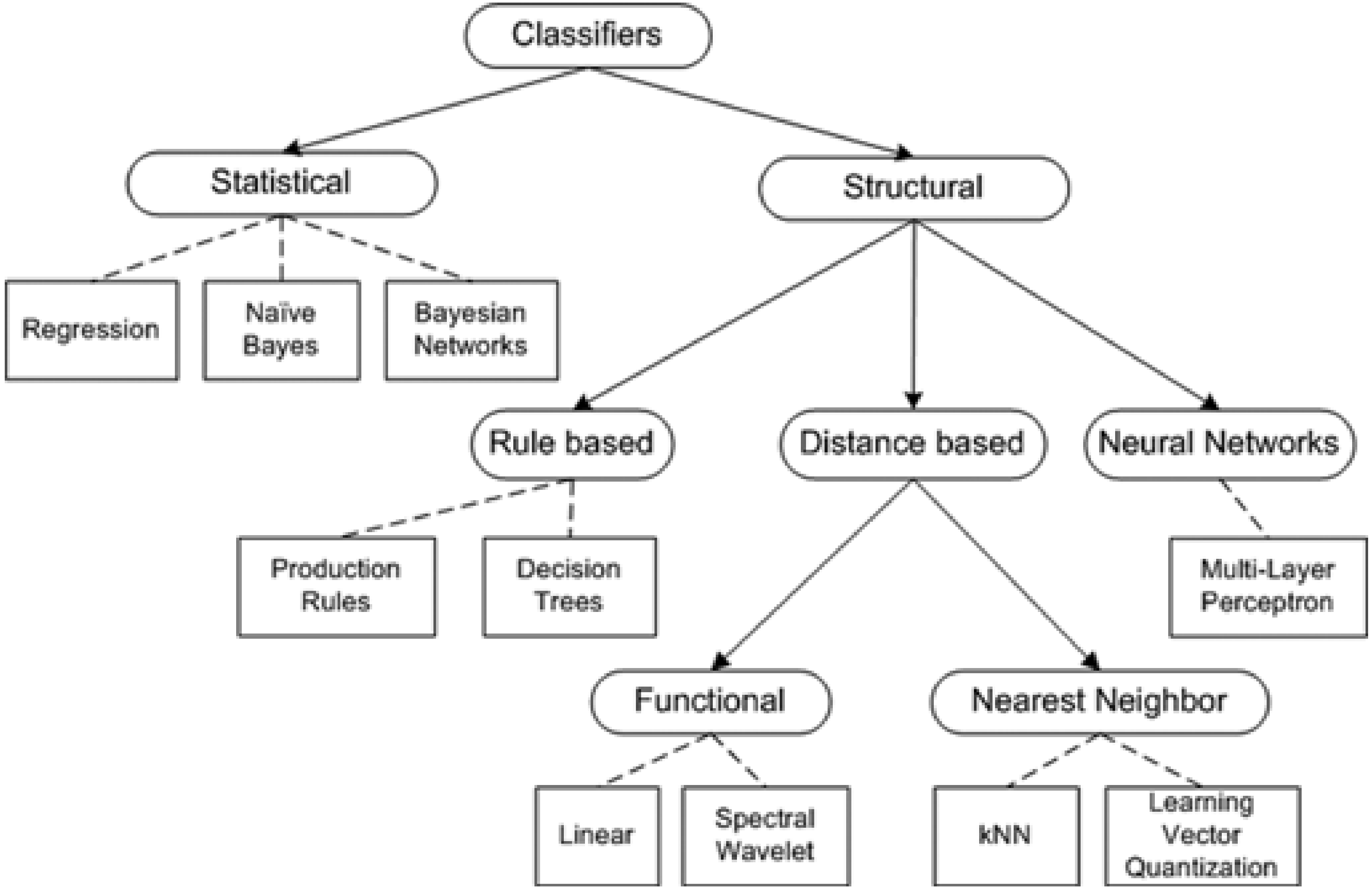
https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime



https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

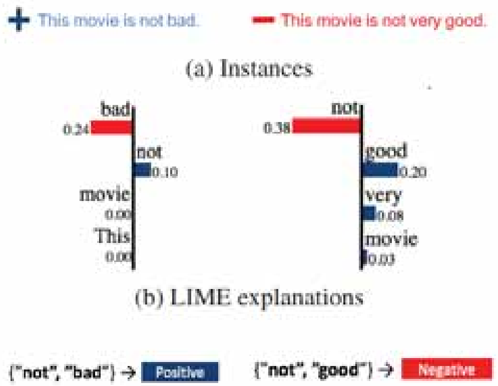
- + very popular,
- + many applications and contributors
- + model agnostic

- - local model behaviour can be unrealistic
- - unclear coverage
- - ambiguity (how to select the kernel width ?)

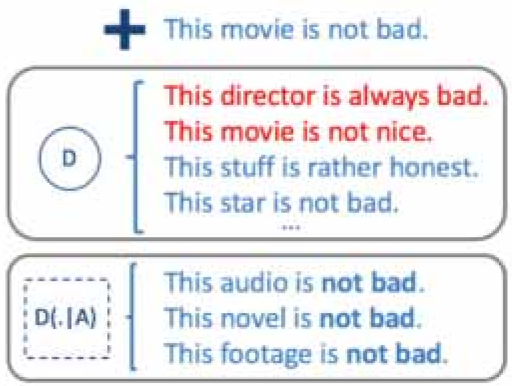


https://stats.stackexchange.com/questions/271247/machine-learning-statistical-vs-structural-classifiers

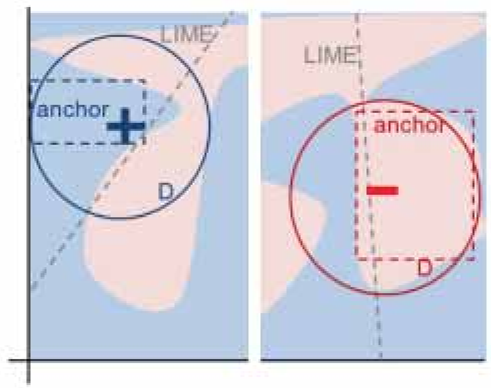
What is the difference to the Follow-up: Anchor



(c) Anchor explanations



(a) $\mathcal{D}$ and $\mathcal{D}(.|A)$



(b) Two toy visualizations

(a) Original image (b) Anchor for "beagle" (c) Images where Inception predicts $P(\text{beagle}) > 90\%$

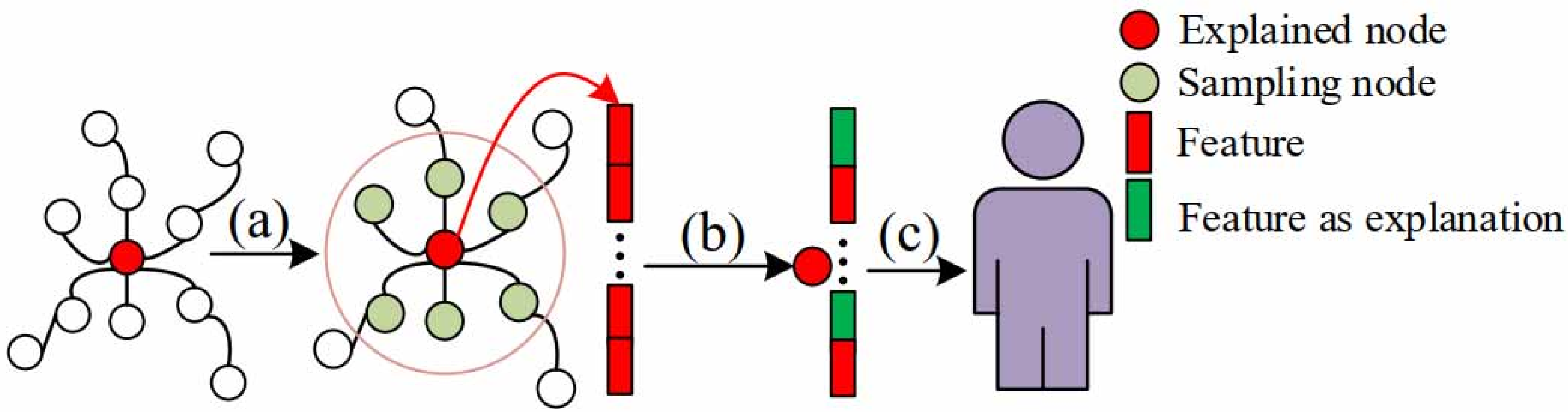
What animal **is featured in this picture ?**	**dog**
What floor is featured in this picture?	dog
What toenail is paired in this flowchart ?	dog
What animal is shown on this depiction ?	dog

(d) **VQA:** Anchor (bold) and samples from $\mathcal{D}(z|A)$

Where is the **dog**?	on the floor
What color is the **wall?**	white
When was this picture taken?	during the day
Why is he lifting his paw?	to play

(e) **VQA:** More example anchors (in bold)

Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Anchors: High-precision model-agnostic explanations. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018 New Orleans. Association for the Advancement of Artificial Intelligence, 1527-1535.



$$\zeta(v) = \operatorname*{argmin}_{g\in G} g(f, \boldsymbol{X}_n)$$

Algorithm 1 Locally nonlinear Explanation: GraphLIME

Input: GNN classifier f, Number of explanation features K
Input: the graph $\mathcal{G}$, the node $\boldsymbol{x}$ being explained
Output: K explanation features

$\boldsymbol{X}_n = N\_hop\_neighbor\_sample(x)$
$\boldsymbol{Z} \leftarrow \{\}$
for all $x_i \in \boldsymbol{X}_n$ **do**
  $y_i = f(\boldsymbol{x}_i)$
  $\boldsymbol{Z} \leftarrow \boldsymbol{Z} \cup (\boldsymbol{x}_i, y_i)$
end for
$\boldsymbol{\beta} \leftarrow$ HSIC Lasso$(\boldsymbol{Z})$ $\triangleright$ with x_i as features, y_i as label
Select top-K features as explanations based on $\boldsymbol{\beta}$

Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin & Yi Chang 2020. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. arxiV 2001.06216v1.

04b BETA (Black Box Explanation through Transparent Approximation)

- BETA is a model agnostic approach to explain the behaviour of an (arbitrary) black box classifier (i.e. a function that maps a feature space to a set of classes) by simultaneously optimizing the accuracy of the original model and interpretability of the explanation for a human.
- Note: Interpretability and accuracy at the same time are difficult to achieve.
- Consequently, users are interactively integrated into the model and can thus explore the areas of black box models that interest them (most).

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 ≥ 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.

$$\underset{\mathcal{R} \subseteq \mathcal{ND} \times \mathcal{DL} \times \mathcal{C}}{\arg\max} \sum_{i=1}^{5} \lambda_i f_i(\mathcal{R})$$

$$\text{s.t. } size(\mathcal{R}) \leq \epsilon_1,\ maxwidth(\mathcal{R}) \leq \epsilon_2,\ numdsets(\mathcal{R}) \leq \epsilon_3$$

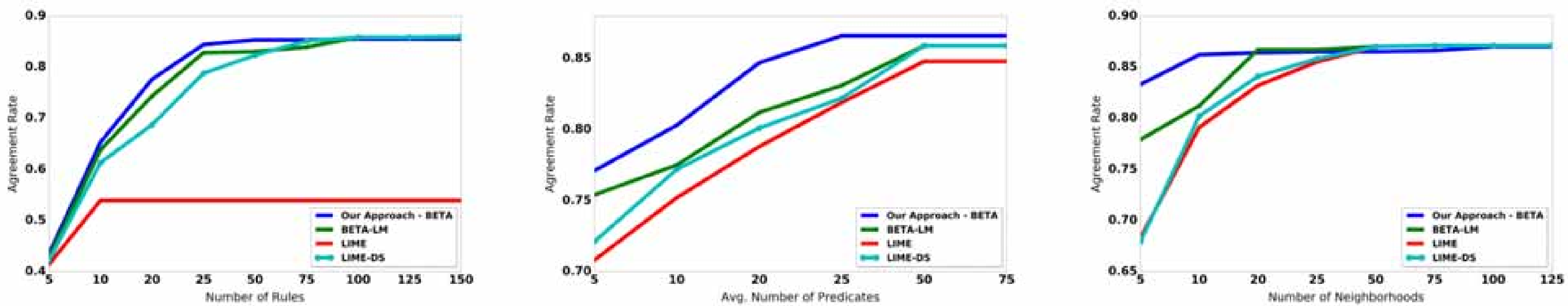
Algorithm 1 Optimization Procedure [5]

Input: Objective f, domain $\mathcal{ND} \times \mathcal{DL} \times \mathcal{C}$, parameter δ, number of constraints k
$V_1 = \mathcal{ND} \times \mathcal{DL} \times \mathcal{C}$
for $i \in \{1, 2 \cdots k+1\}$ **do** ▷ Approximation local search procedure
  $X = V_i$; $n = |X|$; $S_i = \emptyset$
  Let v be the element with the maximum value for f and set $S_i = v$
  while there exists a delete/update operation which increases the value of S_i by a factor of at least $(1 + \frac{\delta}{n^4})$ **do**
    Delete Operation: If $e \in S_i$ such that $f(S_i \backslash \{e\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \backslash e$

    Exchange Operation If $d \in X \backslash S_i$ and $e_j \in S_i$ (for $1 \leq j \leq k$) such that
    $(S_i \backslash e_j) \cup \{d\}$ (for $1 \leq j \leq k$) satisfies all the k constraints and
    $f(S_i \backslash \{e_1, e_2 \cdots e_k\} \cup \{d\}) \geq (1 + \frac{\delta}{n^4}) f(S_i)$, then $S_i = S_i \backslash \{e_1, e_2, \cdots e_k\} \cup \{d\}$
  end while
  $V_{i+1} = V_i \backslash S_i$
end for
return the solution corresponding to $\max\{f(S_1), f(S_2), \cdots f(S_{k+1})\}$

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

Fidelity	$disagreement(\mathcal{R}) = \sum_{i=1}^{M} \lvert\{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{D}, \boldsymbol{x} \text{ satisfies } q_i \wedge s_i, \mathcal{B}(\boldsymbol{x}) \neq c_i\}\rvert$
Unambiguity	$ruleoverlap(\mathcal{R}) = \sum_{i=1}^{M} \sum_{j=1, i \neq j}^{M} overlap(q_i \wedge s_i, q_j \wedge s_j)$ $cover(\mathcal{R}) = \lvert\{x \mid x \in \mathcal{D}, x \text{ satisfies } q_i \wedge s_i \text{ where } i \in \{1 \cdots M\}\}\rvert$
Interpretability	$size(\mathcal{R})$: number of rules (triples of the form (q, s, c)) in $\mathcal{R}$ $maxwidth(\mathcal{R}) = \max_{e \in \bigcup_{i=1}^{M} (q_i \cup s_i)} width(e)$ $numpreds(\mathcal{R}) = \sum_{i=1}^{M} width(s_i) + width(q_i)$ $numdsets(\mathcal{R}) = \lvert dset(\mathcal{R})\rvert$ where $dset(\mathcal{R}) = \bigcup_{i=1}^{M} q_i$ $featureoverlap(\mathcal{R}) = \sum_{q \in dset(\mathcal{R})} \sum_{i=1}^{M} featureoverlap(q, s_i)$



If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

---

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

Else if Risk-Depression=Yes **then** Depression

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

Else if Headaches=Yes **and** Dizziness=Yes, **then** Depression

Else if Doctor-Visits$\geq$ 0.3 **then** Diabetes

Else if Disposition-Tiredness=Yes **then** Depression

Else Diabetes

Notation	Definition	Term
$\mathcal{D}$	Input set of data points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$	Dataset
$\mathbf{x}$	Observed attribute values of a data point	
y	Class label of a data point	
$\mathcal{C}$	Set of class labels in $\mathcal{D}$	
p	(attribute, operator, value) tuple, *e.g.*, Age $\geq$ 50	Predicate
s	Conjunction of one or more predicates, *e.g.*, Age $\geq$ 50 **and** Gender = Female	Itemset
$\mathcal{S}$	Input set of itemsets	
r	Itemset-class pair (s, c)	Rule
$\mathcal{R}$	Set of rules $\{(s_1, c_1), \ldots, (s_k, c_k)\}$	Decision set

https://himalakkaraju.github.io

Himabindu Lakkaraju, Stephen H Bach & Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016. ACM, 1675-1684.

- + model agnostic
- + learns a compact two-level decision set
- + unambiguously

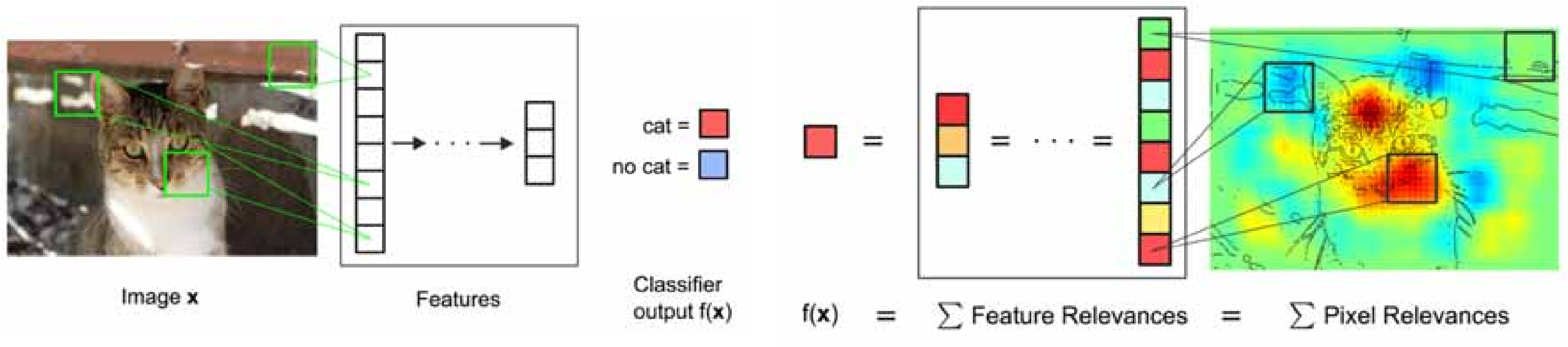
- - not so popular
- - unclear coverage
- - needs care

04c LRP (Layer-wise Relevance Propagation)

- LRP is general solution for understanding classification decisions by pixel-by-pixel (or layer-by-layer) decomposition of nonlinear classifiers (hence the name).
- In a highly simplified way, LRP allows the "thinking processes" of neural networks to run backwards.
- Thereby it becomes comprehensible (for a human) which input had which influence on the respective result,
- e.g. in individual cases how the neural network came to a classification result, i.e. which input contributed most to the gained output.
- Example: If genetic data is entered into a network, it is not only possible to analyze the probability of a patient having a certain genetic disease, but with LRP also the characteristics of the decision.
- Such an approach is a step towards personalised medicine (remember the concept of PM - to provide an individual cancer therapy that is tailored to the particular patient).

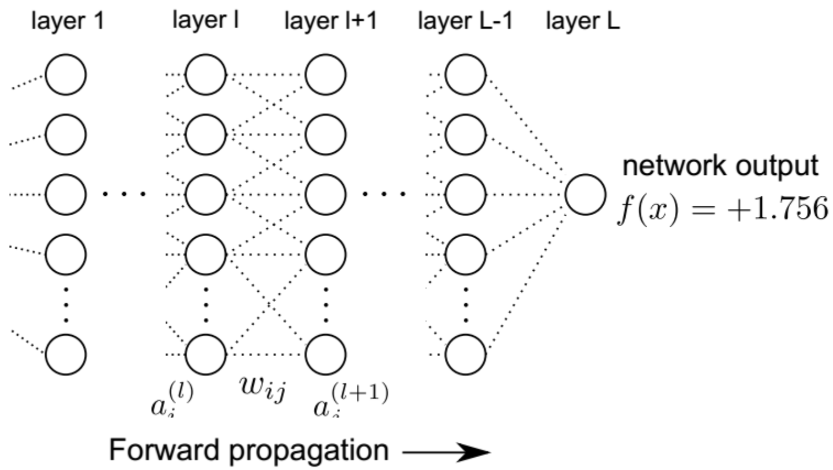
How does LRP work in principle ?

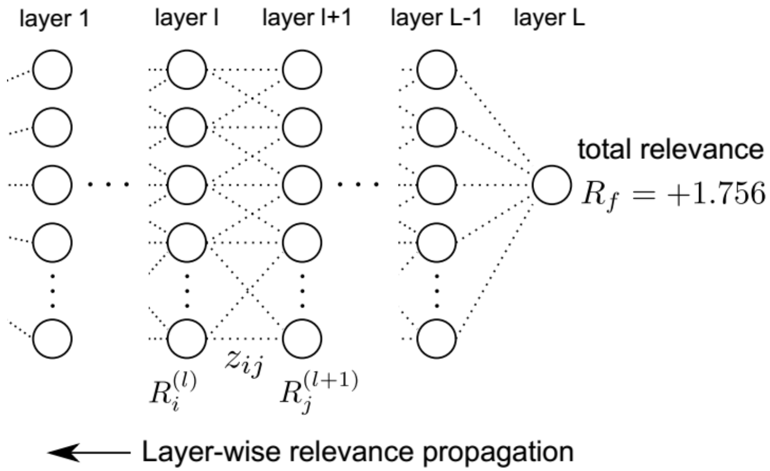
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



$$a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)$$

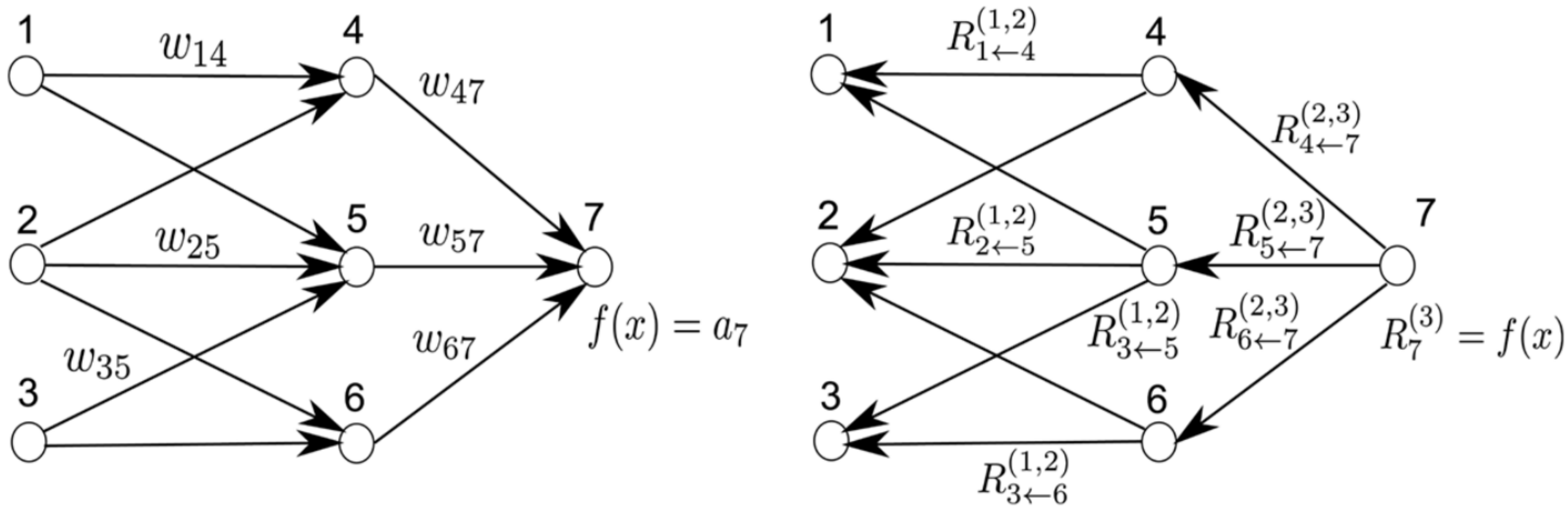
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)}$$





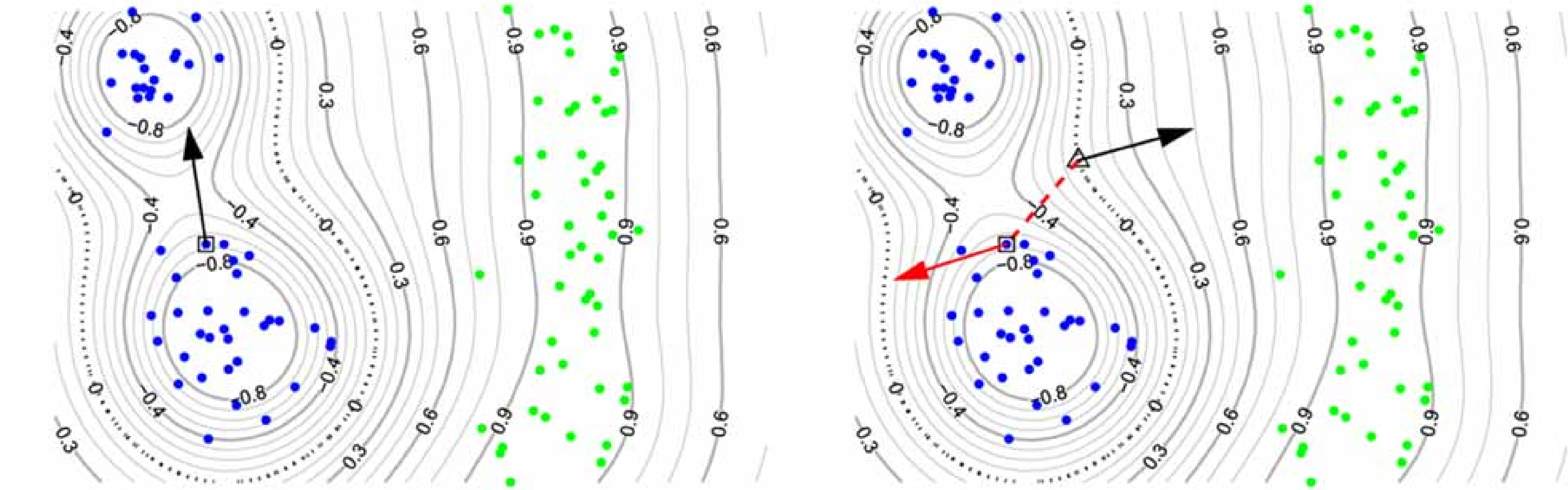
$$R_i = \left\| \frac{\partial}{\partial x_i} f(\mathbf{x}) \right\| \qquad \sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x})$$

How does a NN-classifier during prediction time look like ?



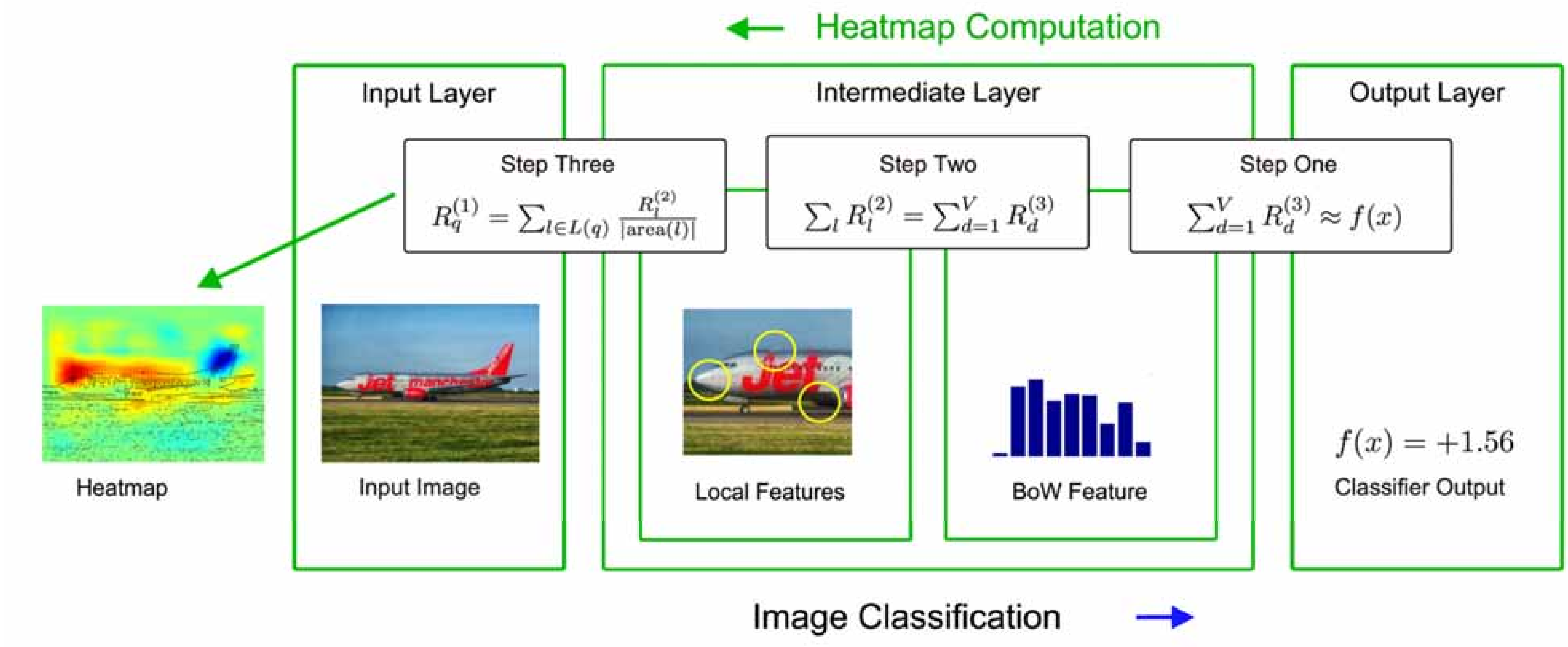
$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_d R_d^{(1)}$$

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.

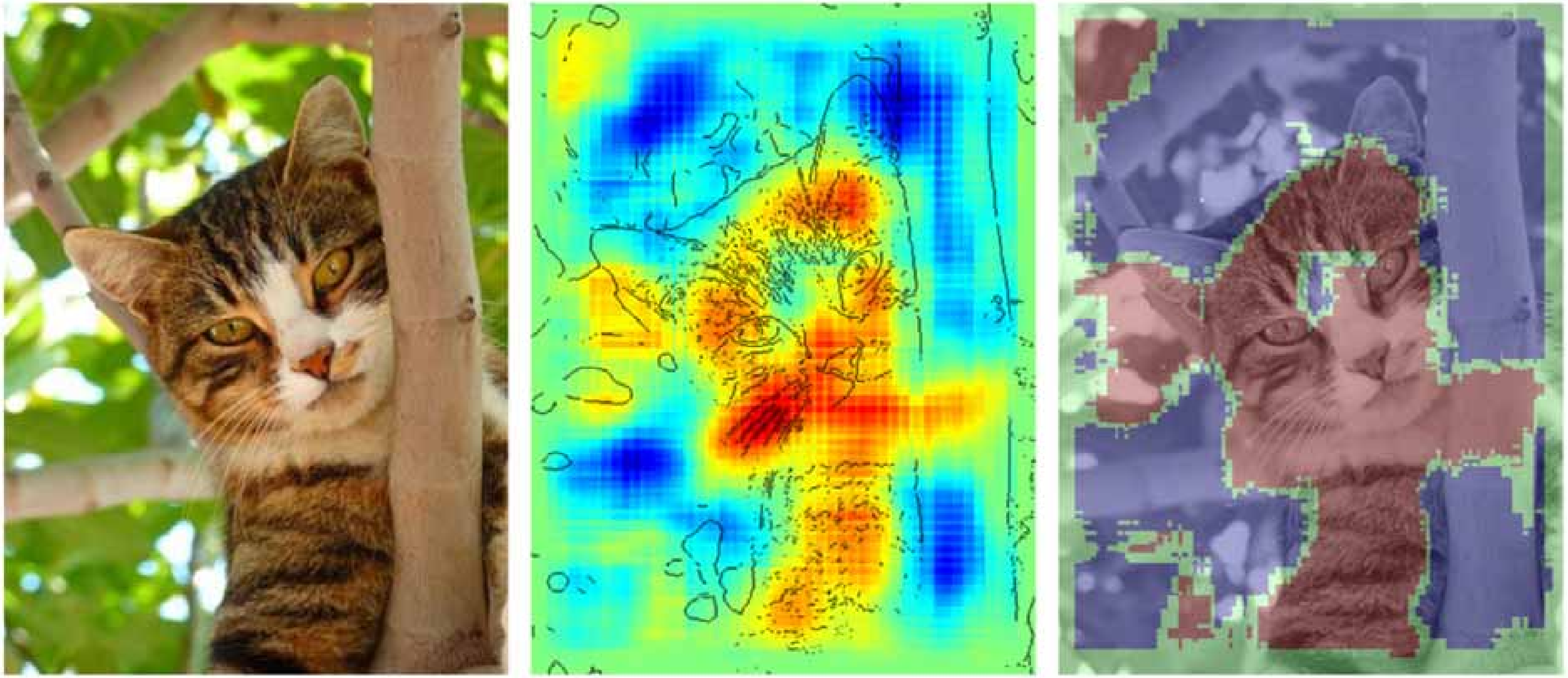


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

What is the output of the LRP ?

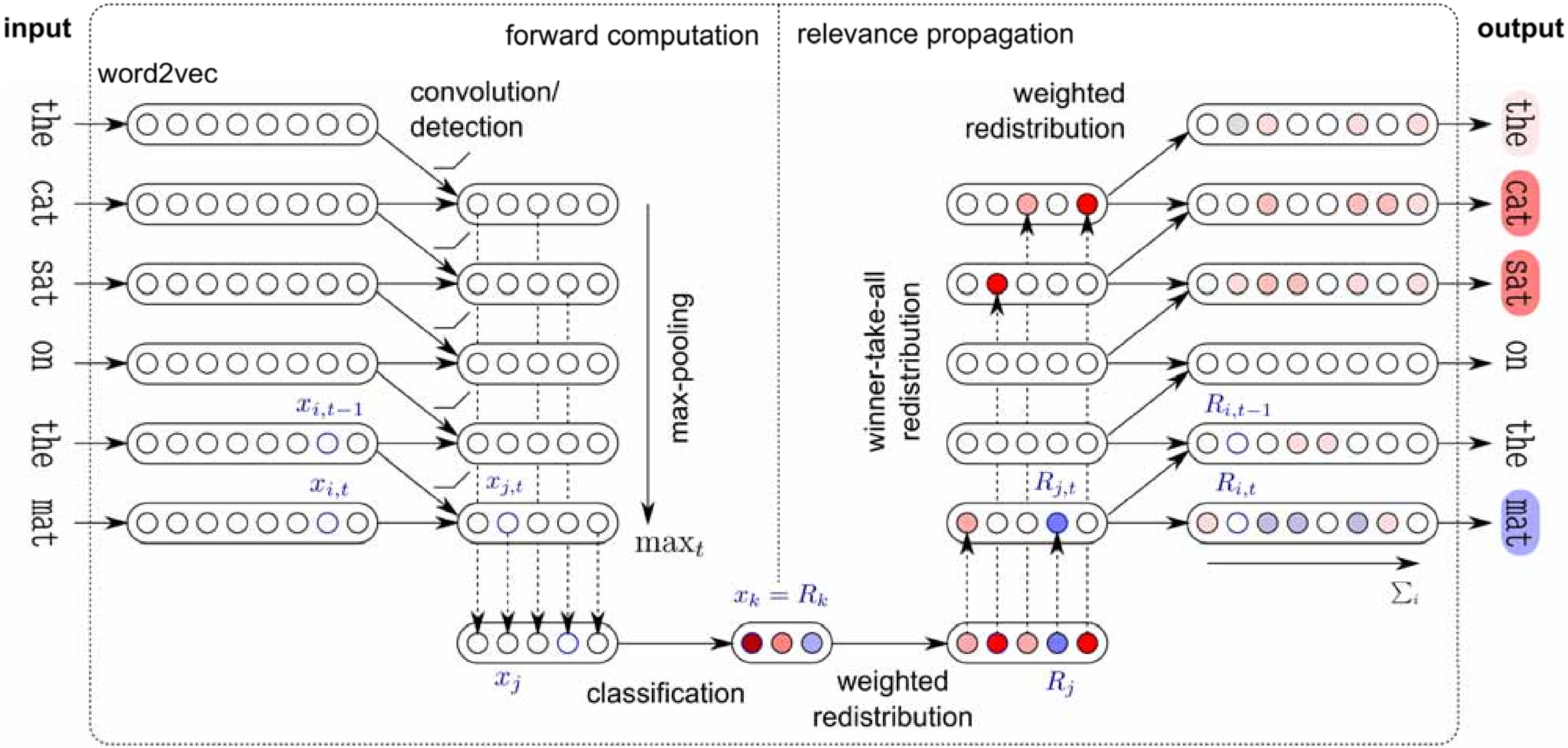


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

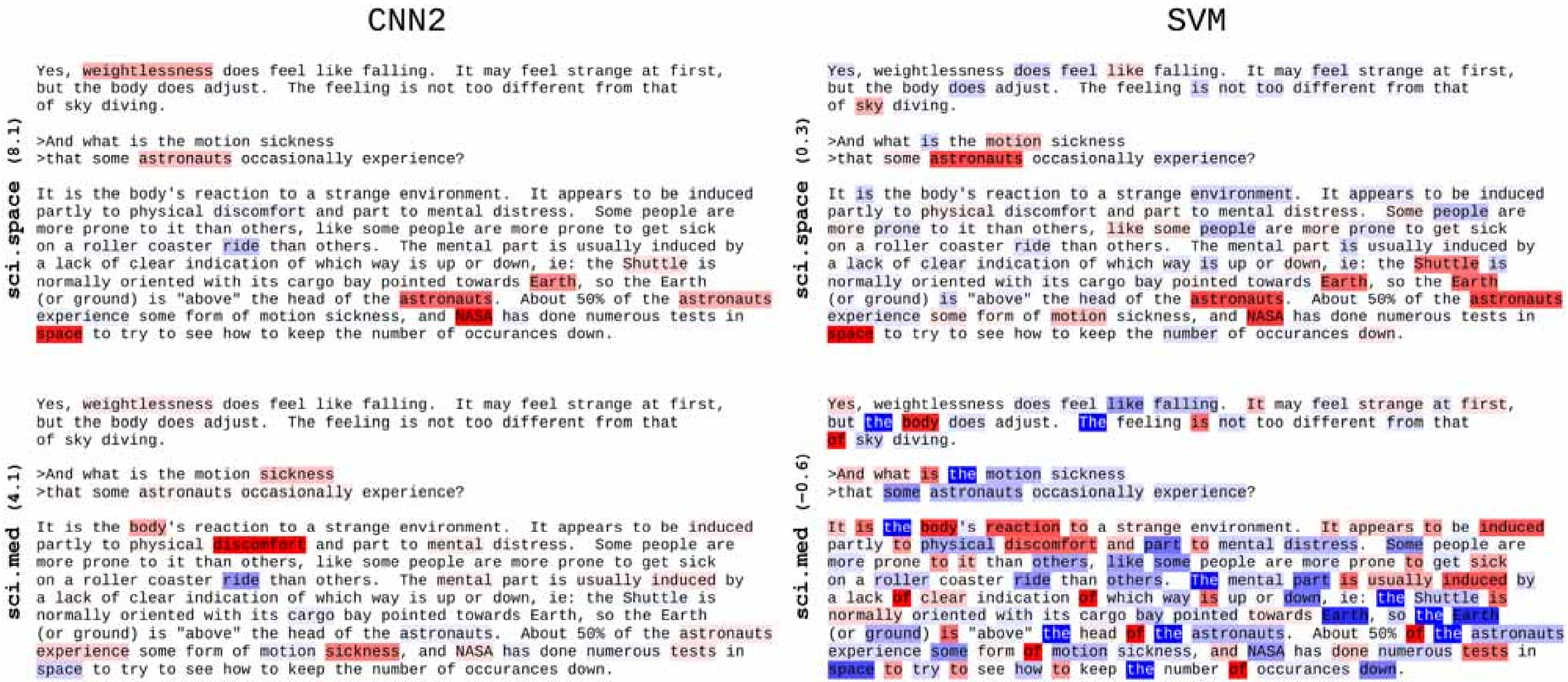


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

What is relevant in a text document?

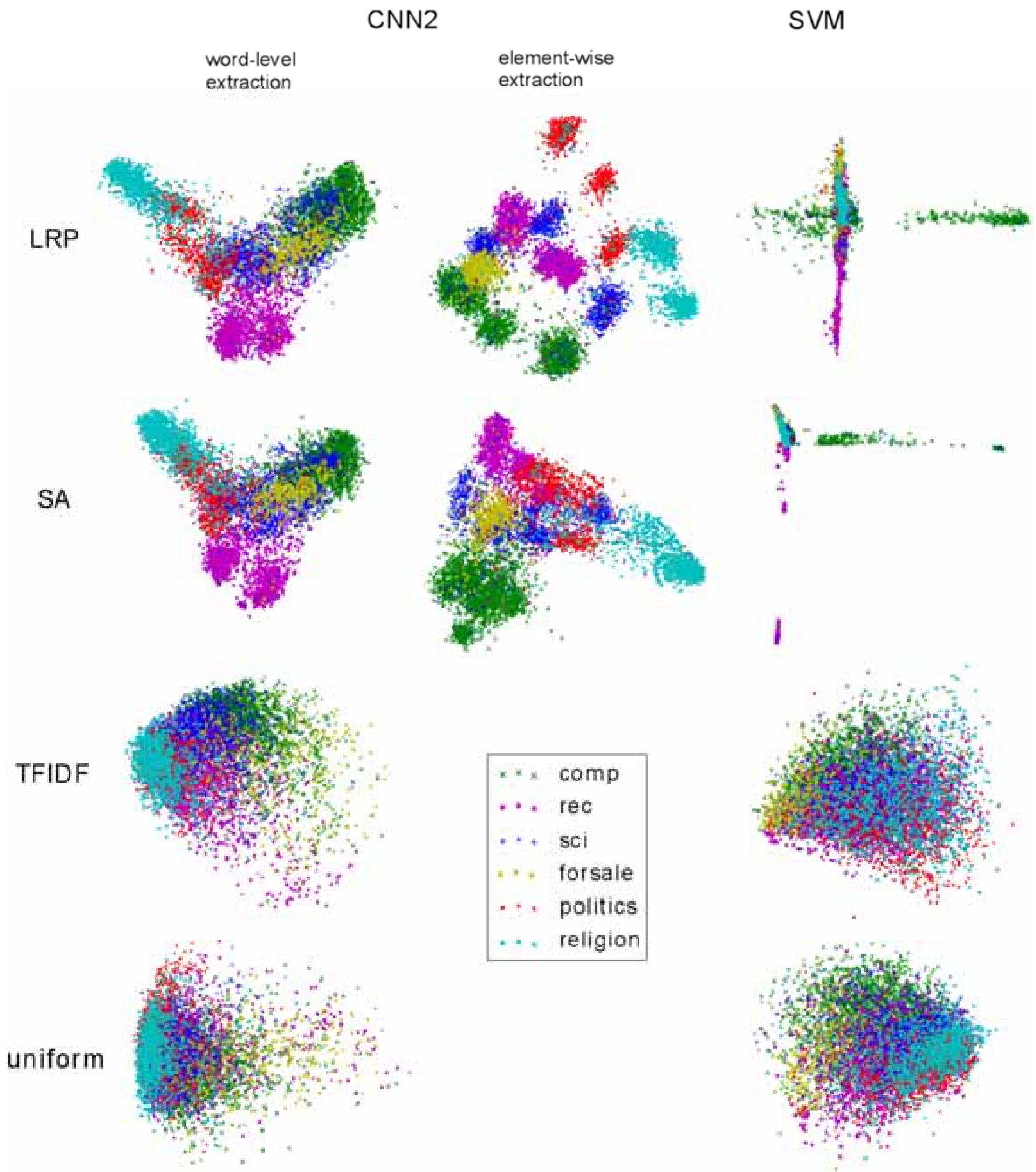


Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller & Wojciech Samek 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12, (8), e0181142, doi:10.1371/journal.pone.0181142.



Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller & Wojciech Samek 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12, (8), e0181142, doi:10.1371/journal.pone.0181142.

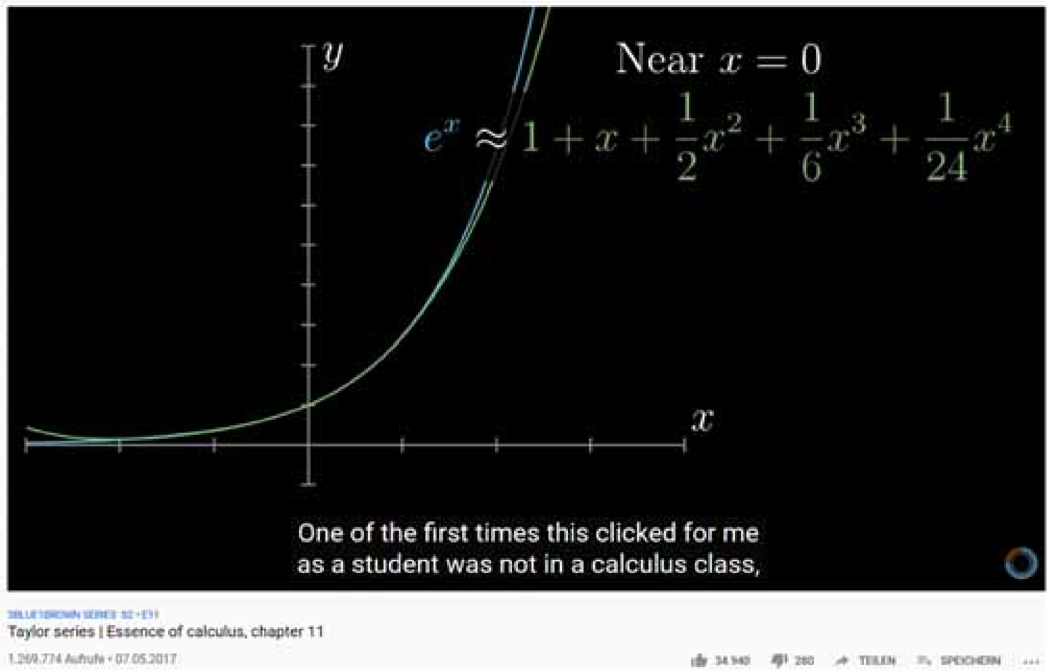
PCA-Projection of the summary vectors



Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller & Wojciech Samek 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12, (8), e0181142, doi:10.1371/journal.pone.0181142.

04d Deep Taylor Decomposition

What are Taylor series ?



https://www.youtube.com/watch?v=3d6DsjIBzJ4

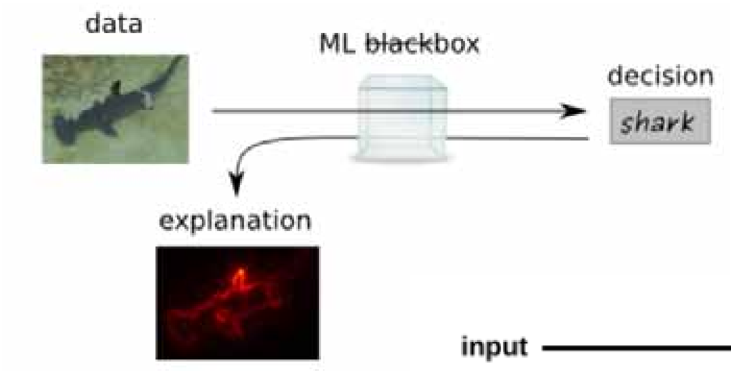
$$f(\boldsymbol{x}) = f(\tilde{\boldsymbol{x}}) + \left(\frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}} \right)^\top \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}}) + \varepsilon = 0 + \sum_p \underbrace{\frac{\partial f}{\partial x_p}\Big|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}} \cdot (x_p - \tilde{x}_p)}_{R_p(\boldsymbol{x})} + \varepsilon,$$

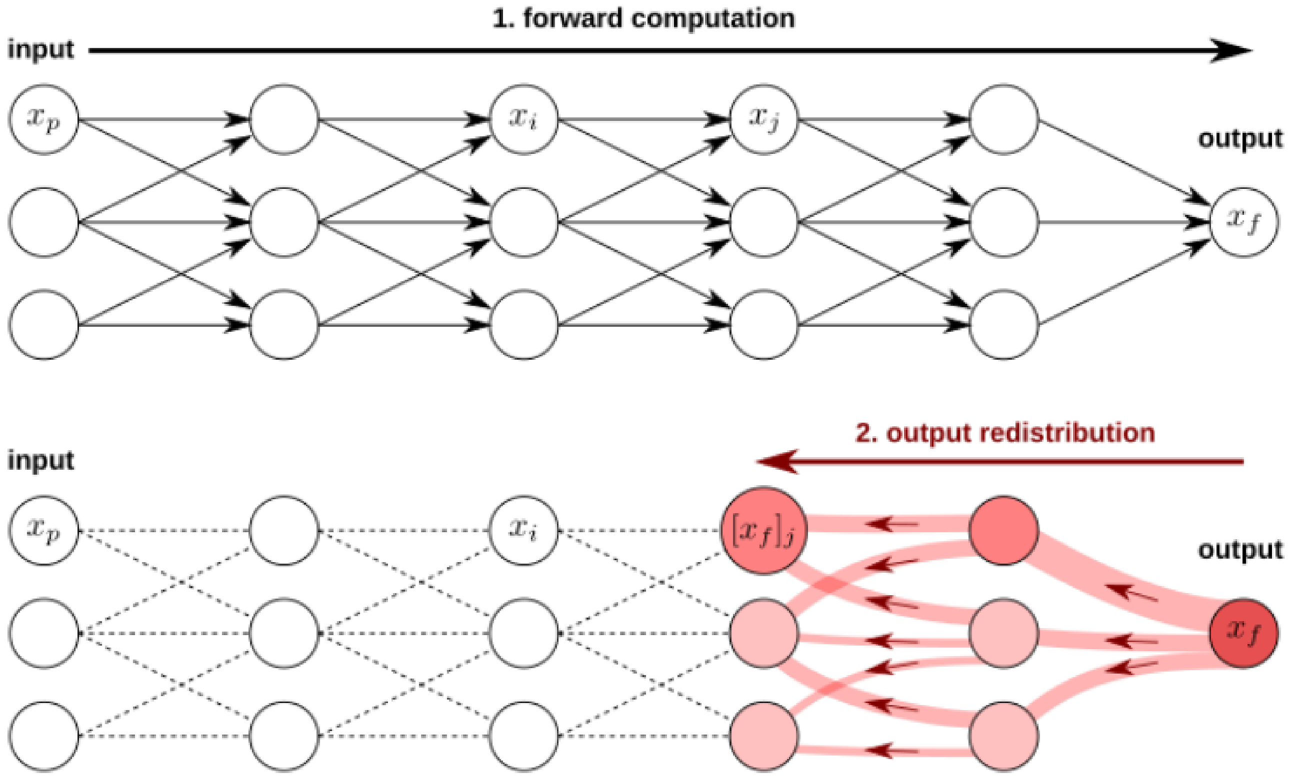
Brook Taylor

Brook Taylor (1685-1731)

Born	18 August 1685 Edmonton, Middlesex, England
Died	29 December 1731 (aged 46) London, England
Residence	England
Nationality	English
Alma mater	St John's College, Cambridge
Known for	Taylor's theorem Taylor series

https://en.wikipedia.org/wiki/Brook_Taylor





http://www.heatmapping.org/deeptaylor

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:

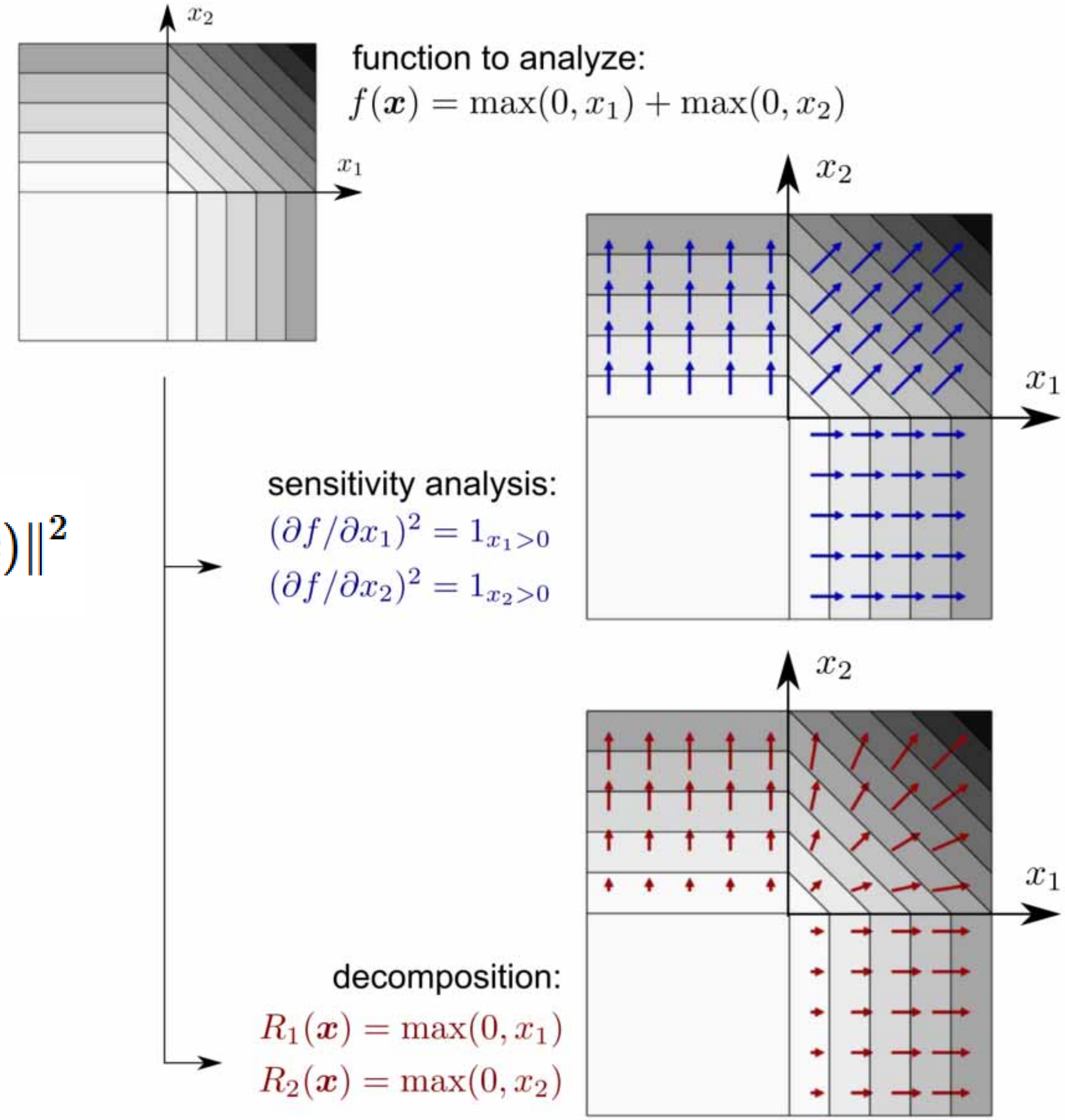
$$\forall\, \boldsymbol{x}: f(\boldsymbol{x}) = \sum_{p} R_p(\boldsymbol{x}).$$

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.

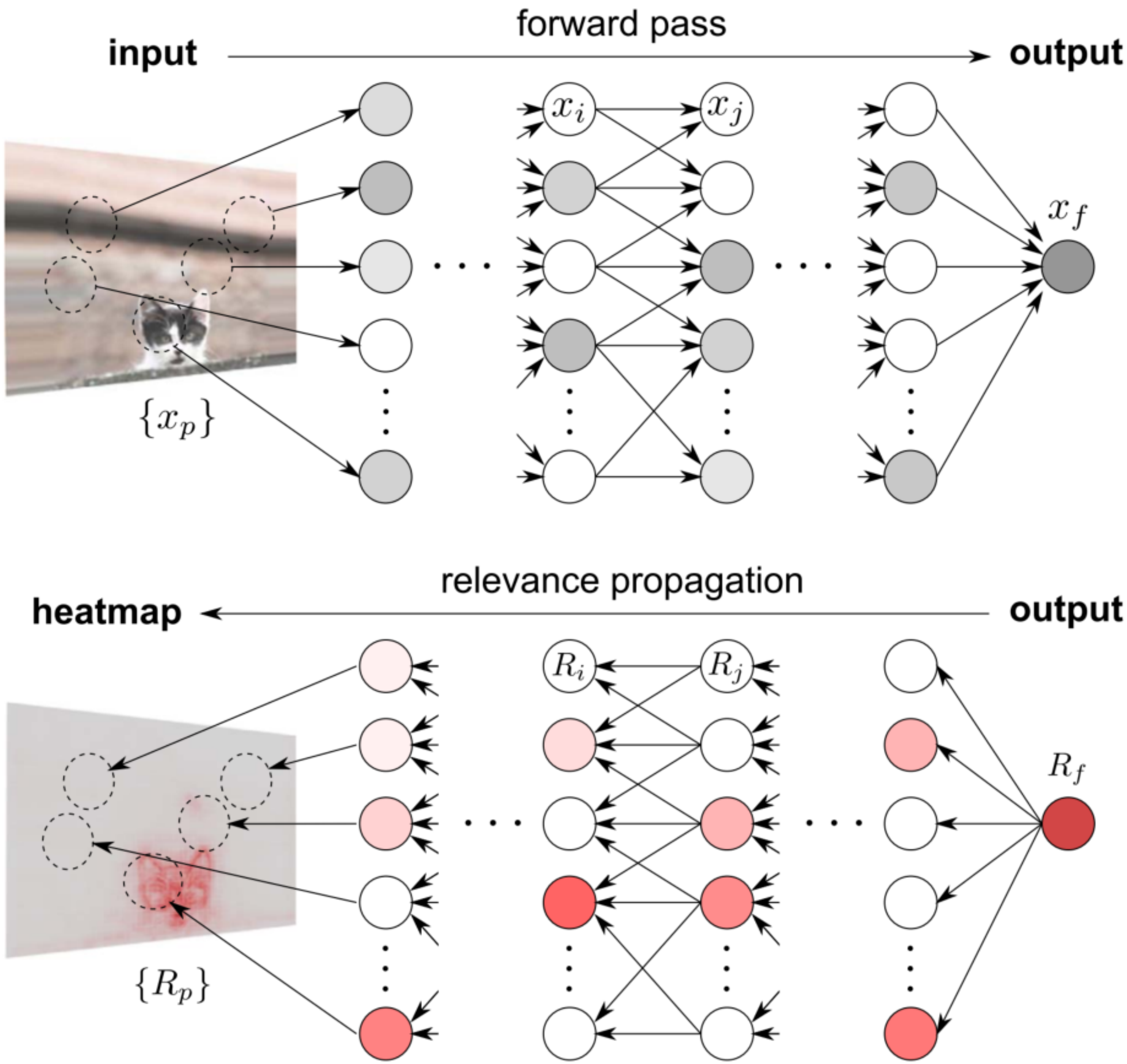
Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.



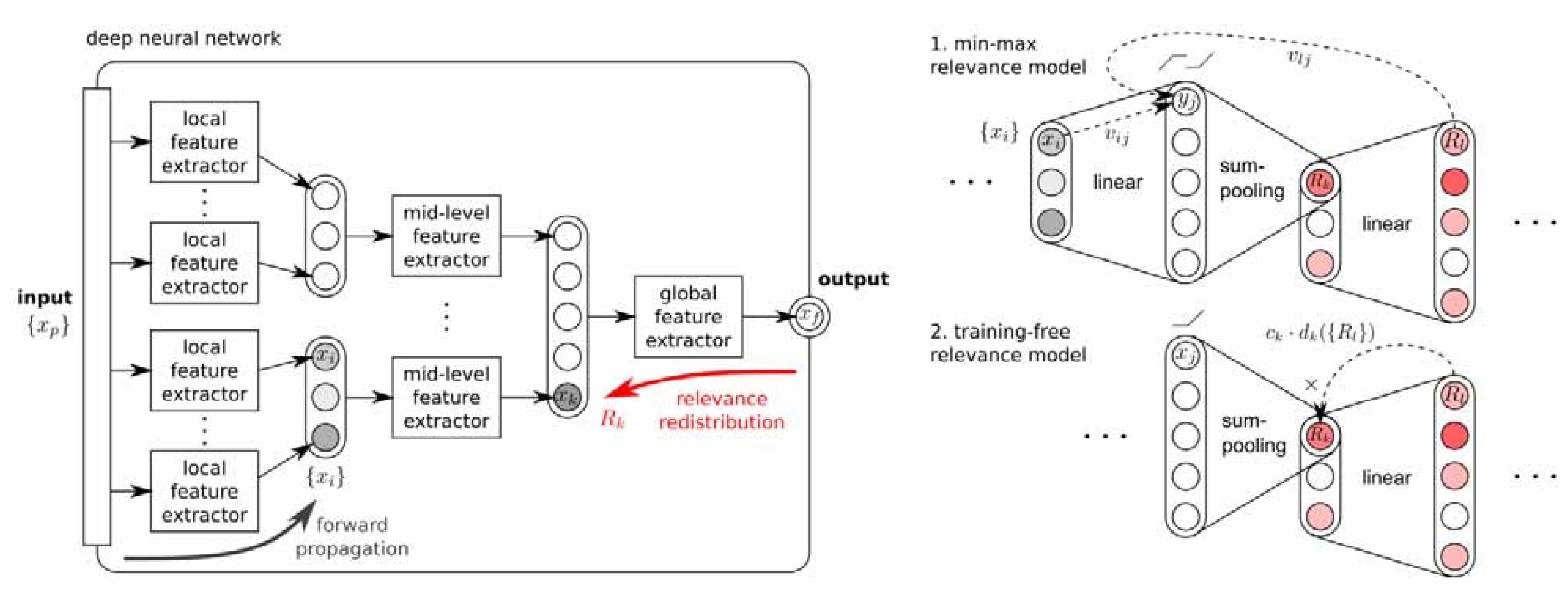
$$\sum_p (\partial f / \partial x_p)^2 = \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|^2$$

$$\sum_p [f(\boldsymbol{x})]_p = f(\boldsymbol{x})$$

How is the computational flow of deep Taylor decomp. in detail ?



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.



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.

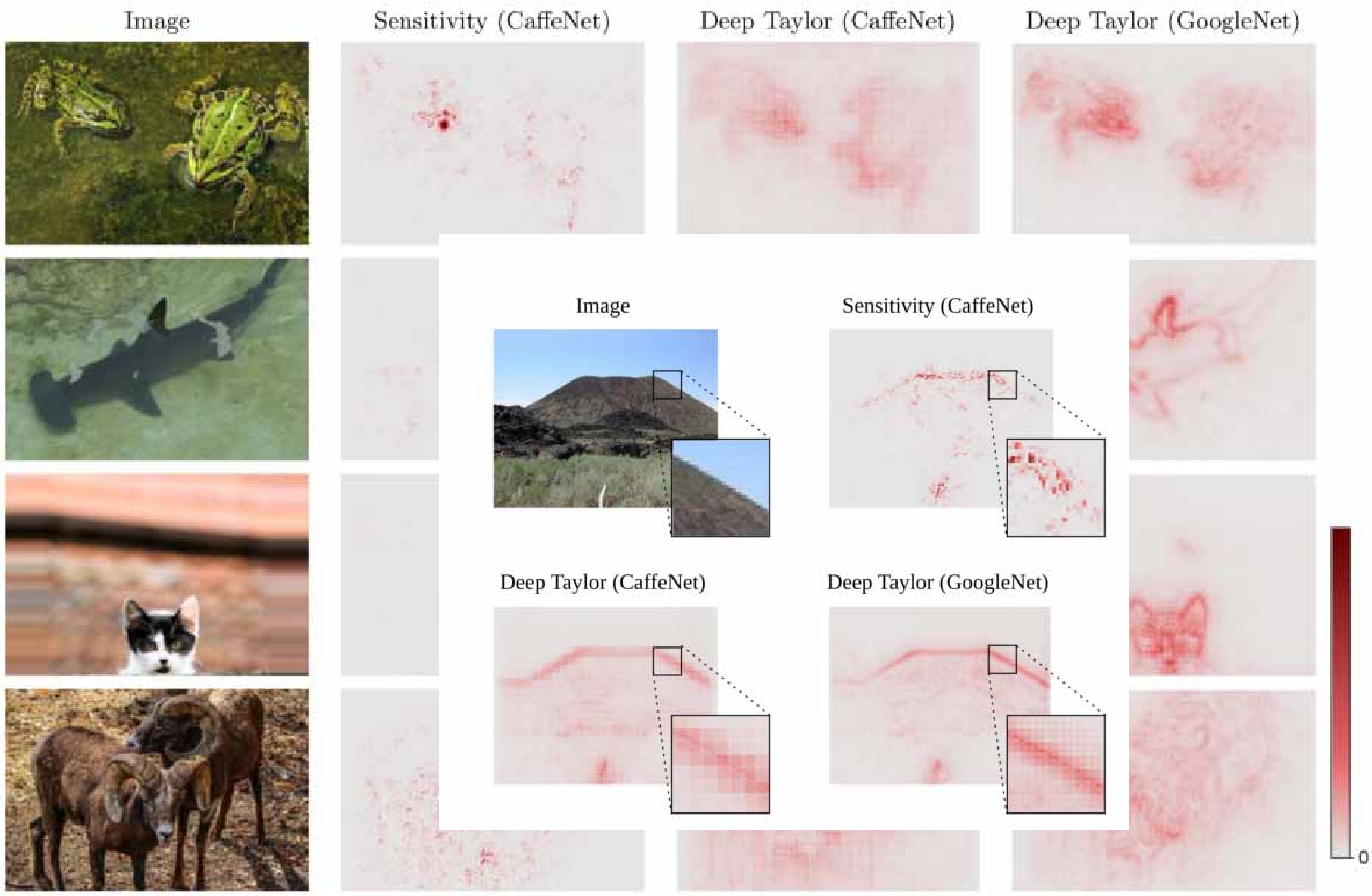
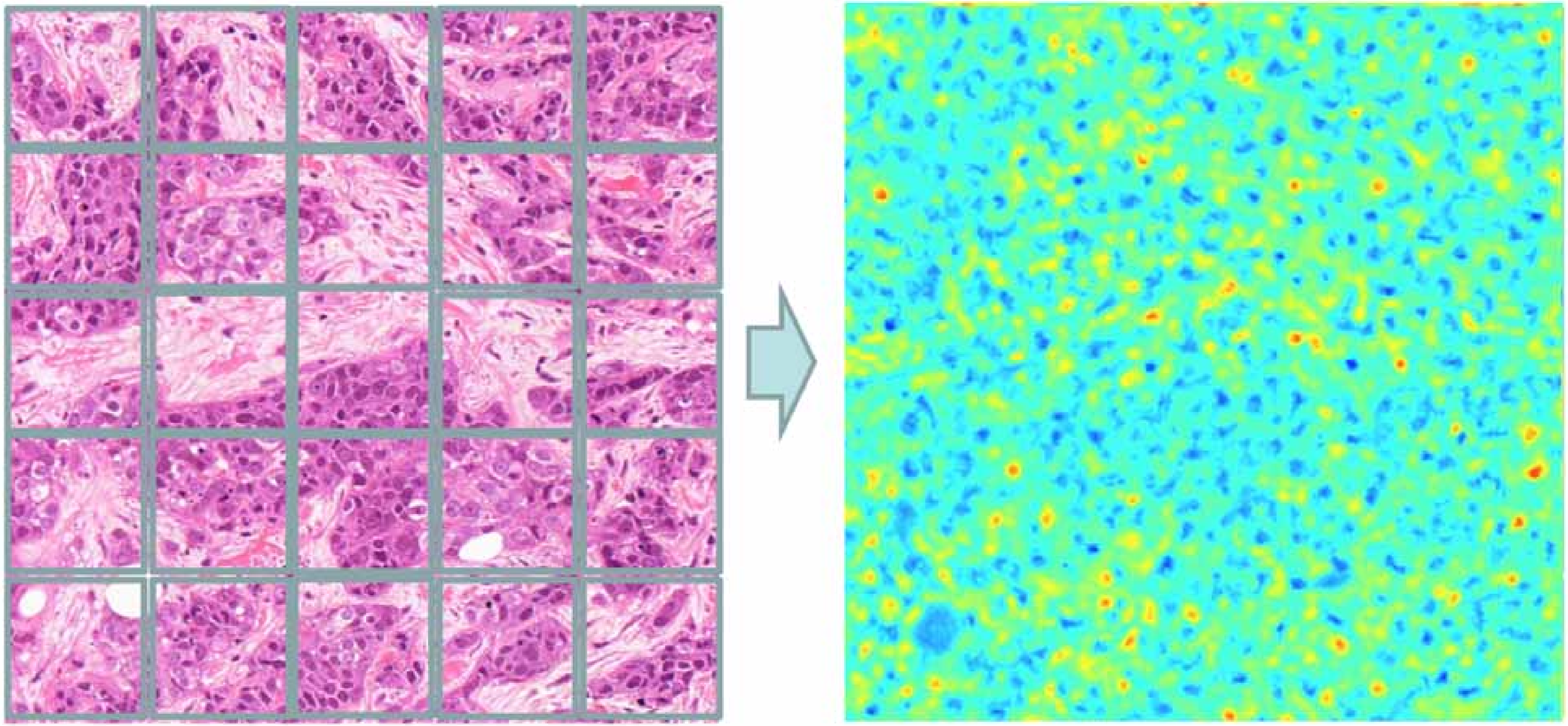


Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
0



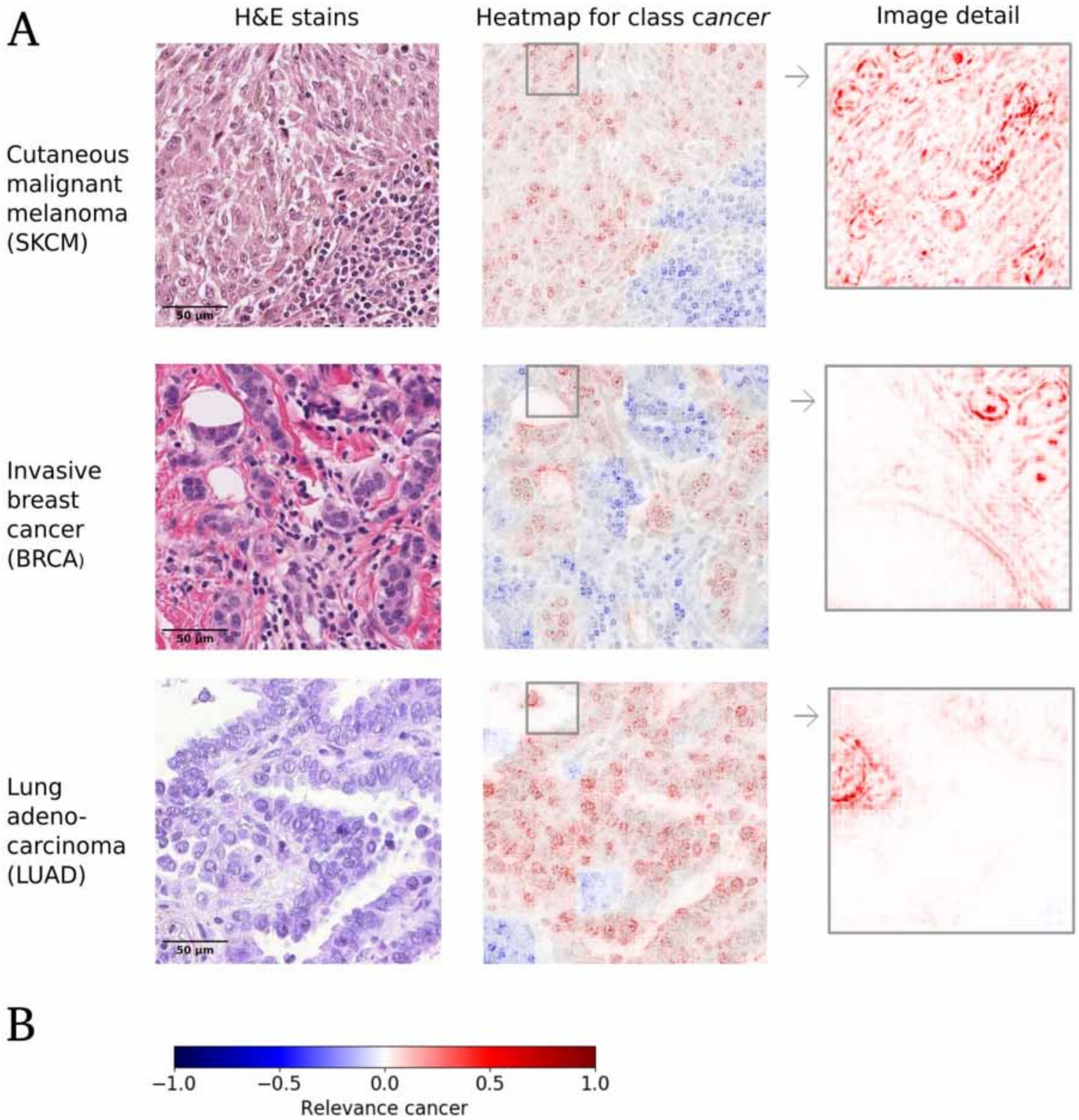
Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller & Alexander Binder 2019. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *arXiv:1908.06943*.

Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Albrecht Stenzinger, Laura Parlow, Jan Budczies & Benjamin Goeppert 2018. Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. *arXiv:1805.11178vl*.

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek & Sebastian Lapuschkin 2019. Towards best practice in explaining neural network decisions with LRP. *arXiv:1910.09840*.

Example: Deep Learning from histopatho images explainable

Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller & Alexander Binder 2020. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Scientific reports, 10, (1), 1-12, doi:10.1038/s41598-020-62724-2.



04e Prediction Difference Analysis

$$p(c|\mathbf{x}_{\setminus i}) = \sum_{x_i} p(x_i|\mathbf{x}_{\setminus i}) p(c|\mathbf{x}_{\setminus i}, x_i)$$

$$p(c|\mathbf{x}_{\setminus i}) \approx \sum_{x_i} p(x_i) p(c|\mathbf{x}_{\setminus i}, x_i)$$

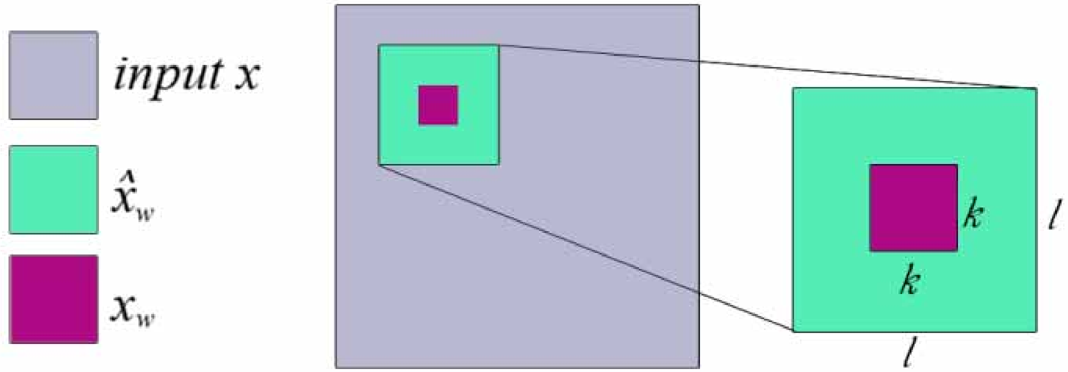
$$\mathrm{WE}_i(c|\mathbf{x}) = \log_2\left(\mathrm{odds}(c|\mathbf{x})\right) - \log_2\left(\mathrm{odds}(c|\mathbf{x}_{\setminus i})\right)$$

Marko Robnik-Šikonja & Igor Kononenko 2008. Explaining Classifications For Individual Instances. *IEEE Transactions on Knowledge and Data Engineering*, 20, (5), 589-600, doi:10.1109/TKDE.2007.190734.

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel & Max Welling 2017. Visualizing deep neural network decisions: Prediction difference analysis. arXiv:1702.04595.

https://github.com/lmzintgraf/DeepVis-PredDiff/blob/master/README.md

https://openreview.net/forum?id=BJ5UeU9xx

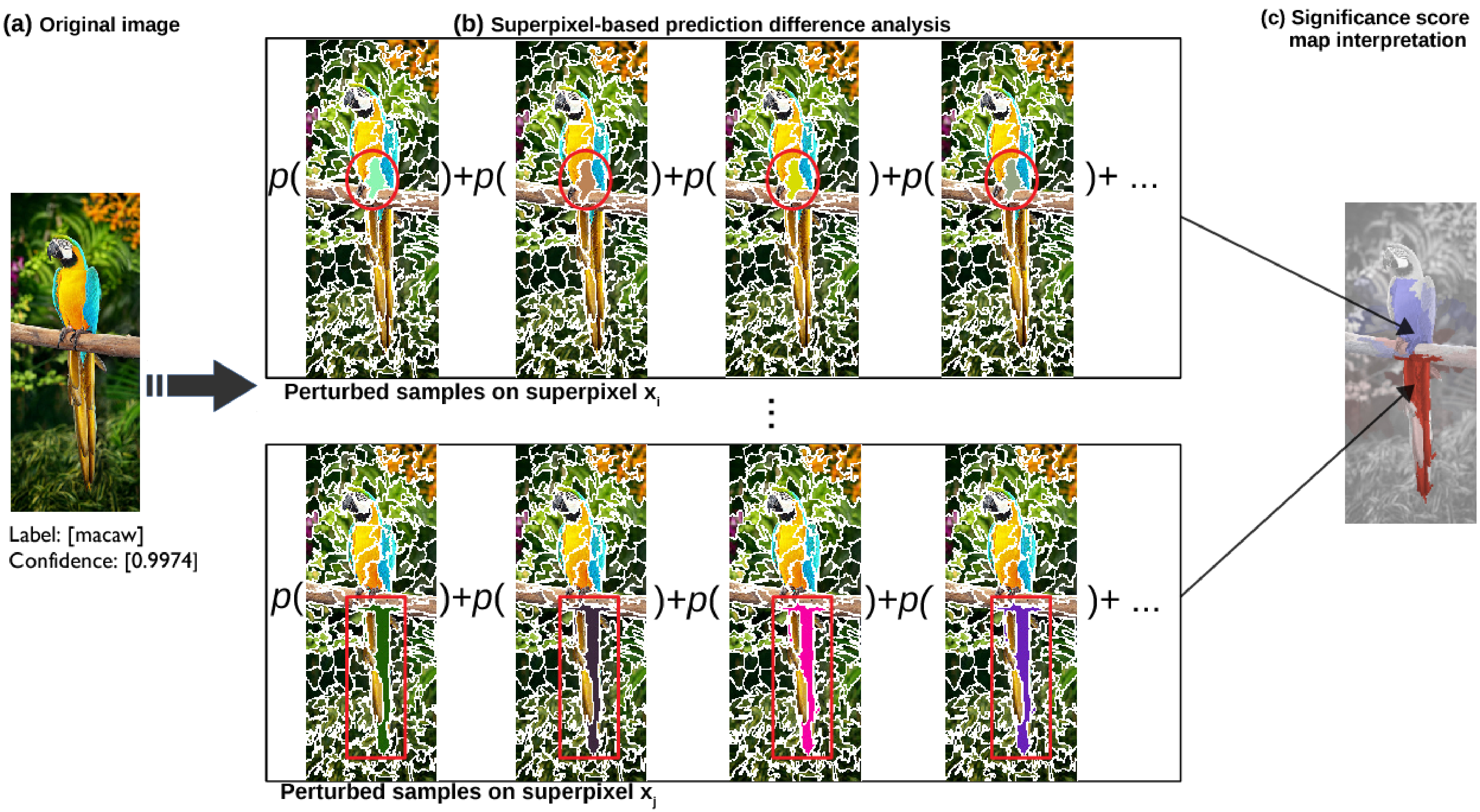


Algorithm 1 Evaluating the prediction difference using conditional and multivariate sampling

Input: classifier with outputs p(c|x), input image $\mathbf{x}$ of size $n \times n$, inner patch size k, outer patch size $l > k$, class of interest c, probabilistic model over patches of size $l \times l$, number of samples S
Initialization: WE = zeros(n*n), counts = zeros(n*n)
for every patch $\mathbf{x}_w$ of size $k \times k$ **in x do**
  $\mathbf{x}' = \text{copy}(\mathbf{x})$
  $\text{sum}_w = 0$
  define patch $\hat{\mathbf{x}}_w$ of size $l \times l$ that contains $\mathbf{x}_w$
  for $s = 1$ **to** S **do**
    $\mathbf{x}'_w \leftarrow \mathbf{x}_w$ sampled from $p(\mathbf{x}_w|\hat{\mathbf{x}}_w \backslash \mathbf{x}_w)$
    $\text{sum}_w \mathrel{+}= p(c|\mathbf{x}')$ ▷ evaluate classifier
  end for
  $p(c|\mathbf{x}\backslash\mathbf{x}_w) := \text{sum}_w / S$
  WE[coordinates of $\mathbf{x}_w$] $\mathrel{+}= \log_2(\text{odds}(c|\mathbf{x})) - \log_2(\text{odds}(c|\mathbf{x}\backslash\mathbf{x}_w))$
  counts[coordinates of $\mathbf{x}_w$] $\mathrel{+}= 1$
end for
Output: WE / counts ▷ point-wise division

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel & Max Welling 2017. Visualizing deep neural network decisions: Prediction difference analysis. arXiv:1702.04595.

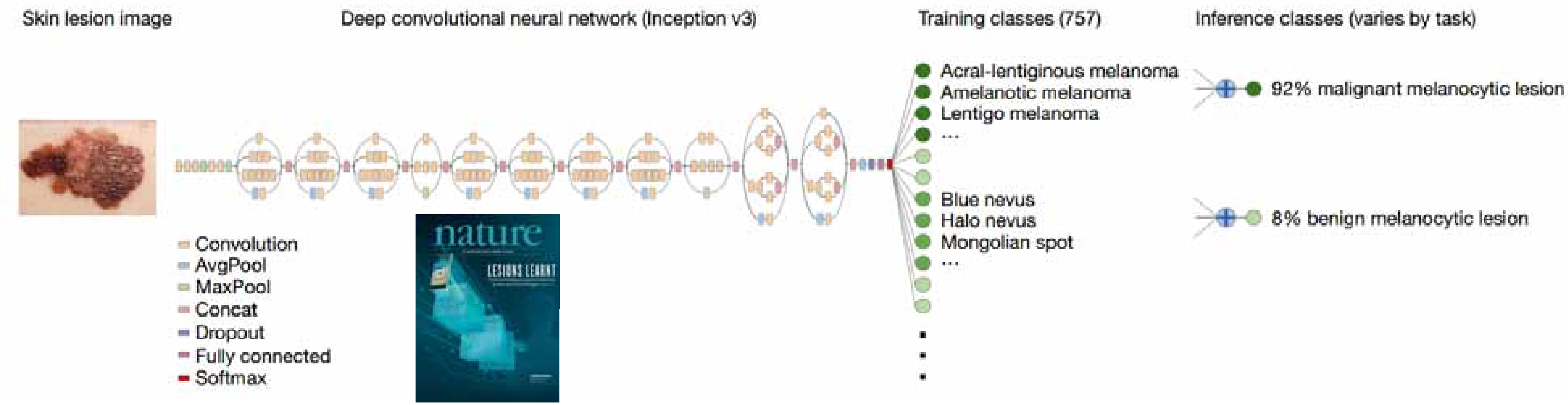
What is Superpixel-based prediction difference analysis ?



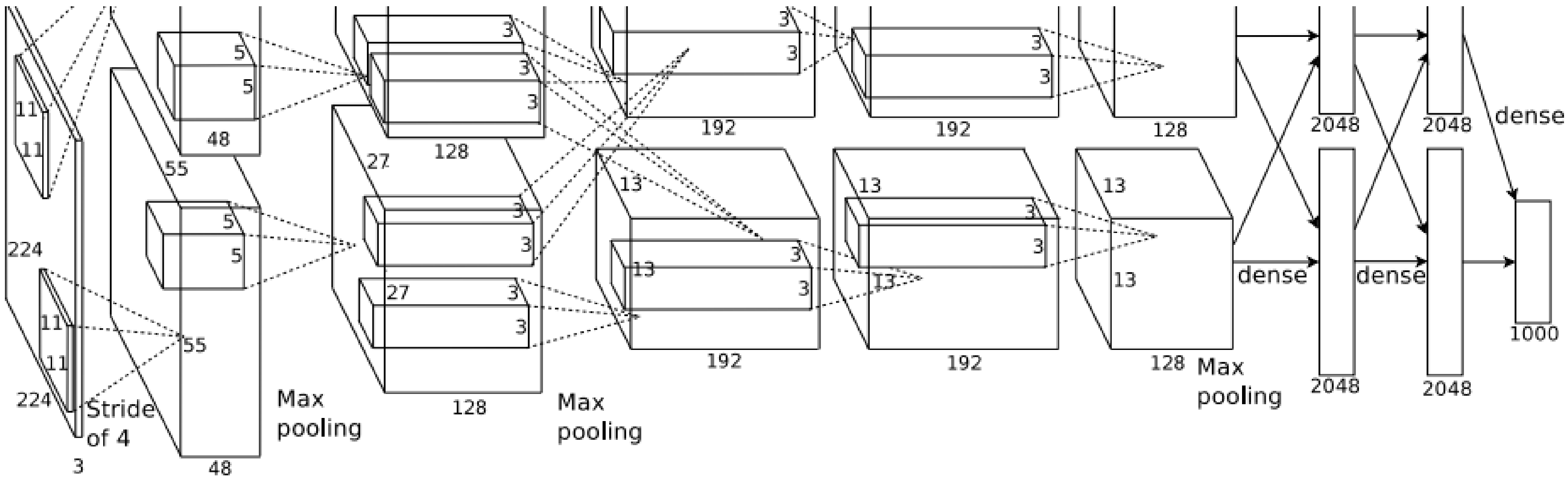
Yi Wei, Ming-Ching Chang, Yiming Ying, Ser Nam Lim & Siwei Lyu. Explain Black-box Image Classifications Using Superpixel-based Interpretation. 2018 24th International Conference on Pattern Recognition (ICPR), 2018. IEEE, 1640-1645.

Digression: Visualizing CNN with Deconvolution

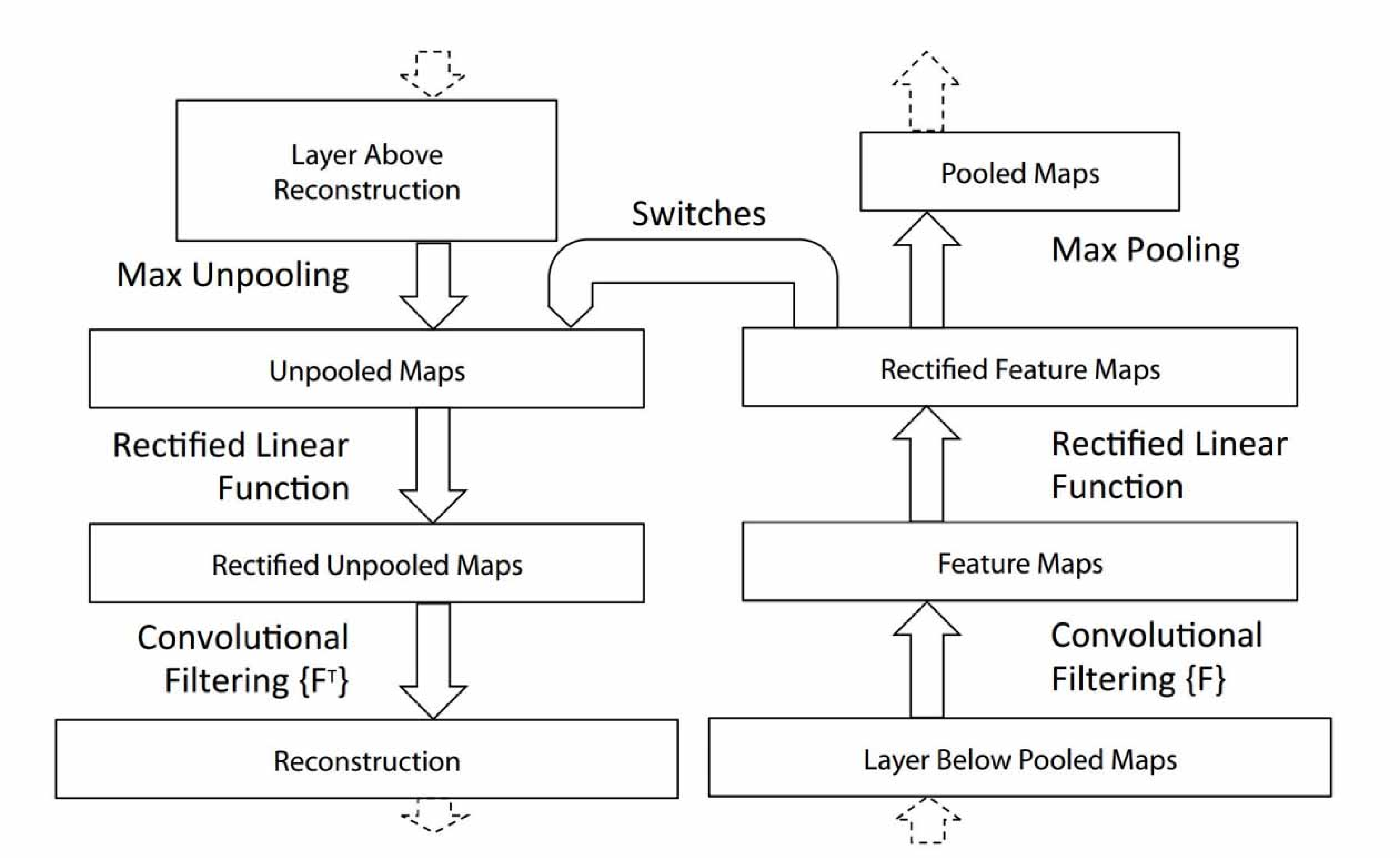
Our typical example to answer the question of why



Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

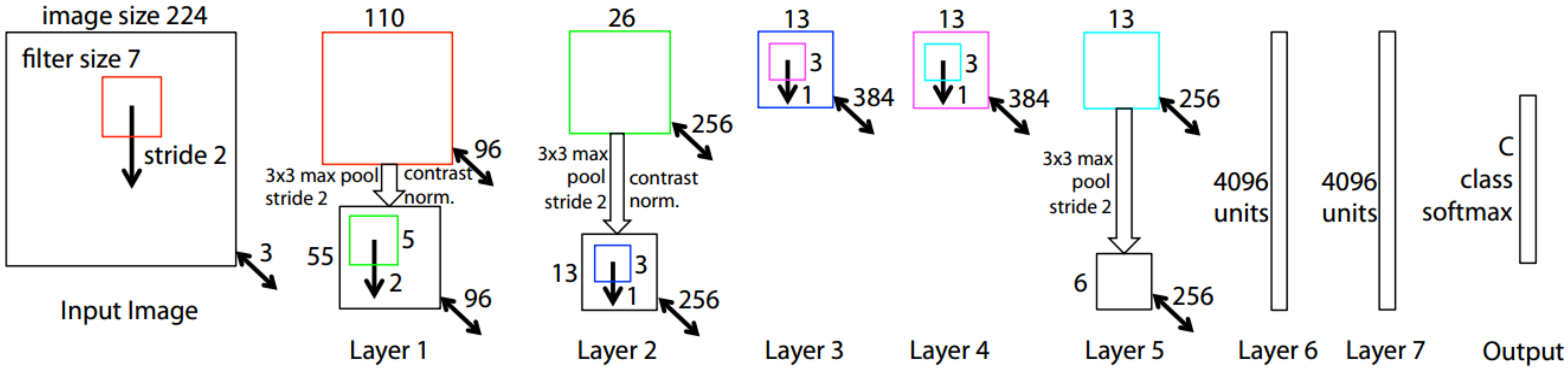


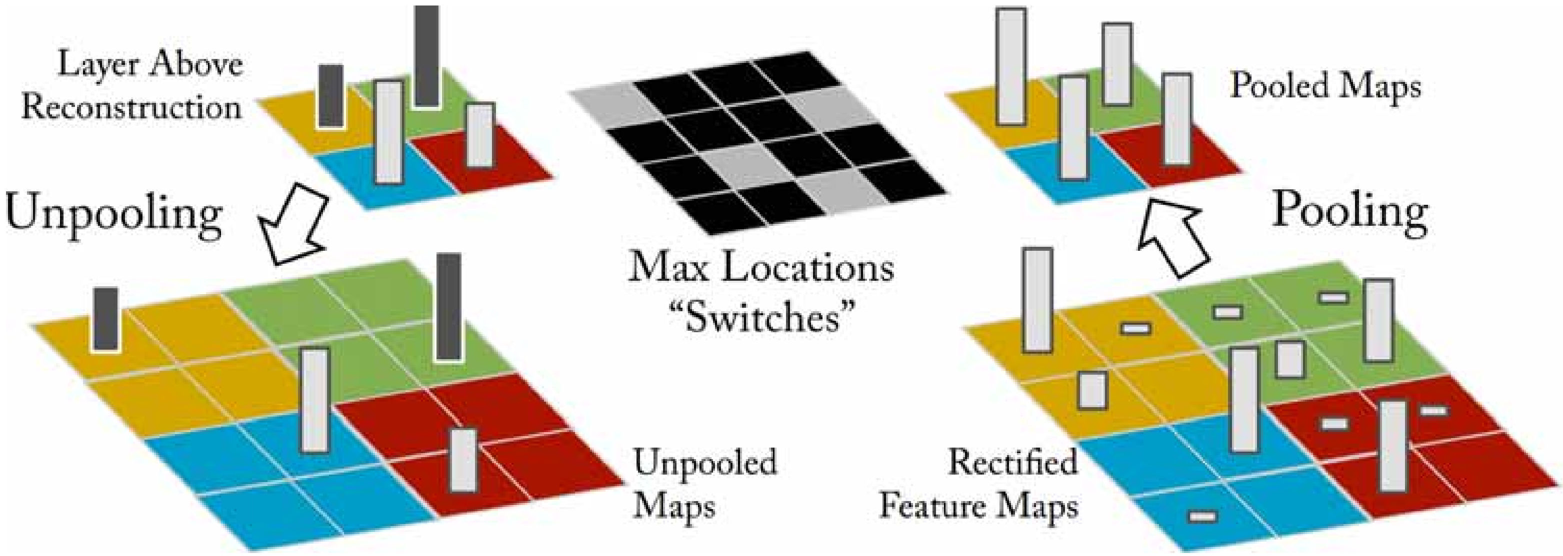
How to get insight into a deep neural network ?



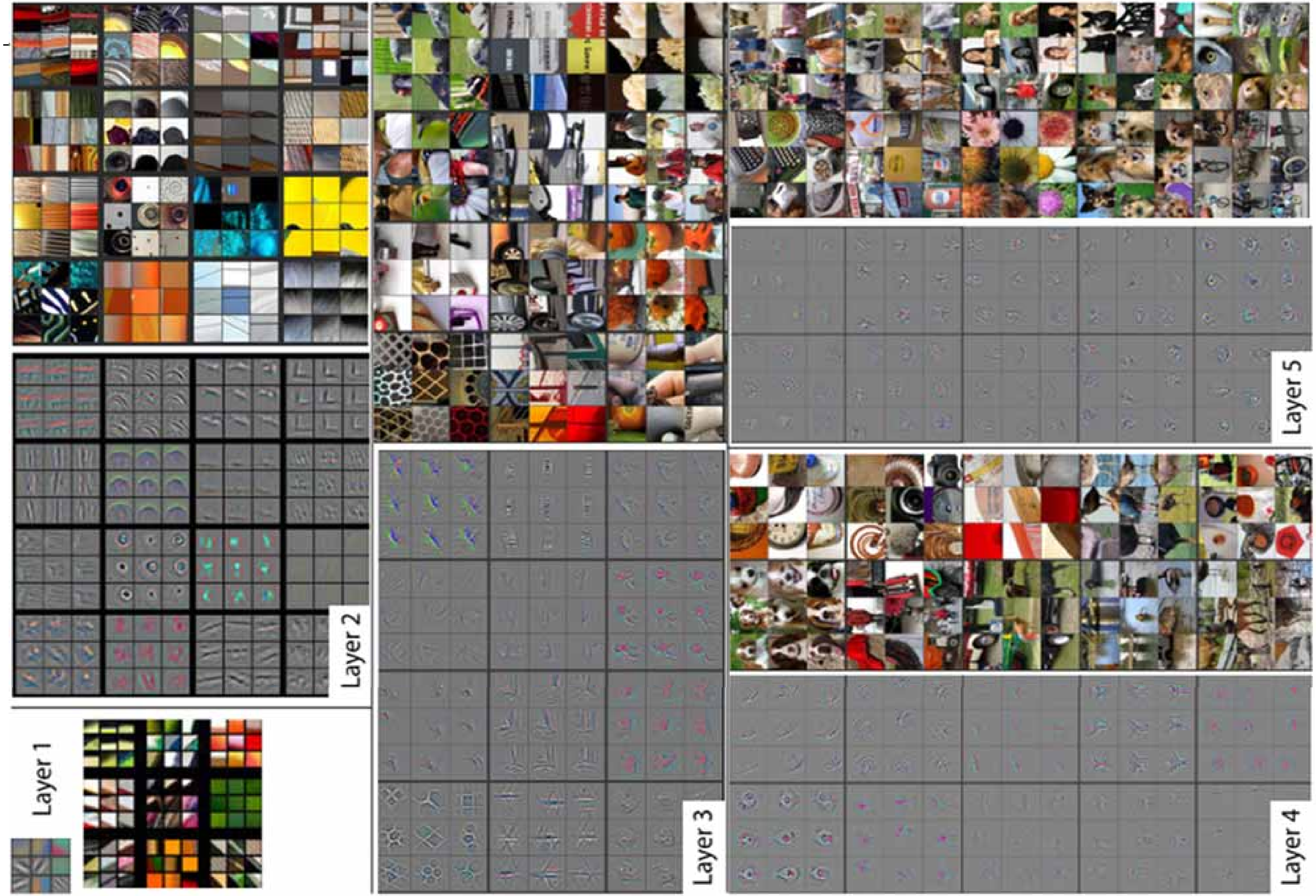
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

Visualizing a Conv Net with a De-Conv Net

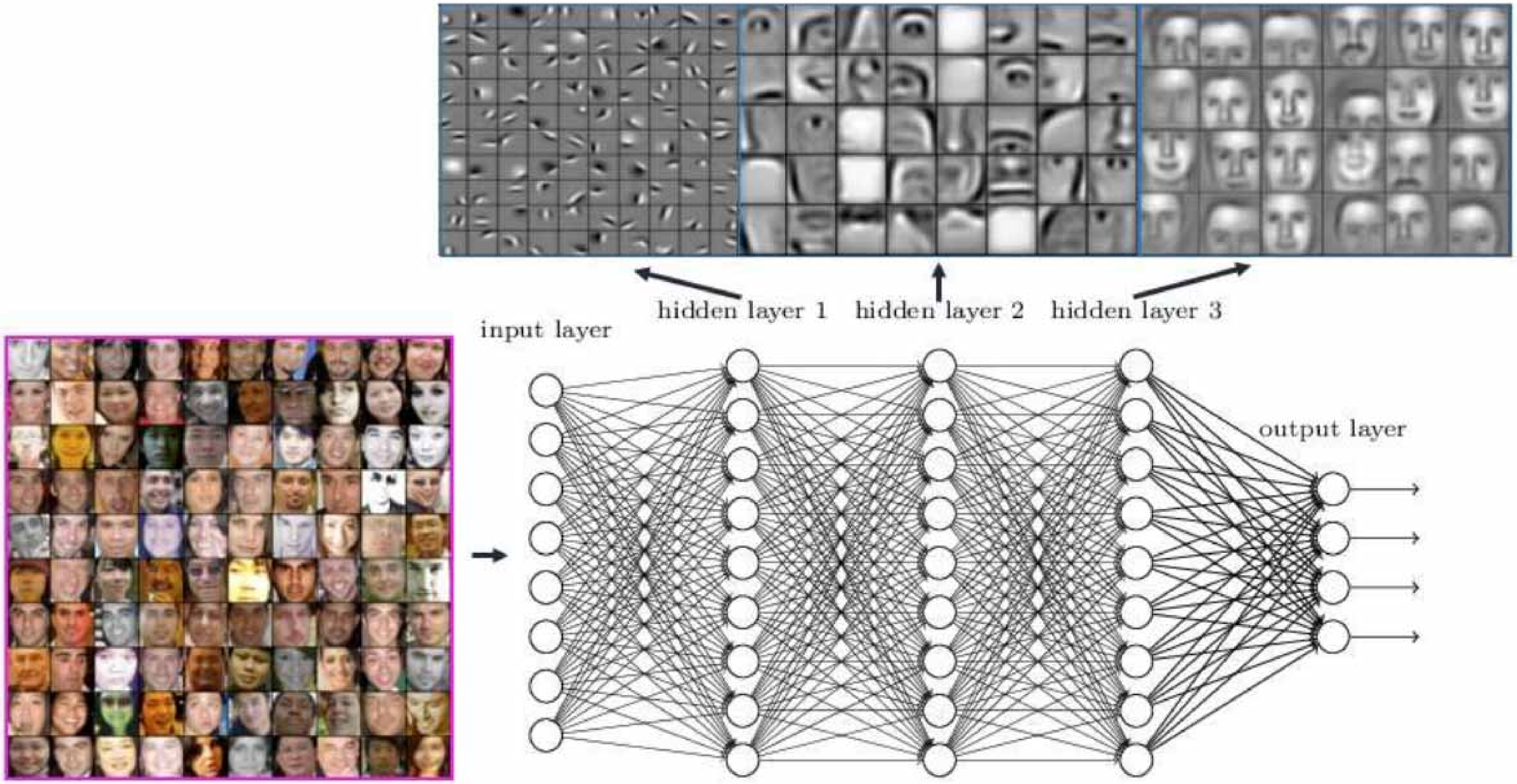




Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



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



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

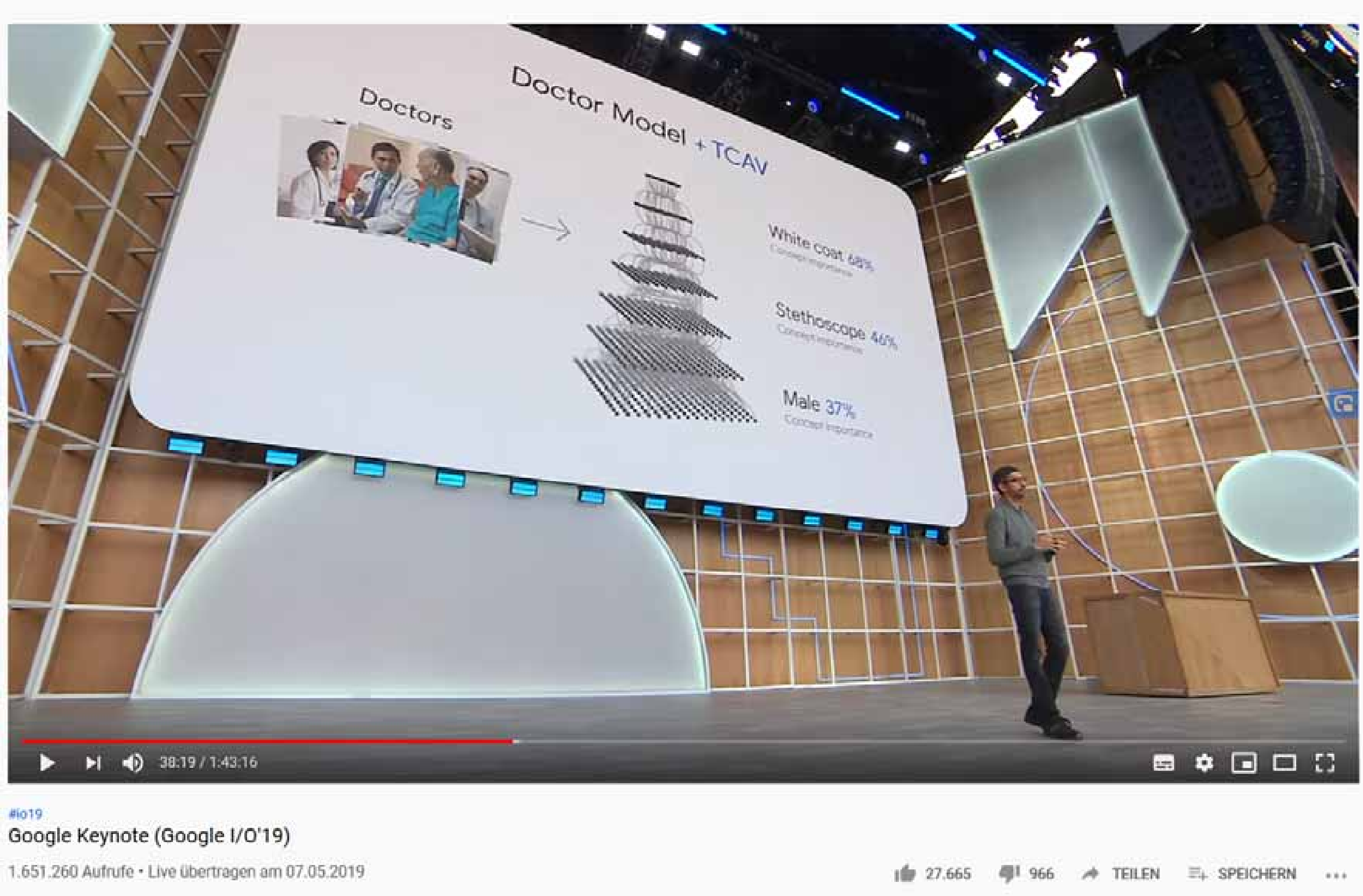
04f Testing with Concept Activation Vectors (TCAV)

- **“It’s not enough to know if a model works, we need to know *how* it works”**

... if Sundar Pichai is saying this ...

Sundar Pichai

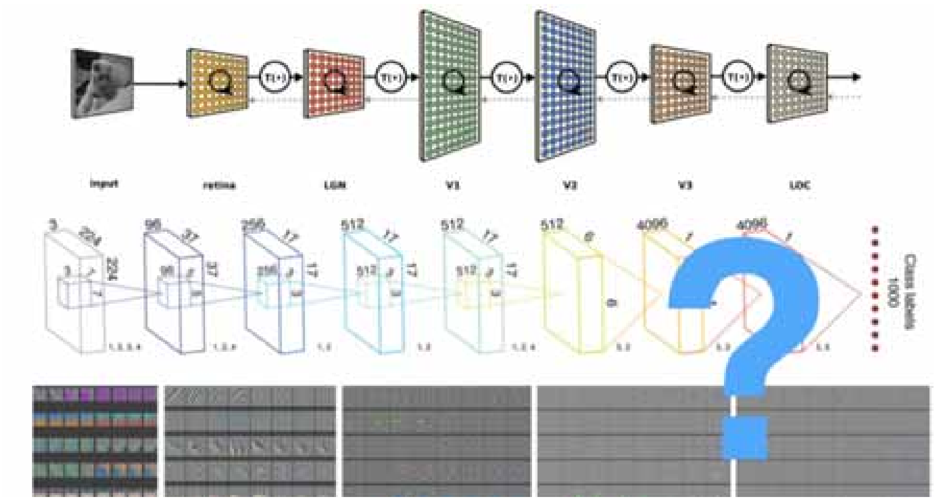
Born	Pichai Sundararajan June 10, 1972 (age 47) Madurai, Tamil Nadu, India
Alma mater	Indian Institute of Technology Kharagpur (BTech) Stanford University (MS) The Wharton School (MBA)
Salary	US$1,881,066 (2018) US$1,333,557 (2017)[1] US$199.7 million[2] (2016)
Title	CEO of Google and Alphabet
Board member of	Alphabet Inc.[3] Magic Leap[4]



https://www.youtube.com/watch?v=lyRPyRKHO8M&t=3408s

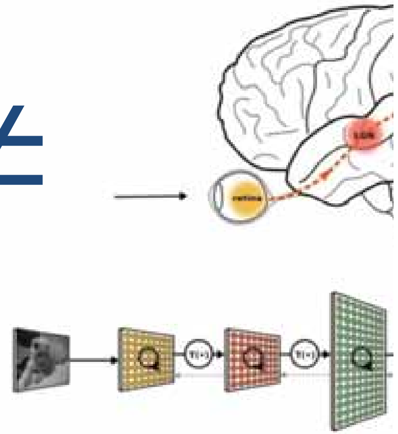
- ML models work on **low-level features** (edges, dots, lines, pixel, circles, …)
- Humans are working on **high-level concepts** (shape, size, color, Gestalt-principles, …)
- Every pixel of an image is a input feature and are just numbers, which do not make sense to humans.
- TCAV enables to provide an explanation that is generally true for a class of interest, beyond one image (global explanation).
- The goal of TCAV is to learn ‘concepts’ from examples.

How does a Concept Activation Vector (CAV) work ?



Yann Lecun, Yoshua Bengio & Geoffrey
Nature, 521, (7553), 436-444, doi:10.1(

$\neq$



$$\frac{\partial h_k(x)}{\partial x_{a,b}}$$

Humans work in another vector space wh
spanned by **implicit knowledge** vectors c
to an unknown set of human interpretabl

$$S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}\left(f_l(x) + \epsilon v_C^l\right) - h_{l,k}\left(f_l(x)\right)}{\epsilon} = \nabla h_{l,}$$

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. In
feature attribution: Quantitative testing with concept activation vectors, ICML, 2018. 2673-2

3.3. Directional Deri

Interpretability metho
of logit values with re
pixels, and compute

where $h_k(\boldsymbol{x})$ is the lo
$\boldsymbol{x}_{a,b}$ is a pixel at pos
the derivative to gaug
to changes in the mag

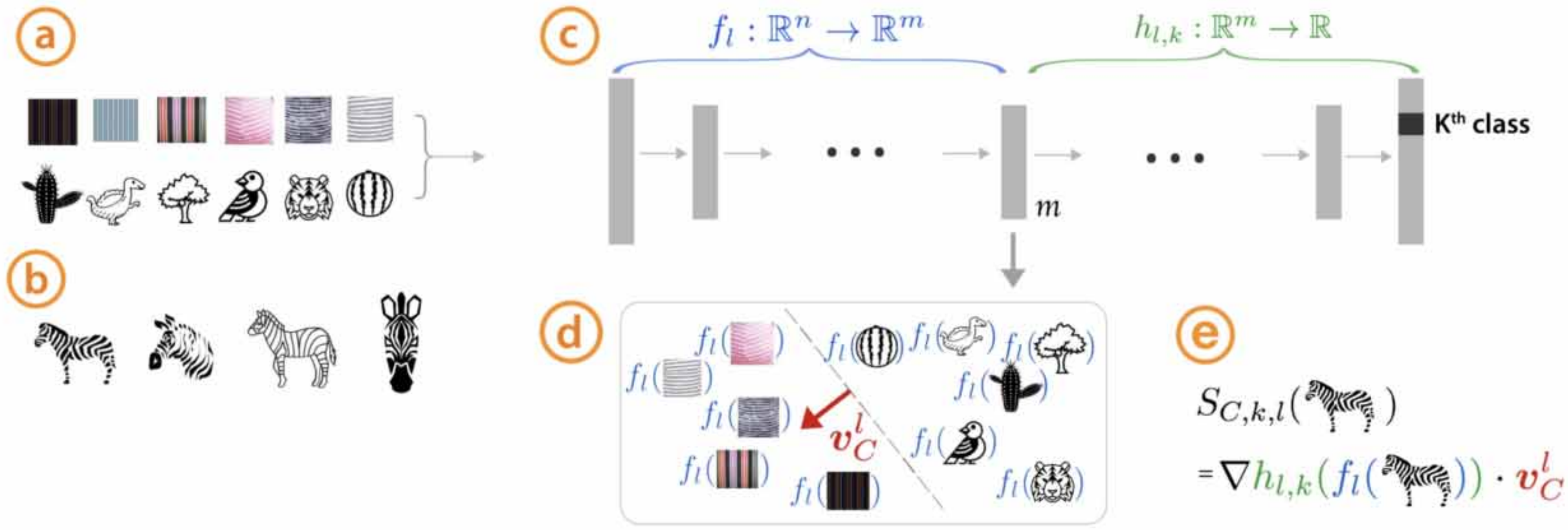
By using CAVs and di
the sensitivity of ML
wards the direction of
If $\boldsymbol{v}_C^l \in \mathbb{R}^m$ is a unit
and $f_l(\boldsymbol{x})$ the activati
tual sensitivity" of cla
the directional deriva

$$S_{C,k,l}(\boldsymbol{x}) = \lim_{\epsilon \to} \qquad = \nabla h$$

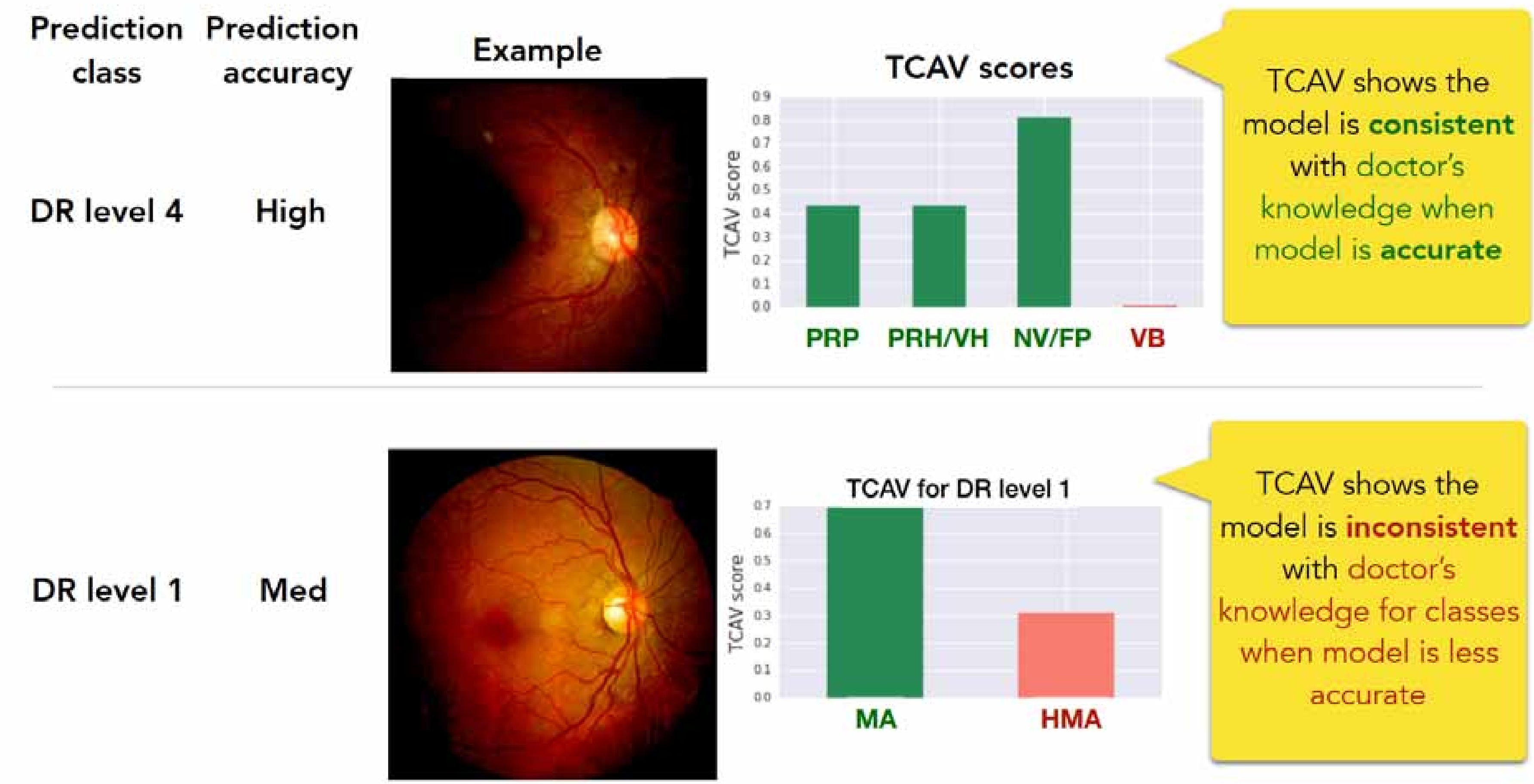
where $h_{l,k} : \mathbb{R}^m \to$
measure the sensitivi
to concepts at any mo
ric (e.g., unlike per-p
scalar quantity compu

3.4. Testing with CA

Testing with CAVs,

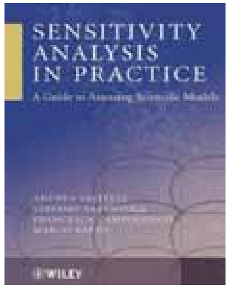


Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682. https://github.com/tensorflow/tcav



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.

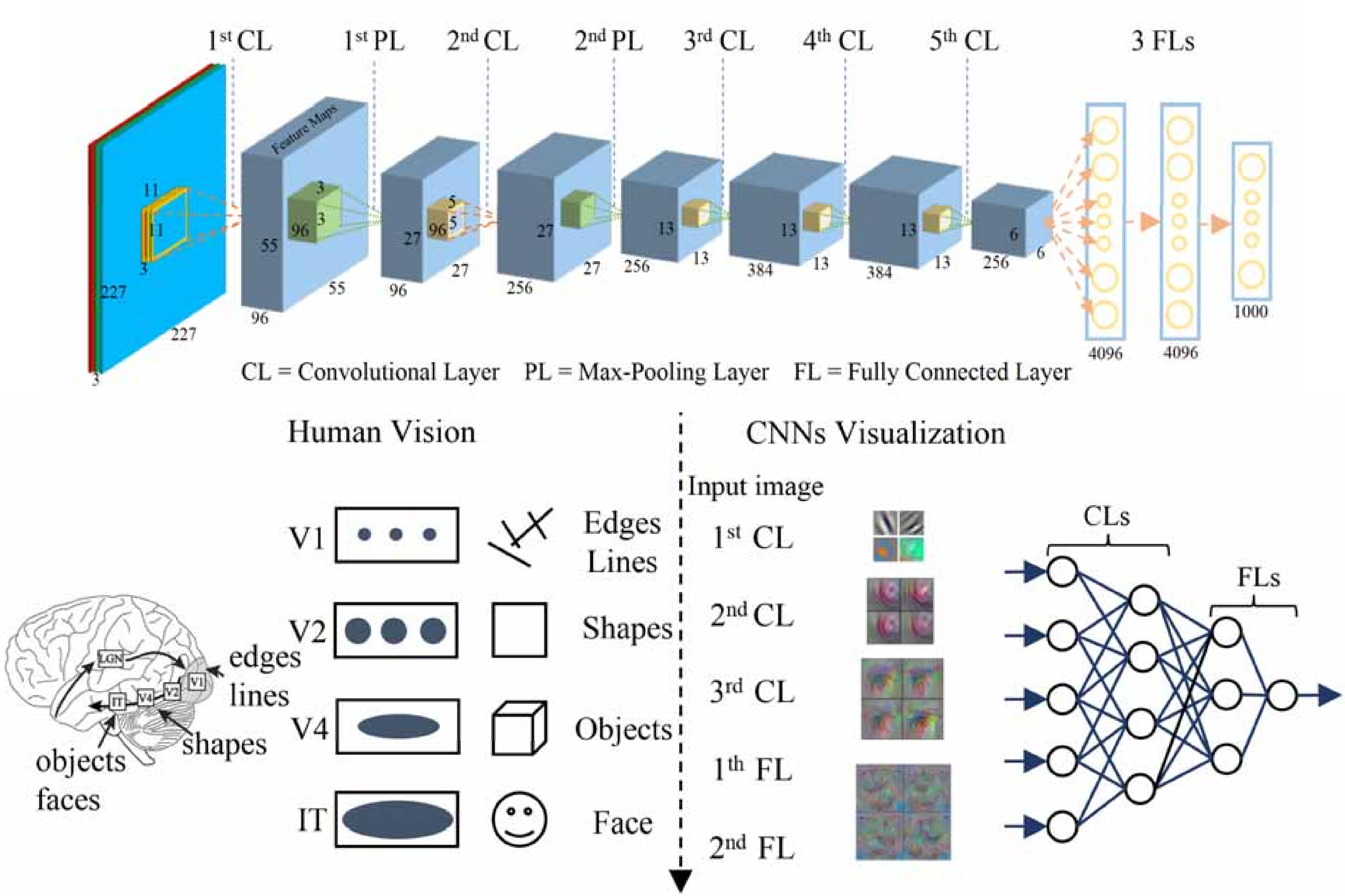
Digression: Sensitivity Analysis



- Sensitivity analysis (SA) is a classic, versatile and broad field with long tradition and can be used for a variety of different purposes, including:
  - Robustness testing (very important for ML)
  - Understanding the relationship between input and output
  - Studying and reducing uncertainty

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo & Marco Ratto 2004. Sensitivity analysis in practice: a guide to assessing scientific models. Chichester, England.

- Remember: NN=nonlinear function approximators using gradient descent to minimize the error in such a function approximation
- To students this seems to be “new” – but it has a long history:
  - Chain rule = back-propagation was invented by Leibniz (1676) and L’Hopital (1696)
  - Calculus and Algebra have long been used to solve optimization problems and gradient descent was introduced by Cauchy (1847)
  - This was used to fuel machine learning in the 1940ies > perceptron – but was limited to linear functions, therefore
  - Learning nonlinear functions required the development of a multilayer perceptron and methods to compute the gradient through such a model
  - This was elaborated by LeCun (1985), Parker (1985), Rumelhart (1986) and Hinton (1986)



Zhuwei Qin, Fuxun Yu, Chenchen Liu & Xiang Chen 2018. How convolutional neural network see the world-A survey of convolutional neural network visualization methods. *arXiv preprint arXiv:1804.11191*.

What are Saliency Maps?

Visualization of Neural Networks Using Saliency Maps

Niels J. S. Mørch+‡ Ulrik Kjems+ Lars Kai Hansen+
Claus Svarer‡ Ian Law‡ Benny Lautrup† Steve Strother♮ Kelly Rehm♮

+Electronics Institute
Technical University of Denmark
DK-2800 Lyngby, Denmark

†Niels Bohr Institute
University of Copenhagen
DK-2100 Copenhagen Ø, Denmark

‡Department of Neurology
National University Hospital, Rigshospitalet
DK-2100 Copenhagen Ø, Denmark

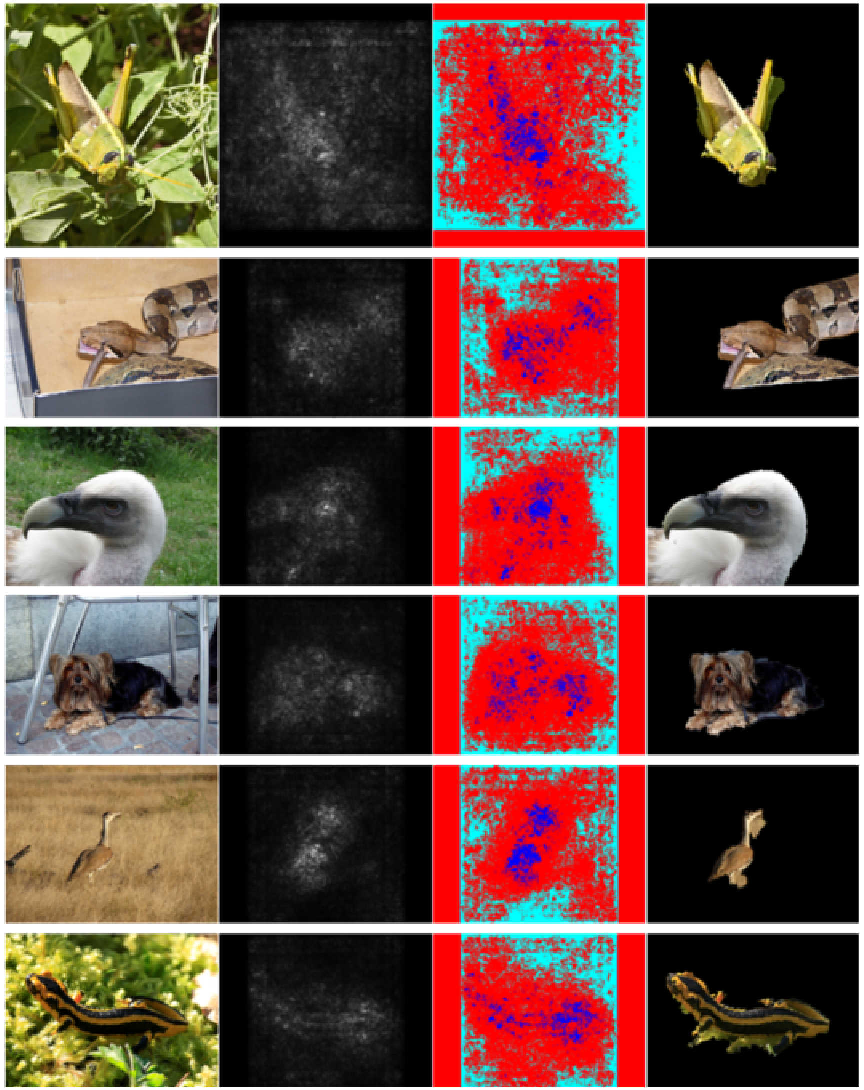
♮PET Imaging Service, Va Medical Center
Radiology and Health Informatics Depts.
University of Minnesota, Minneapolis
Minnesota, 55417, USA

E-Mail : nmorch@ei.dtu.dk

ABSTRACT

The saliency map is proposed as a new method for understanding and visualizing the non-linearities embedded in feed-forward neural networks, with emphasis on the ill-posed case, where the dimensionality of the input-field by far exceeds the number of examples. Several levels of approximations are discussed. The saliency maps are applied to medical imaging (PET-scans) for identification of paradigm-relevant regions in the human brain.
Keywords: saliency map, model interpretation, ill-posed learning, PCA, SVD, PET.

Niels J. S. Morch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Steve Strother & Kelly Rehm. Visualization of neural networks using saliency maps. Proceedings of ICNN'95-International Conference on Neural Networks, 1995 Perth (Australia). IEEE, 2085-2090, doi:10.1109/ICNN.1995.488997.



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

- Let us consider a function f,
- a data point $x = (x1, \dots, xd)$ and the prediction
- $f(x1, \dots, xd)$
- Now, SA measures the local variation of the function along each input dimension:
- $Ri = \left(\frac{\partial f}{\partial xi} | x = x\right)^2$
- With other words, SA produces local explanations for the prediction of a differentiable function f using the squared norm of its gradient w.r.t. the inputs x : $S(x) \, / \, krxfk2$.
- The saliency map S produced with this method describes the extent to which variations in the input would produce a change in the output $S(x) \propto \|\nabla_{\mathbf{x}} f\|^2$

Muriel Gevrey, Ioannis Dimopoulos & Sovan Lek 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, 160, (3), 249-264.

- Given an image classification (ConvNet), we aim to answer two questions:
  - What does a class model look like?
  - What makes an image belong to a class?
- To this end, we visualise:
  - Canonical image of a class
  - Class saliency map for a given image and class
- Both visualisations are based on the class score derivative w.r.t. the input image (computed using back-prop)

Thank you!

Glossar (1/3)

- Ante-hoc Explainability (AHE) = such models are interpretable by design, e.g. glass-box approaches; typical examples include linear regression, decision trees/lists, random forests, Naive Bayes and fuzzy inference systems; or GAMs, Stochastic AOGs, and deep symbolic networks; they have a long tradition and can be designed from expert knowledge or from data and are useful as framework for the interaction between human knowledge and hidden knowledge in the data.
- BETA = Black Box Explanation through Transparent Approximation, developed by Lakkarju, Bach & Leskovec (2016) it learns two-level decision sets, where each rule explains the model behaviour; this is an increasing problem in daily use of AI/ML, see e.g. http://news.mit.edu/2019/better-fact-checking-fake-news-1017
- Bias = inability for a ML method to represent the true relationship; High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting);
- Causability = is a property of a human (natural intelligence) and a measurement for the degree of human understanding; we have developed a causability measurement scale (SCS).
- Decomposition = process of resolving relationships into the consituent components (hopefully representing the relevant interest). Highly theoretical, because in real-world this is hard due to the complexity (e.g. noise) and untraceable imponderabilities on our observations.
- Deduction = deriving of a conclusion by reasoning
- Explainability = motivated by the opaqueness of so called “black-box” approaches it is the ability to provide an explanation on why a machine decision has been reached (e.g. why is it a cat what the deep network recognized). Finding an appropriate explanation is difficult, because this needs understanding the context and providing a description of causality and consequences of a given fact. (German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz)

Glossar (2/3)

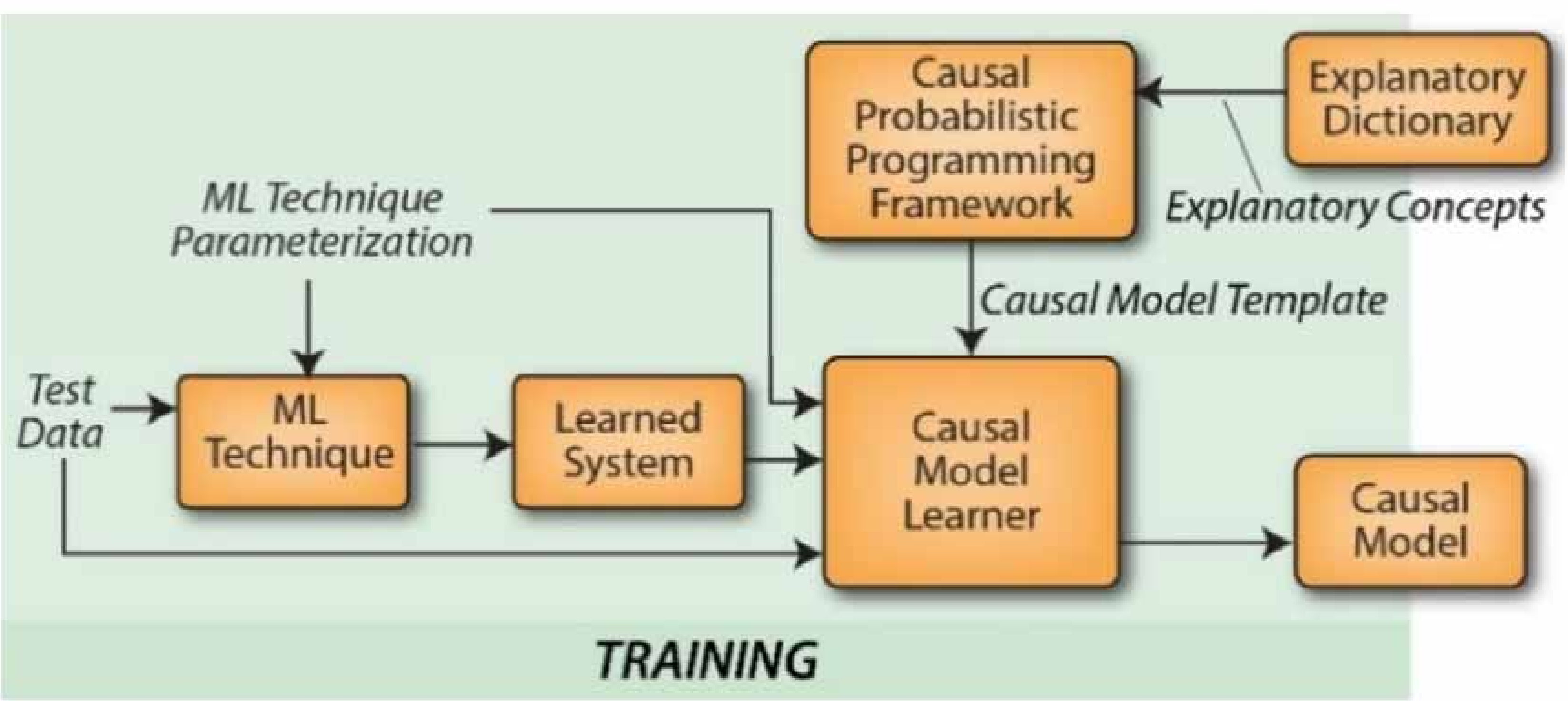
- Explanation = set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questionss (“Why is this a cat?”). (German: Erklärung, Begründung)
- Explanation = set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questionss (“Why is this a cat?”). (German: Erklärung, Begründung)
- Explanatory power = is the ability of a set hypothesis to effectively explain the subject matter it pertains to (opposite: explanatory impotence).
- Explicit Knowledge = you can easy explain it by articulating it via natural language etc. and share it with others.
- European General Data Protection Regulation (EU GDPR) = Regulation EU 2016/679 – see the EUR-Lex 32016R0679 , will make black-box approaches difficult to use, because they often are not able to explain why a decision has been made (see explainable AI).
- Gaussian Process (GP) = collection of stochastic variables indexed by time or space so that each of them constitute a multidimensional Gaussian distribution; provides a probabilistic approach to learning in kernel machines (See: Carl Edward Rasmussen & Christopher K.I. Williams 2006. Gaussian processes for machine learning, Cambridge (MA), MIT Press); this can be used for explanations. (see also: Visual Exploration Gaussian)
- Gradient = a vector providing the direction of maximum rate of change.
- Ground truth = generally information provided by direct observation (i.e. empirical evidence) instead of provided by inference. For us it is the gold standard, i.e. the ideal expected result (100 % true);

Glossar (3/3)

- Interactive Machine Learning (iML) = machine learning algorithms which can interact with – partly human – agents and can optimize its learning behaviour trough this interaction. Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics (BRIN), 3, (2), 119-131.
- Inverse Probability = an older term for the probability distribution of an unobserved variable, and was described by De Morgan 1837, in reference to Laplace's (1774) method of probability.
- Implicit Knowledge = very hard to articulate, we do it but cannot explain it (also tacit knowlege).
- Kernel = class of algorithms for pattern analysis e.g. support vector machine (SVM); very useful for explainable AI
- Kernel trick = transforming data into another dimension that has a clear dividing margin between the classes
- Multi-Agent Systems (MAS) = include collections of several independent agents, could also be a mixture of computer agents and human agents. An excellent pointer of the later one is: Jennings, N. R., Moreau, L., Nicholson, D., Ramchurn, S. D., Roberts, S., Rodden, T. & Rogers, A. 2014. On human-agent collectives. Communications of the ACM, 80-88.
- Post-hoc Explainability (PHE) = such models are designed for interpreting black-box models and provide local explanations for a specific decision and re-enact on request, typical examples include LIME, BETA, LRP, or Local Gradient Explanation Vectors, prediction decomposition or simply feature selection.
- Preference learning (PL) = concerns problems in learning to rank, i.e. learning a predictive preference model from observed preference information, e.g. with label ranking, instance ranking, or object ranking. Fürnkranz, J., Hüllermeier, E., Cheng, W. & Park, S.-H. 2012. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. Machine Learning, 89, (1-2), 123-156.
- Saliency map = image showing in a different representation (usually easier for human perception) each pixel's quality.
- Tacit Knowledge = Knowledge gained from personal experience that is even more difficult to express than implicit knowledge.
- Transfer Learning (TL) = The ability of an algorithm to recognize and apply knowledge and skills learned in previous tasks to novel tasks or new domains, which share some commonality. Central question: Given a target task, how do we identify the commonality between the task and previous tasks, and transfer the knowledge from the previous tasks to the target one? Pan, S. J. & Yang, Q. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/tkde.2009.191.

DARPA Causal Model Induction (CRA)

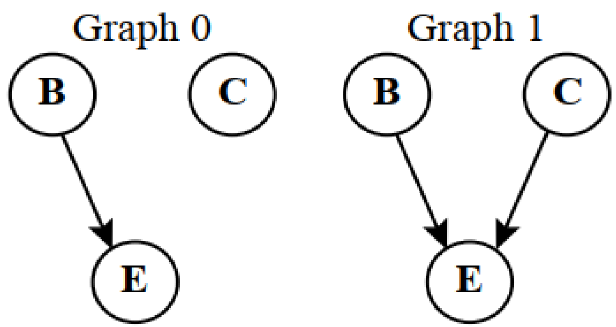
https://www.darpa.mil/attachments/XAIProgramUpdate.pdf



(a)

	c^+	c^-
e^+	$N(e^+, c^+)$	$N(e^+, c^-)$
e^-	$N(e^-, c^+)$	$N(e^-, c^-)$

(b)



Michael D Pacer & Thomas L Griffiths. A rational model of causal induction with continuous causes. Proceedings of the 24th International Conference on Neural Information Processing Systems, 2011. Curran Associates Inc., 2384-2392.

Computer Science > Machine Learning

iNNvestigate neural networks!

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, Pieter-Jan Kindermans

(Submitted on 13 Aug 2018)

In recent years, deep neural networks have revolutionized many application domains of machine learning and are key components of many critical decision or predictive processes. Therefore, it is crucial that domain specialists can understand and analyze actions and pre- dictions, even of the most complex neural network architectures. Despite these arguments neural networks are often treated as black boxes. In the attempt to alleviate this short- coming many analysis methods were proposed, yet the lack of reference implementations often makes a systematic comparison between the methods a major effort. The presented library iNNvestigate addresses this by providing a common interface and out-of-the- box implementation for many analysis methods, including the reference implementation for PatternNet and PatternAttribution as well as for LRP-methods. To demonstrate the versatility of iNNvestigate, we provide an analysis of image classifications for variety of state-of-the-art neural network architectures.

Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)

Cite as: arXiv:1808.04260 [cs.LG]

(or arXiv:1808.04260v1 [cs.LG] for this version)

Bibliographic data

Select data provider: **Semantic Scholar** | Prophy [Disable Bibex(What is Bibex?)]

References (28) Citations (20)

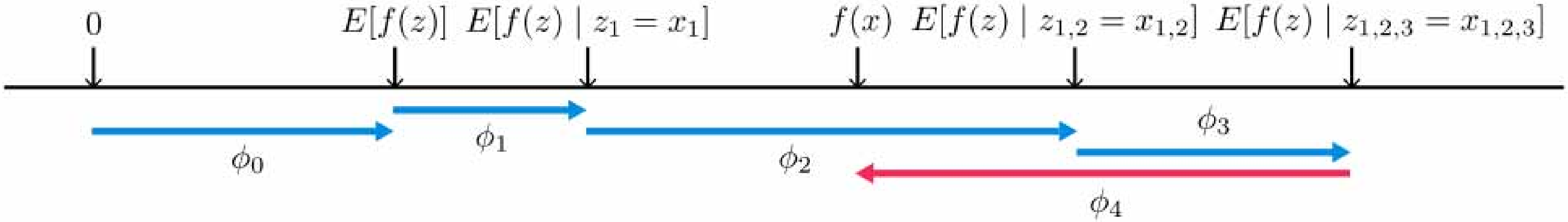
https://github.com/albermax/innvestigate

https://github.com/sebastian-lapuschkin/lrp_toolbox

https://github.com/ArrasL/LRP_for_LSTM

Also Explore:
https://innvestigate.readthedocs.io/en/latest/modules/analyzer.html#module-innvestigate.analyzer.relevance_based.relevance_analyzer

Alernatively: (SHapley Additive exPlanations)



Theorem 2 (Shapley kernel) *Under Definition 1, the specific forms of $\pi_{x'}$, L, and Ω that make solutions of Equation 2 consistent with Properties 1 through 3 are:*

$$\Omega(g) = 0,$$

$$\pi_{x'}(z') = \frac{(M-1)}{(M \textit{ choose } |z'|)|z'|(M - |z'|)},$$

$$L(f, g, \pi_{x'}) = \sum_{z' \in Z} \left[f(h_x(z')) - g(z')\right]^2 \pi_{x'}(z'),$$

where $|z'|$ is the number of non-zero elements in z'.

Scott M. Lundberg & Su-In Lee. A unified approach to interpreting model predictions. *In:* Guyon, Isabelle, Luxburg, Ulrike Von, Bengio, Samy, Wallach, Hanna, Fergus, Rob, Viswanathan, Svn & Garnett, Roman, eds. Advances in Neural Information Processing Systems, 2017 Montreal. NIPS, 4765-4774.

https://github.com/OpenXAIProject/PyConKorea2019-Tutorials