Seminar Explainable AI

Module 4

Overview on Explainable AI Methods

Birds-Eye View

Global-Local-Ante-hoc-Post-hoc

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)

Institute for Medical Informatics/Statistics, Medical University Graz, Austria

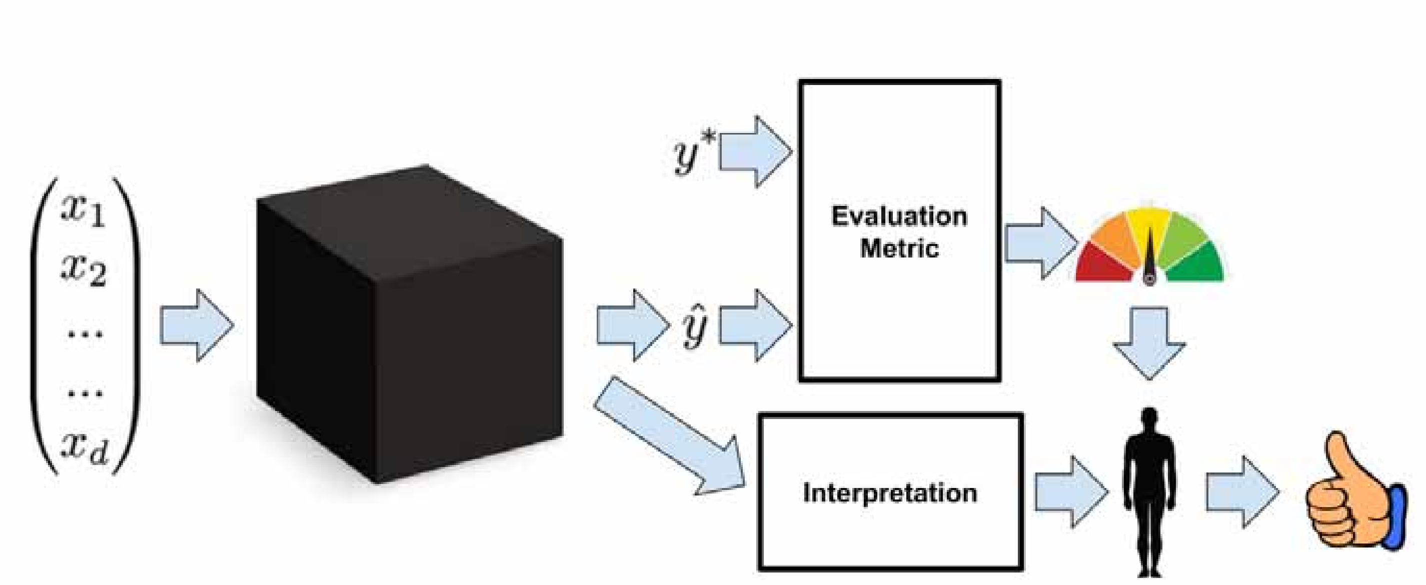
and

Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

- 00 Reflection – follow-up from last lecture
- 01 Basics, Definitions, …
- 02 Please note: xAI is not new!
- 03 Examples for Ante-hoc models (explainable models, interpretable machine learning)
- 04 Examples for Post-hoc models (making the “black-box” model interpretable
- 05 Explanation Interfaces: Future human-AI interaction
- 06 A few words on metrics of xAI (measuring causability)

01 Basics, Definitions, …

Explainability/ Interpretability – contrasted to performance



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

- Inconsistent Definitions: What is the difference between explainable, interpretable, verifiable, intelligible, transparent, understandable … ?

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

- **Trust** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?
- **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

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

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)

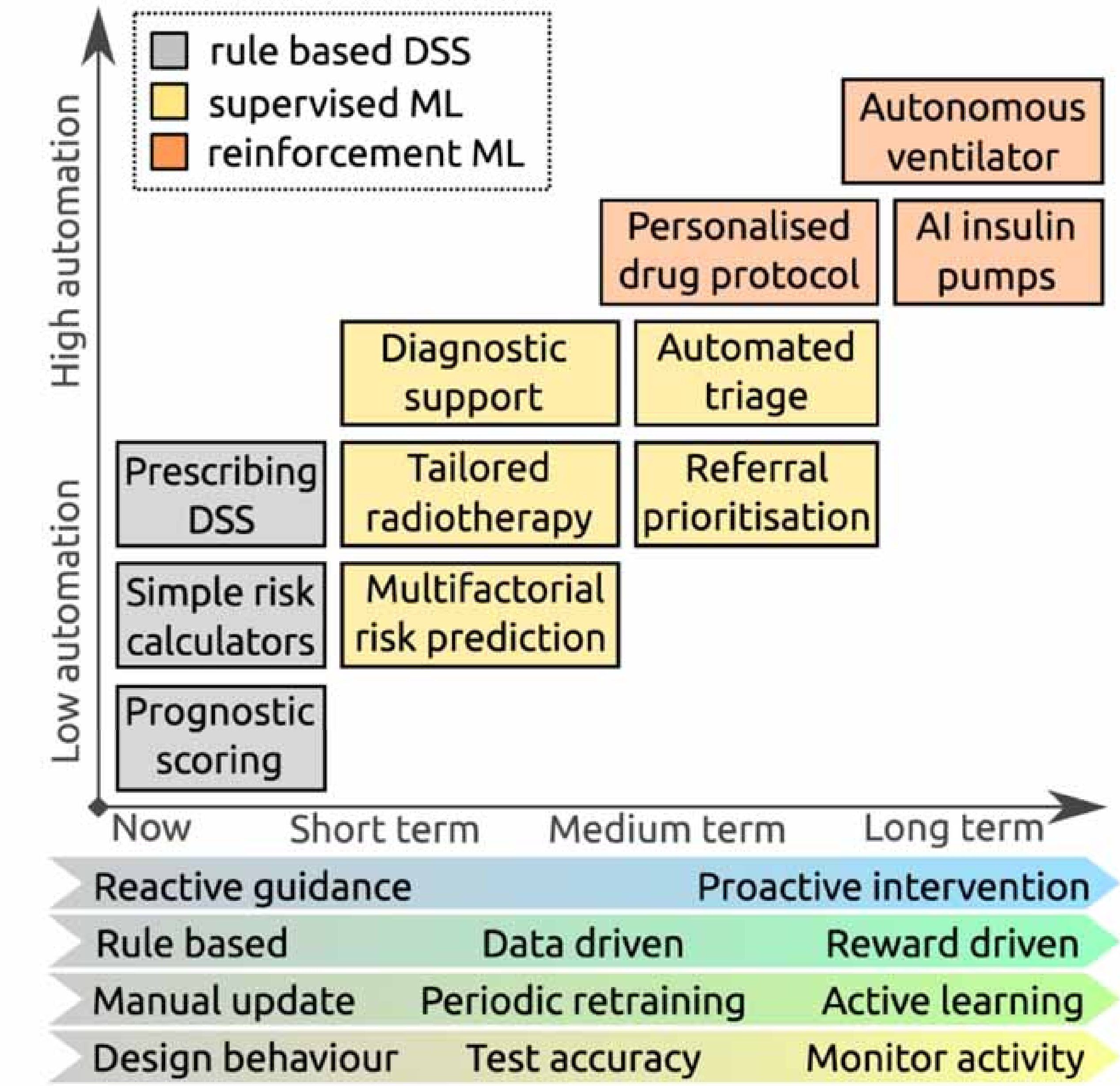
- Explanation = set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questionss (“Why is this a cat?”). (German: Erklärung, Begründung)
- Explanation = set of statements to describe a given 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.

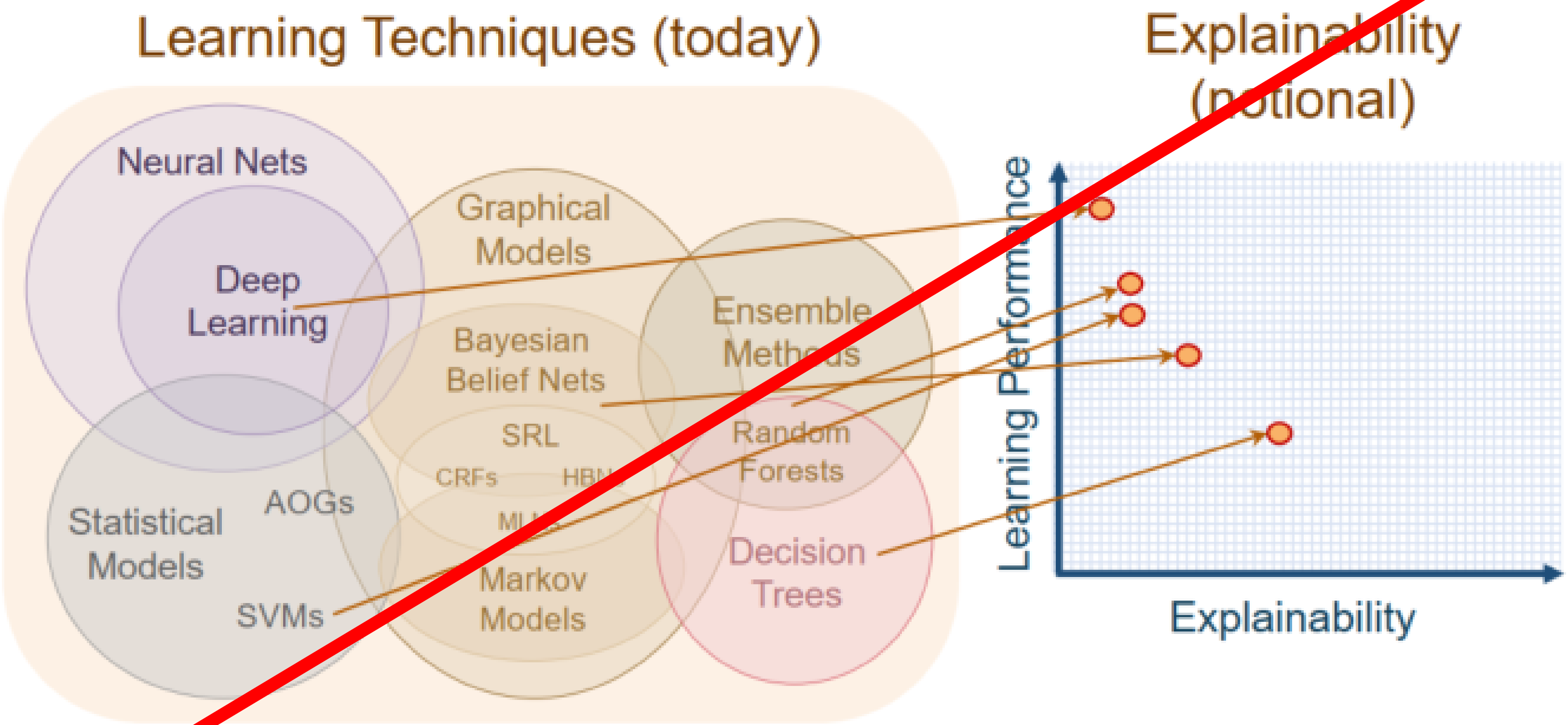
Expected trends in medical AI

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards & Krasimira Tsaneva-Atanasova 2019. Artificial intelligence, bias and clinical safety. BMJ Quality and Safety, 28, (3), 231-237, doi:10.1136/bmjqs-2018-008370.



- **Interpretable 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:
  - 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:
  - Not easy to interpret ("black box")
  - Needs a lot of top-quality training data
  - Can well generalize even from high-dimensional data, with noise and good for images
  - Needs previous knowledge

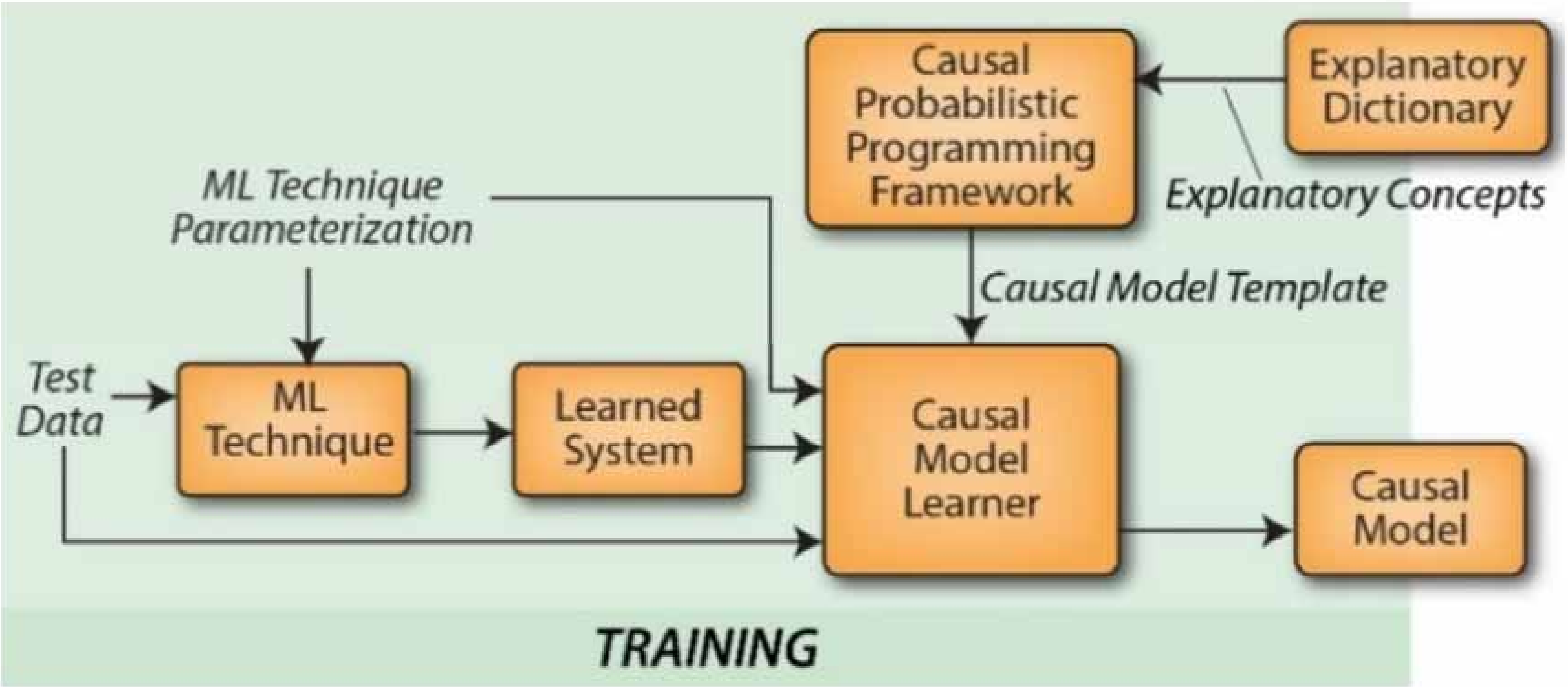


https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

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

DARPA Causal Model Induction (CRA)

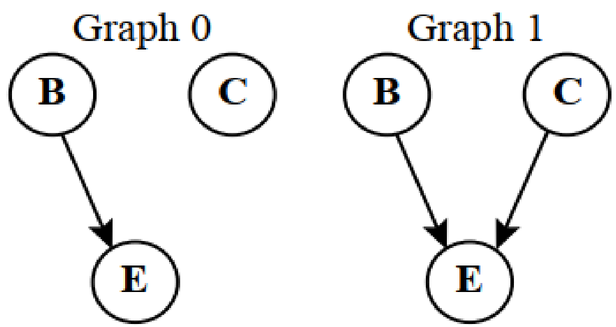
https://www.darpa.mil/attachments/XAIProgramUpdate.pdf



(a)

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

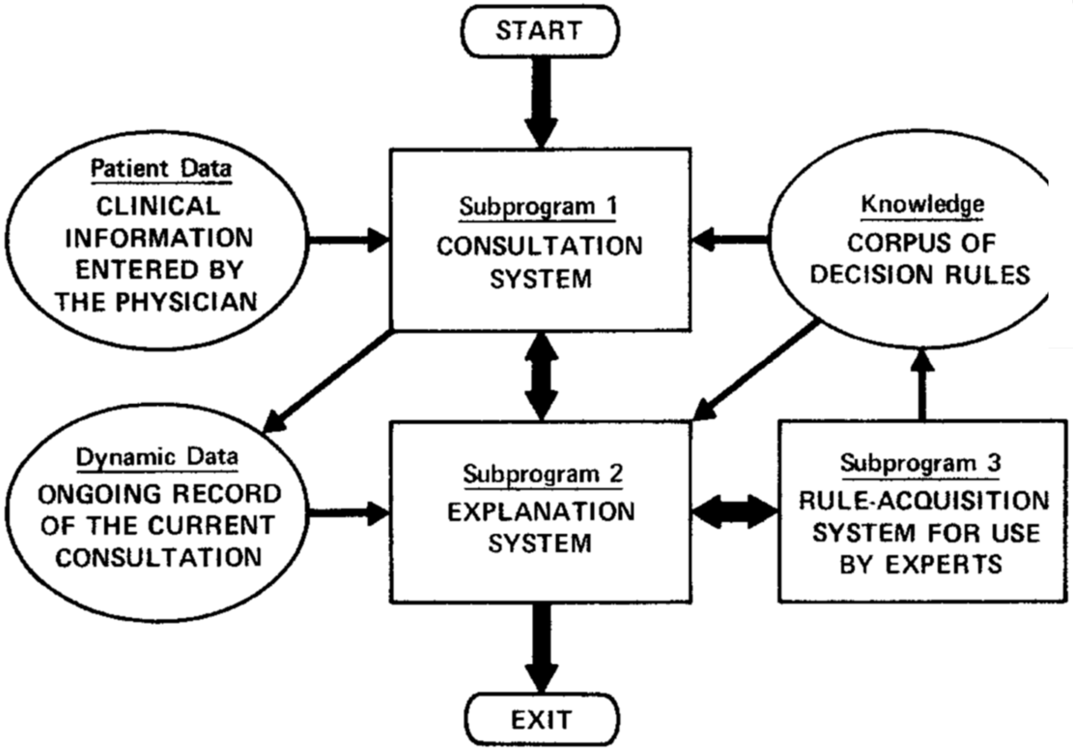
(b)



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.

02 Please note: xAI is not new !

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



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

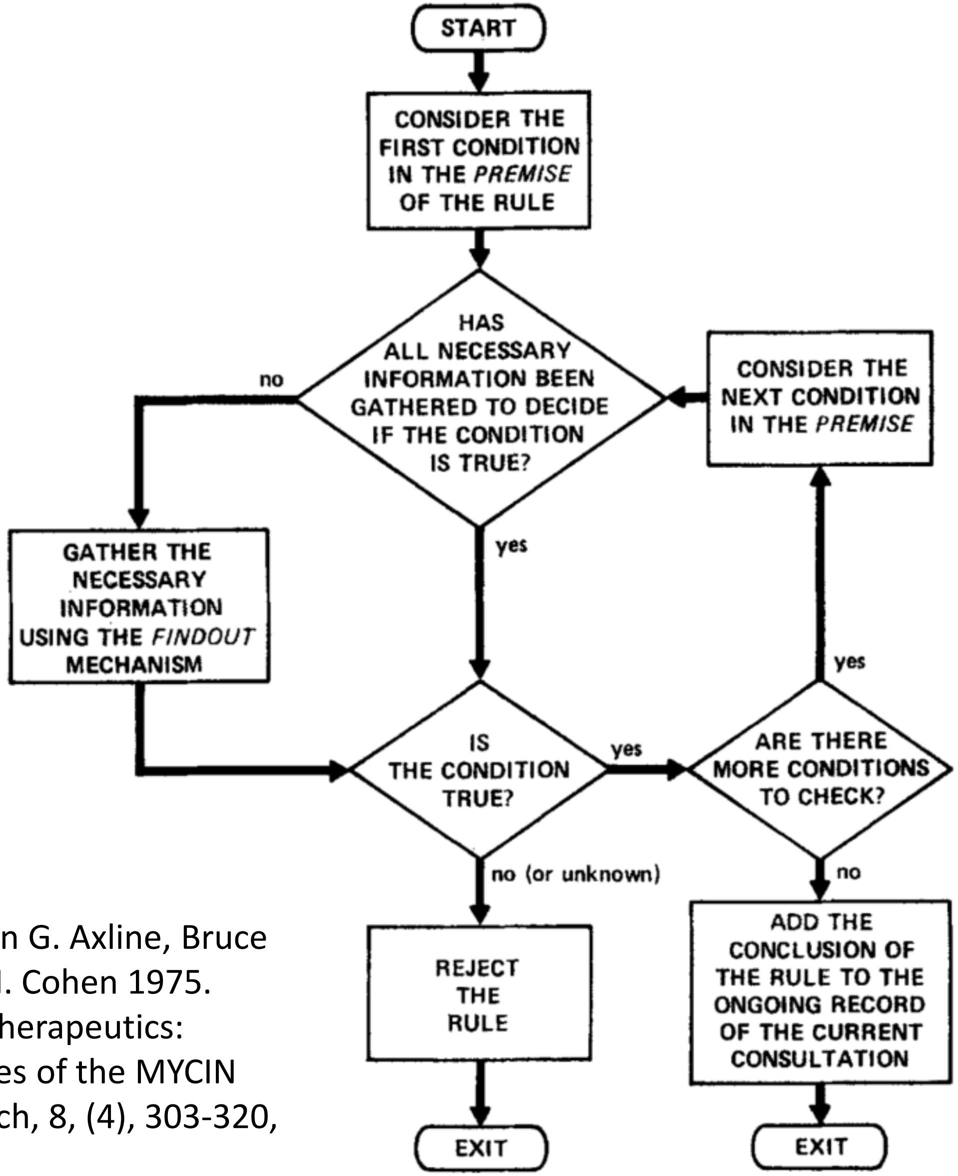
Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System*

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.



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.

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.

Howard L Bleich 1971. The computer as a consultant. New England Journal of Medicine, 284, (3), 141-147.

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 anatomically or temporally, 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.

- AI is actually the oldest field of computer science, aiming at solving task where humans are good (e.g. speech, vision, problem solving, ...)
- Note that while the first goal was to mimic human intelligence across tasks, the success today is only on very narrow AI e.g. solving one specific task, playing a game, driving a car, classifying objects, … due to the advancements in “deep learning” this works well,
- But the best performing methods remain opaque, i.e. are considered as so-called “black-box” models

- 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 (discussion of "bias, fairness, see Module 09)
- Black box Type 1 = too complicated for a human to understand
- Black box Type 2 = proprietary = "secret algorithm"

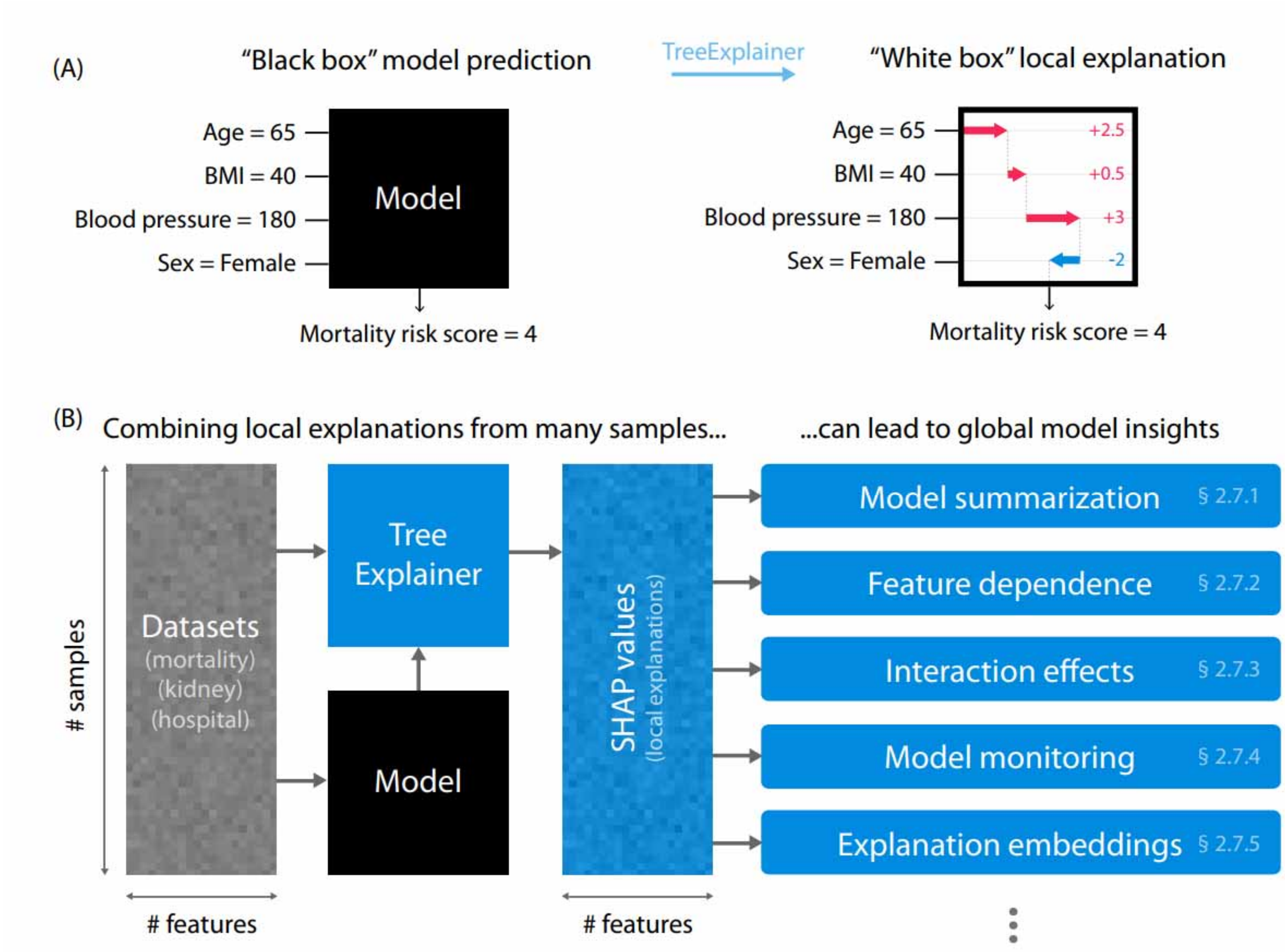
Cynthia Rudin, Caroline Wang & Beau Coker 2018. The age of secrecy and unfairness in recidivism prediction. *arXiv:1811.00731*.

- A black box model could be either
  - (1) a function that is too complicated for any human to comprehend or
  - (2) a function that is proprietary

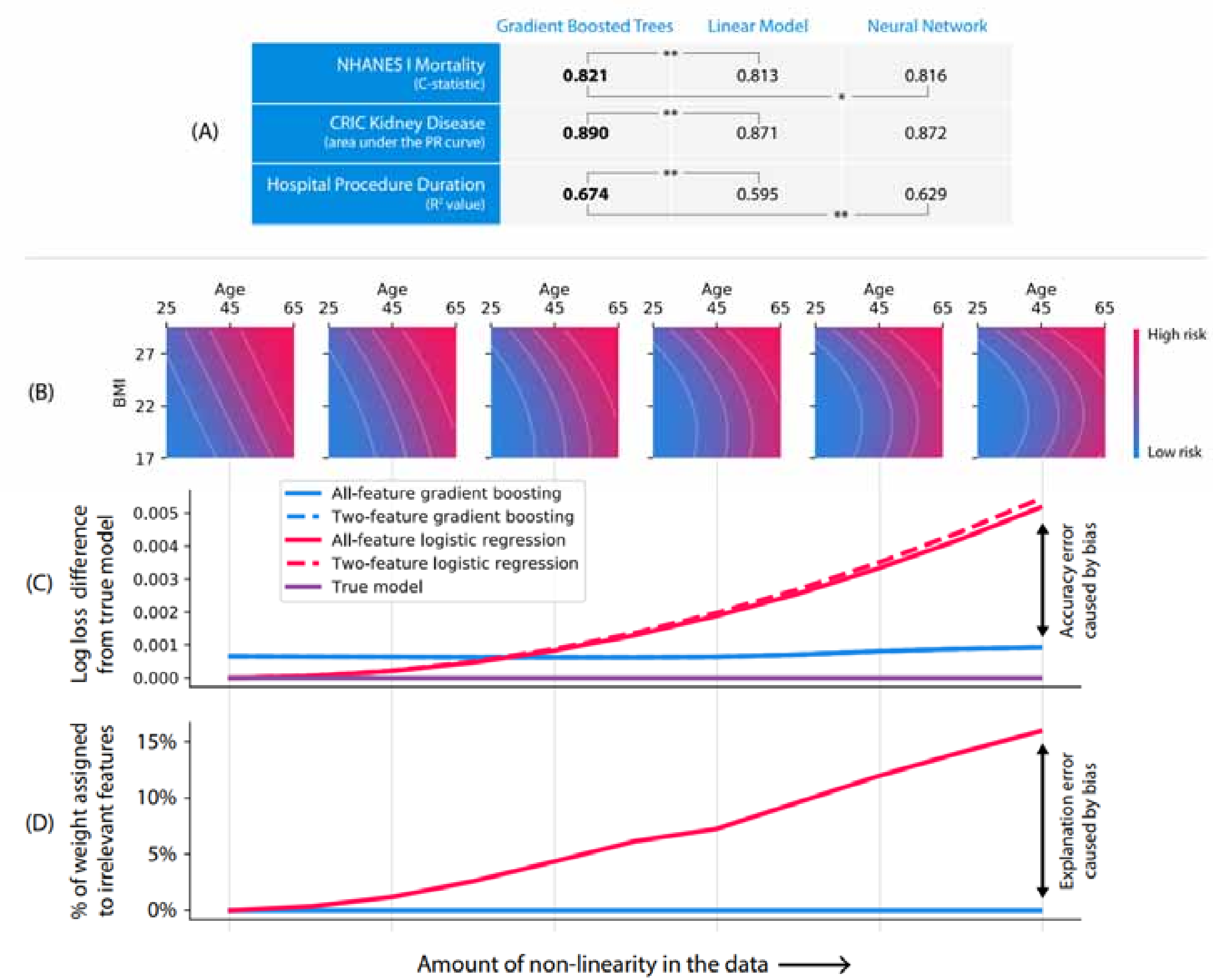
Cynthia Rudin 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, (5), 206-215, doi:10.1038/s42256-019-0048-x.

- **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.

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*.



Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex Degrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal & Su-In Lee 2019. Explainable ai for trees: From local explanations to global understanding. arXiv:1905.04610.

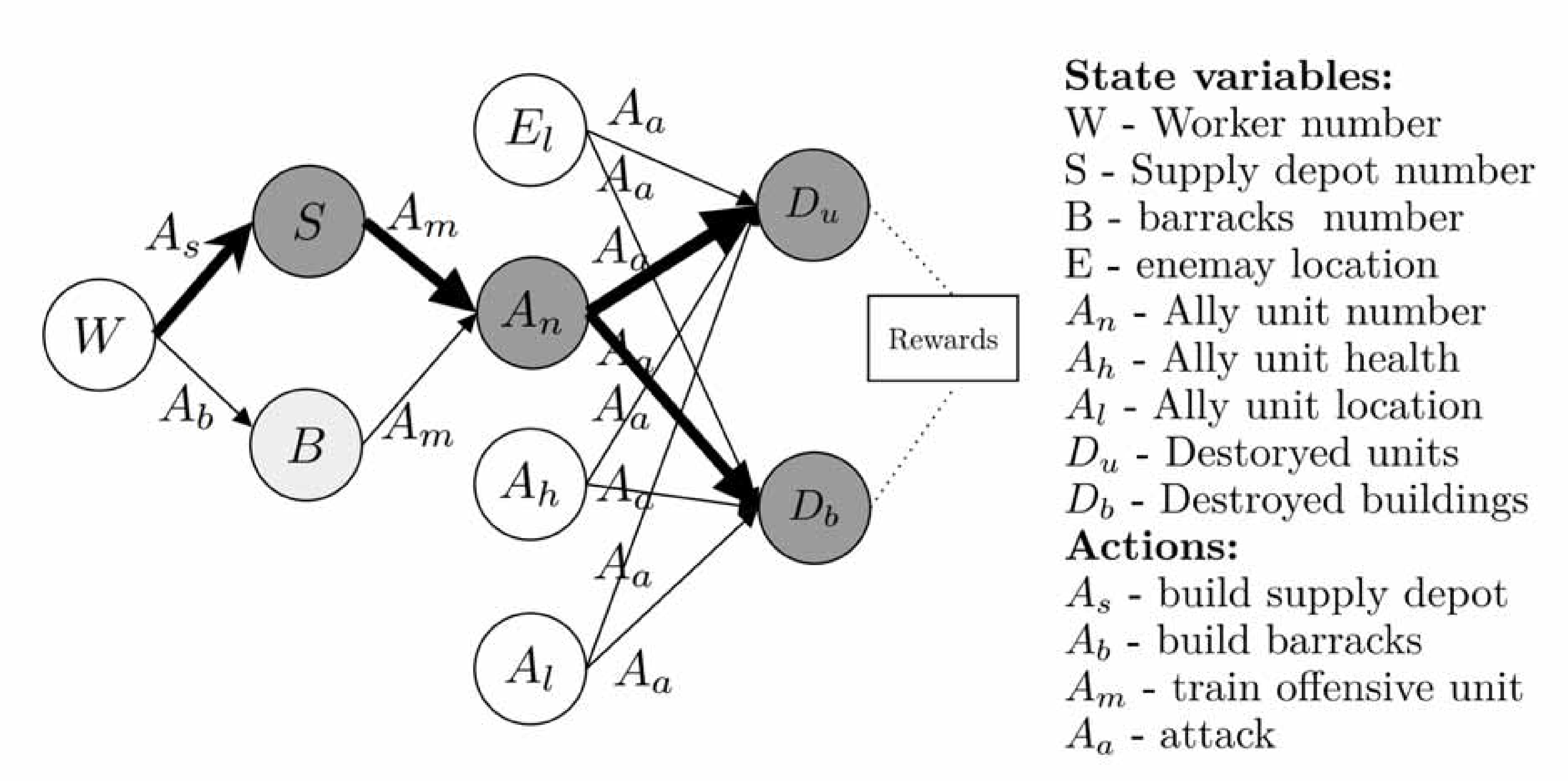


Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex Degrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal & Su-In Lee 2019. Explainable ai for trees: From local explanations to global understanding. arXiv:1905.04610.

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 (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*.



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

Example for an Action Influence Graph of a Starcraft agent

Algorithm 1 Task Prediction:Action Influence Model

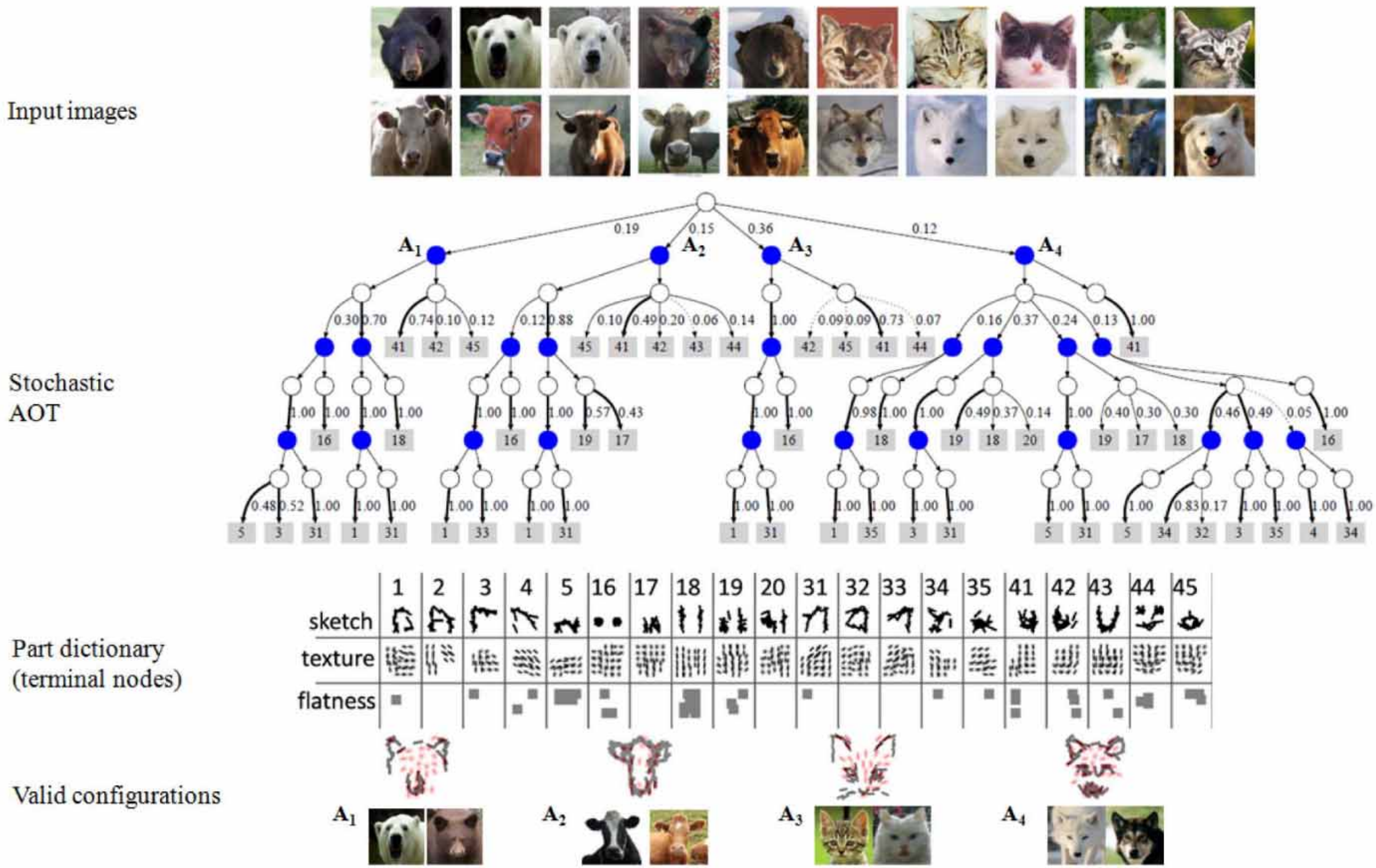
Input: trained regression models $\mathcal{L}$, current state S_t
Output: predicted action a

$\vec{F_p} \leftarrow []$; vector of predicted difference
for every $\widehat{L} \in \mathcal{L}$ **do**
    $P_y \leftarrow \widehat{L} \cdot predict(S_{x.t})$; predict variable S_y at S_{t+1}
    $\vec{F_p} \leftarrow |S_y - P_y|$; difference with actual S_y value
end for
return $max\left(\vec{F_p}\right) \cdot getAction()$

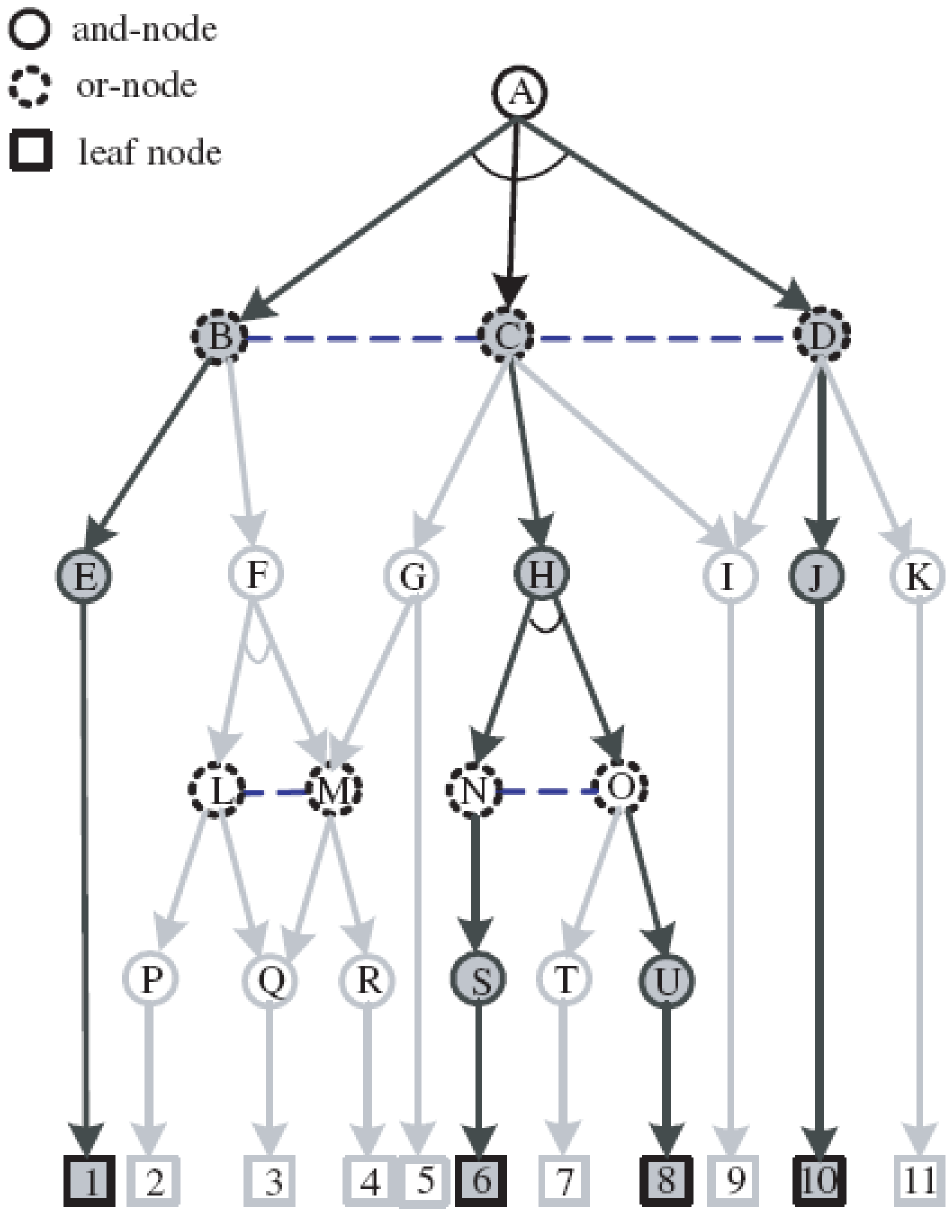
Env - RL		Accuracy (%)			Performance (s)		
	Size	LR	DT	MLP	LR	DT	MLP
Cartpole-PG	4/2	83.8	81.6	86.0	0.007	0.018	0.03
MountainCar-DQN	3/3	69.7	57.8	69.6	0.020	0.037	0.32
Taxi-SARSA	4/6	68.2	74.2	67.9	0.001	0.001	0.49
LunarLander-DDQN	8/4	68.4	63.7	72.1	0.002	0.002	0.33
BipedalWalker-PPO	14/4	56.9	56.4	56.7	0.010	0.015	0.41
Starcraft-A2C	9/4	94.7	91.8	91.4	0.144	0.025	3.33

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

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.



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

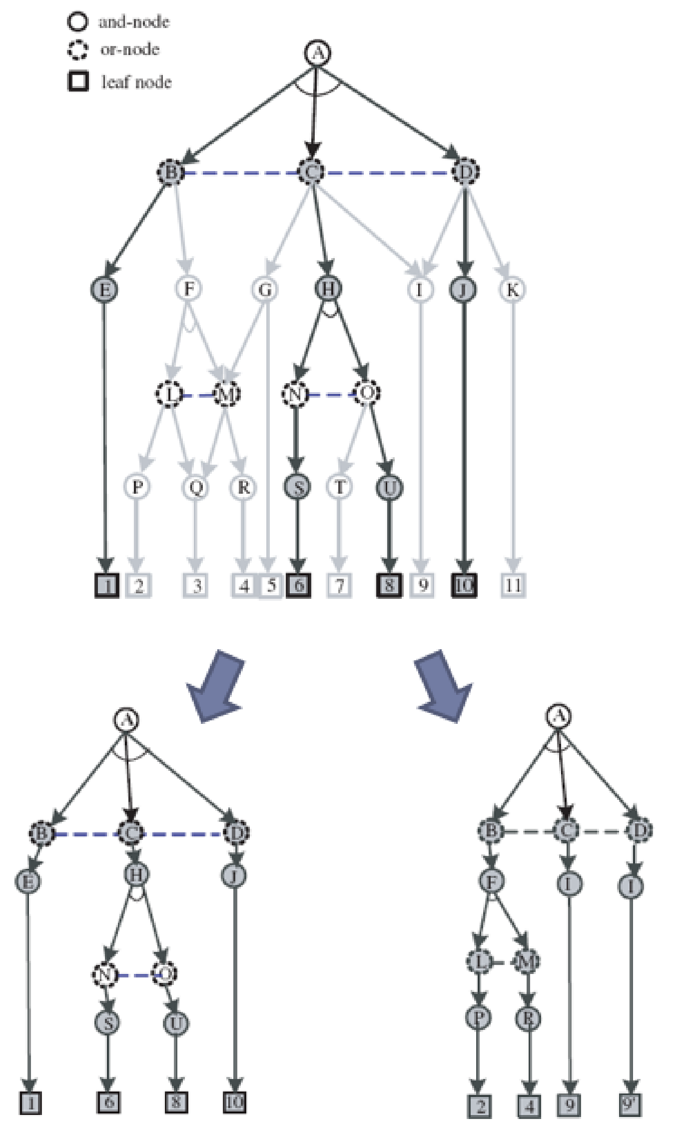
Images credit to Zhaoyin Jia (2009)

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

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

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$



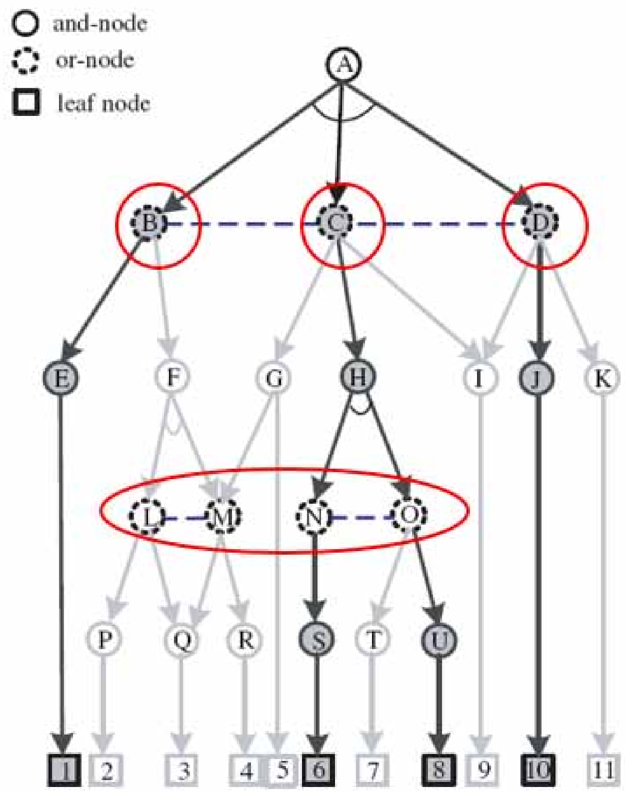
- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

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

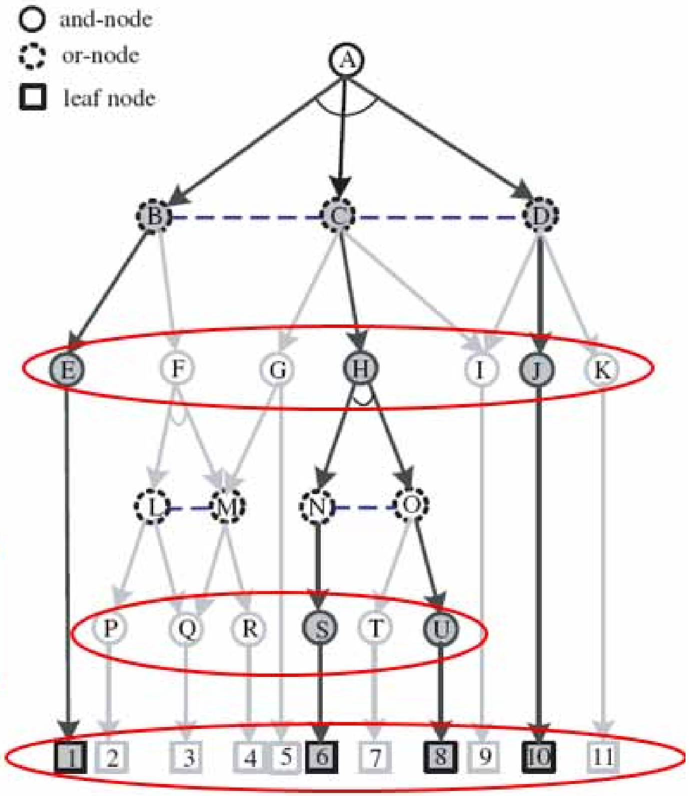
$$\xi(pg)=\boxed{\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))}+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$

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



- 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: $\alpha(t)$



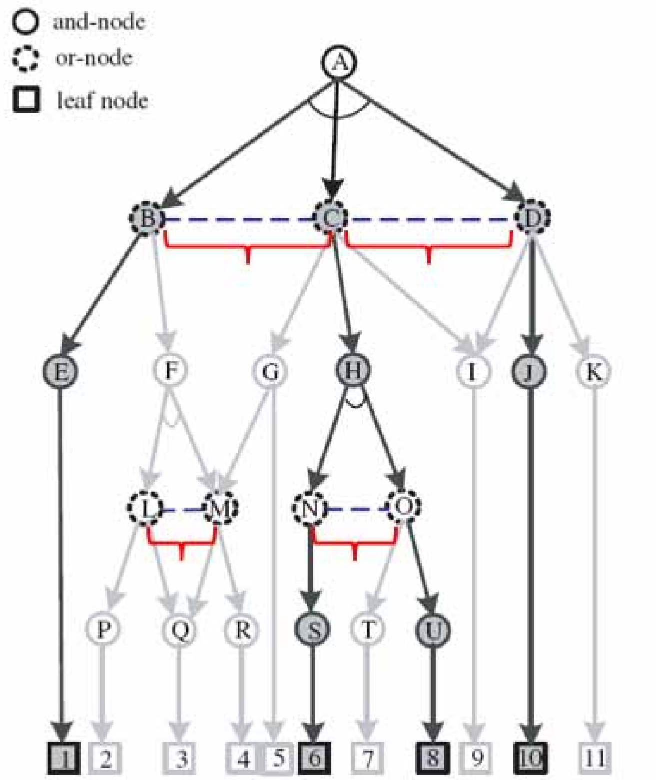
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\boxed{\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))}$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$

Weigh the local compatibility of primitives (geometric and appearance)

- 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: $\alpha(t)$



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

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\boxed{\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})}$$

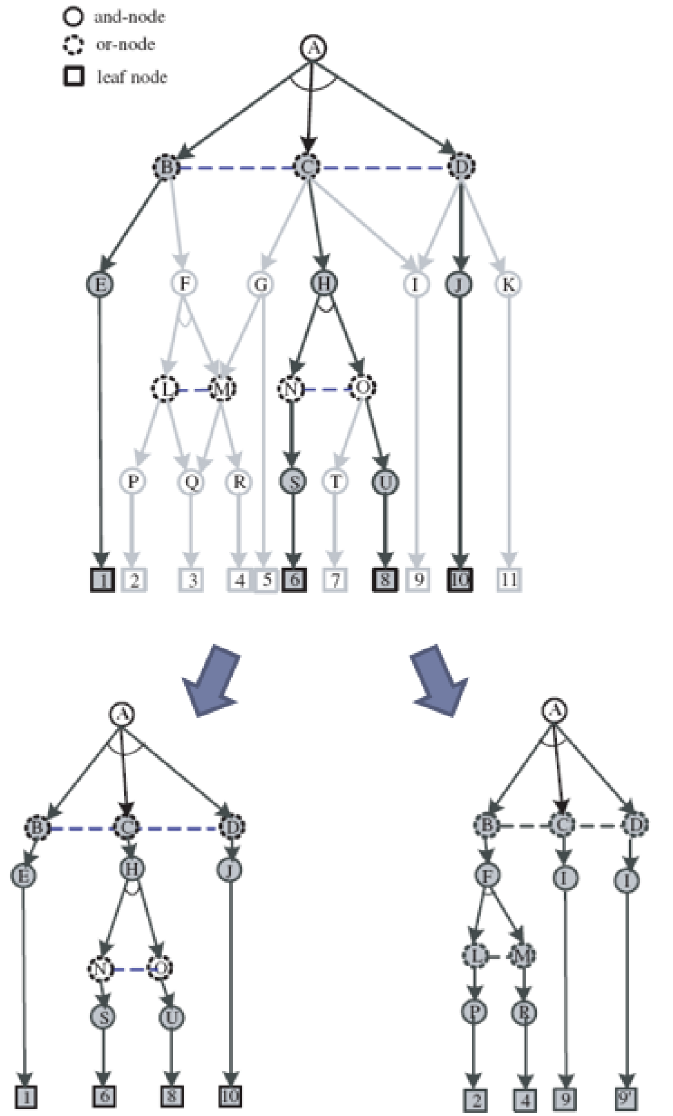
Spatial and appearance between primitives (parts or objects)

- 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: $\alpha(t)$

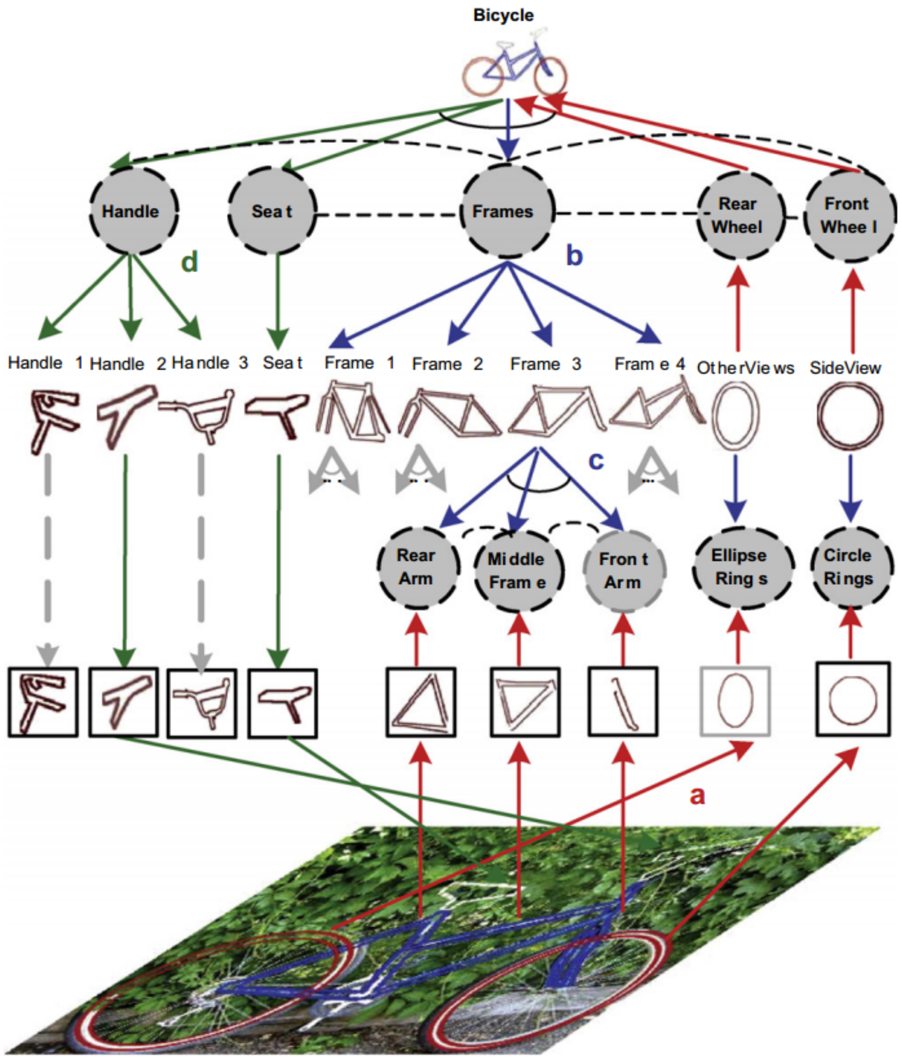
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$



Stochastic graph grammar/comp. object representation



Input: an input image **I**, and a set of constructed And-Or graphs of compositional object categories.
Output: a parsing graph pg, of the scene that consists of the parsing graphs of detected objects.

- Repeat the following steps

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

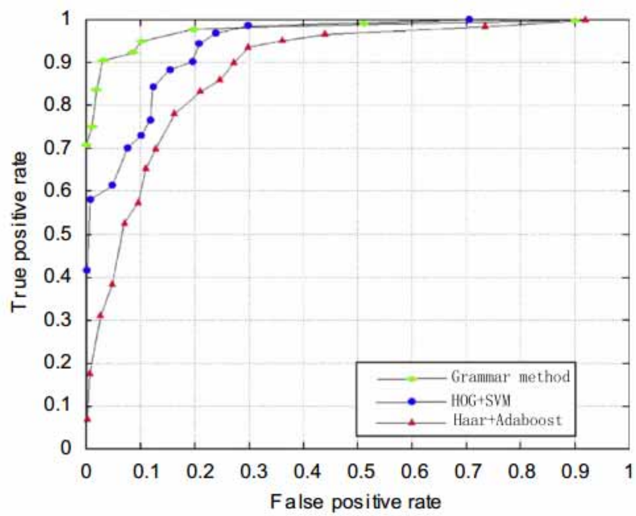
2 Call Bottom-up(A) to update A's **open** list.

 i Detect terminal instances of A from the image.
 ii Bind non-terminal instances of A from its children's **open** or **closed** lists

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

 i Accept hypotheses from A's **open** list to its **closed** list.
 ii Remove (or disassemble) hypotheses from A's **closed** list.
 iii Update the **open** lists for particles that overlap with node A.

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

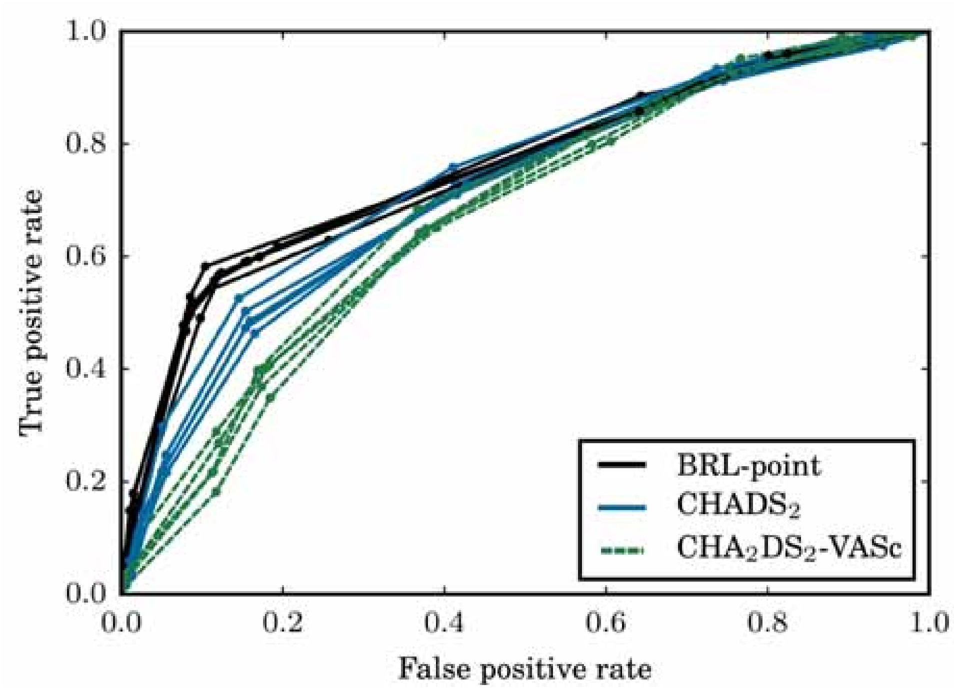


Liang Lin, Tianfu Wu, Jake Porway & Zijian Xu 2009. A stochastic graph grammar for compositional object representation and recognition. Pattern Recognition, 42, (7), 1297-1307, doi:10.1016/j.patcog.2008.10.033.

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.

Example: Interpretable Decision Sets

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

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.

Algorithm 1 Smooth Local Search (SLS) [18]

1: **Input:** Objective f, domain $X = \mathcal{S} \times \mathcal{C}$, parameters δ and δ'
2:
3: $A = \emptyset$
4: $\texttt{OPT} = f(\Phi_X(X, 0))$
5: **for** each element $x \in X$ **do**
6: Estimate $\mathbb{E}[f(\Phi_X(A, \delta) \cup x)] - \mathbb{E}[f(\Phi_X(A, \delta) \setminus x)]$ within an error of $\frac{1}{|X|^2}\texttt{OPT}$
7: Call this estimate $\tilde{\omega}_{A,\delta}(x)$
8: **end for**
9: **for** each element $x \in X \setminus A$ such that $\tilde{\omega}_{A,\delta}(x) > \frac{2}{|X|^2}\texttt{OPT}$ **do**
10: $A = A \cup x$
11: Goto Line 5
12: **end for**
13: **for** each element $x \in A$ such that $\tilde{\omega}_{A,\delta}(x) < \frac{-2}{|X|^2}\texttt{OPT}$ **do**
14: $A = A \setminus x$
15: Goto Line 5
16: **end for**
17: **return** $\Phi_X(A, \delta')$

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, … (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 (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 !

HCAI HUMAN-CENTERED.AI

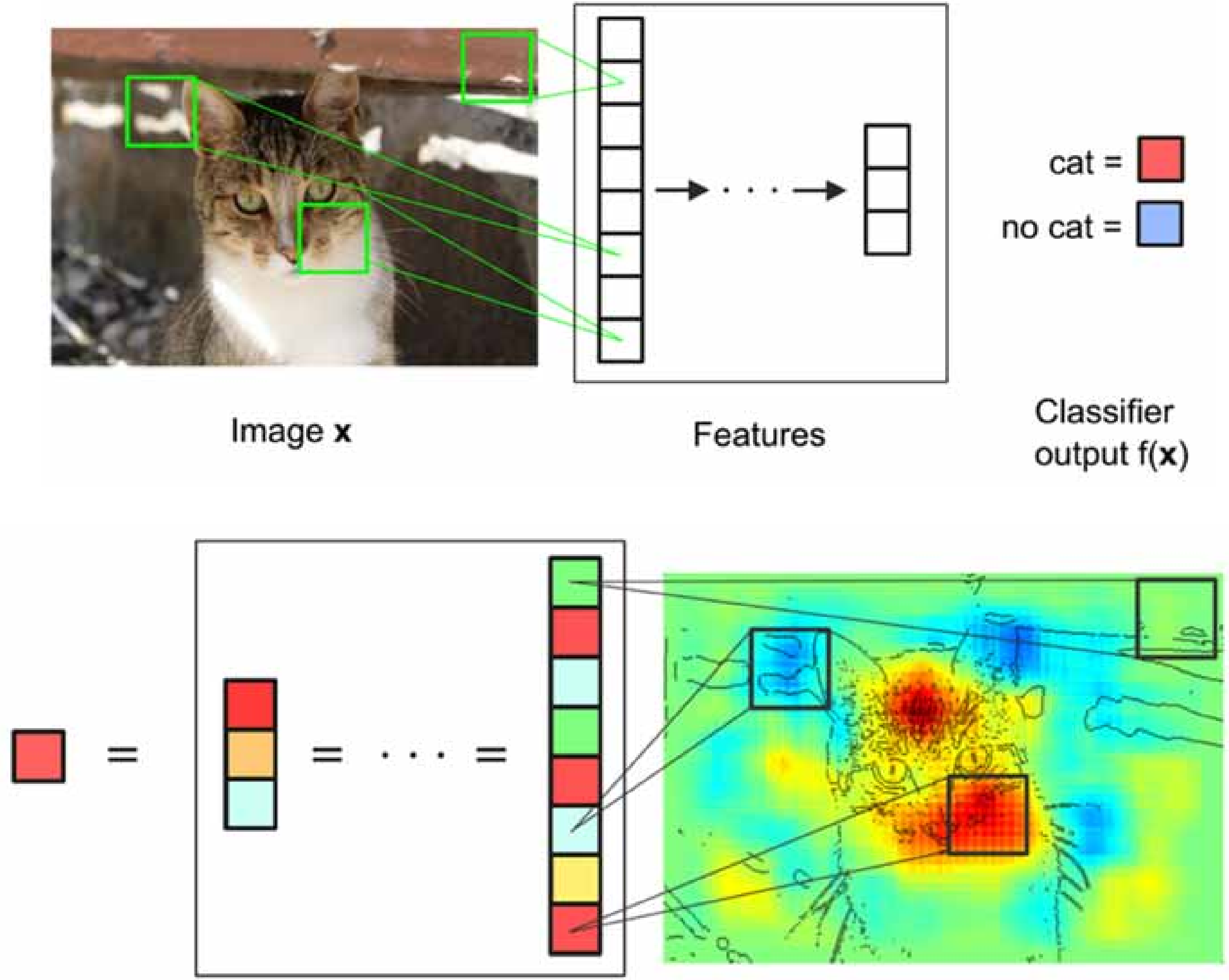
Cynthia Rudin 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, (5), 206-215, doi:10.1038/s42256-019-0048-x.

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

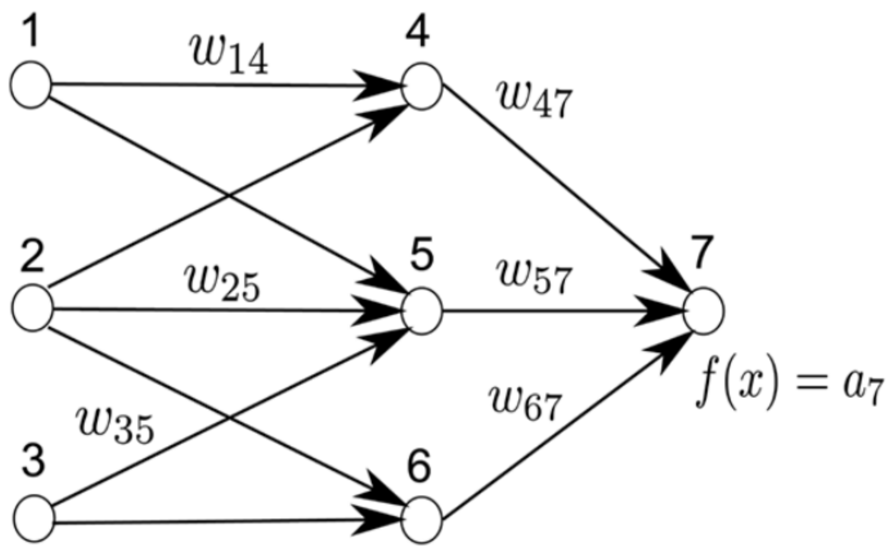
Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying 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.

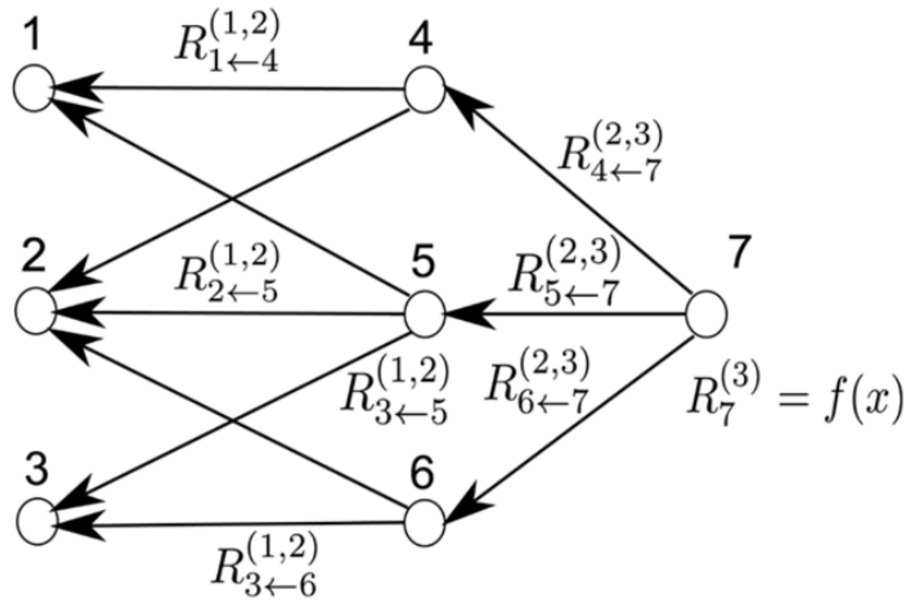


$$f(\mathbf{x}) = \sum \text{Feature Relevances} = \sum \text{Pixel Relevances}$$

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.

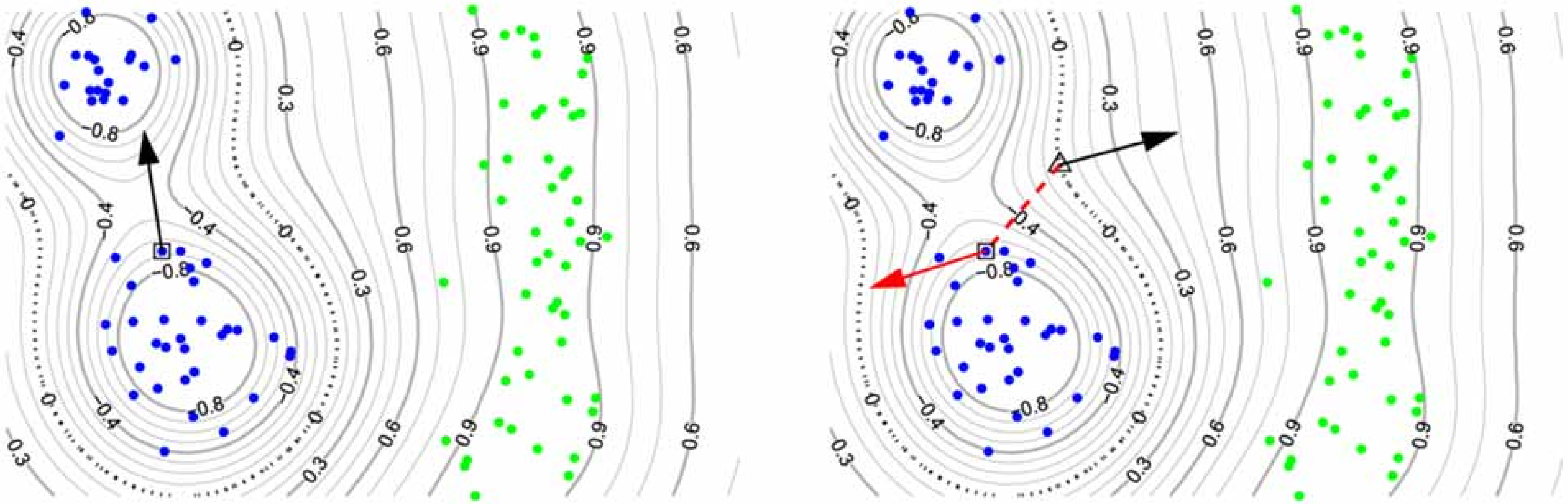
A NN-classifier during prediction time



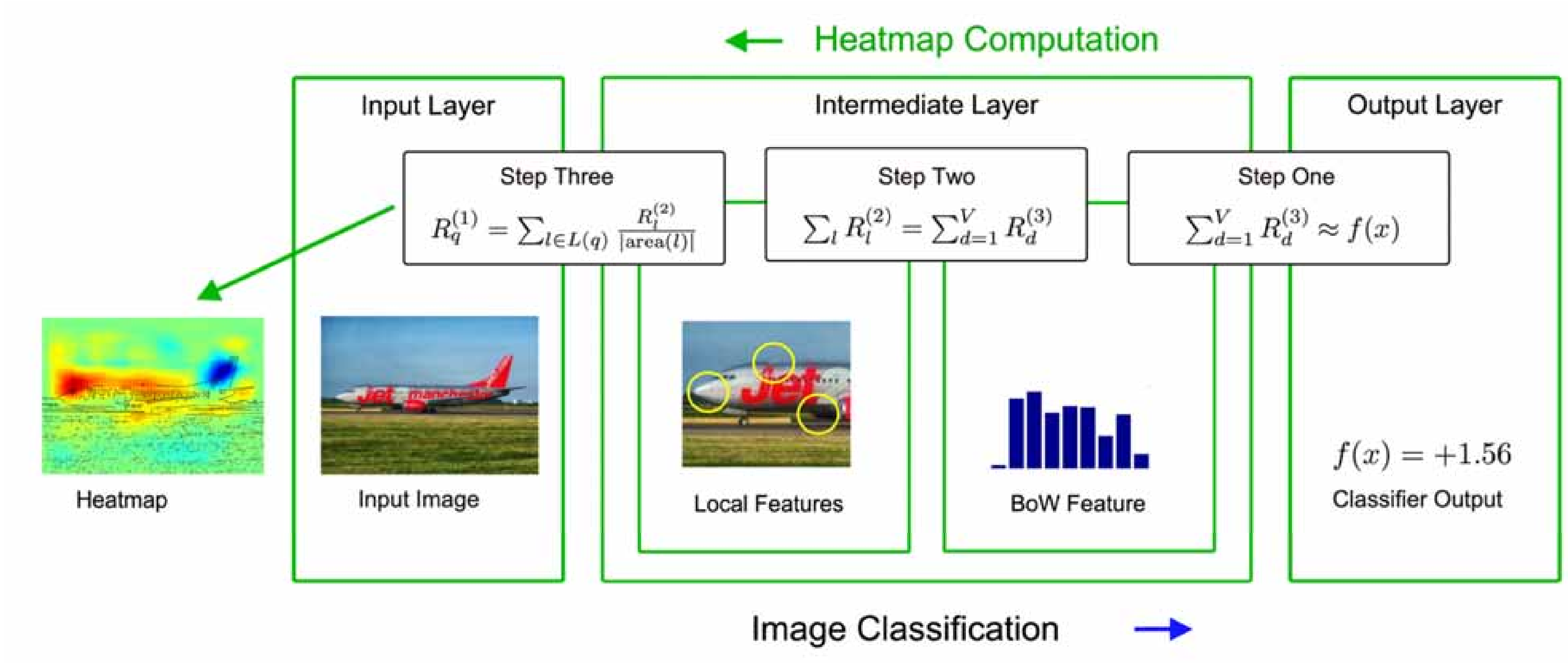


$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)}$$

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.

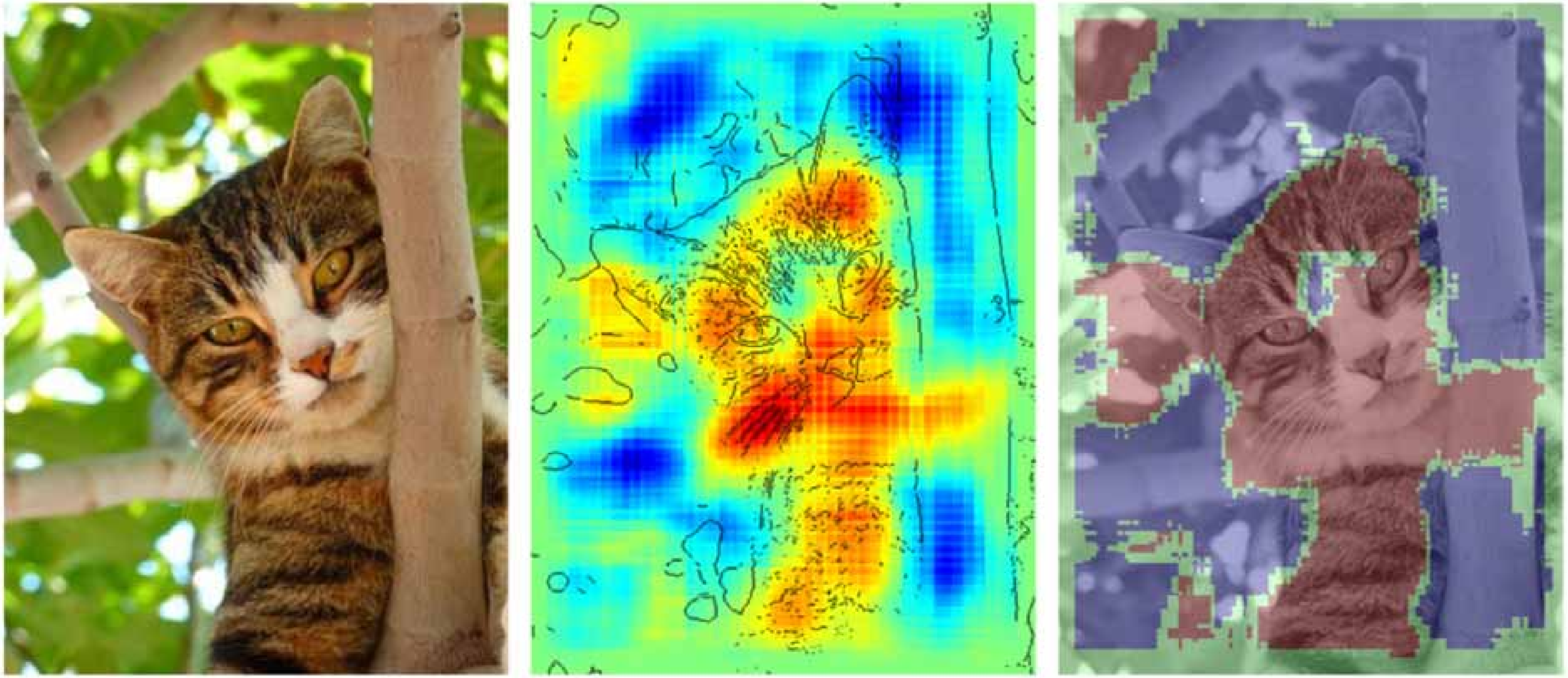


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



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.

Pixel-wise decomposition for bag-of-words features



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.

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:

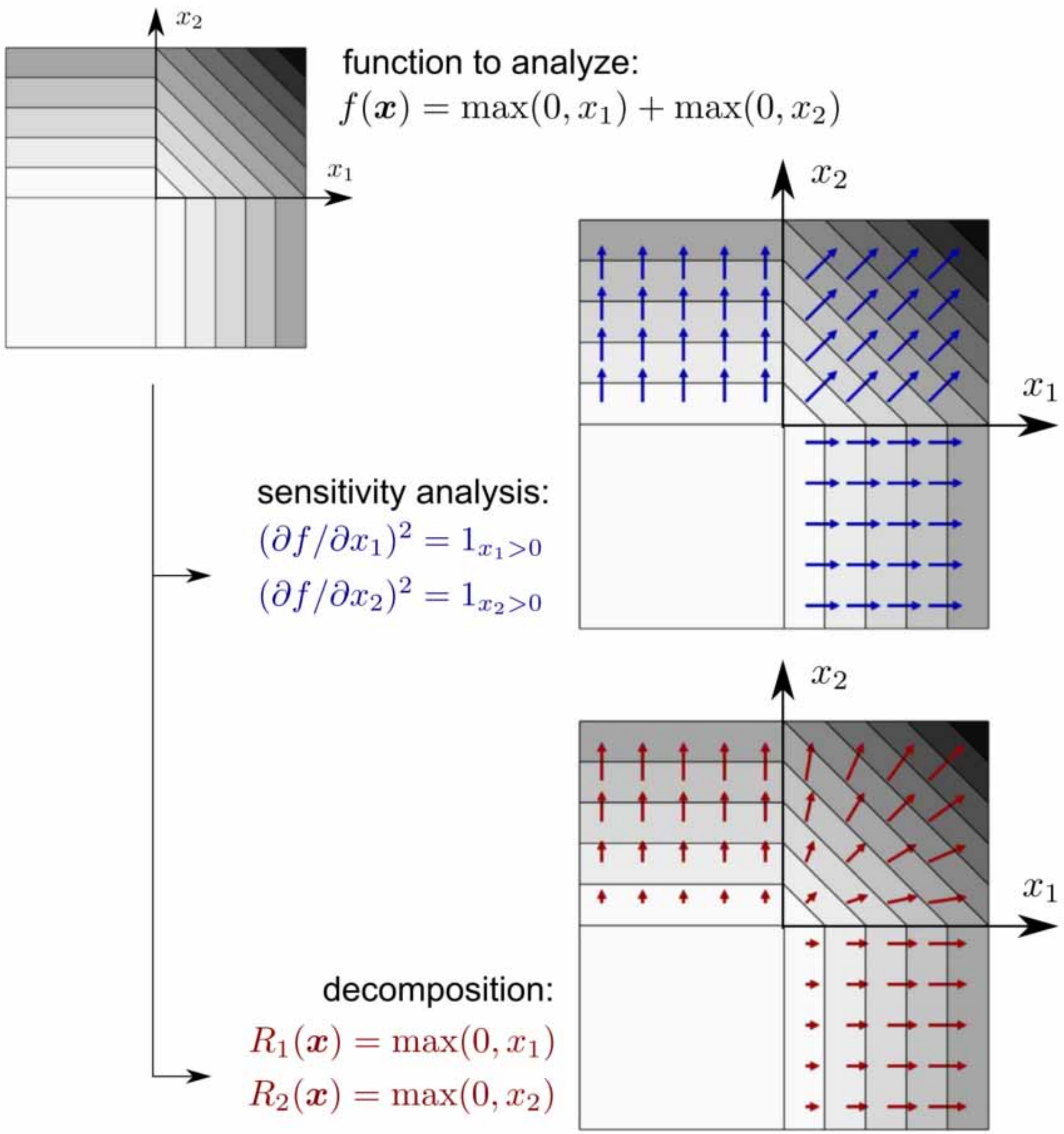
$$\forall\, \boldsymbol{x}: f(\boldsymbol{x}) = \sum_p R_p(\boldsymbol{x}).$$

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

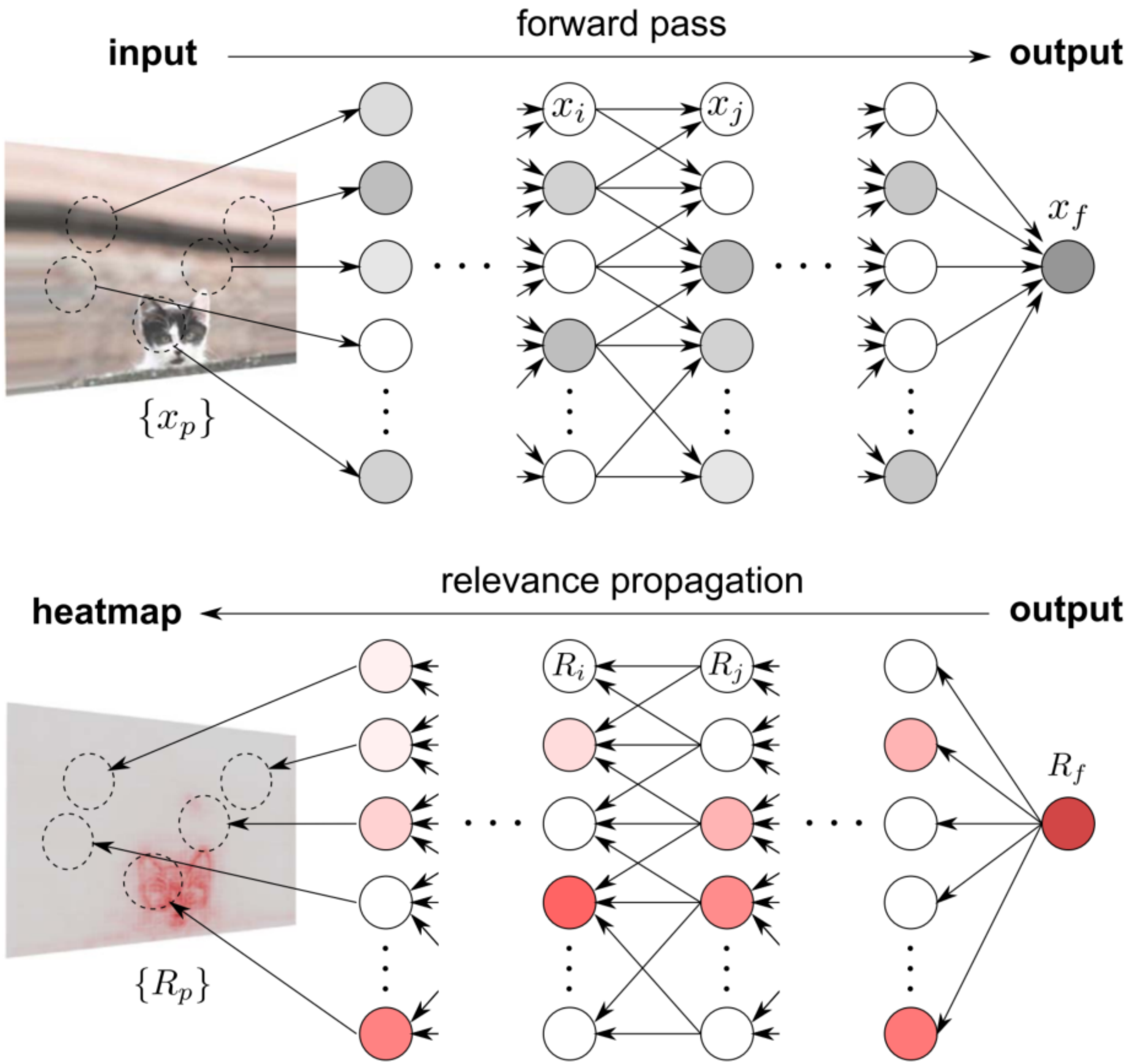
$$\forall\, \boldsymbol{x}, p: R_p(\boldsymbol{x}) \geq 0$$

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.

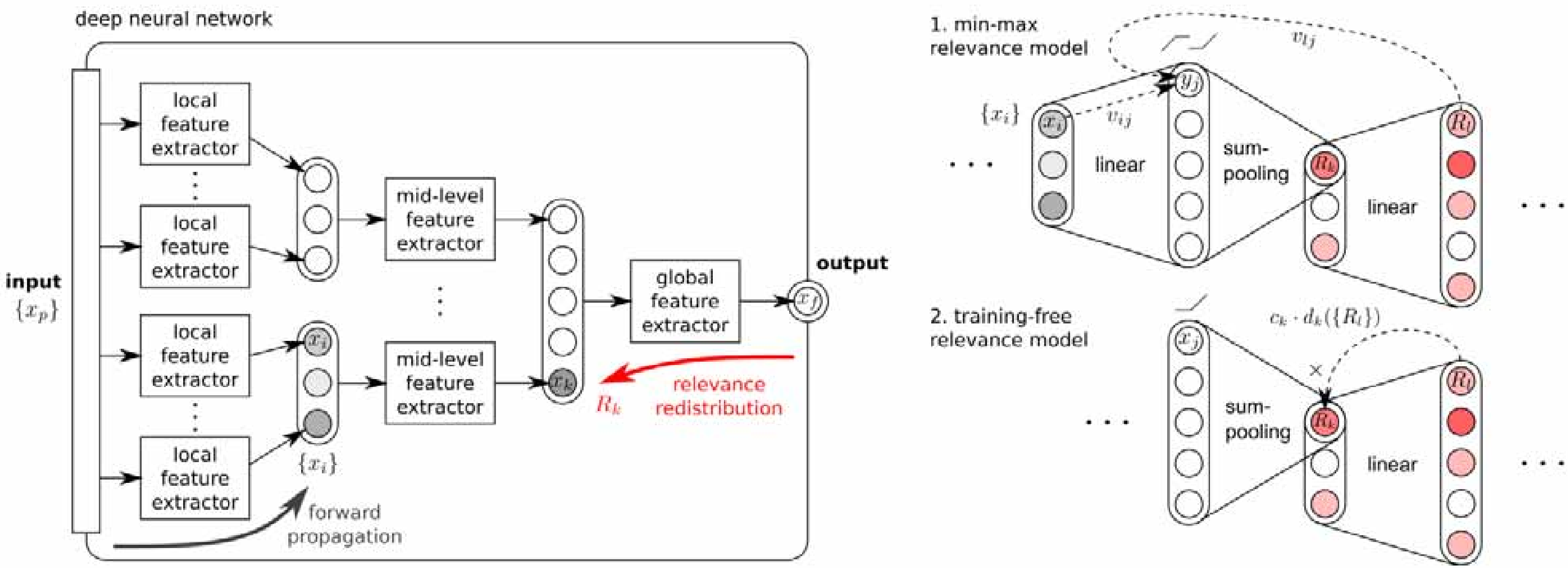
Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.


x_2
x_1
function to analyze:
$f(\boldsymbol{x}) = \max(0, x_1) + \max(0, x_2)$
sensitivity analysis:
$(\partial f / \partial x_1)^2 = 1_{x_1 > 0}$
$(\partial f / \partial x_2)^2 = 1_{x_2 > 0}$
decomposition:
$R_1(\boldsymbol{x}) = \max(0, x_1)$
$R_2(\boldsymbol{x}) = \max(0, x_2)$

Relevance propagation



Relevance Redistribution



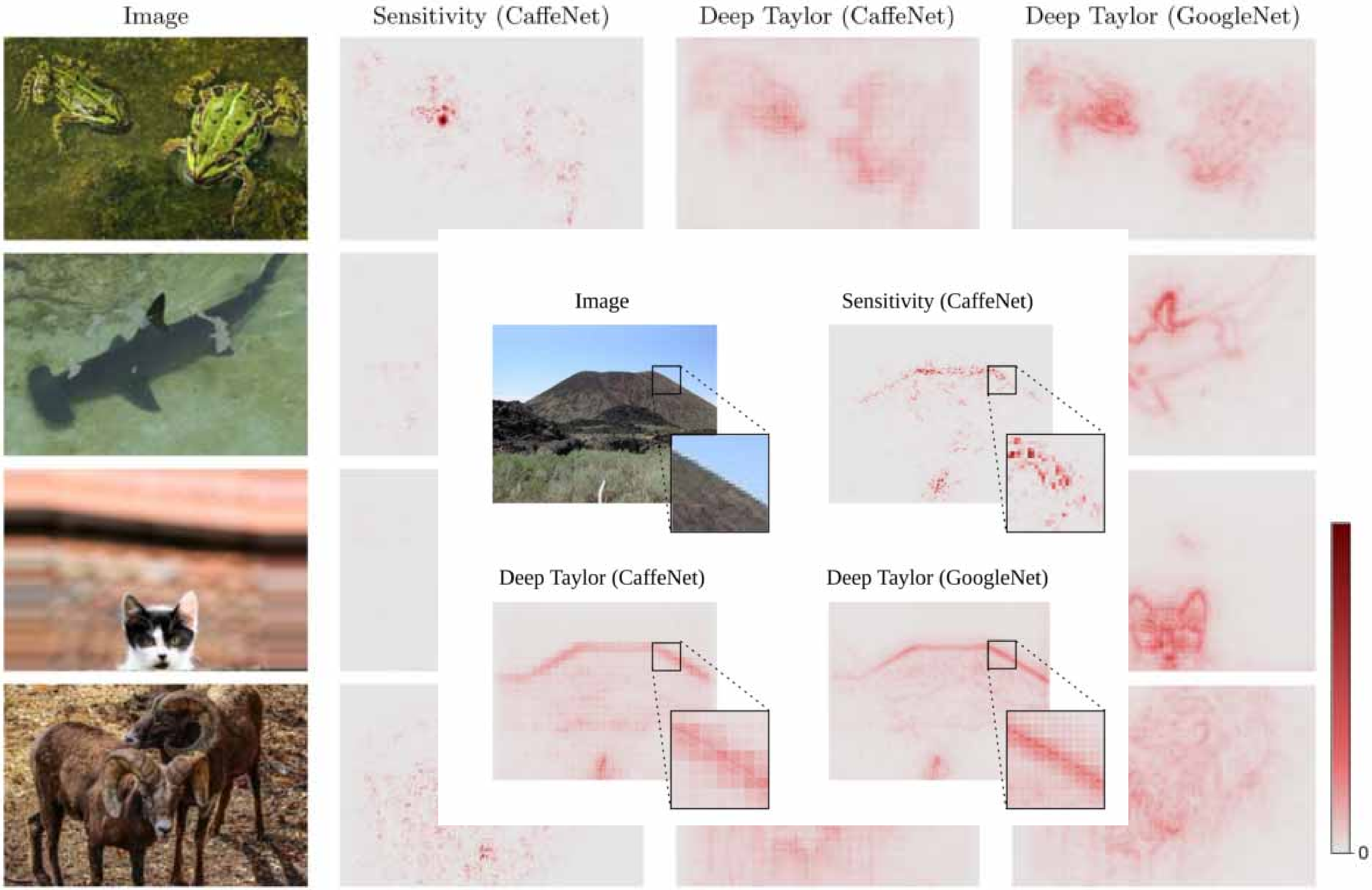
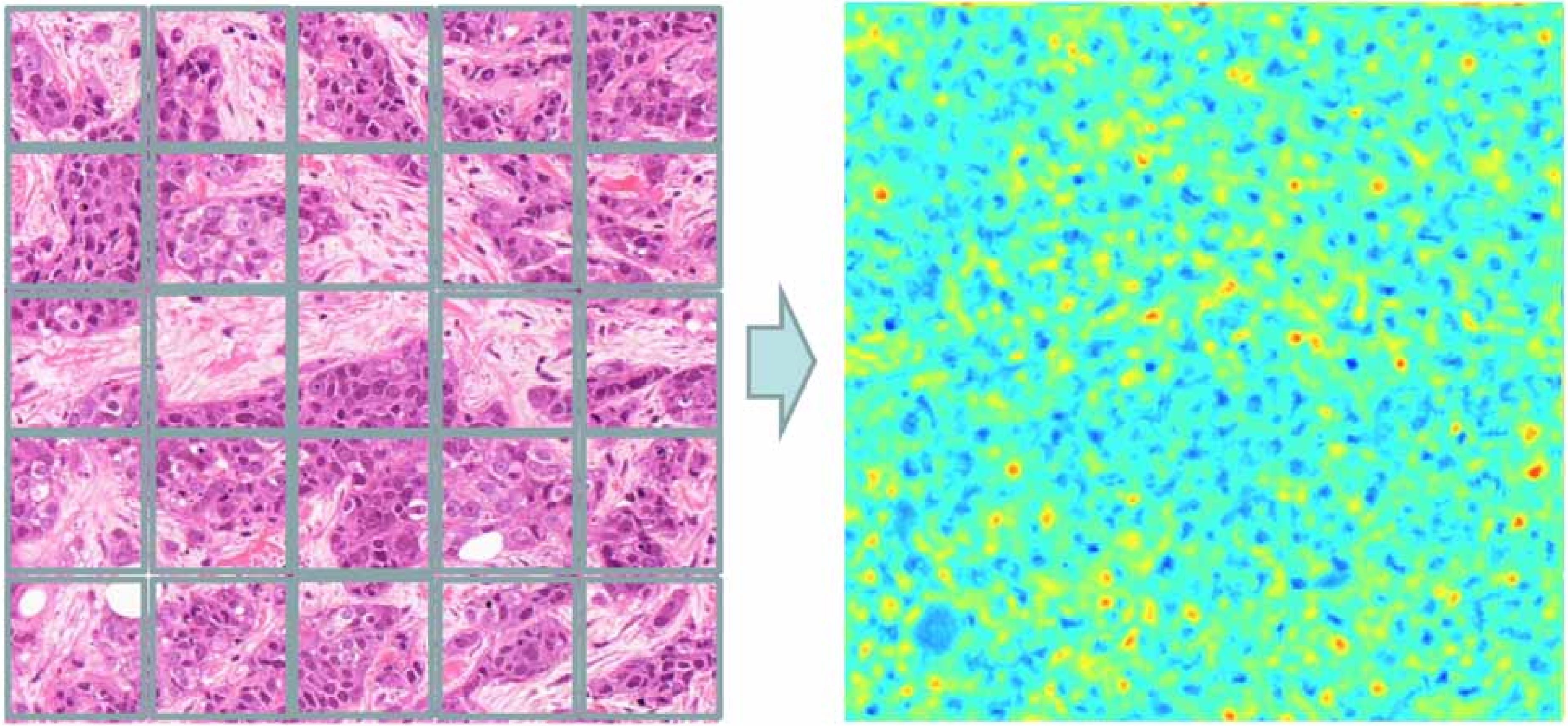
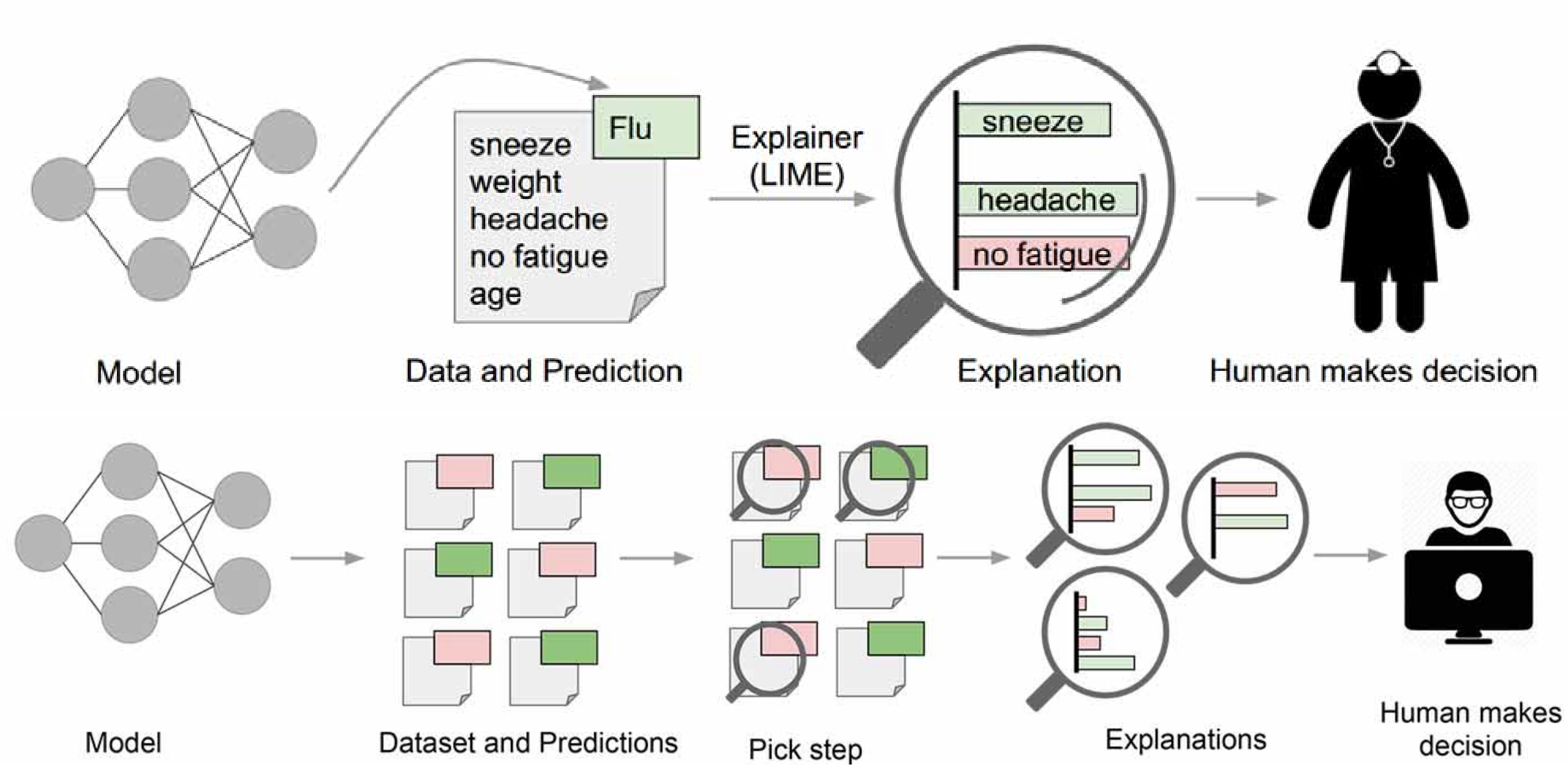


Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
0

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.



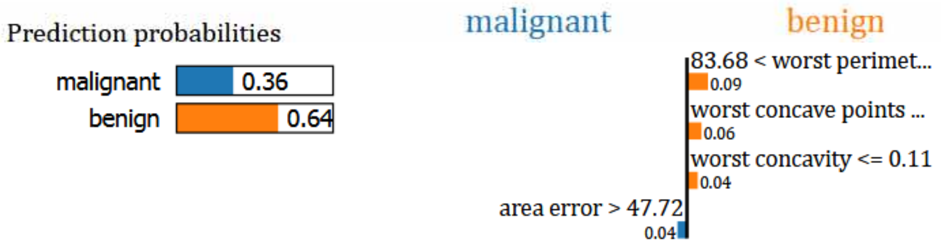
Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

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

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

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

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

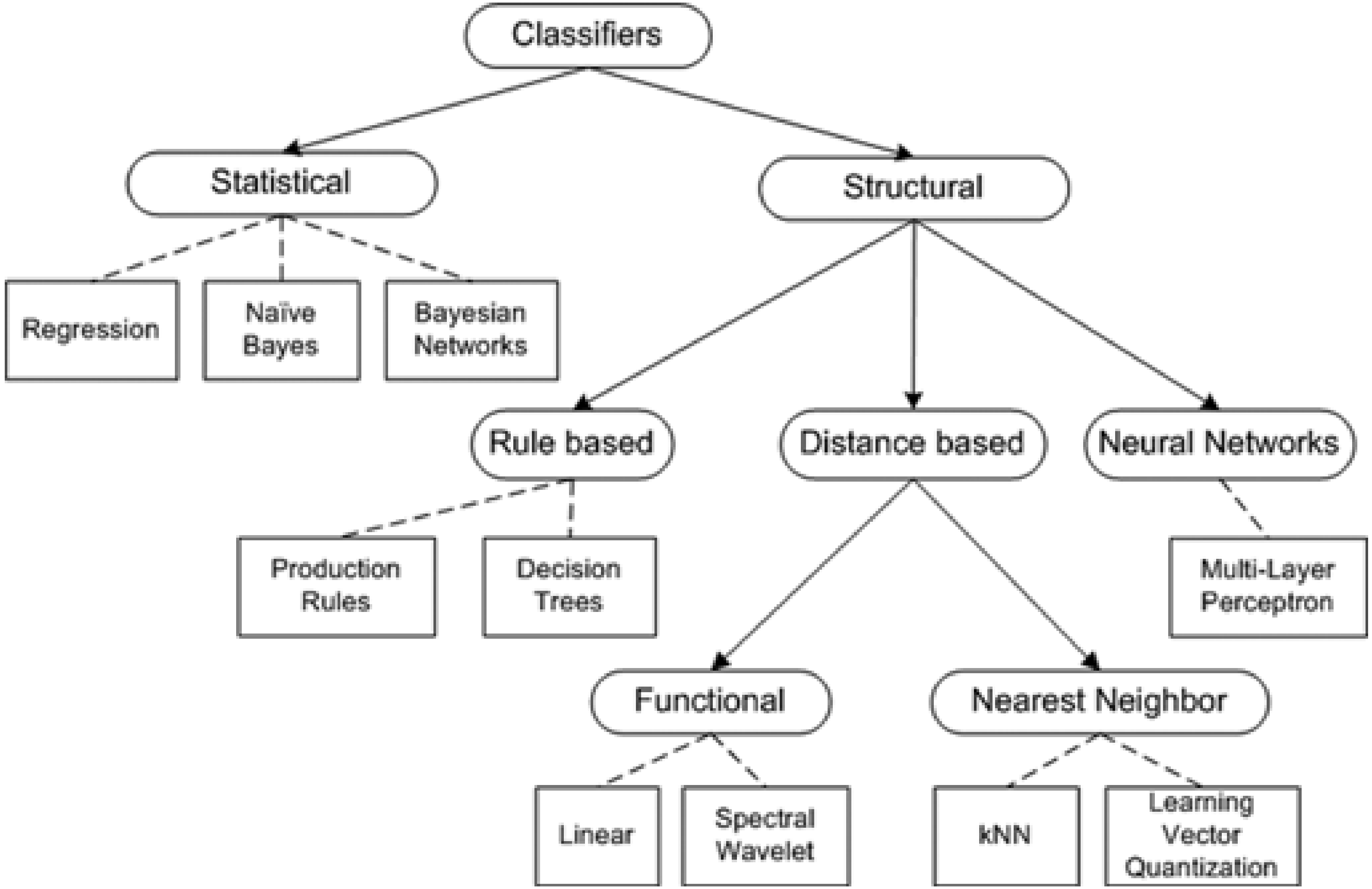


Feature	Value
worst perimeter	95.54
worst concave points	0.03
worst concavity	0.02
area error	106.00

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

If Age < 50 and Male = Yes:

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

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

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

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

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

Default:

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

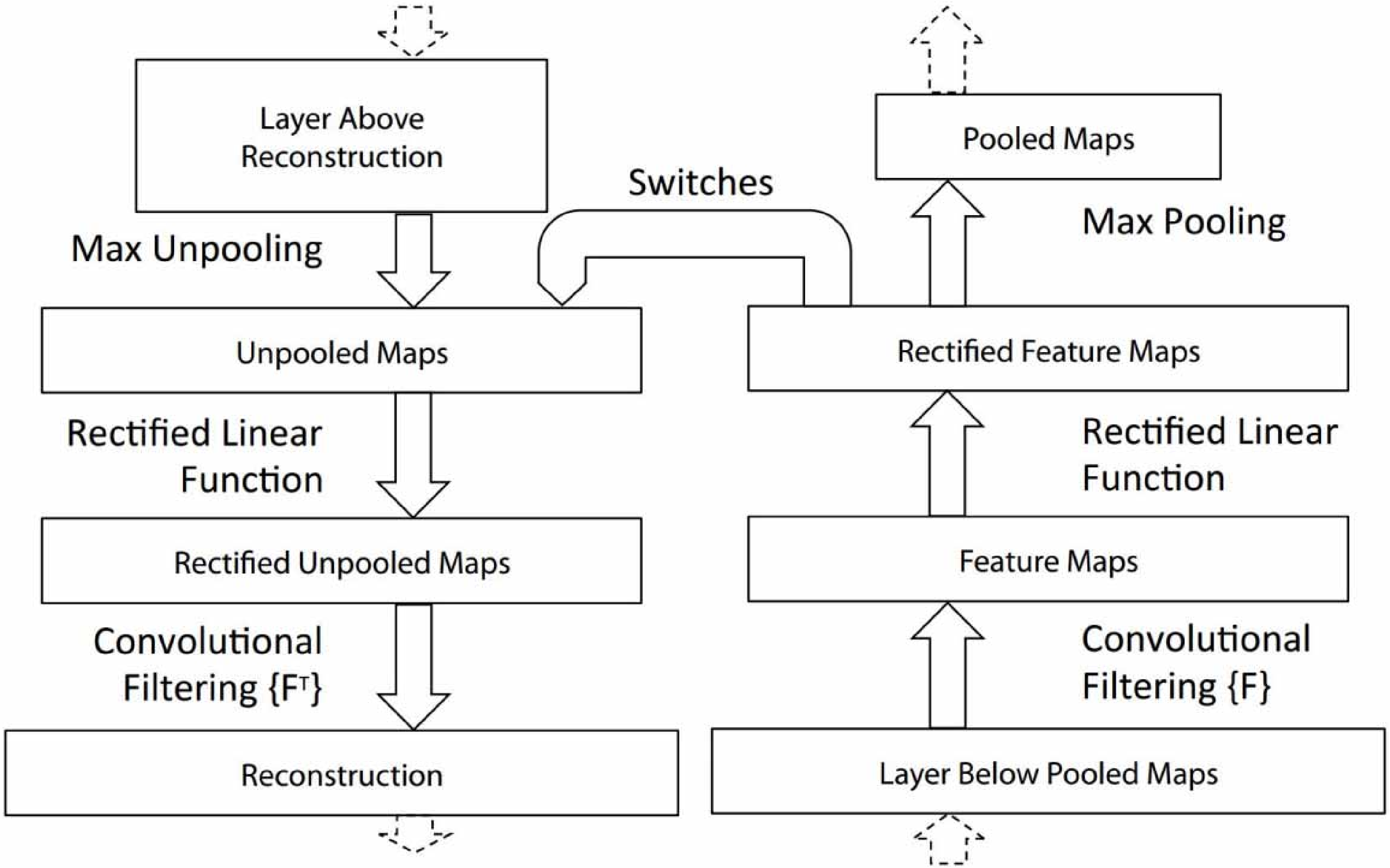
If Past-Depression = No and Weight-Gain = Yes and Tiredness = Yes and Melancholy = Yes, then Depression

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

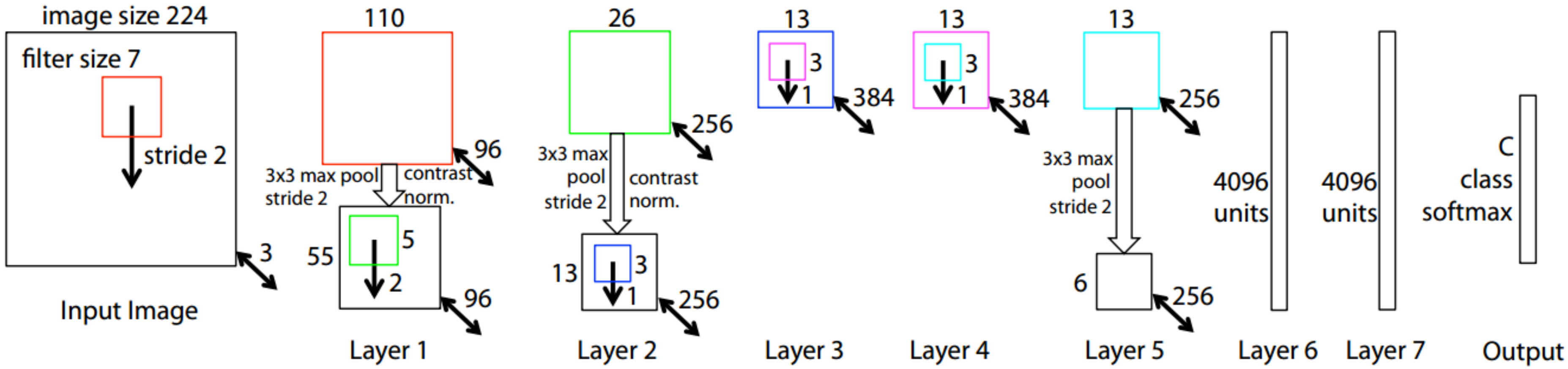
Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

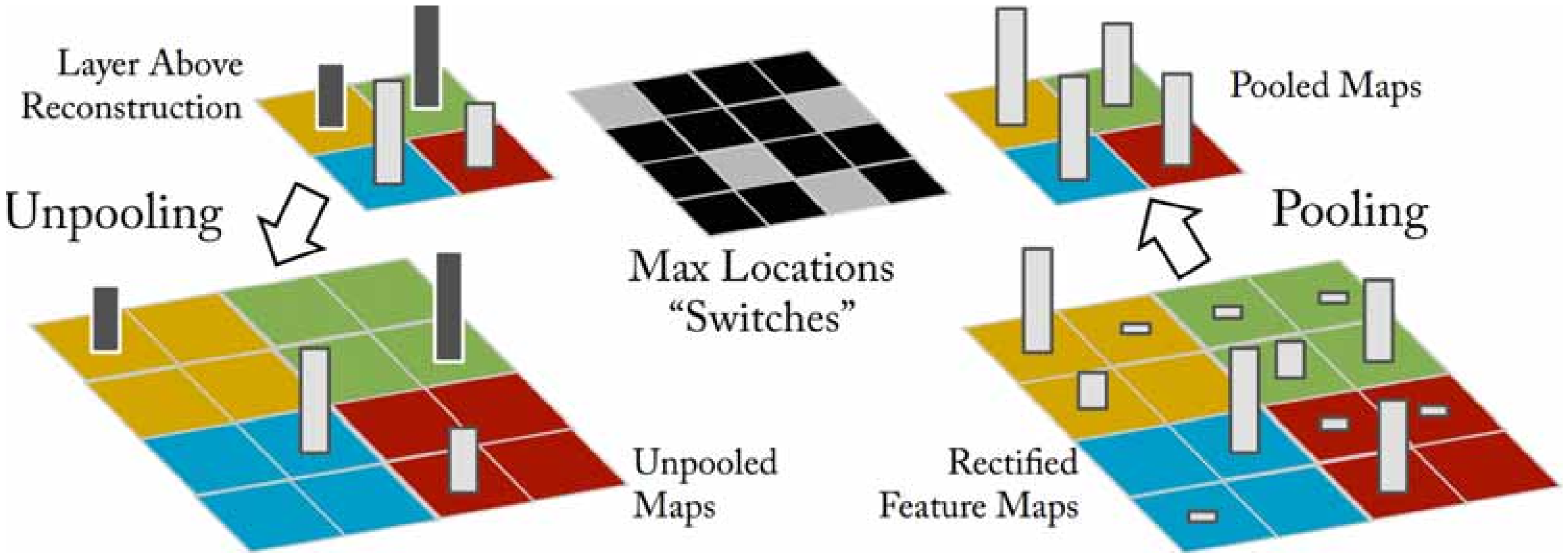
05 Principles of Making Neural Networks transparent

Example: Interpretable Deep Learning Model

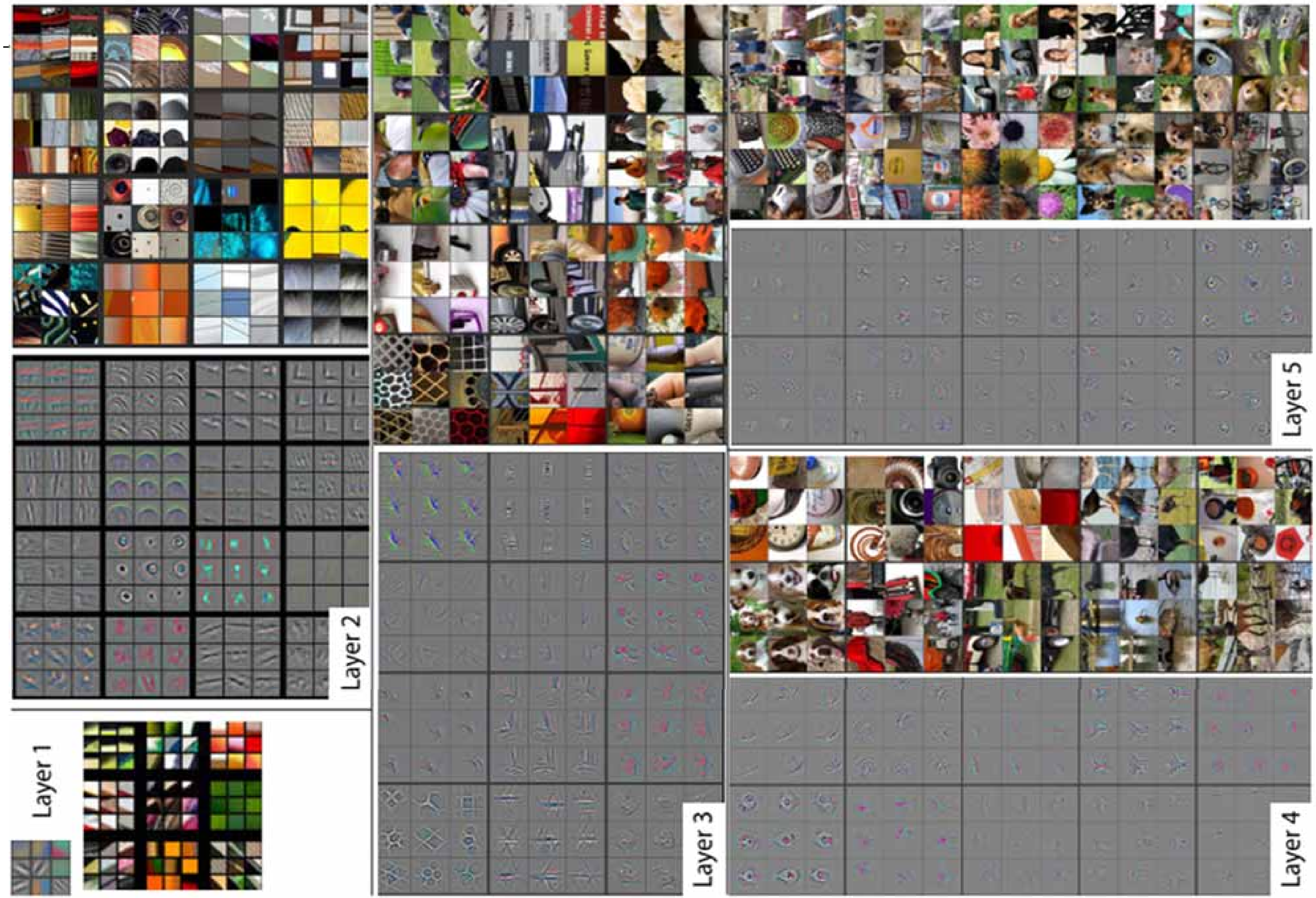


Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.



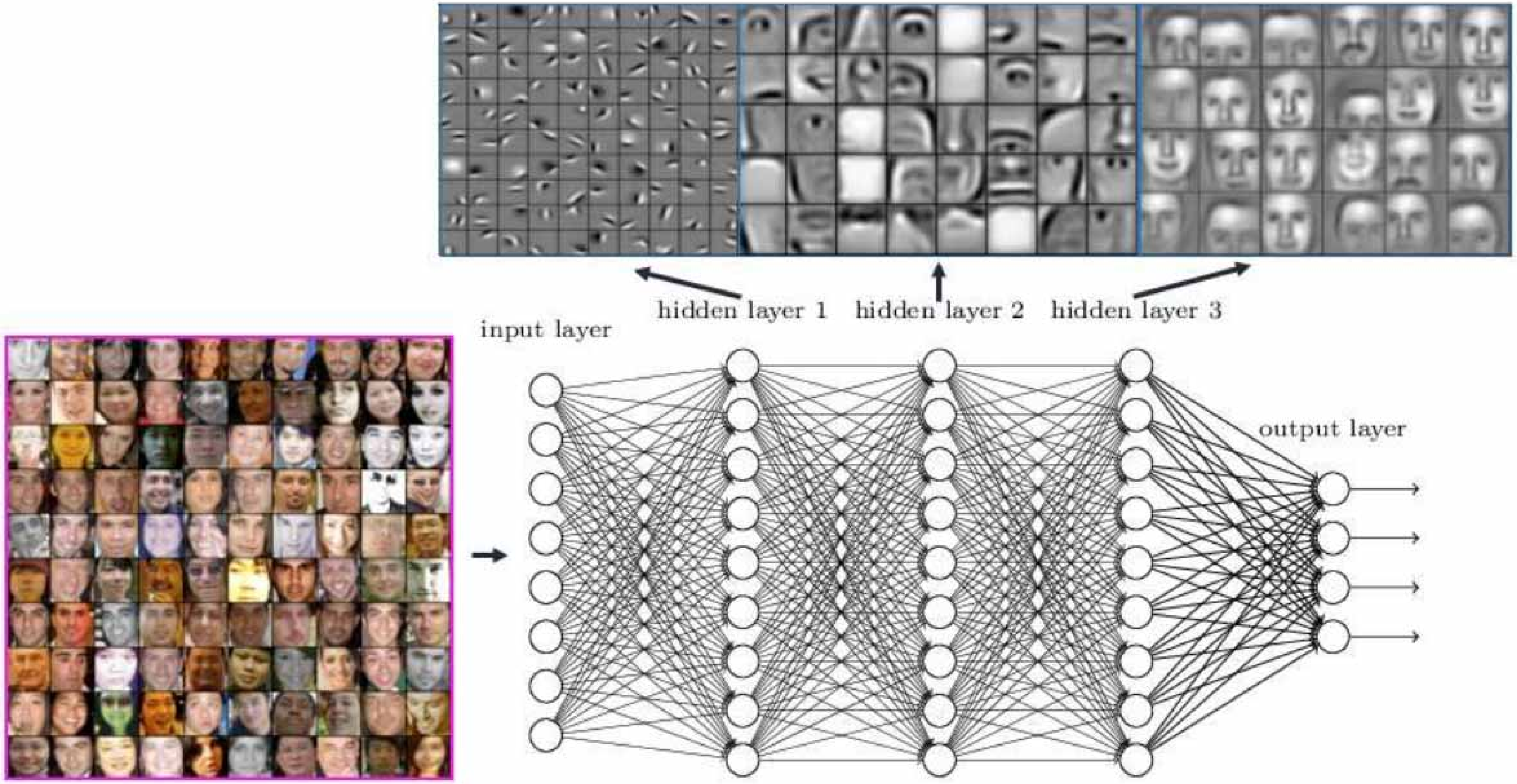


Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



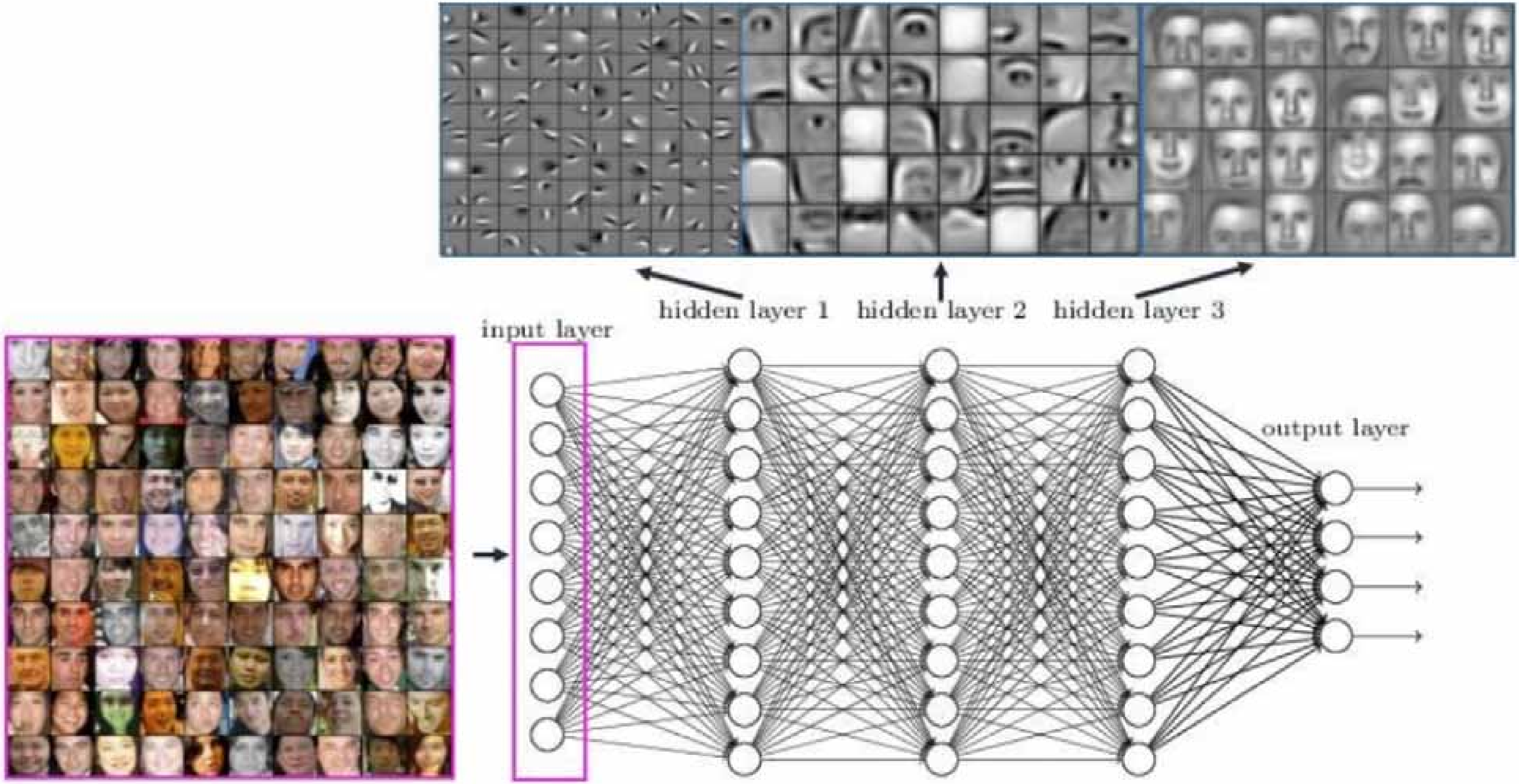
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

The world is compositional (Yann LeCun)

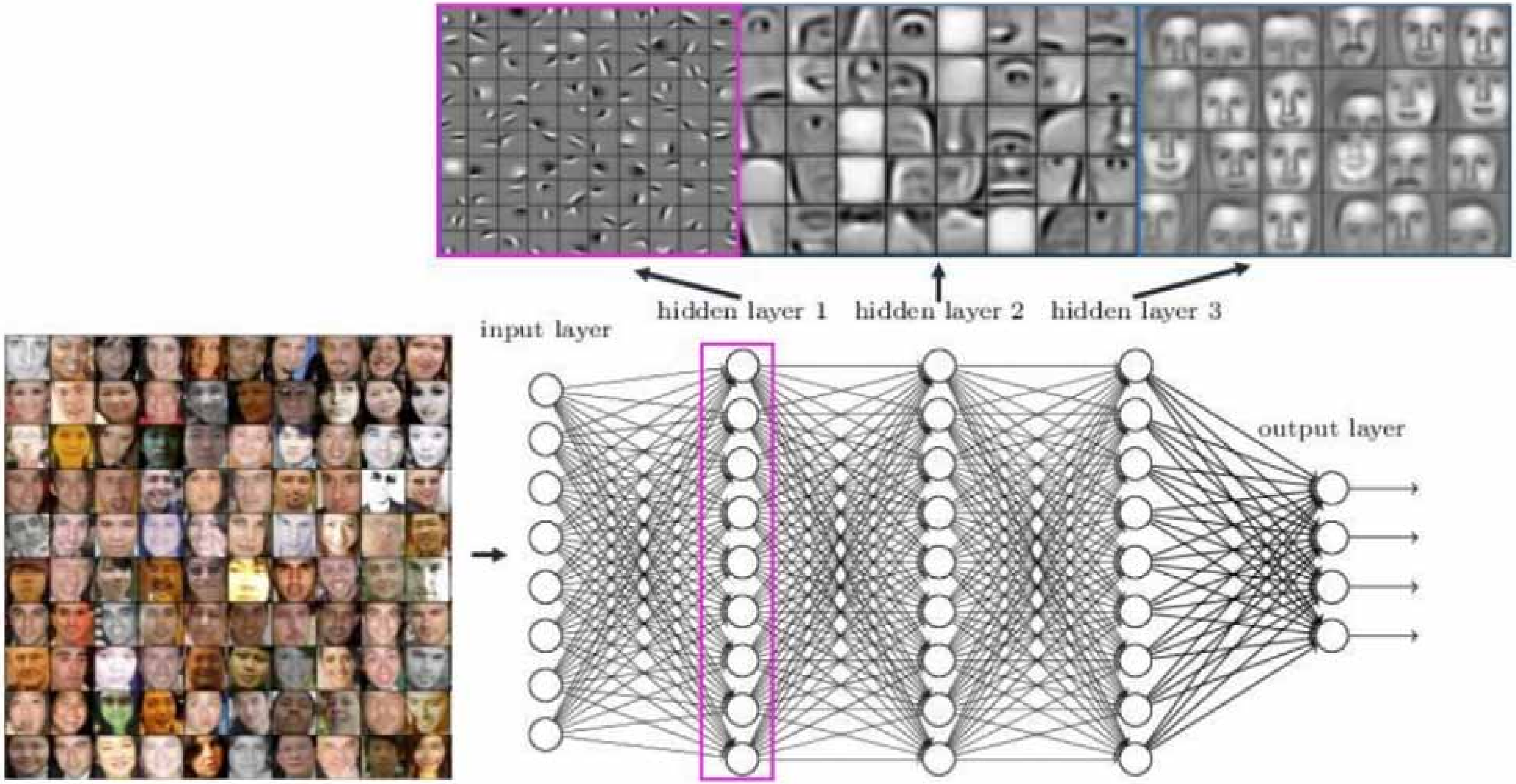


Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

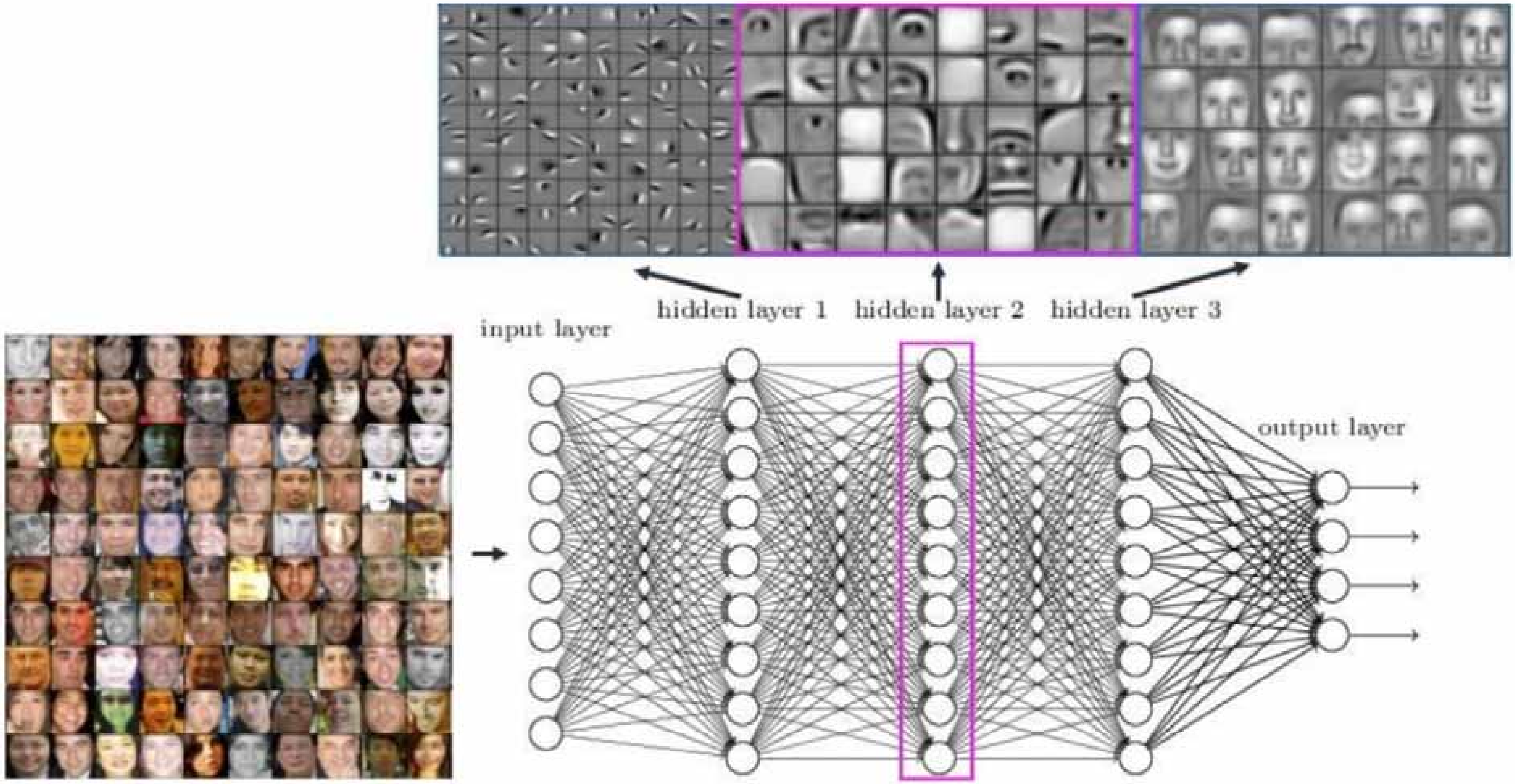
The world is compositional (Yann LeCun)



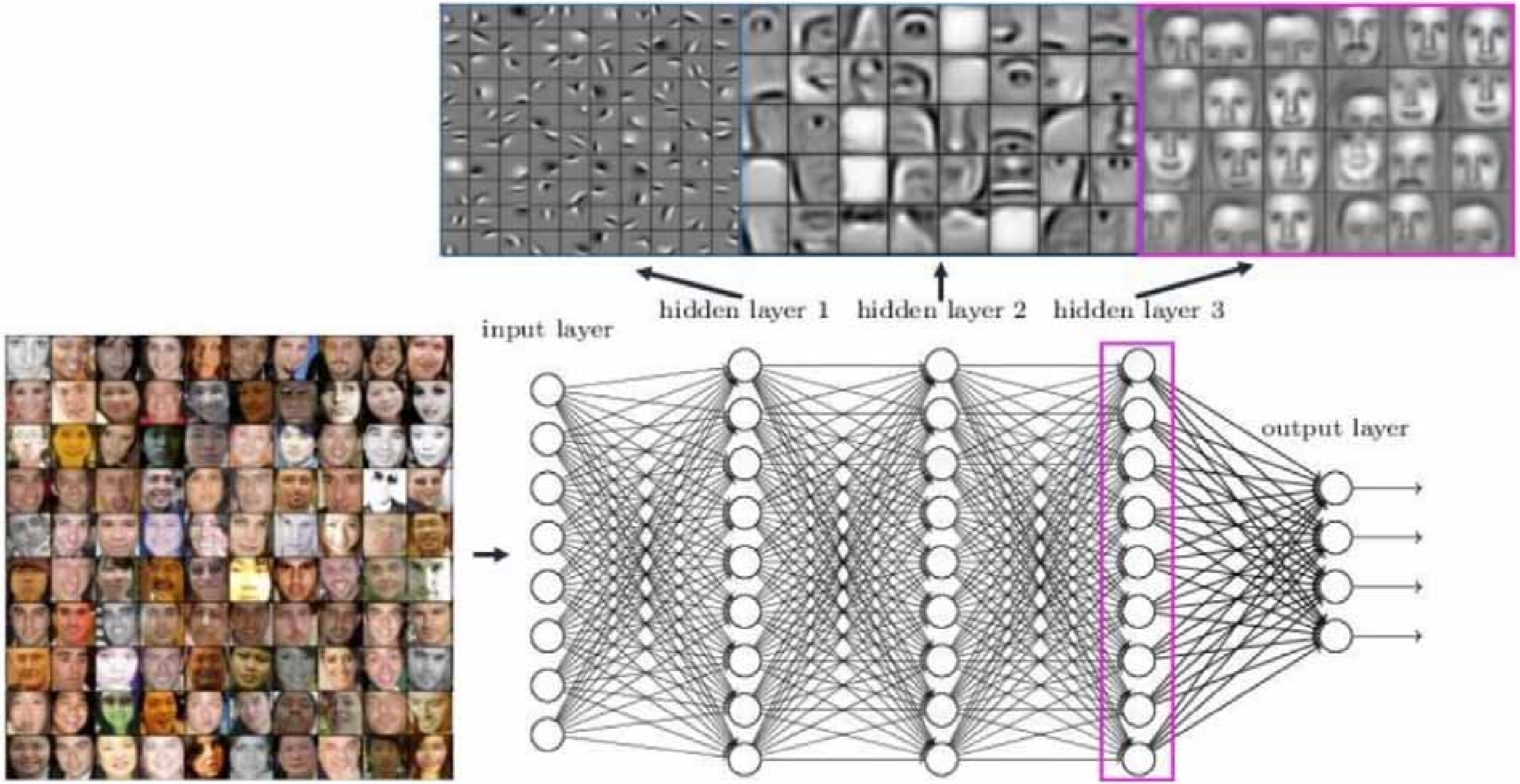
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



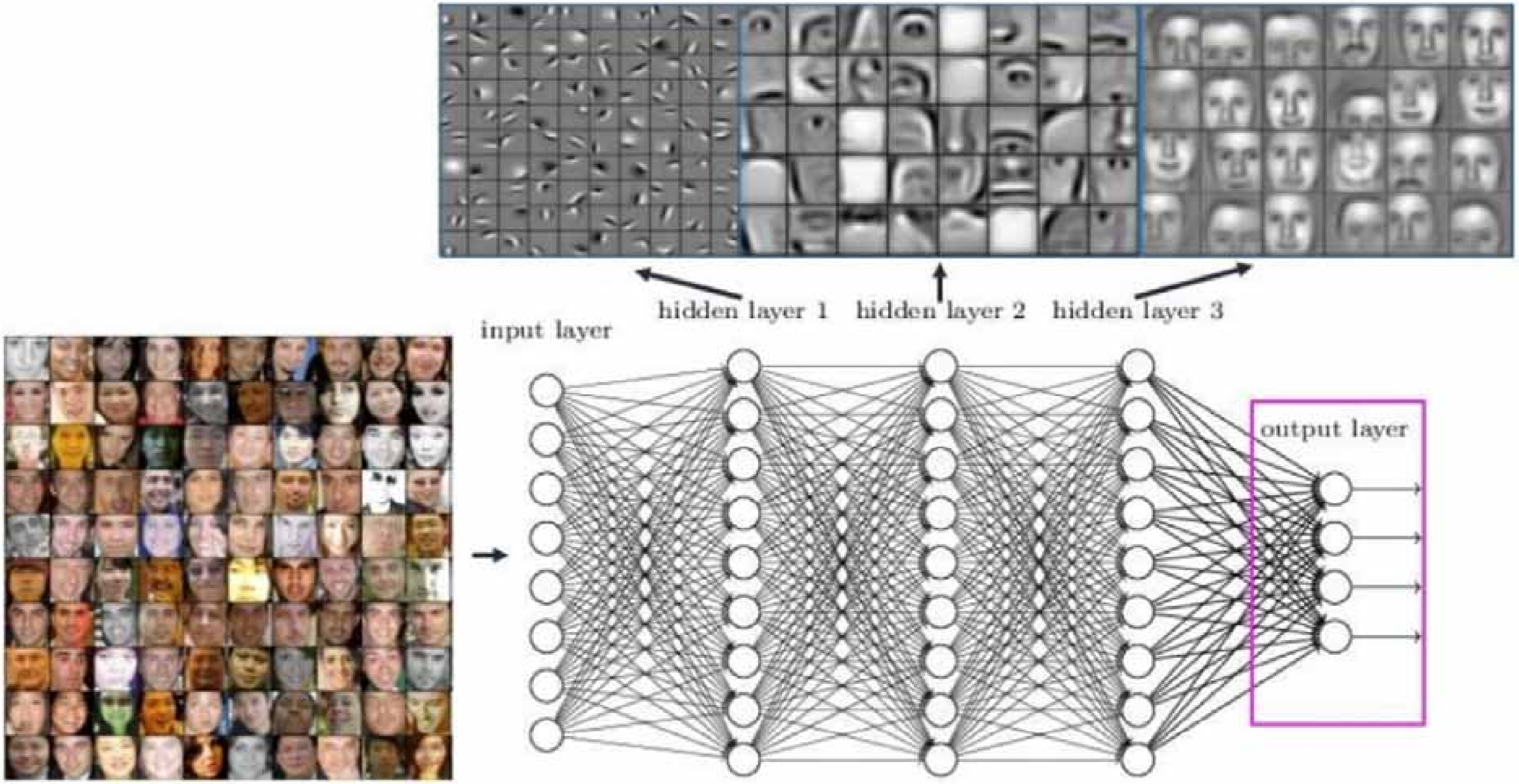
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



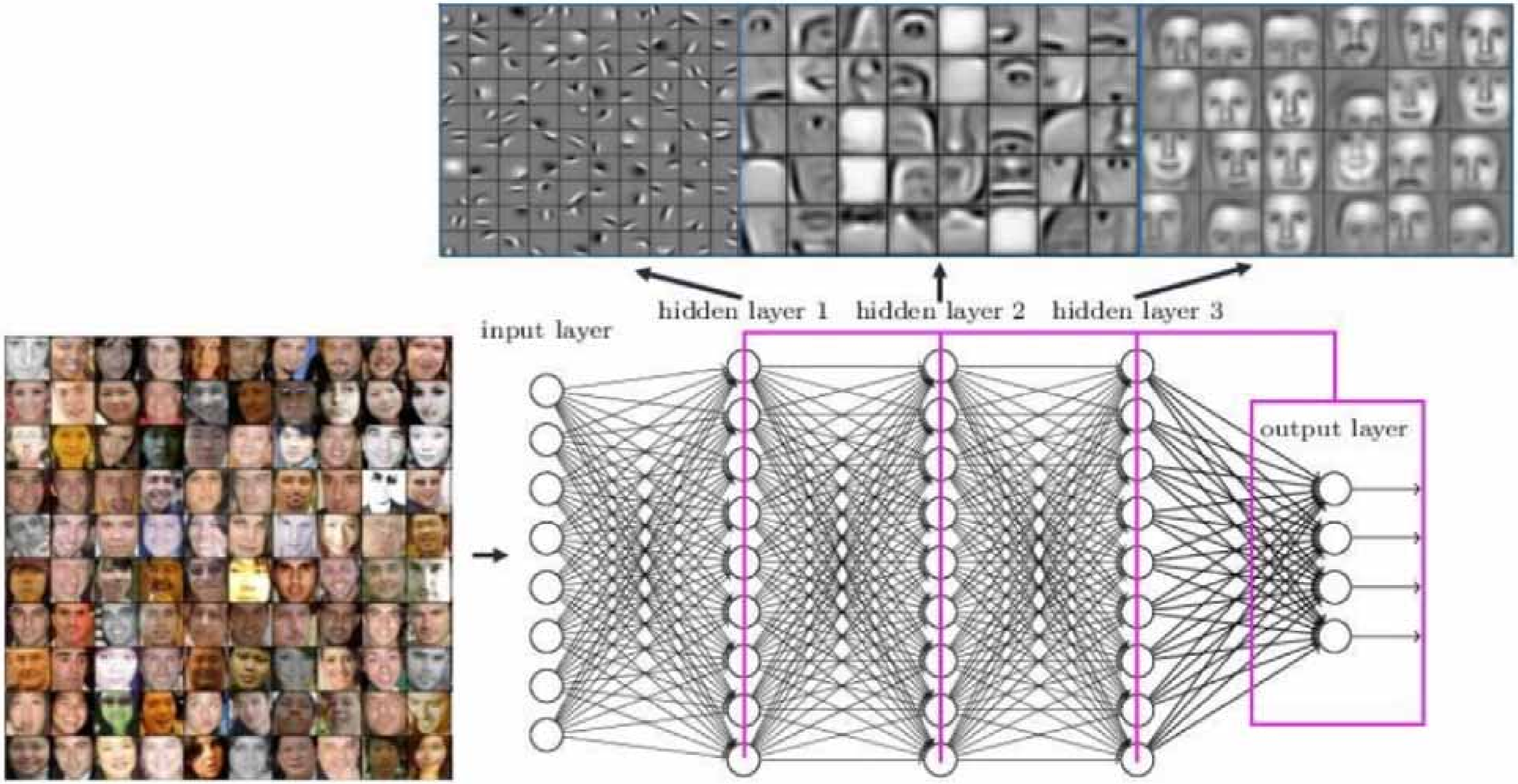
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



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



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

06 Explanation Interfaces: Future Human-AI Interaction

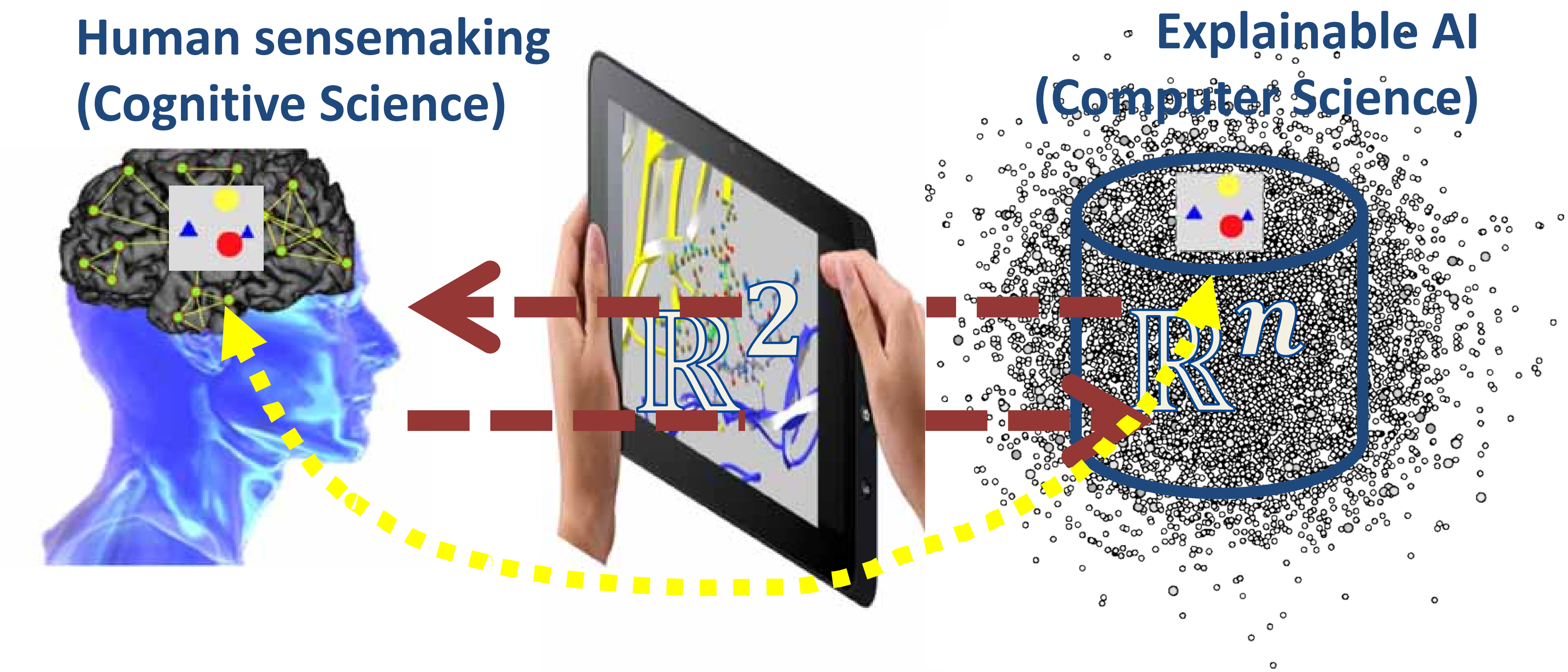
- Explanation is a reasoning process
- Open questions:
  - What is a good explanation?
  - When is it enough (degree of saturation)?
  - Context dependent (Emergency vs. research)
  - How can we measure the degree of comprehensibility of a given explanation
  - (obviously the explanation was good when 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!

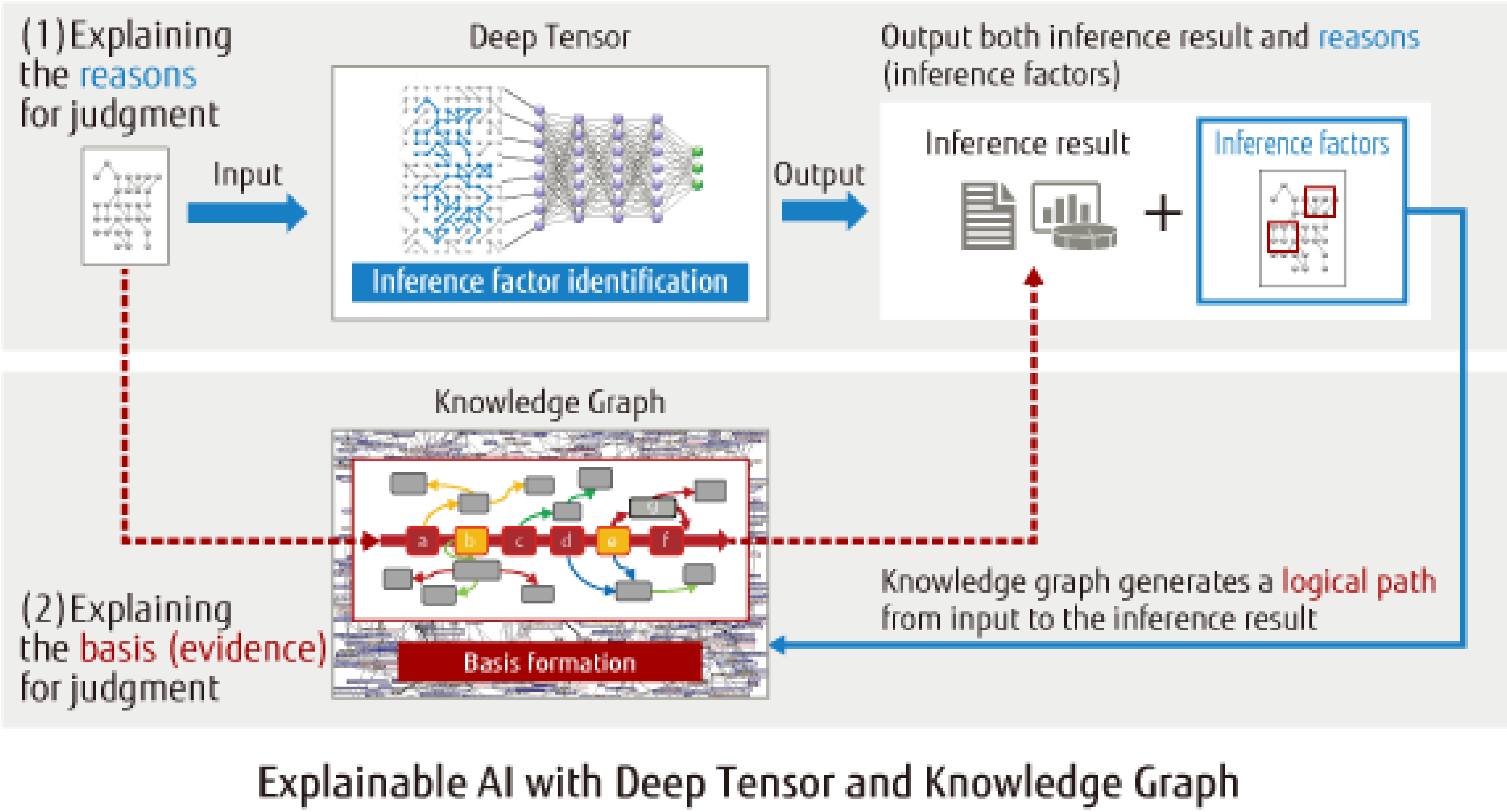
Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland & Patrick Vinck 2018. Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology, 31, (4), 611-627.

Explainability :=
a property of a system
(“the AI explanation)
Causability :=
a property of a person
(“the Human explanation)

Andreas Holzinger et al. 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

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





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

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

- Justification, Explanation and Causality
- Trust > scaffolded by justification of actions (why)
- Please always take into account the inherent uncertainty and incompleteness of medical data!

Alex John London 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report, 49, (1), 15-21, doi:10.1002/hast.973.

Noel C.F. Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, 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

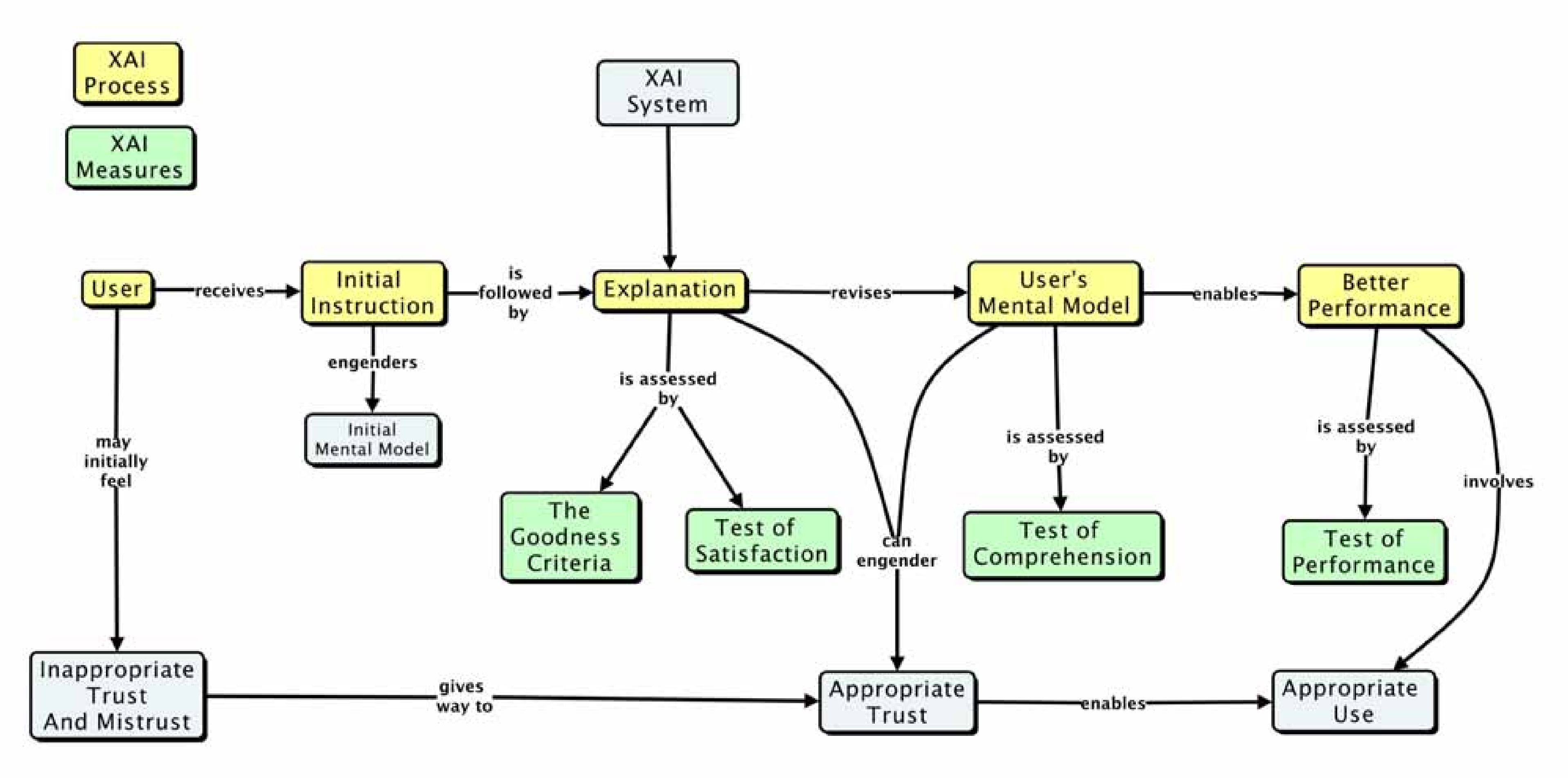
New regulations call for automated decision making systems to provide "meaningful information" on the logic used to reach conclusions [1–4]. Selbst and Powles interpret the concept of "meaningful information" as information that should be understandable to the audience (potentially individuals

07 Metrics of xAI

- Central question: How can we measure whether and to what extent an “explanation” given by a machine has been understood by a human?
- Therefore we need to know:
- (1) the “goodness” of explanations,
- (2) the “satisfaction” of the user
- (3) the “understandability”
- (4) the “trustability”
- (5) the “human-AI interaction”
- Please note that the terms are in “quotation marks” because it is extremely difficult to measure!

Robert R. Hoffman, Shane T. Mueller, Gary Klein & Jordan Litman 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

Conceptual Model of explaining by Hoffman et al. (2018)

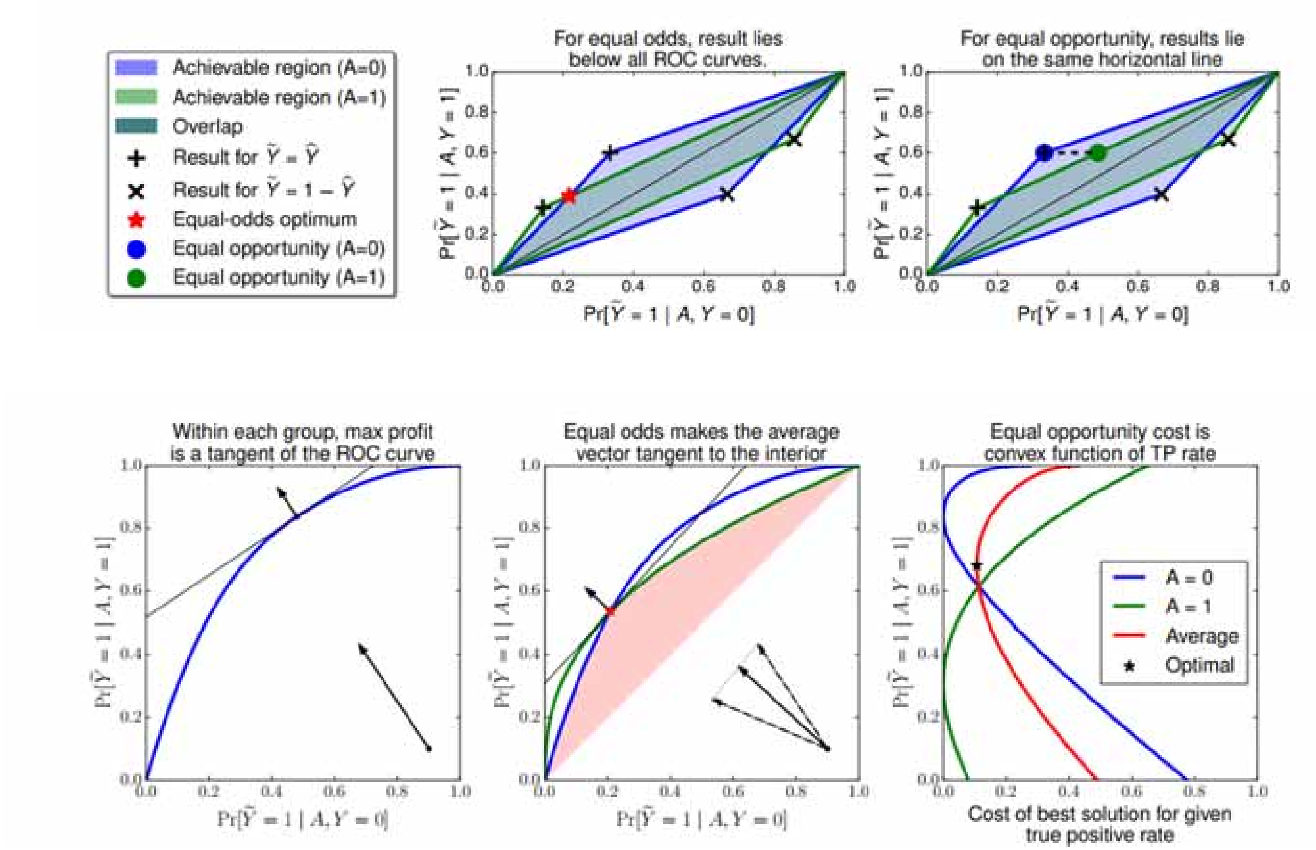


Robert R. Hoffman, Shane T. Mueller, Gary Klein & Jordan Litman 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

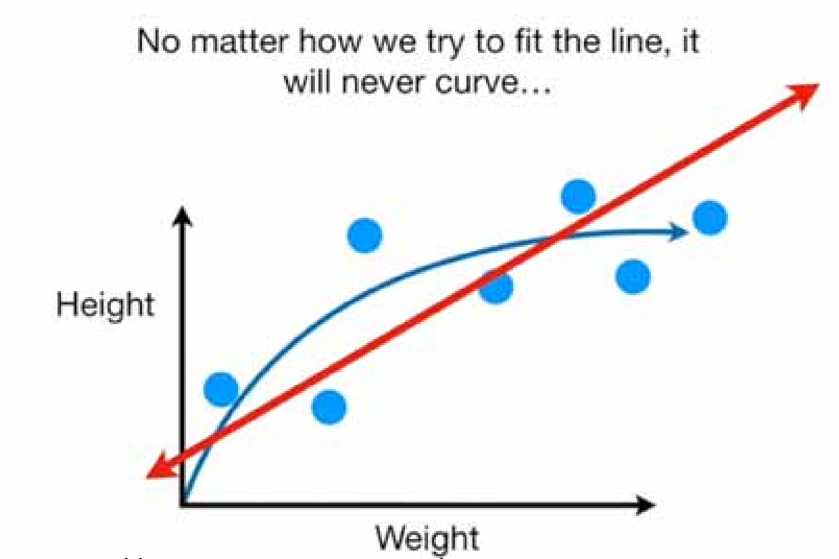
Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey & Gary Klein 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv:1902.01876.

Tim Miller 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38, doi:10.1016/j.artint.2018.07.007.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel. Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012 Cambridge, Massachusetts. Association for Computing Machinery, 214–226, doi:10.1145/2090236.2090255.



Moritz Hardt, Eric Price & Nati Srebro. Equality of opportunity in supervised learning. Advances in neural information processing systems (NIPS 2016), 2016. 3315-3323.



https://www.youtube.com/watch?v=EuBBz3bI-aA

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

> **"There are errors in these systems which propagate very quickly. Because of their scale of their action space – they can be hitting a billion or two billion users per day – that means the costs of getting it wrong are very very high."**
>
> **-Mustafa Suleyman**
> **co-founder DeepMind**

Artificial Intelligence / Robots

Forget Killer Robots— Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.

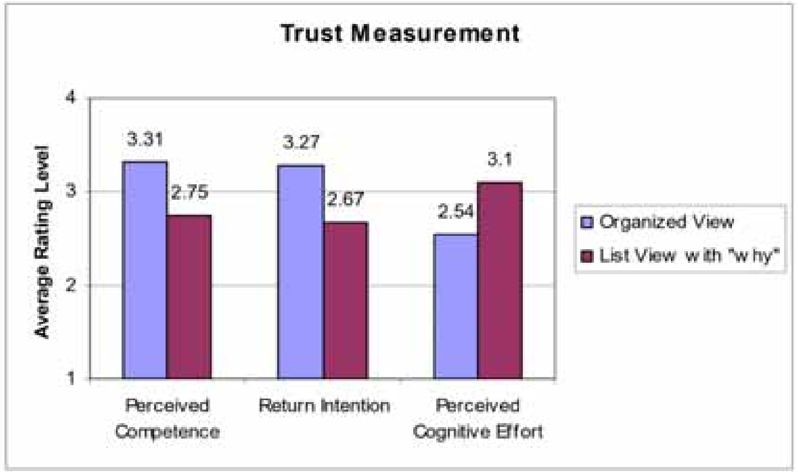
by **Will Knight** Oct 3, 2017

Google's AI chief isn't fretting about super-intelligent killer robots. Instead, John Giannandrea is concerned about the danger that may be lurking inside the machine-learning algorithms used to make millions of decisions every minute.

"The real safety question, if you want to call it that, is that if we give these systems biased data, they will be biased," Giannandrea said before a recent Google conference on the relationship between humans and AI systems.

How to measure “trust”?

Items in the **Cognitive Effort** construct	Mean	
	Organized view	List view with “why”
I easily found the information I was looking for (*reverse scale*);	2.47	3.07
Selecting a product using this interface required too much effort.	2.61	3.14
	Cronbach’s alpha = 0.73	

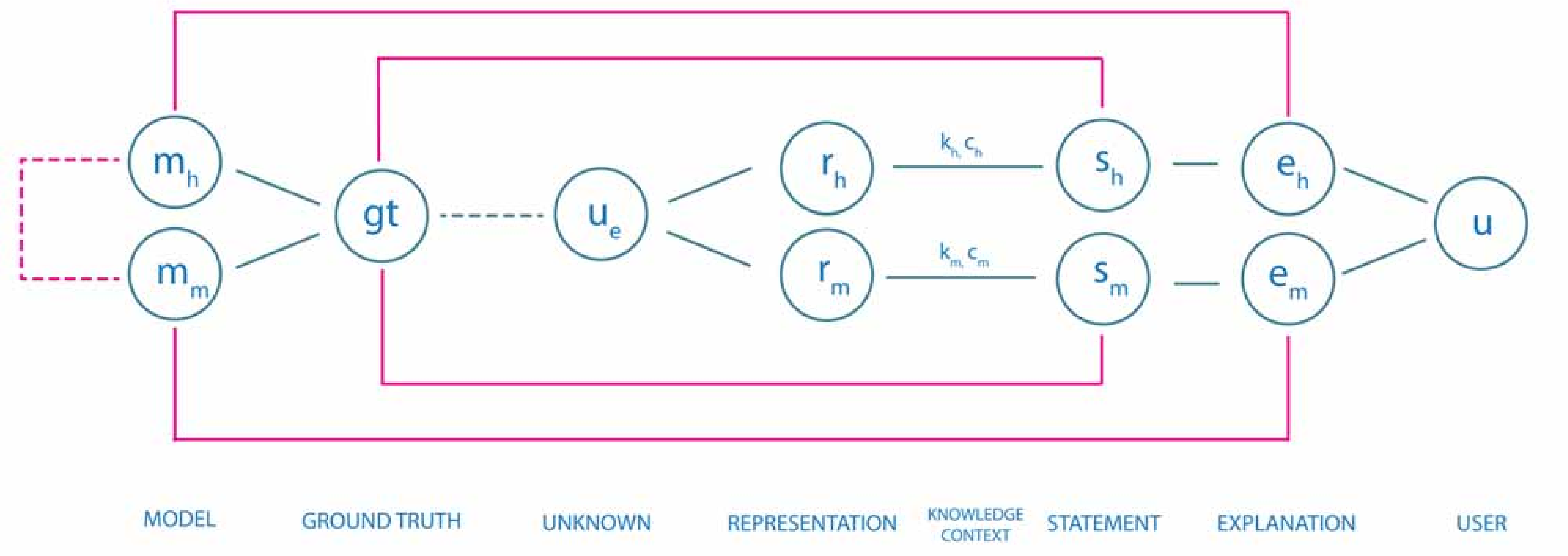


	Perceived Competence	Intention to Return	Cognitive Effort	Completion Time
Perceived Competence	1	.778** (.000)	-.826 ** (.000)	-.018 (.830)
Intention to Return	.778** (.000)	1	-.675** (.000)	-.042 (.619)
Cognitive Effort	-.826 ** (.000)	-.675** (.000)	1	.069 (.414)
Completion Time	-.018 (.830)	-.042 (.619)	.069 (.414)	1

** Correlation is significant at the 0.01 level (2-tailed).

Pearl Pu & Li Chen. Trust building with explanation interfaces. Proceedings of the 11th international conference on Intelligent user interfaces, 2006. ACM, 93-100.

Measuring Causability: Mapping machine explanations with human understanding



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.

- Computational approaches can find in R^n what no human is able to see
- However, still there are many hard problems where a human expert in R^2 can understand the **context** and bring in experience, expertise, knowledge, intuition, …
- Black box approaches can not explain **WHY** a decision has been made …

Thank you!