VISUM Summer School

VISUM

VISion Understanding and Machine intelligence

Interpretability Methodologies for Machine Learning in Medical Imaging

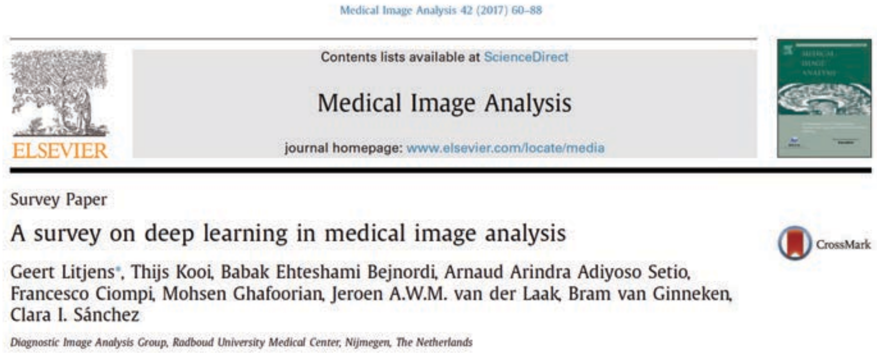
Mauricio Reyes, PhD.

Healthcare Imaging A.I.

University of Bern/ Insel Data Science Center

Table of contents

DL in Medical Imaging


Medical Image Analysis 42 (2017) 60–88

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

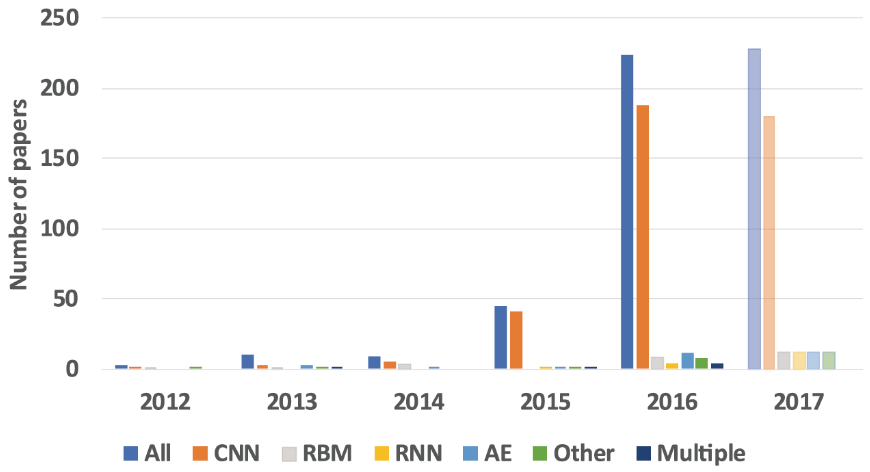
ELSEVIER

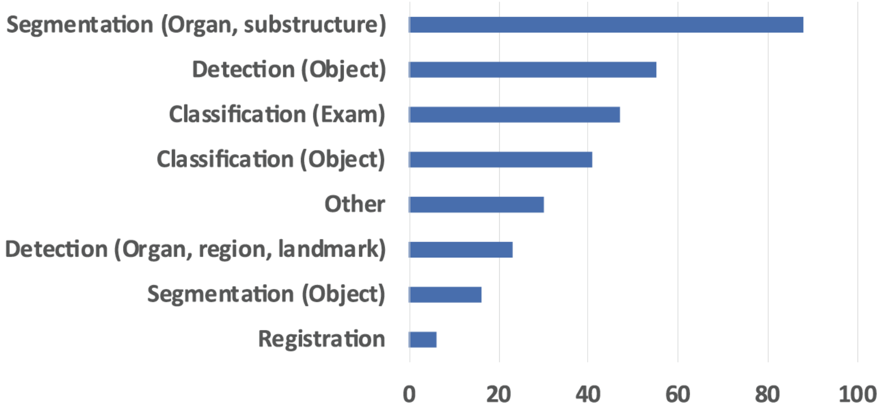
Survey Paper

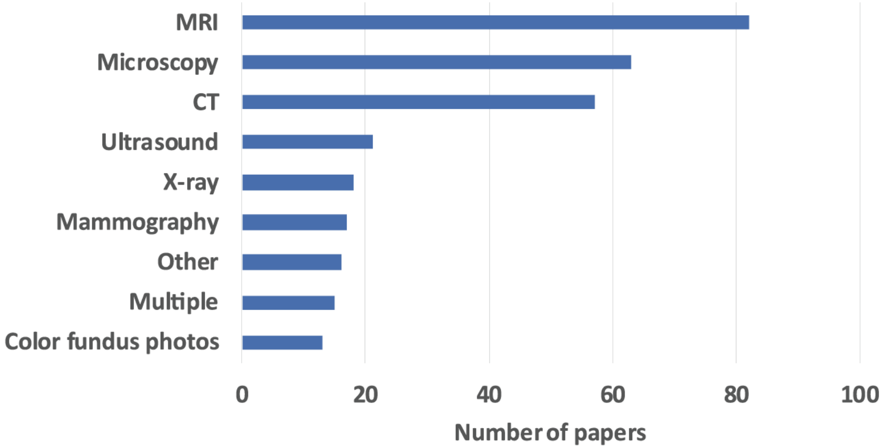
A survey on deep learning in medical image analysis

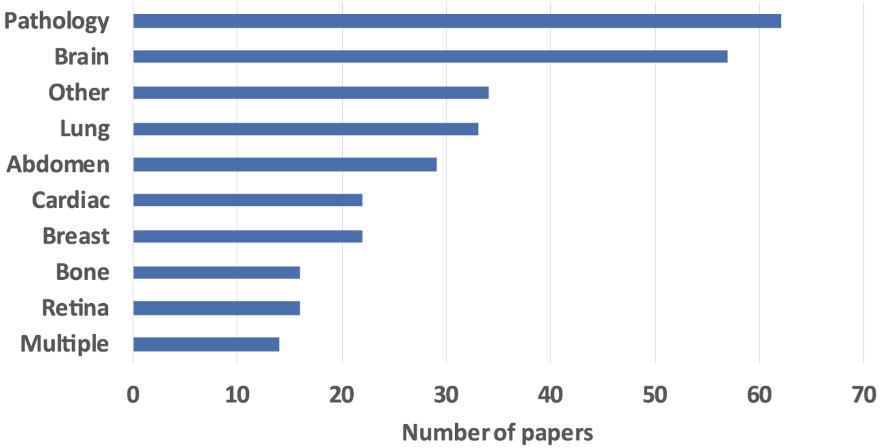
Geert Litjens*, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, Clara I. Sánchez

Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands





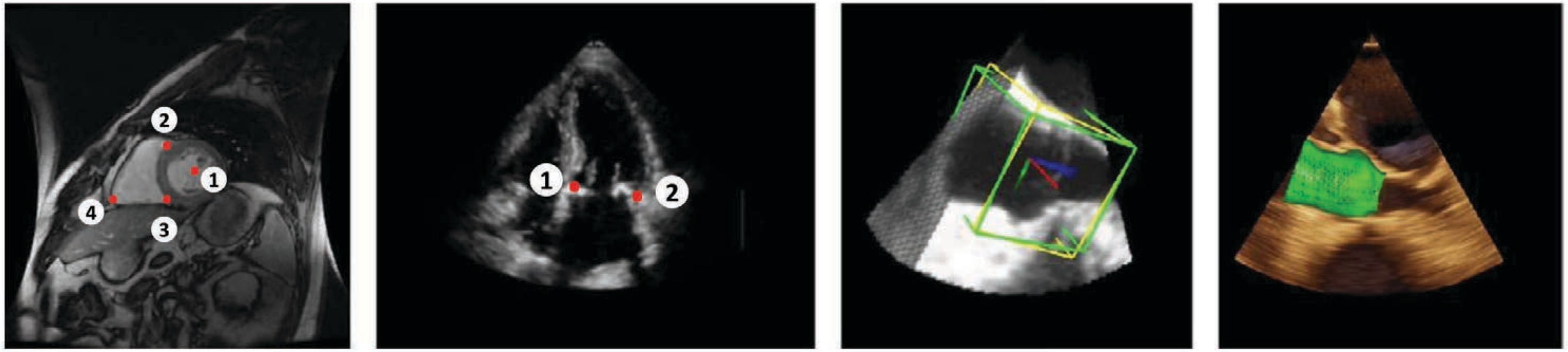


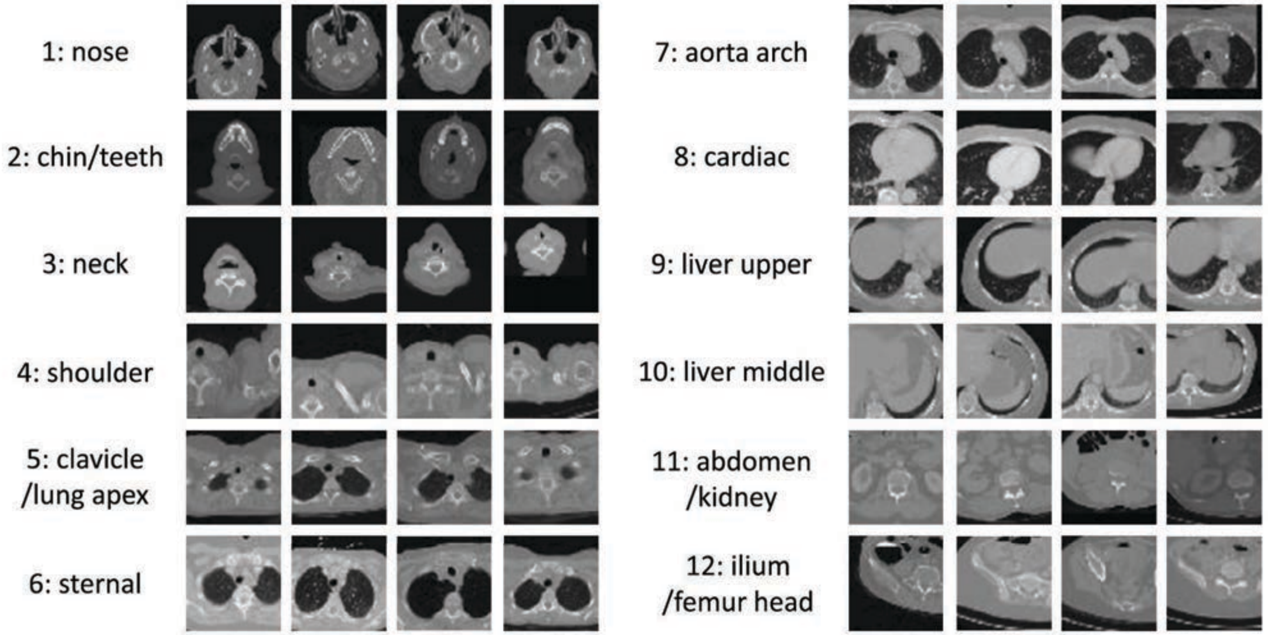


Source: Litjens et al. 2017

DL in Medical Imaging

Medical Image parsing



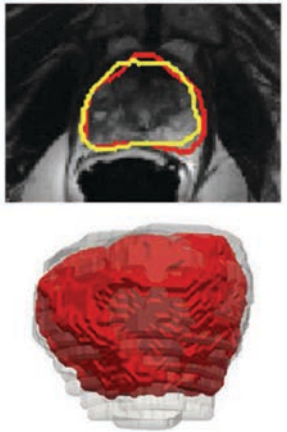


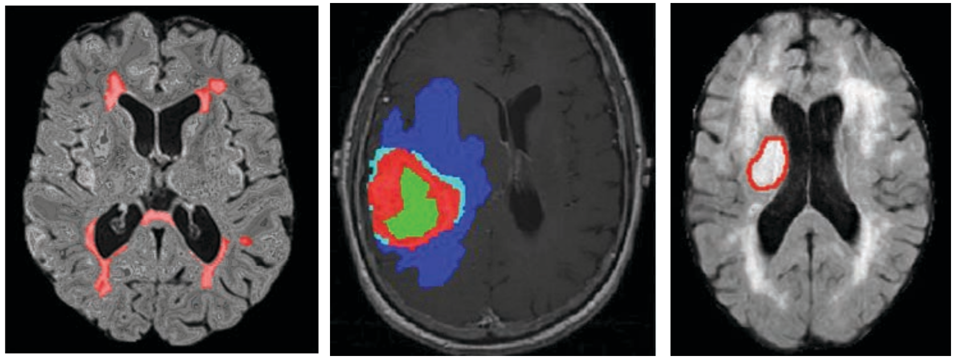
Source: Zhou et al. 2017

DL in Medical Imaging: Versatility

Medical Image Segmentation

Prostate



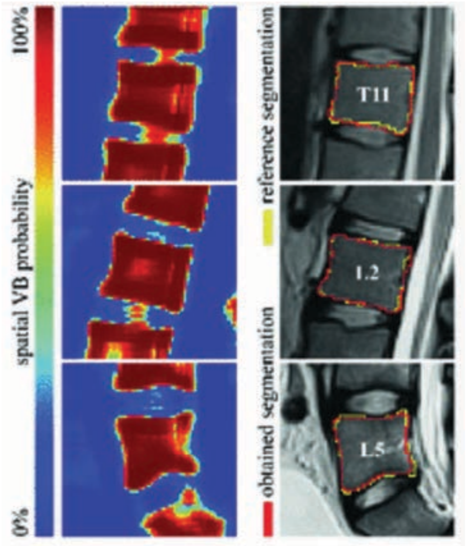


M.S

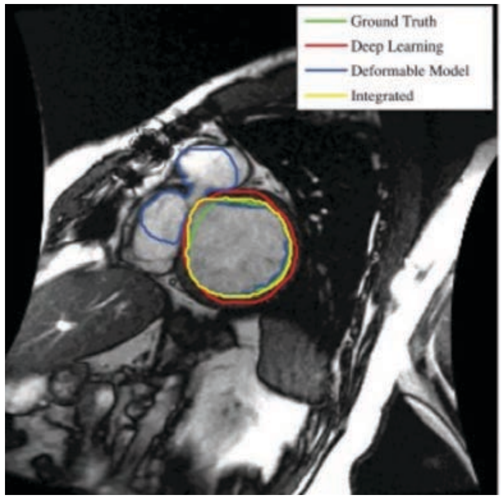
Oncology

Stroke

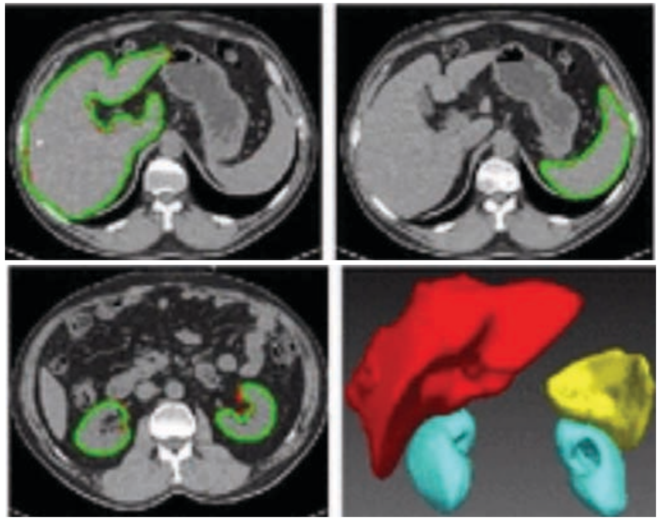
Musculoskeletal



Cardiac



Abdominal

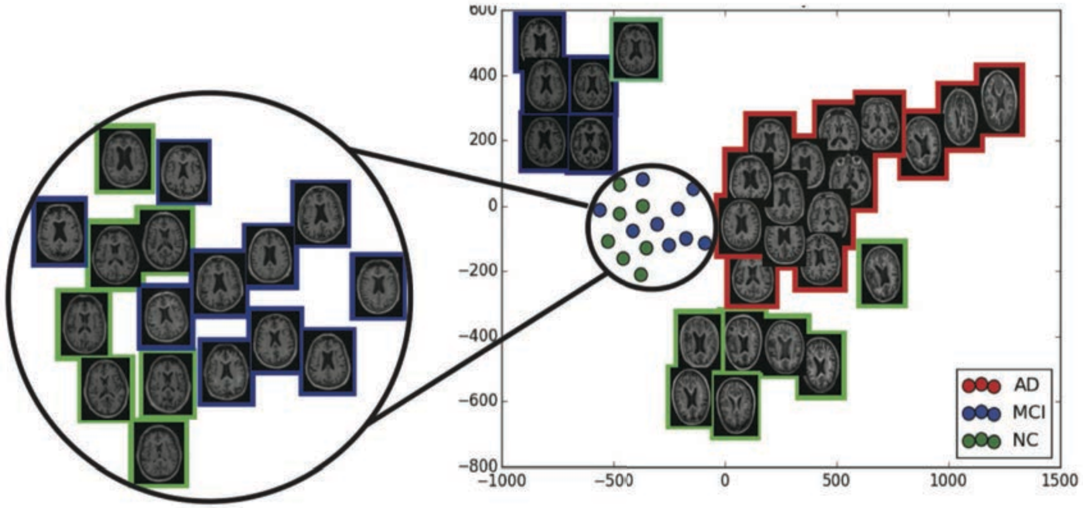


Source: Zhou et al. 2017

DL in Medical Imaging

Medical Image Classification

Chest image classification



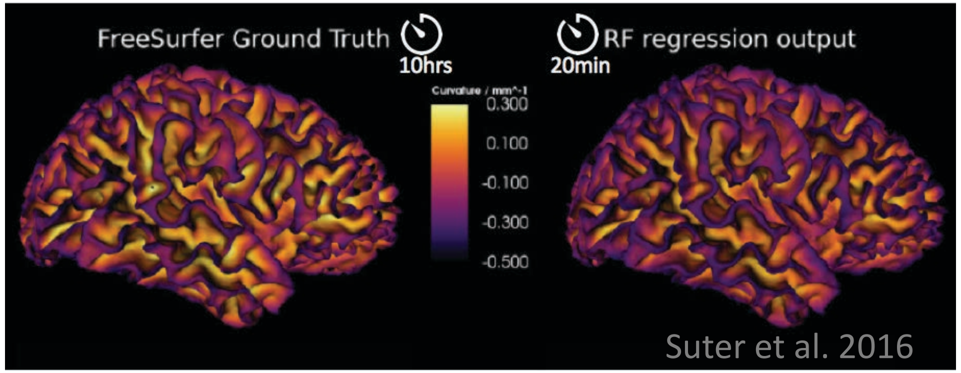
Alzheimer detection

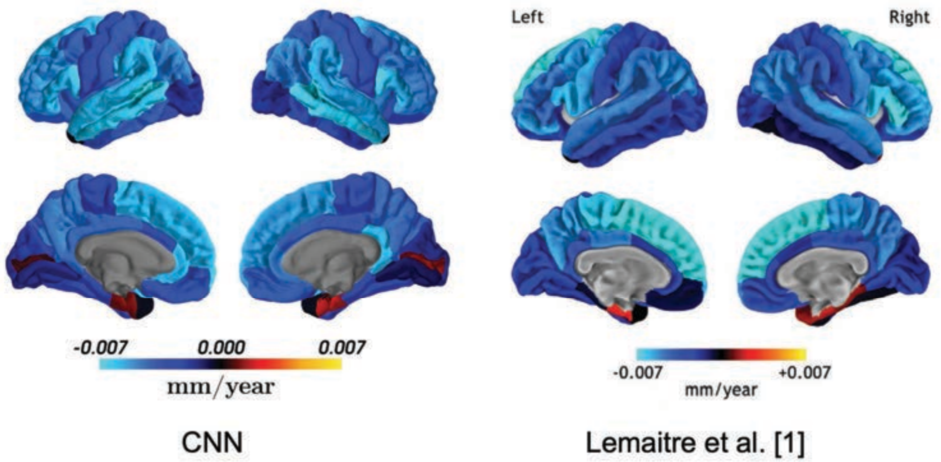
mauricio

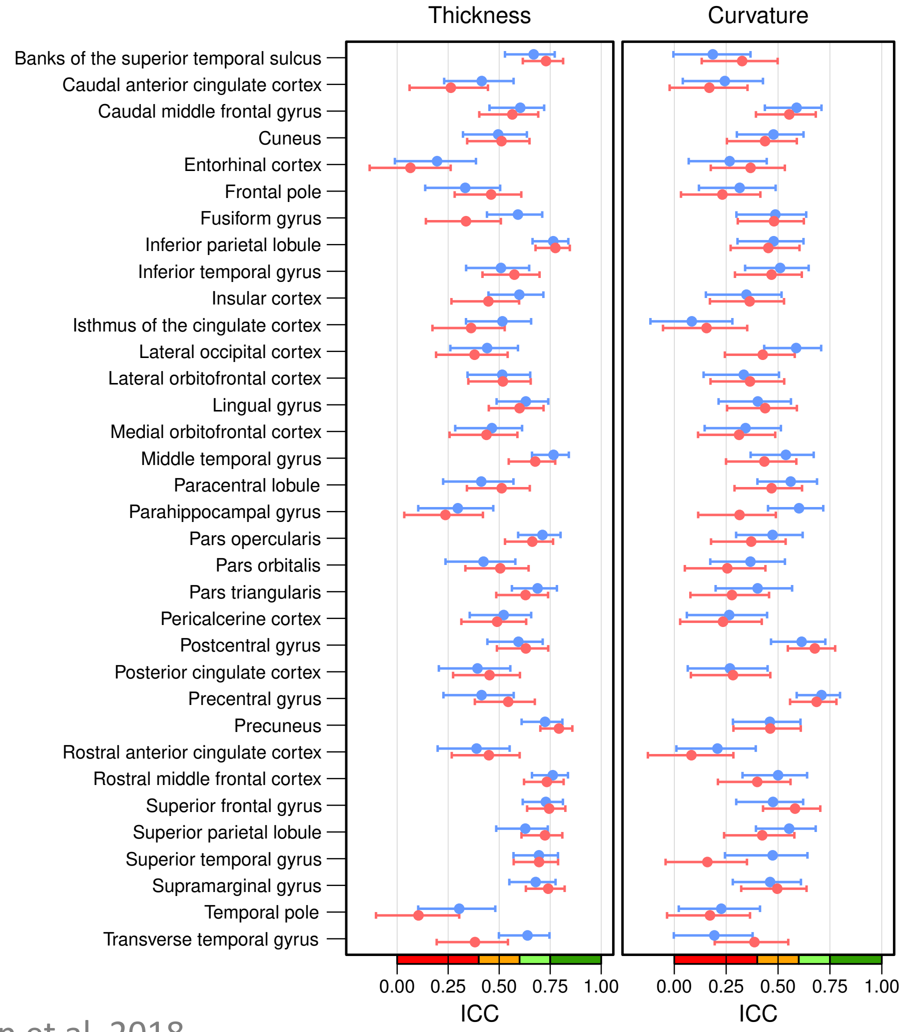
Source: Zhou et al. 2017

DL in Medical Imaging

Medical Image Quantification



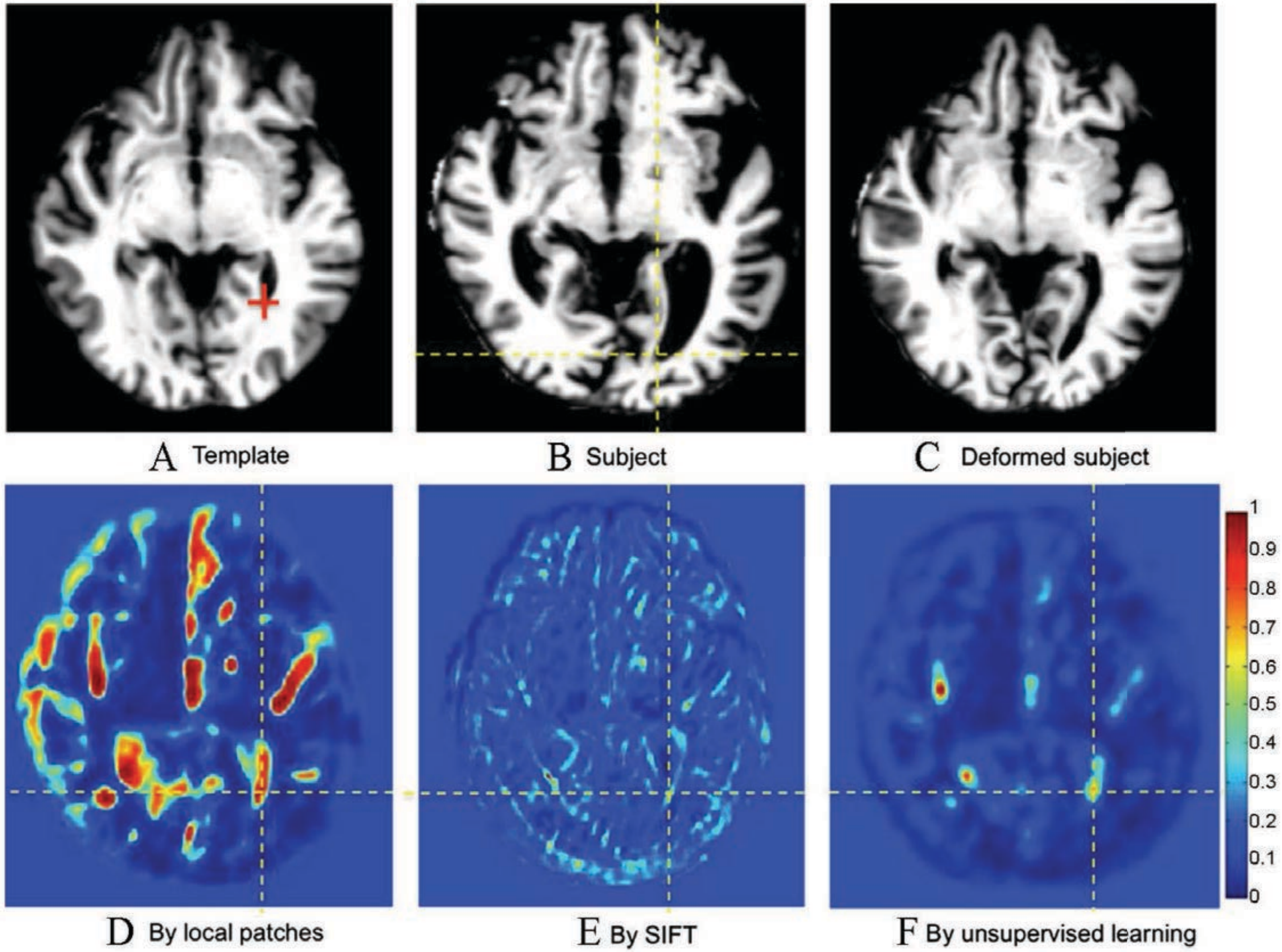




Rebsamen et al. 2018

DL in Medical Imaging

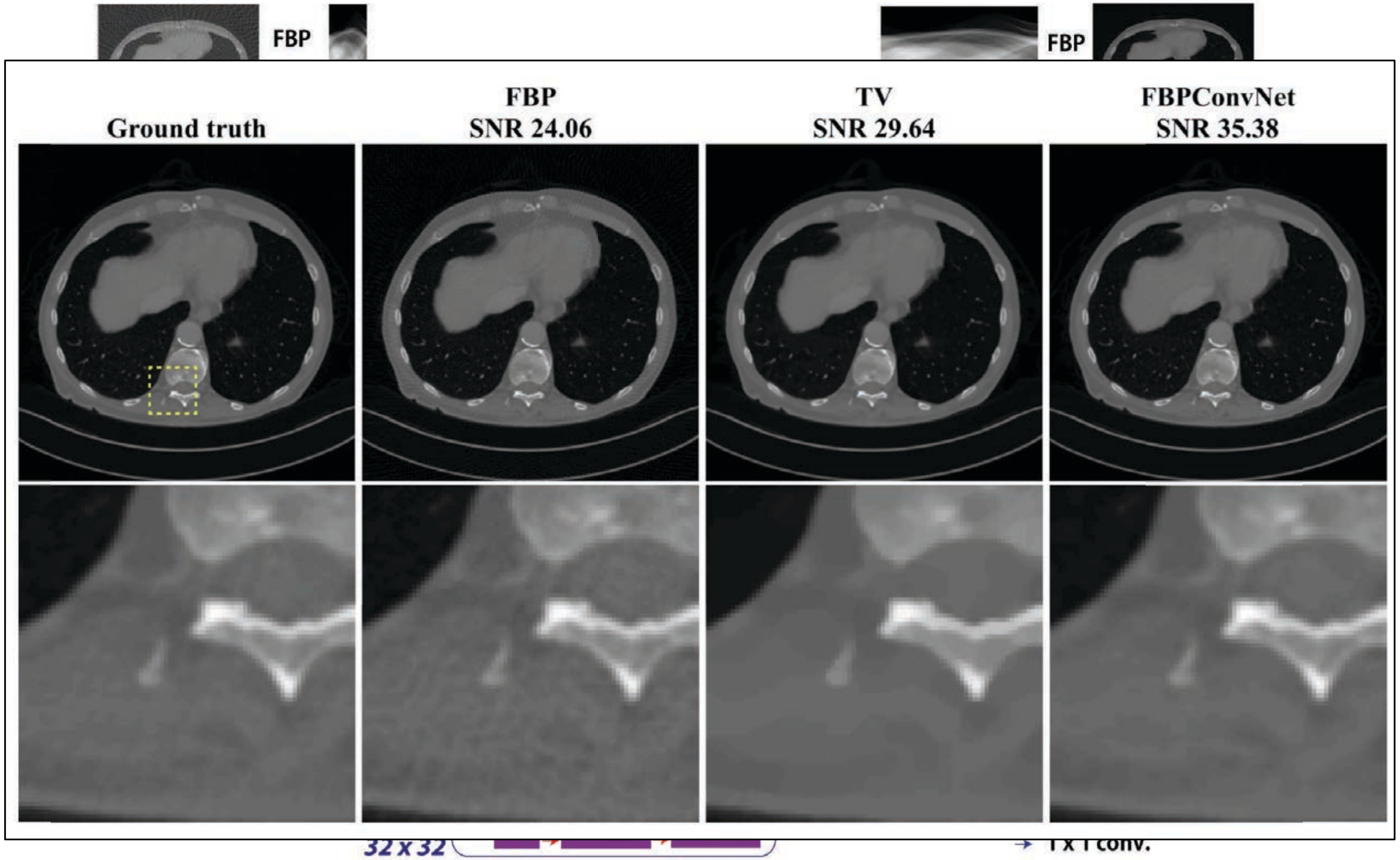
Medical Image Registration



Source: Zhou et al. 2017

DL in Medical Imaging

Medical Image Reconstruction

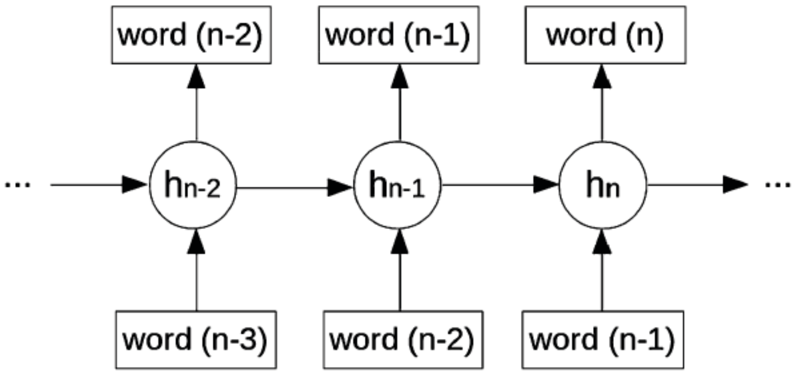


Source: Hwan-Jin TMI et al. 2017

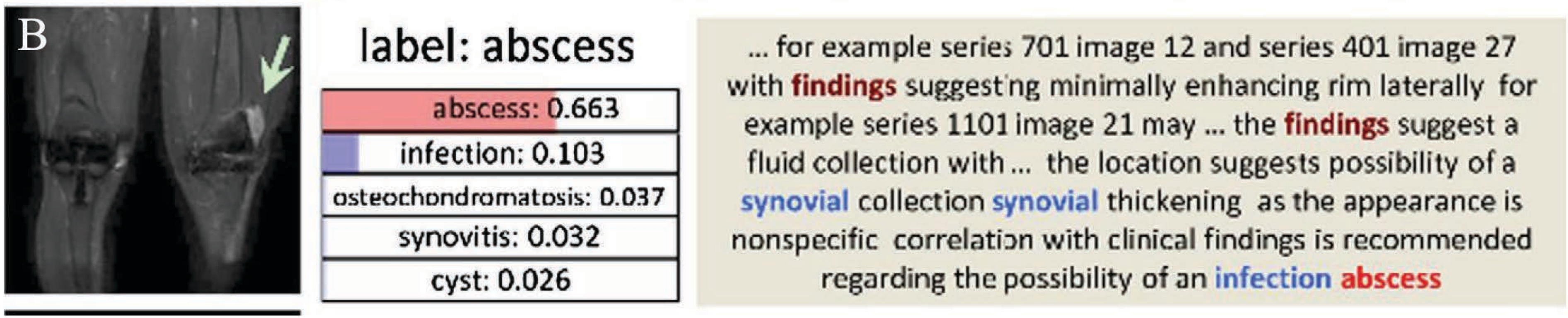
DL in Medical Imaging

Automated Image Interpretation

"heart"		"brain"		"liver"	
lungs	0.526600	t1	0.615066	spleen	0.759884
mediastinum	0.517008	mri	0.595027	gallbladder	0.648075
consolidating	0.486605	sagittal	0.580841	hepatomegaly	0.642022
pa	0.449816	flair	0.565445	gallstones	0.611837
chest	0.433362	t2	0.555053	pancreas	0.608356
infiltrates	0.428404	axial	0.554040	gallstone	0.606063
hyperinflated	0.413326	spgr	0.520954	steatosis	0.601081
cardiomegaly	0.410785	weighted	0.502047	dome	0.594812
hyperlucent	0.400836	technique	0.487768	portal	0.570008
pectus	0.396142	astrocytoma	0.480527	ascites	0.551869
great	0.395712	gbm	0.476956	hepatosplenomegaly	0.540501
ectatic	0.394560	gradient	0.476593	hepatic	0.537453
shifted	0.389205	oligodendroglioma	0.465892	cirrhosis	0.530389
ray	0.389091	postcontrast	0.463686	fatty	0.522134



Word neural embeddings

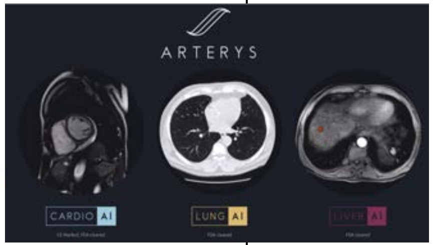


Source: Zhou et al. 2017

Feb. 2018

Arterys Receives First FDA Clearance for Broad Oncology Imaging Suite with Deep Learning

FDA clearance covers all solid tumors. Initial launch will include Liver AI and Lung AI oncology software to empower clinicians to quickly measure and track lesions and nodules in MRI and CT scans



As FDA signals wider AI approval, hospitals have a role to play

By Mike Miliard | May 31, 2018 | 03:27 PM

With more machine learning tools expected to get the go-ahead soon, health systems should be partnering with vendors to supply data for better algorithms.

Credit: FDA.gov

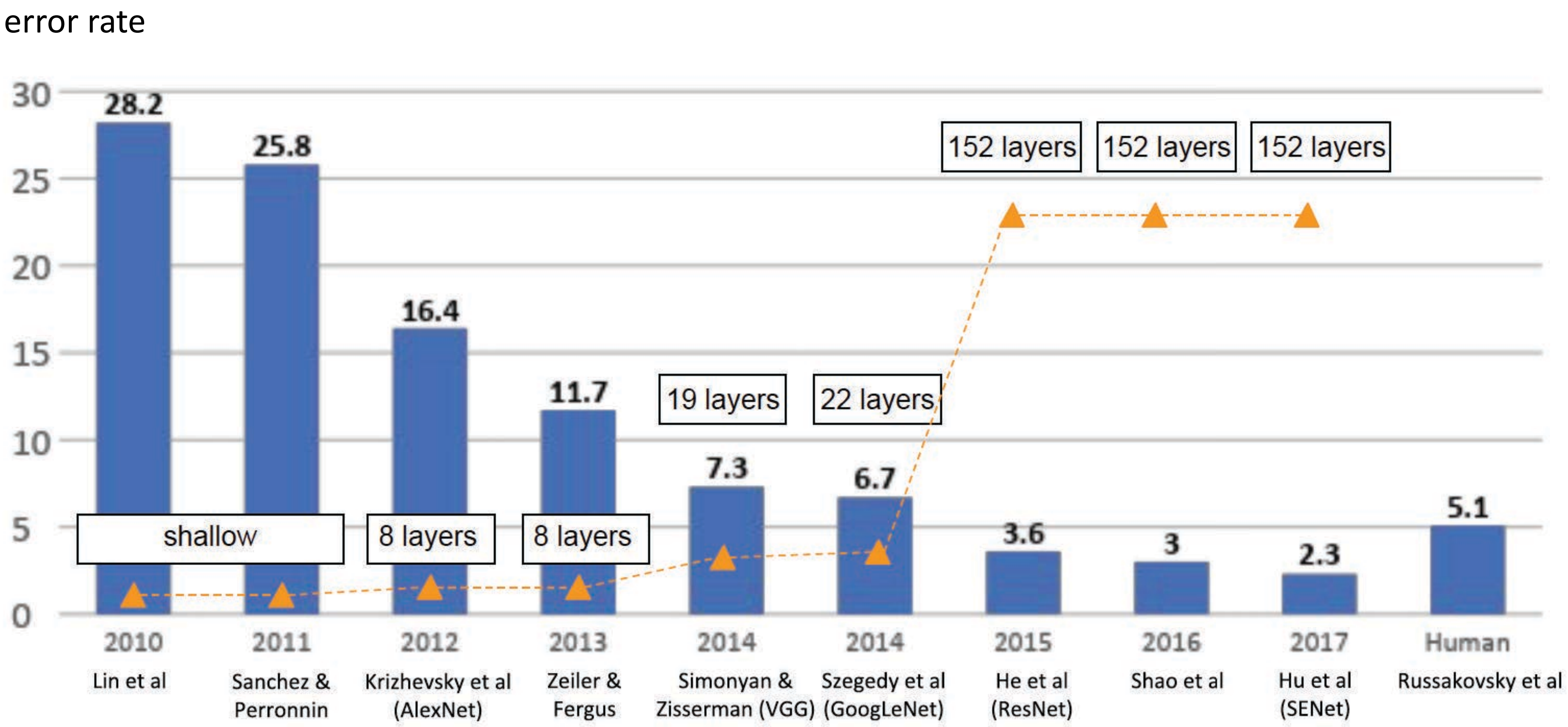
April. 2018

FDA Approves First A.I. Device To Detect Diabetic Retinopathy

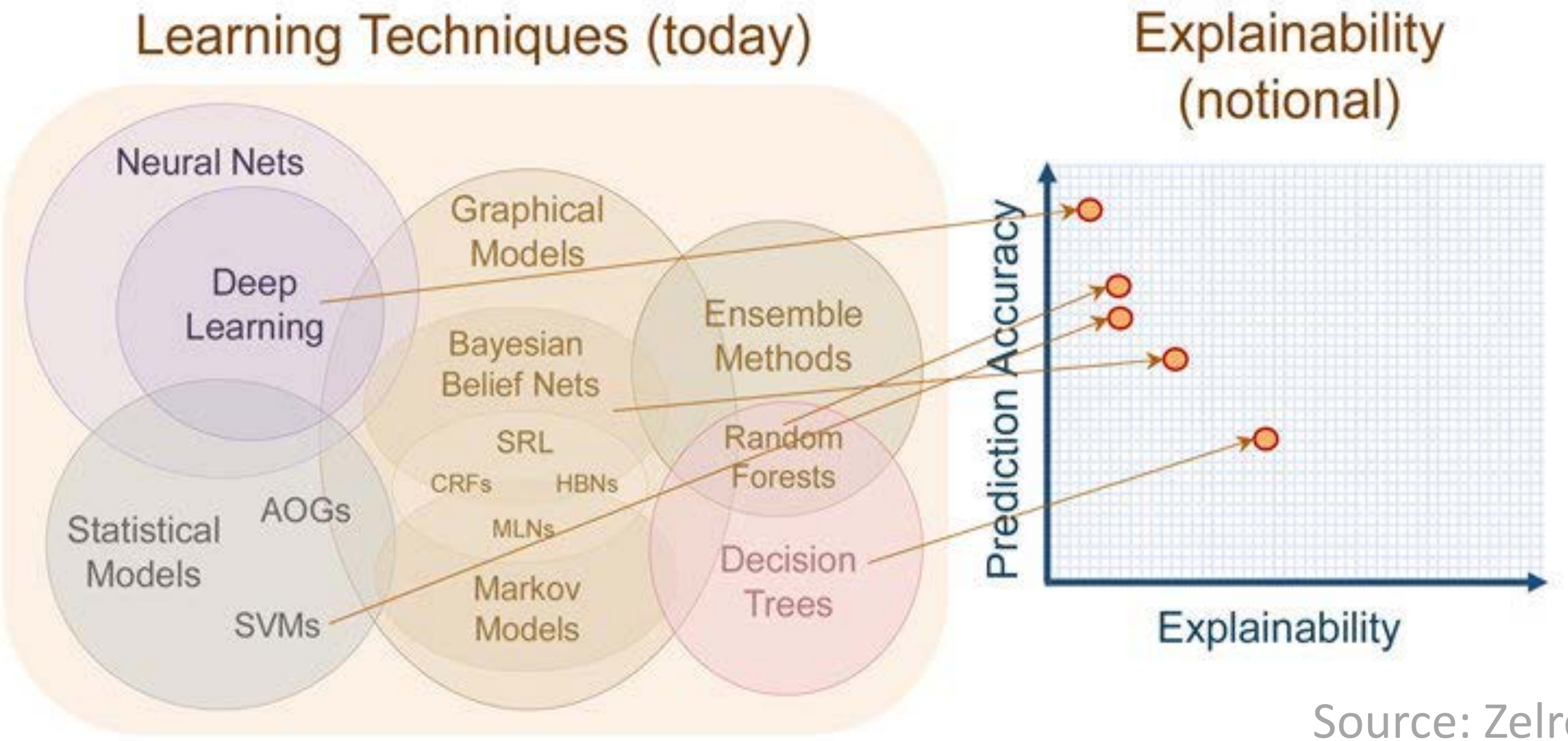
The IDx AI tool in clinical use at University of Iowa Health Care. Credit: YouTube

"A.I. holds enormous promise for the future of medicine, and we're actively developing a new regulatory framework to promote innovation in this space and support the use of AI-based technologies". Scott Gottlieb, FDA commissioner

Going deeper



Explainable A.I



Source: Zelros AI

Definitions

- What is Interpretability?

An interpretable machine learning algorithm is described as one where the **link between the features** used by the machine learning and the **prediction** itself **can be understood by a human**

Another definition

- Produce explanations without sacrificing accuracy
  - Simpler is easier to understand
  - But oversimplified is typically from an accuracy point of view, not interesting

Barocas, S., Friedler, S., Hardt, M., Kroll, J., Venka-Tasubramanian, S., & H. Wallach, H. (2018). The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning.

Explainability and Interpretability

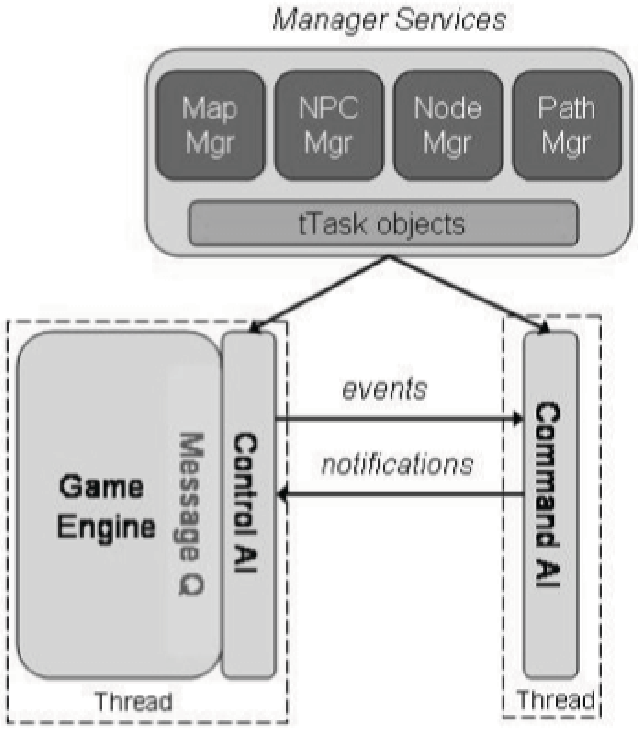
- Used interchangeably
- Explainability → "What's the process behind" (e.g. Apple falling from a tree: gravity)
- Interpretability → "understanding/predicting causal/effect phenomena" (what happens if I cut the apple from the tree: it will fall)

Why do we need Interpretability?

- Accountability
- Trust
- Quality Control
- Quality Assurance
- Ethical aspects
- Data exploration

Early uses of Explainable AI (XAI)

- Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 900–907).



Simulator/Game “Full Spectrum Command”

Early uses of Explainable AI (XAI)

<u>Recorded events:</u>

- Weapon fire
- Unit generation/update
- Task generation/Update
- Friendly Soldier Killed in Action (KIA)
- Friendly Soldier Wounded in Action (WIA)
- Enemy Soldier Killed in Action
- Enemy Soldier Wounded in Action
- Civilian Killed in Action
- Civilian Wounded in Action
- Friendly Unit First Contact with the Enemy
- Platoon Task Started
- Platoon Task Completed

<u>XAI Questions:</u>

- What is the platoon's mission?
- What is the platoon's mission status?
- How is the platoon task organized?
- How many soldiers are in the platoon?
- What is the platoon's ammo status?

Why do we need Interpretability?

- Accountability
  - System Liability

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman, Seth Flaxman

(Submitted on 28 Jun 2016 (v1), last revised 31 Aug 2016 (this version, v3))

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which "significantly affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.

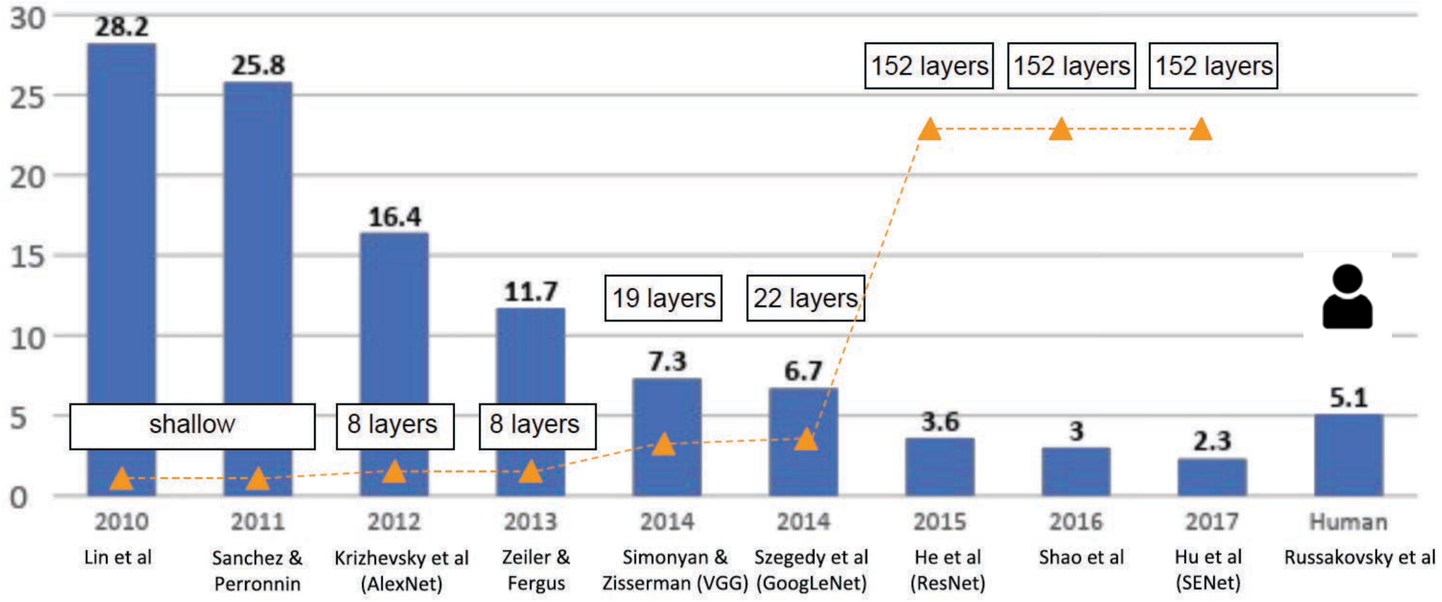
Comments: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY

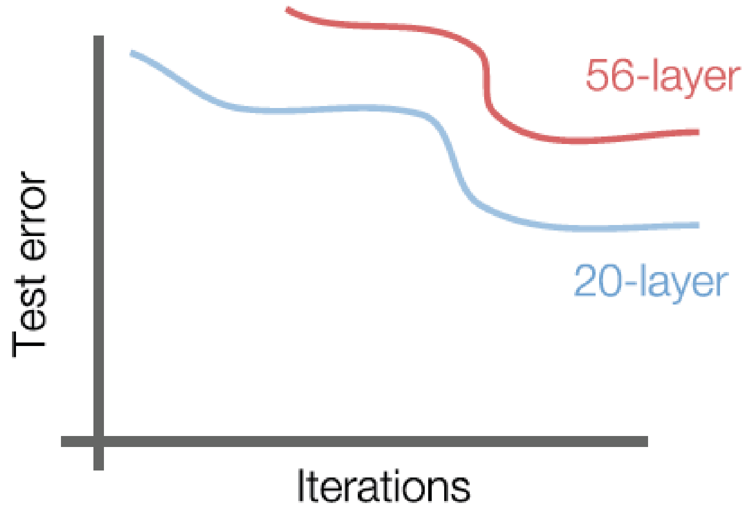
Why do we need Interpretability?

- Quality Assurance: Thorough evaluation of a model’s performance
- Ensuring robustness and knowledge of breaking points of an algorithm

Interpretability is needed to leverage Quality Control (QC) of A.I. systems

Solution featuring "high degrees of freedom"



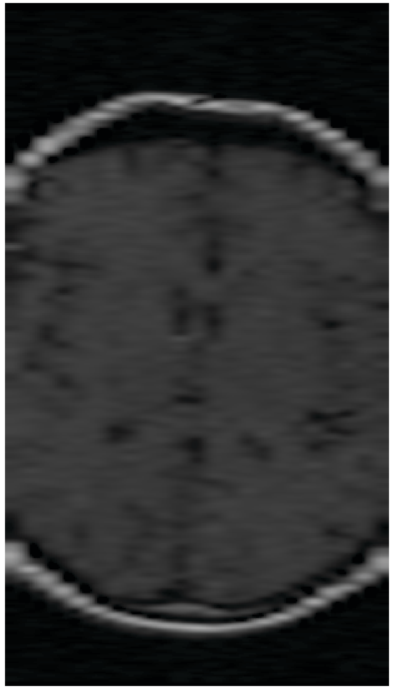


Why do we need Interpretability?

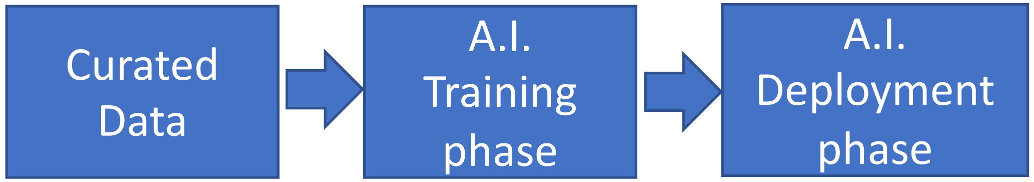
- Quality Control

flawed orientation

flawed orientation



low resolution



A.I. technology is as good as the training data characterizing the task

Example: Automated MS Segmentation – Multi-Center Analysis

SCIENTIFIC REPORTS

OPEN

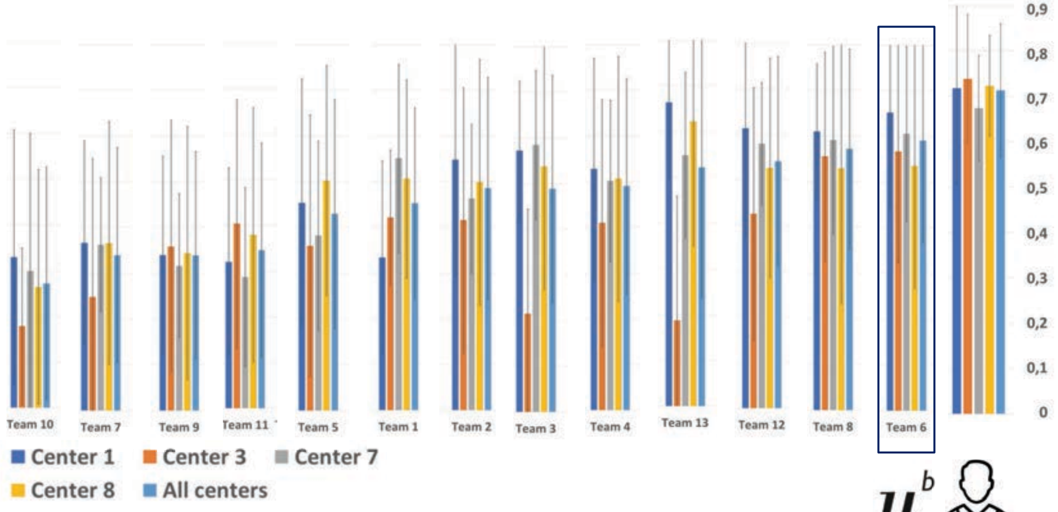
Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure

Received: 20 April 2018
Accepted: 6 August 2018
Published online: 12 September 2018

Olivier Commowick[1], Audrey Istace[2], Michaël Kain[1], Baptiste Laurent[3], Florent Leray[1], Mathieu Simon[1], Sorina Camarasu Pop[4], Pascal Girard[4], Roxana Améli[2], Jean-Christophe Ferré[5,1], Anne Kerbrat[6,1], Thomas Tourdias[7], Frédéric Cervenansky[4], Tristan Glatard[8], Jérémy Beaumont[1], Senan Doyle[9], Florence Forbes[3,10], Jesse Knight[11], April Khademi[12], Amirreza Mahbod[13], Chunliang Wang[13], Richard McKinley[14], Franca Wagner[14], John Muschelli[15], Elizabeth Sweeney[15], Eloy Roura[16], Xavier Lladó[16], Michel M. Santos[17], Wellington P. Santos[18], Abel G. Silva-Filho[17], Xavier Tomas-Fernandez[19], Hélène Urien[20], Isabelle Bloch[20], Sergi Valverde[16], Mariano Cabezas[16], Francisco Javier Vera-Olmos[21], Norberto Malpica[21], Charles Guttmann[22], Sandra Vukusic[2], Gilles Edan[6,1], Michel Dojat[23], Martin Styner[24], Simon K. Warfield[19], François Cotton[2] & Christian Barillot[1]

Scanner	Modality	Matrix	Slices	Voxel resolution (mm)
GE Discovery 3T	Sagittal 3D FLAIR Sagittal 3D T1 Axial 2D DP-T2	512 × 512 512 × 512 512 × 512	224 248 From 28 to 44	0.47 × 0.47 × 0.9 0.47 × 0.47 × 0.6 0.43 × 0.43 × 3 Gap: 0.5
Philips Ingenia 3T	Sagittal 3D FLAIR Sagittal 3D T1 Axial 2D PD-T2	336 × 336 336 × 336 512 × 512	261 200 46	0.74 × 0.74 × 0.7 0.74 × 0.74 × 0.85 0.45 × 0.45 × 3
Siemens Aera 1.5T	Sagittal 3D FLAIR Sagittal 3D T1 Axial 2D PD-T2	256 × 224 256 × 256 320 × 320	128 176 25	1.03 × 1.03 × 1.25 1.08 × 1.08 × 0.9 0.72 × 0.72 × 4 Gap: 1.2
Siemens Verio 3T	Sagittal 3D FLAIR Sagittal 3D T1 Axial 2D PD-T2	512 × 512 256 × 256 240 × 320	144 176 44	0.5 × 0.5 × 1.1 1 × 1 × 1 0.69 × 0.69 × 3

Multi-vendor, multi-center



- Center 3 (orange) → Unknown to models, lowest performance

Why do we need Interpretability?

- Trust
  - Essential in fields such as medicine where technology can play a crucial role
  - In radiology, AI and the potential of restructuring the radiologist’s workflow

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

DeepFool: a simple and accurate method to fool deep neural networks

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard
École Polytechnique Fédérale de Lausanne
{seyed.moosavi,alhussein.fawzi,pascal.frossard} at epfl.ch

Full Citation: Nguyen A, Yosinski J, Clune J. *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*. In Computer Vision and Pattern Recognition (CVPR '15), IEEE, 2015.

**Deep Neural Networks are Easily Fooled:
High Confidence Predictions for Unrecognizable Images**

Anh Nguyen, University of Wyoming, anguyen8@uwyo.edu
Jason Yosinski, Cornell University, yosinski@cs.cornell.edu
Jeff Clune, University of Wyoming, jeffclune@uwyo.edu

Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org

PLOS | MEDICINE — BROWSE PUBLISH ABOUT

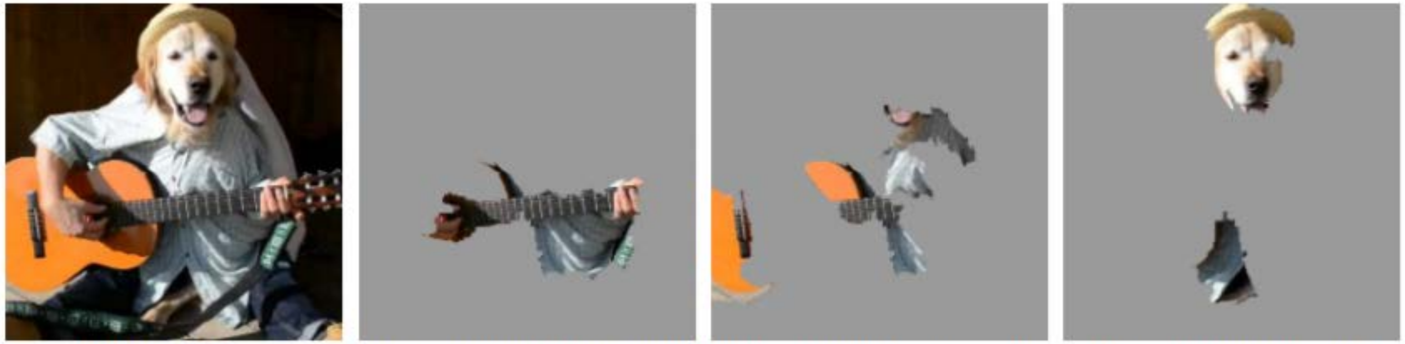
OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, Eric Karl Oermann

Published: November 6, 2018 • https://doi.org/10.1371/journal.pmed.1002683

Explaining DL decisions via attention maps

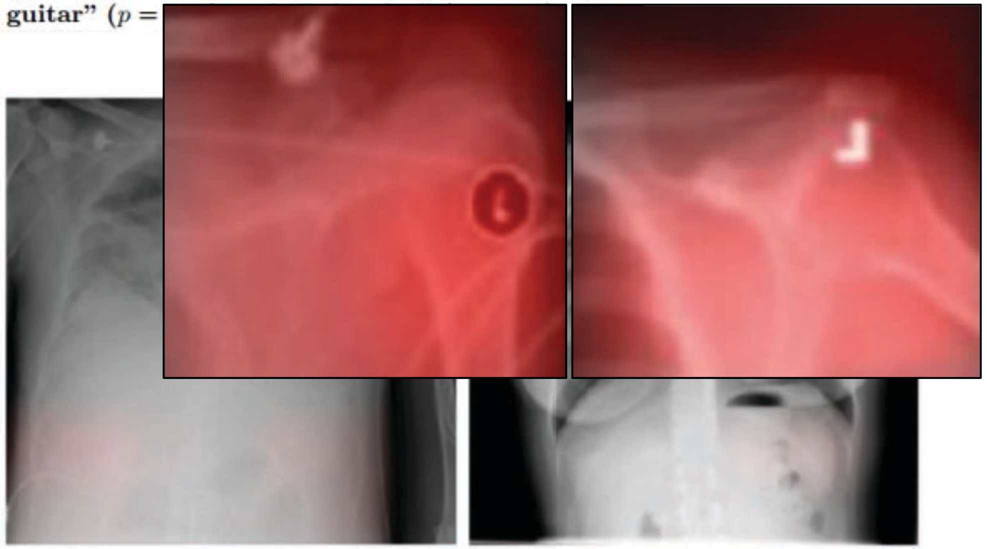


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p =$

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.



DL learned to detect clinical site!

Source: Ribeiro et al. 2017, Zech et al. 2018

Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org

Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

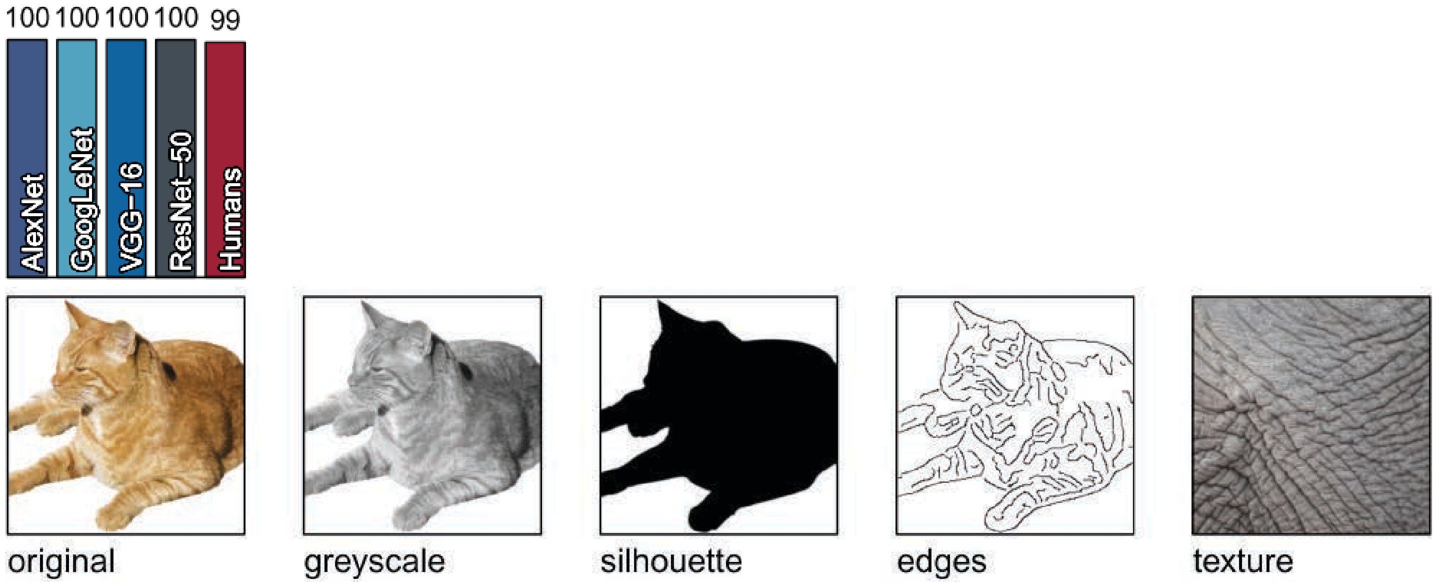
Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org



Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

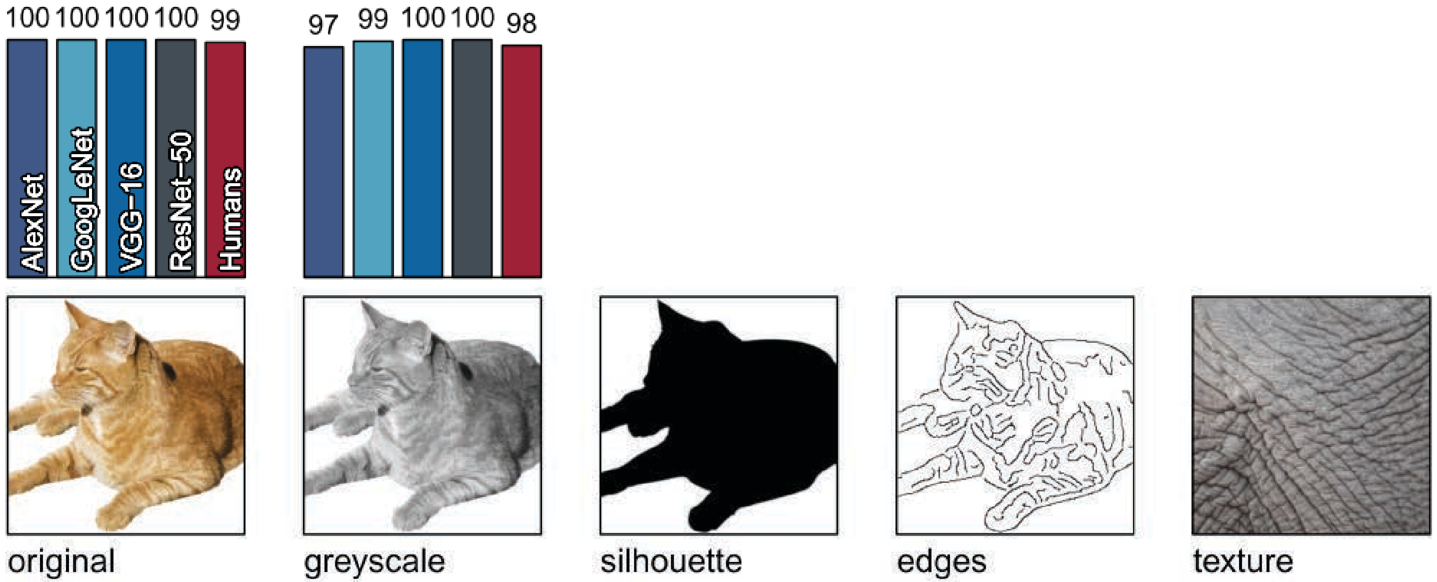
Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org



Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

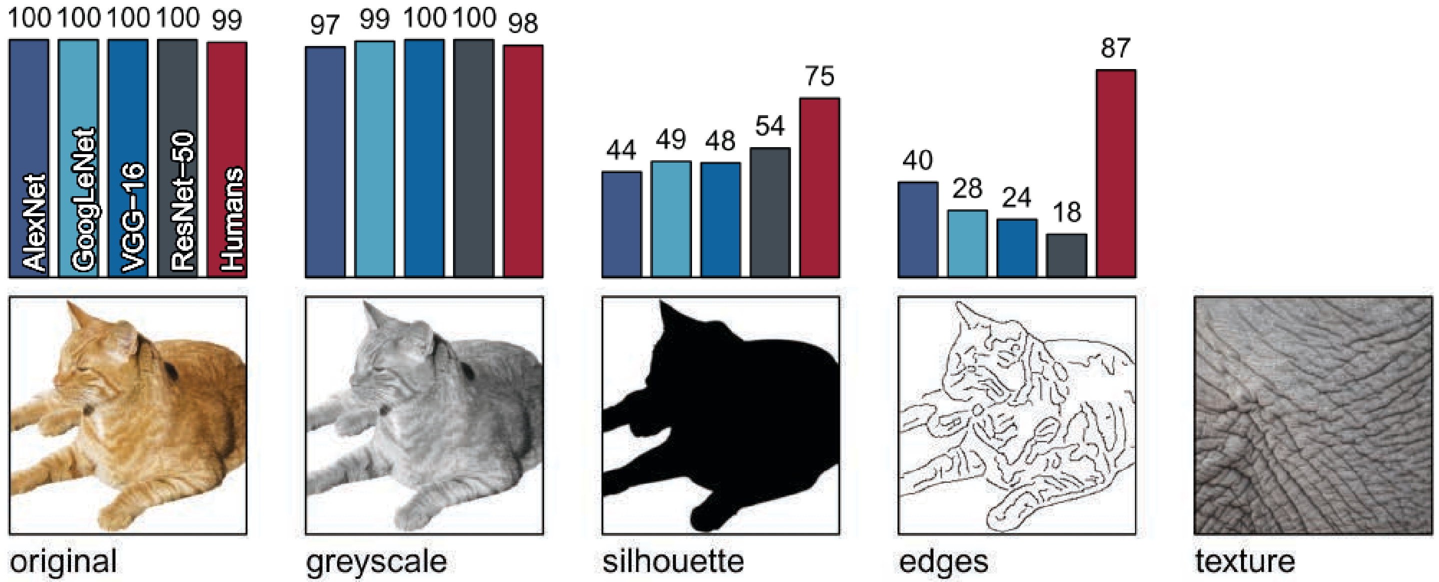
Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org



Published as a conference paper at ICLR 2019

IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Robert Geirhos
University of Tübingen & IMPRS-IS
robert.geirhos@bethgelab.org

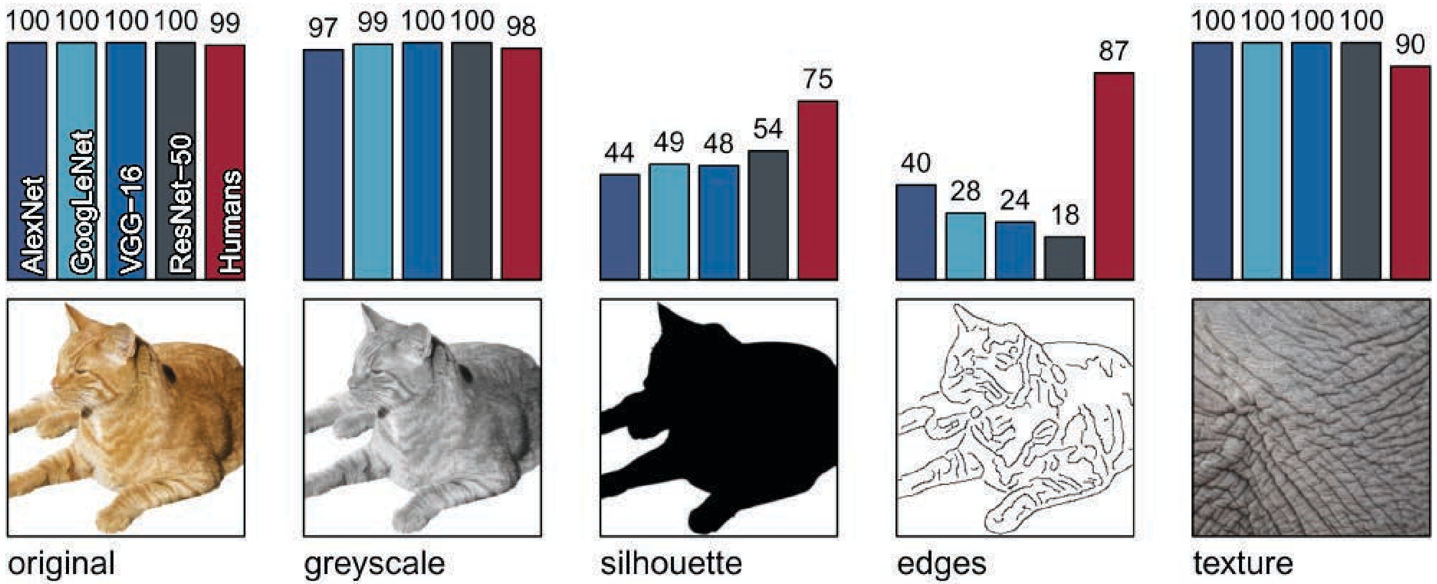
Patricia Rubisch
University of Tübingen & U. of Edinburgh
p.rubisch@sms.ed.ac.uk

Claudio Michaelis
University of Tübingen & IMPRS-IS
claudio.michaelis@bethgelab.org

Matthias Bethge*
University of Tübingen
matthias.bethge@bethgelab.org

Felix A. Wichmann*
University of Tübingen
felix.wichmann@uni-tuebingen.de

Wieland Brendel*
University of Tübingen
wieland.brendel@bethgelab.org



Ethics of AI in Radiology: European and North American Multisociety Statement

Transparency, interpretability, and explainability

Transparency, interpretability, and explainability are ***necessary to build patient and provider trust****. When a radiologist makes a mistake, we want to know why, in part because we want to know whether the mistake is excusable. We want to know whether the mistake reflects malintent or negligence, or occurred due to other factors.*

Similarly, if an algorithm fails or contributes to an adverse clinical event or malpractice, radiologists need to be able to understand ***why*** *it produced the result that it did, and* ***how*** *it reached a decision.*

”

https://www.acr.org/-/media/DSI/Files/PDFs/Ethics-of-AI-in-Radiology-Statement_RFC.pdf?la=en

Enhancing Interpretability of machine lear[redacted]

Taxonomy

- **Black-boxes** operating methods: No need to have “access” to the internal of models. Also referred to as model-agnostic.
- **White-boxes**: These are methods that require access to the internals of the models. Gradient-based models are one example of white box interpretability models.
- By **output**:
  - Visualization, or saliency maps: Provide pixel-wise values reflecting their importance to the performance of the model being interpreted
  - Concepts: summary statement or keyword.
  - Feature importance: explain models by expressing importance of features E.g. high weights of a model.

Let’s check some of these approaches

Christoph Molnar @ChristophMolnar · May 17

Explanation algorithms for neural network predictions are like yogurts in a supermarket. Choice overload.

LRP
LIME
Integrated Gradients
deconvnet
Occlusion
DeepShap
DeepLift
Deep Taylor Decomposition
Gradient* Input
Grad-CAM
Guided Backpropagation
Saliency Map
...

Partial Dependence Plot (PDP) – Friedman et al. 2001

- shows the marginal effect that one or two features have on the predicted outcome of a machine learning model
- PDP can show whether the relationship between the target and a feature is linear, monotonous or more complex.

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)d\mathbb{P}(x_C) \quad = \quad \boxed{\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)})}$$

Feature being explained – variable

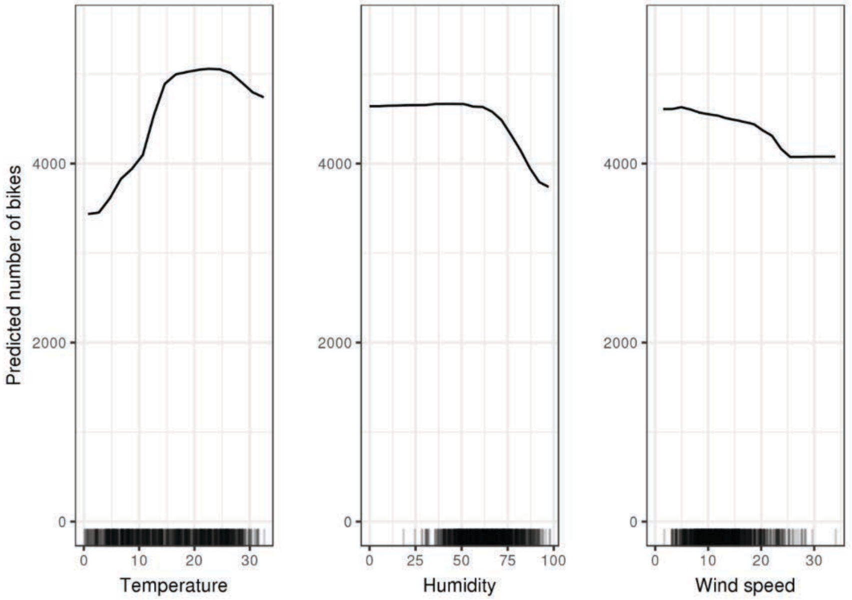
Marginalized over other features – Actual values

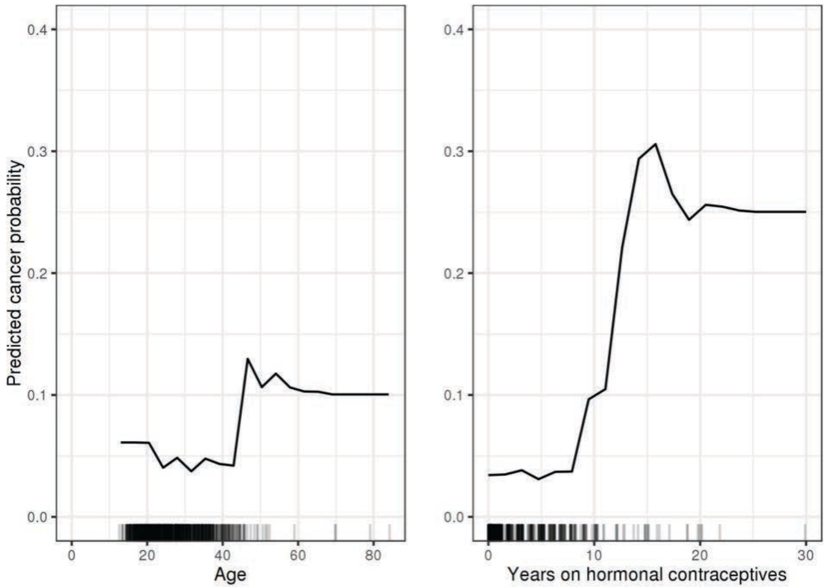
- Global method: The method considers all instances and gives a statement about the global relationship of a feature with the predicted outcome.

Source: Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Partial Dependence Plot (PDP) – Friedman et al. 2001

Example: bike rental – Relation to temperature, humidity, windspeed





<u>Pros/Cons of PDP:</u>

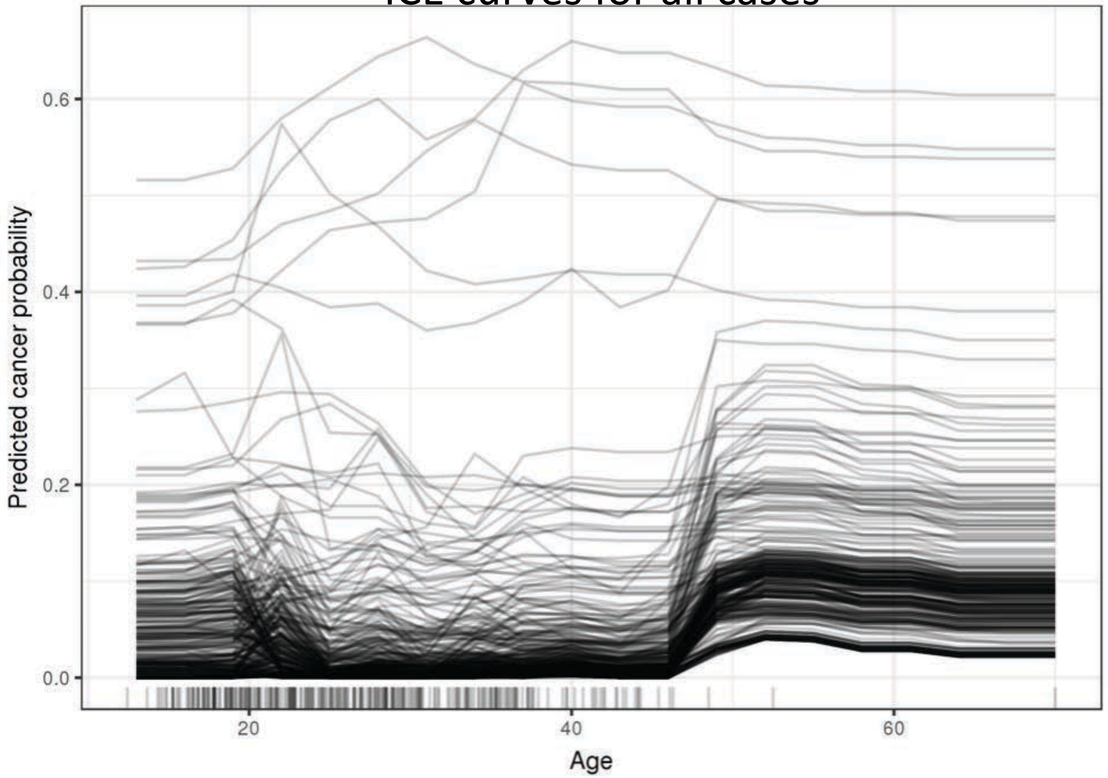
- Simple
- Easy to implement
- Does not account for correlated features
- More than 3 features is hard to visualize
- Doesn't show feature distribution
- Effects might cancel out

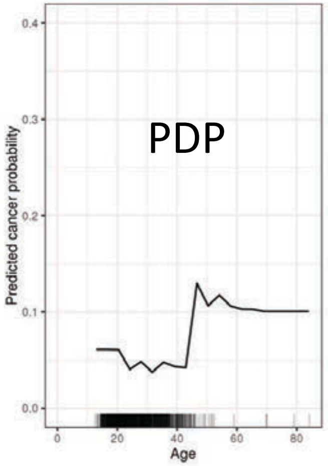
Source: Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Individual Conditional Expectation (ICE)

- ICE shows how the instance's prediction changes when a feature changes. One feature at a time
- PDP is the average of ICE over all features

ICE curves for all cases





Actual values for other features

$$\hat{f}_{x_S}(x_S) = \left\{ \hat{f}\left(x_S, x_C\right) \right\}_{i=1..,N}$$

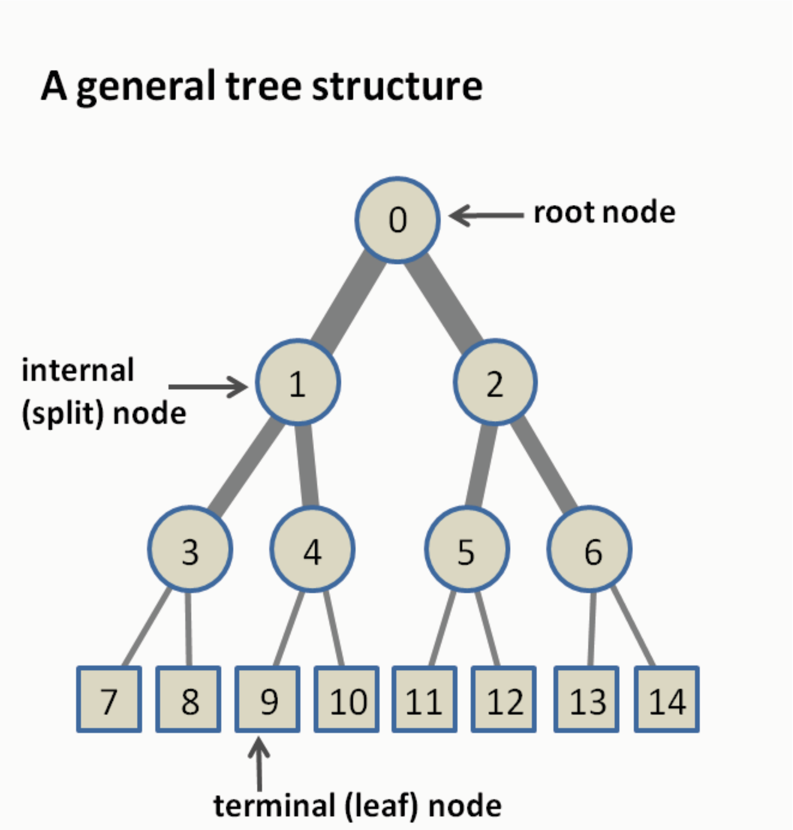
Variable - e.g. grid search

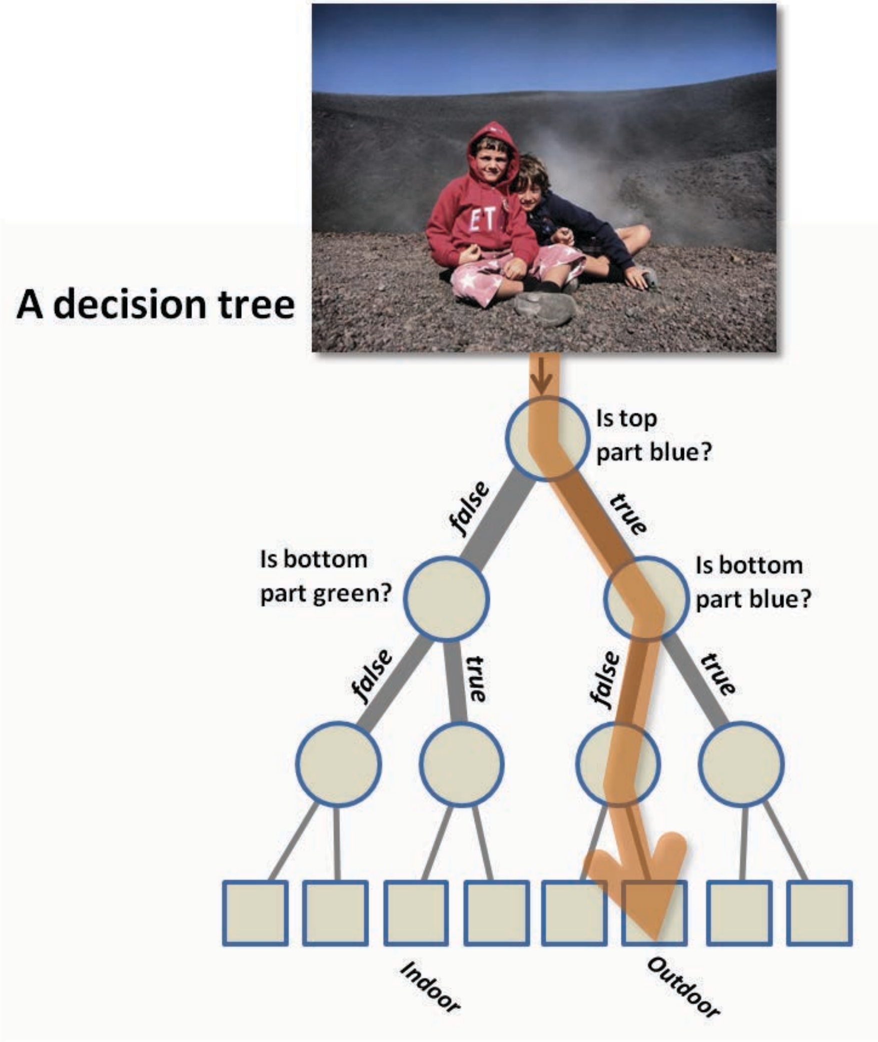
i^{th} case

Source: Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

Decision trees

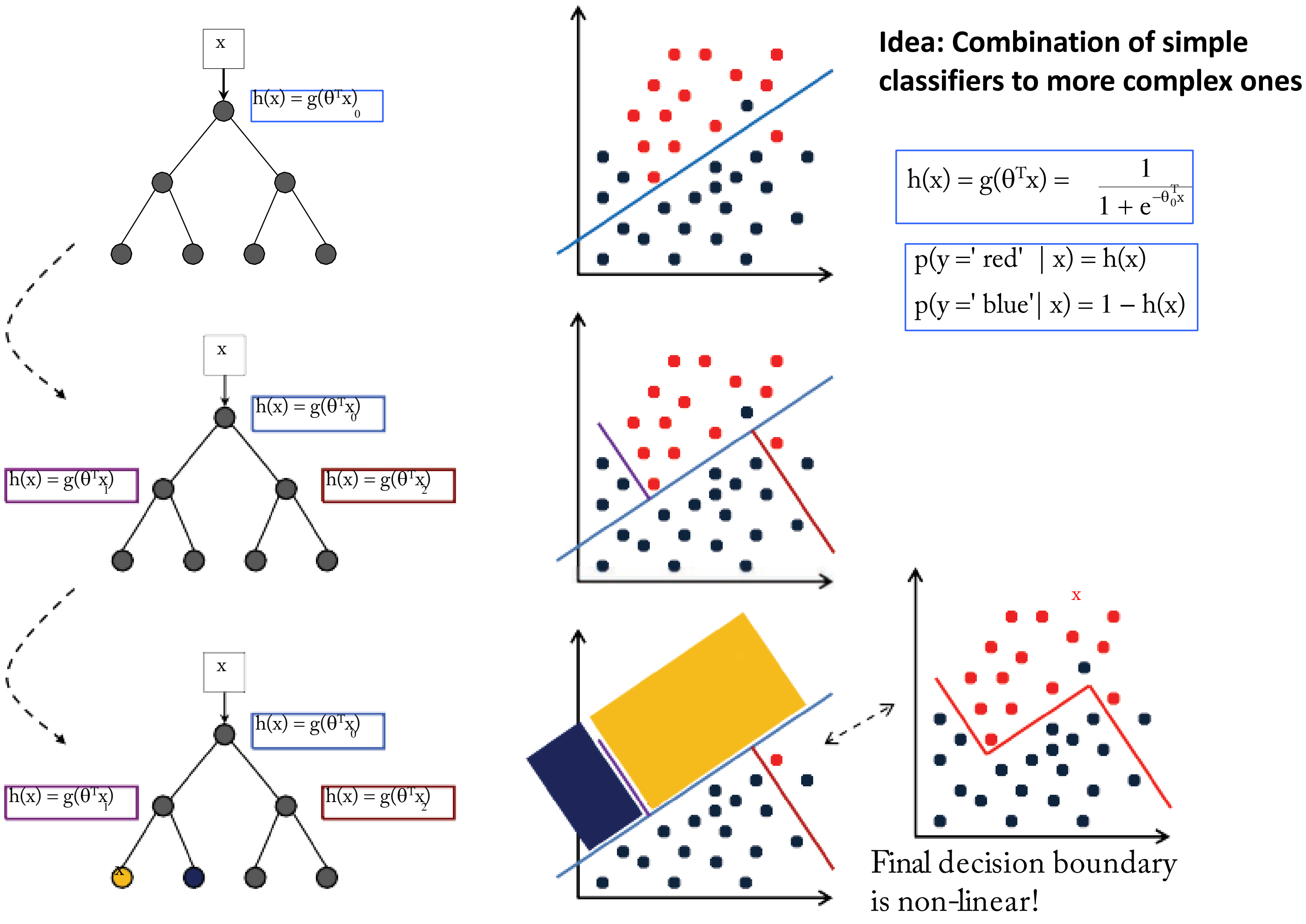
General Concept





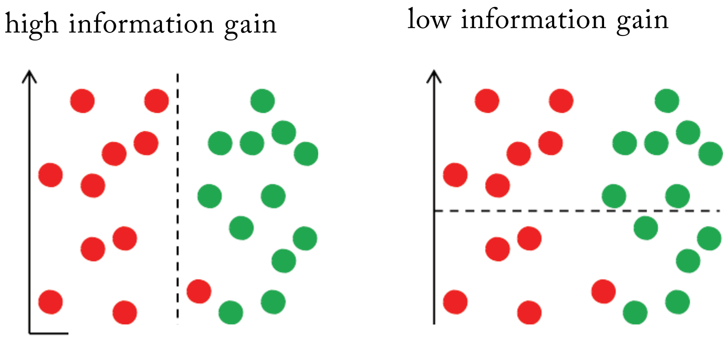
Transitioning from linear to non-linear classifier

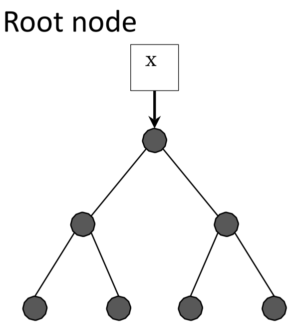
- *Small* trees structure are interpretable
- Good for visualization
- Capture interactions among features
- However, subject to depth (D): number of terminal nodes is 2^D

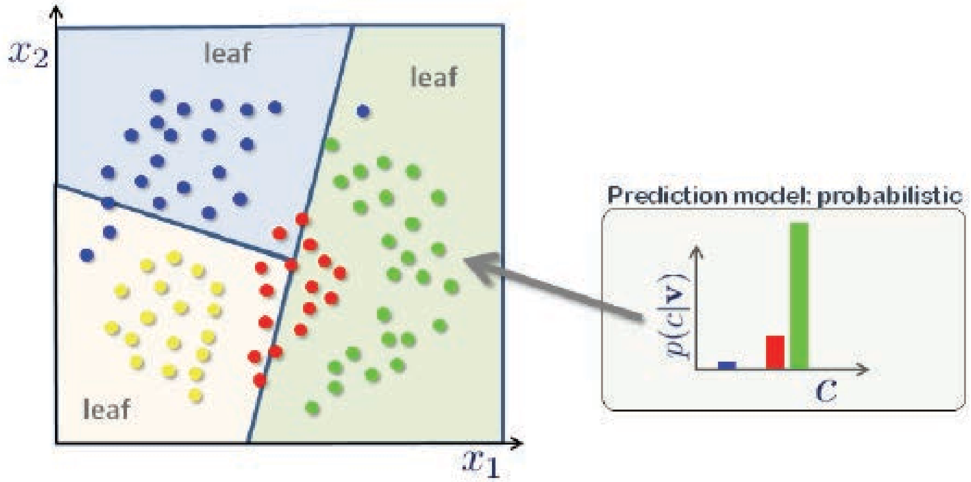


Interpretation of Decision Trees

- A model can be interpreted by analyzing the feature importance
- A feature is important when it maximally reduces impurity of split data
- Tree structure describes feature importance (e.g. root node features are most important)







<u>Information Gain - IG</u>

$$IG = H(\mathcal{S}) - \sum_{i \in \{L,R\}} \frac{|\mathcal{S}^i|}{|\mathcal{S}|} H(\mathcal{S}^i)$$

<u>Entropy</u>

$$H(\mathcal{S}) = \sum_{y \in \mathcal{Y}} p(y) \log p(y).$$

Alternative, <u>Gini Index</u>

$$\text{GI} = 1 - \sum p(y)$$

- IG favors smaller but varied partitions; GI favors larger partitions

Linking note to DL methods

Decision Forests, Convolutional Networks and the Models in-Between

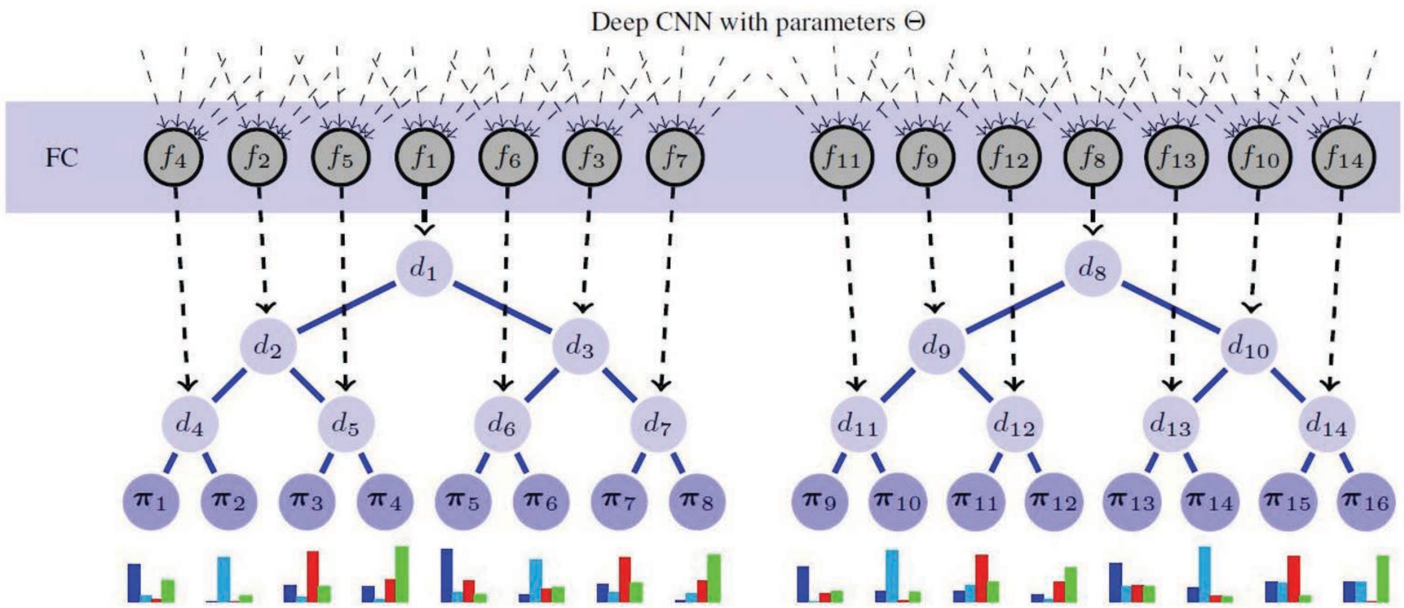
Microsoft Research Technical Report 2015-58

Yani Ioannou[1] Duncan Robertson[2] Darko Zikic[2] Peter Kontschieder[2]
Jamie Shotton[2] Matthew Brown[3] Antonio Criminisi[2]
[1]University of Cambridge, [2]Microsoft Research, [3]University of Bath

Deep Neural Decision Forests

Peter Kontschieder[1] Madalina Fiterau[*,2] Antonio Criminisi[1] Samuel Ro

Microsoft Research[1] Cambridge, UK   Carnegie Mellon University[2] Pittsburgh, PA   Fondazione Bruno Kessler[3] Trento, Italy

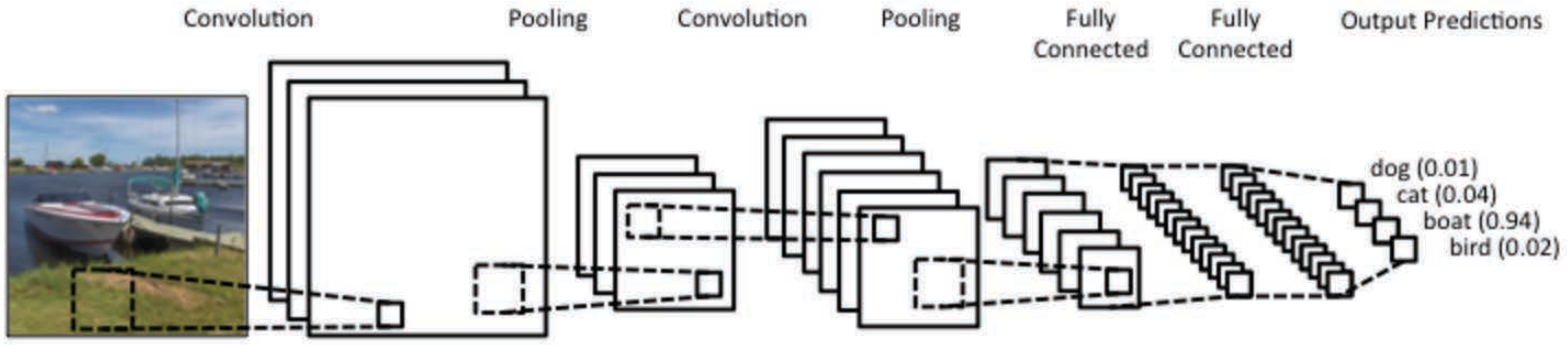


Interpretable DL

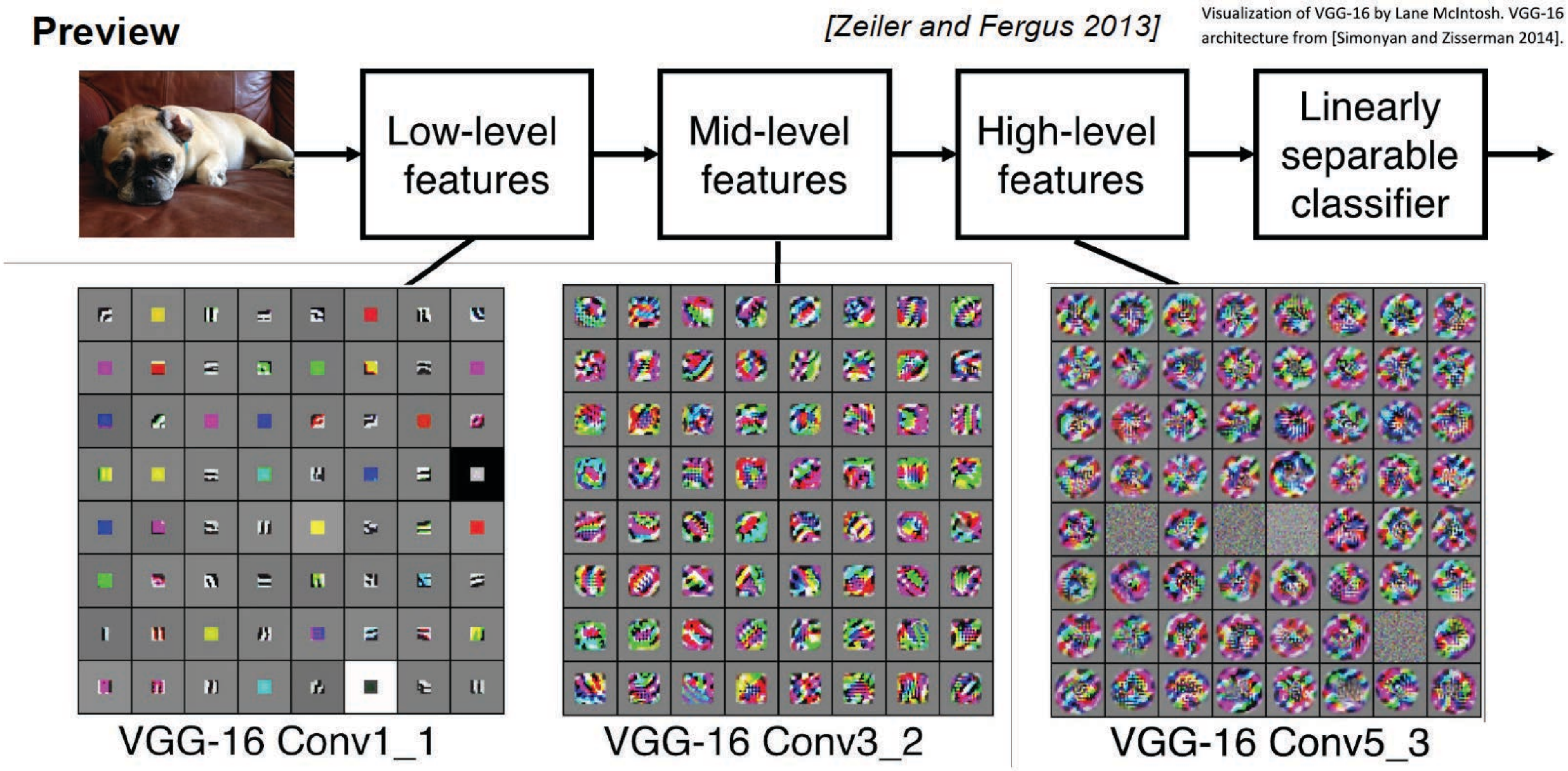
Convolutional Neural Networks

Typical structure:

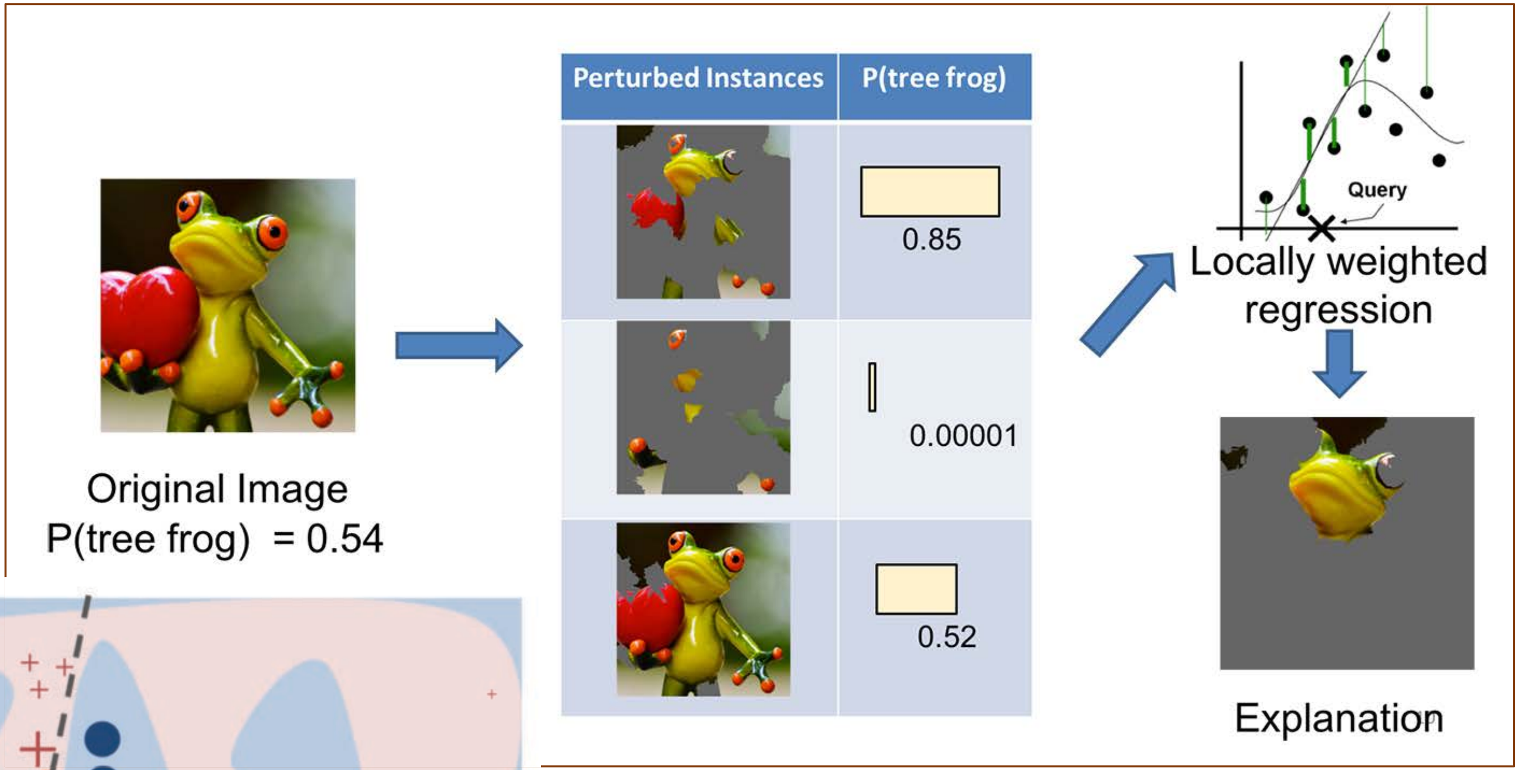
- Convolution layers
- Pooling or subsampling layers
- Fully connected layers

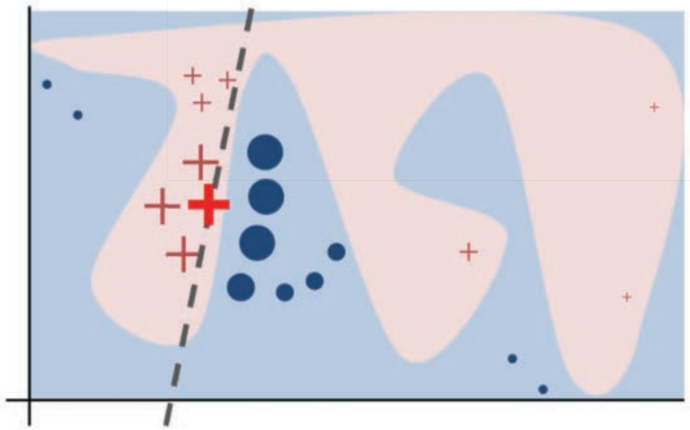


Convolutional Neural Networks



“Why Should I Trust You?” Explaining the Predictions of Any Classifier: Local Interpretable Model-agnostic Explanations (LIME) – Ribeiro et al . 2017

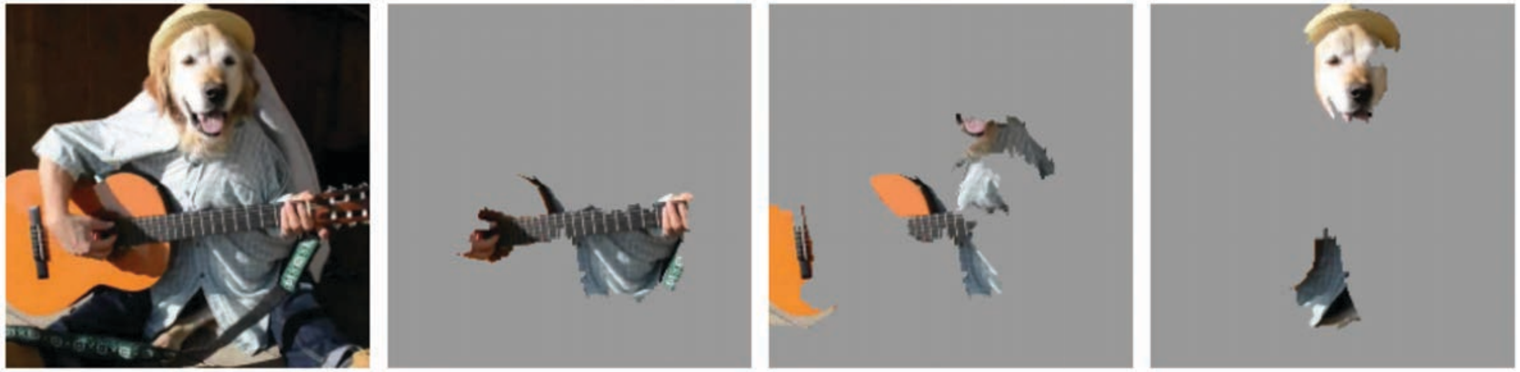




Local Interpretable Model-Agnostic Explanations (LIME)

Source: Ribeiro et al. 2017

"Why Should I Trust You?" Explaining the Predictions of Any Classifier: Local Interpretable Model-agnostic Explanations (LIME) – Ribeiro et al . 2017



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Source: Ribeiro et al. 2017

"Why Should I Trust You?" Explaining the Predictions of Any Classifier: Local Interpretable Model-agnostic Explanations (LIME) – Ribeiro et al . 2017

Local Interpretable Model-Agnostic Explanations (LIME)

Proximity measure

Model being interpreted

complexity measure

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

How well g ~ f

Potentially interpretable model

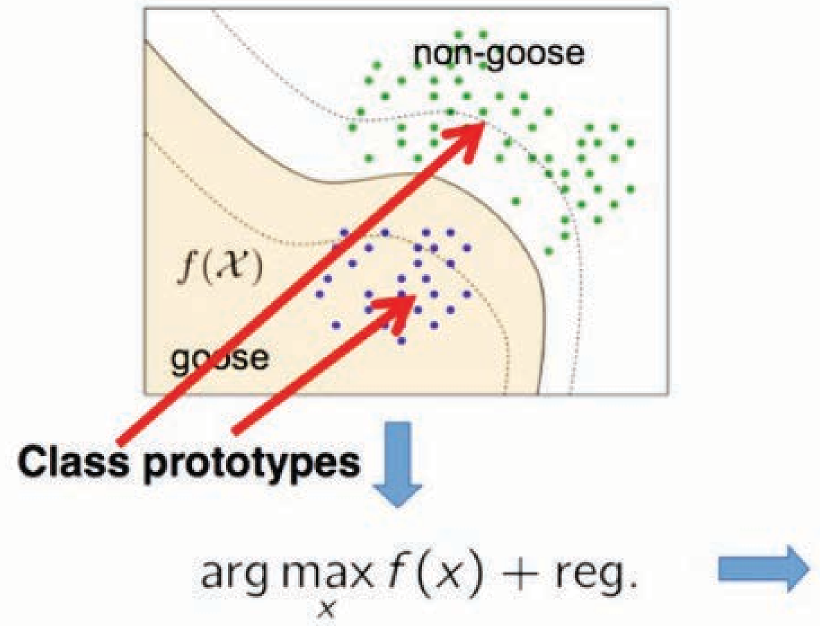
Algorithm 1 Sparse Linear Explanations using LIME

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$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

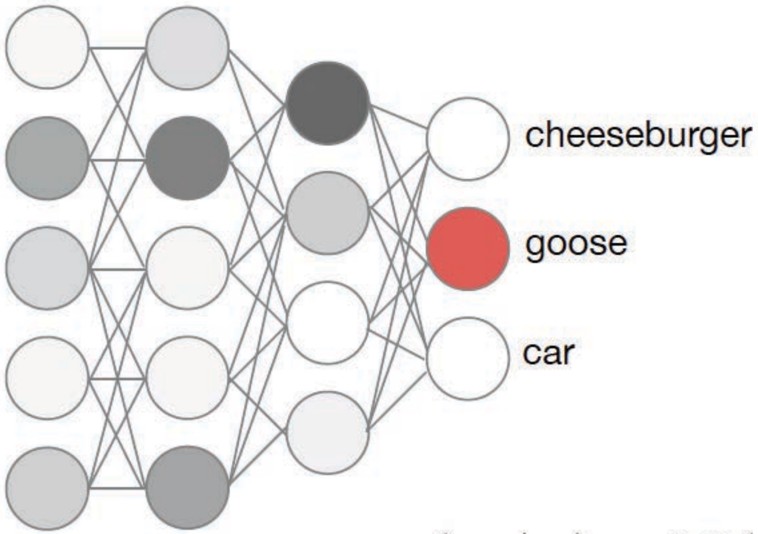
Source: Ribeiro et al. 2017

Explaining by finding prototype class members

- Global explanation technique
- Find pattern maximizing class activation
- Not-necessarily a real image

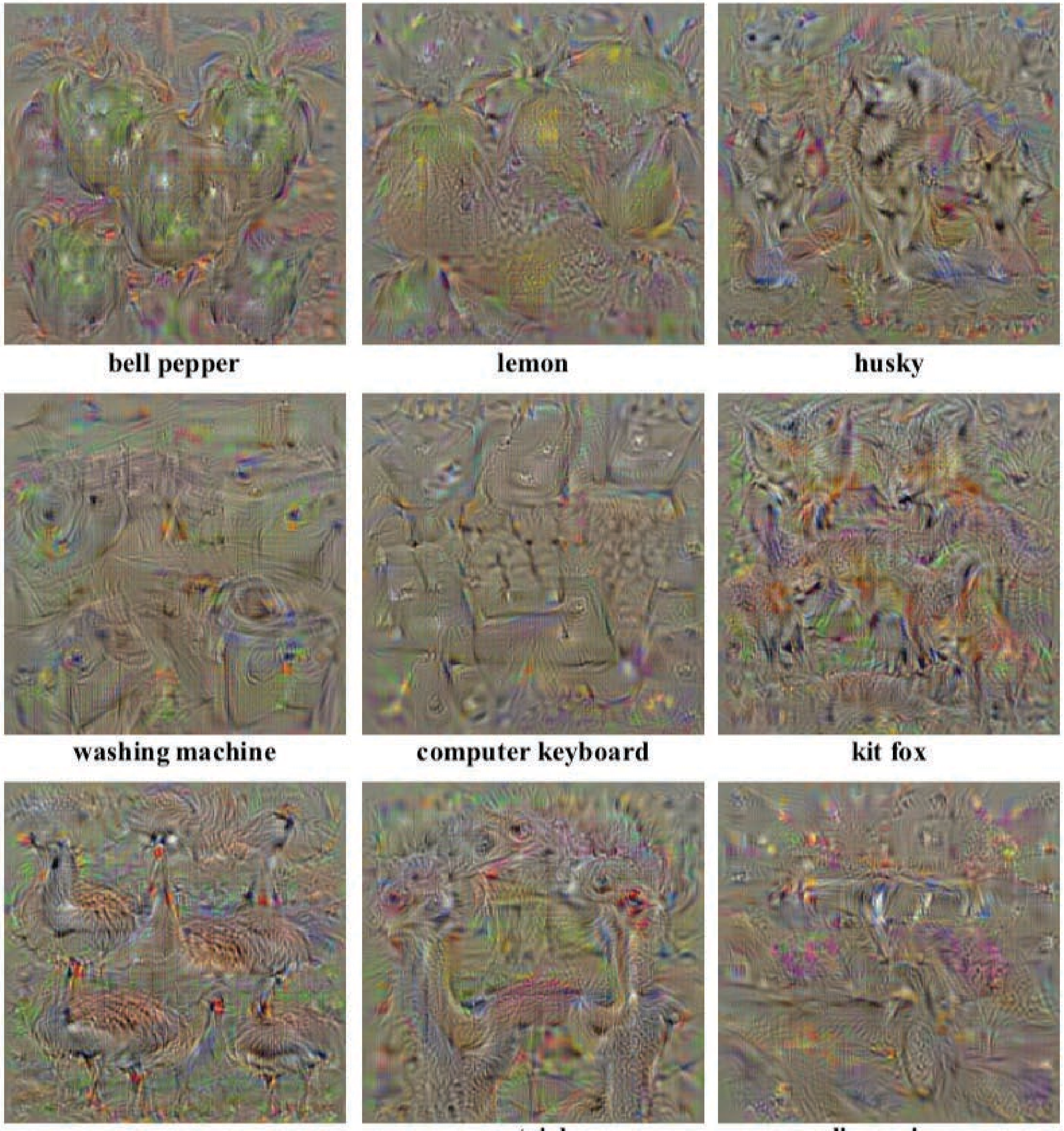


Simonyan et al. 2013



$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \,|\, x) + \lambda \Omega(x)$$

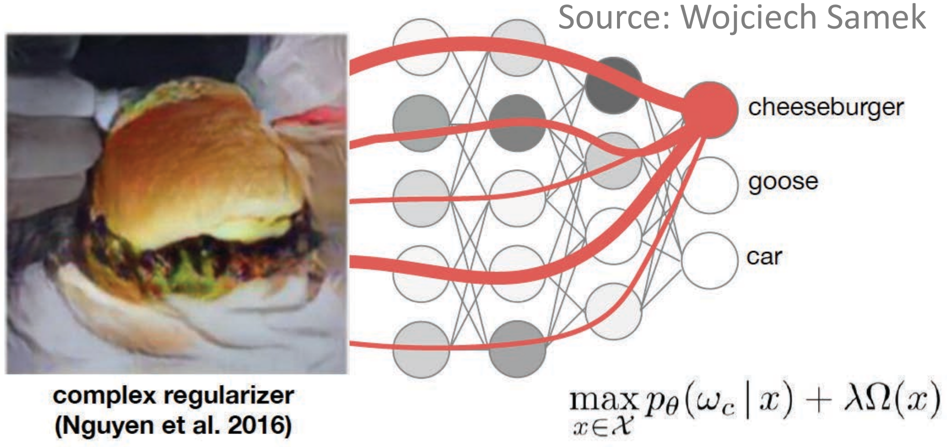
Simonyan et al. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps



Simonyan et al. 2013. Deep Inside Convolutional Networks:
Visualising Image Classification Models and Saliency Maps

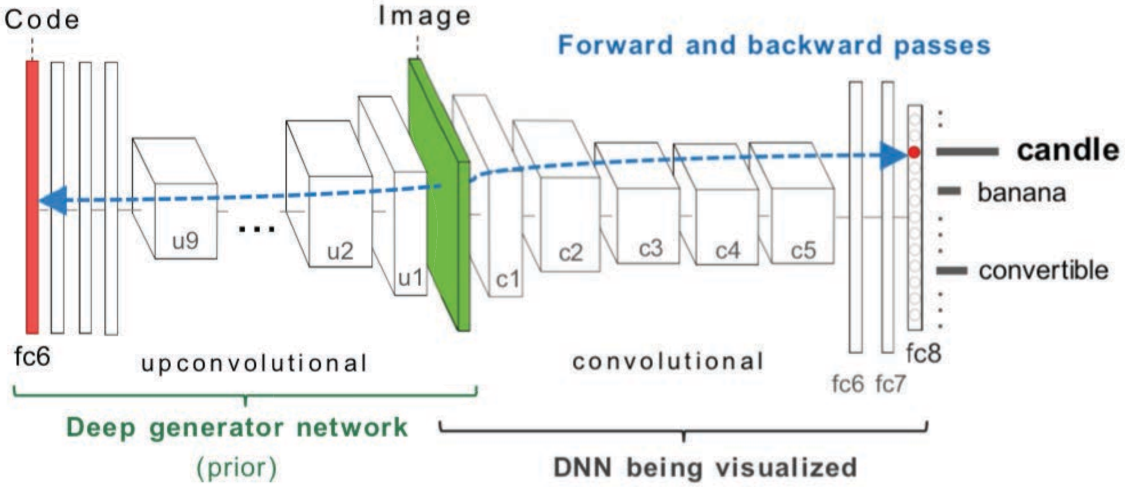
Synthesizing the preferred inputs for neurons in neural networks via deep generator networks
- Nguyen et al. 2016

Source: Wojciech Samek



Regularization via generative adversarials

1.- Forwards pass of gradients
2- Detect maximum activation
3.- Backwark pass of “FC6” into encoder



Feature vector

classifier

generator

$$\widehat{\mathbf{y}^l} = \arg\max_{\mathbf{y}^l}(\Phi_h(G_l(\mathbf{y}^l)) - \lambda\|\mathbf{y}^l\|)$$

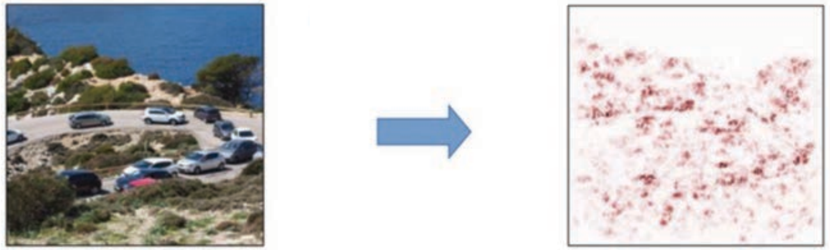
- Prior generator and DNN can be trained separately

Synthesizing the preferred inputs for neurons in neural networks via deep generator networks
- Nguyen et al. 2016

Gradient-based methods

- Idea: magnitude of gradient reflects importance (attribution) of input to output scores

$$R = \sum \left(\frac{\delta S}{\delta x_i}\right)^2$$



Variations:

- Input * Gradient $x * \frac{\delta f}{\delta \mathrm{x}_i}$ → Addresses gradient saturation and reduces diffusivity – Shrikumar et al. 2016
- Deconvnet: inverts direction of applying activation; zero outs negative activations– Zeiler et al. 2014
- Guided-backpropagation: deconvnet + filters out negative forward activations - Springenberg et al. 2015
- Integrated gradients: sums over scaled versions of input; addresses gradient saturation – Sundararajan et al. 2017
- Grad-CAM: includes a neuron importance weight (detailed later) - Selvaraju et al. 2017

Image source: Samek et al. 2018

Differences between backprop, deconvnet and guided-backpropagation

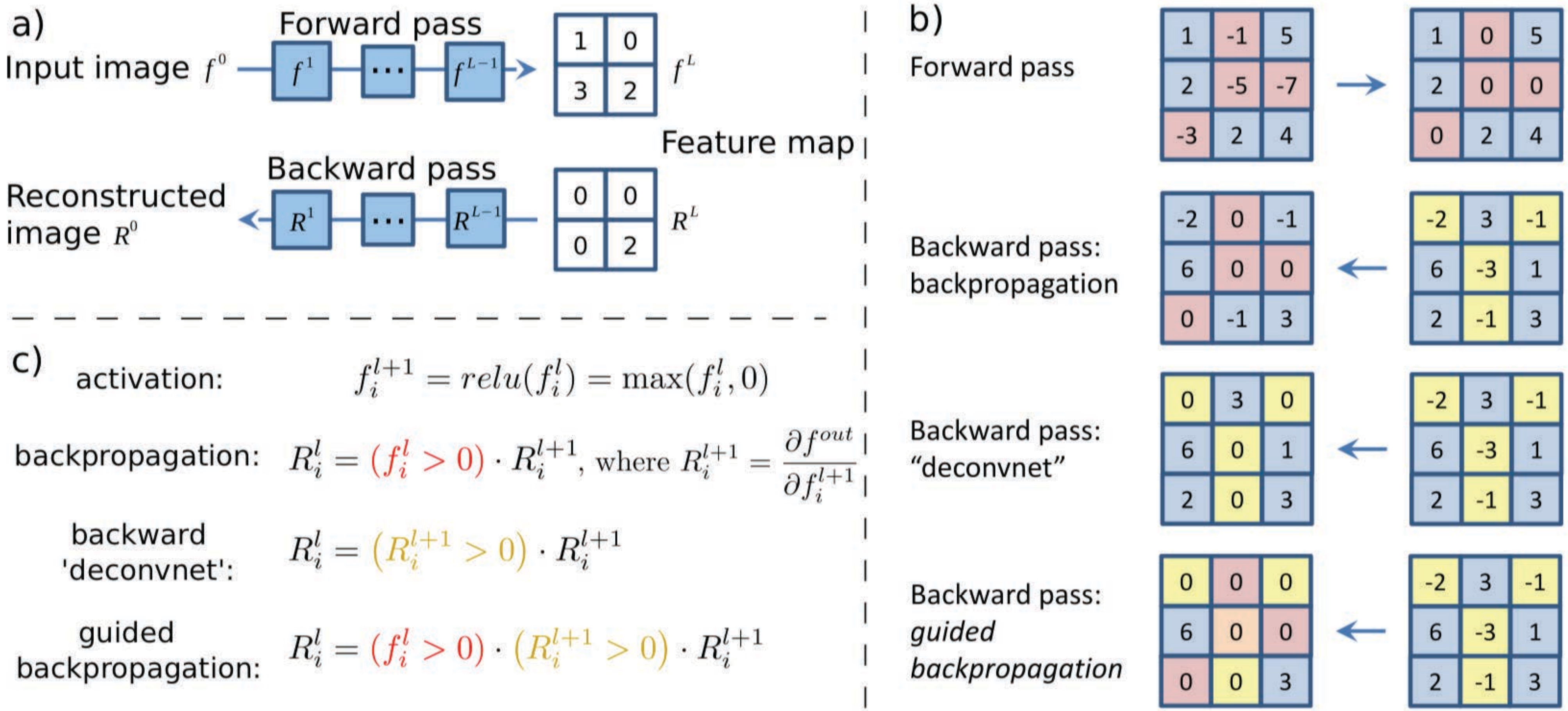
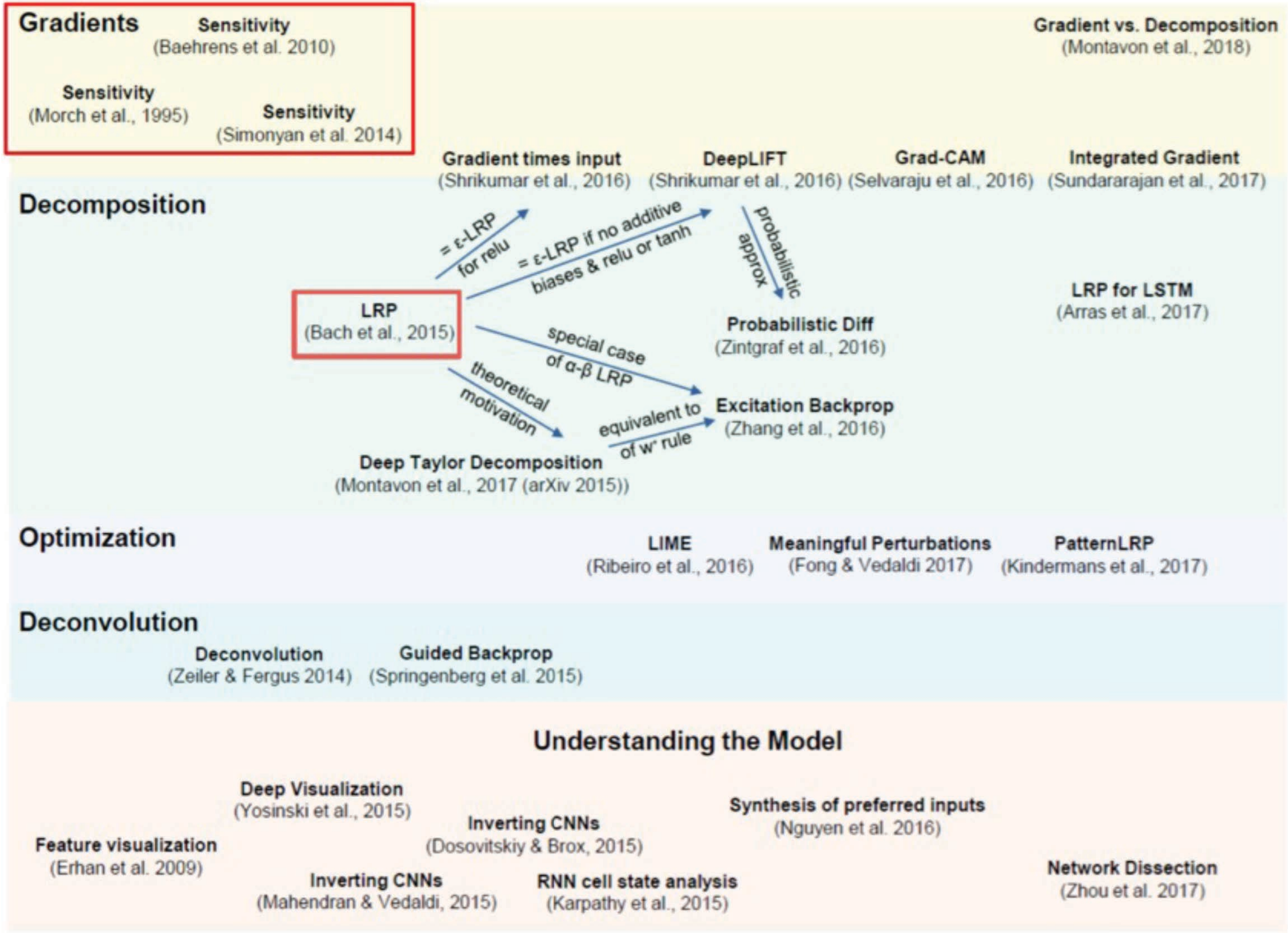


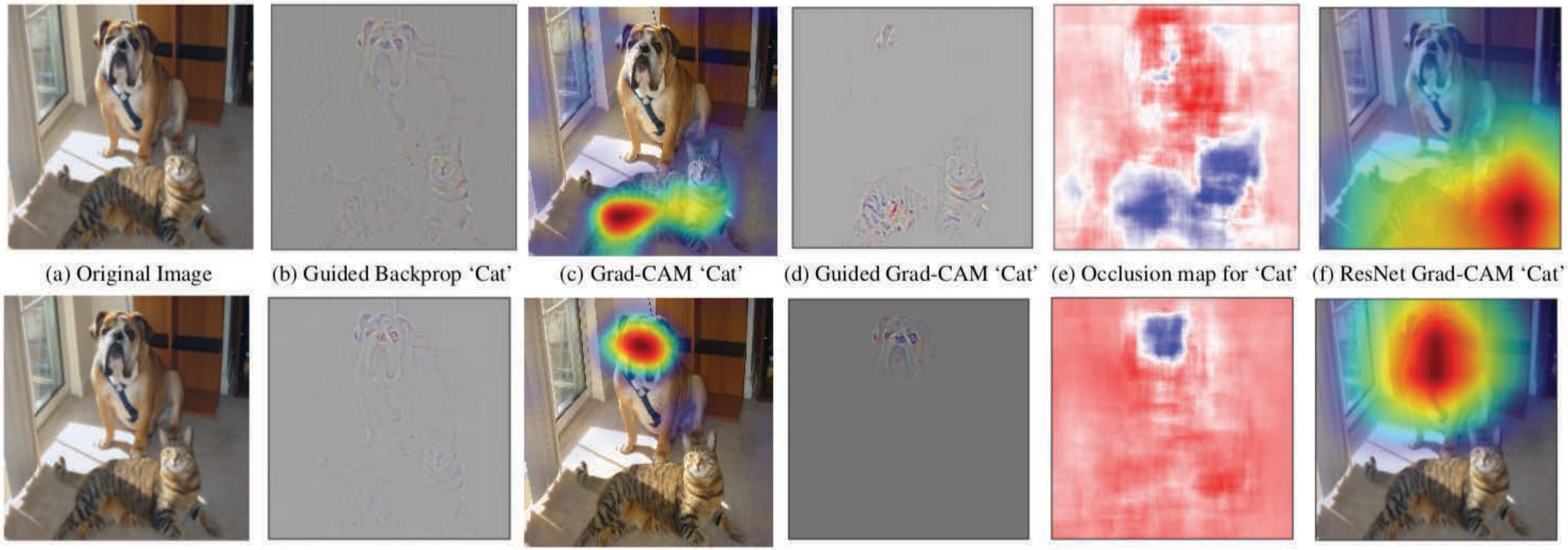
Image source: Springenberg et al. ICLR 2015



Gradients
Sensitivity
(Baehrens et al. 2010)
Sensitivity
(Morch et al., 1995)
Sensitivity
(Simonyan et al. 2014)
Gradient vs. Decomposition
(Montavon et al., 2018)
Gradient times input
(Shrikumar et al., 2016)
DeepLIFT
(Shrikumar et al., 2016)
Grad-CAM
(Selvaraju et al., 2016)
Integrated Gradient
(Sundararajan et al., 2017)
Decomposition
= ε-LRP for relu
= ε-LRP if no additive biases & relu or tanh
probabilistic approx
LRP
(Bach et al., 2015)
LRP for LSTM
(Arras et al., 2017)
Probabilistic Diff
(Zintgraf et al., 2016)
special case of α-β LRP
theoretical motivation
equivalent to of w⁺ rule
Excitation Backprop
(Zhang et al., 2016)
Deep Taylor Decomposition
(Montavon et al., 2017 (arXiv 2015))
Optimization
LIME
(Ribeiro et al., 2016)
Meaningful Perturbations
(Fong & Vedaldi 2017)
PatternLRP
(Kindermans et al., 2017)
Deconvolution
Deconvolution
(Zeiler & Fergus 2014)
Guided Backprop
(Springenberg et al. 2015)
Understanding the Model
Deep Visualization
(Yosinski et al., 2015)
Synthesis of preferred inputs
(Nguyen et al. 2016)
Inverting CNNs
(Dosovitskiy & Brox, 2015)
Feature visualization
(Erhan et al. 2009)
Network Dissection
(Zhou et al. 2017)
Inverting CNNs
(Mahendran & Vedaldi, 2015)
RNN cell state analysis
(Karpathy et al., 2015)

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

- Generalization of CAM: Zhou et al. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.
    - Applicable to any CNN-based (CAM requires conv feature maps → global average pooling → softmax layer)
- Motivation – Good visualization is class-specific <u>and</u> detailed

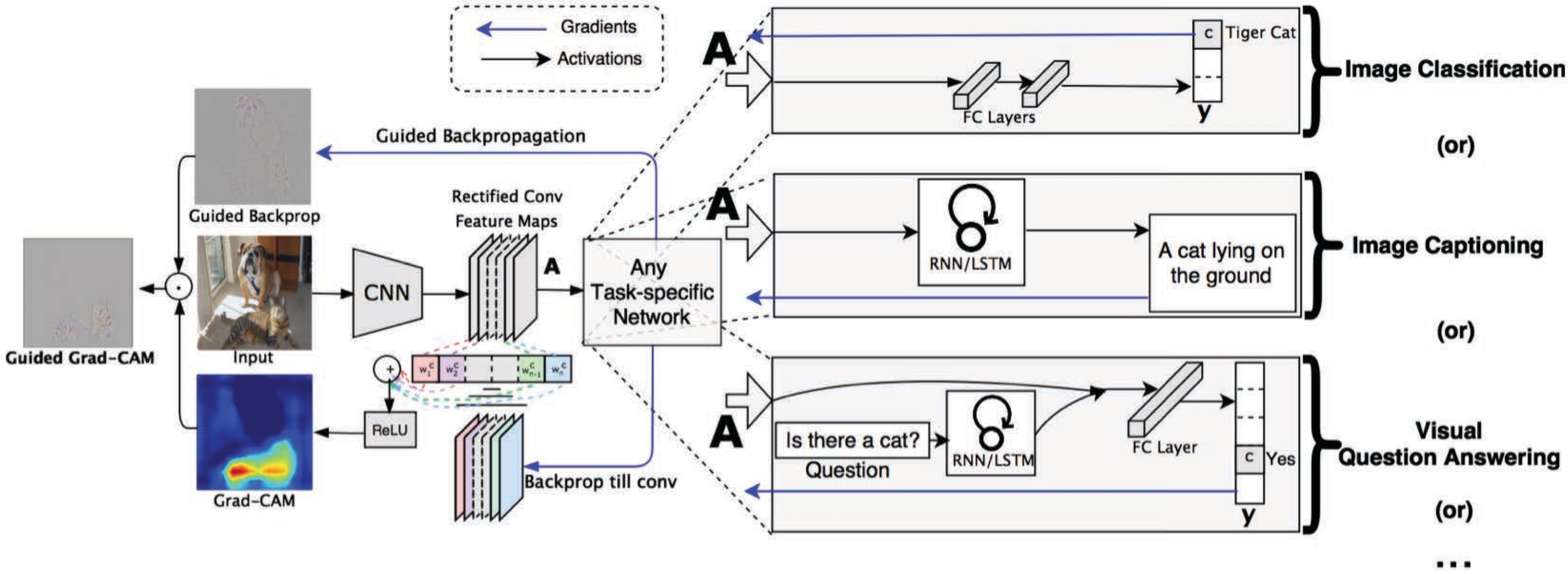


(a) Original Image (b) Guided Backprop ‘Cat’ (c) Grad-CAM ‘Cat’ (d) Guided Grad-CAM ‘Cat’ (e) Occlusion map for ‘Cat’ (f) ResNet Grad-CAM ‘Cat’

(g) Original Image (h) Guided Backprop ‘Dog’ (i) Grad-CAM ‘Dog’ (j) Guided Grad-CAM ‘Dog’ (k) Occlusion map for ‘Dog’ (l) ResNet Grad-CAM ‘Dog’

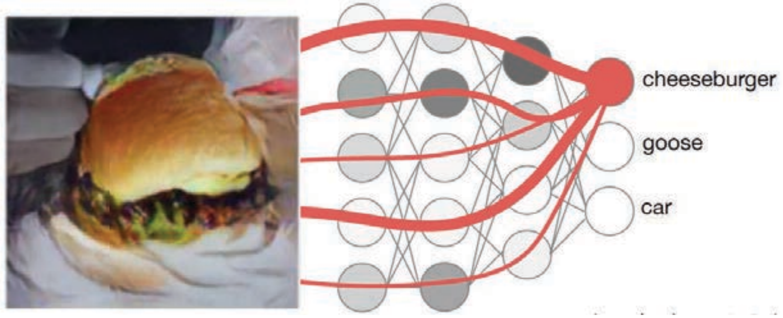
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Guided-Grad-CAM: Uses Grad-CAM ask mask for Guided Backprop



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Single forward and (partial) backward pass



Source: Wojciech Samek

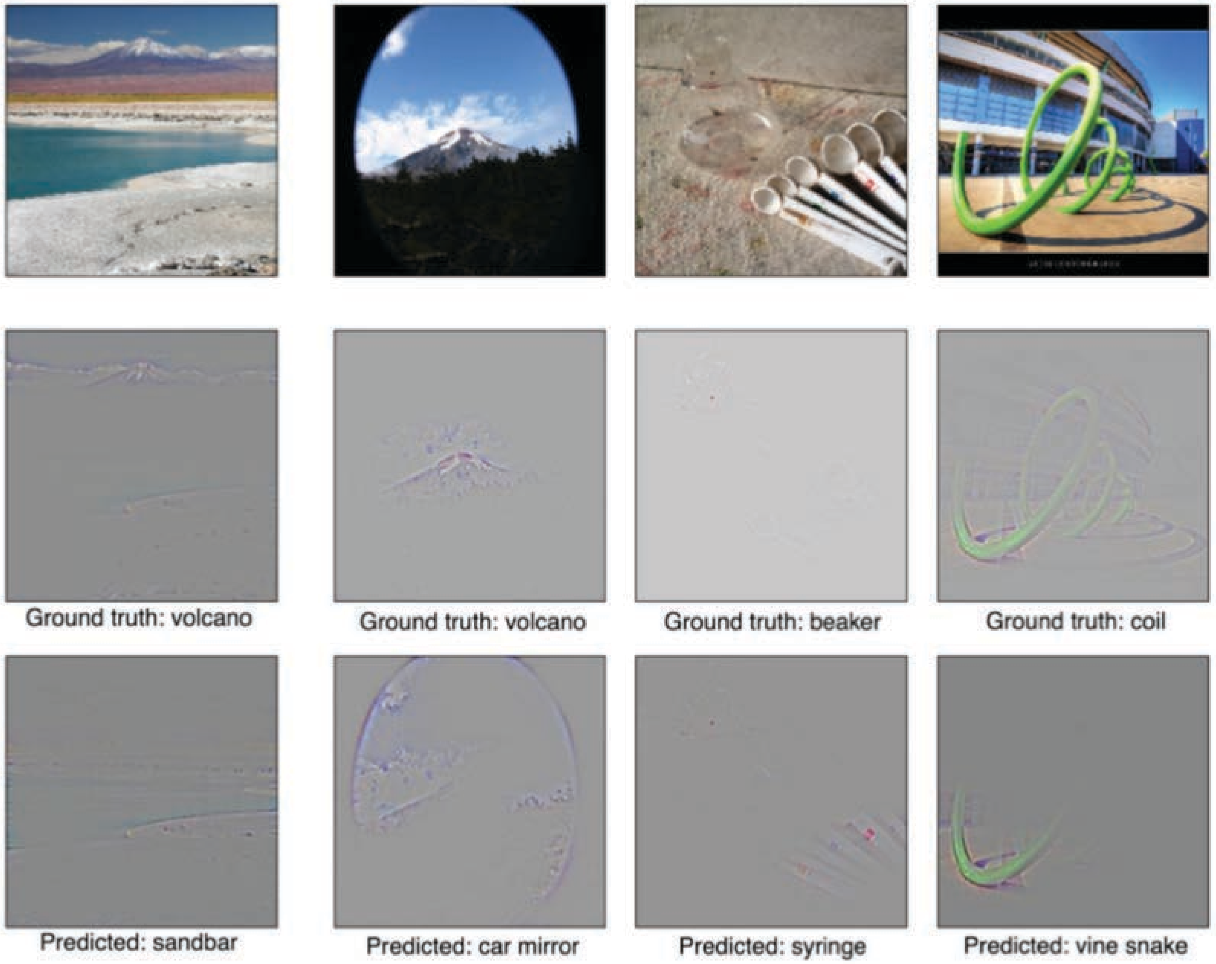
Neuron importance weights
A^k: kth feature map
y^c: score for class c (before softmax)

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k \right)}_{\text{linear combination}}$$

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Failure analysis: Cases where VGG failed and corresponding visualizations



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Human studies

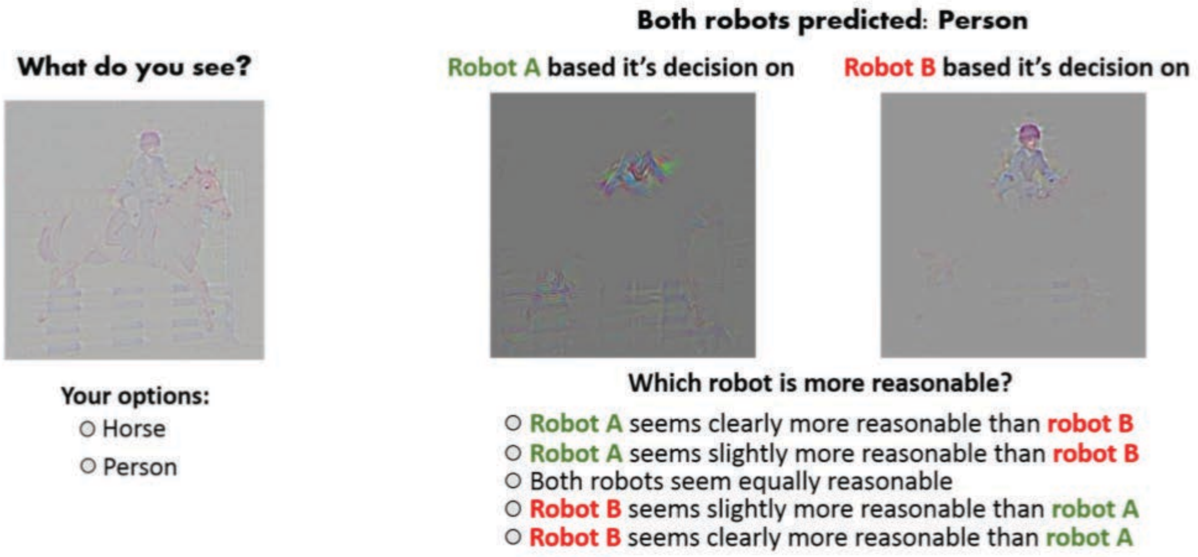
"Which of the two object categories is depicted in the image?"

- N=43, amazon mechanical turk
- Humans recognized correct class in 61.23% vs 44.44% guided backprop

"Which model is more trustable?"

VGG vs. AlexNet

"Remarkably, we find that human subjects are able to identify the more accurate classifier (VGG over AlexNet) despite viewing identical predictions from the two, simply from the different explanations generated from the two."



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Robustness to adversarial examples: Examples where VGG got "fooled"

Boxer: 0.40 Tiger Cat: 0.18
(a) Original image

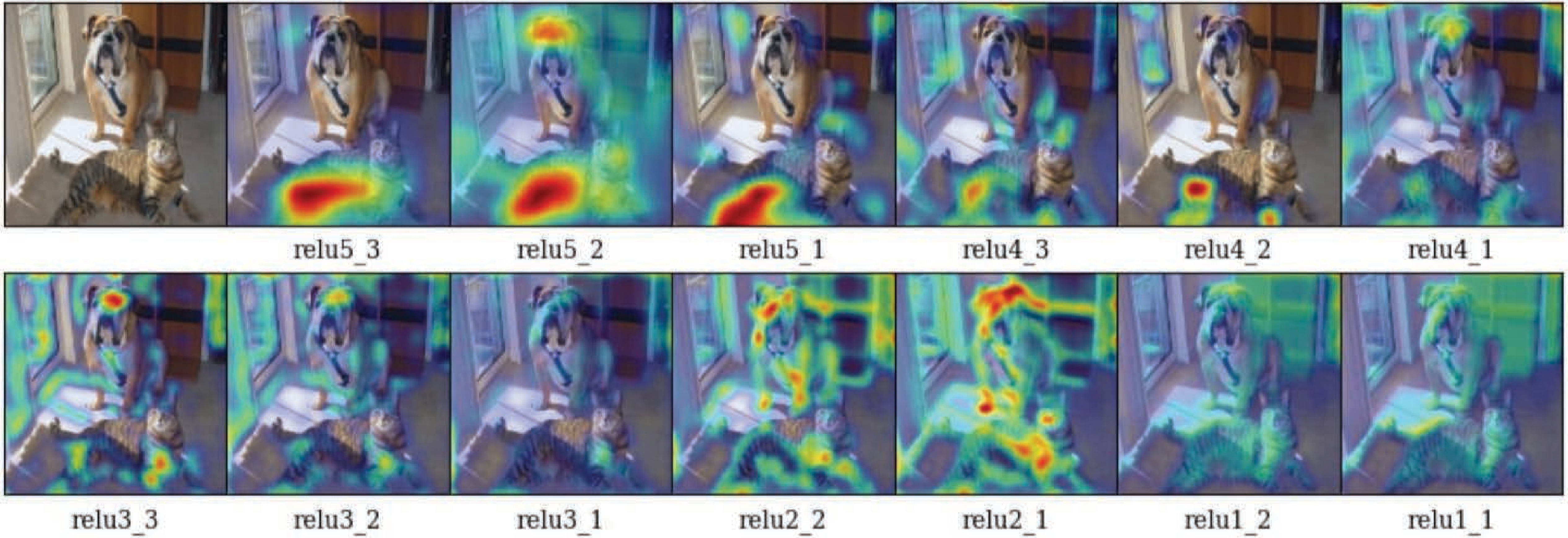
Airliner: 0.9999
(b) Adversarial image

Boxer: 1.1e-20
(c) Grad-CAM "Dog"

Tiger Cat: 6.5e-17
(d) Grad-CAM "Cat"

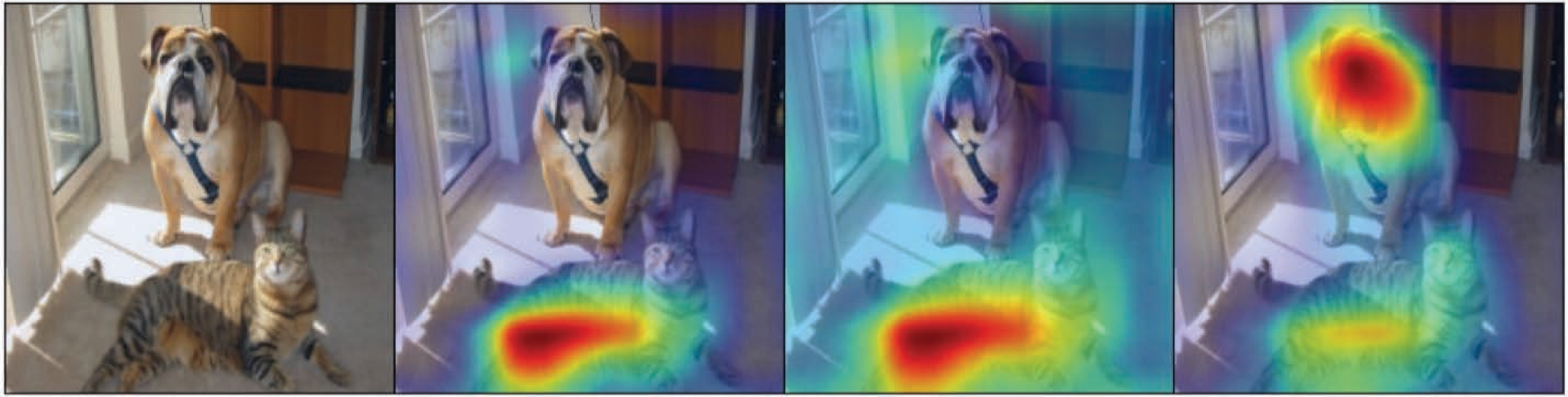
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Layer-wise analysis



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Ablation study



Grad-CAM | Grad-CAM without ReLU | Grad-CAM with absolute gradients

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization - Selvaraju et al. 2017

Effect of global average pooling vs. global max pooling

Grad-CAM

Grad-CAM with Global Max Pooled gradients

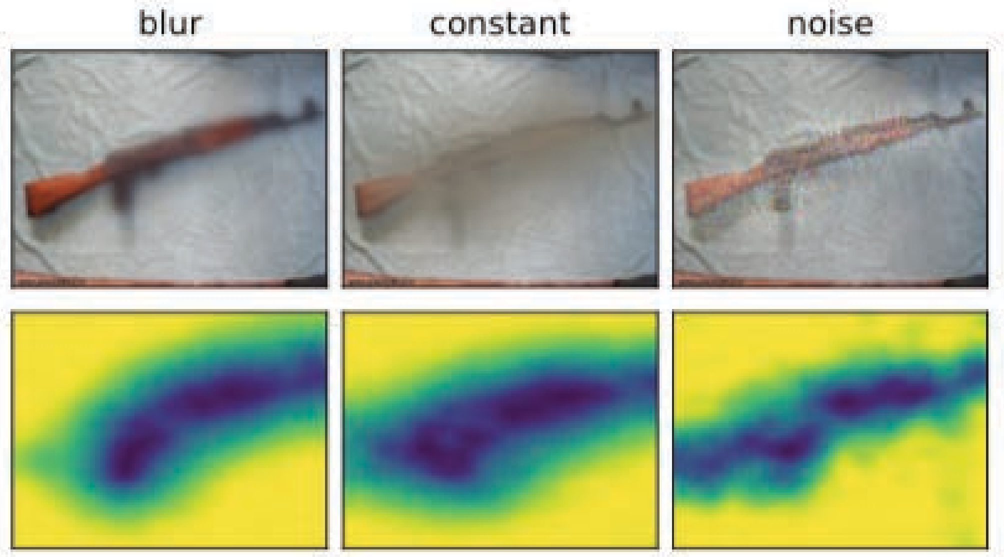
Interpretable Explanations of Black Boxes by Meaningful Perturbation - Fong et al. 2018

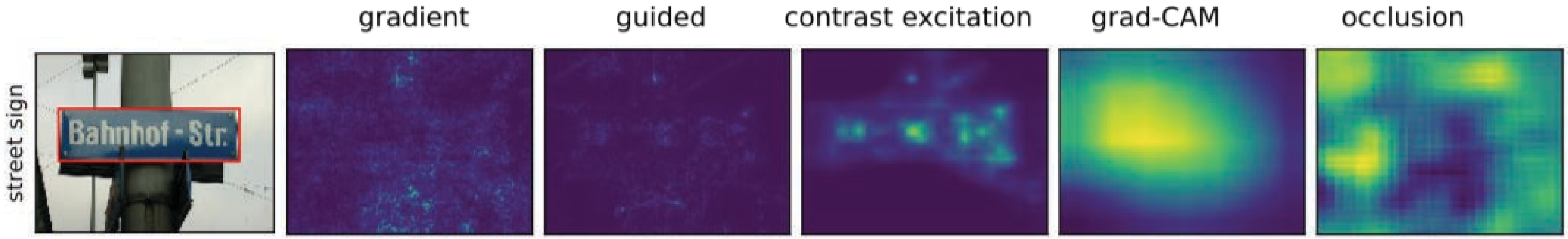
Motivation: Saliency (gradient-based) approaches are not specific enough

Key idea: select the right perturbation* to study the effect on the prediction function f(x)

Meaning of an explanation depends on the meaning of the changes applied to the input x

*Perturbation ←→ type of information deletion





Interpretable Explanations of Black Boxes by Meaningful Perturbation - Fong et al. 2018

Optimal ”information deletion” mask

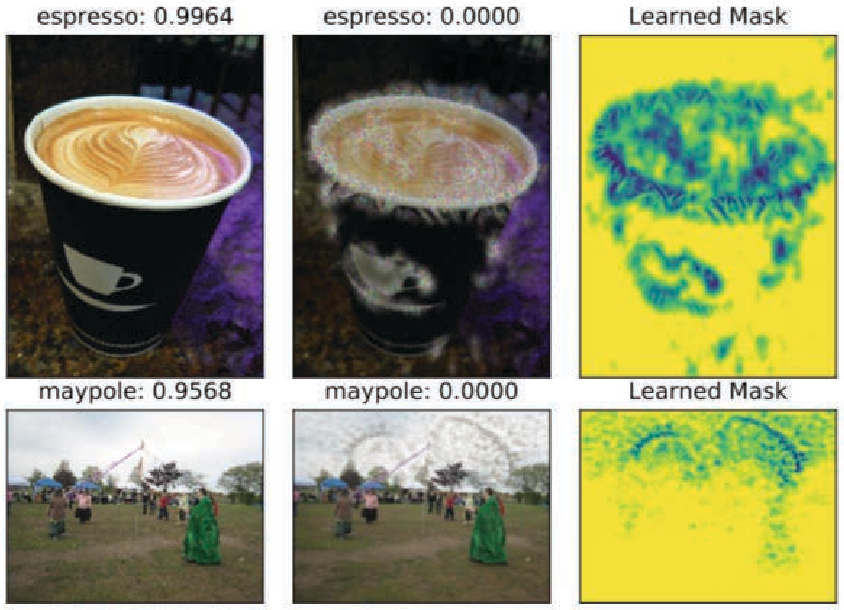
Enforces minimal mask size

perturbator

Input image

$$m^* = \operatorname*{argmin}_{m\in[0,1]^\Lambda} \lambda\|\mathbf{1} - m\|_1 + f_c(\Phi(x_0; m))$$

Predictor for class c being interpreted



Can lead to unnatural perturbations → against explanations

Interpretable Explanations of Black Boxes by Meaningful Perturbation - Fong et al. 2018

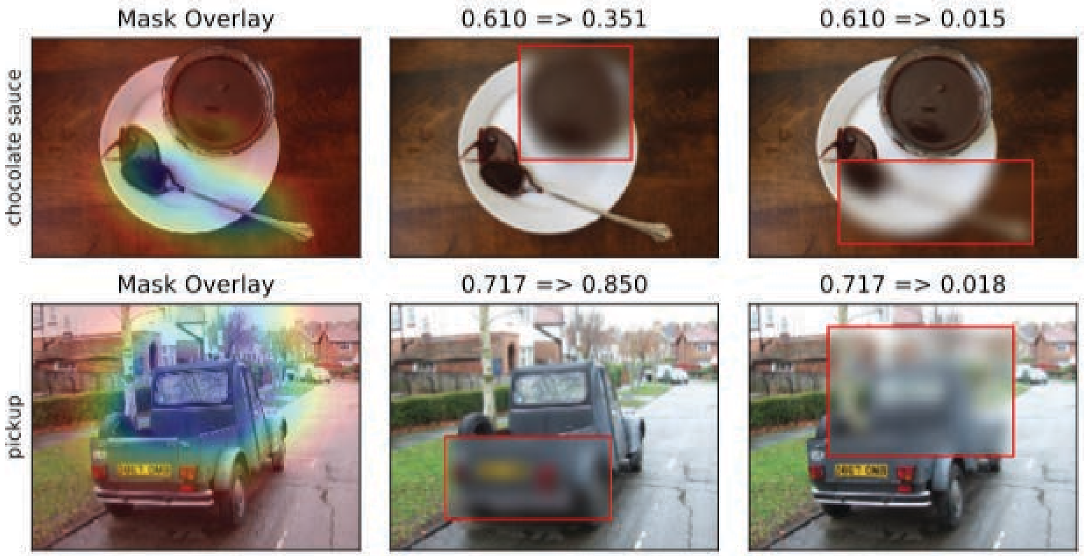
Enforces minimal mask size

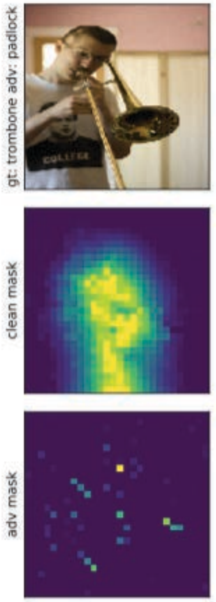
Enforces simplicity of the mask

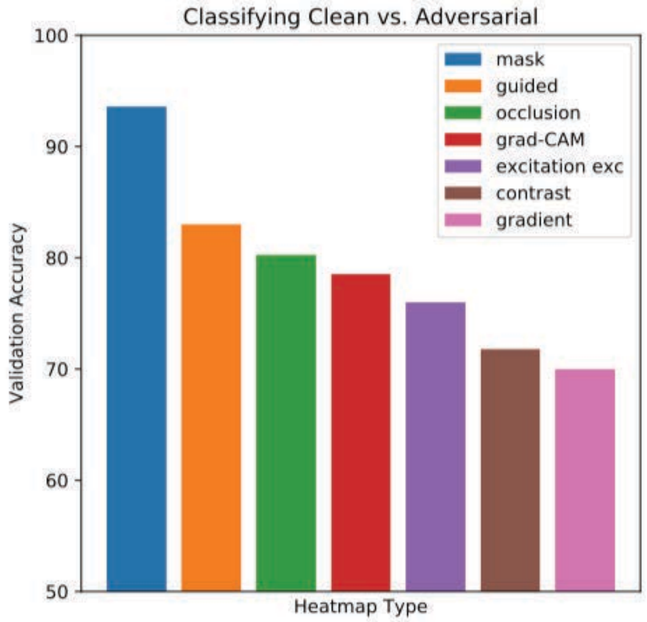
$$\min_{m\in[0,1]^\Lambda} \lambda_1\|\mathbf{1}-m\|_1 + \lambda_2 \sum_{u\in\Lambda}\|\nabla m(u)\|_\beta^\beta$$
$$+ \mathbb{E}_\tau[f_c(\Phi(x_0(\cdot-\tau),m))]$$

Expectation

Add variability to the mask (via jitter)

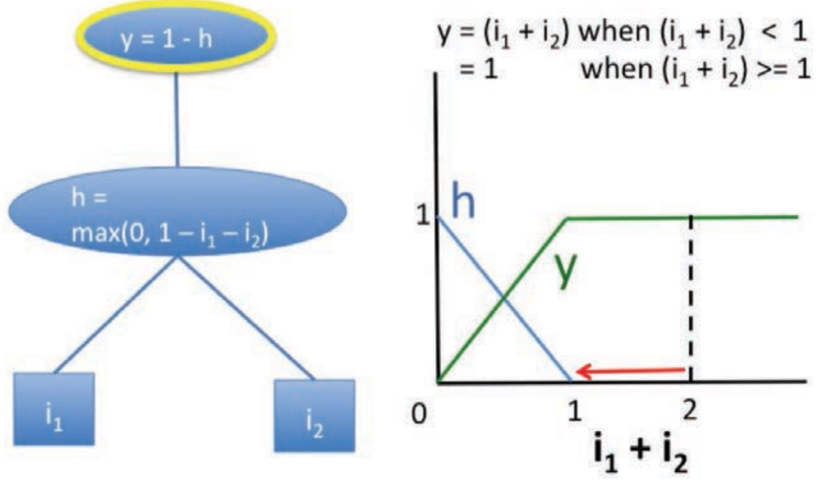


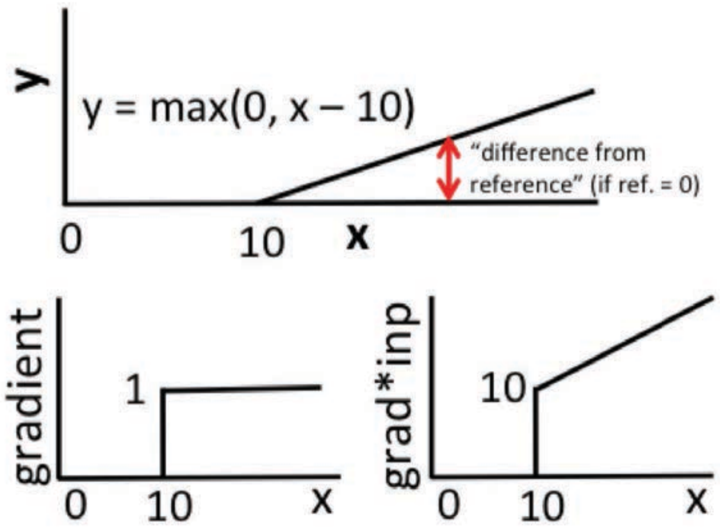




DeepLift - Learning Important Features Through Propagating Activation Differences – Shrikumar et al. 2017

- Also part of the “gradient backpropagation” family of methods
- Motivation: zero gradients not necessarily mean no attribution; consider positive and negative attributions
- Idea: measure importance by assessing difference to a “reference”





DeepLift - Learning Important Features Through Propagating Activation Differences – Shrikumar et al. 2017

- "reference"? → default or neutral state ?(e.g. black image)

Difference between neuron output and its reference value

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t$$

Contribution measure: difference to delta_t attributed to differences in input x

Positive and negative contributions

$$\Delta x_i = \Delta x_i^+ + \Delta x_i^-$$

$$C_{\Delta x_i \Delta t} = C_{\Delta x_i^+ \Delta t} + C_{\Delta x_i^- \Delta t}$$

Dense and conv. layers

$$C_{\Delta x_i^+ \Delta y^+} = 1\{w_i \Delta x_i > 0\} w_i \Delta x_i^+$$

$$C_{\Delta x_i^- \Delta y^+} = 1\{w_i \Delta x_i > 0\} w_i \Delta x_i^-$$

$$C_{\Delta x_i^+ \Delta y^-} = 1\{w_i \Delta x_i < 0\} w_i \Delta x_i^+$$

$$C_{\Delta x_i^- \Delta y^-} = 1\{w_i \Delta x_i < 0\} w_i \Delta x_i^-$$

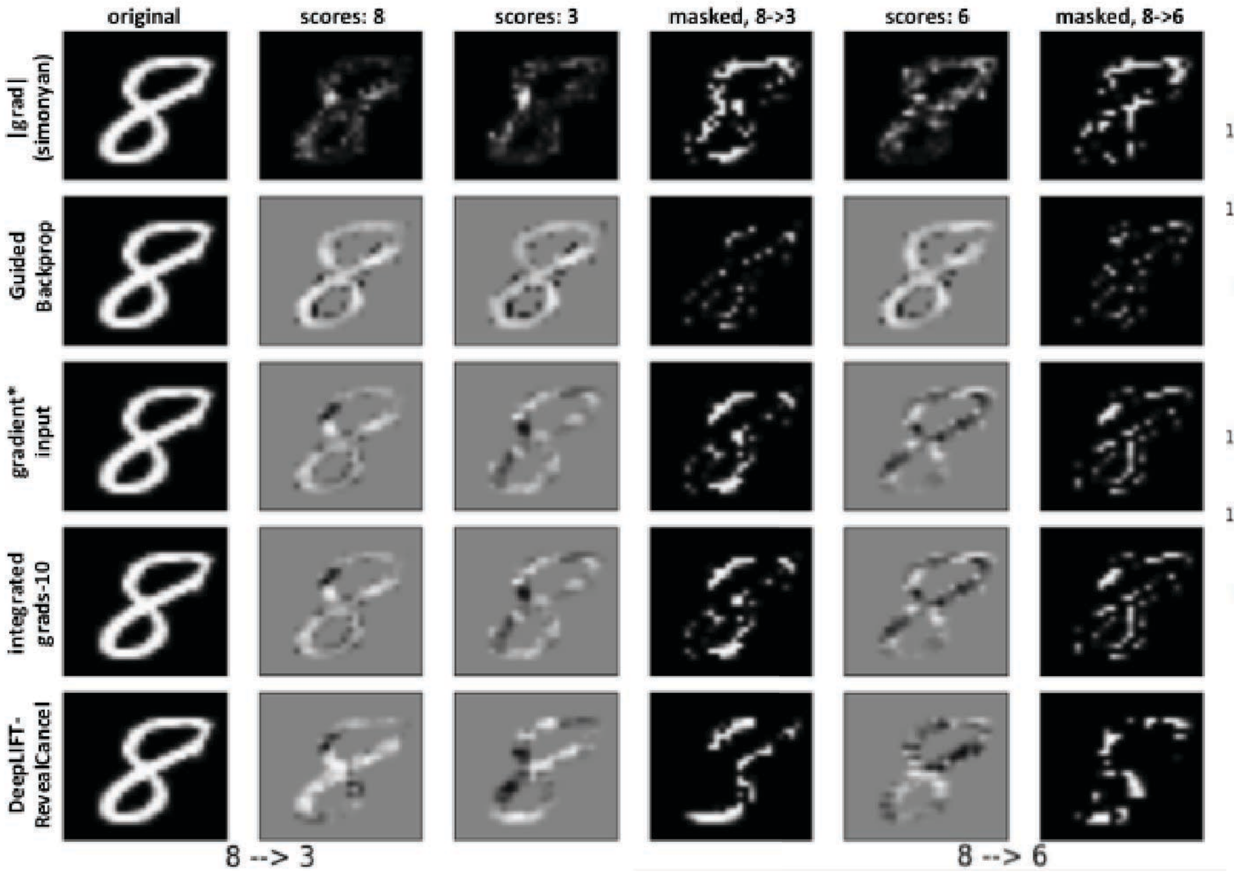
Non-linear (e.g.ReLU)

$$C_{\Delta x \Delta y} = \Delta y$$

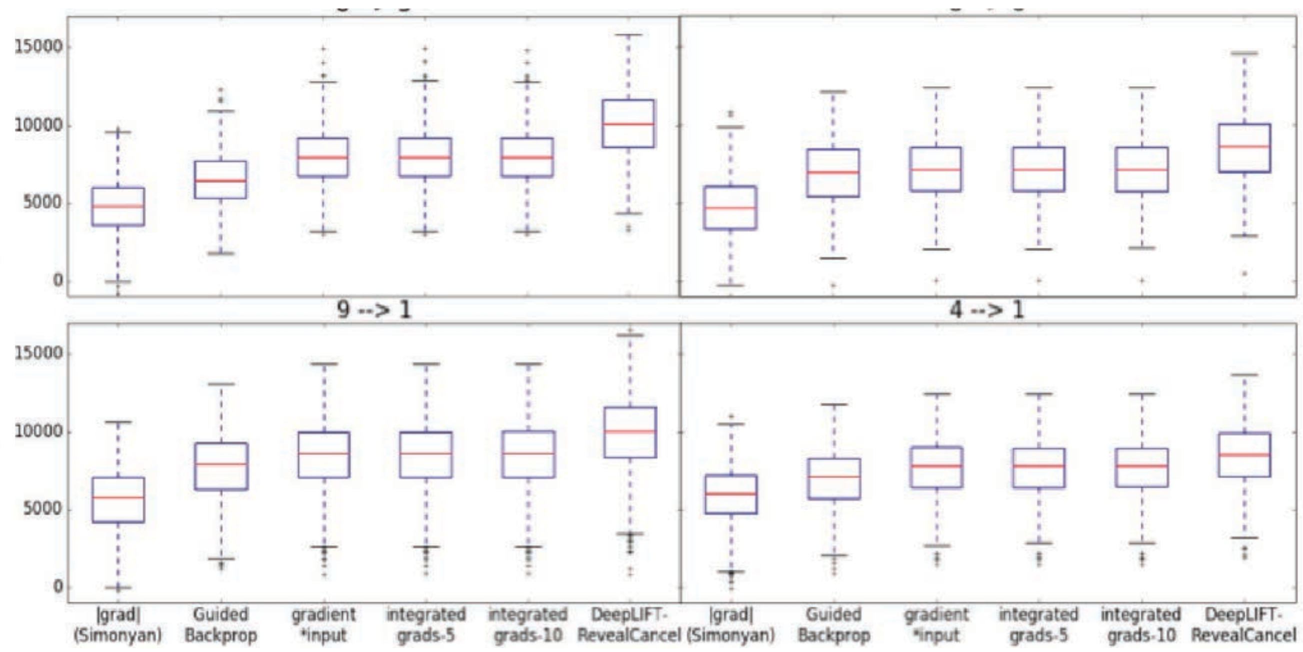
Approximates gradient

DeepLift - Learning Important Features Through Propagating Activation Differences – Shrikumar et al. 2017

Evaluation on MNIST: maximizing transition between digits



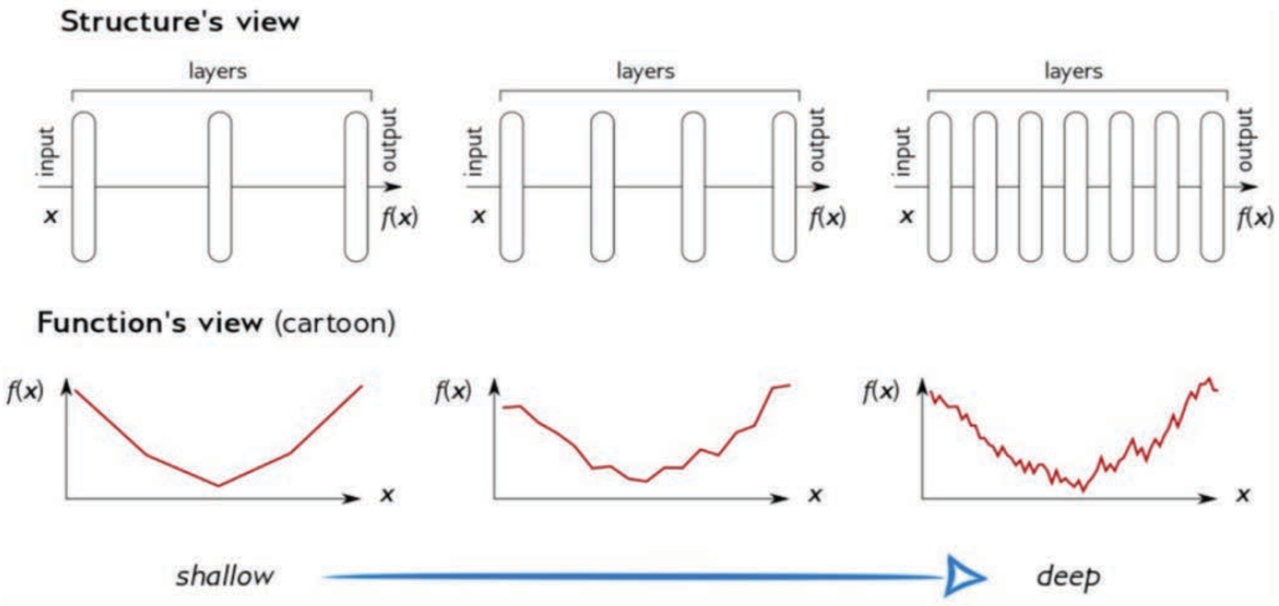
log-odds scores of target vs. original class after the mask is applied



Layer-wise Relevance Propagation (LRP) -

Bach et al., PLOS ONE, 2015

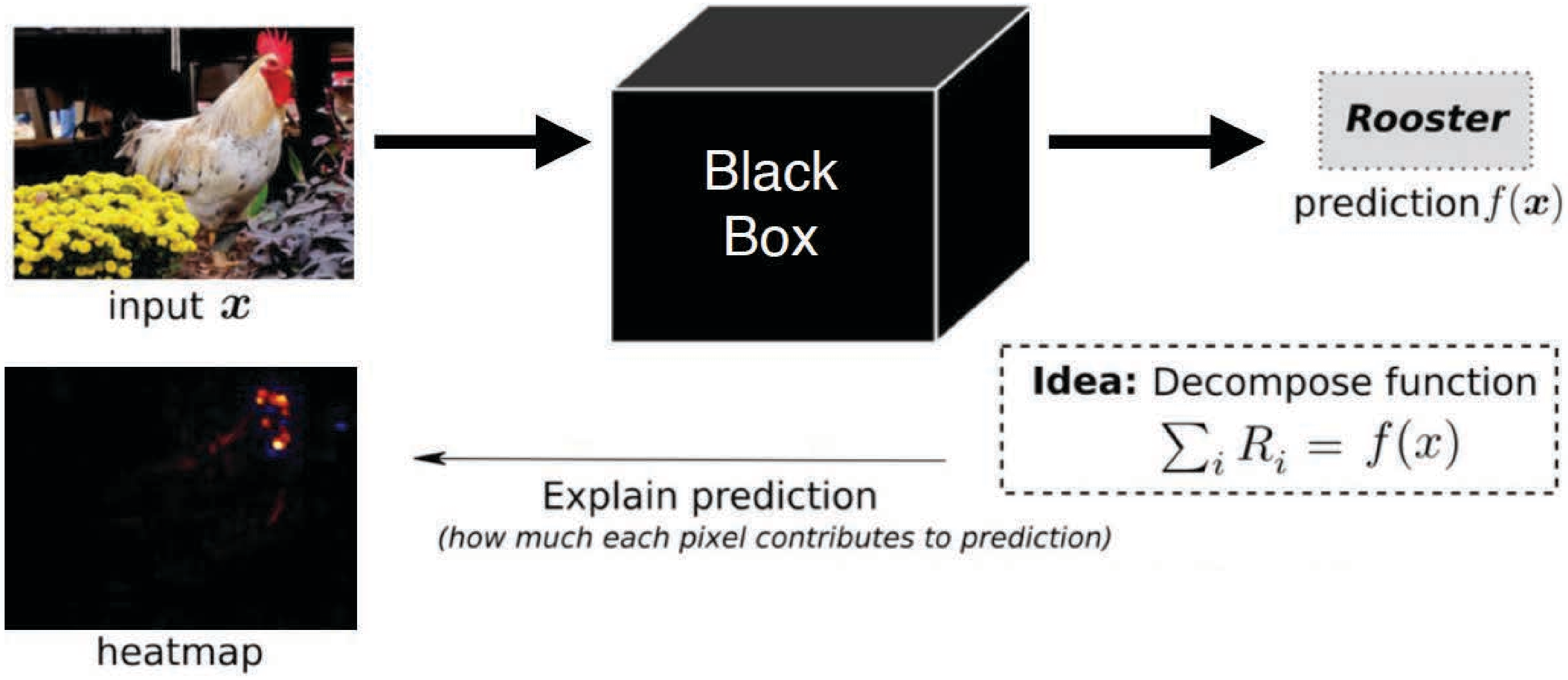
- Motivation: Gradient methods suffering from shattered gradient problem → depth affects reliability of computed input gradients
- Sensitivity explains changes to the prediction function, not the function's value itself.



Source: Wojciech Samek

Layer-wise Relevance Propagation (LRP) -
Bach et al., PLOS ONE, 2015

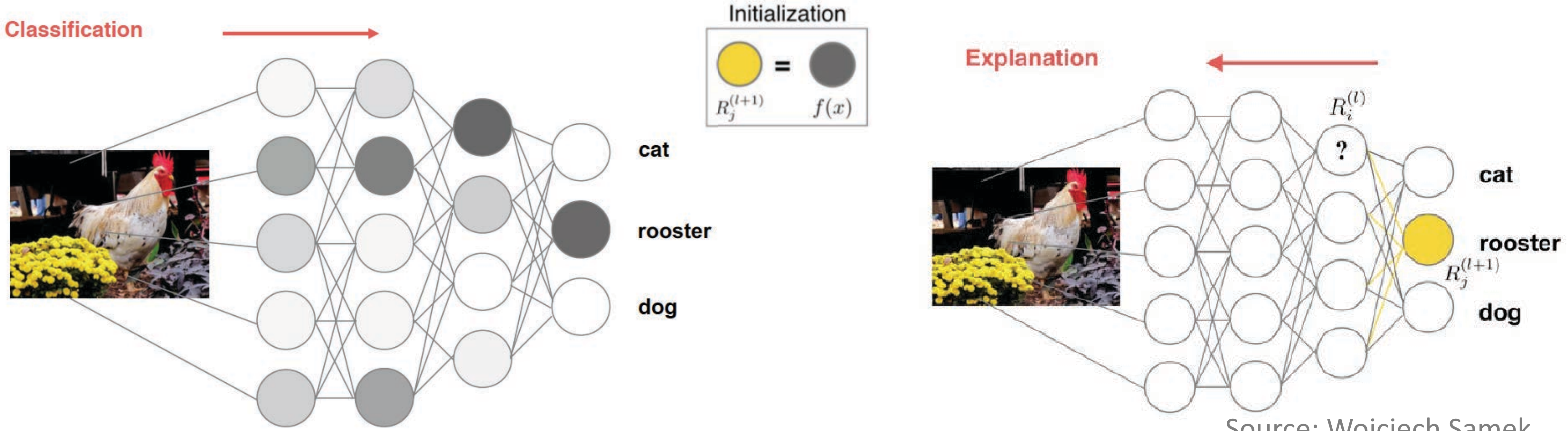
Idea: Decompose prediction function f(x) as a sum of layer-wise relevance values



Source: Wojciech Samek

Layer-wise Relevance Propagation (LRP) -
Bach et al., PLOS ONE, 2015

Idea: Decompose prediction function f(x) as a sum of layer-wise relevance values



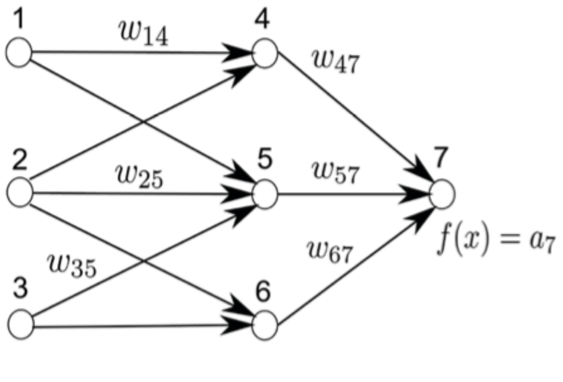
Source: Wojciech Samek

Relevance scores are layer-wise

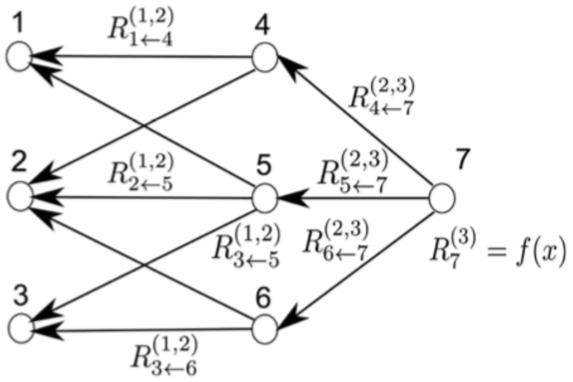
$$f(x) = \cdots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \cdots = \sum_d R_d^{(1)}$$

Layer-wise Relevance Propagation (LRP) -

Bach et al., PLOS ONE, 2015



Testing



Relevance propagation

<u>Conditions</u>

$$R_7^{(3)} = R^{(2,3)}_{4\leftarrow 7} + R^{(2,3)}_{5\leftarrow 7} + R^{(2,3)}_{6\leftarrow 7}$$

$$R_4^{(2)} = R^{(1,2)}_{1\leftarrow 4} + R^{(1,2)}_{2\leftarrow 4}$$

$$R_5^{(2)} = R^{(1,2)}_{1\leftarrow 5} + R^{(1,2)}_{2\leftarrow 5} + R^{(1,2)}_{3\leftarrow 5}$$

$$R_6^{(2)} = R^{(1,2)}_{2\leftarrow 6} + R^{(1,2)}_{3\leftarrow 6}$$

As a function of activations “a” and neuron weights “w”. Example:

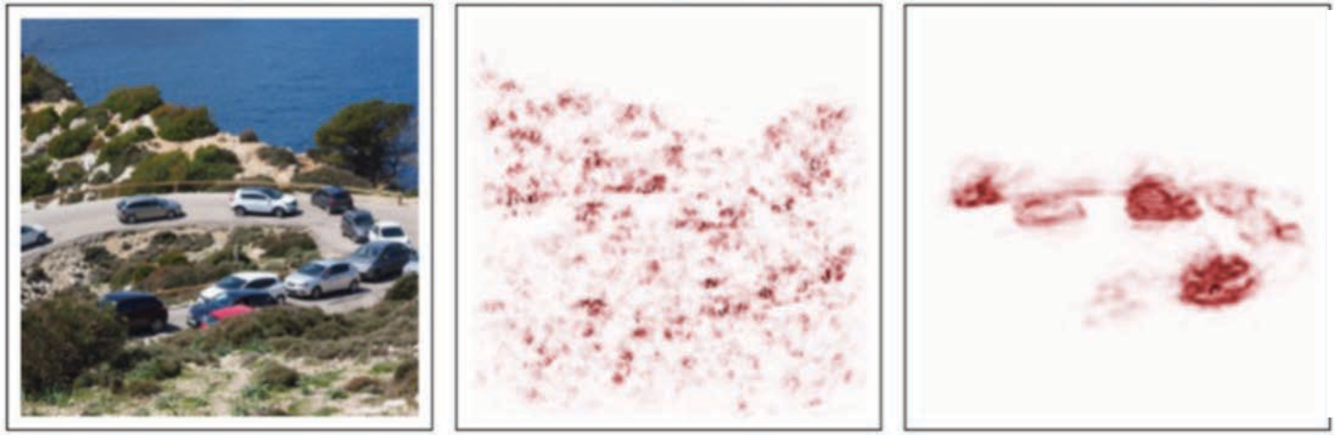
$$R_4^{(2)} = R_4^{(2)} \frac{a_1 w_{14}}{\sum_{i=1,2} a_i w_{i4}} + R_4^{(2)} \frac{a_2 w_{24}}{\sum_{i=1,2} a_i w_{i4}}$$

General formulation

$$\boxed{R^{(l,l+1)}_{i\leftarrow k} = R_k^{(l+1)} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}}$$

Layer-wise Relevance Propagation (LRP) -

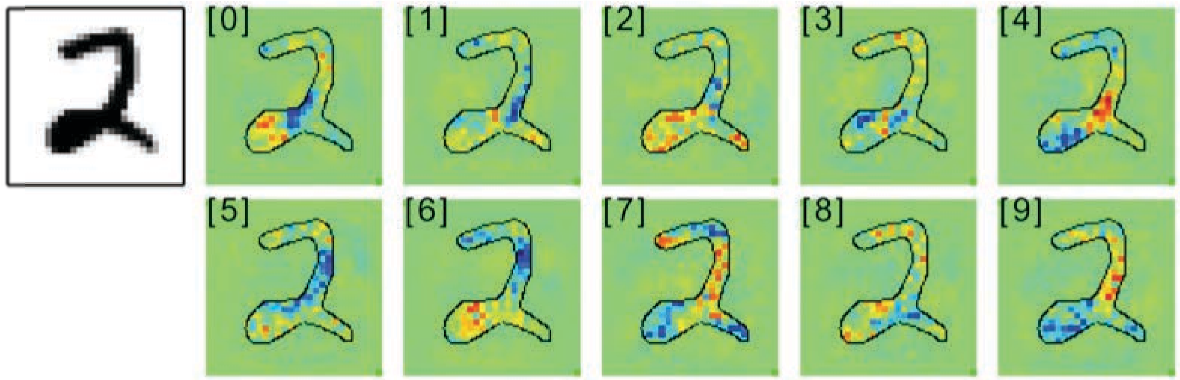
Bach et al., PLOS ONE, 2015

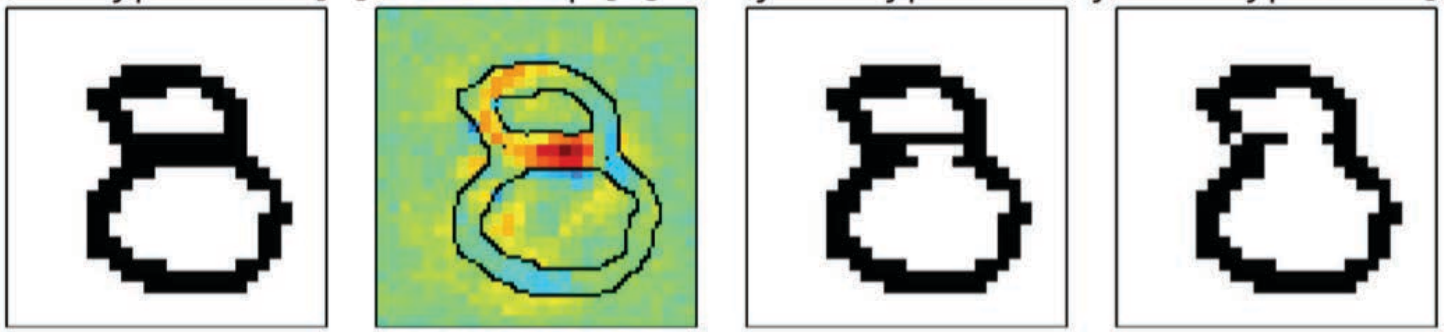


Explains what influences prediction "cars".

Explains prediction "cars" as is.

Interpreting MNIST predictions

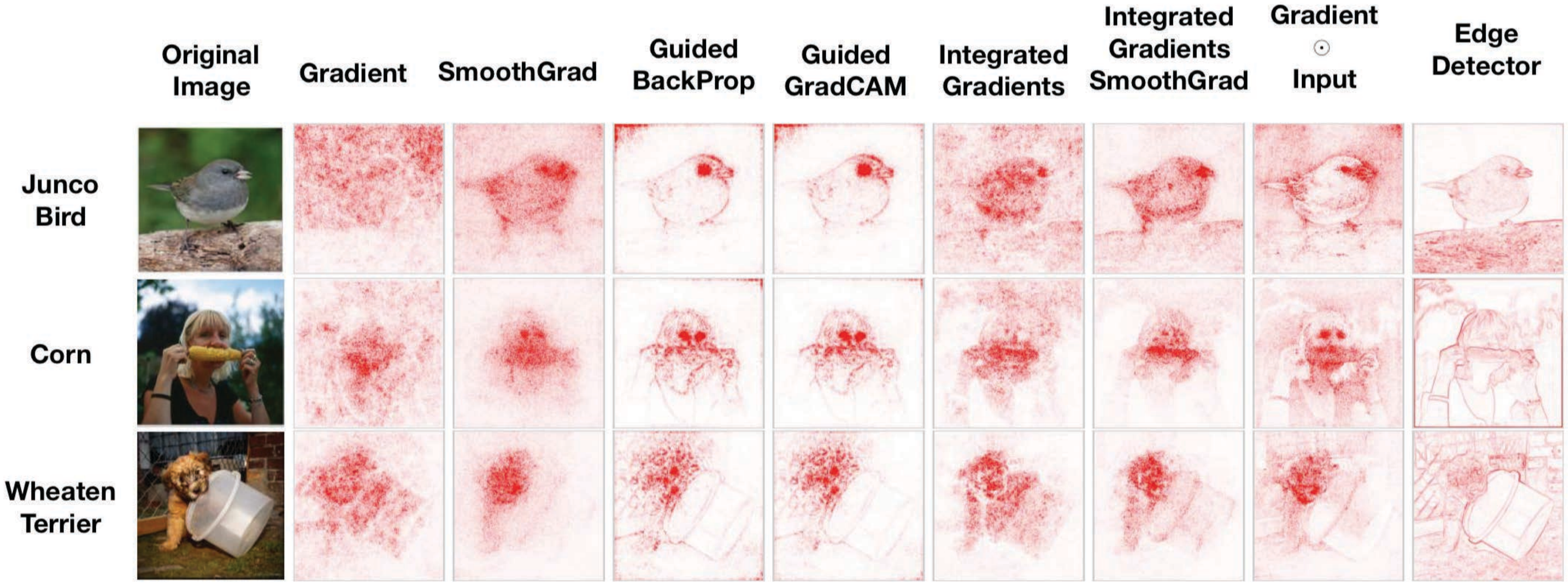




Sanity Checks for Saliency Maps

Julius Adebayo[*], Justin Gilmer[♯], Michael Muelly[♯], Ian Goodfellow[♯], Moritz Hardt[♯†], Been Kim[♯]
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
[♯]Google Brain
[†]University of California Berkeley

Motivation: Objective Evaluation of Interpretability Methods
Observation: Some methods are independent of the model and data processing.



Sanity checks

Sanity Checks for Saliency Maps

Julius Adebayo,[*] Justin Gilmer[♯], Michael Muelly[♯], Ian Goodfellow[♯], Moritz Hardt[♯†], Been Kim[♯]
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
[♯]Google Brain
[†]University of California Berkeley

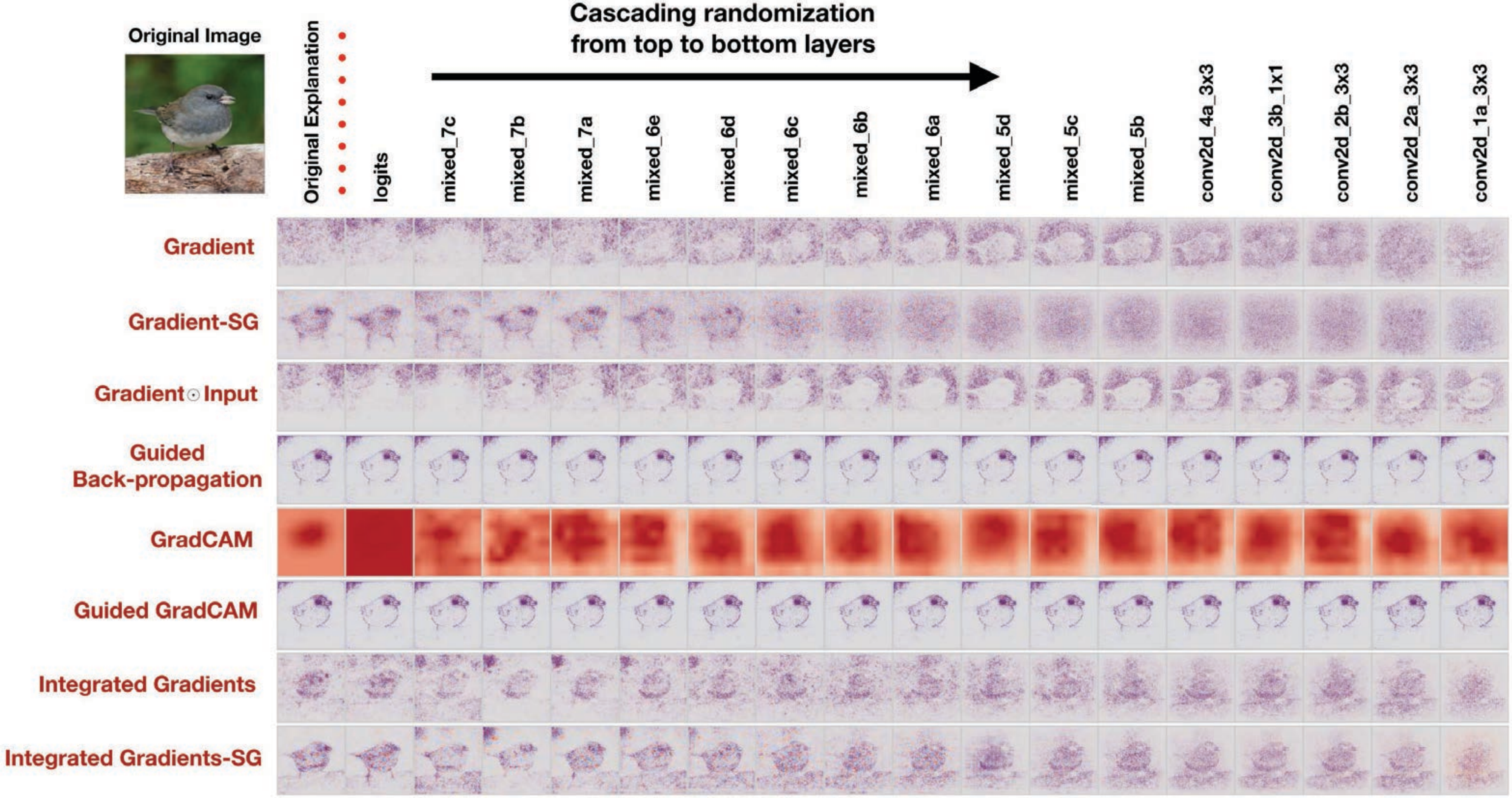
Observation: Some methods are independent of the model and data generating process

- The **model parameter randomization test**: perturbations to model parameters
- **The data randomization test** : perturbations to labels

Other quantitative ways include:
- "Fragility": find adversarial examples that most affect the saliency maps
- As surrogate for image classification, i.e. thresholded saliency maps used as localization ROI

Assessing similarity among maps
- Structural similarity maps – SSIM
- Spearman rank correlation
- Pearson correlation of the histogram of gradients (HOGs)

Sanity Checks for Saliency Maps

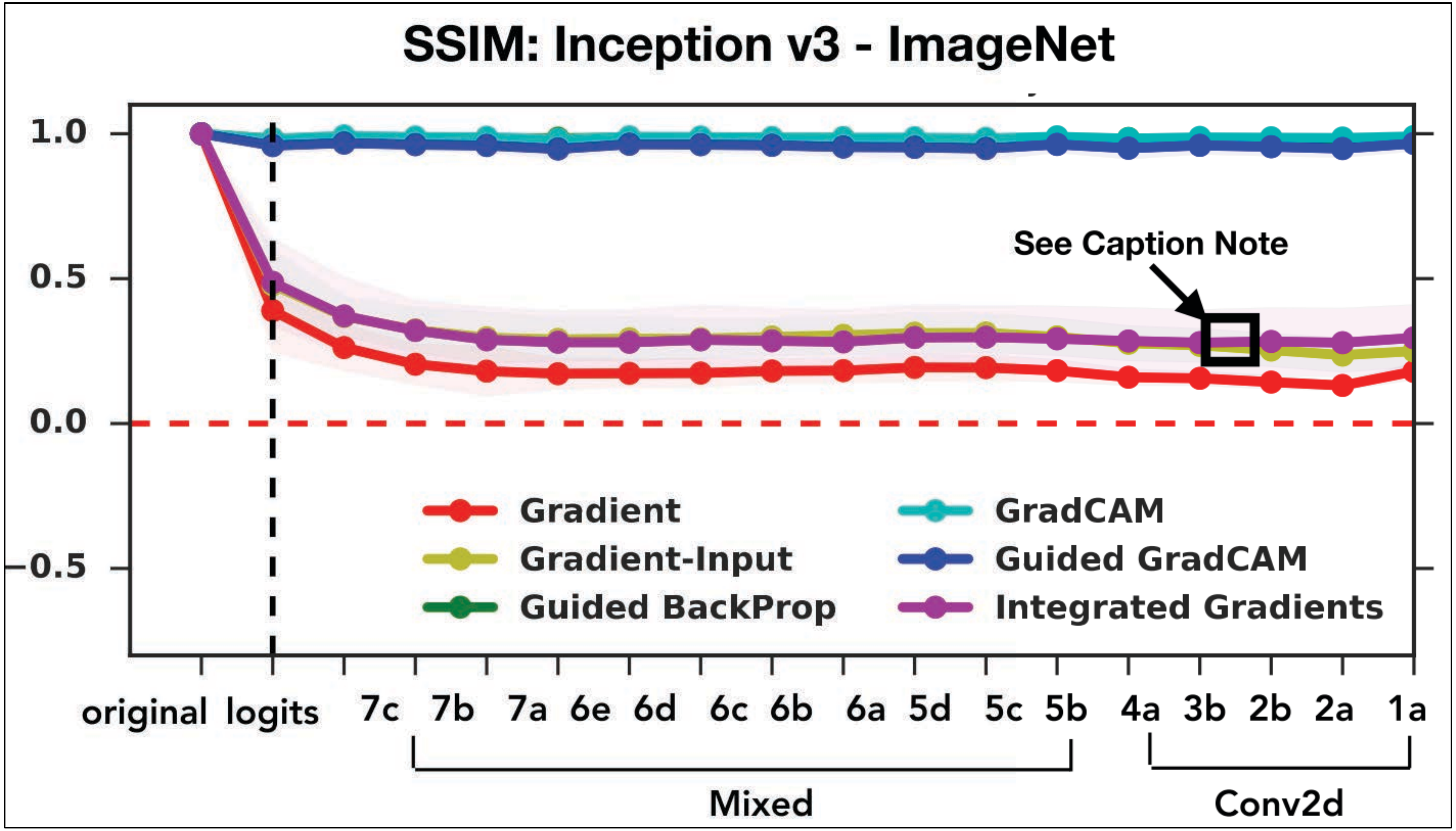


Cascaded randomization on Inception v3

Findings:

Gradients & GradCAM passed the sanity checks, while Guided BackProp & Guided GradCAM fail

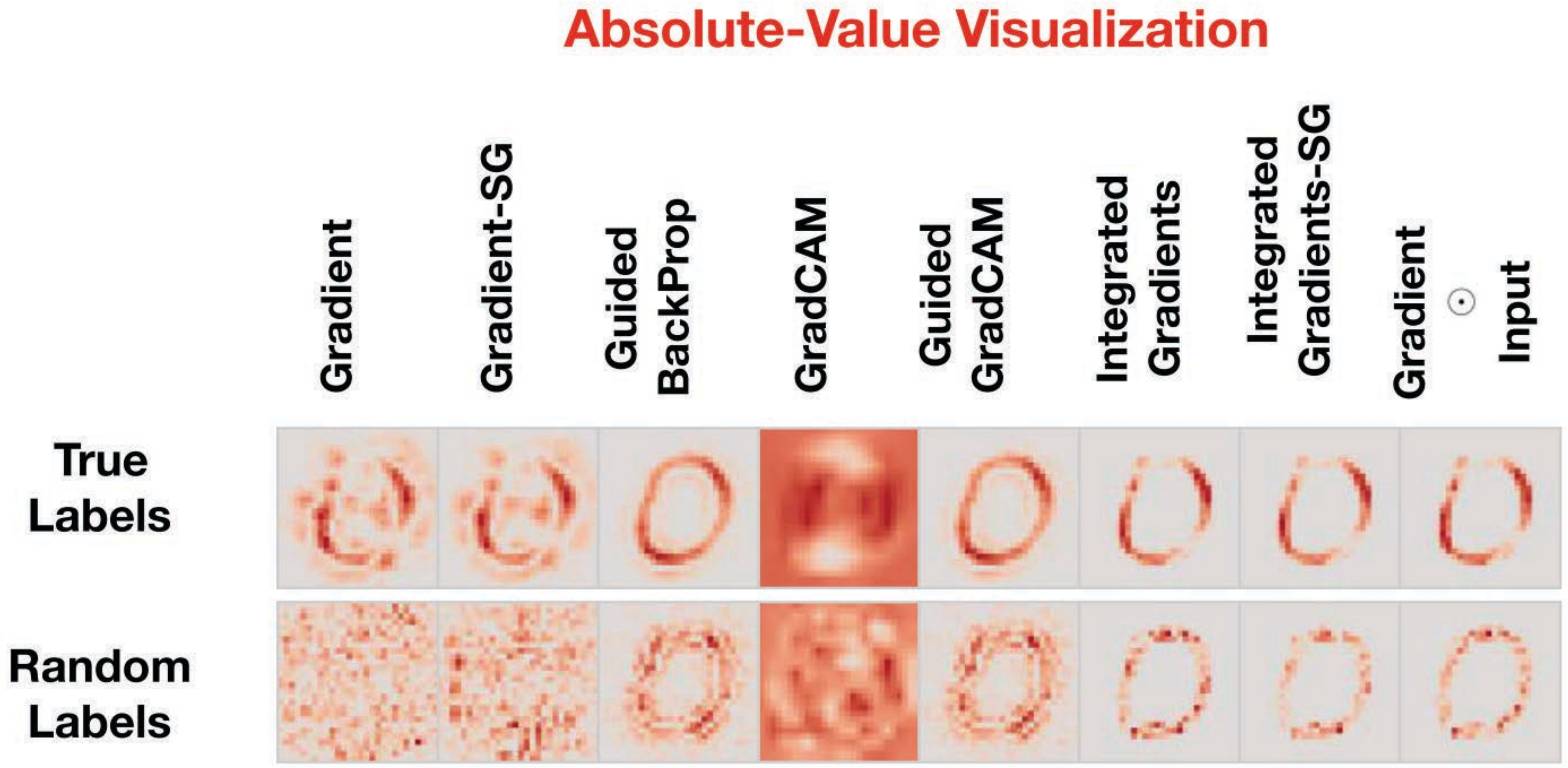
Sanity Checks for Saliency Maps



Cascaded randomization on Inception v3 – SSIM across layers

Findings:
Gradients & GradCAM passed the sanity checks, while Guided BackProp & Guided GradCAM fail

Sanity Checks for Saliency Maps



Explanations of “zero” on true vs randomized labels

Findings:
Gradients & GradCAM passed the sanity checks, while Guided BackProp & Guided GradCAM fail

Sanity Checks for Saliency Maps

Julius Adebayo*, Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley

Other findings/confirmation from previous works:

- LRP and DeepLIFT are equivalent to (input *gradient) for ReLU network w/o positive biases.
- Explanations of methods of the type "input * gradient" are heavily based on input structure → edge detector effect
- Explanations that do not depend on model parameters or training data might still depend on the model architecture, and thus provide some useful information about the prior incorporated in the model architecture.

See more benchmarking here:

- DeepLIFT often approximates Integrated gradients (IG)
- DeepLIFT != IG and fail to produce meaningful results when applied to RNNs with multiplicative interactions (e.g. LSTM)

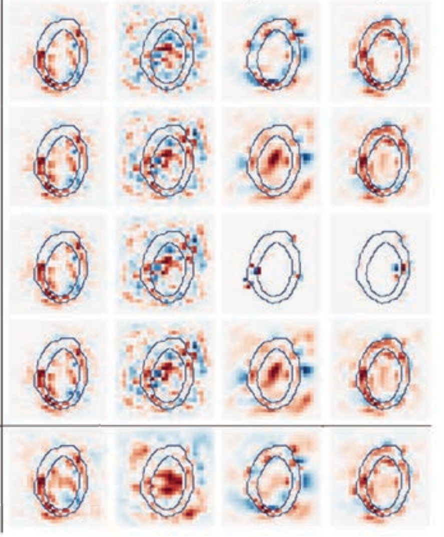
Published as a conference paper at ICLR 2018

TOWARDS ... GRADIENT ... FOR DEEP

Marco Ancona
Department of Co...
ETH Zurich, Swit...
marco.ancona...

Cengiz Öztireli
Department of Co...
ETH Zurich, Swit...
cengizo@inf....

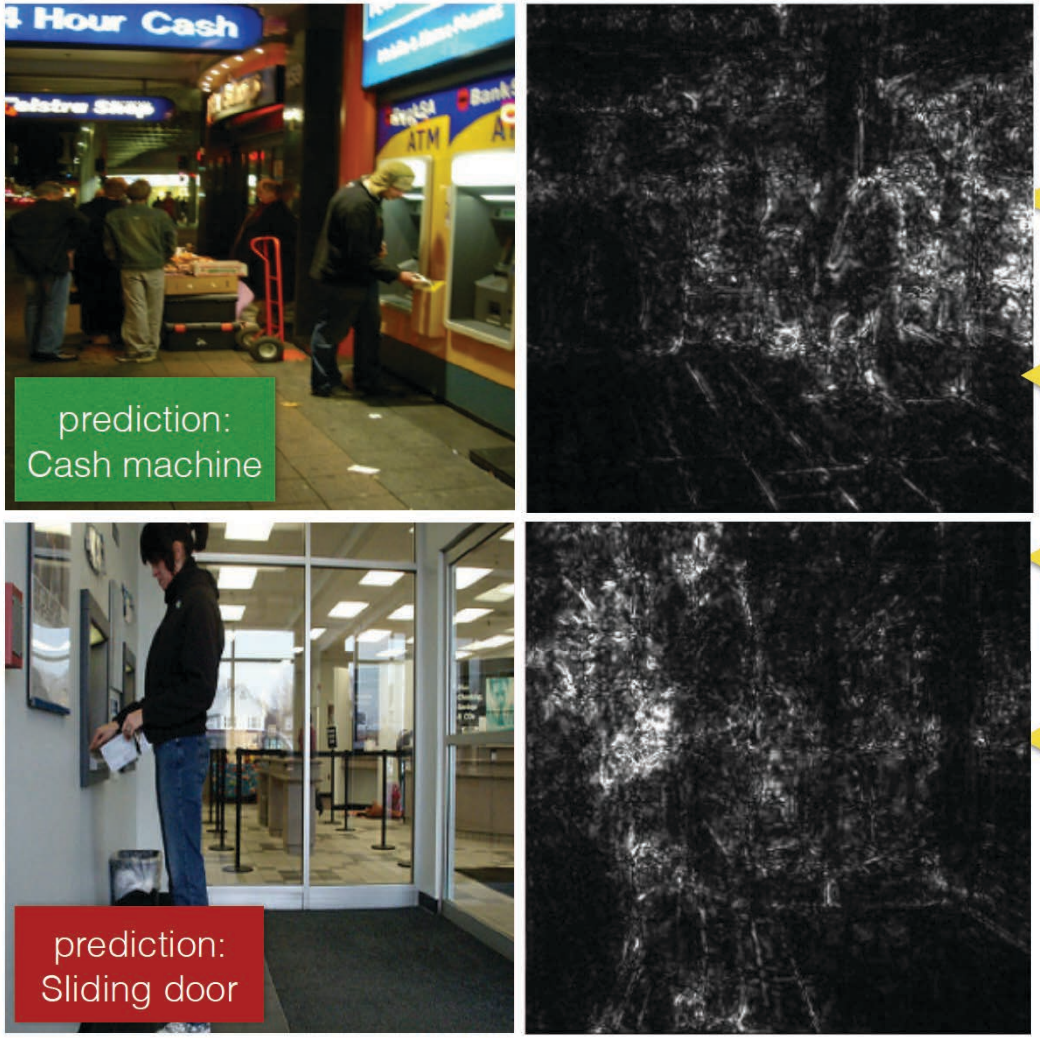
Method	Attribution $R_i^c(x)$	Example of attributions on MNIST (ReLU, Tanh, Sigmoid, Softplus)
Gradient * Input	$x_i \cdot \frac{\partial S_c(x)}{\partial x_i}$	
Integrated Gradient	$(x_i - \bar{x}_i) \cdot \int_{\alpha=0}^{1} \frac{\partial S_c(\tilde{x})}{\partial(\tilde{x}_i)}\Big\vert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$	
ϵ-LRP	$x_i \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z)}{z}$	
DeepLIFT	$(x_i - \bar{x}_i) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \quad g = \frac{f(z) - f(\bar{z})}{z - \bar{z}}$	
Occlusion-1	$S_c(x) - S_c(x_{[x_i=0]})$	



Wrap-up saliency methods

- Many approaches have been proposed!
- Visual evaluation alone can be misleading
- Recent findings on deficiencies and similarities of gradient-based approaches
- (Recall) all these methods require access to the model architecture
- Assessment of methods is an active field of research
- Seemingly lack of integration of expert-knowledge (e.g. more methodological way of defining “reference” points for interpretability)

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017



Were there more pixels on the cash machine than on the person?

Did the 'human' concept matter?
Did the 'glasses' or 'paper' matter?

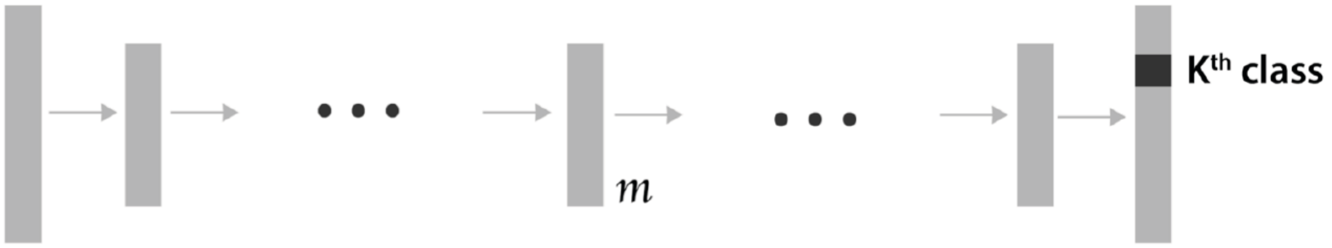
Which concept mattered more?

Is this true for all other cash machine predictions?

Wouldn't it be great if we can **quantitatively** measure how important *any* of these **user-chosen concepts** are?

Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017



Quantitative explanation: how much a concept (e.g., gender, race) was important for a prediction in a trained model.

…even if the concept was not part of the training.

Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

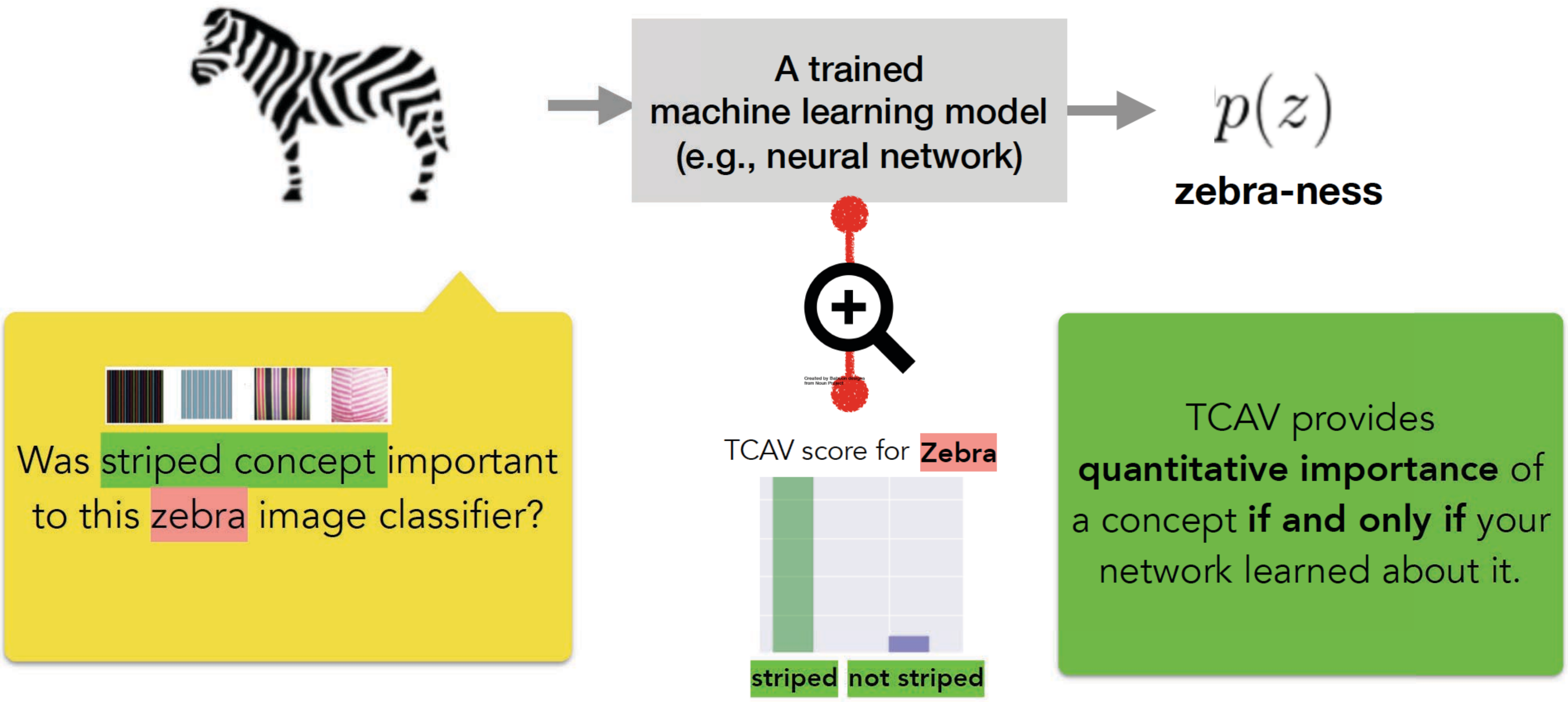


Image source: Kim Been

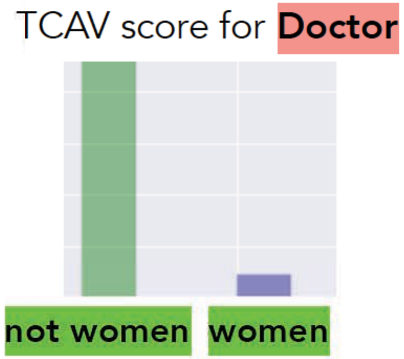
Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

A trained
machine learning model
(e.g., neural network)

$p(z)$

Doctor-ness

Was gender concept important to this doctor image classifier?



TCAV provides **quantitative importance** of a concept **if and only if** your network learned about it.

Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

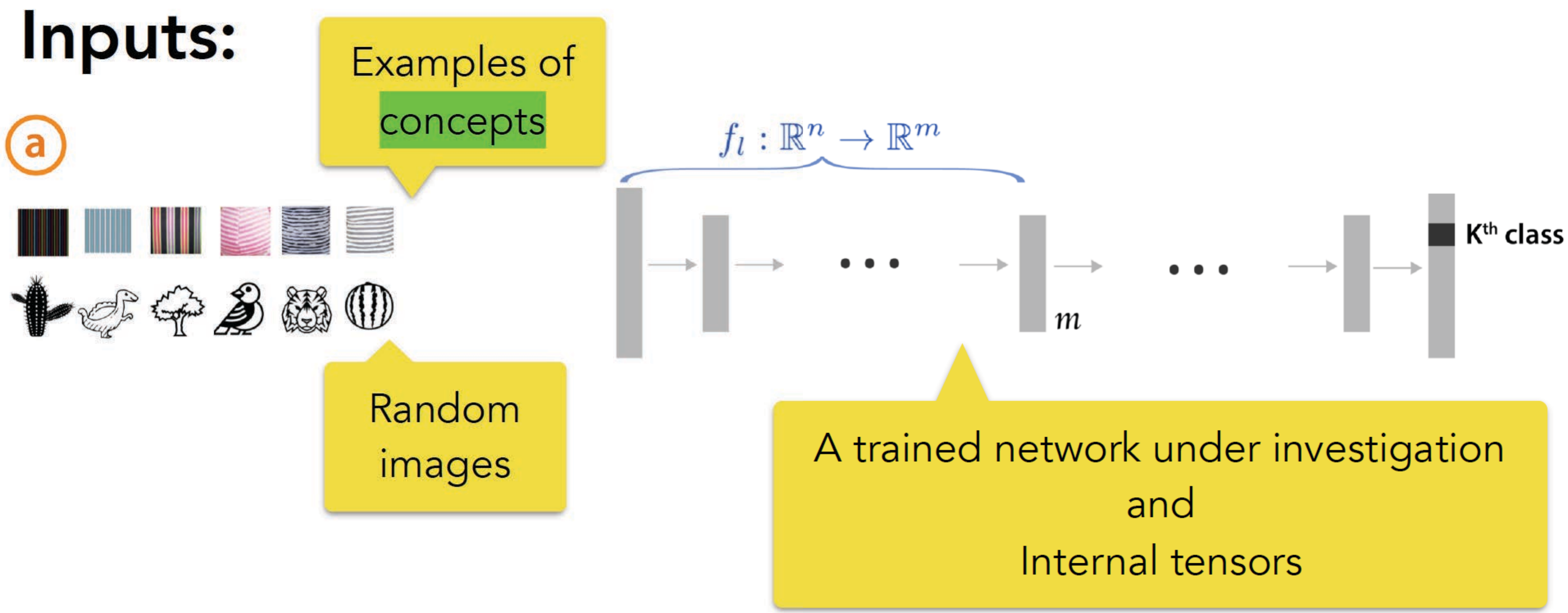


Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

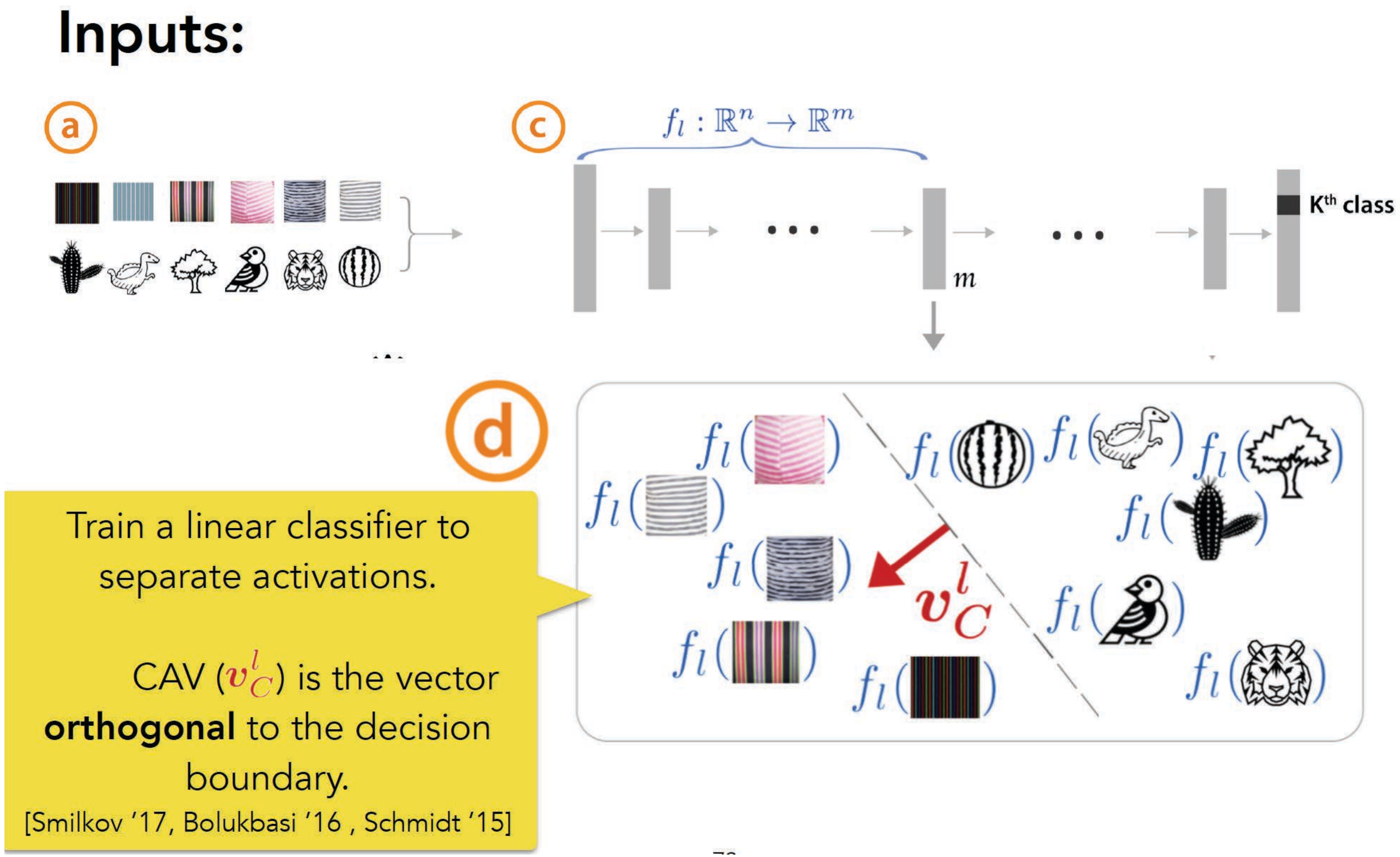
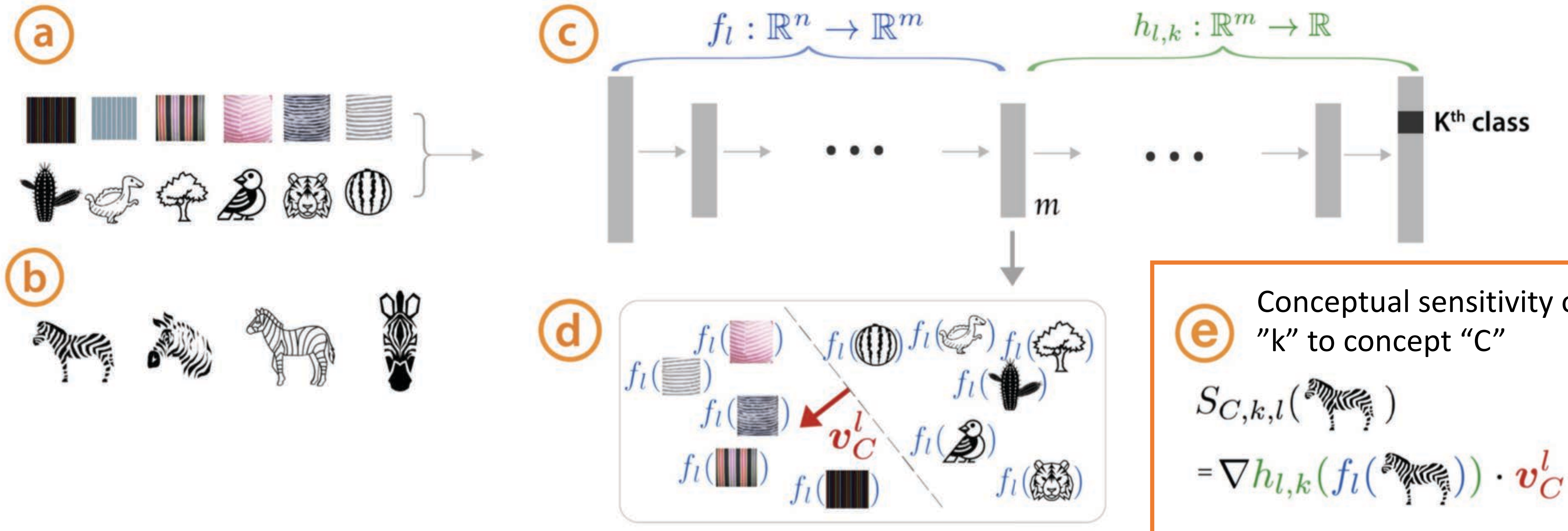


Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

Logit for layer “l” and class “k”



Conceptual sensitivity of class ”k” to concept “C”

$S_{C,k,l}(\cdot)$

$= \nabla h_{l,k}(f_l(\cdot)) \cdot v_C^l$

Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

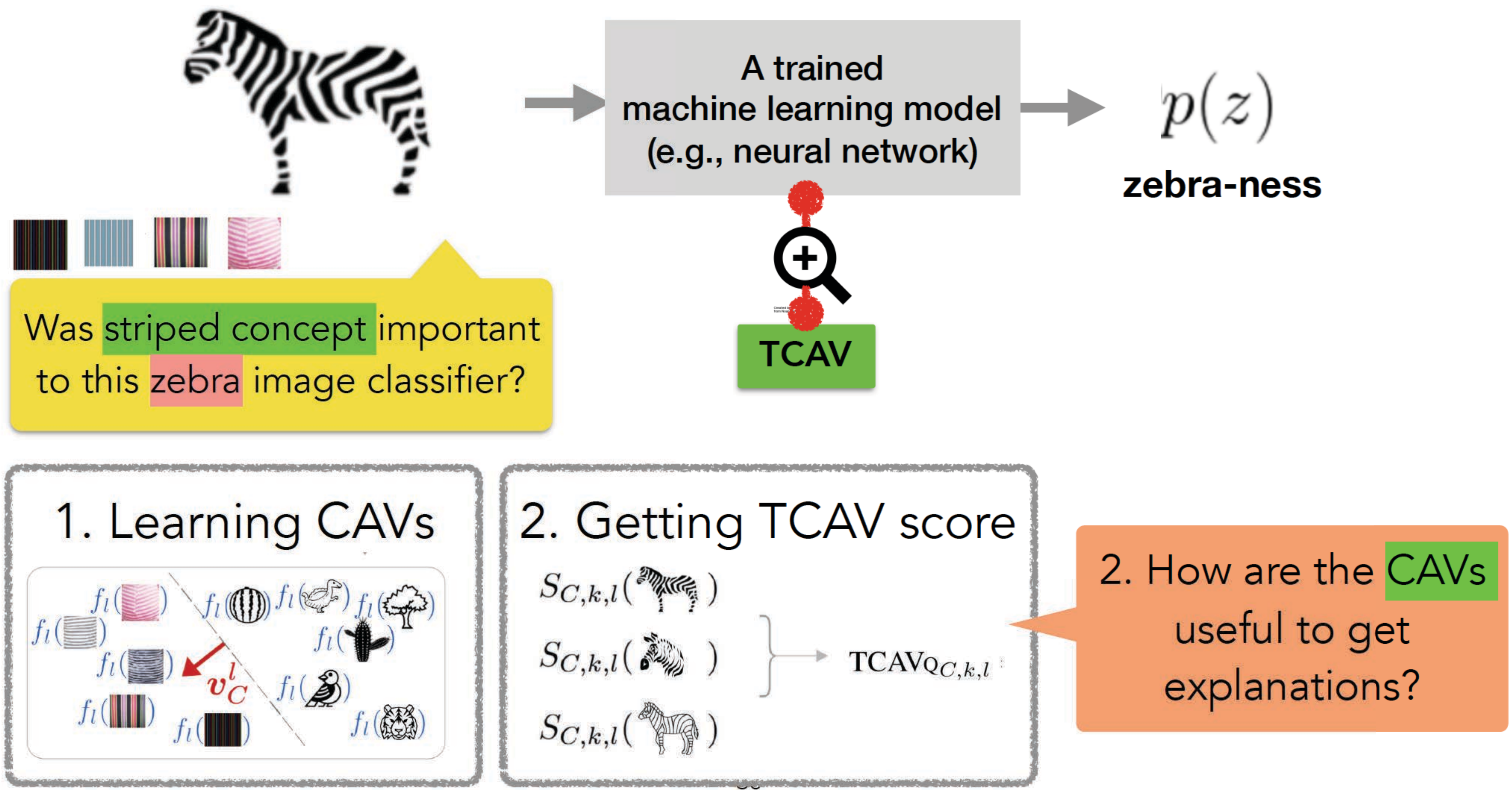
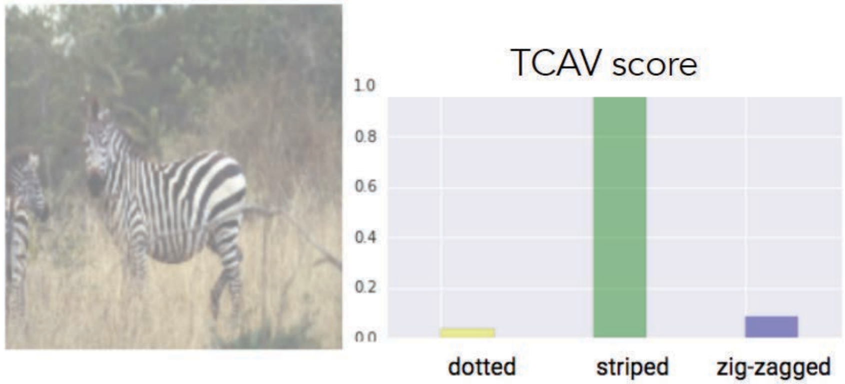


Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

TCAV

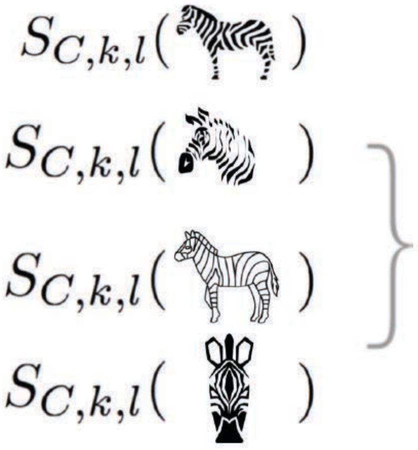


zebra-ness → $\partial p(z)$

striped CAV → $\partial \boldsymbol{v}_C^l$

$$\frac{\partial p(z)}{\partial \boldsymbol{v}_C^l} = S_{C,k,l}(\boldsymbol{x})$$

Directional derivative with CAV



$$\boxed{\mathrm{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}}$$

fraction of k-class inputs whose l-layer activation vector was positively influenced by concept C

Image source: Kim Been

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

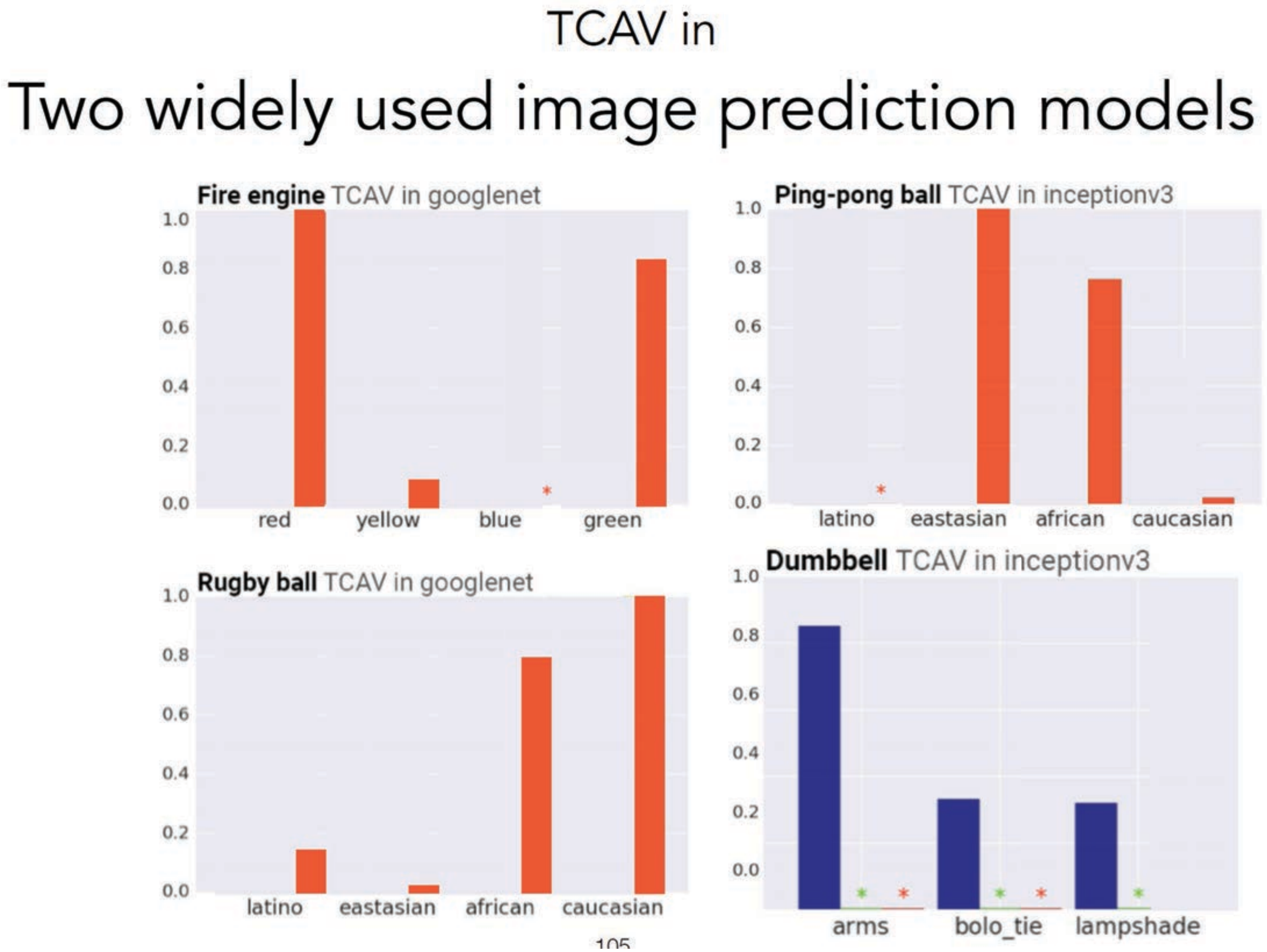


Image source: Kim Been

TCAV facilitates spotting of biases in the datasets

Testing with Concept Activation Vectors – TCAV – Kim et al. 2017

Example application to Diabetic Retinopahty (DR)

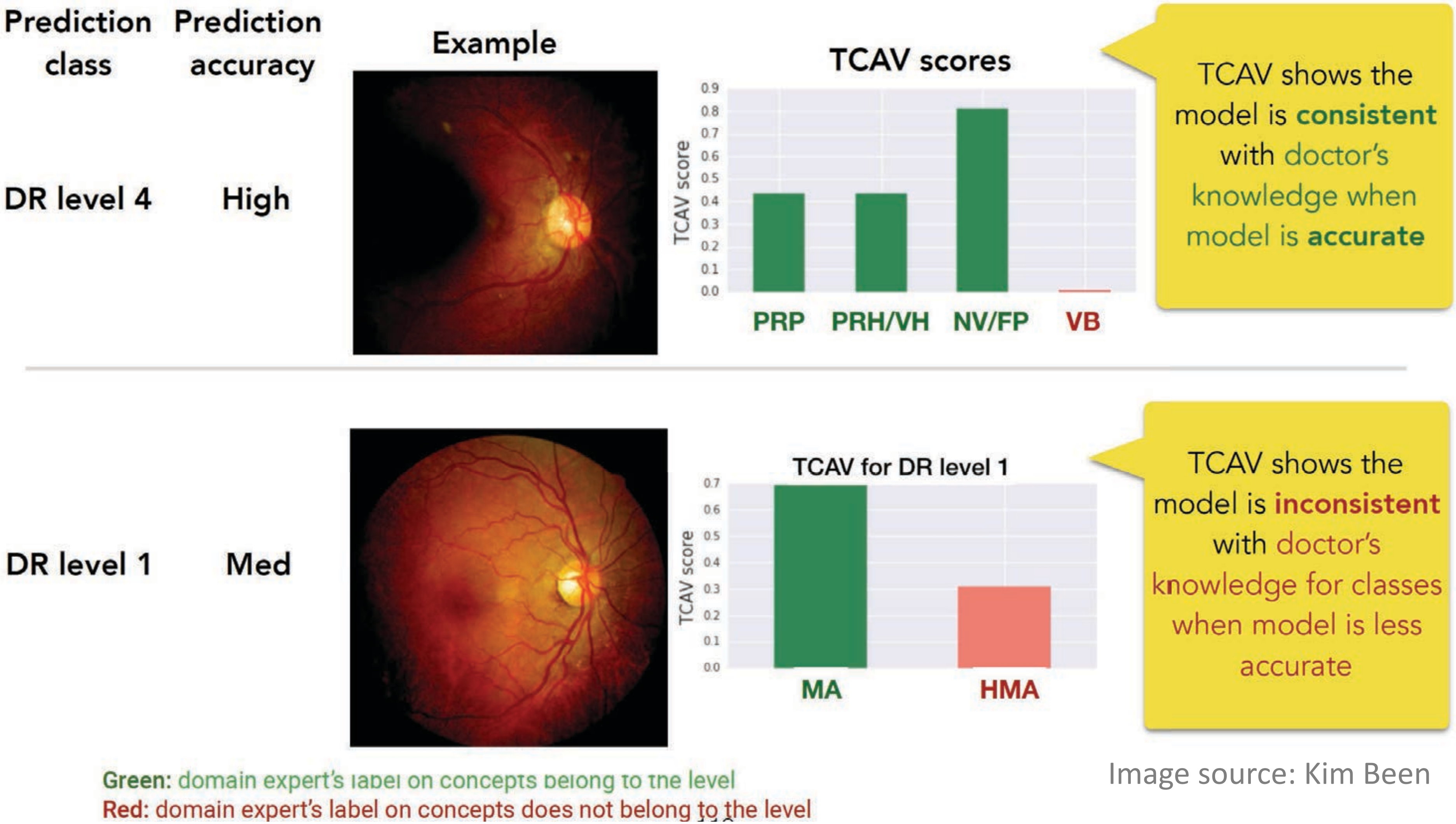
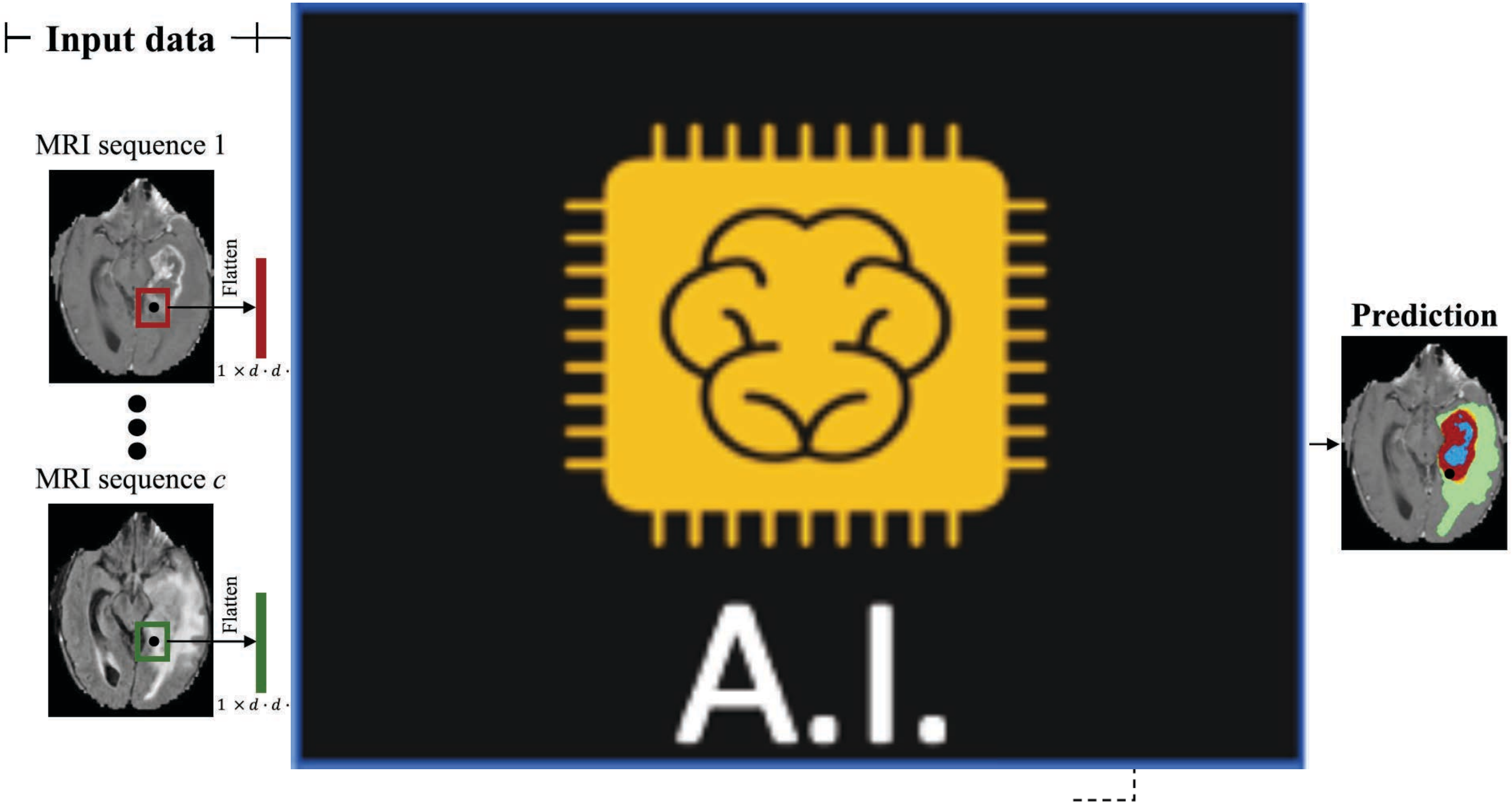


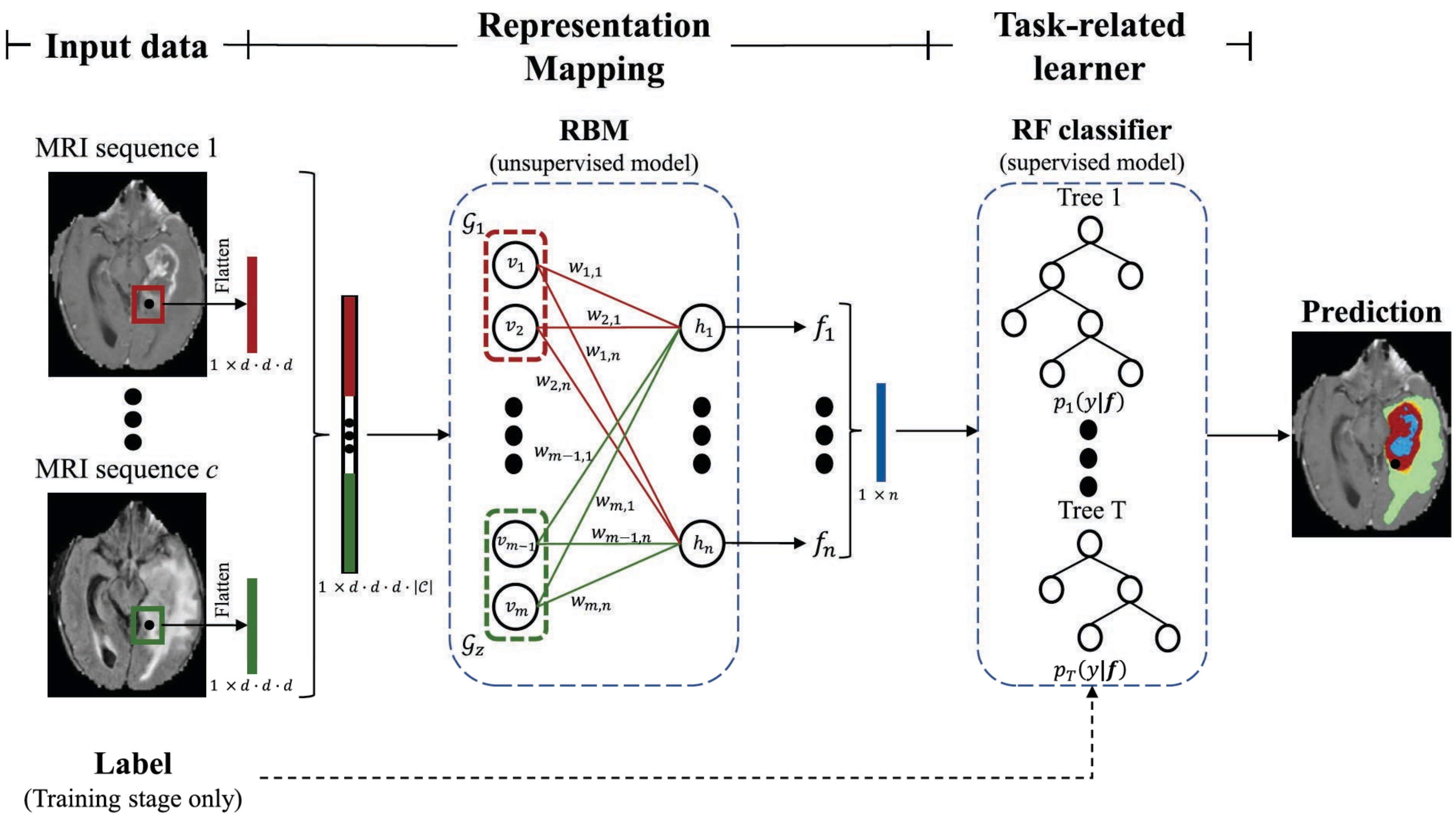
Image source: Kim Been

Enhancing Interpretability of Learned Features in Multisequence-MRI

Example task:
Automated brain tumor segmentation from multisequence MRI



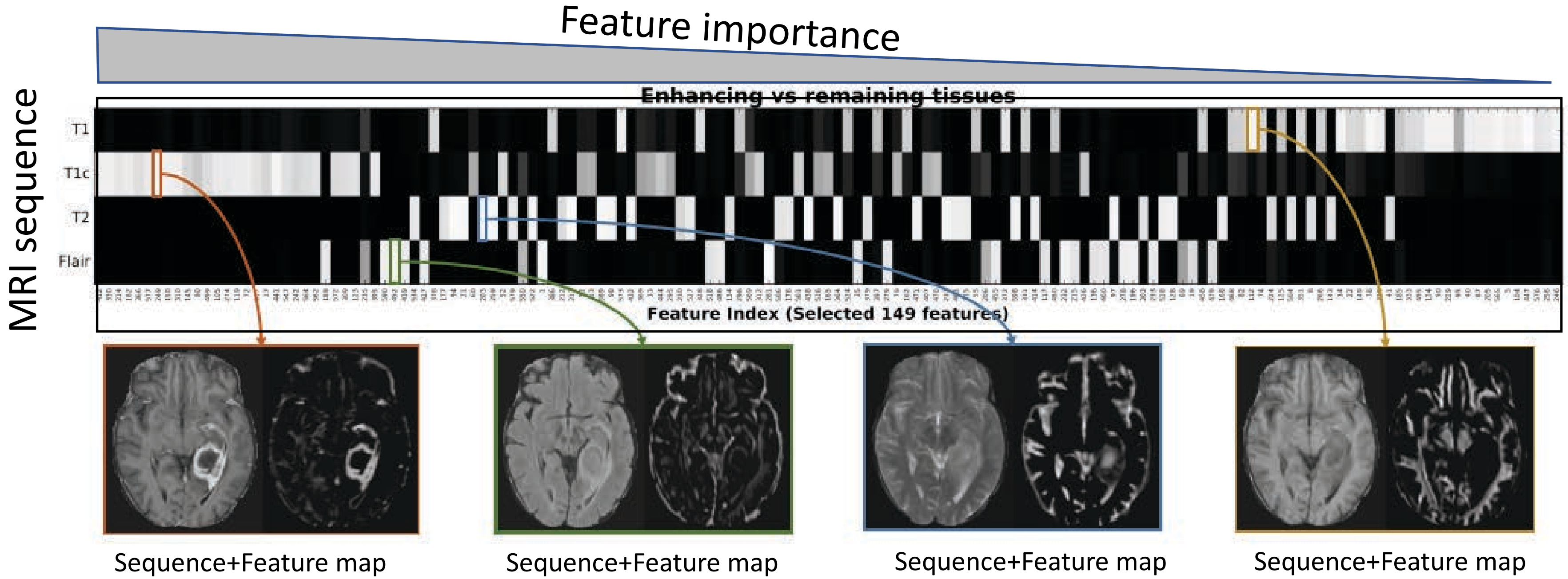
Enhancing Interpretability of Learned Features in Multisequence-MRI



Pereira et al. 2018

Enhancing interpretability of machine learning algorithms

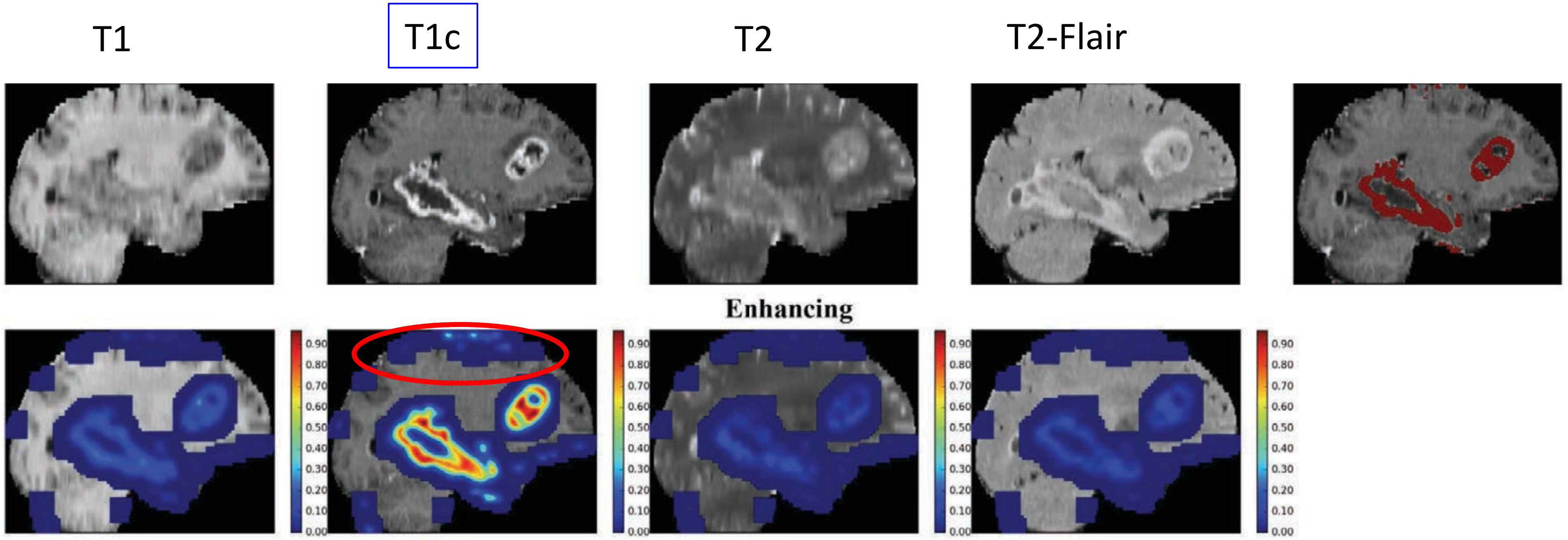
- *How is the machine using the information?*
- *Which sequence is more important?*
- *If it fails, why does it fail?*
- *Quality Certification*



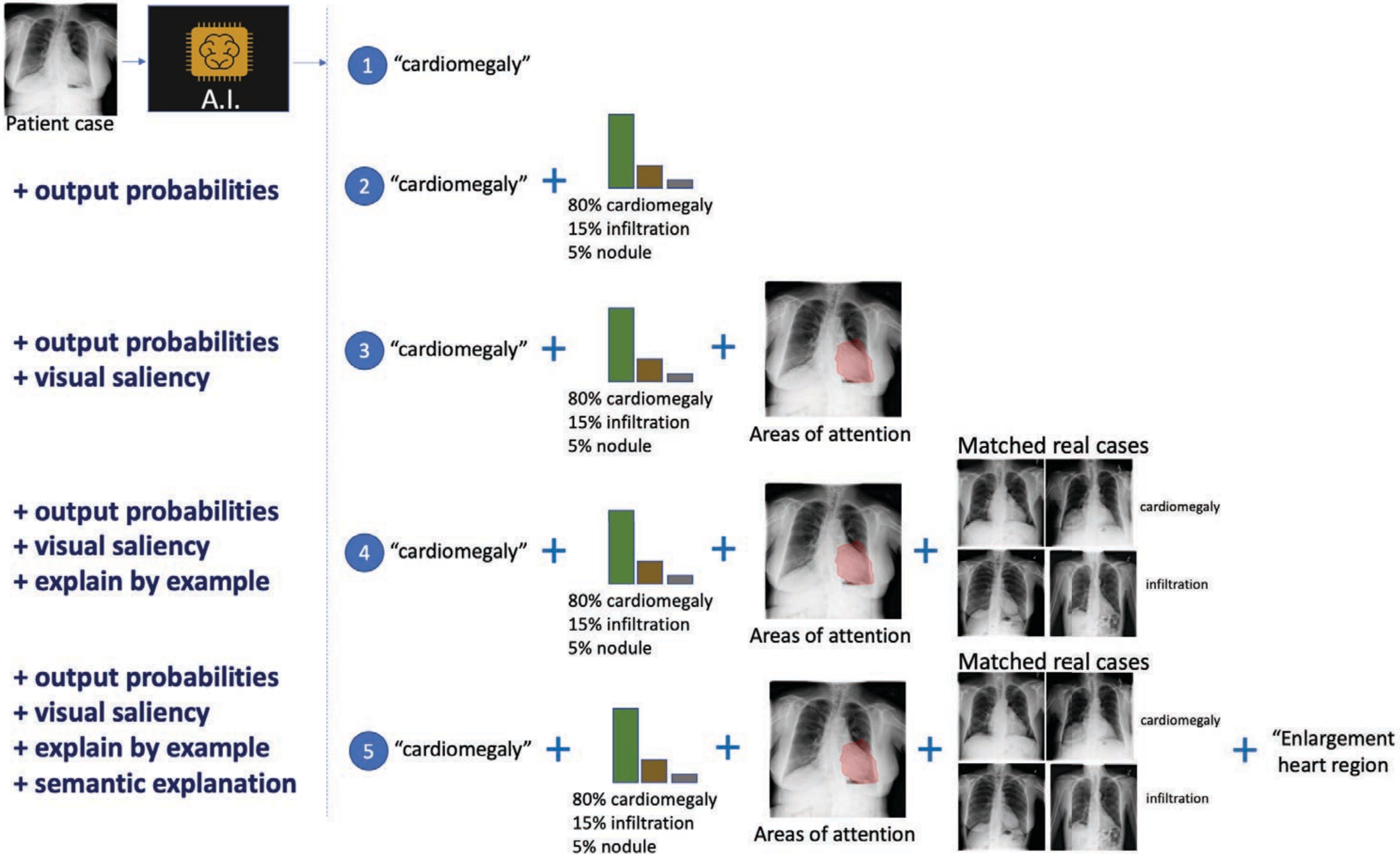
Enhancing interpretability of machine learning algorithms

Which Sequence is used most for the task of tumor-enhancing segmentation?

- *T1c (agrees with clinical practice)*
- *Method also detects bias from pre-processing*



Combining interpretability approaches

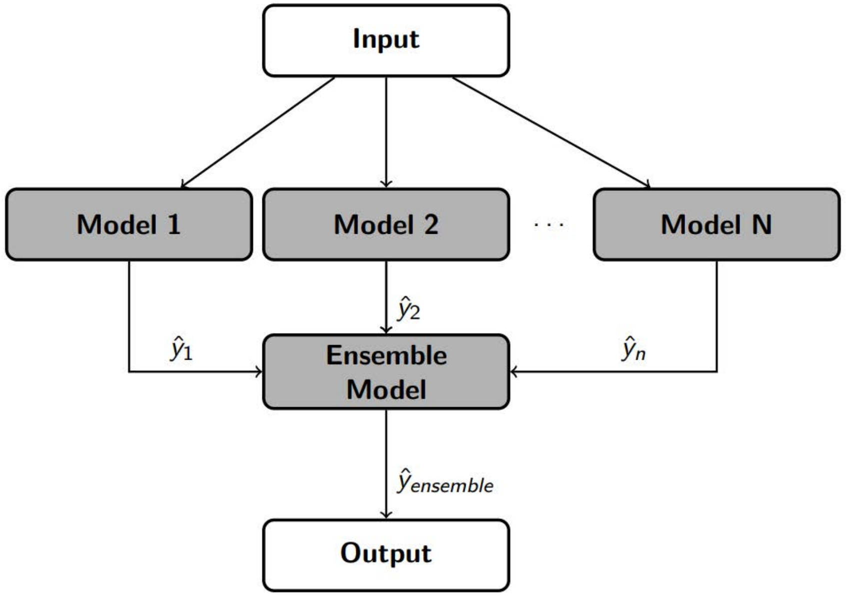


How to produce complementary explanations using an Ensemble Model – Silva et al. 2019 (IJCNN)

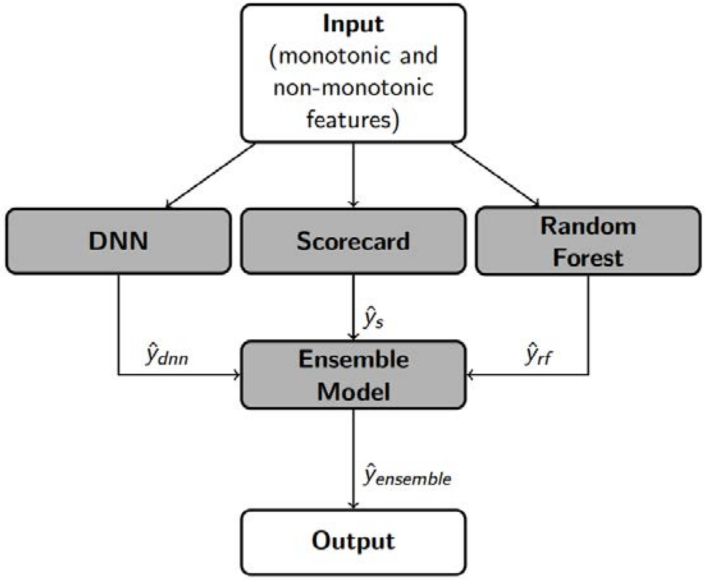
- Motivation – Generating explanations is not easy: model complexity + variability in application domains and target users. Most adequate type of explanation may vary with the example under evaluation.

- Ensemble Model:
  - Different models inside a global one
  - Improvement in quality + diversity of explanations

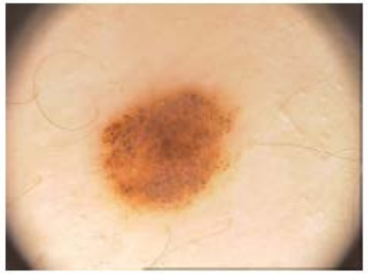
$$E_{global}(x) = \underset{E_n(x)}{\operatorname{argmax}}(corr(E_n(x)))$$



How to produce complementary explanations using an Ensemble Model – Silva et al. 2019 (IJCNN)



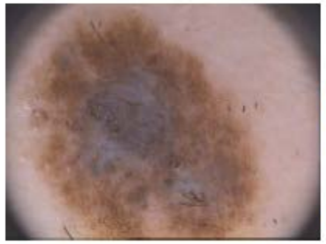
Input image
(**Prediction:** {Common, Atypical})



Random Forest: Lack of regression, presence brown color and lack of white color.

Scorecard: Presence of brown color, lack of atypical dots and asymmetry.

Similar case
Why?: Both images have light and dark brown color and atypical presence of dots/globules.



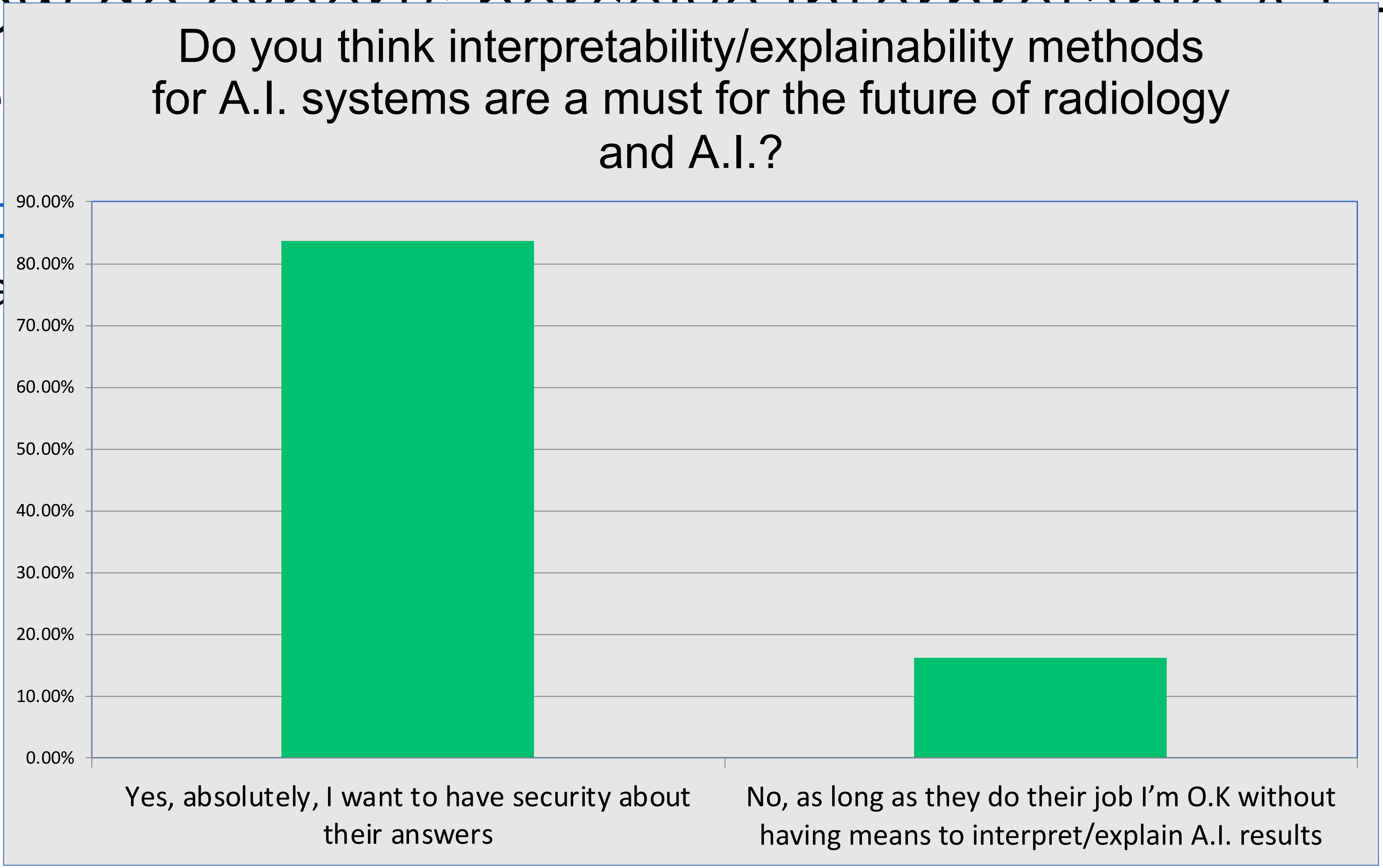
Opponent case
Why?: It doesn't have light brown color or atypical dots/globules. It has blue whitish veil and pigmented network.

How do experts perceive interpretable A.I. –

ne

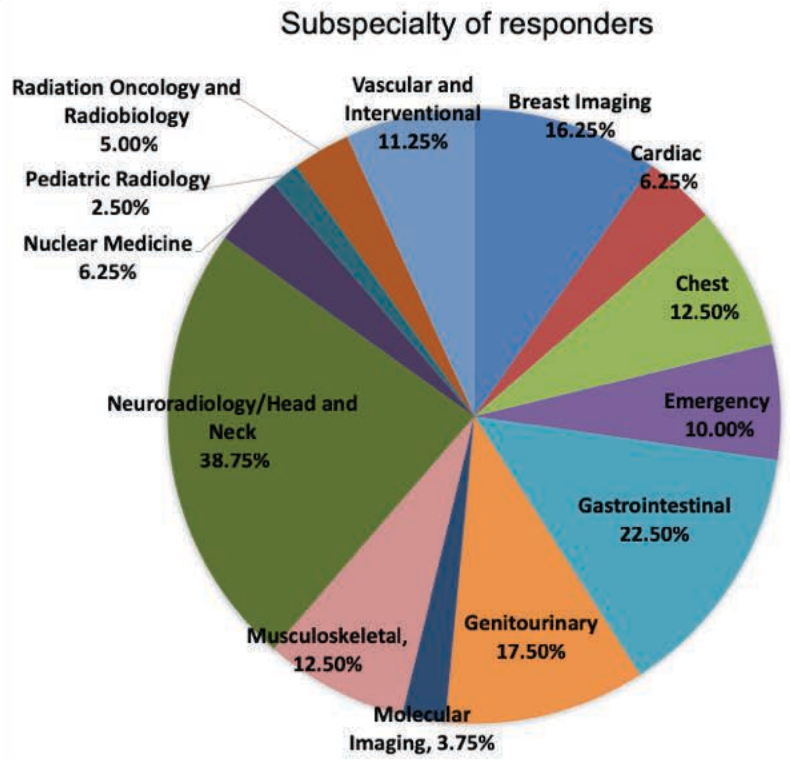
• ht

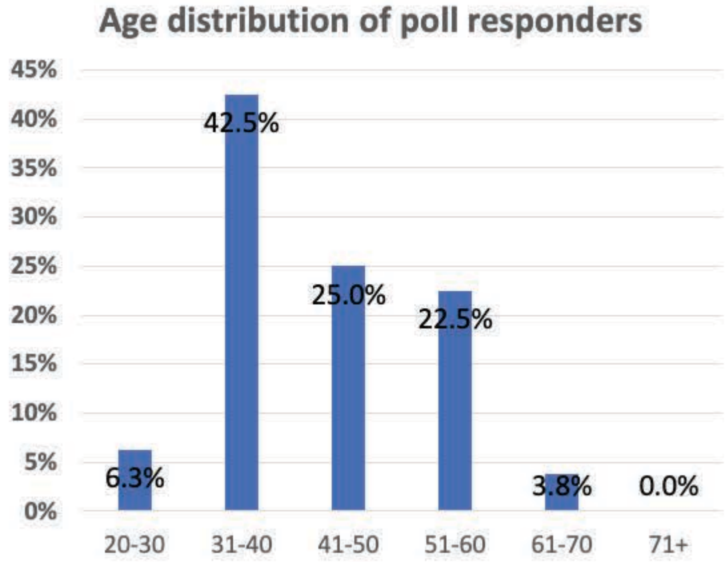
Plea

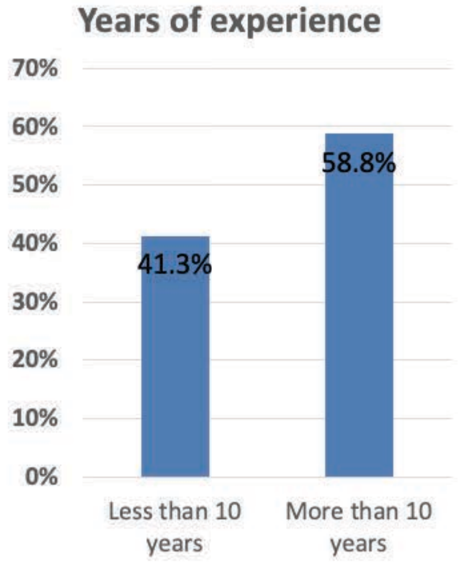


How do experts perceive interpretable A.I. – needed?

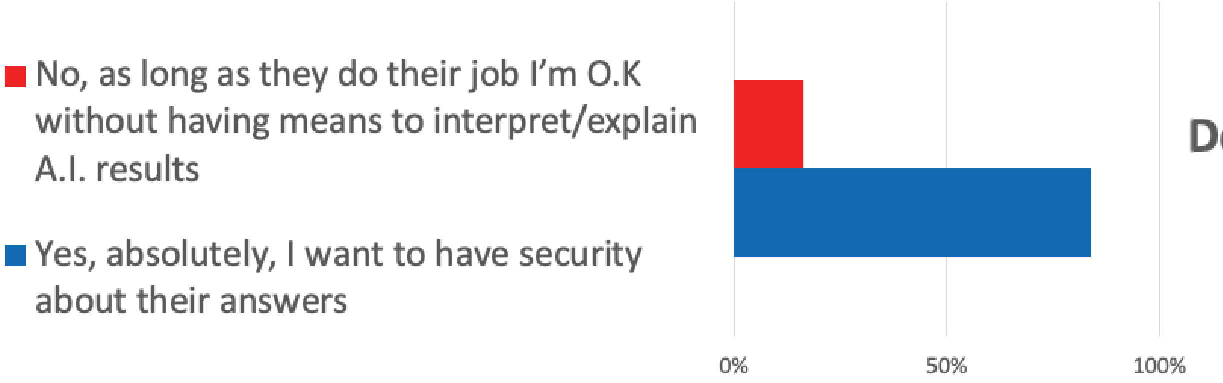
80 responders
Online survey





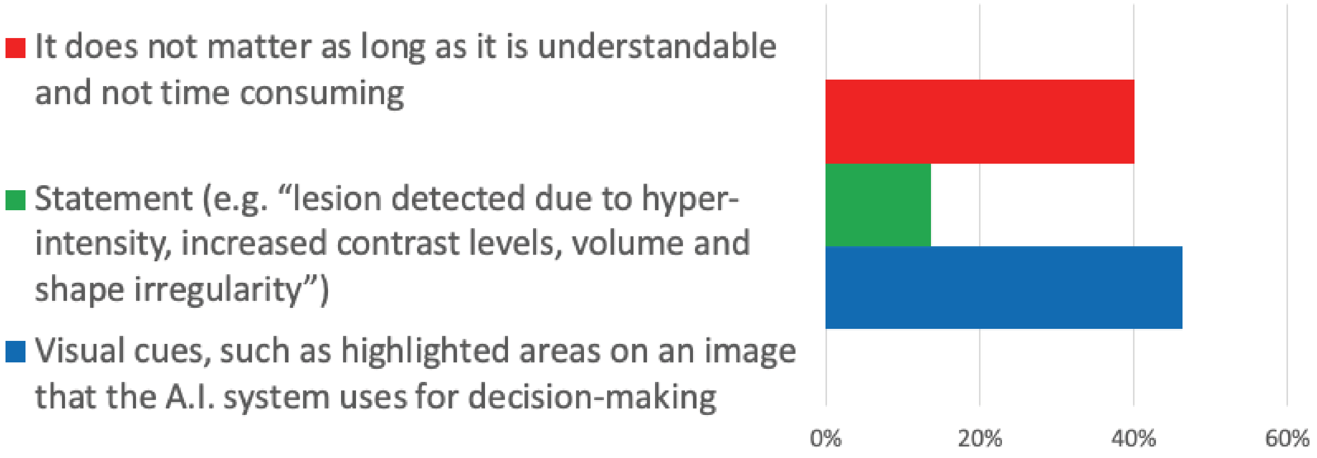


How do experts perceive interpretable A.I. – needed?

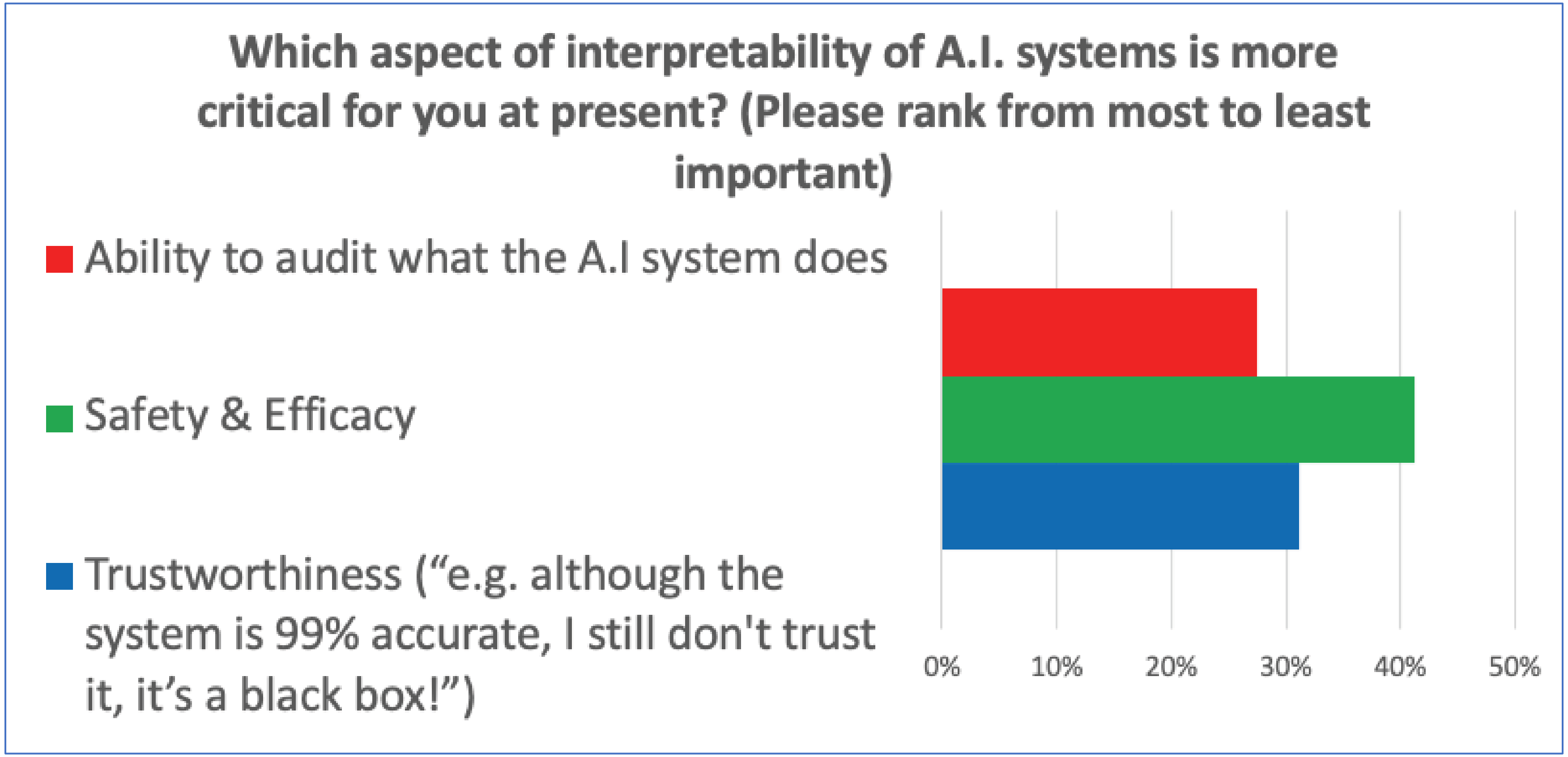


Do you think interpretability/explainability methods for A.I. systems are a must for the future of radiology and A.I.?

What kind of information would you prefer from interpretability methods? (Please rank from most to least important)



How do experts perceive interpretable A.I. – needed?



"Primum non nocere – First, to do no harm"

Improved Interpretability: Deep Learning meets Causal Inference

CONTRIBUTED ARTICLES

The Seven Tools of Causal Inference, with Reflections on Machine Learning

By Judea Pearl
Communications of the ACM, March 2019, Vol. 62 No. 3, Pages 54-60
10.1145/3241036

Counterfactual ("Was it X that caused Y"?)

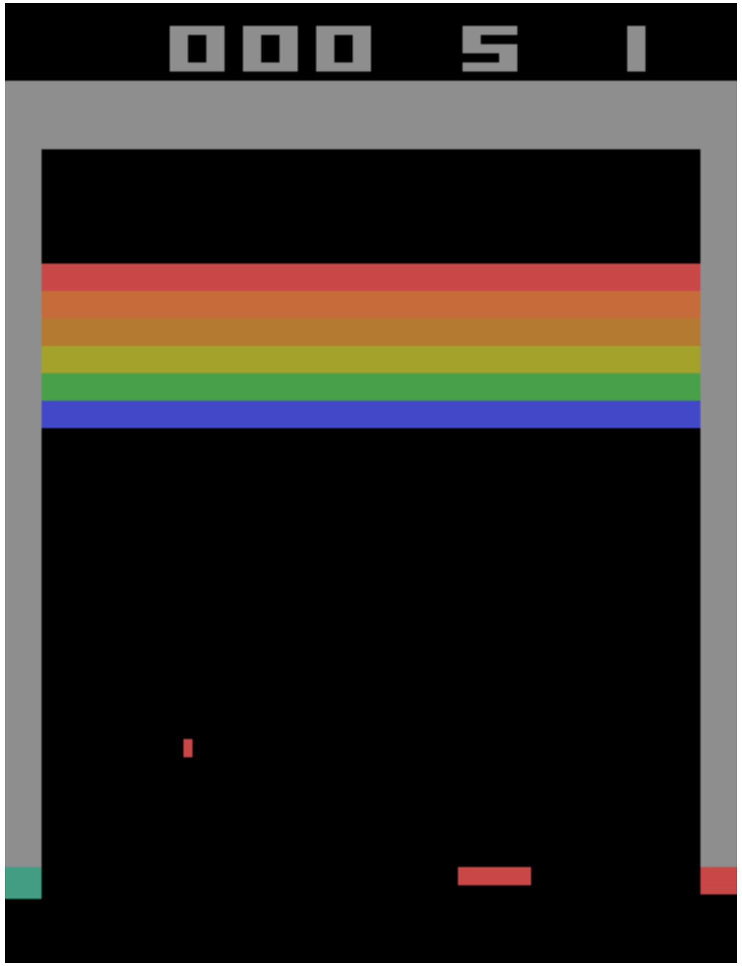
"alternative worlds"

Interventional ("What if?")

Associational ("relation between X and Y")

Counterfactual calls for "alternative worlds" ←→ A.I

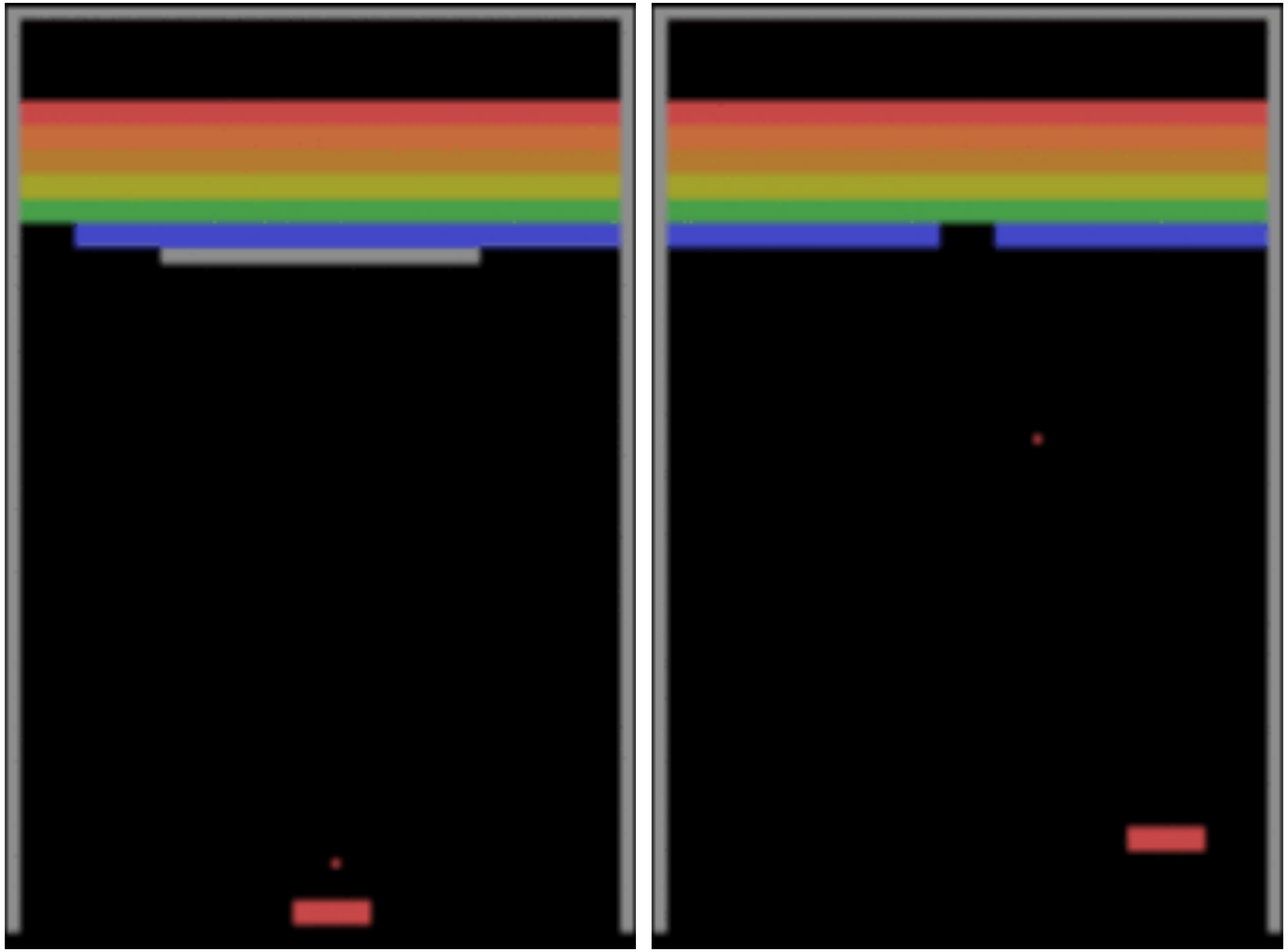
Improved Interpretability: Deep Learning meets Causal Inference



2014 - DeepMind teaches itself to play breakout

..but can the A.I. system adapt to small changes?

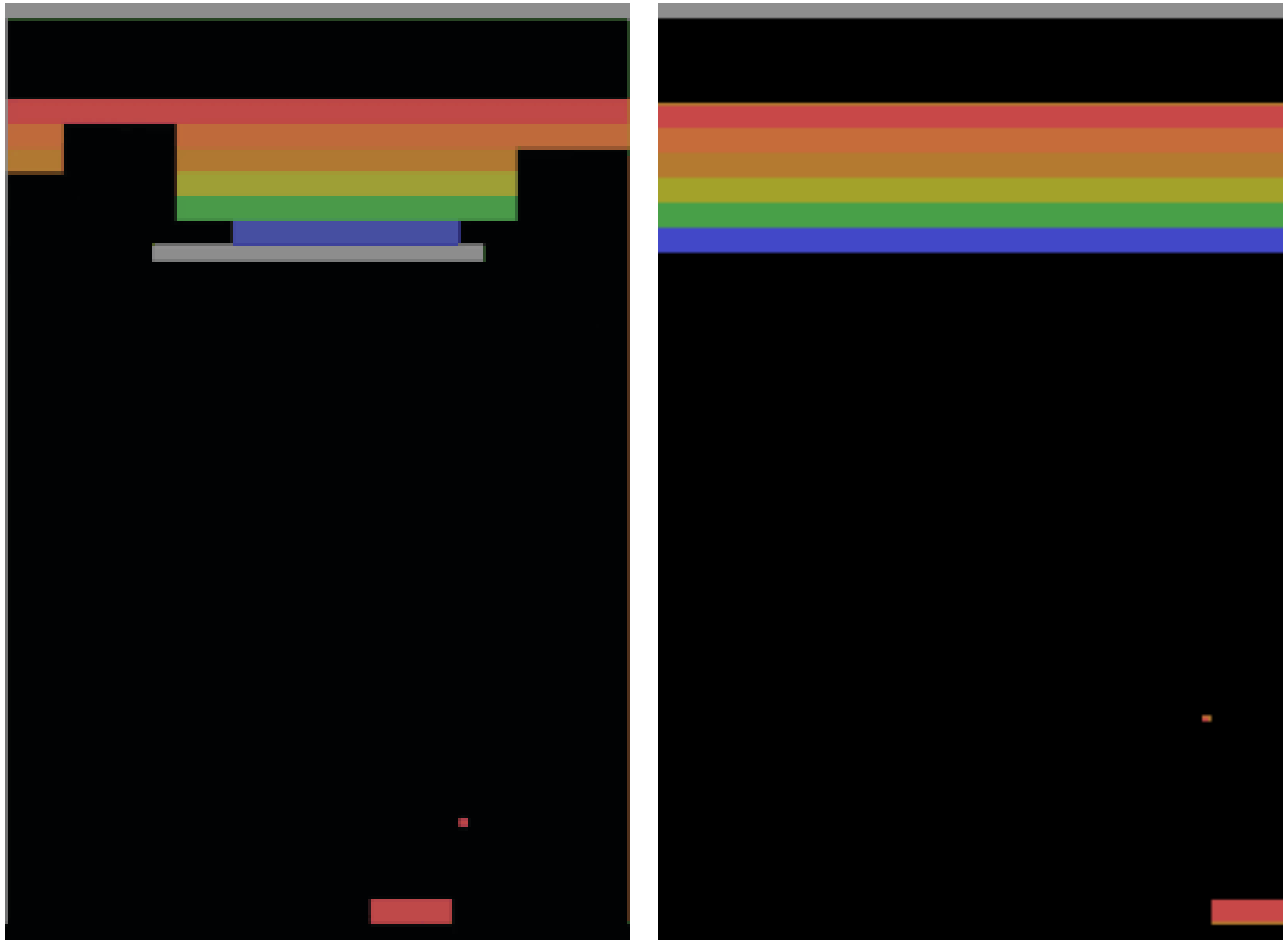
Improved Interpretability: Deep Learning meets Causal Inference



Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky Tom Silver David A. Mély Mohamed Eldawy Miguel Lázaro-Gredilla Xinghua Lou Nimrod Dorfman Szymon Sidor Scott Phoenix Dileep George

Improved Interpretability: Deep Learning meets Causal Inference



Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky Tom Silver David A. Mély Mohamed Eldawy Miguel Lázaro-Gredilla Xinghua Lou Nimrod Dorfman Szymon Sidor Scott Phoenix Dileep George

- Better modeling of cause-effect
- Better generalization through a conceptual representation

- Causal-Effect can leverage **better explanations** of systems

Let’s wrap-up 

- Interpretability of ML systems is not a new topic, but propelled by new findings from the DL community and efforts to safely translate this technology to (medical) applications. Complexity and Prevalence.
- The goal of interpretability is *NOT* to understand every part but to have enough information for the task at hand.
- Interpretability can leverage: auditability, trust, adoptability, understanding of system’s inner workings.
- Lot of work in visualization techniques; active area of research
- Benchmarking of interpretability methods is still an unexplored area.
- Likely a task-dependent combination of approaches will ultimately succeed. Tradeoff *succinct-yet-rich explanation*

Bibliography

- Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. http://doi.org/10.1109/ACCESS.2018.2870052
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. Retrieved from http://arxiv.org/abs/1810.03292
- Barocas, S., Friedler, S., Hardt, M., Kroll, J., Venka-Tasubramanian, S., & H. Wallach, H. (2018). The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15* (pp. 1721–1730). New York, New York, USA: ACM Press. http://doi.org/10.1145/2783258.2788613
- Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 598–617). IEEE. http://doi.org/10.1109/SP.2016.42
- Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning, (Ml), 1–13. http://doi.org/10.1016/j.intell.2013.05.008
- Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., & Cardoso, M. J. (2018). Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. Retrieved from http://arxiv.org/abs/1806.08640
- Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A. P., & Palmer, L. J. (2018). Producing radiologist-quality reports for interpretable artificial intelligence. Retrieved from http://arxiv.org/abs/1806.00340
- Gallego-Ortiz, C., & Martel, A. L. (2016). Interpreting extracted rules from ensemble of trees: Application to computer-aided diagnosis of breast MRI. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*. Retrieved from http://arxiv.org/abs/1606.08288
- Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, *278*(2), 563–77. http://doi.org/10.1148/radiol.2015151169
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). *Explaining Explanations: An Overview of Interpretability of Machine Learning*. Retrieved from https://arxiv.org/pdf/1806.00069.pdf
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. http://doi.org/10.1080/10618600.2014.907095
- Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine*, *38*(3), 50. http://doi.org/10.1609/aimag.v38i3.2741
- Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), Nd Web*. article.

Bibliography (contd.)

- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. JMLR.org. Retrieved from https://dl.acm.org/citation.cfm?id=3305518
- He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36. http://doi.org/10.1038/s41591-018-0307-0
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. http://doi.org/10.1038/s41568-018-0016-5
- Jiang, H., Kim, B., Guan, M., & Gupta, M. (2018). To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems* (pp. 5546–5557). inproceedings.
- Jungo, A., Meier, R., Ermis, E., Herrmann, E., & Reyes, M. (2018). Uncertainty-driven Sanity Check: Application to Postoperative Brain Tumor Cavity Segmentation. In *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)* (Vol. In Press). conference.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Retrieved from http://arxiv.org/abs/1711.11279
- Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017). Learning how to explain neural networks: PatternNet and PatternAttribution, 1–12. http://doi.org/10.1021/jm900403j
- Kleesiek, J., Petersen, J., Döring, M., Maier-Hein, K., Köthe, U., Wick, W., … Biller, A. (2016). Virtual Raters for Reproducible and Objective Assessments in Radiology. *Scientific Reports*, *6*, 25007. http://doi.org/10.1038/srep25007
- Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. http://doi.org/10.3389/fpsyg.2012.00223
- Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., … Walsh, S. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, *14*(12), 749–762. http://doi.org/10.1038/nrclinonc.2017.141
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. http://doi.org/10.1038/s41467-019-08987-4
- Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, *9*(3), 1350–1371. article. http://doi.org/10.1214/15-AOAS848
- Lipton, Z. C. (2016). The Mythos of Model Interpretability. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*. Retrieved from http://arxiv.org/abs/1606.03490
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017, December 1). A survey on deep learning in medical image analysis. *Medical Image Analysis*. Elsevier. http://doi.org/10.1016/j.media.2017.07.005
- Mahapatra, D., Bozorgtabar, B., Thiran, J., & Reyes, M. (2018). Efficient Active Learning for Image Classification and Segmentation using a Sample Selection and Conditional Generative Adversarial Network. In *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2018* (Vol. In Press). inproceedings.
- Maier-Hein, L., Ross, T., Gröhl, J., Glocker, B., Bodenstedt, S., Stock, C., … Maier-Hein, K. (2016). Crowd-Algorithm Collaboration for Large-Scale Endoscopic Image Annotation with Confidence (pp. 616–623). Springer, Cham. http://doi.org/10.1007/978-3-319-46723-8_71
- Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Retrieved from http://arxiv.org/abs/1706.07269

Bibliography (contd.)

- Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2574–2582). IEEE. http://doi.org/10.1109/CVPR.2016.282
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Interpretable machine learning: definitions, methods, and applications*. Retrieved from https://arxiv.org/pdf/1901.04592.pdf
- Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2018). Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation (pp. 655–663). Springer, Cham. http://doi.org/10.1007/978-3-030-00928-1_74
- Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 427–436). IEEE. http://doi.org/10.1109/CVPR.2015.7298640
- Parikh, R. B., Obermeyer, Z., & Navathe, A. S. (2019). Regulation of predictive analytics in medicine. *Science*, *363*(6429), 810–812. http://doi.org/10.1126/science.aaw0029
- Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C. A., & Reyes, M. (2018). Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation. *Medical Image Analysis*, *44*, 228–244. http://doi.org/10.1016/j.media.2017.12.009
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and Measuring Model Interpretability. Retrieved from http://arxiv.org/abs/1802.07810
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?. Explaining the Predictions of Any Classifier. Retrieved from http://arxiv.org/abs/1602.04938
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE. http://doi.org/10.1109/ICCV.2017.74
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Retrieved from http://arxiv.org/abs/1312.6034
- Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, & Fergus, Rob. (2013). Intriguing properties of neural networks. *Eprint arXiv:1312.6199*. Retrieved from http://adsabs.harvard.edu/abs/2013arXiv1312.6199S
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. http://doi.org/10.1038/s41591-018-0300-7
- Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 900–907). inproceedings.
- Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. http://doi.org/10.1371/journal.pmed.1002683

Resources

- iMIMIC websites
  - 2018 edition, check material section https://imimic.bitbucket.io/resources.html
  - 2019 edition (talks and accepted contributions to be posted)

  http://imimic-workshop.com

  - Awesome book from Christoph Molnar: https://christophm.github.io/interpretable-ml-book/

  - INNvestigate - https://github.com/albermax/innvestigate/

Hands-on session

- Use of INNvestigate on two datasets
- Apply methods seen during morning session

VISUM Summer School

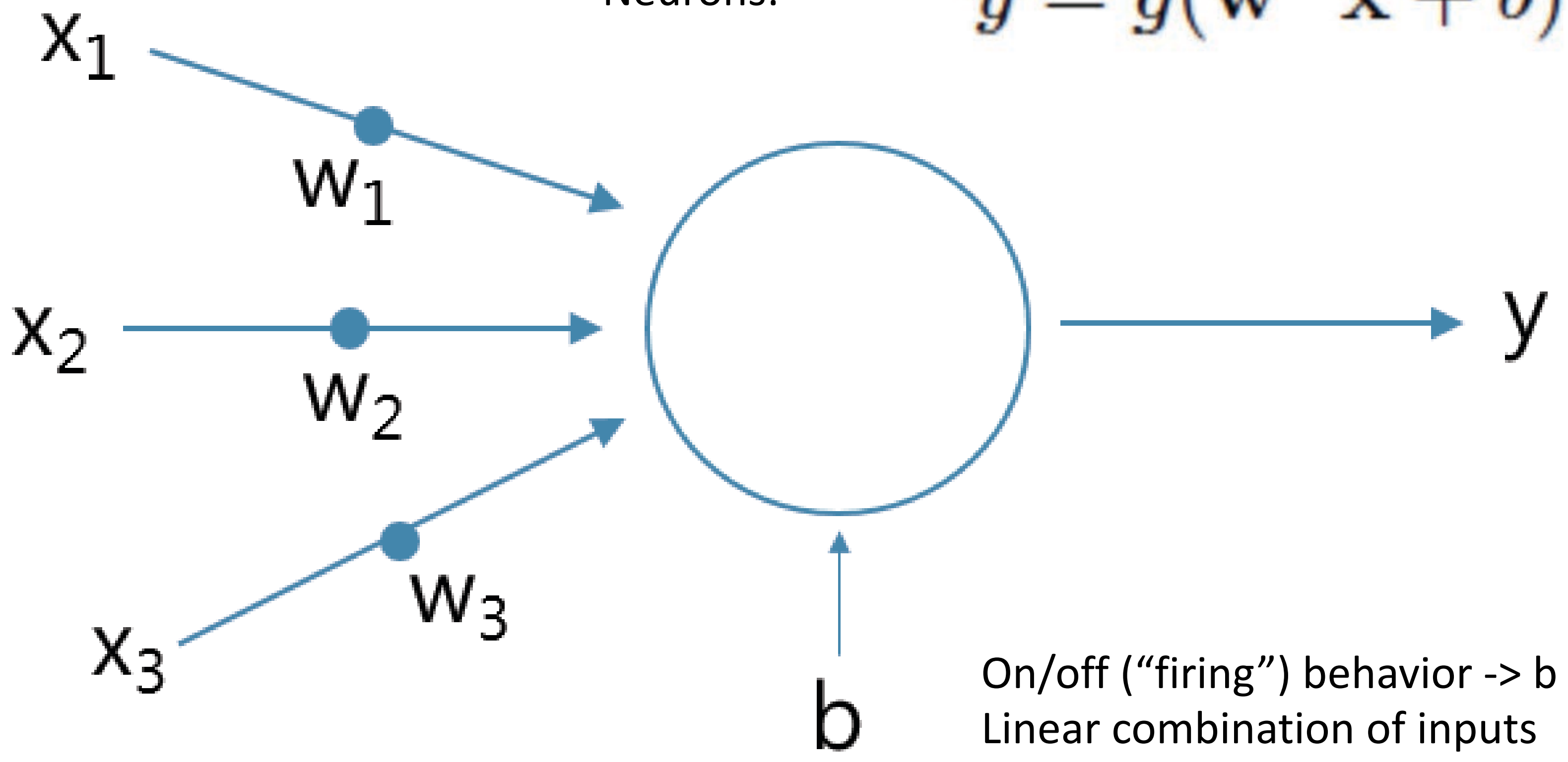
VISUM

VISion Understanding and Machine intelligence

..and a short quiz now

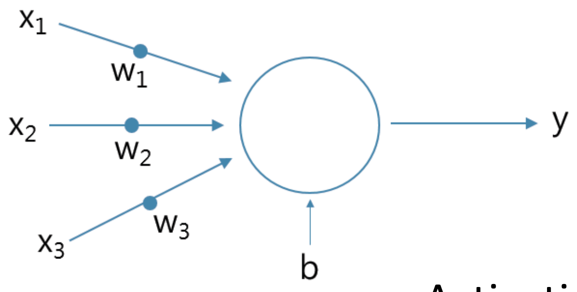
From neuron to neural nets

Neurons: $y = g(\mathbf{w}^\top \mathbf{x} + b)$



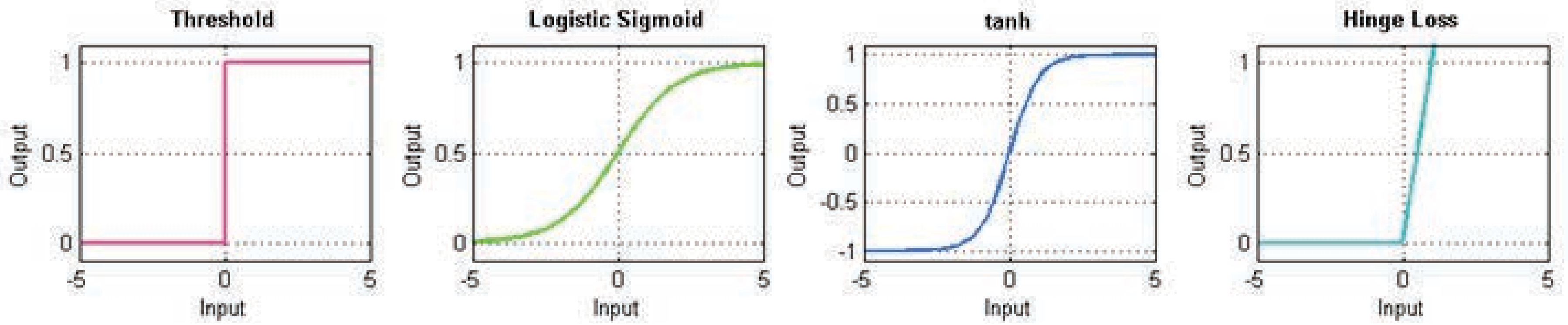
On/off ("firing") behavior -> b value
Linear combination of inputs

Neural Nets: Activation functions

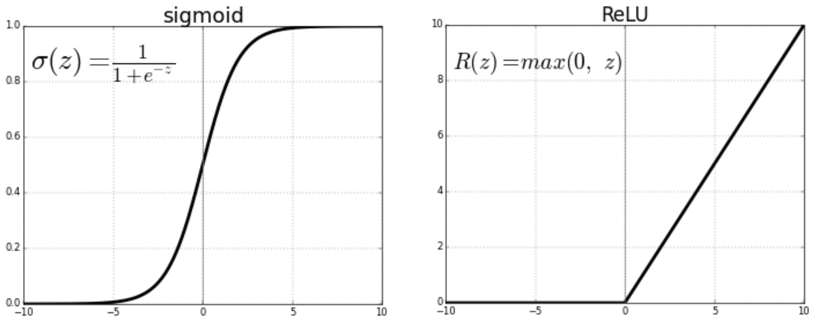


$$y = g(\mathbf{w}^\top \mathbf{x} + b)$$

Activation Function "g"

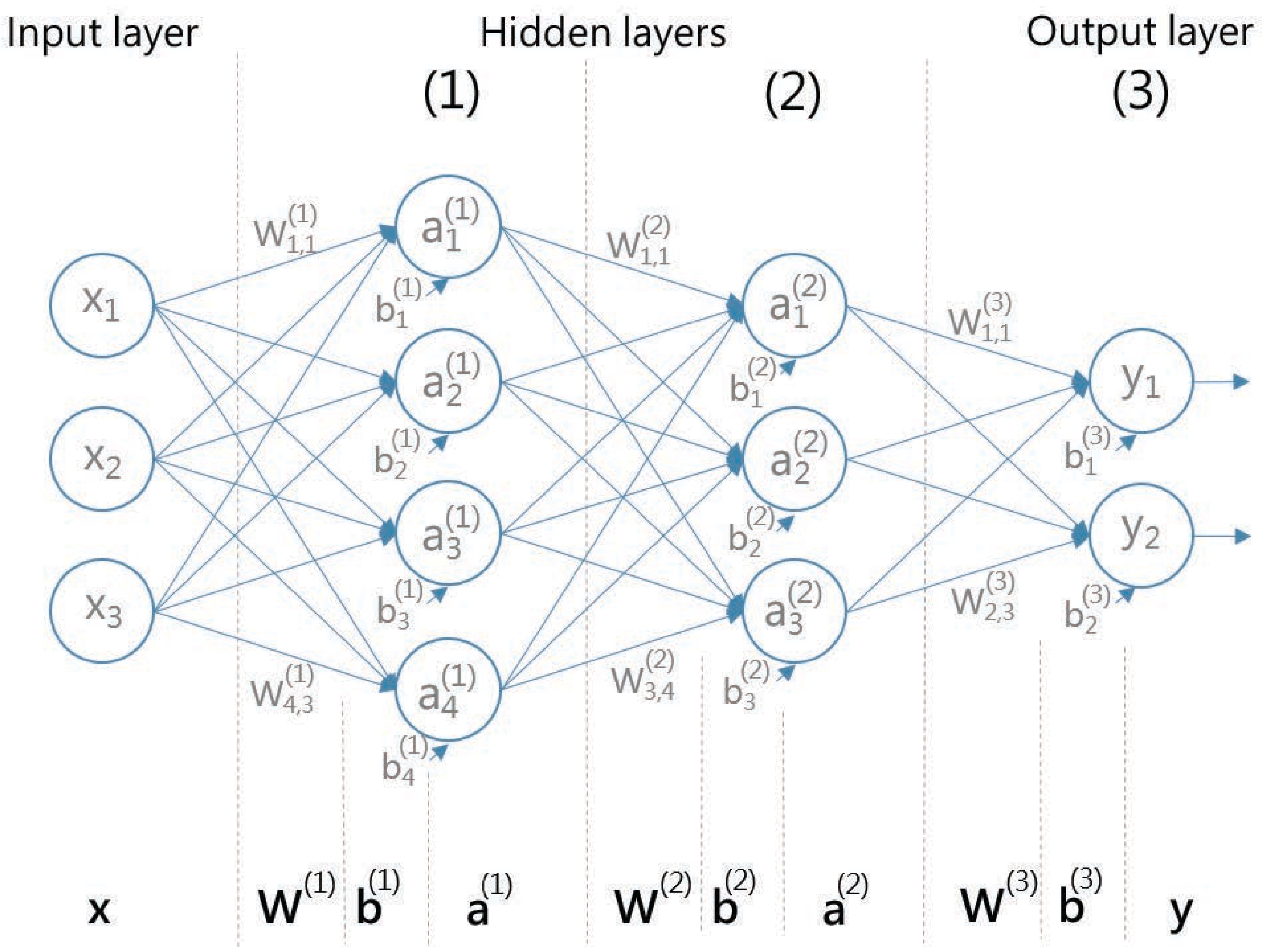


Most used: sigmoid and ReLu



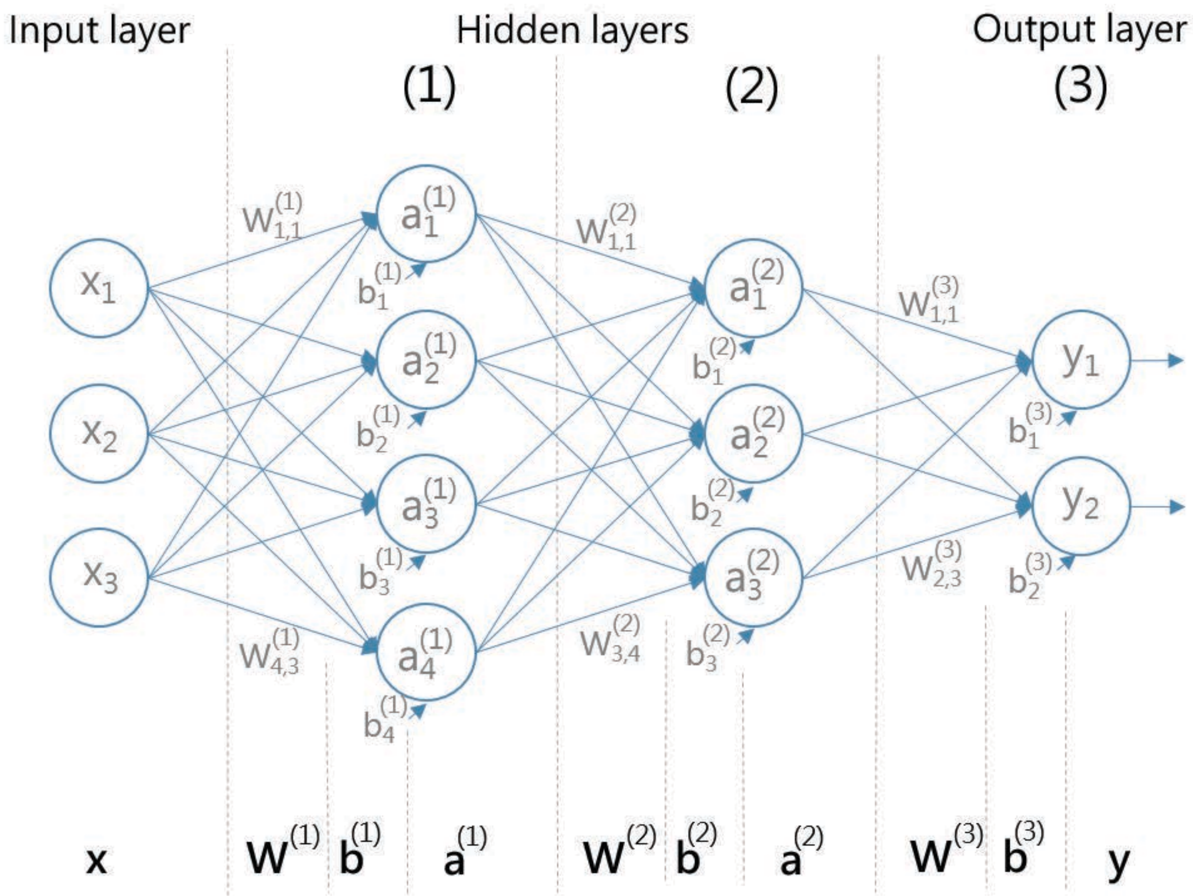
(We'll talk later about issues with sigmoid)

Neural Nets: Layers



Neural Nets: Math behind layers

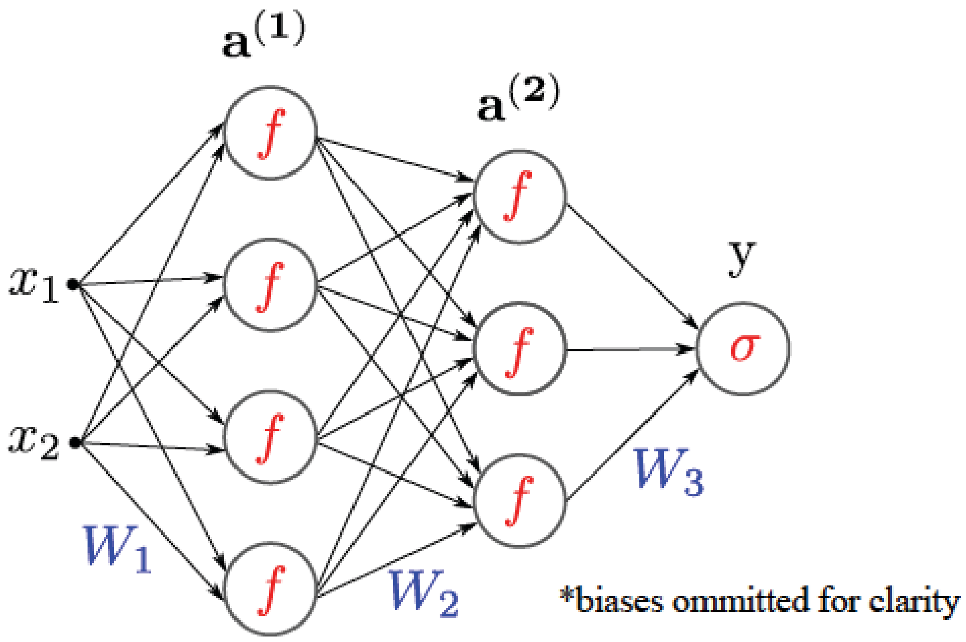
$$\mathbf{a}^{(1)} = g^{(1)}\left(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right) \quad \mathbf{a}^{(l)} = g^{(l)}\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}\right)$$



Multilayer Perceptron (MLP)

Characteristics and Training a MLP

- Adjust weights to minimize error -> supervised
- Backpropagation -> backward propagation of errors
- Gradient Descent
- Fully connected: output of a neuron connects to all downstream neurons



Training is an optimization process

- Gradient descent

Numerical gradient: slow :(, approximate :(, easy to write :) Analytic gradient: fast :), exact :), error-prone :(

In practice: Derive analytic gradient, check your implementation with numerical gradient

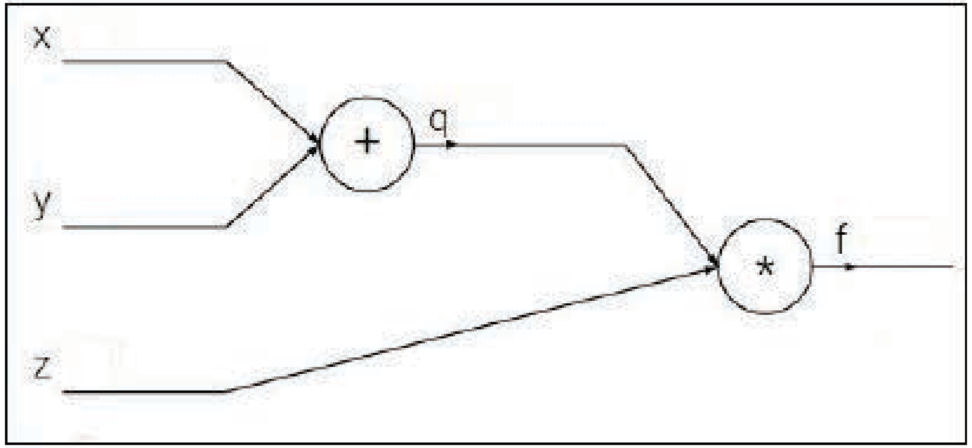
(Kingma & Ba 2015)

- Adam optimization is the most popular to date

Simple example Backpropagation credits: Fei-Fei CS231

Backpropagation: a simple example

$$f(x, y, z) = (x + y)z$$

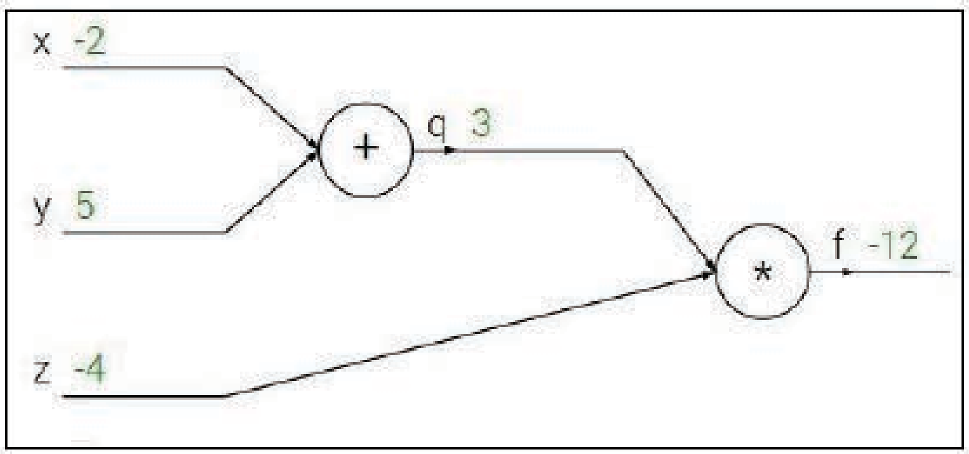


Simple example Backpropagation credits: Fei-Fei CS231

Backpropagation: a simple example

$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4



Simple example Backpropagation credits: Fei-Fei CS231

Backpropagation: a simple example

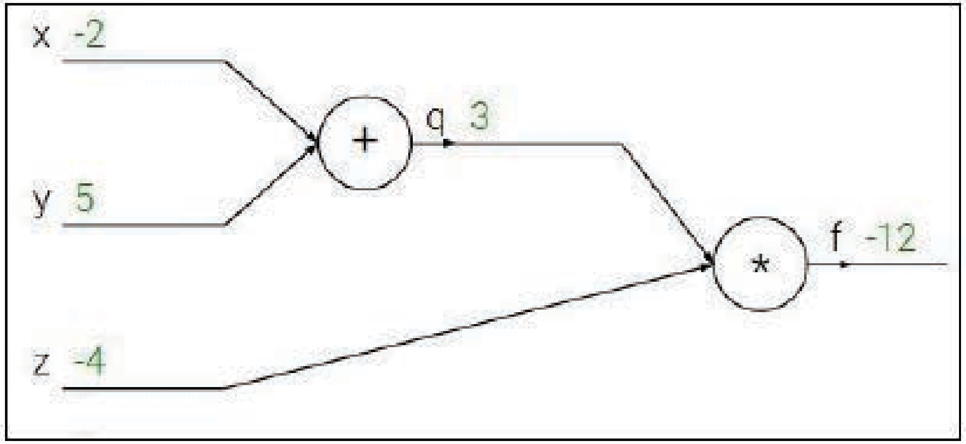
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Simple example Backpropagation credits: Fei-Fei CS231

Backpropagation: a simple example

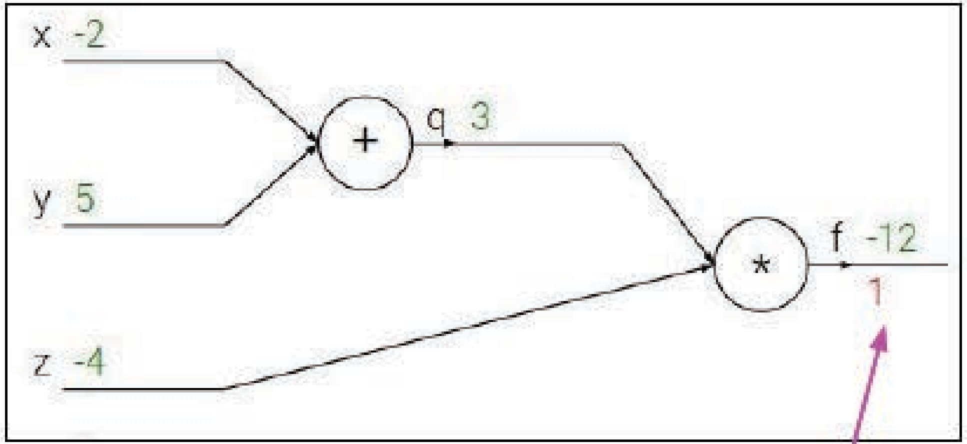
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$\frac{\partial f}{\partial f}$

Simple example Backpropagation credits: Fei-Fei CS231

Backpropagation: a simple example

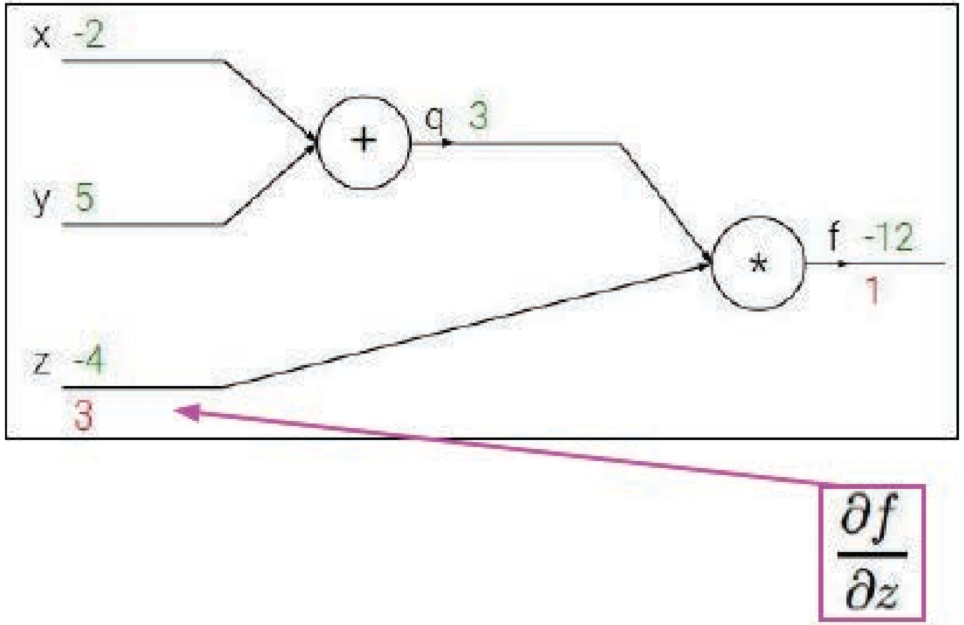
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Simple example Backpropagation credits: Fei-Fei CS231

Backpropagation: a simple example

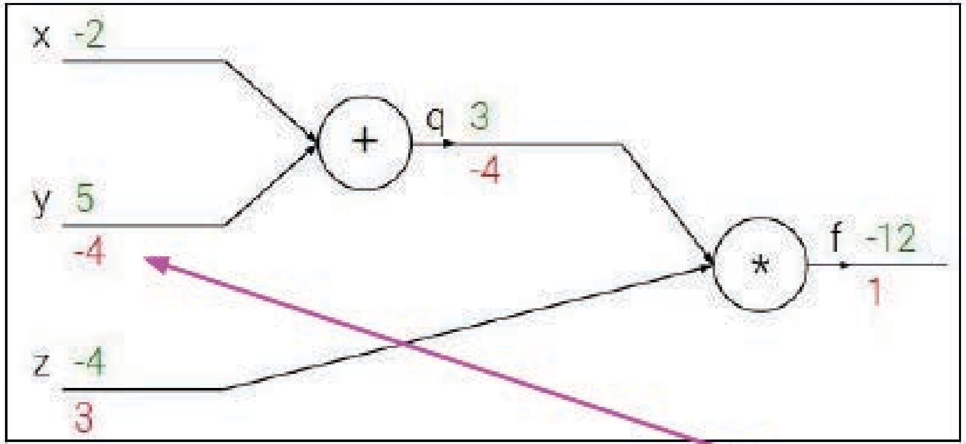
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



$\frac{\partial f}{\partial y}$

Chain rule:

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y}$$

Upstream gradient

Local gradient

Simple example Backpropagation credits: Fei-Fei CS231