From Machine Learning to Explainable AI

Assoc.Prof. Dr. Andreas Holzinger

Holzinger-Group, HCI-KDD, Institute for Medical Informatics/Statistics
Medical University Graz, Austria
&
Institute of interactive Systems and Data Science
Graz University of Technology, Austria

a.holzinger@hci-kdd.org
http://hci-kdd.org

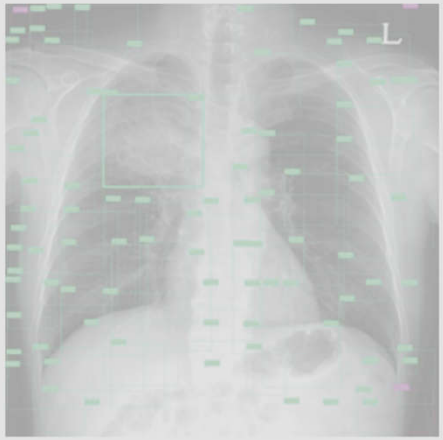
IEEE

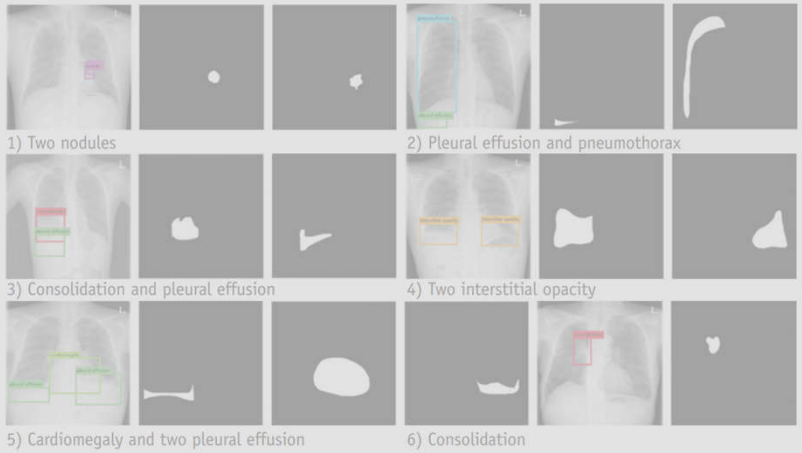
TECHNICKÁ UNIVERZITA V KOŠICIACH

DISA 2018

Houston, we have a problem …

Deep Learning is considered as “black-box” approach





June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo & Namkug Kim 2017. Deep learning in medical imaging: general overview. Korean journal of radiology, 18, (4), 570-584, doi:10.3348/kjr.2017.18.4.570.

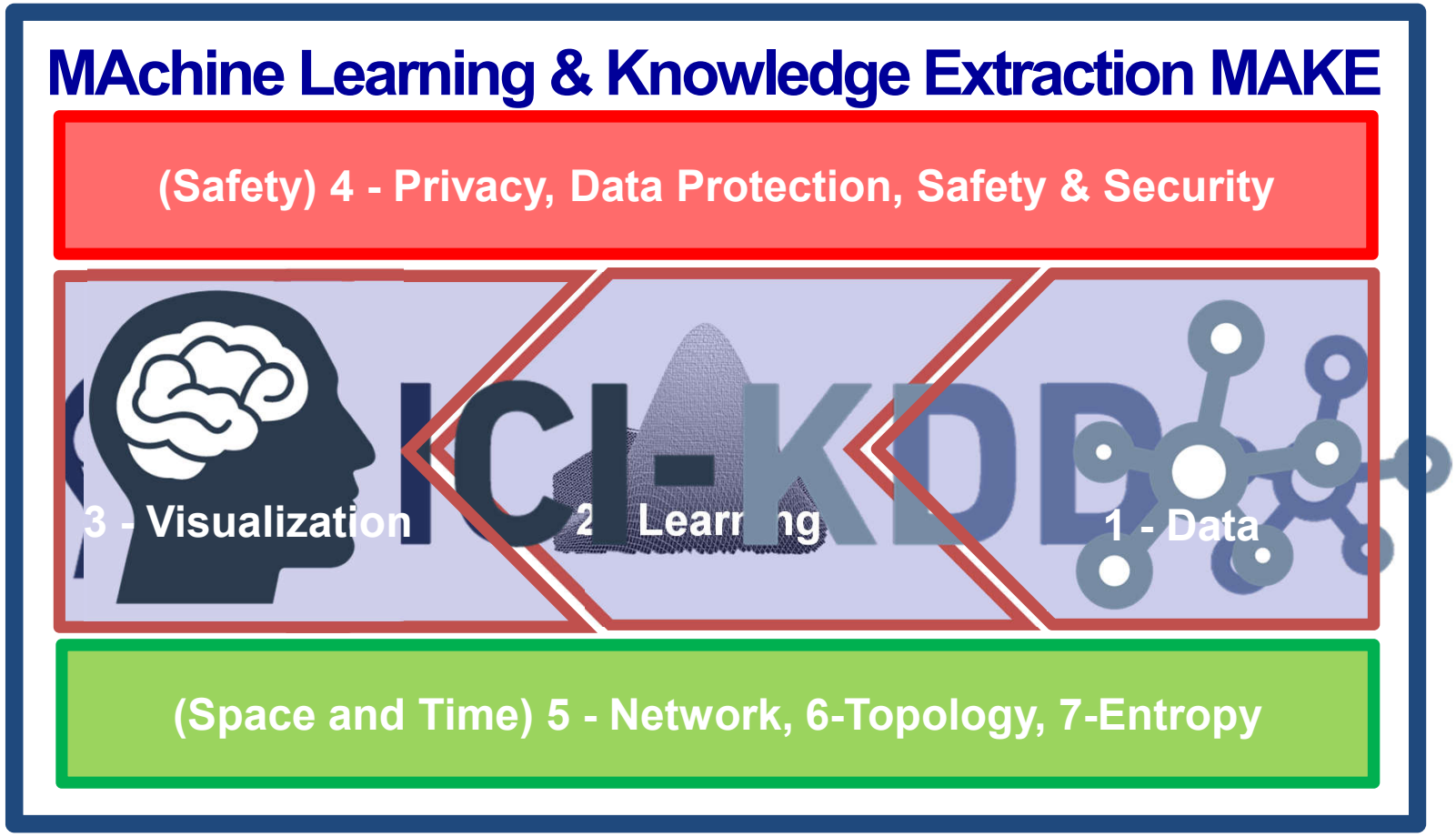
Agenda

- 01 HCI-KDD – integrative ML
- 02 Understanding Intelligence
- 03 Application Area: Health
- 04 automatic ML (aML)
- 05 interactive ML (iML)
- 06 towards explainable AI

01 What is the

approach?

- **ML is a very practical field –**
  algorithm development is at the core –
  however,
  successful ML needs a concerted effort of various topics …



Andreas Holzinger 2017. Introduction to Machine Learning and Knowledge Extraction (MAKE). *Machine Learning and Knowledge Extraction,* 1, (1), 1-20, doi:10.3390/make1010001.

http://www.bach-cantatas.com


http://hci-kdd.org/international-expert-network

CD-MAKE

Cross Domain Conference for Machine Learning and Knowledge Extraction
co-located with ARES 2018, Hamburg, Germany, August 27-30, 2018

About CD-MAKE | Call for Papers | Committee | Special Sessions | Authors Area | Venue & Registration | Call for Sponsors | Contact | Archive | ARES 2018

Welcome

"Augmenting Human Intelligence with Artificial Intelligence"

International IFIP Cross Domain (CD) Conference for Machine Learning & Knowledge Extraction (MAKE)
CD-MAKE 2018

machine learning and knowledge extraction

MDPI Academic Open Access Publishing since 1996

General Information

About CD-MAKE

ARES 2018

CD-MAKE is held in conjunction with the International Conference on Availability,

02 Understanding Intelligence

“Solve intelligence – then solve everything else”

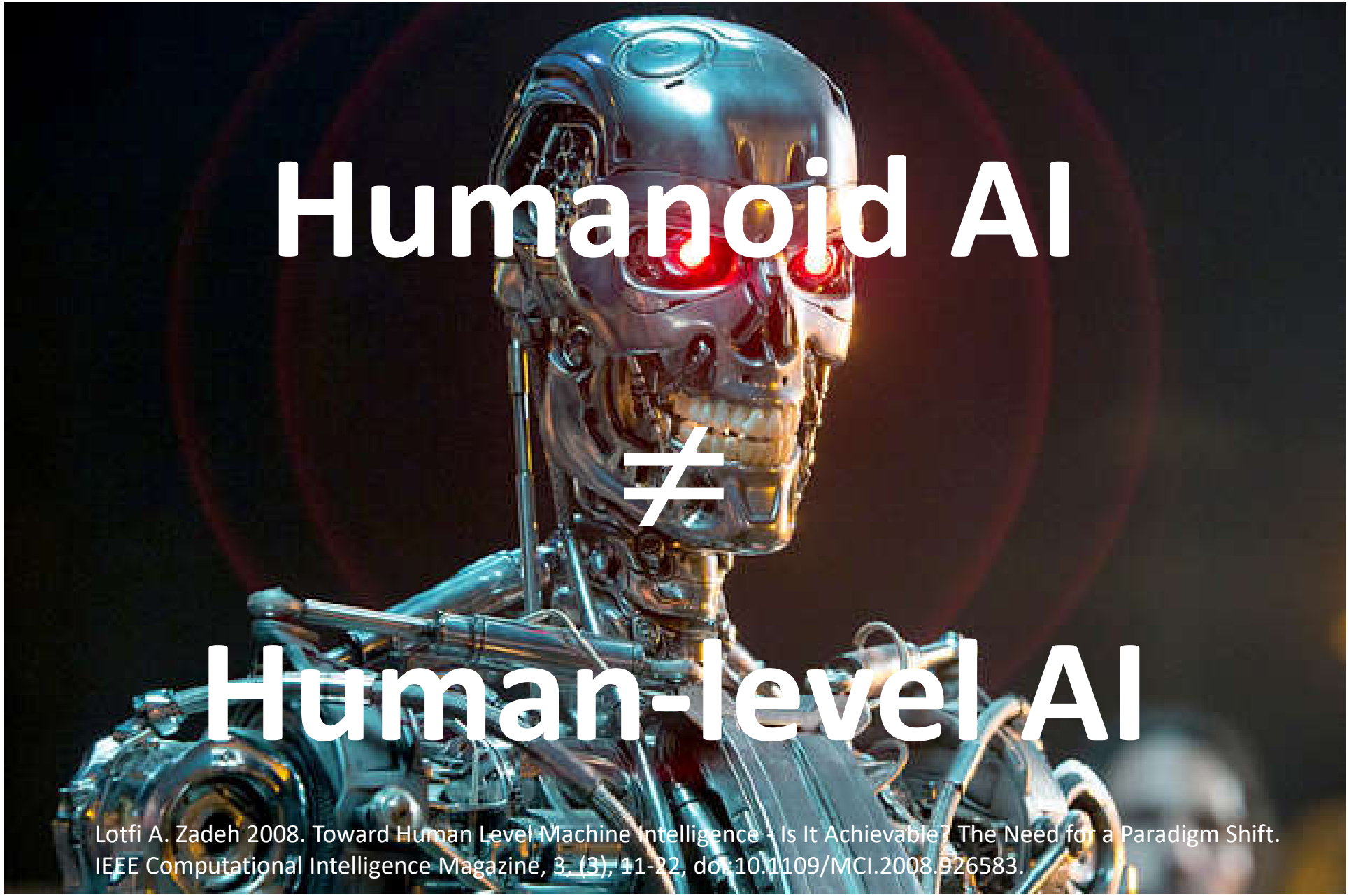
Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

https://youtu.be/XAbLn66iHcQ?t=1h28m54s

Not our Goal: Humanoid AI



Lotfi A. Zadeh 2008. Toward Human Level Machine Intelligence - Is It Achievable? The Need for a Paradigm Shift. IEEE Computational Intelligence Magazine, 3, (3), 11-22, doi: 10.1109/MCI.2008.926583.

Understanding Context

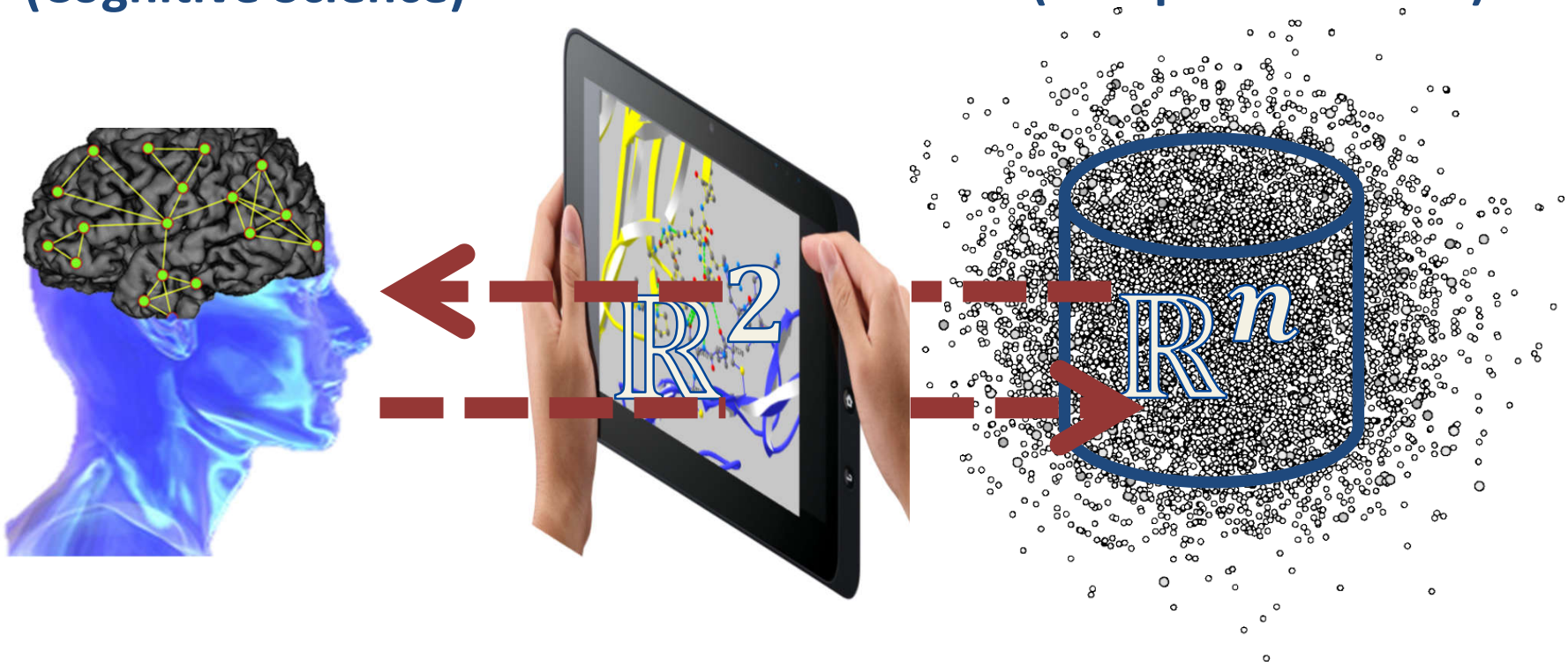
- 1) learn from prior data
- 2) extract knowledge
- 2) generalize, i.e. guessing where a probability mass function concentrates
- 4) fight the curse of dimensionality
- 5) disentangle **underlying explanatory factors of data**, i.e.
- 6) **understand** the data in the **context** of an application domain

Augment human Intelligence with artificial intelligence

HCI-KDD

Human intelligence (Cognitive Science)

Artificial intelligence (Computer Science)



Holzinger, A. (2013). Human–Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Lecture Notes in Computer Science 8127 (pp. 319-328)

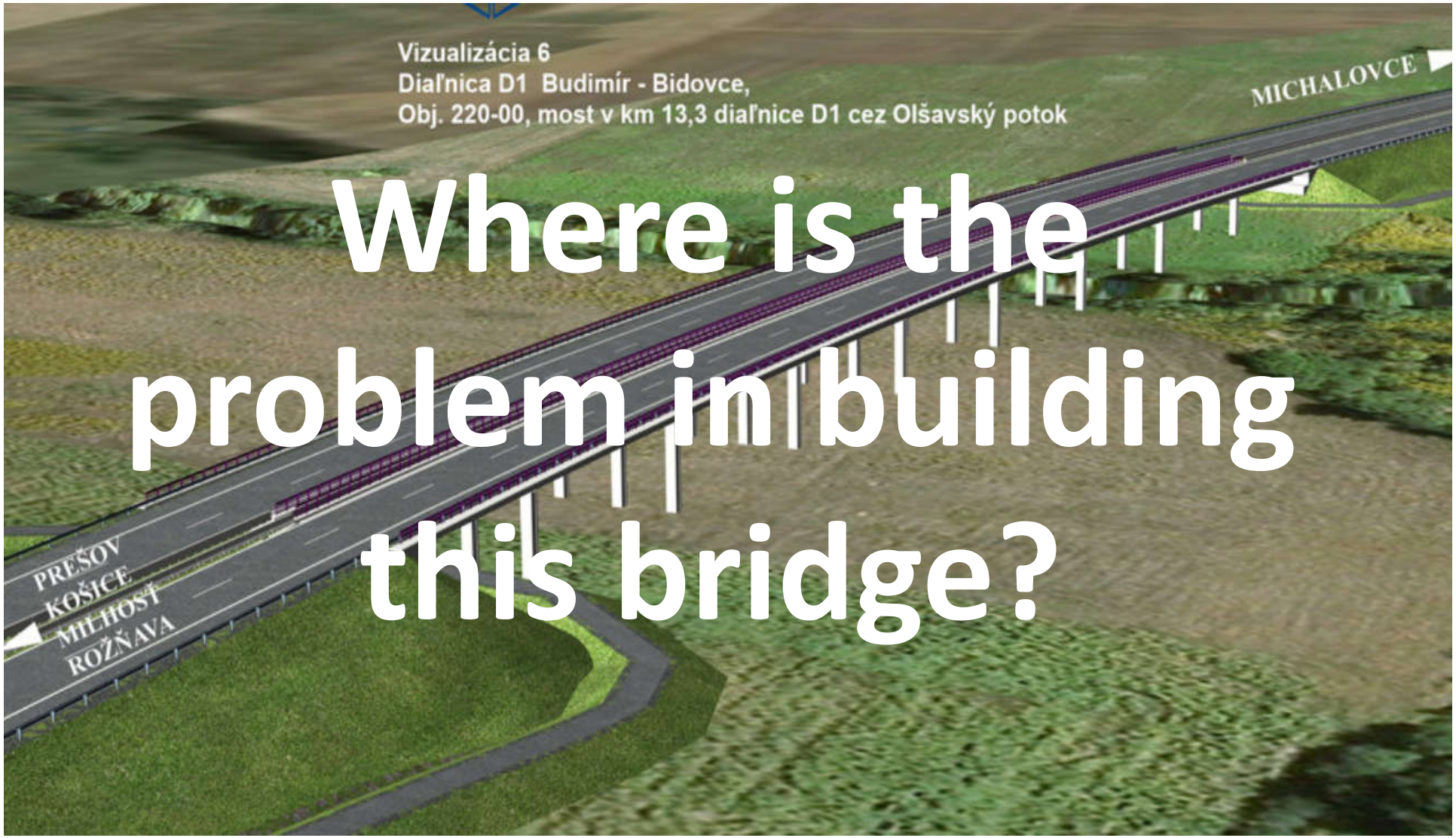
03 Application Area Health Informatics

Why is this application area complex ?

In medicine we have two different worlds …

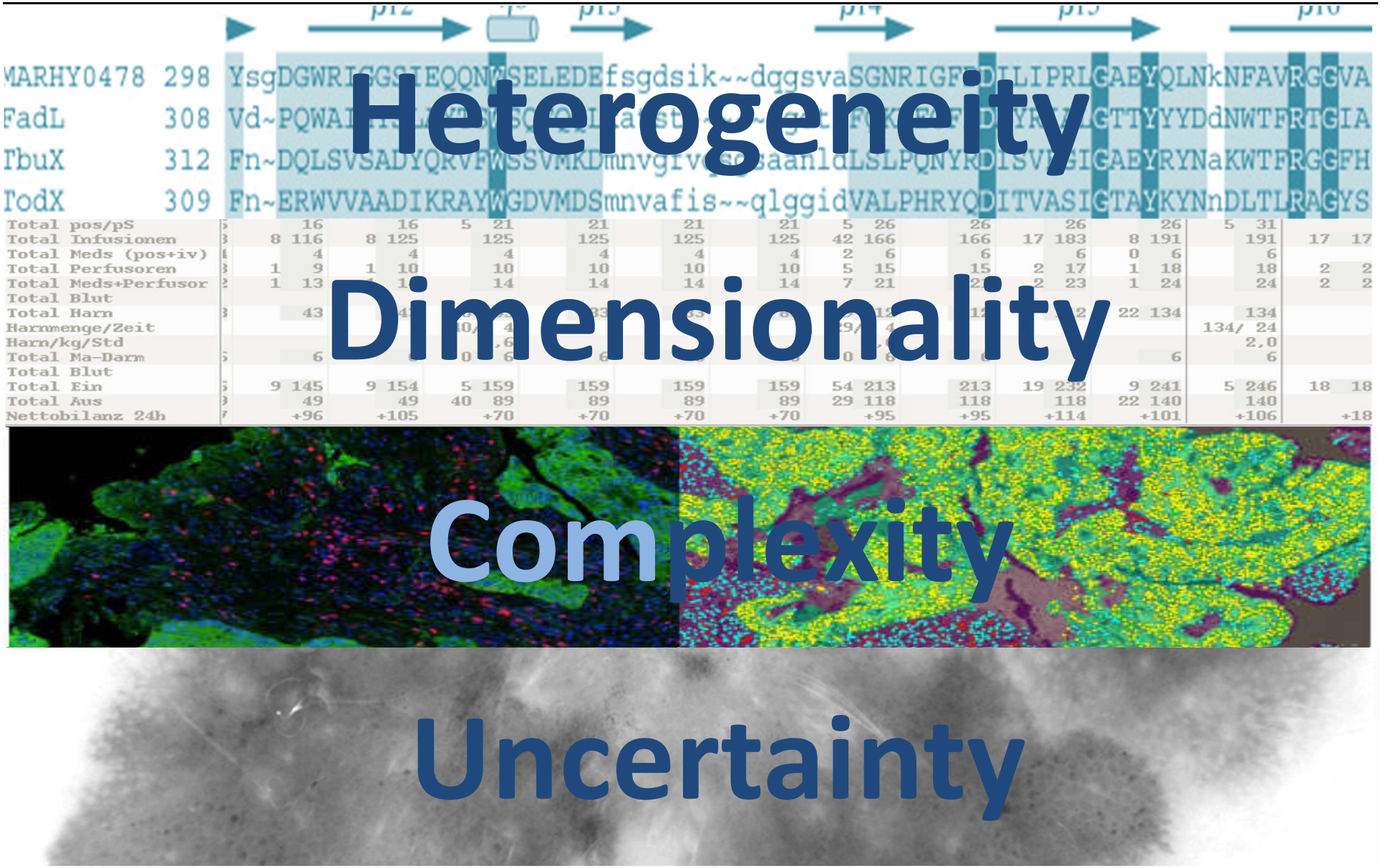
Our central hypothesis: Information may bridge this gap

Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*



Where is the problem in building this bridge?

Main problems …



Holzinger, A., Dehmer, M. & Jurisica, I. 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics, 15, (S6), I1.

Probabilistic Information p(x)

Probability theory is nothing but common sense reduced to calculation …

Pierre Simon de Laplace (1749-1827)

$$\mathcal{D} = x_{1:n} = \{x_1, x_2, \ldots, x_n\}$$

$$p(\mathcal{D}|\theta)$$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

$$posterior = \frac{likelihood * prior}{evidence}$$

The "inverse probability" allows to learn, to infer unknowns and to make predictions

d … *data*

h … *hypotheses*

$\mathcal{H}$ … $\{H_1, H_2, \ldots, H_n\}$ $\forall\, h, d$ …

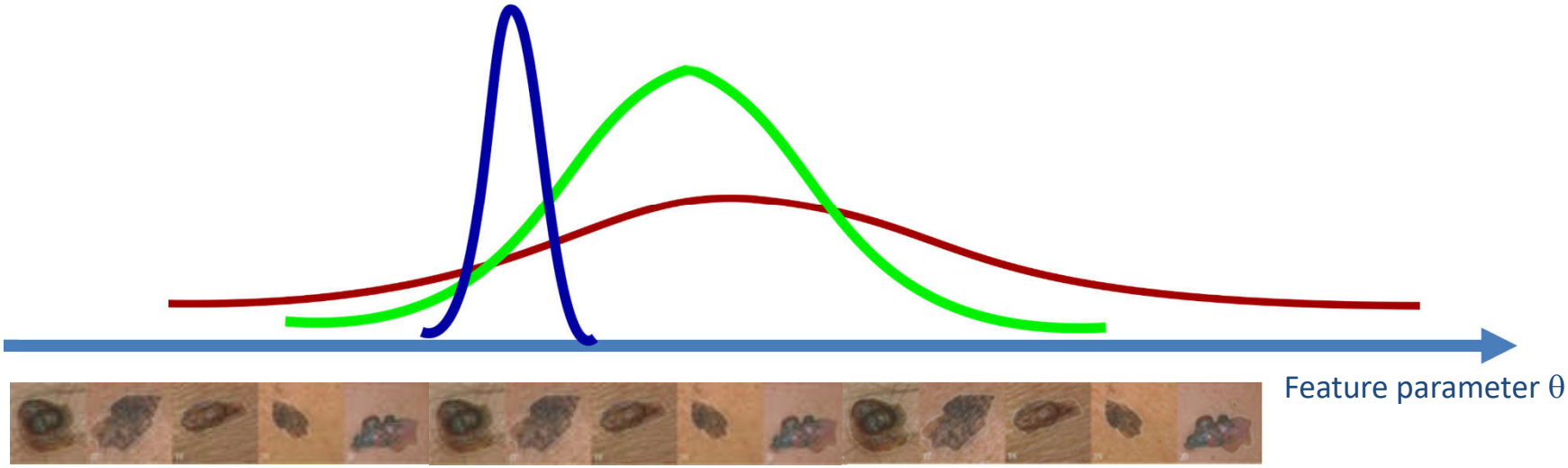
Likelihood

Prior Probability

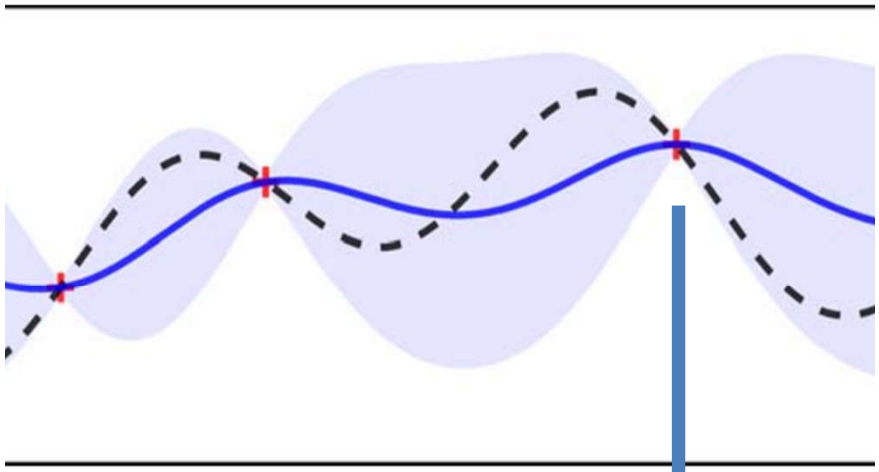
$$p(h|d) = \frac{p(d|h) * p(h)}{\sum_{h \in H} p(d|h')\, p(h')}$$

Posterior Probability

Problem in $\mathbb{R}^n \rightarrow$ complex



Fully automatic → Goal: Taking the human out of the loop

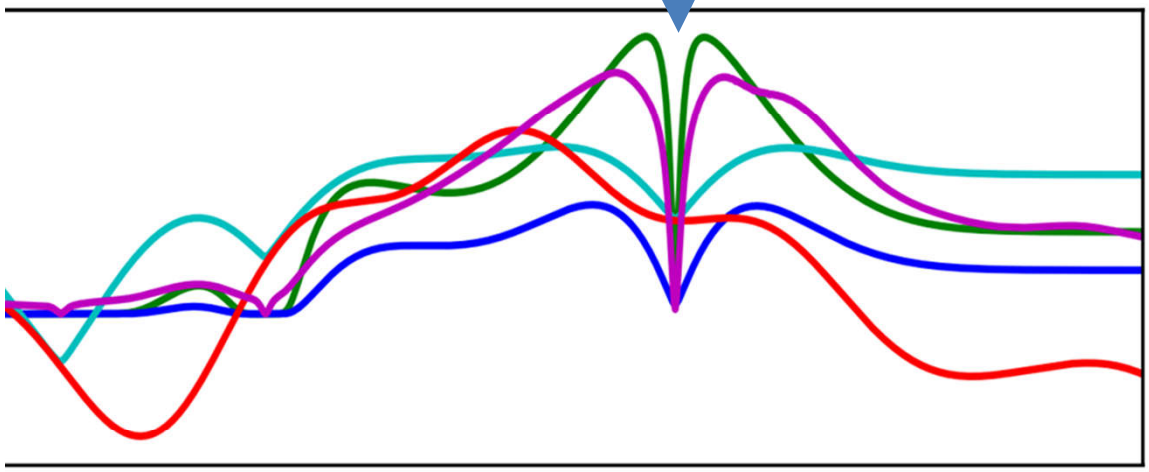


Algorithm 1 Bayesian optimization

1: **for** $n = 1, 2, \ldots$ **do**
2: select new $\mathbf{x}_{n+1}$ by optimizing acquisition function α

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}; \mathcal{D}_n)$$

3: query objective function to obtain y_{n+1}
4: augment data $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (\mathbf{x}_{n+1}, y_{n+1})\}$
5: update statistical model
6: **end for**



- PI — Probability of Improvement
- EI — Expected Improvement
- UCB — Upper Confidence Bound
- TS — Thompson Sampling
- PES — Predictive Entropy Search

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. 2016. **Taking the human out of the loop:** A review of Bayesian optimization. *Proceedings of the IEEE,* 104, (1), 148-175, doi:10.1109/JPROC.2015.2494218.

04 aML best practice examples

Recommender Systems

amazon.co.uk Try Prime

All | glass cutter circular

Shop by Department | Your Amazon.co.uk | Today's Deals | Gift Cards & Top Up | Sell | Help

Amazon.co.uk | Today's Deals | Warehouse Deals | Outlet | Subscribe & Save | Vouchers | Amazon Family | Amazon Prime | Amazon Video | Amazon Student | Mobile Apps

Showing results for "glass cutter circular"

Show results for

DIY & Tools >
- Glass Cutters
- Cold Chisels
- Power Tools

Sports & Outdoors >
- Compasses

+ See All 13

Refine by

Delivery Opt
- Prime
- Free UK De

Brand
- sourcingm
- SODIAL(R

Silverline 101228 Circular Glass Cutter with 65-300 mm Diameter 10 Oct 2014

by Silverline

£7.81 ~~£10.02~~ Prime

Get it by Tomorrow, Sep 5

Eligible for FREE UK Delivery

More buying choices
£6.40 new (22 offers)

42

DIY & Tools: See all 162 items

Highlander 3 Hole Thinsulate Balaclava

by Highlander

£1.99 - £7.00 Prime

More buying choices
£1.99 new (5 offers)

163

Sports & Outdoors: See all 5,918 items

Sanwood® Outdoor Motorcycle Cycling Ski Neck Protecting Lycra Balaclava Full Face Mask

by Phoenix B2C UK

£1.74 - £3.57

More buying choices
£0.01 new (4 offers)

73

Sports & Outdoors: See all 5,918 items

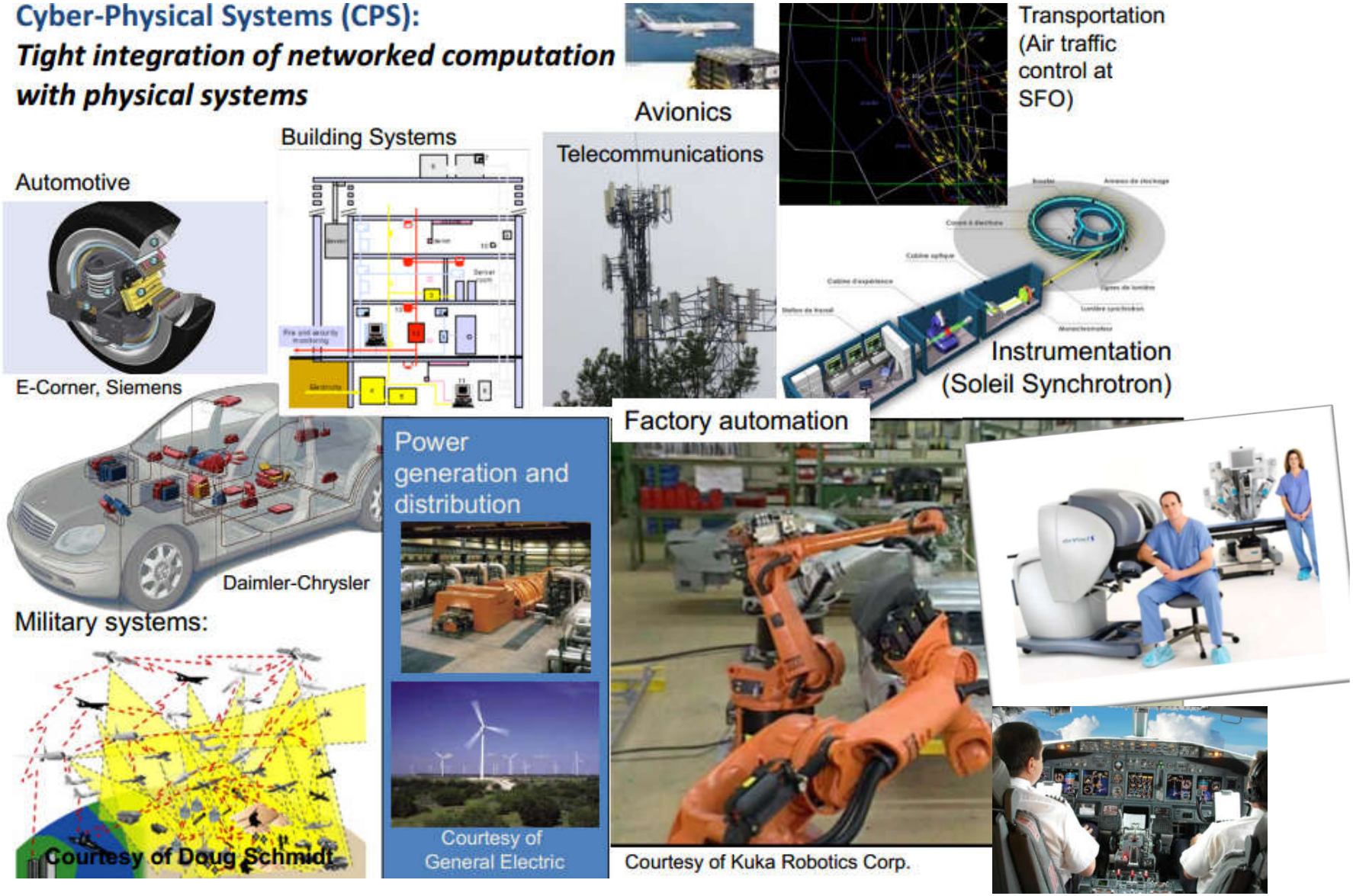
Fully automatic autonomous vehicles (“Google car”)

Guizzo, E. 2011. How google’s self-driving car works. IEEE Spectrum Online, 10, 18.

Open: Why did the car do that? Who is responsible?

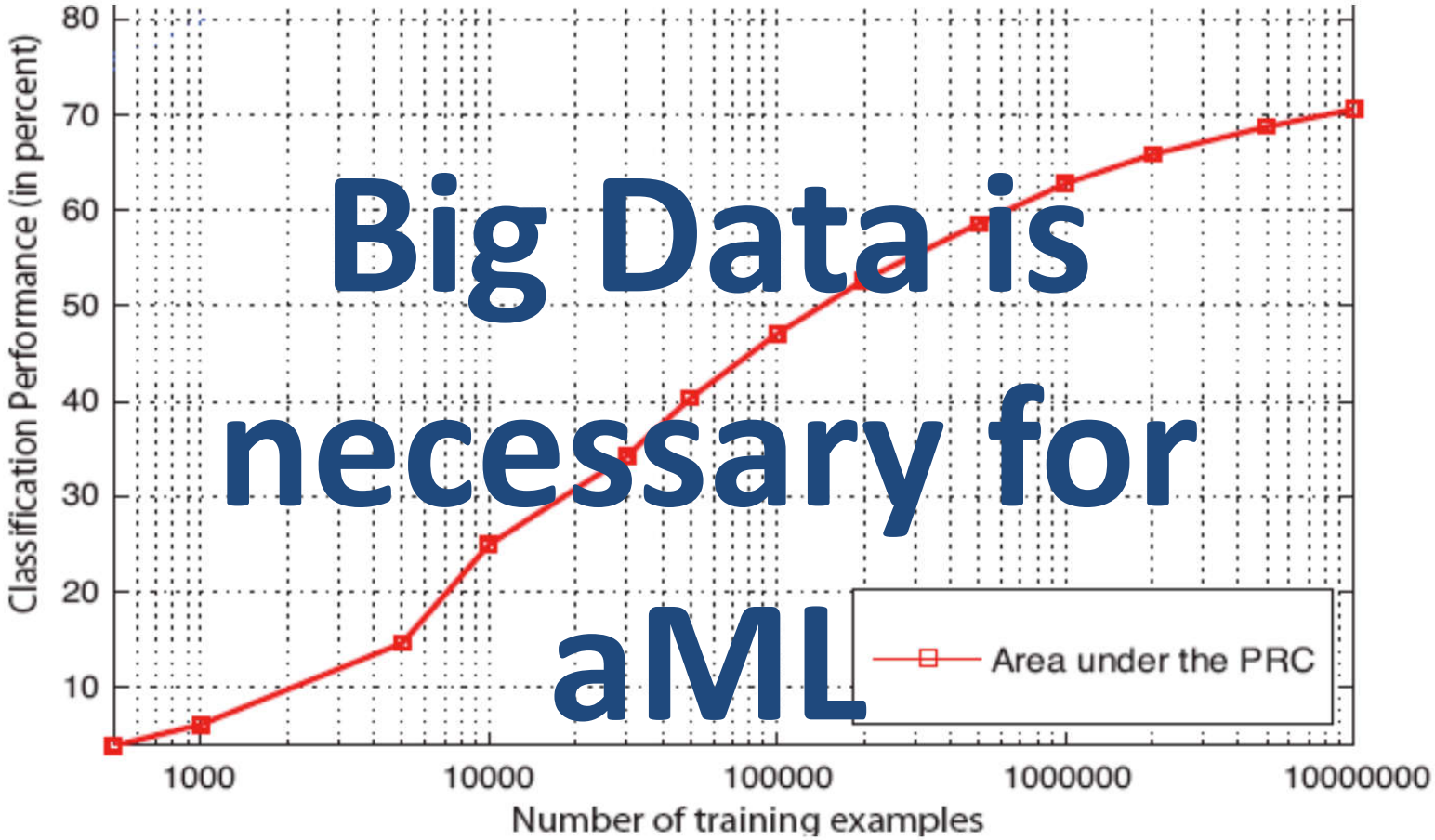
Image Source: http://www.businessinsider.de/who-is-responsible-when-a-driverless-car-crashes-2016-2?r=US&IR=T

… and thousands of industrial aML applications …



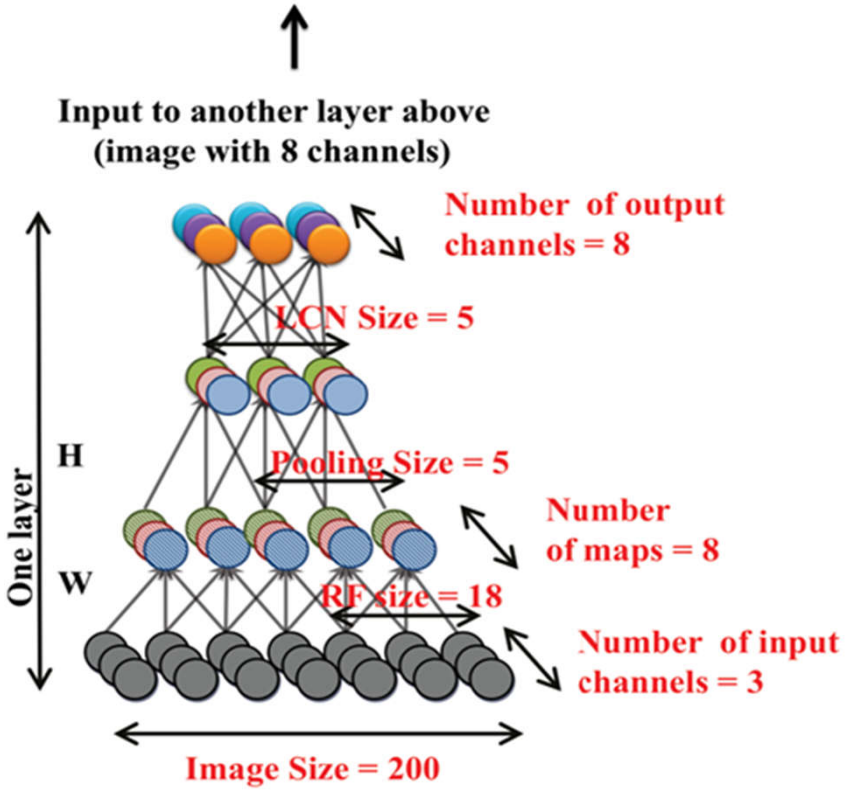
Seshia, S. A., Juniwal, G., Sadigh, D., Donze, A., Li, W., Jensen, J. C., Jin, X., Deshmukh, J., Lee, E. & Sastry, S. 2015. Verification by, for, and of Humans: Formal Methods for Cyber-Physical Systems and Beyond. Illinois ECE Colloquium.

Big Data is necessary for aML !



Sonnenburg, S., Rätsch, G., Schäfer, C. & Schölkopf, B. 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

10 million 200 x 200 px images downloaded from Web



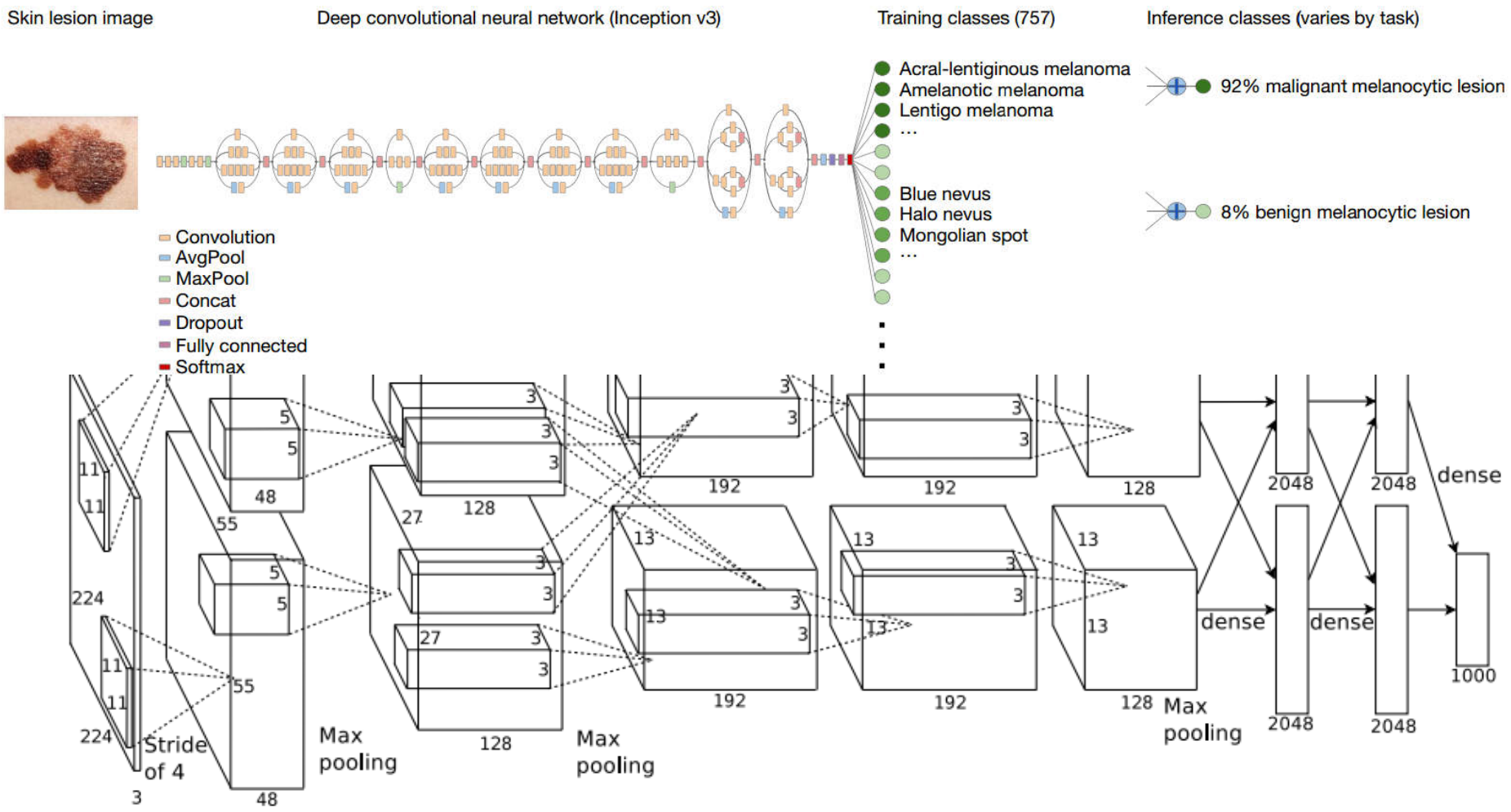
$$x^* = \arg\min_x f(x; W, H), \text{ subject to } ||x||_2 = 1.$$

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J. & Ng, A. Y. 2011. Building high-level features using large scale unsupervised learning. arXiv preprint arXiv:1112.6209.

Le, Q. V. 2013. Building high-level features using large scale unsupervised learning. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing ICASSP.* IEEE. 8595-8598, doi:10.1109/ICASSP.2013.6639343.

1,28 million images …

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



Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

- Sometimes we **do not have "big data"**, where aML-algorithms benefit.
- Sometimes we have
  - **Small amount of data sets**
  - Rare Events – **no training samples**
  - **NP-hard problems, e.g.**
    - Subspace Clustering,
    - k-Anonymization,
    - Protein-Folding, …

Sometimes we (still) need a human-in-the-loop

05 iML

- iML := algorithms which interact with agents*) and can optimize their learning behaviour through this interaction

***) where the agents can be human**

Holzinger, A. 2016. Interactive Machine Learning (iML). Informatik Spektrum, 39, (1), 64-68, doi:10.1007/s00287-015-0941-6.

Sometimes we need a doctor-in-the-loop

Image Source: 10 Ways Technology is Changing Healthcare http://newhealthypost.com Posted online on April 22, 2018

A group of experts-in-the-loop

Image Source: Cisco (2008). Cisco Health Presence Trial at Aberdeen Royal Infirmary in Scotland

A crowd of people-in-the-loop

Image is in the public domain

- **Humans can generalize even from few examples …**
  - They learn relevant representations
  - Can disentangle the explanatory factors
  - Find the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

Even Children can make inferences from little data …

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

Adversarial examples: https://imgur.com/a/K4RWn

See also: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Gamaleldin F. Elsayed*
Google Brain
gamaleldin.elsayed@gmail.com

Shreya Shankar
Stanford University

Brian Cheung
UC Berkeley

Nicolas Papernot
Pennsylvania State University

Alex Kurakin
Google Brain

Ian Goodfellow
Google Brain

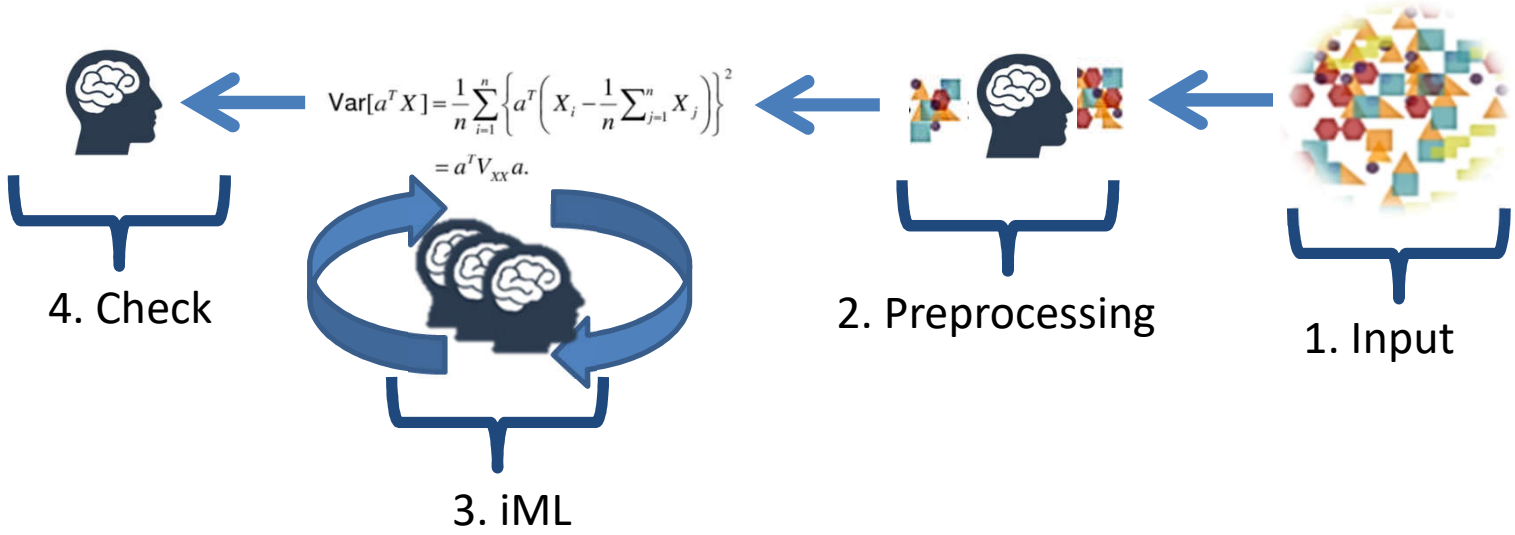
Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

Abstract

Machine learning models are vulnerable to **adversarial examples**: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein 2018. Adversarial Examples that Fool both Human and Computer Vision. arXiv:1802.08195.

Interactive Machine Learning: Human is seen as an agent involved in the actual learning phase, step-by-step influencing measures such as distance, cost functions …



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, doi:10.1007/s40708-016-0042-6.

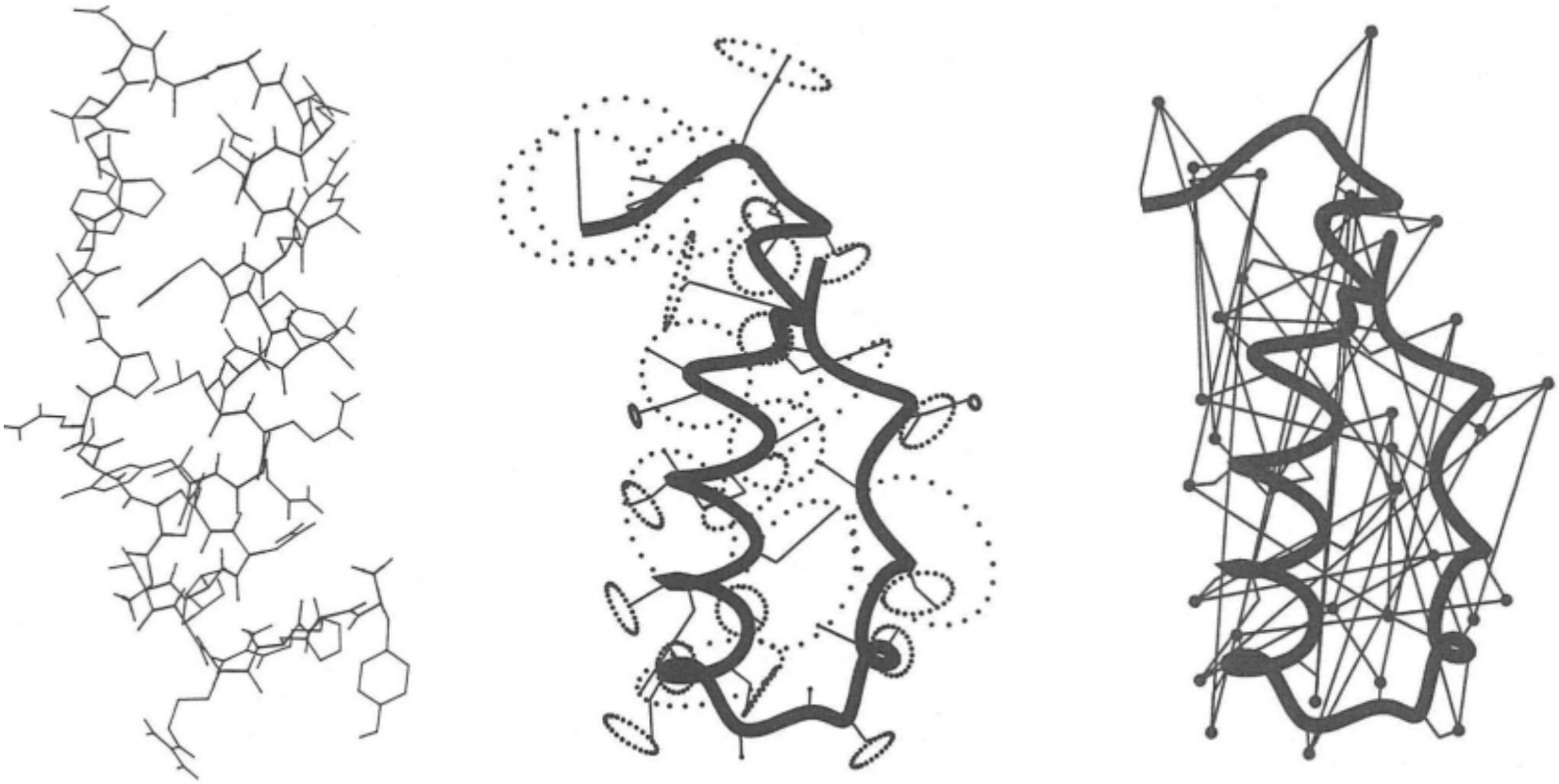
- Example 1: Subspace Clustering
- Example 2: k-Anonymization
- Example 3: Protein Design

Hund, M., Böhm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D. A., Majnaric, L. & Holzinger, A. 2016. Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. Brain Informatics, 1-15, doi:10.1007/s40708-016-0043-5.

Kieseberg, P., Malle, B., Fruehwirt, P., Weippl, E. & Holzinger, A. 2016. A tamper-proof audit and control system for the doctor in the loop. Brain Informatics, 3, (4), 269–279, doi:10.1007/s40708-016-0046-2.

Lee, S. & Holzinger, A. 2016. Knowledge Discovery from Complex High Dimensional Data. In: Michaelis, S., Piatkowski, N. & Stolpe, M. (eds.) Solving Large Scale Learning Tasks. Challenges and Algorithms, Lecture Notes in Artificial Intelligence LNAI 9580. Springer, pp. 148-167, doi:10.1007/978-3-319-41706-6_7.

Example: Protein Folding is a TSP



Bohr, H. & Brunak, S. 1989. A travelling salesman approach to protein conformation. Complex Systems, 3, 9-28

Open the black box to influence the algorithm …

HCI-KDD

http://hci-kdd.org/projects/iml-experiment

Input : ProblemSize, m, β, ρ, σ, q_0
Output: $Pbest$
$Pbest \leftarrow$ CreateHeuristicSolution(ProblemSize);
$Pbest_{cost} \leftarrow$ Cost($Pbest$);
$Pheromone_{init} \leftarrow \frac{1.0}{ProblemSize \times Pbest_{cost}}$;
$Pheromone \leftarrow$ InitializePheromone($Pheromone_{init}$);
while $\neg StopCondition()$ **do**
  for $i = 1$ to m **do**
    $S_i \leftarrow$ ConstructSolution(Pheromone, ProblemSize, β, q_0);
    $Si_{cost} \leftarrow$ Cost(S_i);
    if $Si_{cost} \leq$ Pbest$_{cost}$ **then**
      $Pbest_{cost} \leftarrow Si_{cost}$;
      $Pbest \leftarrow S_i$;
    end
    LocalUpdateAndDecayPheromone(Pheromone, S_i, Si_{cost}, ρ);
  end
  GlobalUpdateAndDecayPheromone(Pheromone, $Pbest$, $Pbest_{cost}$, ρ);
  while $isUserInteraction()$ **do**
    GlobalAddAndRemovePheromone(Pheromone, $Pbest$, $Pbest_{cost}$, ρ);
  end
end
return P_{best};

Holzinger, A., Plass, M., Holzinger, K., Crisan, G., Pintea, C. & Palade, V. 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. Springer Lecture Notes in Computer Science LNCS 9817. 81-95, doi:10.1007/978-3-319-45507-56.

Influence the Pheromone Level by human intuition

HCI-KDD

$$p_{ij} = \frac{[\tau_{ij}]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum_{l \in J_i^k} [\tau(t)]^{\alpha} \cdot [\eta]^{\beta}}$$

- p_{ij} … **probability** of ants that they, at a particular node i, select the route from node $i \rightarrow j$ **("heuristic desirability")**
- $\alpha > 0$ and $\beta > 0$ … the **influence parameters** (α … history coefficient, β …heuristic coefficient) usually $\alpha \approx \beta \approx 2 < 5$
- τ_{ij} … the **pheromone value** for the components, i.e. the amount of pheromone on edge (i, j)
- k … the set of usable components
- J_i … the set of nodes that ant k can reach from v_i (tabu list)
- $\eta_{ij} = \frac{1}{dij}$ … attractiveness computed by a heuristic, indicating the "a-priori **desirability"** of the move

Experimental Game: The travelling Snakesman

Instruction to the Travelling Snakesman NEW versions v1.1 and v2 (as of 18.April 2018)
This page is current as of May, 11, 2018 13:15 CEST

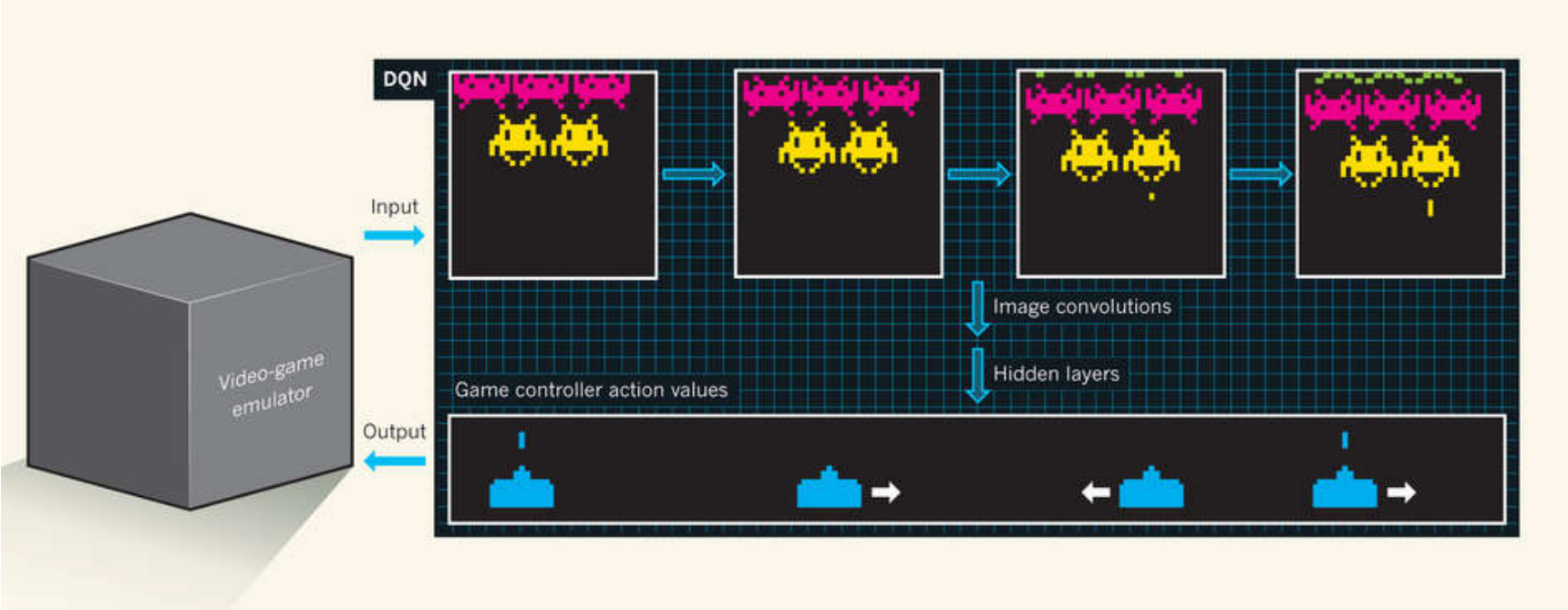
This game uses an iML algorithm for computations in the background. We try to measure, if human interaction with this algorithm leads to better solutions than the algorithm running automatically without any interaction.

https://hci-kdd.org/gamification-interactive-machine-learning

YOU ARE A SNAKE AND YOUR GOAL IS TO EAT ALL APPLES AS FAST AS POSSIBLE! ENJOY PLAYING BOTH VARIANTS!

- You find the links for the Browser and for Android below (just click)
- 1) Enter a name
- 2) Select the level (1 = easy, 3=difficult)
- 3) Press "Play!" With your mouse/touch you direct the snake and your goal is to eat all apples as fast as possible!

If Google is doing their experiments with Games …



Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement learning. **Nature,** 518, (7540), 529-533, doi:10.1038/nature14236

06 Towards Explainable AI

Mastering the game of Go without human knowledge

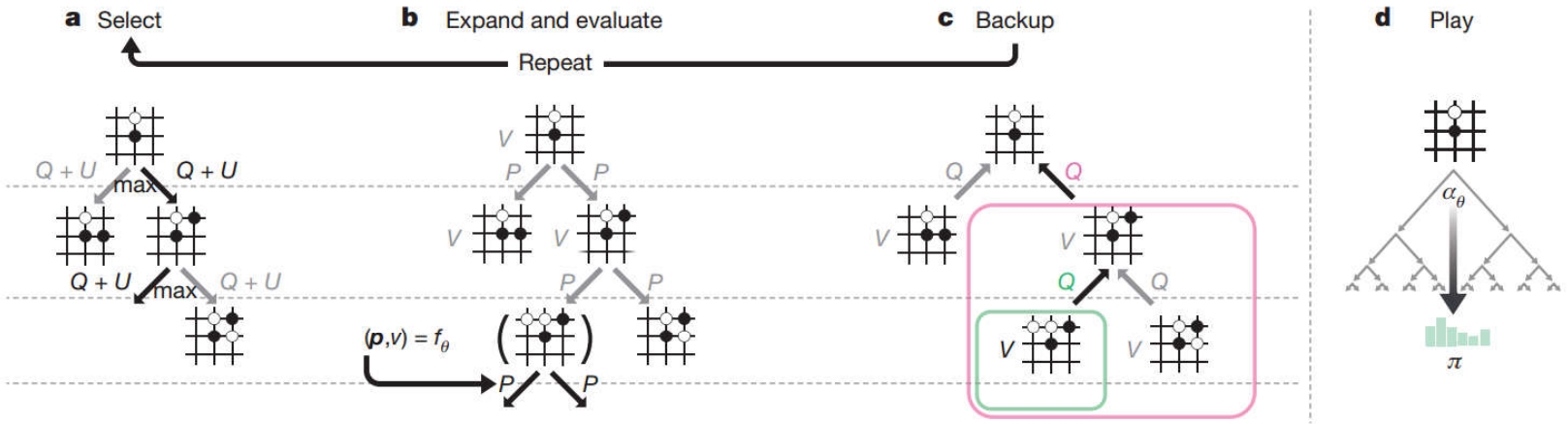


Figure 2 | MCTS in AlphaGo Zero. a, Each simulation traverses the tree by selecting the edge with maximum action value Q, plus an upper confidence bound U that depends on a stored prior probability P and visit count N for that edge (which is incremented once traversed). **b**, The leaf node is expanded and the associated position s is evaluated by the neural network $(P(s, \cdot), V(s)) = f_\theta(s)$; the vector of P values are stored in the outgoing edges from s. **c**, Action value Q is updated to track the mean of all evaluations V in the subtree below that action. **d**, Once the search is complete, search probabilities π are returned, proportional to $N^{1/\tau}$, where N is the visit count of each move from the root state and τ is a parameter controlling temperature.

19 OCTOBER 2017 | VOL 550 | NATURE | 355

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

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

Why did it make this

"It's not a human move. I've never seen a human play this move." (Fan Hui, 2016).

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

Deep Learning Context recognition state-of-the-art

a woman riding a horse on a dirt road

an airplane is parked on the tarmac at an airport

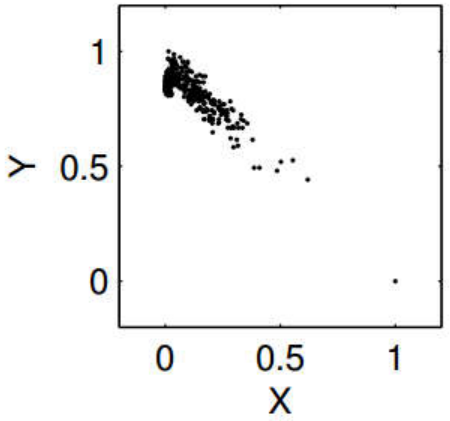
a group of people standing on top of a beach

Andrej Karpathy & Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3128-3137.

Image Captions by deep learning : github.com/karpathy/neuraltalk2

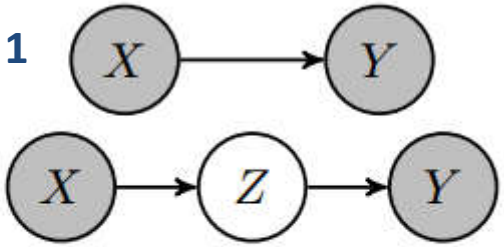
Image Source: Gabriel Villena Fernandez; Agence France-Press, Dave Martin (left to right)

Decide if X → Y, or Y → X using only observed data



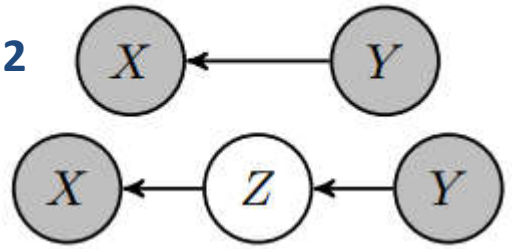
Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.

1



$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

2

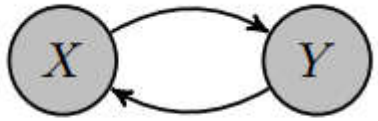


$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

3

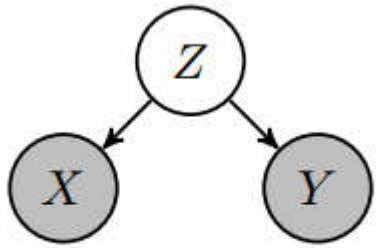
$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

4



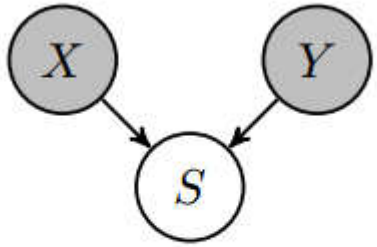
$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

5



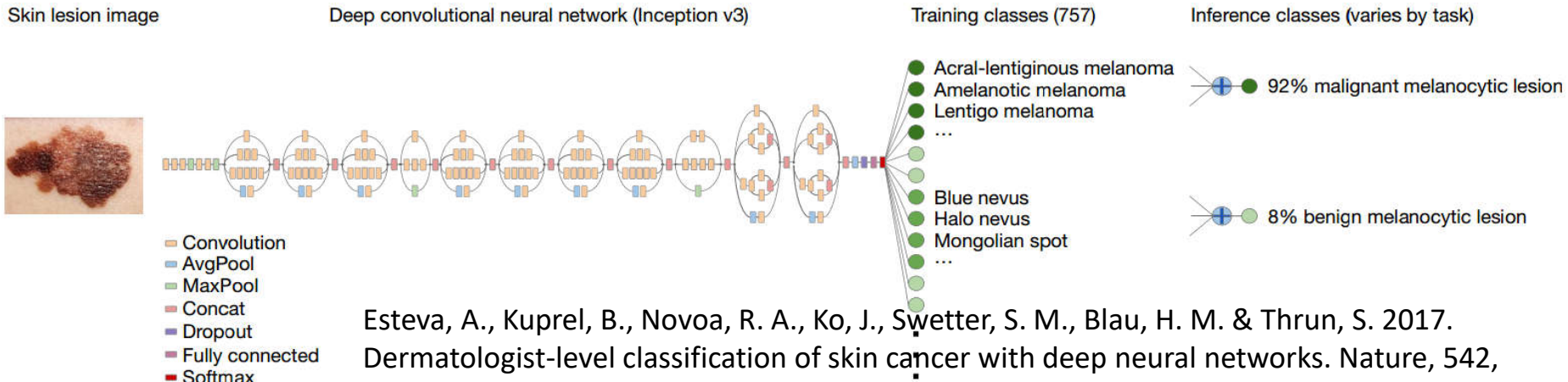
$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

6

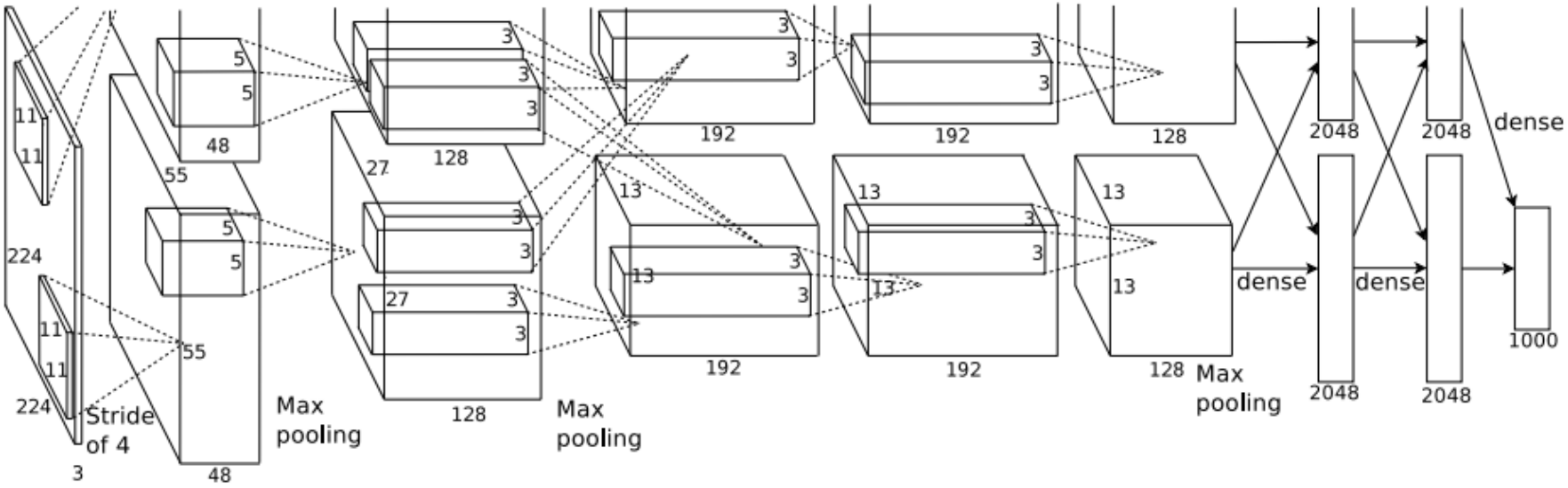


$$\mathbb{P}_{Y\,|\,s} \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x),s} = \mathbb{P}_{Y\,|\,x,s}$$
$$\mathbb{P}_{X\,|\,s} \neq \mathbb{P}_{X\,|\,\mathrm{do}(y),s} = \mathbb{P}_{X\,|\,y,s}$$

Why?



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

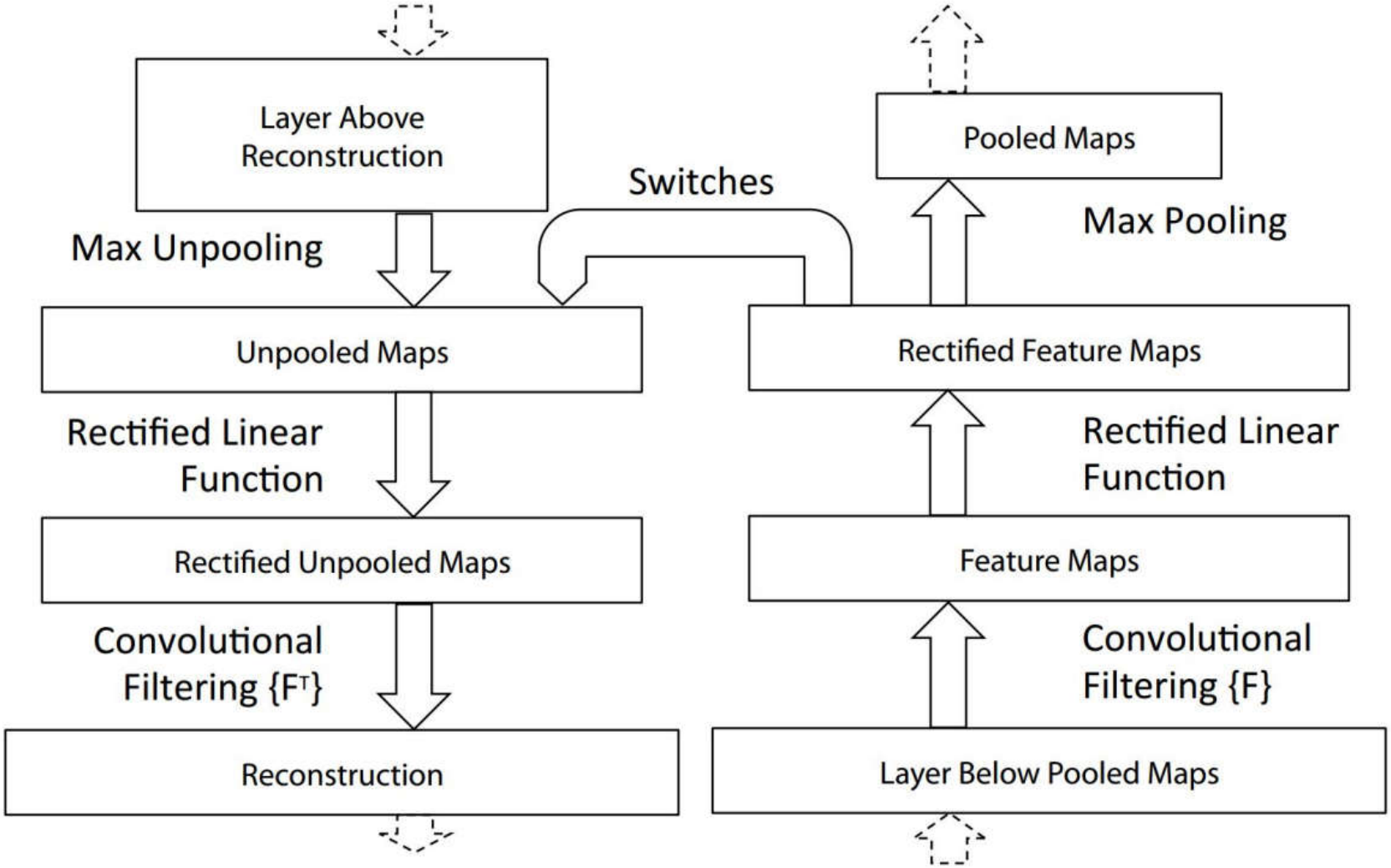


Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

Example: Interpretable Deep Learning Model

HCI-KDD

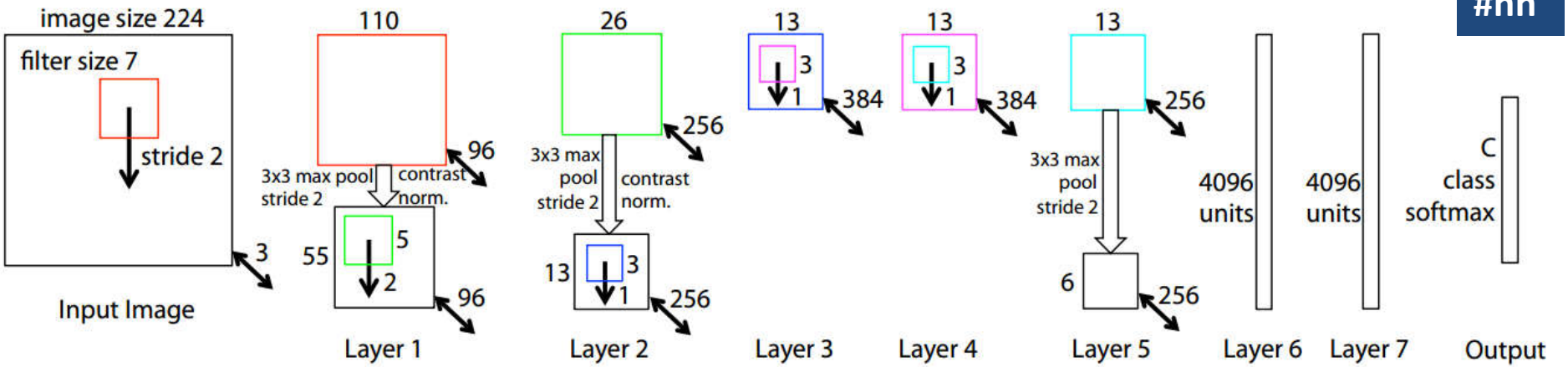
#gg

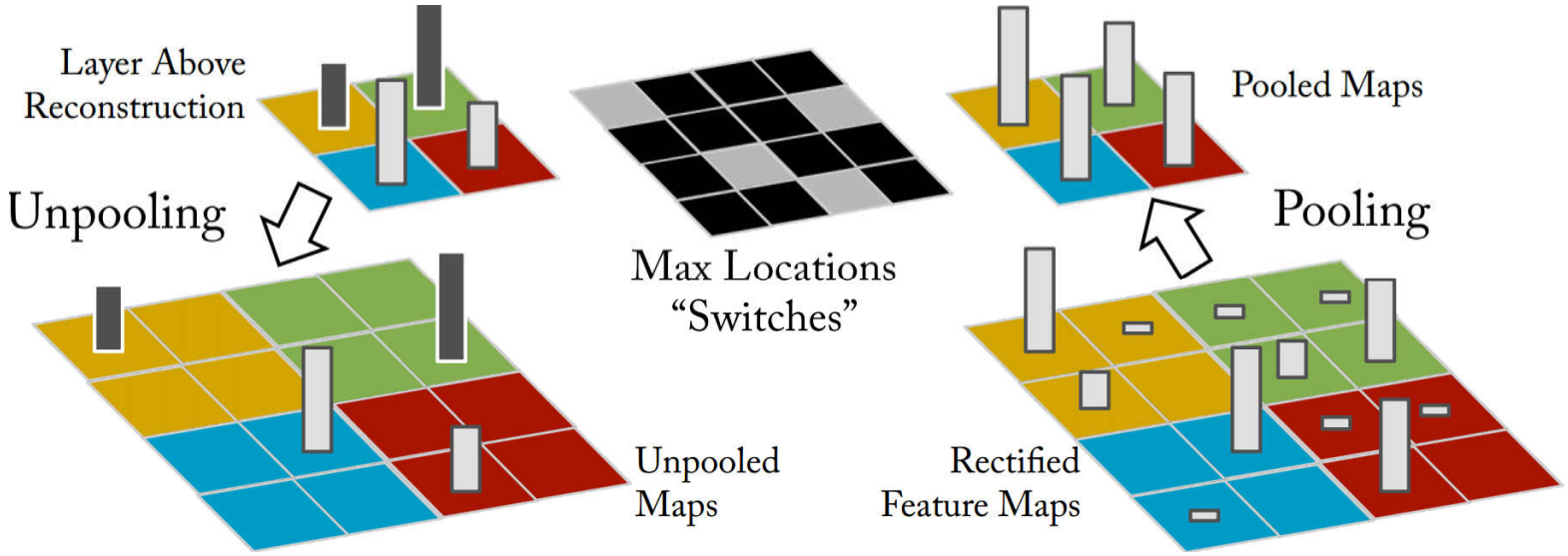


Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

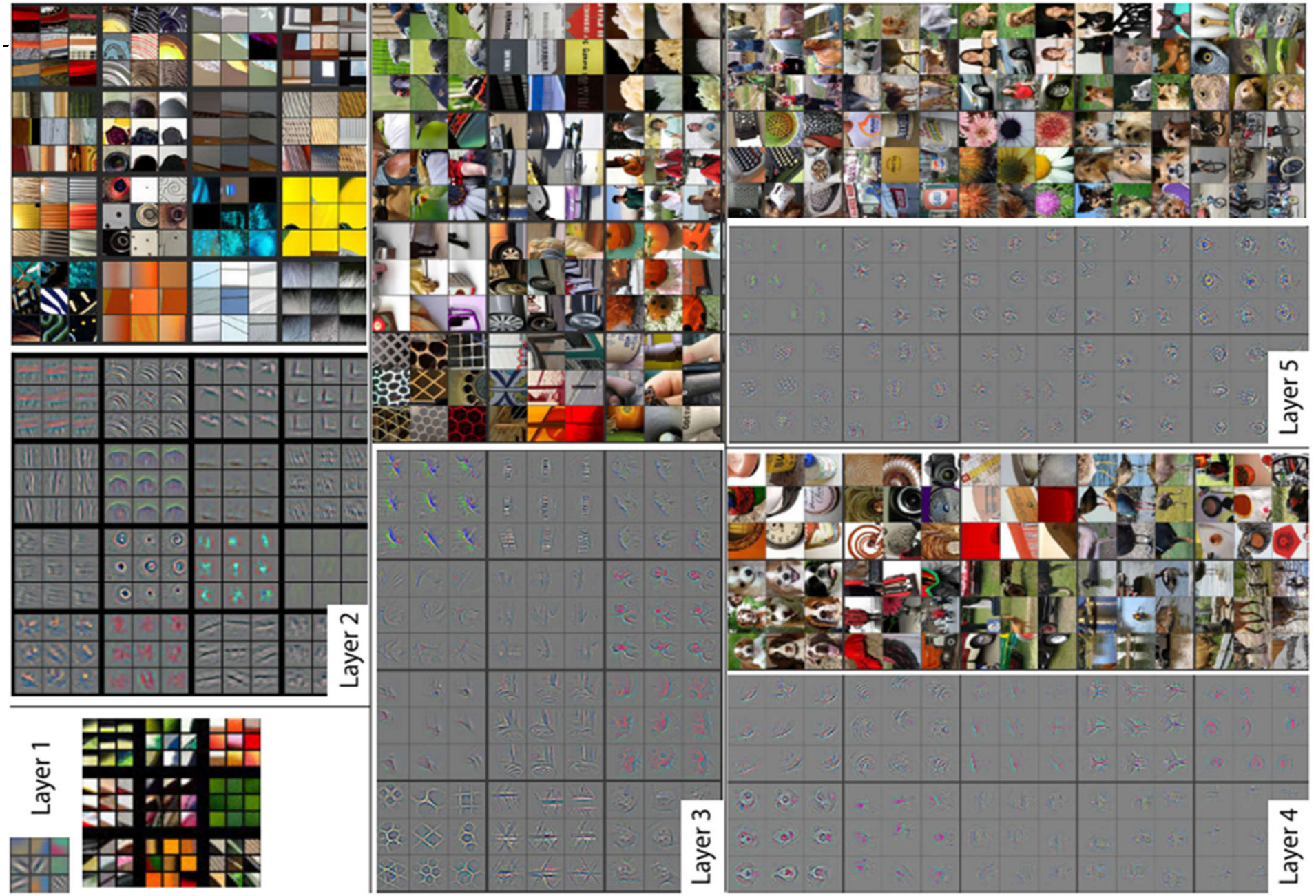
Visualizing a Conv Net with a De-Conv Net

#hh





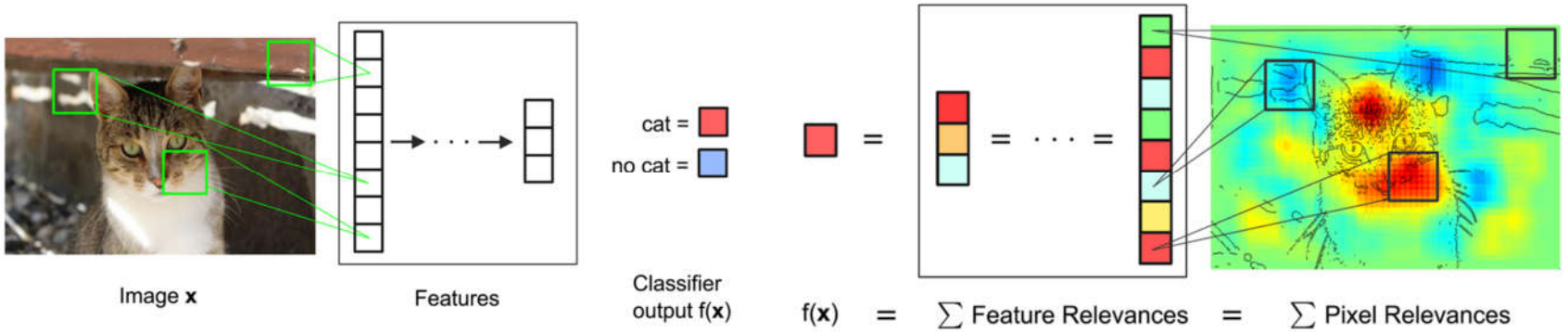
Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



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

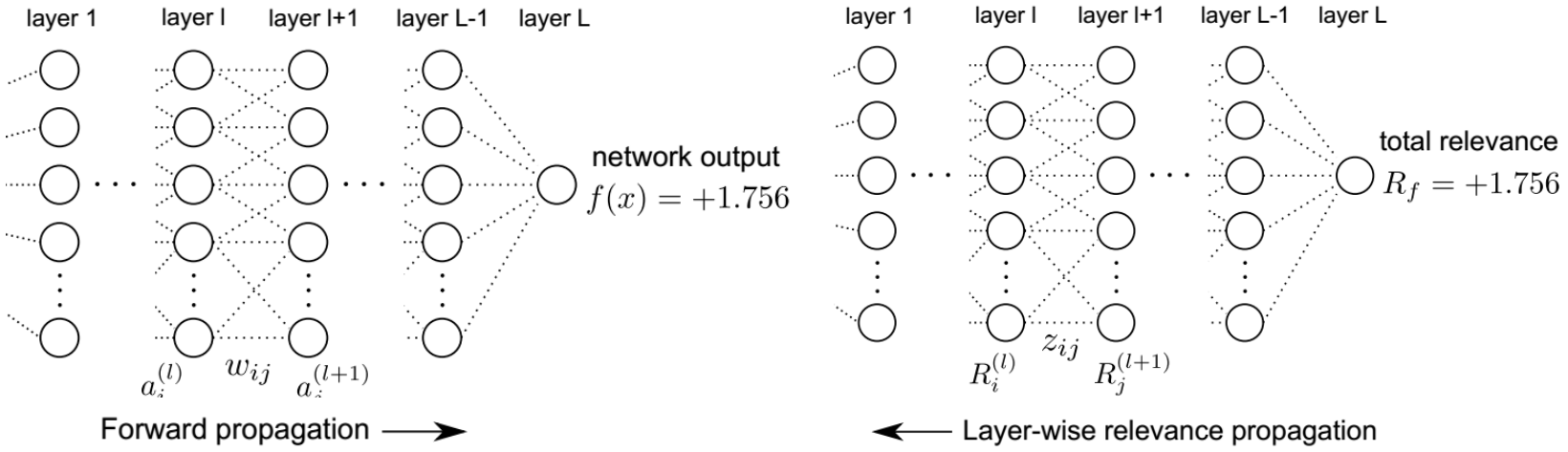
LRP Layer-Wise Relevance Propagation

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



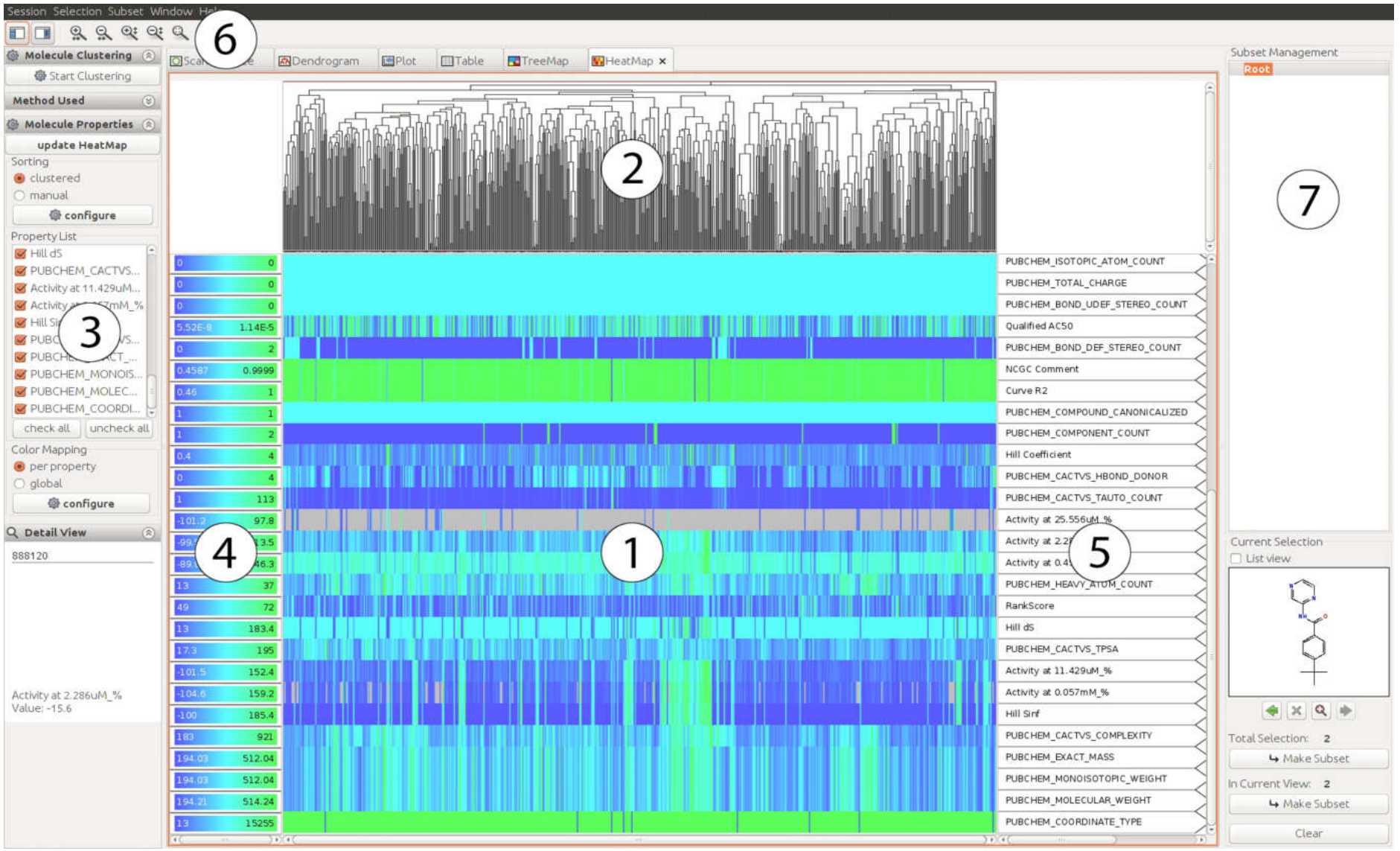
$$a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)$$

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



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

Comprehensible, re-traceable, re-enactable for the user

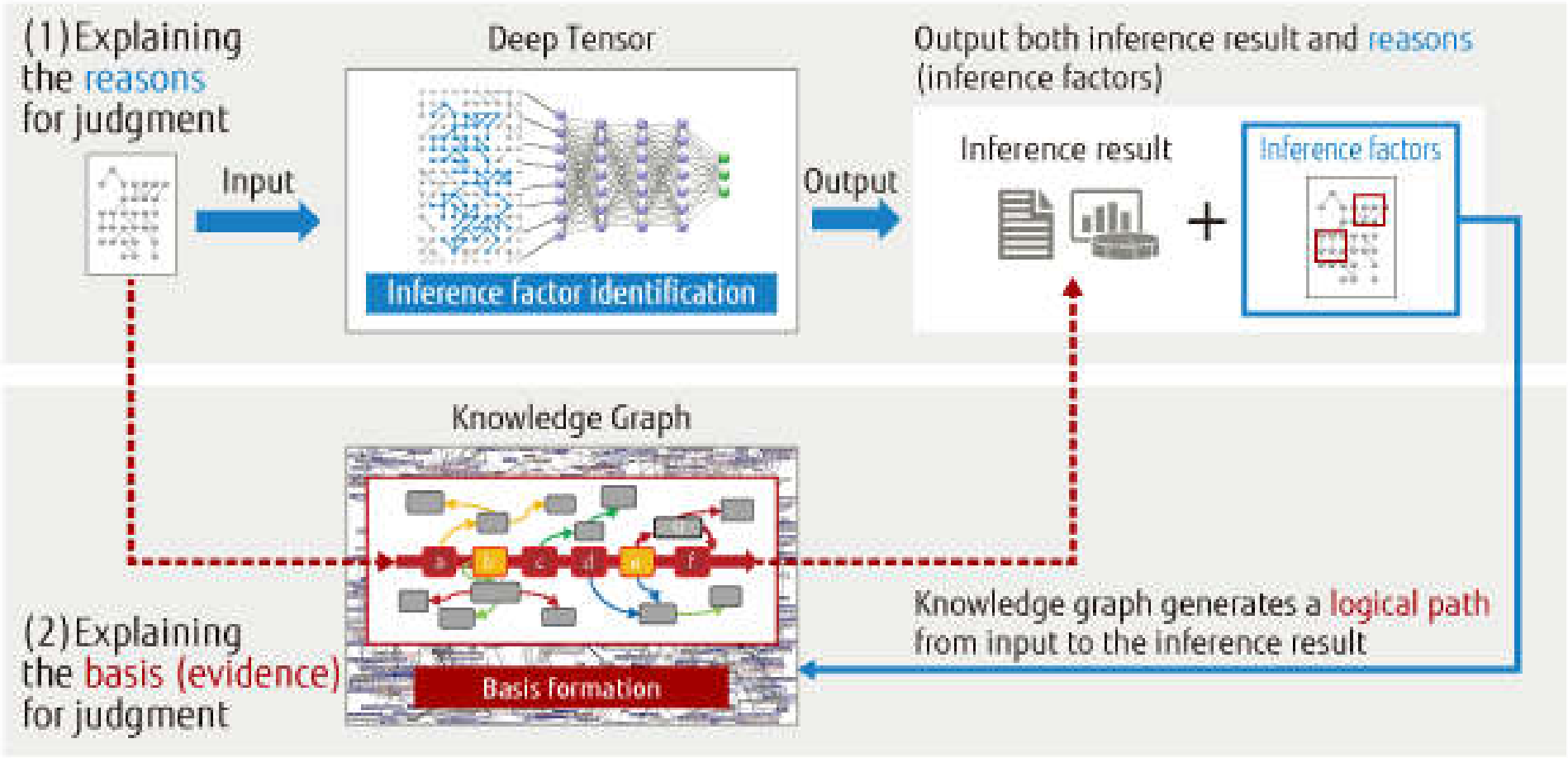


Werner Sturm, Till Schaefer, Tobias Schreck, Andeas Holzinger & Torsten Ullrich. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps In: Borgo, Rita & Turkay, Cagatay, eds. EG UK Computer Graphics & Visual Computing CGVC 2015, 2015 University College London (UCL). Euro Graphics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

Combination of Deep Learning with Ontologies

HCI-KDD

#bb

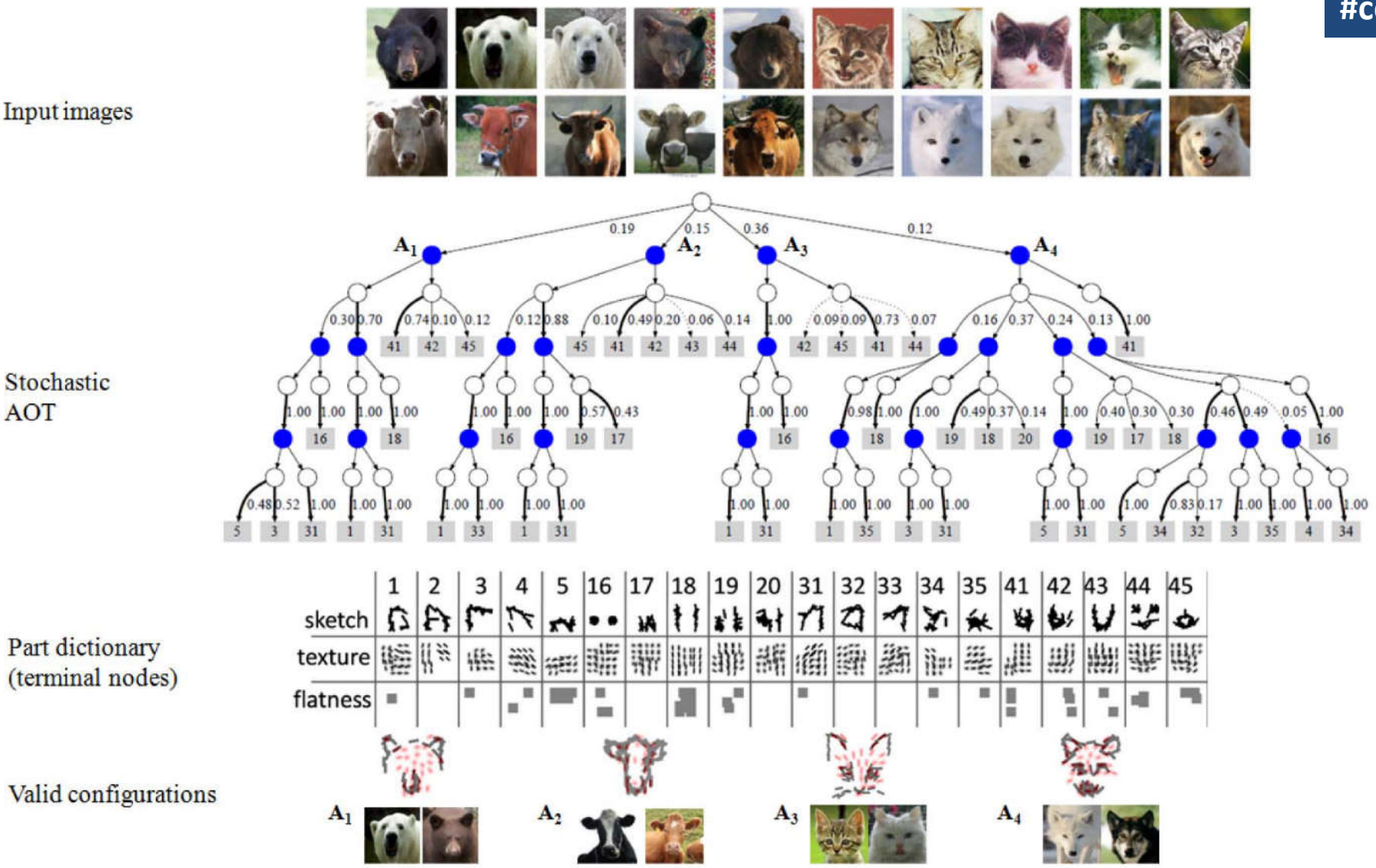


Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger 2018. Explainable AI: the new 42? Springer Lecture Notes in Computer Science LNCS 11015. Cham: Springer, pp 295-303

Stochastic AND-OR Templates for visual objects

HCI-KDD

#cc



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.


Conclusion

- ❶ Computational approaches can find in $\mathbb{R}^N$ what no human would be able to see
- ❷ Complexity – reduction of search space **augment** Human intelligence with AI & v.v.
- ❸ Human expert can understand the **context,** need **effective** mapping $\mathbb{R}^N \rightarrow \mathbb{R}^2$
- ❹ Black box approaches can not explain **WHY** a decision has been made …

Multi-Task Learning …

help to reduce **catastrophic forgetting**

Transfer learning …

is not easy: learning to perform a task by exploiting knowledge acquired when solving previous tasks:

A solution to this problem would have major impact to AI research generally and ML specifically!

Multi-Agent-Hybrid Systems …

collective intelligence and crowdsourcing
client-side federated machine learning – ensures **privacy, data protection, safety & security …**

A final citation attributed to Albert Einstein …

- Computers are fast, accurate and stupid,
- humans are slow, inaccurate and brilliant,
- **together** they are powerful beyond imagination

(Einstein never said that)

https://www.benshoemate.com/2008/11/30/einstein-never-said-that

Ďakujem!