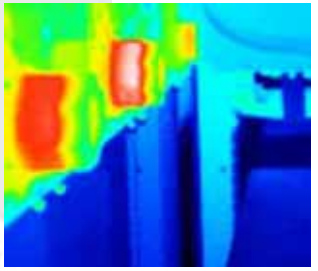
From Explainable AI to Human-Centered AI

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)
Institute for Medical Informatics/Statistics,
Medical University Graz, Austria
and TU Austria (TU Graz & TU Wien)

@aholzin #KandinskyPatterns

20+ M samples



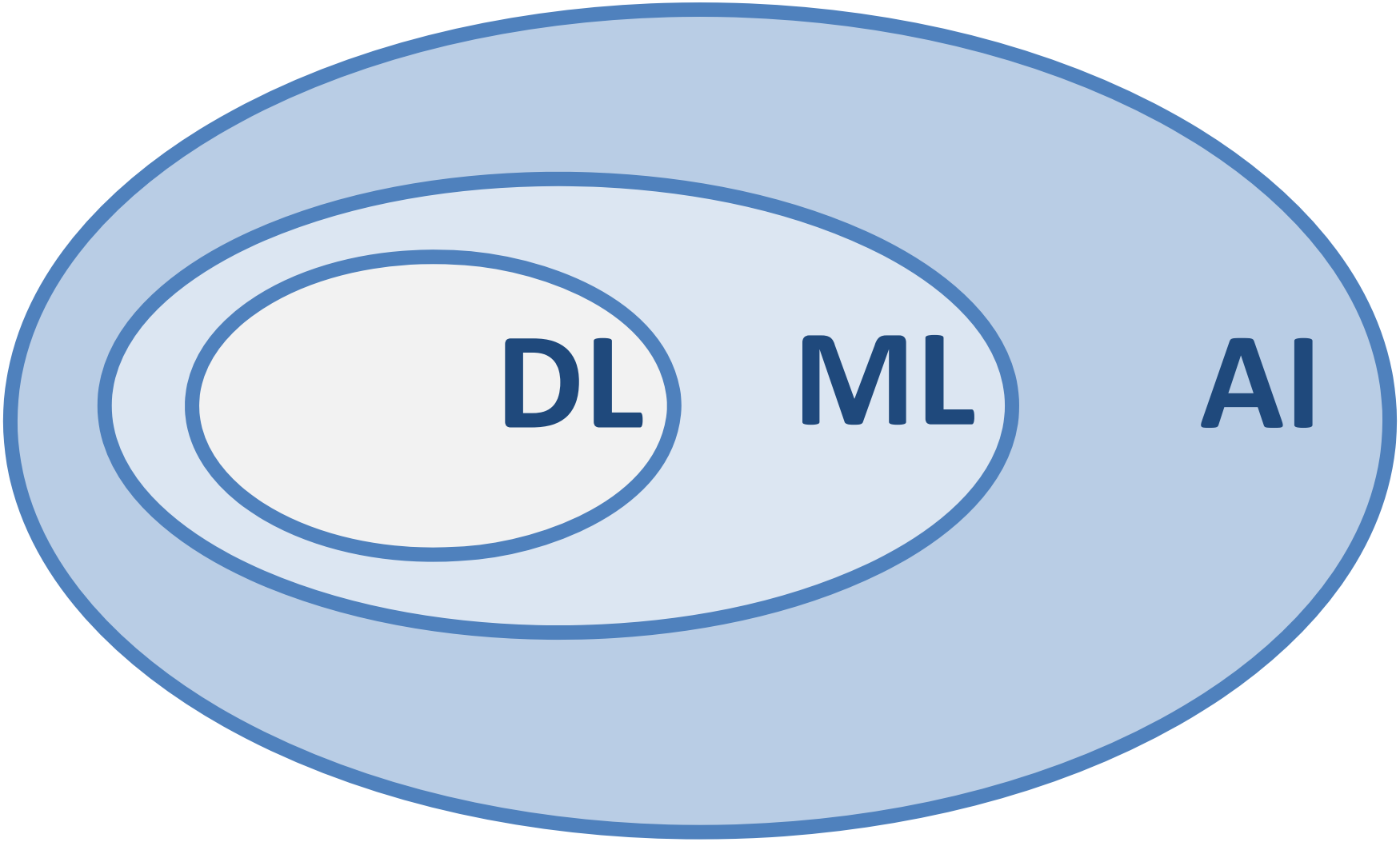

-86
-41
479

Biobank Graz
The largest biobank in Europe
LSH 3
MORGEN EPIC
LifeLines
MORGEN
HP-CVDRF
ACS
NLCS
PBCCBF
PBCCCM
Banu BB
NCT
TUM
BB MUG
BioBank
GRAZ
BioPIC
CEPH
CRB Bichat
BNA DN
CNIO
Biobank 06
ZONCO 3
TBGI
CNESPS
BBI
Surface is proportional to number of samples
Tissue, Blood, DNA
Blood, DNA

Graz, Steiermark, Austria


HELP

Preamble to ensure mutual understanding

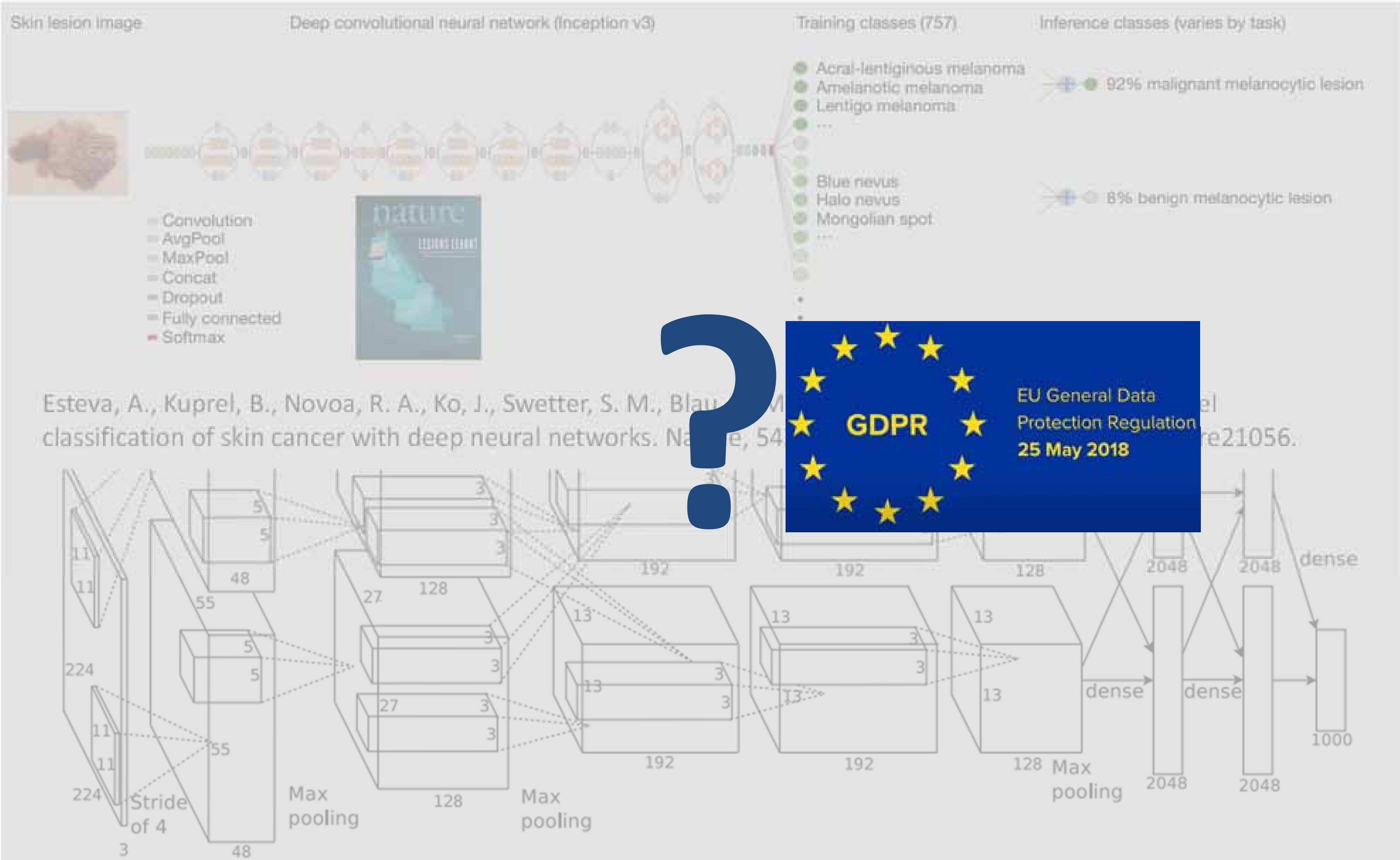


Andreas Holzinger, Peter Kieseberg, Edgar Weippl & A Min Tjoa 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. Springer Lecture Notes in Computer Science LNCS 11015. Cham: Springer, pp. 1-8, doi:10.1007/978-3-319-99740-7_1

Houston, we have a problem … drawbacks of DL

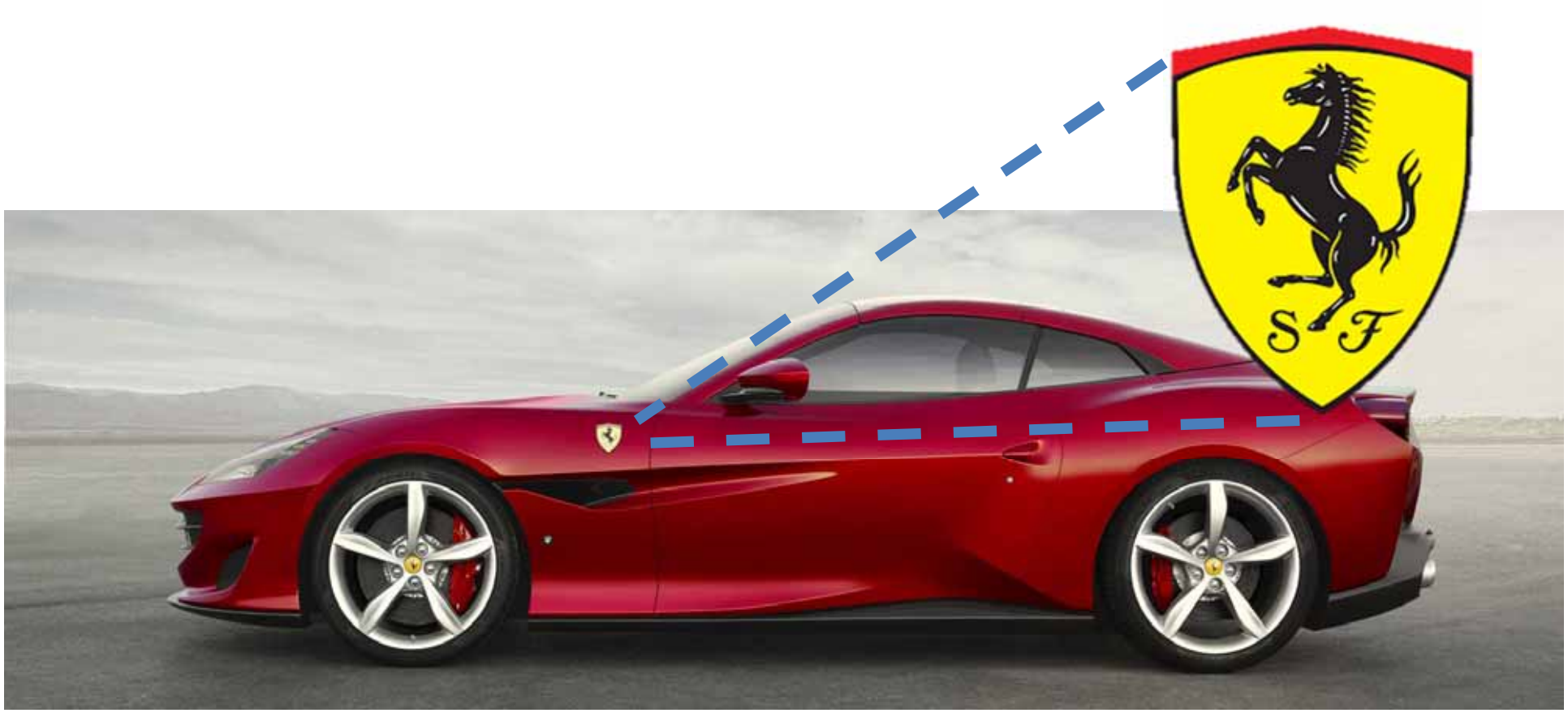
- High-dimensional
- Non-convex
- Resource and data hungry
- Lacking interpretability …

Remember Richard Feynman on the universe: It's not complicated, it's just a lot …


Skin lesion image
Deep convolutional neural network (Inception v3)
Training classes (757)
Inference classes (varies by task)
Acral-lentiginous melanoma
Amelanotic melanoma
Lentigo melanoma
Blue nevus
Halo nevus
Mongolian spot
92% malignant melanocytic lesion
8% benign melanocytic lesion
Convolution
AvgPool
MaxPool
Concat
Dropout
Fully connected
Softmax
nature
Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau
classification of skin cancer with deep neural networks. Na
?
GDPR
EU General Data
Protection Regulation
25 May 2018
Stride of 4
Max pooling
dense
1000

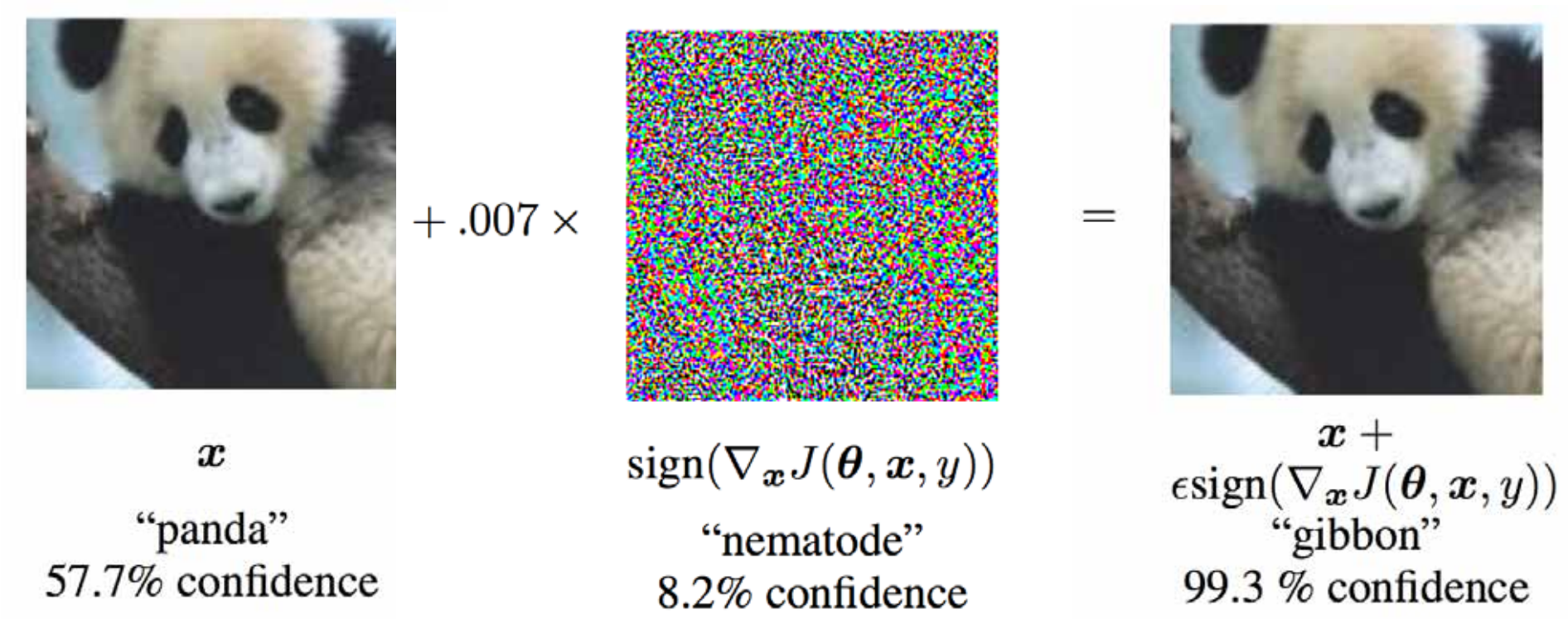
Example for Randy (this is not a joke)

- Result of the classifier: **This is a horse**
- **Why?**



Source: Image is in the public domain

Correlation does not imply Causation



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

Urgent need for explainable AI !

AGENDA

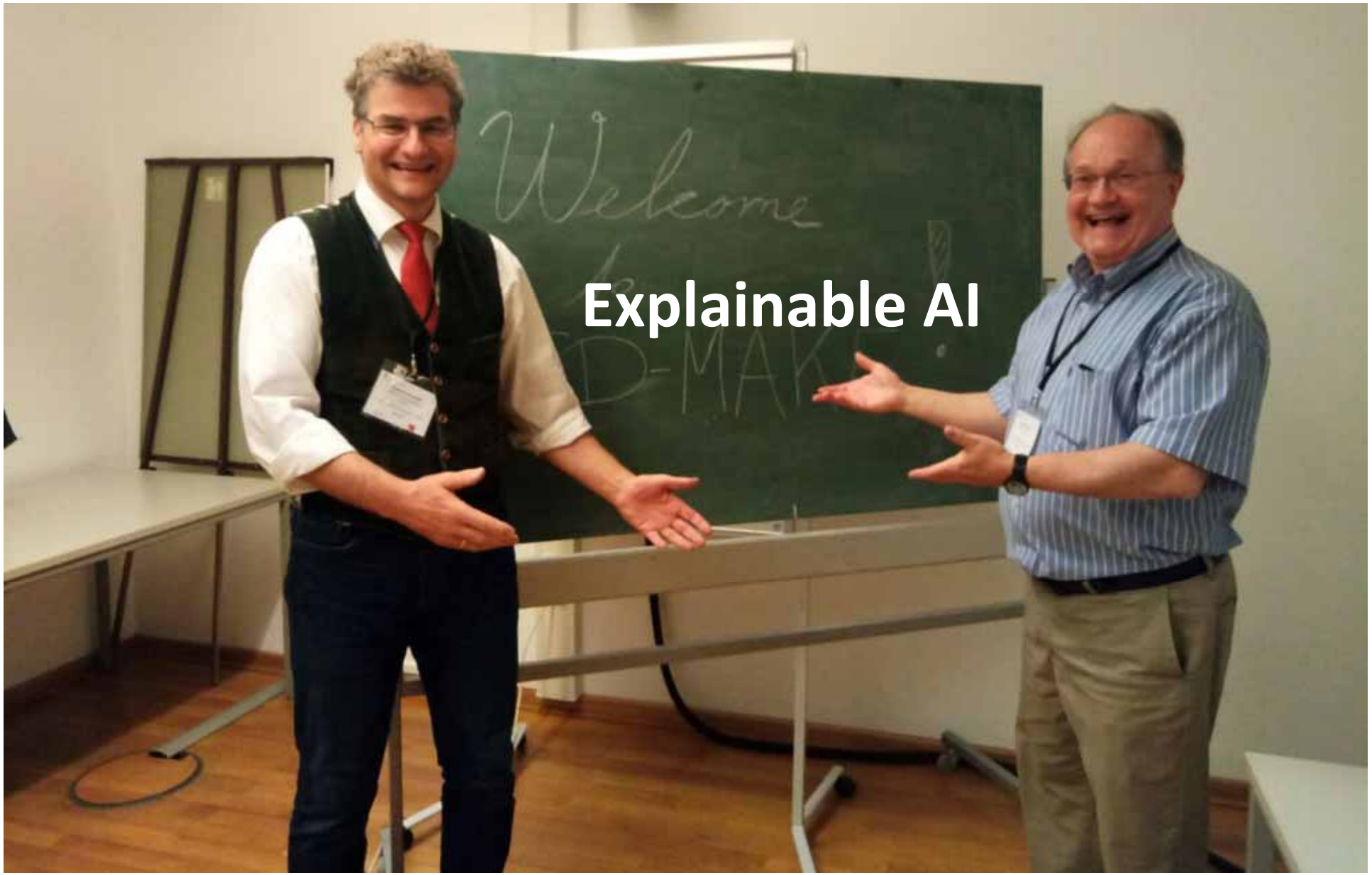
- 00 Motivation
- 01 What is the HCAI approach?
- 02 Application Area: Health Informatics
- 03 Probabilistic Information
- 04 aML
- 05 iML
- 06 Explainable AI (state-of-the-art)
- 07 Towards interpretable ML
- 08 Measuring Machine Intelligence (Kandinsky)

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

Successful ML needs …


https://human-centered.ai/international-expert-network


Lecture Notes in Computer Science
LNCS LNAI LNBI
CD-MAKE 2019
ifip
Cross Domain Conference for Machine Learning and Knowledge Extraction
BMC
Part of Springer Nature
SBA Research
https://cd-make.net
Image with friendly permission of Michael D. Beckwith

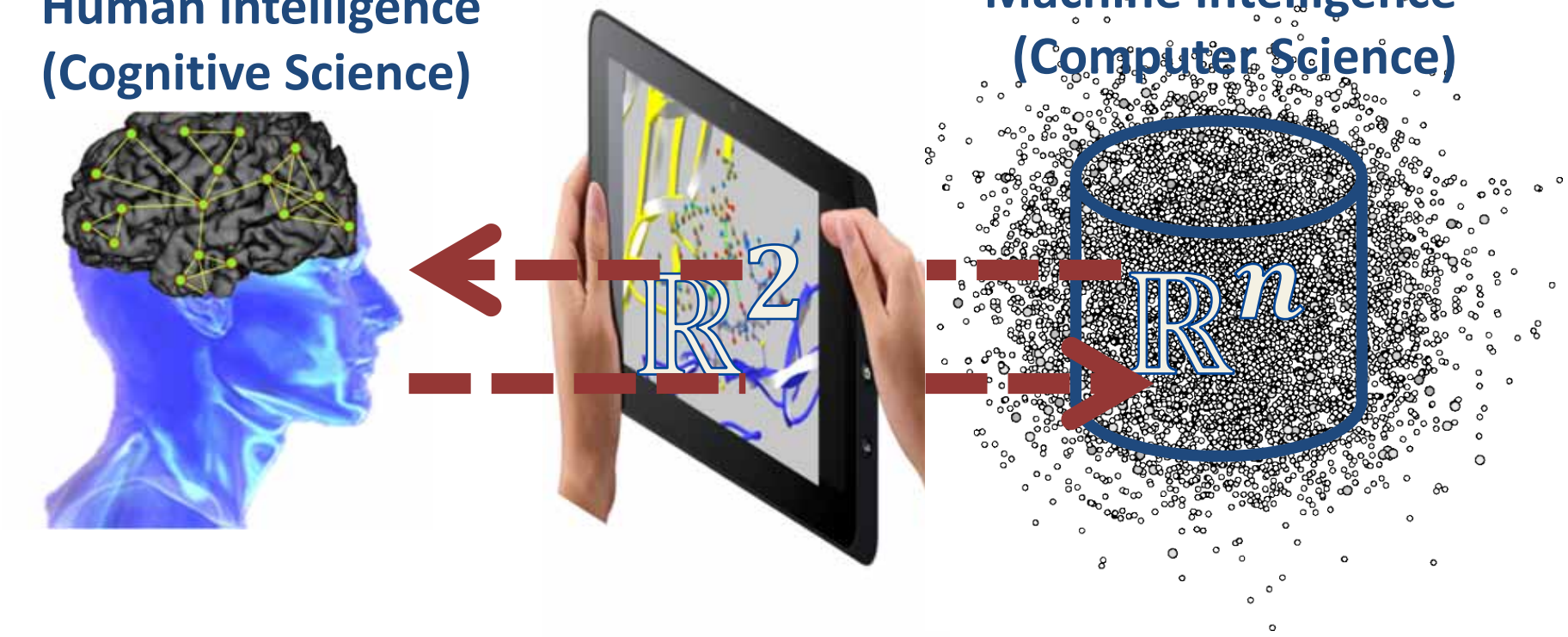

Welcome
Explainable AI

Grand challenge: Mapping from $\mathbb{R}^n$ to $\mathbb{R}^2$

Our goal is that human values are aligned to ensure responsible machine learning

Human intelligence (Cognitive Science)

Machine intelligence (Computer Science)



Andreas Holzinger 2013. Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Lecture Notes in Computer Science LNCS 8127. pp. 319-328, doi:10.1007/978-3-642-40511-2_22.

"Solve intelligence – then solve everything else"

Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

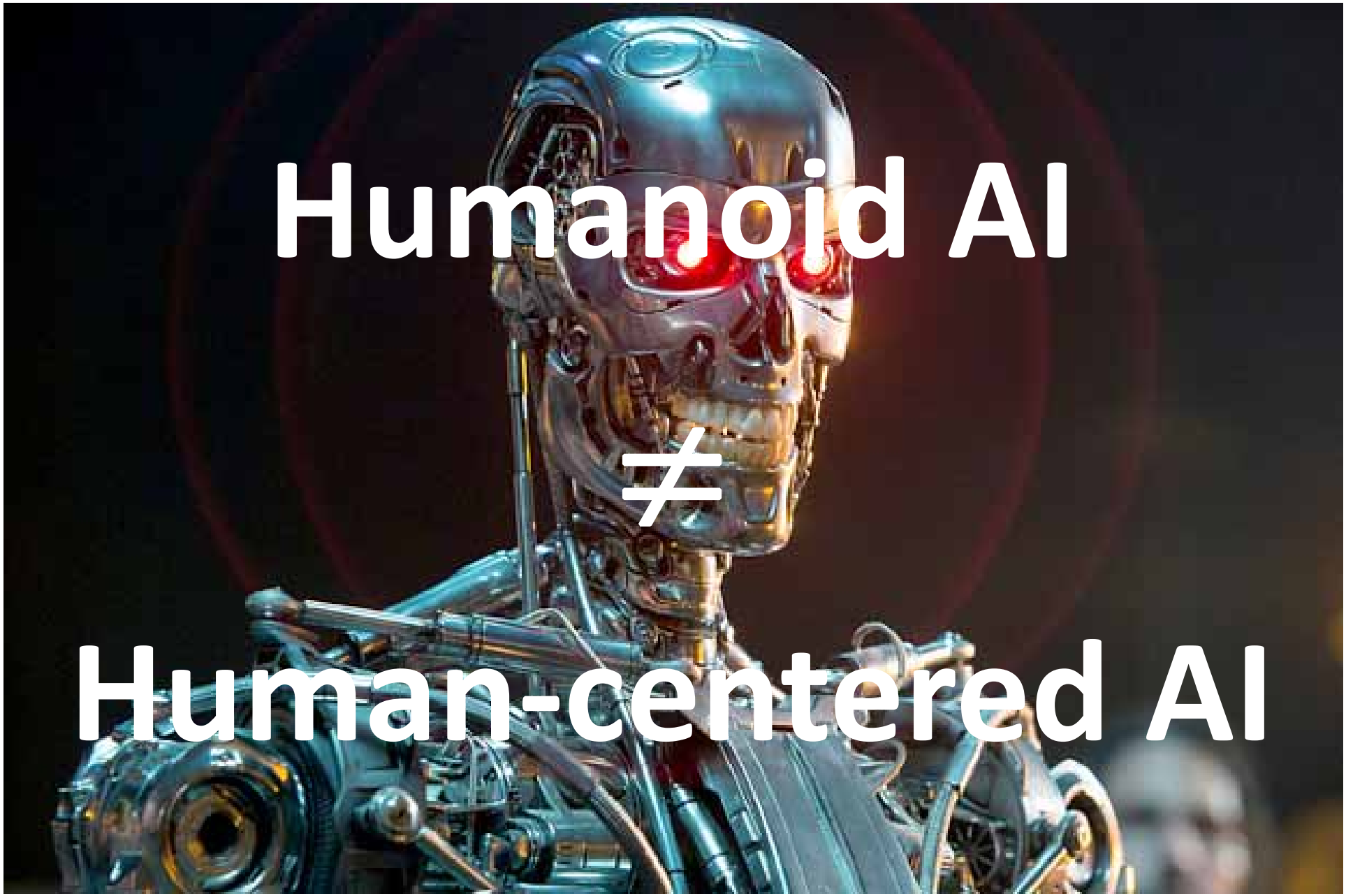
https://youtu.be/XAbLn66iHcQ?t=1h28m54s

To reach a level of usable intelligence we need to …

- 1) learn from prior data
- 2) extract knowledge
- 3) 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

Our goal is: Understanding Context !

Not our Goal



02 Application Area Health Informatics

Why is this application area complex ?

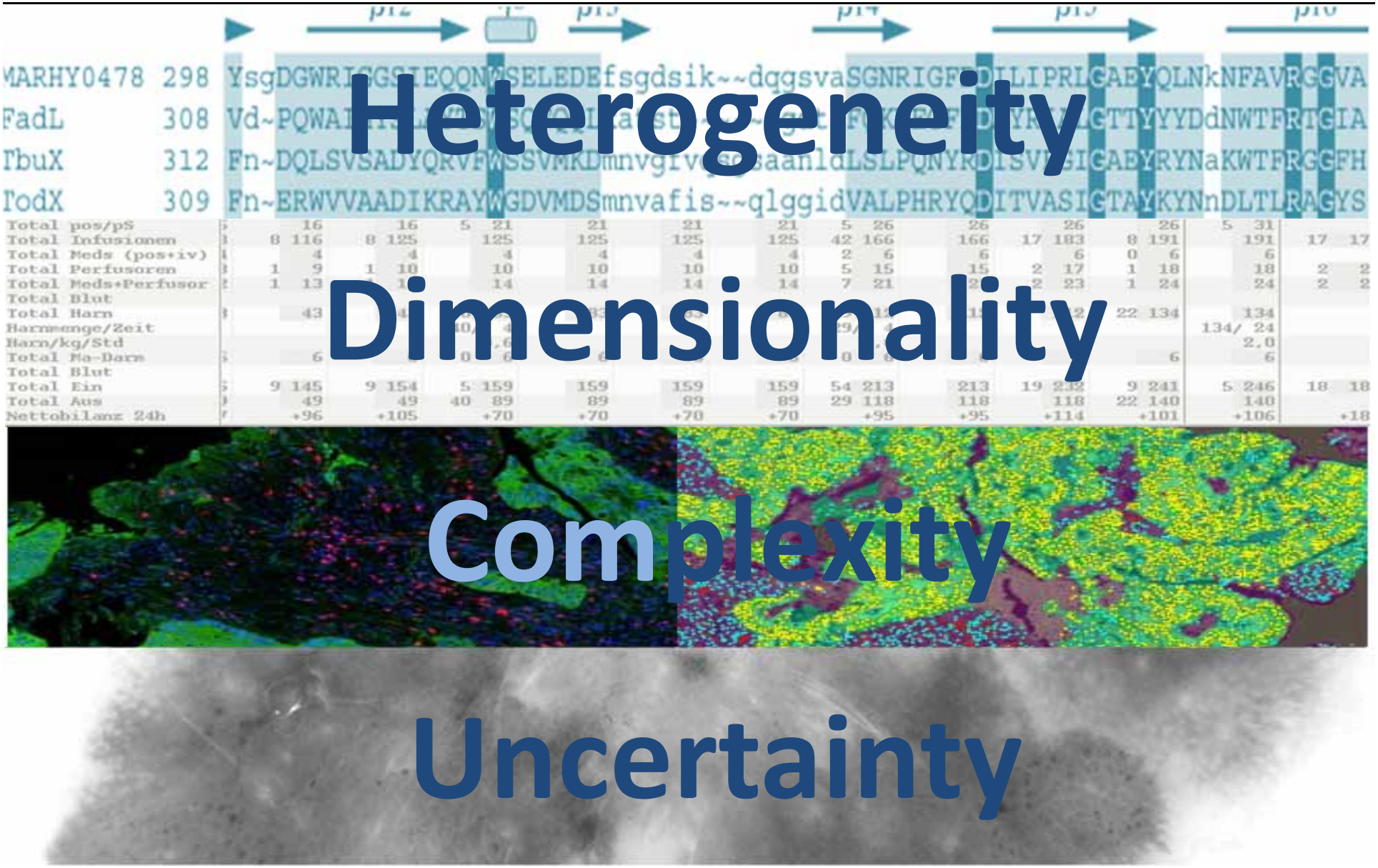
In medicine we have two different worlds …

Our central hypothesis: Information may bridge this gap

Andreas Holzinger & Klaus-Martin Simonic (eds.) 2011. Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer, doi:10.1007/978-3-642-25364-5.

The bridge …

Main problems …



Andreas Holzinger, Matthias Dehmer & Igor Jurisica 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. Springer/Nature BMC Bioinformatics, 15, (S6), I1, doi:10.1186/1471-2105-15-S6-I1.

03 Probabilistic Learning

The true logic of this world is in the calculus of probabilities.
James Clerk Maxwell

Maxwell, J. C. (1850). Letter to Lewis Campbell; reproduced in L. Campbell and W. Garrett, The Life of James Clerk Maxwell, Macmillan, 1881.

Probability theory is nothing but common sense reduced to calculation …

Pierre Simon de Laplace (1749-1827)

The foundation for modern machine learning …

- **1763: Richard Price publishes post hum the work of Thomas Bayes (see next slide)**
- **1781: Pierre-Simon Laplace: Probability theory is nothing but common sense reduced to calculation ...**
- **1812: Théorie Analytique des Probabilités, now known as Bayes' Theorem**

- **Hypothesis** $h \in \mathcal{H}$ (uncertain quantities (Annahmen)
- **Data** $d \in \mathcal{D}$ … measured quantities (Entitäten)
- **Prior probability** $p(h)$ … probability that h is true
  Likelihood $p(d|h)$ … "how probable is the prior"
- **Posterior Probability** $p(h|d)$ … probability of h given d

Pierre Simon de Laplace (1749-1827)

$$p(h|d) \propto p(d|h) * p(h)$$

$$p(h|d) = \frac{p(d|h)p(h)}{p(d)}$$

Learning representations (θ, h) from observed data

Observed data:

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

Feature Parameter: θ **or hypothesis *h*** $h \in \mathcal{H}$

Prior belief ≈ prior probability of hypothesis *h*: $p(\theta)$ $p(h)$

Likelihood ≈ *p(x)* of the data that *h* is true $p(\mathcal{D}|\theta)$ $p(d|h)$

Data evidence ≈ marginal *p(x)* that *h* = true $p(\mathcal{D})$ $\sum_{h \in \mathcal{H}} p(d|h) * p(h)$

Posterior ≈ *p(x)* of *h* after seen ("learn") data *d* $p(\theta|\mathcal{D})$ $p(h|d)$

$$posterior = \frac{likelihood * prior}{evidence} \qquad p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

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

Learning and Inference

d … *data*

h … *hypotheses*

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

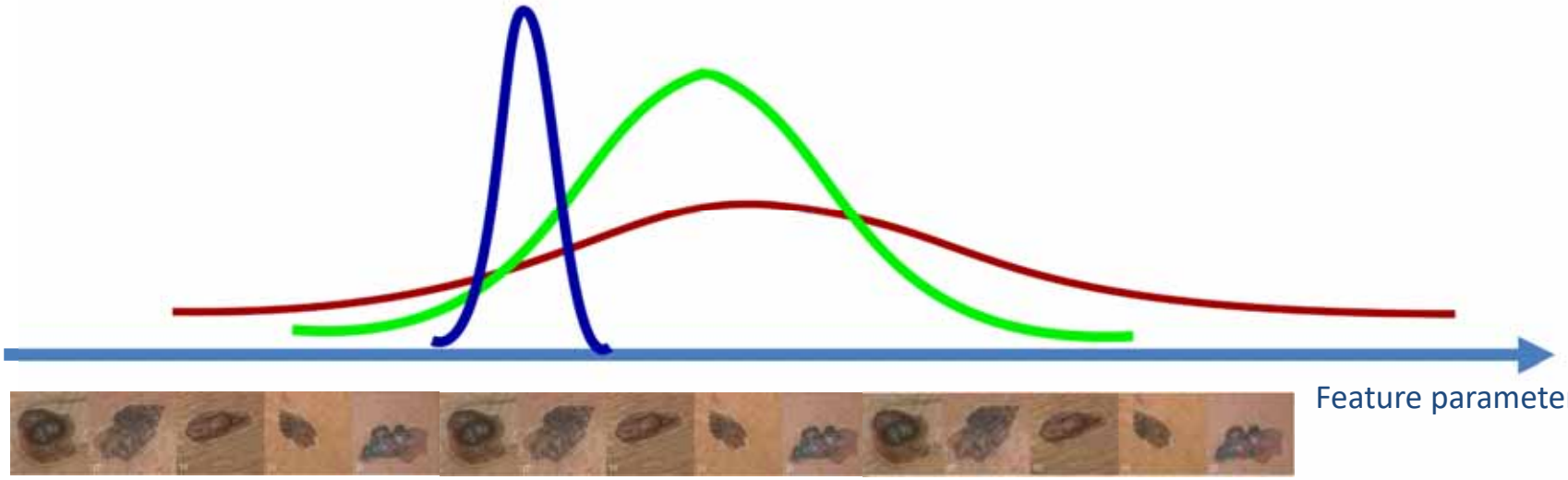
Likelihood

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



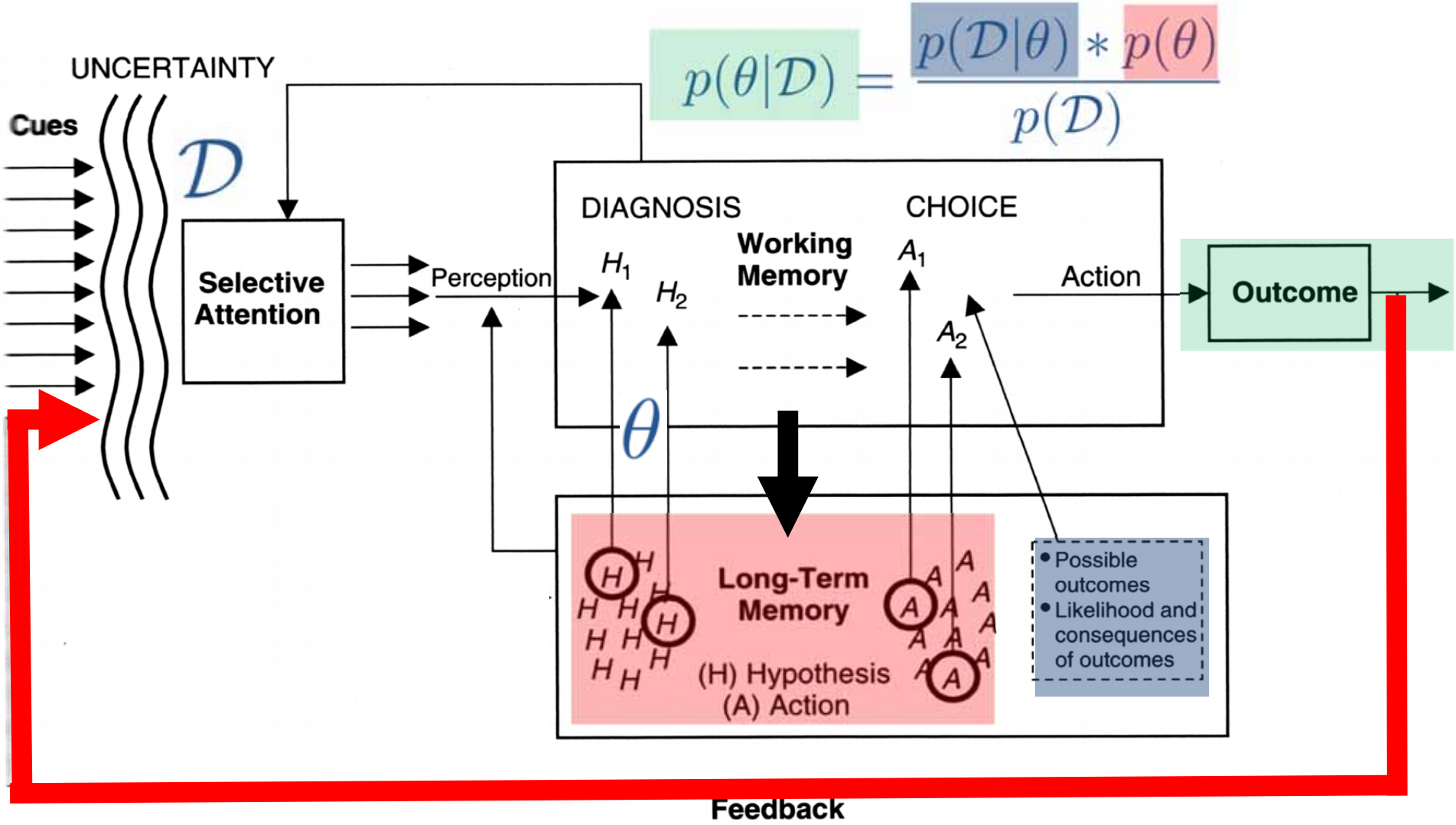
Why is this relevant for medicine?

Reasoning under uncertainty: Decision Making

- Take patient information, e.g., observations, symptoms, test results, -omics data, etc. etc.
- Reach conclusions, and **predict** the future, e.g. how likely will the patient …
- Prior = belief before making a particular observation
- Posterior = belief after making the observation and is the prior for the next observation – intrinsically incremental

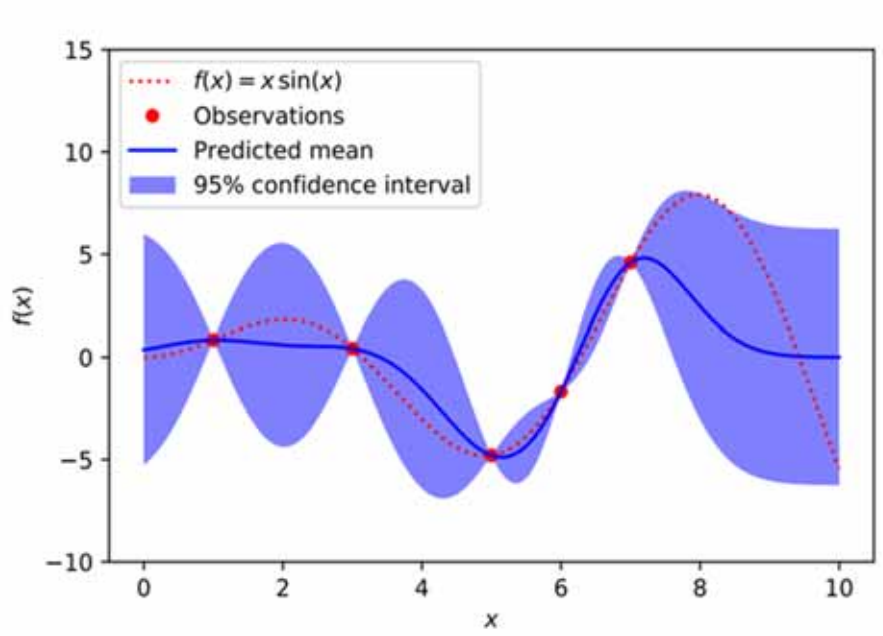
$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}$$

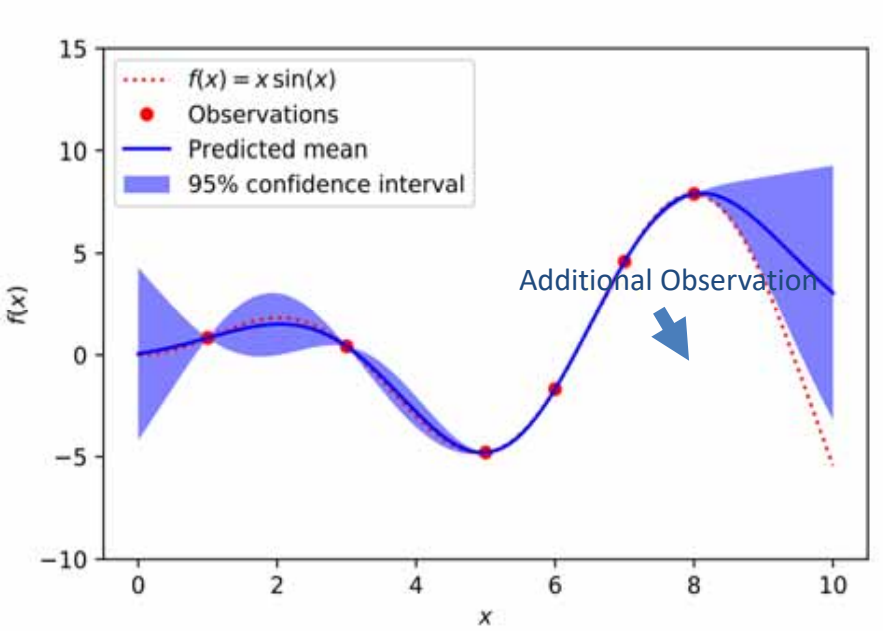
Connection to Cognitive Science: Decision Making



Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill, modified by Holzinger, A.*

Goal: Reduce uncertainty …





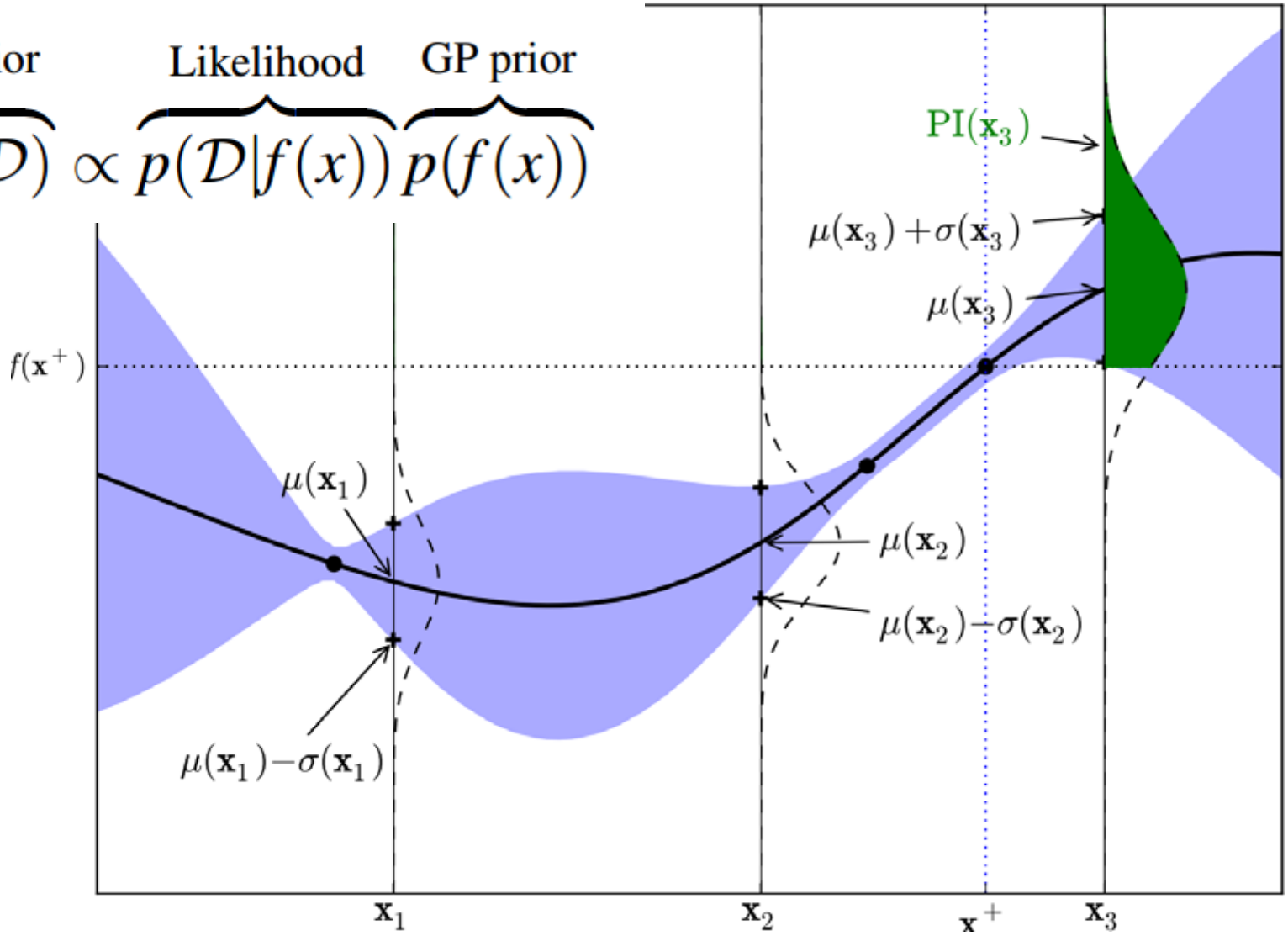
$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

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

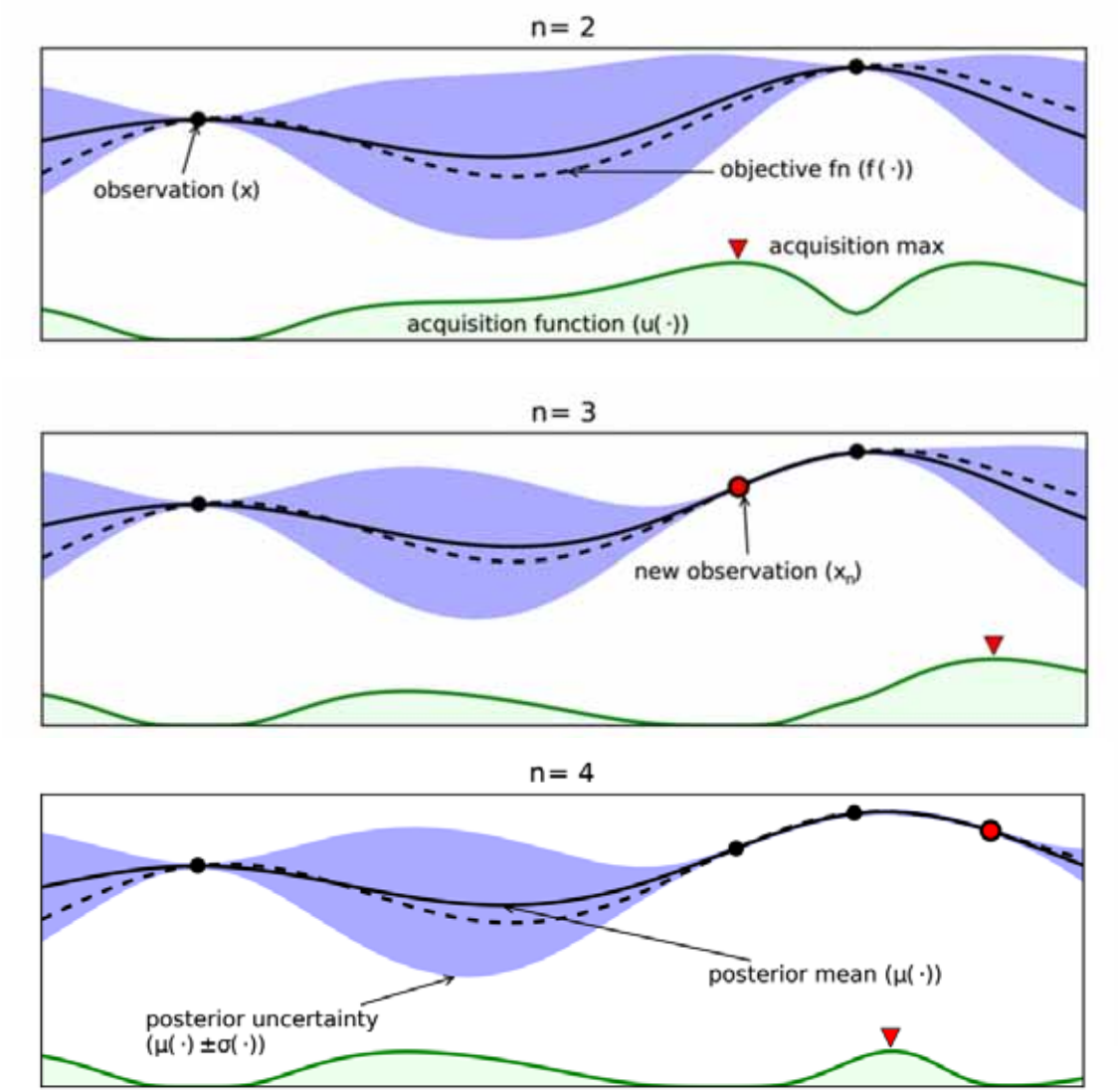
observations occur in a cont. domain, e.g. t or space

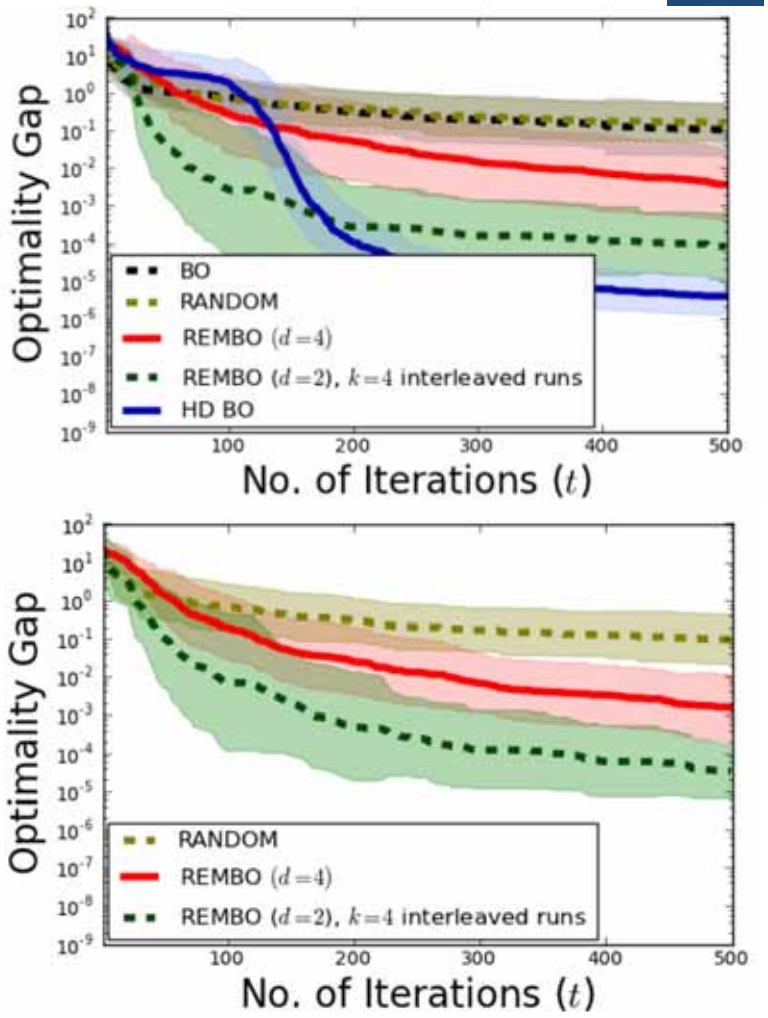
$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}} \overbrace{p(f(x))}^{\text{GP prior}}$$



Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

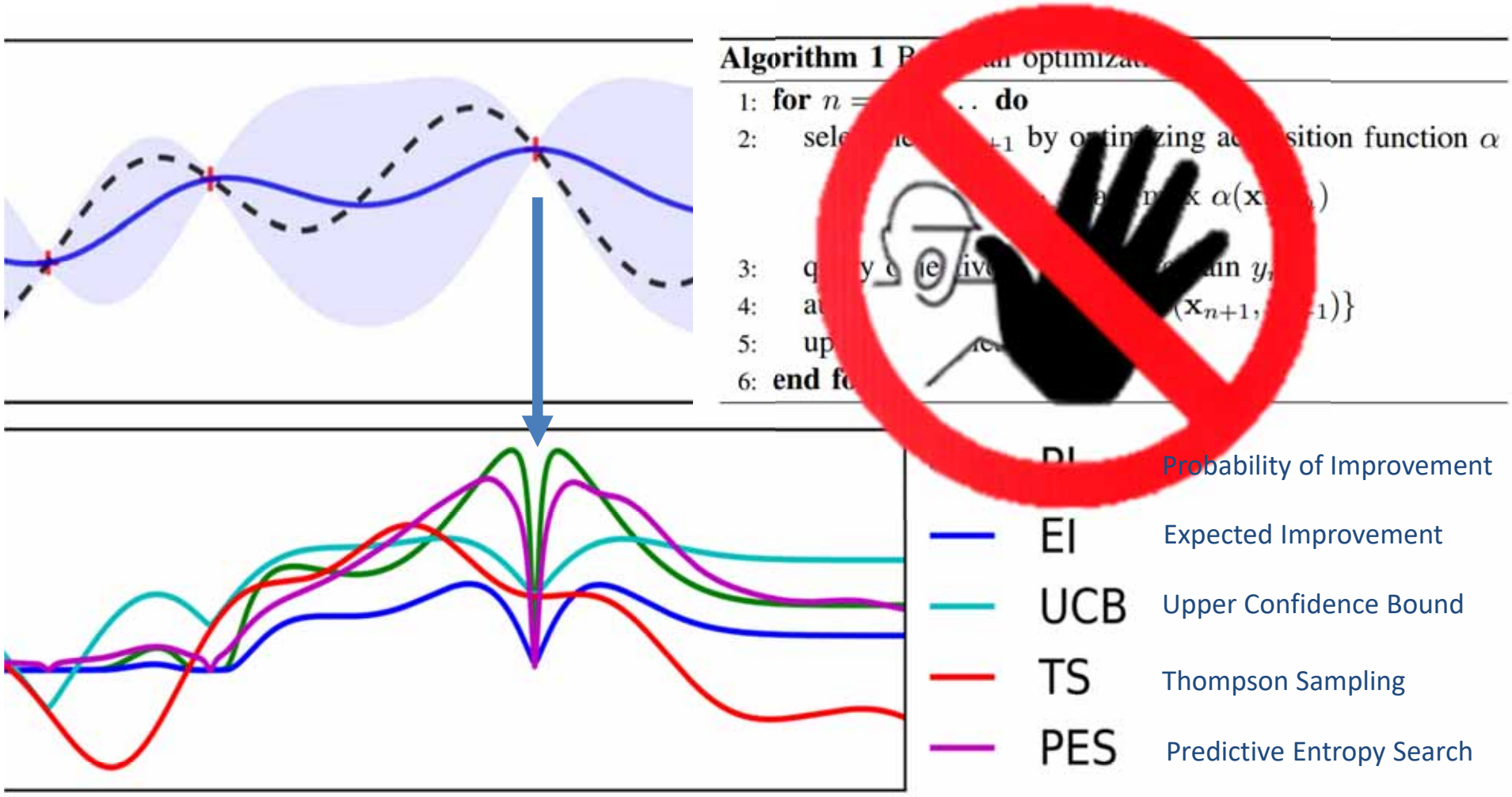
Scaling to high-dimensions is the holy grail in ML





Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research, 55, 361-387, doi:10.1613/jair.4806.

Goal: Taking the human out of the loop

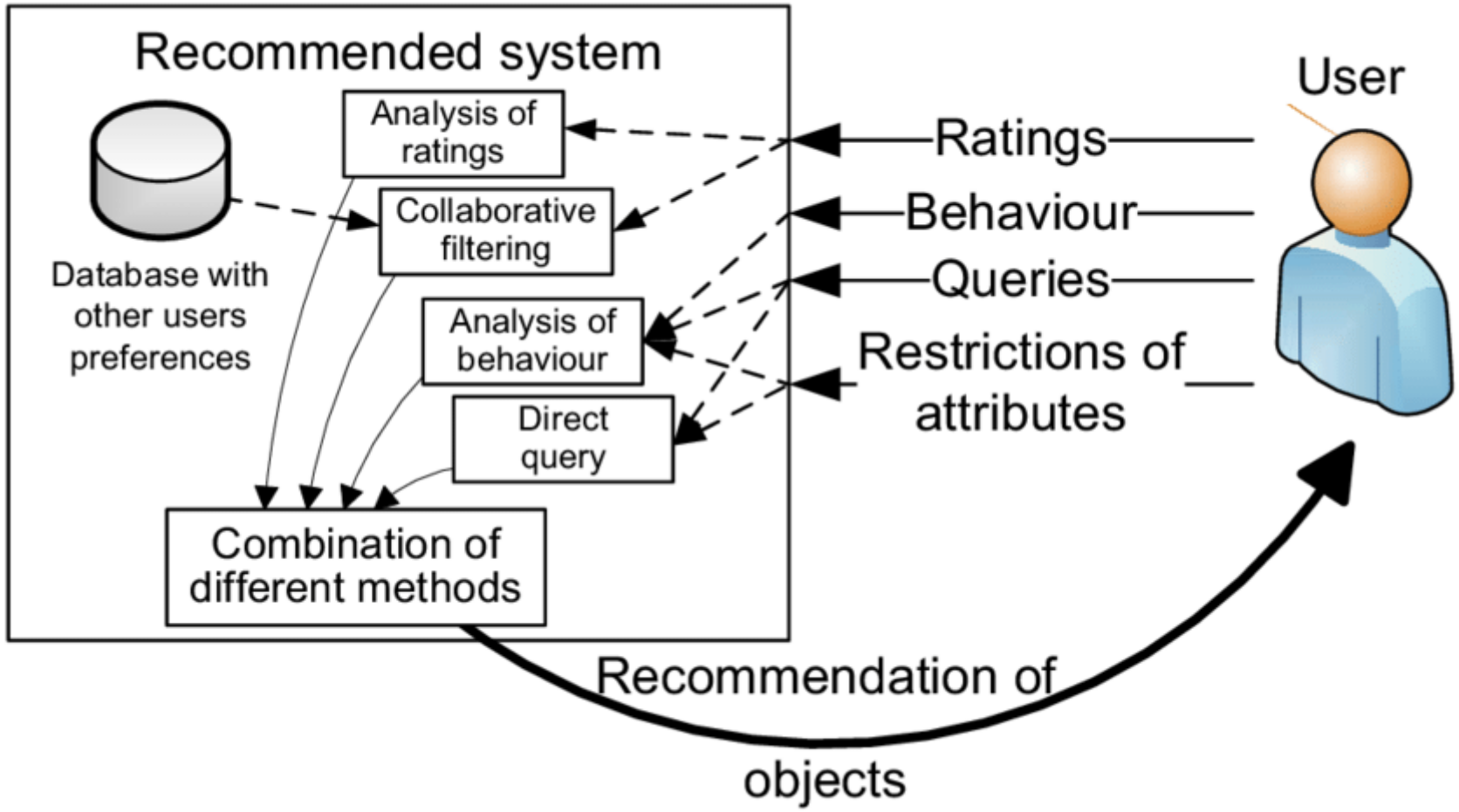


Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. 2016.
Taking the human out of the loop: A review of Bayesian optimization.
Proceedings of the IEEE, 104, (1), 148-175, doi:10.1109/JPROC.2015.2494218.

04 aML

Best practice examples of aML …

Recommender Systems



Alan Eckhardt 2009. Various aspects of user preference learning and recommender systems. DATESO. pp. 56-67.

Fully automatic autonomous vehicles (“Google car”)

a.holzinger@human-centered.ai

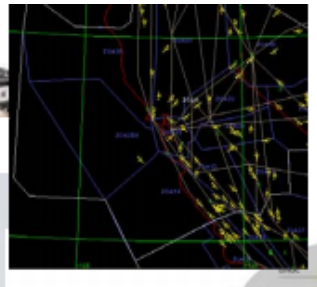
Autonomous aerial vehicle (AAV): passenger drone

https://www.businessinsider.sg/the-worlds-first-passenger-drone-makes-public-flight-in-china-and-you-could-soon-own-one

… and thousands of industrial aML applications …

Cyber-Physical Systems (CPS):
Tight integration of networked computation with physical systems

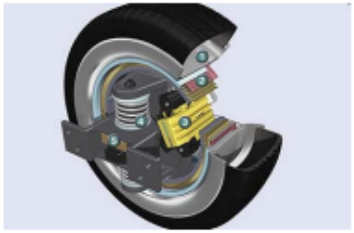
Avionics



Transportation (Air traffic control at SFO)

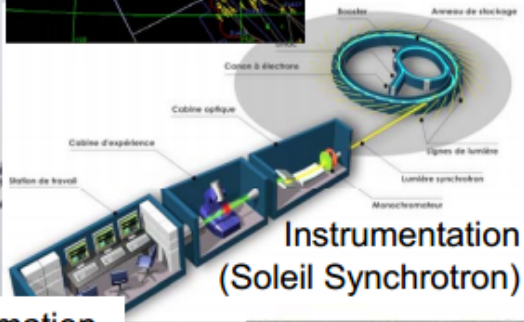
Building Systems

Automotive

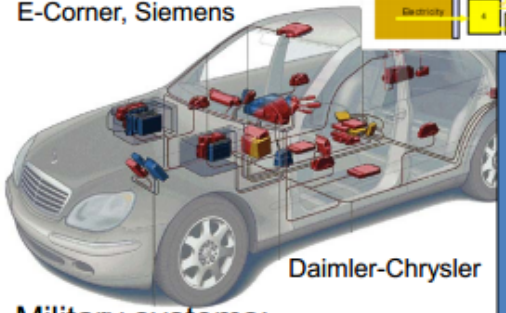


E-Corner, Siemens

Telecommunications



Instrumentation (Soleil Synchrotron)



Daimler-Chrysler

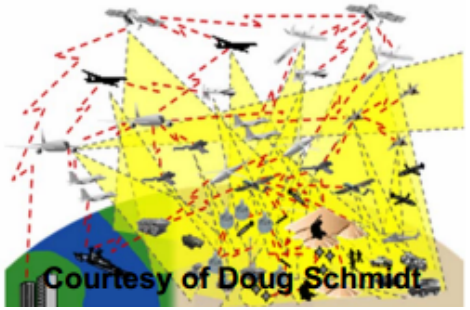
Power generation and distribution

Courtesy of General Electric

Factory automation

Courtesy of Kuka Robotics Corp.

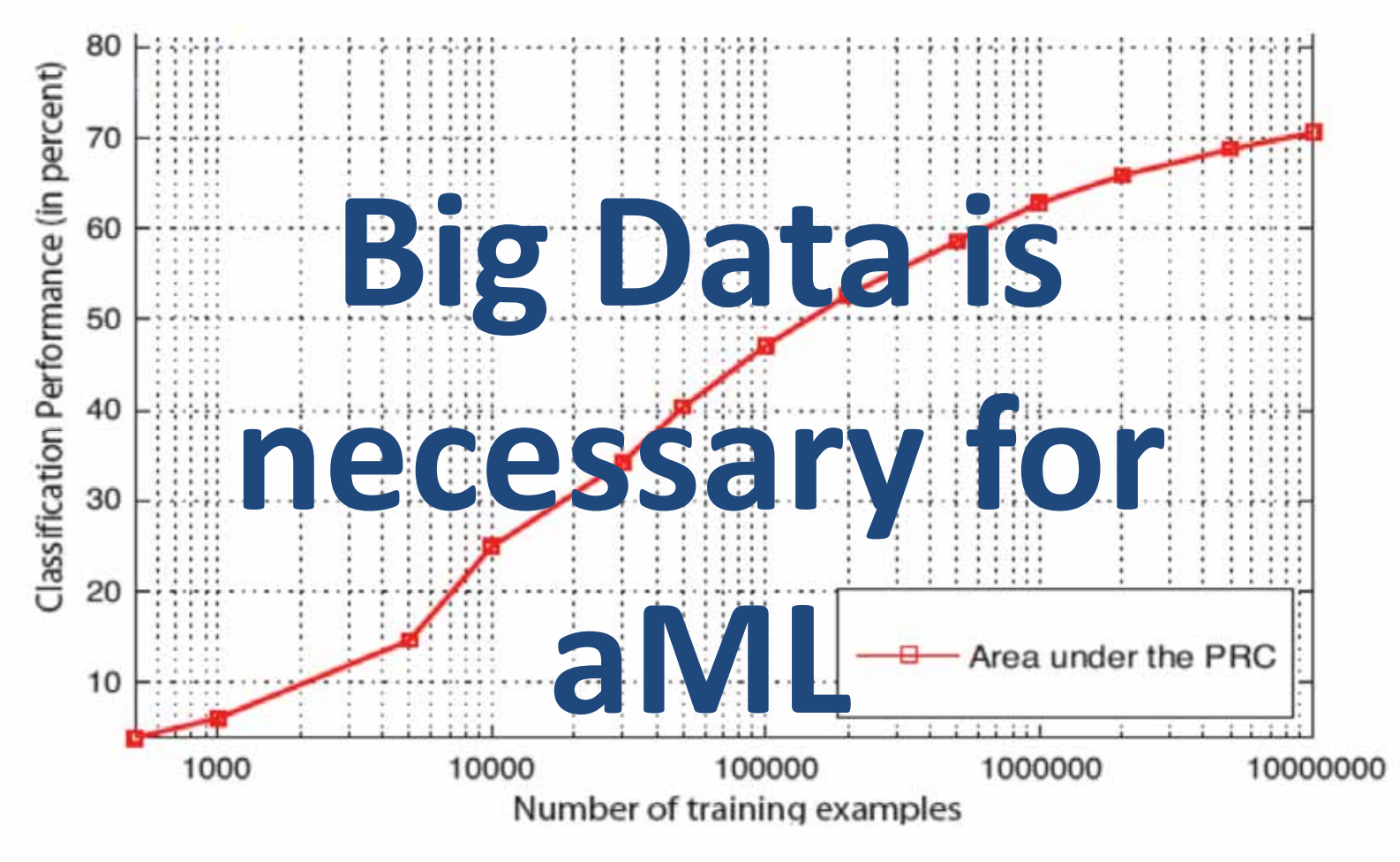
Military systems:



Courtesy of Doug Schmidt

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.

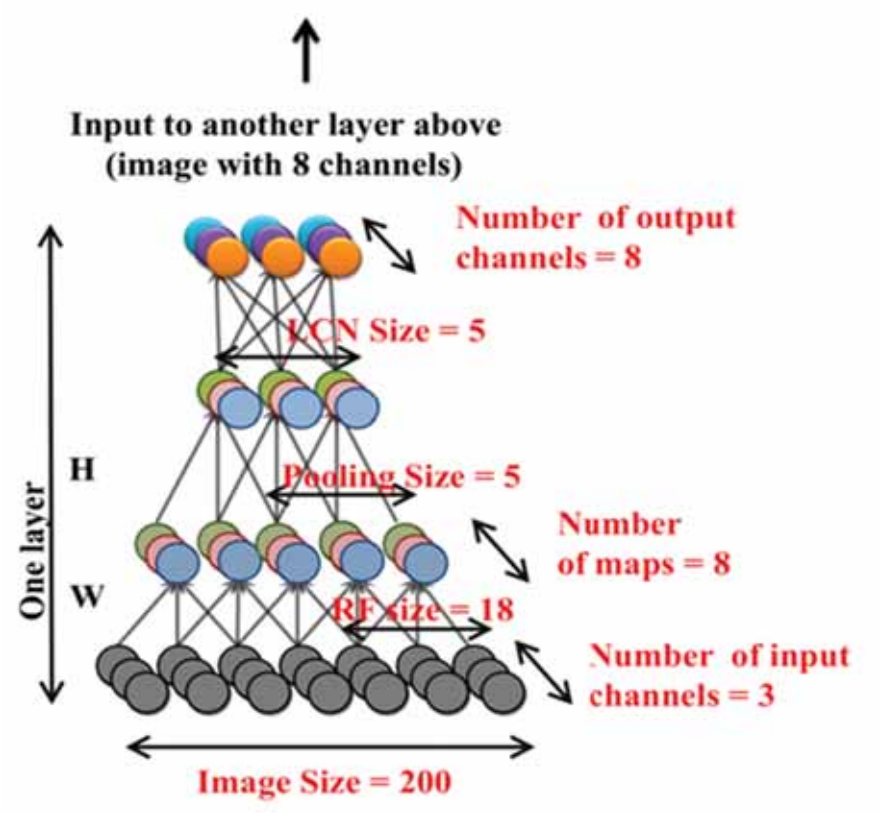
Big Data is necessary for aML !



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 from the 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.

When does aML fail …

- 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, …

Consequently …

05 iML

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

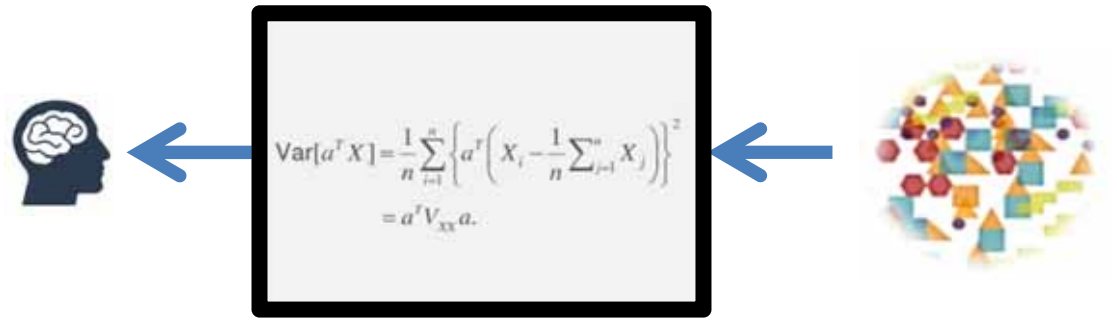
- **Example 1: Subspace Clustering**
- **Example 2: k-Anonymization**
- **Example 3: Protein Design**

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.

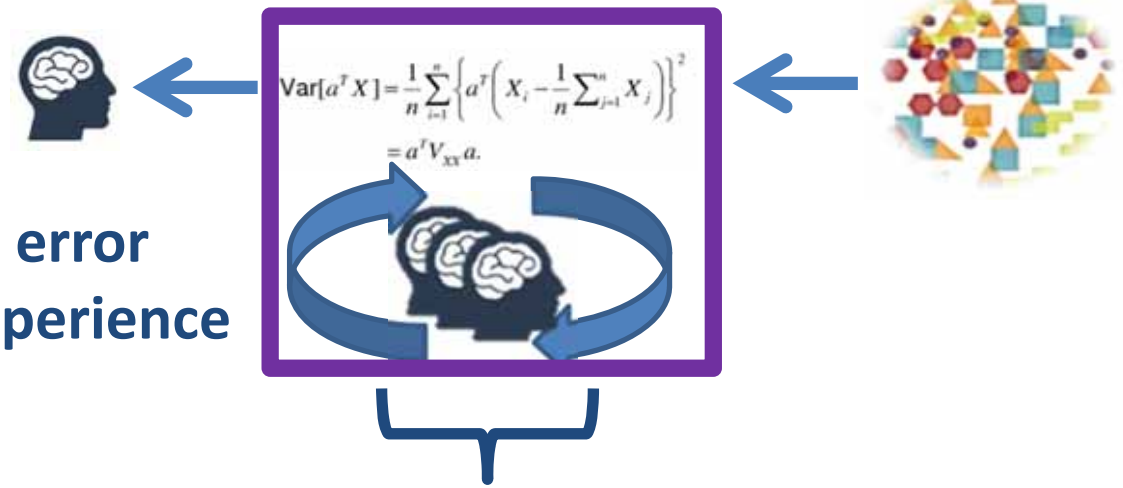
Kieseberg, P., Malle, B., Fruehwirt, P., Weippl, E. & Holzinger, A. 2016. A tamper-proof audit and control system for the doctor in the loop. Brain Informatics, 3, (4), 269–279, doi:10.1007/s40708-016-0046-2.

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.

iML: bringing in human intuition



Generalization error



Generalization error
<u>plus</u> human experience

iML = human inspection – bring in human intuition

Andreas Holzinger et al. 2018. Interactive machine learning: experimental evidence for the human in the algorithmic loop. Springer/Nature Applied Intelligence, doi:10.1007/s10489-018-1361-5.

Why using human intuition?

Humans can generalize from few examples, and …

- understand relevant representations,
- find concepts between $P(x)$ and $P(Y|X)$,
- with a causal link between $Y \rightarrow X$

even Children can make inferences from little, noisy, incomplete data …

Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050

Ok, this is also hard for humans …

See also: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572. Images see: https://imgur.com/a/K4RWn

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.

The state-of-the-art is terrible …

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 dee learning : github.com/karpathy/neuraltalk2

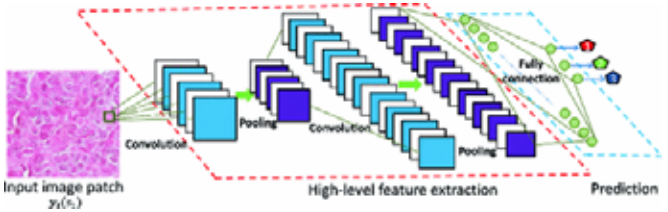
Image Source: Gabriel Villena Fernandez; Agence France-Press, Dave Martin (left to right)

06 Methods of Explainable AI

Why explainable AI?

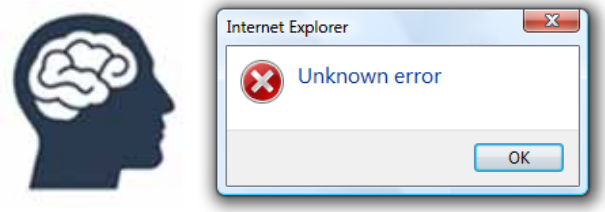
Verify that algorithms/classifiers work as expected …

Wrong decisions can be costly and dangerous …



Understanding the errors …

Detection of bias, weaknesses, unknowns, …



Scientific replicability and causality …

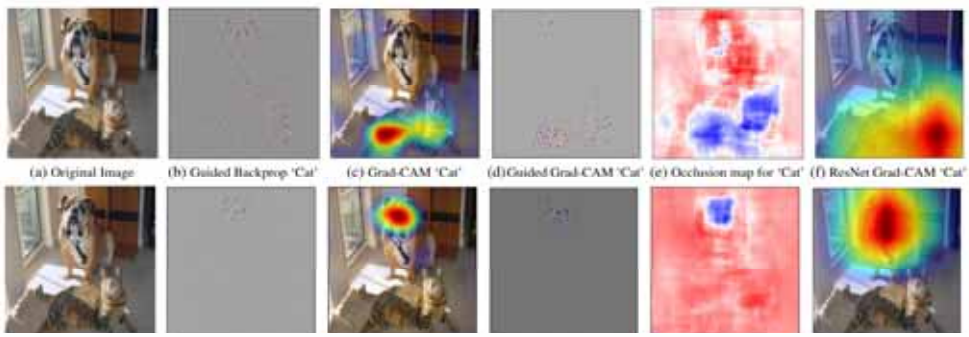
The “why” is often more important than the prediction …

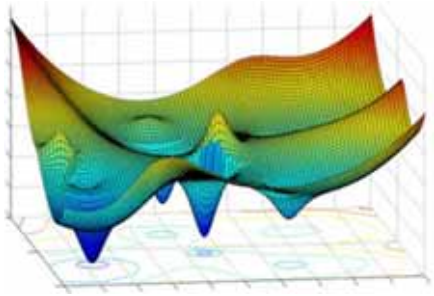
Methods of ex-AI

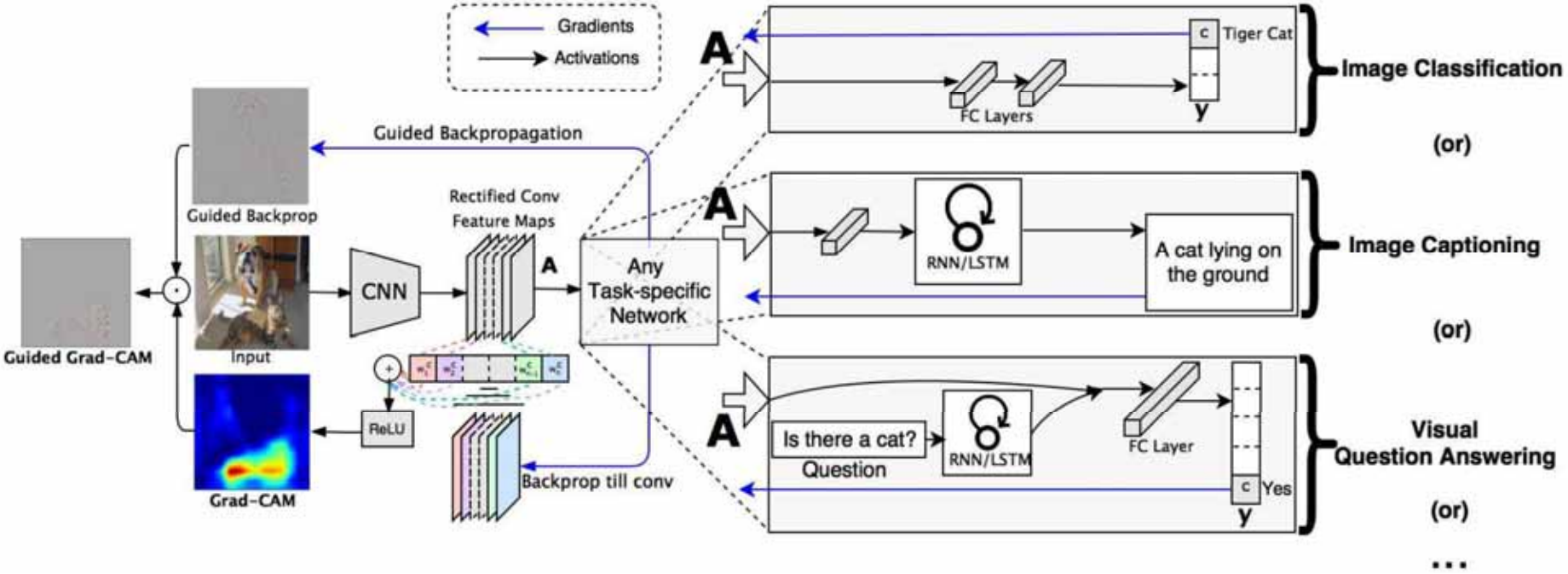
- 1) Gradients
- 2) Sensitivity Analysis
- 3) Decomposition Relevance Propagation (Pixel-RP, Layer-RP, Deep Taylor Decomposition, …)
- 4) Optimization (Local-IME – model agnostic, BETA transparent approximation, …)
- 5) Deconvolution and Guided Backpropagation
- 6) Model Understanding
  - Feature visualization, Inverting CNN
  - Qualitative Testing with Concept Activation Vectors TCAV
  - Network Dissection

Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course at Graz University of Technology
https://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

Example: Gradients

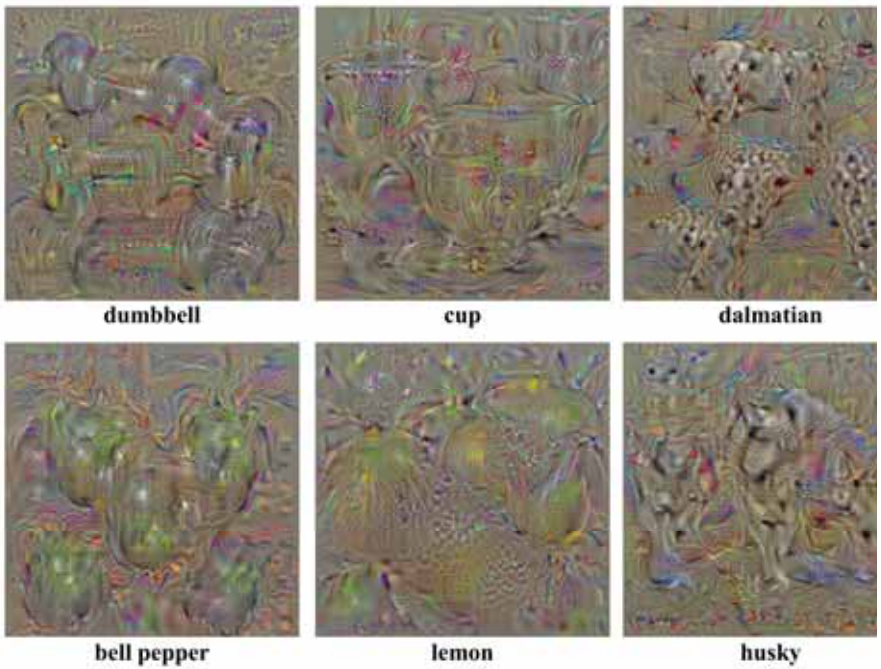


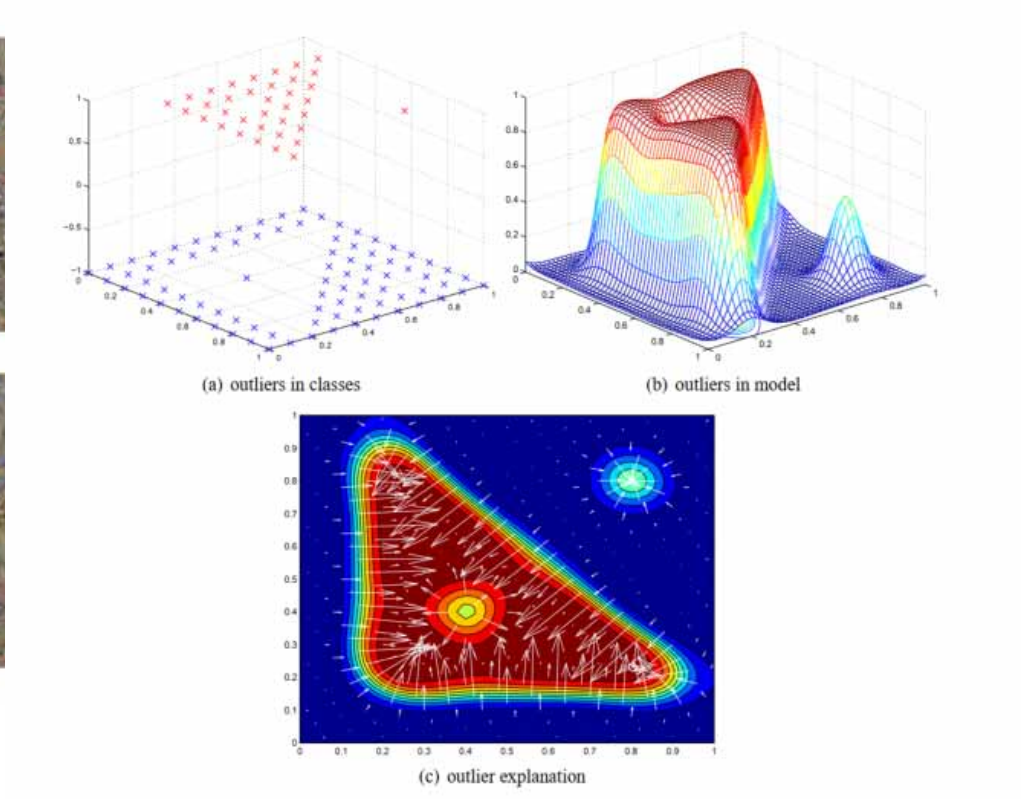




Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. ICCV, 2017. 618-626.

Gradients > Sensitivity Analysis > Heatmapping



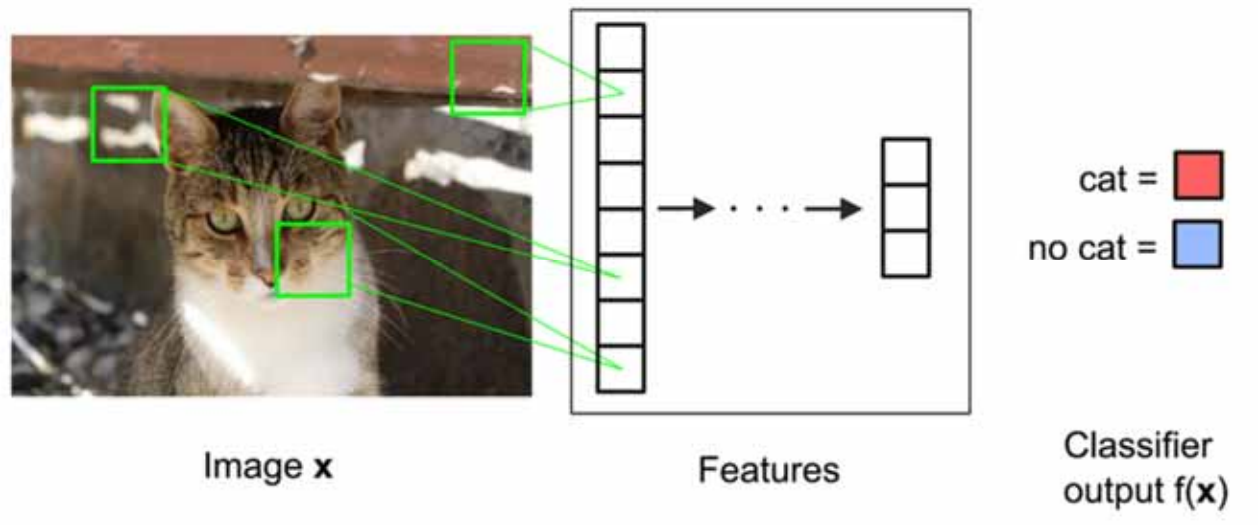


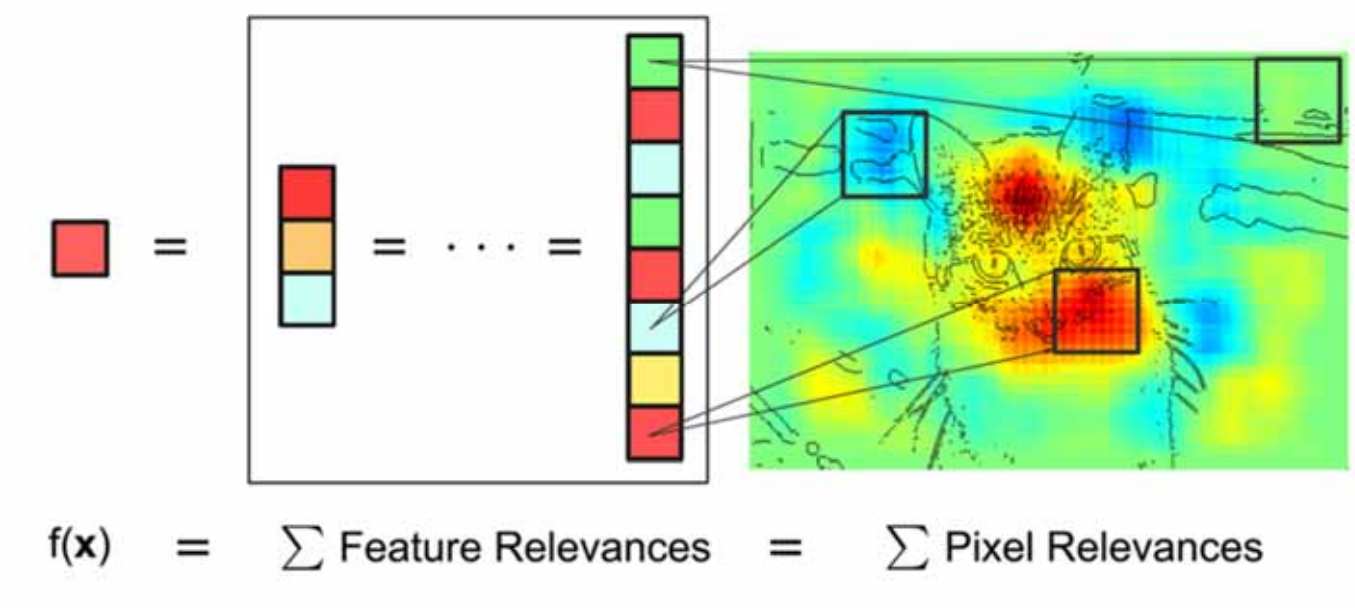
Karen Simonyan, Andrea Vedaldi & Andrew Zisserman 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen & Klaus-Robert Mueller 2010. How to explain individual classification decisions. Journal of machine learning research (JMLR), 11, (6), 1803-1831.

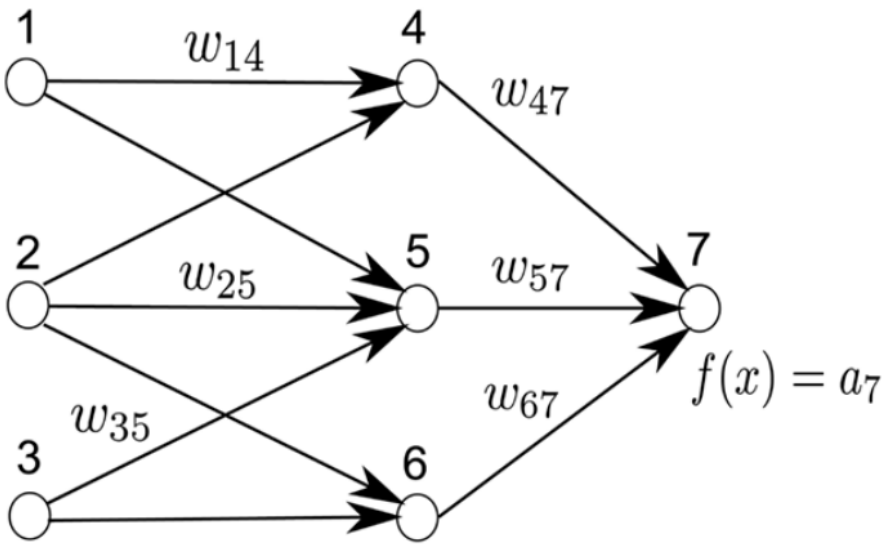
Relevance Propagation

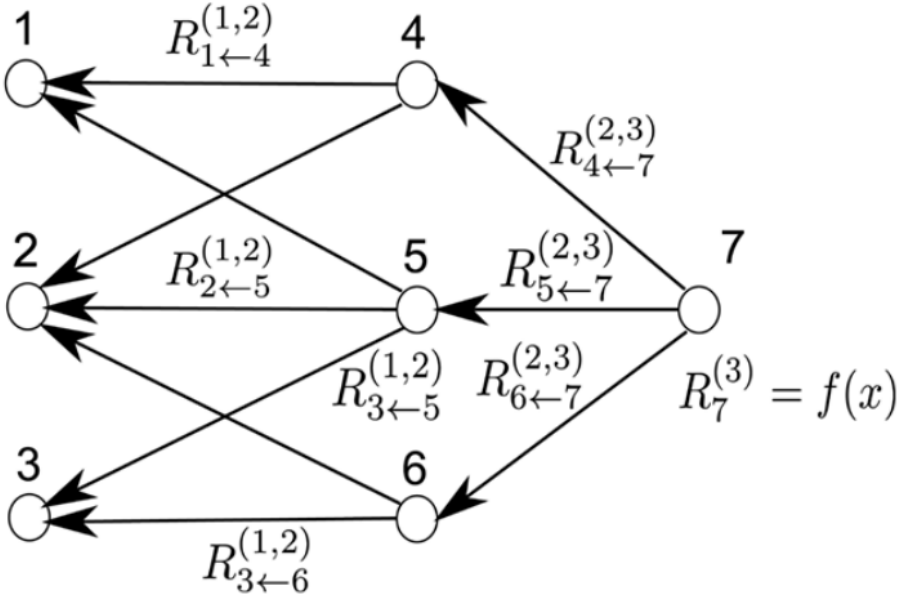
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.





A NN-classifier during prediction time

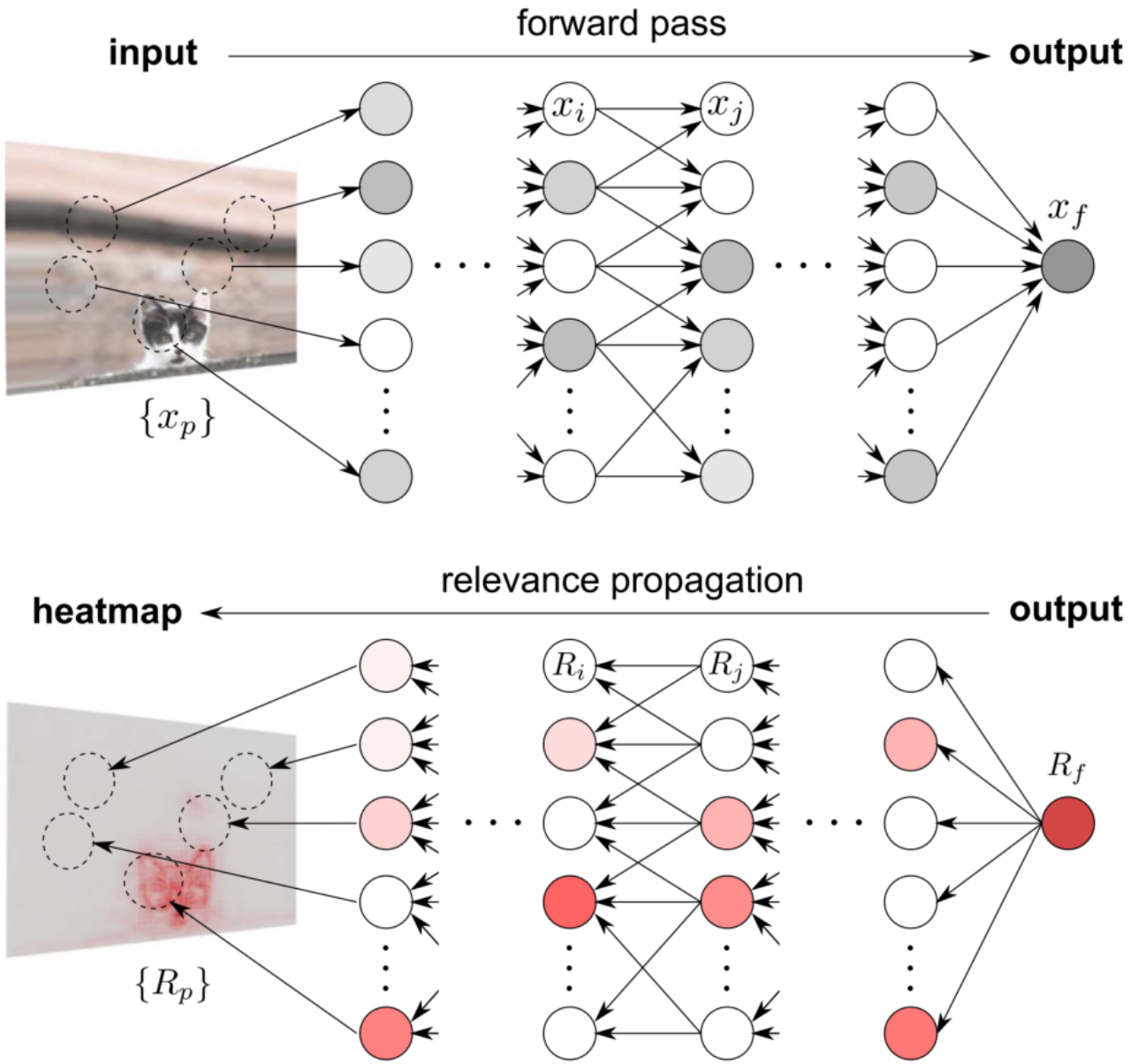




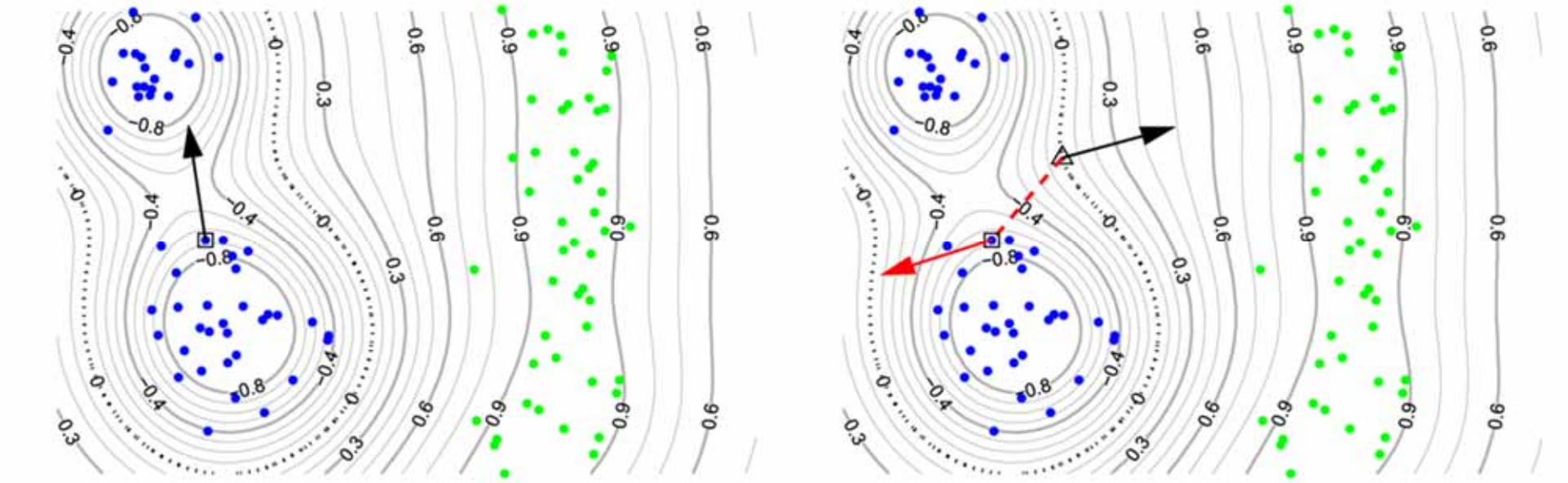
$$f(x) = \cdots = \sum_{d\in l+1} R_d^{(l+1)} = \sum_{d\in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)}$$

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

Relevance propagation

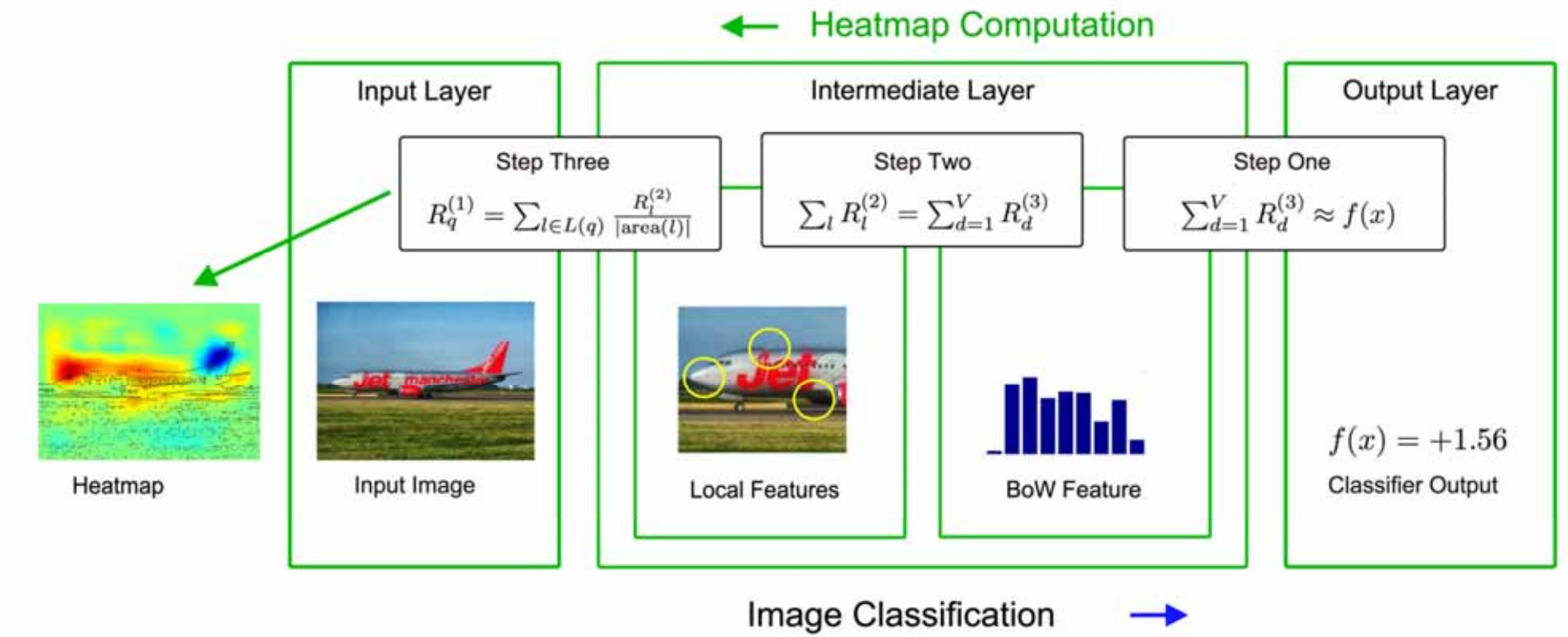


Example Taylor Decomposition



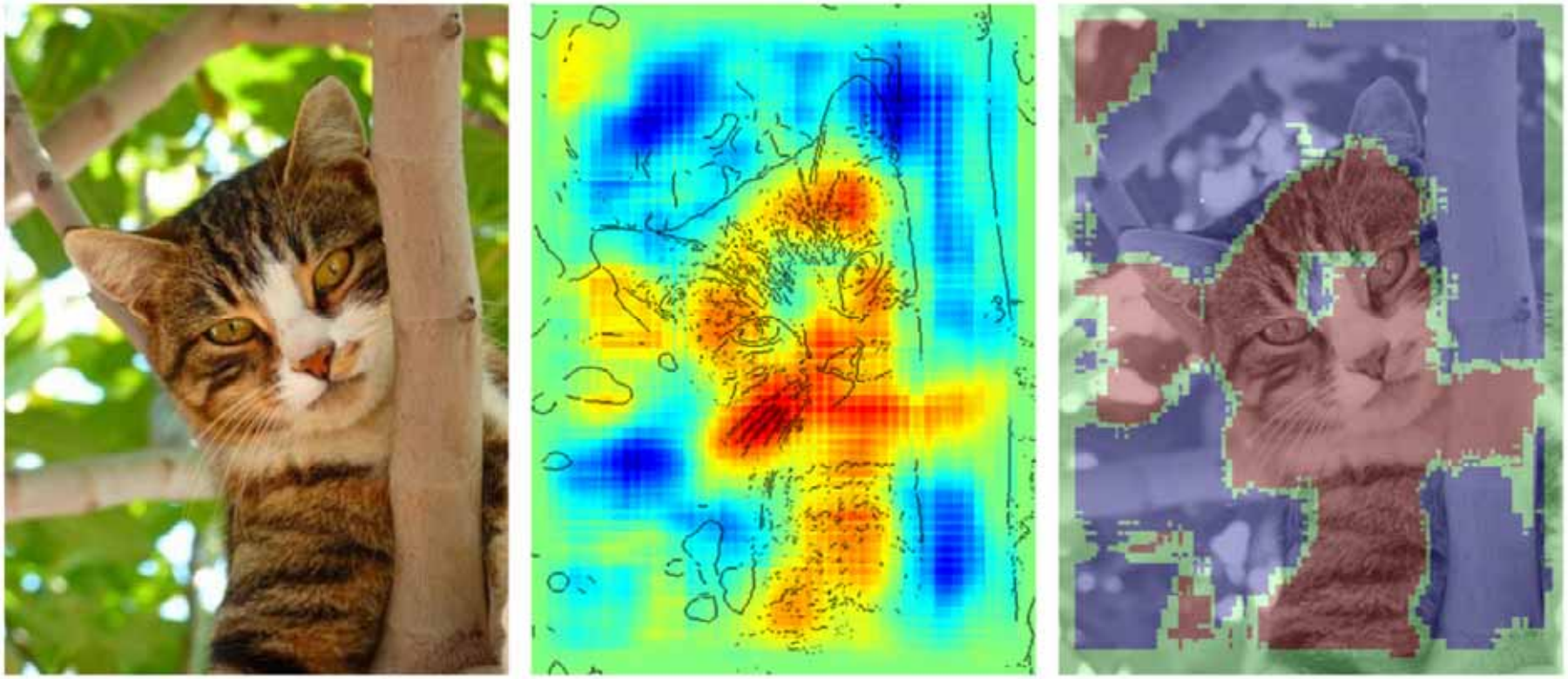
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

Heatmap Computation



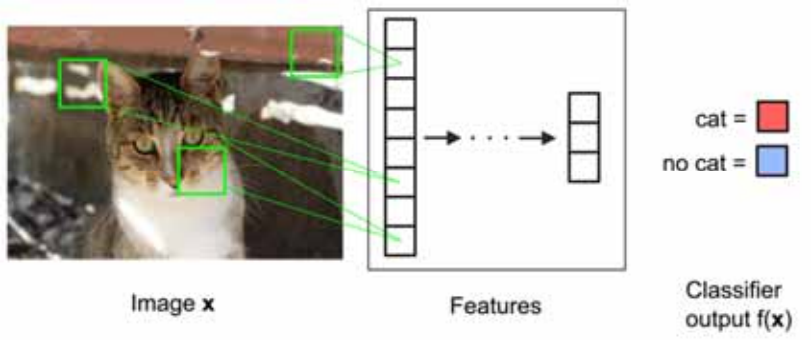
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

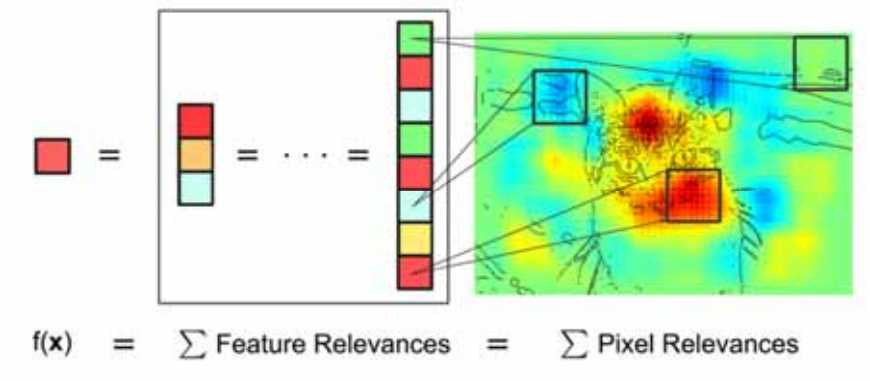
Pixel-wise decomposition for bag-of-words features



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

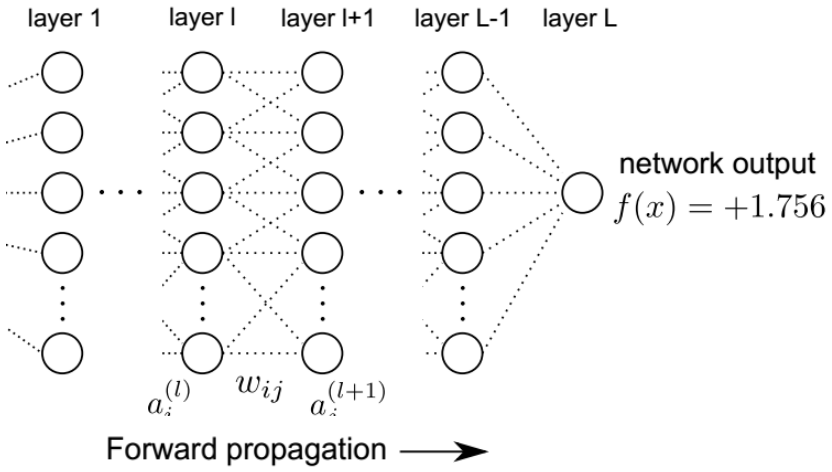
LRP Layer-Wise Relevance Propagation

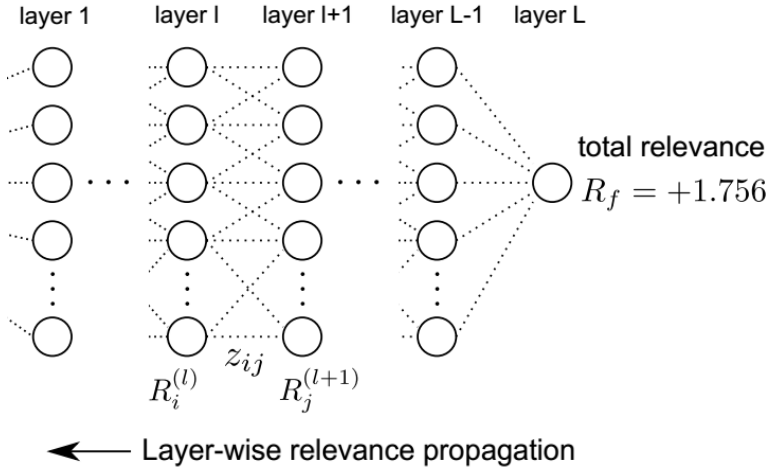




$$a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)$$

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

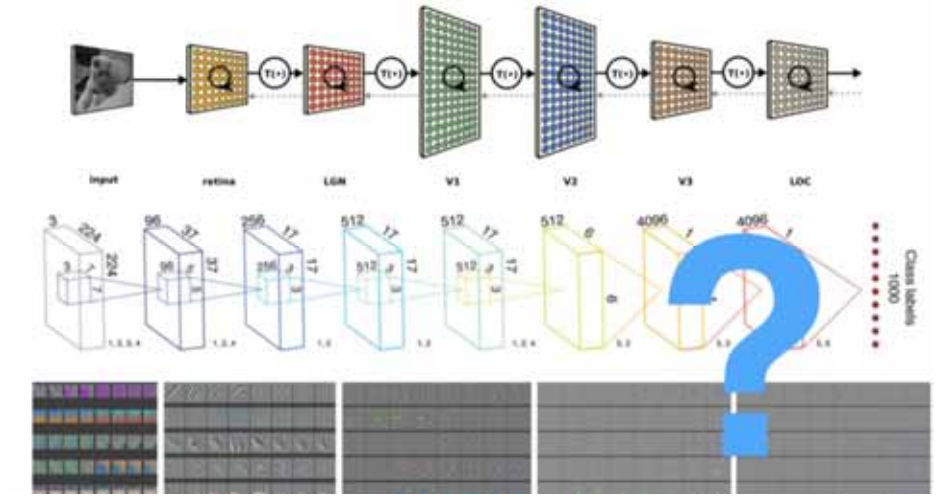




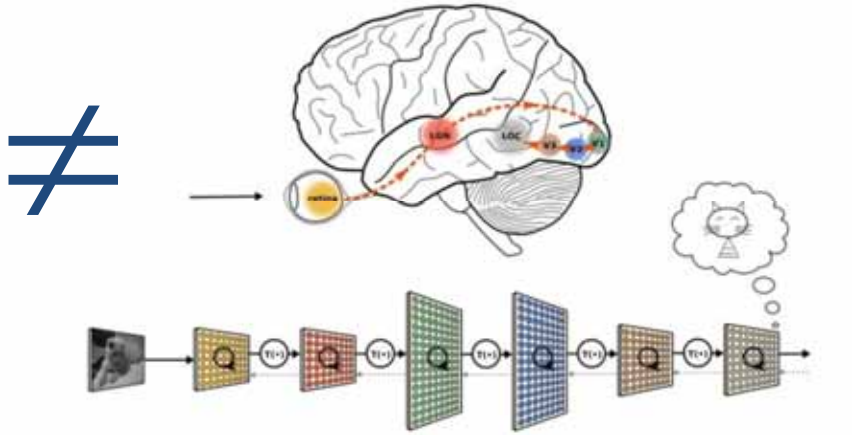
$$R_i = \left|\left|\frac{\partial}{\partial x_i} f(\mathbf{x})\right|\right| \qquad \sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x})$$

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.

Example: Concept Activation Vector (CAV)



Yann Lecun, Yoshua Bengio & Geoffrey Hinton 2015. Deep learning. Nature, 521, (7553), 436-444, doi:10.1038/nature14539.



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

Humans work in another vector space which is spanned by **implicit knowledge** vectors corresponding to an unknown set of human interpretable concepts.

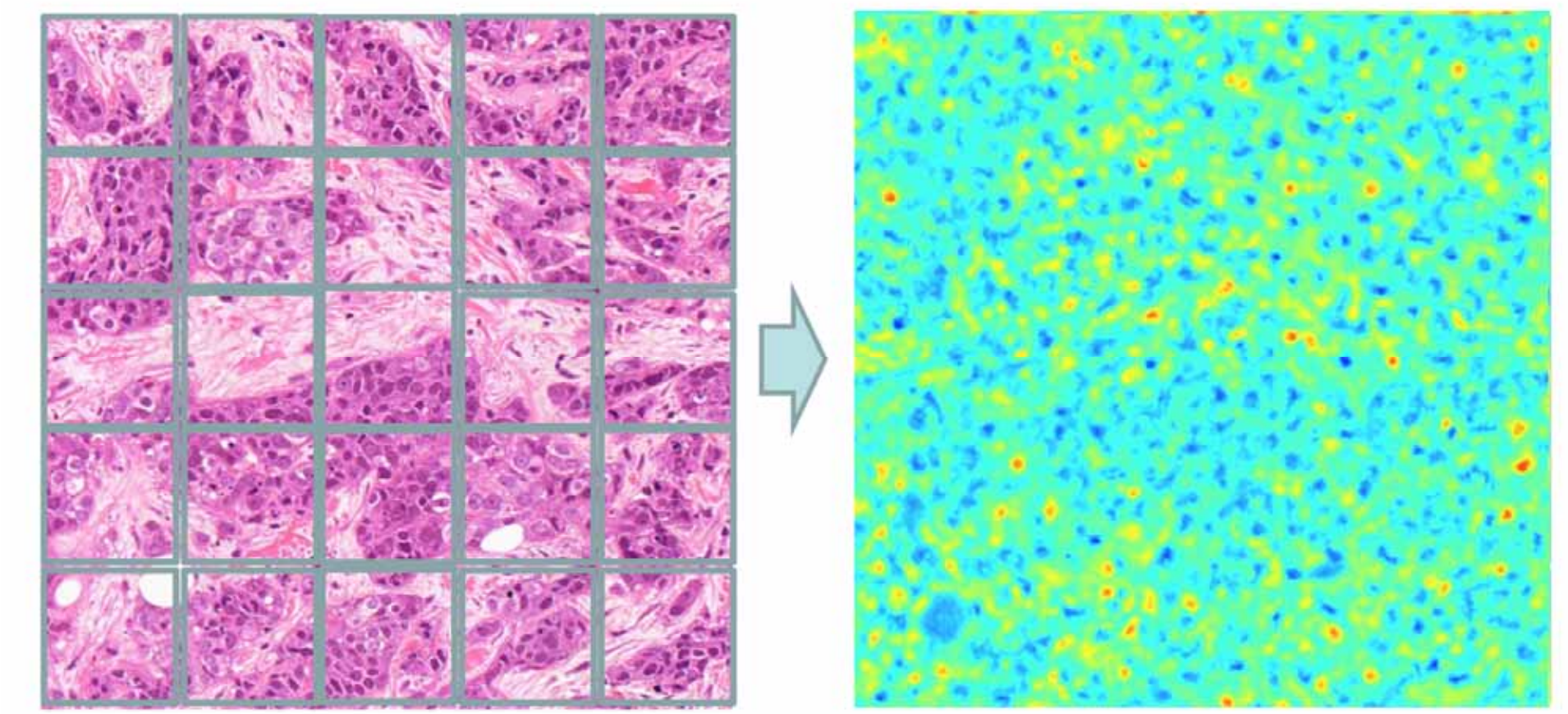
$$S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}\left(f_l(x) + \epsilon v_C^l\right) - h_{l,k}\left(f_l(x)\right)}{\epsilon} = \nabla h_{l,k}\left(f_l(x)\right) \cdot v_C^l$$

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors, ICML, 2018. 2673-2682.

Example 1


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

Example 2 Histopathology

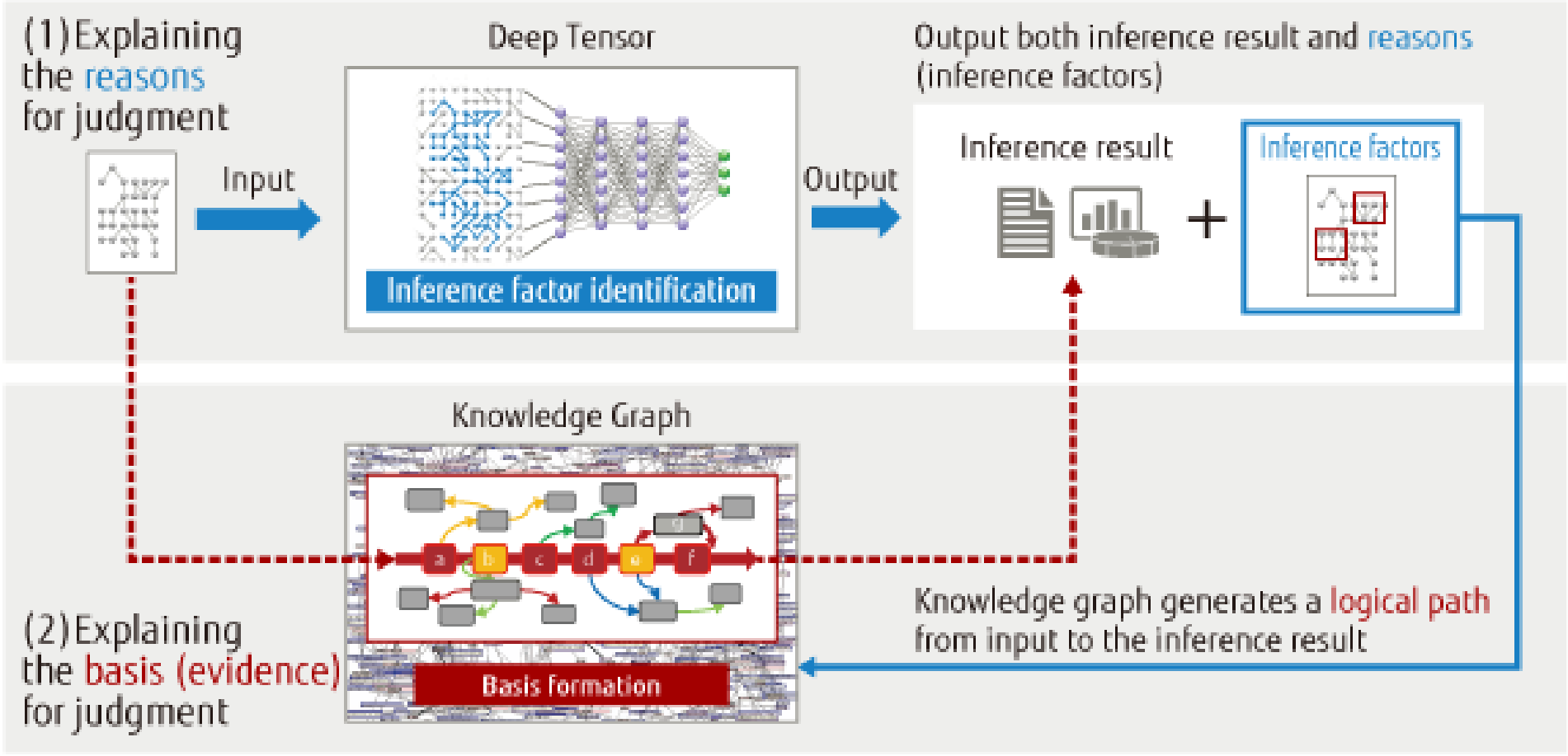


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

07 Towards human interpretable models

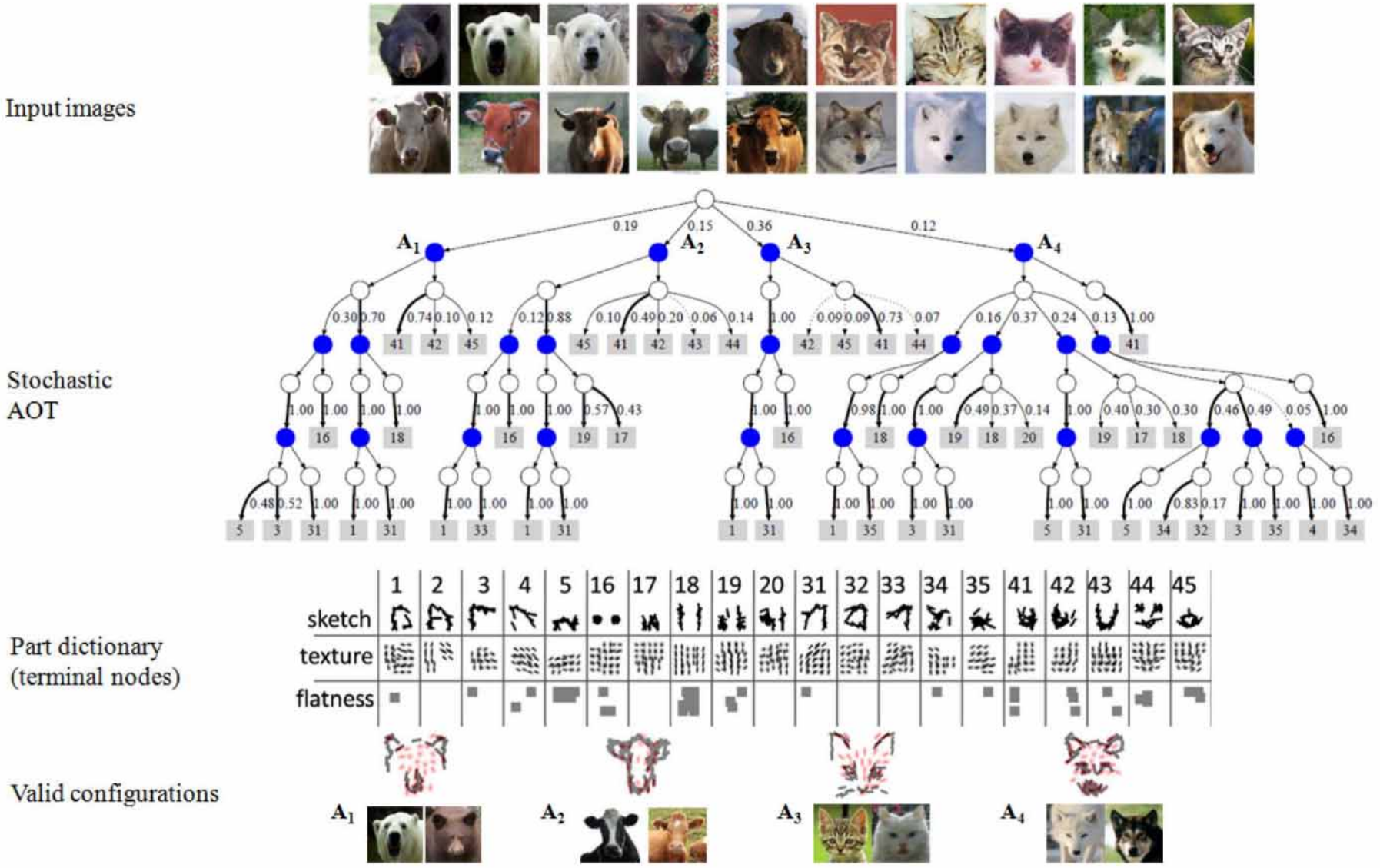
End-users shall be able to retrace the results on demand and we engineers need to understand our own machine learning models!

Combination Probabilistic + Logic approaches



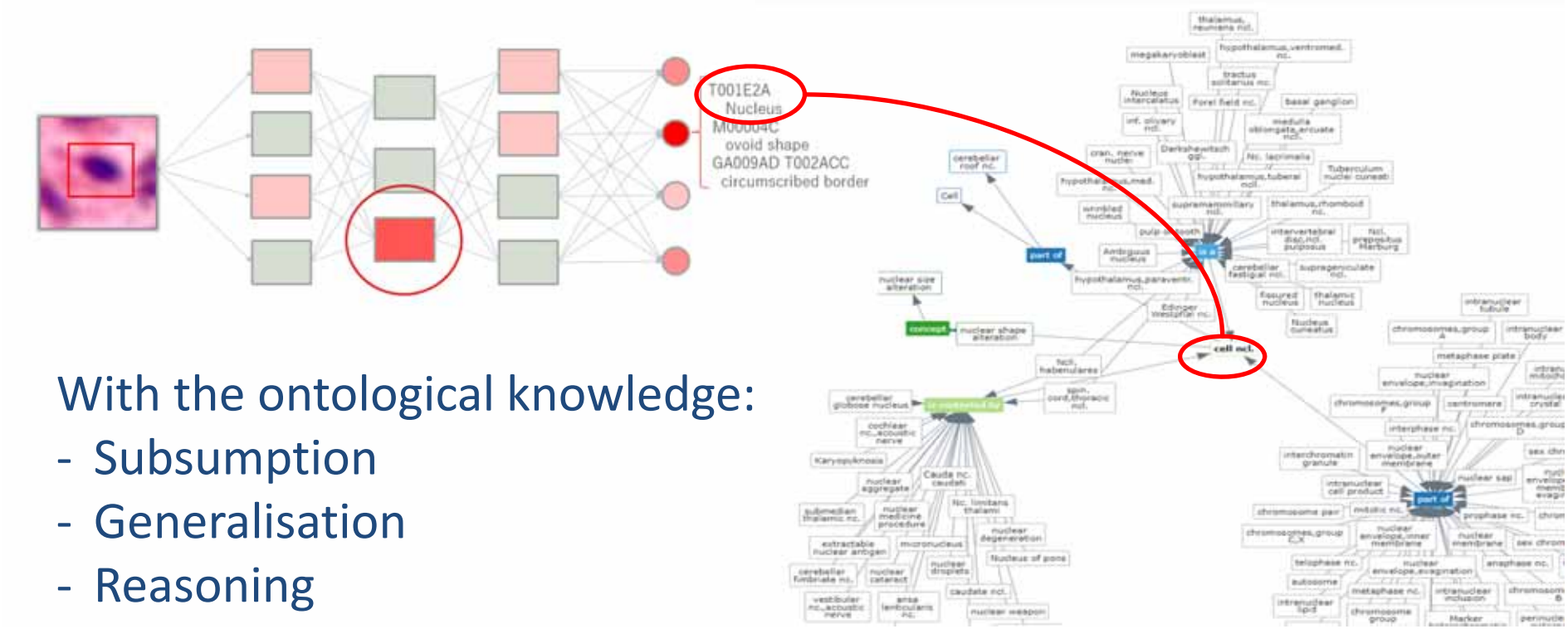
Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger 2018. Explainable AI: the new 42? Springer Lecture Notes in Computer Science LNCS 11015

Stochastic AND-OR Templates for visual objects



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

Connect Terminology and Ontology

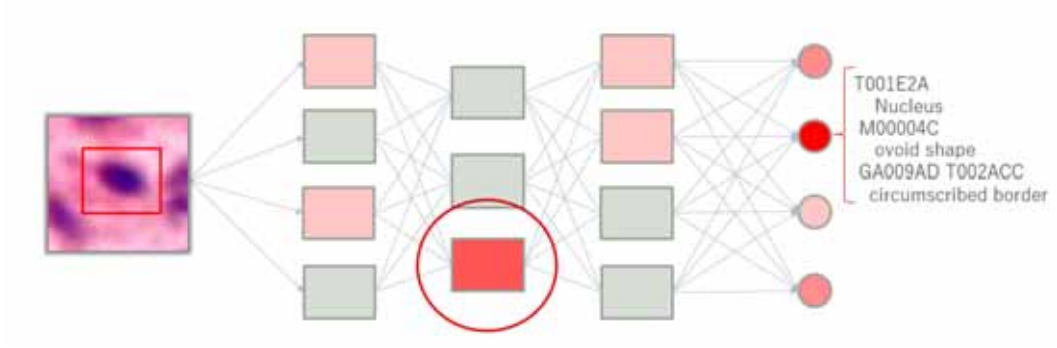


With the ontological knowledge:

- Subsumption
- Generalisation
- Reasoning
- Validation
- Context !

Image Source: unpublished, from our current Project with ID-Berlin

Which nodes have which relevant task?
Which sematic categories have been used at all?



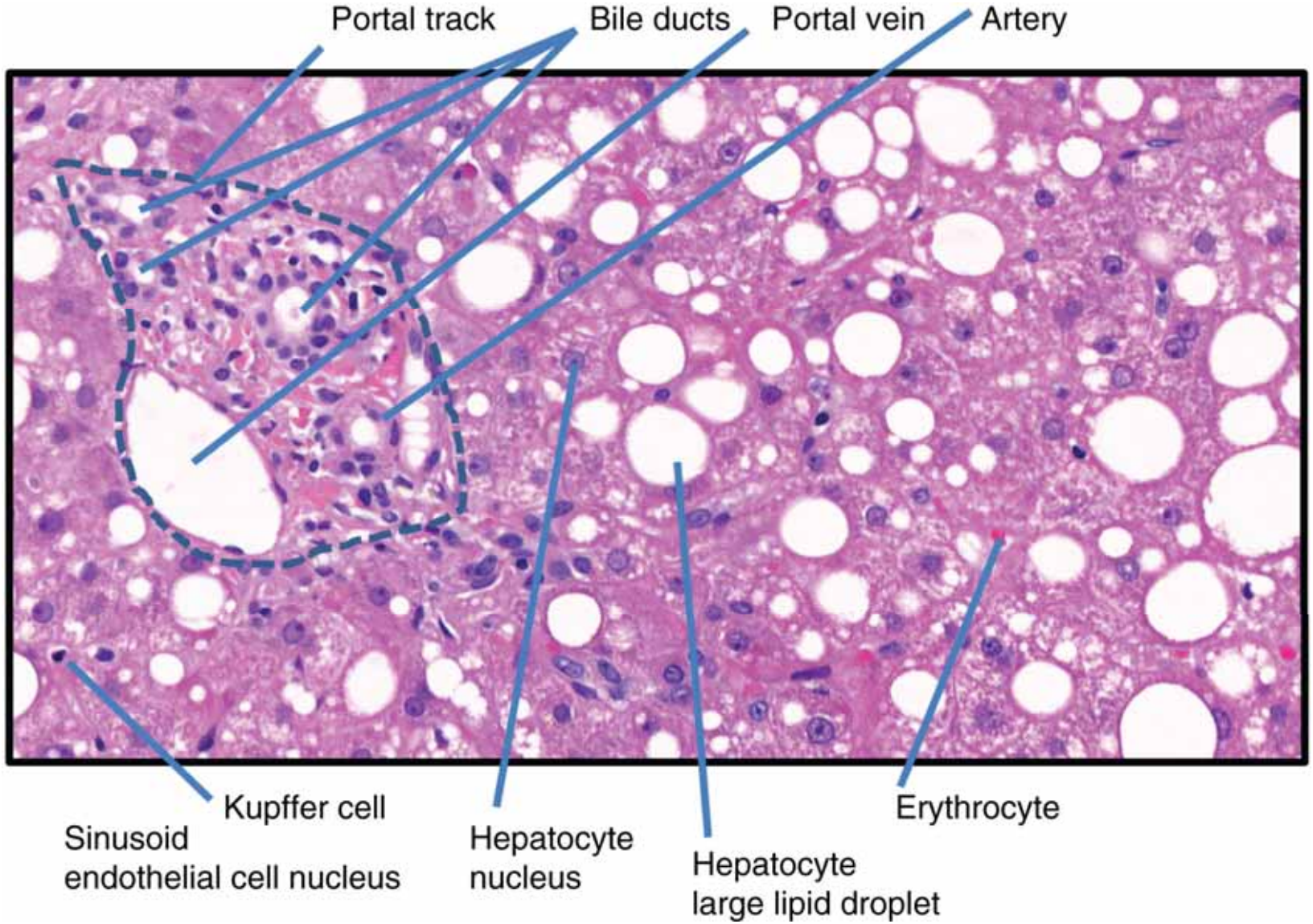
Moreover, ontologies can be used to

- analyze the internal structure
- describe the network
- filter nodes

Image Source: unpublished, from our current Project with ID-Berlin

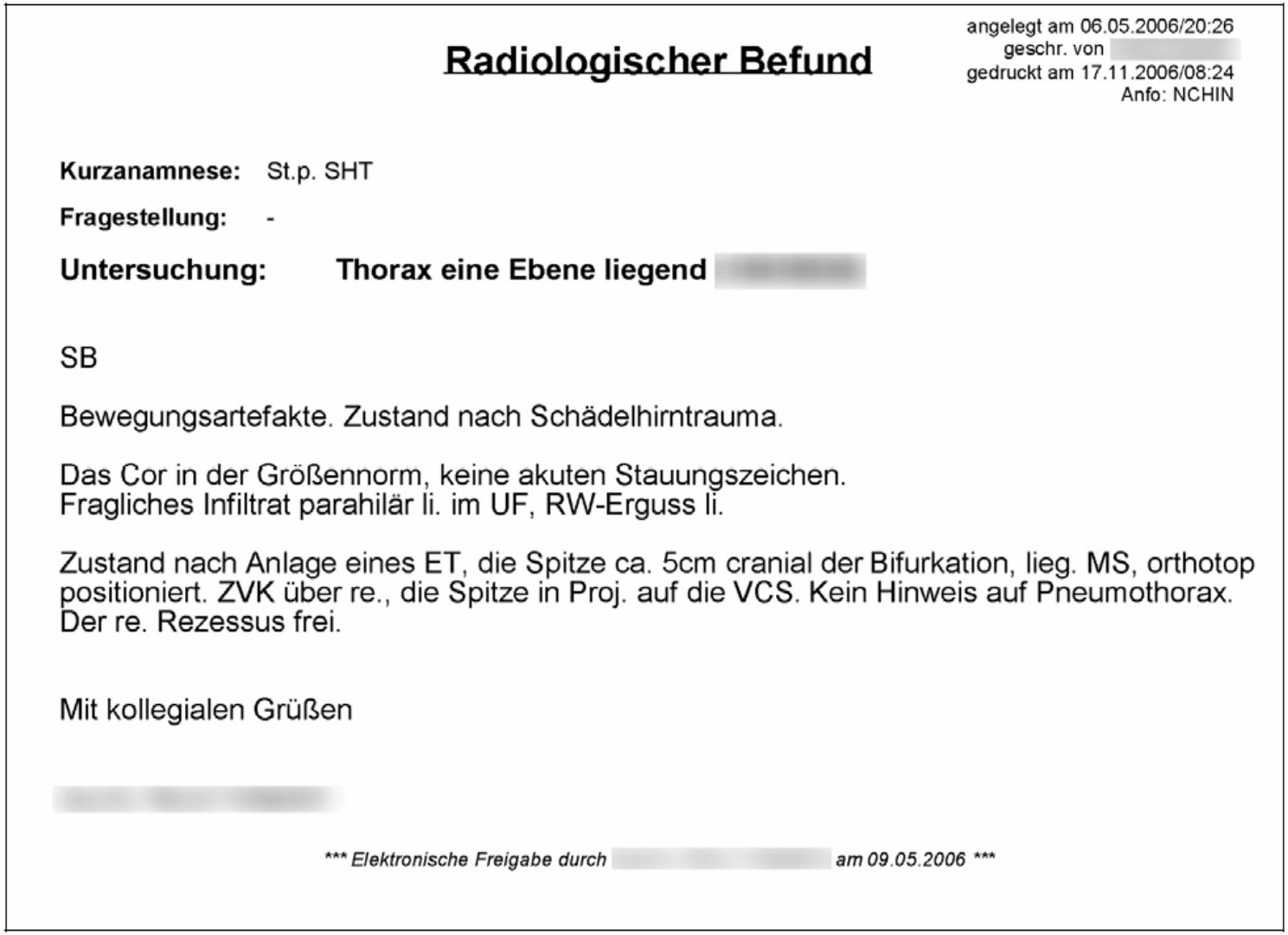
08 Measuring Machine Intelligence KANDINSKY-Project

Motivation: Typical Task of a human Pathologist …



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

The medical report is legally binding – not the image



Radiologischer Befund

angelegt am 06.05.2006/20:26
geschr. von [redacted]
gedruckt am 17.11.2006/08:24
Anfo: NCHIN

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: **Thorax eine Ebene liegend** [redacted]

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

Das Cor in der Größennorm, keine akuten Stauungszeichen.
Fragliches Infiltrat parahilär li. im UF, RW-Erguss li.

Zustand nach Anlage eines ET, die Spitze ca. 5cm cranial der Bifurkation, lieg. MS, orthotop positioniert. ZVK über re., die Spitze in Proj. auf die VCS. Kein Hinweis auf Pneumothorax. Der re. Rezessus frei.

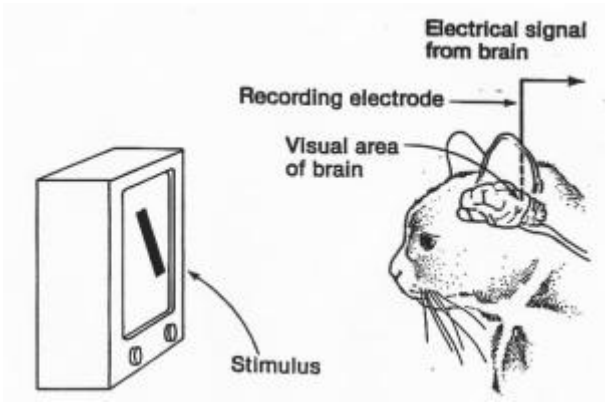
Mit kollegialen Grüßen

[redacted]

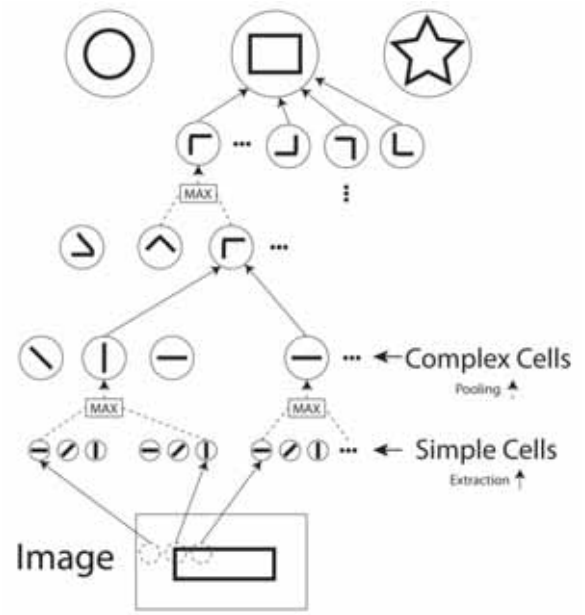
**** Elektronische Freigabe durch* [redacted] *am 09.05.2006 ****

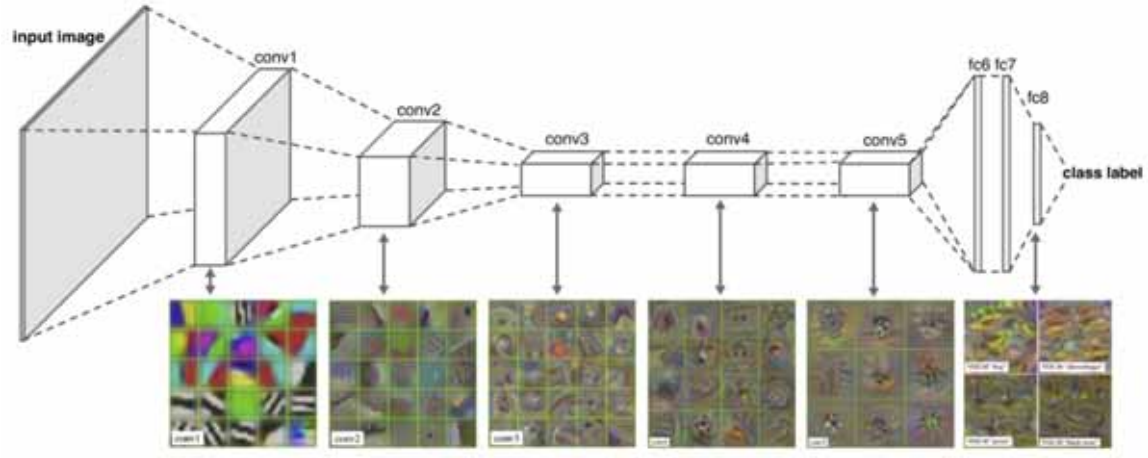
Holzinger, A., Geierhofer, R. & Errath, M. 2007. Semantische Informationsextraktion in medizinischen Informationssystemen. *Informatik Spektrum, 30, (2), 69-78.*

Our world is compositional …

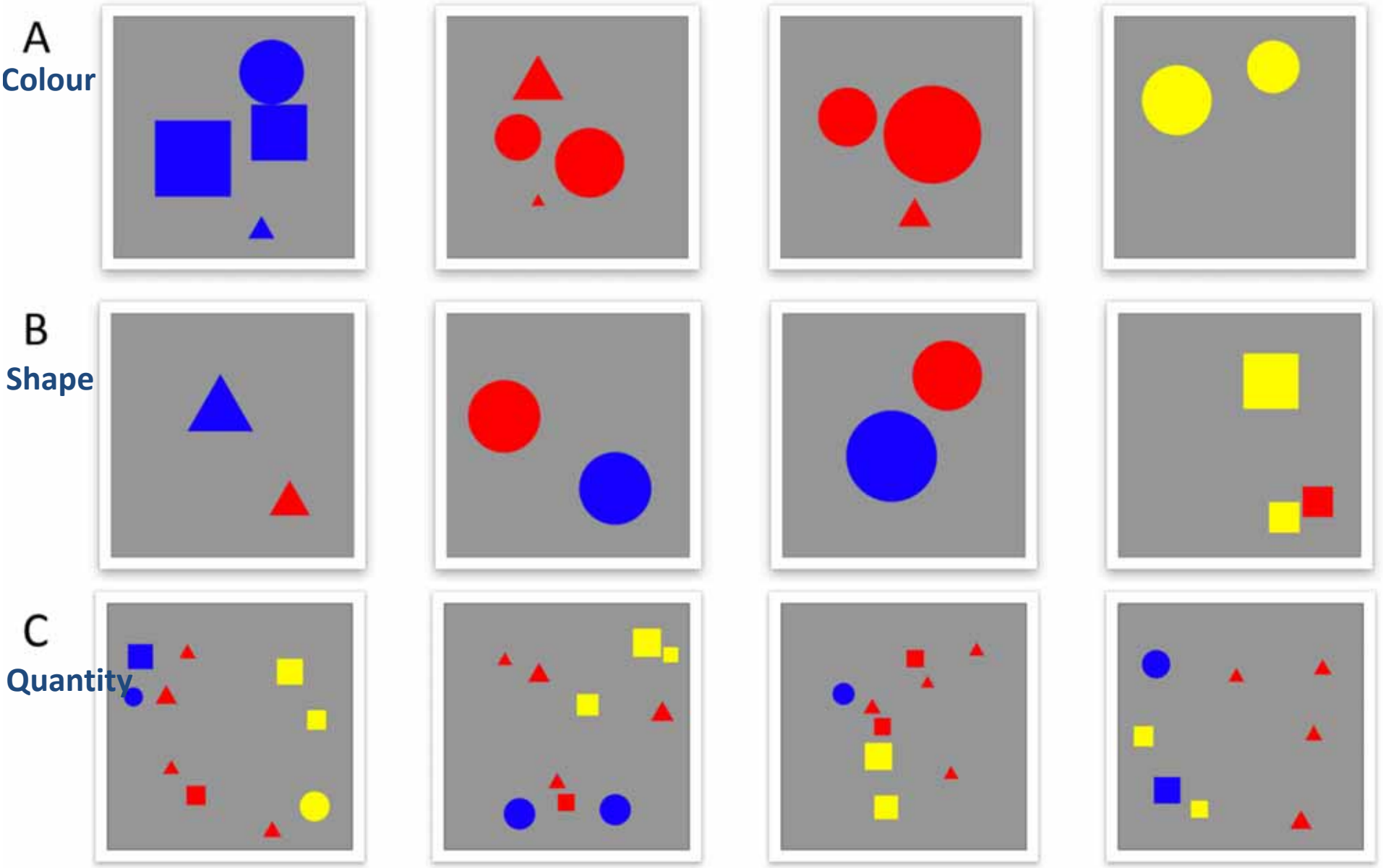


David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837.





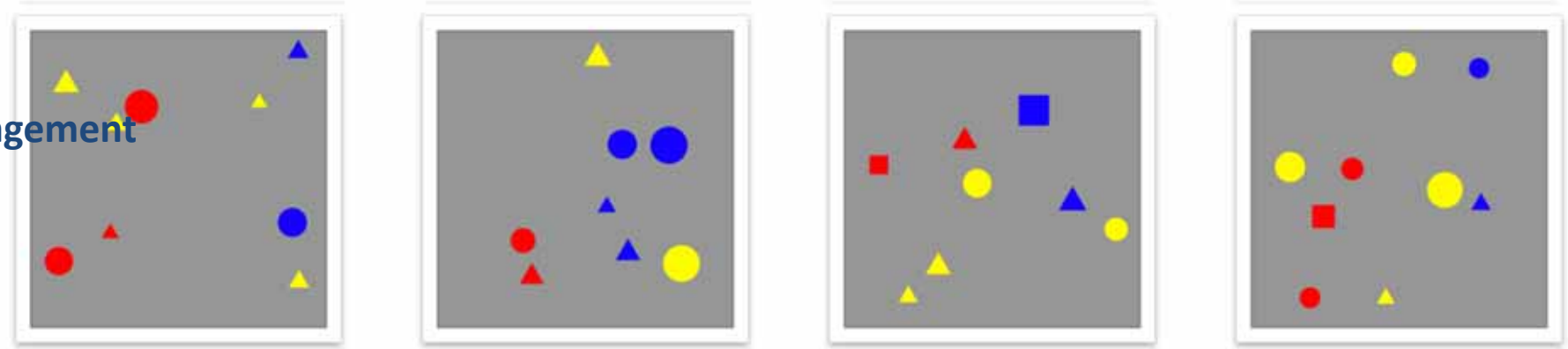
How do humans explain? They use concepts!



How do humans explain? (2/2)

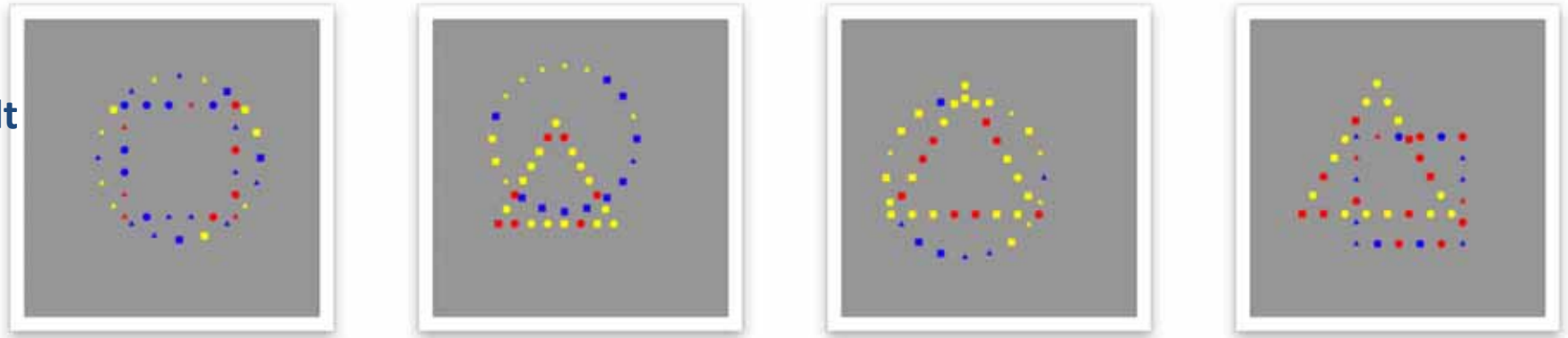
D

Arrangement



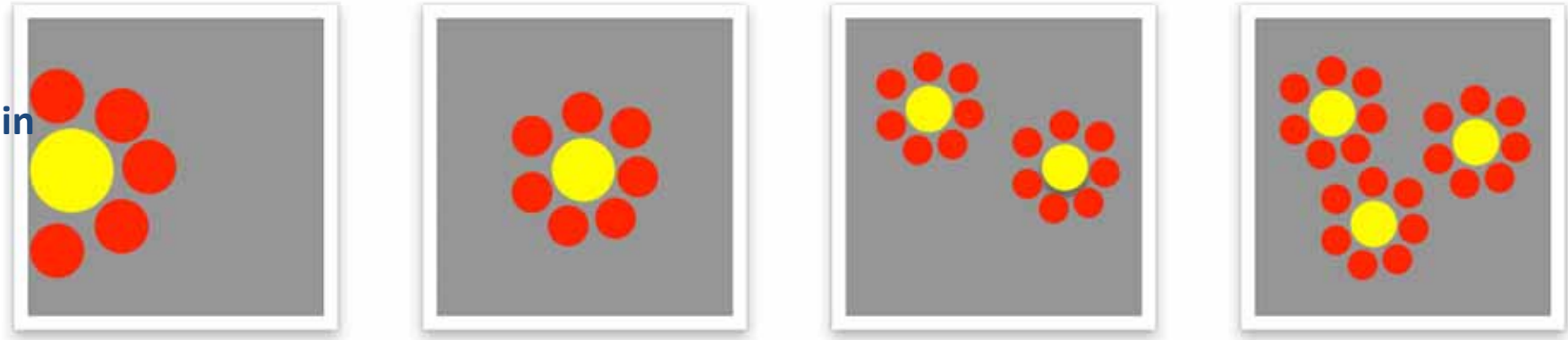
E

Gestalt



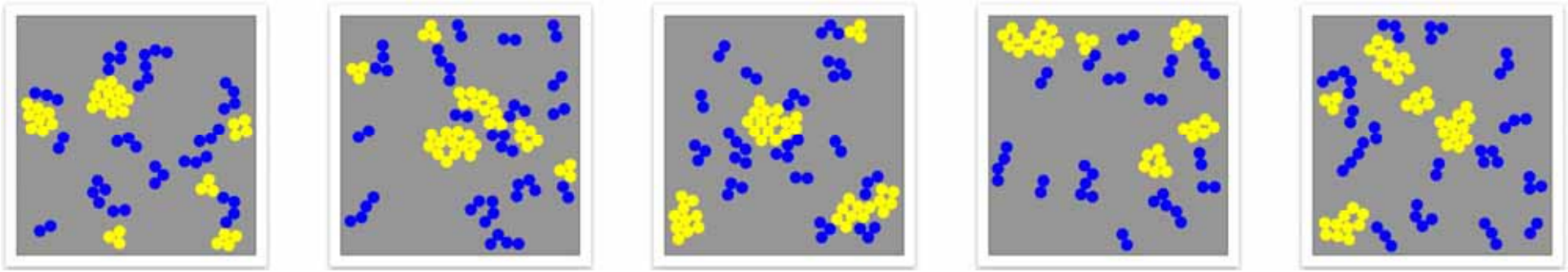
F

Domain

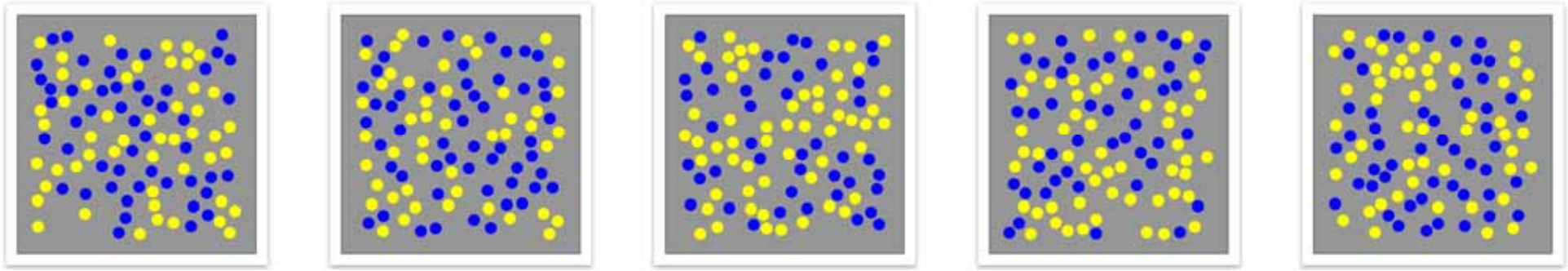


Domain Concept from Pathology

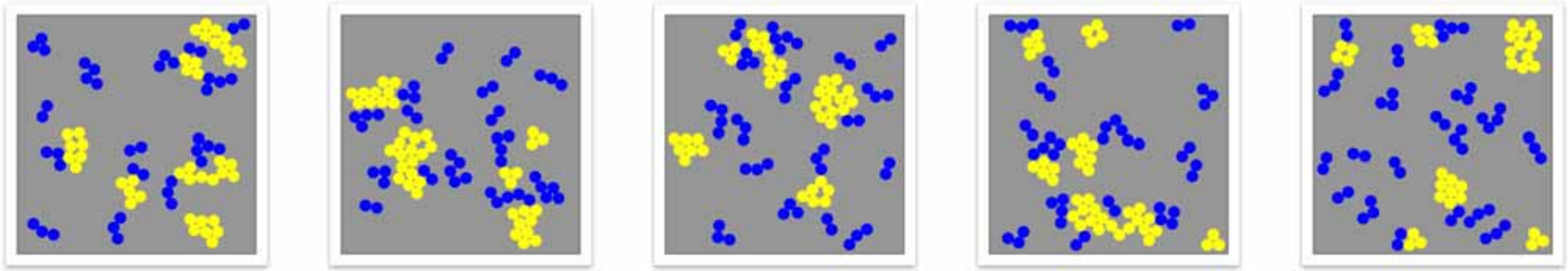
A) True



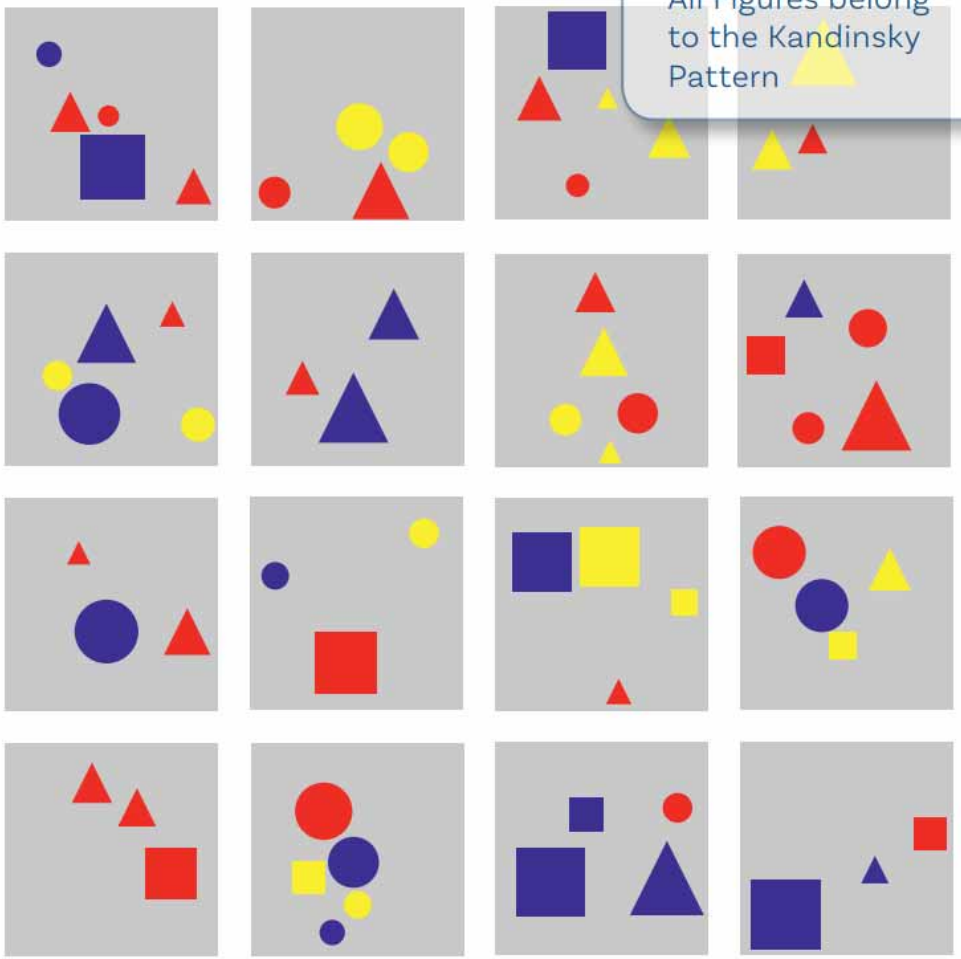
B) False



C) Counterfactual



Part of the pattern



All Figures belong to the Kandinsky Pattern

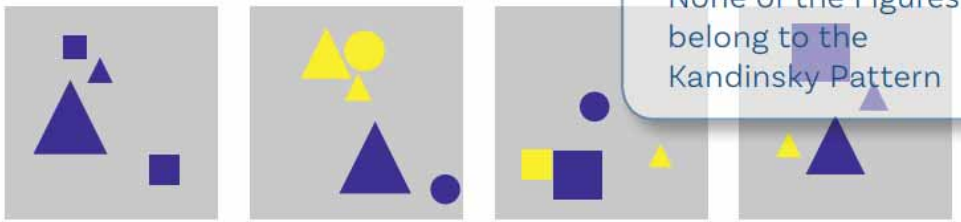
Hypothesis 1

It only contains circles and triangles.

Hypothesis 2

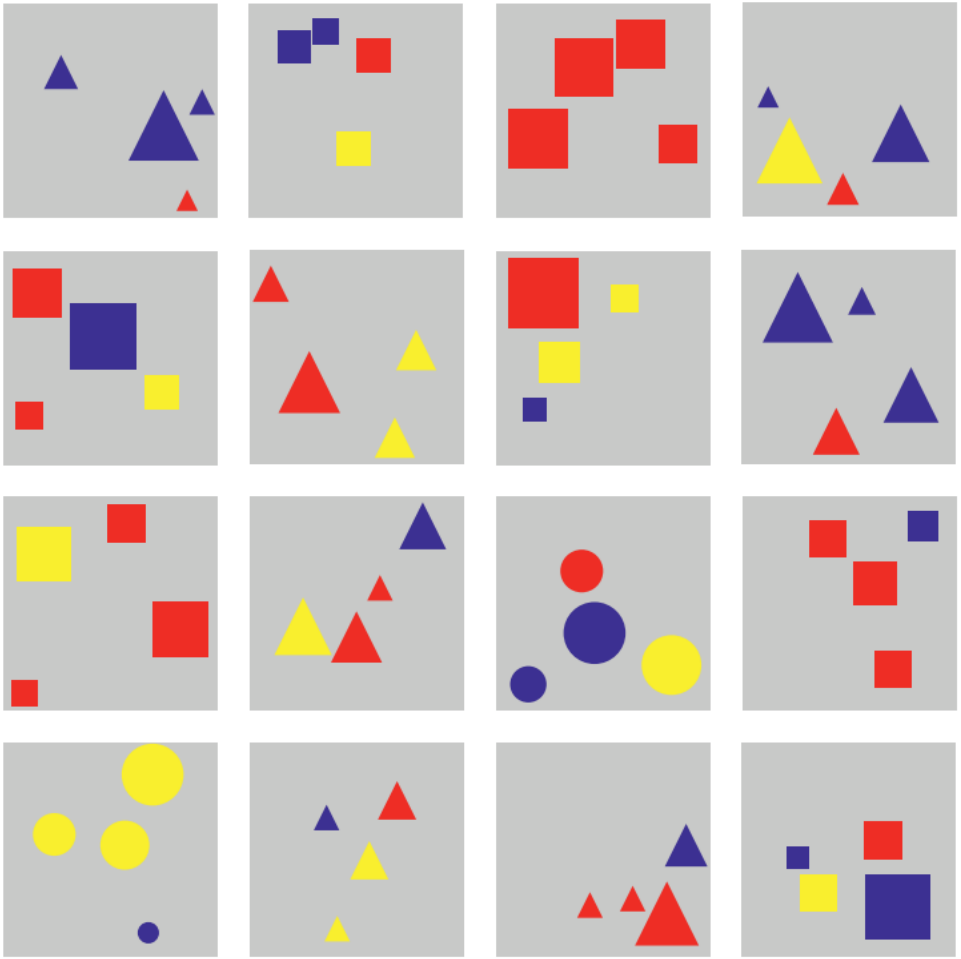
It contains at least a red object.

Not part of the pattern



None of the Figures belong to the Kandinsky Pattern

Part of the pattern



Not part of the pattern

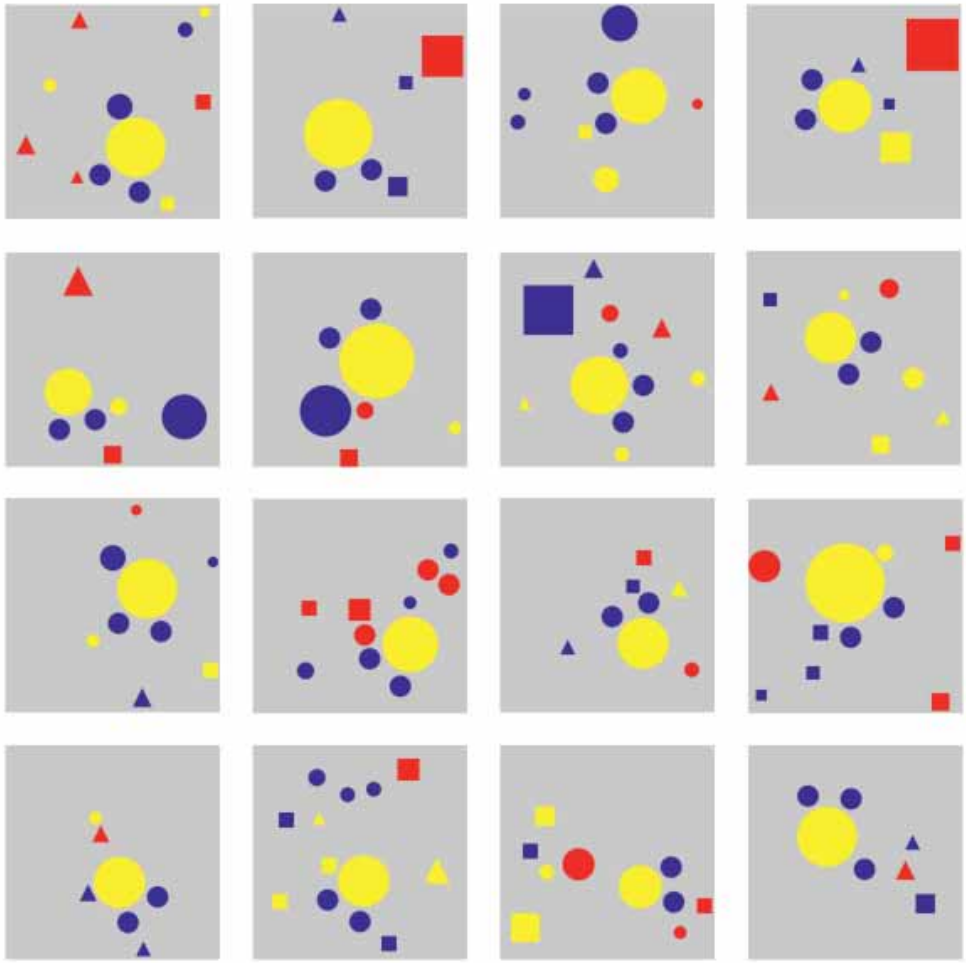
S2 Basic Pattern 2

Title: **All of Same Shape** ->

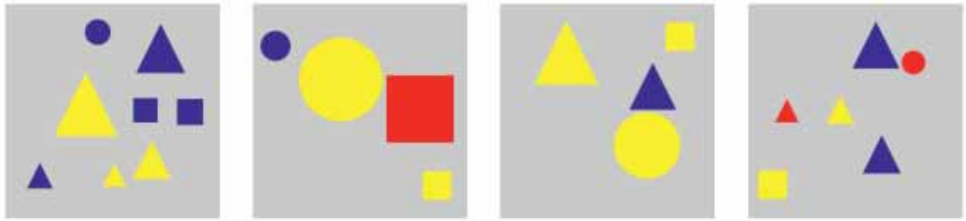
All objects have the same shape.

Hint: Don't be distracted by the colors

Part of the pattern



Not part of the pattern



S8 Basic Pattern 8

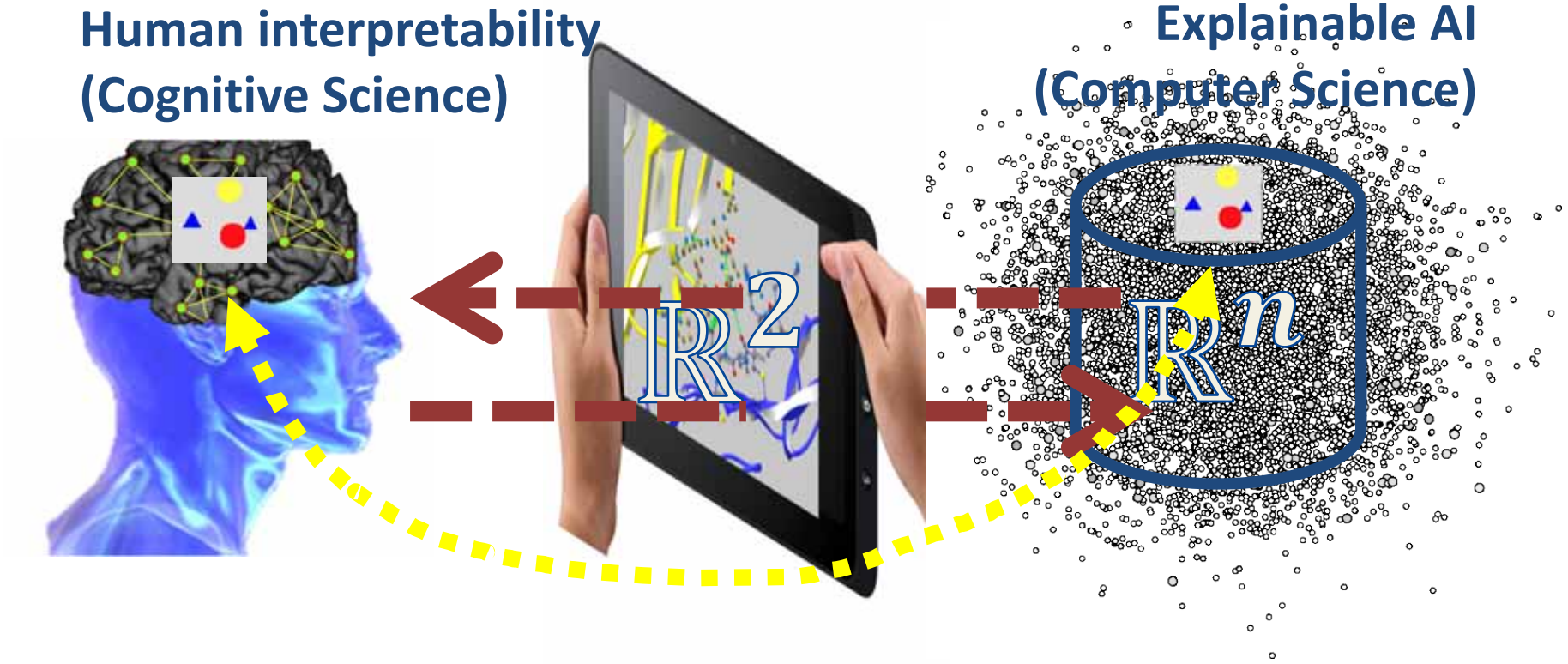
Title: **Mickey Mouse** ->

Every figure contains a pattern which is made out of a big yellow circle and two smaller blue ones and looks like a Mickey Mouse.

Conclusion

Explainable AI needs effective concept mapping !!!

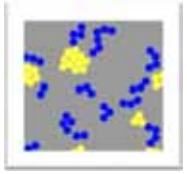
- **Causability := a property of a person (Human)**
- **Explainability := a property of a system (Computer)**



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

Take part in the explainable-AI challenge:

https://human-centered.ai/kandinksy-challenge



Thank you !