

21st IEEE International Conference on Machine Learning and Applications



December 12-14, 2022
Atlantis Hotel, Bahamas

AMLA



150 YEARS
FEATURING
FUTURE
1872 - 2022

UNIVERSITY OF NATURAL RESOURCES AND
LIFE SCIENCES, VIENNA

Human-Centered AI to foster Trustworthy AI

Andreas Holzinger

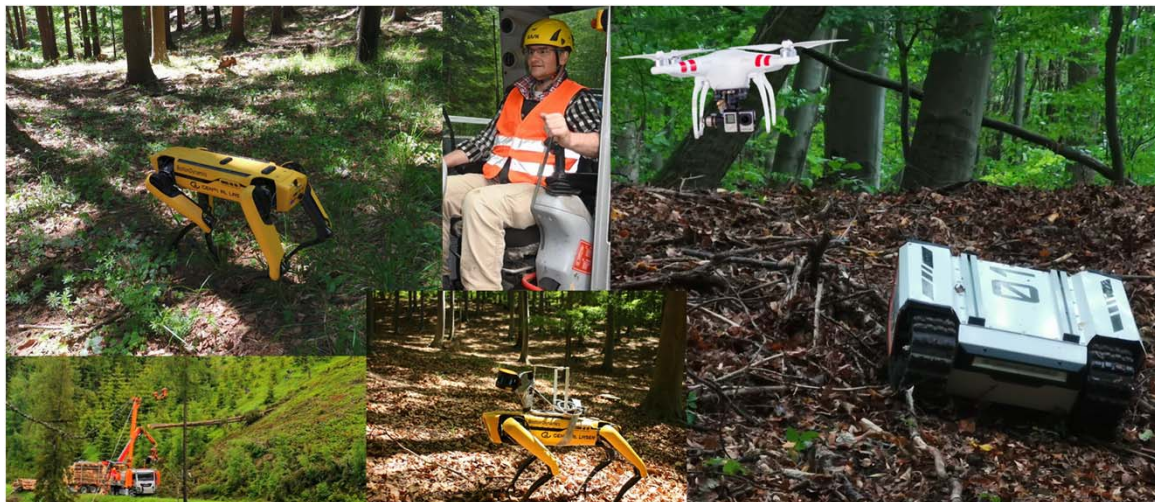
University of Natural Resources and Life Sciences, Vienna
Department of Forest and Soil Sciences
Institute of Forest Engineering
Human-Centered AI Lab

Keynote 21st IEEE Intl. Conference on
Machine Learning & Applications
Nassau, Wed, December 14, 2022, 08:30

andreas.holzinger@human-centered.ai

<https://human-centered.ai>

We work in the world's largest Human-Centered AI Lab



<https://human-centered.ai>

andreas.holzinger AT human-centered.ai

2

HCAI to foster Trustworthy AI, 14.12.2022

Industry 5.0 is here ...



Industry 1.0

1800

mechanization,
water and
steam powers



Industry 2.0

1900

mass production,
electric power,
assembly line



Industry 3.0

2000

computers,
automated
production,
electronics



Industry 4.0

2010

cyber-physical
systems, IoT,
networking,
machine learning



Industry 5.0

2020

human-robot
collaboration,
cognitive systems,
customization

Source BOKU Forest Engineering, 2022

Human-Centered AI (HCAI):=

- a synergistic approach to align AI with human values, ethical principles, and legal requirements to ensure security, safety and trust – to foster “One Health”



<https://boku.ac.at>

Andreas Holzinger, Edgar Weippl, A Min Tjoa & Peter Kieseberg (2021). Digital Transformation for Sustainable Development Goals (SDGs) - a Security, Safety and Privacy Perspective on AI. Springer Lecture Notes in Computer Science, LNCS 12844. Cham: Springer, pp. 1--20, doi:10.1007/978-3-030-84060-0_1.

Acknowledgements



- FWF P-32554 xAI - A reference model of explainable Artificial Intelligence
- EU RIA 826078 FeatureCloud - Trusted digital solutions and Cybersecurity in Health
- EU RIA 874662 HEAP - Human Exposome: digital toolbox for assessing and addressing environmental impact on health
- FFG 879881 EMPAIA – Digital Ecosystem for Pathology Diagnostics with AI Assistance



European
Commission

Horizon 2020
European Union funding
for Research & Innovation



Agenda



- (1) Advances in statistical data-driven machine learning makes AI popular again
- (2) For many applications explainability and robustness are major challenges
- (3) Correlation is not causality, and a human-in-the-loop can (sometimes - not always) bring-in experience and conceptual understanding
- (4) Explainability methods allow to understand why and how a result was achieved
- (5) Causability measures the quality of explanations obtained by (4)

(1) Statistical data-driven machine learning

Machine Learning 101

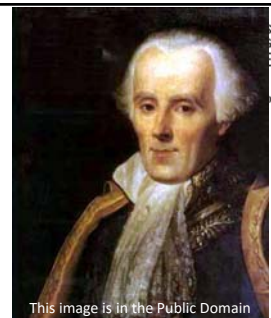
\mathcal{D} ... data $\mathcal{D} = x_{1:n} = \{x_1, x_2, \dots, x_n\}$

θ ... features prior: $p(\theta)$ likelihood: $p(\mathcal{D}|\theta)$

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

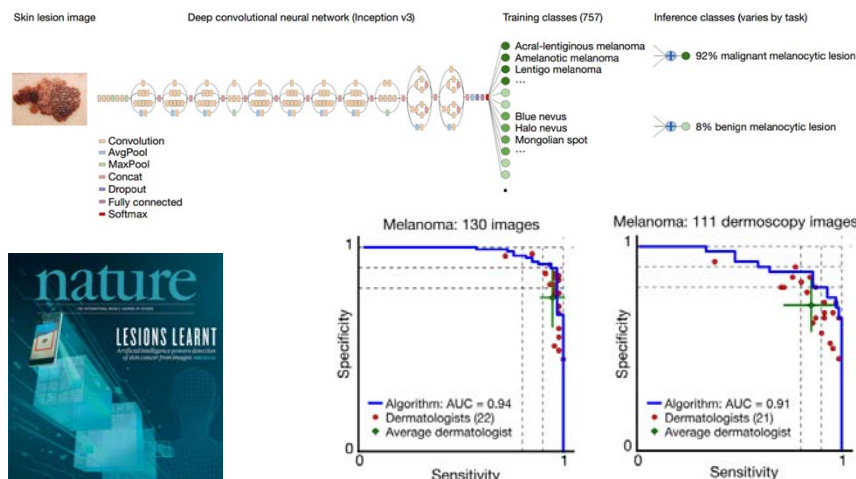
Posterior $\approx p(x)$ of θ after seen ("learned") \mathcal{D} : $p(\theta|\mathcal{D})$

The inverse probability allows us to infer unknowns and to make predictions ...



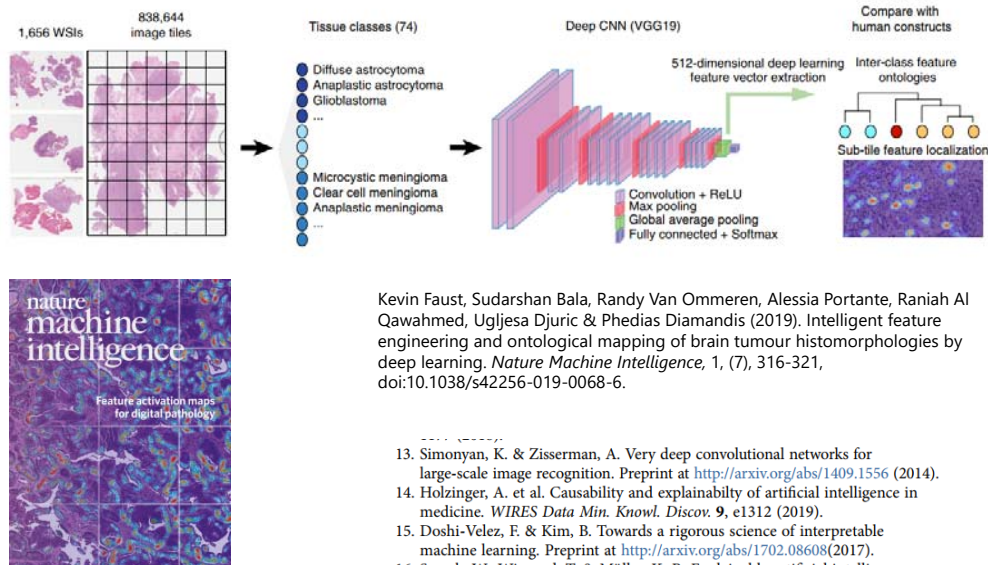
"Il est remarquable qu'une science qui a commencé avec l'ère la prise en compte des jeux de hasard ... aurait dû devenir l'objet le plus important de la connaissance humaine."
Laplace (1781)

Human-Level AI ...



Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, (7639), 115-118, doi:10.1038/nature21056.

Histopathology: Towards Human Level AI



Kevin Faust, Sudarshan Bala, Randy Van Ommeren, Alessia Portante, Raniah Al Qawahmed, Ugljesa Djuric & Phedias Diamandis (2019). Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. *Nature Machine Intelligence*, 1, (7), 316-321, doi:10.1038/s42256-019-0068-6.

13. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at <http://arxiv.org/abs/1409.1556> (2014).
14. Holzinger, A. et al. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* 9, e1312 (2019).
15. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at <http://arxiv.org/abs/1702.08608> (2017).
16. Samal, W., Weng, T. & Miller, K. D. Explainable artificial intelligence.



x
classified as
stop sign
with 57,7 %
confidence



$sign(\nabla_x J(\theta, x, y))$



$x +$
 $\epsilon sign(\nabla_x J(\theta, x, y))$
classified as
max. 100 km/h sign
with 99,3 % confidence

State-of-the-art embodied intelligence



Source: Abbeel, 2021



Xiaofei Wang, Kimin Lee, Kourosh Hakhmaneshi, Pieter Abbeel & Michael Laskin. Skill preferences: Learning to extract and execute robotic skills from human feedback. 5th Conference on Robot Learning (CoRL 2021), (2022) London (UK).

The world best algorithms are lacking robustness

What do we need to reach robustness



- 1) learning from (little) **real-world data**
- 2) extracting **relevant** knowledge
- 3) **generalize**
- 4) fight the curse of **dimensionality**
- 5) disentangle **independent** explanatory factors of data, i.e.
- 6) **causal understanding** of the data in the **context** of an application domain

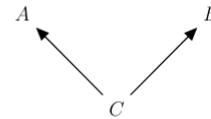
How do we solve this problem ?

(3) Correlation \neq Causality and the Human-in-the-loop

Correlation does not tell anything about causality!

- Hans Reichenbach (1891-1953):
Common Cause Principle
 Links causality with probability:
 - If A and B are statistically dependent,
 there is a C influencing both
 - Whereas:
 - A, B, C ... events
 - p ... probability density

time ↑



$$p(A \cap B) > p(A)p(B)$$

$$p(A \cap B|C) = p(A|C)p(B|C)$$

$$p(A \cap B|\bar{C}) = p(A|\bar{C})p(B|\bar{C})$$

$$p(A|C) > p(A|\bar{C})$$

$$p(B|C) > p(B|\bar{C})$$

$$p(X|Y) \doteq \frac{p(X \cap Y)}{p(Y)}$$

Hans Reichenbach 1956. The direction of time
 (Edited by Maria Reichenbach), Mineola, New York, Dover.

Hitchcock, Christopher and Miklós Rédei, "Reichenbach's Common Cause Principle",
 The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.),
 Online available: <https://plato.stanford.edu/archives/spr2020/entries/physics-Rpcc>

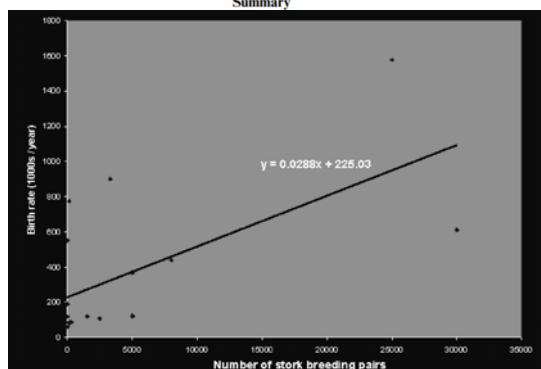
Remember: Correlation is NOT Causality

Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
 Teaching;
 Correlation;
 Significance;
 p-values.

Robert Matthews
 Aston University, Birmingham, England.
 e-mail: rajm@compuserve.com

Summary

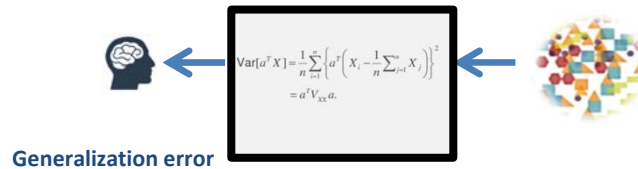


Country	Area (km ²)	Storks (pairs)	Humans (10 ⁶)	Birth rate (10 ³ /yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

Robert Matthews 2000. Storks deliver babies ($p = 0.008$). Teaching Statistics, 22, (2), 36-38.

Human-in-the-Loop



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

What is the human supposed to do ?

Humans can generalize from very few examples

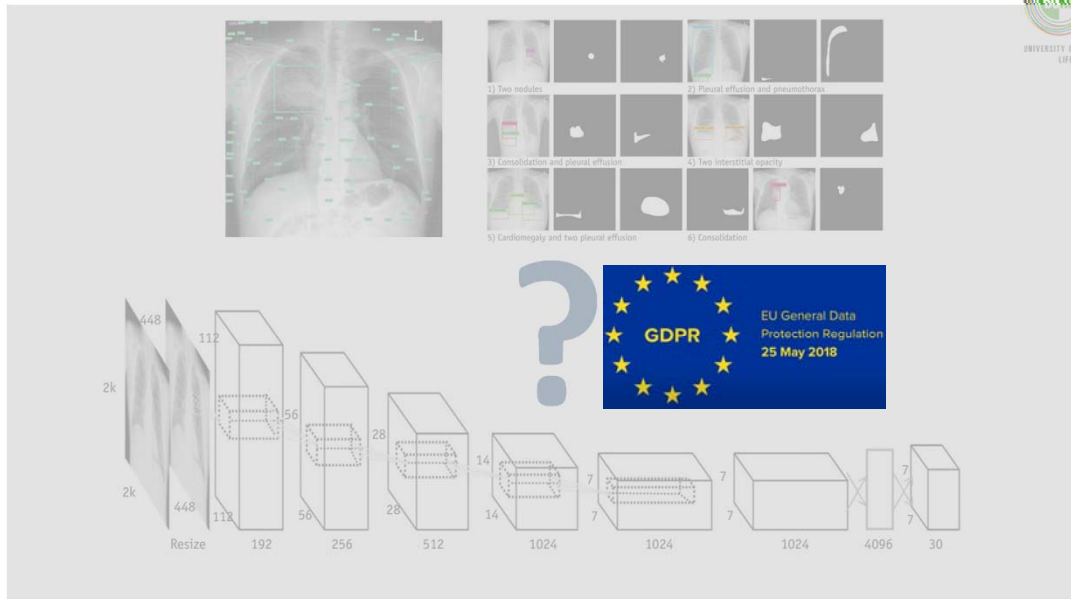


Source: Public Domain, freedesignfile.com

- Humans use abstract concepts
- Humans can make inference from little, noisy, incomplete data
- Humans can set the prior: finding shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

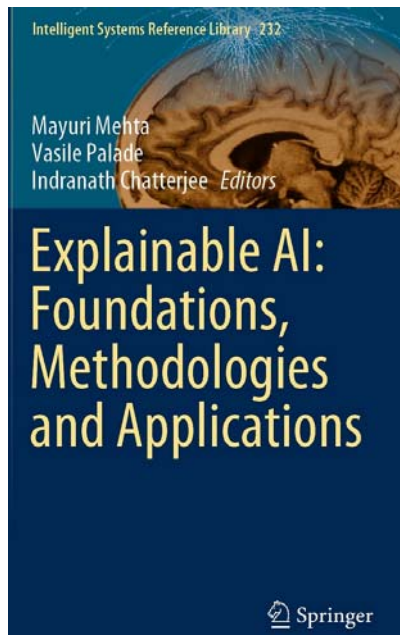
Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum (2015). Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

However, there is another problem



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.

(4) Methods of Explainability



Dr Mayuri Mehta



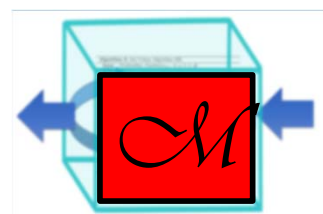
Dr. Indranath Chatterjee
Professor of Computer Engineering at Korea.



Vasile Palade^{1st}
Professor of Artificial Intelligence and Data Science

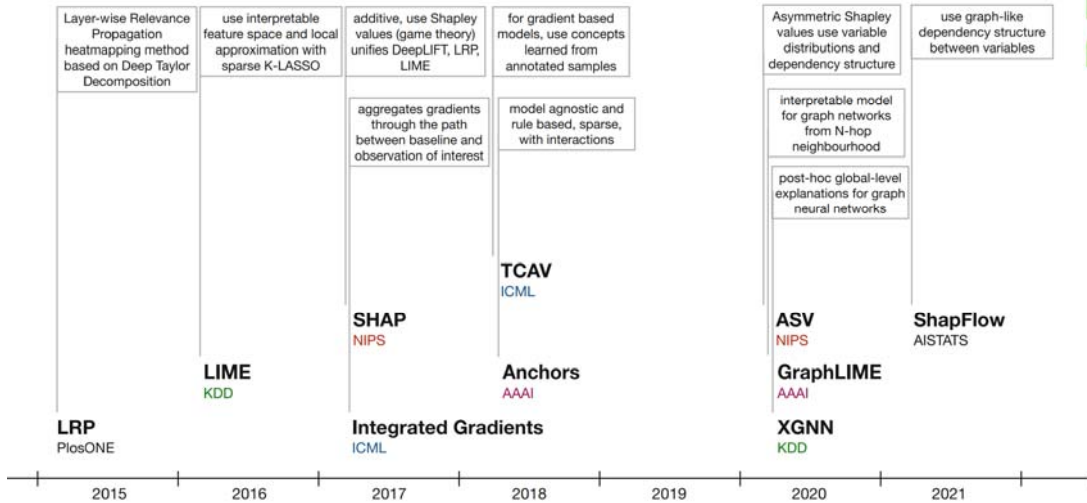
Mayuri Mehta, Vasile Palade & Indranath Chatterjee (eds.)
(2023). Explainable AI: Foundations, Methodologies and
Applications Cham: Springer, doi:10.1007/978-3-031-12807-3.

- **Interpretable Models, = ante-hoc** - the “glass-box” model itself is *ante-hoc* interpretable, e.g., Regression, Naïve Bayes, Decision Trees, Graphs, ...
- **Interpreting Black-Box Models, = post-hoc** - the model is not interpretable and needs a post-hoc interpretability method \mathcal{M}



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

Methods of explainable AI (xAI)



Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek & Wojciech Samek (2022). Explainable AI Methods - A Brief Overview. XXAI - Lecture Notes in Artificial Intelligence LNAI 13200. Cham: Springer, pp. 13--38, doi:10.1007/978-3-031-04083-2_2.

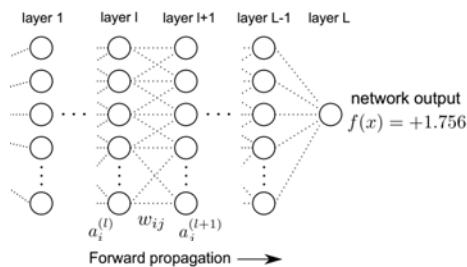
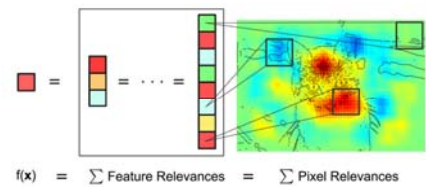
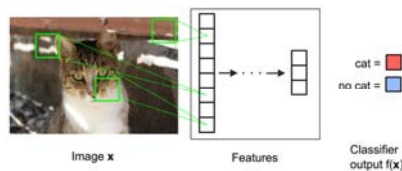
andreas.holzinger AT human-centered.ai

29

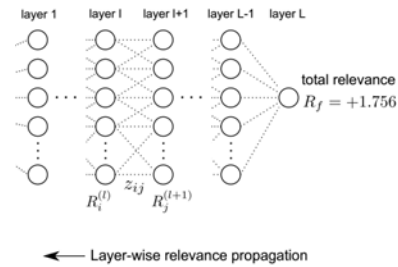
HCAI to foster Trustworthy AI, 14.12.2022

$$f(x) \approx \sum_{d=1}^V R_d$$

$$R_i = \left\| \frac{\partial}{\partial x_i} f(x) \right\|$$



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



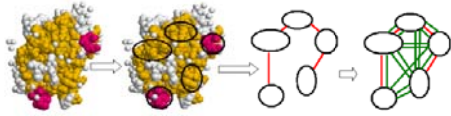
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_i' z_{i'j}} R_j^{(l+1)}$$

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.

andreas.holzinger AT human-centered.ai

30

HCAI to foster Trustworthy AI, 14.12.2022



Karsten M Borgwardt, Cheng Soon Ong, Stefan Schöner, Svn Vishwanathan, Alex J Smola & Hans-Peter Kriegel (2005). Protein function prediction via graph kernels. *Bioinformatics*, 21, (suppl 1), i47-i56.

G ... input graph

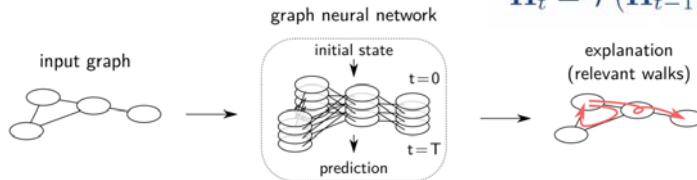
$$G = (\mathcal{V}, \mathcal{E})$$

$$\mathcal{V} = \{v_1, \dots, v_n\}$$

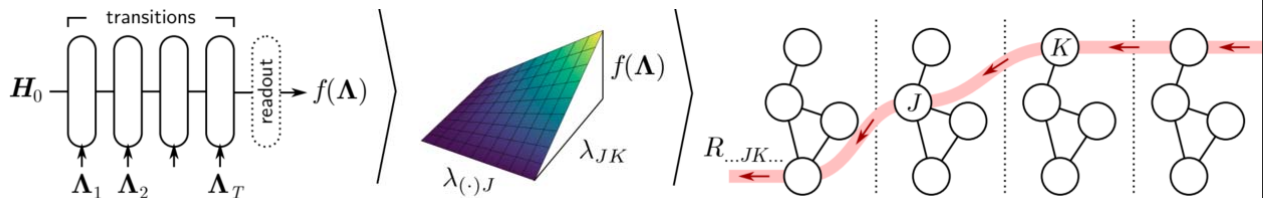
$$\mathcal{E} \subseteq \{(v_i, v_j) | v_i, v_j \in \mathcal{V}\}$$

H_0 ... initial state

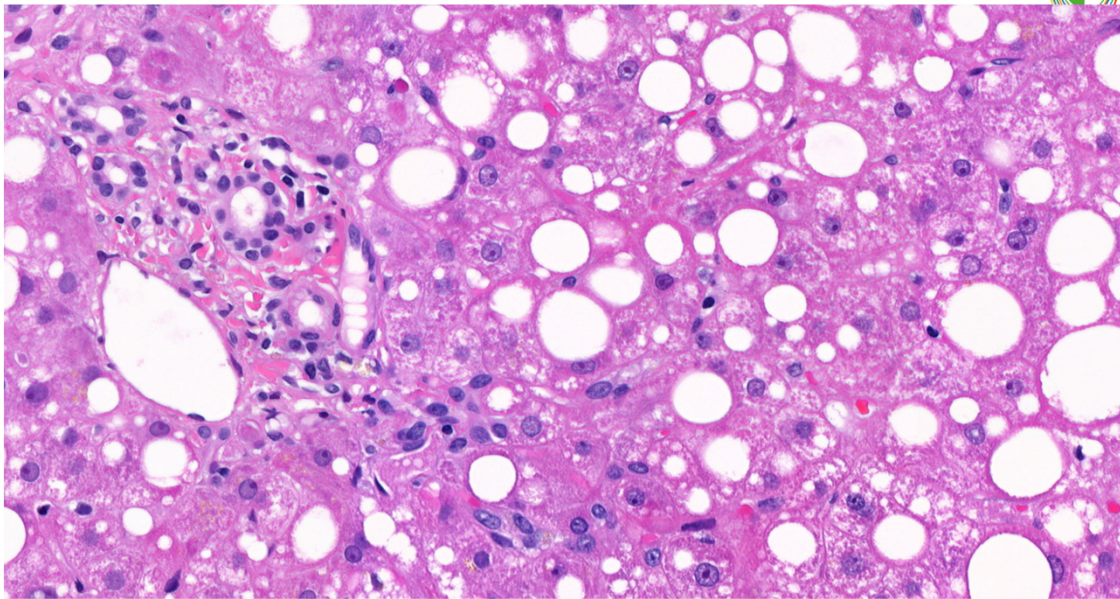
$$H_t = \mathcal{T}(H_{t-1}, \Lambda_t, W_t)$$



Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller & Grégoire Montavon (2020). XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. *arXiv:2006.03589*.



(5) Causability measures the quality of explanations obtained from (4).

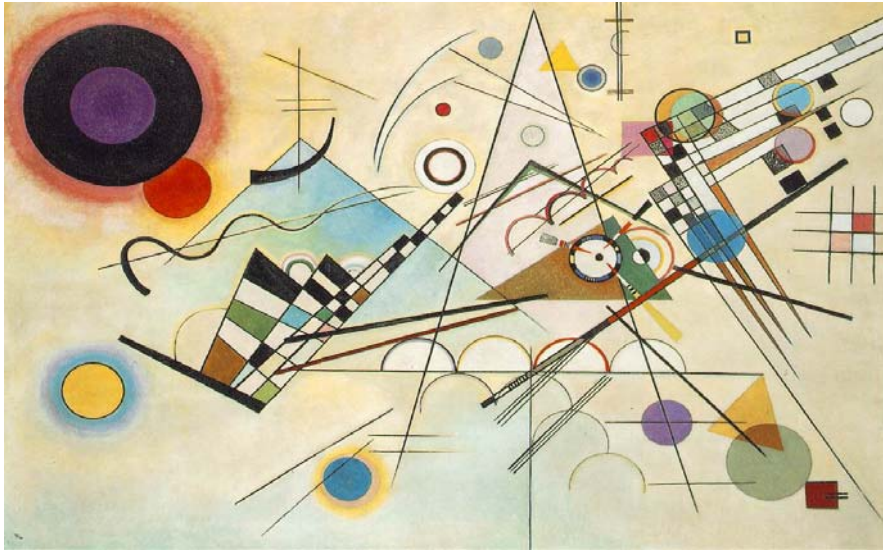


Ground truth

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
 - *Empirical evidence* = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
 - *Empirical inference* = drawing conclusions from empirical data (observations, measurements)
 - *Causal inference* = drawing conclusions about a causal connection based on the conditions of the occurrence of an effect
 - *Causal machine learning* is key to ethical AI in health to model explainability for bias avoidance and algorithmic fairness for decision making

Federico Cabitza, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stoeger & Andreas Holzinger (2023). Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI. Expert Systems with Applications, 213, (3), doi:10.1016/j.eswa.2022.118888.

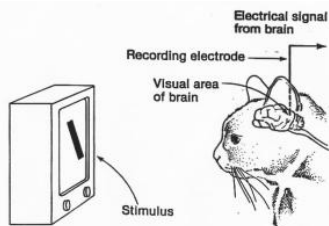
Kandinsky 1923



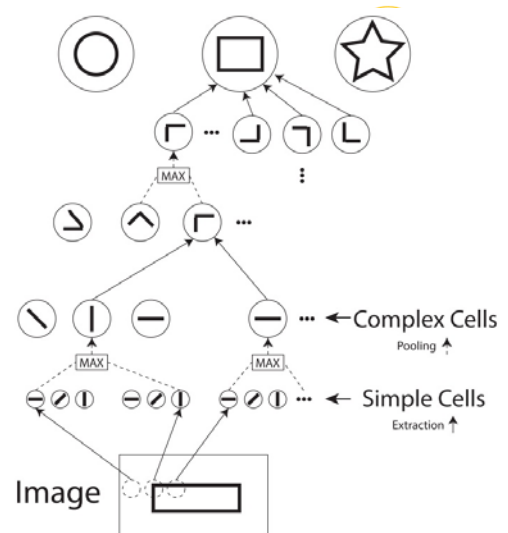
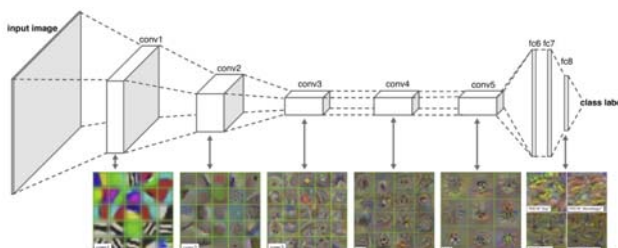
Wassily Kandinsky
(1866 – 1944)

Komposition VIII, 1923, Solomon R. Guggenheim Museum, New York. Source: https://de.wikipedia.org/wiki/Wassily_Kandinsky
Note: Image is in the public domain and is used according to UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

Hubel & Wiesel 1962

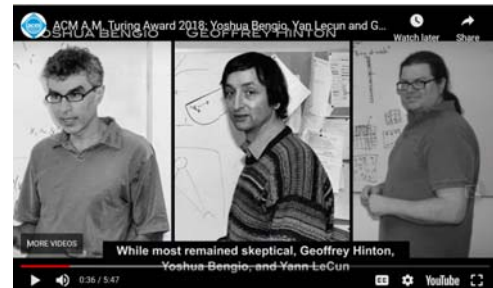
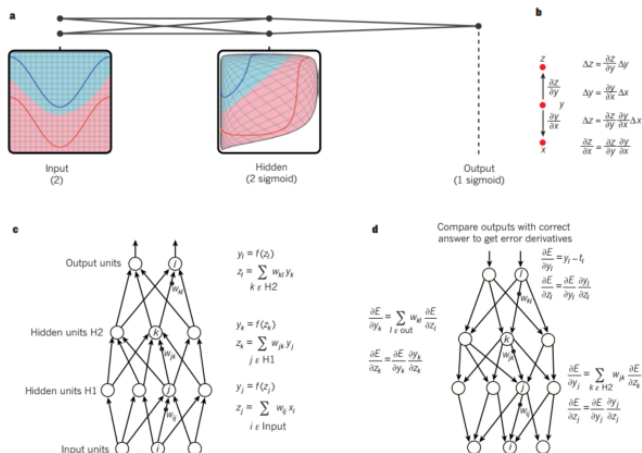


David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837



Source: <https://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/models-of-information-processing-in-the-visual-cortex>

Geoff Hinton, Yann LeCun, Yoshua Bengio Turing Award 2018



<https://awards.acm.org/about/2018-turing>

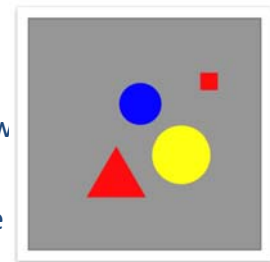
Yann Lecun, Yoshua Bengio & Geoffrey Hinton (2015). Deep learning. Nature, 521, (7553), 436-444, doi:10.1038/nature14539.

Yoshua Bengio, Yann Lecun & Geoffrey Hinton (2021). Deep learning for AI. Communications of the ACM, 64, (7), 58-65, doi:10.1145/3448250.

How do humans explain ?

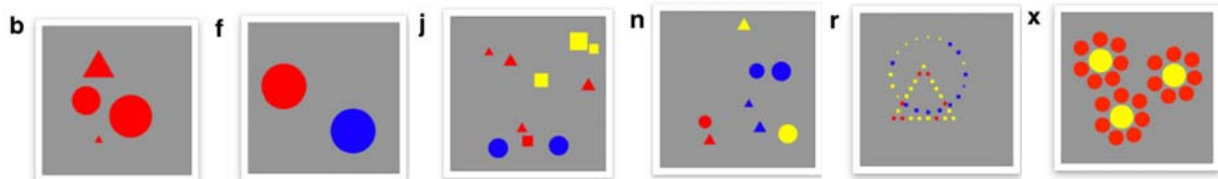
Definition 1: A Kandinsky Figure is ...

- ... a square image containing 1 to n geometric objects.
- Each object is characterized by its shape, color, size and position w
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable



Heimo Mueller & Andreas Holzinger (2021). Kandinsky Patterns. Artificial intelligence, 300, (11), 103546, doi:10.1016/j.artint.2021.103546.

Definition 2 A statement $s(k)$



- about a Kandinsky Figure k is ...
- either a mathematical function $s(k) \rightarrow B$; with $B (0,1)$
- or a *natural language statement* which is true or false
- The evaluation of a natural language statement is always done in a *specific context*.
- we follow **well known concepts from human perception** and linguistic theory.
- If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.

Holzinger, A. & Müller, H. 2020. Verbinden von Natürlicher und Künstlicher Intelligenz: eine experimentelle Testumgebung für Explainable AI (xAI). HMD Praxis der Wirtschaftsinformatik, 57, (1), 33-45, doi:10.1365/s40702-020-00586-y

GO BACK TO LIST OF PATTERNS GO TO NEXT PATTERN

What is Pattern VIII?

Hypothesis 1

There are 4 objects



Hypothesis 2

There is always a triangle



Hypothesis 3

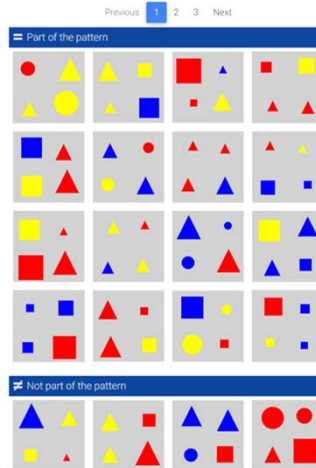
There is more than 1 color



+ NEW HYPOTHESIS

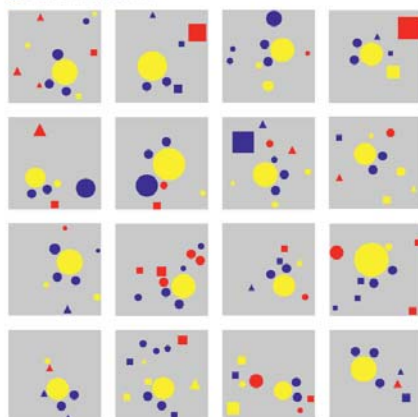
! HINT

? SOLUTION

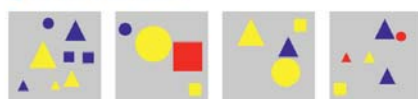


Andreas Holzinger, Michael Kickmeier-Rust & Heimo Mueller 2019. KANDINSKY Patterns as IQ-Test for machine learning. Springer Lecture Notes LNCS 11713. Cham (CH): Springer Nature Switzerland, pp. 1-14, doi:10.1007/978-3-030-29726-8_1.

Part of the pattern



Not part of the pattern

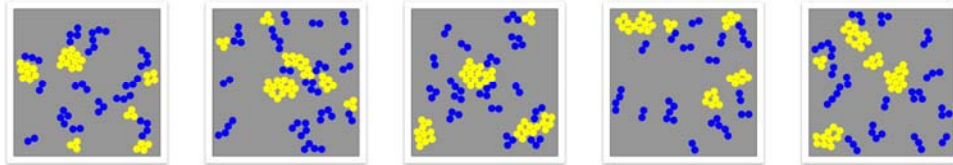


S8 Basic Pattern 8

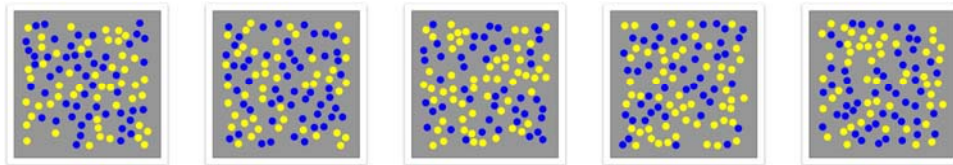
Title: **Mickey Mouse** ->

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.

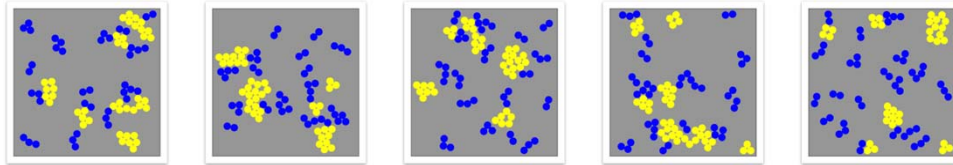
A) True (the cells are smaller and closer together – it is an tumor ...)



B) False



C) Counterfactual (What if the cells are slightly bigger ?)



<https://human-centered.ai/project/kandinsky-patterns>

[Home](#)
[About](#)
[Holzinger Group](#)
[For Experts](#)
[For Students](#)
[Open Work \(2020\)](#)
[Partners](#)
[News](#)
[Q](#)



#KANDINSKYPatterns our Swiss-Knife for the study of explainable-AI



ABSTRACT

KANDINSKYPatterns (yca, named after the famous artist Wassily Kandinsky) are mathematically describable, simple, self-contained, hence controllable test data sets for the development, validation and training of explainability in artificial intelligence (AI) and machine learning (ML). Whilst our *KANDINSKYPatterns* have these computationally manageable properties, they are at the same time easily distinguishable from human observers. Consequently, controlled patterns can be described by both humans and algorithms.

We define a *KANDINSKY Pattern* as a set of *KANDINSKY Figures*, where for each figure an "infallible authority" (ground truth) defines that this figure belongs to the *KANDINSKY Pattern*. With this simple principle we build training and validation data sets for automatic interpretability and context learning.



KANDINSKYPATTERNS AT TEDx



KANDINSKY ARTIFICIAL INTELLIGENCE EXPLANATION CHALLENGE

Here we challenge the international machine learning community to generate machine explanations



KANDINSKY HUMAN INTELLIGENCE EXPLANATION CHALLENGE

Here we challenge any human individual to take part in this experiment and to generate human explanations



HCAI GITHUB REPOSITORY

OPEN STUDENTS THESES

Human-AI Interface DESIGNER
More Projects

LATEST NEWS

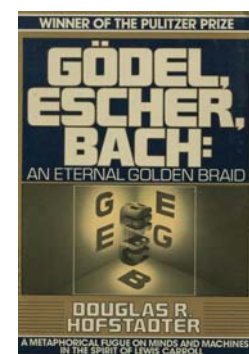
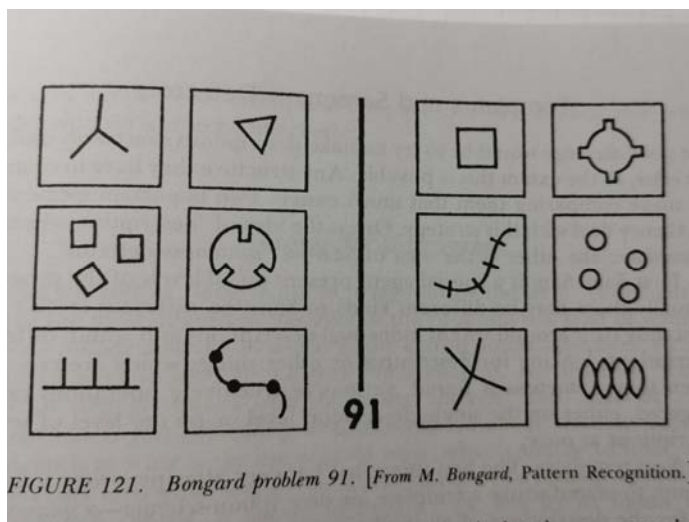

 August 25-28, 2020, Machine Learning & Knowledge Extraction, LNCS 12279 published !
 2020-08-21 - 12:15

Our Springer LNCS 12279 Machine Learning & Knowledge Extraction just been published.

Related Work (citations as of 14.12.2022)

- Bongard-Problems (1967) – 68 (Hofstadter (1979) - 7990)
- CLEVR (2017) - 1622
- Shapeworld (2017) – 48
- CLEVR-Humans (2017) - 499
- SCOOP (2018) - 127
- CLEVRER (2019) -232
- RAVEN (from Raven Progressive Matrices (RPM)) (2019) - 131
- Bongard Logo (2020) - 26
- CURI (2021) – 14
- CLEVRER-Humans (2022) – 0
- CLEVER-XAI (2022) – 23
- Super-CLEVR (2022) - 0

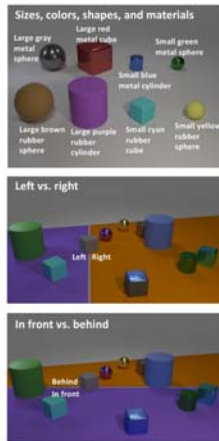
Bongard Problems



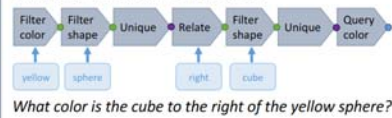
Douglas R. Hofstadter (1979)
 Gödel, Escher, Bach:
 An Eternal Golden Braid,
 New York: Basic Books.

Bongard, M. Mikhail, 1967. The problem of recognition (in Russian), Moscow, Nauka (1970 in English)

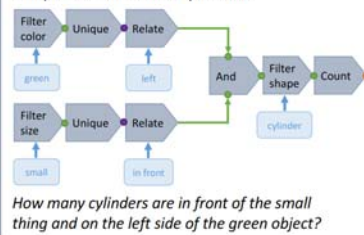
CLEVR = Compositional Language and Elementary Visual Reasoning



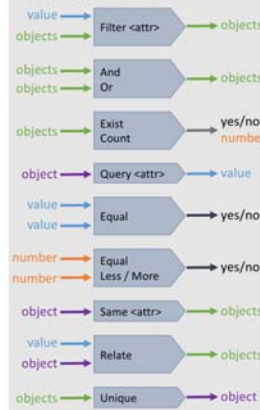
Sample chain-structured question:



Sample tree-structured question:



CLEVR function catalog



Questions in CLEVR test various aspects of visual reasoning including **attributes**, **identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.

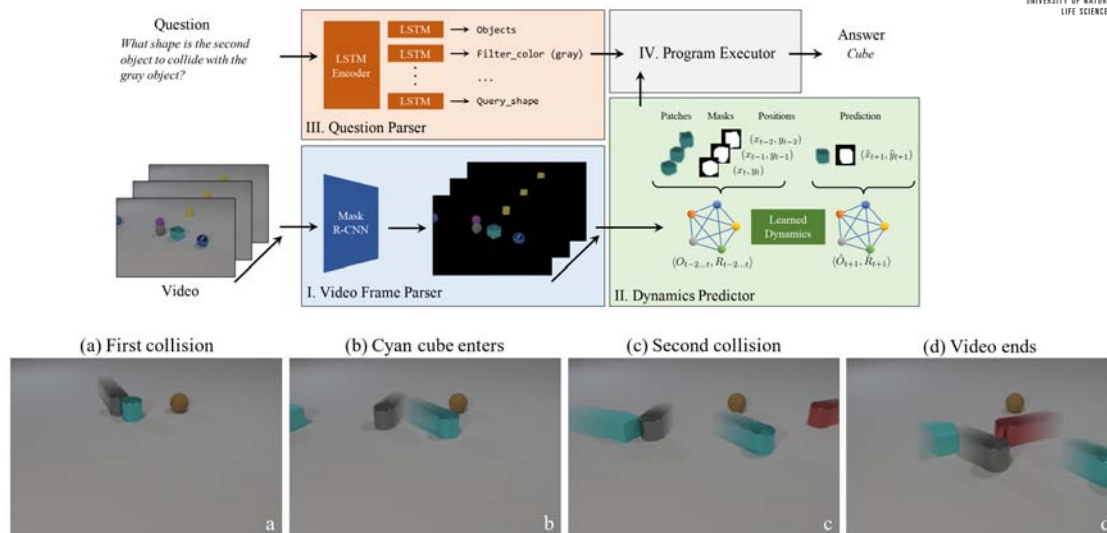


Q: Are there an equal number of large things and metal spheres?
 Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
 Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?

<https://cs.stanford.edu/people/jcjohns/clevr/>

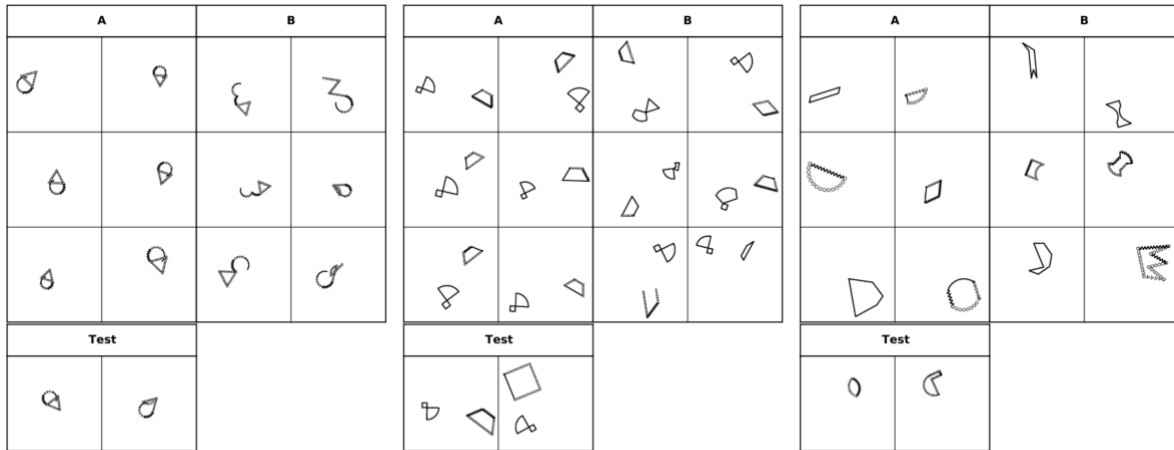
Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick & Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 Hawaii. IEEE.

CLEVRER



Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba & Joshua B. Tenenbaum (2019). CLEVRER: Collision events for video representation and reasoning. arXiv:1910.01442.

Bongard-LOGO



(a) free-from shape problem

(b) basic shape problem

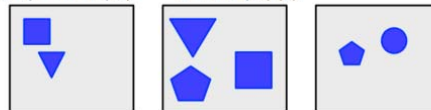
(c) abstract shape problem

Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu & Anima Anandkumar (2020). BONGARD-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning. Advances in Neural Information Processing Systems, 33.

CURI

for-all $x \in S$ ($\text{color?}(x) = \text{"blue"}$) and ($\text{all}(\text{size?}(S) = \text{size?}(x))$)

All objects are blue and have the same size



for-all $x \in S$ ($\text{all}(\text{color?}(x) = \text{color?}(S))$)

All objects in the scene have the same color



exists $x \in S$ ($\text{color?}(x) = \text{"blue"}$) and $\text{all}(\text{shape?}(S_{\{-x\}}) = \text{"square"})$

There exists a blue object in the scene and the rest of the objects are squares



\mathcal{G} : Context Free Grammar

Variables

$x \triangleq$ Object in scene
 $S \triangleq$ All objects

$S_{\{-x\}} \triangleq S/\{x\}$

Quantifiers

for-all

exists

Functions

color? location?

shape? size?

material? all

Operators

and Greater(>)

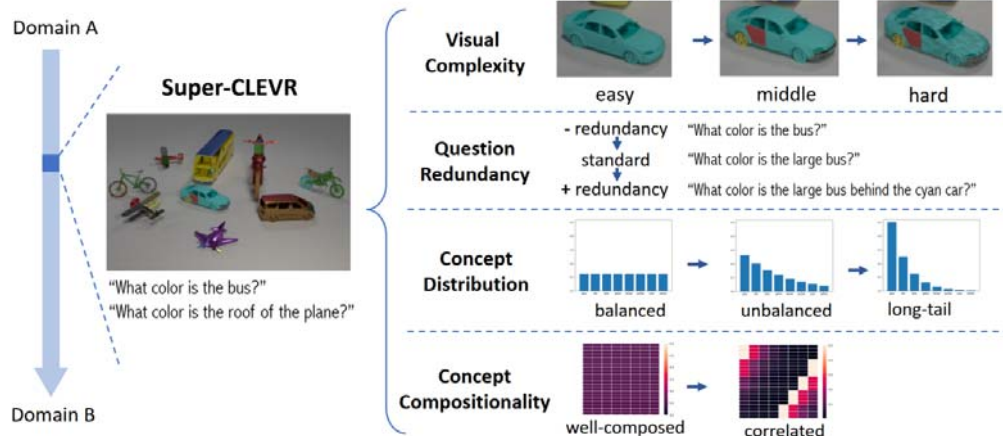
or Lesser(<)

not

=

Ramakrishna Vedantam, Arthur Szlam, Maximilian Nickel, Ari Morcos & Brenden Lake (2020). CURI: A Benchmark for Productive Concept Learning Under Uncertainty. arXiv:2010.02855.

Super-CLEVR



Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme & Alan Yuille (2022). Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning. arXiv preprint, doi:10.48550/arXiv.2212.00259.

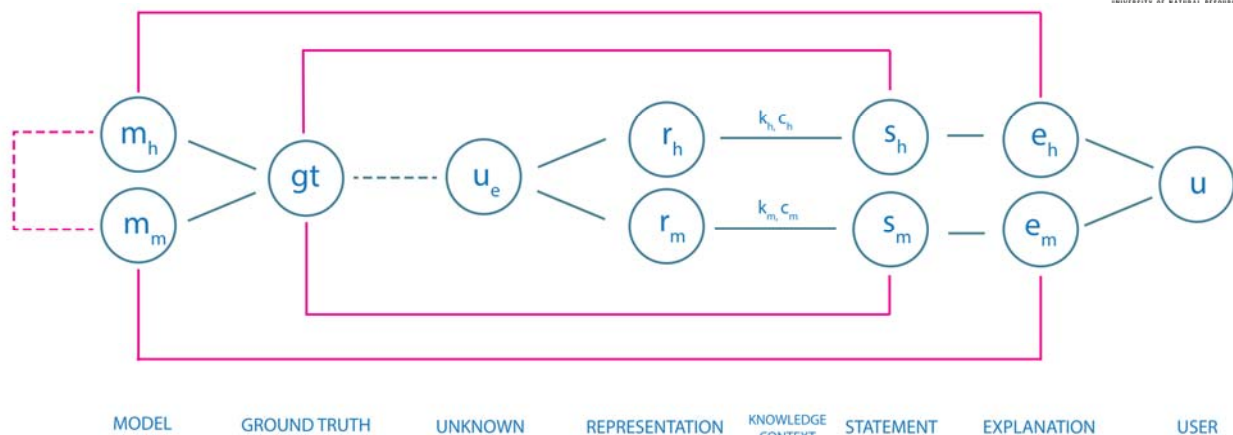
Measuring the quality of Explanations: The Systems Causability Scale

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

- Causability is neither a typo nor a synonym for Causality
- Causa-bil-ity ... in reference to ... Usa-bil-ity.
- While xAI is about implementing transparency and traceability, Causability is about the measurement of the quality of explanations.
- **Explainability** := technically highlights decision relevant parts of machine representations and machine models i.e., parts which contributed to model accuracy in training, or to a specific prediction.
 - Explainability does not refer to a human model!
- **Causability** := the measurable extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency, satisfaction in a specified context of use.
 - Causability does refer to a human model!

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

How can we measure the quality of explanations ?

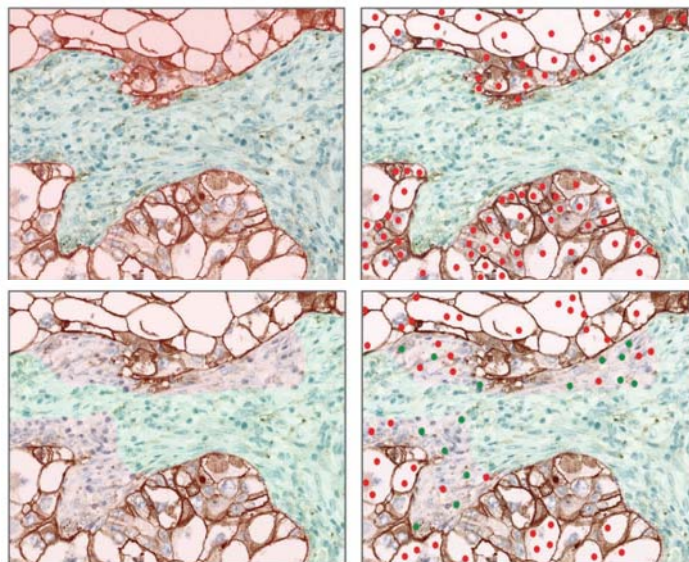


Andreas Holzinger, Andre Carrington & Heimo Müller 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited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

Explainability is the first step

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

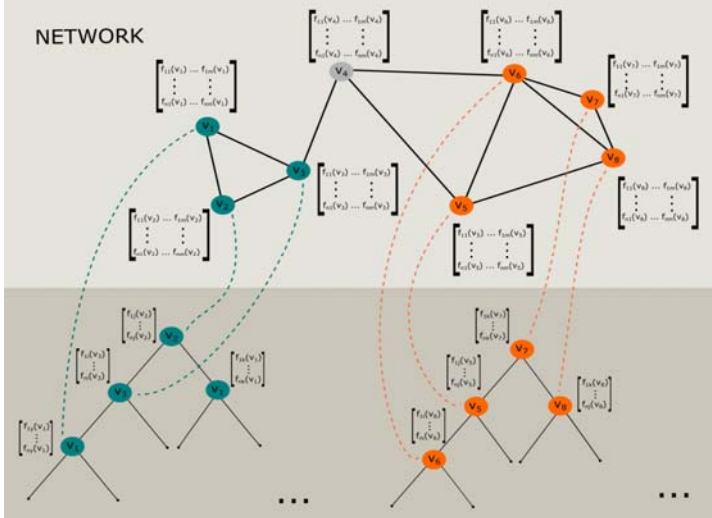
Example



Andreas Holzinger & Heimo Mueller (2021). Toward Human-AI Interfaces to Support Explainability and Causability in Medical AI. *IEEE COMPUTER*, 54, (10), doi:10.1109/MC.2021.3092610.

Actionable Explainable AI – with the expert-in-the-loop

NETWORK



Algorithm 1: Greedy Decision Forest

```

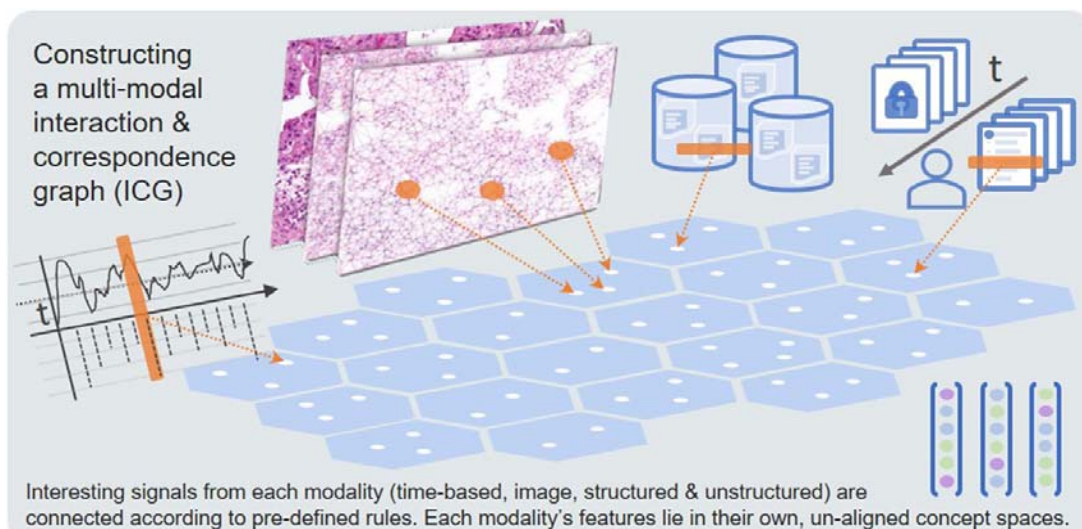
1 Given a Decision Forest:  $\{T_k(x, \Theta_k, X_k^G)\}$ ;
2 Given a graph:  $G = (V, E)$ ;
3  $k = \{1, \dots, ntree\}, t = 1$ ;
4  $mtry_k = \sqrt{\#V}$ ;
5  $X_k^G[t = 1] = X_k^G$ ;
6 while  $t \leq niter$  do
7   for  $k \leftarrow 1$  to  $niter$  do
8      $Perf(T_k[t]) = Performance\ of\ T_k(x; \Theta_k, X_k^G[t])$ ;
9     if  $Perf(T_k[t]) \leq Perf(T_k[t - 1])$  then
10       $T_k[t] = T_k[t - 1]$ ;
11       $X_k^G[t] = X_k^G[t - 1]$ ;
12    else
13       $mtry_k[t] = -$ ;
14       $X_k^G[t] = RandomWalk(G|X_k, mtry_k[t])$ ;
15    end
16  end
17  Sample  $niter$  trees according to  $Perf\{T_k[t]\}$ ;
18   $t++$ ;
19 end
  
```

Bastian Pfeifer, Anna Saranti, Andreas Holzinger (2021). Network Module Detection from Multi-Modal Node Features with a Greedy Decision Forest for Actionable Explainable AI. arXiv:2108.11674.

Conclusio

Explainability needs a framework to ensure common understanding and adaptive Question/Answering Interfaces

Multimodal Causability: enabling *why* and *what-if* ...



Andreas Holzinger, Bernd Malle, Anna Saranti & Bastian Pfeifer (2021). Towards Multi-Modal Causability with Graph Neural Networks enabling Information Fusion for explainable AI. Information Fusion, 71, (7), 28-37, doi:10.1016/j.inffus.2021.01.008.

Great things are only possible in great teams – Merci BOKU



HCAI
HUMAN-CENTERED.AI



andreas.holzinger AT human-centered.ai

61

HCAI to foster Trustworthy AI, 14.12.2022