Human-Centered AI to foster Explainability and Robustness for Trustworthy AI

Andreas Holzinger

Human-Centered AI Lab (HCAI Lab), Medical University Graz, Austria
with effect of March, 1, 2022: University of Natural Resources and Life Sciences Vienna, Austria
and
Explainable AI-Lab, Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada

ORGANIZERS

- **Elizabeth Daly (Workshop Chair)**, IBM Research,
- **Öznur Alkan**, IBM Research, Dublin
- **Stefano Teso**, University of Trento
- **Wolfgang Stammer**, TU Darmstadt

Thanks to my funding organizations

- FWF P-32554 xAI - A reference model of explainable Artificial Intelligence for digital medicine
- EU RIA 826078 FeatureCloud - Trusted digital federated 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

FWF Der Wissenschaftsfonds. European Commission | Horizon 2020 European Union funding for Research & Innovation FFG Österreichische Forschungsförderungsgesellschaft

Agenda

- (0) Motivation ...
- (1) Examples ...
- (2) Challenges ...
- (3) Human-in-the-loop ...
- (4) Explainability ...
- (5) Causability ...

(0) Motivation

Why is AI So Dumb ?

Definitions

- **Trust** := *subjective* belief/assessment incl. security, dependability, integrity, predictability, reliability (always as expectation!)
- **Trustworthy AI** := *ensures* security, safety, privacy, non-discrimination, fairness, accountability (re-traceability, replicability), auditability and environmental well-being, and most of all robustness and explainability
- **Robustness** := to produce reliable results even if the input data is perturbed
- **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, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, Wojciech Samek, Igor Jurisica & Natalia Díaz-Rodríguez (2021). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, (3), 263-278, doi:10.1016/j.inffus.2021.10.007.

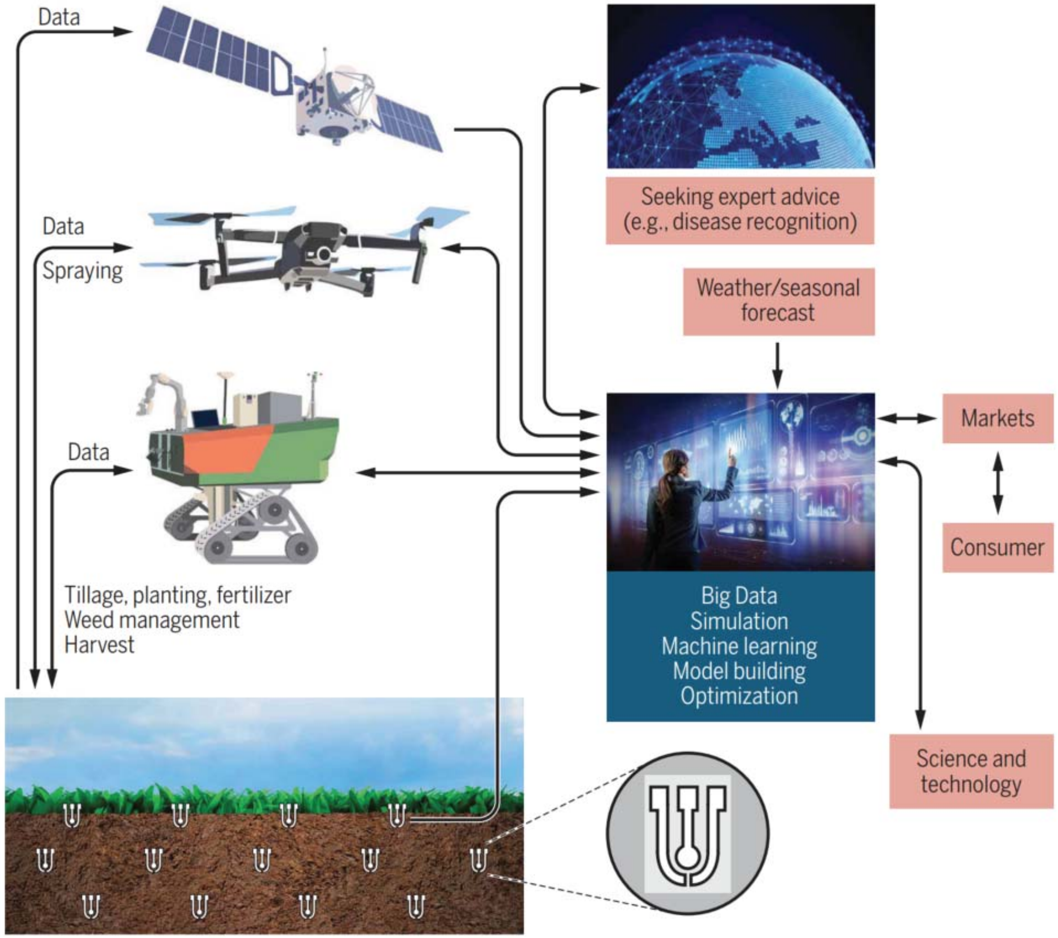
(1) Examples ...

Precision Medicine

Source: Stock licensed

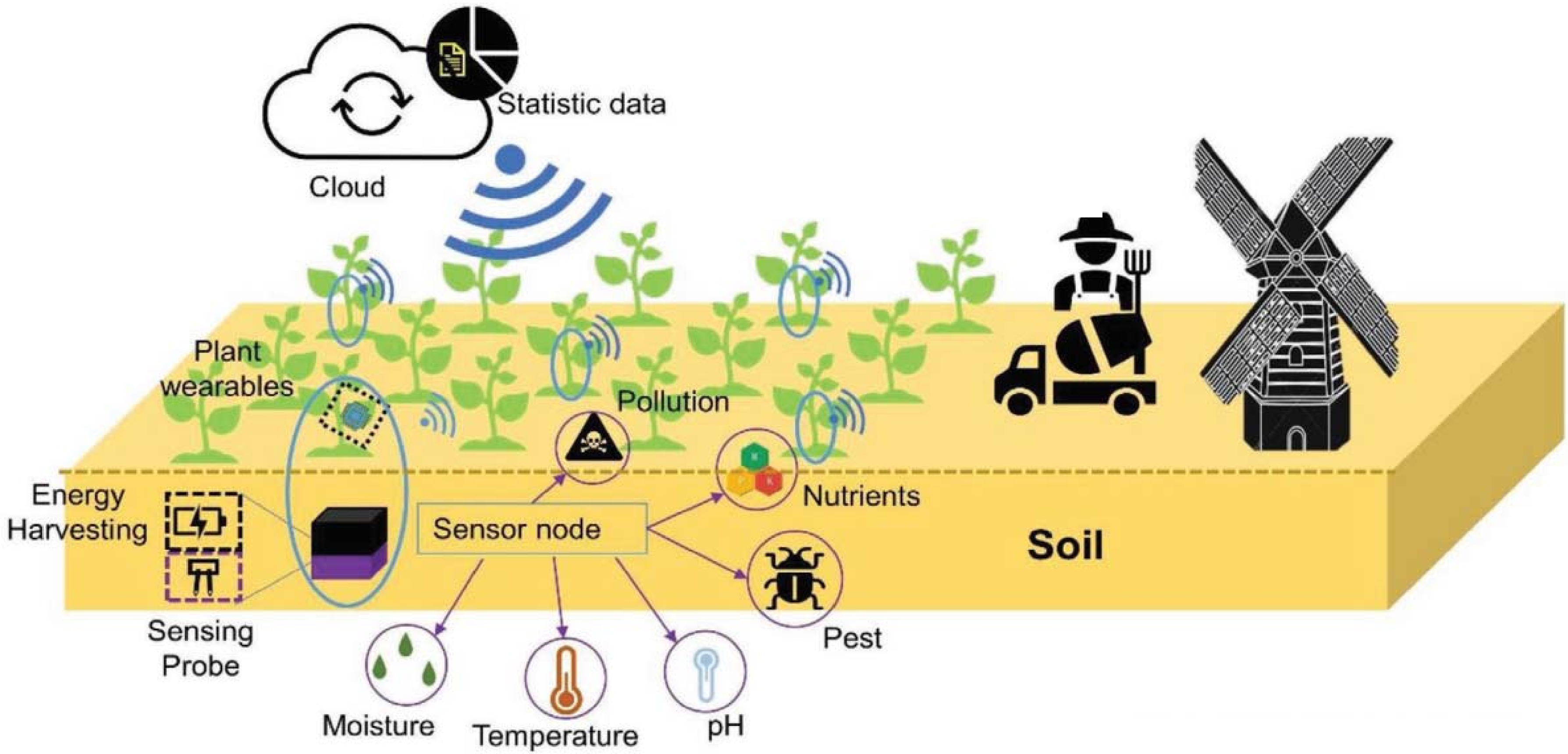
Digital Transformation in Smart Farm and Forest Operations

Example Precision Farming: Data $\mathcal{D}$, Features Θ, ...



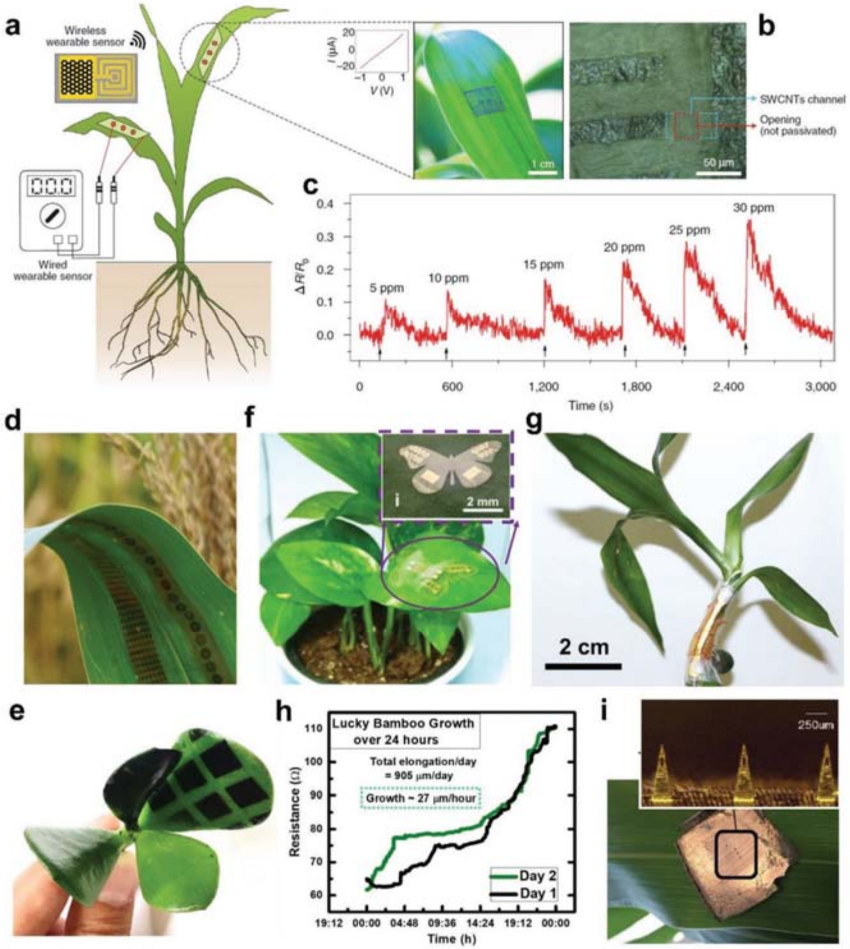
Senthold Asseng & Frank Asche (2019). Future farms without farmers. Science Robotics, 4, (27), 1-2, doi:10.1126/scirobotics.aaw1875.

Example Precision Farming: Data $\mathcal{D}$, Features Θ

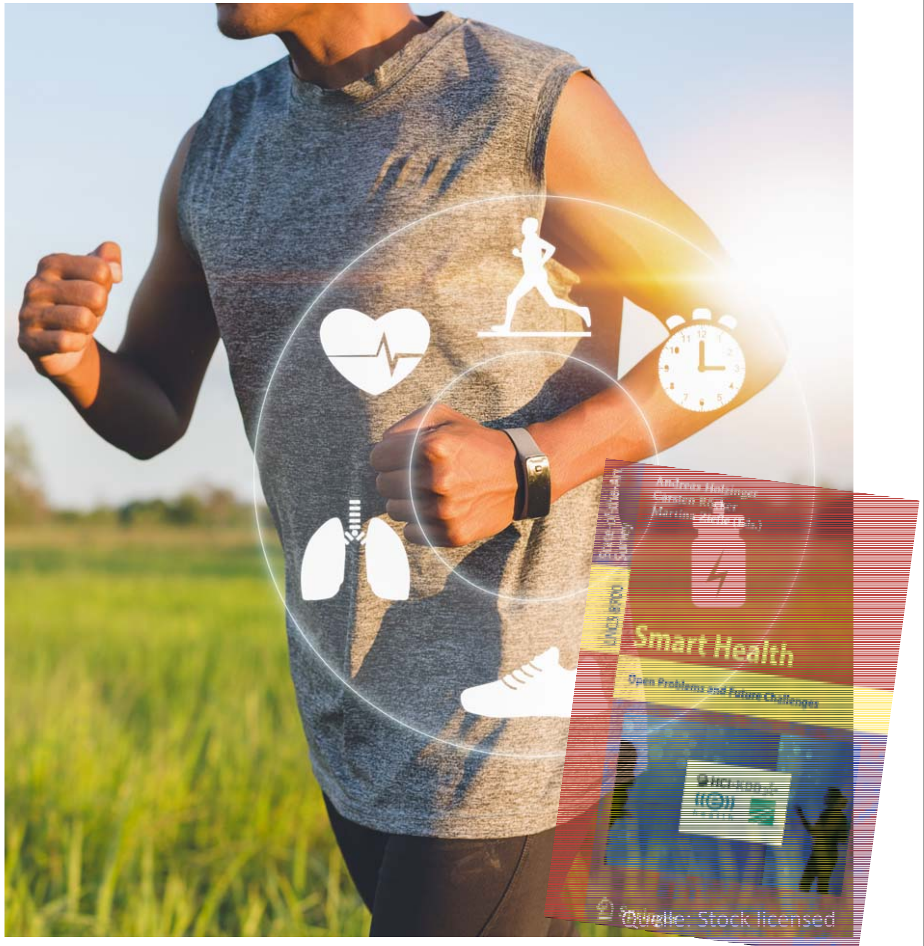


Heyu Yin, Yunteng Cao, Benedetto Marelli, Xiangqun Zeng, Andrew J. Mason & Changyong Cao (2021). Soil Sensors and Plant Wearables for Smart and Precision Agriculture. Advanced Materials, 33, (20), 2007764, doi:10.1002/adma.202007764.

Smart Farming – Smart Health



Quelle: Yin et. al (2021) doi: 10.1002/adma.202007764

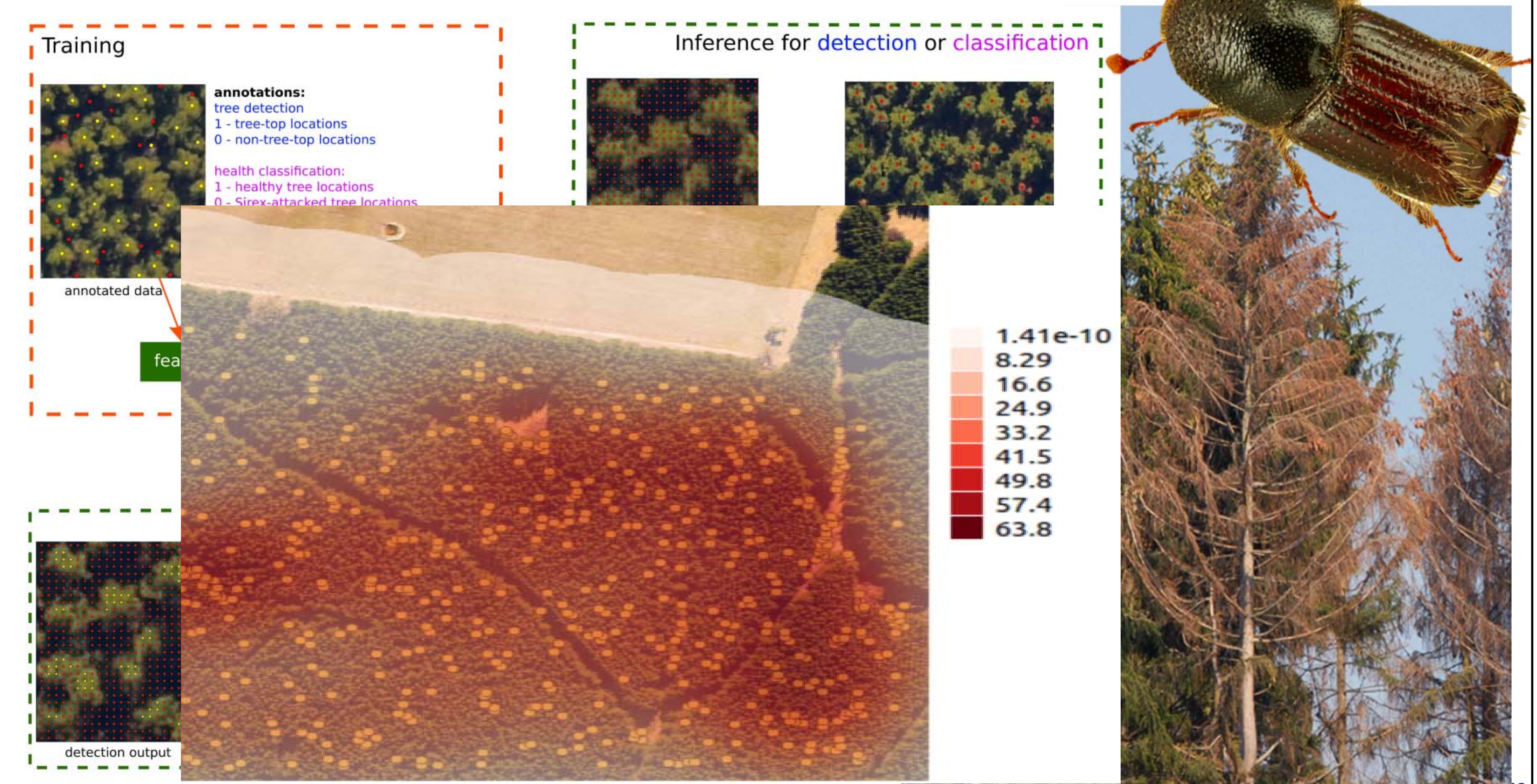


https://human-centered.ai/lncs-8700-smart-health

IMLW @ AAAI'22

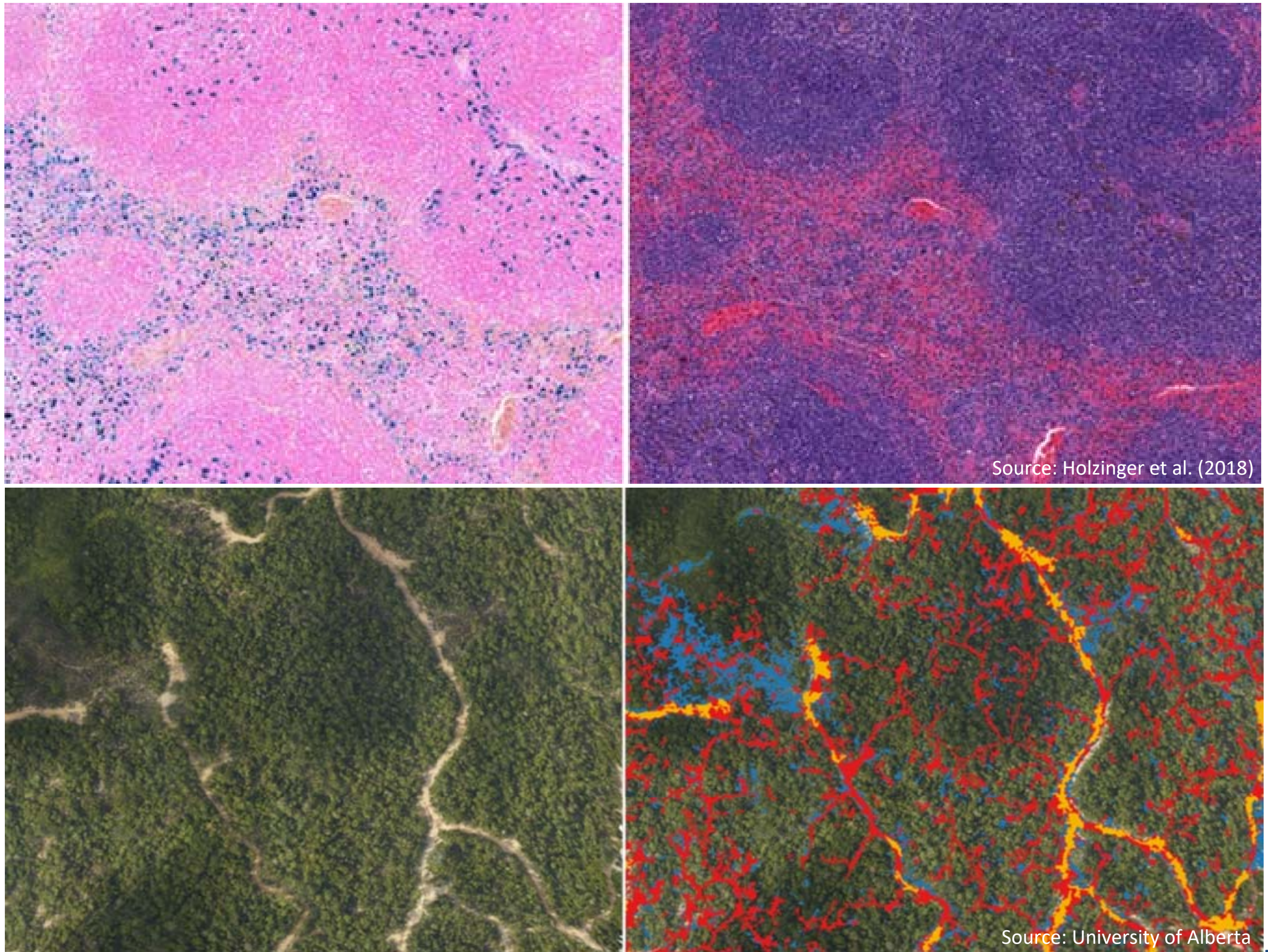
Example Precision Forestry: Daten $\mathcal{D}$, Features Θ, ...

HCAI HUMAN-CENTERED.AI



Tumor growth prediction vs. forest wildfire detection

Fleur Jeanquartier, Claire Jean-Quartier, David Cemernek & Andreas Holzinger (2016). In silico modeling for tumor growth visualization. BMC Systems Biology, 10, (1), 1-15, doi:10.1186/s12918-016-0318-8.



Robert S. Allison, Joshua M. Johnston, Gregory Craig & Sion Jennings (2016). Airborne optical and thermal remote sensing for wildfire detection and monitoring. Sensors, 16, (8), 1310, doi:10.3390/s16081310.

Pierre Simon de Laplace (1749-1827)

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

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

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

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

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

The inverse probability allows us to learn from data, infer unknowns, and 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 (1812)

Reasoning under uncertainty: Decision Making

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

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

Example „Forest Monitoring“

Source: Stock licensed

(2) Challenges ...

IMLW @ AAAI'22

Herausforderungen: „Adversarial Examples“

HCAI HUMAN-CENTERED.AI

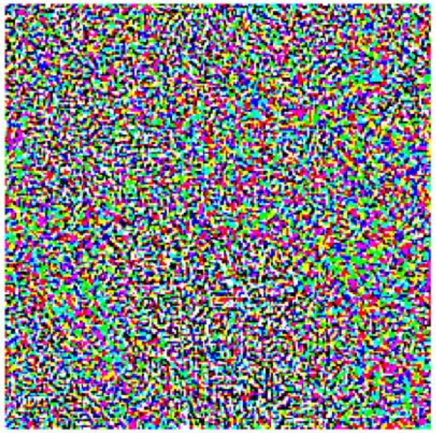
See also: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Robustness & Explainability

classified as

Stop Sign

$+ .007 \times$



$\mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

classified as

Max Speed 100

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy (2014). Explaining and harnessing adversarial examples. arXiv:1412.6572
Traffic Sign Examples Image Credit to Jiefeng Chen & Xi Wu (2019). https://www.altacognita.com/robust-attribution

What is adversarial machine learning ?

HCAI HUMAN-CENTERED.AI

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.

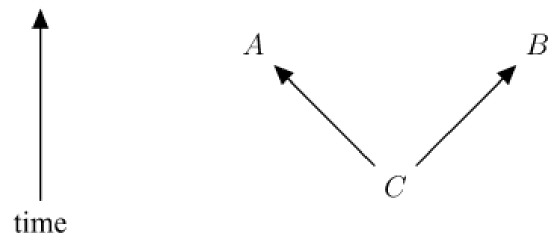
ML 101: What do we need to reach robust intelligence ?

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

(3) Correlation ≠ 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



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

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

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

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

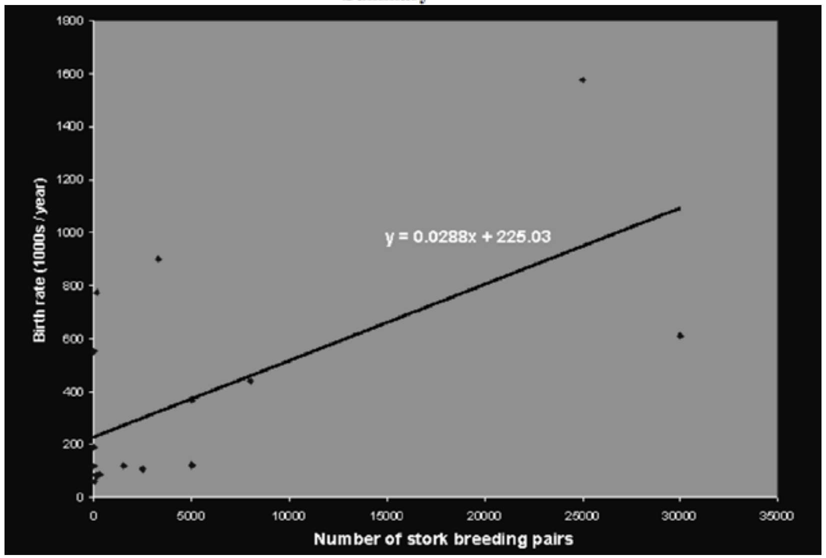
Remember: Correlation is NOT Causality

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.

How can we make AI more robust ?

$$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2 = a^T V_{XX}\, a.$$

Generalization error

Generalization
plus human ex

iML = human inspection – bring in human conceptual knowledge

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.

How do humans generalize from a few examples ?

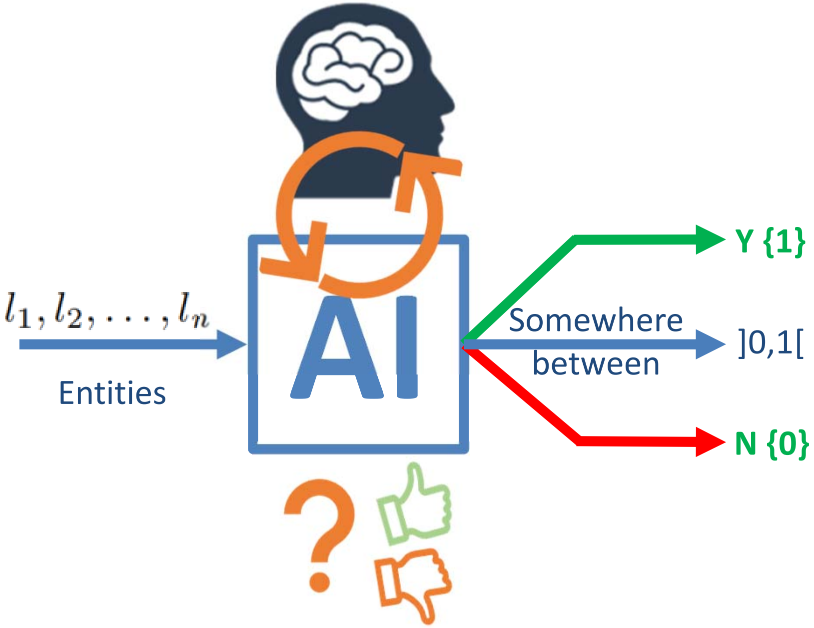
(Sometimes – **not** always!) humans are able ...

- to understand the context
- to make inferences from little, noisy, incomplete data sets
- to learn relevant representations
- to find shared underlying explanatory factors,
- with a causal reasoning $P(Y|X)\ Y \rightarrow X$ (predict cause from effect) or $P(Y|X)\ X \rightarrow Y$ (predict effect from cause)

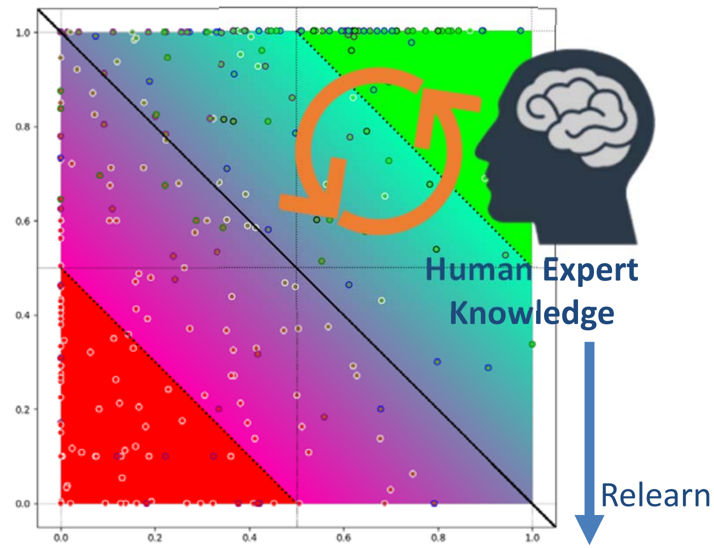
Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths & Noah D. Goodman 2011. How to grow a mind: Statistics, structure, and abstraction. *Science,* 331, (6022), 1279-1285, doi:10.1126/science.1192788.

Enabling Actionable Explainable AI: Adaptive classification

The Problem



Our Solution

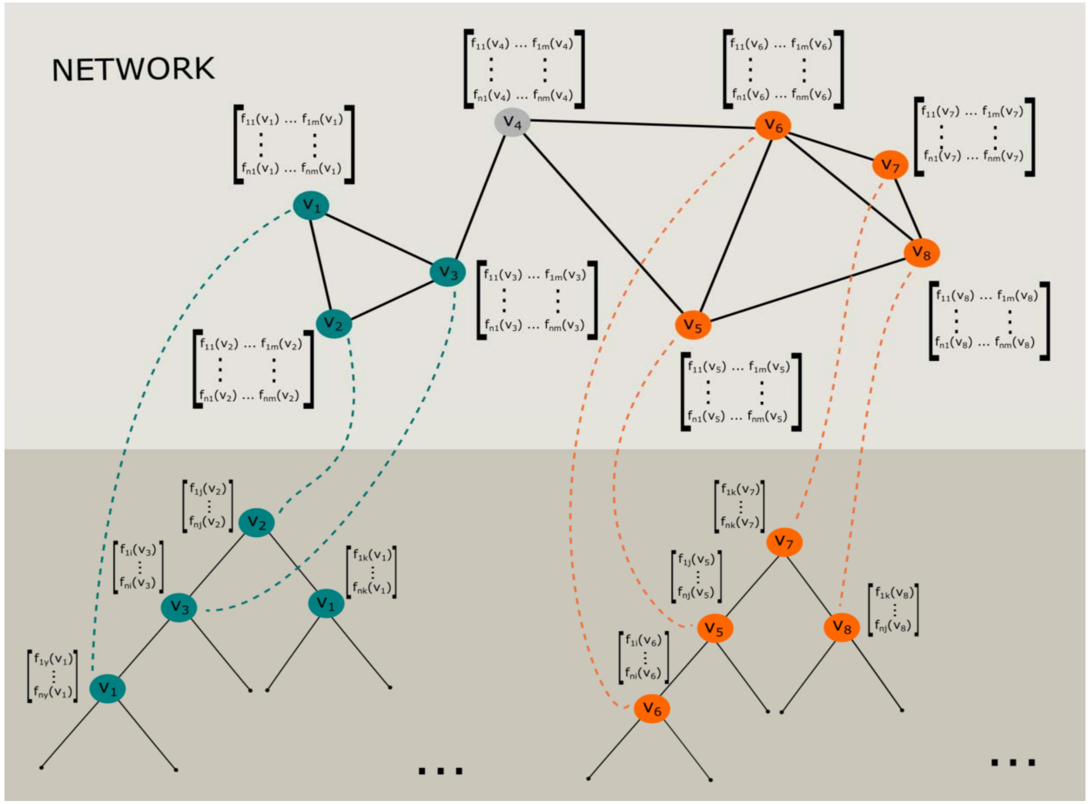


Y: $S_k(x,y) = (\min(1, x^k + y^k - 0.5^k))^{\frac{1}{k}}$

N: $T_k(x,y) = (\max(1, x^k + y^k - 0.5^k))^{\frac{1}{k}}$

Miroslav Hudec, Erika Minarikova, Radko Mesiar, Anna Saranti & Andreas Holzinger (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. *Knowledge Based Systems,* 220, doi:10.1016/j.knosys.2021.106916.

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

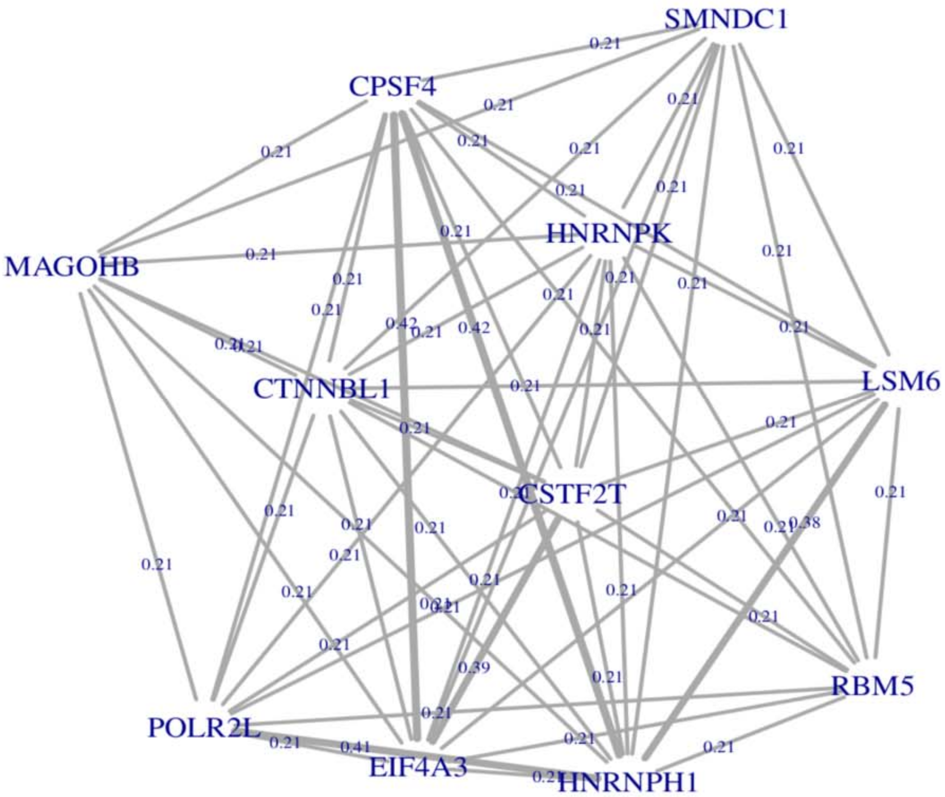


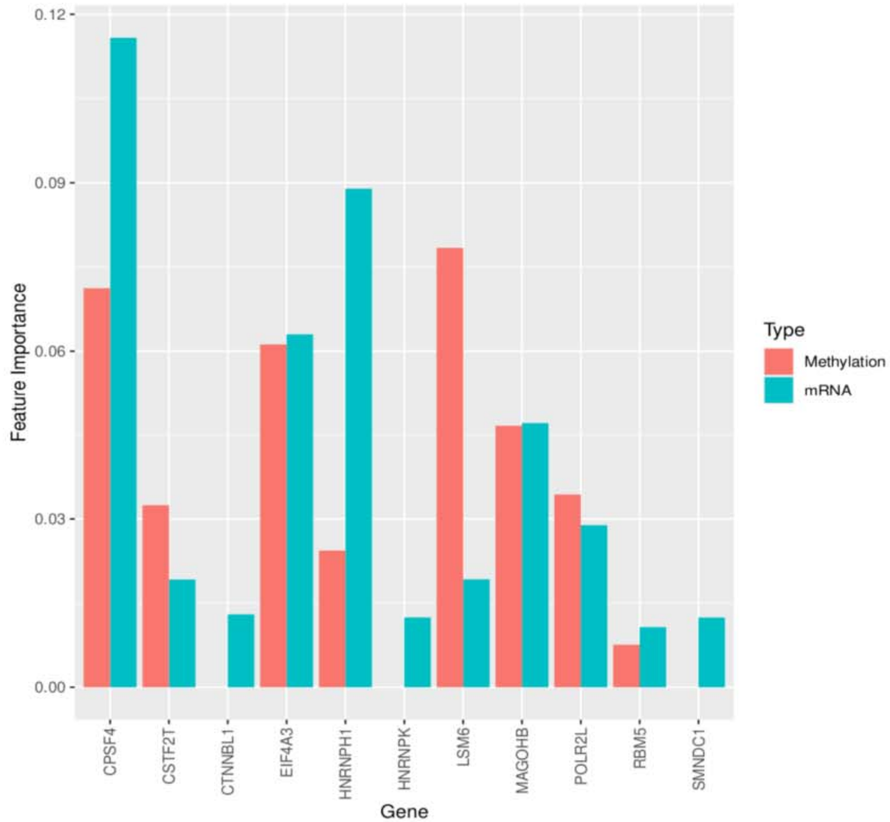
Algorithm 1: Greedy Decision Forest

Given a Decision Forest: $\{T_k(x, \Theta_k, X_k^G)\}$;
Given a graph: $G = (V, E)$;
$k = \{1, \ldots, ntree\}, t = 1$;
$mtry_k = \sqrt{\#V}$;
$X_k^G[t = 1] = X_k^G$;
while $t \leq niter$ **do**
  for $k \leftarrow 1$ **to** $ntree$ **do**
    $Perf(T_k[t]) = Performance\ of\ T_k(x; \Theta_k, X_k^G)$;
    if $Perf(T_k[t]) \leq Perf(T_k[t-1])$ **then**
      $T_k[t] = T_k[t-1]$;
      $X_k^G[t] = X_k^G[t-1]$;
    else
      $mtry_k[t] - -$;
      $X_k^G[t] = RandomWalk(G|X_k, mtry_k[t])$;
    end
  end
  Sample ntree trees according to $Perf\{T_k[t]\}$;
  $t + +$;
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.

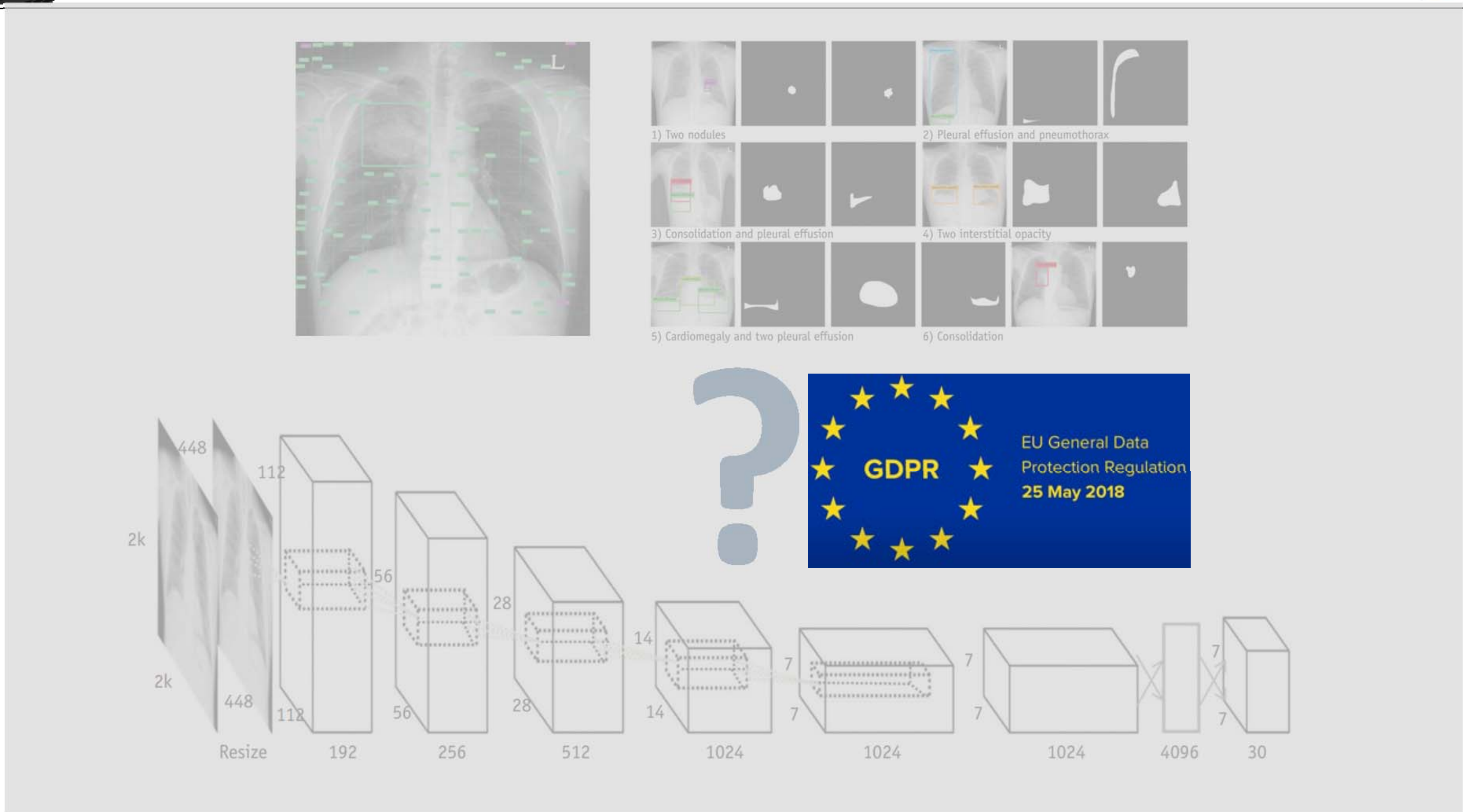
actionable explainable AI – multi-modality in network medicine





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.

New Regulations – Right for explanation



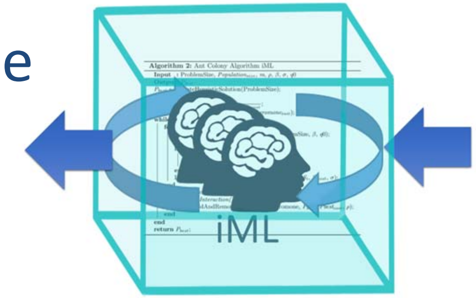
Karl Stoeger, David Schneeberger & Andreas Holzinger (2021). Medical Artificial Intelligence: The European Legal Perspective. Communications of the ACM, 64, (11), doi:10.1145/3458652.

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.

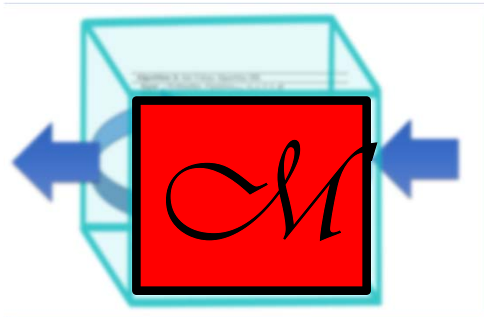
(4) Methods of Explainability

What are interpretable models vs. interpreting models ?

- **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*.

What are typical examples of post-hoc methods of explainable AI ?

- 1) Gradients
- 2) Sensitivity Analysis
- 3) Simple Taylor expansions
- 4) Decomposition and Relevance Propagation (Pixel-RP, Layer-RP, Deep Taylor Decomposition, …)
- 5) Excitation Backpropagation
- 6) Optimization (LIME, BETA, Smooth Grad, ...) BETA transparent approximation, …)
- 7) Deconvolution (Occlusion-based, meaningful perturbations, ...)
- 8) Qualitative Testing with Concept Activation Vectors TCAV

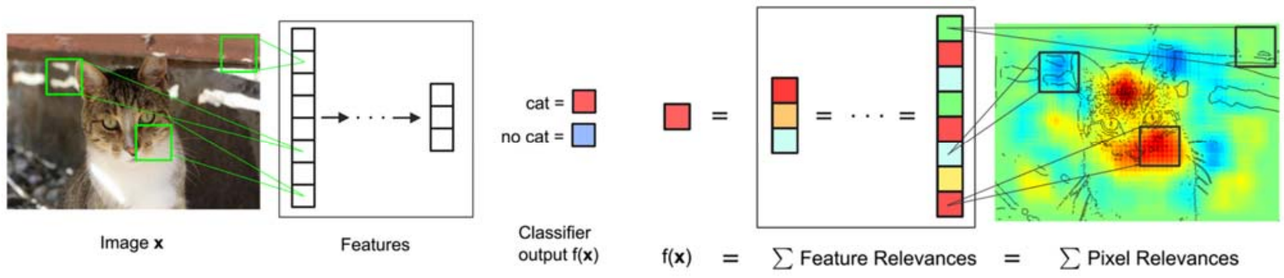
Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course
https://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

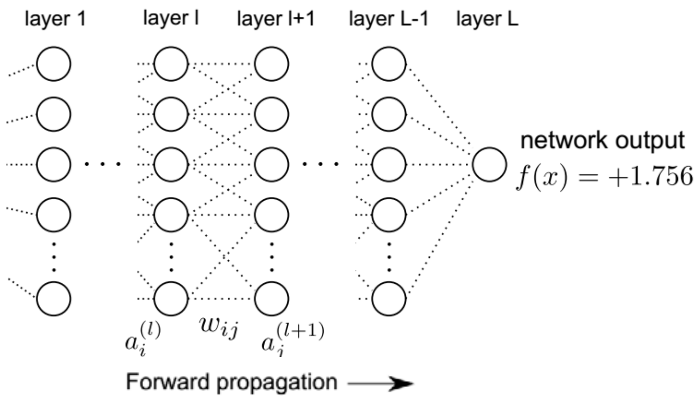
LRP Layer-Wise Relevance Propagation

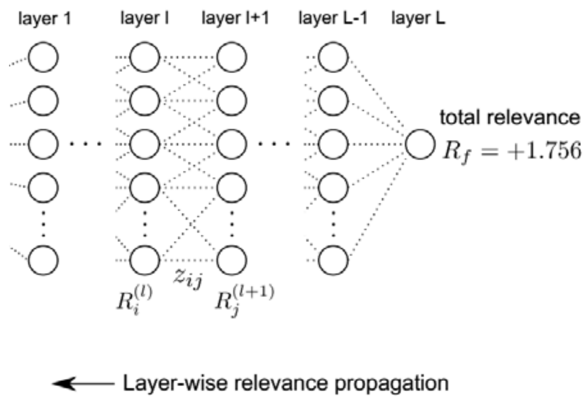
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

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

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



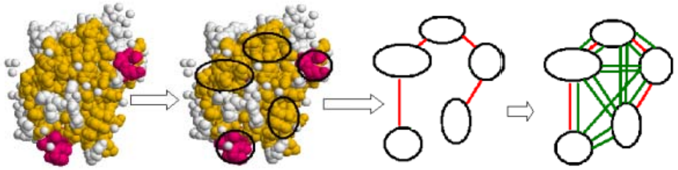




$$\boxed{a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)}$$

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

LRP works also on graphs

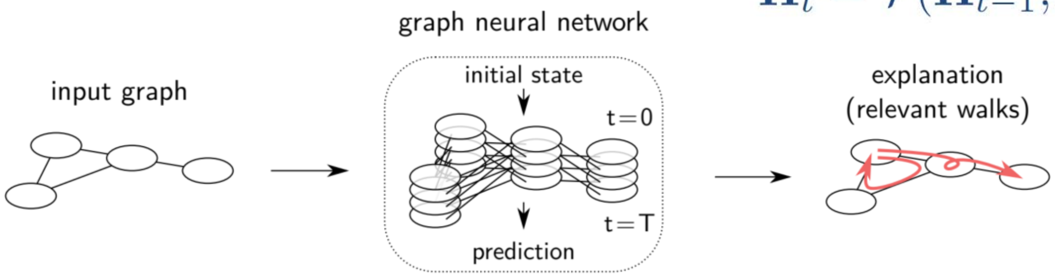


Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, Svn Vishwanathan, Alex J Smola & Hans-Peter Kriegel (2005). Protein function prediction via graph kernels. *Bioinformatics*, 21, (suppl 1), i47-i56.

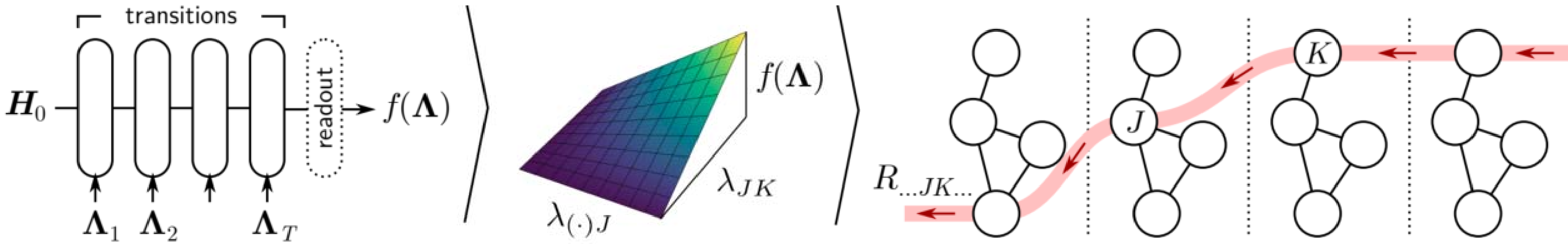
$\mathcal{G}$ ... input graph $\quad G = (\mathcal{V}, \mathcal{E}) \quad \mathcal{V} = \{v_1, ..., v_n\} \quad \mathcal{E} \subseteq \{(v_i, v_j) | v_i, v_j \in \mathcal{V}\}$

$\mathbf{H}_0$ ... initial state

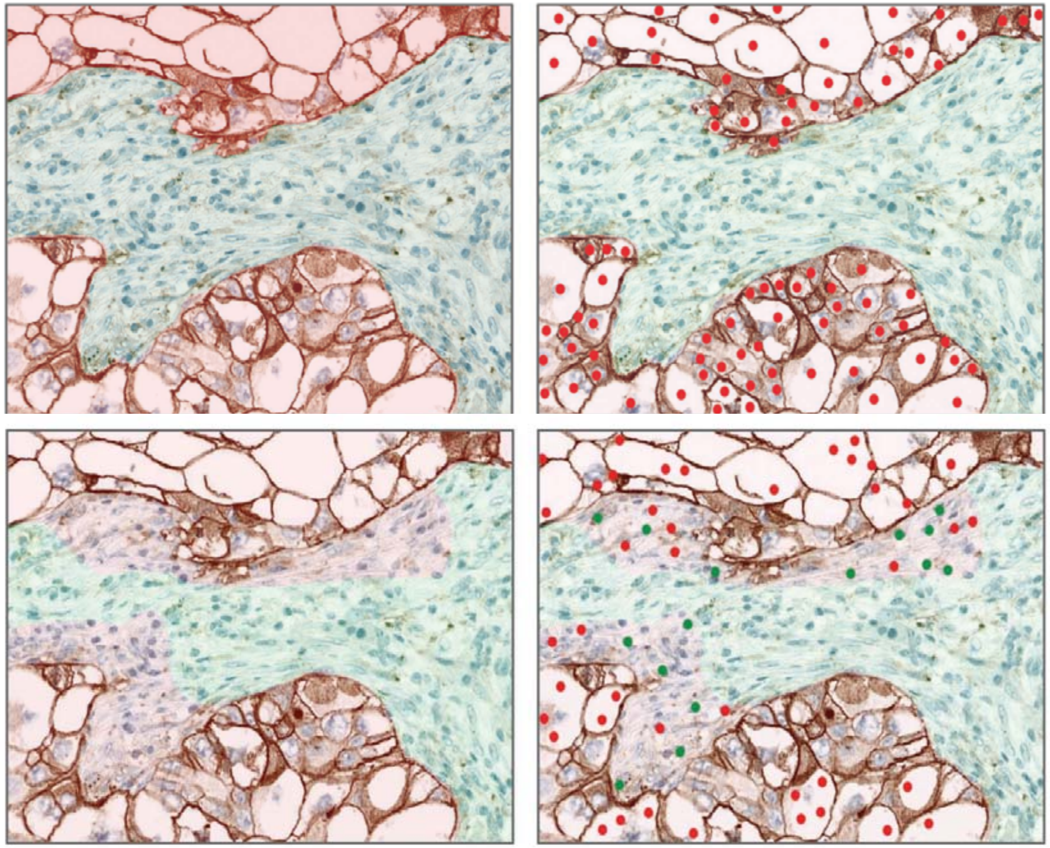
$$\mathbf{H}_t = \mathcal{T}(\mathbf{H}_{t-1}, \mathbf{\Lambda}_t, \mathbf{W}_t)$$



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*.



Medical Example why Explainable AI is important



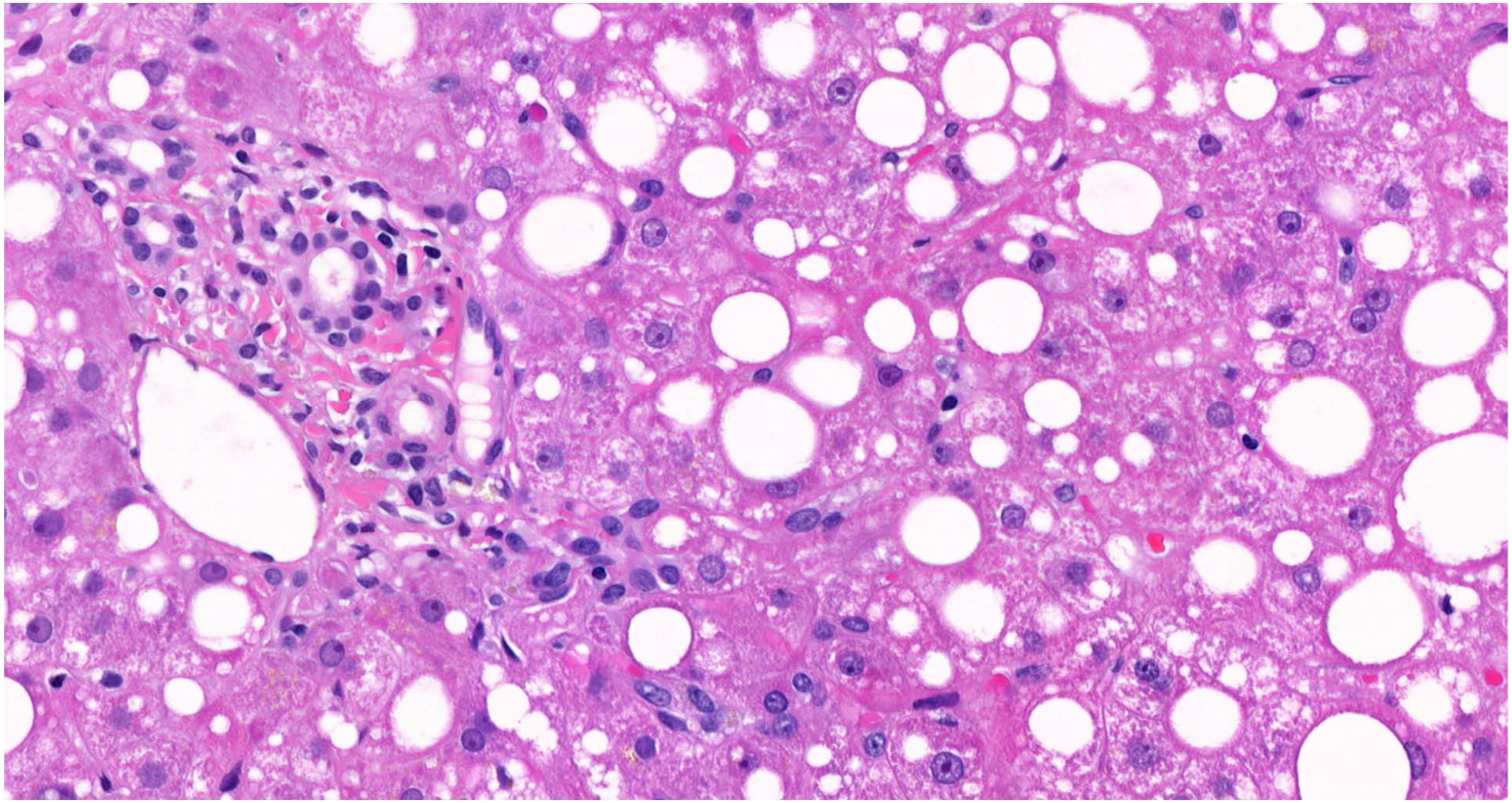
Andreas Holzinger & Heimo Mueller (2021). Toward Human-AI Interfaces to Support Explainability and Causability in Medical AI. *IEEE COMPUTER,* 54, (10), doi:10.1109/MC.2021.3092610.

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

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: How do human pathologists make diagnoses ?

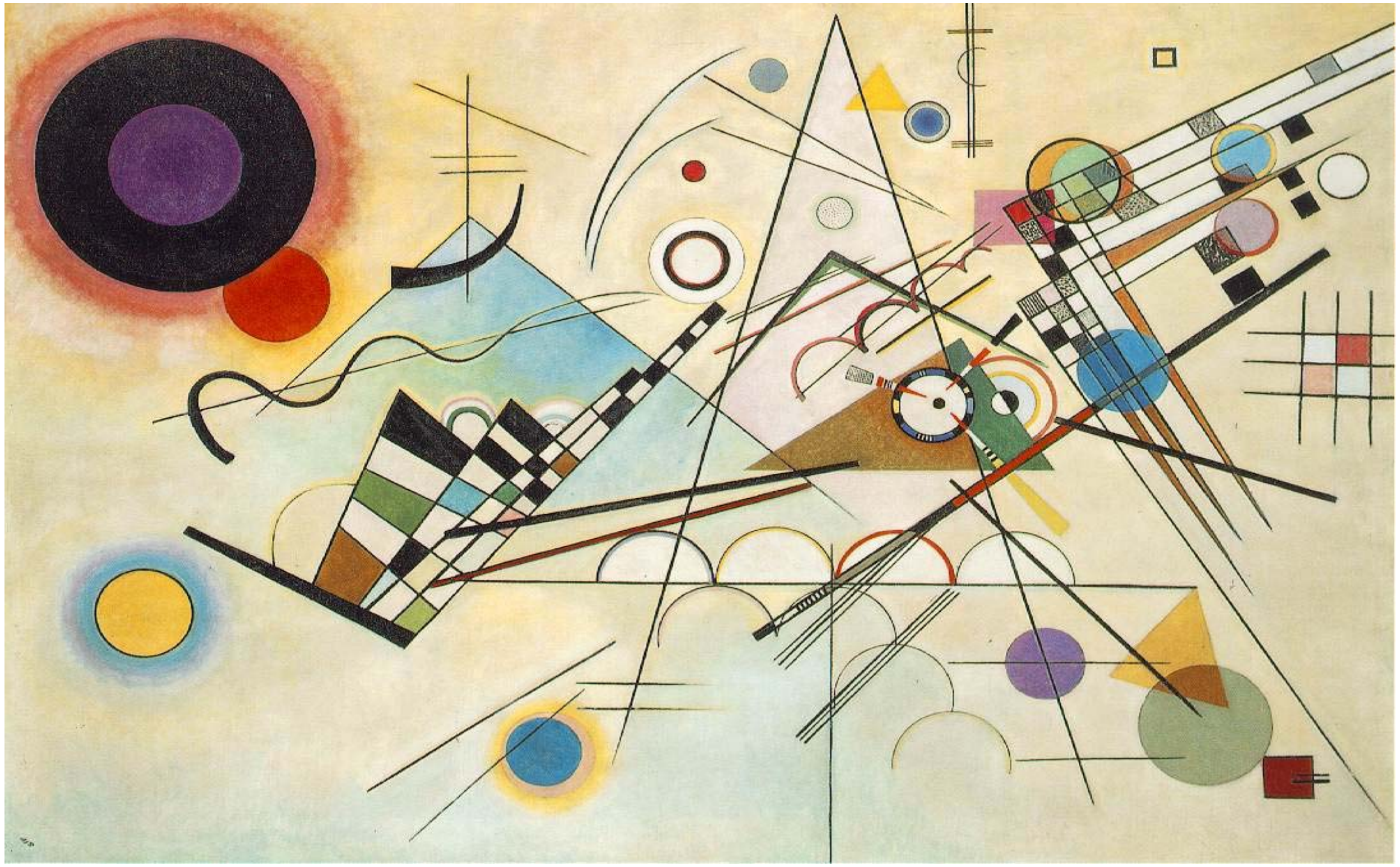


What is ground truth ? Where is the ground truth ?

- := 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

Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, Jiang Bian (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Mach.Intelligence, 2, (7), 369-375, doi:10.1038/s42256-020-0197-y

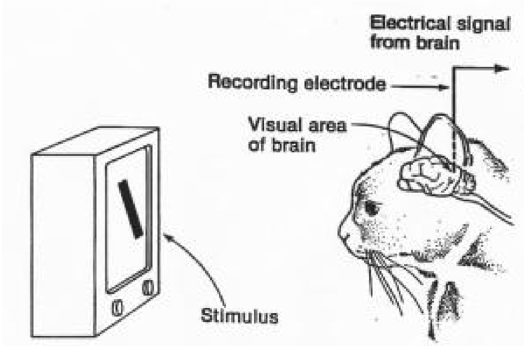
KANDINSKYPatterns – where does the name come from ?



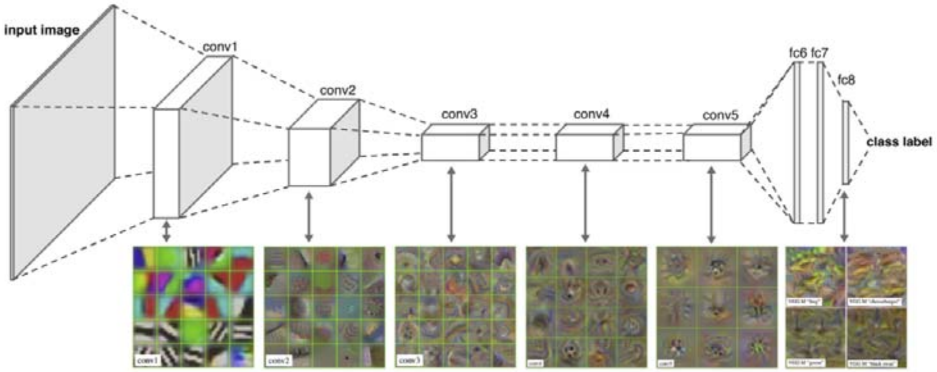
Wassily Kandinsky
(1866 – 1944)

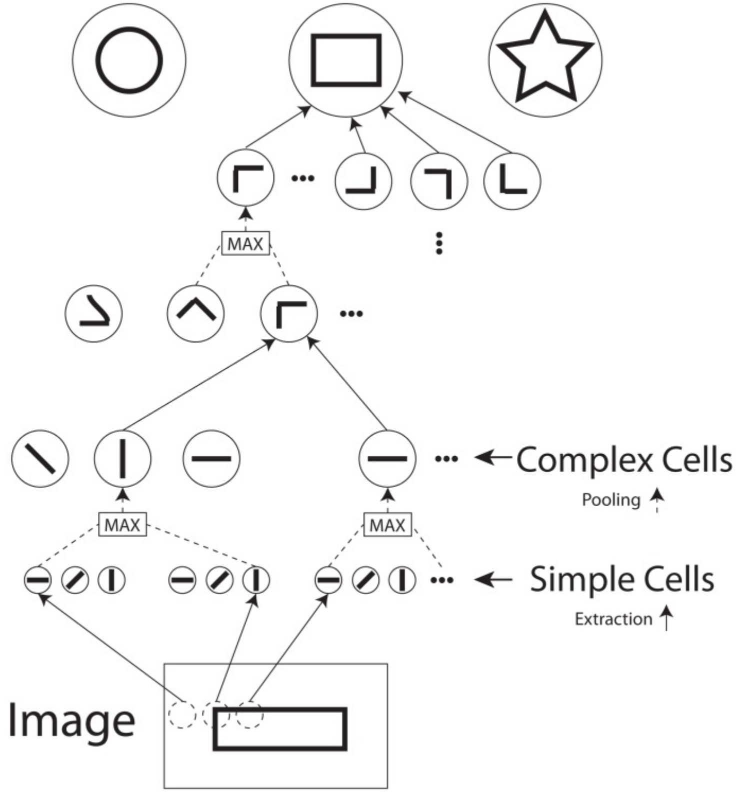
Komposition VIII, 1923, Solomon R. Guggenheim Museum, New York. Source: https://de.wikipedia.org/wiki/Wassily_Kandinsky
Note: Image is in the public domain and is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

Hubel & Wiesel (1962): 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

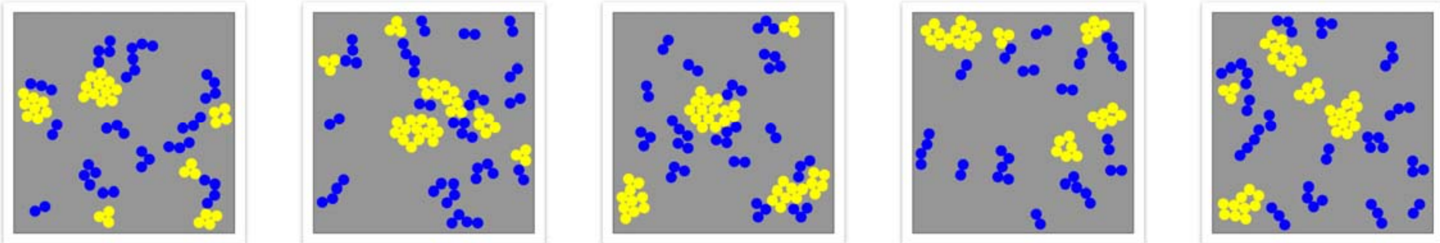




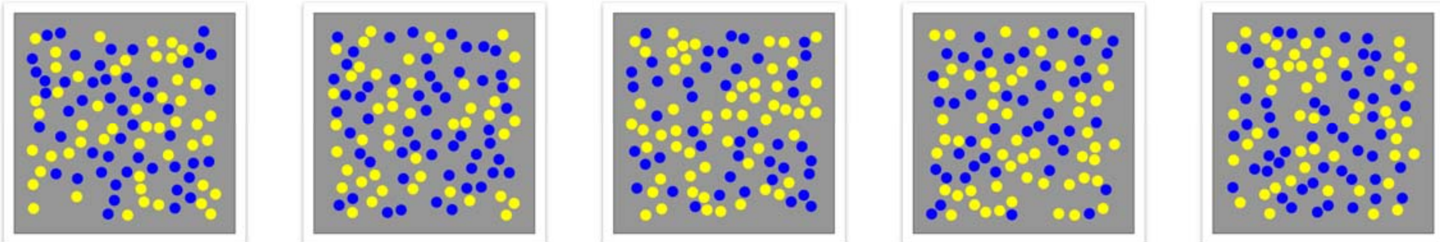
Source: https://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/models-of-information-processing-in-the-visual-cortex

#KANDINSKYPatterns

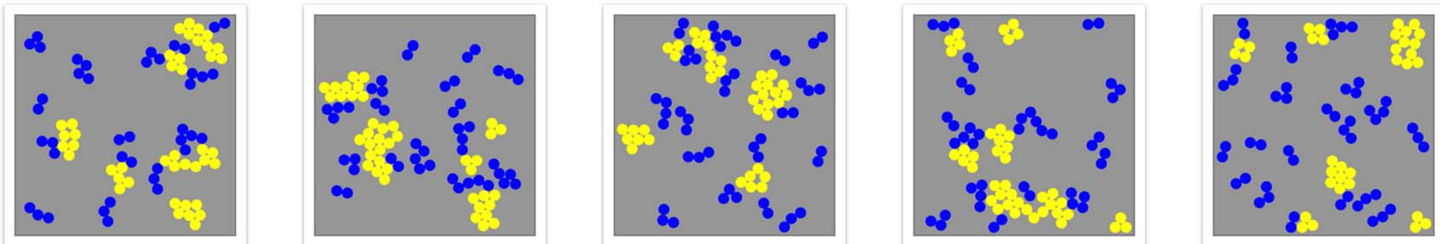
A) True (the cells are smaller and closer together – it is an tumor ...)



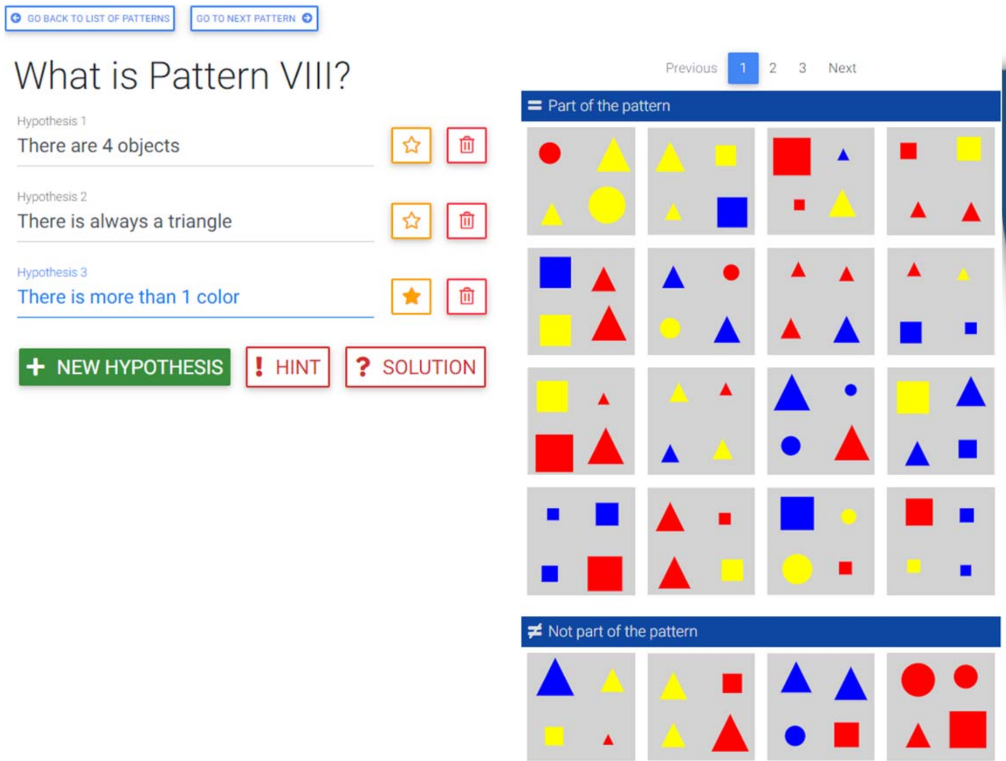
B) False



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



"Intelligence Test for Machines"



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.

Related Work (1): Bongard Problems

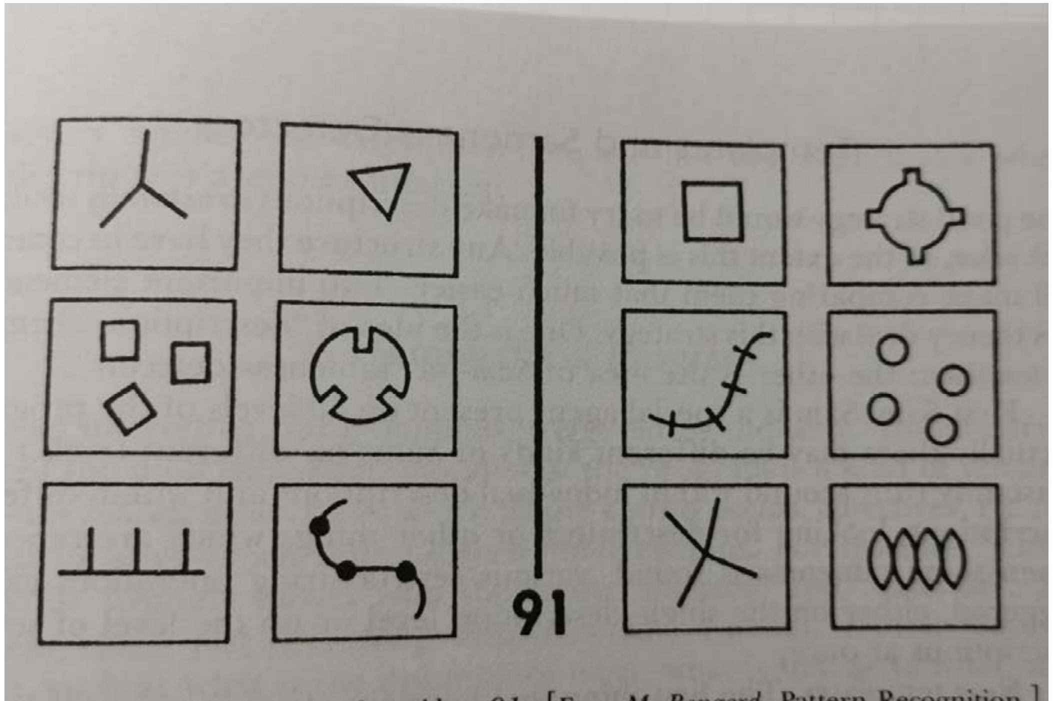
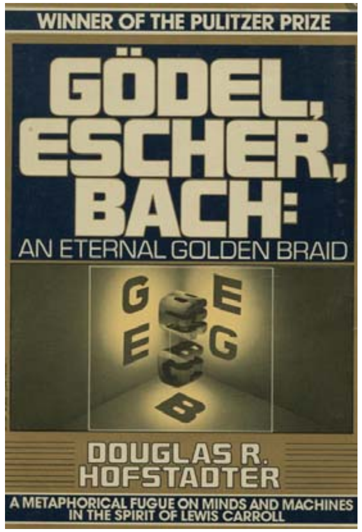


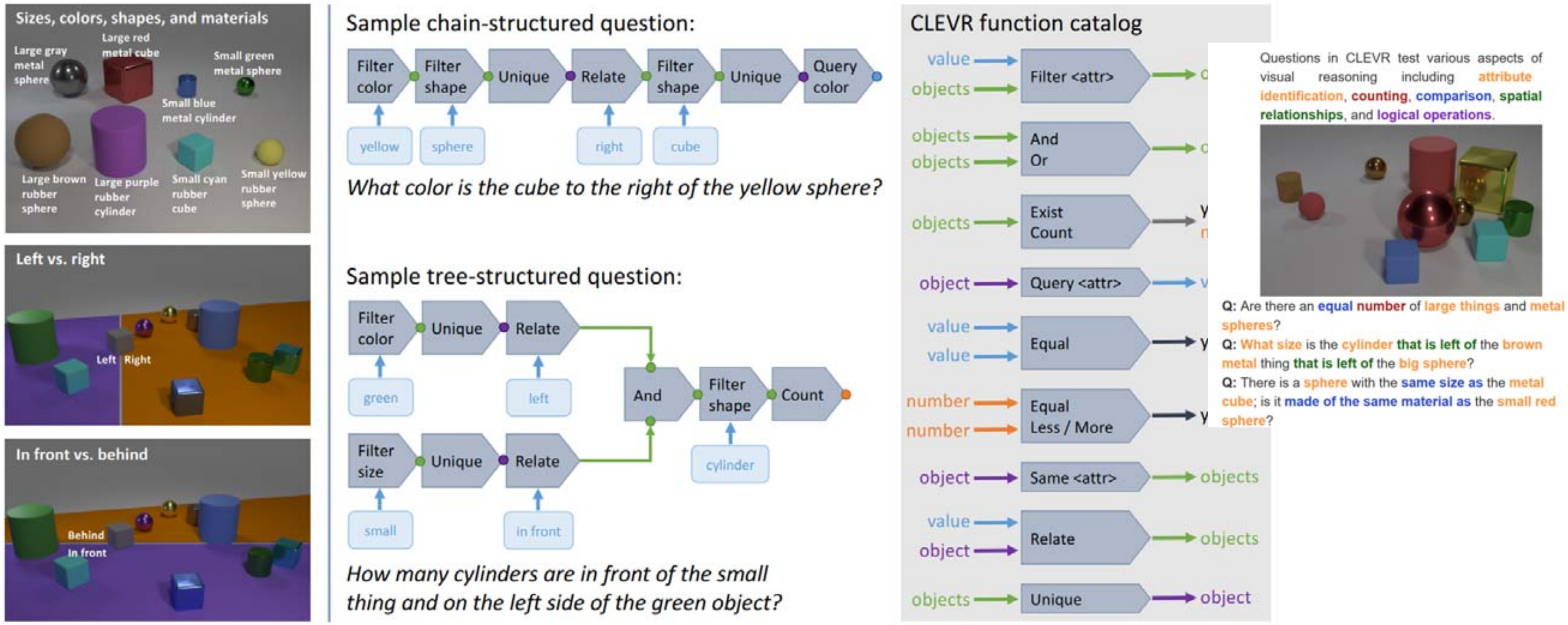
FIGURE 121. Bongard problem 91. [From M. Bongard, Pattern Recognition.]



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)

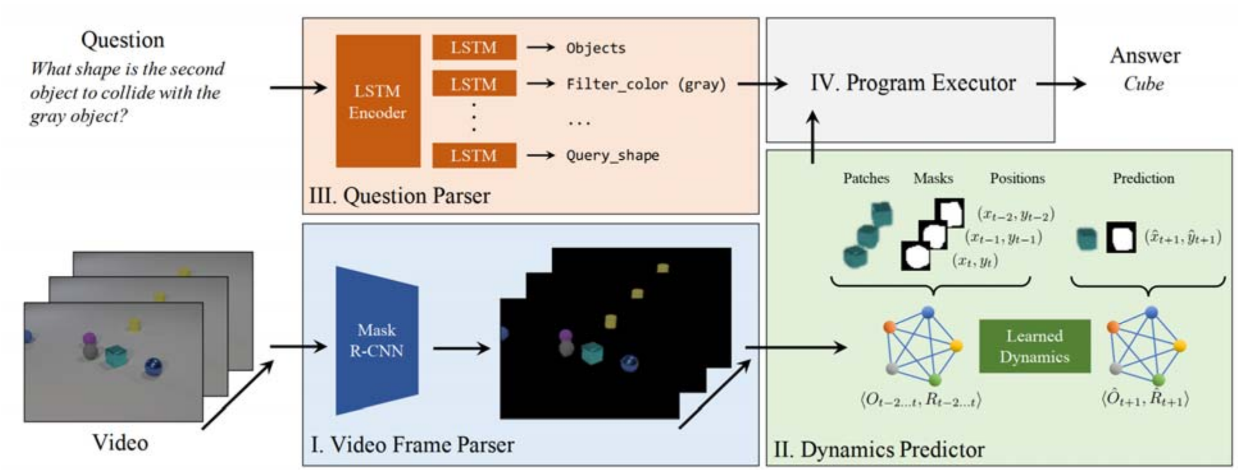
Related Work (2): CLEVR

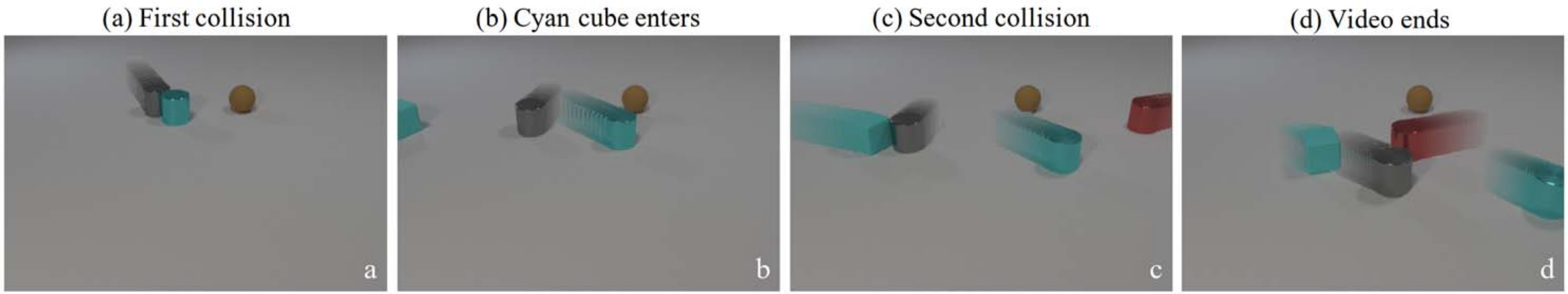


https://cs.stanford.edu/people/jcjohns/clevr/

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.

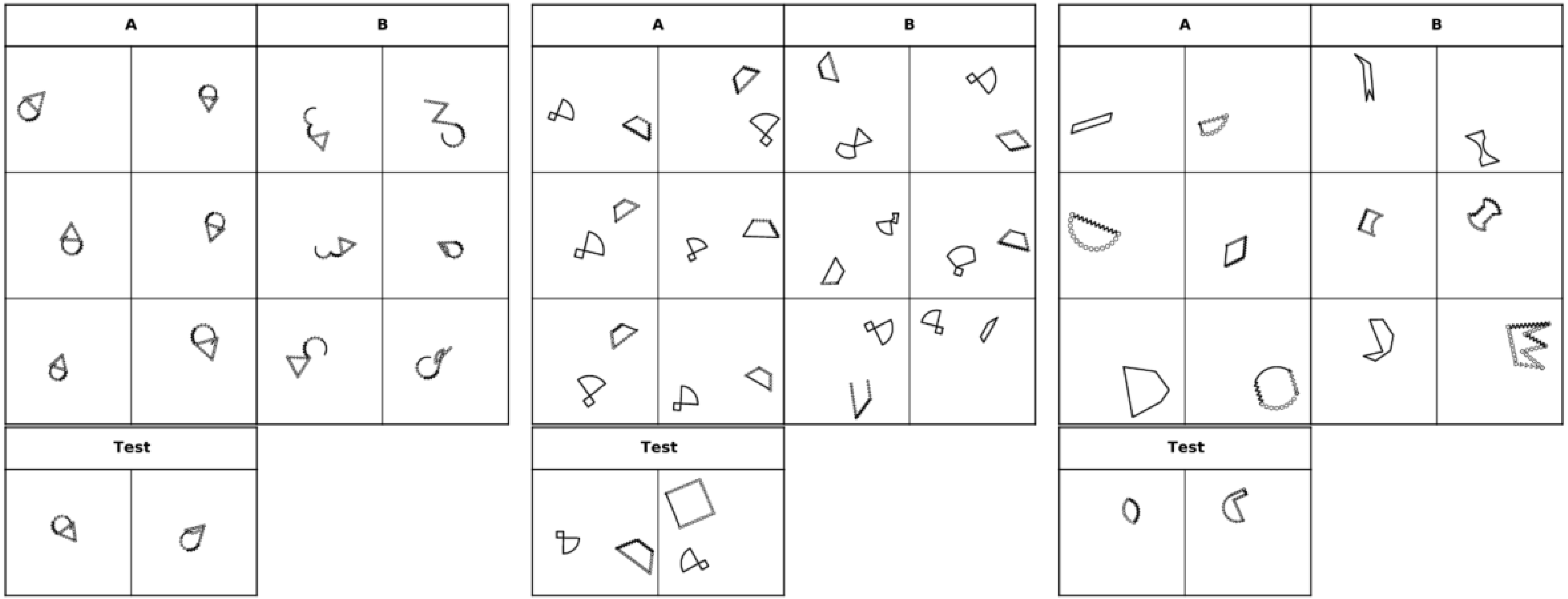
Related Work (3): CLEVRER





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.

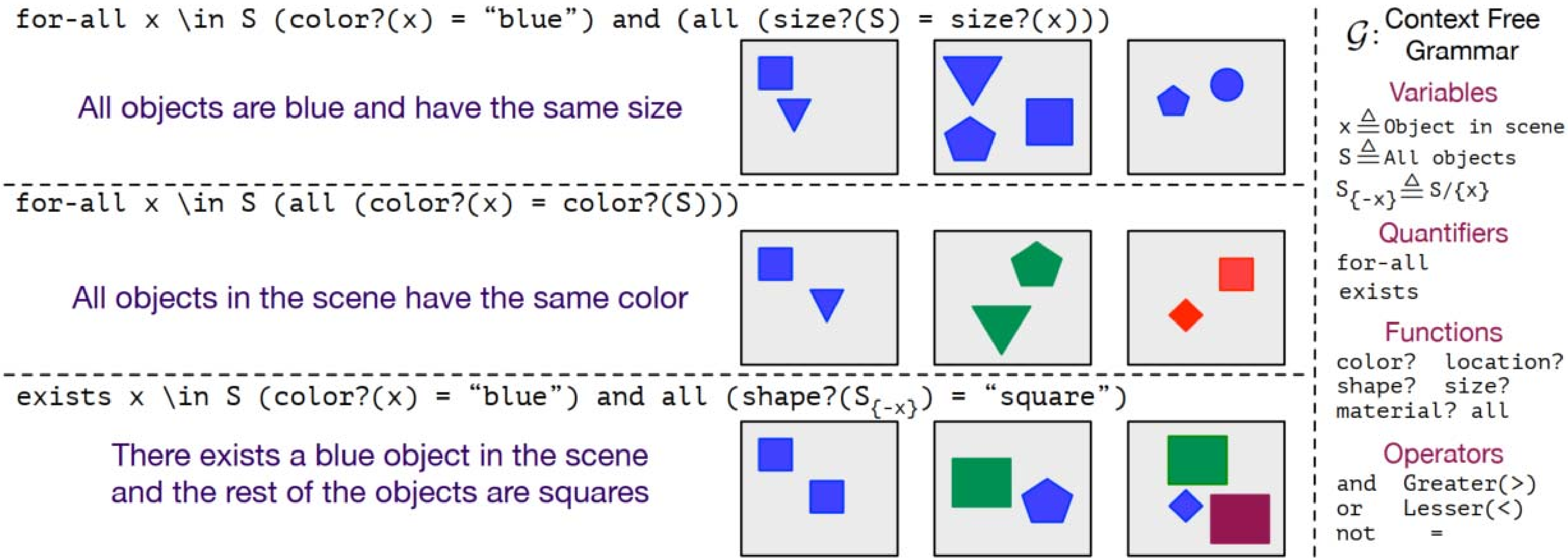
Related Work (4): Bongard-LOGO



(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.

Related Work (5): CURI



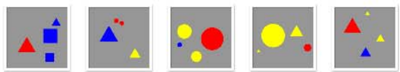
Ramakrishna Vedantam, Arthur Szlam, Maximilian Nickel, Ari Morcos & Brenden Lake (2020).
CURI: A Benchmark for Productive Concept Learning Under Uncertainty. arXiv:2010.02855.

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

Home About Holzinger Group For Experts For Students Open Work (2020) Partners News

HCAI HUMAN-CENTERED.AI

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



ABSTRACT

KANDINSKYPatterns (yes, 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 KANDINSKY Patterns 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.

https://arxiv.org/abs/2103.00519

How can we measure the quality of explanations ?

Measuring the quality of Explanations: The Systems Causa<u>bility</u> Scale

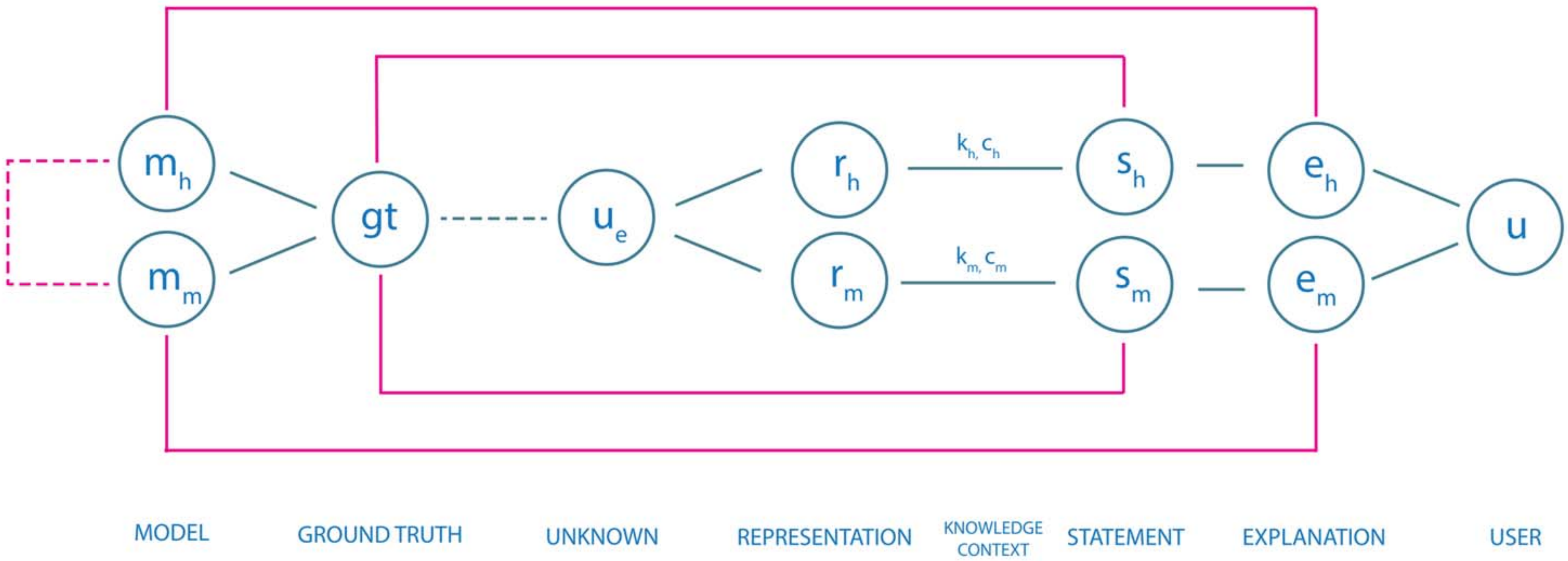
Andreas Holzinger, Andre Carrington & Heimo Müller 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited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

Definitionen: Explainability vs. Causability

- 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.

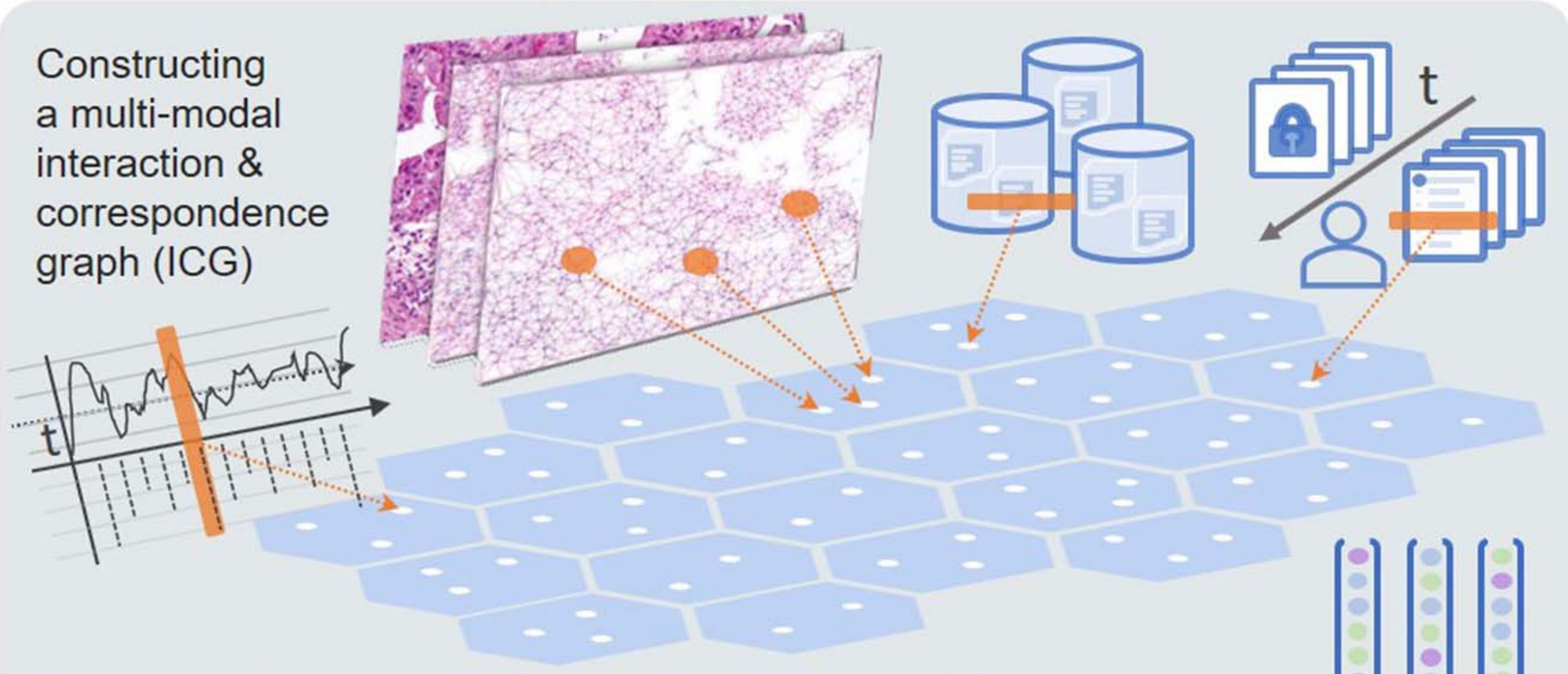
How can we measure the quality of explanations ?



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.

Conclusio

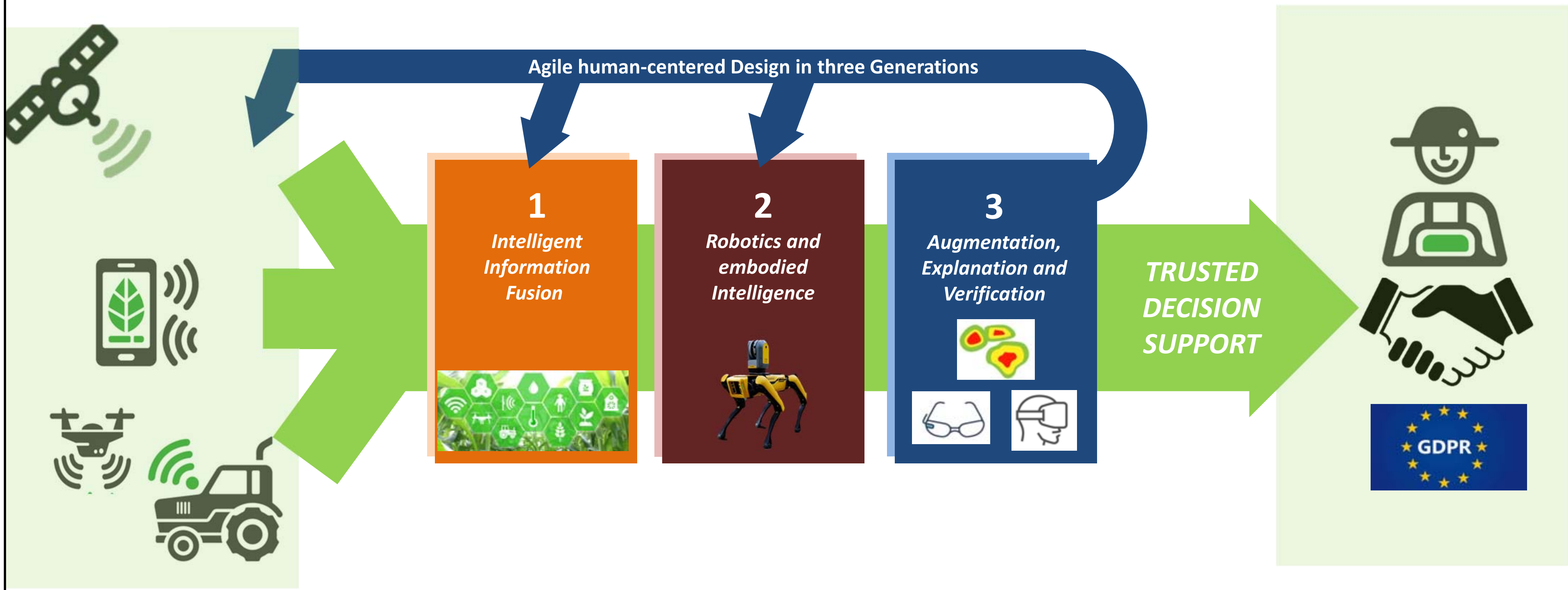
Multimodal Causability: enabling *why* and *what-if* ...



Interesting signals from each modality (time-based, image, structured & unstructured) are connected according to pre-defined rules. Each modality's features lie in their own, un-aligned concept spaces.

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.

Three Frontier Research Areas



Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Carl Orge Retzlaff, Andreas Gronauer, Viktoria Motsch, Christoph Gollob, Karl Stampfer (2022). Digital Transformation in Smart Farm and Forest Operations needs Human-Centered AI: Challenges and Future Directions. Sensors (in print)

Human-Centered AI aligns AI with human values, ethical principles and legal requirements.

Bibliography (selected)

- Andreas Holzinger, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, Wojciech Samek, Igor Jurisica & Natalia Díaz-Rodríguez (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, (3), 263-278, doi:10.1016/j.inffus.2021.10.007.
- Bastian Pfeifer, Afan Secic, Anna Saranti & Andreas Holzinger (2022). GNN-SubNet: disease subnetwork detection with explainable Graph Neural Networks. bioRxiv, 1--8, doi:10.1101/2022.01.12.475995
- Andre M. Carrington, Douglas G. Manuel, Paul W. Fieguth, Tim Ramsay, Venet Osmani, Bernhard Wernly, Carol Benett, Steven Hawken, Matthew Mcinnes, Olivia Magwood, Yusuf Sheikh & Andreas Holzinger (2022). Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Early Access, doi:10.1109/TPAMI.2022.3145392.
- Karl Stoeger, David Schneeberger & Andreas Holzinger (2021). Medical Artificial Intelligence: The European Legal Perspective. Communications of the ACM, 64, (11), doi:10.1145/3458652
- Karl Stoeger, David Schneeberger, Peter Kieseberg & Andreas Holzinger (2021). Legal aspects of data cleansing in medical AI. Computer Law and Security Review, 42, 105587, doi:10.1016/j.clsr.2021.105587
- Jianlong Zhou, Amir H. Gandomi, Fang Chen & Andreas Holzinger (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics, 10, (5), 593, doi:10.3390/electronics10050593
- Heimo Mueller, Michaela T. Mayrhofer, Evert-Ben Van Veen & Andreas Holzinger (2021). The Ten Commandments of Ethical Medical AI. IEEE COMPUTER, 54, (7), 119--123, doi:10.1109/MC.2021.3074263
- Miroslav Hudec, Erika Minarikova, Radko Mesiar, Anna Saranti & Andreas Holzinger (2021). Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. Knowledge Based Systems, 220, 106916, doi:10.1016/j.knosys.2021.106916
- 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
- David Schneeberger, Karl Stoeger & Andreas Holzinger (2020). The European legal framework for medical AI. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer LNCS 12279. Cham: Springer, pp. 209--226, doi:10.1007/978-3-030-57321-8_12
- Anna Saranti, Behnam Taraghi, Martin Ebner & Andreas Holzinger (2020). Property-Based Testing for Parameter Learning of Probabilistic Graphical Models. Springer Lecture Notes in Computer Science LNCS 12279. Cham: Springer, pp. 499--515, doi:10.1007/978-3-030-57321-8-28
- Peter Regitnig, Heimo Mueller & Andreas Holzinger (2020). Expectations of Artificial Intelligence in Pathology. Springer Lecture Notes in Artificial Intelligence LNAI 12090. Cham: Springer, pp. 1-15, doi:10.1007/978-3-030-50402-1-1
- Shane O'sullivan, Simon Leonard, Andreas Holzinger, Colin Allen, Fiorella Battaglia, Nathalie Nevejans, Fijs W.B. Van Leeuwen, Mohammed Imran Sajid, Michael Friebe, Hutan Ashrafian, Helmut Heinsen, Dominic Wichmann & Margaret Hartnett (2020). Anatomy 101 for AI-driven robotics: Explanatory, ethical and legal frameworks for development of cadaveric skills training standards in autonomous robotic surgery/autopsy. The International Journal of Medical Robotics and Computer Assisted Surgery, doi:10.1002/rcs.2020
- Heimo Müller, Peter Regitnig, Peter Ferschin, Anna Saranti & Andreas Holzinger. Classification and Visualization of Patterns in Medical Images. 24th International Conference on Information Visualization (IV), (2020) Melbourne. IEEE, 639--643, doi:10.1109/IV51561.2020.00110.