Mini Course

From Data Science to interpretable AI

Assoc. Prof. Dr. Andreas HOLZINGER (Medical University Graz)

Day 1 > Part 4 > Monday, 17.06.2019

Causal Reasoning and Interpretable AI

This is the version for printing and reading. The lecture version is didactically different.

Overview

Day 1 - Fundamentals

01 Introduction to AI
Machine Learning

02 Data, Information
and Knowledge

03 Decision Making and
Decision Support

04 Causal Reasoning
and Interpretable AI

Keywords

- Causality, Causability
- Causal Reasoning
- Concept activation Vector
- Explainability, Interpretability
- Explainable AI
- Generalization
- Human-in-Control
- Kandinsky Patterns
- Layer-wise relevance propagation

Advance Organizer (1/2)

- **Causality** = fundamental relationship between cause and effect
- **Causability** = similar to the concept of usability the property of a human explanation
- **Collective Intelligence** = shared group (symbolic) intelligence, emerging from cooperation/competition of many individuals, e.g. for consensus decision making;
- **BETA** = Black Box Explanation through Transparent Approximation, developed by Lakkarju, Bach & Leskovec (2016) it learns two-level decision sets, where each rule explains the model behaviour.
- **Decision Making** = central cognitive process in every medical activity, resulting in the selection of a final choice of action out of several alternatives;
- **Etiology** = in medicine (many) factors coming together to cause an illness (see causality)
- **Explainability** = motivated by the opaqueness of so called “black-box” approaches it is the ability to provide an explanation on why a machine decision has been reached (e.g. why is it a cat what the deep network recognized). Finding an appropriate explanation is difficult, because this needs understanding the context and providing a description of causality and consequences of a given fact. (German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz)

Advance Organizer (2/2)

- **Explanation** = set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questionss (“Why is this a cat?”). (German: Erklärung, Begründung)
- **Explainable AI** = Explainability = fundamental technical topic within AI; answering e.g. **why** a decision has been made (is now necessary for GDPR)
- **European General Data Protection Regulation (EU GDPR)** = Regulation EU 2016/679 – see the EUR-Lex 32016R0679 , will make black-box approaches difficult to use, because they often are not able to explain why a decision has been made (see explainable AI).
- **Empirical evidence** = information acquired by observation or by experimentation in order to verify the truth (accurate to reality) or to falsity (inaccurate to reality).
- **Gradient** = a vector providing the direction of maximum rate of change.
- **Ground truth** = Ground truth is information provided by direct observation (empirical evidence) in contrast to information provided by inference.
- **Interpretability** = a relation between formal theories that expresses the possibility of interpreting or translating one into the other
- **Reasoning** = cognitive (thought) processes involved in making medical decisions (clinical reasoning, medical problem solving, diagnostic reasoning;
- **Transparency** = opposite of opacity of black-box approaches, and connotes the ability to understand how a model works (that does not mean that it should always be understood, but that – in the case of necessity – it can be re-enacted

Learning Goals: At the end of this lecture you …

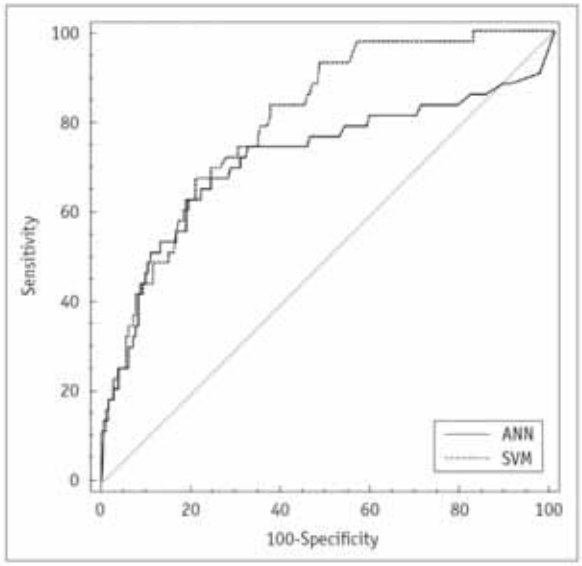
- …have a basic understanding of the difference between correlation and causation ;
- … you have seen some limits of current AI;
- … you know the disadvantages of a “black-box” approach and are aware of the need for explanations;
- … you have a very basic overview on some current methods of explainable AI
- … have seen an example for a specific method of explainable AI;
- … you have seen some future research questions in explainability and interpretability;

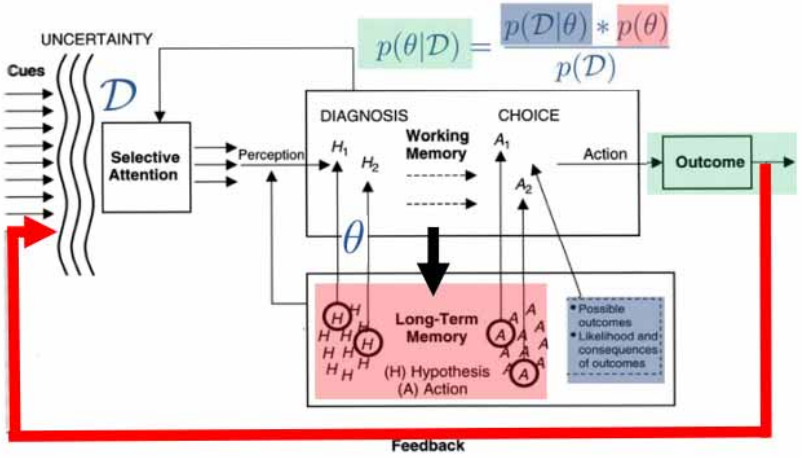
00 Reflection

Five Mainstreams in Machine Learning

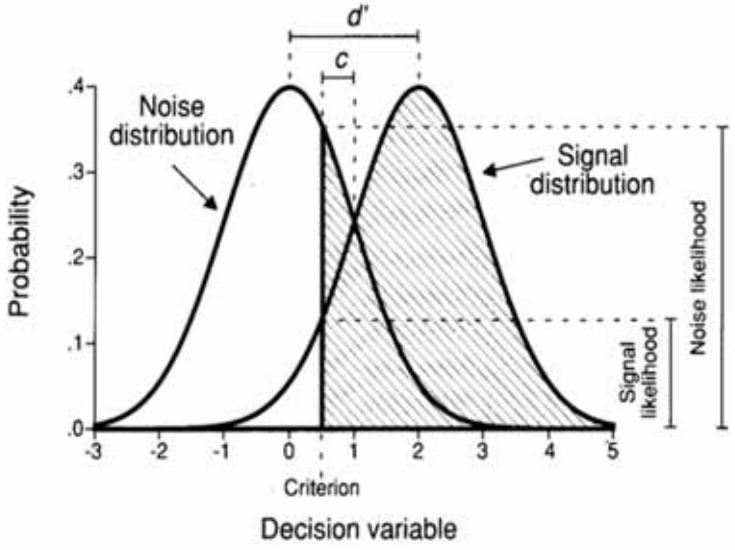
- Symbolic ML
  - First order logic, inverse deduction
  - Tom Mitchell, Steve Muggleton, Ross Quinlan, …
- Bayesian ML
  - Statistical learning
  - Judea Pearl, Michael Jordan, David Heckermann, …
- Cognitive ML
  - Analogisms from Psychology, Kernel machines
  - Vladimir Vapnik, Peter Hart, Douglas Hofstaedter, …
- Connectionist ML
  - Neuroscience, Backpropagation
  - Geoffrey Hinton, Yoshua Bengio, Yann LeCun, …
- Evolutionary ML
  - Nature-inspired concepts, genetic programming
  - John Holland (1929-2015), John Koza, Hod Lipson, …

Reflection from last lectures

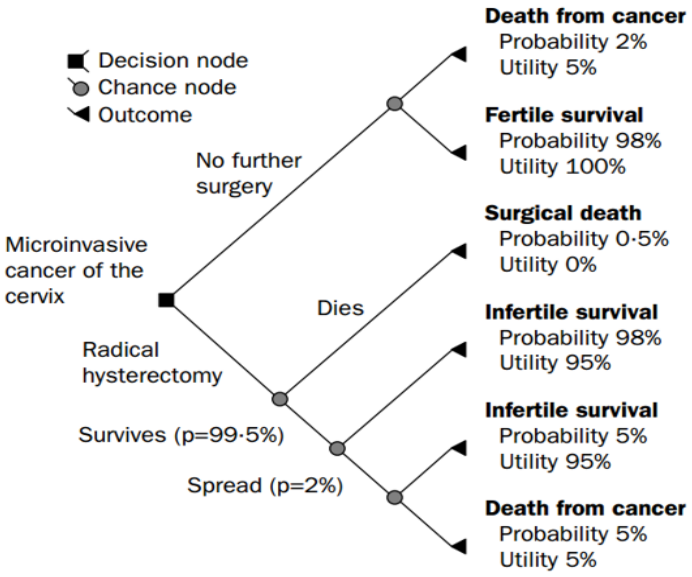




2



3



4

Key Challenges

- Remember: Medicine is an complex application domain – dealing most of the time with **probable information!**
- Some challenges include:
- (a) defining hospital system architectures in terms of generic tasks such as diagnosis, therapy planning and monitoring to be executed for (b) medical reasoning in (a);
- (c) patient information management with (d) minimum uncertainty.
- Other challenges include: (e) knowledge acquisition and encoding, (f) human-ai interface and ai-interaction; and (g) system integration into existing clinical legacy and proprietary environments, e.g. the enterprise hospital information system; to mention only a few.

01 Causality and Decision Making

Causation – beware of counterfactuals!

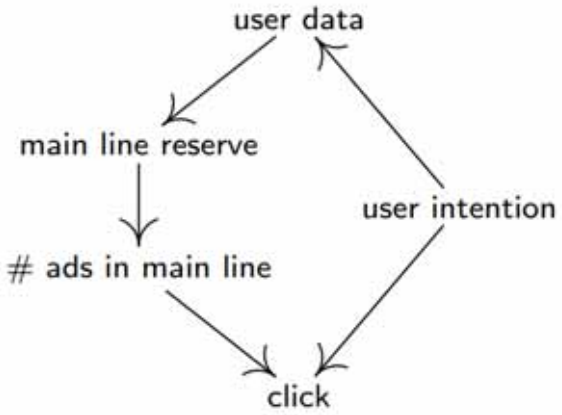
- David Hume (1711-1776): Causation is a matter of perception: observing fire > result feeling heat
- Karl Pearson (1857-1936): Forget Causation, you should be able to calculate correlation
- Judea Pearl (1936-): Be careful with purely empirical observations, instead define causality based on known causal relationships, and **beware of counterfactuals** …

Judea Pearl 2009. Causal inference in statistics: An overview. Statistics surveys, 3, 96-146

Judea Pearl, Madelyn Glymour & Nicholas P. Jewell 2016. Causal inference in statistics: A primer, John Wiley & Sons.

What is a counterfactual?

- Hume again: *"… if the first object had not been, the second never had existed …"*
- Causal inference as a missing data problem
- $x_i := f_i(\text{ParentsOf}_i, \text{Noise}_i)$
- Interventions can only take place on the right side



Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard & Ed Snelson 2013. Counterfactual reasoning and learning systems: The example of computational advertising. The Journal of Machine Learning Research, 14, (1), 3207-3260.

Remember: Correlation is NOT Causality

Dependence vs. Causation

Storks Deliver Babies (*p*= 0.008)

Robert Matthews

Article first published online: 25 DEC 2001

DOI: 10.1111/1467-9639.00013

Teaching Statistics Trust, 2000

Issue

Teaching Statistics

Volume 22, Issue 2

38, June 2000

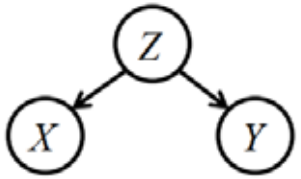
Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	mailto:rajm@compuserve.com	
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.

Correlation does not tell anything about causality!

- Hans Reichenbach (1891-1953): **Common Cause Principle**
- This principle links causality with probability:
  - If X and Y are statistically dependent, there is a Z influencing both
  - whereas:
  - A, B, … events
  - X, Y, Z random variables
  - P … probability measure
  - Px … probability distribution of X
  - p … probability density
  - p(X) .. Density of Px
  - p(x) probability density of Px evaluated at the point x



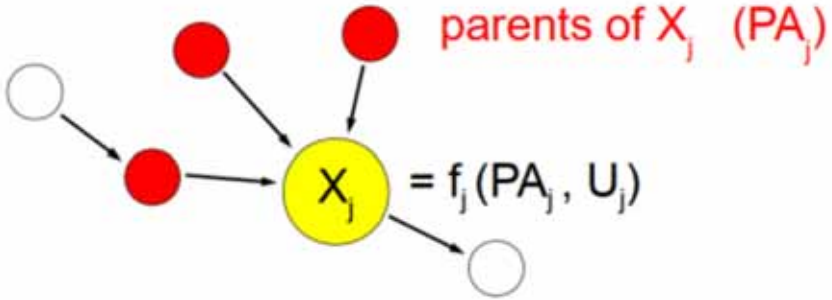
Hans Reichenbach 1956. The direction of time (Edited by Maria Reichenbach), Mineola, New York, Dover.

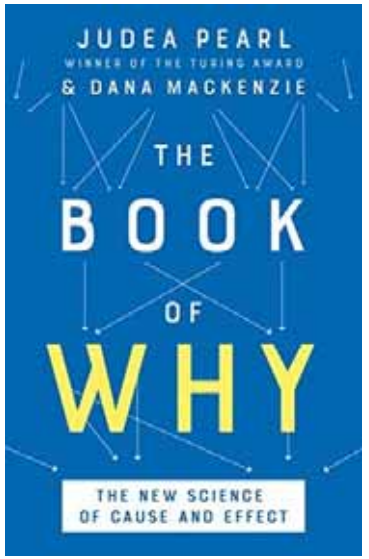
https://plato.stanford.edu/entries/physics-Rpcc/

For details please refer to the excellent book of: Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA). https://mitpress.mit.edu/books/elements-causal-inference

Functional Causal Model

- $X_1, \ldots, X_n$ … set of observables
- Draw a directed acyclic graph G with nodes $X_1, \ldots, X_n$





- Parents = direct causes
- $x_i := f_i(\text{ParentsOf}_i, \text{Noise}_i)$

Remember: Noise means “unexplained (exogenous) data” and is denoted as U_i

Question: Can we recover G from p ?
Answer: under certain assumptions, we can recover an equivalence class containing the correct *G* using conditional independence testing (but there are other problems as well)

Counterfactual Learning

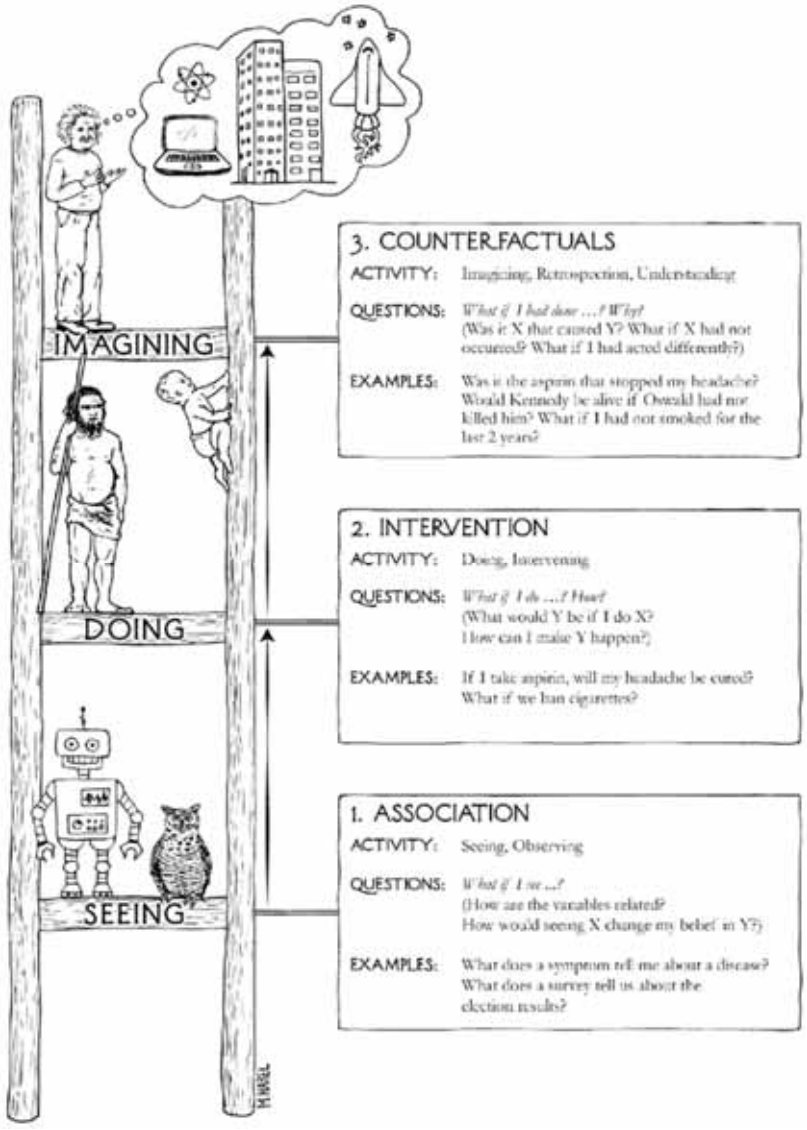


Figure 1.2 from Judea Pearl & Dana Mackenzie 2018. The book of why, New York, Basic Books,
Source: Illustrator: Maayan Harel, http://www.maayanillustration.com

3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done ...? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do ...? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

- **The current data-driven machine learning approach of artificial intelligence misses an essential element of human intelligence:**
- **AI cannot reason why!**

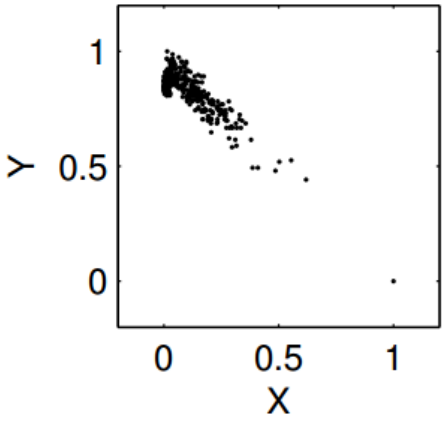
02 Causal Reasoning

- **"How do humans generalize from few examples?"**
  - Learning relevant representations
  - Disentangling the explanatory factors
  - Finding the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

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

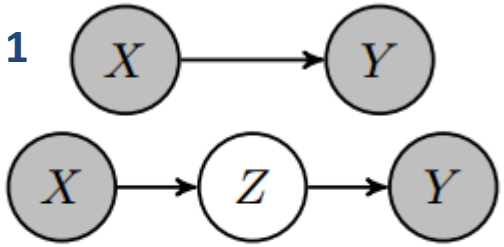
Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

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



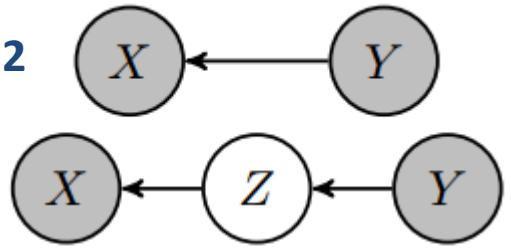
Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.

1



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

2



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

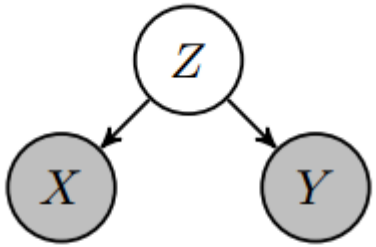
3

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

4

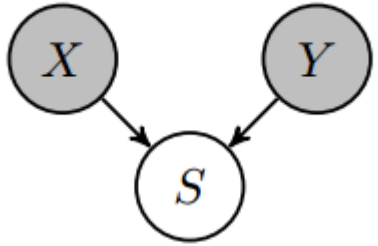
$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

5



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

6



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

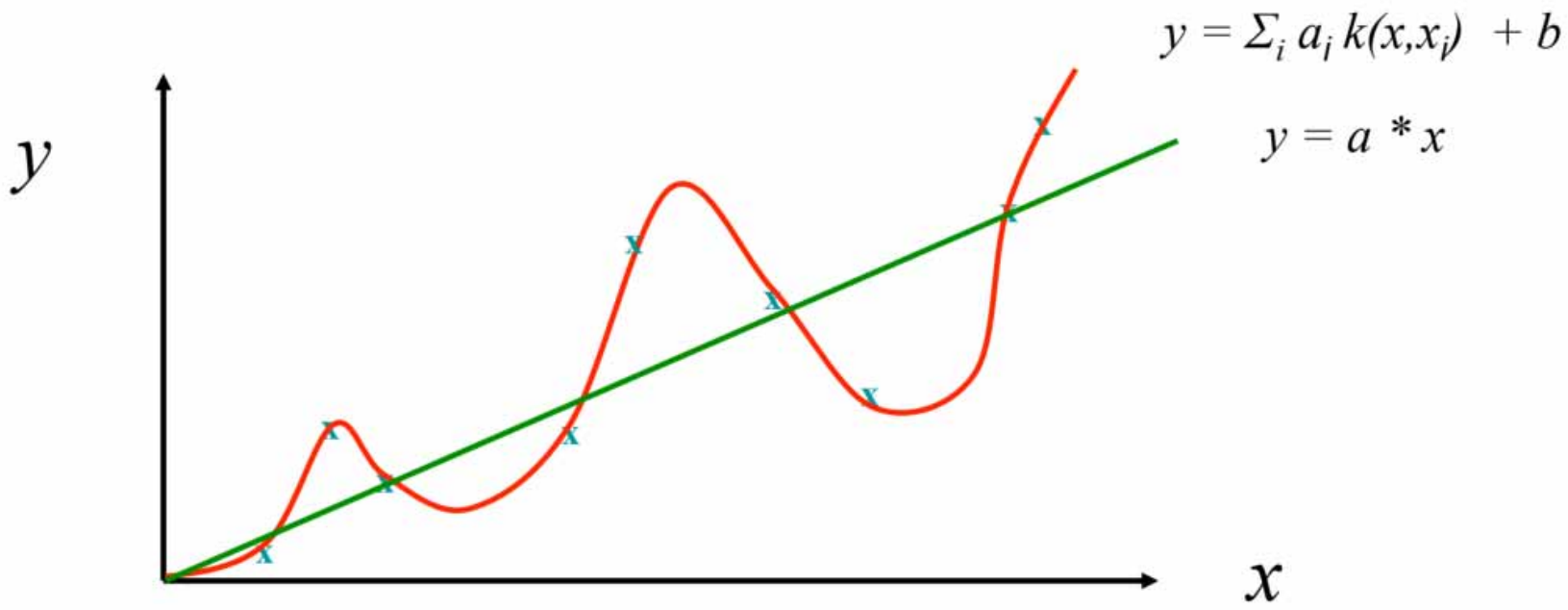
Remember: Reasoning = “Sensemaking”

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, therefore A=C
- **Inductive reasoning** = makes broad generalizations from specific observations
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions
- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.

Important Definition: 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 a conclusion about a causal connection based on the conditions of the occurrence of an effect.
    - Causal inference is an example of causal reasoning.

Empirical Inference Example



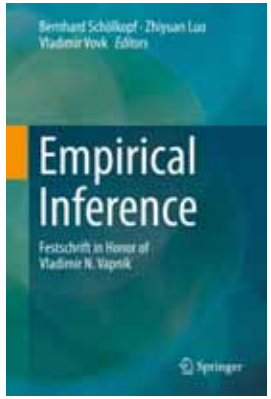
Gottfried W. Leibniz (1646-1716)

Hermann Weyl (1885-1955)

Vladimir Vapnik (1936-)

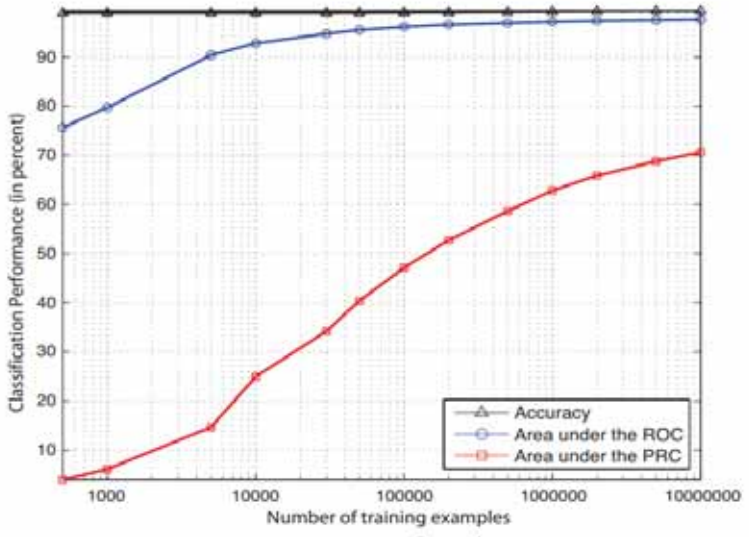
Alexey Chervonenkis (1938-2014)

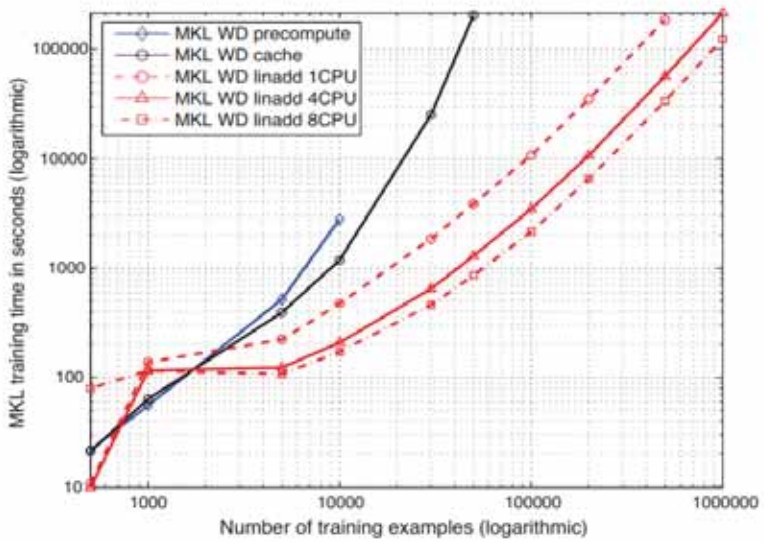
Gregory Chaitin (1947-)



Remember: hard inference problems

- High dimensionality (curse of dim., many factors contribute)
- Complexity (real-world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - *) = Def.: a sequence or collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent





Sören Sonnenburg, Gunnar Rätsch, Christin Schaefer & Bernhard Schölkopf 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

What makes it hard … ?

Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA).

Example 3.4 (Eye disease) There exists a rather effective treatment for an eye disease. For 99% of all patients, the treatment works and the patient gets cured ($B = 0$); if untreated, these patients turn blind within a day ($B = 1$). For the remaining 1%, the treatment has the opposite effect and they turn blind ($B = 1$) within a day. If untreated, they regain normal vision ($B = 0$).

Which category a patient belongs to is controlled by a rare condition ($N_B = 1$) that is unknown to the doctor, whose decision whether to administer the treatment ($T = 1$) is thus independent of N_B. We write it as a noise variable N_T.

Assume the underlying SCM

$$\mathfrak{C}: \begin{array}{lcl} T & := & N_T \\ B & := & T \cdot N_B + (1-T)\cdot(1-N_B) \end{array}$$

with Bernoulli distributed $N_B \sim \mathrm{Ber}(0.01)$; note that the corresponding causal graph is $T \to B$.

Now imagine a specific patient with poor eyesight comes to the hospital and goes blind ($B = 1$) after the doctor administers the treatment ($T = 1$). We can now ask the counterfactual question *"What would have happened had the doctor administered treatment $T = 0$?"* Surprisingly, this can be answered. The observation $B = T = 1$ implies with (3.5) that for the given patient, we had $N_B = 1$. This, in turn, lets us calculate the effect of $do\,(T := 0)$.

To this end, we first condition on our observation to update the distribution over the noise variables. As we have seen, conditioned on $B = T = 1$, the distribution for N_B and the one for N_T collapses to a point mass on 1, that is, δ_1. This leads to a modified SCM:

$$\mathfrak{C}|B=1,T=1: \begin{array}{lcl} T & := & 1 \\ B & := & T\cdot 1+(1-T)\cdot(1-1)=T \end{array} \tag{3.6}$$

Note that we only update the noise distributions; conditioning does not change the structure of the assignments themselves. The idea is that the physical mechanisms are unchanged (in our case, what leads to a cure and what leads to blindness), but we have gleaned knowledge about the previously unknown noise variables *for the given patient*.

Next, we calculate the effect of $do\,(T=0)$ for this patient:

$$\mathfrak{C}|B=1,T=1;do\,(T:=0): \begin{array}{lcl} T & := & 0 \\ B & := & T \end{array} \tag{3.7}$$

Clearly, the entailed distribution puts all mass on $(0,0)$, and hence

$$P^{\mathfrak{C}|B=1,T=1;do(T:=0)}(B=0)=1\,.$$

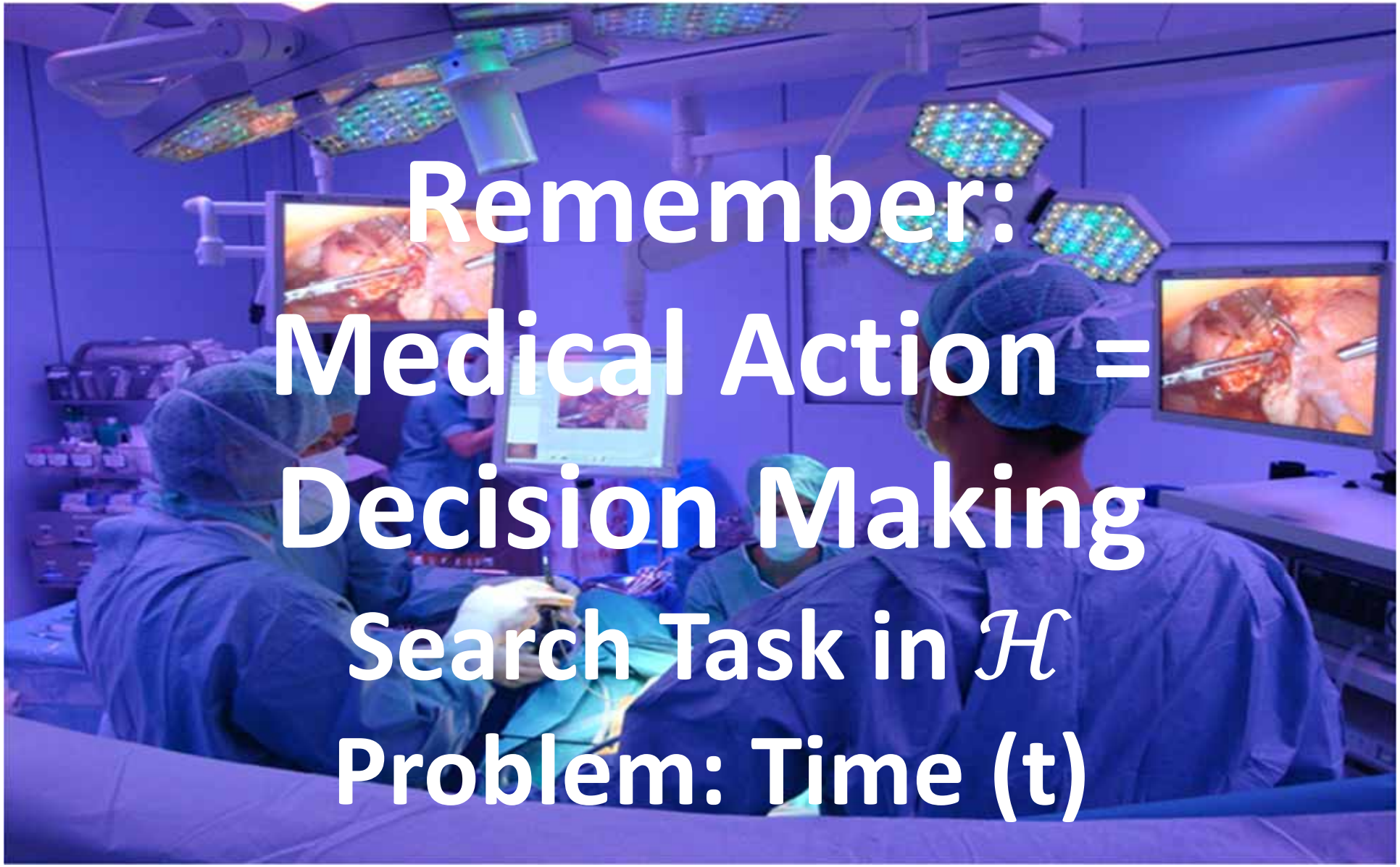
This means that the patient would thus have been cured ($B=0$) if the doctor had not given him treatment, in other words, $do\,(T:=0)$. Because of

$$P^{\mathfrak{C};do(T:=1)}(B=0)=0.99 \quad \text{and}$$
$$P^{\mathfrak{C};do(T:=0)}(B=0)=0.01,$$

however, we can still argue that the doctor acted optimally (according to the available knowledge). □

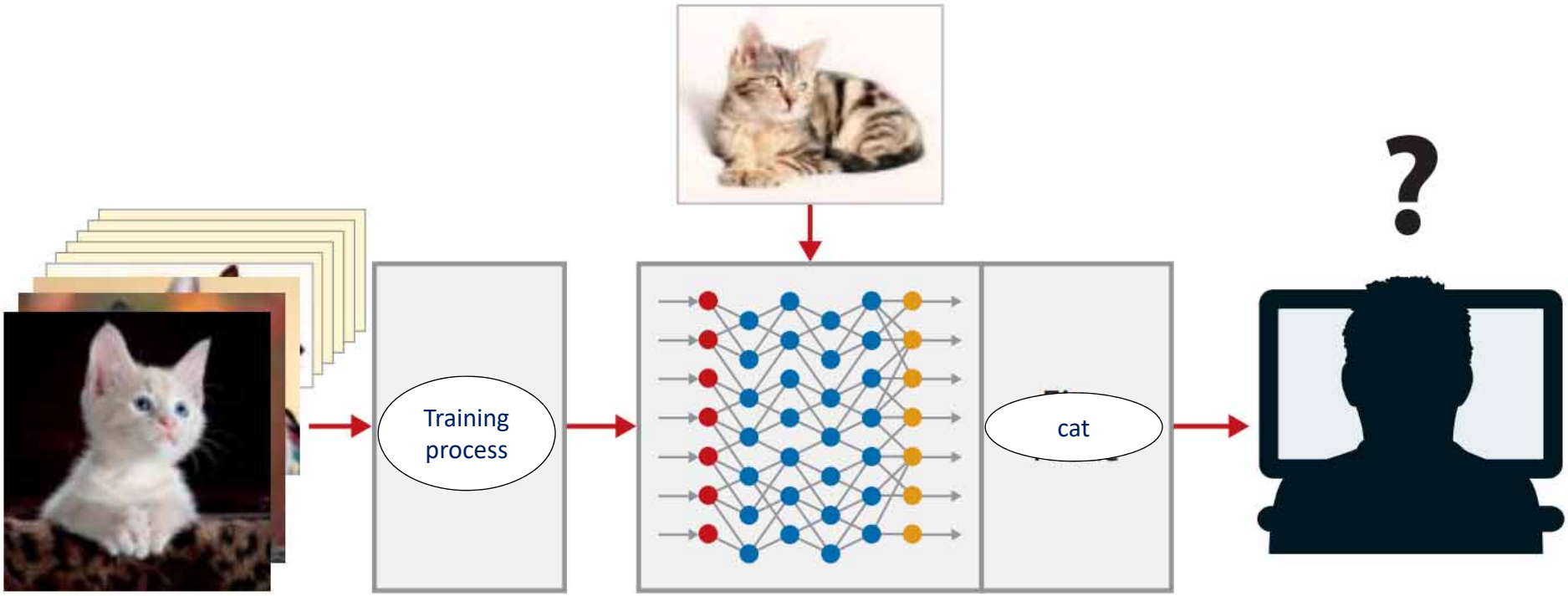
Interestingly, Example 3.4 shows that we can use counterfactual statements to falsify the underlying causal model (see Section 6.8). Imagine that the rare condition N_B can be tested, but the test results take longer than a day. In this case, it is possible that we observe a counterfactual statement that contradicts the measurement result for N_B. The same argument is given by Pearl [2009, p.220, point (2)]. Since the scientific content of counterfactuals has been debated extensively, it should be emphasized that the counterfactual statement here is falsifiable because the noise variable is not unobservable in principle but only at the moment when the decision of the doctor has to be made.

Judea Pearl 2009. *Causality: Models, Reasoning, and Inference (2nd Edition)*, Cambridge, Cambridge University Press.


Remember:
Medical Action =
Decision Making
Search Task in $\mathcal{H}$
Problem: Time (t)

03 Why Interpretability ?

Current state-of-the-art: the “why” is missing!

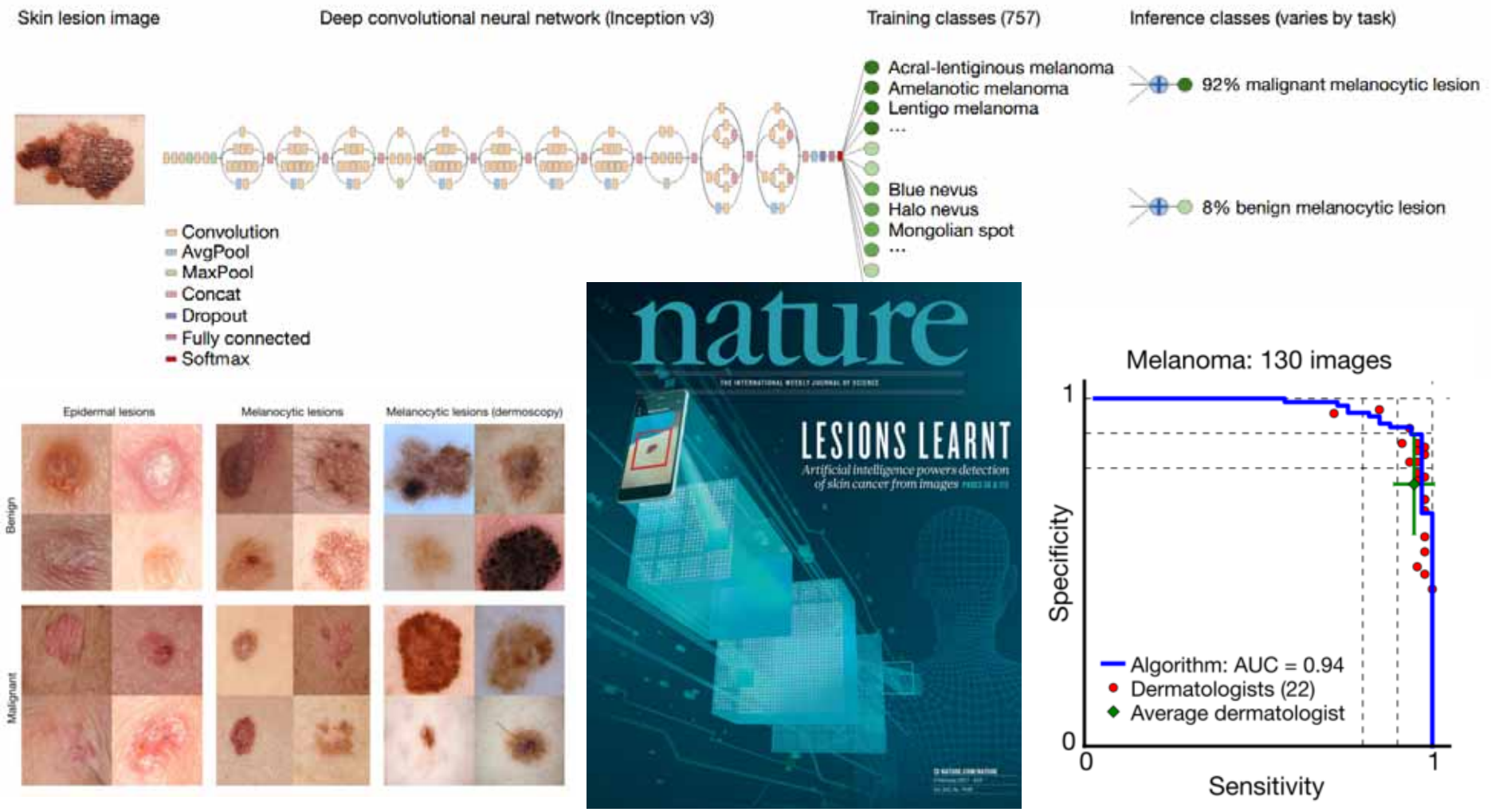


Andreas Holzinger 2018. Interpretierbare KI: Neue Methoden zeigen Entscheidungswege künstlicher Intelligenz auf. c't Magazin für Computertechnik, 22, 136-141.

Example: 1,28 million images …

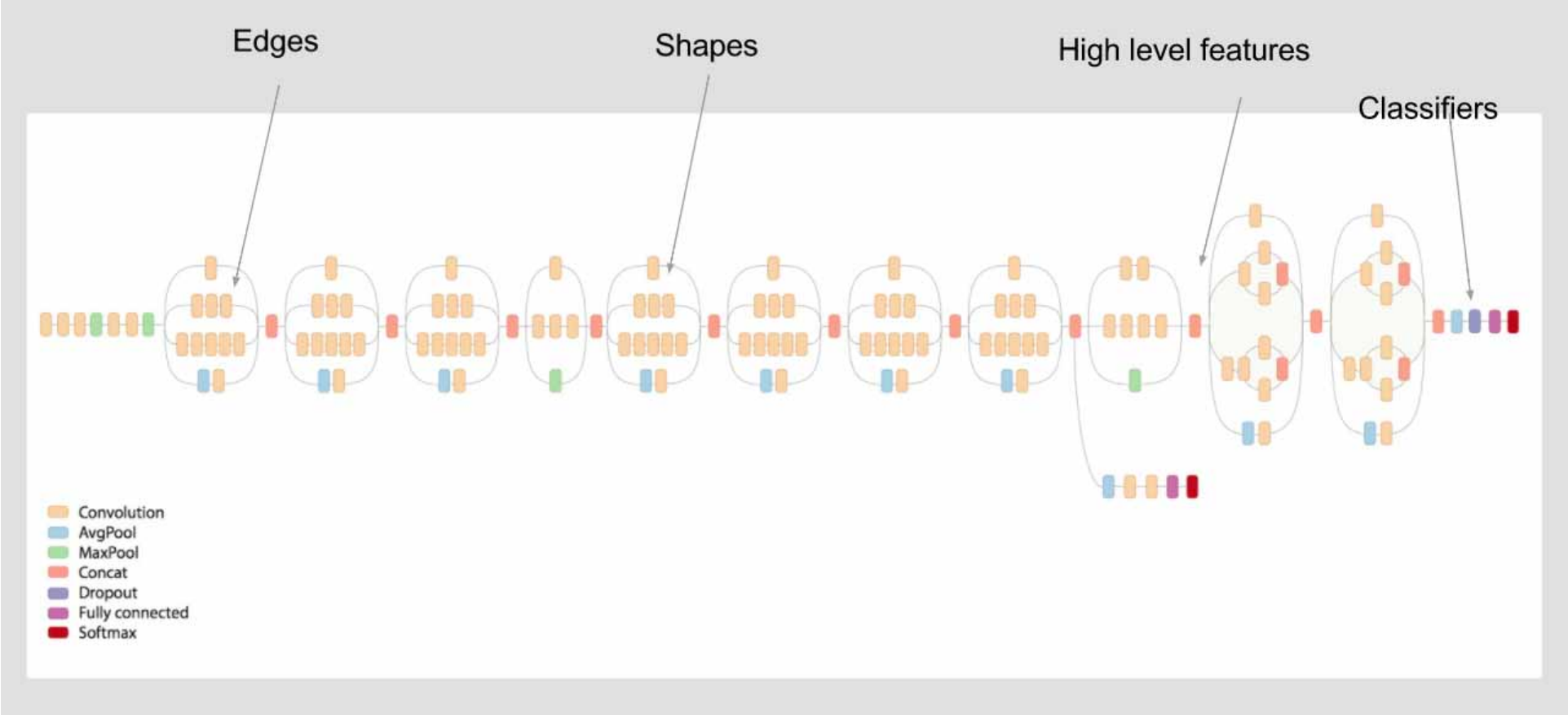
LETTER

Dermatologist-level classification of skin cancer with deep neural networks



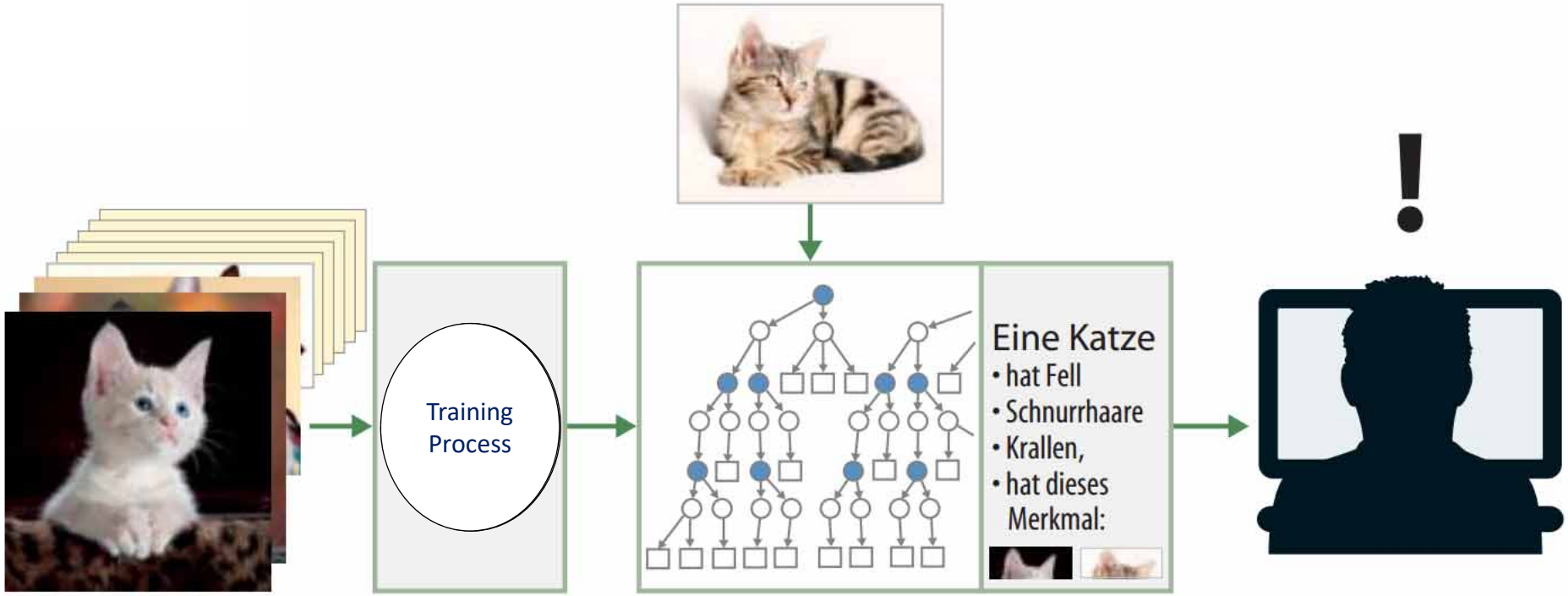
Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

Inside Inception v3: what do the layers learn?



Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens & Zbigniew Wojna. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016. 2818-2826.

Our goal: Explainability and Causability



Andreas Holzinger 2018. Interpretierbare KI: Neue Methoden zeigen Entscheidungswege künstlicher Intelligenz auf. c't Magazin für Computertechnik, 22, 136-141.

04 Methods of Explainable AI

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

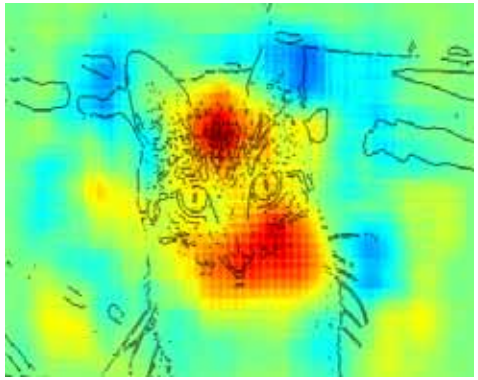
Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course at Graz University of Technology https://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

Example: Explanation by Decomposition (general idea)

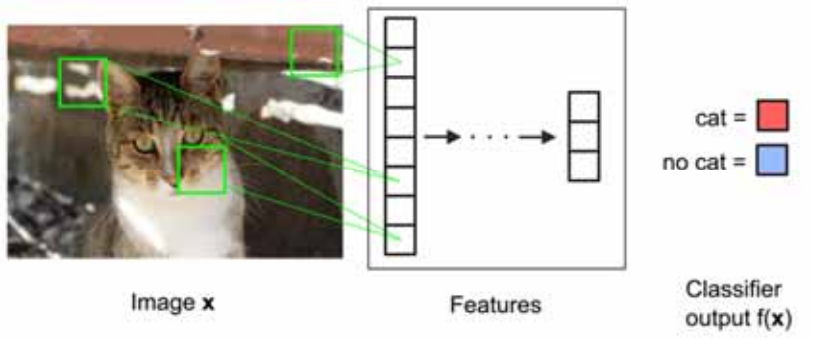
- Given: a prediction $f(x)$ over an input set $x = (x_1, \ldots, x_d)$
- Goal: Computing a relevance score $r_d(x)$ for each input x_d in dimension d

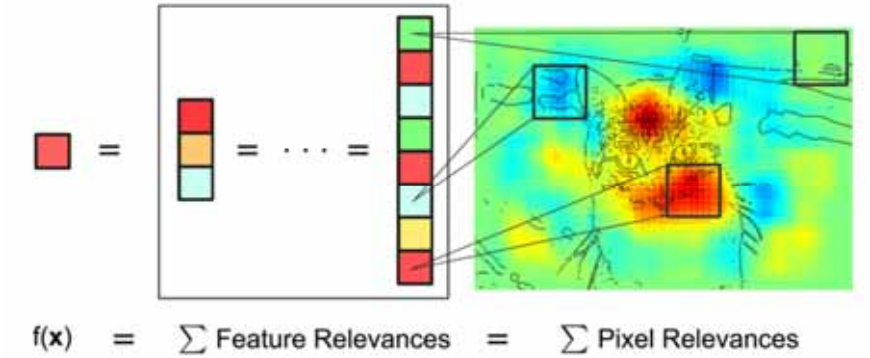
$$f(x) = \sum_{d=1}^{D} r_d(x)$$

- Decompose the prediction depending on the test data
- $r_d(x)$ = ?
- Looking for a linear mapping which can be a meaningful explanation for a human expert



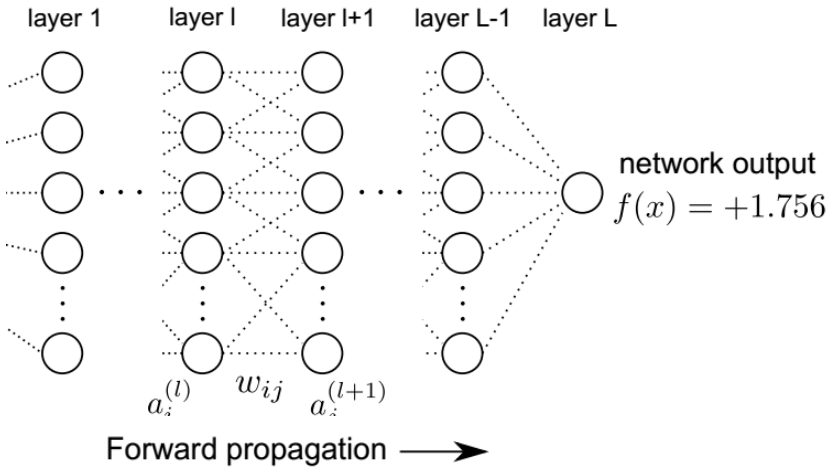
Example: LRP Layer-Wise Relevance Propagation

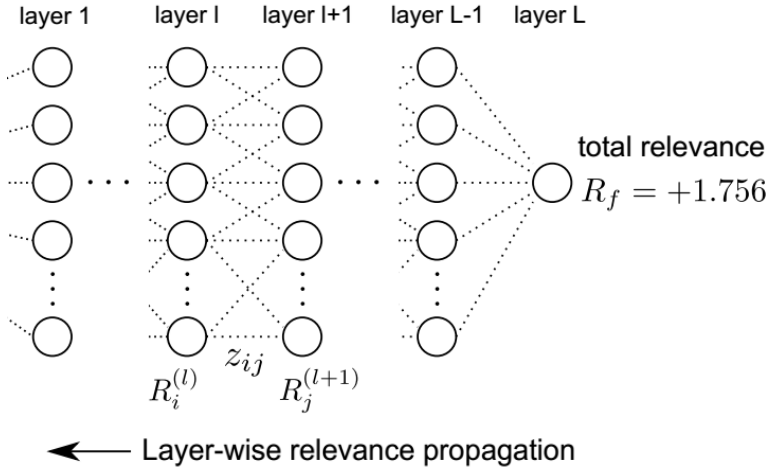




$$a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)$$

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

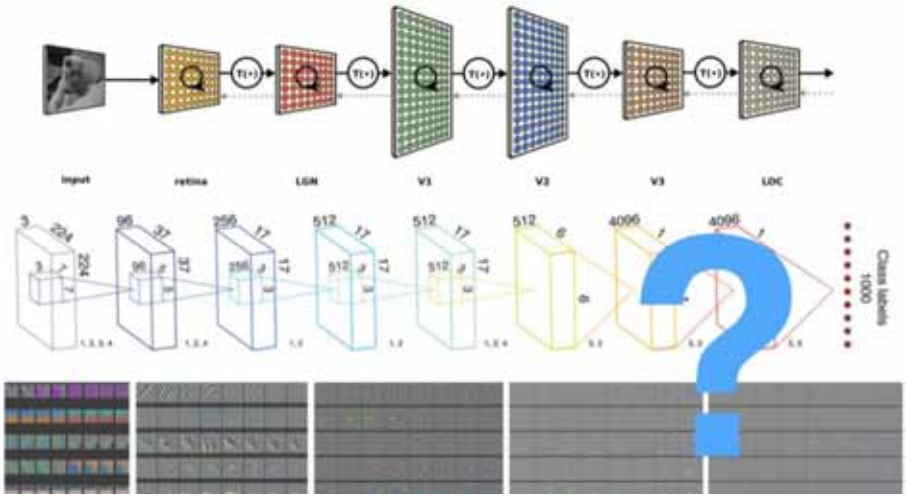




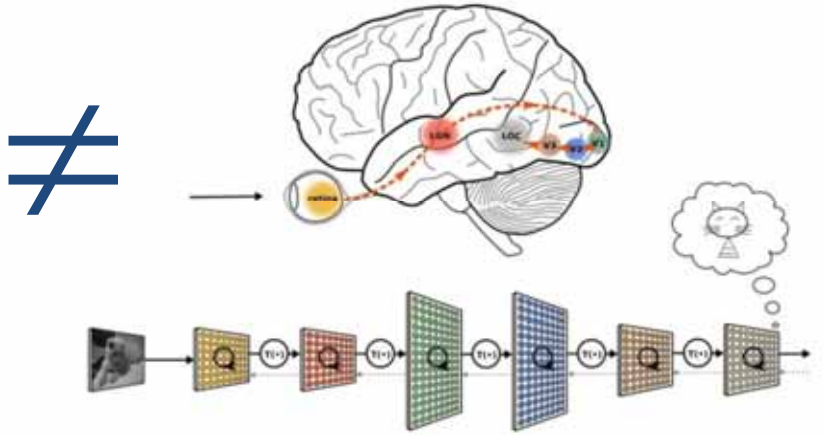
$$R_i = \left|\left|\frac{\partial}{\partial x_i} f(\mathbf{x})\right|\right| \qquad \sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x})$$

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.



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

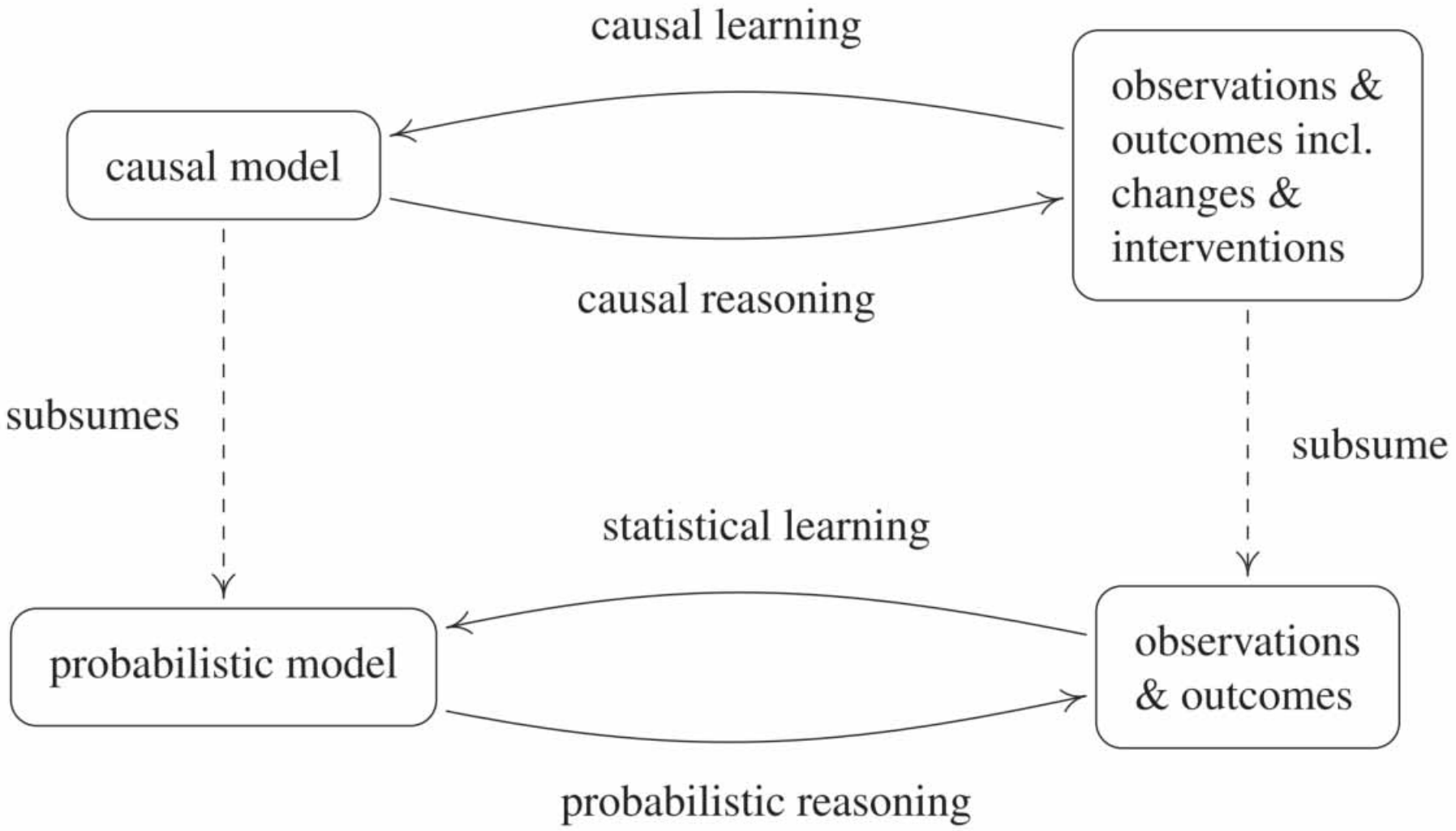
Humans work in another vector space which is spanned by **implicit knowledge** vectors corresponding to an unknown set of human interpretable concepts.

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

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.

05 Interpretability: Mapping AI with Human Intelligence

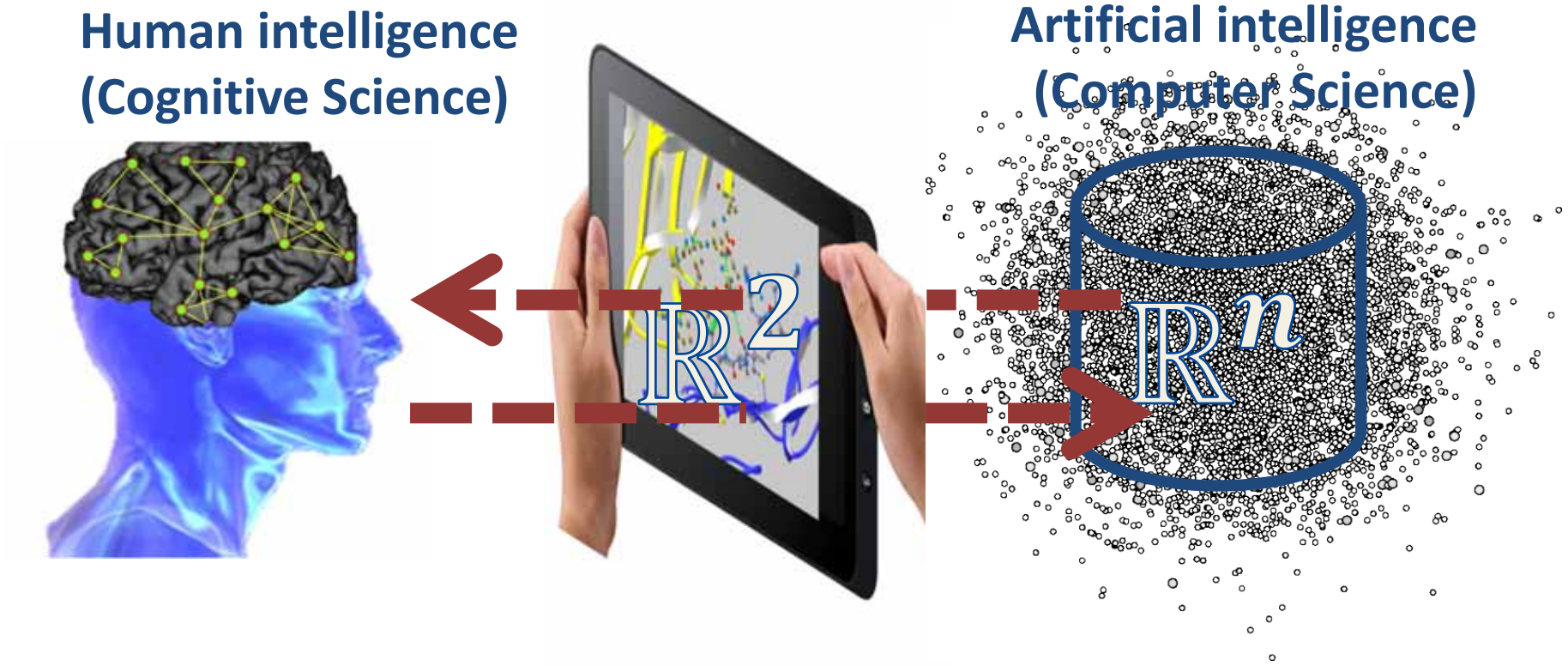
Mapping human intelligence with artificial intelligence



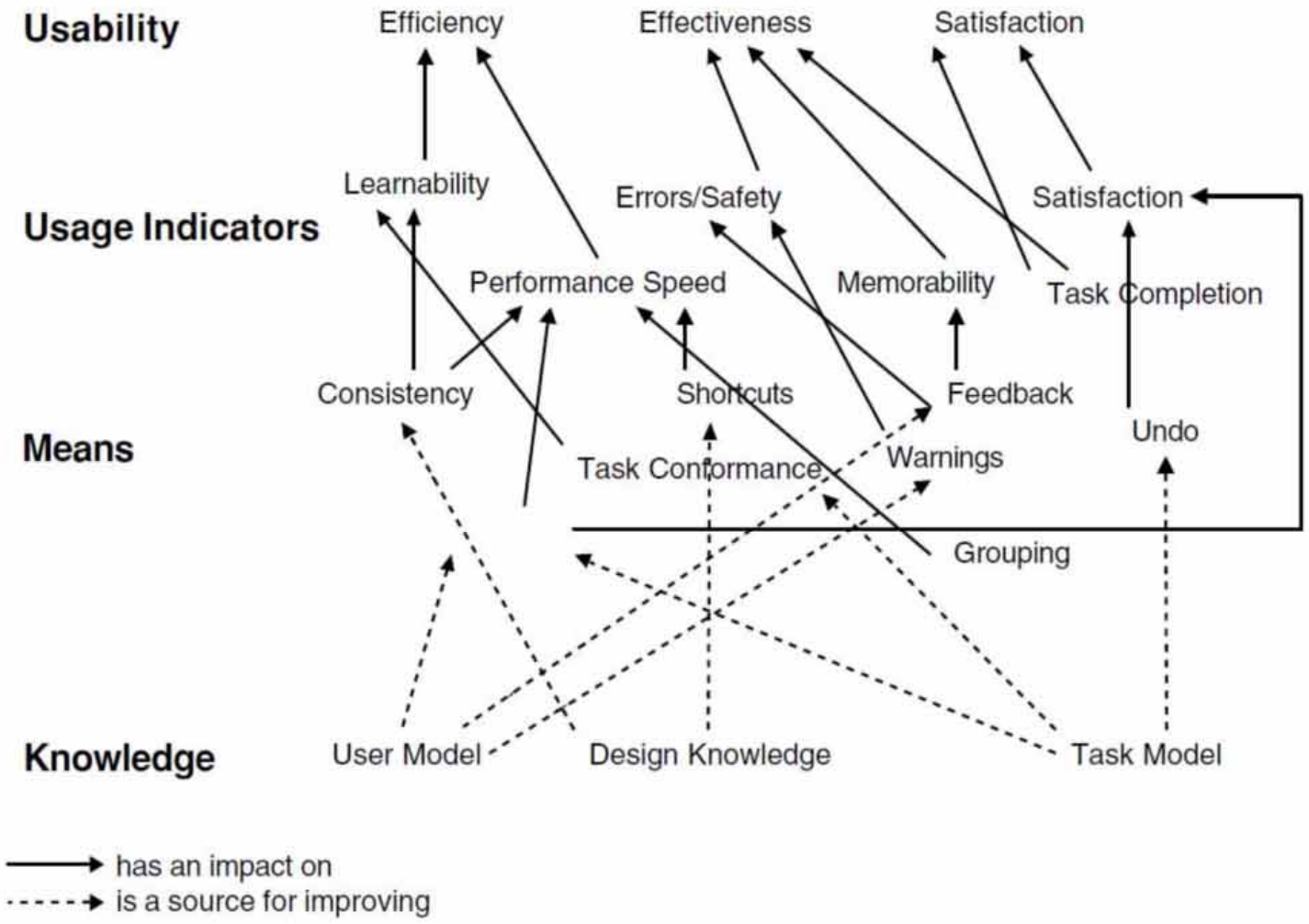
Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA).

Efficient Human-AI interaction needs a “ground truth”

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



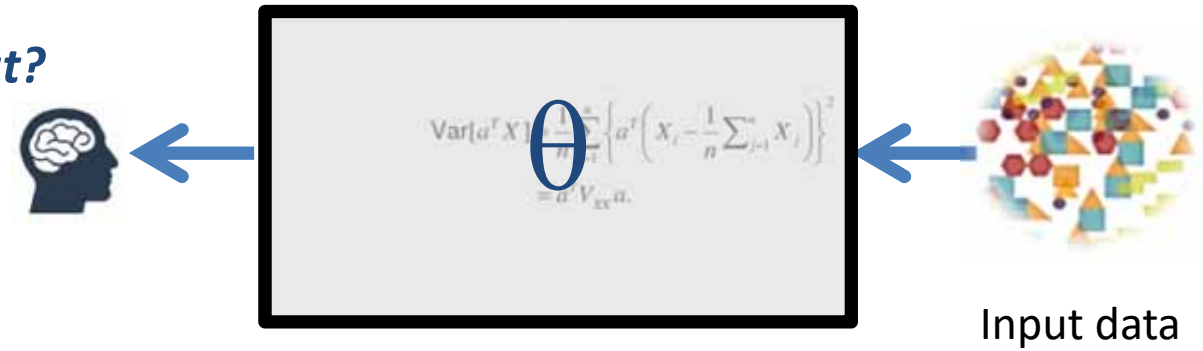
Compare this with usability



Veer, G. C. v. d. & Welie, M. v. (2004) DUTCH: Designing for Users and Tasks from Concepts to Handles. In: Diaper, D. & Stanton, N. (Eds.) *The Handbook of Task Analysis for Human-Computer Interaction. Mahwah (New Jersey), Lawrence Erlbaum, 155-173.*

We need effective tools for Human-AI Interaction

Why did the algorithm do that?
Can I trust these results?
How can I correct an error?

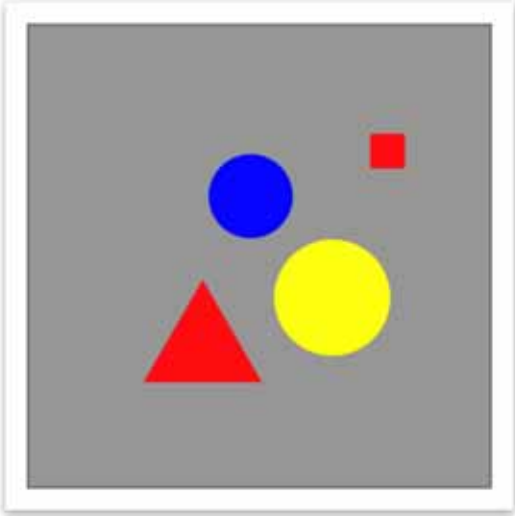


A possible solution

The domain expert can understand why …
The domain expert can learn and correct errors …
The domain expert can re-enact on demand …

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.

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 within this square.
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable by a human observer.

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

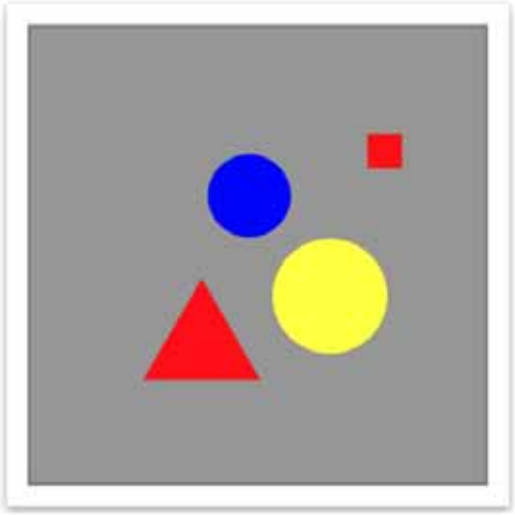
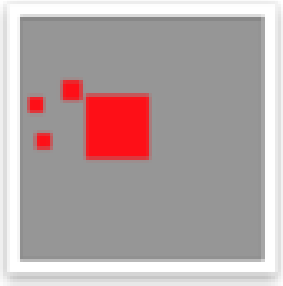
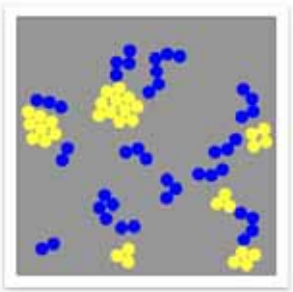
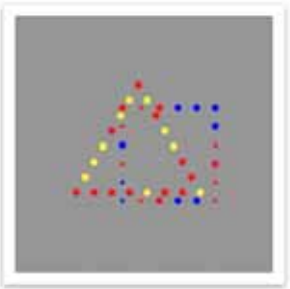
- Remark: The evaluation of a natural language statement is always done in a specific context. In the followings examples we use **well known concepts from human perception** and linguistic theory.
- If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.

Definition 3 A Kandinsky Pattern K …

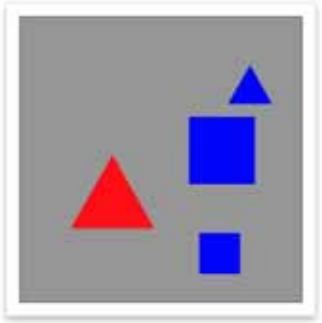
- … is defined as the subset of all possible Kandinsky Figures k with $s(k) \rightarrow 1$ or the natural language statement is true.
- $s(k)$ and a natural language statement are equivalent, if and only if the resulting Kandinsky Patterns contains the same Kandinsky Figures.
- $s(k)$ and the natural language statement are defined as the **Ground Truth** of a Kandinsky Pattern

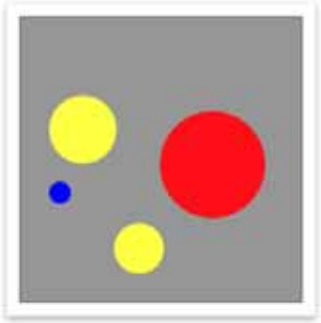
"… the Kandinsky Figure has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share a object …".

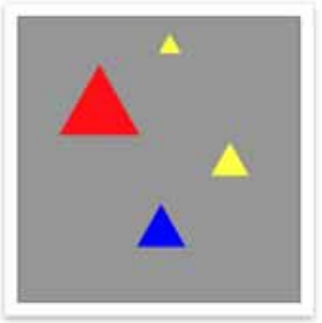
Take part in our Kandinsky Challenge

- https://github.com/human-centered-ai-lab/dat-kandinsky-patterns
- https://human-centered.ai/project/kandinsky-patterns

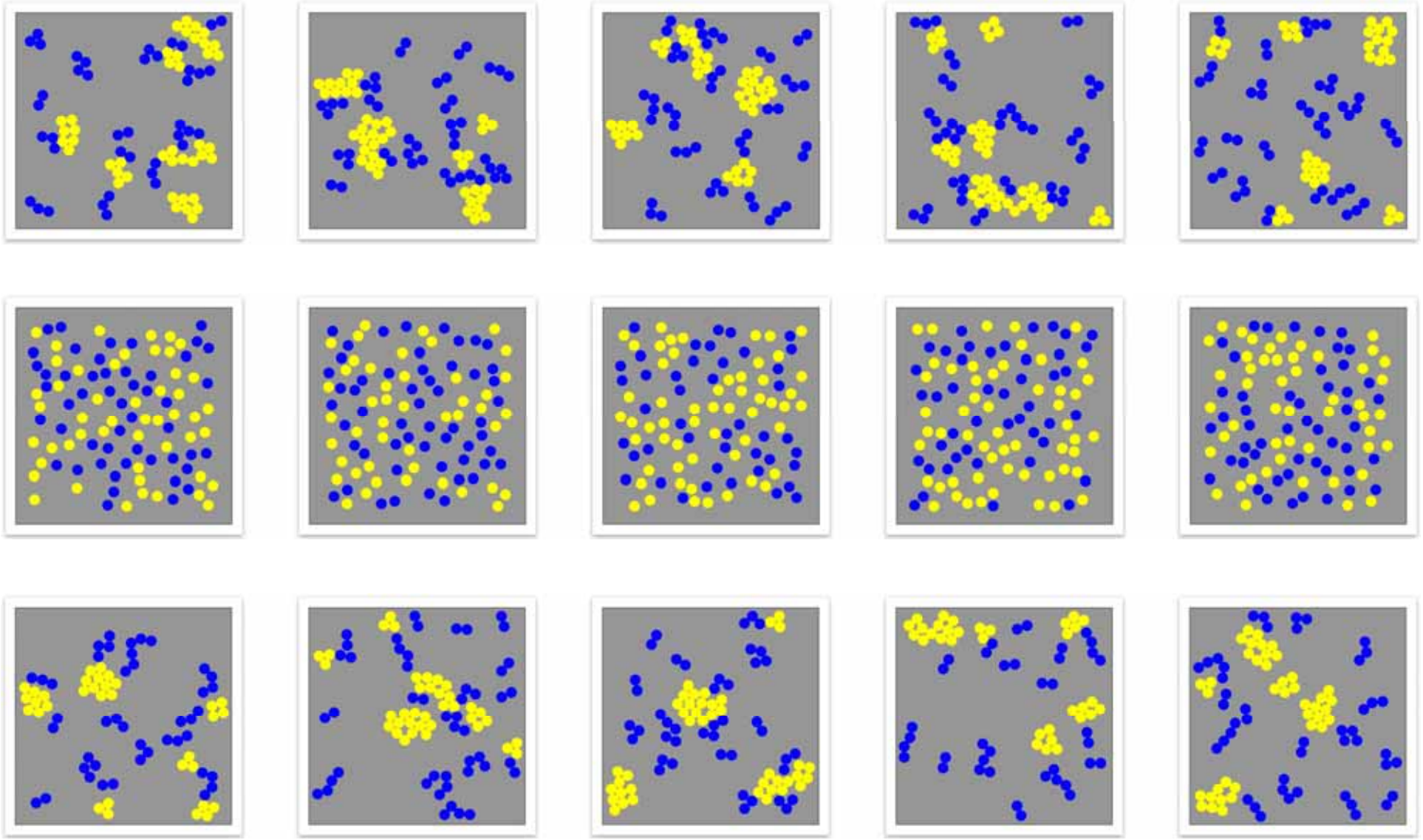




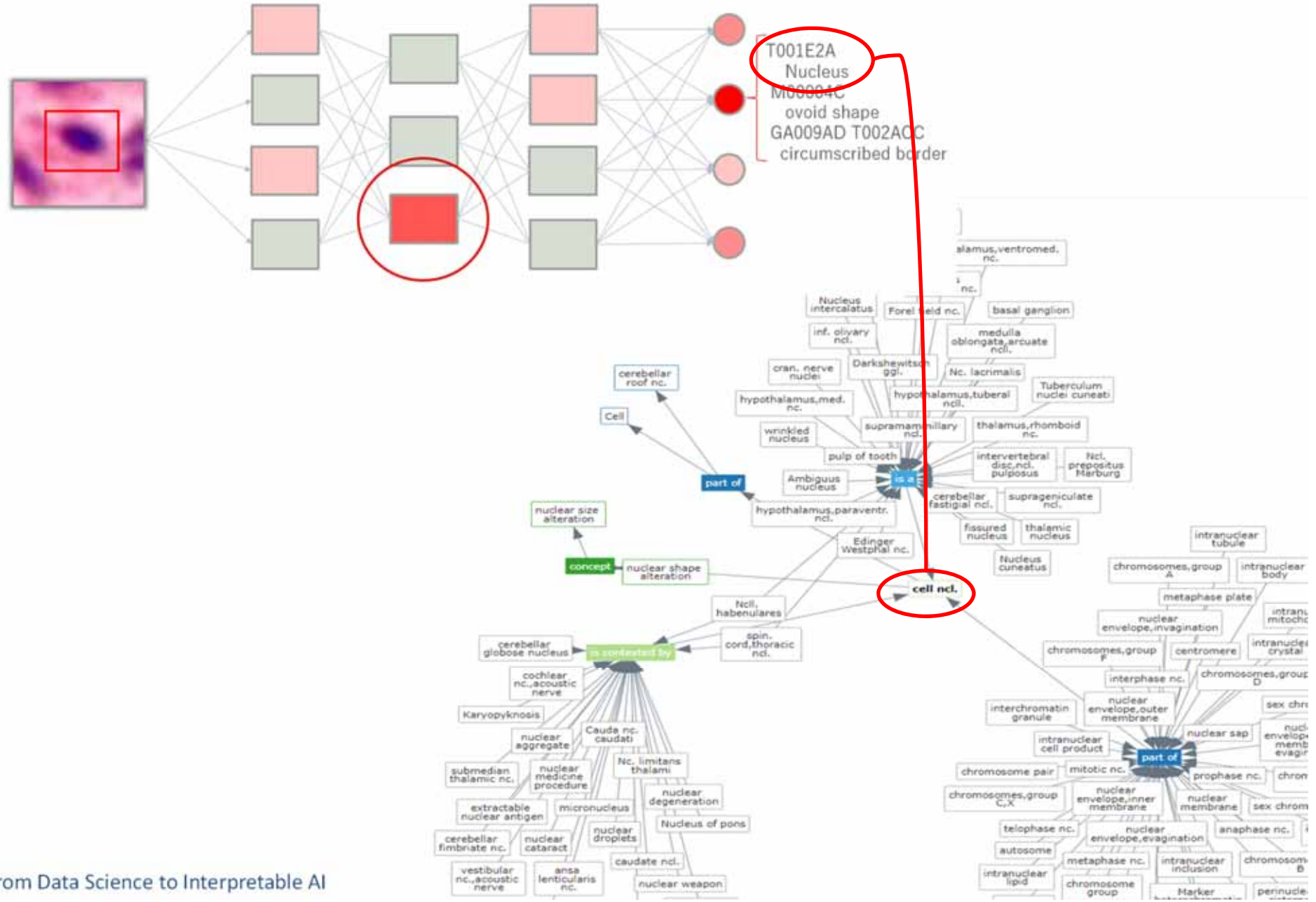


Heimo Müller & Andreas Holzinger 2019. Kandinsky Patterns. arXiv:1906.00657

Example: Patho Challenge



Practice: Combination of statistical learning with ontologies



Conclusion:
Human-in-control

This is compatible to interactive machine learning

- Computational approaches can find in R^n what no human is able to see
- However, still there are many hard problems where a human expert in R^2 can understand the **context** and bring in experience, expertise, knowledge, intuition, …
- Black box approaches can not explain **WHY** a decision has been made …

The fist wave of AI (1943-1975): Handcrafted Knowledge

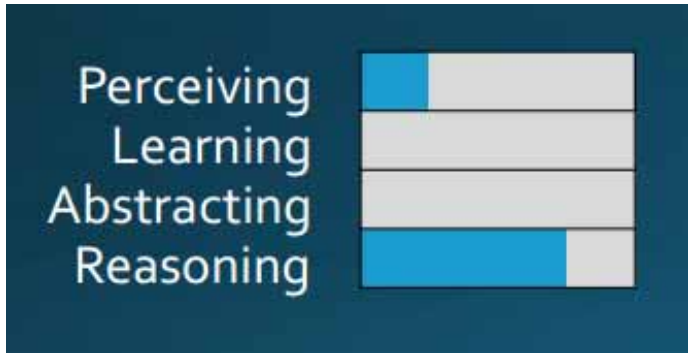


Image credit to John Launchbury

- Engineers create a set of logical rules to represent knowledge (Rule based Expert Systems)
- Advantage: works well in narrowly defined problems of well-defined domains
- Disadvantage: No adaptive learning behaviour and poor handling of $p(x)$

The second wave of AI (1975 –): Statistical Learning

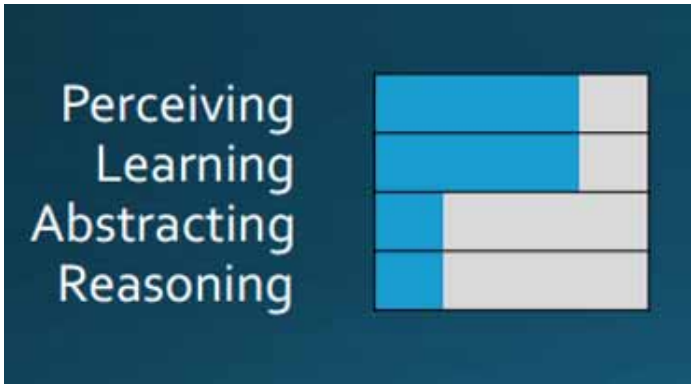


Image credit to John Launchbury

- Engineers create learning models for specific tasks and train them with “big data” (e.g. Deep Learning)
- Advantage: works well for standard classification tasks and has prediction capabilities
- Disadvantage: No contextual capabilities and minimal reasoning abilities

The third wave of AI (?): Adaptive Context Understanding

Image credit to John Launchbury

- A contextual model can perceive, learn and understand and abstract and reason
- Advantage: can use transfer learning for adaptation on unknown unknowns
- Disadvantage: Superintelligence …

Three (selected) dangers and myths about AI generally

- Myth 1a: Superintelligence by 2100 is inevitable!
- Myth 1b: Superintelligence by 2100 is impossible!
- **Fact: We simply don't know it!**

- Myth 2: Robots are our main concern

  Fact: Cyberthreats are the main concern: it needs no body – only an Internet connection

- Myth 3: AI can never control us humans

  Fact: Intelligence is an enabler for control: We control tigers by being smarter …

Human-Centered AI (HCAI) ensures Human-in-control

Thank you!

Questions

Reflection from last lecture

- The Quiz-Slide will be shown during the course