

Mini Course

# Fundamentals of Medical AI

Part 04

Causal Reasoning and Interpretable AI

## From Decision Support Systems to Causability (or an ultrashort history of artificial intelligence)

**Andreas Holzinger**

Human-Centered AI Lab (Holzinger Group)

Institute for Medical Informatics/Statistics, Medical University Graz, Austria  
and

Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada



## Primer on Probability & Information

### Part 1 Theory

01 Introduction to Medical AI and Machine Learning for Health

02 Data, Information and Knowledge

03 Human Decision Making and AI Decision Support

04 Causal Reasoning and Interpretable AI

### Part 2 Practice

05 Methods of Explainable AI

06 Social, Ethical and Legal Aspects of Medical AI

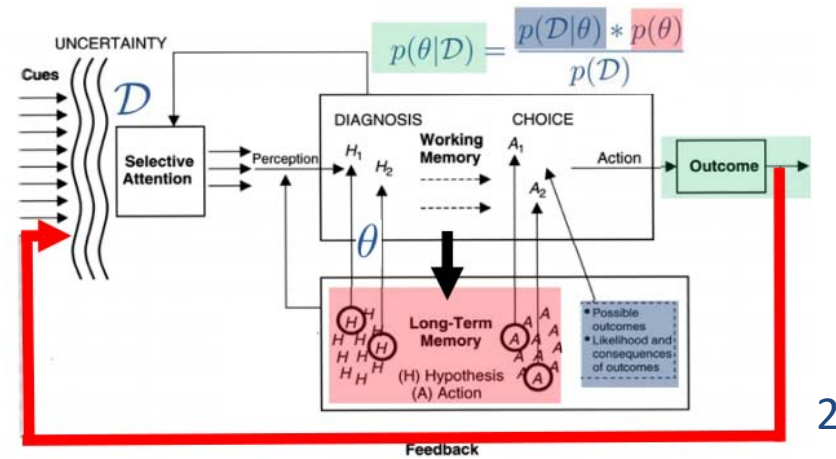
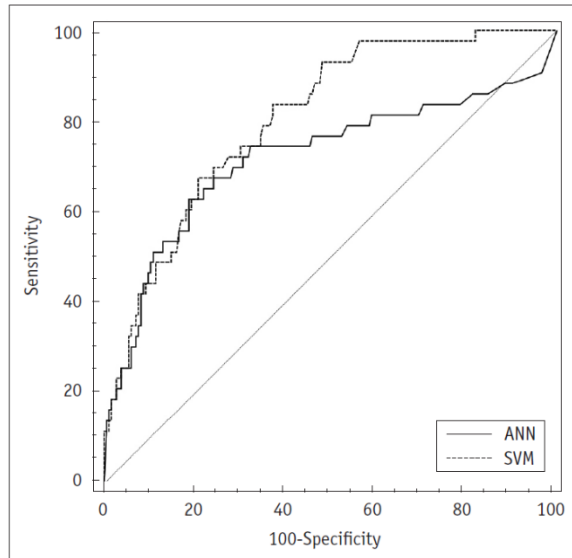
07 Project: Bringing AI into medical workflows

08 Presentation of the developed concepts

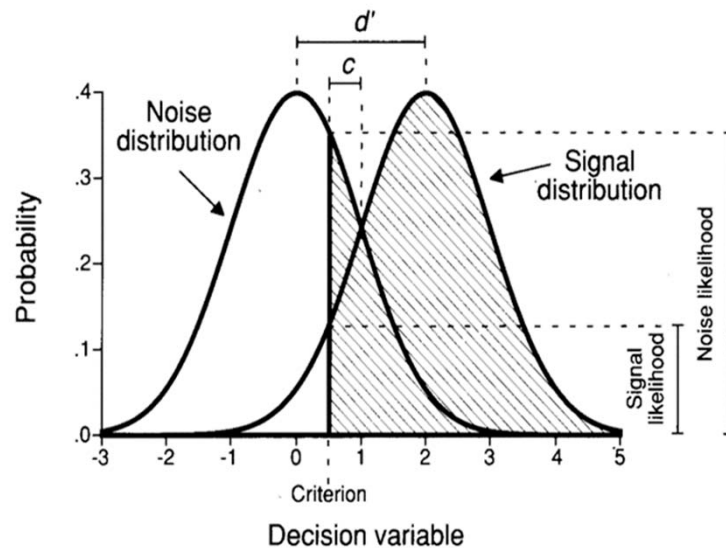
Written Exam

- **00 Reflection – follow-up from last lecture**
- **01 History of DSS = History of AI**
- **02 Causality and Decision Making**
- **03 Medical Communication**
- **04 Causal Reasoning**
- **05 Interpretability**

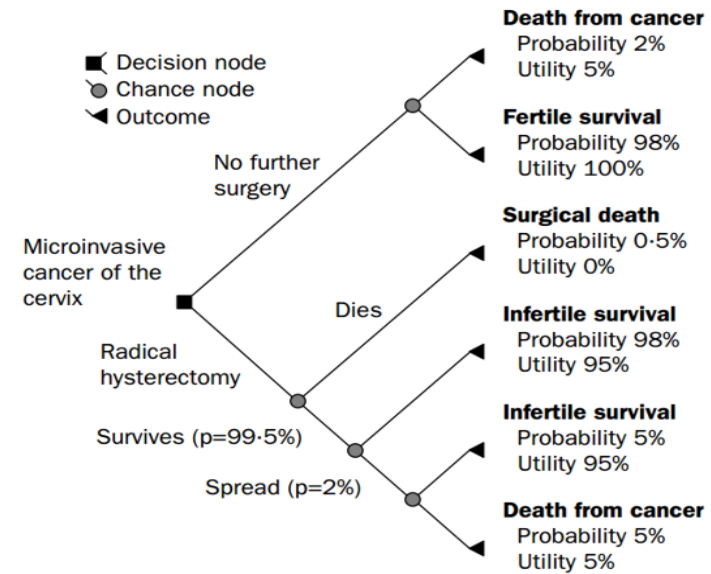
# 00 Reflection



2



3



4

# 01 History of DSS = History of AI

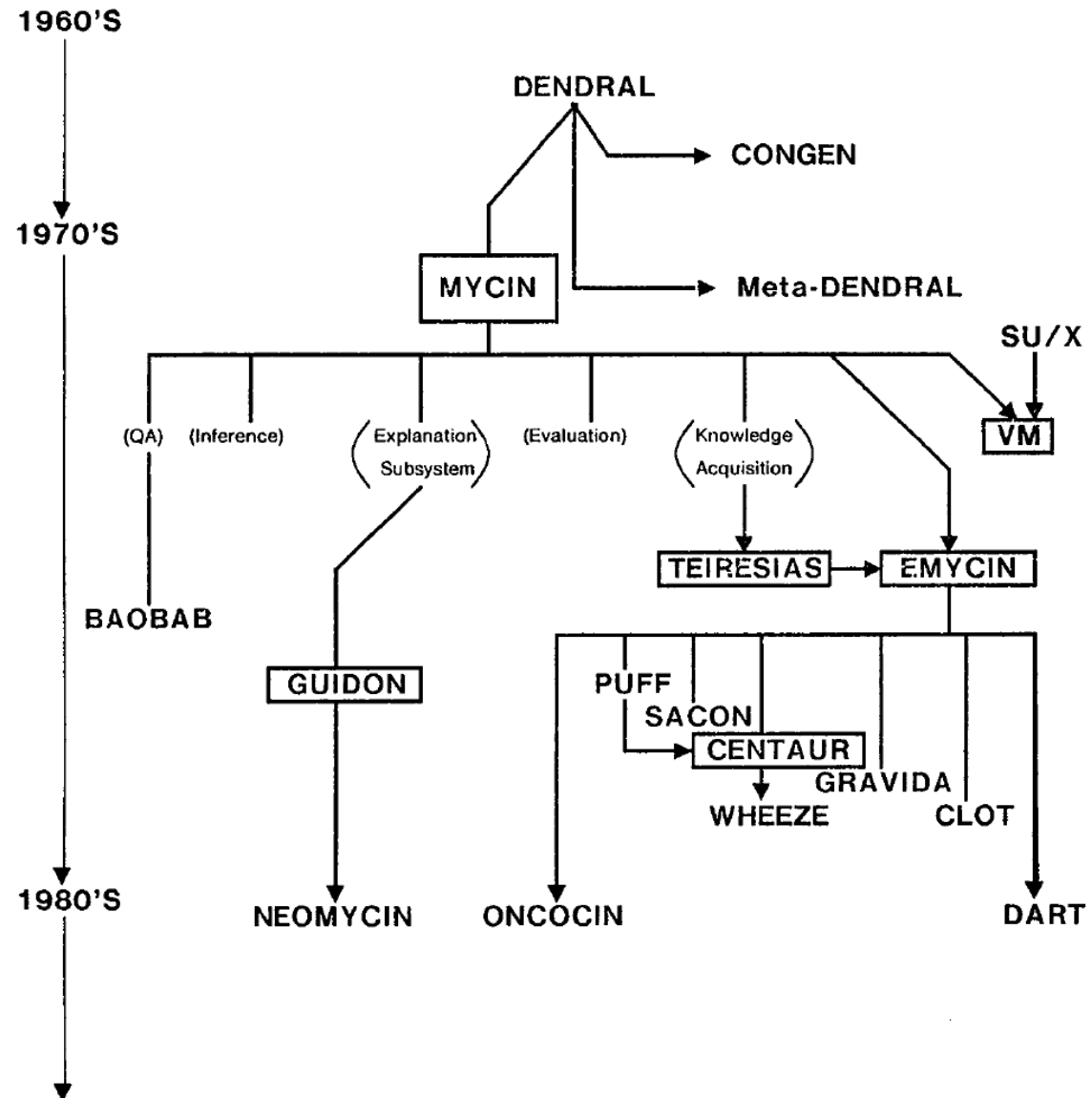
- **1943** McCulloch, W.S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biology, 5, (4), 115-133, doi:10.1007/BF02459570.
- **1950** Turing, A.M. Computing machinery and intelligence. Mind, 59, (236), 433-460.
- 1958 John McCarthy Advice Taker: programs with common sense
- **1959** Samuel, A.L. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3, (3), 210-229, doi:10.1147/rd.33.0210.
- **1975** Shortliffe, E.H. & Buchanan, B.G. 1975. A model of inexact reasoning in medicine. Mathematical biosciences, 23, (3-4), 351-379, doi:10.1016/0025-5564(75)90047-4.
- 1978 Bellman, R. Can Computers Think? Automation of Thinking, problem solving, decision-making ...

- **1960+ Medical Informatics (AI Hype)**
  - Focus on data acquisition, storage, accounting (typ. “EDV”), Expert Systems
  - The term was first used in 1968 and the first course was set up 1978 !
- **1985+ Health Telematics (AI winter)**
  - Health care networks, Telemedicine, CPOE-Systems, ...
- **1995+ Web Era (AI is “forgotten”)**
  - Web based applications, Services, EPR, distributed systems, ...
- **2005+ Success statistical learning (AI renaissance)**
  - Pervasive, ubiquitous Computing, Internet of things, ...
- **2010+ Data Era – Big Data (super for AI)**
  - Massive increase of data – data integration, mapping, ...
- **2020+ Information Era – (towards explainable AI)**
  - Sensemaking, disentangling the underlying concepts, **causality**, ...

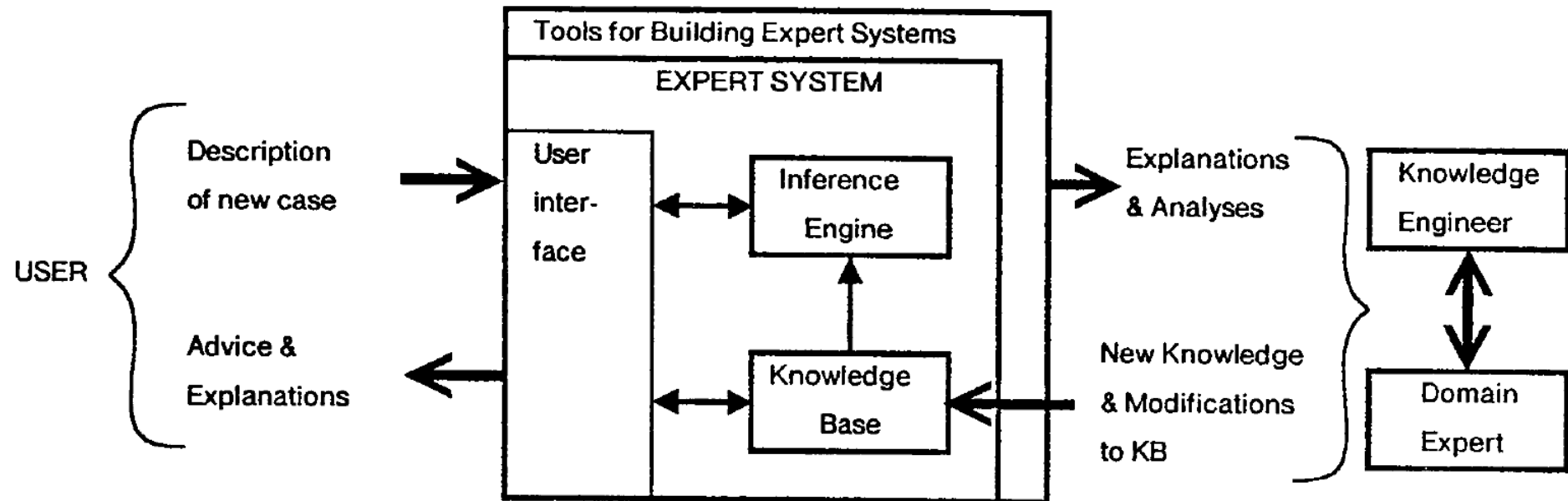


# What is the prototypical DSS (“Expert System”)

Shortliffe, E. H. &  
Buchanan, B. G. (1984)  
*Rule-based expert  
systems: the MYCIN  
experiments of the  
Stanford Heuristic  
Programming Project.*  
Addison-Wesley.

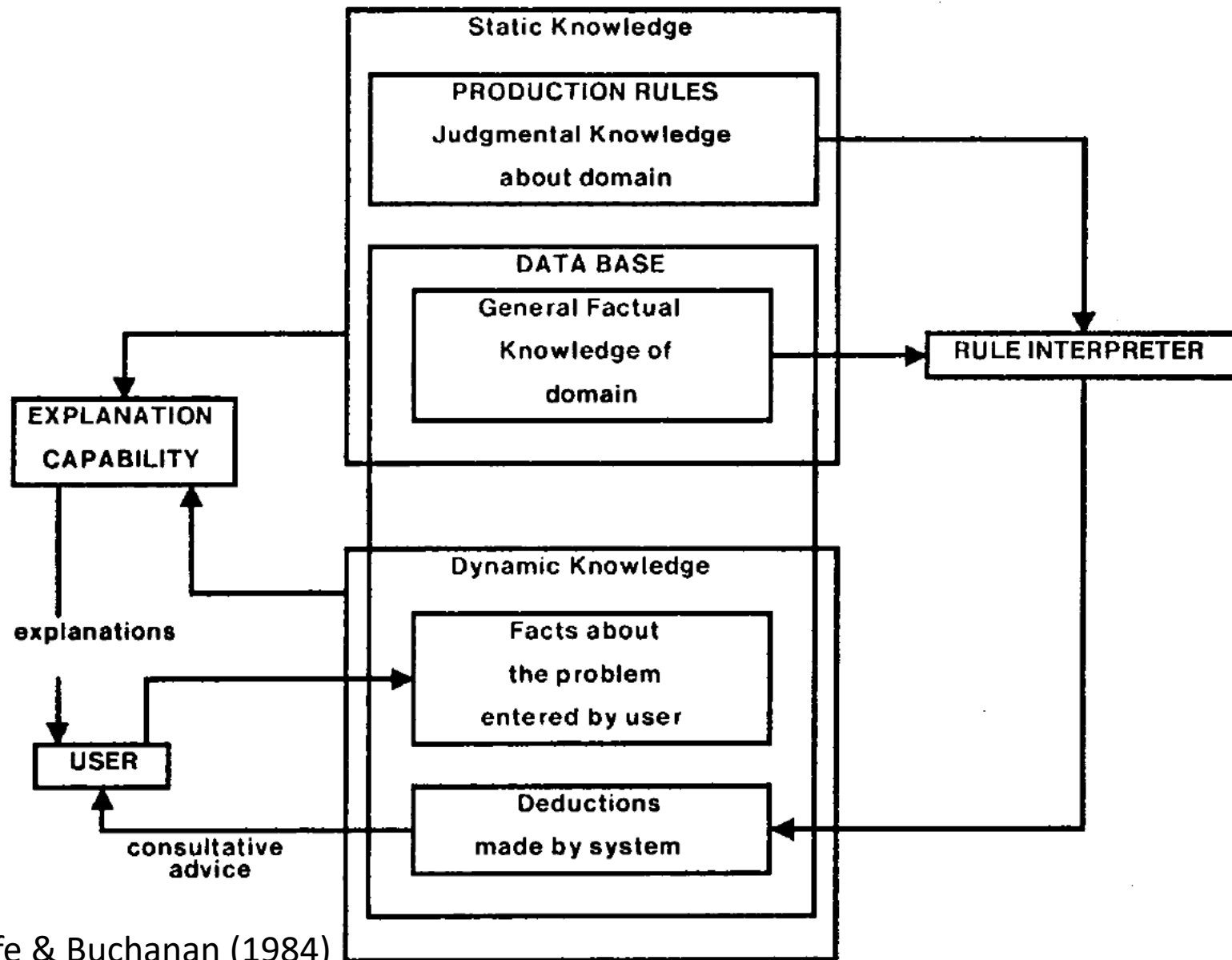


# What is the architecture of a typical DSS ?



Shortliffe, T. & Davis, R. (1975) Some considerations for the implementation of knowledge-based expert systems *ACM SIGART Bulletin*, 55, 9-12.

# What is static Knowledge versus dynamic knowledge ?



- The information available to humans is often imperfect – imprecise - uncertain.
- This is especially in the medical domain the case.
- An **human agent** can cope with deficiencies.
- Classical logic permits only **exact reasoning**:
- IF A is true THEN A is non-false and  
IF B is false THEN B is non-true
- Most real-world problems do not provide this exact information, mostly it is inexact, incomplete, uncertain and/or **un-measurable!**

- MYCIN is a rule-based Expert System, which is used for therapy planning for patients with bacterial infections
- Goal oriented strategy (“Rückwärtsverkettung”)
- To every rule and every entry a certainty factor (CF) is assigned, which is between 0 und 1
- Two measures are derived:
  - MB: measure of belief
  - MD: measure of disbelief
- Certainty factor – CF of an element is calculated by:
$$CF[h] = MB[h] - MD[h]$$
- CF is positive, if more evidence is given for a hypothesis, otherwise CF is negative
- $CF[h] = +1 \rightarrow h$  is 100 % true
- $CF[h] = -1 \rightarrow h$  is 100% false

$h_1$  = The identity of ORGANISM-1 is streptococcus

$h_2$  = PATIENT-1 is febrile

$h_3$  = The name of PATIENT-1 is John Jones

$CF[h_1, E] = .8$  : There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus

$CF[h_2, E] = -.3$  : There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile

$CF[h_3, E] = +1$  : It is definite (1) that the name of PATIENT-1 is John Jones

Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

# Why was MYCIN *no* success in the clinical routine ?

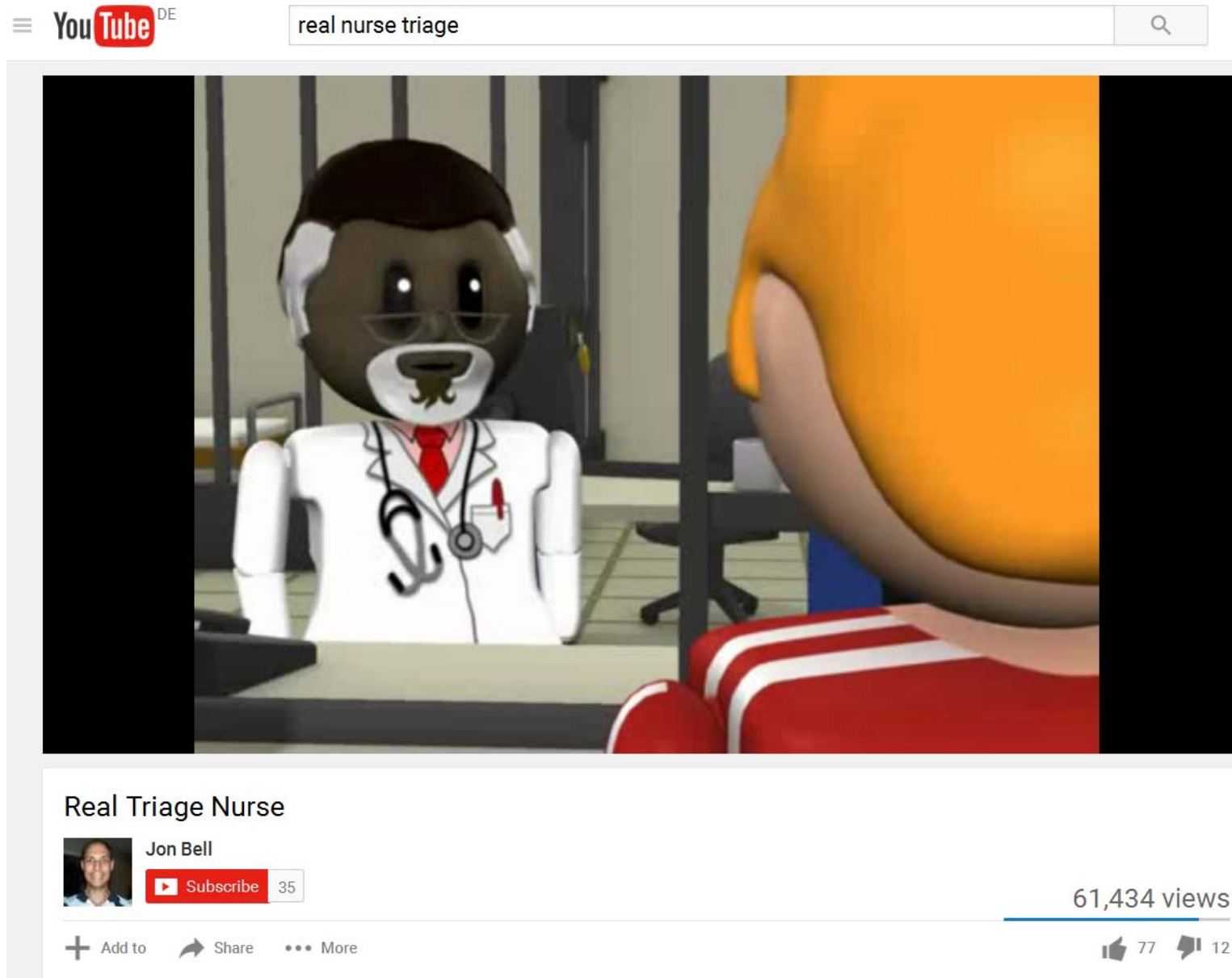




Image credit to Bernhard Schölkopf

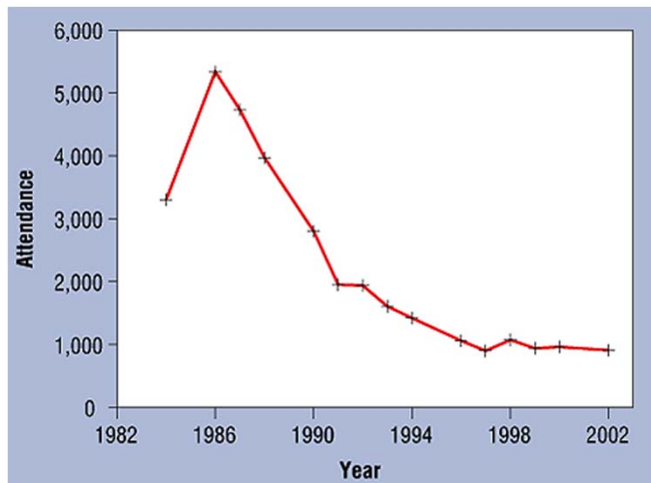




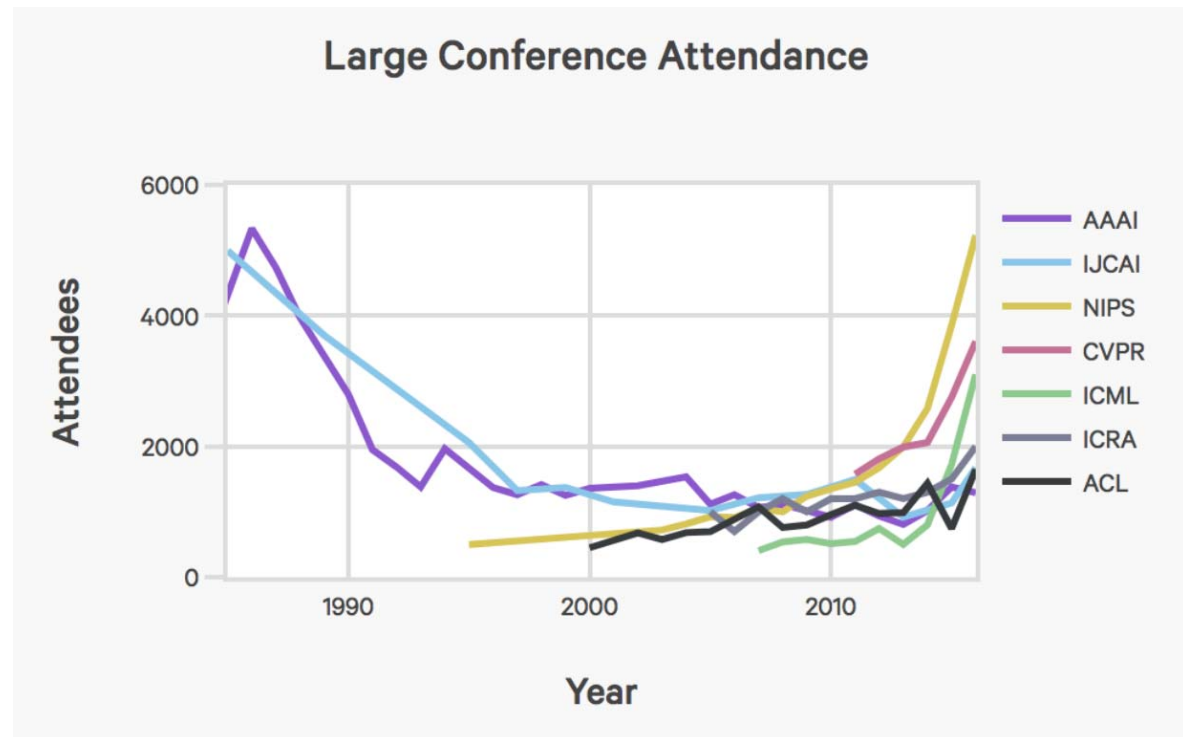
Image credit to Bernhard Schölkopf



<https://blogs.dxc.technology/2017/04/25/are-we-heading-toward-an-ai-winter/>



<https://www.computer.org/cs/mags/ex/2003/03/x3018.html>



<https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0>

# Why is the history of “Deep Learning” interesting for us ?

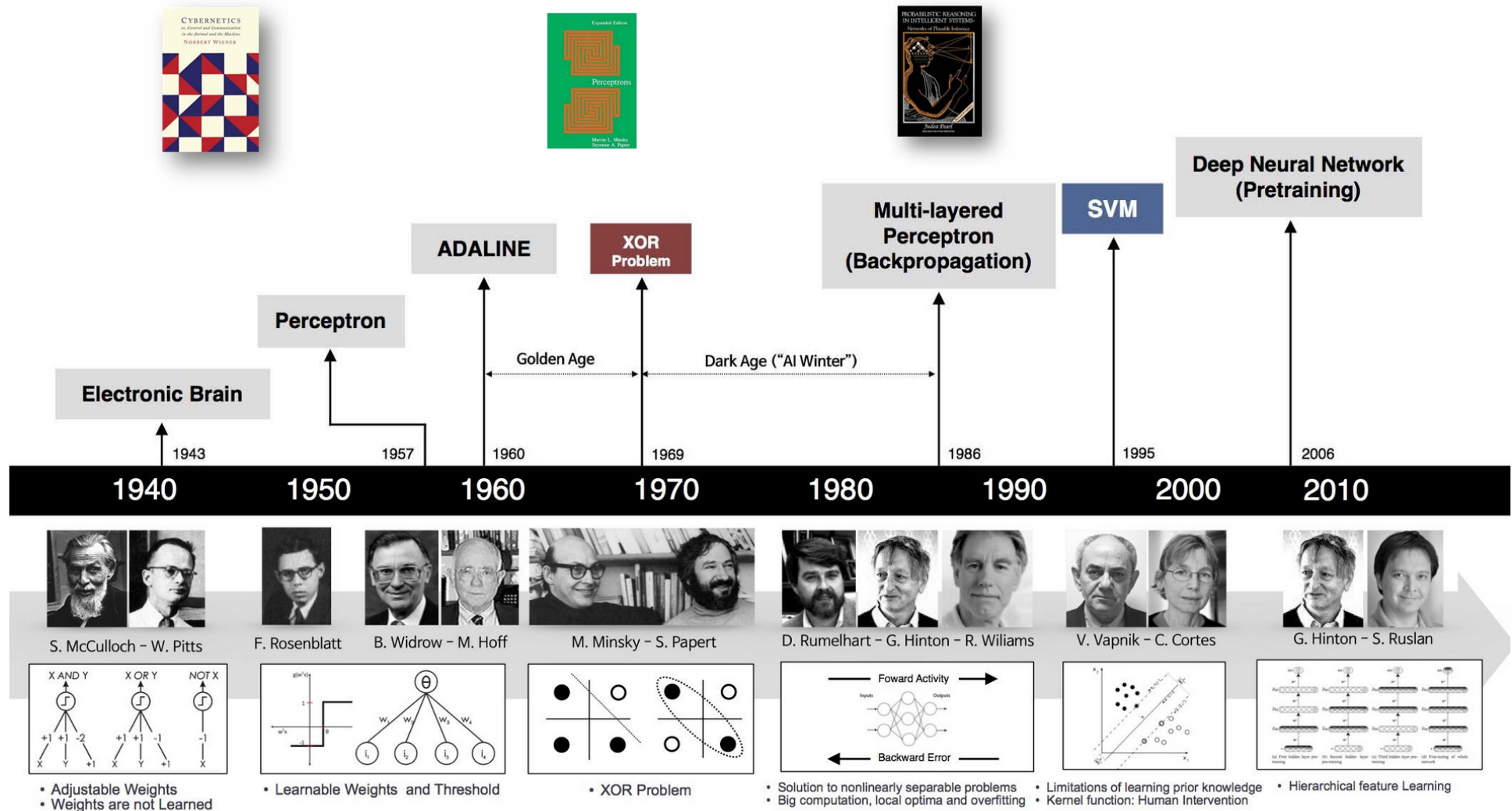


Image source: Andrew Beam, Department of Biomedical Informatics, Harvard Medical School

<https://slides.com/beamandrew/deep-learning-101/#/12>

This image is used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students

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

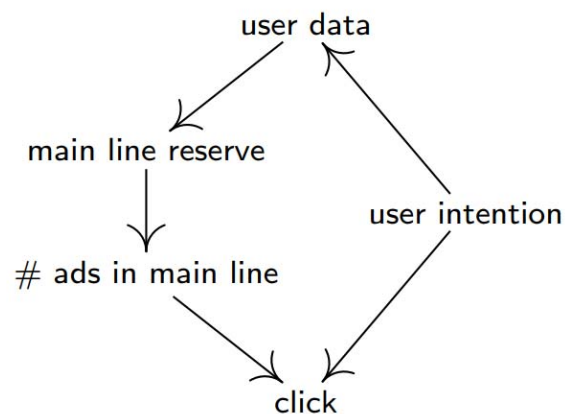
# 02 Causality and Decision Making

- 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.

- 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.



## 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



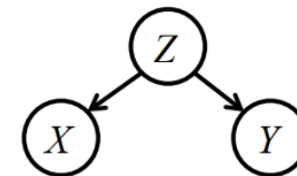
Teaching Statistics  
Volume 22, Issue 2  
38, June 2000

Country	Area (km <sup>2</sup> )	Storks (pairs)	Humans (10 <sup>6</sup> )	Birth rate (10 <sup>3</sup> /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	<a href="mailto:rajm@compuserve.com">mailto:rajm@compuserve.com</a>	
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.

- 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:
    - $A, B, \dots$  events
    - $X, Y, Z$  random variables
    - $P \dots$  probability measure
    - $P_X \dots$  probability distribution of  $X$
    - $p \dots$  probability density
    - $p(X) \dots$  Density of  $P_X$
    - $p(x)$  probability density of  $P_X$  evaluated at the point  $x$



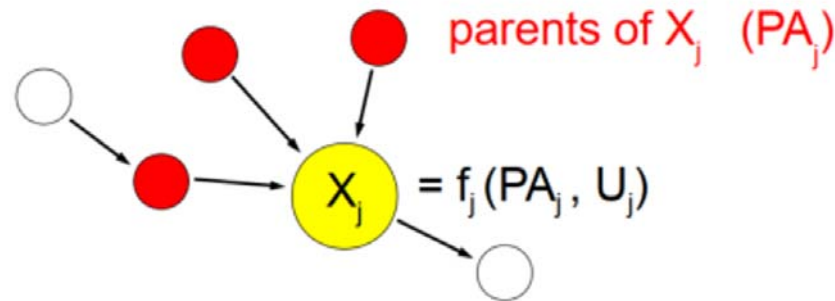
Hans Reichenbach 1956. The direction of time (Edited by Maria Reichenbach), Mineola, New York, Dover.

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

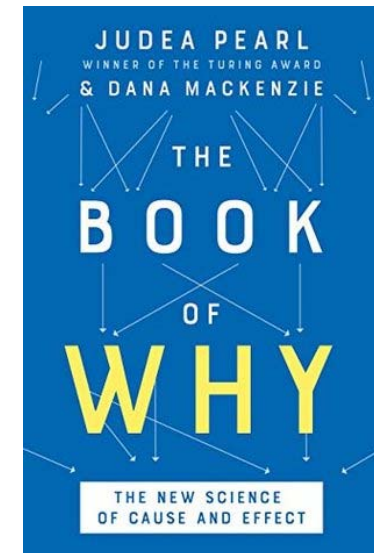
For details please refer to the excellent book of: Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA). <https://mitpress.mit.edu/books/elements-causal-inference>



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



- 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

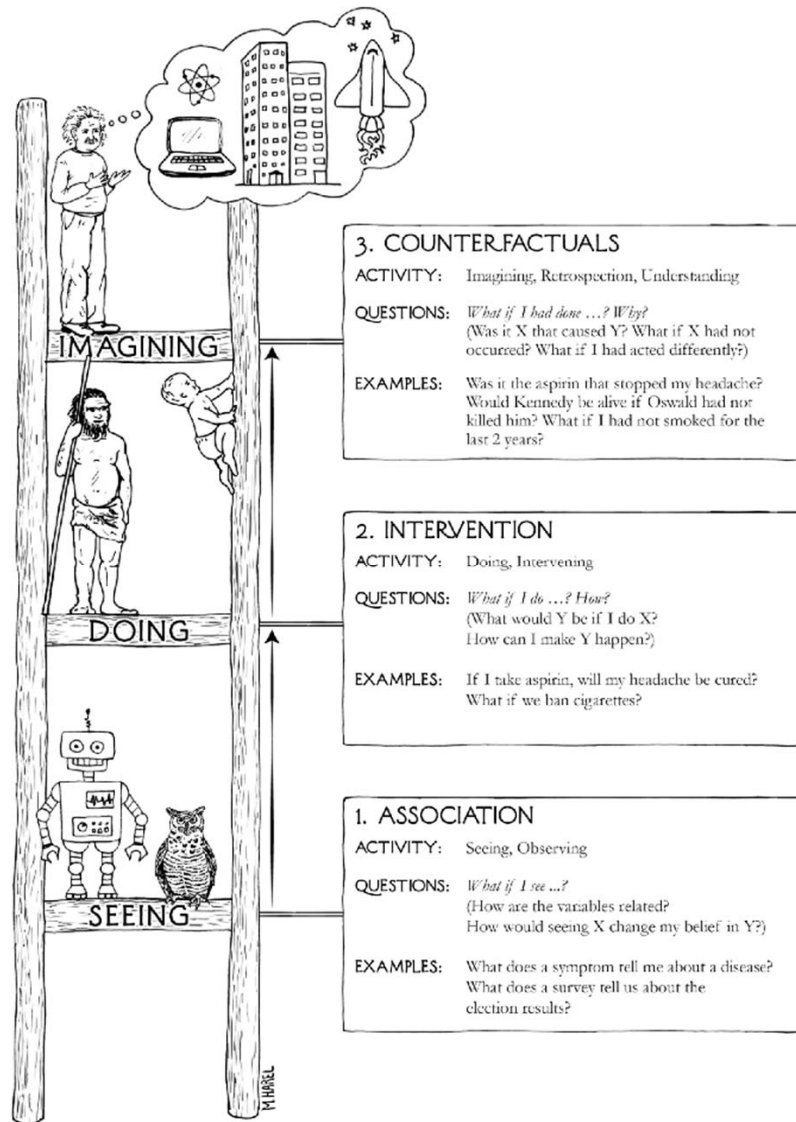


Figure 1.2 from Judea Pearl & Dana Mackenzie 2018. The book of why, New York, Basic Books,  
 Source: Illustrator: Maayan Harel, <http://www.maayanillustration.com>

## 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?)

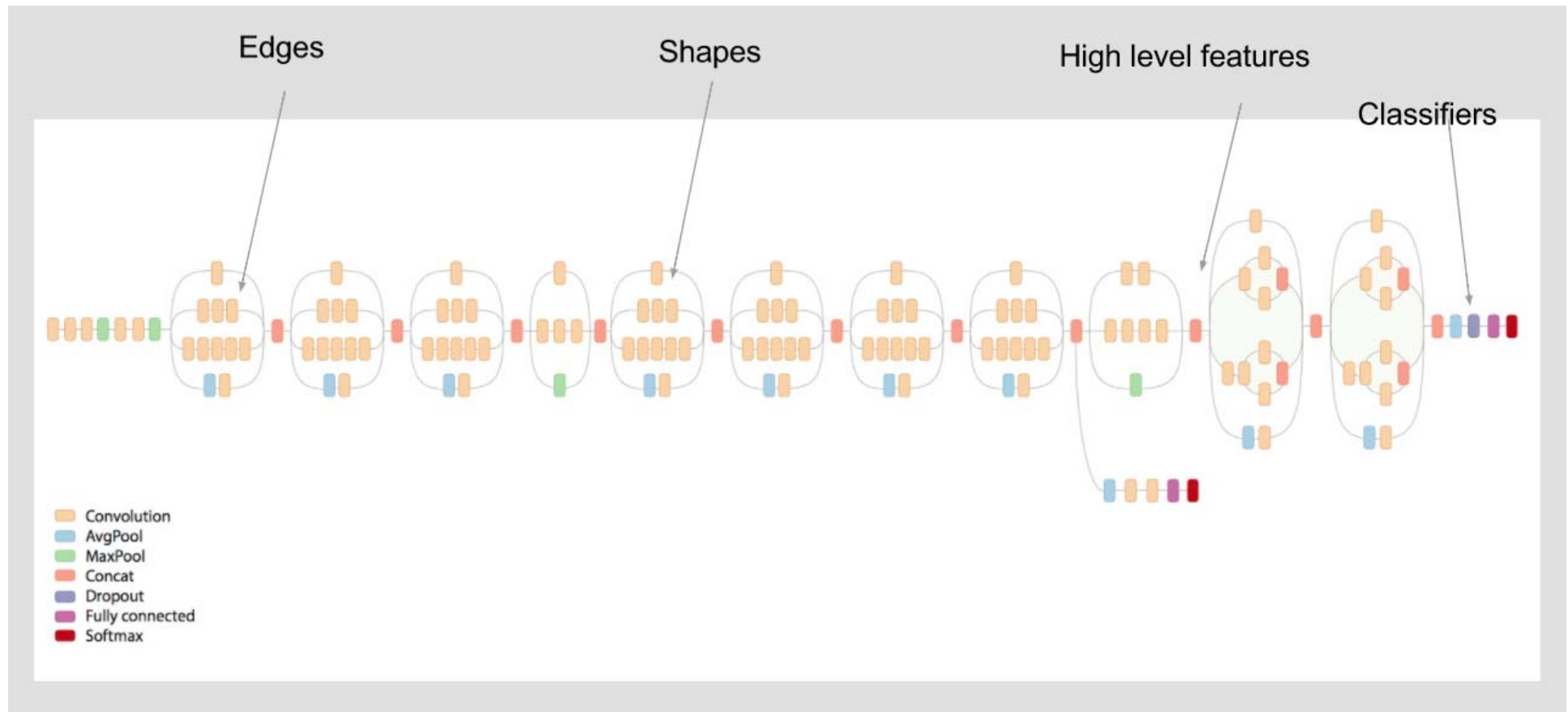
**EXAMPLES:** If I take aspirin, will my headache be cured? What if we ban cigarettes?

## 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?

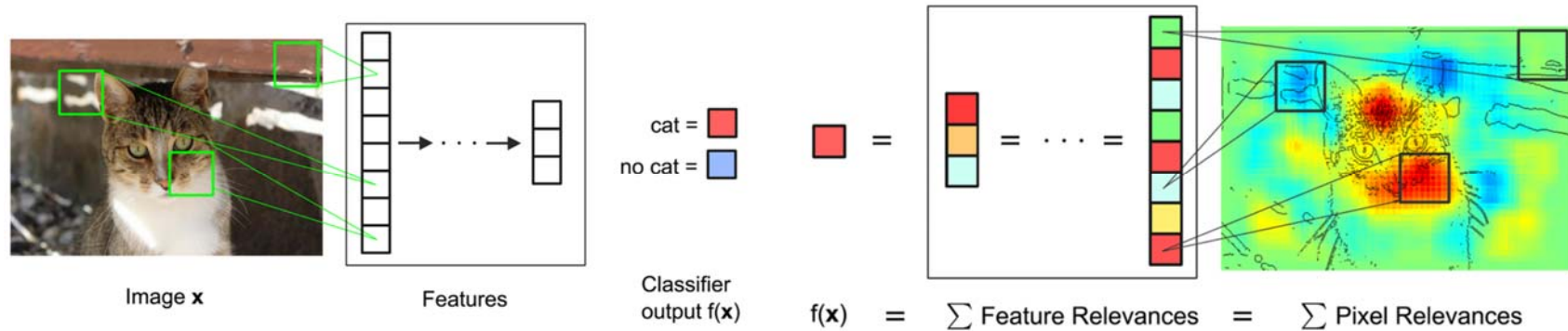


Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens & Zbigniew Wojna. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016. 2818-2826.



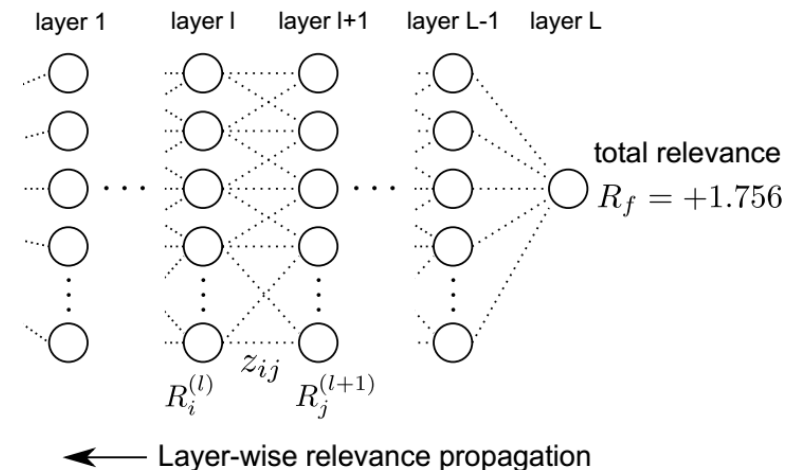
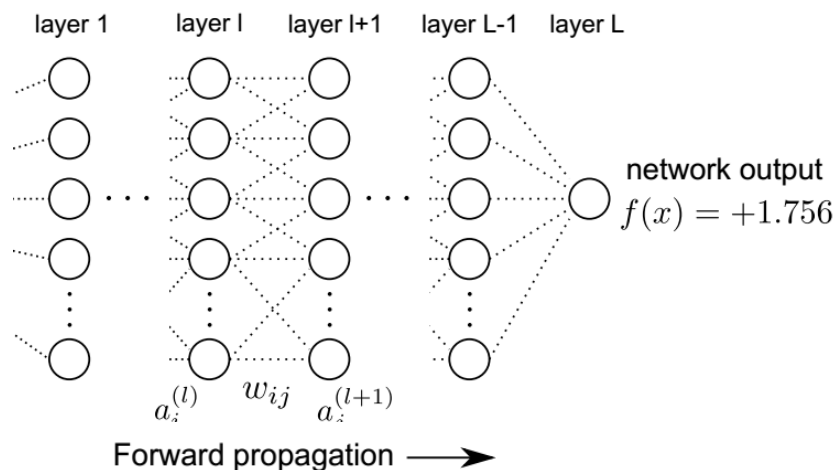
# What can Explainable AI methods do ?

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



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

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



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

# Why is explainable AI only a first step ?

*Why did the algorithm do that?*

*Can I trust these results?*

*How can I correct an error?*



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



Input data

## A possible solution



*The domain expert can understand why ...*

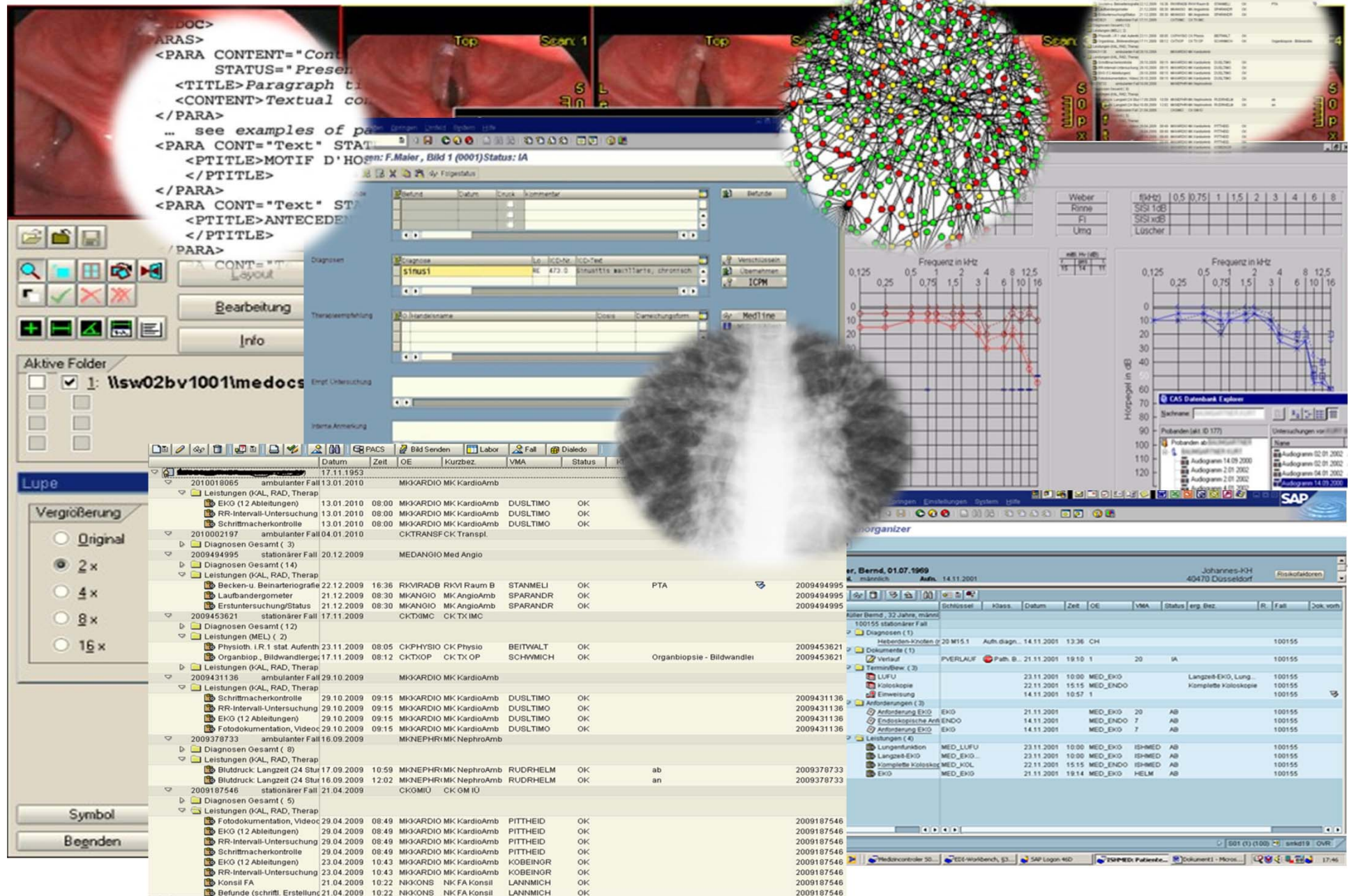
*The domain expert can learn and correct errors ...*

*The domain expert can re-enact on demand ...*

# 03 Medical Communication



# What are the key problems in medical data management ?





**Biomedical R&D data**  
(e.g. clinical trial data)

**Clinical patient data**  
(e.g. EPR, lab, reports etc.)

# The combining link is text

**Health business data**  
(e.g. costs, utilization, etc.)

**Private patient data**  
(e.g. AAL, monitoring, etc.)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity*. Washington (DC), McKinsey Global Institute.



## Radiologischer Befund

angelegt am 06.05.2006/20:21  
geschr. von  
gedruckt am 17.11.2006/08:21  
Anl: NCHB

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: Thorax eine Ebene liegend

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

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

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

Mit kollegialen Grüßen

\*\*\* Elektronische Freigabe durch am 09.05.2006 \*\*\*

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



- ... and requires information exchange



Holzinger, A., Geierhofer, R., Ackerl, S. & Searle, G. (2005). *CARDIAC@VIEW: The User Centered Development of a new Medical Image Viewer*. Central European Multimedia and Virtual Reality Conference, Prague, Czech Technical University (CTU), 63-68.

**Radiologischer Befund**

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

**Kurzanamnese:** St.p. SHT  
**Fragestellung:** -  
**Untersuchung:** Thorax eine Ebene liegend [REDACTED]

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

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

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

Mit kollegialen Grüßen

[REDACTED]

\*\*\* Elektronische Freigabe durch [REDACTED] am 09.05.2006 \*\*\*

**Special Words  
Language Mix  
Abbreviations  
Errors ...**

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

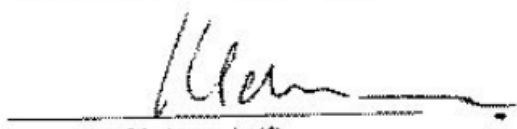
## Why is synonymy and ambiguity such a huge problem?

Untersuchungsbefund / Beschwerden: *perf. Antrumschleimhaut für histol. Untersuchung*  
*Leber 1/3 Jg. 1/3 in a. b. d. in unklare Antrumschleimhaut*  
*in Leber unklare Schleimhaut & Antrumschleimhaut & Antrumschleimhaut*  
*& Antrumschleimhaut & Antrumschleimhaut & Antrumschleimhaut*  
*Leber 1/3 Jg. 1/3 in a. b. d. in unklare Antrumschleimhaut*  
*in Leber unklare Schleimhaut & Antrumschleimhaut & Antrumschleimhaut*

Diagnose: *unklare Antrumschleimhaut. DD: Leber 1/3 Jg. 1/3*

Empfehlung / Therapie: *histol. Schleimhaut in unklare Antrumschleimhaut*  
*Antrumschleimhaut in unklare Antrumschleimhaut*  
*Py. 1/3 Jg. 1/3 in a. b. d. in unklare Antrumschleimhaut*

Mit freundlichen kollegialen Grüßen

  
-Unterschrift-

PDF 10/11/2021

„die Antrumschleimhaut ist durch Lymphozyten infiltriert“  
„lymphozytäre Infiltration der Antrum mukosa“  
„Lymphozyteninfiltration der Magenschleimhaut im Antrumbereich“

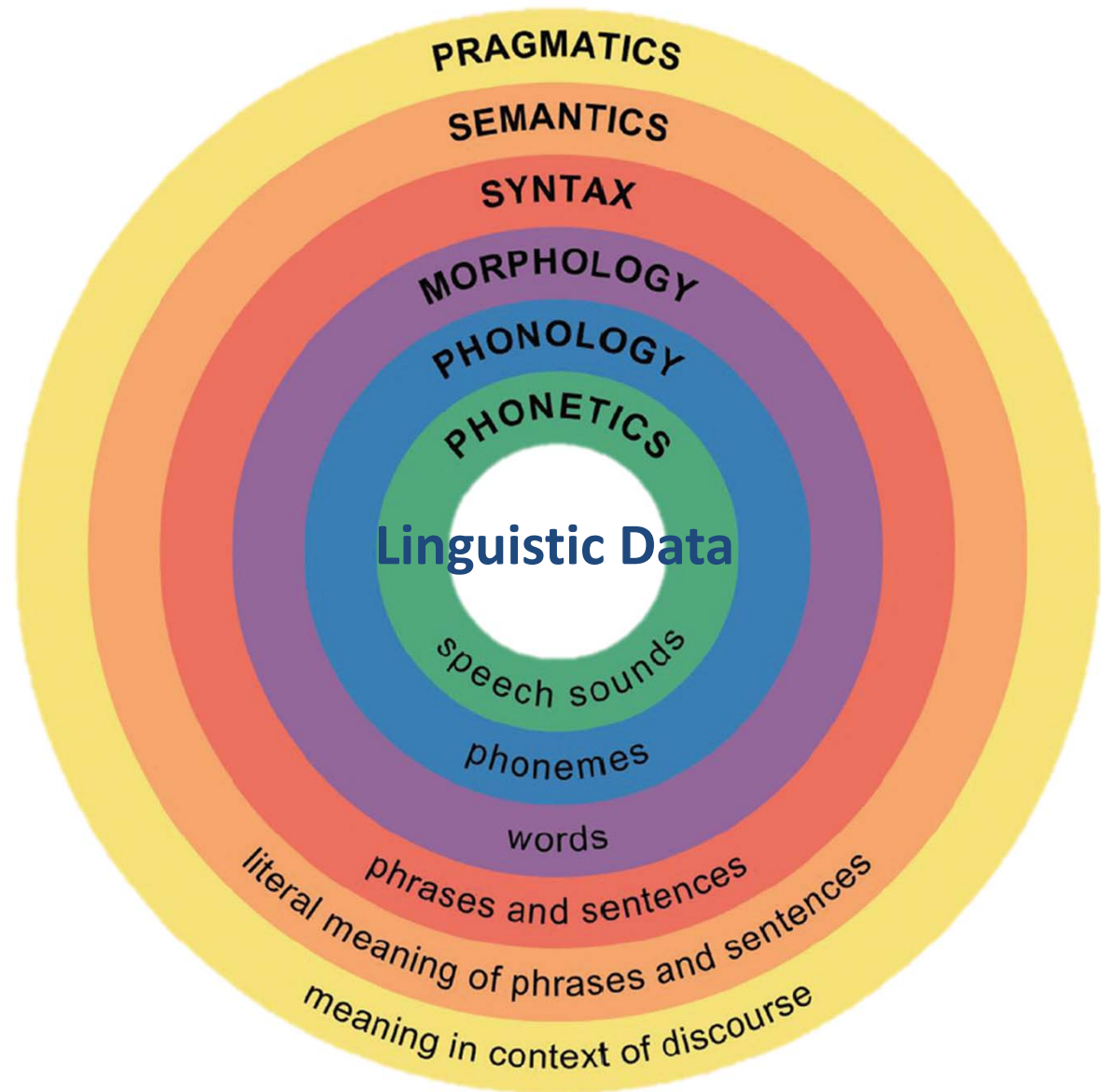
- Syntax
- Semantics
- Pragmatics
- Context
- (Emotion)



"a young boy is holding a  
baseball bat."

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

Image Source: <https://cs.stanford.edu/people/karpathy/deepimagesent/>



Thomas, J. J. & Cook, K. A.  
2005. *Illuminating the path:  
The research and  
development agenda for  
visual analytics*, New York,  
IEEE Computer Society Press.



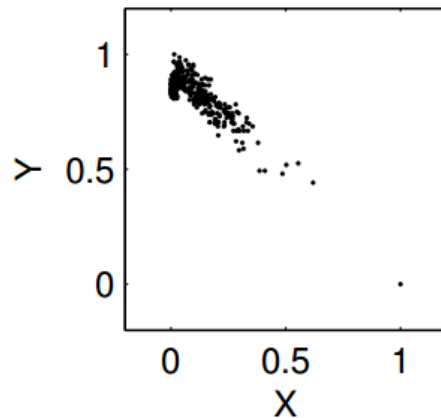
# 04 Causal Reasoning

- “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  $Y \rightarrow X$

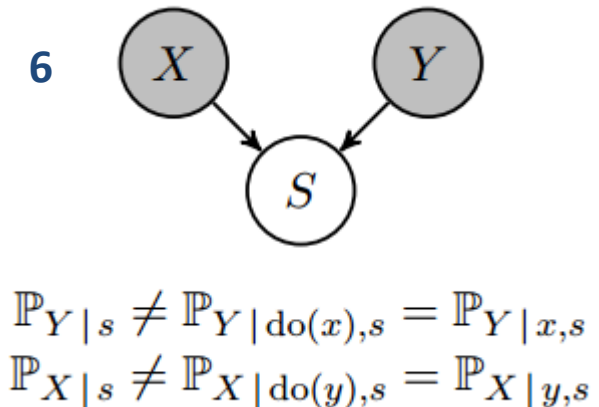
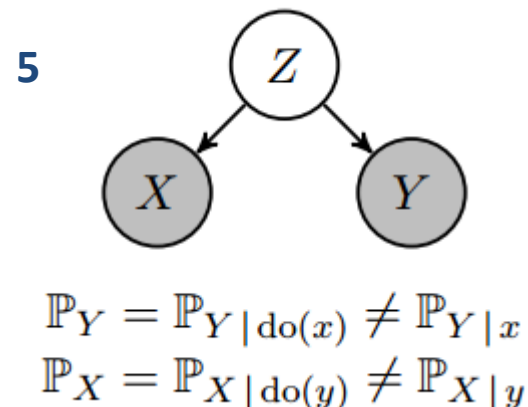
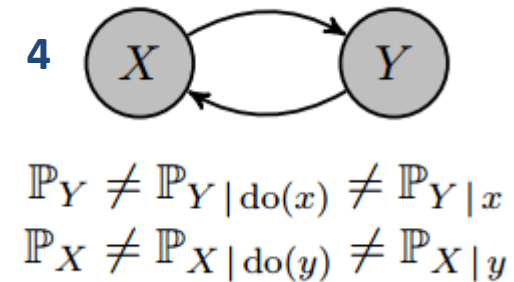
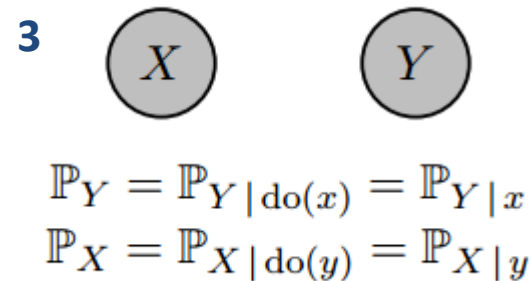
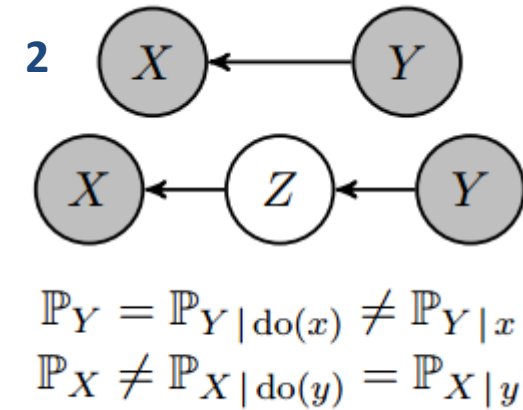
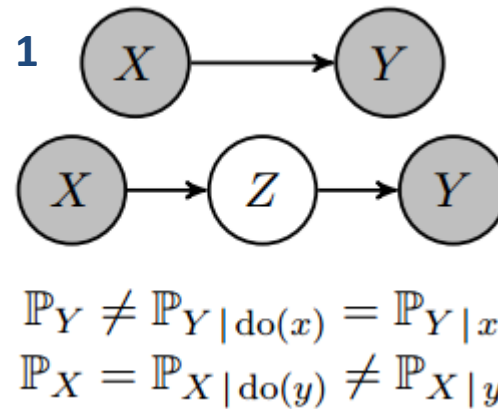
Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

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 $X \rightarrow Y$ , or $Y \rightarrow X$ using only observed data

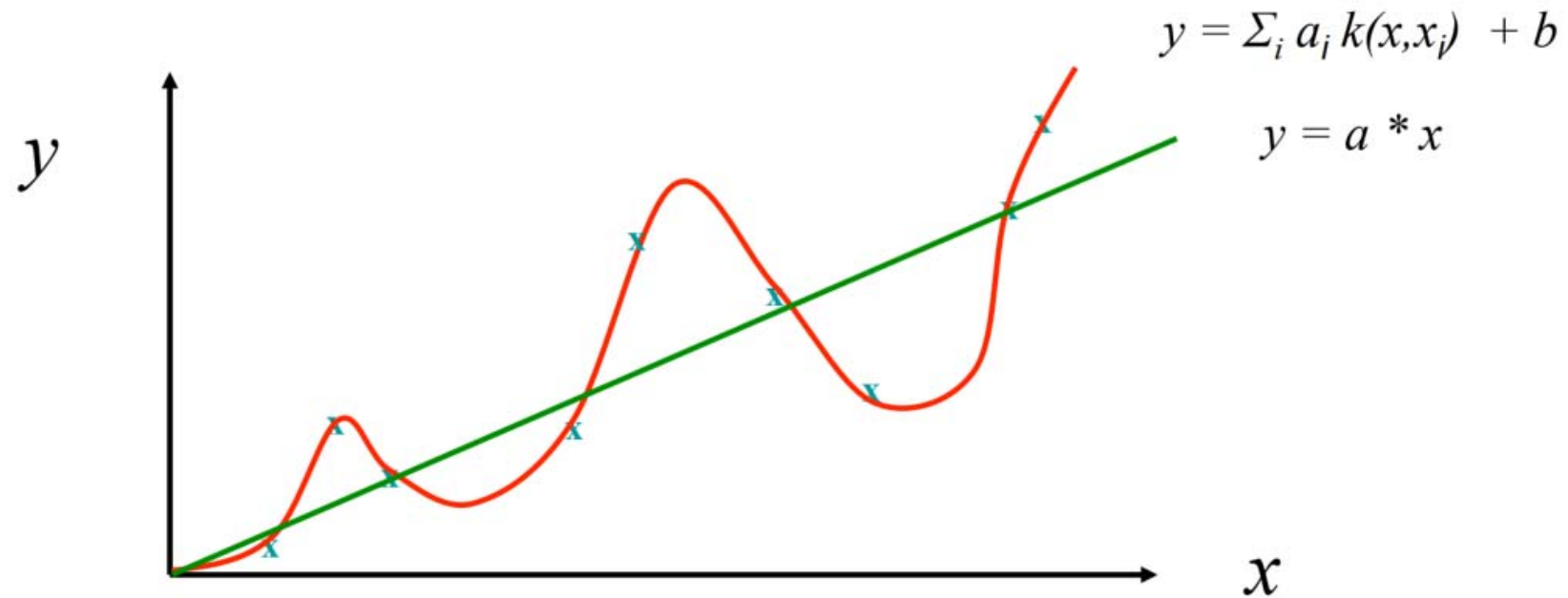


Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.



- **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.

- $:=$  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.



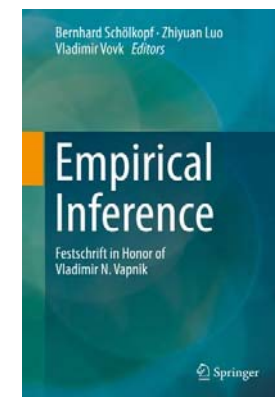
Gottfried W. Leibniz (1646-1716)

Hermann Weyl (1885-1955)

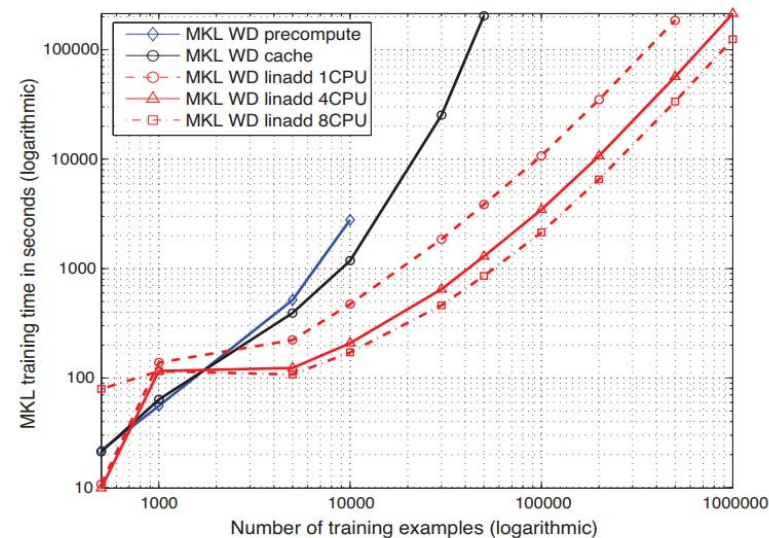
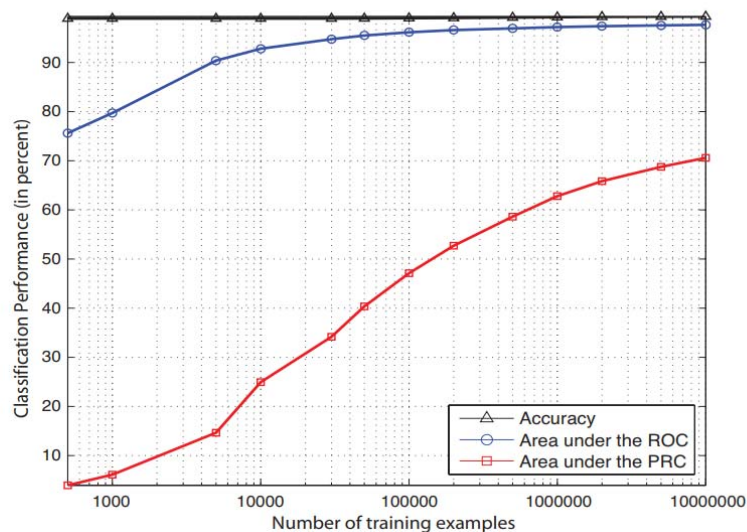
Vladimir Vapnik (1936-)

Alexey Chervonenkis (1938-2014)

Gregory Chaitin (1947-)



- 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.



**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  $N_B \sim \text{Ber}(0.01)$ ; note that the corresponding causal graph is  $T \rightarrow B$ .

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,  $\delta_1$ . This leads to a modified SCM:



$$\begin{aligned} \mathcal{C}|B=1, T=1: \quad T &:= 1 \\ B &:= T \cdot 1 + (1-T) \cdot (1-1) = T \end{aligned} \quad (3.6)$$

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:

$$\mathcal{C}|B=1, T=1; do(T:=0): \quad \begin{aligned} T &:= 0 \\ B &:= T \end{aligned} \quad (3.7)$$

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

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

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

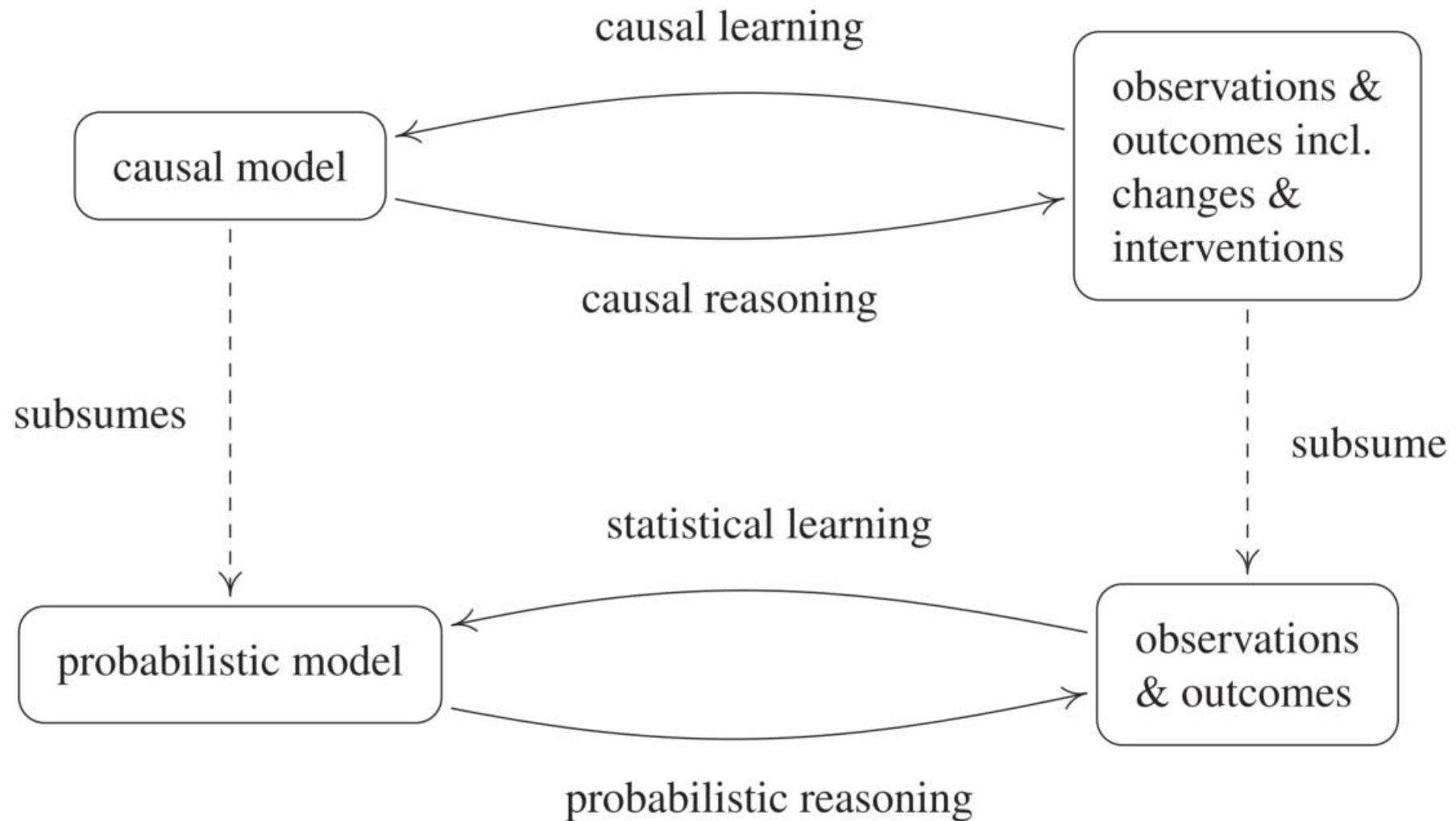
$$\begin{aligned} P^{\mathcal{C}; do(T:=1)}(B=0) &= 0.99 \quad \text{and} \\ P^{\mathcal{C}; do(T:=0)}(B=0) &= 0.01, \end{aligned}$$

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

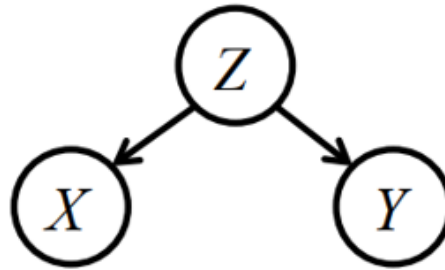
Interestingly, Example 3.4 shows that we can use counterfactual statements to falsify the underlying causal model (see Section 6.8). Imagine that the rare condition  $N_B$  can be tested, but the test results take longer than a day. In this case, it is possible that we observe a counterfactual statement 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.

# 05 Interpretability: Mapping AI with Human Intelligence



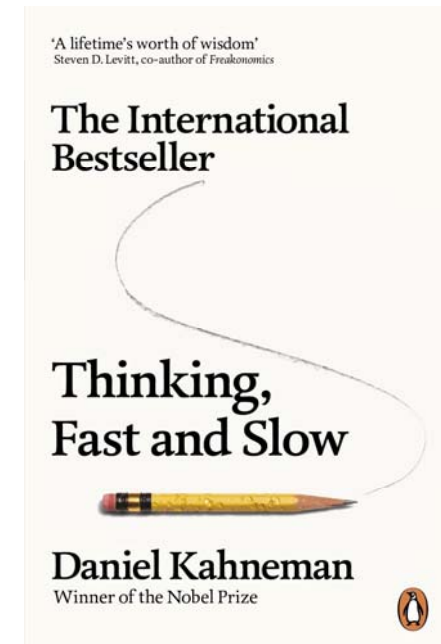
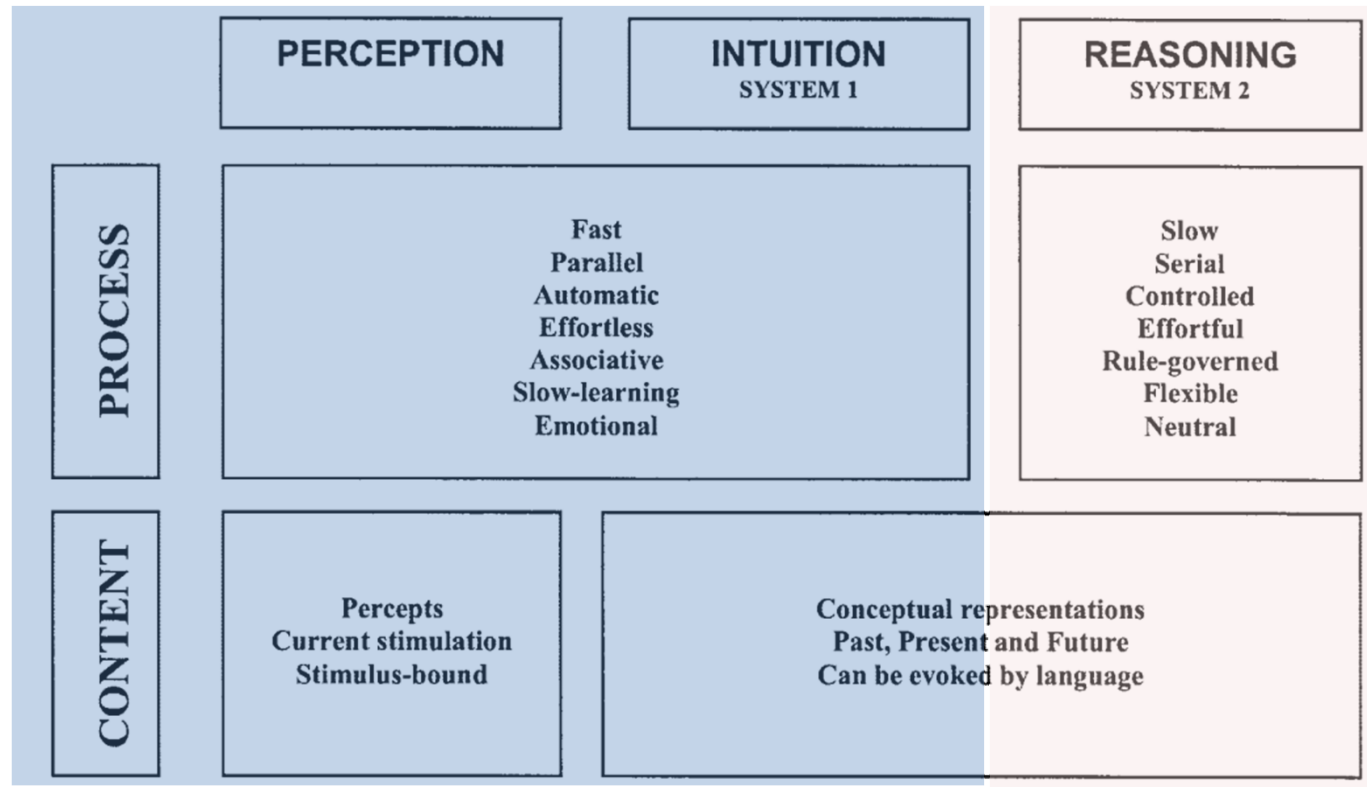
Jonas Peters, Dominik Janzing & Bernhard Schölkopf 2017. Elements of causal inference: foundations and learning algorithms, Cambridge (MA).



# There is no correlation without causation

Hans Reichenbach (1956). The direction of time. New York: Dover.

# Why is this important for deep learning ?



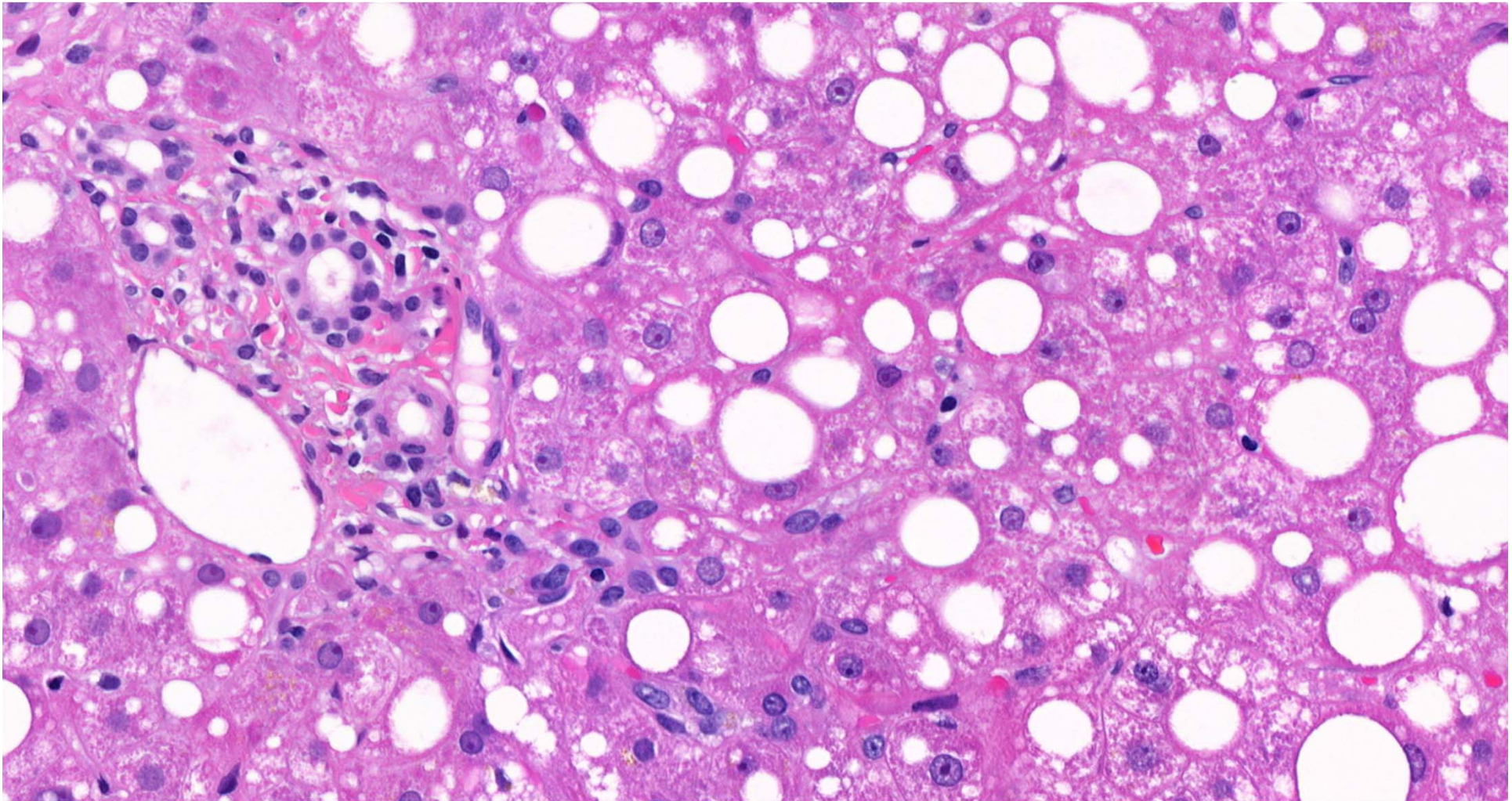
Amos Tversky & Daniel Kahneman 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185, (4157), 1124-1131, doi:10.1126/science.185.4157.1124.

(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,
- in particular between  $P(x)$  and  $P(Y|X)$ , with a causal link between  $Y \rightarrow X$

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.

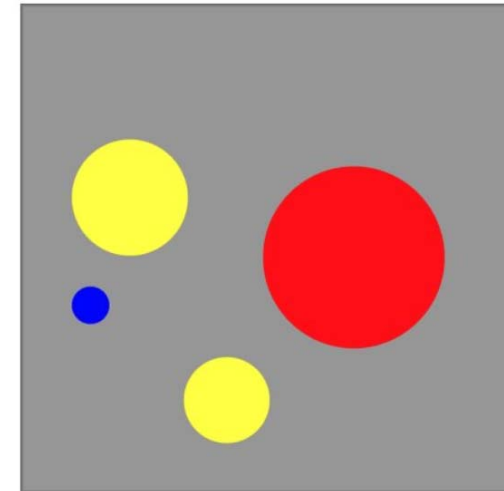
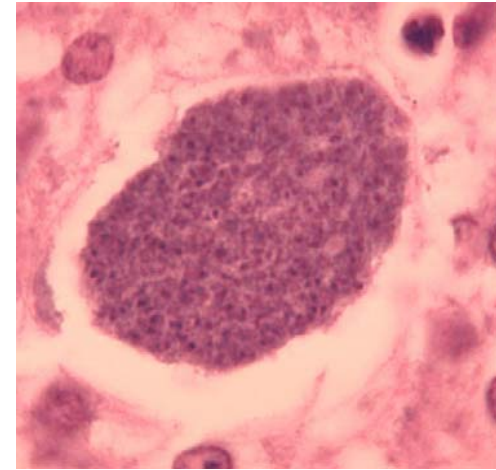






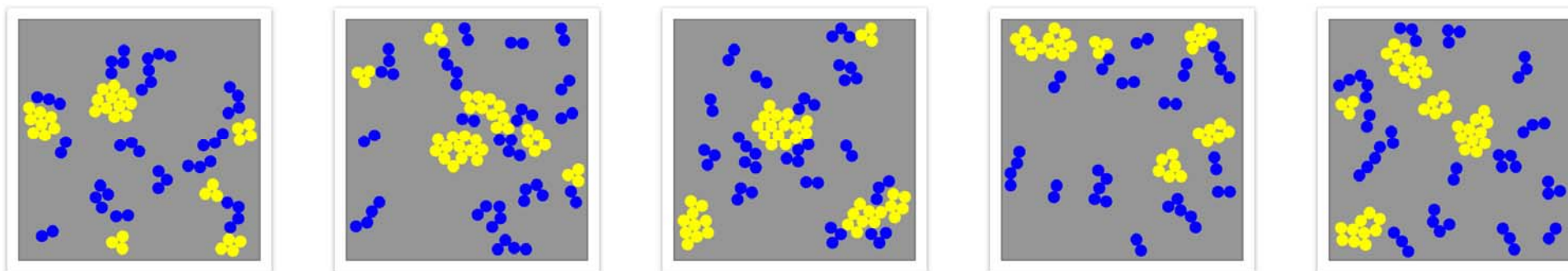
- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = 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.

- 1) ground truth is not always well defined, especially when making a medical diagnosis;
- 2) although human (scientific) models are often based on understanding causal mechanisms,
- today's successful machine models or algorithms are typically based on correlation or related concepts of similarity and distance!

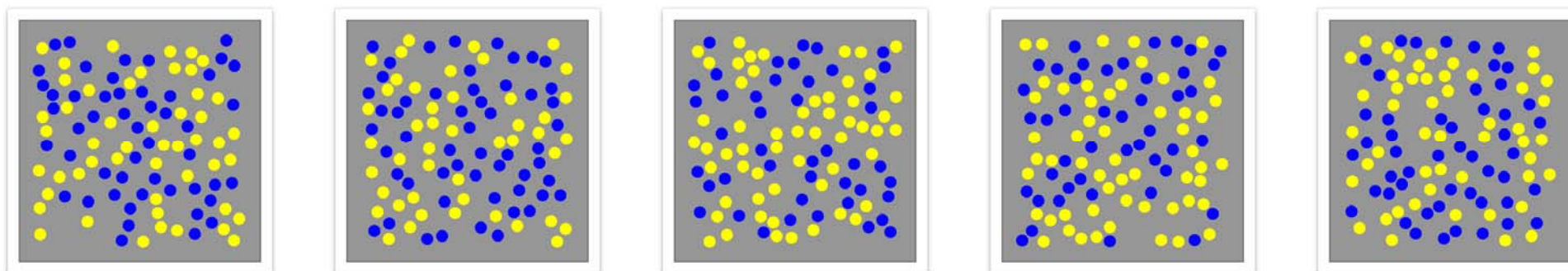


# What are domain concepts from histopathology ?

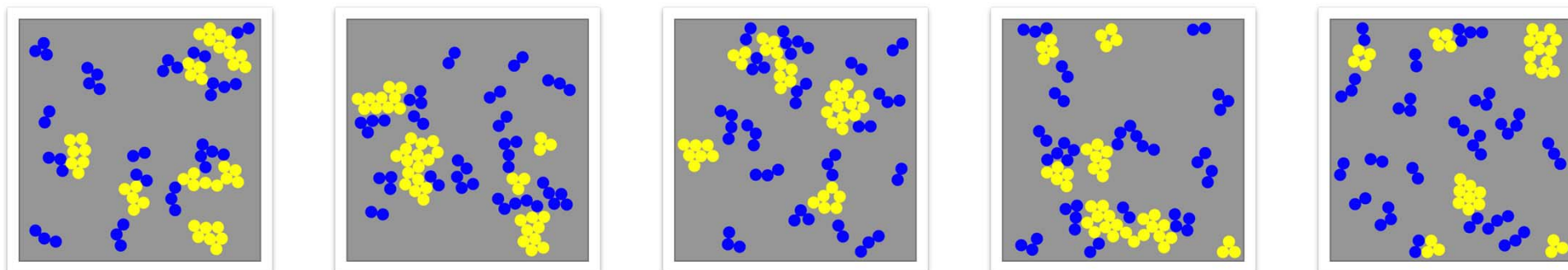
A) True (the cells are smaller and closer together – it is an tumor ...)



B) False



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



# Causality: The art and science of cause and effect

Judea Pearl 2000. Causality: Models, Reasoning, and Inference,  
Cambridge: Cambridge University Press.

# **Causability:**

# **Mapping machine**

# **explanations with**

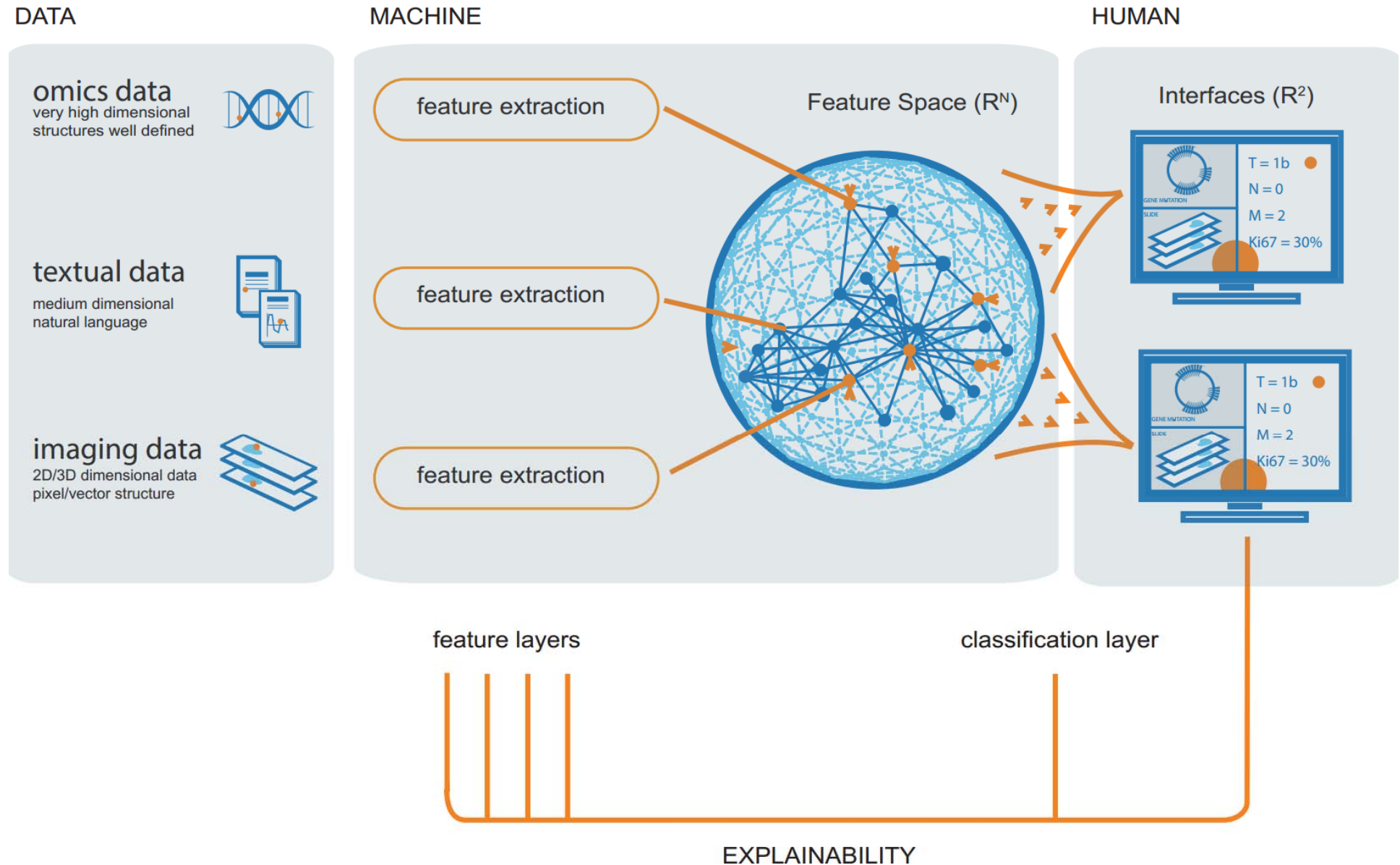
# **human understanding**

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.

# 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

# Conclusion







**HCAI**  
HUMAN-CENTERED.AI

**Thank you!**