



Andreas Holzinger  
185.A83 Machine Learning for Health Informatics  
2019S, VU, 2.0 h, 3.0 ECTS  
Lecture 02 – Dienstag, 19.03.2019



## From Clinical Decision Support to Causal Reasoning and explainable AI

andreas.holzinger AT tuwien.ac.at

<https://human-centered.ai/machine-learning-for-health-informatics-class-2019>



human-centered.ai (Holzinger Group)

1

2019 Machine Learning for Health 02

- Decision support system (DSS)
- MYCIN – Rule Based Expert System
- GAMUTS in Radiology
- Reasoning under uncertainty
- Example: Radiotherapy planning
- Example: Case-Based Reasoning
- Explainable Artificial intelligence
- Re-trace > Understand > Explain
- Transparency > Trust > Acceptance
- Fairness > Transparency > Accountability
- Causality > Causability
- (Some) Methods of Explainable AI

human-centered.ai (Holzinger Group)

2

2019 Machine Learning for Health 02

- Causality** = fundamental relationship between cause and effect
- Causability** = similar to the concept of usability the property of a human explanation
- Case-based reasoning (CBR)** = process of solving new problems based on the solutions of similar past problems;
- Certainty factor model (CF)** = a method for managing uncertainty in rule-based systems;
- CLARION** = Connectionist Learning with Adaptive Rule Induction ON-line (CLARION) is a cognitive architecture that incorporates the distinction between implicit and explicit processes and focuses on capturing the interaction between these two types of processes. By focusing on this distinction, CLARION has been used to simulate several tasks in cognitive psychology and social psychology. CLARION has also been used to implement intelligent systems in artificial intelligence applications.
- Clinical decision support (CDS)** = process for enhancing health-related decisions and actions with pertinent, organized clinical knowledge and patient information to improve health delivery;
- Clinical Decision Support System (CDSS)** = expert system that provides support to certain reasoning tasks, in the context of a clinical decision;
- Collective Intelligence** = shared group (symbolic) intelligence, emerging from cooperation/competition of many individuals, e.g. for consensus decision making;
- Counterfactual** = relating to or expressing what has not happened or is not the case
- Crowdsourcing** = a combination of "crowd" and "outsourcing" coined by Jeff Howe (2006), and describes a distributed problem-solving model; example for crowdsourcing is a public software beta-test;
- Decision Making** = central cognitive process in every medical activity, resulting in the selection of a final choice of action out of several alternatives;
- Decision Support System (DSS)** = is an IS including knowledge based systems to interactively support decision-making activities, i.e. making data useful;

human-centered.ai (Holzinger Group)

3

2019 Machine Learning for Health 02

- DXplain** = a DSS from the Harvard Medical School, to assist making a diagnosis (clinical consultation), and also as an instructional instrument (education); provides a description of diseases, etiology, pathology, prognosis and up to 10 references for each disease;
- Etiology** = in medicine (many) factors coming together to cause an illness (see causality)
- Explainable AI** = Explainability = upcoming fundamental topic within recent AI; answering e.g. **why** a decision has been made
- Expert-System** = emulates the decision making processes of a human expert to solve complex problems;
- GAMUTS in Radiology** = Computer-Supported list of common/uncommon differential diagnoses;
- ILIAD** = medical expert system, developed by the University of Utah, used as a teaching and testing tool for medical students in problem solving. Fields include Pediatrics, Internal Medicine, Oncology, Infectious Diseases, Gynecology, Pulmonology etc.
- Interpretability** = there is no formal technical definition yet, but it is considered as a prerequisite for trust
- MYCIN** = one of the early medical expert systems (Shortliffe (1970), Stanford) to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for patient's body weight;
- 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

human-centered.ai (Holzinger Group)

4

2019 Machine Learning for Health 02

- 00 Reflection – follow-up from last lecture
- 01 Decision Support Systems (DSS)
- 02 History of DSS = History of AI
- 03 Example: Towards Personalized Medicine
- 04 Example: Case Based Reasoning (CBR)
- 05 Causal Reasoning
- 06 Explainability – Causability
- 07 (Some) Methods of Explainable AI

human-centered.ai (Holzinger Group)

5

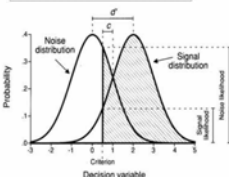
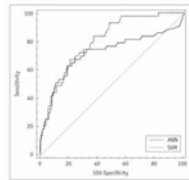
2019 Machine Learning for Health 02



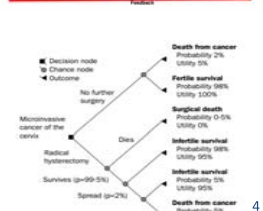
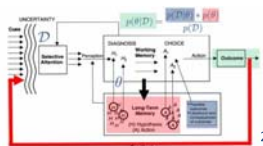
human-centered.ai (Holzinger Group)

6

2019 Machine Learning for Health 02



3

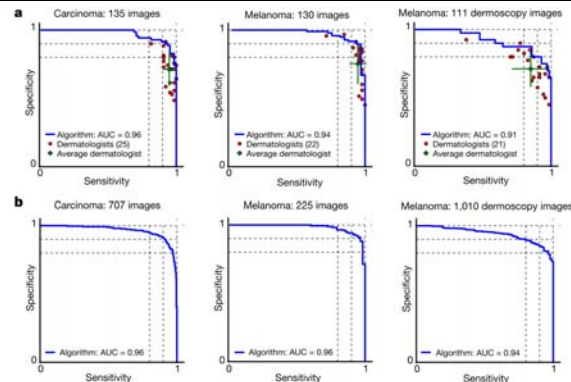


4

human-centered.ai (Holzinger Group)

7

2019 Machine Learning for Health 02



Andre Esteve, 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

human-centered.ai (Holzinger Group)

8

2019 Machine Learning for Health 02

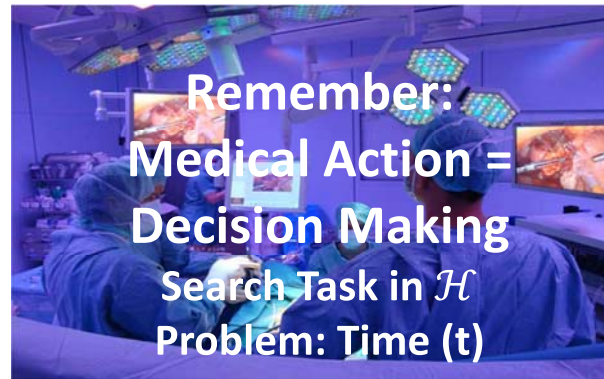
- Remember: Medicine is an complex application domain – dealing most of the time with **probable information!**
- Some challenges include:
- (a) defining hospital system architectures in terms of generic tasks such as diagnosis, therapy planning and monitoring to be executed for (b) medical reasoning 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.

human-centered.ai (Holzinger Group)

9

2019 Machine Learning for Health 02

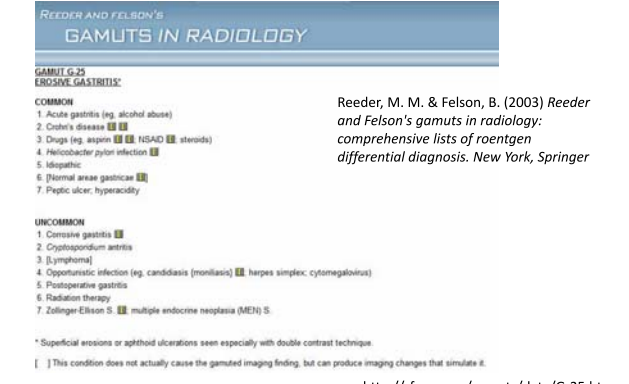
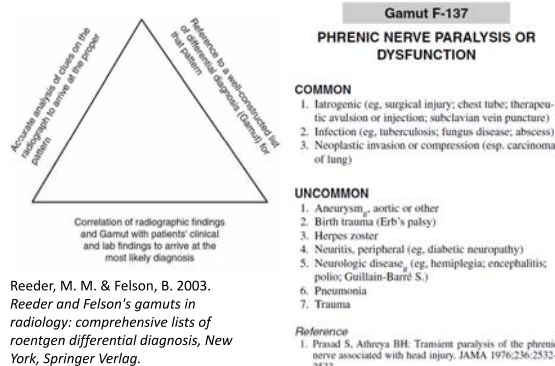
# 01 Decision Support Systems



- 400 BC Hippocrates (460-370 BC), father of western medicine:
  - A medical record should accurately reflect the course of a disease
  - A medical record should indicate the probable cause of a disease
- 1890 William Osler (1849-1919), father of modern western medicine
  - Medicine is a science of uncertainty and an art of probabilistic decision making
- Today
  - Prediction models are based on data features, patient health status is modelled as high-dimensional feature vectors ...

- Clinical guidelines are **systematically** developed documents to assist doctors and patient decisions about appropriate care;
- In order to build DS, based on a guideline, it is **formalized** (transformed from natural language to a logical algorithm), and
- implemented** (using the algorithm to program a DSS);
- To increase the quality of care, they must be linked to a process of care, for example:
  - "80% of diabetic patients should have an HbA1c below 7.0" could be linked to processes such as:
  - "All diabetic patients should have an annual HbA1c test" and
  - "Patients with values over 7.0 should be rechecked within 2 months."
- Condition-action rules** specify one or a few conditions which are linked to a specific action, in contrast to narrative guidelines which describe a series of branching or iterative decisions unfolding over time.
- Narrative guidelines and clinical rules are two ends of a continuum of clinical care standards.

Medlock, S., Opondo, D., Eslami, S., Askari, M., Wierenga, P., de Rooij, S. E. & Abu-Hanna, A. (2011) LERM (Logical Elements Rule Method): A method for assessing and formalizing clinical rules for decision support. *International Journal of Medical Informatics*, 80, 4, 286-295.



Reeder, M. M. & Felson, B. (2003) *Reeder and Felson's gamuts in radiology: comprehensive lists of roentgen differential diagnosis*. New York, Springer

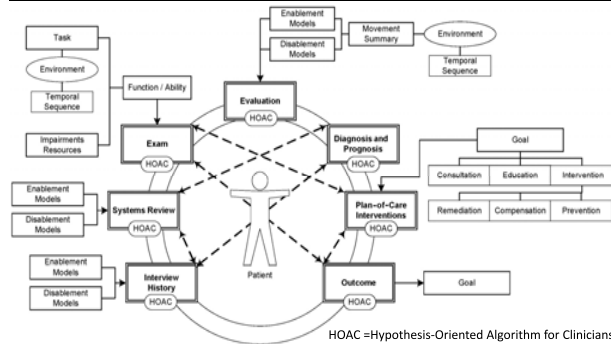
<http://rfs.acr.org/gamuts/data/G-25.htm>



Iserson, K. V. & Moskop, J. C. 2007. Triage in Medicine, Part I: Concept, History, and Types. *Annals of Emergency Medicine*, 49, (3), 275-281.

human-centered.ai (Holzinger Group)

19 Image Source: <http://store.comed-tech.com>



Schenkman, M., Deutsch, J. E. & Gill-Body, K. M. (2006) An Integrated Framework for Decision Making in Neurologic Physical Therapist Practice. *Physical Therapy*, 86, 12, 1681-1702.

human-centered.ai (Holzinger Group)

20

2019 Machine Learning for Health 02

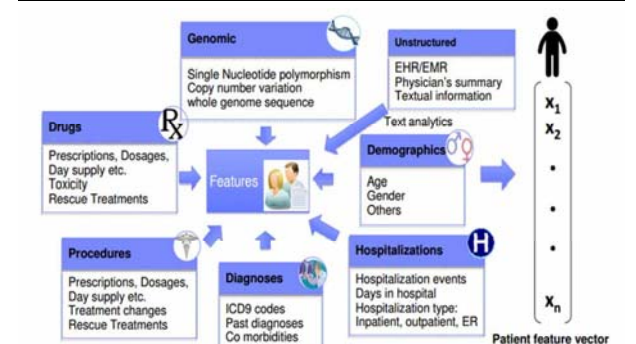


Image credit to Michal Rosen-Zvi

human-centered.ai (Holzinger Group)

21

2019 Machine Learning for Health 02

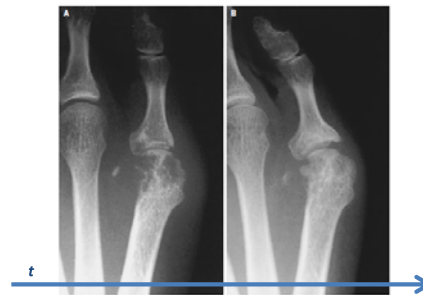


Chao, J., Parker, B. A. & Zvaifler, N. J. (2009) Accelerated Cutaneous Nodulosis Associated with Aromatase Inhibitor Therapy in a Patient with Rheumatoid Arthritis. *The Journal of Rheumatology*, 36, 5, 1087-1088.

human-centered.ai (Holzinger Group)

22

2019 Machine Learning for Health 02



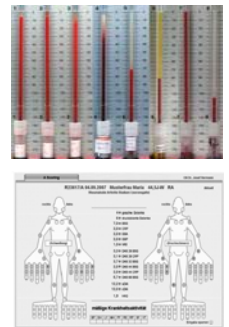
Ikari, K. & Momohara, S. (2005) Bone Changes in Rheumatoid Arthritis. *New England Journal of Medicine*, 353, 15, e13.

human-centered.ai (Holzinger Group)

23

2019 Machine Learning for Health 02

- 50+ Patients per day ~ 5000 data points per day ...
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ time-series data

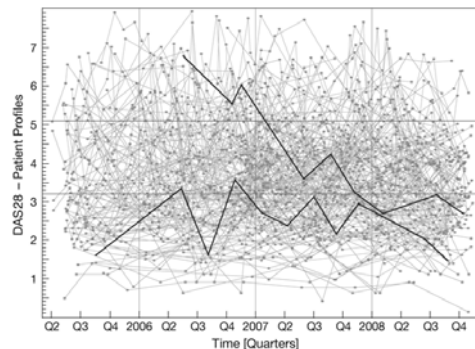


Simonik, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

human-centered.ai (Holzinger Group)

24

2019 Machine Learning for Health 02



Simonik, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

human-centered.ai (Holzinger Group)

25

2019 Machine Learning for Health 02

## Can Computers help doctors to make better decisions?

For reading and discussion: Michael Duerr-Specht, Randy Goebel & Andreas Holzinger 2015. *Medicine and Health Care as a Data Problem: Will Computers become better medical doctors?* In: Holzinger, Andreas, Roecker, Carsten & Ziefle, Martina (eds.) *Smart Health, State-of-the-Art SOTA Lecture Notes in Computer Science LNCS 8700*. Heidelberg, Berlin, New York: Springer, pp. 21-40, doi:10.1007/978-3-319-16226-3\_2.

human-centered.ai (Holzinger Group)

26

2019 Machine Learning for Health 02

Reasoning Process	Human	Computer
<b>Abductive</b> Hypothesis generation	Uniquely capable of complex pattern recognition and creative thought. "the whole is greater than the sum of its parts"	Matches multiple individual correlations from extensive data banks based on preconceived algorithms. Secondary construction of relationships. "the whole equals the sum of its parts"
<b>Inductive</b> Symptom → Disease	Limited database. Subject to biases <ul style="list-style-type: none"> <li>Anchoring bias</li> <li>Confirmation bias</li> <li>Premature closure</li> </ul>	Extensive database. Probability based on Bayesian statistics, no significant bias. Limitation based on available data.
<b>Deductive</b> Disease → Symptoms, Treatment	Limited database. Personal intuition and experience affect decision making.	Extensive database. Application of rules of evidence based medicine with potential biases.

Michael Duerr-Specht, Randy Goebel & Andreas Holzinger 2015. *Medicine and Health Care as a Data Problem: Will Computers become better medical doctors?* In: Holzinger, Andreas, Roecker, Carsten & Ziefle, Martina (eds.) *Smart Health, State-of-the-Art SOTA Lecture Notes in Computer Science LNCS 8700*. Heidelberg, Berlin, New York: Springer, pp. 21-40, doi:10.1007/978-3-319-16226-3\_2.

human-centered.ai (Holzinger Group)

27

2019 Machine Learning for Health 02



monday afternoon  
december 9  
3:45 p.m. / arena  
Chairman  
**DR. D. C. ENGELBART**  
Stanford Research Institute  
Menlo Park, California

a research center  
for augmenting human  
intellect

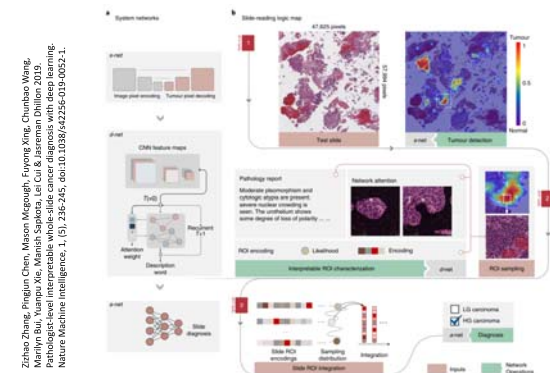
This session is entirely devoted to a presentation by Dr. Engelbart on a computer-based, interactive, multisensory display system which is being developed at Stanford Research Institute under the sponsorship of ARPA, NASA and RADG. The system is being used as an experimental laboratory for investigating principles by which interactive computer aids can augment intellectual capability. The techniques which are being described will, themselves, be used to augment the presentation.

The session will use an on-line, closed circuit television hook-up to the SRI computing system in Menlo Park. Following the presentation remote terminals to the system, in operation, may be viewed during the remainder of the conference in a special room set aside for that purpose.

Source: <https://web.stanford.edu/dept/SUI/library/extra4/sloan/mousesite/dce1968conferenceannouncement.jpg>



Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihls & Kurt Zatloukal 2017. Towards the Augmented Pathologist: Challenges of Explainable AI in Digital Pathology. arXiv:1712.06657.



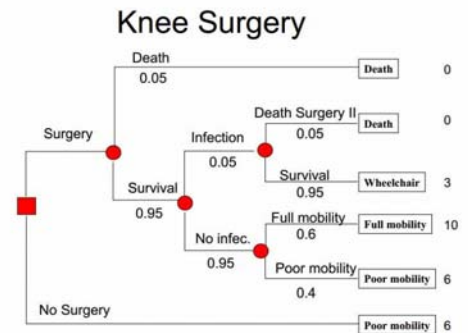
- **Type 1 Decisions:** related to the **diagnosis**, i.e. AI/ML is used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, ...)?
  - What is the probability that this patient has acute appendices, given the signs and symptoms concerning abdominal pain?
- **Type 2 Decisions:** related to **therapy**, i.e. AI/ML is used to select the best therapy on the basis of clinical evidence, e.g.:
  - What is the best therapy for patients of age x and risks y, if an obstruction of more than z % is seen in the left coronary artery?
  - What amount of insulin should be prescribed for a patient during the next 5 days, given the blood sugar levels and the amount of insulin taken during the recent weeks?

Jan H. Van Bemmel & Mark A. Musen 1997. Handbook of Medical Informatics, Heidelberg, Springer.

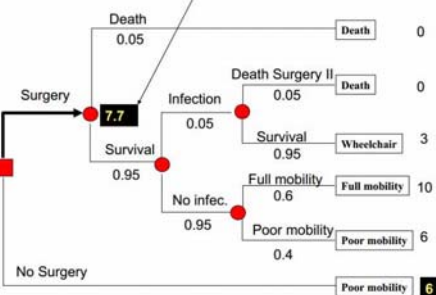


- Example of a Decision Problem
- Soccer player considering knee surgery
- Uncertainties:
- Success: recovering full mobility
- Risks: infection in surgery (if so, needs another surgery and may lose more mobility)
- Survival chances of surgery

Harvard-MIT Division of Health Sciences and Technology  
HST.951J: Medical Decision Support, Fall 2005  
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo



### Expected Value of Surgery



For a single decision variable an agent can select  $D = d$  for any  $d \in \text{dom}(D)$ .

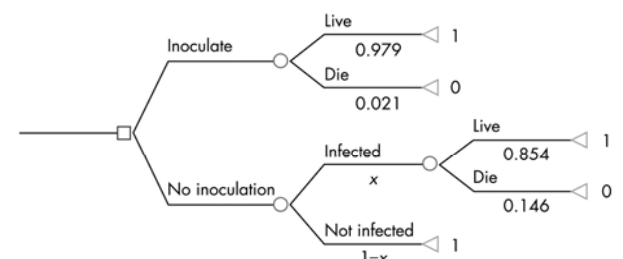
The expected utility of decision  $D = d$  is

$$E(U | d) = \sum_{x_1, \dots, x_n} P(x_1, \dots, x_n | d) U(x_1, \dots, x_n, d)$$

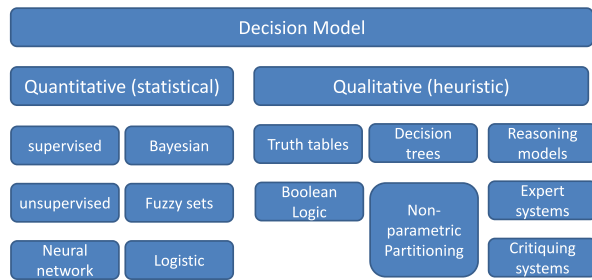
An optimal single decision is the decision  $D = d_{\max}$  whose expected utility is maximal:

$$d_{\max} = \arg \max_{d \in \text{dom}(D)} E(U | d)$$

John Von Neumann & Oskar Morgenstern 1944. *Theory of games and economic behavior*, Princeton, Princeton university press.

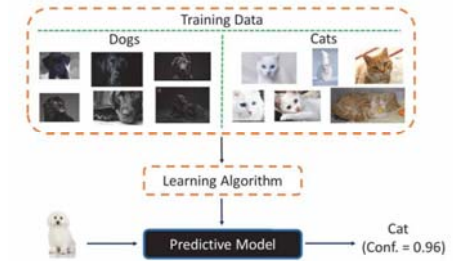


Ferrando, A., Pagano, E., Scaglione, L., Petrinco, M., Gregori, D. & Ciccone, G. (2009) A decision-tree model to estimate the impact on cost-effectiveness of a venous thromboembolism prophylaxis guideline. *Quality and Safety in Health Care*, 18, 4, 309-313.

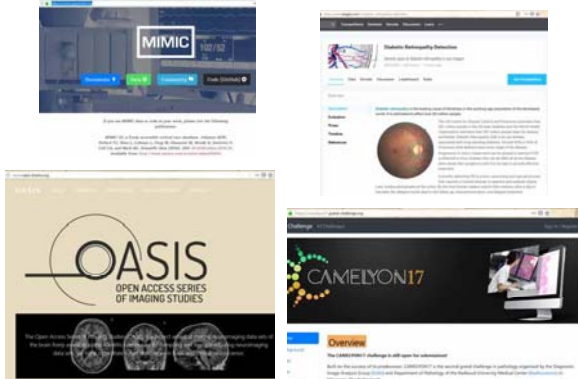


Extended by A. Holzinger after: Bemmle, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

- Need for robust algorithms
- Need for trustworthy, fair and accountable algorithms
- Augmenting the doctor – not replacing them, but let “Chimpanzee”-Work do by algorithms
- Focus of the doctors to cognitively high-end demanding, challenging work
- Double-Check (“look at this corner, maybe there is something relevant”)
- Many of the questions of medical doctors need causal explanations “the why” !!

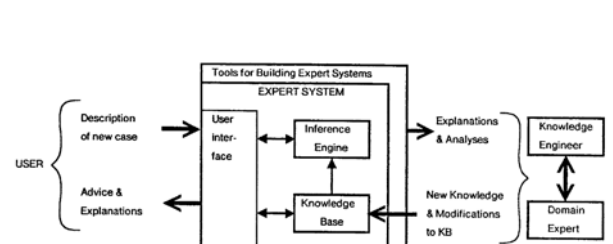
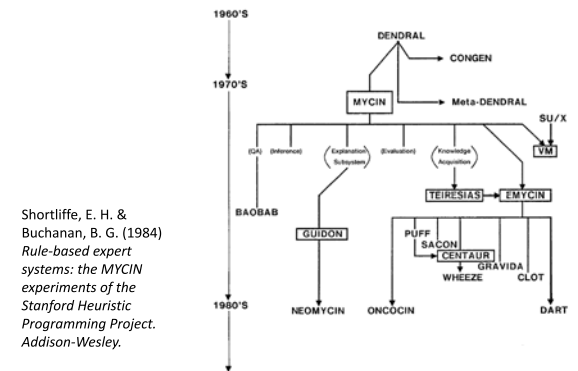


Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. Thirty-First AAAI Conference on Artificial Intelligence, 2017.

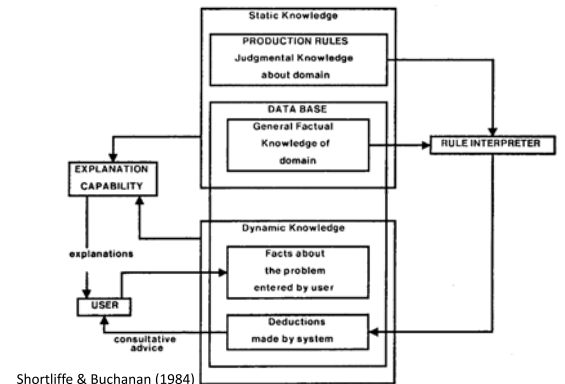


## 02 History of DSS = History of AI

- **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 ...



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



Shortliffe & Buchanan (1984)

- 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.



Image credit to Bernhard Schölkopf

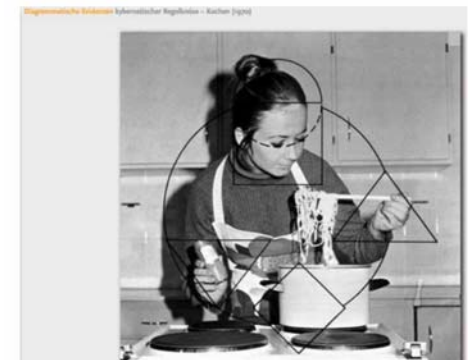
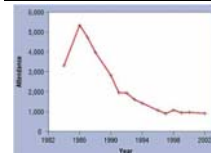


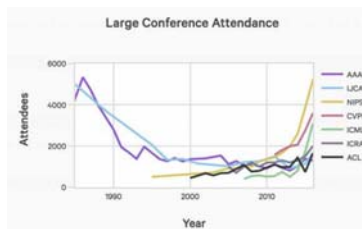
Image credit to Bernhard Schölkopf



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

AAAI = AAAI Conference on Artificial Intelligence:  
<https://aaai.org/Conferences/AAAI-20/>

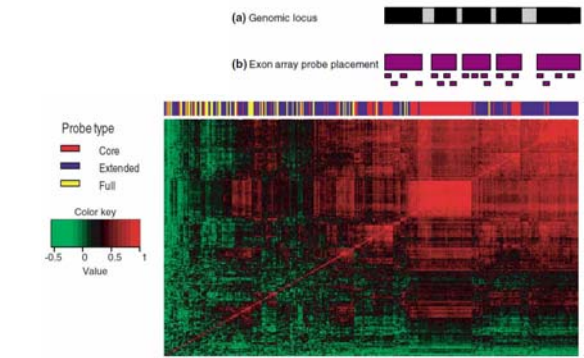
International Joint Conference on Artificial Intelligence:  
<https://ijcai20.org/>



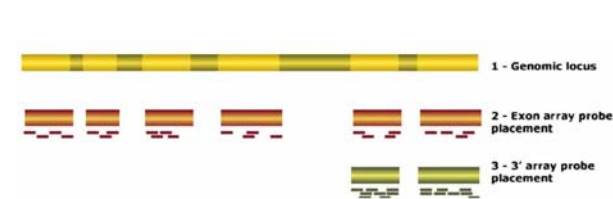
<https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0>

## 03 Example: P4-Medicine

James Hendler 2008. Avoiding another AI winter. *IEEE Intelligent Systems*, 23, (2), 2-4, doi:10.1109/MIS.2008.20.

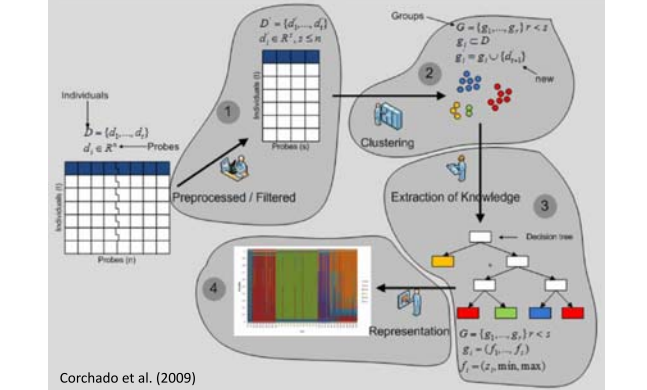


Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biology*, 8, 5, R82.



Exon array structure. Probe design of exon arrays. (1) Exon-intron structure of a gene. Gray boxes represent introns, rest represent exons. Introns are not drawn to scale. (2) Probe design of exon arrays. Four probes target each putative exon. (3) Probe design of 30expression arrays. Probe target the 30end of mRNA sequence.

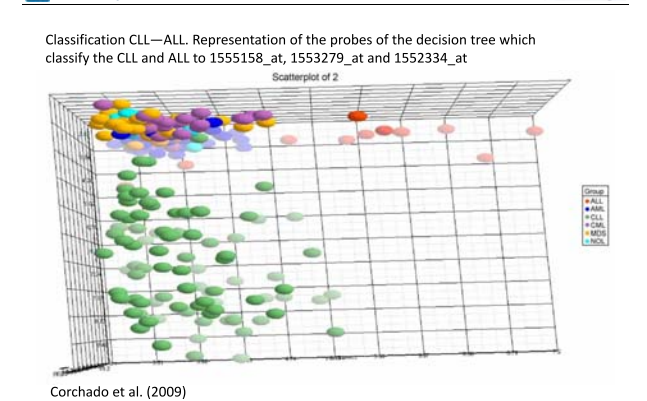
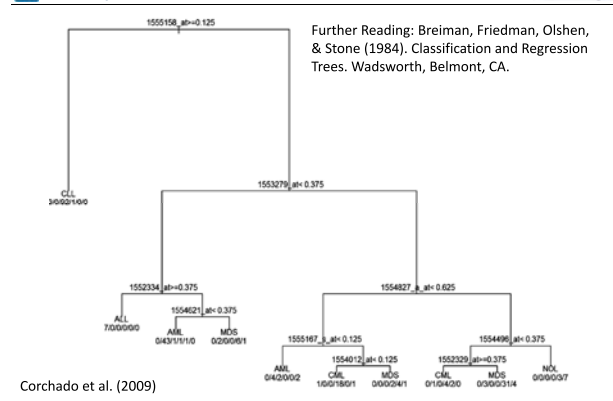
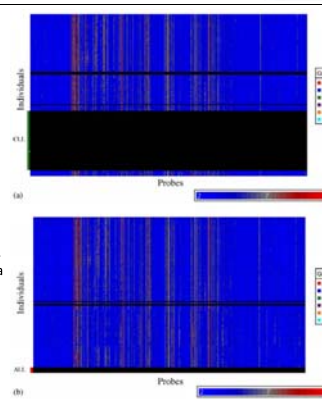
Corchado, J. M., De Paz, J. F., Rodriguez, S. & Bajo, J. (2009) Model of experts for decision support in the diagnosis of leukemia patients. *Artificial Intelligence in Medicine*, 46, 3, 179-200.



A = acute, C = chronic,  
L = lymphocytic, M = myeloid

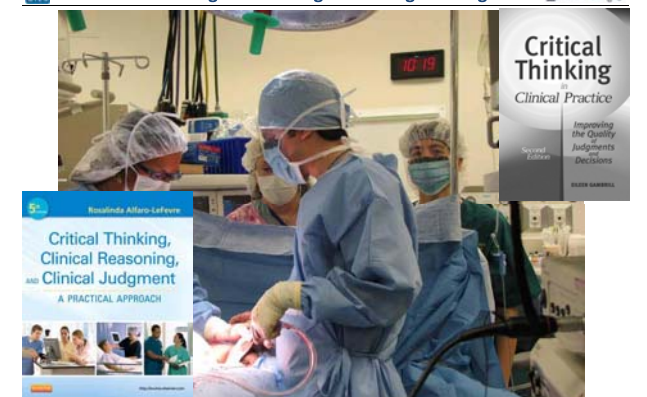
- ALL = cancer of the blood AND bone marrow caused by an abnormal proliferation of lymphocytes.
- AML = cancer in the bone marrow characterized by the proliferation of myeloblasts, red blood cells or abnormal platelets.
- CLL = cancer characterized by a proliferation of lymphocytes in the bone marrow.
- CML = caused by a proliferation of white blood cells in the bone marrow.
- MDS (Myelodysplastic Syndromes) = a group of diseases of the blood and bone marrow in which the bone marrow does not produce a sufficient amount of healthy cells.
- NOL (Normal) = No leukemias

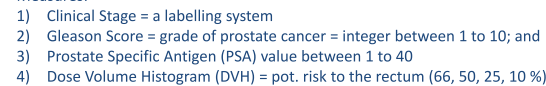
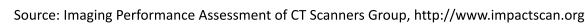
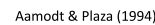
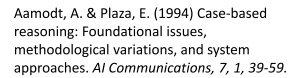
Corchado et al. (2009)



- The model of Corchado et al. (2009) combines:
  - 1) methods to **reduce the dimensionality** of the original data set;
  - 2) pre-processing and data filtering techniques;
  - 3) a clustering method to classify patients; and
  - 4) extraction of knowledge techniques
- The system reflects how human experts work in a lab, but
  - 1) **reduces the time** for making predictions;
  - 2) **reduces the rate of human error**; and
  - 3) **works with high-dimensional data** from exon arrays

# 04 Example: Case Based Reasoning (CBR)





2019 Machine Learning for Health 02



The graph illustrates the fuzzy membership functions for Gleason Score. The x-axis represents the Gleason Score from 0 to 13, and the y-axis represents the Membership value from 0 to 1.4. Three fuzzy sets are defined: 'Low', 'Medium', and 'High'. The 'Low' set has a membership of 1.0 for scores 0-4 and then decreases. The 'Medium' set increases from score 3 to a peak of 1.0 at score 7 and then decreases. The 'High' set increases from score 6 to a membership of 1.0 at score 10 and remains constant thereafter. An inset histology image shows prostate tissue.

2019 Machine Learning for Health 02

2019 Machine Learning for Health 02

## 2019 Machine Learning for Health 02

## ■ “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.

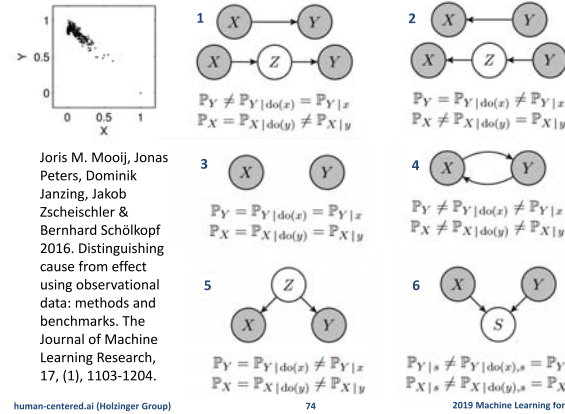
Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

## ■ Important Definition: Ground truth

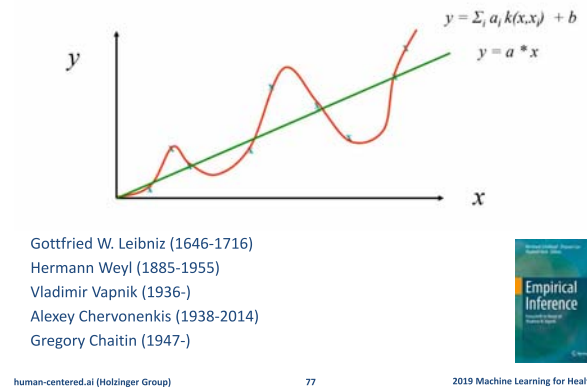
- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
- Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
- Empirical inference =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

## 06 Explainability

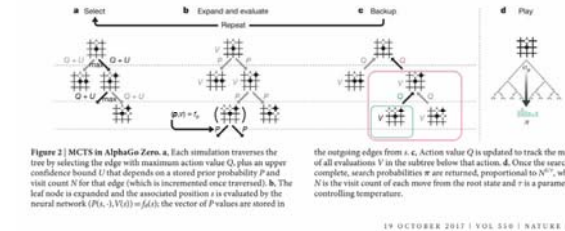
## Decide if $X \rightarrow Y$ , or $Y \rightarrow X$ using only observed data



## Empirical Inference Example



## Mastering the game of Go without human knowledge



$$(p, v) = f_\theta(s) \text{ and } l = (z - v)^2 - \pi^\top \log p + c \|\theta\|^2$$

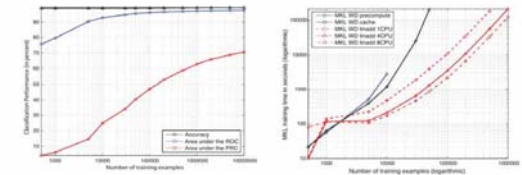
David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel & Demis Hassabis 2017. Mastering the game of go without human knowledge. Nature, 550, (7676), 354-359, doi:10.1038/nature24270.

## 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.

## 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.

## Why did it make this decision???



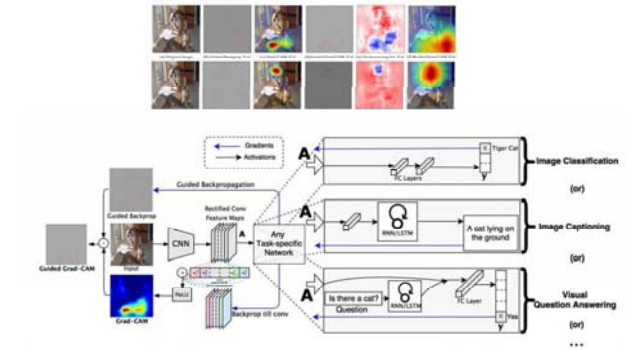
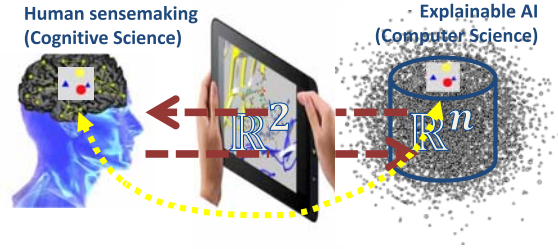
David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis 2016. Mastering the game of Go with deep neural networks and tree search. Nature, 529, (7587), 484-489, doi:10.1038/nature16961.



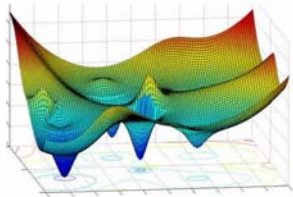
**Explainability :=**  
a property of a system  
("the AI explanation)  
**Causability :=**  
a property of a person  
("the Human explanation)

Andreas Holzinger et al. 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

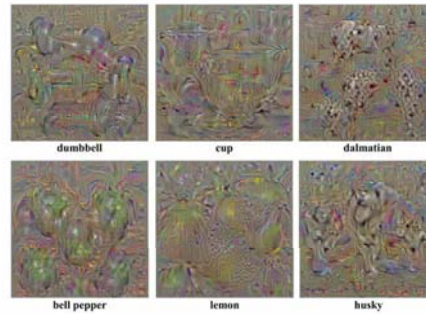
- Causability := a property of a person (Human)
- Explainability := a property of a system (Computer)



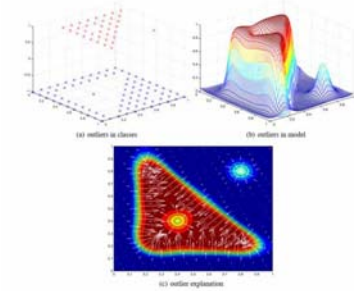
Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. ICCV, 2017. 618-626.



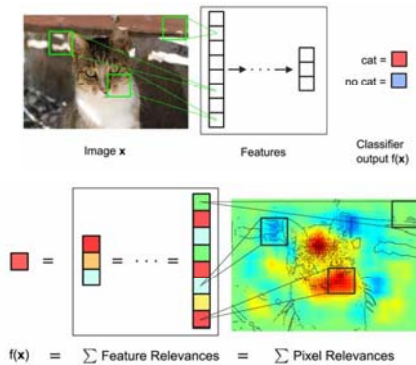
<https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient/a/the-gradient>



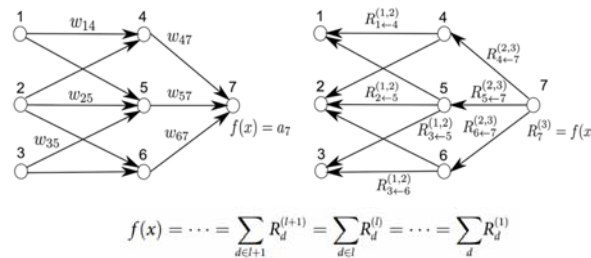
Karen Simonyan, Andrea Vedaldi & Andrew Zisserman 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.



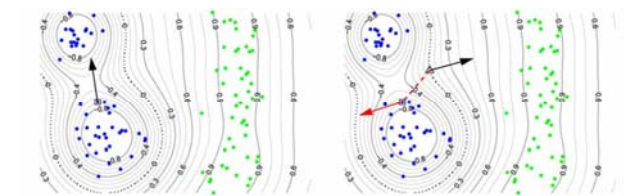
David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen & Klaus-Robert Mueller 2010. How to explain individual classification decisions. Journal of machine learning research (JMLR), 11, (6), 1803-1831.



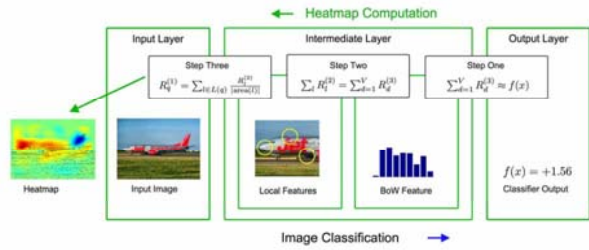
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS one, 10, (7), e0130140. doi:10.1371/journal.pone.0130140.



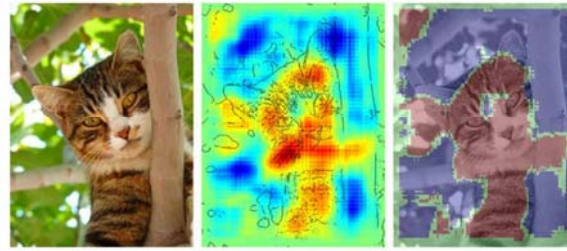
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS one, 10, (7), e0130140. doi:10.1371/journal.pone.0130140.



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.



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.



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.

**Definition 1.** A heatmapping  $R(x)$  is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

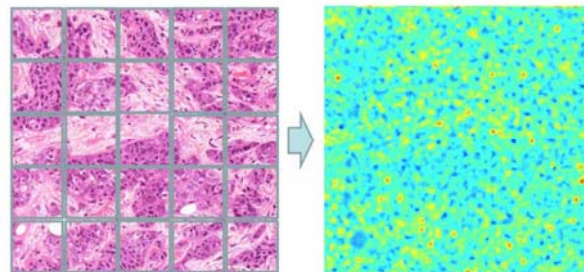
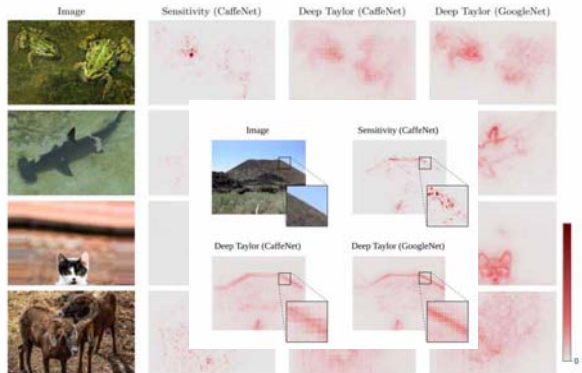
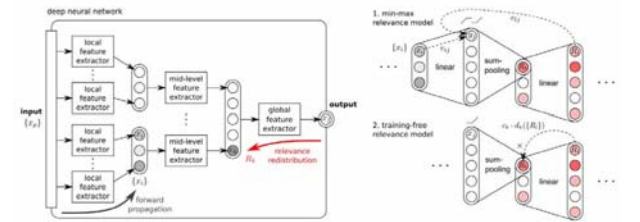
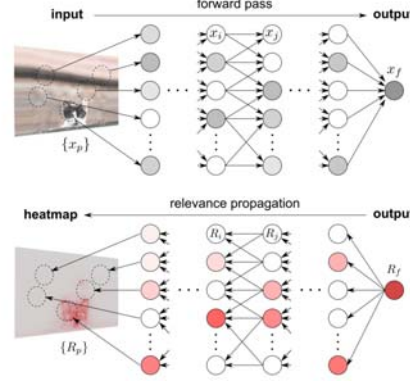
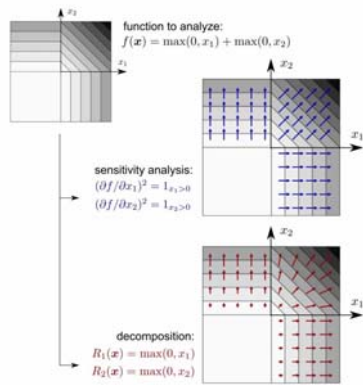
$$\forall x: f(x) = \sum_p R_p(x).$$

**Definition 2.** A heatmapping  $R(x)$  is *positive* if all values forming the heatmap are greater or equal to zero, that is:

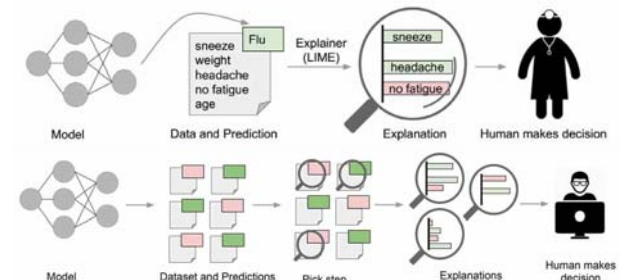
$$\forall x, p: R_p(x) \geq 0$$

**Definition 3.** A heatmapping  $R(x)$  is *consistent* if it is conservative and positive. That is, it is consistent if it complies with Definitions 1 and 2.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.

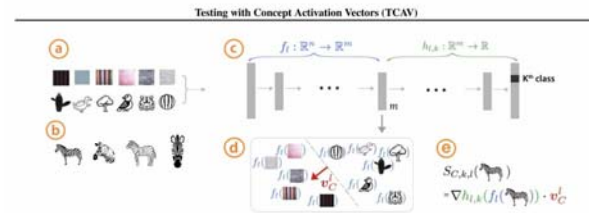
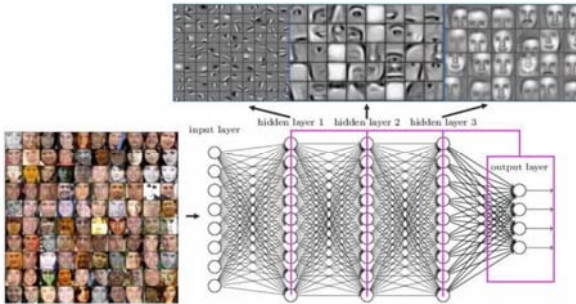
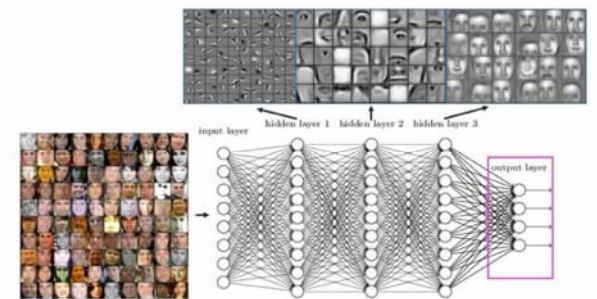
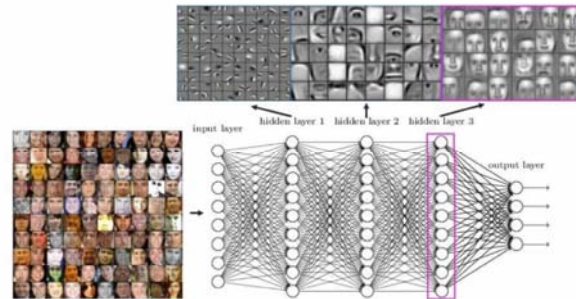
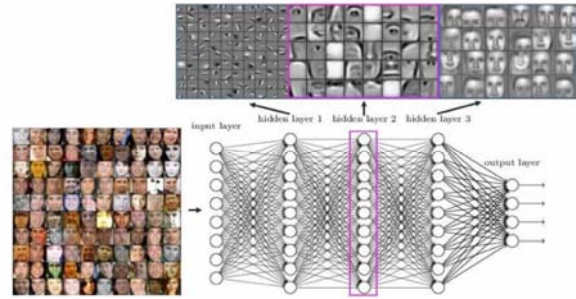


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

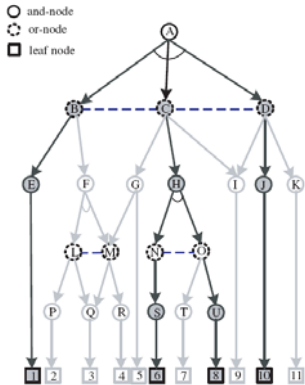




**Figure 1. Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., "striped"), and random examples (a), labeled training-data examples for the studied class (zebras) (b), and a trained network (c), TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer (d). The CAV is the vector orthogonal to the classification boundary ( $v_C$ , red arrow). For the class of interest (zebras), TCAV uses the directional derivative  $S_{C,k,l}(x)$  to quantify conceptual sensitivity (e).

<https://github.com/tensorflow/tcav>

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viégas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.



### Algorithm for this framework

- Top-down/bottom-up computation

### Generalization of small sample

- Use Monte Carlos simulation to synthesis more configurations

### Fill semantic gap

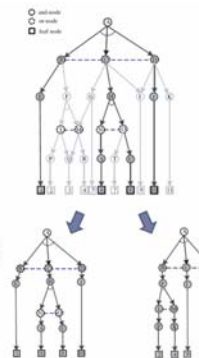
Images credit to Zhaoyin Jia (2009)

- Terminal (leaf) node:  $T(pg)$
- And-Or node:  $V^{or}(pg), V^{and}(pg)$
- Set of links:  $E(pg)$
- Switch variable at Or-node:  $w(t)$
- Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^{and}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{or}(pg) \rightarrow T(pg)} \lambda_v(\alpha(t))$$

$$+ \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$



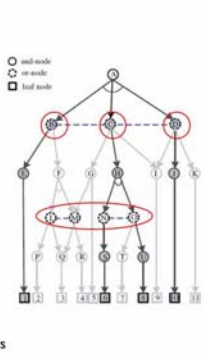
- Terminal (leaf) node:  $T(pg)$
- And-Or node:  $V^{or}(pg), V^{and}(pg)$
- Set of links:  $E(pg)$
- Switch variable at Or-node:  $w(t)$
- Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^{and}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{or}(pg) \rightarrow T(pg)} \lambda_v(\alpha(t))$$

$$+ \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$

SCFG: weigh the frequency at the children of or-nodes

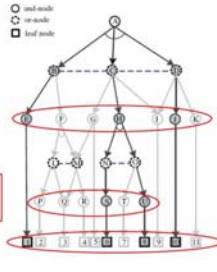


- Terminal (leaf) node:  $T(pg)$
- And-Or node:  $V^{or}(pg), V^{and}(pg)$
- Set of links:  $E(pg)$
- Switch variable at Or-node:  $w(t)$
- Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \sim T(pg)} \lambda_t(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$

Weigh the local compatibility of primitives (geometric and appearance)

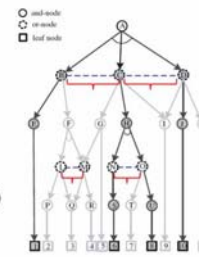


- Terminal (leaf) node:  $T(pg)$
- And-Or node:  $V^{or}(pg), V^{and}(pg)$
- Set of links:  $E(pg)$
- Switch variable at Or-node:  $w(t)$
- Attributes of primitives:  $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \sim T(pg)} \lambda_t(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$

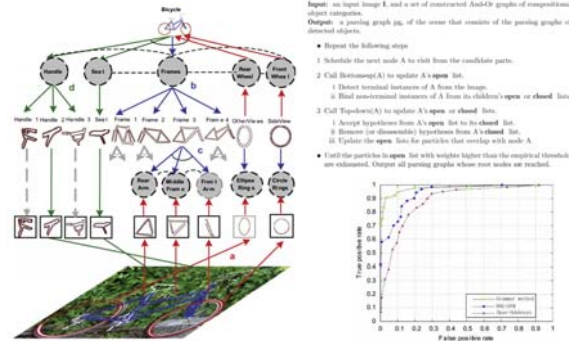
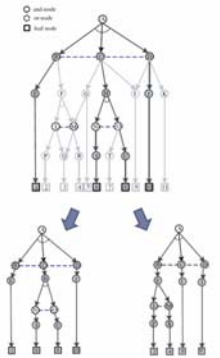
Spatial and appearance between primitives (parts or objects)



- Terminal (leaf) node:  $T(pg)$
- And-Or node:  $V^{or}(pg), V^{and}(pg)$
- Set of links:  $E(pg)$
- Switch variable at Or-node:  $w(t)$
- Attributes of primitives:  $\alpha(t)$

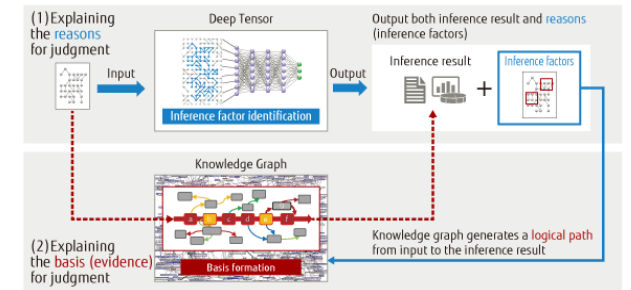
$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \sim T(pg)} \lambda_t(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$



Liang Lin, Tianfu Wu, Jake Porway & Zijian Xu 2009. A stochastic graph grammar for compositional object representation and recognition. Pattern Recognition, 42, (7), 1297-1307. doi:10.1016/j.patcog.2008.10.033.

## Future Work



Explainable AI with Deep Tensor and Knowledge Graph

[http://www.fujitsu.com/jp/images/artificial-intelligence-en\\_tcm102-3781779.png](http://www.fujitsu.com/jp/images/artificial-intelligence-en_tcm102-3781779.png)

- What is a good explanation?
- (obviously if the other did understand it)
- Experiments needed!
- What is explainable/understandable/intelligible?
- When is it enough (Sättigungsgrad – you don't need more explanations – enough is enough)
- But how much is it ...

- Justification, Explanation and Causality
- Trust > scaffolded by justification of actions (why)
- Please always take into account the inherent uncertainty and incompleteness of medical data!

Alex John London 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report, 49, (1), 15-21, doi:10.1002/hast.973.

**Teaching Meaningful Explanations**

Noel C. F. Codella,\* Michael Hind,\* Karthikeyan Natesan Ramamurthy,\* Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei & Aleksandra Mojsilovic 2018. Teaching Meaningful Explanations. arXiv:1805.11648v1 [cs.AI] 29 May 2018

IBM Research  
Yorktown Heights, NY 10598  
(ncodell1,hind,mnatesa,scan,adhuram,kvrvarsh,dwei,alekmojs@us.ibm.com)

**Abstract**

The adoption of machine learning in high-stakes applications such as healthcare and law has lagged in part because predictions are not accompanied by explanations comprehensible to the domain user, who often holds ultimate responsibility for decisions and outcomes. In this paper, we propose an approach to generate such explanations in which training data is augmented to include, in addition to features and labels, explanations elicited from domain users. A joint model is then learned to produce both labels and explanations from the input features. This simple idea ensures that explanations are tailored to the complexity expectations and domain knowledge of the consumer. Evaluation spans multiple modeling techniques on a simple game dataset, an image dataset, and a chemical color dataset, showing that our approach is generalizable across domains and algorithms. Results demonstrate that meaningful explanations can be reliably taught to machine learning algorithms, and in some cases, improve modeling accuracy.

**1 Introduction**

New regulations call for automated decision making systems to provide “meaningful information” on the logic used to reach conclusions [1–3]. Selbst and Powles interpret the concept of “meaningful information” as information that should be understandable to the audience (potentially individuals



# Conclusion

This image is in the Public Domain

human-centered.ai (Holzinger Group)

136

2019 Machine Learning for Health 02

- 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 ...

human-centered.ai (Holzinger Group)

137

2019 Machine Learning for Health 02

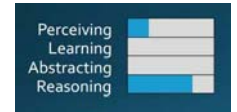


Image credit to John Launchbury

- 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  $p(x)$

human-centered.ai (Holzinger Group)

138

2019 Machine Learning for Health 02



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

human-centered.ai (Holzinger Group)

139

2019 Machine Learning for Health 02



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 ...

human-centered.ai (Holzinger Group)

140

2019 Machine Learning for Health 02

- 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 (Holzinger Group)

141

2019 Machine Learning for Health 02



# Thank you!