185.A83 Machine Learning for Health Informatics
2020S, VU, 2.0 h, 3.0 ECTS

Andreas Holzinger, Rudolf Freund
Marcus Bloice, Florian Endel, Anna Saranti

Lecture 01 – Introduction

From health informatics to ethical responsible medical AI

Contact: andreas.holzinger AT tuwien.ac.at
@aholzin #KandinskyPatterns

https://human-centered.ai/lv-185-a83-machine-learning-for-health-informatics-class-of-2021

Remarks

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

For deeper learning effect, please always try to answer the question in the title bar first.

If you need a refresher on the basics of probability and information theory, please refer to slide deck 00.

This course follows a Research-Based Teaching Style

At the end of this lecture you will …

- … see why an integrative approach is important
- … understand that machine learning can help medicine
- … be fascinated by the possibilities of machine learning
- ... disillusioned at the limits of machine learning
- … be aware of the complexity of the health domain
- ... realize why a human-in-the-loop is sometimes crucial
- ... identify the main challenges in this area
- … recognize some ideas of future work

Advance Organizer (1/2)

- **Bioinformatics** = discipline, as part of biomedical informatics, at the interface between *bio*logy and *infor*mation science and mathema*tics*; processing of biological data;
- **Biomarker** = a characteristic (e.g. body-temperature (fever) as a biomarker for an infection, or proteins measured in the urine) as an indicator for normal or pathogenic biological processes, or pharmacologic responses to a therapeutic intervention;
- **Biomedical data** = compared with general data, it is characterized by large volumes, complex structures, high dimensionality, evolving biological concepts, and insufficient data modeling practices;
- **Biomedical Informatics** = 2011-definition: similar to medical informatics but including the optimal use of biomedical data, e.g. from genomics, proteomics, metabolomics;
- **Classical Medicine** = is both the science and the art of healing and encompasses a variety of practices to maintain and restore health;
- **Genomics** = branch of molecular biology which is concerned with the structure, function, mapping & evolution of genomes;
- **interactive Machine Learning** = defined as algorithms that can interact with both computational agents and human agents and can optimize their learning behaviour through these interactions, by bringing in a human-into-the-loop
- **Machine Learning** = addresses the question of how to design algorithms that improve automatically through experience from big data - doing it automatically (aML) without a human-in-the-loop
- **Medical Informatics** = 1970-definition: "… scientific field that deals with the storage, retrieval, and optimal use of medical information, data, and knowledge for problem solving and decision making"; - see the better 2011-definition by the AMIA
- **Metabolomics** = study of chemical processes involving metabolites (e.g. enzymes). A challenge is to integrate proteomic, transcriptomic, and metabolomic information to provide a more complete understanding of living organisms;
- **Molecular Medicine** = emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs;

Advance Organizer (2/2)

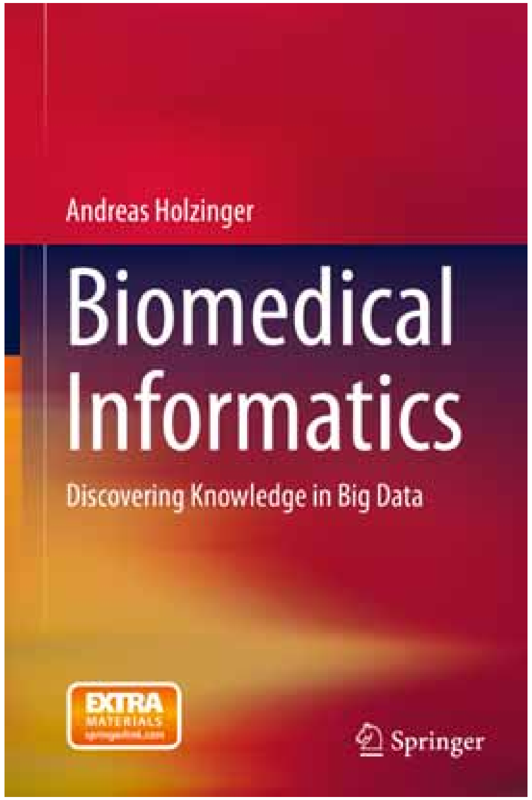
- **Omics data** = data from e.g. genomics, proteomics, metabolomics, etc.
- **Pervasive Computing** = similar to ubiquitous computing (Ubicomp), a post-desktop model of Human-Computer Interaction (HCI) in which information processing is integrated into every-day, miniaturized and embedded objects and activities; having some degree of “intelligence”;
- **Pervasive Health** = all unobtrusive, analytical, diagnostic, supportive etc. information functions to improve health care, e.g. remote, automated patient monitoring, diagnosis, home care, self-care, independent living, etc.;
- **Proteome** = the entire complement of proteins that is expressed by a cell, tissue, or organism;
- **Proteomics** = field of molecular biology concerned with determining the proteome;
- **P-Health Model** = Preventive, Participatory, Pre-emptive, Personalized, Predictive, Pervasive (= available to anybody, anytime, anywhere);
- **Space** = a set with some added structure;
- **Technological Performance** = machine “capabilities”, e.g. short response time, high throughput, high availability, etc.
- **Time** = a dimension in which events can be ordered along a time line from the past through the present into the future;
- **Translational Medicine** = based on interventional epidemiology; progress of Evidence-Based Medicine (EBM), integrates research from basic science for patient care and prevention;
- **Von-Neumann-Computer** = a 1945 architecture, which still is the predominant machine architecture of today (opp.: Non-Vons, incl. analogue, optical, quantum computers, cell processors, DNA and neural nets (in silico));

Agenda

- (0) A few definitions – for mutual understanding
- (1) Machine Learning Health examples
- (2) A brief look at the application area health
- (3) Statistical Machine Learning
- (4) Automatic Machine Learning (aML)
- (5) Interactive Machine Learning (iML)
  and why we need the human-in-the-loop
- (6) Explainable AI and Methods of Explainability
- (7) Causability – Measuring the Quality of (6)

(0) A few definitions first

- *Health* := state of physical, mental and social well-being
- *Medicine* := art, science, practice of patient diagnosis, prognosis, prevention, treatment of injury or disease.
- *Informatics* := study of information processing (Schrödinger [1]: Life is Information Processing)
- *Biomedical informatics* (BMI) := interdisciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific problem solving, and decision making, motivated by efforts to improve human health [2]

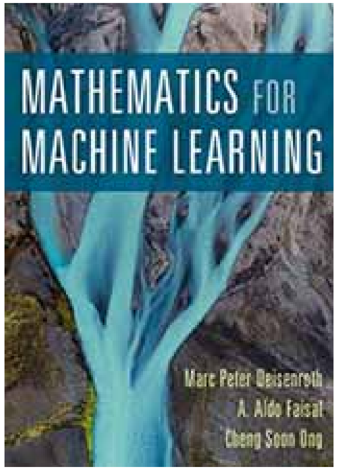


Andreas Holzinger (2014). Biomedical Informatics: Discovering Knowledge in Big Data, New York, Springer, doi:10.1007/978-3-319-04528-3

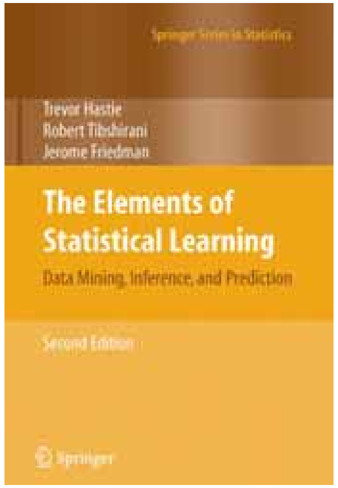
[1] Erwin Schrödinger (1944). What Is Life? The Physical Aspect of the Living Cell, Dublin, Dublin Institute for Advanced Studies at Trinity College.

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3534470 (American Association of Medical Informatics, AMIA)

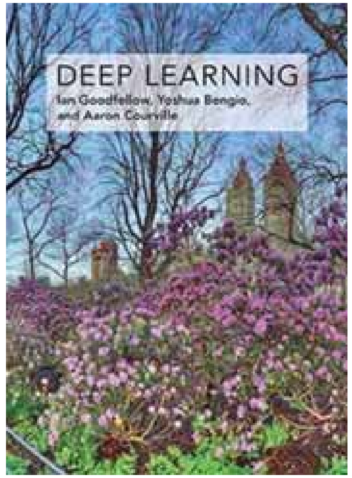
- *AI* := not clearly defined “umbrella term” of “making intelligent machines”
- *ML* := the workhorse of AI - algorithms that improve by learning from data
- *DL* := a particular family of neural networks - currently very successful



Marc Peter Deisenroth, A. Aldo Faisal & Cheng Soon Ong (2020). Mathematics for machine learning, Cambridge (UK), Cambridge University Press, doi:10.10179781108679930. https://mml-book.github.io



Trevor Hastie, Robert Tibshirani & Jerome Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, New York, Springer, doi:10.1007/978-0-387-84858-7 https://web.stanford.edu/~hastie/ElemStatLearn



Ian Goodfellow, Yoshua Bengio & Aaron Courville (2016). Deep Learning, Cambridge (MA), MIT Press. https://www.deeplearningbook.org

What was the first definition of Machine Learning ?

SOME STUDIES IN
MACHINE LEARNING USING
THE GAME OF CHECKERS

by A. L. Samuel

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3, (3), 210-229.

Field of Study that gives computers the ability to **learn** [from Data] **without explicitly being programmed …**

Introduction

The studies reported here have been co
a digital computer to behave in a way w
animals, would be described as involvin
this is not the place to dwell on the imp
cedures, or to discourse on the philosoph
very large amount of work, now done b
its demands on the intellect but does, ne
We have at our command computers wi
and with sufficient computational speed
techniques, but our knowledge of the ba
is still rudimentary. Lacking such know
methods of problem solution in minute a
and costly procedure. Programming com
should eventually eliminate the need for
ming effort.

Memoriam

AI Magazine Volume 11 Number 3 (1990) (© AAAI)

In Memoriam

Arthur Samuel: Pioneer in Machine Learning

Arthur Samuel (1901–1990) was a pioneer of artificial intelligence research. From 1949 through the late 1960s, he did the best work in making computers learn from their experience. His vehicle for this work was the game of checkers.

Programs for playing games often fill the role in artificial intelligence research that the fruit fly Drosophila

Samuel was a modest man, and the importance of his work was widely recognized only after his retirement from IBM in 1966, in part because he didn't relish the politics that were required to have his research more vigorously followed up on. He was also realistic about the large difference between what had been accomplished in understanding intellectual mechanisms and what would be required to reach human-level intelligence.

Samuel's papers on machine learn-

strate the power of electronic computers. He didn't finish the program while he was at the university of Illinois, perhaps because the computer wasn't finished in time.

In 1949, Samuel joined IBM's Poughkeepsie Laboratory, where he worked on IBM's first stored program

McCarthy, J. & Feigenbaum, E. A. 1990. In Memoriam: Arthur Samuel: Pioneer in Machine Learning. AI Magazine, 11, (3), 10.

What is Standardized Medicine versus Personalized Medicine ?

EBM CPG

Standardized Medicine

Person-alized Medicine

GBM GPM

Pervasive Health

Preventive Health Integration

EBM = Evidence Based Medicine
CPG = Clinical Practice Guideline
GBM = Genome Based Medicine
GPM = Genetic Polymorphism


nature International weekly journal of science

Journal content
- Journal home
- Advance online publication
- Current issue
- Nature News
- Archive
- Supplements
- Web focuses
- Podcasts
- Videos
- News Specials

Journal information
- About the journal
- For authors
- Online submission
- Nature Awards
- Nature history

NPG services
- Advertising
- work@npg
- Reprints & permissions
- For librarians
- Authors & referees

NPG resources
- Gateways & databases

Books and Arts

A reality check for personalized medicine

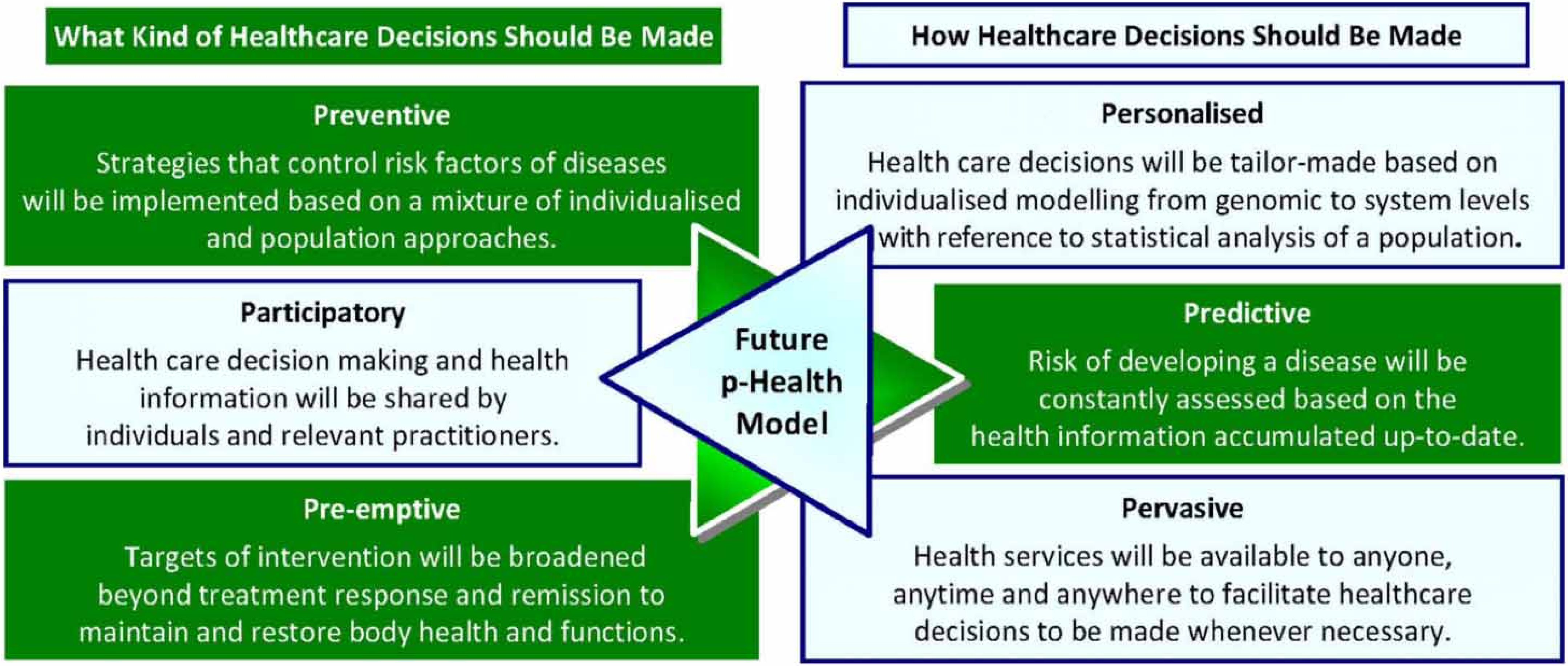
Muin J. Khoury, James Evans & Wylie Burke

Bringing genetic information into health care is welcome but its utility in the clinic needs to be rigorously reviewed, caution Muin J. Khoury, James Evans and Wylie Burke.

BOOK REVIEWED
Personal Genomics and Personalized Medicine

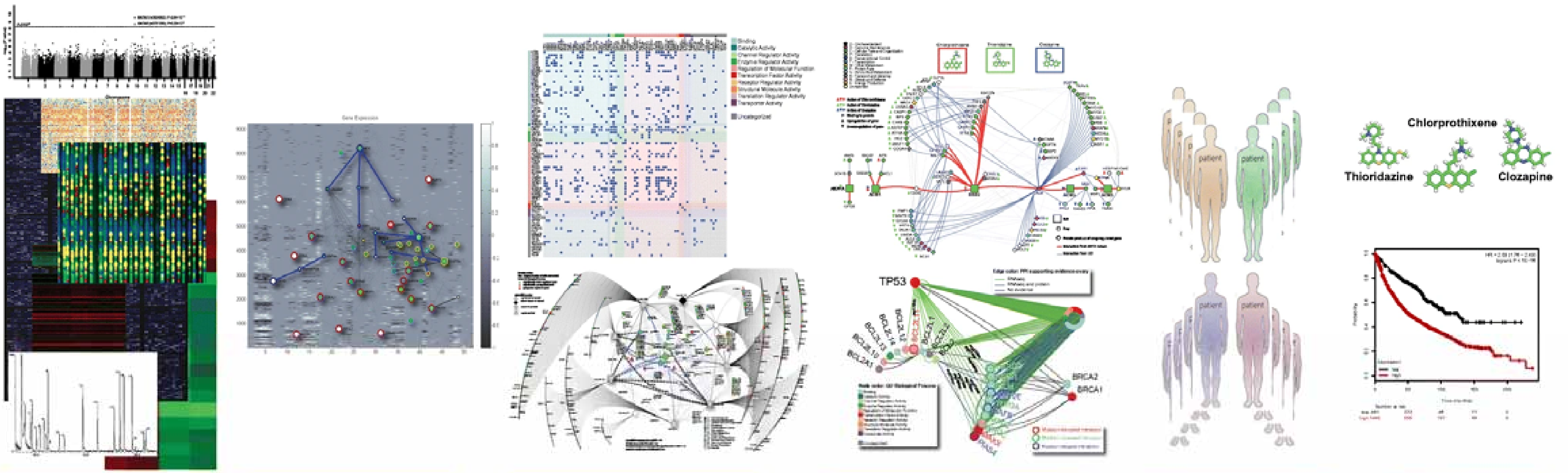
Rui Chen & Michael Snyder (2013). Promise of personalized omics to precision medicine. Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 5, (1), 73-82, doi:10.1002/wsbm.1198.

Preventive, Participatory, Pre-emptive, Pervasive, Predictive, Personalized ?



Yuan-Ting Zhang & Carmen C. Y. Poon (2010). Editorial Note on Bio, Medical, and Health Informatics. IEEE Transactions on Information Technology in Biomedicine, 14, (3), 543-545, doi:10.1109/TITB.2010.2049597

How is Stratified Medicine tailoring the treatment to patient subgroups ?



Omics profiles across genome, proteome, metabolome can be analyzed separately or combined to find differentially expressed entities

Network relationships link relevant entities within each data layer and identify better biomarkers

Layers of annotated networks; annotated with tissue, disease, network properties can further characterize potential biomarkers

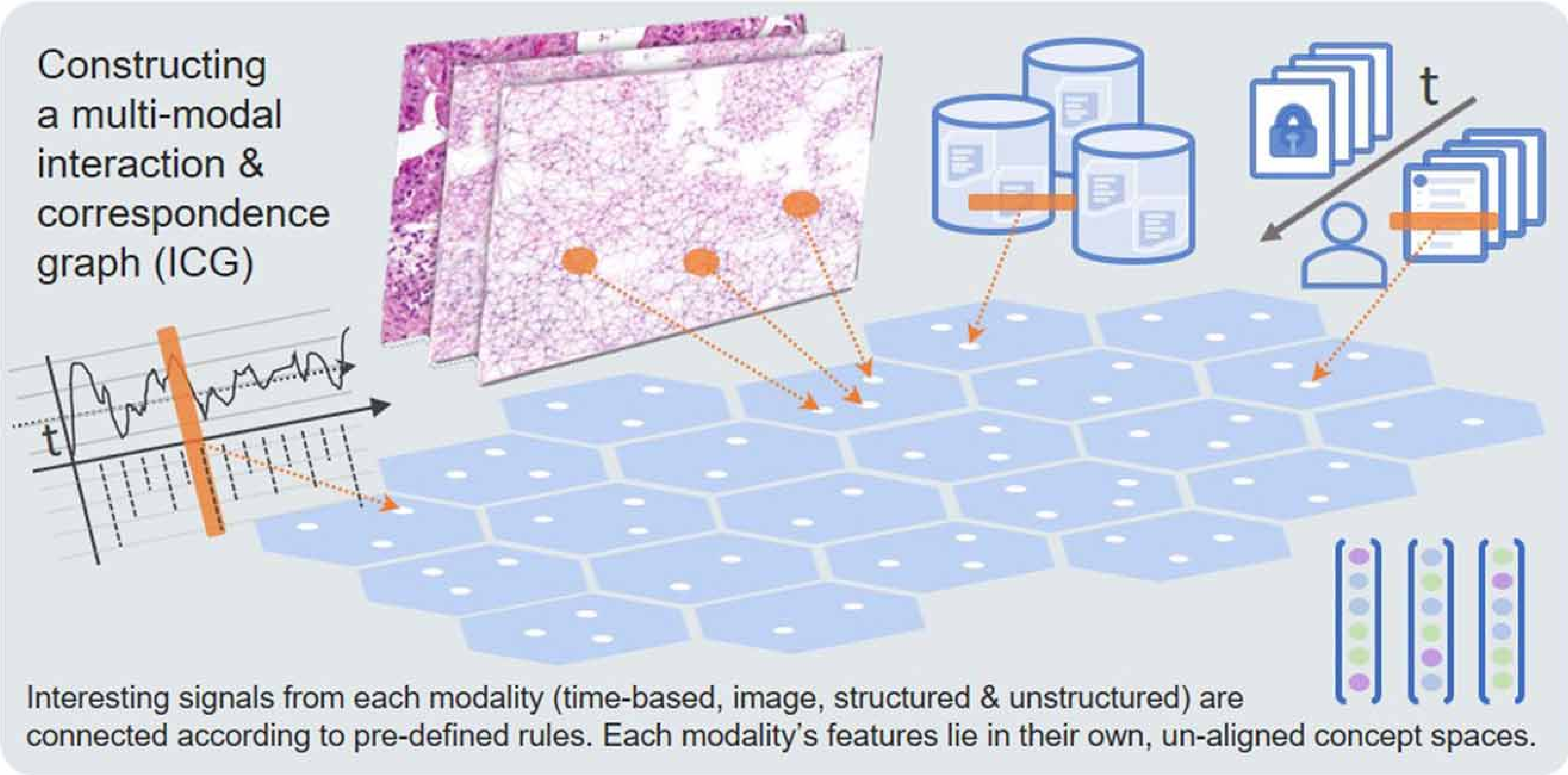
Discovered relationships across data layers identify combined biomarkers, drug mechanism of action and create explainable disease models

Combined biomarkers identify clinically-relevant patient subgroups

Treatment tailored to patient subgroups results in improved patient outcomes

Andreas Holzinger, Benjamin Haibe-Kains & Igor Jurisica (2019). Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. European Journal of Nuclear Medicine and Molecular Imaging, 46, (13), 2722-2730, doi:10.1007/s00259-019-04382-9

How can multi-modal representations enable joint learning ?



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

(1) Machine learning health examples

ARTIFICIAL INTELLIGENCE IN MEDICINE

Where Do We Stand?

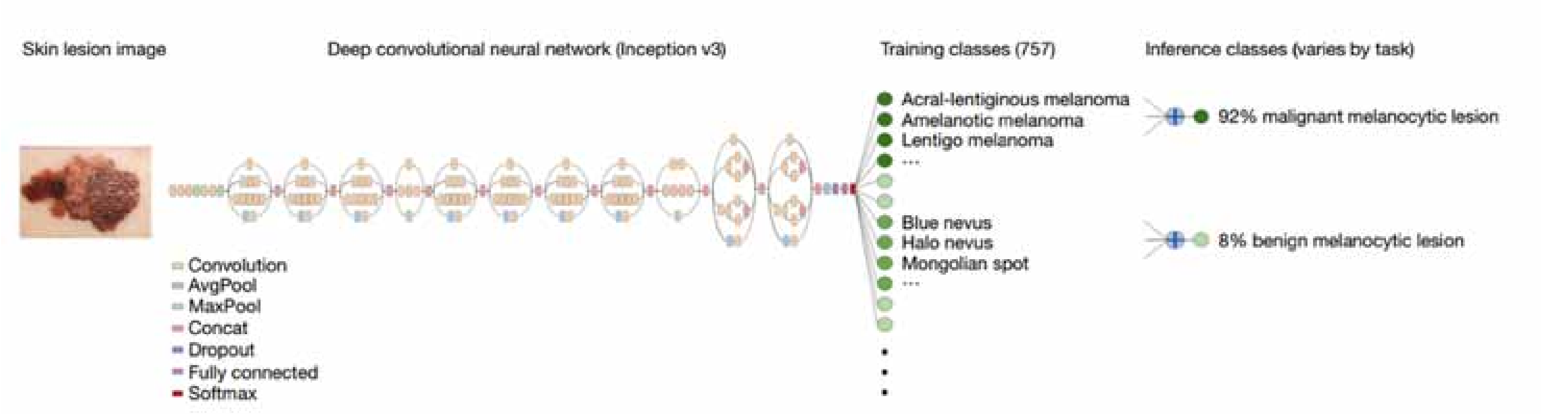
AFTER hearing for several decades that computers will soon be able to assist with difficult diagnoses, the practicing physician may well wonder why the revolution has not occurred. Skepticism at this point is understandable. Few, if any, programs currently have active roles as consultants to physicians. The story behind these unfulfilled expectations is instructive and, we believe, offers hope for the future.

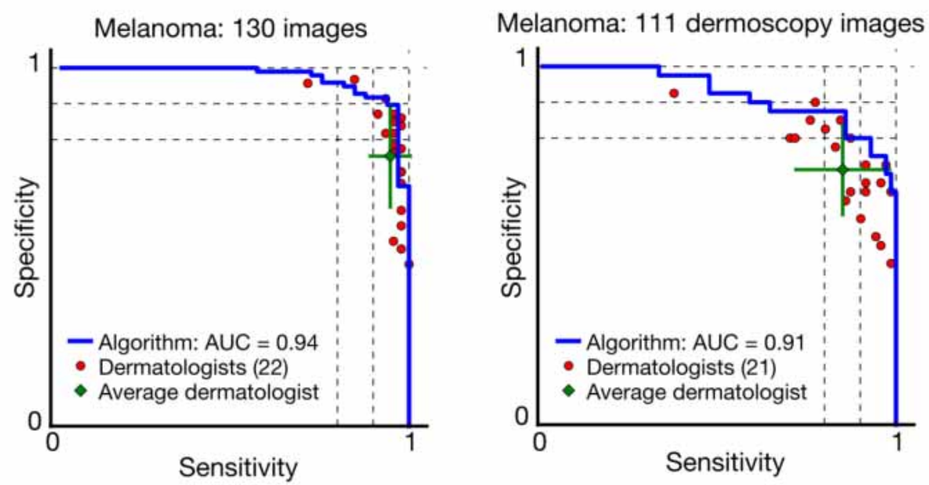
William B. Schwartz, Ramesh S. Patil & Peter Szolovits **(1987).** Artificial Intelligence in Medicine Where Do We Stand? New England Journal of Medicine, 316, (11), 685-688, doi:10.1056/NEJM198703123161109.

WHAT DOES THE FUTURE HOLD?

In 1970 an article in the *Journal* predicted that by the year 2000 computers would have an entirely new role in medicine, acting as a powerful extension of the physician's intellect.[33] At the halfway point, how realistic does this projection seem? It is now clear that great progress has been made in understanding how physicians solve difficult clinical problems and in implementing experimental programs that capture at least a portion of human expertise. On the other hand, it has become increasingly apparent that major intellectual and technical problems must be solved before we can produce truly reliable consulting programs. Nevertheless, assuming continued research, it still seems possible that by the year 2000 a range of programs will be available that can greatly assist the physician. It seems highly unlikely that such a goal will be achieved much before that time.

Example Dermatology: Towards Human-Level AI

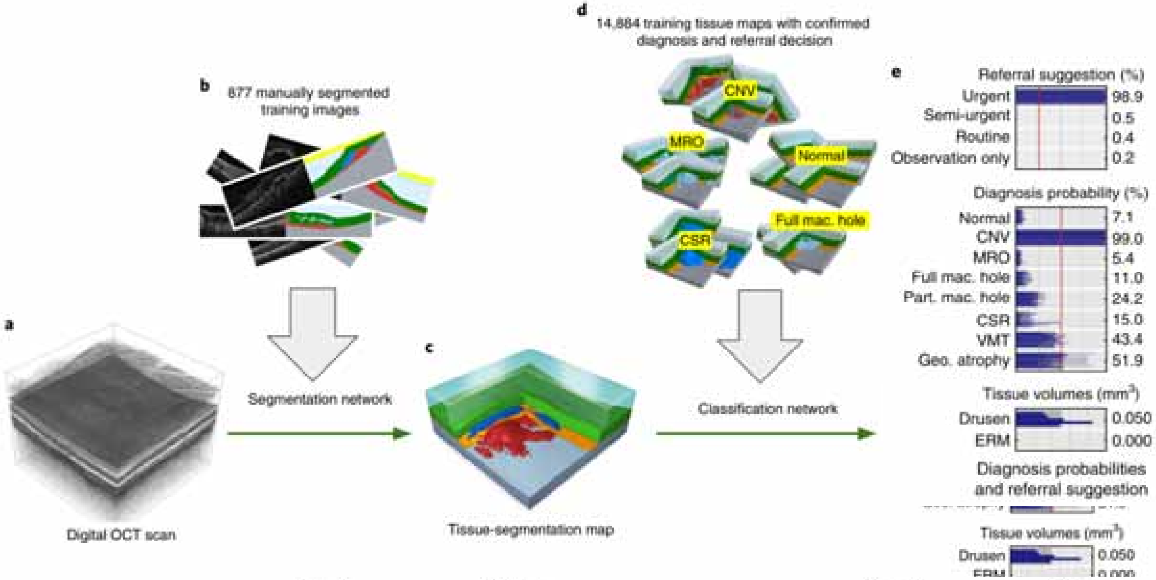


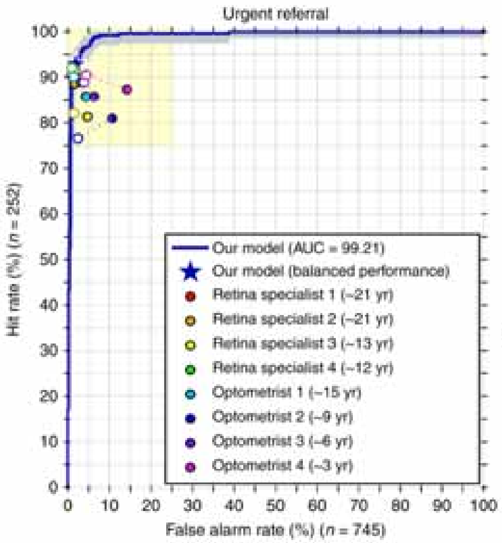


Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

Example Retinopathy: Towards Human-level AI

Jeffrey De Fauw et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24, (9), 1342-1350



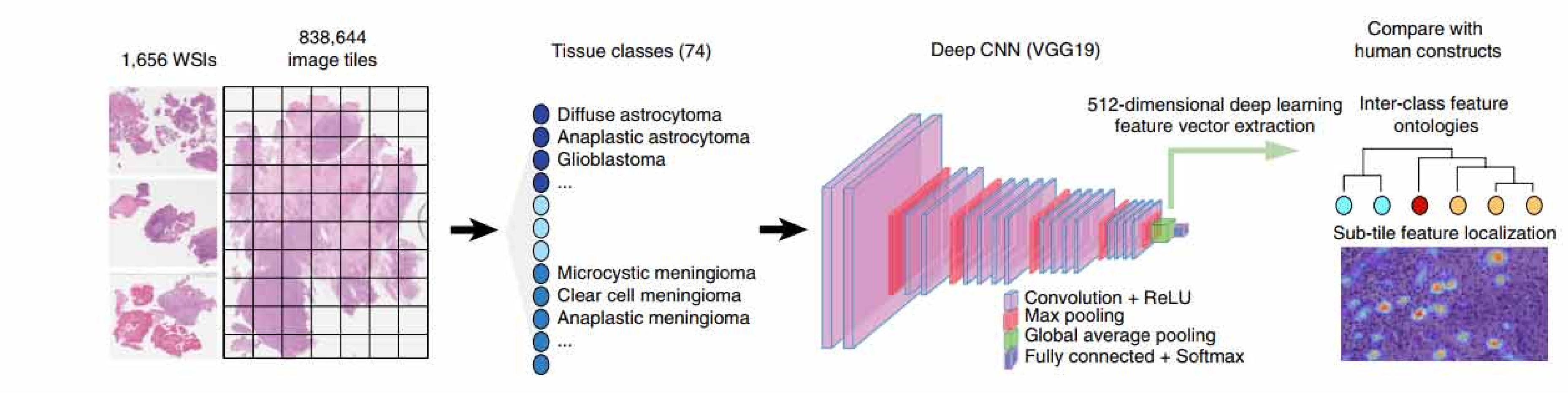


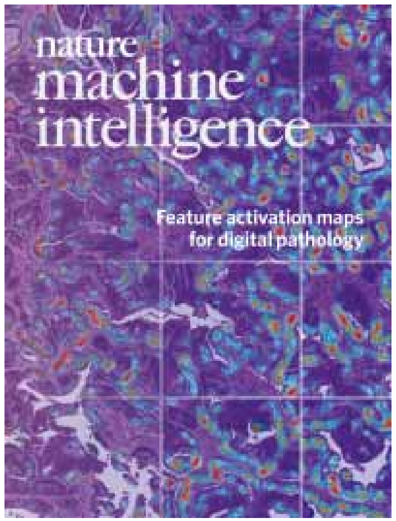
	Urgent	Semi-urgent	Routine	Observation
Urgent	234	5	13	0
Semi-urgent	3	225	2	0
Routine	10	2	250	4
Observation	1	1	14	233

Gold standard referral

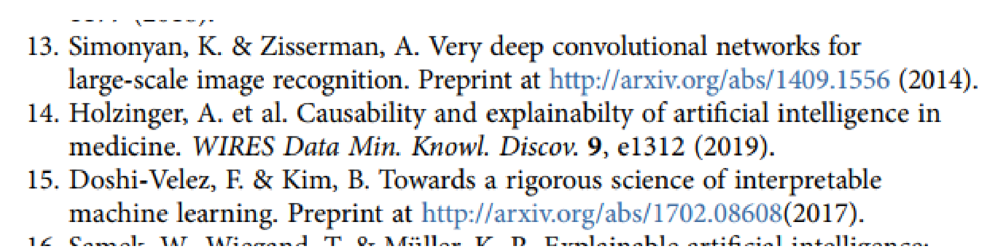
	Urgent	Semi-urgent	Routine	Observation
Urgent	228	4	20	0
Semi-urgent	3	223	4	0
Routine	2	7	254	3
Observation	1	1	10	237

Example Histopathology: Towards Human Level AI





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


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

What is the difference between humanoid AI and human-level AI ?

Lotfi A. Zadeh 2008. Toward Human Level Machine Intelligence - Is It Achievable? The Need for a Paradigm Shift. IEEE Computational Intelligence Magazine, 3, (3), 11-22, doi:10.1109/MCI.2008.926583.

Why is Machine Learning in health enormously progressing ?

- Progress is driven by the explosion in the availability of **big data** and **low-cost computation.**
- **Health concerns everyone ...**

Michael I. Jordan & Tom M. Mitchell (2015). Machine learning: Trends, perspectives, and prospects. Science, 349, (6245), 255-260, doi:10.1126/science.aaa8415.

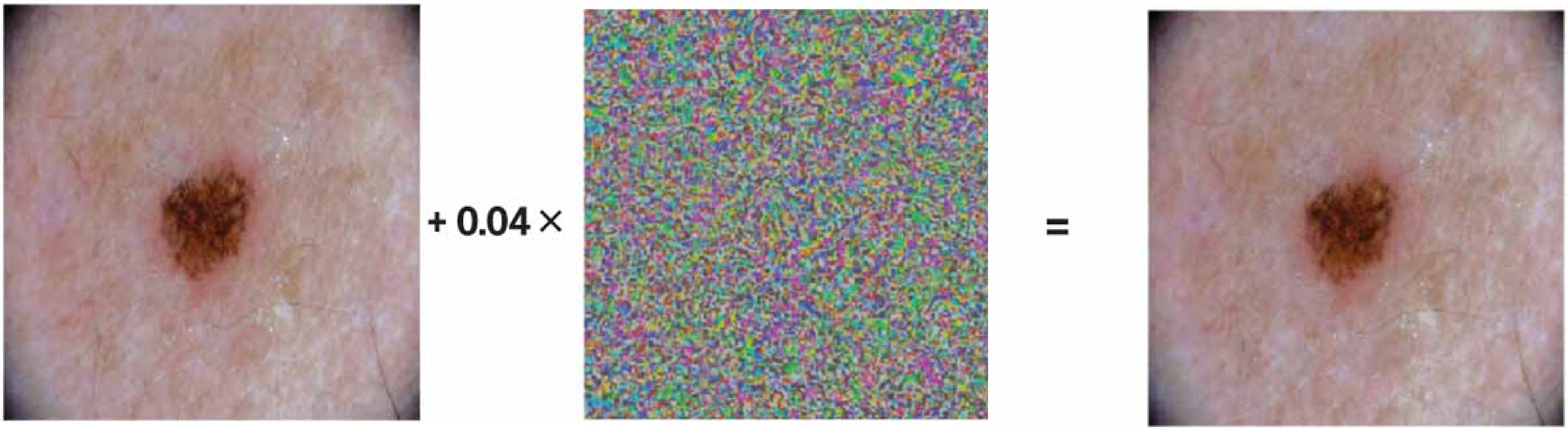
Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni & Anna Goldenberg (2019). Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine, 25, (9), 1337-1340, doi:10.1038/s41591-019-0548-6.

(1) Why can AI solve some tasks better than humans ?

(2) How did AI get these results in the first place ?

(3) What happens if I change, replace, disturb, remove, ... input data ?

Robustness & Explainability

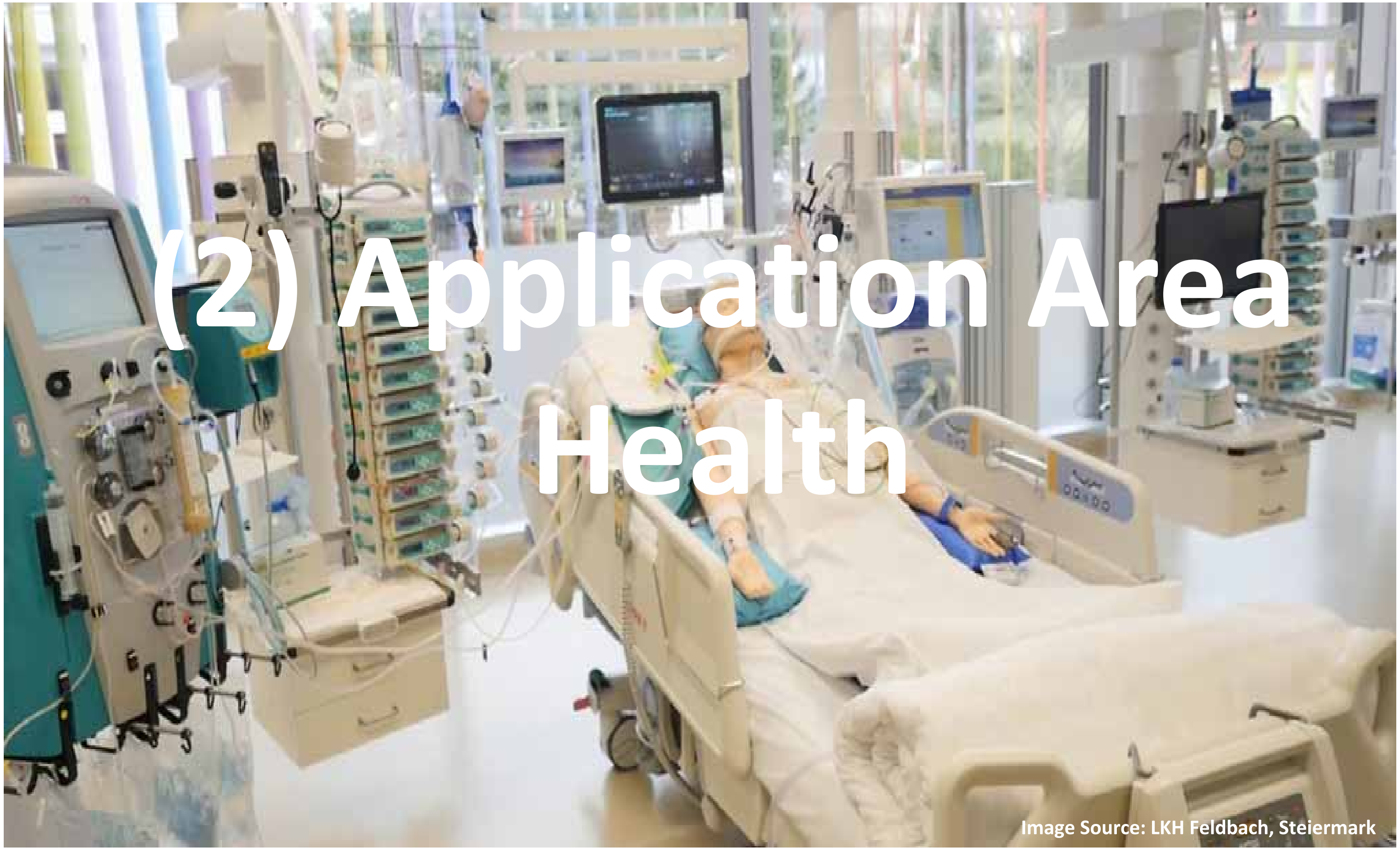


x $\quad$ $sign(\nabla_x J(\theta, x, y))$ $\quad$ $x + \varepsilon\, sign(\nabla_x J(\theta, x, y))$

Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam & Isaac S. Kohane (2019). Adversarial attacks on medical machine learning. *Science*, 363, (6433), 1287-1289, doi:10.1126/science.aaw4399

https://github.com/sgfin/adversarial-medicine

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

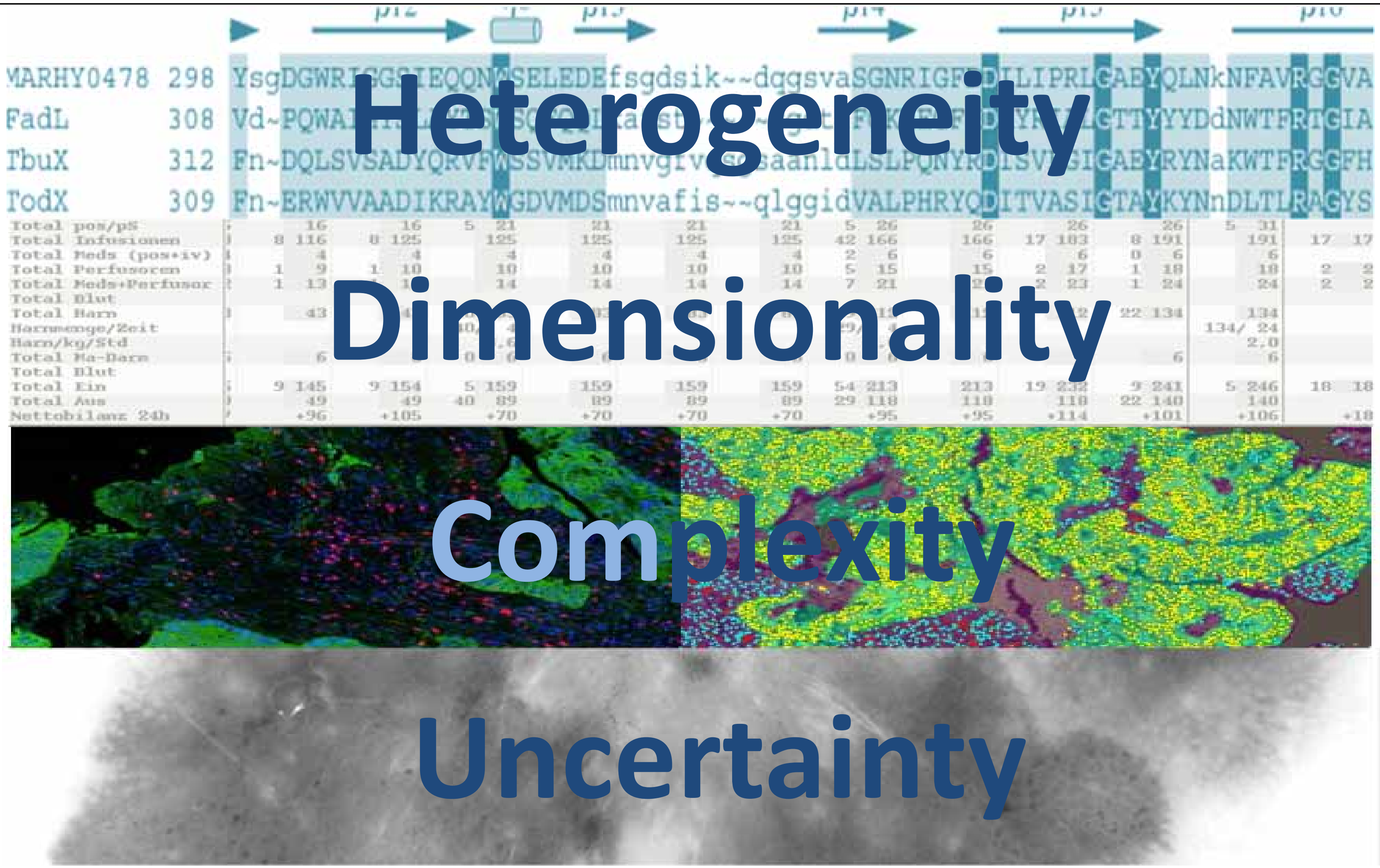

(2) Application Area
Health
Image Source: LKH Feldbach, Steiermark

Why is health a complex application domain ?

Our central hypothesis:
Information may bridge this gap

Andreas Holzinger & Klaus-Martin Simonic (eds.) 2011. Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer, doi:10.1007/978-3-642-25364-5.

What are the main problems in “building a bridge”?



Andreas Holzinger, Matthias Dehmer & Igor Jurisica 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. Springer/Nature BMC Bioinformatics, 15, (S6), I1, doi:10.1186/1471-2105-15-S6-I1.

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

(3) Probable Information and Statistical Machine Learning

What is probability p (x) ?

- $Probability\ p(x)$ is the formal study of laws of chance and managing uncertainty; allows to measure (many) events
  - Frequentist* view: coin toss
  - Bayesian* view: probability as a measure of belief (this is what made machine learning successful)
  - $p(x)$ = 1 means that all events occur for certain
  - Information is a measure for the reduction of uncertainty
  - If something is 100 % certain its uncertainty = 0
  - Uncertainty is max. if all choices are equally probable (I.I.D = independent and identically distributed)
  - Uncertainty (as information) sums up for independent sources: $\sum_x p(x = X) = 1$

*) Bayesian vs. Frequentist - please watch the excellent video of Kristin Lennox (2016): https://www.youtube.com/watch?v=eDMGDhyDxuY

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

Pierre Simon de Laplace, 1812

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

Pierre Simon de Laplace (1749-1827)

Probability Theory is nothing, but common sense reduced to calculation ...

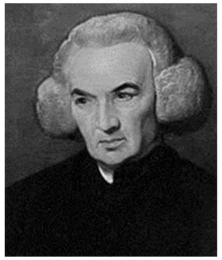
Pierre-Simon Laplace (1781). Mémoire sur les probabilités. Mémoires de l'Académie Royale des sciences de Paris, 1778, 227-332.

Pierre-Simon Laplace 1825. Philosophical Essay on Probabilities: Translated 1995 from the fifth French edition of 1825 With Notes by Andrew I. Dale, New York, Springer Science.

Who inspired Laplace for his work ?

N.B. This image is obviously not Thomas Bayes

Thomas Bayes
1701 - 1761



Richard Price
1723 -1791

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances (Postum communicated by Richard Price). Philosophical Transactions, 53, 370-418.

$$p(x_i) = \sum P(x_i, y_j)$$

$$p(x_i, y_j) = p(y_j|x_i)P(x_i)$$

Bayes' Rule is a corollary of the Sum Rule and Product Rule:

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

Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. Biometrika, 45(3/4), 293-315.

Why is the work of Laplace so important for us?

What is the simplest mathematical operation for us?

$$p(x) = \sum_x (p(x,y)) \tag{1}$$

How do we call repeated adding?

$$p(x,y) = p(y|x) * p(y) \tag{2}$$

Laplace (1773) showed that we can write:

$$p(x,y) * p(y) = p(y|x) * p(x) \tag{3}$$

Now we introduce a third, more complicated operation:

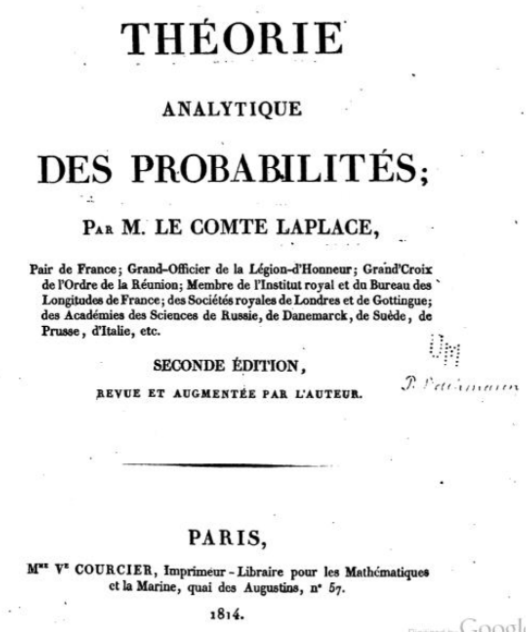
$$\frac{p(x,y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \tag{4}$$

We can reduce this fraction by $p(y)$ and we receive what is called Bayes rule:

$$p(x,y) = \frac{p(y|x) * p(x)}{p(y)} \qquad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \tag{5}$$

What are the fundamentals of the work of Bayes-Price-Laplace ?

- **1763: Richard Price publishes post hum the work of Thomas Bayes**
- **1781: Pierre-Simon Laplace: Probability theory is nothing, but common sense reduced to calculation ...**
- **1812: Théorie Analytique des Probabilités, now known as Bayes' Theorem, should be correctly named as Bayes-Price-Laplace T.**



THÉORIE

ANALYTIQUE

DES PROBABILITÉS;

PAR M. LE COMTE LAPLACE,

Pair de France; Grand-Officier de la Légion-d'Honneur; Grand'Croix de l'Ordre de la Réunion; Membre de l'Institut royal et du Bureau des Longitudes de France; des Sociétés royales de Londres et de Gottingue; des Académies des Sciences de Russie, de Danemarck, de Suède, de Prusse, d'Italie, etc.

SECONDE ÉDITION,

REVUE ET AUGMENTÉE PAR L'AUTEUR.

PARIS,

Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques et la Marine, quai des Augustins, n° 57.

1814.

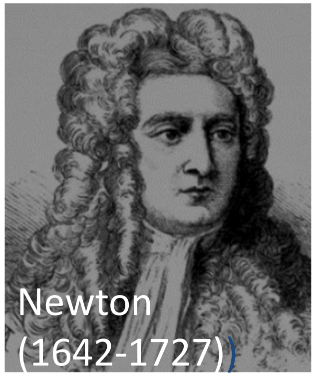
https://archive.org/details/thorieanalytiqu01laplgoog

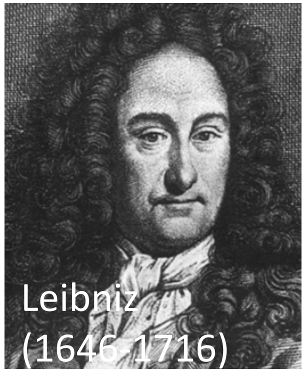
$$p(h|d) \propto p(d|h) * p(h)$$

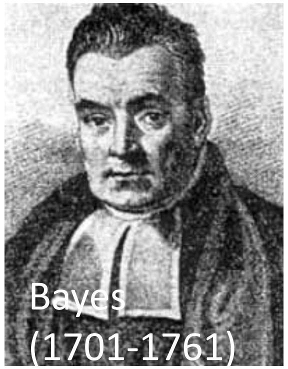
- **Hypothesis** $h \in \mathcal{H}$ (uncertain quantities (Annahmen)
- **Data** $d \in \mathcal{D}$ … measured quantities (Entitäten)
- **Prior probability** $p(h)$ … probability that h is true
  Likelihood $p(d|h)$ … "how probable is the prior"
- **Posterior Probability** $p(h|d)$ … probability of h given d

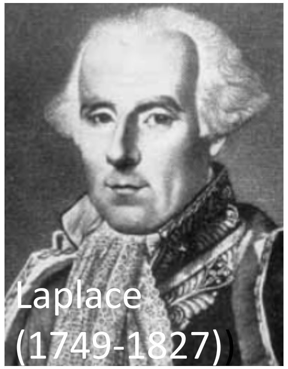
$$p(h|d) = \frac{p(d|h)p(h)}{p(d)}$$

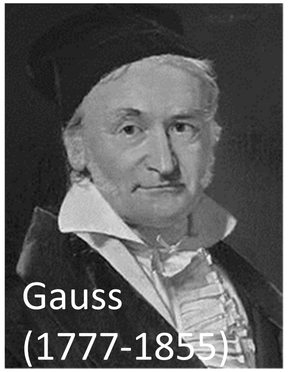
Who are our giants in the foundations of statistical machine learning?











- **Newton, Leibniz, … developed calculus – mathematical language for describing and dealing with rates of change**
- **Bayes, Laplace, … developed probability theory - the mathematical language for describing and dealing with uncertainty**
- **Gauss generalized those ideas**

How does probabilistic ("Bayesian") machine learning work ?

$$p(x_i) = \sum P(x_i, y_j)$$

$$p(x_i, y_j) = p(y_j|x_i)P(x_i)$$

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances (Postum communicated by Richard Price). Philosophical Transactions, 53, 370-418.

Bayes' Rule is a corollary of the Sum Rule and Product Rule:

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

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis})P(\text{data}|\text{hypothesis})}{\sum_h P(\text{h})P(\text{data}|\text{h})} \qquad P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(D|\theta, m)$ likelihood of parameters θ in model m

$P(\theta|m)$ prior probability of θ

$P(\theta|D, m)$ posterior of θ given data D

Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. Biometrika, 45(3/4), 293-315.

d … *data*

h … *hypotheses*

$\mathcal{H}$ … $\{H_1, H_2, \dots, H_n\}$ $\quad \forall\, h, d$ …

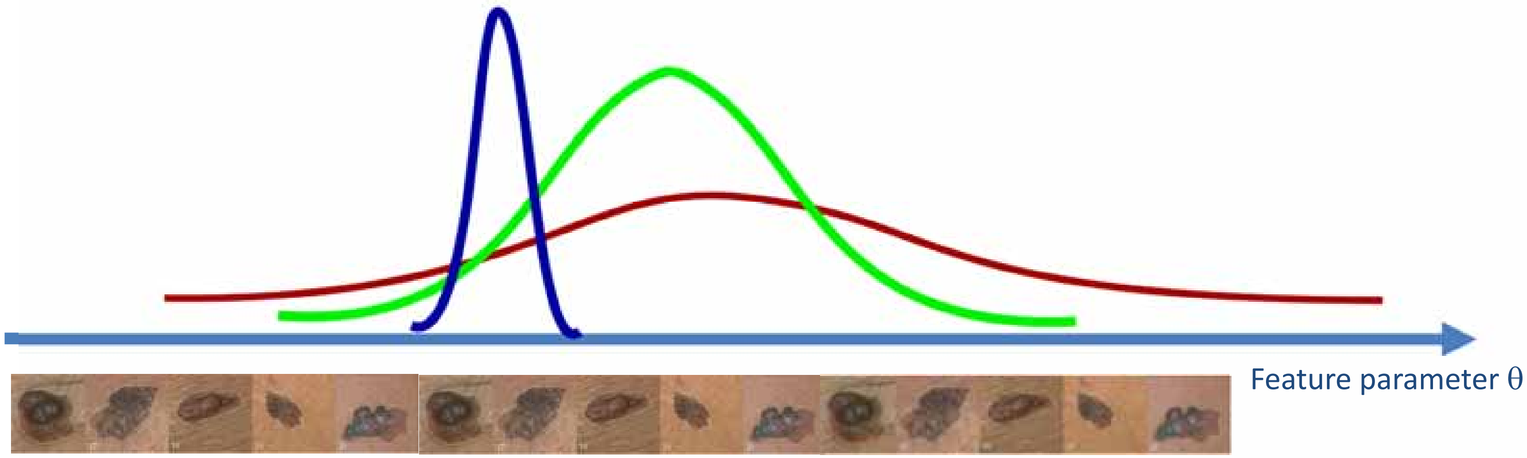
Likelihood

Prior Probability

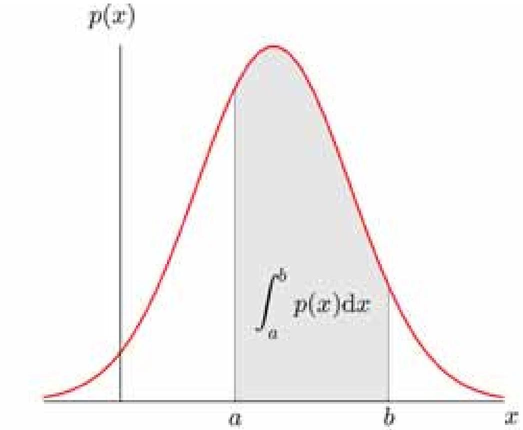
$$p(h|d) = \frac{p(d|h) * p(h)}{\sum_{h \in H} p(d|h')\, p(h')}$$

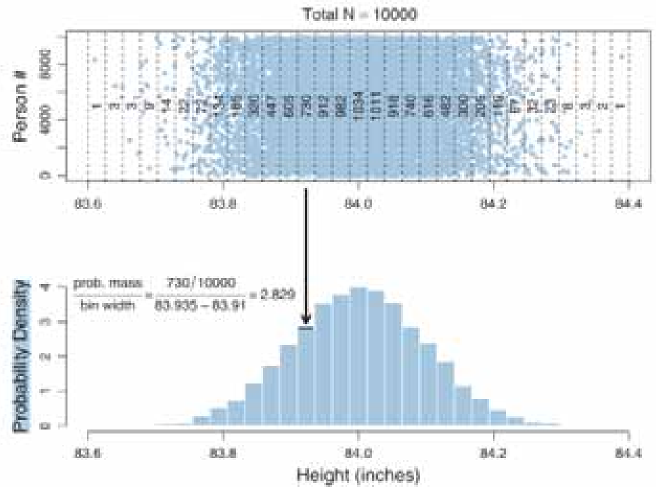
Posterior Probability

Problem in $\mathbb{R}^n \rightarrow$ complex

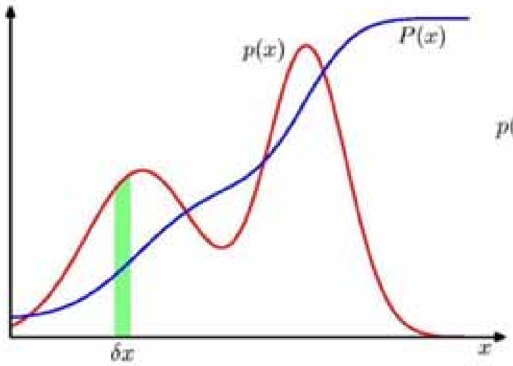


What is the Probability Density Function vs. Probability Distribution ?





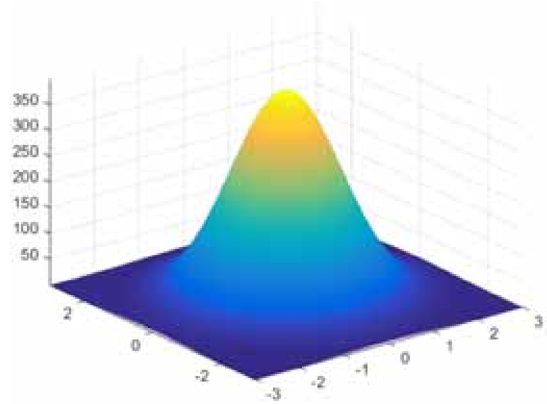
John Kruschke 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, Amsterdam et al., Academic Press.



$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

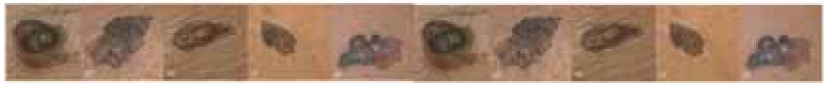
$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$



https://brilliant.org/wiki/multivariate-normal-distribution

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

$$p(\mathcal{D}|\theta)$$



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

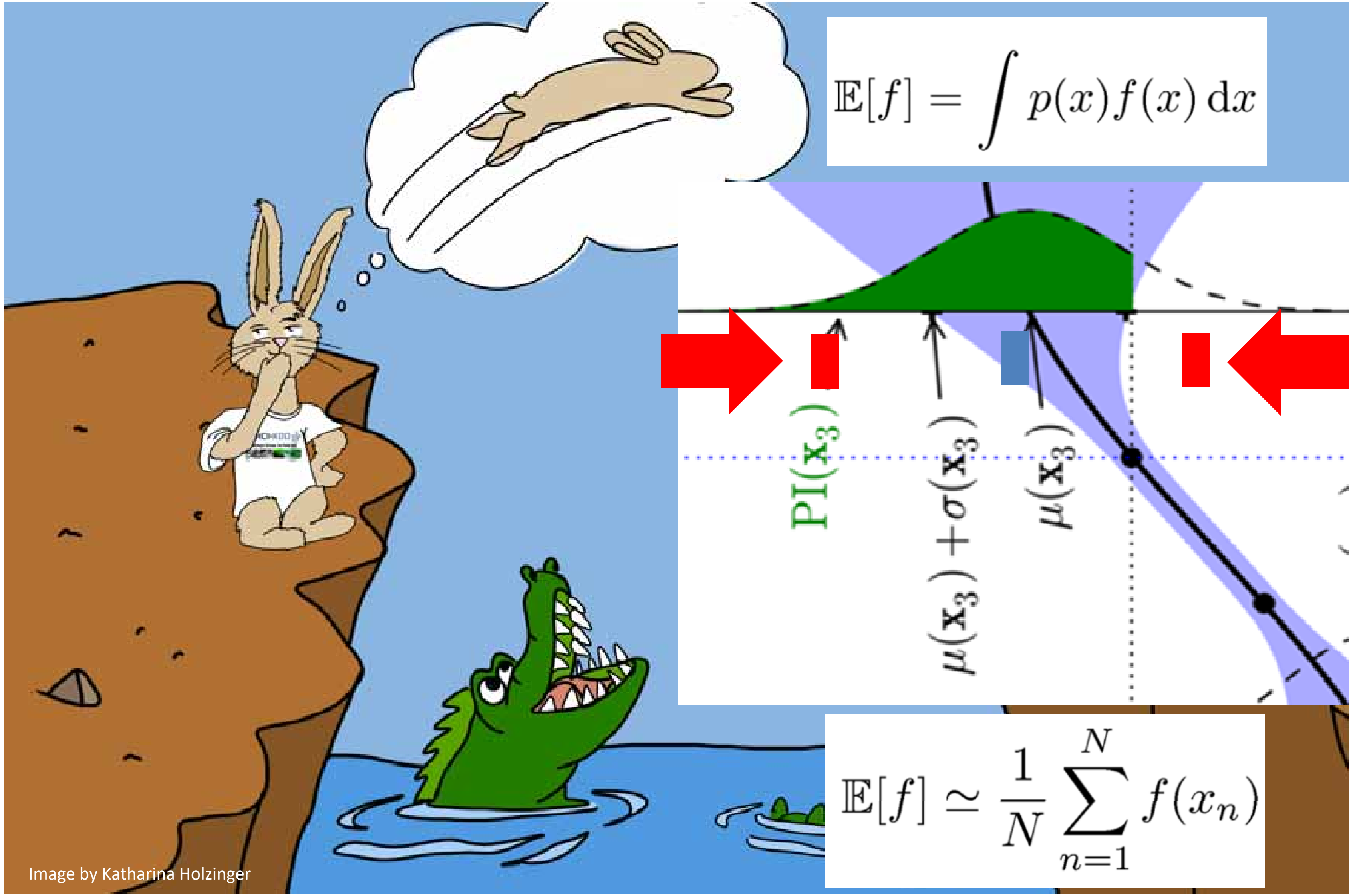
$$posterior = \frac{likelihood * prior}{evidence}$$

The inverse probability allows to learn from data, infer unknowns, and make predictions

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

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

Expectation
and
Expected Utility Theory



$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$
PI($\mathbf{x}_3$)
$\mu(\mathbf{x}_3)+\sigma(\mathbf{x}_3)$
$\mu(\mathbf{x}_3)$
$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$
Image by Katharina Holzinger

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

The expected utility of decision $D = d$ is

http://www.eoht.info/page/Oskar+Morgenstern

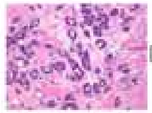
$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d) U(x_1,\ldots,x_n,d)$$

An optimal single decision is the decision $D = dmax$ whose expected utility is maximal:

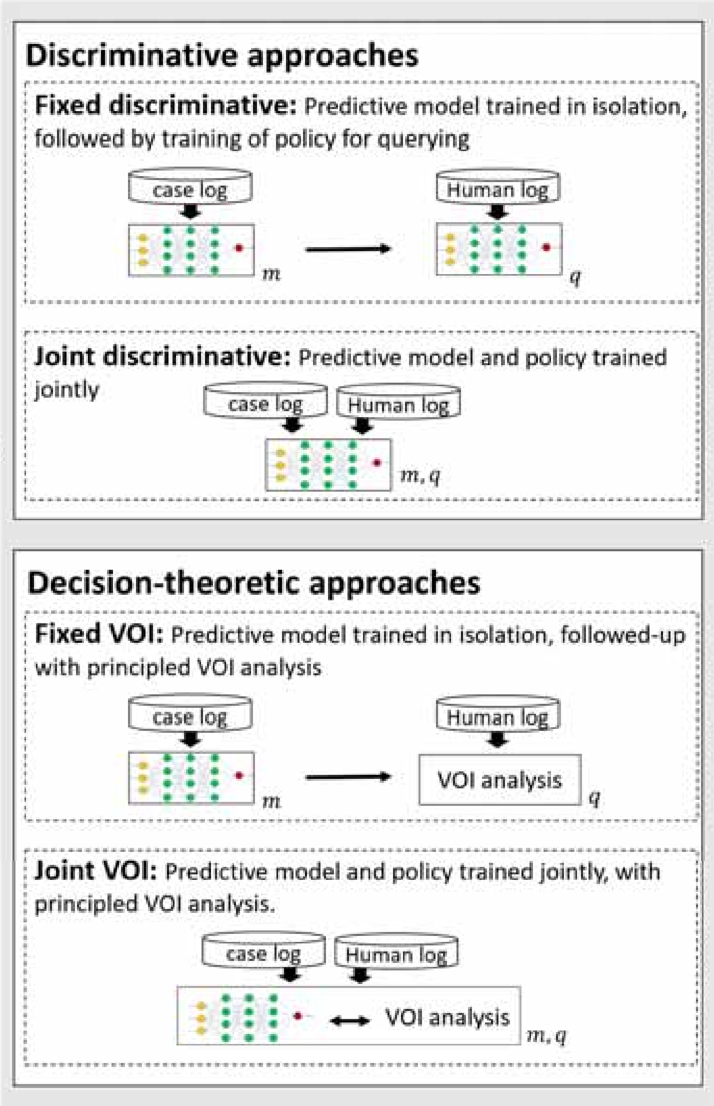
$$d_{\max} = \arg \max_{d \in \mathrm{dom}(D)} E(U \mid d)$$

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

Example: Learning to complement humans



Input



$$\mathbb{E}_{(x,y,h)\sim P}\Big[q_\theta(x)\left(u(y,m_\theta(x,h))-c\right) + (1-q_\theta(x))\left(u(y,m_\theta(x))\right)\Big].$$

$$u_{\text{nq}} = \max_{\hat{y}\in\mathcal{Y}}\left(\sum_{y\in\mathcal{Y}} p_\alpha(y|x)u(\hat{y},y)\right)$$

Decision to consult human expert (pay cost)

$$u_{\text{q}} = \mathbb{E}_{h\sim p_\beta(h|x)}\left[\max_{\hat{y}\in\mathcal{Y}}\left(\sum_{y\in\mathcal{Y}} p_\gamma(y|x,h)u(\hat{y},y)\right)\right] - c$$

Classification decision

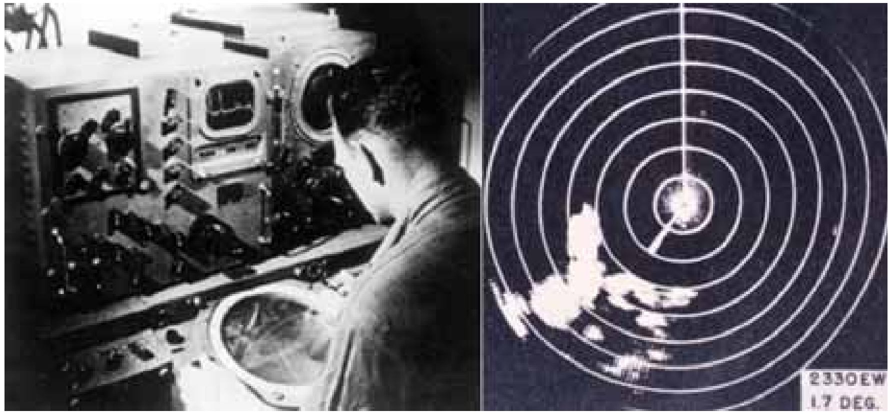
Algorithm 1 Joint VOI training

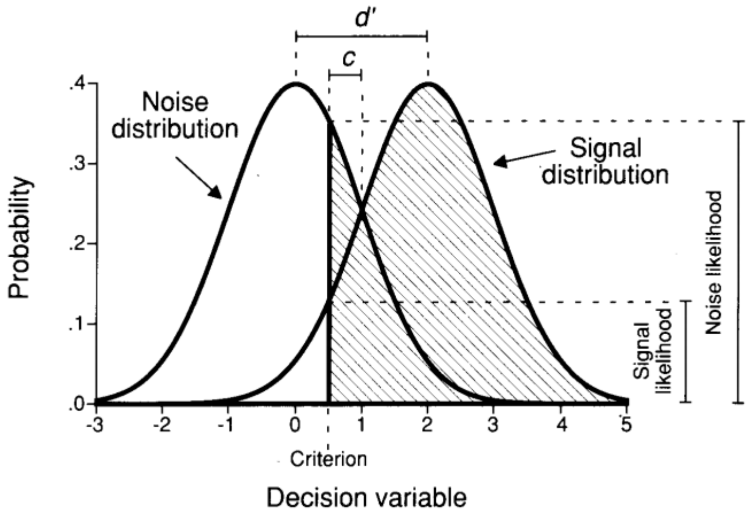
for T iterations **do**
  Sample a minibatch $B \subseteq [n]$
  for $i \in B$ **do**
    for $\hat{y} \in \mathcal{Y}$ **do**
      $u_{\text{nq}}(\hat{y}) = \sum_{y\in\mathcal{Y}} p_\alpha(y|x_i)u(\hat{y},y)$
    end for
    $u_{\text{nq}} = \sum_{\hat{y}\in\mathcal{Y}} \frac{u_{\text{nq}}(\hat{y})\exp(u_{\text{nq}}(\hat{y}))}{\sum_{y'\in\mathcal{Y}}\exp(u_{\text{nq}}(y'))}$
    for $\hat{y} \in \mathcal{Y}$ **do**
      $u_{\text{q}}(\hat{y},h) = \sum_{y\in\mathcal{Y}} p_\gamma(y|x_i,h)u(\hat{y},y)$
    end for
    $u_{\text{q}} = \sum_{h\in\mathcal{Y}} p_\beta(h|x)\sum_{\hat{y}} \frac{u_{\text{q}}(\hat{y},h)\exp(u_{\text{q}}(\hat{y},h))}{\sum_{y'\in\mathcal{Y}}\exp(u_{\text{q}}(y',h))}$
    $q = \frac{\exp(u_{\text{q}})}{\exp(u_{\text{q}})+\exp(u_{\text{nq}})}$
    $\ell^i_{\text{combined}} = \ell(q\,p_\gamma(\cdot|x_i,h_i)$
      $+(1-q)p_\alpha(\cdot|x_i)) + qc$
  end for
  Backpropagate $\frac{1}{|B|}\sum_{i\in B}\ell^i_{\text{combined}}$
  Every t iterations: update calibrators
end for

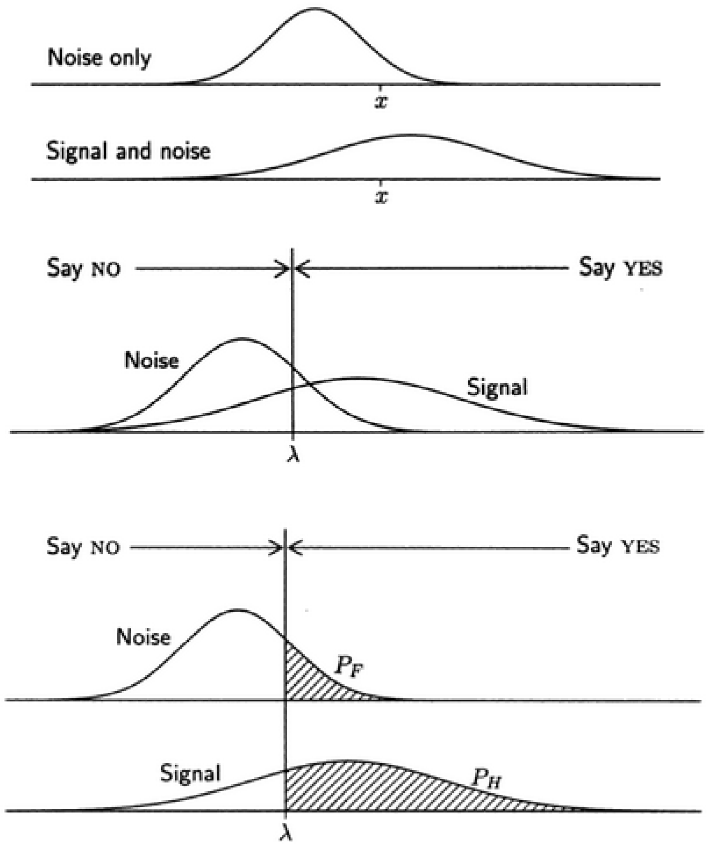
Bryan Wilder, Eric Horvitz & Ece Kamar (2020). Learning to complement humans. arXiv:2005.00582.

What was the origin of probabilistic decision making ?

Image source: Staffordshire University Computing Futures Museum http://www.fcet.staffs.ac.uk/jdw1/sucfm/malvern.htm

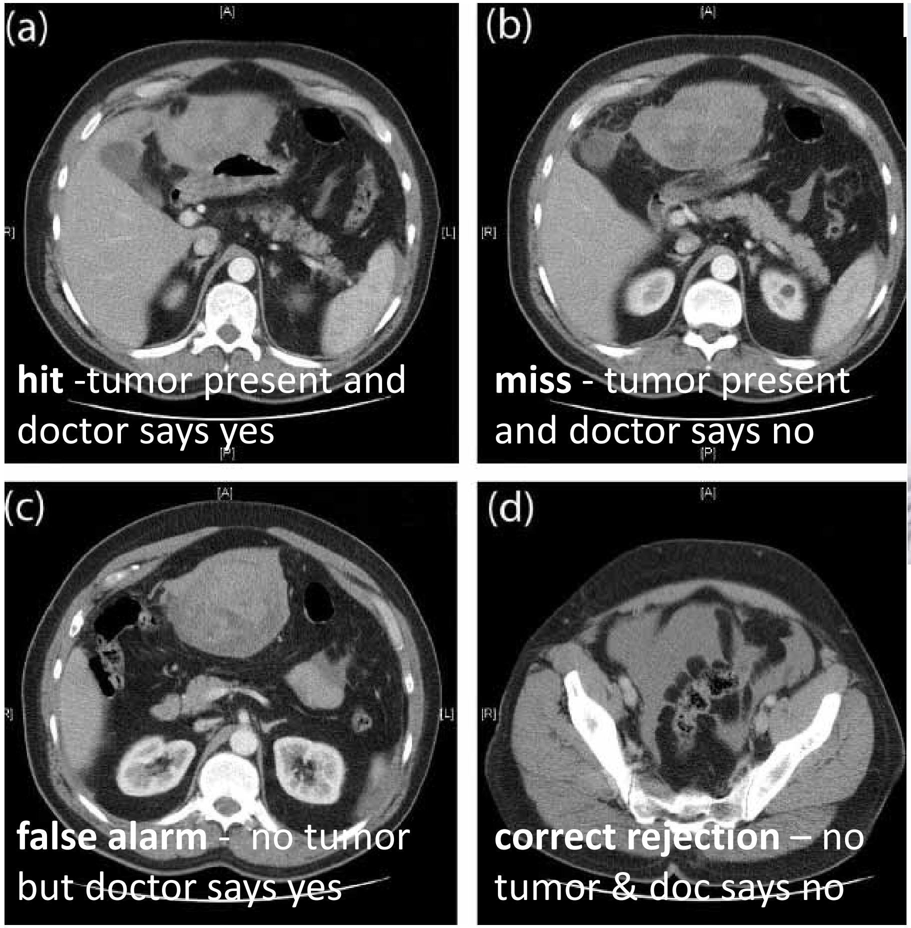






Stanislaw, H. & Todorov, N. 1999. Calculation of signal detection theory measures. Behavior research methods, instruments, & computers, 31, (1), 137-149.

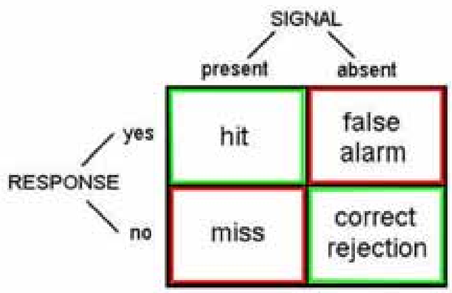
How does this work in medical decision making ?



Two doctors, with equally good training, looking at the same CT scan, will have the same information … but they may have a **different bias/criteria!**

What does a correct rejection mean?

Remember: Two doctors, with equally good training, looking at the same CT scan data, will have the same information … but they may gain different knowledge due to *bias/criteria*.



Positive = identified and negative = rejected
True positive = correctly identified (hit)
False positive = incorrectly identified, false alarm, type I error
True negative = correctly rejected (correct rejection)
False negative = incorrectly rejected, miss, type II error

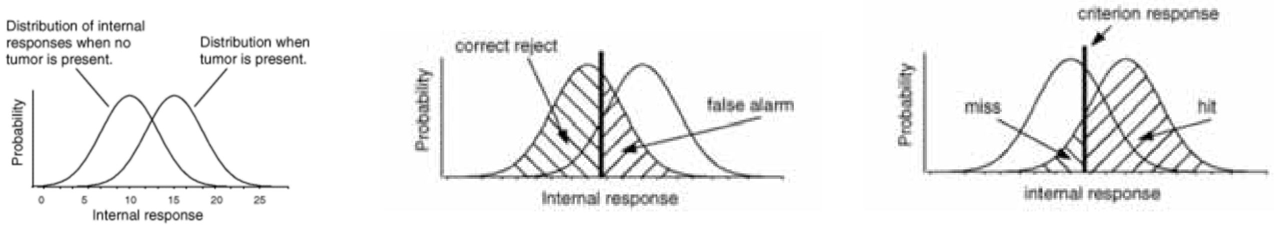
sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

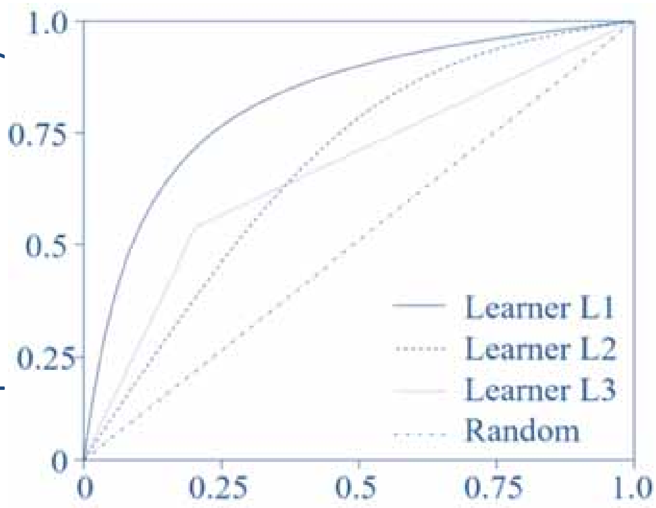
https://en.wikipedia.org/wiki/Sensitivity_and_specificity



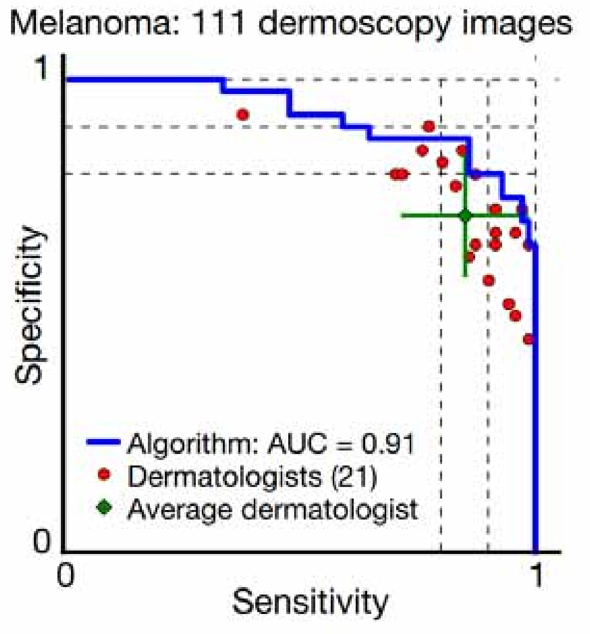
For an example see: Braga & Oliveira (2003) Diagnostic analysis based on ROC curves: theory and applications in medicine. *Int. Journal of Health Care Quality Assurance, 16, 4, 191-198.*
And please look up the Wikipedia page:

Why do we need specificity and sensitivity ?

True positive = Sensitivity

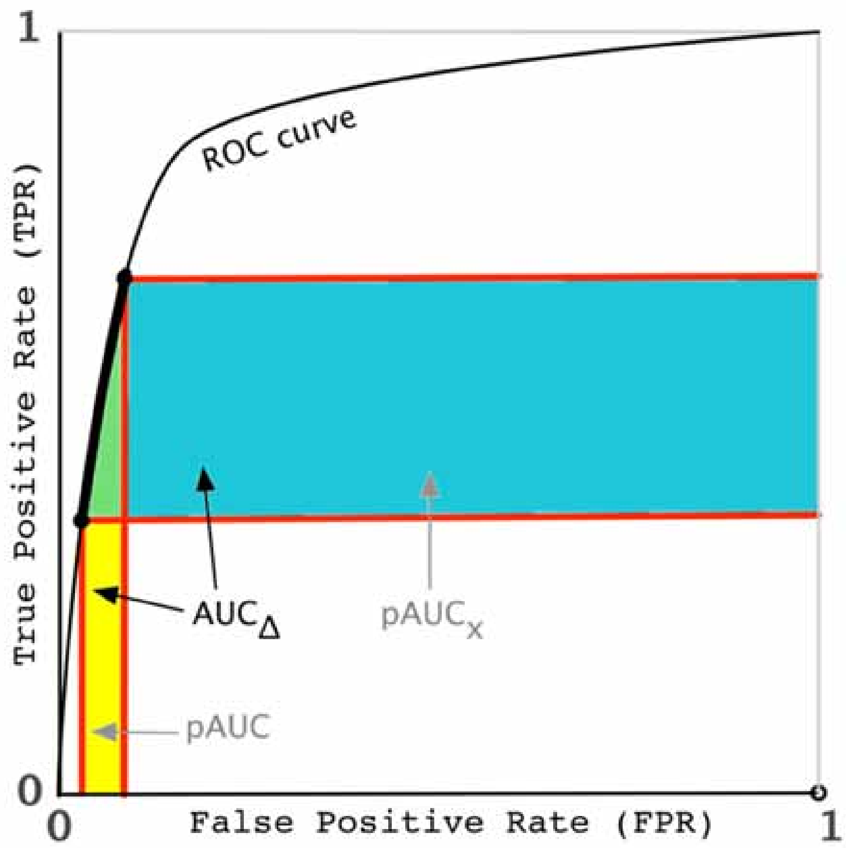


False positive = Specificity



Andrew P. Bradley 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, (7), 1145-1159, doi:http://dx.doi.org/10.1016/S0031-3203(96)00142-2

What is the advantage of the concordant pAUC ?



André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr & Douglas G. Manuel 2020. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. Springer/Nature BMC Medical Informatics and Decision Making, 20, (1), 4, doi:10.1186/s12911-019-1014-6.
https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1014-6

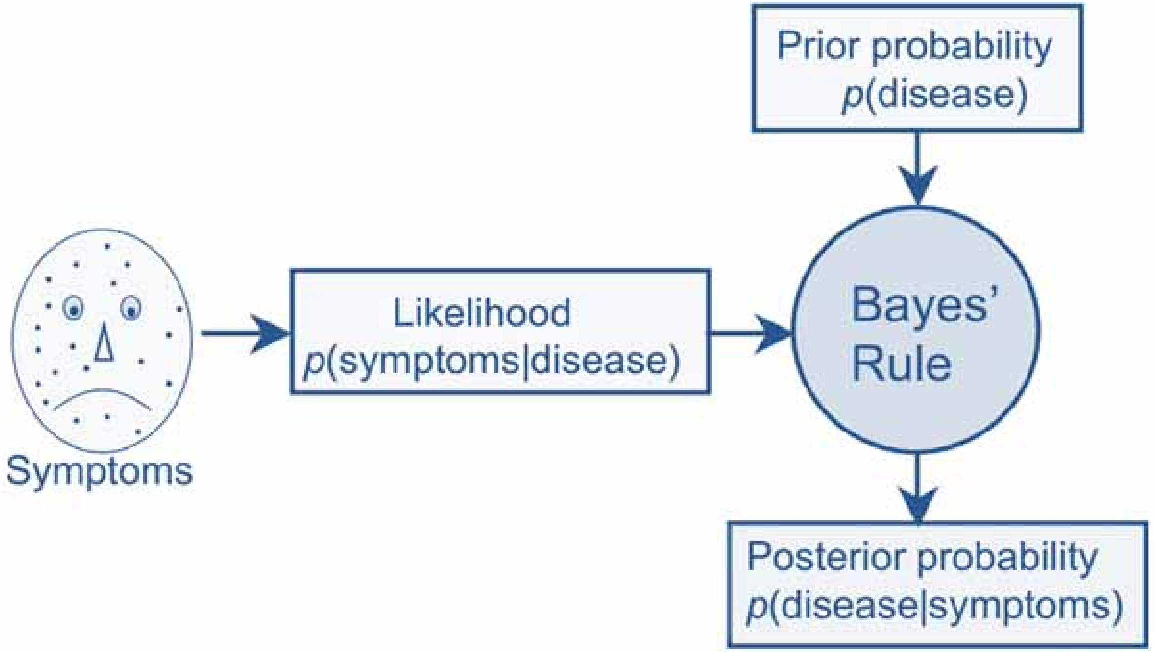
Why do we need to take care of ROC for understanding & interpretation ?

- Optimal performance is critical for decision-making
- common performance measures may be too general or too specific.
- AUC too general because including unrealistic decision thresholds.
- Accuracy, sensitivity or the F1 score are measures at a single threshold that reflect an individual single probability or predicted risk, rather than a range of individuals or risk.
- Deep ROC examines groups of probabilities or predicted risks for more insightful analysis.
- that can improve model selection in some cases and
- provide interpretation and assurance for patients in each risk group

Andre M. Carrington, Douglas G. Manuel, Paul W. Fieguth, Tim Ramsay, Venet Osmani, Bernhard Wernly, Carol Benett, Steven Hawken, Matthew Mcinnes, Olivia Magwood, Yusuf Sheikh & Andreas Holzinger (2021). Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding and Interpretation. https://arxiv.org/abs/2103.11357

https://github.com/Big-Life-Lab/deepROC

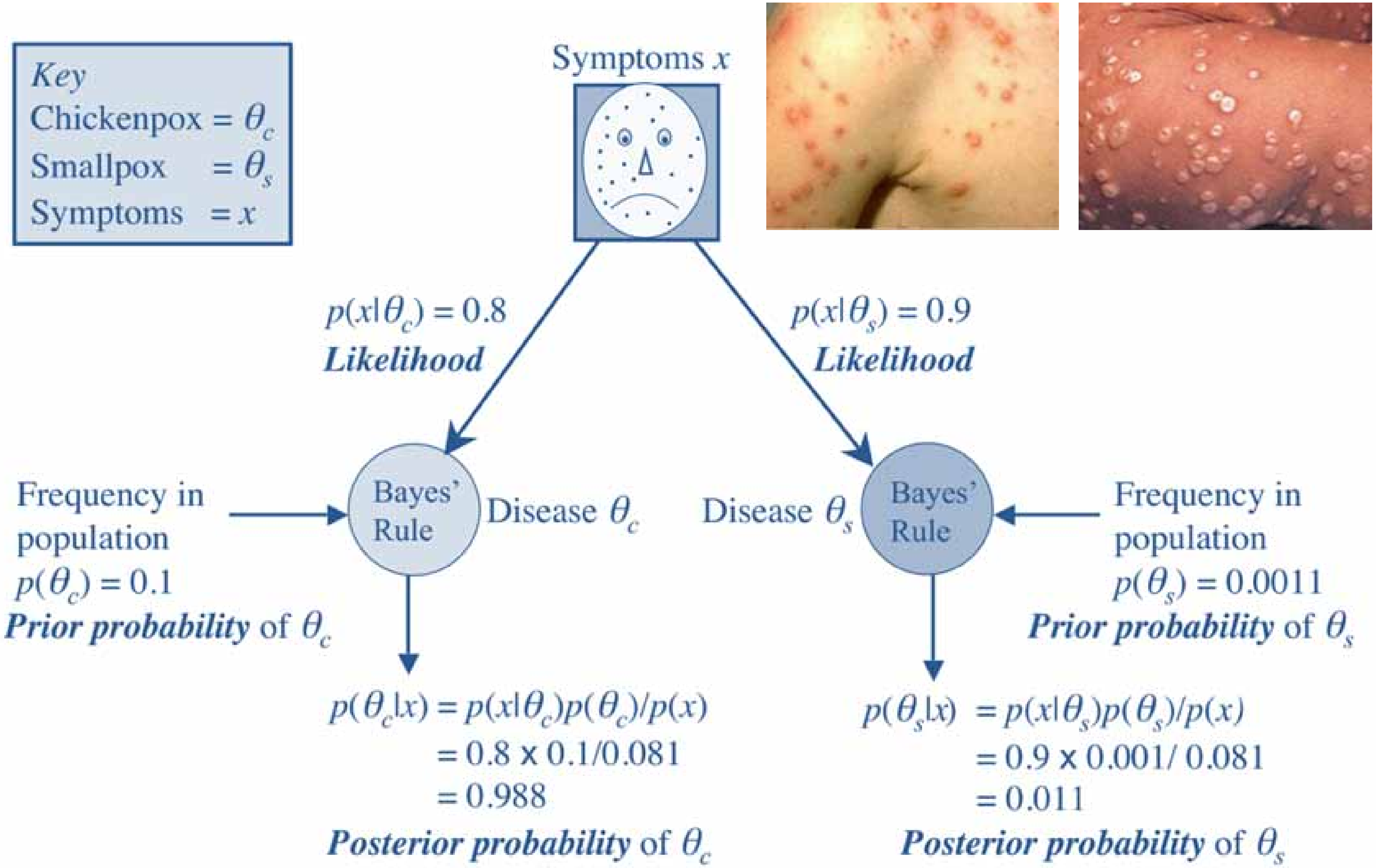
What is the principle of Bayesian diagnostic in medicine ?



$$p(\text{disease}|\text{symptoms}) = \frac{p(\text{symptoms}|\text{disease})p(\text{disease})}{p(\text{symptoms})}$$

James V. Stone 2013. Bayes' rule: a tutorial introduction to Bayesian analysis. Sebtel Press.

How does Bayesian inference work ?



James V. Stone 2013. Bayes' rule: a tutorial introduction to Bayesian analysis. Sebtel Press.

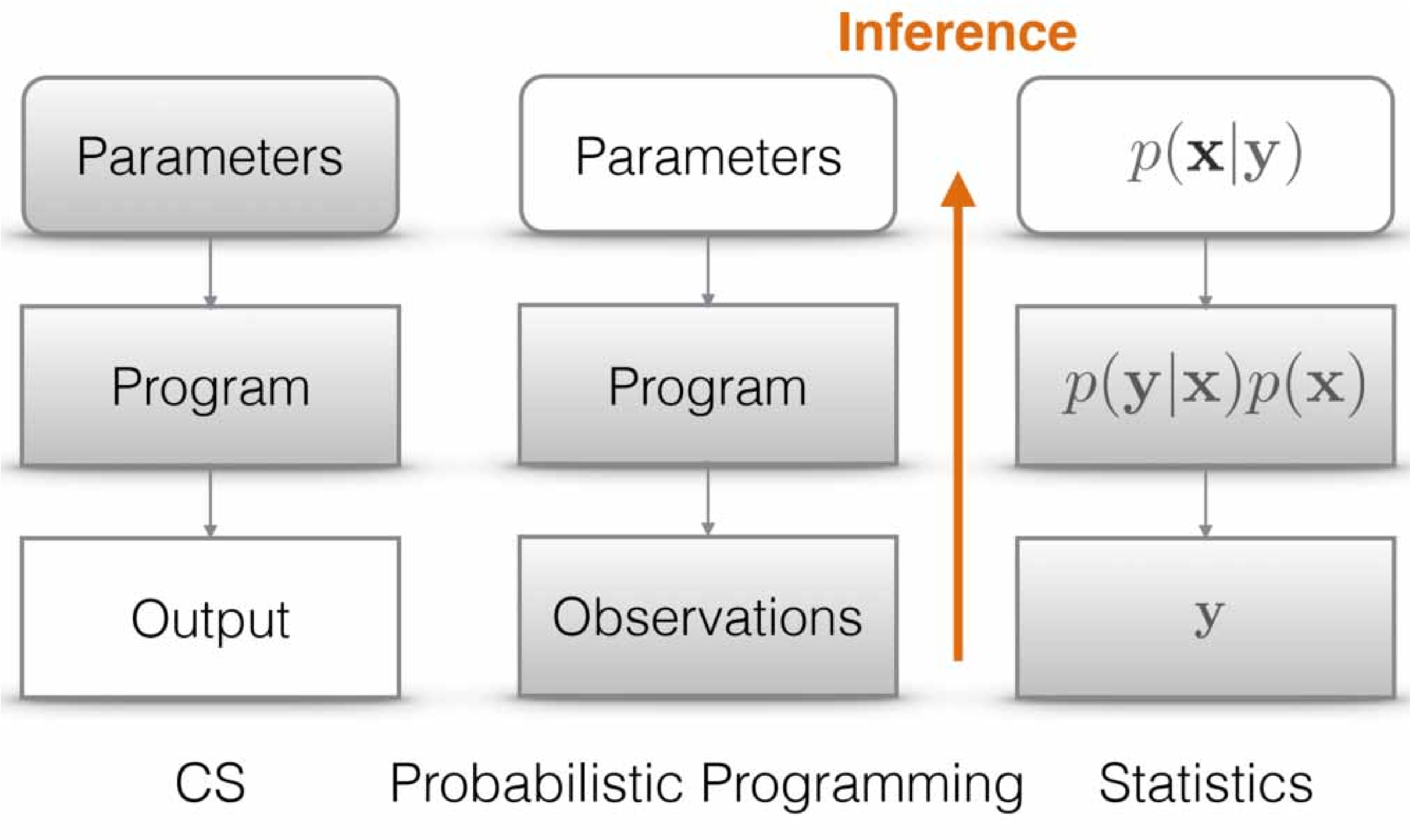
- Your MD has bad news and good news for you.
- Bad news first: You are tested positive for a serious disease, and the test is 99% accurate if you are infected (T)
- Good news: It is a rare disease, striking 1 in 10,000 (D)
- **How worried would you now be?**

$$posterior\ p(x) = \frac{likelihood * prior\ p(x)}{evidence} \qquad p(h|d) = \frac{p(d|h)p(h)}{p(d)}$$

$$p(T = 1|D = 1) = p(d|h) = 0{,}99\ \ and$$
$$p(D = 1) = p(h) = 0{,}0001$$

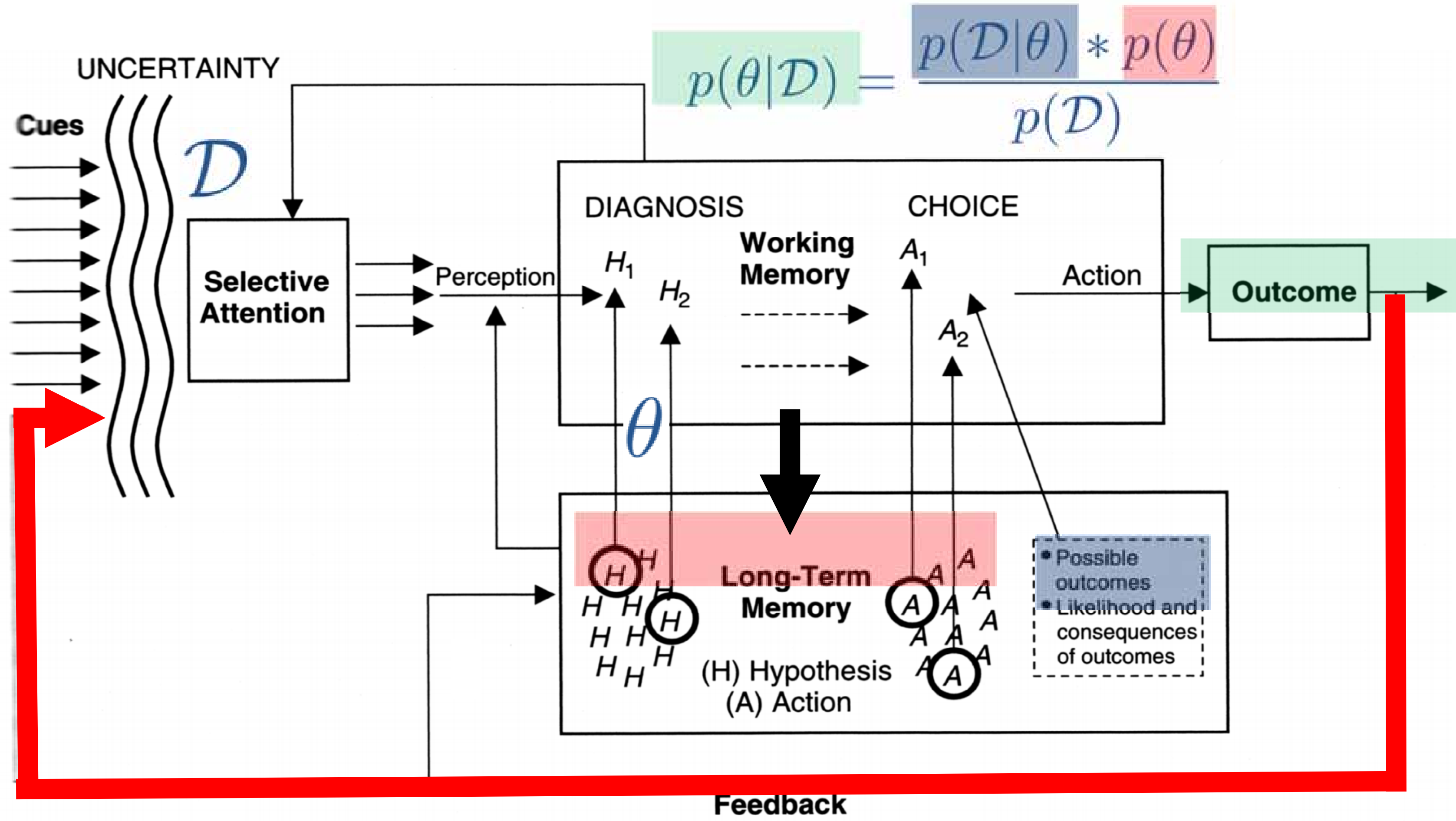
$$p(D = 1 \mid T = 1) = \frac{(0{,}99)*(0{,}0001)}{(1-0{,}99)*(1-0{,}0001)+0{,}99*0{,}0001} = \mathbf{0{,}0098}$$

What is Inference ?



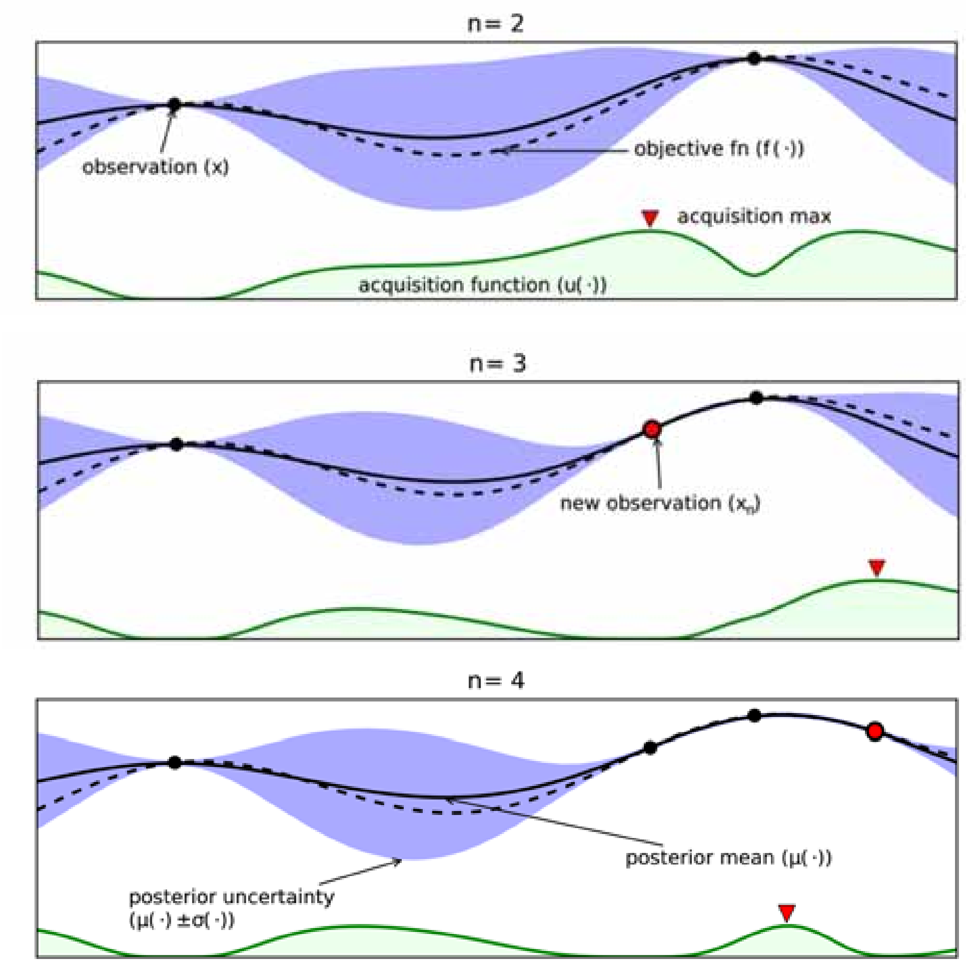
Jan-Willem Van De Meent, Brooks Paige, Hongseok Yang & Frank Wood 2018. An introduction to probabilistic programming. arXiv preprint arXiv:1809.10756.

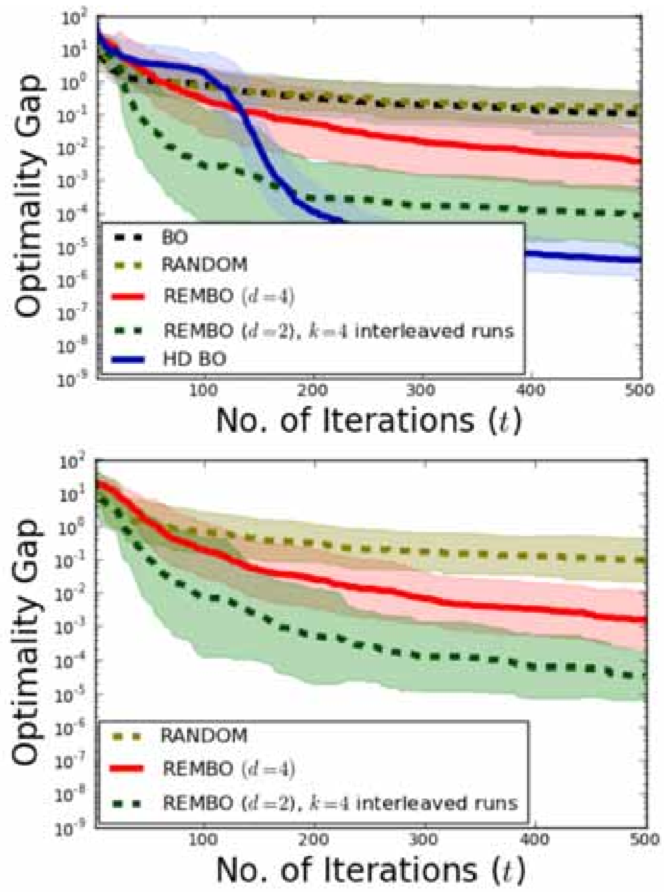
What is the Connection to Cognitive Science/Decision Making ?



Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill, modified by Holzinger, A.*

Can we scale into the high dimensions ?

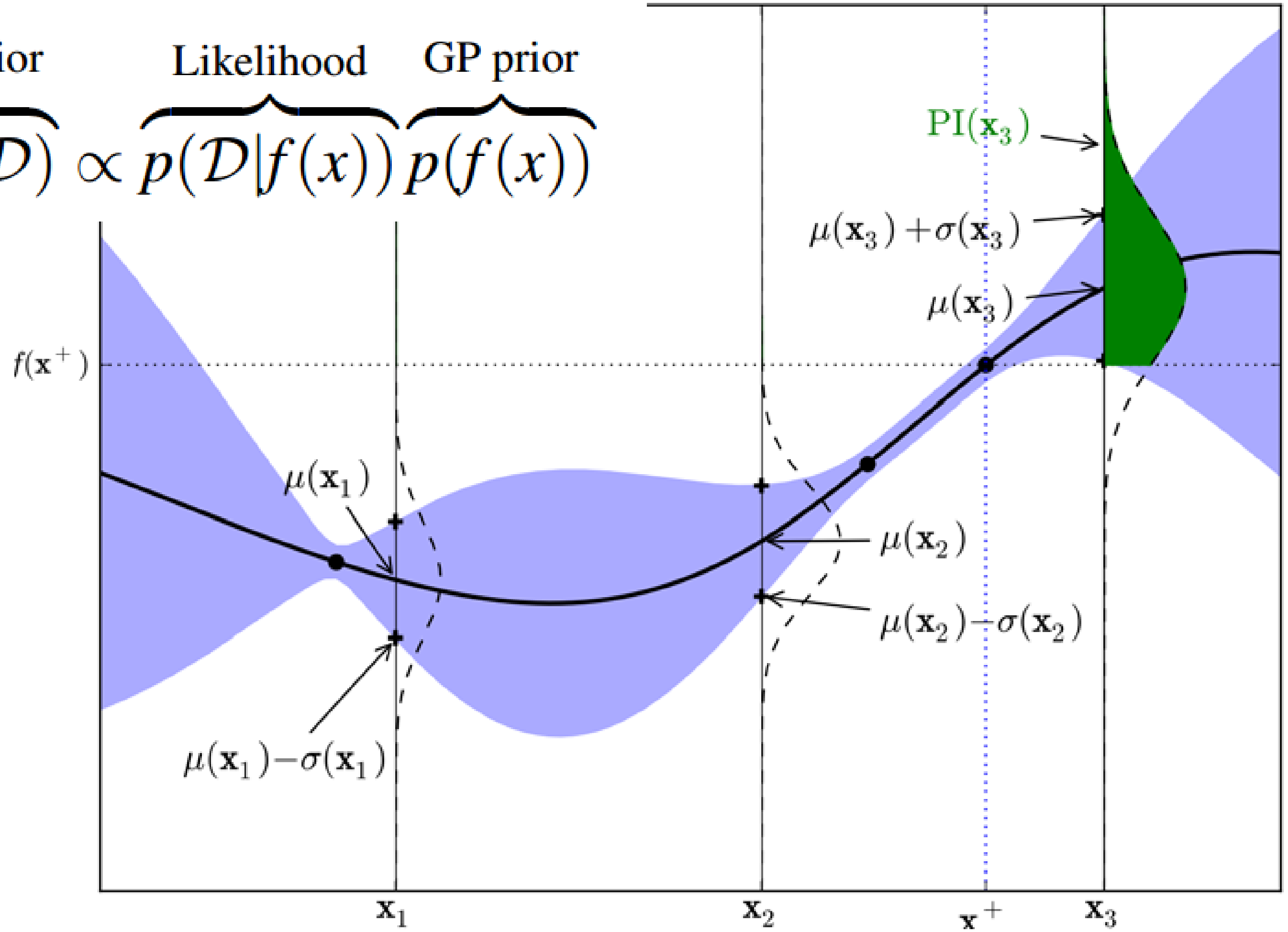




Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research, 55, 361-387, doi:10.1613/jair.4806.

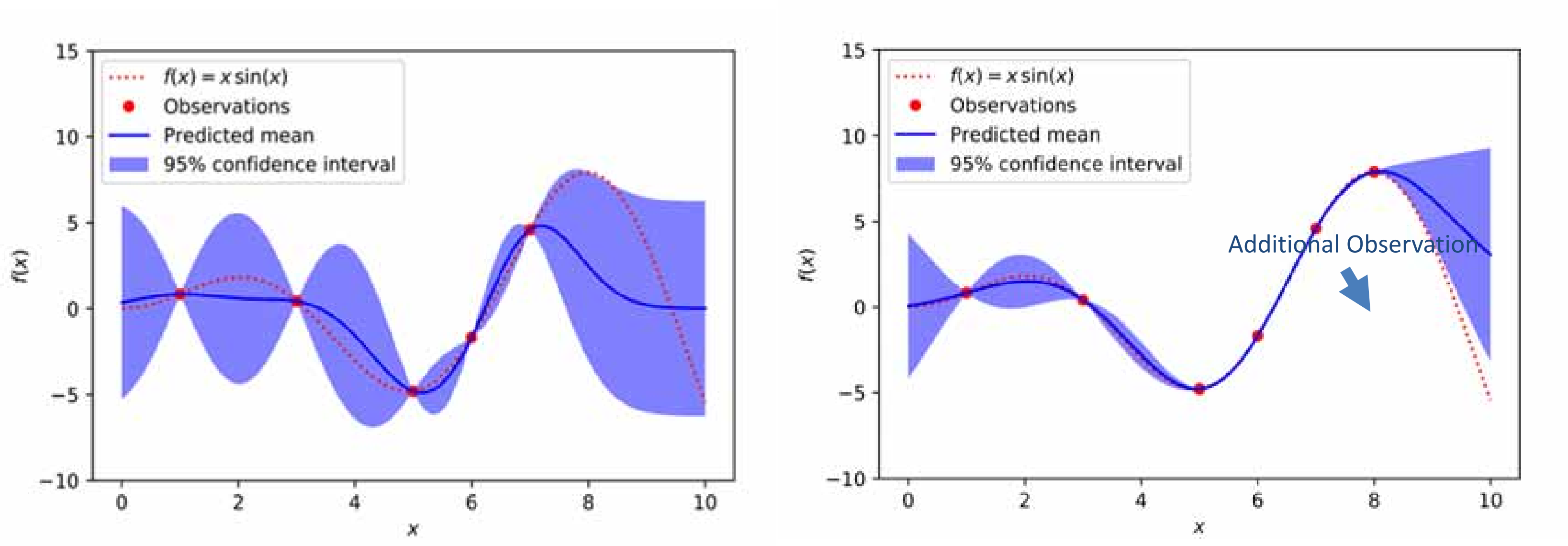
What is a Gaussian process ?

$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}}\,\overbrace{p(f(x))}^{\text{GP prior}}$$



Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

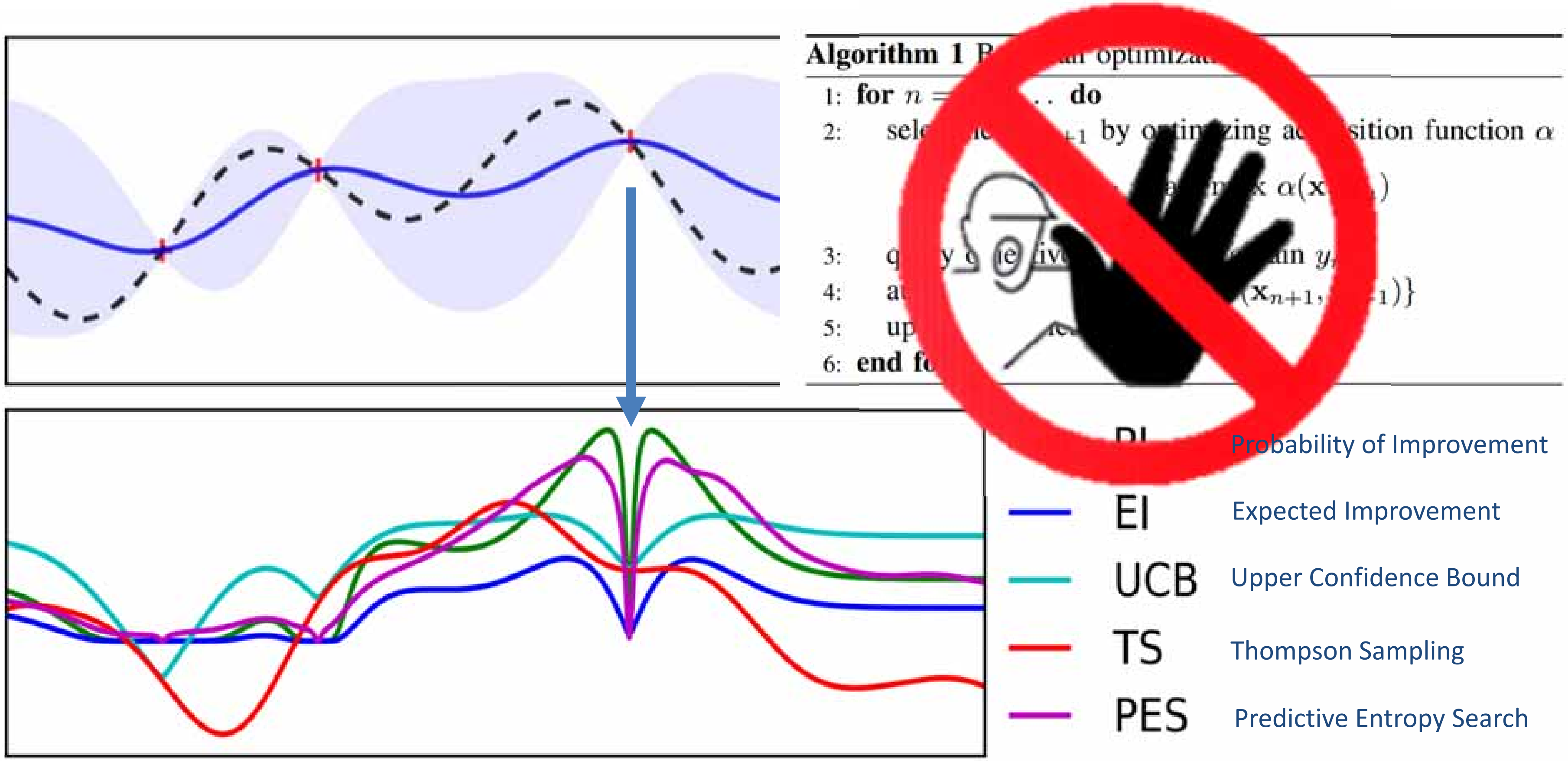
How can we reduce uncertainty?



$$\mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x \qquad \mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Holzinger, A. 2017. Introduction to Machine Learning and Knowledge Extraction (MAKE). Machine Learning and Knowledge Extraction, 1, (1), 1-20, doi:10.3390/make1010001.

Why do we take the human-out-of-the-loop?



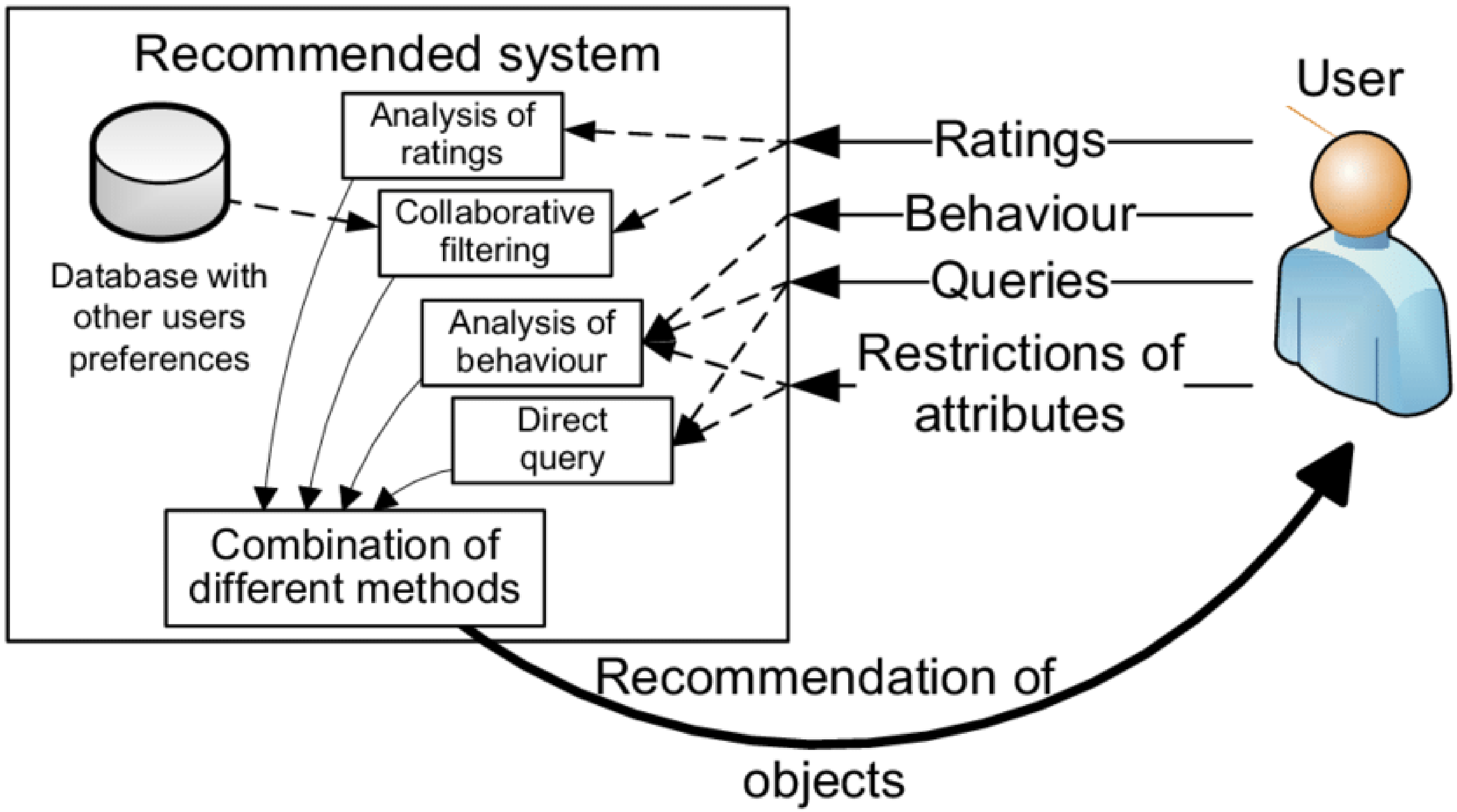
Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. 2016. **Taking the human out of the loop:** A review of Bayesian optimization. *Proceedings of the IEEE,* 104, (1), 148-175, doi:10.1109/JPROC.2015.2494218.

(4) aML

What does level 5 autonomy mean?

Guizzo, E. 2011. How google's self-driving car works. IEEE Spectrum Online, 10, 18.

What does a recommender system do?

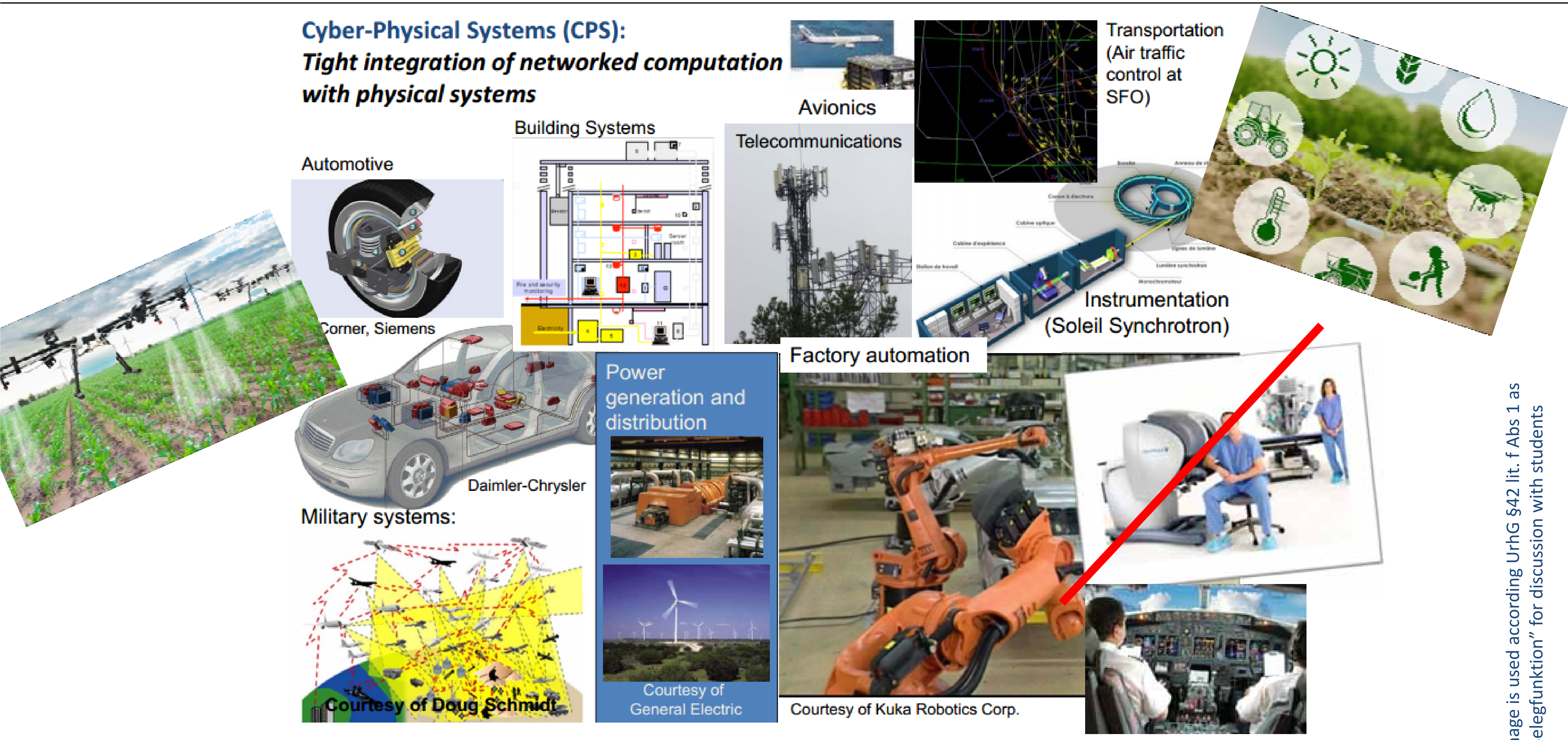


Alan Eckhardt 2009. Various aspects of user preference learning and recommender systems. DATESO. pp. 56-67.

Where can autonomous machine learning help ?

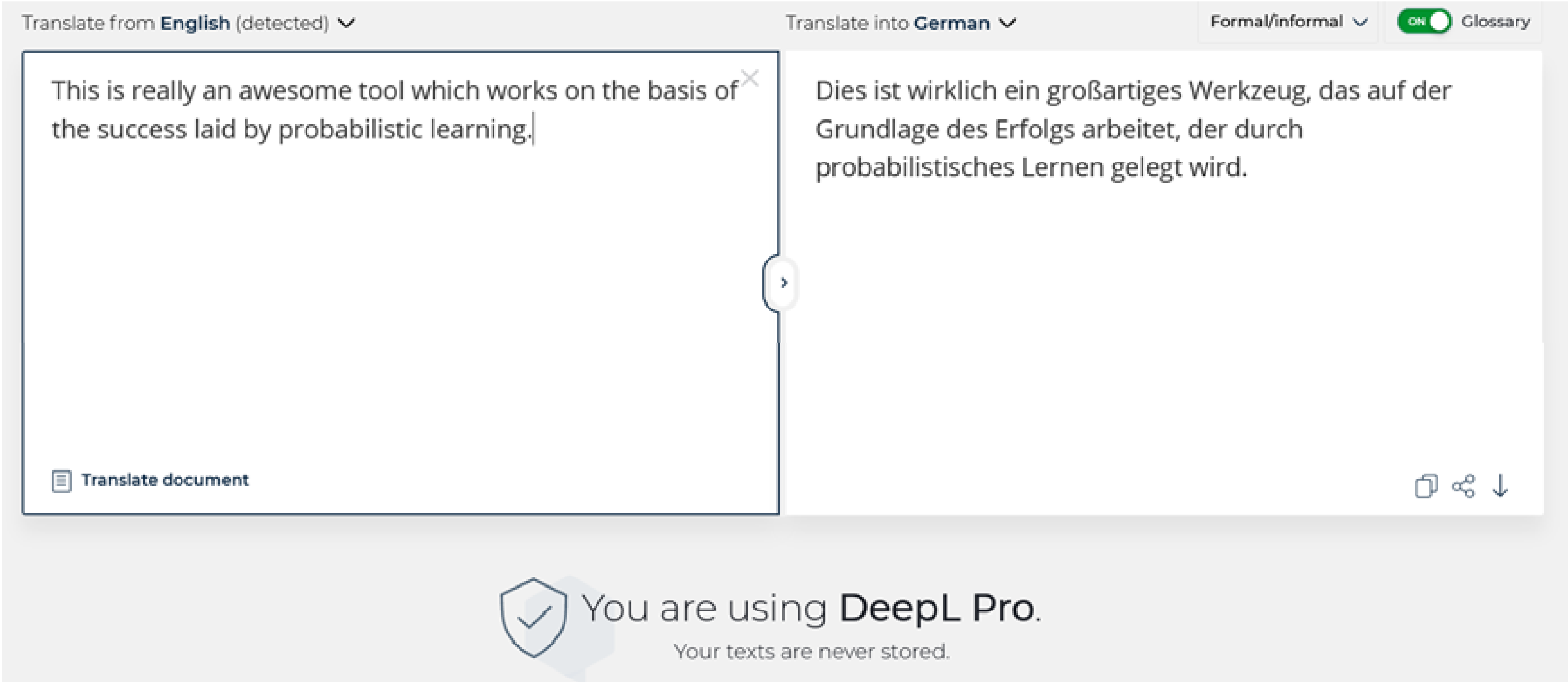
Cyber-Physical Systems (CPS):

Tight integration of networked computation with physical systems

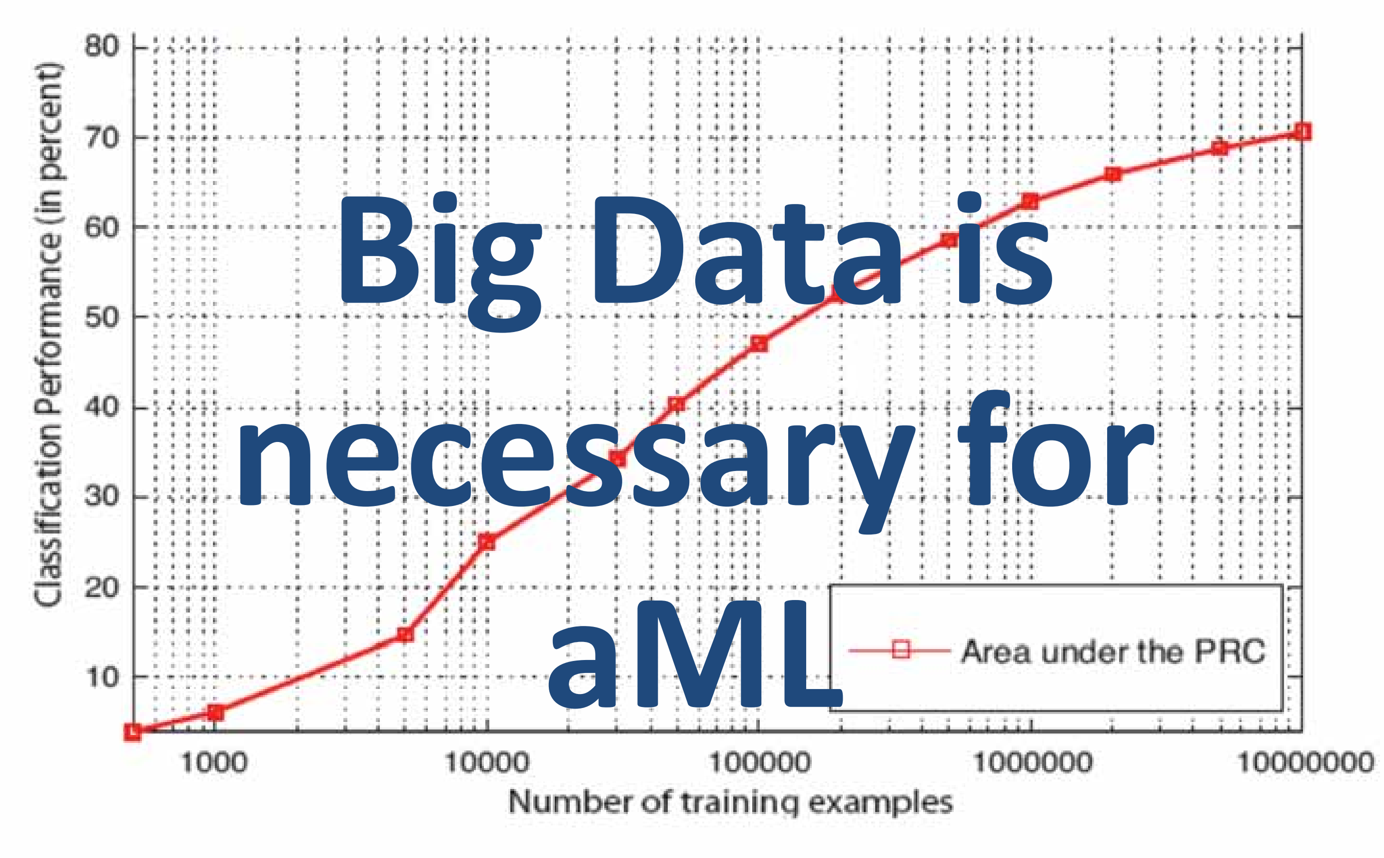


Seshia, S. A., Juniwal, G., Sadigh, D., Donze, A., Li, W., Jensen, J. C., Jin, X., Deshmukh, J., Lee, E. & Sastry, S. 2015. Verification by, for, and of Humans: Formal Methods for Cyber-Physical Systems and Beyond. Illinois ECE Colloquium.

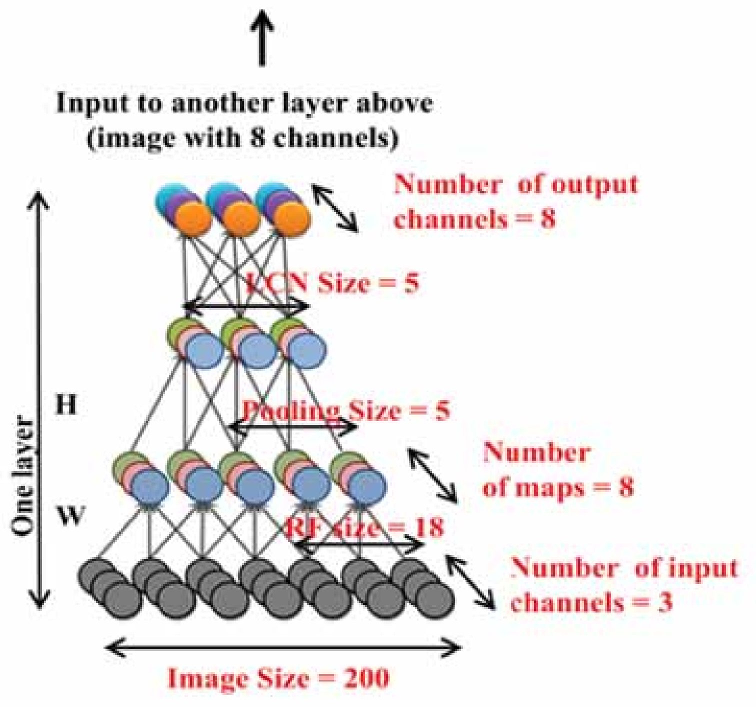
What is a best practice example for the success of Deep Learning ?



Dzmitry Bahdanau, Kyunghyun Cho & Yoshua Bengio (2014). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*.



Sonnenburg, S., Rätsch, G., Schäfer, C. & Schölkopf, B. 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.



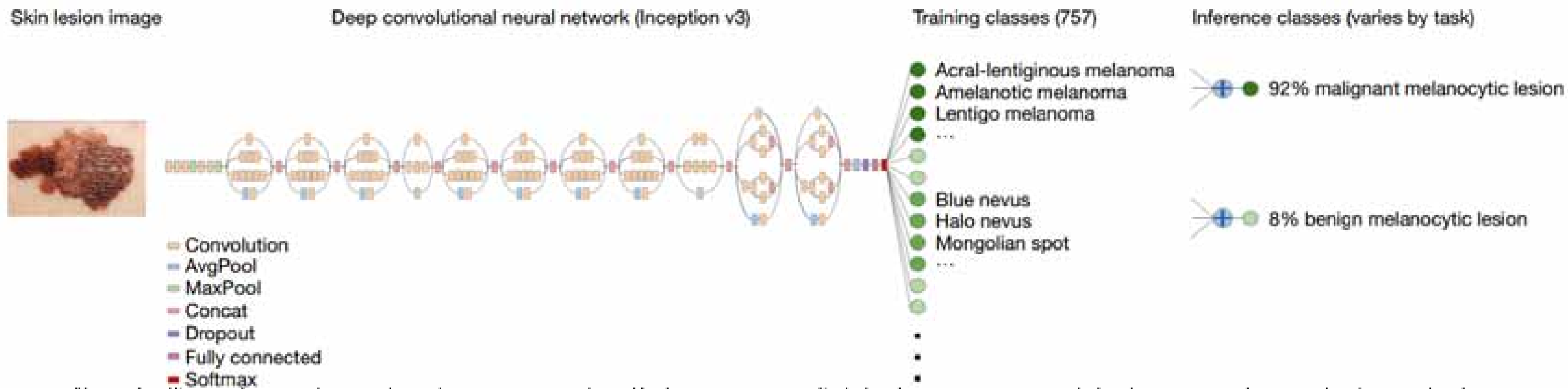
$$x^* = \arg\min_x f(x; W, H), \ \text{subject to } ||x||_2 = 1.$$

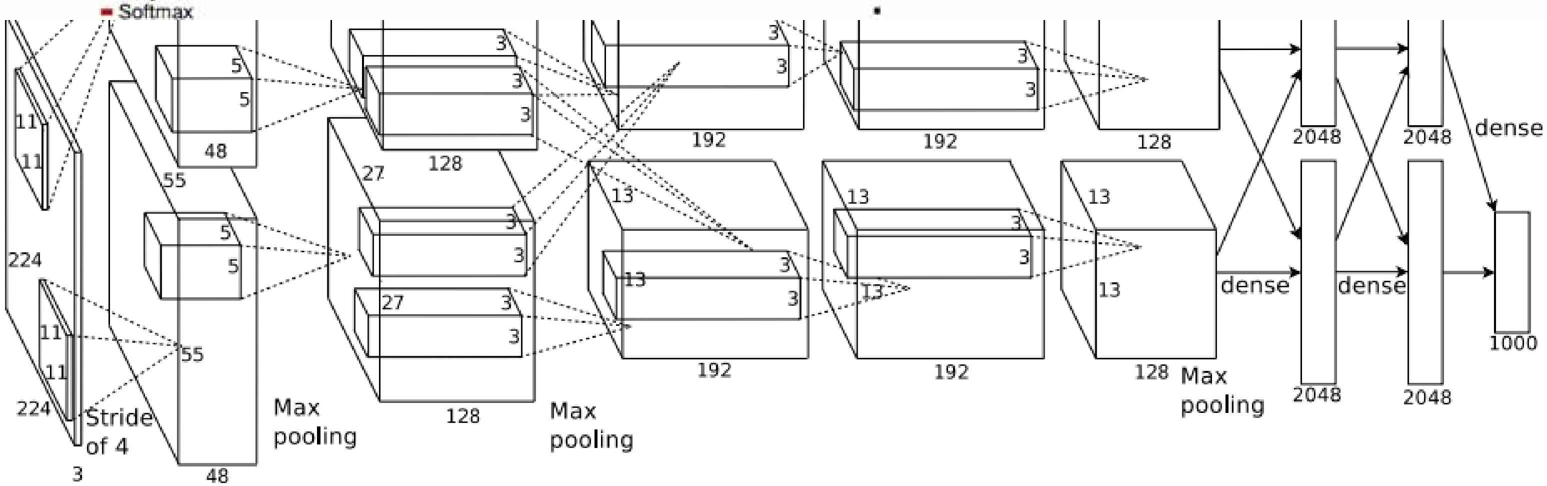
Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J. & Ng, A. Y. 2011. Building high-level features using large scale unsupervised learning. arXiv preprint arXiv:1112.6209.

Le, Q. V. 2013. Building high-level features using large scale unsupervised learning. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing ICASSP.* IEEE. 8595-8598, doi:10.1109/ICASSP.2013.6639343.

When does deep learning fail ?

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.





Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

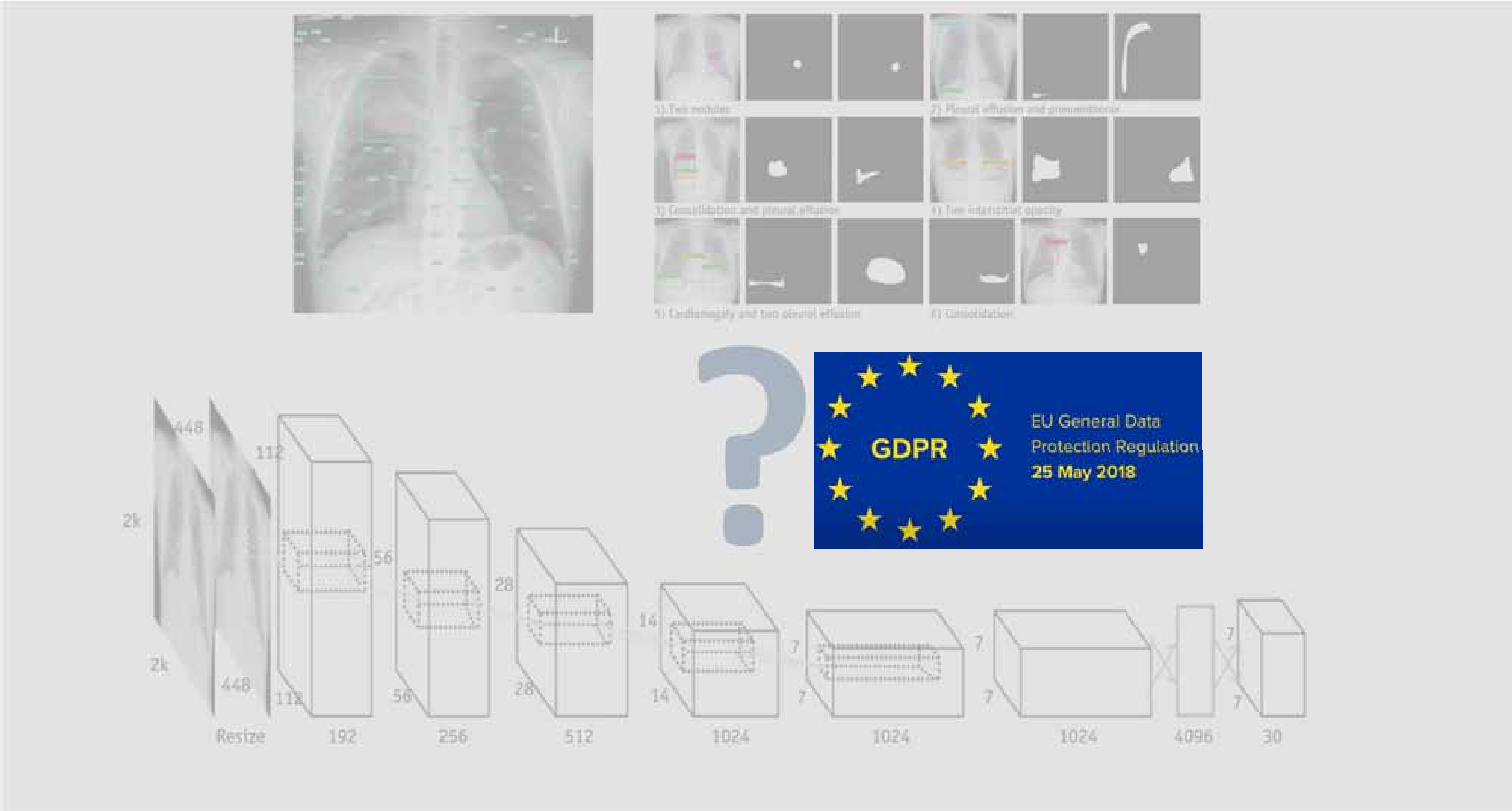
- Sometimes we **do not have "big data",** where aML-algorithms benefit.
- Sometimes we have
  - **Small amount of data sets**
  - Rare Events – **no training samples**
  - **NP-hard problems, e.g.**
    - Subspace Clustering,
    - k-Anonymization,
    - Protein-Folding, …

https://human-centered.ai/project/iml

What is now our big problem with deep learning approaches ?

- High dimensionality (curse of dim., many factors contribute)
- Complexity of medical problems (medical world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Sensitive to small disturbances (noise, bias, one-pixel attacks, …)
- 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
- However, most of all …

Why are such approaches difficult in terms of the EU GDPR ?



June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo & Namkug Kim 2017. Deep learning in medical imaging: general overview. Korean journal of radiology, 18, (4), 570-584, doi:10.3348/kjr.2017.18.4.570.

05 iML

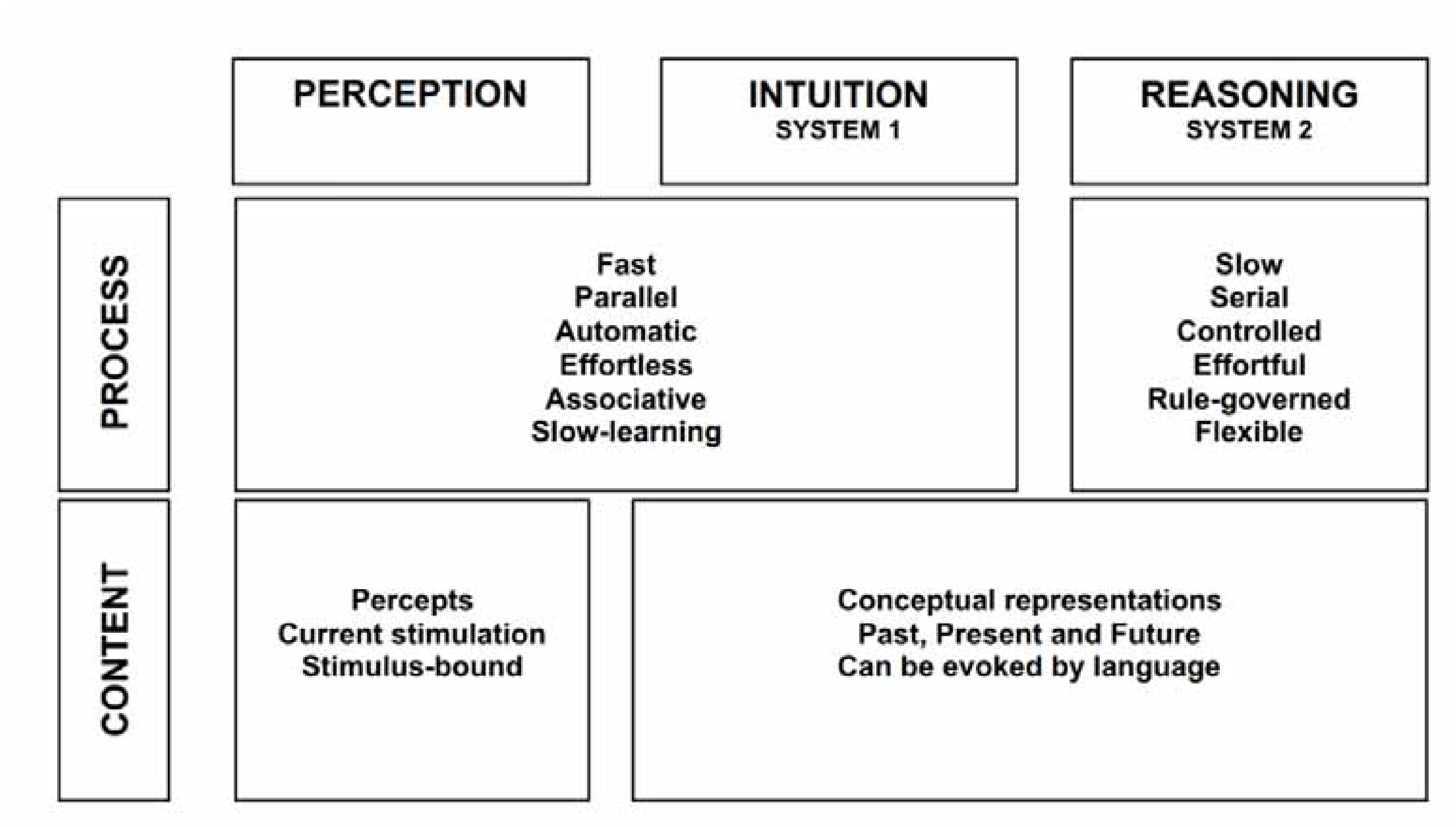
- iML := algorithms which interact with agents***)** and can optimize their learning behaviour through this interaction

***) where the agents can be human**

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

Why do we need – sometimes - a human-in-the-loop ?

Why are humans Bayesian reasoners ?



This was presented on December, 8, 2002 as Nobel Prize Lecture by Daniel Kahneman from Princeton University, an has later been published as:
Daniel Kahneman 2003. Maps of bounded rationality: Psychology for behavioural economics. American economic review, 93, (5), 1449-1475, doi:10.1257/000282803322655392.

What is Deductive vs Inductive vs Abductive Reasoning ?

- **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, conclusion: 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 …
  - 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.

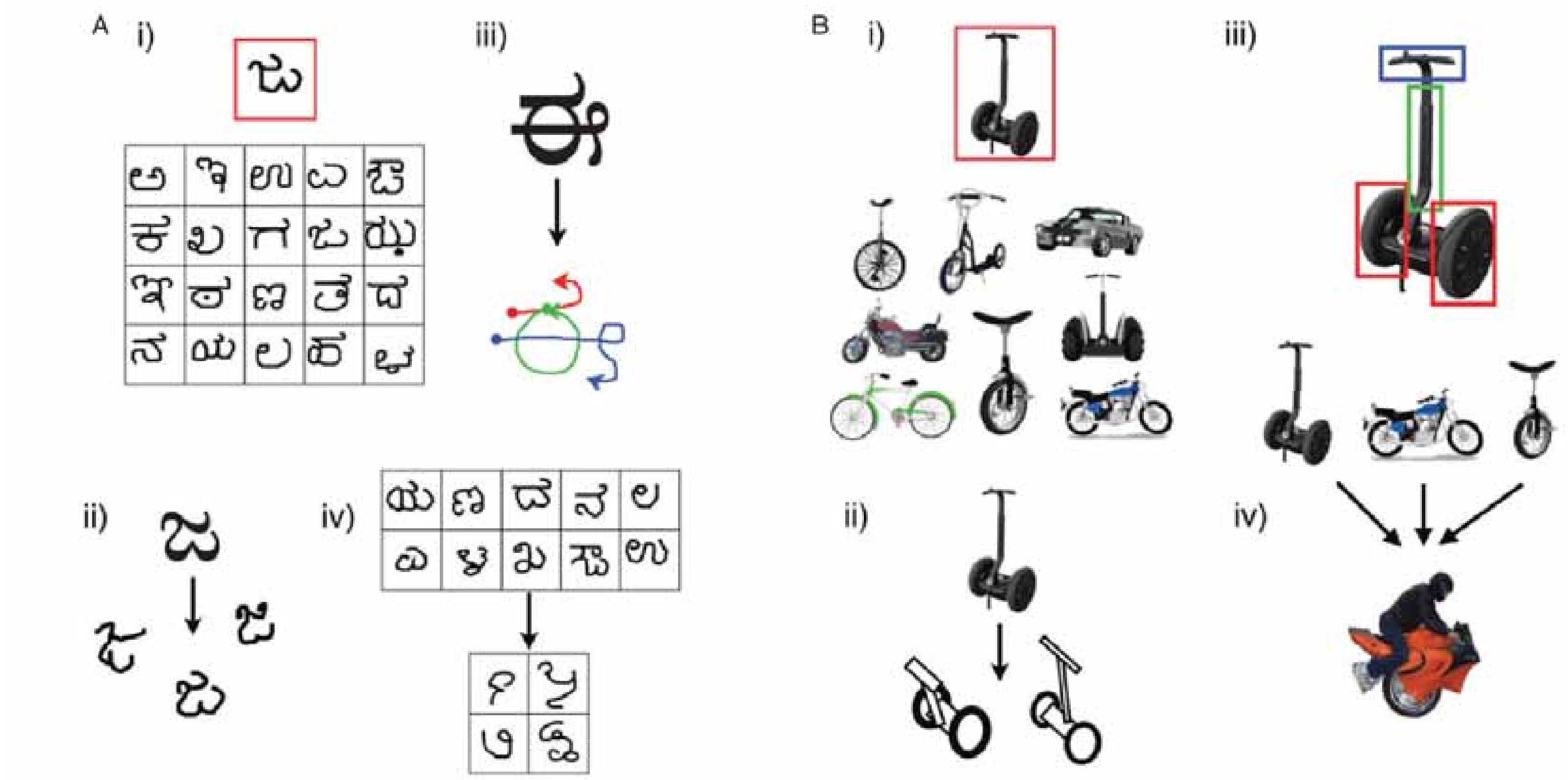
Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger 2018
Explainable AI: the new 42? Springer Lecture Notes in Computer Science LNCS 11015. pp. 295-303,doi:10.1007/978-3-319-99740-7_21.

What is 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.

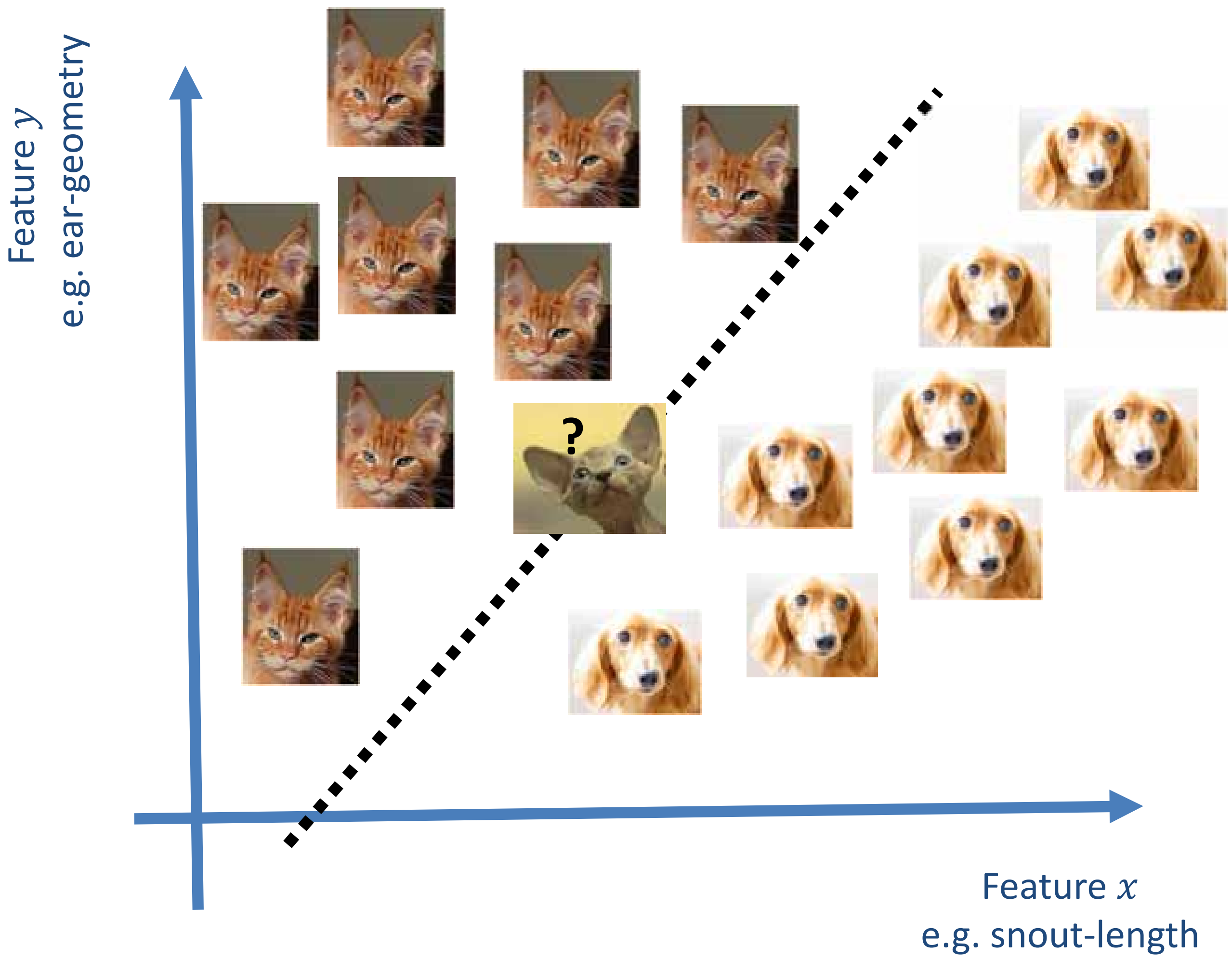
Judea Pearl, Madelyn Glymour & Nicholas P. Jewell (2016). Causal inference in statistics: A primer, John Wiley & Sons.

Even Children can make inferences from little, noisy, incomplete data …

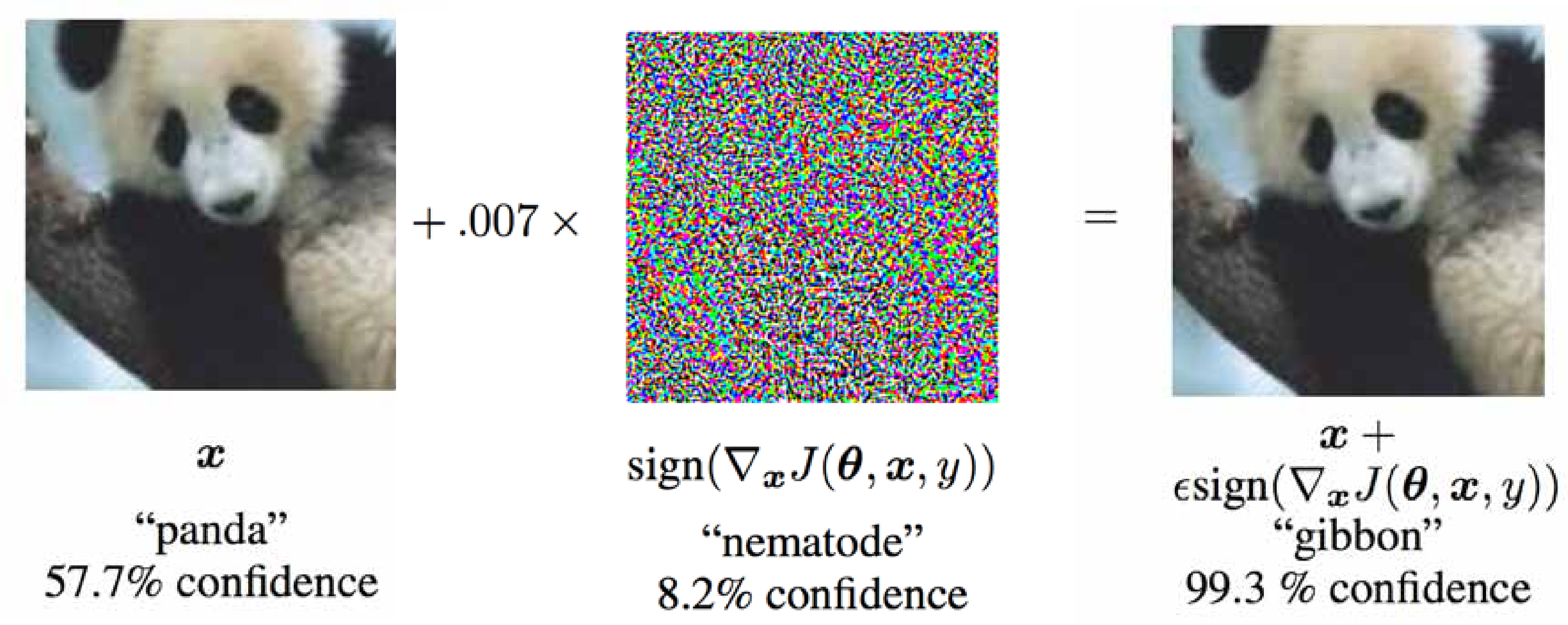


Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050

What can we do with rare examples in typical classification tasks ?



What is the difference between Correlation and Causation ?



Ian Goodfellow, Patrick Mcdaniel & Nicolas Papernot 2018. Making machine learning robust against adversarial inputs. *Communications of the ACM,* 61, (7), 55-66, doi:10.1145/3134599.

Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein 2018. Adversarial Examples that Fool both Human and Computer Vision. *arXiv:1802.08195*.

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

What is adversarial machine learning ?

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Gamaleldin F. Elsayed* Google Brain gamaleldin.elsayed@gmail.com

Shreya Shankar Stanford University

Brian Cheung UC Berkeley

Nicolas Papernot Pennsylvania State University

Alex Kurakin Google Brain

Ian Goodfellow Google Brain

Jascha Sohl-Dickstein Google Brain jaschasd@google.com

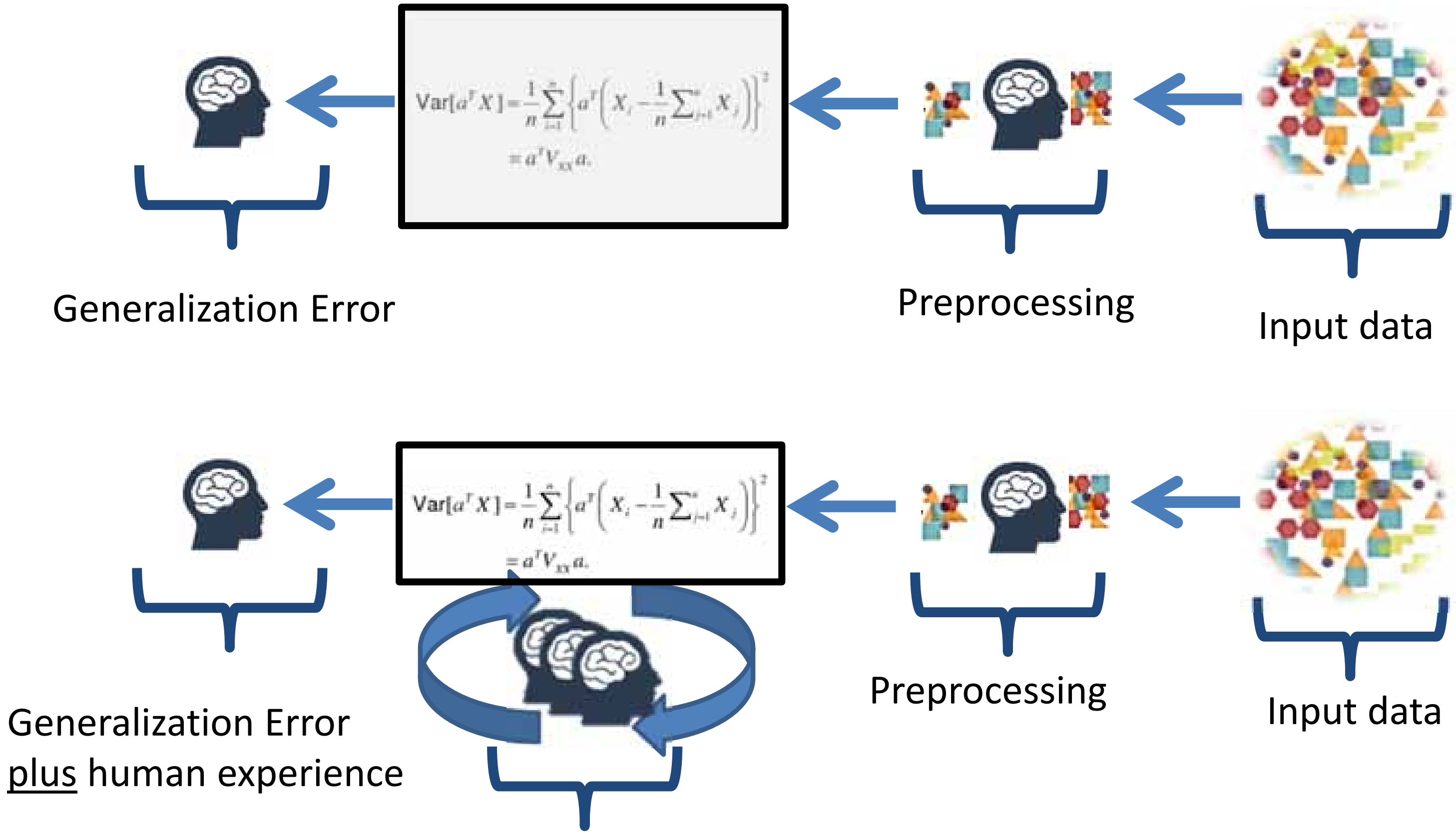
Abstract

Machine learning models are vulnerable to **adversarial examples**: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein 2018. Adversarial Examples that Fool both Human and Computer Vision. arXiv:1802.08195.

Correlation ≠ Causality Why we need the Human-in-the-loop

Where is the benefit of the human-in-the-loop ?

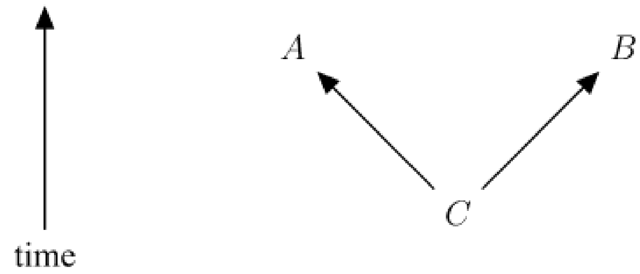


iML = human inspection – bring in human "intuition" – abstract concept learning and context understanding !

Andreas Holzinger 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? *Brain Informatics,* 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

Correlation does not tell anything about causality!

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



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

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

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

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

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

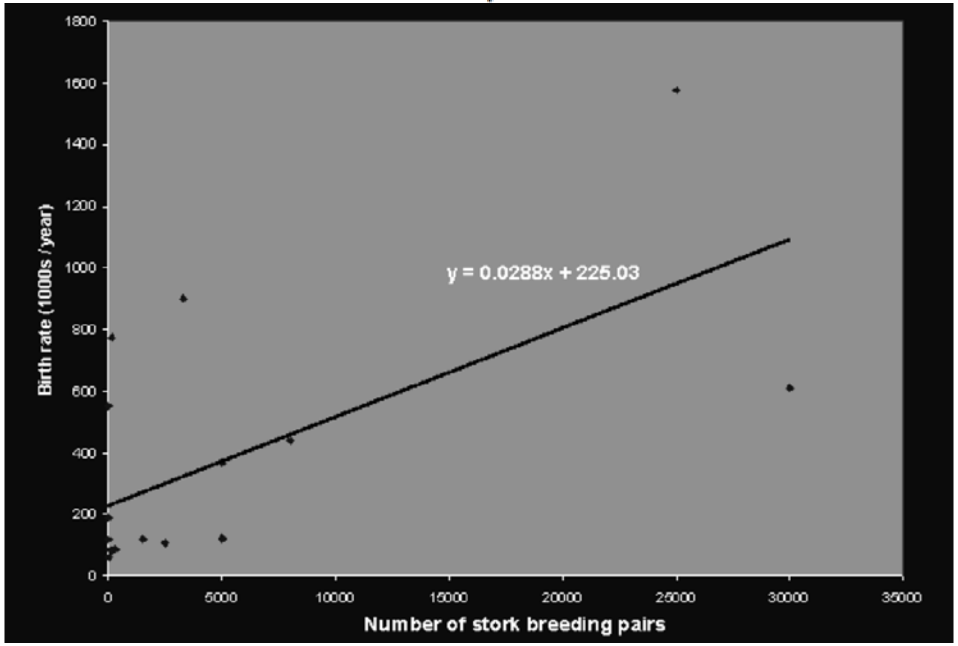
Remember: Correlation is NOT Causality

Storks Deliver Babies ($p = 0.008$)

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

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

Summary

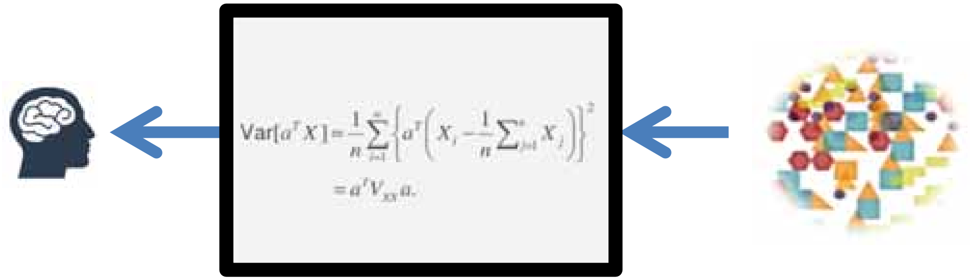


Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15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

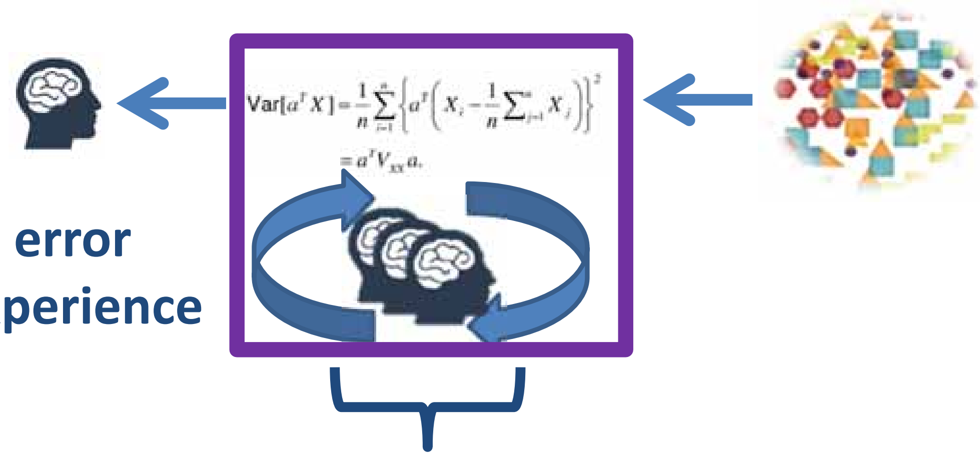
Table 1. Geographic, human and stork data for 17 European countries

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

How can we make AI more robust ?



Generalization error



Generalization error
<u>plus</u> human experience

iML = human inspection – bring in human conceptual knowledge

Andreas Holzinger et al. 2018. Interactive machine learning: experimental evidence for the human in the algorithmic loop. Springer/Nature Applied Intelligence, doi:10.1007/s10489-018-1361-5.

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

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

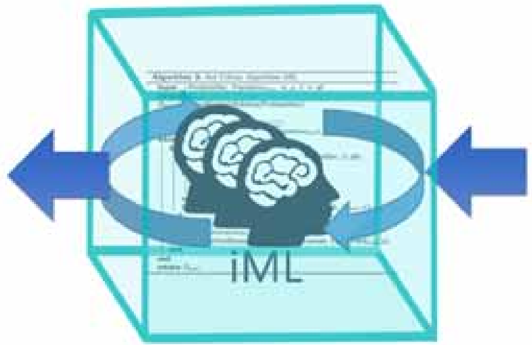
Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths & Noah D. Goodman 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

(6) explainable AI and Methods of Explainability

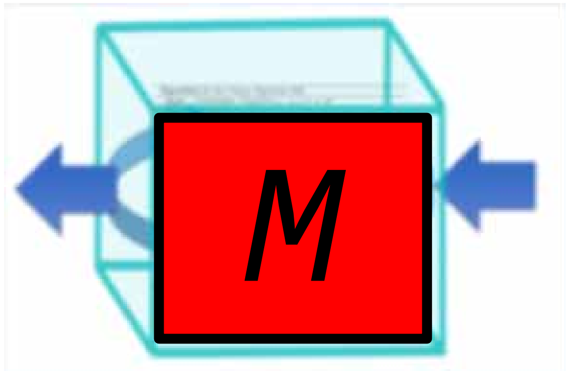
- **Trust** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?
- **Causality** - inferring causal relationships from pure observational data has been extensively studied (Pearl, 2009), however it relies strongly on prior knowledge
- **Transferability** – humans have a much higher capacity to generalize, and can transfer learned skills to completely new situations; compare this with e.g. susceptibility of CNNs to adversarial data (please remember that we rarely have iid data in real world
- **Informativeness** - for example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision
- **Fairness and Ethical decision making** – interpretations for the purpose of assessing whether decisions produced by algorithms conform to ethical standards, avoiding bias and misconceptions ..

Zachary C. Lipton 2016. The mythos of model interpretability. arXiv:1606.03490.

- **Interpretable Models, = ante-hoc -** the "glass-box" model itself is *ante-hoc* interpretable, e.g. Regression, Naïve Bayes, Decision Trees, Graphs, …

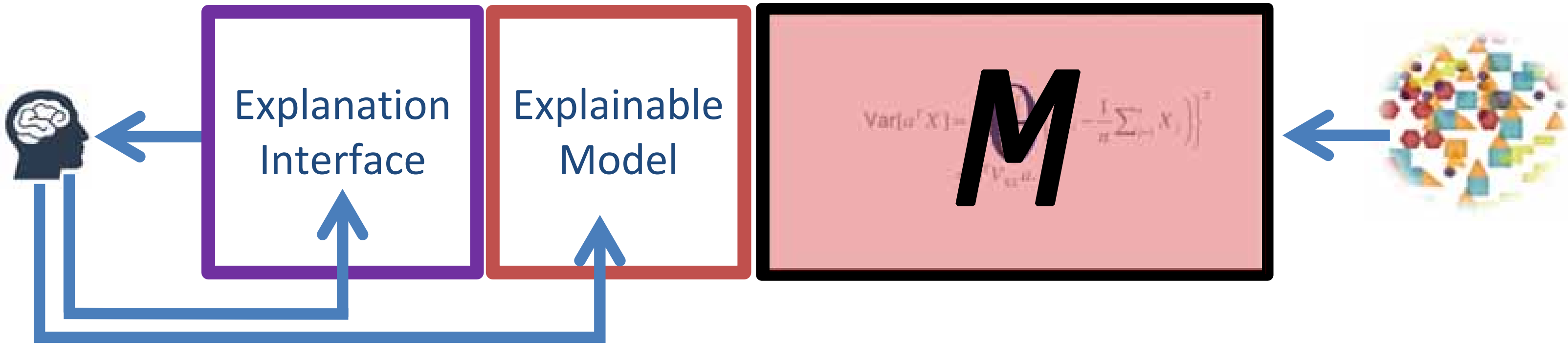


- **Interpreting Black-Box Models, = post-hoc -** the model is not interpretable and needs a post-hoc interpretability method ***M***



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

Why is explainable AI only a first step in human-centered AI ?



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.

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

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

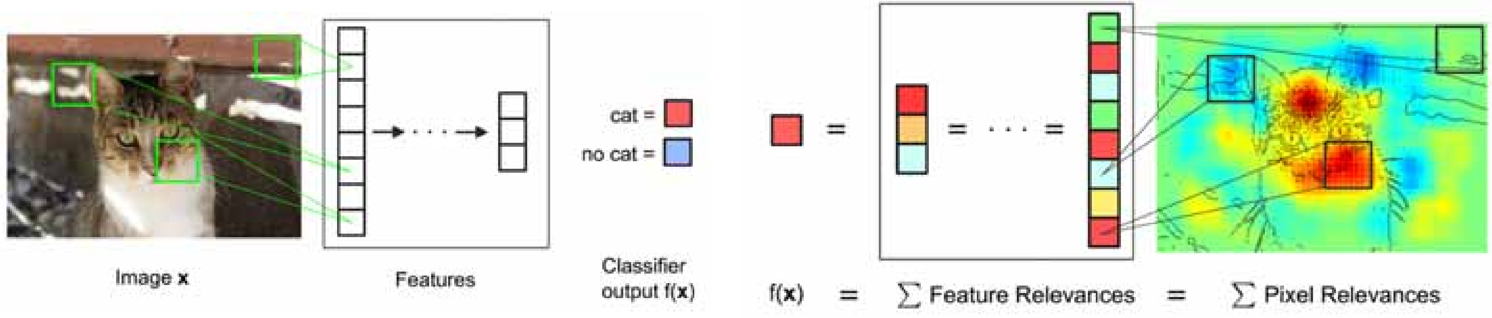
Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course
https://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

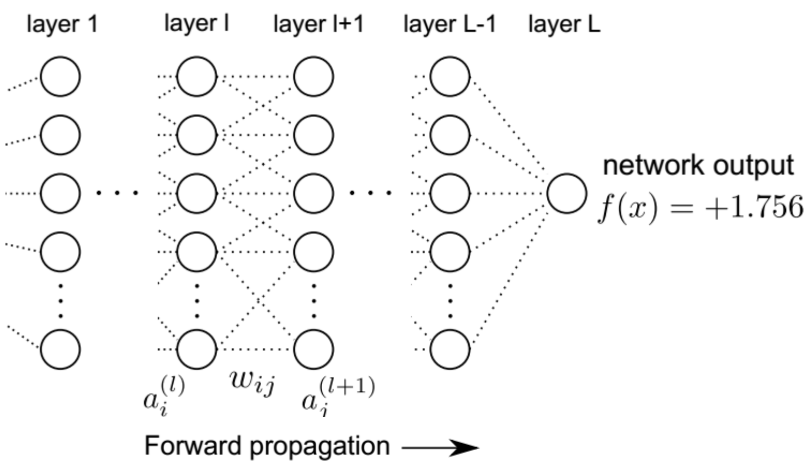
LRP Layer-Wise Relevance Propagation

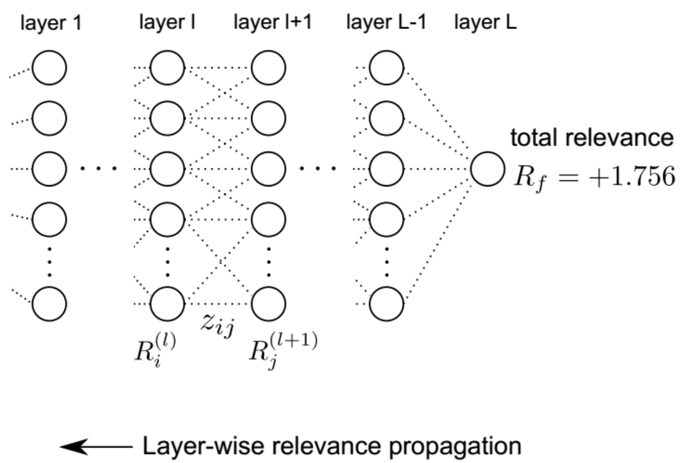
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

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

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



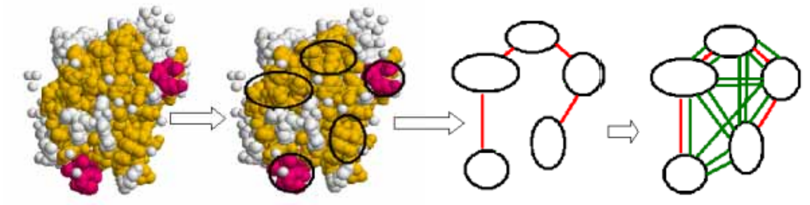




$$\boxed{a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)}$$

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

LRP works also on graphs

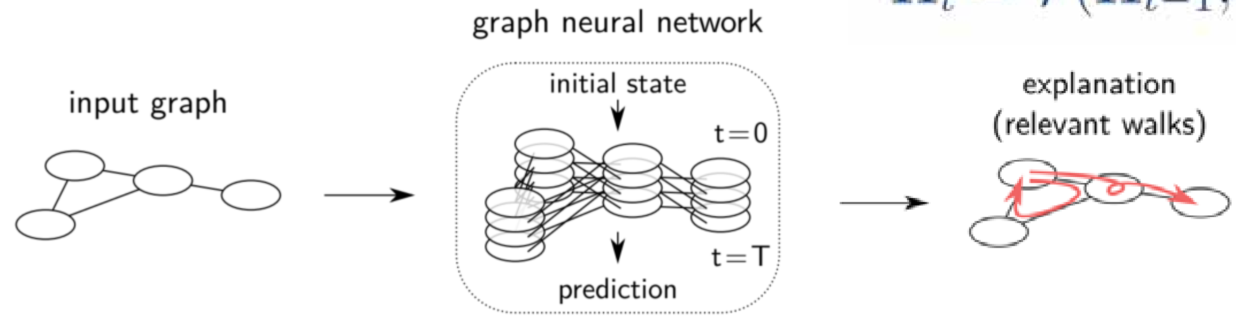


Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, Svn Vishwanathan, Alex J Smola & Hans-Peter Kriegel (2005). Protein function prediction via graph kernels. *Bioinformatics*, 21, (suppl 1), i47-i56.

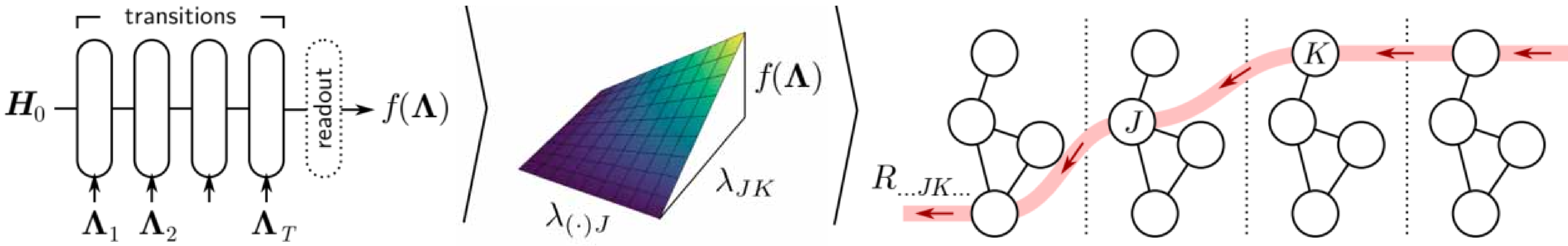
$\mathcal{G}$ ... input graph $\quad G = (\mathcal{V}, \mathcal{E}) \quad \mathcal{V} = \{v_1, \dots, v_n\} \quad \mathcal{E} \subseteq \{(v_i, v_j) | v_i, v_j \in \mathcal{V}\}$

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

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

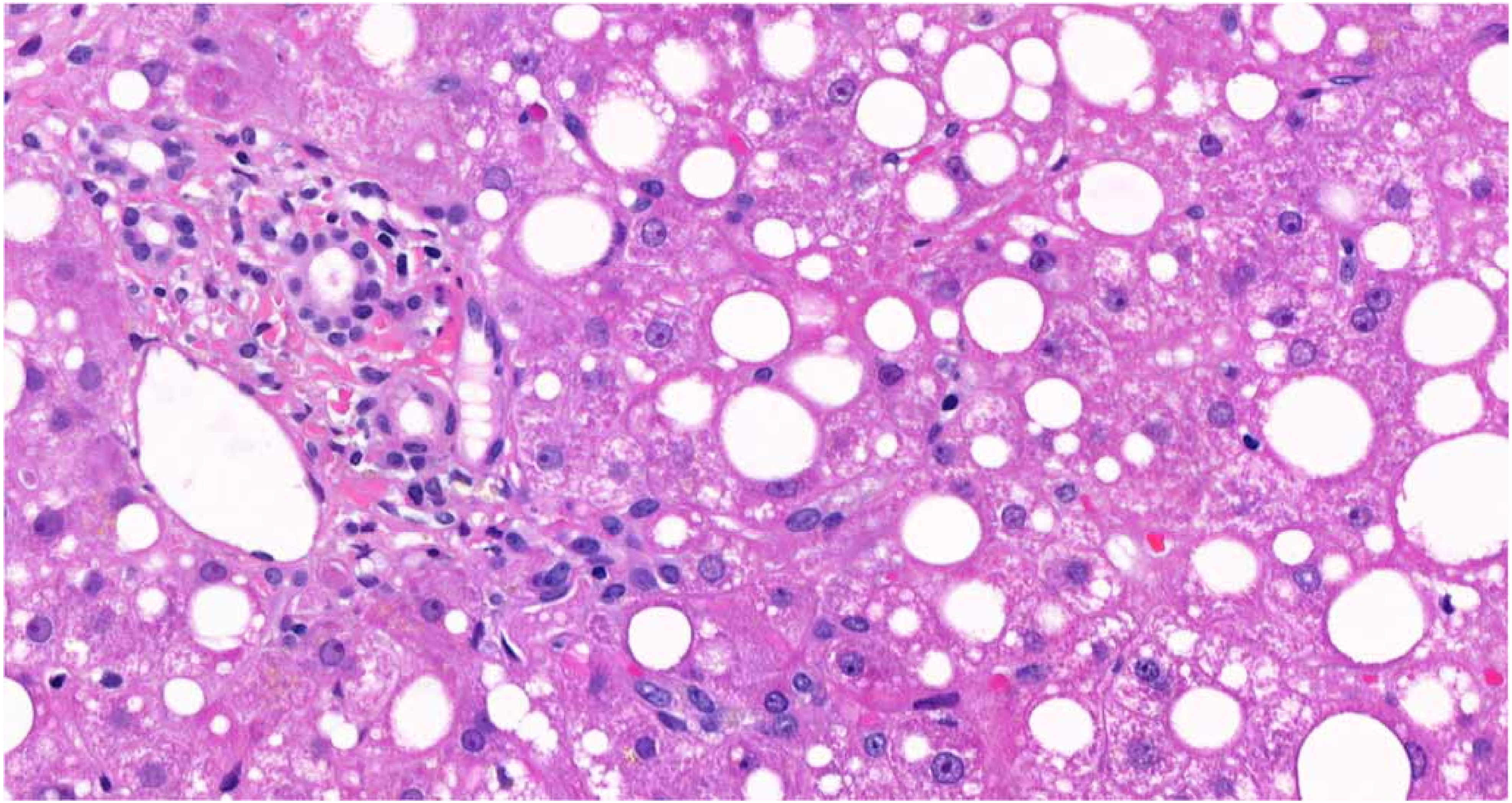


Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller & Grégoire Montavon (2020). XAI for Graphs: Explaining Graph Neural Network Predictions by Identifying Relevant Walks. *arXiv:2006.03589*.



(7) Causability measures the quality of explanations obtained from (6).

Example: How do human pathologists make diagnoses ?

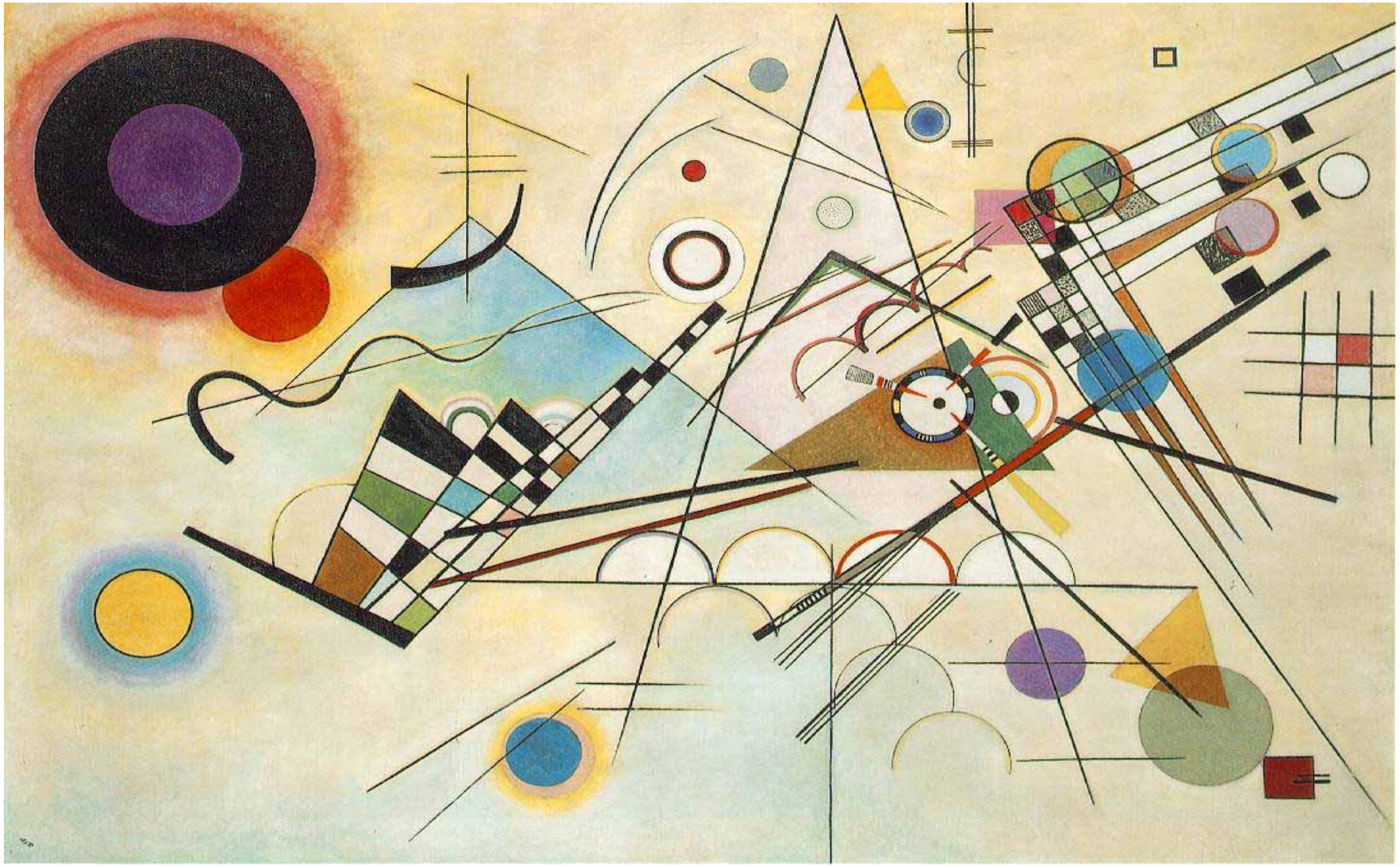


What is ground truth ? Where is the ground truth ?

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

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

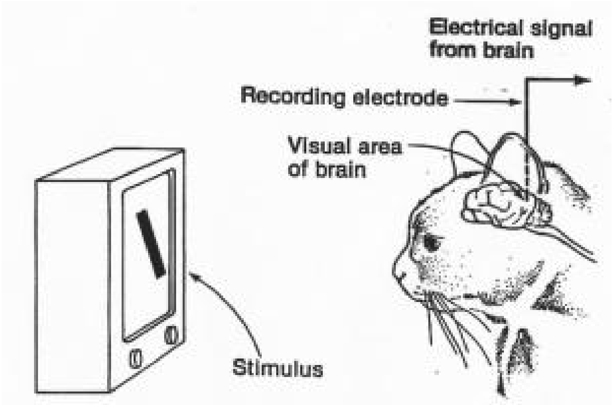
KANDINSKYPatterns – where does the name come from ?



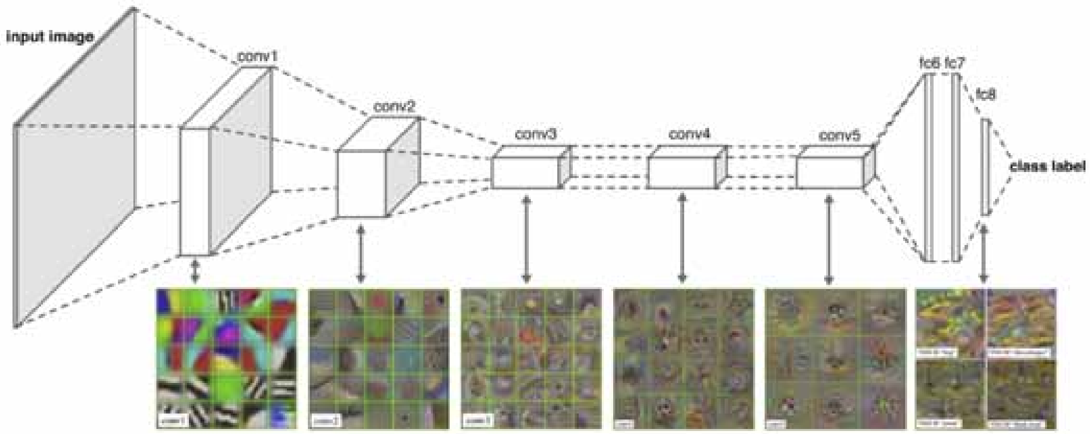
Wassily Kandinsky (1866 – 1944)

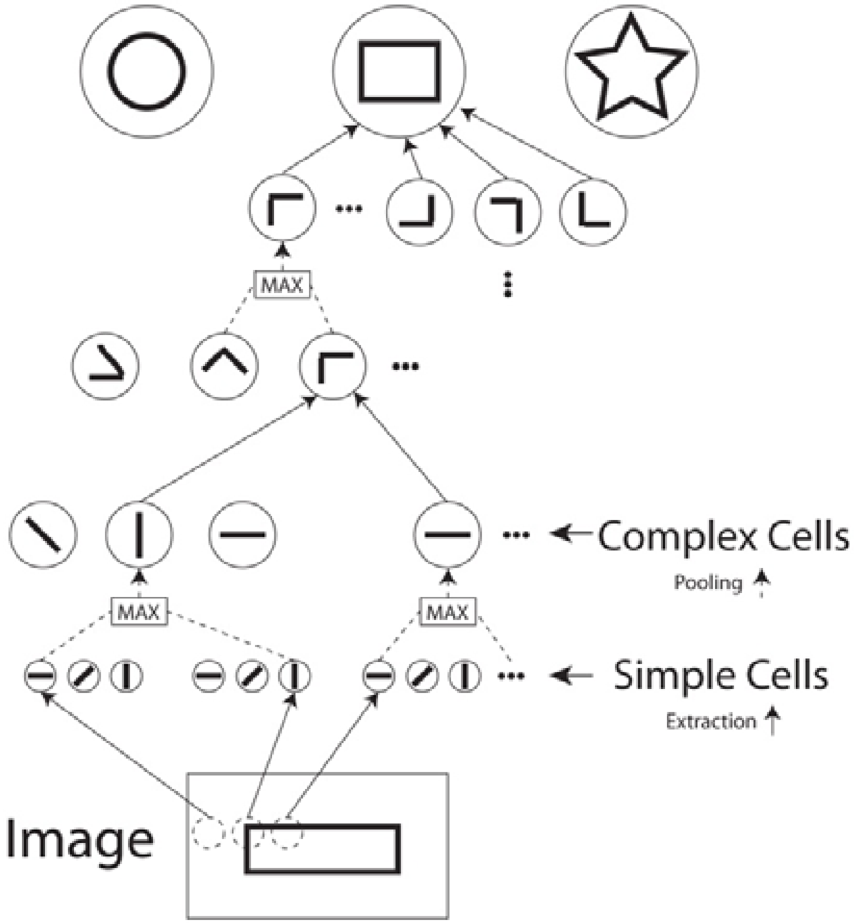
Komposition VIII, 1923, Solomon R. Guggenheim Museum, New York. Source: https://de.wikipedia.org/wiki/Wassily_Kandinsky
Note: Image is in the public domain and is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

Hubel & Wiesel (1962): Our world is compositional !



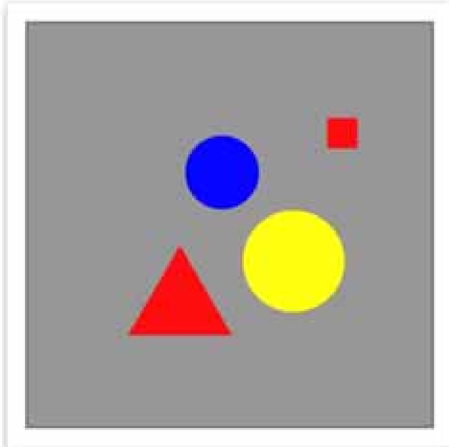
David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837





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

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.

Related Work (1): Bongard Problems

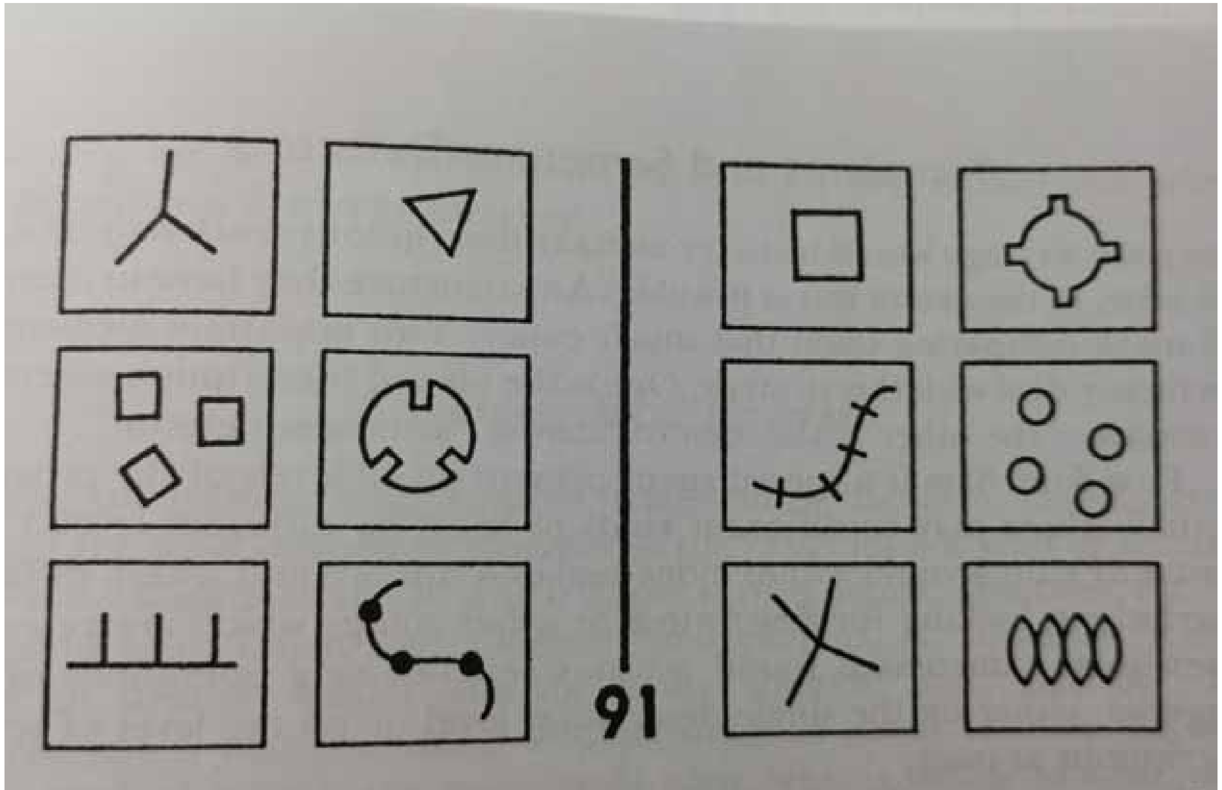
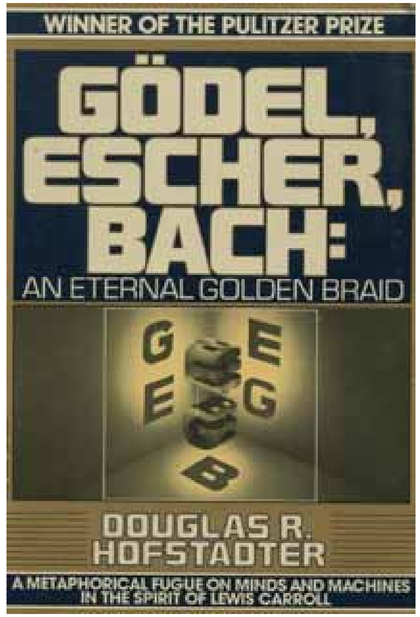


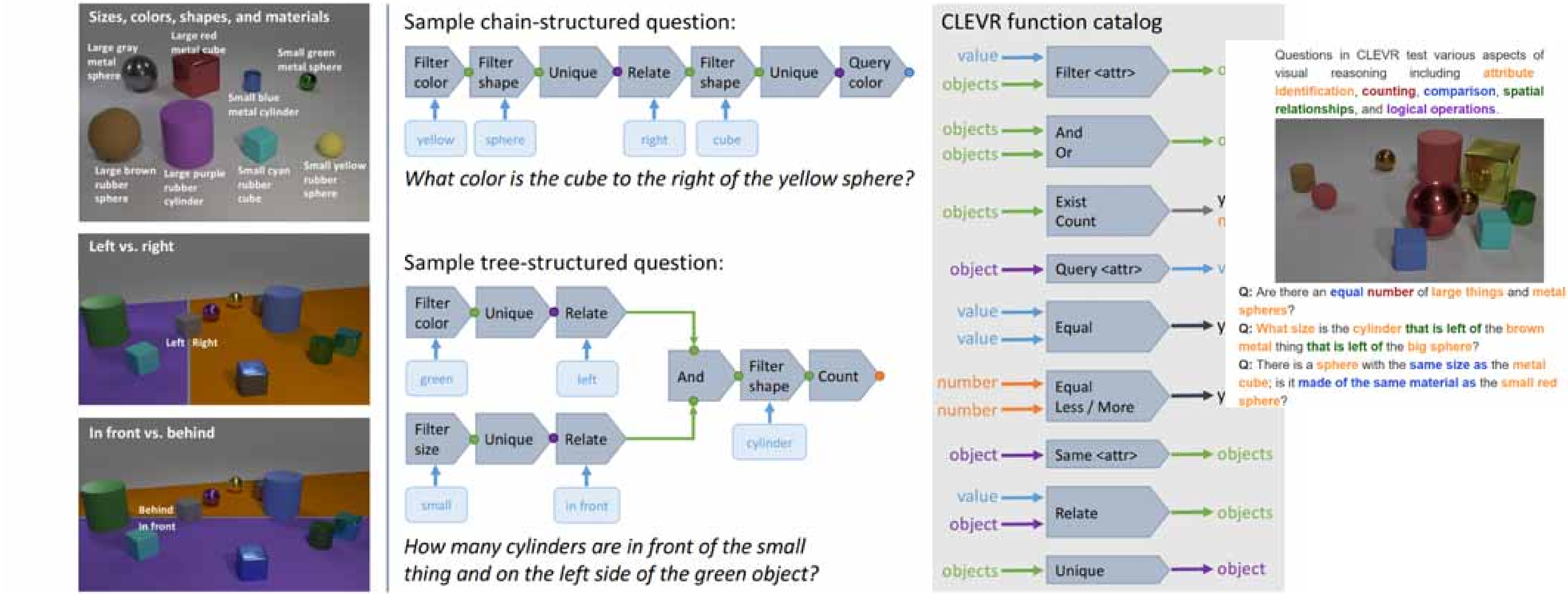
FIGURE 121. Bongard problem 91. [From M. Bongard, Pattern Recognition.]



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

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

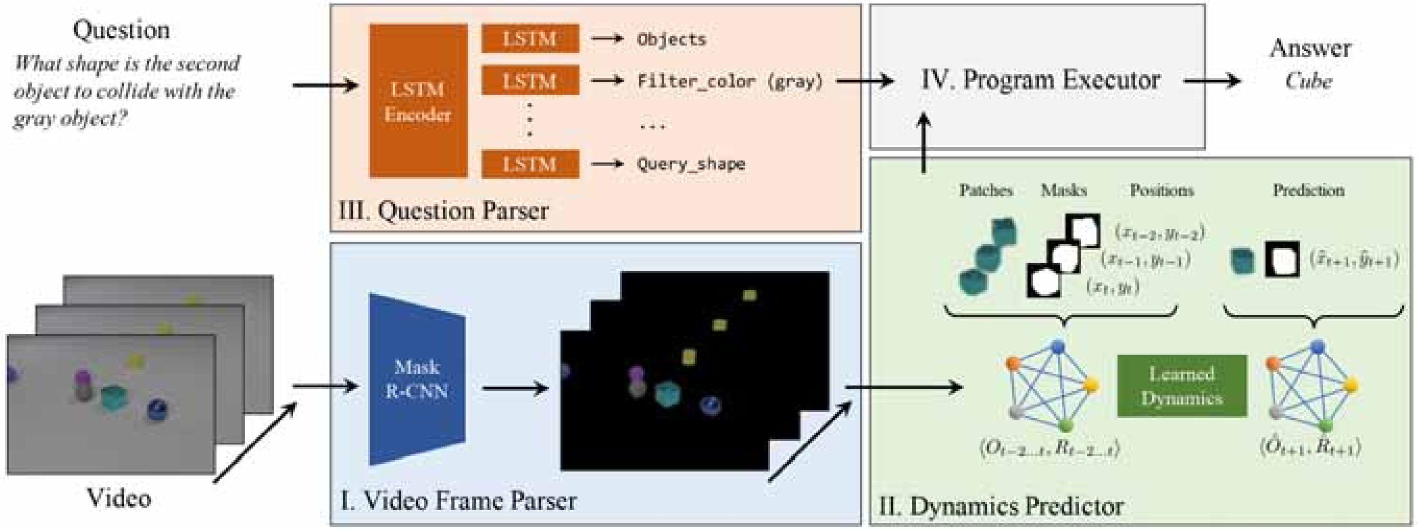
Related Work (2): CLEVR

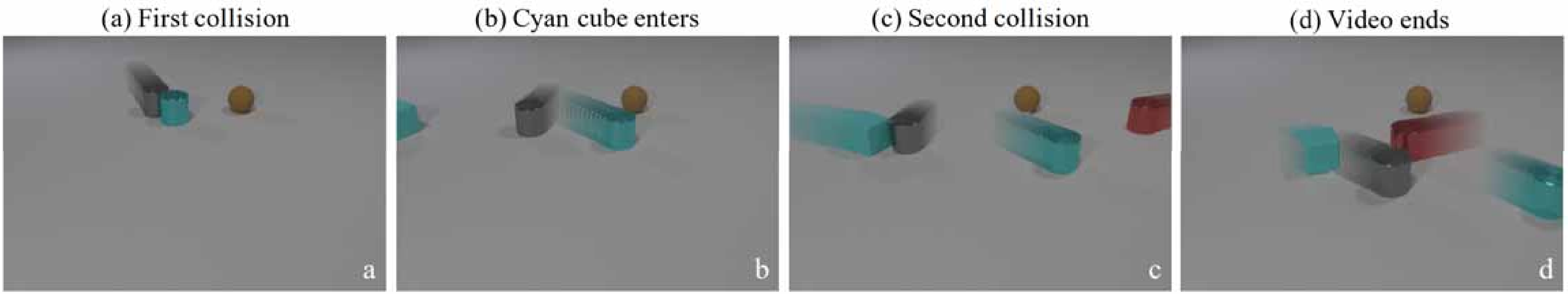


https://cs.stanford.edu/people/jcjohns/clevr/

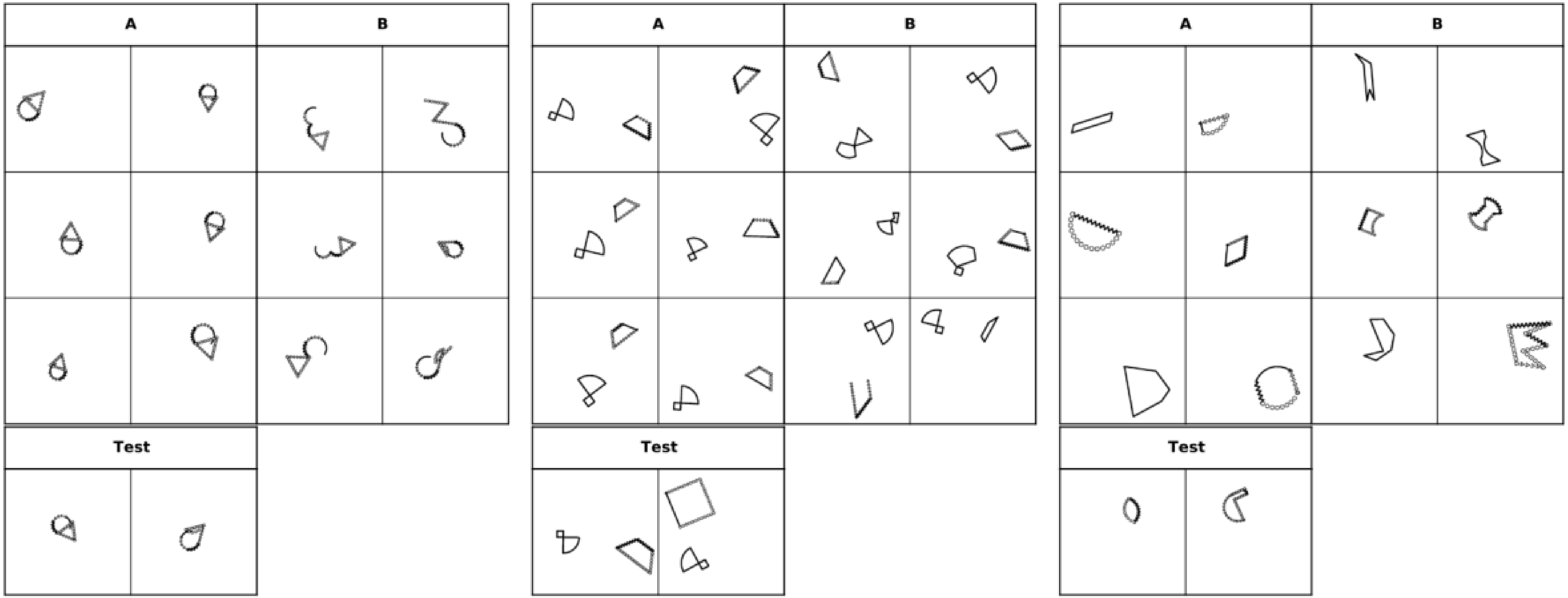
Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick & Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 Hawaii. IEEE.

Related Work (3): CLEVRER





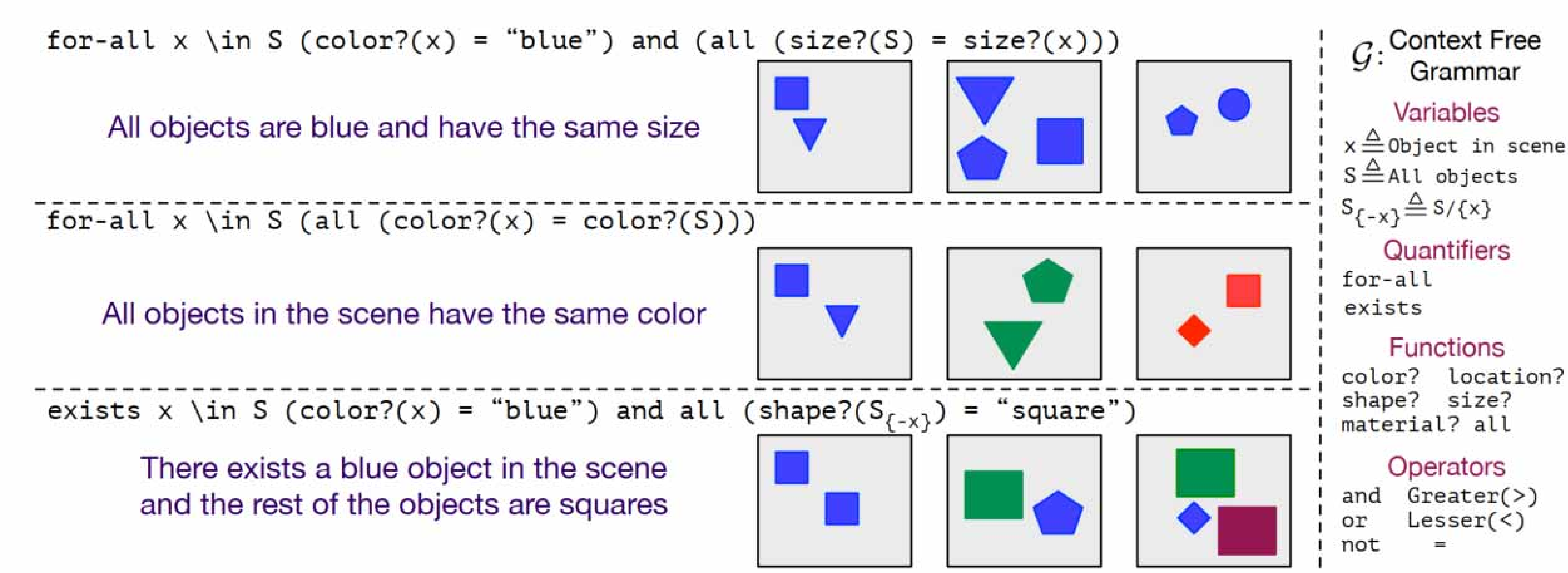
Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba & Joshua B. Tenenbaum (2019). CLEVRER: Collision events for video representation and reasoning. arXiv:1910.01442.



(a) free-from shape problem (b) basic shape problem (c) abstract shape problem

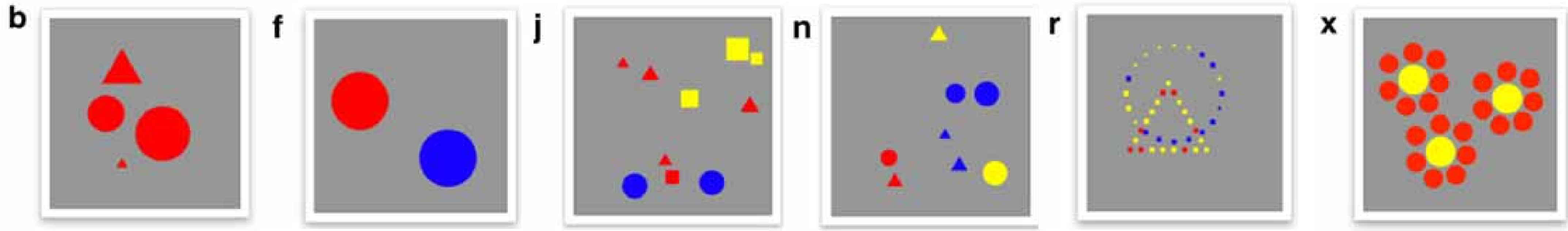
Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu & Anima Anandkumar (2020). BONGARD-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning. Advances in Neural Information Processing Systems, 33.

Related Work (5): CURI



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

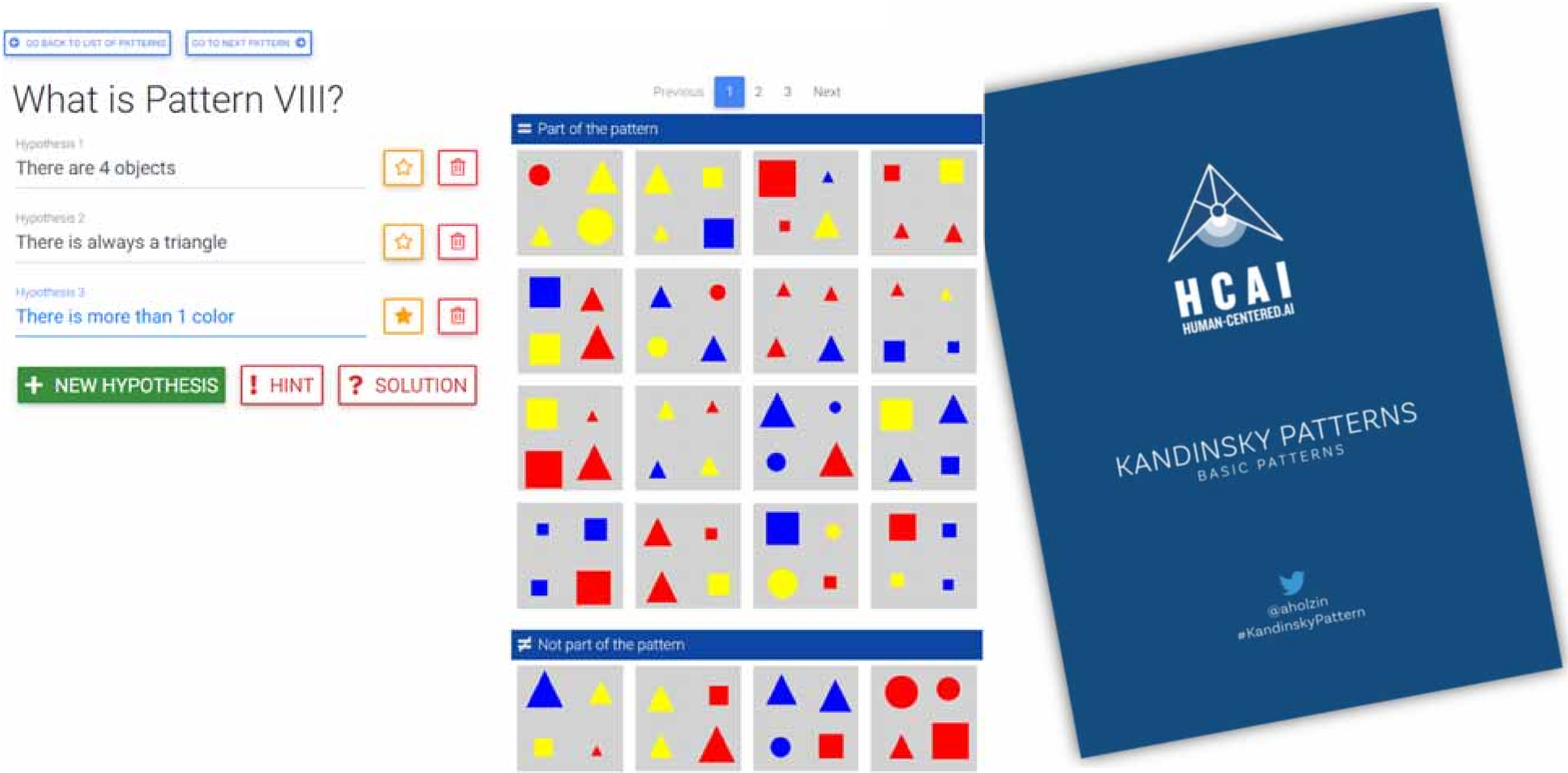
Definition 2 A statement $s(k)$



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

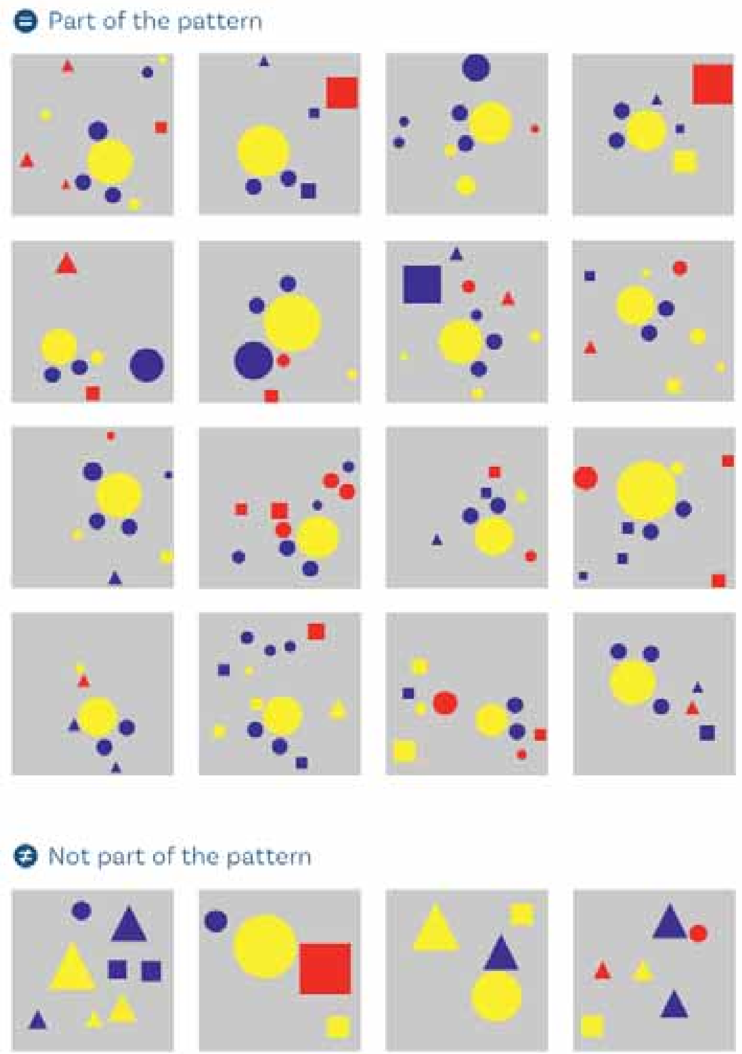
Holzinger, A. & Müller, H. 2020. Verbinden von Natürlicher und Künstlicher Intelligenz: eine experimentelle Testumgebung für Explainable AI (xAI). HMD Praxis der Wirtschaftsinformatik, 57, (1), 33-45, doi:10.1365/s40702-020-00586-y

“Intelligence Test for Machines”



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

Difficult Concepts



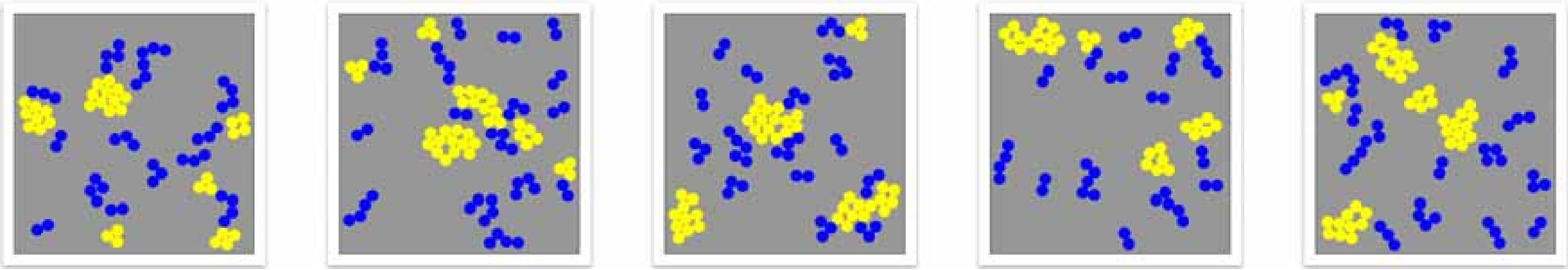
S8 Basic Pattern 8

Title: **Mickey Mouse** ->

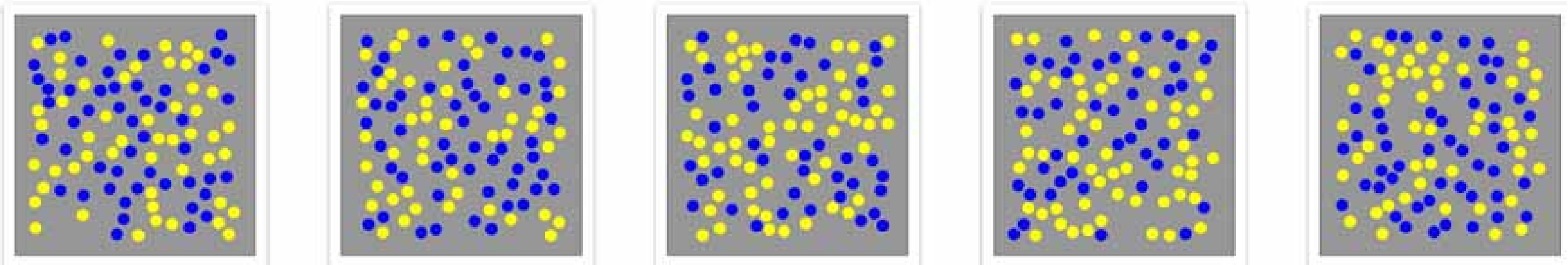
Every figure contains a pattern which is made out of a big yellow circle and two smaller blue ones and looks like a Mickey Mouse.

#KANDINSKYPatterns

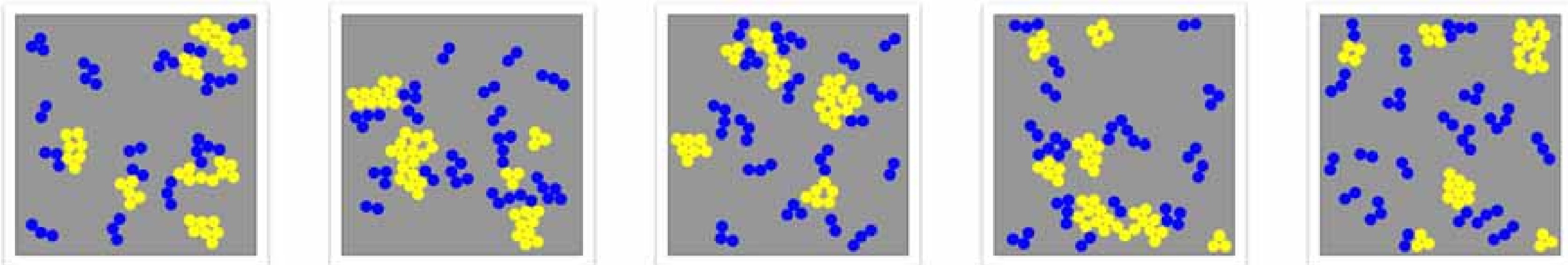
A) True (the cells are smaller and closer together – it is an tumor ...)



B) False



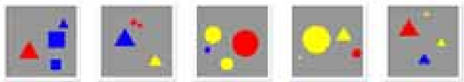
C) Counterfactual (What if the cells are slightly bigger ?)



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

HCAI HUMAN-CENTERED.AI

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



ABSTRACT

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

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

KANDINSKYPATTERNS AT TEDX

KANDINSKY ARTIFICIAL INTELLIGENCE EXPLANATION CHALLENGE

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

KANDINSKY HUMAN INTELLIGENCE EXPLANATION CHALLENGE

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

HCAI GITHUB REPOSITORY

OPEN STUDENTS THESES

Human-AI Interface DESIGNER
More Projects

LATEST NEWS

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

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

https://arxiv.org/abs/2103.00519

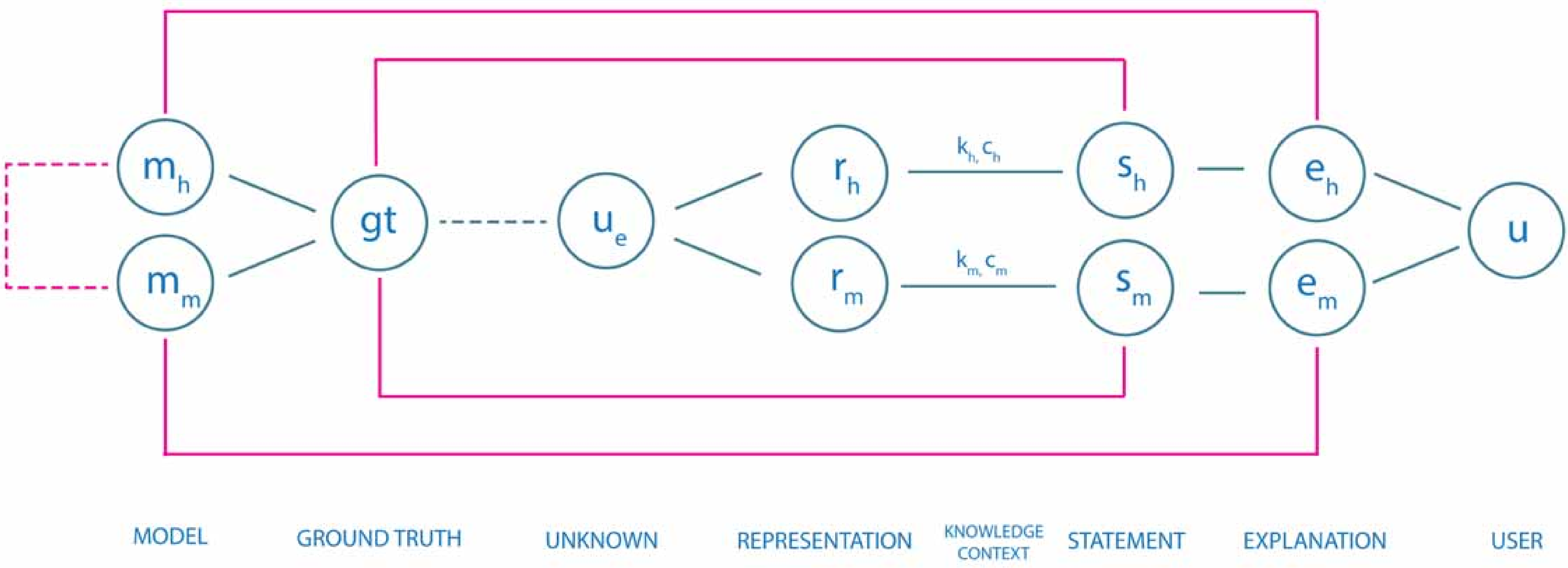
Measuring the quality of Explanations: The Systems Causa<u>bility </u>Scale

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

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

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

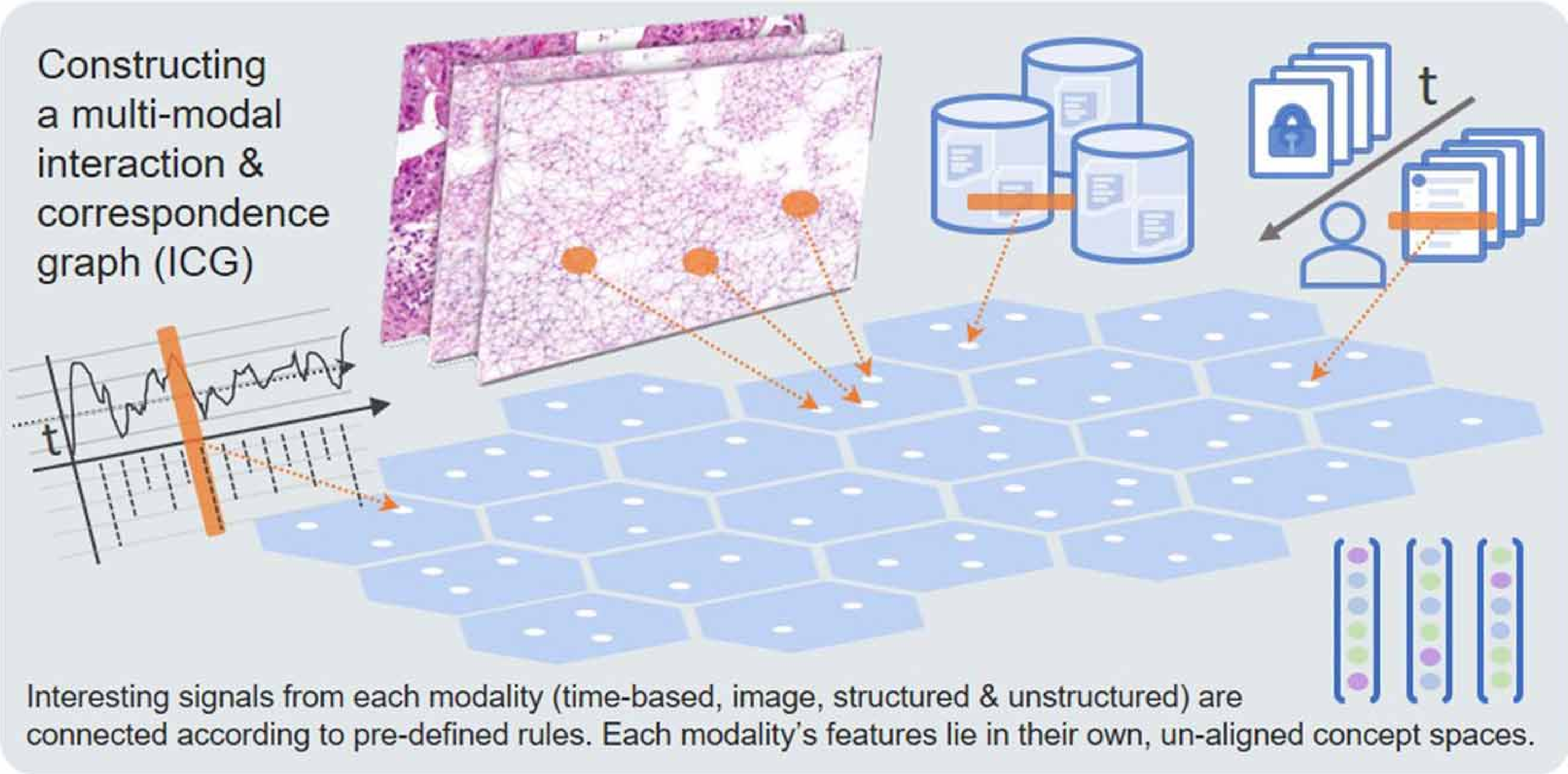
How can we measure the quality of explanations ?



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

Conclusio

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



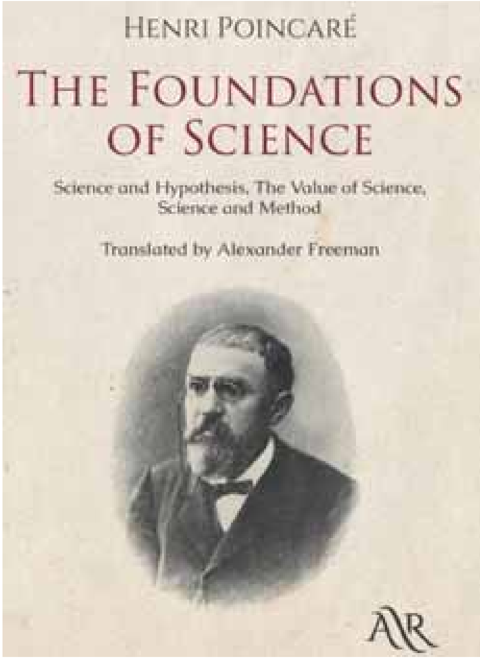
Andreas Holzinger, Bernd Malle, Anna Saranti & Bastian Pfeifer (2021). Towards Multi-Modal Causability with Graph Neural Networks enabling Information Fusion for explainable AI. Information Fusion, 71, (7), 28-37, doi:10.1016/j.inffus.2021.01.008.

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

Thank you very much !

Appendix

- *“The most interesting facts are*
- *those which can be used several times, those which have a chance of recurring …*
- *which, then, are the facts that have a chance of recurring?*
- *In the first place, **simple** facts.”*



Henri Poincare (1854-1912), Sciences et Methods (1908), (1913)