Assoc.Prof. Dr. Andreas Holzinger

185.A83 Machine Learning for Health Informatics
2020S, VU, 2.0 h, 3.0 ECTS
Andreas Holzinger, Marcus Bloice, Florian Endel, Anna Saranti
Lecture 01 - Week 12

From health informatics to ethical responsible medical AI

Contact: andreas.holzinger AT tuwien.ac.at

https://human-centered.ai/machine-learning-for-health-informatics-class-2020

What will we learn in this course ?

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

LV 185.A83 Machine Learning for Health Informatics (Class of 2020)

Andreas HOLZINGER, Marcus BLOICE, Florian ENDEL, Anna SARANTI

Study Code: 066 936 Master program Medical Informatics

https://tiss.tuwien.ac.at/curriculum/public/curriculum.xhtml?dswid=9468&dsrid=253&key=56089&semester=NEXT

Semester hours: 2.0 h; ECTS-Credits: 3.0; Type: VU Lecture and Exercises with Python

ECTS-Breakdown (sum=75 h, corresponds with 3 ECTS, where 1 ECTS = 25 h workload):

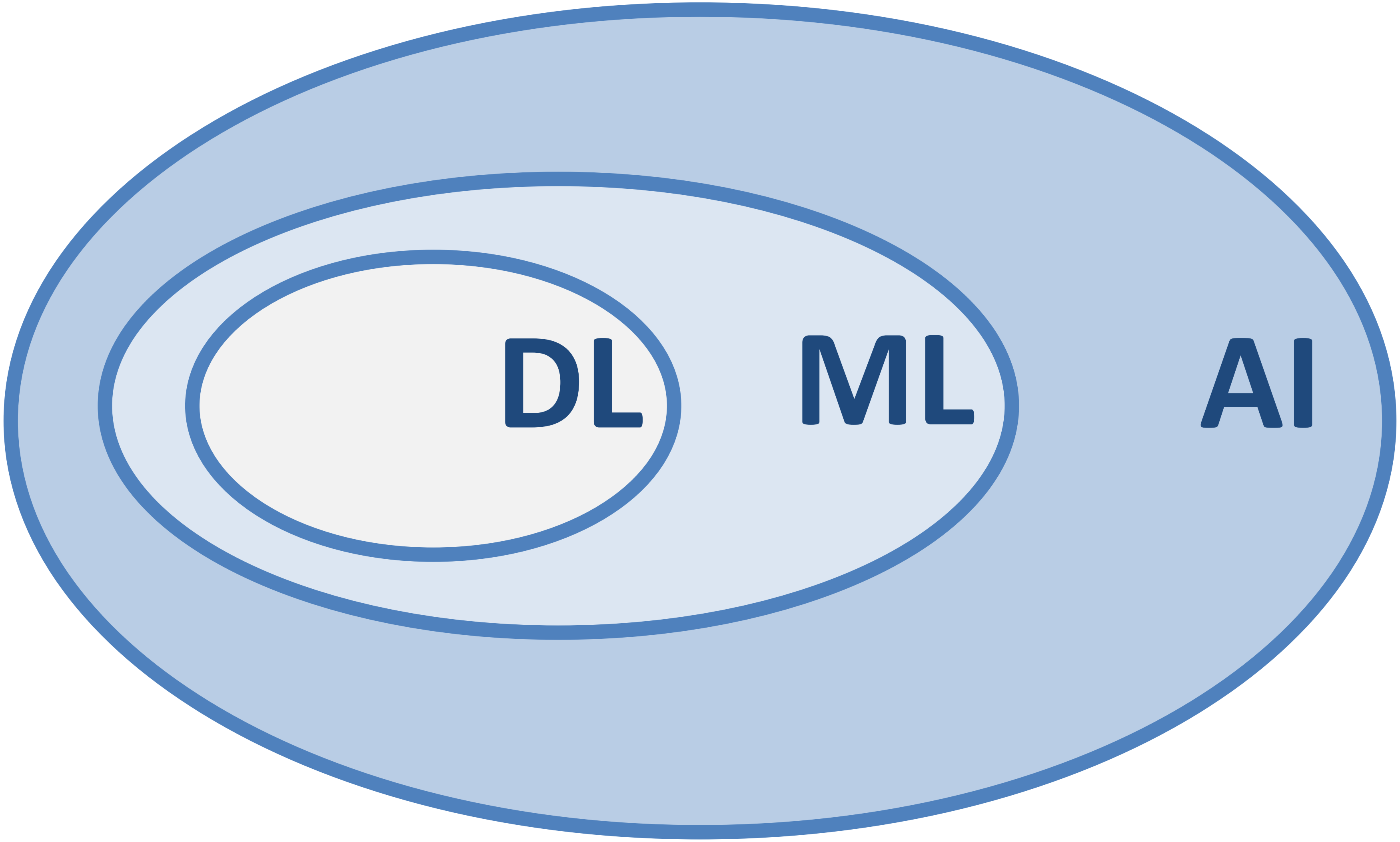
Presence during lecture	8 * 3 h	24 h
Preparation before and after lecture	8 * 1 h	08 h
Preparation of assignments and presentation	28 h + 2 h	30 h
Written exam including preparation	1 h + 12 h	13 h
TOTAL students' workload		**75 h**

Class URL: https://human-centered.ai/machine-learning-for-health-informatics-class-2020

Class Schedule for 2020 (subject to change: please check class URL for any changes):

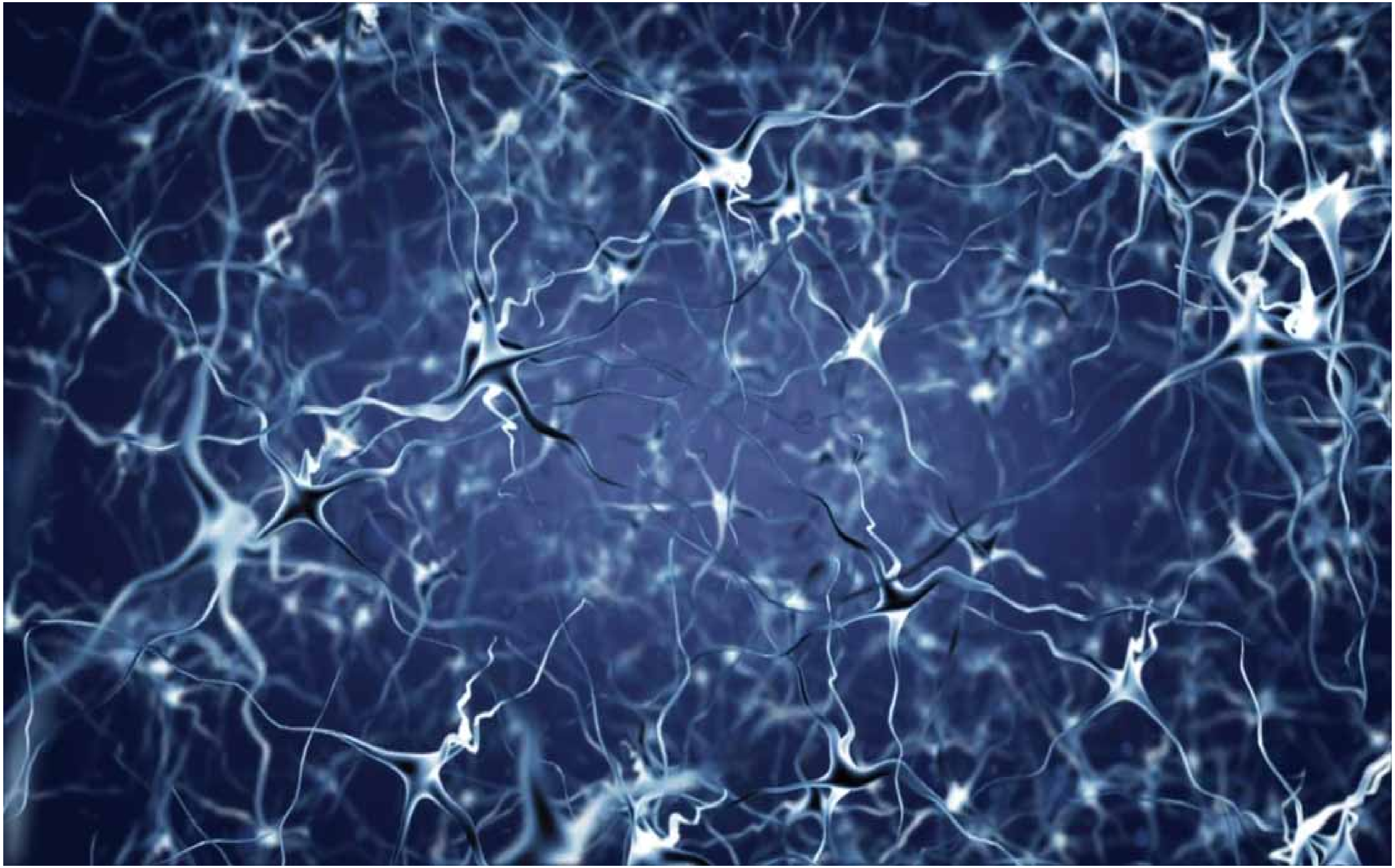
Nr	*Week*	*Topic*
01	12	Introduction and overview: From health informatics to ethical responsible medical AI
02	13	Data for machine learning, Probabilistic information and entropy

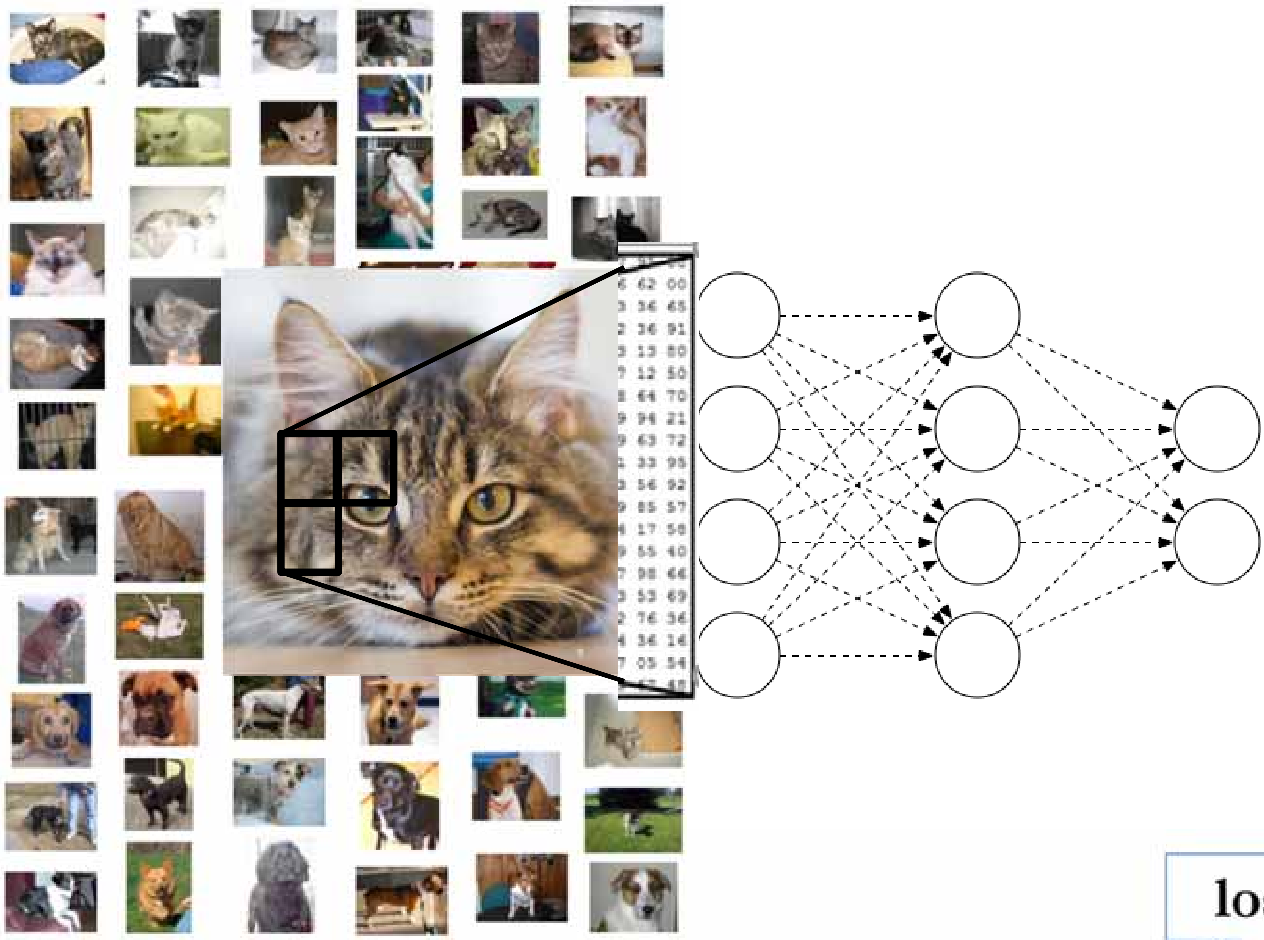
Before we start …



Andreas Holzinger, Peter Kieseberg, Edgar Weippl & A Min Tjoa 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. Springer Lecture Notes in Computer Science LNCS 11015. Cham: Springer, pp. 1-8, doi:10.1007/978-3-319-99740-7_1

Are artificial neural networks the same as biological neural networks ?





92 % cat

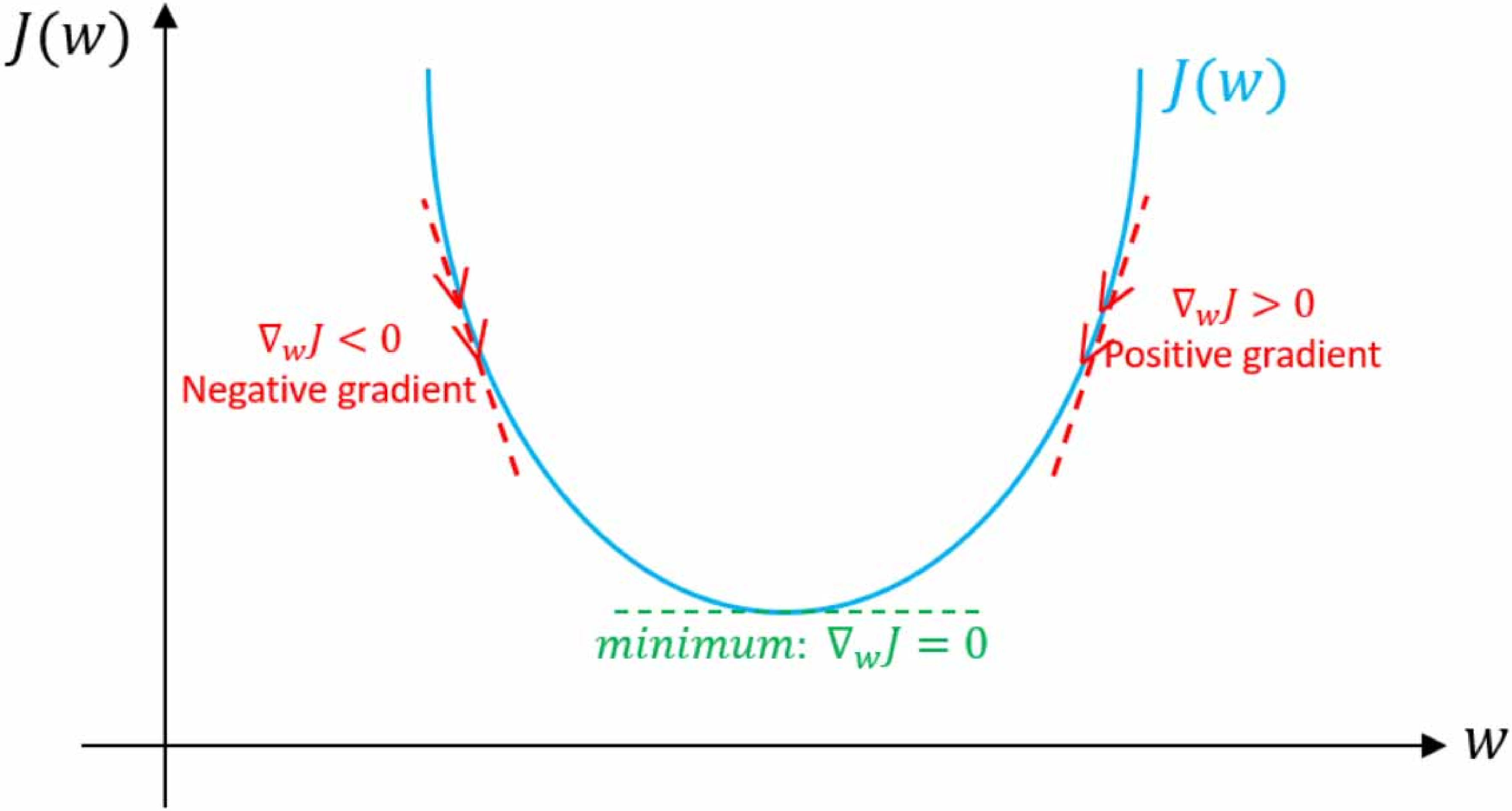
8 % dog

loss function

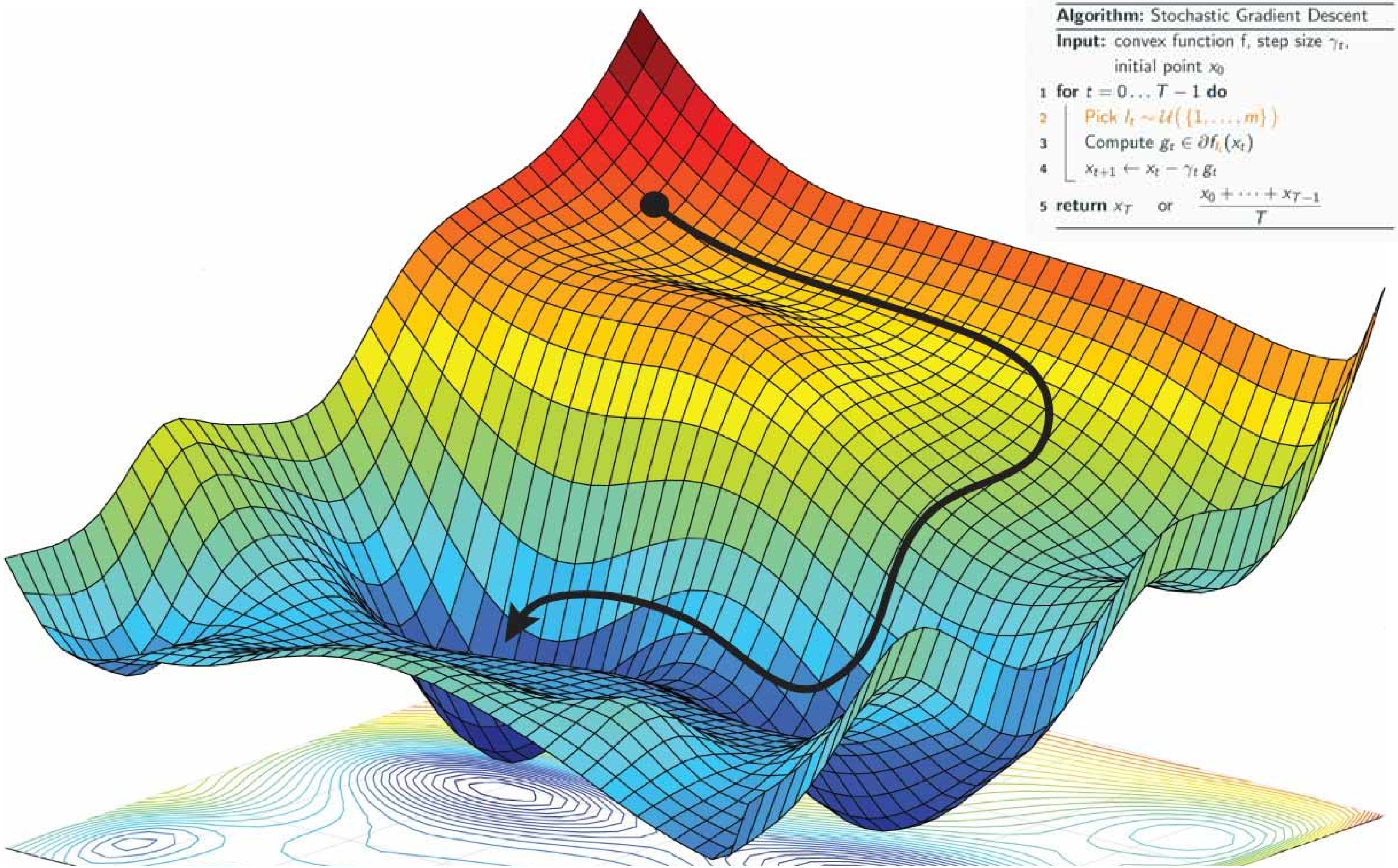
model

$$\min_{x} \sum_{i=1}^{N} f(x; y_i)$$

training examples


$J(w)$
$J(w)$
$\nabla_w J < 0$
Negative gradient
$\nabla_w J > 0$
Positive gradient
$minimum\colon\ \nabla_w J = 0$
w

What about non-convex and non-linear problems?



Algorithm: Stochastic Gradient Descent

Input: convex function f, step size γ_t, initial point x_0

for $t = 0 \ldots T-1$ **do**

  Pick $I_t \sim \mathcal{U}(\{1, \ldots, m\})$

  Compute $g_t \in \partial f_{I_t}(x_t)$

  $x_{t+1} \leftarrow x_t - \gamma_t g_t$

return x_T or $\frac{x_0 + \cdots + x_{T-1}}{T}$

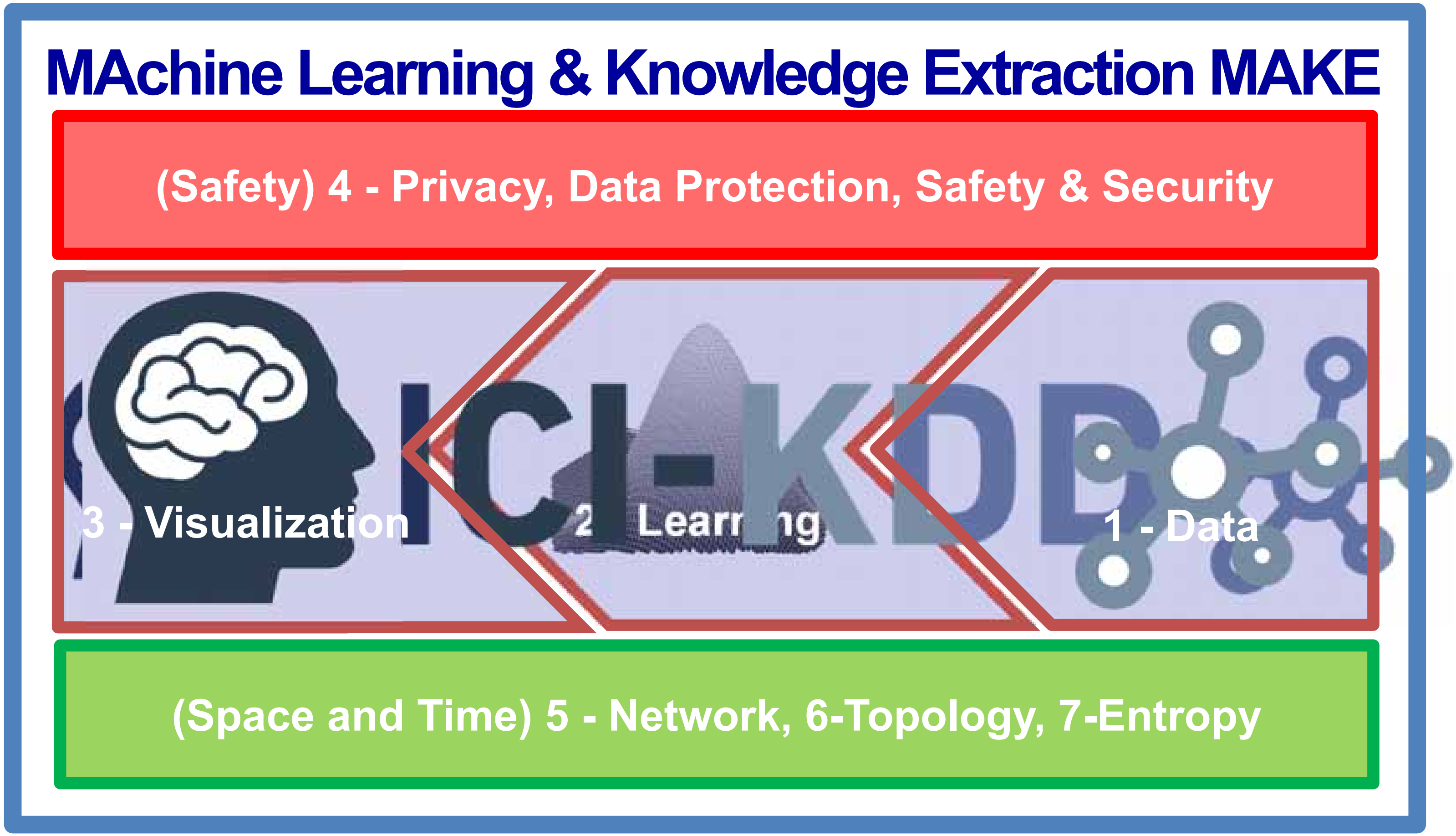
$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \mathbb{E}_i[f_i(x)] = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

For a tutorial on Machine Learning with Python please look up:
https://graz.pure.elsevier.com/de/publications/a-tutorial-on-machine-learning-and-data-science-tools-with-python

Ok, but now let's start …

- 01 Integrative ML: Human-Centered AI
- 02 Application Area Health
- 03 Probabilistic Learning
- 04 Automatic Machine Learning (aML)
- 05 Interactive Machine Learning (iML)
- 06 “Explainable AI”
- Conclusion and future outlook

01 What is the HCAI approach?



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

https://human-centered.ai/explainable-ai-2020

"Solve intelligence – then solve everything else"

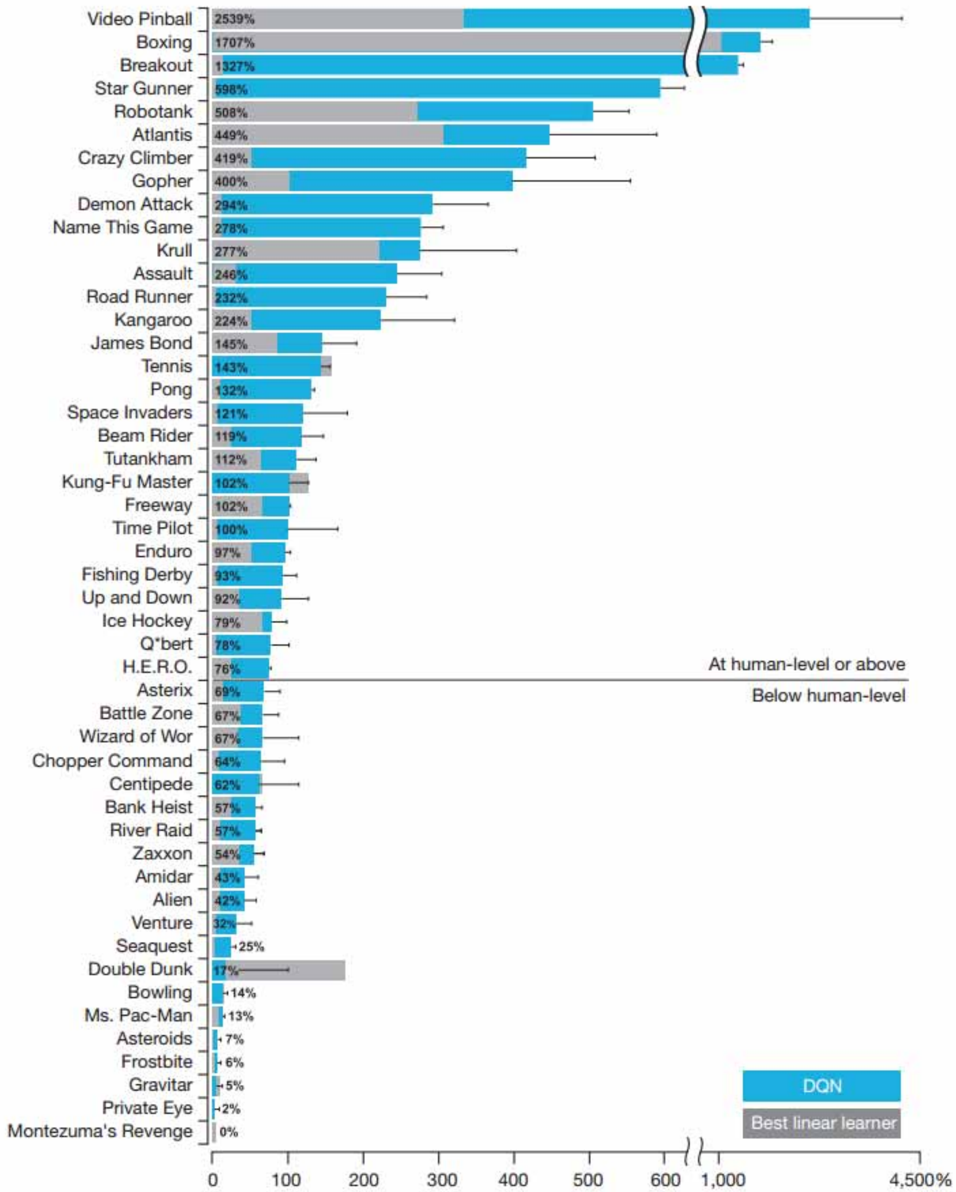
https://youtu.be/XAbLn66iHcQ?t=1h28m54s

Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

Compare your best ML algorithm with a seven-year-old child …

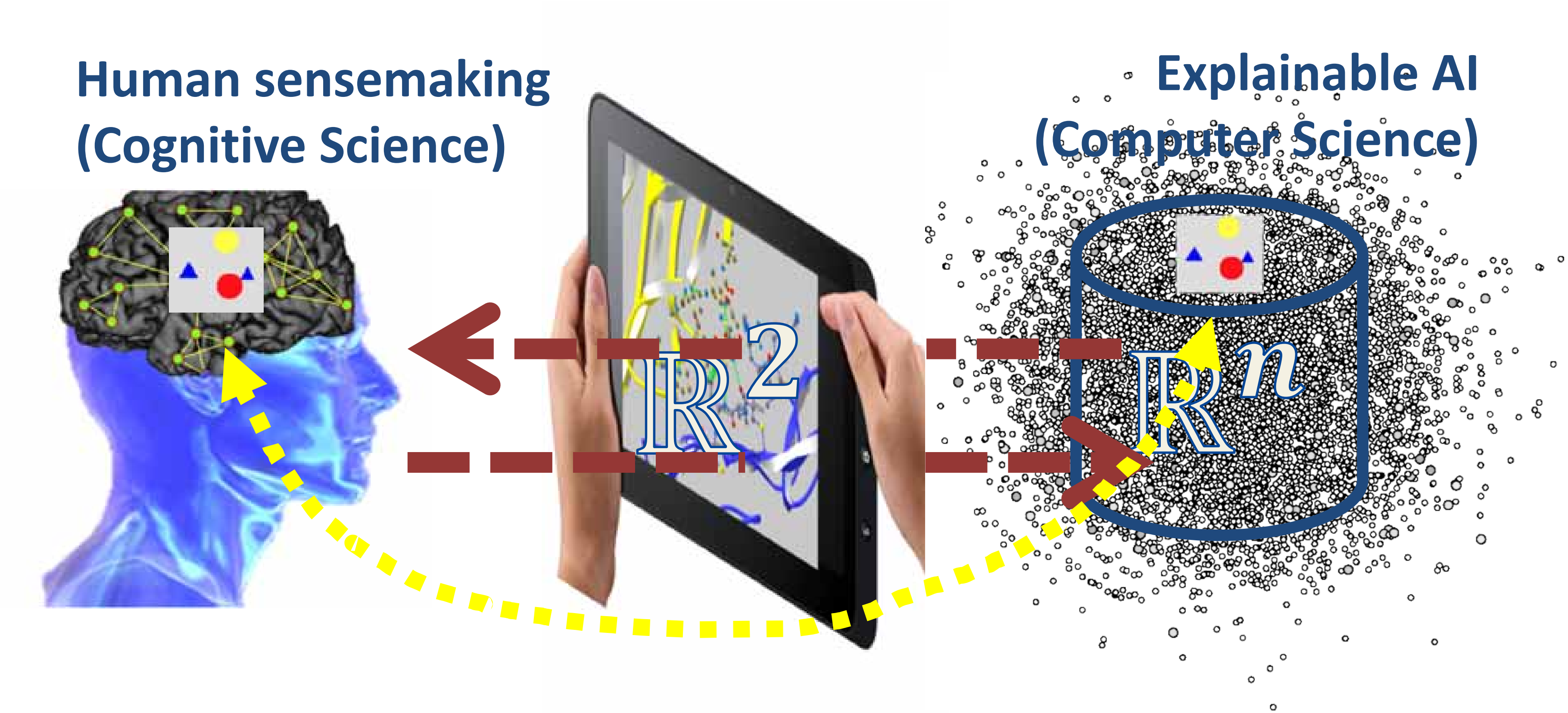


Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement learning. Nature, 518, (7540), 529-533, doi:10.1038/nature14236

- 1) **learn** from prior data
- 2) **extract** knowledge
- 2) **generalize,** (e.g. guessing where a probability mass function concentrates)
- 4) fight the curse of **dimensionality**
- 5) **disentangle** underlying explanatory factors of data, i.e.
- 6) **understand** the results in the **context** of an application domain (sensemaking)

Our goal: Understanding Context !

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

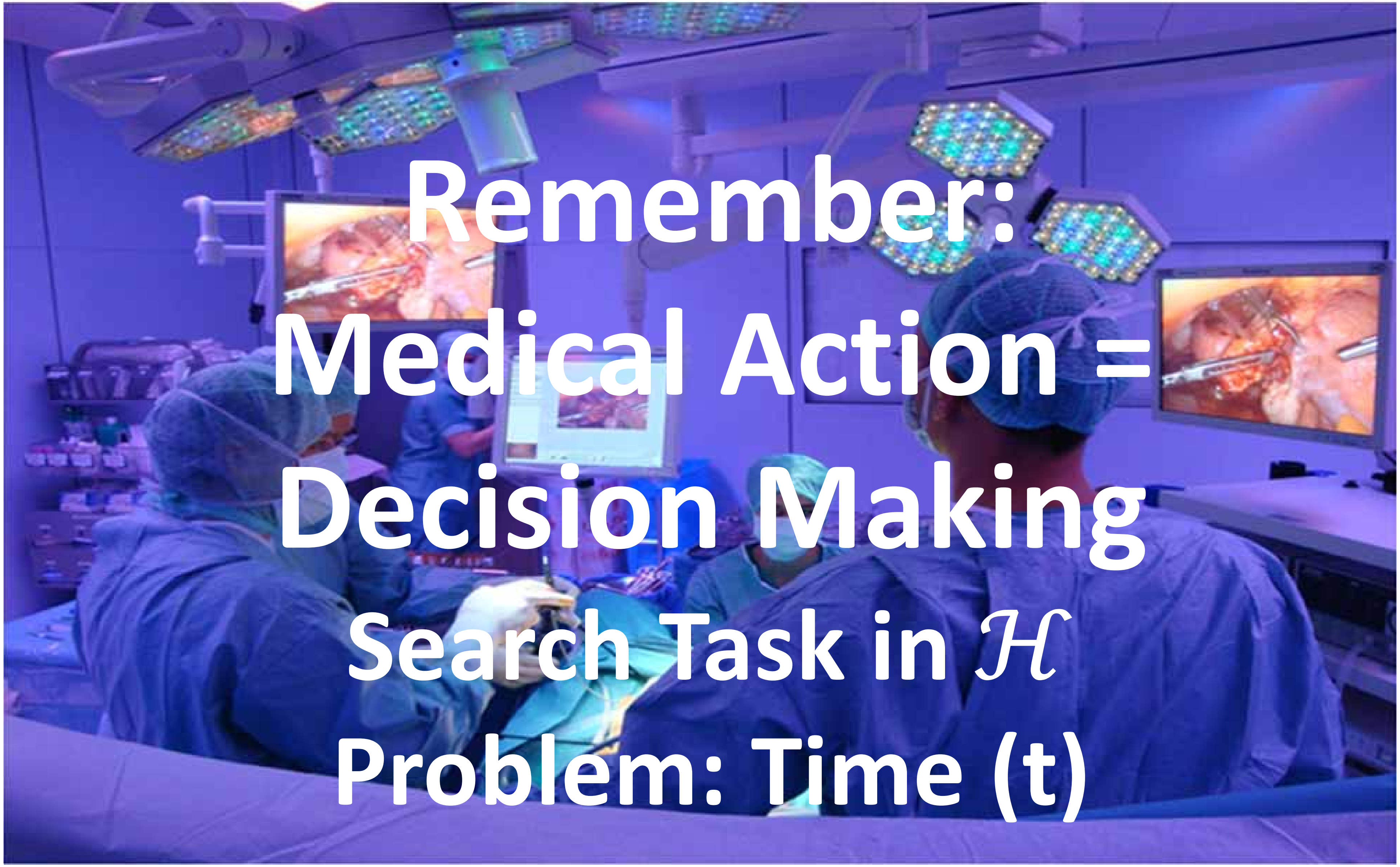


Andreas Holzinger et al. 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

Source: Image is in the public domain and is used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students

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.

02 Application Area Health Informatics



Remember:

Medical Action =

Decision Making

Search Task in $\mathcal{H}$

Problem: Time (t)

Why is time so important and in machine learning often ignored ?

What does it mean to make a “de-cision” ?

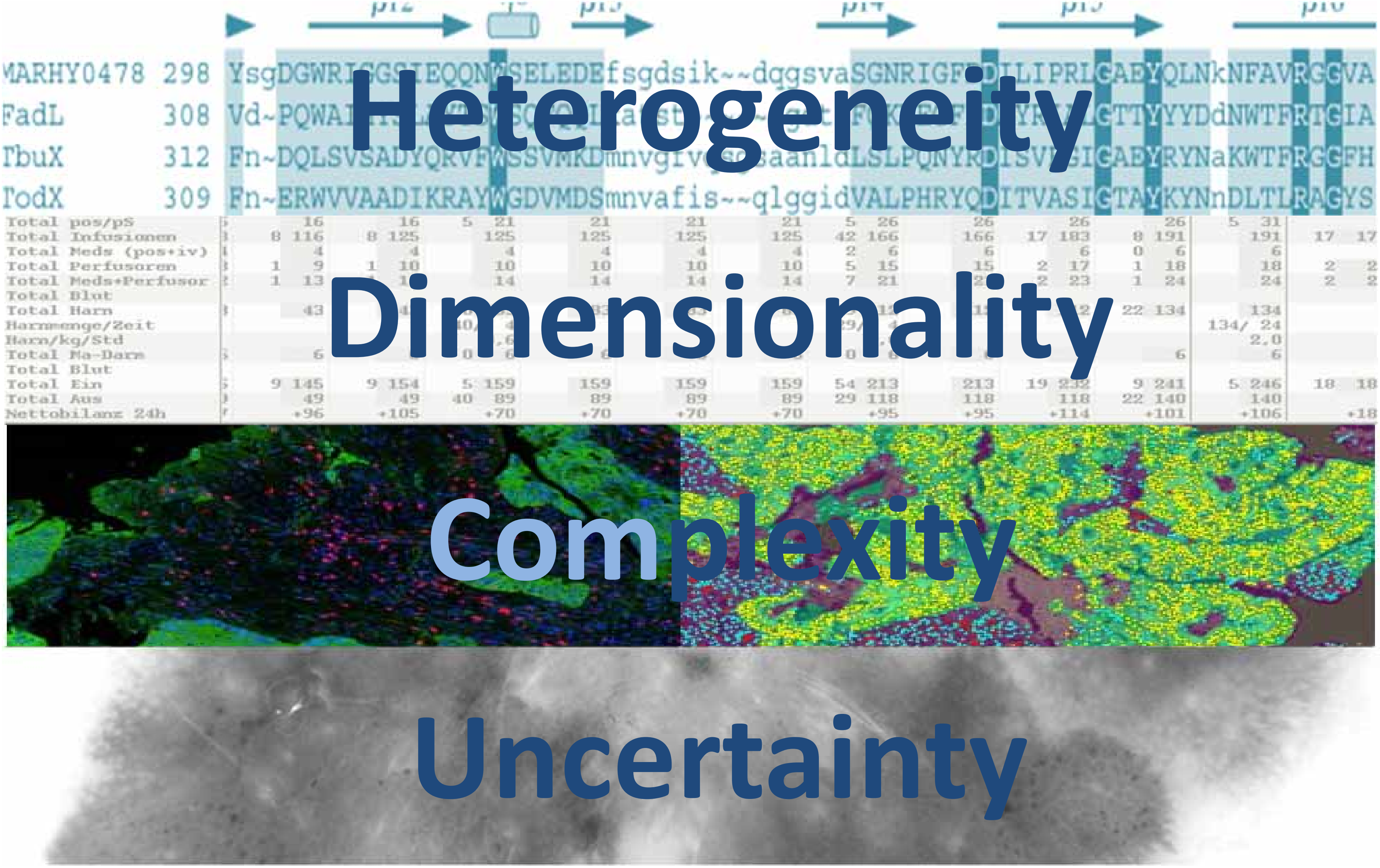
Source: Cisco (2008). Cisco Health Presence Trial at Aberdeen Royal Infirmary in Scotland

- 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 …

The images on this slide are used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students

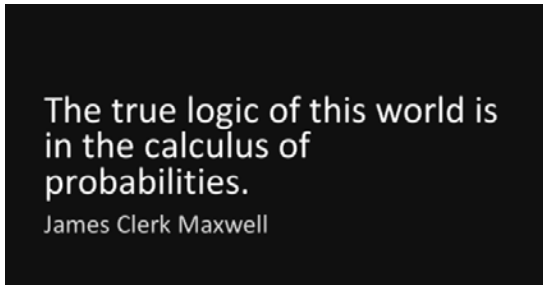
Our central hypothesis:
Information may bridge this gap

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.



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.

03 Probabilistic Learning



Maxwell, J. C. (1850). Letter to Lewis Campbell; reproduced in L. Campbell and W. Garrett, The Life of James Clerk Maxwell, Macmillan, 1881.

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

$$\hat{y} = \hat{f}(\mathbf{x}) = \operatorname*{argmax}_{c=1}^{C} p(y = c|\mathbf{x}, \mathcal{D})$$

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.

Pierre Simon de Laplace (1749-1827)

- **1763: Richard Price publishes post hum the work of Thomas Bayes (see next slide)**
- **1781: Pierre-Simon Laplace: Probability theory is nothing, but common sense reduced to calculation ...**
- **1812: Théorie Analytique des Probabilités, now known as Bayes' Theorem**

- **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

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

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

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:

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

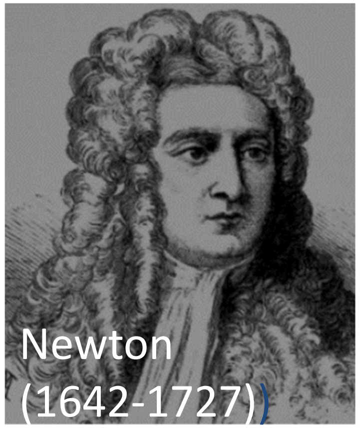
Now we introduce a third, more complicated operation:

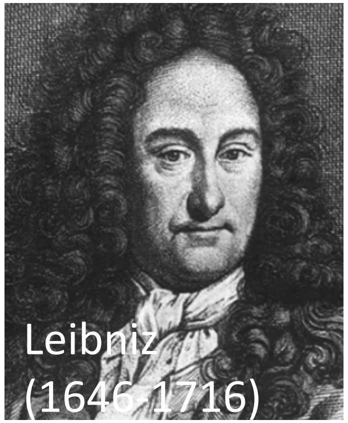
$$\frac{p(x, y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \tag{4}$$

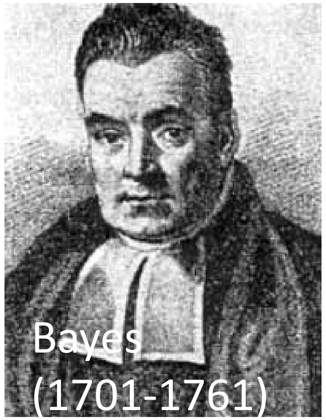
We can reduce this fraction by $p(y)$ and we receive what is called Bayes rule:

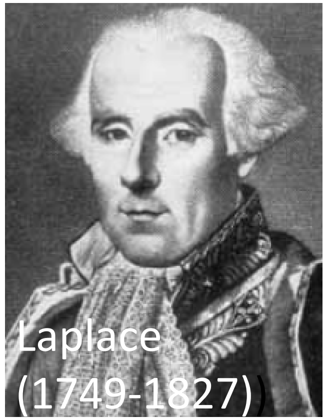
$$p(x, y) = \frac{p(y|x) * p(x)}{p(y)} \qquad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \tag{5}$$

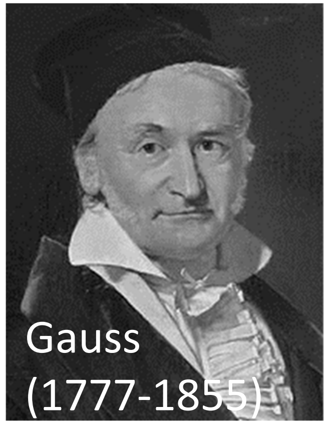
Who are our giants for machine learning?











- **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**

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

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

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:

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

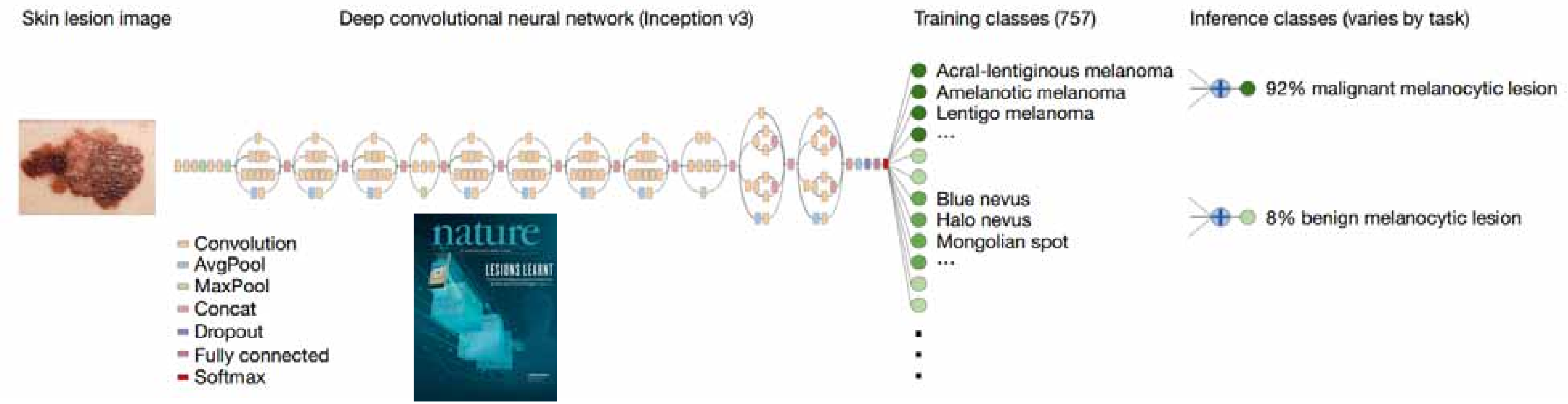
$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis})P(\text{data}|\text{hypothesis})}{\sum_{\text{h}} P(\text{h})P(\text{data}|\text{h})} \qquad P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

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

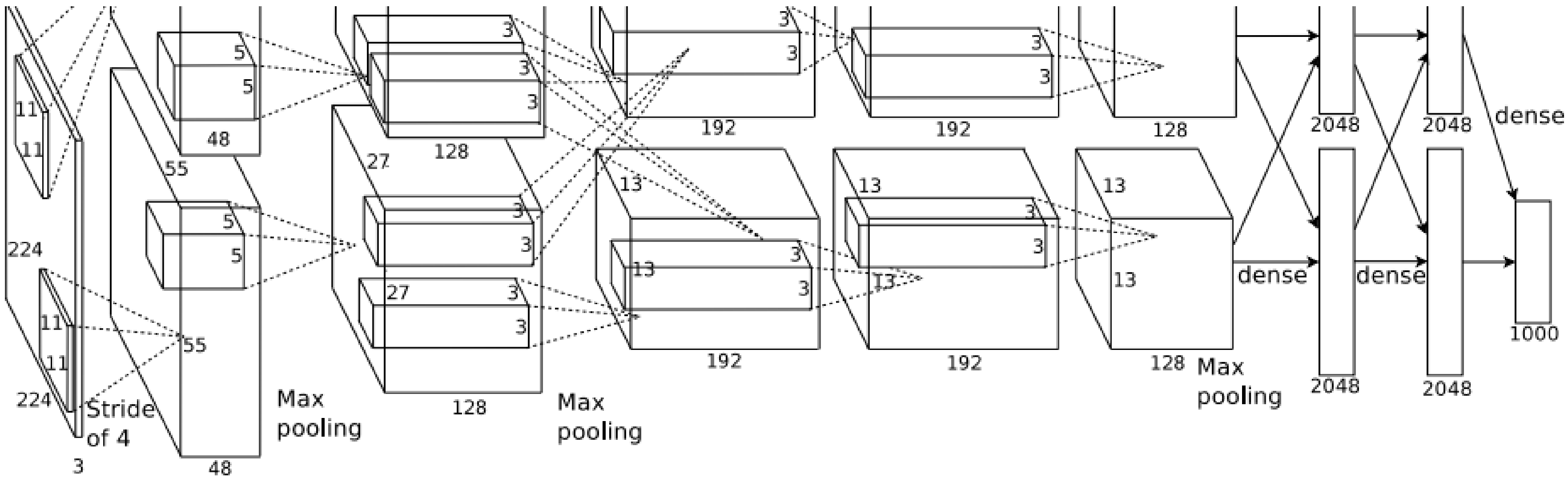
$P(\theta|m)$ prior probability of θ

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

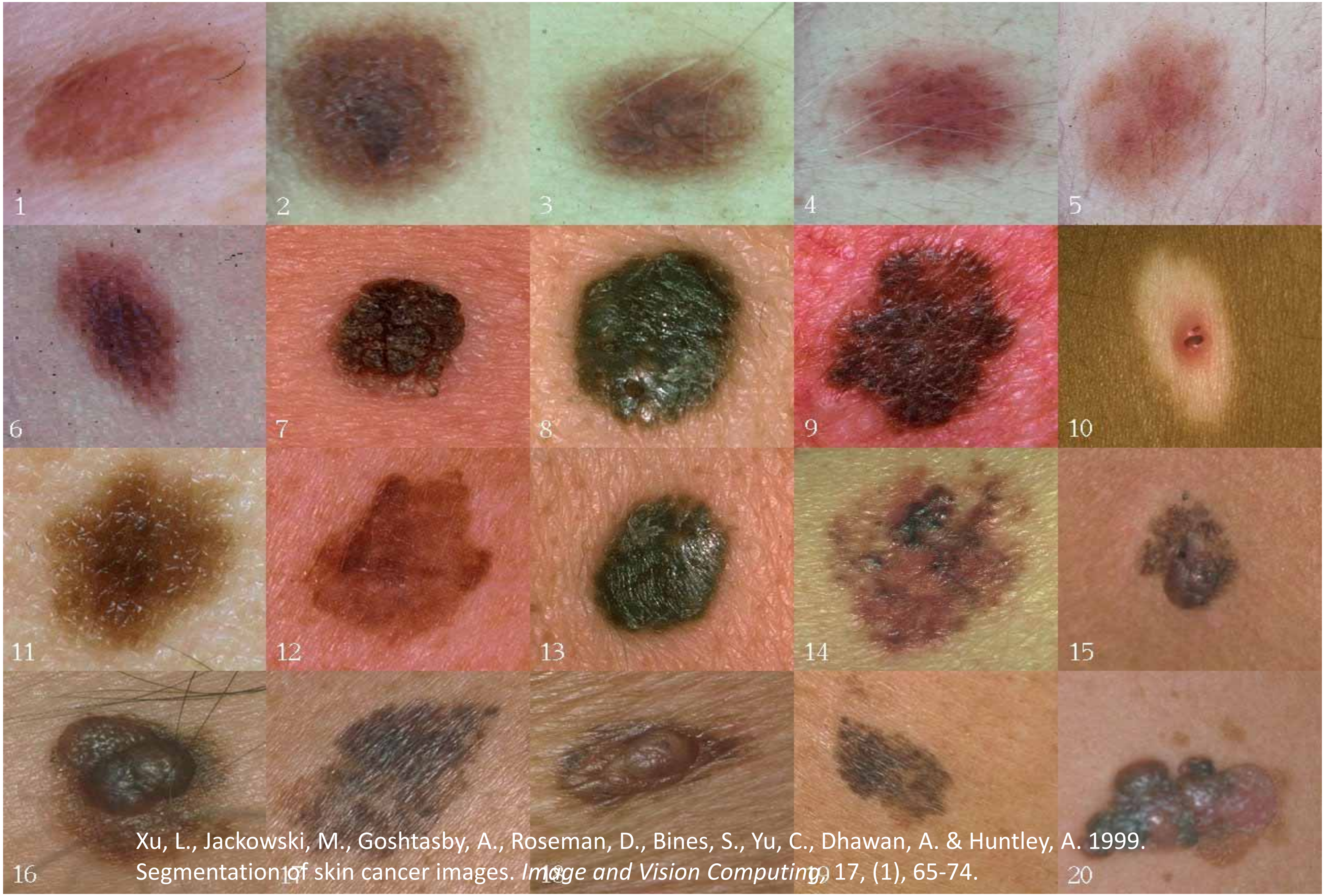
Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. Biometrika, 45(3/4), 293-315.



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.

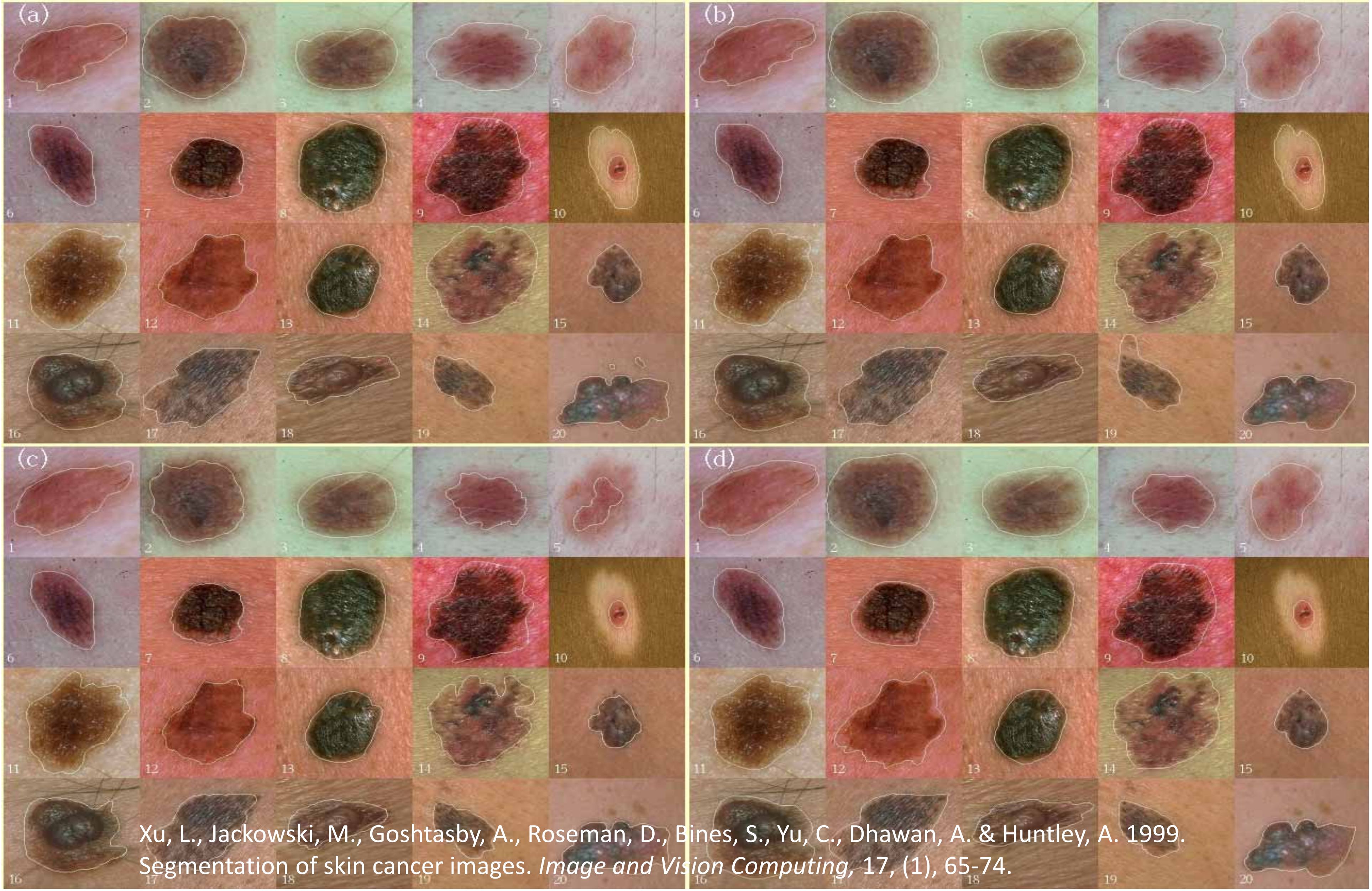


What is needed for successful machine learning?



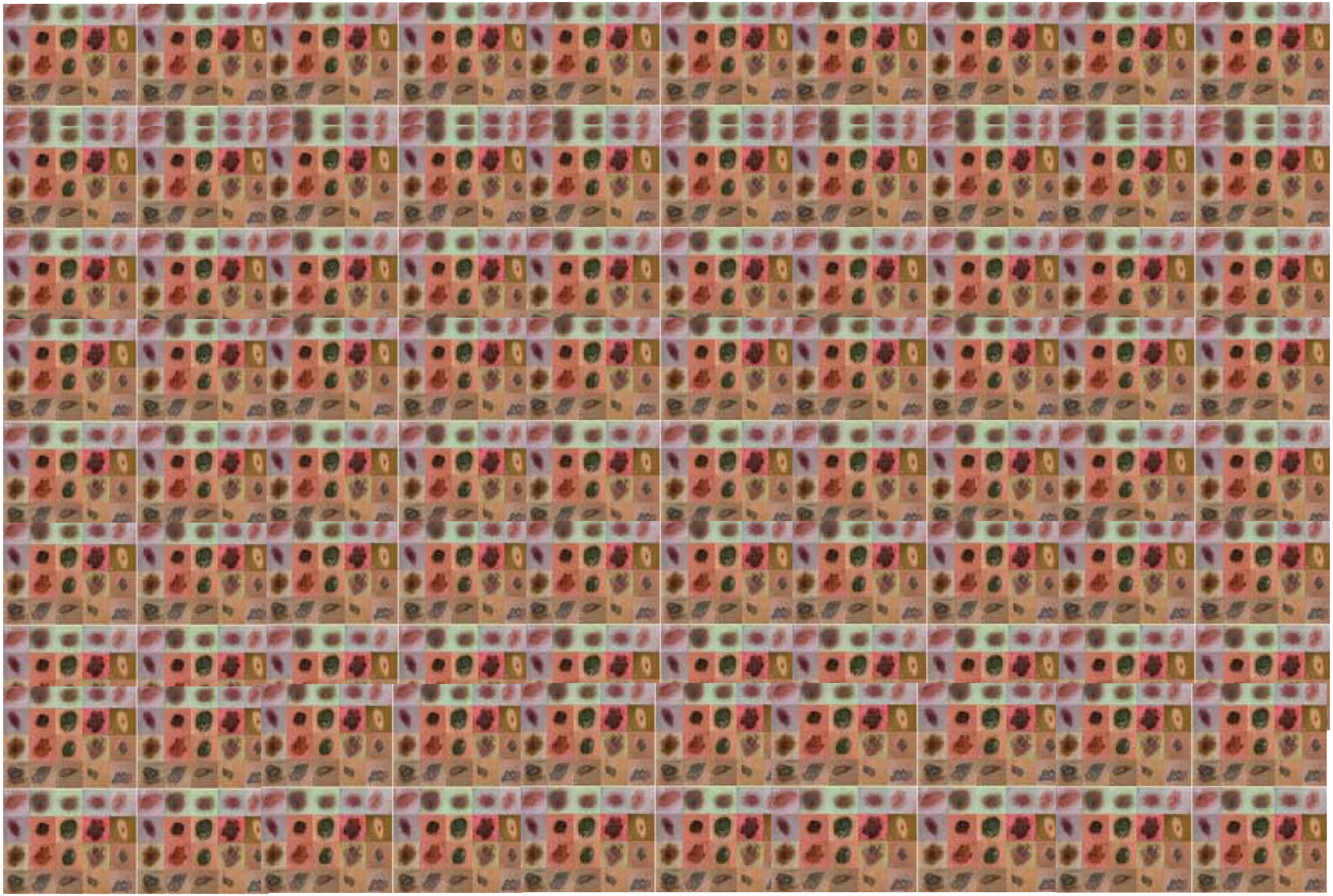
Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A. & Huntley, A. 1999. Segmentation of skin cancer images. *Image and Vision Computing*, 17, (1), 65-74.

Why do we need top quality data for machine learning?

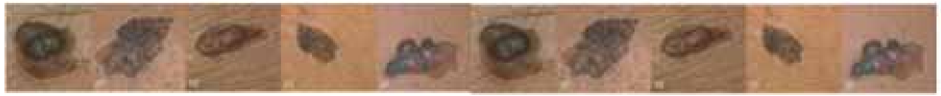


Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A. & Huntley, A. 1999. Segmentation of skin cancer images. *Image and Vision Computing*, 17, (1), 65-74.

Why do wee need so much training data for machine learning?



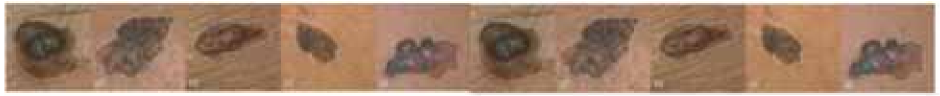
$$\mathcal{D} = x_{1:n} = \{x_1, x_2, \ldots, x_n\}$$



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

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

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

Observed data: 

≈ Training data: $\mathcal{D} = x_{1:n} = \{x_1, x_2, \ldots, x_n\}$

Feature Parameter: θ **or hypothesis** ***h*** $h \in \mathcal{H}$

Prior belief ≈ prior probability of hypothesis ***h*****:** $p(\theta)$ $p(h)$

Likelihood ≈ ***p(x)*** **of the data that** ***h*** **is true** $p(\mathcal{D}|\theta)$ $p(d|h)$

Data evidence ≈ marginal ***p(x)*** **that** ***h*** **= true** $p(\mathcal{D})$ $\sum_{h \in \mathcal{H}} p(d|h) * p(h)$

Posterior ≈ ***p(x)*** **of** ***h*** **after seen ("learn") data** ***d*** $p(\theta|\mathcal{D})$ $p(h|d)$

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

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

Why is this relevant for medicine?

- 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

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

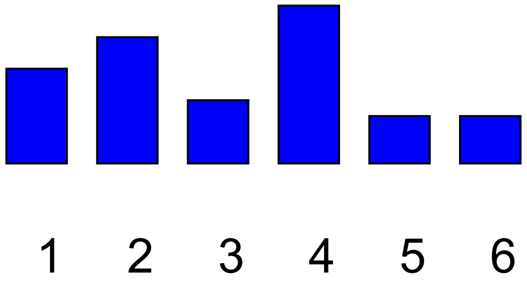
Probabilistic Decision Making

"It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge"
Pierre Simon de Laplace, 1812

- $Probability\ p(x)$ 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

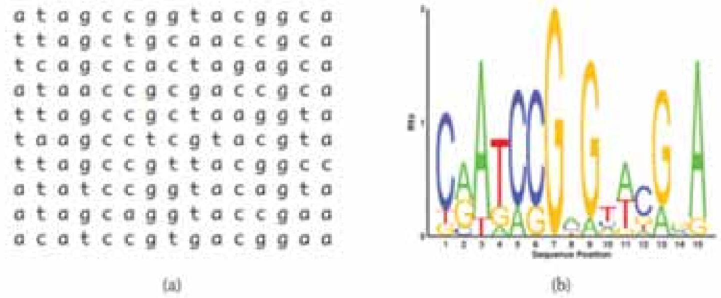
- Discrete distributions:



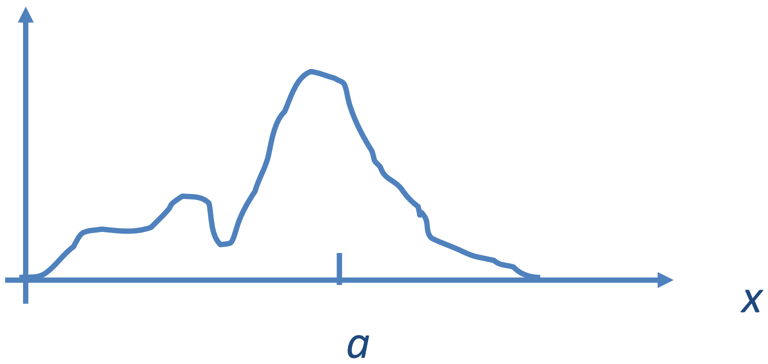
$$\sum_i P(X = x_i) = 1$$

Name	n	K	x
Multinomial	-	-	$\mathbf{x} \in \{0, 1, \ldots, n\}^K, \sum_{k=1}^K x_k = n$
Multinoulli	1	-	$\mathbf{x} \in \{0, 1\}^K, \sum_{k=1}^K x_k = 1$ (1-of-K encoding)
Binomial	-	1	$x \in \{0, 1, \ldots, n\}$
Bernoulli	1	1	$x \in \{0, 1\}$

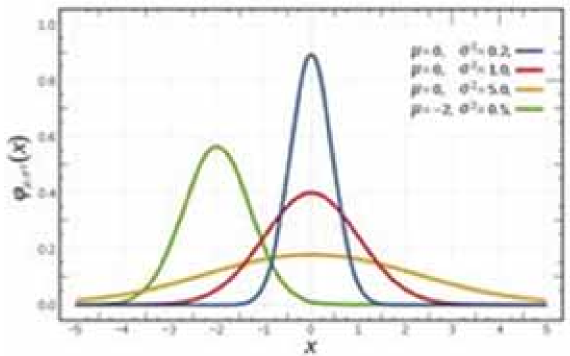
Table 2.1 Summary of the multinomial and related distributions.



- Continuous: Probability density function (PDF) vs Cumulative Density Function (CDF):

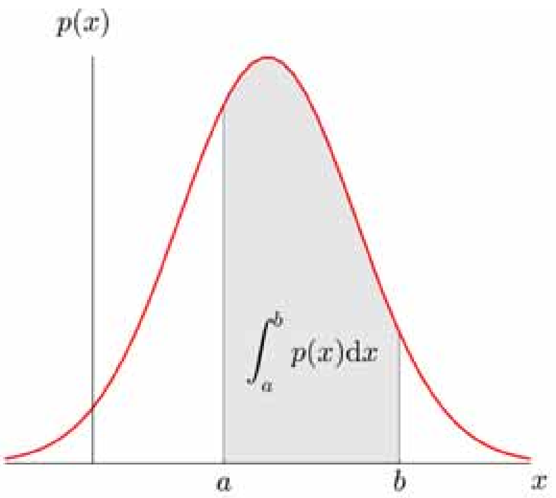


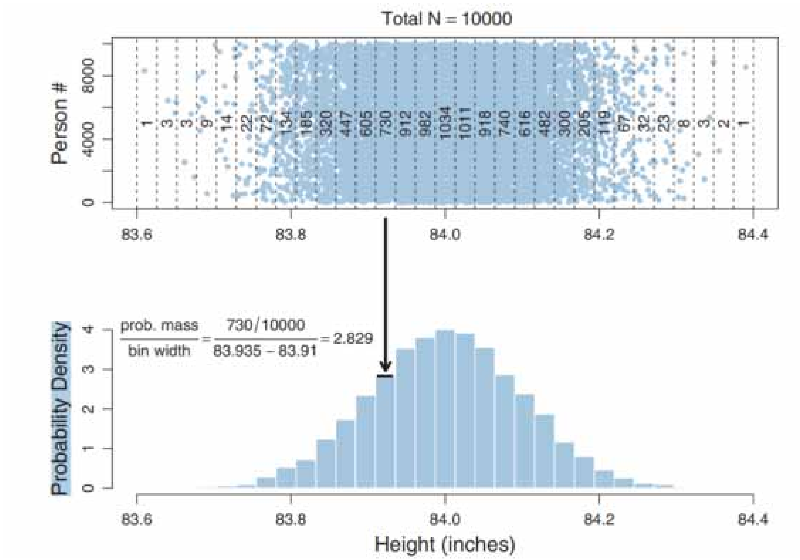
$$P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau$$



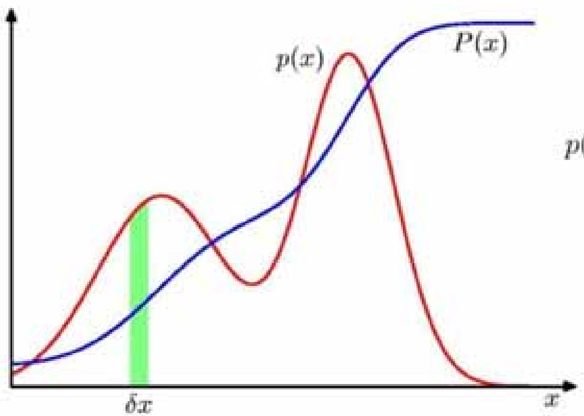
$$\mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Probability Density Function and Probability Distribution





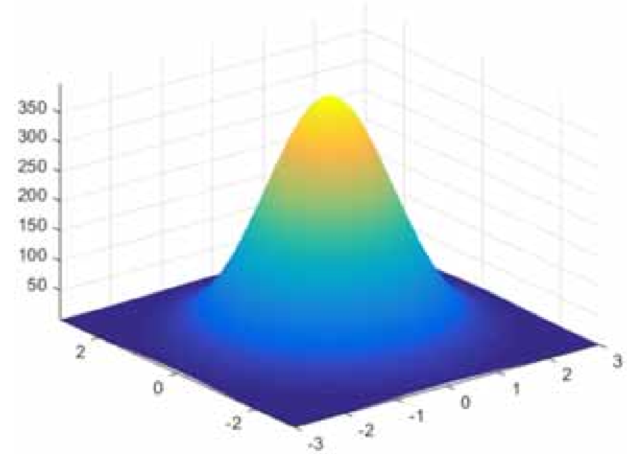
John Kruschke 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, Amsterdam et al., Academic Press.



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

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

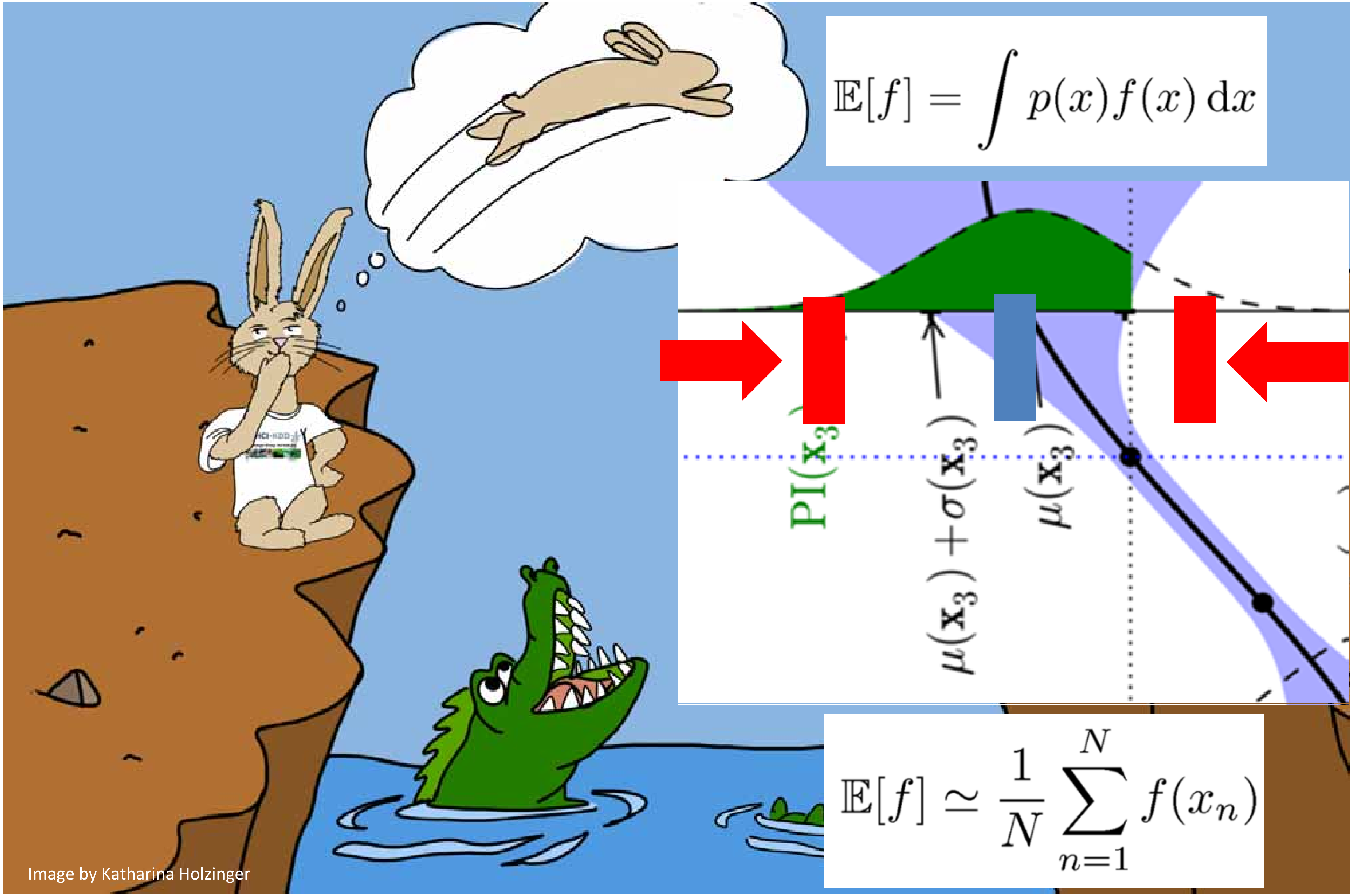
$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$



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

Expectation and Expected Utility Theory

Why does uncertainty matter ?



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

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

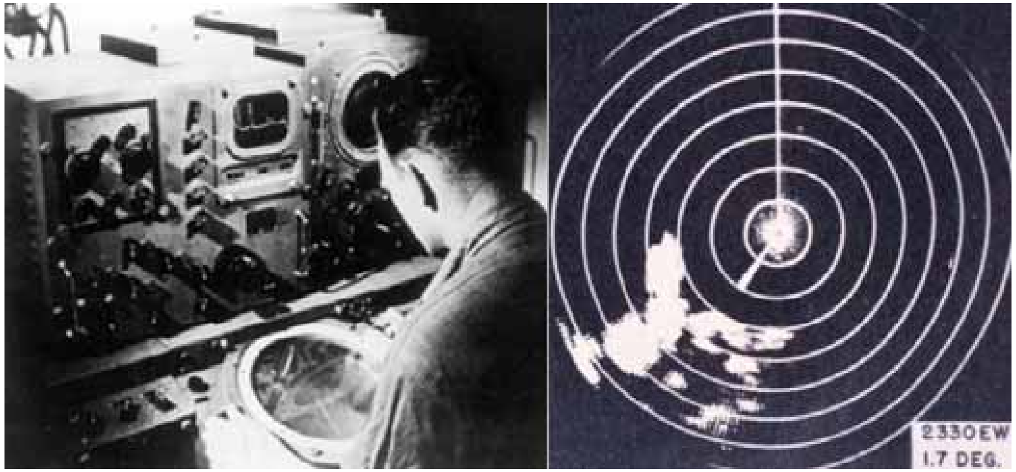
An optimal single decision is the decision $D = dmax$ whose expected utility is maximal:

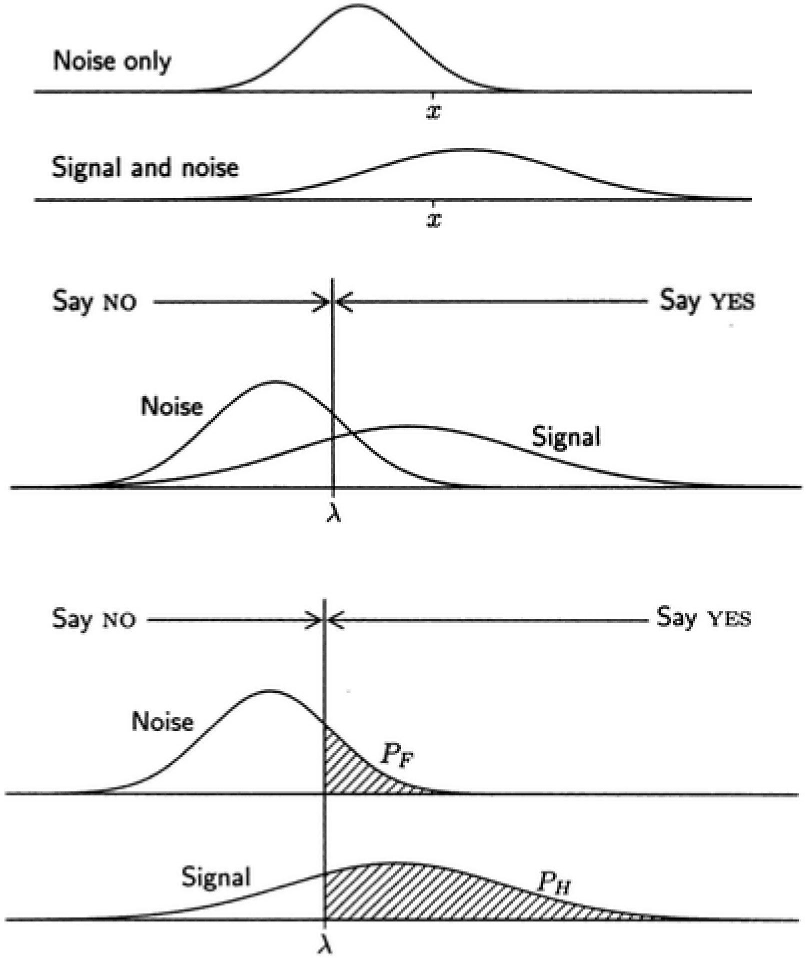
$$d_{\max} = \arg \max_{d \in \mathrm{dom}(D)} E(U \mid d)$$

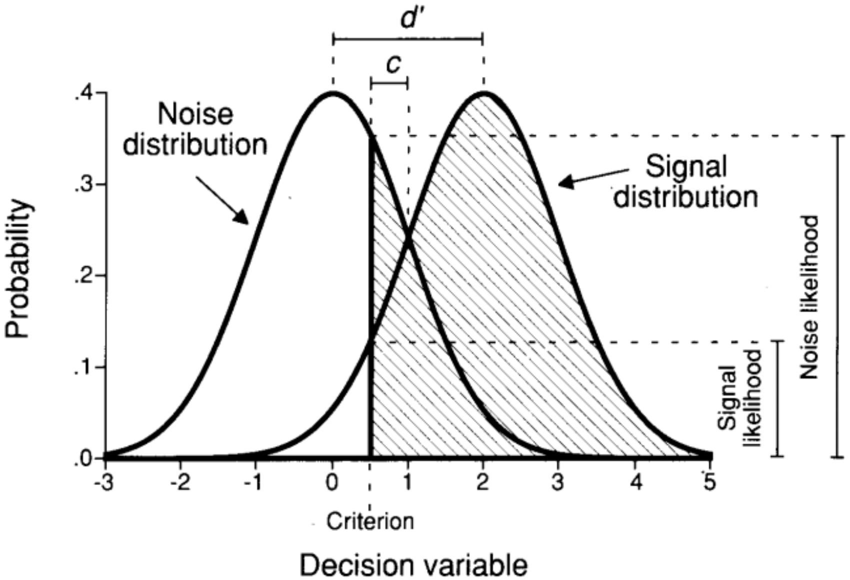
John Von Neumann & Oskar Morgenstern 1944. Theory of games and economic behavior, Princeton university press.

What was the origin of probabilistic decision making ?

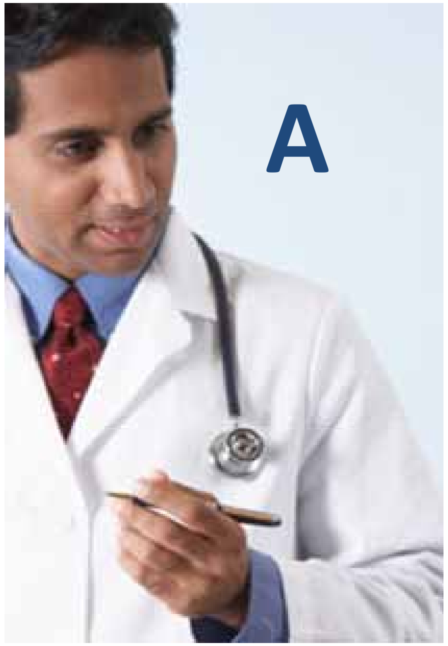
Image source: Staffordshire University Computing Futures Museum http://www.fcet.staffs.ac.uk/jdw1/sucfm/malvern.htm

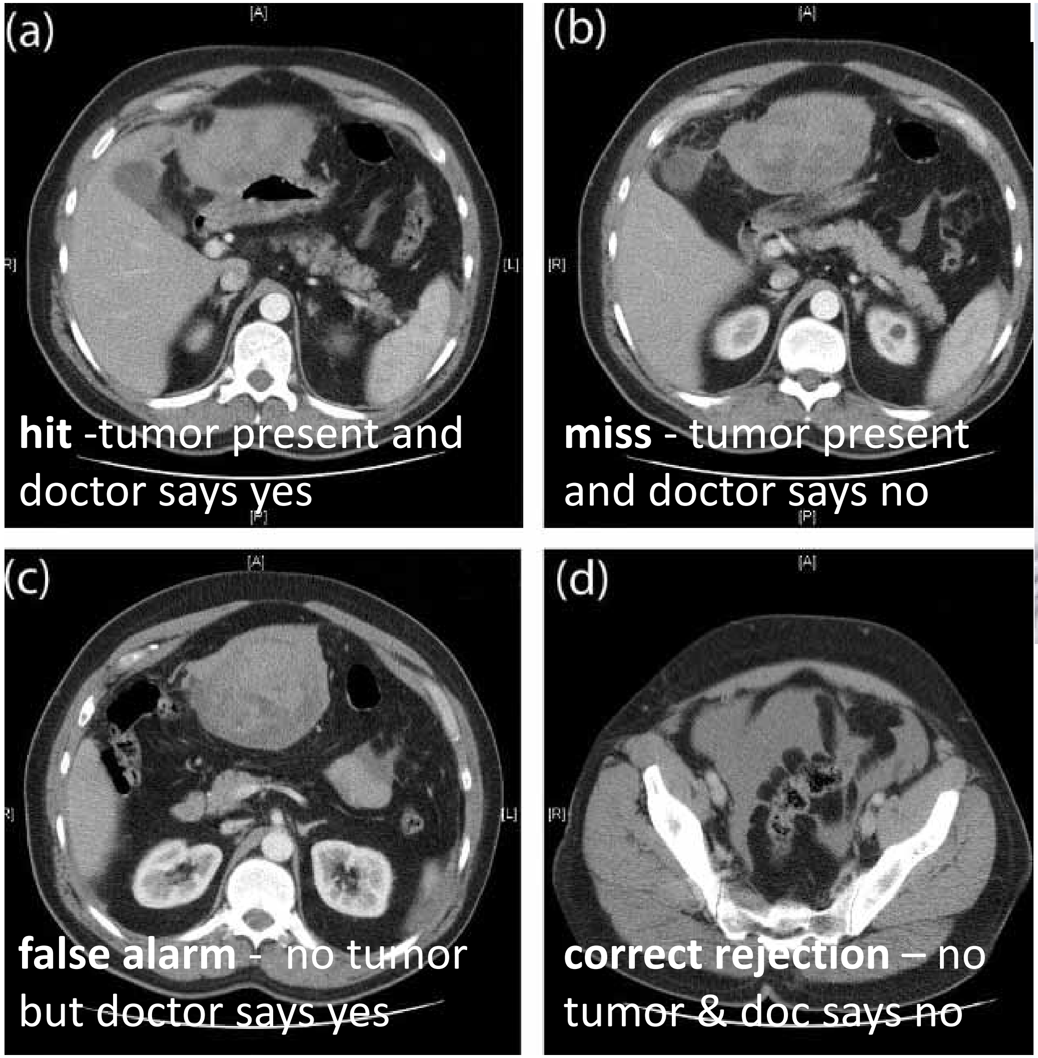






Stanislaw, H. & Todorov, N. 1999. Calculation of signal detection theory measures. Behavior research methods, instruments, & computers, 31, (1), 137-149.

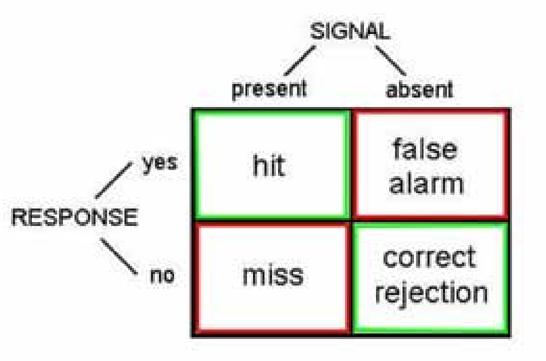




Two doctors, with equally good training, looking at the same CT scan, will have the same information … but they may have a **different bias/criteria!**

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

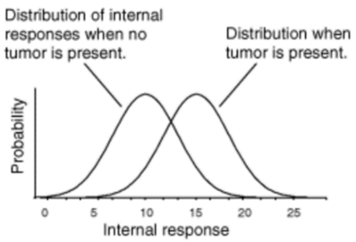
sensitivity, recall, hit rate, or true positive rate (TPR)

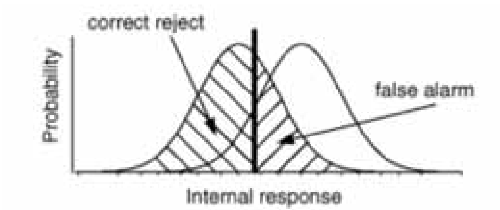
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

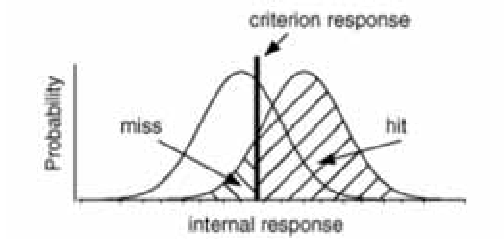
specificity, selectivity or true negative rate (TNR)

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

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

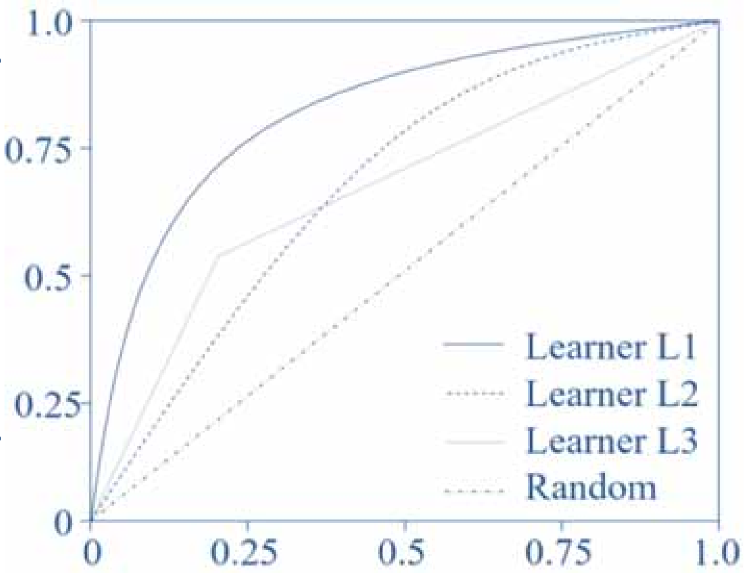






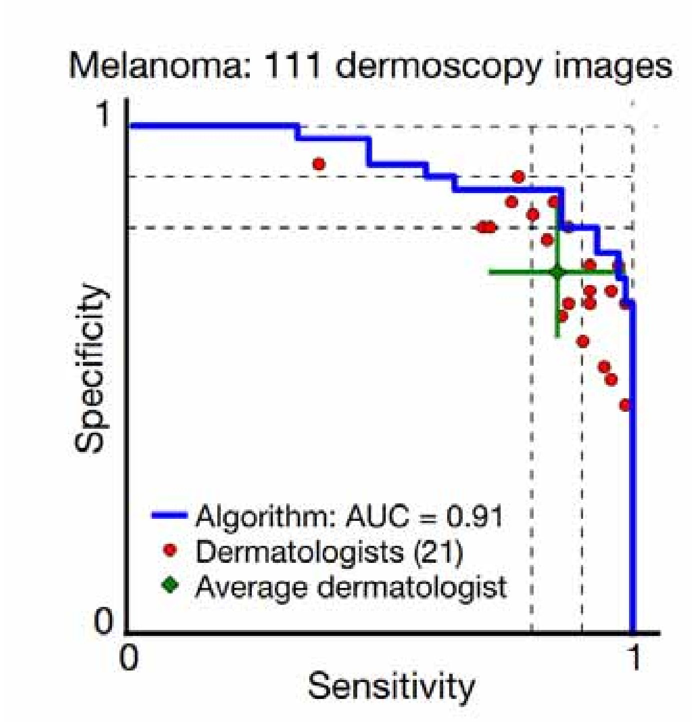
For an example see: Braga & Oliveira (2003) Diagnostic analysis based on ROC curves: theory and applications in medicine. *Int. Journal of Health Care Quality Assurance, 16, 4, 191-198.*

And please look up the Wikipedia page:



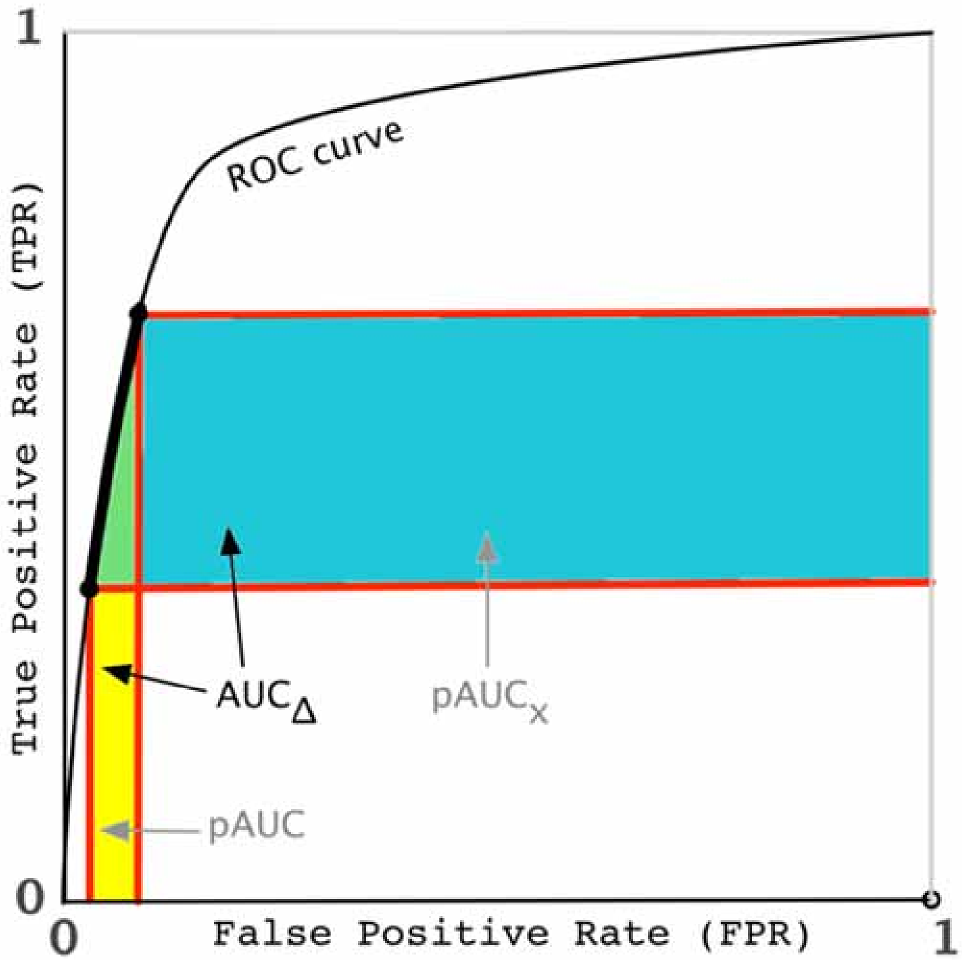
True positive = Sensitivity

False positive = Specificity

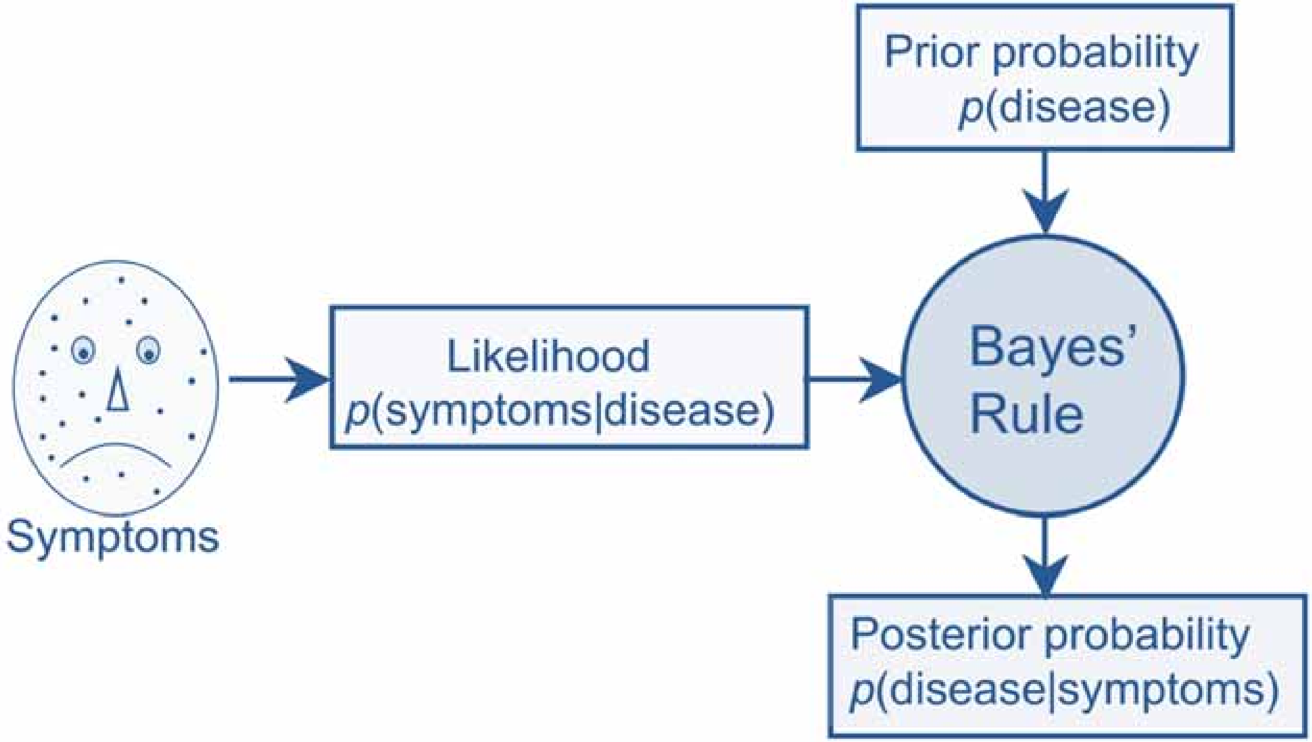


Andrew P. Bradley 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, (7), 1145-1159, doi:http://dx.doi.org/10.1016/S0031-3203(96)00142-2.

What is the advantage of the concordant pAUC ?



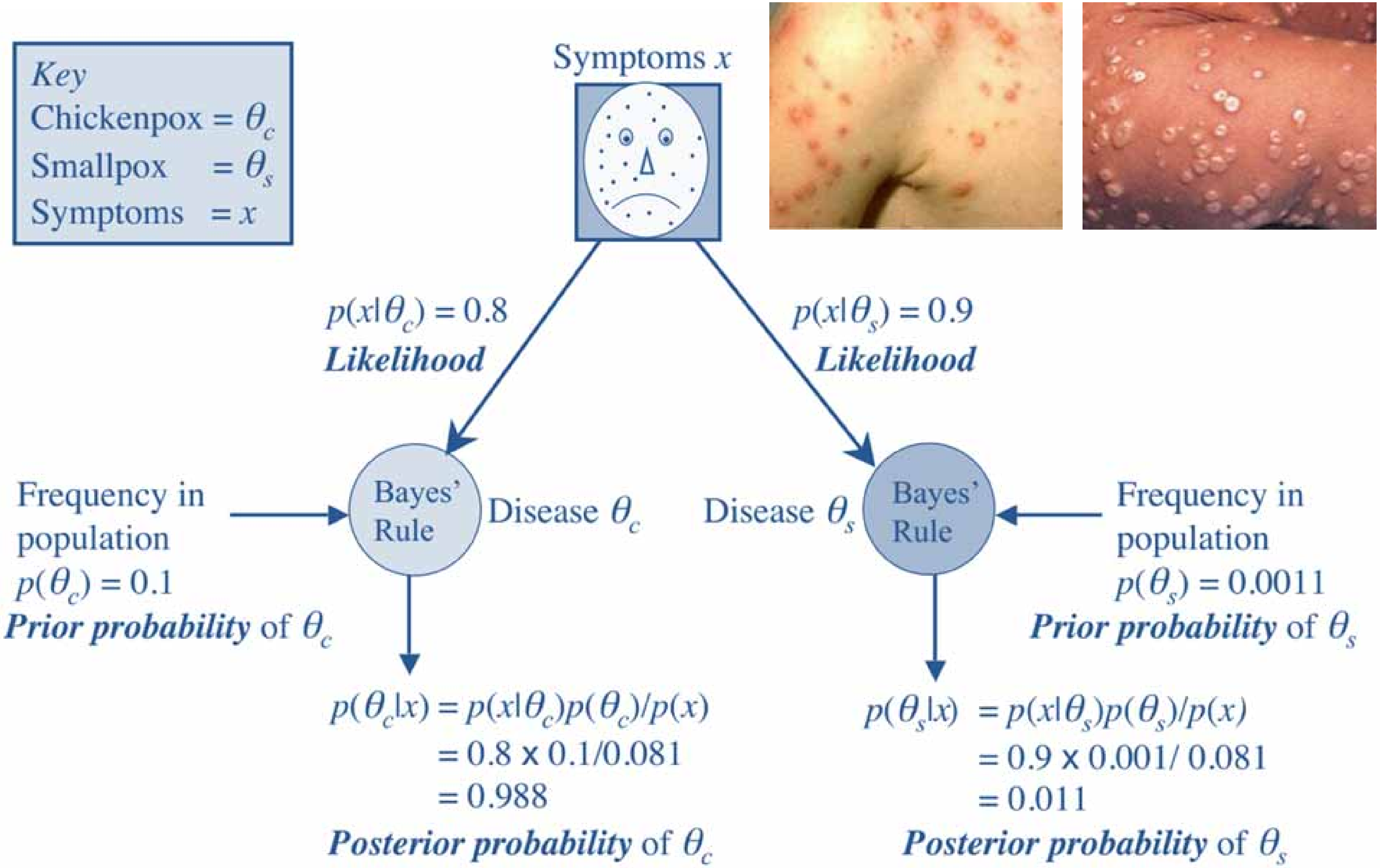
André M. Carrington, Paul W. Fieguth, Hammad Qazi, Andreas Holzinger, Helen H. Chen, Franz Mayr & Douglas G. Manuel 2020. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. Springer/Nature BMC Medical Informatics and Decision Making,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



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

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?**

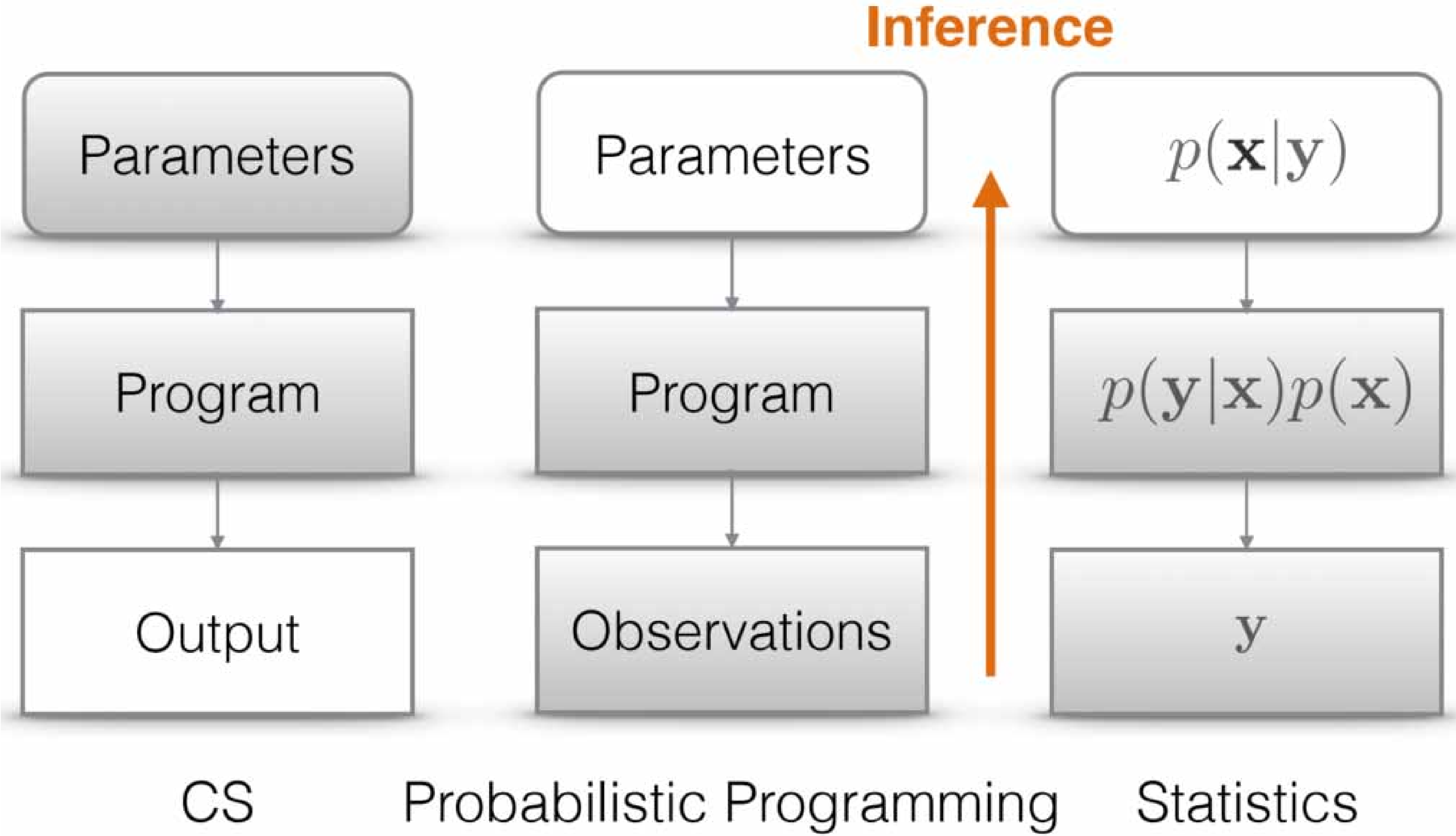
$$posterior\ p(x) = \frac{likelihood * prior\ p(x)}{evidence}$$

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

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

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

$$\mathbf{= 0{,}0098}$$



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

How can we calculate the posterior probability?

d … $data$

h … $hypotheses$

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

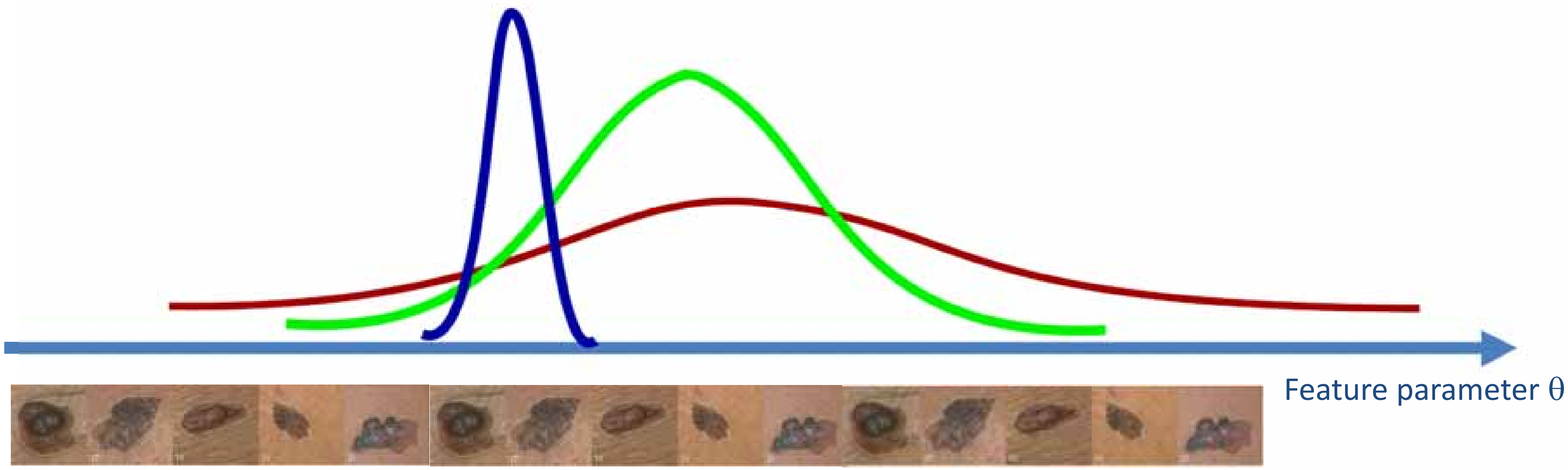
Likelihood

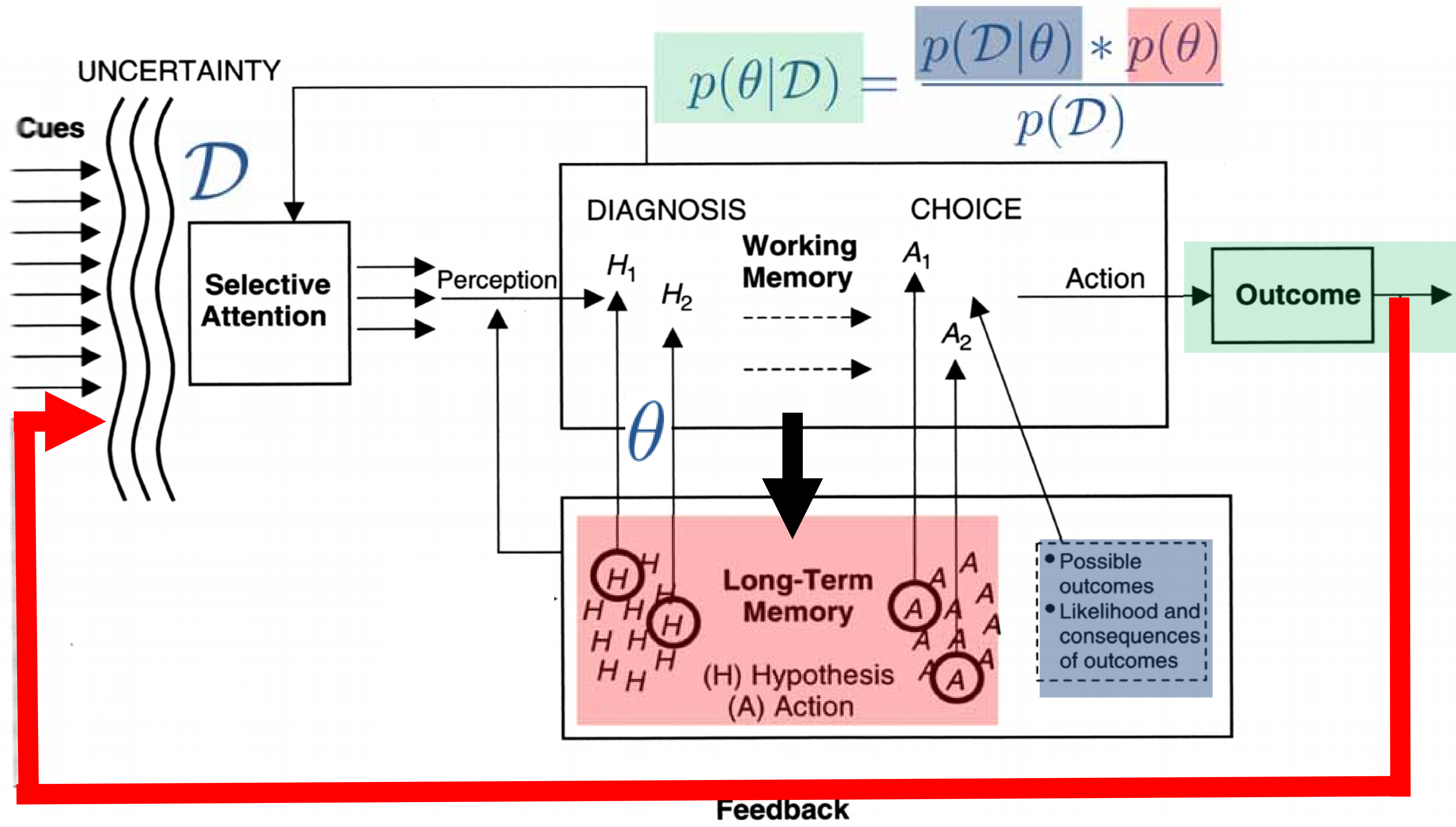
Prior Probability

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

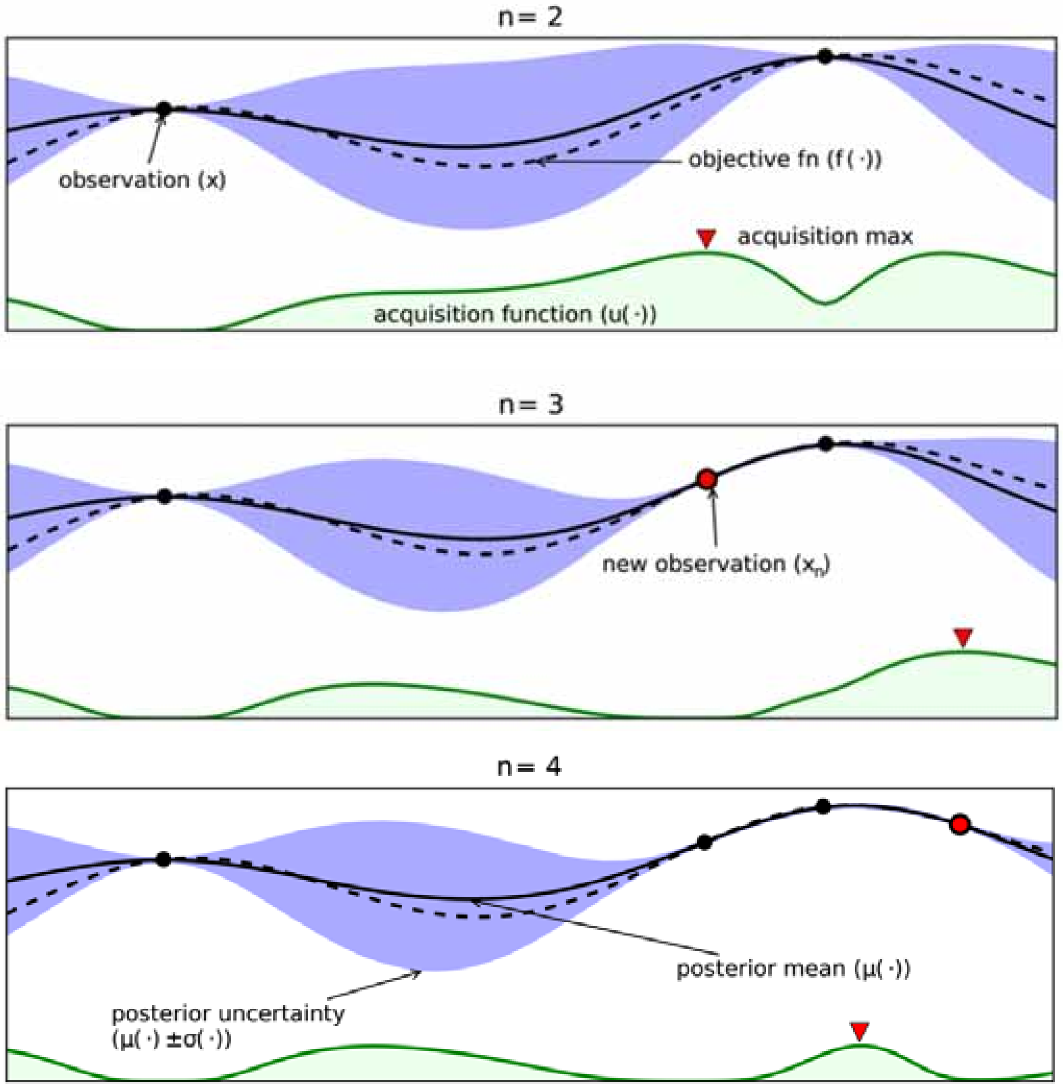
Posterior Probability

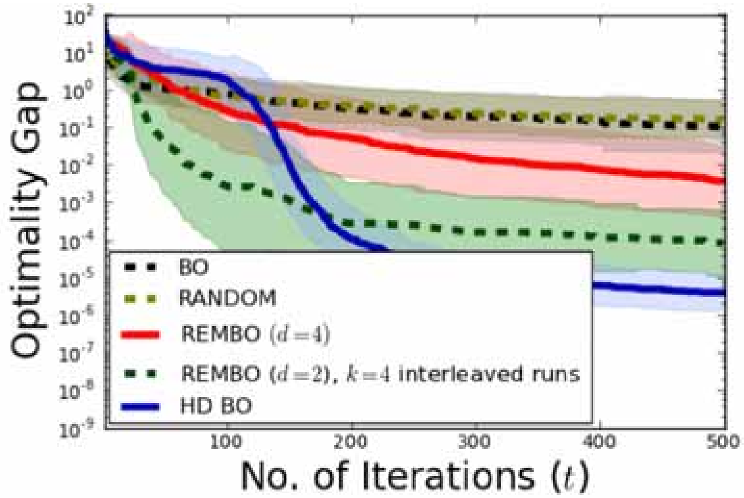
Problem in $\mathbb{R}^n \rightarrow$ complex

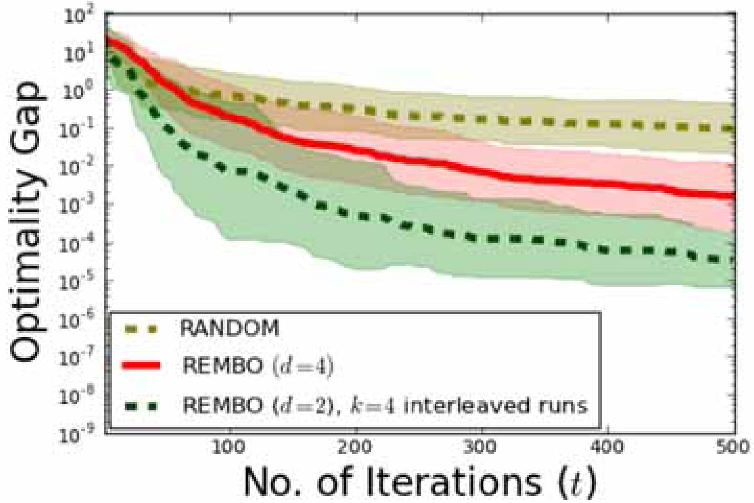




Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill, modified by Holzinger, A.*



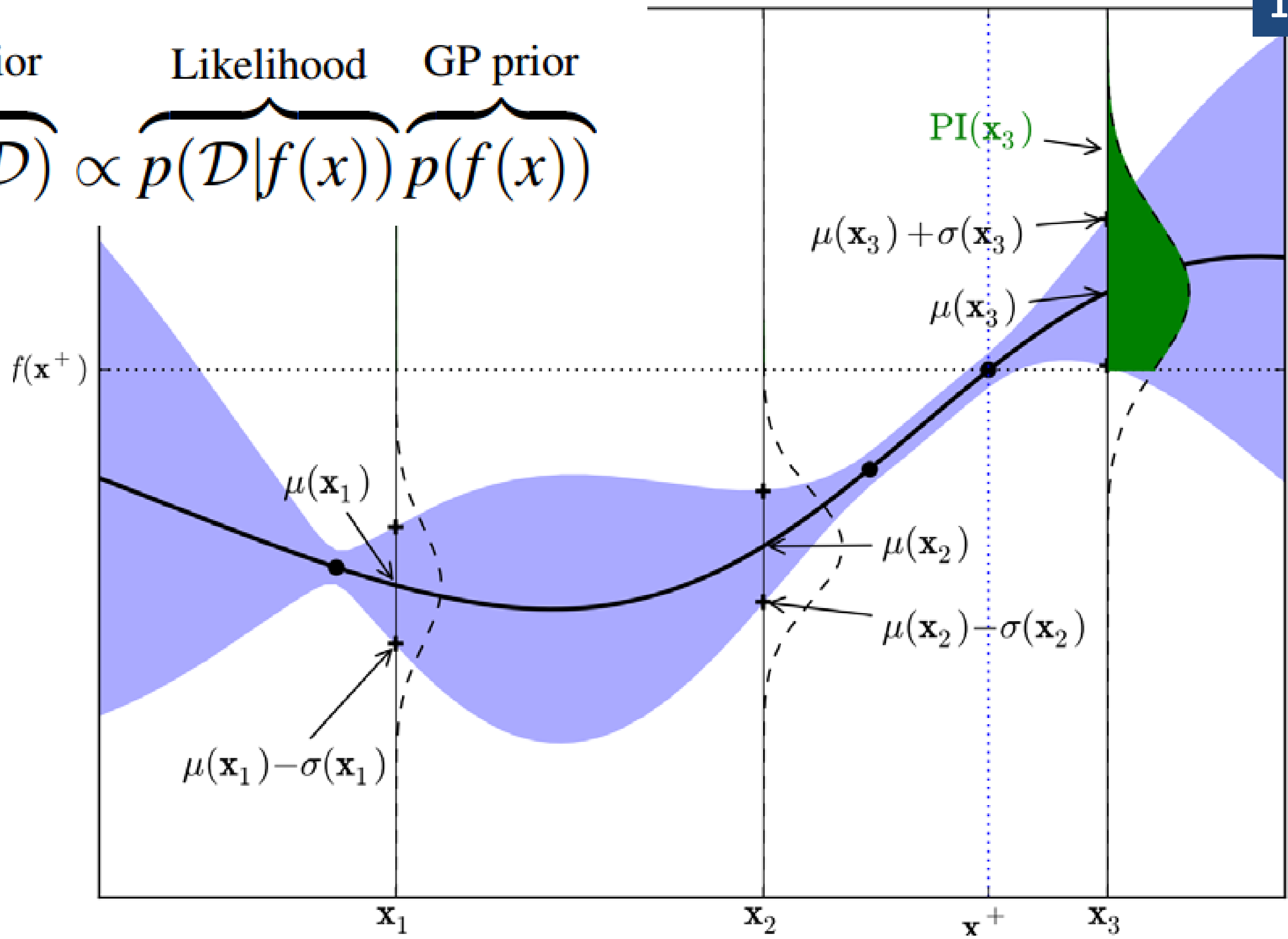




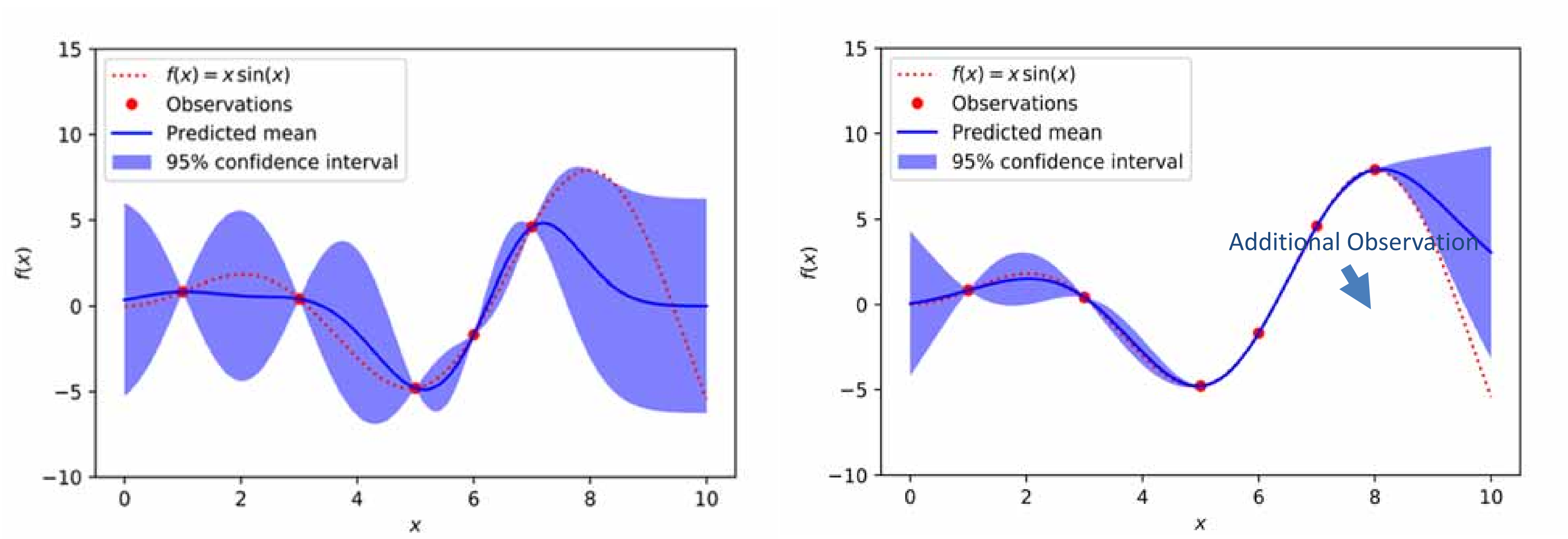
Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research, 55, 361-387, doi:10.1613/jair.4806.

GP = distribution, observations occur in a cont. domain, e.g. t or space

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



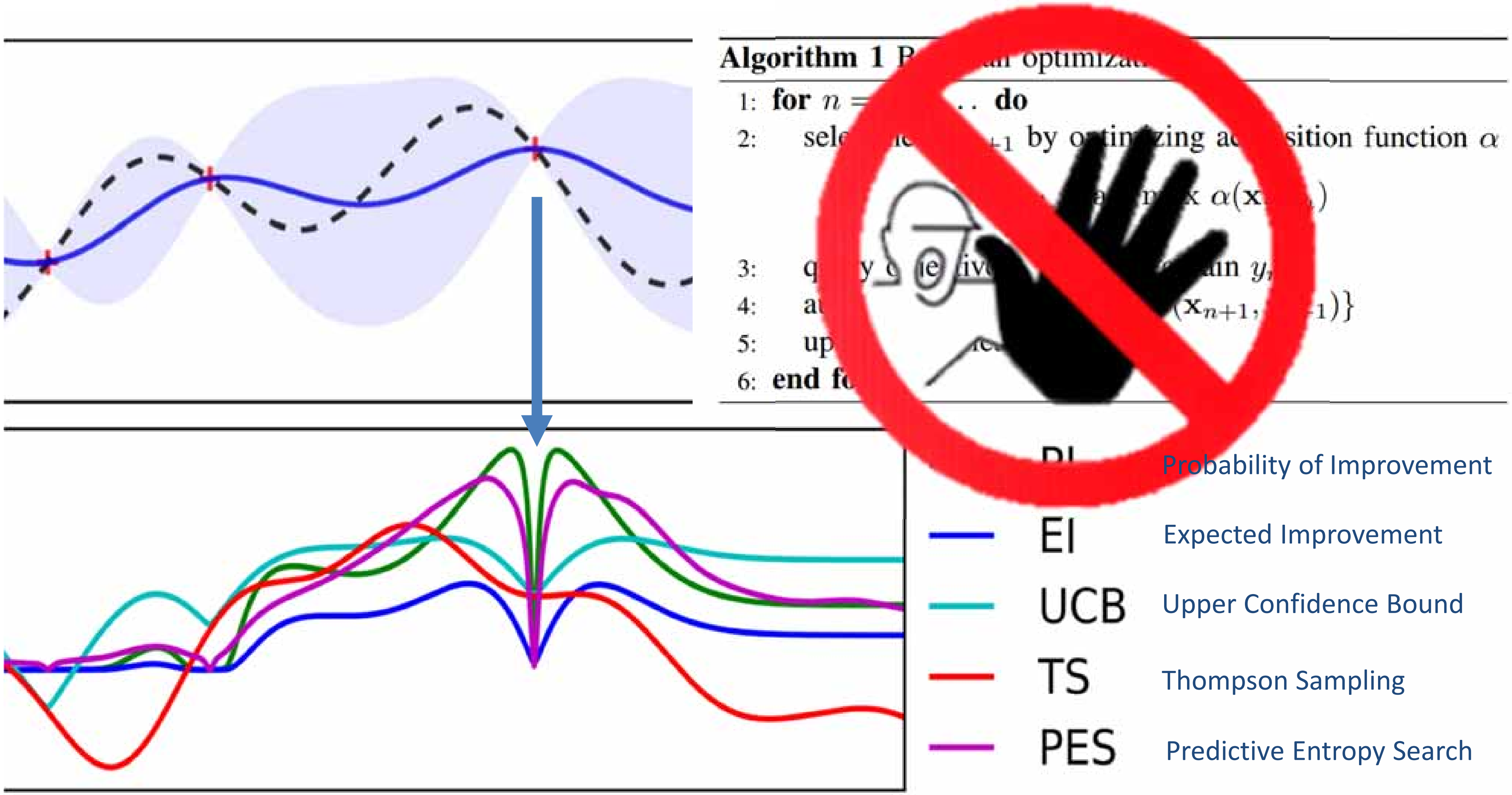
Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.



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

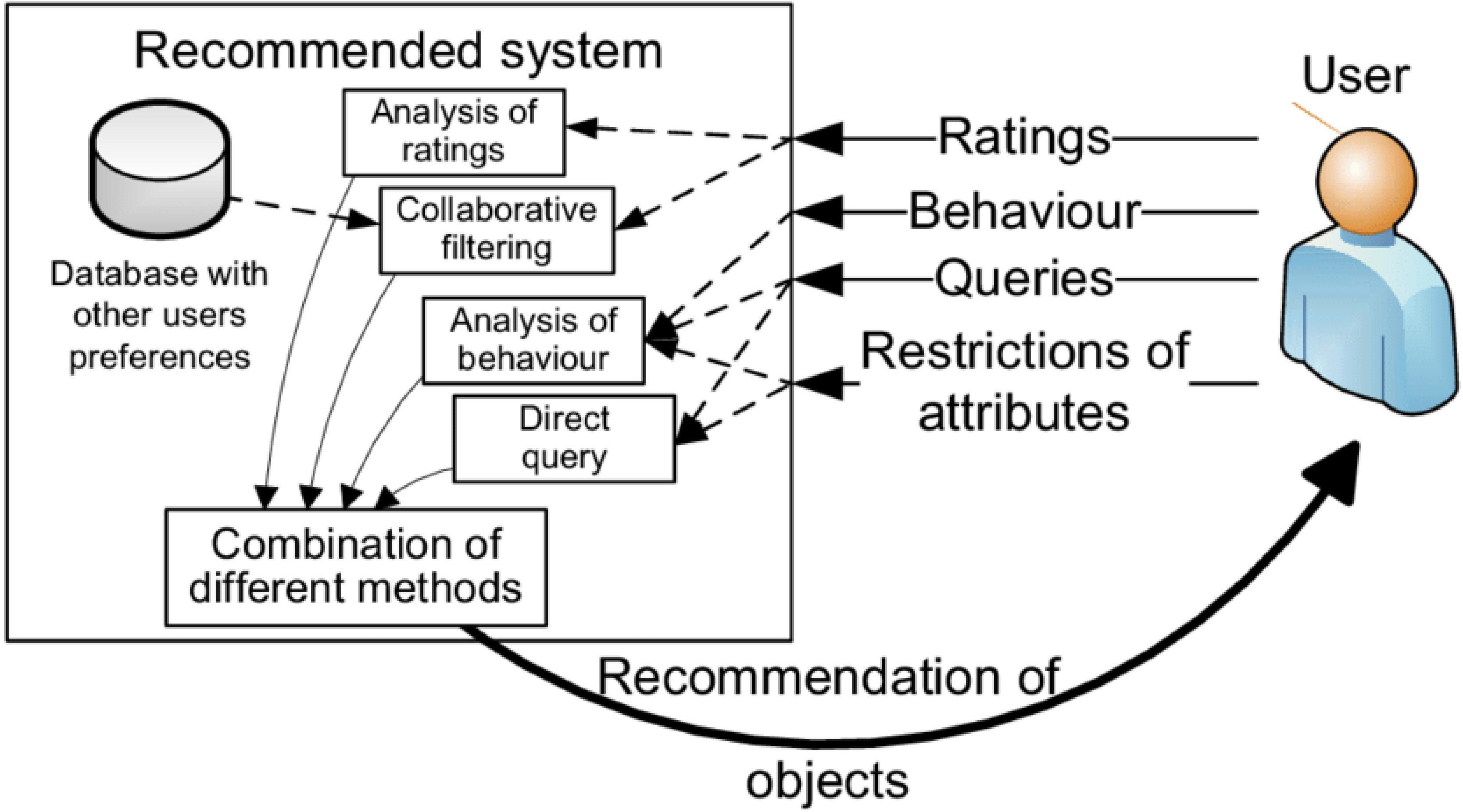
Holzinger, A. 2017. Introduction to Machine Learning and Knowledge Extraction (MAKE). Machine Learning and Knowledge Extraction, 1, (1), 1-20, doi:10.3390/make1010001.

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.

04 aML



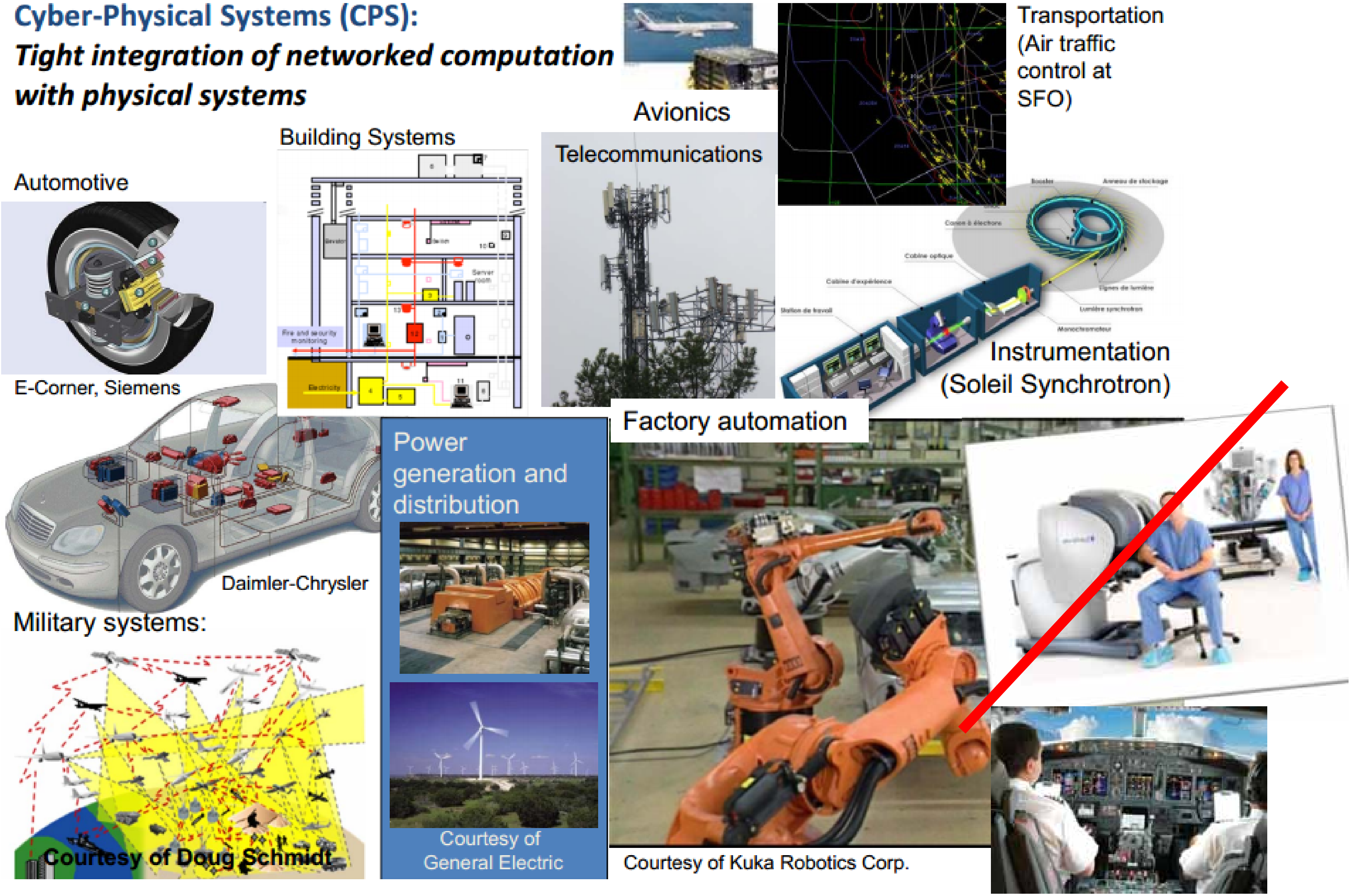
Alan Eckhardt 2009. Various aspects of user preference learning and recommender systems. DATESO. pp. 56-67.

What does level 5 autonomy mean?

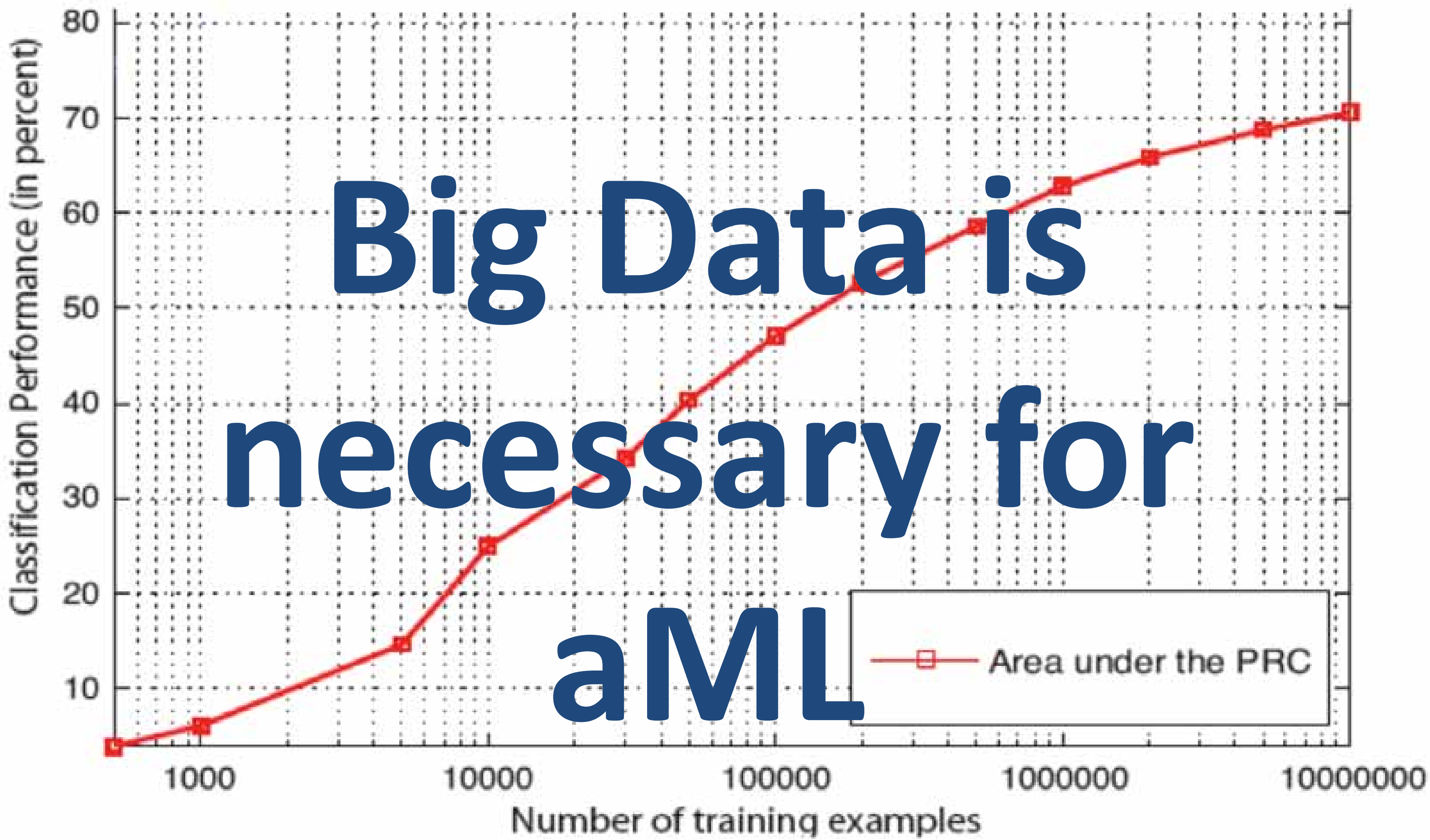
Guizzo, E. 2011. How google's self-driving car works. IEEE Spectrum Online, 10, 18.

Cyber-Physical Systems (CPS):

Tight integration of networked computation with physical systems

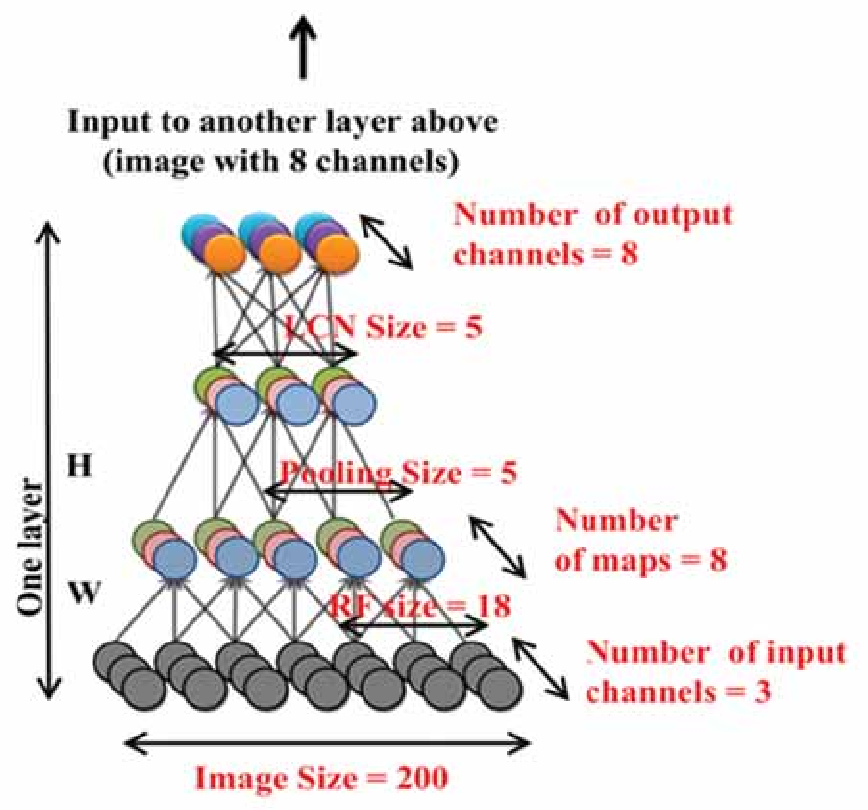


Seshia, S. A., Juniwal, G., Sadigh, D., Donze, A., Li, W., Jensen, J. C., Jin, X., Deshmukh, J., Lee, E. & Sastry, S. 2015. Verification by, for, and of Humans: Formal Methods for Cyber-Physical Systems and Beyond. Illinois ECE Colloquium.



Big Data is necessary for aML

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

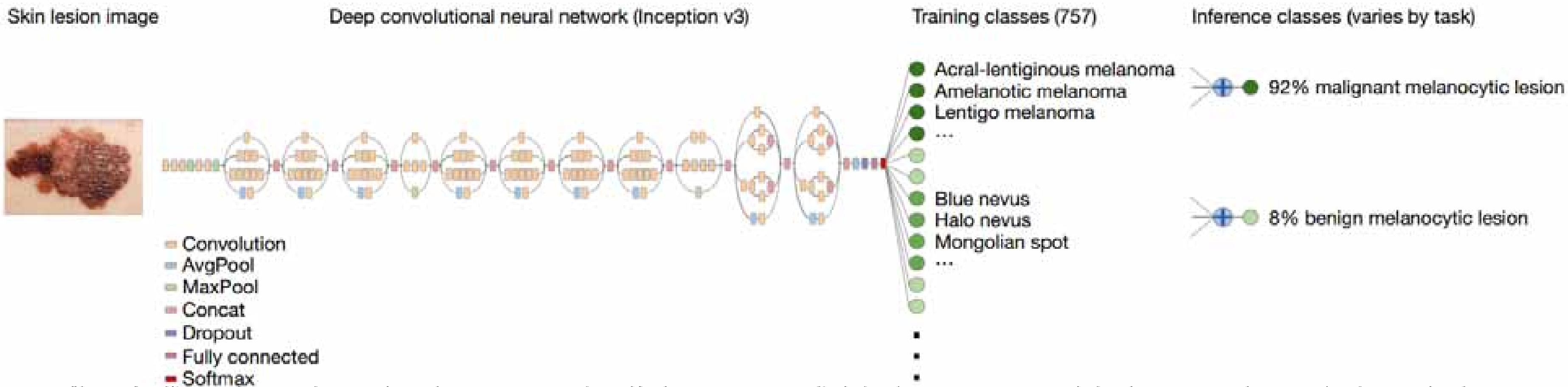


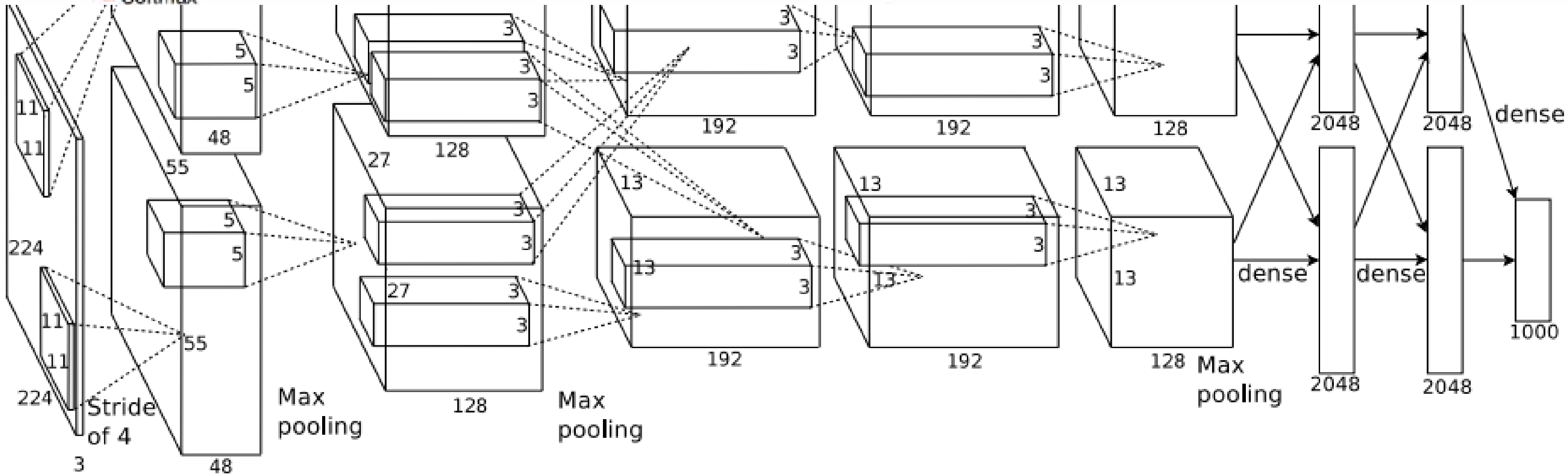
$$x^* = \arg\min_x f(x; W, H), \text{ subject to } ||x||_2 = 1.$$

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.

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.



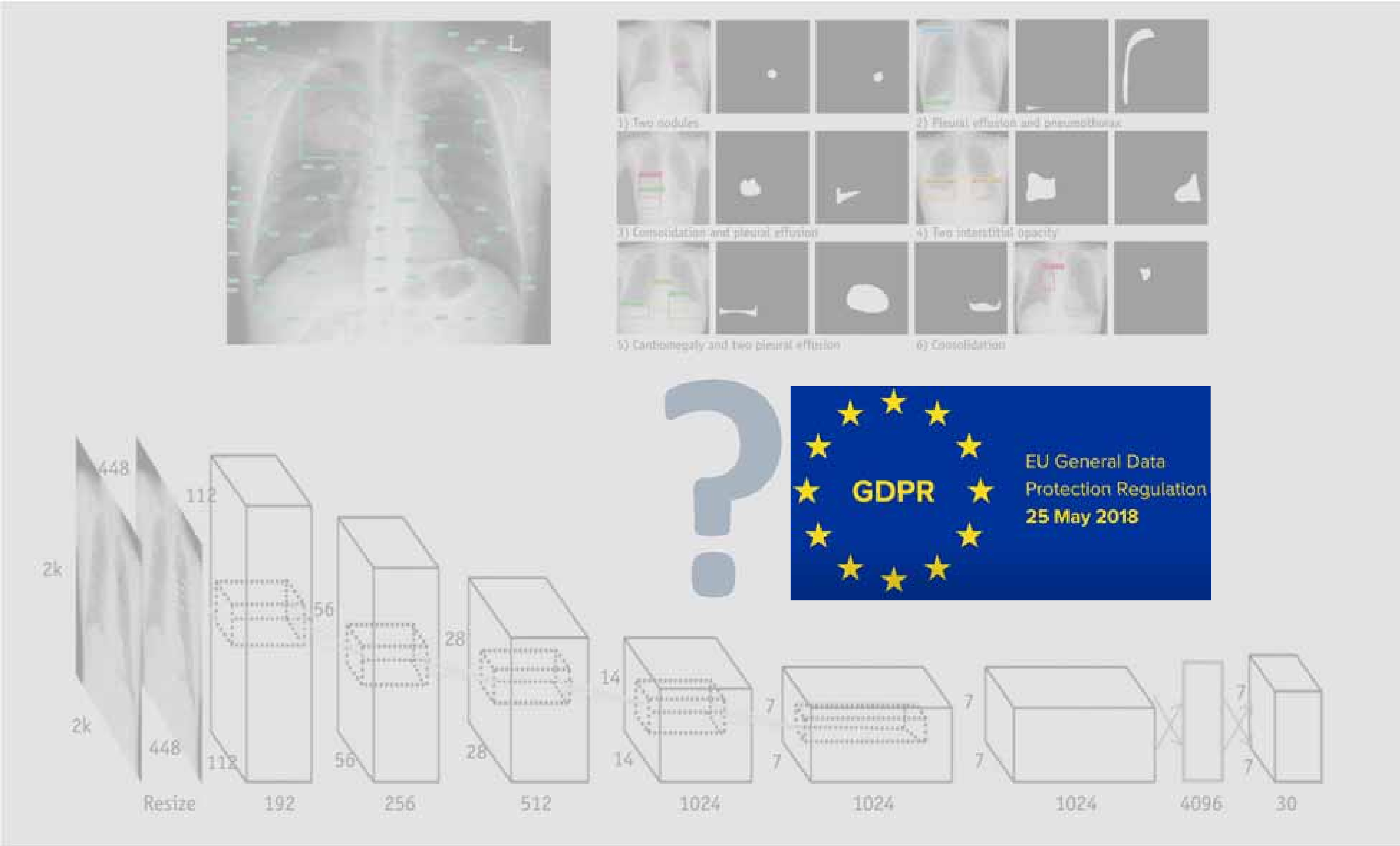


Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

- 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, …

- 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

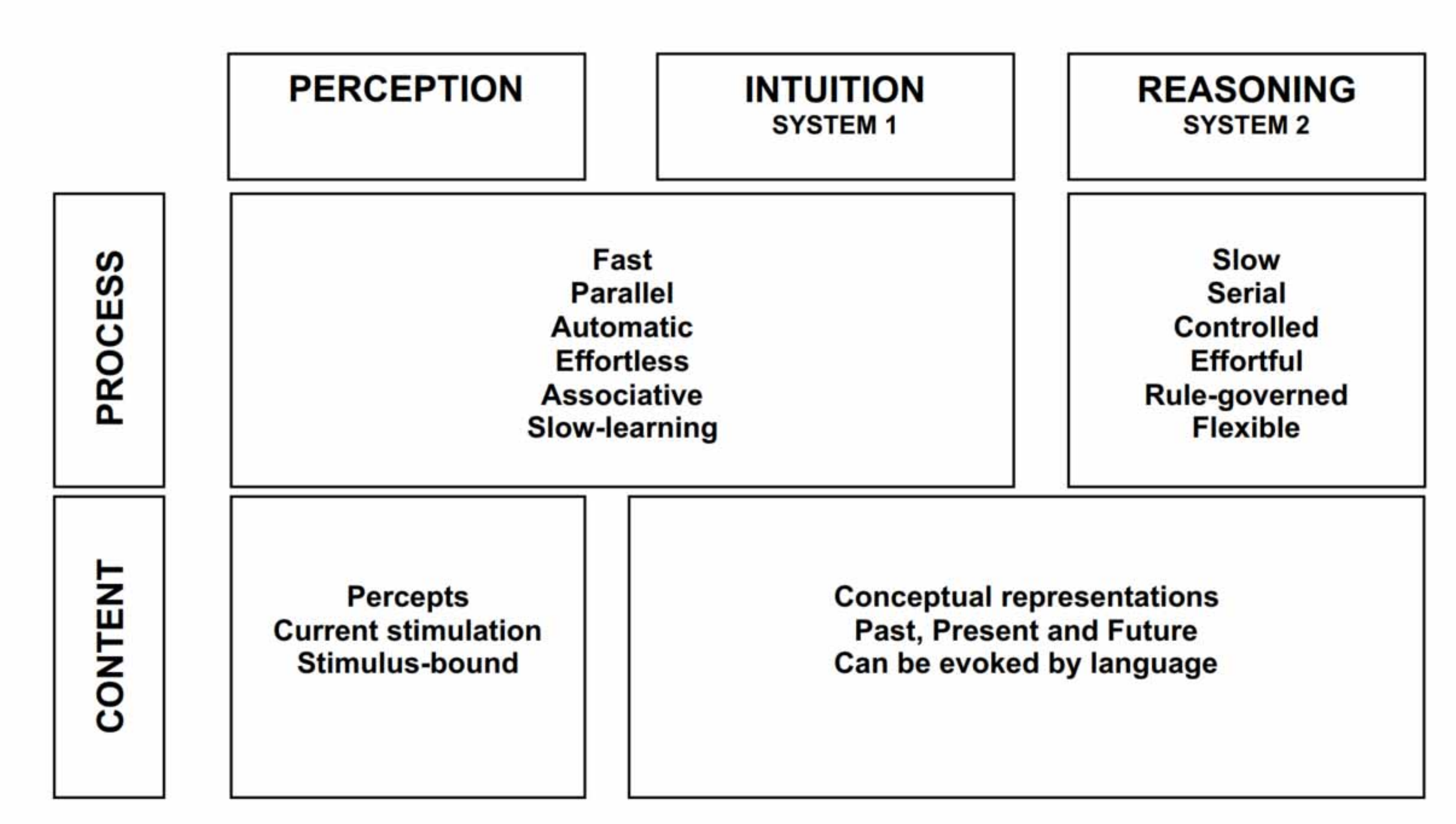
***) where the agents can be human**

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

Humans can understand abstract concepts !

Image Source: 10 Ways Technology is Changing Healthcare http://newhealthypost.com Posted online on April 22, 2018
Image is used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students

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.

- **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.

- := 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.

- **Humans can generalize even from few examples …**
  - They can learn relevant representations
  - Can disentangle the explanatory factors
  - Find the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

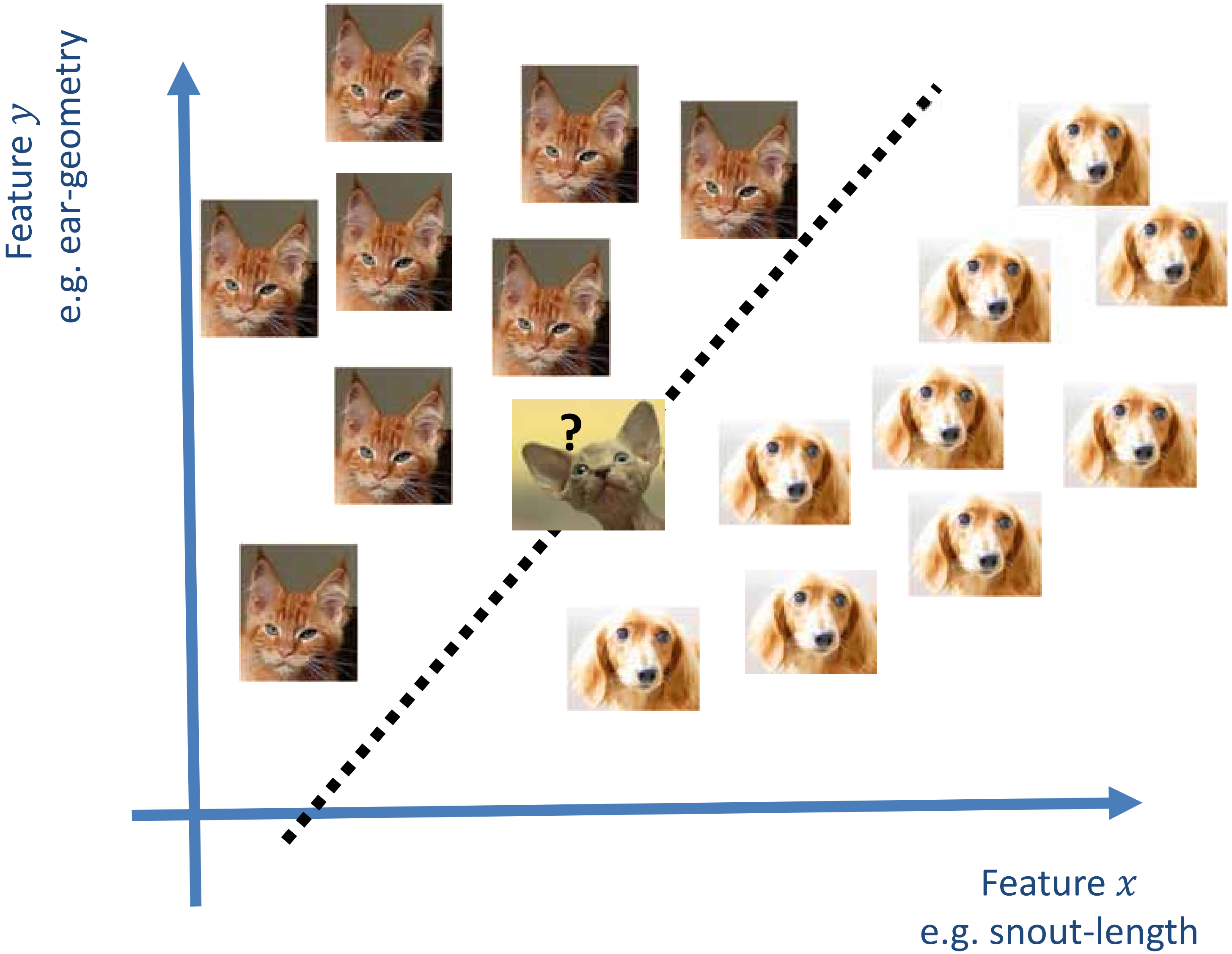
Yoshua Bengio, Aaron Courville & Pascal Vincent 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

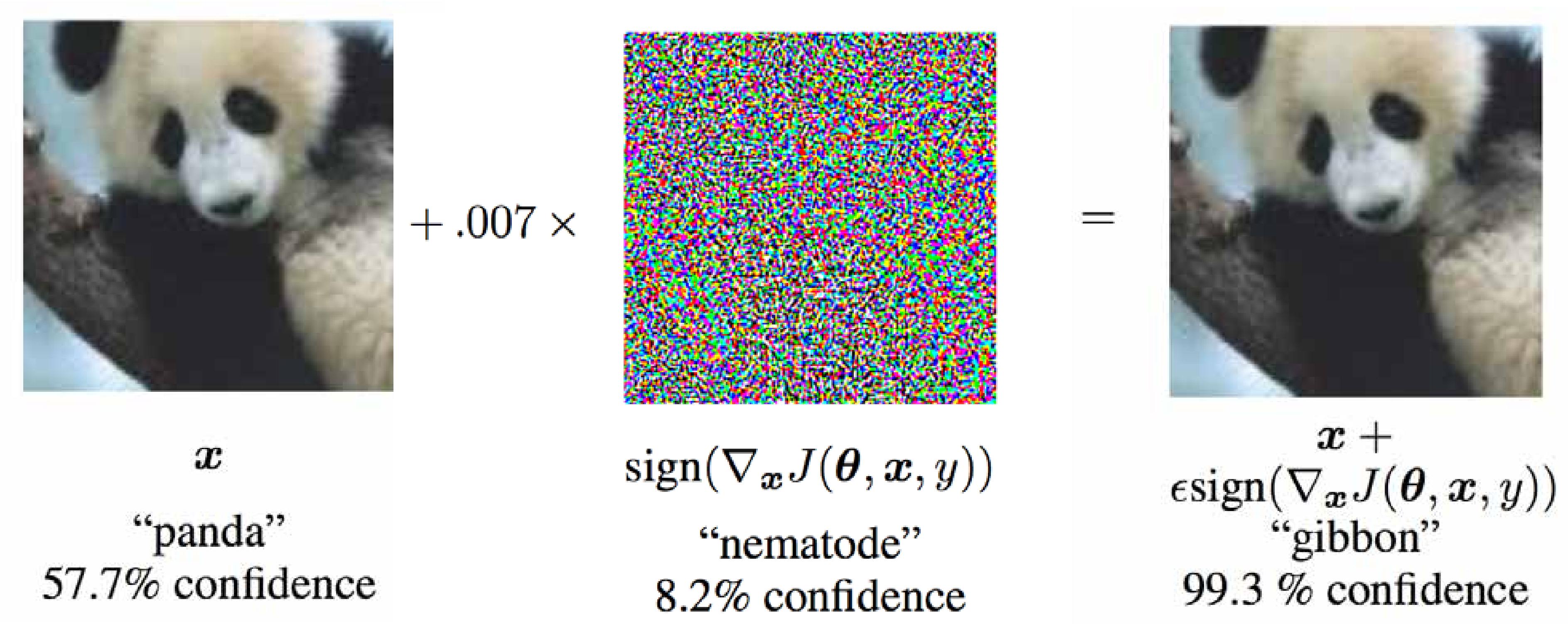
Even Children can make inferences from little, noisy, incomplete data …

This image is in the public domain, Source: freedesignfile.com
Image is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

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

In classification tasks what can we do with rare examples?





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

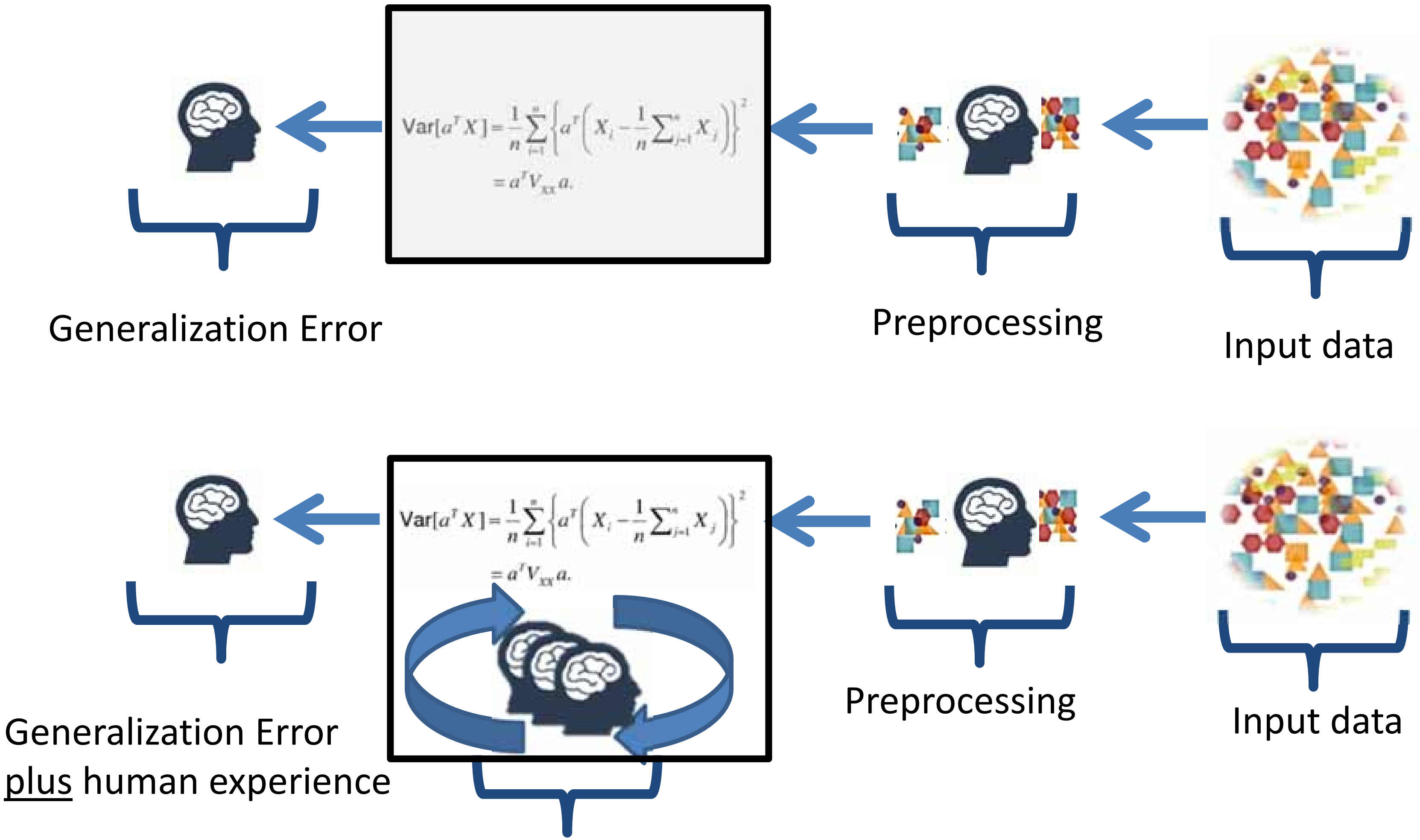
Jascha Sohl-Dickstein
Google Brain
jaschasd@google.com

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.

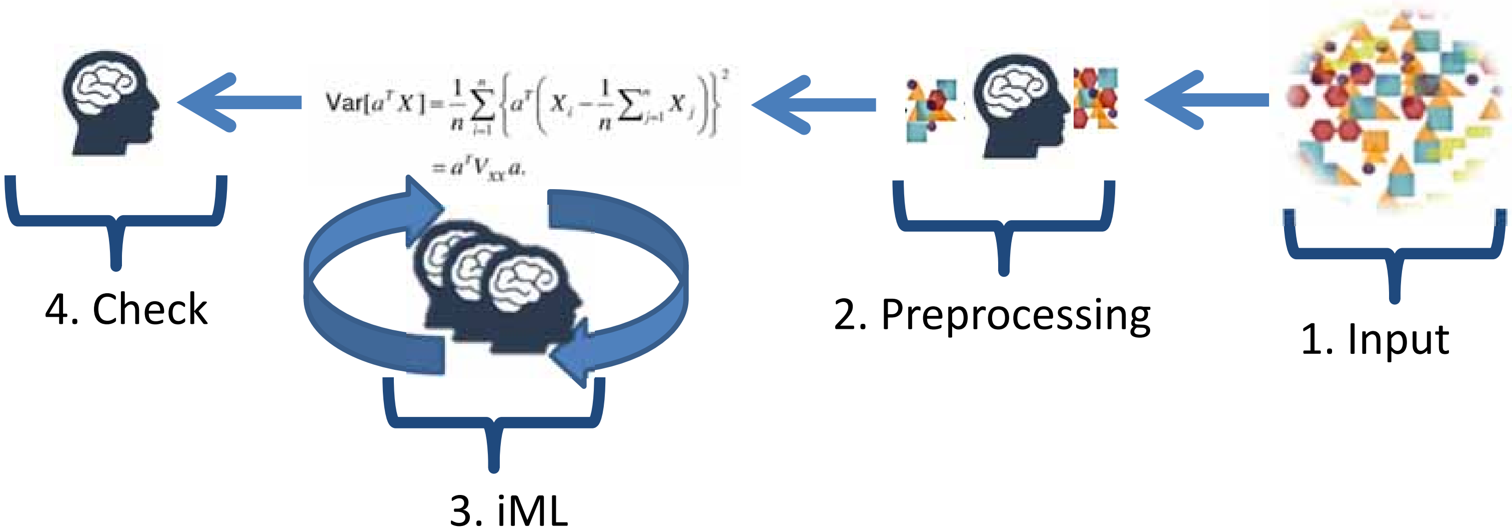
Where is the benefit of the human-in-the-loop ?



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.

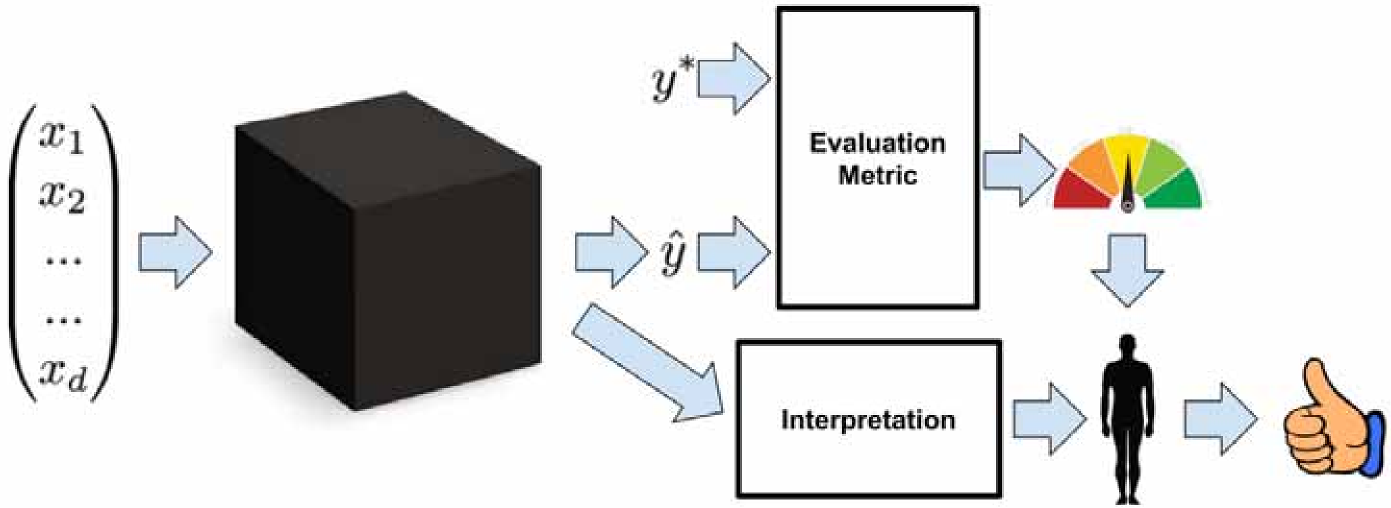
Interactive Machine Learning: Human is seen as an agent involved in the actual learning phase, step-by-step influencing measures such as distance, cost functions …



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.

06
"explainable AI"

Term coined by Dave Gunning, DARPA, see:
David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.



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

- Inconsistent Definitions: What is the difference between explainable, interpretable, verifiable, intelligible, transparent, understandable … ?

Zachary C. Lipton 2018. The mythos of model interpretability. ACM Queue, 16, (3), 31-57, doi:10.1145/3236386.3241340

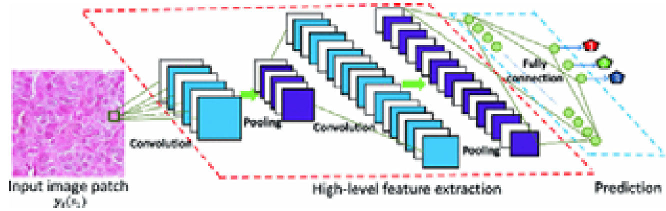
- **Trust** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?
- **Causality** - inferring causal relationships from pure observational data has been extensively studied (Pearl, 2009), however it relies strongly on prior knowledge
- **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.

End-users shall be able to retrace the results on demand and we engineers need to understand our own machine learning models!

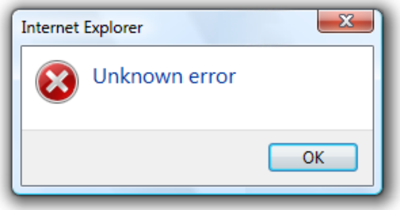
Verify that algorithms/classifiers work as expected …

Wrong decisions can be costly and dangerous …



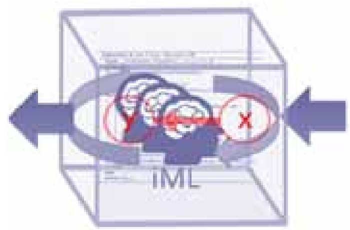
Understanding the errors …

Detection of bias, weaknesses, unknowns, …



Scientific replicability and causality …

The "why" is often more important than the prediction …



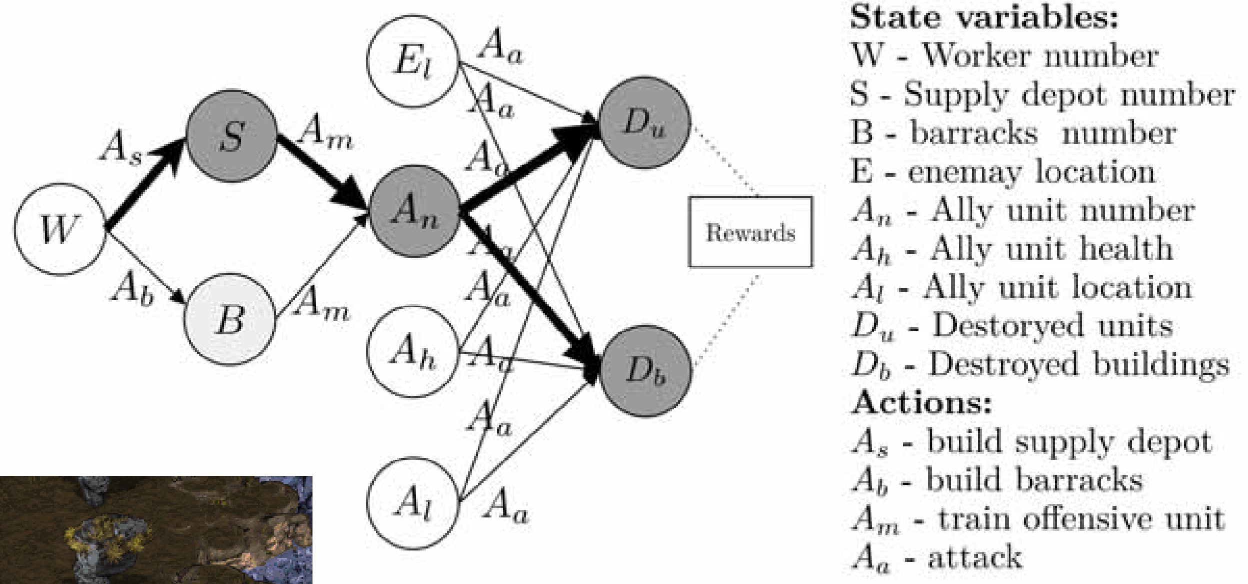
Andreas Holzinger 2018. From Machine Learning to Explainable AI. 2018 World Symposium on Digital Intelligence for Systems and Machines (IEEE DISA). pp. 55-66, doi:10.1109/DISA.2018.8490530.

- **Post-Hoc** (latin) = after- this (event), i.e. such approaches provide an explanation for a specific solution of a "black-box" approach, e.g. LIME, BETA, LRP, …
- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.
- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems, e.g. decision trees, linear regression, random forests, ...

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*.

- **Interpretable Models,** the model itself is already interpretable, e.g.
  - Regression
  - Naïve Bayes
  - Random Forests
  - Decision Trees/Graphs
  - …
- **Interpreting Black-Box Models** (the model is not interpretable and needs a post-hoc interpretability method – like a combustion engine ;-) e.g.:
  - Decomposition
  - LIME/BETA
  - LRP
  - …

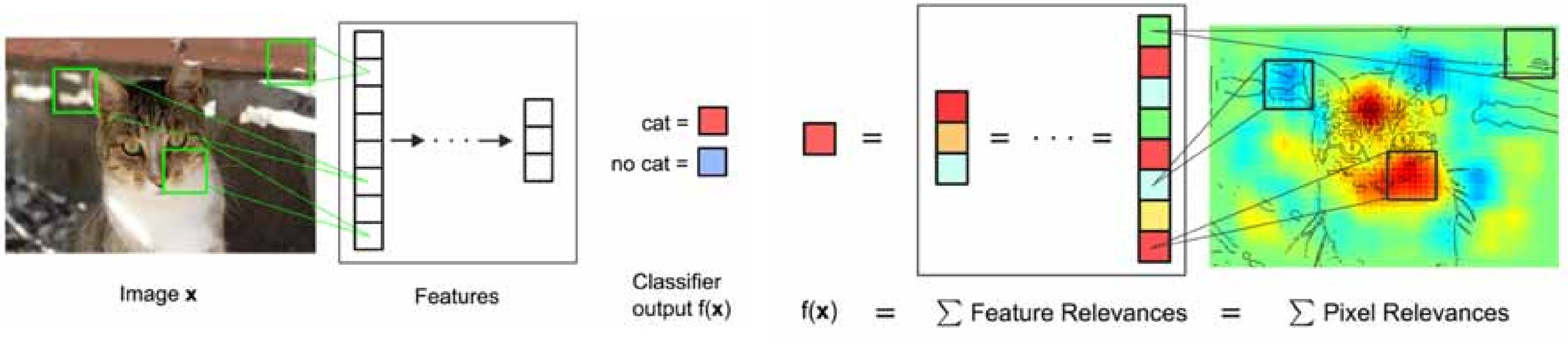
What is a typical example for the ante-hoc approach ?



https://eecs.wsu.edu/~ala/cdtldms/reports/maf_report.htm

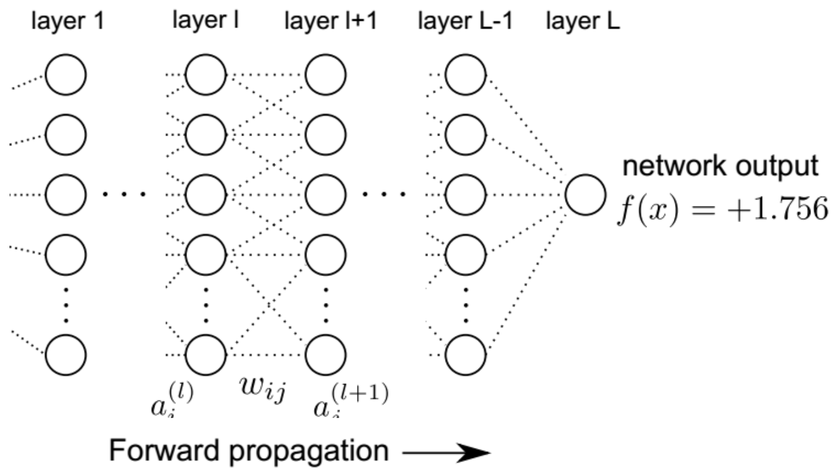
Prashan Madumal, Tim Miller, Liz Sonenberg & Frank Vetere 2019. Explainable Reinforcement Learning Through a Causal Lens. arXiv preprint arXiv:1905.10958.

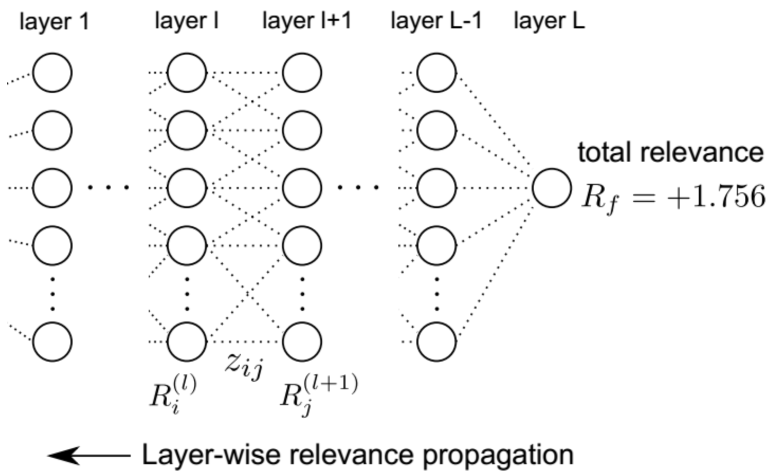
What is a typical example for a post-hoc method ?



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

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

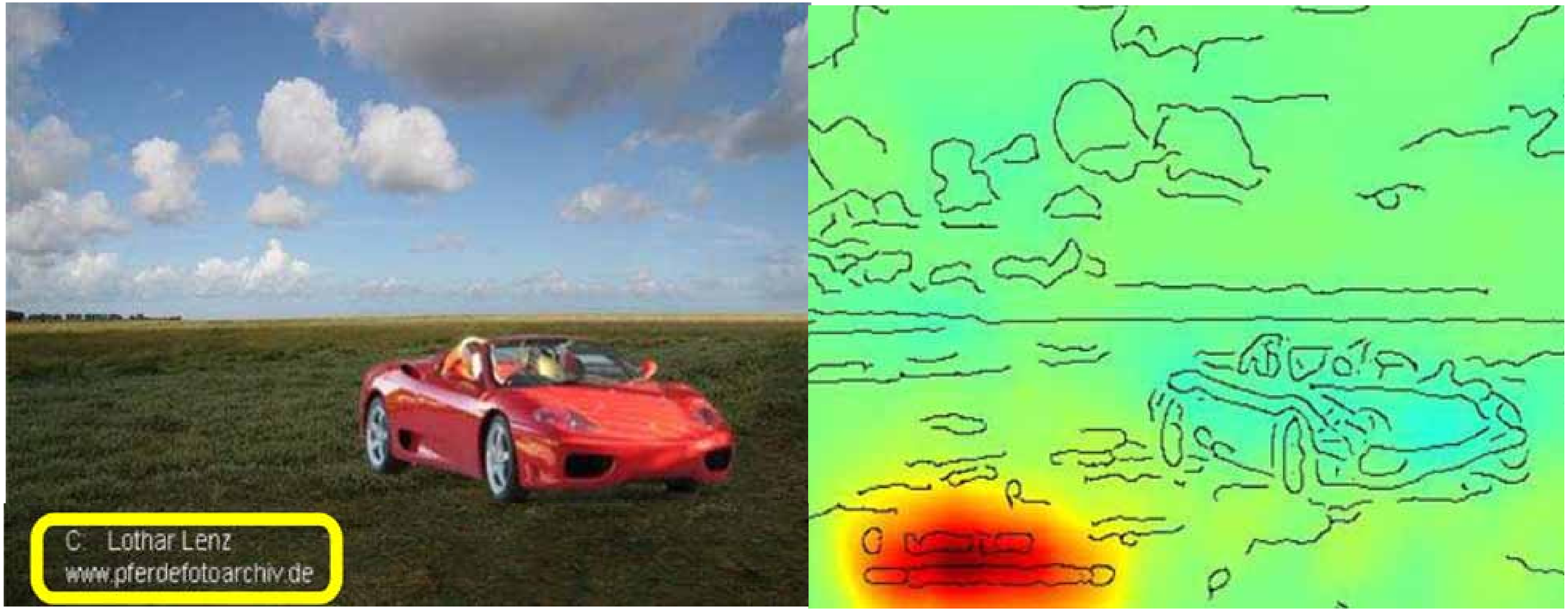




$$R_i = \left\|\frac{\partial}{\partial x_i} f(\mathbf{x})\right\| \qquad \sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x})$$

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

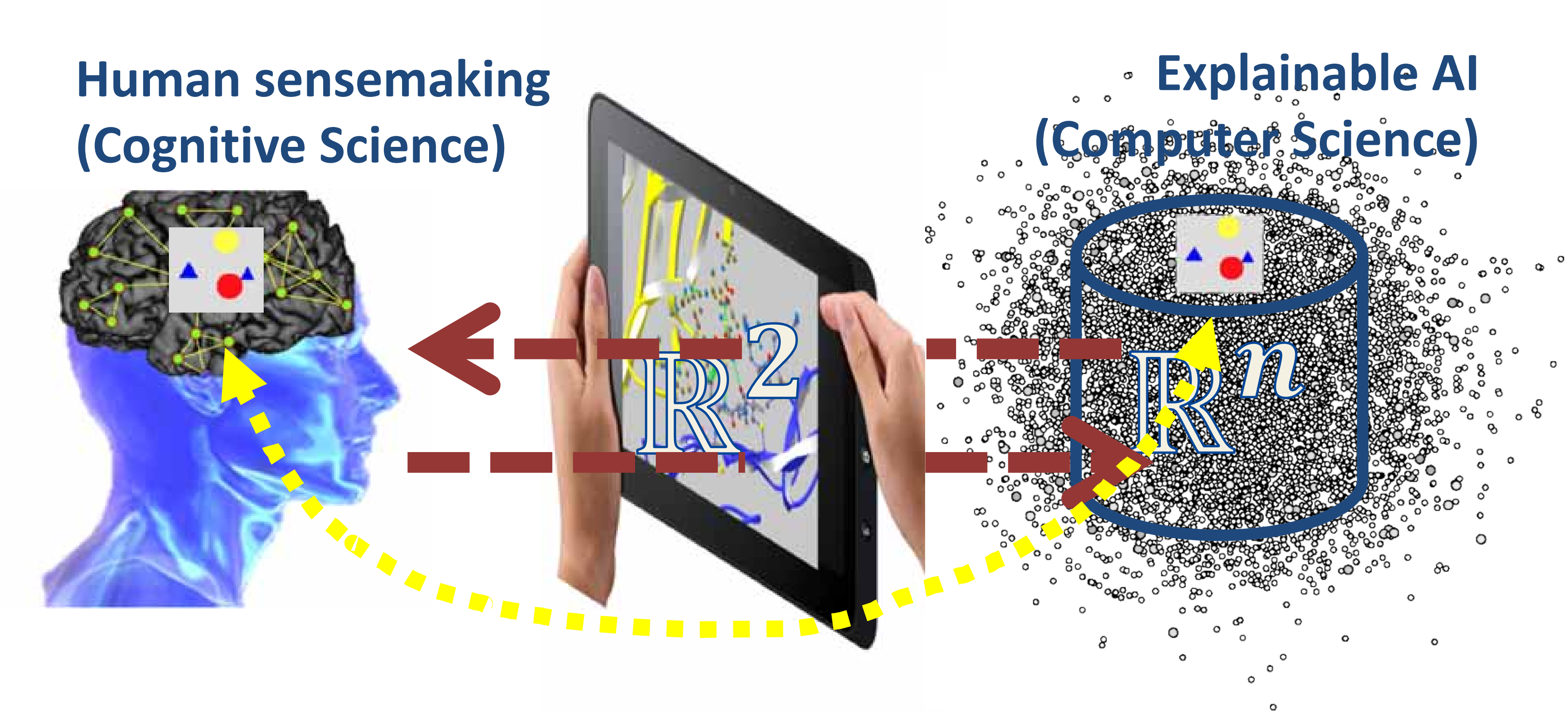
98 % A horse on a meadow



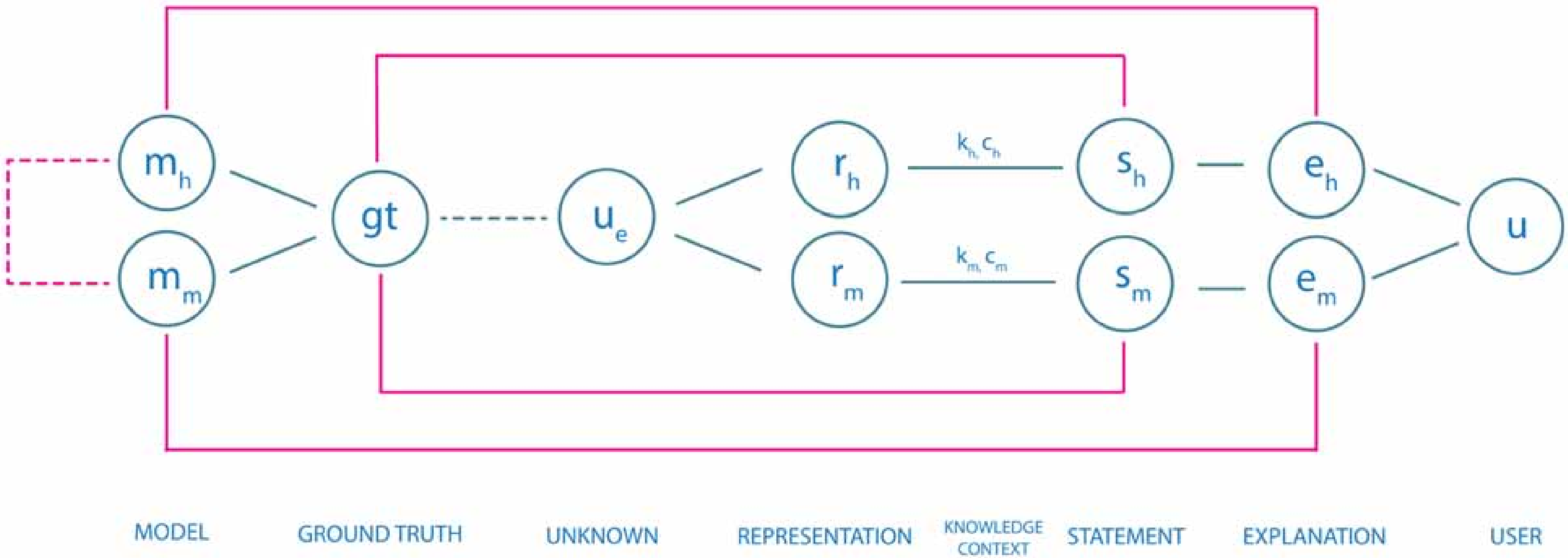
Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek & Klaus-Robert Müller 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications, 10, (1), doi:10.1038/s41467-019-08987-4.

A final note on Measuring Causability: Mapping machine explanations with human understanding

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

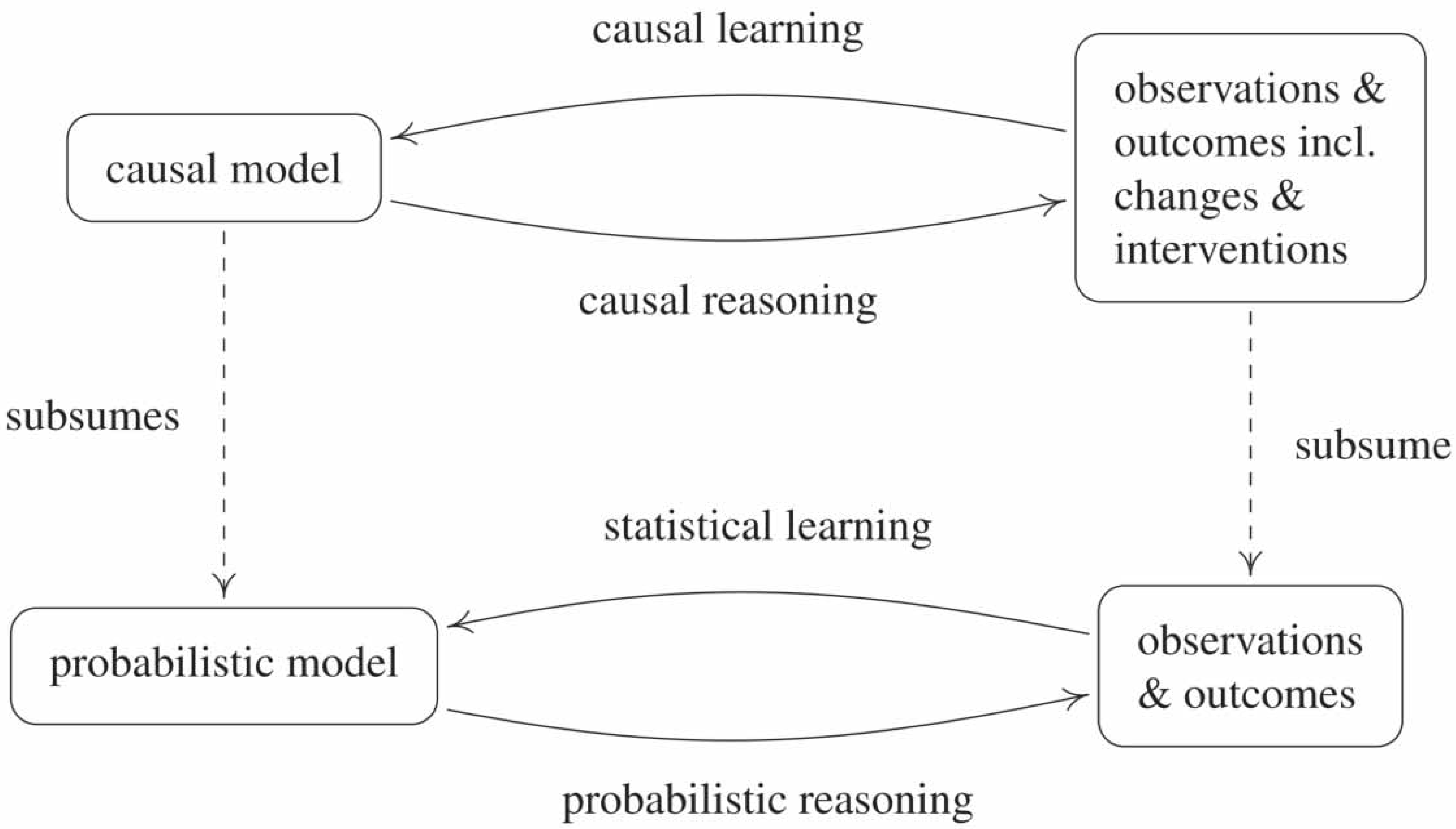


Andreas Holzinger et al. 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.



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., https://link.springer.com/article/10.1007/s13218-020-00636-z

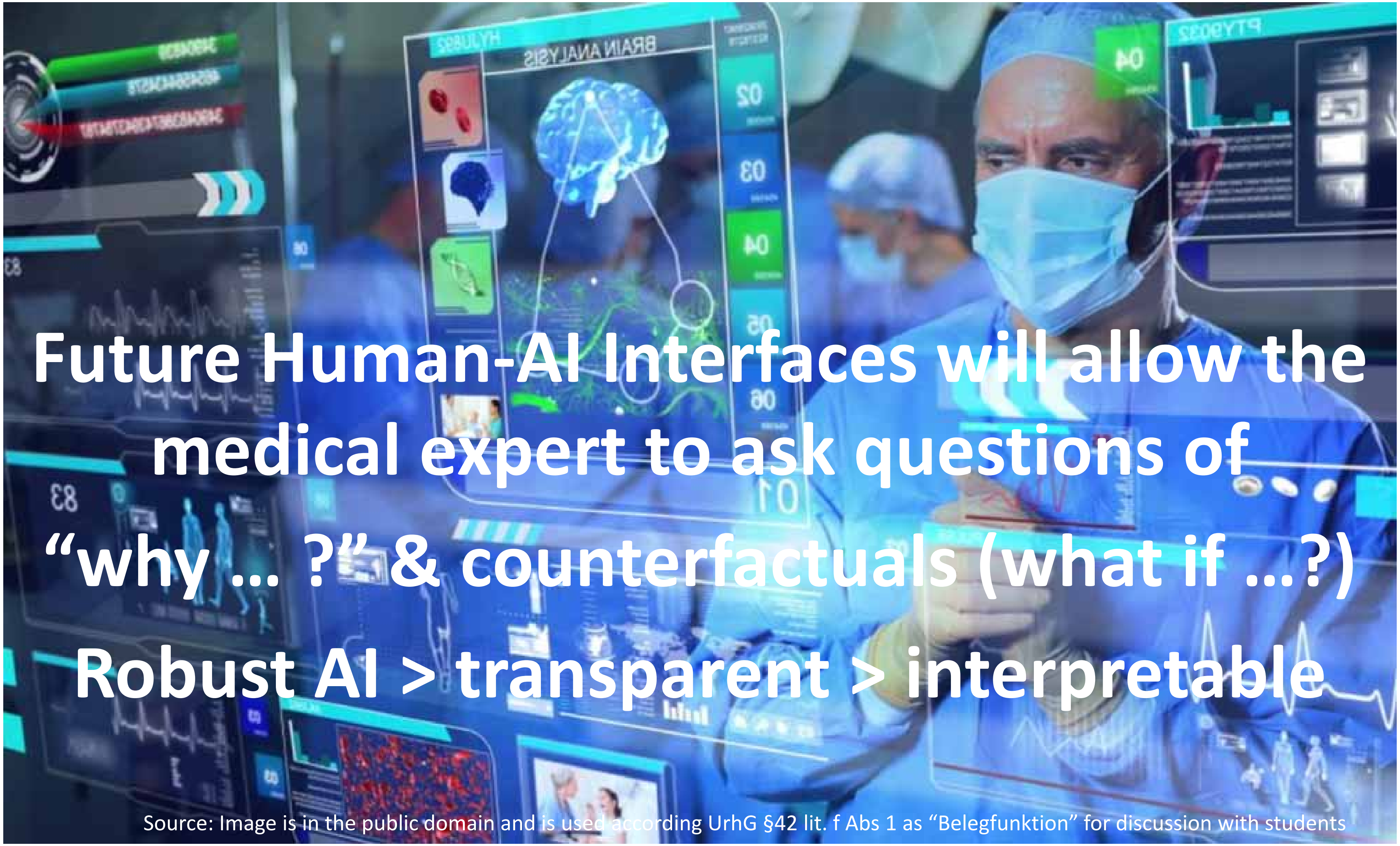
Why is Causality going beyond explainability ?



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

Conclusion and Future Outlook

What should future Human-AI interfaces ensure ?



Future Human-AI Interfaces will allow the medical expert to ask questions of "why … ?" & counterfactuals (what if …?)

Robust AI > transparent > interpretable

Source: Image is in the public domain and is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

- Current AI does not generalize well,
- can not learn from few examples,
- do not infer causal relationships.

We need robust AI to

- reduce costs and limitations

Fostering acceptance and trust, enabling ethical, legal and socially responsible medical AI

Replicability, Retraceability, Explainability & Causability

- Computers are fast, accurate and stupid,
- humans are slow, inaccurate and brilliant,
- **together** they are powerful beyond imagination (attributed to Albert Einstein)

(Einstein never said that)

https://www.benshoemate.com/2008/11/30/einstein-never-said-that