Human-Centered AI Research Seminar (Part 1)

Andreas Holzinger

Human-Centered AI (Holzinger Group)
Institute for Medical Informatics/Statistics, Medical University Graz, Austria
and
Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

@aholzin #KandinskyPatterns

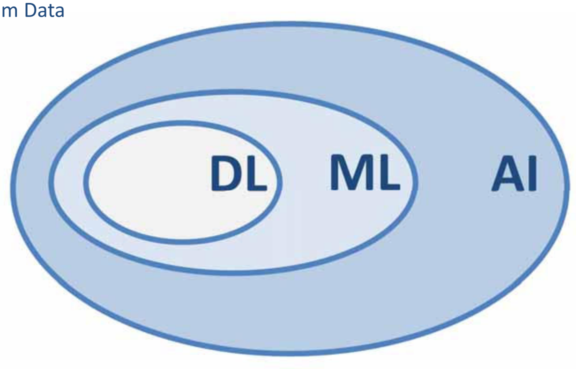
Course Homepage: https://human-centered.ai/hcai-research-seminar-2020/

Remark

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

Preamble to ensure mutual understanding

AI = Artificial Intelligence (in German: KI, Künstliche Intelligenz)
ML = Machine Learning (in German: ML, Maschinelles Lernen)
DL = Deep Learning
aML = automatic (autonomous) ML
iML = interactive ML
HCI = Human-Computer Interaction
KDD = Knowledge Discovery from Data
HCAI = Human-centered AI
HAII = Human AI Interfaces
Ex-AI = explainable AI (also XAI)



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

A word at the very beginning

- The “best” is the enemy of the “good” – whenever you try to be “perfect” – there is the danger that you finalize nothing*) …”

*) zero, nada, null

François-Marie Arouet (1694 – 1778)
known as “Voltaire”

Our motto …

Science is to test crazy ideas – Engineering is put these ideas into Business!

Learning Goals

- At the end of this research seminar you should
- … be aware of the HCAI approach
- … know some current hot topics of AI/ML
- … have an overview on possible research topics
- … be familiar with MSc/PhD requirements
- … understand how to carry out scientific research
- … know how to write scientific papers
- … most of all: getting started with your work

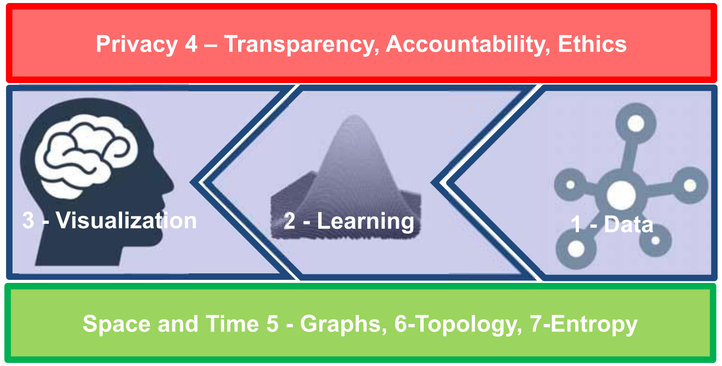
Agenda

- 01 What is the HCAI approach?
- 02 Application Area: Health
- 03 Probabilistic Information
- 04 Gaussian Processes
- 05 Automatic Machine Learning (aML)
- 06 Interactive Machine Learning (iML)
- 07 Explainable AI (Why explainability?)
- 08 #KandinskyPatterns Framework

01 What is the HCAI approach?

- ML is a very practical field – algorithm development is at the core – however, successful ML needs a concerted effort of various topics …

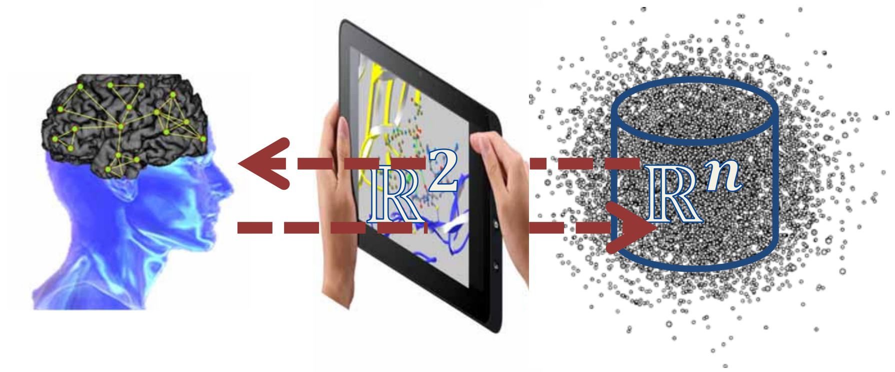
Human-Centered Artificial Intelligence …


Privacy 4 – Transparency, Accountability, Ethics
3 - Visualization
2 - Learning
1 - Data
Space and Time 5 - Graphs, 6-Topology, 7-Entropy

Andreas Holzinger 2013. Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Heidelberg, Berlin, New York: Springer, pp. 319-328, doi:10.1007/978-3-642-40511-2_22

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

Artificial Intelligence needs the Human-in-control



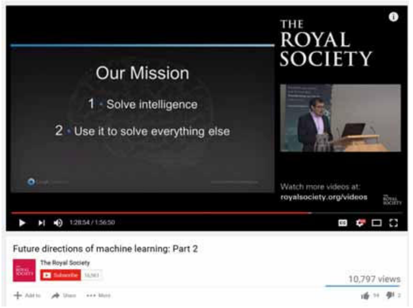
Our goal is that human values are aligned to ensure responsible machine learning

Not our Goal: Humanoid AI

This image is in the public domain

What is intelligence and is Intelligence enough?

“Solve intelligence – then solve everything else”



THE ROYAL SOCIETY

Our Mission

1 · Solve intelligence

2 · Use it to solve everything else

Watch more videos at: royalsociety.org/videos

1:28:54 / 1:56:50

Future directions of machine learning: Part 2

The Royal Society

10,797 views

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

Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

What makes a machine intelligent? Cross-cutting issues

- To hear, to see, to talk, to smell, taste, touch, …
  - Speech recognition, computer vision, natural language processing (olfactory, gustatory sensors)
- To store, to memorize, to represent, to access, …
  - Knowledge representation, semantic networks, ontologies, information retrieval
- To learn from data, to extract knowledge, …
  - Improve with experience from previous events
- **To reason, to understand, to reflect, …**
  - Logic AND Bayesian inference, contextual understanding, ground truth for explanation framework

To reach a level of usable intelligence we need to …

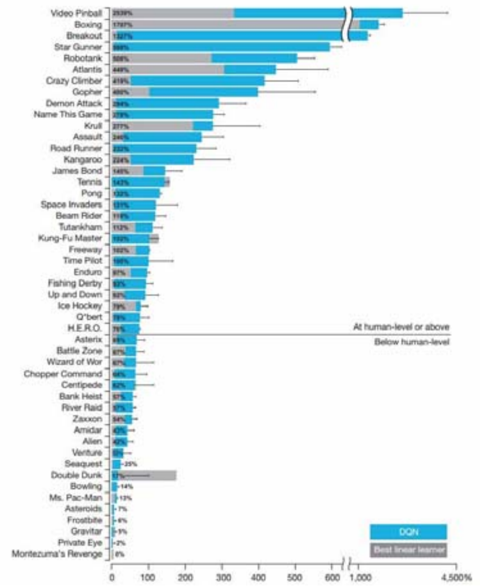
- 1) learn from prior data
- 2) extract knowledge
- 3) generalize, i.e. guessing where a probability mass function concentrates
- 4) fight the curse of dimensionality
- 5) disentangle **underlying explanatory factors of data**, i.e.
- 6) **understand** the data in the **context** of an application domain

Our goal is understanding context !

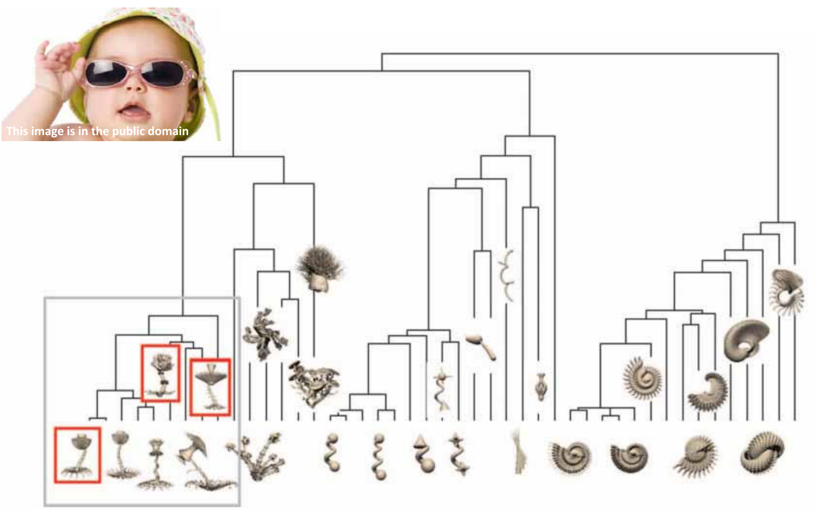
How far are we already ?

Now, compare your best Machine Learning 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



Learning complex concepts from a few examples



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

What if we succeed? (Singularity Irving J. Good)

Image credit to http://history.computer.org/pioneers/good.html

"An ultra-intelligent machine could design even better machines; there would then unquestionably be an **"intelligence explosion*"** and the intelligence of man would be left far behind …

It is curious that this point is made so seldom … outside of science fiction."

Irving John Good, Trinity College, Oxford, 1965
Colleague of Alan Turing in Bletchley Park

Good, I. J. 1966. Speculations Concerning the First Ultraintelligent Machine*. In: Franz, L. A. & Morris, R. (eds.) Advances in Computers. Elsevier, pp. 31-88, doi:10.1016/S0065-2458(08)60418-0

https://web.archive.org/web/20010527181244/http://www.aeiveos.com/~bradbury/Authors/Computing/Good-IJ/SCtFUM.html

*) https://intelligence.org/ie-faq

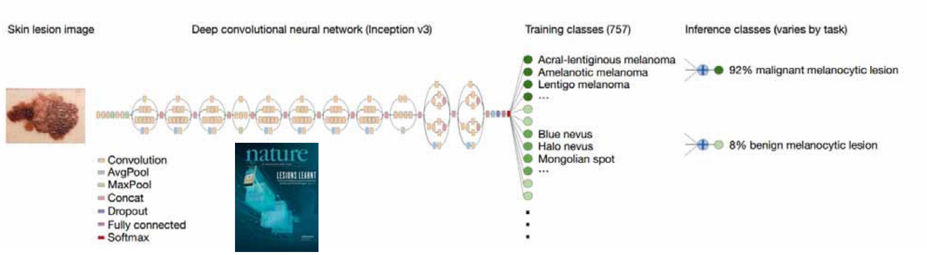
Today ML is enormously progressing …

- Progress is driven by the explosion in the availability of **big data** and **low-cost computation.**
- **Health is amongst the biggest challenges**

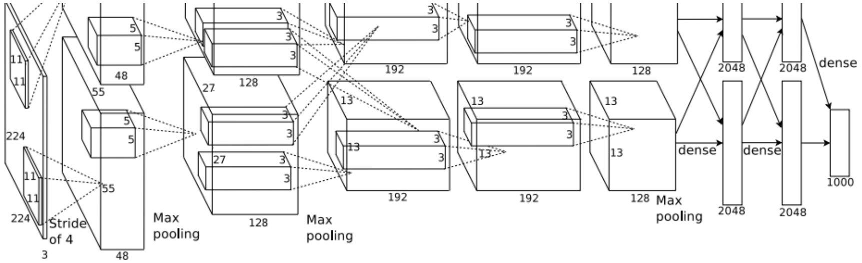
Jordan, M. I. & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. Science, 349, (6245), 255-260.

What it takes to end an AIDS epidemic
Polar bears suffer through lean summers
Sperm produced in ovary of mutant fish
Science
AAAS
nature
LEARNING CURVE
Self-taught AI software attains human-level performance in video games
A GIANT IN THE EARLY UNIVERSE
TELEPORTATION FOR TWO

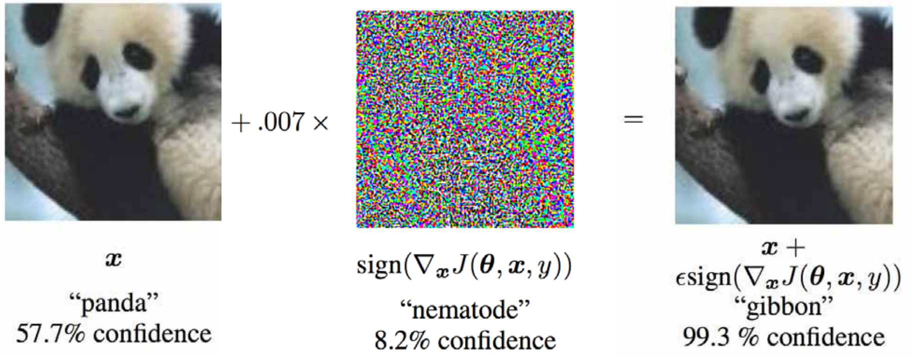
Deep Learning today is very successful



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.



Correlation does not imply Causation


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

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

Urgent need for explainable AI !

02 Application Area Health Informatics

Image Source: LKH Feldbach, Steiermark

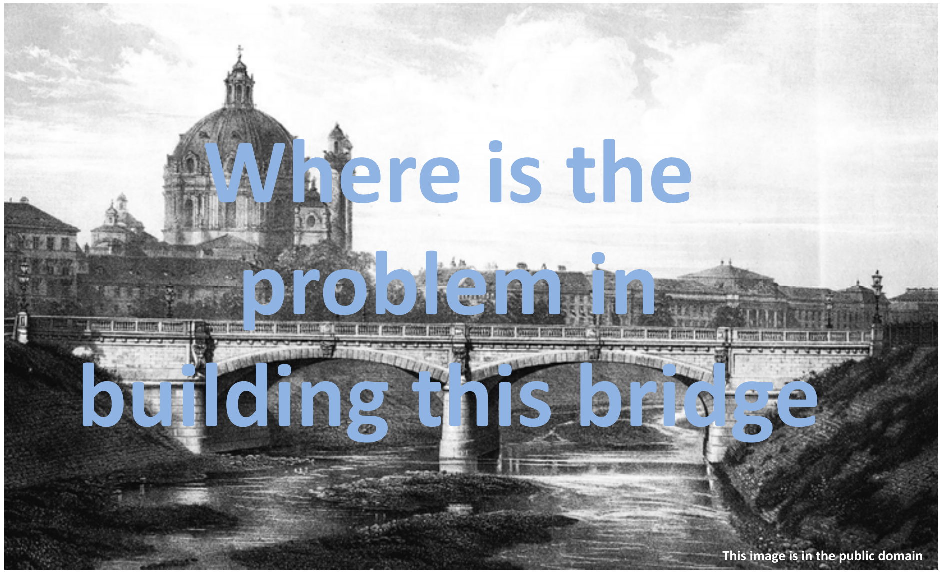
Health is a complex area

Why is this application area complex ?

In medicine we have two different worlds …

Our central hypothesis: Information may bridge this gap

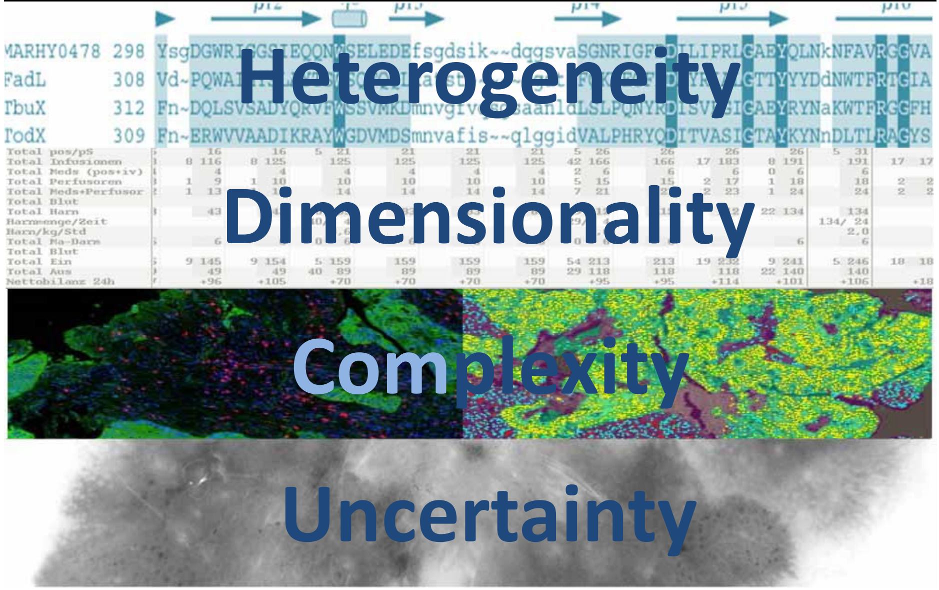
Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*



Where is the problem in building this bridge

This image is in the public domain

Main problems …



Heterogeneity

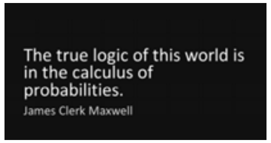
Dimensionality

Complexity

Uncertainty

Holzinger, A., Dehmer, M. & Jurisica, I. 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics, 15, (S6), I1.

03 Probabilistic Learning



The true logic of this world is in the calculus of probabilities.

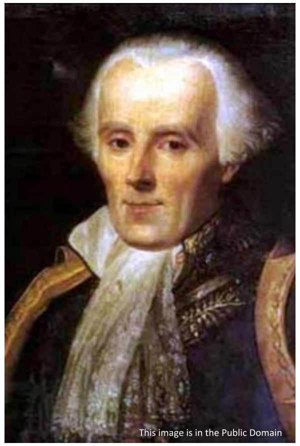
James Clerk Maxwell

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

The foundation for modern machine learning …

- **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


This image is in the Public Domain

Pierre Simon de Laplace (1749-1827)

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

Repetition of Bayes - on the work of Laplace

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:

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

Learning representations (*θ*, *h*) from observed data

Observed data:

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

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* $p(\theta|\mathcal{D})$ $p(h|d)$

$$posterior = \frac{likelihood * prior}{evidence} \qquad p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

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

Learning and Probabilistic Inference (Prediction)

04

d … *data* $\mathcal{H}$ … $\{H_1, H_2, ..., H_n\}$ $\forall\, h, d$ …

h … *hypotheses*

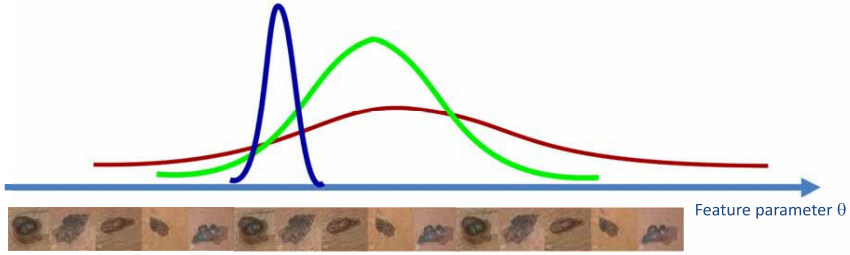
Likelihood

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



Bayesian Learning from data

05

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

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

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

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

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

Let the data do the work!

06

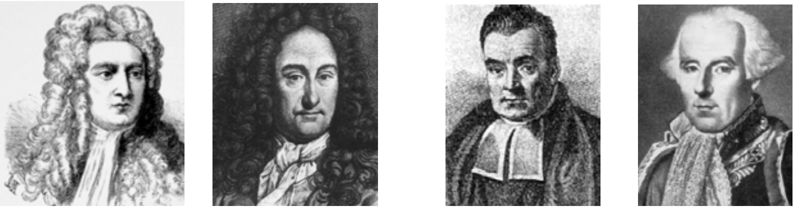
$$\max_{\mathbf{x} \in \mathcal{A} \subset \mathbb{R}^d} f(\mathbf{x})$$

$$p(h|d) \propto p(\mathcal{D}|\theta) * p(h)$$

$$p(f(x)|\mathcal{D}) \propto p(\mathcal{D}|f(x)) * p(f(x)$$

- Machine Learning is the development of algorithms which can **learn from data**
- assessment of **uncertainty,** making **predictions**
- **Automating automation - getting computers to program themselves – let the data do the work!**

Analogies



- **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**

Gaussian processes let integrating expert knowledge

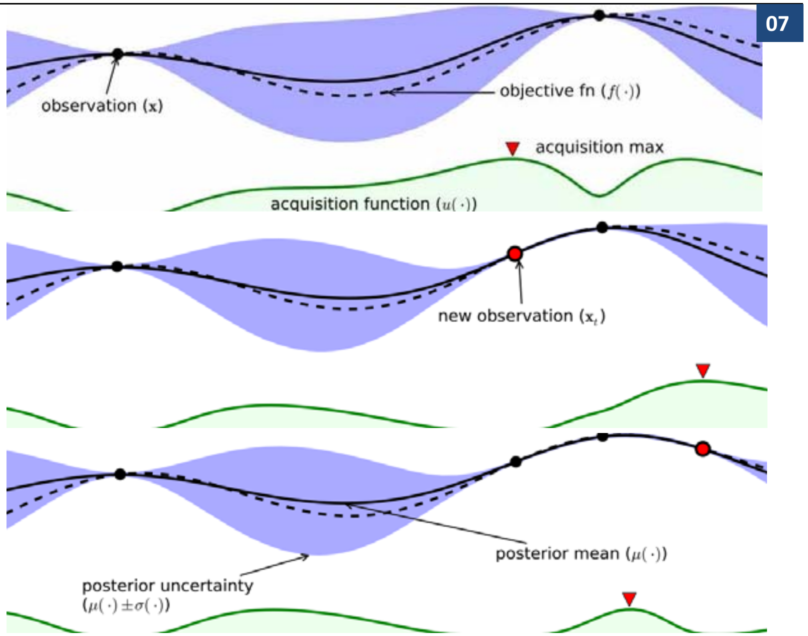
This image is in the public domain

04 Gaussian Processes

From Bayesian Optimization to Gaussian Process (GP) approximation

07

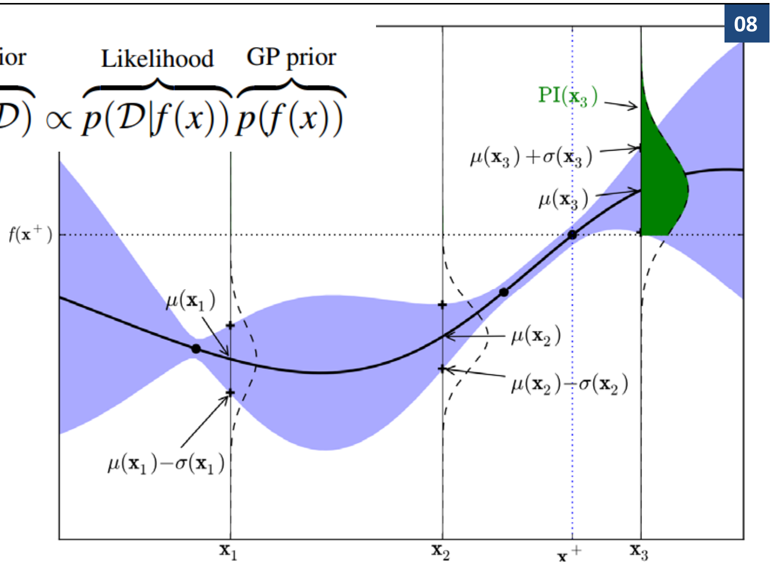
Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning arXiv:1012.2599.



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

08

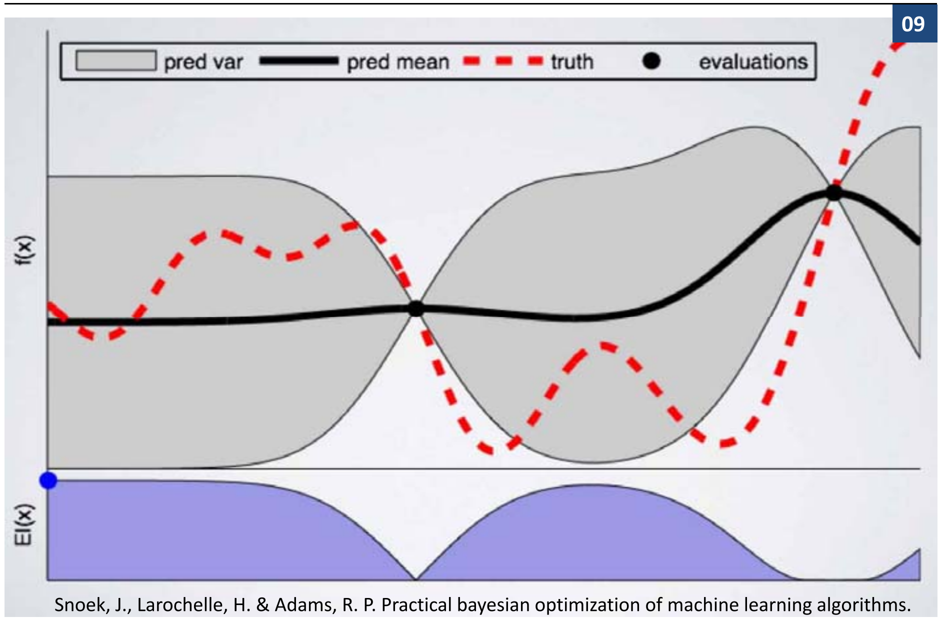
$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}}\overbrace{p(f(x))}^{\text{GP prior}}$$



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.

Demo on how Bayesian Optimization works ...

09



Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2012. 2951-2959.

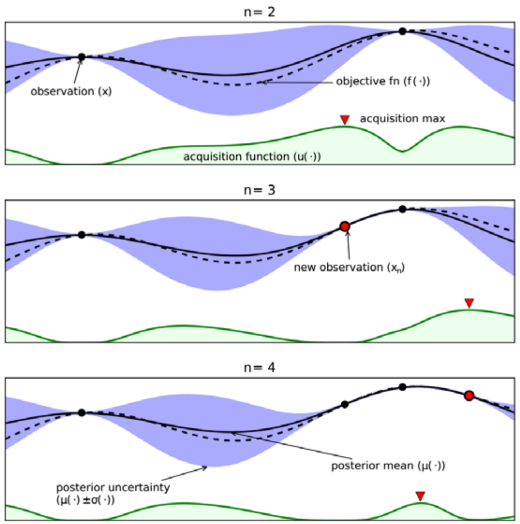
Why is this relevant for medicine?

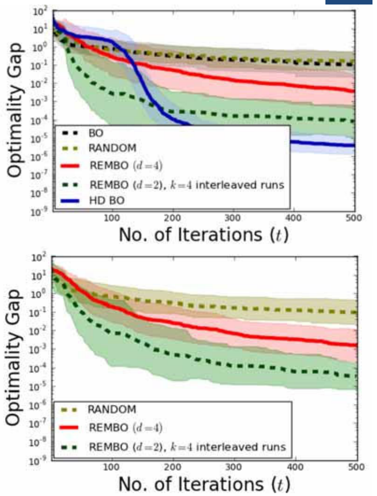
Reasoning under uncertainty: Decision Making

- 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

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

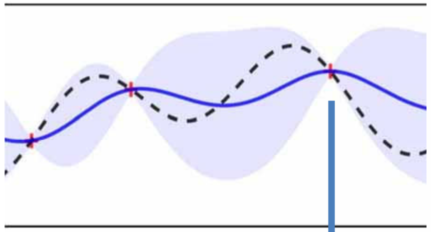
Scaling to high-dimensions is the holy grail in ML





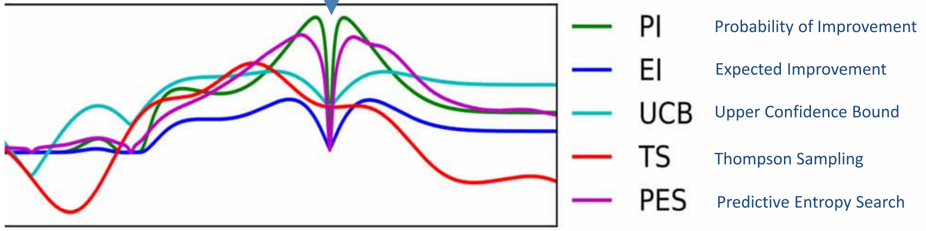
Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research, 55, 361-387, doi:10.1613/jair.4806.

Fully automatic → Goal: Taking the human out of the loop



Algorithm 1 Bayesian optimization

```
1: for $n = 1, 2, \ldots$ do
2:     select new $\mathbf{x}_{n+1}$ by optimizing acquisition function $\alpha$
           $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}; \mathcal{D}_n)$
3:     query objective function to obtain $y_{n+1}$
4:     augment data $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (\mathbf{x}_{n+1}, y_{n+1})\}$
5:     update statistical model
6: end for
```



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.

What do we learn from this …

… big data is good for automatic Machine Learning

and the grand goal of aML is …

05 automatic (autonomous) Machine Learning aML

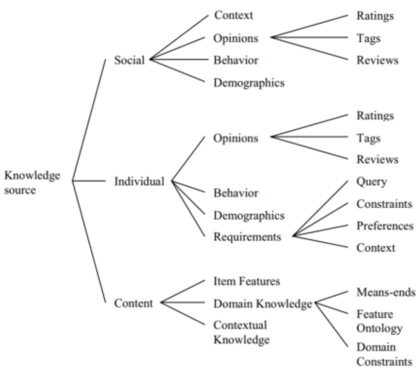
Everything is machine learning …

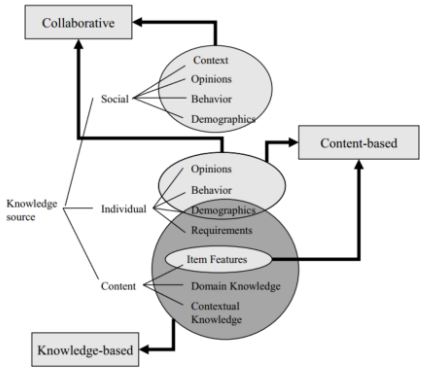
- Today most ML-applications are using automatic Machine Learning (aML) approaches
- automatic Machine Learning (aML) := algorithms which interact with agents and can optimize their learning behaviour trough this interaction

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

Best practice examples of aML

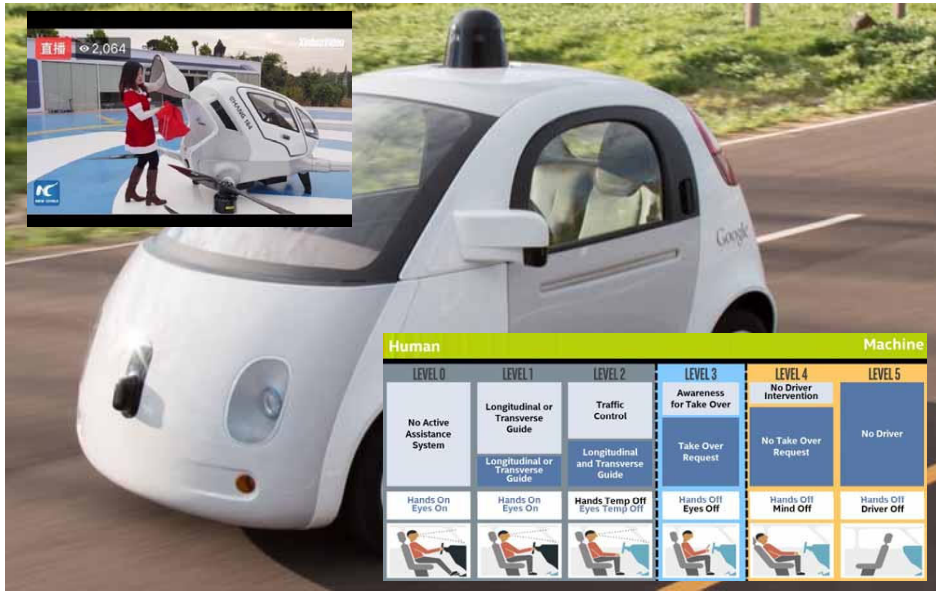
Example for aML: Recommender Systems





Francesco Ricci, Lior Rokach & Bracha Shapira 2015. Recommender Systems: Introduction and Challenges. Recommender Systems Handbook. New York: Springer, pp. 1-34, doi:10.1007/978-1-4899-7637-6_1.

Fully automatic autonomous vehicles (“Google car”)



LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
No Active Assistance System	Longitudinal or Transverse Guide	Traffic Control	Awareness for Take Over	No Driver Intervention	No Driver
	Longitudinal or Transverse Guide	Longitudinal and Transverse Guide	Take Over Request	No Take Over Request	
Hands On Eyes On	Hands On Eyes On	Hands Temp Off Eyes Temp Off	Hands Off Eyes Off	Hands Off Mind Off	Hands Off Driver Off

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

… and thousands of industrial aML applications …

When the heater is on, the LEDs on the device light up and the surface is unsafe to touch. The surface can cause burns.

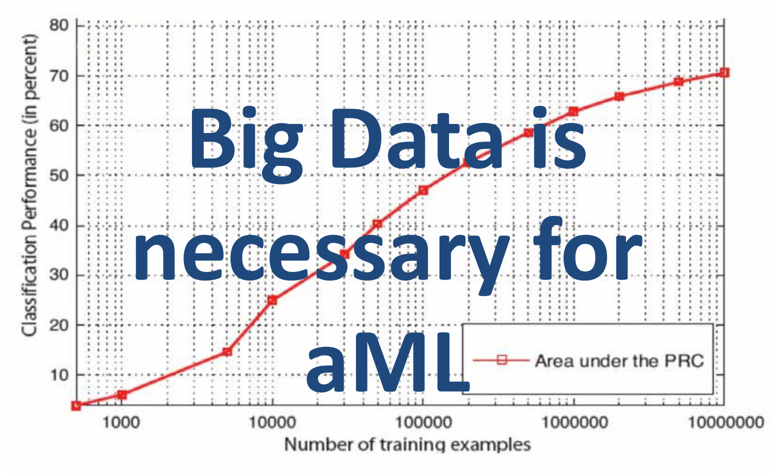
Wenn die Heizung eingeschaltet ist, leuchten die LEDs am Gerät auf und die Oberfläche ist nicht berührungssicher. Die Oberfläche kann Verbrennungen verursachen.

When the heater is on, the LEDs on the device light up and the surface is unsafe to touch. The surface can cause burns.

Wenn die Heizung eingeschaltet ist, leuchten die LEDs am Gerät auf und die Oberfläche ist berührungsempfindlich. Die Oberfläche kann Verbrennungen verursachen.

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.

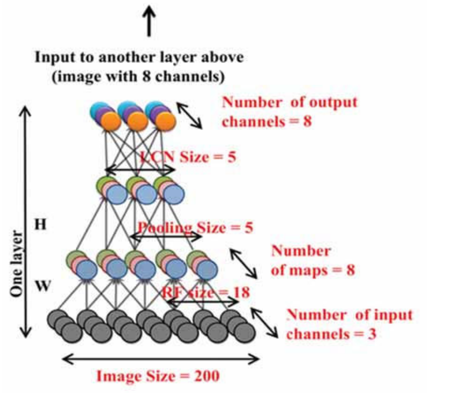
Big Data is necessary for aML !



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.

10 million 200 x 200 px images from the Web



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

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 aML fail …

- 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, …

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

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

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

Consequently …

Sometimes we (still) need a human-in-the-loop

06 interactive Machine Learning (iML) human-in-the-loop

Definition of iML (Holzinger – 2016)

A6

- 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.

Sometimes we need a doctor-in-the-loop

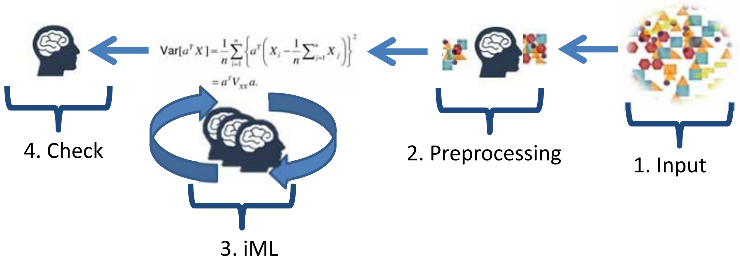
Image Source: 10 Ways Technology is Changing Healthcare http://newhealthypost.com Posted online on April 22, 2018

A group of experts-in-the-loop

A crowd of people-in-the-loop

iML: bringing the human-in-the-loop

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 …


$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2 = a^T V_{XX} a.$
4. Check
3. iML
2. Preprocessing
1. Input

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

Three examples for the usefulness of iML …

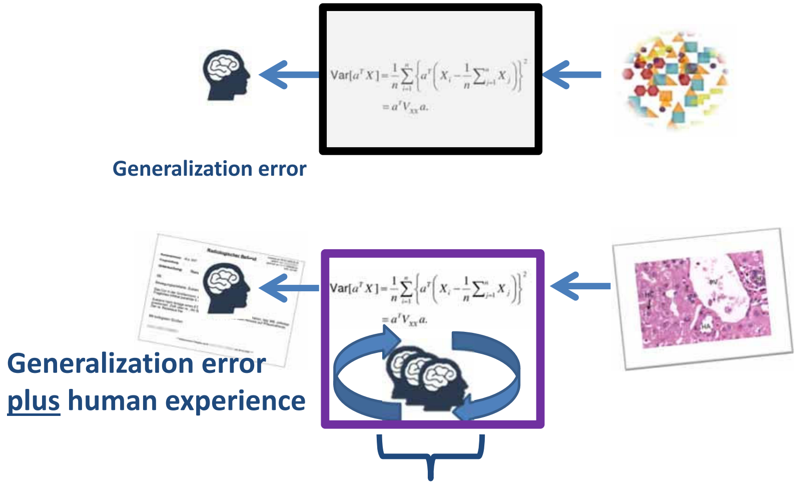
- **Example 1: Subspace Clustering**
- **Example 2: k-Anonymization**
- **Example 3: Protein Design**

Hund, M., Böhm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D. A., Majnaric, L. & Holzinger, A. 2016. Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. Brain Informatics, 1-15, doi:10.1007/s40708-016-0043-5.

Kieseberg, P., Malle, B., Fruehwirt, P., Weippl, E. & Holzinger, A. 2016. A tamper-proof audit and control system for the doctor in the loop. Brain Informatics, 3, (4), 269–279, doi:10.1007/s40708-016-0046-2.

Lee, S. & Holzinger, A. 2016. Knowledge Discovery from Complex High Dimensional Data. In: Michaelis, S., Piatkowski, N. & Stolpe, M. (eds.) Solving Large Scale Learning Tasks. Challenges and Algorithms, Lecture Notes in Artificial Intelligence LNAI 9580. Springer, pp. 148-167, doi:10.1007/978-3-319-41706-6_7.

Humans can deal with abstract concepts


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

Generalization error

Generalization error
plus human experience

iML = human inspection – bring in human intuition

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.

Why using human intuition?

Humans can generalize from few examples, and …

- understand relevant representations,
- find abstract concepts 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.

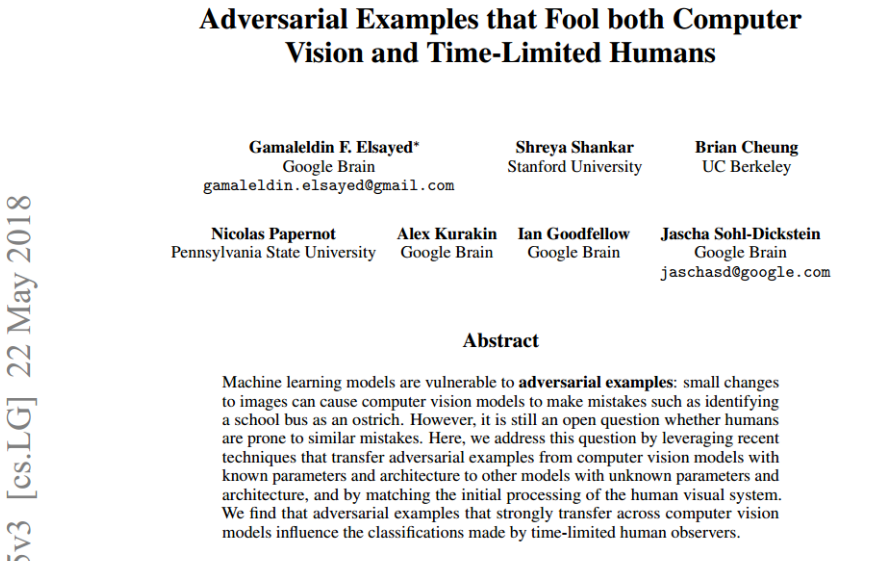
Human Intuition

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

This image is in the public domain, Source: freedesignfile.com

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

Adversarial > Deep Learning > Deep Fake



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

5v3 [cs.LG] 22 May 2018

Abstract

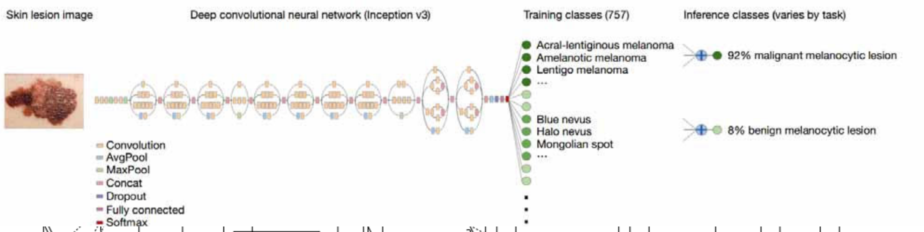
Machine learning models are vulnerable to **adversarial examples**: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question 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.

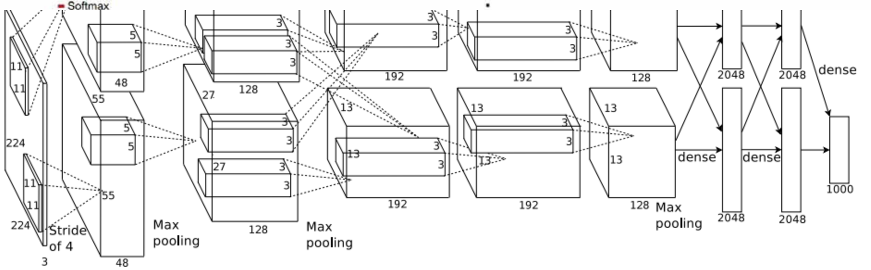
Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein 2018. Adversarial Examples that Fool both Human and Computer Vision. arXiv:1802.08195.

07 Why Explainability?

Deep Convolutional Neural Network Pipeline

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.

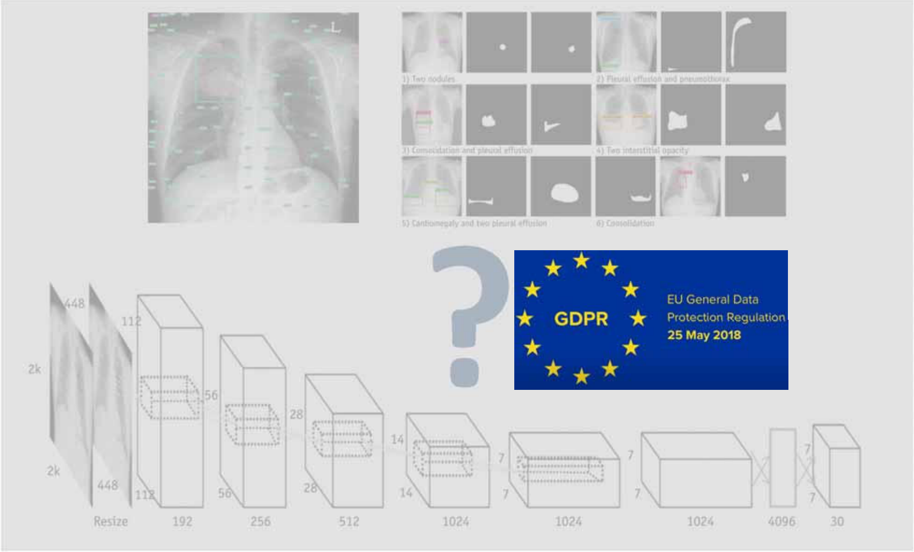




Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

Houston, we have a problem …

Source: NASA, image is in the public domain

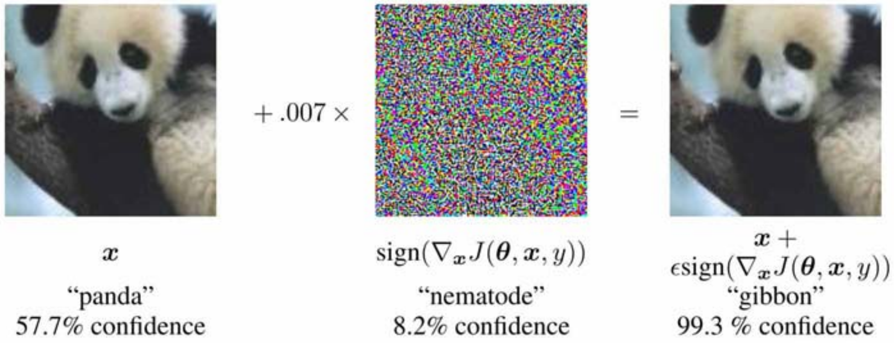


June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo & Namkug Kim 2017. Deep learning in medical imaging: general overview. Korean journal of radiology, 18, (4), 570-584, doi:10.3348/kjr.2017.18.4.570.

Deep Learning is not complicated, it's just a lot of it

- **Non-convex:** difficult to set up, to train, to optimize, needs a lot of expertise, error prone
- **Resource intensive** (GPU’s, cloud CPUs, federated learning, …)
- **Data intensive,** needs often millions of training samples …
- **Transparency lacking,** do not foster trust and acceptance among end-user, legal, ethical and social aspects make “black box” results difficult

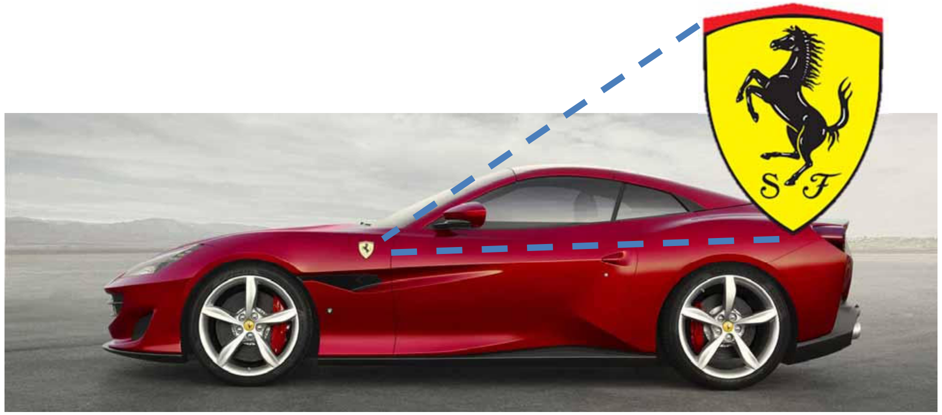
Example: Adversarial examples


$+ .007 \times$
$=$
$\boldsymbol{x}$
“panda”
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence

See also: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572, and see more examples: https://imgur.com/a/K4RWn

Example: Classifier Errors

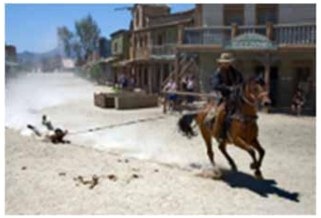
- Result of the classifier: **This is a horse**
- **Why is this a horse?**



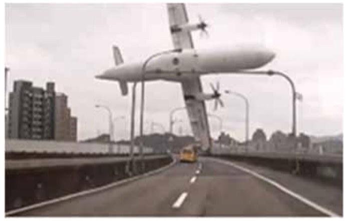
Source: Image is in the public domain

Example: (Automatic) Context understanding …

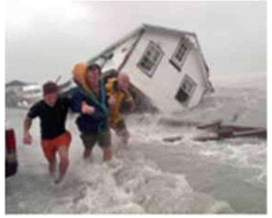
Image Captions by deep learning: State-of-the-Art of the Stanford Machine Learning Group



a woman riding a horse on a dirt road



an airplane is parked on the tarmac at an airport

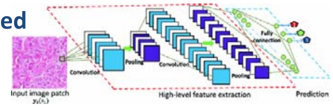


a group of people standing on top of a beach

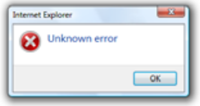
Andrej Karpathy, Justin Johnson & Li Fei-Fei 2015. Visualizing and understanding recurrent networks. arXiv:1506.02078.

Benefits …

Verify that algorithms/classifiers work as expected
Wrong decisions can be costly and dangerous …

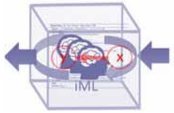


Understanding the weaknesses and errors
Detection of bias – bring in human intuition to know the error …



Unknown error

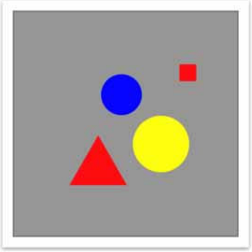
Scientific replicability and causality
The "why" is often more important than the prediction …



Andreas Holzinger 2018. Explainable AI (ex-AI). Informatik-Spektrum, 41, (2), 138-143, doi:10.1007/s00287-018-1102-5.

08 Exploration Environment for Explaninable AI: #KandinskyPatterns

Definition 1: A Kandinsky Figure is …



- … a square image containing 1 to n geometric objects.
- Each object is characterized by its shape, color, size and position within this square.
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable by a human observer.

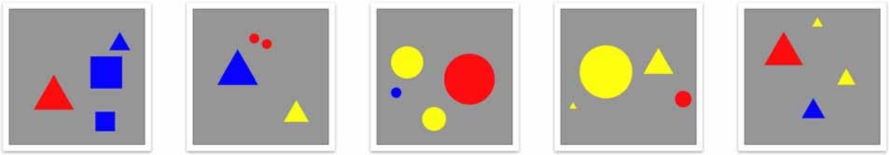
Definition 2 A statement $s(k)$

- about a Kandinsky Figure k is …
- either a mathematical function $s(k) \rightarrow B;$ with $B\ (0,1)$
- or a $natural\ language\ statement$ which is true or false

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

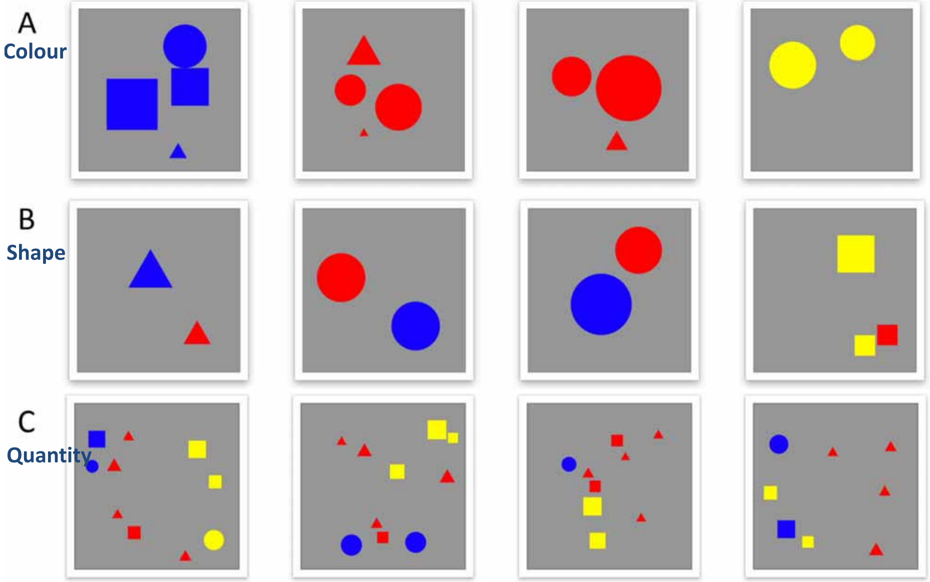
Definition 3 A Kandinsky Pattern K …

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

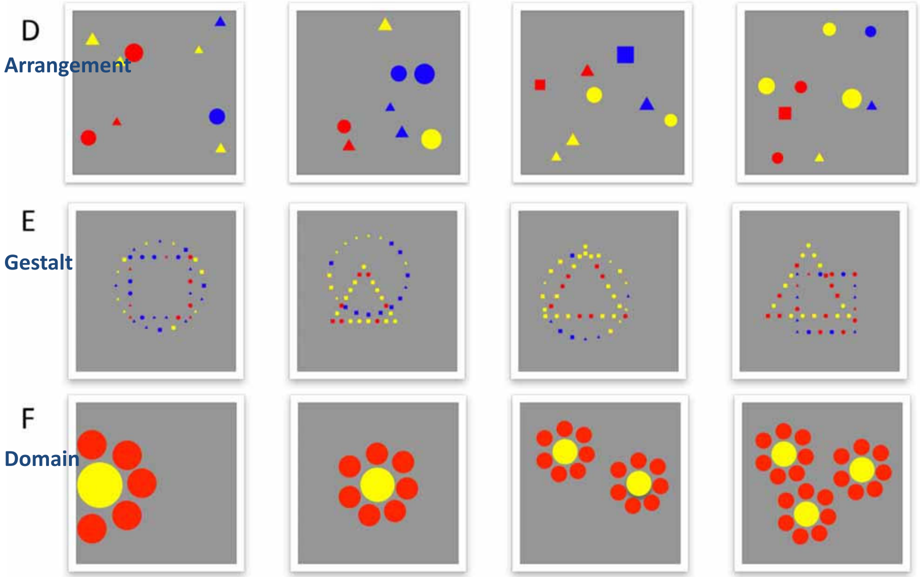


"… the Kandinsky Figure has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share a object …".

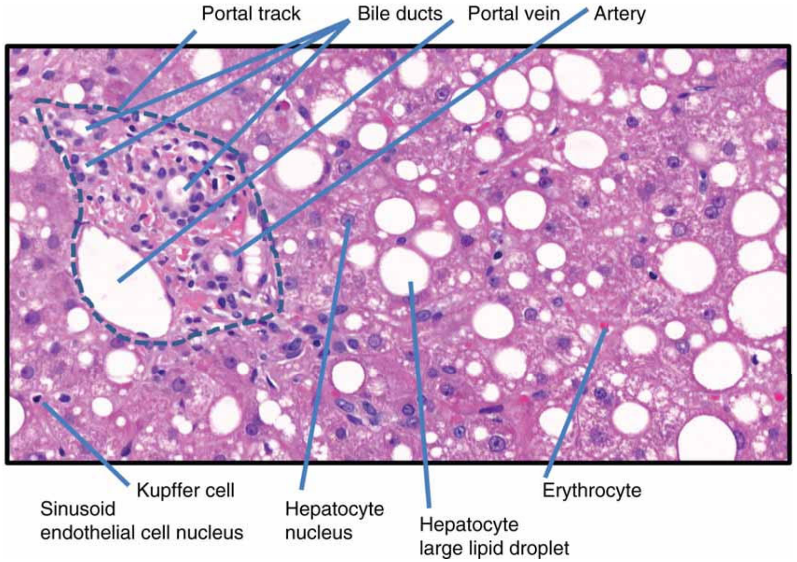
How do humans explain? They use concepts!



How do humans explain? (2/2)

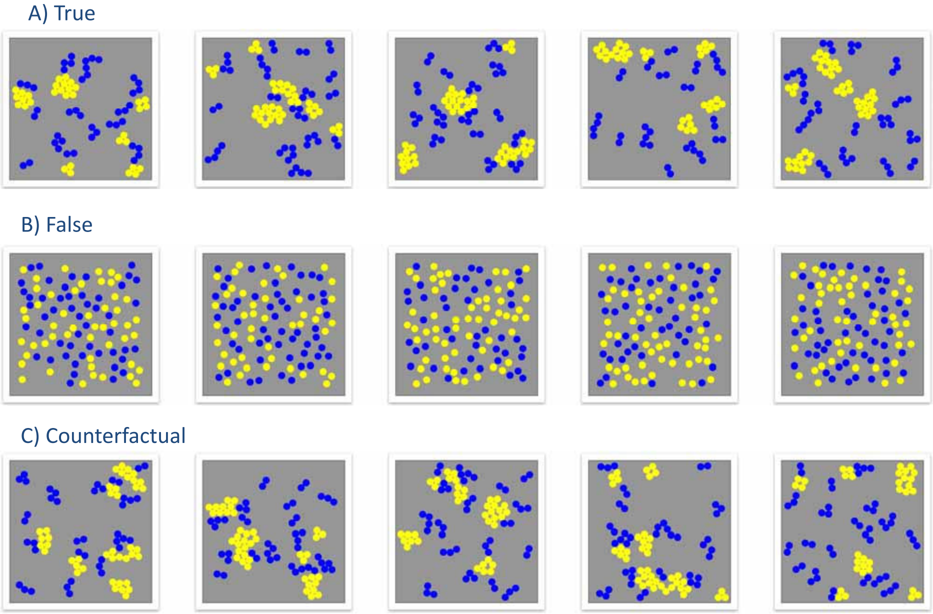


Motivation: Typical Task of a human Pathologist …



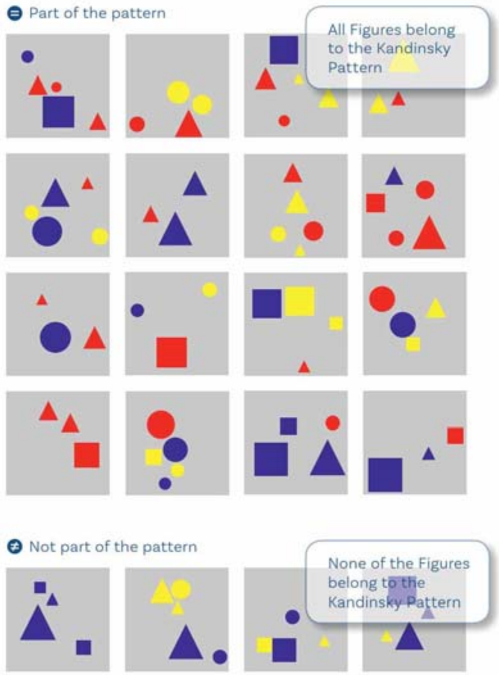
Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

Domain Concept from Pathology


A) True

B) False

C) Counterfactual


Part of the pattern

All Figures belong to the Kandinsky Pattern

Not part of the pattern

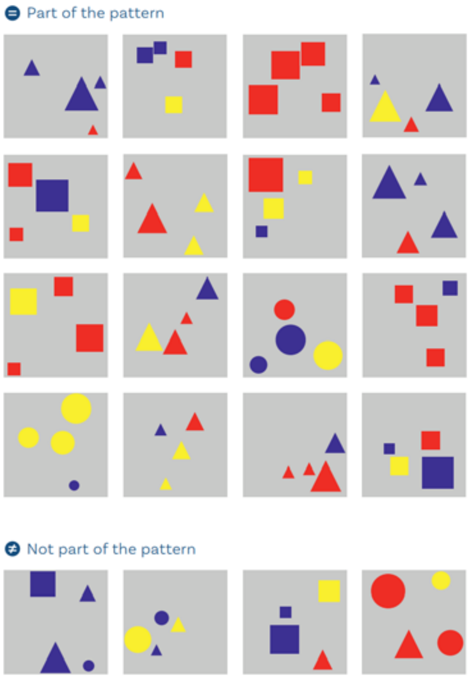
None of the Figures belong to the Kandinsky Pattern

Hypothesis 1

It only contains circles and triangles.

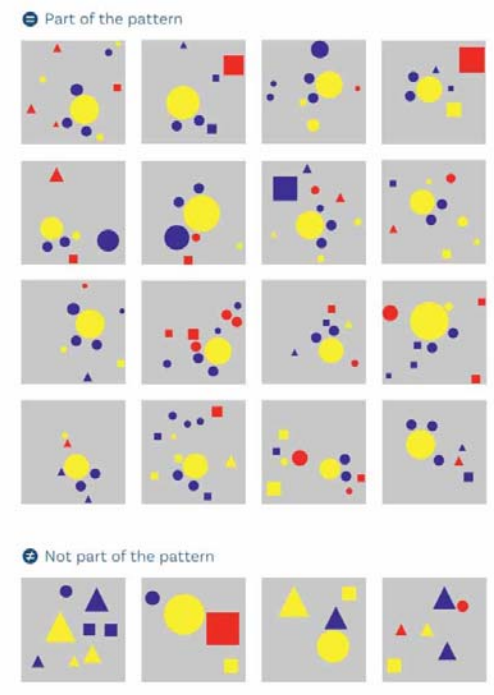
Hypothesis 2

It contains at least a red object. ✓


Part of the pattern

Not part of the pattern

S2 Basic Pattern 2
Title: **All of Same Shape ->**
All objects have the same shape.
Hint: Don't be distracted by the colors

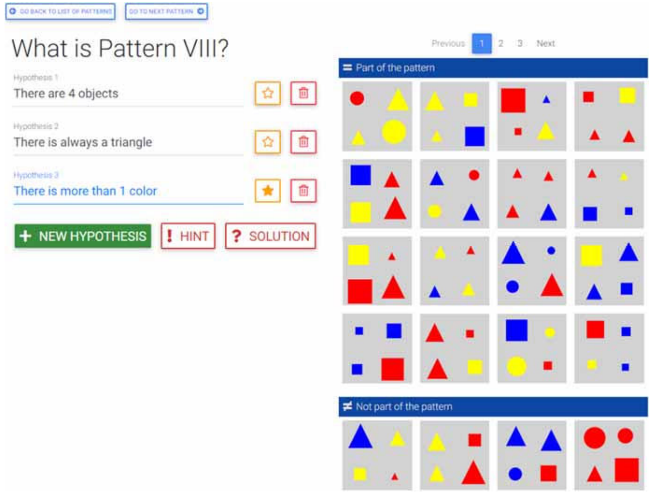


S8 Basic Pattern 8

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.

Intelligence Test for Machines



What is Pattern VIII?

Hypothesis 1: There are 4 objects

Hypothesis 2: There is always a triangle

Hypothesis 3: There is more than 1 color

+ NEW HYPOTHESIS | ! HINT | ? SOLUTION

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.

Visit the #KANDINSKYpatterns homepage:

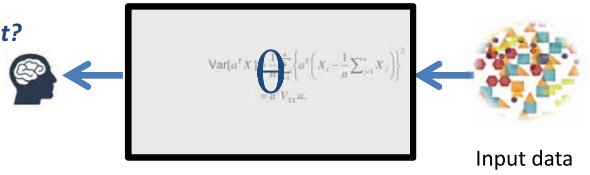
https://human-centered.ai/project/kandinsky-patterns

#KANDINSKYpatterns @aholzin

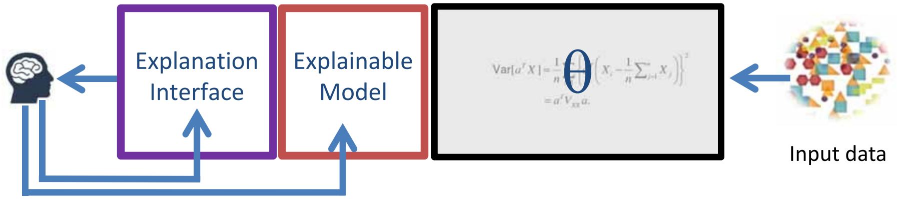
Conclusion

We need effective Human-AI mapping

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



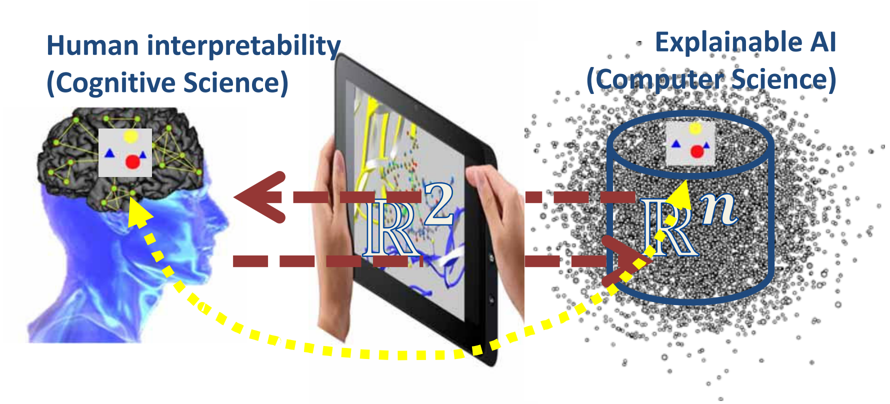
We contribute to …



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

Explainable AI needs effective concept mapping !!!

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



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.