Mini Course

Fundamentals of Medical AI

Part 01

Introduction to Medical AI and Machine Learning for Health Informatics

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)

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

and

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

Course Overview

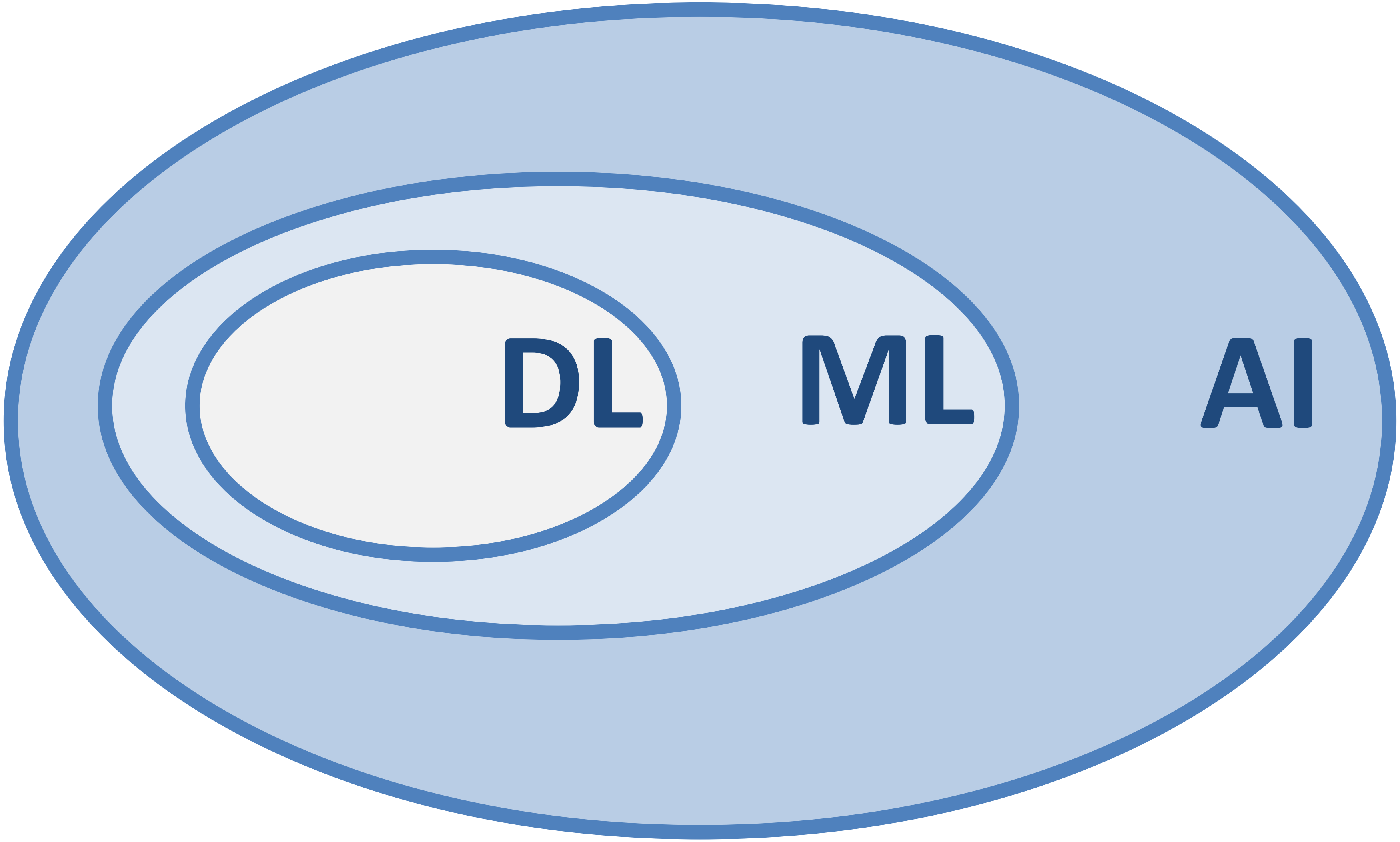
Primer on Probability & Information

Part 1 Theory	Part 2 Practice
01 Introduction to Medical AI and Machine Learning for Health	05 Methods of Explainable AI
02 Data, Information and Knowledge	06 Social, Ethical and Legal Aspects of Medical AI
03 Human Decision Making and AI Decision Support	07 Project: Bringing AI into medical workflows
04 Causal Reasoning and Interpretable AI	08 Presentation of the developed concepts

Written Exam

- 01 Success Stories: Human-Level performance in Medical AI
- 02 Why is biomedicine a complex application area ?
- 03 Probabilistic Learning
- 04 automatic Machine Learning aML
- 05 interactive Machine Learning iML
- Conclusion

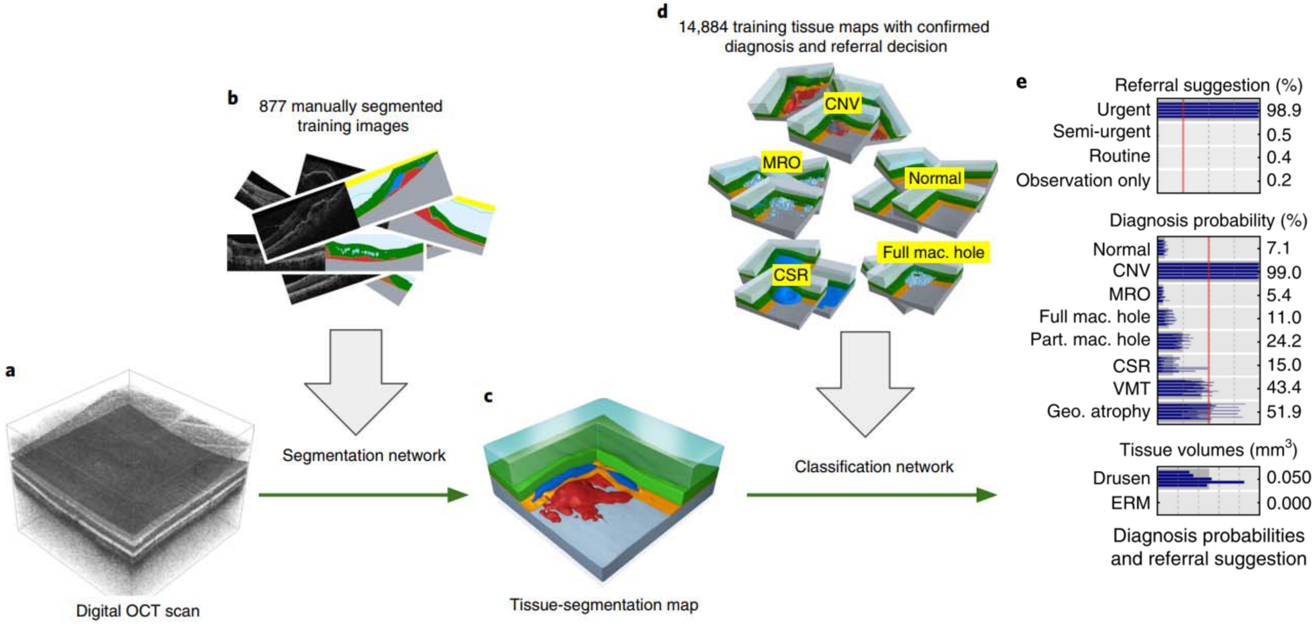
What is AI, ML, DL ?

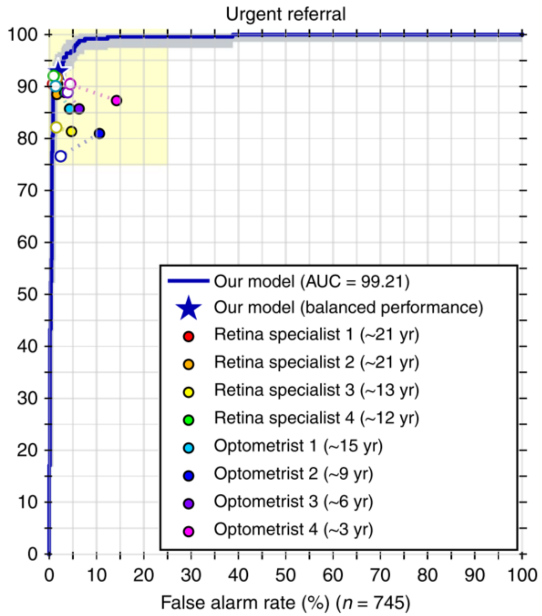


Andreas Holzinger, Peter Kieseberg, Edgar Weippl & A Min Tjoa 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. *Lecture Notes in Computer Science LNCS 11015*. Cham: Springer, pp. 1-8, doi:10.1007/978-3-319-99740-7-1.

01 Success Stories: Human-Level performance in Medical AI

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



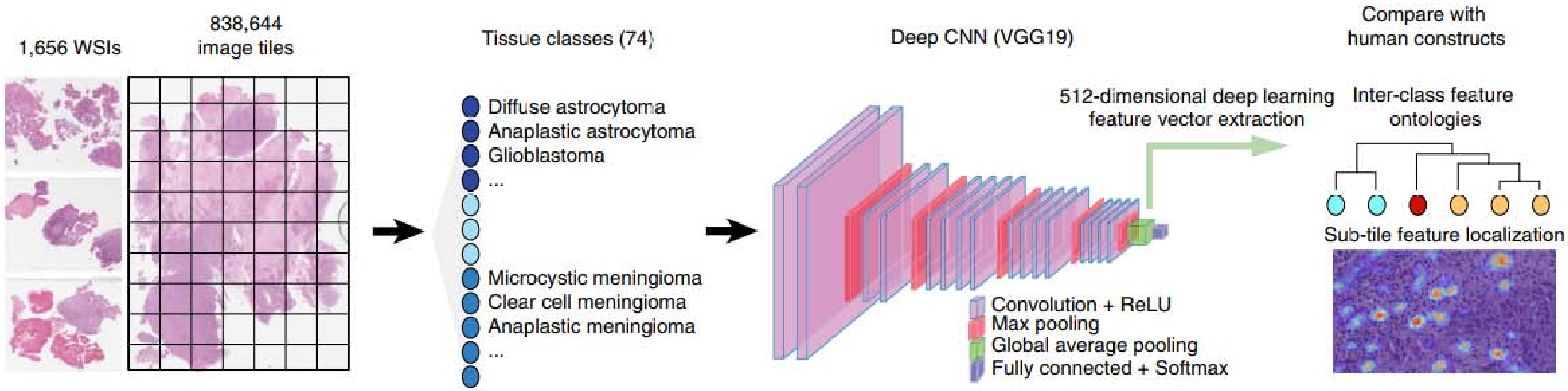


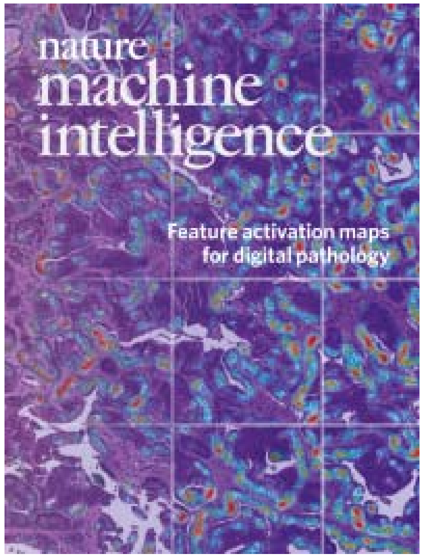
	Urgent	Semi-urgent	Routine	Observation
Urgent	234	5	13	0
Semi-urgent	3	225	2	0
Routine	10	2	250	4
Observation	1	1	14	233

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

NATURE MACHINE INTELLIGENCE ARTICLES



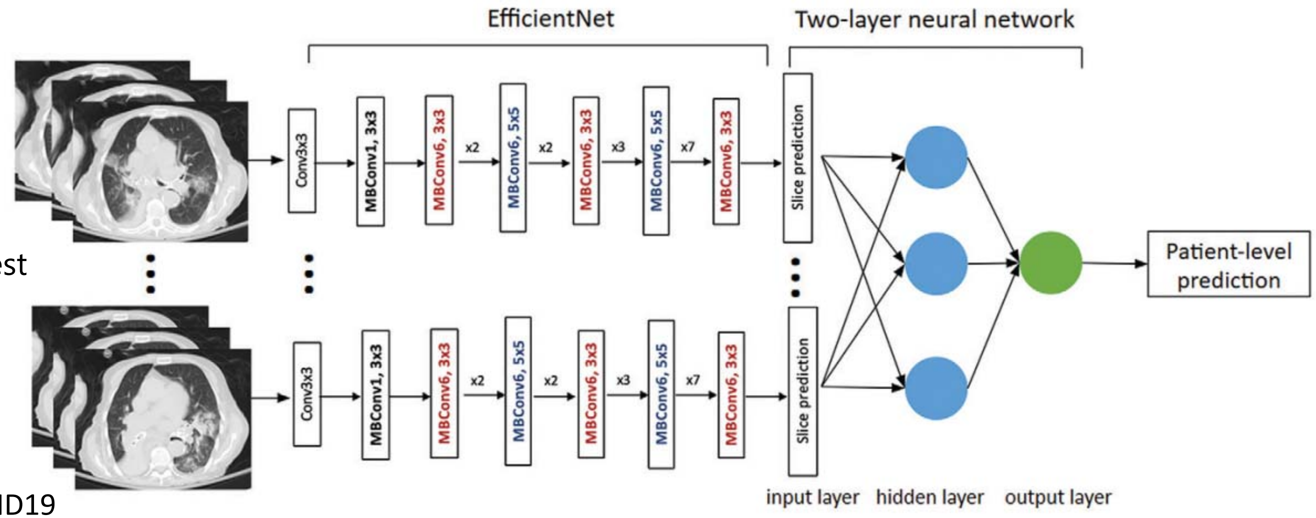


Kevin Faust, Sudarshan Bala, Randy Van Ommeren, Alessia Portante, Raniah Al Qawahmed, Ugljesa Djuric & Phedias Diamandis (2019). Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. *Nature Machine Intelligence,* 1, (7), 316-321, doi:10.1038/s42256-019-0068-6.

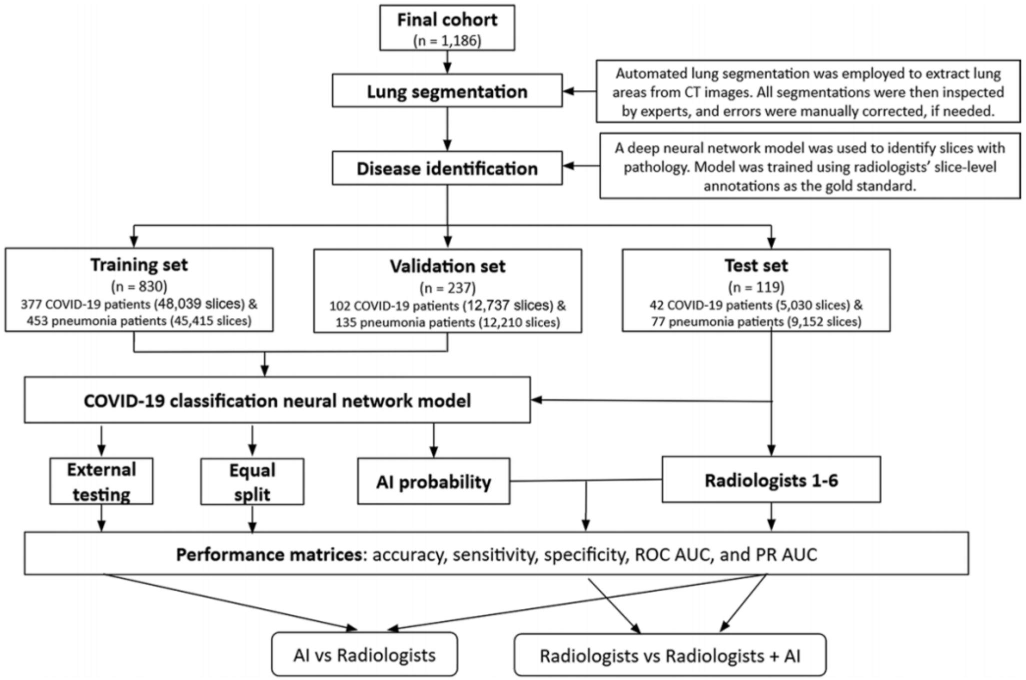
13. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556 (2014).
14. Holzinger, A. et al. Causability and explainabilty of artificial intelligence in medicine. *WIRES Data Min. Knowl. Discov.* **9**, e1312 (2019).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

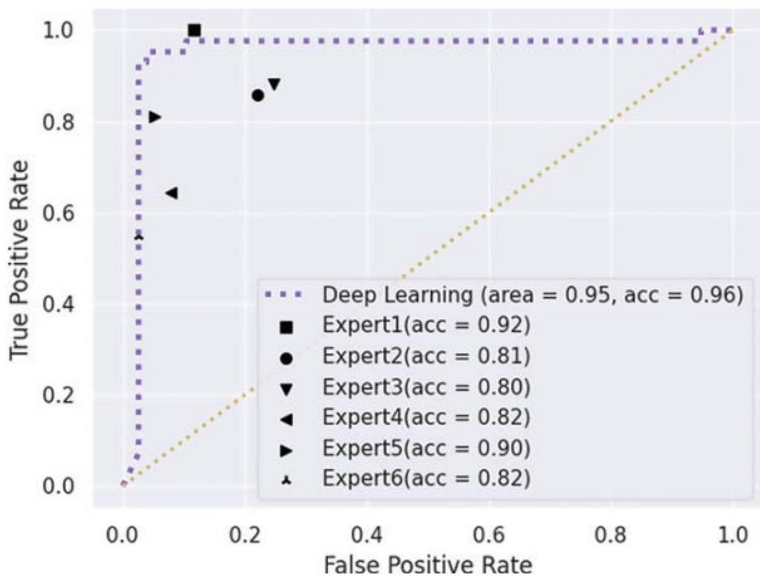
Example: COVID-19 Chest-CT (data augmentation)

Harrison X. Bai, Robin Wang, Zeng Xiong, Ben Hsieh, Ken Chang, Kasey Halsey, Thi My Linh Tran, Ji Whae Choi, Dong-Cui Wang & Lin-Bo Shi (2020). AI augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT. *Radiology,* 296, (3), 156-165, doi:10.1148/radiol.2020201491.



https://github.com/robinwang08/COVID19



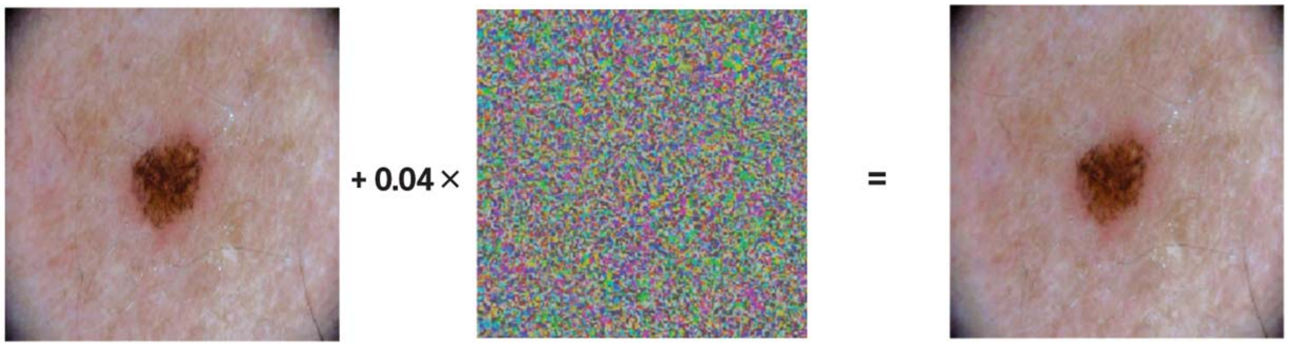


<u>Why</u> can AI solve some tasks better than humans?

<u>Why</u> does AI achieve such results ?

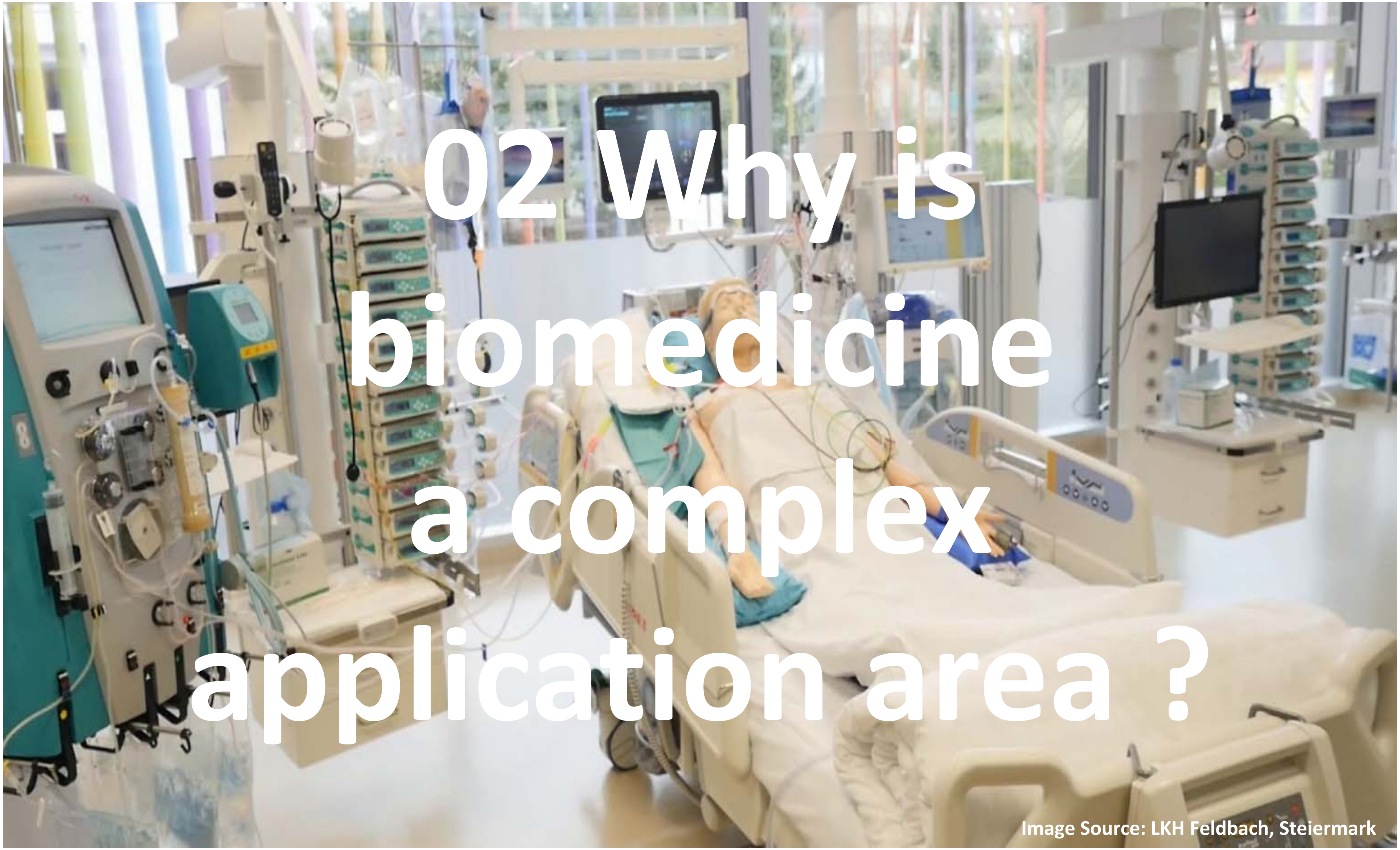
<u>What if</u> the input data changes counterfactually ?

Robustness & Interpretability



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.

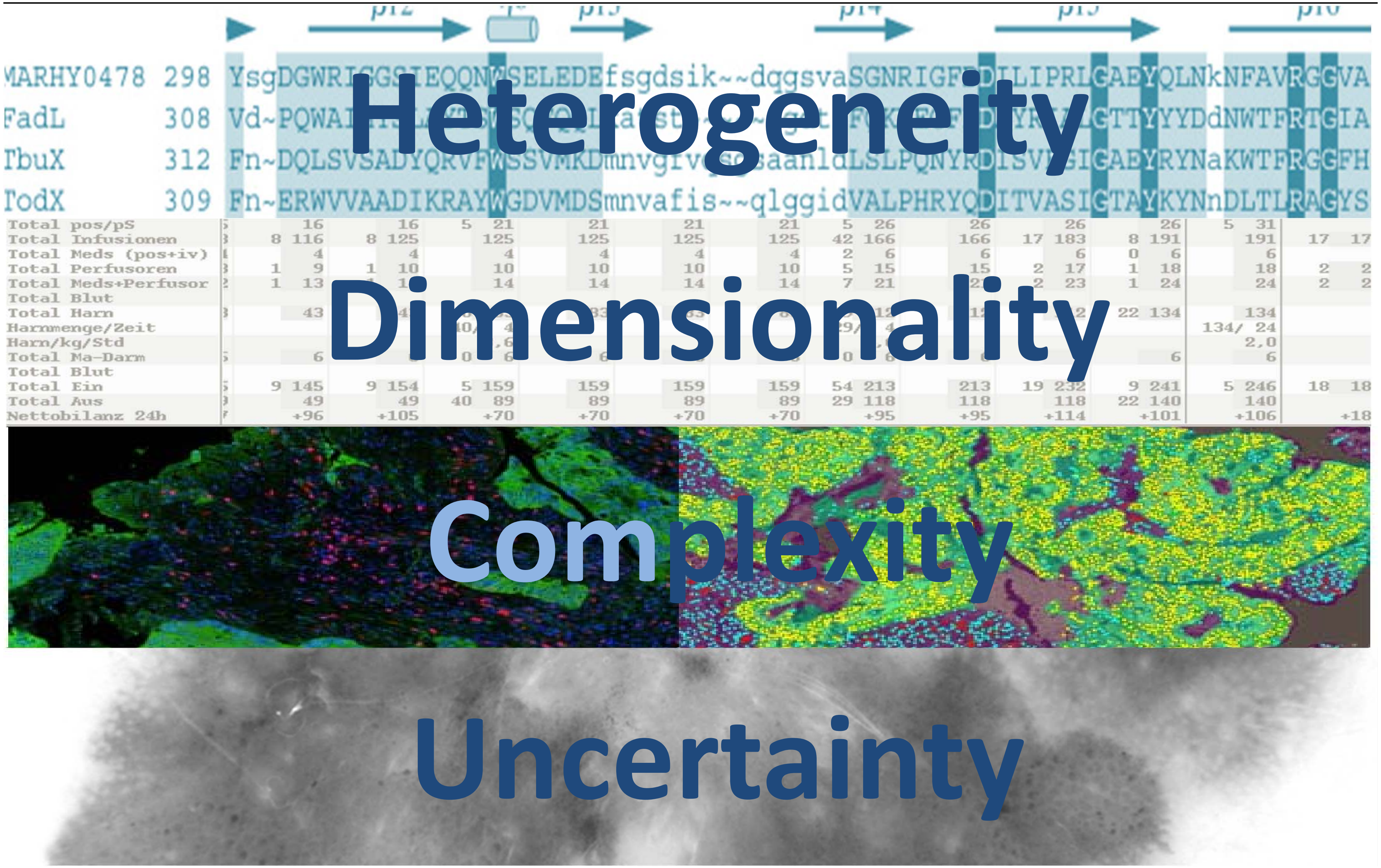
- 1) learning from few **data**
- 2) extracting **knowledge**
- 3) **generalize**
- 4) fight the curse of **dimensionality**
- 5) disentangle the **underlying explanatory factors of data**, i.e.
- 6) **causal understanding** of the data in the **context** of an application domain


02 Why is biomedicine a complex application area ?
Image Source: LKH Feldbach, Steiermark

Our central hypothesis:
Information may bridge this gap

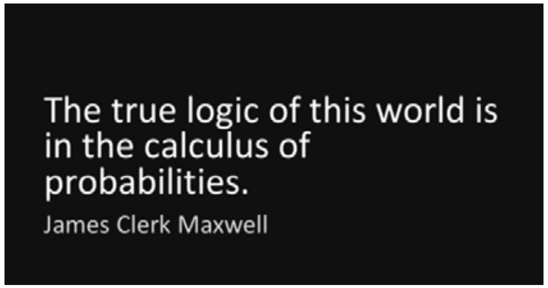
Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*

Where is the problem in building this bridge ?



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

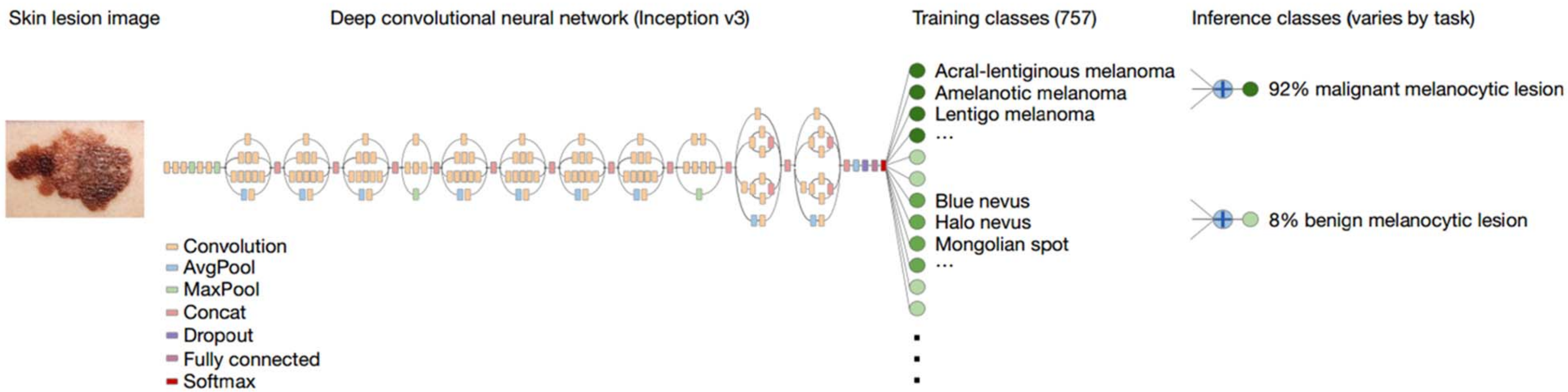


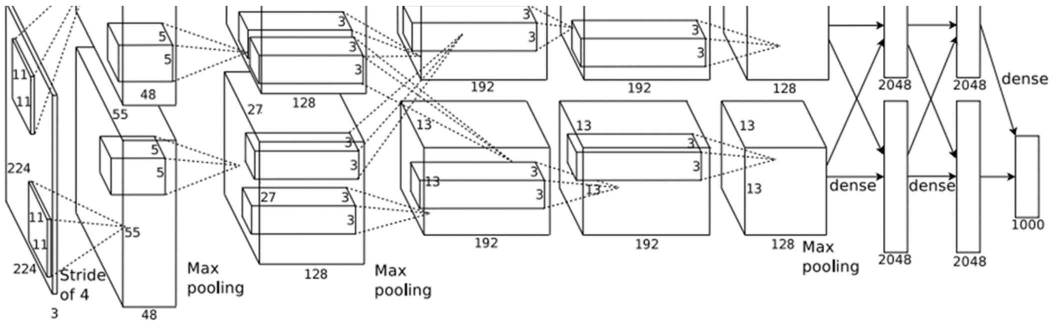
Maxwell, J. C. (1850). Letter to Lewis Campbell; reproduced in L. Campbell and W. Garrett, The Life of James Clerk Maxwell, Macmillan, 1881.

- 1957+ Medical Informatics - "AI Hype"
- 1985+ Health Telematics - "AI winter"
- 1995+ Web Era - AI is "forgotten"
- 2005+ Practical success of probabilistic learning due to increasing "big data" - "AI renaissance"
- 2010+ Data Era – Big Data = "super for AI"
- 2020+ Beyond Explainable AI – Causability "explainable medicine"

An algorithm is learning from **experience *E***
on **task *T*** with **performance measure *P*,**
if its performance on task *T* as measured by *P*,
improves with experience *E*

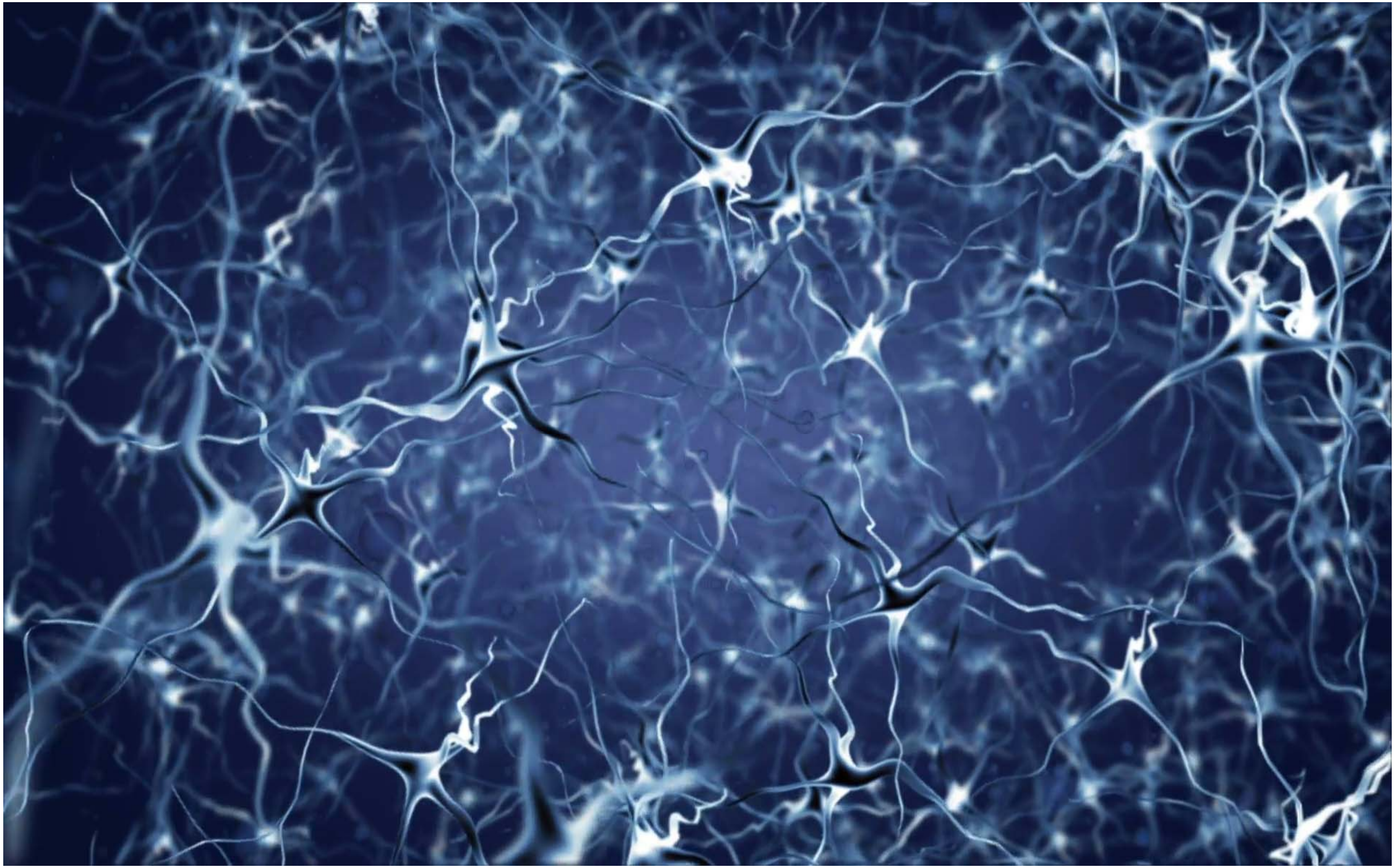
Tom M. Mitchell (1997). *Machine learning,* New York, McGraw Hill.



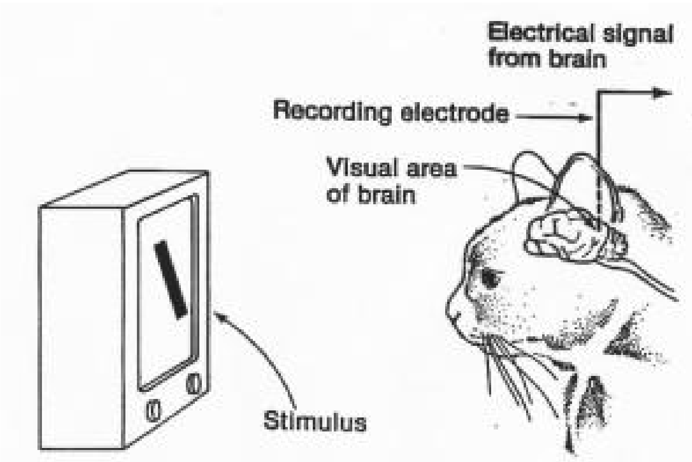


Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

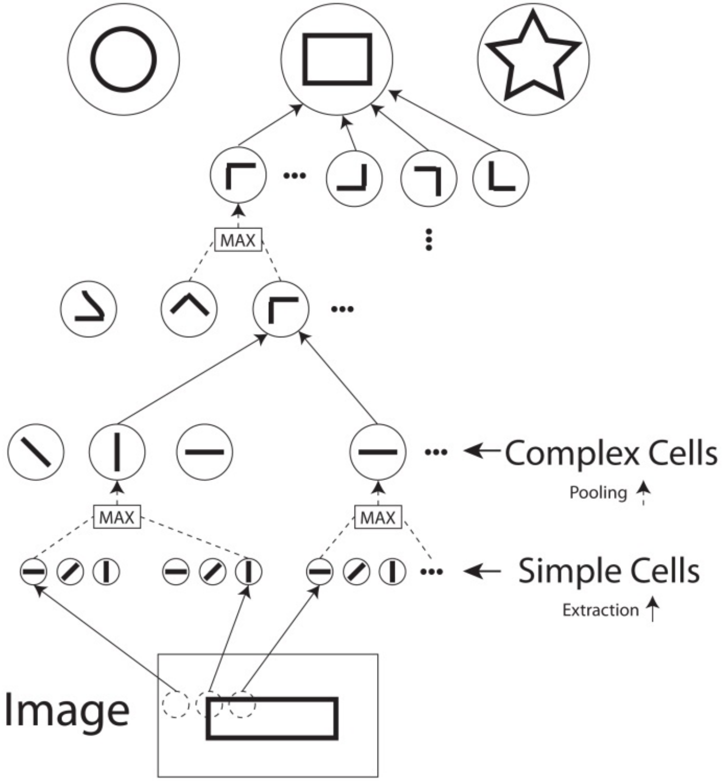
What was the inspiration for this kind of learning ?

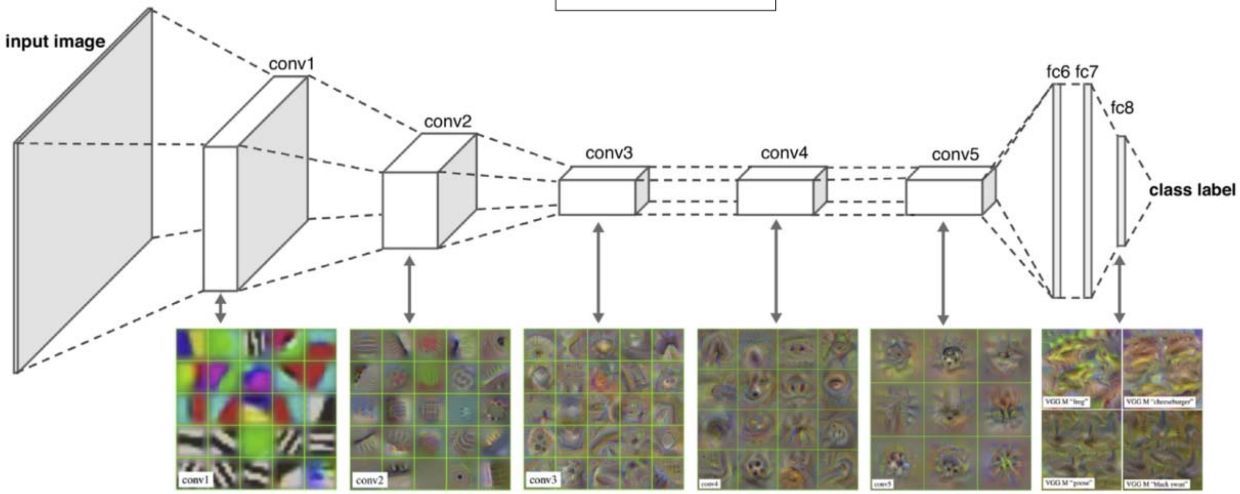


Why does ML with digital computrs work at all ?

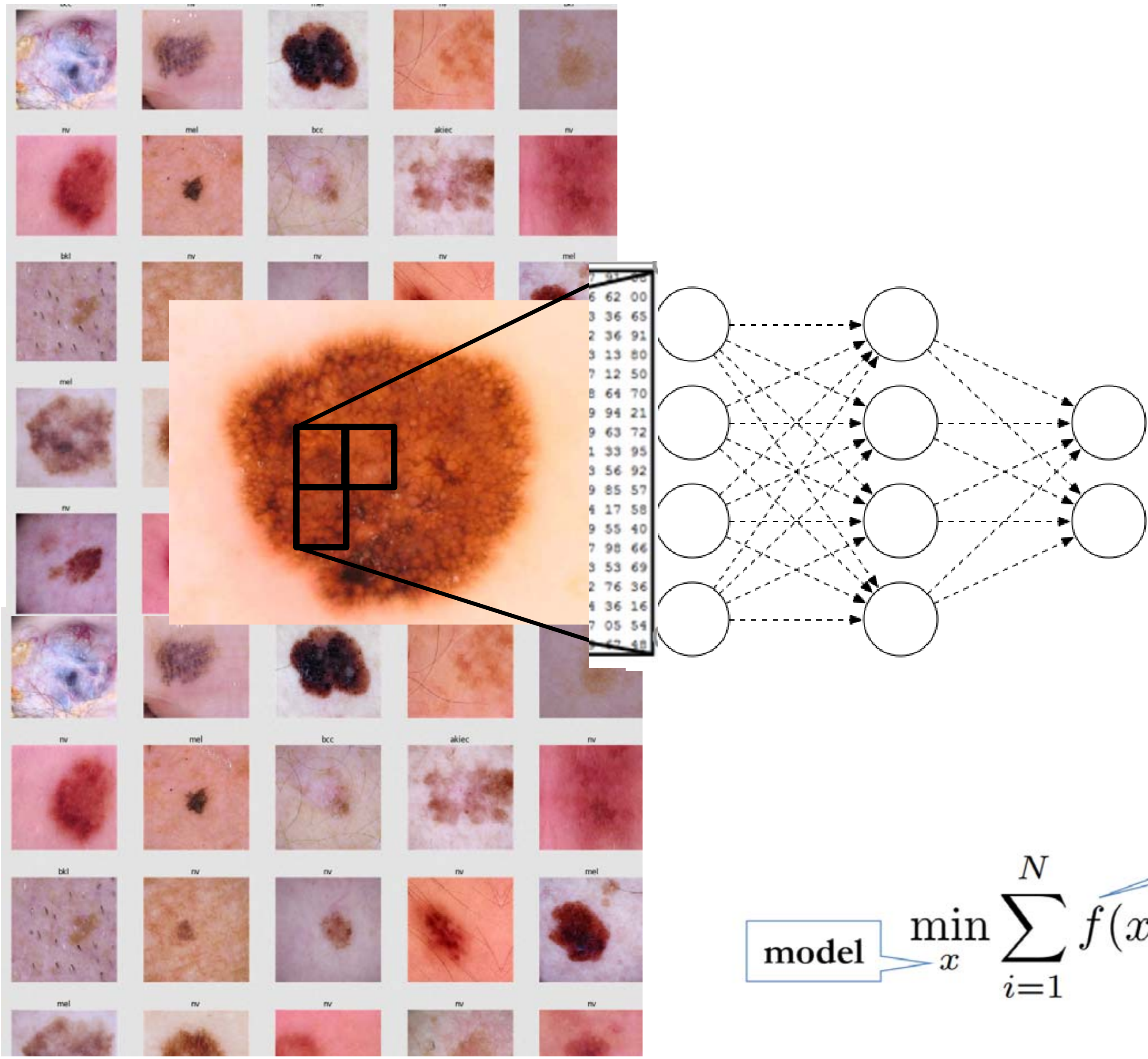


David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837





What is “deep learning” technically?



92 % malign

8 % benign

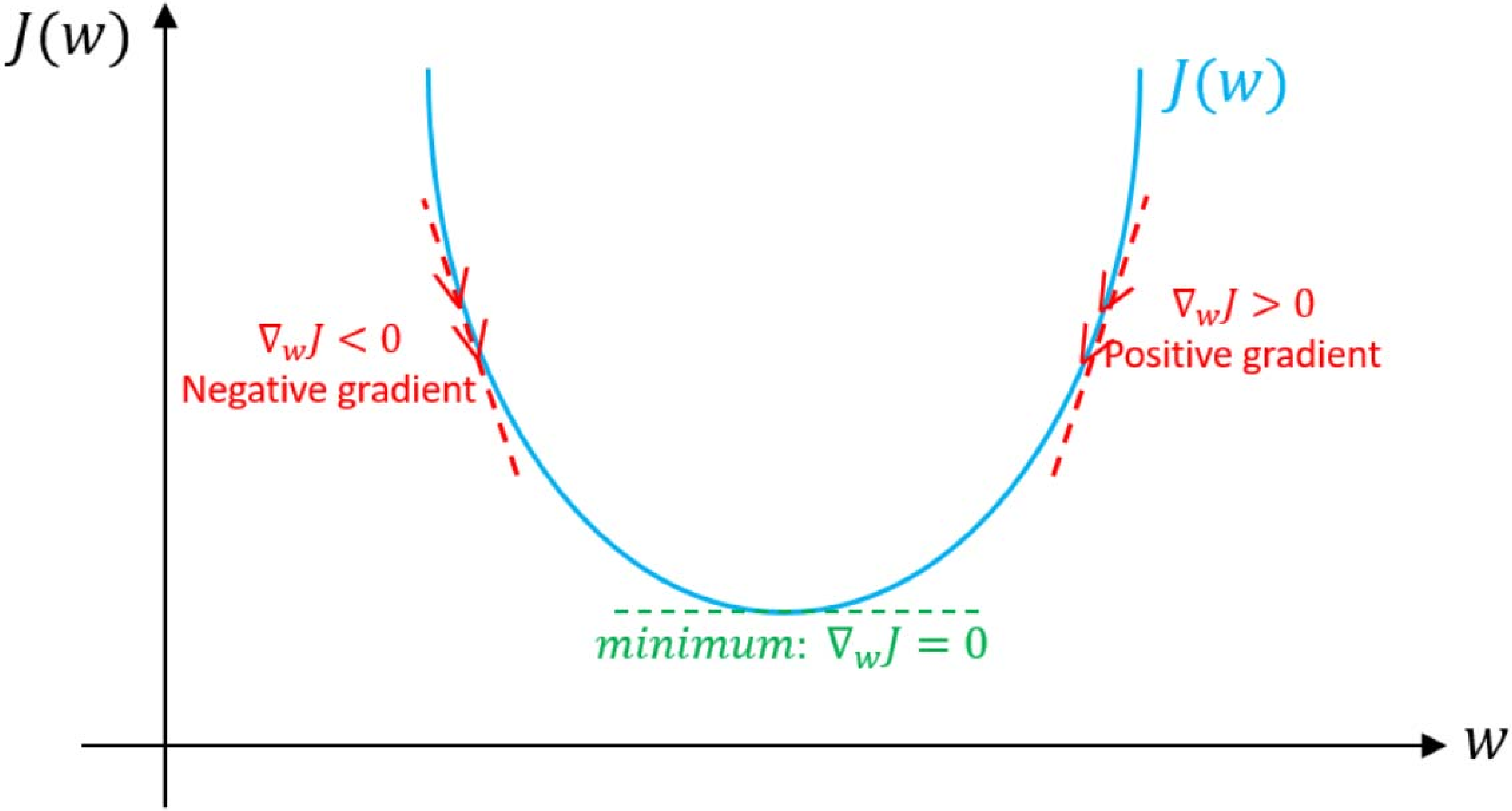
loss function

training examples

model

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

How do you find the min from a convex function ?



How does stochastic gradient descent work ?

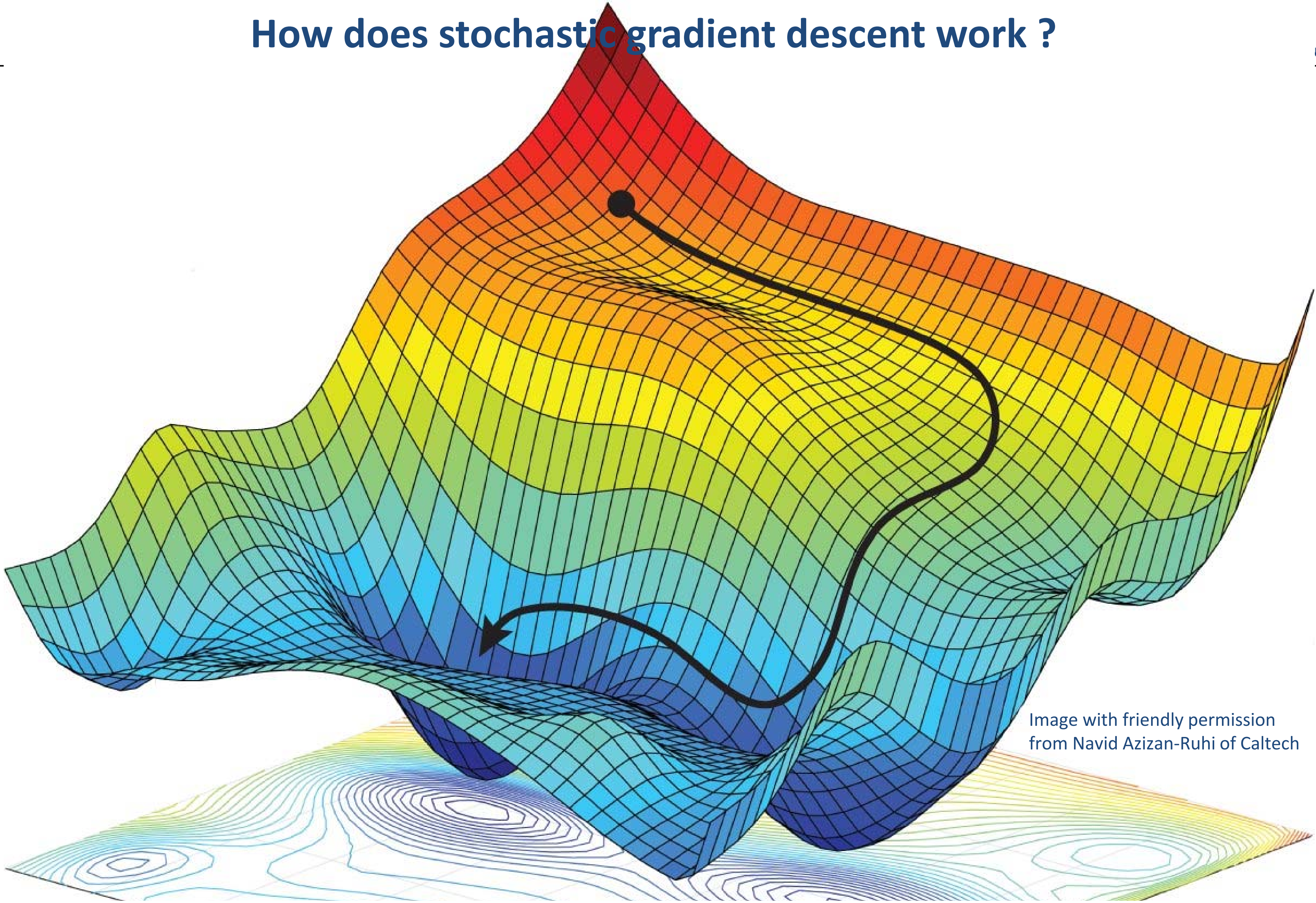


Image with friendly permission from Navid Azizan-Ruhi of Caltech

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

Algorithm 8.3: Stochastic gradient descent

Initialize $\boldsymbol{\theta}$, η;
repeat
  Randomly permute data;
  for $i = 1 : N$ **do**
    $\mathbf{g} = \nabla f(\boldsymbol{\theta}, \mathbf{z}_i)$;
    $\boldsymbol{\theta} \leftarrow \mathrm{proj}_{\Theta}(\boldsymbol{\theta} - \eta \mathbf{g})$;
    Update η;
until *converged*;

Kevin P. Murphy (2012). Machine learning: a probabilistic perspective, Cambridge (MA), MIT press, chapter 8.5.2.3, page 264

Let’s go back to the early roots of this success

- **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

Pierre Simon de Laplace (1749-1827)

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

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:

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

How to learn representations θ, h from observed data?

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

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

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

How to make probabilistic Inferences (Predictions) ?

04

d … *data*

h … *hypotheses*

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

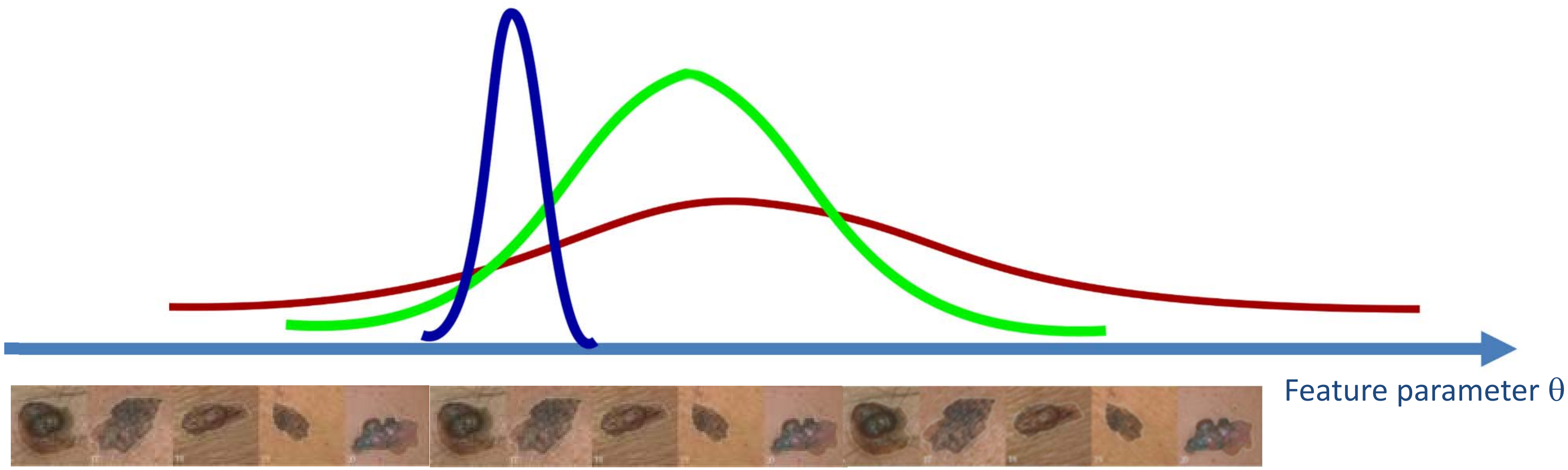
Likelihood

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

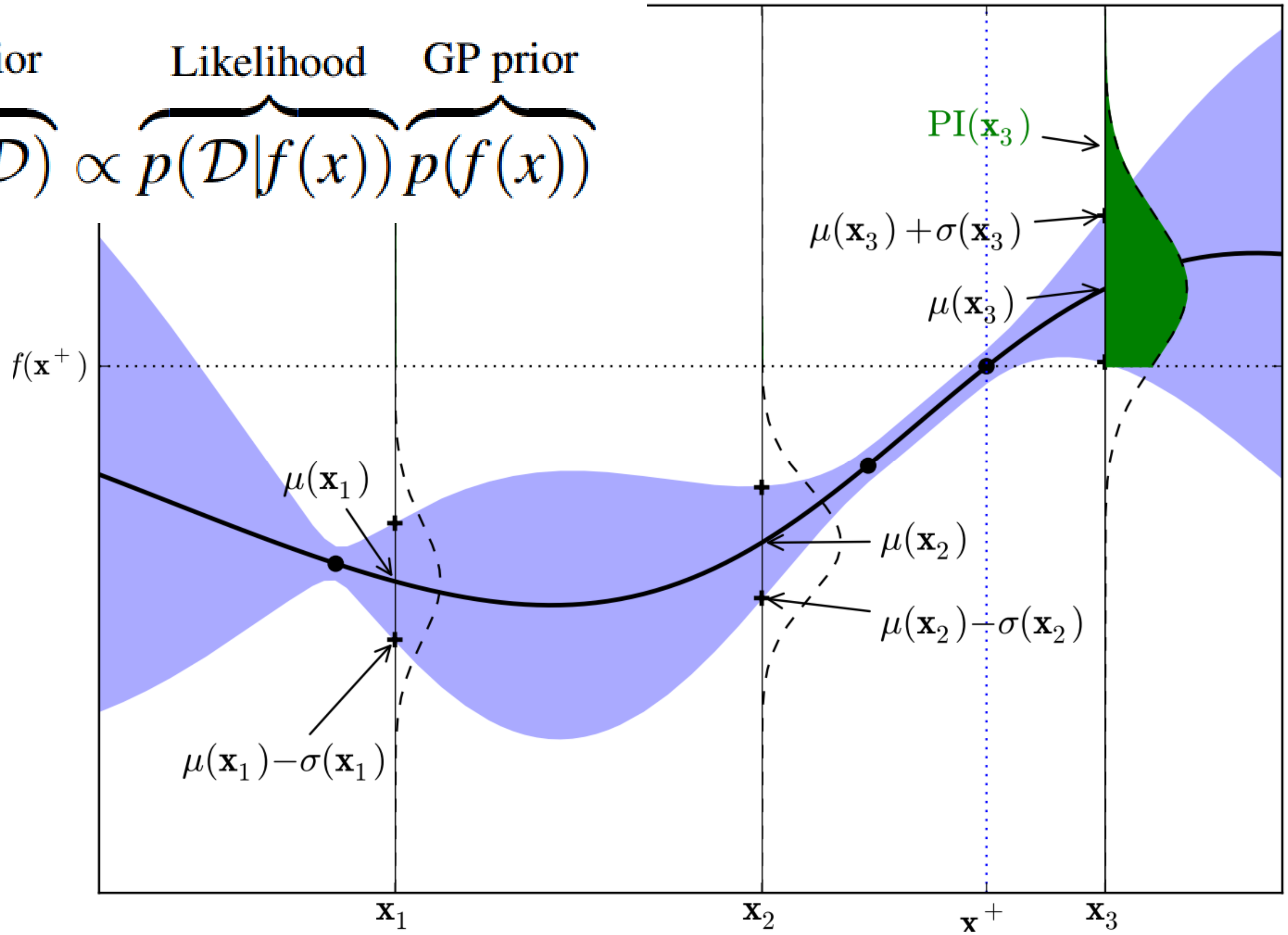


This was relevant for medicine long before machine learning!

- 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

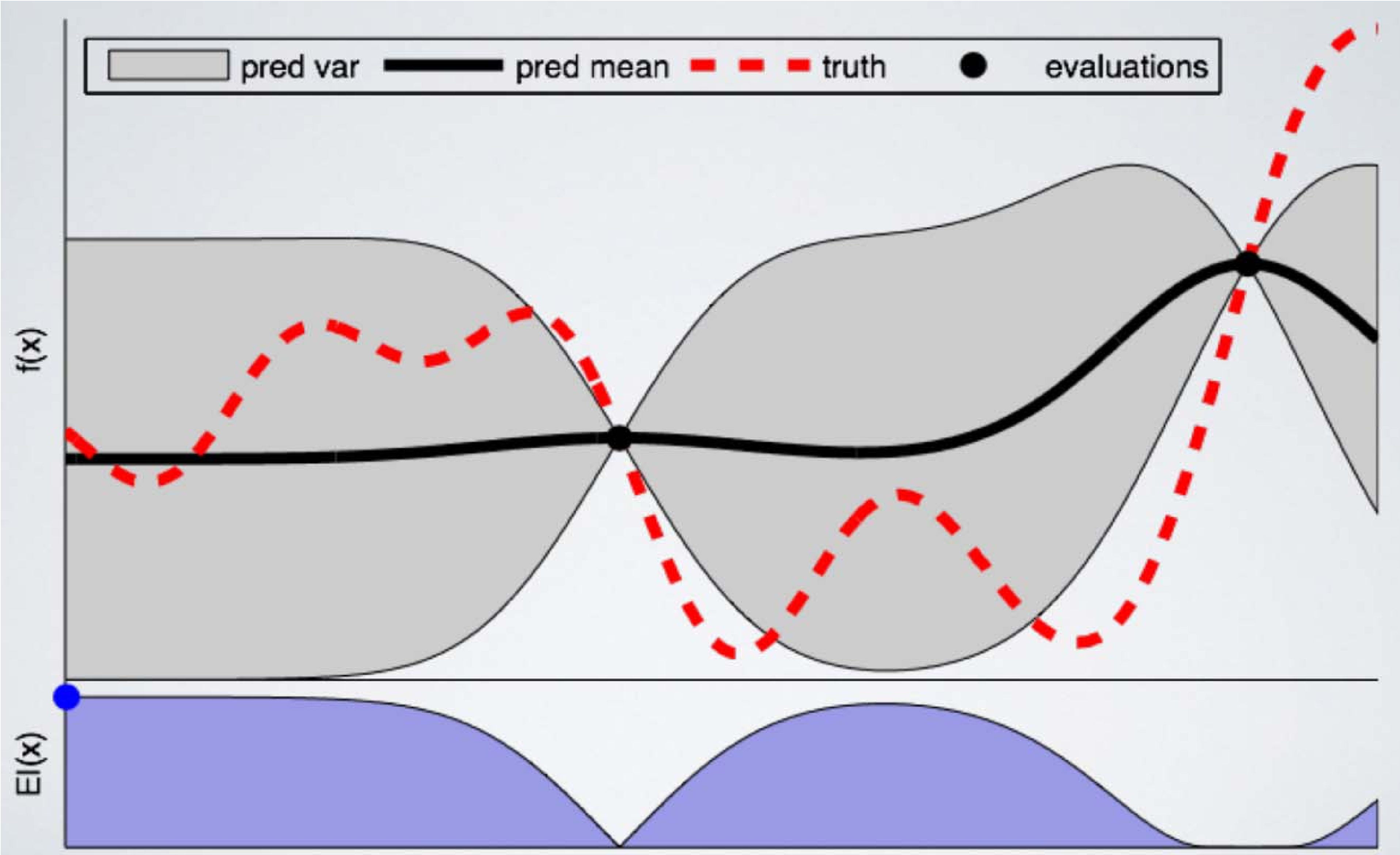
$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}$$

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



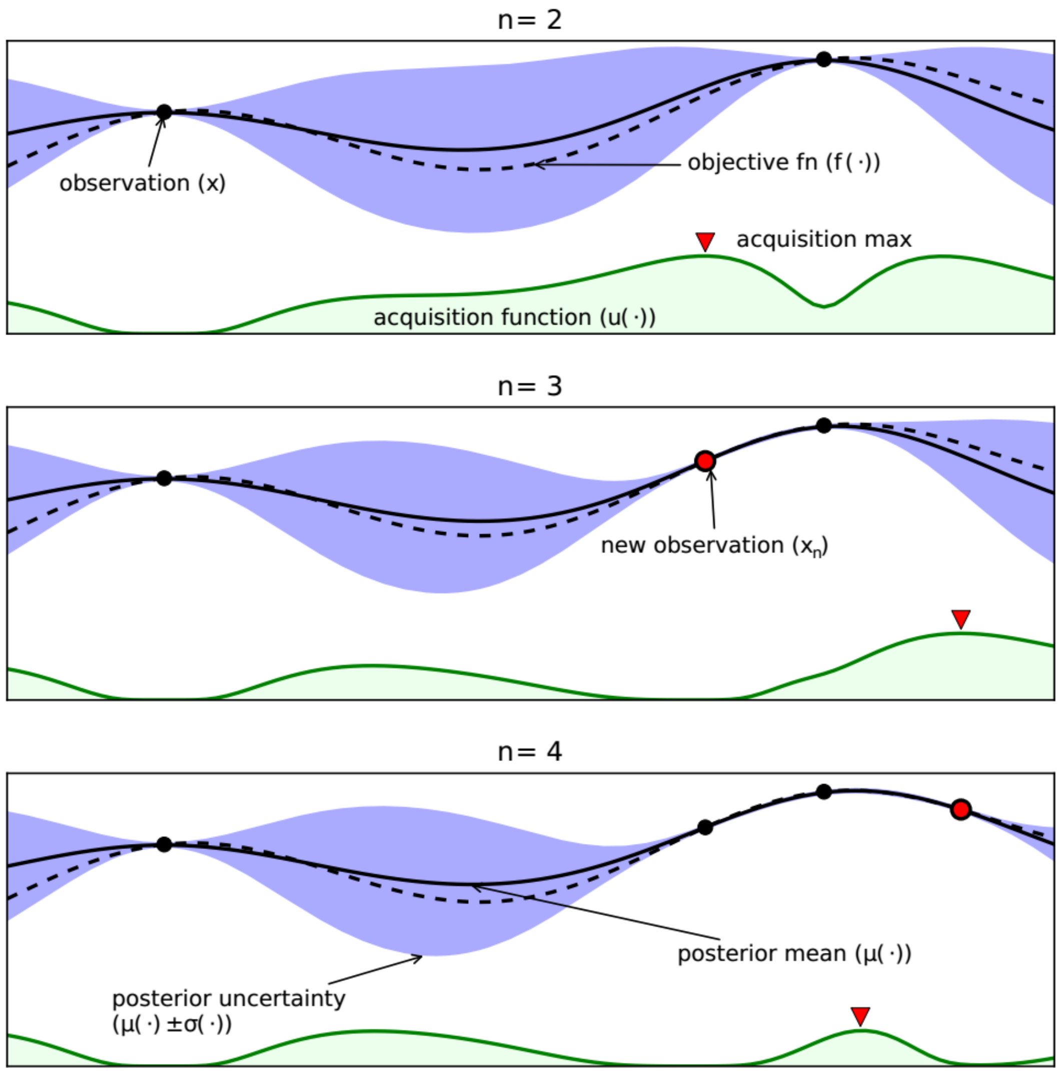
Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

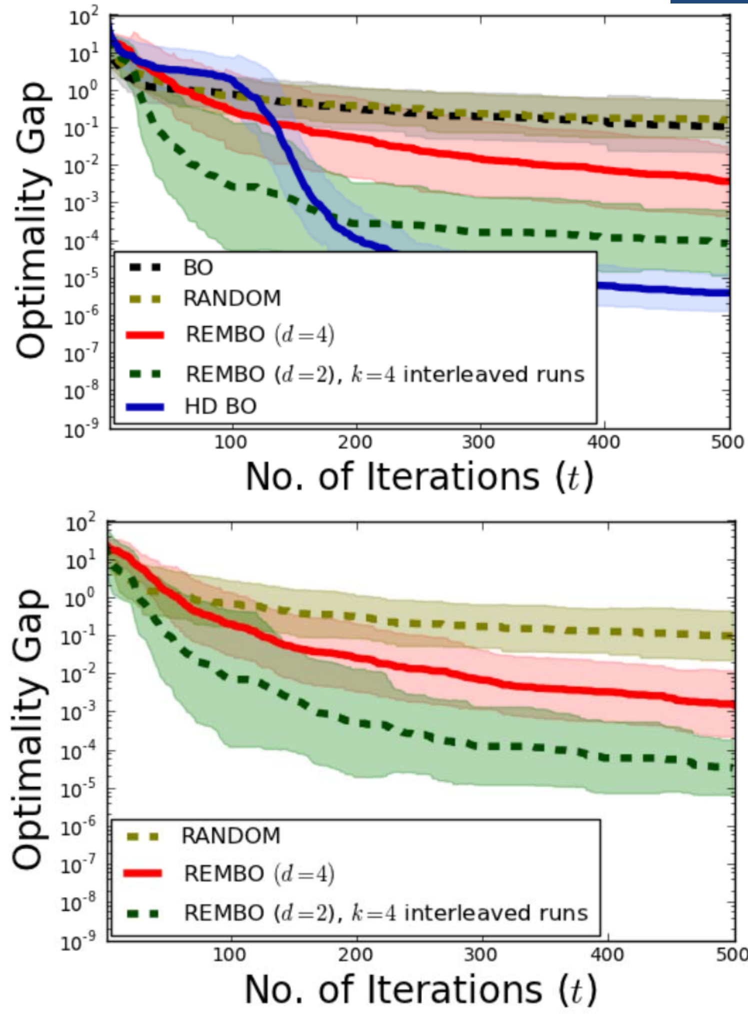
How does “automatic” learning from data work ?



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

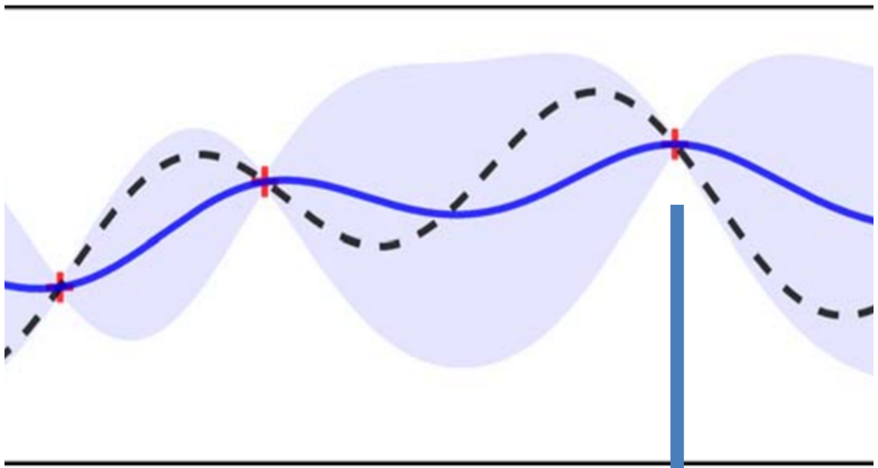
Does this scale well into high-dimensions ?





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 the main goal of automatic machine learning ?

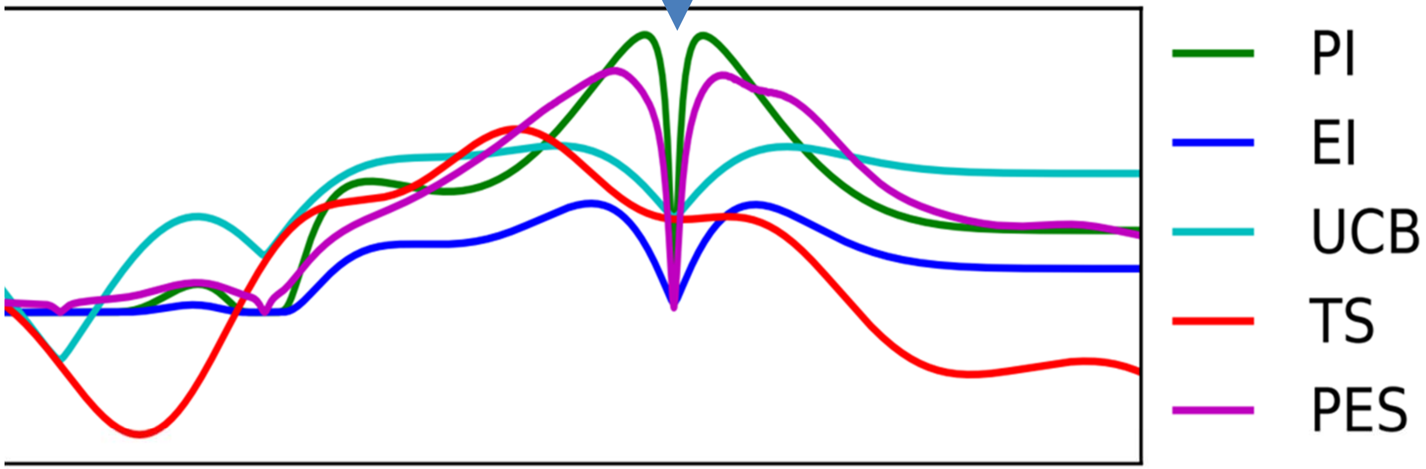


Algorithm 1 Bayesian optimization

1: **for** $n = 1, 2, \ldots$ **do**
2: select new $\mathbf{x}_{n+1}$ by optimizing acquisition function α

$$\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}; \mathcal{D}_n)$$

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**

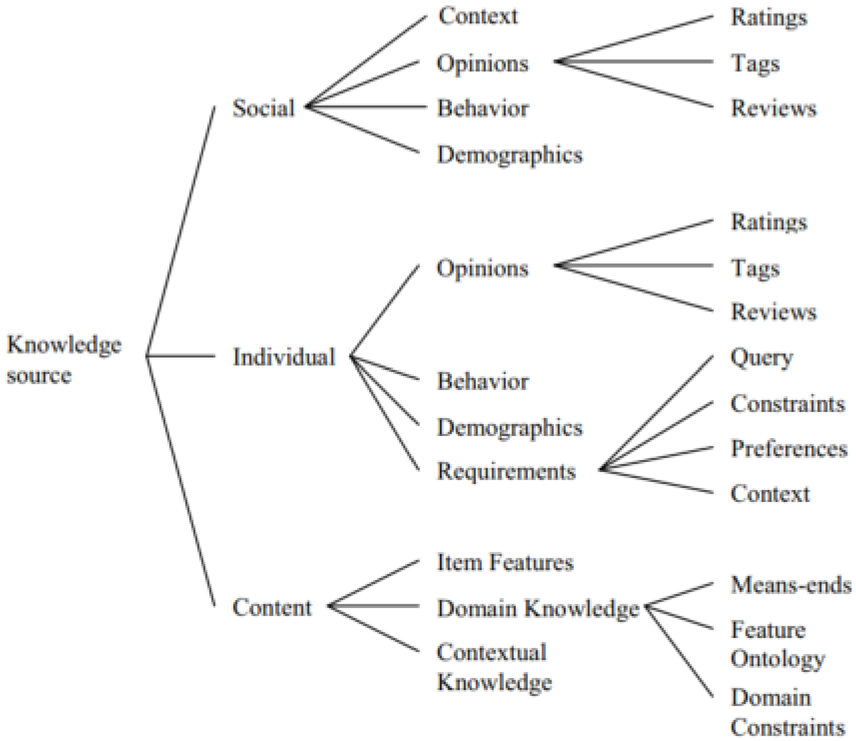


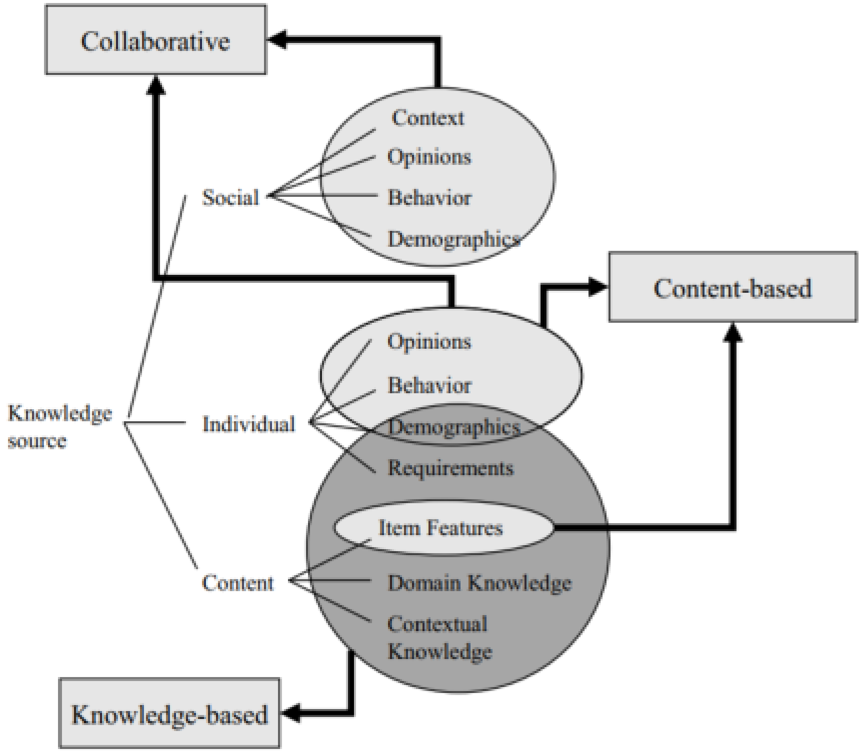
Probability of Improvement
Expected Improvement
Upper Confidence Bound
Thompson Sampling
Predictive Entropy Search

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

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.

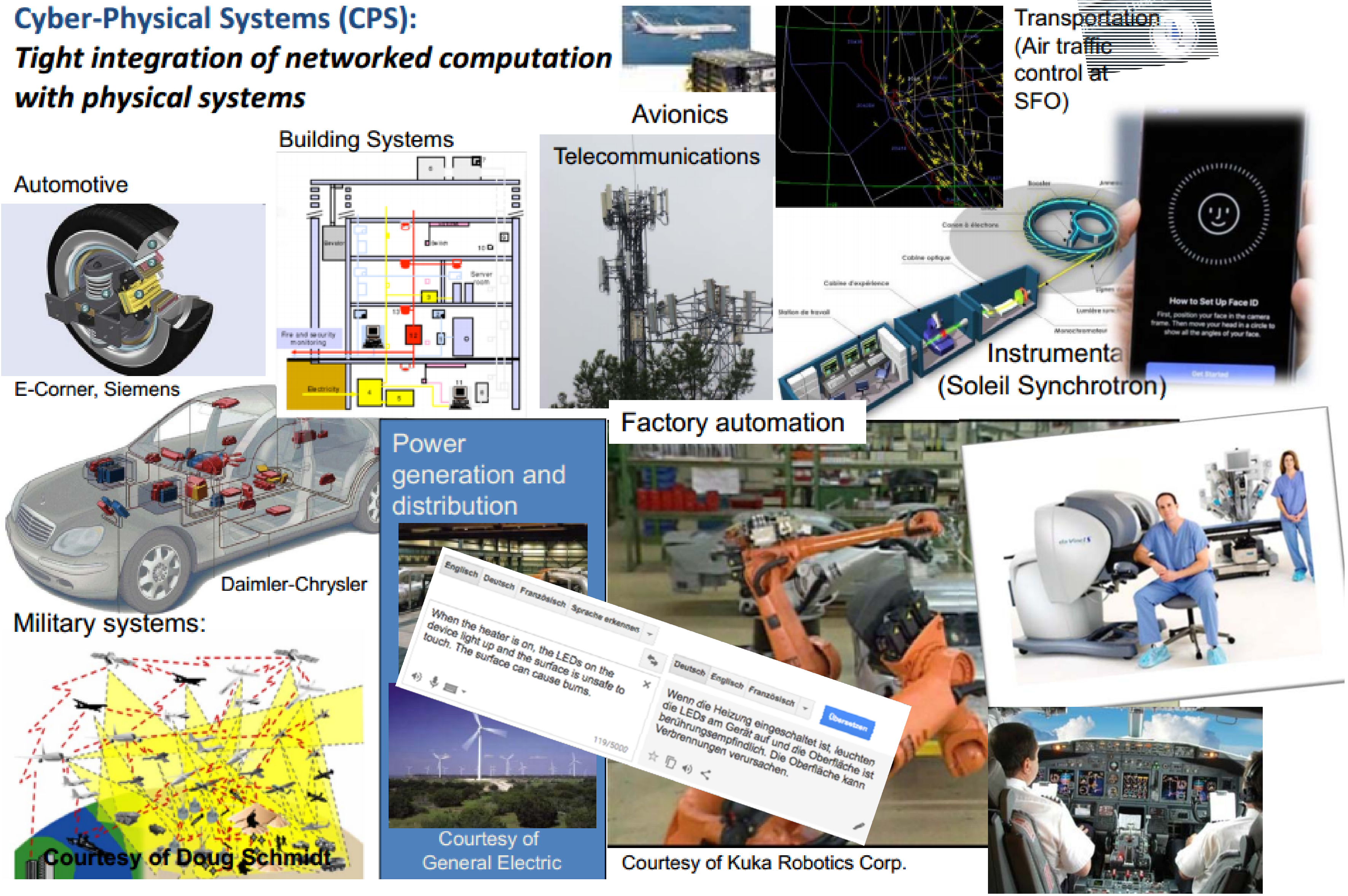
Fully automatic autonomous vehicles (“Google car”)

HCAI
HUMAN-CENTERED.AI

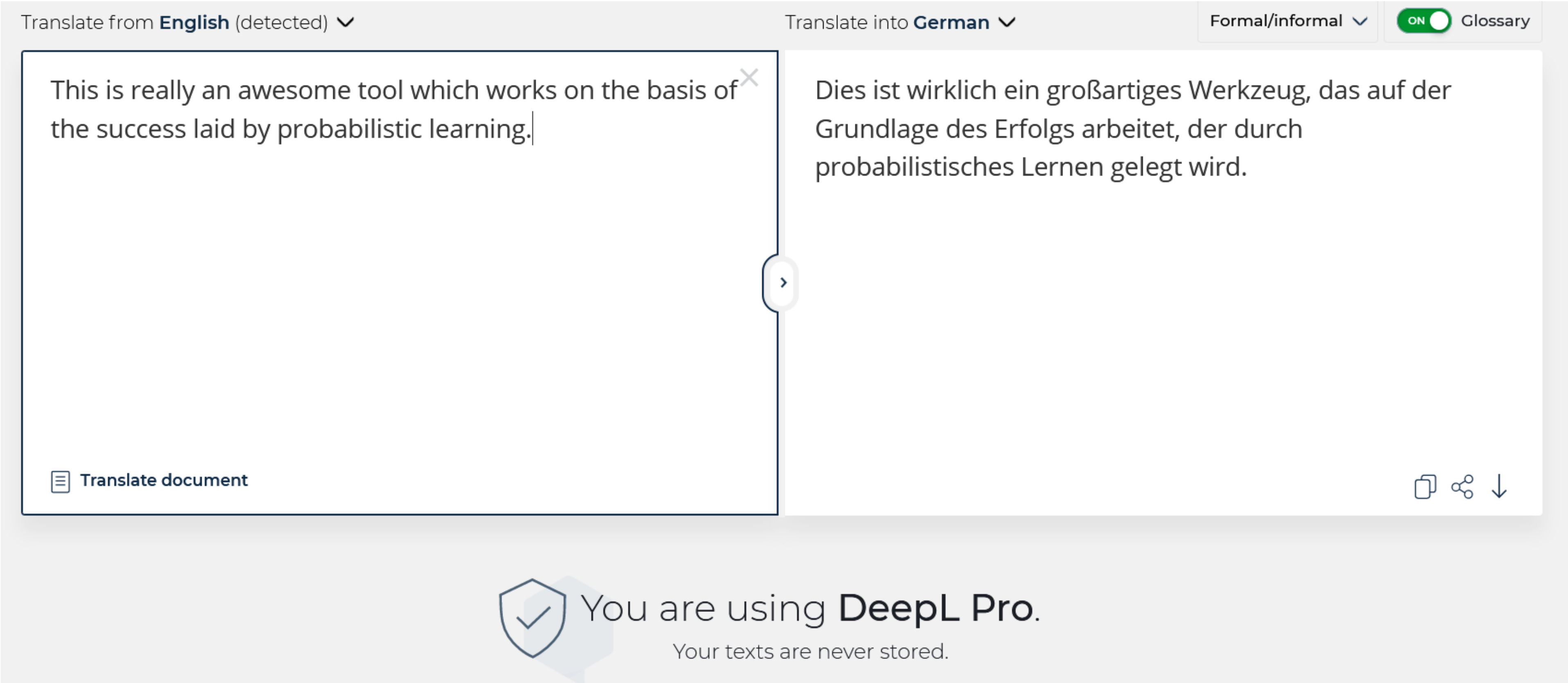
Guizzo, E. 2011. How Google’s self-driving car works. IEEE Spectrum Online, 10, 18.

… and thousands of industrial aML applications …

Cyber-Physical Systems (CPS):
Tight integration of networked computation with physical systems

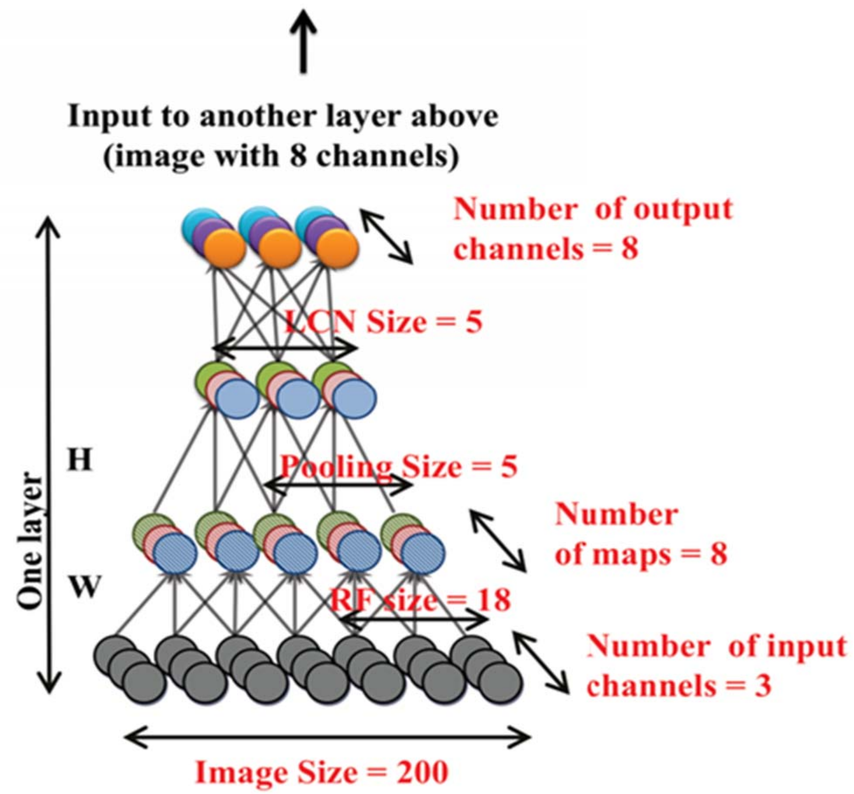


Seshia, S. A., Juniwal, G., Sadigh, D., Donze, A., Li, W., Jensen, J. C., Jin, X., Deshmukh, J., Lee, E. & Sastry, S. 2015. Verification by, for, and of Humans: Formal Methods for Cyber-Physical Systems and Beyond. Illinois ECE Colloquium.



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

What does Deep Learning need ?

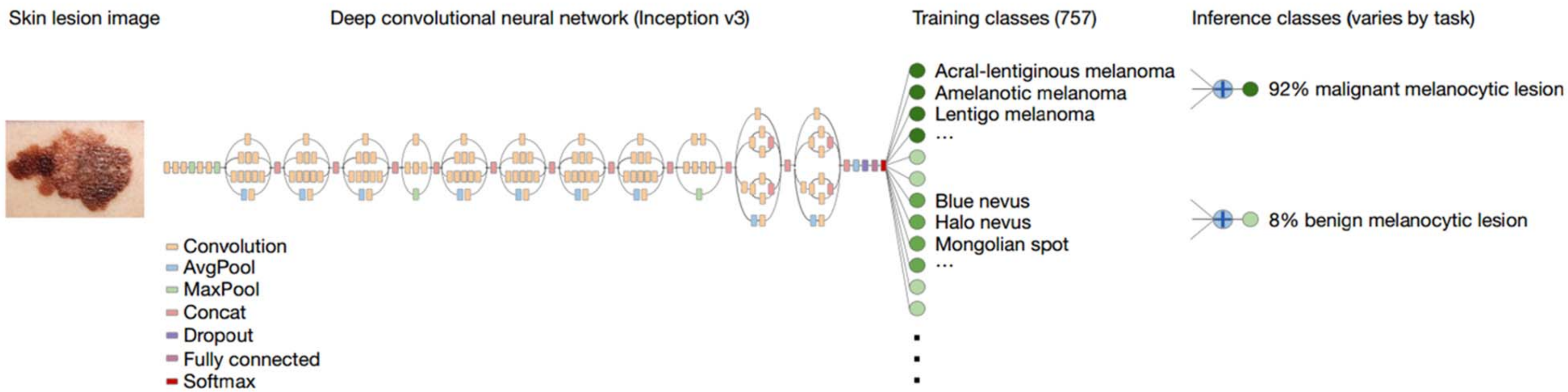


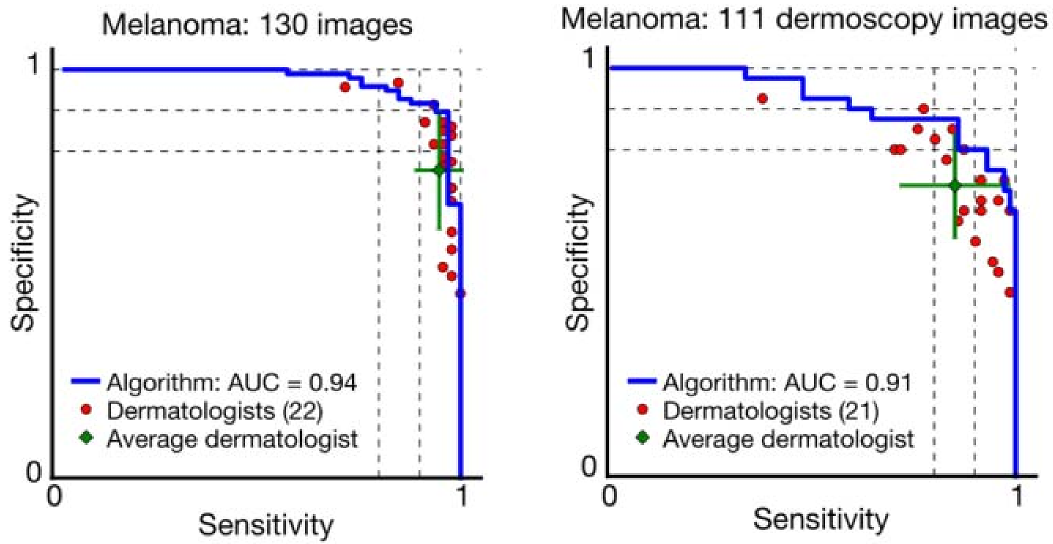
$$x^* = \arg\min_x f(x; W, H), \text{ subject to } ||x||_2 = 1.$$

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.

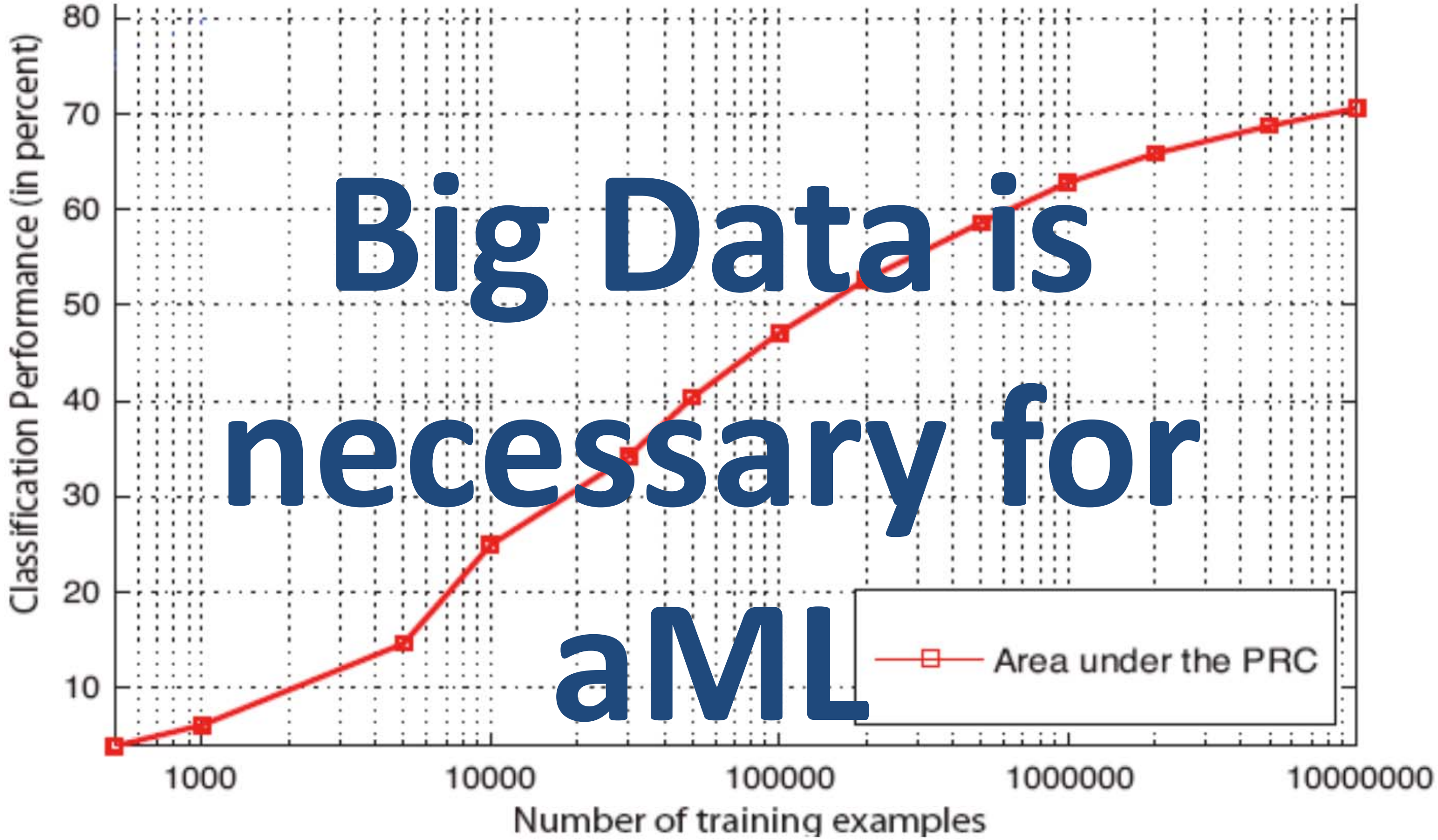
Le, Q. V. 2013. Building high-level features using large scale unsupervised learning. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing ICASSP.* IEEE. 8595-8598, doi:10.1109/ICASSP.2013.6639343.

Remember our success example in dermatology





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.



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.

05 iML

- Sometimes we **do not have "big data",** where aML-algorithms benefit.
- Sometimes we have
  - **Small amount of data sets ("little data")**
  - 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.

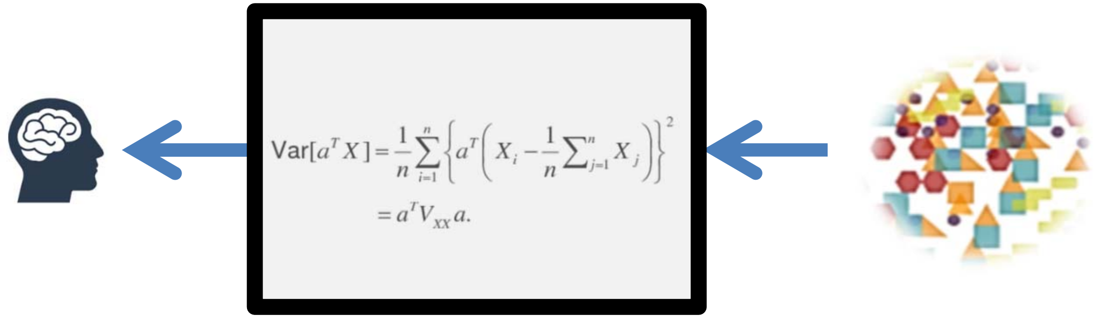
Sometimes we (still) need a human-in-the-loop

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

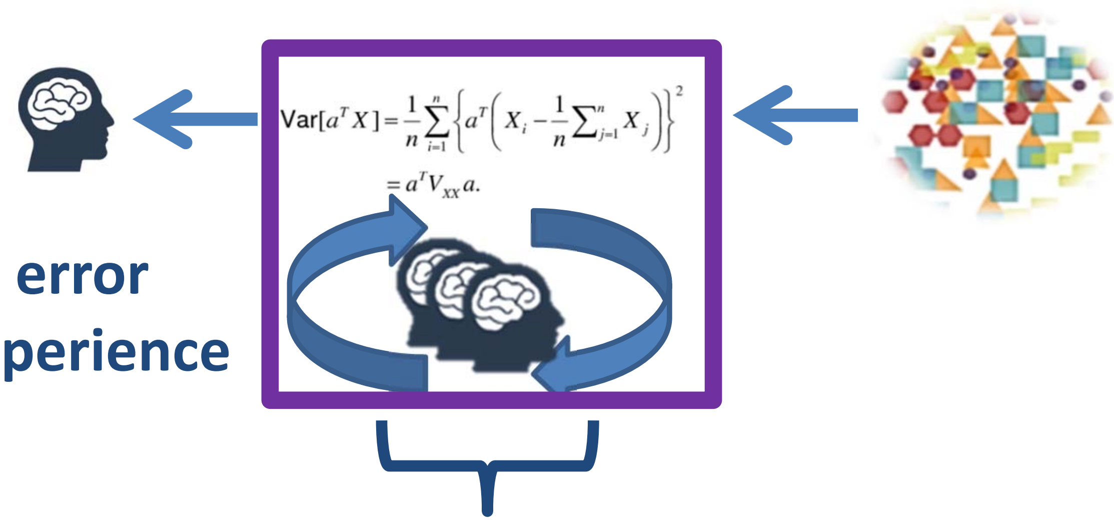
***) where the agents can be human**

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

What is the human in the loop supposed to do ?



Generalization error



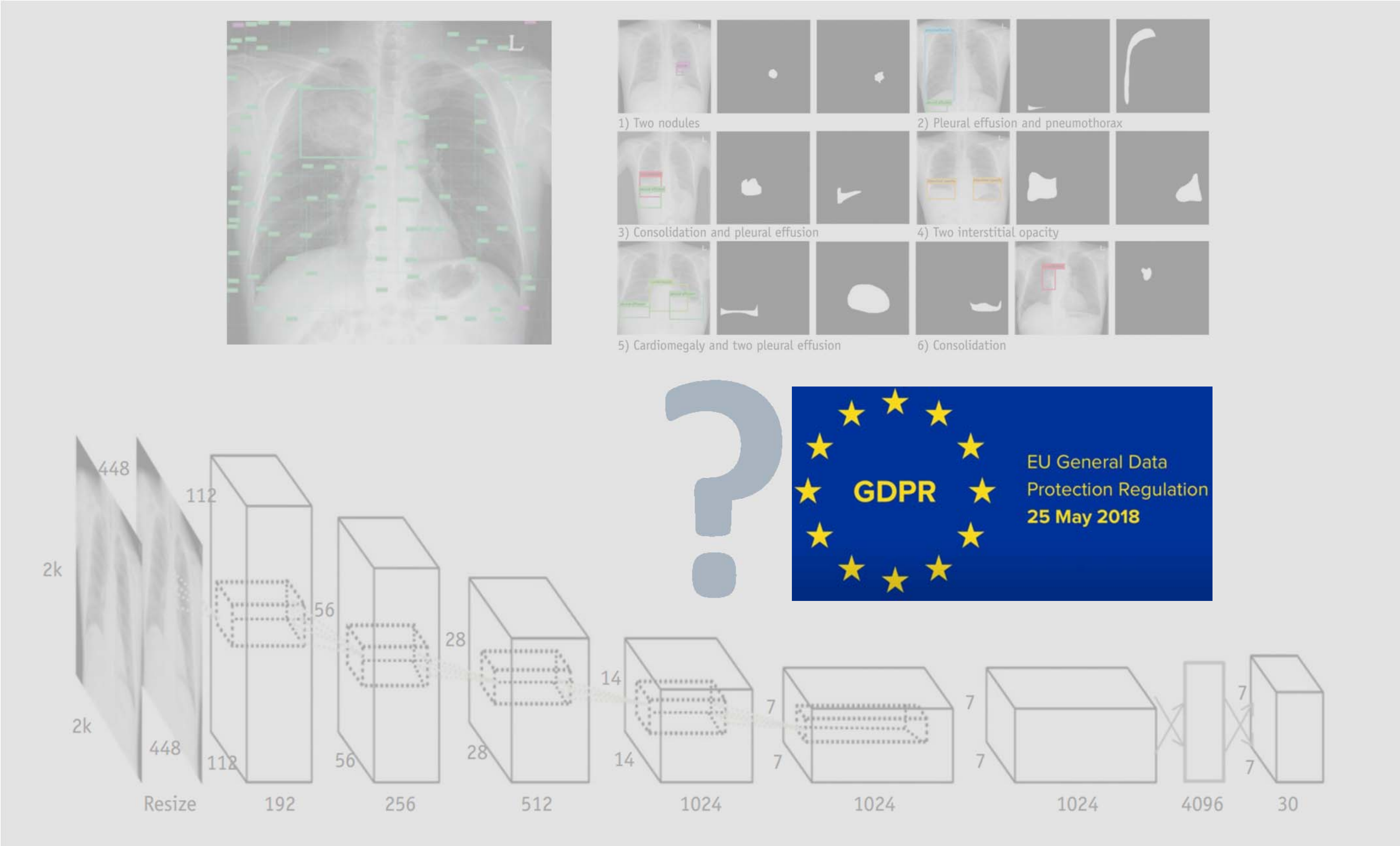
Generalization error
<u>plus</u> human experience

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.

Conclusion

What is problematic with the successful methods ?



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.

- **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, <u>legal</u> aspects make "black box" difficult

- Computational approaches can find in R^n what no human is able to see
- However, still there are many hard problems where a human expert in R^2 can understand the **context** and bring in experience, expertise, knowledge, intuition, …
- Black box approaches can not explain **WHY** a decision has been made …

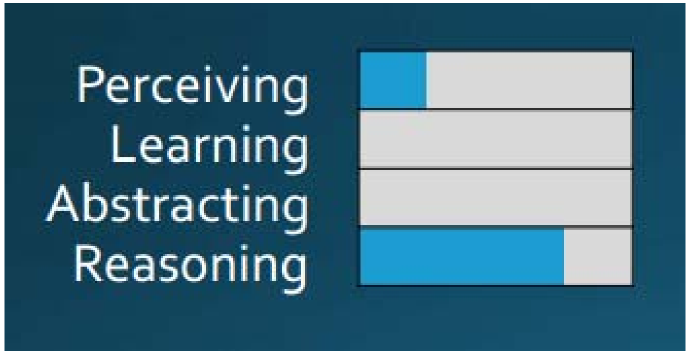


Image credit to John Launchbury, DARPA

- Engineers create a set of logical rules to represent knowledge (Rule based Expert Systems)
- Advantage: works well in narrowly defined problems of well-defined domains
- Disadvantage: No adaptive learning behaviour and poor handling of $p(x)$

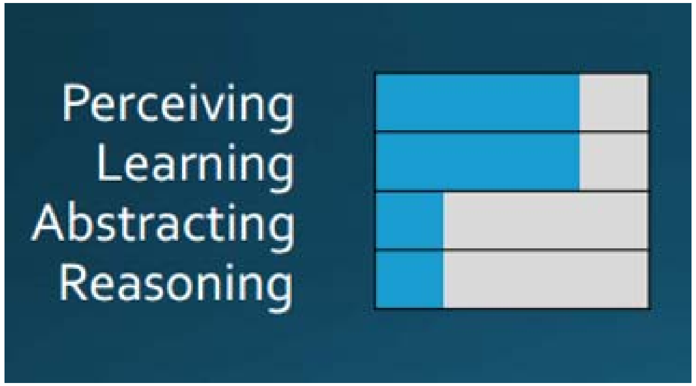


Image credit to John Launchbury, DARPA

- Engineers create learning models for specific tasks and train them with "big data" (e.g. Deep Learning)
- Advantage: works well for standard classification tasks and has prediction capabilities
- Disadvantage: No contextual capabilities and minimal reasoning abilities

Image credit to John Launchbury, DARPA

- A contextual model can perceive, learn and understand and abstract and reason
- Advantage: can use transfer learning for adaptation on unknown unknowns
- Disadvantage: Superintelligence …

- Myth 1a: Superintelligence by 2100 is inevitable!
- Myth 1b: Superintelligence by 2100 is impossible!
- **Fact: We simply don’t know it!**
- Myth 2: Robots are our main concern
- **Fact: Cyberthreats are the main concern – it needs no body – only an Internet connection!**
- Myth 3: AI will never control us humans
- **Fact: Intelligence is an enabler for control: We control tigers just by being smarter ...**

Human-Centered AI (HCAI)
ensures
Human-in-control

HCAI
HUMAN-CENTERED.AI

Thank you!