

Mini Course From Data Science to interpretable AI



Assoc. Prof. Dr. Andreas HOLZINGER (Medical University Graz)

Day 1 > Part 1 > Monday, 17.06.2019

**Introduction to
AI/Machine Learning**

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

- Austrian Representative in IFIP TC 12 “Artificial Intelligence”
- Member of IFIP WG 12.9 “Computational Intelligence”

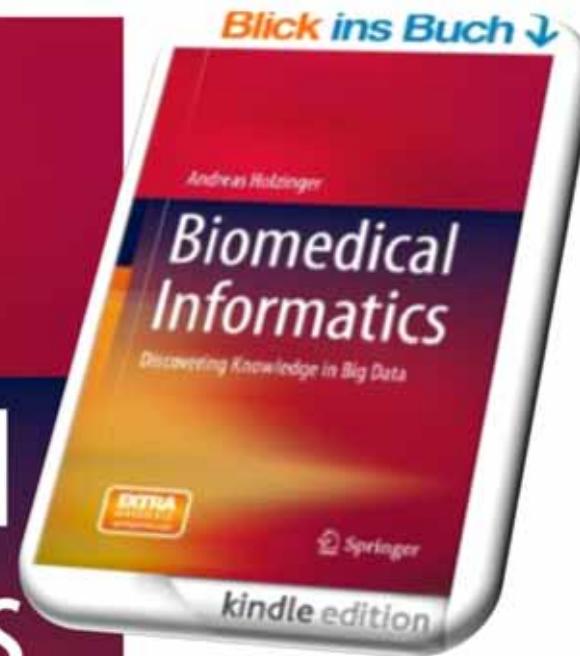
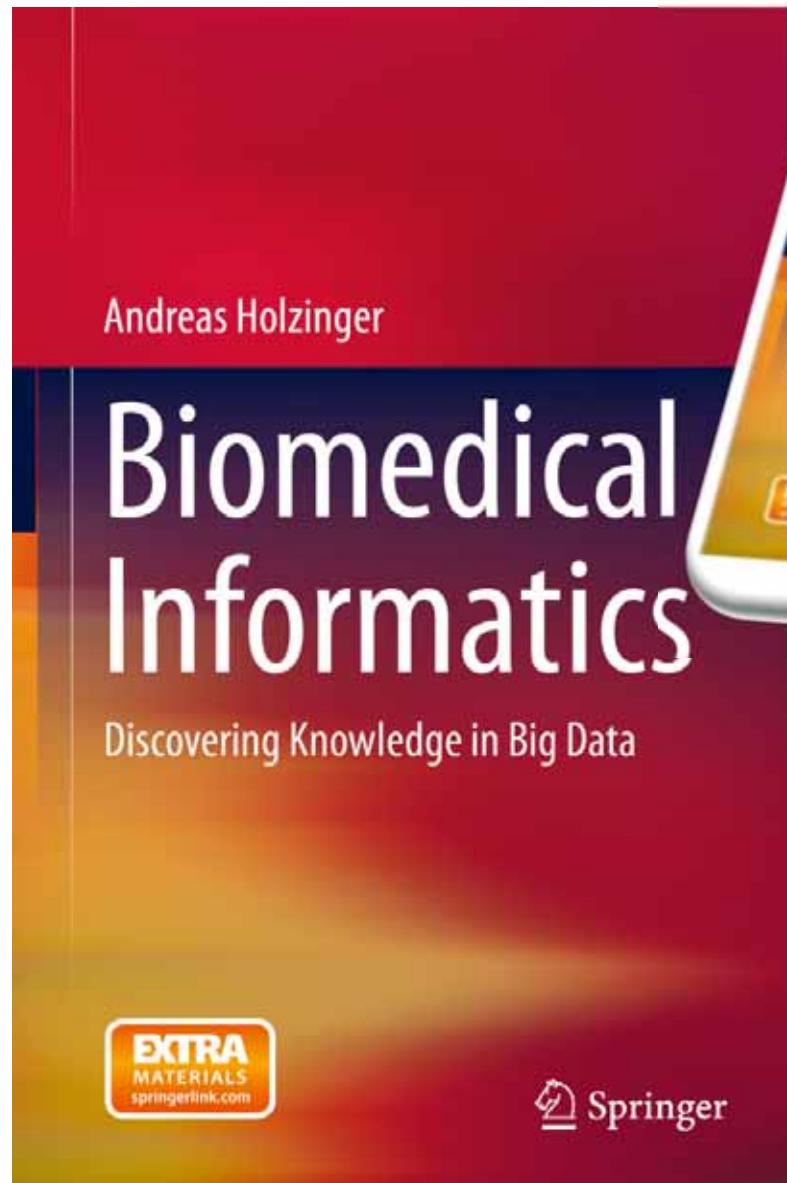
- PhD in Cognitive Science 1998
- Habilitation Computer Science 2003
- Lead Human-Centered AI (Holzinger Group)
- Personal Homepage: <https://www.aholzinger.at>

- Visiting Professor for Machine Learning
in Health Informatics: TU Vienna, Univ. Verona,
UCL London, RWTH Aachen
- 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](https://arxiv.org/abs/1712.09923)
- Andreas Holzinger, 2018. Explainable AI (ex-AI). Informatik-Spektrum, [doi:10.1007/s00287-018-1102-5](https://doi.org/10.1007/s00287-018-1102-5)
- 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](https://doi.org/10.1002/widm.1312)



- At the end of this mini course you will ...
- ... be fascinated to see our world in **data sets**;
- ... understand the differences between
data, information and knowledge
- ... be aware of some problems and challenges in
biomedical informatics
- ... understand the importance of the concept of
probabilistic information $p(x)$
- ... know what **AI/Machine Learning** can (not) do
- ... have some fundamental insight into medical
information science for **decision making**

Background Reading



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

Primer on Probability & Information

Day 1 - Fundamentals

01 Introduction to AI
Machine Learning



02 Data, Information
and Knowledge



03 Decision Making and
Decision Support



04 Causal Reasoning and
Interpretable AI

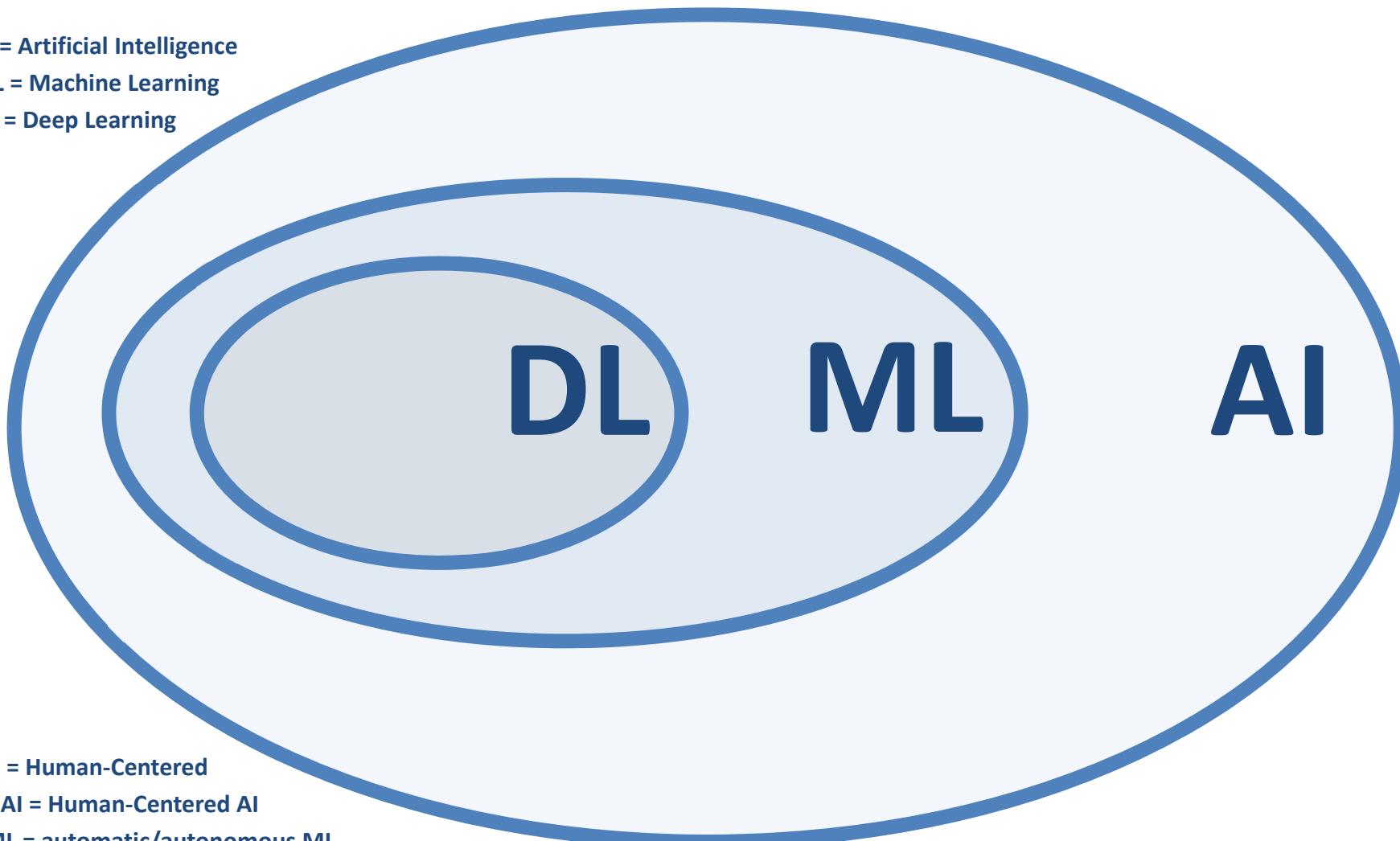
- **01 What is HCAI ?**
- **02 Application Area: Health Informatics**
- **03 Probabilistic Information**
- **04 Automatic Machine Learning**
- **05 Interactive Machine Learning**
- **06 Causality, Explainability, Interpretability**

Abbreviations

AI = Artificial Intelligence

ML = Machine Learning

DL = Deep Learning



HC = Human-Centered

HCAI = Human-Centered AI

aML = automatic/autonomous ML

iML = interactive ML, interpretable ML

KDD = Knowledge Discovery from Data

ExAI = explainable AI

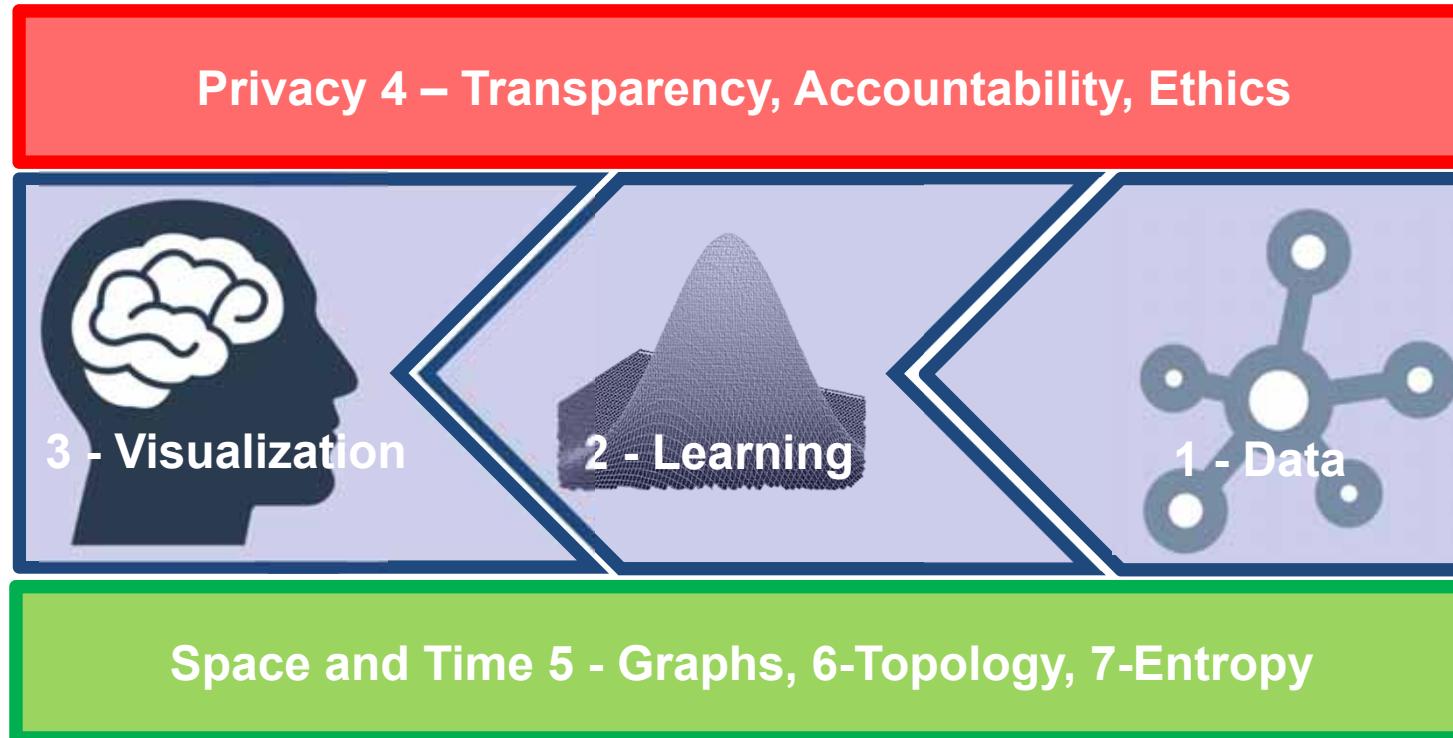
Andreas Holzinger, Peter Kieseberg, Edgar Weippl & A Min Tjoa 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. Springer Lecture Notes in Computer Science LNCS 11015. Cham: Springer, pp. 1-8, doi:[10.1007/978-3-319-99740-7_1](https://doi.org/10.1007/978-3-319-99740-7_1)

01 What is

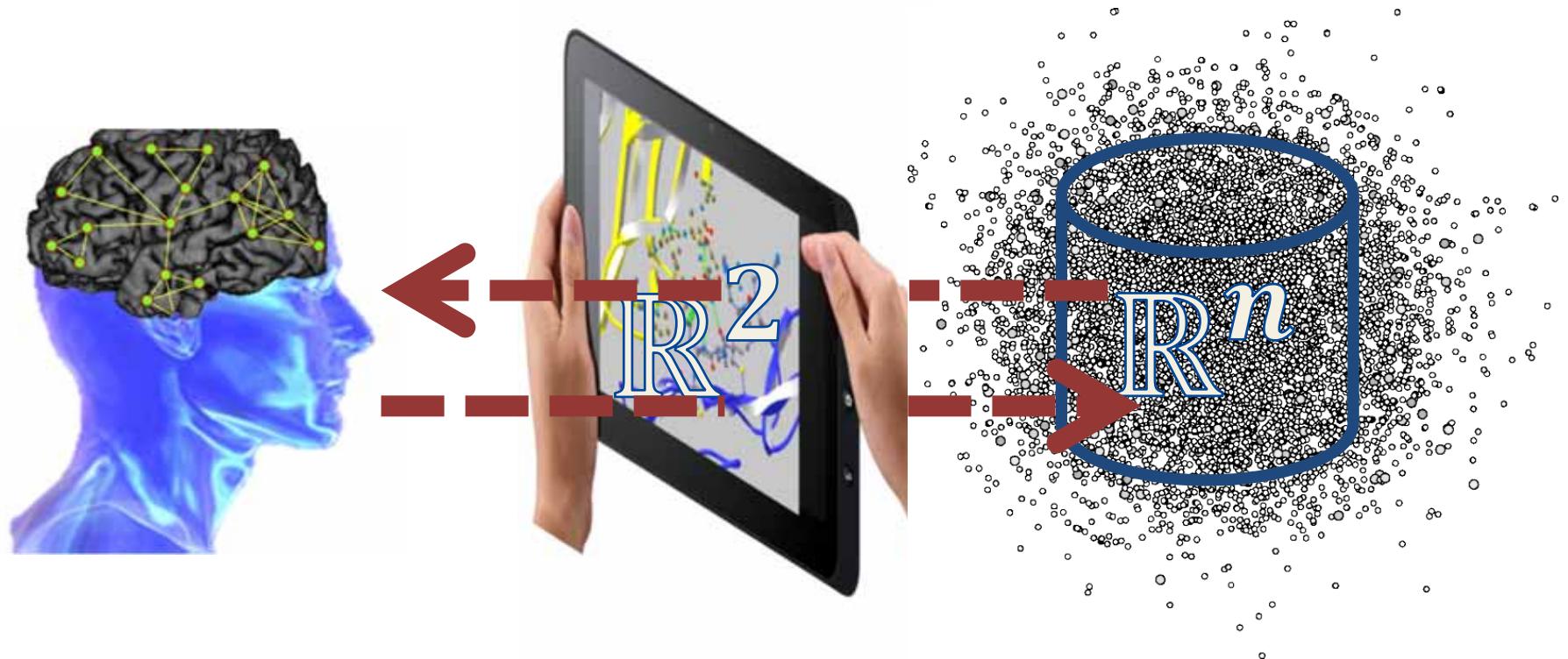


- **ML is a very practical field – algorithm development is at the core – however, successful ML needs a concerted effort of various topics ...**



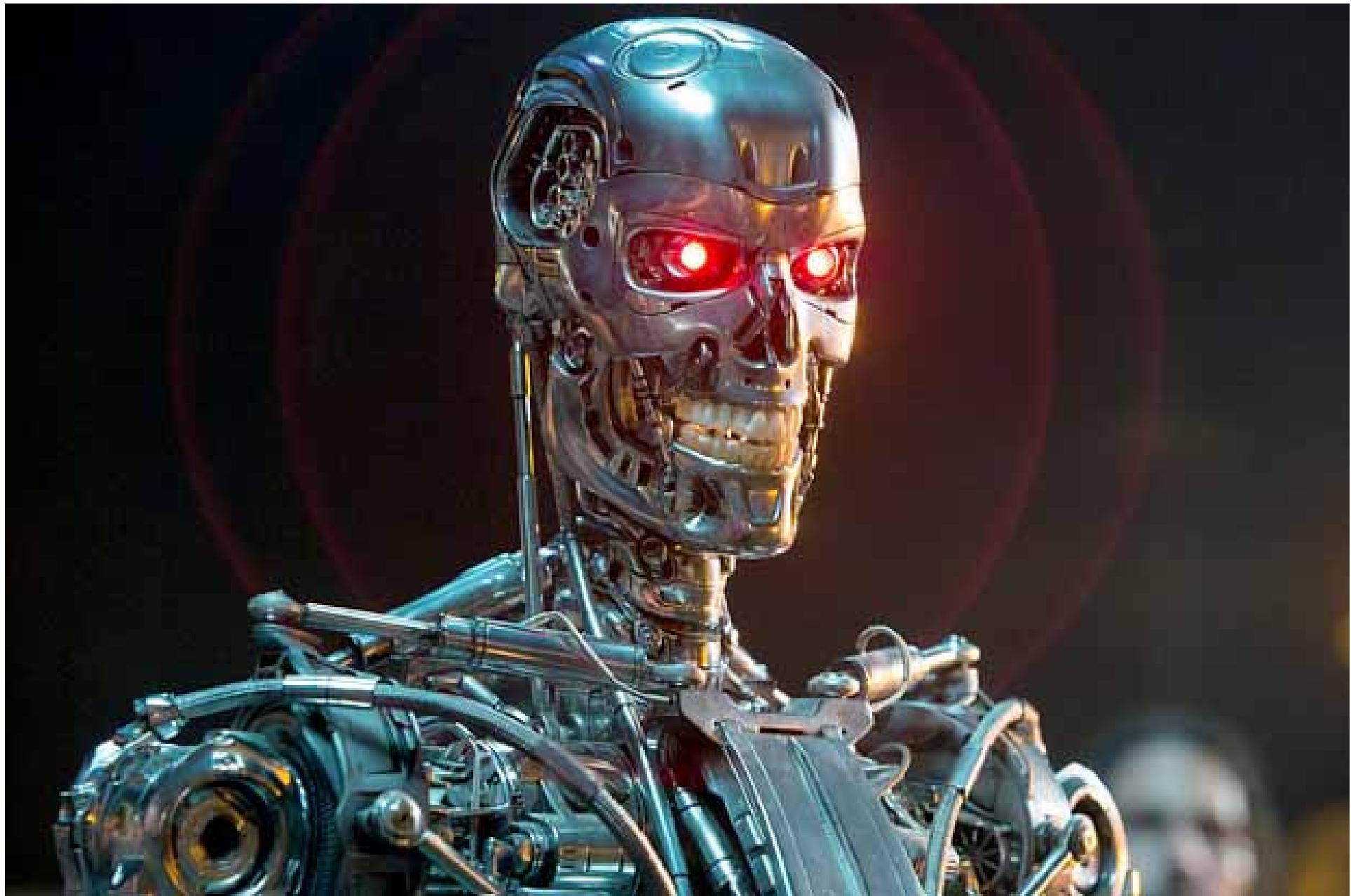


Andreas Holzinger 2013. Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Heidelberg, Berlin, New York: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



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

Not our Goal: Humanoid AI

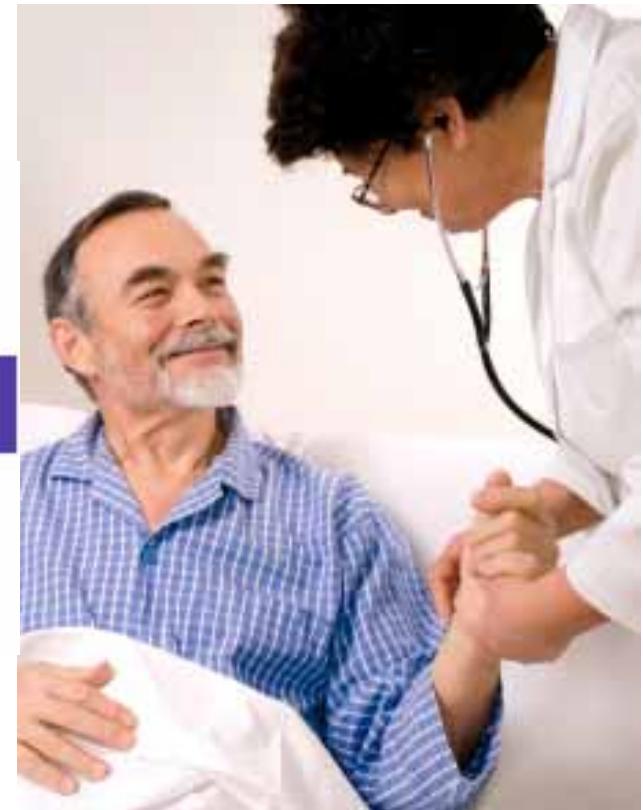


- 1) learn from prior data
- 2) extract knowledge
- 3) generalize, i.e. guessing where a probability mass function concentrates
- 4) fight the curse of dimensionality
- 5) disentangle **underlying explanatory factors of data**, i.e.
- 6) **understand** the data in the **context** of an application domain

02 Application Area Health Informatics



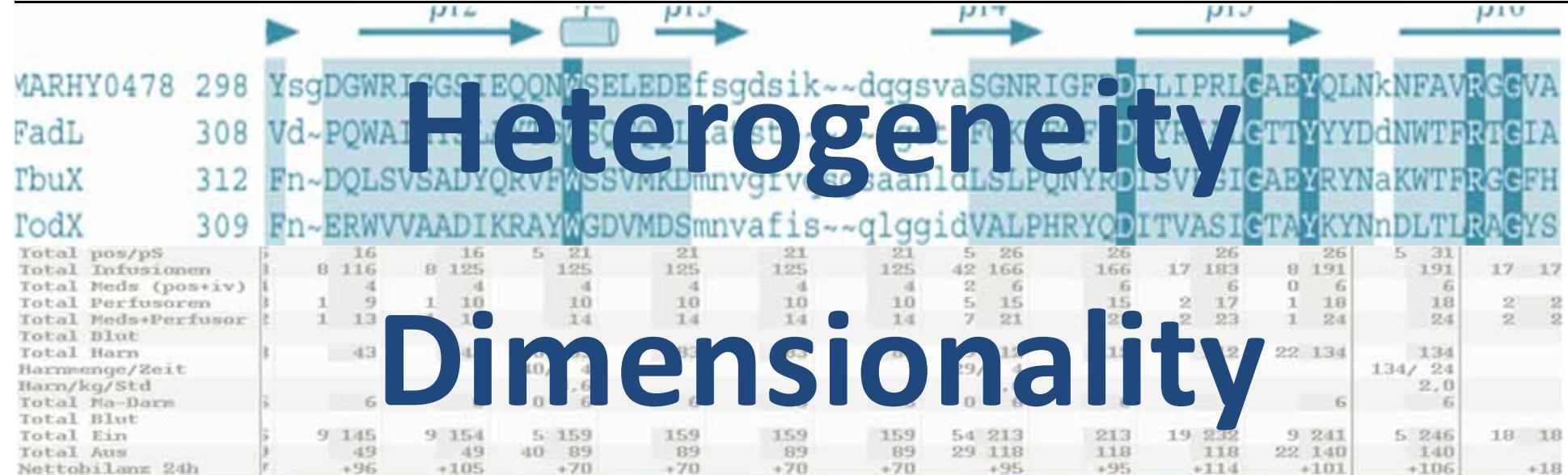
Why is this application area complex ?



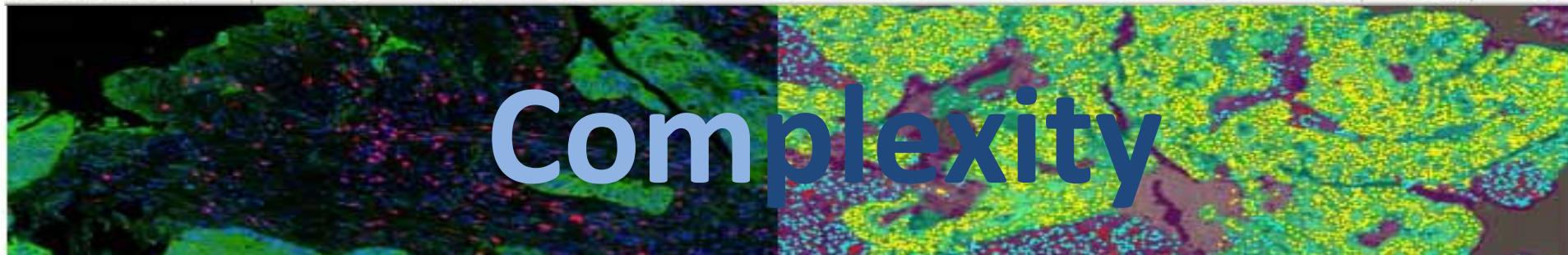
Our central hypothesis: Information may bridge this gap

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





Heterogeneity Dimensionality



Uncertainty

03 Probabilistic 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)

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

What is the simplest mathematical operation for us?

$$p(x) = \sum_x (p(x, y)) \quad (1)$$

How do we call repeated adding?

$$p(x, y) = p(y|x) * p(y) \quad (2)$$

Laplace (1773) showed that we can write:

$$p(x, y) * p(y) = p(y|x) * p(x) \quad (3)$$

Now we introduce a third, more complicated operation:

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

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

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

Observed data:



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

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

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

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

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

Posterior $\approx p(x)$ of h after seen (“learn”) data d $p(\theta|\mathcal{D})$ $p(h|d)$

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

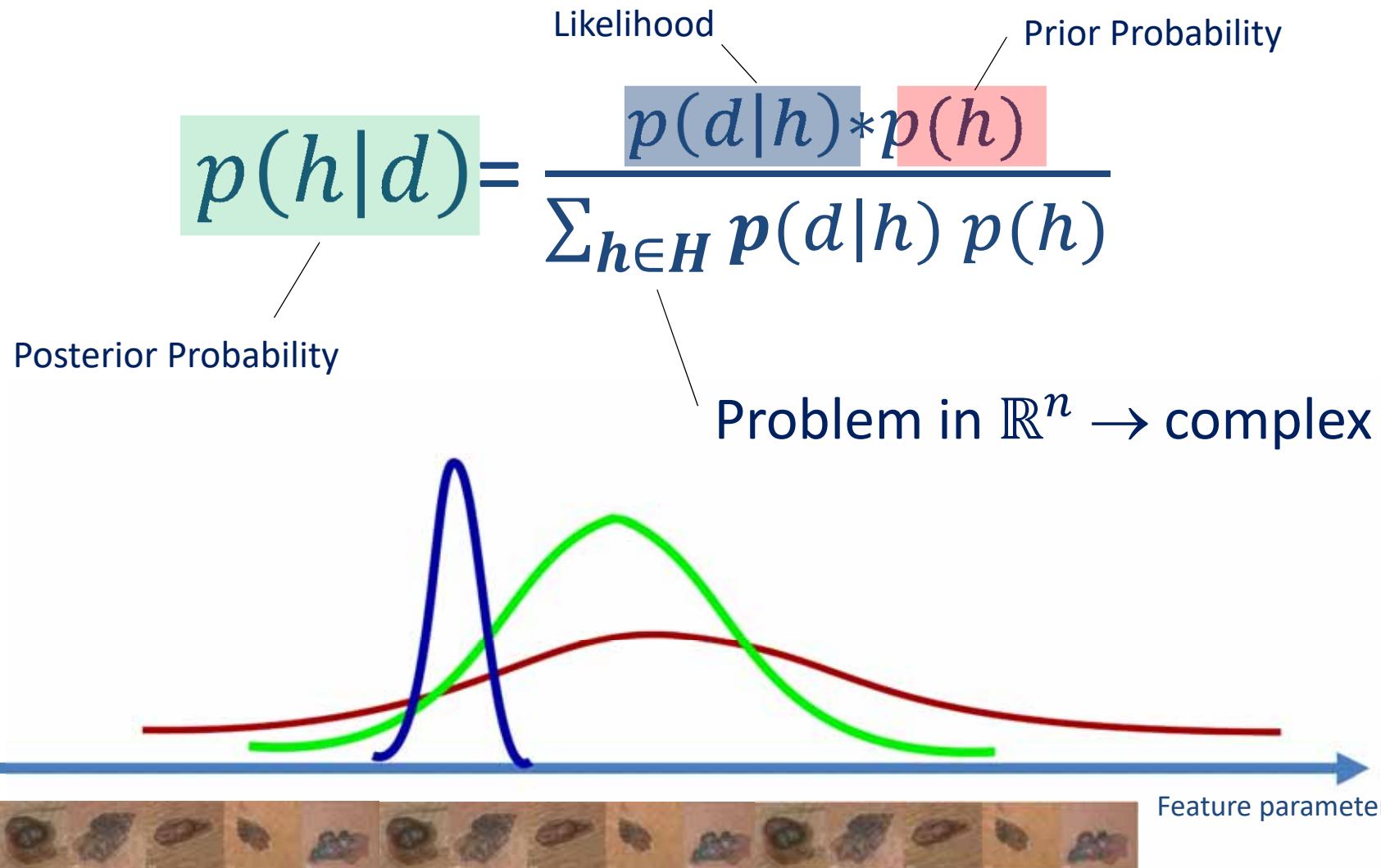
$$p(h|d) = \frac{p(d|h) * p(h)}{\sum_{h \in H} p(d|h) p(h)}$$

d ... data

$\mathcal{H} \dots \{H_1, H_2, \dots, H_n\}$

$\forall h, d \dots$

h ... hypotheses

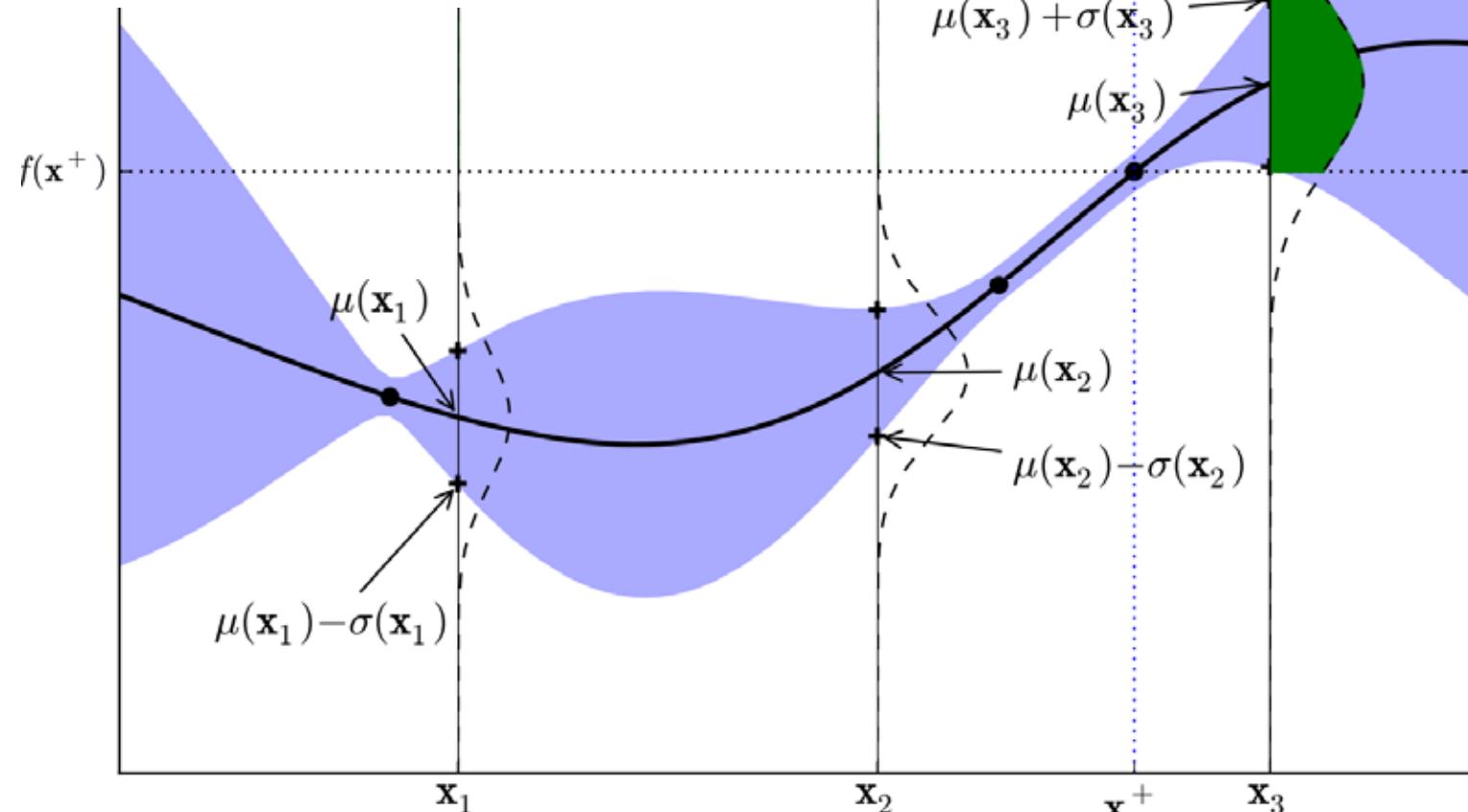


Why is this relevant for medicine?

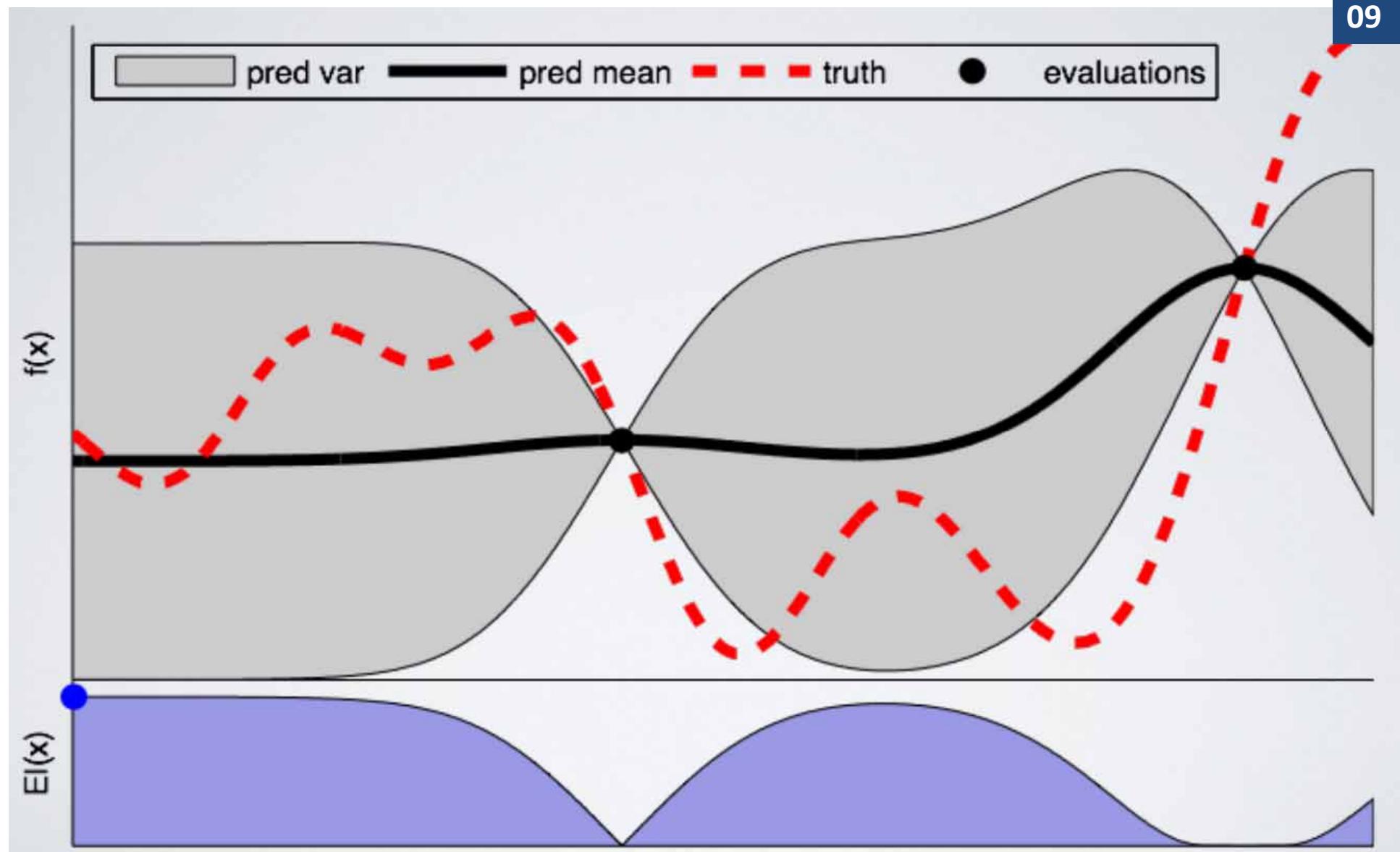
- Take patient information, e.g., observations, symptoms, test results, -omics data, etc. etc.
- Reach conclusions, and **predict** into the future, e.g. how likely will the patient be ...
- Prior = belief before making a particular observation
- Posterior = belief after making the observation and is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

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

$$\text{GP posterior} \quad p(f(x)|\mathcal{D}) \propto \text{Likelihood} \underbrace{p(\mathcal{D}|f(x))}_{p(f(x))}$$

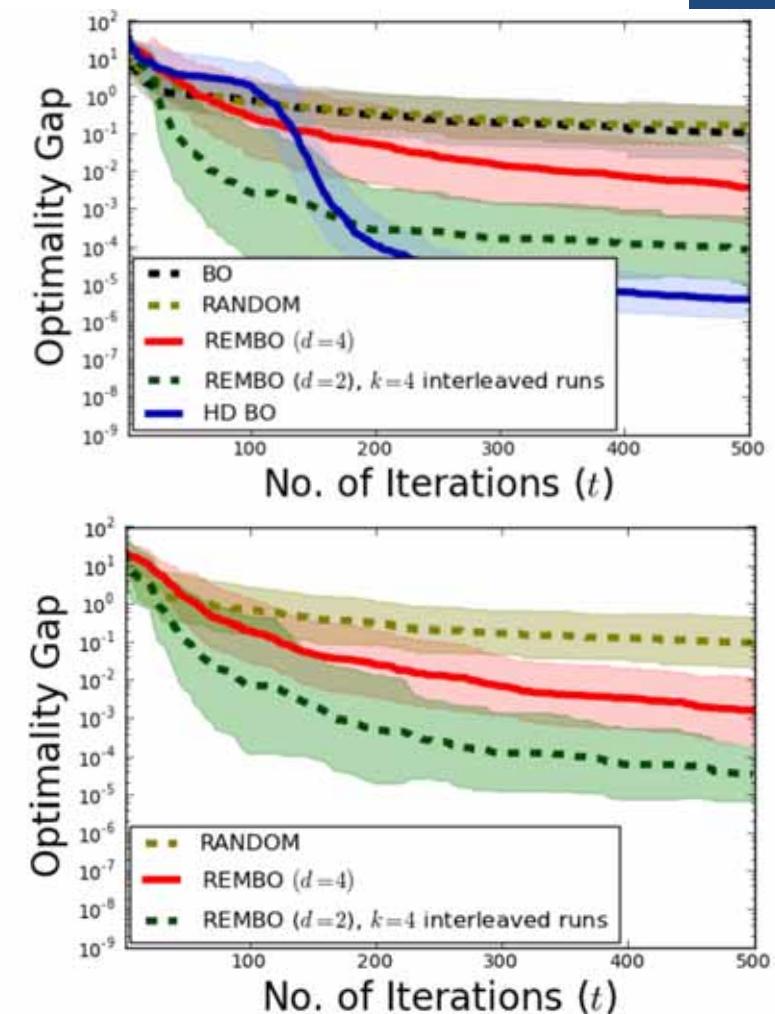
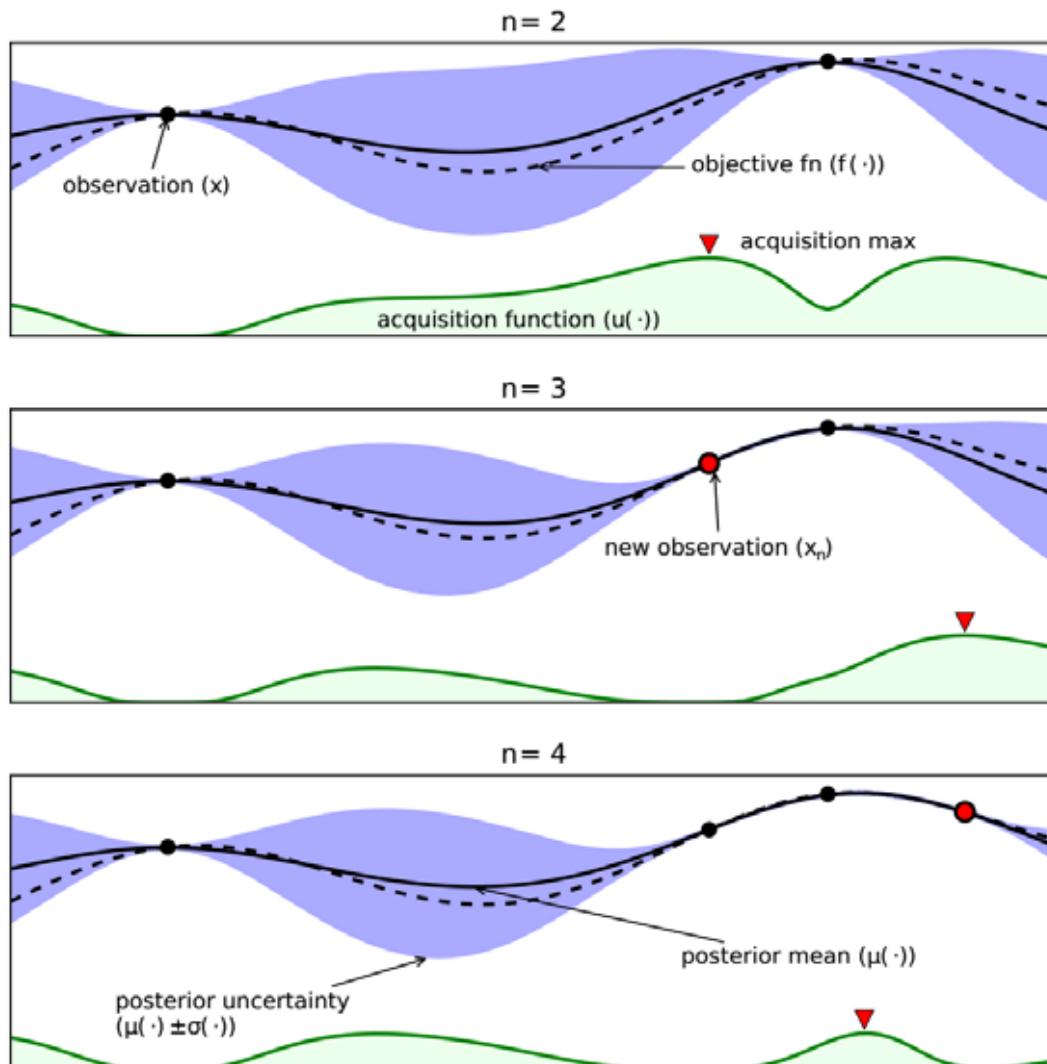


Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

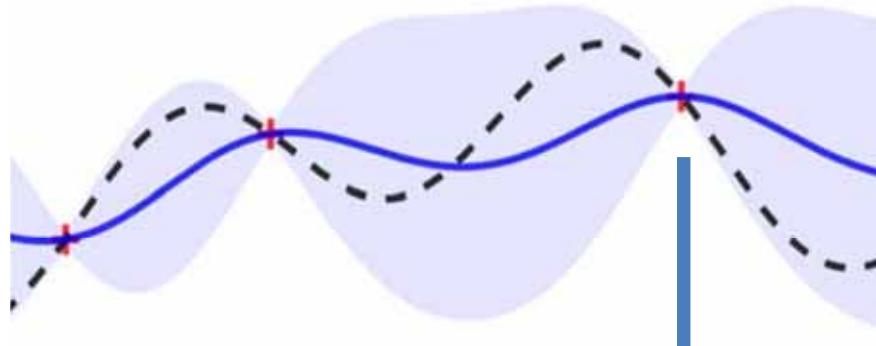


Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2012. 2951-2959.

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.



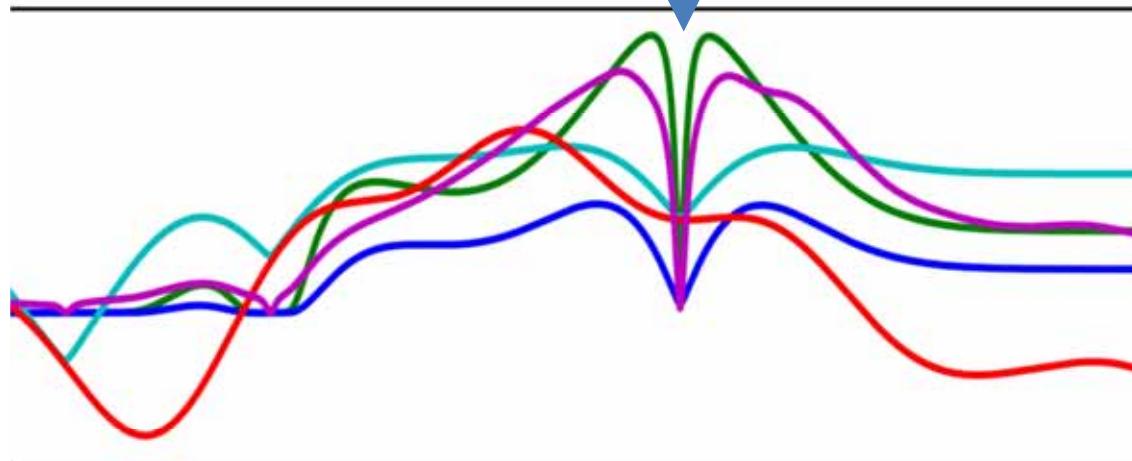
Algorithm 1 Bayesian optimization

```

1: for  $n = 1, 2, \dots$  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

```



- PI Probability of Improvement
- EI Expected Improvement
- UCB Upper Confidence Bound
- TS Thompson Sampling
- PES Predictive Entropy Search

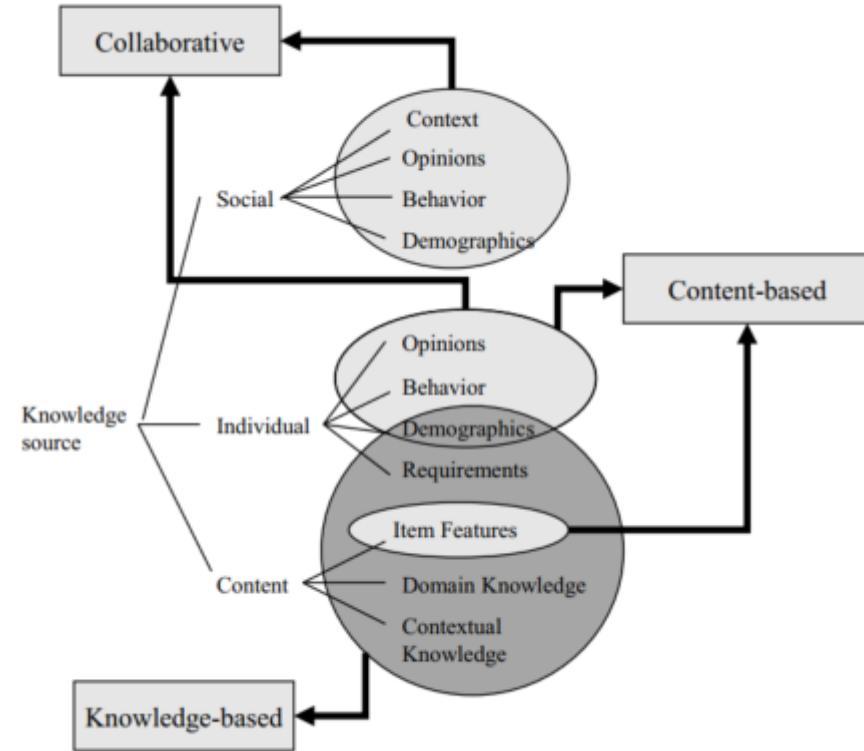
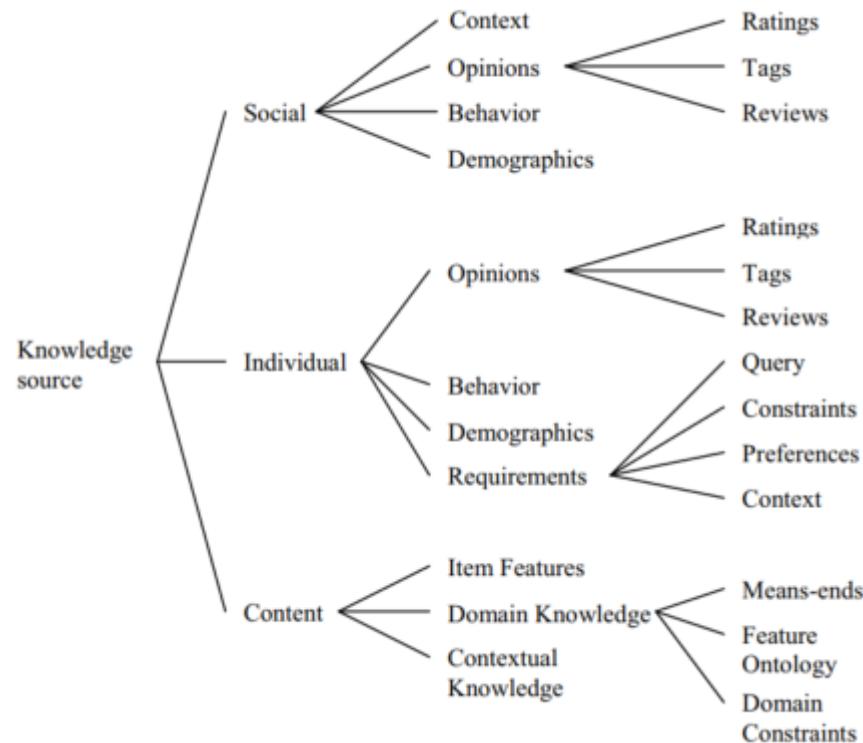
Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. 2016.

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”)



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

Fully automatic autonomous vehicles

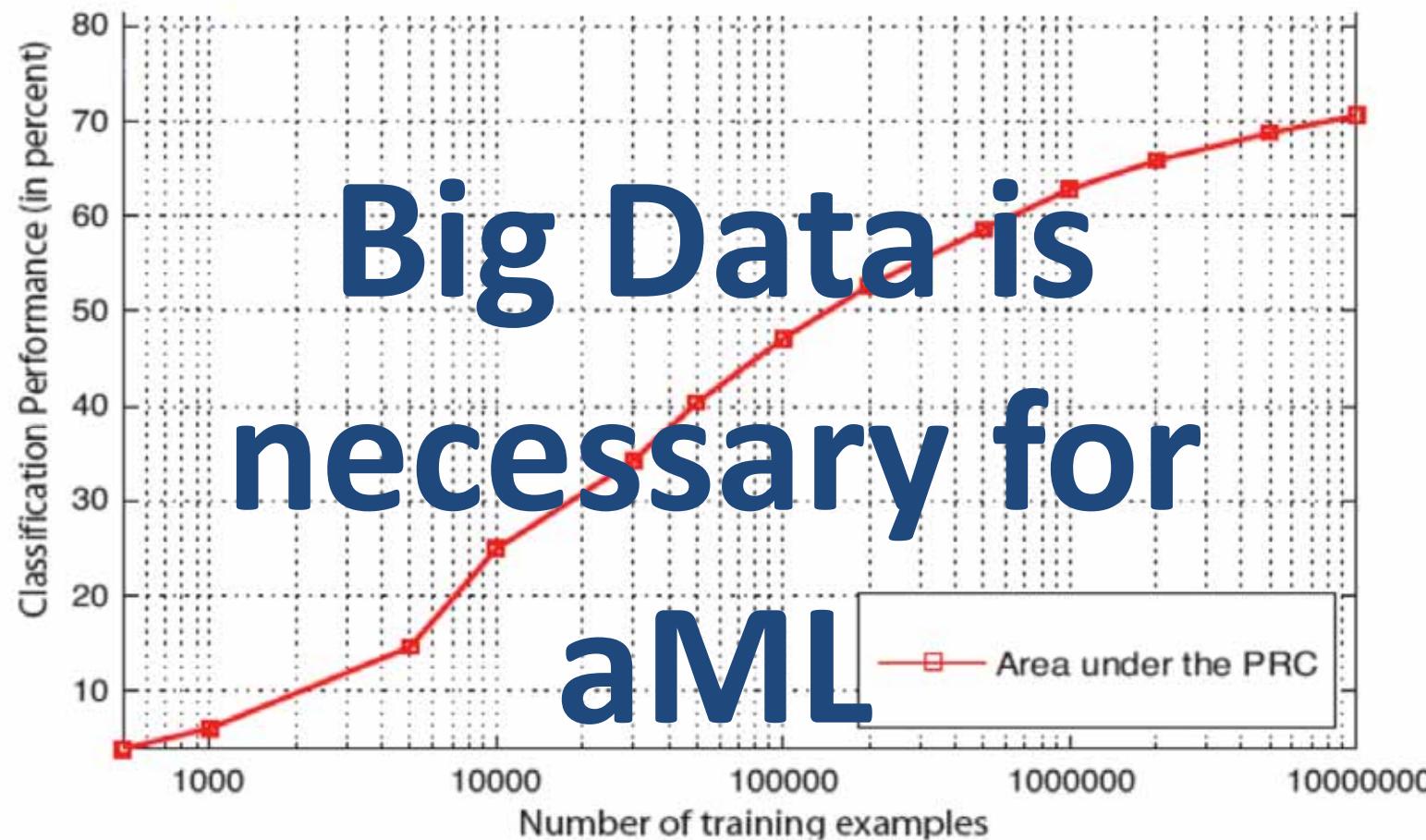


Guizzo, E. 2011. How Google's self-driving car works. IEEE Spectrum Online, 10, 18.

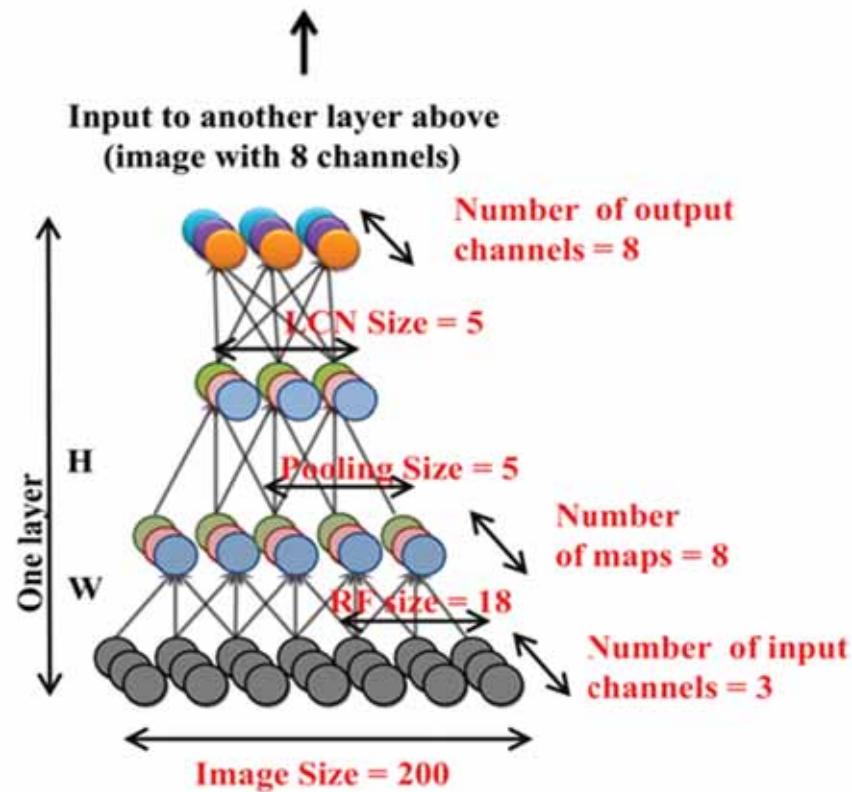
... and thousands of industrial aML applications ...



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.



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.

- 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.

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

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.

Sometimes we need a doctor-in-the-loop



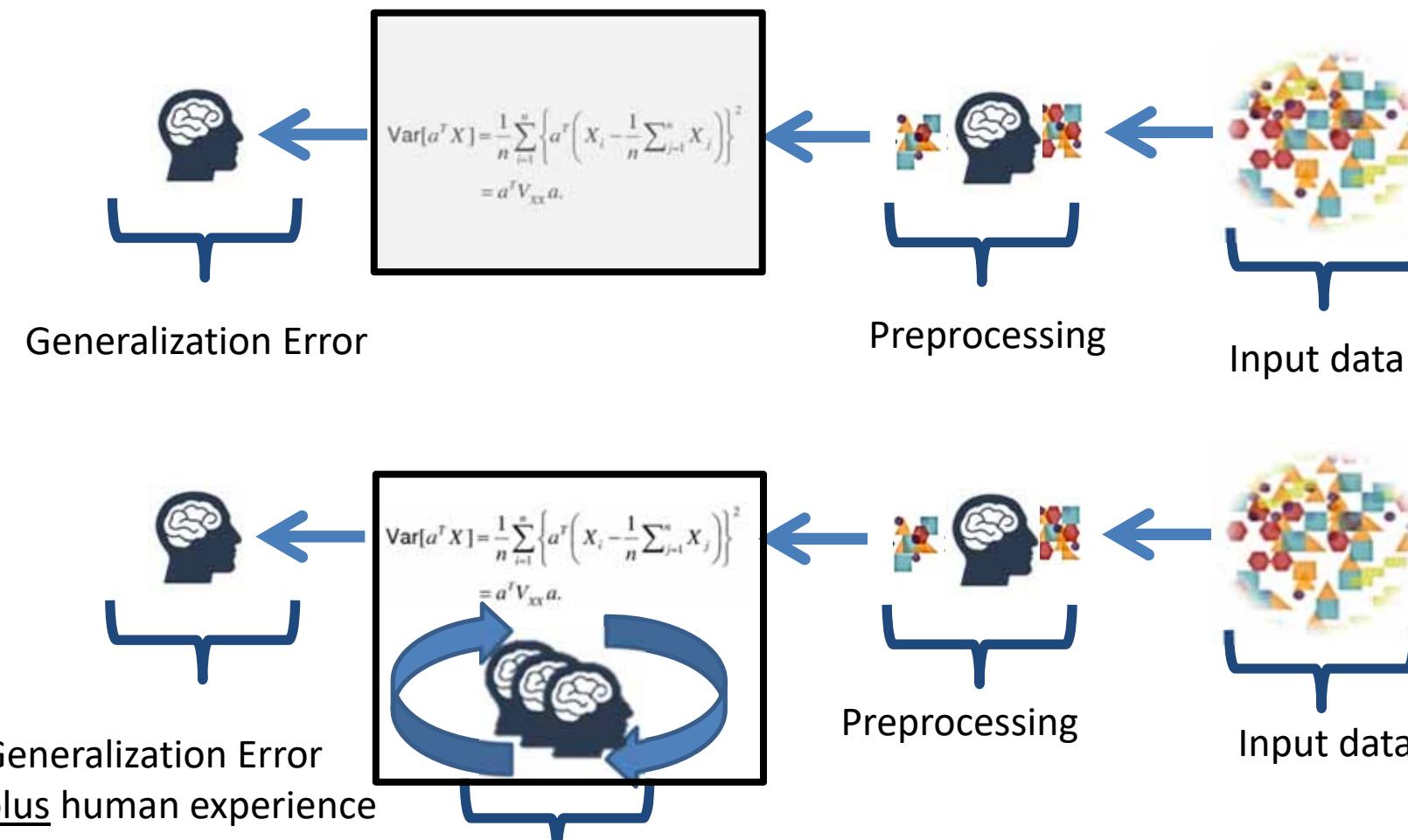
Image Source: 10 Ways Technology is Changing Healthcare <http://newhealthypost.com> Posted online on April 22, 2018

A group of experts-in-the-loop



A crowd of people-in-the-loop





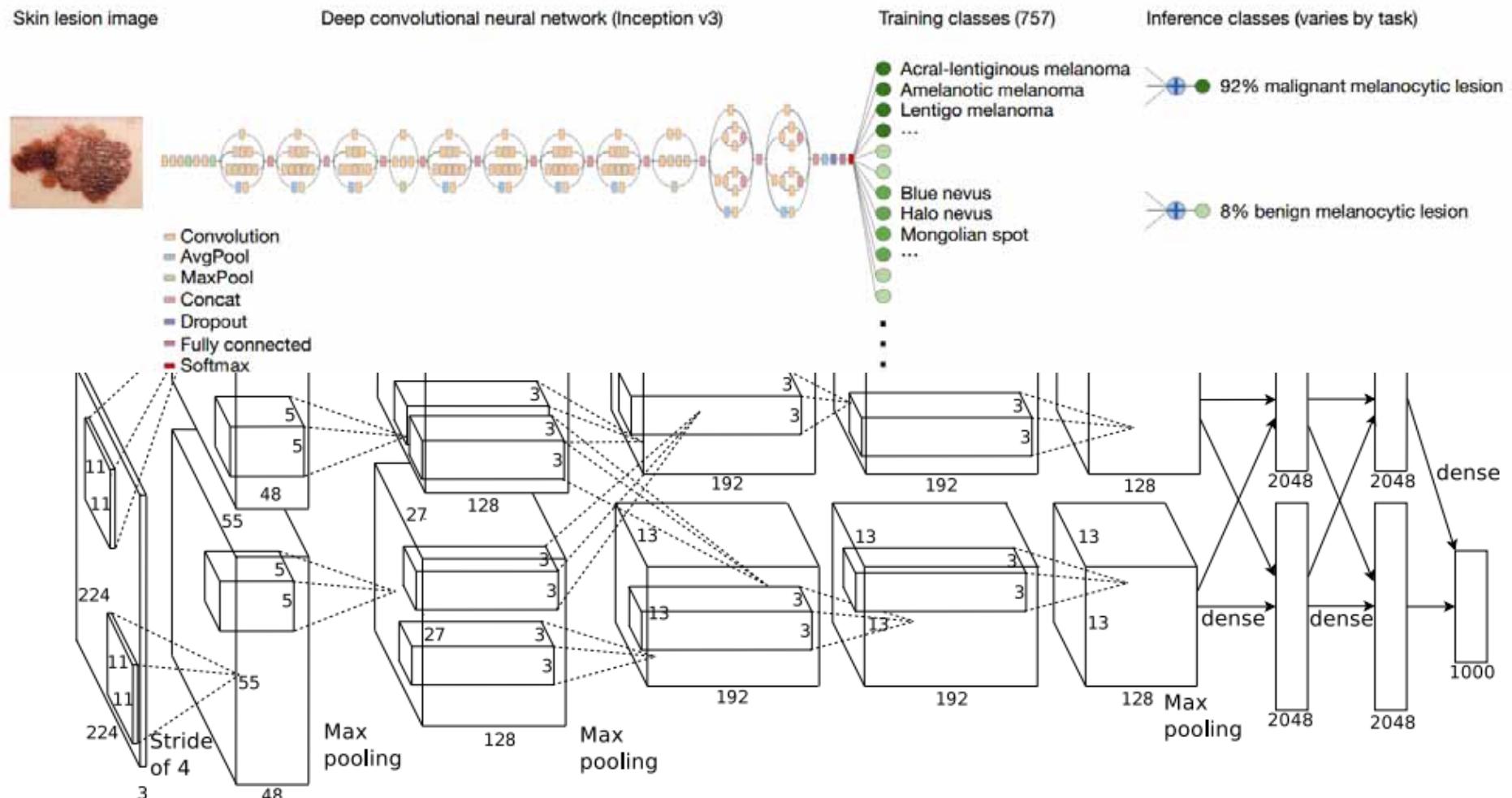
iML = human inspection – bring in human intuition

Andreas Holzinger et al. (2017) A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. arXiv:1708.01104.

06 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.

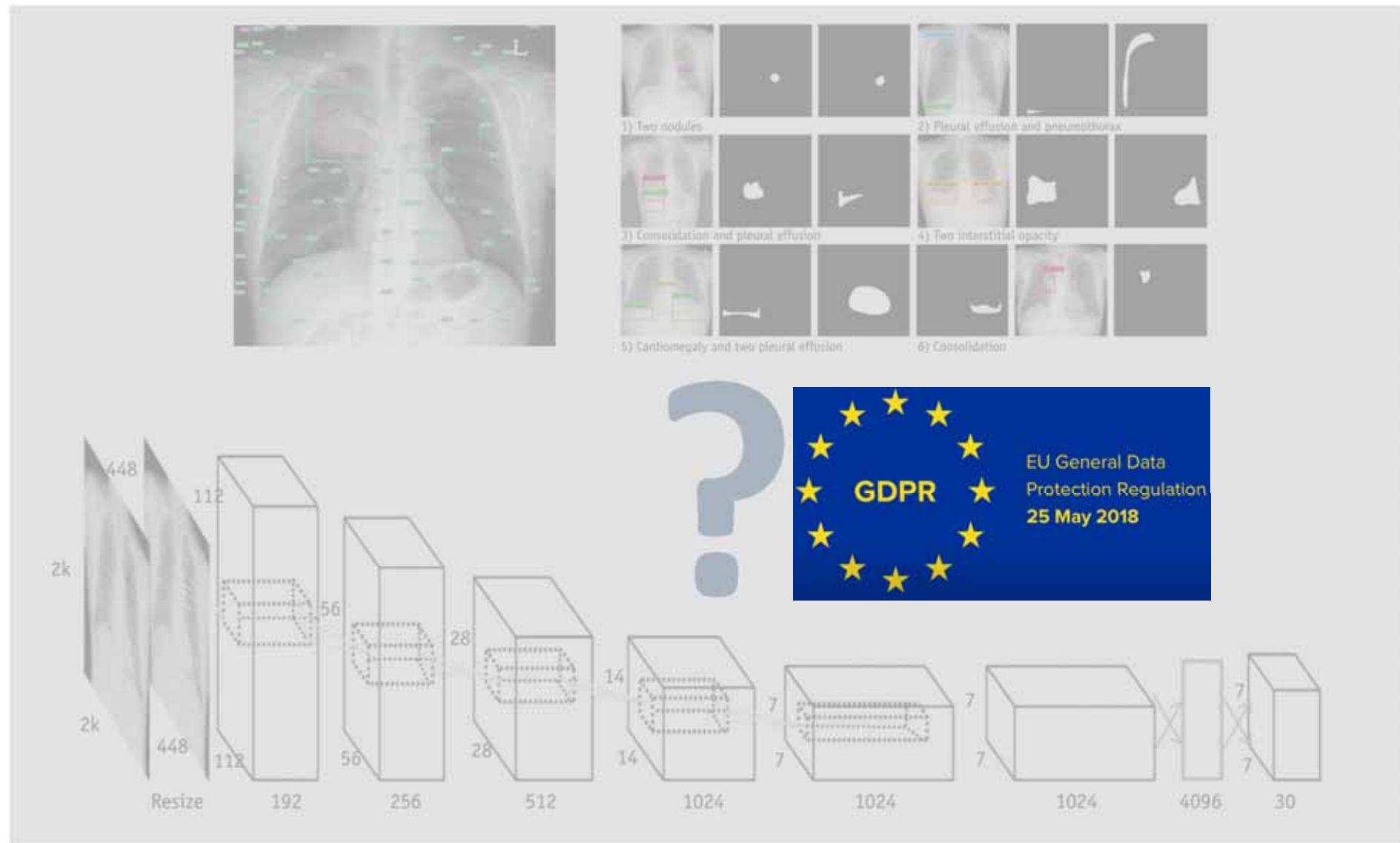


Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

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.

- **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 aspects make “black box” difficult

Example: Adversarial examples



x
“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_x J(\theta, x, y))$
“nematode”
8.2% confidence

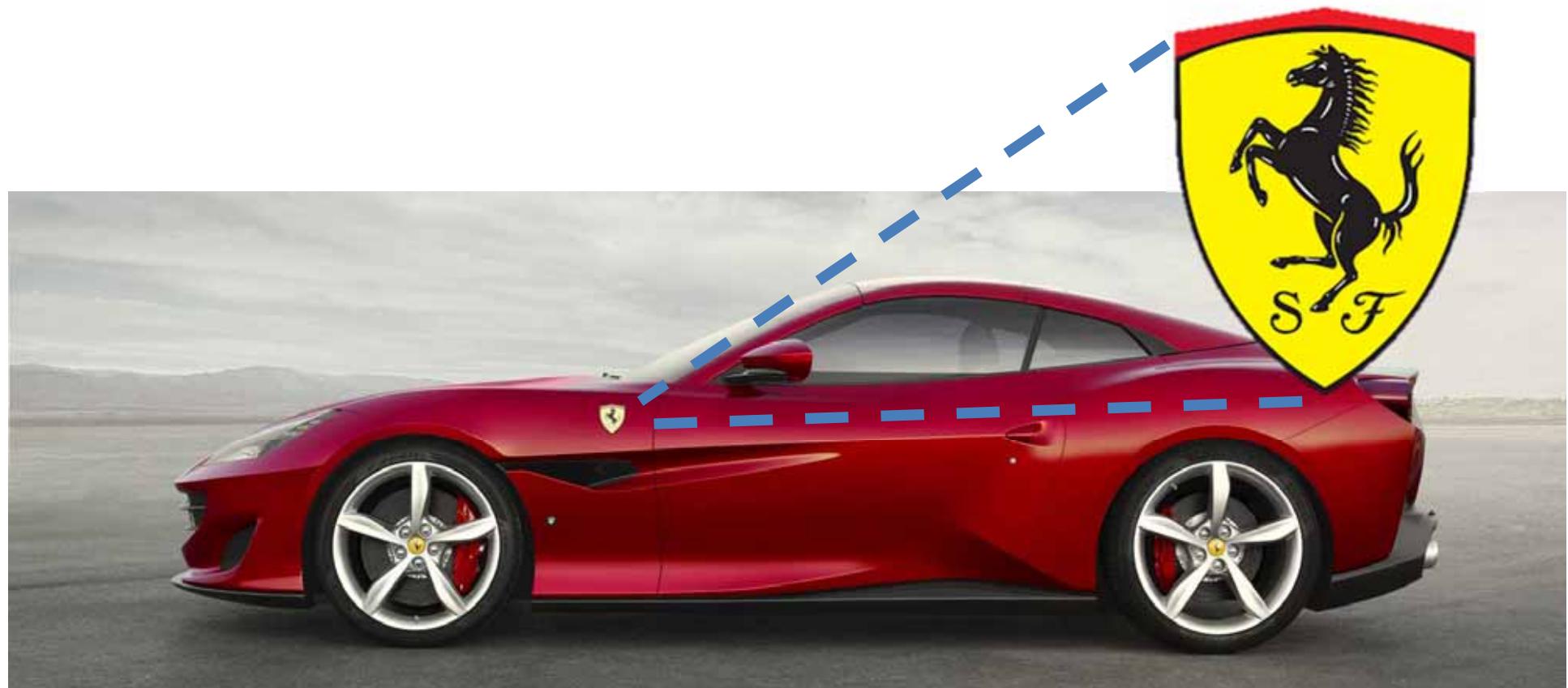
=



$x +$
 $\epsilon \text{sign}(\nabla_x J(\theta, x, y))$
“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>

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



Source: Image is in the public domain

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

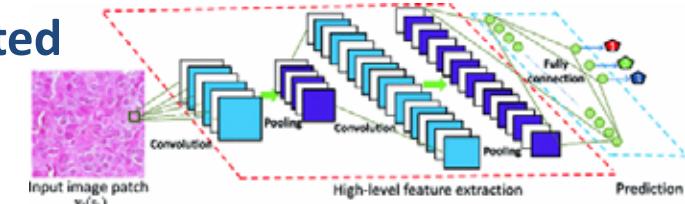


a group of people standing on
top of a beach

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

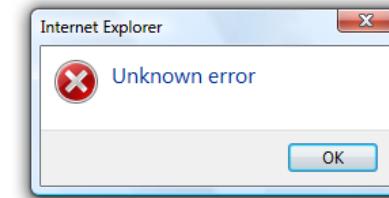
Verify that algorithms/classifiers work as expected

Wrong decisions can be costly and dangerous ...



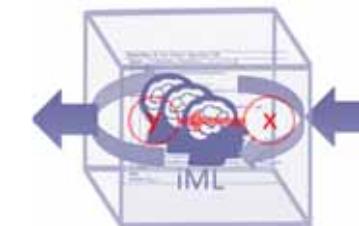
Understanding the weaknesses and errors

Detection of bias – bring in human intuition to know the error ...



Scientific replicability and causality

The “why” is often more important than the prediction ...

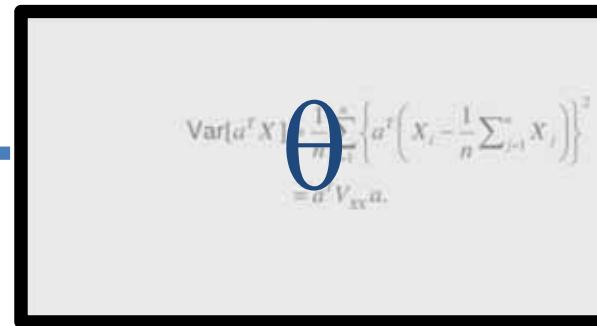


Andreas Holzinger 2018. Explainable AI (ex-AI). Informatik-Spektrum, 41, (2), 138-143, doi:10.1007/s00287-018-1102-5.

Conclusion

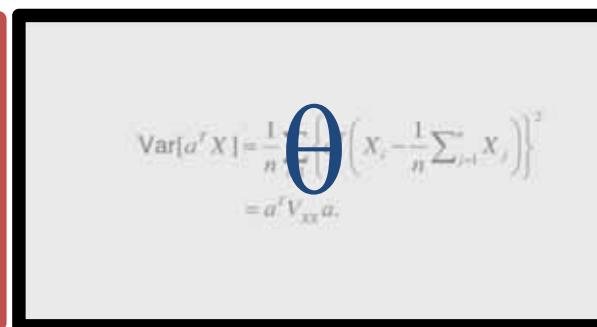
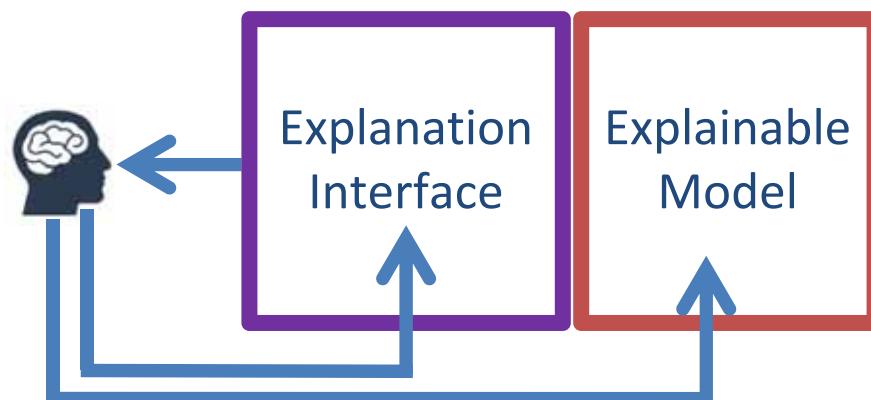
We need effective Human-AI mapping

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



Input data

We contribute to ...



Input data

The domain expert can understand why ...

The domain expert can learn and correct errors ...

The domain expert can re-enact on demand ...



Thank you!

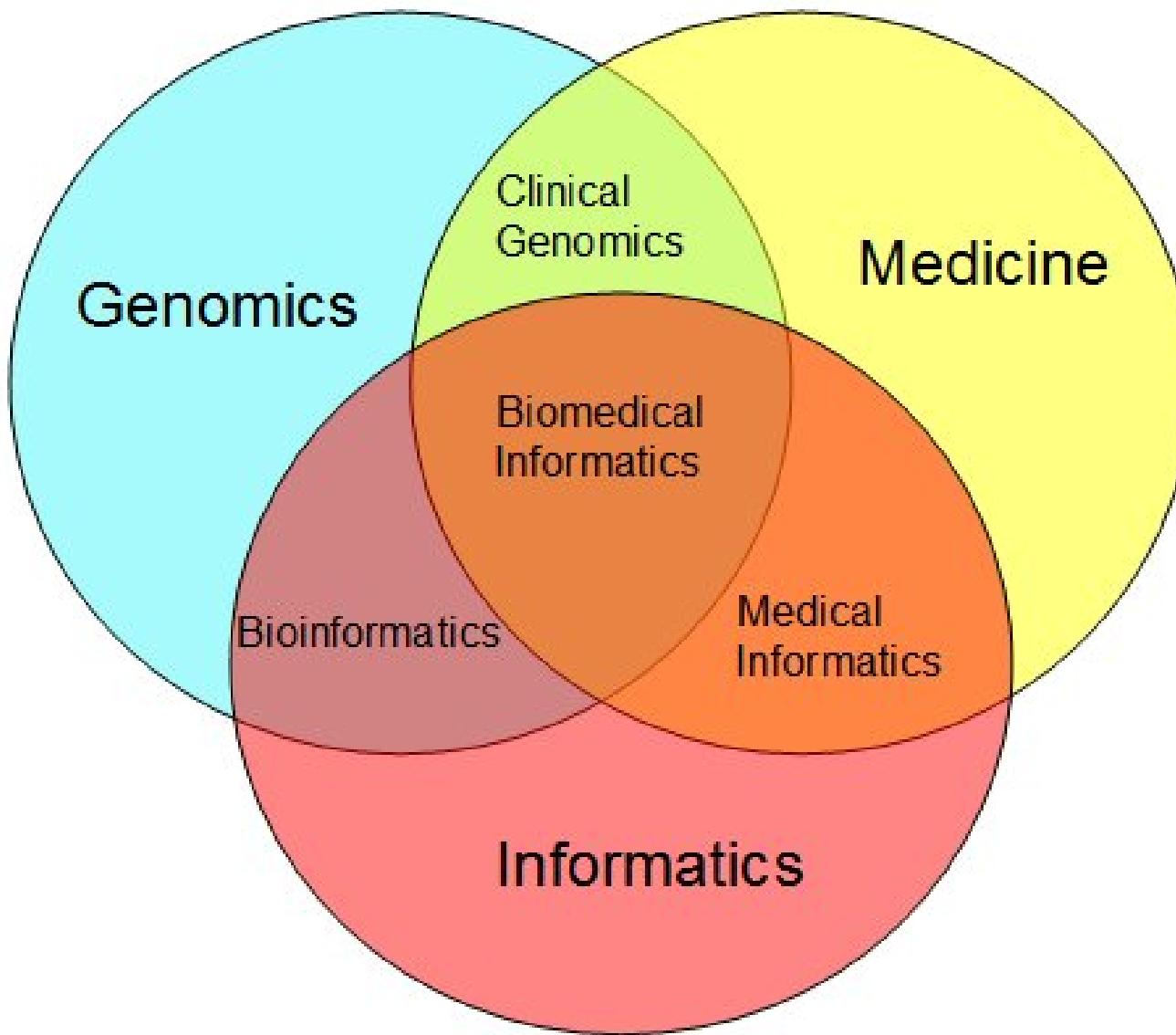


Appendix



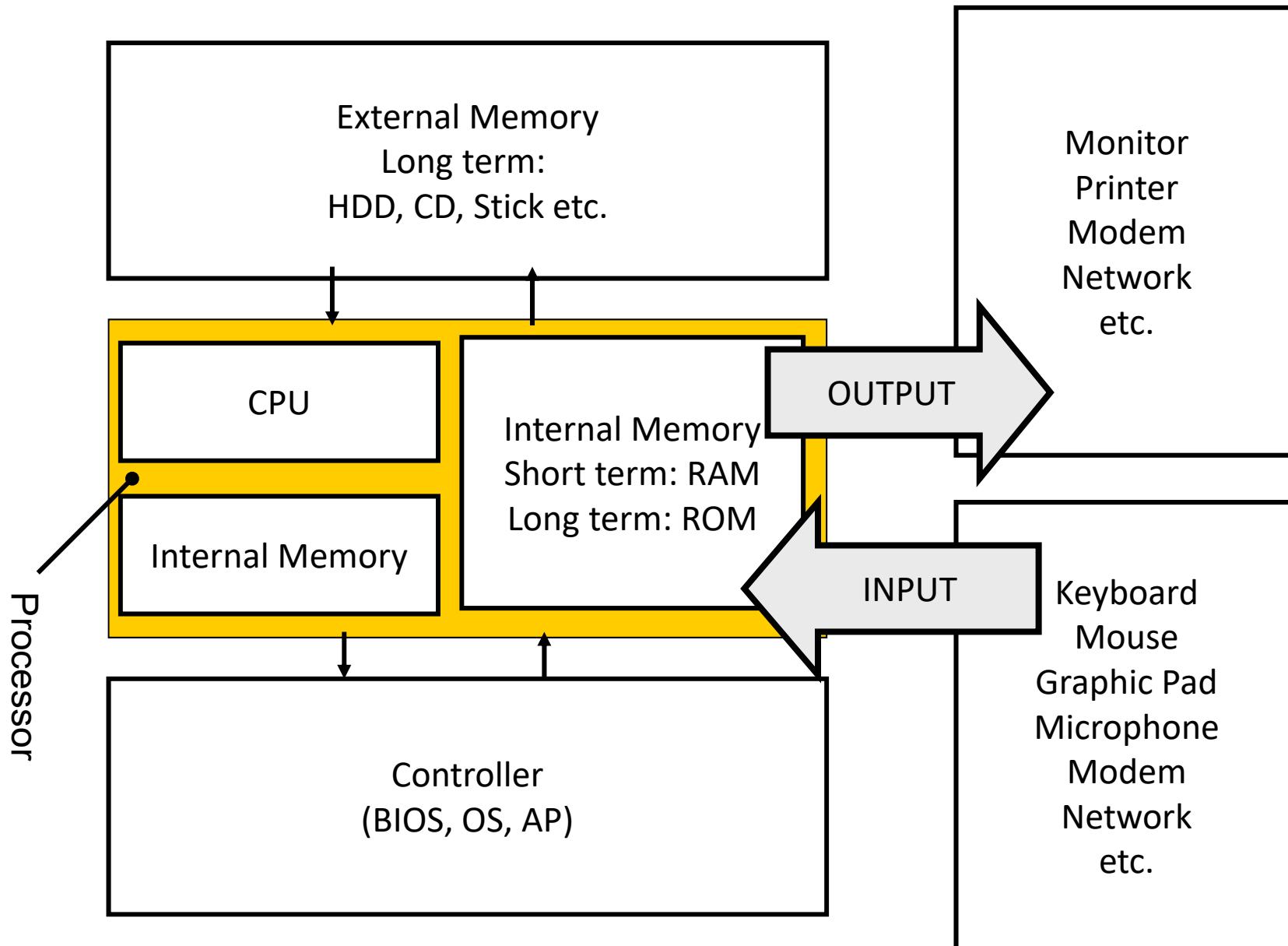
- ***Biomedical informatics (BMI) is the interdisciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific problem solving, and decision making, motivated by efforts to improve human health***

Shortliffe, E. H. (2011). Biomedical Informatics: Defining the Science and its Role in Health Professional Education. In A. Holzinger & K.-M. Simonic (Eds.), *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058* (pp. 711-714). Heidelberg, New York: Springer.

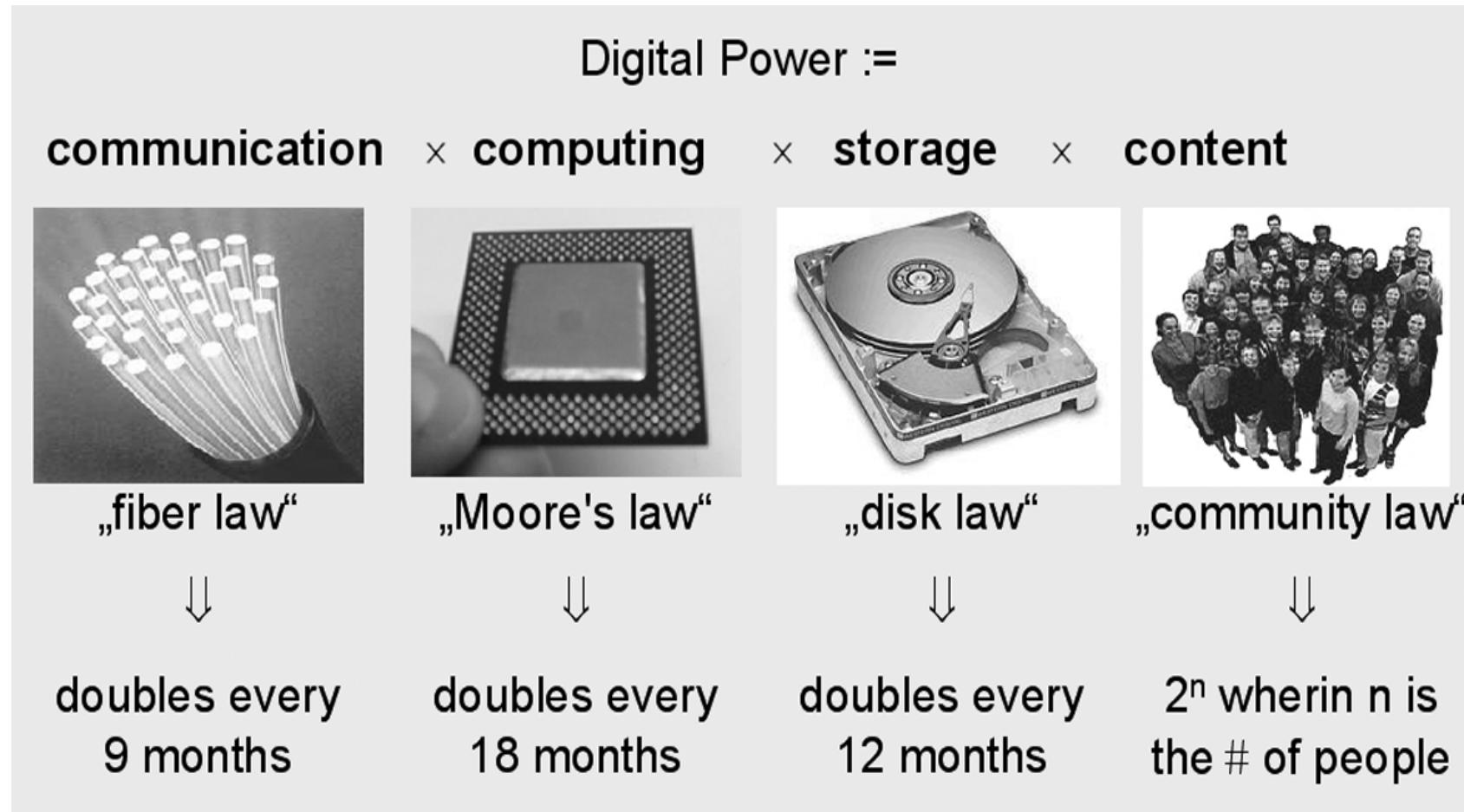


<http://www.bioinformaticslaboratory.nl/twiki/bin/view/BioLab/EducationMIK1-2>

Computer: Von-Neumann Architecture

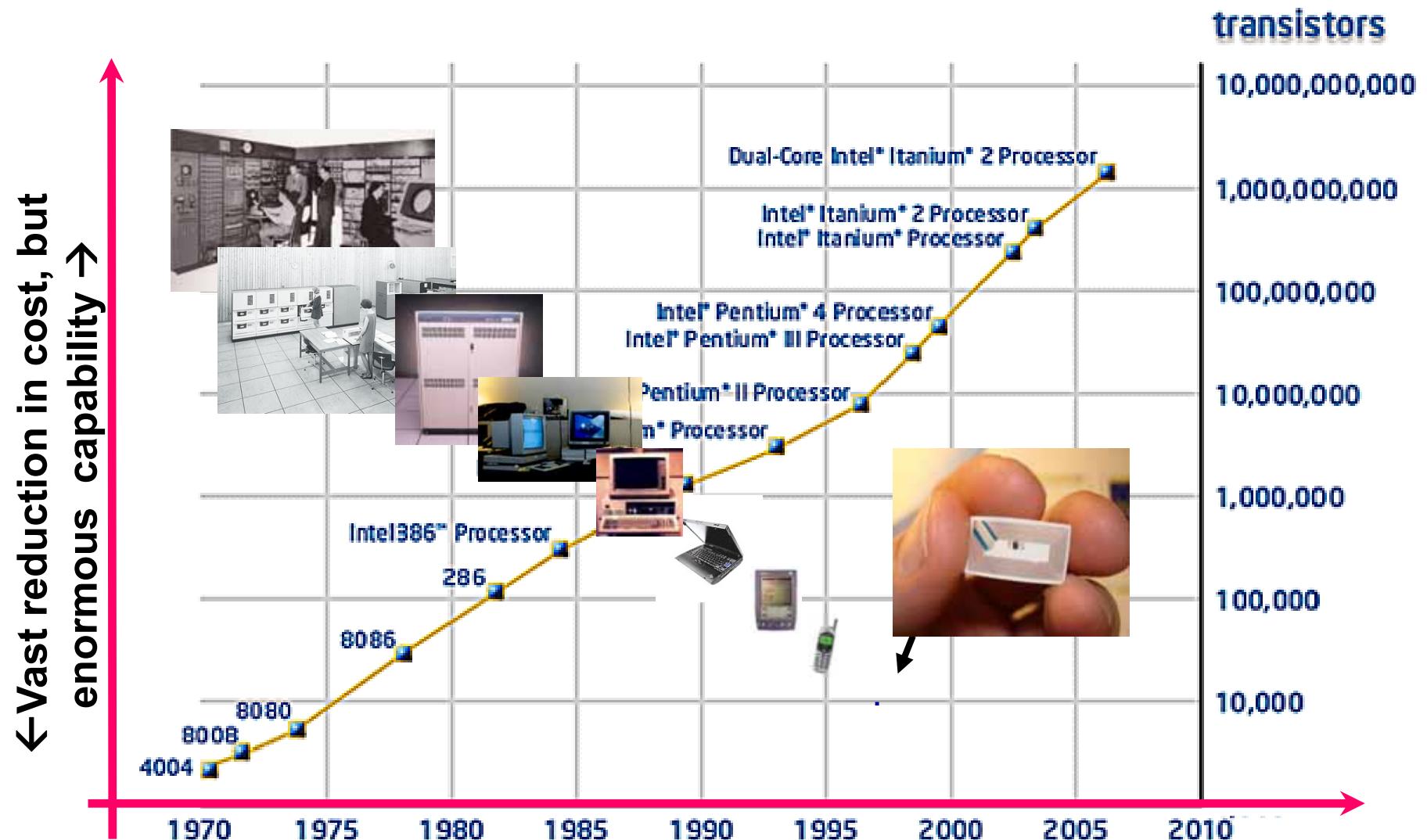


Gordon E. Moore (1965, 1989, 1997)

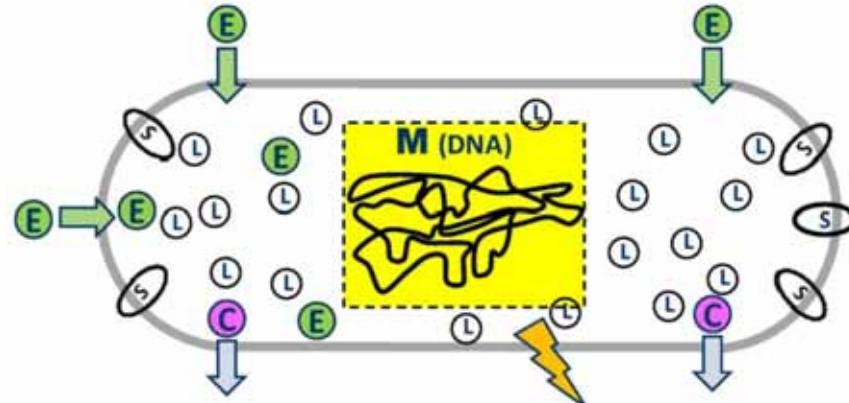


Holzinger, A. 2002. *Basiswissen IT/Informatik Band 1: Informationstechnik. Das Basiswissen für die Informationsgesellschaft des 21. Jahrhunderts*, Wuerzburg, Vogel Buchverlag.

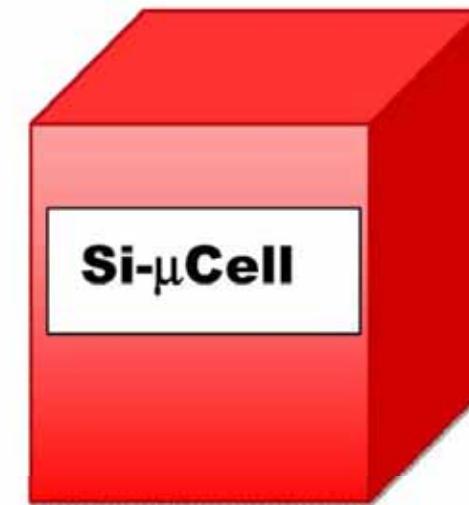
Computer cost/size versus Performance



Cf. with Moore (1965), Holzinger (2002), Scholtz & Consolvo (2004), Intel (2007)



Memory:	10^7 bit
Logic:	$>10^6$ bit
Power:	10^{-13} W
Heat:	10^{-6} W/cm ²
Energy/task*:	10^{-10} J
Task time*:	2400s=40min

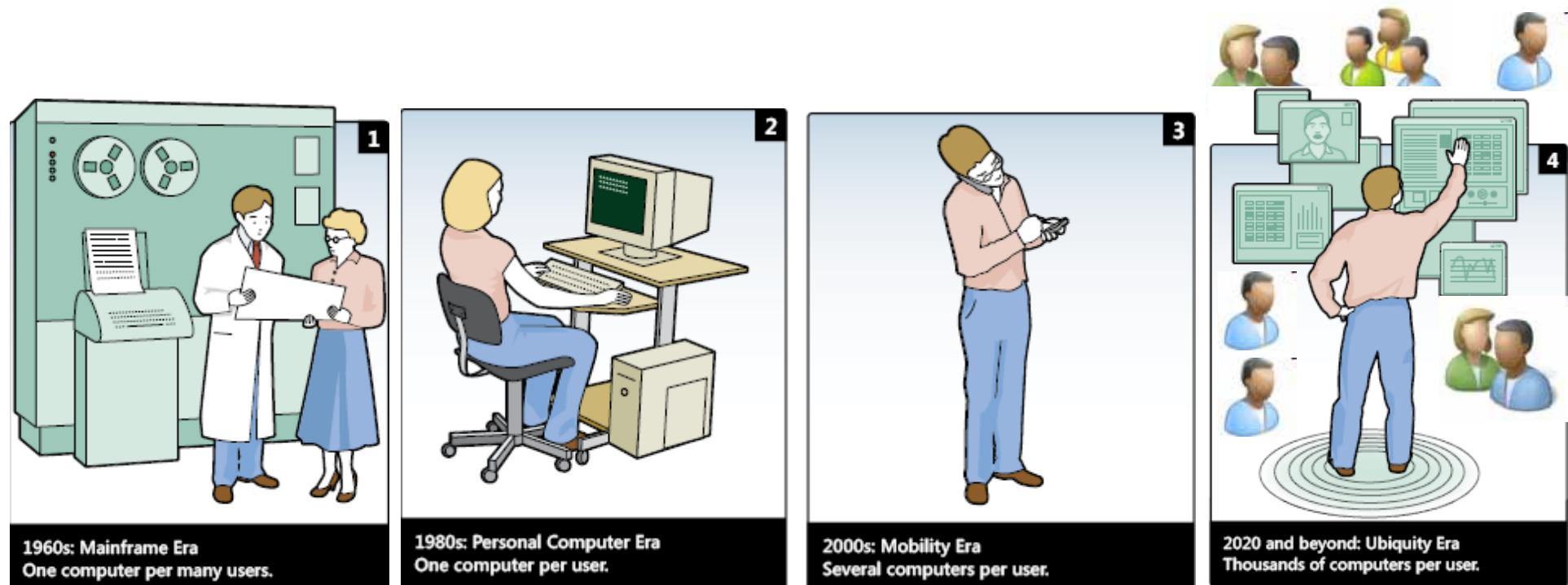


Memory:	$\sim 10^4$ bit
Logic:	$\sim 300\text{--}150,000$ bit
Power:	$\sim 10^{-7}$ W
Heat:	~ 1 W/cm ²
Energy/task*:	$\sim 10^{-2}$ J
Task time*:	510,000 s \sim 6 days

*Equivalent to 10^{11} output bits

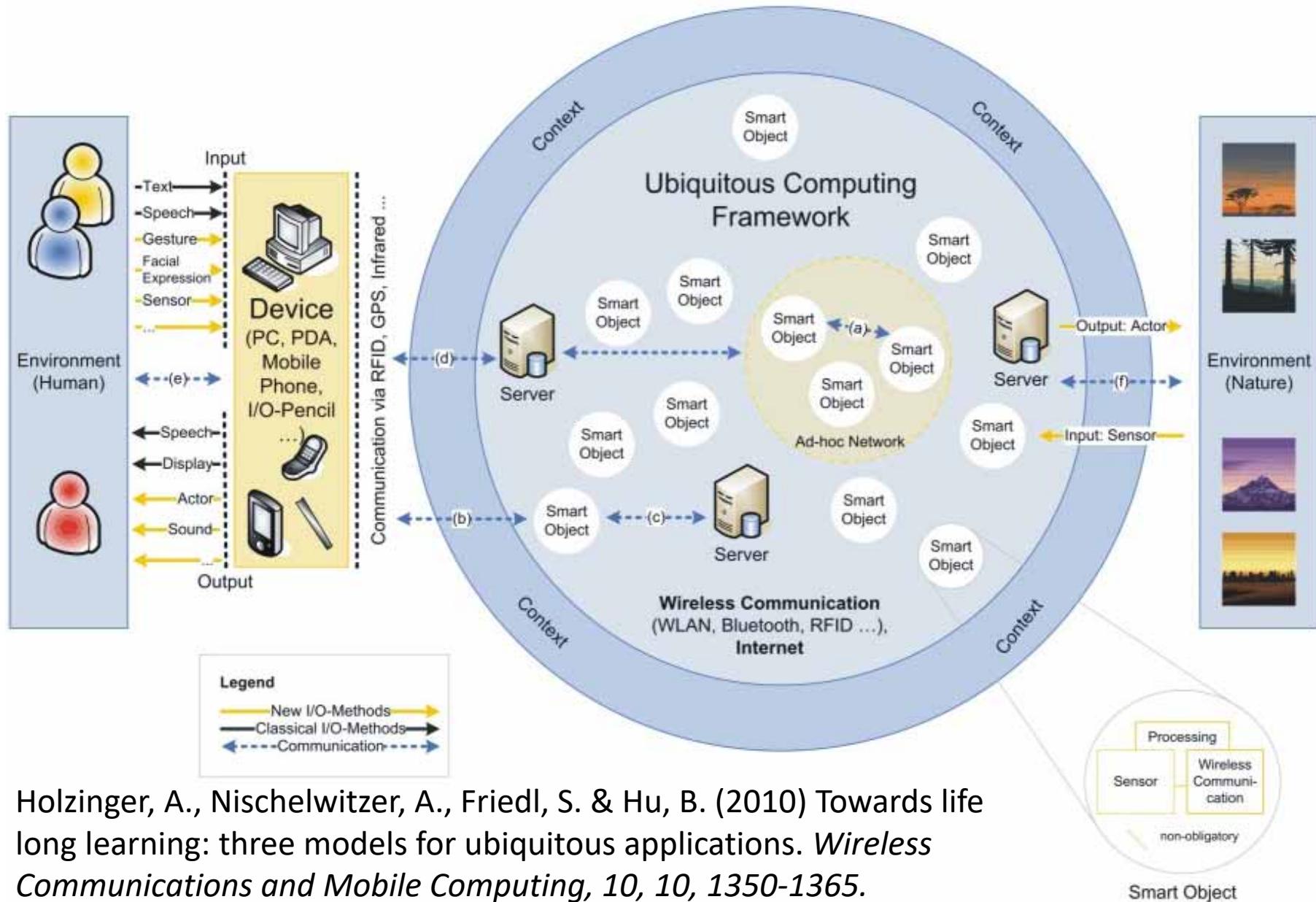
Cavin, R., Lugli, P. & Zhirnov, V. 2012. Science and Engineering Beyond Moore's Law. *Proc. of the IEEE*, 100, 1720-49 (L=Logic-Protein; S=Sensor-Protein; C=Signaling-Molecule, E=Glucose-Energy)

- ... using technology to augment human capabilities for structuring, retrieving and managing information



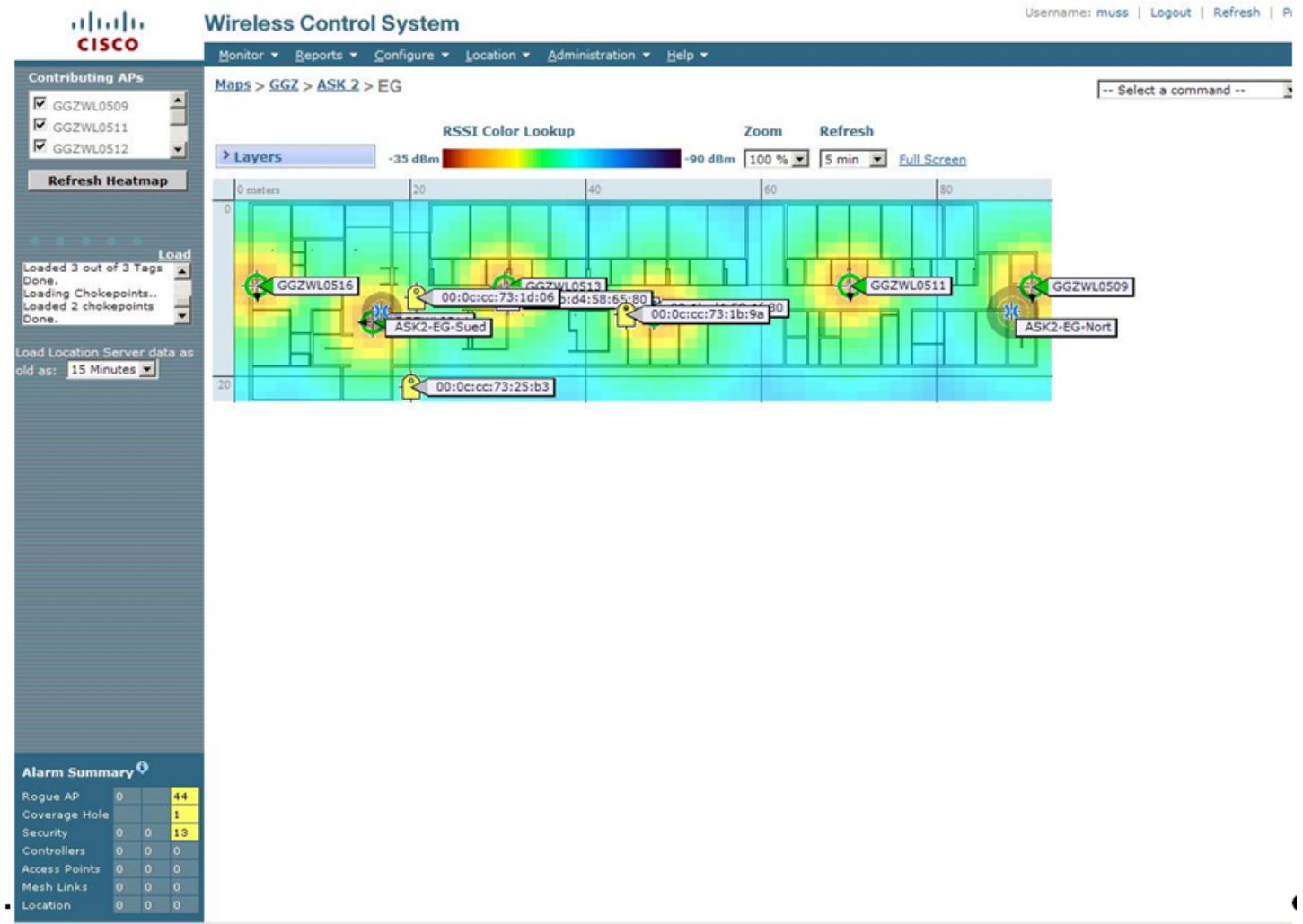
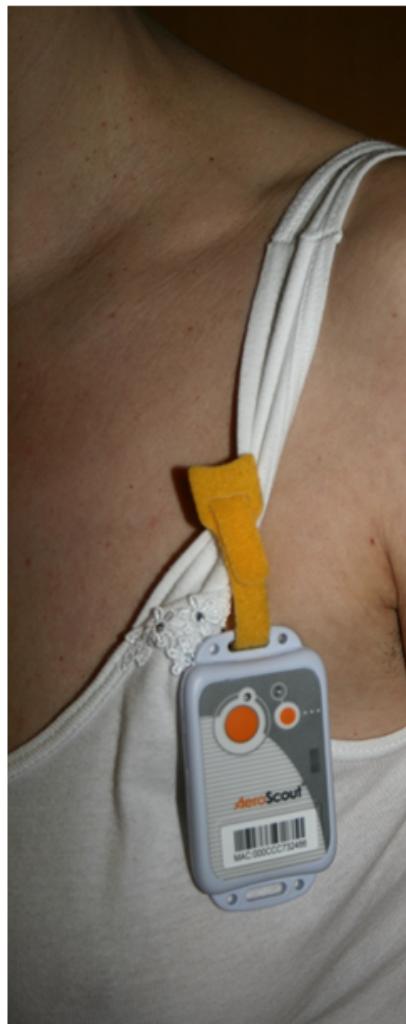
Harper, R., Rodden, T., Rogers, Y. & Sellen, A. (2008) *Being Human: Human-Computer Interaction in the Year 2020*. Cambridge, Microsoft Research.

Ubiquitous Computing – Smart Objects

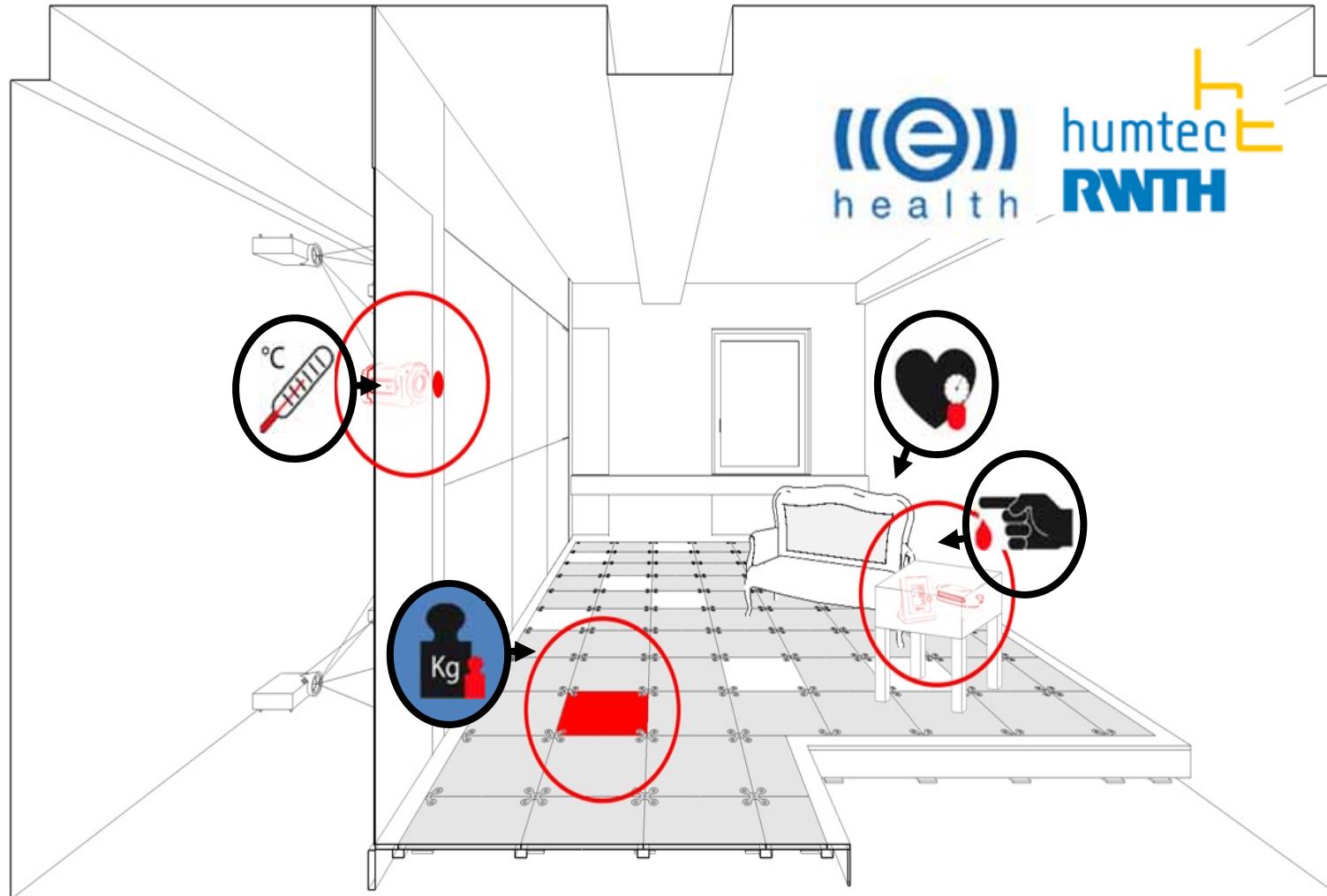


Holzinger, A., Nischelwitzer, A., Friedl, S. & Hu, B. (2010) Towards life long learning: three models for ubiquitous applications. *Wireless Communications and Mobile Computing*, 10, 10, 1350-1365.

Slide 1-34 Example: Pervasive Health Computing

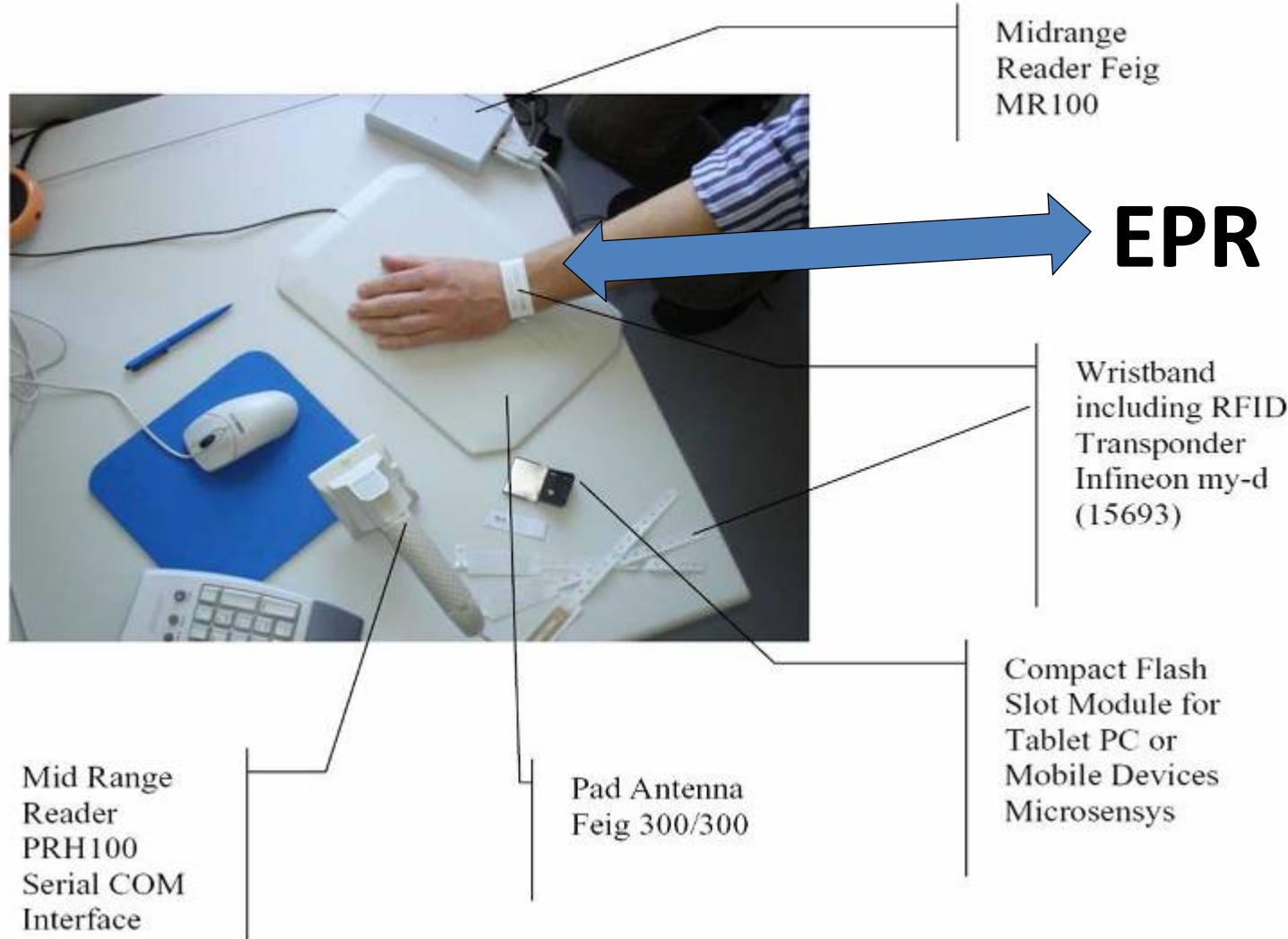


Holzinger, A., Schaupp, K. & Eder-Halbedl, W. (2008) An Investigation on Acceptance of Ubiquitous Devices for the Elderly in an Geriatric Hospital Environment: using the Example of Person Tracking In: *Lecture Notes in Computer Science (LNCS 5105)*. Heidelberg, Springer, 22-29.



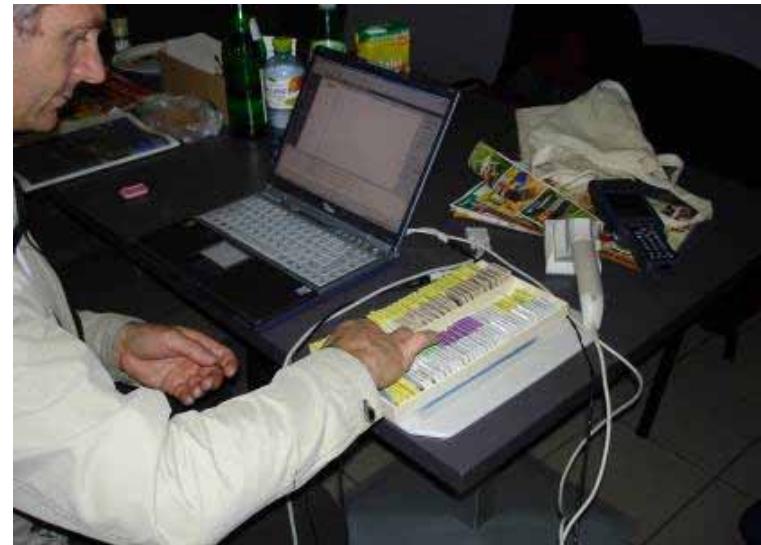
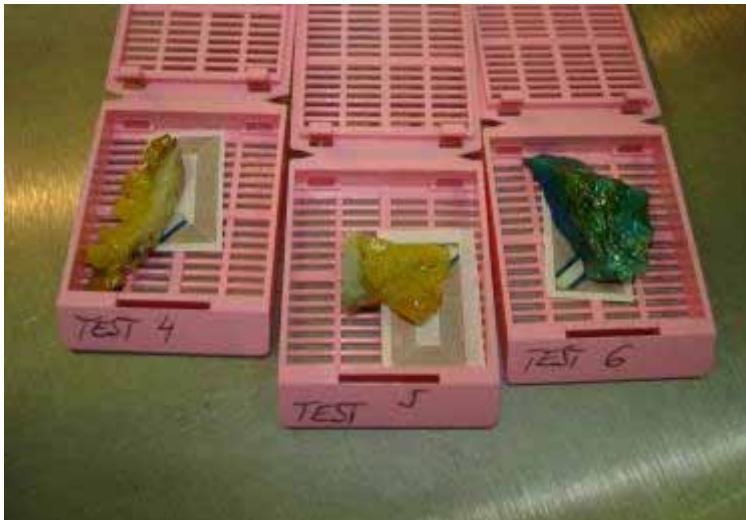
Alagoez, F., Valdez, A. C., Wilkowska, W., Ziefle, M., Dorner, S. & Holzinger, A. (2010) From cloud computing to mobile Internet, from user focus to culture and hedonism: The crucible of mobile health care and Wellness applications. *5th International Conference on Pervasive Computing and Applications (ICPCA). IEEE*, 38-45.

Example Pervasive Computing in the Hospital



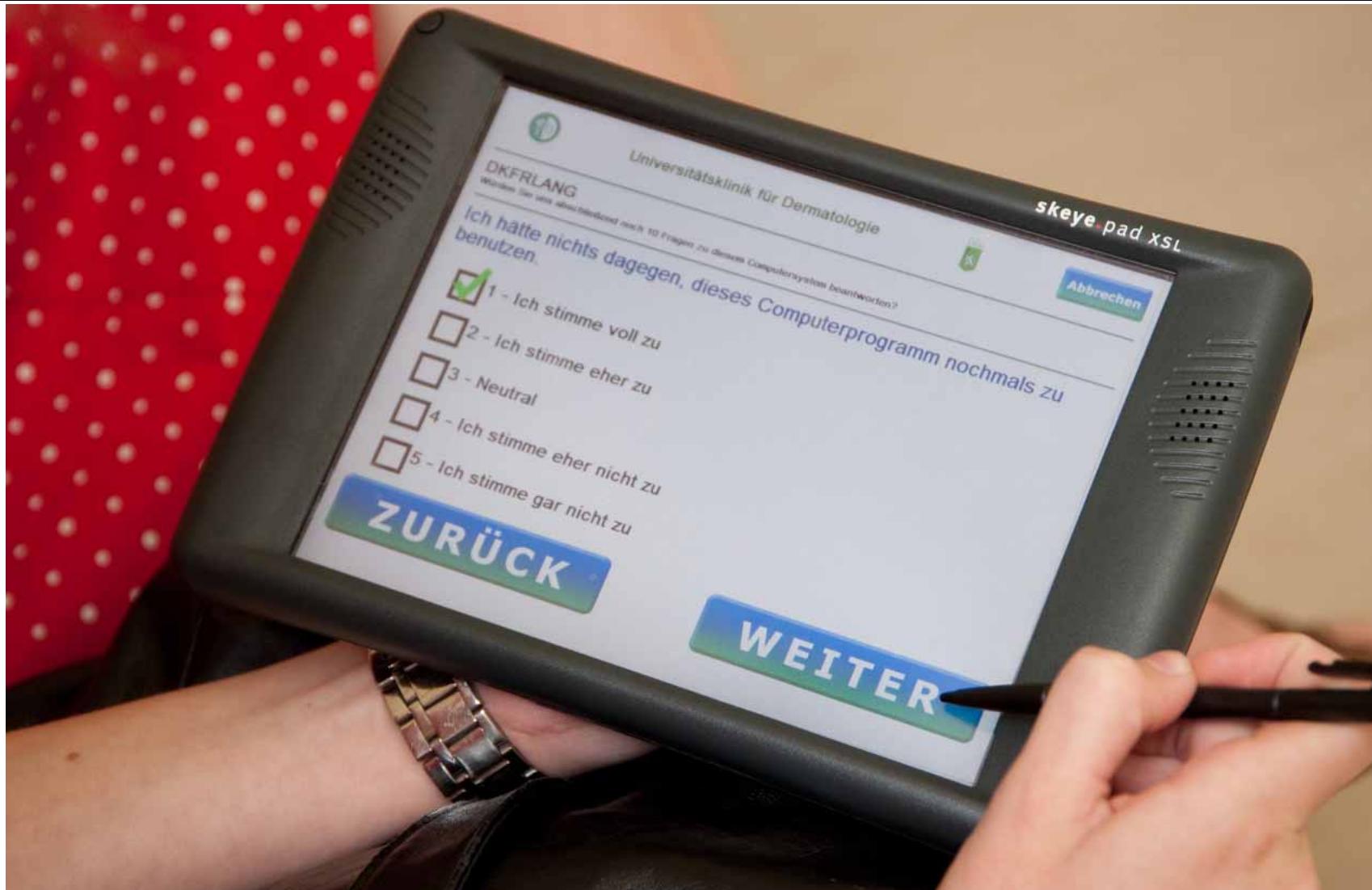
Holzinger, A., Schwaberger, K. & Weitlaner, M. (2005) Ubiquitous Computing for Hospital Applications: RFID-Applications to enable research in Real-Life environments *29th Annual IEEE International Computer Software & Applications Conference (IEEE COMPSAC), 19-20.*

Smart Objects in the pathology



Holzinger et al. (2005)

The medical world is mobile (Mocomed)



Holzinger, A., Kosec, P., Schwantzer, G., Debevc, M., Hofmann-Wellenhof, R. & Frühauf, J. 2011. Design and Development of a Mobile Computer Application to Reengineer Workflows in the Hospital and the Methodology to evaluate its Effectiveness. *Journal of Biomedical Informatics*, 44, 968-977.

1970 Turning Knowledge into Data

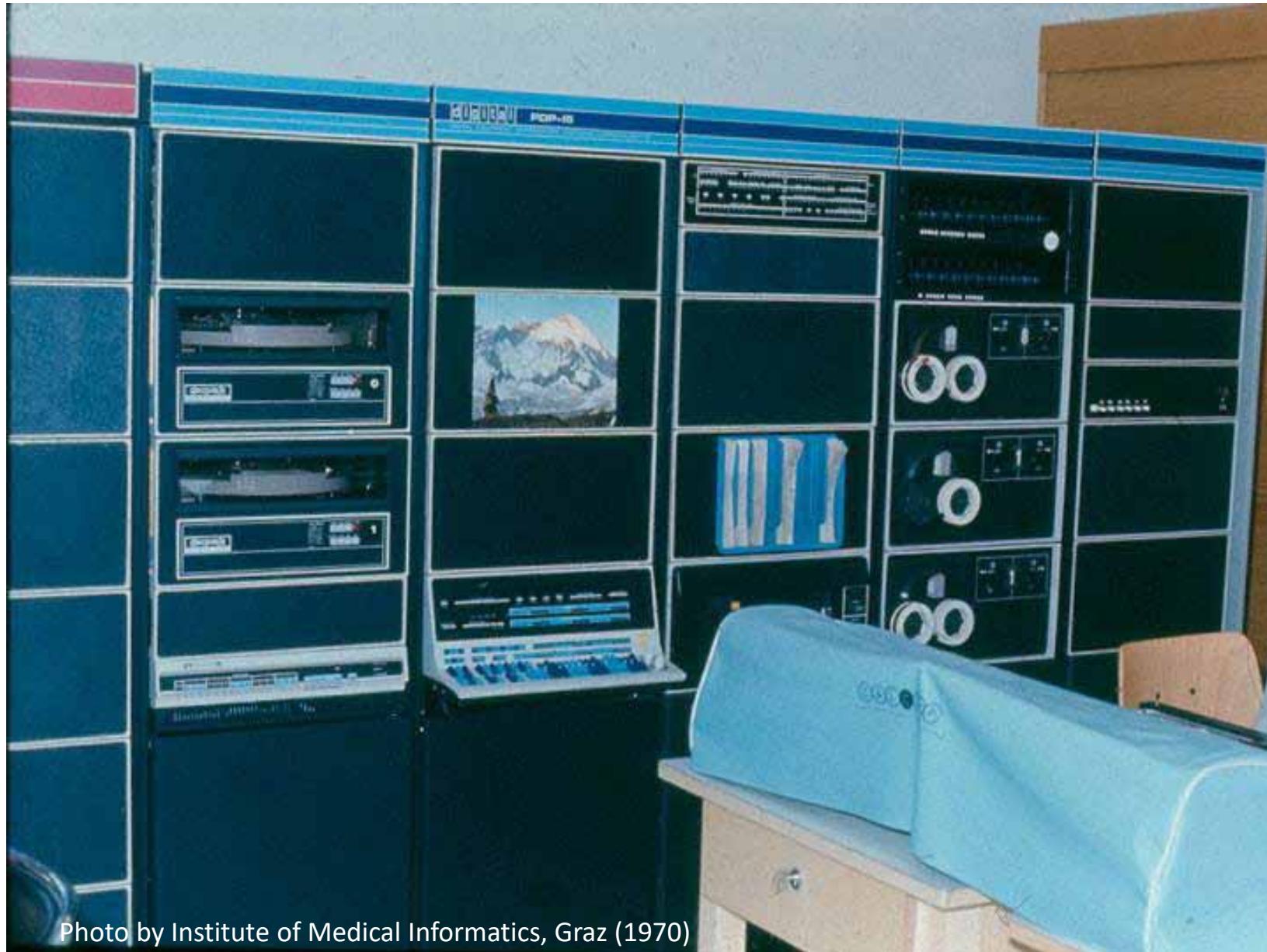


Photo by Institute of Medical Informatics, Graz (1970)