Artificial Intelligence for Health Care

Georg Dorffner
Section for Artificial Intelligence & Decision Support, CeMSIIS

Medical University of Vienna

Artificial Intelligence (AI) – a historical perspective

- Since the 1950s: „Process that require intelligence when performed by humans“
- Many sub-areas, such as Reasoning, Knowledge Representation, Natural Language Processing, Machine Learning (ML)
- The latter (ML) entered the scene pretty late, Neural Networks (since the 1980s, although much older) were initially not really considered AI
- Symbolic vs. subsymbolic, grounding, situatedness, etc.

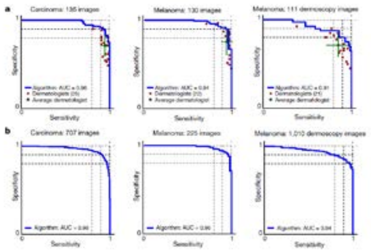
- Today: AI is almost synonymous with ML
  (and almost synonymous with deep learning / neural networks)
- Is this the final breakthrough for AI?
- And will it revolutionize healthcare?

AI for Healthcare

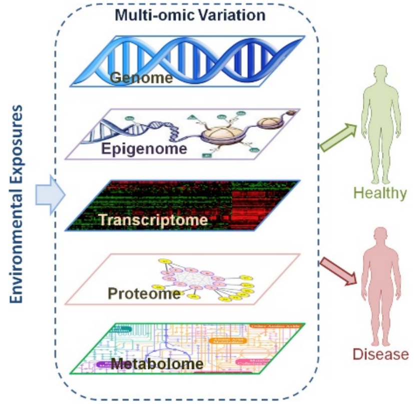
Health Monitoring

https://breathe.ersjournals.com/content/13/2/e27

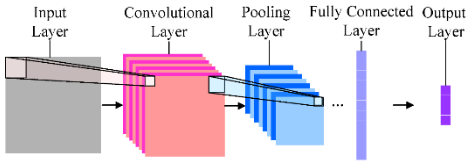
Diagnostics



Personalized Medicine



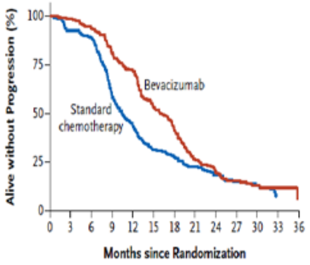
AI



Drug discovery

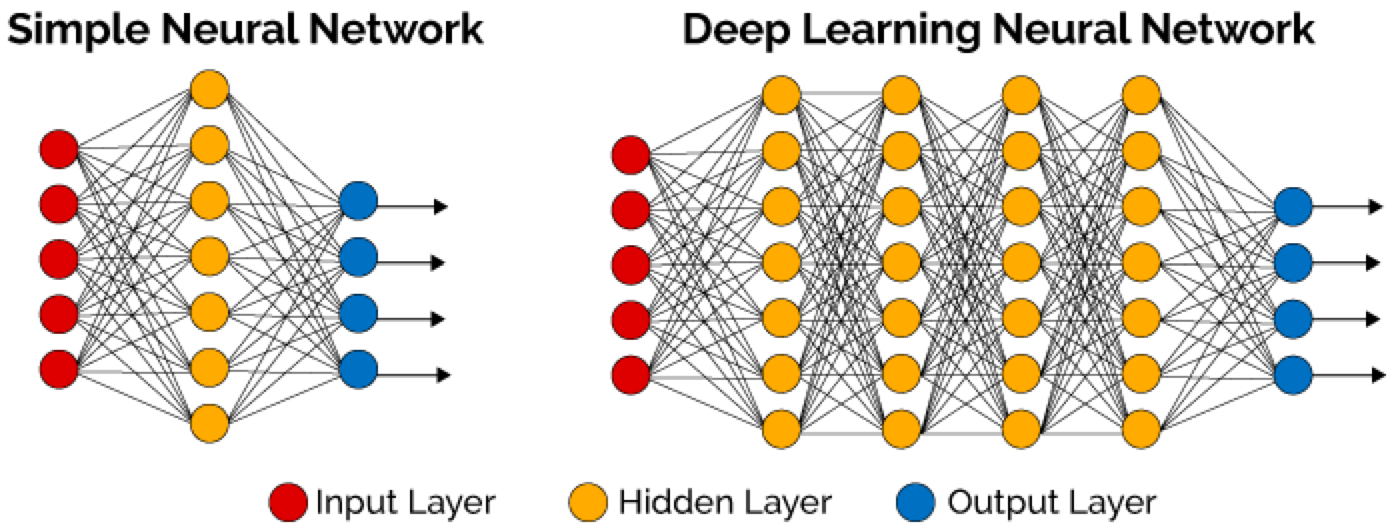
knowledge extraction from text

Prediction



The main ideas behind DL (1)

- A multilayer network learns new features (a linear combination plus the sigmoid)
- These are such that the final classification is linear (perceptron at the output)
- Idea: add higher-level features (a linear combination of first level features, plus sigmoid; and so on)

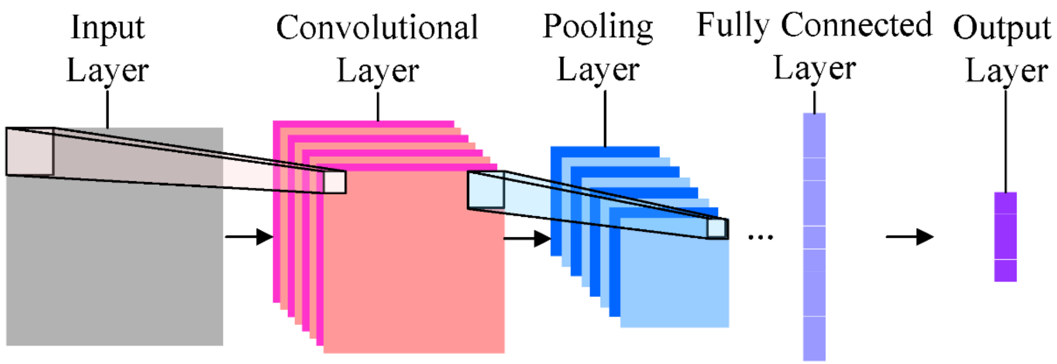


https://hackernoon.com/log-analytics-with-deep-learning-and-machine-learning-20a1891ff70e

The main ideas behind DL (2)

- The first idea particularly makes sense for complex inputs like images or signals
- Problem: Features can occur at many places/times (need to be translation invariant)
- Idea: Have units focus on small parts of the input, but move that part over the entire input
- Akin to convolution (or filtering)

→ **„convolutional neural network“**

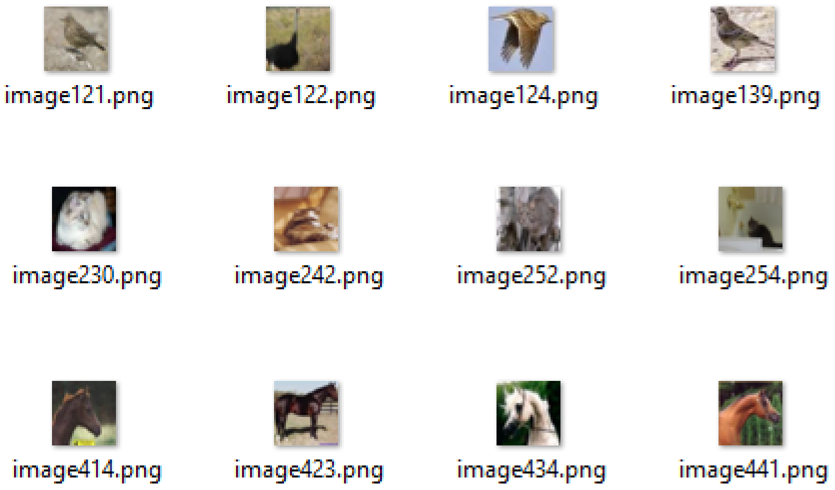


http://www.mdpi.com/2078-2489/7/4/61

The main ideas behind DL (3)

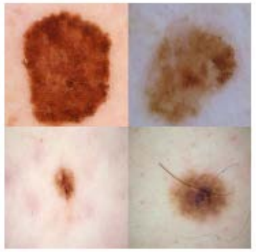
- Visual recognition, or recognition based on signals, share many features on the lowest levels („universal features“)
- A pre-trained network only needs to be adapted to specific domain-specific features (on higher levels)

→ **„transfer learning“**



Pre-training with general images

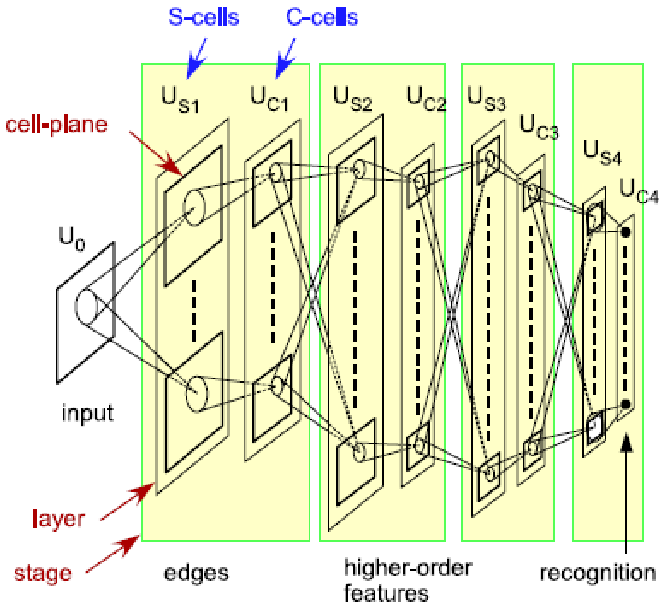
Fine-tuning to specific images

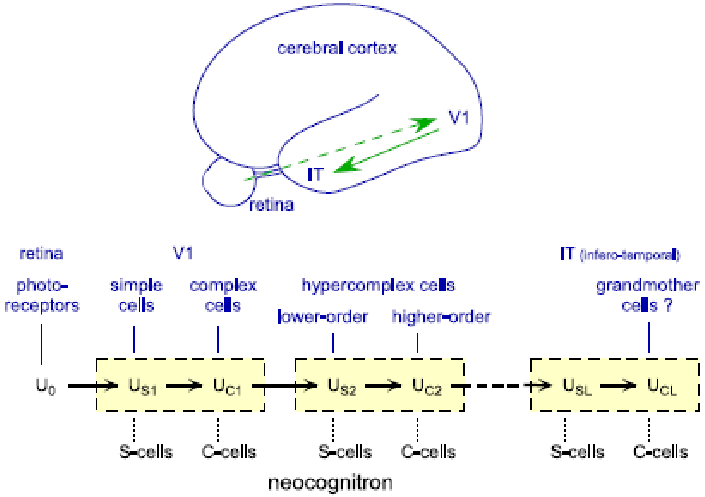


Esteva et al., Nature, 2017

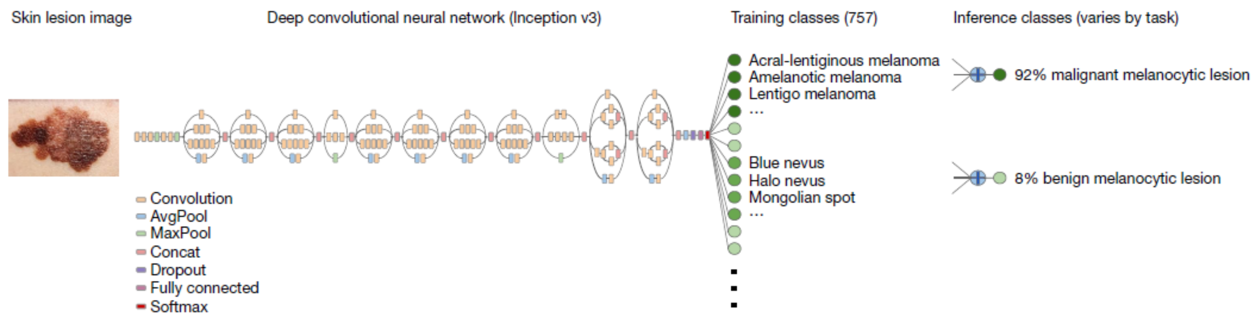
CNN and the visual system (brain research)

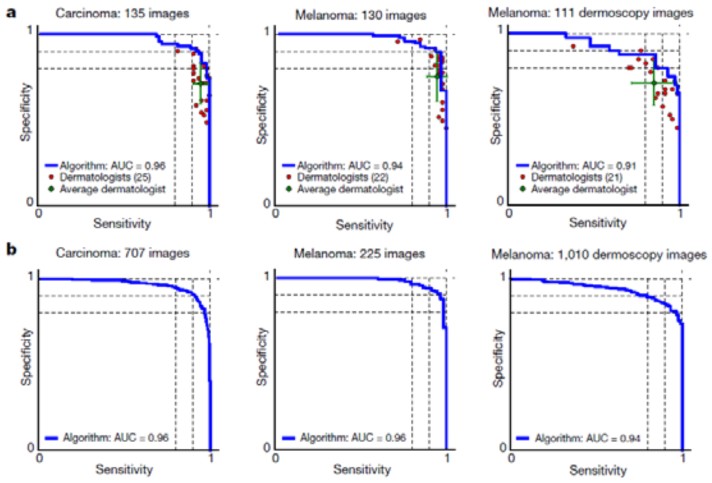
- Neocognitron (Fukushima, 1980)





Medical Example 1: Diagnostics in Dermatology





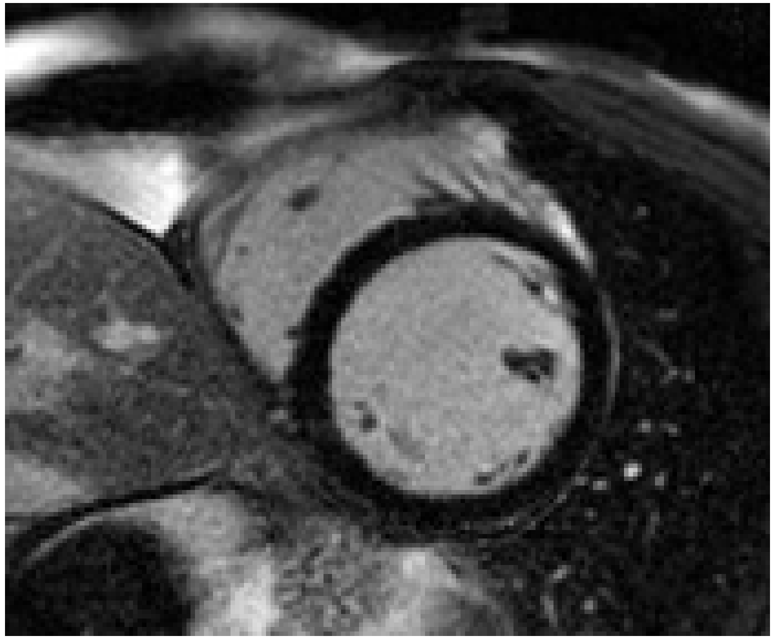
Esteva et al.
Nature, 2017

- Achieves (better than) human-level performance in recognizing melanoma

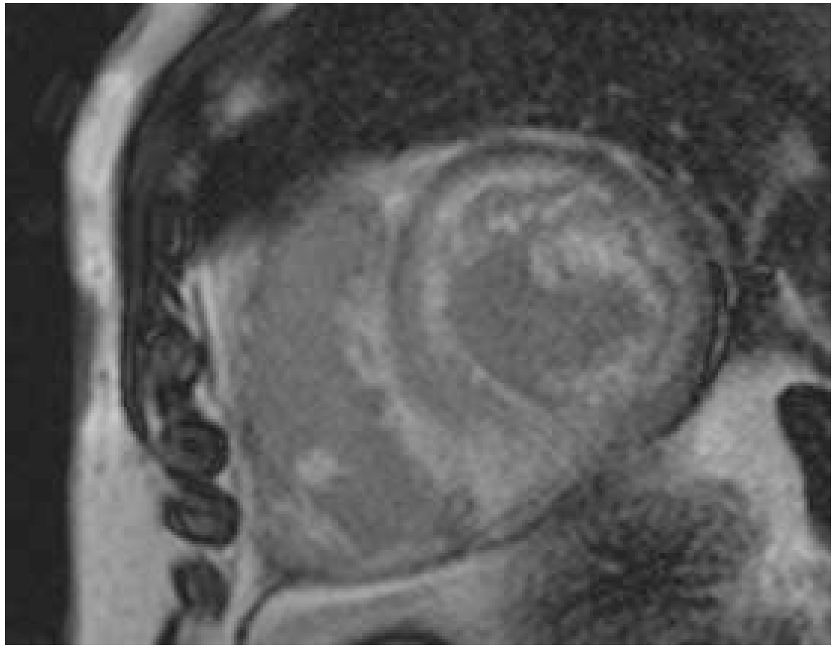
Example 2: Diagnostics in Cardiac MRI

- Backbone in the diagnostic work-up of patients with suspected cardiac amyloidosis
- Cons: Early disease states difficult to assess, lack of experience in non-specialized centers

Healthy



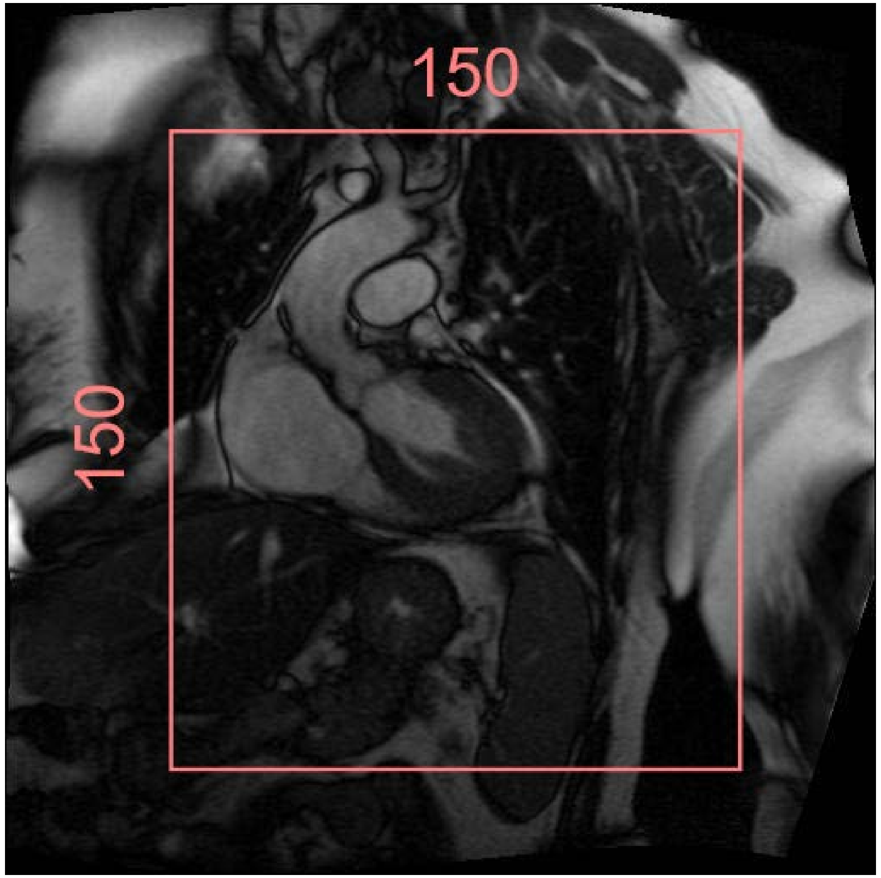
Cardiac amyloidosis



Cooperation mit Diana Bonderman, Julia Mascherbauer et al., MedUniv Vienna

Dataset description

- 510 patients with cardiac disorders admitted to AKH-Wien (2008-2019)
  - 82 (biopsy proven amyloidosis; true positives)
  - 428 (hard (healthy) + soft
    (other cardiac conditions) negatives)
- Three MRI imaging protocols
  - Late gadolinium enhancement (LGE; static)
    - 16,343 images (2,598 positives)
  - T1 mapping (static)
    - 30,630 images (7,649 positives)
  - CINE retro images (dynamic)
    - 309,702 images (53,878 positives)



Results: feature extraction (pre-trained networks) + logistic regression

Agibetov et al., 2018/19 (work in progess)

Imaging modality	**pretrained CNN**	**Image classification**	**Patient classification**
LGE	VGG16	0.86 +/- 0.06	0.95 +/- 0.05
	ResNet50		
	DenseNet121	0.86 +/- 0.08	0.94 +/- 0.07
MOLLI	VGG16	0.83 +/- 0.04	0.91 +/- 0.03
CINE	VGG16	0.8 +/- 0.09	0.9 +/- 0.1

10-fold CV; ROC AUC mean +/- SD. Patient classification – mean prediction over all images of the patient

Results: Fine-tune pre-trained networks

Imaging modality	pretrained CNN	Image classification	Patient classification
LGE	VGG16	0.92 +/- 0.05	0.96 +/- 0.04
	ResNet50		
	DenseNet121	0.89 +/- 0.06	0.95 +/- 0.06
MOLLI	VGG16	0.9 +/- 0.06	0.92 +/- 0.06
	DenseNet121	0.86 +/- 0.05	0.92 +/- 0.05
CINE	VGG16	0.87 +/- 0.09	0.9 +/- 0.06

10-fold CV; ROC AUC mean +/- SD. Patient classification – mean prediction over all images of the patient

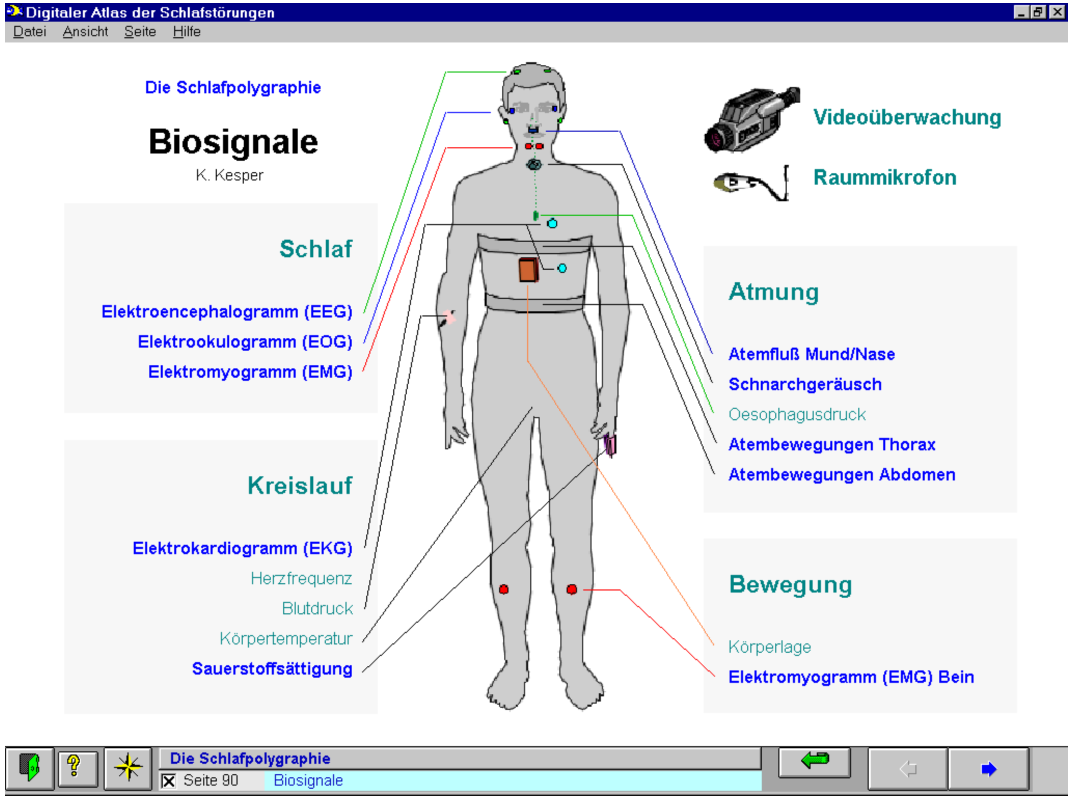
Results: Training CNN (standard Krizhevsky, 2013 architecture style) from scratch

Imaging modality	Image classification	Patient classification
LGE	0.88 +/- 0.05	0.94 +/- 0.06
MOLLI	0.87 +/- 0.01	0.92 +/- 0.05
CINE	0.84 +/- 0.11	0.89 +/- 0.08

10-fold CV; ROC AUC mean +/- SD. Patient classification – mean prediction over all images of the patient

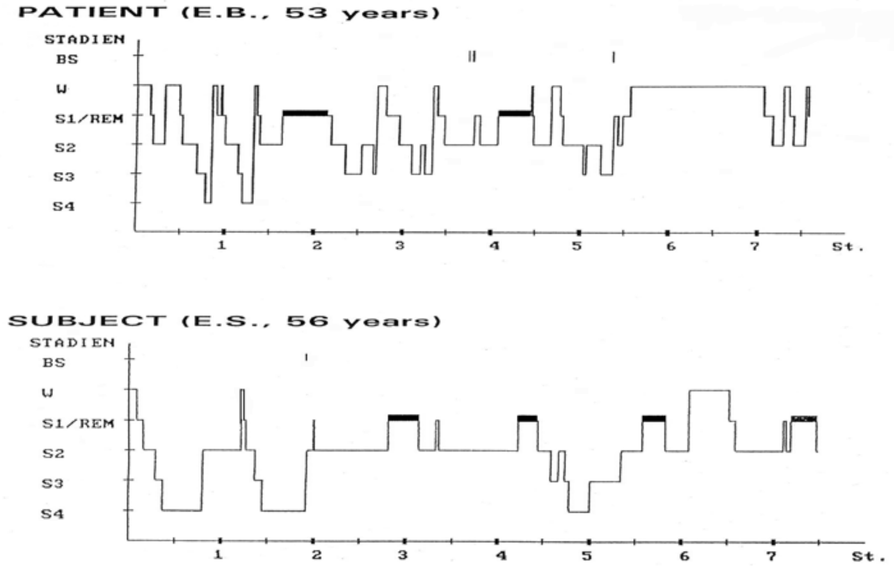
Example 3: Sleep analysis

Measurement: Polysomnography

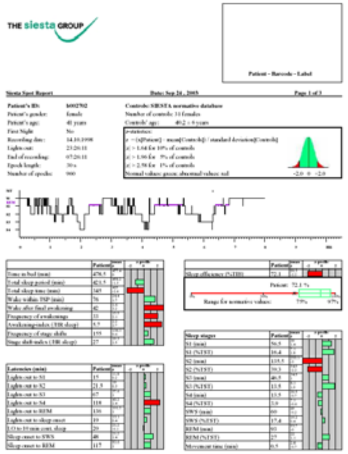


Source: DGSM

Analysis: Sleep profile (stages)



Report



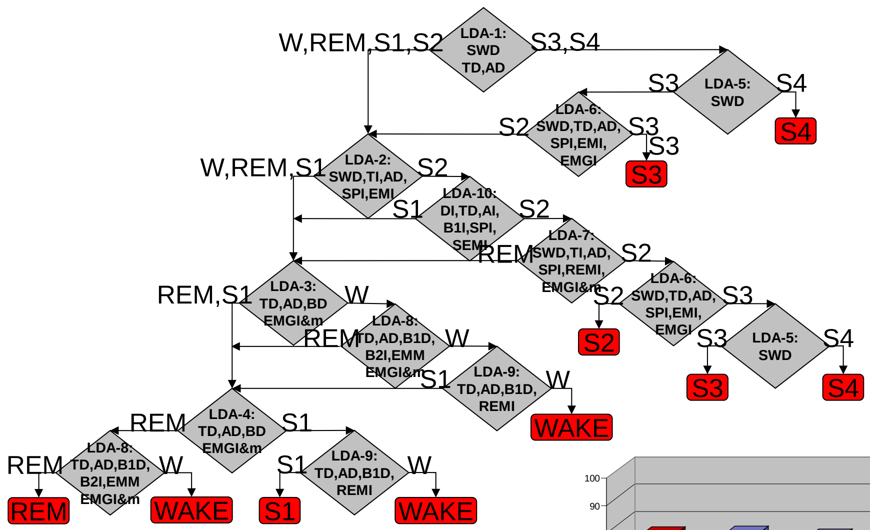
Diagnosis

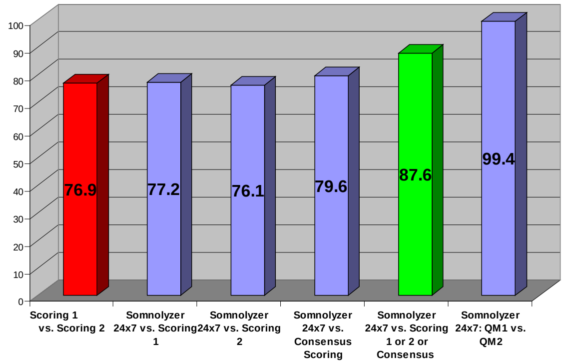
(Sleep apnea,
Periodic leg movements,
etc.)

Somnolyzer

Classical Somnolyzer:

(Anderer et al., The Siesta Group)

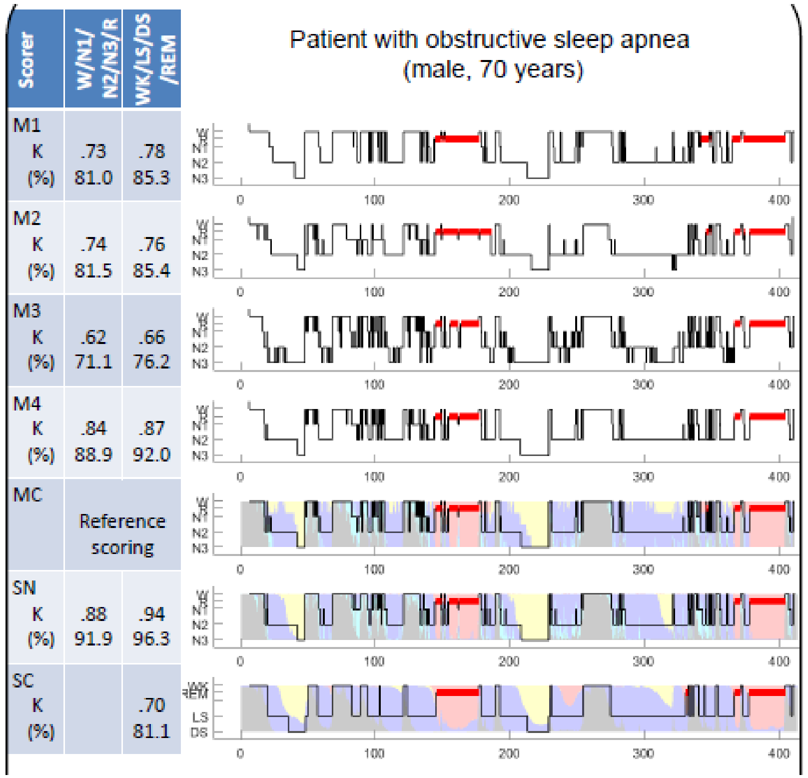




Anderer et al., Neuropsychobiology, 2010

Somnolyzer mit Deep Learning

(Anderer et al., Philips):



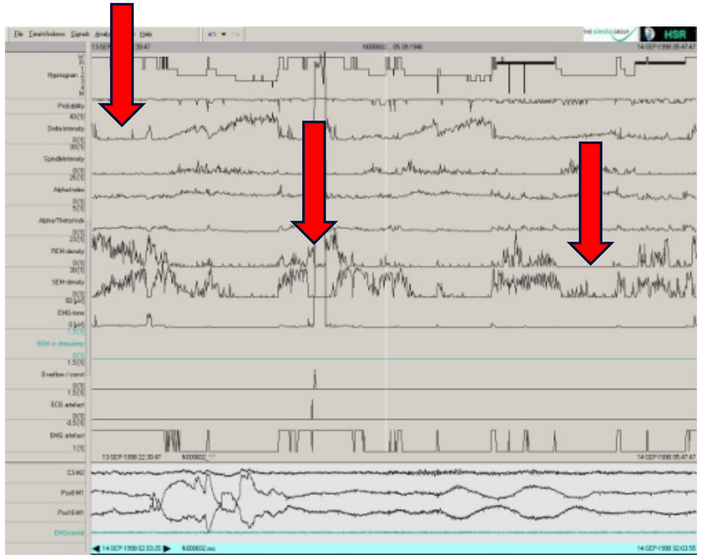
Hypnodensities as learnable concept

Optimal use by personnel– the Expert Review

- User should not trust the system blindly
- A structured process helps to still remain efficient

- Additional advantages:
  - Rare pathologies can be treated specially
  - User does not lose „feeling“ for data
  - Bigger trust

- Downside:
  - Many users change to much
  - Acceptance reduced, since scoring is a big part of job description
  - Many users fear for their job



Can we trust those methods?

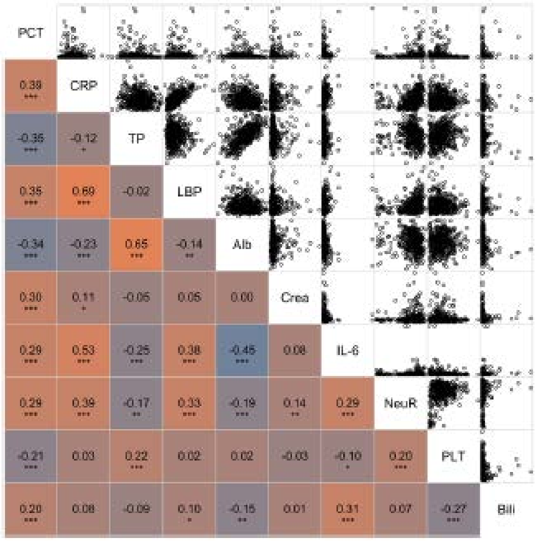
- 3 cases:
  - „ground truth“ is given through **outside criterion** (gold standard)
    AI can be better than human experts (e.g. dermatology)
  - „ground truth“ is given by gold standard, but **no human expertise**
    AI can contribute new modes of diagnosis (e.g. cardiac MRI)
  - „ground truth“ is in the eye of the **expert, with inter-rater variability**
    AI can be as good as human experts in this spectrum,
    or as good as a large number of experts (e.g. sleep medicine)
- There is no good reason to mistrust such systems, as long as
  - The final decision is with the human
  - It was made sure that data was representative
  - Only the reduced „diagnostic report“ task is left to AI

Example 4: Prediction of bacteraemia

Feature	Unit	Missing	No bacteraemia	Bacteraemia	p–value	ROCs (95%CI)
PCT	ng/ml	1.5%	0.3 (0.1–1.0)	1.6 (0.4–5.4)	<0.001*	0.729 (0.679–0.779)
CRP	mg/dl	0.4%	12.9 (7.9–20.4)	15.0 (9.6–22.8)	0.020	0.569 (0.512–0.626)
LBP	µg/ml	1.1%	23.1 (15.6–35.4)	29.7 (19.7–44.25)	<0.001*	0.610 (0.553–0.667)
IL-6	pg/ml	1.7%	42.8 (19.2–99.9)	49.6 (28.9–130.0)	0.028	0.566 (0.508–0.623)
Fib	mg/dl	6.0%	607 (446–752)	613 (490–714)	0.875	0.505 (0.447–0.563)
SI	µg/dl	1.7%	26.0 (14.8–57.0)	21.0 (12.3–46.8)	0.042	0.561 (0.502–0.619)
TP	g/l	2.4%	61.6 (55.8–67.5)	60.5 (53.2–65.8)	0.062	0.556 (0.498–0.614)
ALAT	U/L	1.9%	25.0 (15.0–45.0)	33.0 (18.0–64.0)	0.005	0.585 (0.526–0.643)
Alb	g/l	2.4%	31.3 (27.9–35.1)	29.5 (25.6–33.1)	<0.001*	0.603 (0.546–0.673)
Bili	mg/dl	3.9%	0.6 (0.5–1.0)	0.8 (0.6–1.4)	<0.001*	0.616 (0.558–0.673)
γ–GT	U/L	2.4%	63 (30–130)	102 (44–260)	<0.001*	0.617(0.559–0.675)
Crea	mg/dl	0.4%	0.9 (0.8–1.3)	0.9 (0.8–1.3)	0.707	0.511 (0.452–0.570)
LDH	U/L	4.3%	220 (168–312)	200 (157–286)	0.161	0.543 (0.485–0.602)
Hb	g/dl	2.4%	10.0 (9.0–11.7)	10.0 (9.1–10.9)	0.340	0.529 (0.473–0.584)
Plt	G/l	2.4%	208 (116–326)	190 (134–279)	0.426	0.524 (0.457–0.580)
WBC	G/l	2.4%	8.7 (5.4–13.6)	8.7 (5.5–12.6)	0.629	0.486 (0.428–0.543)
NeuR	%	7.1%	75.5 (63.9–82.8)	79.5 (67.2–86.6)	0.021	0.570 (0.510–0.631)
EosR	%	4.3%	0.9 (0.2–2.5)	0.8 (0.2–1.9)	0.308	0.531 (0.472–0.589)
IL-10**	pg/ml	0.0%	2.2 (1.4–4.6)	3.2 (1.7–7.3)	0.002	0.589 (0.532–0.645)
IL-17a**	pg/ml	0.0%	0.8 (0.0–3.1)	2.7 (0–7.5)	<0.001*	0.601 (0.542–0.660)
MIP-1b**	pg/ml	0.0%	52.1 (29.7–82.5)	72.05 (43.6–134.7)	<0.001*	0.615 (0.557–0.673)

Ratzinger et al., Scientific Reports, 2018

- Features: laboratory values
- 400 patients, all major ML methods
- No improvement over known single predictor (procalcitonin)
- Deep Learning would not advance this either



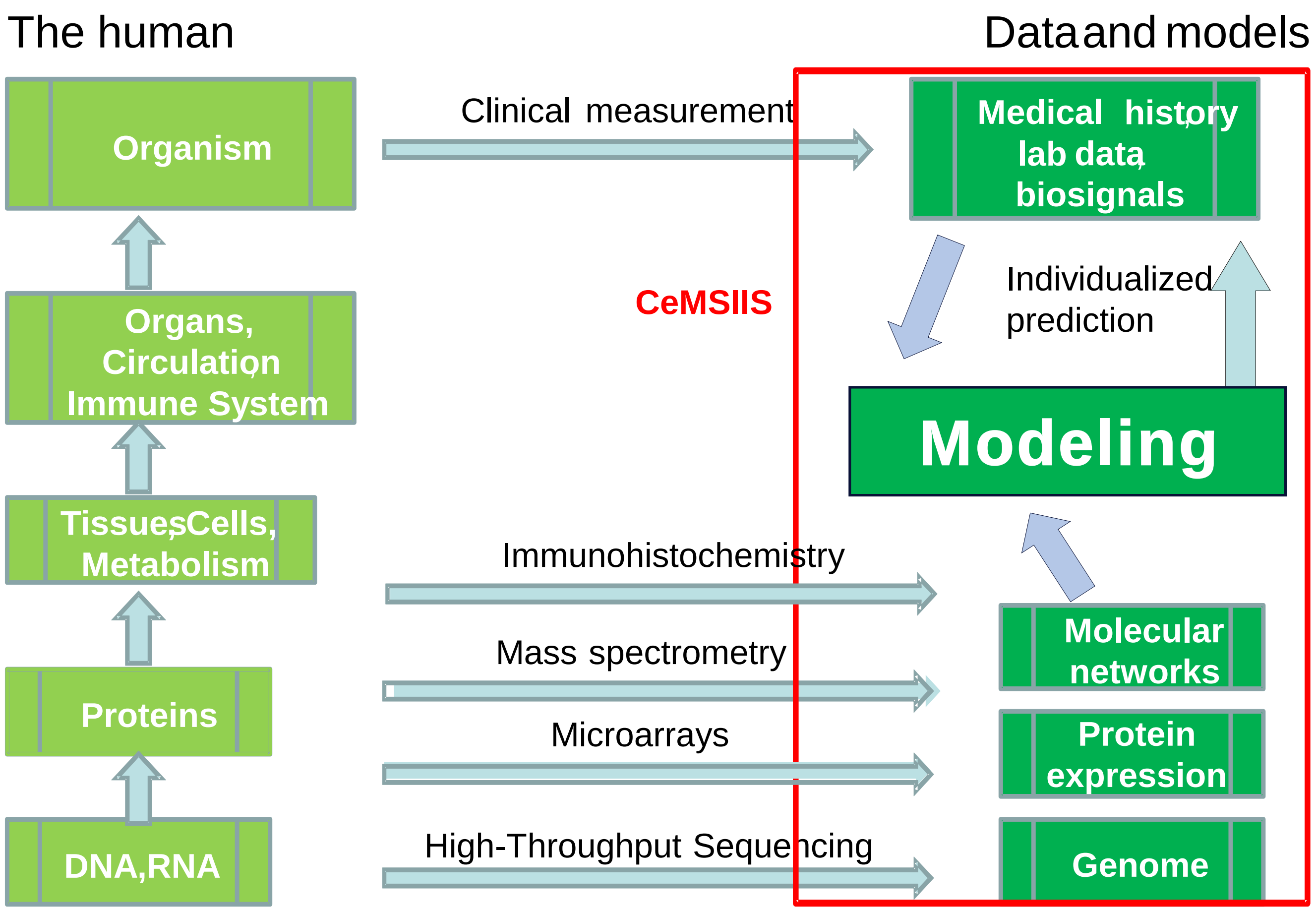
- AI is no panacea
- Yet it raises sensitivity toward data science

How intelligent is that?

- Neural networks learn complex relationships between input and output data
- In most cases, data comes from a narrow domain and must cover or represent this domain sufficiently
- For every decision going beyond such a narrow domain, especially the connection with results from other domains, and a patient‘s history, AI still has ist limits

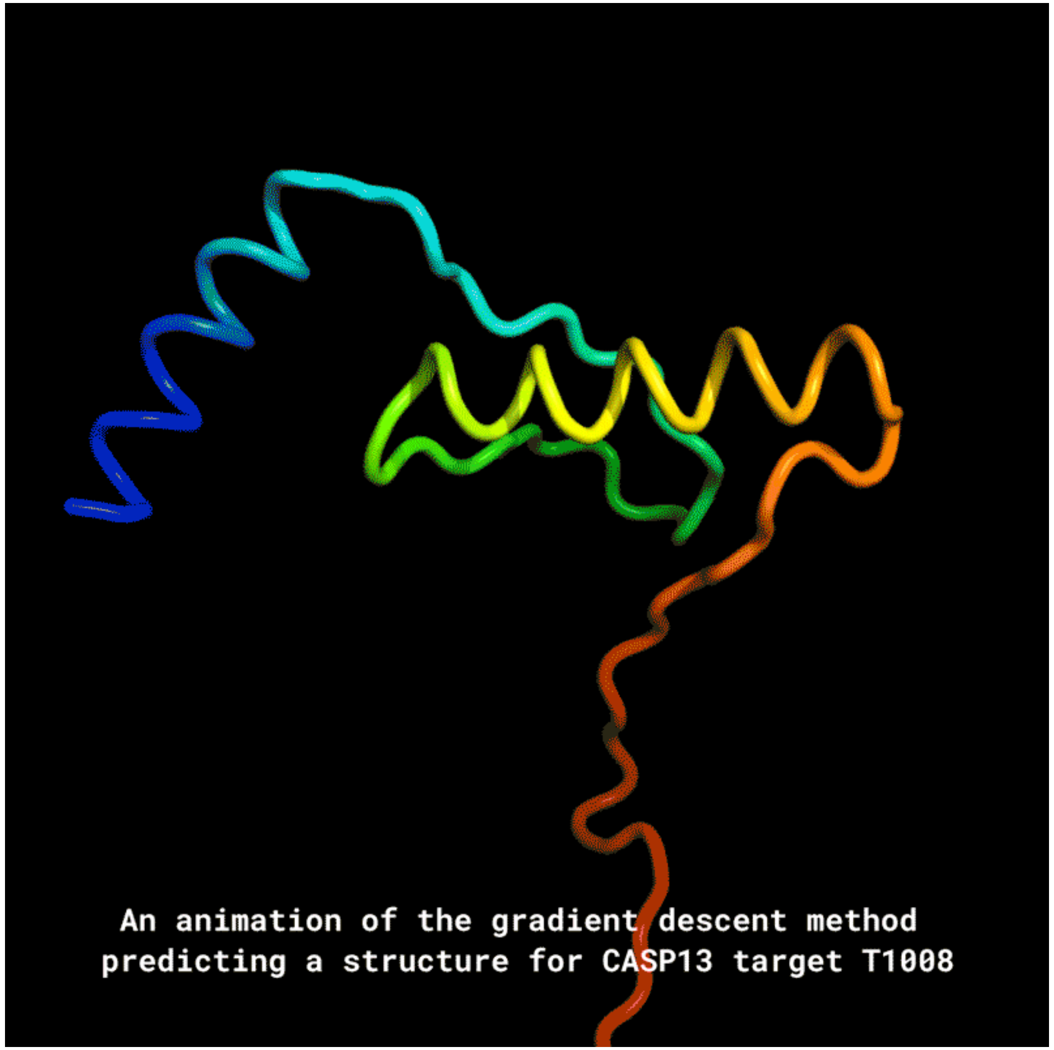
- Question: Can AI replace the physician?
- Answer: Not generally, only with respect to pinpointed (mostly diagnostic) tasks
  There is still plenty of room for human intelligence: Empathy, Communication, Reasoning, Creativity

Data Science for Personalized Medicine



Alpha-Fold

- Predicting structure of proteins from amino acid sequence
- Extremely complex, astronomic number of possibilites
- Machine learning has been applied to this for decades
- Recent success equals that of Alpha Go



https://deepmind.com/blog/alphafold/

Some missing link(s)

- While many applications in health care do not need „big data“, many other promises (e.g. personalized medicine) depend on it
- But: necessary data often does not exist (or is not big enough)
- Ideal dataset:
  - Complete genome, proteome, etc.
  - All organic functions and lab values
  - Complete medical history
  - Environmental and social factors
  - Longitudinal observations
- Many reasons (ethics, data protection, but also insufficient technology) still prevent this

Conclusion

- Deep learning leads to true advances in medicine, from diagnostics to monitoring to prediction
- Big goals like all-encompassing Personalised Medicine need AI
- However, these systems have a narrow expertise
- „General“ artificial intelligence still far away

- Physicians do not need to feel threatened, if they focus on their human capabilities (general intelligence, empathy, reasoning, etc.)
- But physicians need digital know-how and assessment experience