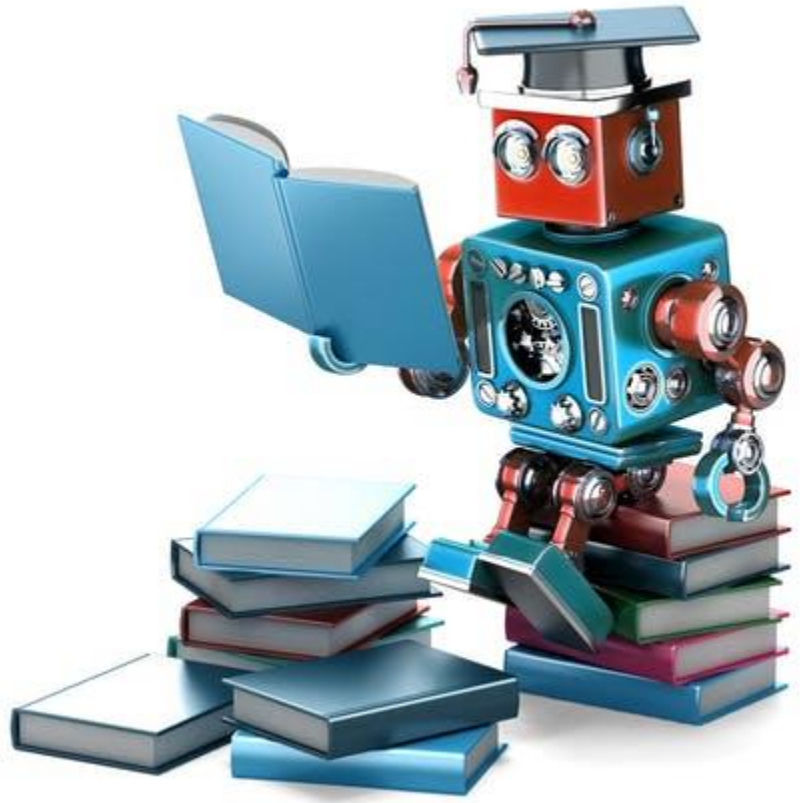
Deep Learning & XAI

Professor Thomas B. Moeslund
Aalborg University

AALBORG UNIVERSITY
DENMARK

Agenda

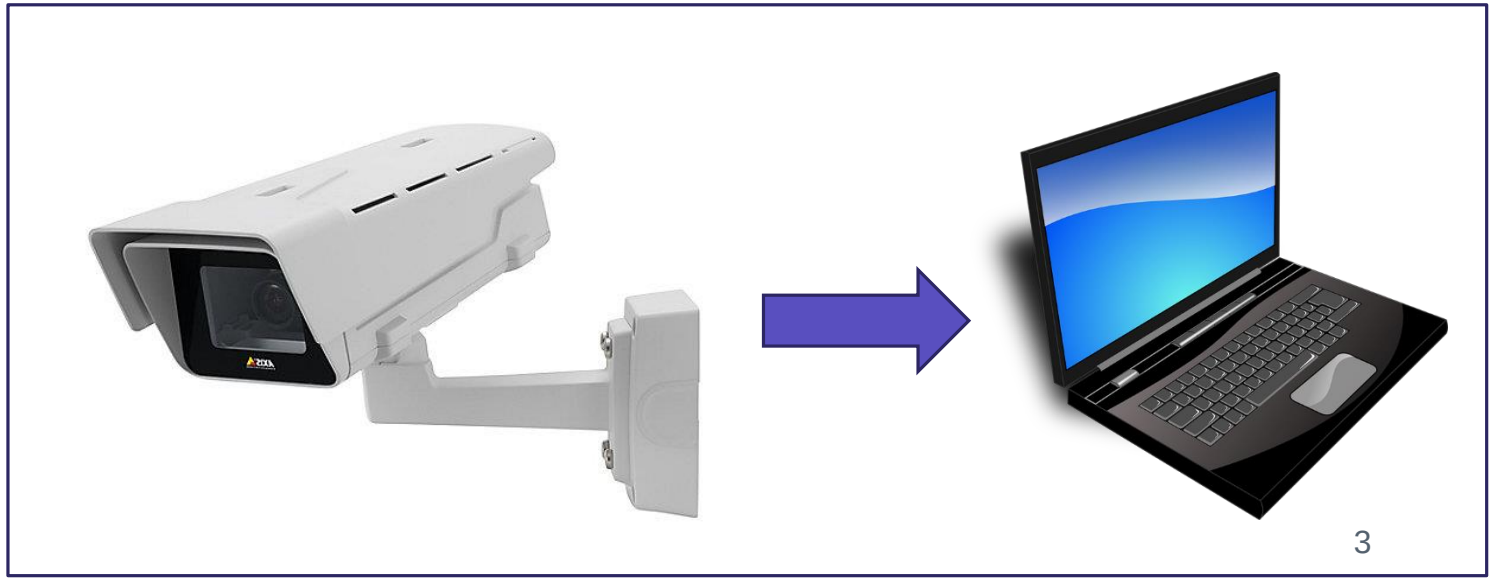
- Who am I?
- What is AI?
- Deep Learning
- Problems with AI
  - XAI
- Will the machines take over the world?
- Q&A



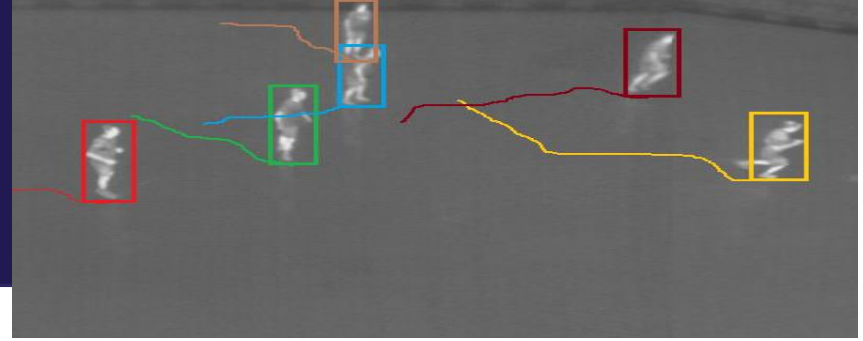
Background

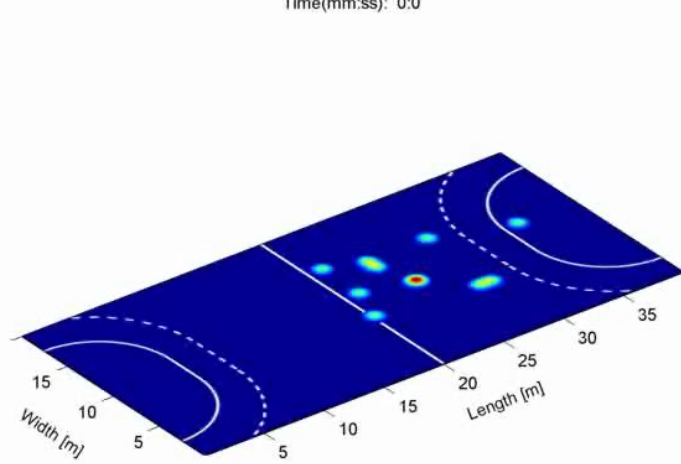
Who am I?

- M.Sc. EE. 1996, PhD 2003. AAU
- Research: Computer Vision
  - AI, pattern recognition, machine learning, deep learning, AI



The lab

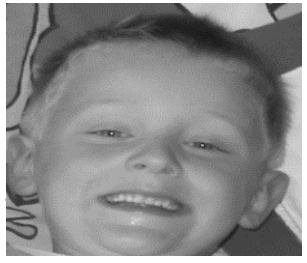


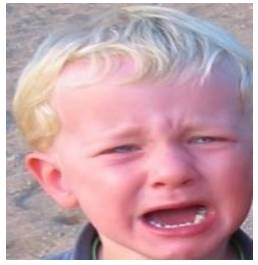


Is there anybody out there?

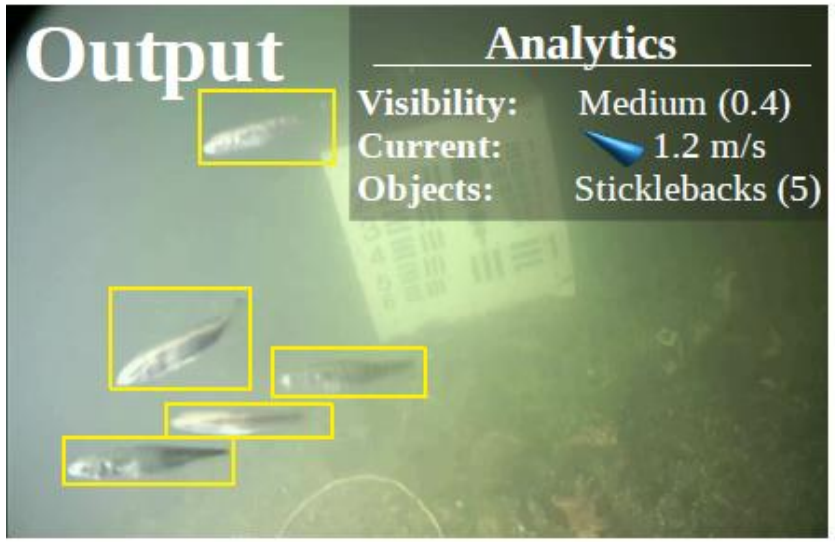
Who are they?

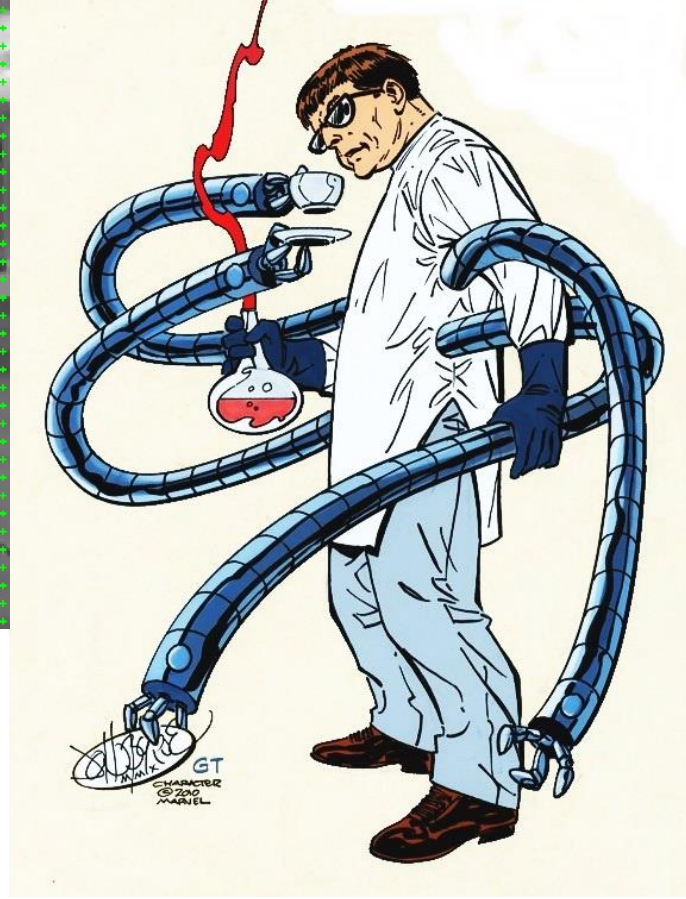
What are they doing?

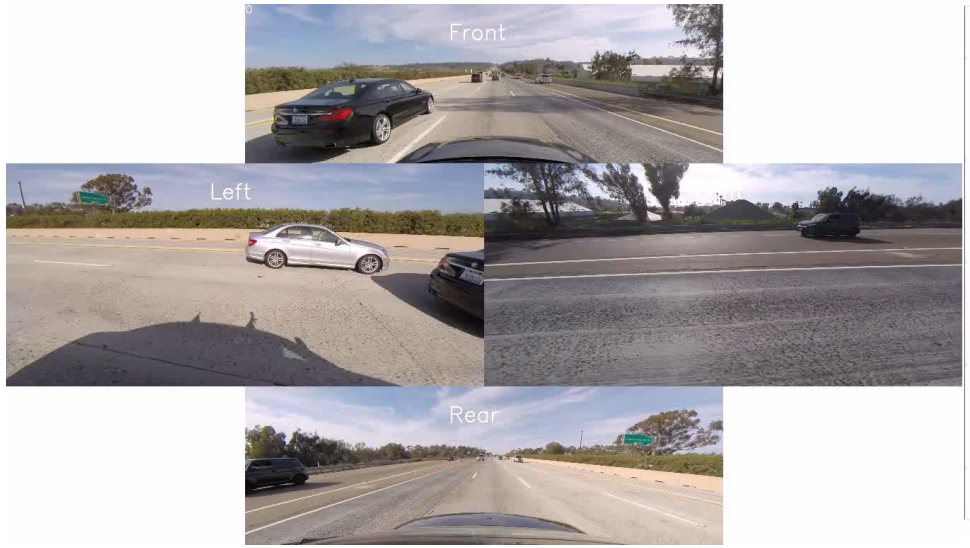




The lab


Output
Analytics
Visibility: Medium (0.4)
Current: 1.2 m/s
Objects: Sticklebacks (5)

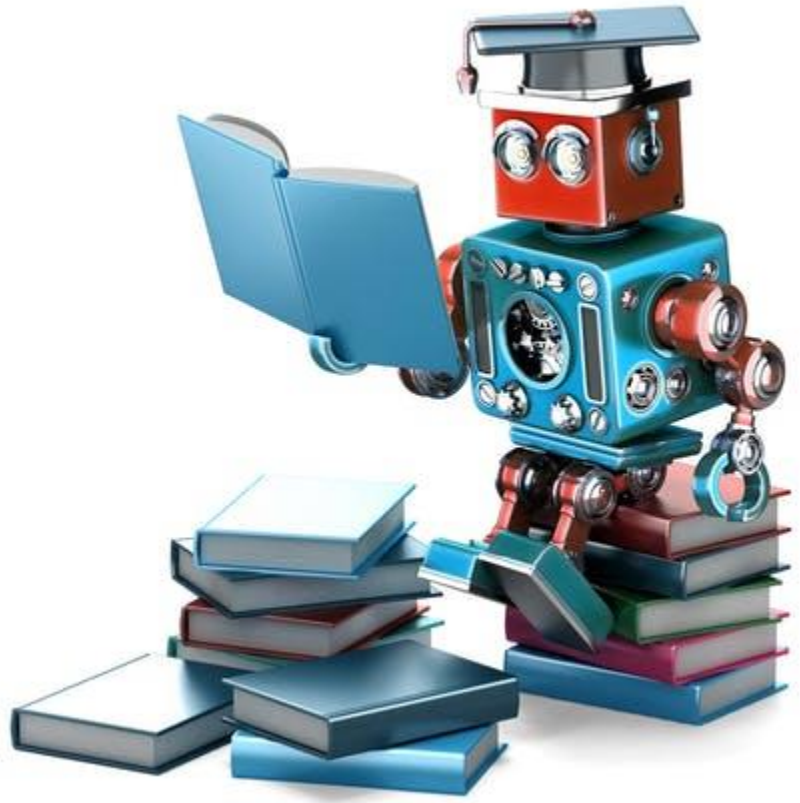



Front
Left
Rear


40
20
0
−20
−40
X Distance [m]
−5
0
5
Y Distance [m]

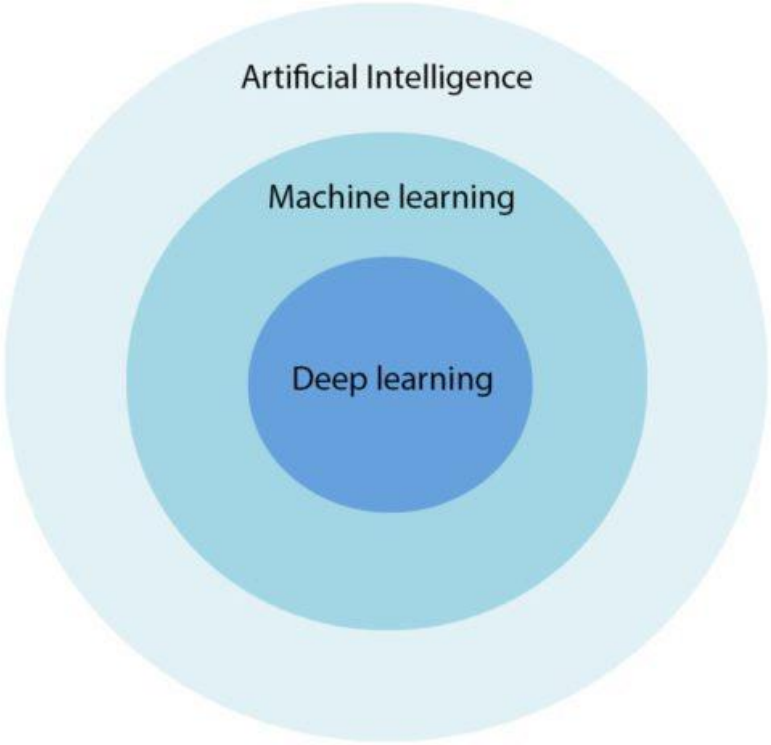
Agenda

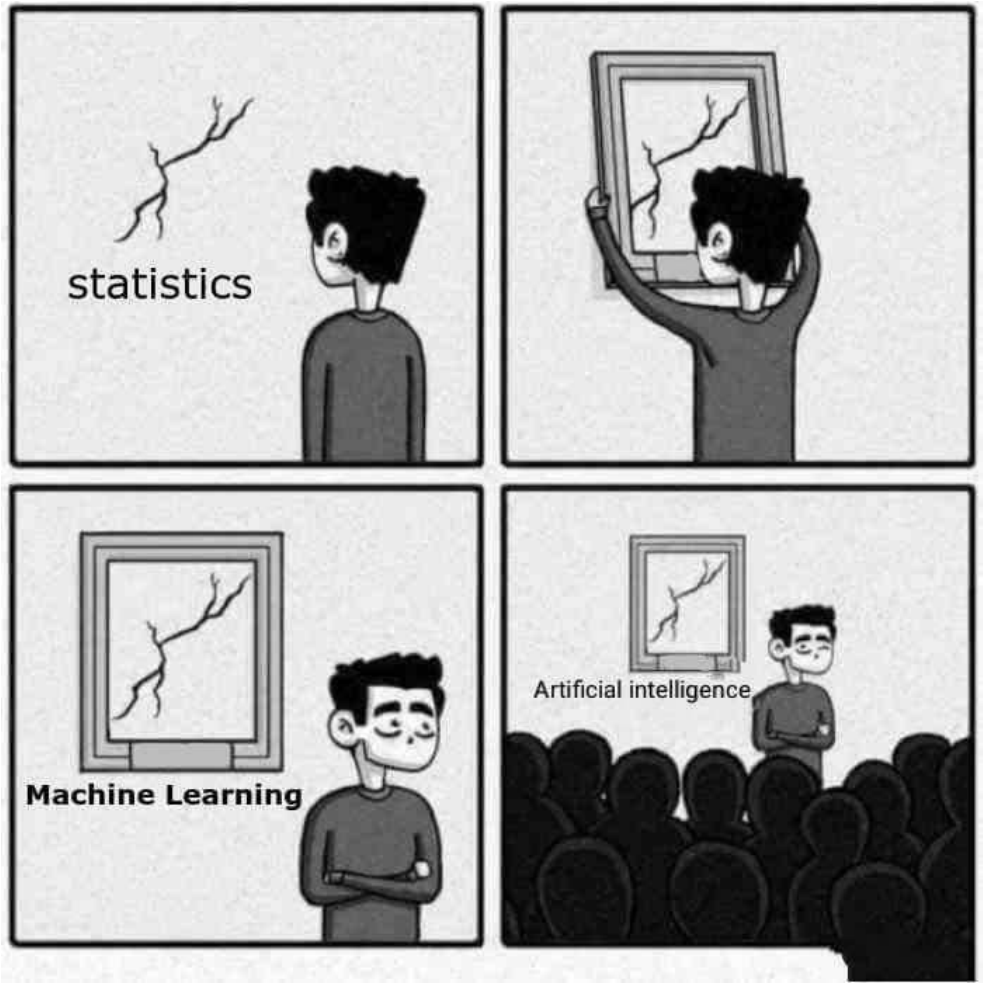
- Who am I?
- What is AI?
- Deep Learning
- Problems with AI
  - XAI
- Will the machines take over the world?
- Q&A



What is AI?

- AI: ~1940s
- Machine Learning: ~1990s
- Deep Learning: ~2010s





What is AI?

EU:
April 2018

"**Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.** AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."

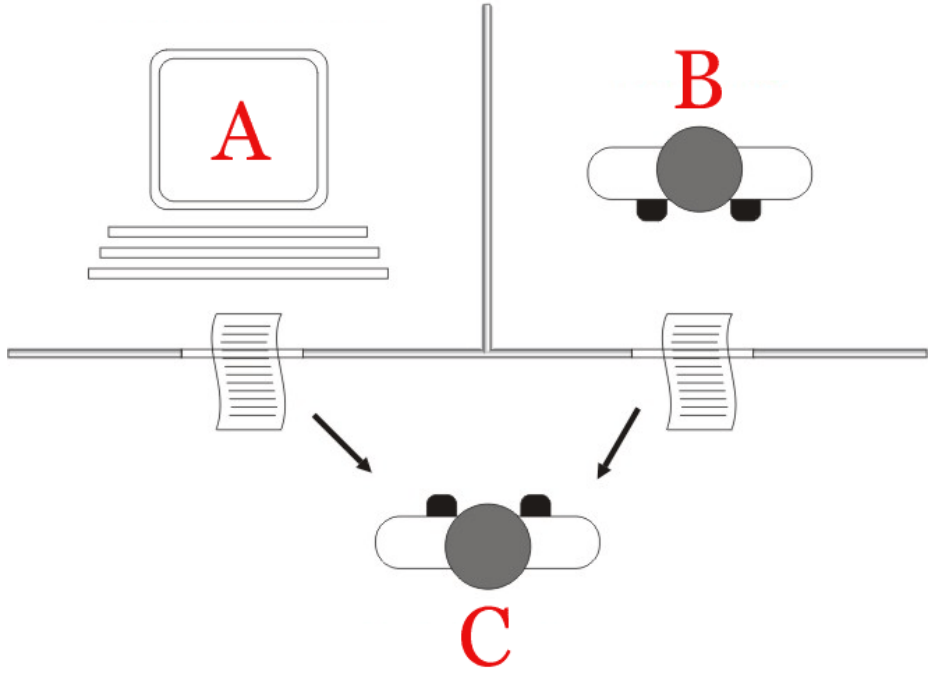
EU:
Dec 2018

"**Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions.** As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

What is AI?

Intelligence

- What does ”intelligence” mean?
- Is a machine intelligent?
- Alan Turing (1912 - 1954)
- The Turing-Test:

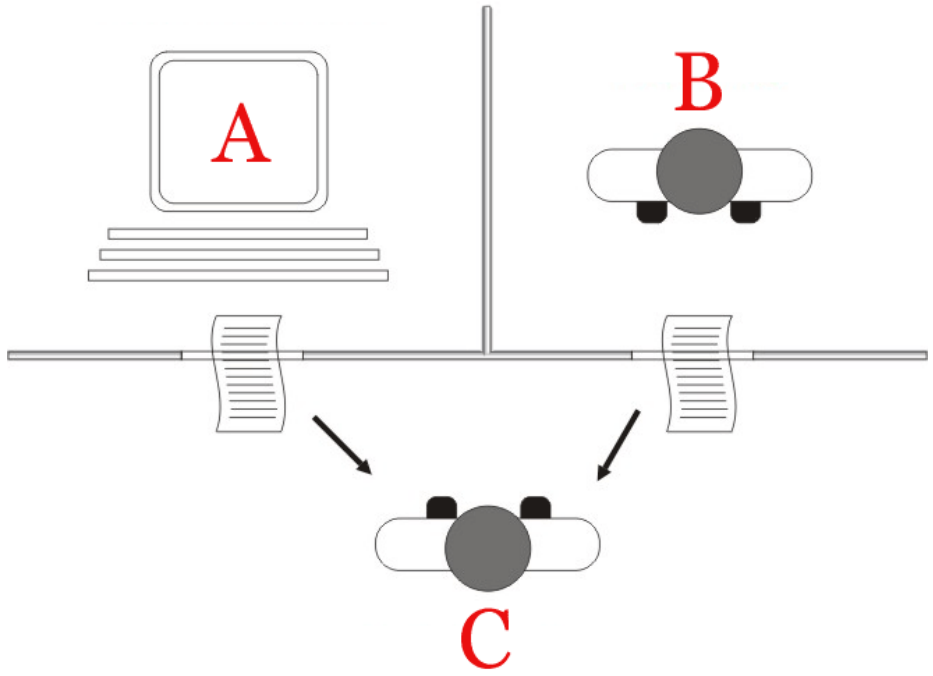


AALBORG UNIVERSITY
DENMARK

What is AI?

Intelligence

- What does ”intelligence” mean?
- Is a machine intelligent?
- Alan Turing (1912 - 1954)
- The Turing-Test:

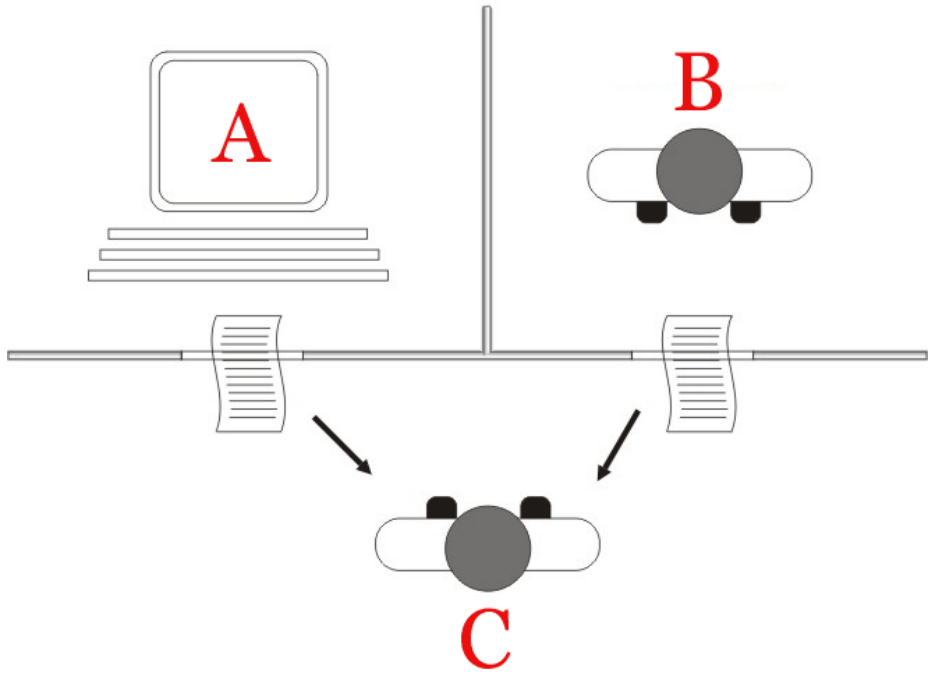


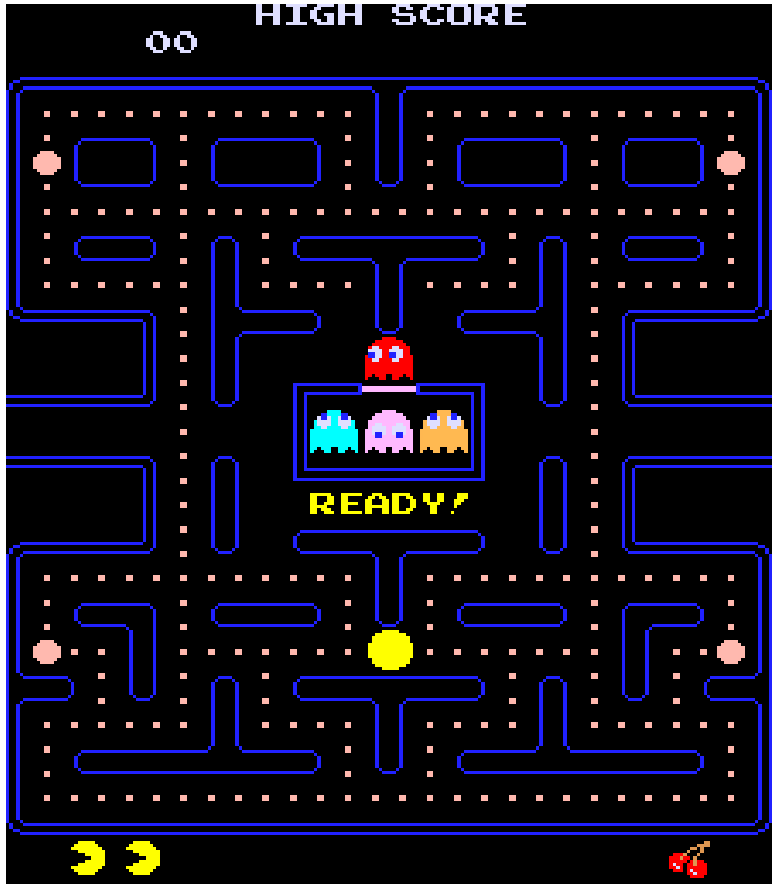
Arthur C. Clarke (1917 - 2008):
” Any sufficiently advanced technology is indistinguishable from magic”

What is AI?

Intelligence

- What does ”intelligence” mean?
- Is a machine intelligent?
- Alan Turing (1912 - 1954)
- The Turing-Test:



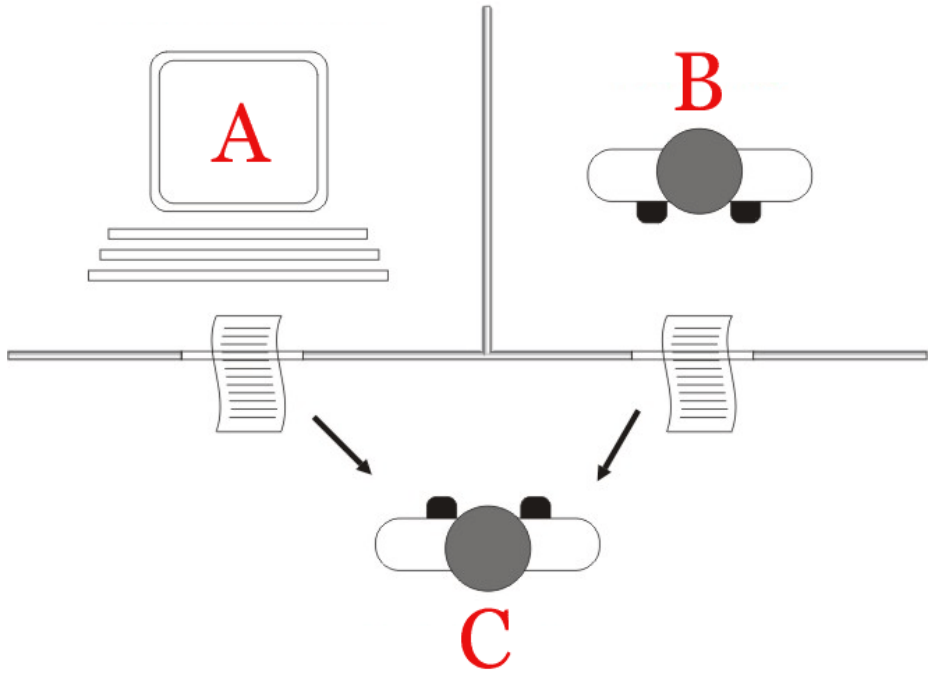


AALBORG UNIVERSITY
DENMARK

What is AI?

Intelligence

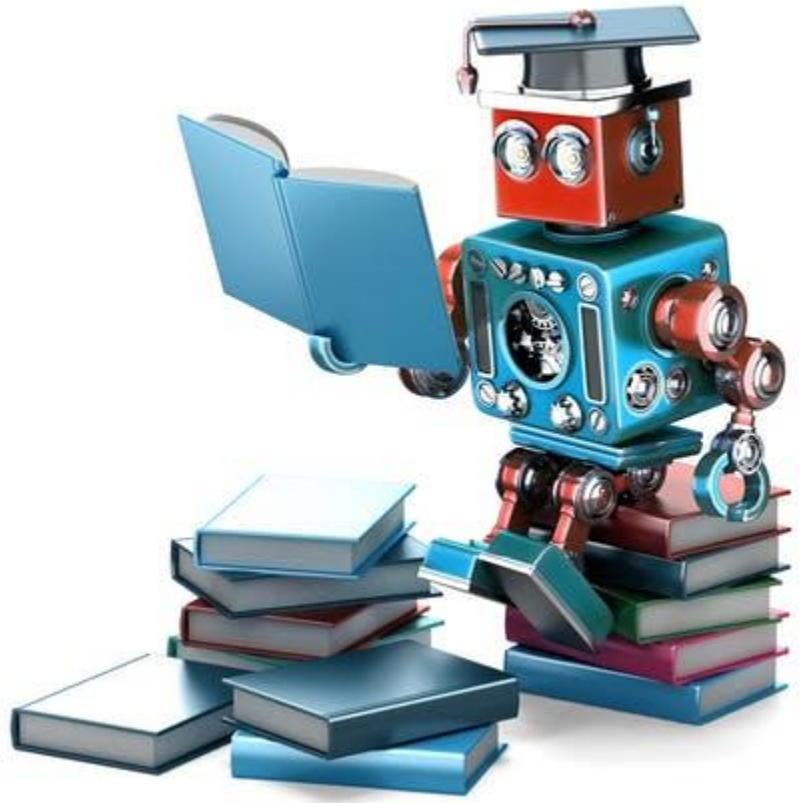
- What does ”intelligence” mean?
- Is a machine intelligent?
- Alan Turing (1912 - 1954)
- The Turing-Test:



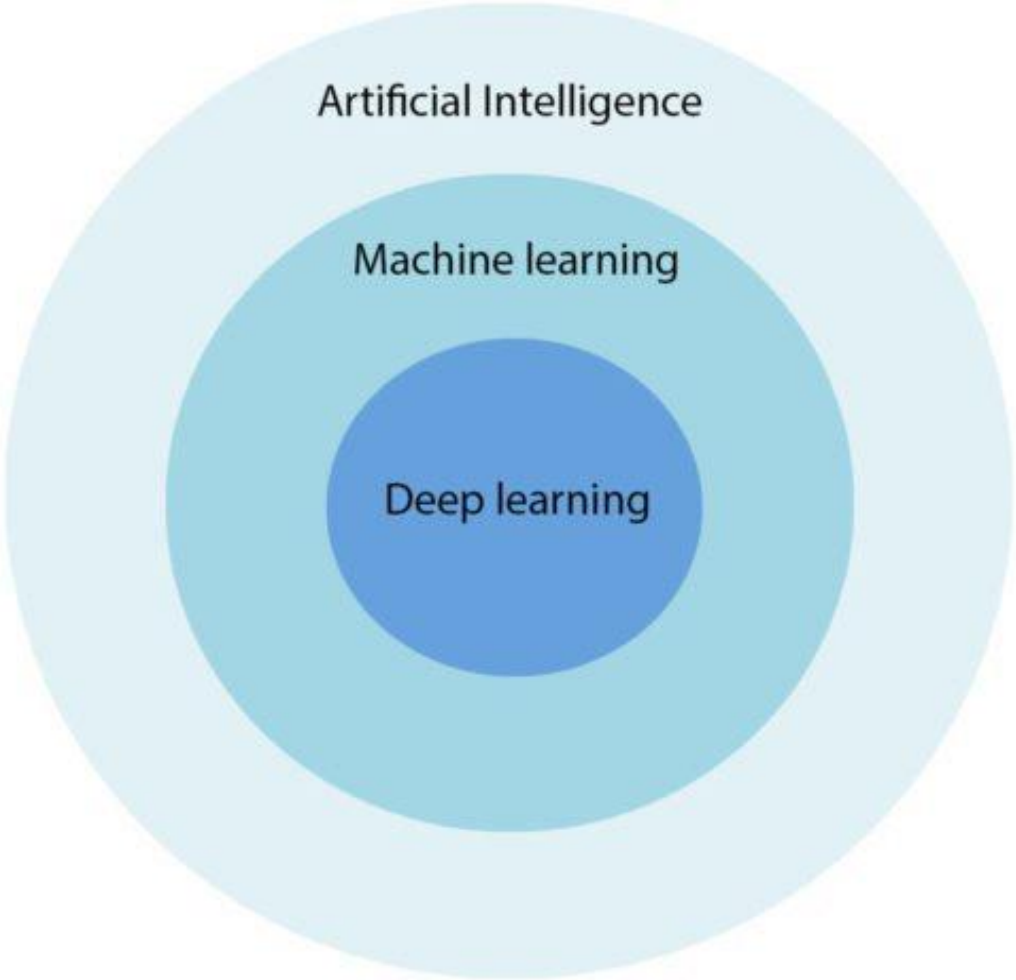
AALBORG UNIVERSITY
DENMARK

Agenda

- Who am I?
- What is AI?
- Deep Learning
- Problems with AI
  - XAI
- Will the machines take over the world?
- Q&A

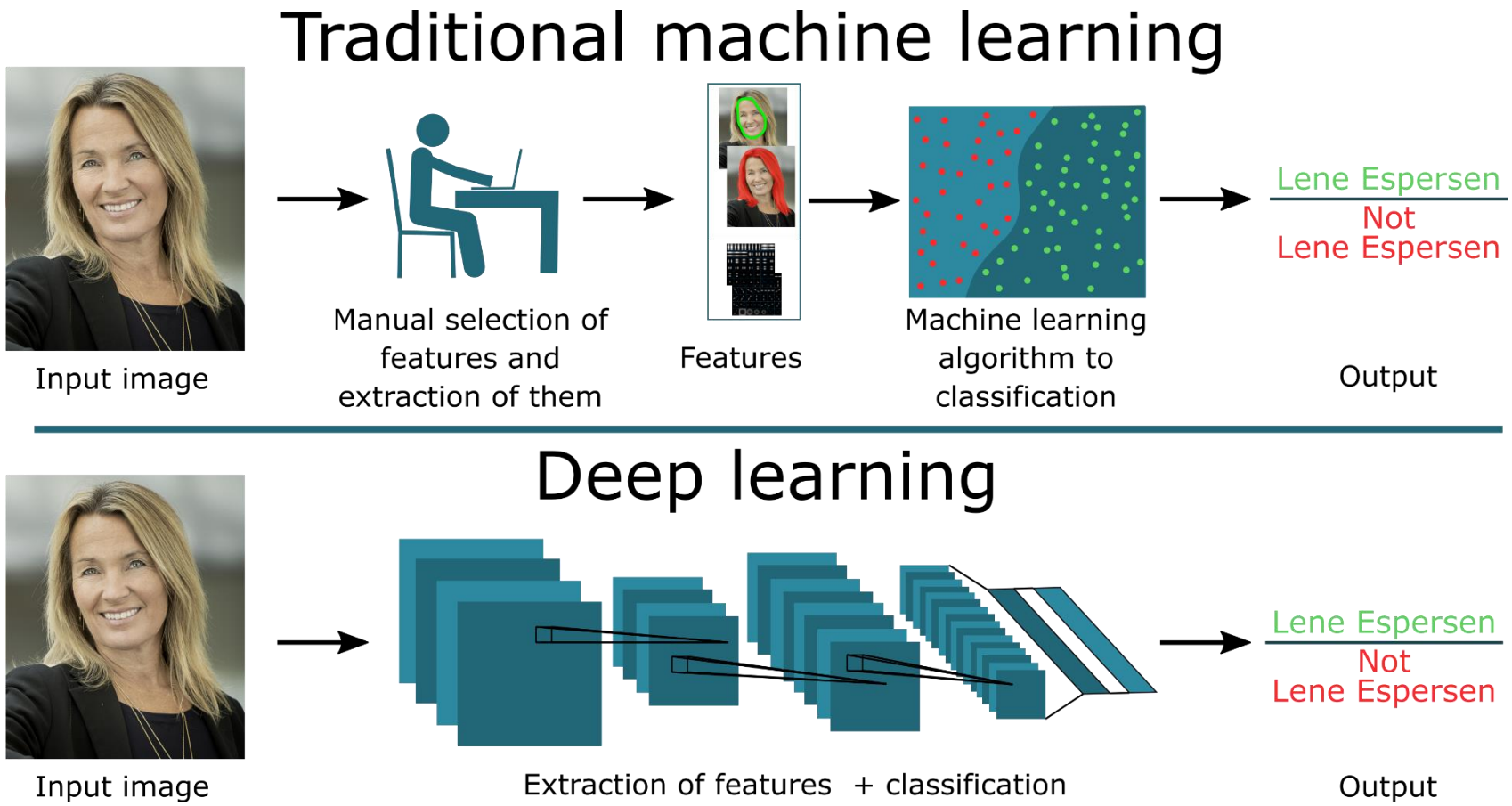


Deep Learning



Deep learning

From machine learning to deep learning

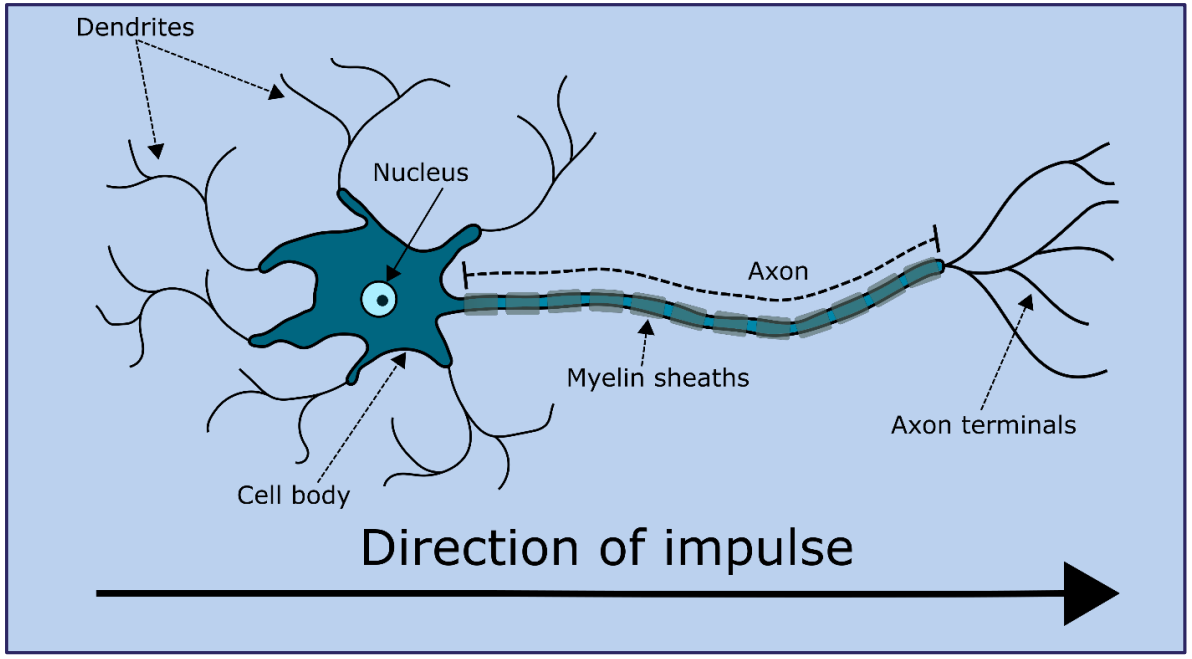


AALBORG UNIVERSITY
DENMARK

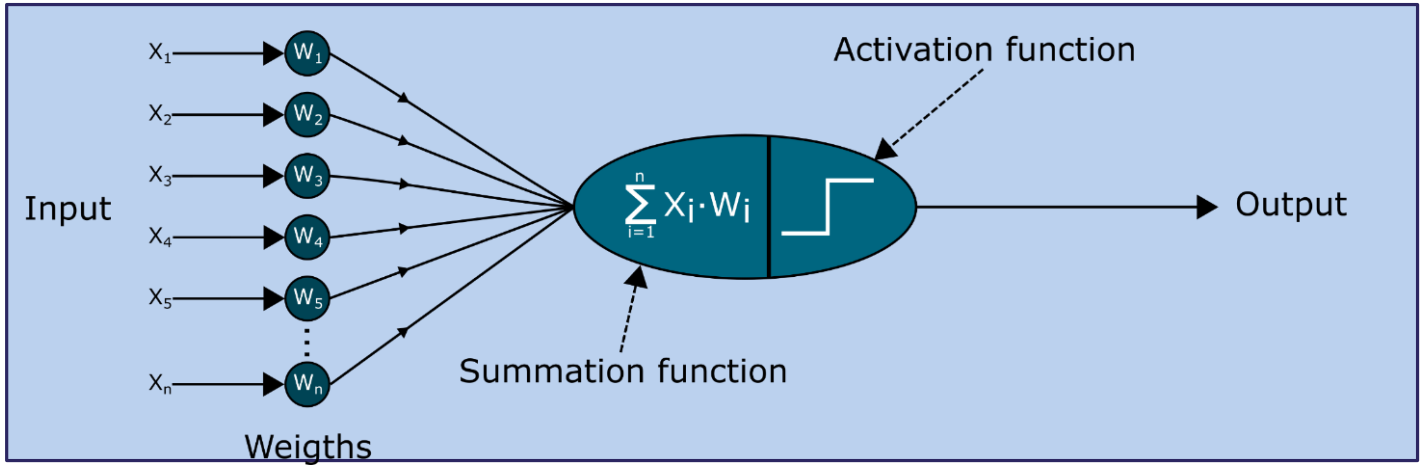
Deep learning

Artificial neural network

Real neuron

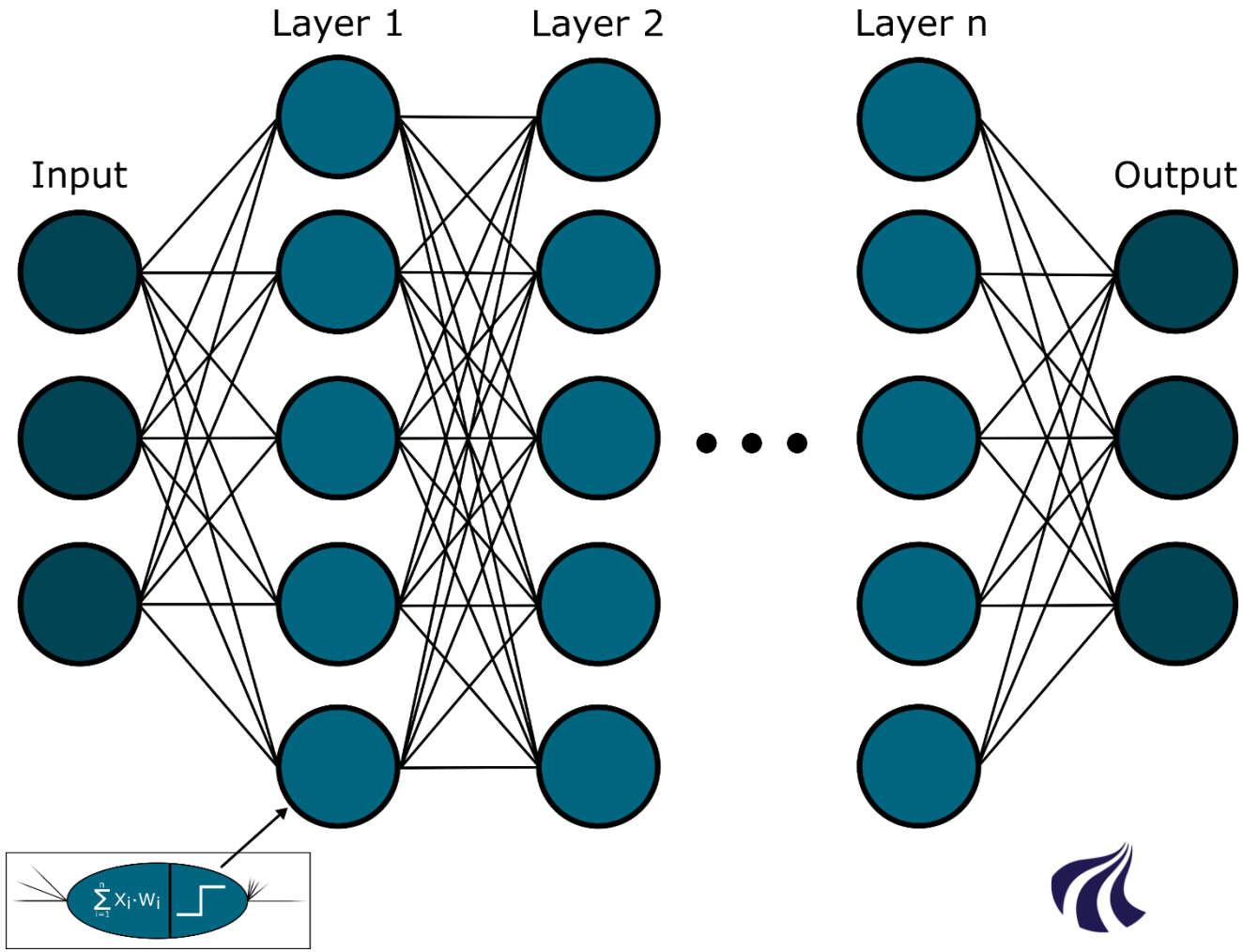


Artificial neuron



Deep learning

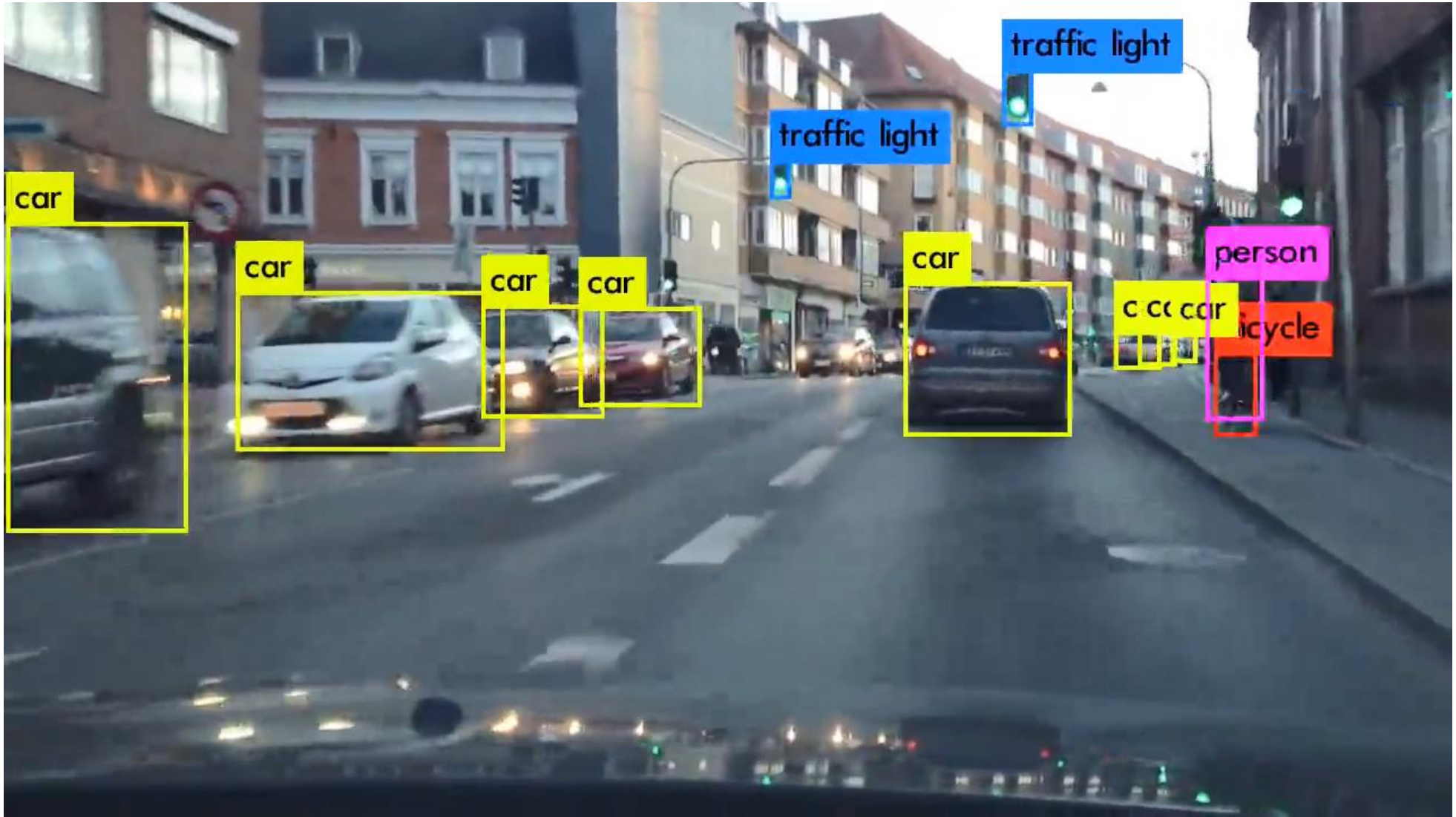
Artificial neural network



- The number of layers describe the depth of the network
- 5+ => Deep…

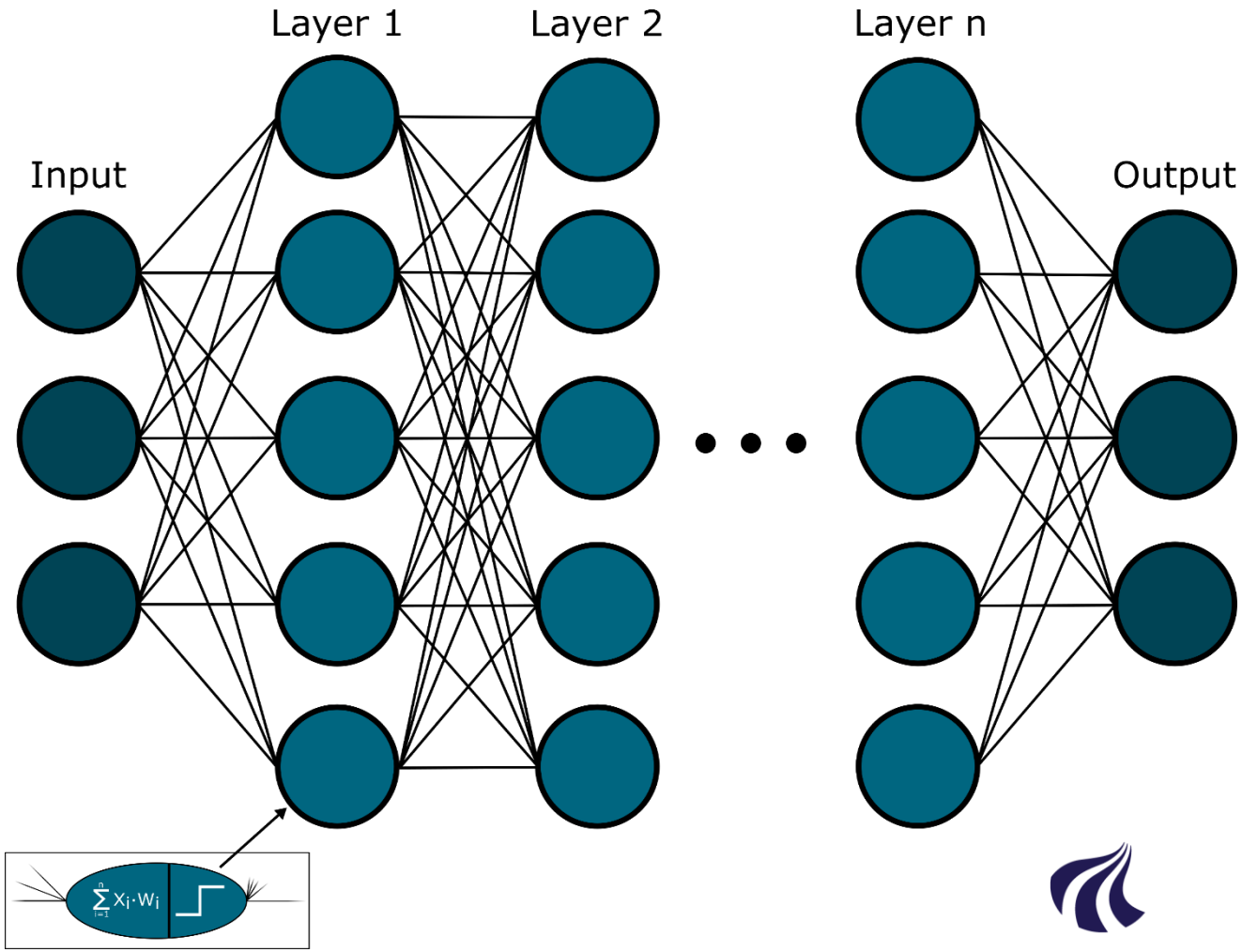
AALBORG UNIVERSITY
DENMARK

It works… !



Deep learning

Why now? (AlexNet, GPU, Data)



- Training:
  - Present many known input-output relations

→ ”Per”

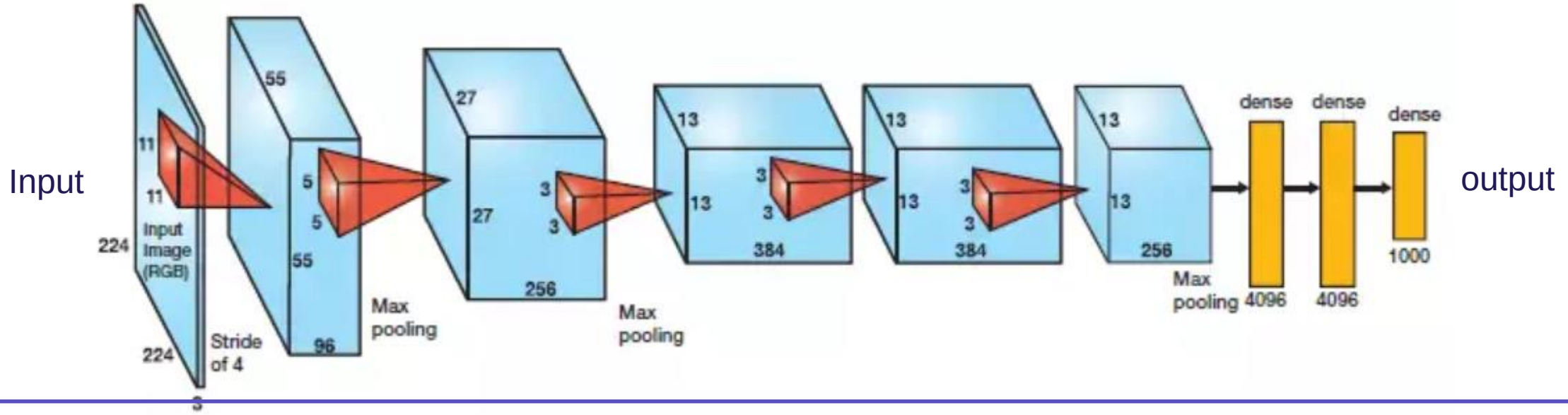
→ ”Lene”

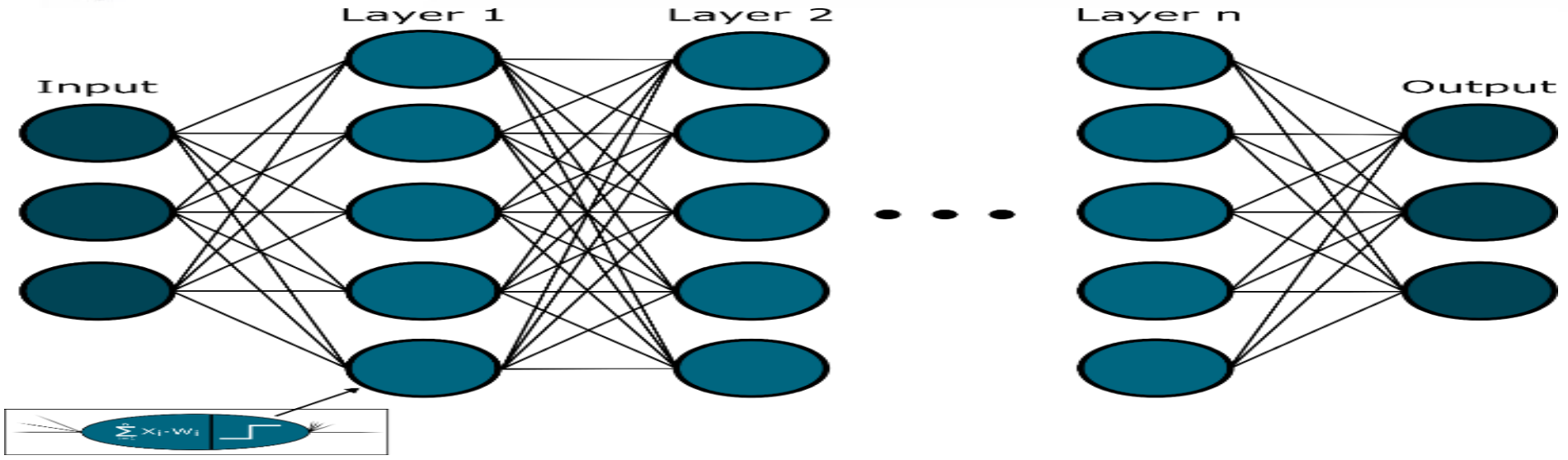
→ ”Cat”

  - Let the network iterate
    - Many calculations

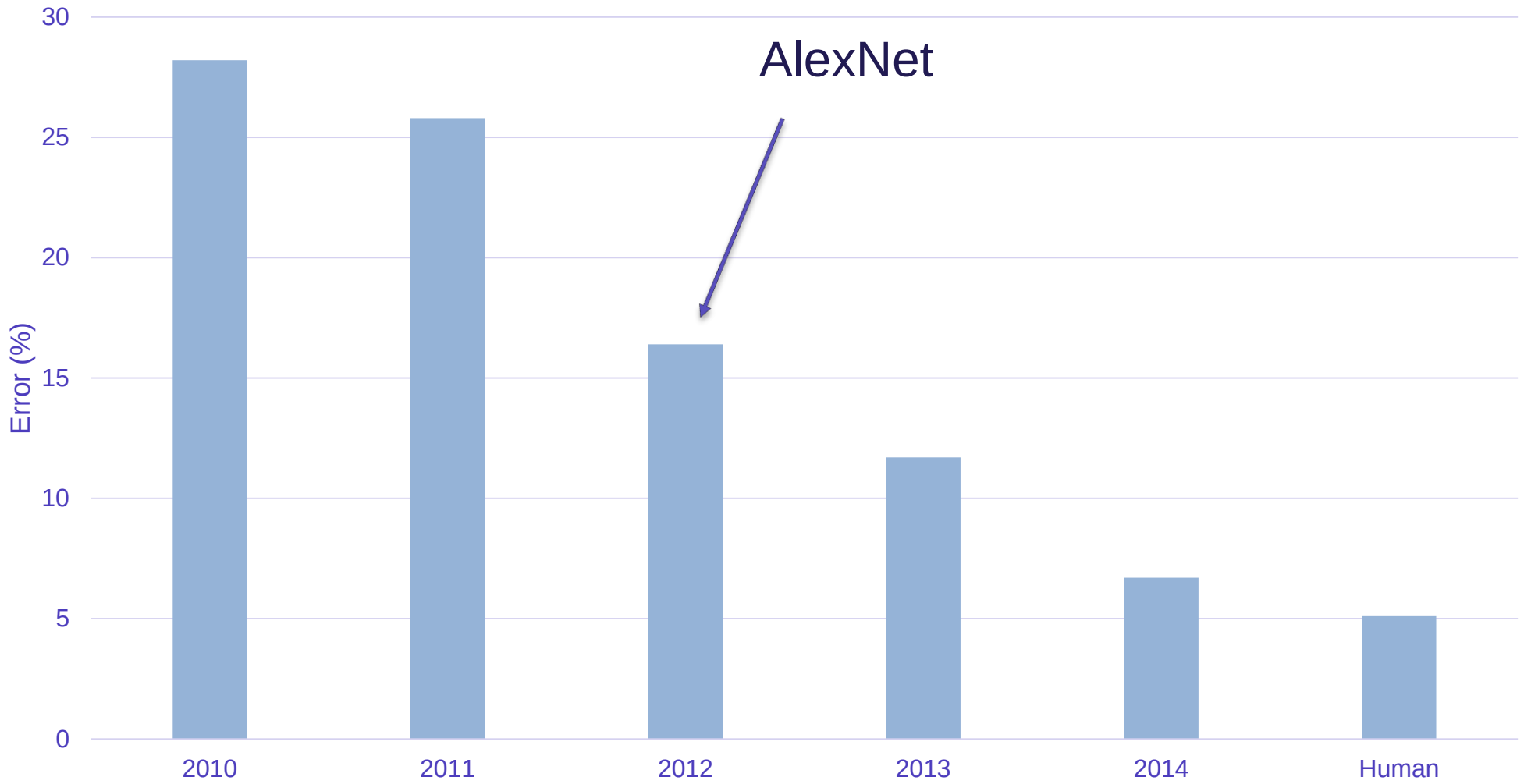
AALBORG UNIVERSITY
DENMARK

AlexNet (2012)

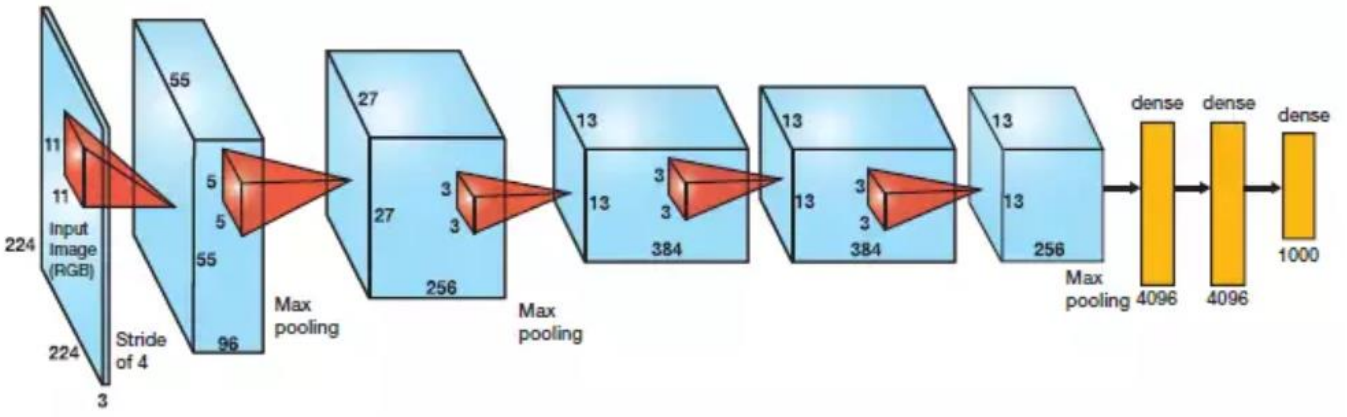


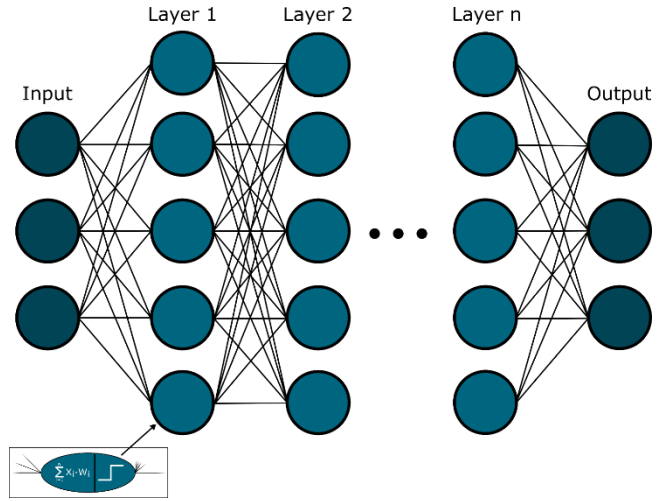


“World Cup” in image recognition



AlexNet (2012)





In	W1	L1	W2	L2	W3	L3	W4	L4	W5	L5	W6	L6	W7	L7	W8	Out
150 E3	45 E9	300 E3	60 E9	200 E3	13 E9	65 E3	4 E9	65 E3	3 E9	40 E3	160 E6	4 E3	16 E6	4 E3	4 E6	1 E3

Neurons = 800.000

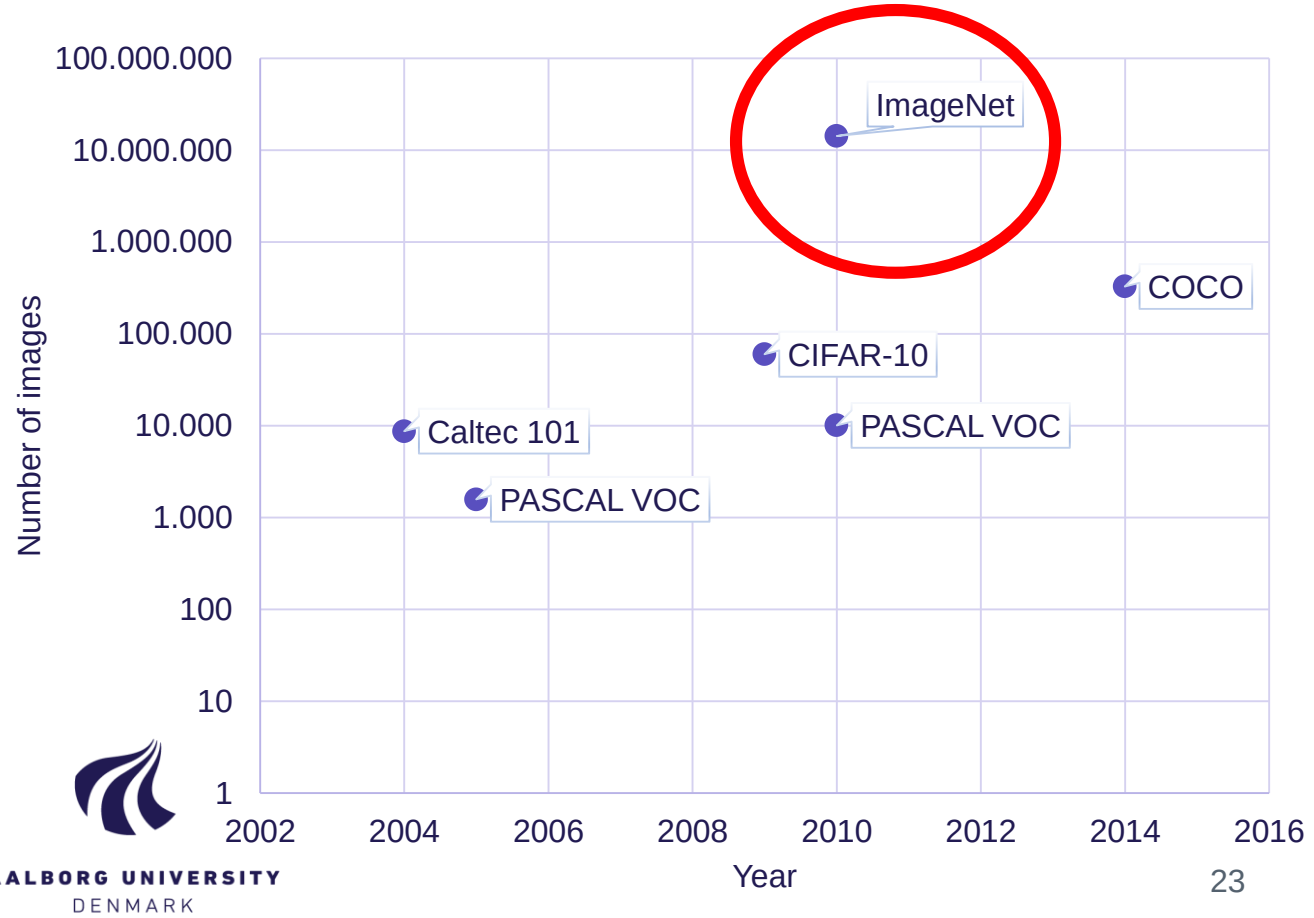
Weights = 120.000.000.000

Free parameters = 60.000.000

Our brain: 100.000.000.000 neurons

Data

- ImageNet
  - 14 million images
  - 20,000 classes
  - Images from Internet (Flickr)
  - Human labels from Amazon Mechanical Turk

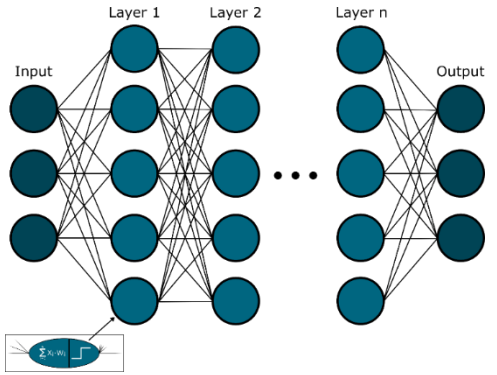


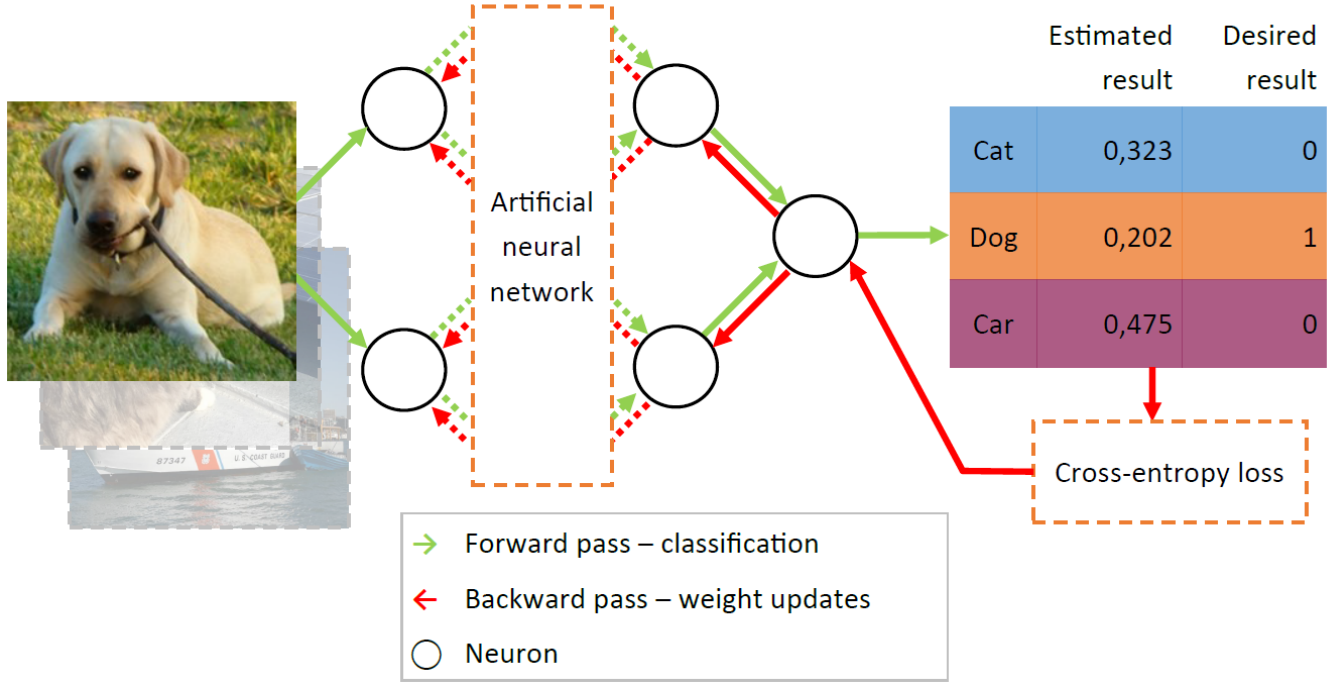
AALBORG UNIVERSITY
DENMARK

Deep Learning

How does it work?

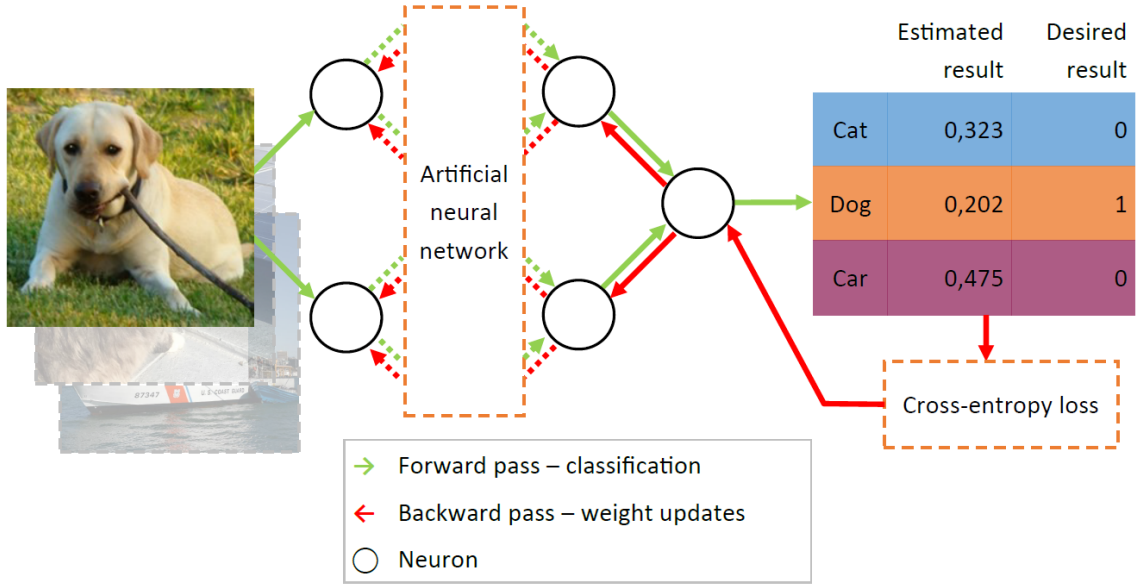
- Network is initialized with random weights
- Input is supervised, i.e. with manually annotated labels
- Define loss function to describe prediction error
- For each training image
  - Calculate Loss
  - Back propagation
    - Update all weights
- Repeat (until Loss is small)





Minimizing the loss function

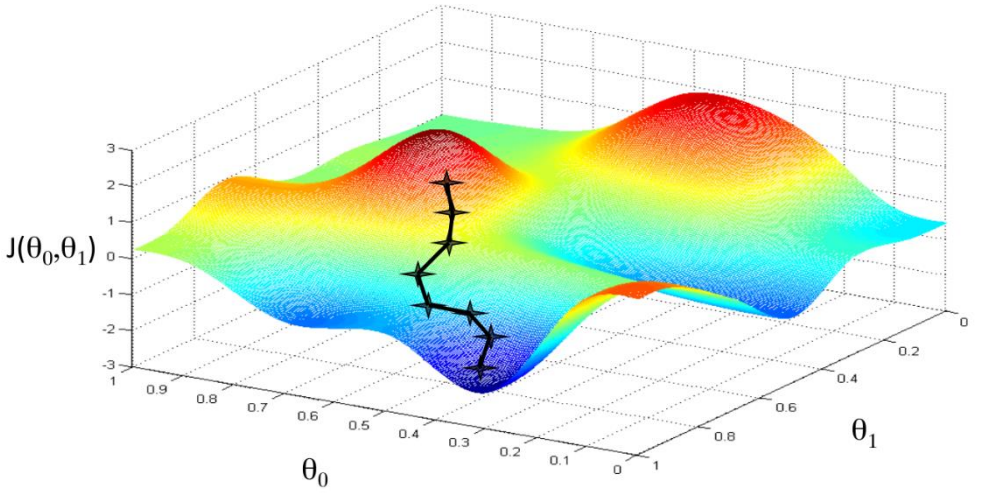
- Finding the values of the weights



AALBORG UNIVERSITY
DENMARK

Minimizing the loss function

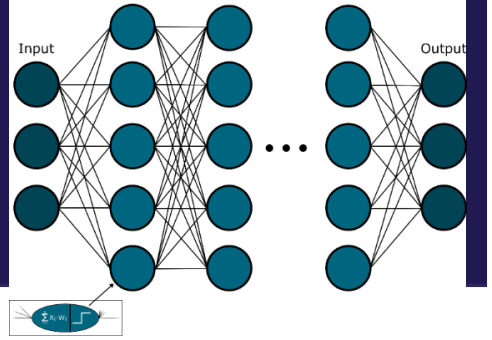
- Brute force
  - Par: 60.000.000
  - Res: 100
  - 100^60.000.000 = INF
- Global approach
  - Follow the slope…
    - Gradient Descent
    - Res: 3
    - 3^60.000.000 = INF
- Local approach
  - One weight at a time…
  - Back propagation

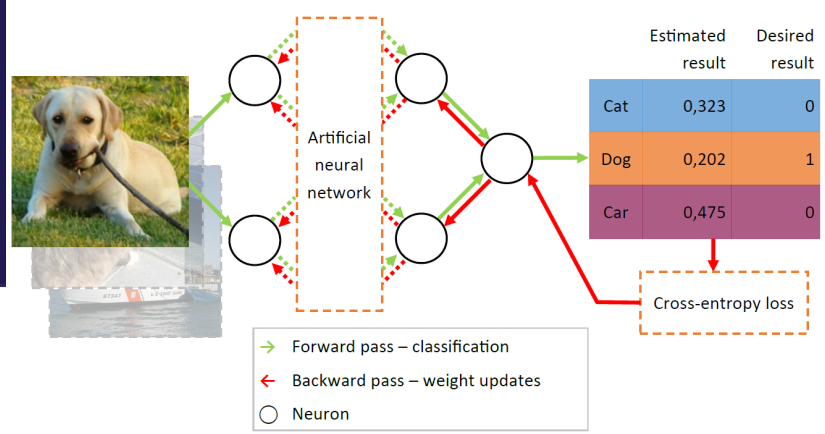


AALBORG UNIVERSITY
DENMARK

Minimizing the loss function

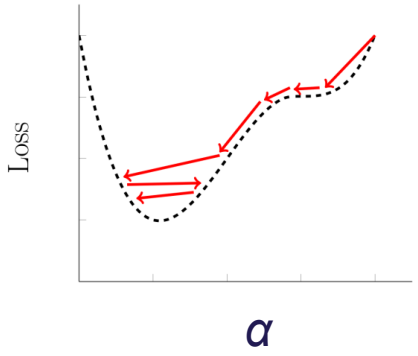
Back propagation



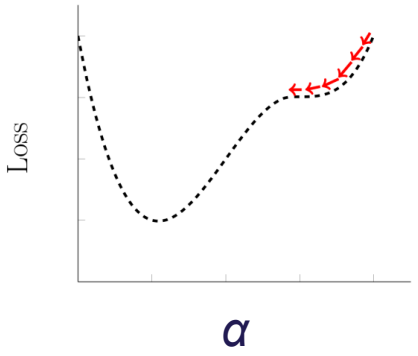


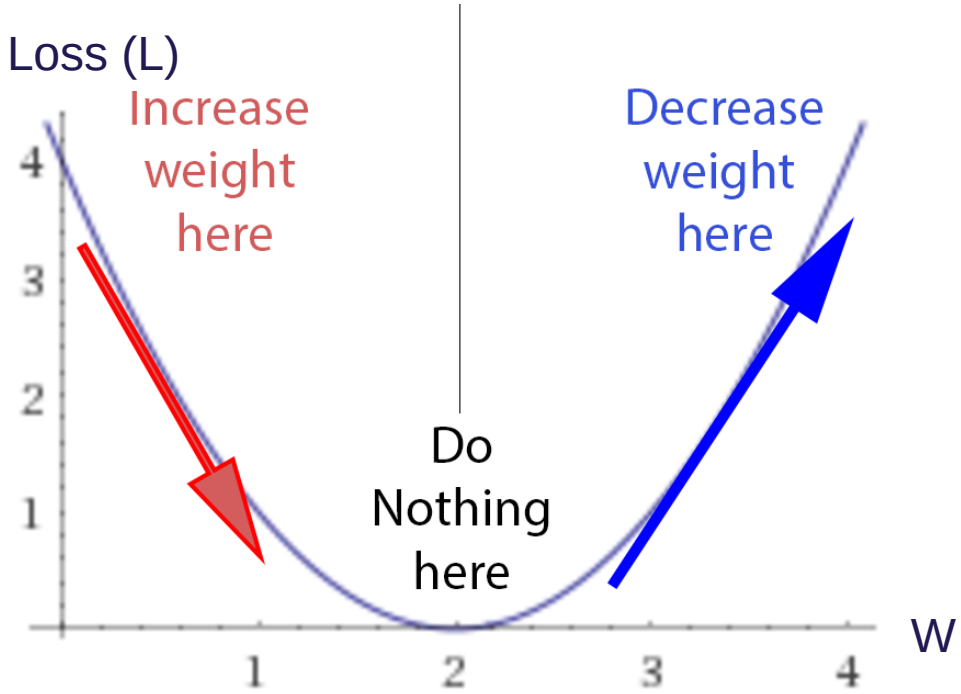
- Imagine we fix all weights, except one: W
- Imagine we know the gradient: $\frac{\partial L}{\partial W}$
- $W_{New} = W_{Old} - \alpha \cdot \frac{\partial L}{\partial W}$
- α is the Learning Rate

High Learning Rate



Low Learning Rate

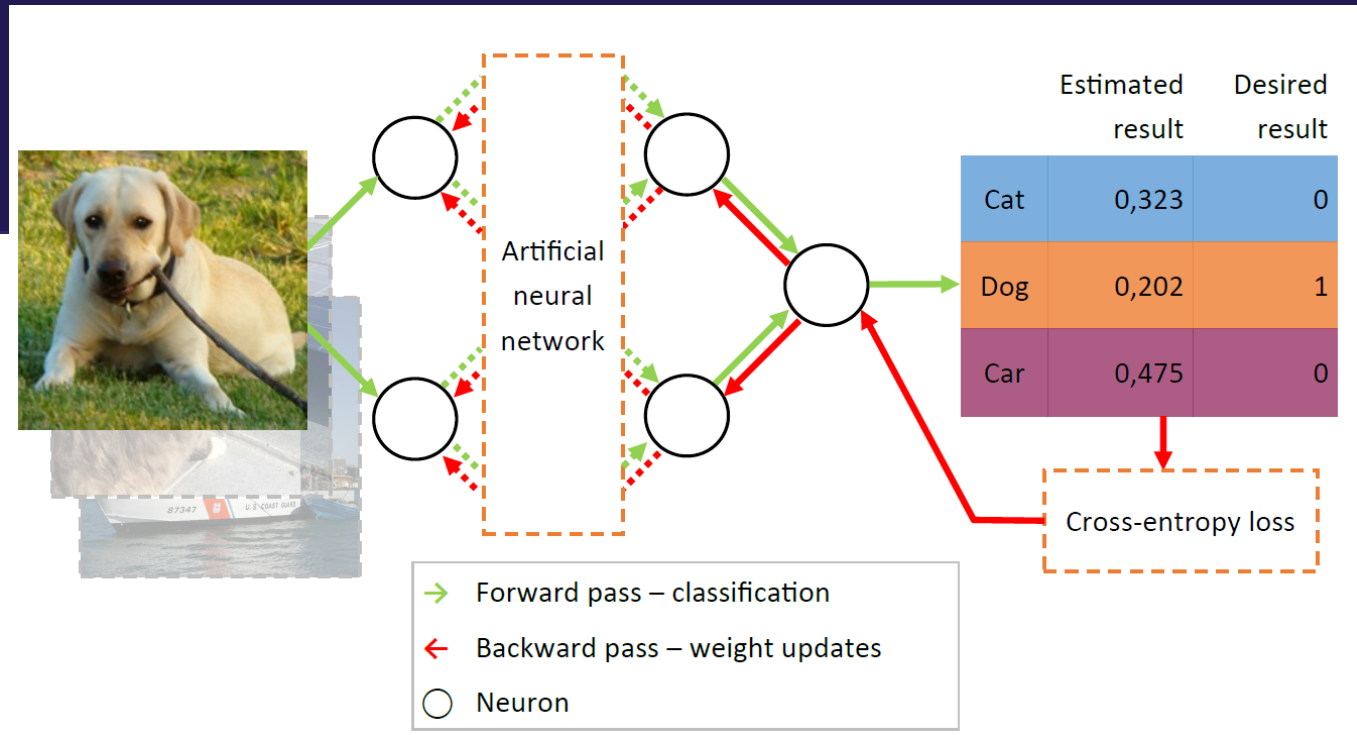




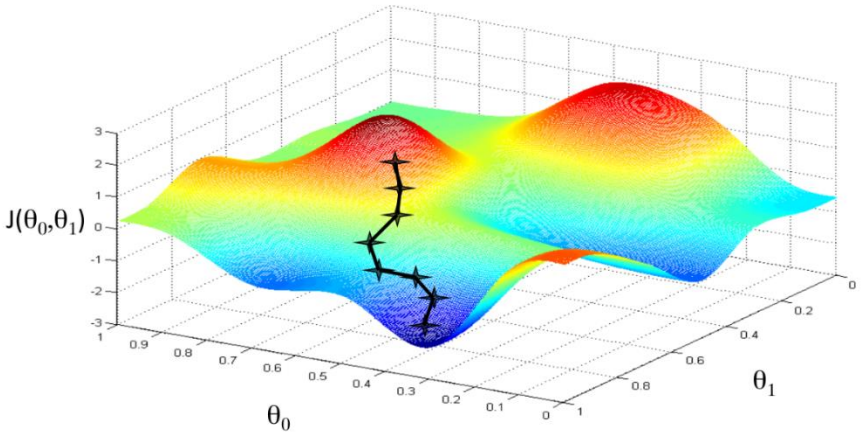
AALBORG UNIVERSITY
DENMARK

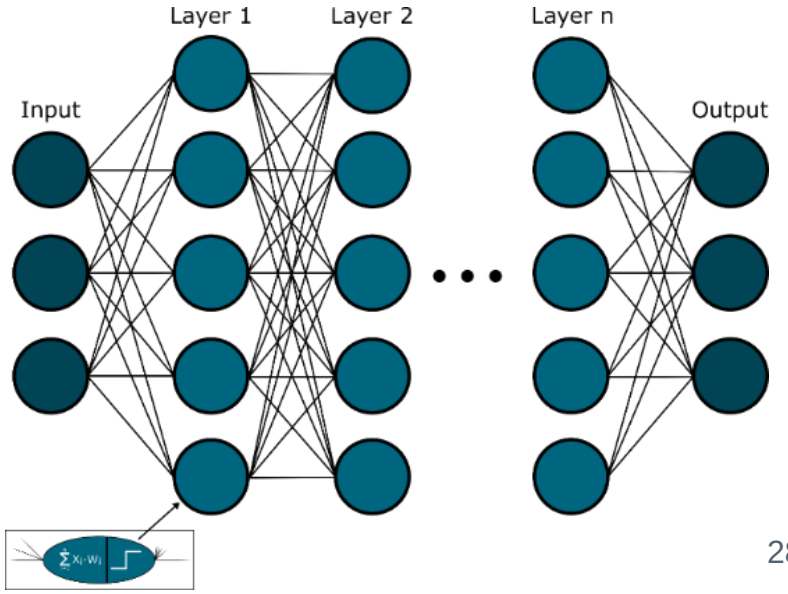
Deep Learning

How does it work?



- Initialize weights
- For each training image
  - Forward pass
  - Calculate Loss
  - Back propagation
    - Update all weights
- Repeat (until Loss is small)

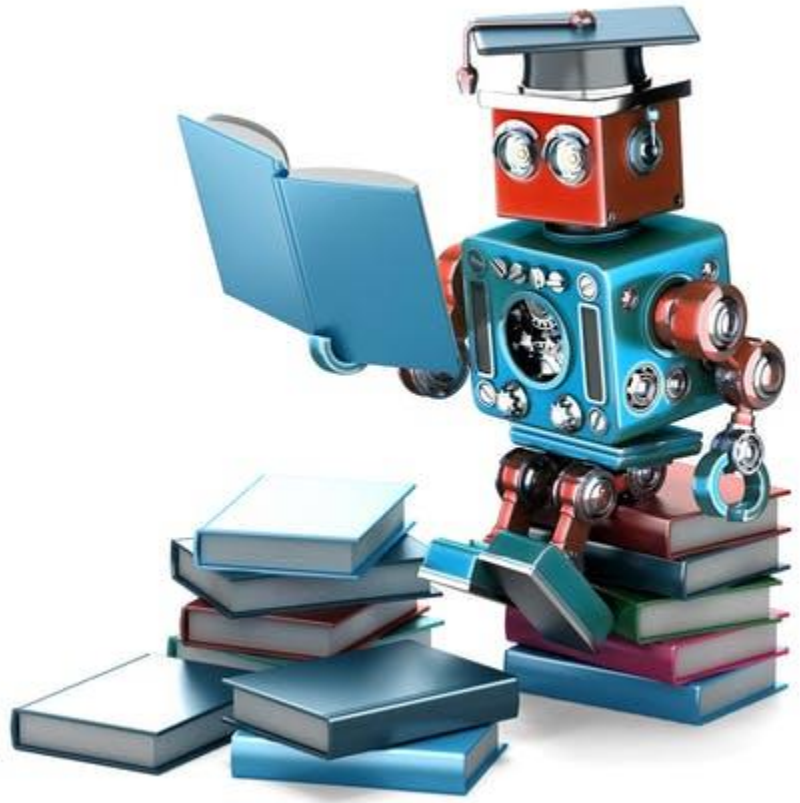




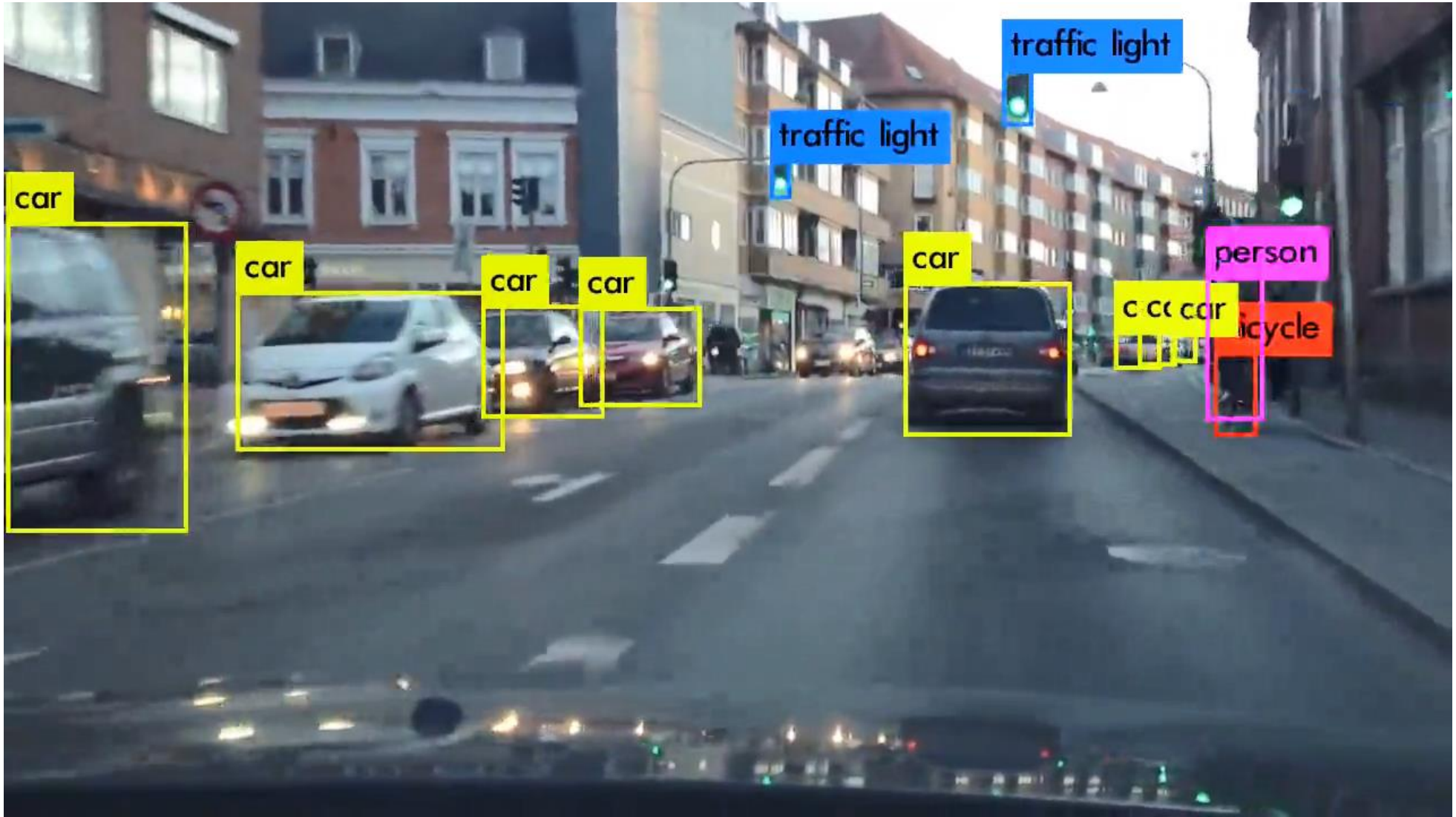
AALBORG UNIVERSITY
DENMARK

Agenda

- Who am I?
- What is AI?
- Deep Learning
- Problems with AI
  - XAI
- Will the machines take over the world?
- Q&A



It works… !



It works… !

- 1997: IBM’s Deep Blue beats Garry Kasparov in chess

- 2016: Google’s DeepMind beats Lee Sedol in GO

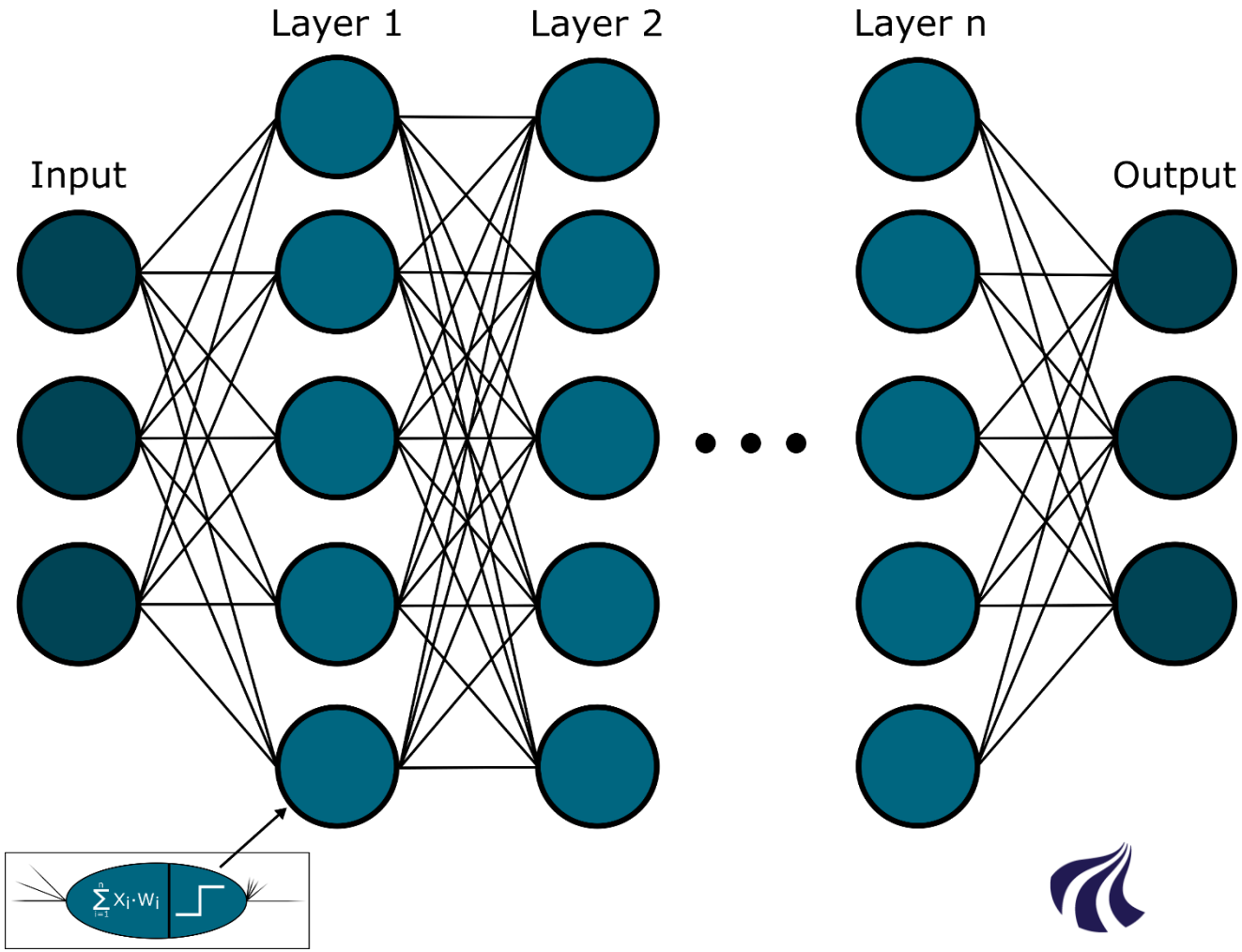
- 2019: Elon Musk’s OpenAI beats world’s top pro team in Dota 2

It works… !

- … and will therefore be used in many places…

Deep learning

Artificial neural network



Neurons = 800.000

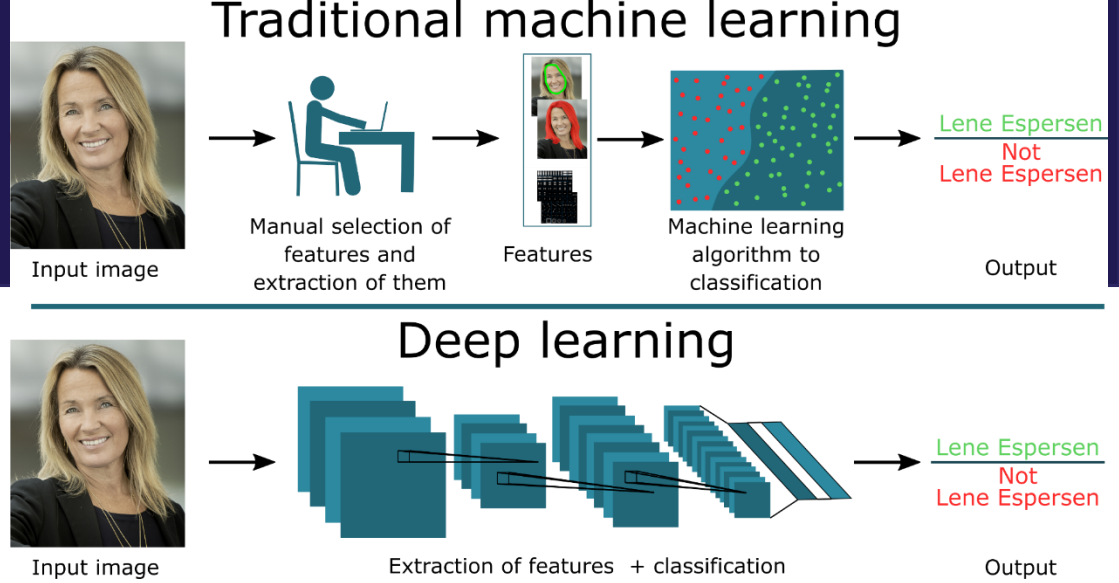
Weights = 120.000.000.000

Free parameters = 60.000.000

Our brain: 100.000.000.000 neurons

AALBORG UNIVERSITY
DENMARK

The Black box



- Deep Learning - It works ☺
- But with 60.000.000 parameters it becomes a black box

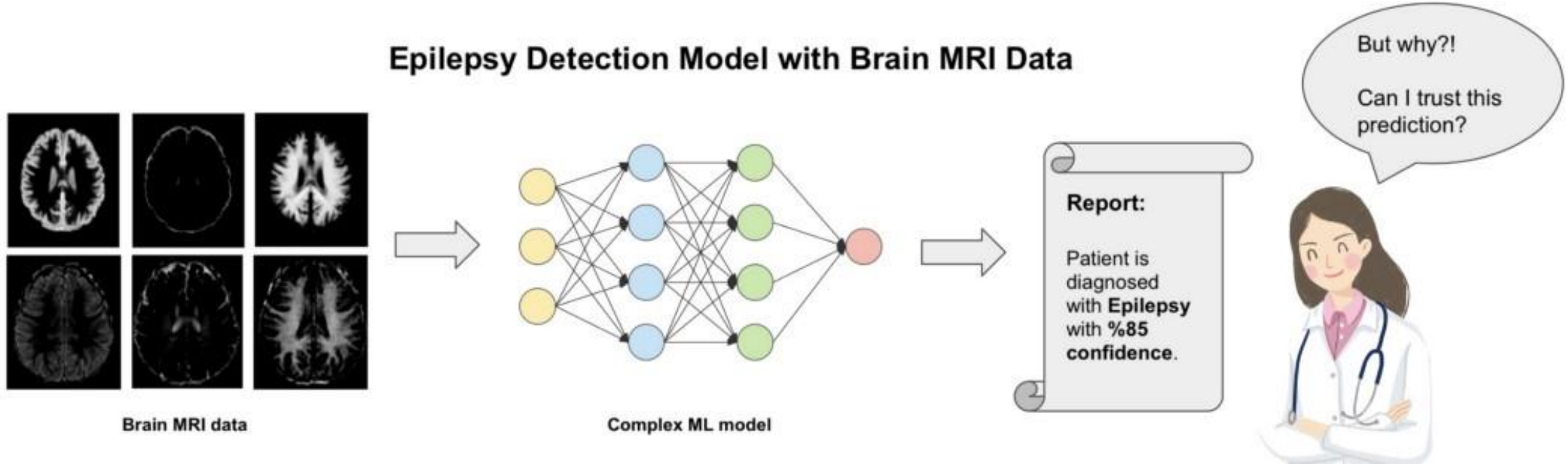
- What if we *WANT* to know?

- **XAI research: Black box ➡ Glass box**

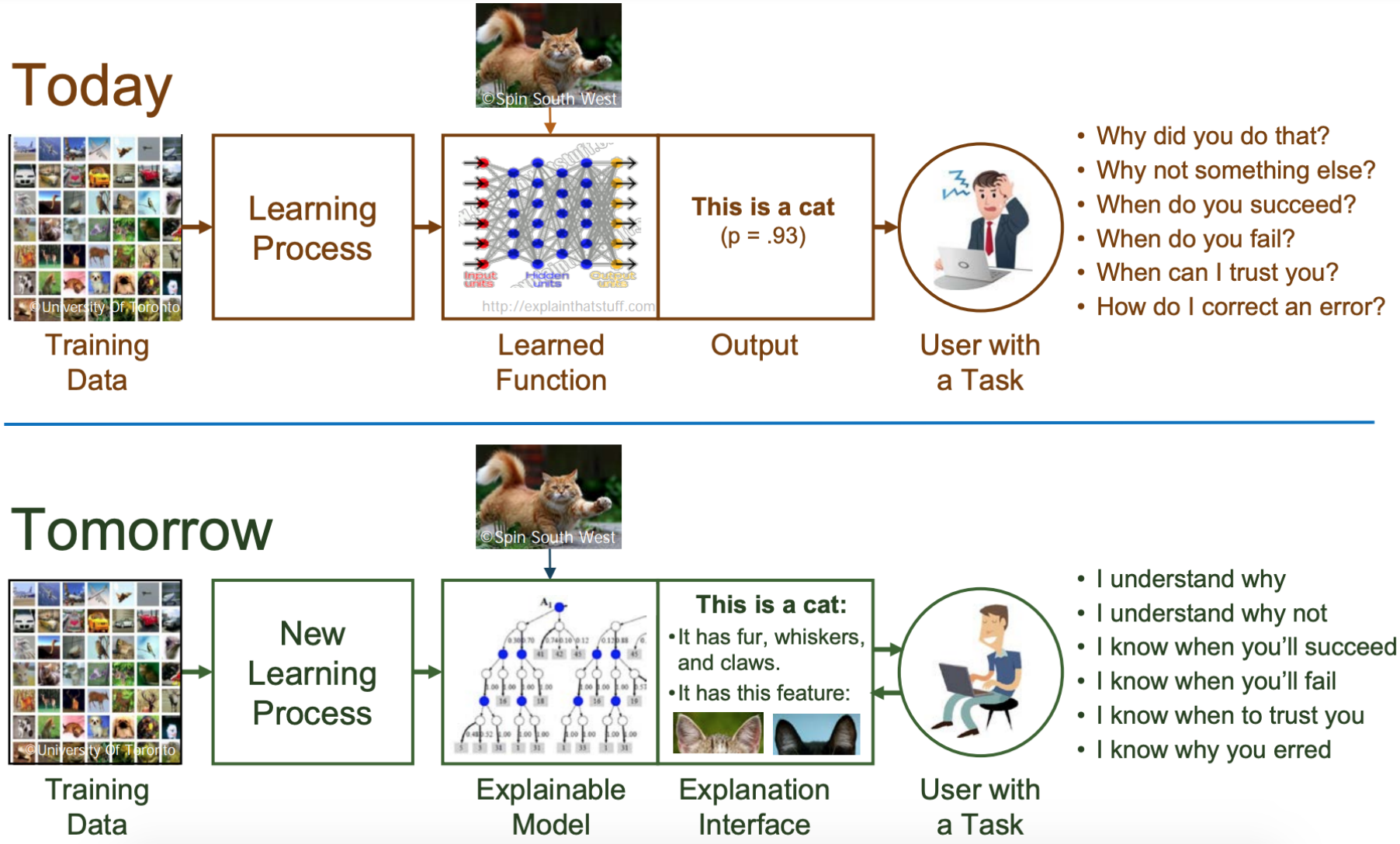
AALBORG UNIVERSITY
DENMARK

The black box

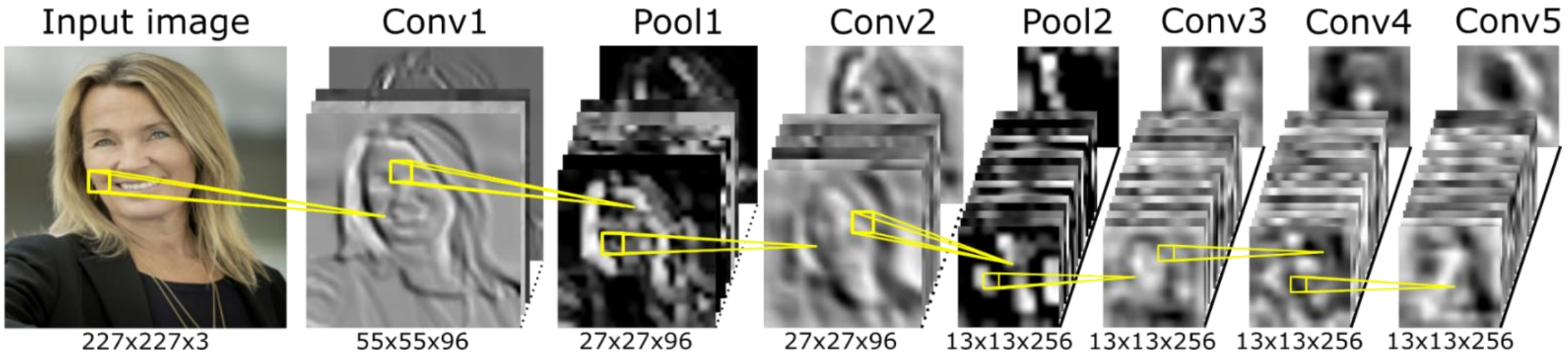
- **XAI research: Black box ⇒ Glass box**

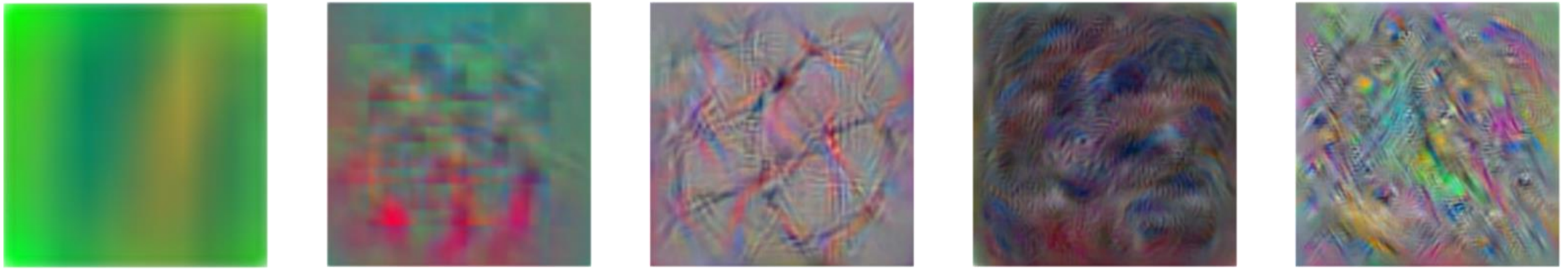


Black box to Glass box



The Black box





Other motivators for XAI…

- Problem
  - Unbalanced training data
  - Unfair

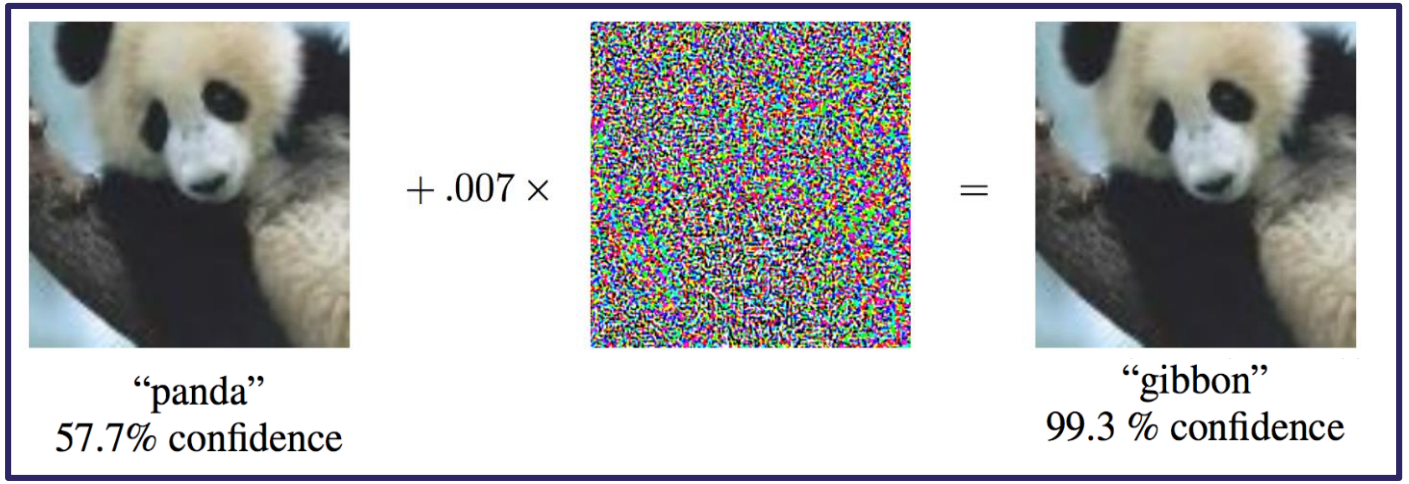
Other motivators for XAI… Fake images

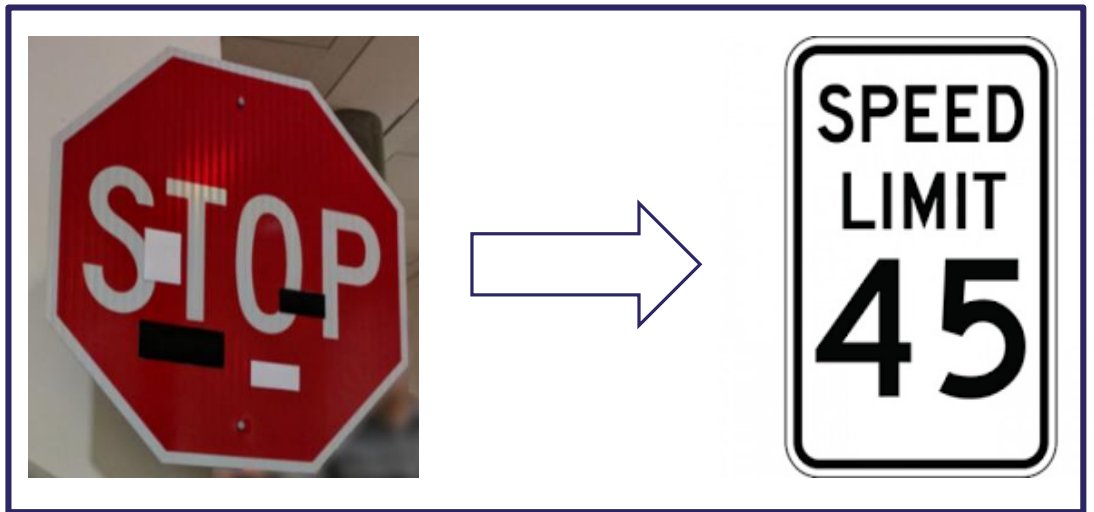
Other motivators for XAI… Fake images

Other motivators for XAI… Style transfer

Other motivators for XAI…

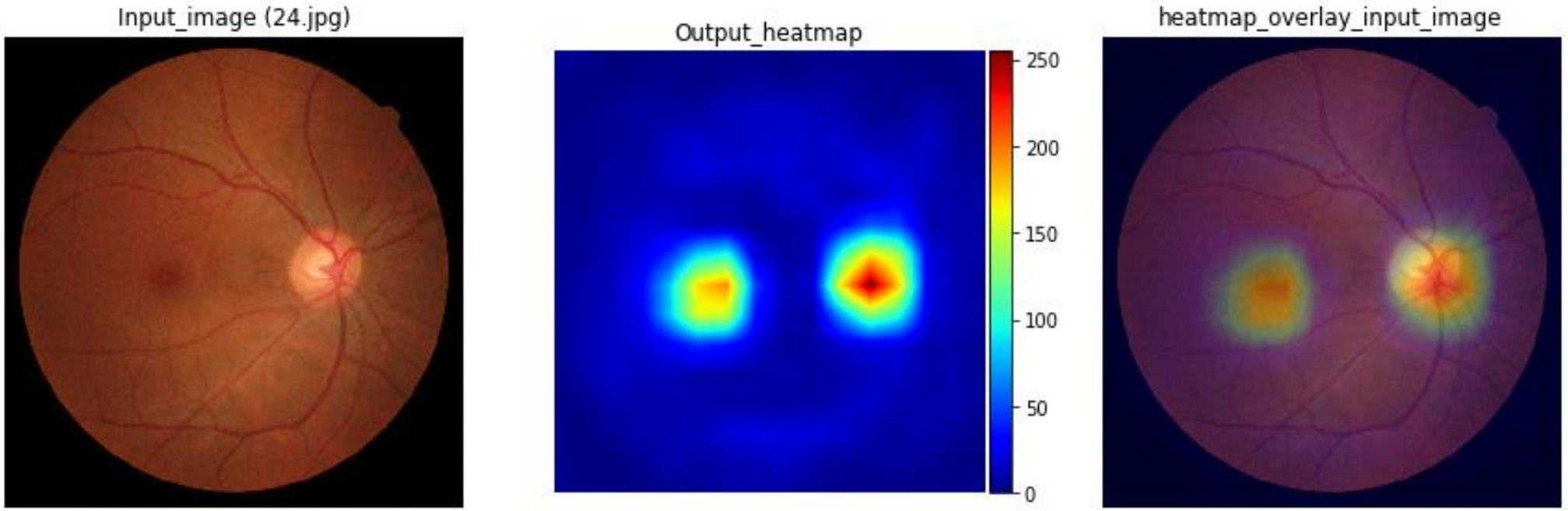
- AI hacking
  - (Adversarial attacks)
- Safety





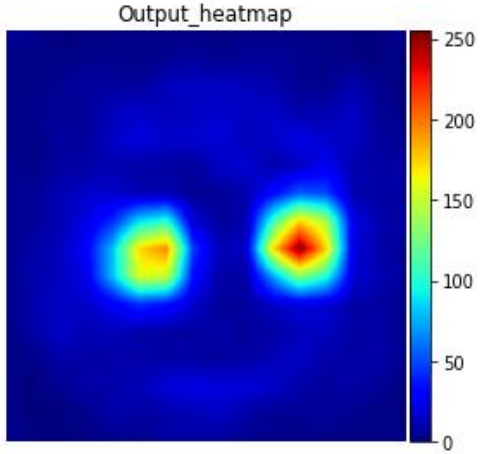
Other motivators for XAI…

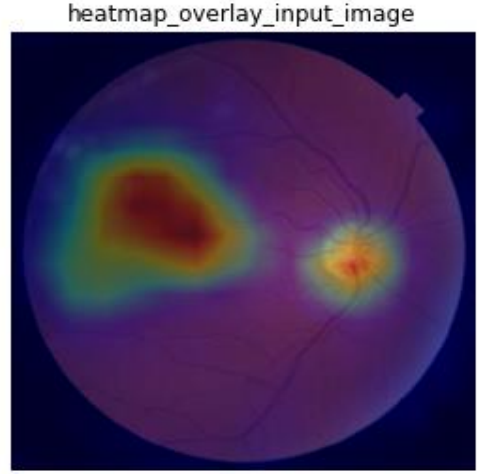
Technical research in XAI



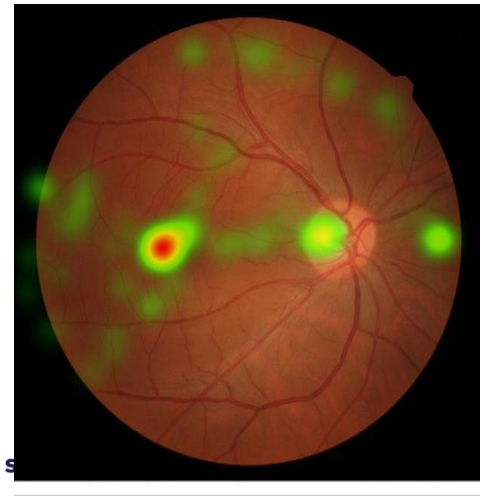
Technical research in XAI

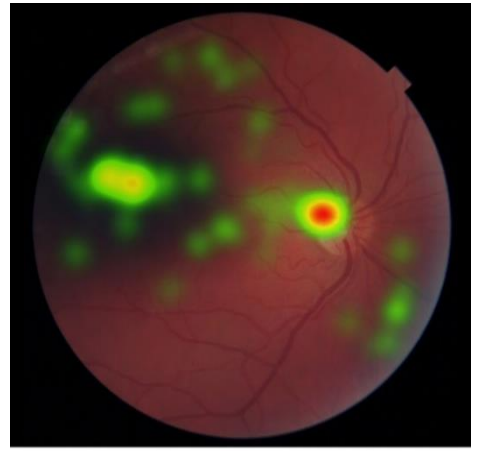
AI algorithm:





Eyetracking:





AALBORG UNIVERSITY
DENMARK

XAI - National level

- Denmark should have a **common ethical and human-centred basis for artificial intelligence**
- Danish researchers should research and develop artificial intelligence
- Danish businesses should achieve growth through developing and using artificial intelligence
- The public sector should use artificial intelligence to offer world-class services

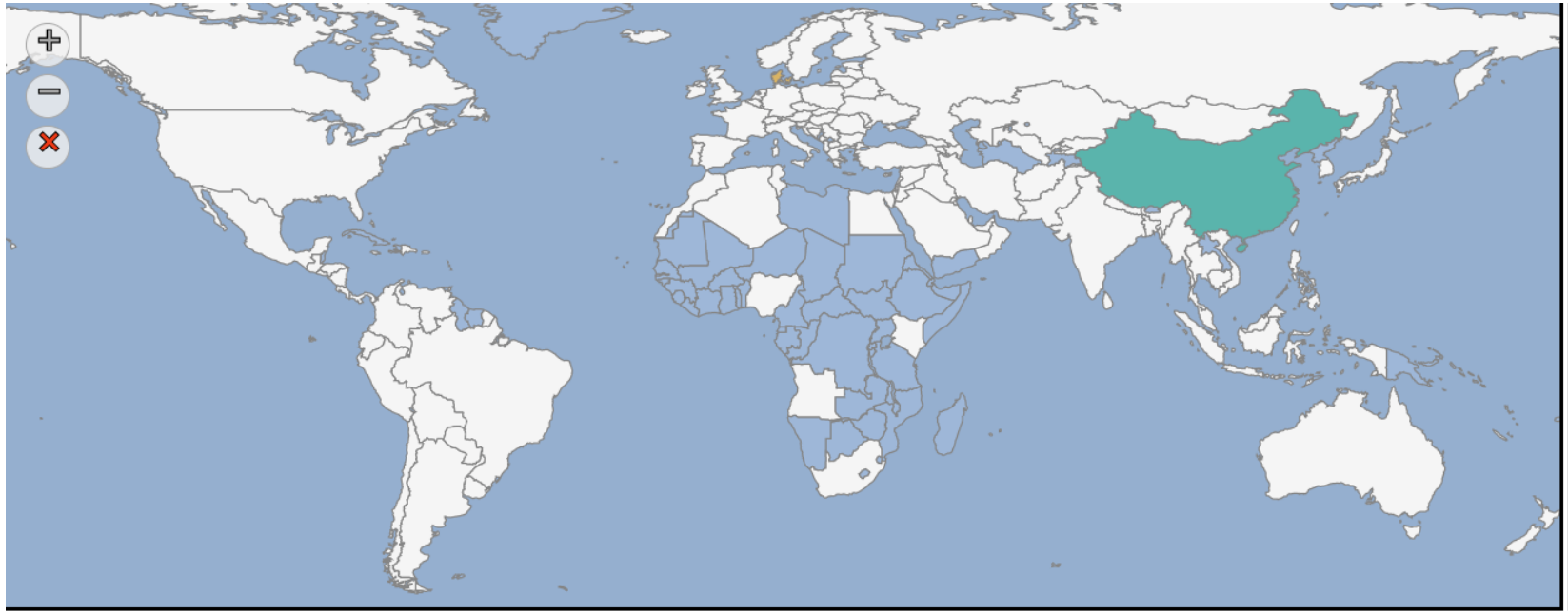
XAI - EU level

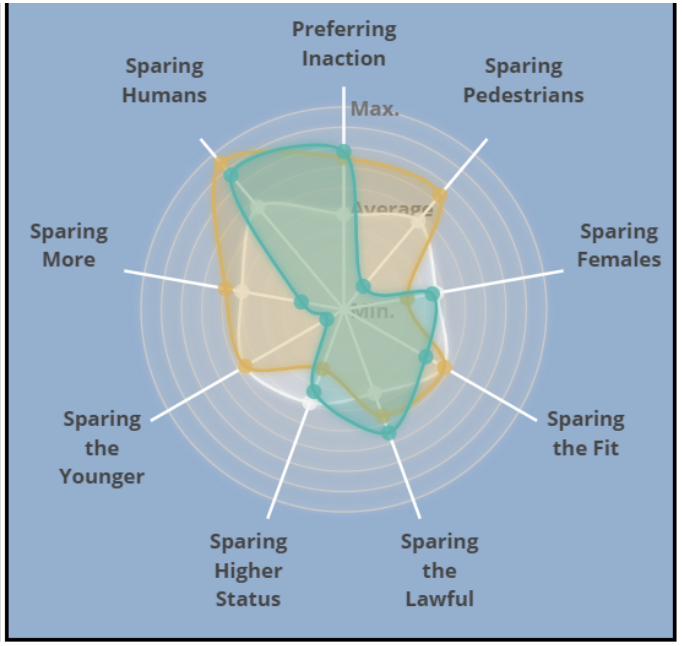
European Commission's High-Level Expert Group on Artificial Intelligence released a set of guidelines for trustworthy AI. The guidelines can be broken into seven categories:

- Human agency and oversight
- Robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

XAI & Ethics…

Moral machine: https://www.moralmachine.net/





The gray area is the world average.

Denmark is most similar to **Norway**, and most different from **Venezuela**

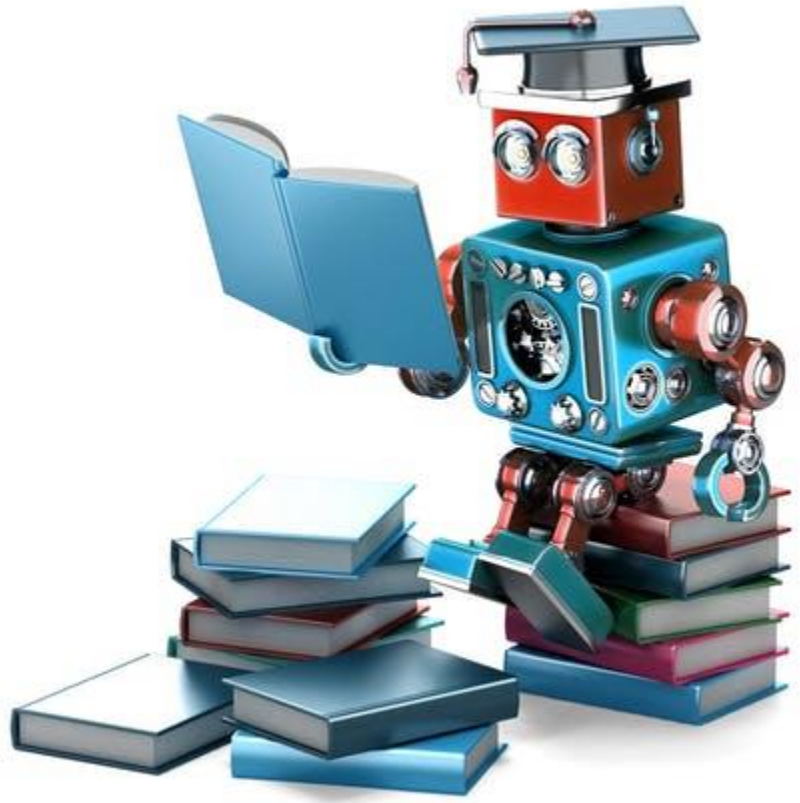
China is most similar to **Thailand**, and most different from **Azerbaijan**

Denmark and China are very different

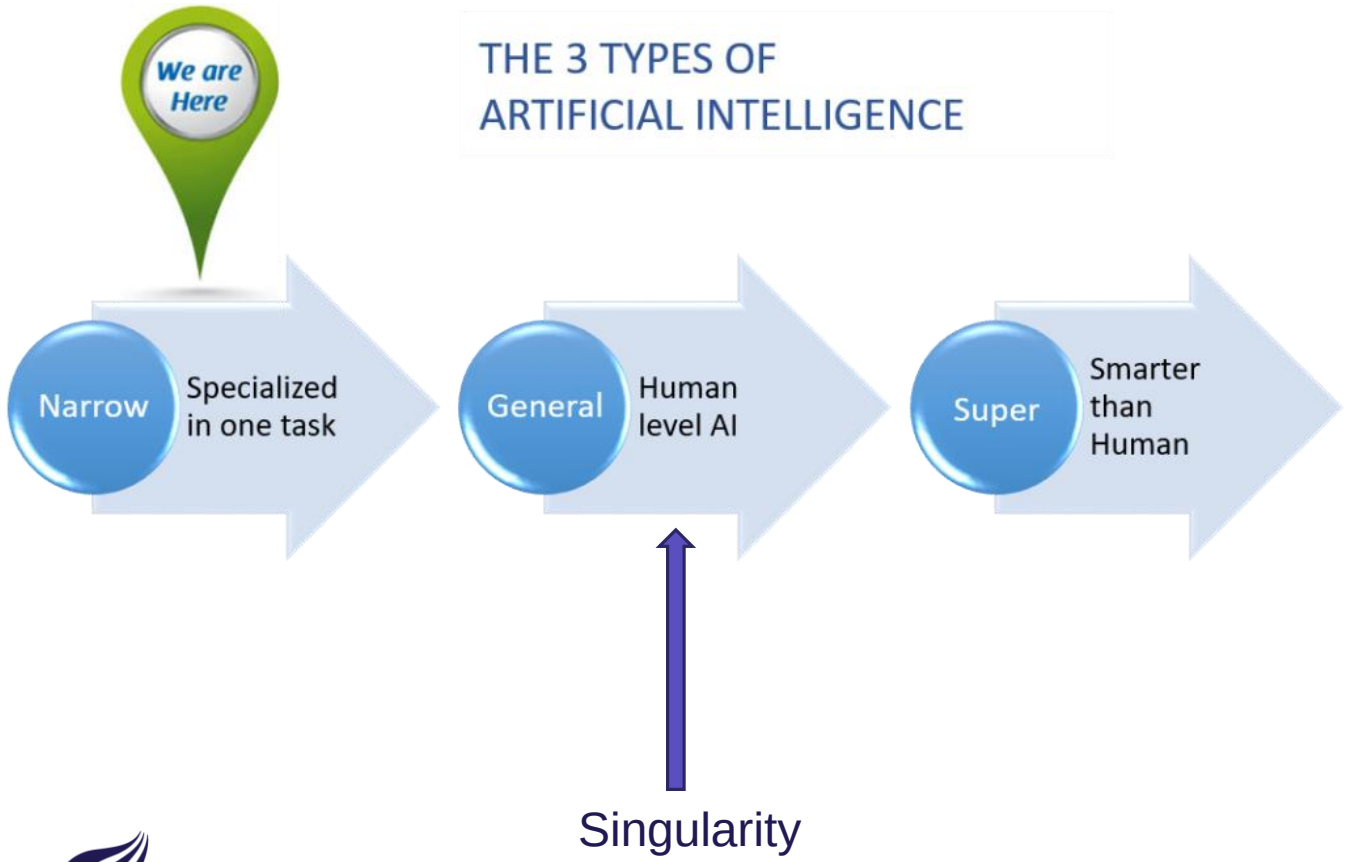
World Ranking (out of 117 Countries)	Preferring Inaction	Sparing Pedestrians	Sparing Females	Sparing the Fit	Sparing the Lawful	Sparing Higher Status	Sparing the Younger	Sparing More	Sparing Humans
Denmark	8th	16th	102nd	60th	28th	100th	67th	35th	4th
China	6th	116th	61st	85th	9th	74th	115th	113th	14th

Agenda

- Who am I?
- What is AI?
- Deep Learning
- Problems with AI
  - XAI
- Will the machines take over the world?
- Q&A

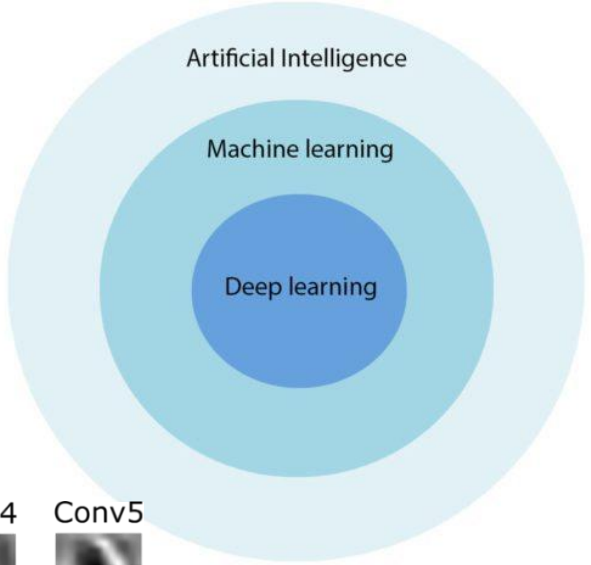


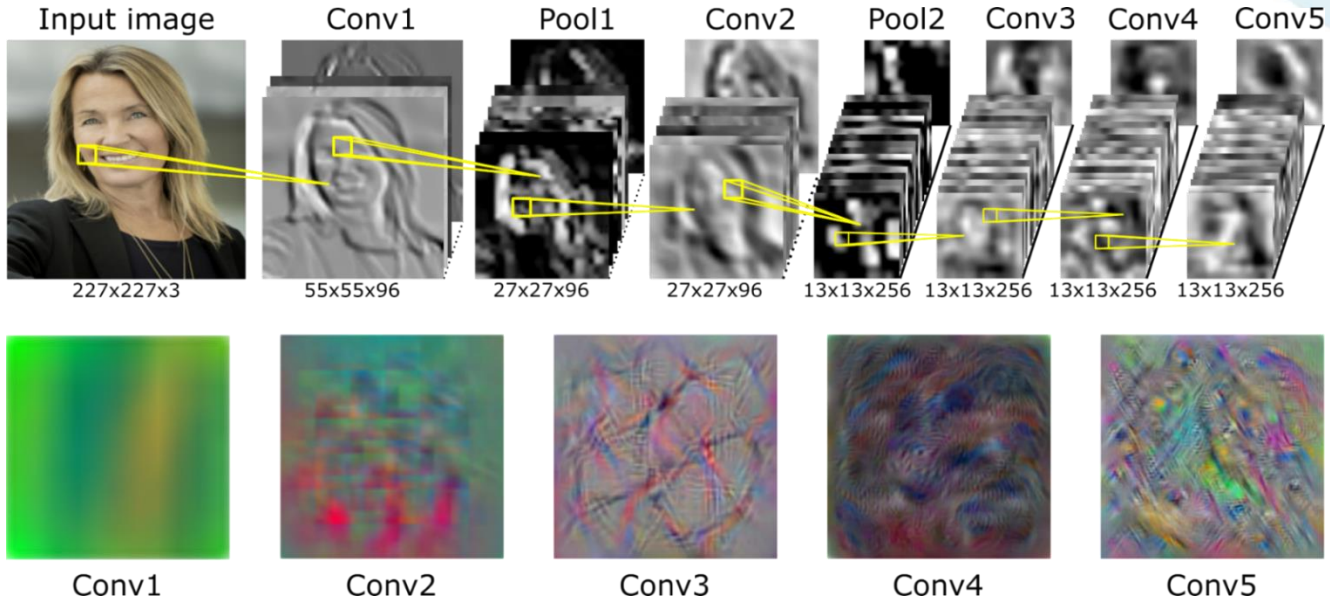
Will machines take over the world?

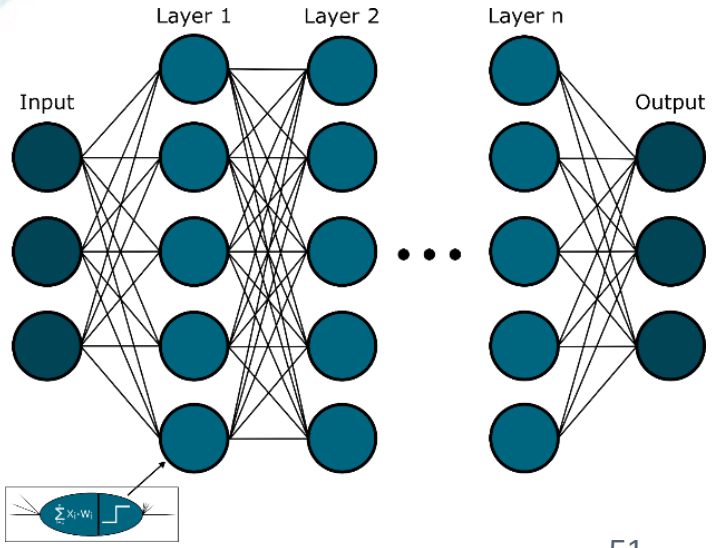


Conclusion

- Deep Learning & XAI are ”hot topics”
- Many free parameters => ”black box”
- More and more applications
- (when) will machines take over the world?







Conclusion

- Deep Learning & XAI are ”hot topics”
- Many free parameters => ”black box”
- More and more applications
- (when) will machines take over the world?