clean-AI

Making Artificial Intelligence Energy-efficient

Prof. Dr. Christoph Meinel
Dean, Institute Director, and CEO
Hasso Plattner Institute, Germany

Why Do We Need AI – Machine Learning?
To Managing the Data Flood …

2020

Every minute **400 hours** of video are uploaded on YouTube

325.000 new malware files every day

35+ billion IoT-devices collect data

2.3+ billion people use smartphones and collect data

Medical scan of a single organ creates **10 GB** of raw data each second

Amazon offers **550+ million** products

Energy-efficient AI

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

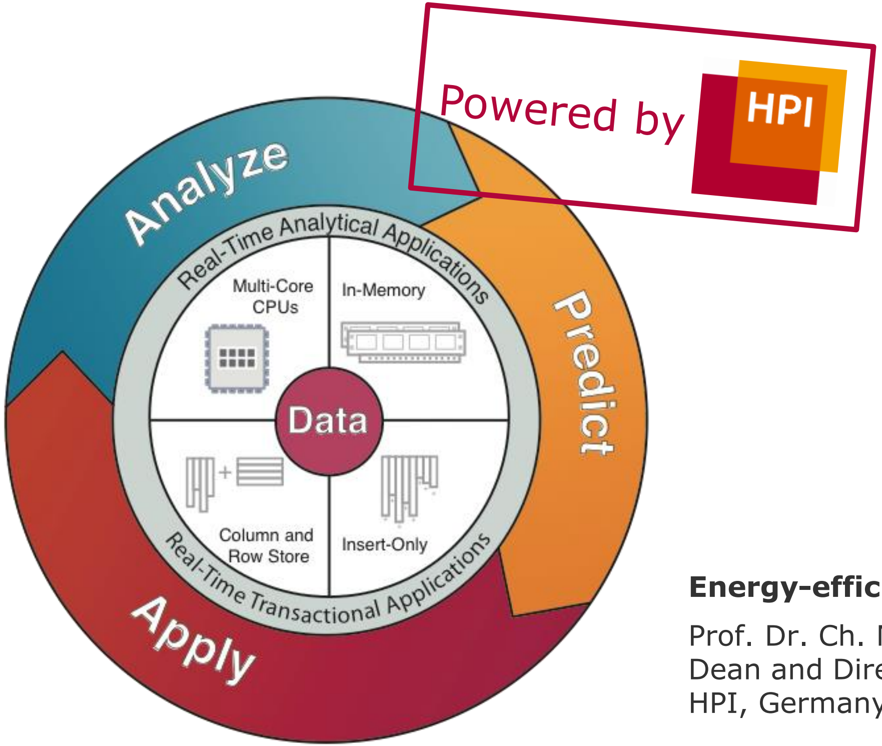
What is Machine Learning?

- Machines learn from data “without” programming
- Computers become able to “see”, “read”, “listen”, “understand” and “interact”

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Why has Machine Learning been so Successful Lately?

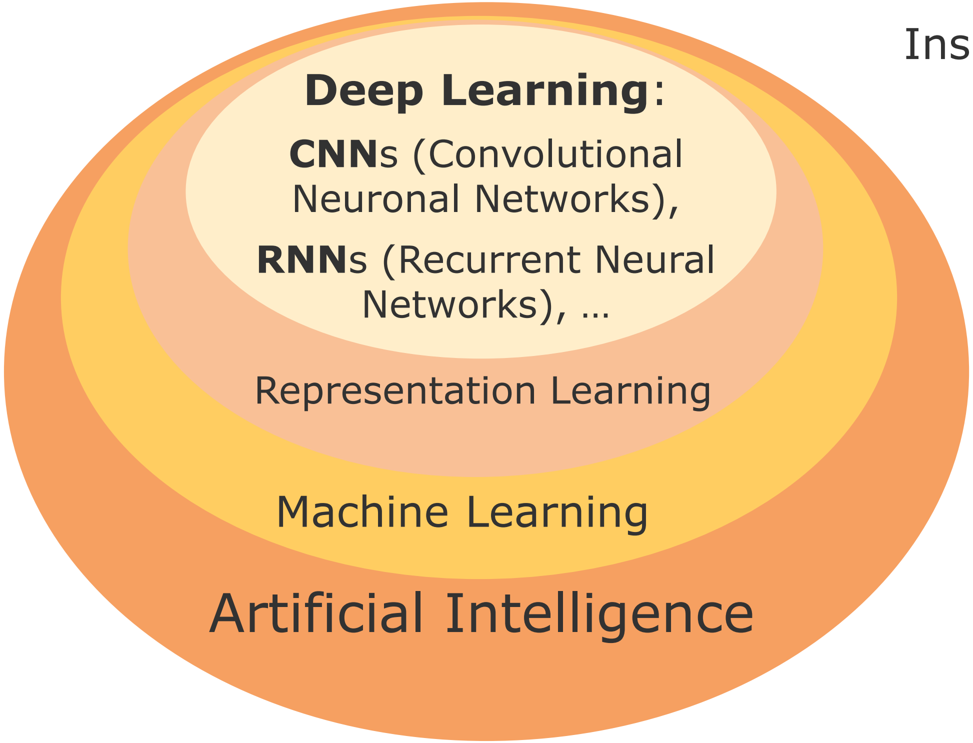
- **Big data** available (Cloud Applications, IoT, Social Media)
- Significant **hardware** improvement in (Multicore, GPU)
- **In-memory data management** for real-time Big Data analysis
- **Cloud computing** (unlimited access to processing power anytime and everywhere)
- **Deep learning algorithms**



Energy-efficient AI

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

But Let´s Start This Story at the Beginning: Artificial Intelligence: What Is **Deep Learning**?



Insights about Deep Learning:

- **Hierarchically** learning features from large scale data
- Machine learning is **data** driven
- Deep Learning progress is driven by **scale**

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Deep Learning – the Current State

Neuronal Networks can...

- classify images extremely well, better than humans [1]
- beat the strongest human players in the game Go [2]
- generate realistic looking images of non-existing people [3]
- and solve a variety of other tasks

Sample from ImageNet Dataset

Generated by thispersondoesnotexist.com

[1] He, Kaiming, et al. "Delving Deep into Rectifiers: Surpassing Human-level Performance on Imagenet Classification." Proceedings of the IEEE International Conference on Computer Vision. 2015.

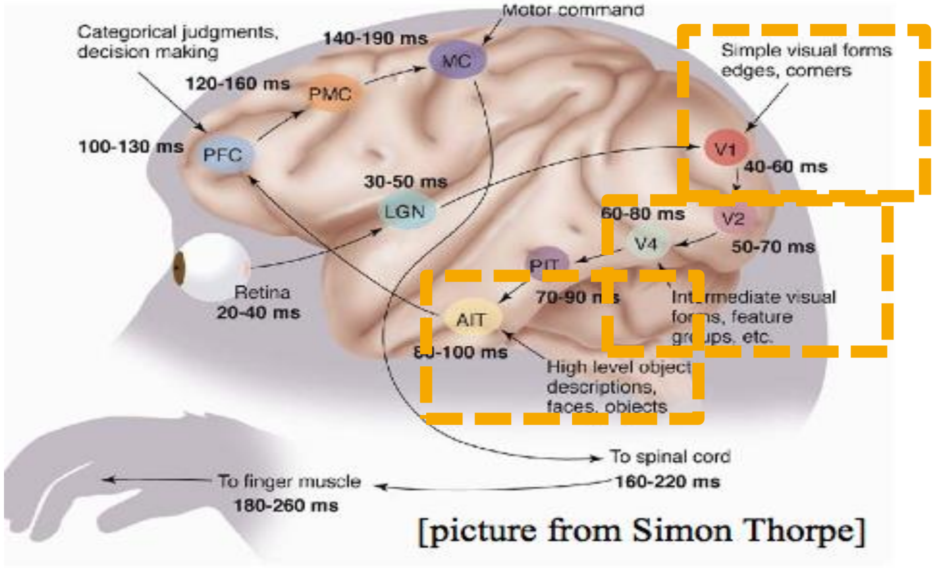
[2] Silver, David, et al. "Mastering the Game of Go without Human Knowledge." nature 550.7676 (2017): 354-359.

[3] Karras, Tero, et al. "Analyzing and Improving the Image Quality of Stylegan." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

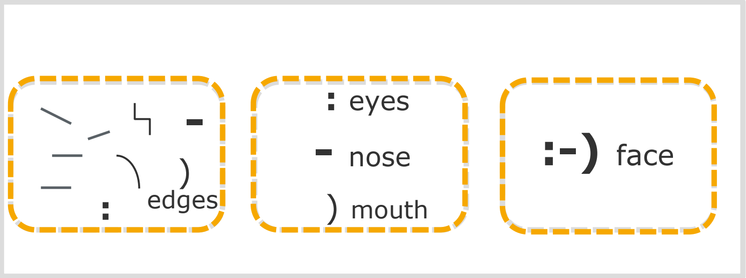
Deep Learning
Inspired by Nature – The Visual Cortex is Hierarchical

The ventral (recognition) pathway in the visual cortex has multiple stages:

- Retina - LGN - V1 - V2 - V4 - PIT..., lots of intermediate representations



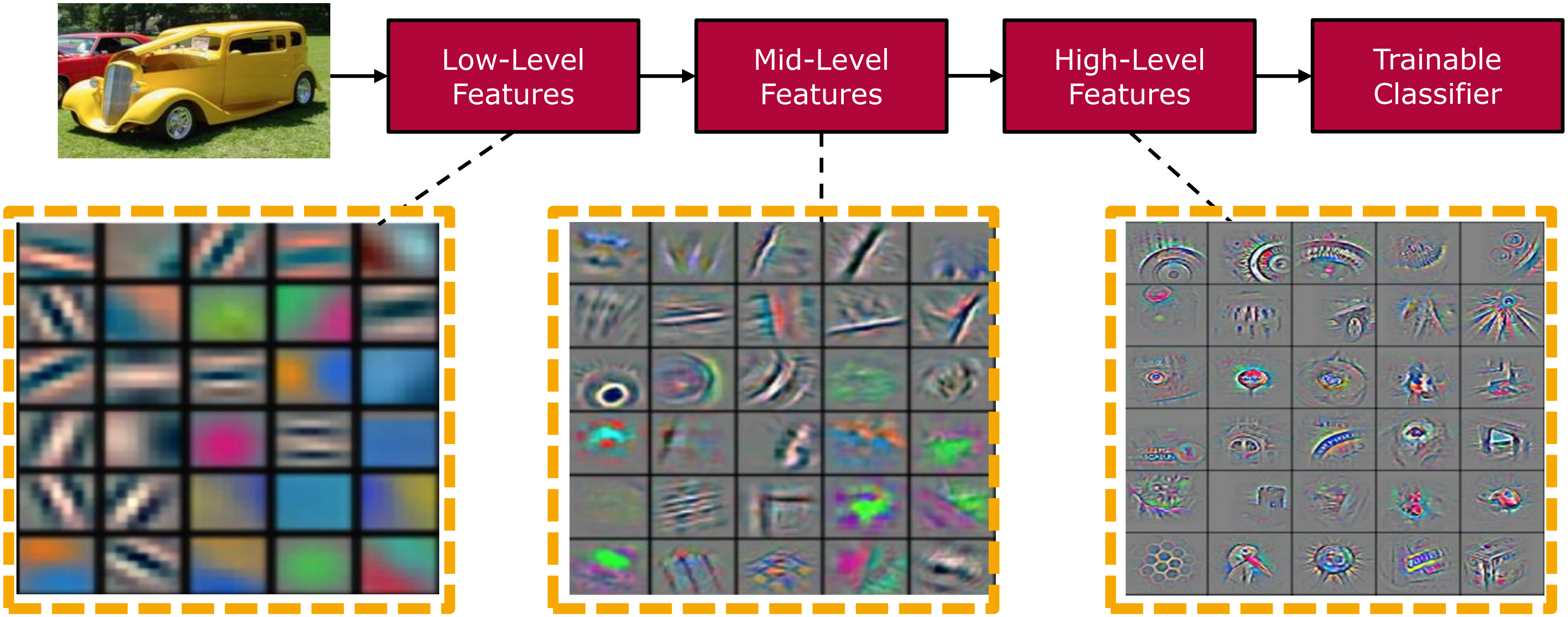
[Gallant & Van Essen]



Energy-efficient AI

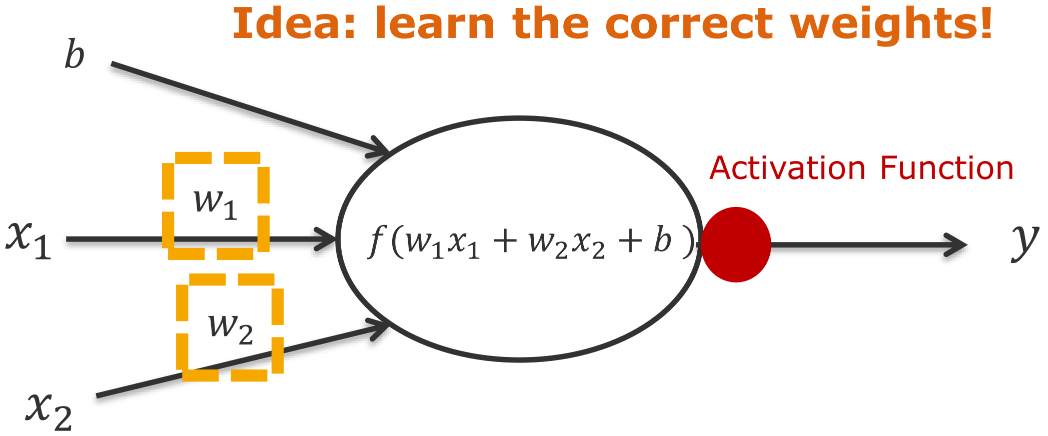
Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Deep Learning is Hierarchical Feature Learning: Learning Higher Abstractions



(Zeiler and Fergus 2013)

Basic **Artificial Neuronal Network**
A Perceptron with an Activation Function



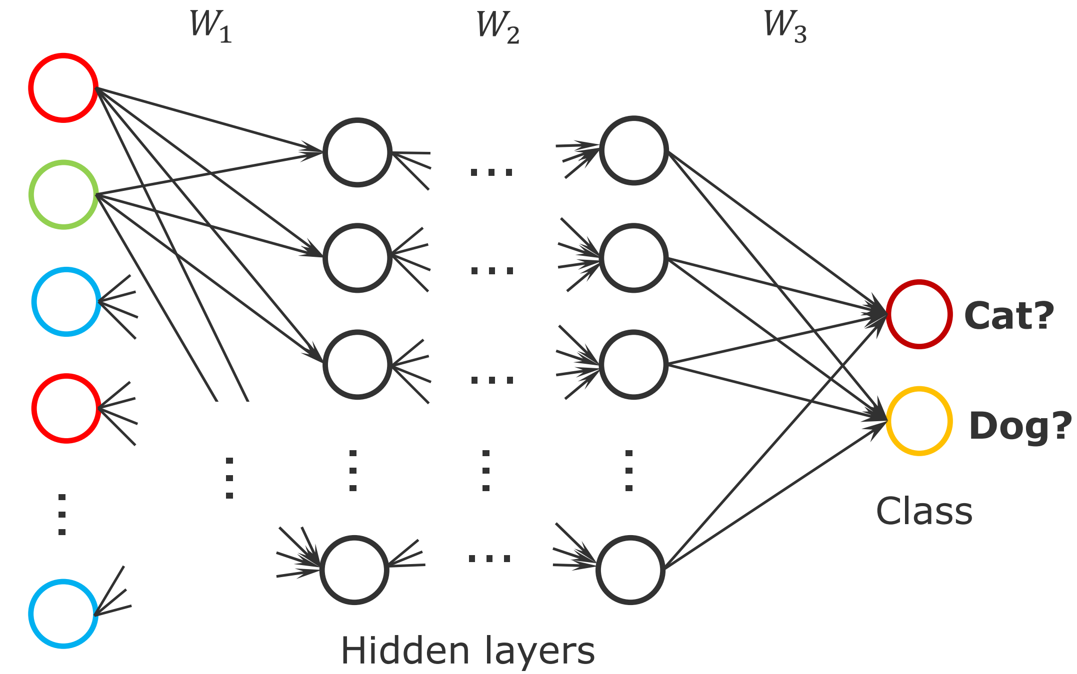
Output of neuron $y = f(w_1x_1 + w_2x_2 + b)$

$$y = \sum w_i x_i + b \quad (convolution)$$

Energy-efficient AI

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Basic **Artificial Neuronal Network**
A Neural Network with More Neurons



Chaining (many) simple functions can approximate any desired function


[1]


[2]

Energy-efficient AI

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Basic **Artificial Neuronal Network**
A Neural Network with More Neurons

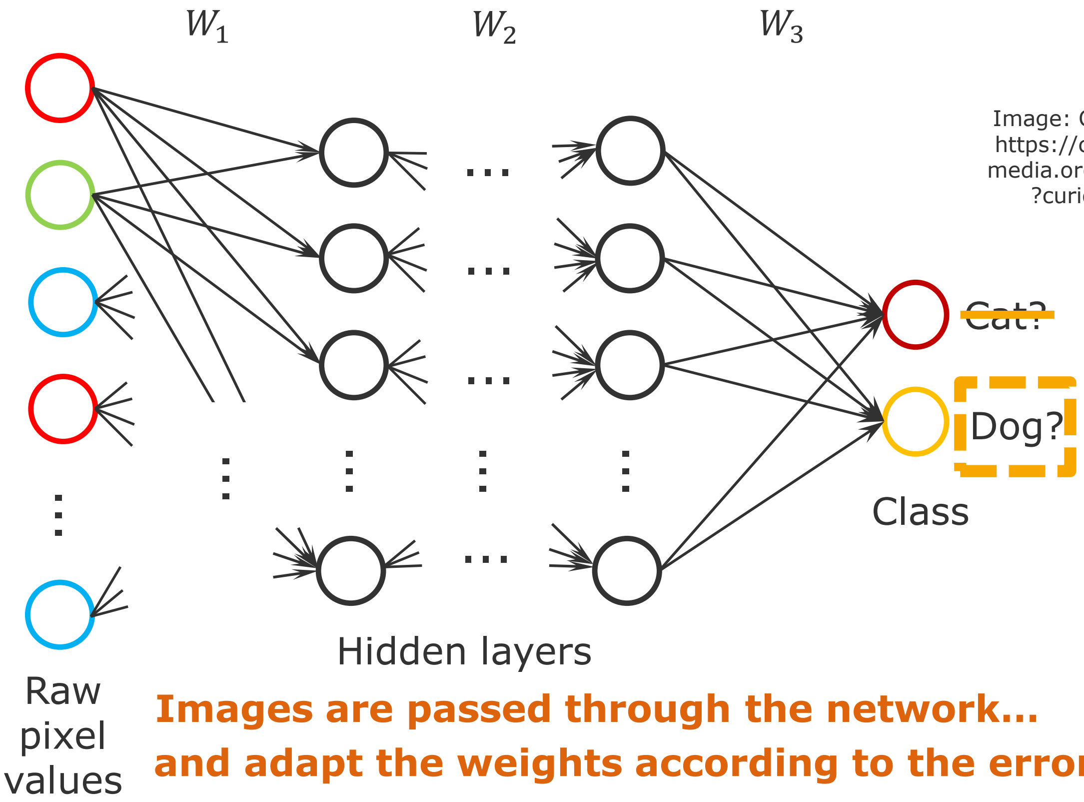


Image: CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=610053

Wrong.

Images are passed through the network...
and adapt the weights according to the error

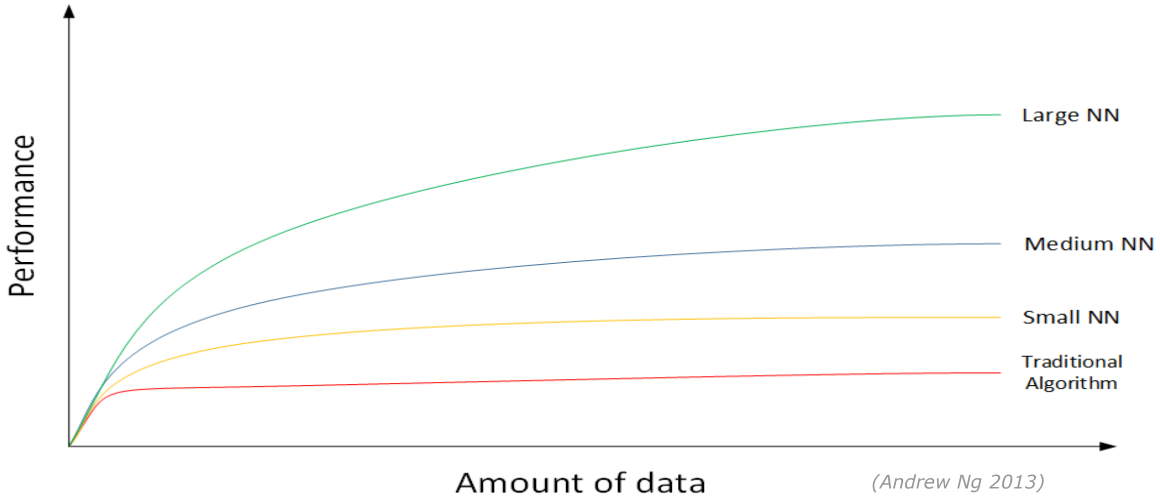
Convolutional Neuronal Networks

- **Convolutions**
  - share weights over the whole image / feature map
  - similar to a sliding window
  - more **efficient** than connecting **all** inputs to **all** outputs
- Thus, networks for images often use Convolutions:
  Convolutional Neuronal Networks – CNNs

Scaling of Convolutional Networks

Networks become **more accurate** with a **larger size** (and **more data**)

- CNNs often have hundreds of layers nowadays



(Andrew Ng 2013)

Deep Learning Models **Spend Lots of Energy**

The **CNN ResNet-152** surpasses Human performance on image classification tasks …

…but needs **240 MB** of storage and **11.3 billion** floating point operations

Therefore such networks need to run on powerful servers in the cloud

Florian Hirzinger – www.fh-ap.com - CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6212634

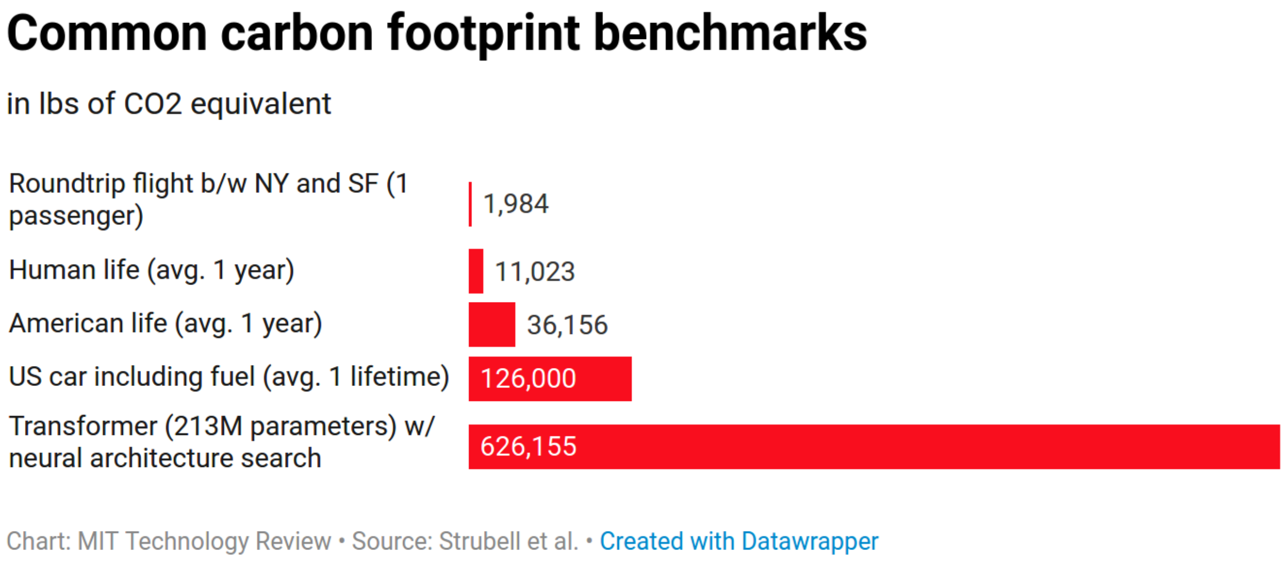
Sample from ImageNet Dataset

Energy-efficient AI

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Deep Learning Models **Spend Lots of Energy**

An extreme example during training of very large CNNs:

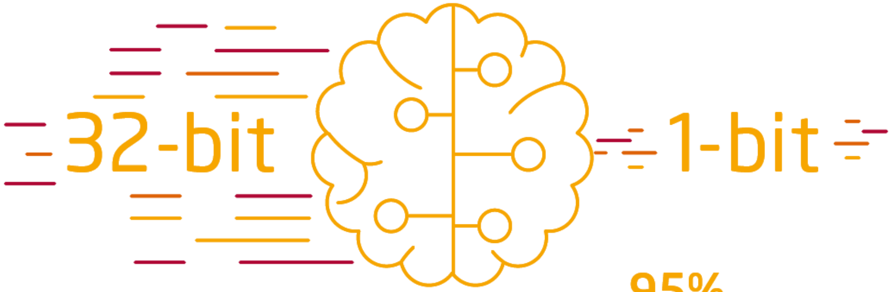


Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. July 2019

HPI clean-IT Initiative – Energy-efficient AI Training: Deep Learning with **Binary Neuronal Networks**

Binary neuronal networks – BNNs

- State of the art deep neuronal networks are trained and operate on 32-bit models
- Design and Training of deep neuronal networks on binary-level (1-bit) is possible



95%
electricity saving

Energy-efficient AI

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Reduced Energy Requirements Save the Environment … and Allow to Run AI Models on Mobile Devices

To save the environment **energy-efficient** deep learning models are needed

How to reduce energy requirements?

- Lower number of operations while raising their energy-efficiency
- Lower memory requirements

Potential:

- If deployed on a large scale: huge **energy savings**
- Models can be run on **mobile and embedded devices**

https://commons.wikimedia.org/wiki/File:Raspberry_Pi_3_B%2B_(39906369025).png

Prof. Dr. Ch. Meinel
Dean and Director
HPI, Germany

Approaches to Achieve More Energy Efficient Deep Learning Models

Knowledge distillation

- distills a large model (teacher) into a small model (student)

Network pruning techniques

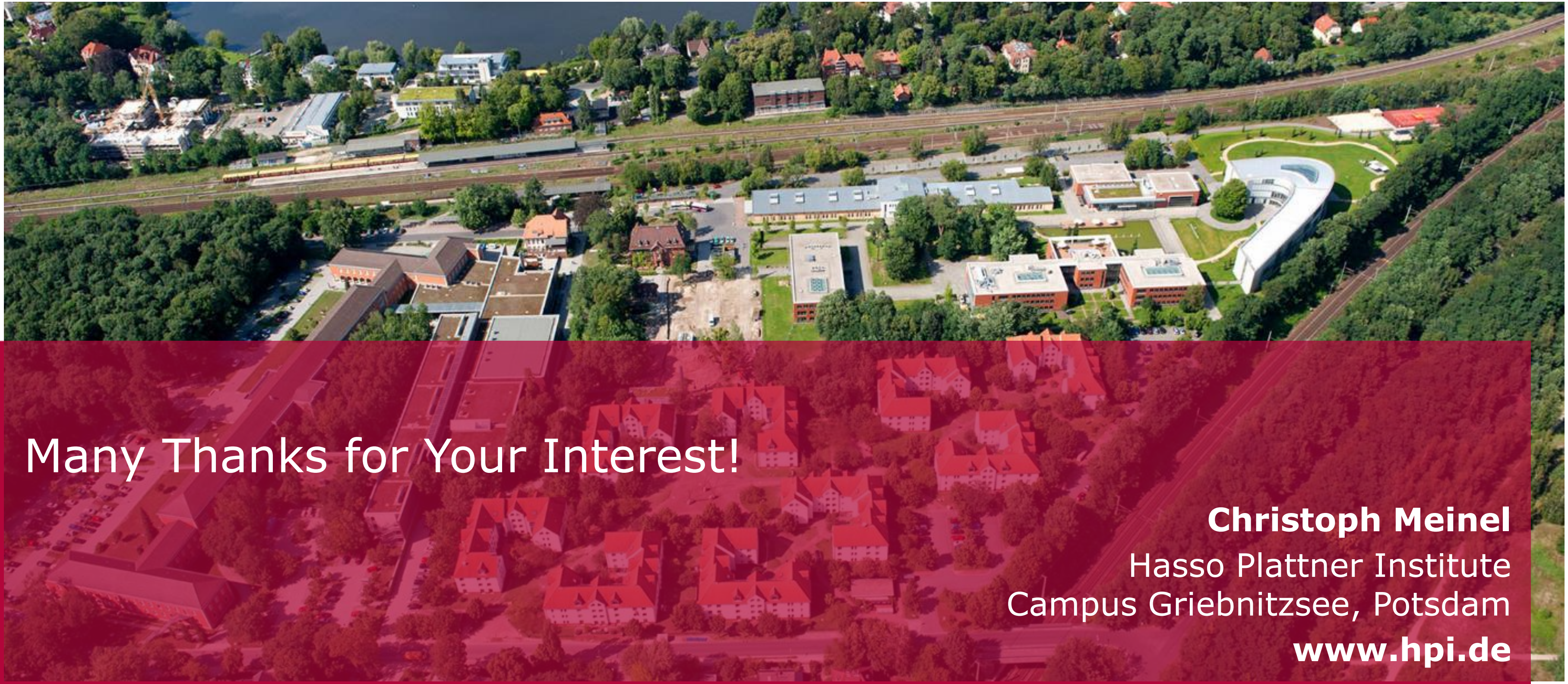
- remove non-essential weights

Compact network designs

- use layer structures with less weights and operations

Low-bit quantization

Quantizes 32-bit floating point weights to a lower bit-width, e.g. 2-bit: +1, +0.3, -0.3, -1



Many Thanks for Your Interest!

Christoph Meinel
Hasso Plattner Institute
Campus Griebnitzsee, Potsdam
www.hpi.de