Christof Teuscher

ECE 410/510: Hardware for AI and ML

Hardware for AI and ML: an Overview

Portland State University
Department of Electrical and Computer Engineering (ECE)
www.teuscher-lab.com
teuscher@pdx.edu

teuscher Lab teuscher-lab.com

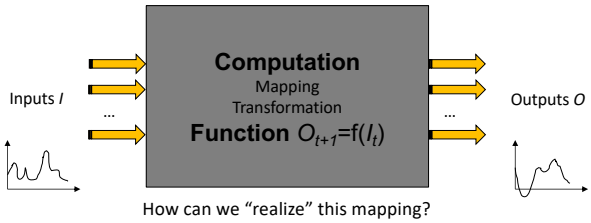
Portland State UNIVERSITY

What is a computer?

teuscher Lab teuscher-lab.com

Portland State UNIVERSITY

What is computation? (1)

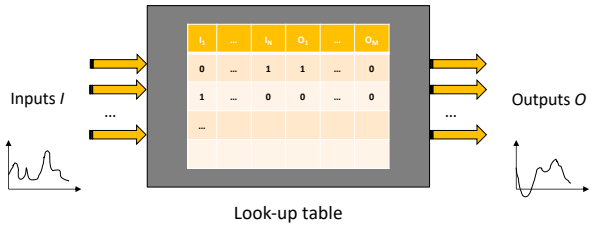


How can we "realize" this mapping?

teuscher Lab teuscher-lab.com

Portland State UNIVERSITY

What is computation? (2)

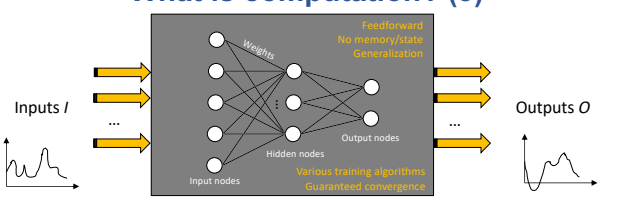


Look-up table
Simple, fast

teuscher Lab teuscher-lab.com

Portland State UNIVERSITY

What is computation? (3)

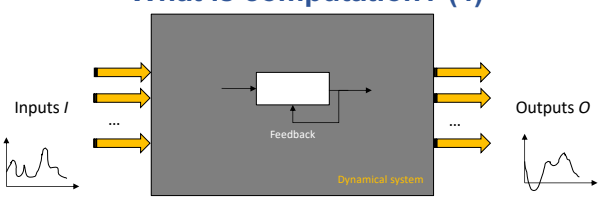


Neural network
Universal function approximator
(Hornik, Stinchcombe, White, 1989, https://doi.org/10.1016/0893-6080(89)90020-8)

teuscher Lab teuscher-lab.com

Portland State UNIVERSITY

What is computation? (4)

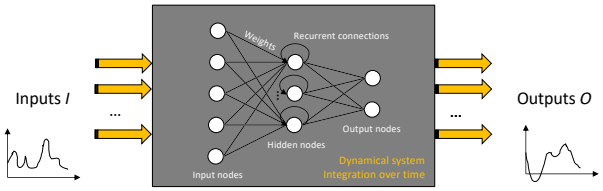


Memory is needed in a lot of cases
Temporal signal processing (audio, video)
Timeseries prediction
Control tasks

teuscher Lab teuscher-lab.com

Portland State UNIVERSITY

What is computation? (5)

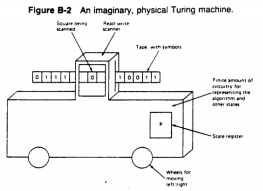


Recurrent neural networks (RNN)

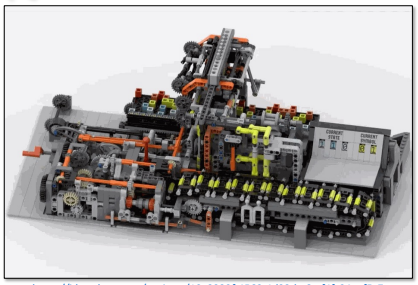
Training is challenging, various algorithms exist (backpropagation through time), convergence not guaranteed, many weights, slow learning. RNNs are also universal function approximators

(Schaefer & Zimmerman, 2007, https://doi.org/10.1142/S0129065707001111)

Turing Machine (TM)



- Turing's aim was to define the simplest possible device that could perform any computation that could be performed by a human.
- TMs are a model of computability, they do not model the physical process of a computer.



https://ideas.lego.com/projects/10a3239f-4562-4d23-ba8e-f4fc94eef5c7

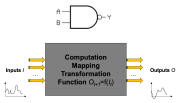
Turing Machine (TM)

"What does a Turing machine look like? You can certainly imagine some crazy looking machine, but a better approach is to look in a mirror."

— Charles Petzold, The Annotated Turing

Intrinsic vs designed computation (1)

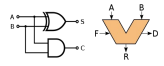
- **Designed computation:**
  - top-down
  - we design devices and circuits for our purpose
  - information processing that is "useful"
- **Intrinsic computation:**
  - inherent dynamics of a dynamical system
  - not necessarily "useful"



What does a falling apple "compute?"

- Everything computes something, the question is what and how it can be harnessed for "useful" purposes.
- *"What natural substrates are complex enough to support controllable and useful information processing?"* (Crutchfield et al., 2010, https://doi.org/10.1063/1.3492712)

Intrinsic vs designed computation (2)



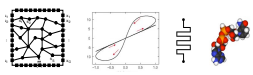
Desired computation

Computing architecture

How to obtain a desired computation?

Intrinsic dynamics

Device network



Intrinsic vs designed computation (3)

"Evolvable hardware" variant

Nanocell Logic Gates for Molecular Computing

James M. Tour, William L. Van Zandt, Christopher P. Husband, Summer M. Husband, Lauren S. Wilson, Paul D. Franzon, and David P. Nackashi

(Tour et al., 2002, http://doi.org/10.1109/TNANO.2002.804744)

- Network trained by switching NDR switches ON/OFF at post fabrication.
- Omnipotence is assumed (one knows and can control all switches)
- Tasks: logic functions, 1 bit-adder
- In simulation only

Self-assembled network of metallic particles, with reprogrammable negative differential resistance (NDR)

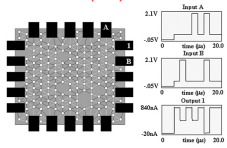
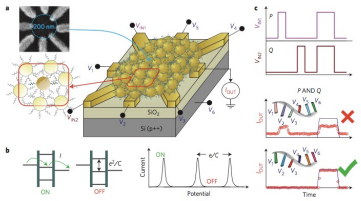


Fig. 5. NAND gate is demonstrated in this nanocell using the SPICE interface model and the $I(V)$ curve shown in Fig. 3. The plots show the $V(t)$ and $I(t)$ with the accompanying NAND Truth Table logic.

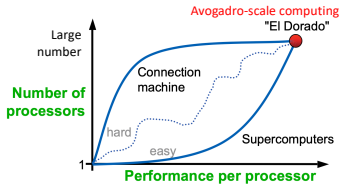
Intrinsic vs designed computation (4)



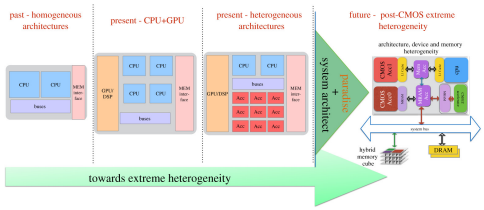
(Bose et al., 2015, https://doi.org/10.1038/nnano.2015.207)

What kind of computing machinery would an extraterrestrial build?

The goals and drivers have changed



The goals and drivers have changed



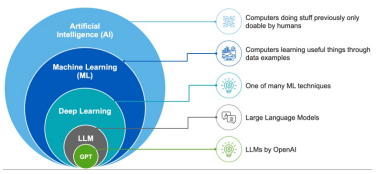
https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061

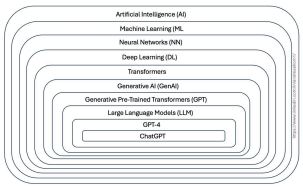
AI vs ML

- **Machine Learning (ML)** is one technique used to achieve **Artificial Intelligence** (AI).
- All ML is AI, but not all AI is ML.
- AI is the destination (e.g., intelligent machines), while ML is one path to get there (learning from data).

- **Artificial Intelligence (AI):** Broader concept encompassing machines that can perform tasks requiring human intelligence. Ultimate goal: Artificial General Intelligence (AGI)
- **Machine Learning (ML):** Subset of AI focused specifically on algorithms that improve through experience. Example: neural networks

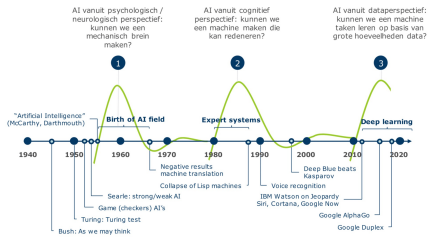
AI vs ML





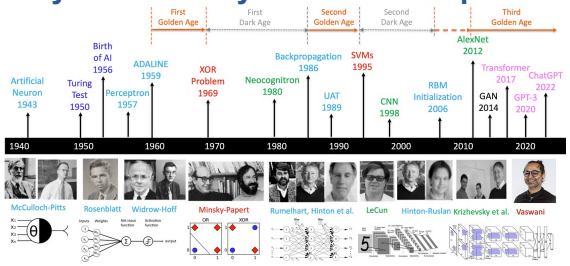
Source: Tignis

A very brief history of AI



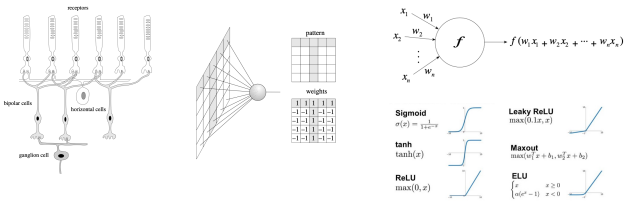
Strong AI = AGI = Artificial General Intelligence

A very brief history of AI with deep learning



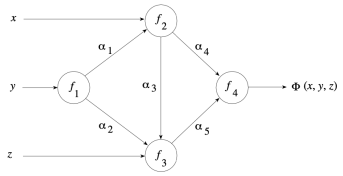
https://medium.com/@lmpo/a-brief-history-of-ai-with-deep-learning-26f7948bc87b

An abstract neuron



Rojas, 1986

A neural network



Rojas, 1986

Special types of neural networks

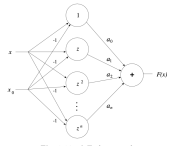


Fig. 1.16. A Taylor network

$$F(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n + \cdots,$$

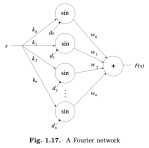


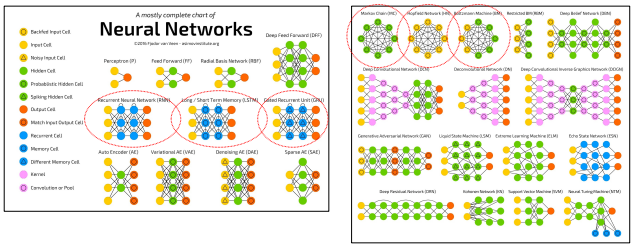
Fig. 1.17. A Fourier network

Figure 1.17 shows how a Fourier series can be implemented as a neural network. If the function F is to be developed as a Fourier series it has the form

$$F(x) = \sum_{i=0}^{\infty} (a_i \cos(ix) + b_i \sin(ix)). \qquad (1.2)$$

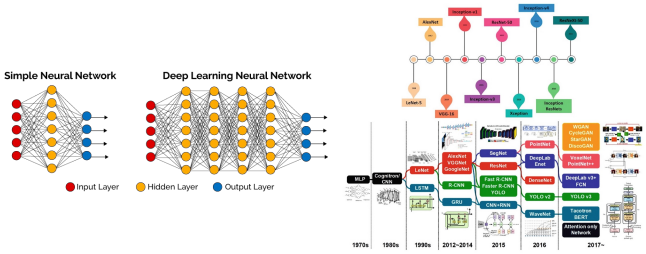
Rojas, 1986

Types of neural networks

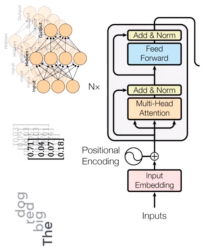


(Fjodor van Veen, 2016, asimovinstitute.org)

Deep neural networks

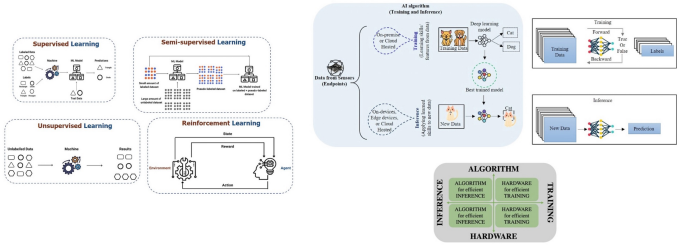


Transformers are deep neural networks



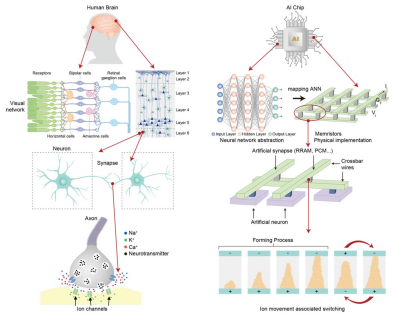
https://builtin.com/artificial-intelligence/transformer-neural-network

Learning



What is harder? Inference or training?

Mishra et al., 2023



Source: "Bridging Biological and Artificial Neural Networks with Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges"

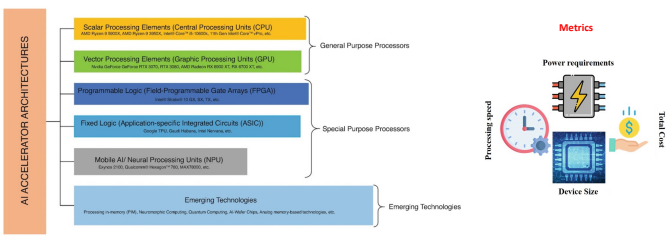
Neuromorphic Computing with Large Scale Spiking Neural Networks

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2025 doi:10.20944/preprints202503.1505.v1

Table 3. Comparison of leading neuromorphic hardware platforms in terms of neuron capacity, power efficiency, and scalability.

Hardware Platform	Technology	Neuron Capacity	Power Efficiency	Scalability
IBM TrueNorth	Digital	10^6 neurons	High	Moderate
Intel Loihi	Digital	10^5 neurons	Very High	High
SpiNNaker	Digital	10^6 neurons	Moderate	High
BrainScaleS	Analog	10^4 neurons	Low	Low
Tianjic	Hybrid	10^5 neurons	High	High

Hardware for AI/ML



Mishra et al., 2023

Hardware for AI/ML

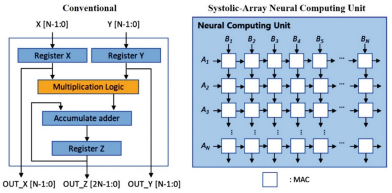
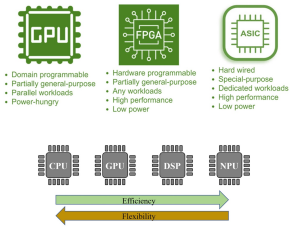


Fig. 19 A comparison of MAC operations performed on conventional and neural computing units.

Mishra et al., 2023

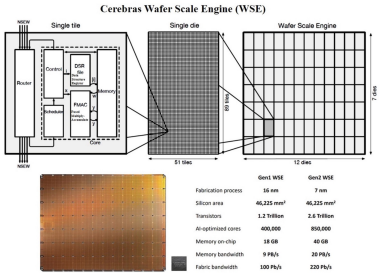
Hardware for AI/ML

Processor	Power Consumption	Strengths	Limitations
CPU	High	→ Flexible → General-purpose processing → Complex instructions and tasks → System management	→ Possible memory access bottlenecks → Few cores (4–16)
GPU	High	→ Parallel cores (~1000 s of cores) → High-performance AI processing	→ Power consumption → Large footprint
FPGA	Medium	→ Configurable logic gates → Flexible → In-field re-programmability	→ Programming complexity
ASIC	Low	→ Custom logic designed with libraries → Faster processing → Small footprint	→ Fixed function → Expensive custom design
Vision Processing Unit (VPU)	Ultra-low	→ Dedicated image and vision co-processor → Small footprint	→ Limited dataset and batch size → Limited network support
Tensor Processing Unit (TPU)	Low to medium	→ Specialized tool support → Optimized for TensorFlow	→ Proprietary design → Limited framework support



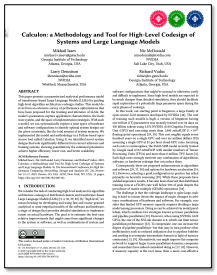
Mishra et al., 2023

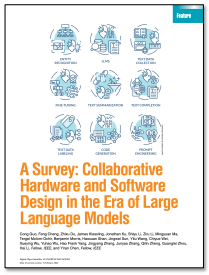
Hardware for AI/ML



Mishra et al., 2023

Co-design & LLM





(HW/SW) Co-design

- **Classical approach**: in college, different technologies are taught in different classes.
- **Hardware-software co-design:**
  - Started in the 1990s.
  - Its core idea is the concurrent designs of hardware and software components of complex electronic systems.
- **More broadly**:
  - The concurrent design of hardware and software across all layers of the compute stack.

- **What are possible metrics and trade-offs?**

Deep co-design across the entire stack

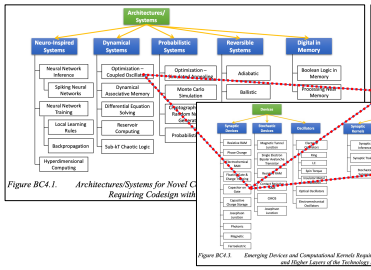
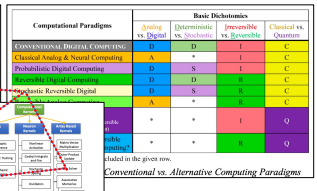


Figure BC4.1. Architectures/Systems for Novel C… Requiring Codesign with…



Conventional vs. Alternative Computing Paradigms

Figure BC4.3. Emerging Devices and Computational Kernels Requiring Codesign between the Device Layer and Higher Layers of the Technology Stack

Food for thought

- Why is executing a neural network on a general purpose computer inefficient?
- What circuitry/processor would you design to find the shortest path in large graphs?
  - What if the graph is dense?
  - What if the graph is sparse?
- What circuitry/processor would you design to recognize QR codes for an embedded camera?
  - Recognition needs to be fast and low-power.