6.5930/1

Hardware Architectures for Deep Learning

Introduction and Applications

February 5, 2024

Joel Emer and Vivienne Sze

Massachusetts Institute of Technology
Electrical Engineering & Computer Science

ACM's Celebration of 50 Years of the ACM Turing Award (June 2017)

"Compute has been the oxygen of deep learning"

– Ilya Sutskever, Research Director of Open AI

Why is Deep Learning Hot Now?

Big Data Availability

GPU Acceleration

New ML Techniques

facebook **136,000** photos uploaded every minute

YouTube **500 hours** of video uploaded every minute

Walmart **2.5 Petabytes** of customer data hourly

ImageNet Challenge

IMAGENET

Image Classification Task

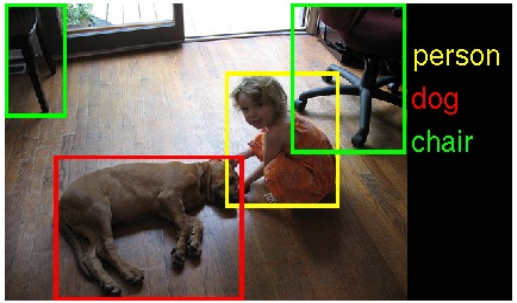
1.2M training images

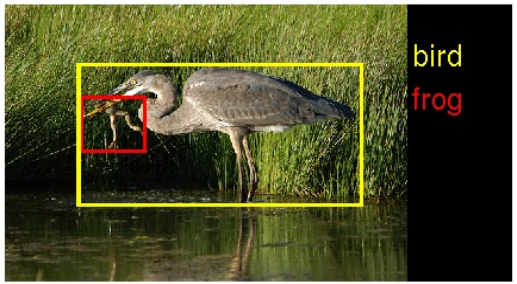
1000 object categories

Object Detection Task

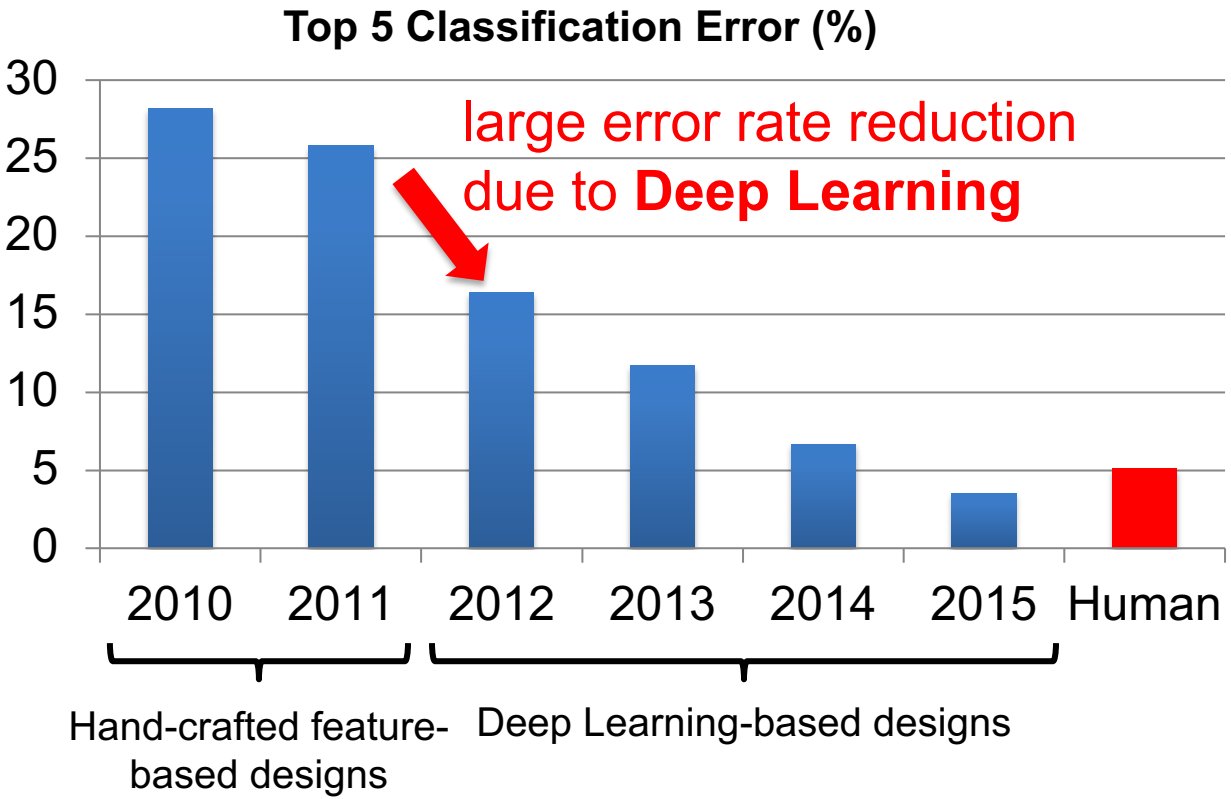
456k training images

200 object categories



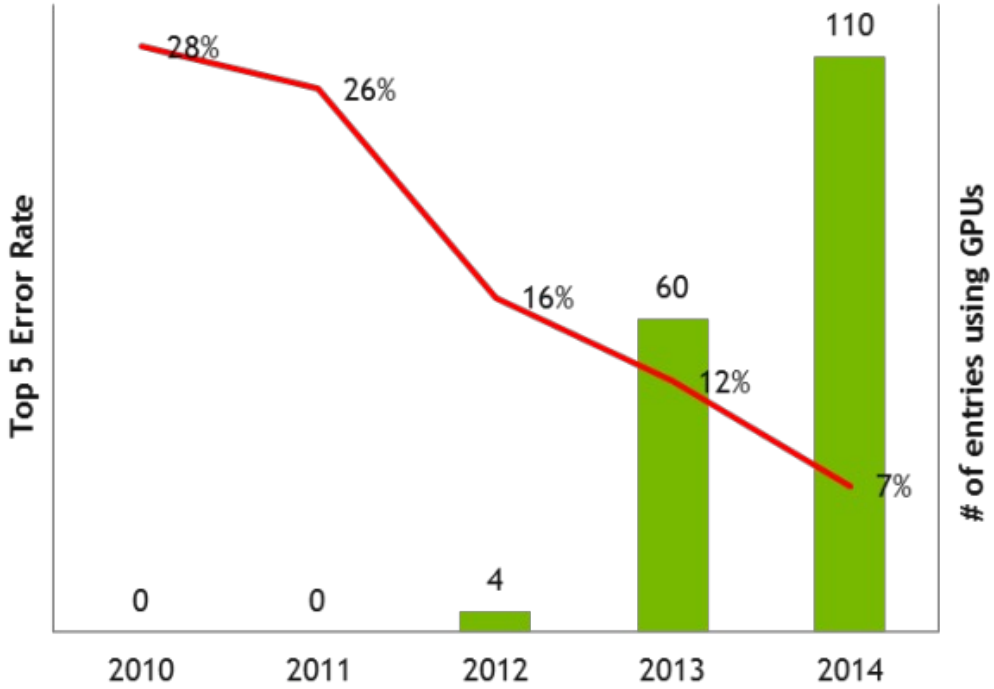


ImageNet: Image Classification Task



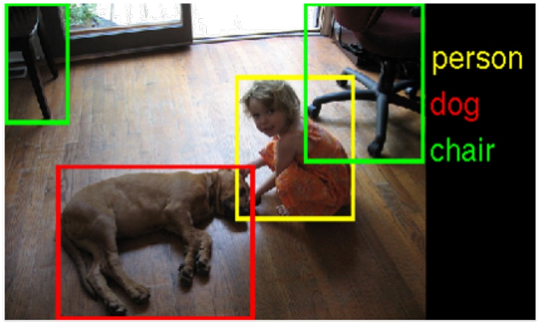
[**Russakovsky**, *IJCV* 2015]

GPU Usage for ImageNet Challenge



Deep Learning on Images and Video

- Image Classification
- Object Detection
- Object Localization
- Image Segmentation
- Action Recognition
- Image Generation



Computer Vision (6.8301/0$_{[6.819/869]}$)

Deep Learning for Speech and Language

- **Speech Recognition**
- **Natural Language Processing**
- **Speech Translation**
- **Audio Generation**

Speech ($6.8620_{[6.345]}$), Natural Language Processing ($6.8610/1_{[6.864/806]}$)

Deep Learning on Games

Google DeepMind AlphaGo (2016)

Training on 160,000 games required 3 weeks on 50 GPUs!

During play (***inference***) AlphaGo used 40 search threads, 48 CPUs, and 8 GPUs.

(A distributed version used 40 search threads, 1,202 CPUs and 176 GPUs.)

Deep Learning on Games

Deepmind's AlphaStar defeats top professional players at StarCraft II (January 2019)

https://www.deepmind.com/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii

StarCraft, considered to be one of the most challenging Real-Time Strategy (RTS) games and one of the longest-played e-sports of all time, has emerged by consensus as **a "grand challenge" for AI research**.

Requires overcoming research challenges including **Game Theory, Imperfect Information, Long term planning, Real-time, and Large action space**

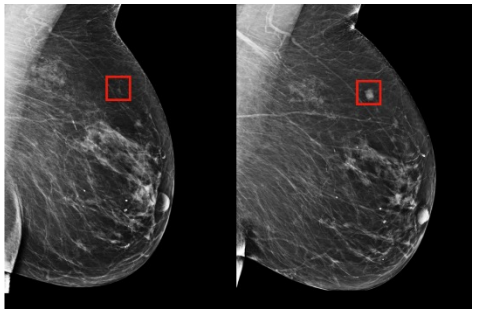
Training required 14 days, using 16 TPUs for each agent!

Each agent experienced up to 200 years of real-time play.

Deep Learning for Medical Applications

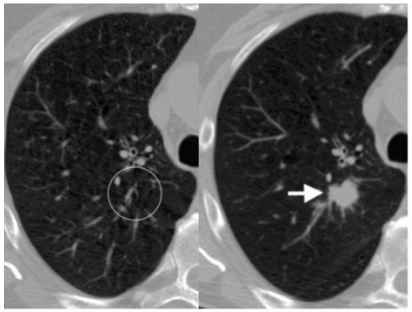
Cancer Detection

Breast



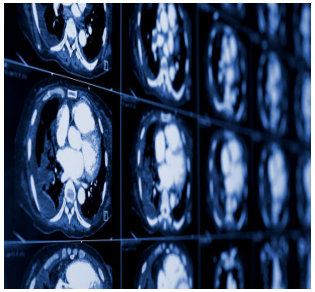
[**Yala**, *Radiology* 2019]

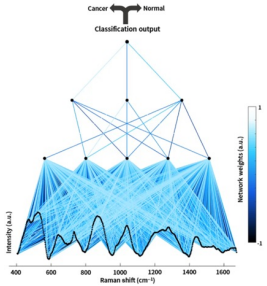
Lung



[**Mikael**, *JCO* 2023]

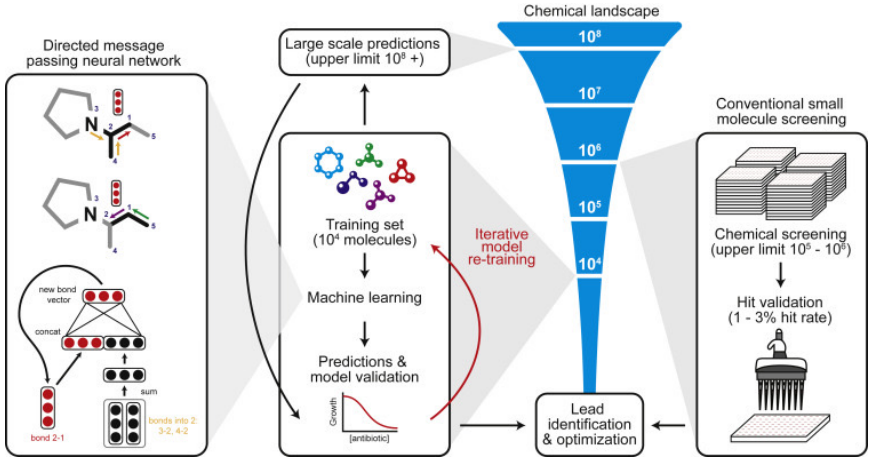
Brain





[**Jermyn**, *JBO* 2016]

Antibiotic Discovery



[**Stokes**, *Cell* 2020]

Machine Learning for Healthcare 6.7930$_{[6.871]}$

Deep Learning in Biology

- Structure of protein provides insight on its function, which are important for development of disease treatments (e.g., proteins associated with SARS-CoV-2)

- Current approaches involve years of work and multi-million dollars equipment

- **Protein Folding Problem:** Predict 3D structure of protein based on 1D sequence of amino acids.
- Estimated 10^{300} possibilities → ***Grand challenge in Biology for the past 50 years!***

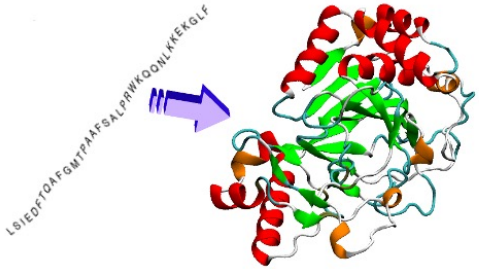
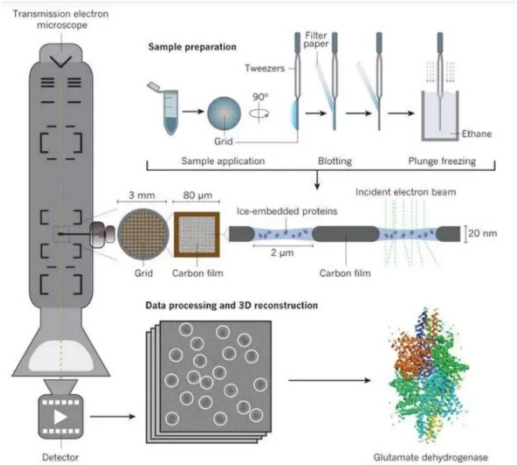


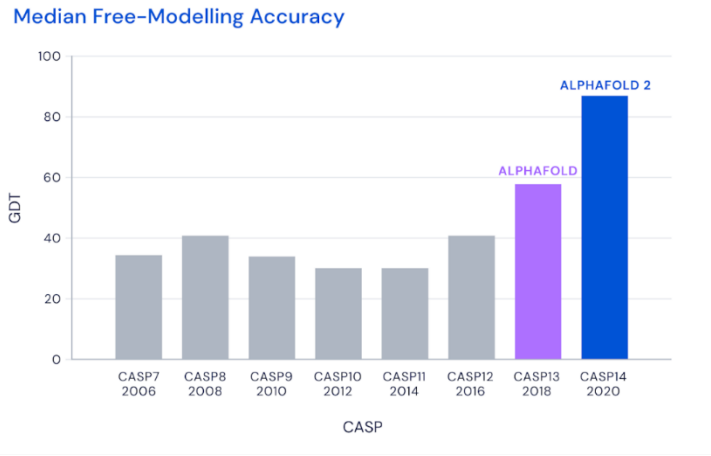
Image source: Dill et al., Annu. Rev. Biophys. 2008 37:289

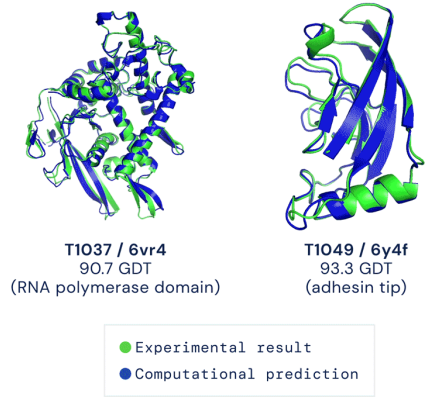


Cryo-electron microscopy
Image source: Quora

Deep Learning in Biology

DeepMind's AlphaFold2 recognized as solution to protein folding challenge (November 2020)





Trained on publicly available data consisting of ~170,000 protein structures together with large databases containing protein sequences of unknown structure.
It uses approximately 16 TPUv3s (which is 128 TPUv3 cores or roughly equivalent to ~100-200 GPUs) run over a few weeks.

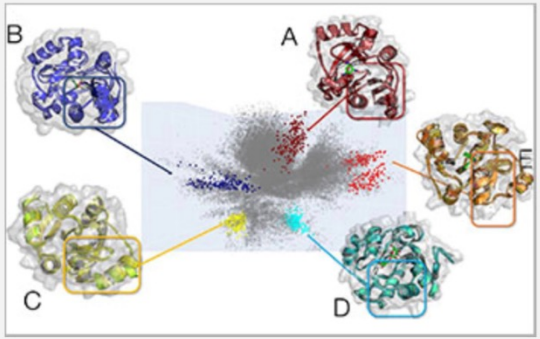
Source: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

Deep Learning in Fighting COVID-19

Research News

AI accelerating drug discovery to fight COVID-19

Deep learning, drug docking and molecular dynamics simulations identify ways to shut down virus



AI-driven molecular dynamics simulations may lead to new drugs to treat coronavirus.

"A project **using some of the most powerful supercomputers on the planet** -- …-- is running millions of simulations, training a machine learning system to identify the factors that might make a molecule a good candidate…."

https://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=300481&org=NSF&from=news

Large Language Models: ChatGPT

Computing Cost of ChatGPT (Training)

- ChatGPT is based on a variant of GPT-3 [**Brown,** *NeurIPS* 2020]
- GPT-3 has 96-layers, 175 billion parameters and requires $3.14x10^{23}$ FLOPS of computing for training
- It would take **355 years** to train GPT-3 on a Tesla V100 GPU
- It would cost **~$4,600,000** to train GPT-3 on using the lowest cost GPU cloud provider

Source: https://lambdalabs.com/blog/demystifying-gpt-3

Computing Cost of ChatGPT (Inference)

Elon Musk @elonmusk · Dec 5, 2022
Replying to @sama
What's the average cost per chat?
2,677 1,828 52.3K

Sam Altman
@sama

Replying to @elonmusk

average is probably single-digits cents per chat; trying to figure out more precisely and also how we can optimize it

2:46 AM · Dec 5, 2022 https://twitter.com/sama/status/1599671496636780546

Chief Disruptor @ahmedsalims · Dec 5, 2022
Replying to @sama
Will it be FREE forever?
13 7 306

Sam Altman
@sama

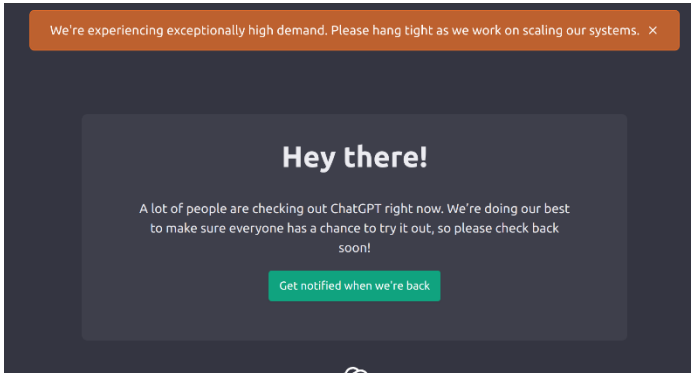
Replying to @ahmedsalims

we will have to monetize it somehow at some point; the compute costs are eye-watering

2:38 AM · Dec 5, 2022 https://twitter.com/sama/status/1599669571795185665

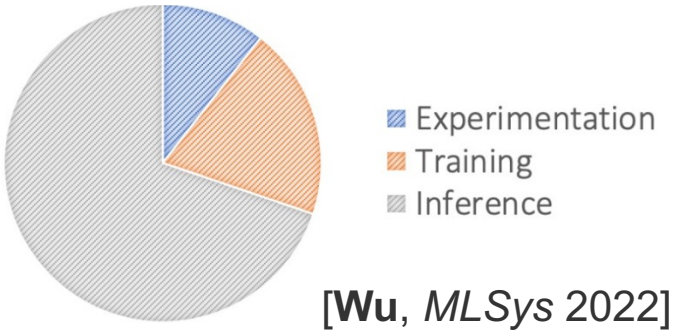
Estimated monthly cost of **$1.5 to $8 million!**

Source: https://medium.com/swlh/3-questions-puzzled-me-about-openais-chatgpt-and-here-is-what-i-learned-1dda74b5f6db



Compute for Training versus Inference

- Training **high cost per iteration**, but **low frequency**
- Inference **low cost per iteration**, but **high frequency**
- Regarding computing at Google:
  - "Across all three years [2019-2021], about three-fifths of the ML energy use was for inference, and two-fifths were for training" [**Patterson**, *Computer* 2022]
- Regarding computing at Meta:
  - "…trillions of inference per day across Meta's data centers. … we observe a rough power capacity breakdown of 10:20:70 for AI infrastructures devoted to the three key phases — Experimentation, Training, and Inference"



[**Wu**, *MLSys* 2022]

Huge Financial Investment in GPUs

Home > News > Components > Graphics Cards

Zuckerberg's Meta Is Spending Billions to Buy 350,000 Nvidia H100 GPUs

In total, Meta will have the compute power equivalent to 600,000 Nvidia H100 GPUs to help it develop next-generation AI, says CEO Mark Zuckerberg.

By Michael Kan January 18, 2024

Source: https://www.pcmag.com/news/zuckerbergs-meta-is-spending-billions-to-buy-350000-nvidia-h100-gpus

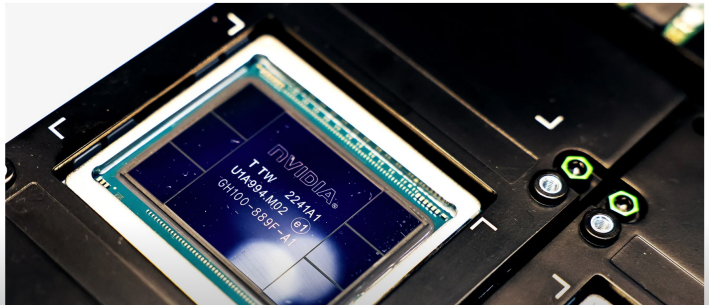
GPU Shortage

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS POLITICS SCIENCE SECURITY MERCH SIGN IN SUBSCRI

PARESH DAVE BUSINESS AUG 24, 2023 6:00 AM

Nvidia Chip Shortages Leave AI Startups Scrambling for Computing Power

Trimming profits, delaying launches, begging friends. Companies are going to extreme lengths to make do with shortages of GPUs, the chips the heart of generative AI programs.



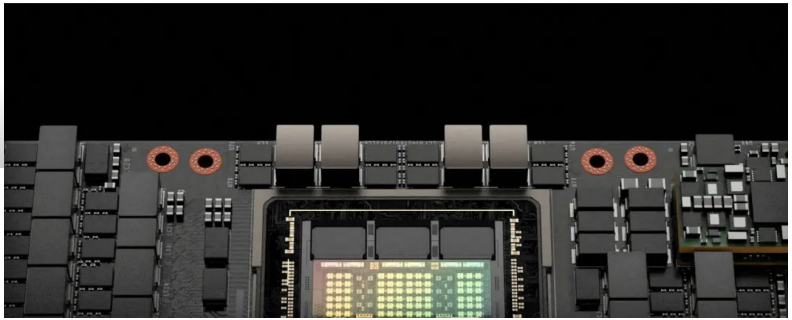
Source: https://www.wired.com/story/nvidia-chip-shortages-leave-ai-startups-scrambling-for-computing-power/

The New York Times

The Desperate Hunt for the A.I. Boom's Most Indispensable Prize

To power artificial-intelligence products, start-ups and investors are taking extraordinary measures to obtain critical chips known as graphics processing units, or GPUs.

Share full article 42

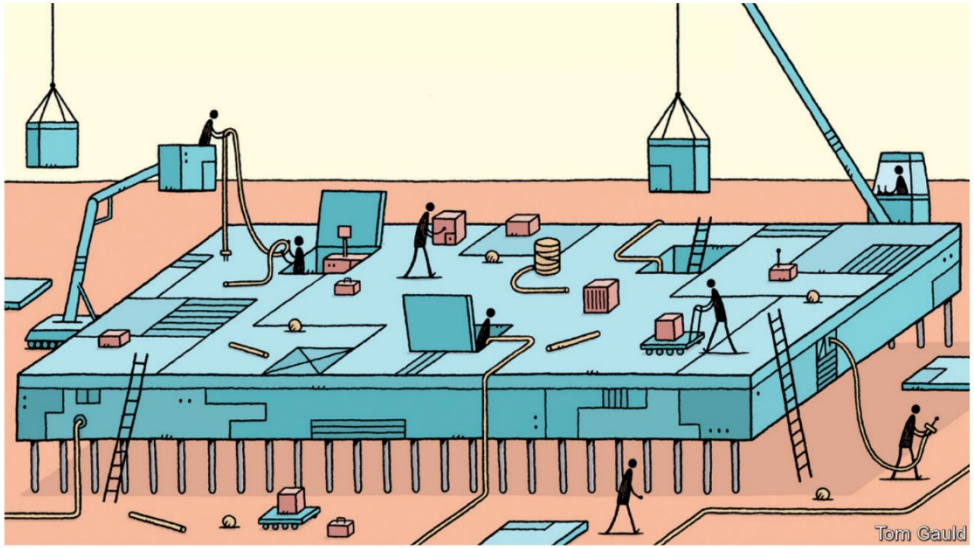


Source: https://www.nytimes.com/2023/08/16/technology/ai-gpu-chips-shortage.html

Compute Demands Growing Exponentially

The Economist

Today Weekly edition Menu Search My account



Technology Quarterly

Jun 11th 2020 edition ›

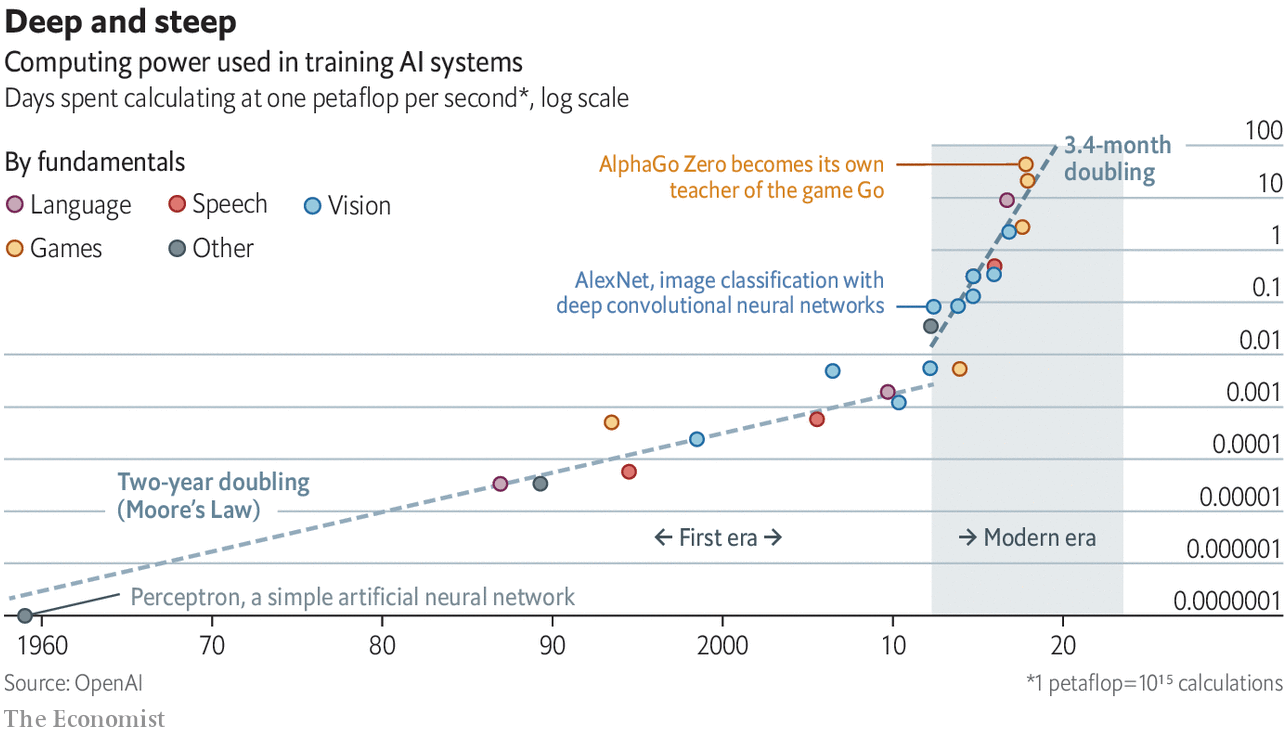
Computing hardware

The cost of training machines is becoming a problem

Increased complexity and competition are part of it

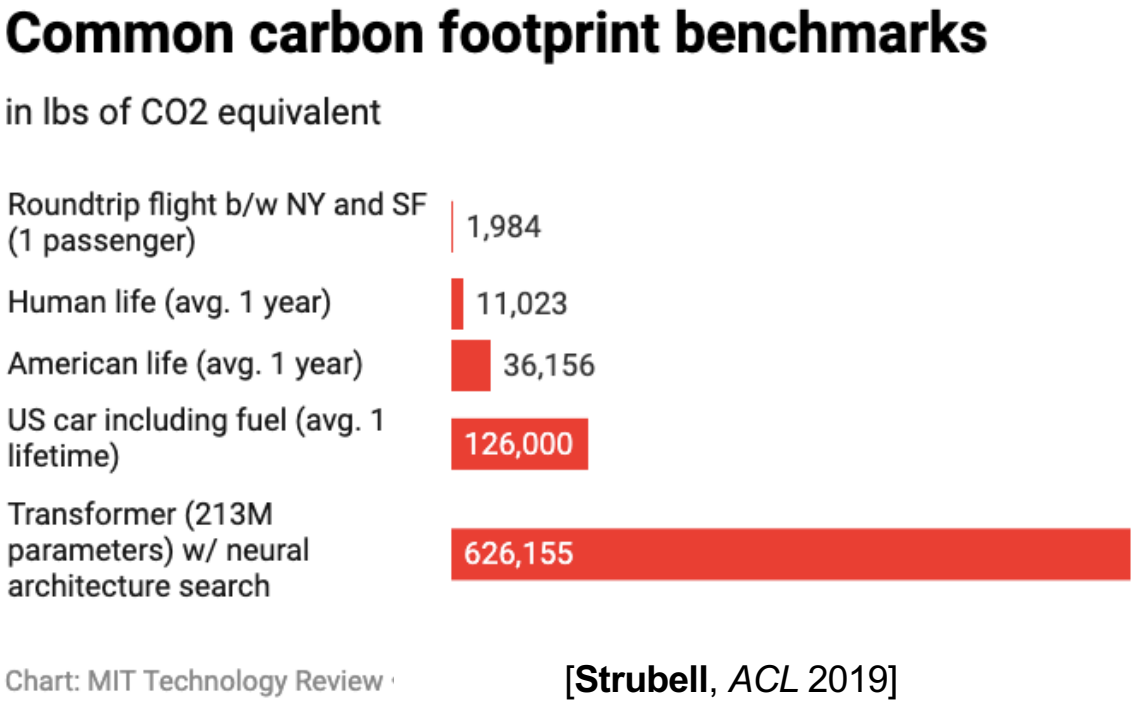
Compute Demands Growing Exponentially

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



Source: https://www.economist.com/technology-quarterly/2020/06/11/the-cost-of-training-machines-is-becoming-a-problem

Compute Demands for Deep Neural Networks



[**Strubell**, *ACL* 2019]

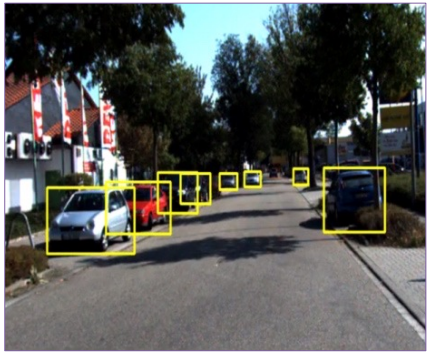
Processing at “Edge” instead of the “Cloud”

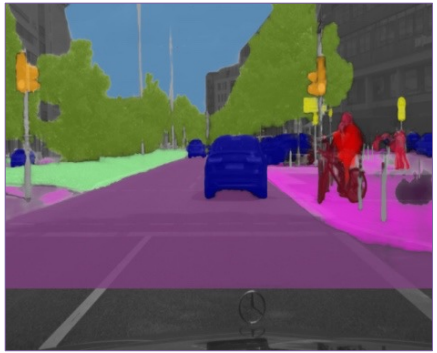
Communication

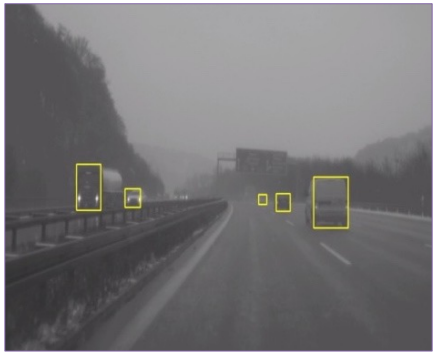
Privacy

Latency

Deep Learning for Self-Driving Cars







Compute Challenges for Self-Driving Cars

JACK STEWART TRANSPORTATION 02.06.18 08:00 AM

SELF-DRIVING CARS USE CRAZY AMOUNTS OF POWER, AND IT'S BECOMING A PROBLEM

Shelley, a self-driving Audi TT developed by Stanford University, uses the brains in the trunk to speed around a racetrack autonomously.

NIKKI KAHN/THE WASHINGTON POST/GETTY IMAGES

WIRED

(Feb 2018)

Cameras and radar generate ~6 gigabytes of data every 30 seconds.

Prototypes use around 2,500 Watts. Generates wasted heat and some prototypes need water-cooling!

Carbon Footprint of Self-Driving Cars

"[T]rillions of inference per day across Facebook's data centers" [Source: Wu et al. 2021]

Autonomous vehicles (AVs) w/ 10 deep neural network (DNN) inferences at 60 Hz on 10 cameras:

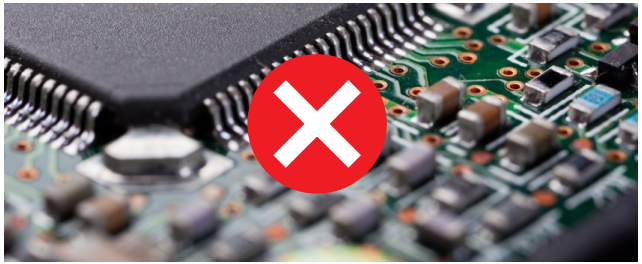
One AV: 21.6 million inferences per hour driven

One billion AVs: **21.6 quadrillion** inferences per hour driven!

[**Sudhakar**, IEEE *Micro* 2023]

Existing Processors Consume Too Much Power

< 1 Watt

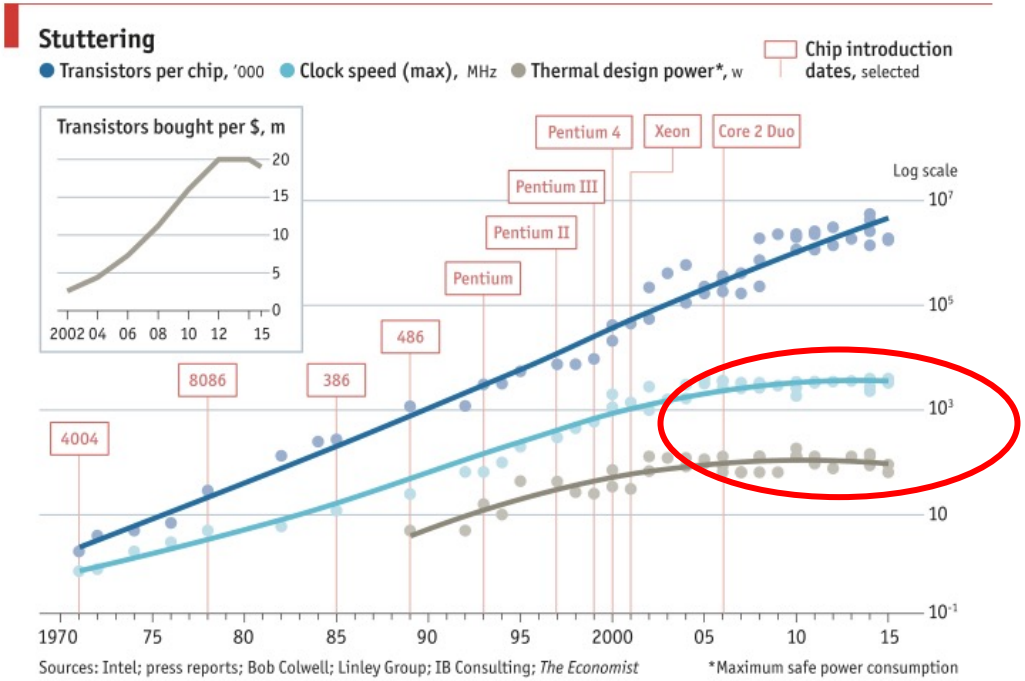


> 10 Watts

Need for Specialized Computing

Slow down of Moore's Law and Dennard Scaling

General purpose microprocessors not getting faster or more efficient



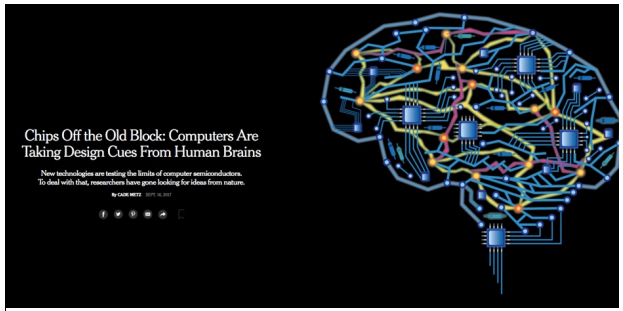
Domain specific hardware needed for significant improvement in **speed** and **energy-efficiency** → **redesign computing hardware from ground up!**

Challenges and Opportunities

- Define the domain and degree of flexibility required
- Requires understanding of domain, variants of algorithms (identify which ones are important), and resulting workloads
- Handle heterogenous computing at the system level
- Handle heterogenous devices (emerging device technology)
- May have tighter resource constraints since hardware cannot be used for other applications
- Co-design across algorithms and hardware
- Domain specific languages to program specialized hardware
- Tools for rapid evaluation and prototyping

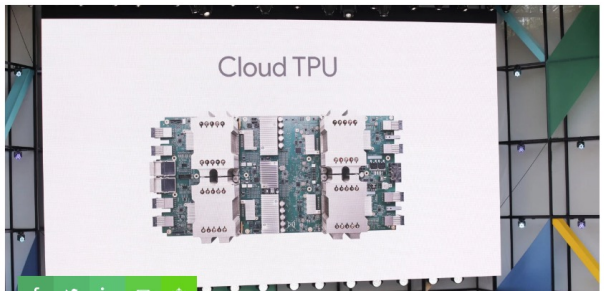
Software Companies are Building HW

The New York Times



Google's custom TPU machine learning accelerators are now available in beta

Frederic Lardinois @fredericl / Feb 12, 2018 Comment



Chips Off the Old Block: Computers Are Taking Design Cues From Human Brains
(September 16, 2017)

After training a speech-recognition algorithm, for example, Microsoft offers it up as an online service, and it actually starts identifying commands that people speak into their smartphones. **G.P.U.s are not quite as efficient during this stage of the process. So, many companies are now building chips specifically to do what the other chips have learned.**

Google built its own specialty chip, a Tensor Processing Unit, or T.P.U. Nvidia is building a similar chip. **And Microsoft has reprogrammed specialized chips from Altera**, which was acquired by Intel, so that it too can run neural networks more easily.

HW Beyond Cloud Computing

WIRED

sk Says Tesla Is Building Its Own Chip for Autopilot

TOM SIMONITE BUSINESS 12.08.17 01:09 PM

MUSK SAYS TESLA IS BUILDING ITS OWN CHIP FOR AUTOPILOT

Elon Musk disclosed plans for Tesla to design its own chip to power its self-driving function.
NASA/ALAMY

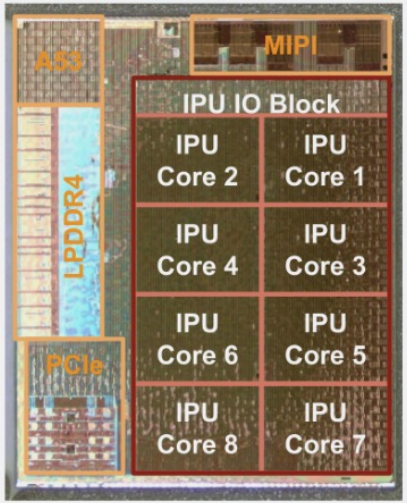
ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

TWO SOCS IS BETTER THAN ONE —

Surprise! The Pixel 2 is hiding a custom Google SoC for image processing

Google's 8-core Image Processing Unit will be enabled with Android 8.1.

RON AMADEO - 10/17/2017, 9:00 AM



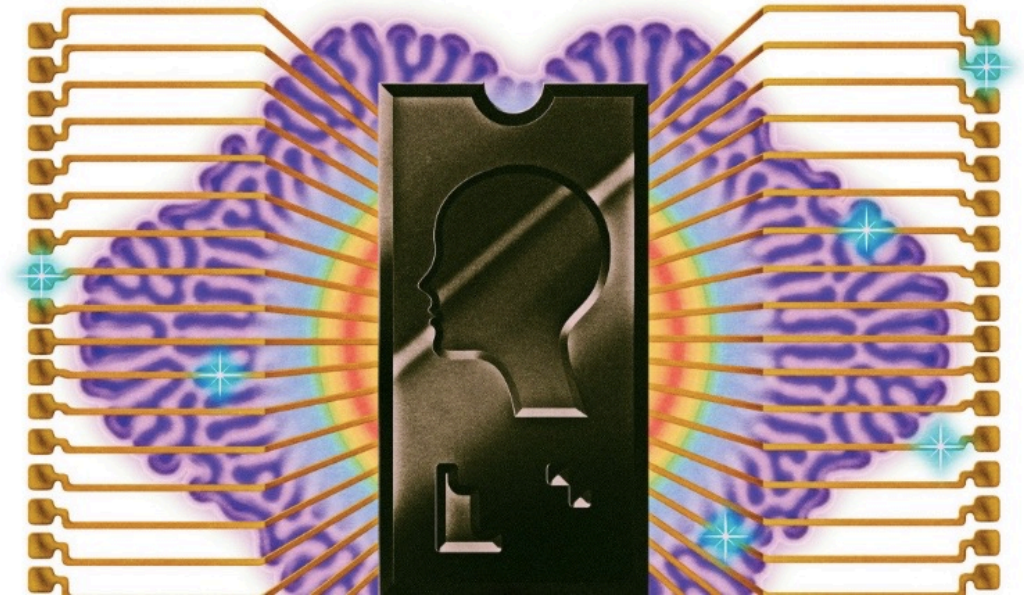
Enlarge / Google's Pixel Visual Core, an SoC designed for image processing and machine learning.

[Also Nvidia, Intel, Qualcomm…]

Startups Building Custom Hardware

The New York Times

By CADE METZ JAN. 14, 2018



Big Bets On A.I. Open a New Frontier for Chips Start-Ups, Too. *(January 14, 2018)*

"Today, **at least 45 start-ups are working on chips** that can power tasks like speech and self-driving cars, and at least five of them have raised more than $100 million from investors. **Venture capitalists invested more than $1.5 billion in chip start-ups last year**, nearly doubling the investments made two years ago, according to the research firm CB Insights."

Class Overview

Course Outline

- Overview of Deep Neural Networks (DNNs)
- DNN Development Resources
- DNNs on Programmable Hardware
- DNN Accelerator Architecture
- DNN Model and Hardware Co-Design
- Advanced Technologies for DNN

Takeaways

- Know the key computations used by DNNs
- Be familiar with how DNN computations are mapped to various hardware platforms
- Understand the tradeoffs between various architectures and platforms
- Be able to evaluate different DNN accelerator implementations with benchmarks and comparison metrics
- Have an appreciation for the utility of various optimization and approximation approaches
- Be able to distill the key attributes of recent implementation trends and opportunities

Course Objective

By the end of this course, we want you to be able understand a new design in terms of attributes like:

- Order of computation
- Partitioning of computation
- Flow of data for computation
- Data movement in the storage hierarchy
- Data attribute specific optimizations
- Exploiting algorithm/hardware co-design
- Degree of flexibility

Course Staff and Contact Info

- **Instructors** (Office hours by request)
  - Joel Emer (jsemer@mit.edu)
  - Vivienne Sze (sze@mit.edu)
- **Teaching Assistants** (Office hours TBD)
  - Tanner Andrulis (andrulis@mit.edu)
  - Zoey Song (zhiye@mit.edu)
  - Michael Gilbert (gilbertm@mit.edu)
- **Schedule**
  - Lectures: MW 1PM-2:30PM – E25-111
  - Recitations: F 11AM-12PM – 32-155
    - Review Lectures
      - ML & Computer Architecture Basics
    - Tools for Labs
- **Course Website:** http://csg.csail.mit.edu/6.5930/

Joel Emer

Vivienne Sze

Tanner Andrulis

Zoey Song

Michael Gilbert

Slide Contributors: Yu-Hsin Chen and Tien-Ju Yang (http://eyeriss.mit.edu/tutorial.html)

Course Requirements and Materials

- Pre-requisites
  - $6.3000_{[6.003]}$ (Signal Processing) or $6.3900_{[6.036]}$ (Intro to Machine Learning)
  - $6.1910_{[6.004]}$ (Computation Structures)
- We will use Python and PyTorch
  - PyTorch website: https://pytorch.org/
  - Introduction to PyTorch Code Examples: https://cs230.stanford.edu/blog/pytorch/
- Course Textbook/Readings
  - Book "Efficient Processing of Deep Neural Networks"
    - https://doi.org/10.1007/978-3-031-01766-7 (download free on MIT network)
    - *We welcome feedback (including errata) on Piazza thread*
  - Selected papers published in past few years.
- Course Handouts (uploaded on website)

Related Classes

- Computer System Architecture ($6.5900_{[6.823]}$)
- Digital Circuits and Systems ($6.6010_{[6.374]}$)
- Digital Image Processing ($6.7010_{[6.344]}$)

- Machine Learning ($6.3900_{[6.036]}$/ $6.7900_{[6.867]}$)
- Computer Vision ($6.8301_{[6.819]}$/$6.8300_{[6.869]}$)

Recitations

- Machine Learning Review (Feb 9)
- Computer Architecture Review (Feb 16, Feb 23)
- Tools
  - PyTorch (Feb 9)
  - Accelergy (Mar 8)
  - Timeloop (Mar 15)
  - Sparseloop (April 5)
- Lab Overview (on same week as Lab released)
- *Note: Above dates are tentative. Will announce updates on website and Piazza.*

Class Participation

- We ask that during the semester each student either ask or answer **at least a total of five questions** related to the lectures in Piazza.

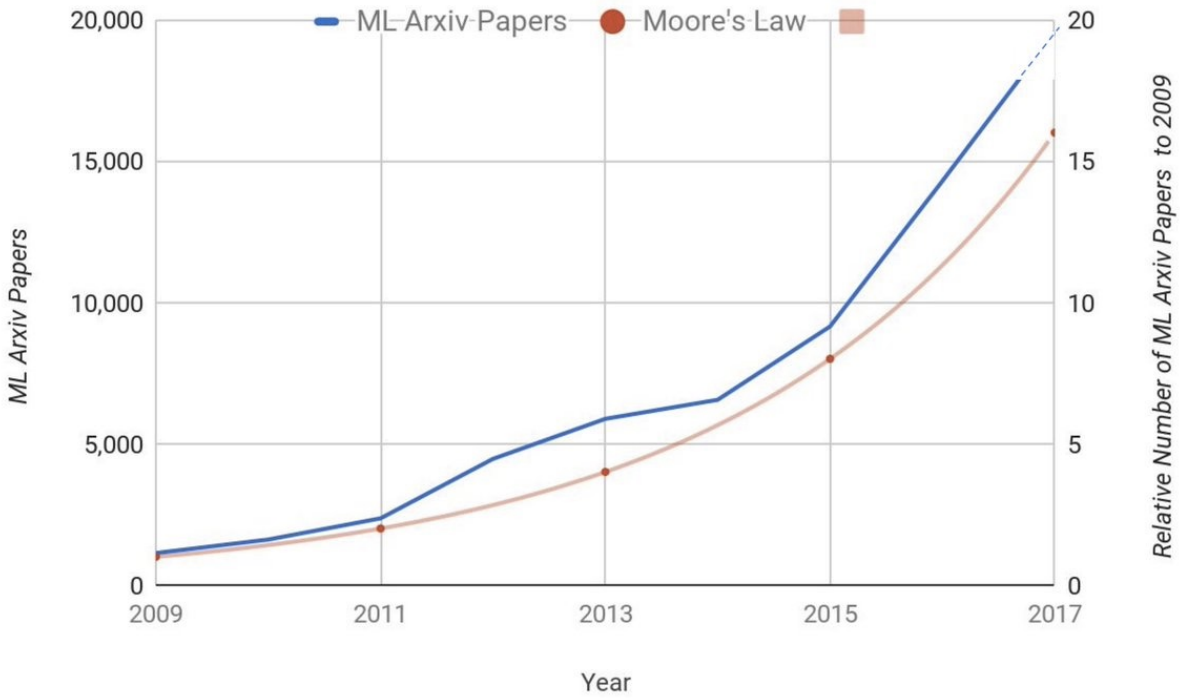
- Multiple (distinct) answers to a single question are permitted.

- If you ask a question in class, please also submit the question to Piazza (for grading purposes). Note: providing the lecturer's answer doesn't count.

Labs

- Lab 1: Inference + DNN Model Design (Release Today)
  - Release: Feb 6 – Due: Feb 26 (3 weeks)
- Lab 2: Kernel + Tiling Optimization
  - Release: Feb 26 – Due: March 11 (2 weeks)
- Lab 3: Hardware Design & Mapping
  - Release: March 11 – Due: April 3 (2.5 weeks + spring break)
- Lab 4: Sparsity
  - Release: April 3 – Due: April 22 (2.5 weeks)
- Lab 5: Compute In Memory (CiM) **(6.5931 students only)**
  - Release: April 22 – Due: May 13 (3 weeks)

ML Arxiv Papers per Year



"Number of ML Arxiv papers published per year beat Moore's Law!… "
– Jeff Dean (Google)

Paper Review

- We will be forming a program committee of the (fictional) *Deep Learning Hardware (DLH)* Conference 2024
  - Learn how to read papers (critique and extract information)
  - Use concepts from class to analyze and gain deep understanding of paper
  - Enable wider coverage of papers
  - Gain insight of how paper decisions are made

- Schedule
  - March 22 – Papers assigned (check for undetected conflicts, e.g., advisor or past co-authors)
  - April 19 – Reviews due (Phase I)
  - April 26 – Online Paper discussion plus decision complete (Phase II)

Quizzes

- Opportunity to apply analytical skills to prior work
- You will be provided with a paper as pre-read
- During the quiz, you will be asked to answer questions related to the content and topic area of the paper
- You will be evaluated based on your ability to apply the principles from the class
- Quizzes will be held during recitations
  - April 12
  - May 3

Design Project (6.5930 students only)

- Project
  - Choose from a list of suggested projects
  - May propose own project (requires formal proposal and approval by course staff)
- Use tools from labs
  - PyTorch, Accelergy, Timeloop, Sparseloop
- Teams of two or three (depending on class size)
- Schedule
  - March 20 – List of projects released
  - April 8 – Submit project selection (or proposal)
  - May 6 – Project Report Due
    - Poster session or short presentation (depending on class size)

Example Design Projects

- Hardware design study of well-known DNN accelerators (e.g., TPU, NVDLA) and analyze architectural tradeoffs though modeling
- Analyze co-design approaches to gain deeper understanding of impact on accuracy and efficiency
- Evaluate impact of new technologies (RRAM, Optical)
- Extend tools for improved design exploration and analysis capabilities

Goal: Apply concepts and tools from class!

Assignments and Grading

- **Grading**

	6.5930 (grad)	6.5931 (undergrad)
Class Participation	5%	5%
Labs 1 to 4	40%	45%
Lab 5	-	10%
Final project	20%	-
Two Quizzes	25%	30%
Paper review	10%	10%

- All assignments are due by **11:59PM ET on the due date** (submitted online)
- Labs are to be completed individually, although discussion of course concepts covered in the laboratories is encouraged. Please carefully review collaboration policy at http://csg.csail.mit.edu/6.5930/collaboration.html

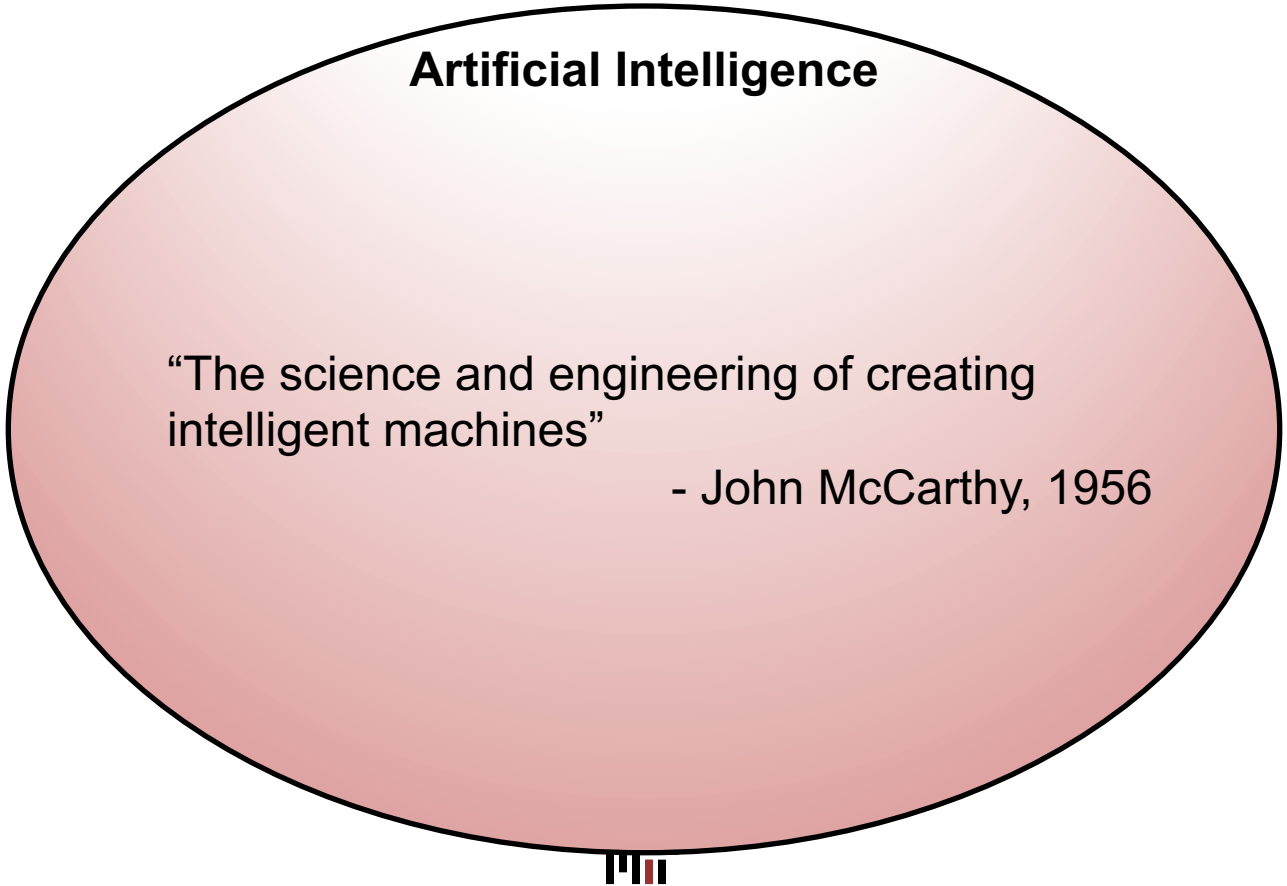
Late Policy

- **Late Policy for Labs**
  - You should always submit your labs on time. Nonetheless, since unexpected situations, like illnesses, might occur, you have a budget of **5 late days** throughout the semester.
  - The budget is spent in increments of 1 day
  - You do not need to inform us about your use of your budget. The course staff will keep track of the days you have spent.
  - *You will receive zero credit once your budget is expended* ***(Note: please contact course staff regarding extenuating circumstances).*** *However, you must complete all assignments to pass the course (even if you get zero credit).*

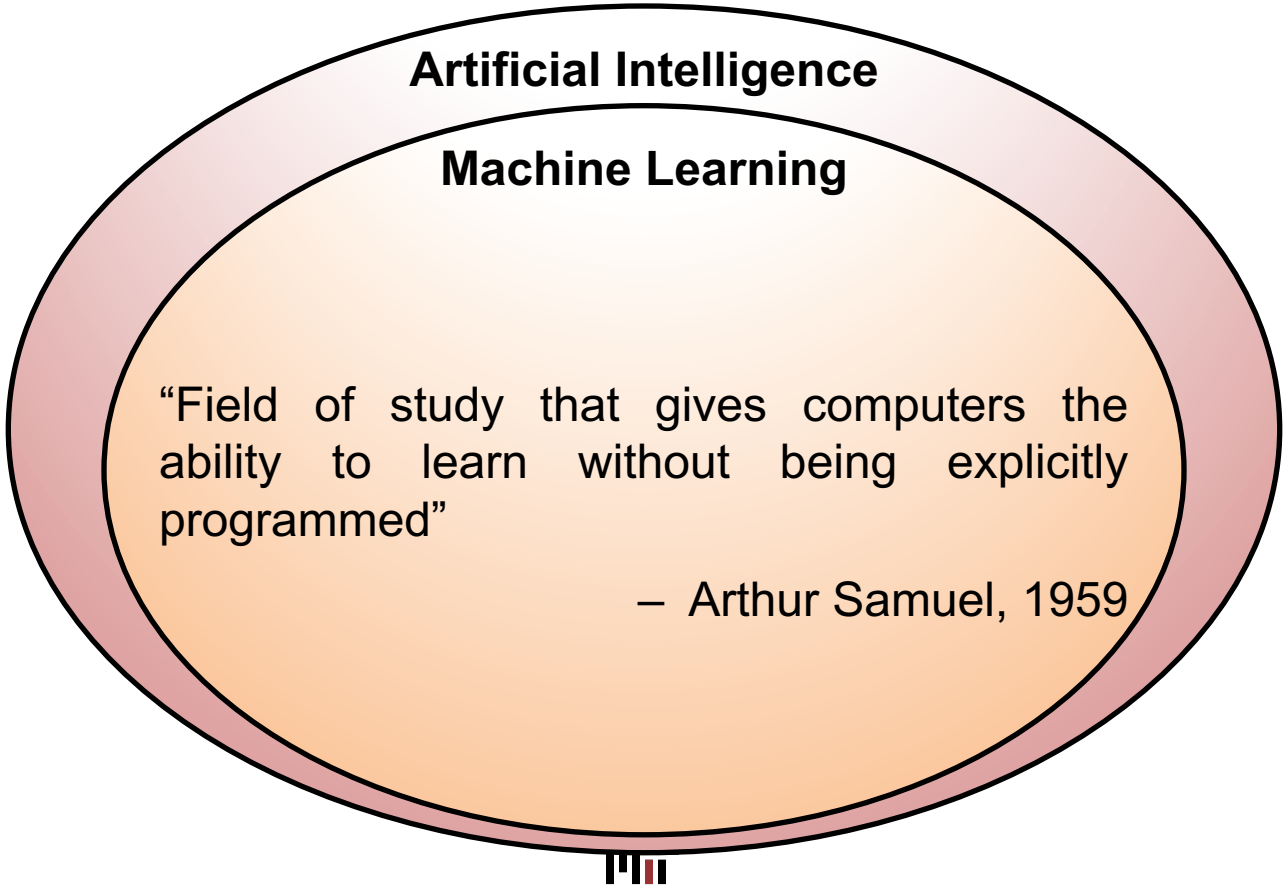
- *No late days for paper reviews and project due to tight timeline*

Background of Deep Neural Networks

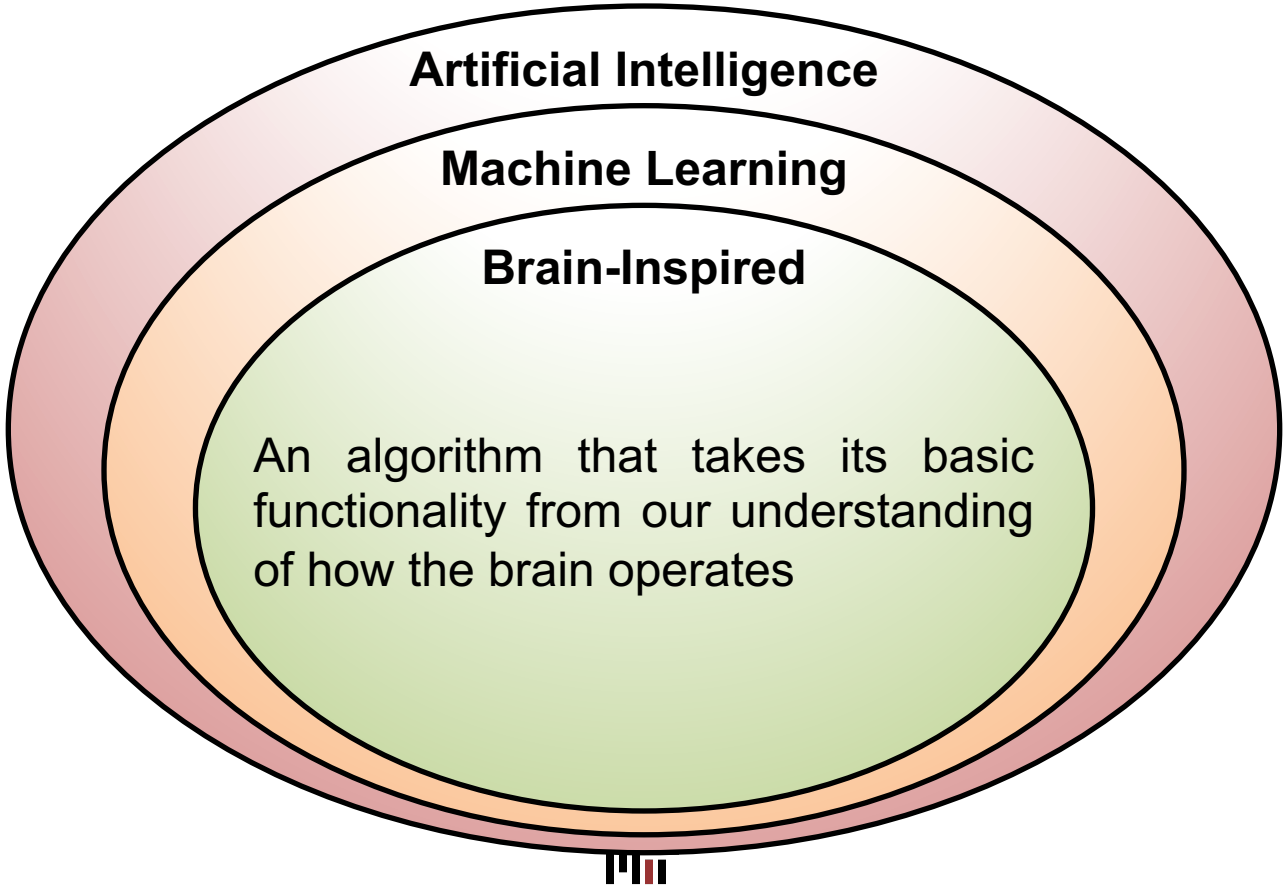
Artificial Intelligence



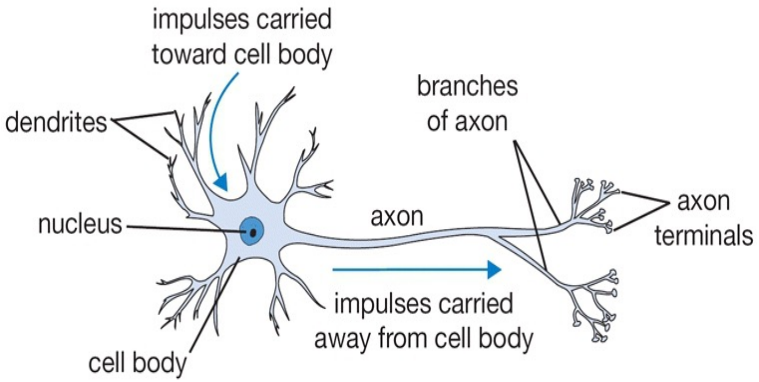
AI and Machine Learning

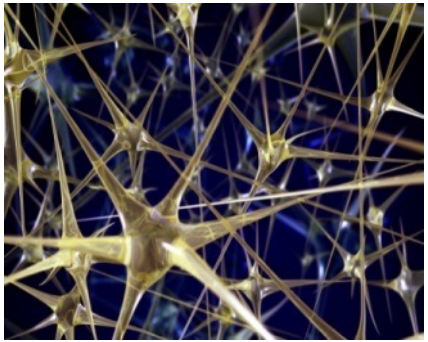


Brain-Inspired Machine Learning



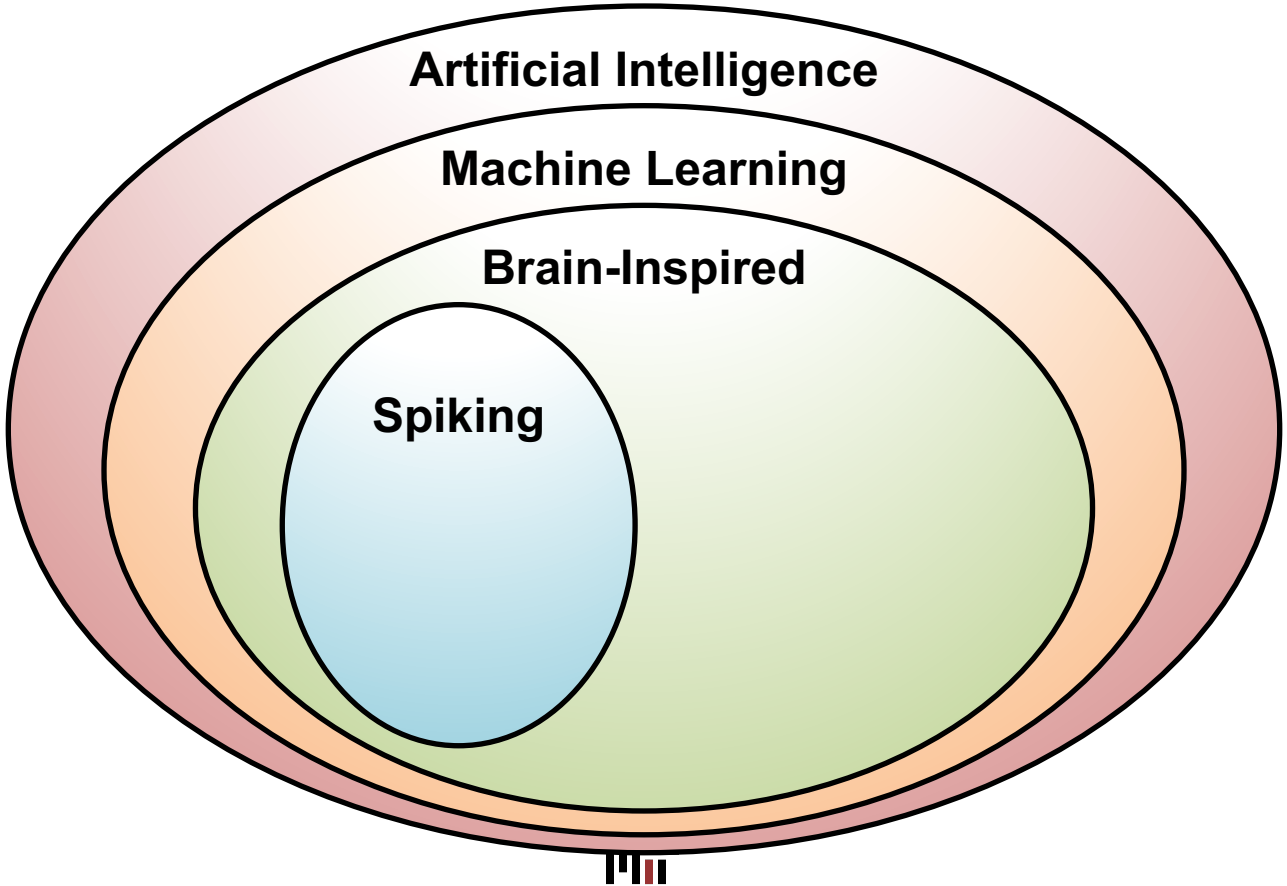
How Does the Brain Work?





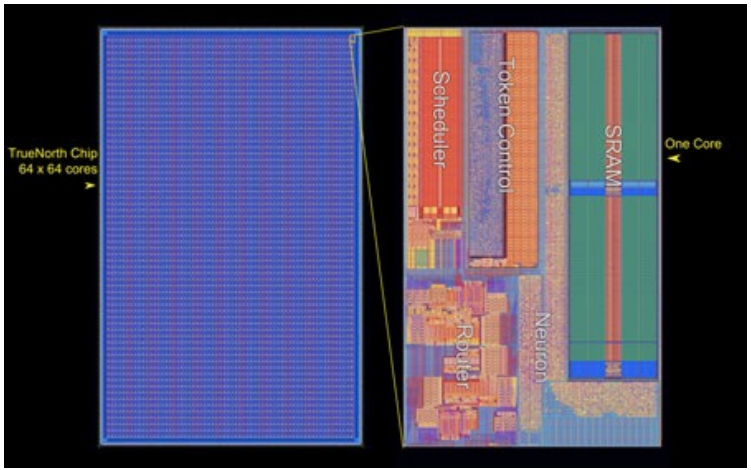
- The basic computational unit of the brain is a **neuron**
  → 86B neurons in the brain
- Neurons are connected with nearly **10^{14} – 10^{15} synapses**
- Neurons receive input signal from **dendrites** and produce output signal along **axon**, which interact with the dendrites of other neurons via **synaptic weights**
- Synaptic weights – learnable & control influence strength

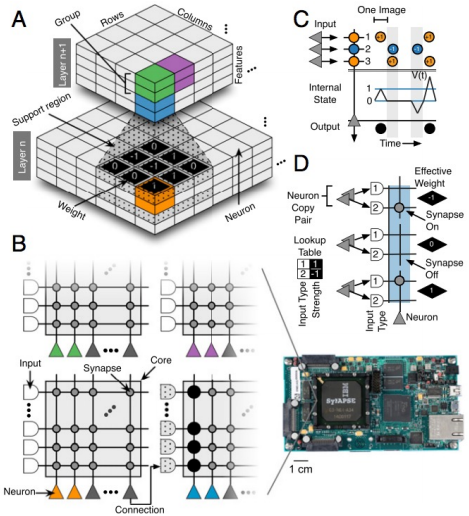
Spiking-based Machine Learning



Spiking Architecture

- Brain-inspired - Integrate and fire
- Example: IBM TrueNorth

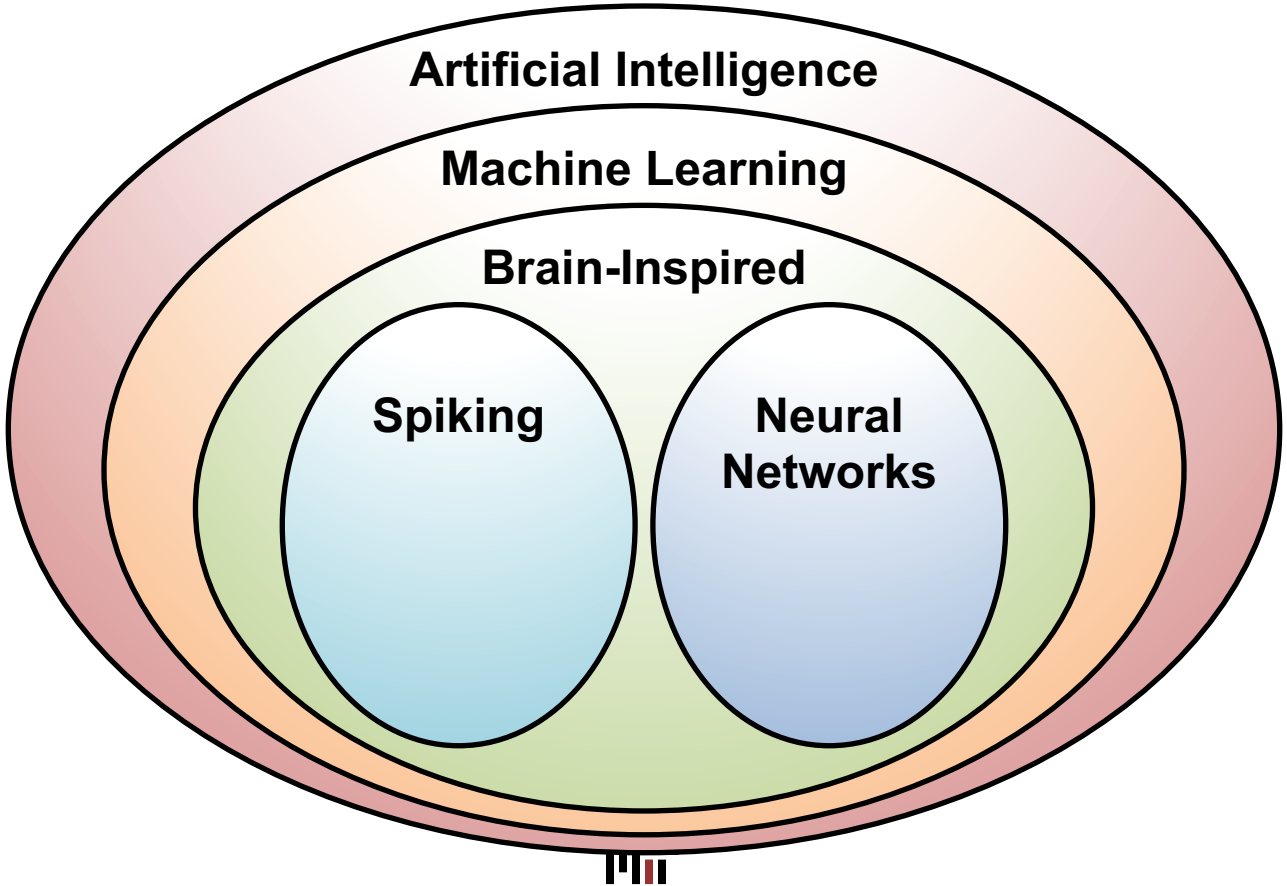




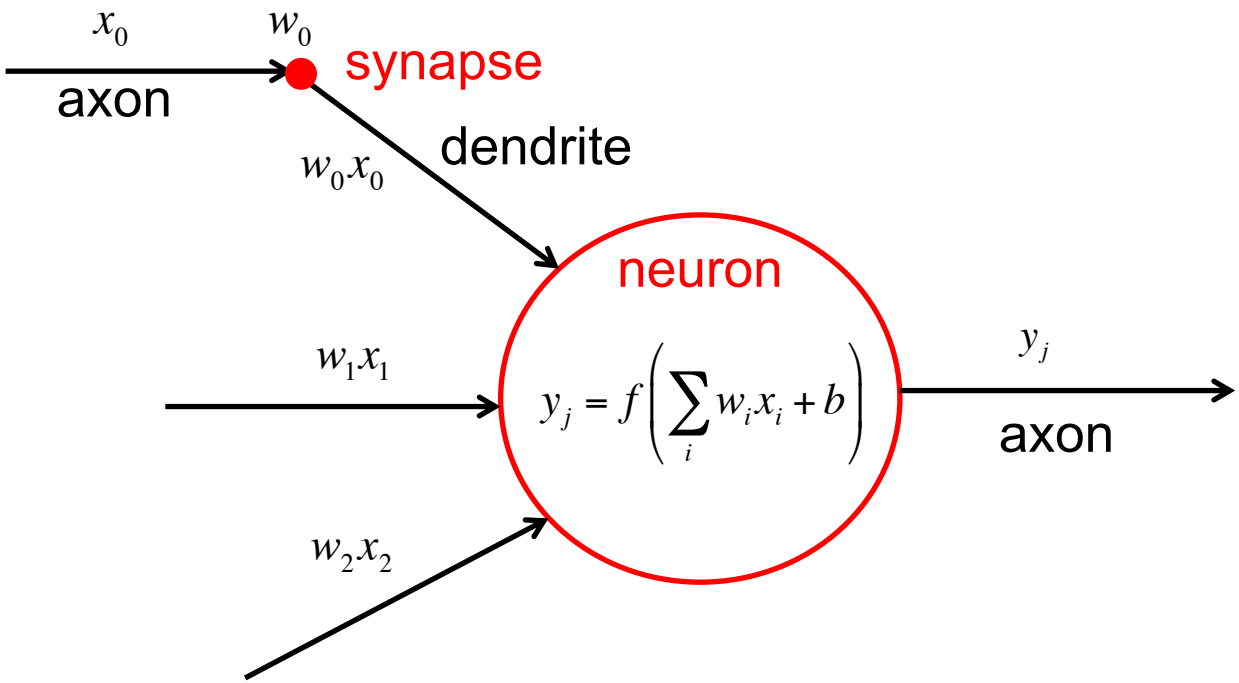
[**Merolla**, *Science* 2014; **Esser**, *PNAS* 2016]

http://www.research.ibm.com/articles/brain-chip.shtml

Machine Learning with Neural Networks

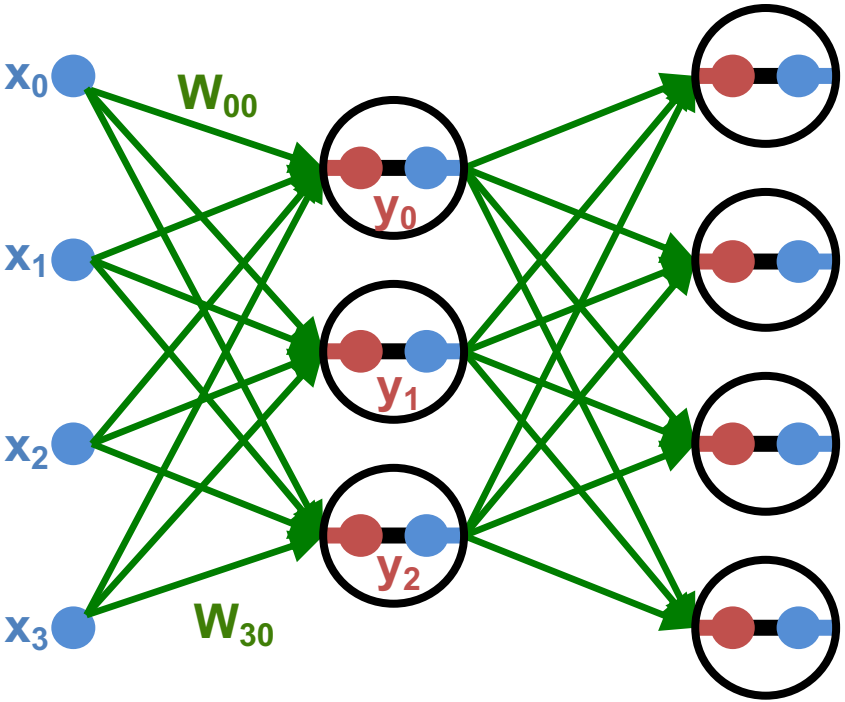


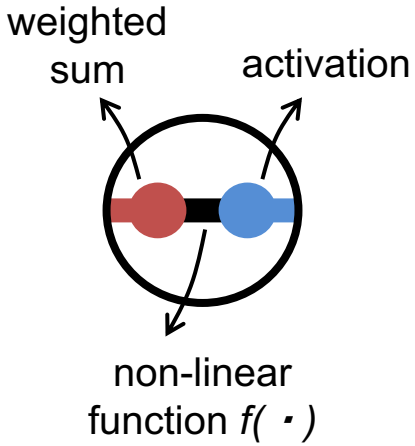
Neural Networks: Weighted Sum



Modified Image Source: Stanford

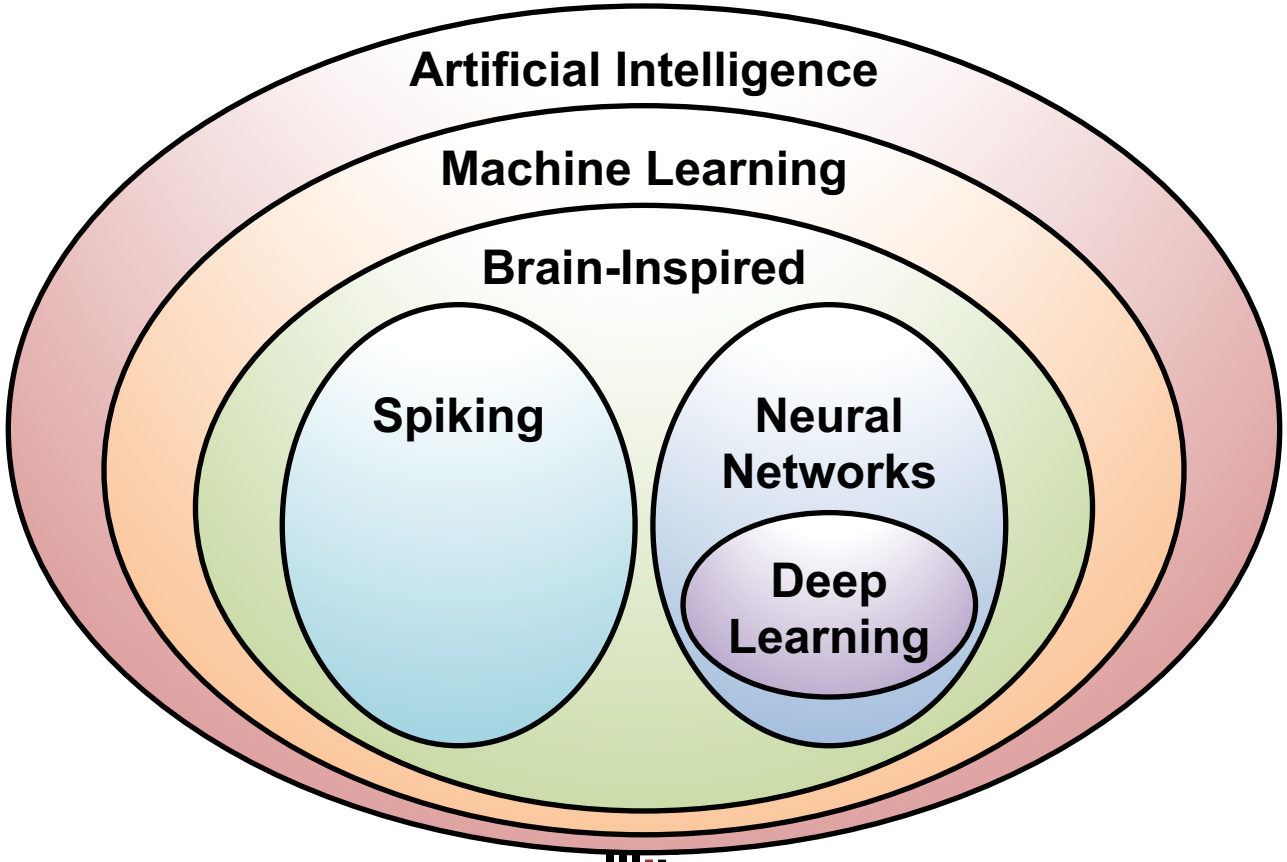
Many Weighted Sums



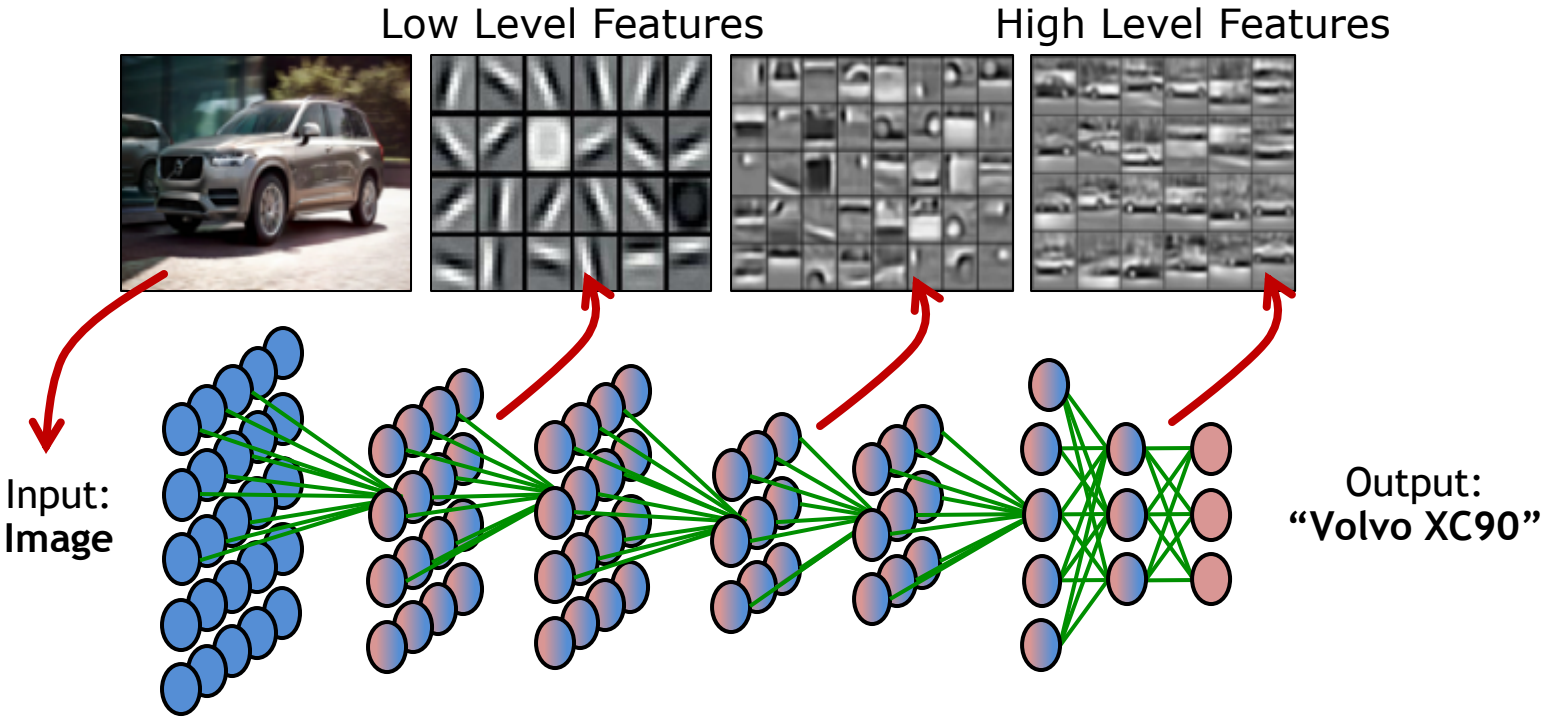


$$y_j = f\left(\sum_{i=0}^{3} W_{ij} \times x_i\right)$$

Deep Learning



Deep Learning Network



Modified Image Source: [**Lee**, *CACM* 2011]

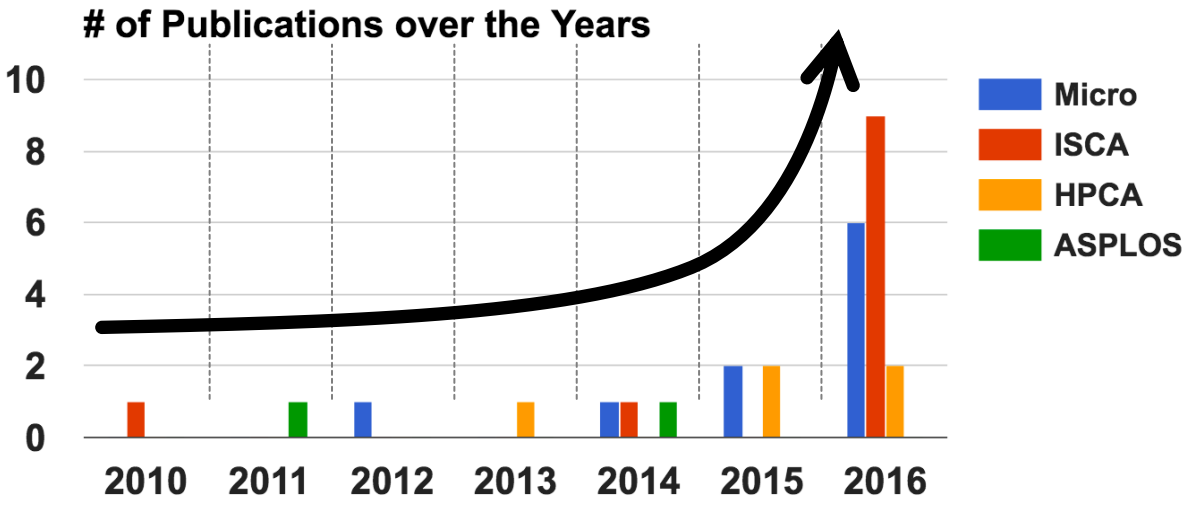
DNN Timeline

- 1940s: Neural networks were proposed
- 1960s: Deep neural networks were proposed
- 1989: Neural network for recognizing digits (LeNet)
- 1990s: Hardware for shallow neural nets
  - Example: Intel ETANN (1992)
- 2011: Breakthrough DNN-based speech recognition
  - Microsoft real-time speech translation
- 2012: DNNs for vision supplanting traditional ML
  - AlexNet for image classification
- 2014+: Rise of DNN accelerator research
  - Examples: Neuflow, DianNao, etc.

Deep Learning Platforms

- CPU
  - Intel, ARM, AMD…
- GPU
  - NVIDIA, AMD…
- Fine Grained Reconfigurable (FPGA)
  - Microsoft BrainWave
- Coarse Grained Programmable/Reconfigurable
  - Wave Computing, Plasticine, Graphcore…
- Application Specific
  - Neuflow, *DianNao, Eyeriss, Cnvlutin, SCNN, TPU, …

DNN-focused Publications at Architecture Conferences



HPCA 2023

Session 1A: Neural Networks and Accelerators 1

Session 4A: Neural Networks and Accelerators 2

Session 7A: Neural Network and Accelerators 3

Also new conferences focusing on addressing compute challenges for ML/DNN, e.g., MLSys, AICAS, TinyML

Next Lecture:

Overview of Deep Neural Networks

Survey

https://forms.gle/YRraBubjSFdgWbr48

Due: Wed, Feb 7