Петербургский институт ядерной физики им. Б. П. Константинова
Национального исследовательского центра «Курчатовский институт»

Artificial Neural Networks in High Energy Physics data processing (succinct survey) and probable future development

Andrey Ye. Shevel

Outline

- Initial remarks
- Case of history
- Artificial Neural Networks (ANN) in High Energy Physics
- ANN root functions, gradient descent
- Challenges, Cost of Training
- Stewardship of scientific data
- Probable trends in ANN development for HEP

Initial remarks

- Artificial Neural Network (ANN) is part of Machine Learning (ML) which in turn is part of Artificial Intelligence (AI)
- Artificial Intelligence is NOT equal to Artificial Intellect. Opposite opinion is kind of anthropomorphism.
- ANN is in use at High Energy Physics several tens of years.

Machine (computer program) learning

- “A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.”
  - From: Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, New York.

Cybernetics is not directly related to ANN

- However the book from W. Ross Ashby (1903-1972) "Introduction to Cybernetics" where he paid much attention to
  - animal nervous system and mentioned fundamental results from
    - I.P. Pavlov (1849-1936) experiments
    - A.A. Markov (1856-1922) chains theory

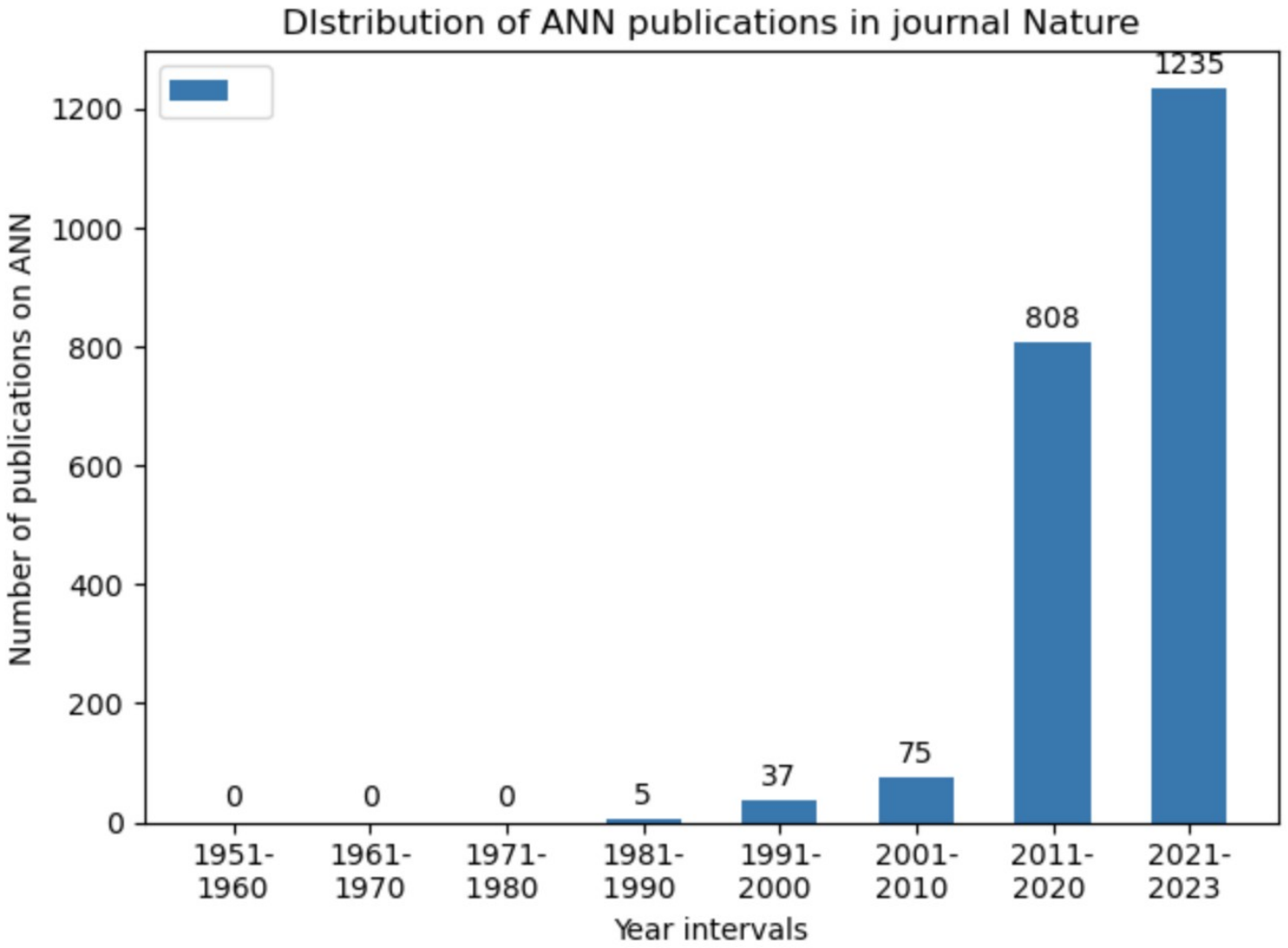
inspired the common interests for Artificial Intelligence.

The nervous system is the organ for surviving any animal.

Persons in AI workshop 1956

From: https://spectrum.ieee.org/dartmouth-ai-workshop

Publications in Nature.com where ANN has been mentioned



Journal Nature is published from November 4, 1869

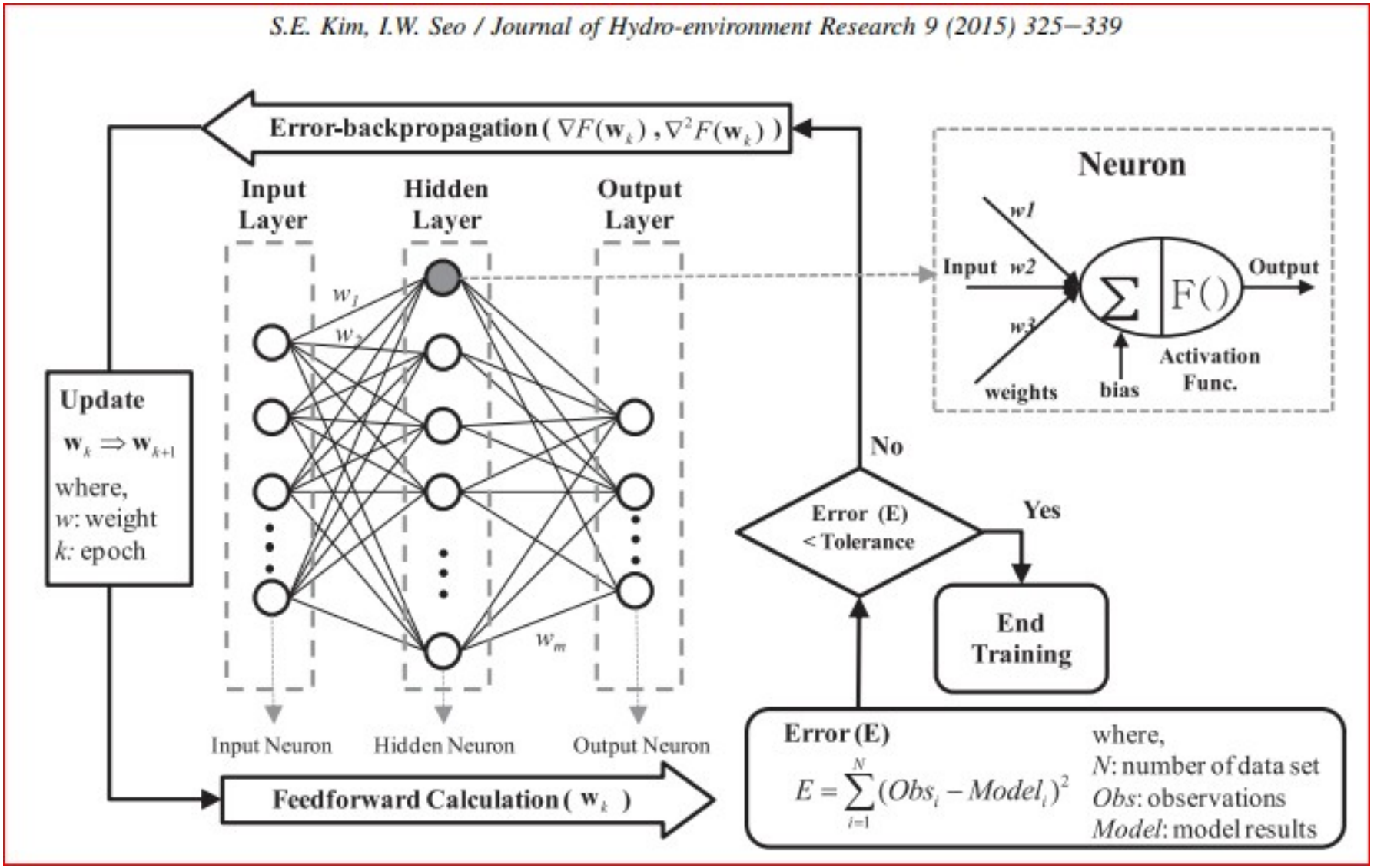
ANN in HEP

- Denby B. «Neural networks and cellular automata in experimental high energy physics» // Computer Physics Communications, Volume 49, Issue 3, June 1988, Pages 429-448 // https://doi.org/10.1016/0010-4655(88)90004-5
- D. Goldner et al “Artificial Neural Networks as a Level-2 Trigger for the H1 Experiment: Status of the Hardware Implementation” // International Journal of Modern Physics C Vol. 06, No. 04, pp. 541-548 (1995) https://doi.org/10.1142/S012918319500040X
- P Kokkas et al “The neural network first level trigger for the DIRAC experiment” // NI&M Vol 471, Issue 3, 1 October 2001, Pages 358-367 // https://www.sciencedirect.com/science/article/abs/pii/S016890020100852X
- In many areas of experiments “Machine Learning in High Energy Physics Community White Paper” // Kim Albertsson et al 2018 J. Phys.: Conf. Ser. 1085 022008 (73 authors)
- Popular software toolkits: TensorFlow, Theano, PyTorch Caffe2, and “TMVA - Toolkit for Multivariate Data Analysis (it not only ANN)” – is part of ROOT.
- Hardware: CPU, GPU, FPGA, TPU, ...

The rise of ANN and ANN model scaling

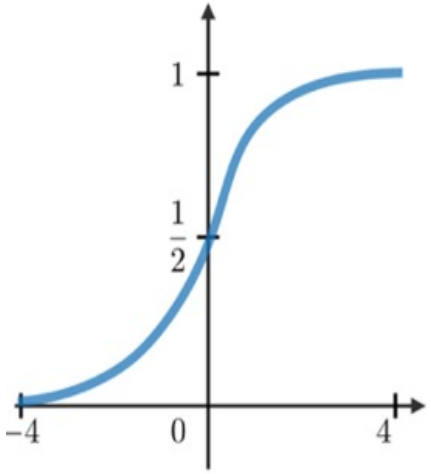
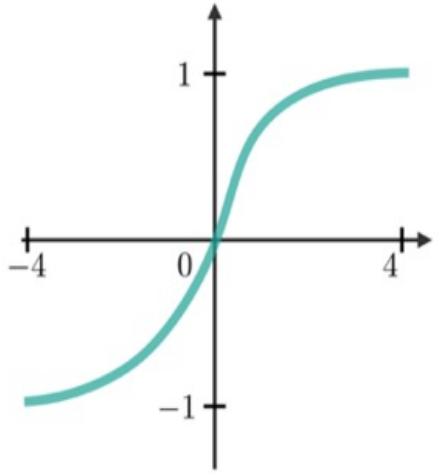
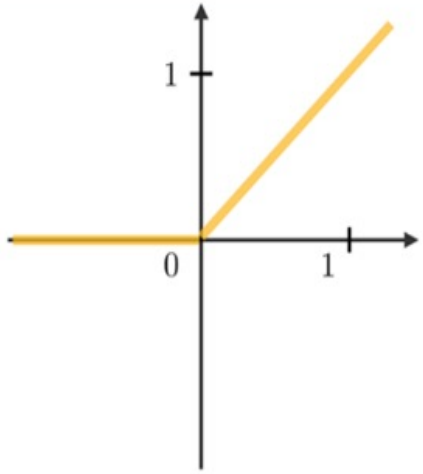
- The rise of interest to ANN is rooted in
  - Computing power growth (including clouds)
    - scaling across 86,400 CPU cores — a virtual cluster with nearly 5 peak petaflops of performance.
      - https://www.hpcwire.com/2020/11/20/azure-scaled-to-record-86400-cores-for-molecular-dynamics/
  - Recorded data volume growth
  - Special equipment dedicated to build up ANN
  - Conferences, competitions, open data sets
- Theories:
  - Approximation of math function.
  - All algorithms for learning [respectively, optimization] do equally well at generalization performance [cost of the found solution] when averaged over all possible problems.

ANN Training algorithm



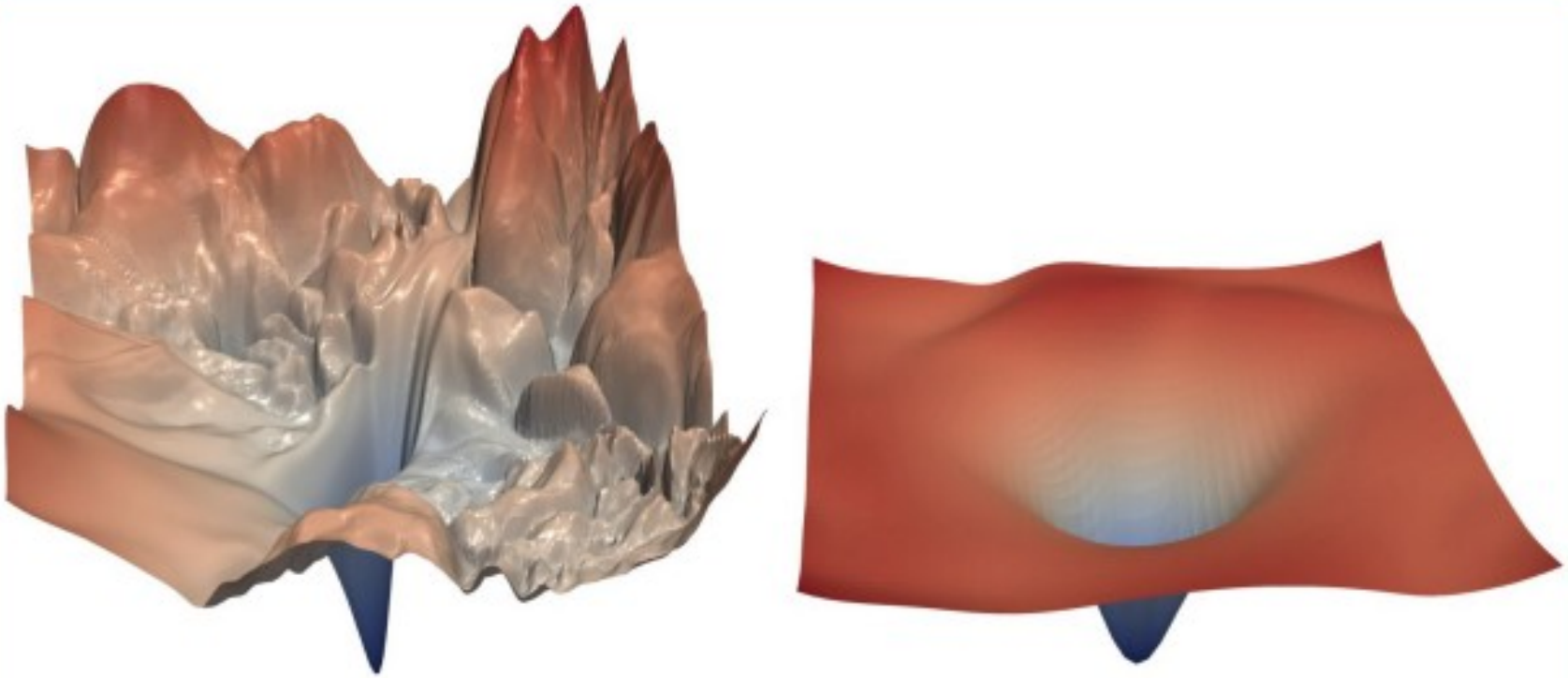
Popular algorithm is Stochastic Gradient Descent (SGD)

Examples of Activation functions

$g(z) = \frac{1}{1+e^{-z}}$	$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$g(z) = \max(0, z)$
		

From left to right: sigmoid, tanh, ReLu

Example of surface

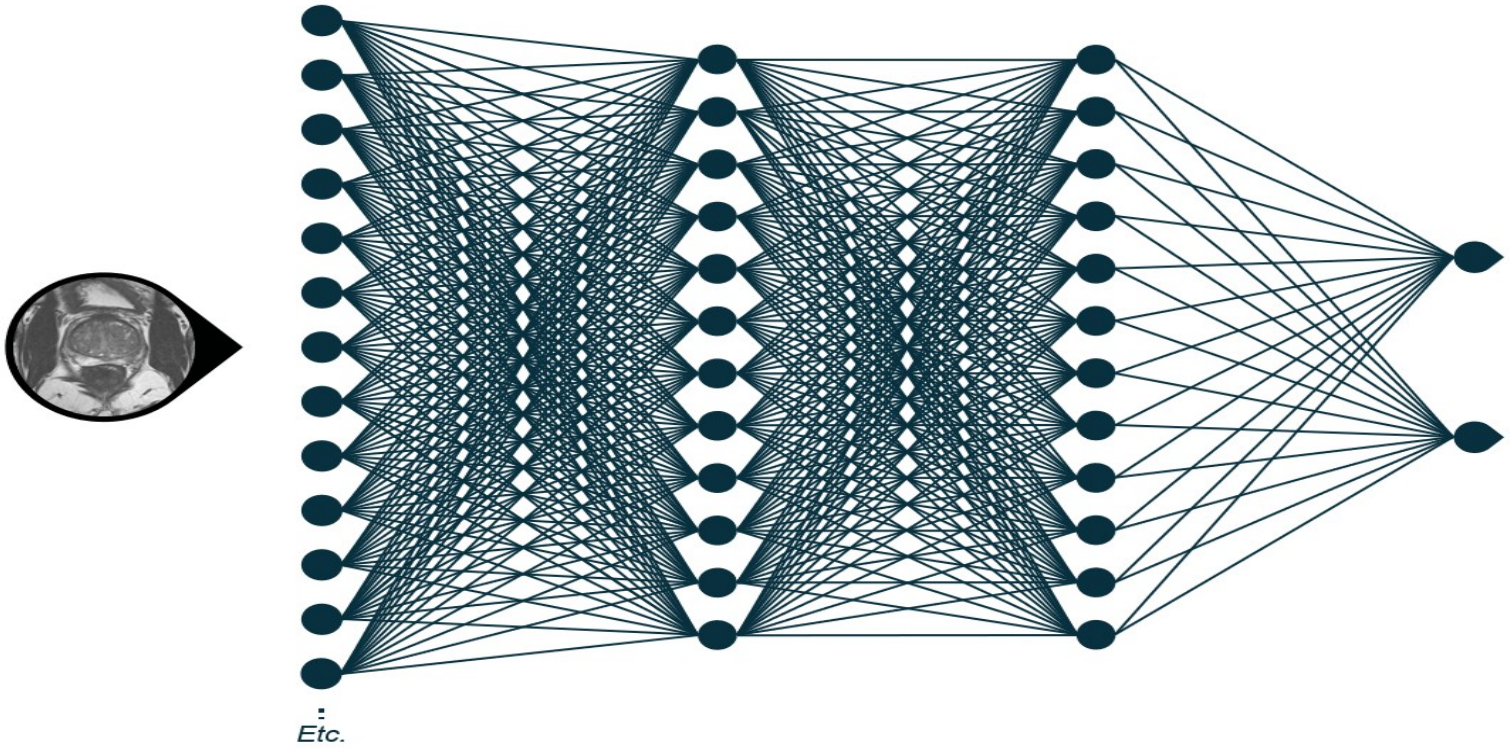


*Hao Li, and et al, Visualizing the Loss Landscape of Neural Nets. NIPS, 2019.

ANN complexity

DEEP LEARNING (DL)

How many weights are there in a neural network?



12 168 weights in total

Quantib

Challenges for simple ANN

- Data preparation for training a model
  - Is very important
- Cost of training
  - Increase accuracy N times might require N^4 rise in cost
- Overfitting:
  - Overfitting happens when a model becomes too good at recognizing patterns in the training data and becomes too specific to that data. This means that it won't perform well on new, unseen data.
- Underfitting:
  - Underfitting is the opposite of overfitting. It occurs when a model is too simple and doesn't have enough capacity to learn the patterns in the data.
- Vanishing gradient:
  - Gradient becomes small or zero
- Hyper-parameters have to be set before training a model and can't be learned from the data.
  - The Number of hidden layers
  - The number of neurons in hidden layers
  - And other hyper-parameters

Data preparation

- IID: Independent and Identically Distributed.
- Data preparation
  - Normalization (or standardization)
  - Decorrelation
  - Cleaning up
  - Dividing the data: training, validation, testing sets

Normalization Technique	Formula	When to Use
Linear Scaling	$x' = (x - x_{min})/(x_{max} - x_{min})$	When the feature is more-or-less uniformly distributed across a fixed range.
Clipping	if x > max, then x' = max. if x < min, then x' = min	When the feature contains some extreme outliers.
Log Scaling	x' = log(x)	When the feature conforms to the power law.
Z-score	x' = (x - μ) / σ	When the feature distribution does not contain extreme outliers.

From https://developers.google.com/machine-learning/data-prep/transform/normalization?hl=en

ANN learning results

- Learning process is resulted in ANN with tuned weights. That ANN could help discover anomaly or internal not obvious relations In between fragments of data in large data pool.
- Obviously that ANN trained on appropriate data pool is able to speed up many HEP tasks e.g. determine elementary particle type in experimental data.
- Very good if it’s possible to use already trained ANN (learning transfer).

Distributed ANN ensemble - federated learning

- Federated learning is the concept to cover collaborative learning scenarios among organizations where is described the privacy-preserving decentralized collaborative machine learning techniques.
- It is considered feature and sample space of the data parties which may not be identical.
- Is it similar to the team of human brains?

Challenges for large scale generative ANN

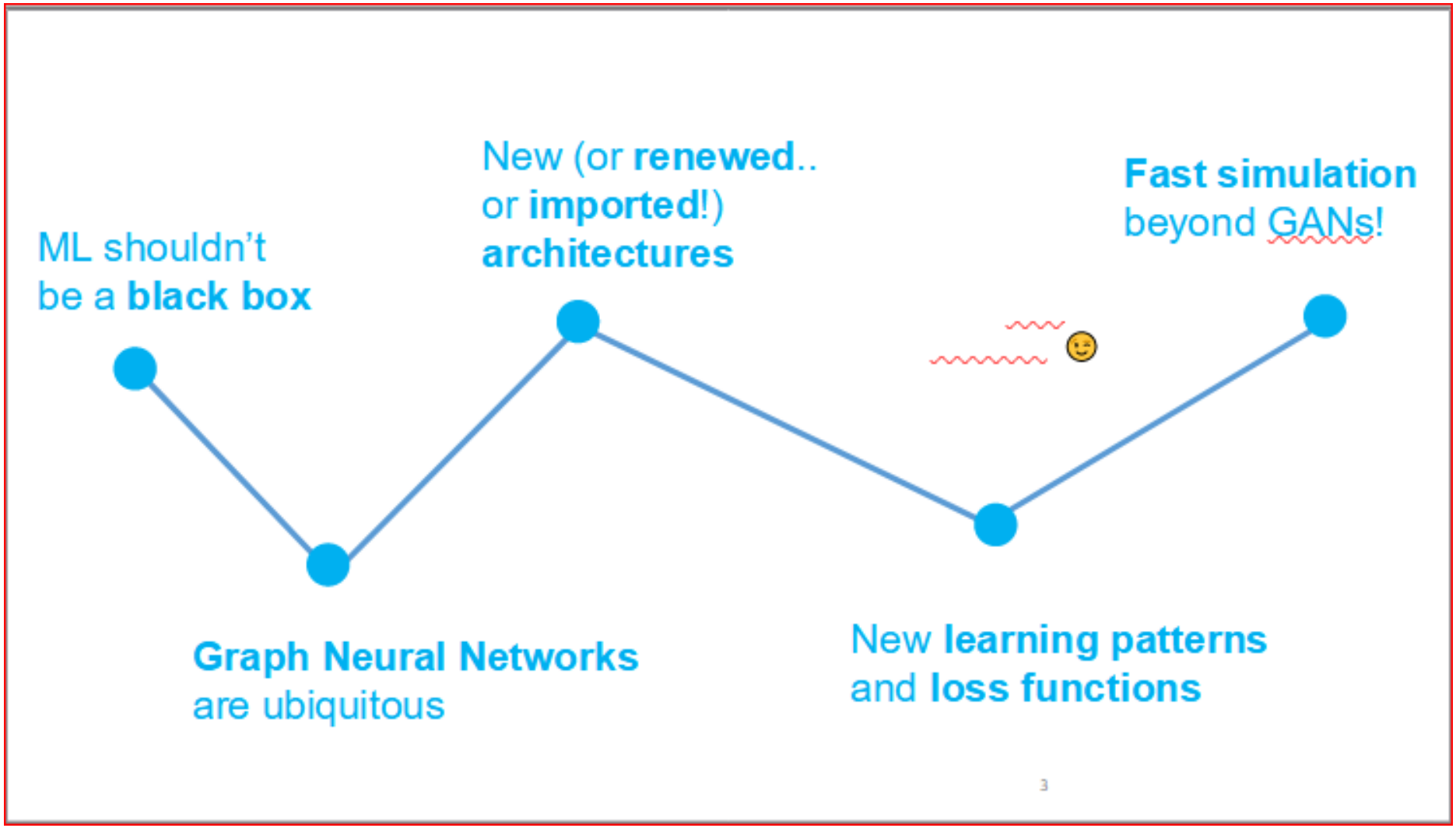
- Data preparation for training is very important
- Explainability/interpretability
- Hallucinations (appeared in Large Language Models – LLM)
- Cost of training (may be tens of $M and more)
  - The cite from OpenAI developer experience [ref 6 in spare slides]: Model performance depends most strongly on scale, which consists of three factors: the number of model parameters N (excluding embeddings), the size of the dataset D, and the amount of compute C used for training. Within reasonable limits, performance depends very weakly on other architectural hyper-parameters such as depth vs. width.

Four principles of Explainable AI

- Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
- Meaningful: A system provides explanations that are understandable to the intended consumer(s).
- Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.
- Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

  - This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8312

ML & ANN at CHEP-2023



Stefano Dal Pra, Jana Schaarschmidt, Sofia Vallecorsa, Sandro Wenzel

Open access to experimental data & FAIR principles

- Open access to experimental data has to be guaranteed in 2025 (from https://www.science.org/doi/epdf/10.1126/science.adg8142).
- Findable,
- Accessible,
- Interoperable,
- Reusable
  - These principles are designed to guide data resources, tools, vocabularies, and infrastructures to assist discovery and reuse by third parties. The principles are related but independent and separable, and they define characteristics that contemporary data resources should exhibit to assist discovery and reuse by third parties.
  - Nikil Ravi et al & Ian Foster “FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy”

Probable future development trends

- It is possible to expect two big trends for ANN application in HEP experiments and around:
  - Interactive Language Models (LM) for large installations (detectors, computing systems, etc) with prompt engineering.
  - Data analysis with ANN implemented foundation model(s) for HEP [like it was implemented for biology (AlphaFold2), business (Bloomberg), and a range of theorem provers in mathematics]

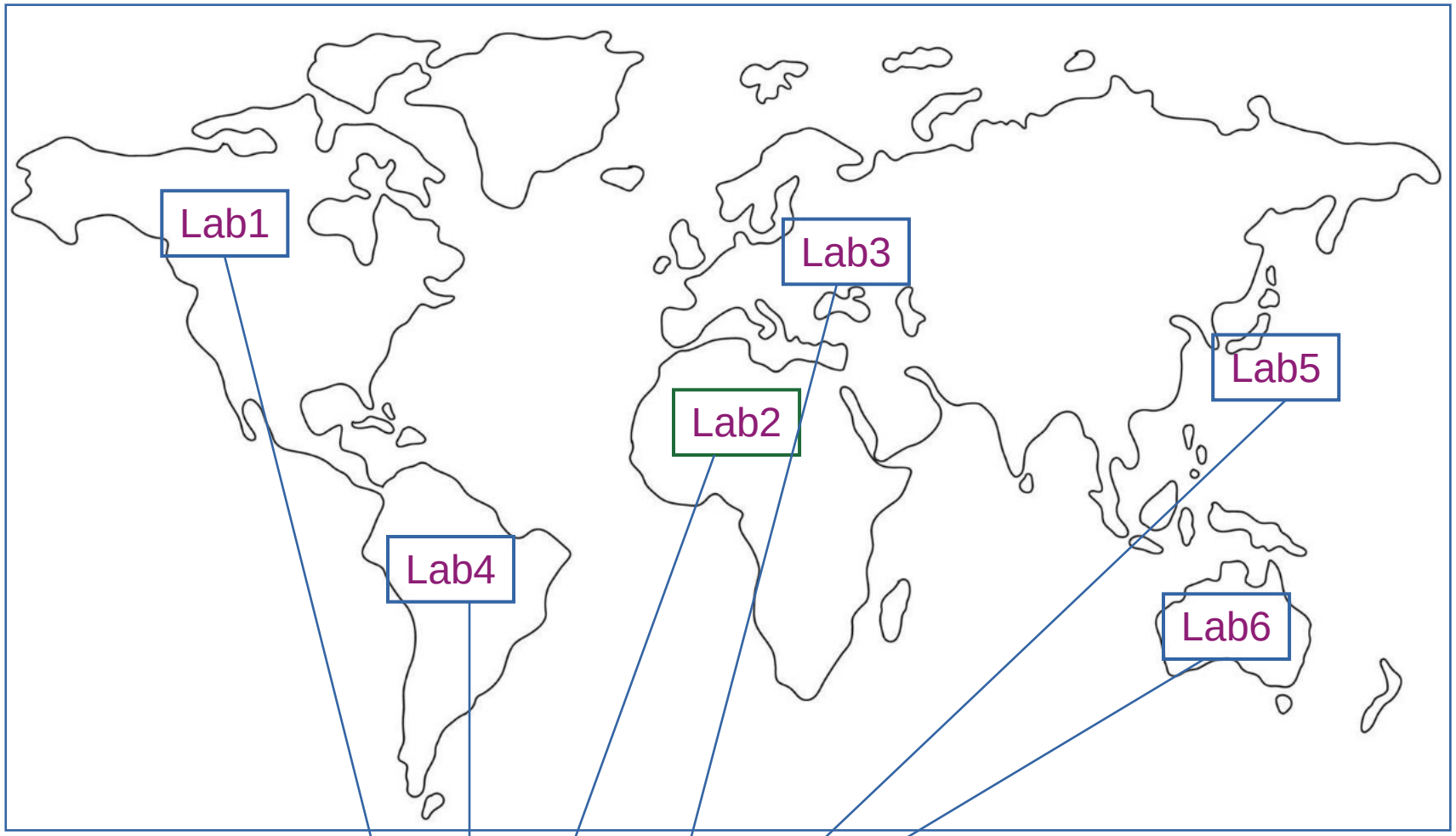
ANN for large experimental system

1. The problem
    1. A large volume of documentation (technical descriptions, administrative orders, operating manuals, etc), as well as a volume of logs (automatic and semi-automatic log records) about the functioning of the entire system.
    2. A meaningful analysis (obtaining an answer to a specific question based on all available data) of such a large amount of data (hundreds of GB or more) is a non-trivial task, which in many cases turns out to be labor- and time-consuming.
2. Possible development
    1. It seems reasonable to undertake the development of a special expert system (SES) using ANN technology, which could provide the operator (system administrator) with effective assistance in the described task.
3. Possible implementation
    1. The using of local computing facilities with appropriate model with regular re-training to take into account new log records.

Idea of Foundation Federated ANN for global experimental data analysis

1. The problem
    1. The experimental data is distributed around different laboratories in the World
    2. The experimental data might be from related areas (elementary particle, nuclear, astro physics, and other related experiments).
    3. In each laboratory the analysis might be implemented in form of local ANN which is trained on local experimental data
    4. The global data analysis might use the centralized ANN which in turn can use a number of local ANN from related experimental areas.
2. Possible implementation
    1. The use of federated ANN architecture.
    2. There are hardware examples in data centers for AI [ref 1-4 in spare slides] for ANNs.
    3. May be possible to use existing WLCG framework.

Foundation Federated ANN for HEP



Nearest needs for ANN in HEP and around

- Developments of the the methods and procedures to build up foundation models for physics.
- Security issues: how to minimize viruses threat in ANN.
- The theories about architectures and features of ANN which are applicable to physics experiment environment.
- It seems that each large experimental setup and/or system has to be accompanied with appropriate ANN.

Final remarks

- Globally we don’t need to be afraid general AI based on machine deep learning.
- Main threat for humanity is human “deep stupidity”.

Thank you!

Spare slides

References

1) Graphcode launches wafer-on-wafer ‘Bow’ IPU
https://www.hpcwire.com/2022/03/03/graphcore-launches-wafer-on-wafer-bow-ipu/

2) NVIDIA AI Platform Delivers Big Gains for Large Language Models https://developer.nvidia.com/blog/nvidia-ai-platform-delivers-big-gains-for-large-language-models/

3) Training 175B Parameter Language Models at 1000 GPU scale with Alpa and Ray
https://www.anyscale.com/blog/training-175b-parameter-language-models-at-1000-gpu-scale-with-alpa-and-ray

4) An in-depth look at Google’s first Tensor Processing Unit (TPU)

    1) https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu
    https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

5) ALCF AI Testbed https://www.alcf.anl.gov/alcf-ai-testbed

6) Jared Kaplan et al “Scaling Laws for Neural Language Models” arXiv:2001.08361v1 [cs.LG] 23 Jan 2020 (p.30) https://arxiv.org/abs/2001.08361

7) Types of artificial neural networks https://en.wikipedia.org/wiki/Types_of_artificial_neural_networks

8) Artificial Intelligence for High-Energy Physics https://doi.org/10.1142/12200

AI Conferences and related sites

- Conference on Neural Information Processing Systems (NeurIIPS) - https://nips.cc
- The International Conference on Learning Representations (ICLR) https://iclr.cc/
- The International Conference on Machine Learning (ICML) https://icml.cc
- Models, datasets, etc (over 120K models, 20K datasets, and 50K demos) https://huggingface.co
- Competitions Kaggle.com

Popular models for ANN

- Feed Forward Neural Networks (FFNN)
- Convolutional (CNN)
- Recurrent (RNN)
- Reinforcement Learning (RLNN)
- Reinforcement Learning with Human Feedback (RLHF)
- [Variational]AutoEncoder (AE/VAE)
- Generative Adversarial NN
- Graph NN
- Physics-informed neural networks
- Also pls see [ref 7 in spare slides]

Popular Languages for AI

- Python https://www.python.org
- R https://www.r-project.org
- Julia https://julialang.org
- RUST https://www.rust-lang.org

Open data sets for ANN training

- Datasets https://paperswithcode.com/datasets?mod=physics&page=1
- Supersymmetric Particles Data Set
  https://www.kaggle.com/datasets/omidbaghchehsaraei/supersymmetric-particles-data-set
- ML Physics Portal http://mlphysics.ics.uci.edu
- List of datasets for machine-learning research
  https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

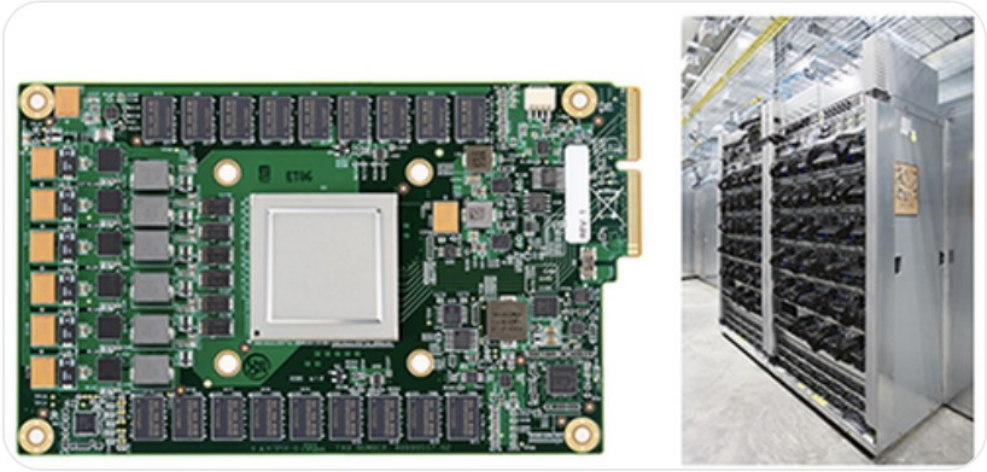
Popular AI tools for HEP and other info

- Best AI Tools Directory https://www.insidr.ai/ai-tools/
- Machine learning with ROOT https://root.cern/manual/tmva/
- CERN Inter-experimental Machine Learning (IML) Working Group https://iml.web.cern.ch/homepage

Tensor Processing Unit (TPU)

Google 2013
Special chip: ASIC
28nm, 700 MHz, 40W, PCIe Gen3x16

Available over Google cloud as well.



Google's first Tensor Processing Unit (TPU) on a printed circuit board (left); TPUs deployed in a Google datacenter (right)

From https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

BOW POD 256

From https://www.graphcore.ai/products/bow-pod256

ANN Installations

- Argonne Leadership Computing Facility (ALCF) AI Testbed
- SambaNova
- Cerebras

Company	Parameters	Reference
Tesla	7,360 * A100 GPUs	https://www.hpcwire.com/2022/08/16/tesla-gooses-its-gpu-powered-ai-super-is-dojo-next/
NERSC's Perlmutter	6,144 * A100 GPUs	https://docs.nersc.gov/systems/perlmutter/architecture/ , https://www.hpcwire.com/2022/08/16/tesla-gooses-its-gpu-powered-ai-super-is-dojo-next/
Nvidia's own in-house A100 system	4,480 * A100 GPUs	https://www.hpcwire.com/2022/08/16/tesla-gooses-its-gpu-powered-ai-super-is-dojo-next/
RSC Meta	6,080 * A100 GPUs	https://www.hpcwire.com/2022/01/24/metas-massive-new-ai-supercomputer-will-be-worlds-fastest/

Examples of Foundation ANN

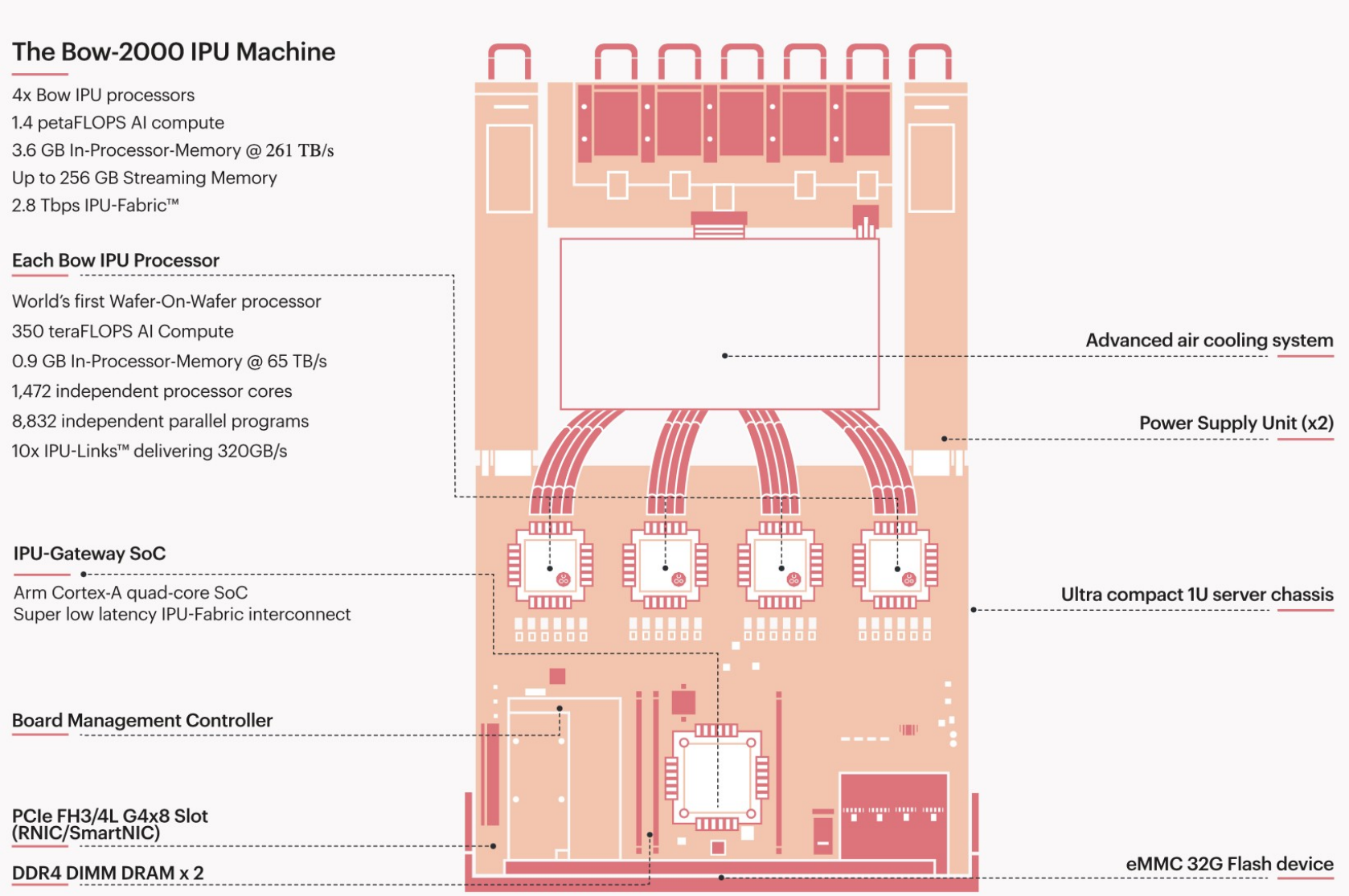
- GPT (3&4) (OpenAI.com)
- DALL-E2 (OpenAI.com)
- ALPHA-Fold-2 (Google DeepMind)
- BloombergGPT [https://doi.org/10.48550/arXiv.2303.17564]
- LlaMA (Meta AI) [https://doi.org/10.48550/arXiv.2302.13971]

- Modern Large Language Models (LLMs) have billions of parameters, are trained on trillions of tokens, and cost millions of dollars.

Nvidia

	H100 SXM	H100 PCIe
FP64	**34 TFLOPS**	**26 TFLOPS**
FP64 Tensor Core	**67 TFLOPS**	**51 TFLOPS**
FP32	**67 TFLOPS**	**51 TFLOPS**
TF32 Tensor Core	**989 TFLOPS***	**756 TFLOPS***
BFLOAT16 Tensor Core	**1,979 TFLOPS***	**1,513 TFLOPS***
FP16 Tensor Core	**1,979 TFLOPS***	**1,513 TFLOPS***
FP8 Tensor Core	**3,958 TFLOPS***	**3,026 TFLOPS***
INT8 Tensor Core	**3,958 TOPS***	**3,026 TOPS***
GPU memory	**80GB**	**80GB**
GPU memory bandwidth	**3.35TB/s**	**2TB/s**
Decoders	**7 NVDEC** **7 JPEG**	**7 NVDEC** **7 JPEG**
Max thermal design power (TDP)	**Up to 700W (configurable)**	**300-350W (configurable)**
Multi-Instance GPUs	**Up to 7 MIGS @ 10GB each**	

From https://resources.nvidia.com/en-us-gpu-resources/h100-datasheet-24306?lx=CPwSfP



From https://www.graphcore.ai/products/bow-2000

Graphcore Bow Pod256 hardware

IPU-Machines	64x Bow-2000 blades
IPUs	256x Bow IPU processors (4 in each Bow-2000)
IPU-Cores™	376,832
Worker threads	2.26 million
AI compute	89.234 petaFLOPS AI (FP16.16) compute
	22.309 petaFLOPS FP32 compute
Memory	Up to 8422.4 GB (includes 230.4 GB In-Processor-Memory (64x 3.6 GB per Bow-2000) and 8192 GB Streaming Memory (64x 64 GB DIMM x2 per Bow-2000)

From https://docs.graphcore.ai/projects/bow-pod256-datasheet/en/latest/product-description.html

From https://www.graphcore.ai/products/bow-2000