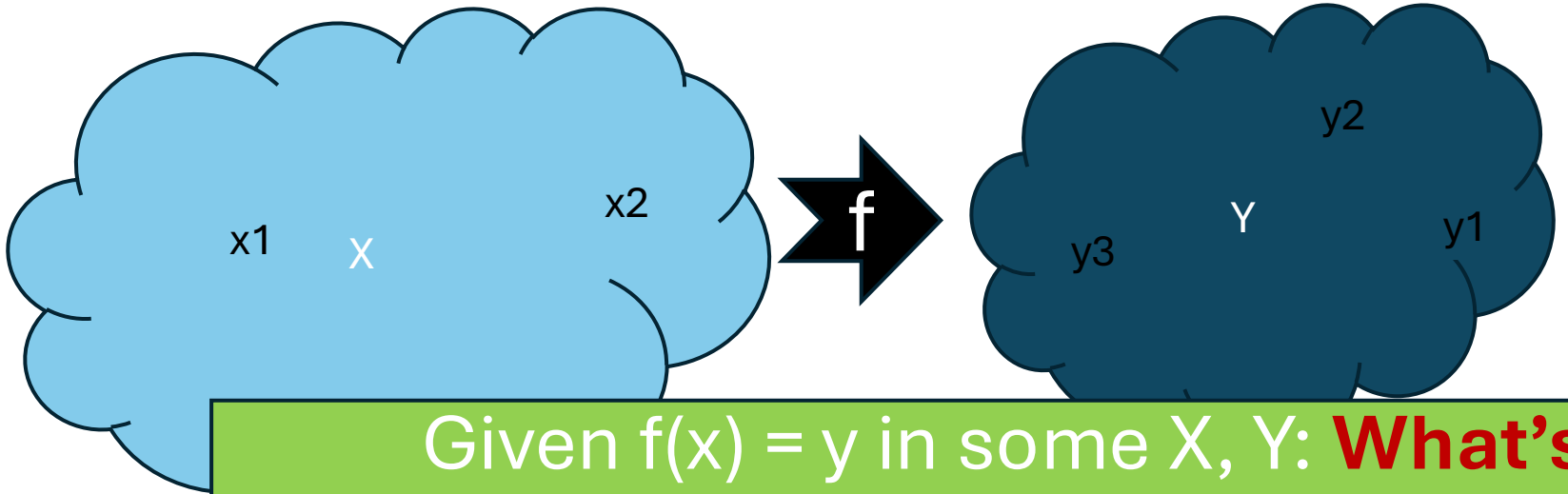
A whirlwind tour through the wonderful world of ML and modern processors

Konrad Handrick

Workshop 1: stepping stone in AI

November 2025

Pursuit



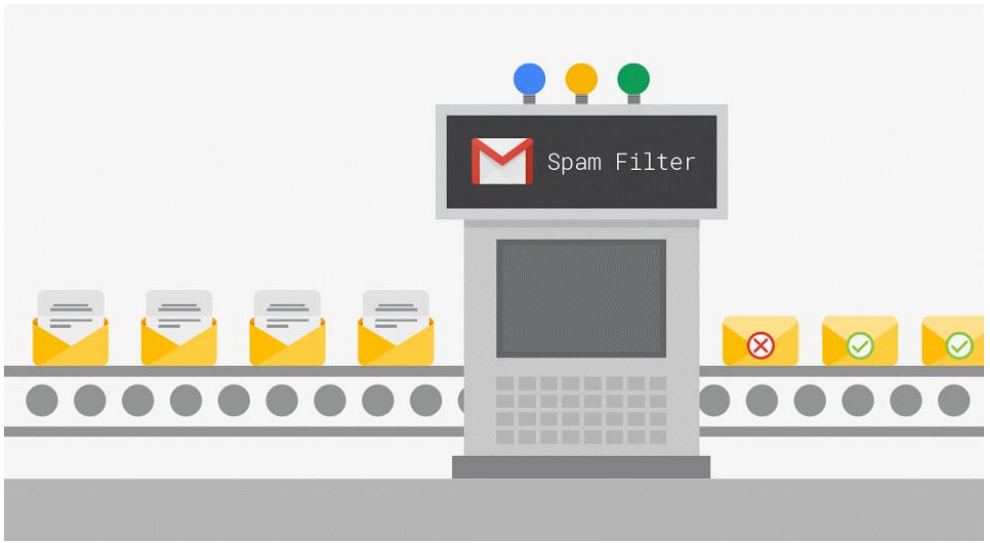
Given f(x) = y in some X, Y: **What's f?**

1. Can we automate finding f?
   -> Training

2. Having found f: How do we make it **fast**?
   -> Inference optimization (our TODO)

Applications

classification



Left : https://chools.in/wp-content/uploads/NLP-1.gif
Right: https://www.ge.com/news/sites/default/files/Reports/2020-03/GIF.gif,

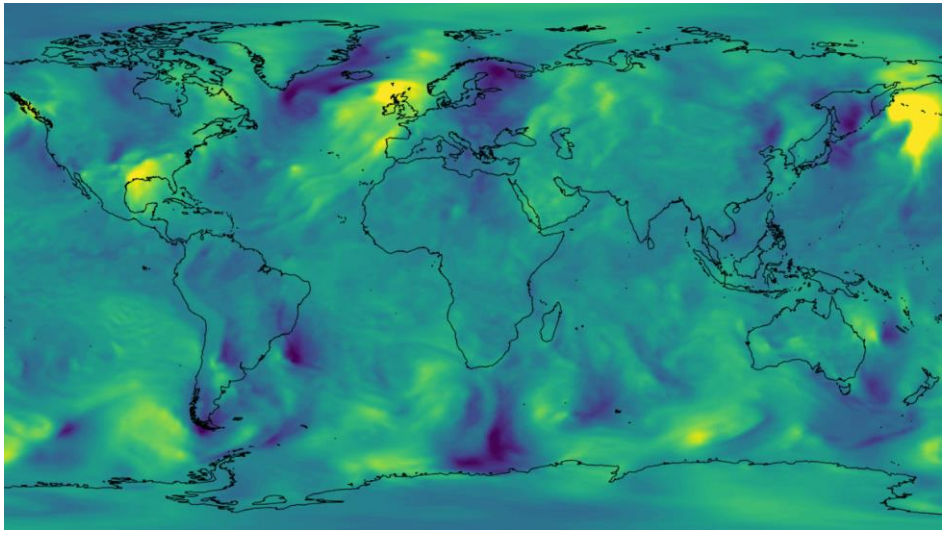
Applications

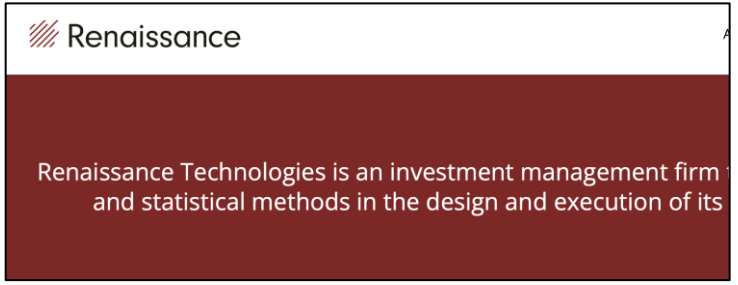
classification

Predicting categorical outputs from input data using labeled examples. The goal is to predict a class label from a list of possibilities.

Applications

prediction

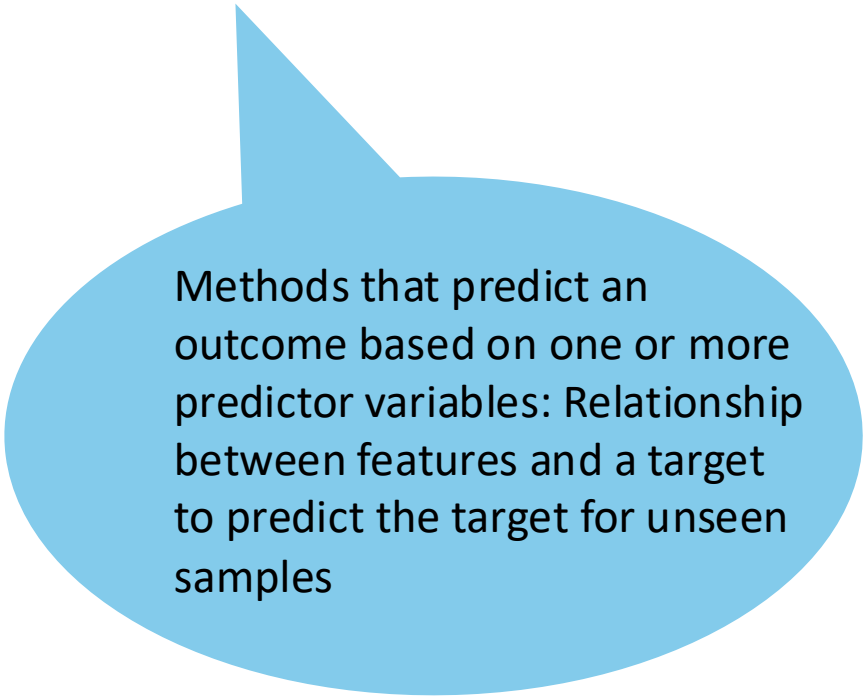




Left: https://yale-threesixty.transforms.svdcdn.com/production/European-AI-Weather-Forecast_ECMWF-HEADER.gif
Right: https://www.rentec.com/

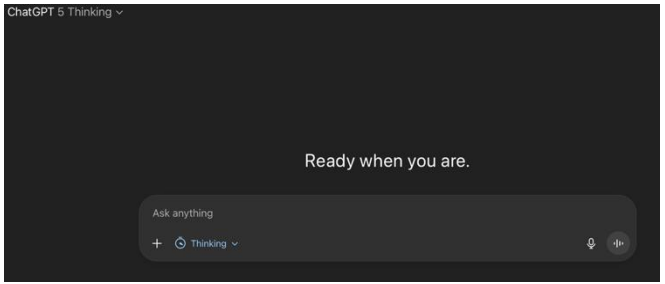
Applications

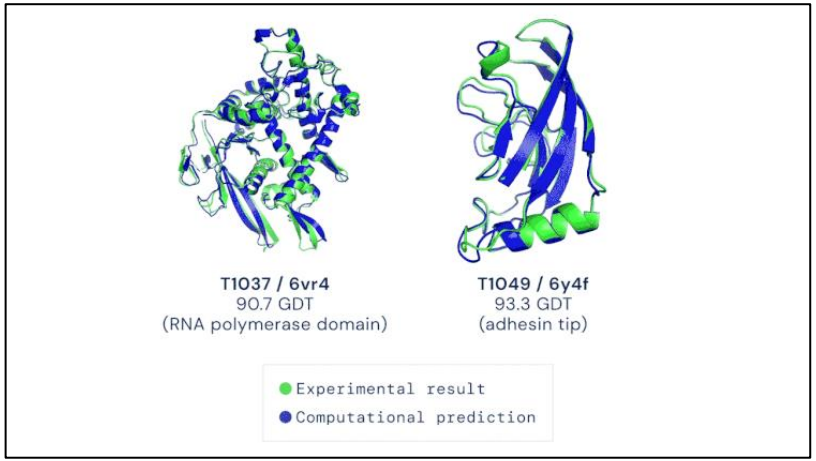
classification | prediction



Applications

generation

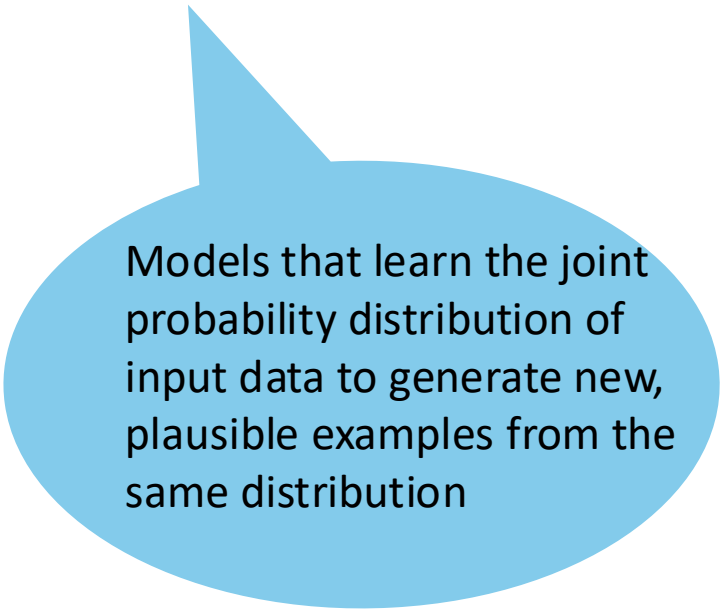




Left: https://www.chat.com
Right: https://images.squarespace-cdn.com/content/v1/5a6ef51551a584bd47153b61/beb1be2e-82b7-4cbc-94a3-33fb8a560045/62277caa7b7ddd67e5c7a0ad_Fig%2B2.gif,

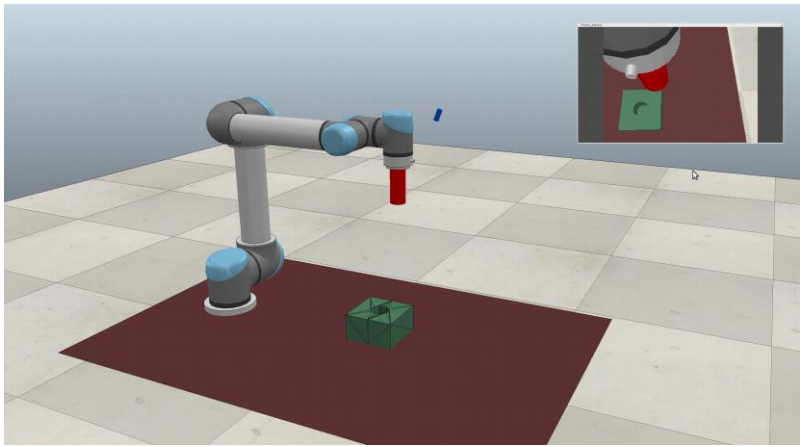
Applications

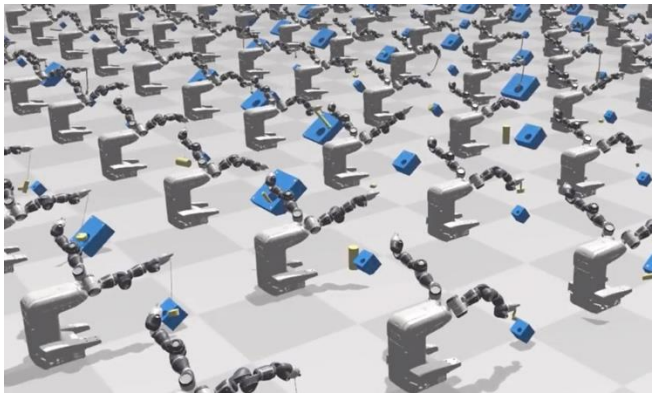
classification | prediction | generation



Applications

Reinforcement Learning (RL)

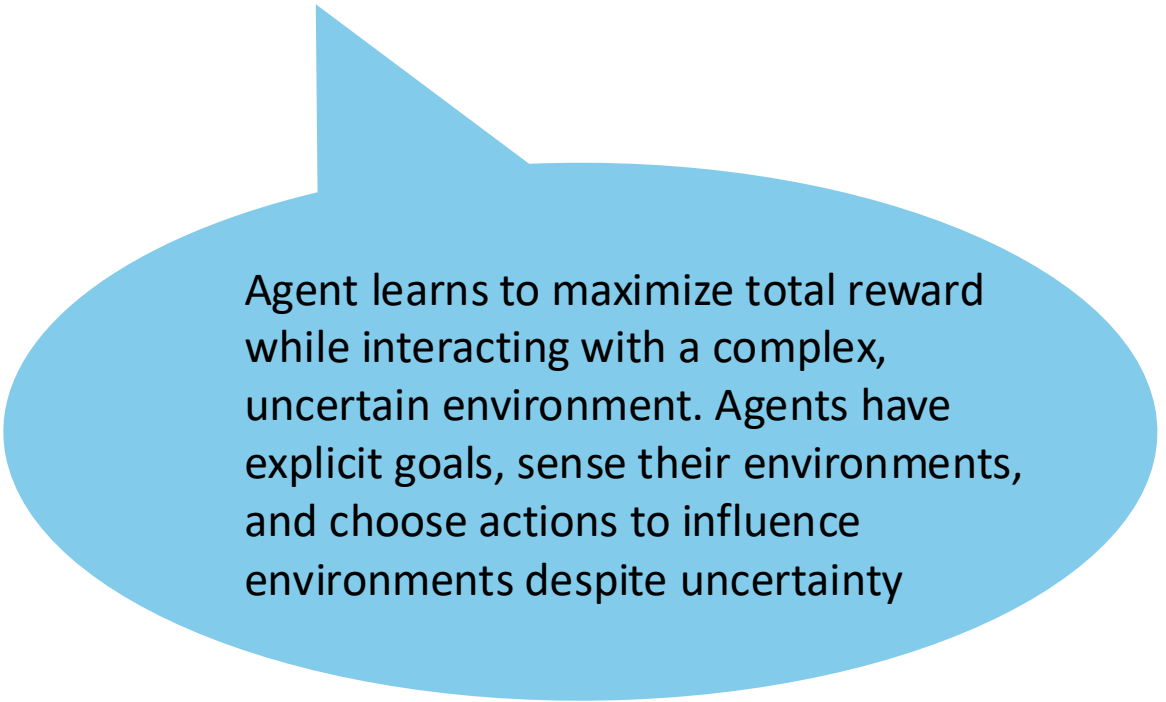




Left: https://dtransposed.github.io/blog/2020/10/21/Robotic-Assembly/
Right: https://spectrum.ieee.org/nvidia-brings-robot-simulation-closer-to-reality-by-making-humans-redundant

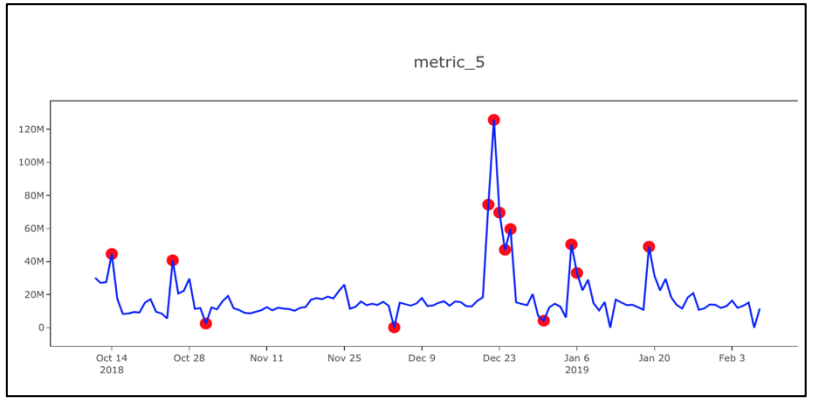
Applications

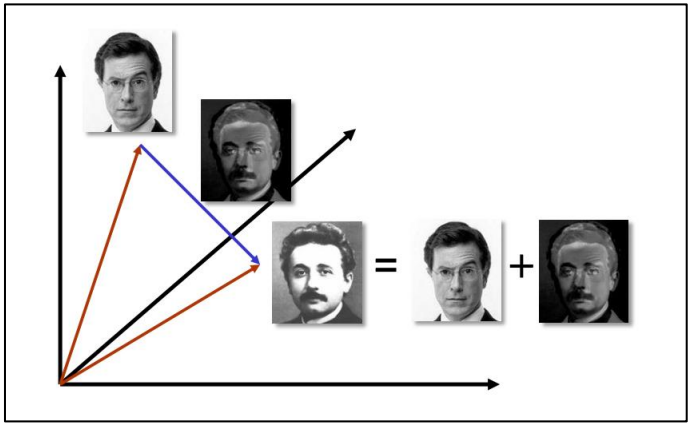
classification | prediction | generation | RL

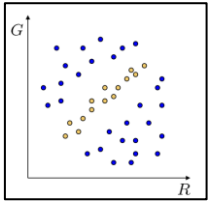


Applications

unsupervised learning



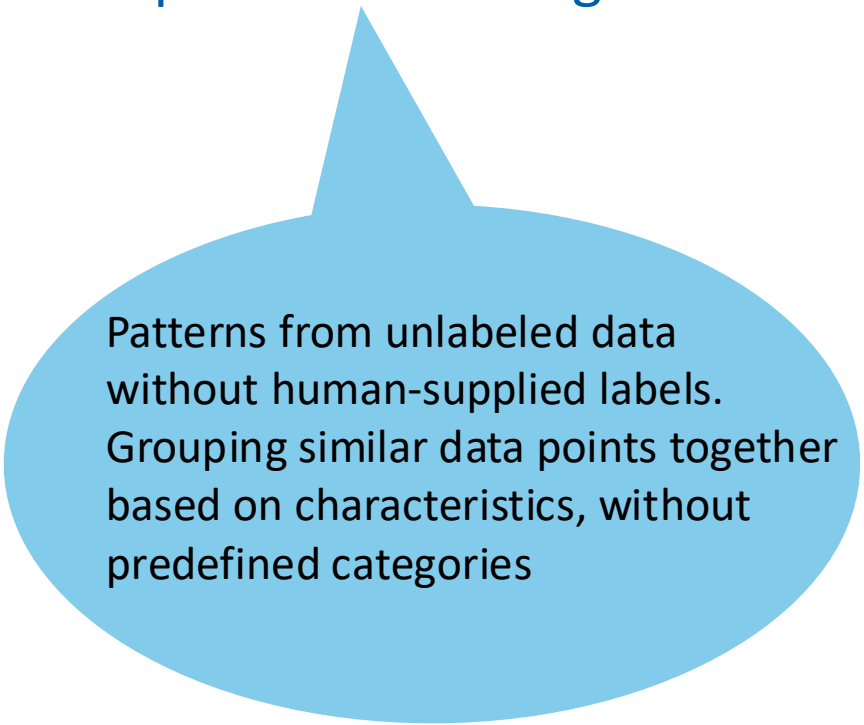




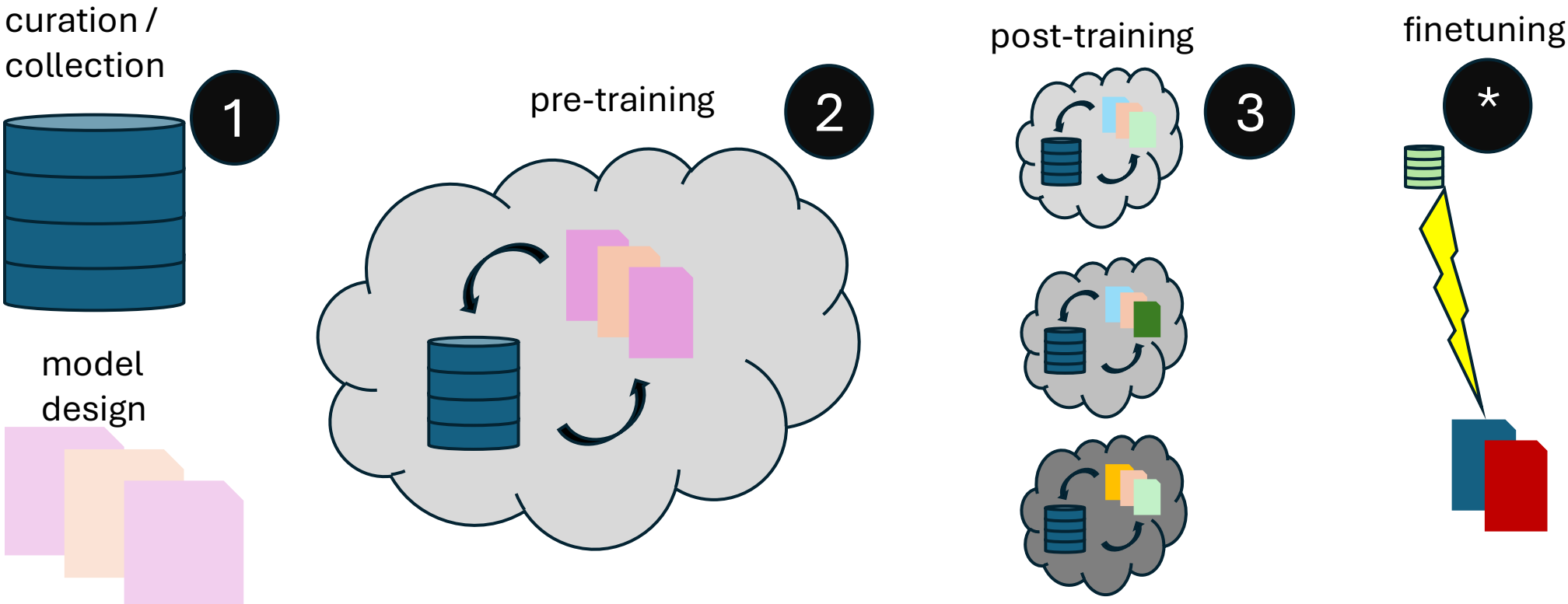
Left: https://www.analyticsvidhya.com/blog/2022/05/an-end-to-end-guide-on-anomaly-detection/
Right: https://www.cs.toronto.edu/~guerzhoy/320/lec/pca.pdf

Applications

classification | prediction | generation | RL | unsupervised learning

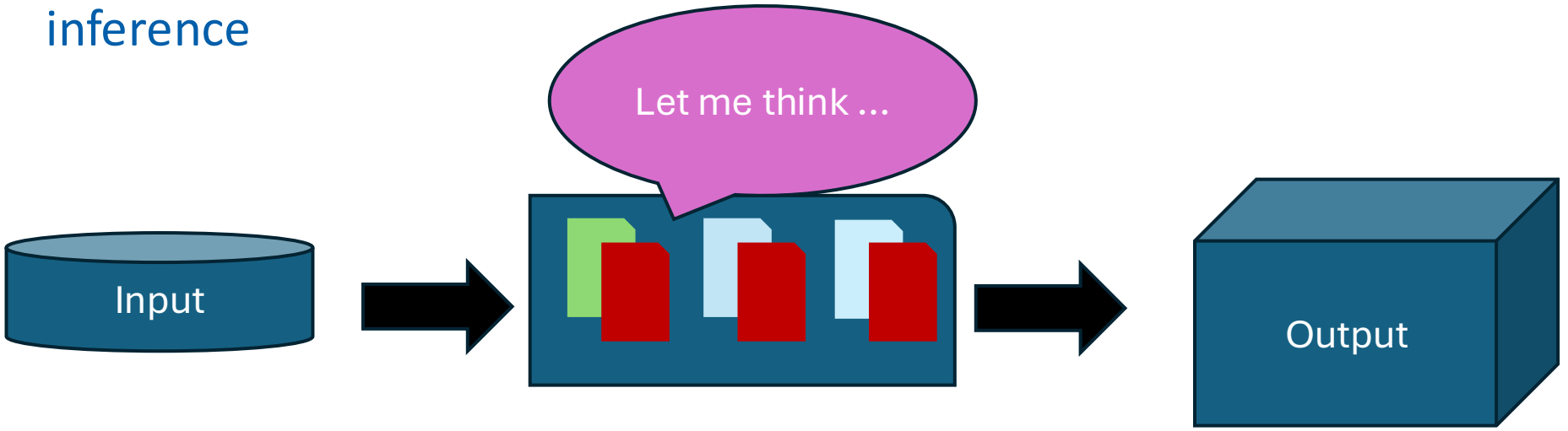


Learning?


curation / collection
1
model design
pre-training
2
post-training
3
finetuning
*

Intelligence?

inference



AI takeaways & glossary

Takeaways

- models are architected by vibe, depending on data and problem
- distinction: training, posttraining, inference
- five domains for AI models
- we're interested in inference

Glossary

Model: Sequence of operations that map an input to an output.

Training: Algorithmic procedure to find an unknown mapping between input and output.

Inference: Receive output from a model.

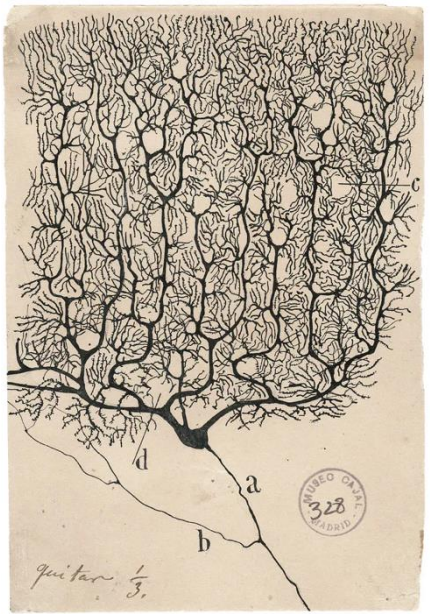
Pretraining: Data-intensive first training of a model in the most general setting possible.

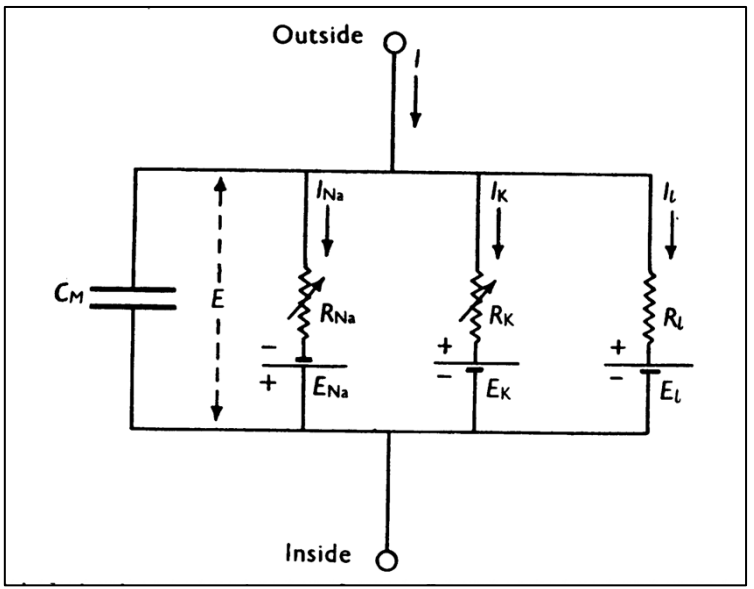
Post-training: Collection of procedures to adapt model to given problem at hand.

Reinforcement Learning: Set up model as an agent in a dynamic environment and, depending on its reaction to changes in the environment, modify it.

Intelligence -- physiology -> physics

We're going to save us time thinking about what intelligence actually means from the Bible to modern psychological thought and immediately go what natural sciences came up with.





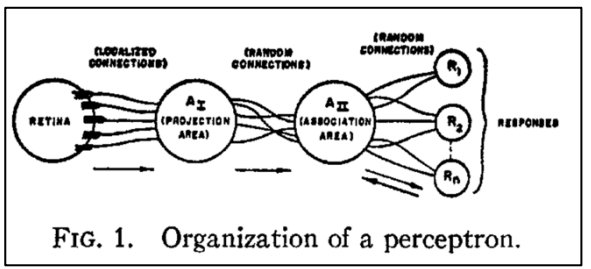


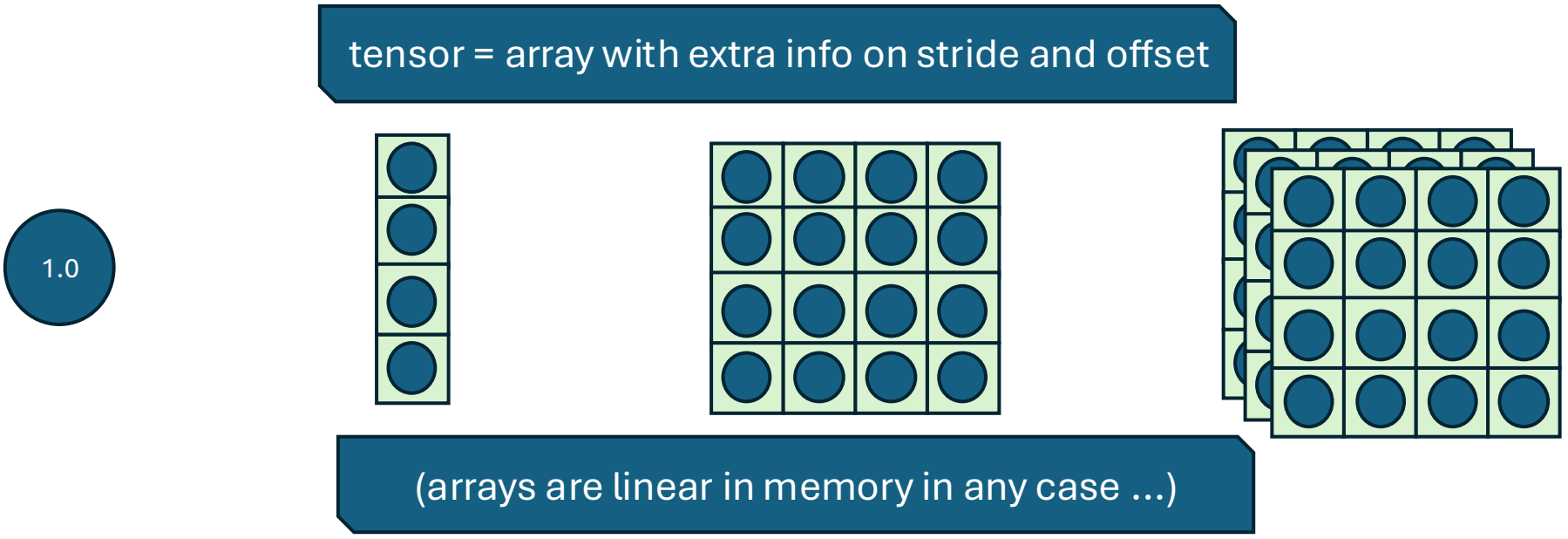
FIG. 1. Organization of a perceptron.

Left: https://publicdomainreview.org/collection/illustrations-of-the-nervous-system-golgi-and-cajal/
Center: https://pmc.ncbi.nlm.nih.gov/articles/PMC1392413/
Right: https://psycnet.apa.org/record/1959-09865-001

Interlude

The language of modern science and engineering is linear algebra. What's that about?

tensor = array with extra info on stride and offset



0-dimensional tensor / scalar

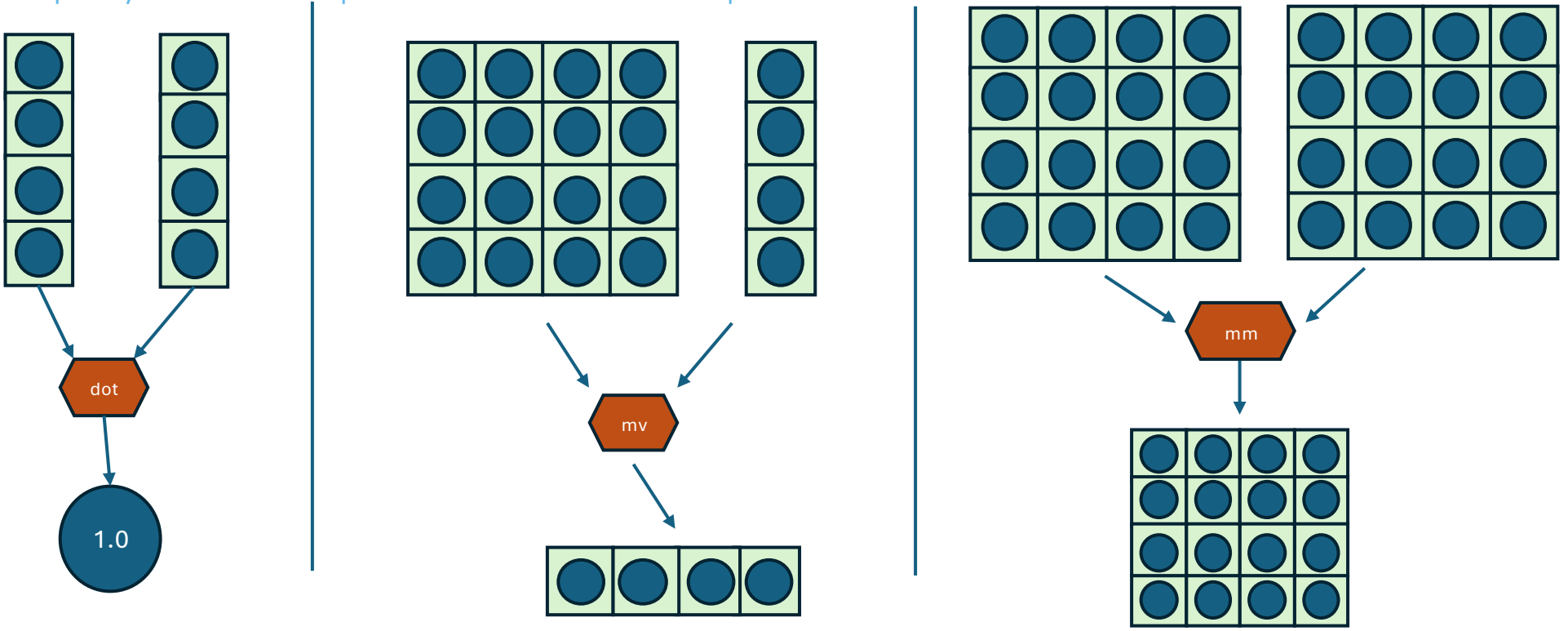
1-dimensional tensor / vector

2-dimensional tensor / matrix

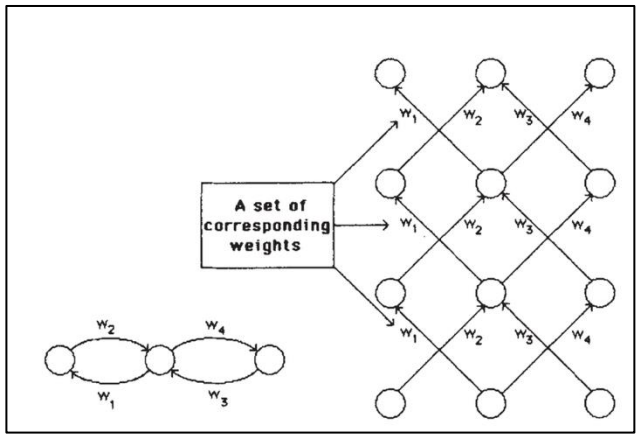
3-dimensional tensor

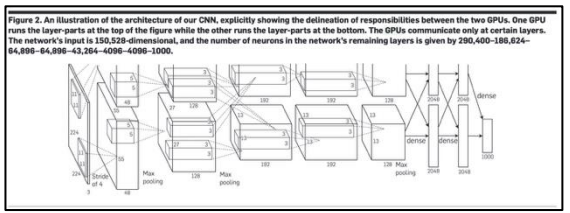
Interlude

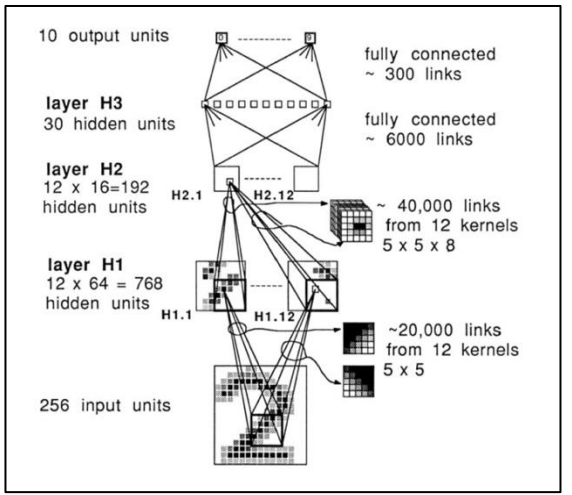
For software people, one could consider these operations similar to assembly programming:
Complexity arises from compositions of these and similar operation.



Intelligence – physics -> theory







Left: https://www.nature.com/articles/323533a0
Center: https://dl.acm.org/doi/10.1145/3065386
Right: https://ieeexplore.ieee.org/document/6795724

Intelligence – systems engineering

Attention Is All You Need

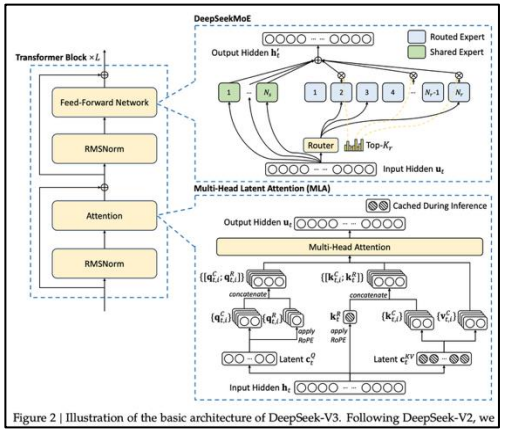


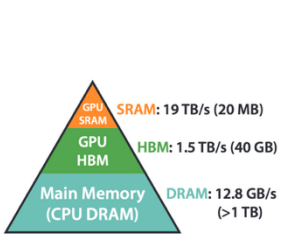
Figure 2 | Illustration of the basic architecture of DeepSeek-V3. Following DeepSeek-V2, we

DeepSeek-V3 Technical Report

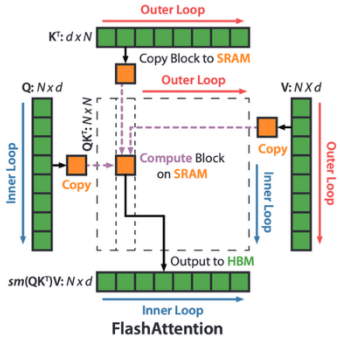
DeepSeek-AI

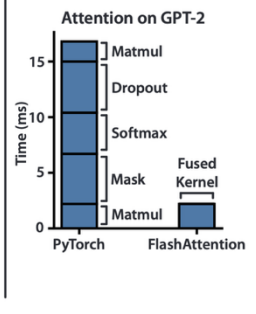
research@deepseek.com

FLASHATTENTION: Fast and Memory-Efficient Exact Attention with IO-Awareness



Memory Hierarchy with Bandwidth & Memory Size





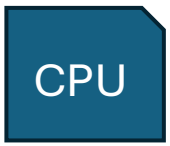
Left: https://arxiv.org/abs/1706.03762
Center: https://arxiv.org/abs/2412.19437
Right: https://arxiv.org/abs/2205.14135

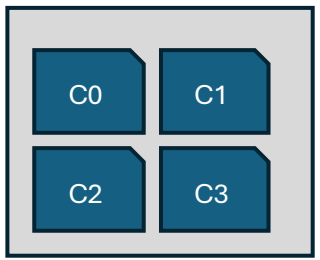
ML & Intelligence – Takeaways

Takeaways

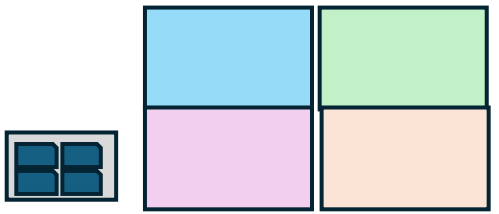
- we've moved quickly from raw chemical mechanisms to making intelligence a systems engineering challenge
- ML is linear algebra (LA)
- ... and LA is highly parallel

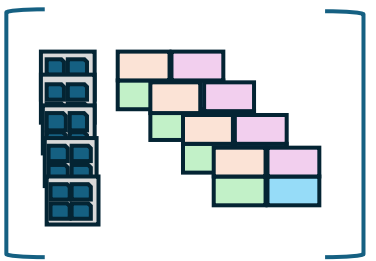
Processors – Taxonomy / Evolution

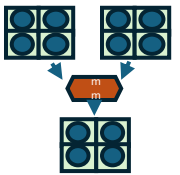




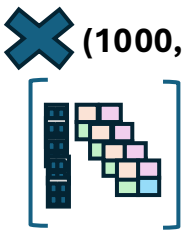
x188 (RTX 6000 Pro Blackwell)
= 26064 threads







I'm also pretty good at GEMM



Let me work on my Task

Let C0 serve the frontend …

Let SM 0 work on the upper part of the frame …

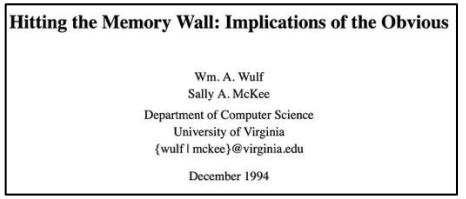
Let me render this movie …

Let me train this model …

Processors – Modernity

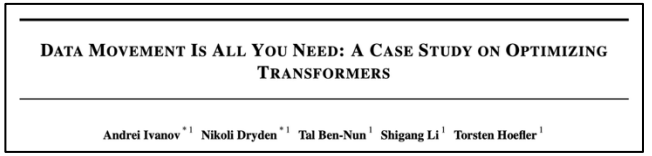
- Rule of thumb: processor speed (FLOPs) doubles every 18 months, bus bandwidth (B/c) only every 36 months!

- weird architectures: TPU, NPU, XPU …
- NVL (Mellanox, 2017 acquisition)
- -> up to several TB/s on a single node

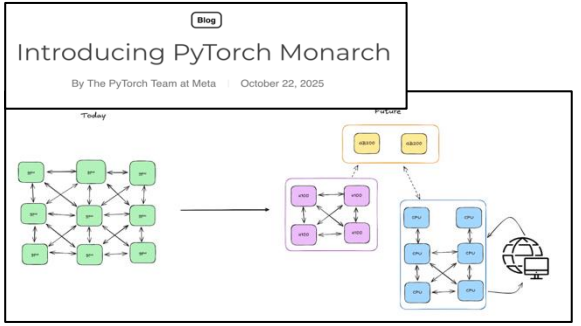

Hitting the Memory Wall: Implications of the Obvious

Wm. A. Wulf
Sally A. McKee
Department of Computer Science
University of Virginia
{wulf | mckee}@virginia.edu

December 1994


DATA MOVEMENT IS ALL YOU NEED: A CASE STUDY ON OPTIMIZING TRANSFORMERS

Andrei Ivanov [*1] Nikoli Dryden [*1] Tal Ben-Nun [1] Shigang Li [1] Torsten Hoefler [1]

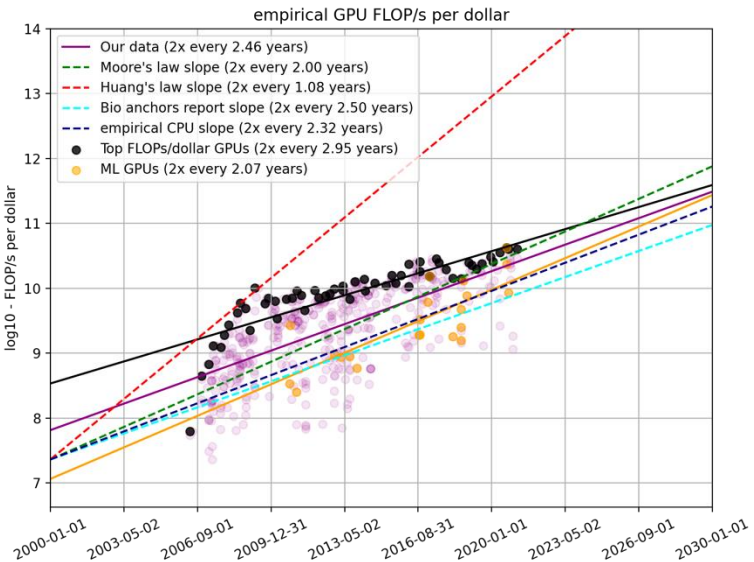


Top: https://dl.acm.org/doi/10.1145/216585.216588
Center: https://arxiv.org/abs/2007.00072
Bottom: https://pytorch.org/blog/introducing-pytorch-monarch/

Processors

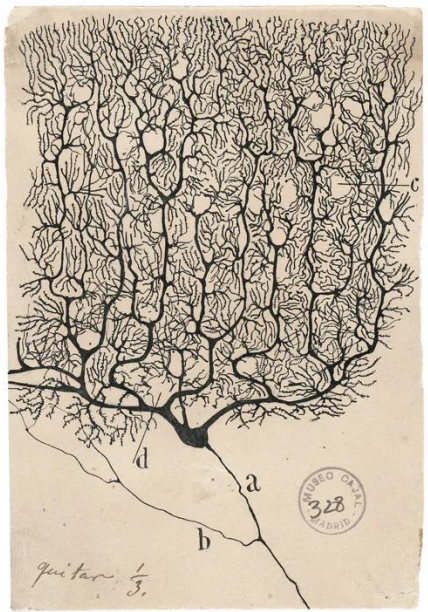
- Moore’s Law: transistor count doubles every 18 months
  - coudn’t keep up the pace, physics become icky at <10nm
  - need to rethink; see data movement
- Heard of Jensen’s Law?

Jensen Huang,
NVIDIA founder



Left: https://en.wikipedia.org/wiki/Jensen_Huang
Right: https://epoch.ai/blog/trends-in-gpu-price-performance/

Revisiting – Limits of computation



https://publicdomainreview.org/collection/illustrations-of-the-nervous-system-golgi-and-cajal/

System	#Units (neurons or transistors)	FLOPs	Power / [W]	data movement multiple
human brain	60-100 x 1e9	[1e15, 1e18]	~20	< 1e2
CPU	20 x 1e9	[1e12, 1e13]	[65, 265]	1e2-1e3
GPU node (8x A100)	1.6 x 1e14	~1e16	[3.2, 6.4] x 1e3	1e2-1e4

KEY TAKEAWAY: We matched the brain's power in terms of FLOPs. We're pretty bad at organizing dataflow ourselves though!

Processors – Takeaways & glossary

Takeaways

- innovation in image rendering very fortunately translates well for general computation
- processing speed develops at twice the speed of data movement speed
- lots of different architectures (micro + macro) but most use: CPU + GPU

Glossary

Unit, computation: FLOPs (floating point operations / cycle (/second))

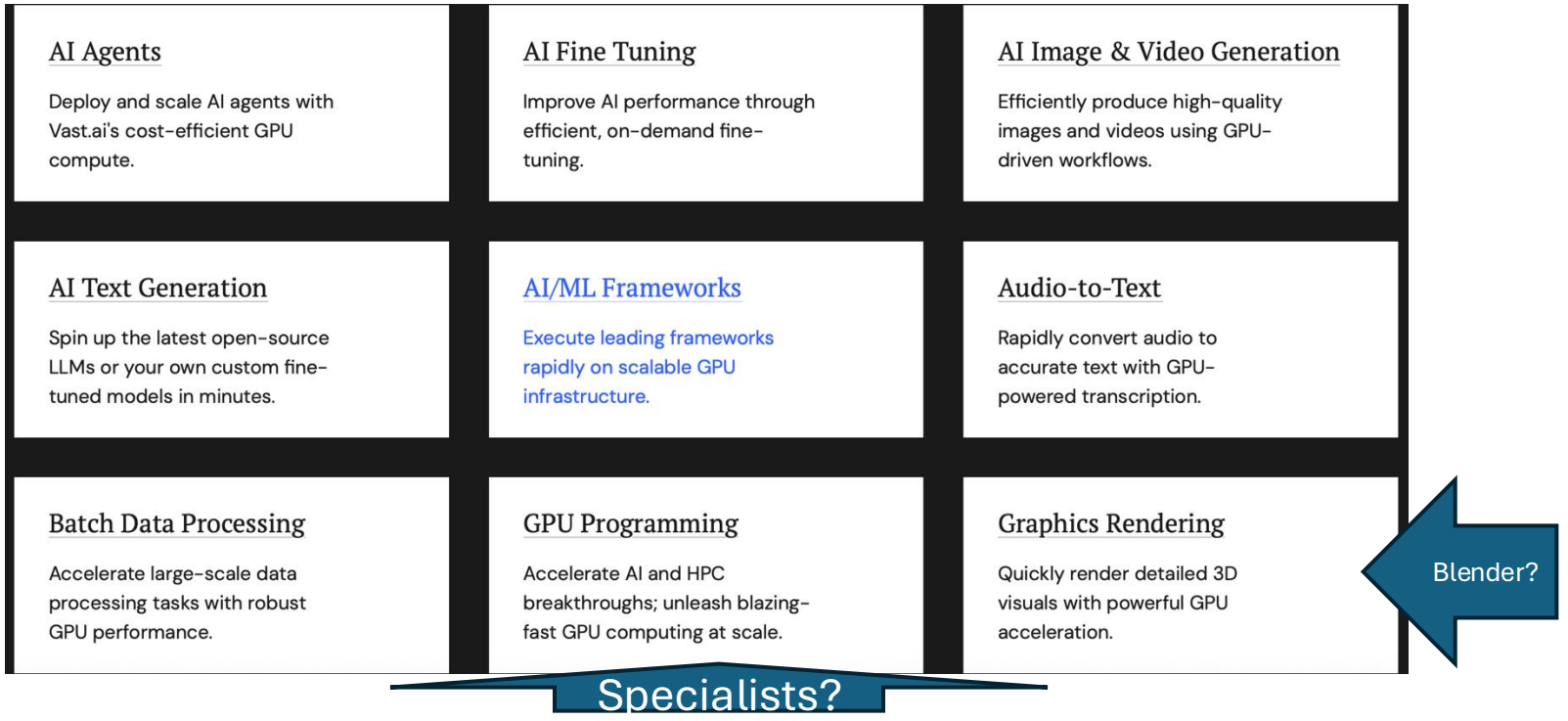
Unit, data movement: B/c (Bytes / cycle or Bytes / second)

Hierarchical parallelism: distributing tasks among machines which themselves expose different parallelism for the tasks at hand

Economics & implications of ML & HPC

- currently: generative AI leads by long margin
- other applications are there, not as much traction
- for our specific HW, more use cases
    - HPC: scientific computing, rendering services
    - different AI workloads, generative and more analytics-heavy

More applications for our infra



https://www.vast.ai/

Economics & implications of ML & HPC

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698

https://www.top500.org/lists/top500/2025/06/

Economics & implications of ML & HPC

Compute Cluster	H100 Equivalents	Owner	Country
xAI Colossus Memphis Phase 1	100000	xAI	U.S.
Tesla Cortex Phase 1	50000	Tesla	U.S.
Lawrence Livermore NL El Capitan Phase 2	44143	U.S. Department of Energy	U.S.
Anonymized Chinese System	30000	N/A	China
Meta GenAI 2024a	24576	Meta AI	U.S.
Meta GenAI 2024b	24576	Meta AI	U.S.

https://www.visualcapitalist.com/the-worlds-most-powerful-ai-supercomputers/

Implications – Political race

AI Factories

AI Factories leverage the supercomputing capacity of the EuroHPC Joint Undertaking to develop trustworthy cutting-edge generative AI models.

Trump's Road to Riyadh: The Geopolitics of AI and Energy Infrastructure

by Guy Laron

Top: https://digital-strategy.ec.europa.eu/en/policies/ai-factories
Bottom: https://americanaffairsjournal.org/2025/08/trumps-road-to-riyadh-the-geopolitics-of-ai-and-energy-infrastructure/

Today's takeaways

1. ML = training + inference. We: The latter.
2. Science went: Drawing pictures -> Sys Eng
3. Core operations: Linear Algebra (LA)
4. Challenge: Data movement
5. We spent trillions to bring data faster to processors.

In the next workshops we'll see

1. What's an LLM? And how do we distribute it?
2. How do we optimize serving an LLM? What are ideal choices for multi-GPU setups for a given model?
3. Advanced topics: What's RAG, LoRA, VLM/LLM? And how do we integrate this at stepping stone?
4. Not mandatory: GPU perf programming, compiling models

Fragen?

Links

- https://www.stepping-stone.ch/
- https://www.stoney-backup.com/
- https://www.stoney-cloud.com/
- https://www.stoney-mail.com/
- https://www.stoney-meet.com/
- https://www.stoney-office.com/
- https://www.stoney-services.com/
- https://www.stoney-storage.com/
- https://www.stoney-wiki.com/

stepping stone AG
Wasserwerkgasse 7
CH-3011 Bern

Telefon: +41 31 332 53 63
www.stepping-stone.ch
info@stepping-stone.ch