

Hardware pre umelú inteligenciu

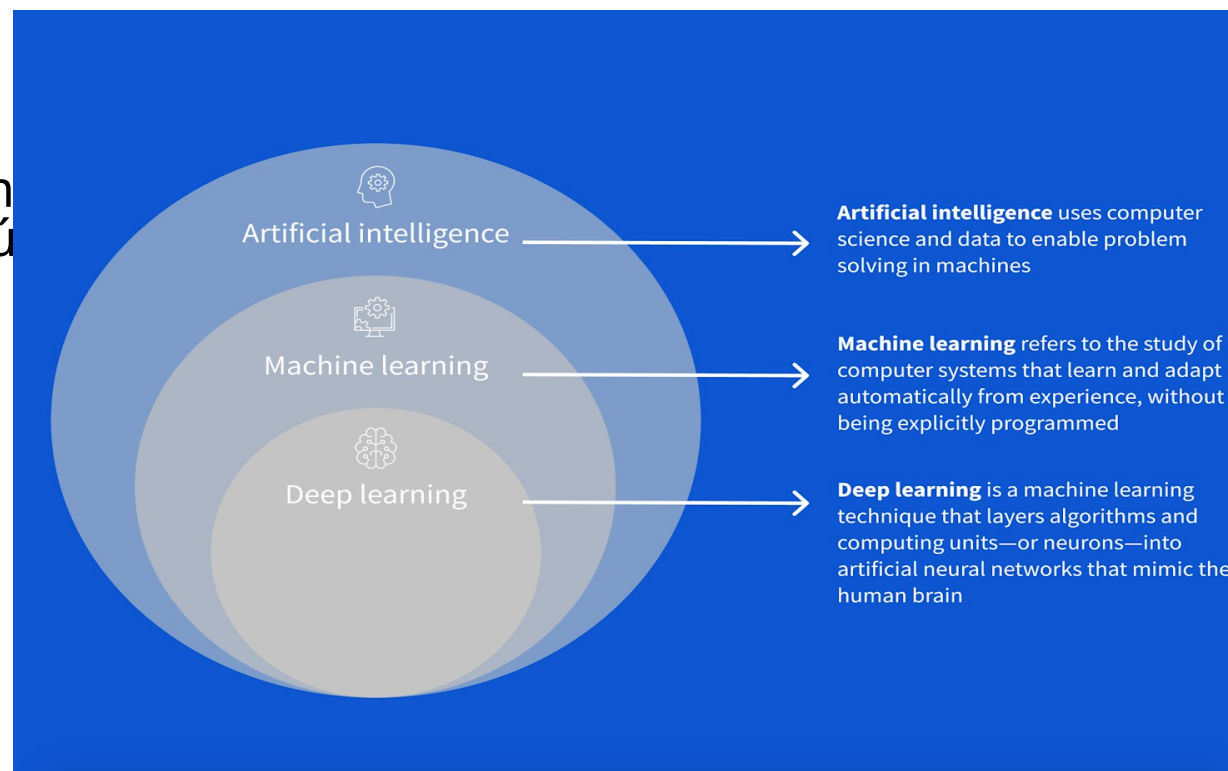
Jozef Janotík, 2.ročník

Čo je to umelá inteligencia?

- Umelá inteligencia je oblasť informatiky, ktorá sa zaoberá vývojom systémov schopných simulovať ľudskú inteligenciu.

- Symbolic AI
- SubSymbolic AI

- automatizácia procesov
- Riešenie ťažko programovateľných úloh



Minimax algoritmus – implementácia umelej inteligencie

- **Daný stav s:**
- **MAX** si vyberá akciu $a \in \text{ACTIONS}(s)$, ktorá produkuje najvyššiu hodnotu $\text{MIN-VALUE}(\text{RESULT}(s,a))$
- **MIN** si vyberá akciu $a \in \text{ACTIONS}(s)$, ktorá produkuje najnižšiu hodnotu $\text{MAX-VALUE}(\text{RESULT}(s,a))$

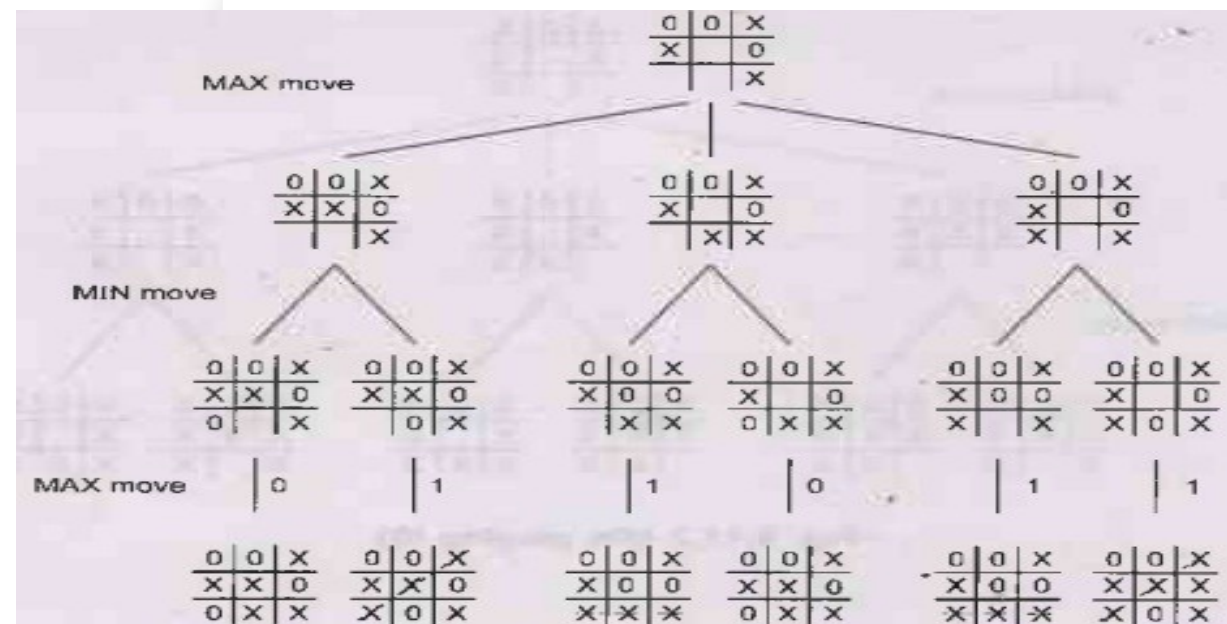
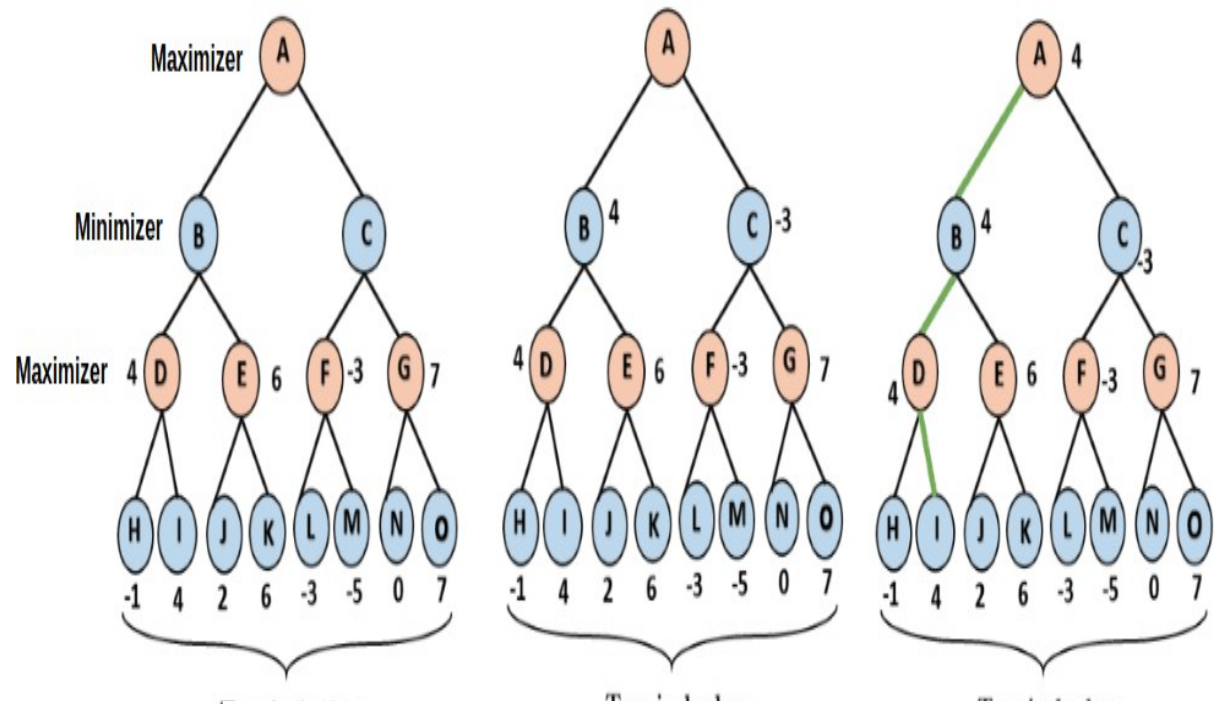
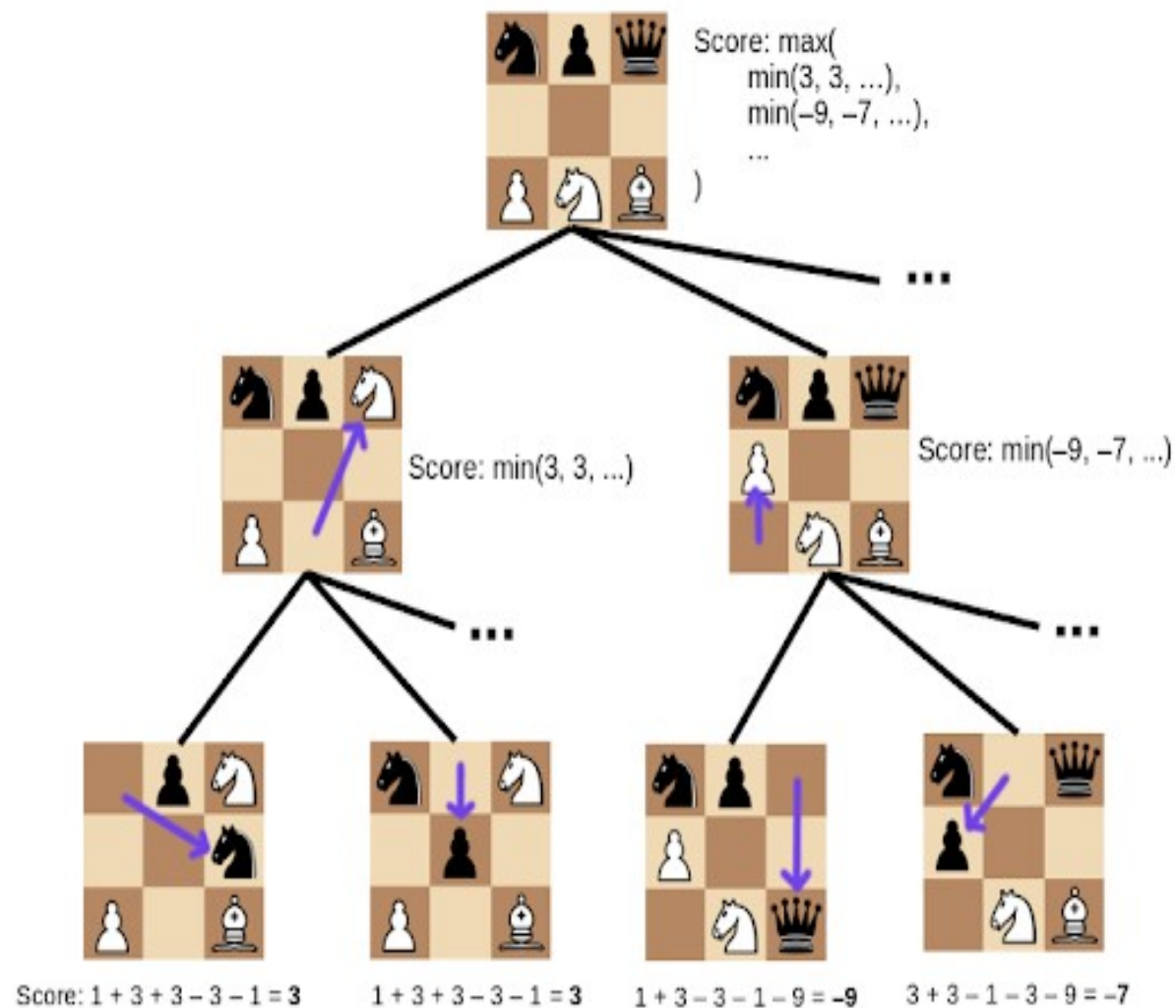


Fig. 5.11.4 Min playing (0)

Minimax

- Minimax prehľadáva strom možných ťahov v hre
- Výpočtová náročnosť závisí od hĺbky prehľadávania
- Nízka hĺbka → stačí CPU (laptop, mobil)
- Vysoká hĺbka → silný CPU, prípadne GPU alebo cloud
- Kombinácia s neurónovou sieťou zvyšuje nároky na výpočty
- Používa sa optimalizácia, aby sa znížil počet prehľadávaných stavov (napr. pruning)



Machine learning

Machine learning (strojové učenie) je oblasť umelej inteligencie, kde sa modely **učia z dát namiesto toho, aby boli pevne naprogramované pravidlami.**

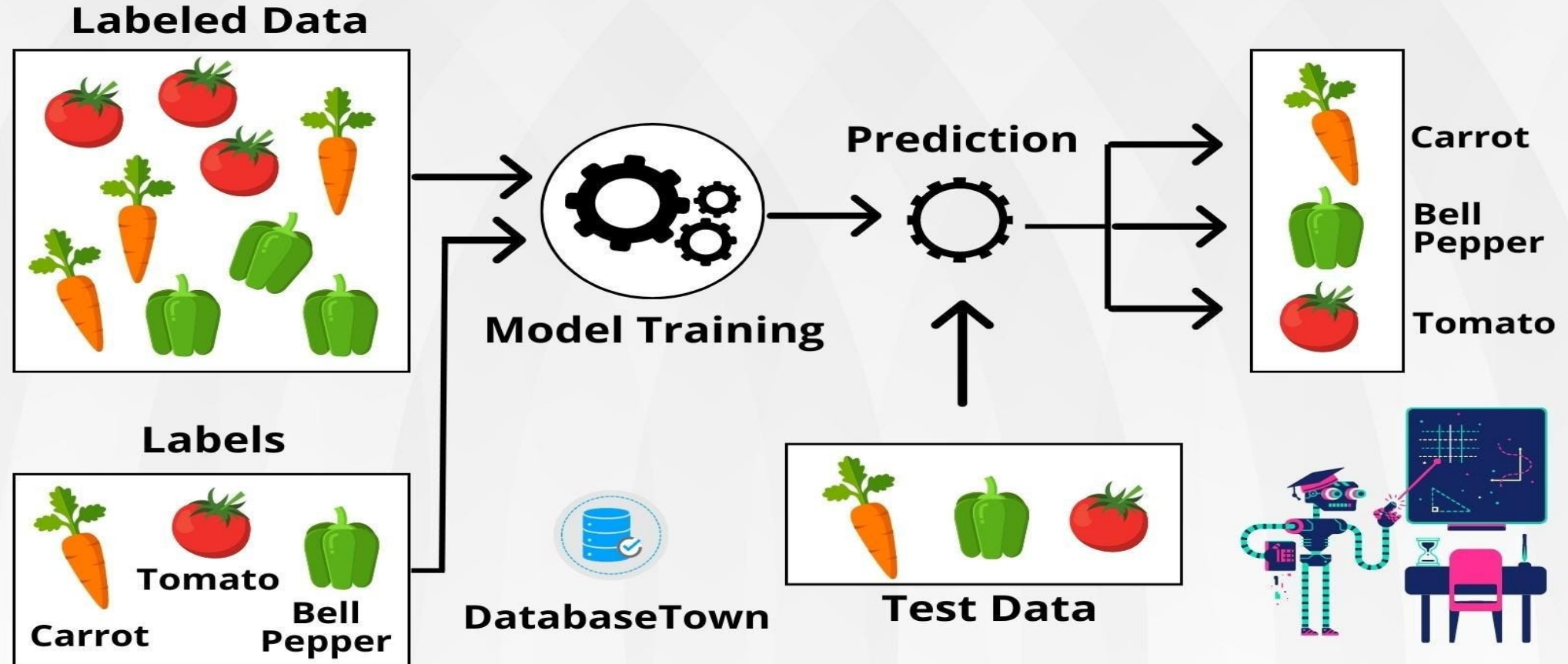
1.Návrh modelu

2.Tréning Modelu

3.Úprava parametrov (rozlišujú sa naučené parametre a hyperparametre)

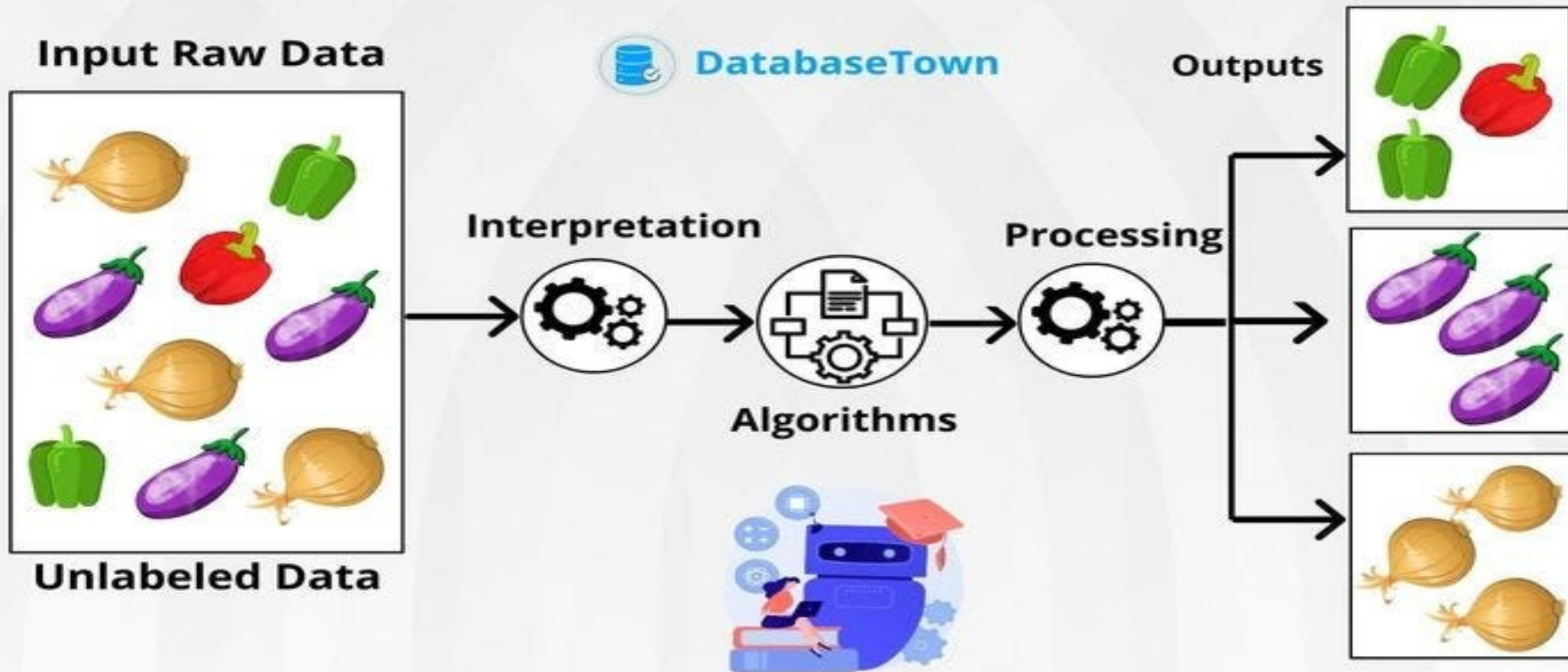
SUPERVISED LEARNING

Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.



UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.



REINFORCEMENT LEARNING

- Učenie na základe trestov a odmiern



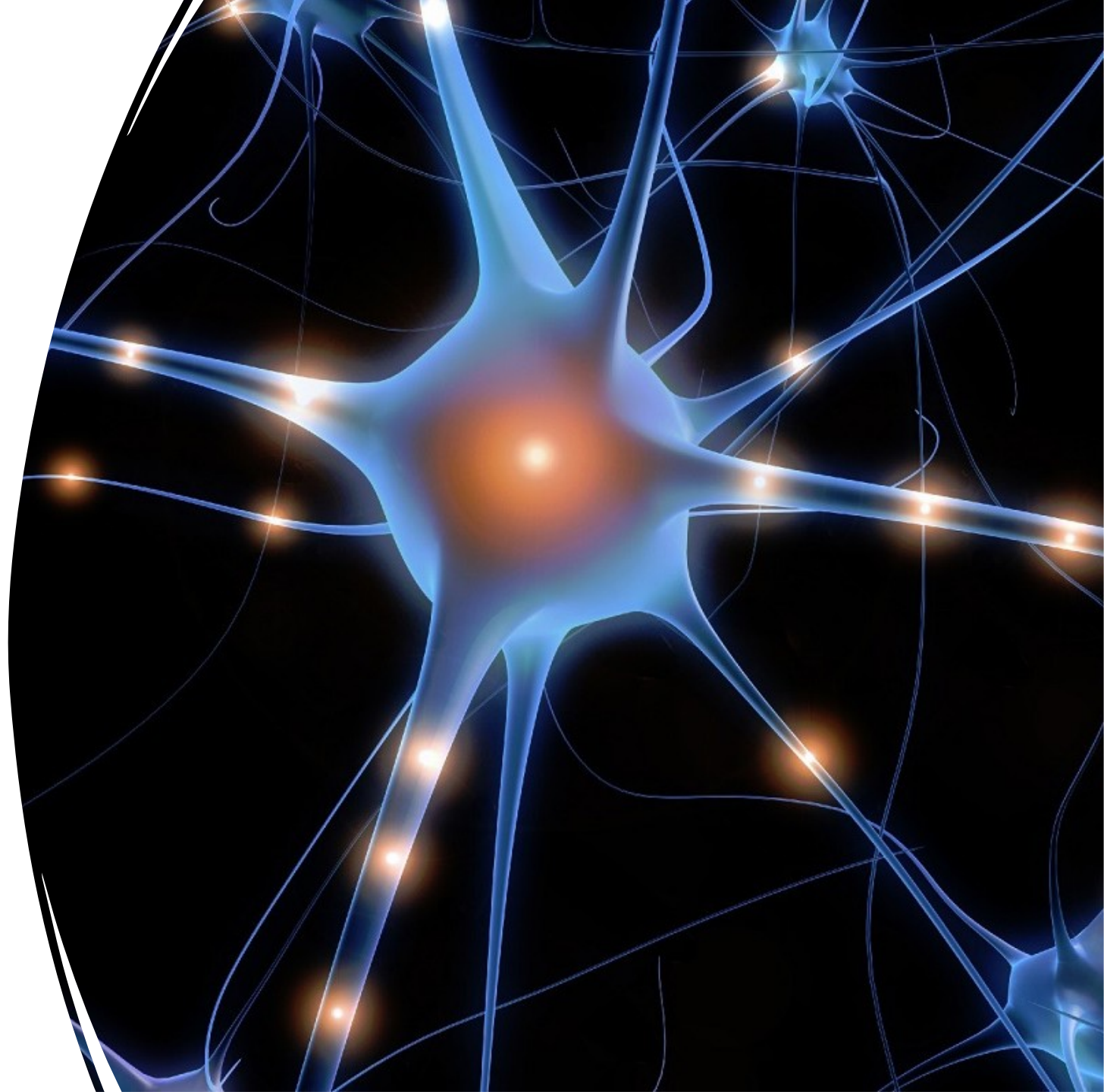
DEEP LEARNING

Neurónové siete sú inšpirované mozgom, ale sú **veľmi zjednodušený matematický model**

Základom siete je **neurón (uzol)**

Tréning prebieha úpravou **parametrov siete (váh a biasov)**

Neuróny sú prepojené **váhami (weights), ktoré určujú silu spojenia**

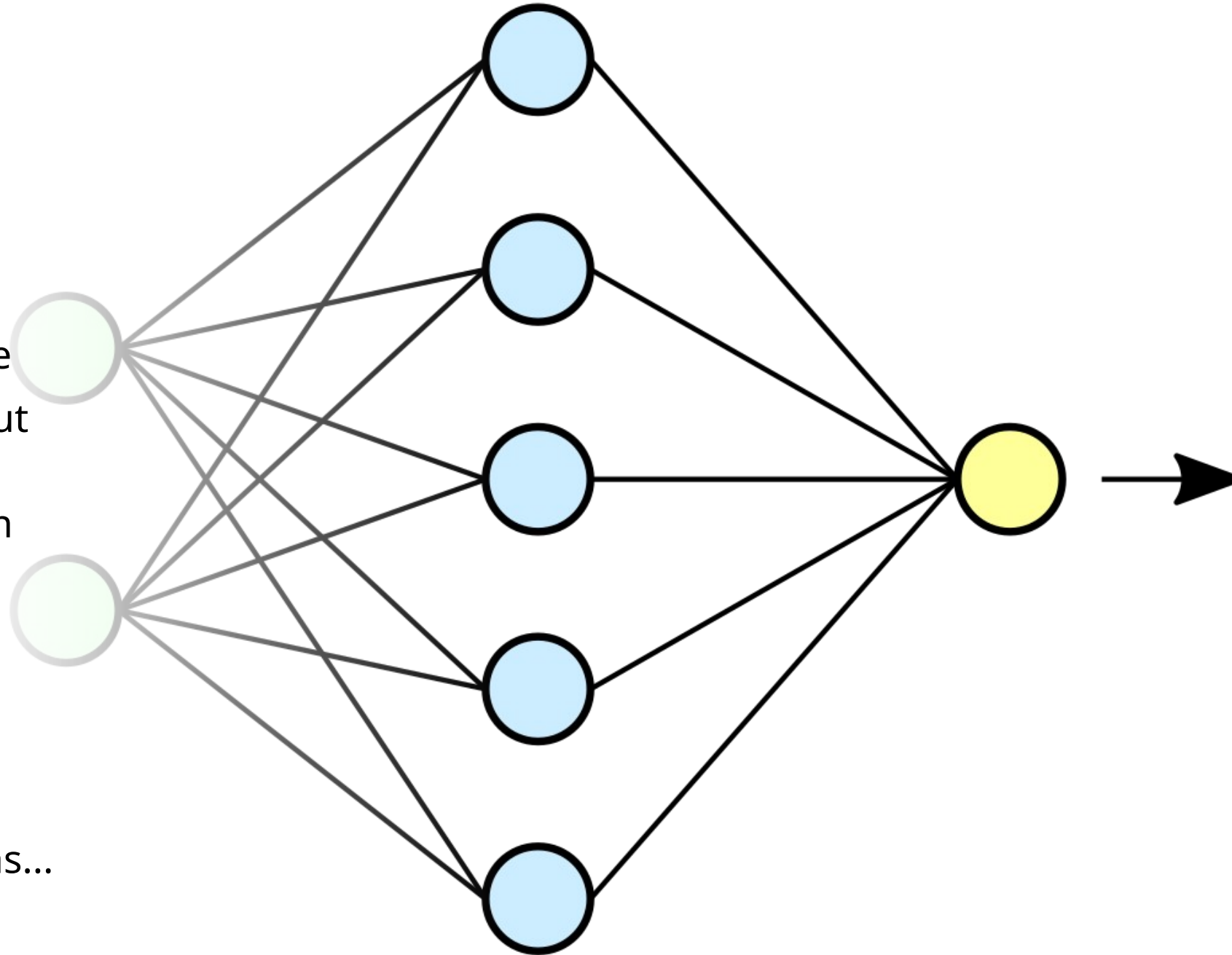


DEEP LEARNING

- Štruktúra neurónovej siete

- Vstupná vrstva (input layer)
- Skryté vrstvy (hidden layers)
- Výstupná vrstva (output layer)

- Softvéry knižnice:
Tensorflow, Pytorch, Keras...



AI čipy

- GPU
 - TPU
 - NPU
 - ASIC AI ČIPY
 - FPGAs
-
- Všetky tieto čipy sú postavené z tranzistorov, ktoré tvoria základné elektronické obvody.
 - Rozdiel medzi nimi je v tom, že sú optimalizované na rôzne typy výpočtov.



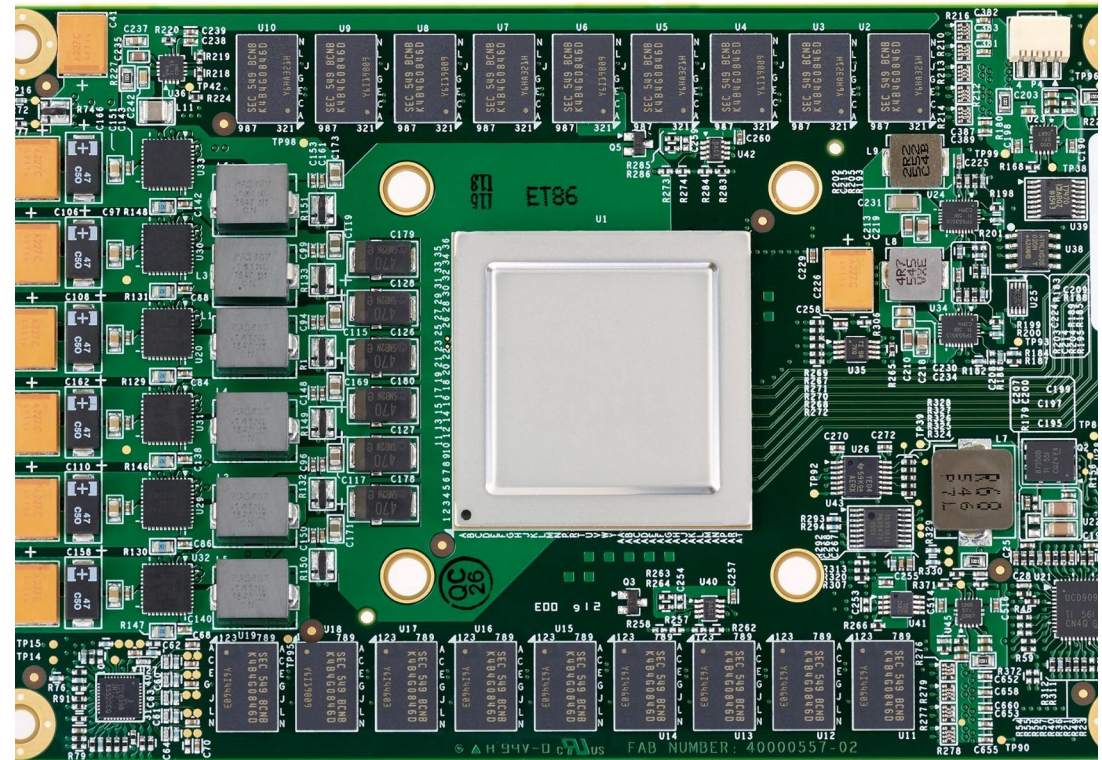
GPU

- Univerzálny paralelný chip
- Vysoká flexibilita
- Nižšia efektívnosť
- SIMD jadrá architektúra
- Najlepšia podpora pre PyTorch
- Funguje všade – cloud, lokálne servery, spotrebiteľské zariadenia



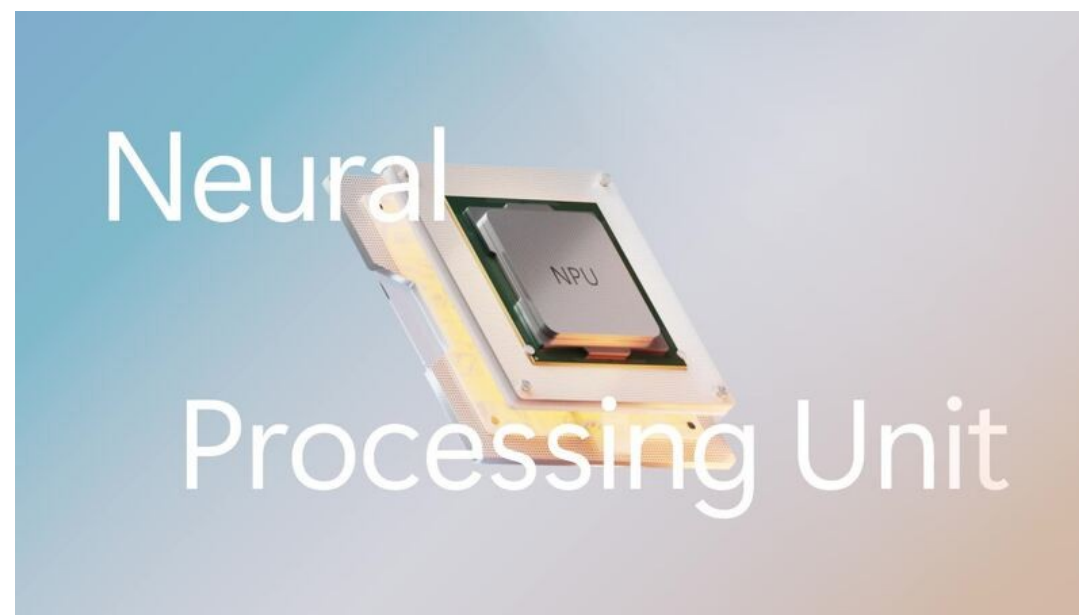
TPU

- Účelovo navrhnutý AI akcelerátor pre výpočty s maticami a tenzormi
Optimalizovaný pre vysokovýkonné úlohy hlbokého učenia
- Najlepší výkon dosahuje s TensorFlow/JAX + XLA
- architektúra systolických polí



NPU

- NPU = špecializovaný AI čip určený hlavne pre "edge AI"
- Extrémne nízka spotreba energie
- Vyššia bezpečnosť



ASIC AI ČIPY

- navrhnuté na jednu konkrétnu úlohu
- nedajú sa po výrobe preprogramovať (fixný hardvér)
- ASIC AI čipy (Application-Specific Integrated Circuit)
- Nízka latencia
- Energetická efektívnosť



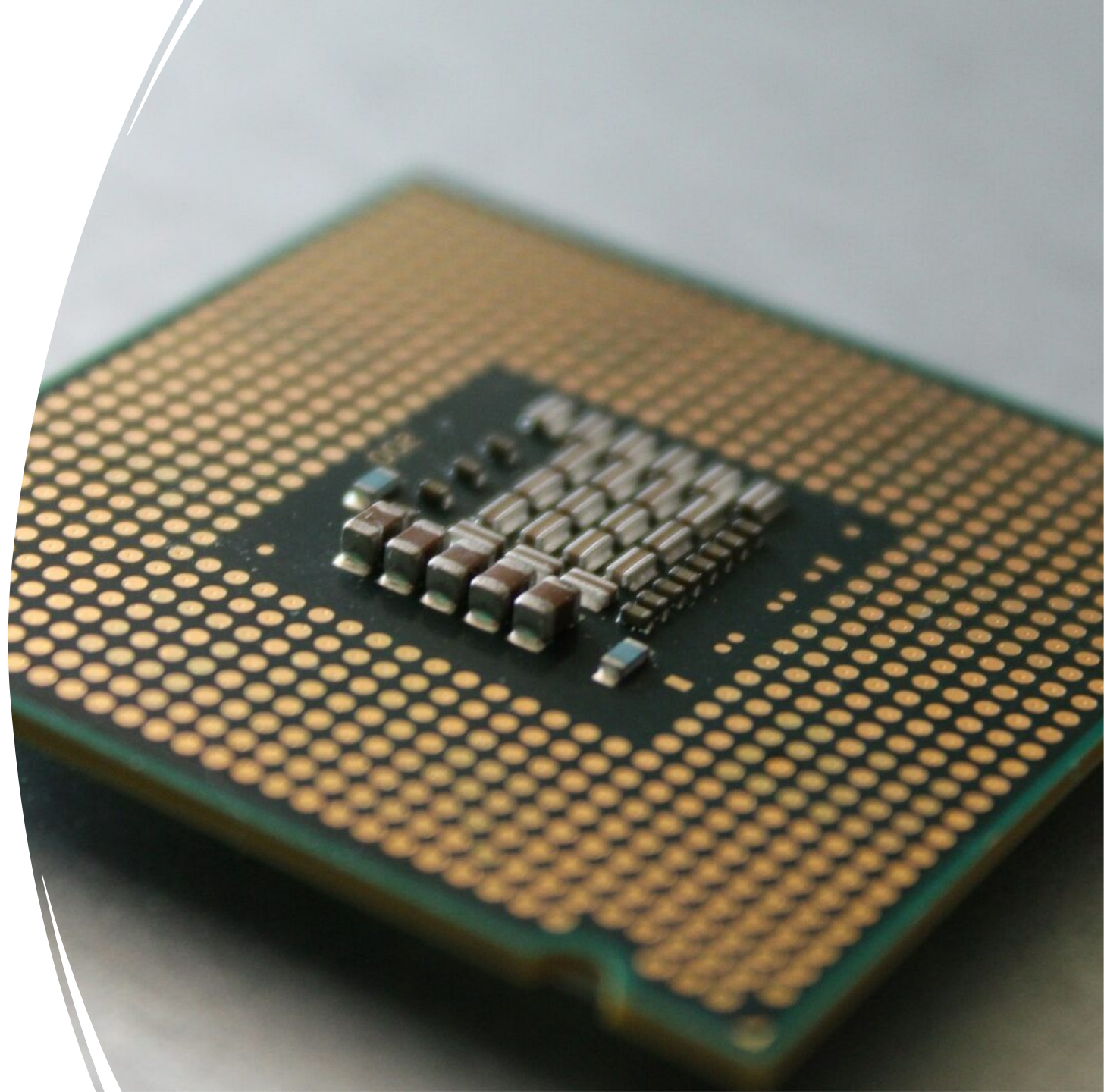
FPGIs (Field-Programmable Gate Array)

- FPGA sú programovateľné hardvérové čipy, ktoré si vieš po výrobe nakonfigurovať podľa potreby.
- Na rozdiel od ASIC čipov, ktoré sú pevne navrhnuté a už sa nedajú meniť, FPGA sa dajú opakovane preprogramovať.
- Používajú sa hlavne na:
- testovanie a vývoj nových čipov
- špeciálne výpočty
- prototypovanie AI hardvéru



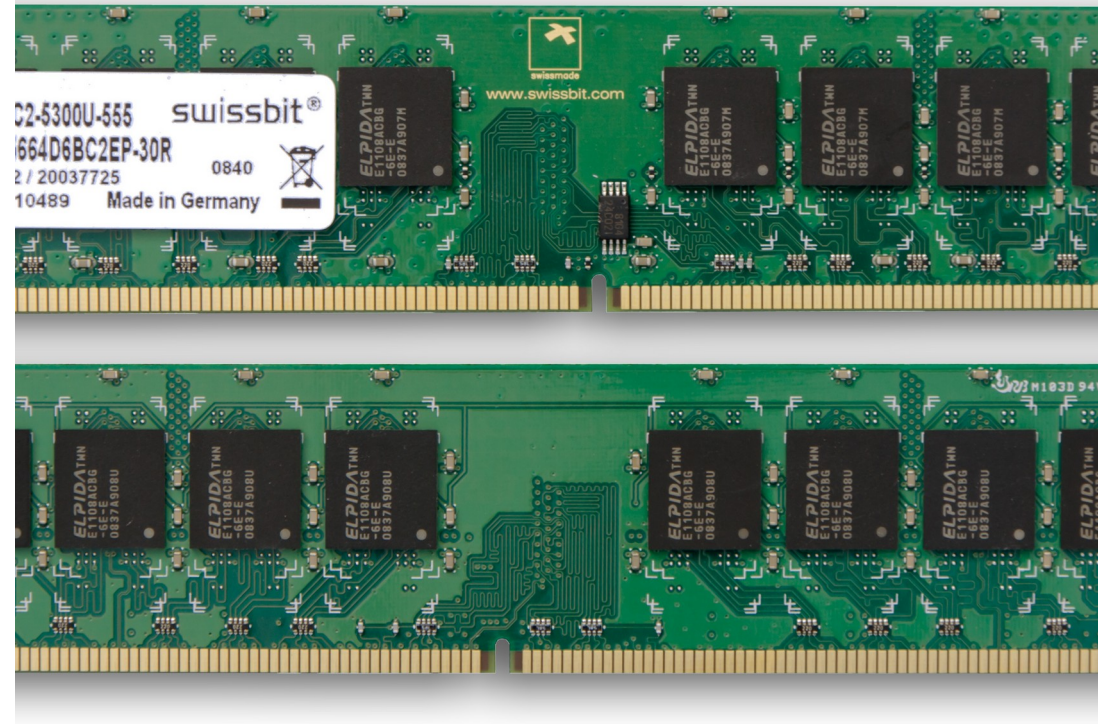
CPU

- „mozog počítača“
- základná súčasť každej počítačovej infraštruktúry
- vykonáva hlavné riadiace a výpočtové úlohy
- koordinuje ostatné komponenty vrátane GPU



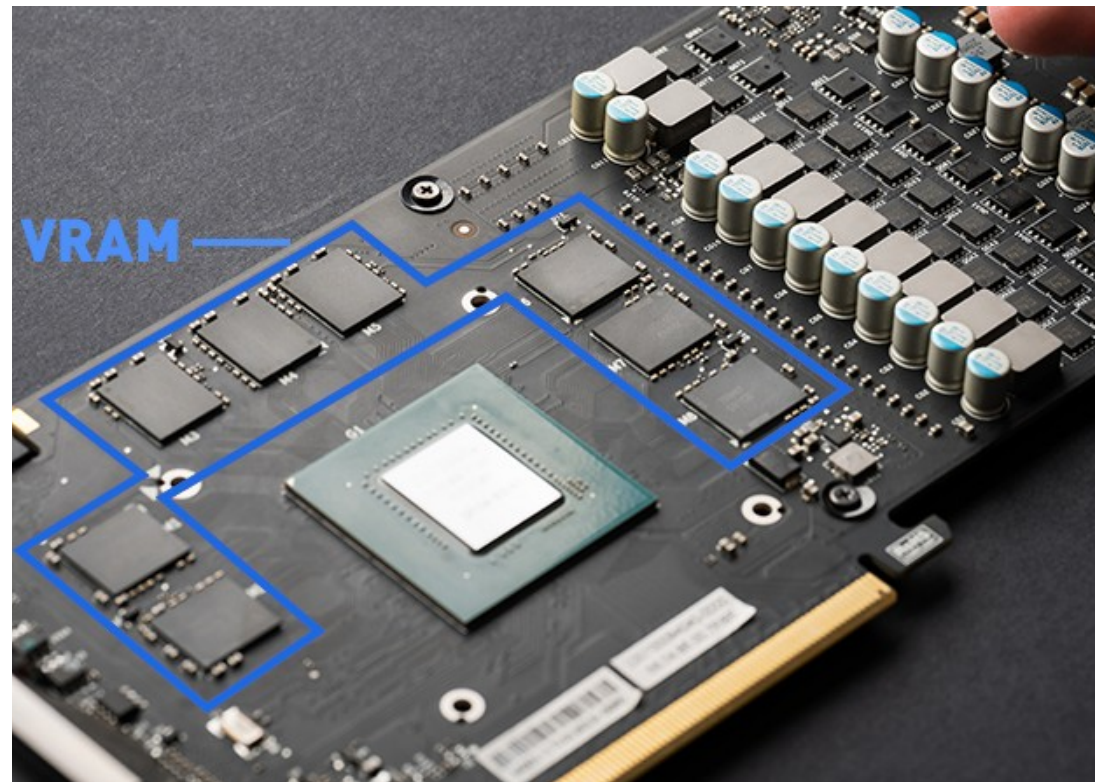
RAM

- hlavný typ pamäte pre AI aj bežné počítače
- rýchle, dočasné ukladanie dát



VRAM (Video Random Access Memory)

- špecializovaná pamäť používaná v GPU
- určená pre grafiku a AI výpočty
- podporuje vysoký stupeň paralelného spracovania
- vhodná pre náročné AI úlohy
- drahšia než RAM
- má menšiu kapacitu než bežná RAM



HBM (High Bandwidth Memory)

- pôvodne vyvinutá pre vysokovýkonné výpočty (HPC)
- navrhnutá na veľmi rýchly prenos dát medzi výpočtovými jednotkami
- veľmi dôležitá pre AI aplikácie
- extrémne vysoká rýchlosť
- veľmi drahá
- ideálna pre náročné AI systémy



Nevolatilná pamäť

- nevolatilná pamäť uchováva dáta aj bez napájania (napr. SSD, HDD)
- v AI systémoch slúži na dlhodobé uchovávanie dát a modelov
- nie je vhodná na aktívne výpočty a prenos dát

SSD vs HDD

SSD



Solid State Drive

HDD



Hard Disk Drive

NVIDIA DGX Spark 3.0

- **Špecifikácie**
- 128 GB zdieľanej pamäte
- Veľmi svižný disk SSD poslúži kapacitou 4000 GB
- podpora modelov do približne 100 miliárd parametrov
- CUDA-X a NVIDIA AI softvér v základe



AI Clouding

- AI beží v cloude (ChatGPT, Gemini, Claude na GPU/TPU clustroch)
- Špecializovaná infraštruktúra (GPU, rýchla pamäť, siete, orchestration)
- Škálovanie (tisíce GPU podľa potreby – elastic compute)
- Nevýhoda: vyššia latencia
- Výhoda: vysoký výkon bez vlastného hardvéru



Minulost' vs budúcnosť

- Nedostatok dát, výpočtová technika veľmi drahá
- 1990–1995 vznik world wide web
- 2000–2010 prudký pokles ceny hardvéru
- lacnejší výpočtový výkon(Moore's law)
- lacné dátové úložiská (Kryder's law)
- masívne investície do AI a výskumu (Google, Microsoft, Amazon, NVIDIA)

Zoznam použitej literatúry(ISO 690)

- DIVE INTO DEEP LEARNING. *Dive into Deep Learning* [online]. Dostupné na: <https://d2l.ai/> [cit. 2026-04-09]
- IBM. *AI Hardware* [online]. IBM Think. Dostupné na: <https://www.ibm.com/think/topics/ai-hardware> [cit. 2026-04-09]
- IBM. *AI Hardware* [online]. IBM Research. Dostupné na: <https://research.ibm.com/topics/ai-hardware> [cit. 2026-04-09]
- INTEL. *Artificial Intelligence (AI) Hardware* [online]. Dostupné na: <https://www.intel.com/content/www/us/en/learn/ai-hardware.html> [cit. 2026-04-09]
- HARVARD UNIVERSITY. *Prednášky z predmetu: Artificial Intelligence (lecture notes)* [online]. 2024. [cit. 2026-04-09]

- Ďakujem za pozornosť!