

Mašīnmācīšanās un superdatori

Dr. phys. Rūdolfs Livanovičs

EuroCC HPC kompetences centrs



Kas ir mašīnmācīšanās (machine learning, ML)?

- **Mērķis:** datu apstrādes sistēmas, kas apgūst likumsakarības no datiem, samazinot nepieciešamību pēc tiešas programmēšanas konkrētu problēmu līmenī.
- **Dažādas apmācības pieejas:**
 - *Supervised learning*: speciāli marķēti dati tiek izmantoti lai virzītu apmācības procesu konkrētiem pielietojumiem (klasifikācija, regresija, utt.).
 - *Unsupervised learning*: algoritmi pēc vispārīgiem principiem atrod konkrētas likumsakarības nemarķētos datos (klasterizācija, dimensiju redukcija, utt.).
 - *Reinforcement learning*: mācīšanos nodrošina stohastisku mēģinājumu korekcijas, modelis tiek "atalgots" par pareiziem lēmumiem un mācās no kļūdām.
- Reālas sistēmas bieži izmanto kombināciju no šīm pieejām.

Vai mašīnmācīšanās ir "mākslīgais intelekts"?

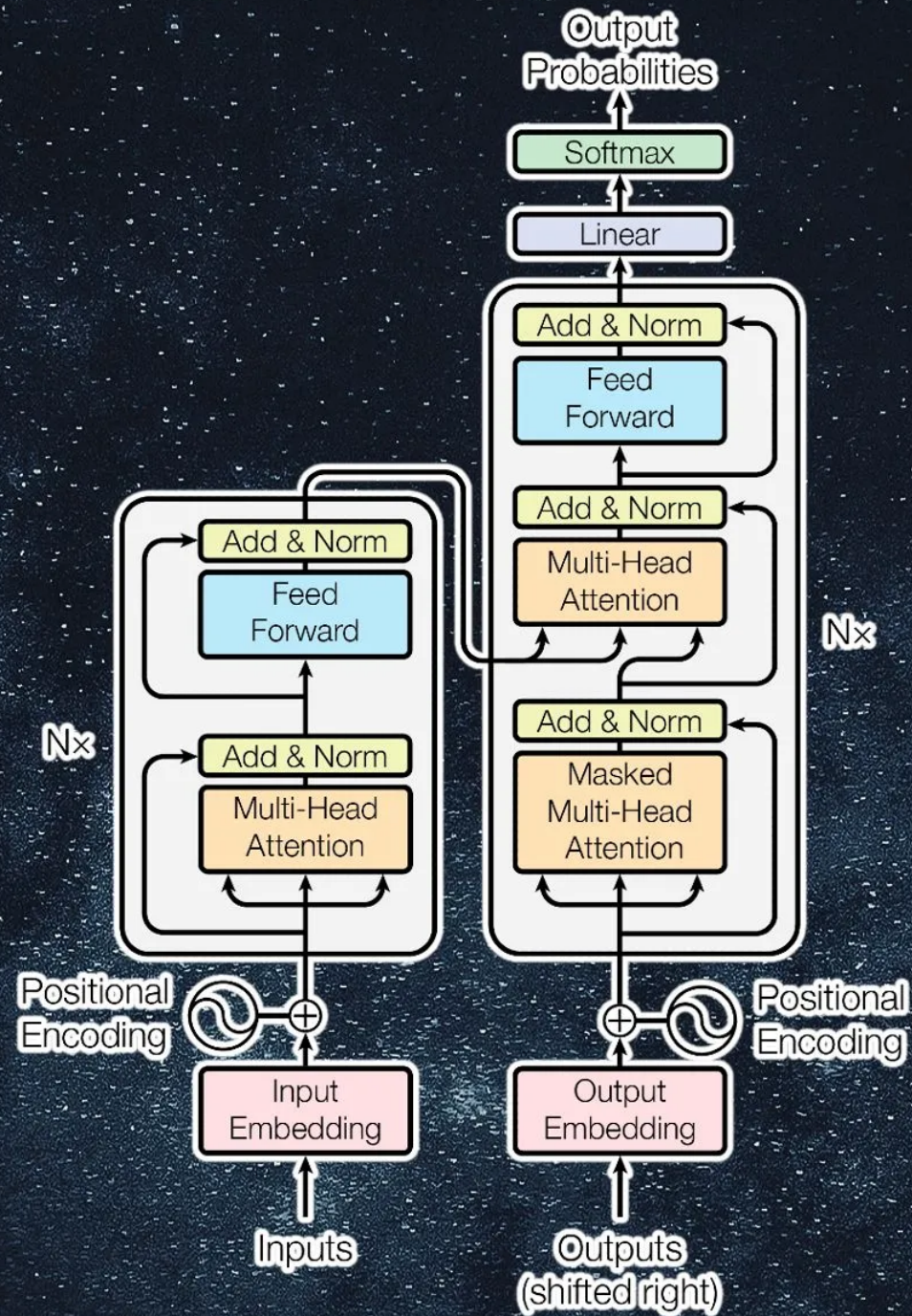
- **Mākslīgais intelekts (artificial intelligence)** fundamentāli nozīmē divas dažādas lietas:
 - Sistēmas, kas izpilda uzdevumus, kuriem tradicionāli ir nepieciešams cilvēka intelekts. No šī neizriet, ka pašas sistēmas ir "inteliģentas" - spēlēt šahu tikpat labi var arī deterministiski izrēķinot visus iespējamus gājienus.
 - **Artificial general intelligence (AGI)**: sistēmas, kas ir "inteliģentas" plašākā nozīmē, bez piesaistes konkrētiem uzdevumiem.
- Nav vispārpieņemtas definīcijas tam, ko nozīmē vai kā identificēt AGI.
- Nav atzīta tehniska pamata uzskatīt, ka esošās mašīnmācīšanās sistēmas ir daļa no vai ceļš uz AGI.

Mašīnmācīšanās vēsture

- **1950.-1960. gadi:** pirmā dambretes programma (*Arthur Samuel*), maksliģo neironu tīklu (*artificial neural networks*) jēdziens, perceptrons (*Frank Rosenblatt*).
- **1980. gadi:** multislāņu perceptroni un efektīvas apmācības metodes (backpropagation, SGD), pirmie text-to-speech un rokraksta atpazīšanas modeļi.
- **1990. gadi:** SVM (*support vector machines*) un RNN (*recurrent neural networks*) modeļi, LSTM (RNN ar atmiņu), IBM Deep Blue uzvar Kasparovu šahā.
- **2000. gadi:** CNN (*convolutional neural network*) ļauj izveidot praktiskus attēlu atpazīšanas un klasifikācijas modeļus. Pirmie *deep learning* piemēri.
- **2010. gadi:** praktiski *deep learning* piemēri pateicoties arvien pieaugošajām datorjaudām un GPU izmantošanai. ImageNet, ResNet, GAN un difūzijas modeļi. Transfomeru arhitektūra, LLM un GPT.

Mašīnmācīšanās mūsdienās

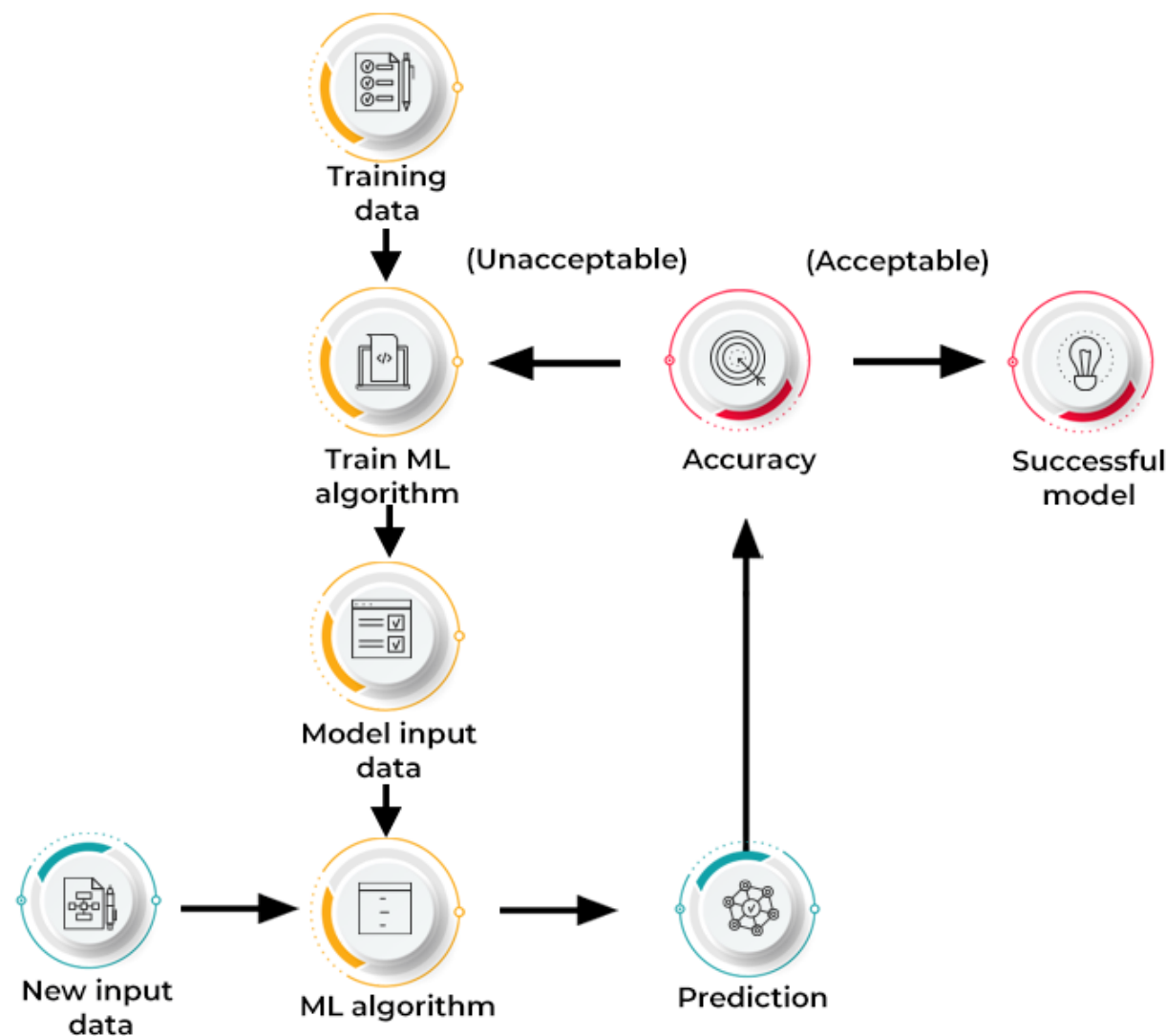
- Vairums praktisko pielietojumu izmanto transformer arhitektūru - LLM (GPT, Claude, DeepSeek, utt.), TTS/STT, laika sēriju un sekvenču analīzes modeļi, u.c.
- Izņēmums ir attēlu atpazīšanas un klasifikācijas modeļi, kur joprojām dominē CNN arhitektūras (YOLO, EfficientNet, u.c.)
- Sekojot datortehnikas progresam, modeļi ir dramatiski lielāki.



Kas jāzina pirms mēģināt lietot ML?

- Pietiekoša apjoma un kvalitātes datiem ir fundamentāla nozīme.
- Ne mazāk svarīga ir spēja sistemātiski validēt rezultātu.
- ML pievienotā vērtība ir iespēja nerakstīt konkrētu algoritmu specifiskai problēmai – nevajag to lietot tur, kur šādi algoritmi jau eksistē.

HOW DOES MACHINE LEARNING WORK?

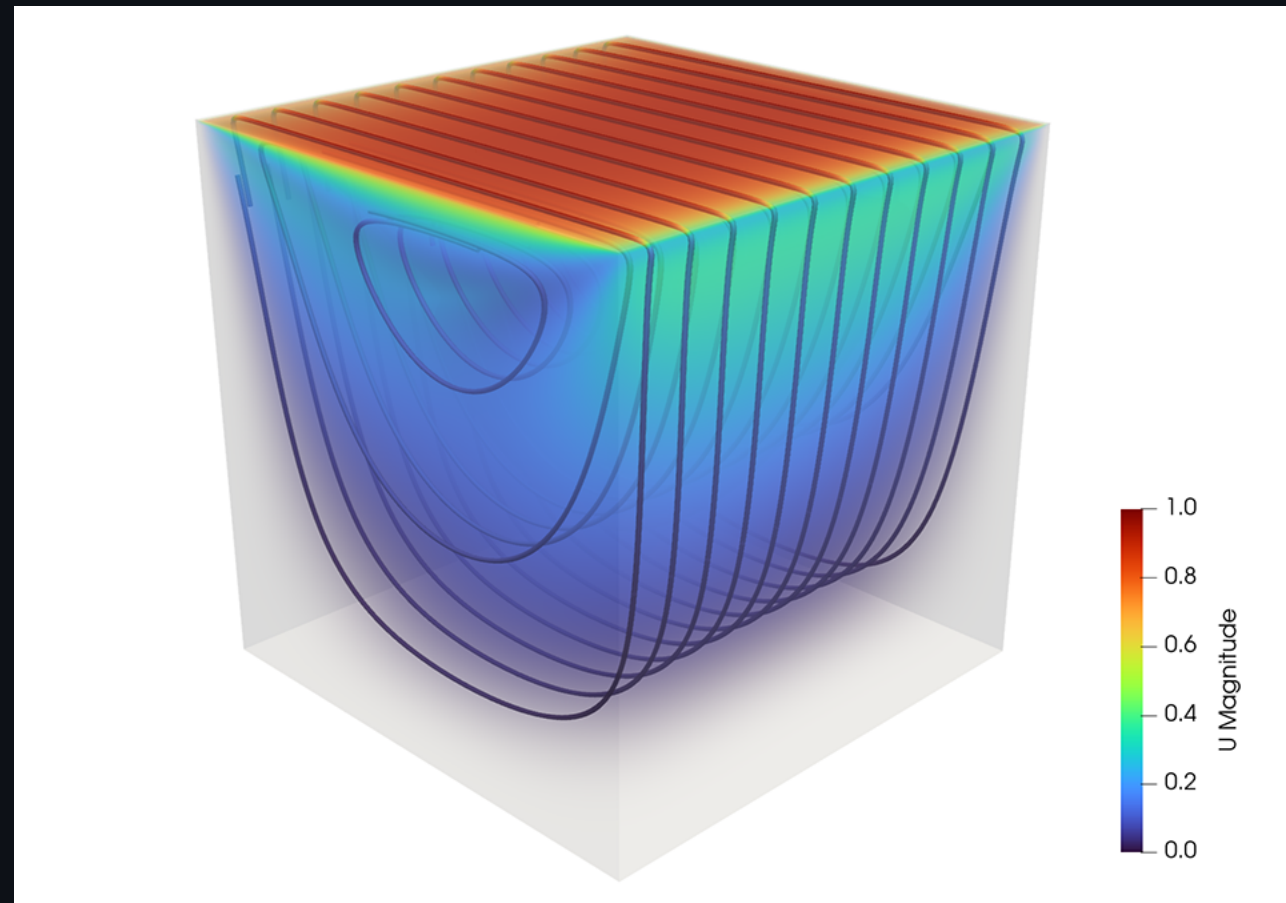


Kas ir HPC?

- **High Performance Computing:** augstas veiktspējas datortehnika + attiecīgi optimizēta un paralelizēta programmatūra.
- **Pielietojums:** skaitļošanas problēmām, kas pārsniedz personālo datoru resursus un ir sadalāmas nosacīti neatkarīgās daļās.
- **HPC vs. cloud:** HPC arhitektūras tipiski ietver ātrus tīkla savienojumus un specializētus risinājumus augstas veiktspējas pielietojumiem.
- Lielākie HPC klasteri ir pēc jaudas daudzkārt pārsniedz publiski pieejamus mākoņskaitļošanas pakalpojumus.

Tipiskie HPC pielietojumi

- **Matemātiskā modelēšana:** dabas un inženiertehnisko procesu modeļi zinātniskie un industriāliem mērķiem.
- **Lielo datu analīze (big data):** biznesa procesu un patērētāju datu analīze, utt.
- **Procesu vadība:** tranzakciju apstrāde maksājumu sistēmās, infrastruktūras sensoru datu analīze, u.c.



Kāpēc HPC ML kontekstā?

- **Trenēt modeļus ir dārgi:** mūsdienu ML arhitektūras un pielietojumi sasniedz panākumus primāri pateicoties milzīgam datu apjomam apmācības procesā.
- **Izmēram ir nozīme:** lielāki modeļi bieži ir fundamentāli labāki pat pie vienādas arhitektūras, treniņdatiem un validācijas.
- **Apmācība prasa daudz vairāk resursus nekā izmantošana:** tipisku ML modeļu inference ir salīdzinoši lēta, kamēr pirmreizējās apmācības process bieži prasa HPC līmeņa resursus.
- Tipiski ML modeļu apmācība mūsdienās notiek uz GPU-bāzētiem HPC klasteriem.

Cik dārgi ir uztrenēt modeli?

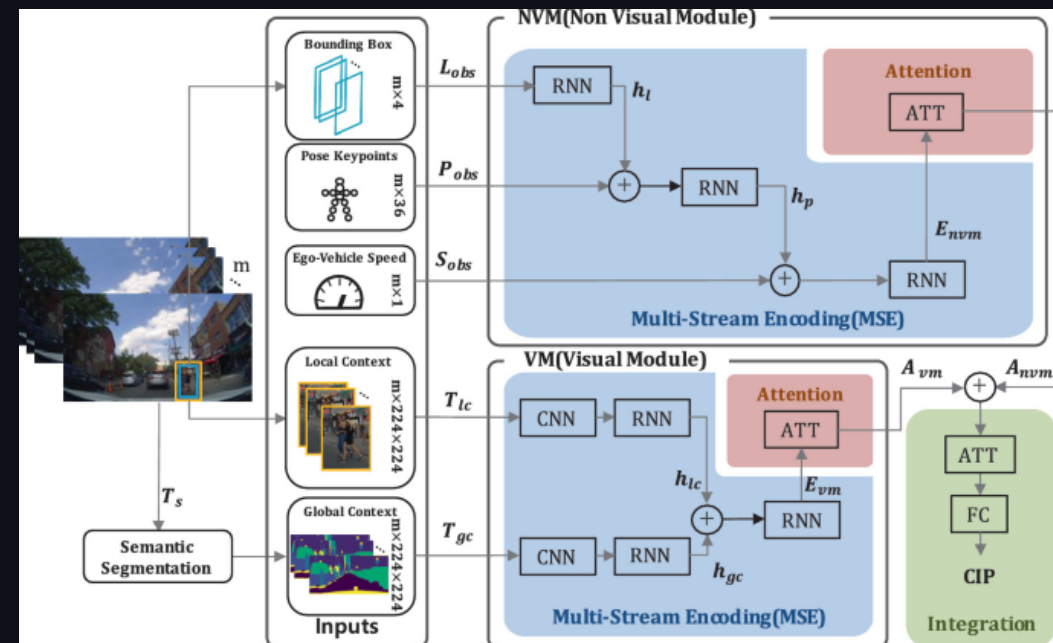
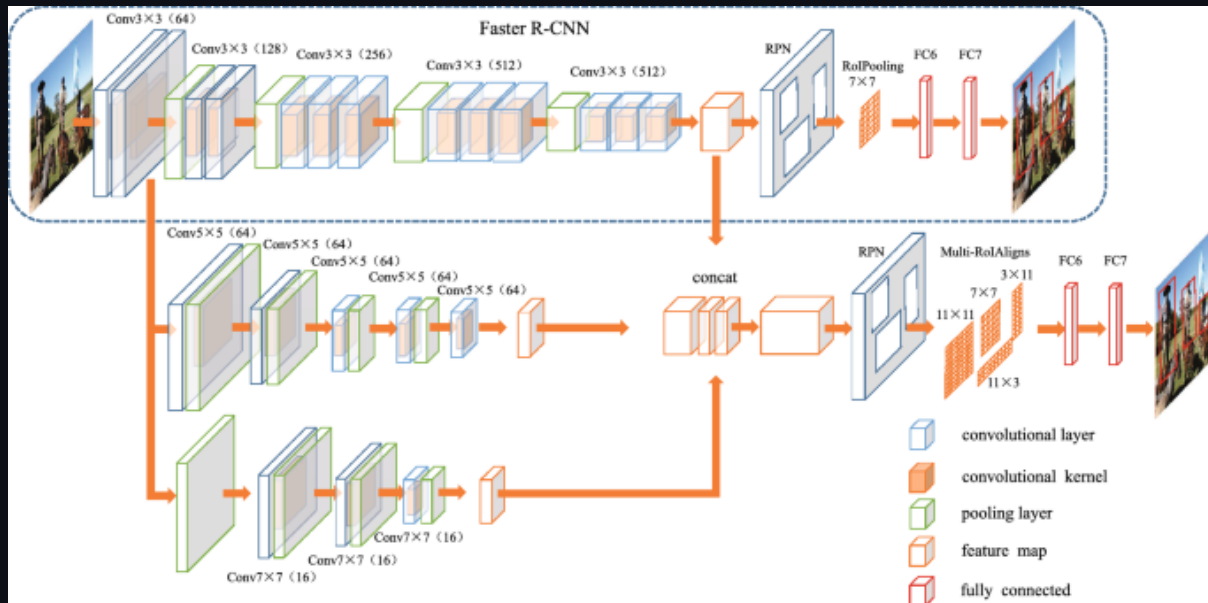
- **Nelieli attēlu klasifikatori:** 0.1-10 GPU stundas (1-100\$).
- **Vidēja izmēra CNN (BERT-base 110M):** 100-1000 GPU stundas (5-20K\$).
- **Liela mēroga LLM (GPT-3 175B):** 1-10M GPU stundas (5-20M\$).
- **GPT-4+ klases modeļi:** 10K+ GPU izmēra klasteri, kas strādā mēnešiem (100M\$+)

Reālistiski, tikai mazus un ļoti specializētus modeļus ir jēga trenēt no nulles.

Vairumā pielietojumu racionālāk ir pieskaņot publiski pieejamus modeļus – fine tuning ar QLoRa, adapteri, embedding vektori, prompt engineering, utt.

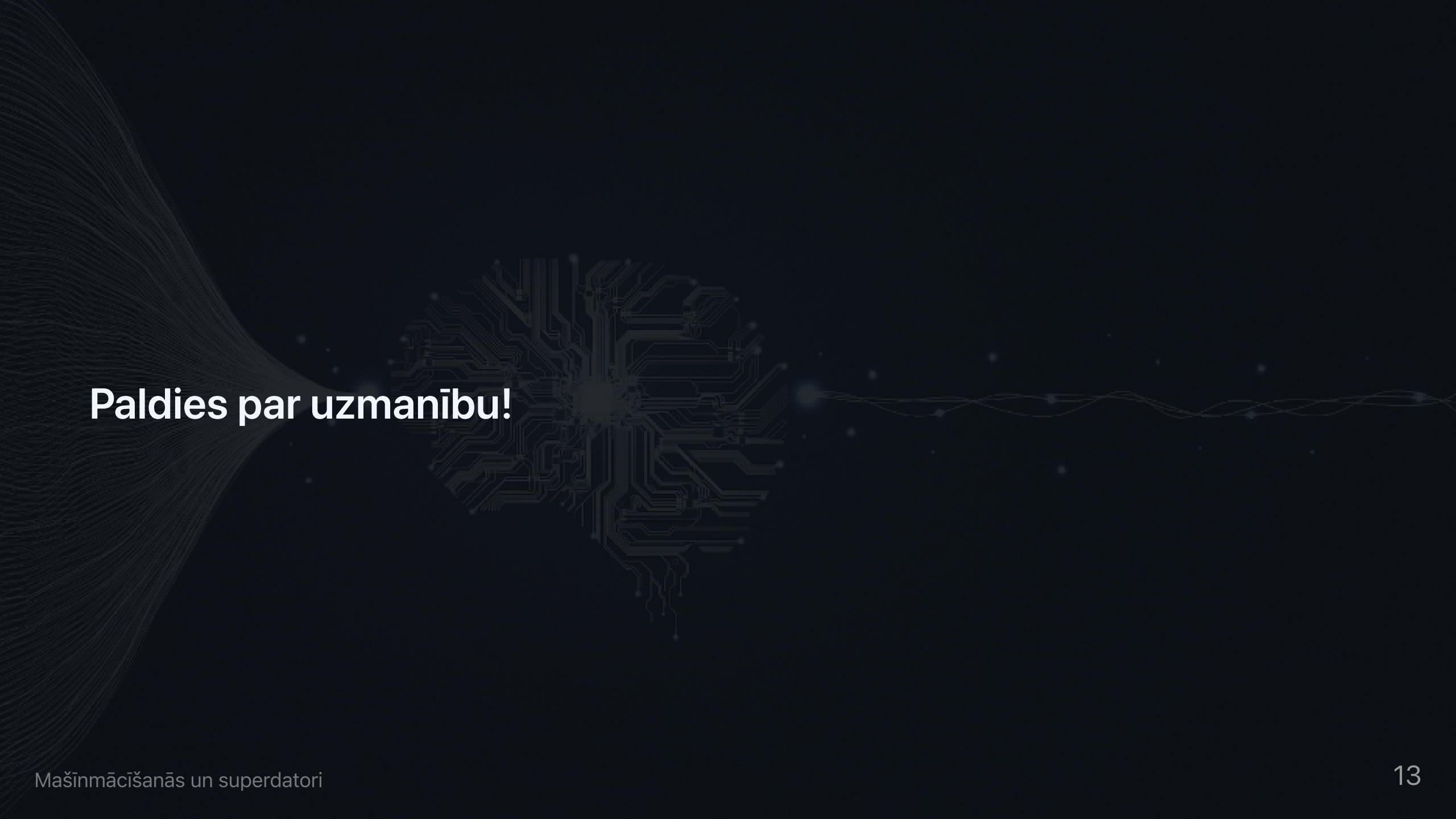
EuroCC demo case

- Lai ilustrētu HPC praktisku izmantošanu modeļu pielāgošanā un tālākā lietošanā, EuroCC ietvaros veidojam ML demo case.
- Gājēju detektors: balstoties uz ieejas video datiem, atpazīt gājēju atrašanās vietas un paredzēt to tālāko uzvedību.
- Par pamatu ņemam Faster-RCNN vision modeli priekš detekcijas.
- Tālākai klasifikācijai RNN vai LSTM modeļi (vēl testējam).



Kopsavilkums

- ML/AI ir spēcīgs instruments datu apstrādei un analīzei, taču tā nav panaceja.
- Lai produktīvi izmantotu mašīnmācīšanās metodes, vispirms ir jābūt skaidrai izpratnei par vēlamu rezultātu un jāiegūst pietiekoša apjoma un kvalitātes dati.
- HPC var nodrošināt vajadzīgo skaitļošanas jaudu modeļu apmācībai un pielāgošanai.
- Tradicionālā mākoņskaitļošana (*cloud computing*) tipiski ir ērtāka izvēle ikdienas inferencei modeļu tālākā izmantošanā.

The background is dark with abstract, glowing circuit-like patterns. A central, faint silhouette of a brain is visible, composed of intricate circuitry. To the left, there are wavy, organic lines. To the right, a thin, wavy line with small dots extends across the frame.

Paldies par uzmanību!