nedeljainformatike v8.0

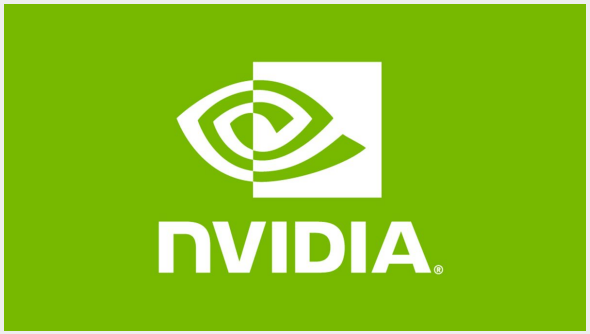
Kako Nvidia podstiče revoluciju u veštačkoj inteligenciji (AI)

Nikola Spasojević
Nvidia

Matematička gimnazija

17. 05. 2022.

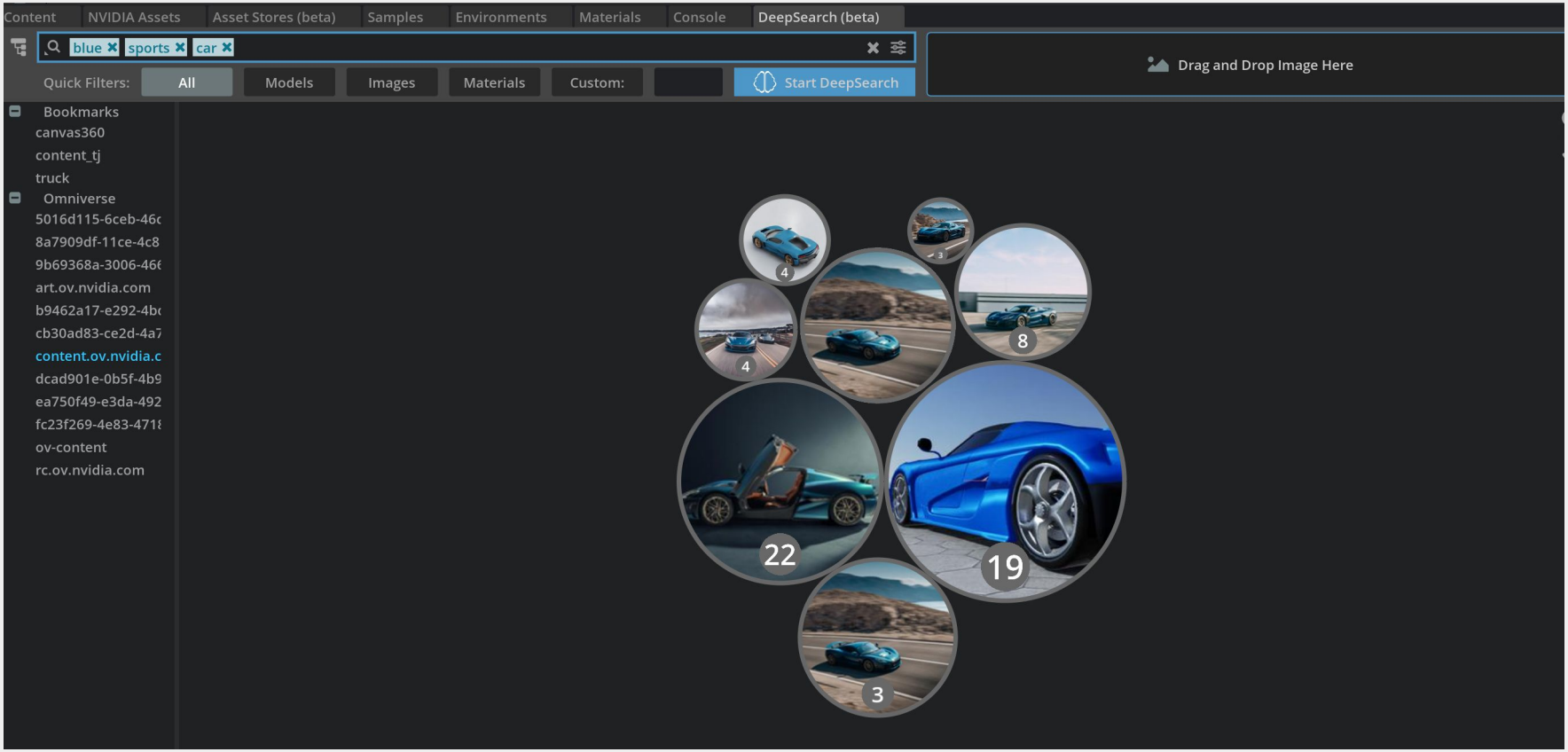
Uvod: Gde sam radio kao softverski inženjer



Trenutni projekti u Nvidia

Omniverse:

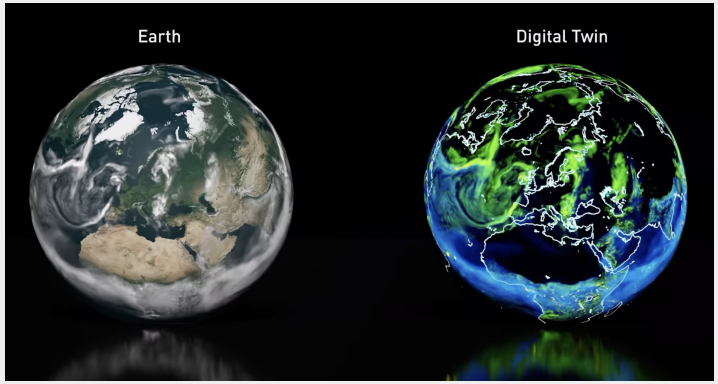
- Pretraga 3D objekata sa tekstom



Trenutni projekti u Nvidia

Digital Twins:

- Virtualne fabrike
- “Veštačke” datoteke za samovozeče automobile
- Simulacija klime



Novi razvoji u AI

Large Language Model (LLM)

- ChatGPT
- Github Copilot

Open Source Machine Learning

- HuggingFace Model Hub

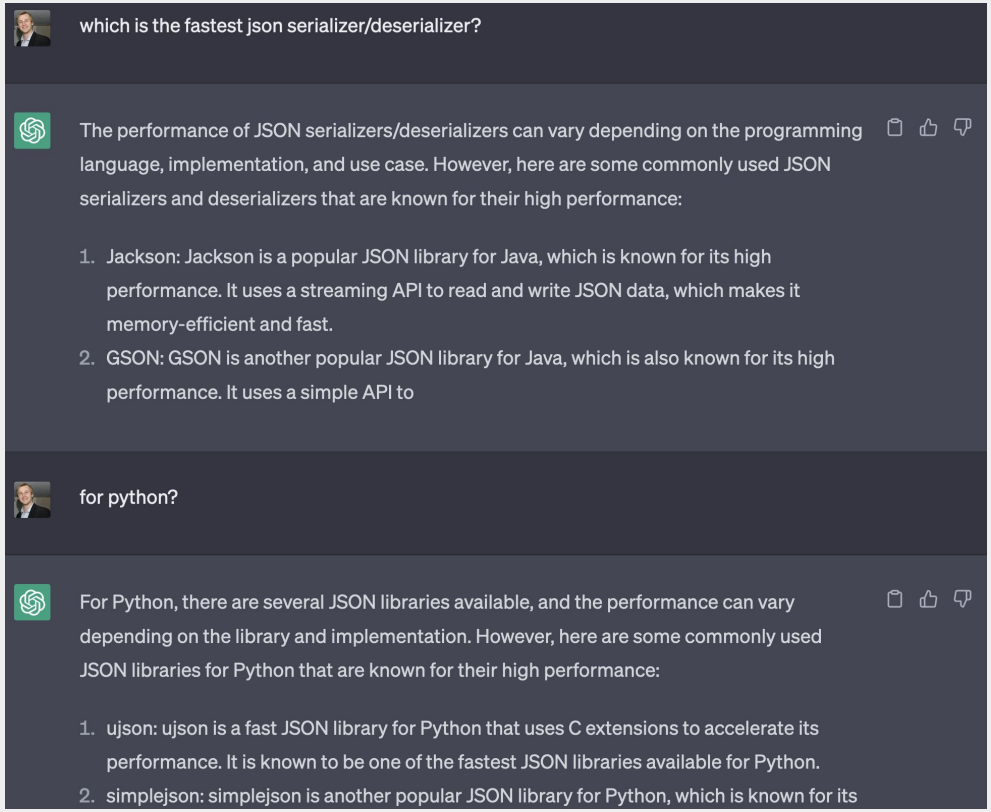
Stable Diffusion

- Text-to-Image models

Kako softverski inženjer koristi AI u svakodnevnici

[ChatGPT]

- Chatbot koji govori ljudksim jezikom (natural language)
- Razume “kontekst” teksta i slika



Kako softverski inženjer koristi AI u svakodnevnici

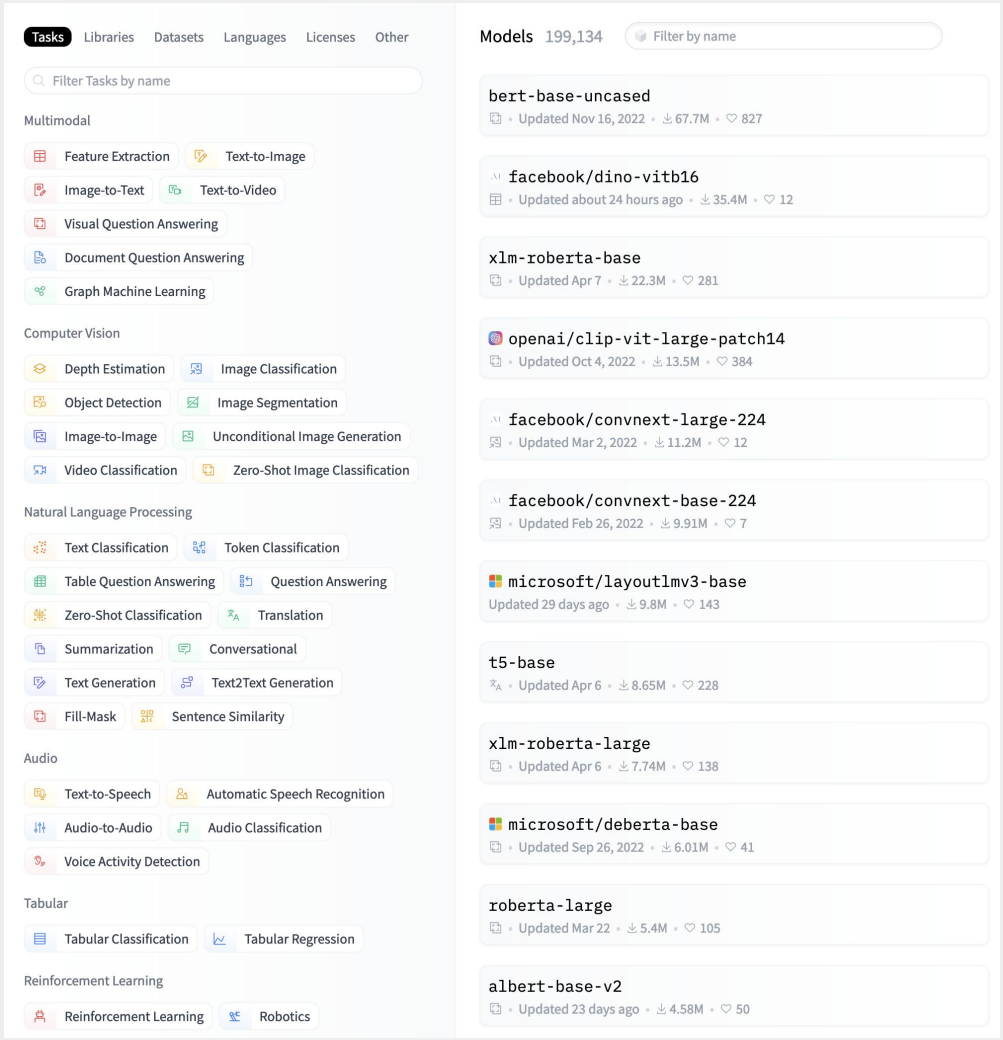
[Github Copilot]

- Ljudksim jezikom preporučuje kod/opis celih funkcija
- Uči “kontekst” tvog koda

```python
def binary_search(arr: list, target: int) -> int:
    """
    Binary search implementation.
    """
    low = 0
    high = len(arr) - 1

    while low <= high:
        mid = (low + high) // 2

        if arr[mid] == target:
            return mid
        elif arr[mid] < target:
            low = mid + 1
        elif arr[mid] > target:
            high = mid - 1

    return -1
```

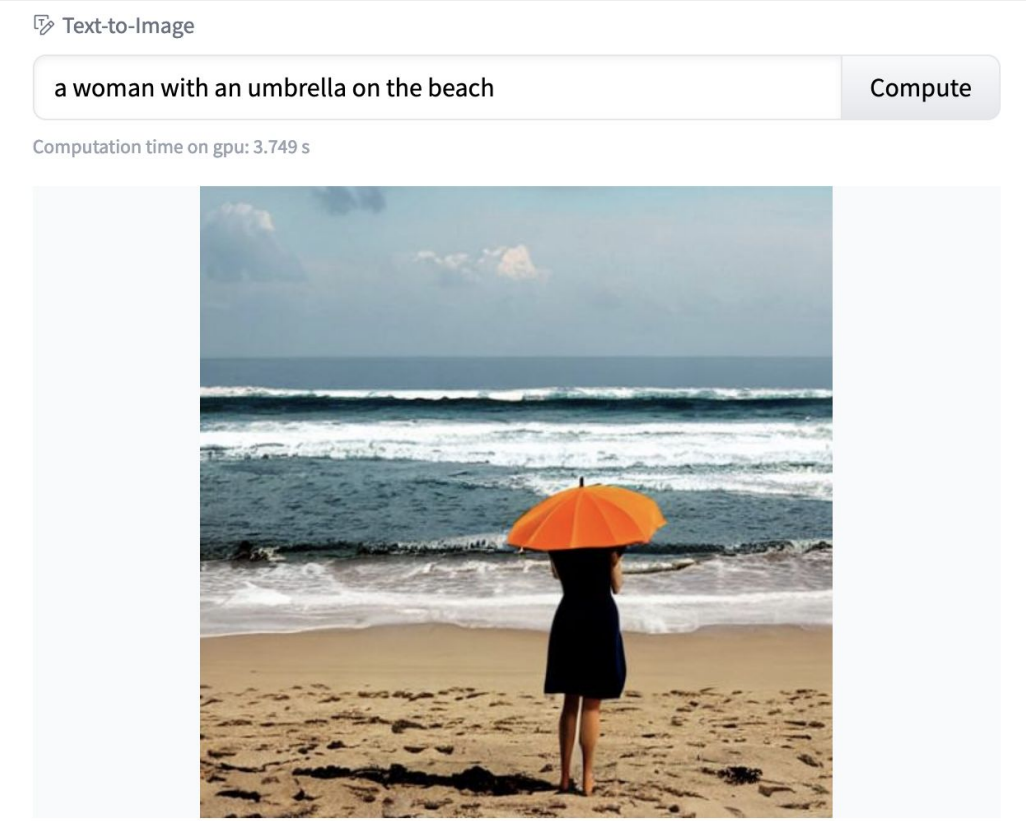
Kako softverski inženjer koristi AI u svakodnevnici

HuggingFace Model Hub - “Demokratizacija” AI modela



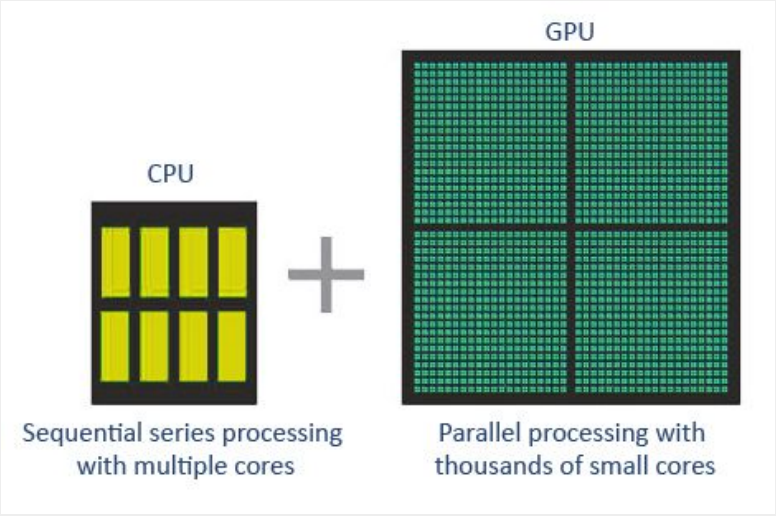
Stable Diffusion (Text-to-Image models)

- Model koji razume “kontekst” u slici



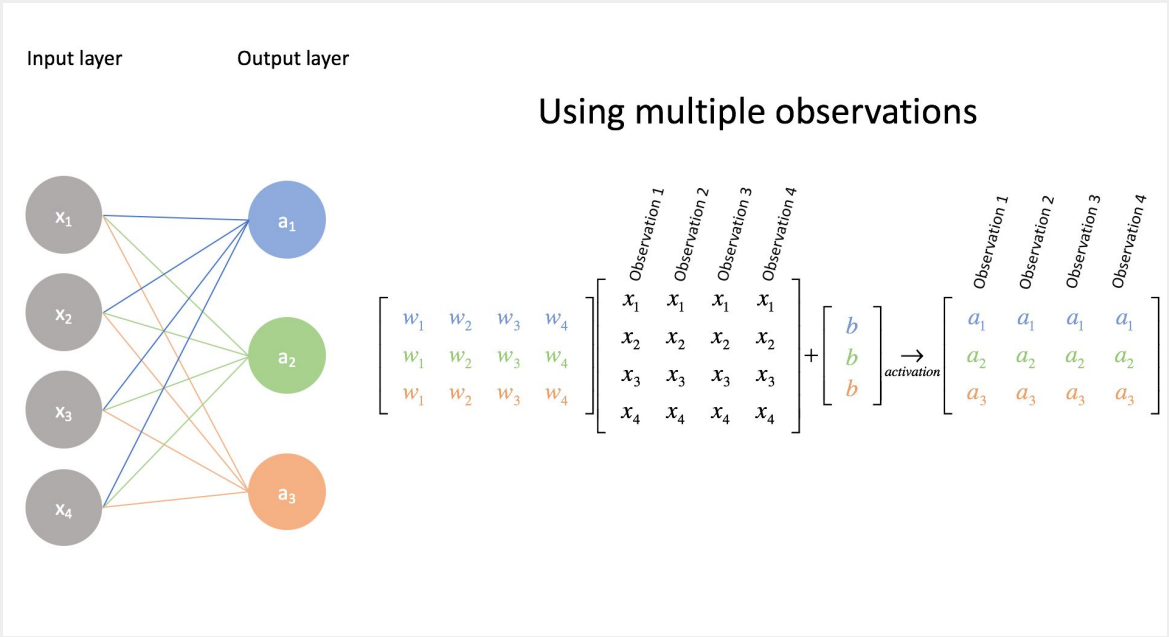
Kako je ovo moguće?

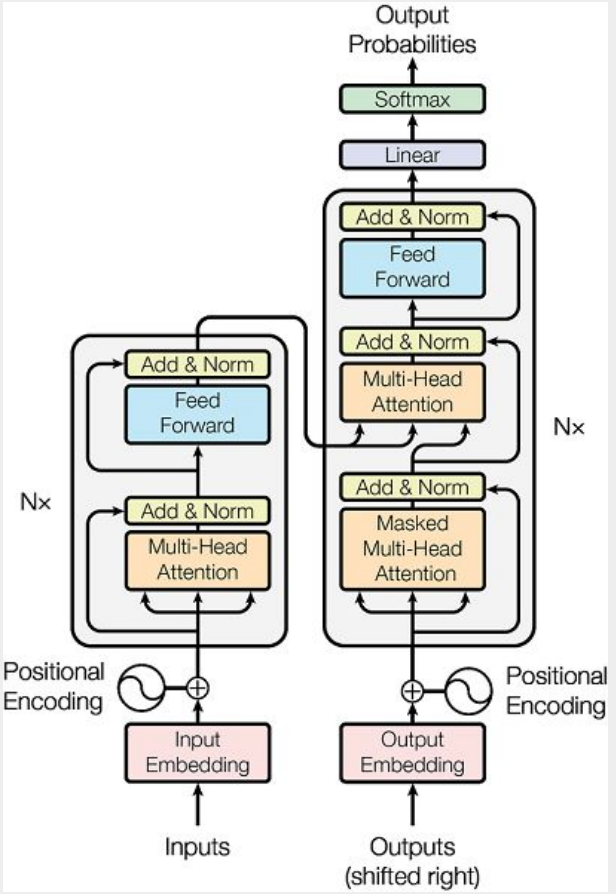
- Šta je GPU?
  - Dizajnirano primarno za paralelizaciju matematičkih operacija
  - Za razliku od CPU-a, koji radi operacije sekvencijalno
  - Originalno korišćeno za grafiku i video



Kako je ovo moguće?

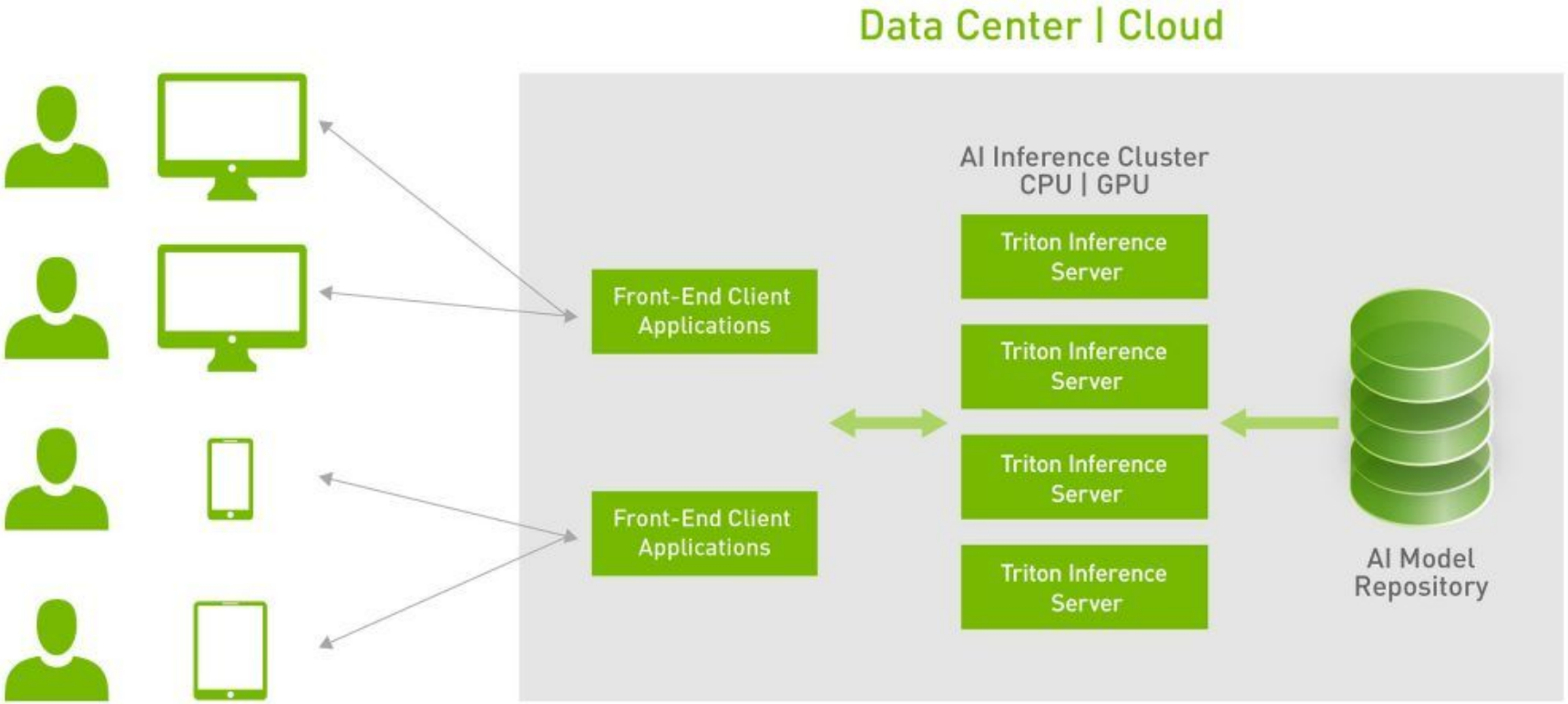
- Zašto se koristi GPU za “treniranje” AI modela?
  - Jer modeli su samo matrice za koje tražimo tačne “weights”
  - Možemo i da paralelizujemo više GPU-a





Deploying a model into production

Nije samo dovoljno trenirati model, treba ih “deploy”-ovati i “serve”-irati:



Deploying a model into production

Zašto je ovo najteži deo?

- Veliki su modeli
  - “Mala” verzija GPT4-a ima 100 milione parametra
  - Ne može svaki kompjuter/GPU da ih “drži” u (RAM) memoriju
  - Treba vremena da se izračuna rezultat
  - Treba imati dobru arhitekture sistema da podnosi mnogo zahteva (npr. google prima stotine hiljade svake sekunde)
- Skupi su GPU-i

Mala napomena

- Programiranje je nova pismenost:
  - Trebamo svi biti spremni za digitalni svet
  - Sa kompjuterima možemo mnogo više da postignem
  - Copilot/leetcode kombinacija za učenje programiranja

- AI je već svuda oko nas:
  - Neće AI da nam ukrade poslove, nago ljudi koje koriste AI
  - Bitno je da znamo kako AI modeli funkcionišu
  - Za šta su sposobni
  - Za šta nisu sposobni

Hvala na pažnji!

Pitanja?