

# Kako Nvidia podstiče revoluciju u veštačkoj inteligenciji (AI)

Nikola Spasojević  
Nvidia

Matematička gimnazija

17. 05. 2022.



THOMSON REUTERS



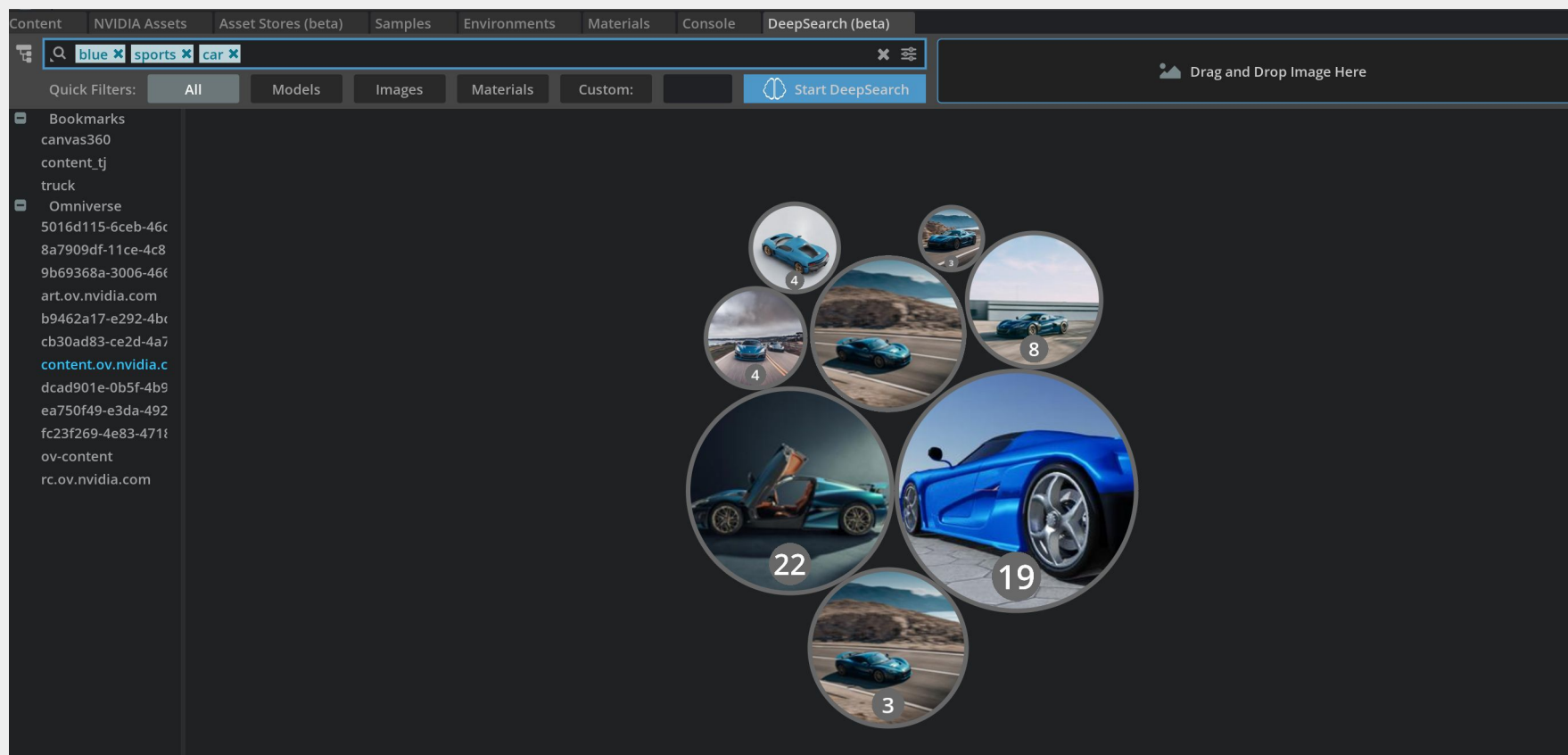
amazon alexa

CREDIT SUISSE 



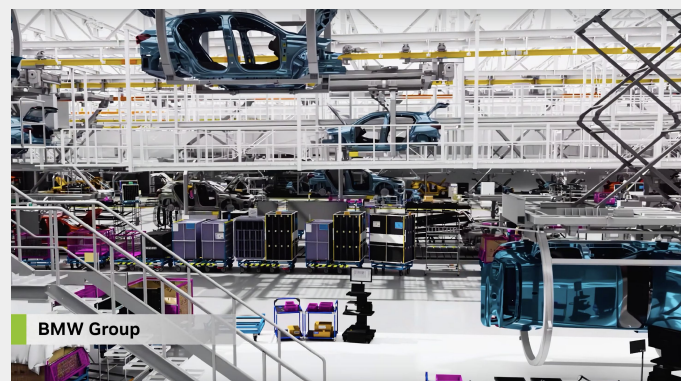
## Omniverse:

- Pretraga 3D objekata sa tekstom



## Digital Twins:

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





## Large Language Model (LLM)

- [ChatGPT](#)
- [Github Copilot](#)

## Open Source Machine Learning

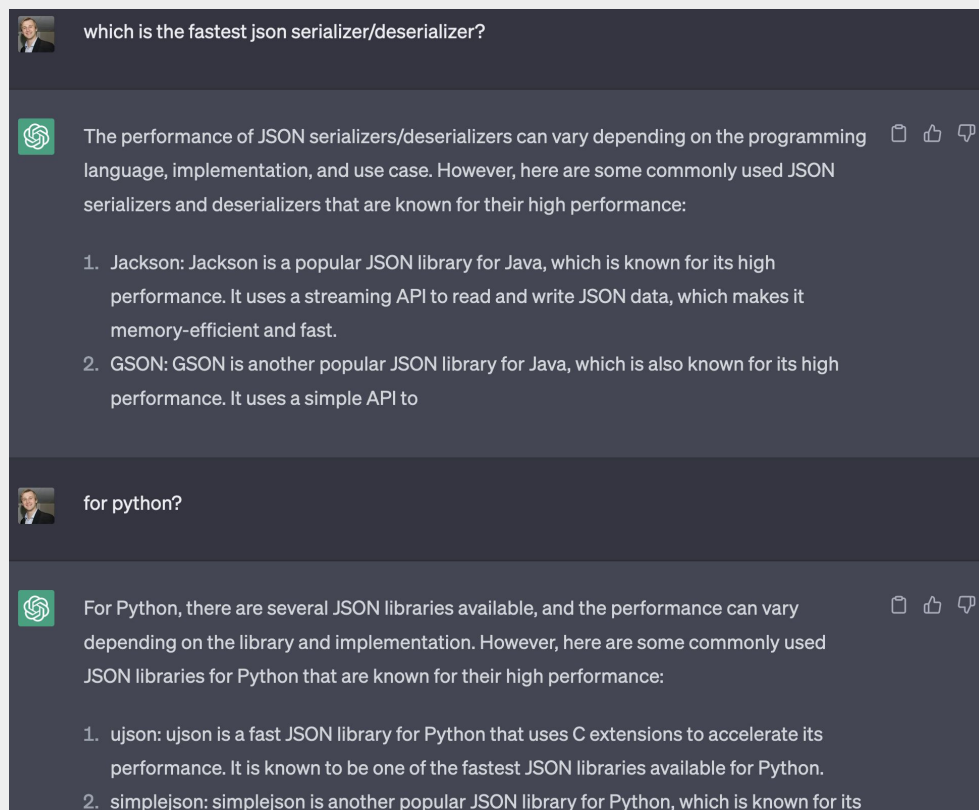
- [HuggingFace Model Hub](#)

## Stable Diffusion

- [Text-to-Image models](#)

## ChatGPT

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



## [Github Copilot](#)

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

```
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 svakodnevnicu



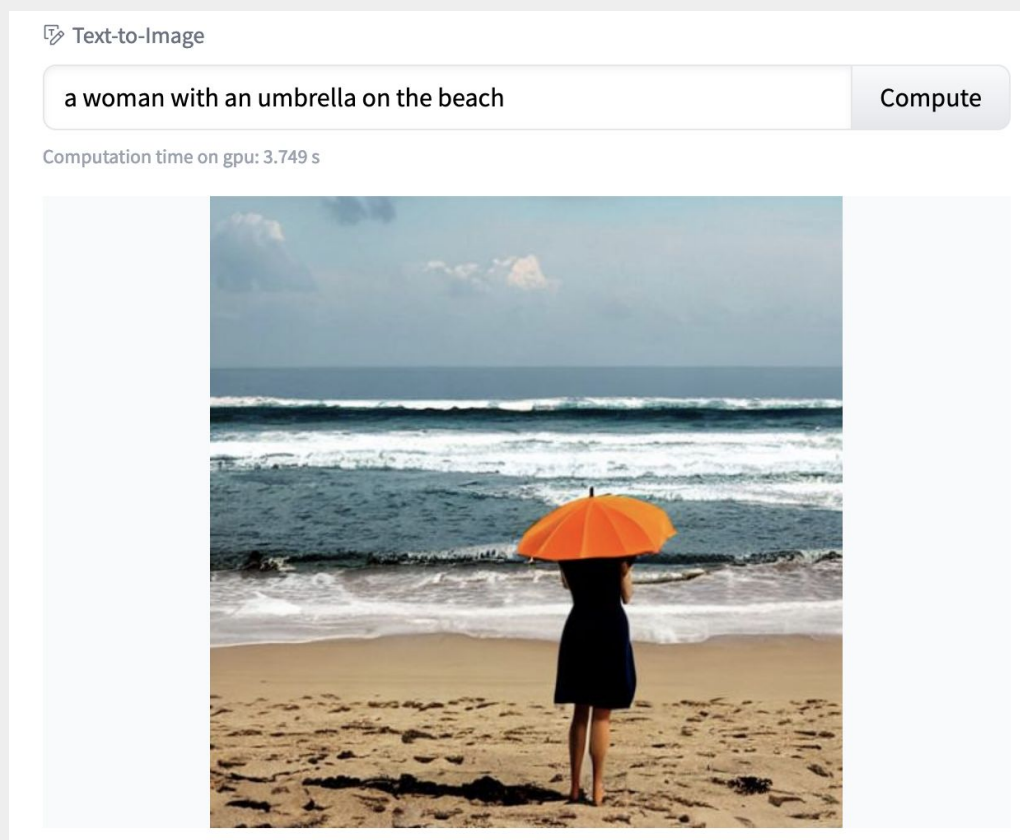
## [HuggingFace Model Hub](https://huggingface.co) - “Demokratizacija” AI modela

The screenshot displays the HuggingFace Model Hub interface. On the left, there are navigation tabs for 'Tasks', 'Libraries', 'Datasets', 'Languages', 'Licenses', and 'Other'. Below these, there are search filters and a list of tasks categorized into Multimodal, Computer Vision, Natural Language Processing, Audio, Tabular, and Reinforcement Learning. On the right, a list of models is shown, including 'bert-base-uncased', 'facebook/dino-vitb16', 'xlm-roberta-base', 'openai/clip-vit-large-patch14', 'facebook/convnext-large-224', 'facebook/convnext-base-224', 'microsoft/layoutlmv3-base', 't5-base', 'xlm-roberta-large', 'microsoft/deberta-base', 'roberta-large', and 'albert-base-v2'. Each model entry includes its name, update date, download count, and heart count.



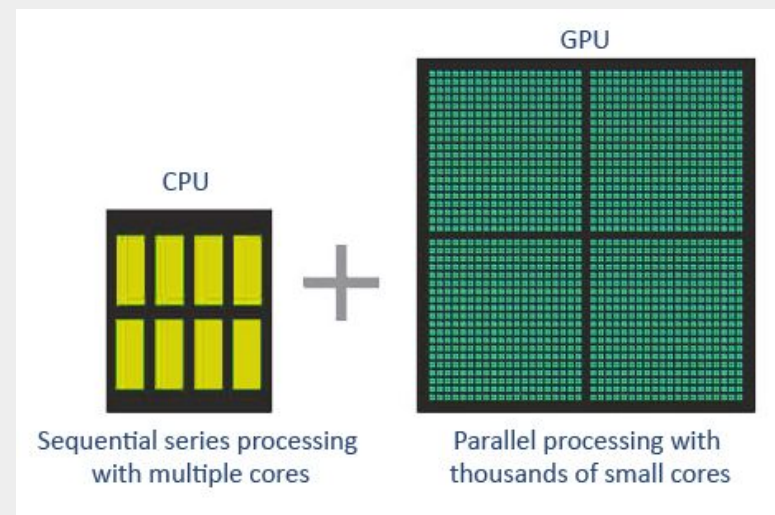
## Stable Diffusion (Text-to-Image models)

- Model koji razume “kontekst” u slici



## • Šta je GPU?

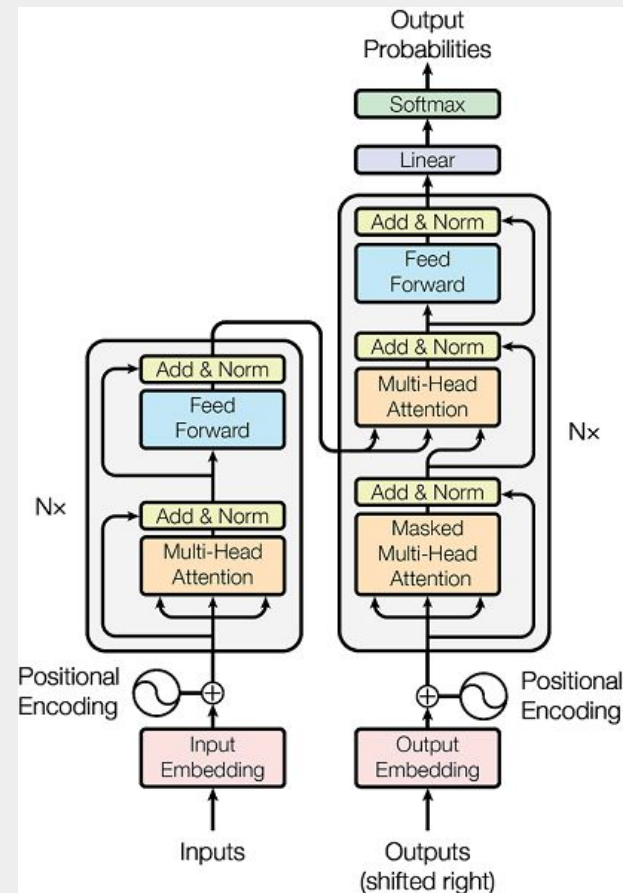
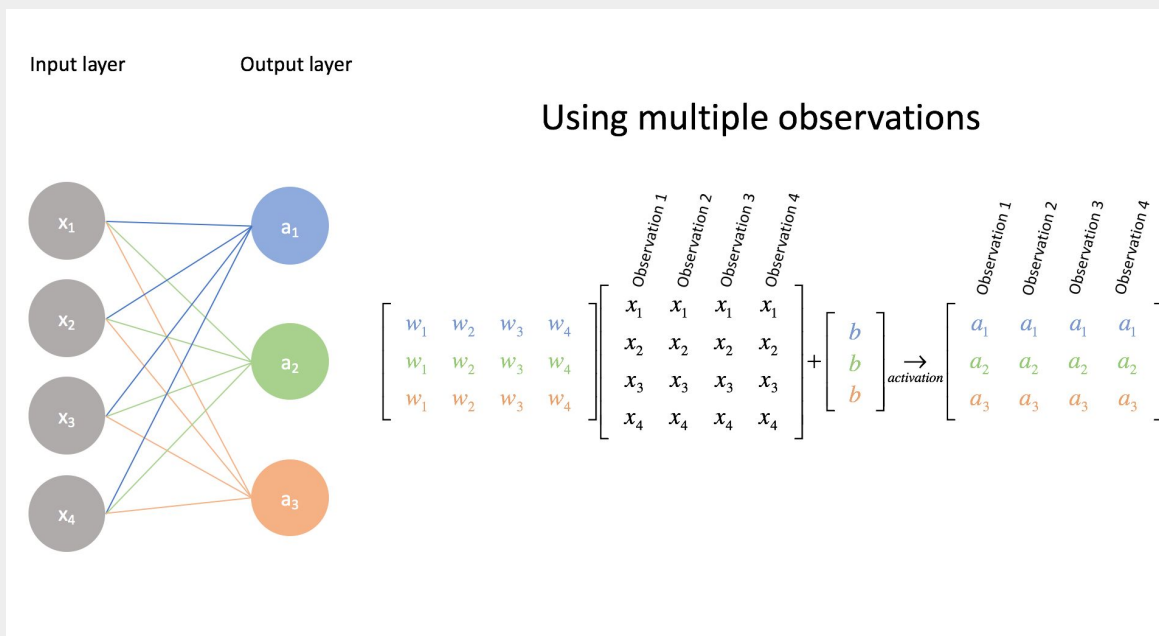
- Dizajnirano primarno za paralelizaciju matematičkih operacija
- Za razliku od CPU-a, koji radi operacije sekvencijalno
- Originalno korišćeno za grafiku i video



# Kako je ovo moguće?

- Zašto se koristi GPU za “treniranje” AI modela?

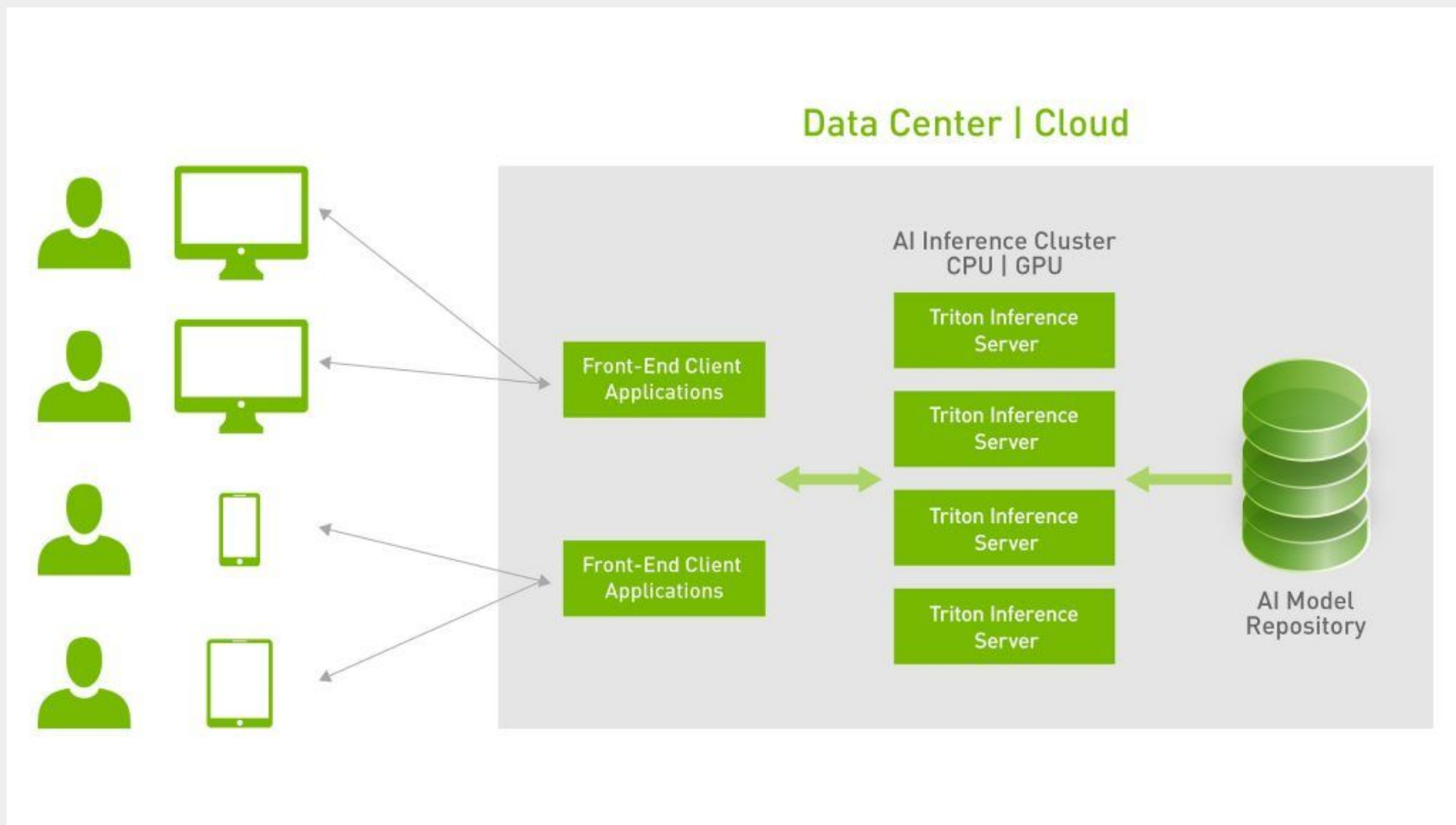
- Jer modeli su samo matrice za koje tražimo tačne “weights”
- Možemo i da paralelizujemo više GPU-a



# Deploying a model into production



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





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 arhitekturu sistema da podnosi mnogo zahteva (npr. google prima stotine hiljade svake sekunde)
- Skupi su GPU-i

- 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?