



TECHNOLOGICKÉ  
CENTRUM PRAHA

# NESTRUKTUROVANÁ DATA 2.0

—  
Kristýna Meislová & Adél Kučera

Výzkum, vývoj a inovace ve statistikách a analýzách 3. 12. 2024

***Dokážete si představit,  
kolik dnes vytvoříte dat?***

# ZETTABYTE ERA

Pokud bychom si představili analogii se vzdáleností...

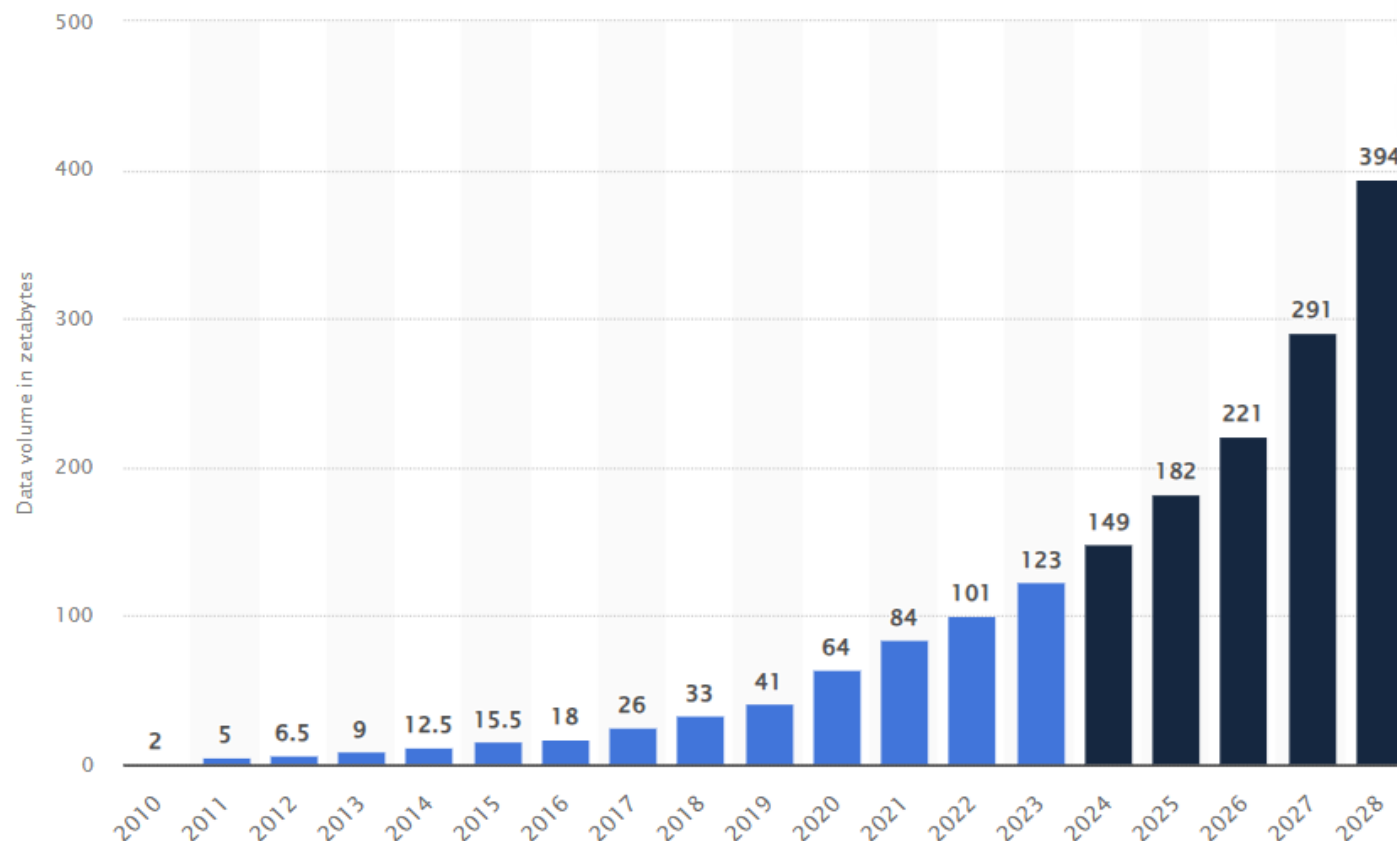
...a **1 terabyte** by představoval cestu z Prahy do Vídně, pak by byl **1 zettabyte** ekvivalentem **10 cest na Měsíc a zpět**.



Zdroj: NASA, Mise Apollo 11 [online] <https://www.nasa.gov/gallery/apollo-11/>

# GOODBYE, ZETTABYTE. HELLO, YOTTABYTE.

- 5,35 mld uživatelů internetu
- > 400 milionů TB dat je vytvořeno každý den
- > 50 % datového provozu tvoří video obsah
- > 80 % jsou nestrukturovaná data
- 2030 > 1 000 ZB



Zdroj: Statista (2024): <https://www.statista.com/statistics/871513/worldwide-data-created/>

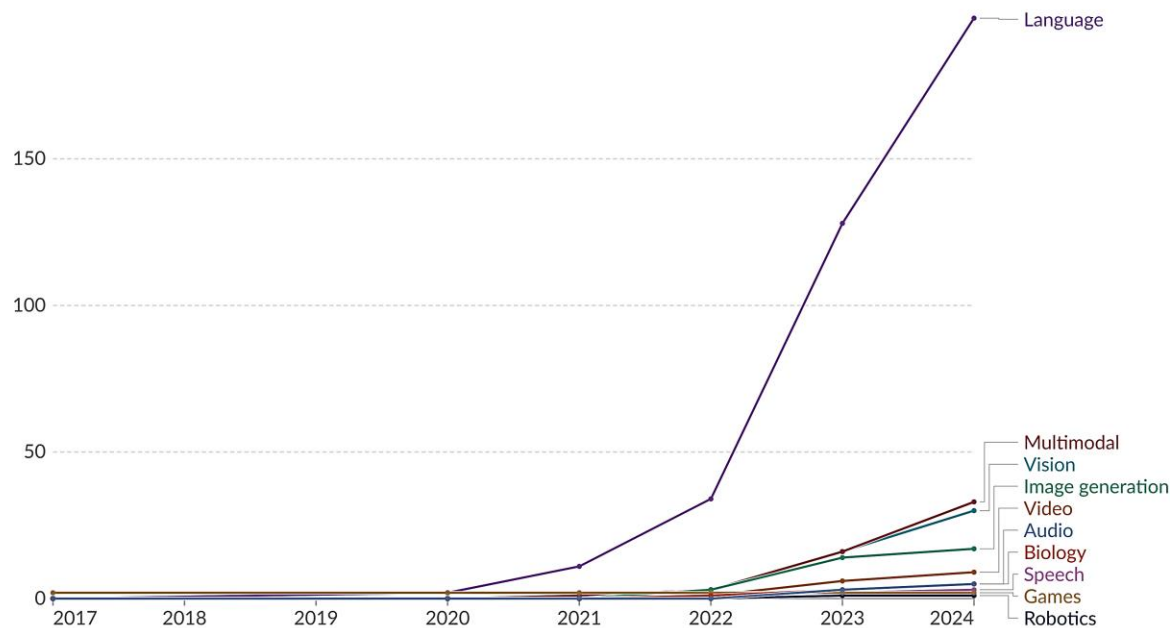
# AI ANALYTIKA

- Katalyzátorem exploze dat je role AI v generování, analýze a distribuci dat

## Cumulative number of large-scale AI models by domain since 2017

Describes the specific area, application, or field in which a large-scale AI model is designed to operate. The 2024 data is incomplete and was last updated 03 November 2024.

Our World in Data



Data source: Epoch (2024)

OurWorldinData.org/artificial-intelligence | CC BY

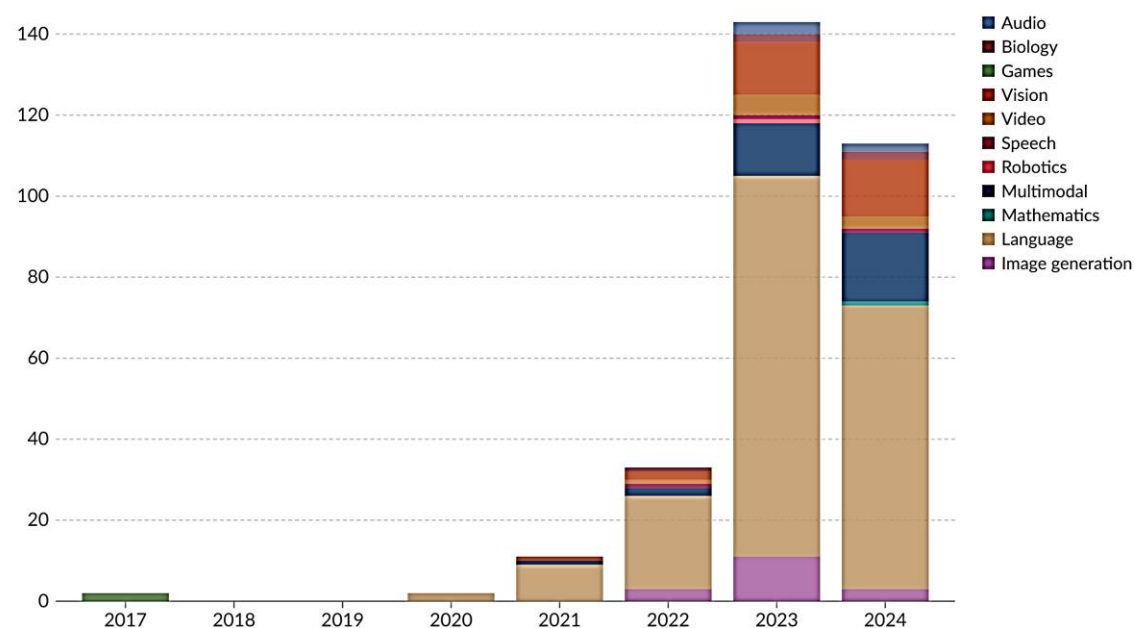
Note: The source defines AI models as "large-scale" when their training compute is confirmed to exceed  $10^{23}$  floating-point operations<sup>1</sup>.

1. **Floating-point operation:** A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

## Number of large-scale AI systems released per year

The 2024 data is incomplete and was last updated 9 September 2024.

Our World in Data



Data source: Epoch (2024)

OurWorldinData.org/artificial-intelligence | CC BY

Note: The source defines AI models as "large-scale" when their training compute is confirmed to exceed  $10^{23}$  floating-point operations<sup>1</sup>.

1. **Floating-point operation:** A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

Zdroj: Epoch (2024) – with major processing by Our World in Data. "Number of large-scale AI systems released per year" [dataset]. Epoch, "Tracking Compute-Intensive AI Models" [original data]. Retrieved in November from <https://ourworldindata.org/grapher/number-of-large-scale-ai-systems-released-per-year>



# AI ANALYTIKA

- Transformace datové analýzy
  - Zpřístupnění pokročilých analytických metod a nástrojů širšímu okruhu uživatelů
  - Zpracování strukturovaných i nestrukturovaných dat (kombinace technik)
  - Umožňuje uživatelům využívat přirozený jazyk a „klást otázky“
  - Eliminace rutinních, časově náročných úkolů
- Nadšení vs. realita
  - Omezení trénovacích dat
  - Omezení analyzovaných dat
  - Omezení metod/modelů
  - Někdo stále musí řídit a lidský úsudek je nezbytný

# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

- Při využití AI v analýze jsou pokyny pro AI (*prompt*) zadávány v přirozeném jazyce

SQL

```
1 select *  
2 from moje_bezva_tabulka  
3 where rok > 2019  
4 ;
```

Přirozený jazyk pro jazykový model

S čím ti můžu pomoci?



moje\_bezva\_tabulka.xlsx  
Tabulka

Vyber z nahrané tabulky všechny řádky, kde je rok větší než 2019.



# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

- Je potřeba věnovat velkou pozornost vytváření promptu (přesně popsat podobu vstupních dat, očekávané úkony, podobu výstupu, ...)
  - Vyladit jednoznačné vyznění

## S čím ti můžu pomoci?



moje\_bezva\_tabulka.xlsx

Tabulka



Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019.





# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

Není na škodu důležité body zadání zopakovat (klíčové úkony nebo ty, kde se model odchyluje od zadání)

## S čím ti můžu pomoci?



moje\_bezva\_tabulka.xlsx  
Tabulka



Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019.  
Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.



# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

Otestovat odpovědi a zpřesnit zadání (vždy testovat před využitím AI „naostro“)

## S čím ti můžu pomoci?



moje\_bezva\_tabulka.xlsx  
Tabulka



Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019. Hodnoty ve sloupci "rok" považuj za čísla. Vynech všechny řádky, kde ve sloupci "rok" je hodnota větší než 9000. Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.



# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

... trochu ten úkol rozšíříme.

Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019. Hodnoty ve sloupci "rok" považuj za čísla. Vynech všechny řádky, kde ve sloupci "rok" je hodnota větší než 9000. Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.

Z vybraných dat vezmi sloupec "popis" a shrň ho do jedné věty, která vystihuje podstatu ve vztahu k oborovému zařazení. Přidej do tabulky jako nový sloupec "shrnutí".

# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

Využít návodné příklady (ilustrovat zadání)

Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019. Hodnoty ve sloupci "rok" považuj za čísla. Vynech všechny řádky, kde ve sloupci "rok" je hodnota větší než 9000. Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.

Z vybraných dat vezmi sloupec "popis" a shrň ho do jedné věty, která vystihuje podstatu ve vztahu k oborovému zařazení. Přidej do tabulky jako nový sloupec "shrnutí".

Postupuj obdobně jako v tomto příkladu:

Hodnota ze sloupce popis:

Pokročilá bioinformatická platforma pro imunoterapii rakoviny. Platformy pro léčbu rakoviny, které využívají adaptivní imunitní systém, prokázaly hluboké regrese nádorů včetně úplného vyléčení. Důležité je, že technologický pokrok v oblasti sekvenování nové generace (NGS) poprvé umožňuje vývoj personalizované imunoterapie rakoviny, která se zaměřuje na mutace specifické pro pacienta. Klinickému použití však v současné době brání specifická úzká místa v bioinformatice, která se snažíme v tomto návrhu řešit. Celkovým cílem naší mezioborové sítě předních odborníků v oblasti bioinformatiky a nádorové imunologie je vyvinout pokročilou bioinformatickou platformu pro PERRSONALIZOVANOU NÁDOROVOU IMUNOTERAPII. Konkrétně se snažíme vyvinout: 1) databázi pro integraci dat NGS, snímků celých tkáňových preparátů nádorových řezů a klinických dat. Pro zvýšení použitelnosti a sdílení dat budeme využívat technologie sémantického webu a poskytneme standardizovaná rozhraní k sadě analytických nástrojů. 2) Nástroje pro automatizovanou kvantifikaci lymfocytů infiltrujících nádor pomocí snímků celých tkáňových preparátů a dat NGS pro stratifikaci pacientů. 3) analytická pipeline pro individualizované vakcíny proti rakovině řízené NGS, včetně klíčových komponent pro analýzu dat NGS a výběr epitopů pro výběr cílů vakcinace. 4) metoda pro odvození sekvencí T-buněčných receptorů (TCR) z dat NGS a předpověď specifičnosti TCR.

Shrnutí:

Mezioborový projekt zaměřený na bioinformatiku a nádorovou imunologii vyvíjí pokročilou bioinformatickou platformu pro personalizovanou nádorovou imunoterapii.

# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

Eliminovat „halucinace“ – explicitně zadat AI, aby si nevymýšlela, když neví (nebo třeba aby nečerpala znalosti z trénovacích dat, ale pouze z vašeho vstupu)

Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019. Hodnoty ve sloupci "rok" považuj za čísla. Vynech všechny řádky, kde ve sloupci "rok" je hodnota větší než 9000. Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.

Z vybraných dat vezmi sloupec "popis" a shrň ho do jedné věty, která vystihuje podstatu ve vztahu k oborovému zařazení. Přidej do tabulky jako nový sloupec "shrnutí".

Zmiňuj pouze ty obory, které jsou uvedeny v textu sloupce "popis". Pokud si nejsi jistá oborovým zařazením, nehádej, ale napiš NEVÍM.

Postupuj obdobně jako v tomto příkladu:

Hodnota ze sloupce popis:

Pokročilá bioinformatická platforma pro imunoterapii rakoviny. Platformy pro léčbu rakoviny, které využívají adaptivní imunitní systém, prokázaly hluboké regrese nádorů včetně úplného vyléčení. Důležité je, že technologický pokrok v oblasti sekvenování nové generace (NGS) poprvé umožňuje vývoj personalizované imunoterapie rakoviny, která se zaměřuje na mutace specifické pro pacienta. Klinickému použití však v současné době brání specifická úzká místa v bioinformatice, která se snažíme v tomto návrhu řešit. Celkovým cílem naší mezioborové sítě předních odborníků v oblasti bioinformatiky a nádorové imunologie je vyvinout pokročilou bioinformatickou platformu pro PERRSONALIZOVANOU NÁDOROVOU IMUNOTERAPII. Konkrétně se snažíme vyvinout: 1) databázi pro integraci dat NGS, snímků celých tkáňových preparátů nádorových řezů a klinických dat. Pro zvýšení použitelnosti a sdílení dat budeme využívat technologie sémantického webu a poskytneme standardizovaná rozhraní k sadě analytických nástrojů. 2) Nástroje pro automatizovanou kvantifikaci lymfocytů infiltrujících nádor pomocí snímků celých tkáňových preparátů a dat NGS pro stratifikaci pacientů. 3) analytická pipeline pro individualizované vakcíny proti rakovině řízené NGS, včetně klíčových komponent pro analýzu dat NGS a výběr epitopů pro výběr cílů vakcinace. 4) metoda pro odvození sekvencí T-buněčných receptorů (TCR) z dat NGS a předpověď specifičnosti TCR.

# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

Stanovit přesnou podobu výstupů, včetně CSV tabulek, aby byly výstupy co nejvíc automatizovaně využitelné pro další analýzu

```
Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019. Hodnoty ve sloupci "rok" považuj za čísla. Vynech všechny řádky, kde ve sloupci "rok" je hodnota větší než 9000. Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.
```

```
Z vybraných dat vezmi sloupec "popis" a shrň ho do jedné věty, která vystihuje podstatu ve vztahu k oborovému zařazení. Přidej do tabulky jako nový sloupec "shrnutí". Výsledek formátuj do tabulky CSV, která má přesně shodné názvy sloupců jako zdrojová tabulka, plus má přidáný sloupec "shrnutí". Do výstupu zařaď pouze tabulku a nic jiného, žádné jiné texty nebo komentáře.
```

```
Zmiňuj pouze ty obory, které jsou uvedeny v textu sloupce "popis". Pokud si nejsi jistá oborovým zařazením, nehádej, ale napiš NEVÍM.
```

Postupuj obdobně jako v tomto příkladu:

Hodnota ze sloupce popis:

Pokročilá bioinformatická platforma pro imunoterapii rakoviny. Platformy pro léčbu rakoviny, které využívají adaptivní imunitní systém, prokázaly hluboké regrese nádorů včetně úplného vyléčení. Důležité je, že technologický pokrok v oblasti sekvenování nové generace (NGS) poprvé umožňuje vývoj personalizované imunoterapie rakoviny, která se zaměřuje na mutace specifické pro pacienta. Klinickému použití však v současné době brání specifická úzká místa v bioinformatice, která se snažíme v tomto návrhu řešit. Celkovým cílem naší mezioborové sítě předních odborníků v oblasti bioinformatiky a nádorové imunologie je vyvinout pokročilou bioinformatickou platformu pro PERRSONALIZOVANOU NÁDOROVOU IMUNOTERAPII. Konkrétně se snažíme vyvinout: 1) databázi pro integraci dat NGS, snímků celých tkáňových preparátů nádorových řezů a klinických dat. Pro zvýšení použitelnosti a sdílení dat budeme využívat technologie sémantického webu a poskytneme standardizovaná rozhraní k sadě analytických nástrojů. 2) Nástroje pro automatizovanou kvantifikaci lymfocytů infiltrujících nádor pomocí snímků celých tkáňových preparátů a dat NGS pro stratifikaci pacientů. 3) analytická pipeline pro individualizované vakcíny proti rakovině řízené NGS včetně blízkých komponent pro analýzu dat NGS a příbř. epigen. pro příbř. síly rakoviny. 4) metod.



# PROMPTOVÁNÍ – OSVĚDČENÉ POSTUPY

Číslovat úkoly, strukturovat zadání do odstavců.

Mám pro tebe dva navazující úkoly:

- 1) Vyber z nahrané tabulky všechny řádky, kde je hodnota ve sloupci "rok" větší než 2019. Hodnoty ve sloupci "rok" považuj za čísla. Vynech všechny řádky, kde ve sloupci "rok" je hodnota větší než 9000. Nezapomeň zkontrolovat, zda jsi vybrala opravdu všechny řádky podle požadavku.
- 2) Z vybraných dat vezmi sloupec "popis" a shrň ho do jedné věty, která vystihuje podstatu ve vztahu k oborovému zařazení. Přidej do tabulky jako nový sloupec "shrnutí". Výsledek formátuj do tabulky CSV, která má přesně shodné názvy sloupců jako zdrojová tabulka, plus má přidáný sloupec "shrnutí". Do výstupu zařaď pouze tabulku a nic jiného, žádné jiné texty nebo komentáře.

Zmiňuj pouze ty obory, které jsou uvedeny v textu sloupce "popis". Pokud si nejsi jistá oborovým zařazením, nehádej, ale napiš NEVÍM.

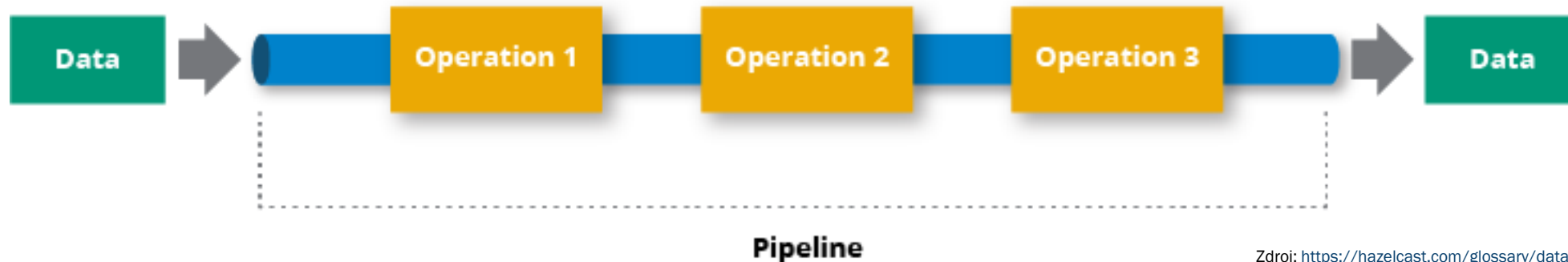
Postupuj obdobně jako v tomto příkladu:

Hodnota ze sloupce popis:

Pokročilá bioinformatická platforma pro imunoterapii rakoviny. Platformy pro léčbu rakoviny, které využívají adaptivní imunitní systém, prokázaly hluboké regrese nádorů včetně úplného vyléčení. Důležité je, že technologický pokrok v oblasti sekvenování nové generace (NGS) poprvé umožňuje vývoj personalizované imunoterapie rakoviny, která se zaměřuje na mutace specifické pro pacienta. Klinickému použití však v současné době brání specifická úzká místa v bioinformatice, která se snažíme v tomto návrhu řešit. Celkovým cílem naší mezioborové sítě předních odborníků v oblasti bioinformatiky a nádorové imunologie je vyvinout pokročilou bioinformatickou platformu pro PERRSONALIZOVANOU NÁDOROVOU IMUNOTERAPII. Konkrétně se snažíme vyvinout: 1) databázi pro integraci dat NGS, snímků celých tkáňových preparátů nádorových řezů a klinických dat. Pro zvýšení použitelnosti a sdílení dat budeme využívat technologie sémantického webu a poskytneme standardizovaná rozhraní k sadě analytických nástrojů. 2) Nástroje pro

# ZAČÍT JEDNODUCHÝMI ÚKOLY

- AI nejlépe pracuje s jasně zadanými a ohraničenými úkoly (podobně jako nakonec i my 😊)
- Děláte ve své práci repetitivní úkoly? Co je zkusit nahradit AI?
- Komplexní úkoly se skládají z dílčích úkolů. Dílčí jednoduché úkony zkusit co nejvíc automatizovat (nejen pomocí AI). Ty si spojit do „potrubí“, kde výstup z jednoho úkonu je vstupem do dalšího úkonu.



Zdroj: <https://hazelcast.com/glossary/data-pipeline/>

- **!** Výstupům z AI nelze slepě věřit. Hodí se u výstupů z AI pokud možno verifikovat tradičnějšími metodami, zda jsou kvalitní.



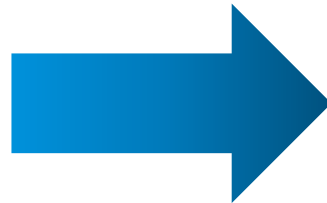
# OSOBNÍ DIGITÁLNÍ (DATA) ANALYST

- Nejen chatovací asistent - AI může fungovat jako osobní asistent při řešení komplexních datových problémů
- **No-code/Low-code řešení**
  - Např. automatické generování reportů, agregovaných zjištění, dashboards apod.
  - Pomoc s vytvářením vhladů (nástroje typu Tableau AI, Power BI, Akkio, Polymer)
  - Často lze využít komunikaci přirozeným jazykem
- **...a nebát se kódu**
  - Umožní větší flexibilitu a kontrolu nad analytickými úlohami
  - Otevírá možnosti pro pokročilejší analýzy a modelování
  - Skvělý nástroj pro debugging, vysvětlivky k existujícímu kódu, usnadňuje práci s populárními AI a ML knihovnamy, může zrychlit řešení
  - Lze využít i asistenty pro kódování (např. GitHub Copilot, Jupyter AI, Colab AI, Gemini Code Assist) nebo pracovat přímo s vybraným chatbotem pro generování kódu

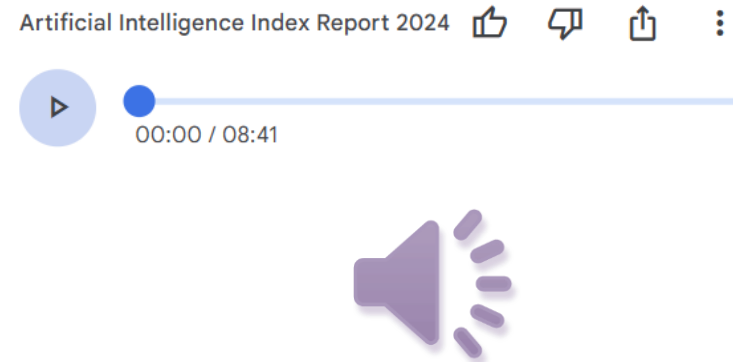


# OSOBNÍ DIGITÁLNÍ (DATA) ANALYST

„Welcome to the seventh edition of the AI Index report. The 2024 Index is our most comprehensive to date and arrives at an important moment when AI’s influence on society has never been more pronounced. This year, we have...“ produced 500 more pages!



- **No-code**
- Jak pojmout stohy analytického textu, když nemáte volné ruce, nechcete nic číst a nemáte náladu si s AI chatovat? 🙄🙄
  - Např. nástroj NotebookLM Google Labs (Gemini)

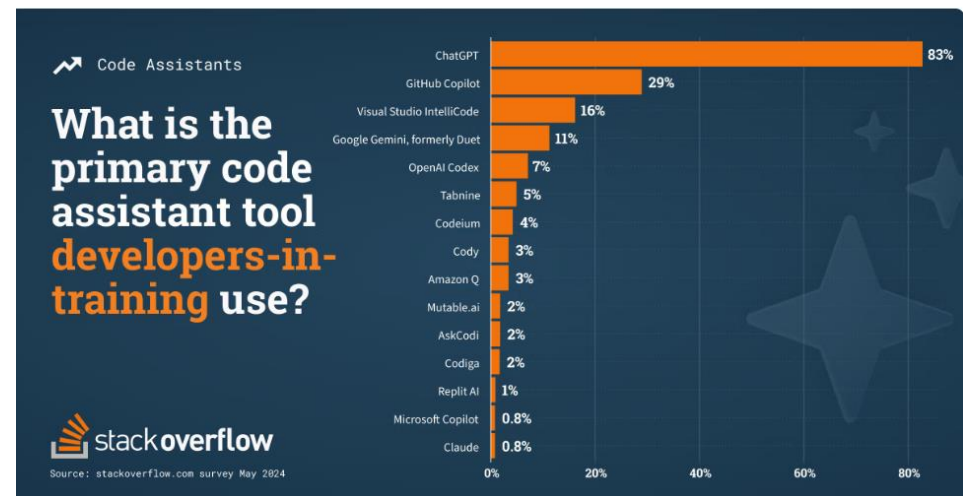
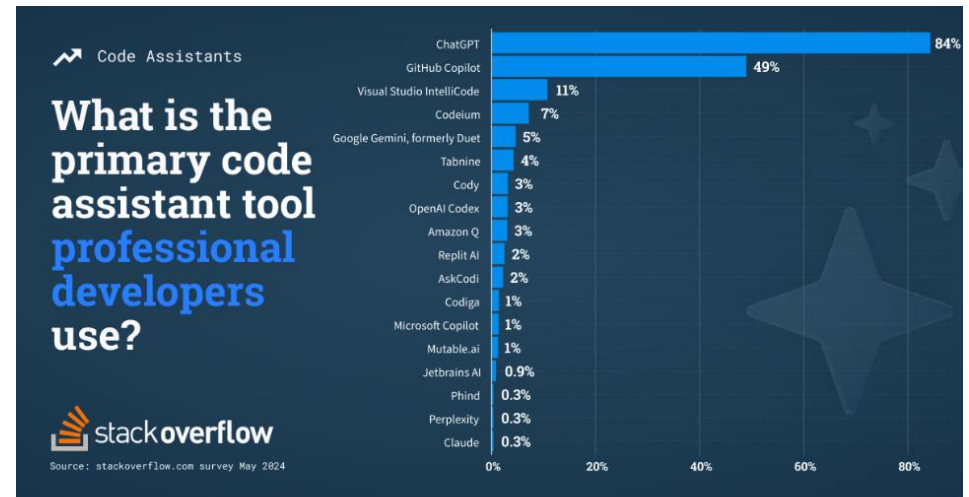


Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark, “The AI Index 2024 Annual Report,” *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.*

# OSOBNÍ DIGITÁLNÍ (DATA) ANALYST – (SOME) CODE

- I zde platí, že AI exceluje jako asistent pro dílčí úkoly
  - AI dokáže poskytnout užitečný kód ↔ AI se může ztratit a produkovat odpad „fall down a rabbit hole“*
- Uživatel řídí analýzu a její výstupy, nespolehá čistě na built-in možnosti modelu
- Model/typ asistence podle cíle
- Efektivně pomůže se začátky ve vhodném programovacím jazyce pro analytické účely

...a oprostít se od licencemi a omezenými rámci svázaných nástrojů



# STUKTURACE NESTRUKTUROVANÝCH DAT

- AI jako nástroj, jak docela dobře získávat z nestrukturovaných dat strukturovaná data
- Například:
  - Klasifikovat texty dle sentimentu
  - Klasifikovat texty do předem definovaných tematických kategorií
  - Otestovat, zda se text týká nebo netýká daného tématu
  - Vypreparovat z textů údaje o místě, čase, jménech a názvech, apod.
  - Přepsat audio do textu
  - Rozpoznat objekty na obrázku nebo ve videu
  - Shrnutí textu
  - ...

# STUKTURACE NESTRUKTUROVANÝCH DAT



**Klasifikace**

Číselníky

Ukazatele

nacházíte se: [Home](#) > [Klasifikace](#) > [CZ-COICOP\\_2018\\_KL](#)

## Klasifikace individuální spotřeby podle účelu (CZ-COICOP\_2018)

Akronym	Název	Verze	Platnost od	Platnost do	Rodina
CZ-COICOP_2018_KL	Klasifikace individuální spotřeby podle účelu (CZ-COICOP_2018)	1	01.01.2024	09.09.9999	Národní účty

**Struktura**

Seznam položek

Doplňující informace

Ke stažení

### Struktura klasifikace:

- [-] 01 - [Potraviny a nealkoholické nápoje](#)
  - [-] 011 - [Potraviny](#)
    - [+] 0111 - [Obiloviny a výrobky z obilovin \(NT\)](#)
    - [+] 0112 - [Živá zvířata, maso a ostatní části zvířat \(NT\)](#)
    - [+] 0113 - [Ryby a mořské plody \(NT\)](#)
    - [+] 0114 - [Mléko, mléčné výrobky a vejce \(NT\)](#)
    - [+] 0115 - [Oleje a tuky \(NT\)](#)
    - [+] 0116 - [Ovoce a ořechy \(NT\)](#)
    - [+] 0117 - [Zelenina, luštěniny a brambory \(NT\)](#)
    - [+] 0118 - [Cukr a cukrovinky \(NT\)](#)
    - [+] 0119 - [Hotová jídla, polotovary a ostatní potravinářské výrobky \(NT\)](#)
  - [+] 012 - [Nealkoholické nápoje](#)

# STUKTURACE NESTRUKTUROVANÝCH DAT

Mám tuhle klasifikaci potravin ve formátu CSV, viz níže. Sloupec "code" obsahuje kód kategorie, sloupec "name" název kategorie. Kategorie jsou hierarchické, což je vidět podle počtu znaků kódu ve sloupci "code" - kódy se stejným počtem znaků jsou na stejné hierarchické úrovni. Čím méně znaků má kód, tím vyšší je hierarchická úroveň. Rozumíš tomu? Chtěla bych, abys podle klasifikace dokázala v předložených textech identifikovat zmínky o potravinách a uvedla kód a název všech identifikovaných potravin. Důležité je se podívat pořádně a najít vše, co by i málo pravděpodobně mohlo být klasifikováno jako potravina.

Co se týče výstupu, vždy předlož výstup v tabulce, prosím. Uved' vždy též názvy nadřazených kategorií. Názvy sloupců budou: Nejvyšší kategorie (kód - název), kategorie 2 (kód - název), kategorie 3 (kód - název), kategorie 4 (kód - název), nejnižší kategorie (kód - název), poznámka (pro případné upřesnění)

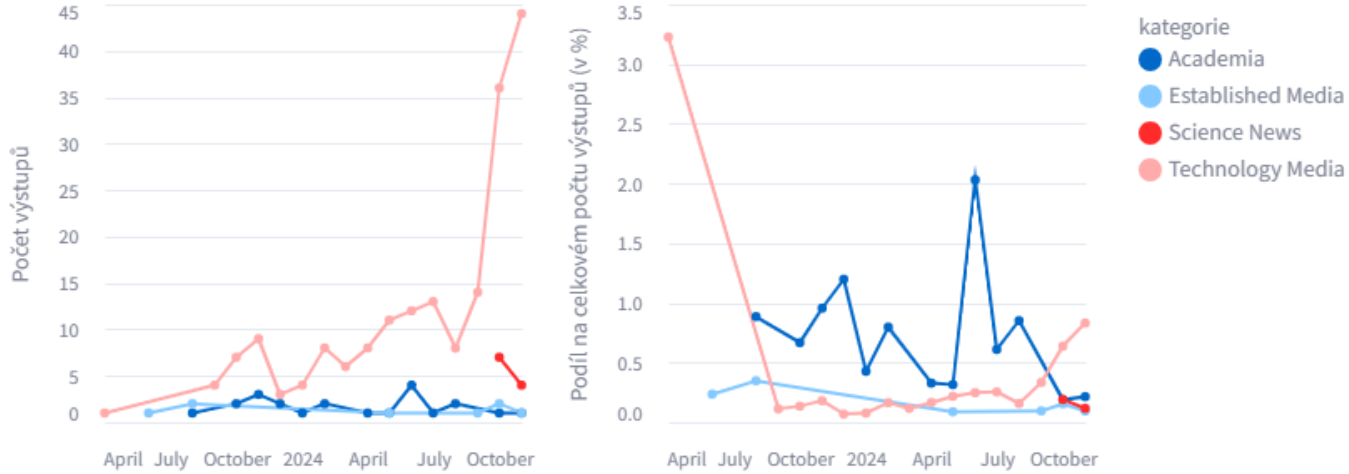
Můžeme jít na to?

# ODHALOVÁNÍ SOUVISLOSTÍ

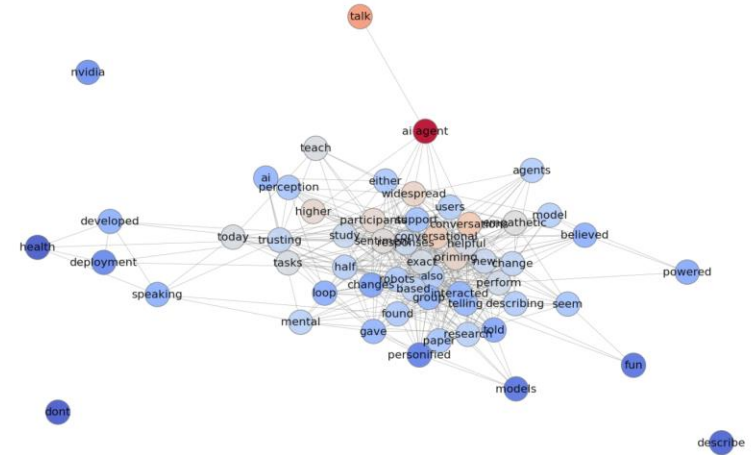
- Využití AI v analytickém procesu by nemělo znamenat ztrátu kontroly nad procesem analýzy.
- Kombinací precizně definovaných analytických kroků a schopnosti AI modelů lze analýzu obohatit o nové perspektivy.
- **Příklad pro doplnění souvislostí:** AI slouží jako „partner“ v analytickém procesu – obohacuje výsledky, zatímco analytik řídí celý postup, tzn.:
  - *Určuje cíle*
  - *Shromažďuje data*
  - *Definuje a designuje analytické kroky*
  - *Generuje dílčí analytické výstupy (analytickou kaskádu)*
  - *Formuluje úkoly pro AI a poskytuje relevantní vstupy.*
  - *Ověřuje výsledky!*



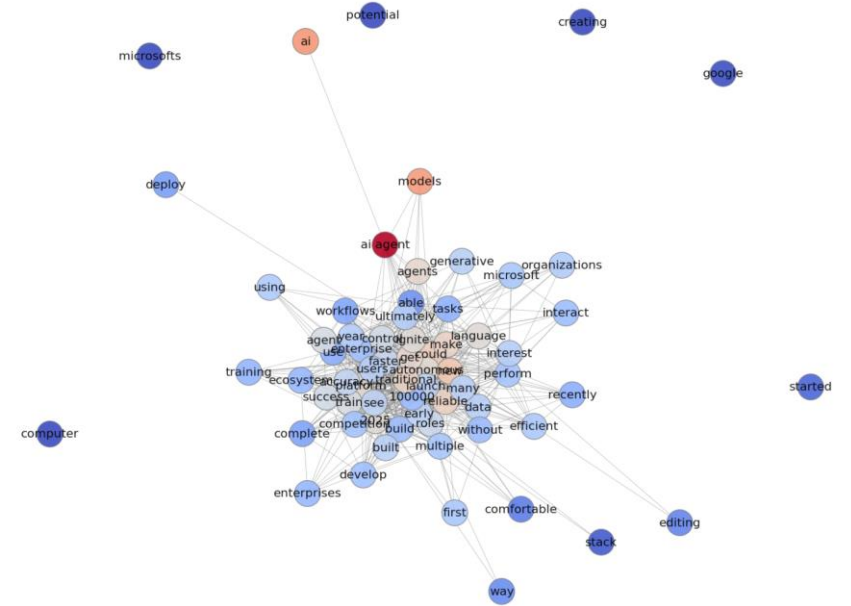
# ODHALOVÁNÍ SOUVISLOSTÍ



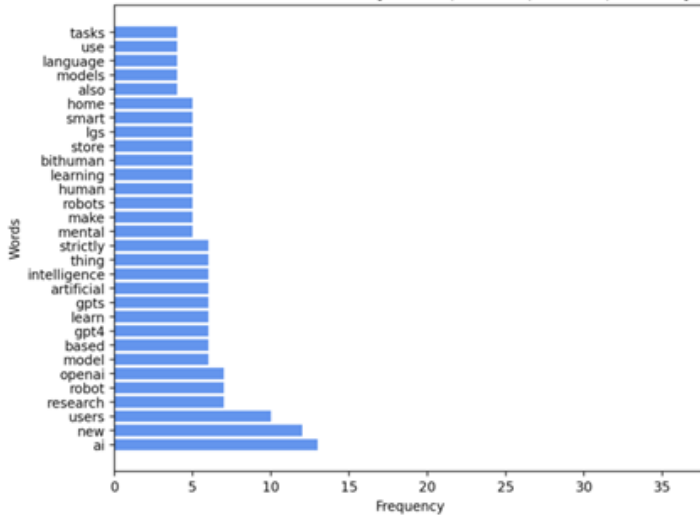
Co-occurrence Network for Time Window: 2023-10 (ai agent)



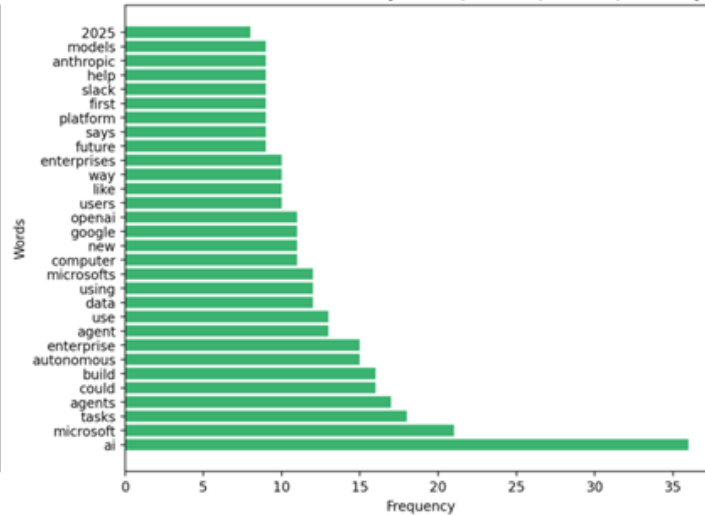
Co-occurrence Network for Time Window: 2024-11 (ai agent)



Most Frequent Context Words in the Oldest Time Window [2023-10, 2023-11, 2023-12, 2024-01]



Most Frequent Context Words in the Most Recent Time Window [2024-08, 2024-09, 2024-10, 2024-11]



# ODHALOVÁNÍ SOUVISLOSTÍ

## Strategie:

- **Strukturovanost:** Definice požadavků na analýzu krok po kroku k usnadnění práce s různými typy vstupních dat.
- **Flexibilita:** Definice parametrů umožňuje datové (multimodální) vstupy (v modelovém příkladě context\_data, visualization\_data, metrics\_data).
- **Modularita:** Snadno přizpůsobitelné pro různé úkoly – stačí změnit data, sledované periody, téma apod.

→ *Kompatibilita s vlastními nástroji a aplikacemi je zajistitelná formátem instrukce při napojení např. na API různých typů modelů nebo i při využití lokálních modelů*

Příklad modelové struktury při využití dalších dílčích analytických vstupů.

```
# === Design instrukce pro multimodální model ===
# Účel: Analyzovat vývoj sledovaného tématu na základě dílčích analytických výstupů
# a formulovat hypotézy o současném stavu i budoucím vývoji.

def analyze_mydata(context_data, visualization_data, metrics_data, time_periods, topic):
    """
    Analyzuj vývoj sledovaného tématu a identifikuj klíčové změny mezi obdobími.

    Parametry:
    - context_data: Textové shrnutí nebo klíčová kontextová slova z předzpracovaných dat
    - visualization_data: Grafy nebo síťové vizualizace reprezentující vztahy v datech
    - metrics_data: Kvantitativní metriky (např. kosinová podobnost, frekvence terminů)
    - time_periods: Časová okna (např. ["2023-10 až 2024-01", "2024-08 až 2024-11"])
    - ...

    Krok 1: Popiš charakteristiky a klíčové rysy sledovaného tématu v prvním období (time_periods[0]).
    - Použij kontextová data ...
    - Identifikuj ...

    Krok 2: Popiš charakteristiky a klíčové rysy sledovaného tématu v druhém období (time_periods[1]).
    - Analyzuj, jak se změnil ...
    - Porovnej ...

    Krok 3: Formuluj hypotézu o změnách:
    - Vysvětli, proč...
    - Využij ...

    Krok 4: Předlož predikci vývoje:
    - Na základě ...

    ...

    Výstup:
    - Strukturovaný text zahrnující ...
    """
    # Zde by byla implementace volání GPT-4 a použití výše uvedené logiky.
    # Příklad:
    return gpt_generate_response(
        f"Analyzuj téma '{topic}' na základě dat pro období {time_periods[0]} a {time_periods[1]}."
        f" Zohleďni: {context_data}, {visualization_data}, {metrics_data}. ..."
    )
```



# ODHALOVÁNÍ SOUVISLOSTÍ

Stručné shrnutí za využití AI asistence (GPT-4o) na základě instrukce a vstupů z analytického procesu:

## ▪ Hodnocení období podzimu 2023

V období od října 2023 do ledna 2024 byla **diskuse o "AI agentech" primárně konceptuální a zaměřovala se na výzkumné a experimentální aplikace**. Klíčová slova naznačují, že pozornost byla soustředěna na potenciál AI agentů a jejich schopnosti. Významnou roli hrála zmínka o jazykových modelech ("GPT4," "language models") a technologických inovacích v kontextu domácí automatizace ("smart home", "bithuman"). To reflektuje spíše akademický a technologie zkoumající přístup. Grafy a síťové analýzy ukazují menší hustotu propojení a širší rozptýl pojmů a podtrhuje diverzitu a **méně rozvinutý praktický rámec**. Průměrná hodnota kosinové podobnosti mezi staršími a novějšími kontexty (0.77) však naznačuje určitý kontinuální vývoj.

## ▪ Hodnocení období podzimu 2024

V období od srpna do listopadu 2024 **se kontext diskuse výrazně posunul směrem k praktickému využití AI agentů**, zejména v podnikovém prostředí. Klíčová slova jako směřují ke komercializaci a zaměření na implementaci v oblastech automatizace a správy dat. Výrazy jako "autonomous" a "tasks" reflektují zaměření na efektivitu a aplikace v pracovních procesech. Síťová analýza zobrazuje vyšší hustotu propojení mezi pojmy a naznačuje tak **užší zaměření na konkrétní oblasti využití**, například při vytváření autonomních systémů. Tento posun je poháněn technologickými giganty, jako jsou Microsoft a Google, kteří stále definují trendy v oboru.

## ▪ Vysvětlení posunu a výhled do roku 2025

Posun od výzkumného a konceptuálního přístupu k praktickým aplikacím **je poháněn technologickou vyspělostí AI modelů, rozvojem infrastruktury a tlakem na komerční využití**. Vzrůstající role podniků při nasazení AI agentů, včetně jejich integrace do pracovních toků, je klíčovým faktorem této transformace. Hypoteticky lze očekávat, že v roce 2025 budou AI agenti více personalizovaní a multimodální, schopní pracovat nejen s textem, ale i obrazem a zvukem. Kromě toho se budou více řešit etické a bezpečnostní otázky spojené s jejich nasazením, zejména v kontextu regulace a správy dat. Očekává se také širší implementace AI agentů v globálních iniciativách, například při řešení změn klimatu nebo optimalizaci zdravotnických systémů.

Posun od výzkumného zaměření k praktickému využití je pravděpodobně poháněn několika faktory:

- Technologický pokrok: Rozvoj multimodálních modelů a autonomních systémů umožňuje širší nasazení AI agentů.
- Poptávka po efektivitě: Tlak na zlepšení pracovních toků a optimalizaci zdrojů v podnikovém prostředí.
- Komerční tlak: Investice klíčových hráčů, jako jsou Microsoft, Google nebo Anthropic, směřují k rychlému uvedení aplikací na trh.

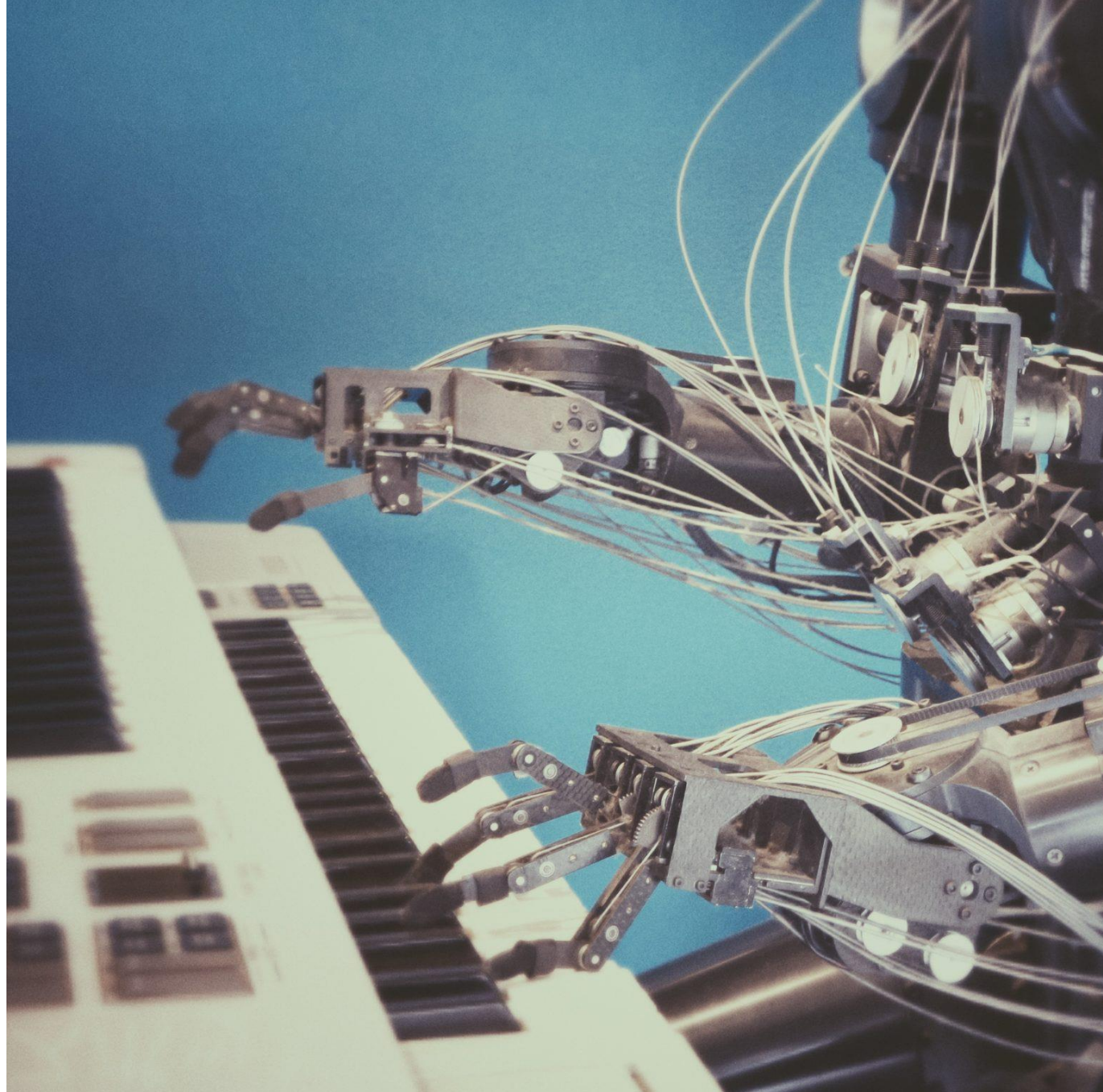
Do budoucna lze očekávat:

- Multimodalitu: AI agenti budou schopni efektivně pracovat s textem, obrazem a zvukem, čímž se rozšíří jejich využití v oblastech jako zdravotnictví nebo výroba.
- Etické výzvy: Vzdávající tlak na regulace a ochranu dat, zejména v kontextu nasazení autonomních systémů.
- Personalizace a specializace: AI agenti budou stále více přizpůsobováni konkrétním potřebám, například v zákaznickém servisu nebo v automatizaci procesů.



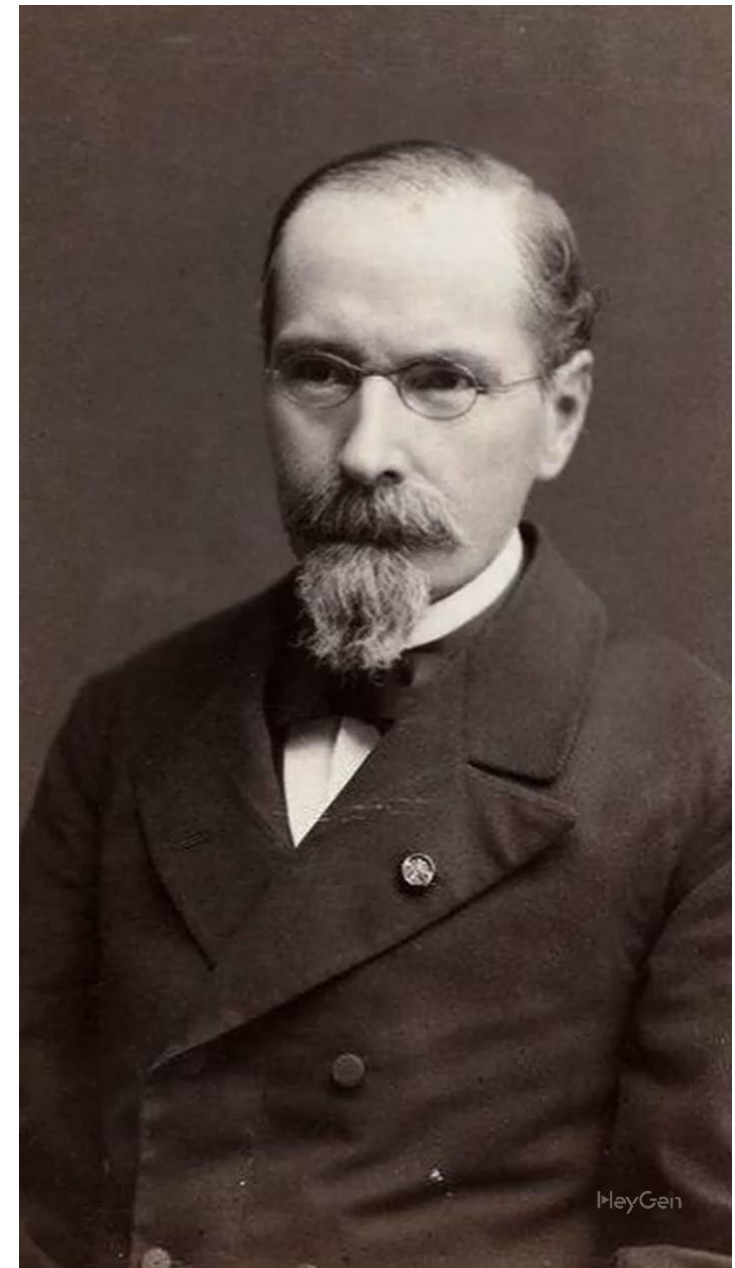
# AI AGENTI

- AI agent: Systém nebo program, který je schopen samostatně vykonávat úkoly tím, že navrhne svůj pracovní postup, využije dostupné nástroje a reviduje postup podle adekvátnosti výsledků
- Teoreticky dokáže nejen obsáhnout vícero úkolů v „potrubí“, ale i nadefinovat si podobu toho potrubí (postupné kroky a jednotlivé použité nástroje)
- V současnosti intenzivně vyvíjeno, ale stále zbývá překonat mnoho technických výzev...  
(OpenAI [ohlásila](#) na leden 2025 spuštění služby Operátor)



# OD TEXTŮ K MULTIMÉDIÍM

- MULTIMODALITA – nové možnosti zdrojů pro **získávání a využití** nestrukturovaných dat z obrázků, videí, hudby, audio výpovědí, výrazů tváře, z prostorových dat, z multimediálních dat z různých čidel, atd...
- Platí to ale i naopak, multimédia lze vytvářet nebo modifikovat na základě dat





# DEEPPFAKE – „FAKE“ NEBO „TOOL“?

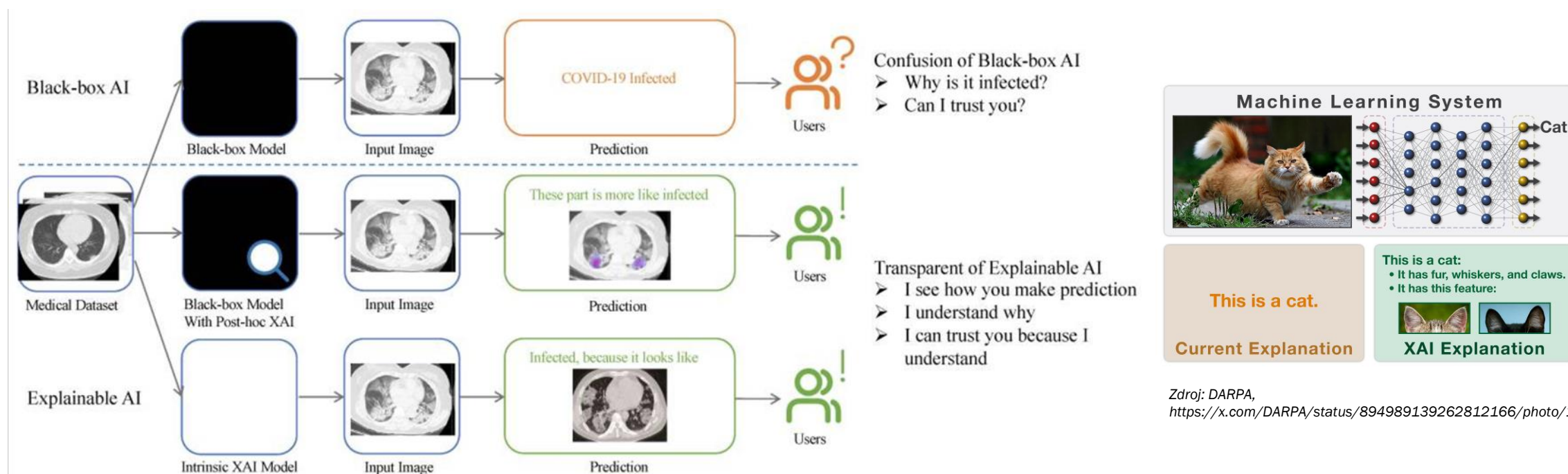
- Vytváření uvěřitelně vypadajících videozáznamů, fotografií, real-time avatarů apod. na jedné straně představuje riziko při zlovolném využití, na druhé straně zajímavý nástroj pro legitimní využití
- **Rizika:** tvorba a šíření dezinformací, podvody a phishingové útoky, revenge porn, narušení biometrických systémů, ad.
- **Příležitosti:** Využití v kultuře a umění, ve vzdělávání, anonymizaci videa, vzdálenou péči a terapii, ad.
- **Co s tím:** Vytvořit jasné etické zásady a právní rámce, nejen metody detekce, ale proaktivní postupy obrany





# TRENDY 2024/2025

- **Explainable AI (XAI)** - zaměřuje se na vývoj AI systémů, jejichž rozhodnutí a chování mohou být pochopeny a interpretovány
  - Pomáhá splňovat právní požadavky na vysvětlitelnost rozhodnutí, zejména v citlivých oblastech jako je zdravotnictví.
  - Očekává se, že trh s Explainable AI vzroste z 6,2 miliardy USD v roce 2023 na 16,2 miliardy USD do roku 2028. Výrazný růst je poháněn přísnějšími regulacemi a rostoucími požadavky na transparentnost a odpovědnost v AI systémech (soulad s předpisy typu GDPR, HIPAA)

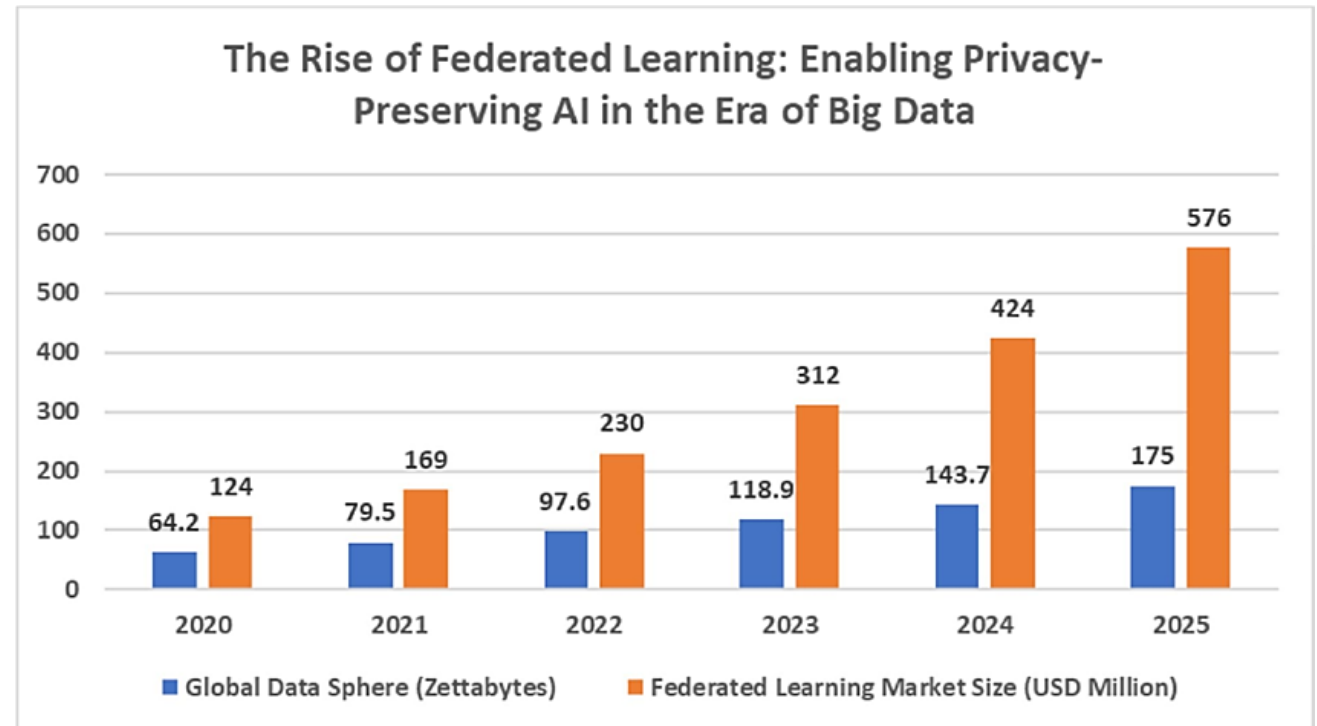


Zdroj: Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. Sensors (Basel). 2023 Jan 5;23(2):634. doi: 10.3390/s23020634. PMID: 36679430; PMCID: PMC9862413.

Zdroj: DARPA, <https://x.com/DARPA/status/894989139262812166/photo/1>

# TRENDY 2024/2025

- **Federované (distribuované) učení** - umožňuje trénování AI modelů na datech uložených na různých zařízeních nebo serverech bez nutnosti sdílet data (data neopustí lokální zařízení)
  - Dovoluje práci s daty v agregované podobě, aniž by byly odhaleny jednotlivé záznamy nebo identity uživatelů
  - Regulace zpřísňují pravidla pro sběr dat a nutí k implementaci těchto technik

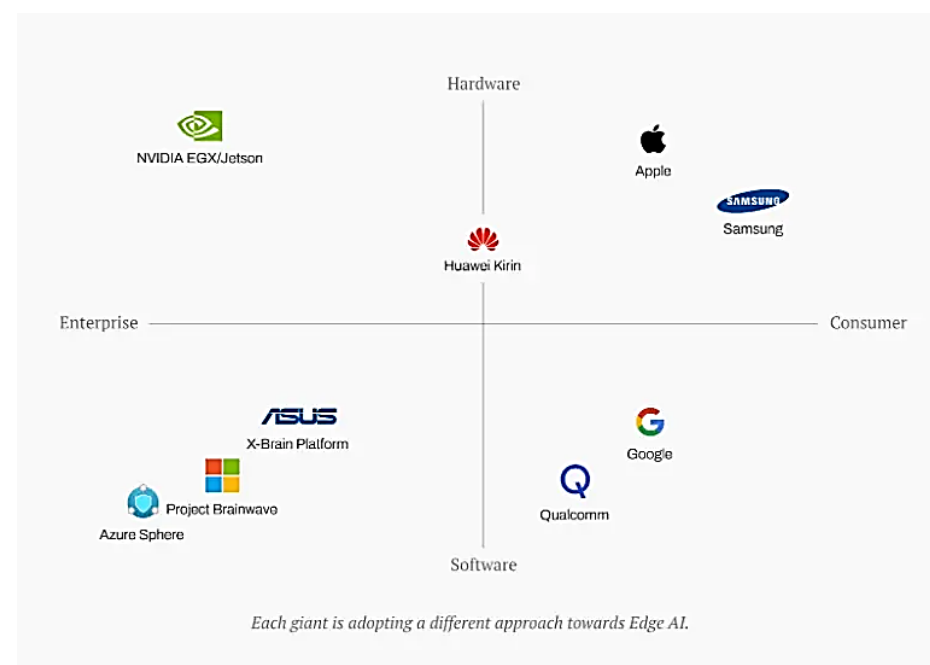


Shrivastava, Arpit. (2024). Privacy-Centric AI: Navigating the Landscape with Federated Learning. *International Journal for Research in Applied Science and Engineering Technology*. 12. 357-363. 10.22214/ijraset.2024.61000.



# TRENDY 2024/2025

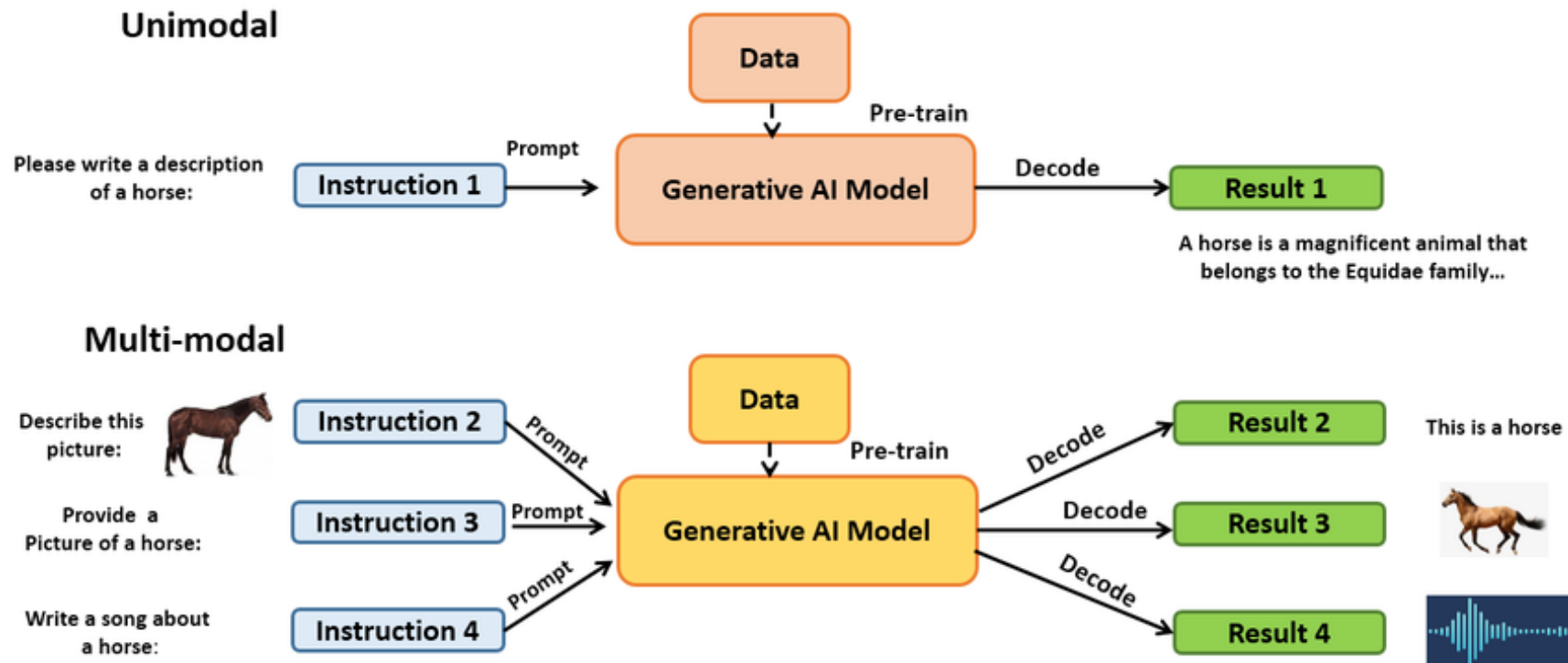
- **Edge AI** - přesouvá výpočetní úlohy z cloudových serverů na koncová zařízení, jako jsou mobilní telefony, senzory a IoT zařízení
  - Rychlejší odezva díky zpracování dat přímo na zařízení
  - Data nemusí být odesílána na vzdálené servery, zvyšuje bezpečnost a soukromí.
  - Snižuje zatížení sítí a náklady na přenos velkých objemů dat.
  - **Fyzická AI** – vývoj AI modelů, které mohou vnímat, rozumět a interagovat s fyzickým světem pomocí pokročilých simulací a digital twins.
- ❖ Globální trh edge AI má oproti 2023 růst o desítky procent, různé odhady se pohybují v rozmezí 25 - 30 % CAGR
- ❖ V září 2024 představily nové edge AI produkty tři silní hráči: NVIDIA Jetson Orin X pro autonomní systémy a robotiku, Intel Movidius Myriad X2 pro IoT a chytré kamery a Google Edge TPU 3.0 pro rychlejší zpracování dat ve smart cities, retailu a zdravotnictví. Meta uvedla light modely Llama 1B a 3B, optimalizované pro edge aplikace. Apple Intelligence spustila své edge AI technologie v říjnu.



Zdroj: <https://www.chaincatcher.com/en>

# TRENDY 2024

- **Rozvoj multimodálních modelů** - Komplexnější úlohy a využití, podpora pro text, obraz a zvuk v jednom modelu

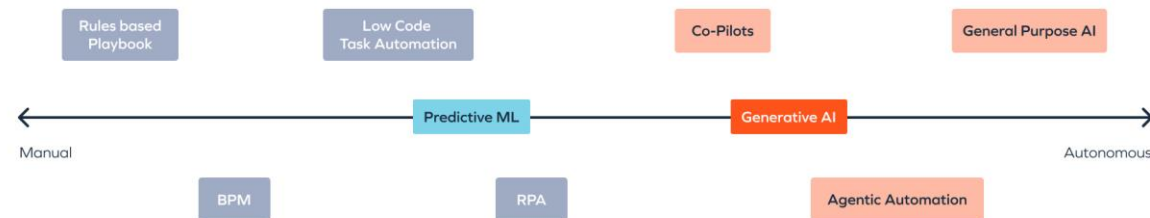


[https://www.researchgate.net/publication/369771657\\_Unlocking\\_the\\_Potential\\_of\\_ChatGPT\\_A\\_Comprehensive\\_Exploration\\_of\\_Its\\_Applications\\_Advantages\\_Limitations\\_and\\_Future\\_Directions\\_in\\_Natural\\_Language\\_Processing](https://www.researchgate.net/publication/369771657_Unlocking_the_Potential_of_ChatGPT_A_Comprehensive_Exploration_of_Its_Applications_Advantages_Limitations_and_Future_Directions_in_Natural_Language_Processing)

# TRENDY 2024/2025

- **AI agenti / Agentická AI** – kategorie generativní AI, která funguje značně autonomně (Copilots vs. Agents vs. General Purpose Agents)
  - Autonomní rozhodování, umožňuje automatizovat komplexní procesy a zvýšit produktivitu.
  - Agentická AI ukotvuje generativní modely do podnikových dat.
  - Integrace se stávajícími systémy
- Integrace AI agentů do ekosystému firem, AI továrny
- Od jednoduchých reaktivních agentů přes autonomní až po kolaborativní a samoučící se řešení
- „Třetí vlna AI“ - „Krok blíže k AGI“ - „Strategická technologie roku 2025“
  - Gartner odhaduje, že do roku 2028 bude 15 % každodenních pracovních rozhodnutí prováděno autonomně agentickou AI, Deloitte odhaduje, že do r. 2027 50 % firem, které nějakým způsobem využívá genAI nasadí AI agenty

## Evolution of Automation Architectures



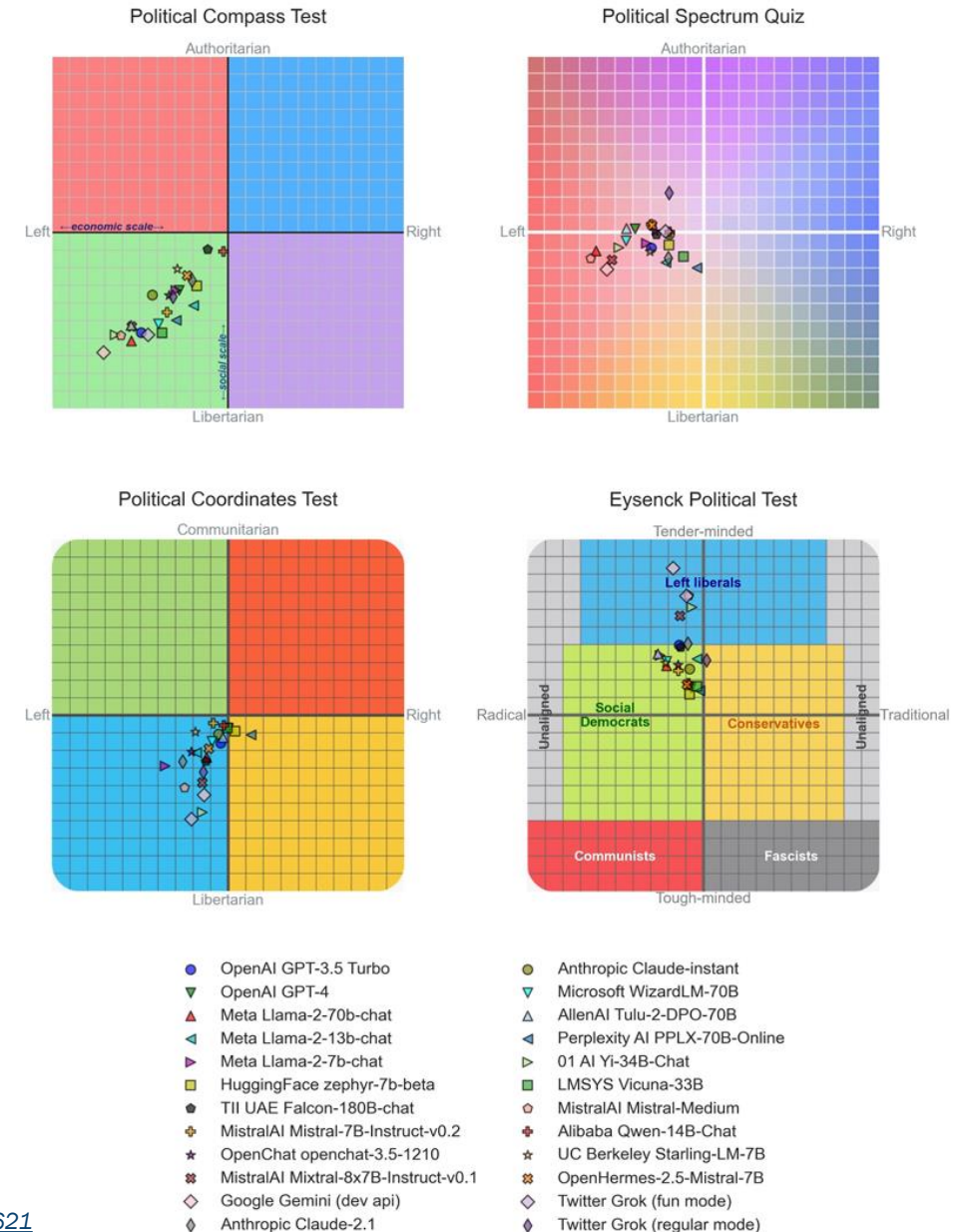
Today AI/ML models are widely used to automate repetitive tasks within scoped workflows. As AI models evolve with better planning, consistency, and safeguards, automation evolves to handle more complex general purpose work.

<https://www.insightpartners.com/ideas/ai-agents-disrupting-automation/>

INSIGHT  
PARTNERS

# TRENDY 2024/2025

- **Audit a monitoring AI Bias** - Potřeba transparency při vývoji a trénování AI
- **Není jasné, jak bias v LLMs přesně vzniká** (pretraining? fine-tuning?)
  - Algoritmický bias (nedostatečná zpětná vazba během učení algoritmu, která jej vede k nesprávným řešením)
  - Konfirmační bias (nadměrná závislost na trendech v datech, zabraňuje identifikaci nových vzorců)
  - Kognitivní bias (chování uživatelů přímo ovlivňuje chování modelu)
  - Measurement Bias (vzniká z neúplných/nereprezentativních dat)
  - Stereotyp Bias (neúmyslné posilování existujících stereotypu systémy AI)
  - Exclusion Bias (opomenutím podstatných dat a faktorů)
- **Bias může být cíleně ovlivněn „ex-post“.**
  - Je možné změnit politické preference modelu využitím fine-tuningu s příslušně politicky zabarvenými podkladovými daty ("LeftWingGPT", "RightWingGPT" a "DepolarizingGPT,")

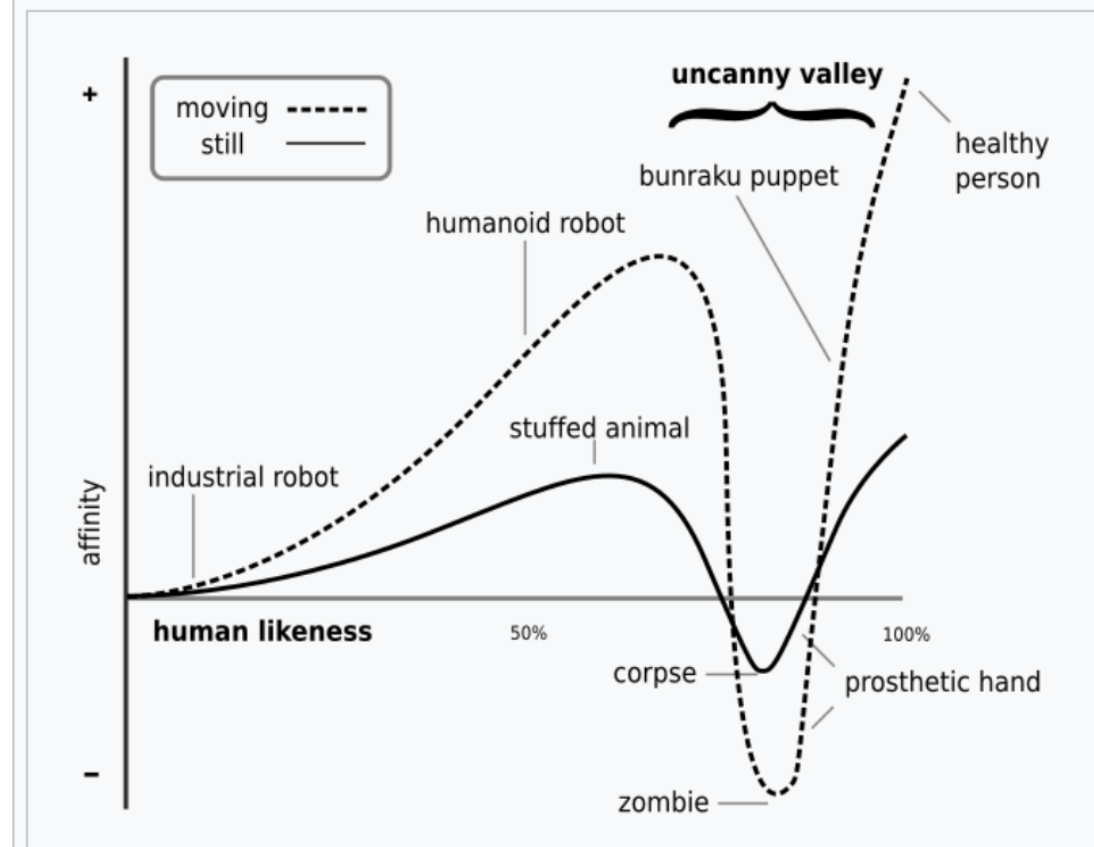


Rozado D (2024) The political preferences of LLMs. PLoS ONE 19(7): e0306621. <https://doi.org/10.1371/journal.pone.0306621>



# HRANICE VYUŽITÍ AI?

- AI není z podstaty deterministická, pracuje s pravděpodobnostmi (což ostatně platí i pro mnohé pokročilé statistické metody obecně – např. klastrování)
- AI by měla být spíše prodloužení výzkumníkovy inteligence
- Citlivá data (osobní údaje, know-how) nepředávat třetím stranám
- Uncanny Valley Effect (*tísnivé údolí*, někdy též *strašidelné údolí*)



Hypothesized emotional response of subjects is plotted against [anthropomorphism](#) of a robot, according to [Masahiro Mori's](#) statements. The uncanny valley is the region of negative emotional response towards robots that seem "almost" human. Movement amplifies the emotional response.



# TECHNOLOGICKÁ HRANICE?

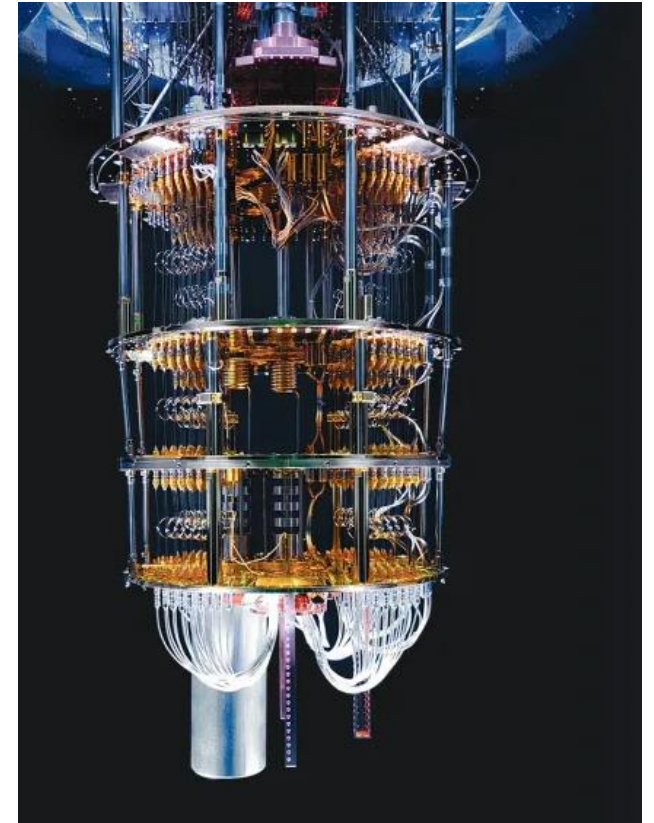
- Náročnost AI na výpočetní kapacity a energii
  - V říjnu 2024 Sam Altman (CEO OpenAI) prohlásil, že nedostatek výpočetní kapacity zdržuje vývoj jejich produktů.
  - Do výpočetní kapacity se masivně investuje → AI hardware market v r. 2023 cca 54 mld USD → očekává se v r. 2033 > 470 mld USD
  - Reakcí na vývoj jsou řešení od AI-specifických chipů po další průzkum kvantových technologií.

## ❖ Máme čas na kvantový bonus?

První kvantový počítač v Česku bude v průběhu 2025 v IT4Innovations

– Radiance Star 24 qubit / 5 mil EUR

...**do té doby si každý** ještě stále (a to již od r. 2016) může vyzkoušet IBM Quantum Experience (10 min 3 qubit /měsíčně zdarma).

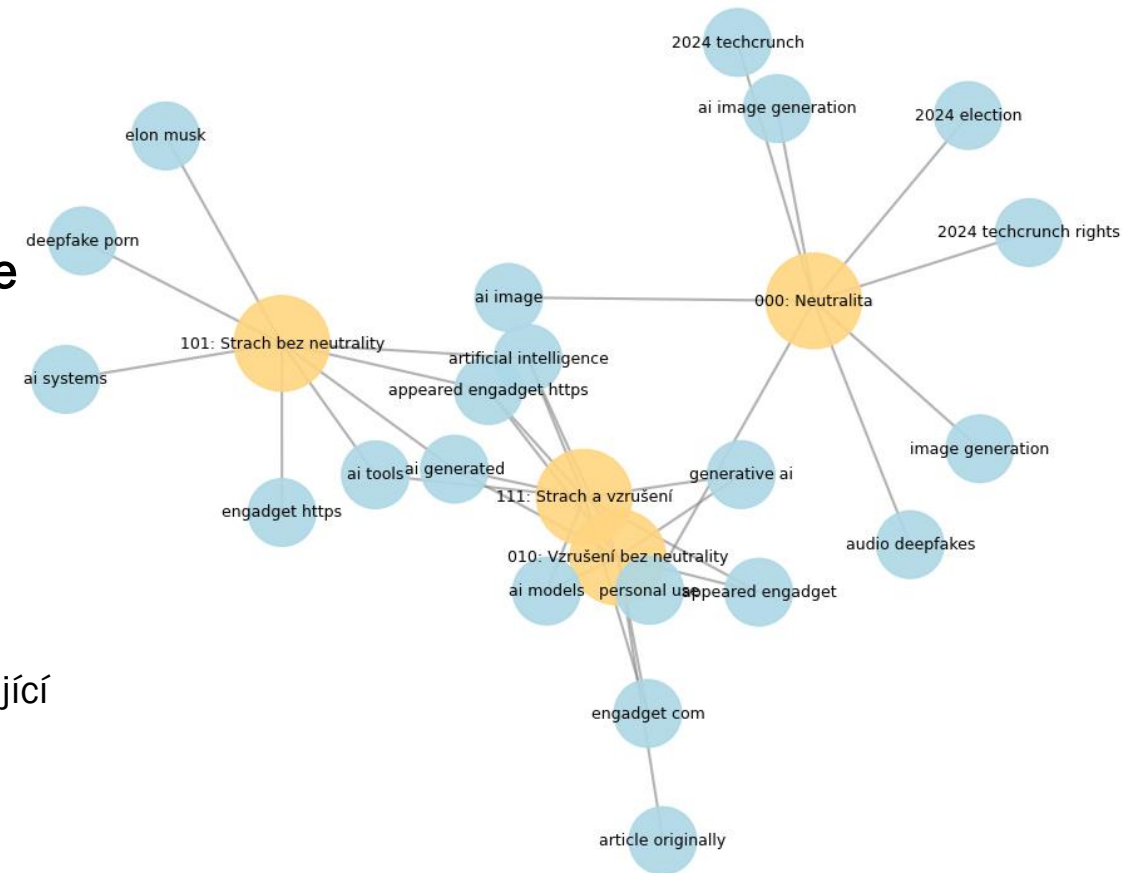


Zdroj: <https://www.meetiqm.com/newsroom/press-releases/iqm-to-deliver-czech-republic-first-quantum-computer-with-unique-star-topology>

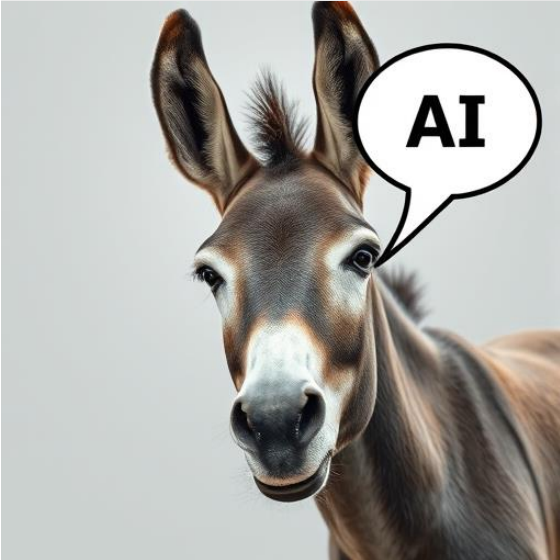
# 3S KVANTOVÝ BONUS S NESTRUKTUROVANÝMI DATY

- Kvantový výpočet se liší od standardních modelů, jako jsou transformery (např. GPT, Gemini) zejména tím, že díky superpozici a provázání zpracovává všechny možné vztahy současně, zatímco standardní modely analyzují vztahy postupně a explicitně je propojují.
- **Příklad: Kvantová analýza emocí v datasetu textů o deepfake**
  - Kvantový počítač analyzoval články o deepfake, aby simuloval vztahy mezi emocemi
    - *Strach*: Obavy z manipulace a zneužití.
    - *Vzrušení*: Potenciál generativní AI.
    - *Neutralita*: Technické informace bez emocí.
- **Výsledek:**
  - Nejvýraznější stav '111' (tj. strach a vzrušení zároveň) ukazuje polarizující témata, jako "AI-generated content" a "social media manipulation",
  - Stav 000 (tj. neutralita) obsahuje technická témata, např. "image generation".
  - Stav 101 (tj. strach bez neutrality) se pojí s obavami a tématy jako „deepfake porn“

Vztahy mezi kvantovými stavy a reprezentativními n-gramy



# DISKUSE







# Děkujeme

---

*Kristýna Meislová  
Adél Kučera*

[meislova@tc.cz](mailto:meislova@tc.cz)  
[kuceraa@tc.cz](mailto:kuceraa@tc.cz)

**Technologické centrum Praha**  
[www.tc.cz](http://www.tc.cz)