

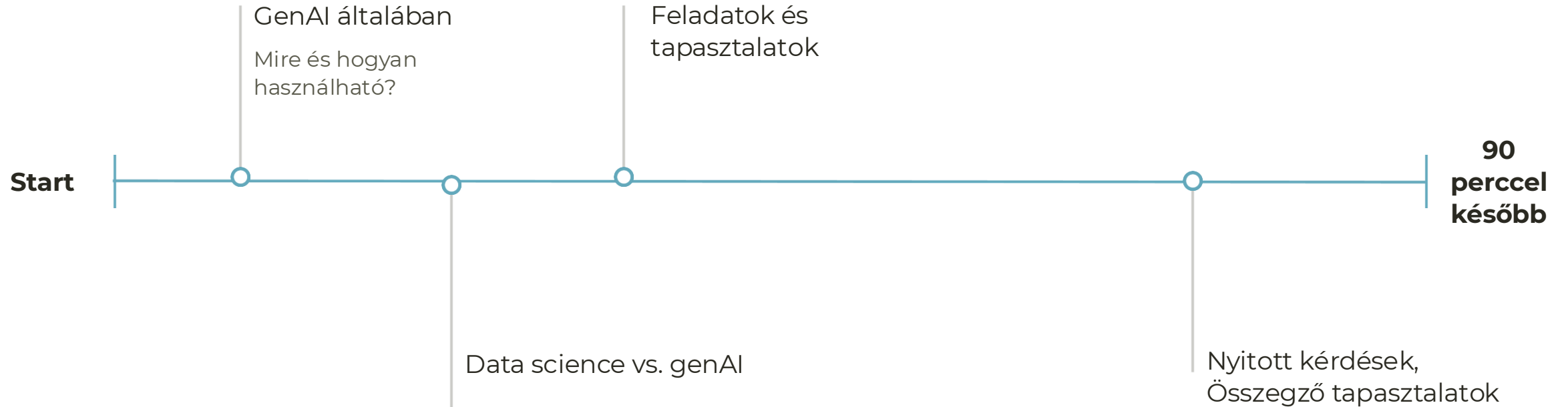
Adattudományi és adatbányászati alkalmazások

Feladatok, buktatók, megoldások

Csaba GÁSPÁR
BME-TMIT / dmlab

2025.03.26. – Prompt Engineering tárgy

Előadásom menetrendje



A photograph of a space shuttle launching from a launchpad. The shuttle is ascending vertically, leaving a massive, billowing plume of white and orange smoke and fire. The launchpad structure is visible in the foreground, and the sky is a mix of blue and white clouds. Overlaid on the center of the image is the text "CHAT GPT - 2022. OKTÓBER 30." in a bold, white, sans-serif font.

CHAT GPT - 2022. OKTÓBER 30.



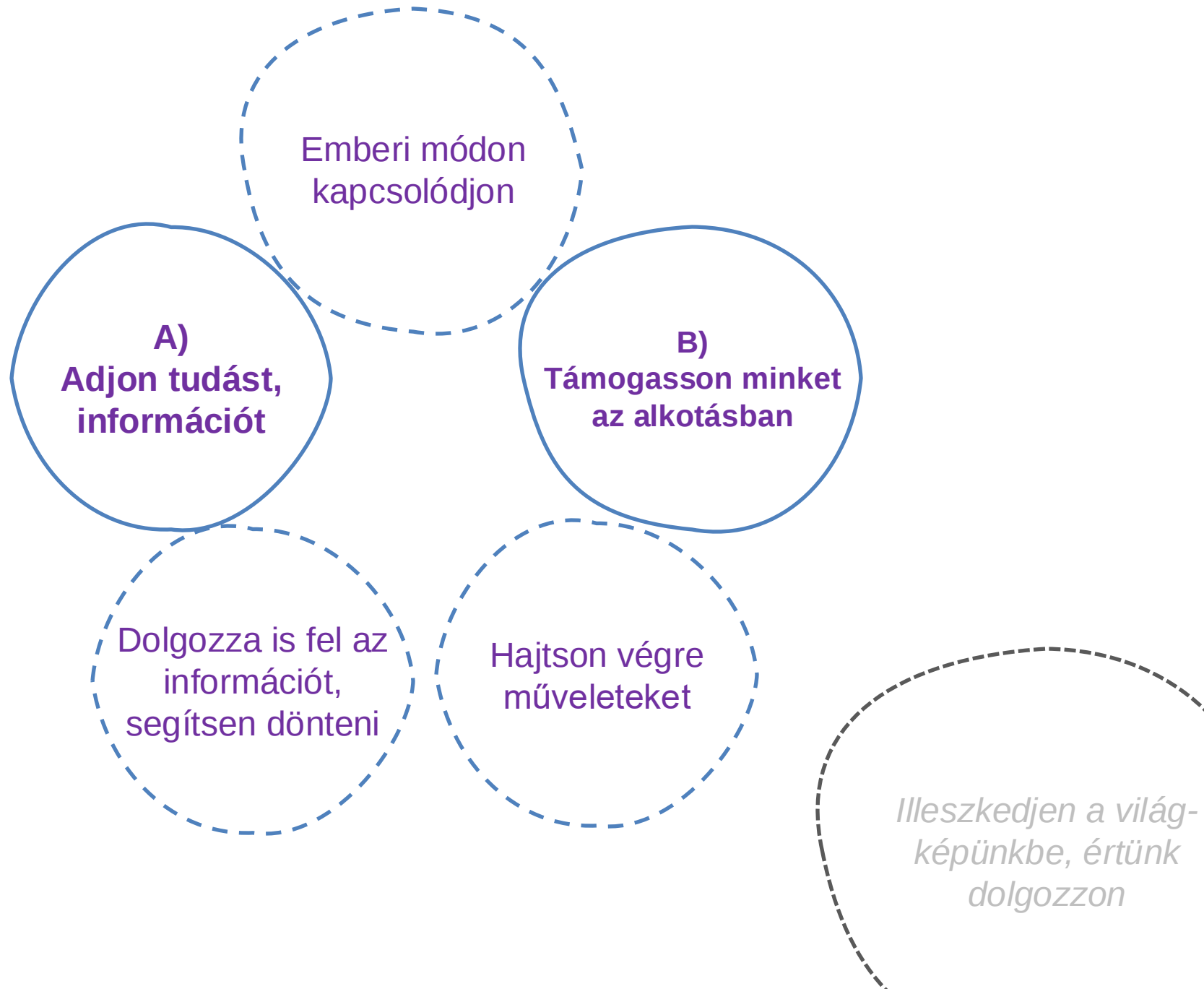
SPANYOLOK



VIKINGEK

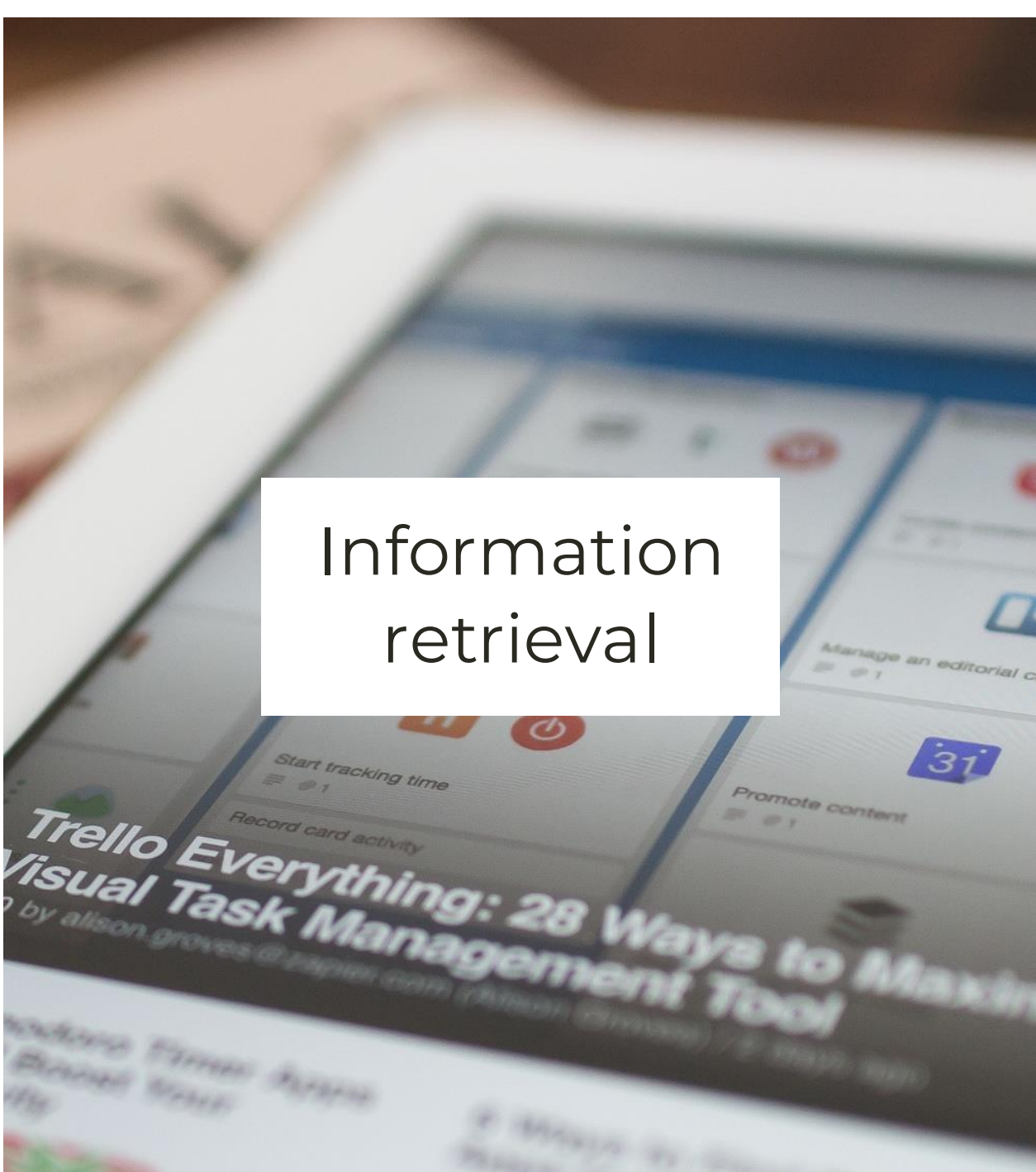


Elvárásaink az új generatív AI megoldásoktól



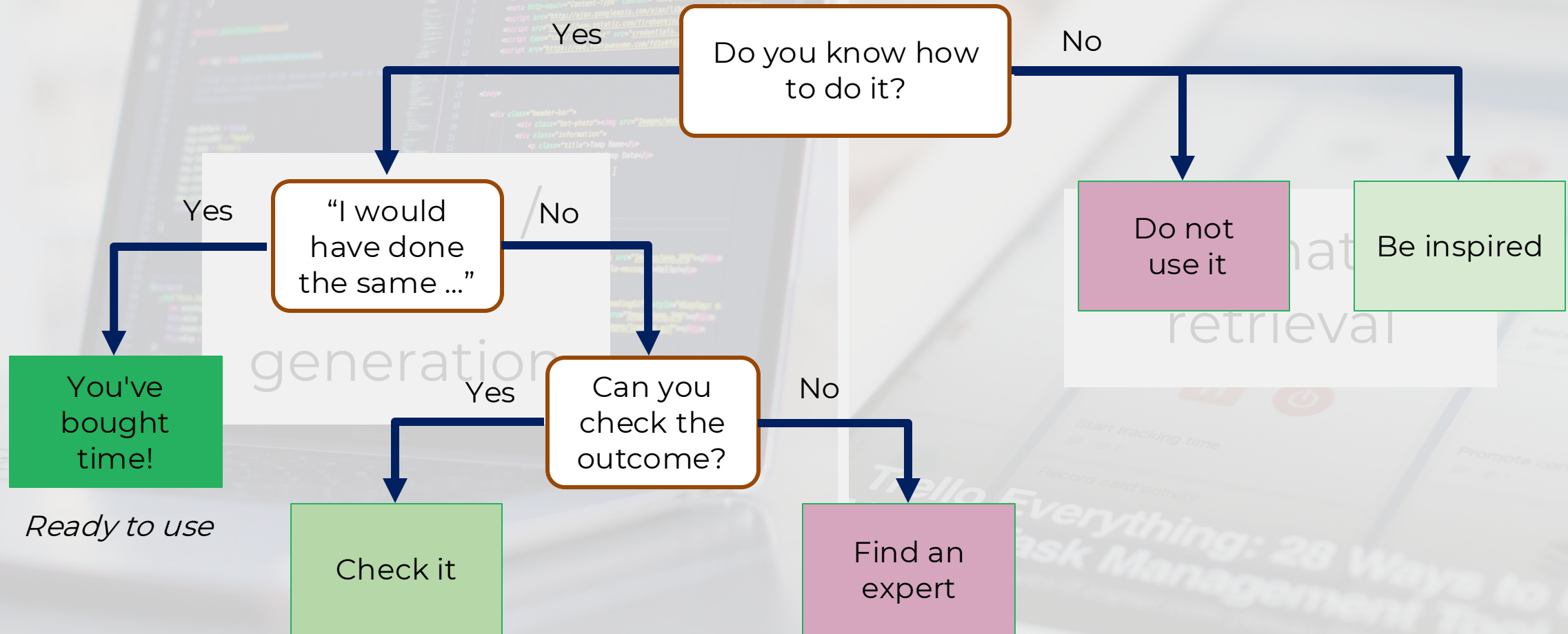
A close-up photograph of a laptop screen. The screen shows a code editor with a dark background and colorful syntax highlighting. The code appears to be HTML and JavaScript. A white rectangular box is overlaid on the lower-left portion of the screen, containing the text 'Content / Code generation'.

Content /
Code
generation

A close-up photograph of a tablet screen. The screen displays a Trello interface with various cards and lists. A white rectangular box is overlaid on the upper-right portion of the screen, containing the text 'Information retrieval'.

Information
retrieval

When to use genAI?

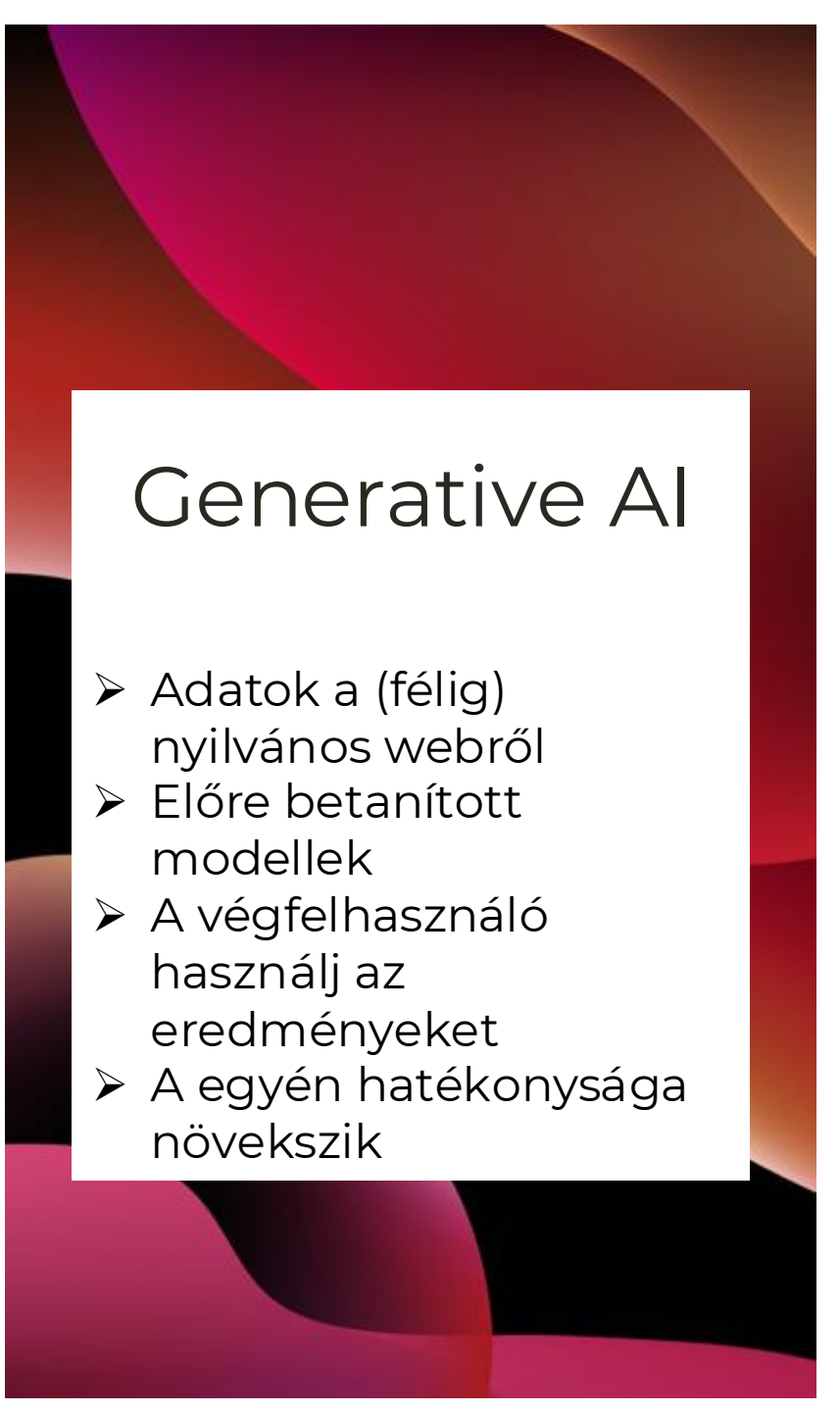




Data science and machine learning

- Céges adatokon
- Modellek betanítása fontos feladat
- Cég felhasználja az eredményeket
- A cég hatékonysága növekszik

AI vs DS



Generative AI

- Adatok a (félig) nyilvános webről
- Előre betanított modellek
- A végfelhasználó használja az eredményeket
- A egyén hatékonysága növekszik

Artificial Intelligence

Is the field of study

Machine Learning

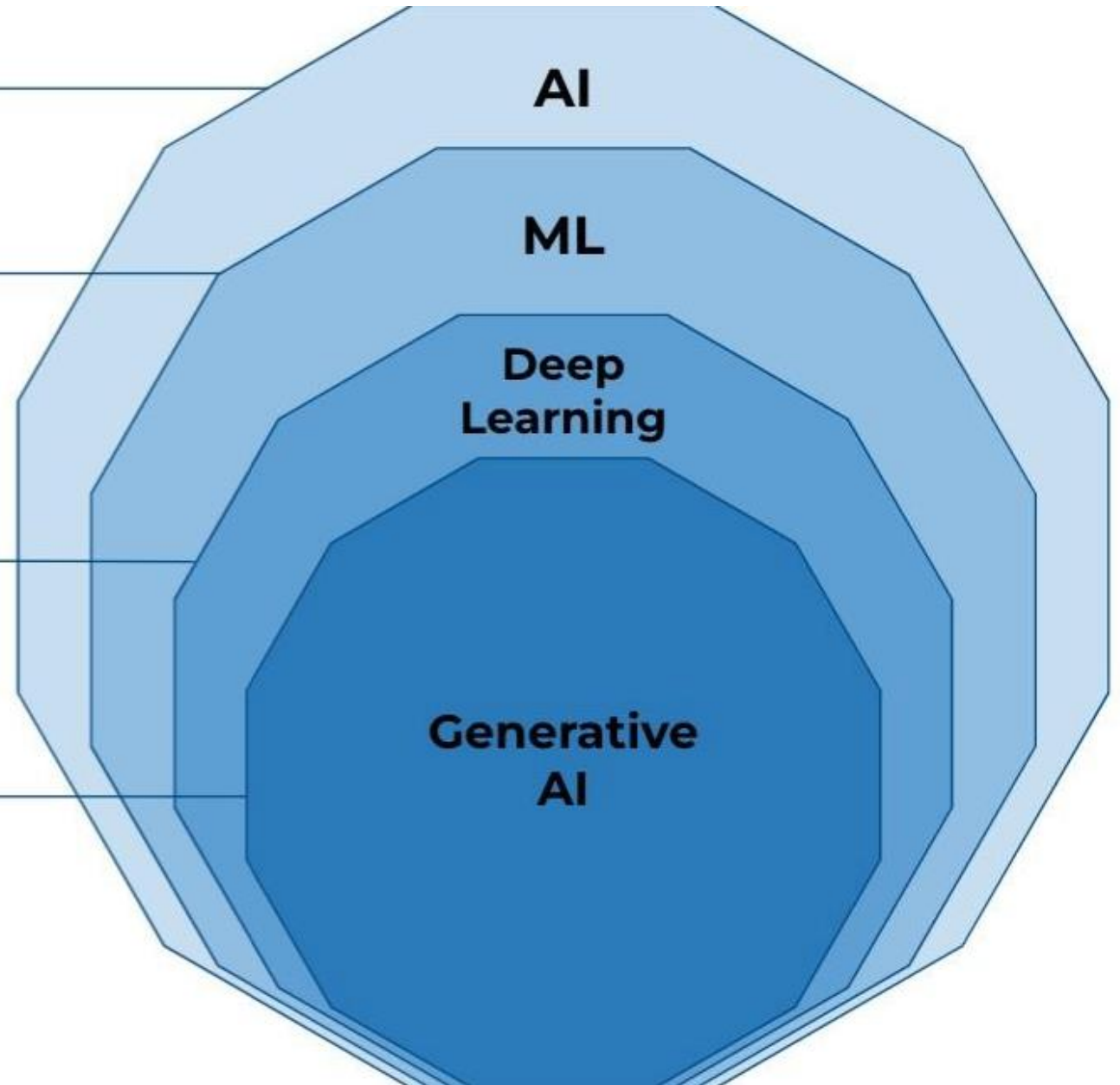
Is a branch of AI that focus on the creation of intelligent machines that learn from data. Another very well know branch inside AI is **Optimization**.

Deep Learning

Is a subset of Machine Learning methods, based on **Artificial Neural Networks**.
Examples: CNNs, RNNs

Generative AI

A type of ANNs that generate data that is similar to the data it was trained on.
Examples: GANs, LLMs



1.) Kitalálja-e mit kell csinálni?



Name of Ship		TITANIC		Date of Departure	Apr 10 1912		191	Where bound	NEW YORK								
Part of Departure		SOUTHAMPTON		Steamship Line		WHITE STAR LINE											
NAMES AND DESCRIPTIONS OF BRITISH PASSENGERS EMBARKED AT THE PORT OF SOUTHAMPTON																	
Passenger Number	NAME OF PASSENGER	SEX	AGE	Part of whole Passage here mentioned to Land	RELATION	CLASS OF PASSENGER							Country of Birth	Country of Residence			
						First	Second	Third	Fourth	Fifth	Sixth	Seventh			Eighth		
1881/1	James Allen	M	34		THIRD CLASS	NEW YORK	5800005	1625							USA		
1881/2	Alfred Harris	M	34														
1881/3	Robert Rams	M	34														
1881/4	William Rams	M	34														
1881/5	John Rams	M	34														
1881/6	John Rams	M	34														
1881/7	John Rams	M	34														
1881/8	John Rams	M	34														
1881/9	John Rams	M	34														
1881/10	John Rams	M	34														
1881/11	John Rams	M	34														
1881/12	John Rams	M	34														
1881/13	John Rams	M	34														
1881/14	John Rams	M	34														
1881/15	John Rams	M	34														
1881/16	John Rams	M	34														
1881/17	John Rams	M	34														
1881/18	John Rams	M	34														
1881/19	John Rams	M	34														
1881/20	John Rams	M	34														
1881/21	John Rams	M	34														
1881/22	John Rams	M	34														
1881/23	John Rams	M	34														
1881/24	John Rams	M	34														
1881/25	John Rams	M	34														
1881/26	John Rams	M	34														
1881/27	John Rams	M	34														
1881/28	John Rams	M	34														
1881/29	John Rams	M	34														
1881/30	John Rams	M	34														
1881/31	John Rams	M	34														
1881/32	John Rams	M	34														
1881/33	John Rams	M	34														
1881/34	John Rams	M	34														
1881/35	John Rams	M	34														
1881/36	John Rams	M	34														
1881/37	John Rams	M	34														
1881/38	John Rams	M	34														
1881/39	John Rams	M	34														
1881/40	John Rams	M	34														
1881/41	John Rams	M	34														
1881/42	John Rams	M	34														
1881/43	John Rams	M	34														
1881/44	John Rams	M	34														
1881/45	John Rams	M	34														
1881/46	John Rams	M	34														
1881/47	John Rams	M	34														
1881/48	John Rams	M	34														
1881/49	John Rams	M	34														
1881/50	John Rams	M	34														
1881/51	John Rams	M	34														
1881/52	John Rams	M	34														
1881/53	John Rams	M	34														
1881/54	John Rams	M	34														
1881/55	John Rams	M	34														
1881/56	John Rams	M	34														
1881/57	John Rams	M	34														
1881/58	John Rams	M	34														
1881/59	John Rams	M	34														
1881/60	John Rams	M	34														
1881/61	John Rams	M	34														
1881/62	John Rams	M	34														
1881/63	John Rams	M	34														
1881/64	John Rams	M	34														
1881/65	John Rams	M	34														
1881/66	John Rams	M	34														
1881/67	John Rams	M	34														
1881/68	John Rams	M	34														
1881/69	John Rams	M	34														
1881/70	John Rams	M	34														
1881/71	John Rams	M	34														
1881/72	John Rams	M	34														
1881/73	John Rams	M	34														
1881/74	John Rams	M	34														
1881/75	John Rams	M	34														
1881/76	John Rams	M	34														
1881/77	John Rams	M	34														
1881/78	John Rams	M	34														
1881/79	John Rams	M	34														
1881/80	John Rams	M	34														
1881/81	John Rams	M	34														
1881/82	John Rams	M	34														
1881/83	John Rams	M	34														
1881/84	John Rams	M	34														
1881/85	John Rams	M	34														
1881/86	John Rams	M	34														
1881/87	John Rams	M	34														
1881/88	John Rams	M	34														
1881/89	John Rams	M	34														
1881/90	John Rams	M	34														
1881/91	John Rams	M	34														
1881/92	John Rams	M	34														
1881/93	John Rams	M	34														
1881/94	John Rams	M	34														
1881/95	John Rams	M	34														
1881/96	John Rams	M	34														
1881/97	John Rams	M	34														
1881/98	John Rams	M	34														
1881/99	John Rams	M	34														
1881/100	John Rams	M	34														

Train				
	PassengerId	Survived	Pclass	Name
1	1	0	3	Braund, Mr. Owen Harris
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)
3	3	1	3	Heikkinen, Miss. Laina
4	4	1	1	Futrelle, Mrs. Jacques Heath

Mellékelem a Titanic adathalmazt, végezz rajta részletes adatelőkészítést, majd hozz létre rajta egy hatékony osztályozási modellt, ami nagy megbízhatósággal előrejelzi, ki éli túl a katasztrófát.

Azonnal kezdem az adatelőkészítést. [↗]

Az adatok előkészítése megtörtént, és a tanító- és tesztadalmazok is létrejöttek. A következő lépésben egy osztályozási modellt hozok létre, amely előrejelzi, ki élte túl a katasztrófát. Elsőként egy Random Forest modellt alkalmazok, és kiértékelem a teljesítményét. [↗]

A Random Forest modell teljesítményének eredményei:

- **Pontosság (Accuracy):** 80.45%
- **Klasszifikációs riport:**
 - 0 (Nem élte túl): Precízió: 82%, Recall: 88%, F1-érték: 85%
 - 1 (Túlélte): Precízió: 78%, Recall: 68%, F1-érték: 73%
- **Konfúziós mátrix:**
 - Helyes besorolás: 97 (nem élte túl), 47 (túlélte)
 - Hibás besorolás: 13 (nem élte túlként), 22 (túlélteként)

Ha szeretnéd, finomhangolhatjuk a modellt, más algoritmusokat próbálhatunk ki, vagy készíthetünk további elemzéseket. [↗]



Message ChatGPT



1.) Kitalálja-e mit kell csinálni?

Látszólag jól megoldja a feladatot...

DE

- Jól ismert, publikus adathalmaz
- Tutorial gondolatmenet – hibás

“Valaki tudja, valaki tanítja” effektus

1.) Kitalálja-e mit kell csinálni?

- Vertex AI (Google technológia) a data science versenyeken
- ChatGPT mint data scientist: Műegyetemi órák, mint feladatok
- GitHub Copilot: ügyesen támogat – de melyik környezetben



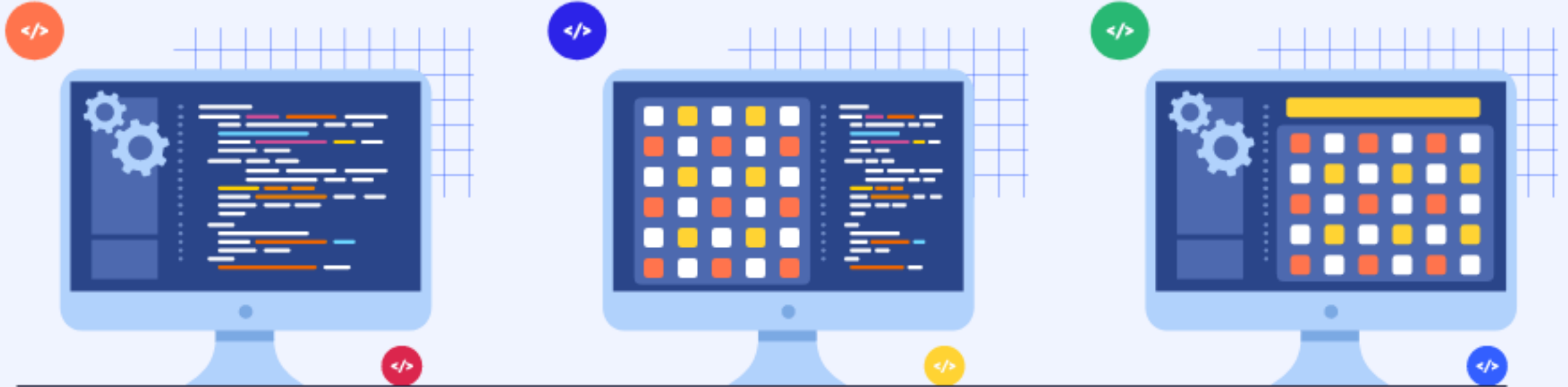
Konklúzió:

Junior már használhatja, de őmaga nincs junior szinten

Custom Coding

Low Code

No Code



SPSS
SAS
(S)
...
R
....
Python (+pandas+sklearn)
Spark(Java/Python/Scala/R)

Programnyelv
integráció

AutoML

BigData / Scale up
megoldások

SAS Enterprise Miner / Guide
IBM SPSS Modeler
(Oracle Data Miner)
Altair RapidMiner
KNime

2.) Beszélgessünk az adatunkkal



- ChatGPT 4o – Filefeltöltés + python futtatás
- Technológia építhető rá – de miért nincs?
- Demo: Melyik országból jön az adat?

Hogyan fejlesztünk
csináljunk adhoc
riportokat chatGPTvel?



Beszélgessünk
az adatokról...

Tranzakciós adataim vannak a Tesco-tól.
Vajon melyik országból származik?

	customer_id	visit_date	visit_spend	daynum
0	2	2010-04-01	5.97	0
1	2	2010-04-06	12.71	5
2	2	2010-04-07	34.52	6
3	2	2010-04-12	7.89	11
4	2	2010-04-14	17.17	13
5	...	...	...	...
6	7376	2011-05-15	4.86	409
7	7376	2011-05-24	56.55	418
8	7376	2011-05-26	60.25	420
9	7376	2011-06-09	159.15	434
10	7376	2011-06-19	5.98	444

Szeretném megtudni, hány ember adata van az adathalmazban. És az adathalmaznak hány sora van.

1359 tok

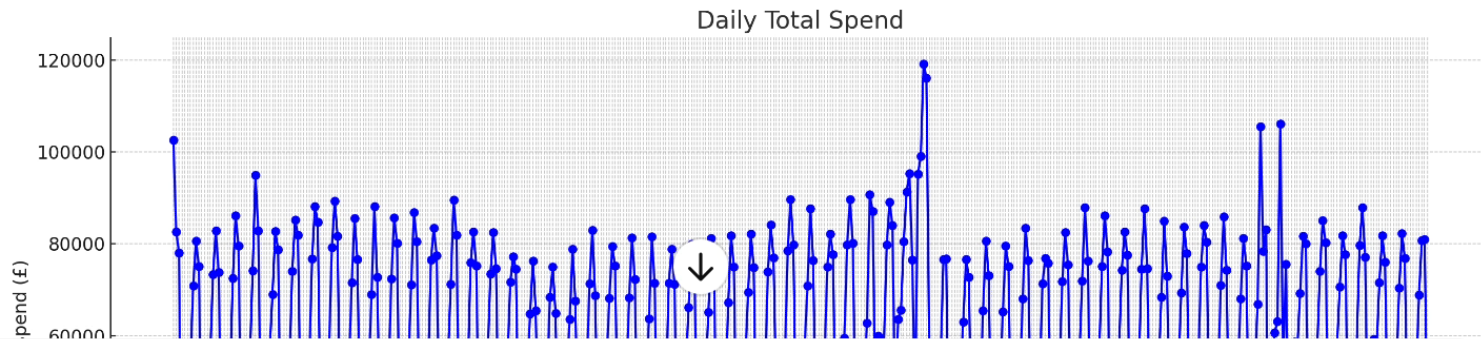


Az adathalmazban összesen 596,733 sor van. A szerkezete alapján a következő oszlopokat tartalmazza: ``customer_id``, ``visit_date``, ``visit_spend``, és ``daynum``.

Most megnézem, hogy hány különböző vásárló szerepel az adatok között: [↗]

Az adathalmazban összesen 5,000 különböző vásárló van szerepelve. Ha bármi további kérdésed vagy elemzési igényed van az adatokkal kapcsolatban, szívesen segítek! [↗]

Kérlek számold ki egy táblázatba, hogy melyik nap összesen mennyit költöttek az emberek. Majd kérlek ábrázold ezt a diagrammot egy linechart segítségével.



Message ChatGPT



2.) Beszélgessünk az adatunkkal



- ChatGPT 4o – Filefeltöltés + python futtatás
- Technológia építhető rá – de miért nincs?
- Demo: Melyik országból jön az adat?

Konklúzió:

Technológián és szemléletváltás egyszerre szükséges

Three main patterns of use



Agent

Who I talk to, and ask for knowledge, and ask for different actions



Co-author

Who helps in the creation/ generation of content



Procedure

Who solves a task from time to time, independently, without supervision

Insurance business examples

3.) Struktúrált kimenet GPT modellekből

```
from openai import OpenAI
from pydantic import BaseModel, Field
from typing import List

import instructor

class Character(BaseModel):
    name: str
    age: int
    fact: List[str] = Field(..., description="A list of facts about the character")

# enables `response_model` in create call
client = instructor.from_openai(
    OpenAI(
```

```
{
  "name": "Christopher Columbus",
  "age": 30,
  "fact": [
    "He was an Italian explorer and adventurer who is credited with leading the first European expedition to the New World.",
    "Columbus made four voyages across the Atlantic Ocean, sponsored by King Ferdinand II of Aragon and Queen Isabella I of Castile.",
    "In 1492, Columbus landed on the island of Guanahani in the Bahamas, which he named San Salvador.",
    "During his exploration, Columbus encountered various indigenous peoples, including the Taino, Arawak, and Carib.",
    "Columbus made contact with the ruler of the island of Cuba, the Caonabo, but was unable to establish a peace treaty due to his own cultural misunderstandings."
  ]
}
```

```
response_model=Character,
)
print(resp.model_dump_json(indent=2))
```

Konklúzió:

- GPT modellekből strukturált adat
- API vagy lokálisan futtatható GPT modellek (*ollama* – saját gépen futó GPT)

4.) One-shot osztályozás

One-shot classification:

- Előre tanított modellből külön tanítás nélkül
- Példánkban: Férfi vagy női (magyar) név
- Modellek: ChatGPT, ollama



4. terület

Férfi és női nevek szétválasztása



```
from langchain_ollama import ChatOllama
```

```
llm = ChatOllama(  
    model="llama3.2",  
    temperature=0,  
)
```

```
def get_type3(x):  
    messages = [  
        (  
            "system",  
            "Egy magyar nyelvész vagy, aki férfi és női keresztnéveket kap. \  
            Add meg, hogy az adott név férfi, női, esetleg mindkettő nemhez tartozik. \  
            A kimeneteden csak a nő, férfi, mindkettő, egyiksem szavak lehetnek",  
        ),  
        ("human", f"{x}"),  
    ]  
    ai_msg = llm.invoke(messages)  
    valasz=ai_msg.content.lower()  
    v=valasz.split(" ")[0]  
    if v.lower()=="női":  
        v="nő"  
    if v.lower() in ['nő', "férfi", "mindkettő", "egyiksem"]:  
        return v.lower()  
    else:  
        print(x, "->", valasz)  
        return "none"
```

Adatok:

Hazai top 60 női és 60 férfi név

Ollama:

Accuracy: 92%

ChatGPT API

Accuracy: 93%

100.000 lekérdezés: 23 nap
6.000 lekérdezés: 1.38 nap
120 lekérdezés: 40 perc

4.) One-shot osztályozás

One-shot classification:

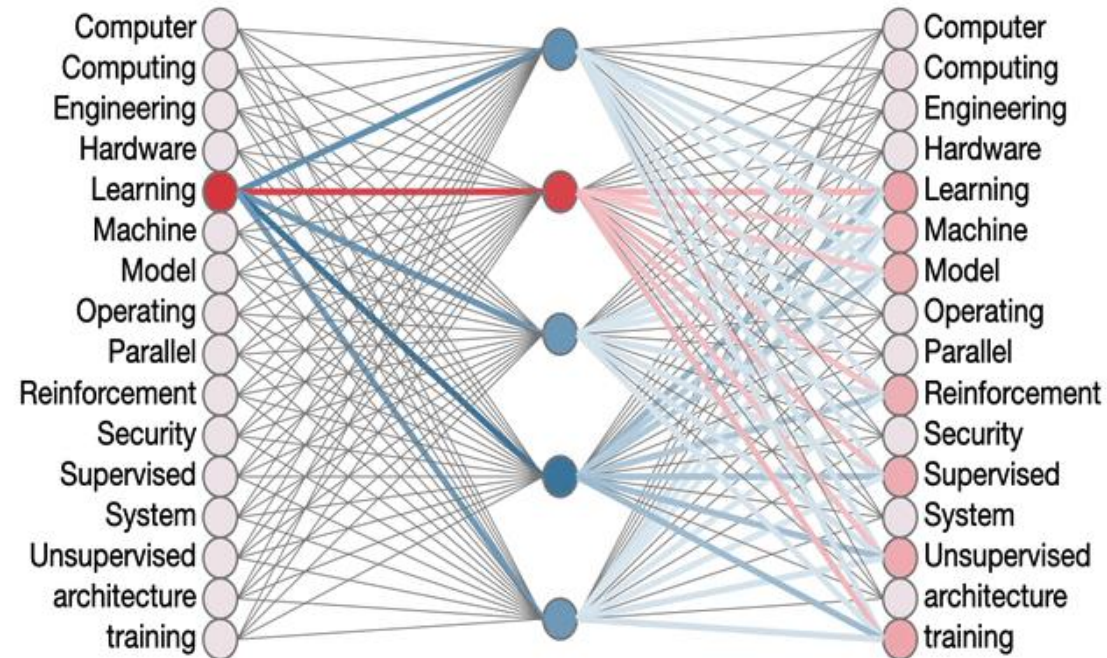
- Előre tanított modellből külön tanítás nélkül
- Példánkban: Férfi vagy női (magyar) név
- Modellek: ChatGPT, ollama

Tapasztalatok

- Végrehajtási idő gondok
- Public web content bias: BBC hírek
bekategorizálása



5.) Embedding, mint transzformáció



5.) Embedding, mint transzformáció

```
from langchain_ollama import OllamaEmbeddings

embeddings_model = OllamaEmbeddings(model="llama3.2")
```

```
embeddings = embeddings_model.embed_documents(
    list(df['Name'].values)
)
```

```
logreg = LogisticRegression(random_state=42, solver="saga", penalty="l1", max_iter=5000, C=100)
logreg = LogisticRegression()
```

```
df['pred'] = cross_val_predict(logreg, embeddings, df['sex_F'], cv=2, method='predict_proba')[:,1]
```

```
from sklearn.metrics import roc_auc_score
from sklearn.metrics import accuracy_score
```

```
roc_auc_score(df['sex_F'], df['pred'])
```

```
0.8597427476737821
```

```
accuracy_score(df['sex_F'], df['pred'] > 0.5)
```

```
0.7520661157024794
```

Adatok:

Hazai top 60 női és 60 férfi név

Ollama:

75%-ban eltalálja
(ROC AUC: 0.86)

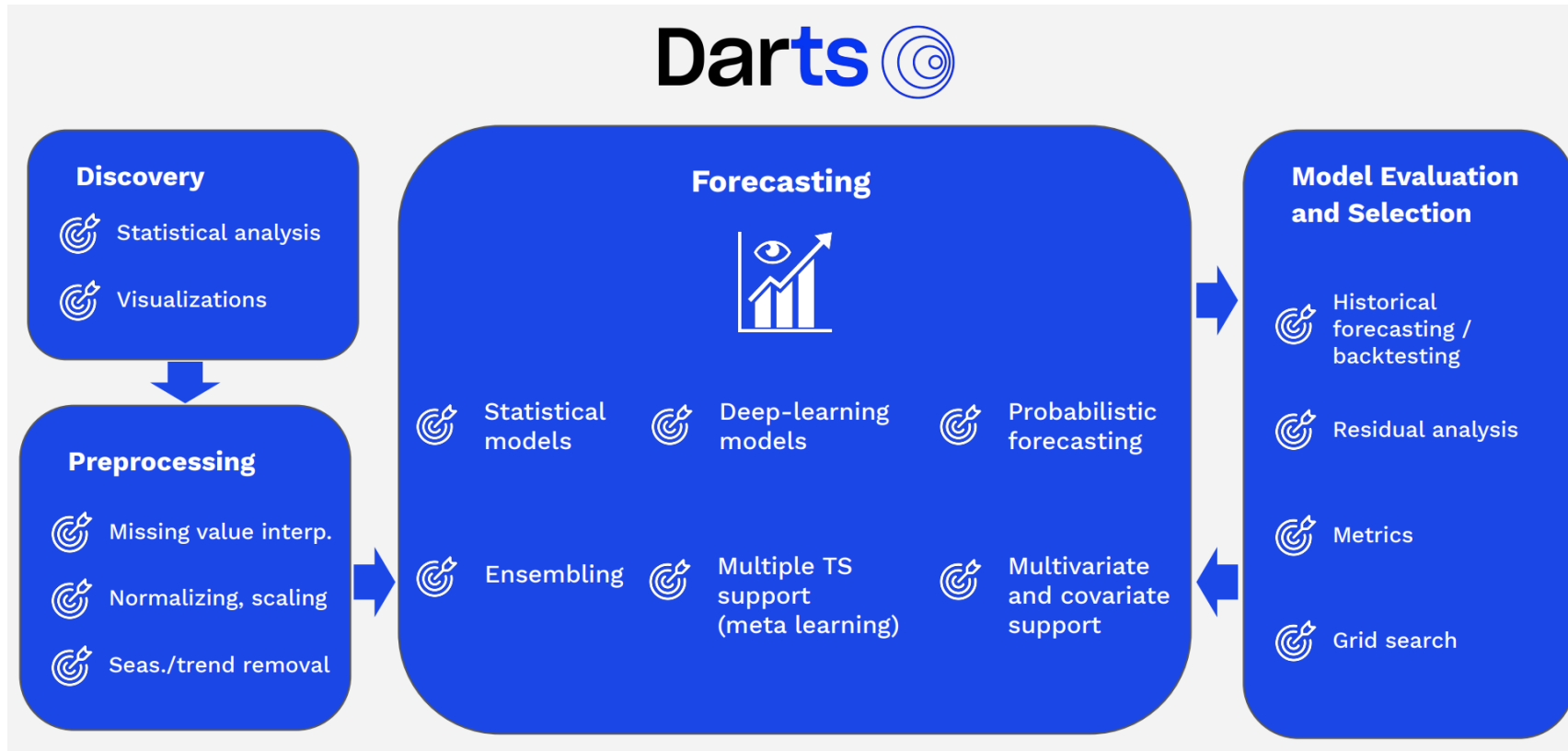
ChatGPT API

86%-ban eltalálja
(ROC AUC: 0.92)

Különleges esetek:

- Nara
- Zente
- Luca

6.) Idősor előrejelzés

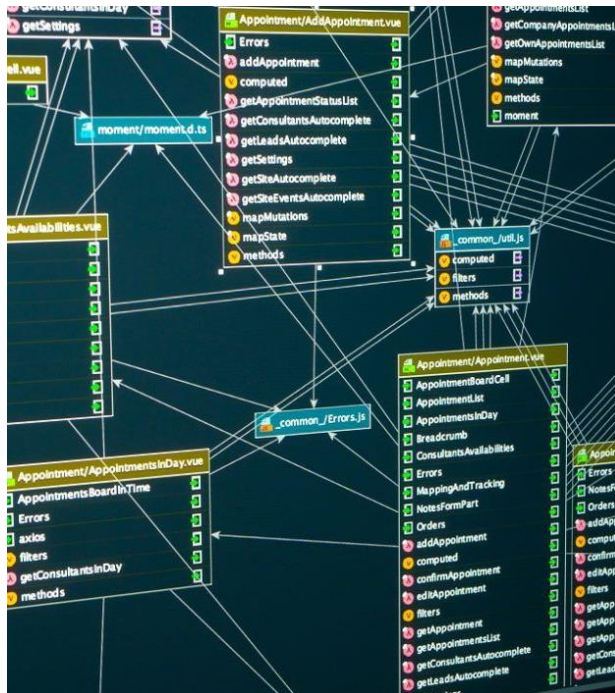


Adat:

Hazai energiatőzsde adatok

Tapasztalatok:

- Végre! Tényleg jobb, mint a klasszikus
- Transformer modellek
- Recurrens neuron hálók
- Előtanított neuronhálók



IT rész-
területek
előnyben

Adatokba még
nem lát bele

Leírt
gondolatmenet
ritka – hogyan
tanuljon?

Nem jobb kell,
hanem új
Ez egy romboló
innováció

Thank you for attention!



Csaba GÁSPÁR
Data scientist
gaspar@tmit.bme.hu
[+36 \(20\) 823 4154](tel:+36208234154)