

Machine Learning Kedvcsináló Kezdőknek



Készítette: Kovács Gyula
Budapest ML Fórum 2022

HELLO!

Kovács Gyula

**Data Science személyi edző
(coach)**

www.kovacsgyulacoach.hu



lancreakcio.clementine.hu/

Láncreakció

27. Humánus-e a személyzetis az adatbányában? 0:00 / 45:39

DATA SCIENCE PODCAST

ITT IS ELÉRED

EPIZODOK

- 27 Humánus-e a személyzetis az adatbányában? **LEJÁTSZÁS**
- 26 Média-értés adat-alapon, vagy miért podcastolunk? **LEJÁTSZÁS**
- 25 Válogatáskazetták a Spotify korából **LEJÁTSZÁS**
- 24 Néha az adatelemzők is személyzetre viszik a fájó fogú beteget **LEJÁTSZÁS**
- 23 Mikortól gyűlölik meg a matekot a gyerekek? **LEJÁTSZÁS**

TÖVÉBB A TÖRTEHEZ

Tartalom

Bemutakozás

1-3 oldal



01

4-32 oldal

Fogalmak tisztázása

- + *Mi a Data Science és mi a gépi tanulás?*
- + *Miért van szükség tanuló algoritmusokra*
- + *Felügyelt és felügyelet nélküli algoritmusok*
- + *CRISP, AutoML, MLOps*

02

32-51 oldal

Neurális hálók, NLP, üzleti alkalmazások

- + *Deep Learning, avagy neurális hálók elmélete*
- + *Miért népszerű a Deep Learning*
- + *NLP fejlődése, avagy a neurális hálók alkalmazása a szövegelemzésben*
- + *GPT3 és AlphaCode*

Fogalmak tisztázása

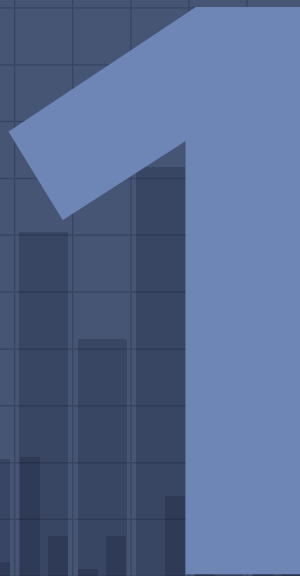
Miért van szükség tanuló algoritmusokra

Mi a Data Science és mi a gépi tanulás?

Felügyelt és felügyelet nélküli algoritmusok

Lehet rasszista egy algoritmus?

CRISP, AutoML, MLOps



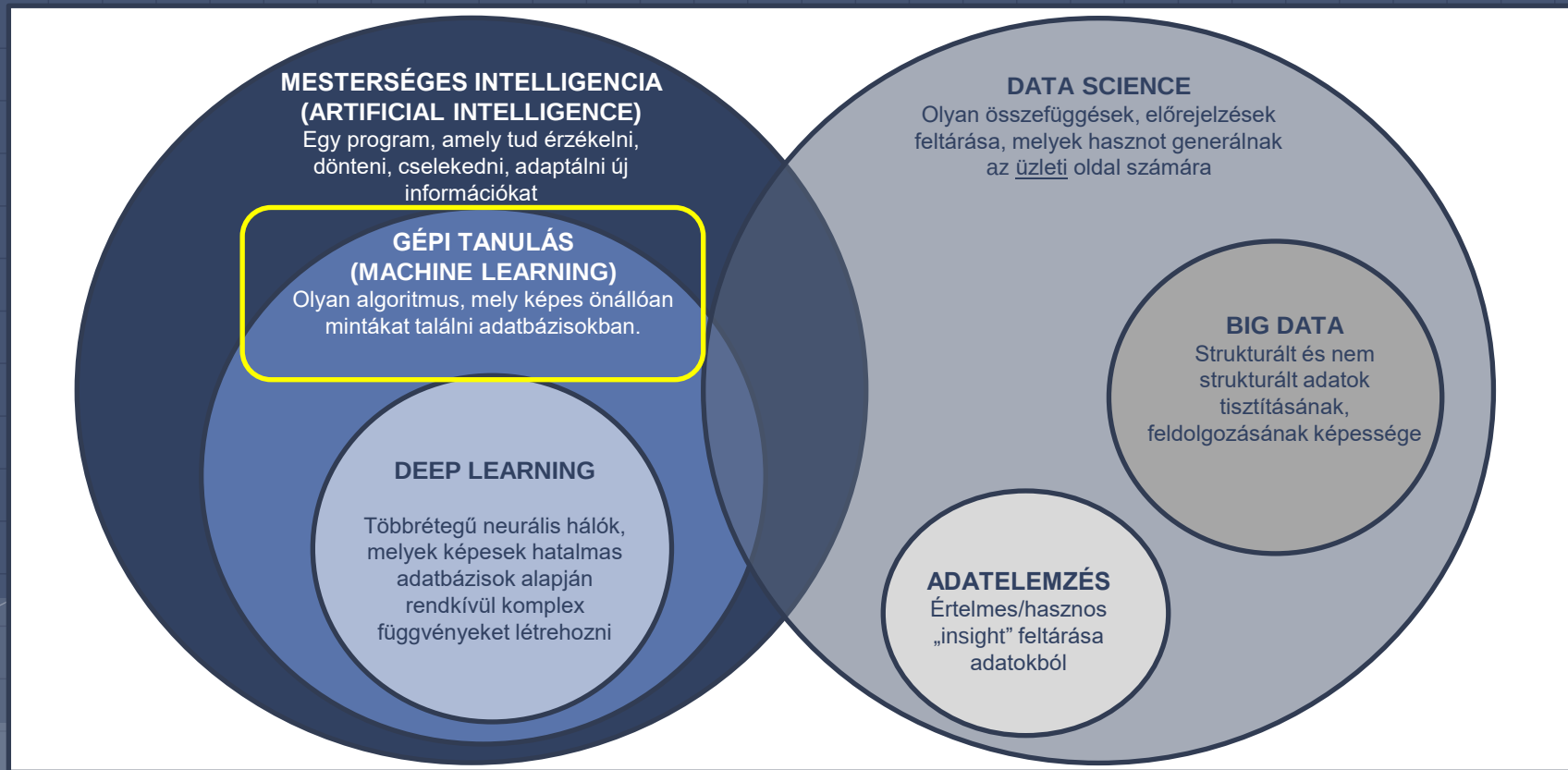
Miért van szükség gépi tanuló algoritmusokra?

- Nagyon ritkák az ilyen egyértelmű szabályok, de ha vannak, akkor ezeket a konkurencia is megtalálja
- Az adatbázisok mérete napjainkban „hatalmas” – rengeteg oszlop és rekord, lehetetlen minden szabályt felfedezni
- Az emberi agy lineáris összefüggéseket képes felismerni – a nagy adatbázisokban azonban nemcsak „lineáris” összefüggések vannak

Gépi tanuló algoritmusok*

A gépi tanulás a mesterséges intelligencia egyik ága, olyan rendszerekkel foglalkoznak, melyek tanulni képesek. A gyakorlatban ez azt jelenti, hogy a rendszer példa adatok, minták alapján képes önállóan (vagy emberi segítséggel) szabályszerűségeket/szabályokat felismerni/meghatározni. A rendszer tehát nem csupán betanulja „kívülről” a mintákat, hanem **képes ezek alapján olyan általánosításra**, ami a tanulási szakasz végeztével ismeretlen adatokra vonatkozólag is „helyes” döntéseket tud hozni.

Fogalmak tisztázása







a tanító adatbázison ellenőrizni tudjuk, hogy a modell mennyire pontosan találja el a célváltozó értékét

Felügyelt és felügyelet nélküli algoritmusok

Felügyelt
algoritmusok

horizontális

Input változók (x1, x2, ...)					y	Y = f (x1, x2, ...)	P(Y)
Hazai átlagosan hány gólt lőtt	Vendégcsapat átlagosan hány gólt lőtt	Hazai csapat formája	Vendégcsapa formája	Formák különbsége	Haza csapat győzött?	Modell output	Tipp valószínűsége
1,0	1,2	8	4	4	1	1	53,5%
0,6	2,0	4	8	-4	0	0	70,2%
1,4	0,6	8	6	2	0	0	54,4%
0,6	1,6	3	4	-1	1	0	54,4%
1,0	1,0	4	9	-5	0	0	70,2%
2,8	1,4	10	6	4	0	1	53,5%
0,8	2,2	7	7	0	0	0	54,4%

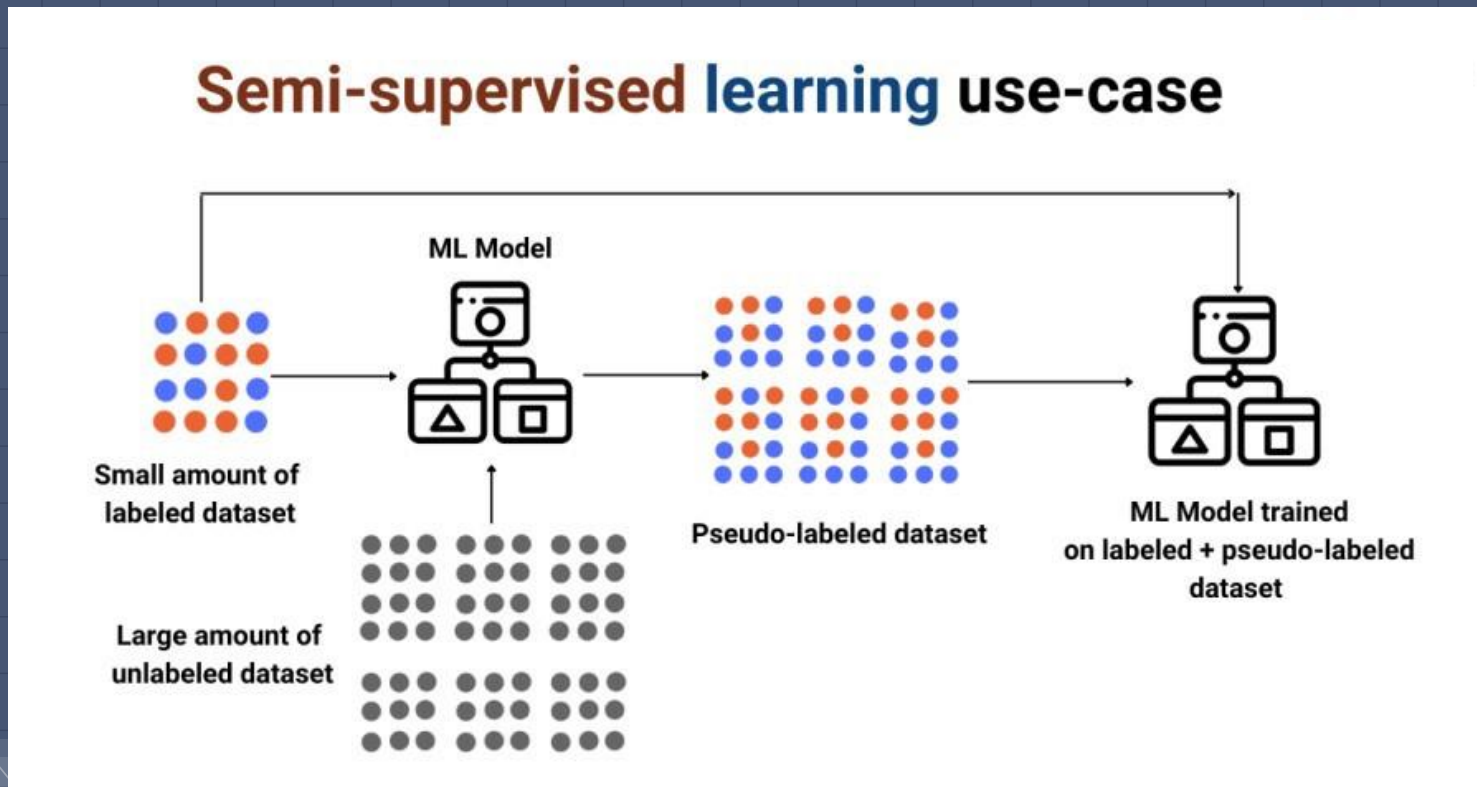
Felügyelet nélküli
algoritmusok

Input változók (x1, x2, ...)					Y = f (x1, x2, ...)
Hazai átlagosan hány gólt lőtt	Vendégcsapat átlagosan hány gólt lőtt	Hazai csapat formája	Vendégcsapa formája	Formák különbsége	Modell output
1,0	1,2	8	4	4	cluster-1
1,0	1,0	7	7	0	cluster-2
0,8	3,2	1	13	-12	cluster-3
2,0	2,4	7	13	-6	cluster-3
0,6	2,0	4	8	-4	cluster-3
1,6	0,8	10	7	3	cluster-1
1,4	0,6	8	6	2	cluster-1

vertikális

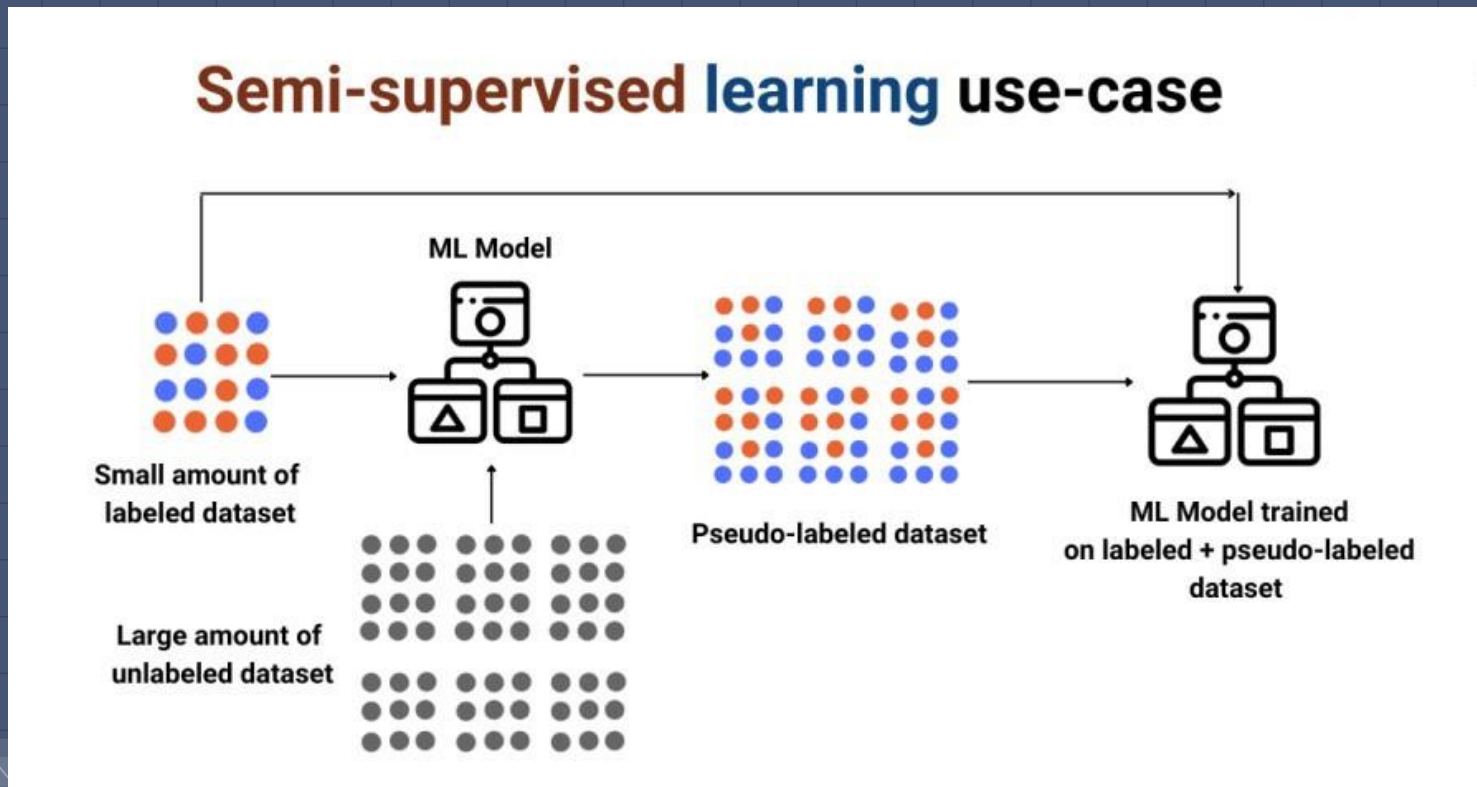
Semi-supervised tanulás – amikor részben címkézett a tanuló adatbázis

12



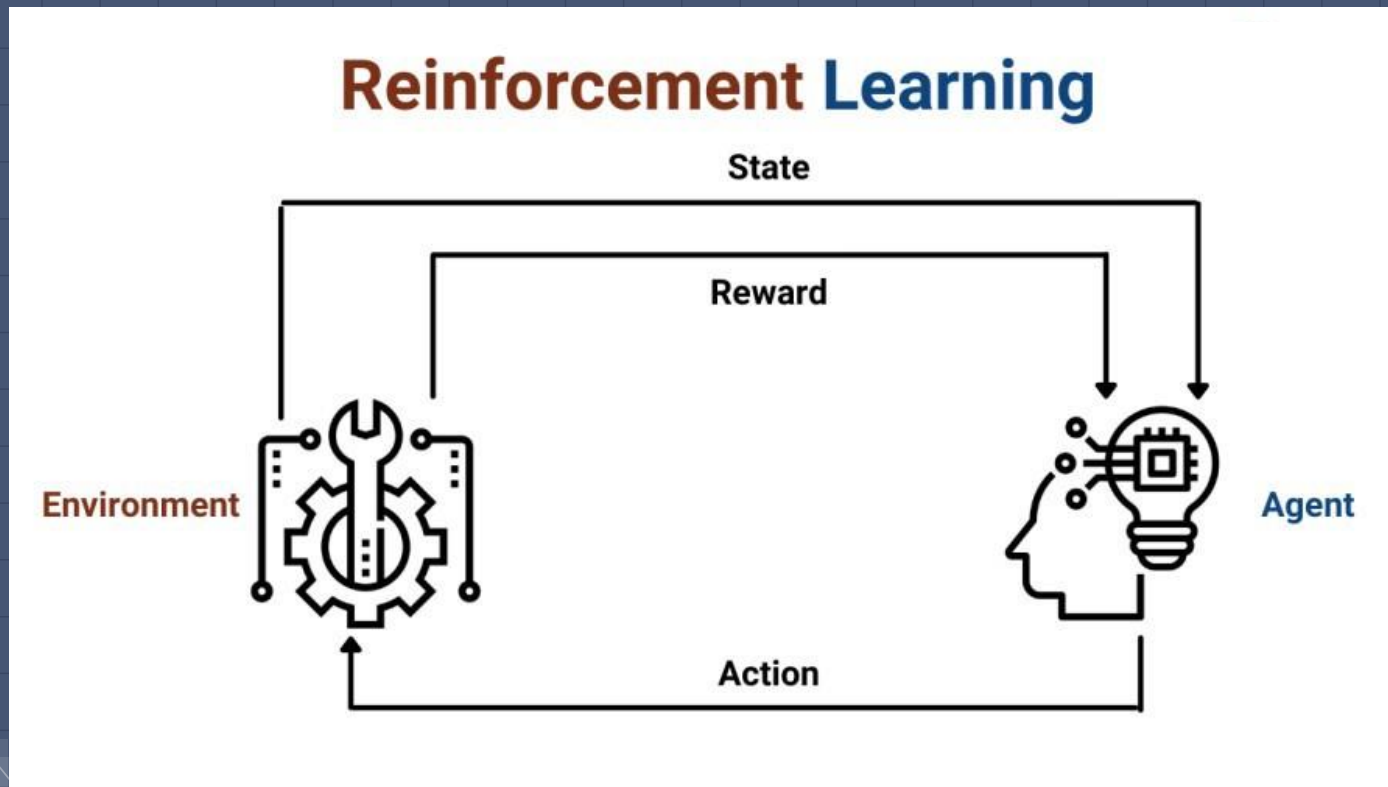
Semi-supervised tanulás – amikor részben címkézett a tanuló adatbázis

13

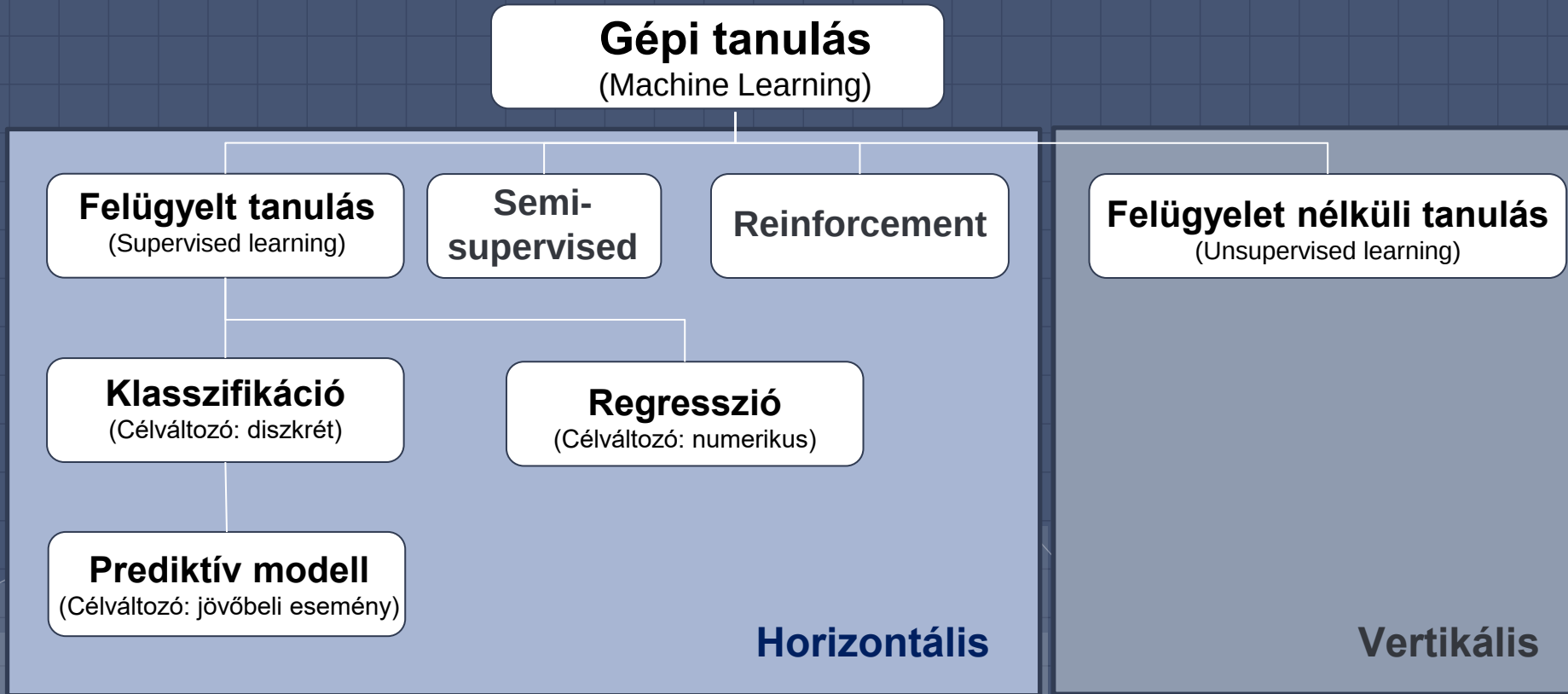


Reinforcement (megerősítésses) tanulás – amikor nem lehet címkézni a tanító adatbázist

14



Gépi tanuló algoritmusok felosztása



Felügyelt tanító algoritmusok

Illesztéses modellek

- (Lineáris regresszió)
- Logisztikus regresszió
- Support Vector Machine

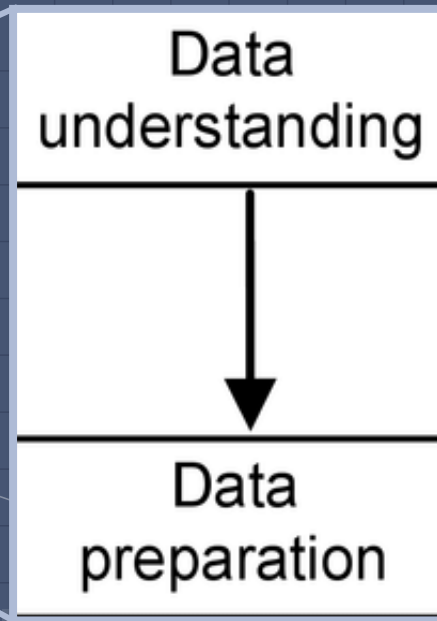
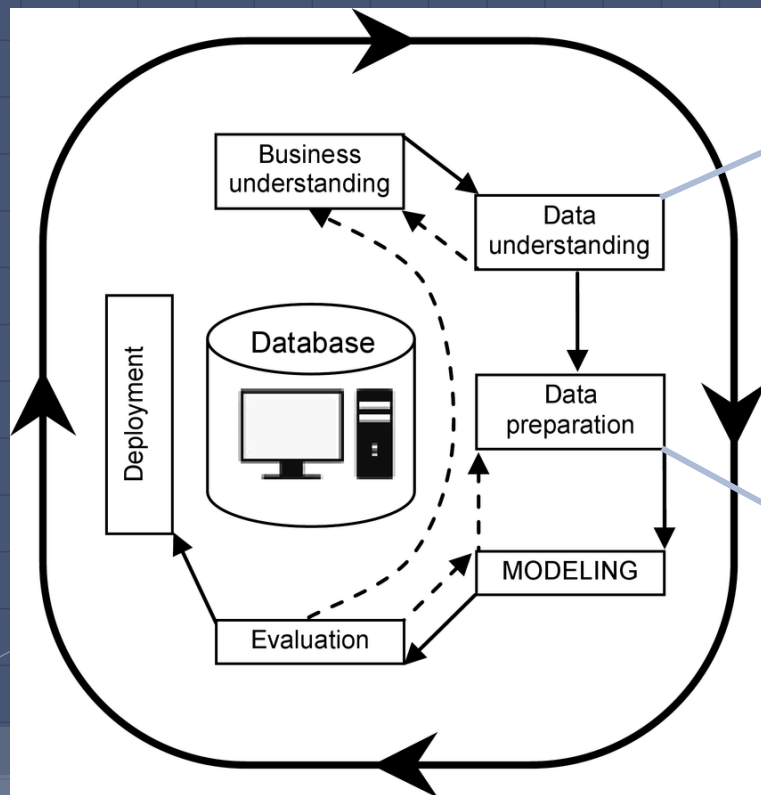
Rekurzív modellek

- Döntési fa algoritmusok (ID3, C5.0, Chaid, CART)
- Véletlen erdők (Random forest)
- GBM modellek

Neurális hálók

- Perceptron modell
- Gradient descent
- Backpropagation
- Deep neural network: RBM, NPLM, CNN

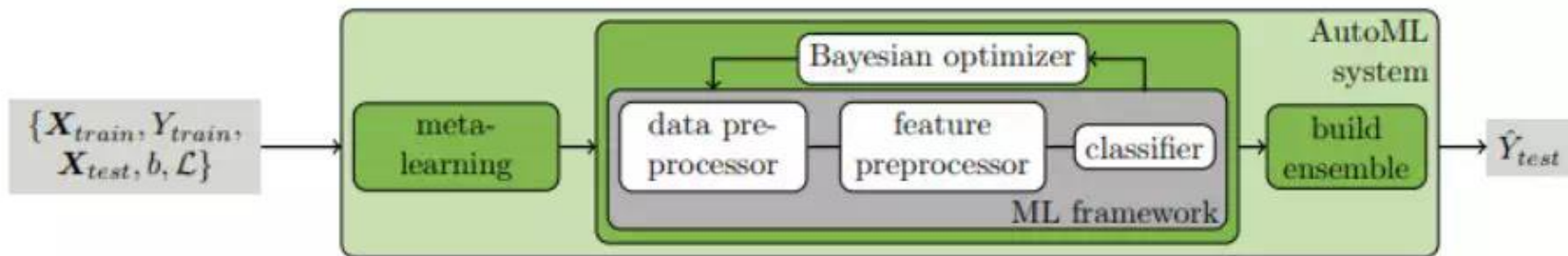
CRISP-DM módszertan – „boldog békeidők”



80-90%

Mi az AutoML?

Az adatelemzés és modellezés folyamatának (pl. CRISP) teljeskörű (vagy majdnem teljeskörű) automatizálása.



Auto-sklearn pipeline

Legnépszerűbb AutoML keretrendszerek

- **Auto-SKLearn** - *Auto-SKLearn egy AutoML szoftver csomag scikit-learn könyvtárban (egy kulcsfontosságú könyvtár a Python programozási nyelvhez, amelyet általában a gépi tanulási projektekben használnak)*
- **MLBox** - *MLBox egy Python könyvtár*
- **TPOT** - *genetikai algoritmust használ az automatizáláshoz*
- **H2O AutoML** – *„in-memory” ML platform mind az R mind a Python nyelvet támogatja*
- ...

MLOps (Machine Learning Operations)

- **Mi a DevOps?** A DevOps a fejlesztői és az üzemeltetői teamek a hagyományosnál szorosabb együttműködését jelenti, ebben a struktúrában a fejlesztői team a ciklus elejétől eszközökkel támogatja az üzemeltetést (deploy scriptek, automatikus diagnosztikai eszközök, terhelés-, és performanciatesztelési eszközök) mely a ciklus előtt, közben, és után is hasznos visszajelzéseket küld.
- **Mi az MLOps?** Az MLOps – gépi tanulási műveletek vagy gépi tanuláshoz készült DevOps – lehetővé teszi az adattudományi és az informatikai csapatok együttműködését, hogy így felgyorsítsák a modellfejlesztést és üzembe helyezést gépi tanulási modellek figyelésével, ellenőrzésével és irányításával.

Mit kell tudnia egy MLOps eszköznek?

- Adatmenedzsment
- Modellezés és design
- Kifejlesztett ML modell alkalmazása (integrálása IT környezetbe) és karbantartása
- End-to-end életciklus menedzsment
- Projekt menedzsment

Miről volt szó az első részben?

- Miért van szükség tanuló algoritmusokra
- Mi a Data Science, és mi a gépi tanulás?
- Felügyelt és felügyelet nélküli algoritmusok
- Lehet-e rasszista egy algoritmus?
- CRISP, AutoML, MLOps – gépi tanulás automatizálása

Neurális hálók, NLP, üzleti alkalmazások

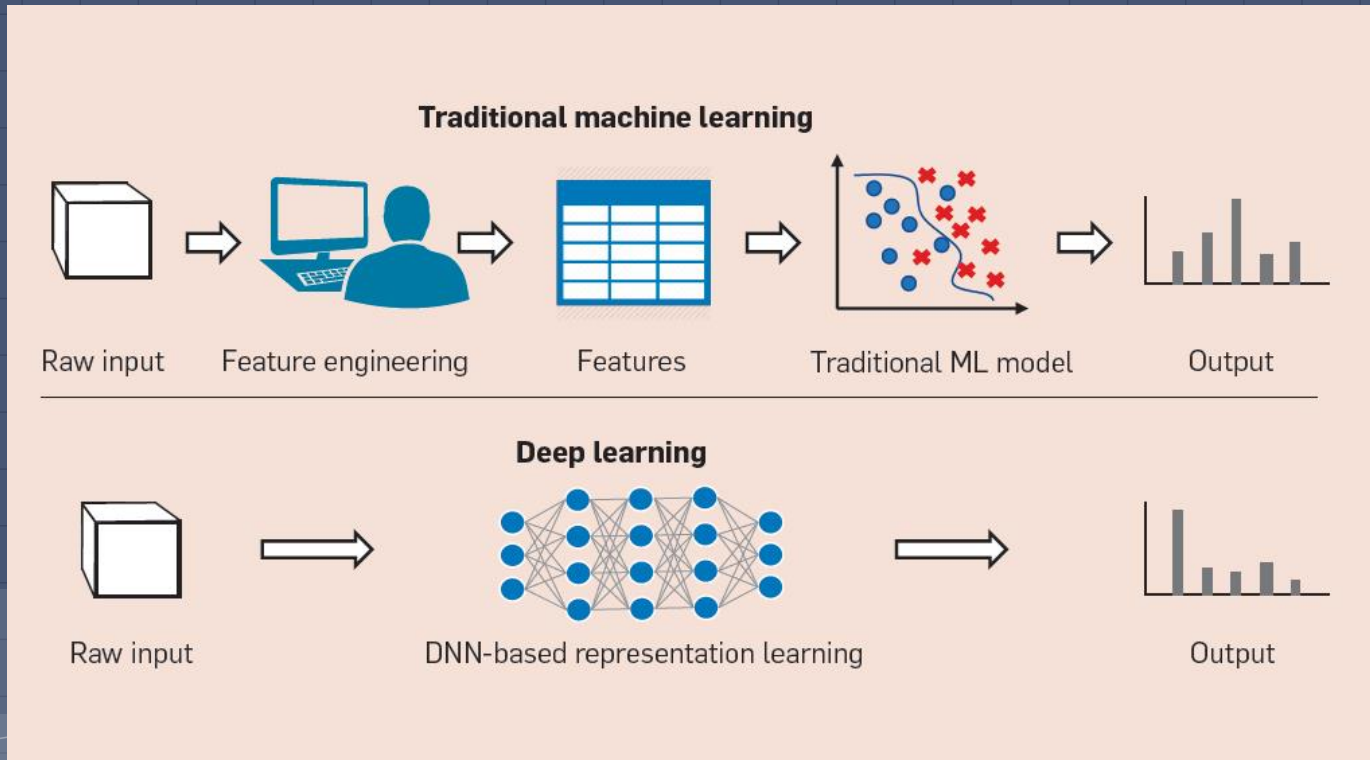
Deep Learning, avagy neurális hálók elmélete

Miért népszerű a Deep learning

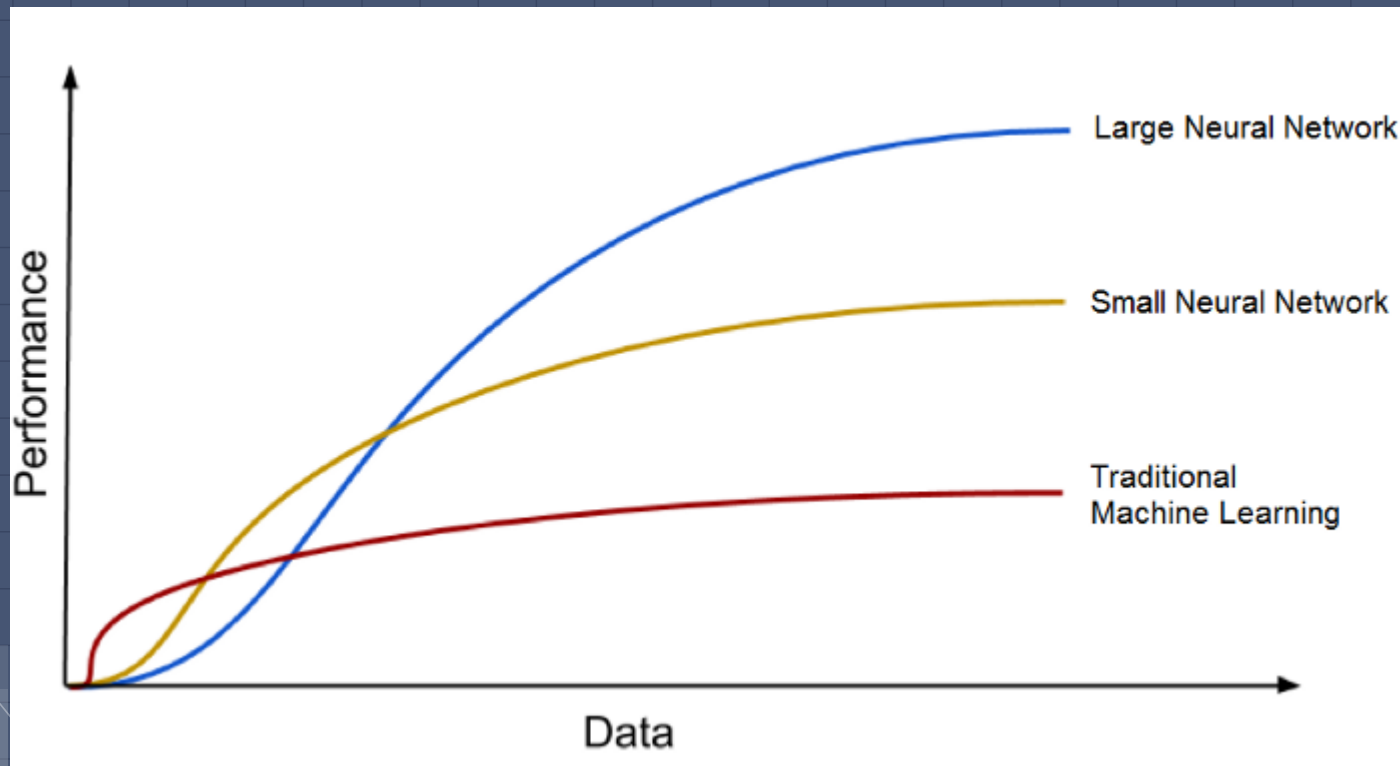
NLP fejlődése, avagy a neurális hálók alkalmazása a szövegelemzésben



Miért terjedt el a deep learning I.



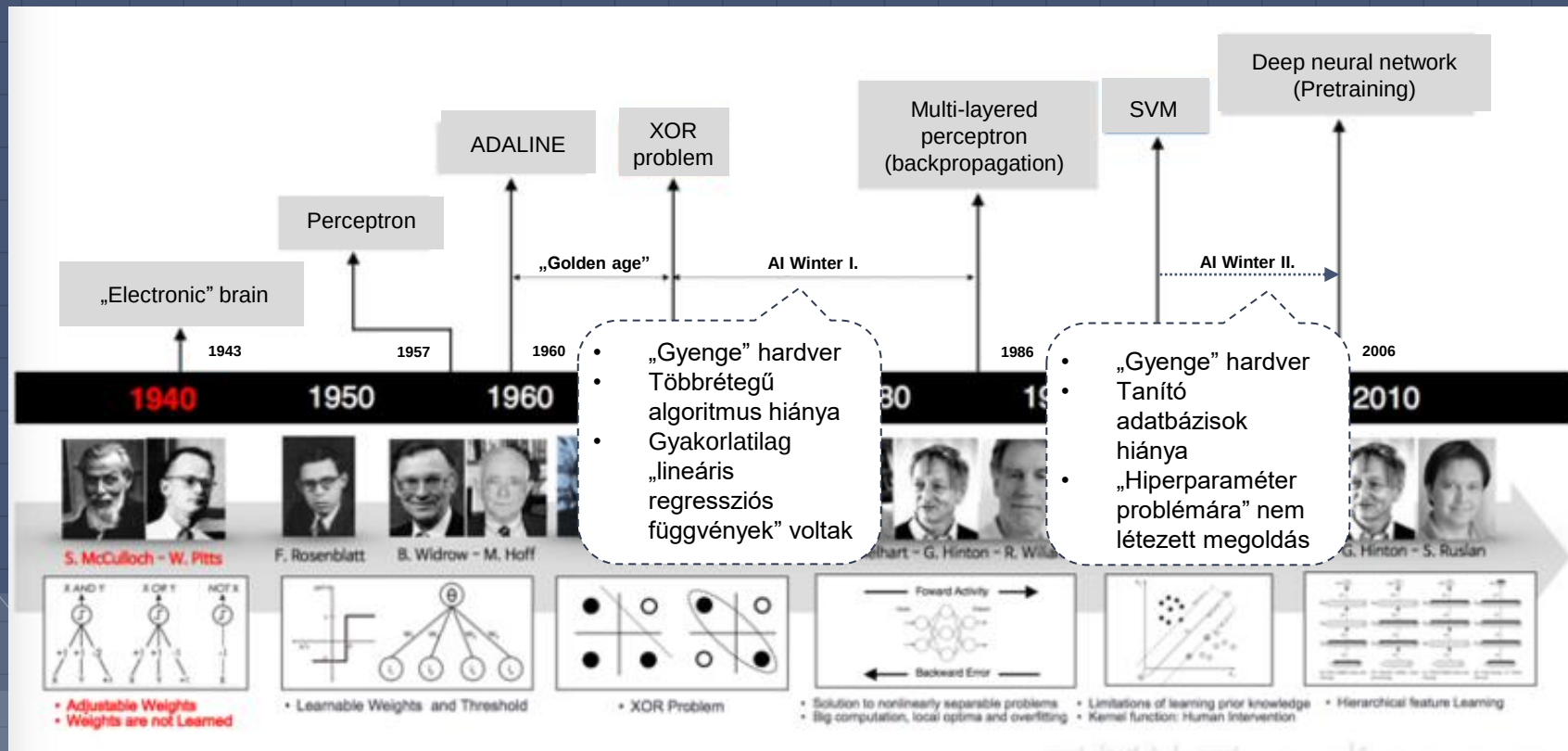
Miért terjedt el a deep learning? II.



Mit jelent egy neurális hálózat tanítása?

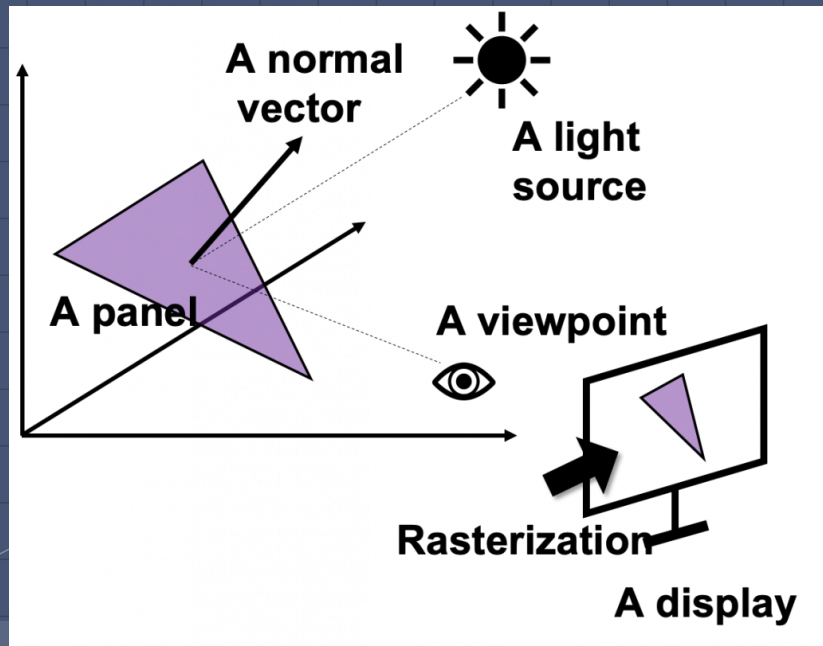
- **Neuronháló esetén a paraméterek: a háló súlyai**
- **A fő metaparaméterek pedig: neuronok száma, rétegek száma, aktivációs függvények (mikor legyen aktív egy neuron és mikor nem)**
 - Ezekkel a háló mérete állítható
 - Nagyobb háló bonyolultabb döntési felületeket képes kialakítani – de növeli a túltanulás veszélyét
 - Különböző metaparaméter-értékek mellett különböző modelleket kapunk
- **Tanítás: forward propagation, backpropagation (hibák visszacsatolás a tanulási folyamatba)**

Neurális hálók története



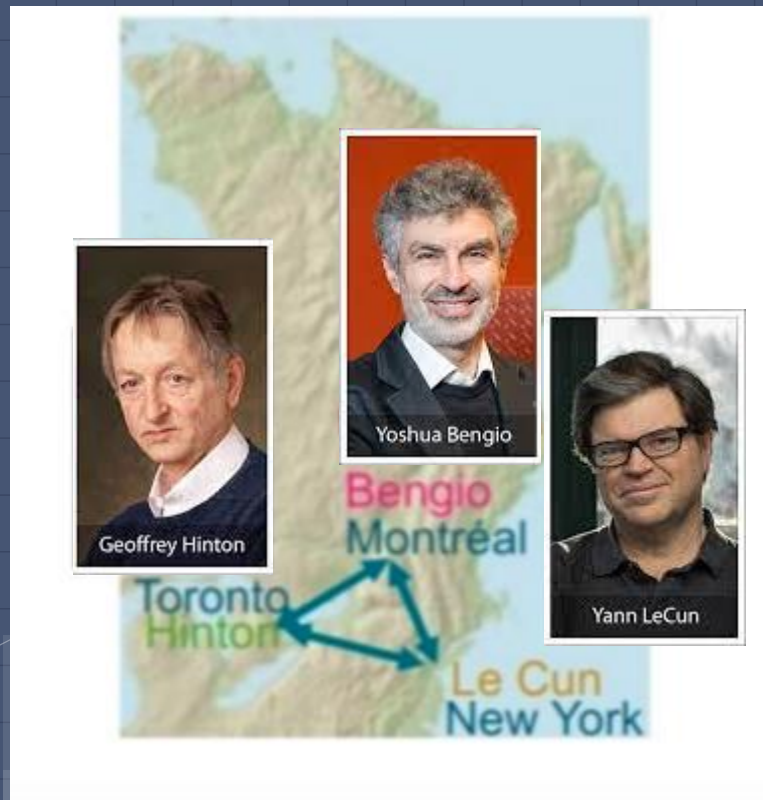
Videójátékok (2000-es évektől)

Massive parallel linear transformations



- CGI-k megjelenése Hollywood-ban (Jurassic park, Toy Story)
- 3D videó játékok
- Xbox megjelenése benne a NV20 GPU (2001)
- Videókártyák megjelenése (és egyben a paralel számolás megjelenése) hatalmas lökést adott a neurális hálók fejlődésének

Deep Learning „keresztapák” (2003–2006)



- Hinton – RBM (Restricted Boltzmann Machine)
- Bengio – NPLM (Neural Probabilistic Language Model)
-> NLP
- LeCun – CNN vagy ConvNet (Convolutional Neural Network)
-> képfeldolgozás

Publikus adatbázisok (2022)

- **MNIST adatbázis:** 60.000 kép (tanító adatbázis) és 10.000 kép (teszt adatbázis)
- **Kinetics-700:** 700.000 videó
- **Google Open Images:** több mint 10M kép
- **IMDB-Wiki:** 5M címkézett kép
- **Amazon Reviews:** 45M Amazon termék értékelés
- **Önvezető autó adatbázisok**
-

Néhány tipikus alkalmazása a Neurális hálóknak

- **Arcfelismerés**
- **Forecasting/prediction** (pl. időjárás előrejelzés, tőzsdei árfolyamok előrejelzése)
- **Szociális média viselkedés elemzés** (cél a felhasználó aktivitásának fokozása)
- **Űrkutatás** (pl. autopilot üzemmódban az utasok biztonságának maximalizálása)
- **Egészségügy** (CNN képfeldolgozó algoritmus, gyógyszerkutatás)
- **NLP**

NLP

„A természetes nyelvek feldolgozása (Natural Language Processing, NLP) olyan mesterséges intelligencia, amely segít a számítógépeknek abban, hogy megértse, tolmácsolja és feldolgozza a természetes emberi nyelveket.”

NLP alkalmazási területei

- **Érzelem detektálás** (közösségi média hozzászólások elemzése, termékvélemények, ..)
- **Fordító programok** (pl. 2019-ben a Facebook superhuman teljesítményt nyújtott)
- **Szövegkinyerés** (releváns információk kinyerése szövegből, pl. termék, ügyfél neve, ...)
- **Chatbotok**
- **Témakategorizálás** (email-ek kategorizálása)
- ...

NLP mit keres egy ML fórumon? – nagy adatbázis

- **Korpusz:** egy adott nyelven összegyűjtött szöveg (és/vagy hanganyag) adatbázis
- **Mitől jó egy korpusz?** Hatalmas méretű, Jó minőségű „Reprezentatív”

Year	Name of the Corpus	Size (in words)
1960s - 70s	Brown and LOB	1 Million words
1980s	The Birmingham corpora	20 Million words
1990s	The British National corpus	100 Million words
Early 21 st century	The Bank of English corpus	650 Million words

NLP mit keres egy ML fórumon? – GPT3

- **OpenAI (Musk) és DeepMind (Google)** – „open éra” két meghatározó szereplője
- **GPT3 (2020)** – 175Mrd paraméterből álló neurális háló (transformer). Ehhez kifejlesztettek egy GP (generative pre-training) eljárást, mely hatalmas szöveg adatbázisokat képes előfeldolgozni emberi beavatkozás nélkül (semi-supervised learning)

Dataset	# Tokens	Weight in Training Mix
Common Crawl (open source)	410 milliárd	60%
WebText2	19 milliárd	22%
Books1	12 milliárd	8%
Books2	55 milliárd	8%
Wikipedia	3 milliárd	3%

AI a „spájzban” van – leváltható az ember az adattudományban?

- DeepMind az év elején jelentette be, az általa fejlesztett AlphaCode egy programozási versenyen (Codeforces) a középmezőnyben végzett (az indulók 46%-át előzte meg)
- Az AlphaCode-nak a versenyfeladatok megoldása során először a természetes nyelven (ti. angolul) megfogalmazott feladatkiírást kellett megértenie, majd arra a kritikus gondolkodás, valamint persze logika és az algoritmizálás szabályait alkalmazva hozta létre a megoldást képező kódokat.