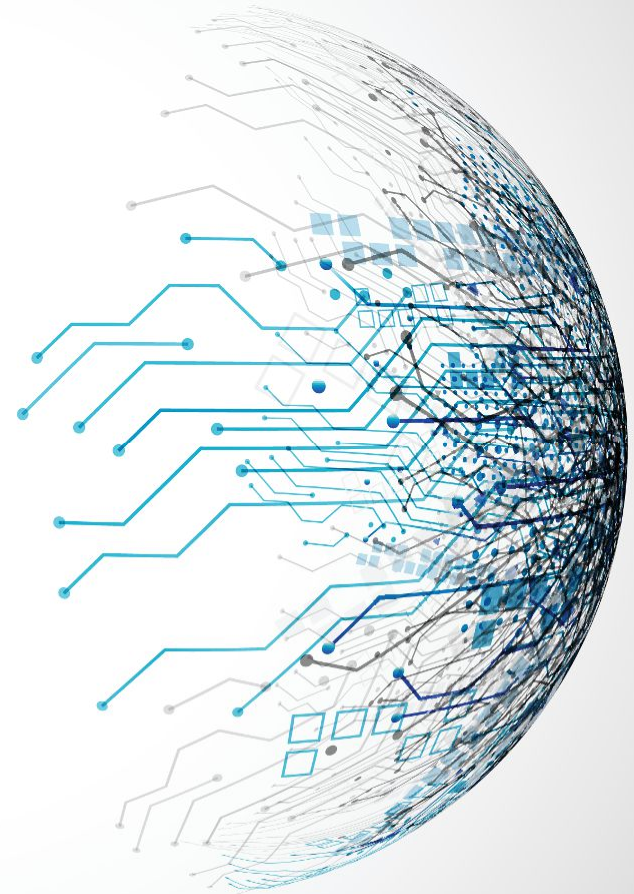


A deep learning iparosodása

Gyires-Tóth Bálint



Bemutakozás



Gyires-Tóth Bálint, Ph.D.

Gyires-Tóth Bálint (toth.b@tmit.bme.hu)

- Jelfeldolgozás
- Idősor modellezés
- Gépi tanulás és mély tanulás 2007 óta

E-mail: toth.b@tmit.bme.hu

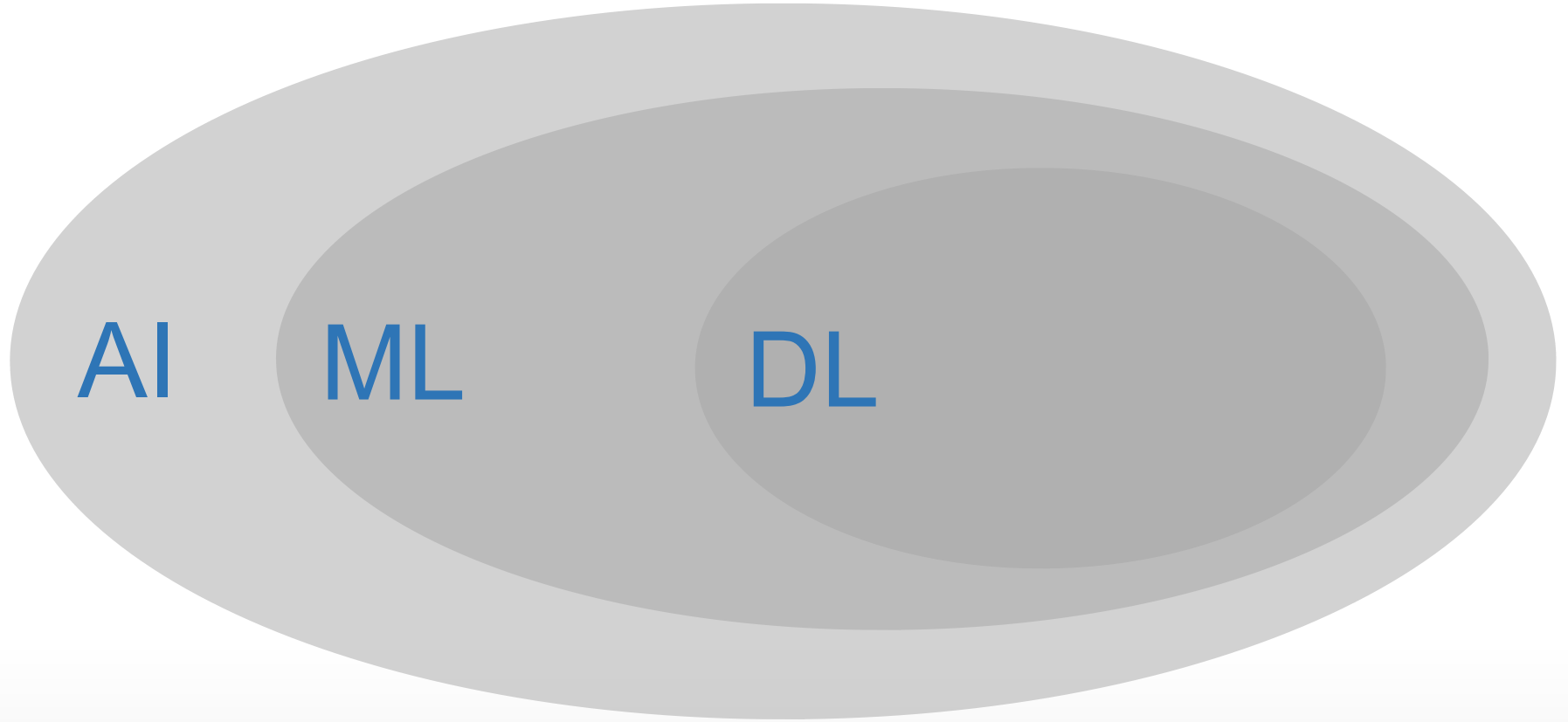
LinkedIn: <https://www.linkedin.com/in/gyires-toth/>

Twitter: <https://twitter.com/bapalto>

Web: <http://smartlab.tmit.bme.hu/dr-gyires-toth-balint>

- SmartLab @ BME TMIT
- NVidia Deep Learning Institute
Certified Instructor & University Ambassador

AI, Machine Learning, Deep Learning



De mi is az a deep learning?

Biológiai indíttatású **gépi tanuló algoritmusok** struktúrált összesége, ahol **több rétegen** keresztül próbáljuk az **adatok különböző szintű absztrakcióit megtanulni és modellezni**.

Gyakorlatban a deep learning kifejezést elsősorban a **mély neuronhálókkal** kapcsolatban használják.

Miben új?

Új algoritmusok + nagymennyiségű adat + GPU

+ **alacsonyabb „belépési küszöb”**

+ **erősen nyílt kutatói közösség**



A deep learning alapú AI a következő nagy lehetőség a növekedésre!

Számos területen már bizonyított...

~ Hallás

~ Látás

~ Beszéd

~ Döntés

A network graph with nodes and edges, overlaid with a white rectangular box containing text.

Skálázhatóság ?= Iparosodás

Hardver architektúra – consumer grade

1-8 GPU workstation (min. 64 GB RAM, 2-4 TB HDD)

NVIDIA TITAN X, 1080Ti, TITAN Xp, Titan V, 2080Ti

- Titan V: 12 teraFLOP (FP32)
- Tensor Core: 110 Tflops (FP16, deep learning)
- Max. 12 GB GPU RAM

Felhő szolgáltatások (AWS, Google Cloud, Azure, Oracle, stb.)

Hardver architektúra – Server grade

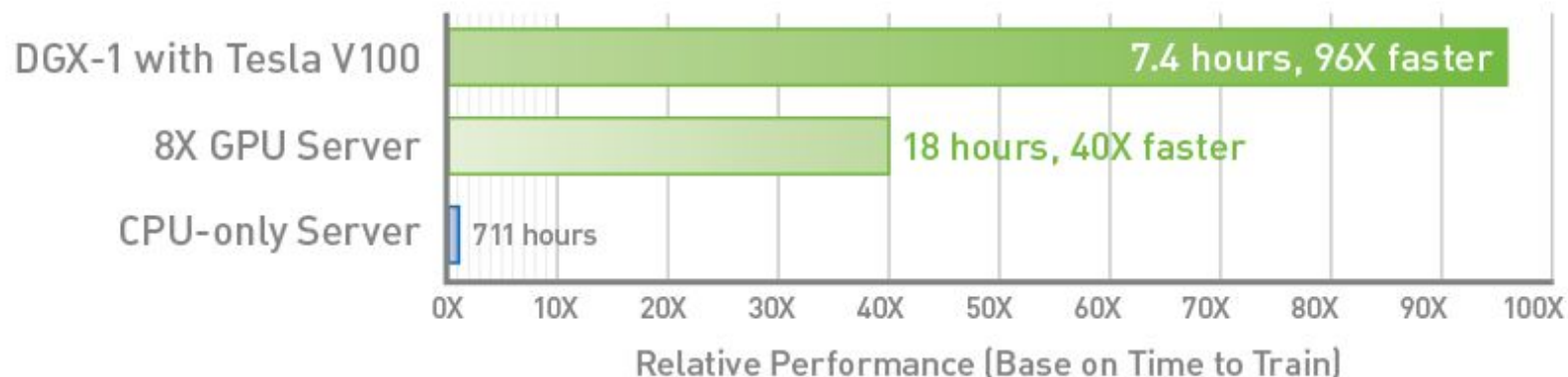


DGX Station (28 FP32, 56 FP16, 480 Tflop DL), \$69.000

DGX-1 V100 (56 FP32, 112 FP16, 960 Tflop DL), \$149.000

DGX-2 V100 (112 FP32, 224 FP16, 2 Pflops DL), \$399.000

NVIDIA DGX-1 Delivers 96X Faster Training



Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz

November 13-ai eredmény: ResNet50 90 epoch tanítás 3.7 perc alatt 2176 db V100-as GPU-n.

Forrás: Hiroaki Mikami, Hisahiro Suganuma, Pongsakorn U-chupala, Yoshiki Tanaka, Yuichi: Kageyama: ImageNet/ResNet-50 Training in 224 Seconds

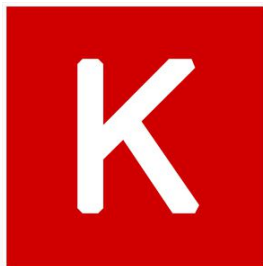
DGX SATURNV



Szoftver architektúra – DL keretrendszerek

Python alapú fejlesztés

- TensorFlow (Google)
- Keras (Google)
- PyTorch (Facebook AI Research, Uber „Pyro” probabilistic programming)
- Microsoft CNTK, Chainer, Theano, Matlab DL Toolbox, etc.
- Caffe, Kaldi

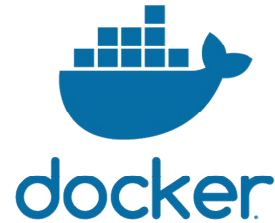


Szoftver architektúra - Skálázhatóság

Docker: Container alapú fejlesztés az egyszerűbb fejlesztés, tesztelés, éles futtatás és skálázhatóság céljából.

Kubernetes: a container alapú rendszerek automatizált menedzsmentje.

Apache Spark 2.3 + Kubernetes 1.7+: az elosztott adatfeldolgozás és modellezés menedzsmentje. Az MLlib könyvtárban van deep learning modul.



Rapids AI

Pandas (2008) → Dask (2015) → Rapids AI (2018)

100 millió adatpont lineáris transzformációja

Techonológia	Sebesség
Pandas (CPU)	3957.1 sec
Dask (CPU paralell)	788.55 sec
Rapids (GPU)	2.103 sec

CPU: Dell T7910 Xeon E5-2640V4x2

GPU: NVIDIA Titan Xp

Rapids GPU DataFrame –Pandas-ból

```
import pandas as pd
from cudf.dataframe import DataFrame
pdf = pd.DataFrame({'a': [0, 1, 2, 3],
                     'b': [0.1, 0.2, None, 0.3]})
df = DataFrame.from_pandas(pdf)
print(df)
```

Rapids GPU DataFrame – read_csv

```
import cudf

df = cudf.read_csv('data.csv', delimiter=',',
                   names=['col1', 'col2'],
                   dtype=['int64', 'int64'])
```

Rapids GPU DataFrame – apply_rows

```
import cudf

import numpy as np

df = cudf.dataframe.DataFrame()

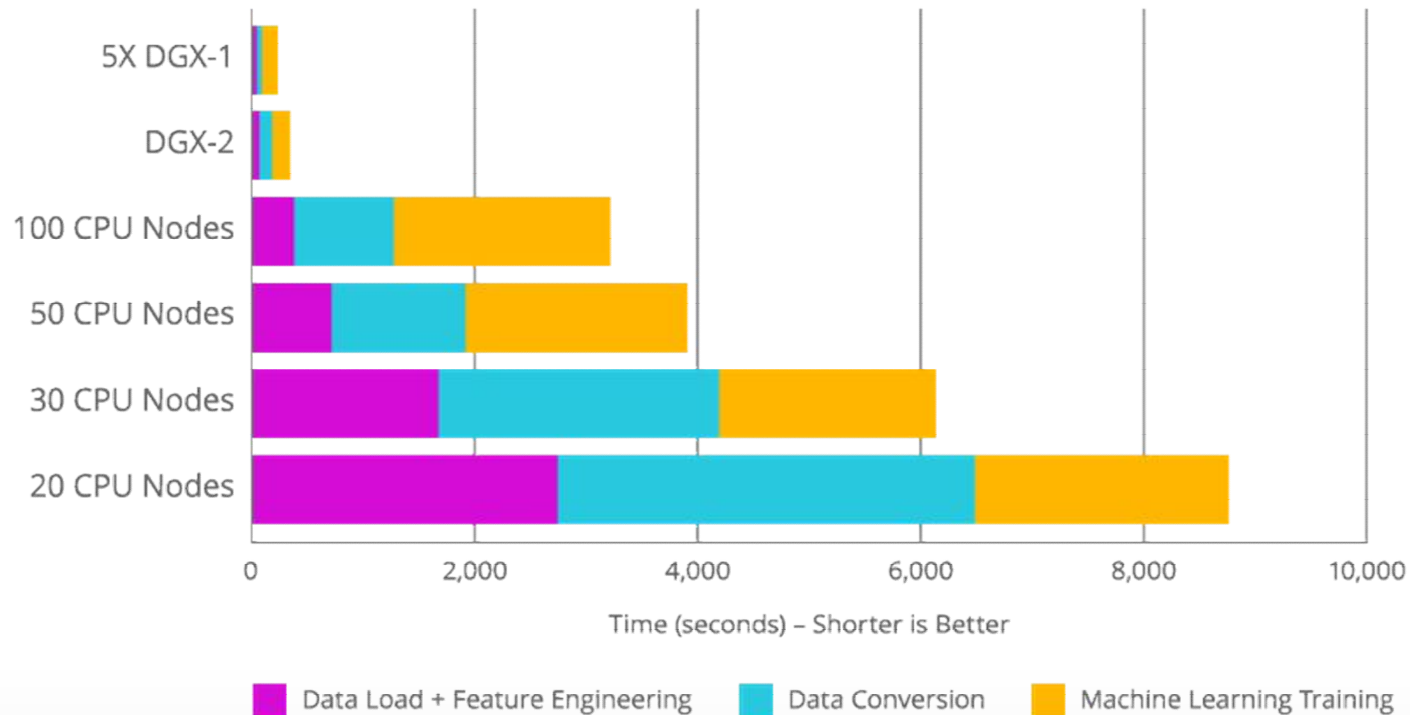
nelem = 3

df['in1'], df['in2'], df['in3'] = np.arange(nelem), np.arange(nelem), np.arange(nelem)

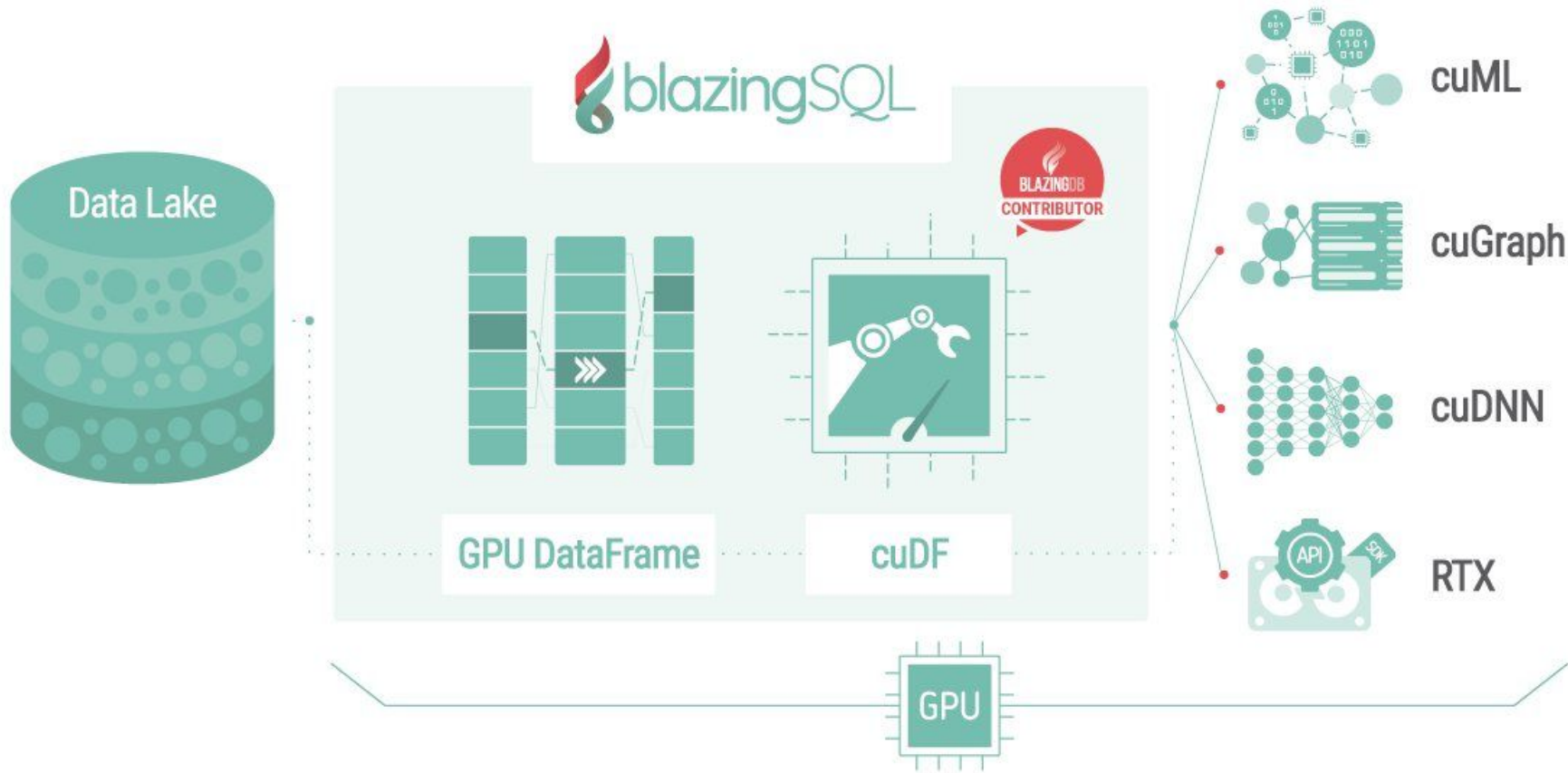
def kernel(in1, in2, in3, out1, out2, kwarg1, kwarg2):
    for i, (x, y, z) in enumerate(zip(in1, in2, in3)):
        out1[i] = kwarg2 * x - kwarg1 * y
        out2[i] = y - kwarg1 * z

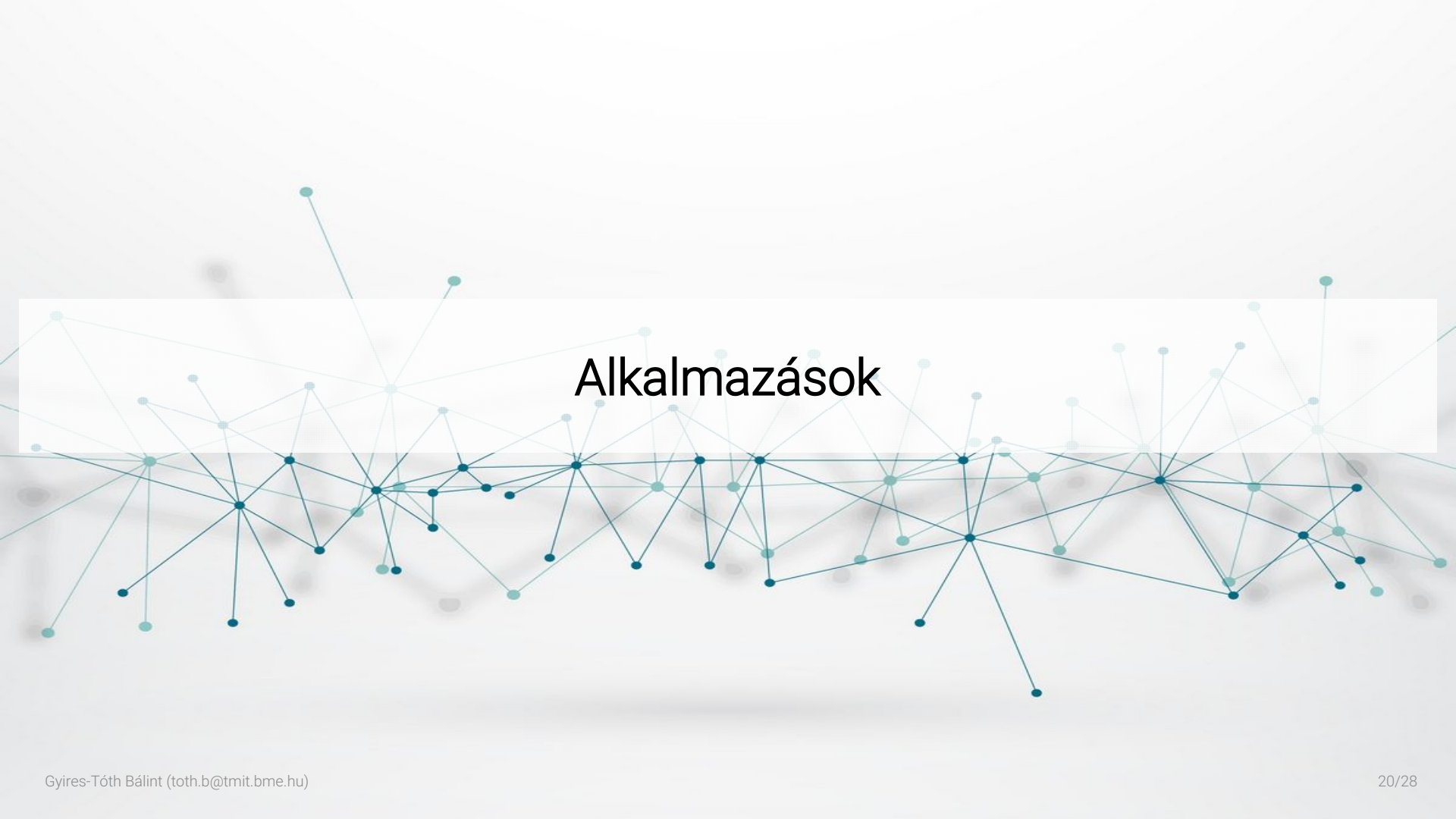
df.apply_rows(kernel, incols=['in1', 'in2', 'in3'],
               outcols=dict(out1=np.float64, out2=np.float64),
               kwargs=dict(kwarg1=3, kwarg2=4))
```

Rapids AI



Rapids AI + Blazing SQL



An abstract network diagram with numerous nodes (dots) and connecting lines (edges) spread across the slide. The nodes are colored in shades of teal and blue. A white rectangular box is centered horizontally, containing the title text.

Alkalmazások

Alkalmazás 1

Idősor klasszifikáció.

CÉL:

Okostelefon alapú gesztus felismerés

BEMENET:

200 x 7 adatpont (1-2 seconds)

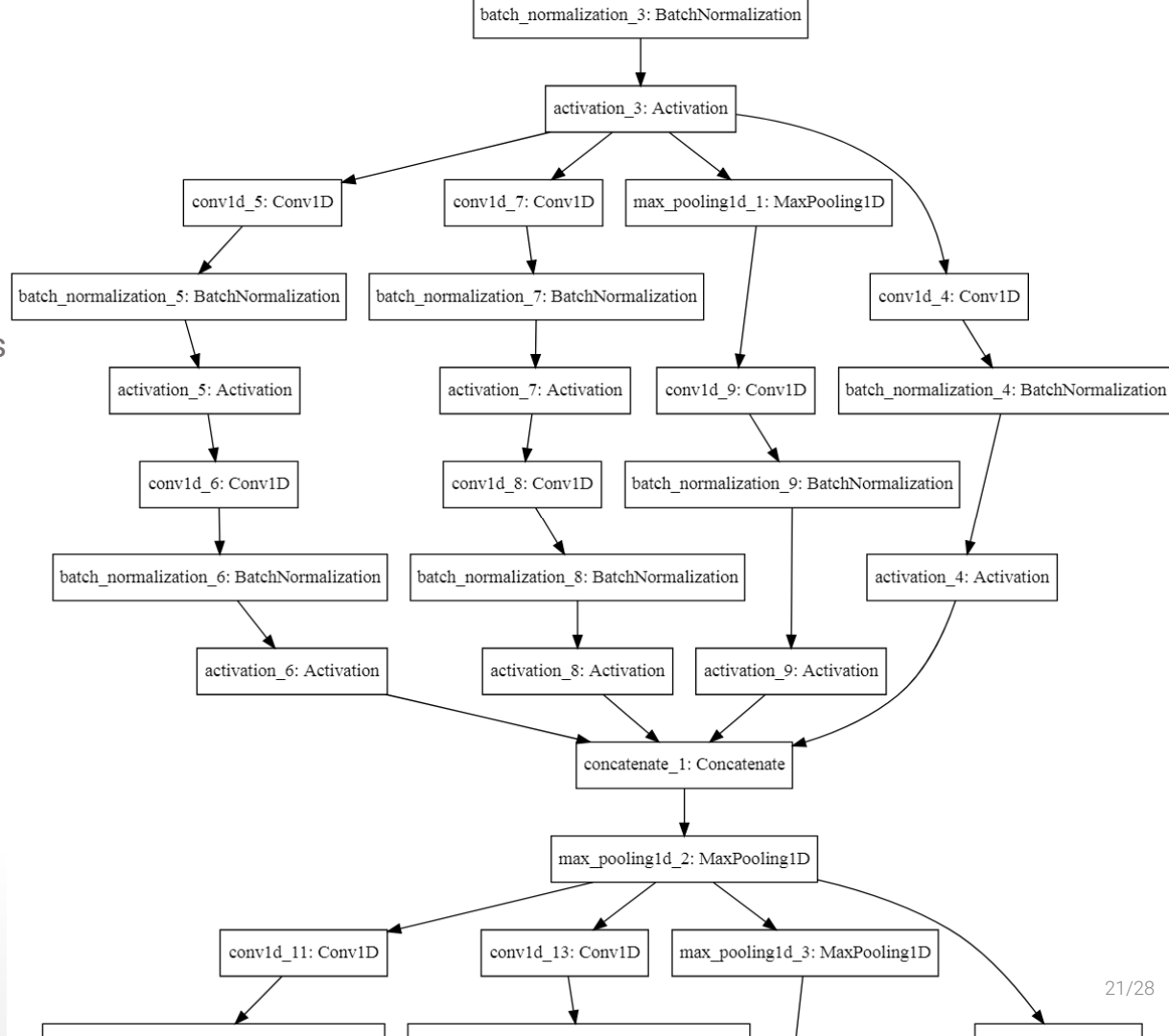
KIMENET:

Bináris osztályozás

>99% pontosság.

~1000 óra adat 100 embertől

Kb. 26000 GPU óra (~3 GPU év) 12
Tflops-os GPU-kon.



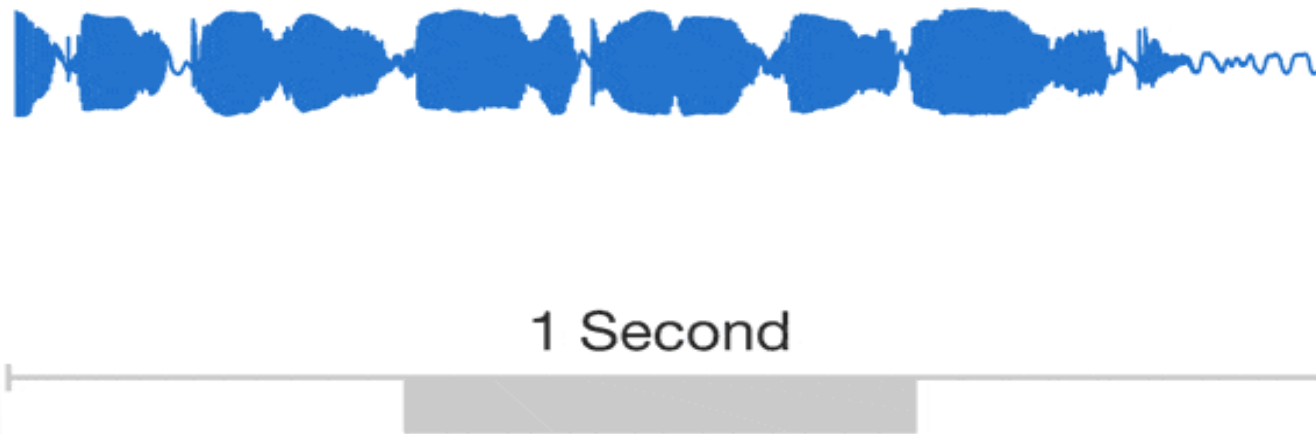
Alkalmazás 1

#tanítható paraméterek: 632834

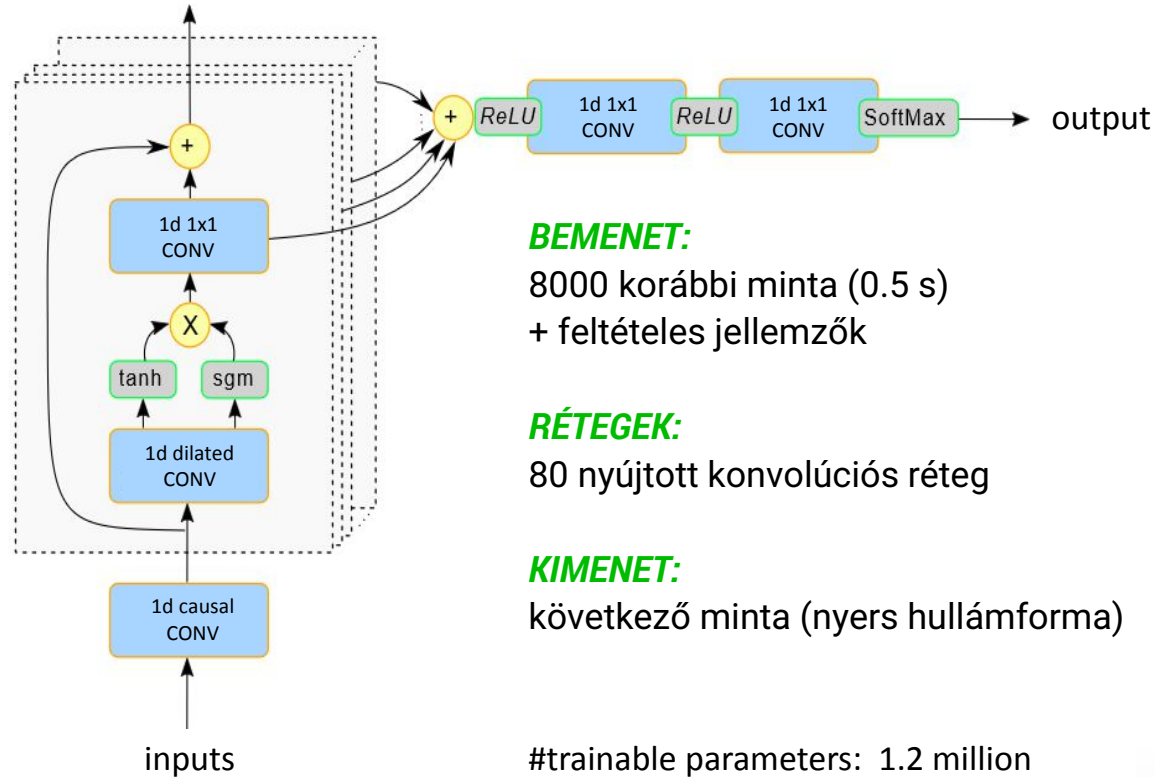
#tanítható paraméterek: 286498

Alkalmazás 2 - WaveNet

- Google DeepMind, 2016 szeptember (később: Deep Voice, Baidu)
- Új típusú Text-to-Speech eljárás



Alkalmazás 2 - WaveNet



(DeepMind WaveNet alapján)

Alkalmazás 2 - WaveNet hangminták

Google DeepMind – angol



SmartLab – magyar



Alkalmazási területek

Google Duplex, call center-ek, vasútállomás, képernyő felolvasók, okostelefonok, automotive, stb.

Alkalmazás 2 - WaveNet számokban

Tanítás

25 óra stúdiófelvétel

Régi megközelítés: 4 hét 12 Tflops GPU-n

Új megközelítés: 1 hét DGX-1: ~1800\$ szerverbérlet

Predikció

1 minta generálása 7.5 ms

1 másodperc generálása 12 másodperc

Összefoglalás

Deep Learning jövőbemutató eredményei

Feketedoboz működéstől való félelem

Skálázhatóság + Iparosodás

Alkalmazott kutatás és ipari alkalmazások

Köszönöm a figyelmet!

Gyires-Tóth Bálint
toth.b@tmit.bme.hu



Jelen kutatás az Emberi Erőforrások Minisztériuma által az Új Nemzeti Kiválóság Program keretében meghirdetett Bolyai+ Felsőoktatási Fiatal Oktatói, Kutatói Ösztöndíj támogatásával készült (ÚNKP-18-4-BME-394).



SmartLab
Intelligent Interactions

