Microsoft

Mesterséges Intelligencia Tudásplatform

Szent István Egyetem
Gödöllő, 2019. szeptember 11.

Varga Gábor
Magyarországi stratégiai technológiai tanácsadó
linkedin.com/in/gaborvar

Delhi: Facial recognition system helps trace 3,000 missing children in 4 days

PTI | Apr 22, 2018, 19:23 IST

A- A+

NEW DELHI: Nearly 3,000 missing children have been traced in four days, thanks to the facial recognition system (FRS) software that the Delhi Police is using on a trial basis to track down such children. The identities of the missing children have been established and efforts are on to help them reunite with their families.

A mesterséges intelligencia megítélése polarizált

A közbizalom megteremtése az MI-vel szemben

Mit értünk mesterséges intelligencia alatt?

Olyan technológiát, amely érzékel, tanul, következtet, és ezzel kiterjeszti az emberek és szervezetek képességeit. (Microsoft)

„AI is whatever hasn't been done yet.”
(Tesler’s theorem)

Vagyis:
Önmagát falja fel a MI fogalma. Nem érdemes vitatkozni arról, mi tartozik bele, úgyis változik.

Közkézen forgó MI fogalmak

- Részterületek, de nem szinonimák
  - Gépi tanulás
  - Deep learning (mélytanulás) - főleg képekhez
  - Neurális hálózatok
  - Matematikai algoritmusok, amelyekben van tanulási elem
  - Stb.

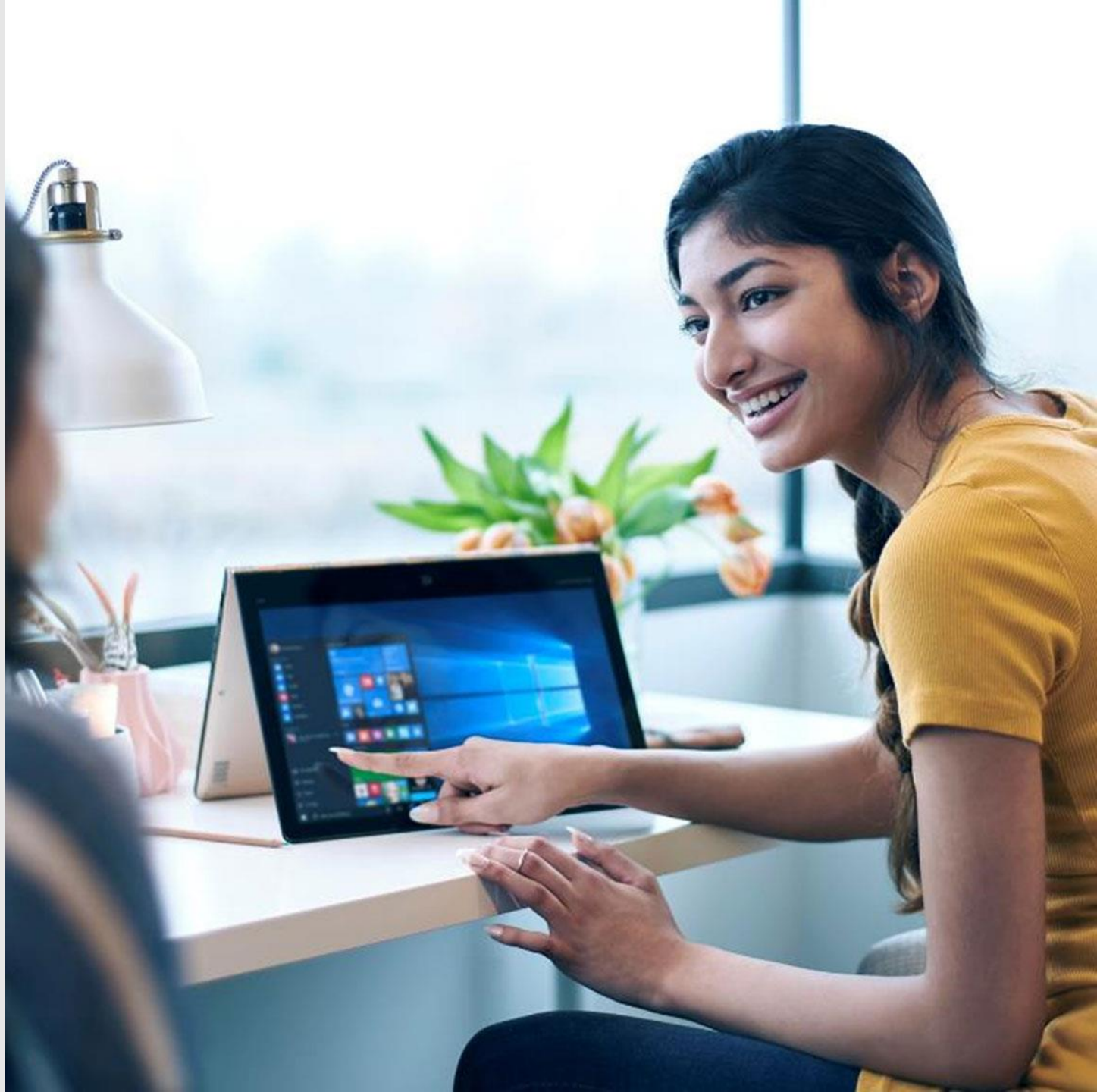
- Szűk AI és általános AI
  - Microsoft has announced a $1 billion investment in OpenAI as part of a shared ambition to achieve the holy grail of artificial general intelligence (AGI). (Aug 2019)

Kissinger, Eric Schmidt on AI revolution:

This revolution is unstoppable. Attempts to halt it would cede the future to that element of humanity more courageous in facing the implications of its own inventiveness.

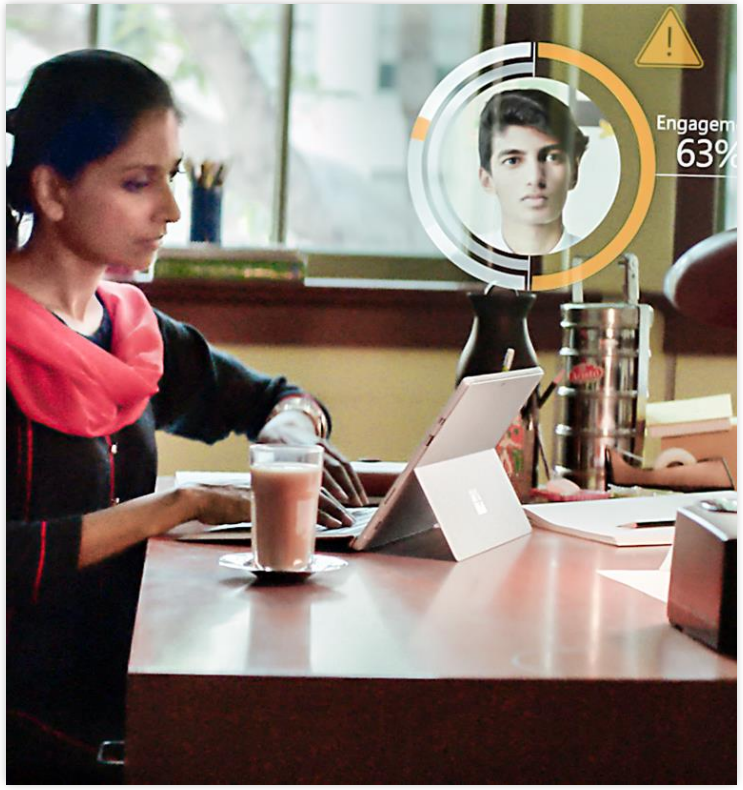
Instead, we should accept that AI is bound to become increasingly sophisticated and ubiquitous, and ask ourselves: How will its evolution affect human perception, cognition, and interaction? What will be its impact on our culture and, in the end, our history?

From <https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/>



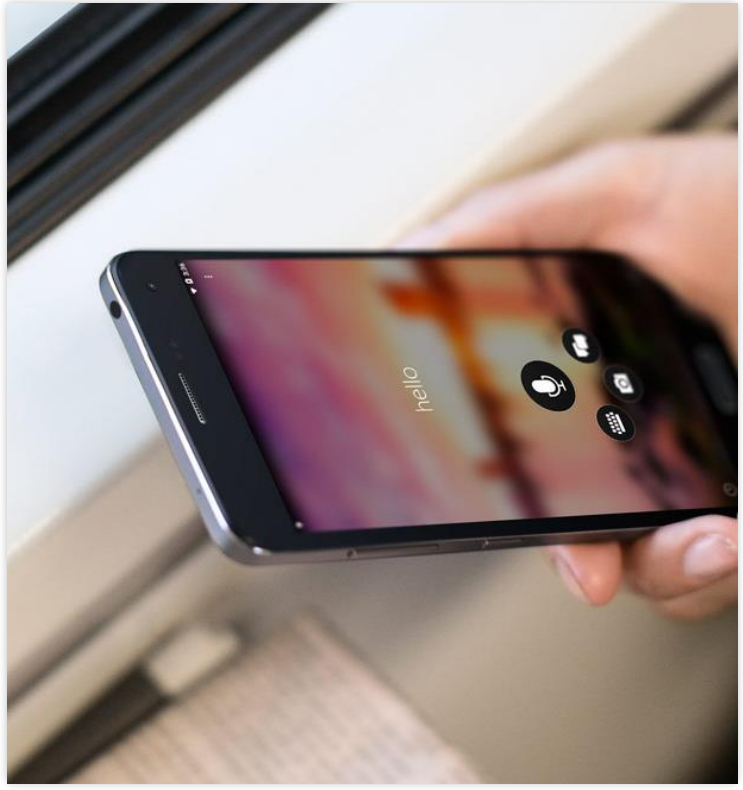
Mesterséges intelligencia mögötti trendek

Számítási felhő



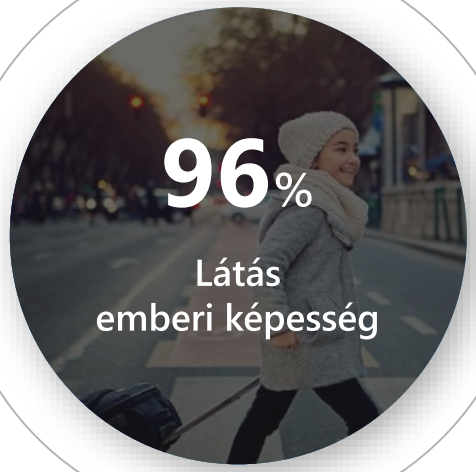
Adat

Algoritmusok



Microsoft AI - az emberi képességek szintjén

Látás



2016

Beszéd

2017

2018

Nyelv

2018

McKinsey elemzés
Notes from the AI frontier: Modeling the impact of AI on the world economy

„at the global average level of adoption and absorption implied by our simulation, AI has the potential to deliver additional global economic activity of **around $13 trillion by 2030, or about 16 percent higher cumulative GDP compared with today**. This amounts to 1.2 percent additional GDP growth per year. If delivered, this impact would compare well with that of other general-purpose technologies through history."

https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-AI-frontier-modeling-the-impact-of-ai-on-the-world-economy

Elemzés: How AI can enable a Sustainable Future

„Agriculture Annual GDP contribution +$350bn by 2030”

„The UN’s Food and Agriculture Organization (FAO) forecasts that global agricultural production must more than double by 2050 to prevent mass food shortages. AI forms a central part of the technological innovations that are transforming agricultural production by responding to growing demand in a way that limits social and ecological trade-offs.”

Microsoft

in association with pwc

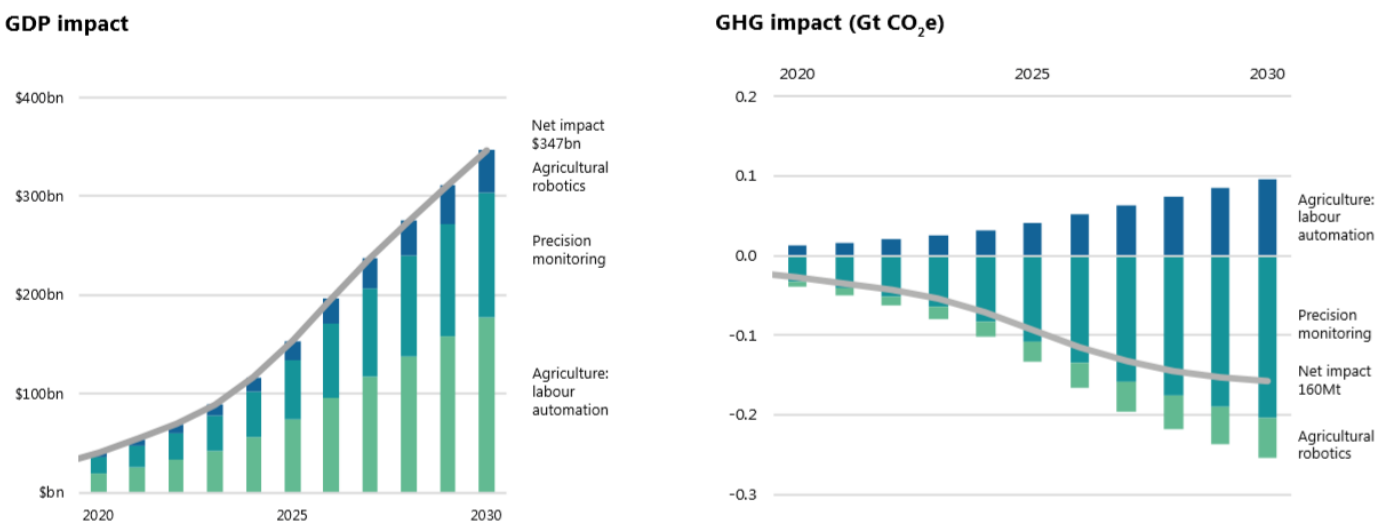
Elérhető: https://www.pwc.co.uk/sustainability-climate-change/assets/pdf/how-ai-can-enable-a-sustainable-future.pdf

Mezőgazdaságról részletek a 27. oldalon

The key AI levers we study within agriculture are:

- Agricultural robotics: This includes AI robotics that are programmed to carry out agricultural tasks autonomously with optimal timing. For example, an autonomous tractor picking fruit only when ripe.
- Precision monitoring of environmental conditions for agriculture and forestry: This includes utilizing field sensors to precisely measure the impact of environmental factors and inputs on agricultural and forestry activities, and provide agri-advisory services. Examples include monitoring local weather conditions to predict the impact on yield and tailor required inputs.
- Land-use planning and management: This involves using AI for mapping agricultural and forestry activities over time for better farm management and better enforcement of regulation.
- Monitoring of crop, soil and livestock health: This deals with monitoring conditions of agriculture (e.g. crop health, prevalence of pests, disease among livestock) to inform better management of crop habitats, and of livestock. For example, monitoring and identification of pests in real time to inform use of pesticides, including volume needed, specific locations on a farm that pesticides are needed etc.

Figure 9: Global impact of environmental AI in the agriculture sector on GDP and GHG emissions in the "Expansion" scenario



Source: PwC analysis

Note: This excludes the impact of land-use planning and monitoring on GHG emissions. Refer to Section 'Wider Impacts and areas for further exploration' for details of the off-model assessment of those impacts.

https://www.pwc.co.uk/sustainability-climate-change/assets/pdf/how-ai-can-enable-a-sustainable-future.pdf

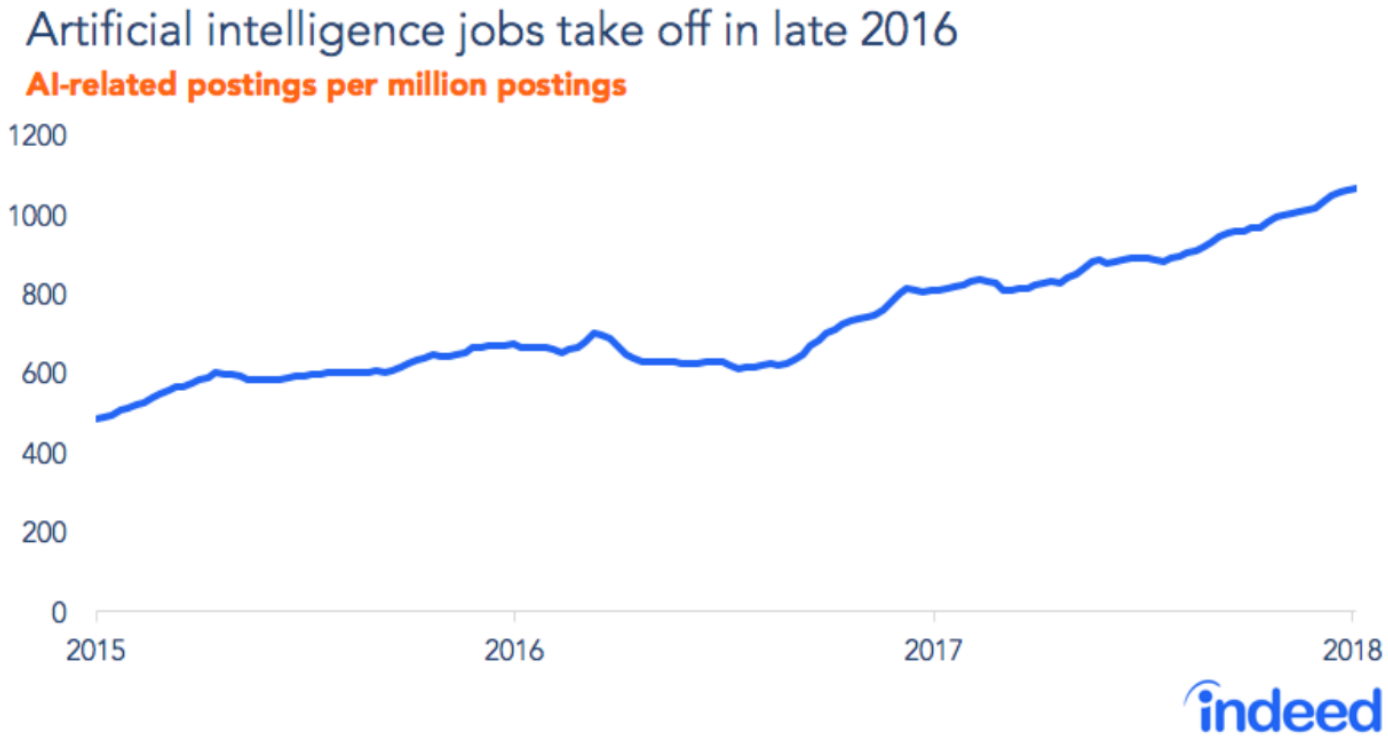
Mit jelent ez az életpálya szemszögéből?

„AI is expected to create 2.3 million jobs by 2020, replacing the 1.8 millions it will eliminate”, according to a Gartner report.

„Demand for workers with artificial intelligence (AI) skills has rapidly increased in the past 18 months, according to a new report from job search site Indeed. However, job seeker interest in these roles has leveled off, suggesting that competition for tech talent in this area is going to heat up quickly.”

https://www.techrepublic.com/article/demand-for-ai-talent-exploding-here-are-the-10-most-in-demand-jobs/

Employer demand for AI skills is soaring



https://www.hiringlab.org/2018/03/01/demand-ai-talent-rise/

Mesterséges Intelligencia Koalíció

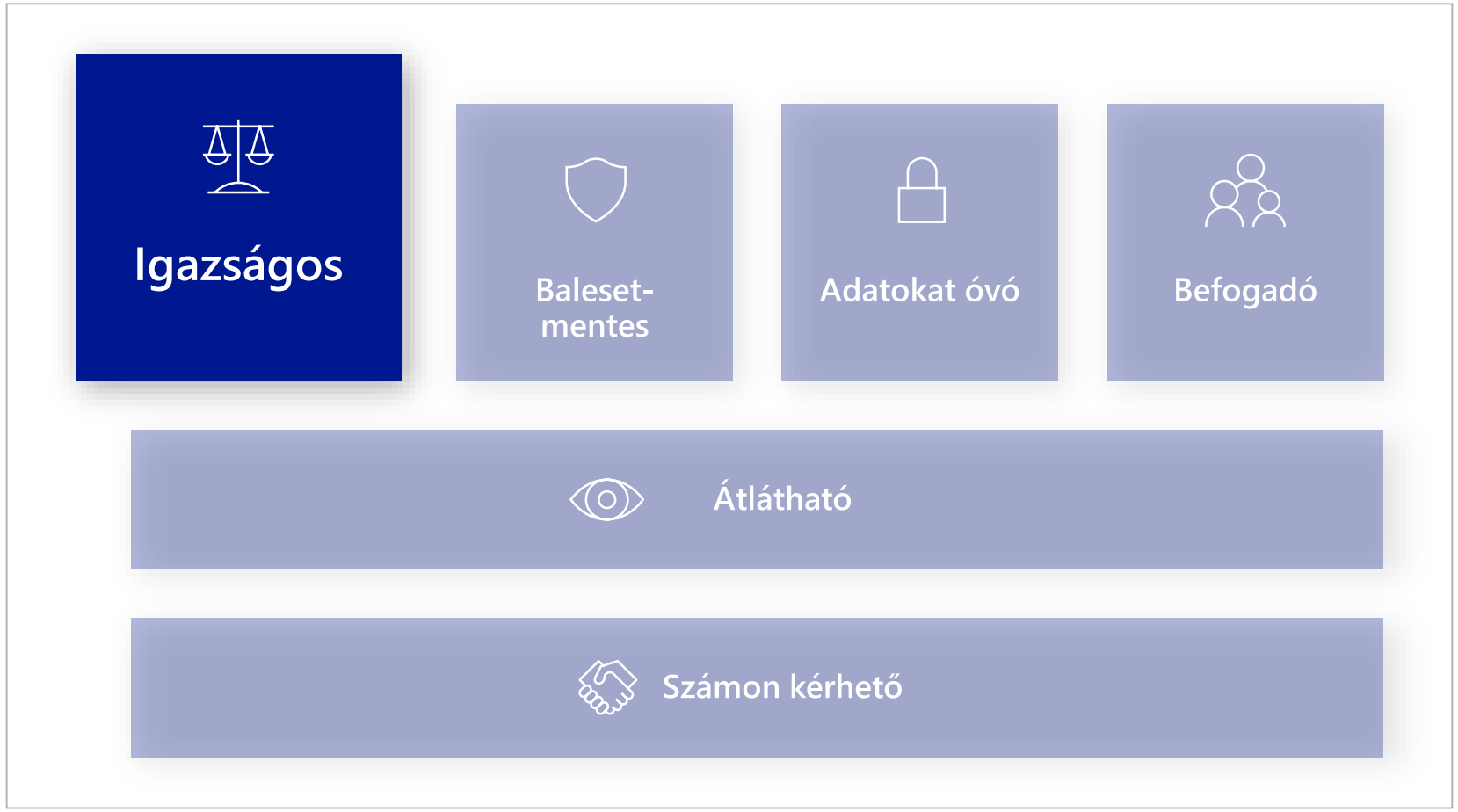
Az innovációs és technológiai miniszter által kezdeményezett Magyarországi Mesterséges Intelligencia Koalíció (MI Koalíció) célja, hogy Magyarország a mesterséges intelligencia fejlesztések és alkalmazások terén kerüljön az európai élvonalba, és váljon a nemzetközi MI közösség fontos tagjává.

Tevékenység munkacsoportokban

- Adatpiac és felhő környezet
- Intézményrendszer, kiválósági központ
- Szabályozás

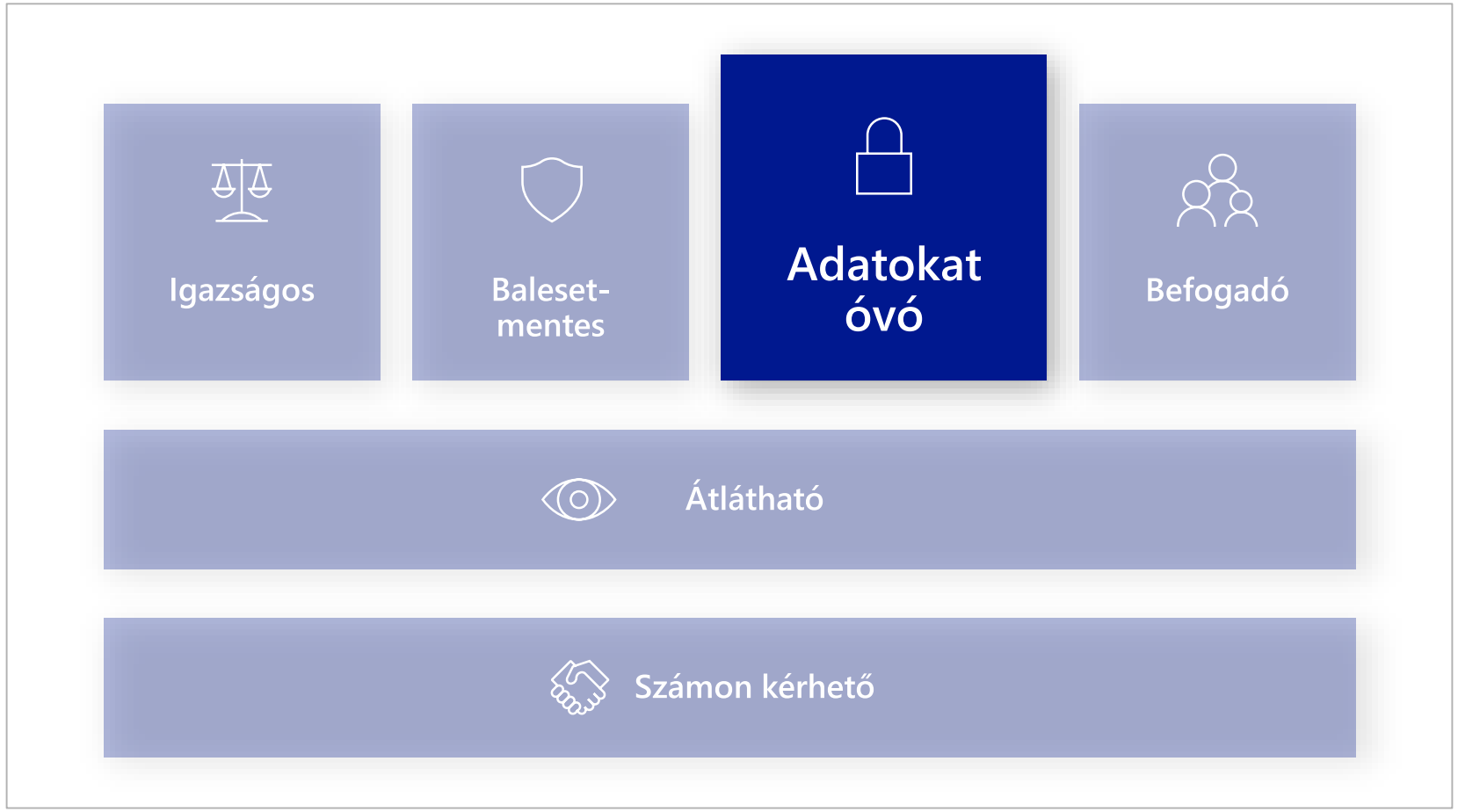
Milyen az etikus mesterséges intelligencia

Az etikus mesterséges intelligencia

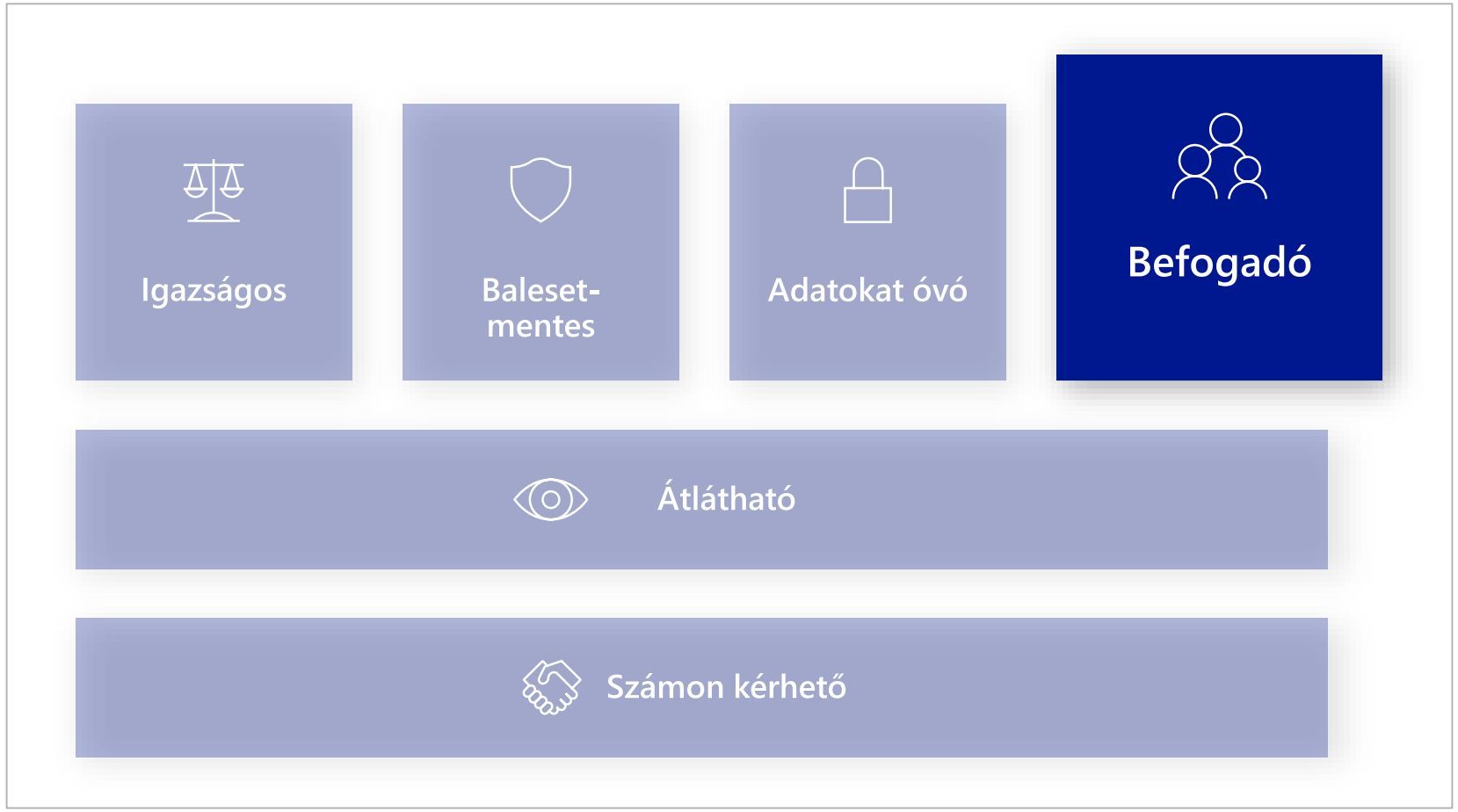


Az etikus mesterséges intelligencia

Az etikus mesterséges intelligencia

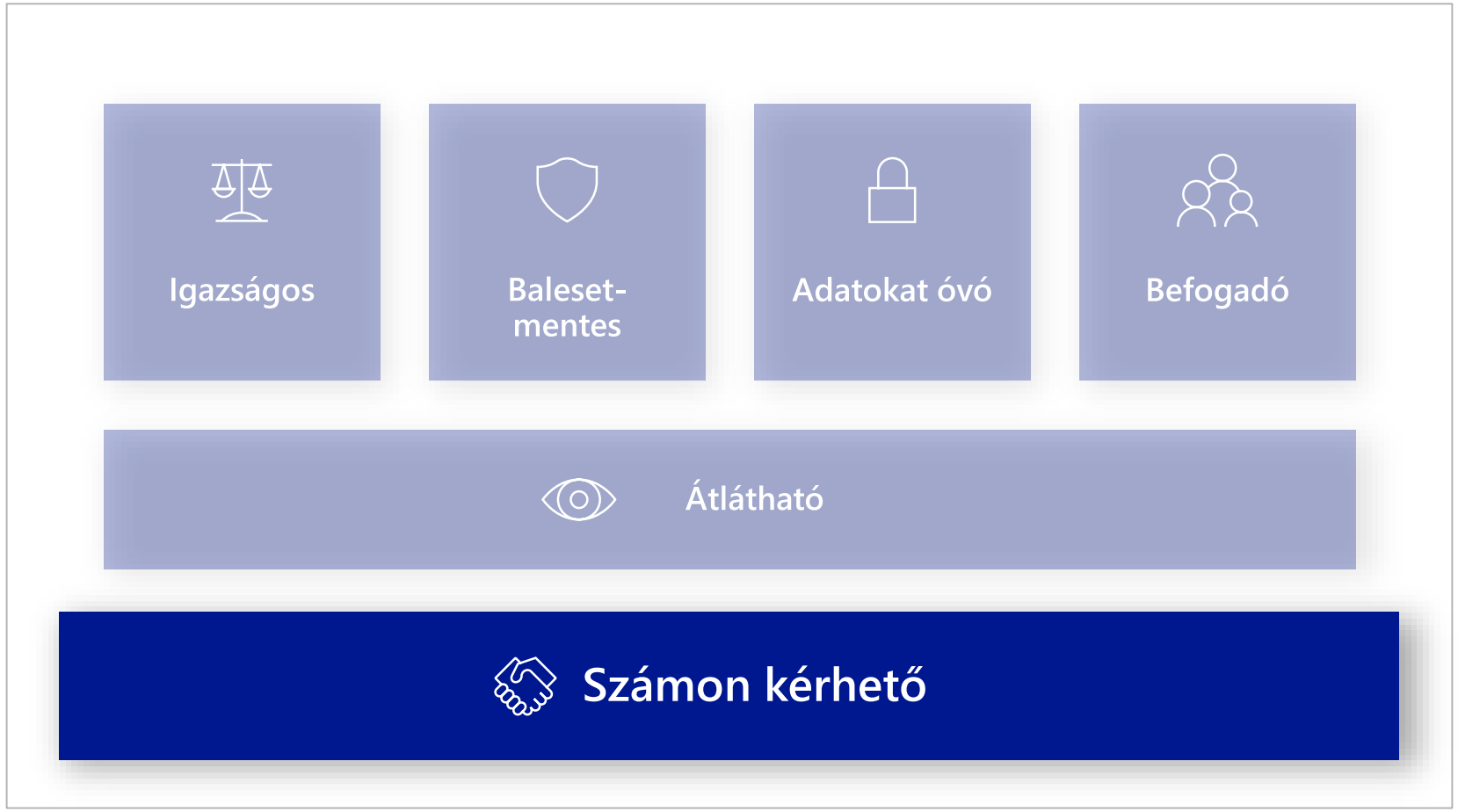


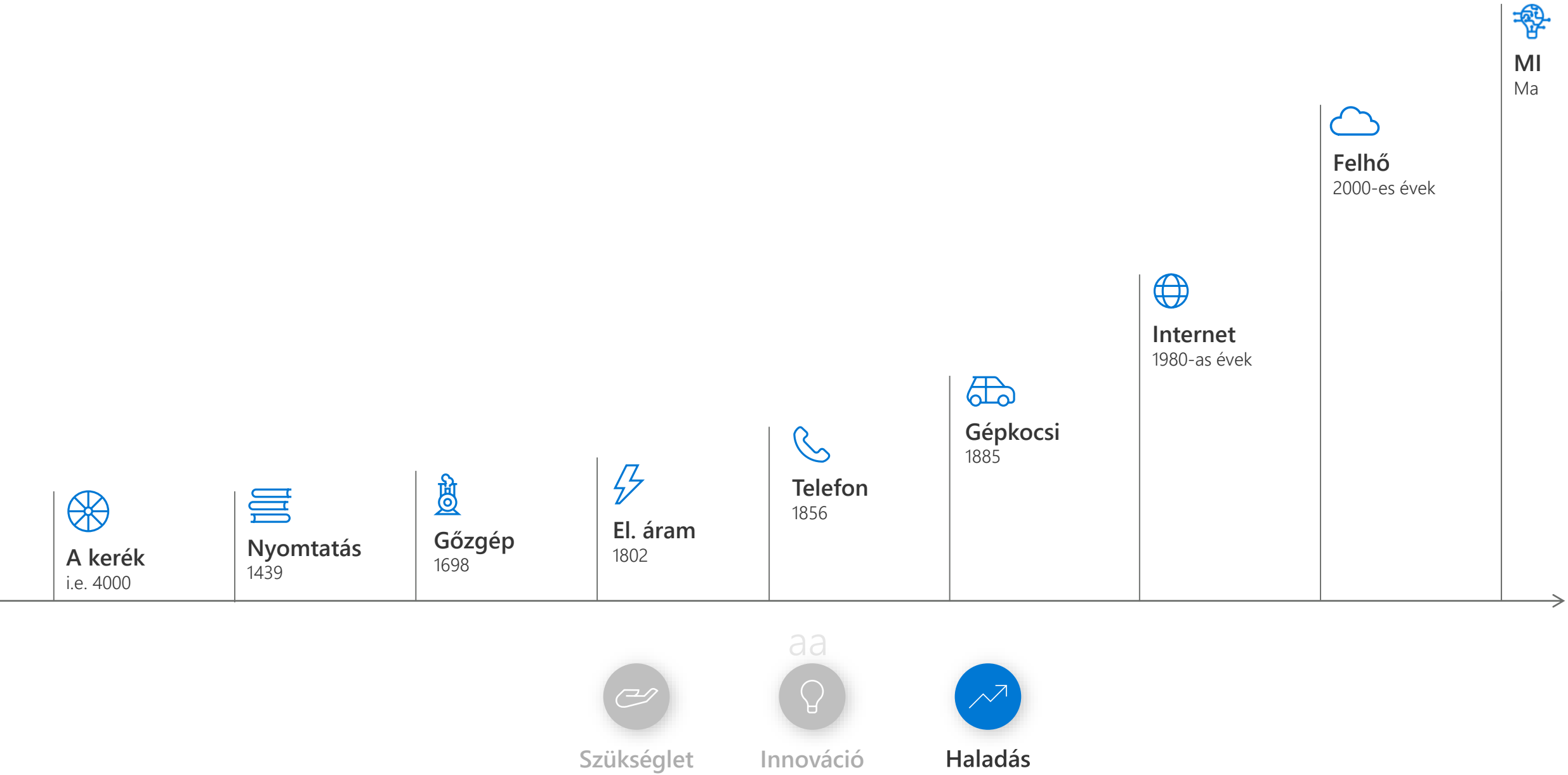
Az etikus mesterséges intelligencia



Az etikus mesterséges intelligencia

Az etikus mesterséges intelligencia




A kerék
i.e. 4000
Nyomtatás
1439
Gőzgép
1698
El. áram
1802
Telefon
1856
Gépkocsi
1885
Internet
1980-as évek
Felhő
2000-es évek
MI
Ma
aa
Szükséglet
Innováció
Haladás

Miért kemény dió a mesterséges intelligencia?

Soha nem látott veszélyeket kell kezelnünk

- Etika
- Robotok uralma
- Szabadságjogok

Ez így volt minden korábbi innovációnál is

- Elektromosság: nincsen hozzá érzékszervünk
- Kerék: a féket is fel kellett találni

Feladataink

Sokat kell beszélni a MI-ről

Uniós vagy magyar vagy iparági vagy ön-szabályozás?

Analógia: motorizáció a XIX. század végétől mostanáig

- KRESZ
- oktatás társadalmi méretekben
- ittas vezetés
- töréstesztek
- műszaki vizsga
- piros zászlós jelzőőr autó előtt
- biztonsági öv – és a több évtizedes meggyőzés
- stb.
- Lesz ezeknek megfelelője az MI világában?

Európai szabályozás

EU vezérelv: Egységes piac (Digital Single Market)

Kifejezetten tiltja a tagállami határokon belüli tárolás megkövetelését: EU 2018/1807 rendelet a nem személyes adatok szabad áramlásának keretéről

Személyes adatok védelme (GDPR) nem alkalmaz földrajzi korlátokat

Kutatási adatokra speciális szabályok alkothatók

GDPR célhoz kötöttség – de rendelkezik az eltérő célú adatkezelés összeegyeztethetőségéről is (8. cikk)

Európai Bankhatóság (EBA): nincs országon belüli tárolási követelmény

A felhő igazi előnyeit csak országonkénti aprózódás nélkül lehet realizálni: Kínával versenyezve a nemzeti szintű aprózódás megengedhetetlen

Mesterséges intelligencia szabályozása

Óvatosan kell haladni

Érzékeny pontokon káros a korlátlan verseny

- Az etika keresése fontosabb, mint a gyors bevezetés
- Pl arcfelismerés
- Szabályozásra van szükség

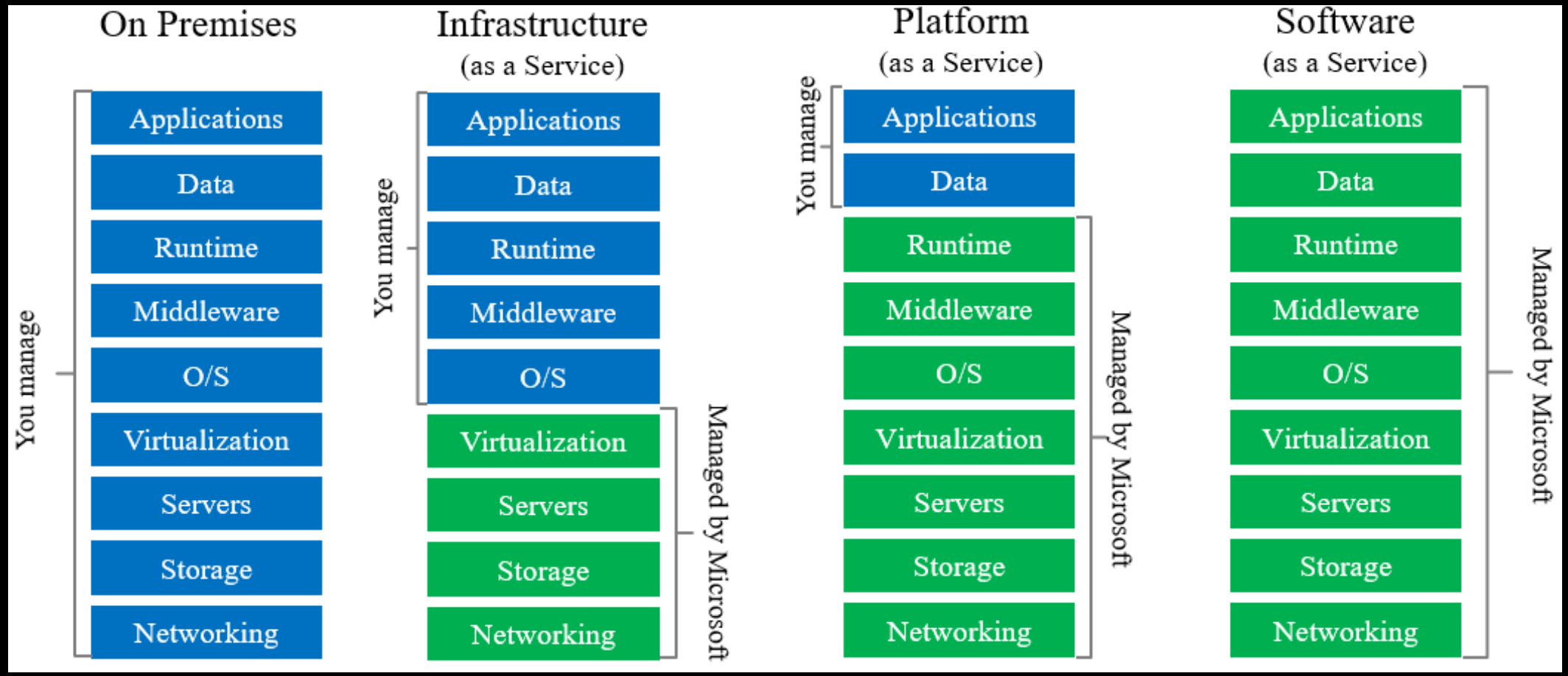
Érthetőség vagy dokumentálhatóság ?

- Elvi akadály: Az intelligencia velejárója a saját döntés
- Gyakorlati : minden tanulási adatot el kellene tárolni
- Nem dokumentálható minden MI algoritmus
  - Legyen inkább megmagyarázható vagy érthető
- Pl: fogyasztóvédelem, tervezői felelősség, hitelhez jutás, álláshoz jutás, stb.

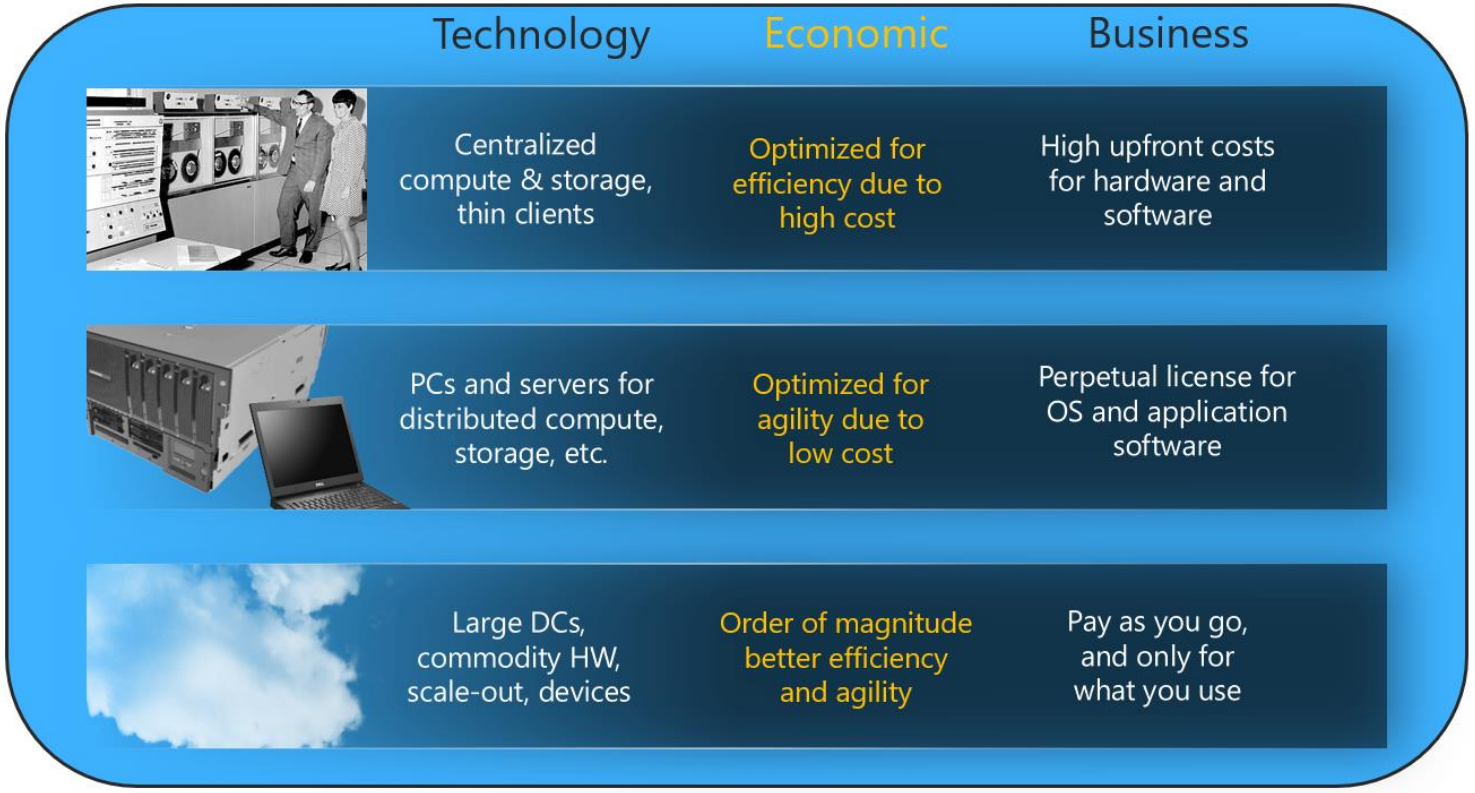
Microsoft

A számítási felhő

Cloud definition



Order of magnitude improvement



Initially, 3 network effects interacted:

1) **Supply Side Economies of Scale**
   Transition from OEM to ODM model

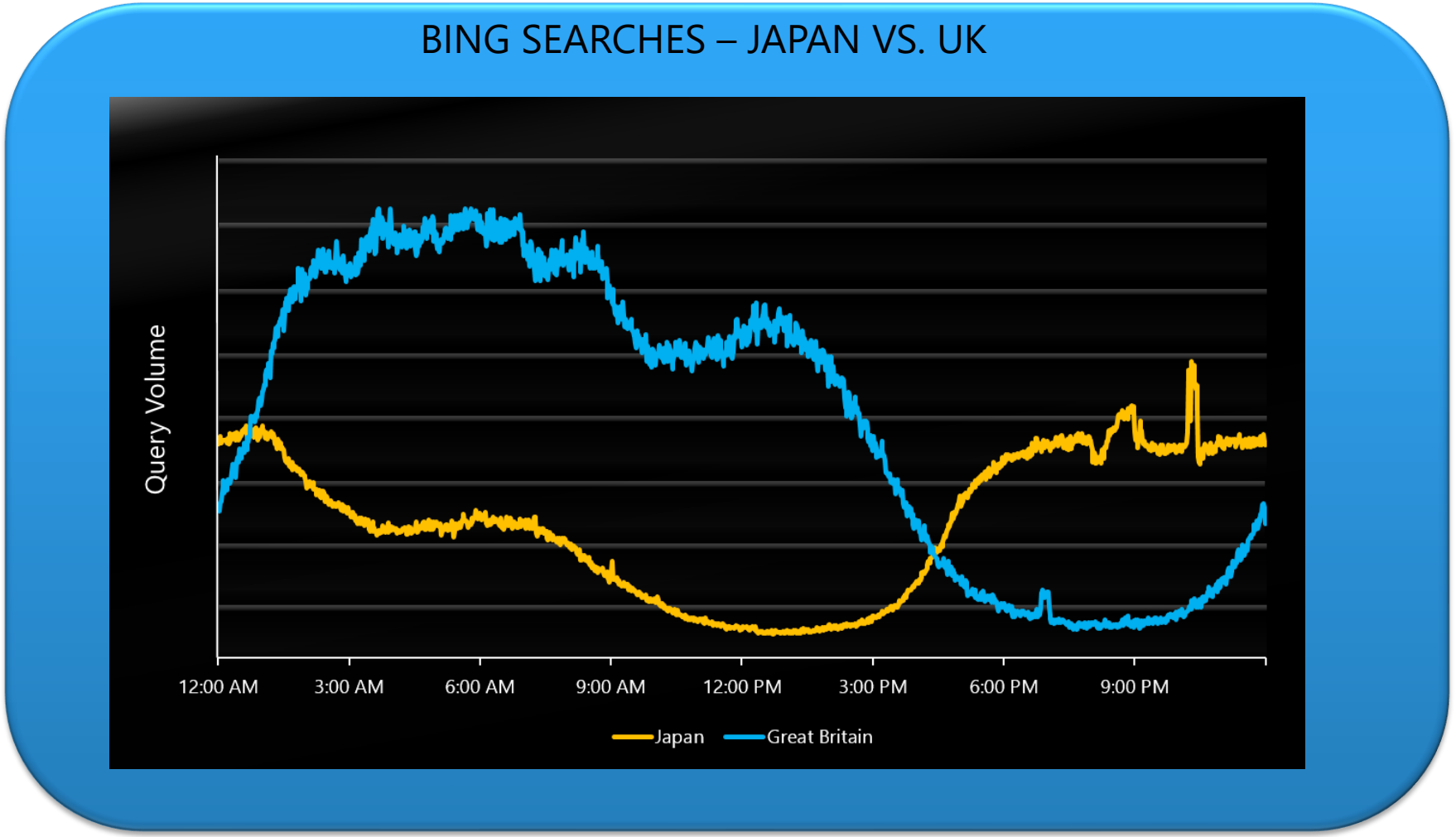
2) **Demand Side Economies of Scale**
   "smoothen the variabilities"

3) **Massive Multitenancy**
   costs grow sublinear
   Cloud native apps break 50yr paradigm

Source: Microsoft Cloud Economics Whitepaper, EU, 2010

Demand Side Economies of Scale

Time of Day

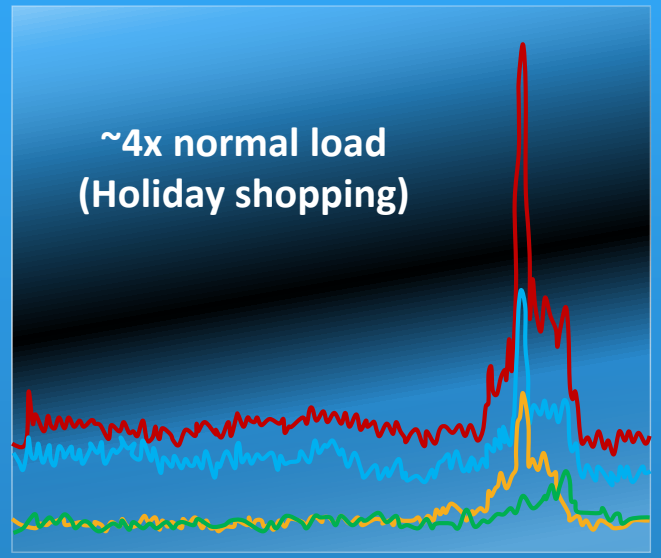


Source: Microsoft

Demand Side Economies of Scale

Industry Variability

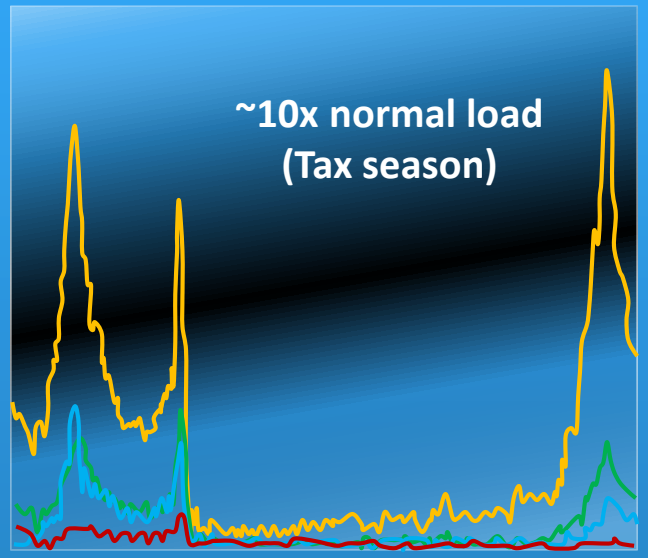
Retail Market



January Dec/Jan

- target.com
- walmart.com
- toysrus.com
- barnesandnoble.com

Income Tax



January Dec/Jan

- turbotax.com
- taxcut.com
- hrblock.com
- taxact.com

Source: Alexa

What is HyperScale ?

TOP 10 Sites for June 2019

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
3	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**Frontera** - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR , Dell EMC Texas Advanced Computing Center/Univ. of Texas United States	448,448	23,516.4	38,745.9	

Cloud data center:
~ 100.000 dual-socket servers
All include FPGA's

9.288 dual socket servers + GPU

4.320 dual socket servers + GPU

40.960 single-socket servers

16.000 dual-socket servers +GPU

8.064 dual-socket servers + GPU

Silicon Technologies for Computing

More Flexible

More Efficient

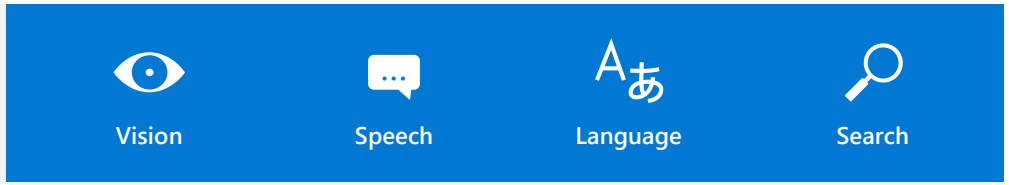
	Perf/W			
CPUs	1X	Today's standard, most programmable, good for services changing rapidly	Conventional programming	C/C++
Manycore CPUs	3X	Many simple cores (10s to 100s per chip), useful if software can be fine-grain parallel, difficult to maintain.		
GPUs	5-30X	Good for data parallelism by merged threads (SIMD), High memory bandwidth, power hungry	Alternative programming	CUDA
FPGAs	5-30X	Most radical fully programmable option. Good for streaming/irregular parallelism. Power efficient but currently need to program in H/W languages.		Verilog
Structured ASICS	20-100X	Lower-NRE ASICs with lower performance/efficiency. Includes domain-specific (programmable) accelerators.	Can't change functionality	Verilog
Custom ASICs	> 100X	Highest efficiency. Highest NRE costs. Requires high volume. Good for functions in very widespread use that are stable for many years.		

Microsoft

Microsoft AI platform

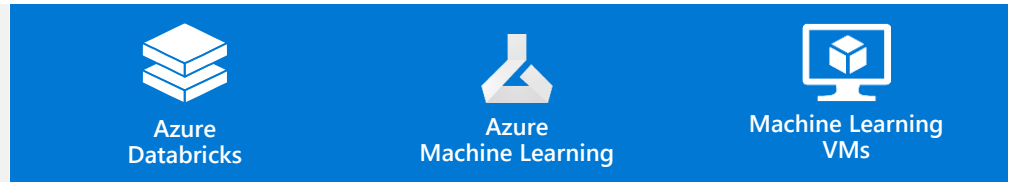
Machine Learning on Azure

Sophisticated pretrained models
To simplify solution development

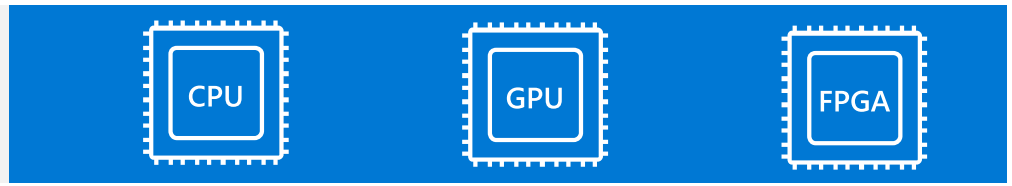


Popular frameworks
To build advanced deep learning solutions

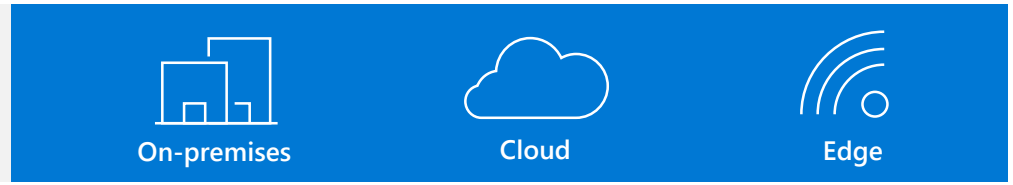
Productive services
To empower data science and development teams



Powerful infrastructure
To accelerate deep learning

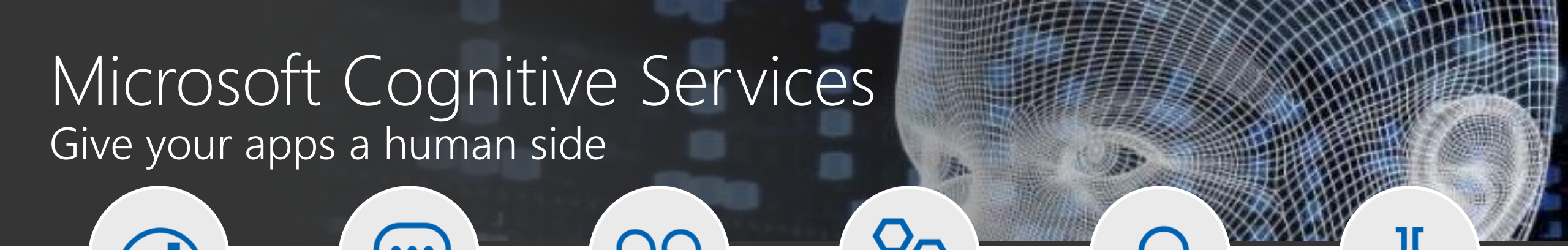


Flexible deployment
To deploy and manage models on intelligent cloud and edge



Microsoft Cognitive Services

Give your apps a human side



Vision

Computer Vision

Content Moderator

Emotion/Face

Video Indexer

Speech

Bing Speech

Language

Bing Spell Check

Text Analytics

Translator Text & Speech

Knowledge

QnA Maker

Search

Bing Autosuggest

Bing Image Search

Bing News Search

Bing Video Search

Bing Web Search

Bing Entity Search

Labs

CUSTOMIZATION

Custom Vision Service

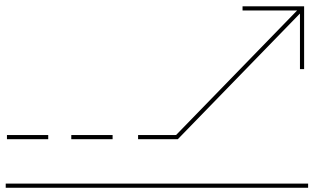
Custom Speech Service

Language Understanding

Custom Decision Service

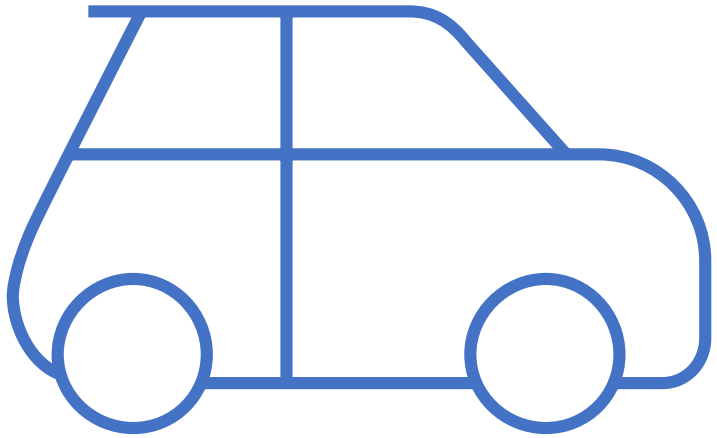
Bing Custom Search

Microsoft



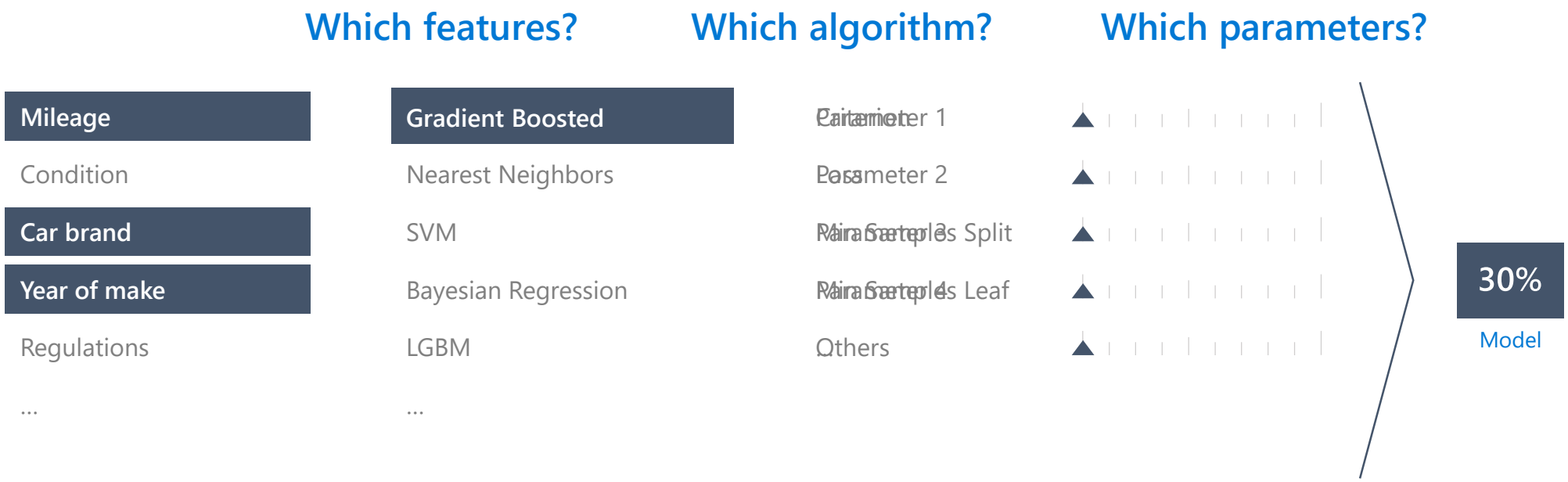
Azure Automated Machine Learning ‘simplifies’ the creation and selection of the optimal model

Machine Learning Problem Example



How much is this car worth?

Model Creation Is Typically Time-Consuming



Model Creation Is Typically Time-Consuming

Which features?

- Mileage
- Condition
- Car brand
- Year of make
- Regulations
- …

Which algorithm?

- Gradient Boosted
- Nearest Neighbors
- SVM
- Bayesian Regression
- LGBM
- …

Which parameters?

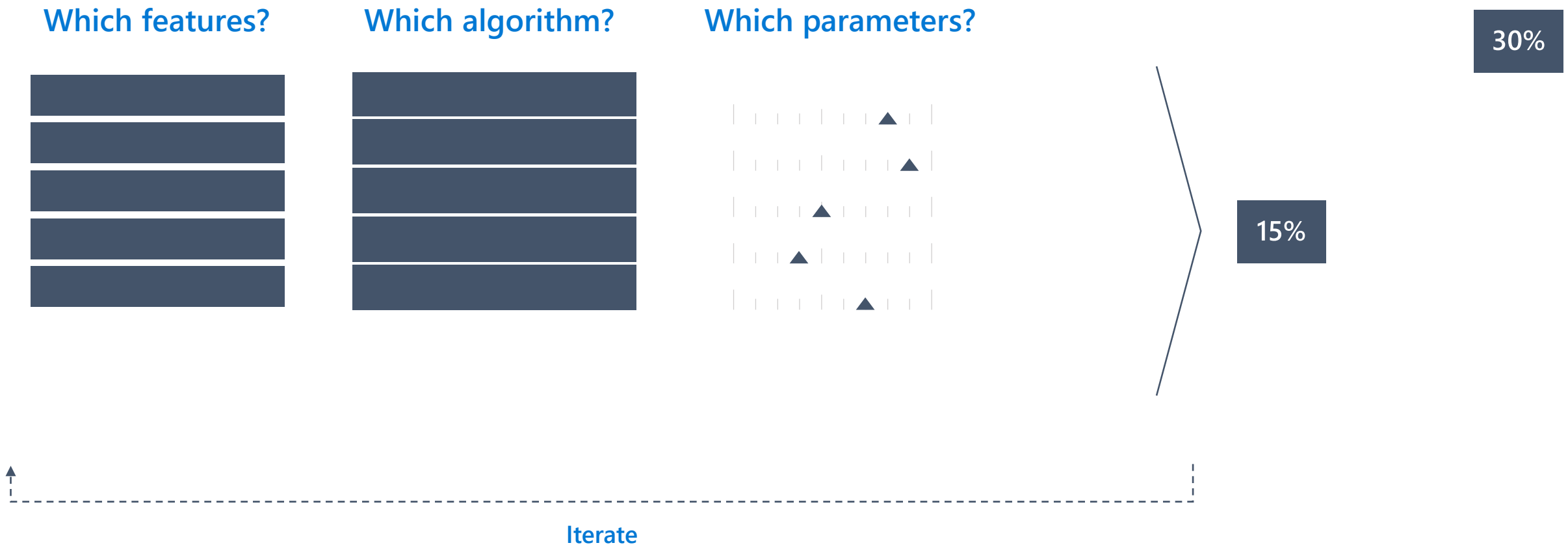
- Number of Neighbors
- Weights
- Min Samples Split
- Min Samples Leaf
- Others

30%

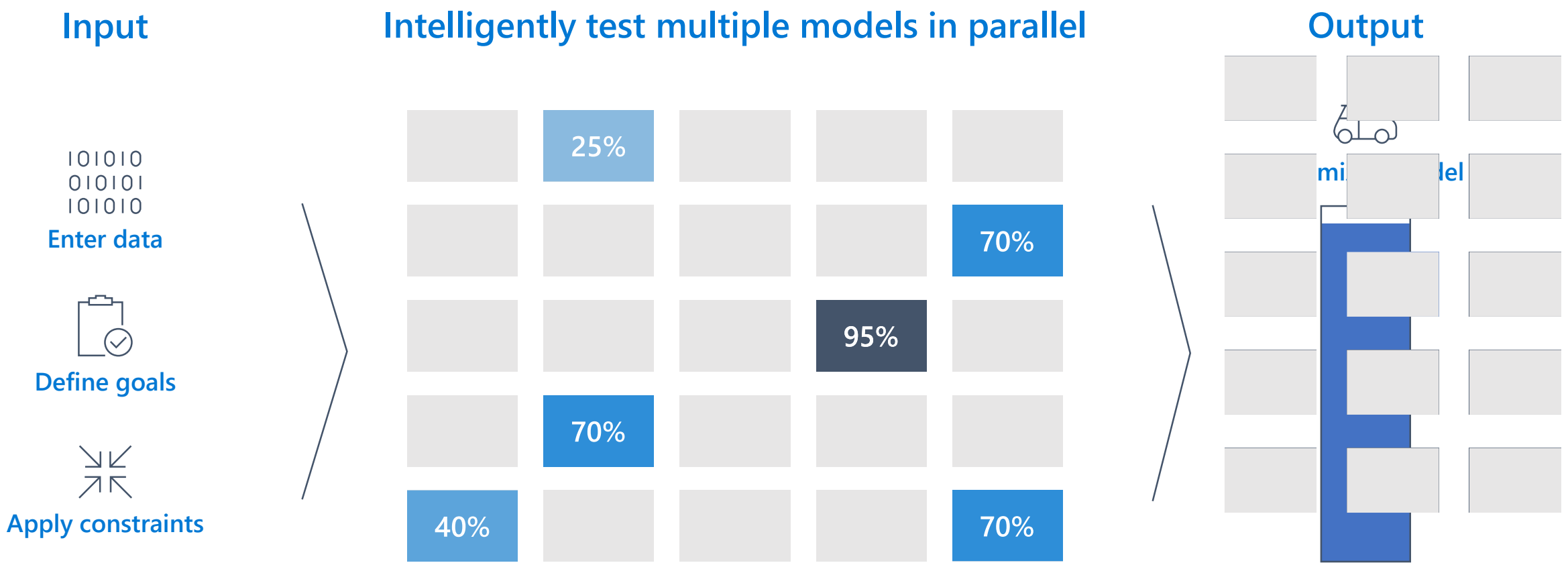
Model

Iterate

Model Creation Is Typically Time-Consuming



Automated ML Accelerates Model Development



The power of Deep Learning on FPGA

Performance

Excellent inference at low batch sizes

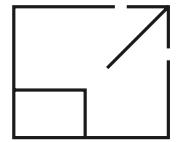
Ultra-low latency | 10x < CPU/GPU

Flexibility

Rapidly adapt to evolving ML

Inference-optimized numerical precision

Exploit sparsity, deep compression

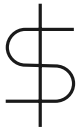


Scale

World's largest cloud investment in FPGAs

Multiple Exa-Ops of aggregate AI capacity

Runs on Microsoft's scale infrastructure



Low cost

$0.21/million images on Azure FPGA

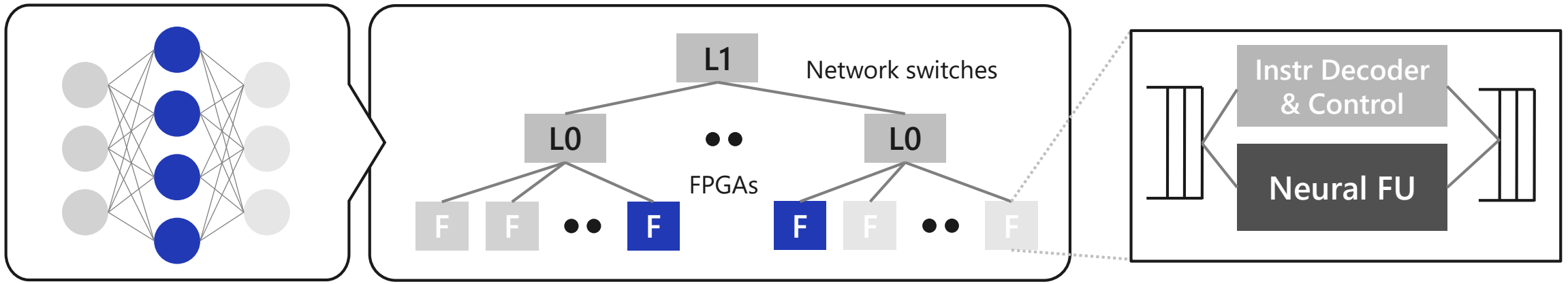
Project BrainWave

A Scalable FPGA-Powered DNN Serving Platform

Fast: Ultra-low latency, high-throughput serving of DNN models at low batch sizes

Flexible: Future proof, adaptable to fast-moving AI space and evolving model types

Friendly: Turnkey deployment of TensorFlow/CNTK/Caffe/etc.



Pretrained DNN Model
in TensorFlow, CNTK, etc.

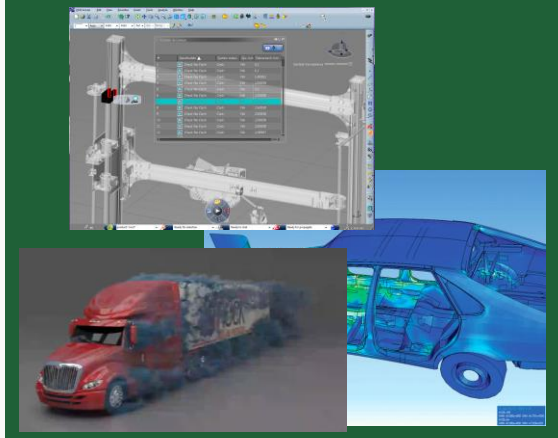
Scalable DNN Hardware
Microservice

BrainWave
Soft DPU

Azure GPUs

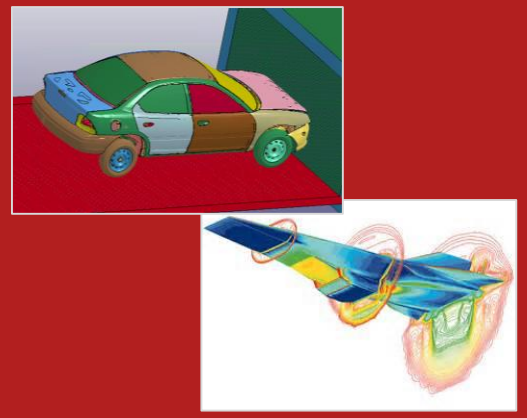
Broad Range of GPU Scenarios

Visualization

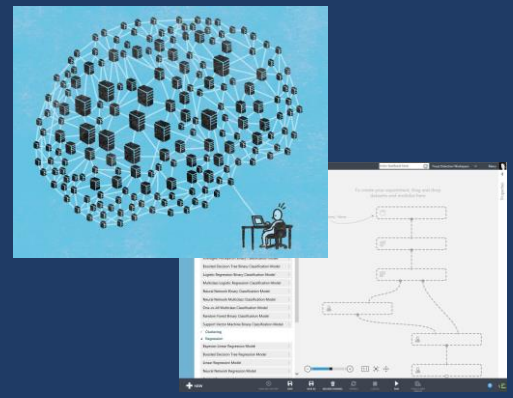


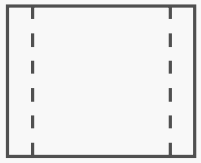
Rendering

HPC/Simulation



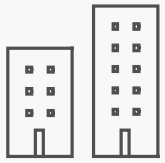
Deep-Learning/AI



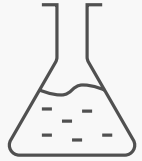


Media, Entertainment & Gaming

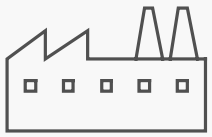
Healthcare & Research



Financial Services

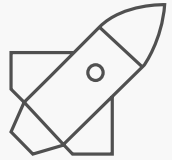


Oil & Gas and Sciences

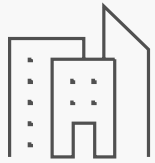


Manufacturing

Automotive



Aerospace



Retail

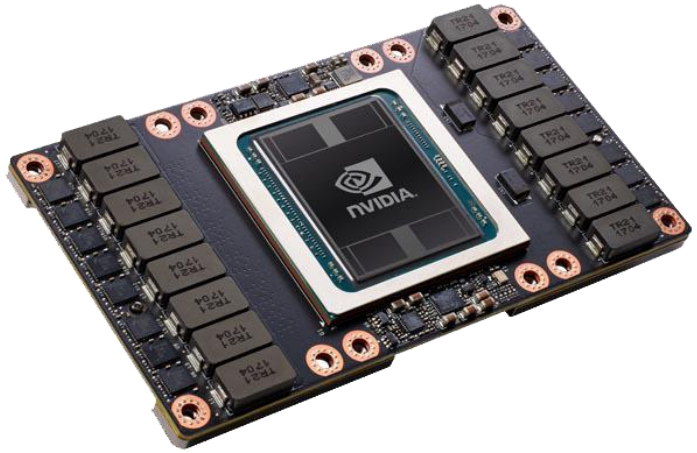
http://aka.ms/AzureNDv2GPUs

In Preview

ND v2 – Volta Generation GPU Compute

- Excellent for accelerating machine learning and HPC workloads
- Volta SXM GPU instances – 8X NVIDIA V100 GPUs interconnected with NVLink mesh
- Tensor Core technology to deliver over 100 TFLOPS of deep learning performance
- Skylake based processor with premium storage support (SSD backed)
- Specs:
  - 640 NVIDIA Tensor Core
  - FP64 - 7.8 TFLOPS of double precision floating point performance
  - FP32 – 15.7 TFLOPS of single precision performance
  - GPU Memory 16 GB
  - 300 GB/s GPU interconnect through NVLink

	ND40s_v3
Cores	40 cores
GPU	8 x V100 SXM
Memory	768 GB
Local Disk	~1.3 TB SSD
Network	Azure Network + NVLink GPU interconnect



http://aka.ms/AzureNVv2GPUs

In Preview

NV v2 – Updated GPU Visualization Platform

- Get faster results for the your graphic intensive 2D and 3D applications with visualization optimized GPU instances featuring NVIDIA Tesla M60 GPUs
- Broadwell based CPU processor with doubled memory from previous generation (up to 448 GB)
- Premium storage support (SSD backed)
- Grid license included with each GPU instance
- Specs:
  - 2048 NVIDIA CUDA cores per GPU
  - 36 H.264 1080p30 streams
  - GPU Memory 8 GB/GPU

NVIDIA Quadro Virtual Workstation Driver

Azure NV/NVIDIA Tesla M60 GPUs

Azure Virtual Machines

	NV6s_v2	NV12s_v2	NV24s_v2
Cores	6	12	24
GPU	1 x M60	2 x M60	4 x M60
Memory	112 GB	224 GB	448 GB
Local Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network
GRID Licenses	1	2	4

Full Lineup of GPU Families

GPU Accelerated Compute Family			
	NC	NC v2	NC v3
Cores	6, 12, 24	6, 12, 24	6, 12, 24
GPU	1, 2, or 4 K80 GPU	1, 2, or 4 P100 GPU	1, 2, or 4 V100 GPU
Memory	56/112/224 GB	112/224/448 GB	112/224/448 GB
Local Disk	~380/~680/~1.5 TB SSD	~700/~1.4/~3 TB SSD	~700/~1.4/~3 TB SSD
Network	Azure Network + InfiniBand (largest size only)		

GPU Accelerated Deep Learning Family			GPU Visualization Family	
	ND v1	ND v2	NV v1	Nv v2
Cores	6, 12, 24	40	6, 12, 24	6, 12, 24
GPU	1, 2, or 4 P40 GPU	8 V100 SXM GPU	1, 2, or 4 M60 GPU	1, 2, or 4 M60 GPU
Memory	112/224/448 GB	768 GB	56/112/224 GB	112/224/448 GB
Local Disk	~700/~1.4/~3 TB SSD	~1.3 TB SSD	~380/~680/~1.5 TB SSD	~700/~1.4/~3 TB SSD
Network	Azure Network + InfiniBand (largest size only)	Azure Network + NVLink GPU interconnect	Azure Network	Azure Network

Azure Batch

Azure Batch

Batch pools

Configure and create VMs to cater for any scale: tens to thousands

Automatically scale the number of VMs to maximize utilization

Choose the VM size most suited to your application

Batch jobs and tasks

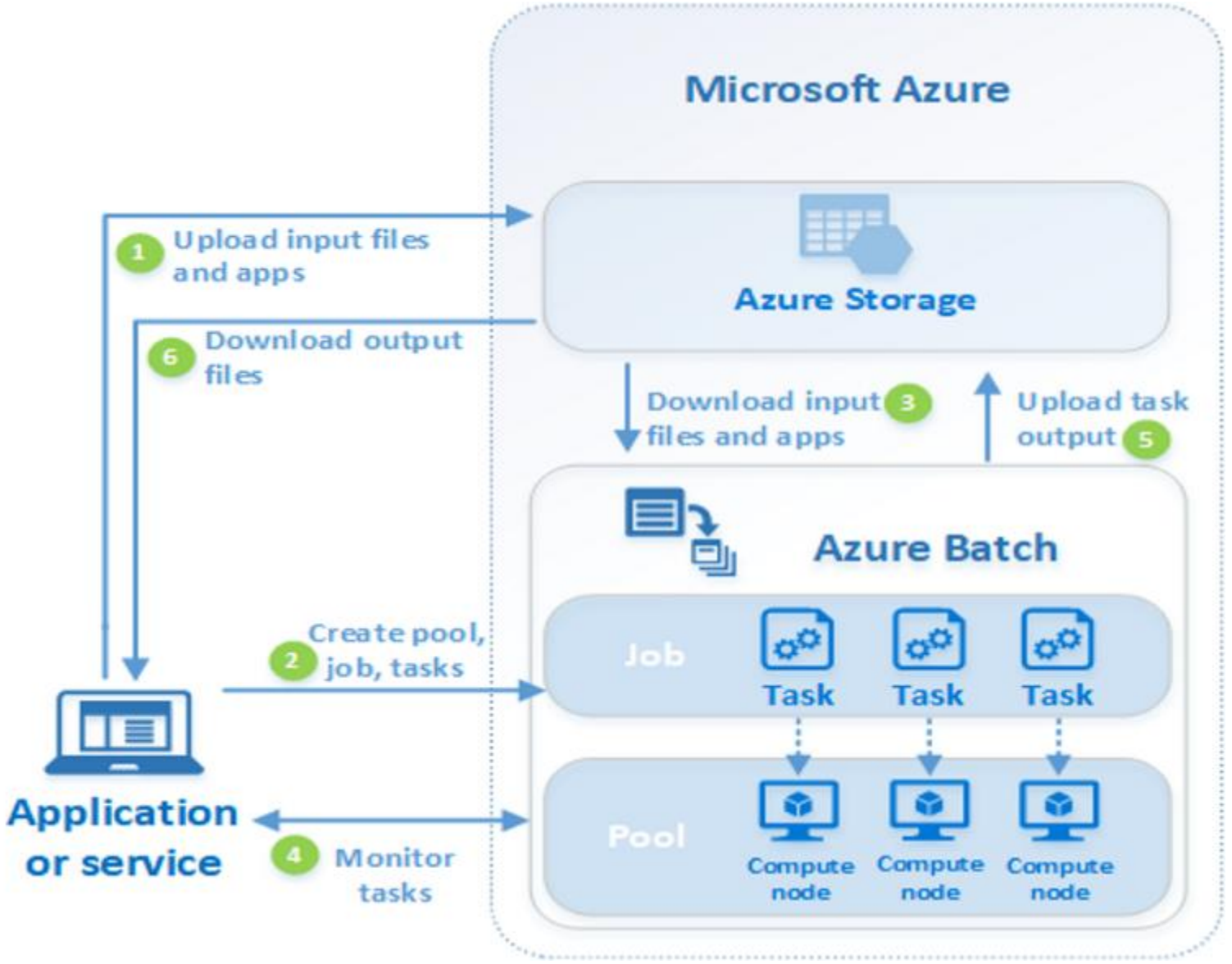
Task is a unit of execution; task = application command line (EXE, BAT, CMD, PS1, etc.)

Jobs are created and tasks are submitted to a pool. Next, tasks are queued and assigned to VMs

Any application, any execution time; run applications unchanged

Automatic detection and retry of frozen or failing tasks

How Azure Batch works – intrinsically parallel example



Azure Batch Supported development technologies

Command-Line

Azure CLI

Azure PowerShell

Languages

.NET

Java

Node.js

Python

REST

Batch Service REST API

Batch Management REST API

Azure Batch Step-by-Step Tutorials and Training

- **Learn how to run compute-intensive workloads on Batch.**
  - Parallel file processing with .NET SDK
  - Parallel file processing with Python SDK
  - Scene rendering with Arnold
  - Parallel R simulation
  - Tutorial: Trigger a Batch job using Azure Functions

- **Free Pluralsight Video Training**
  - Microsoft Azure Batch, Getting Started

AI a **közjó szolgálatában**

AI a bolygónkért

AI a fogyatékossággal élő emberekért

AI a humanitárius segítségnyújtásért


Microsoft
Köszönöm a figyelmet!
https://www.microsoft.com/AI