

The background is a dark, abstract composition. It features a dense field of small, glowing blue and white spheres, resembling a particle simulation or a data visualization. Overlaid on this are several translucent, wavy, ribbon-like structures that flow across the frame. The overall aesthetic is high-tech and futuristic, with a focus on light and motion.

Azure AI Infrastructure

Running AI Workloads at Scale

Jarek Kazmierczak
MTC Silicon Valley

A visualization of mixed reality showing a mountain range with a digital point cloud overlay. The point cloud consists of numerous small, glowing blue and white dots that form the shape of the mountains, appearing to be projected onto the real-world scene.

Mixed
Reality

A visualization of artificial intelligence featuring a realistic image of Mount Everest. Overlaid on the image are two data points: 'Elevation 25,643'' with a line pointing to the peak, and 'Temperature -22°C' with a line pointing to a small white box on the mountain's slope.

Artificial
Intelligence

A visualization of quantum computing with a dark blue background filled with abstract geometric patterns. These include a network of interconnected nodes and lines, a complex polyhedron with a blue and white striped ribbon passing through it, and various mathematical symbols like y , z , ϕ , and ψ scattered throughout.

Quantum
Computing



Amplifying human ingenuity with intelligent technology



Reasoning

Learn and form conclusions
with imperfect data



Understanding

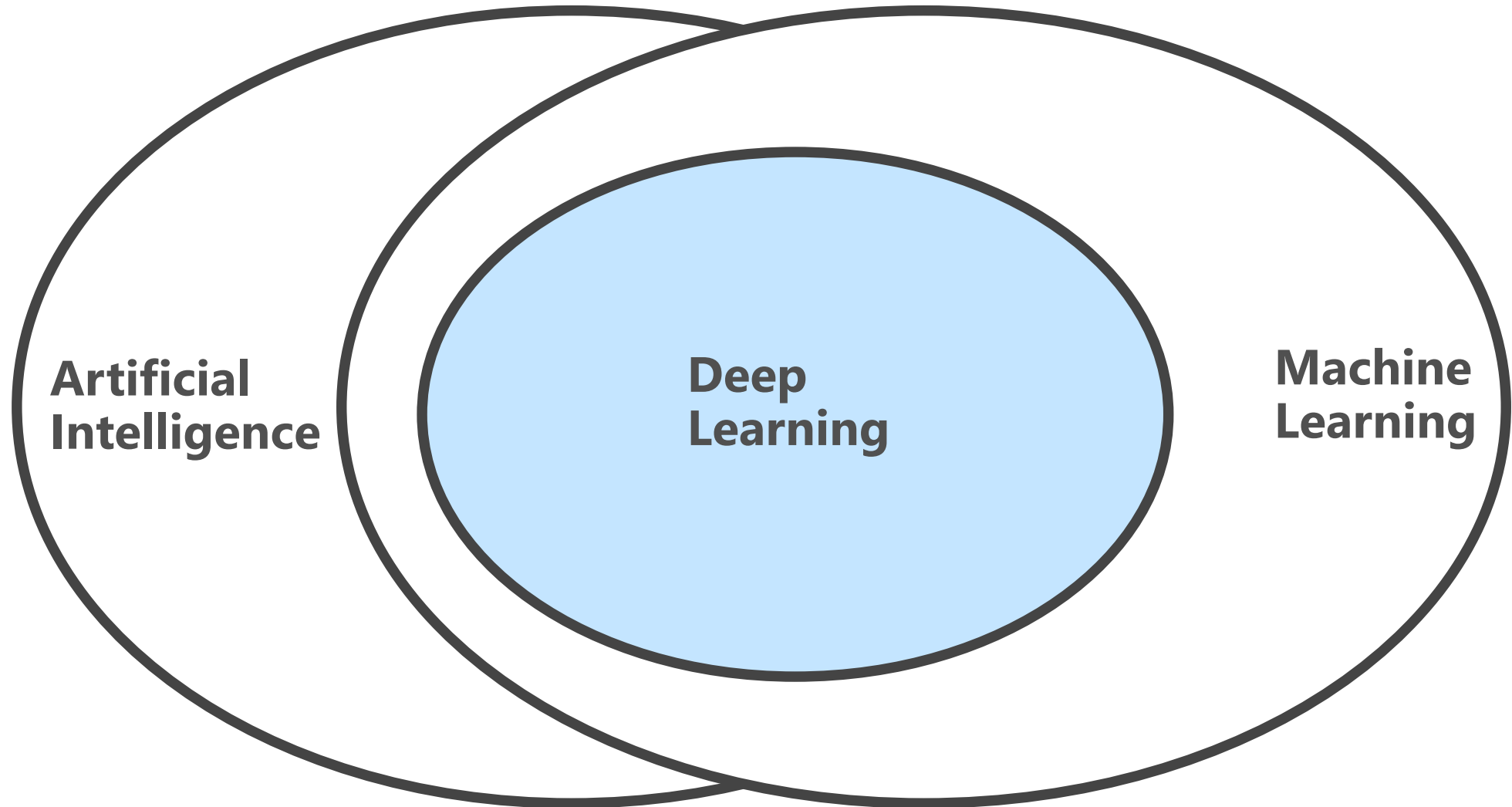
Interpret meaning of data
including text, voice, images



Interacting

Interact with people
in natural ways

Artificial Intelligence, Machine Learning and Deep Learning

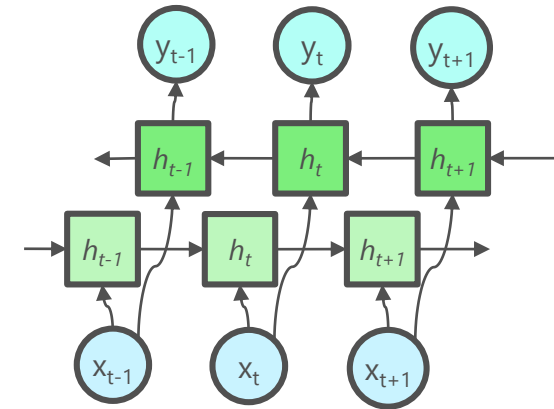


The Rise of Deep Learning in ML

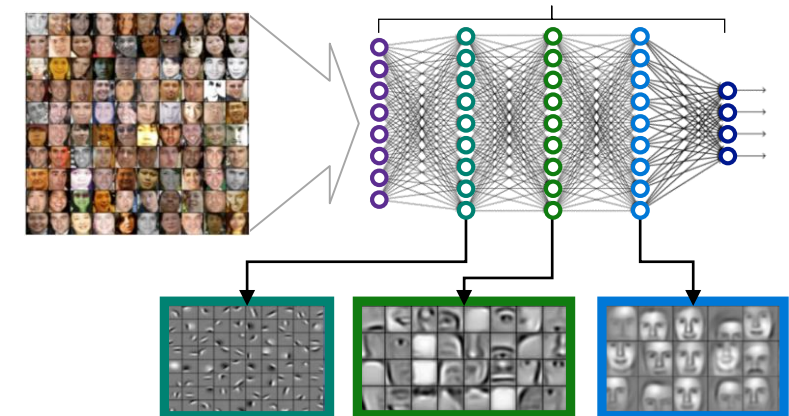
Deep neural networks have enabled major advances in machine learning and AI

Computer vision
Language translation
Speech recognition
Question answering
And more...

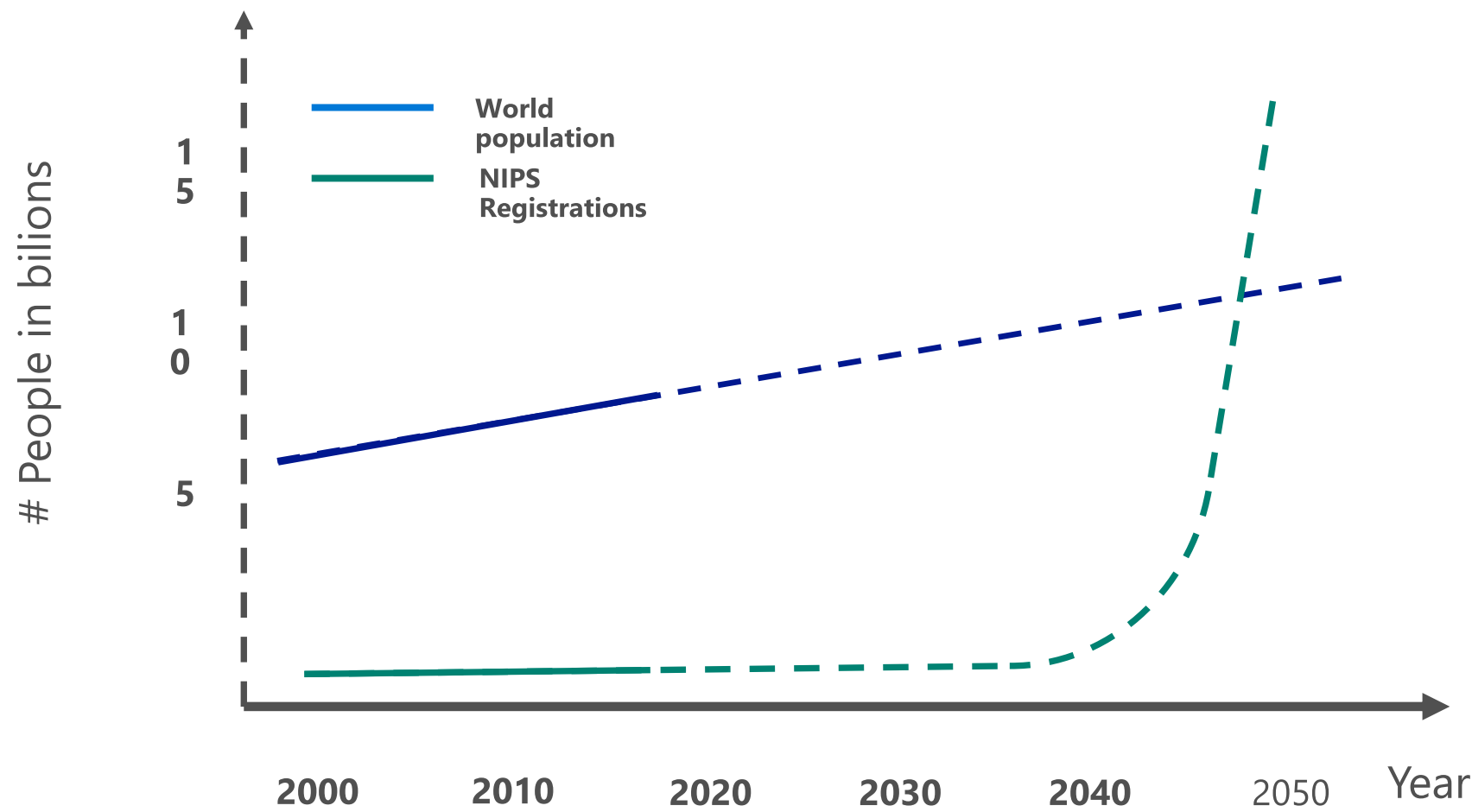
Recurrent Neural Networks



Convolutional Neural Networks



Deep Learning Hype



Microsoft AI Platform

Services

CONVERSATIONAL AI

Bot Framework

TRAINED SERVICES

Cognitive Services

CUSTOM SERVICES

Azure Machine Learning

Infrastructure

DATA

Cosmos
DB

SQL
DB

SQL
DW

Data
Lake

Spark

DSVM

Batch
AI

ACS

Edge

COMPUTE

CPU, FPGA, GPU

Tools

CODING & MANAGEMENT TOOLS

VS Tools
for AI

Azure ML
Studio

Azure ML
Workbench

Others (PyCharm, Jupyter Notebooks...)

DEEP LEARNING FRAMEWORKS

3rd Party

Cognitive
Toolkit

TensorFlow

Caffe

Others (Scikit-learn, MXNet, Keras,
Chainer, Gluon...)

Microsoft AI Platform

Services

CONVERSATIONAL AI

Bot Framework

TRAINED SERVICES

Cognitive Services

CUSTOM SERVICES

Azure Machine Learning

Infrastructure

DATA

Cosmos
DB

SQL
DB

SQL
DW

Data
Lake

Spark

DSVM

Batch
AI

ACS

Edge

COMPUTE

CPU, FPGA, GPU

Tools

CODING & MANAGEMENT TOOLS

VS Tools
for AI

Azure ML
Studio

Azure ML
Workbench

Others (PyCharm, Jupyter Notebooks...)

DEEP LEARNING FRAMEWORKS

3rd Party

Cognitive
Toolkit

TensorFlow

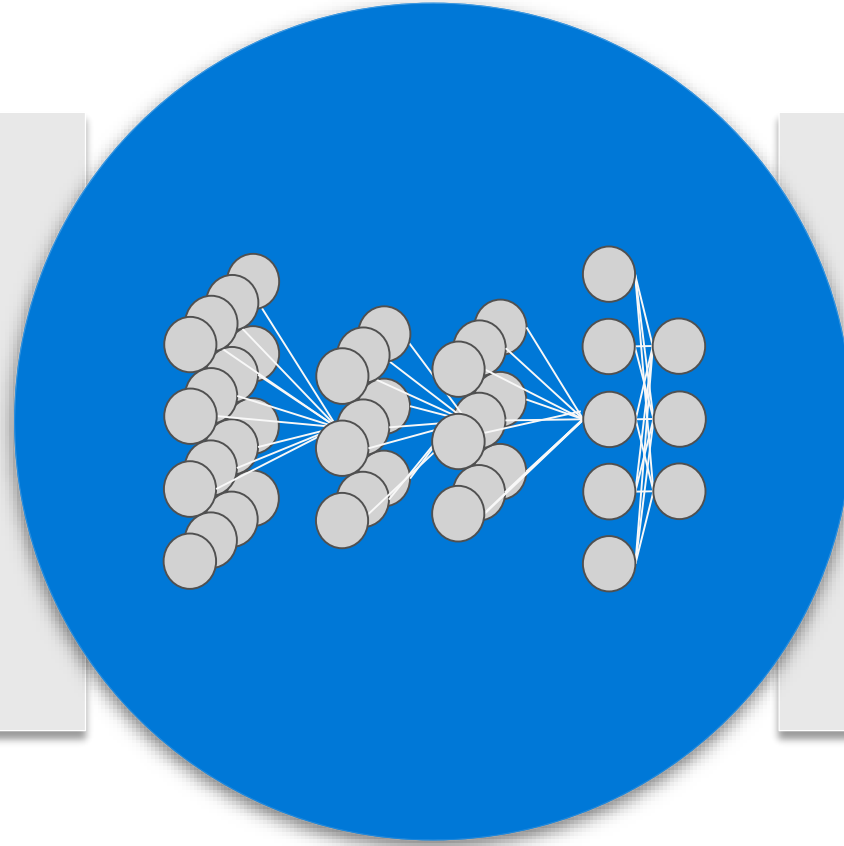
Caffe

Others (Scikit-learn, MXNet, Keras,
Chainer, Gluon...)

Deep Learning Demands Compute Power

TRAINING

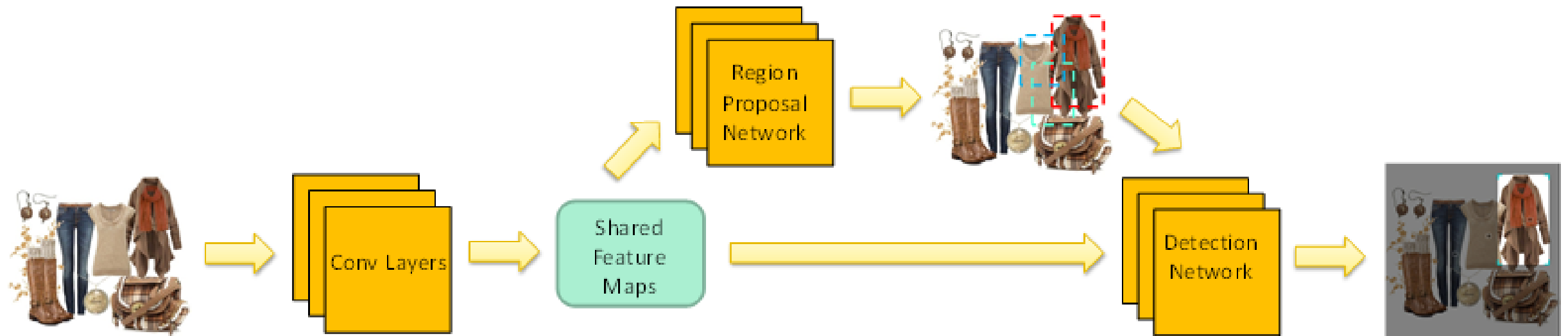
Billions of TFLOPS/TOPS per training run



INFERENCE

Billions of FLOPS/OPS per inference

Massive scale inference



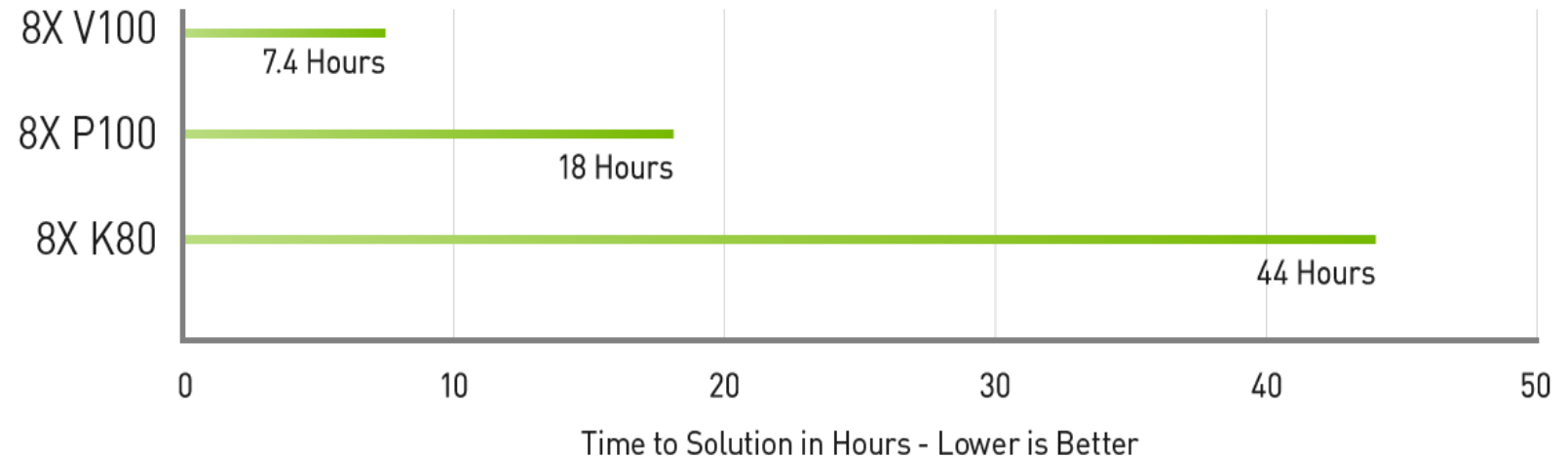
Computational capacity – CPU, GPU, ASICs

Processor type	Operations per cycle
CPU	a few
CPU with vector extensions (e. AVX-512)	tens
GPU	tens of thousands
ASIC/FPGA	hundreds of thousands

GPU Power

- Tesla P40
 - 12 Teraflops per second
 - 47 INT8 operations per second (TOPS)
 - 24 GB GPU memory
 - 346 GB/s Memory Bandwidth
- Volta 100
 - 120 Teraflops per second (Deep Learning)
 - 640 Tensor Cores
 - 900 GB/s Memory Bandwidth

Deep Learning Training in One Workday



Server Config: Dual Xeon E5-2699 v4, 2.6GHz | 8x Tesla K80, Tesla P100 or Tesla V100 | V100 performance measured on pre-production hardware. | ResNet-50 Training on Microsoft Cognitive Toolkit for 90 Epochs with 1.28M ImageNet dataset

Every Deep
Learning
Framework is GPU
Accelerated

TORCH



CAFFE



THEANO



MATCONVNET



MOCHA.JL



PURINE



MINERVA



MXNET*



BIG SUR



TENSORFLOW



WATSON



CNTK



Compute Virtual Machines (NC)

	NC6	NC12	NC24	NC24r
Cores	6	12	24	24
GPU	1 K80 GPU (1/2 Physical Card)	2 K80 GPUs (1 Physical Card)	4 K80 GPUs (2 Physical Cards)	4 K80 GPUs (2 Physical Cards)
Memory	56 GB	112 GB	224 GB	224 GB
Disk	~380 GB SSD	~680 GB SSD	~1.5 TB SSD	~1.5 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand



Next-Gen GPU Compute VM: NC_v2

	NC6s_v2	NC12s_v2	NC24s_v2	NC24rs_v2
Cores	6	12	24	24
GPU	1 x P100 GPU	2 x P100 GPU	4 x P100 GPU	4 x P100 GPU
Memory	112 GB	224 GB	448 GB	448 GB
Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand



Next-Gen GPU Deep Learning VM: ND

	ND6s	ND12s	ND24s	ND24rs
Cores	6	12	24	24
GPU	1 x P40	1 x P40	4 x P40	4 x P40
Memory	112 GB	224 GB	448 GB	448 GB
Disk	~700 GB SSD	~1.4 TB SSD	~3 TB SSD	~3 TB SSD
Network	Azure Network	Azure Network	Azure Network	InfiniBand



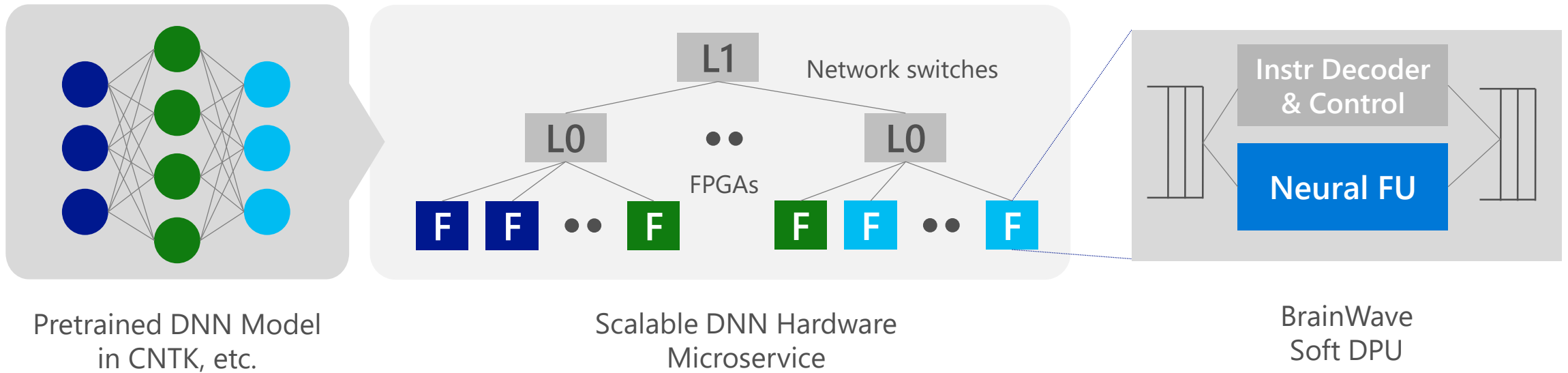
Project BrainWave

A Scalable FPGA-powered DNN Serving Platform

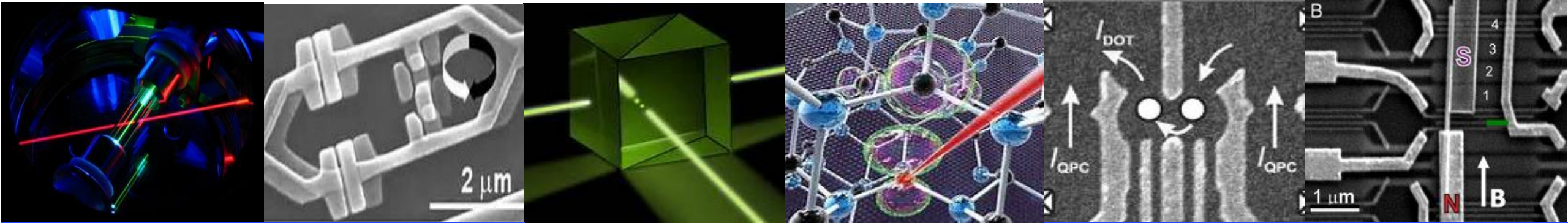
Fast: ultra-low latency, high-throughput serving of DNN models at low batch sizes

Flexible: adaptive numerical precision and custom operators

Friendly: turnkey deployment of CNTK/Caffe/TF/etc



Quantum hardware technologies



Ion
traps

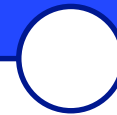
Super-
conductors

Linear
optics

NV
centers

Quantum
dots

Topological



Microsoft AI Platform

Services

CONVERSATIONAL AI

Bot Framework

TRAINED SERVICES

Cognitive Services

CUSTOM SERVICES

Azure Machine Learning

Infrastructure

DATA

Cosmos
DB

SQL
DB

SQL
DW

Data
Lake

Spark

DSVM

Batch
AI

ACS

Edge

COMPUTE

CPU, FPGA, GPU

Tools

CODING & MANAGEMENT TOOLS

VS Tools
for AI

Azure ML
Studio

Azure ML
Workbench

Others (PyCharm, Jupyter Notebooks...)

DEEP LEARNING FRAMEWORKS

3rd Party

Cognitive
Toolkit

TensorFlow

Caffe

Others (Scikit-learn, MXNet, Keras,
Chainer, Gluon...)

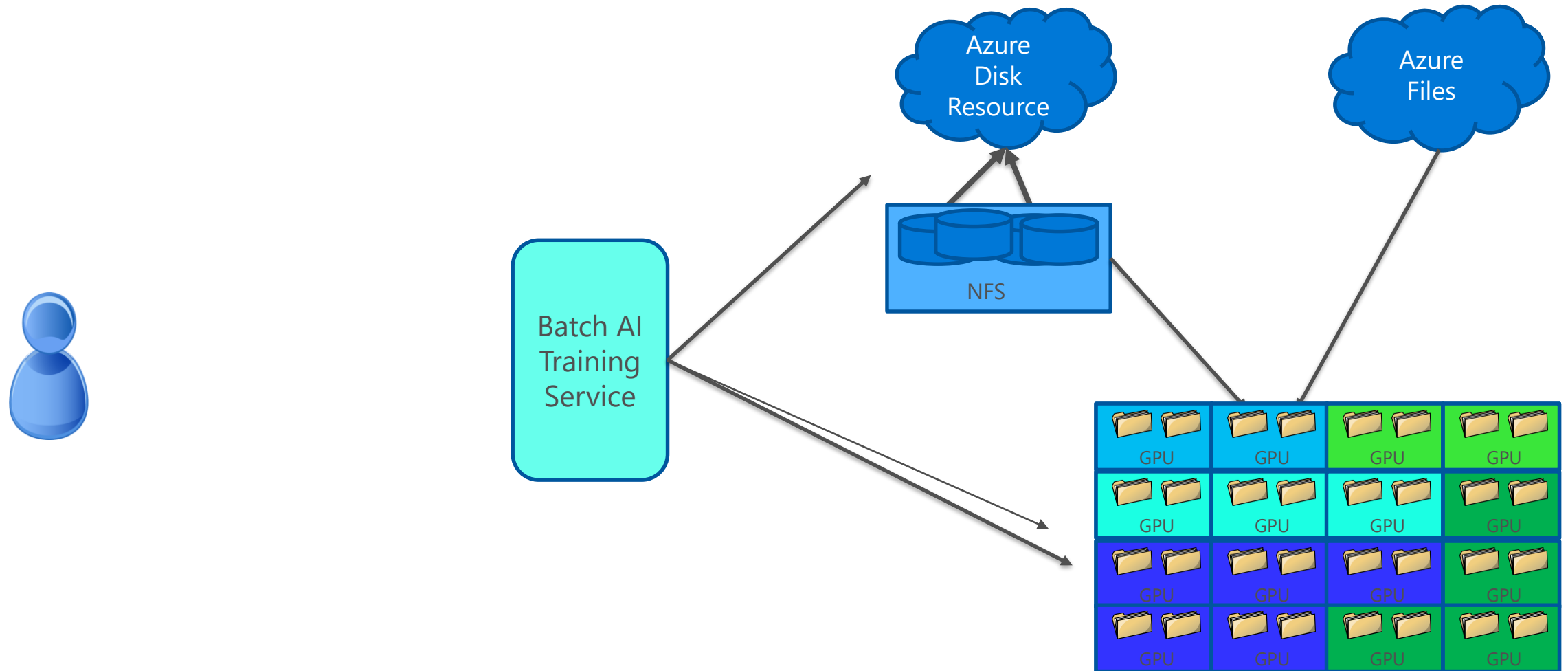
Challenges of Training at Scale

- Deploy virtual machines
- Install Drivers and dependencies
- Manage Cost
- Queue Work
- Handle monitoring and Failures
- Keep things secure
- Training on compliance restricted data
- Cleanup when done

Azure Batch AI Training Service

- Managed Service
- Supports Role Based Access Control
- Hierarchical Quota Management
- Easily Provision VMs at scale
- Load based automatic scaling
- Run experiments in Parallel
- Run in Containers or directly on VM
- Run any toolkit (CNTK, Tensorflow, Caffee, Chainer...)
- Only compute cost. Service is free

Experience Specialized for Learning

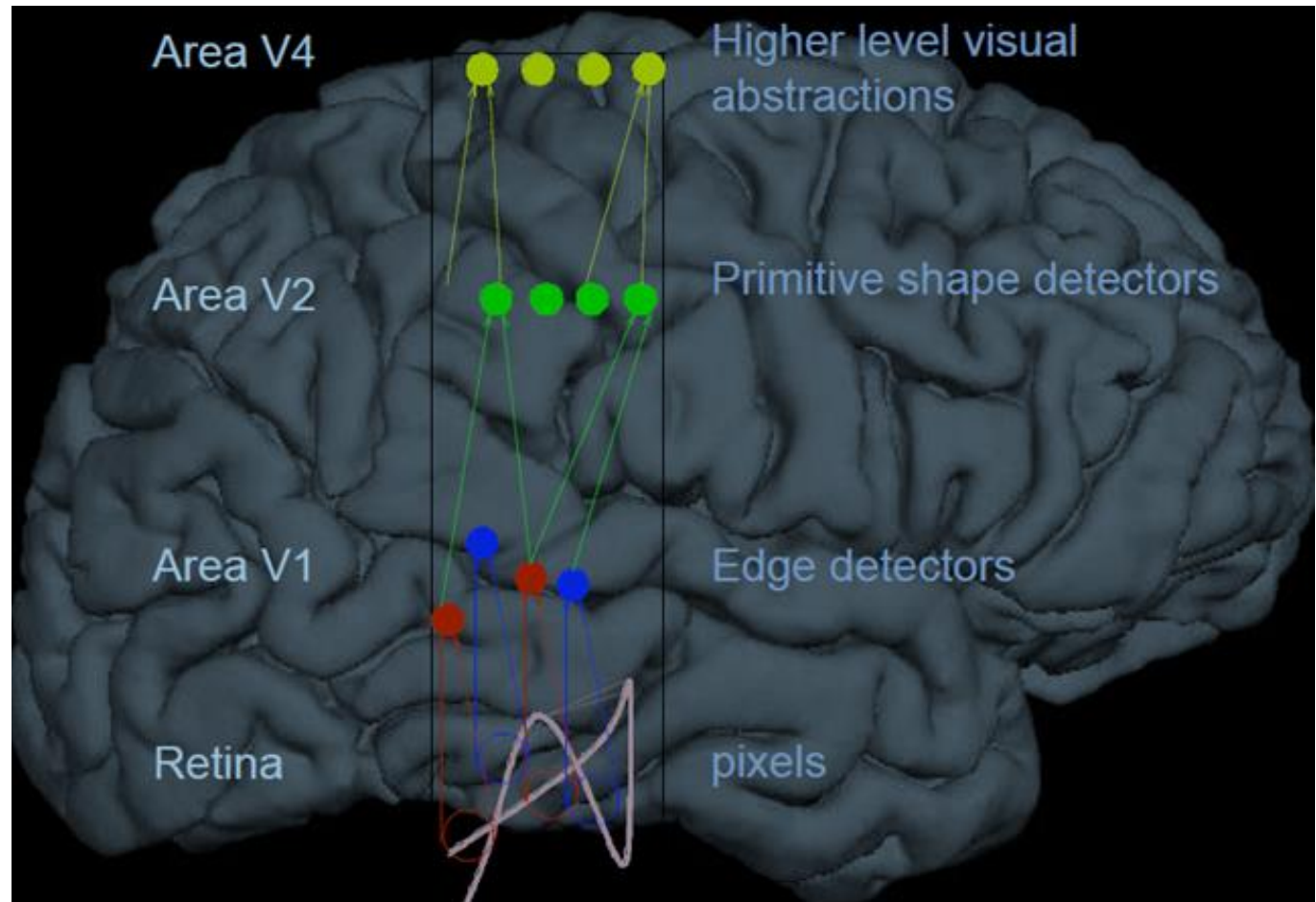




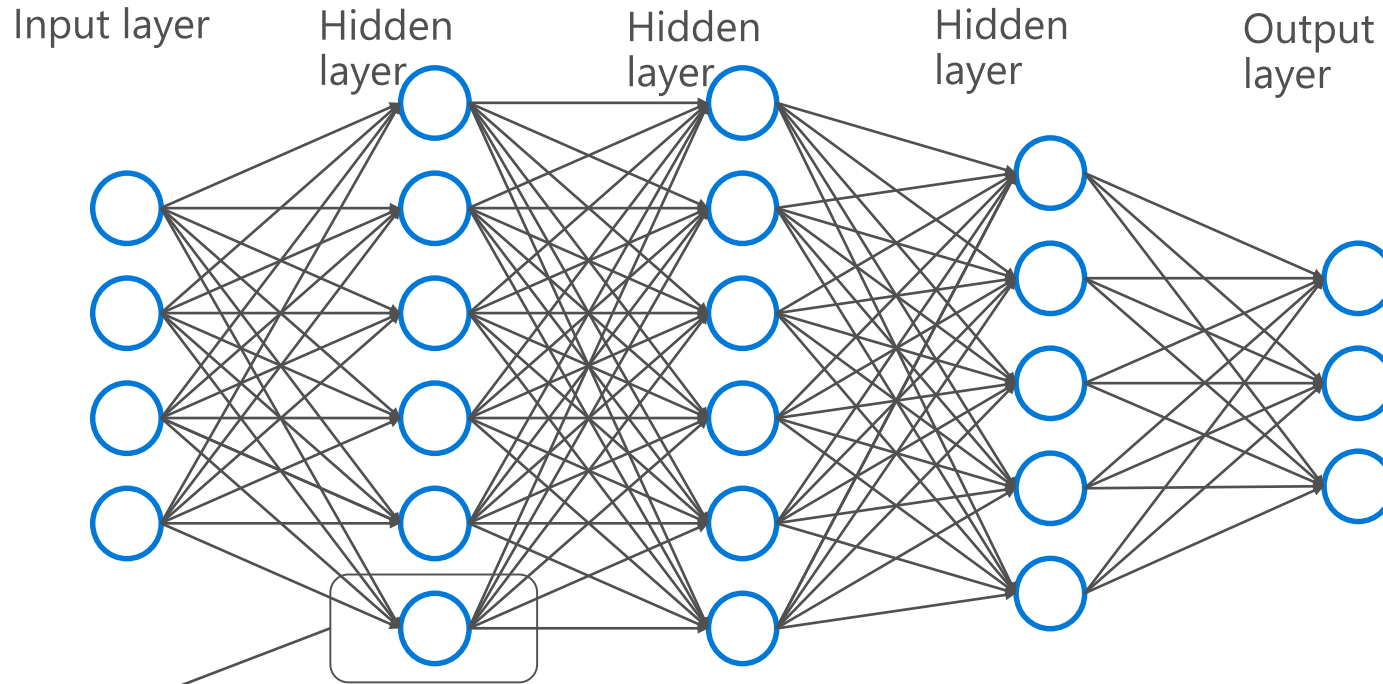
Appendix

Why go deep?

- Deep learning algorithms attempt multiple levels of representation of increasing complexity/abstraction
- Brains have a deep architecture
- Deep Learning has been successful in tasks that have been a challenge for "traditional ML"

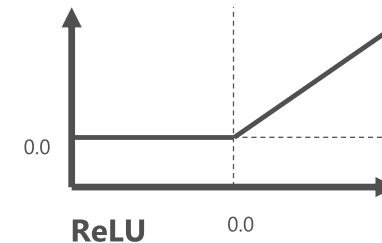
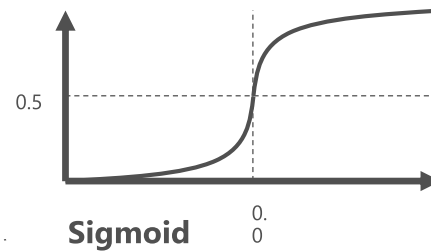


Artificial Neural Networks - ANNs



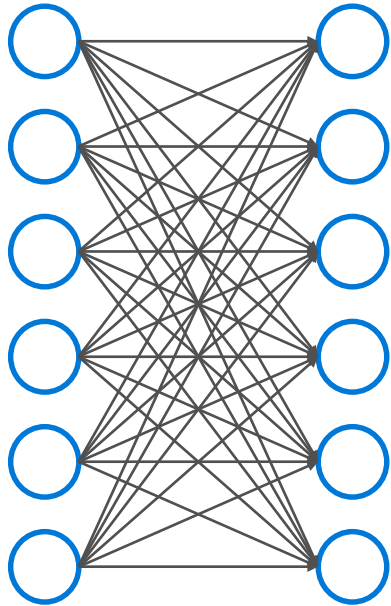
$$a_j = \sigma \left(\sum_{i=1}^n w_i x_i + b_j \right)$$

Activation functions



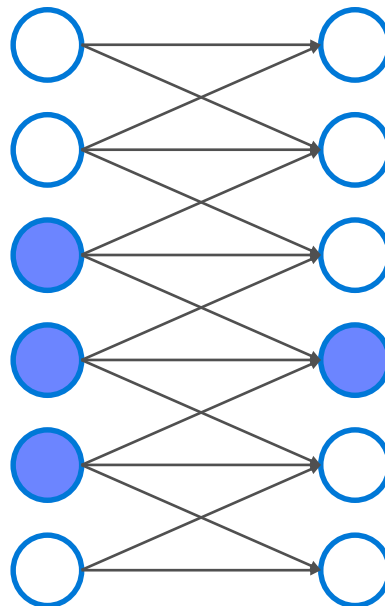
Neurons can connect in various ways ...

"Dense"



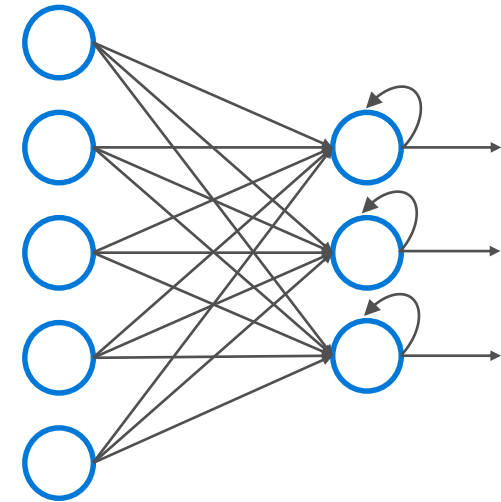
Fully **C**onconnected **N**eural **N**etworks

"Sparse"



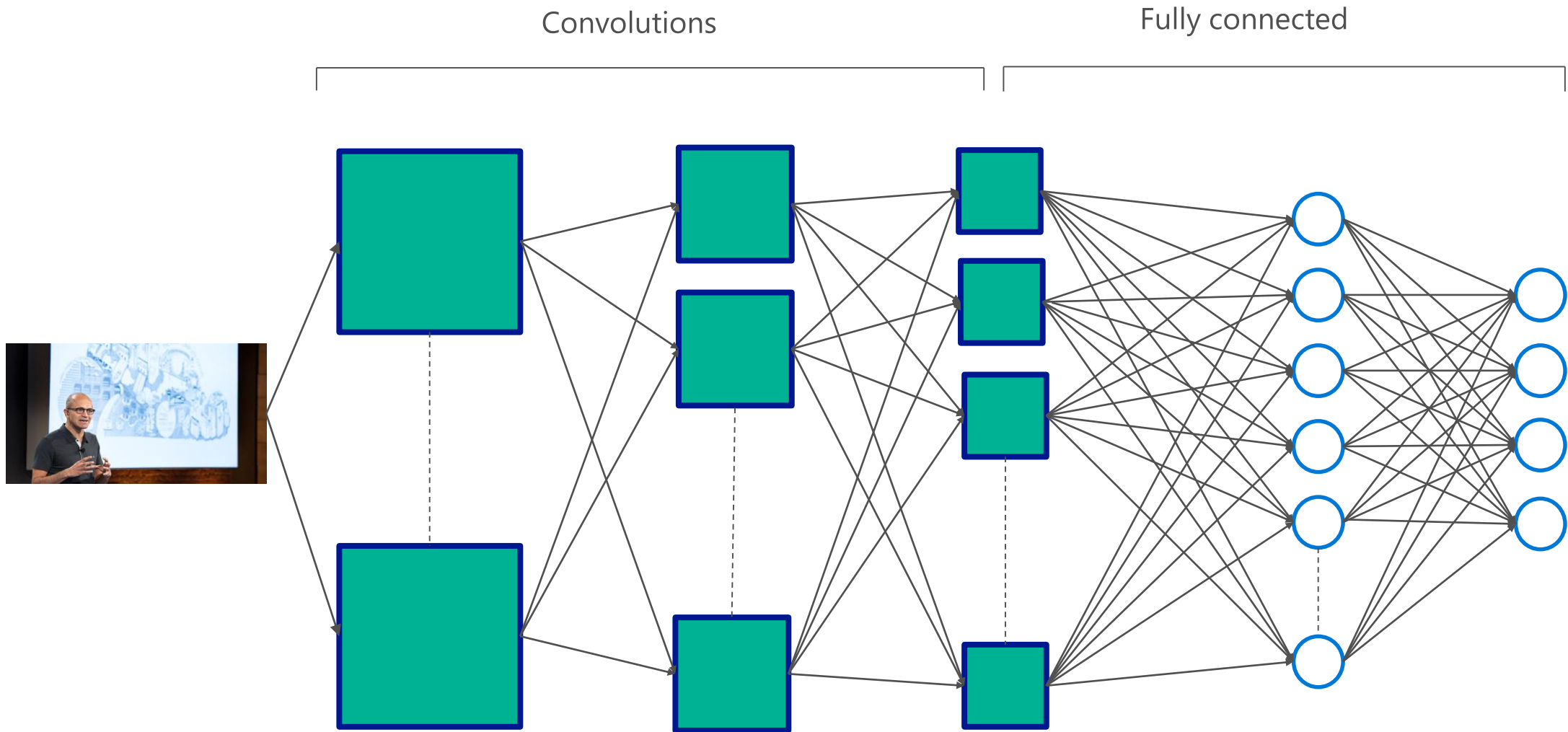
Convolutional **N**eural **N**etworks

"Feedback loops"



Recurrent **N**eural **N**etworks

... and can be arranged in layers

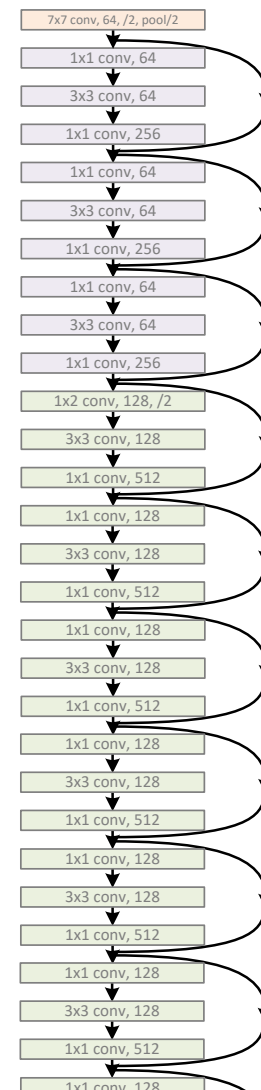


ResNet: 152 layers, and 1001 layers later on
MSRA's ResNet won the 1st places in ImageNet classification, ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in [ILSVRC](#) & [COCO](#) competitions 2015

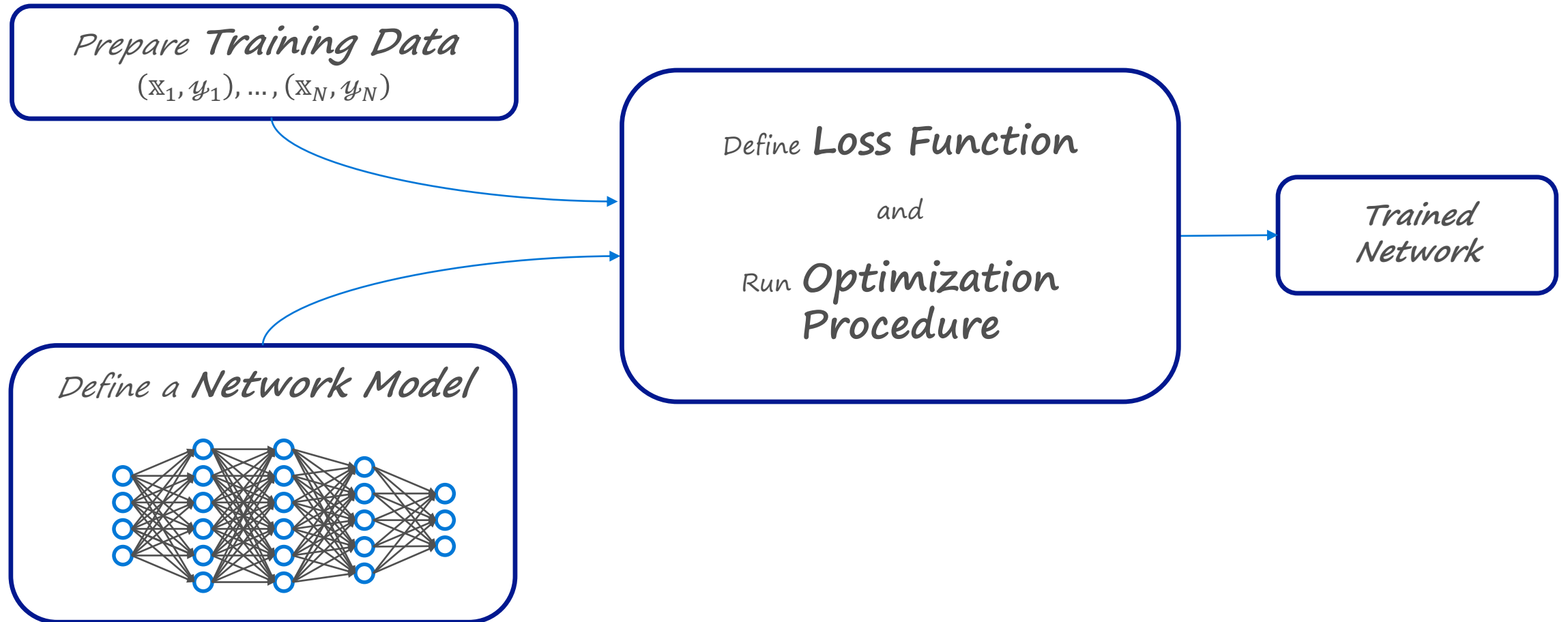
A bar chart illustrating the decline in error rates for ImageNet classification over time. The y-axis represents the error rate percentage, ranging from 0 to 30. The x-axis shows the years from 2010 to 2015. A dashed horizontal line at 5.8% indicates human performance. The error rates for various models are shown as blue bars, with the specific error rate labeled above each bar.

Year	Model	Error Rate (%)
2010	NEC	28.2
2011	Xerox	25.8
2012	AlexNet (8 layers)	16.4
2013	Clarifai (8 layers)	11.7
2014	GoogleNet (22 layers)	6.7
2015	ResNet (152 layers)	3.5

Human Performance: 5.8%



ANN Learning



Stochastic Gradient Descent and Backpropagation

Loss function

$$\min_w \sum_{i=1}^N f(x_i, y_i; w)$$

Stochastic Gradient Descent (SGD)

$$g(w_t) = \nabla f(x_i, y_i; w_t)$$

$$w_{t+1} = w_t - \eta_t g(w_t)$$

