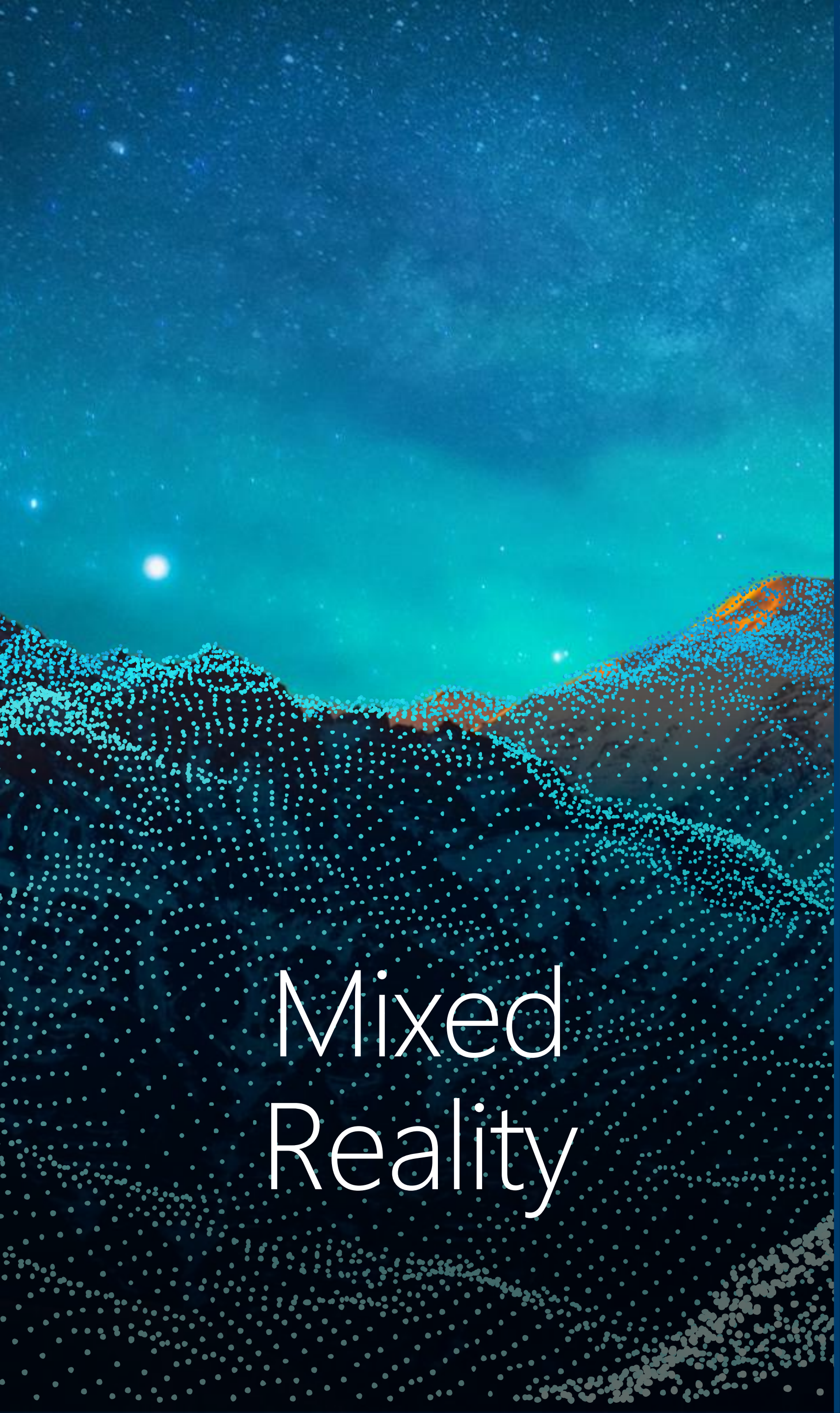
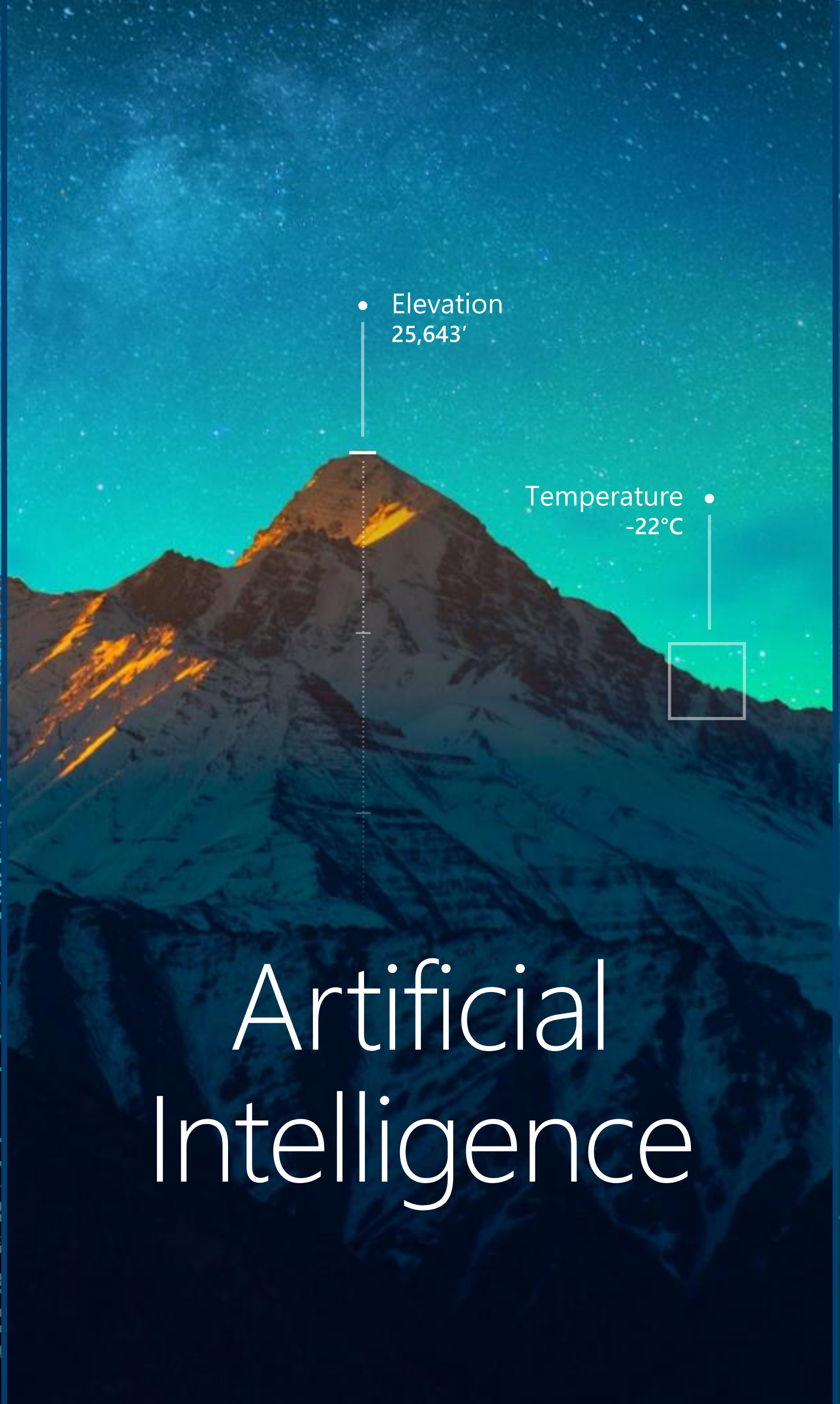
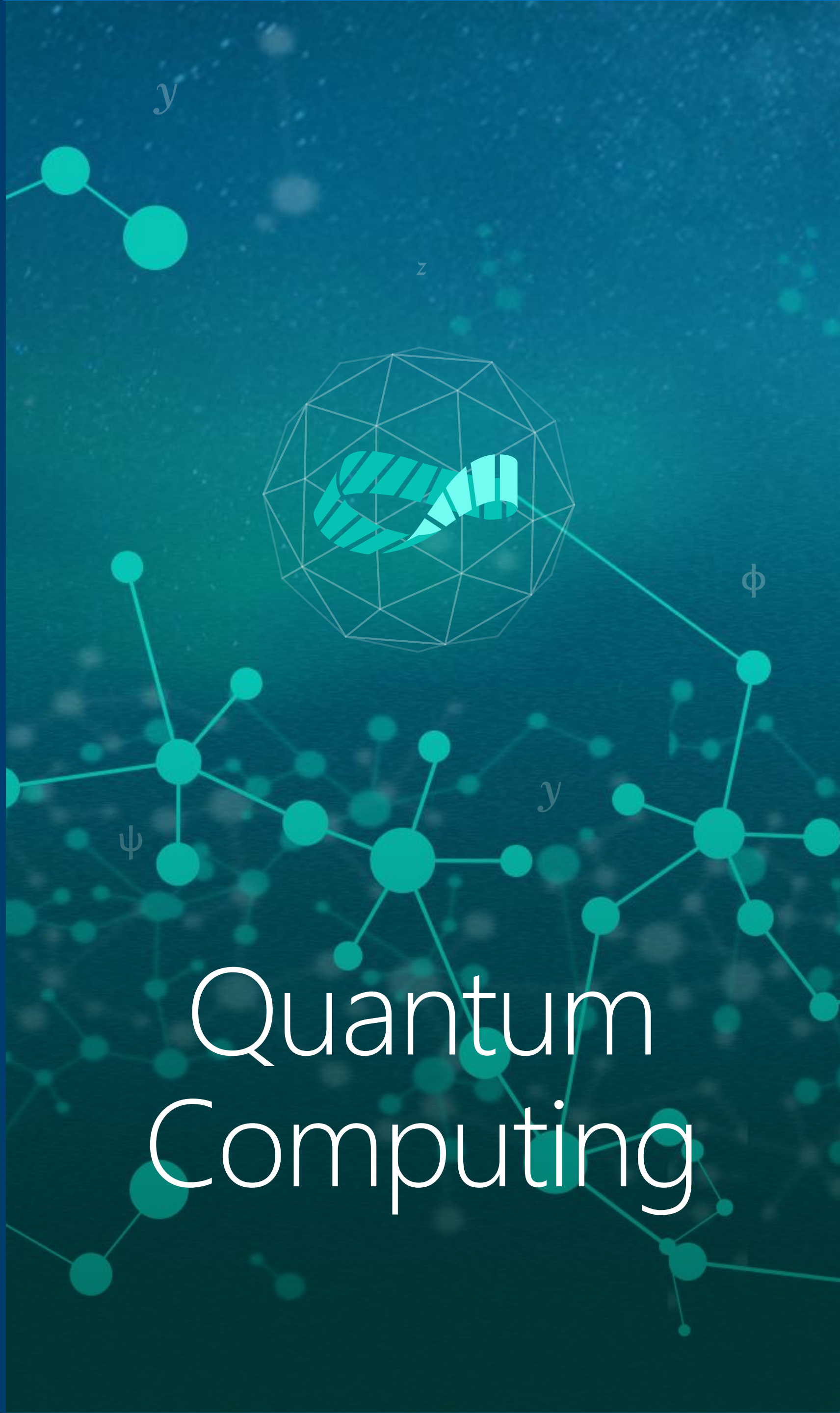
# Azure AI Infrastructure

## Running AI Workloads at Scale

Jarek Kazmierczak
MTC Silicon Valley


Mixed
Reality


Elevation
25,643′
Temperature
-22°C
Artificial
Intelligence


Quantum
Computing

# Amplifying human ingenuity with intelligent technology

## Reasoning

Learn and form conclusions with imperfect data

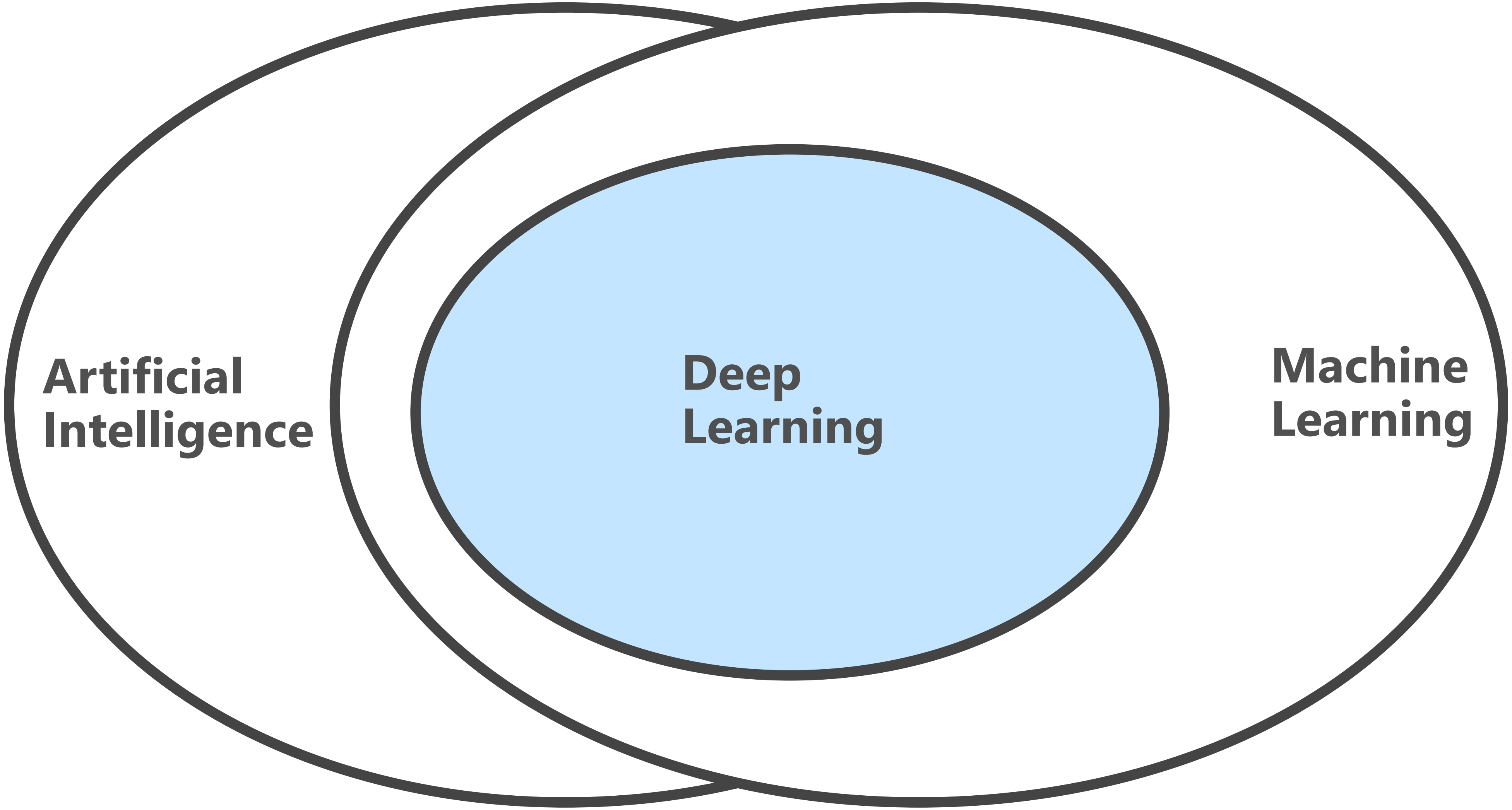
## Understanding

Interpret meaning of data including text, voice, images

## Interacting

Interact with people in natural ways

# Artificial Intelligence, Machine Learning and Deep Learning

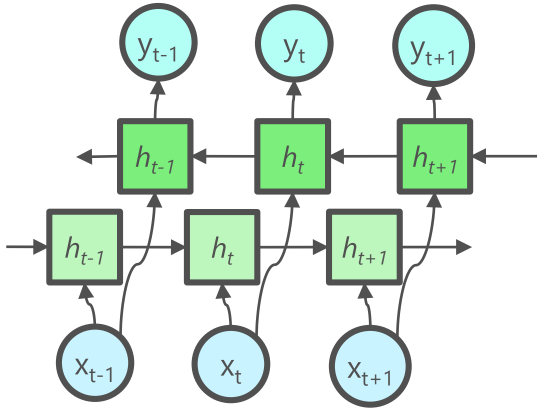


# The Rise of Deep Learning in ML

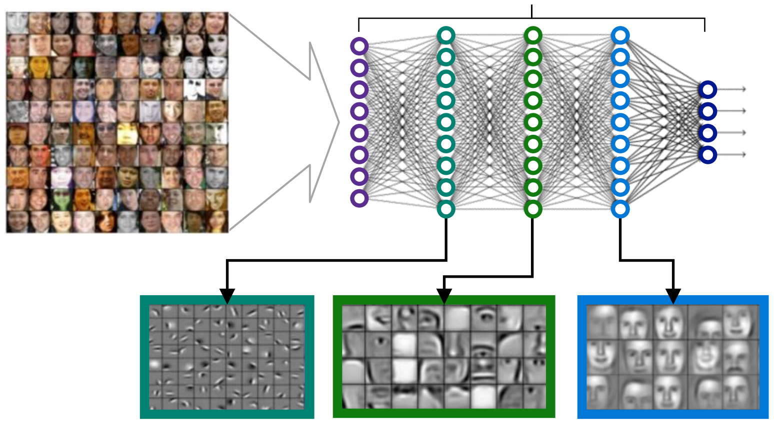
**Deep neural networks have enabled major advances in machine learning and AI**

- Computer vision
- Language translation
- Speech recognition
- Question answering
- And more…

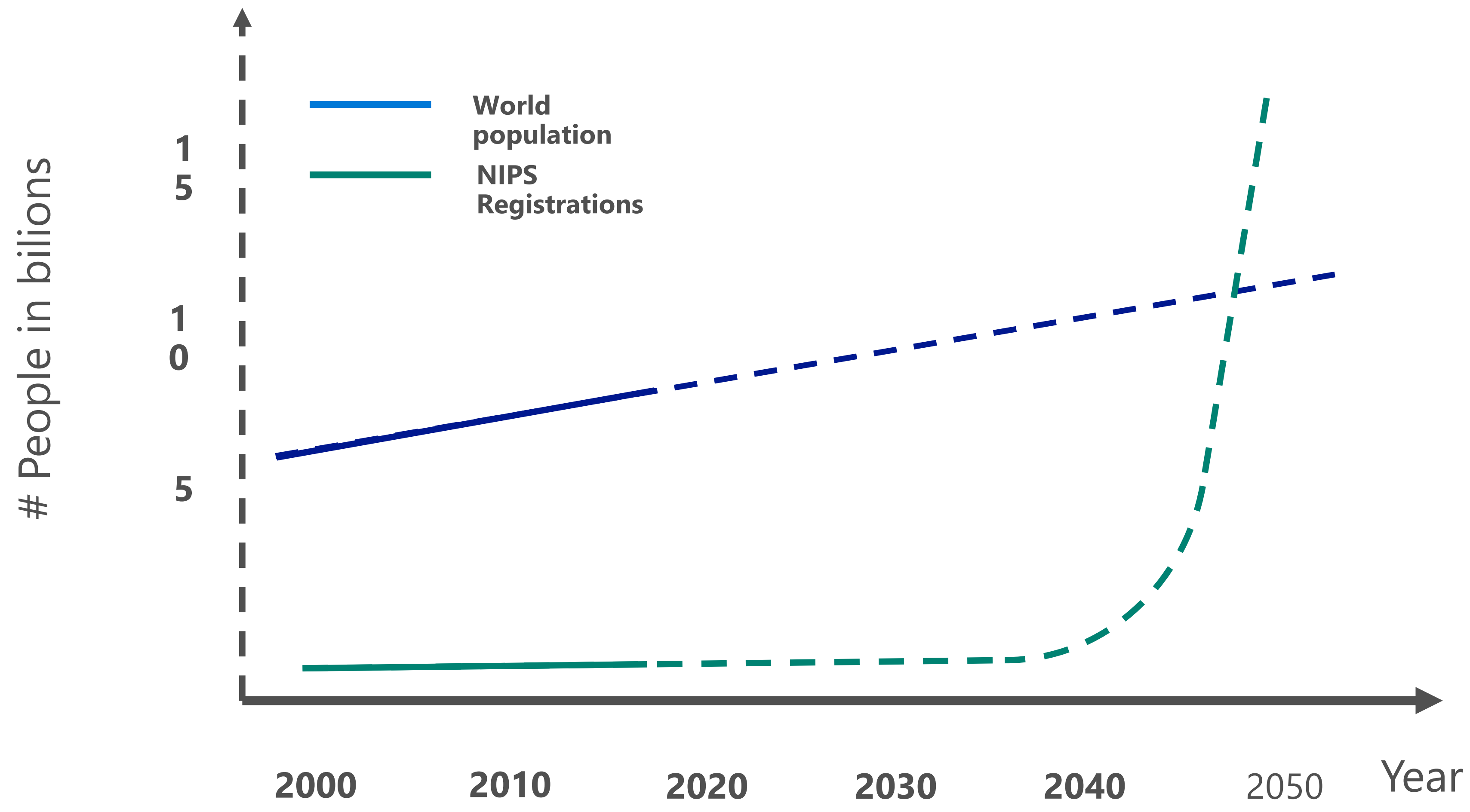
## Recurrent Neural Networks



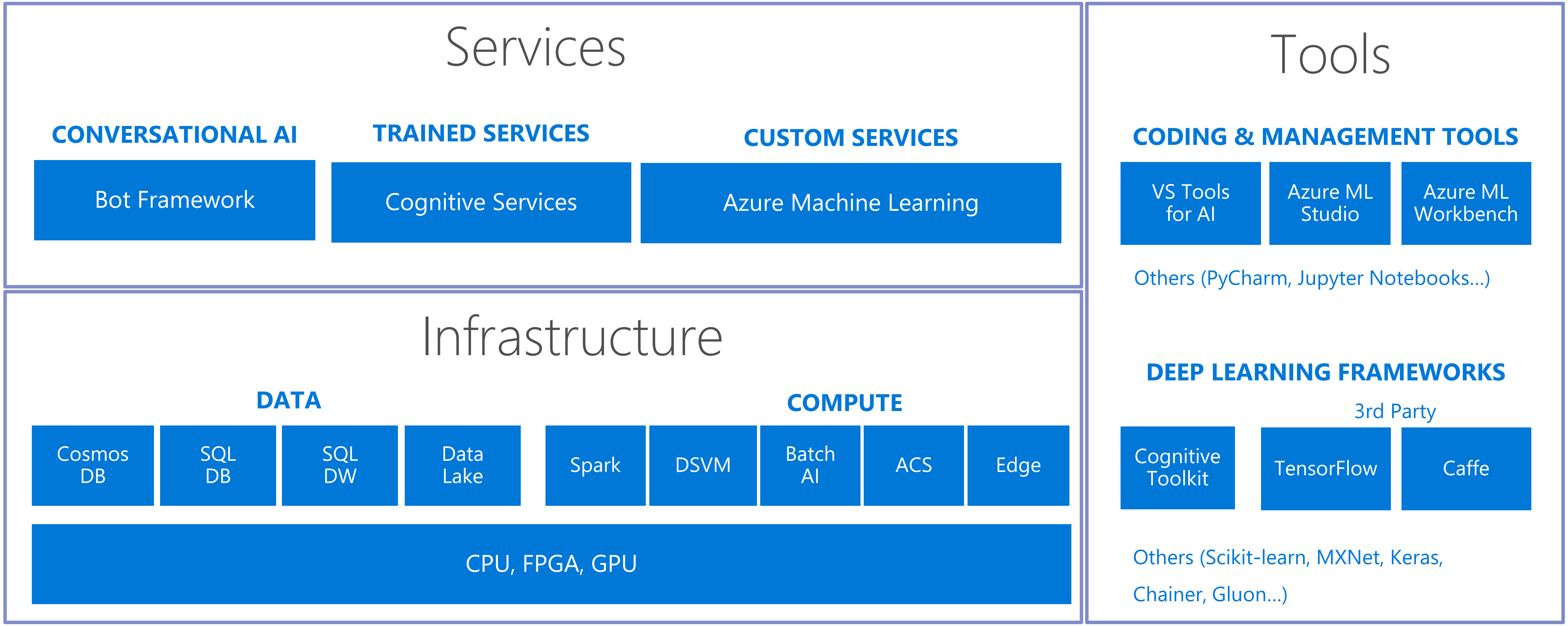

## Convolutional Neural Networks



# Deep Learning Hype



# Microsoft AI Platform



# 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

### COMPUTE
Spark
DSVM
Batch AI
ACS
Edge

CPU, FPGA, GPU

## Tools

### CODING & MANAGEMENT TOOLS
VS Tools for AI
Azure ML Studio
Azure ML Workbench

Others (PyCharm, Jupyter Notebooks…)

### DEEP LEARNING FRAMEWORKS
Cognitive Toolkit

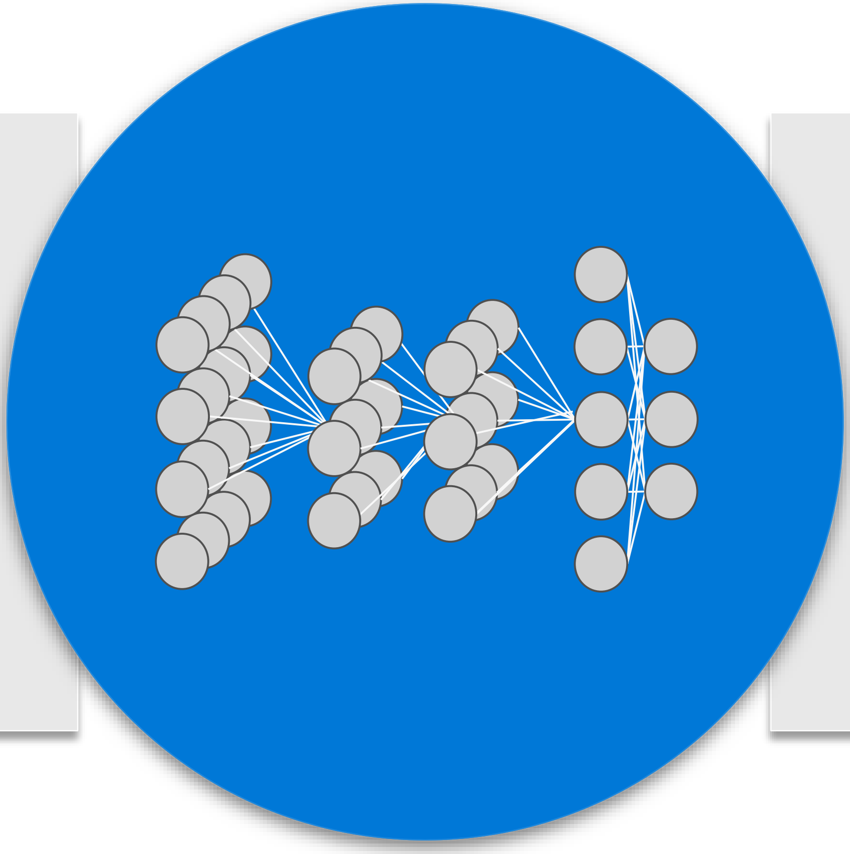
3rd Party
TensorFlow
Caffe

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

# Deep Learning Demands Compute Power

## TRAINING

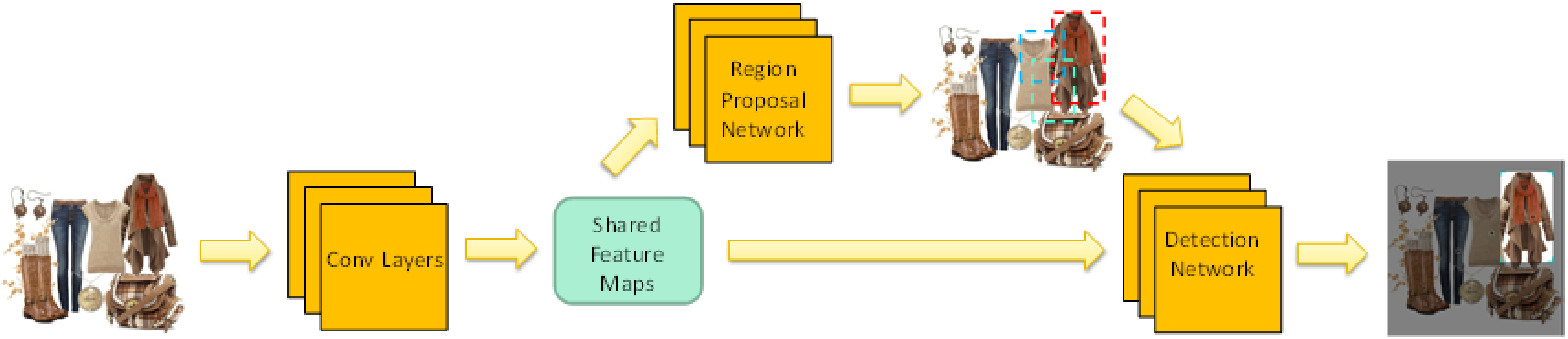
Billions of TFLOPS/TOPS per training run



## INFERENECE

Billions of FLOPS/OPS per inference

# Massive scale inference



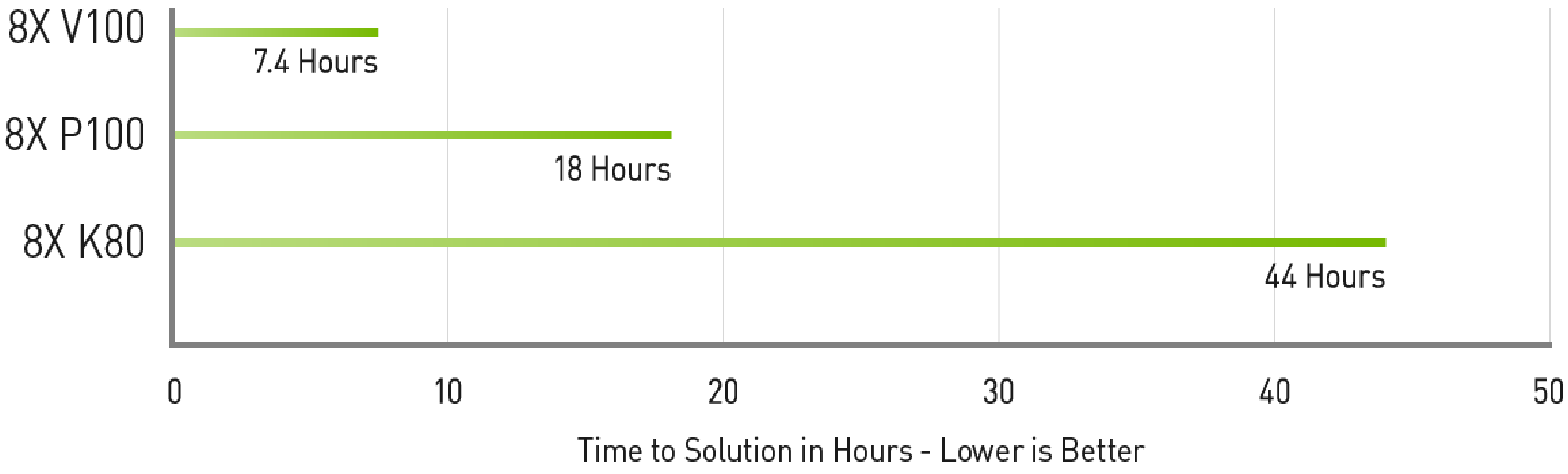
# Computational capacity – CPU, GPU, ASICs

| 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 |
| NYU facebook | Berkeley UNIVERSITY OF CALIFORNIA |
| THEANO | MATCONVNET |
| Université de Montréal | UNIVERSITY OF OXFORD |
| MOCHA.JL | PURINE |
| Massachusetts Institute of Technology | NUS National University of Singapore |
| MINERVA | MXNET* |
| NYU Microsoft | UNIVERSITY of WASHINGTON Carnegie Mellon University |
| BIG SUR | TENSORFLOW |
| facebook | Google |
| WATSON | CNTK |
| IBM | Microsoft |

# 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 |

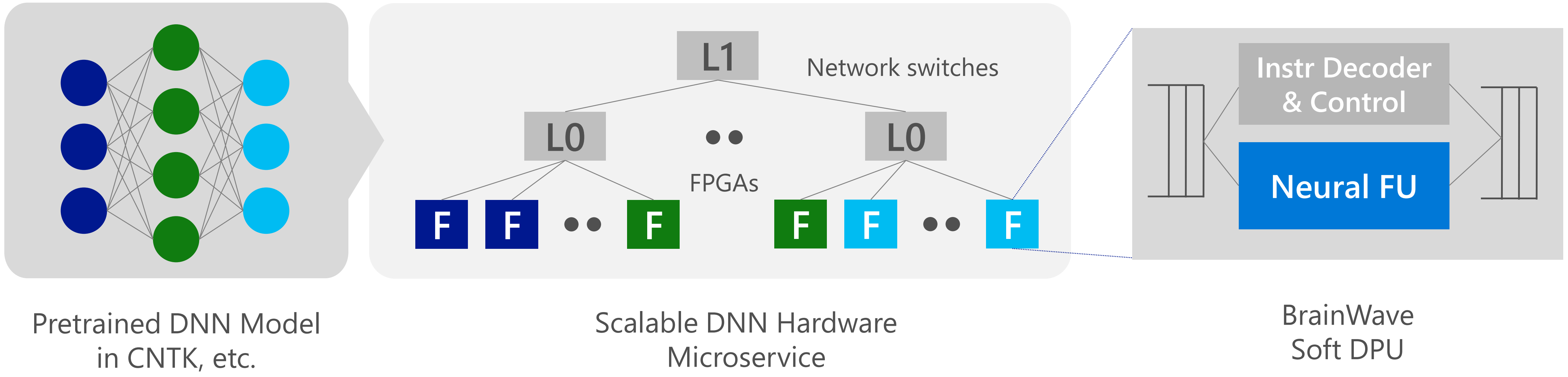
# Project BrainWave

## A Scalable FPGA-powered DNN Serving Platform

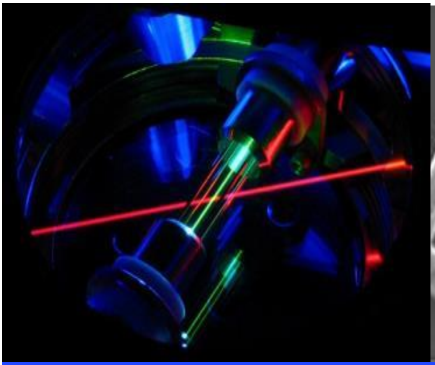
**Fast:** ultra-low latency, high-throughput serving of DNN models at low batch sizes

**Flexible:** adaptive numerical precision and custom operators

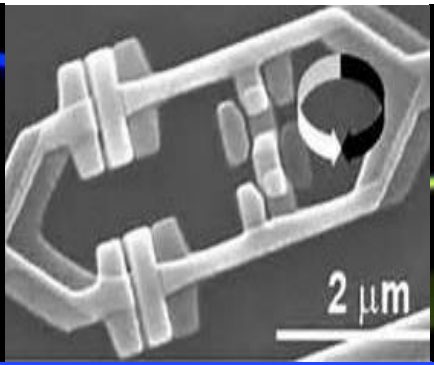
**Friendly:** turnkey deployment of CNTK/Caffe/TF/etc



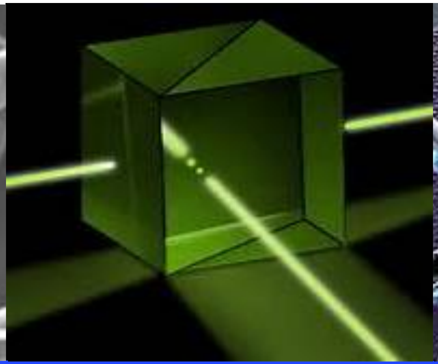
# Quantum hardware technologies



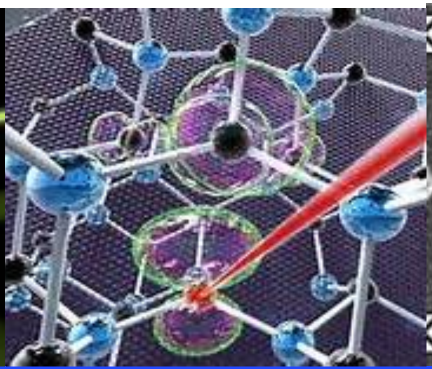
Ion traps



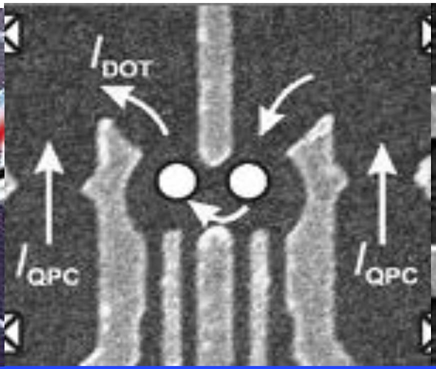

Super-conductors



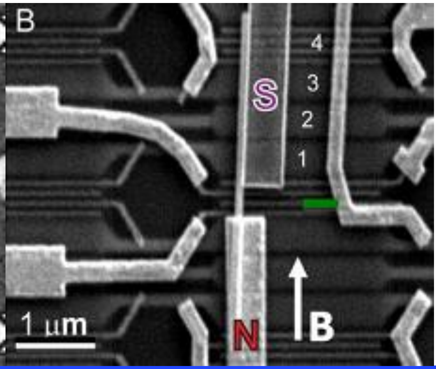
Linear optics



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

### COMPUTE
- Spark
- DSVM
- Batch AI
- ACS
- Edge

CPU, FPGA, GPU

## Tools

### CODING & MANAGEMENT TOOLS
- VS Tools for AI
- Azure ML Studio
- Azure ML Workbench

Others (PyCharm, Jupyter Notebooks...)

### DEEP LEARNING FRAMEWORKS
- Cognitive Toolkit

3rd Party
- TensorFlow
- Caffe

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

# Challenges of Training at Scale

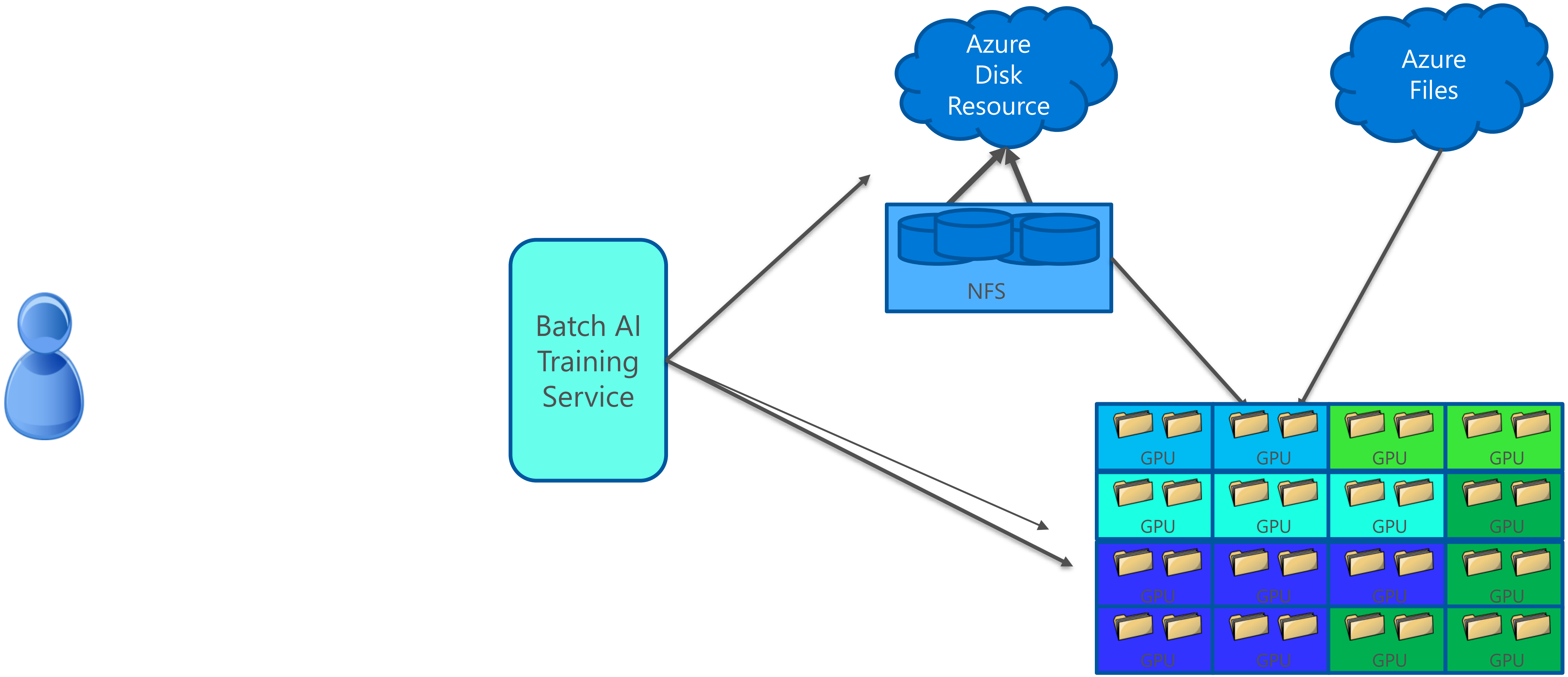
- Deploy virtual machines
- Install Drivers and dependencies
- Manage Cost
- Queue Work
- Handle monitoring and Failures
- Keep things secure
- Training on compliance restricted data
- Cleanup when done

# 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



Microsoft

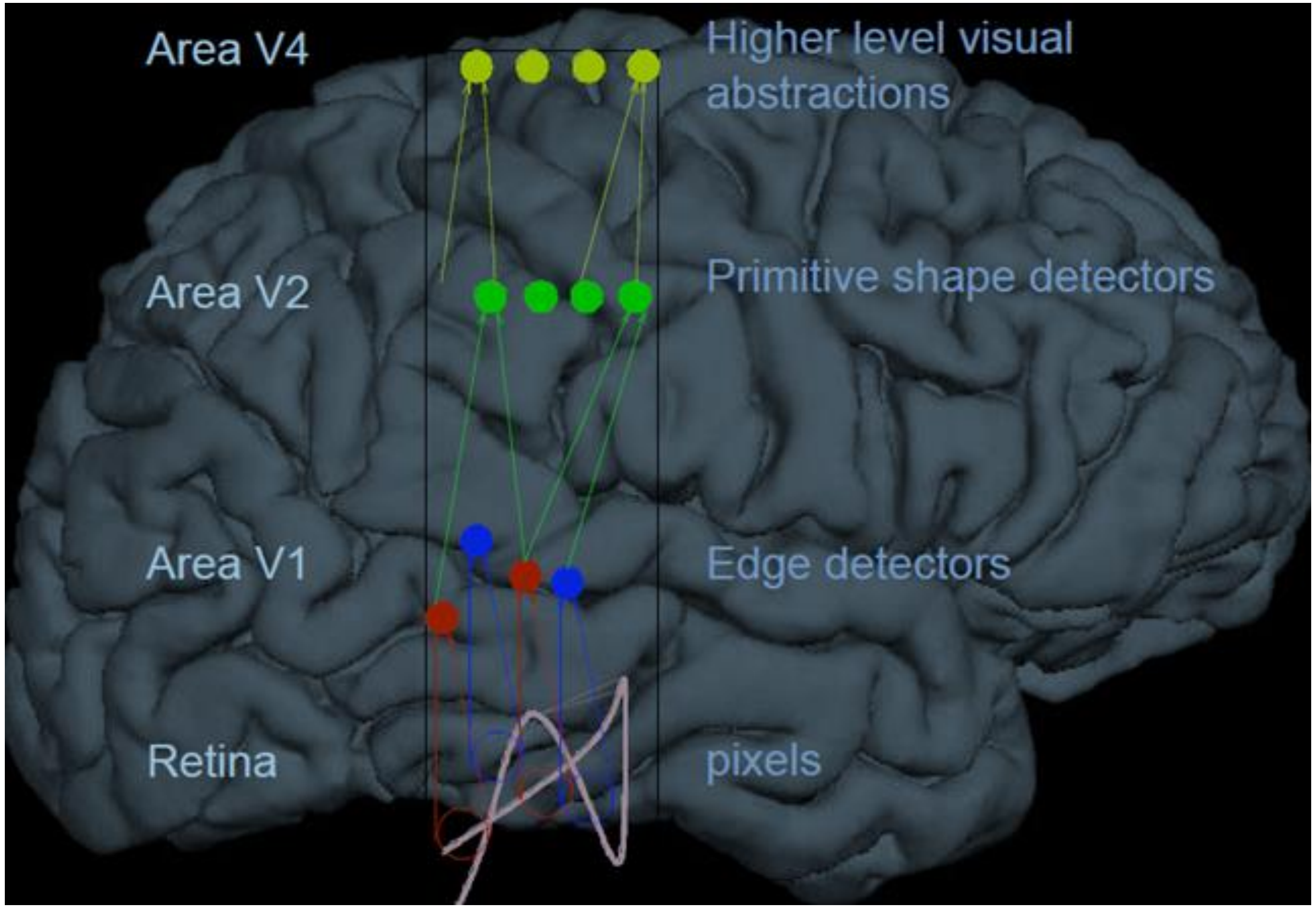
# Appendix

# Why go deep?

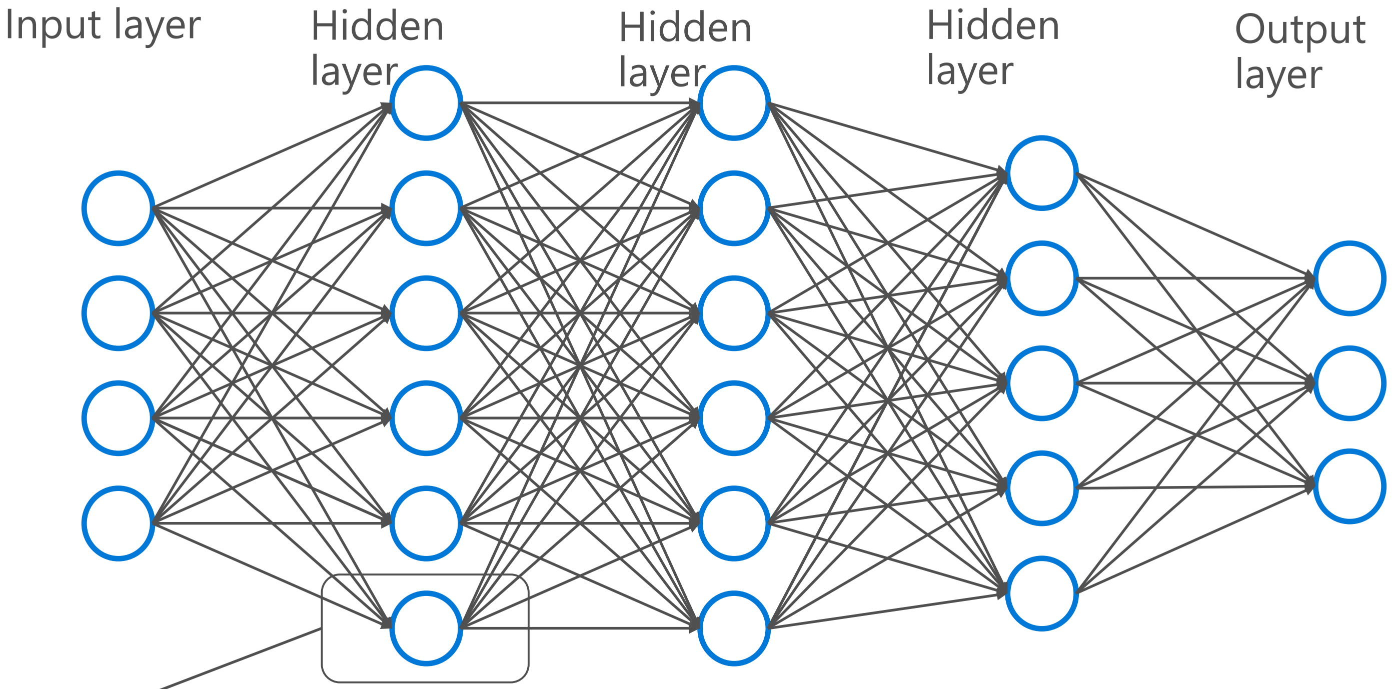
- Deep learning algorithms attempt multiple levels of representation of increasing complexity/abstraction

- Brains have a deep architecture

- Deep Learning has been successful in tasks that have been a challenge for “traditional ML”

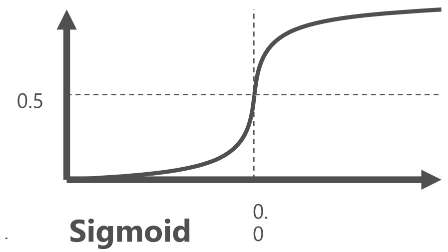


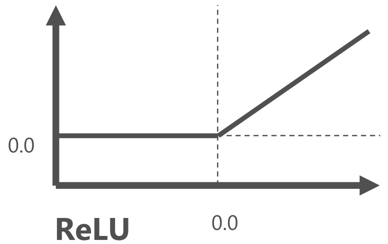
# Artificial Neural Networks - ANNs




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

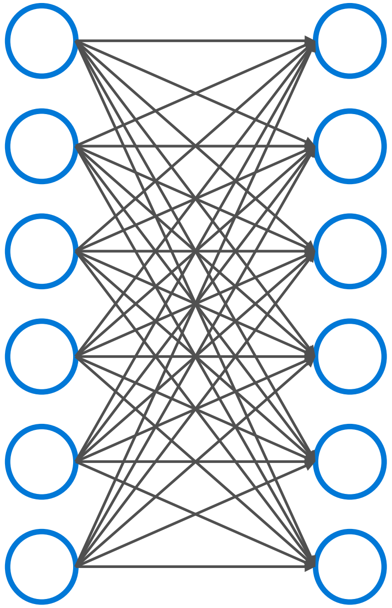
**Activation functions**



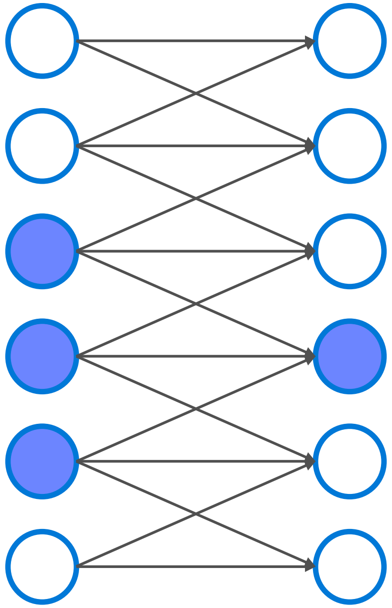
# Neurons can connect in various ways …

"Dense"



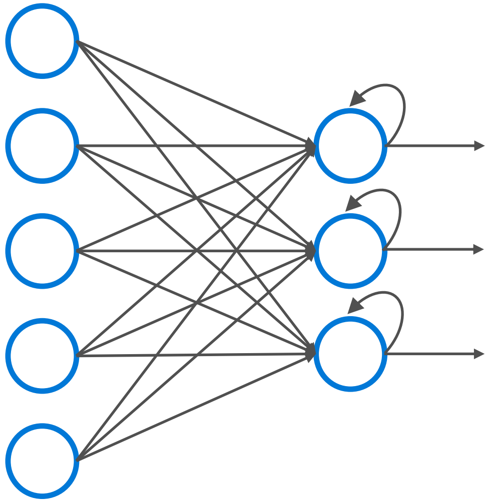
**F**ully **C**onnected **N**eural **N**etworks

"Sparse"



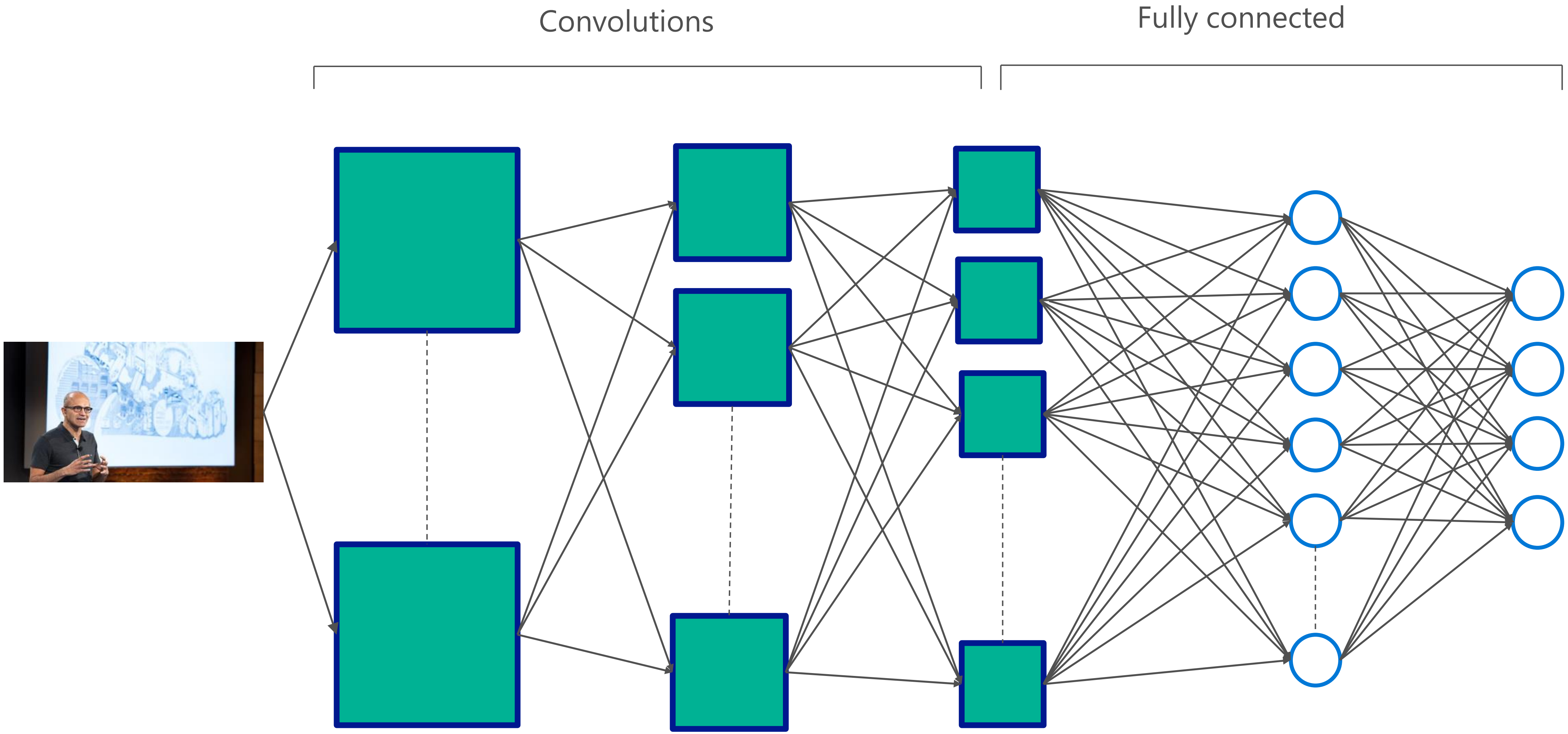
**C**onvolutional **Neural** **N**etworks

"Feedback loops"



**Recurrent** **Neural** **N**etworks

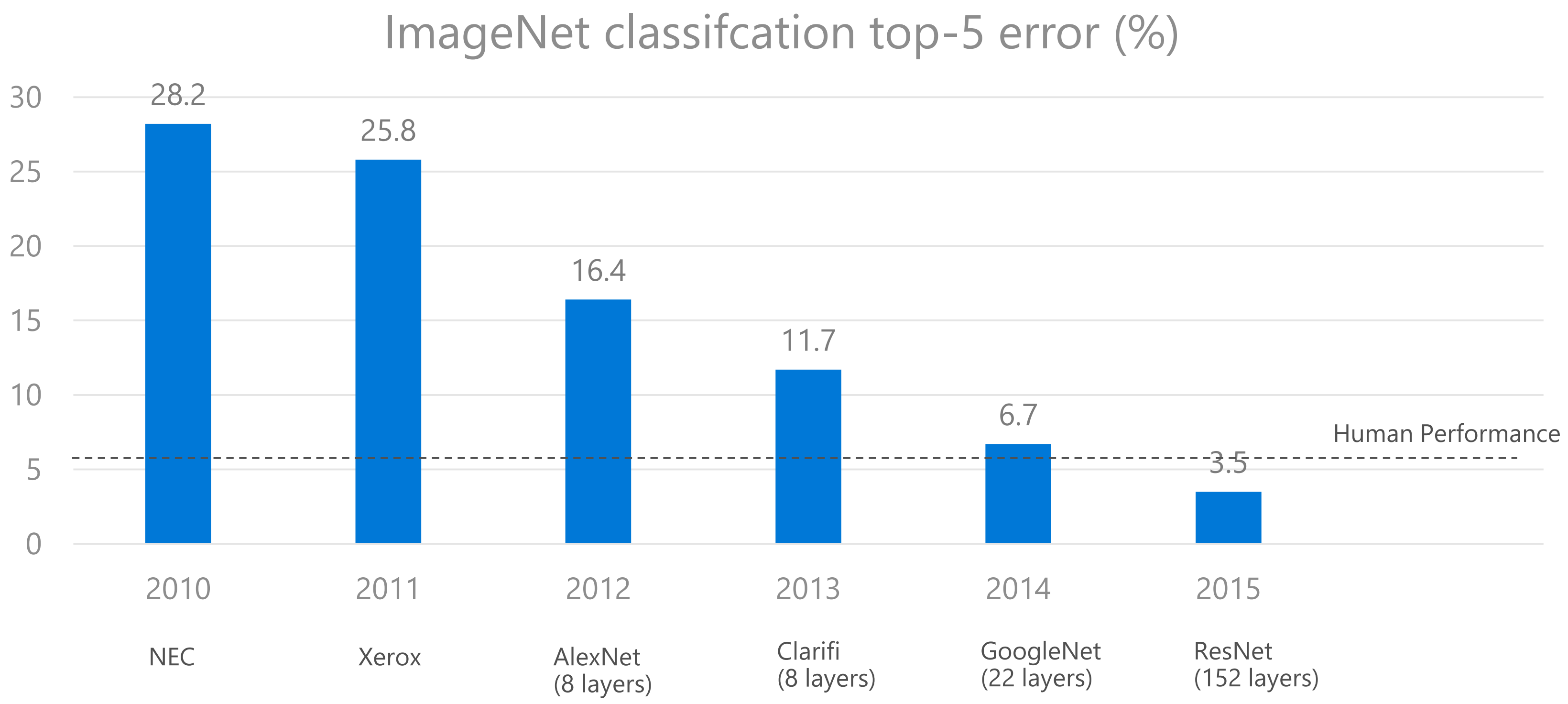
# … and can be arranged in layers

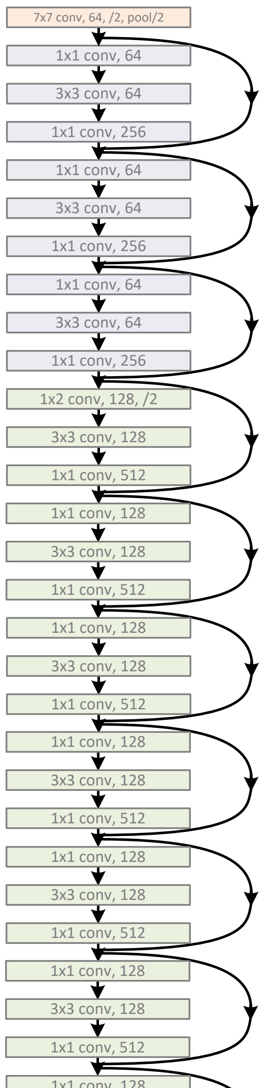


# Revolution of depth

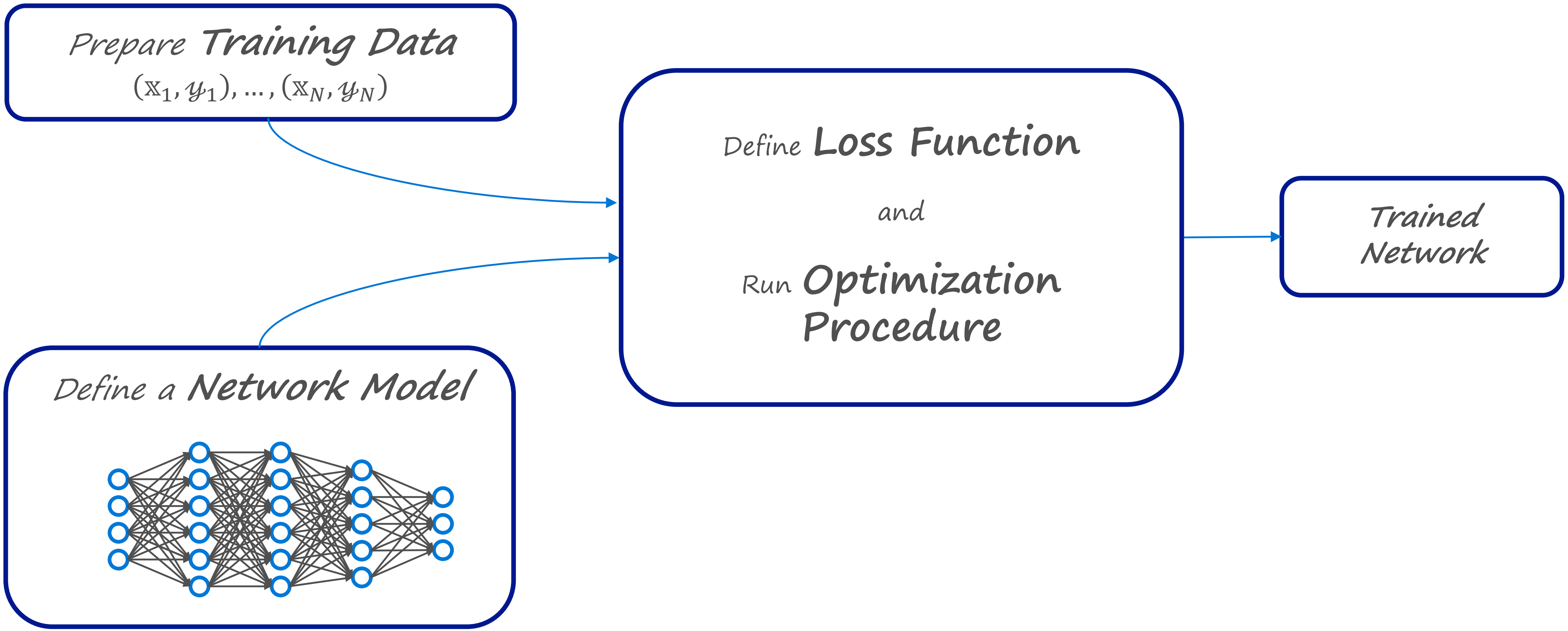
## ResNet: 152 layers, and 1001 layers later on

*MSRA's ResNet won the 1st places in ImageNet classification, ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation in [ILSVRC](ILSVRC) &[COCO](COCO) competitions 2015*



# ANN Learning



# Stochastic Gradient Descent and Backpropagation

Loss function

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

Stochastic Gradient Descent (SGD)

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

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