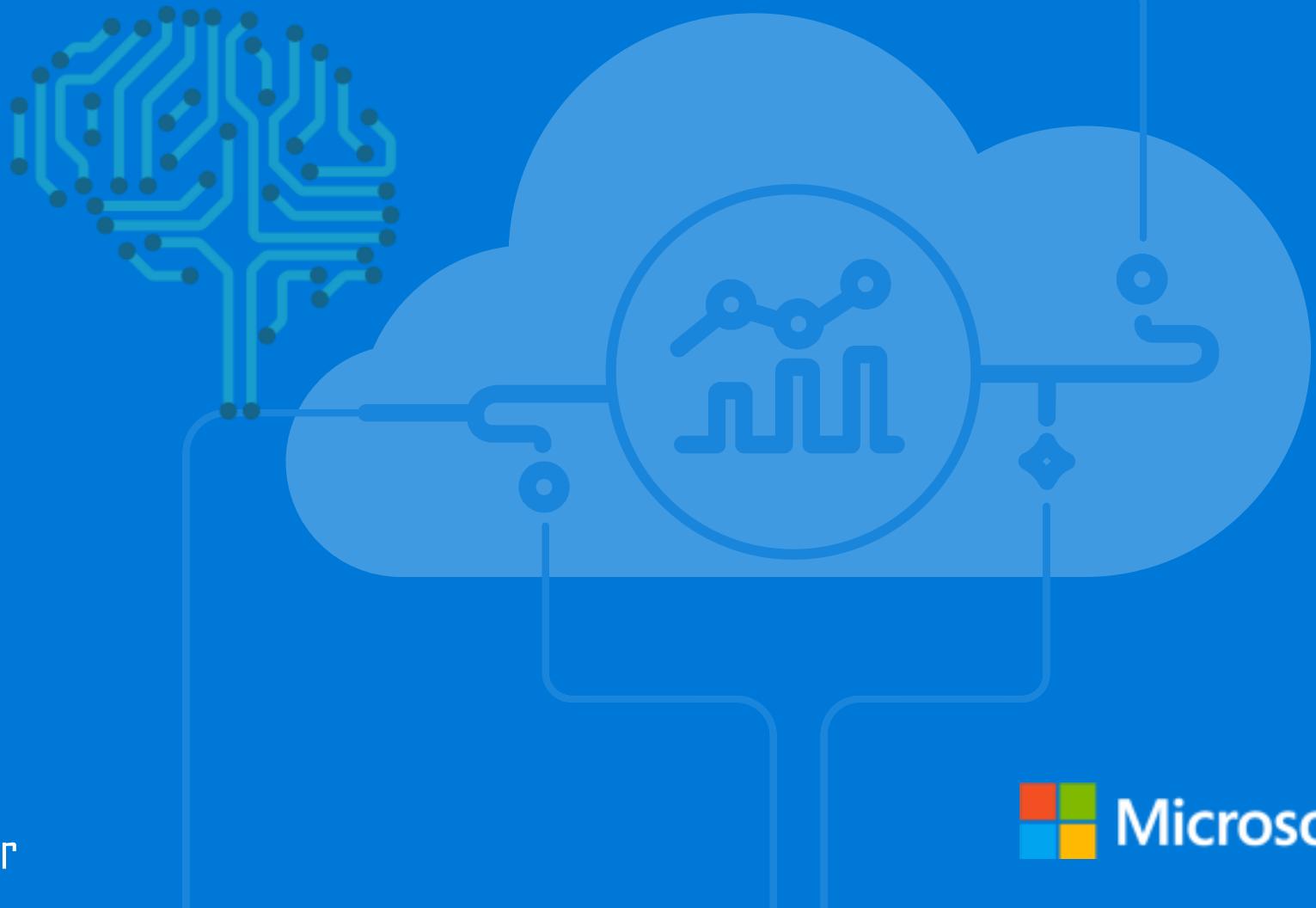
Deep Learning on Azure



Tim Scarfe

Data Solution Architect / Data Scientist

tim.scarfe@microsoft.com / @ecsquendor

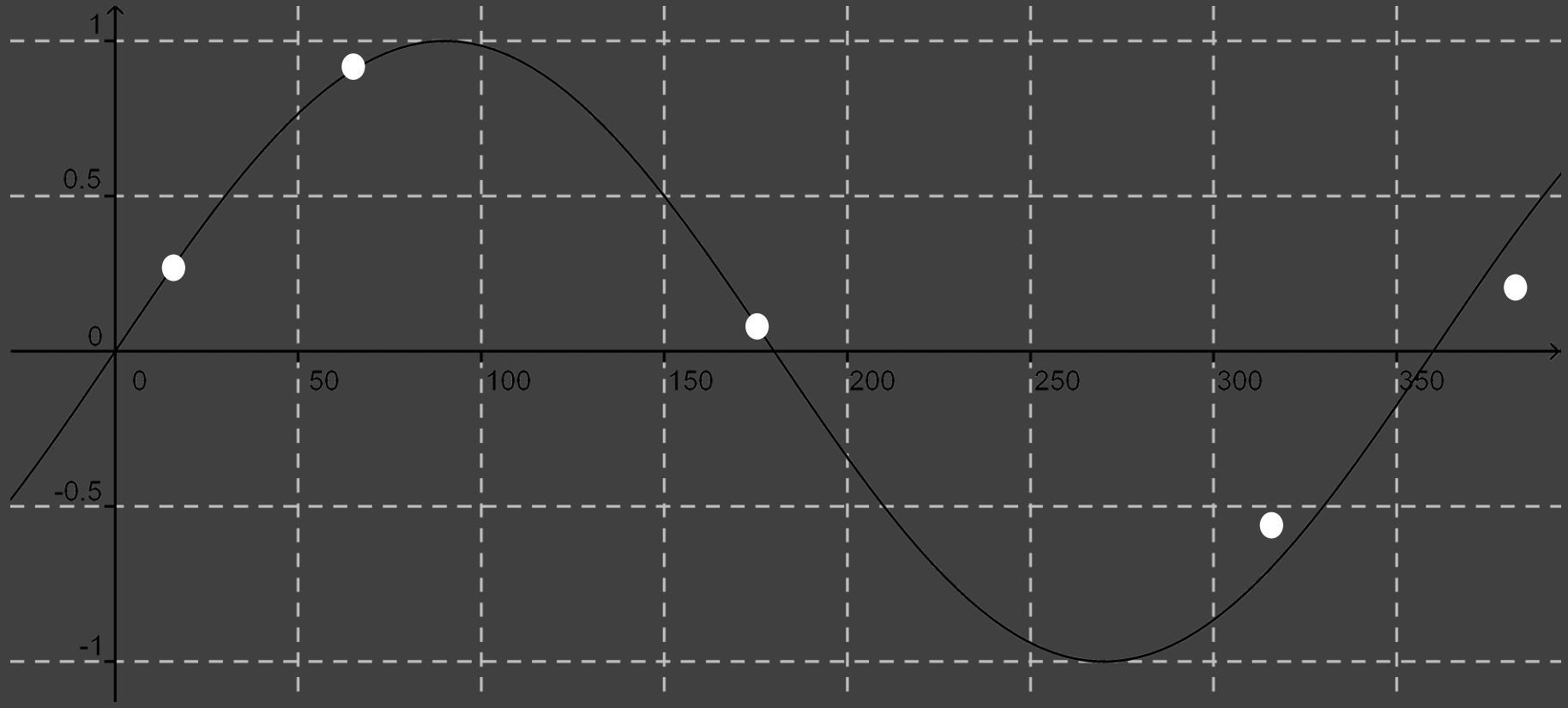
Microsoft

WHAT IS MACHINE LEARNING?

Machine learning is about a machine learning from previous experience so that it can perform some task better *based on that experience*.

Microsoft

F(X)=



Machine learning approximates a function, which we may only have a few examples of.

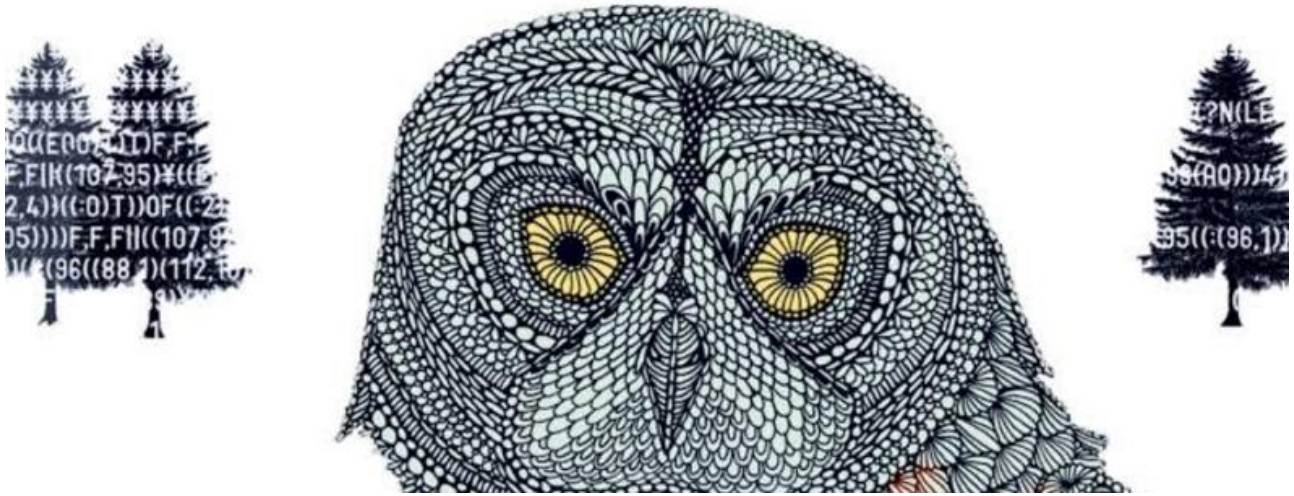
A SIMPLE SINSOIDAL FUNCTION

DATA + COMPUTE EXPLOSION = SMART AI

NICK BOSTROM

SUPERINTELLIGENCE

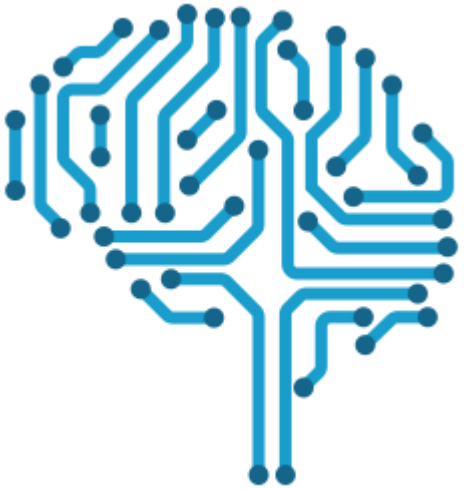
Paths, Dangers, Strategies



DON'T WORRY ABOUT GENERAL INTELLIGENCE

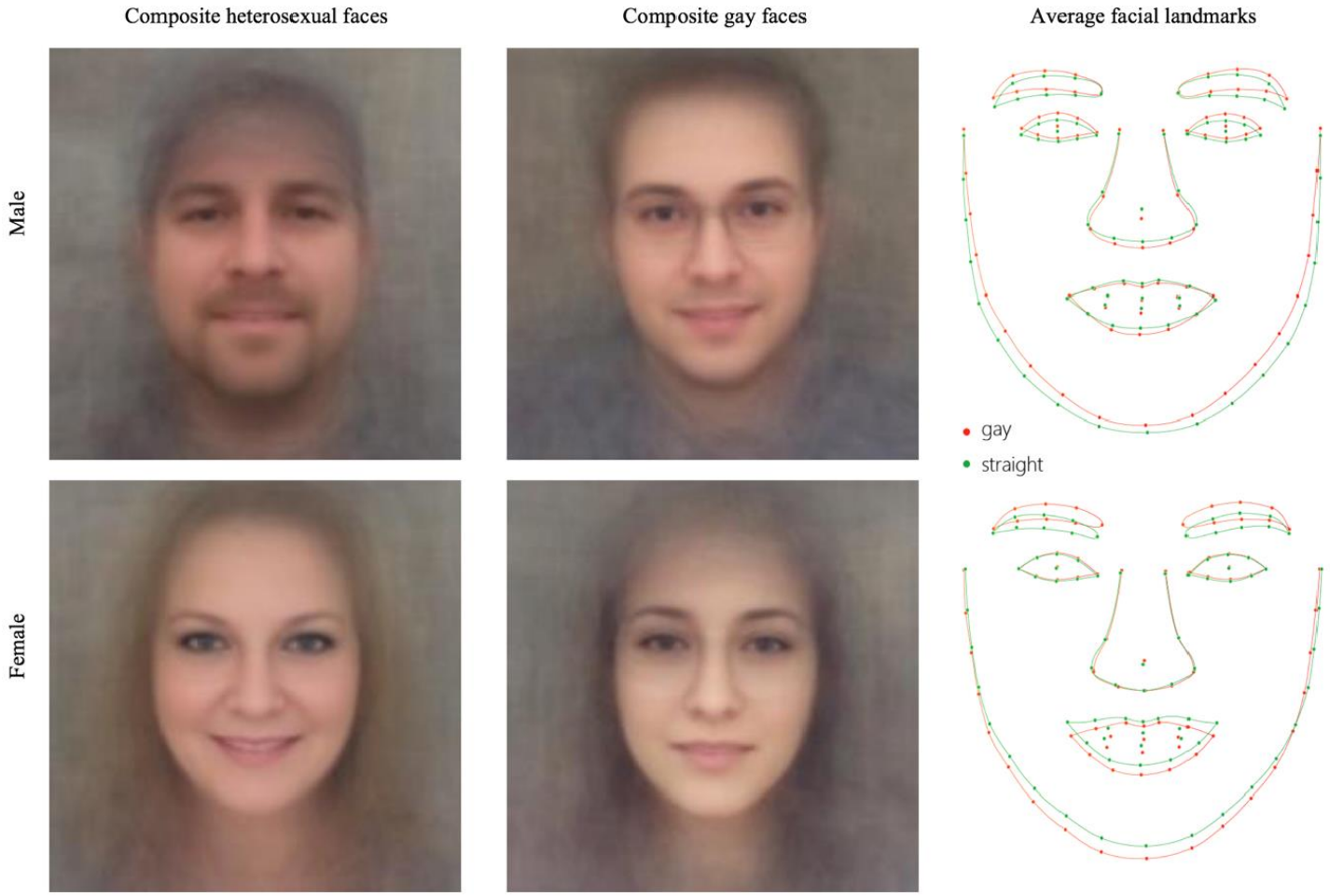
Microsoft

1. Discrimination
2. Opaque AI
3. Data is not neutral
4. Manipulating markets and consumers/voters
5. Lack of human connection
6. Engineers are not philosophers (moral reasoning)
7. Privacy
8. Innocent till proven guilty?



DO WORRY (A LOT) ABOUT ETHICS IN AI

https://osf.io/zn79k/



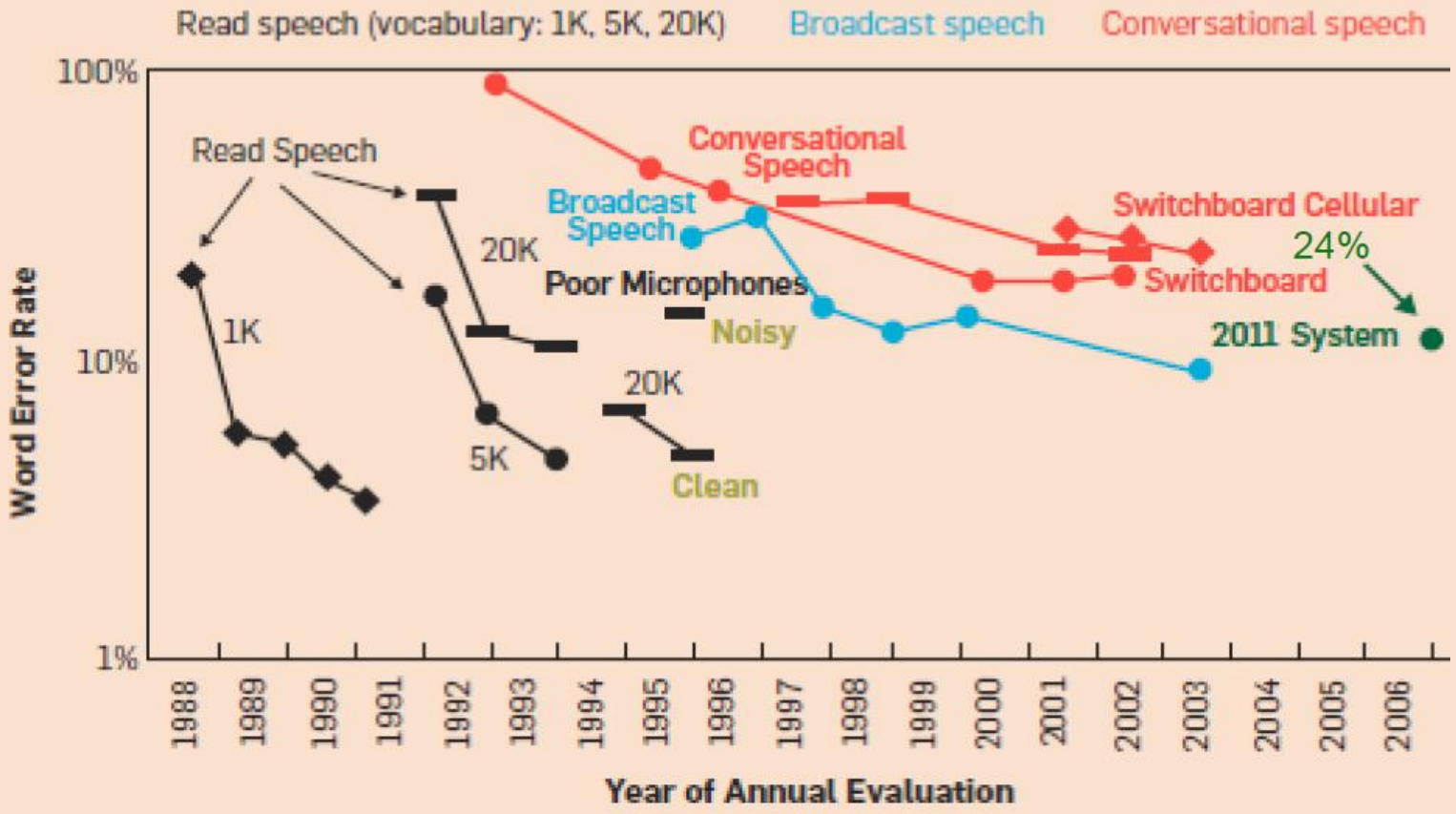
394

395 *Figure 4*. Composite faces and the average facial landmarks built by averaging faces classified as most and least likely to be gay.

2017 Wang, Y., & Kosinski, M.

PHYSIOGNOMY IS BACK?


"Our goal is to democratise AI to empower every person and every organisation to achieve more."
Satya Nadella

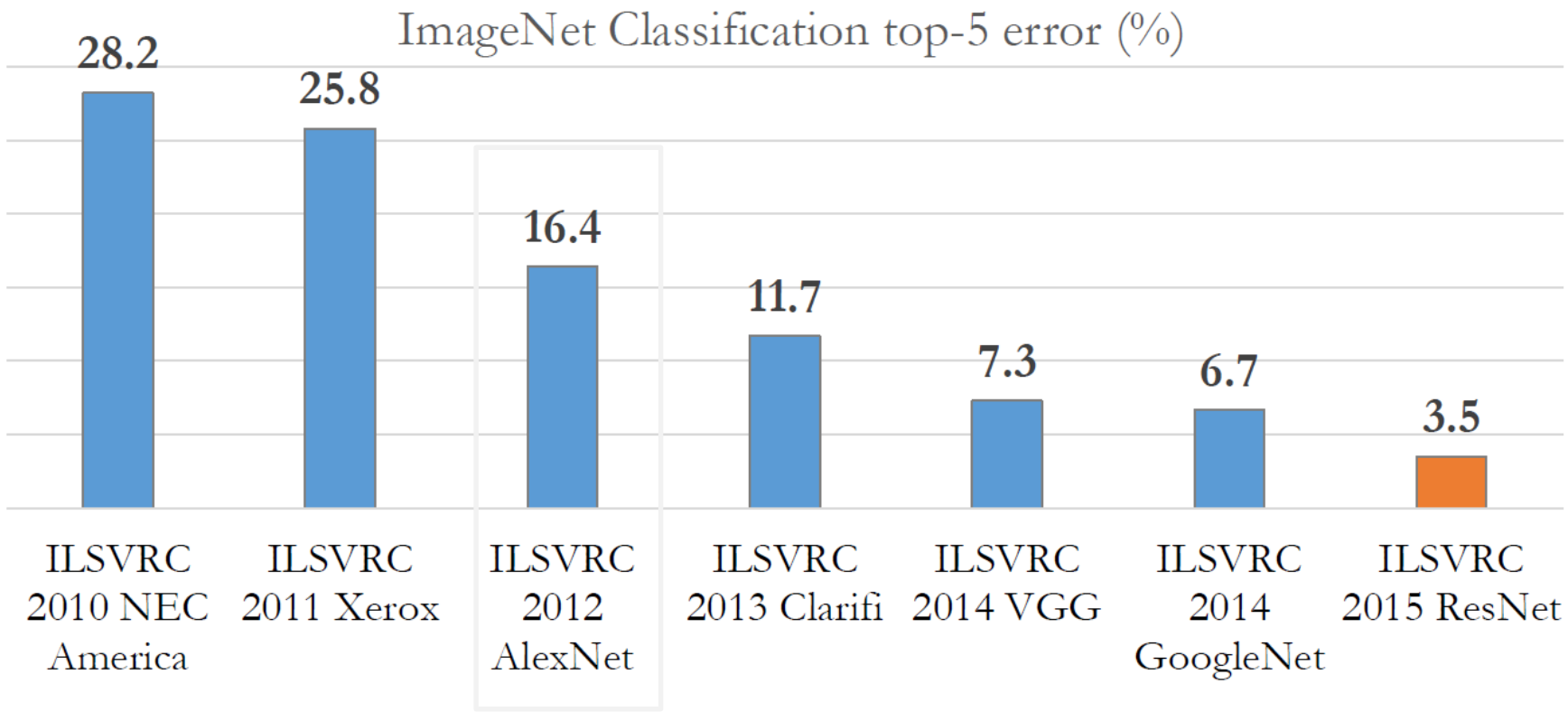


2017: ~5%!

Human error: ~5%

IMPROVEMENTS IN SPEECH RECOGNITION

IMPROVEMENTS IN COMPUTER VISION



First deep learning system

2017: ~2.2%

ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Microsoft

Machine Learning/AI Stack at Microsoft

Project "Vienna" (2018)

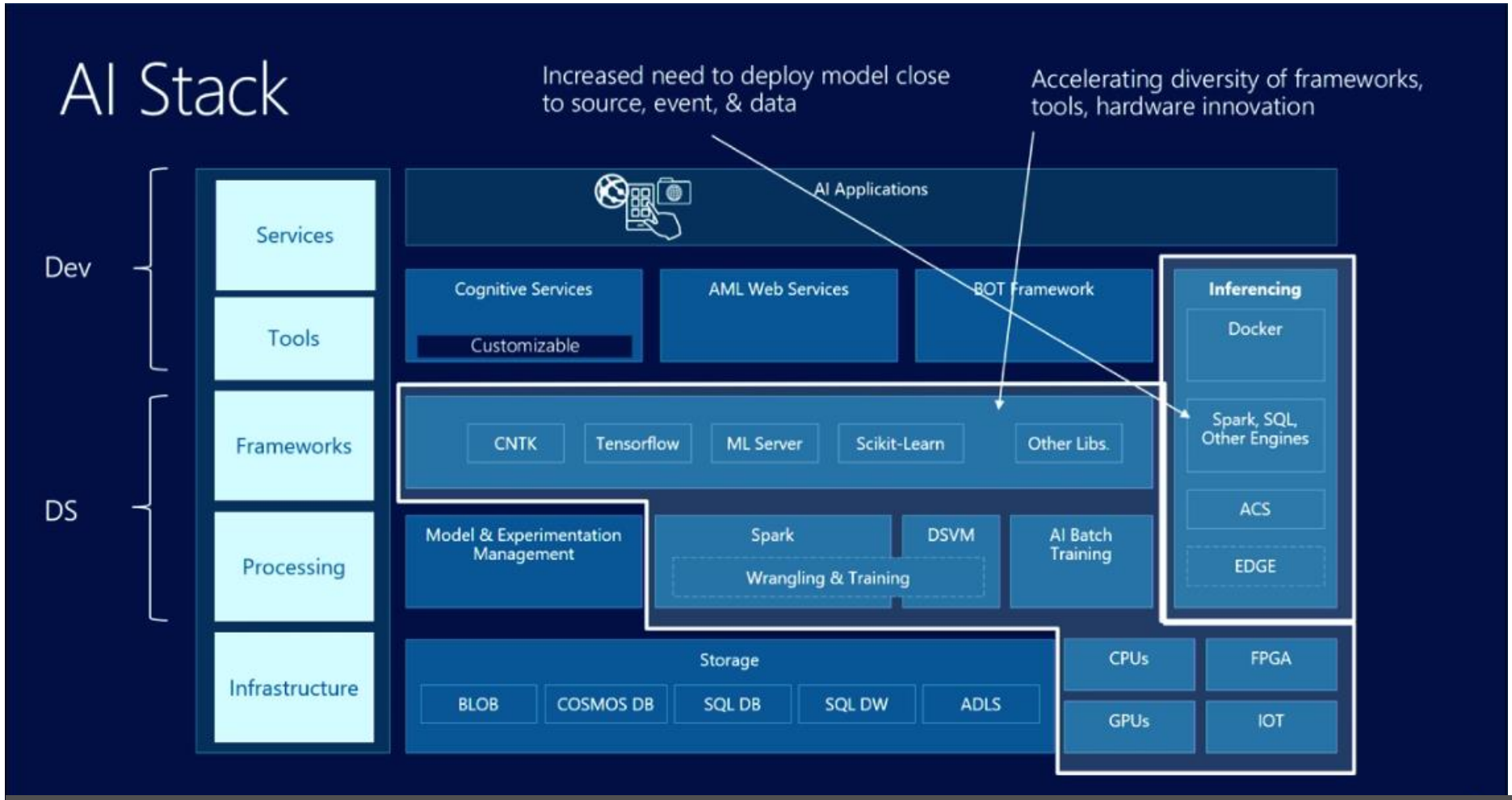
Cognitive Services | AzureML | Framework/Runtime | CNTK

Basic discriminative models → Novel prediction architectures

SaaS (REST)

PaaS (Drag/Drop)

Code-first, hybrid execution model



Microsoft Confidential - Project "Vienna" will support all frameworks and execution environments on-prem and cloud (cloud/containers/Spark/inproc)

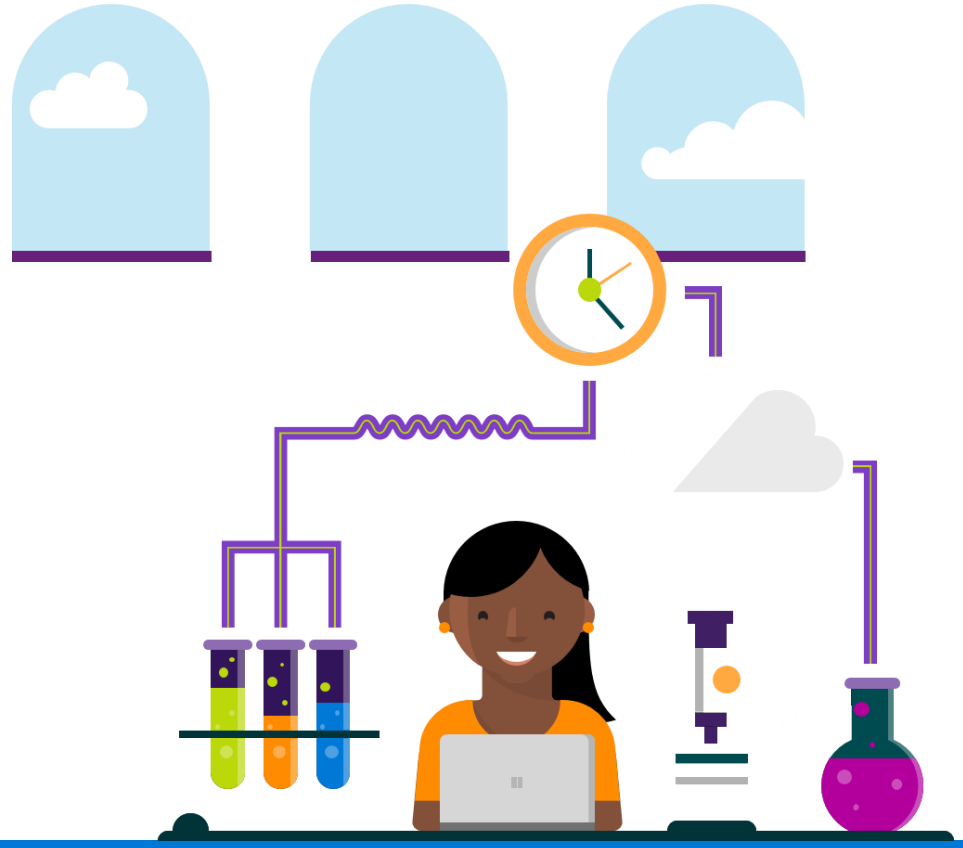
Microsoft

https://www.microsoft.com/en-us/research/blog/microsoft-unveils-project-brainwave/

Microsoft

PROJECT BRAINWAVE

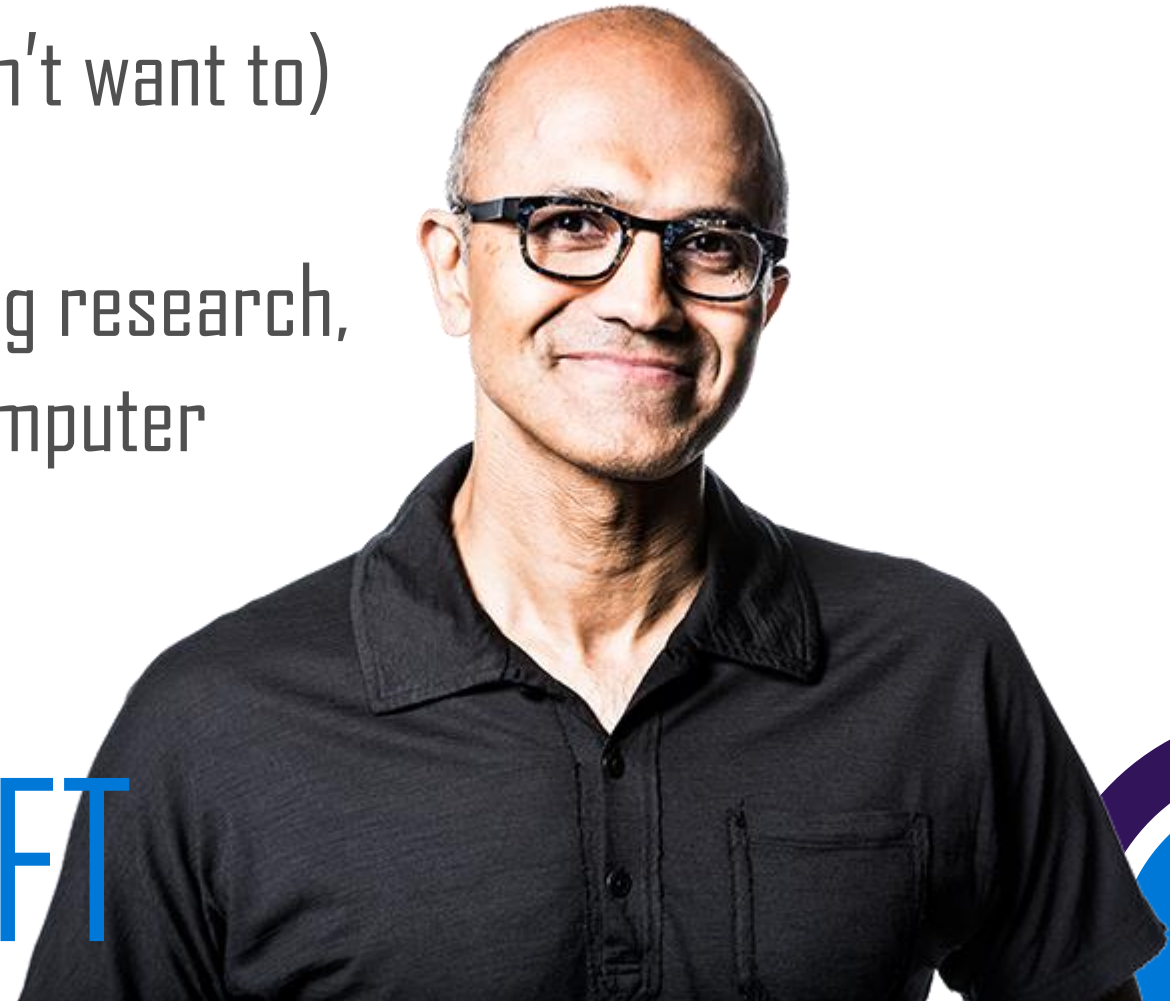
https://batchaitraining.azure.com



AZURE BATCH AI TRAINING SERVICE

Microsoft

- We are #1 contributors to open source
- Platinum member of the Linux foundation
- We support all main deep learning frameworks (you don't have to use CNTK if you don't want to)
- CNTK is 100% open source
- We are world-leading on deep learning research, especially speech recognition and computer vision.



THE NEW MICROSOFT

Images/Sequences

Regression

Supervised

Tell the machine the correct answer

Classification

Anomaly detection

Embeddings

Unsupervised

Tell the machine nothing, let it observe the world

Clustering

Dimensionality Reduction

Agent-based learning

Reinforcement Learning

Only tell machine if it was right or wrong

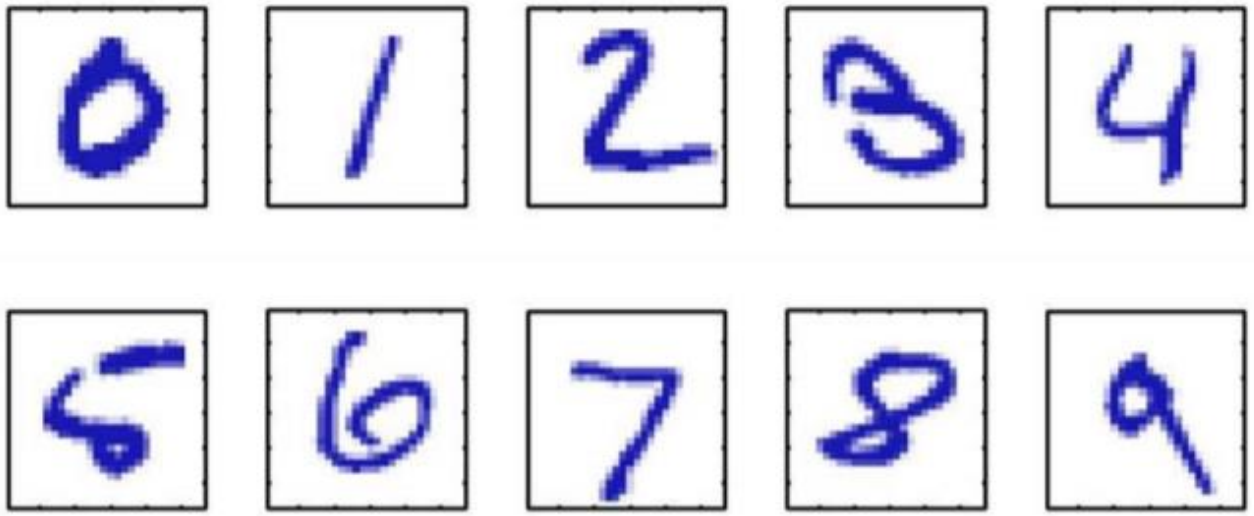
Autonomous cars

TYPES OF MACHINE LEARNING

- Approximate a function which maps from signals (image) to labels (has-cat)
- This “decision function” can predict missing labels (has-cat) on new, previously unseen signals.

MANDATORY CAT EXAMPLE

MNIST Digit Classification



Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Signals

Labels

DISTILLED CONCEPTS OF DEEP LEARNING

Microsoft

- The networks have **many levels** of depth
- Machine learns a hierarchy of representations
- **No feature extraction** required

Traditional ML

Hand crafted features

Feature Extractor → Trainable Classifier

Deep Learning

Representations are hierarchical and trained automatically

Low Level Features → Mid Level Features → High Level Features → Trainable Classifier

"DEEP" = THE ENTIRE MACHINE IS TRAINABLE

- Unlike other (frequentist) ML algorithms; you can map from *anything to anything*

SEQUENCES
SPEECH
LANGUAGE
IMAGES
[...] YOU NAME IT

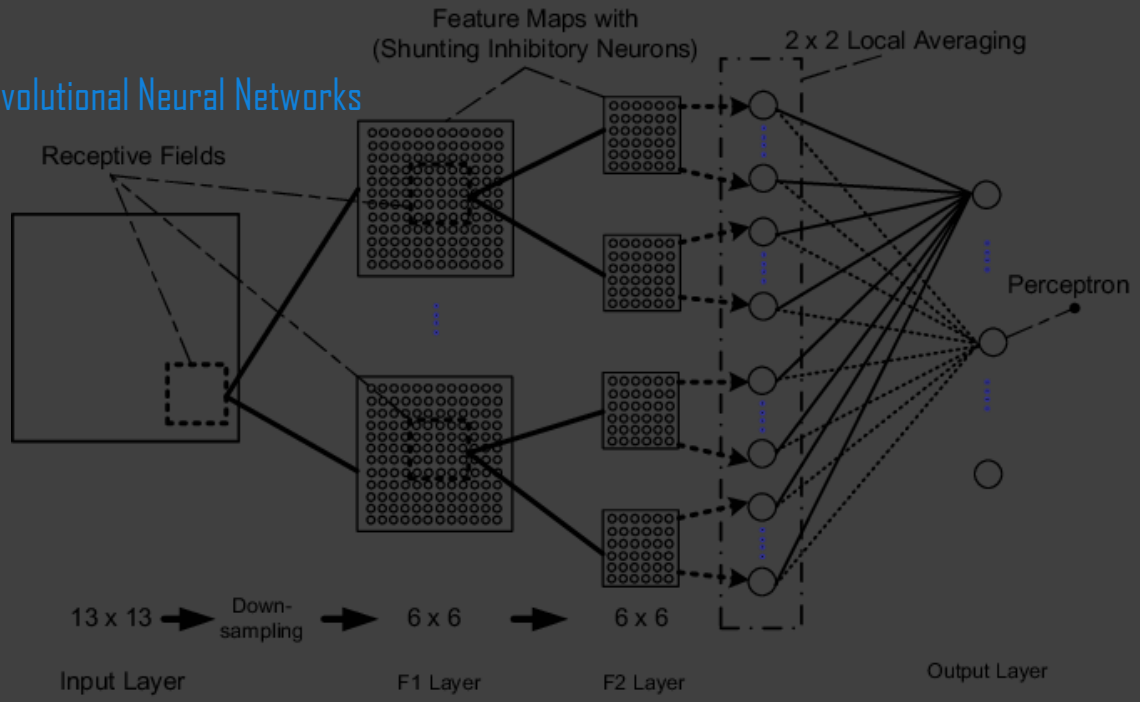
We learn this spacial manifold or "mapping"

SEQUENCES
SPEECH
LANGUAGE
IMAGES
[...] YOU NAME IT

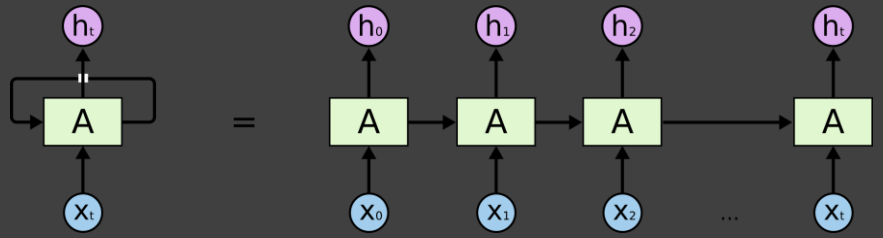
UNIVERSAL FUNCTION APPROXIMATORS

- Unlike other algorithms, NNs can encode useful and obvious relationships in the data domain
  - Local spatial dependencies (vision)
  - Time dependencies (language, speech)

Convolutional Neural Networks

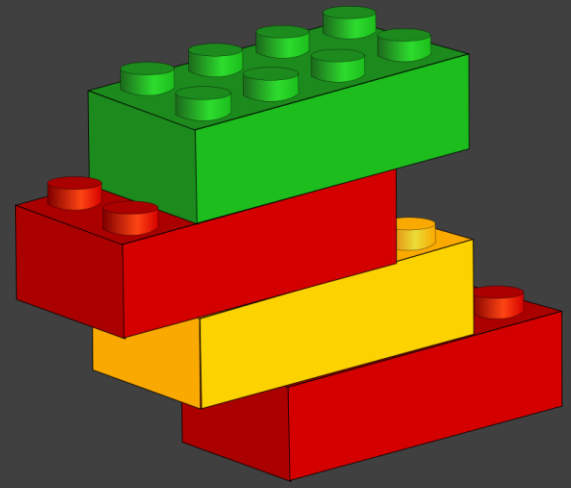


Recurrent Neural Networks



CAPTURE DATA-DOMAIN FEATURES DIRECTLY

- Pre-existing deep learning components can be composed together in an entirely new way, like existing software development
- Deep Learning research is very applied compared to other ML research
- Most innovations are based on finding new architectures or composing existing networks together in new ways
- The people who work on DL tend to be software type people
- Frameworks make DL ubiquitously accessible to anyone who knows a bit of Python, and an end-to-end prediction architecture is contained inside one modality
- Very innovative architectures are possible with deep learning, for example GANs, Autoencoders (explained later).



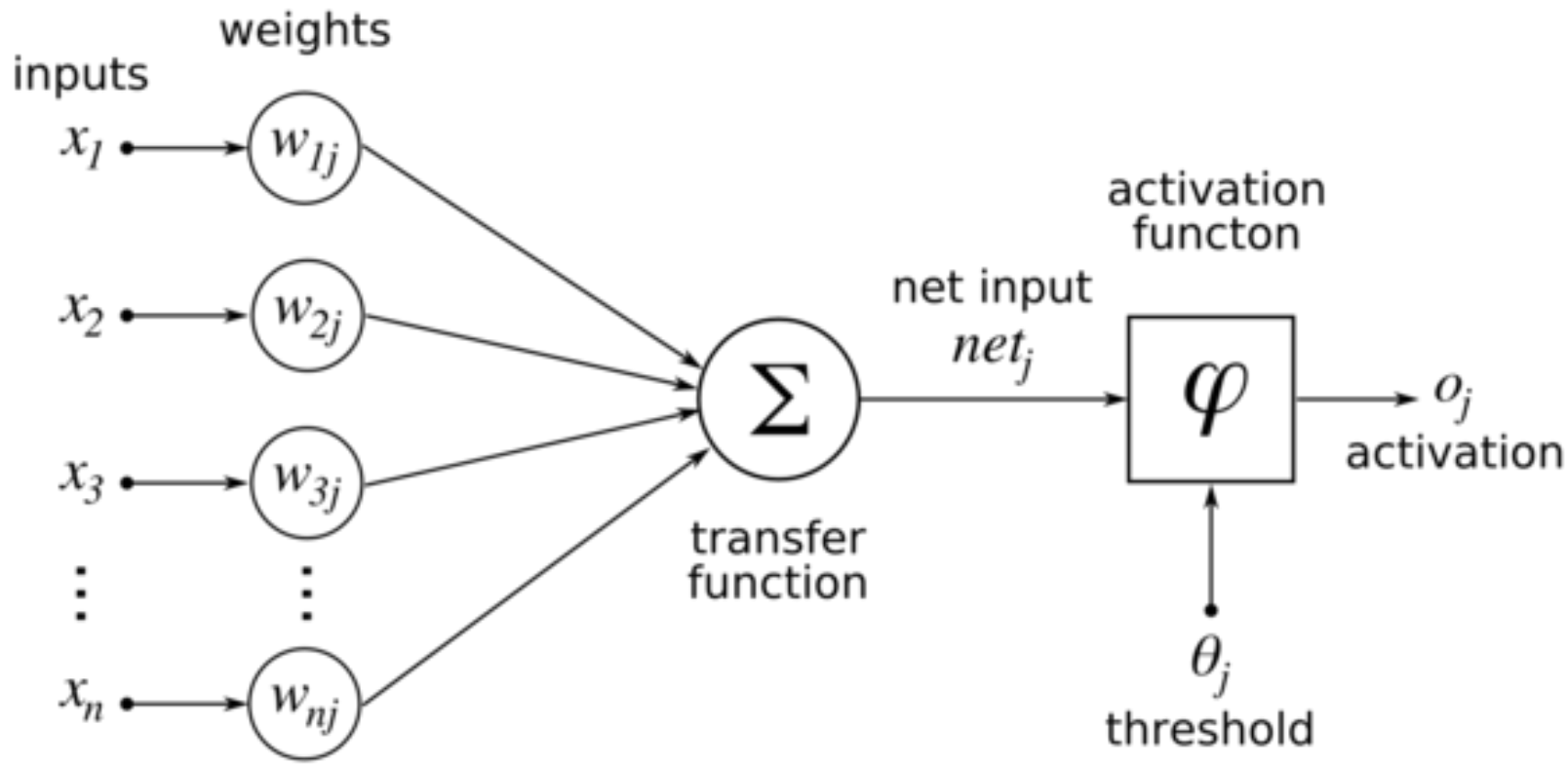
COMPOSABILITY / MORE LIKE SOFTWARE THAN SCIENCE

DEEP LEARNING CONCEPTS

- "Deep" Learning = Neural Networks
- An old technology, but seen differently now
- Universal function approximators (flexible prediction)
- Less emphasis on feature extraction
- Got seriously popular after 2012 due to data+compute explosion
- Particularly good for vision, speech, RL and NLP due to flexible prediction architecture
- Convergence of ML and software development

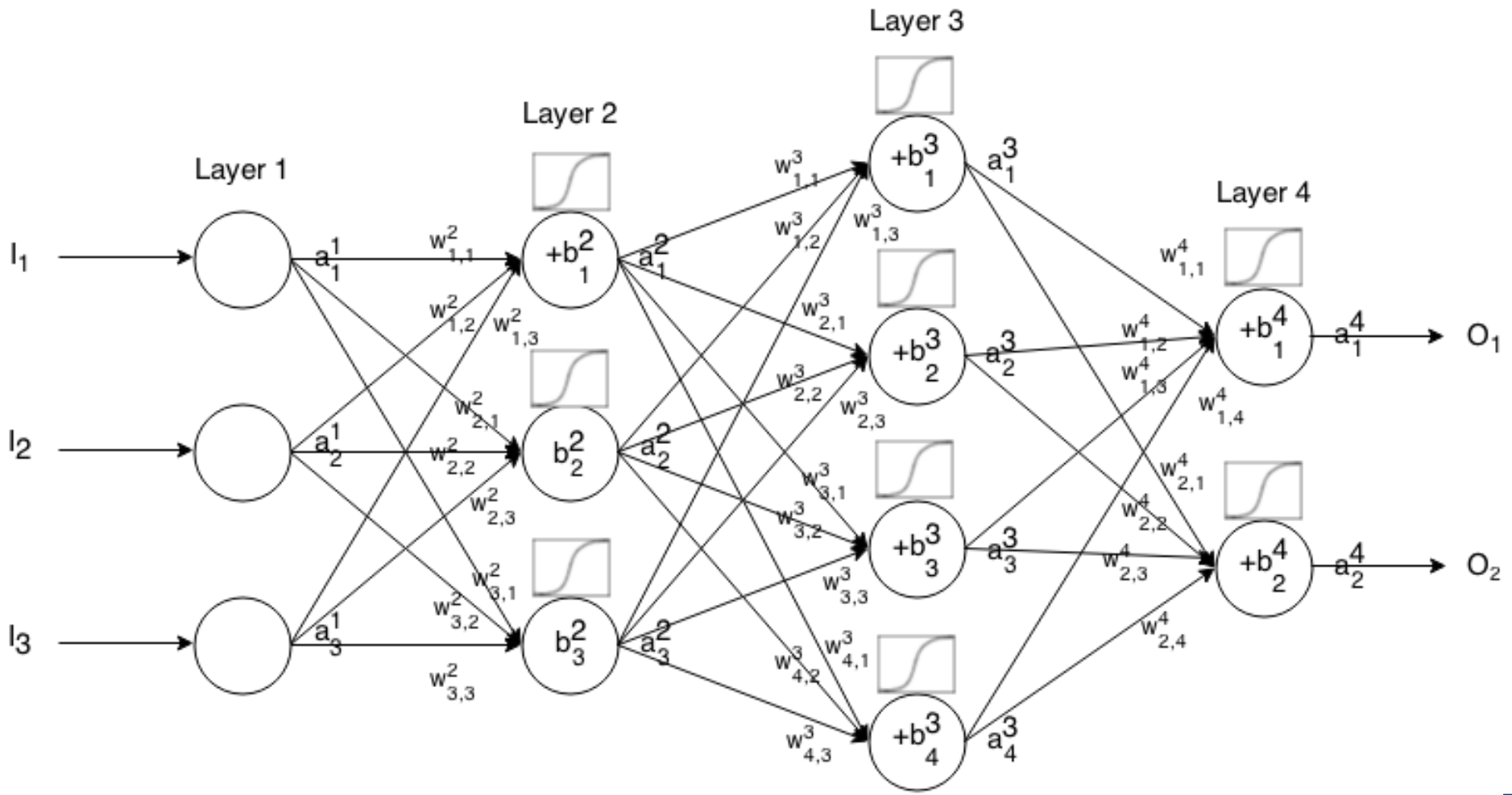
Microsoft

WHAT ARE NEURAL NETWORKS?

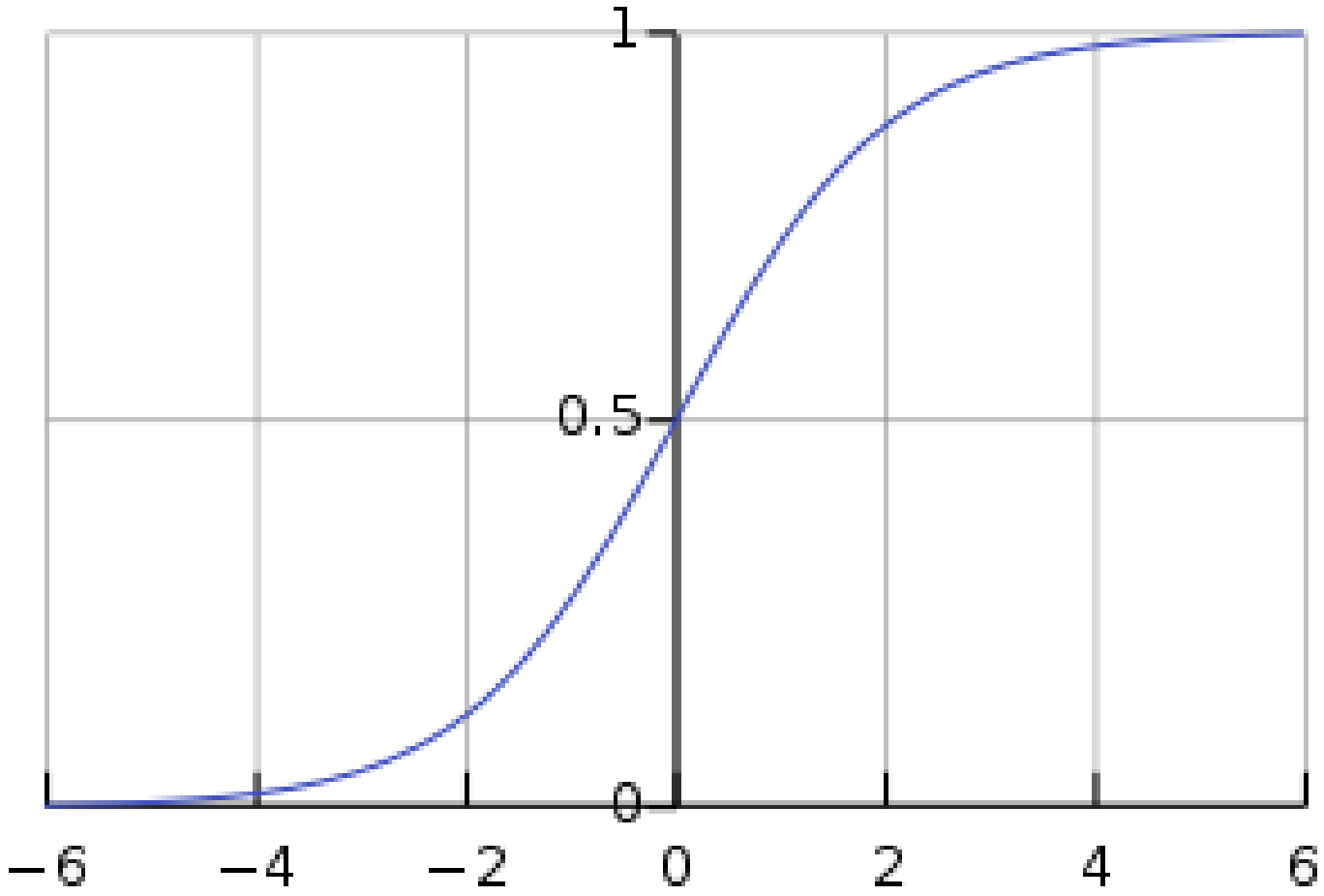


Microsoft

WHAT ABOUT "DEEP" NEURAL NETWORKS?

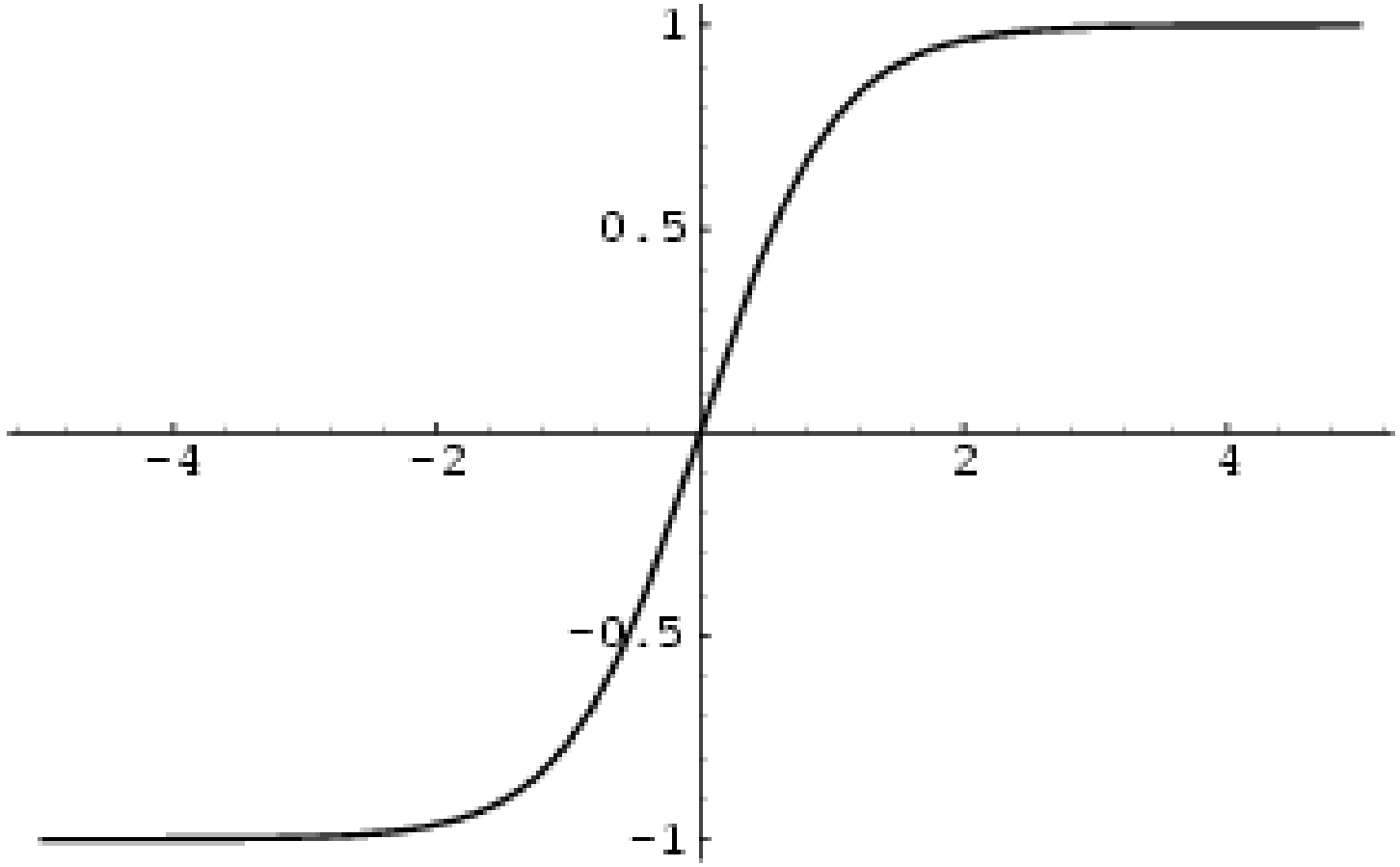


Microsoft



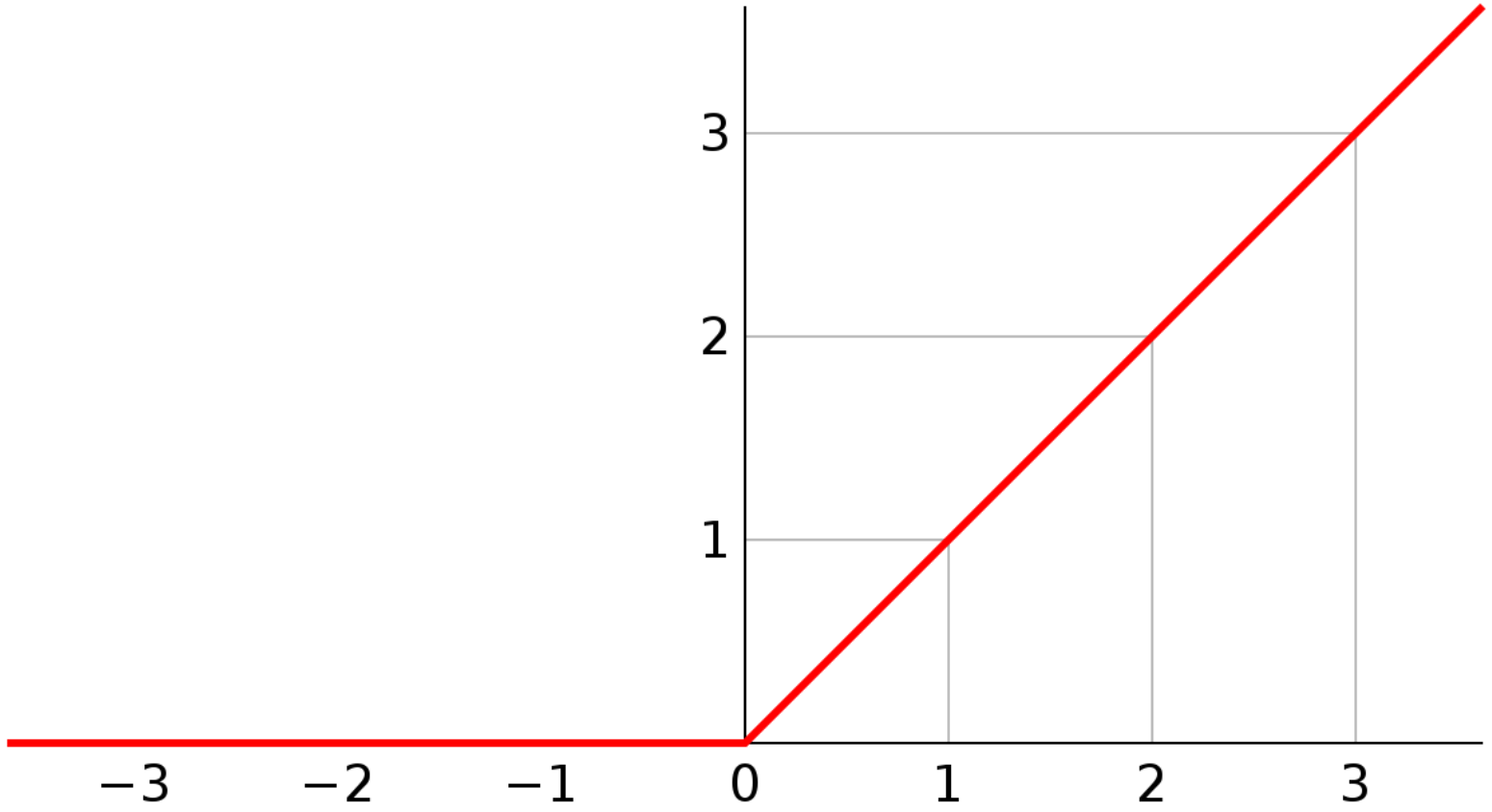
$$\frac{e^x}{e^x + 1}$$

SIGMOID SQUASHING FUNCTION

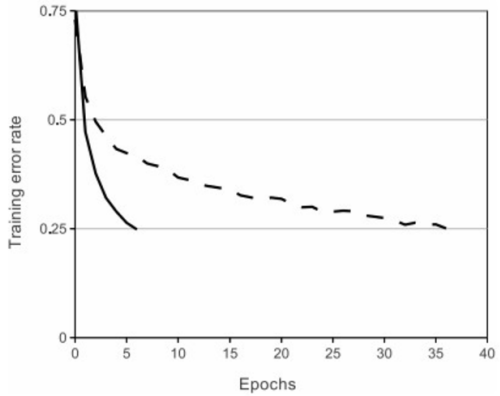


$$\frac{1 - e^{-2x}}{1 + e^{-2x}}$$

TANH SQUASHING FUNCTION



$$f(x) = x^+ = \max(0, x)$$



RELU SQUASHING FUNCTION

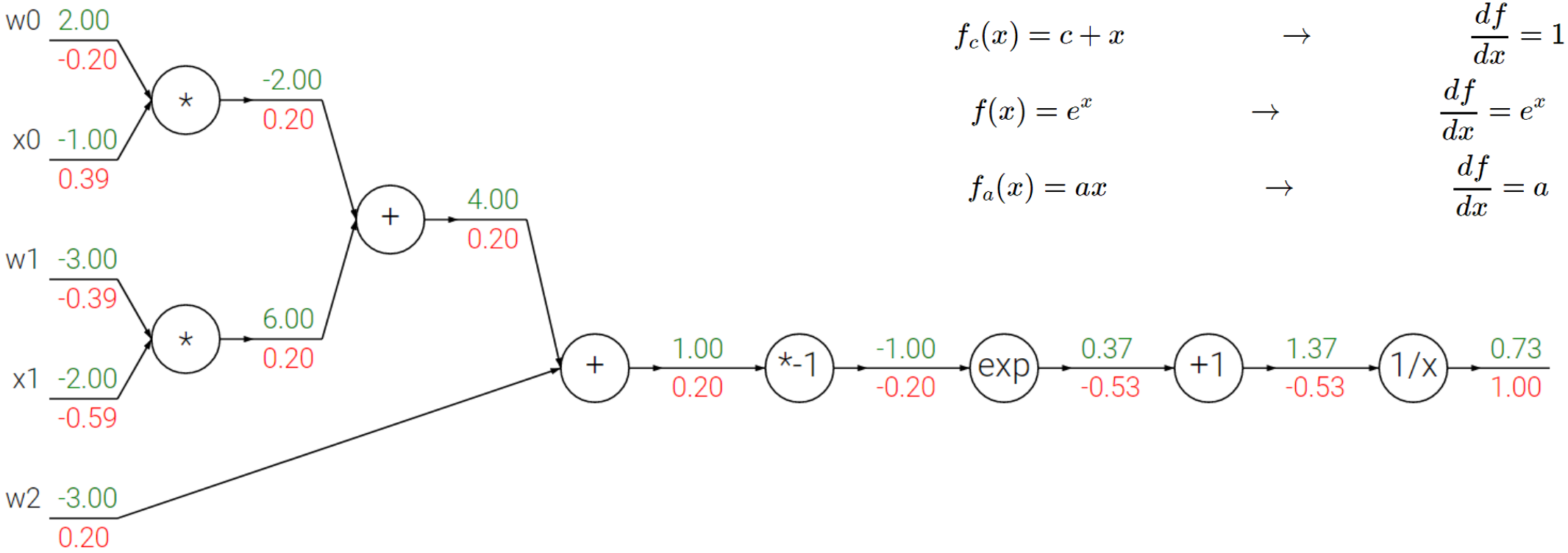
BACKPROPAGATION

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$



$$\frac{\partial E}{\partial w_{jk}}$$

Example circuit for a 2D neuron with a sigmoid activation function. The inputs are [x0,x1] and the (learnable) weights of the neuron are [w0,w1,w2]. As we will see later, the neuron computes a dot product with the input and then its activation is softly squashed by the sigmoid function to be in range from 0 to 1.

Microsoft

$$\Delta w_{jk} = \eta * [\, x_j * (o_k - t_k) * o_k * (1 - o_k) \,]$$

$$\frac{\partial E}{\partial w_{jk}}$$

learning rate

error e_k

derivative of output activation $\varphi_k{}'$

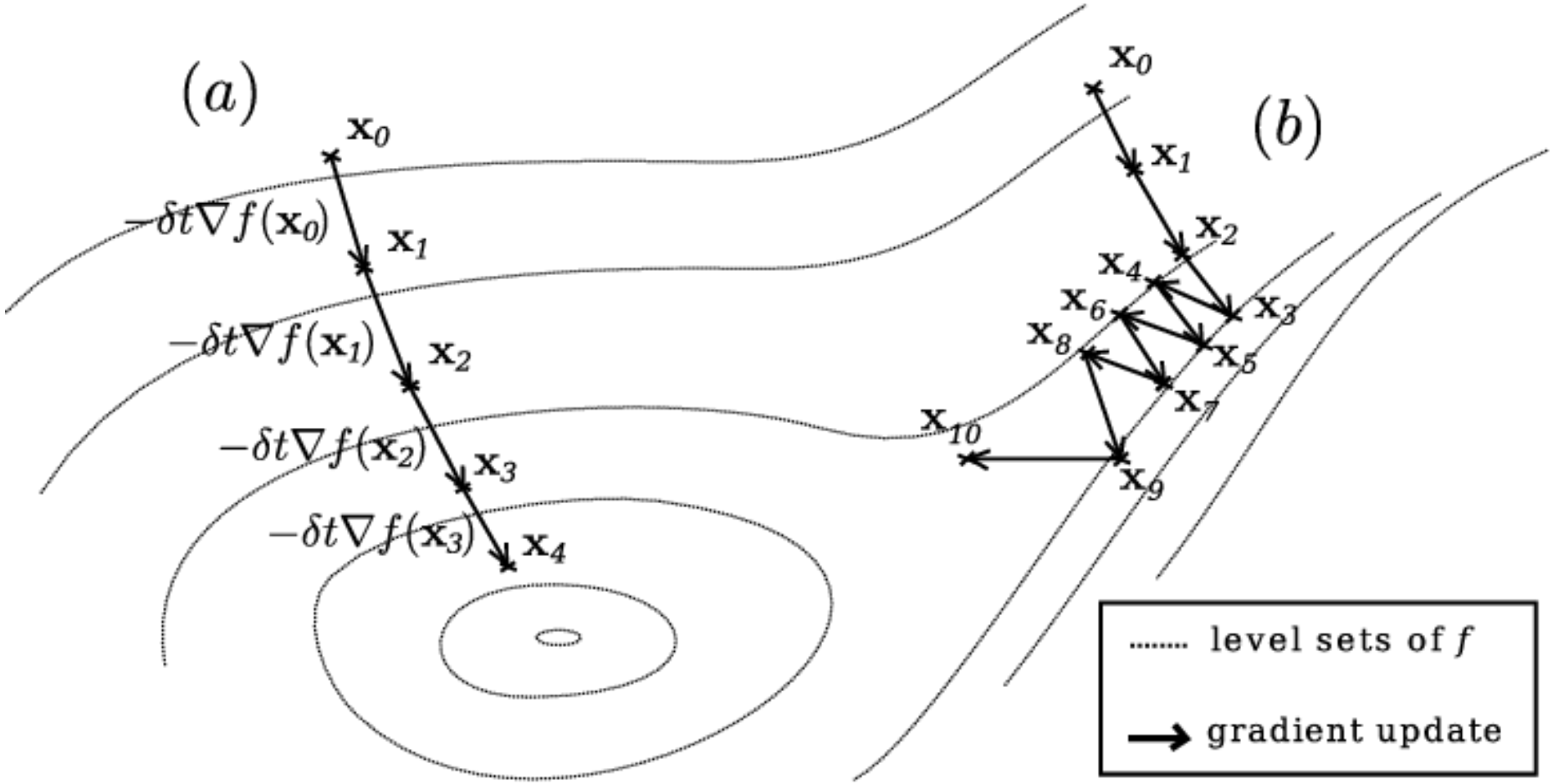
signal δ_k

WEIGHT UPDATE

```
loop maxEpochs times
  for-each training item
    get target values
    compute output values
    compute the gradient of each weight
    use gradient to compute delta for each weight
    update each weight using its delta
  end-for
end-loop
```

BACKPROP ALGORITHM

OPTIMIZATION/GRADIENT DESCENT



WEIGHT UPDATES IN EPOCHS / TRAINING DATA SPLIT INTO MINIBATCHES

Microsoft

DEMO

NEURAL NETWORK PLAYGROUND

Microsoft

DEMO

https://github.com/ecsplendid/deep-learning-cntk-talk

CNTK IRIS DEMO

Microsoft

WHAT IS CNTK?

- DECLARITIVELY DESCRIBE AND TRAIN DEEP NEURAL NETWORKS
- DOES ALL THE HARD WORK FOR YOU
- 80% INTERNAL MS DL WORKLOADS USE CNTK
- 1ST CLASS ON LINUX, WINDOWS, DOCKER
- BRAINSCRIPT, DON'T NEED TO KNOW HOW TO CODE
- C#, PYTHON, R, COMMANDLINE
- KERAS BINDINGS
- V2.2 JUST RELEASED
- OPERATIONALISE WITH THE AZUREML-CLI OR AZURE FUNCTIONS NOW THAT WE SUPPORT C#?!

Microsoft

Microsoft

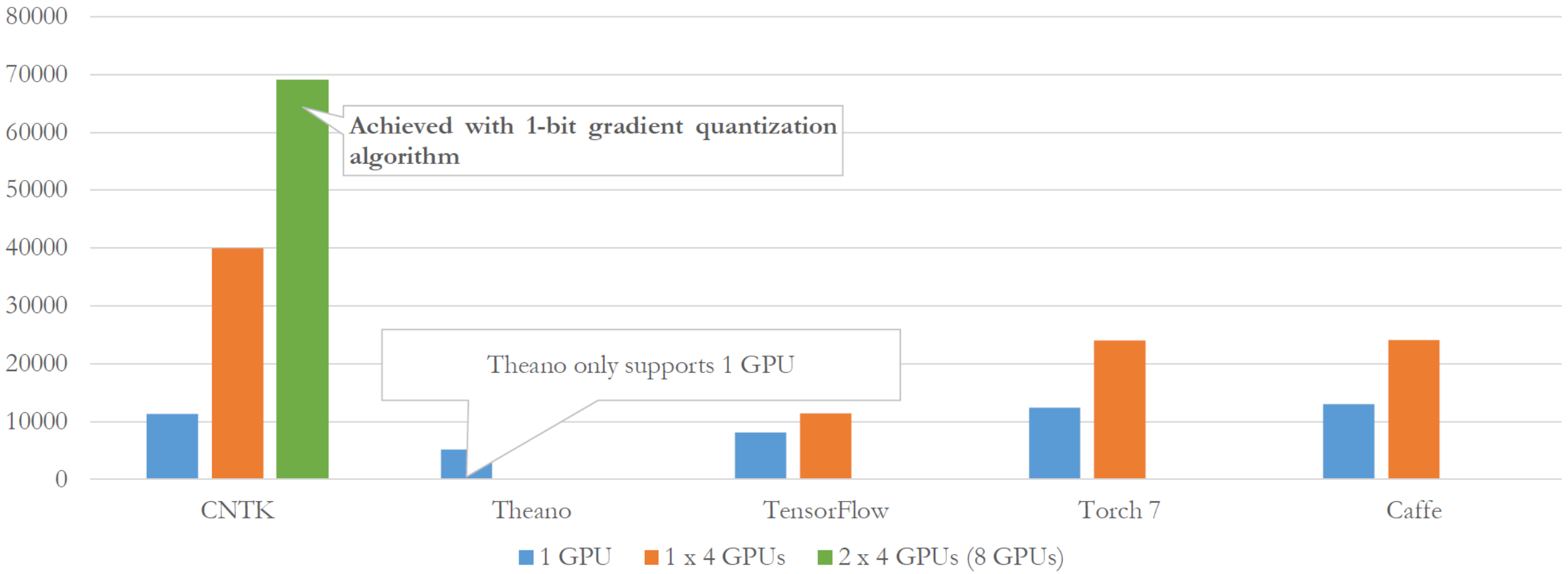
http://dlbench.comp.hkbu.edu.hk/
Benchmarking by HKBU, Version 8
Single Tesla K80 GPU, CUDA: 8.0 CUDNN: v5.1

Caffe: 1.0rc5(39f28e4)
CNTK: 2.0 Beta10(1ae666d)
MXNet: 0.93(32dc3a2)
TensorFlow: 1.0(4ac9c09)
Torch: 7(748f5e3)

	Caffe	CNTK	MxNet	TensorFlow	Torch
FCN5 (1024)	55.329ms	**51.038ms**	60.448ms	62.044ms	52.154ms
AlexNet (256)	36.815ms	**27.215ms**	28.994ms	103.960ms	37.462ms
ResNet (32)	143.987ms	**81.470ms**	84.545ms	181.404ms	90.935ms
LSTM (256) (v7 benchmark)	-	**43.581ms** (44.917ms)	288.142ms (284.898ms)	- (223.547ms)	1130.606ms (906.958ms)

THE FASTEST TOOLKIT

Microsoft



MOST SCALABLE TOOLKIT (2016)

INSTALLING CNTK

- GOOGLE "CNTK INSTALL" (WITH BING)
- USE THE "SCRIPT DRIVEN INSTALLATION"

Microsoft

WHEN TO USE DEEP LEARNING FRAMEWORKS

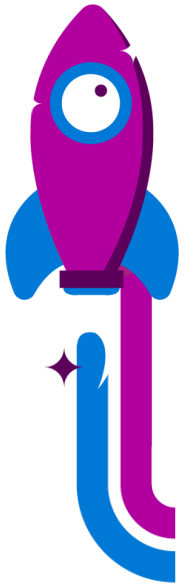
- Sequence modelling (speech, language, time-series)
- Complex vision tasks (localisation, detection)
- Novel prediction architectures
- Generative models
- Reinforcement learning
- … and many more!

Microsoft

DEEP LEARNING ON AZURE CLOUD

- Data Science Virtual Machine (Ubuntu and Windows)
- Batch AI Training Service
- FPGA/ASICS are coming soon
- AzureML supports discriminative neural networks i.e. regression and classification
- R Server includes a trained deep learning image featurizer

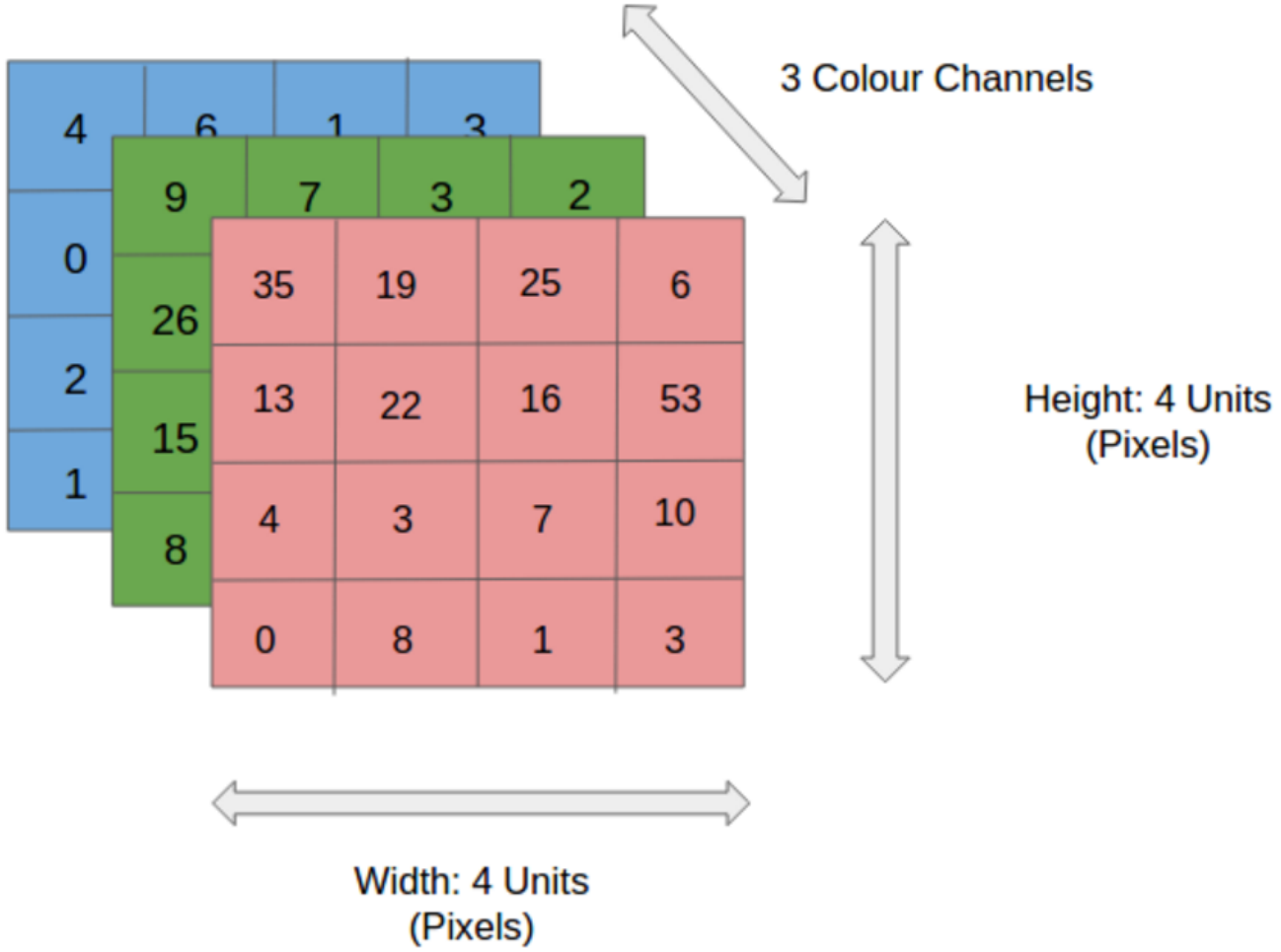
Microsoft



How do convolutional neural networks work?

Microsoft

PREPARE DATASET OF IMAGES

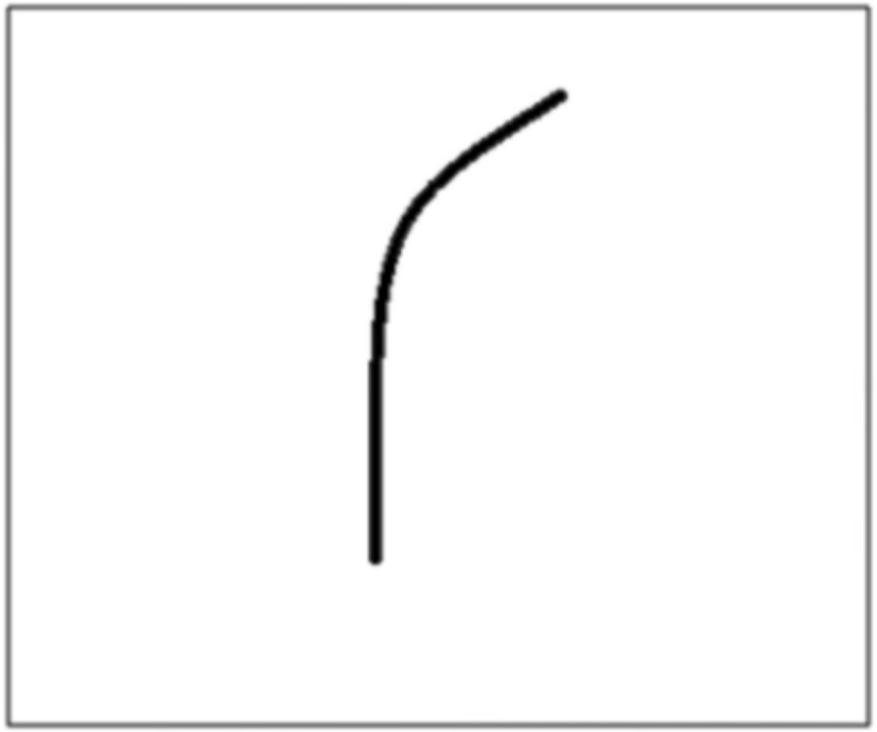


Microsoft

Microsoft

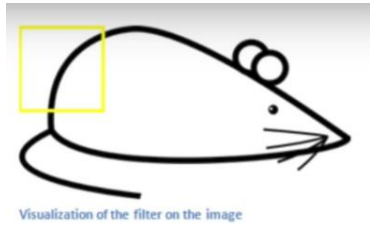
0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter



Visualization of a curve detector filter

CONVOLUTION FILTER



Visualization of the filter on the image

$$(50*30)+(50*30)+(50*30)+(20*30)+(50*30) = 6600$$

Visualization of the receptive field

0	0	0	0	0	0	30
0	0	0	0	50	50	50
0	0	0	20	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0

Pixel representation of the receptive field

*

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter

CONVOLUTION FILTER MATCH

Visualization of the filter on the image

MULTIPLY AND SUMMATION = 0

0	0	0	0	0	0	0
0	40	0	0	0	0	0
40	0	40	0	0	0	0
40	20	0	0	0	0	0
0	50	0	0	0	0	0
0	0	50	0	0	0	0
25	25	0	50	0	0	0

Pixel representation of receptive field

*

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

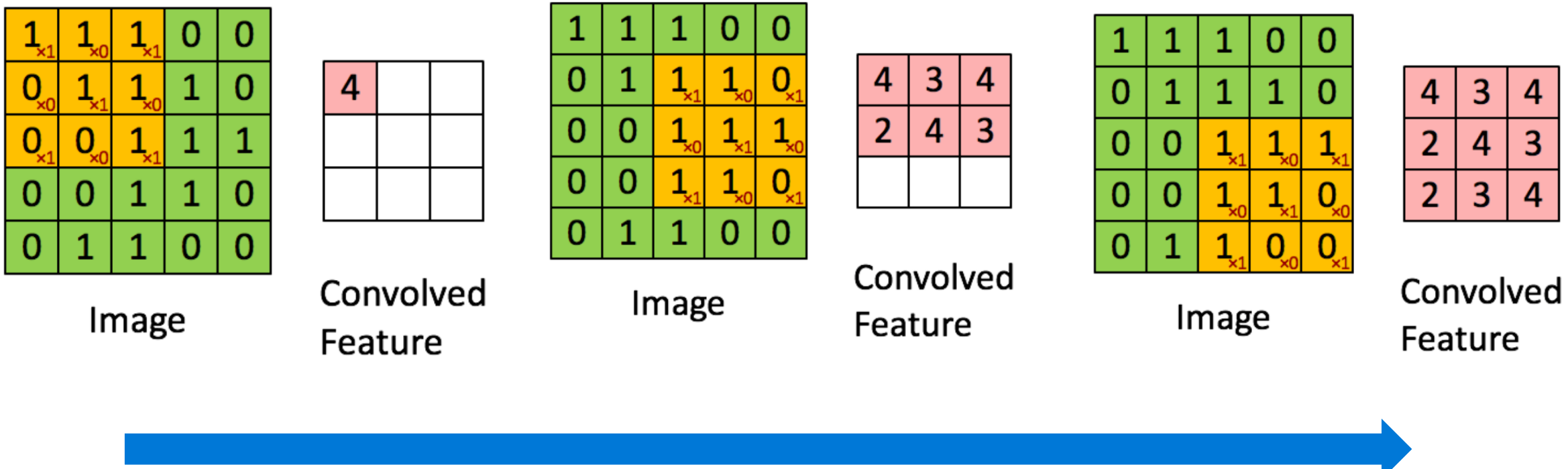
Pixel representation of filter

CONVOLUTION FILTER NO MATCH

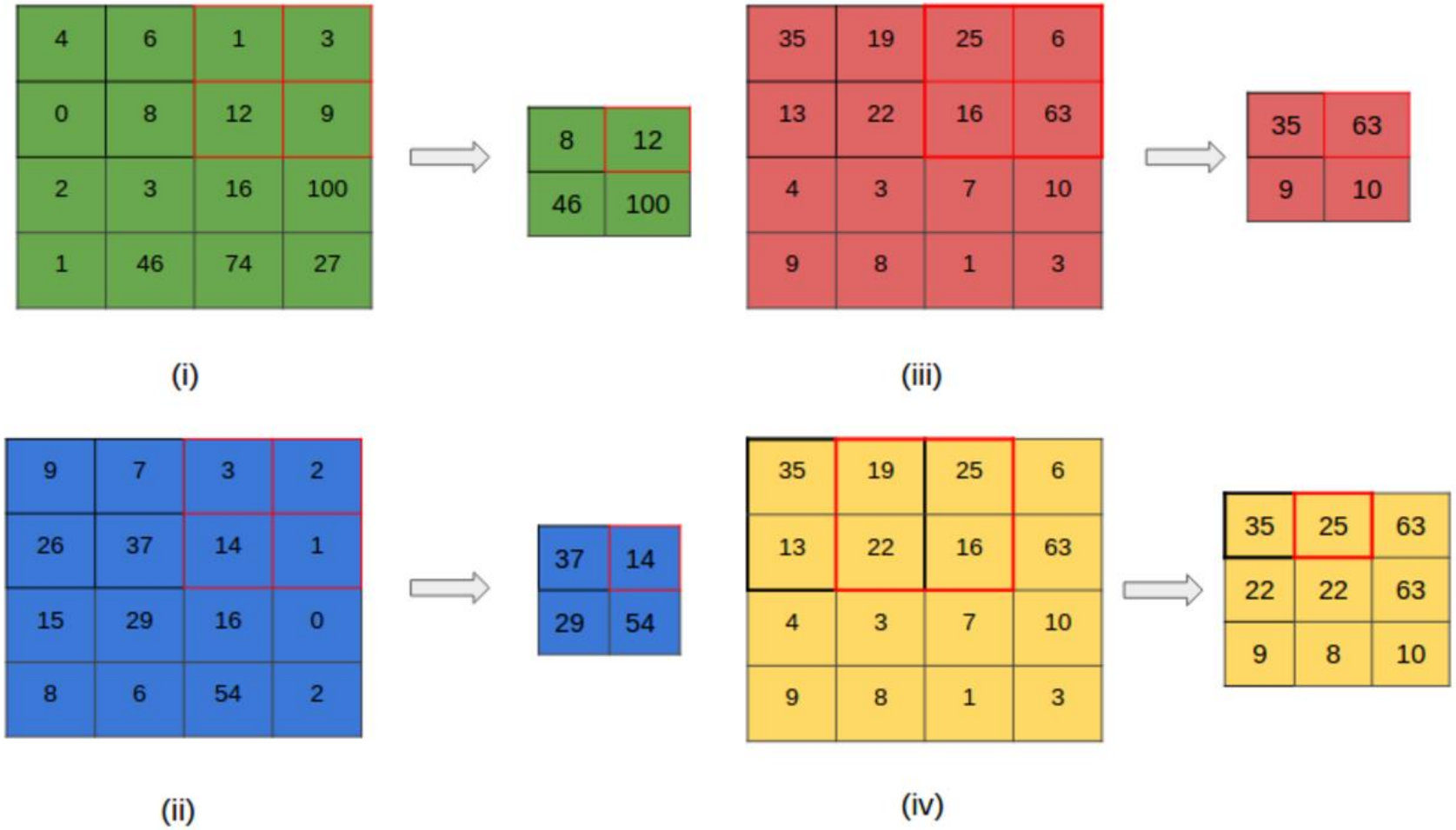
DEMO

CONVOLUTION FILTERING IN MATLAB

Microsoft

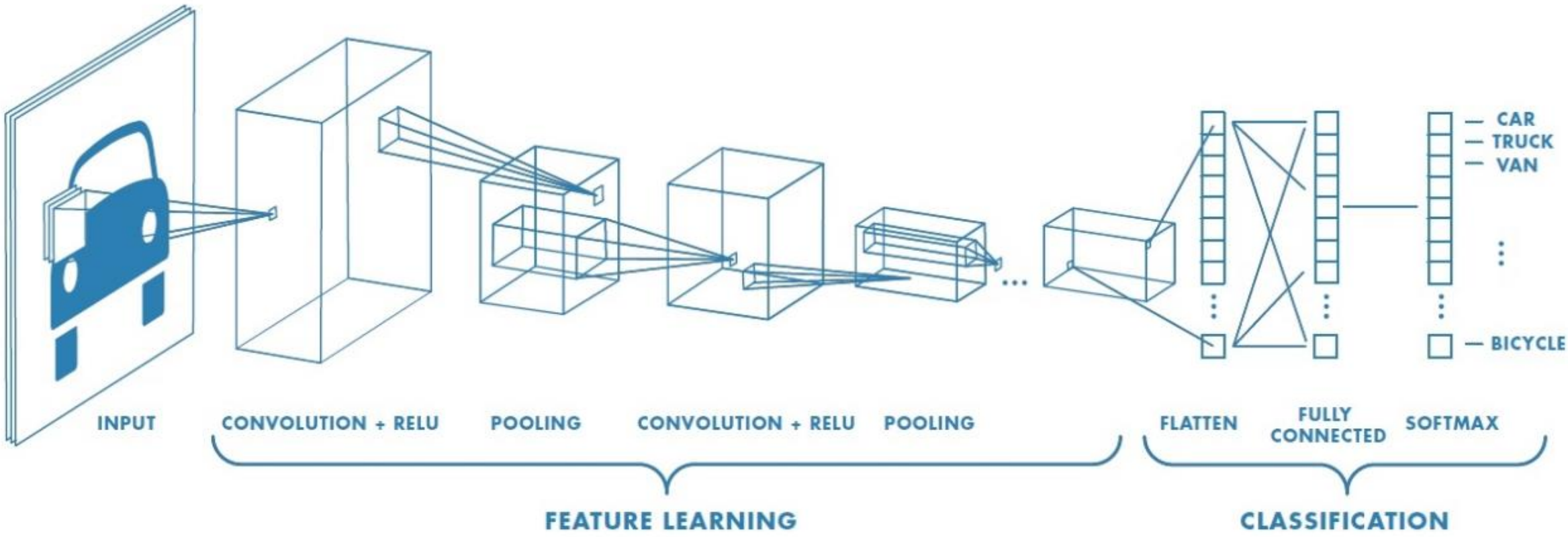


CONVOLUTION


4 6 1 3
0 8 12 9
2 3 16 100
1 46 74 27
8 12
46 100
(i)
35 19 25 6
13 22 16 63
4 3 7 10
9 8 1 3
35 63
9 10
(iii)
9 7 3 2
26 37 14 1
15 29 16 0
8 6 54 2
37 14
29 54
(ii)
35 19 25 6
13 22 16 63
4 3 7 10
9 8 1 3
35 25 63
22 22 63
9 8 10
(iv)

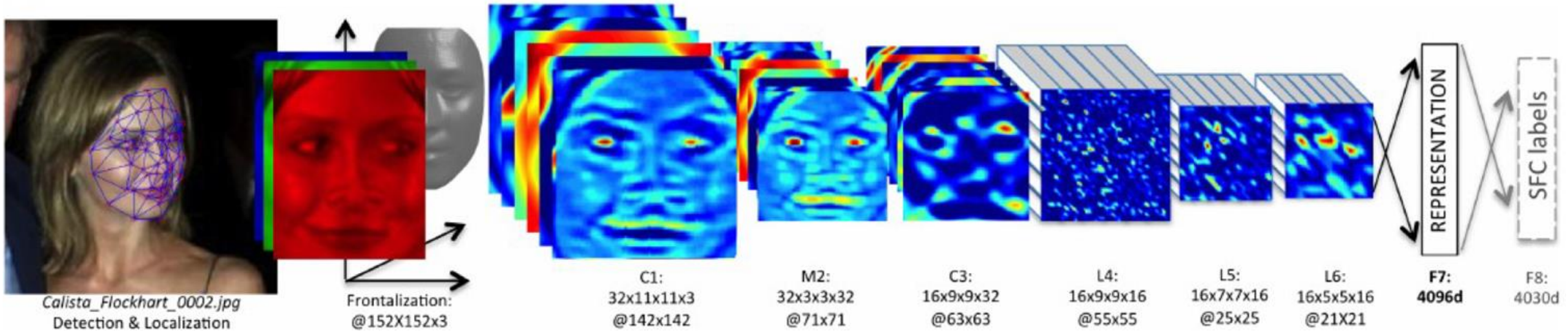
POOLING

Image Classification Example



DEEP FACE

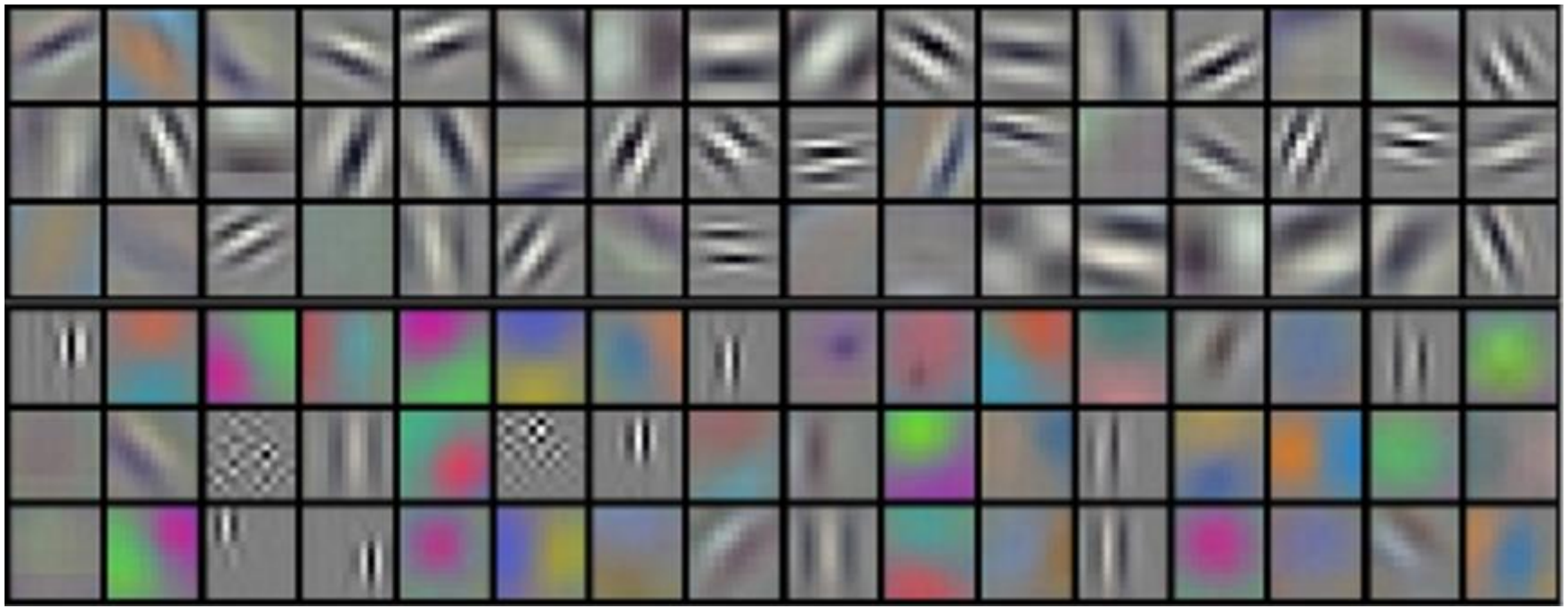
Taigman et al. CVPR 2014



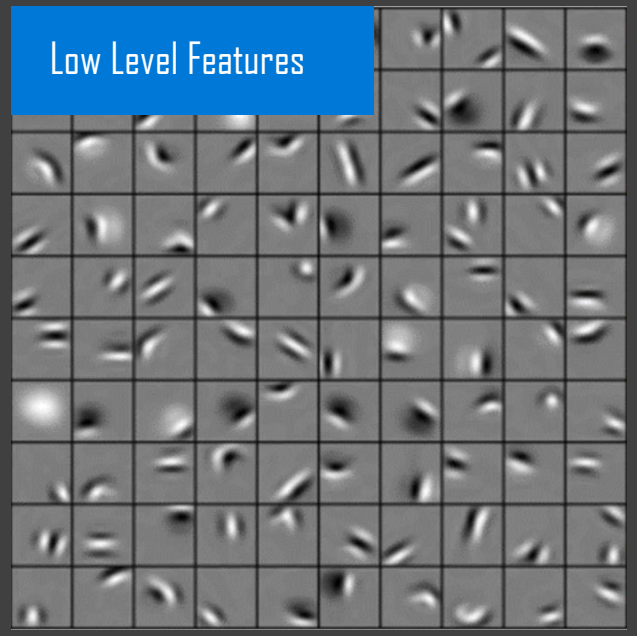
- Alignment
- CNN
- Metric Learning

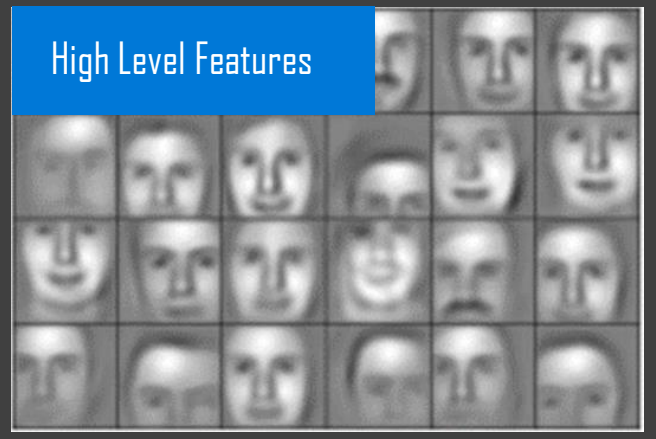
Microsoft

VISUALISING THE FILTERS



Microsoft

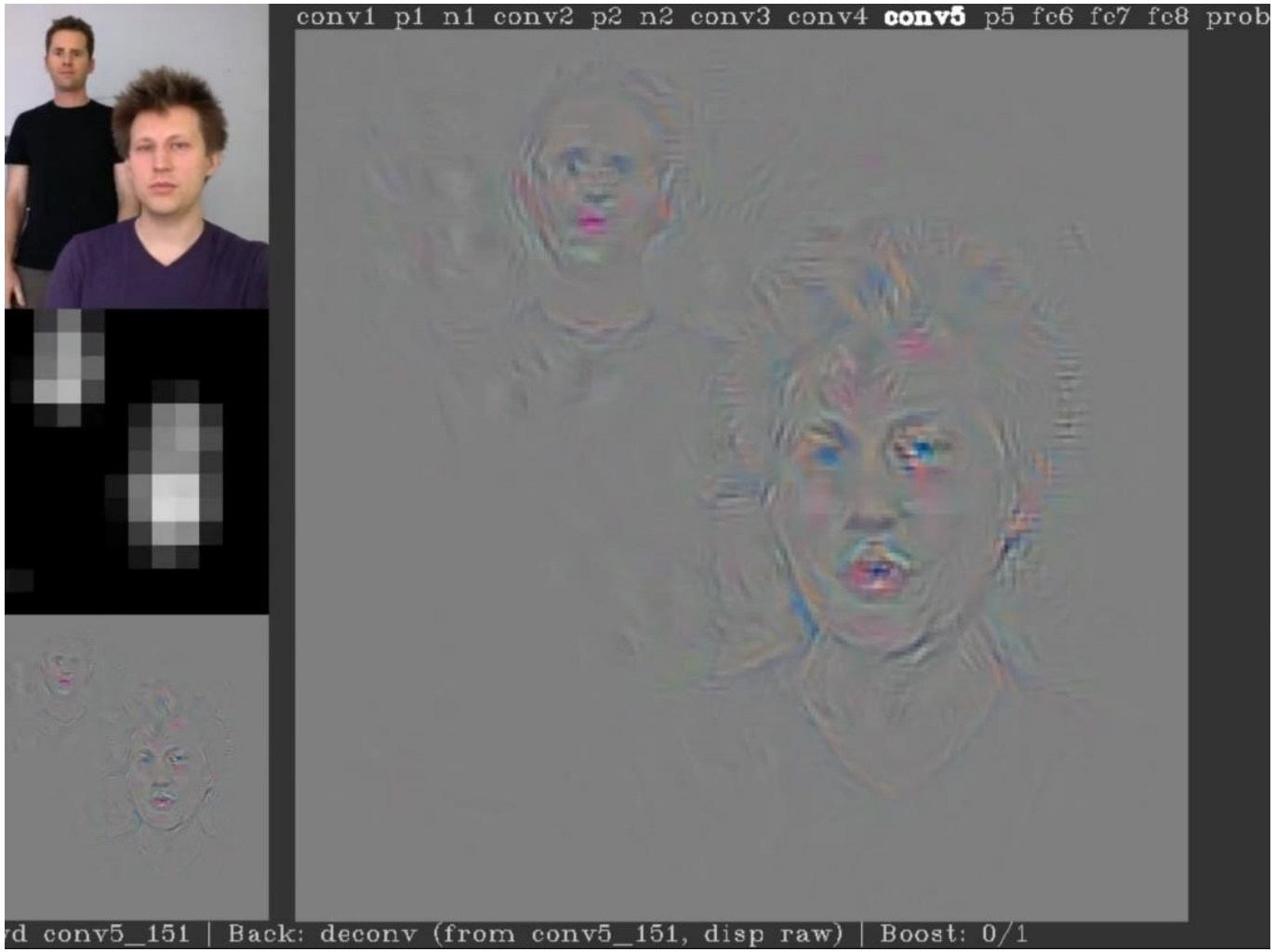




IF IT'S "DEEP" - IT HAS MORE THAN ONE LAYER OF FEATURE TRANSFORMATION

Microsoft

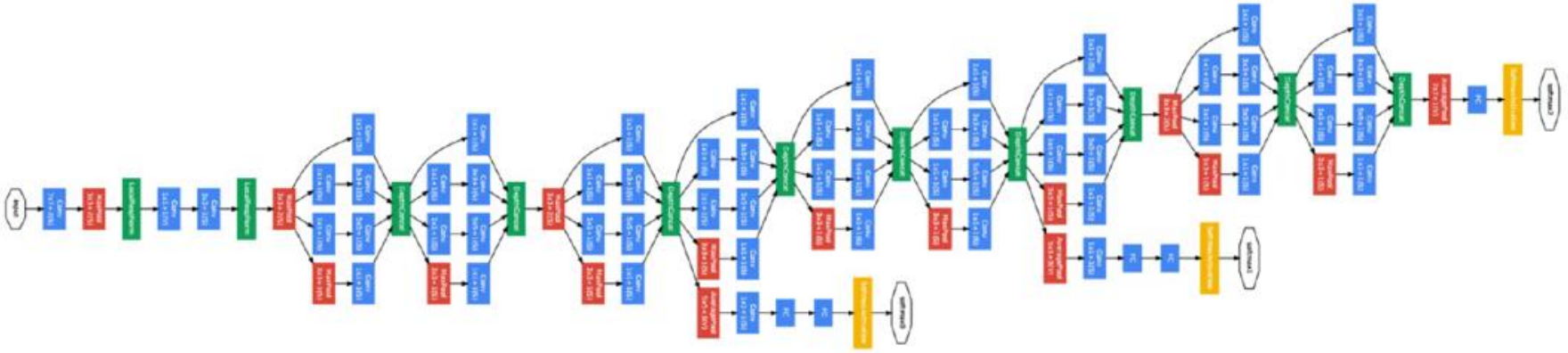
DEEP VISUALISATION TOOLBOX



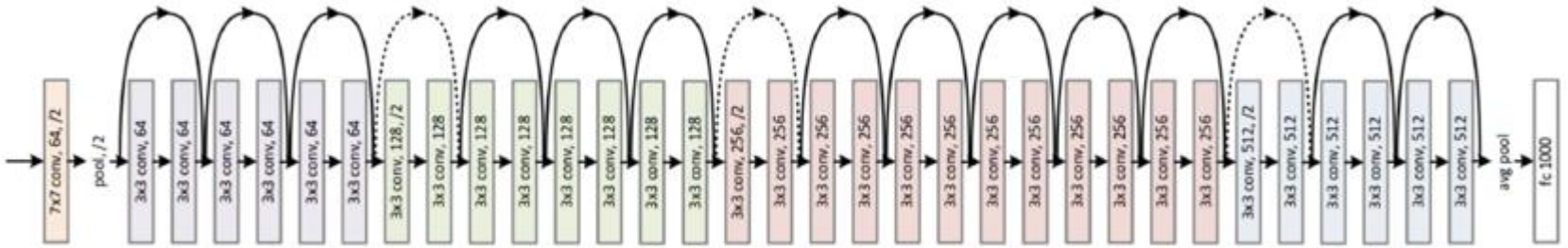
https://www.youtube.com/watch?v=AgkfIQ4IGaM

Microsoft

VGGNet, Oxford Uni, 2014



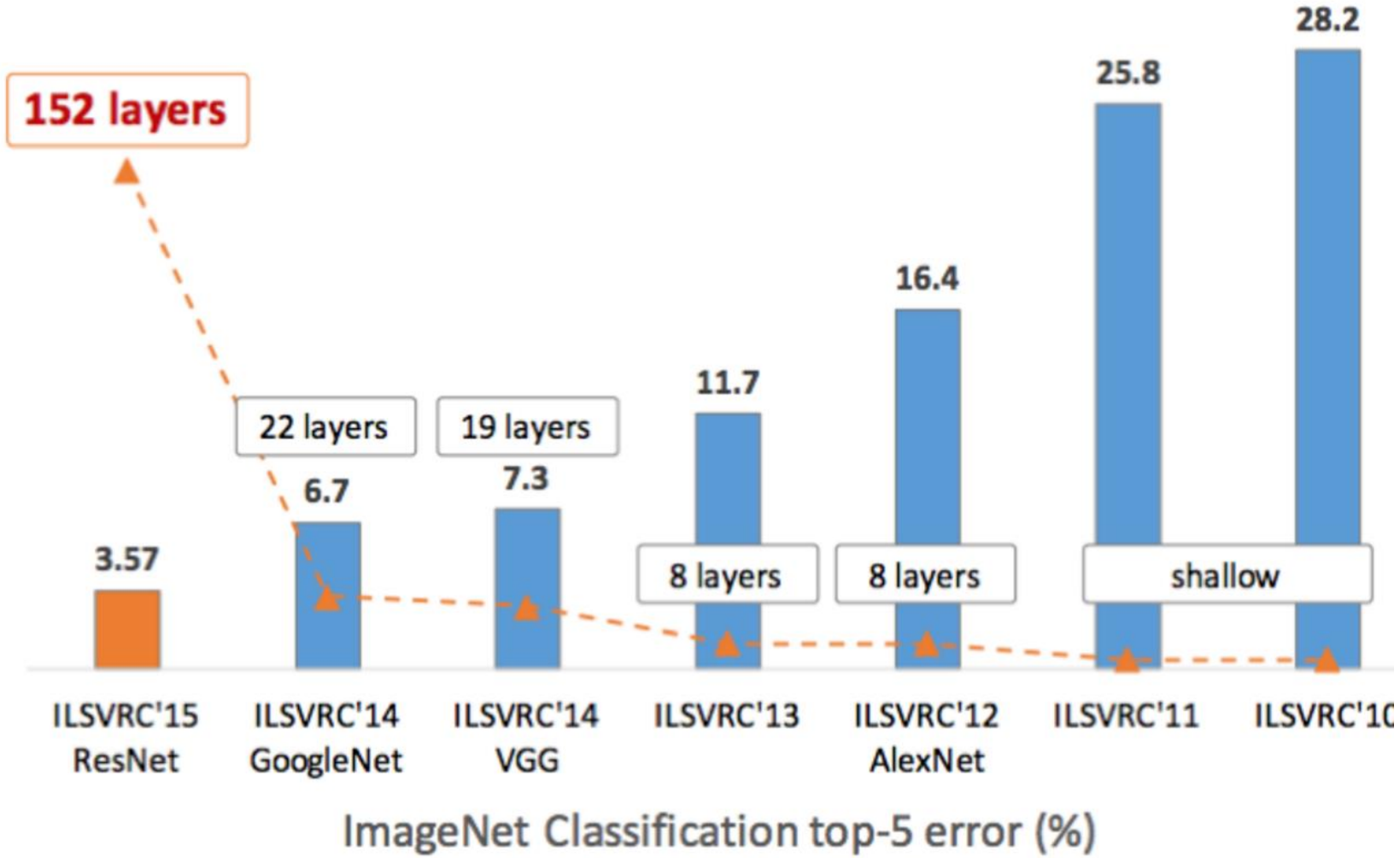
GoogLeNet, Google, 2015



ResNet, Microsoft, 2016

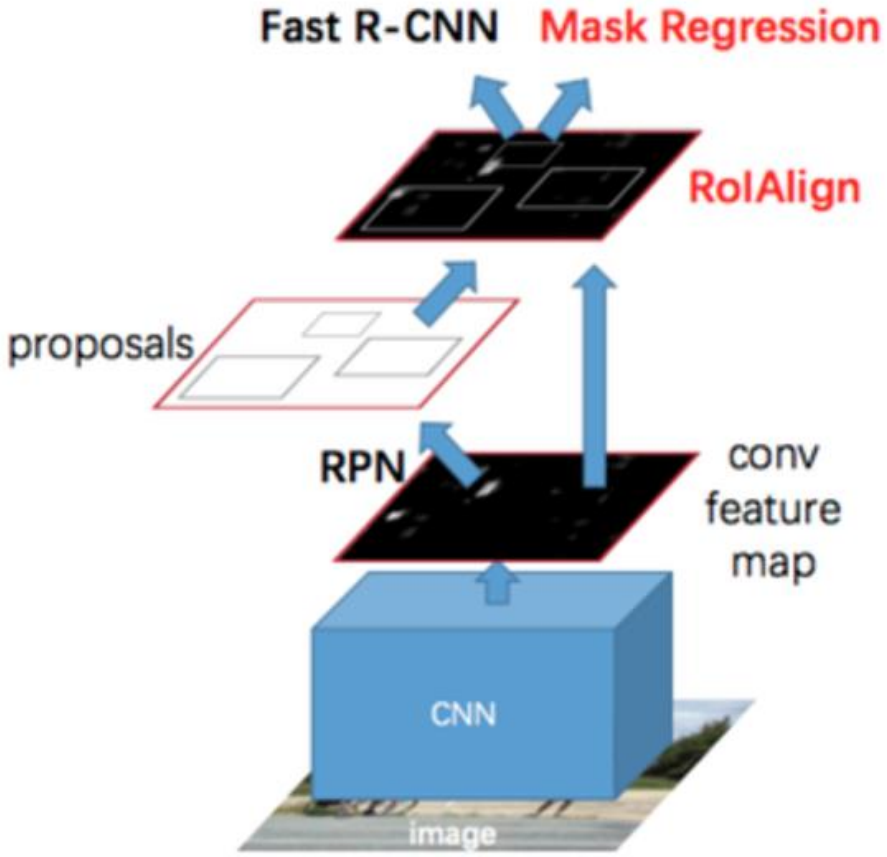
WINNING ARCHITECTURES

RESOLUTION OF DEPTH



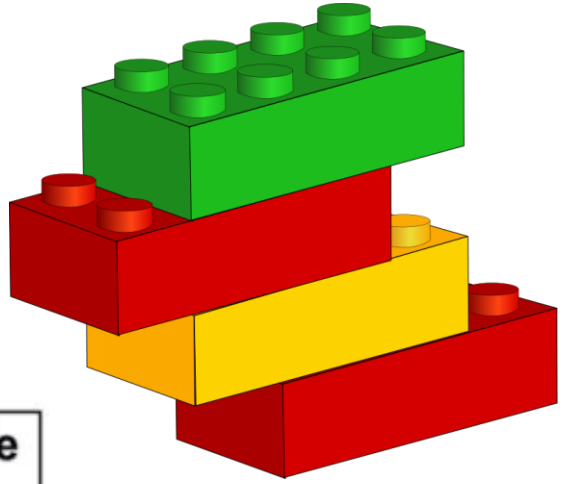
Microsoft

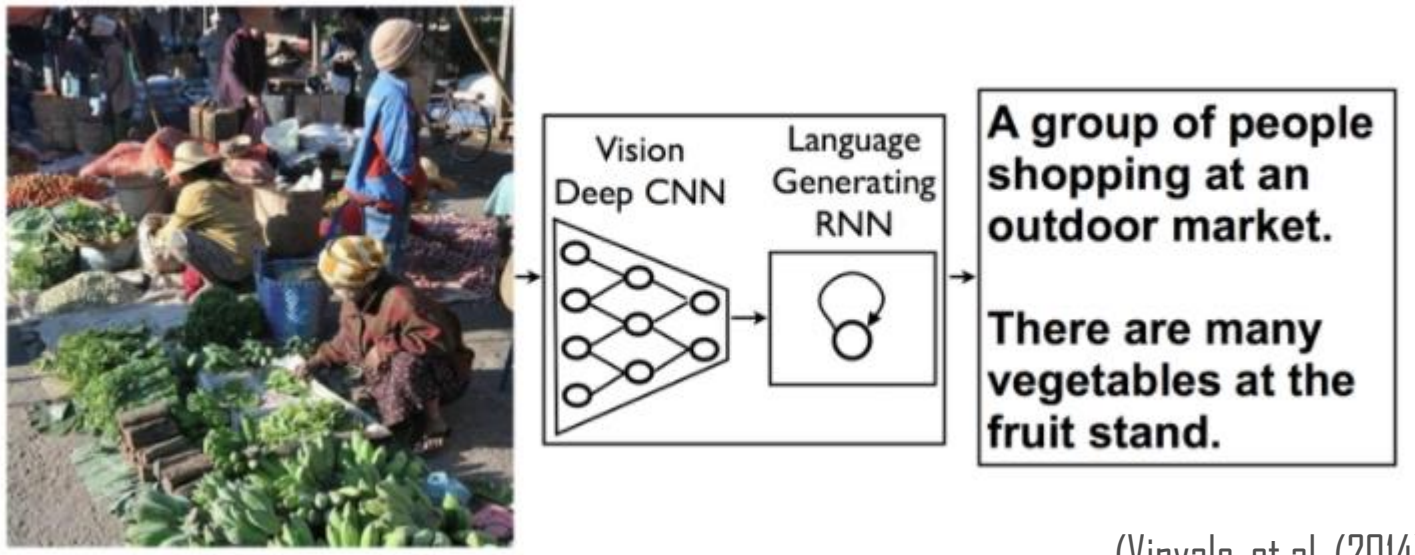
MASK R-CNN ARCHITECTURE (2017)



ARCHITECTURE COMPOSABILITY

BUILDING PREDICTIVE ARCHITECTURES LIKE LEGO BLOCKS





(Vinyals, et al. (2014))

Microsoft

Microsoft

COUNTING SHEEP

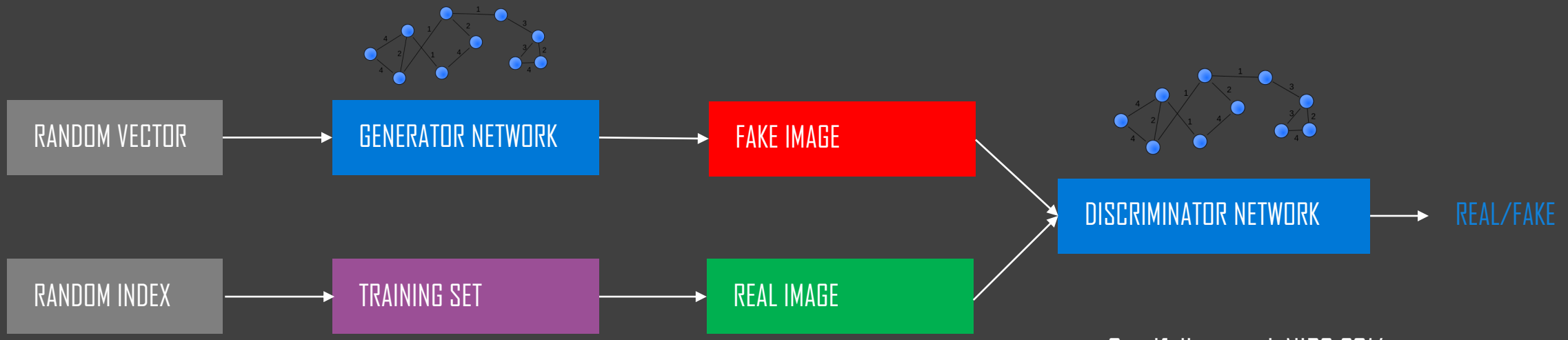
IS DEEP LEARNING A FAD?

- The way we think about neural networks now is totally different to 30 years ago
- Previous frequentist algorithms were just learning weighted combinations of hand-crafted features
- NNs learn a hierarchy of representations which work really well in many domains
- Before we used to talk about classification and regression, now we talk about *predictive architectures, many of which are game-changing.*

Microsoft

TWO INNOVATIVE DL ARCHITECTURES

- An example; generative adversarial networks
- Game-theoretical approach to deep learning

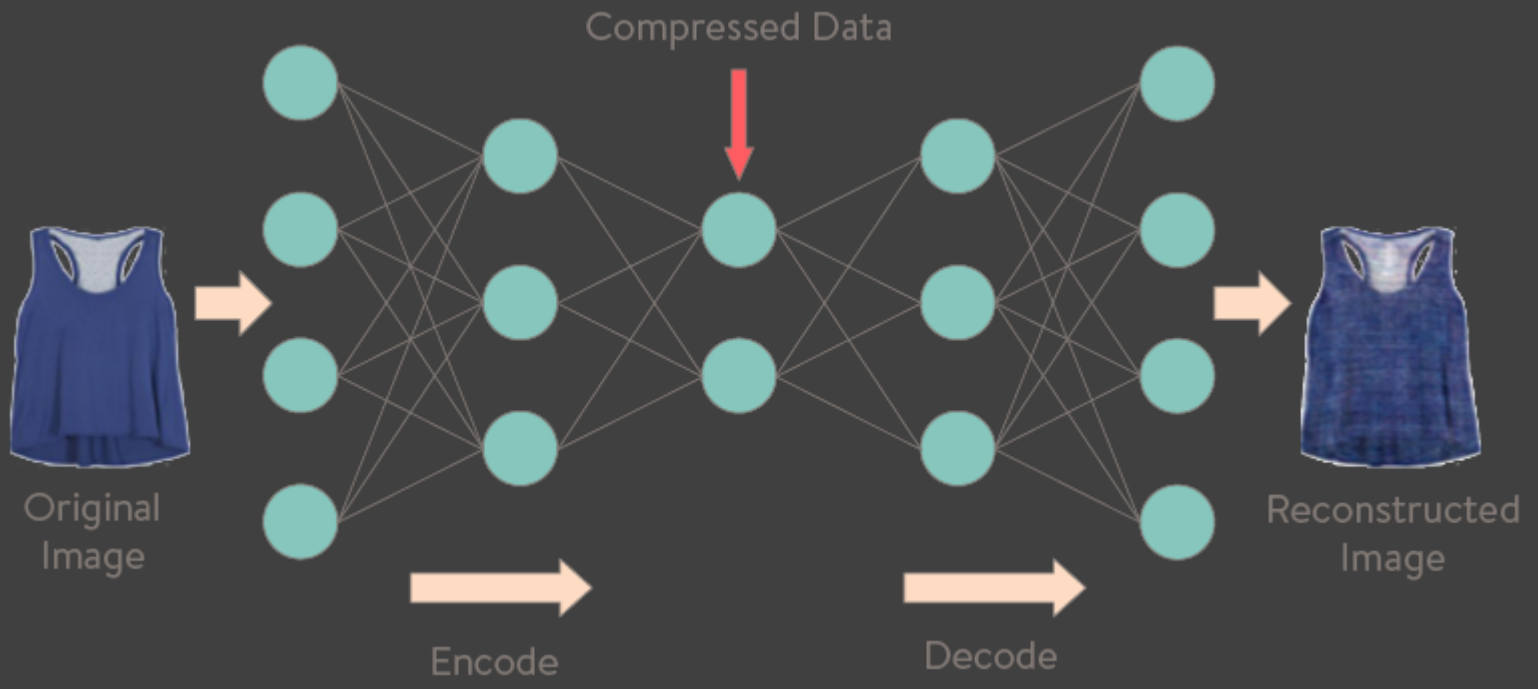


Goodfellow et al. NIPS 2014

VERY ADVANCED NOVEL ARCHITECTURES (GANS)

IMAGES GENERATED WITH GANS

- Autoencoders are an unsupervised method to compress information into an "embedding" similar to unsupervised dimensionality reduction.



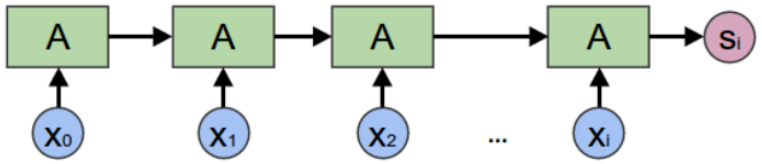
AUTOENCODERS

PARADGM-SHIFT 

- Three narratives currently exist to describe deep learning
  - Neuroscience
  - Probabilistic
  - Manifold
- The *differentiable programming* narrative is emerging

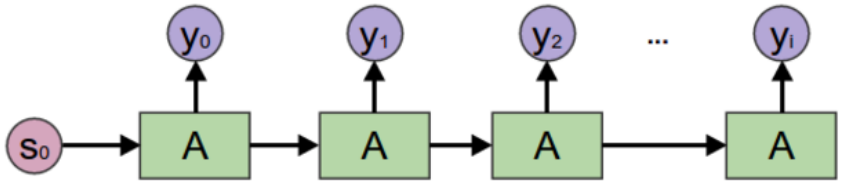
FUNCTIONAL PROGRAMMING IN DL

http://colah.github.io/posts/2015-09-NN-Types-FP/



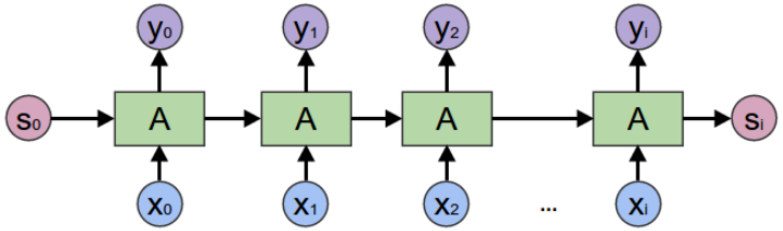
fold = Encoding RNN

Haskell: `foldl a s`



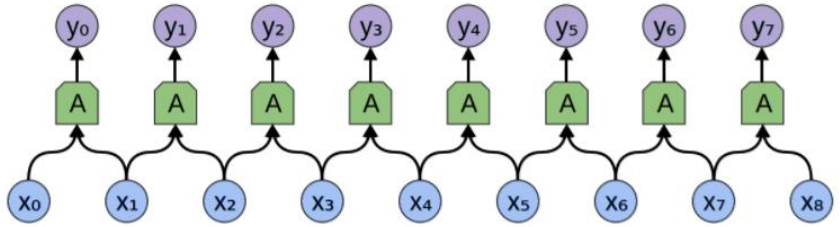
unfold = Generating RNN

Haskell: `unfoldr a s`



Accumulating Map = RNN

Haskell: `mapAccumR a s`



Windowed Map = Convolutional Layer

Haskell: `zipWith a xs (tail xs)`

Microsoft

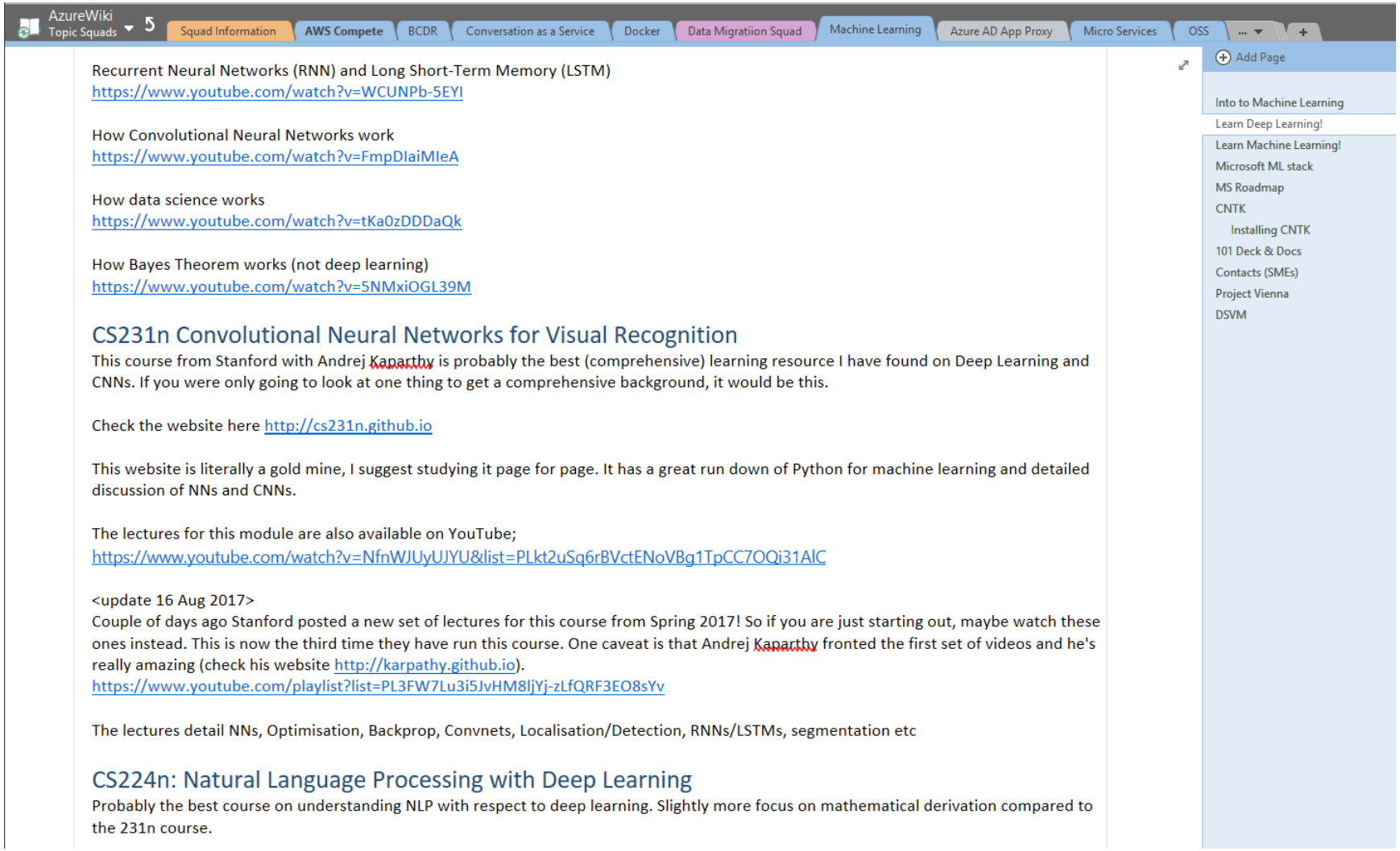
DEMO

CNTK MNIST DEMO

Microsoft

LEARNING MORE; AZUREWIKI, ML DISTRIBUTION LIST

machine-learning@microsoft.com



Microsoft

THANK YOU

tim.scarfe@microsoft.com