Deep Learning: Overview and Hardware Implementation

MSc. Manuel Ríos

Universidad Rafael Landívar

October 17th, 2017

Outline

What is what?

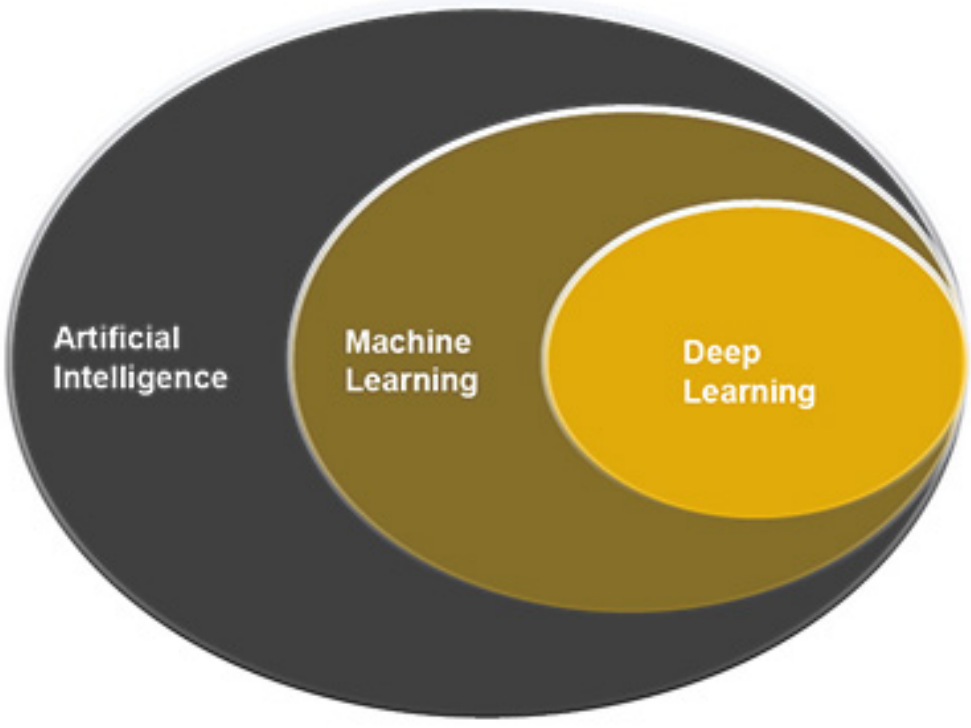
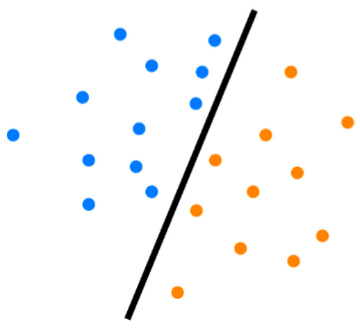


Figure 1 : Deep Learning is a sub-field of Machine Learning, which in turn is a sub-field of Artificial Intelligence.

Outline

Supervised, Unsupervised and Reinforcement Learning

Supervised	Unsupervised	Reinforcement
Learning known patterns	Learning unknown patterns	Generating data Learning patterns



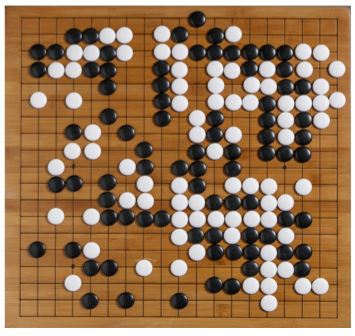


Figure 2 : Types of Machine Learning.

Supervised Learning: Classification and Regression

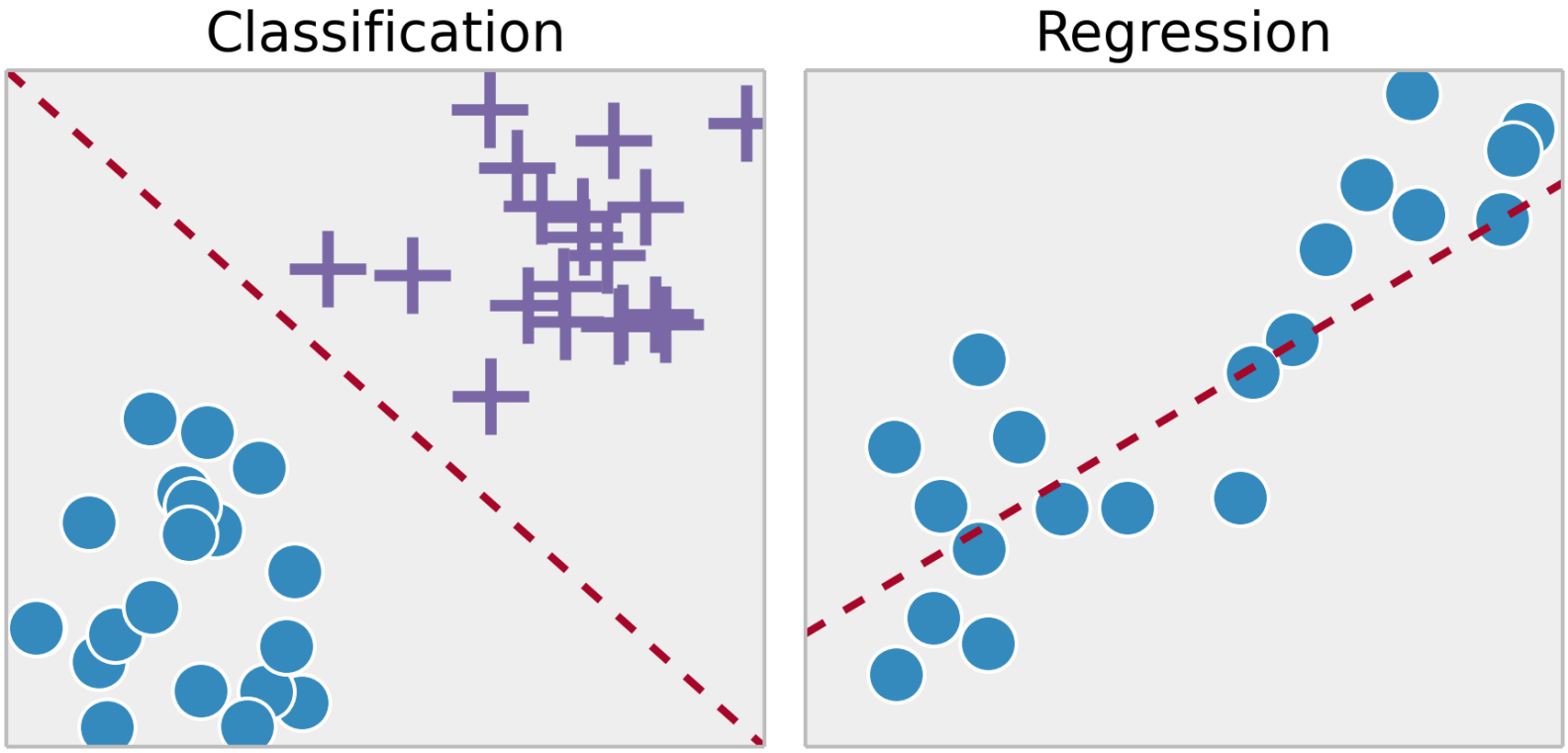


Figure 3 : Classification and Regression.

Classification Datasets

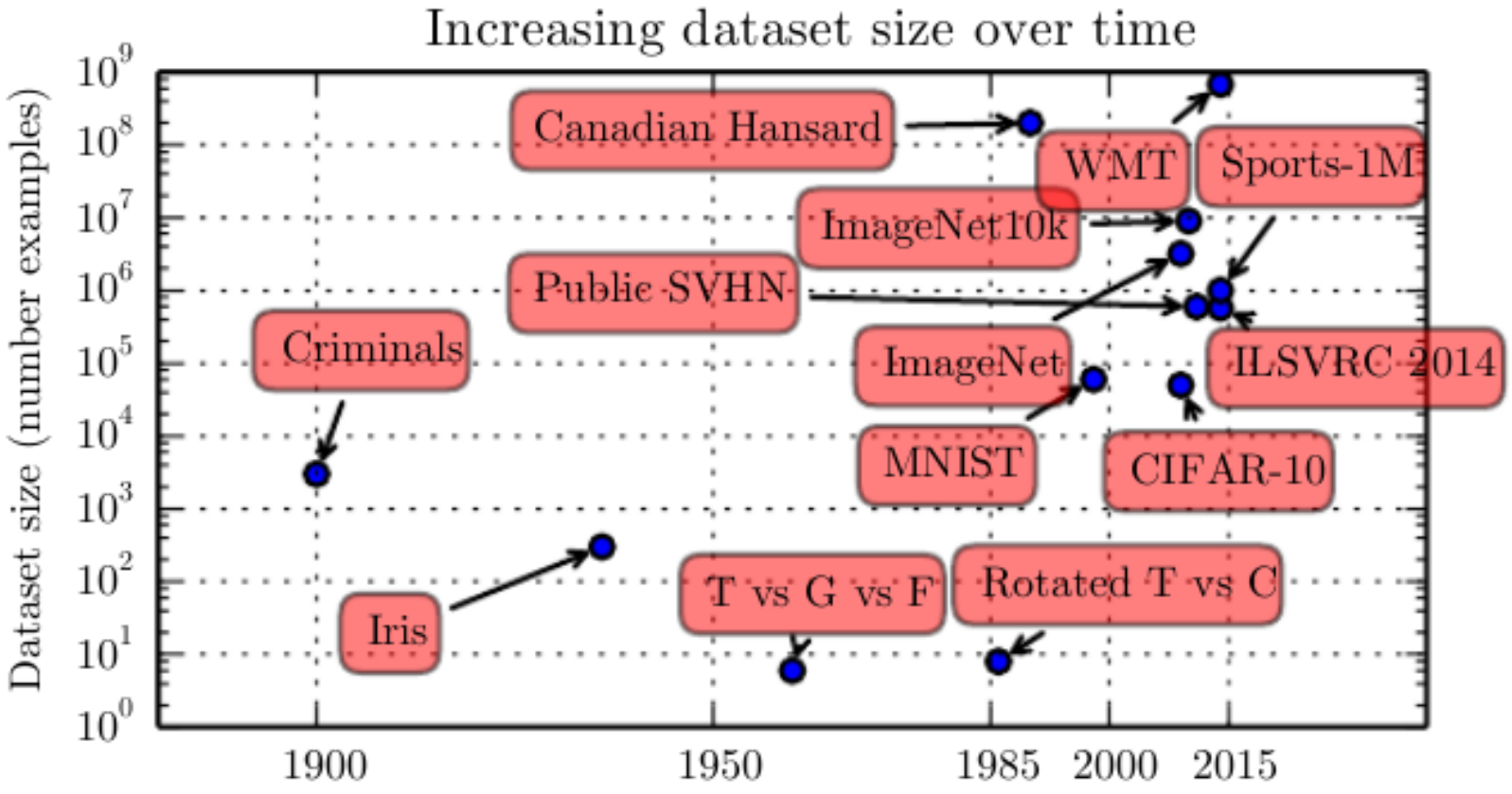


Figure 4 : Dataset sizes.

Classification Datasets

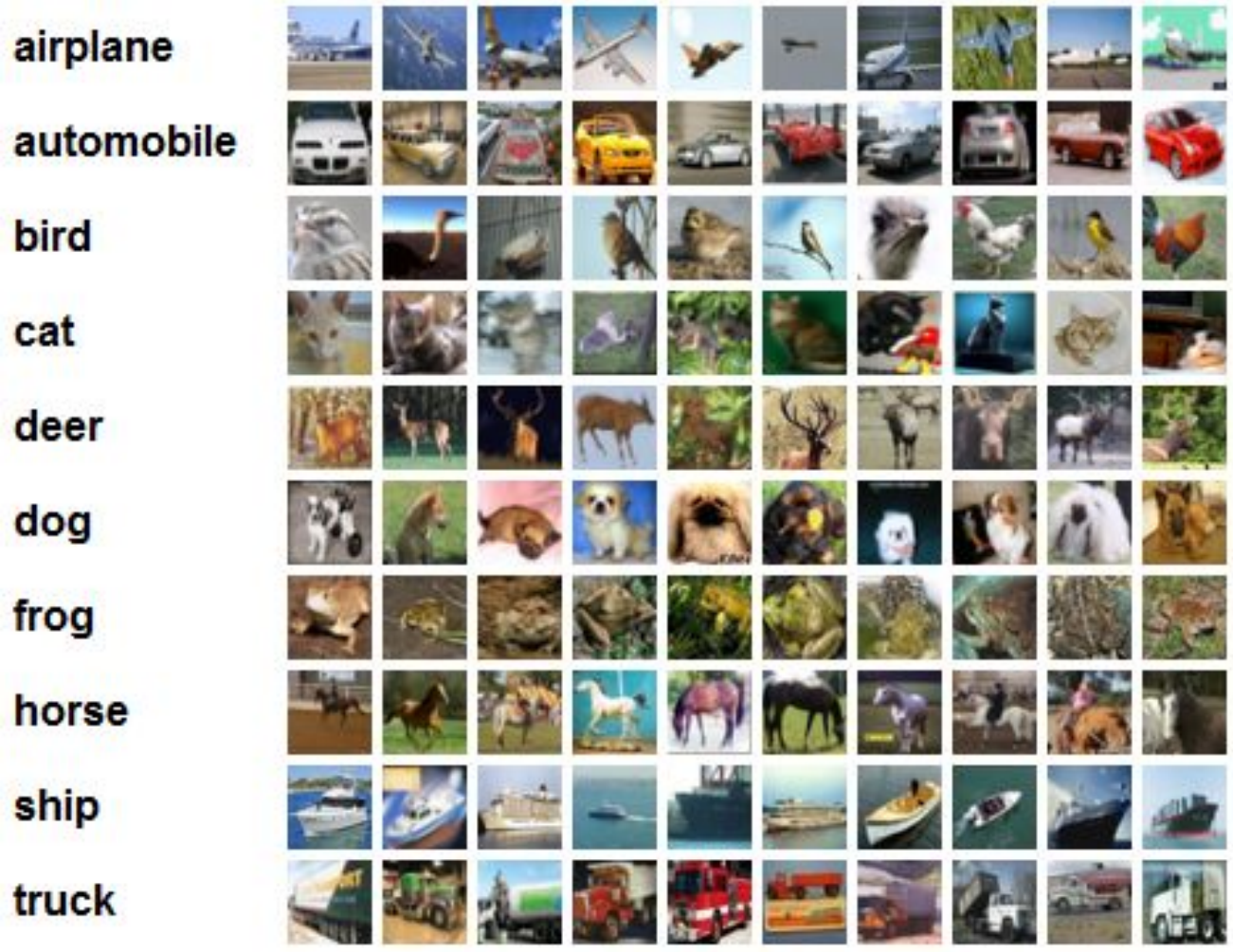


Figure 5 : CIFAR-10 Dataset.

Classification Datasets

Training set $\rightarrow$ to fit the parameters [i.e., weights]

Validation set $\rightarrow$ to tune the parameters [i.e., architecture]

Test set $\rightarrow$ to assess the performance [i.e., generalization and predictive power]

batch size = the number of training examples in one forward/backward pass.

Outline

Three Waves of AI

- Cybernetics: biological learning and perceptron.
- Conectionism: backpropagation, one or two hidden layers.
- Deep Learning

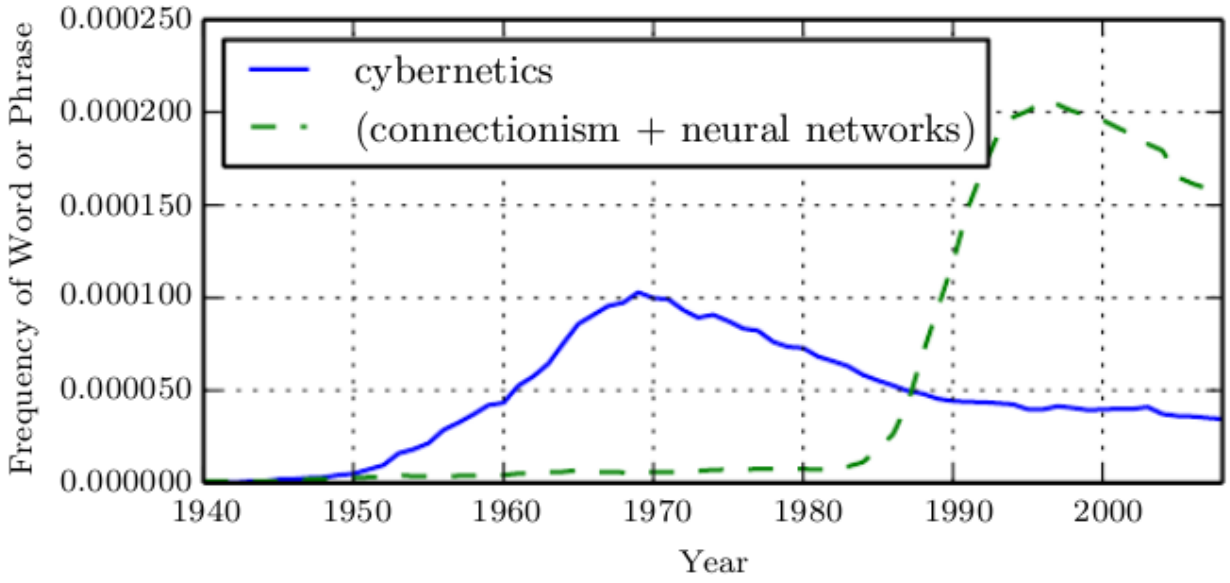


Figure 6 : Three historical waves of artificial neural nets.

Deep Learning

Deep Learning

"...one should not view deep learning as an attempt to simulate the brain. Modern deep learning draws inspiration from many fields, especially applied math fundamentals like linear algebra, probability, information theory, and numerical optimization."

Ian Goodfellow

research scientist at Google Brain

The Perceptron

- Mathematical model of a biological neuron.
- Creates a single output based on sum of many weighted inputs.
- Uses an activation function to create (slightly) non-linear behavior.

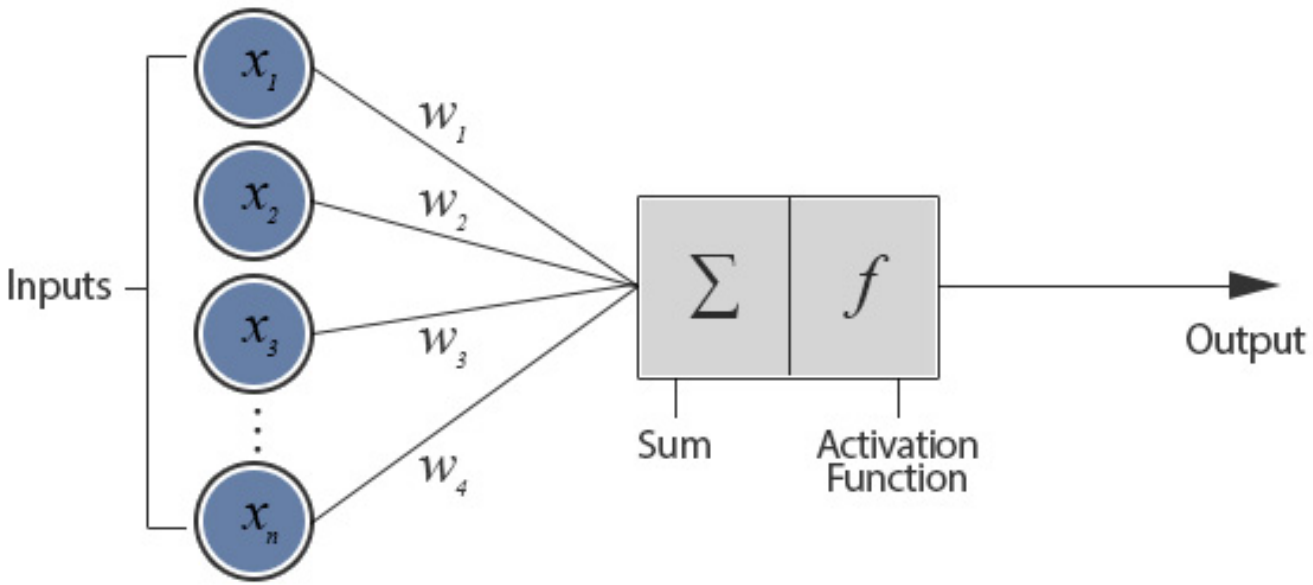


Figure 7 : The perceptron.

The Perceptron: Activation Function

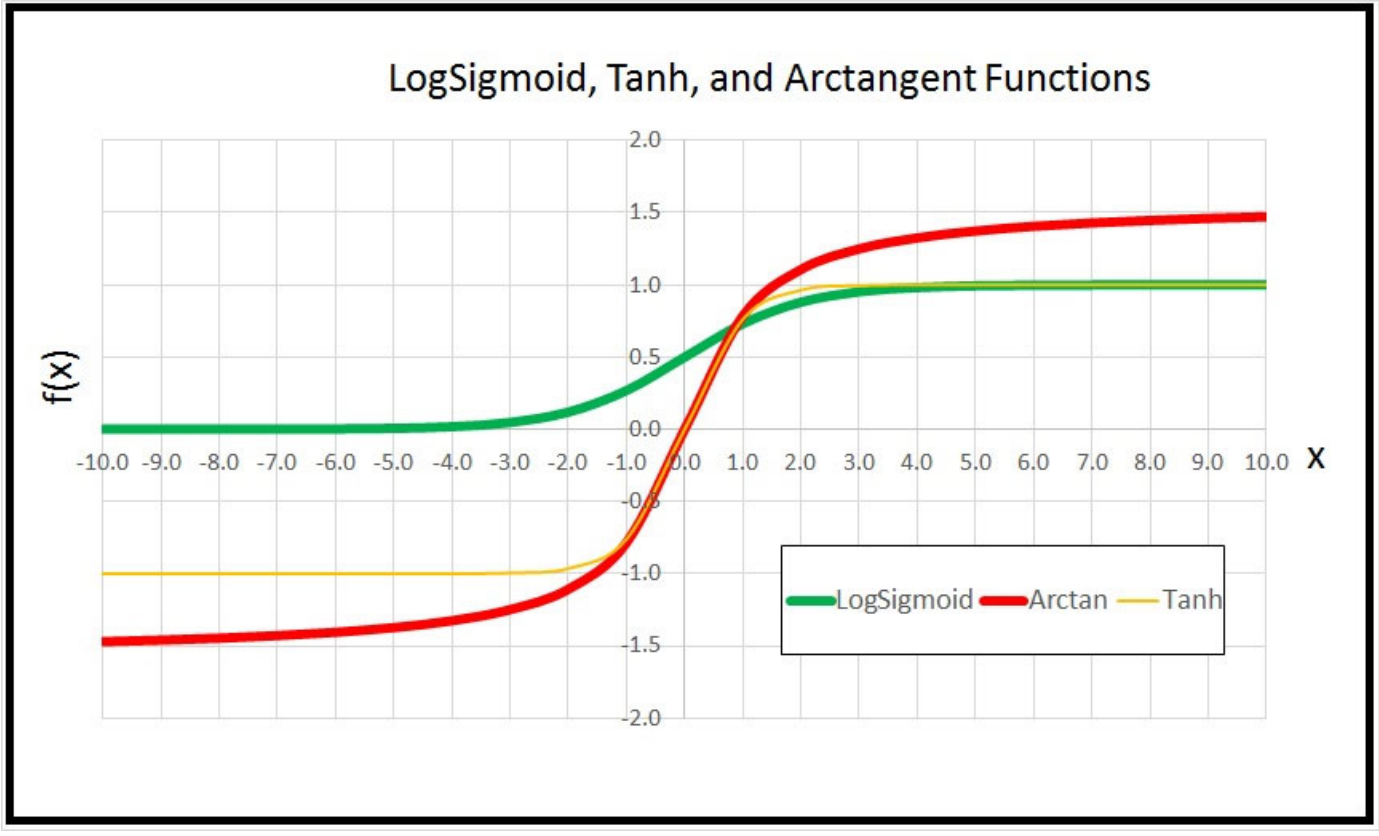


Figure 8 : Different activation functions.

Cost Function

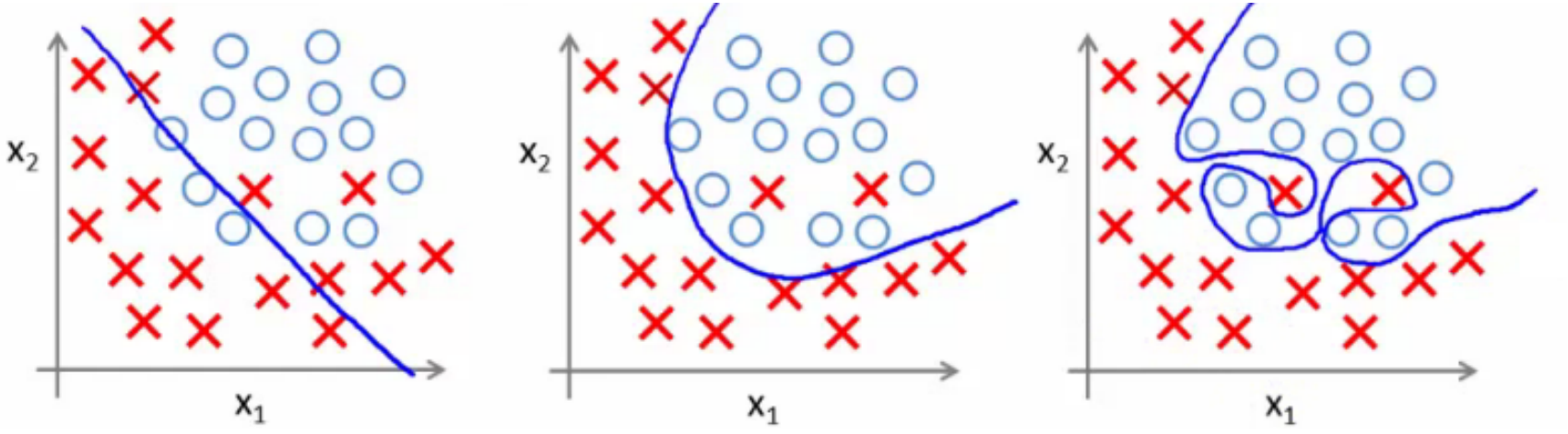
- An evaluation function that measures how well the machine learning algorithm maps the target function that it's striving to guess.
- Tells us that the rate at which the weight or bias learns is controlled by the error in the output.
- Most common used cost function is Cross Entropy.

$$C_0 = -\frac{1}{n}\sum_x [y \ln a + (1-y)\ln(1-a)]$$

where:

$$a = \sigma(x)$$

Regularization



$$L1 \rightarrow C = C_0 + \frac{\lambda}{n} \sum_{i=1}^{k} |w_i| \qquad L2 \rightarrow C = C_0 + \frac{\lambda}{n} \sum_{i=1}^{k} |w_i|^2$$

Gradient Descent

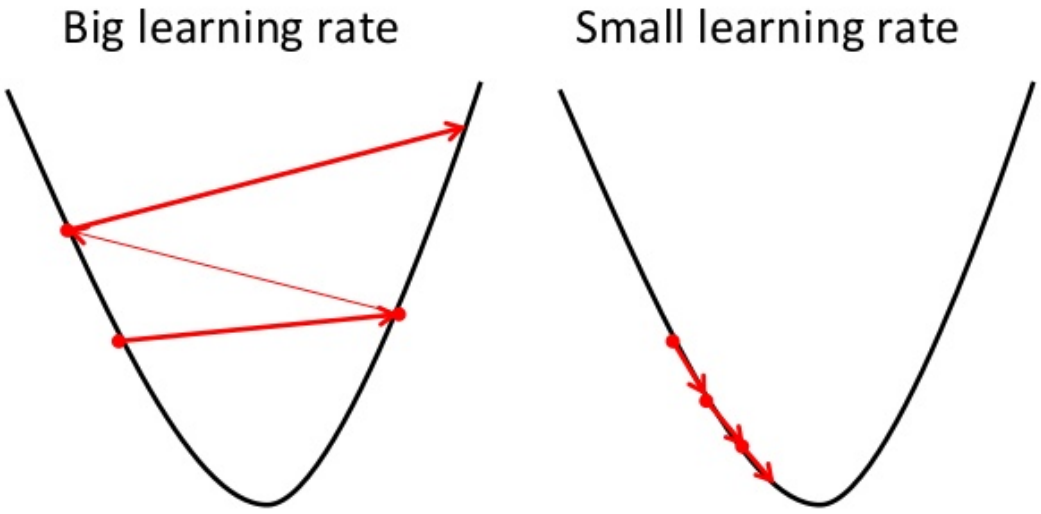


Figure 9 : Gradient descent and learning rates.

$$W_j = W_j - \mu \nabla C$$

Neural Networks

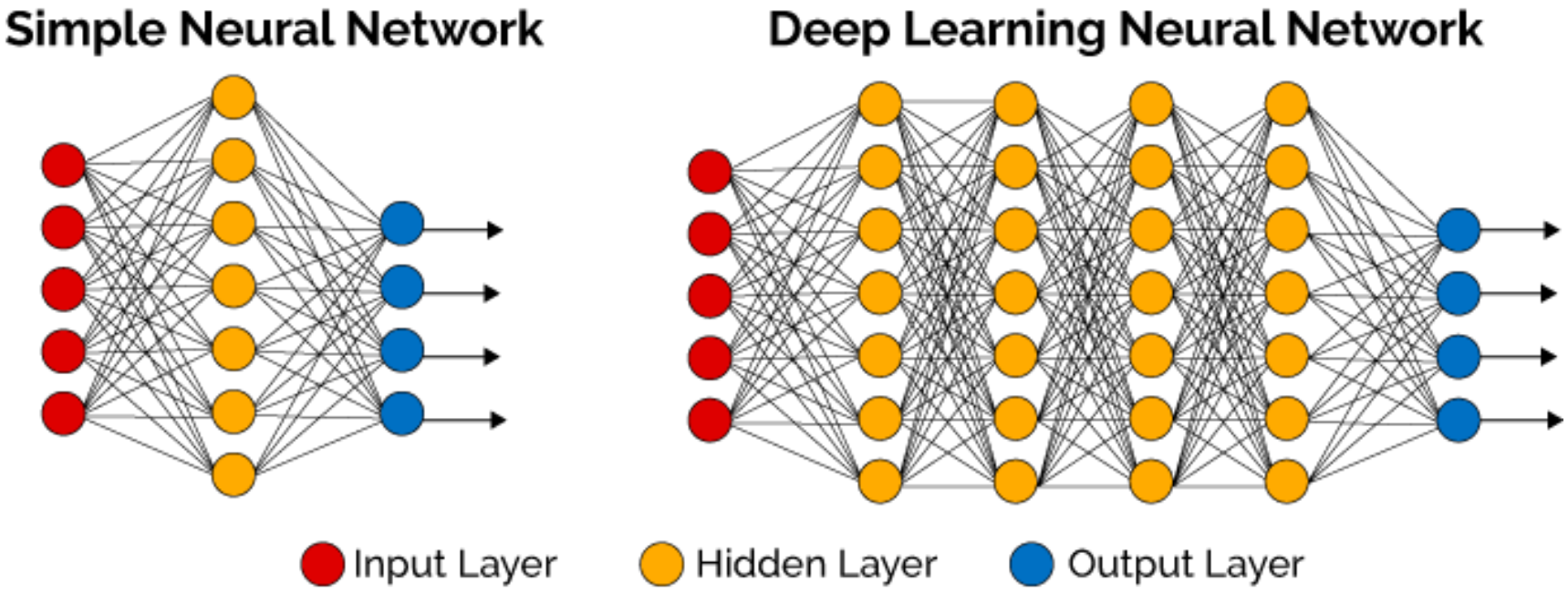


Figure 10 : Neural Networks and Deep Learning Neural Network.

Neural Networks Example 1

Example 1

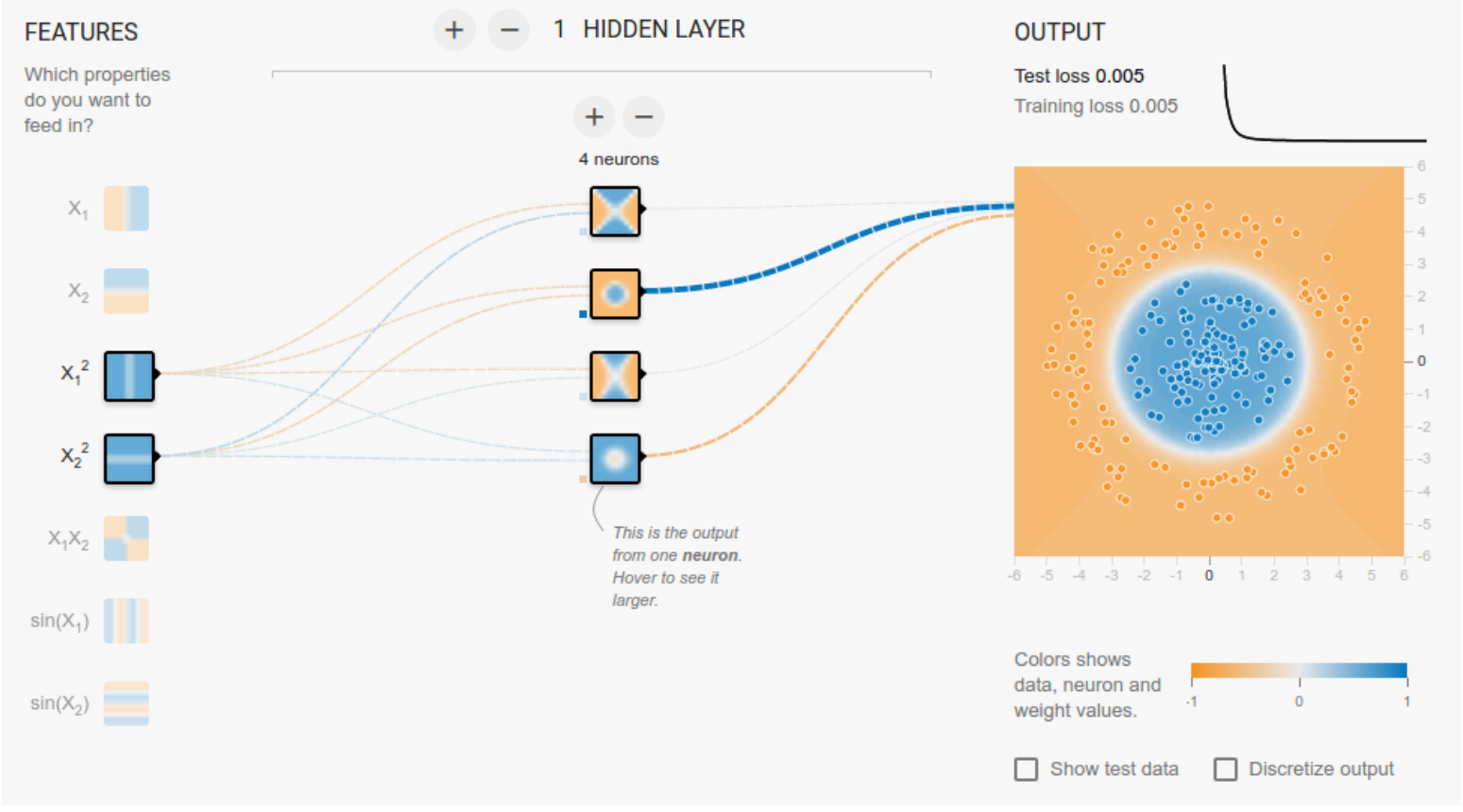


Figure 11 : Neural network data classifier.

Neural Networks Example 2

Example 2

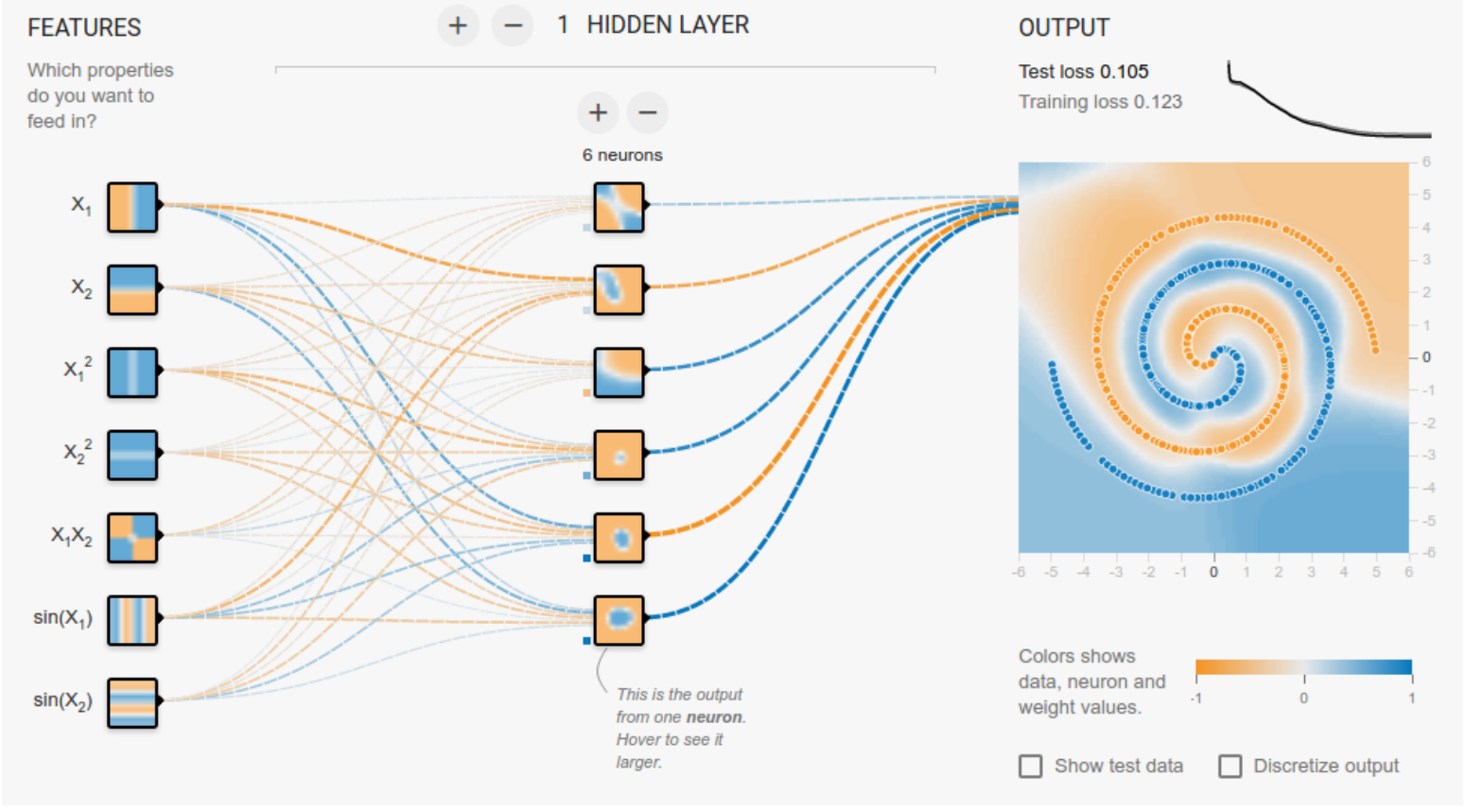
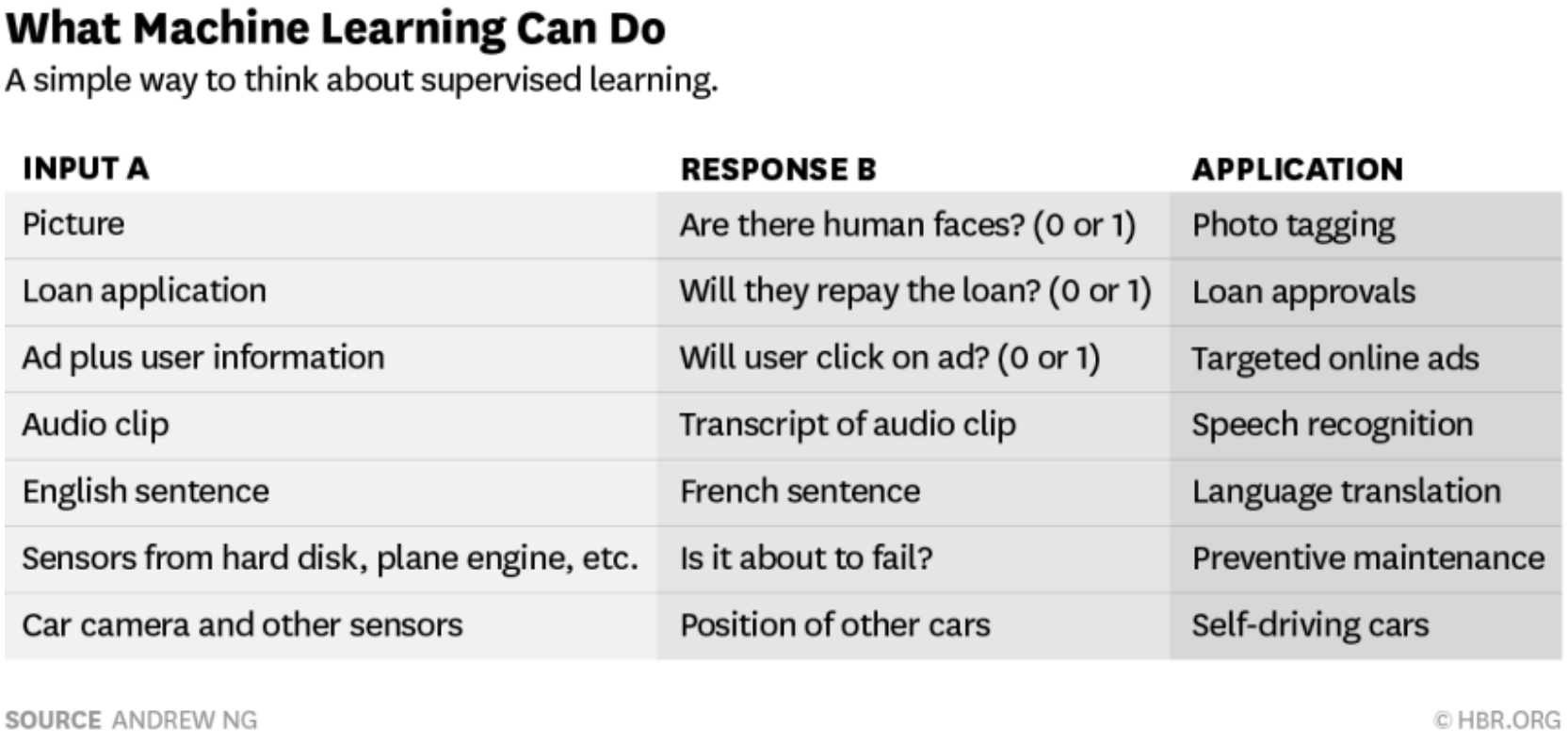


Figure 12 : Neural network data classifier.

Outline

When to apply AI?



What Machine Learning Can Do

A simple way to think about supervised learning.

INPUT A	RESPONSE B	APPLICATION
Picture	Are there human faces? (0 or 1)	Photo tagging
Loan application	Will they repay the loan? (0 or 1)	Loan approvals
Ad plus user information	Will user click on ad? (0 or 1)	Targeted online ads
Audio clip	Transcript of audio clip	Speech recognition
English sentence	French sentence	Language translation
Sensors from hard disk, plane engine, etc.	Is it about to fail?	Preventive maintenance
Car camera and other sensors	Position of other cars	Self-driving cars

SOURCE ANDREW NG

© HBR.ORG

Figure 13 : Applications for AI depends on the model: *INPUT A* $\rightarrow$ *RESPONSE B*.

When to apply AI?

When to apply AI?

"If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future."

Andrew Ng

former VP & Chief Scientist of Baidu, Co-Founder of Coursera

Top Companies Investing in AI, by Industry

SOFTWARE AND IT	FINANCIAL SERVICES	MANUFACTURING	AUTOMOTIVE	INDUSTRIAL AUTOMATION
Google	Bloomberg	Abb	Tesla	Bosch
Microsoft	American Express	National Instruments	Ford	Siemens
Amazon	TD Ameritrade	Toshiba	GM	Rockwell Automation
Facebook	Deutsche Bank	GE	Toyota	Honeywell
Mitre	Goldman Sachs			
IBM	BNP Paribas			
Deloitte				

HEALTHCARE	TELECOMMUNICATIONS	RETAIL	SEMICONDUCTORS	INTERNET
GE Healthcare	Huawei	Walmart	Intel	Google
Nuance	Nokia	Gamestop	Texas Instruments	Facebook
Mayo Clinic	BT Group	Target	Microchip Technology	LinkedIn
Siemens Healthcare	Orange S.A.	Rakuten	Altera	Amazon
McKesson	Nippon	Best Buy	Imagination Technologies	
Massachusetts General Hospital	Verizon	Barnes & Noble	ARM	
			Mellanox	
			Qualcomm	

Figure 14 : Top companies investing in AI, by industry.

Companies Investing in AI, by Categories

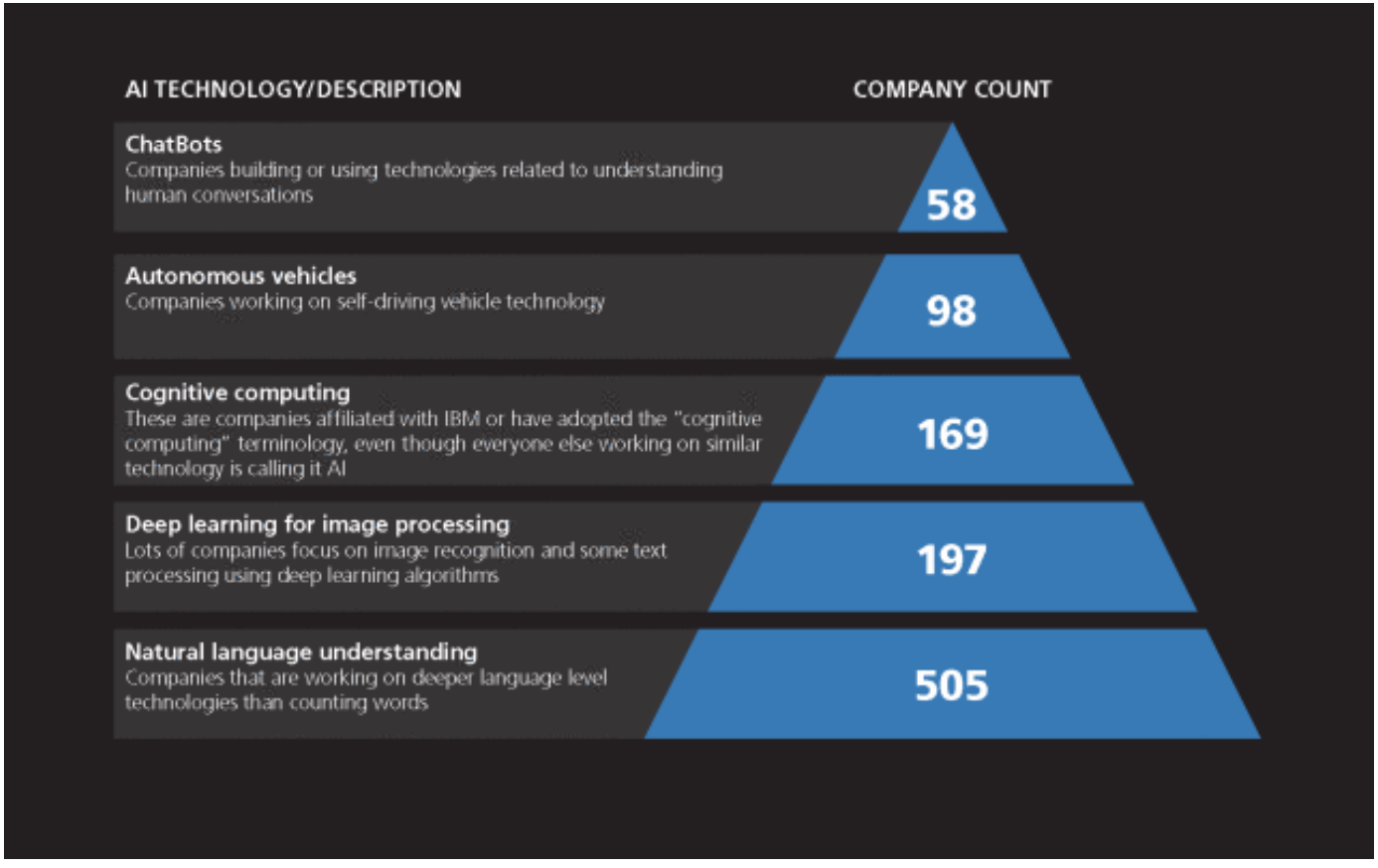


Figure 15 : Subcategories of AI technologies and companies investing.

Convolutional Neural Networks, CNN

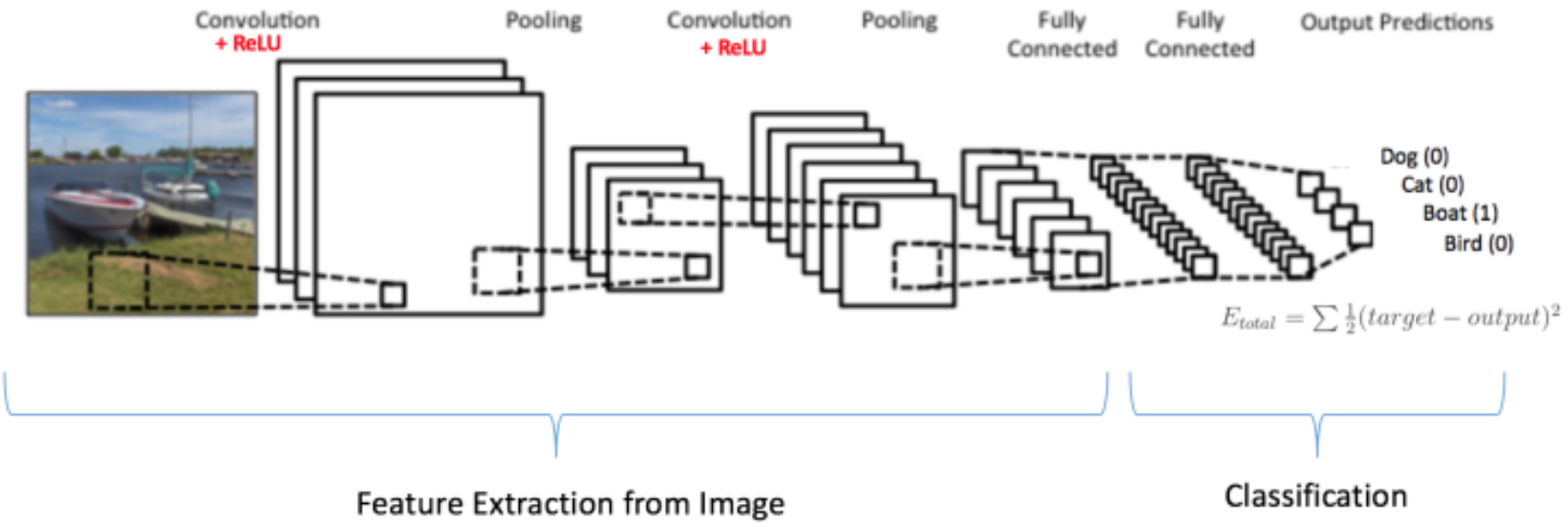


Figure 16 : Convolutional Neural Network, LeNet Architecture

Convolutional Neural Networks, CNN

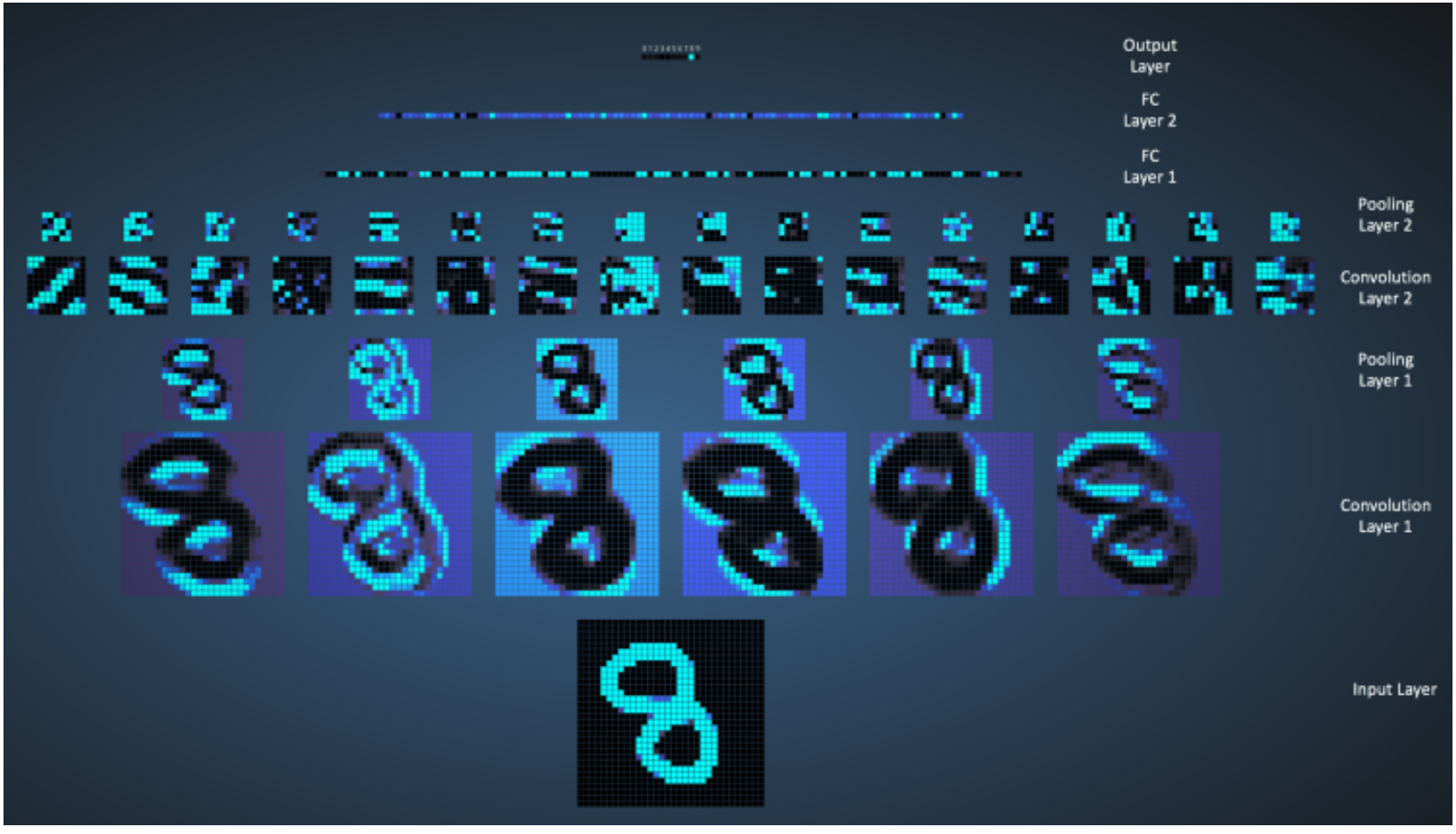


Figure 17 : Convolutional Neural Network, MNIST

Recurrent Neural Networks, RNN

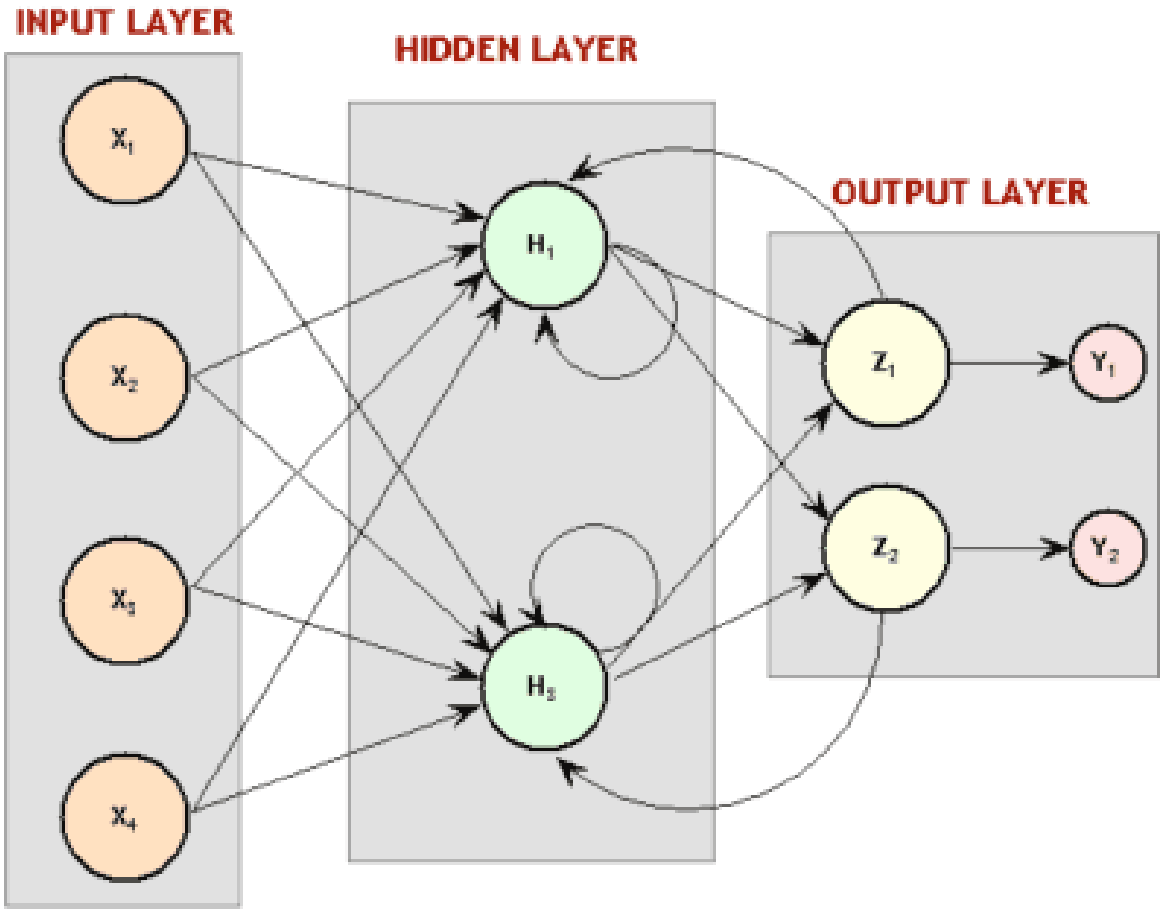


Figure 18 : Recurrent Neural Network, RNN

Recurrent Neural Networks, RNN

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.
```

Figure 19 : A 3-layer RNN with 512 hidden nodes on each layer can learn English by using Shakespeare works.

Outline

Different AI Frameworks

	Caffe	TORCH	theano	TensorFlow
Language	C++, Python	Lua	Python	Python
Pretrained	Yes	Yes	Yes (Lasagne)	Inception
Parallel GPUs: Data	Yes	Yes	Yes	Yes
Parallel GPUs: Model	No	Yes	Experimental	Yes (best)
Readable Source Code	Yes (C++)	Yes	No	No
Good at RNN	No	Mediocre	Yes	Yes (best)
Higher-Level APIs	No	No	Keras	Keras, TFLearn

Figure 20 : Comparison between different AI frameworks

Different AI Hardware Platforms

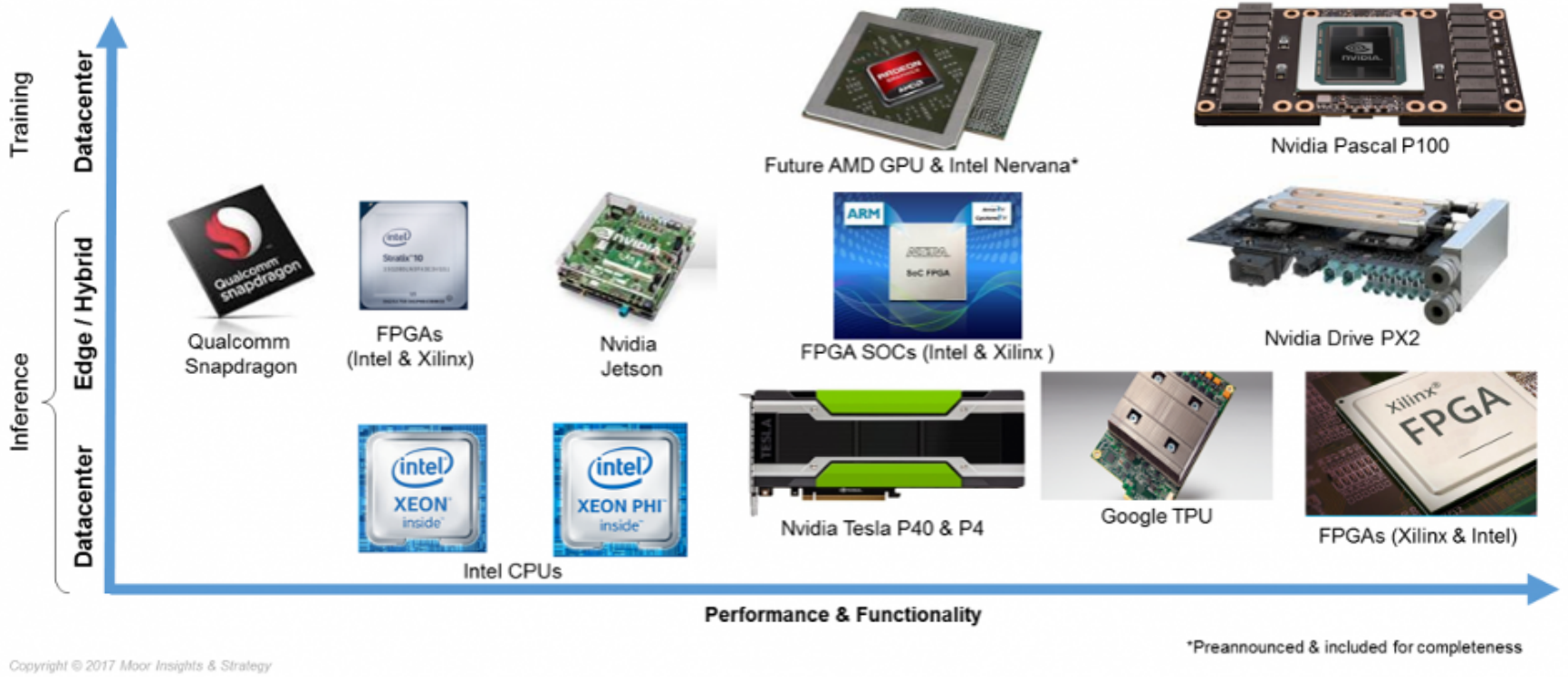


Figure 21 : Comparison between different AI hardware platforms

Training vs. Inferencing

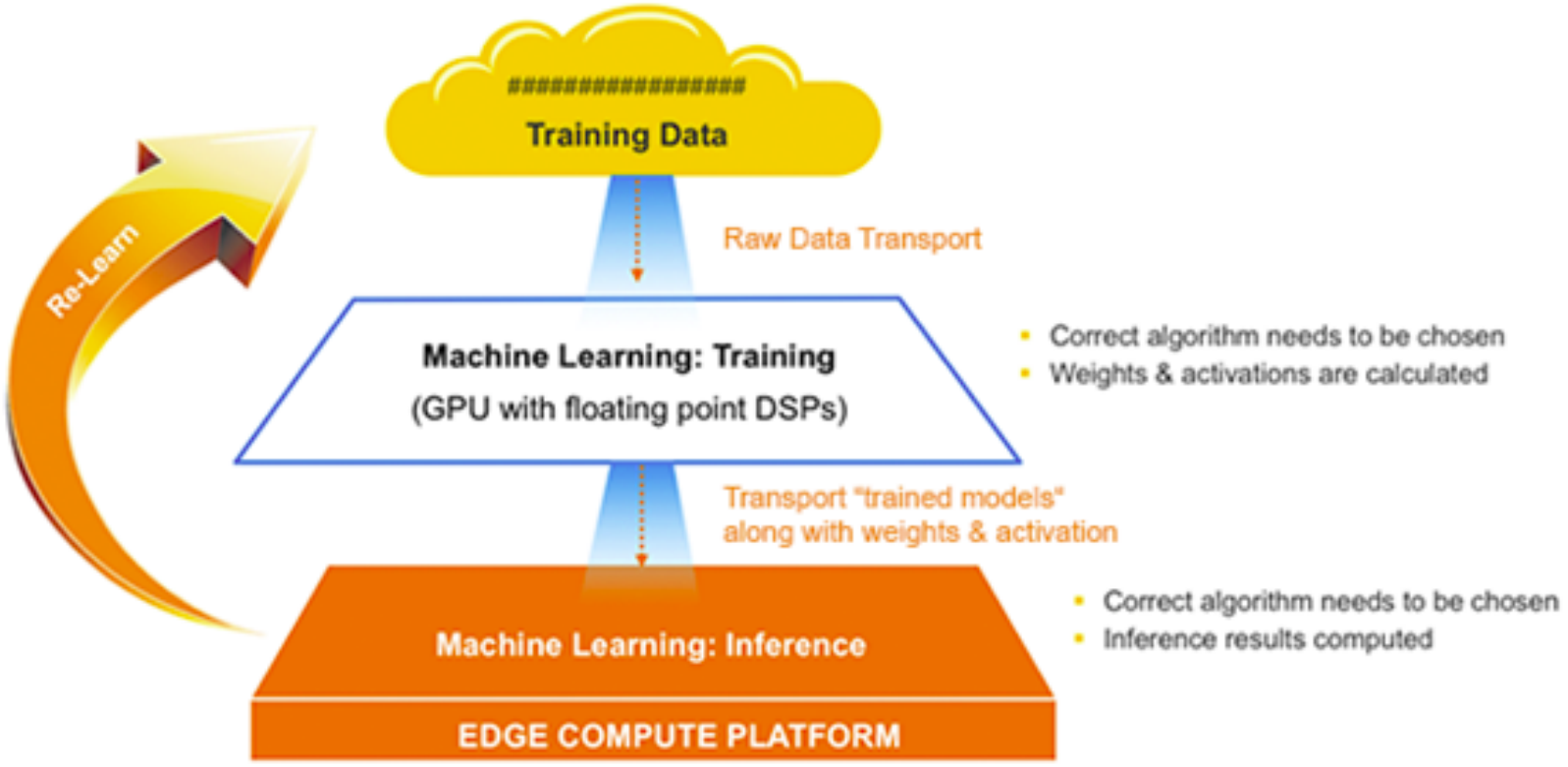
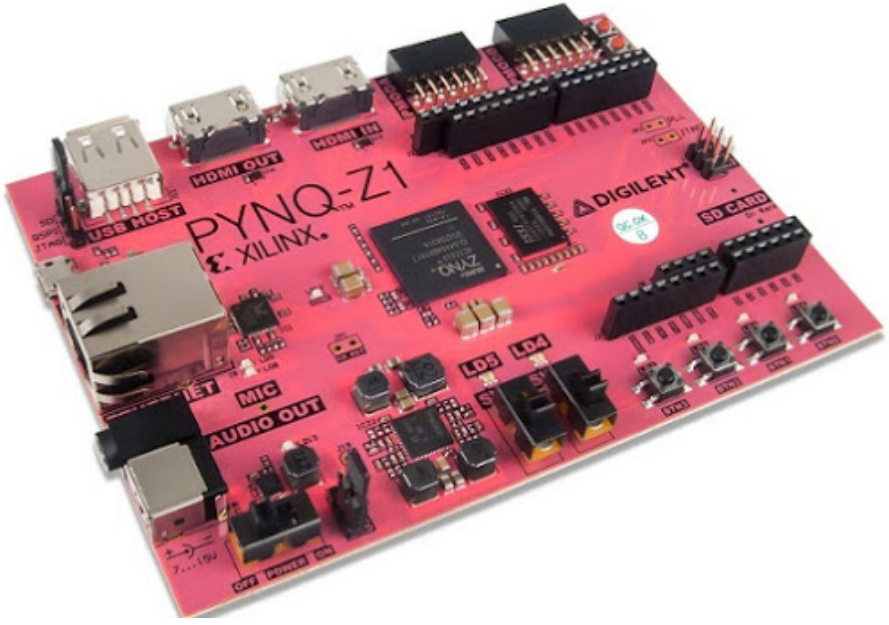


Figure 22 : Training in the cloud/datacenter & Inferencing at the edge.

FPGA Hardware Platform

The PYNQ-Z1 Board

XILINX ALL PROGRAMMABLE™ ZYNQ™



Processor: Dual-Core ARM® Cortex®-A9
FPGA: 1.3 M reconfigurable gates
Memory: 512MB DDR3 / FLASH
Storage: Micro SD card slot
Video: HDMI In and HDMI Out
Audio: Mic in, Line Out
Network: 10/100/1000 Ethernet
Expansion: USB Host connected to ARM PS
Interfaces: 1x Arduino Header, 2x Pmod (49 GPIO)
GPIO: 16 GPIO (65 in total with Arduino and Pmods)
Other I/O: 6x User LEDs, 4x Pushbuttons, 2x Switches
Dimensions: 3.44" x 4.81" (87mm x 122mm)

Figure 23 : The PYNQ-Z1 Board.

FPGA Hardware Platform

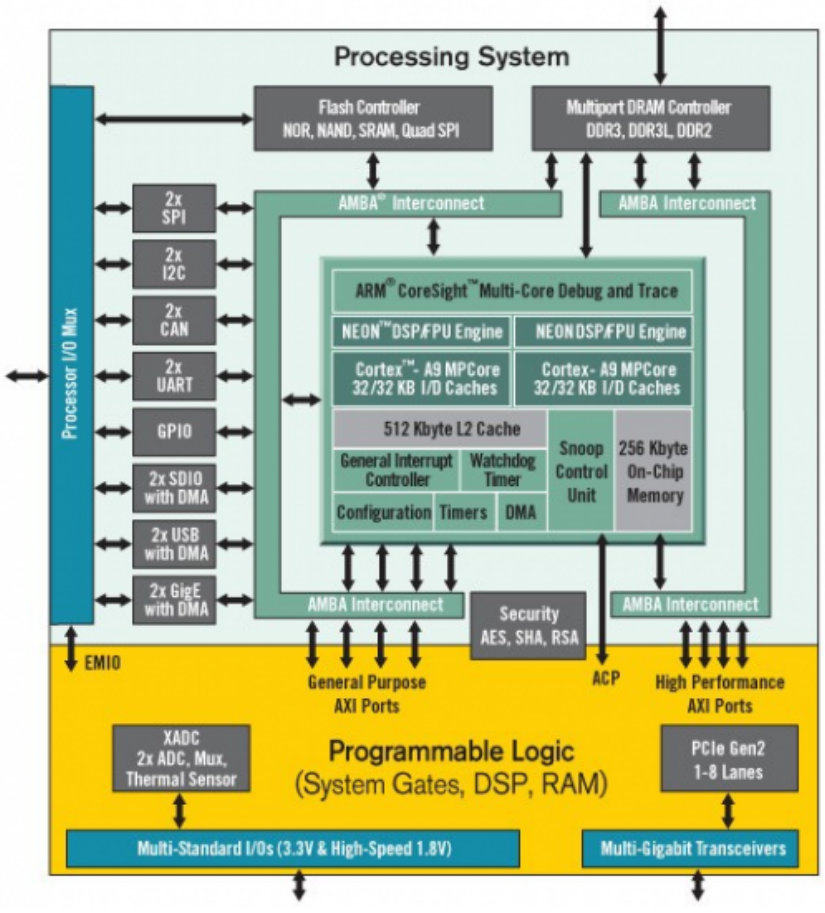


Figure 24 : The Zynq APSoC architecture.

Binarized Neural Networks

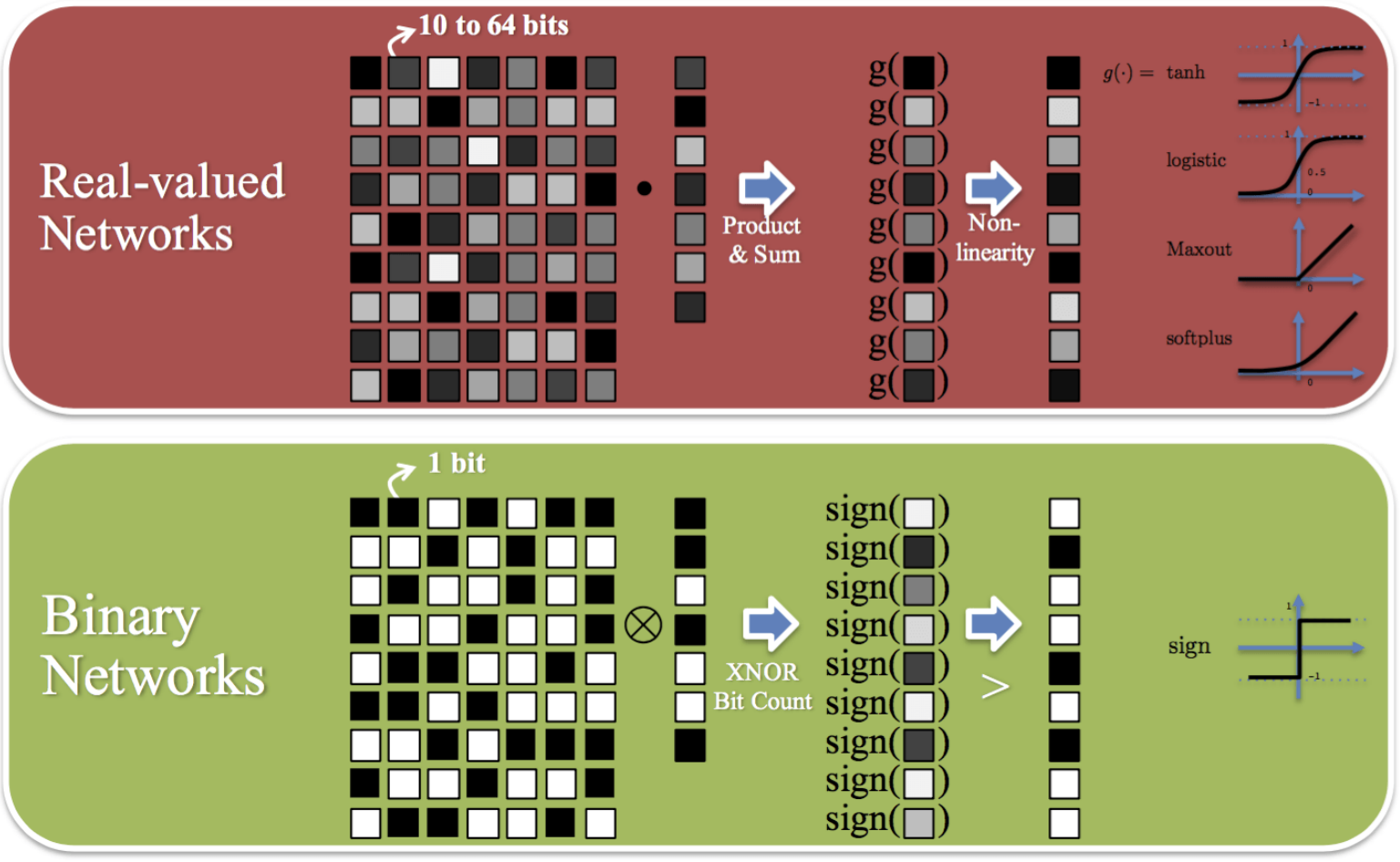


Figure 25 : Real-valued neural networks and binary neural networks .