Introduction

Chia-Chi Tsai (蔡家齊)
cctsai@gs.ncku.edu.tw
AI System Lab
Department of Electrical Engineering
National Cheng Kung University

Outline

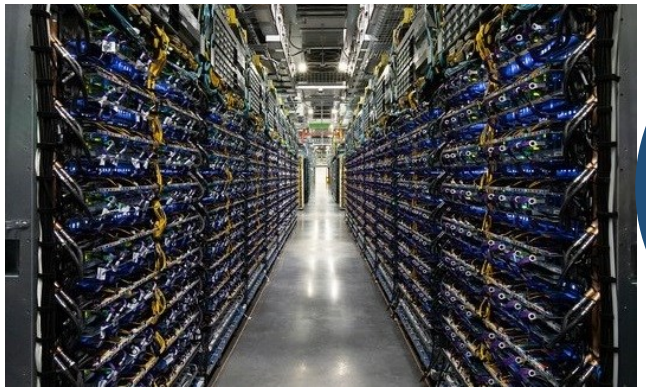
- Introduction
- AI, ML and DL

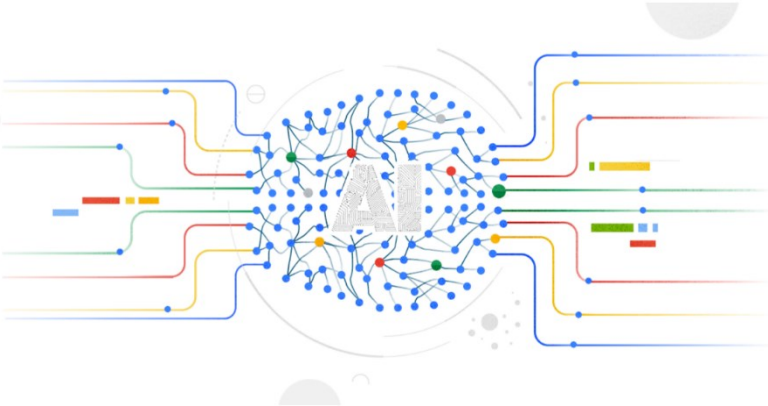
The Virtuous Circle of Deep Learning

Bigger Data

Larger Model

More Compute

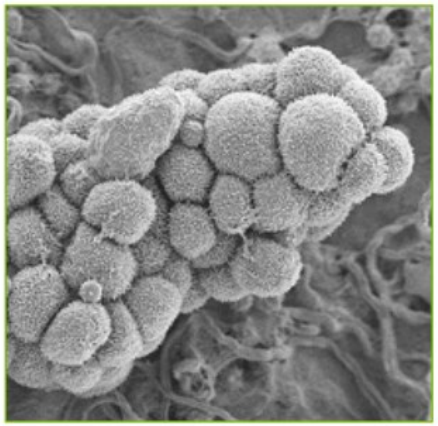




Application Domains Scaling

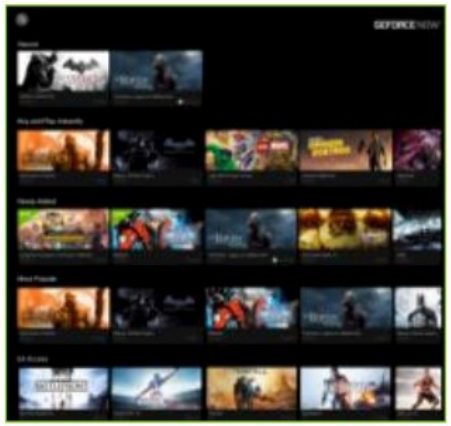
INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery



MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

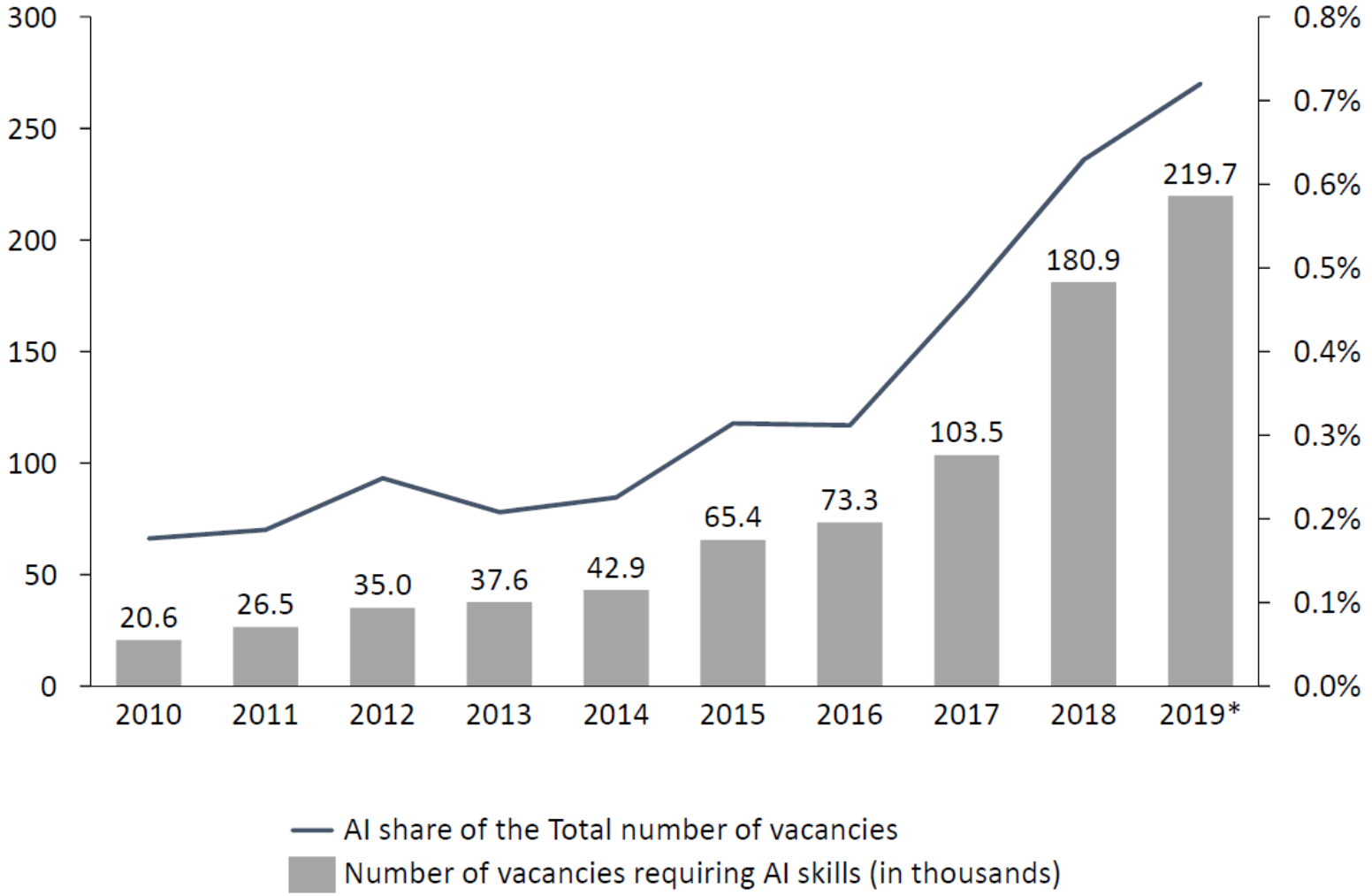
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

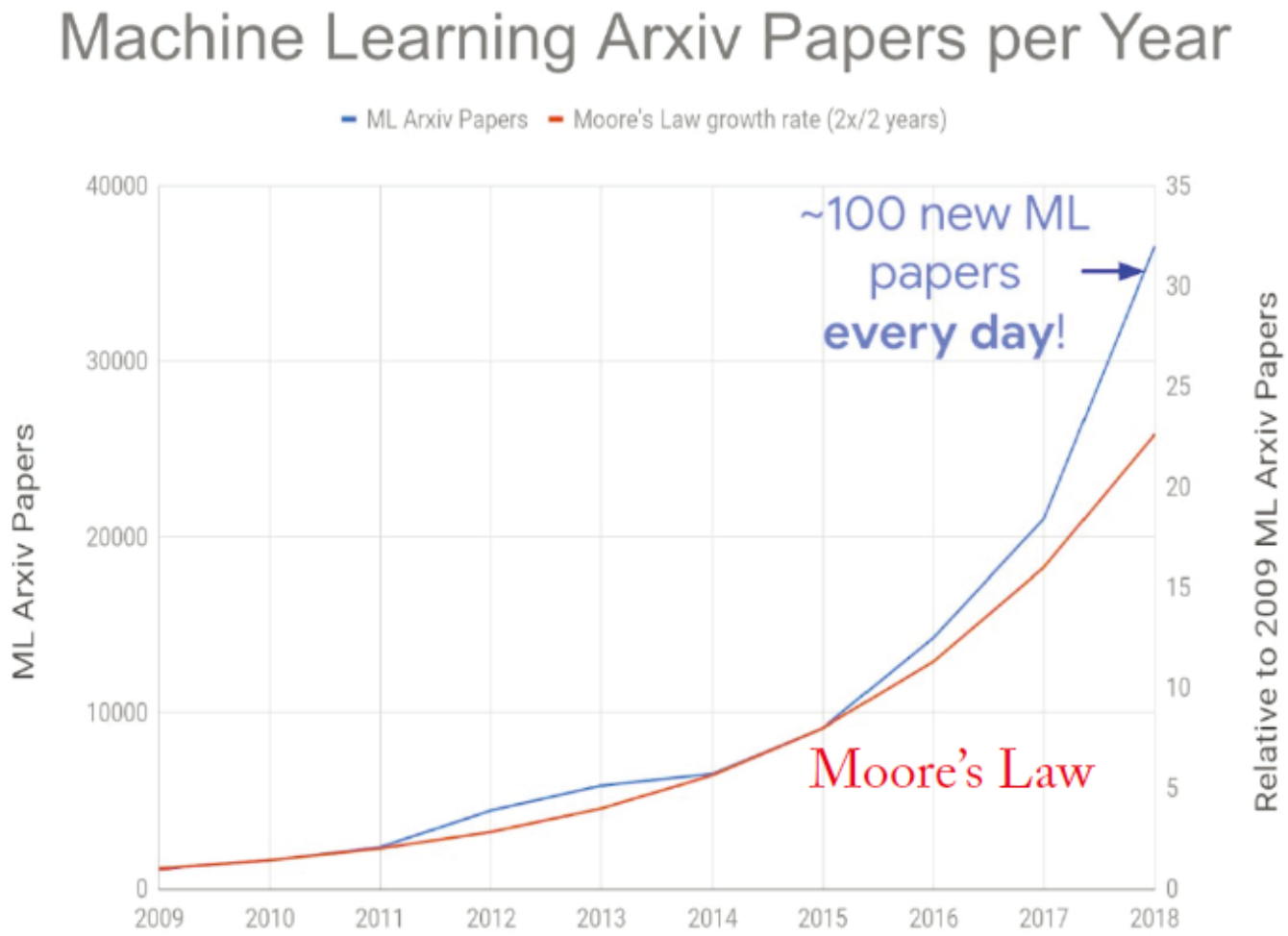
Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Deep Learning Jobs Scaling



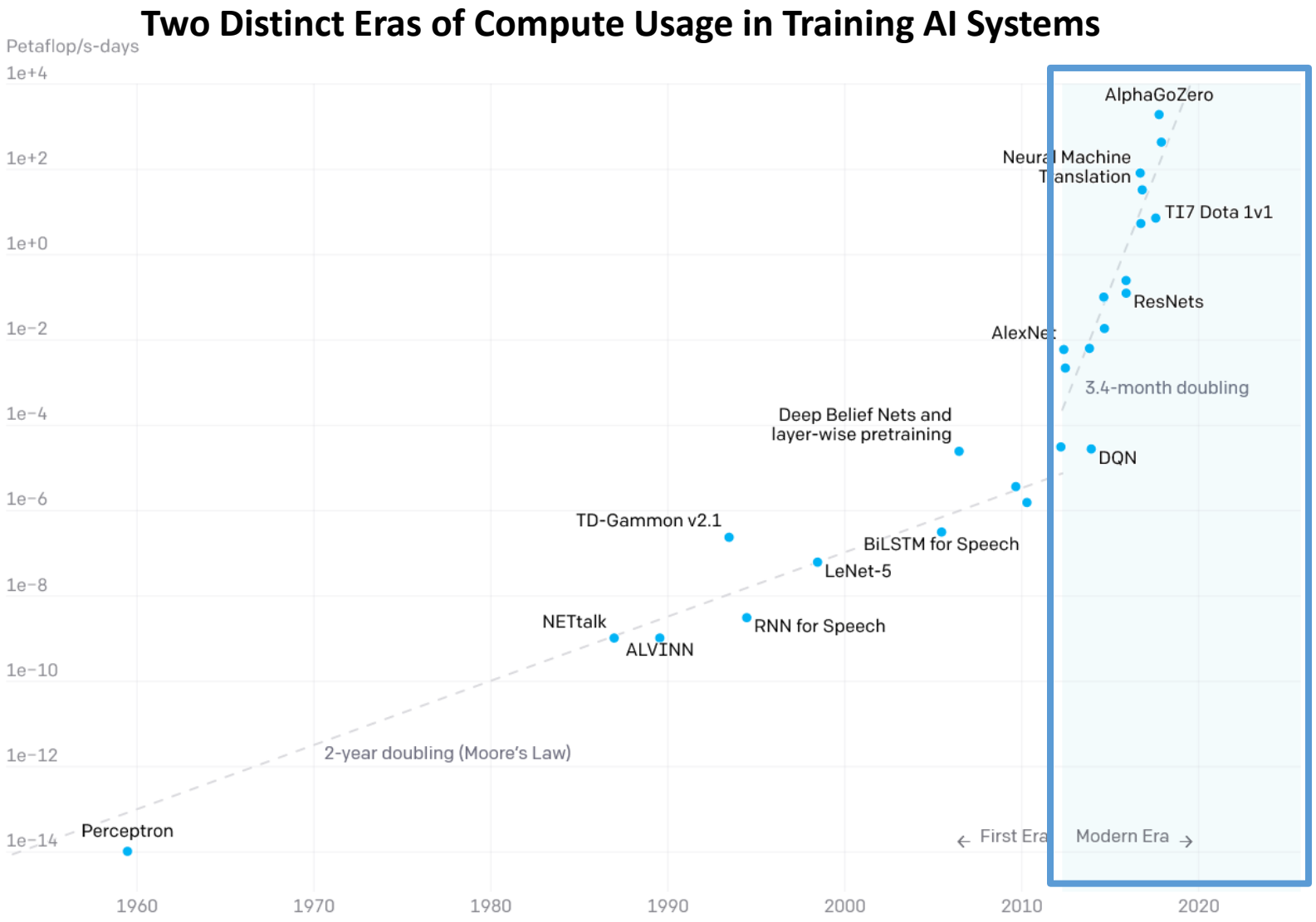
Alekseeva, L., Azar, J., Gine, M., Samila, S., & Taska, B. (2021). The demand for AI skills in the labor market. *Labour Economics*, *71*, 102002.

Deep Learning Papers Scaling



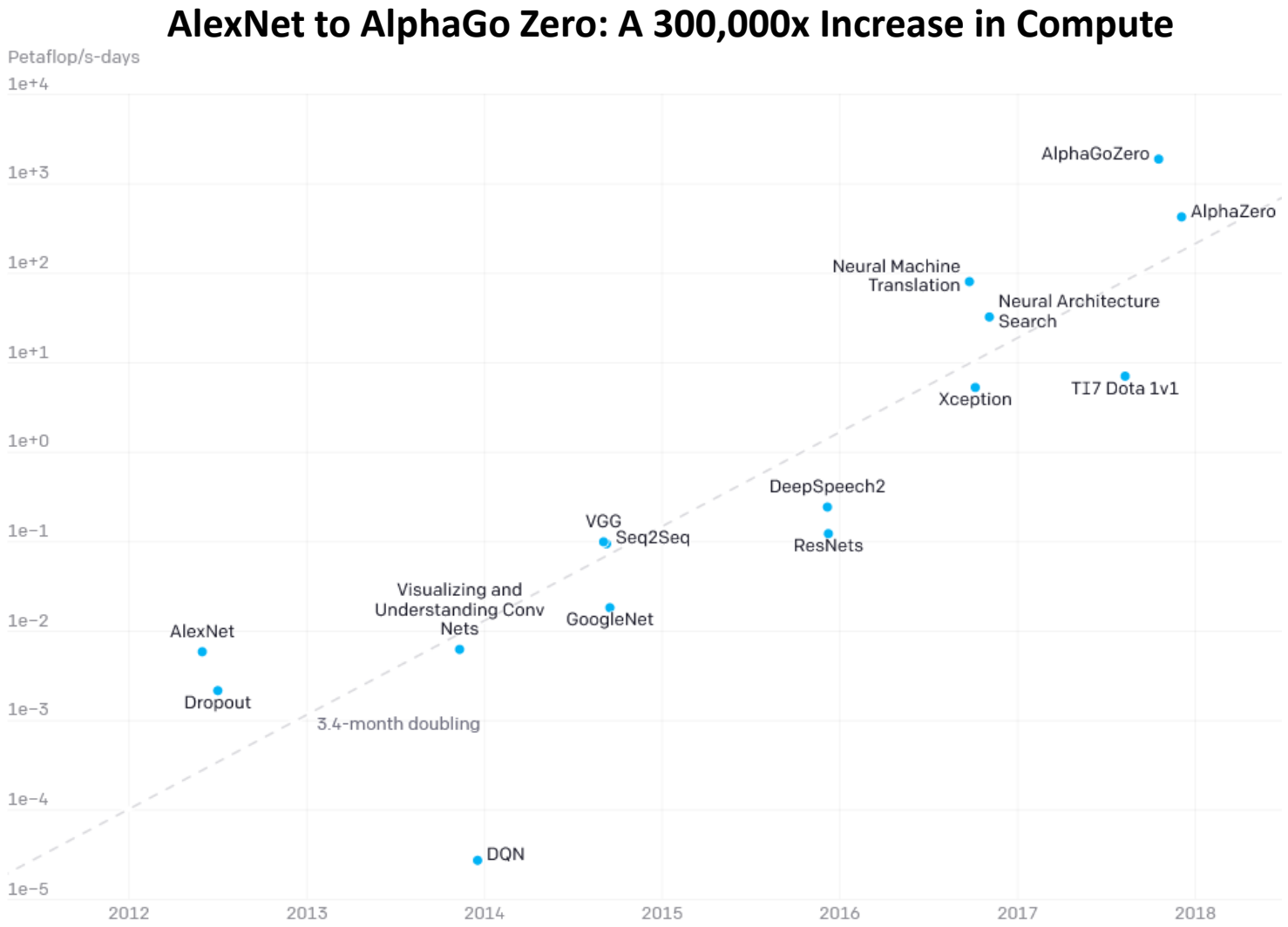
Dean, J., Patterson, D., & Young, C. (2018). A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, *38*(2), 21-29.

Deep Learning Models Scaling



From: OpenAI

Deep Learning Models Scaling



From: OpenAI

Deep Learning Hardware Scaling

Table 1: 90-epoch training time and single-crop validation accuracy of ResNet-50 for ImageNet reported by different teams.

Team	Hardware	Software	Minibatch size	Time	Accuracy
He *et al.* [5]	Tesla P100 × 8	Caffe	256	29 hr	75.3 %
Goyal *et al.* [4]	Tesla P100 × 256	Caffe2	8,192	1 hr	76.3 %
Codreanu *et al.* [3]	KNL 7250 × 720	Intel Caffe	11,520	62 min	75.0 %
You *et al.* [10]	Xeon 8160 × 1600	Intel Caffe	16,000	31 min	75.3 %
This work	Tesla P100 × 1024	Chainer	32,768	15 min	74.9 %

Akiba, T., Suzuki, S., & Fukuda, K. (2017). Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*.

Increasing Demand for Computing

Augmented Reality

Mobile Gaming

4K Video

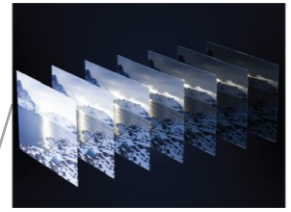
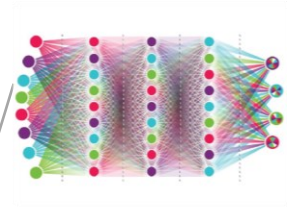


Image Processing

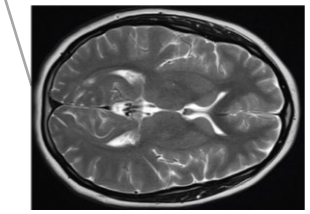
Object Tracking

Obstacle Avoidance



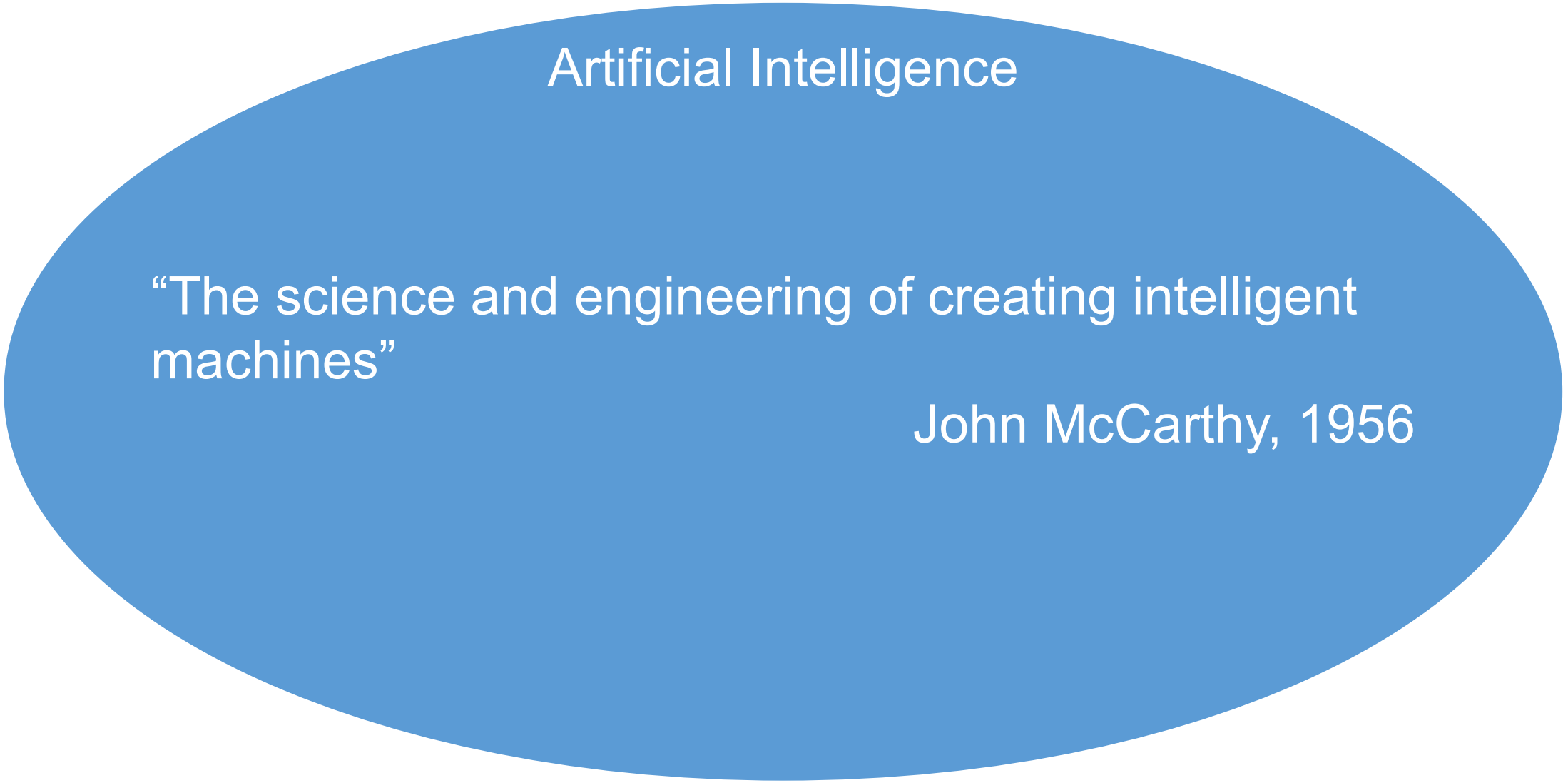
Machine Learning

Genomics

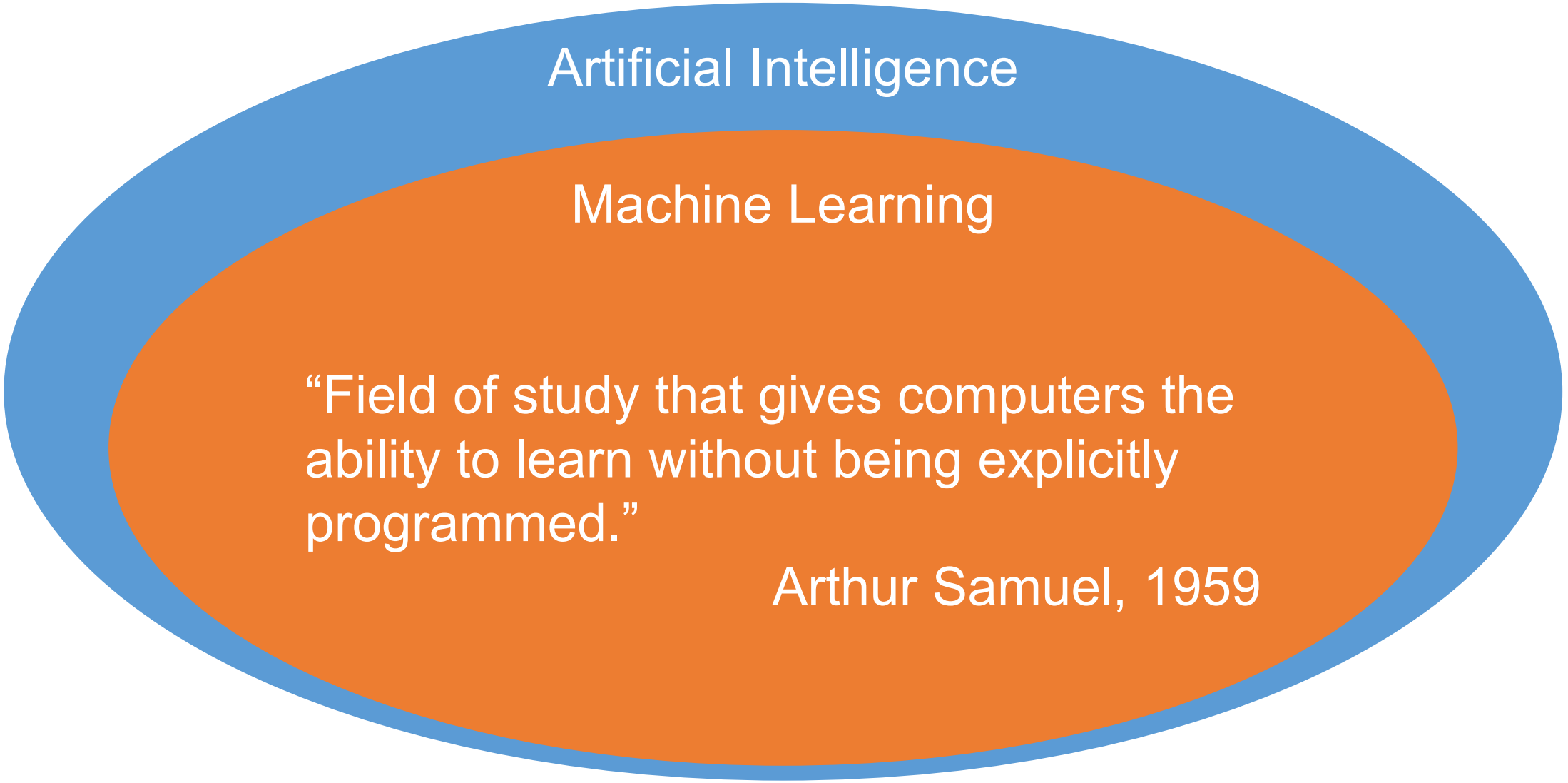


Medical Imaging

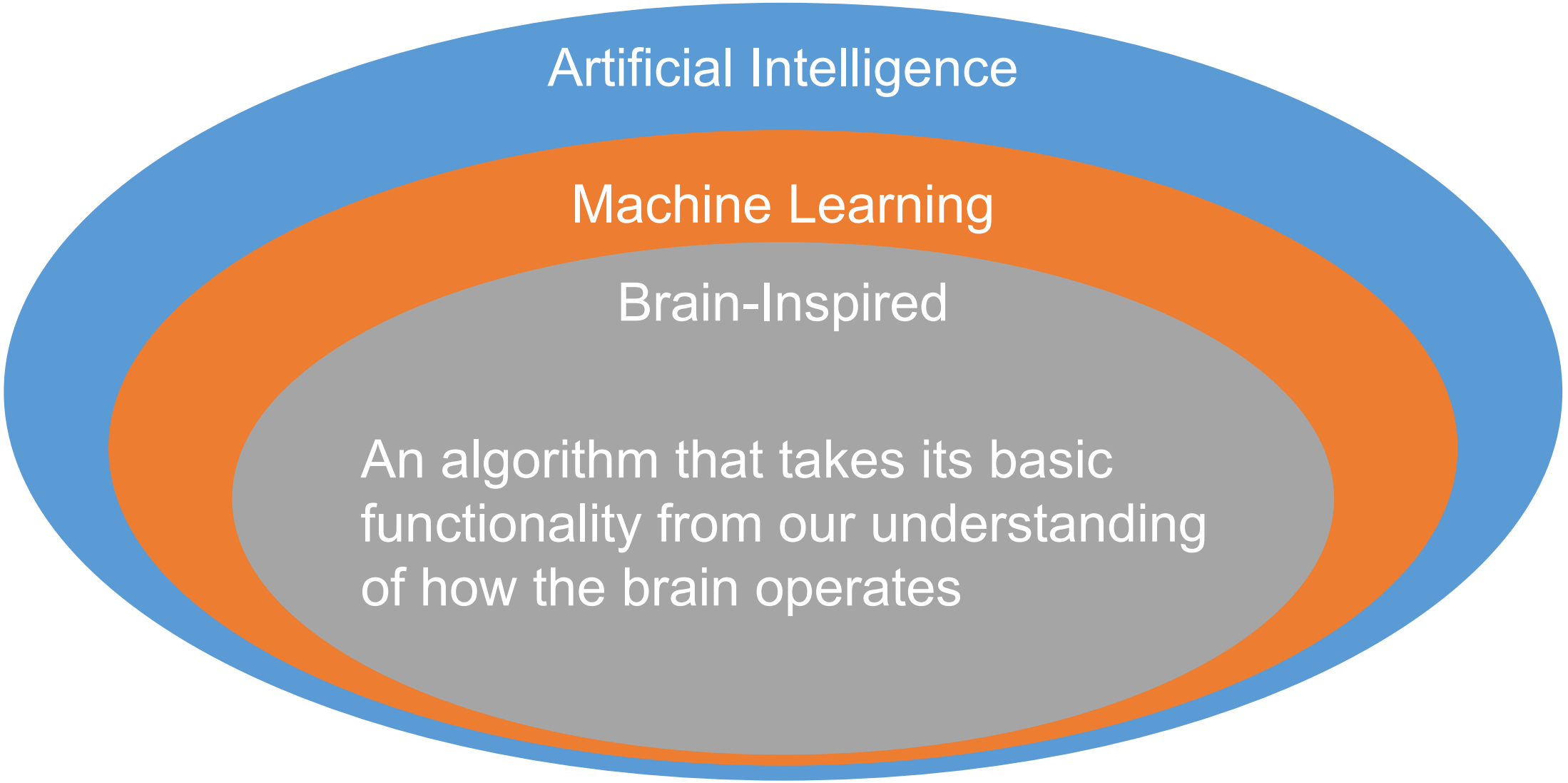
Artificial Intelligence (AI)



Machine Learning (ML)

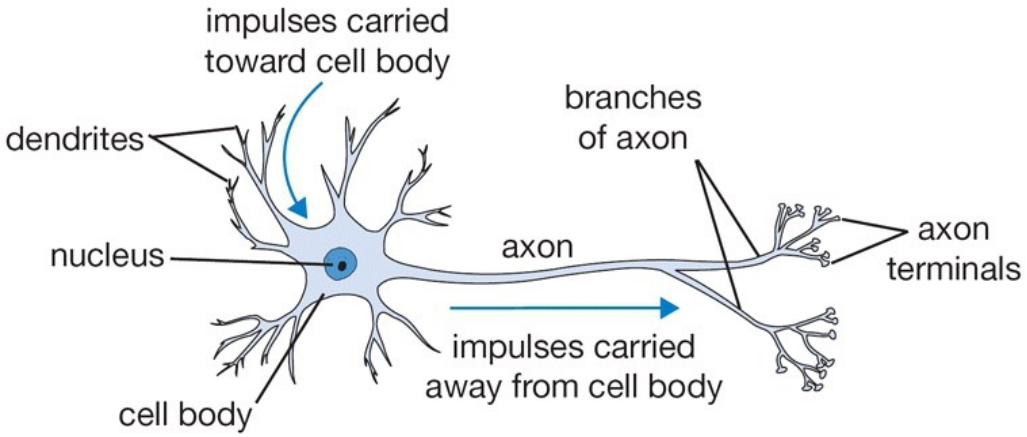


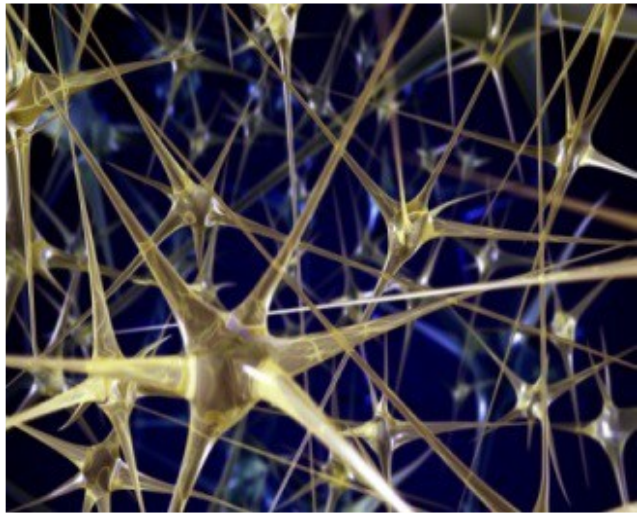
Deep Learning (DL)



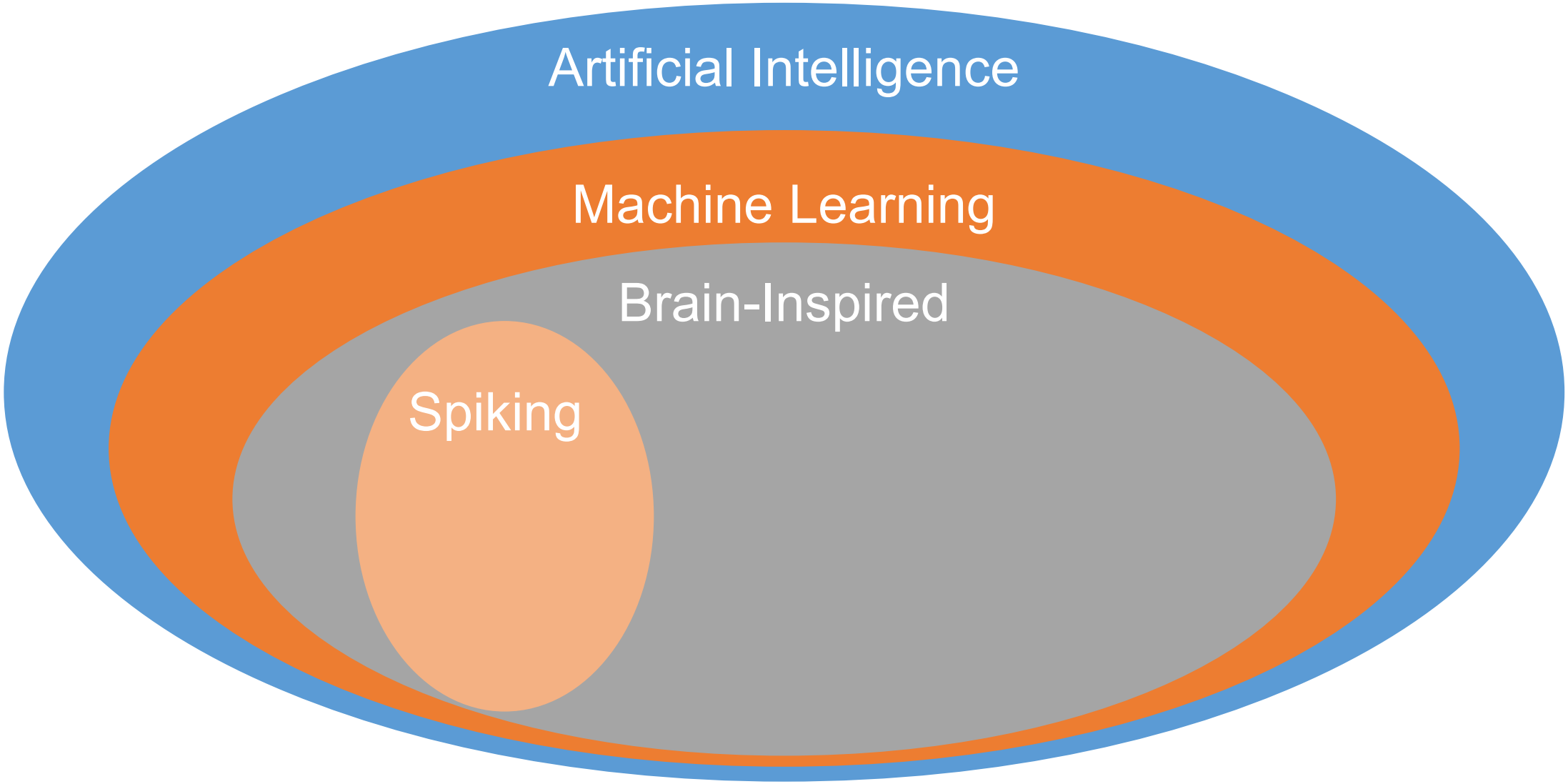
How Does the Brain Work?

- The basic computational unit of the brain is a **neuron**
  - 86B neurons in the brain
- Neurons are connected with nearly **10^{14} – 10^{15} synapses**
- Neurons receive input signal from **dendrites** and produce output signal along **axon**, which interact with the dendrites of other neurons via **synaptic weights**
- Synaptic weights – learnable & control influence strength



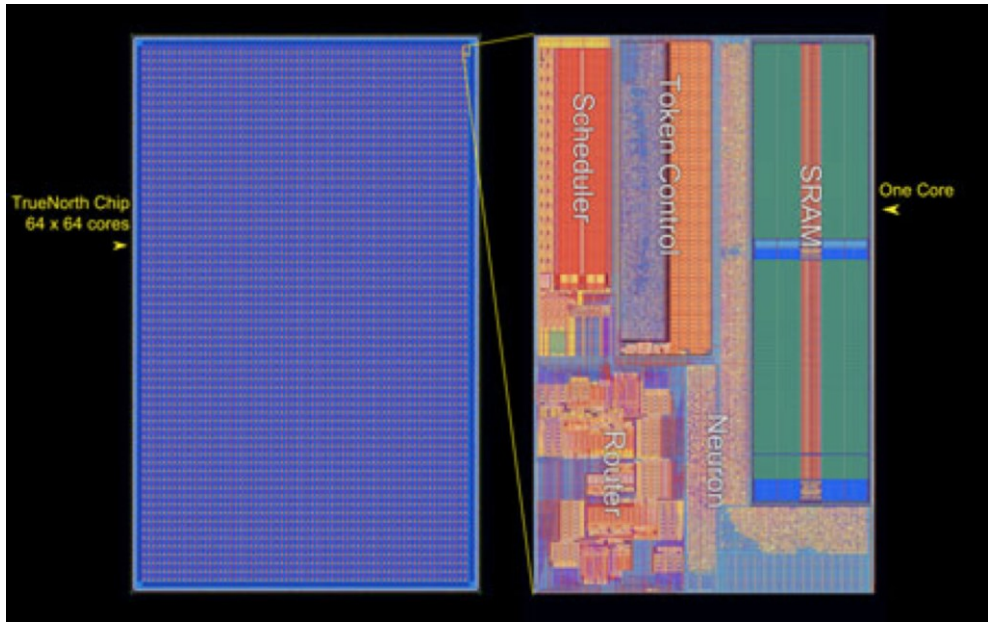


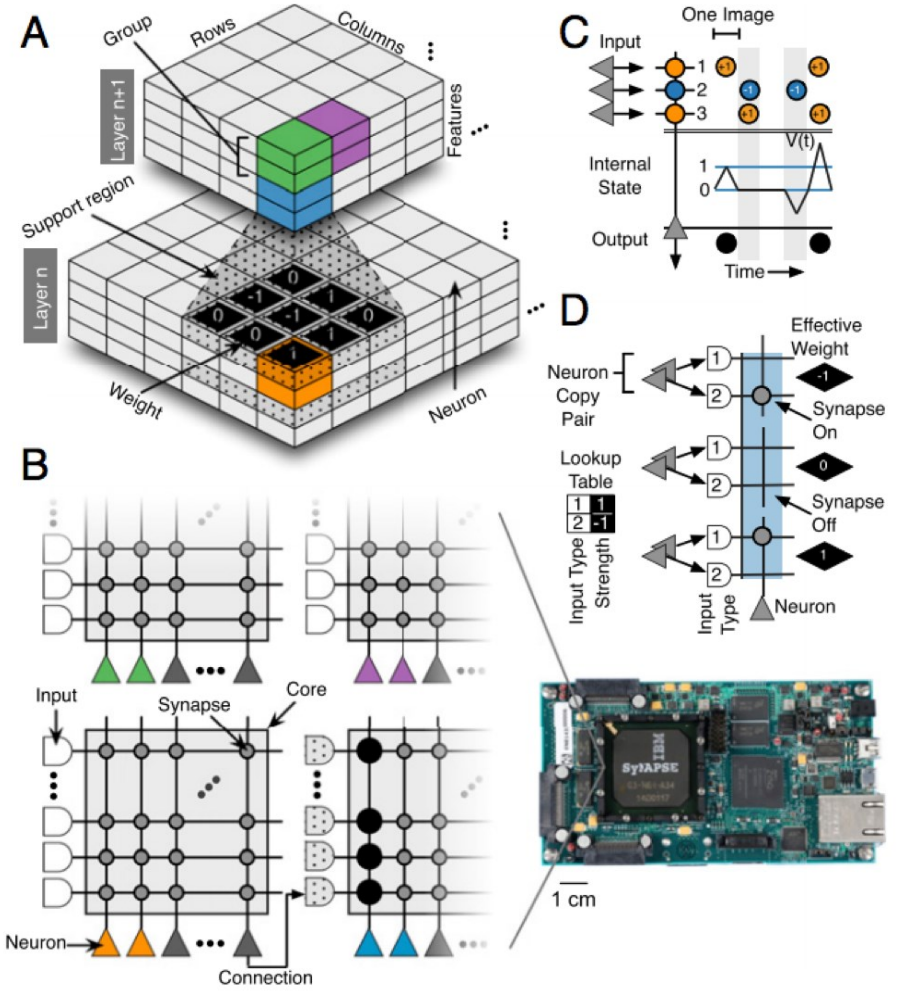
Spiking-based Machine Learning



Spiking Architecture

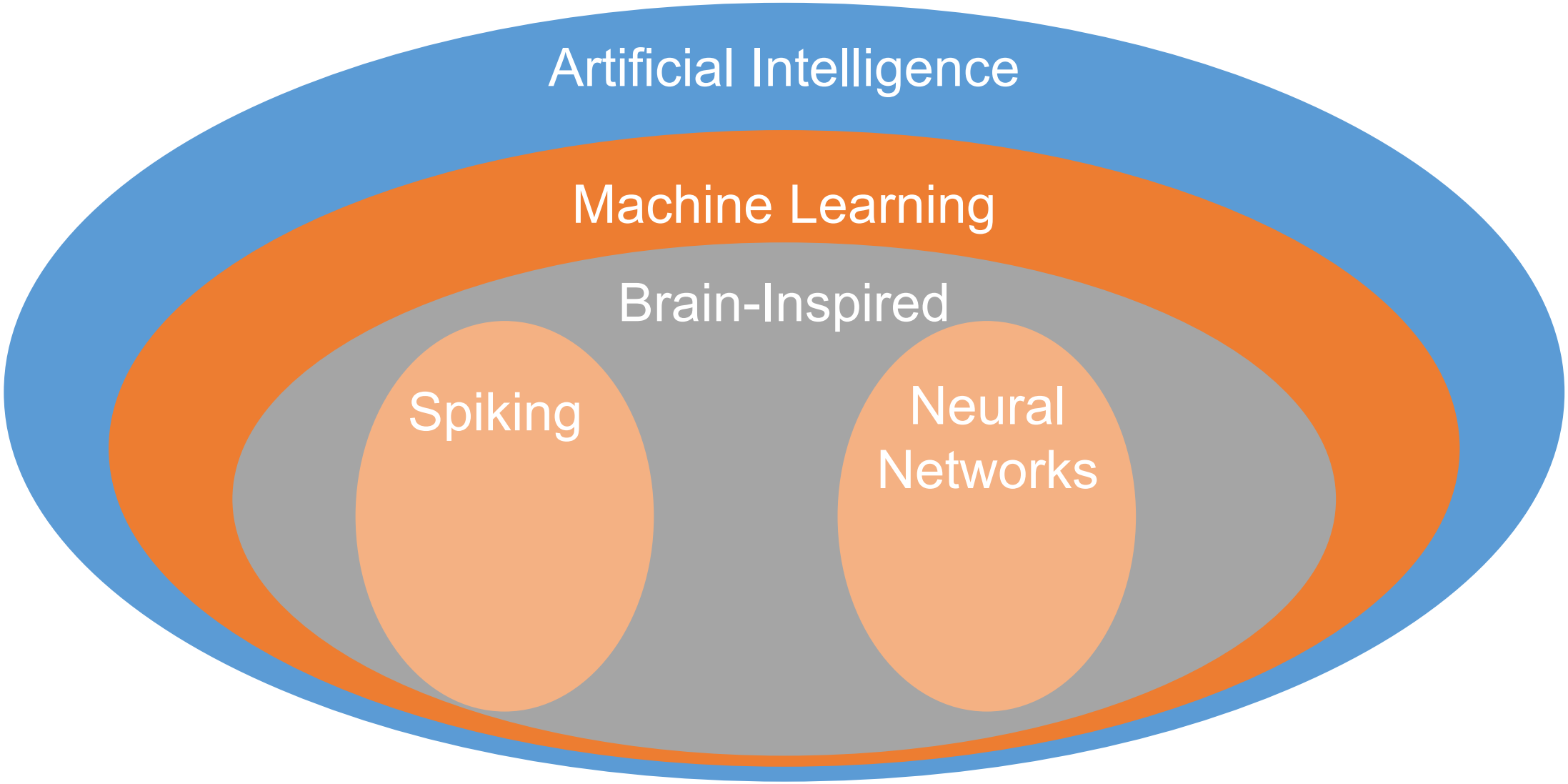
- Brain-inspired
- Integrate and fire
- Example: IBM TrueNorth



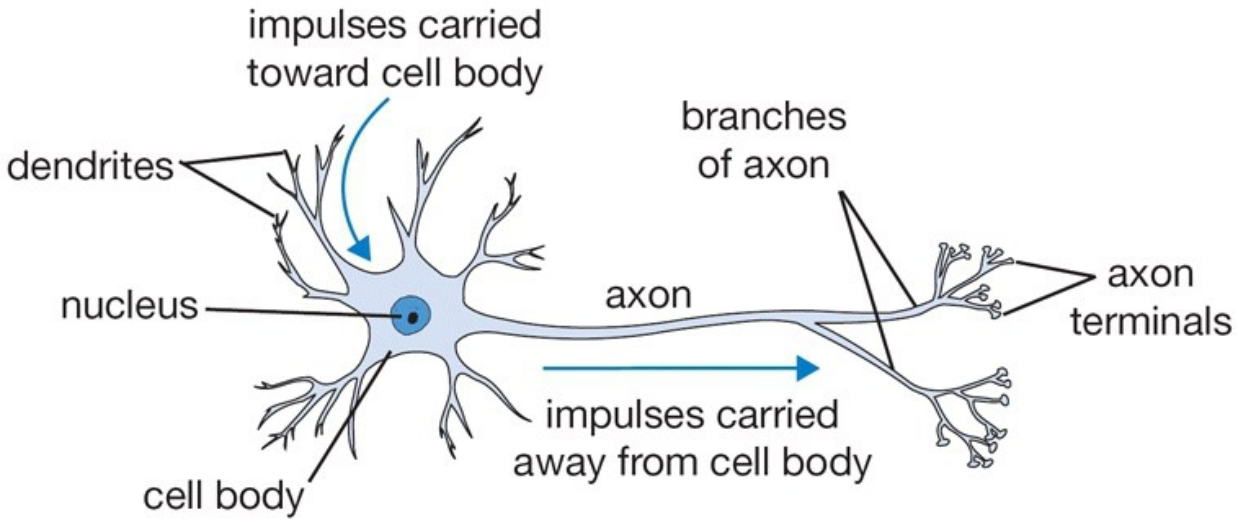


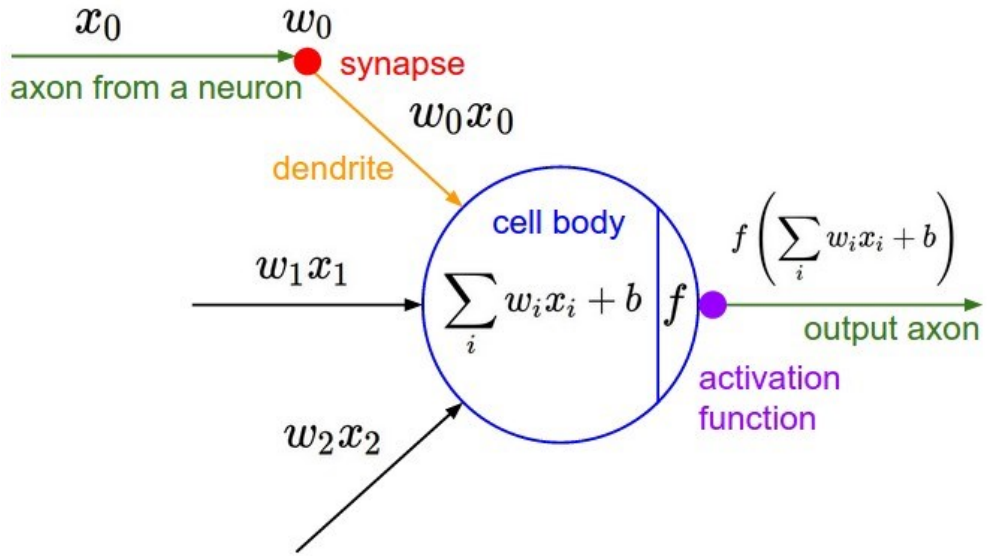
Merolla et al., Science 2014; Esser et al., PNAS 2016

Machine Learning with Neural Networks

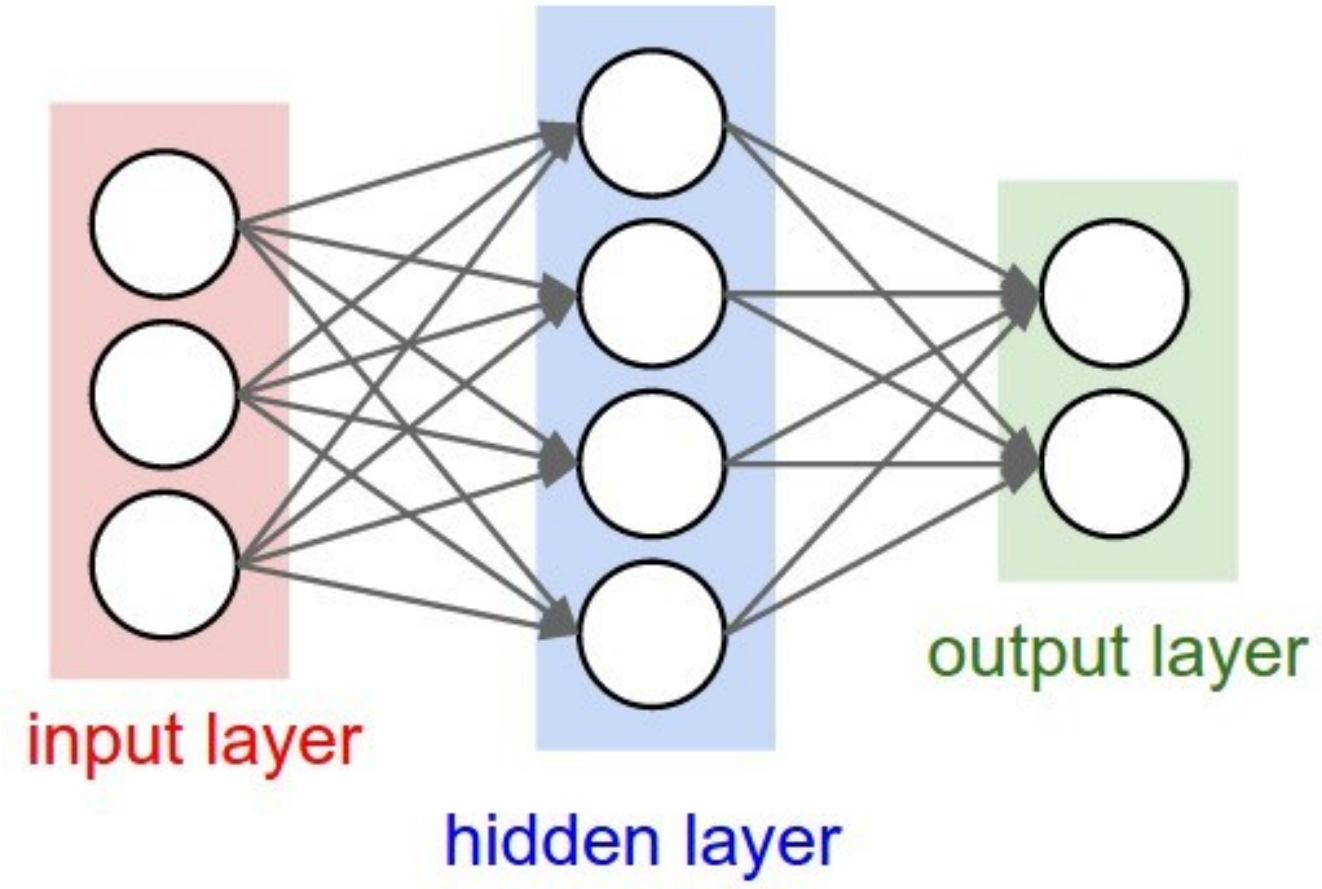


Neural - Weighted Sum

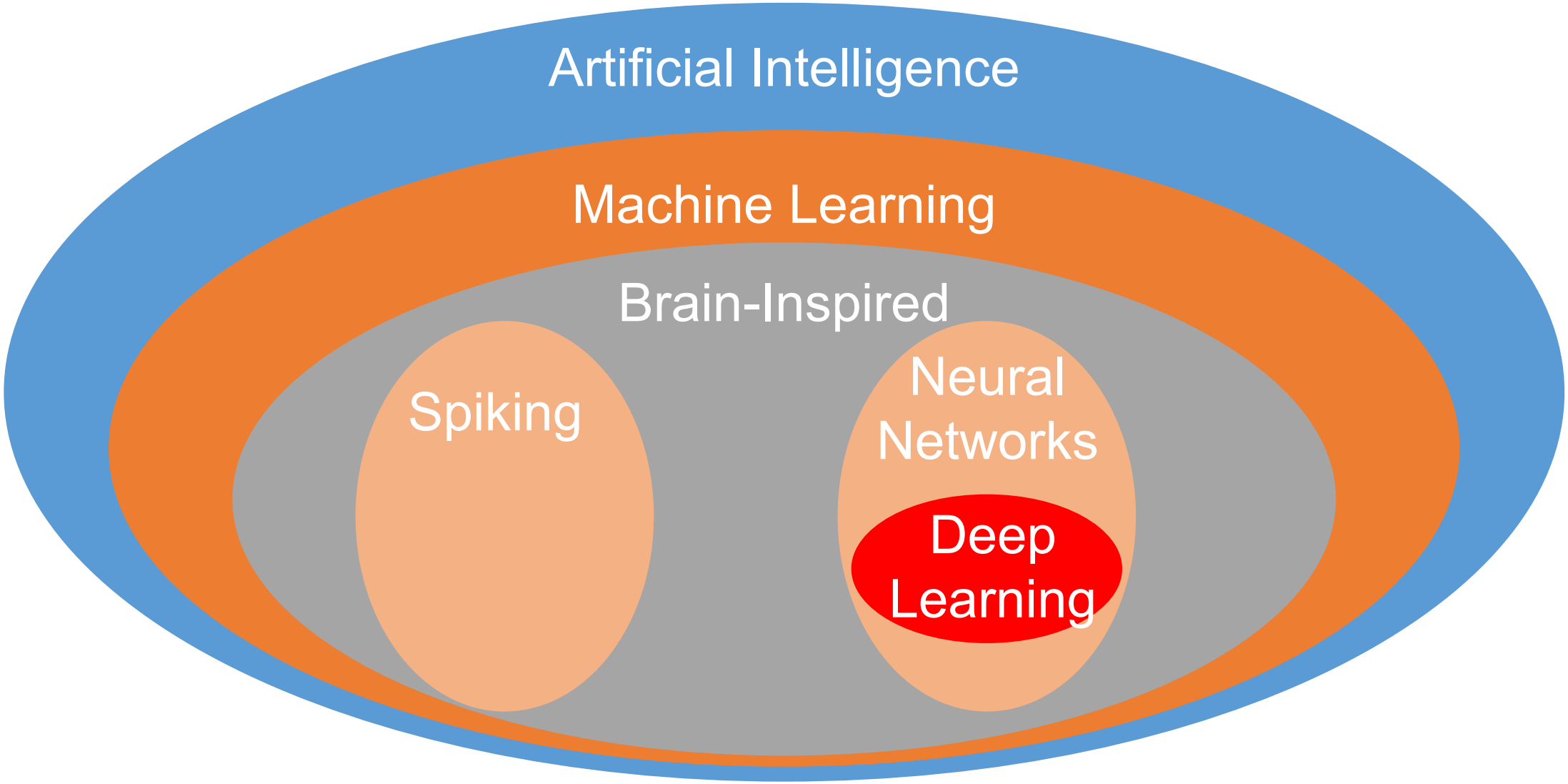




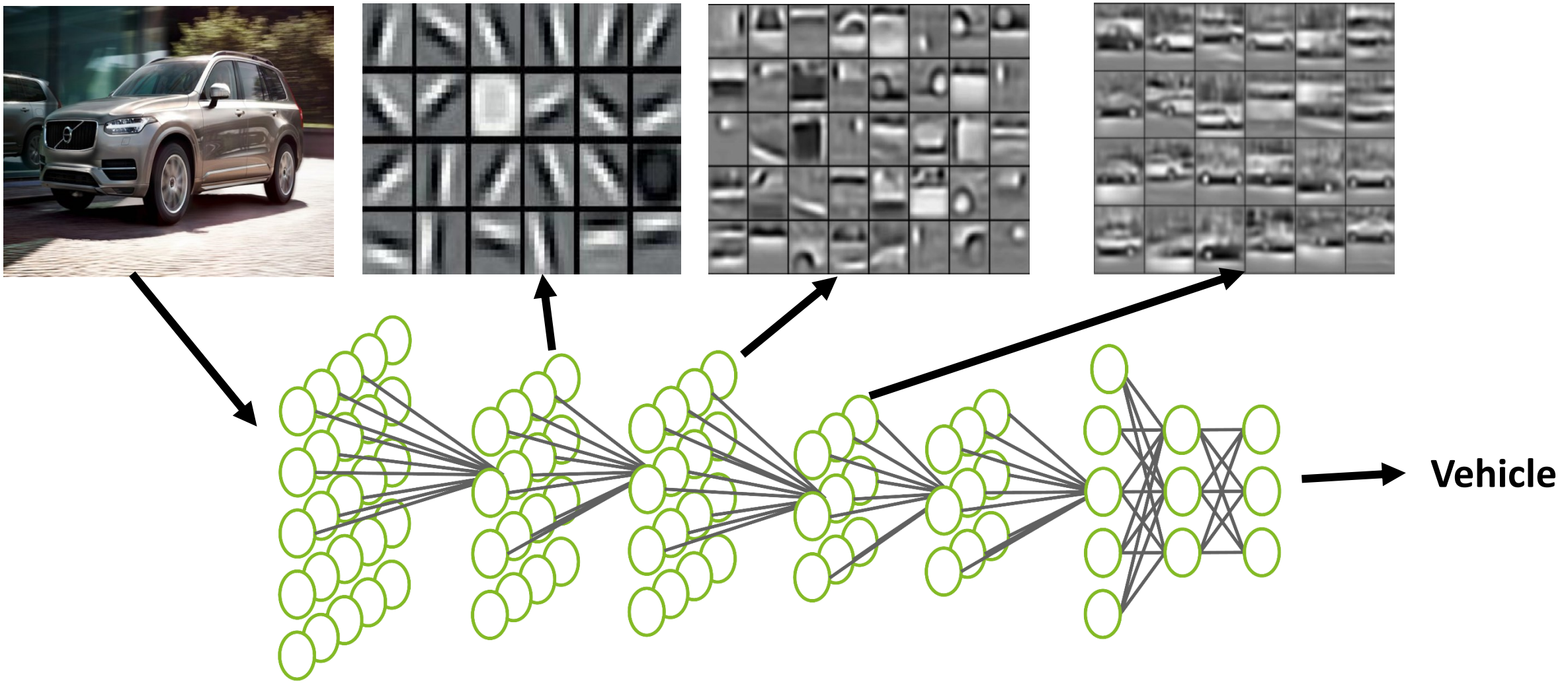
Neural Network - Many Weighted Sums



Machine Learning with Neural Networks

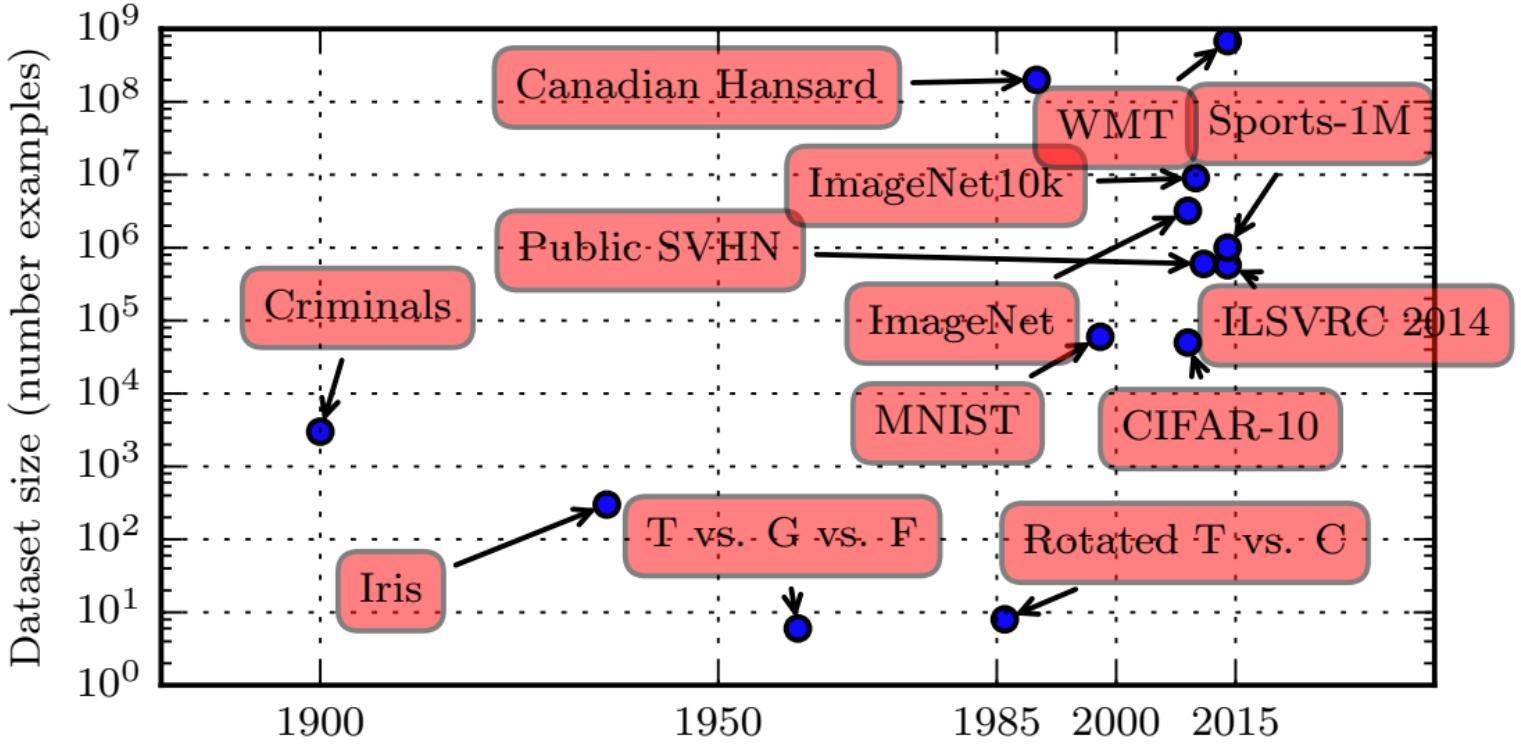


Deep Learning Example



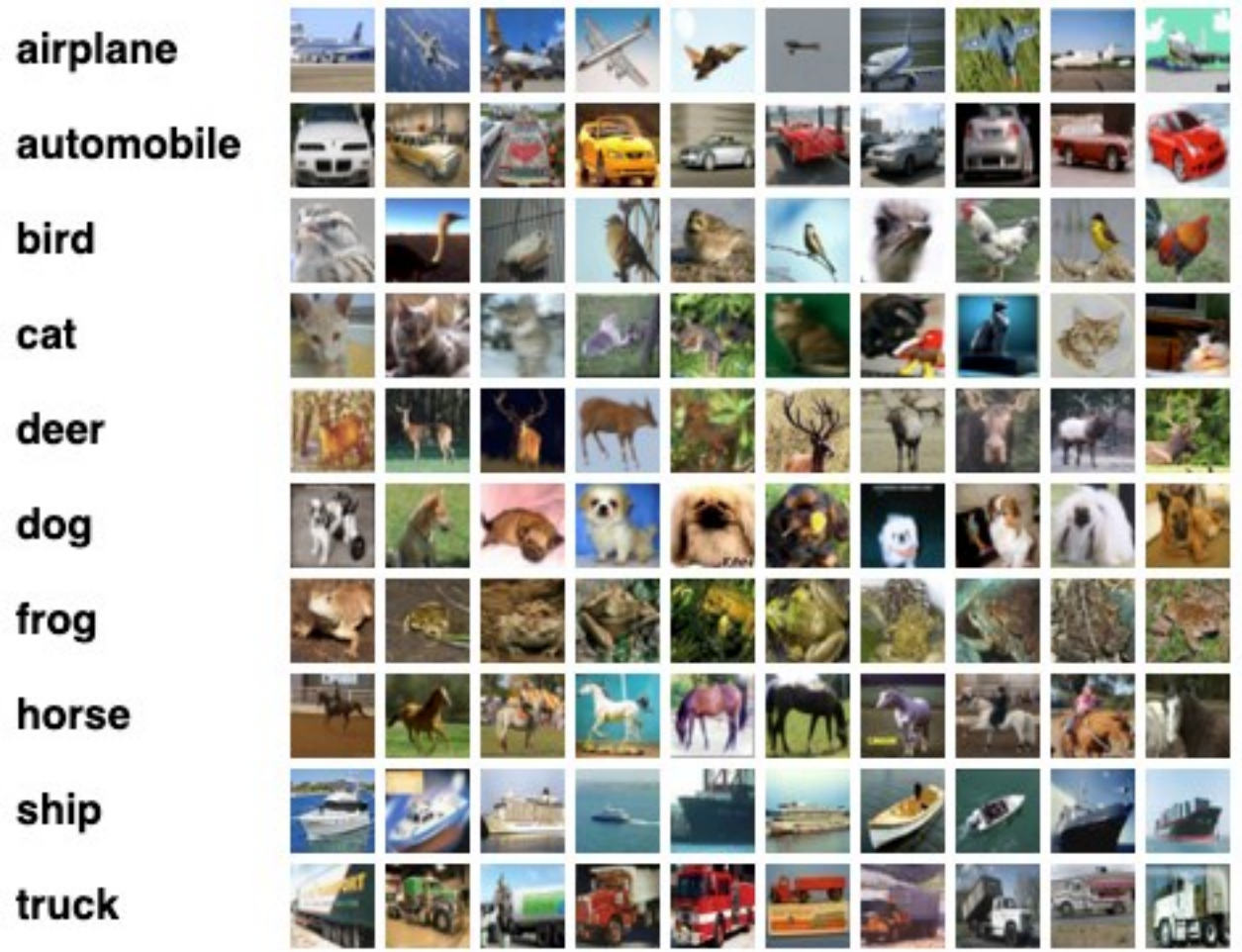
Increasing Dataset Sizes

- The size of datasets has expanded remarkably over time.
- The age of “Big Data” has made machine learning easier.



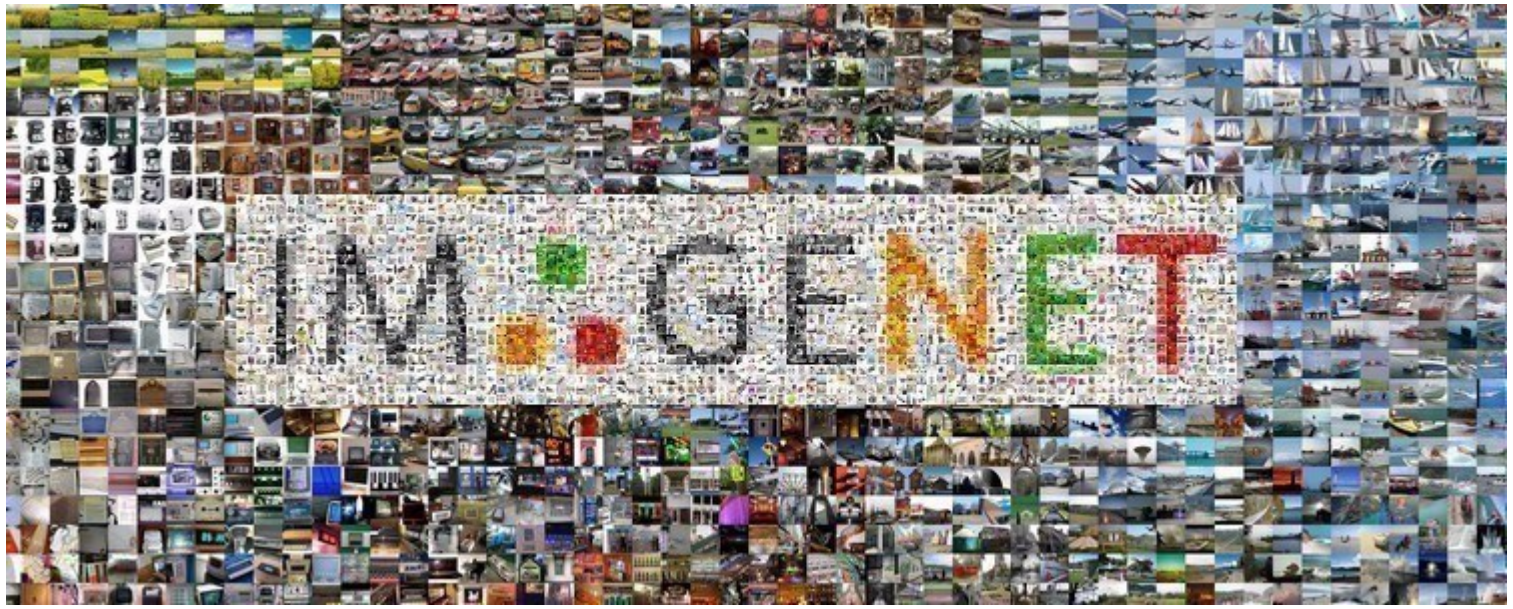
Increasing Dataset Sizes: CIFAR10

- 60,000 32x32 images
  - 50,000 training
  - 10,000 test
- 10 classes
  - 6,000 images/class



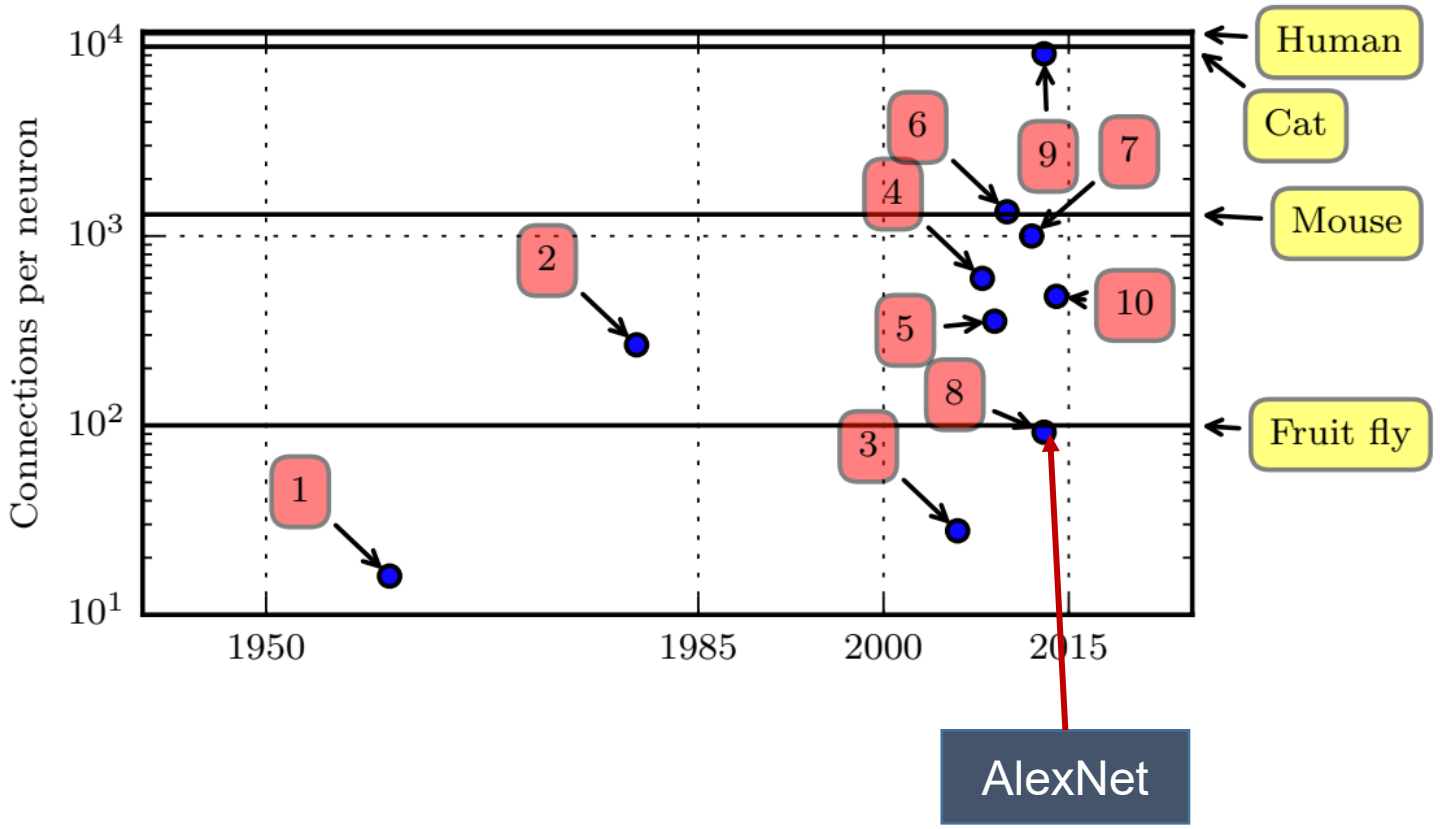
Increasing Dataset Sizes: ImageNet

- Varied in dimensions and resolution images
  - 1,281,167 training
  - 50,000 validation
  - 100,000 test
- 1000 classes



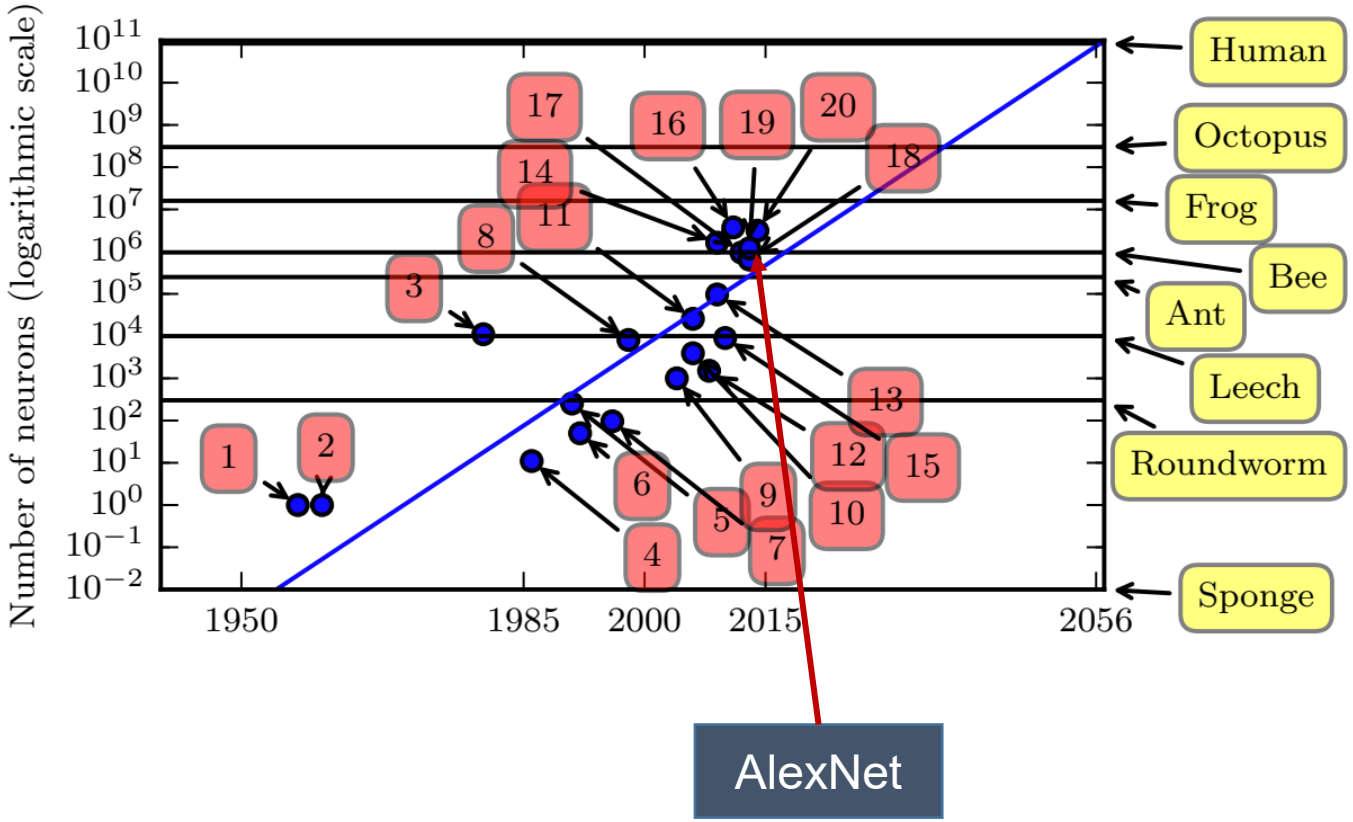
Increasing Model Sizes

- Model size:
  - # of connections / neuron
  - # of neurons
- Largely due to the availability of faster hardware (CPUs and GPUs)
- Expect to continue



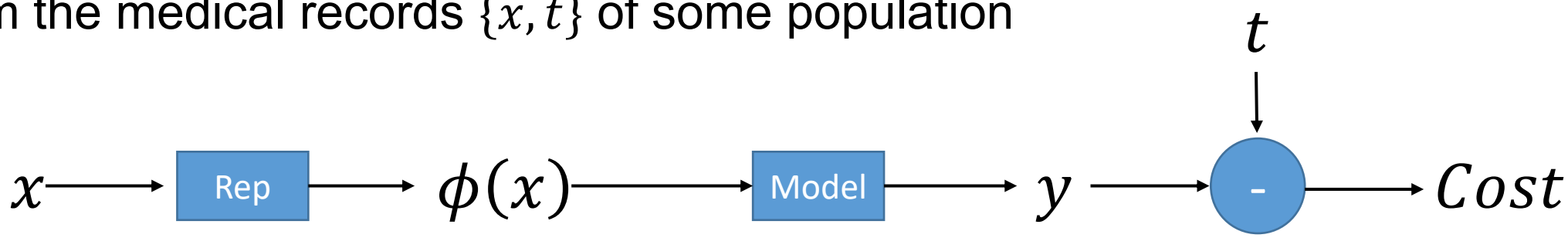
Increasing Model Sizes

- Model size:
  - # of connections / neuron
  - # of neurons
- Largely due to the availability of faster hardware (CPUs and GPUs)
- Expect to continue



Machine Learning

- Acquiring knowledge by extracting **patterns** from **raw data**
- Example: To predict a person's wellness *t* from their MRI scan x by learning patterns from the medical records $\{x, t\}$ of some population



  - x: *MRI scan*
  - $\phi(x)$: *data representation of MRI* scan
  - $y \in (0,1)$: model prediction with parameter w

$$y = f_w(\phi(x)) \triangleq \sigma(w^T\phi(x)), \text{ where } \sigma(x) = \frac{1}{1+e^{-x}}$$

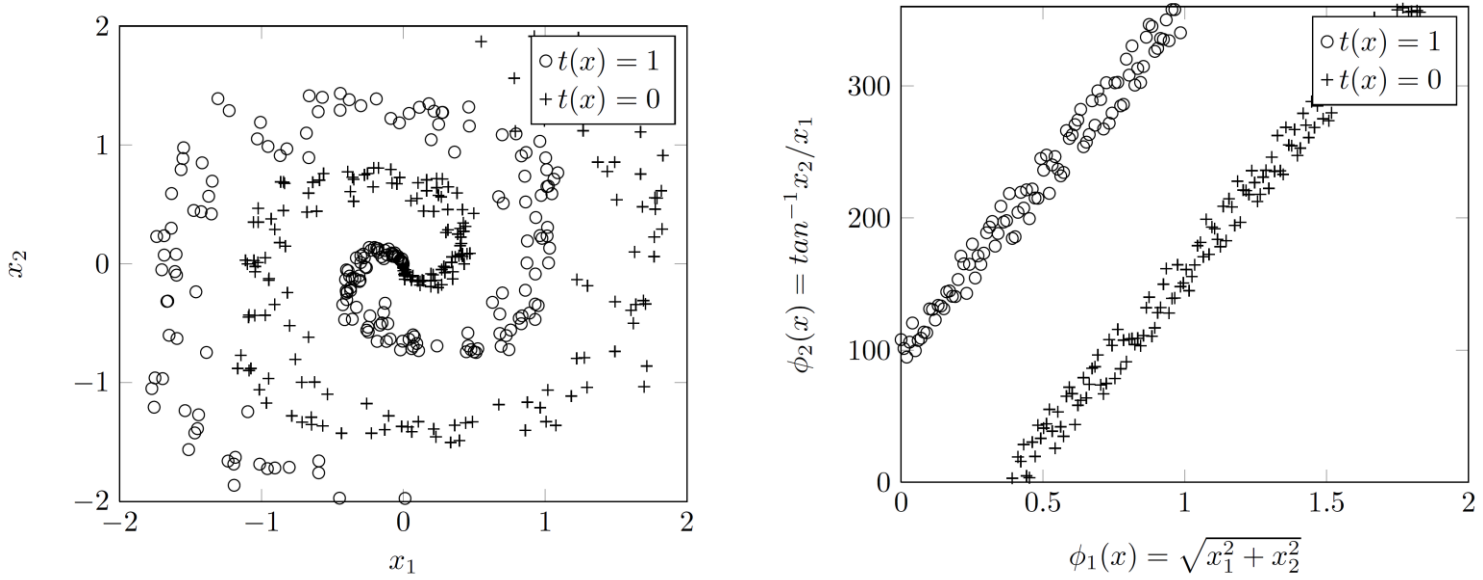
  - $t \in \{0,1\}$: ground-truth result associated with input x
- Cost: some distance between y and t (e.g.$\|y - t\|_2^2$), which is to be minimized w.r.t. w over the $\{x, t\}$ pairs

Machine Learning

- Essentially, we want to find a function $f_w(\phi(x))$to approximate $t(x)$
- In the present example, $f_w(\phi(x))$ bears a probabilistic interpretation of $p(t = 1|x; w)$
- The setting here is termed supervised learning as the ground-truth result t is given for each x

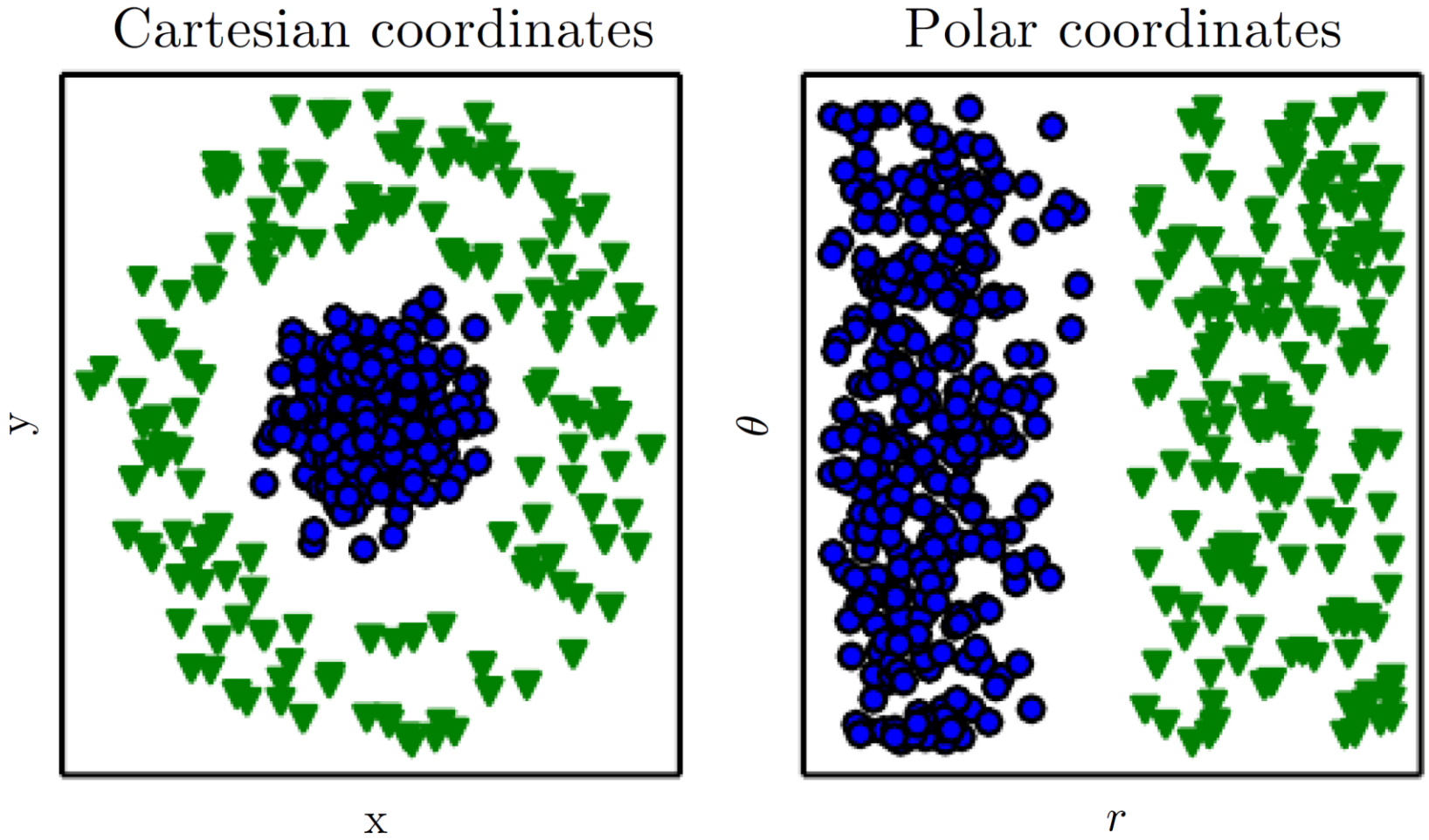
Data Representation - $\phi(x)$

- Data representation can critically determine the prediction performance



- In classic machine learning, hand-designed features are usually used
- For many tasks, it is however difficult to know what features should be used

Another Example of Data Representation



Deep Learning

- A machine learning approach whose data representation is based on building up a **hierarchy of concepts**, with each concept defined through its relation to simpler concepts
- Using the previous example, this amounts to learning a function of the following form

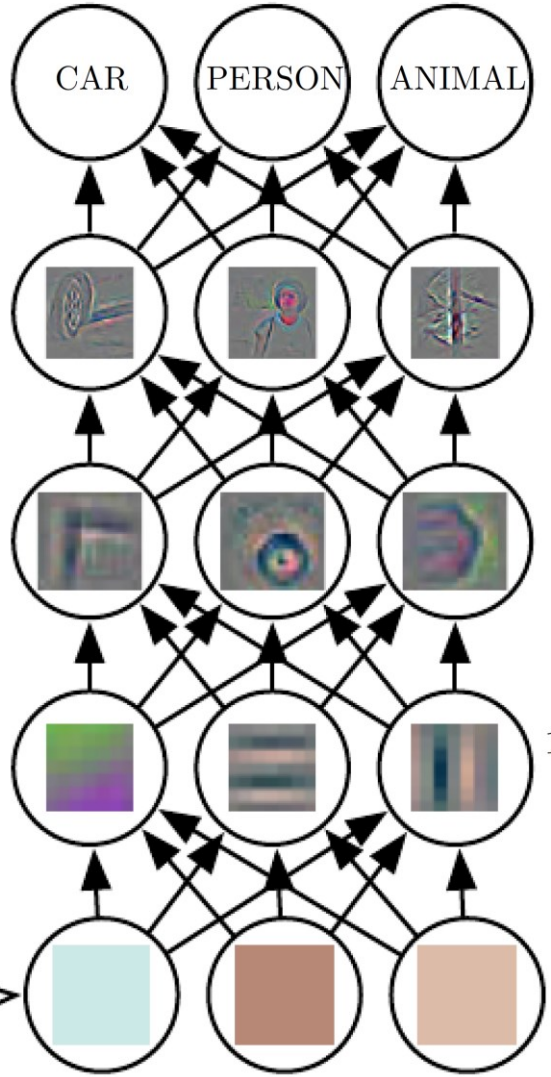
$$f_{w,\theta_n,\theta_{n-1},\ldots,\theta_1}(x) = \sigma(w^T \phi_{\theta_n}(\phi_{\theta_{n-1}}(\ldots \phi_{\theta_1}(x))))$$

where $w, \theta_n, \theta_{n-1}, \ldots, \theta_1$ are model parameters

Hierarchy of concepts/features

- $\phi_\theta(\cdot)'s$ are generally vector-valued functions, e.g. $\phi_\theta(x) = \sigma(\theta_x)$
- Such a deep model allows to construct a complicated function $f(x)$ from nested composition of simpler functions $\phi_\theta(\cdot)'s$

Example: Feedforward Deep Networks



$$\sigma(\phi_{\theta_3}(\phi_{\theta_2}(\phi_{\theta_1}(x))))$$

$$\phi_{\theta_3}(\phi_{\theta_2}(\phi_{\theta_1}(x)))$$

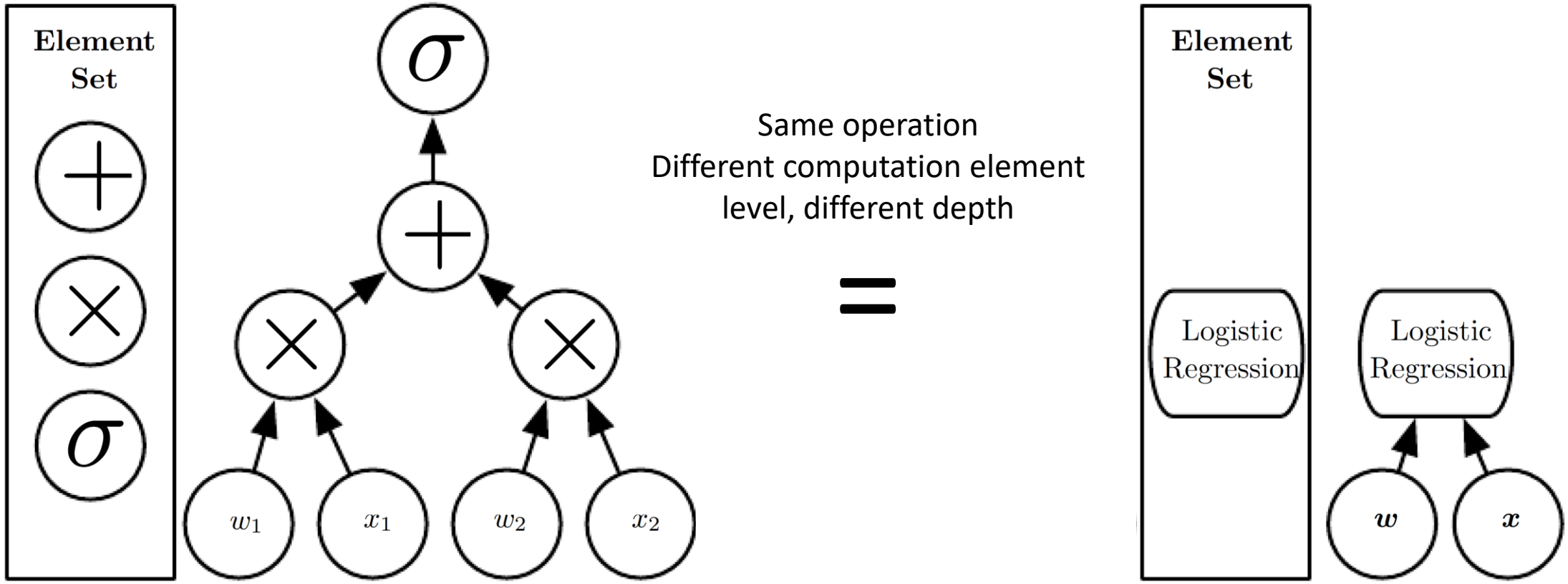
$$\phi_{\theta_2}(\phi_{\theta_1}(x))$$

$$\phi_{\theta_1}(x)$$

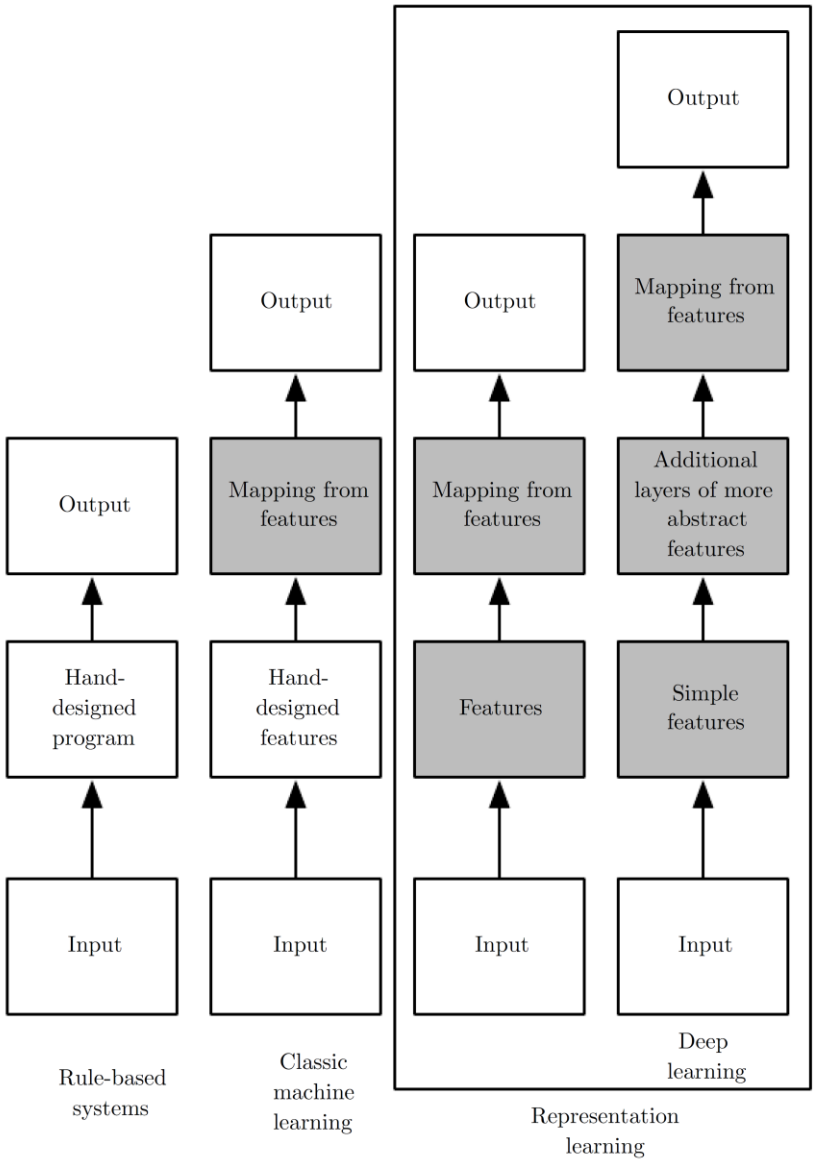
$$x$$

Computational Graphs

- Mapping an input to an output, each node performs an operation



Different AI Systems

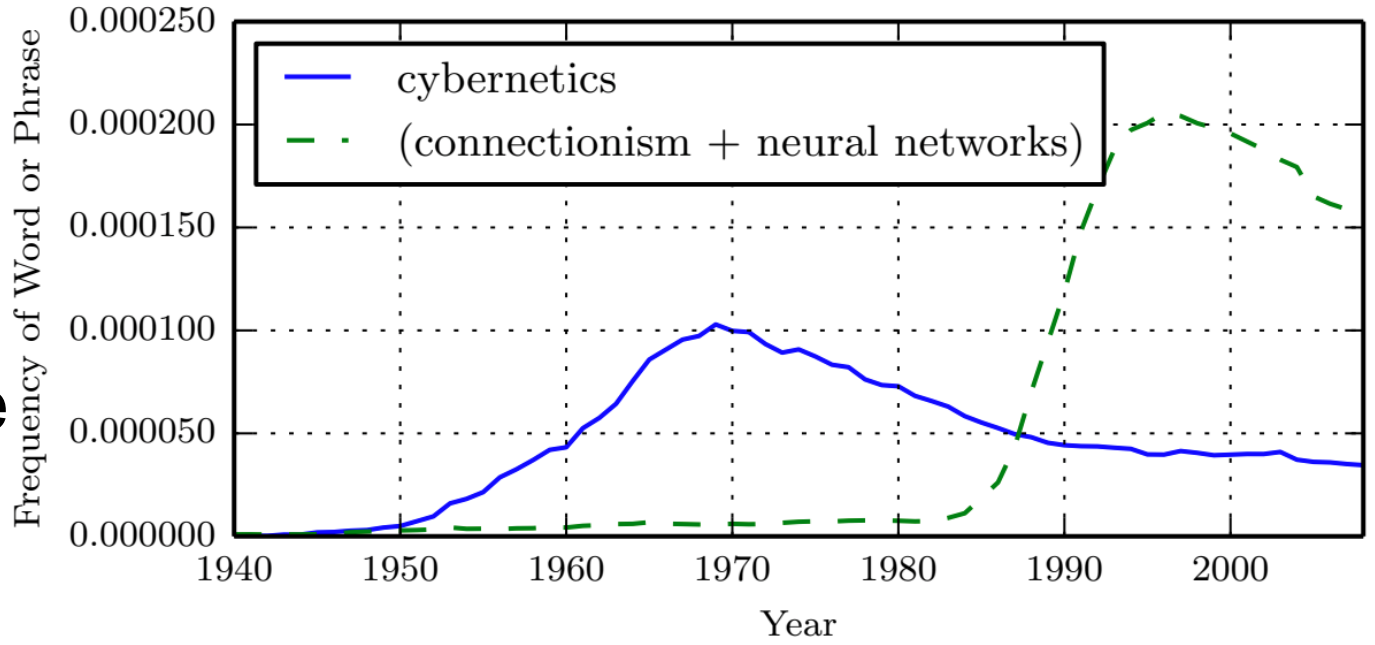


History of Deep Learning

- **Cybernetics** (1940s~1960s): Systems inspired by biological brains
  - Perceptron (Rosenblatt, 1958, 1960), Adaptive Linear Element, ADALINE (Widrow and Hoff, 1960)
- **Connectionism** (1980s~1990s): Connected simple computational units
  - Neocognition (Fukushima, 1980); Recurrent Neural Networks (Rumelhart et al., 1986); Convolutional Neural Networks (LeCun et al., 1998); Long Short-Term Memory (Hochreiter and Schmidhuber, 1997)
- **Deep Learning** (2006s~): Deeper networks and deep generative models
  - Deep Belief Networks (Hinton et al., 2006); Deep Boltzmann Machine (Salakhutdinov et al, 2009); Neural Turing Machine (Graves et al., 2014); Variational Autoencoder (Kingma et al.,2014); Generative Adversarial Networks (Goodfellow et al. 2014)

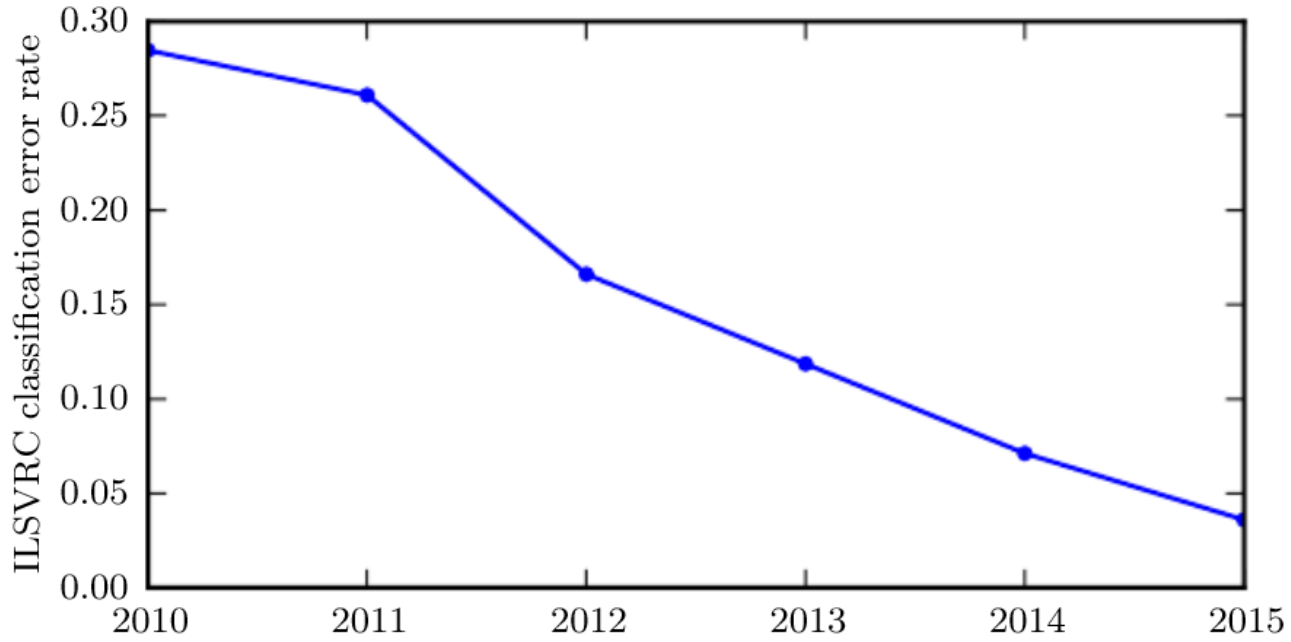
Many Names of Deep Learning

- Three waves of neural networks
  - Cybernetics
    - (1940s – 1960s)
  - Connectionism
    - (1980s – 1990s)
  - Deep learning
    - (2006s – now)
- Diminished role of neuroscience
  - Simply do not have enough information about the brain



Recent Milestone

- ImageNet Large Scale Visual Recognition Challenge



Top-5 error rate reduced from 26% to be less than 5%