CS 758: Advanced Topics in Computer Architecture

Lecture #11: Intro to DNNs

Professor Matthew D. Sinclair

Some of these slides were developed by Tushar Krishna at Georgia Tech and Joel Emer and Vivienne Sze at MIT

Slides enhanced by Matt Sinclair

AI/ML Applications

"AI is the new electricity" – Andrew Ng

Object Detection

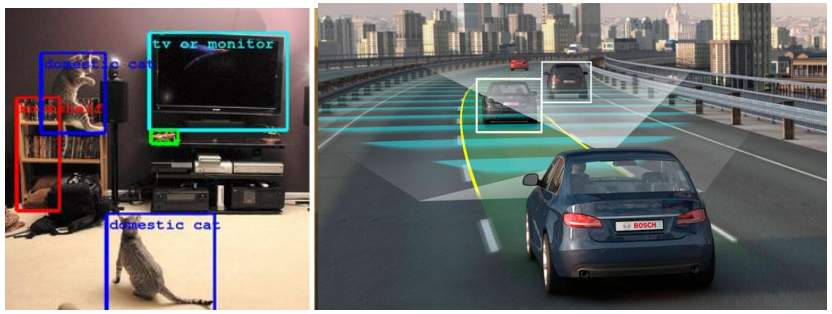
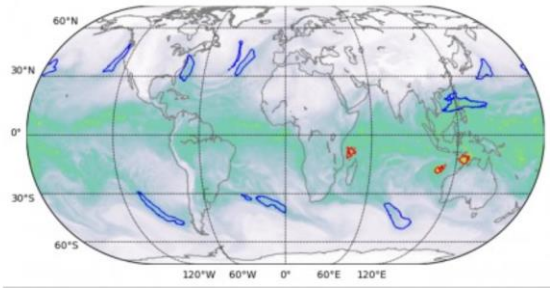
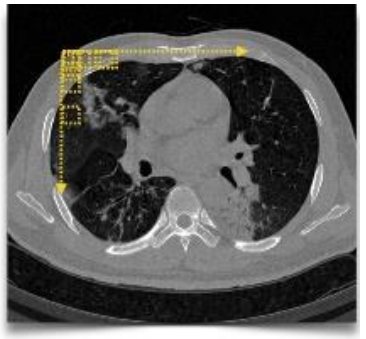


Image Segmentation

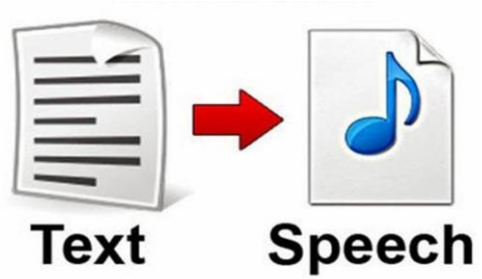


Medical Imaging



Speech Recognition

Text to Speech

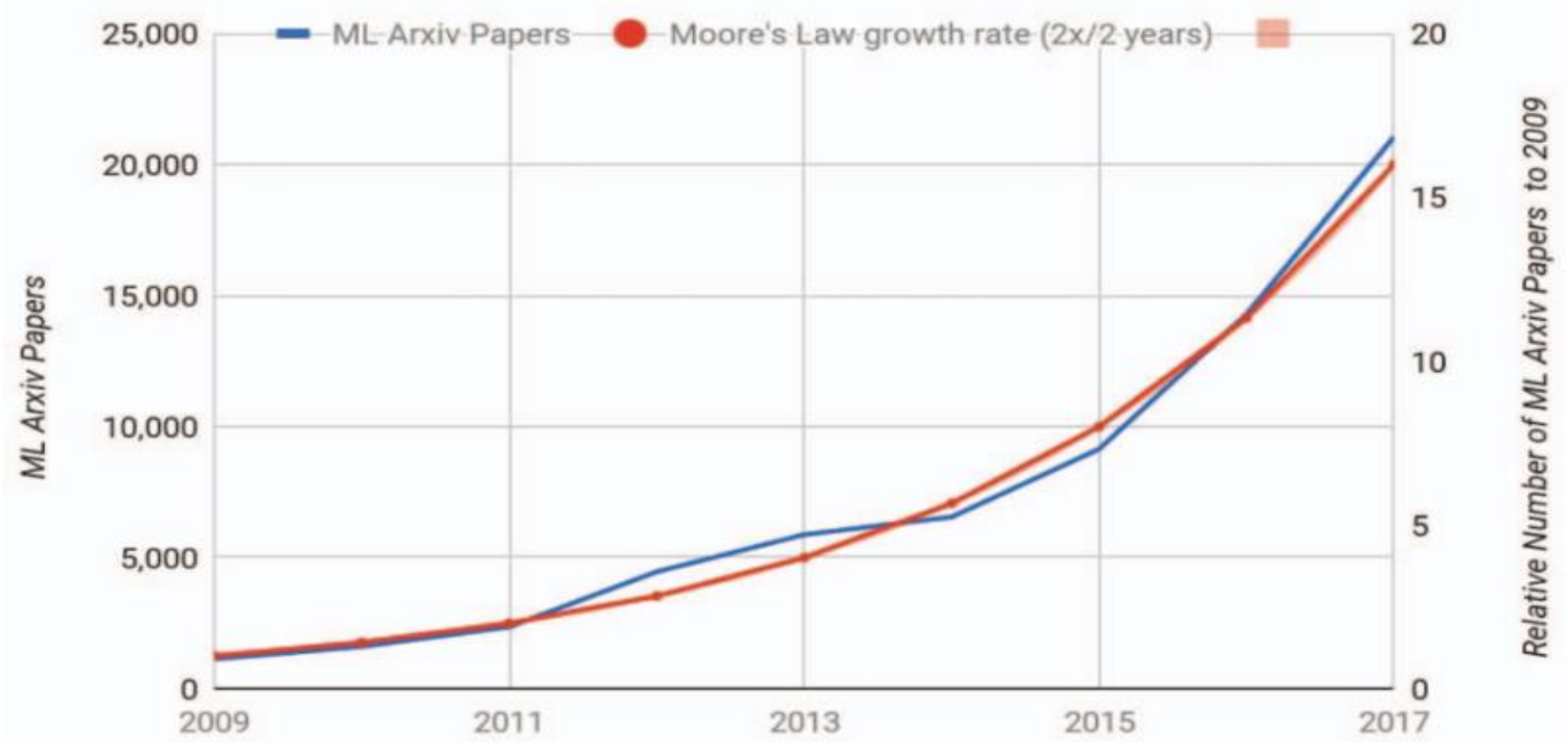


Recommendations

Games

Autonomous Agents, Intelligent Personal Assistants, …

ML is HOT

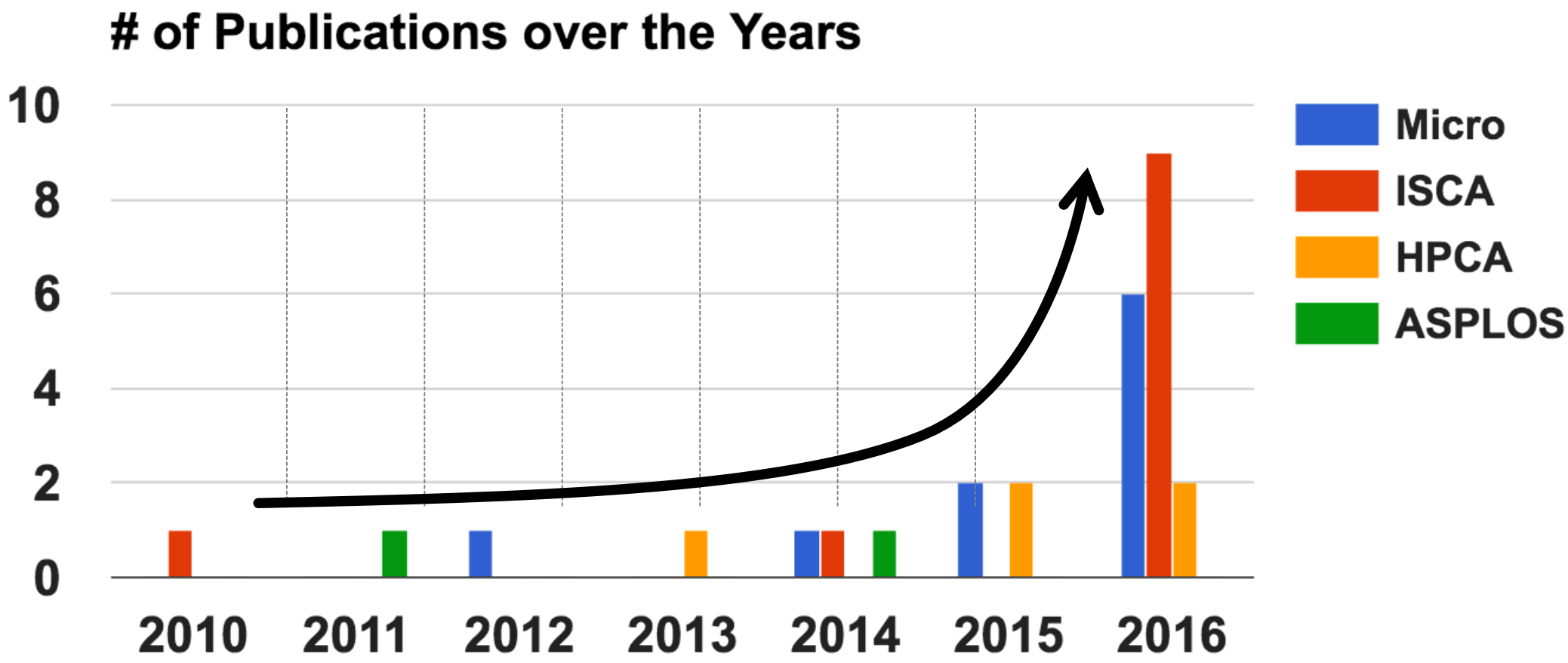


Source: Jeff Dean, Google

Over 1.5B invested in AI chip startups in 2017 alone

Not clear what the endgame will be

Publications at Architecture Conferences



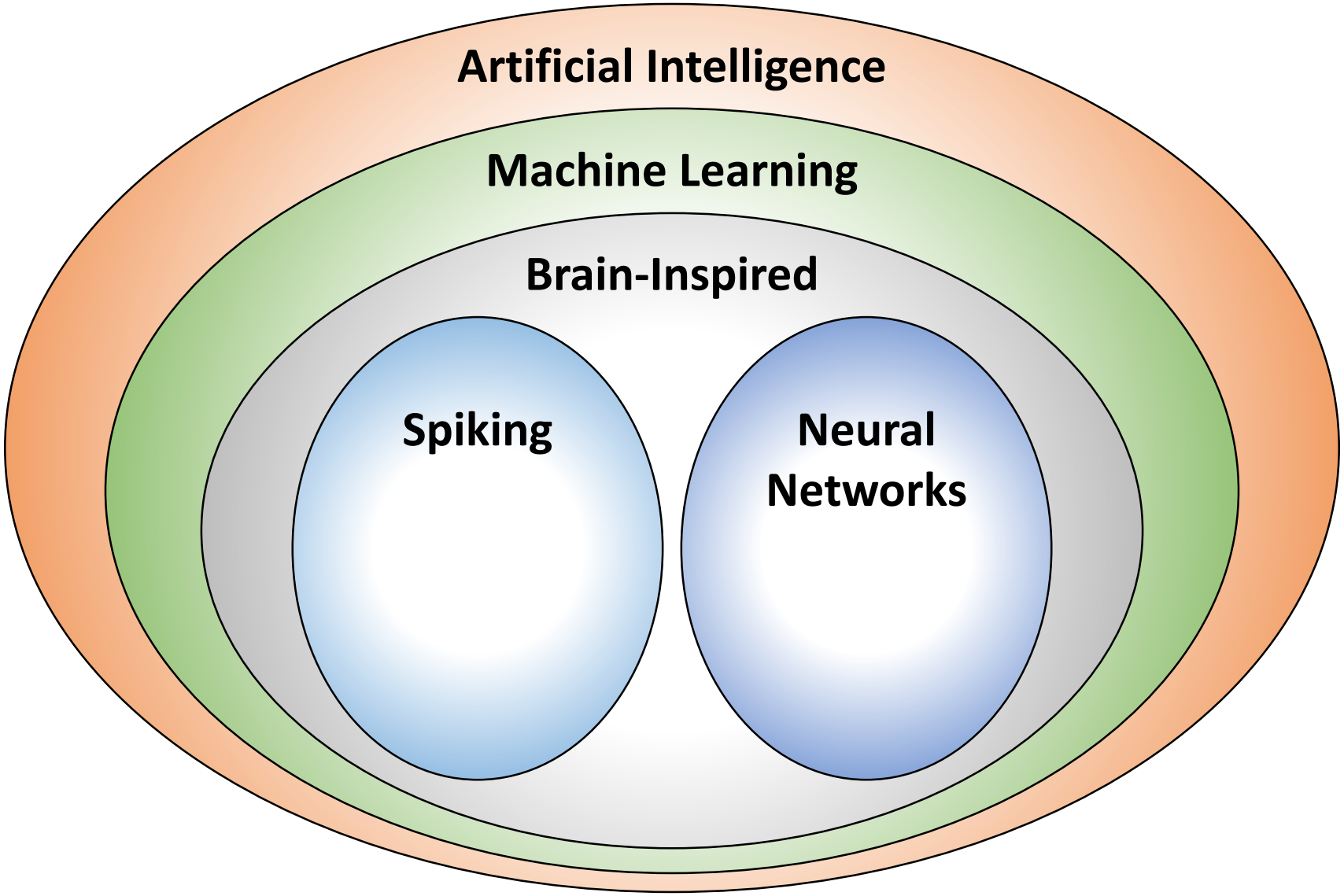
Slide Courtesy: Joel Emer and Vivienne Sze

Key Questions to Answer

- Separating the Hype from the (Remaining) Opportunity

- Computer Architecture perspective
  - The role of HW Acceleration in the AI/ML Boom
    - Understand Compute and Memory behavior of Deep Learning Workloads
    - Understand Limitations of Current Platforms (CPU and GPU)
    - What about Moore's Law?
  - Opportunities and Challenges with custom HW Accelerators
    - Techniques to reduce computation
    - Techniques to to reduce data movement
    - Techniques to to reduce memory overhead
    - Performance vs Energy vs $$

Machine Learning with Neural Networks



Slide Courtesy: Joel Emer and Vivienne Sze

NeurON: Weighted Sum

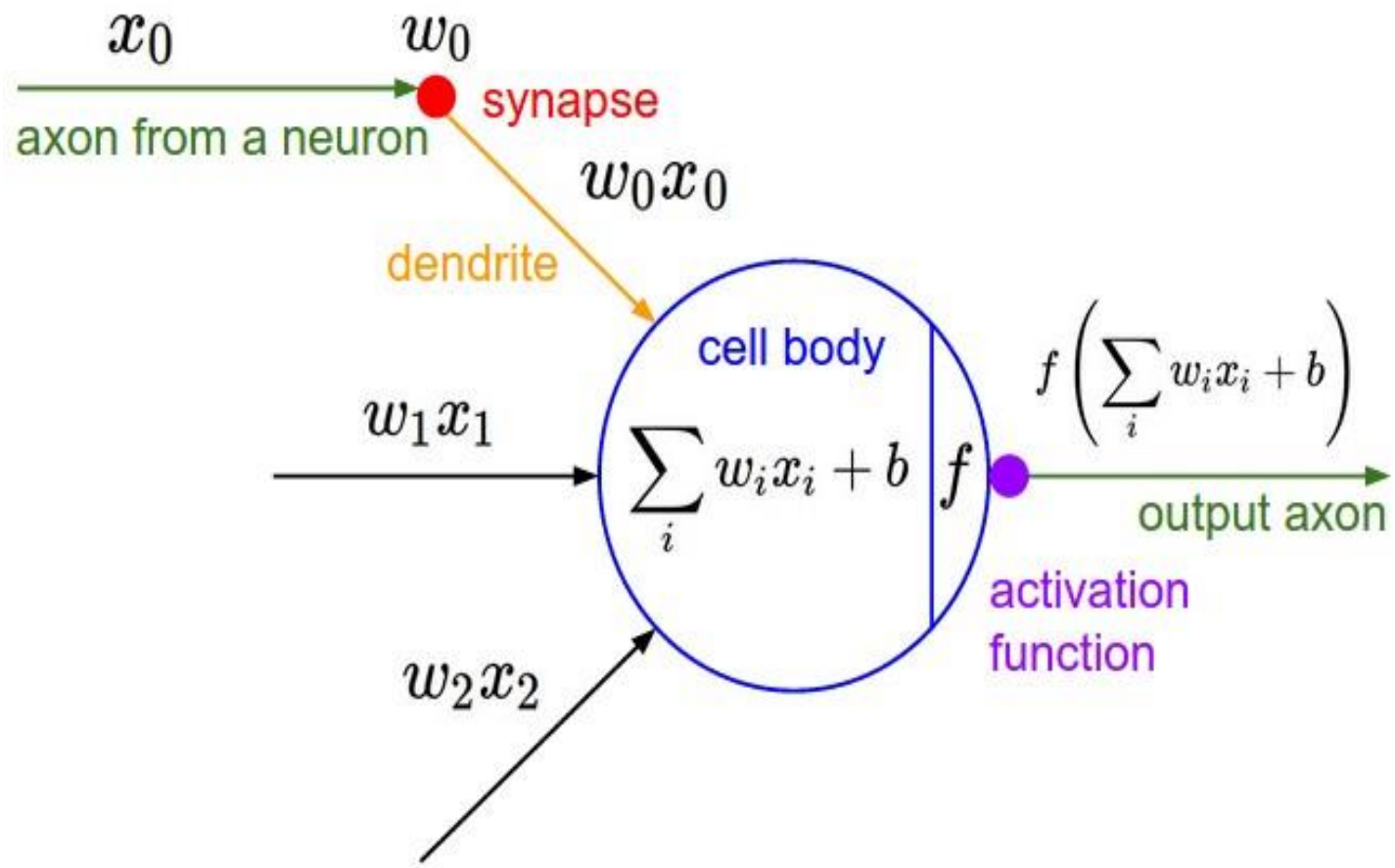
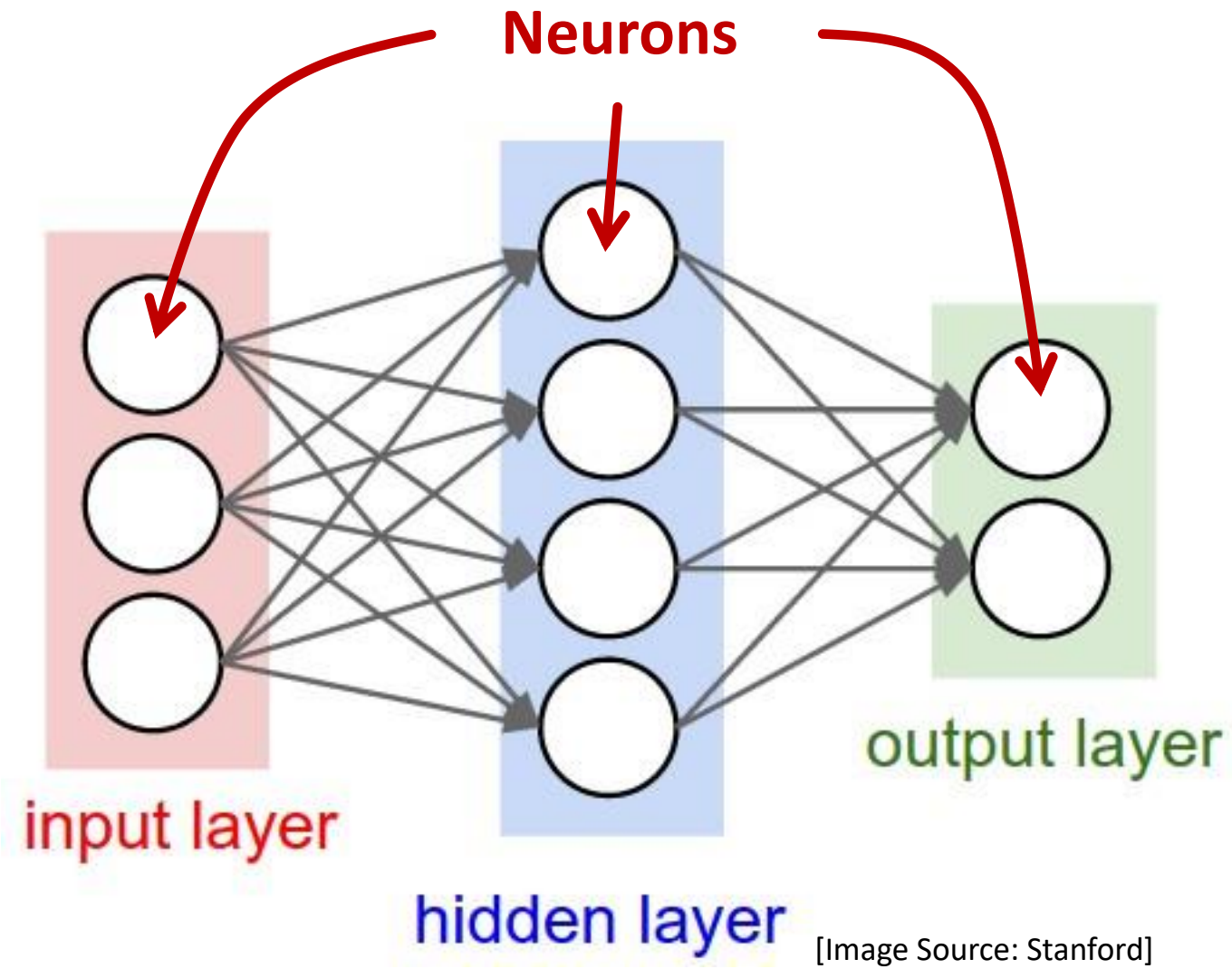


Image Source: Stanford

Slide Courtesy: Joel Emer and Vivienne Sze

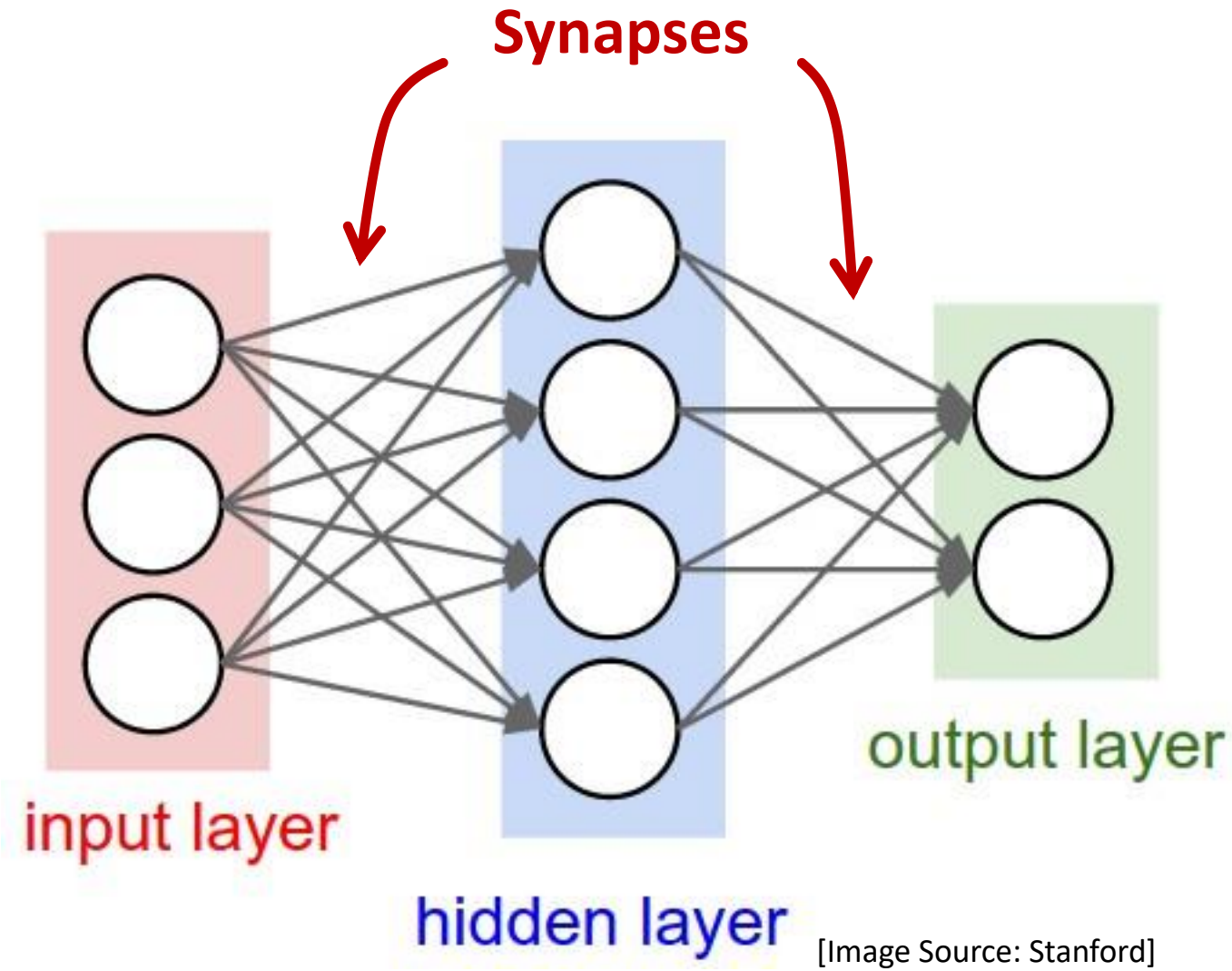
NEURAL NETWORK



[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

NEURAL NETWORK



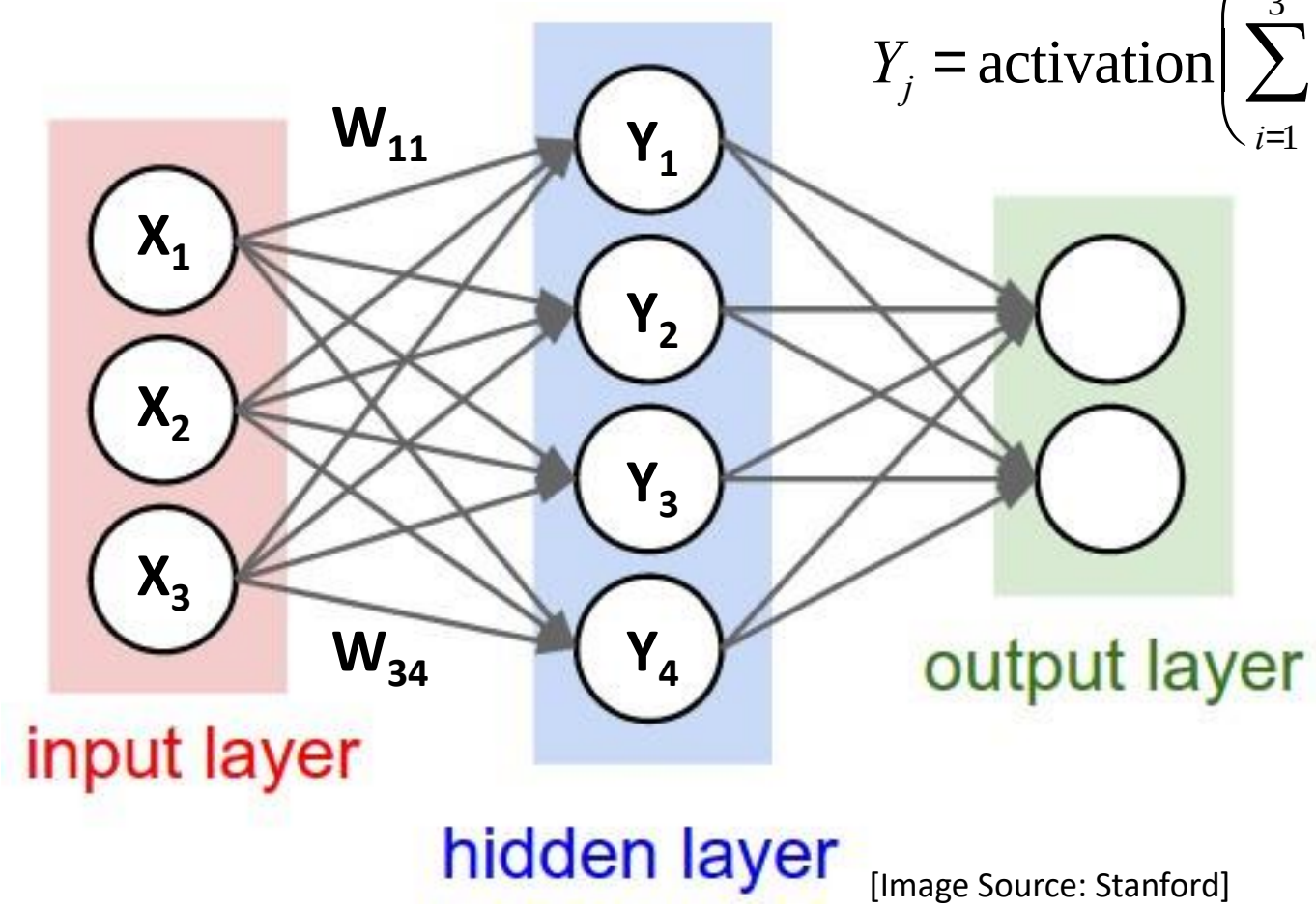
[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

NEURAL NETWORK: Multiple weighted sums

Each **synapse** has a **weight** for neuron **activation**

$$Y_j = \text{activation}\left(\sum_{i=1}^{3} W_{ij} \times X_i\right)$$



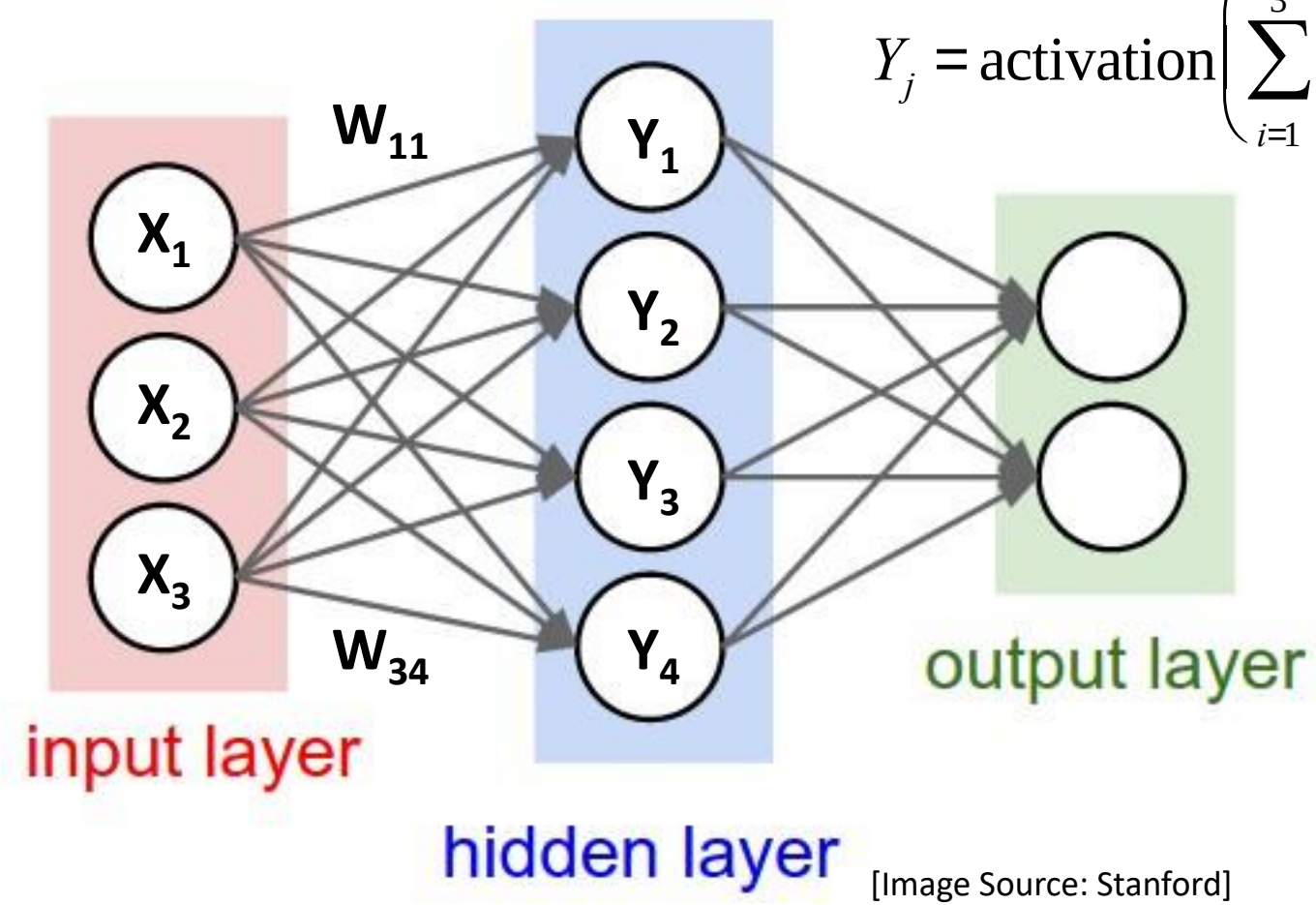
[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

NEURAL NETWORK: Multiple weighted sums

Weight Sharing: multiple synapses use the **same weight value**

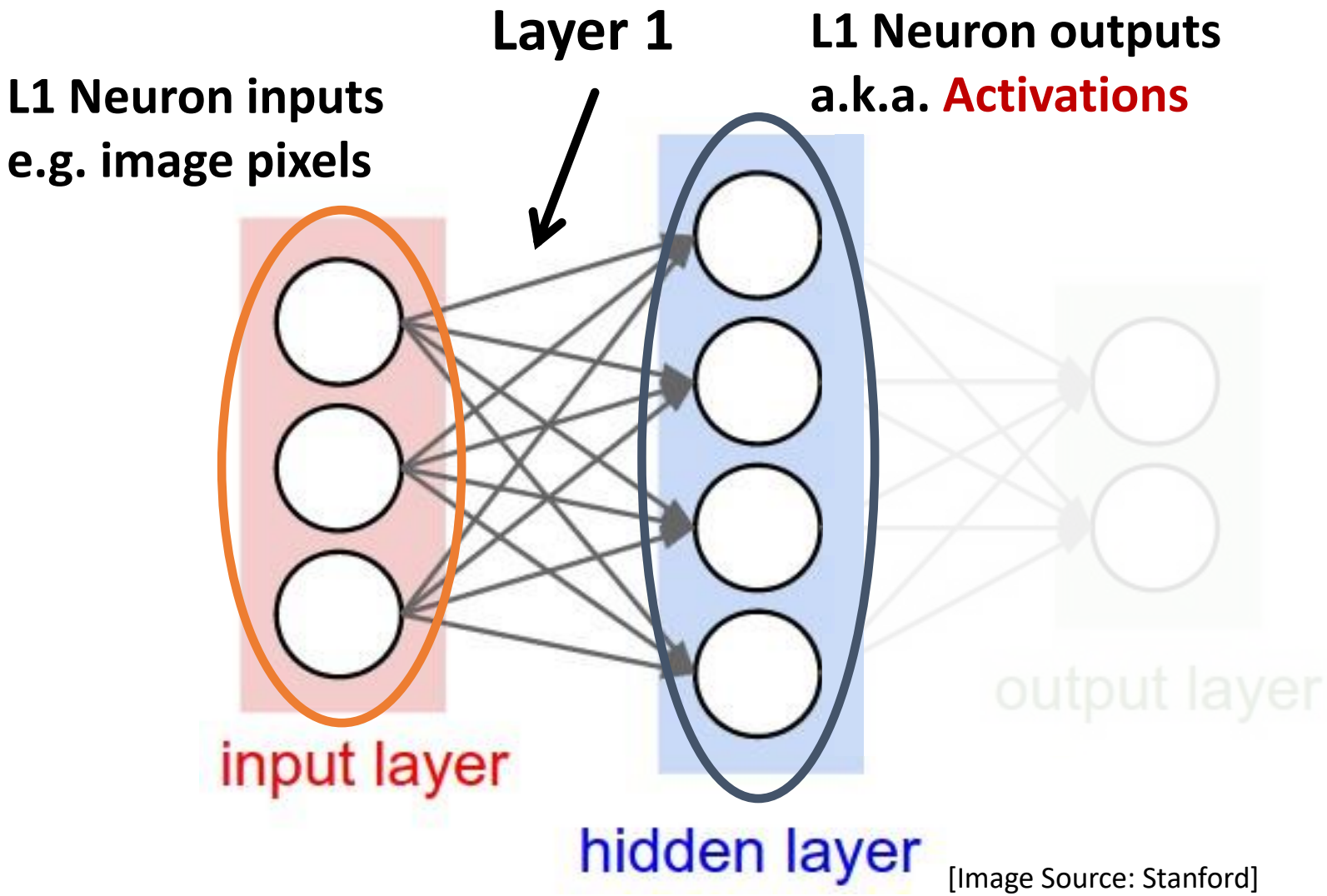
$$Y_j = \text{activation}\left(\sum_{i=1}^{3} W_{ij} \times X_i\right)$$



[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

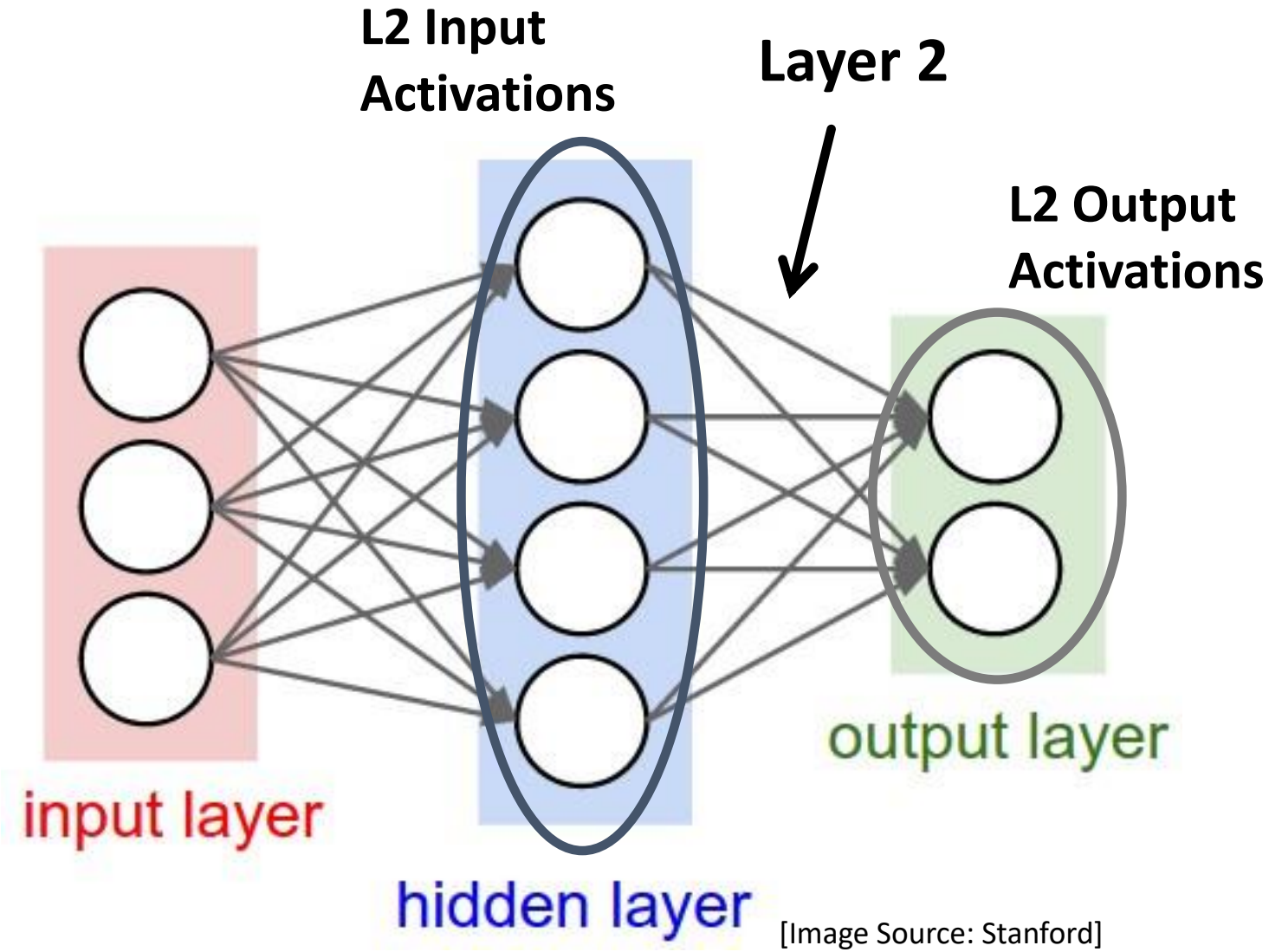
NEURAL NETWORK Terminology



[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

NEURAL NETWORK Terminology

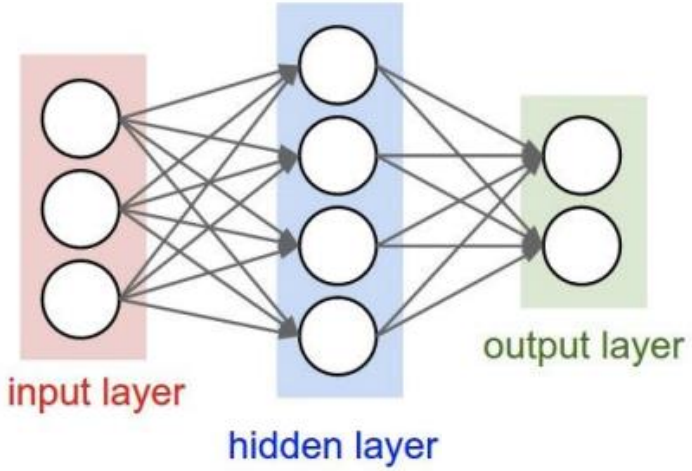


[Image Source: Stanford]

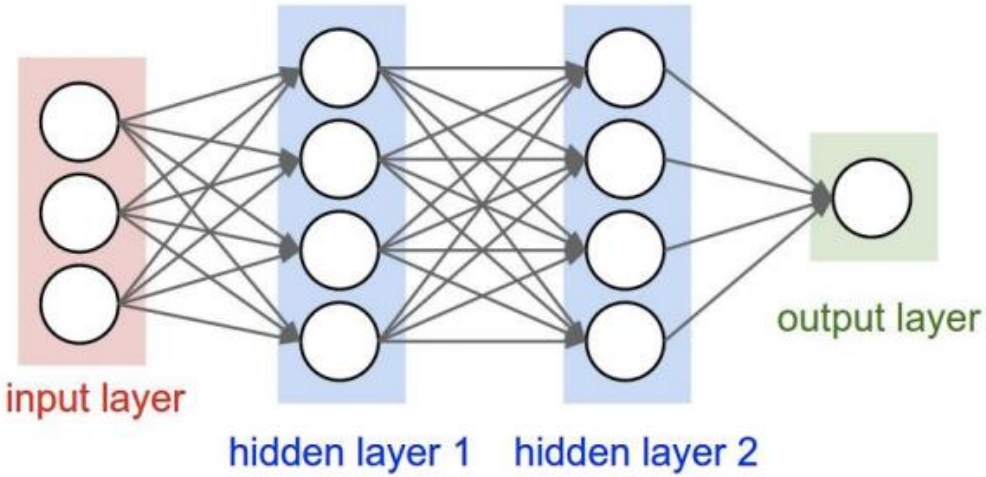
Slide Courtesy: Joel Emer and Vivienne Sze

NEURAL NETWORK Terminology

A **layer** can refer to a set activations or a set of weights.
In this class, we use **layer to refer to a set of weights**.



"2-layer Neural Net", or
"1-hidden-layer Neural Net"



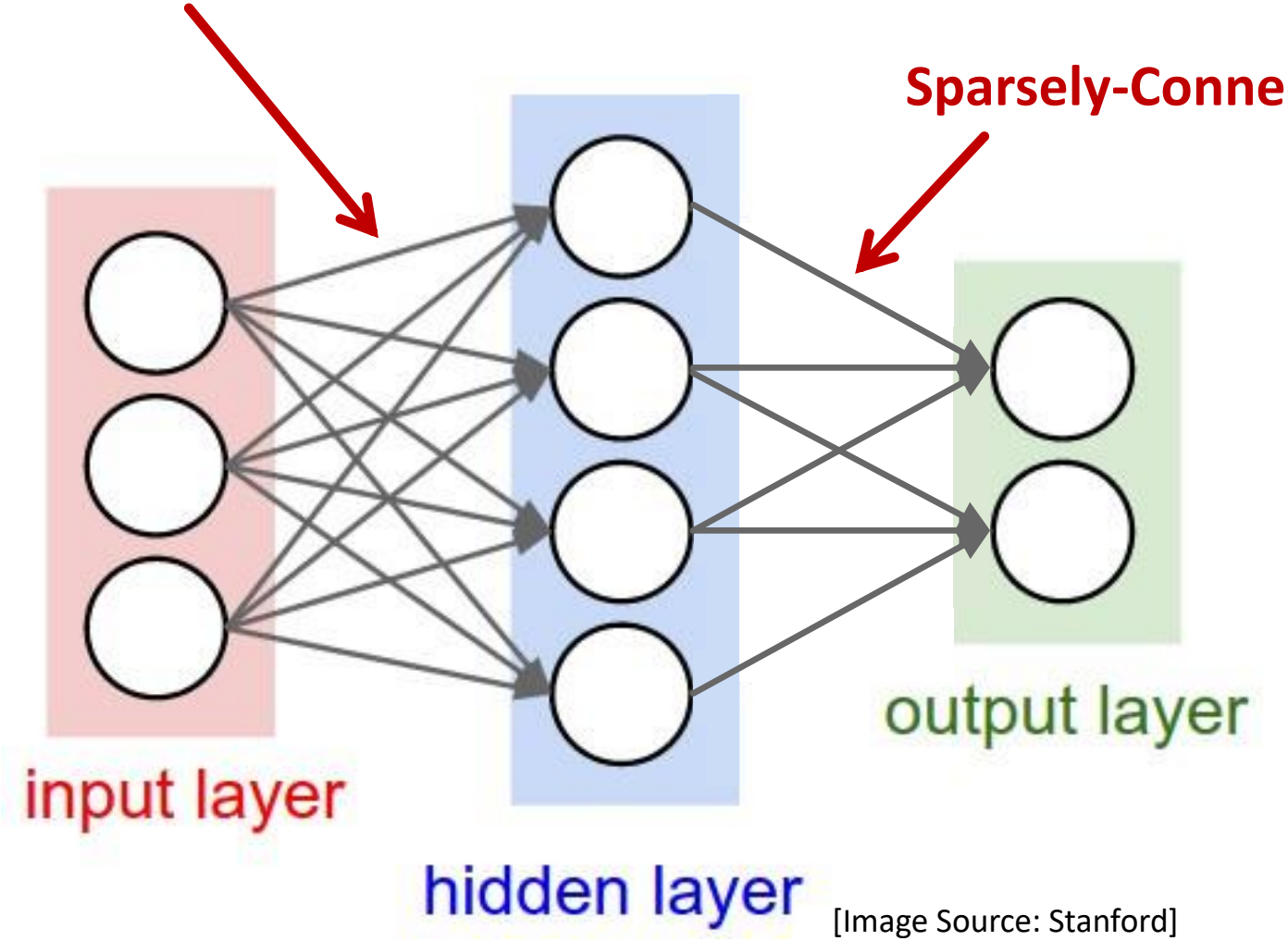
"3-layer Neural Net", or
"2-hidden-layer Neural Net"

[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

NEURAL NETWORK Terminology

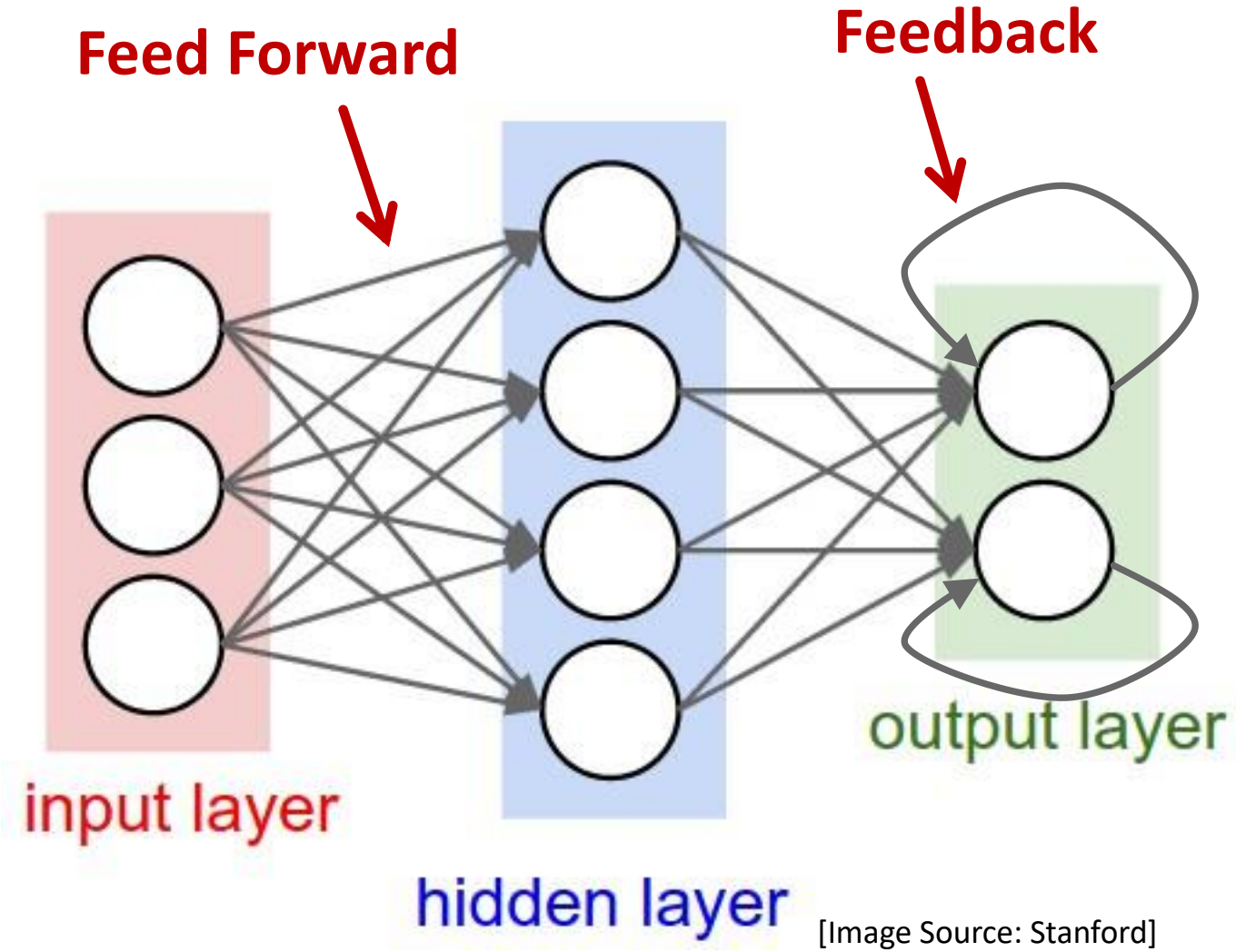
Fully-Connected: all input neurons connected to all output neurons



[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

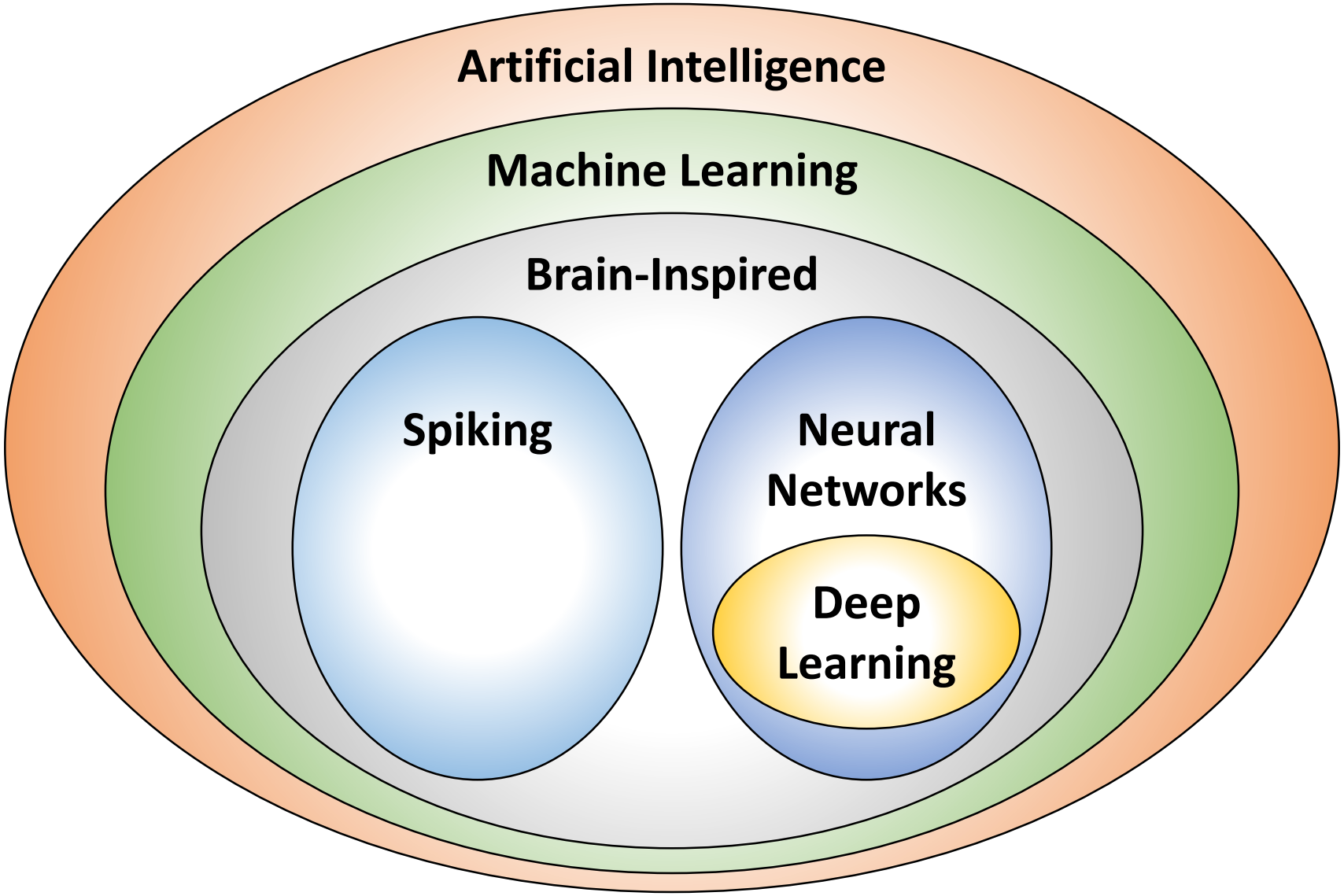
NEURAL NETWORK Terminology



[Image Source: Stanford]

Slide Courtesy: Joel Emer and Vivienne Sze

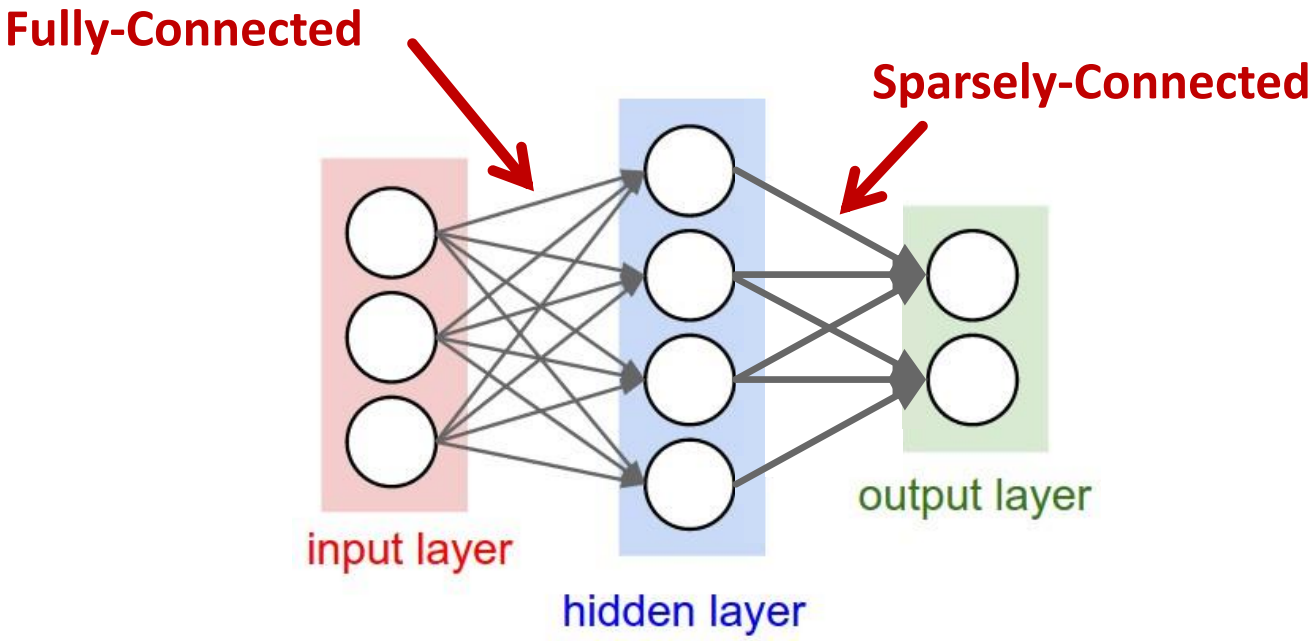
Deep Learning



Slide Courtesy: Joel Emer and Vivienne Sze

Popular Types of DEEP NEURAL NETWORKS

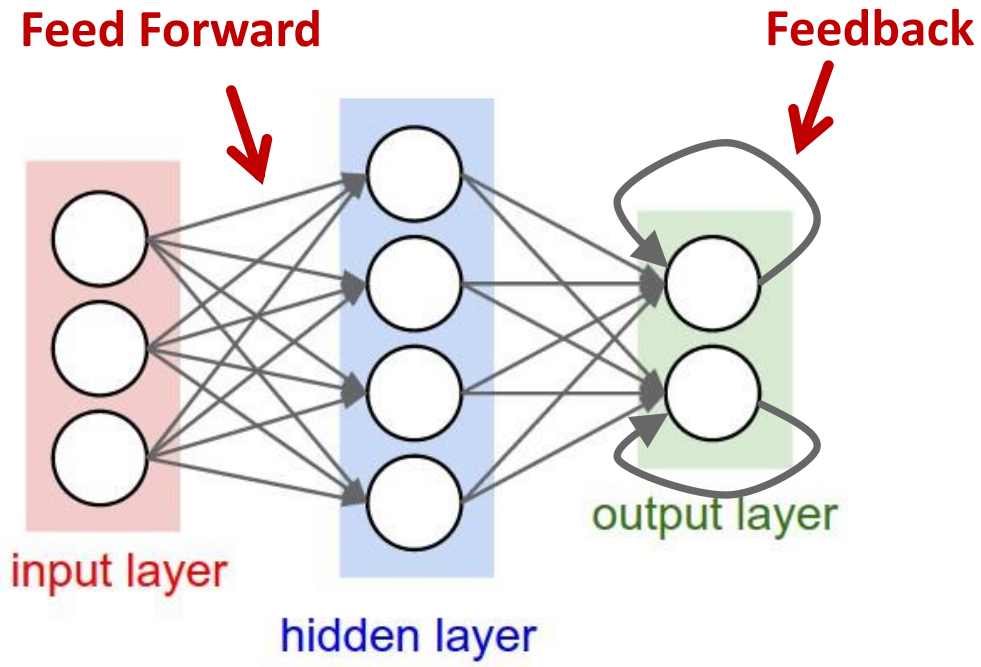
- Fully-Connected NN
  - feed forward, a.k.a. multilayer perceptron (MLP)
- Convolutional NN (CNN)
  - feed forward, sparsely-connected w/ weight sharing



Slide Courtesy: Joel Emer and Vivienne Sze

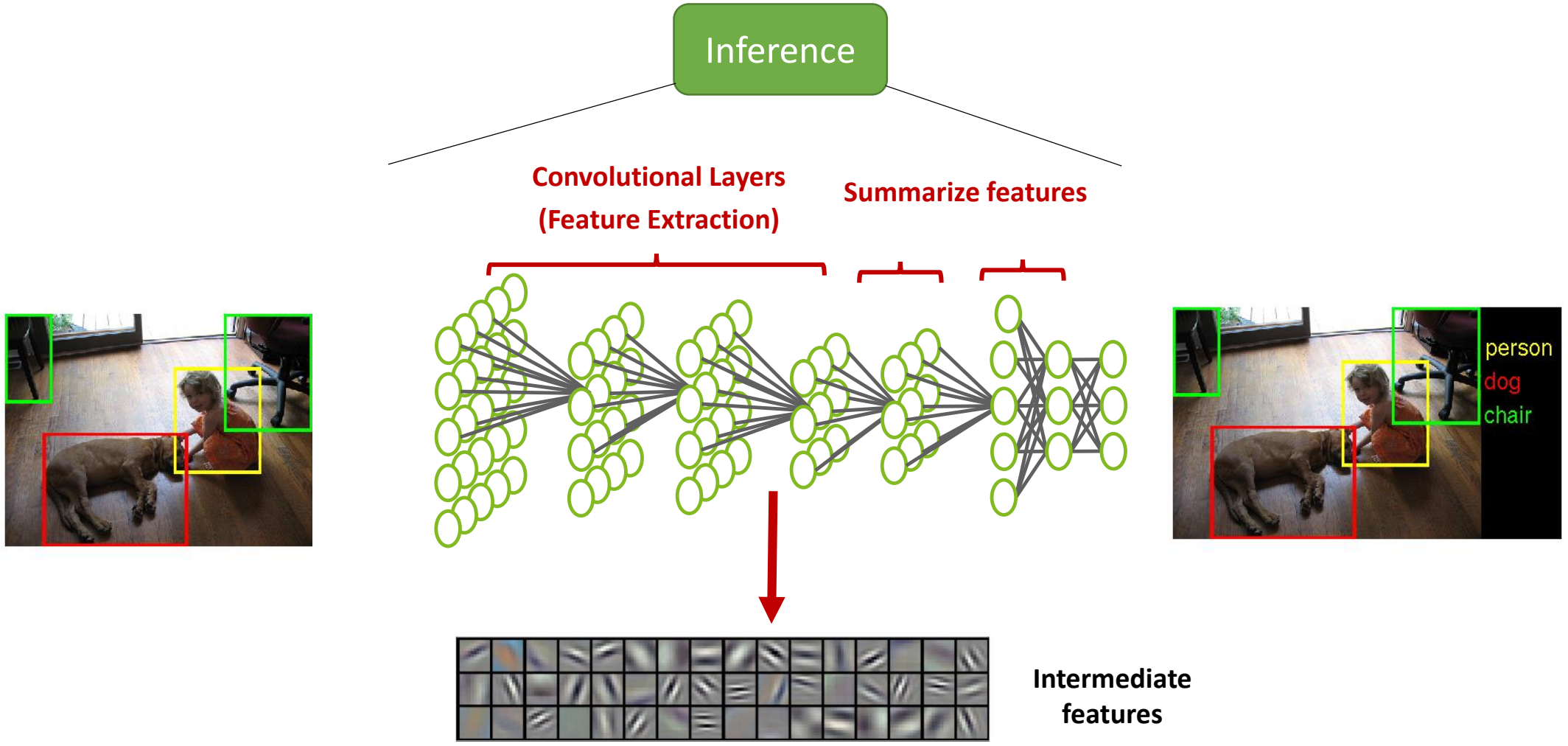
Popular Types of DEEP NEURAL NETWORKS

- Recurrent NN (RNN)
  - feedback
- Long Short-Term Memory (LSTM)
  - feedback + storage

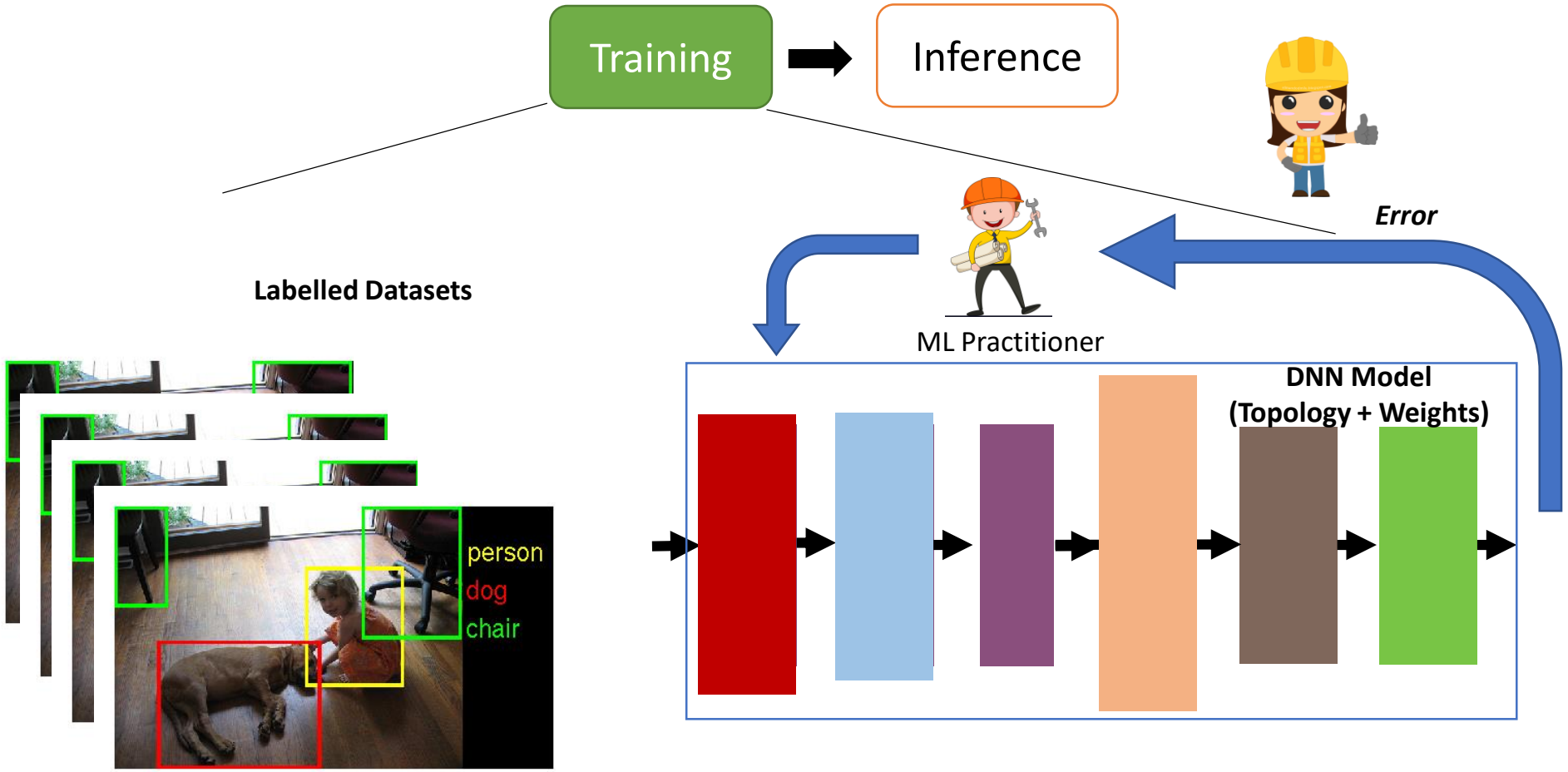


Slide Courtesy: Joel Emer and Vivienne Sze

Deep Learning Landscape



Deep Learning Landscape



Why is deep learning hot now?

Big Data Availability

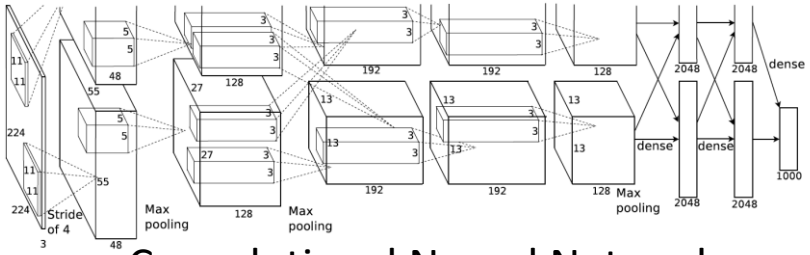
New ML Techniques

GPU Acceleration

facebook **350M** images uploaded per day

YouTube **300 hours** of video uploaded every minute

Walmart **2.5 Petabytes** of customer data hourly



Convolutional Neural Network

Slide Courtesy: Joel Emer and Vivienne Sze

How deep learning started: ImageNet Challenge

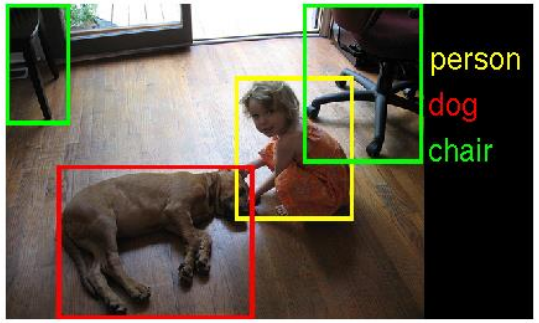
IMAGENET

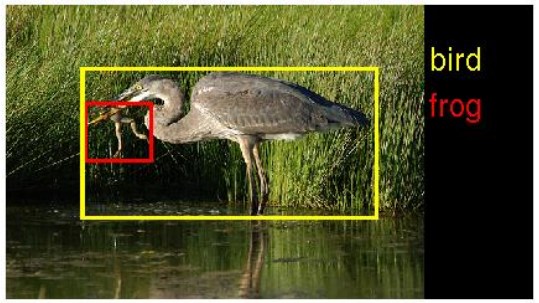
Image Classification Task:

1.2M training images • 1000 object categories

Object Detection Task:

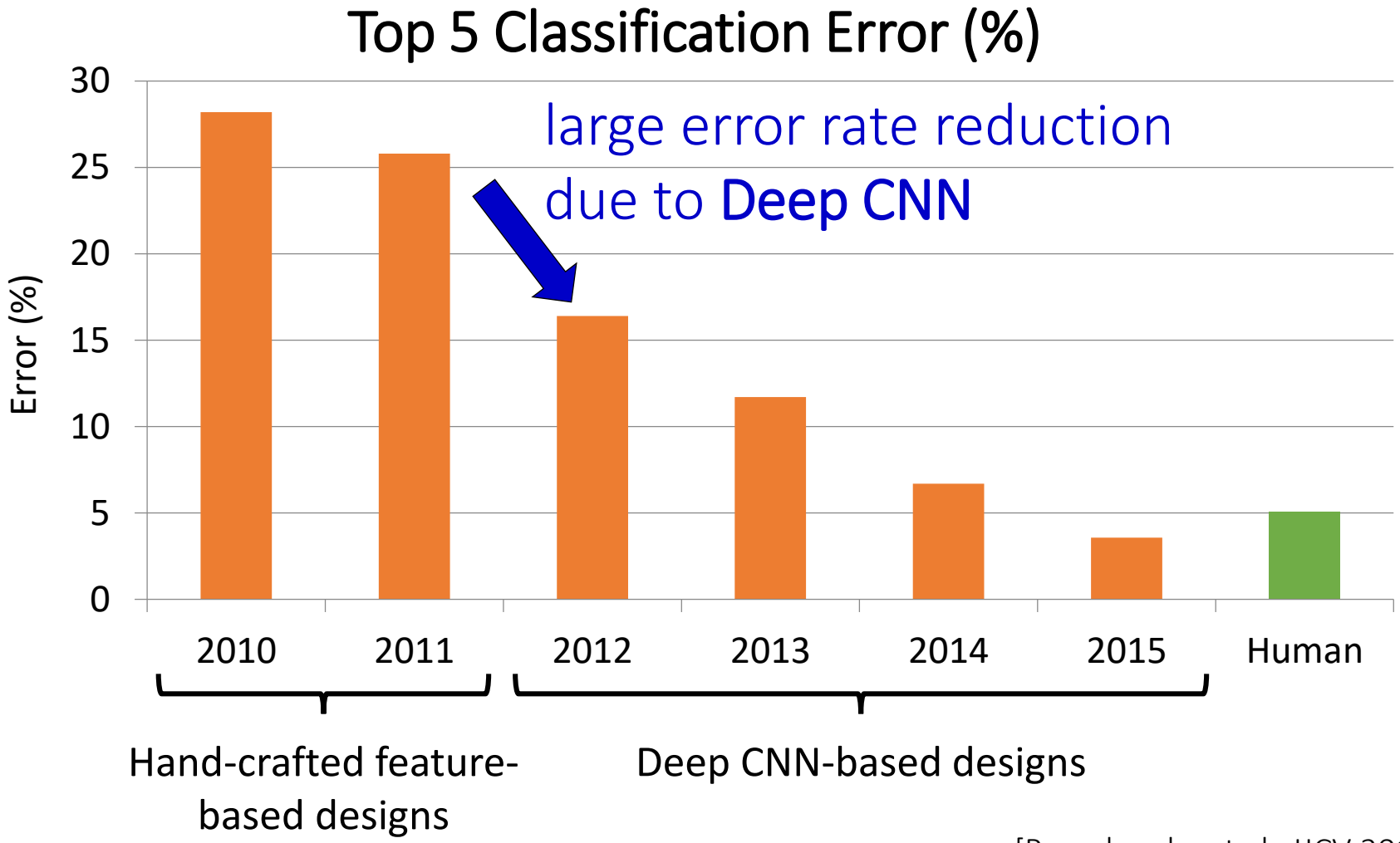
456k training images • ***200 object categories***





Slide Courtesy: Joel Emer and Vivienne Sze

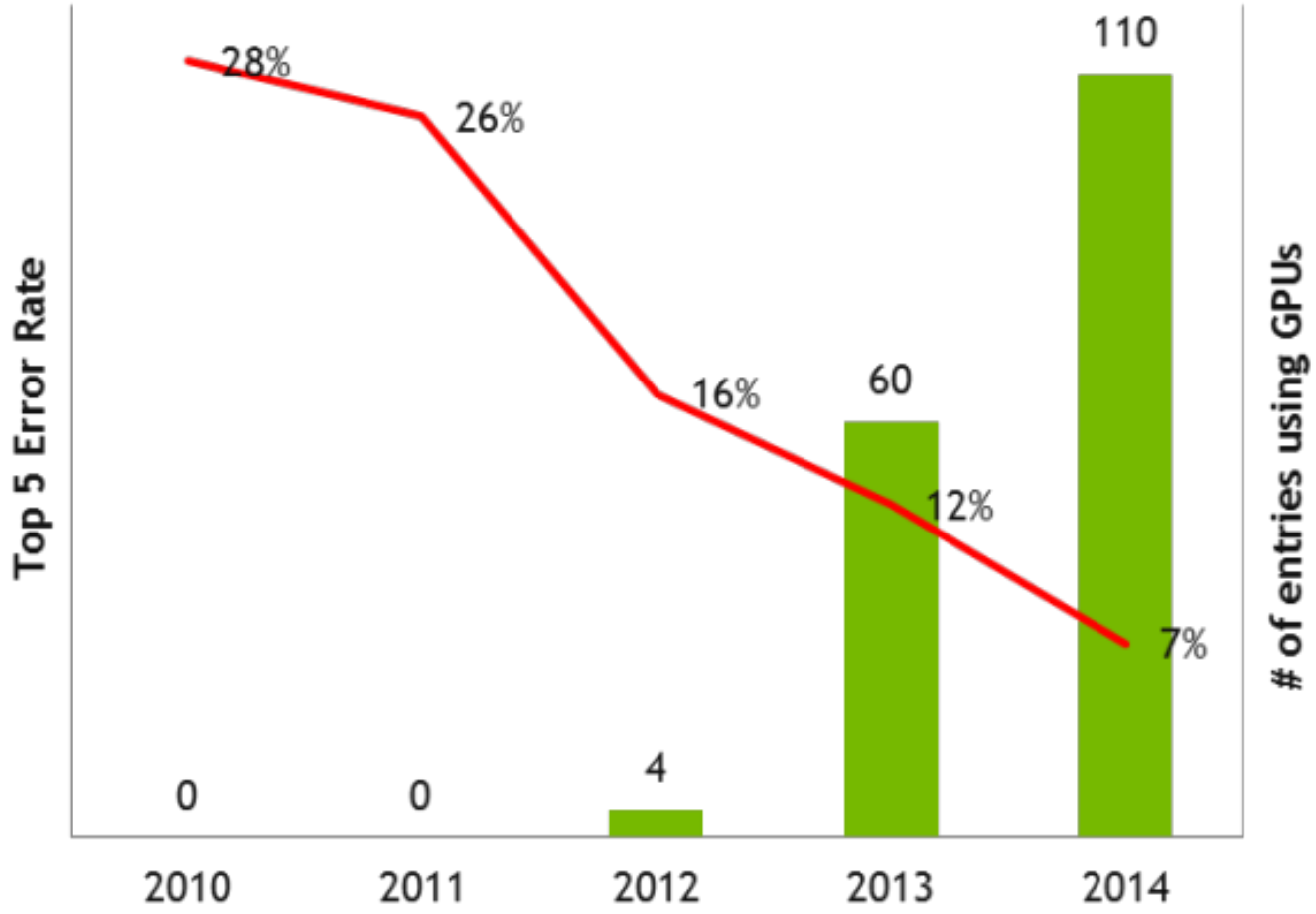
ImageNet: Image Classification Task



[Russakovsky et al., IJCV 2015]

Slide Courtesy: Joel Emer and Vivienne Sze

GPU Usage for ImageNet Challenge



Slide Courtesy: Joel Emer and Vivienne Sze

Mature Applications

- **Image**
  - Classification: image to object class
  - Recognition: same as classification (except for faces)
  - Detection: assigning bounding boxes to objects
  - Segmentation: assigning object class to every pixel
- **Speech & Language**
  - Speech Recognition: audio to text
  - Translation
  - Natural Language Processing: text to meaning
  - Audio Generation: text to audio
- **Games**

Slide Courtesy: Joel Emer and Vivienne Sze

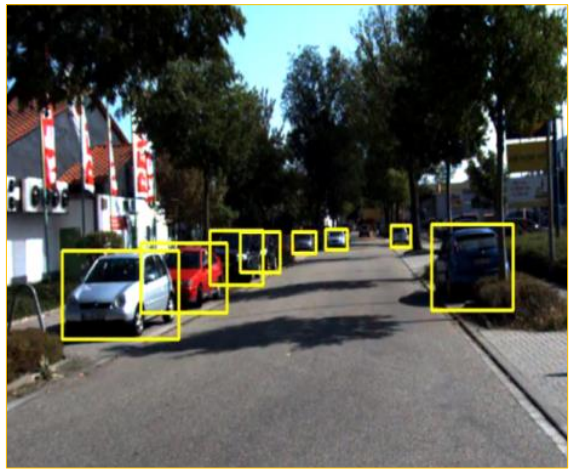
Emerging Applications

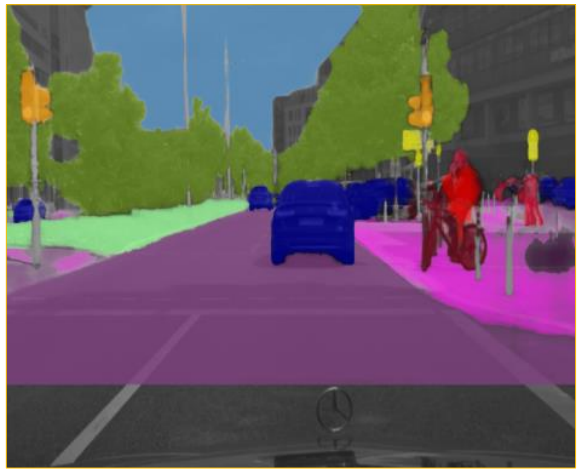
- **Medical** (Cancer Detection, Pre-Natal)
- **Finance** (Trading, Energy Forecasting, Risk)
- **Infrastructure** (Structure Safety and Traffic)
- Weather Forecasting and Event Detection

http://www.nextplatform.com/2016/09/14/next-wave-deep-learning-applications/

Slide Courtesy: Joel Emer and Vivienne Sze

Deep Learning for Self-driving Cars

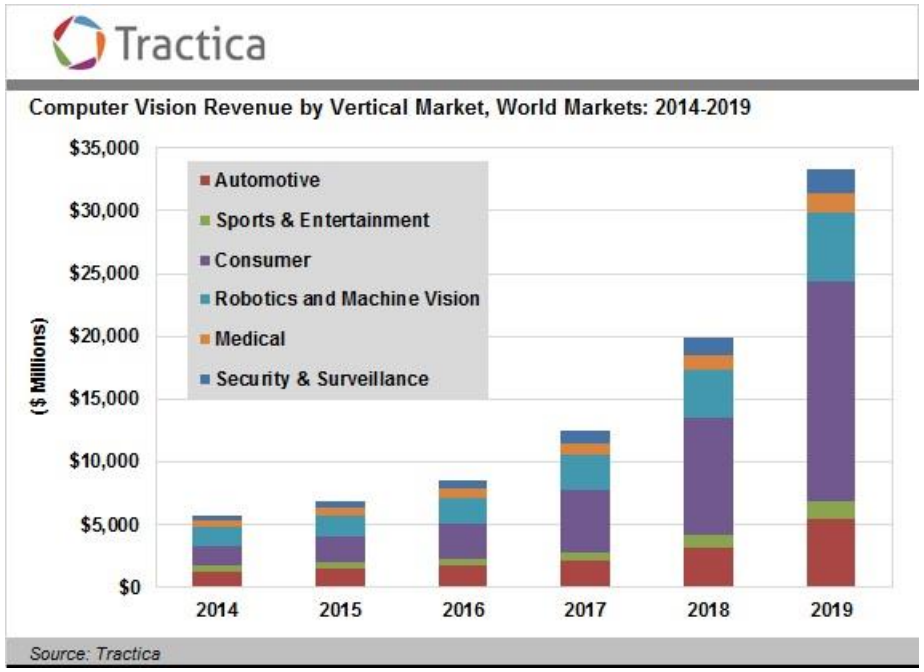





Slide Courtesy: Joel Emer and Vivienne Sze

Opportunities

$500B Market over 10 Years!



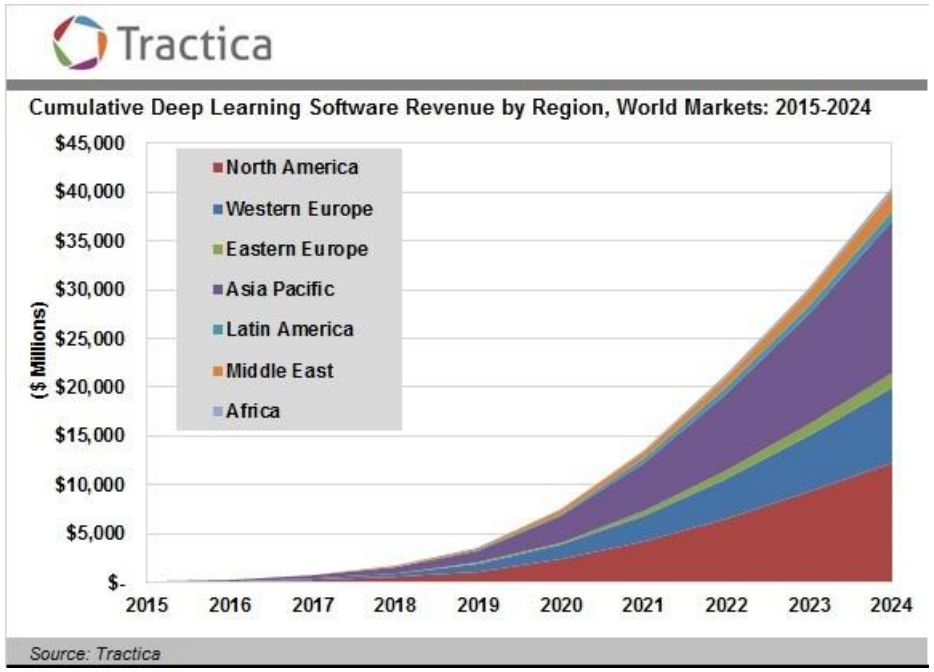


Image Source: Tractica

Slide Courtesy: Joel Emer and Vivienne Sze

Opportunities

From EE Times – September 27, 2016

”Today the job of training machine learning models is limited by compute, if we had faster processors we’d run bigger models…in practice we train on a reasonable subset of data that can finish in a matter of months. We could use improvements of several orders of magnitude – 100x or greater.”

- Greg Diamos, Senior Researcher, (formerly) SVAIL, Baidu

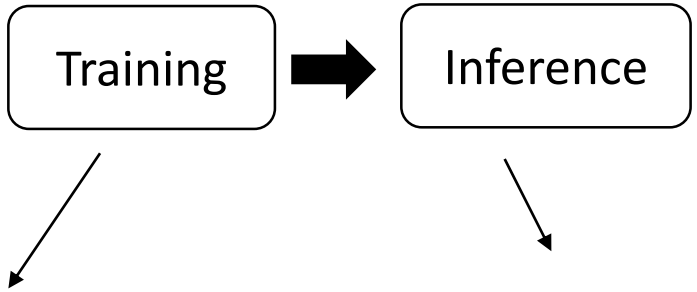
Slide Courtesy: Joel Emer and Vivienne Sze

COMPUTATION Platforms

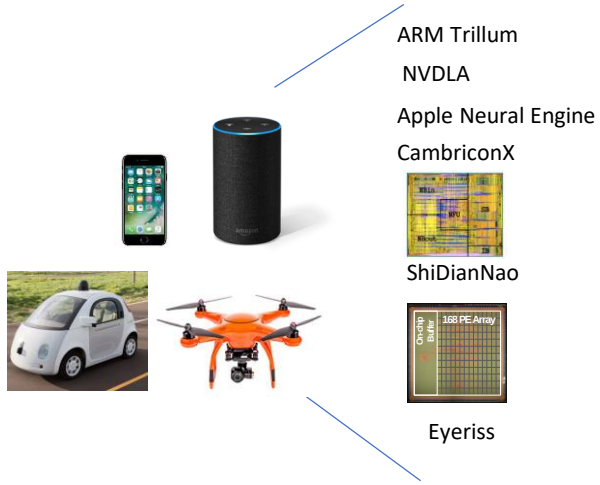
- CPU
  - Intel, ARM, AMD…
- GPU
  - NVIDIA, AMD…
- Fine Grained Reconfigurable (FPGA)
  - Microsoft BrainWave
- Coarse Grained Programmable/Reconfigurable
  - Wave Computing, Plasticine, Graphcore…
- Application Specific
  - Neuflow, *DianNao, Eyeriss, Cnvlutin, SCNN, TPU, …

Slide Courtesy: Joel Emer and Vivienne Sze

Computation Platforms

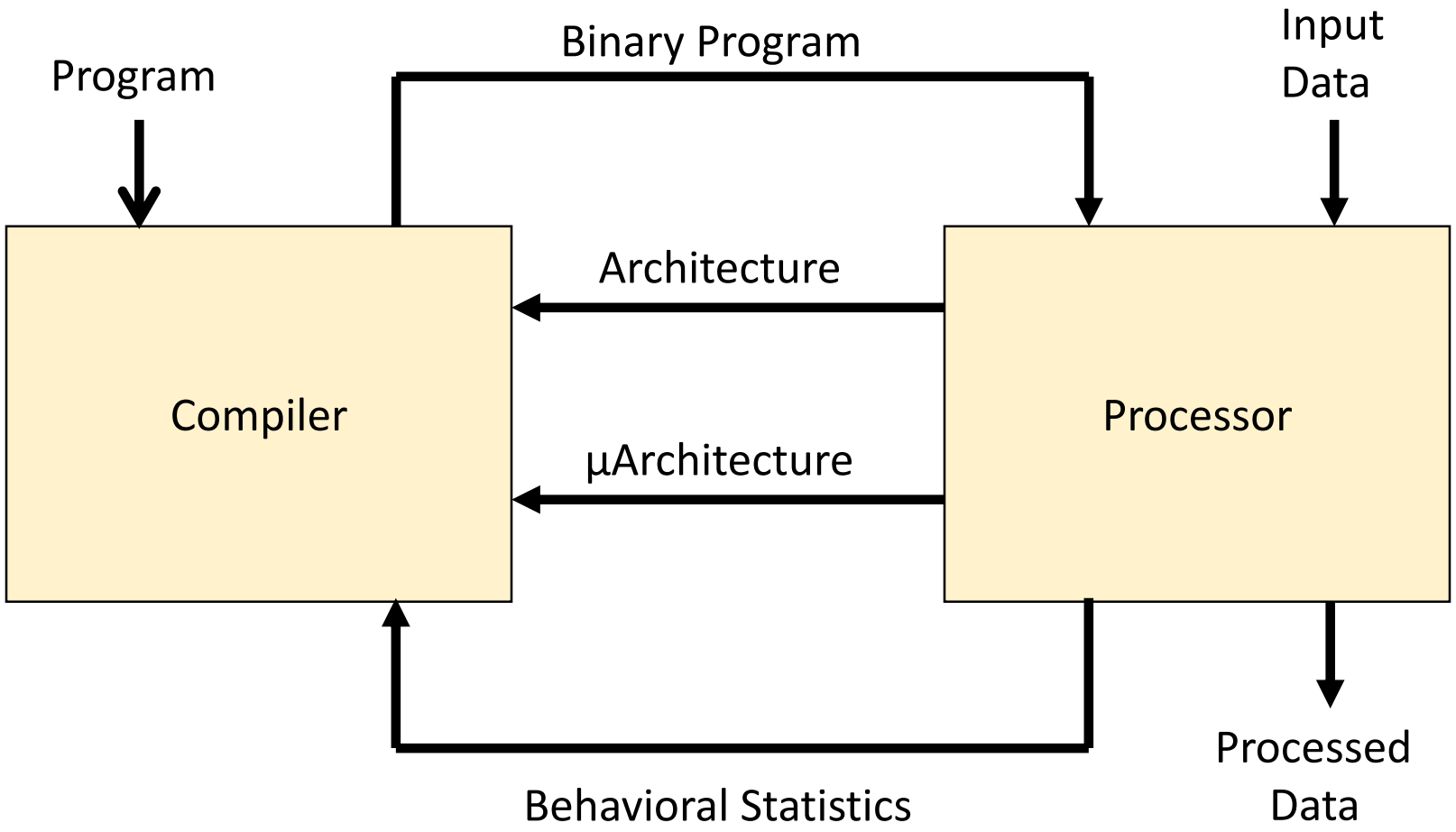


HPC cluster
(CPU+GPU+FPGA)



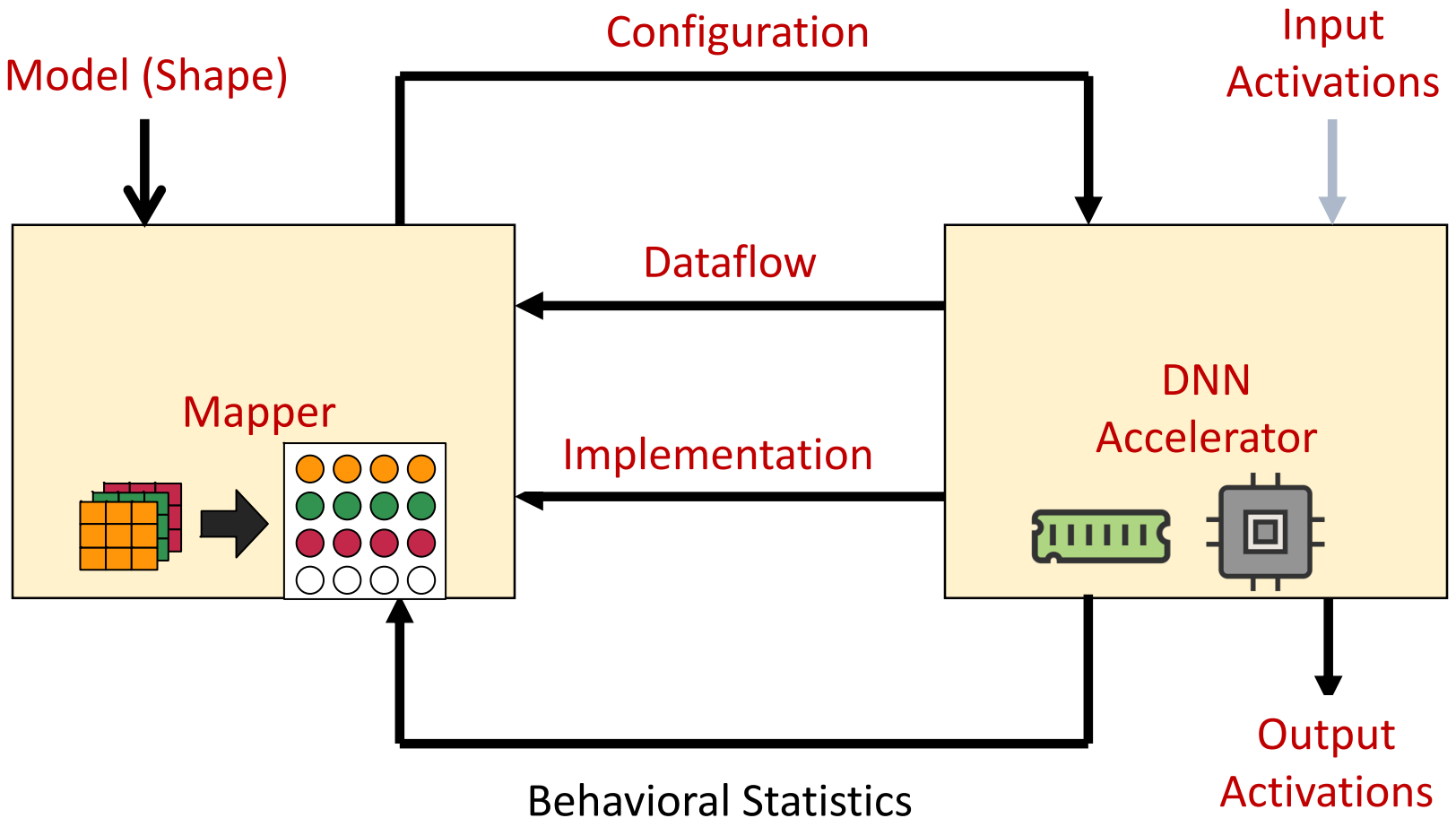
Accelerators
(ASIC)

CPU Compute Model



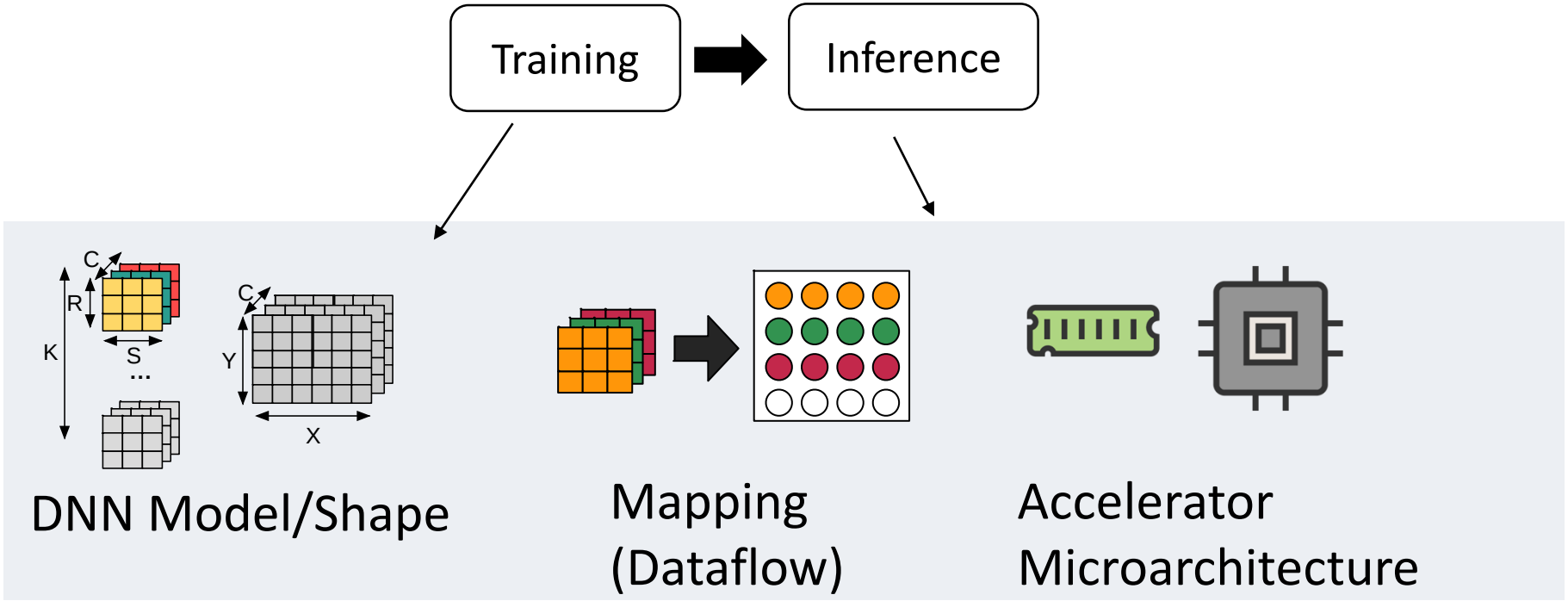
Slide Courtesy: Joel Emer and Vivienne Sze

DNN Compute Model



Slide Courtesy: Joel Emer and Vivienne Sze

Challenges in Design and Deployment



DNN Timeline

- 1940s: Neural networks were proposed
- 1960s: Deep neural networks were proposed
- 1989: Neural network for recognizing digits (LeNet)
- 1990s: Hardware for shallow neural nets
  - Example: Intel ETANN (1992)
- 2011: Breakthrough DNN-based speech recognition
  - Microsoft real-time speech translation
- 2012: DNNs for vision supplanting traditional ML
  - AlexNet for image classification
- 2014+: Rise of DNN accelerator research
  - Examples: Neuflow, DianNao, etc.

Slide Courtesy: Joel Emer and Vivienne Sze