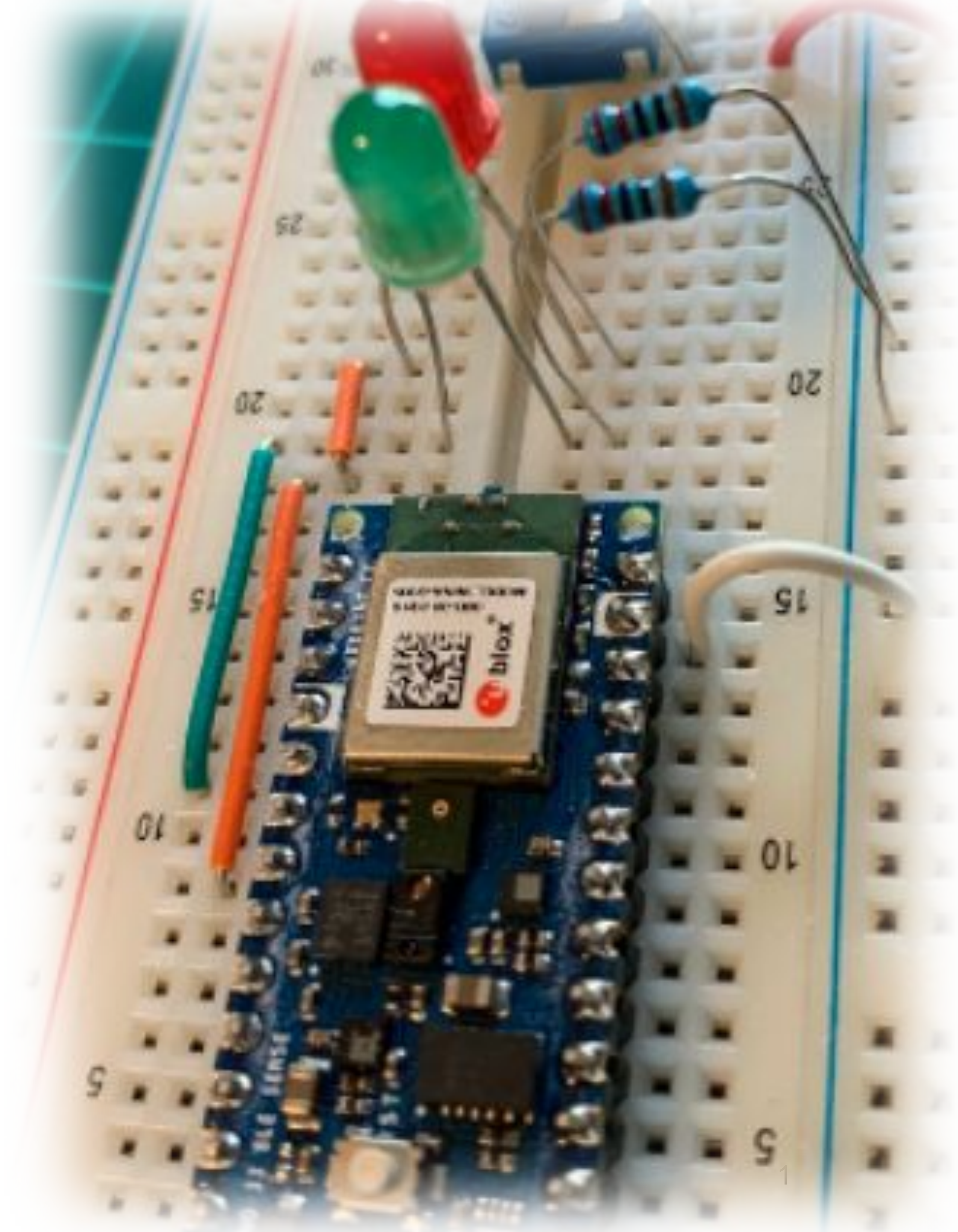
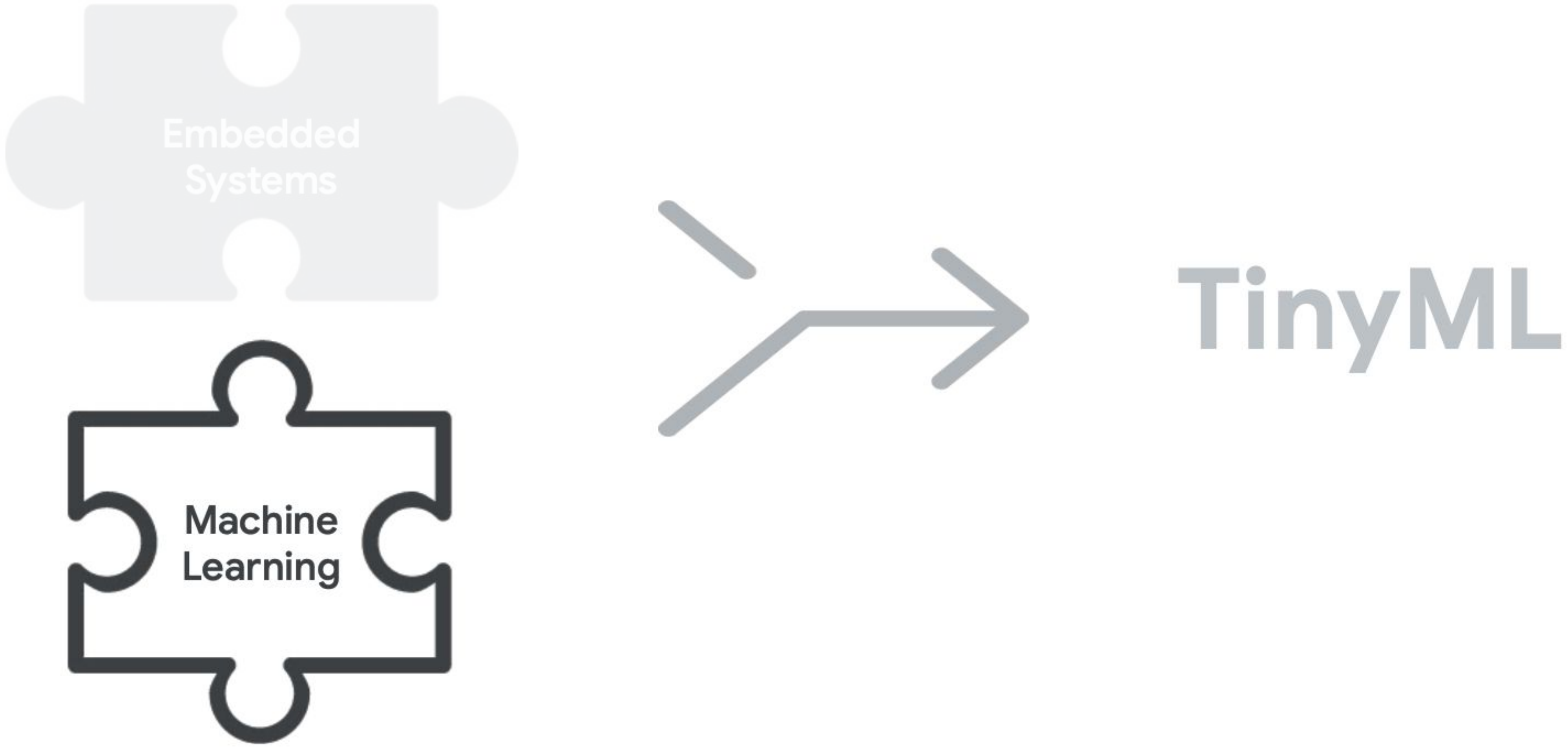
IESTI01 – TinyML

Embedded Machine Learning

4. TinyML Challenges:
   - Machine Learning

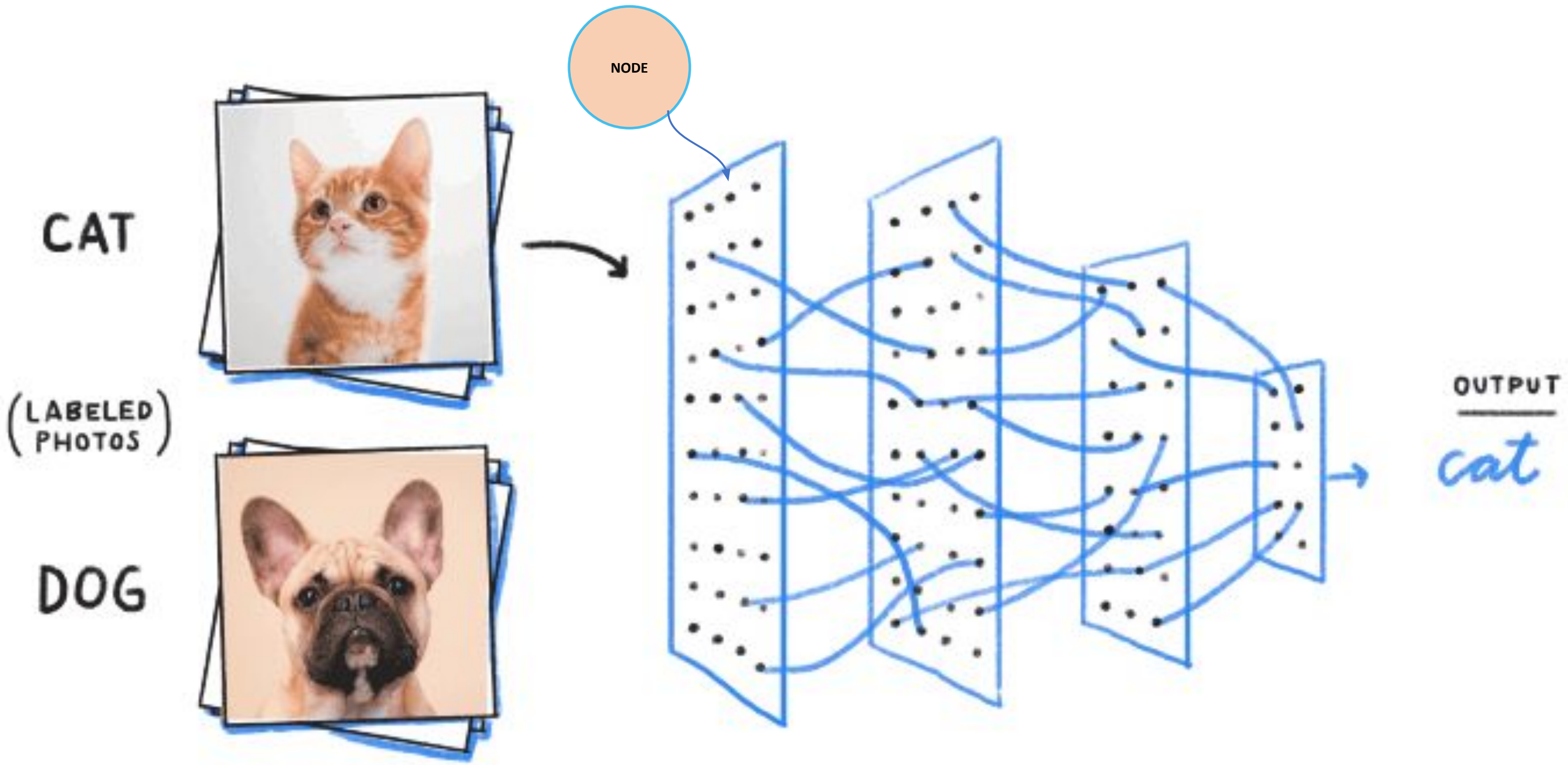
Prof. Marcelo Rovai
UNIFEI



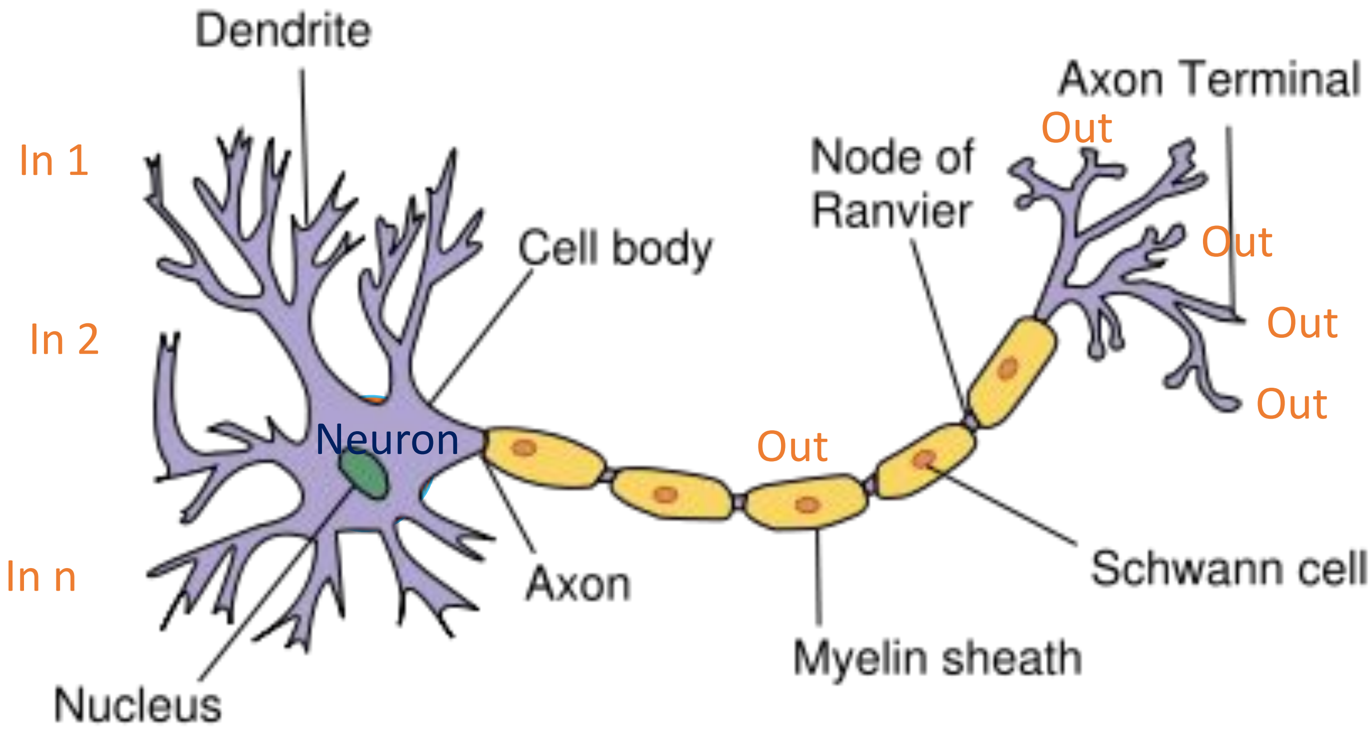

Embedded
Systems
Machine
Learning
TinyML

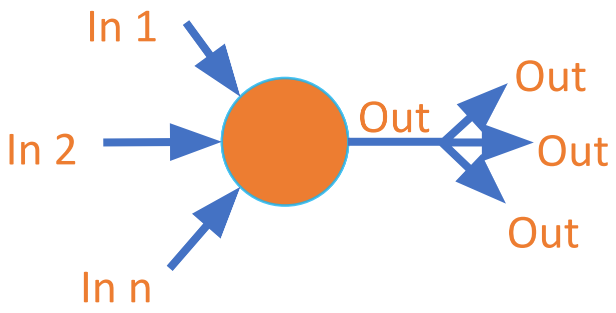
(Deep) Machine Learning

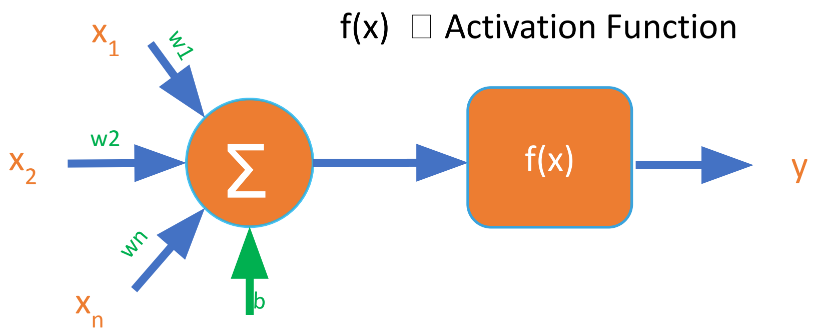
Deep Learning: Subset of Machine Learning in which multilayered neural networks learn from vast amounts of data



Neuron (Perceptron)



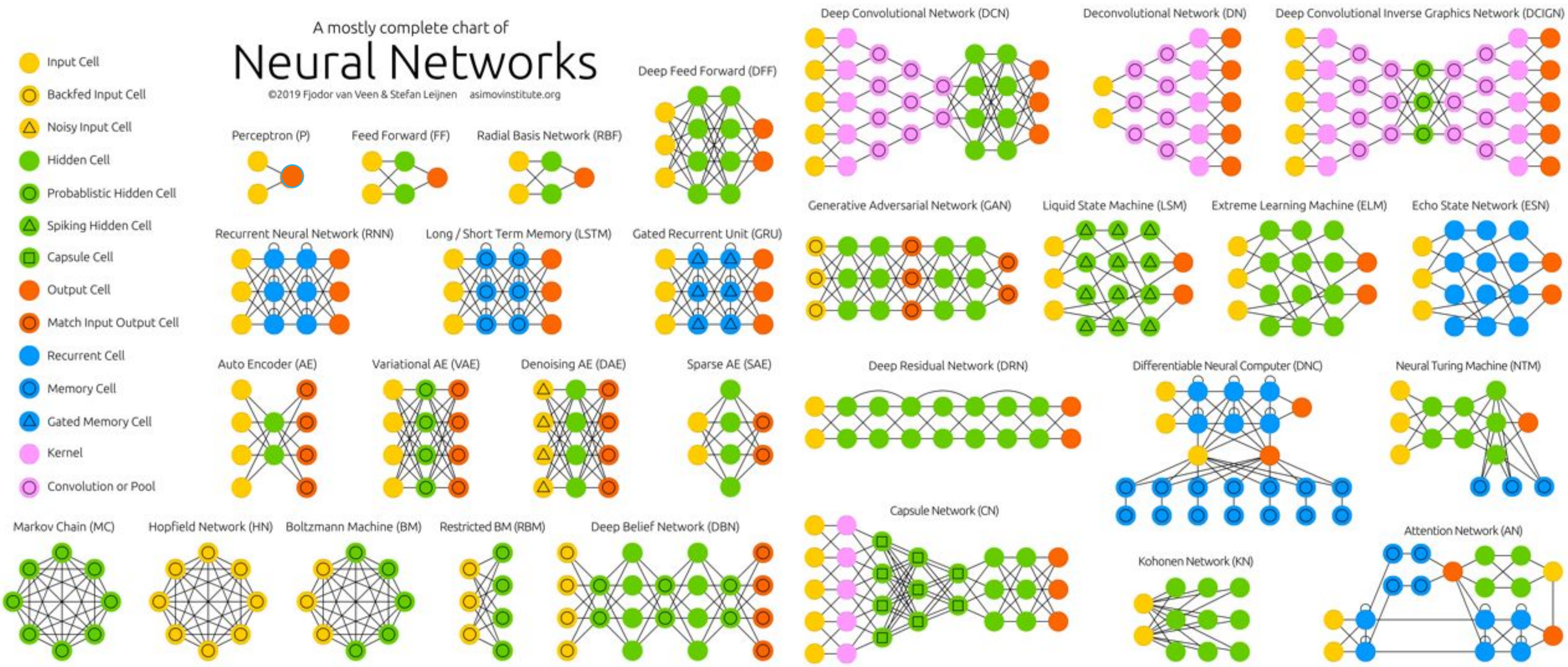




Parameters

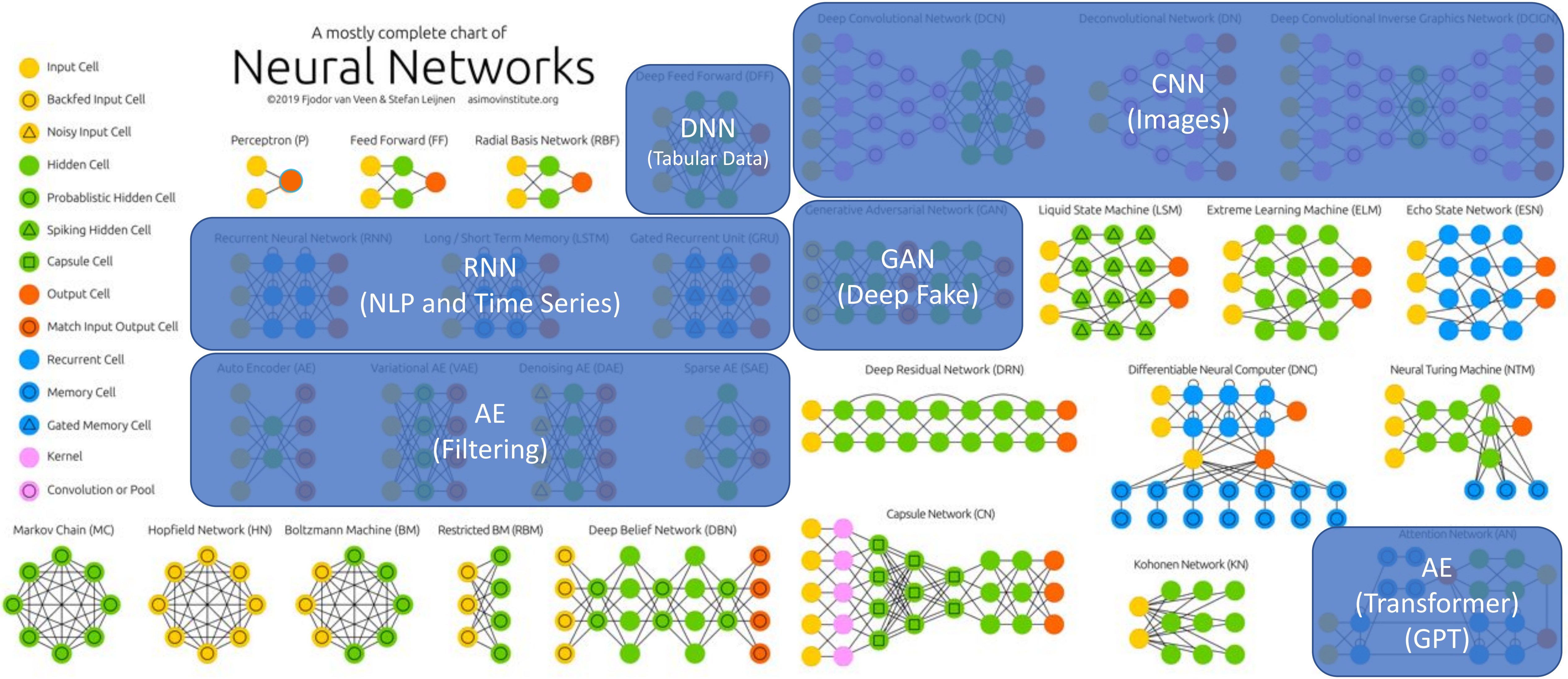
$$y = f\left(\sum_{i=1}^{n} x_i w_i + b\right)$$

The Neural Network Model Architecture



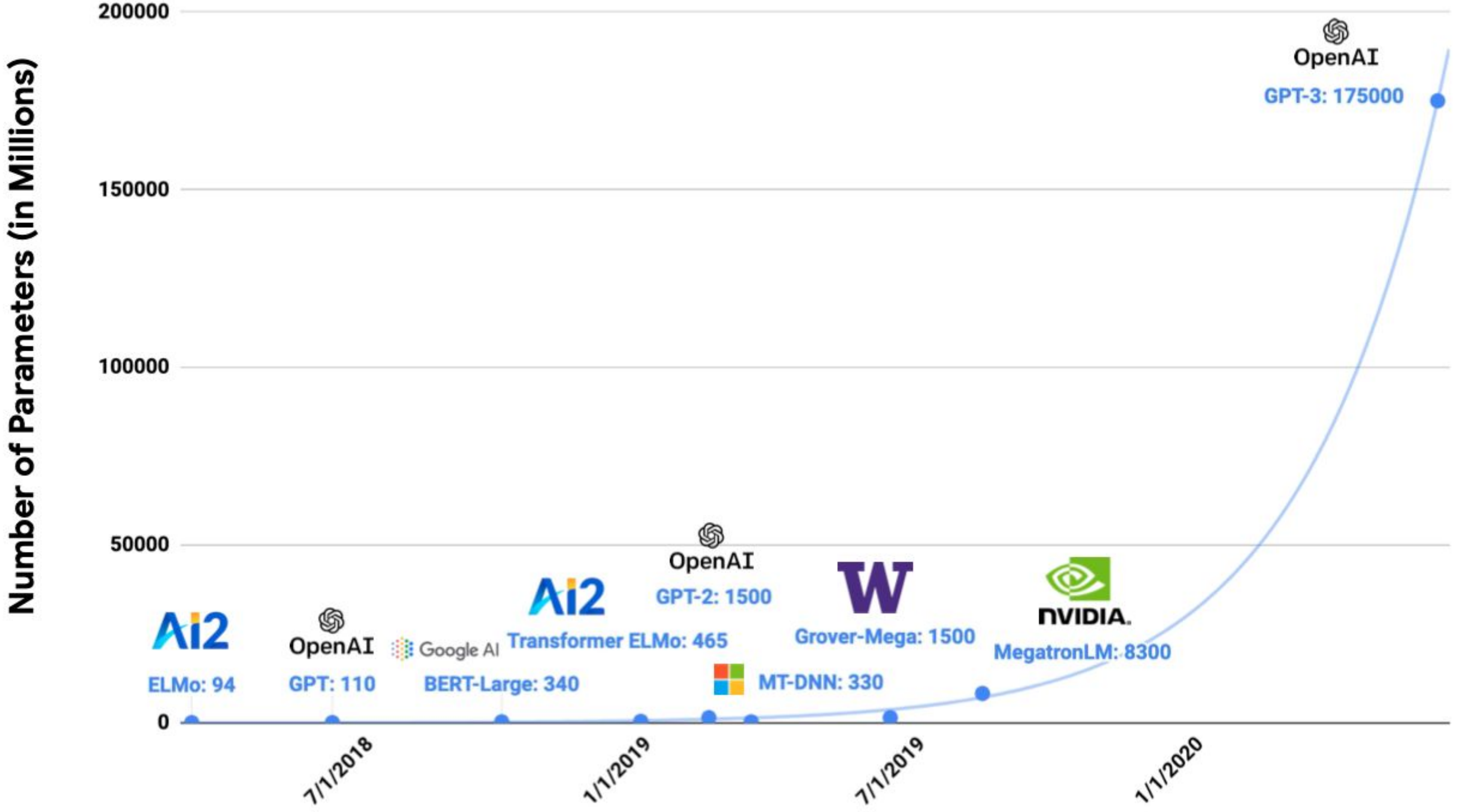
https://www.asimovinstitute.org/neural-network-zoo/

The Neural Network Model Architecture

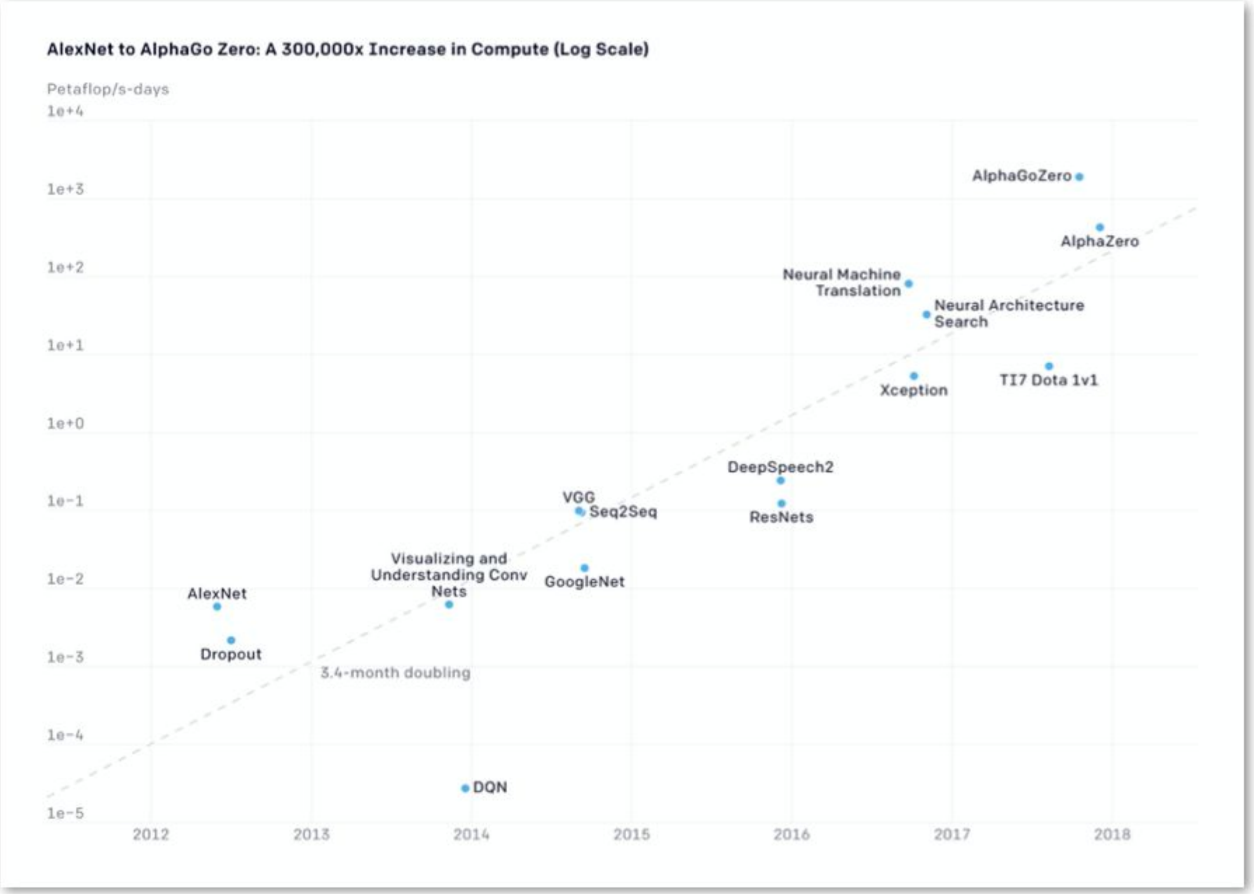


https://www.asimovinstitute.org/neural-network-zoo/

ML Model Size Growth



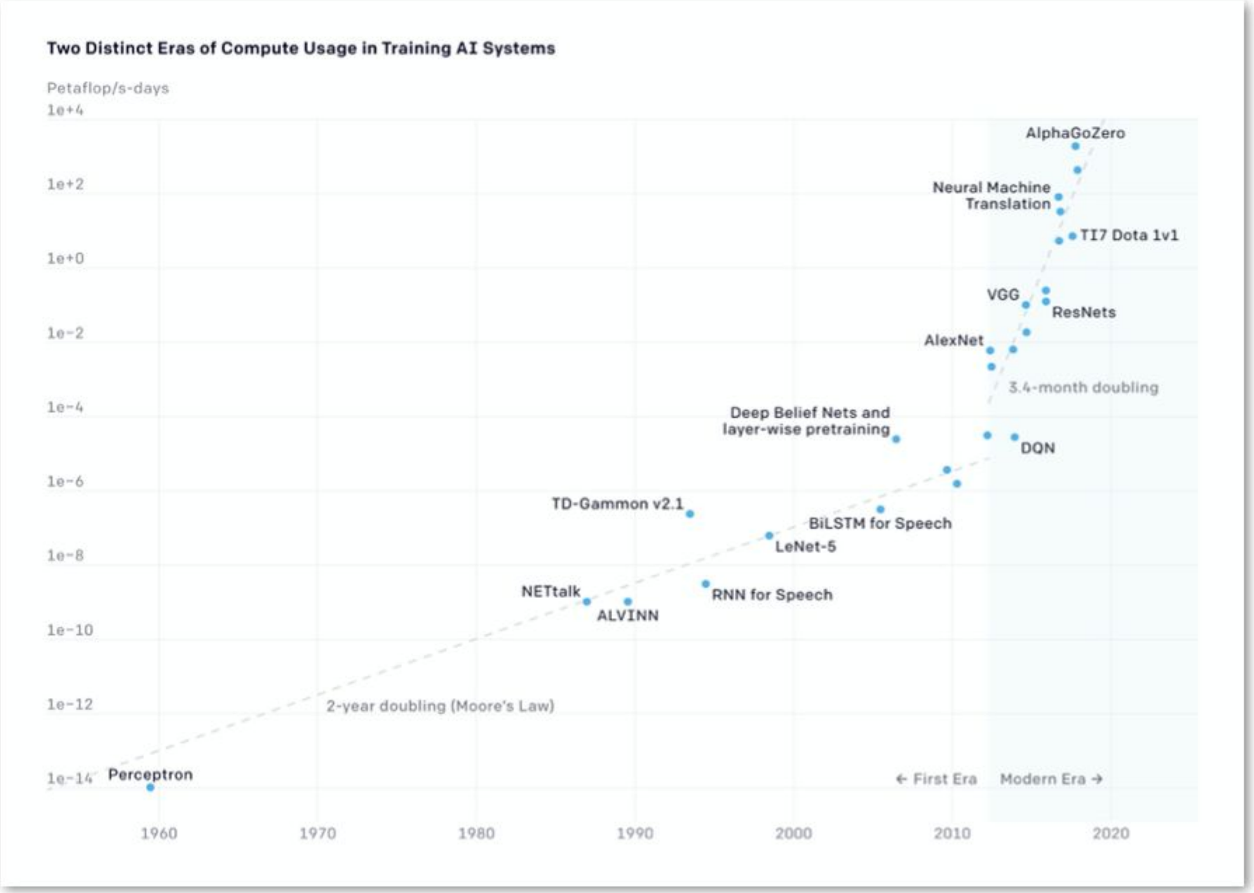
ML Compute Needs (2012 to Present Day)



Source: https://openai.com

In recent years, **computing needs grew by 300,000x** to train the machine learning models that are widely deployed in the industry

ML Compute Needs (from the 1960s)



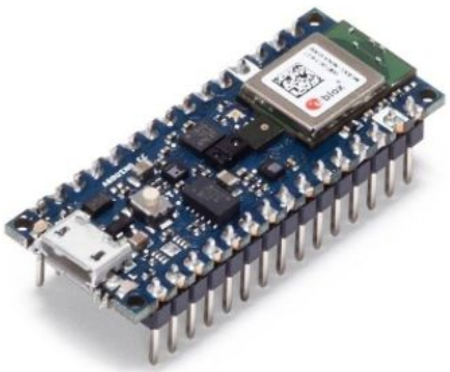
Source: https://openai.com

In recent years, the amount of computing needed has grown remarkably fast.

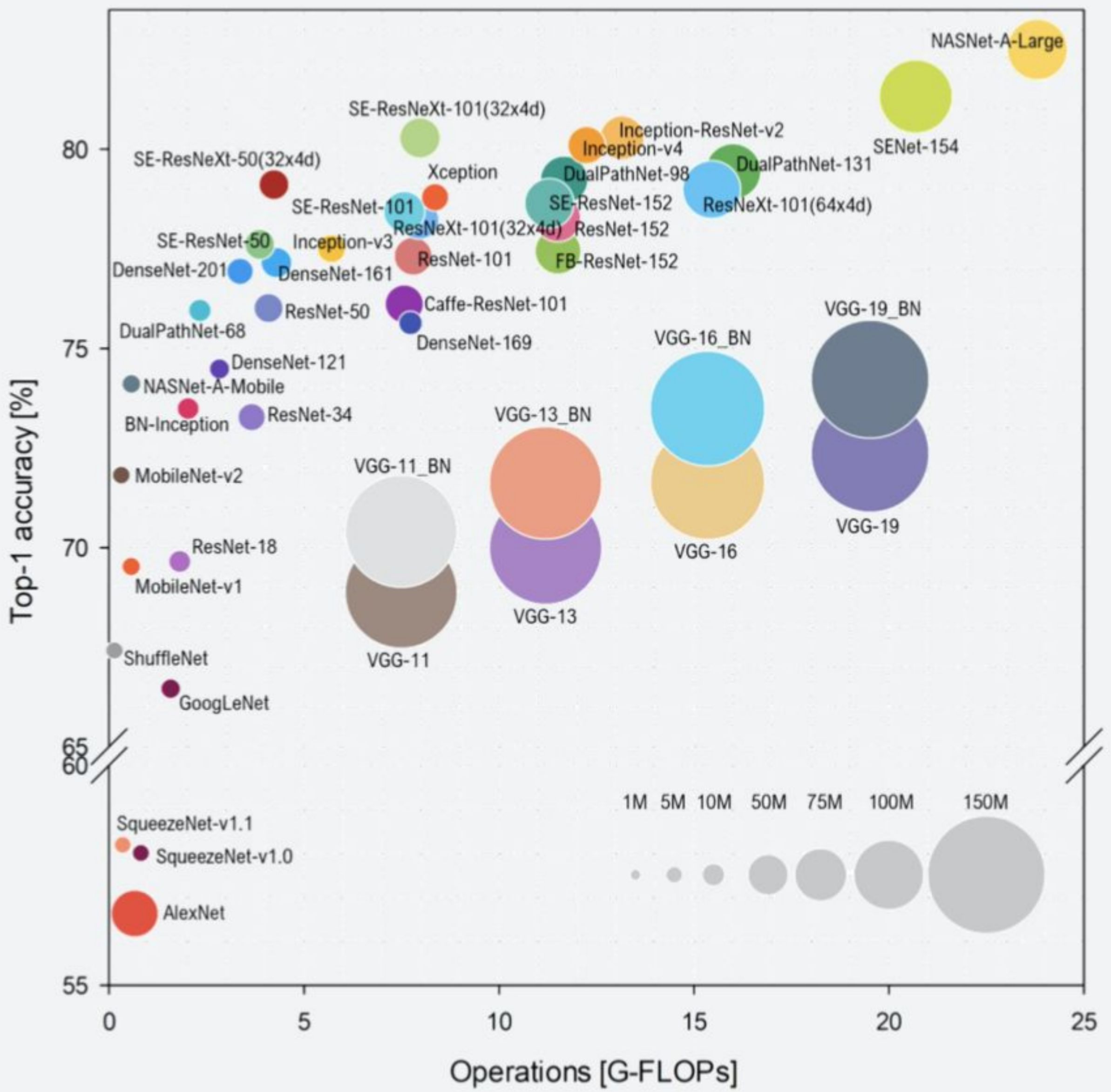
Compute requirements are **doubling nearly every 3 to 4 months**

Cloud TPU

TinyML



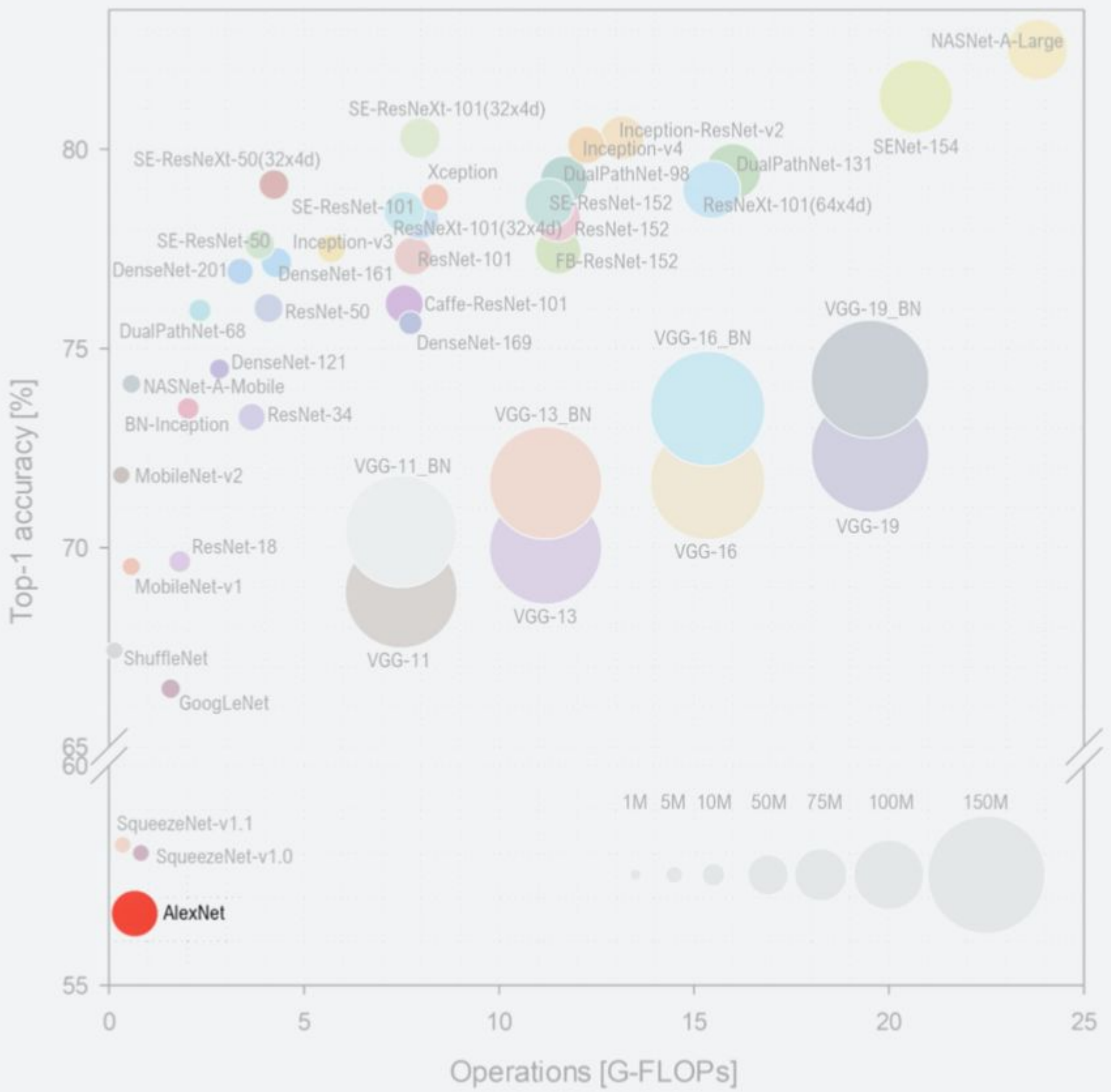
ML Model Evolution



Source: S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018

ML Model Evolution

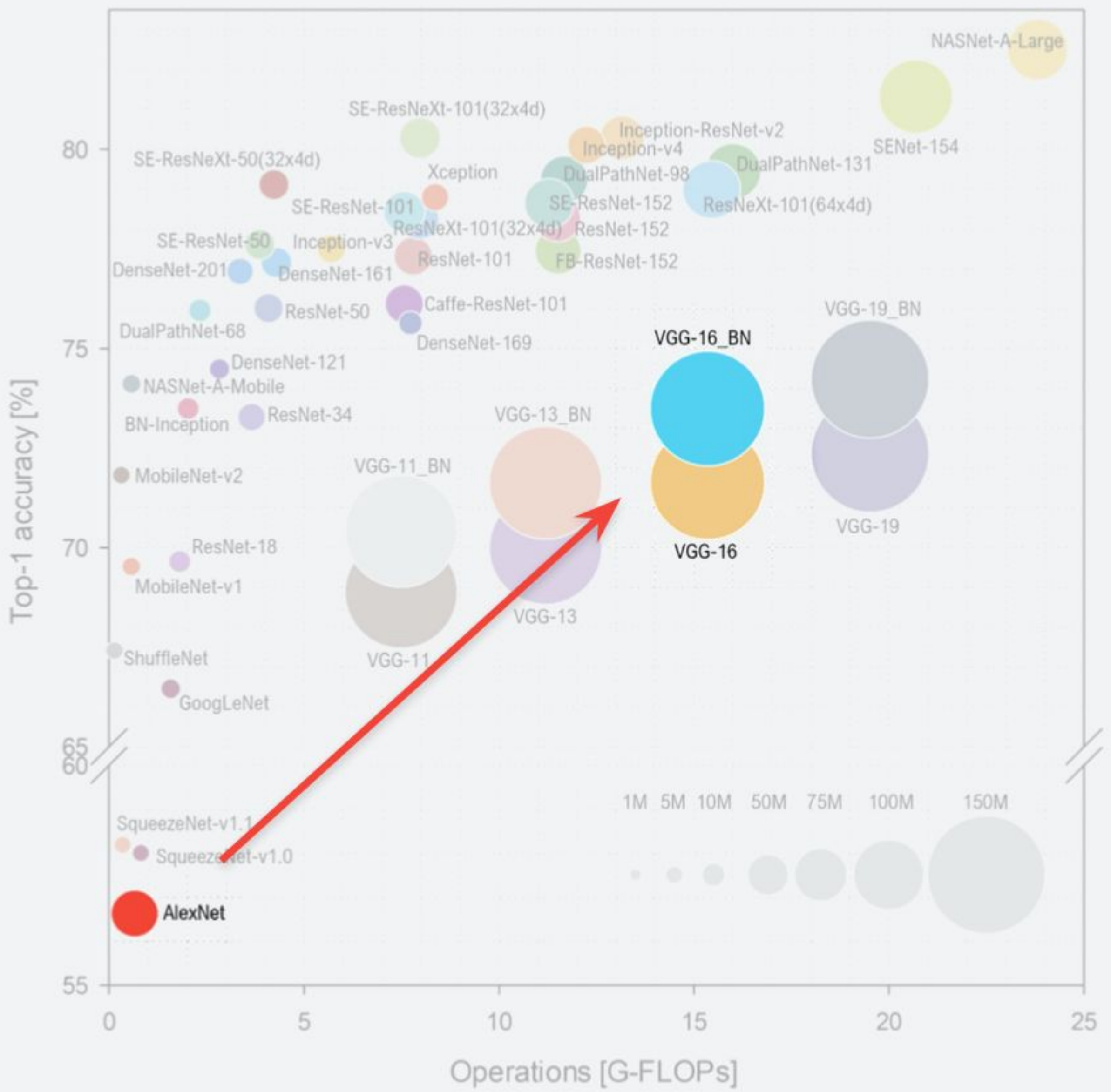
- **AlexNet** (2012)
  - 57.1% accuracy
  - 61MB in size



Source: S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018

ML Model Evolution

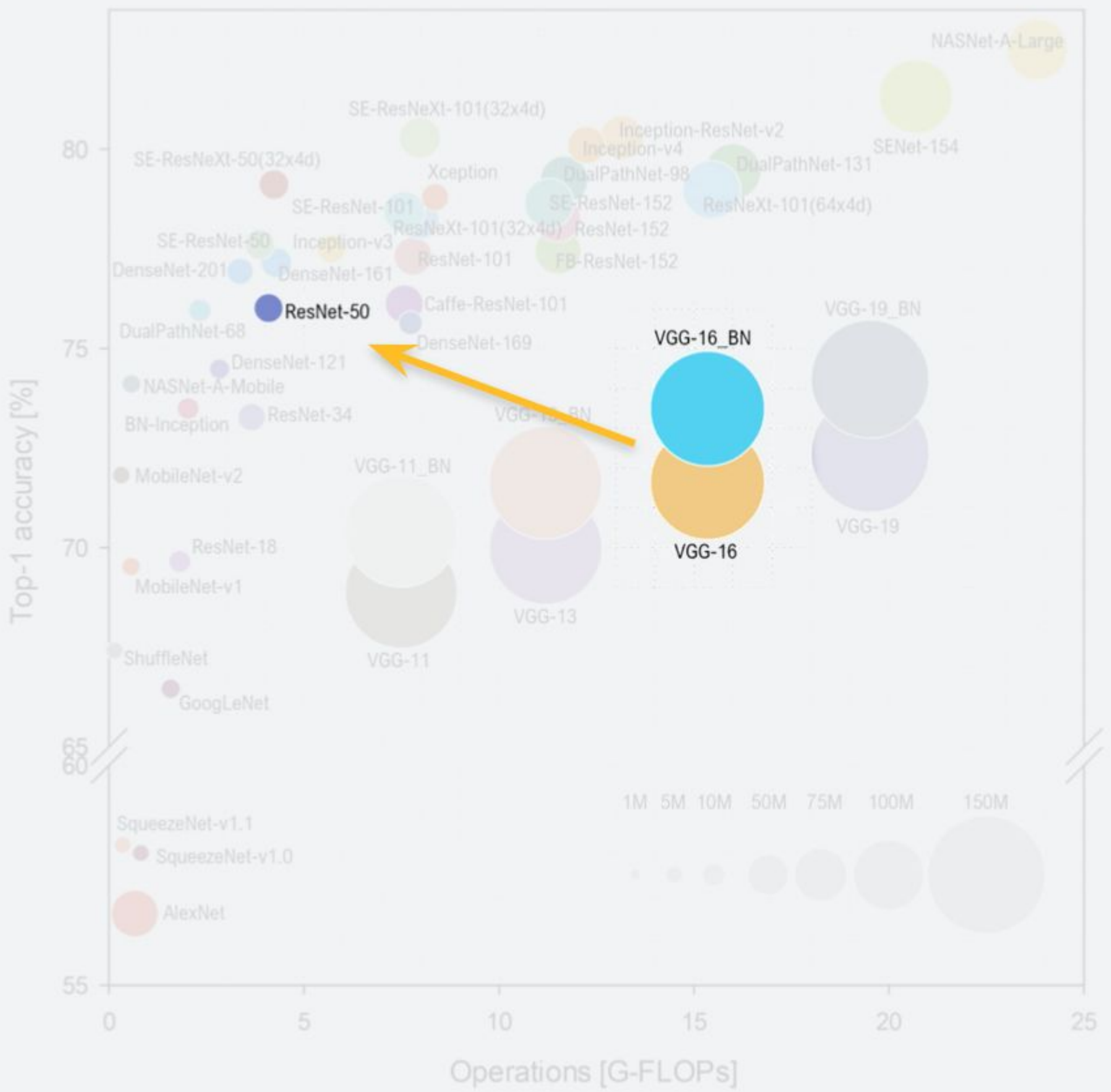
- **VGGNet** (2014) [VGG-16]
  - **71.5%** accuracy
  - **528MB** in size



Source: S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018

ML Model Evolution

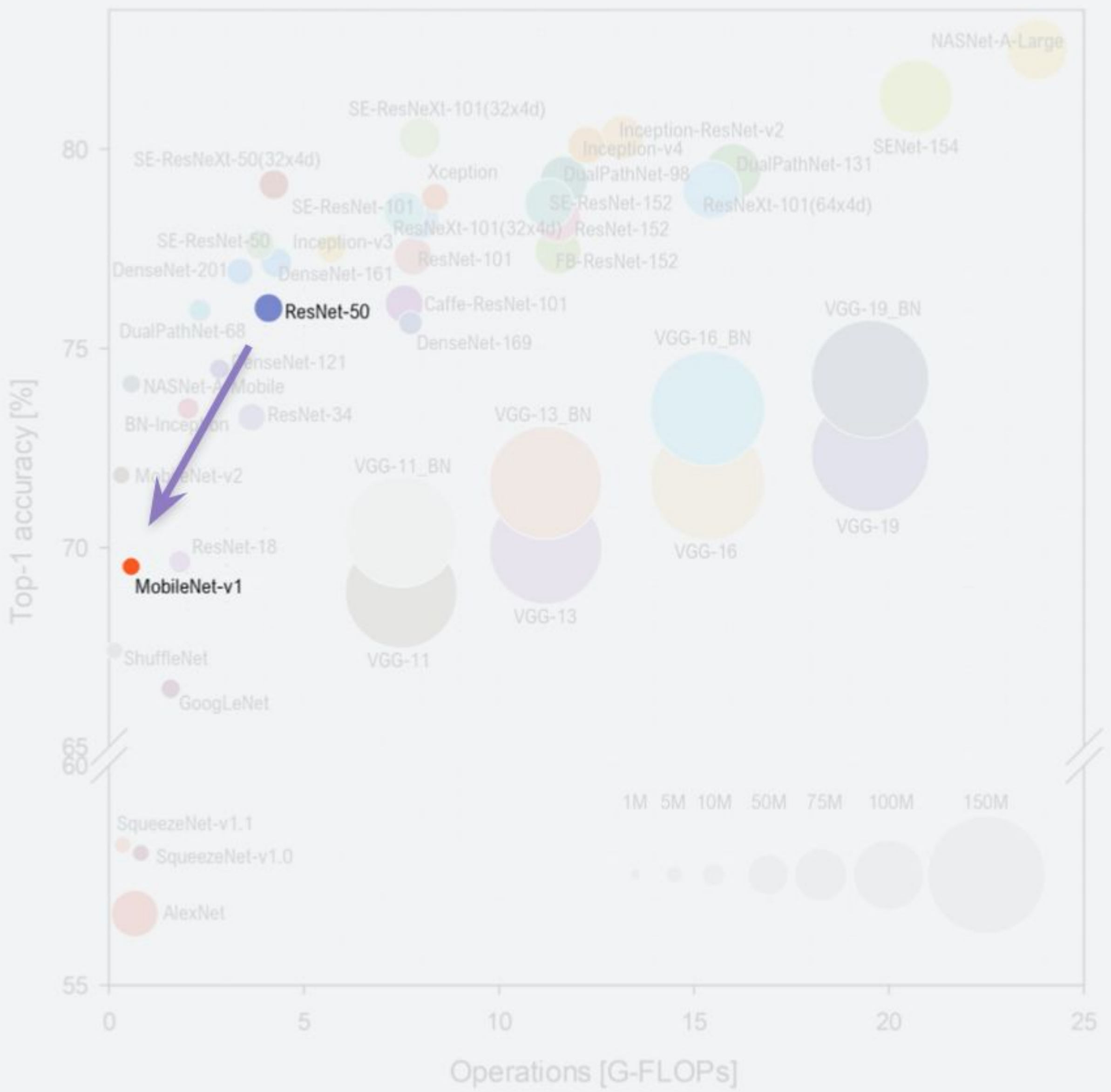
- **ResNet** (2015)
  - **75.8%** accuracy
  - **22.7MB** in size



Source: S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018

ML Model Evolution

- **MobileNet** (2015)
  - ***MobileNetv1***
    - **70.6%** accuracy
    - **16.9MB** in size



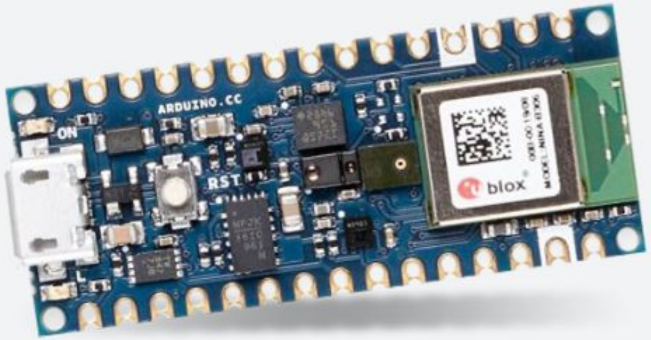
Source: S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018

ML Model Evolution

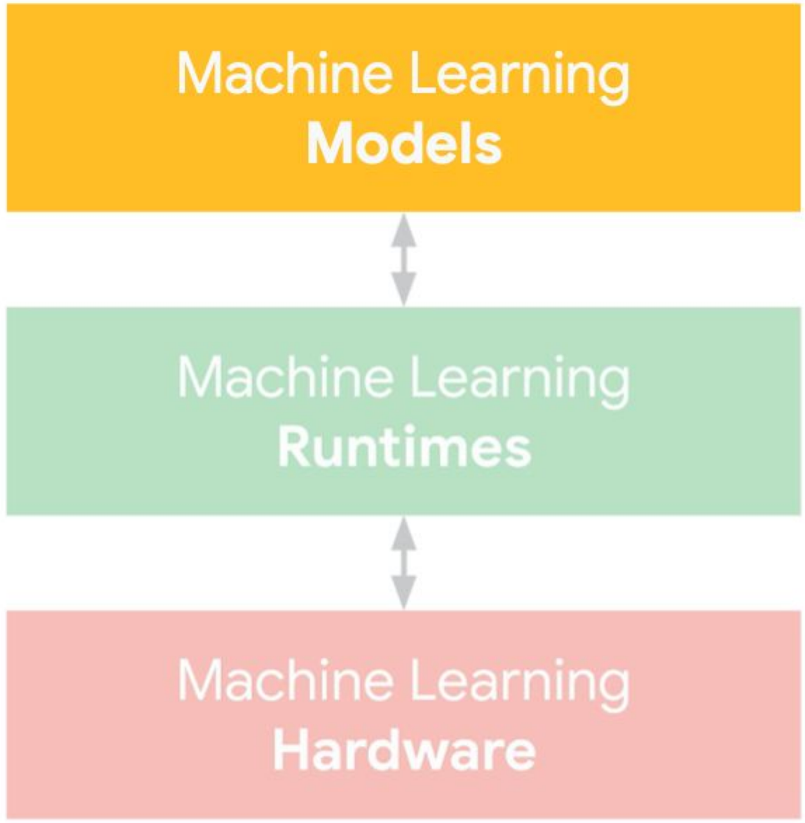
- **MobileNet** (2015)
  - ***MobileNetv1***
    - 70.6% accuracy
    - 16.9MB in size

Problem:

Our board (in your kit for Course)
only has **256KB** of RAM (memory)
yet ***MobileNetv1*** needs **16.9MB**!



Source: S. Bianco, R. Cadene, L. Celona, and P. Napoletano, “Benchmark analysis of representative deep neural network architectures,” *IEEE Access*, vol. 6, pp. 64 270–64 277, 2018



Model Compression Techniques

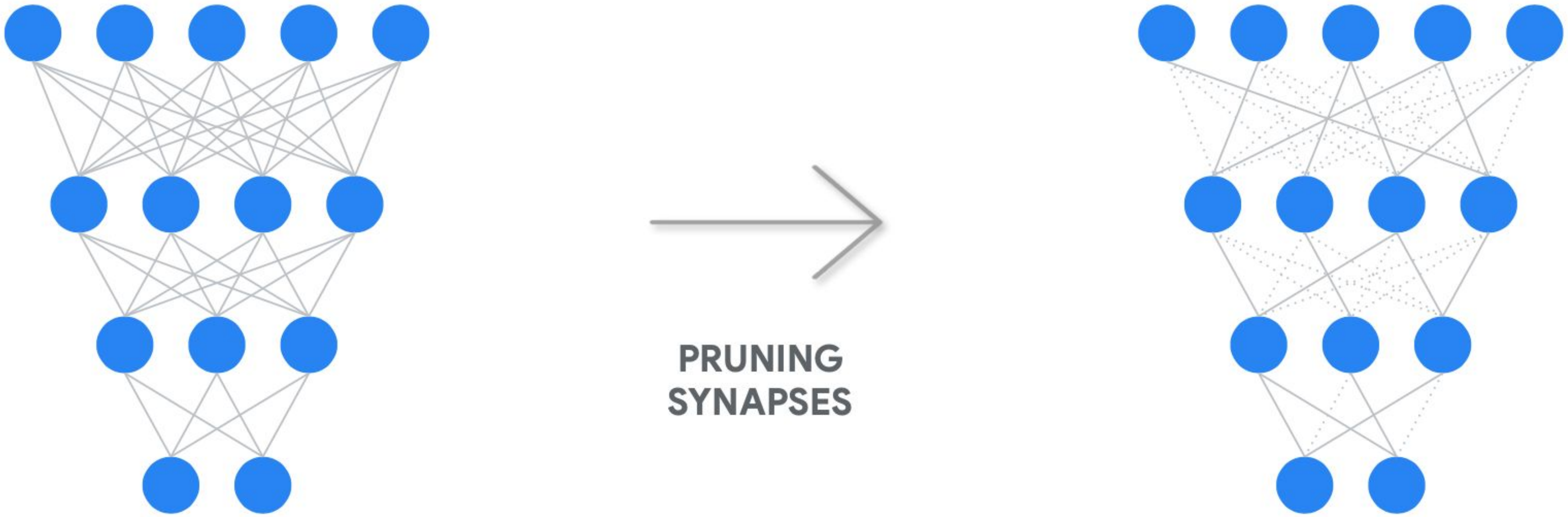
Pruning

Quantization

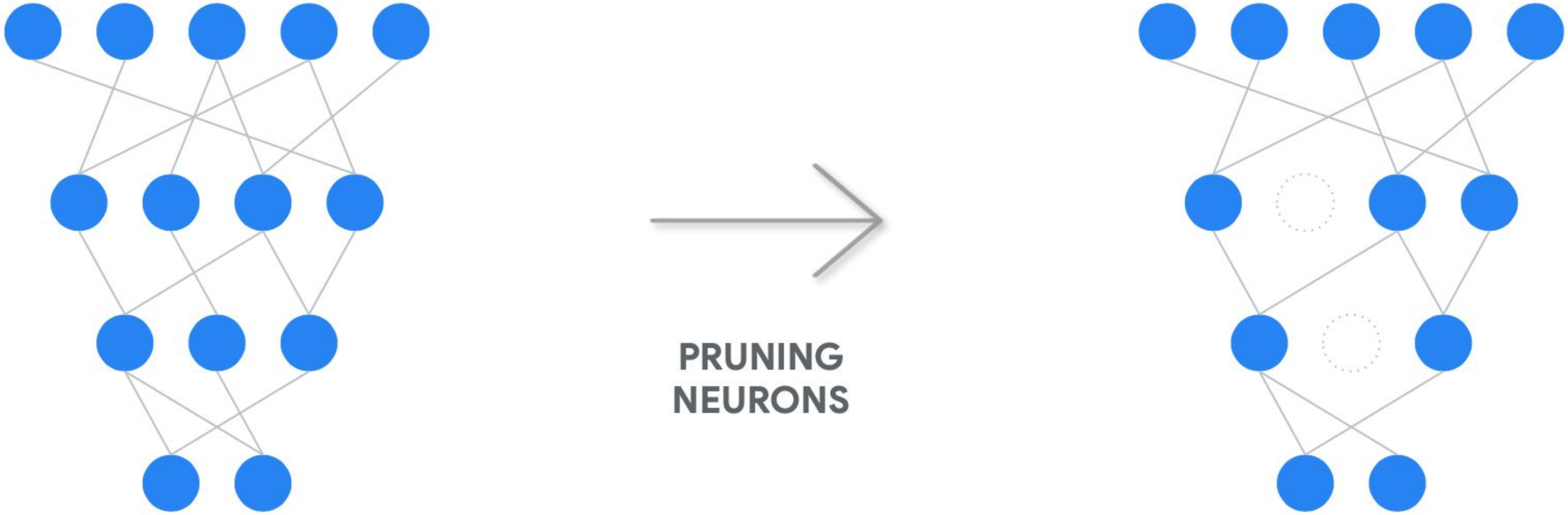
Knowledge Distillation

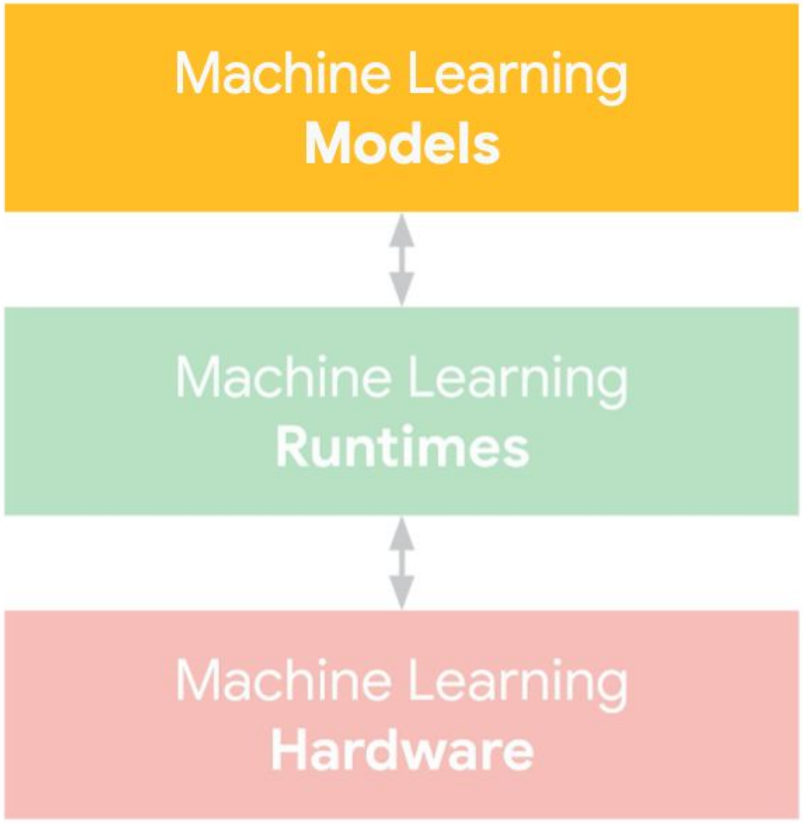
...

Pruning



Pruning





Model Compression Techniques

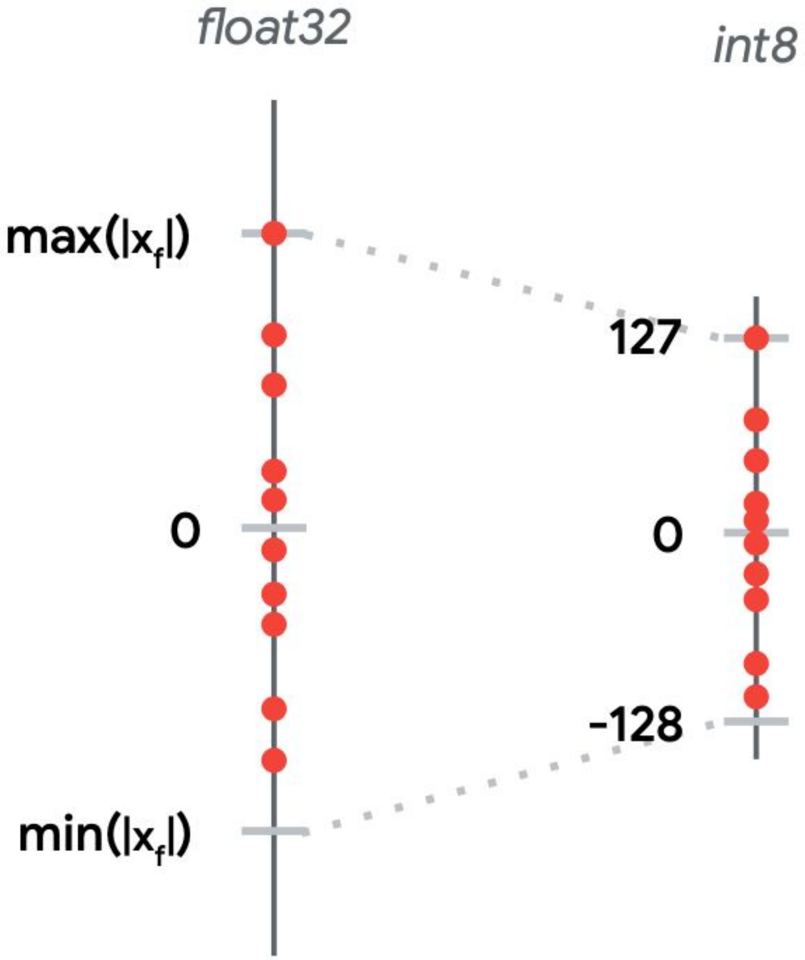
Pruning

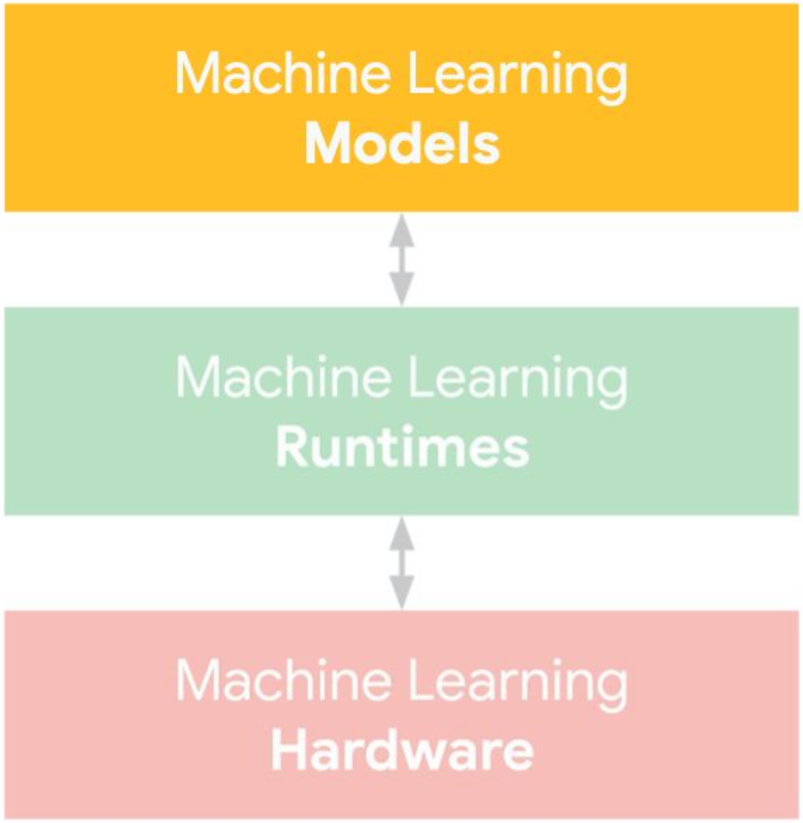
Quantization

Knowledge Distillation

...

Quantization





Model Compression Techniques

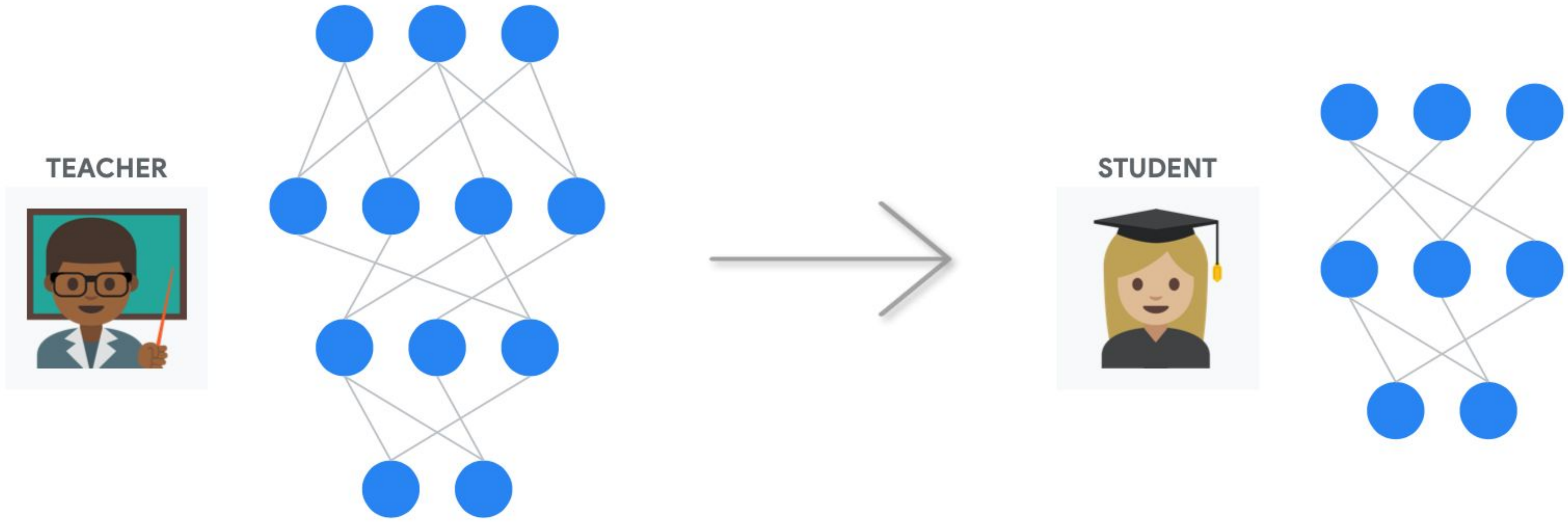
Pruning

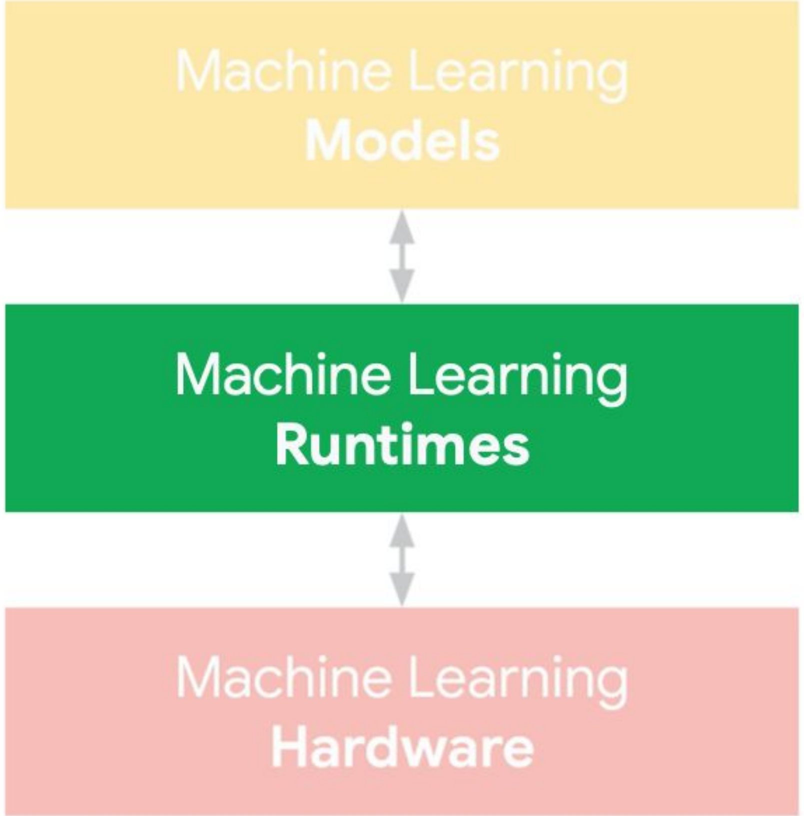
Quantization

Knowledge Distillation

...

Knowledge Distillation





TensorFlow

[TF Video]

TensorFlow

Less memory

Less compute power

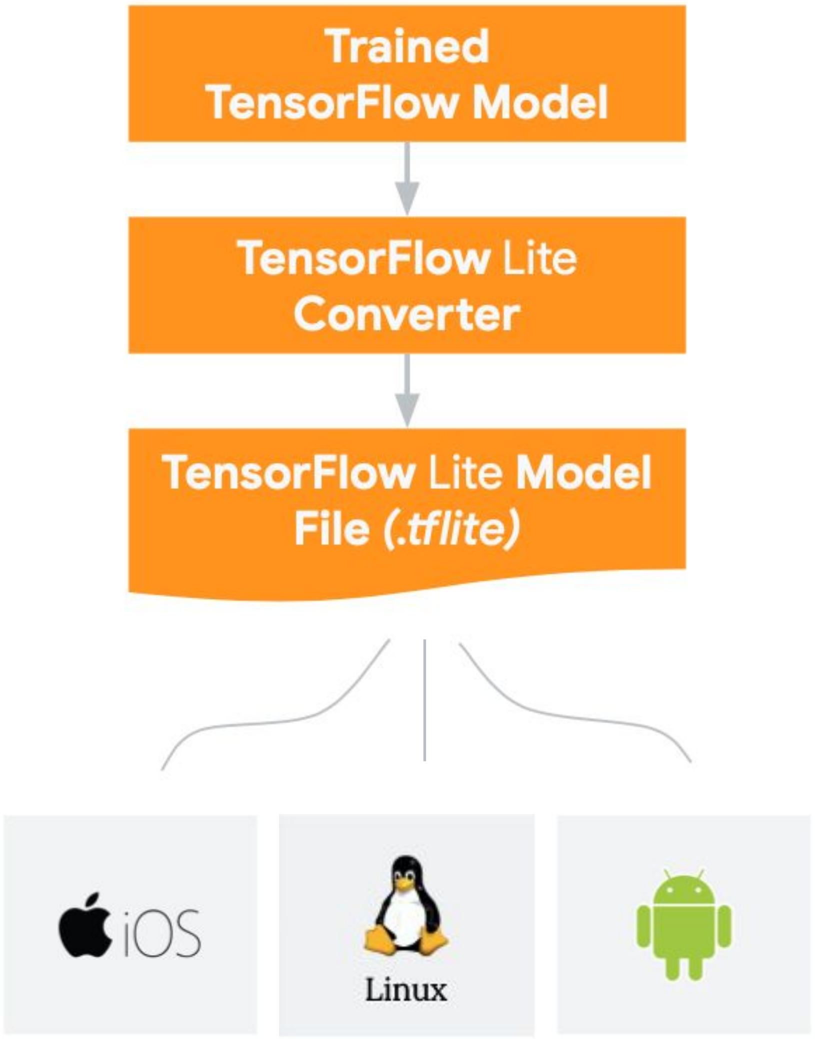
Only focused on *inference*

TensorFlow Lite

Key Differences

	TensorFlow	TensorFlow Lite
Topology	Variable	Fixed
Weights	Variable	Fixed
Binary Size	Unimportant	High Priority
Distributed Compute	Needed	Not Needed
Developer Background	ML Researcher	Application Developer

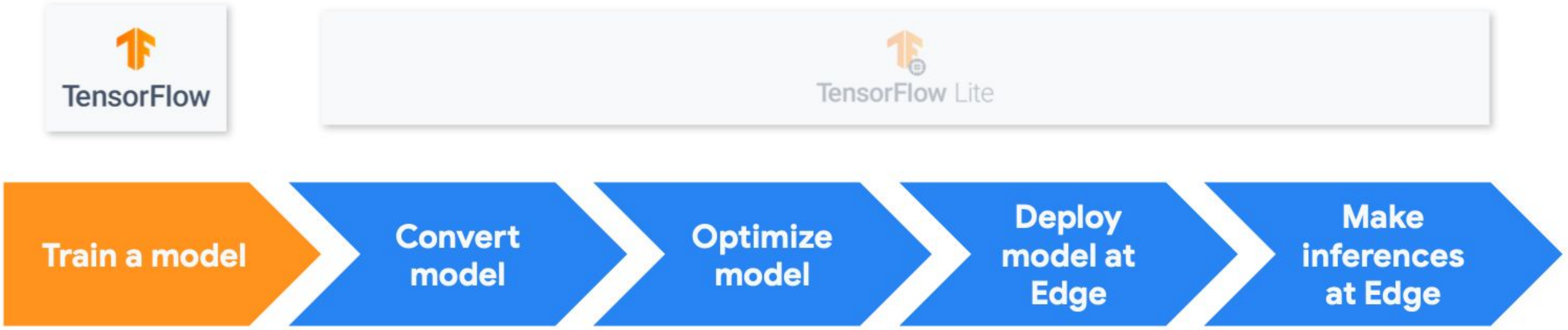
Architecture

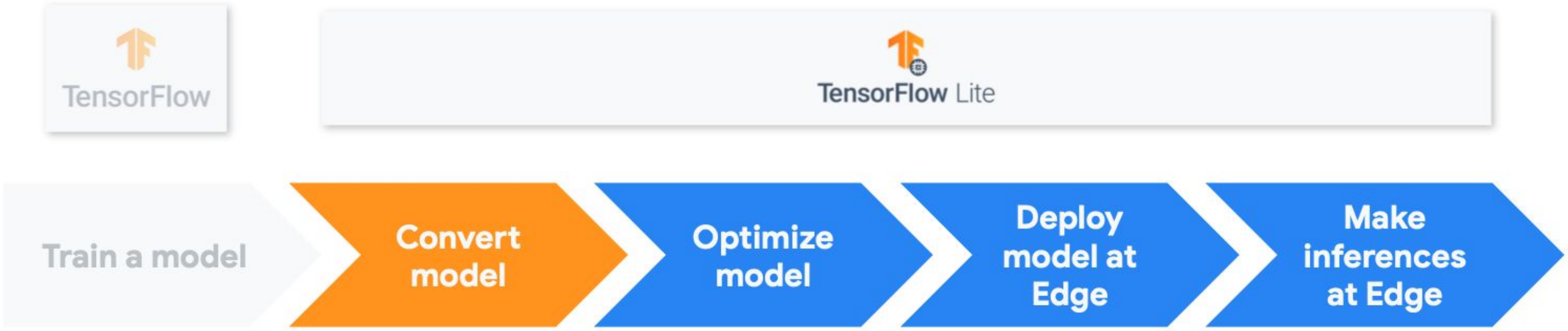


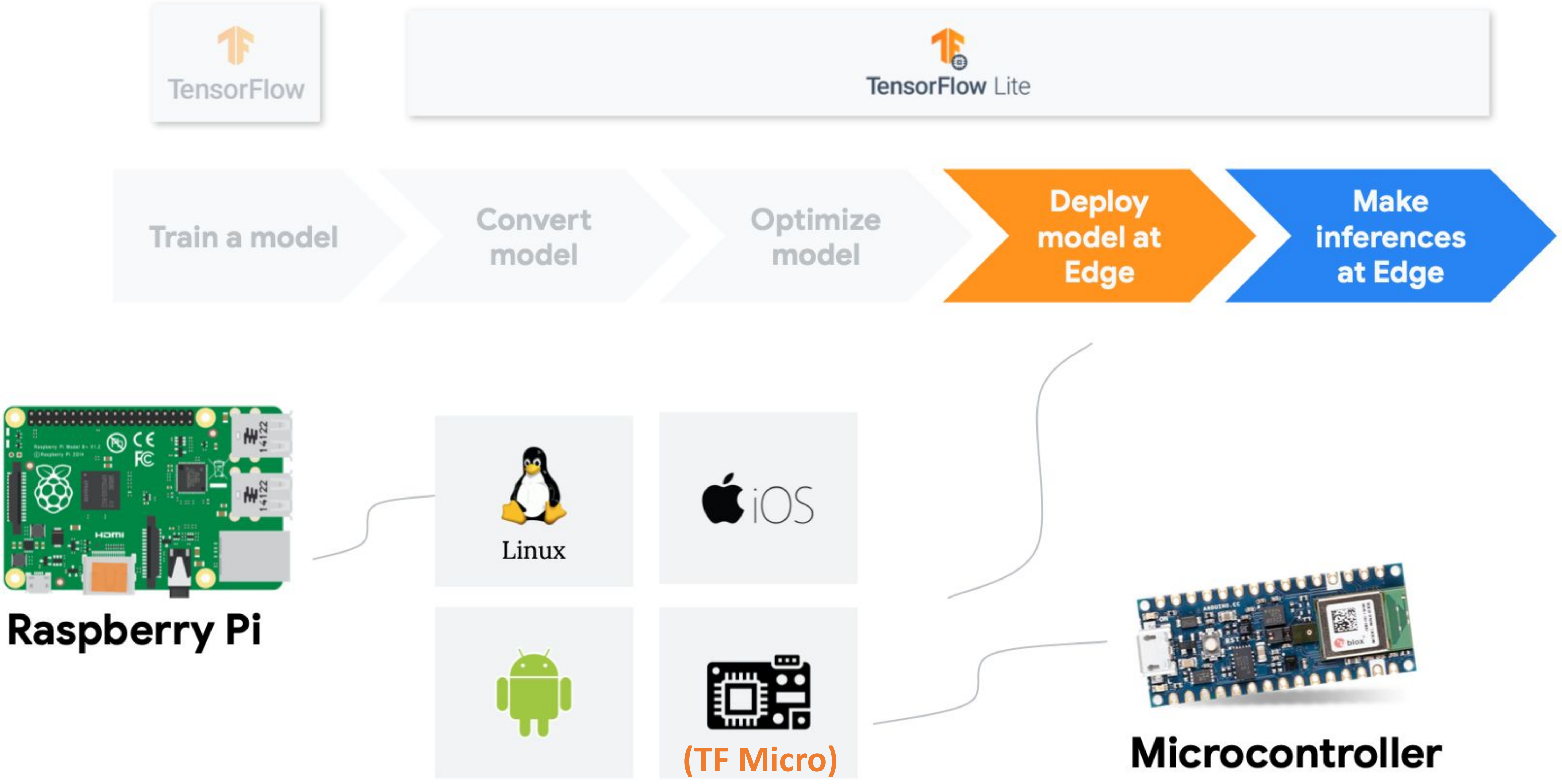
Even less memory

Even less compute power

Also, only focused on *inference*


TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge

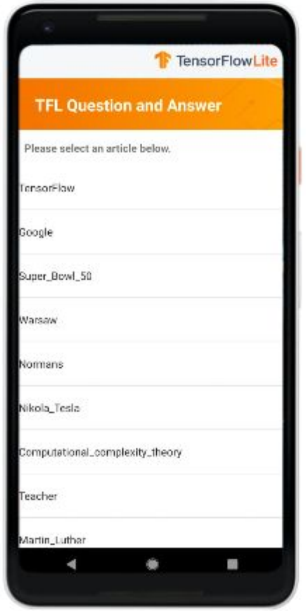

TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge

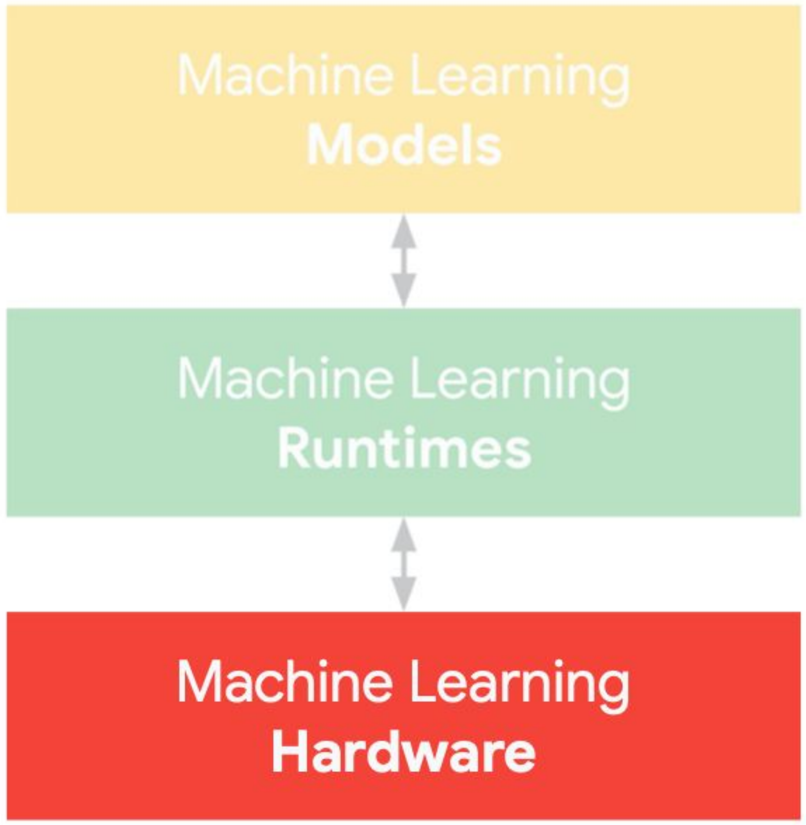

TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge
Linux
iOS
(TF Micro)
Raspberry Pi
Microcontroller

TensorFlow

TensorFlow Lite

- Train a model
- Convert model
- Optimize model
- Deploy model at Edge
- **Make inferences at Edge**

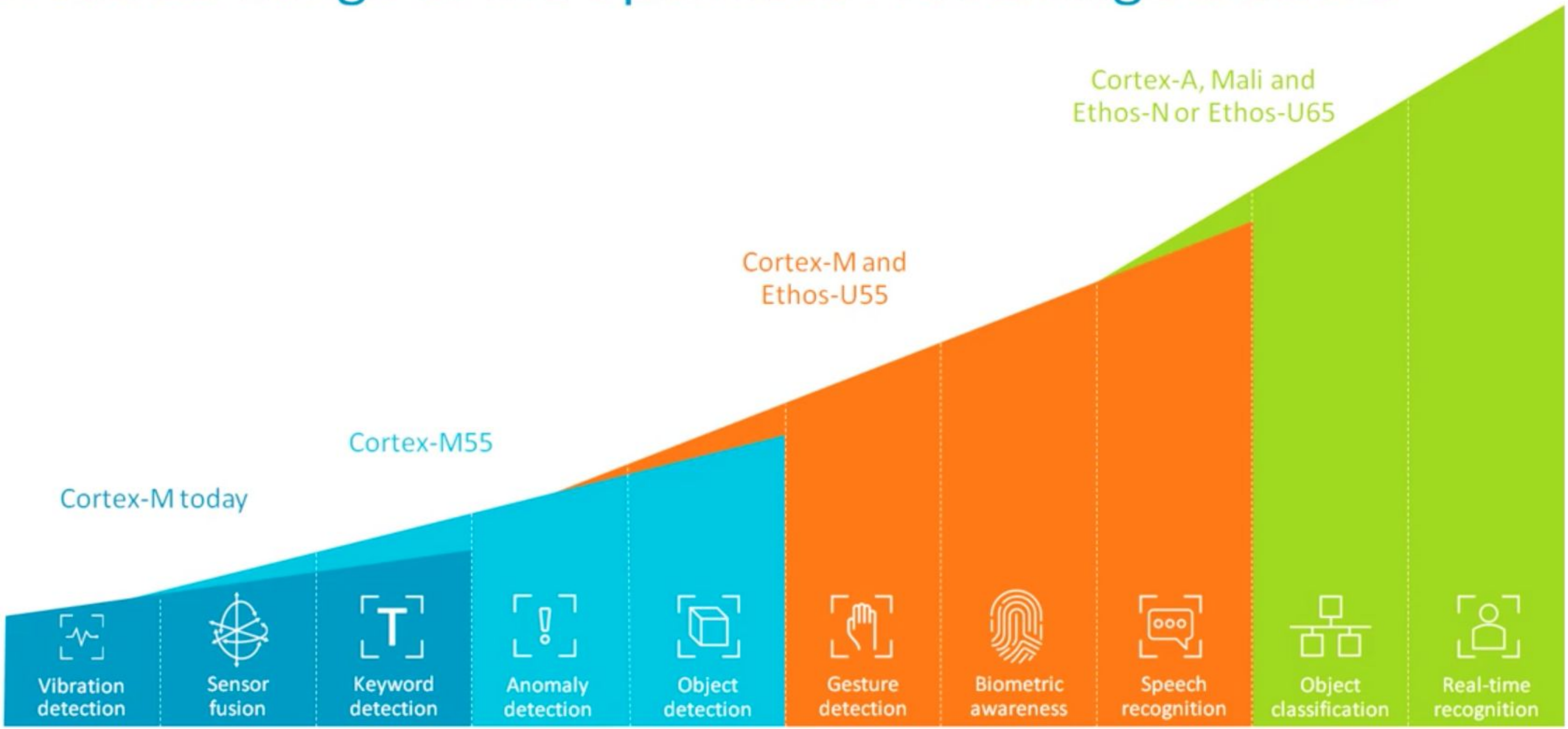




"Energy-efficient On-device Processing for Next-generation Endpoint ML"

By Tomas Edso, Senior Principal Engineer (ML), ARM
At tinyML Summit 2020 presentation

Broadest Range of ML-optimized Processing Solutions



arm

Reading Material

Main references

- Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning
- Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
- Introduction to Embedded Machine Learning - Coursera/Edge Impulse
- Computer Vision with Embedded Machine Learning - Coursera/Edge Impulse
- Fundamentals textbook: "Deep Learning with Python" by François Chollet
- Applications & Deploy textbook: "TinyML" by Pete Warden, Daniel Situnayake
- Deploy textbook "TinyML Cookbook" by Gian Marco Iodice

I want to thank Shawn Hymel and Edge Impulse, Pete Warden and Laurence Moroney from Google, Professor Vijay Janapa Reddi and Brian Plancher from Harvard, and the rest of the TinyMLedu team for preparing the excellent material on TinyML that is the basis of this course at UNIFEI.

The IESTI01 course is part of the **TinyML4D**, an initiative to make TinyML education available to everyone globally.

Thanks