EECE-4710 "IoT and TinyML"

Introduction to IoT and TinyML

Cris Ababei

MARQUETTE UNIVERSITY

BE THE DIFFERENCE.

1

1

What Is IoT?

2

2

What is IoT?

Internet of Things (IoT): is the network of physical objects or "things"—devices, vehicles, buildings and other items embedded with electronics, software, sensors, and network connectivity—that enables these objects to collect and exchange data.

3

3

Various Names, One Concept

- M2M (Machine to Machine)
- "Internet of Everything" (Cisco Systems)
- "World Size Web" (Bruce Schneier, "security guru")
- "Skynet" (Terminator movie)

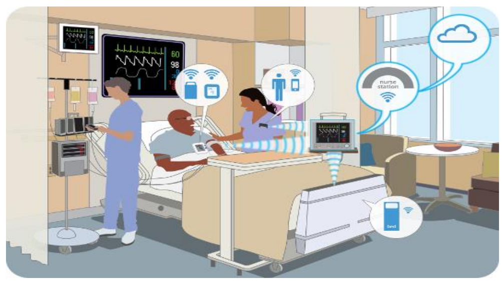
4

4

Where is IoT? Everywhere!

Smart Appliances

Wearable Tech

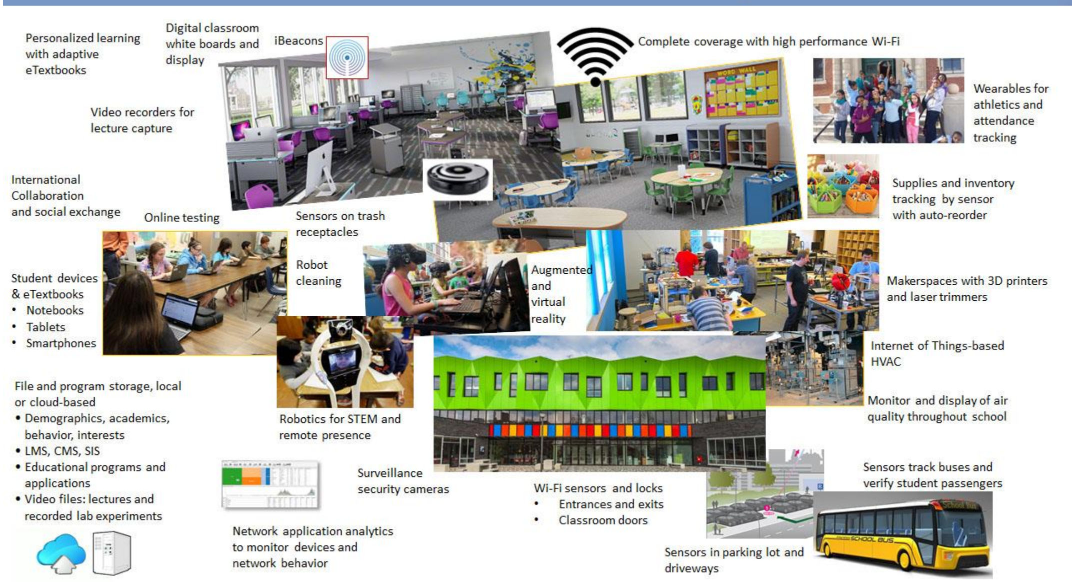


Healthcare

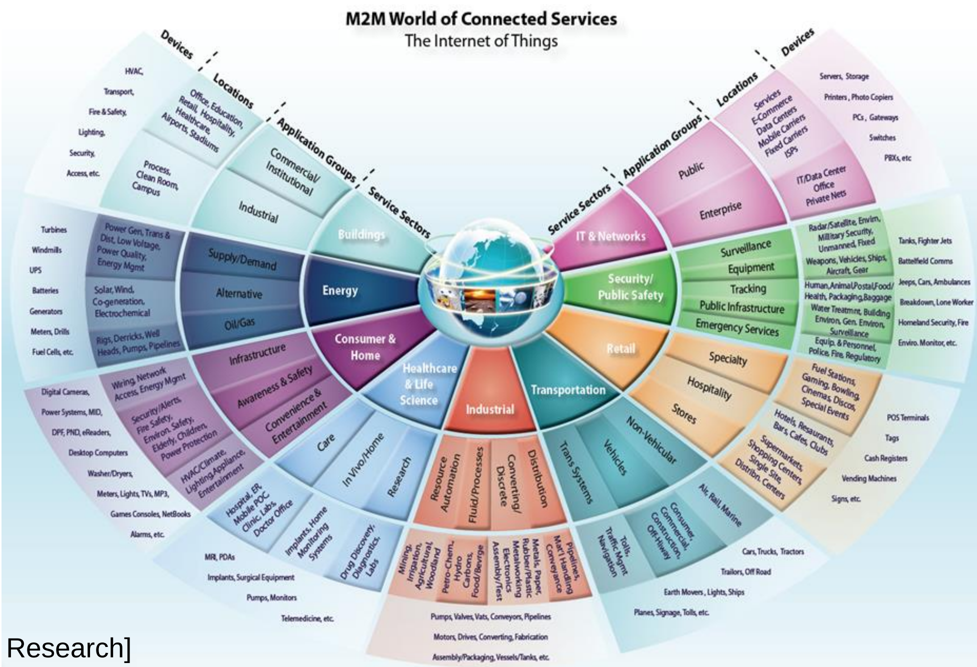
5

5

The Smart *Internet of Things* School



6



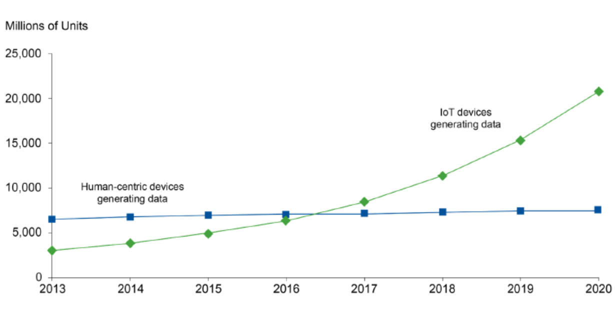
[Beecham Research]

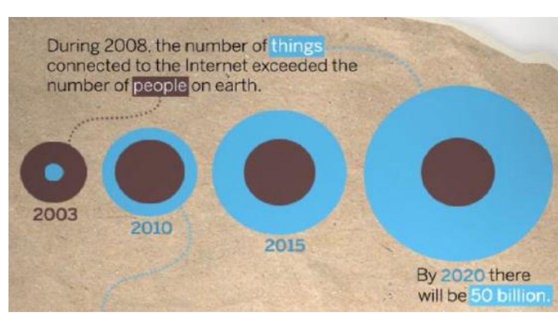
7

7

The IoT Market

- As of 2013, 9.1 billion IoT units
- Expected to grow to 28.1 billion IoT devices by 2020
- Revenue growth from $1.9 trillion in 2013 to $7.1 trillion in 2020





8

As device sensors proliferate across every company's value chain – from new product development through inspection, tracking, and delivery – tinyML is surfacing to provide actionable insights, transforming business as we know it. There are sound economic reasons for all this interest and activity. McKinsey researchers predict IoT will have a potential economic impact of US $4-11 trillion by 2025, identifying manufacturing as the largest vertical (US $1.2-3.7 trillion).

Source: https://www.forbes.com/sites/sap/2021/11/08/meet-tinyml-the-latest-machine-learning-tech-having-an-outsize-business-impact/

9

9

Technological Challenges of IoT

- Scalability
- Technological Standardization
- Inter operability
- Discovery
- Software complexity
- Data volumes and interpretation
- Power Supply
- Interaction and short-range communication
- Wireless communication
- Fault tolerance

10

10

Why be Concerned about IoT?

- It is just another computer, right?
  - All of the same issues we have with access control, vulnerability management, patching, monitoring, etc.
  - Imagine your network with 1,000,000 more devices
  - Any compromised device is a foothold on the network
  - Are highly portable devices captured during vulnerability scans?
  - Where is your network perimeter?
  - Are consumer devices being used in areas – like health care – where reliability is critical?

11

11

Criticisms and Controversies of IoT

Scholars and social observers and pessimists have doubts about the promises of the ubiquitous computing revolution, in the areas as:

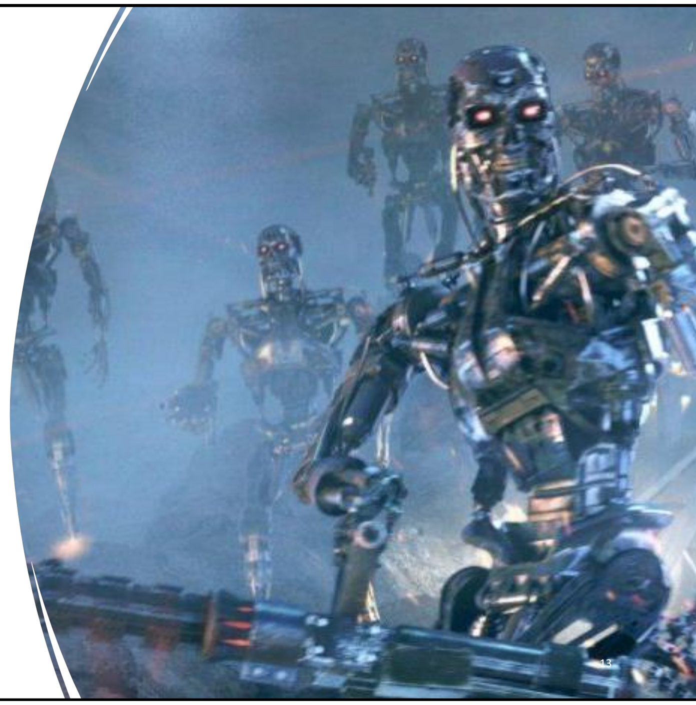
- Privacy
- Security
- Autonomy and Control
- Social control
- Political manipulation
- Environmental impact
- Influences human moral decision making

12

12

Threat vs. Opportunity

- If misunderstood and misconfigured, IoT poses risk to our data, privacy, and safety
- If understood and secured, IoT will enhance communications, lifestyle, and delivery of services



13

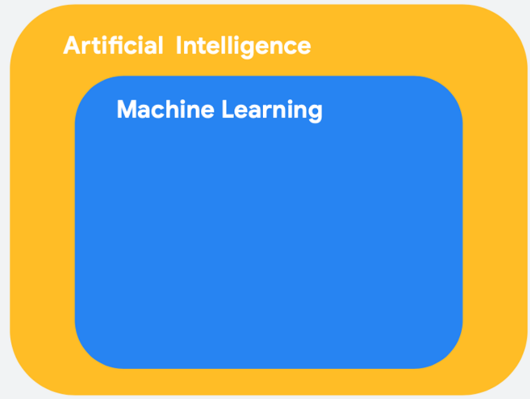
What Is Machine Learning (ML)?

14

14

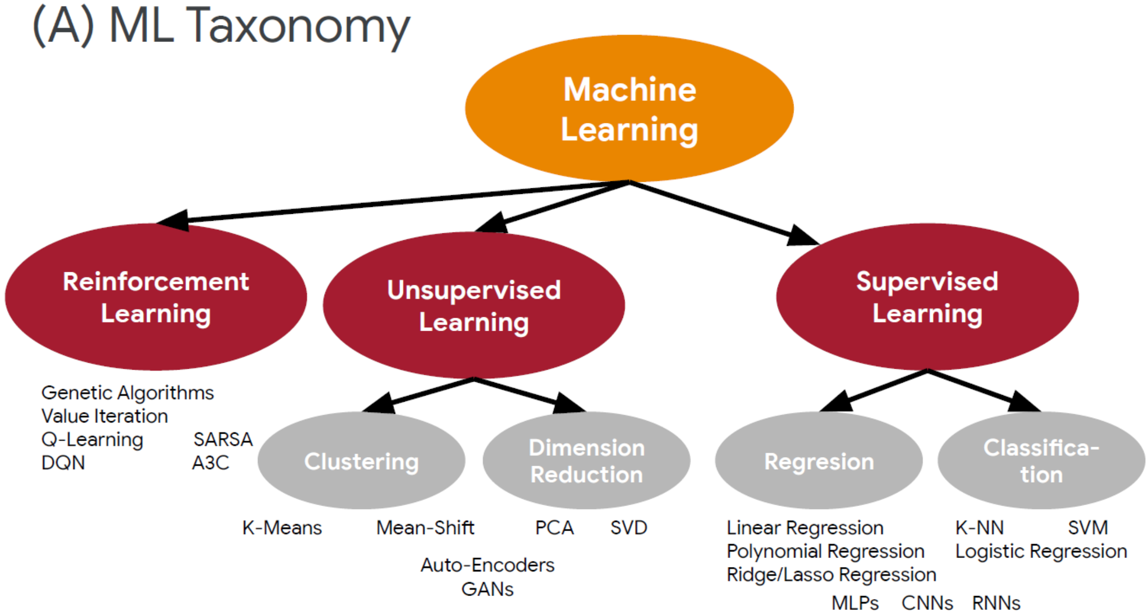
What is Machine Learning?

1. **Machine Learning** is a subfield of **Artificial Intelligence** focused on developing algorithms that learn to **solve problems by analyzing data for patterns**



15

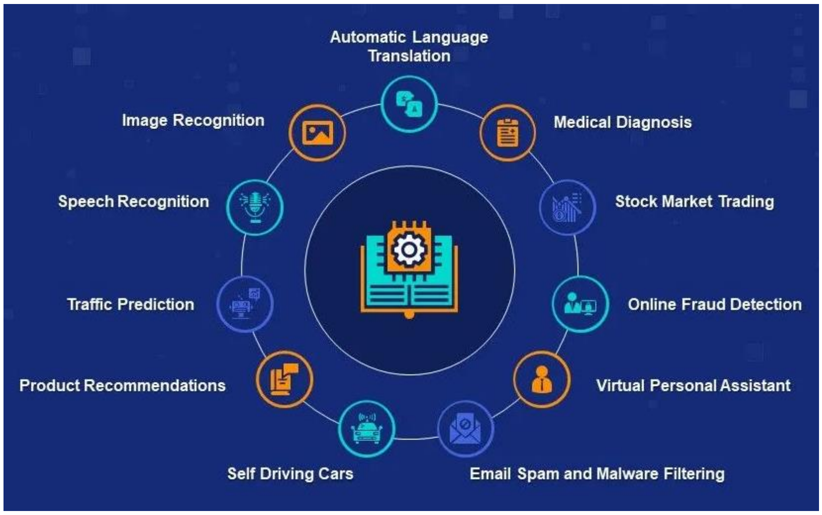
(A) ML Taxonomy



For detailed discussion of this (with examples) see: https://a2r-lab.org/files/CS249_F20_Slides.pdf 16

16

Applications of Machine Learning

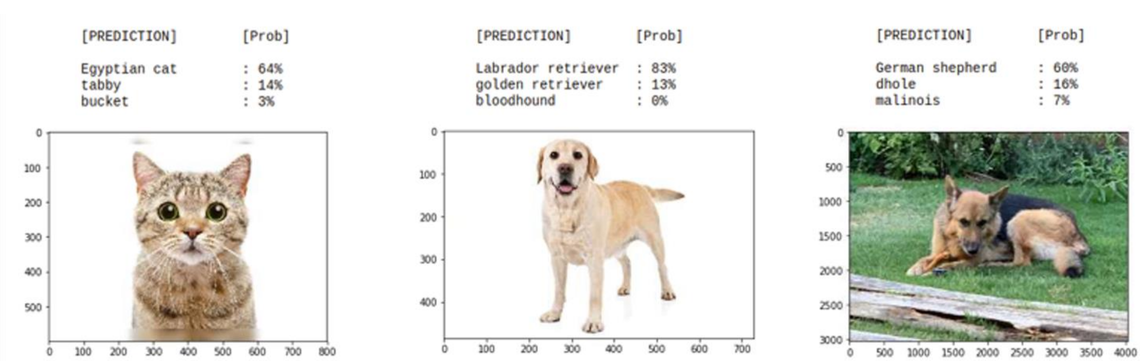


Source: https://swisscognitive.ch/2021/03/18/applications-of-machine-learning

17

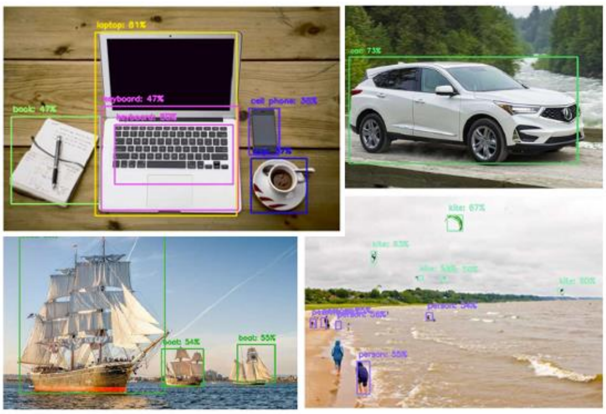
17

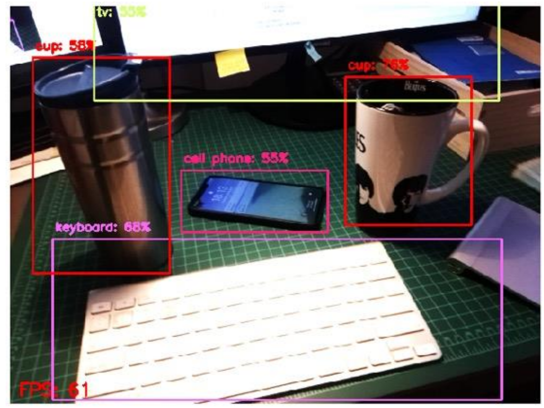
Image Classification



18

Object Detection





Photos

Live Video

19

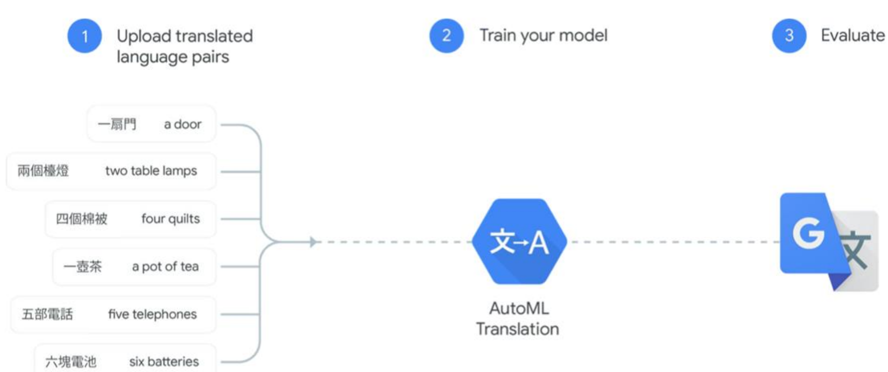
Segmentation

20

Pose Estimation

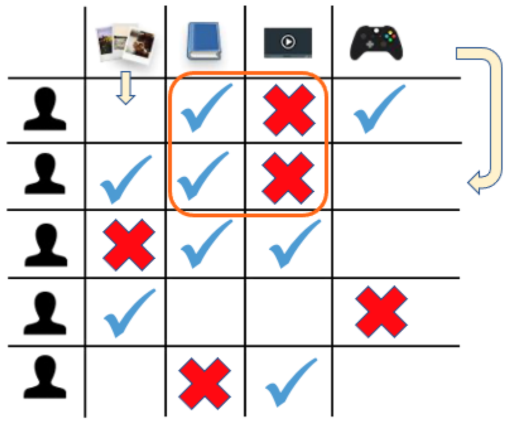
21

Machine Translation



22

Recommendations



Dedicated ML Application examples

- Image Classification
- Object Detection
- Pose Estimation
- Voice Recognition
- Gesture Recognition
- Anomaly Detection
- Natural Language Processing (NLP)

Dedicated TinyML Application Examples

- Image Classification (Camera)
- Object Detection (〃)
- Pose Estimation
- Voice Recognition (Microphone)
- Gesture Recognition (Accelerometer)
- Anomaly Detection (〃)
- Natural Language Processing (NLP)

25

IoT 1.0: Internet of Things

26

26


IoT 2.0: Intelligence on Things
Google
Assistant
nest
10:09
396
13
03
27

27


IoT 2.0: Intelligence on Things
Challenges:
Bandwidth
Reliability
Latency
Privacy
Energy
nest
10:09
396
13
03
28

28

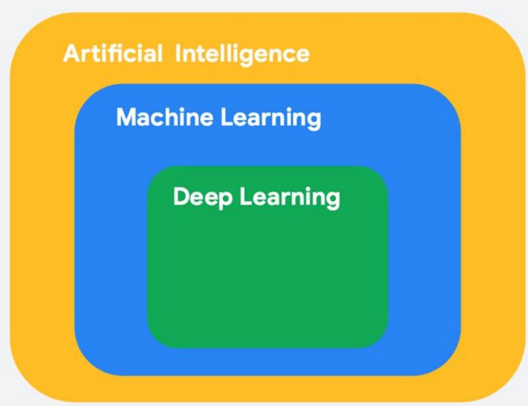
Deep Learning

29

29

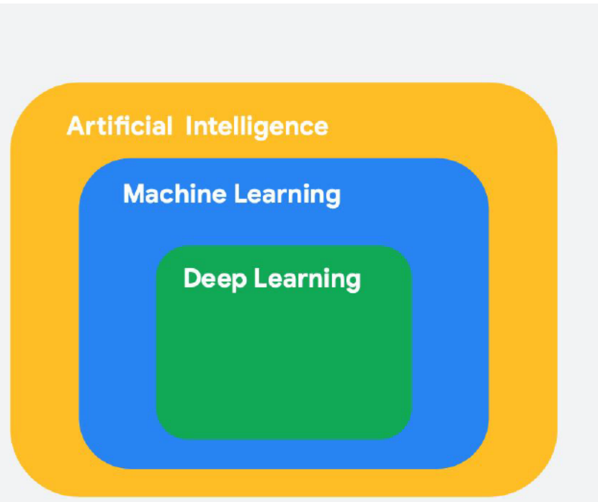
What is (Deep) Machine Learning?

1. Machine Learning is a subfield of Artificial Intelligence focused on developing algorithms that learn to solve problems by analyzing data for patterns
2. **Deep Learning** is a type of Machine Learning that leverages **Neural Networks** and **Big Data**



For detailed discussion of this (with examples) see: https://a2r-lab.org/files/CS249_F20_Slides.pdf

30



AI: Any technique that enables computers to mimic human behavior

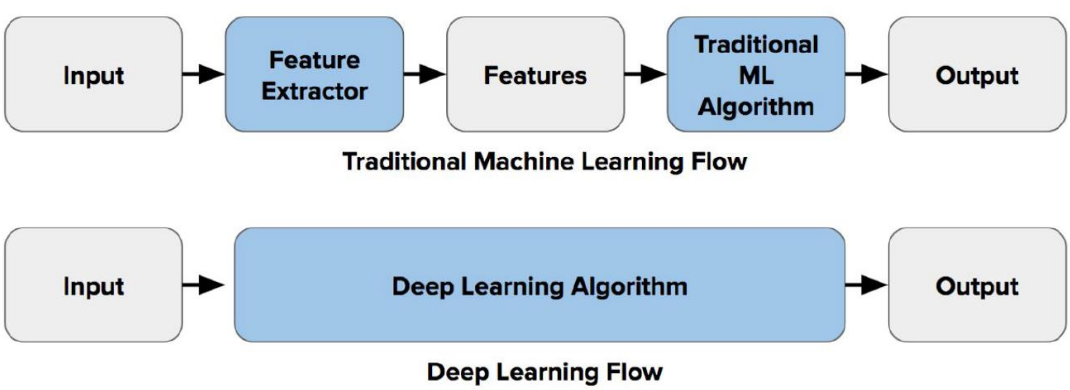
ML: Ability to learn without explicitly being programed

DL: Extract patterns from data using neural networks

31

31

Deep vs Classical Learning



https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
https://cv-tricks.com/cnn/understand-resnet-alexnet-vgg-inception/

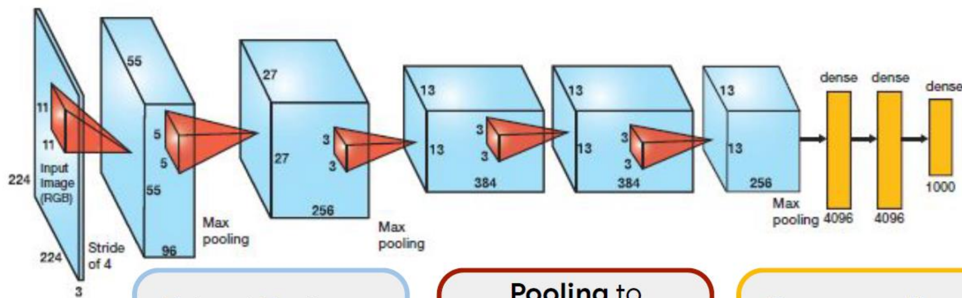
"Deep Learning automates the design and interactions of features by constructing deep networks of nonlinearly activated, connected neurons"

32

32

The Rise of Deep Learning: AlexNet

AlexNet



Extract **Features** from data with **Convolutions**

Pooling to **summarize** for higher level features

Dense nonlinear layers to **classify** the final features

https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
https://cv-tricks.com/cnn/understand-resnet-alexnet-vgg-inception/

3

33

Deep Supervised Learning

1. **Collect** LOTS of (unbiased) data!
2. **Preprocess** the data and **design** your model
3. **Train** your model (in the cloud)
4. **Evaluate** your model and improve hyper parameters
5. **Deploy** and efficient inference engine for your model

Use a machine learning framework (e.g., TensorFlow)

34

From Deep Learning to TinyML

How tiny is Tiny?

Table 4: Memory of CNN models on platforms (MB)

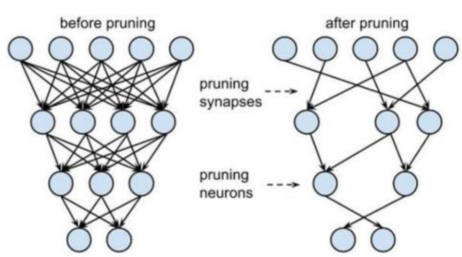
Type/Platform	AlexNet	VGGNet	GoogleNet	ResNet
Weights & Biases	233	528	26	97
Data	8	110	53	221
Workspace	11	168	46	79

Our board only has 256Kb of RAM!

https://arxiv.org/pdf/1709.09503.pdf

How Can We Compress Things? Pruning and Quantization

Pruning removes the least important “stuff” from the model



Pruning removes the least important “stuff” from the model

Table 2: MobileNet sparse vs dense results

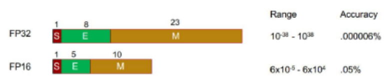
Width	Sparsity	NNZ params	Top-1 acc.	Top-5 acc.
1.0	0%	4.21M	70.6%	89.5%
	50%	2.13M	69.5%	89.5%
	75%	1.09M	67.7%	88.5%
	90%	0.46M	61.8%	84.7%
	95%	0.25M	53.6%	78.9%

https://openreview.net/forum?id=S1lN69AT-

Very small accuracy penalty!

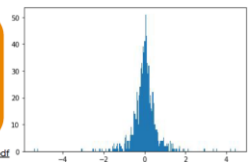
How Can We Compress Things? Pruning and Quantization

Quantization compresses the numerical representation



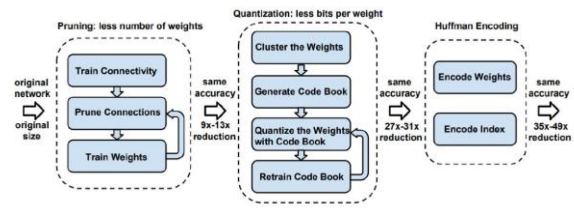
So can we do better?

Turns out weights are often even more densely packed than that!
https://drive.google.com/file/d/1Pn-lNOty-P1fX6G66oa6wvsEn46Xyx8Y



https://media.nips.cc/Conferences/2015/tutorialslides/Dally-NIPS-Tutorial-2015.pdf

Quantization needs to be optimized for each model!

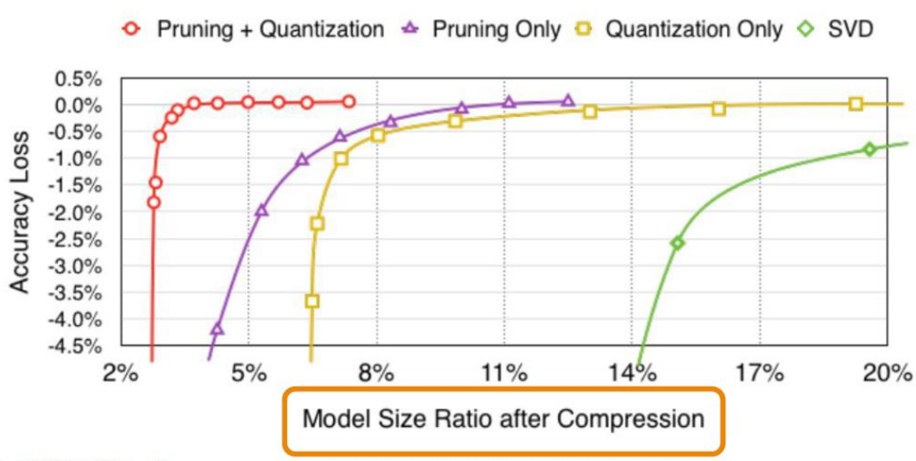


https://arxiv.org/pdf/1510.00149.pdf

37

37

Pruning and Quantization to the rescue!



https://arxiv.org/pdf/1510.00149.pdf

38

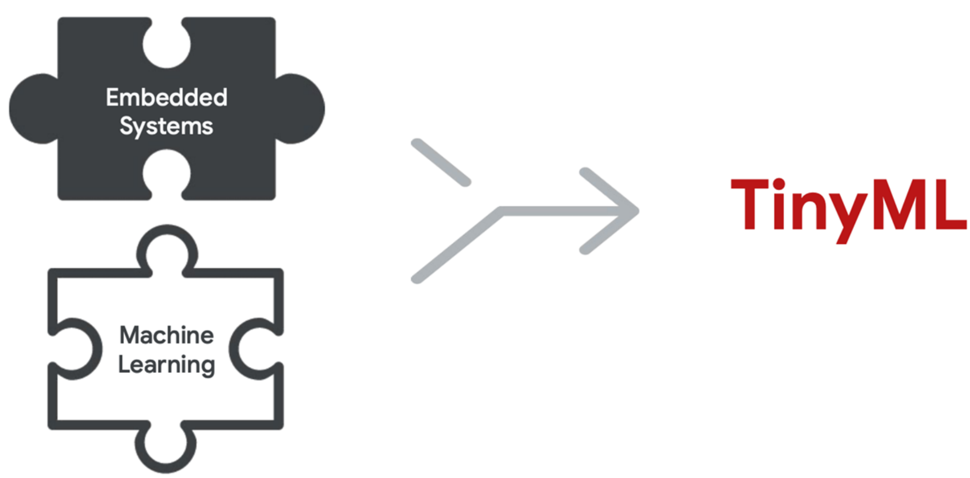
38

What Is TinyML?

39

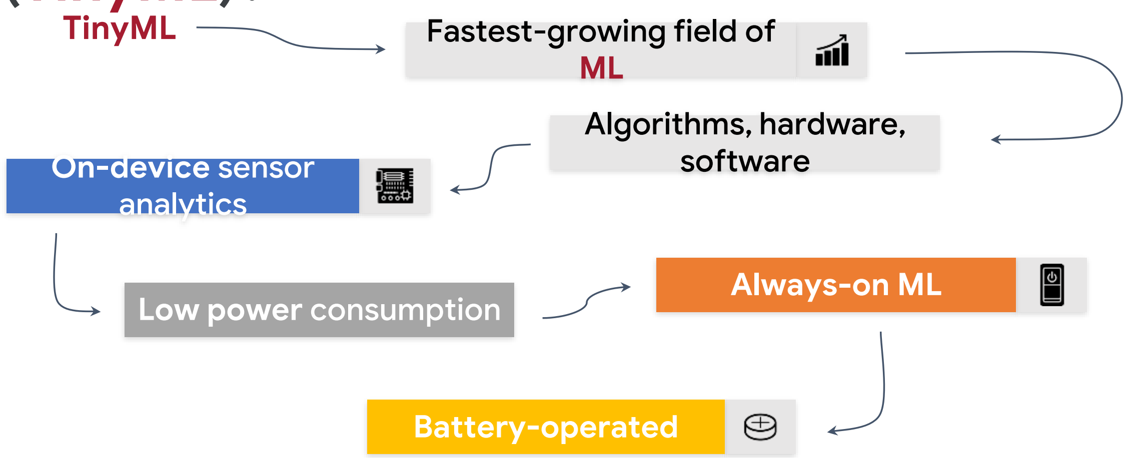
39

What Makes TinyML?



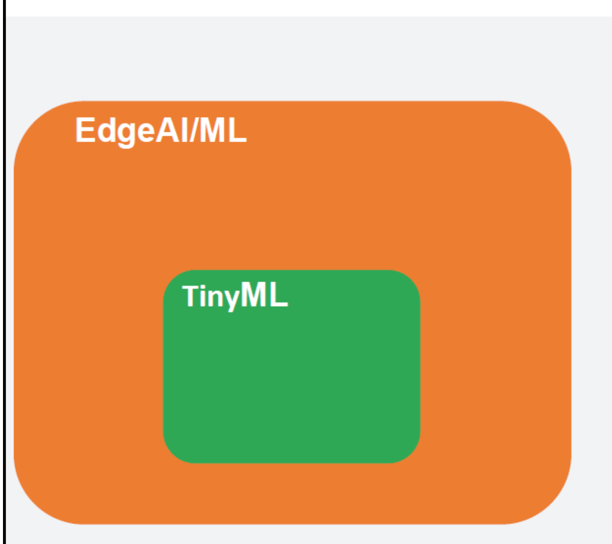
40

What is Tiny Machine Learning (TinyML)?



41

41

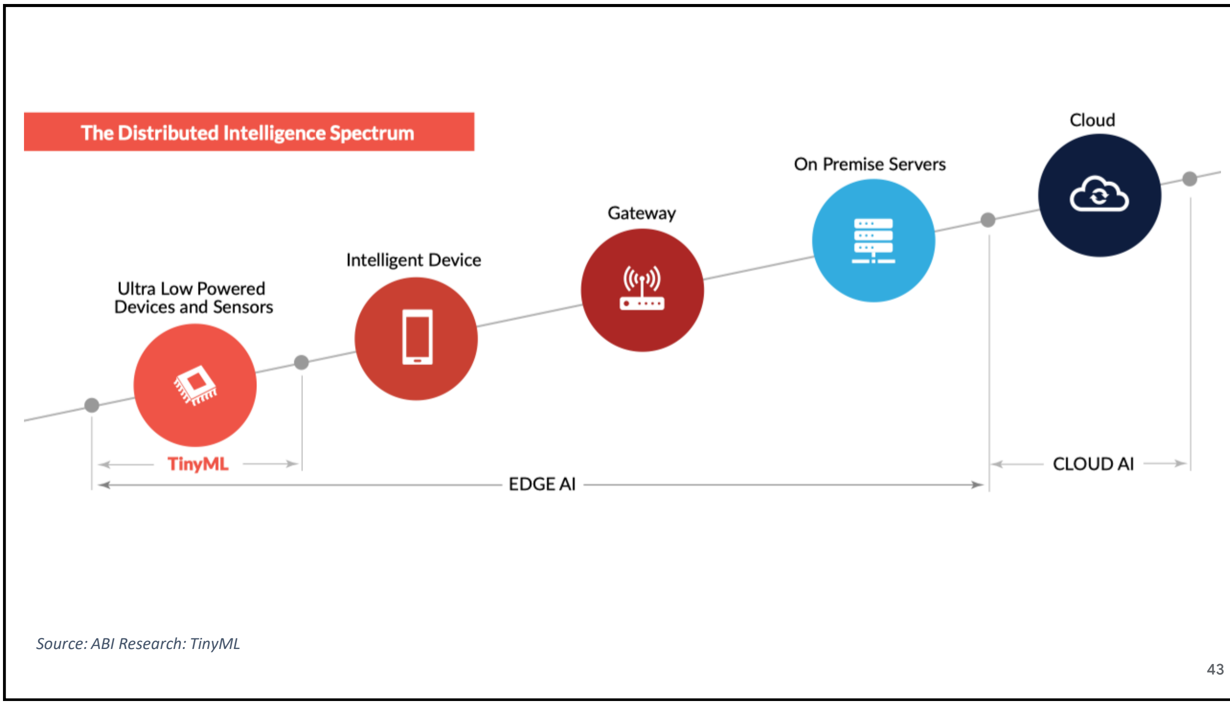


Edge AI (or Edge ML) is the processing of Artificial Intelligence algorithms on edge, that is, on users' devices. The concept derives from **Edge Computing**, which starts from the same premise: data is stored, processed, and managed directly at the Internet of Things (IoT) endpoints.

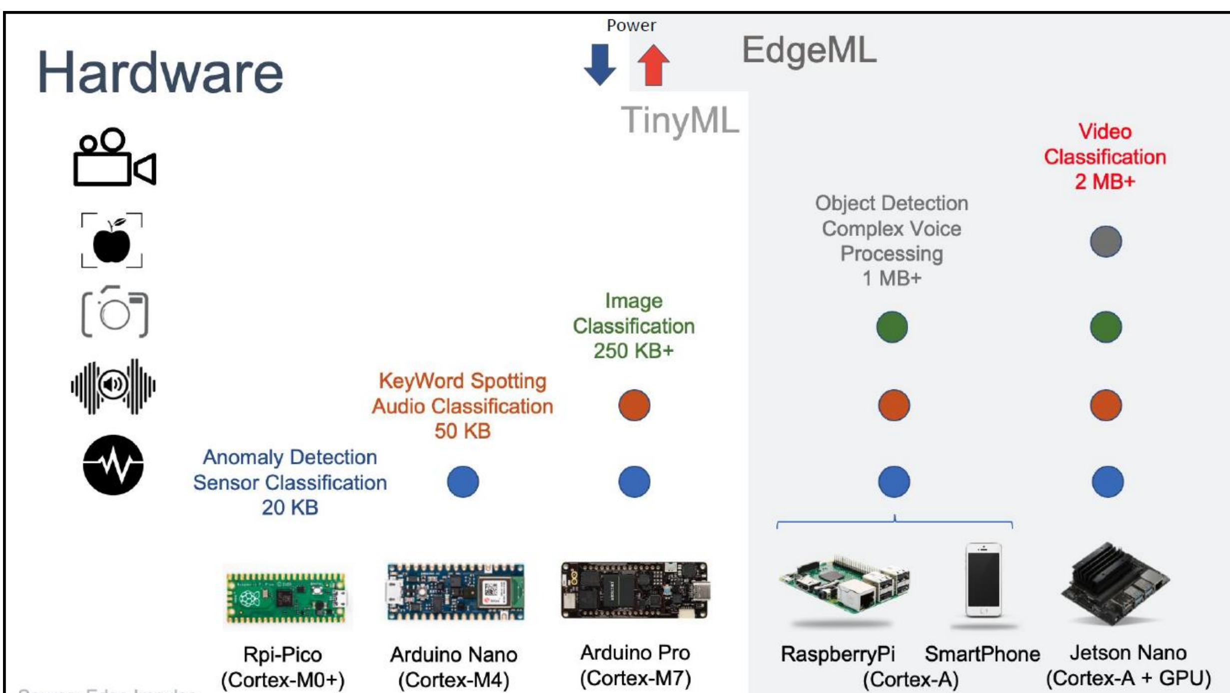
TinyML is a subset of EdgeML, where sensors are generating data with ultra-low power consumption (batteries), so that we can ultimately deploy machine learning continuously ("always on devices")

42

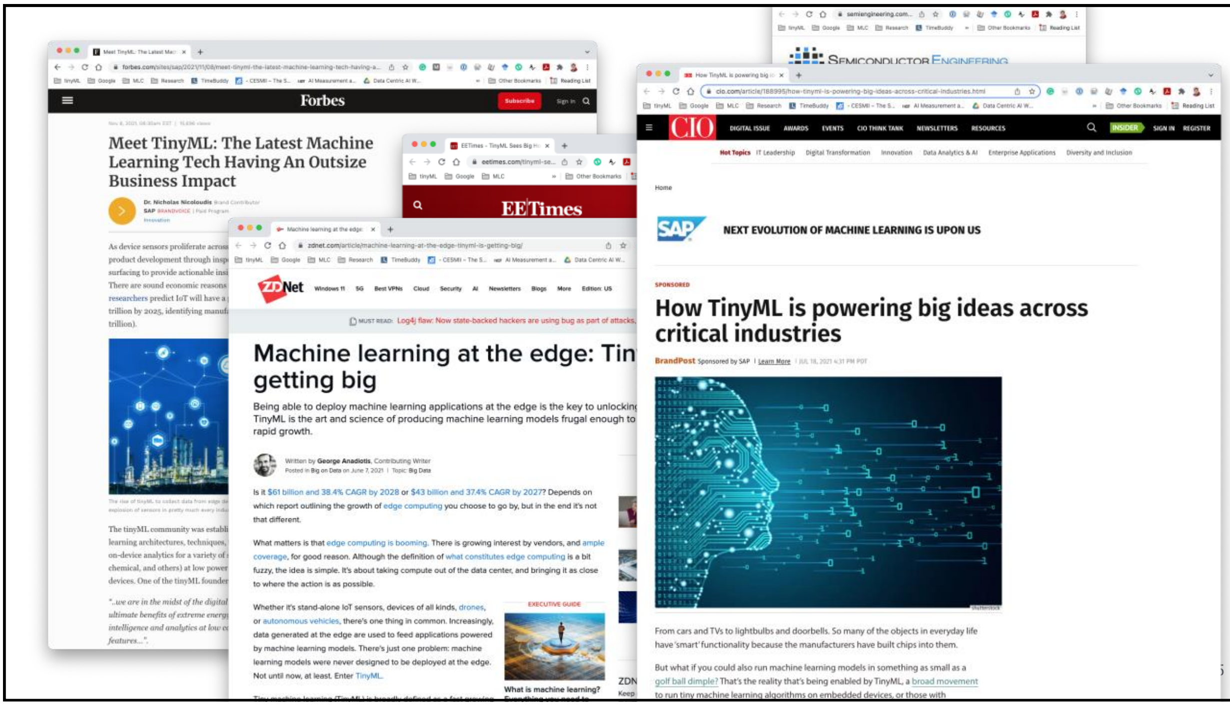
42


The Distributed Intelligence Spectrum
Cloud
On Premise Servers
Gateway
Intelligent Device
Ultra Low Powered
Devices and Sensors
TinyML
EDGE AI
CLOUD AI
Source: ABI Research: TinyML
43

43


Hardware
Power
EdgeML
TinyML
Video
Classification
2 MB+
Object Detection
Complex Voice
Processing
1 MB+
Image
Classification
250 KB+
KeyWord Spotting
Audio Classification
50 KB
Anomaly Detection
Sensor Classification
20 KB
Rpi-Pico
(Cortex-M0+)
Arduino Nano
(Cortex-M4)
Arduino Pro
(Cortex-M7)
RaspberryPi
SmartPhone
(Cortex-A)
Jetson Nano
(Cortex-A + GPU)
Source: Edge Impulse

44



45

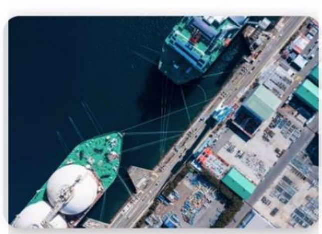
Applications of TinyML

Predictive Maintenance

Motion, current, audio and camera

- → **Industrial**
- → White goods
- → Infrastructure
- → Automotive

Asset Tracking & Monitoring



Motion, temp, humidity, position, audio and camera

- → Logistics
- → Infrastructure
- → Buildings
- → **Agriculture**

Human & Animal Sensing

Motion, radar, audio, PPG, ECG

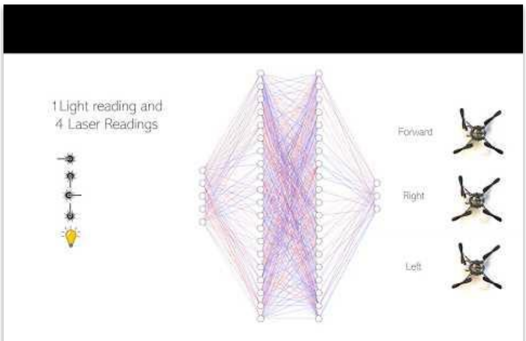
- → **Health**
- → Consumer
- → Industrial

For examples see:
https://github.com/Mjrovai/UNIFEI-IESTI01-TinyML-2022.1/blob/main/00_Curse_Folder/1_Fundamentals/Class_01/IESTI01_TinyML_class_1.pdf

46

46

TinyRL: Autonomous Navigation on Nano Drone



[ICRA'21]

[IROS'21]

47

TinyML Is All About Sensor Data Intelligence
Endpoints Have Sensors, Tons of Sensors

Motion Sensors
Gyroscope, radar, magnetometer, accelerator

Acoustic Sensors
Ultrasonic, Microphones, Geophones, Vibrometers

Environmental Sensors
Temperature, Humidity, Pressure, IR, etc.

Touchscreen Sensors
Capacitive, IR

Image Sensors
Thermal, Image

Biometric Sensors
Fingerprint, Heart rate, etc.

Force Sensors
Pressure, Strain

Rotation Sensors
Encoders

...

48

No Good Data Left Behind

5 Quintillion

bytes of data produced every day by IoT

<1%

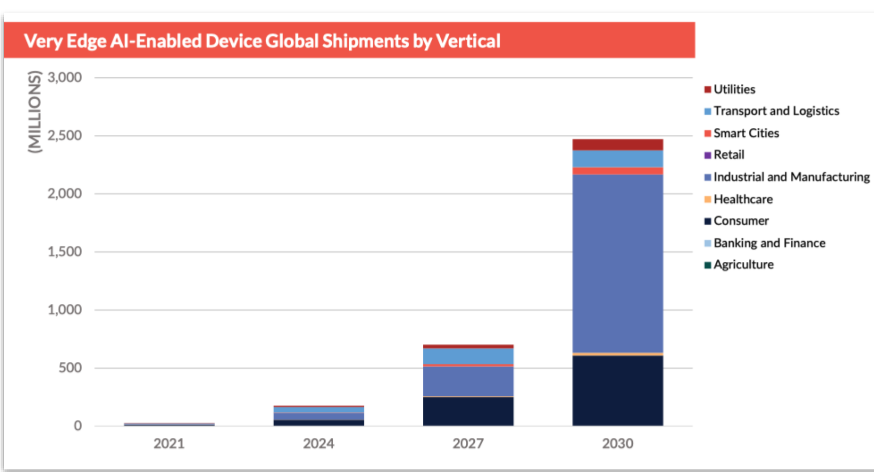
of unstructured data is analyzed or used at all

Source: Harvard Business Review, What's Your Data Strategy?, April 18, 2017
Cisco, Internet of Things (IoT) Data Continues to Explode Exponentially.
Who Is Using That Data and How?, Feb 5, 2018

49

49

Market Forecast



Source: ABI Research: TinyML

50

50

>250 Billion

MCUs today

51

51

Comparing Power

BIG
GPU / CPU

300W
NVIDIA Tesla K80

SMALL

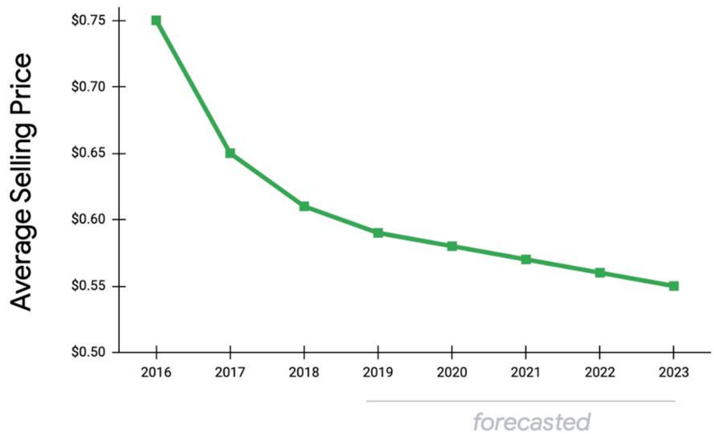
3.64W
Apple A12

Neural Decision Processor
Always-on deep learning speech/audio recognition
Ultra low power, 128KB SRAM, 12-pin, 2.52mm^2

140 μW
Syntiant NDP100

52

MCU Pricing Forecast



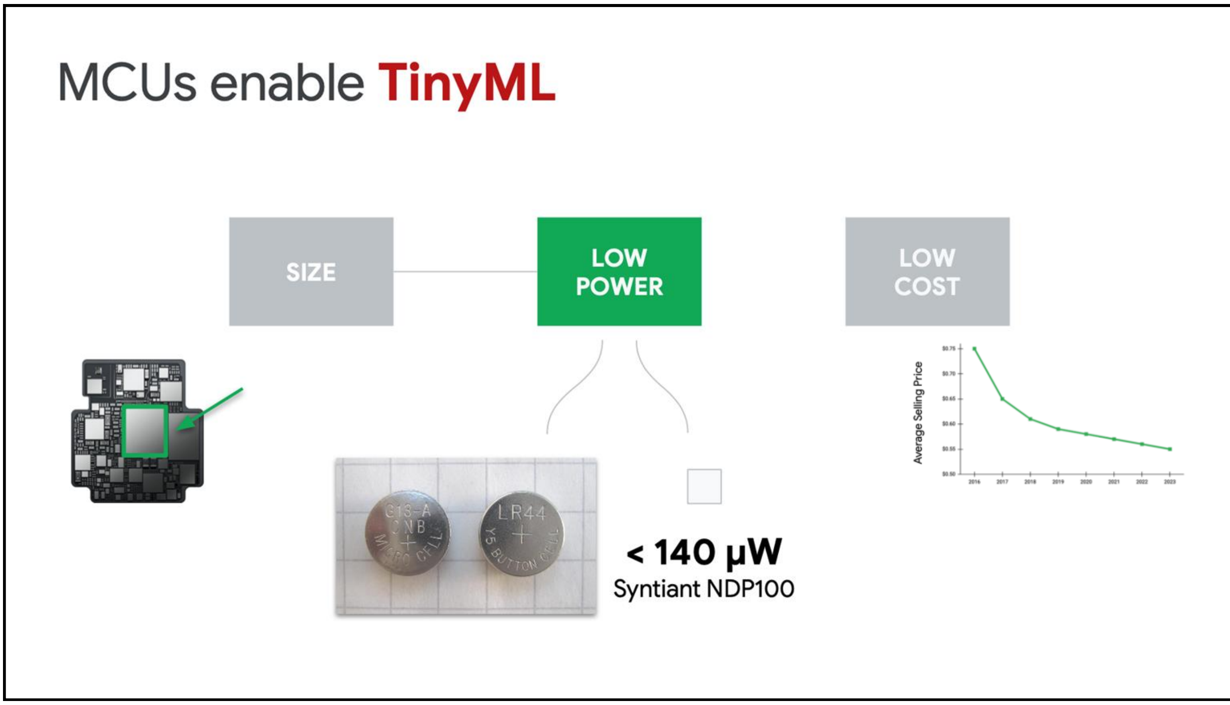
Source: IC Insights

53

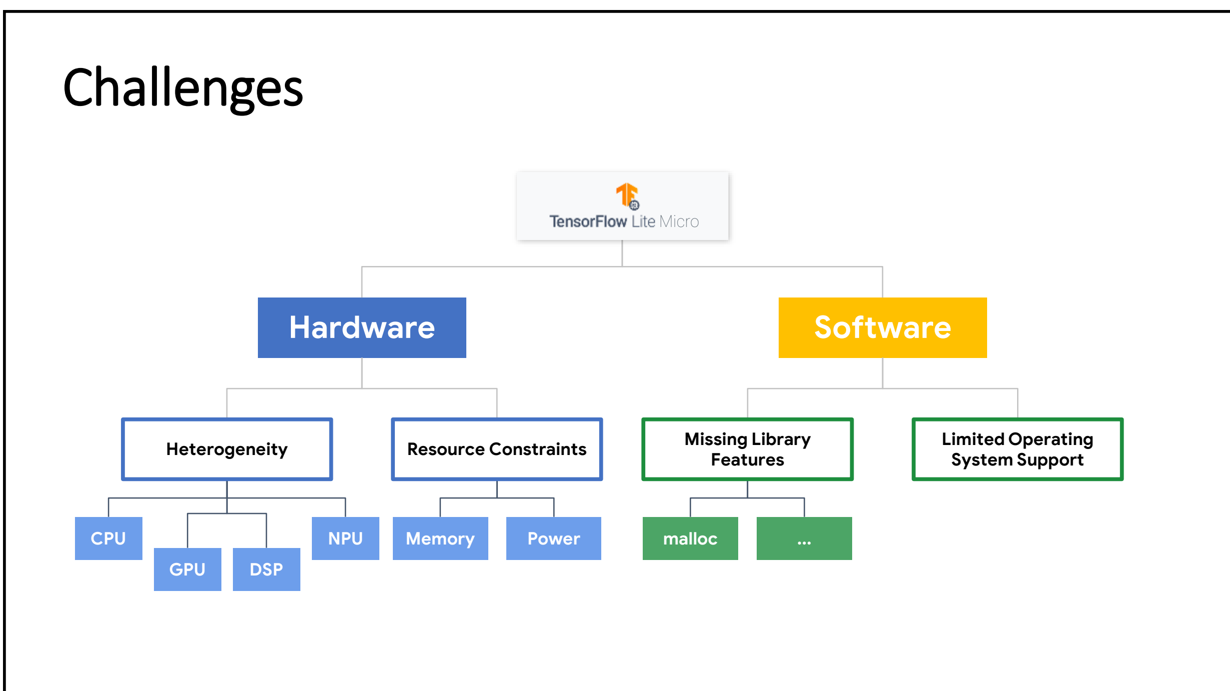
	Board	MCU / ASIC	Clock	Memory	Sensors	Radio
	Himax WE-I Plus EVB	HX6537-A 32-bit EM9D DSP	400 MHz	2MB flash 2MB RAM	Accelerometer, Mic, Camera	None
	Arduino Nano 33 BLE Sense	32-bit nRF52840	64 MHz	1MB flash 256kB RAM	Mic, IMU, Temp, Humidity, Gesture, Pressure, Proximity, Brightness, Color	BLE
	SparkFun Edge 2	32-bit ArtemisV1	48 MHz	1MB flash 384kB RAM	Accelerometer, Mic, Camera	BLE
	Espressif EYE	32-bit ESP32-D0WD	240 MHz	4MB flash 520kB RAM	Mic, Camera	WiFi, BLE

54

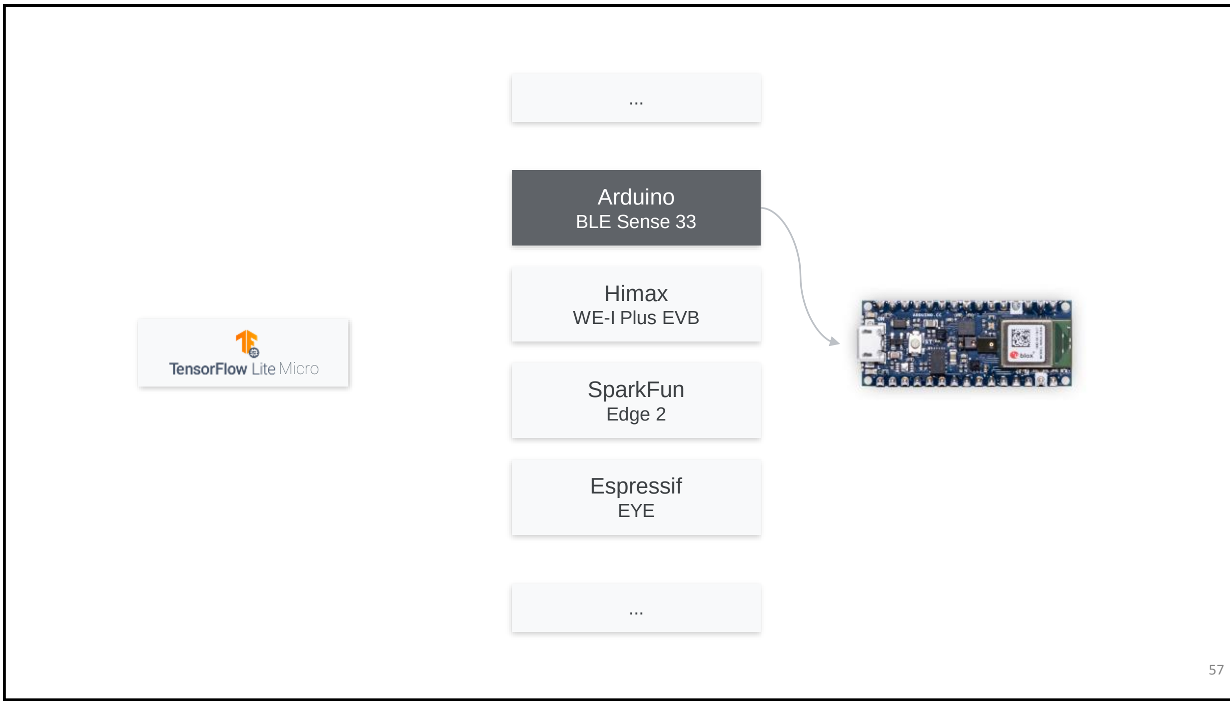
54


MCUs enable TinyML
SIZE
LOW POWER
LOW COST
Average Selling Price
< 140 µW
Syntiant NDP100

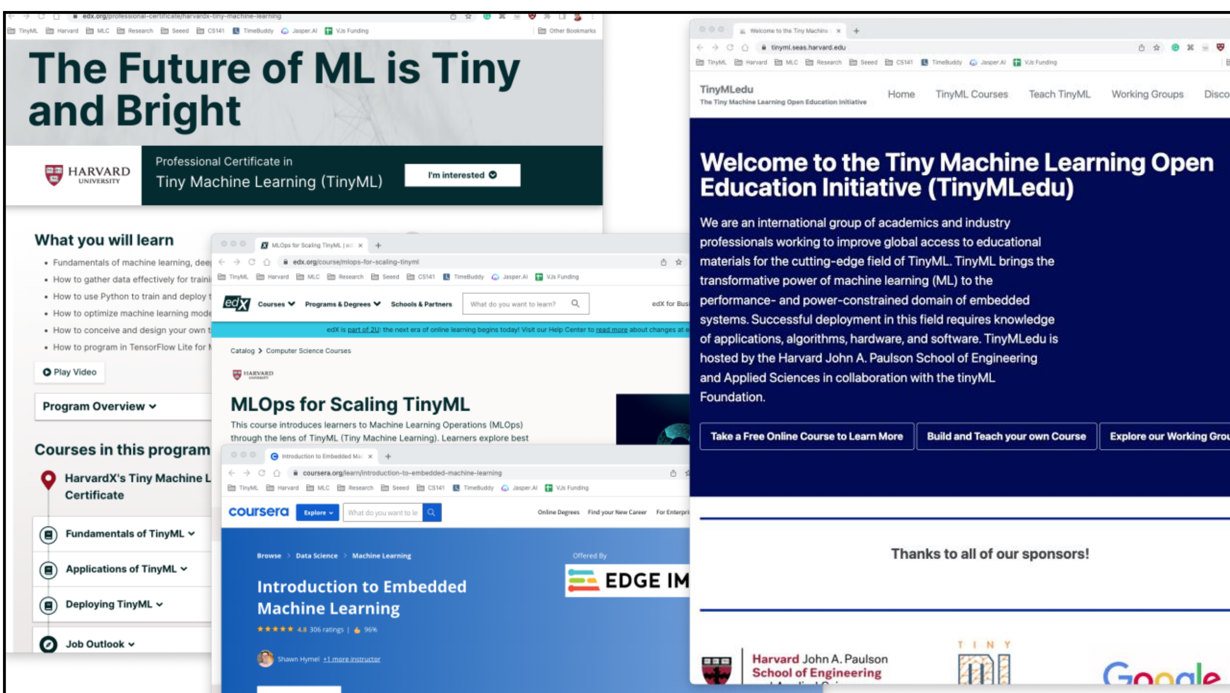
55


Challenges
TensorFlow Lite Micro
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Power
malloc
...

56


...
Arduino
BLE Sense 33
Himax
WE-I Plus EVB
SparkFun
Edge 2
Espressif
EYE
...
TensorFlow Lite Micro
57

57


The Future of ML is Tiny and Bright
HARVARD UNIVERSITY
Professional Certificate in
Tiny Machine Learning (TinyML)
I'm interested
What you will learn
Play Video
Program Overview
Courses in this program
HarvardX's Tiny Machine L
Certificate
Fundamentals of TinyML
Applications of TinyML
Deploying TinyML
Job Outlook
MLOps for Scaling TinyML
This course introduces learners to Machine Learning Operations (MLOps) through the lens of TinyML (Tiny Machine Learning). Learners explore best
coursera
Introduction to Embedded Machine Learning
EDGE IM
Enroll for Free
TinyMLedu
The Tiny Machine Learning Open Education Initiative
Home
TinyML Courses
Teach TinyML
Working Groups
Welcome to the Tiny Machine Learning Open Education Initiative (TinyMLedu)
We are an international group of academics and industry professionals working to improve global access to educational materials for the cutting-edge field of TinyML. TinyML brings the transformative power of machine learning (ML) to the performance- and power-constrained domain of embedded systems. Successful deployment in this field requires knowledge of applications, algorithms, hardware, and software. TinyMLedu is hosted by the Harvard John A. Paulson School of Engineering and Applied Sciences in collaboration with the tinyML Foundation.
Take a Free Online Course to Learn More
Build and Teach your own Course
Explore our Working Group
Thanks to all of our sponsors!
Harvard John A. Paulson School of Engineering

59

Final Notes

- This is a rapidly changing field!
- If you take a break of a few months from it, you will have a lot of catch up to do.
- Expect changes, differences from TextBook, and often “upgrades”!

60

60

Credits

- A previous edition of this course was developed in collaboration with Dr. Susan C. Schneider of Marquette University.
- We are very grateful and thank all the following professors, researchers, and practitioners for jump-starting courses on TinyML and for sharing their teaching materials:
- Prof. Marcelo Rovai - TinyML - Machine Learning for Embedding Devices, UNIFEI
  - https://github.com/Mjrovai/UNIFEI-IESTI01-TinyML-2022.1
- Prof. Vijay Janapa Reddi - CS249r: Tiny Machine Learning, Applied Machine Learning on Embedded IoT Devices, Harvard
  - https://sites.google.com/g.harvard.edu/tinyml/home
- Prof. Rahul Mangharam – ESE3600: Tiny Machine Learning, Univ. of Pennsylvania
  - https://tinyml.seas.upenn.edu/#
- Prof. Brian Plancher - Harvard CS249r: Tiny Machine Learning (TinyML), Barnard College, Columbia University
  - https://a2r-lab.org/courses/cs249r_tinyml/

61

61

References

- Additional references from where information and other teaching materials were gathered include:
- Applications & Deploy textbook: “TinyML” by Pete Warden, Daniel Situnayake
  - https://www.oreilly.com/library/view/tinyml/9781492052036/
- Deploy textbook “TinyML Cookbook” by Gian Marco Iodice
  - https://github.com/PacktPublishing/TinyML-Cookbook
- Jason Brownlee
  - https://machinelearningmastery.com/
- TinyMLedu
  - https://tinyml.seas.harvard.edu/
- Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
  - https://www.edx.org/professional-certificate/harvardx-tiny-machine-learning
- Introduction to Embedded Machine Learning - Coursera/Edge Impulse
  - https://www.coursera.org/learn/introduction-to-embedded-machine-learning
- Computer Vision with Embedded Machine Learning - Coursera/Edge Impulse
  - https://www.coursera.org/learn/computer-vision-with-embedded-machine-learning

62