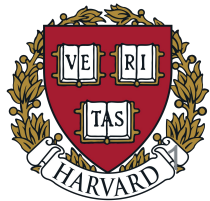
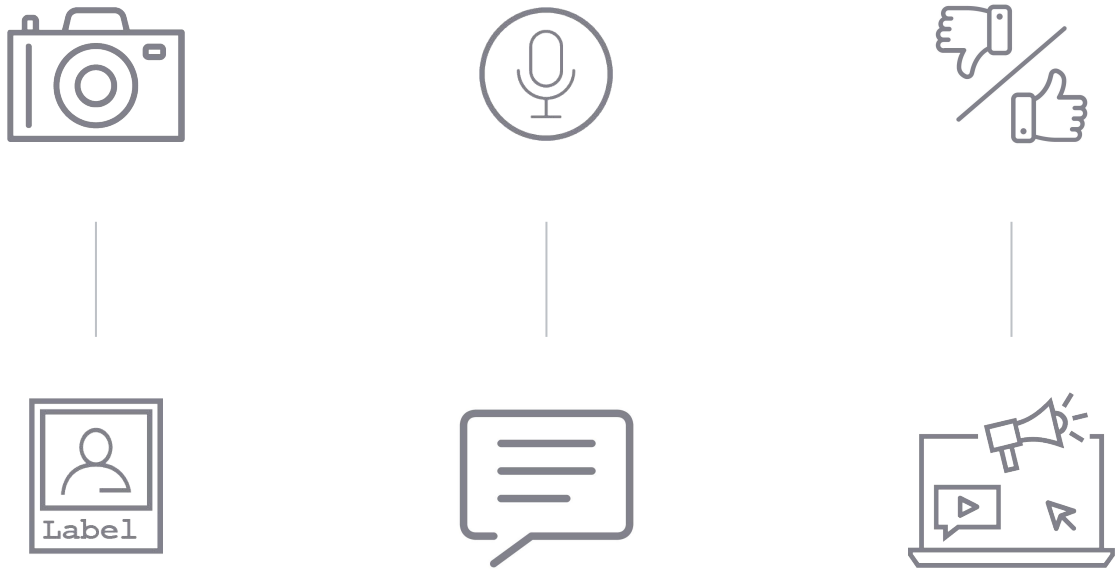
TinyML:
The Future of Machine Learning is Tiny and Bright

***Vijay Janapa Reddi**, Ph. D. | Associate Professor |*
John A. Paulson School of Engineering and Applied Sciences | Harvard University |
Web: http://scholar.harvard.edu/vijay-janapa-reddi

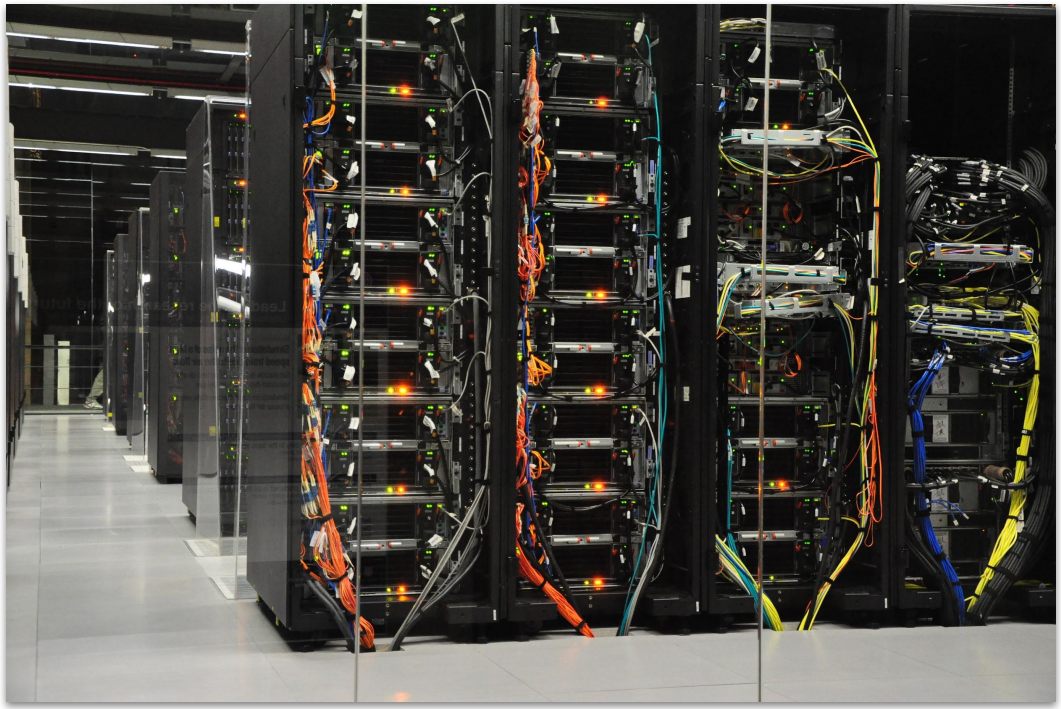
African Regional Workshop on SciTinyML: Scientific Use of Machine Learning on Low-Power Devices



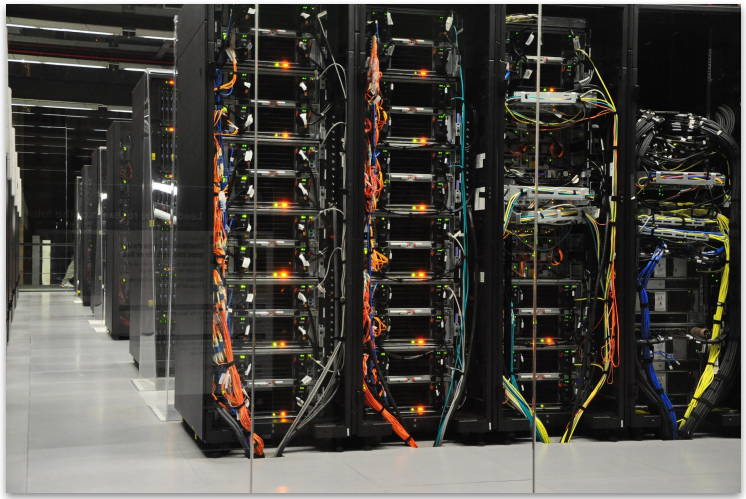
Applications of Machine Learning



Cloud

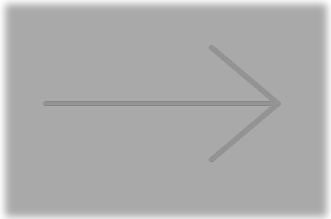


Cloud



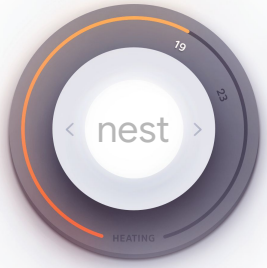
Mobile

Mobile

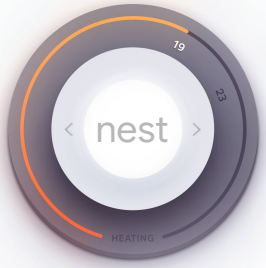


Google Assistant

Google Assistant



IoT 1.0: Internet of Things

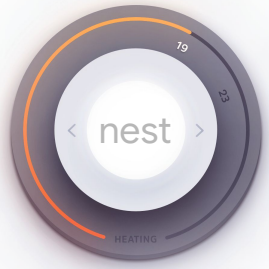


IoT 2.0: Intelligence on Things

Google Assistant

IoT 2.0: Intelligence on Things

Bandwidth
Reliability
Latency
Privacy
Energy



What Is TinyML

TinyML

What is Tiny Machine Learning (**TinyML**)?

TinyML

What is Tiny Machine Learning (**TinyML**)?

TinyML

Fastest-growing field of **ML**

What is Tiny Machine Learning (**TinyML**)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

What is Tiny Machine Learning (**TinyML**)?

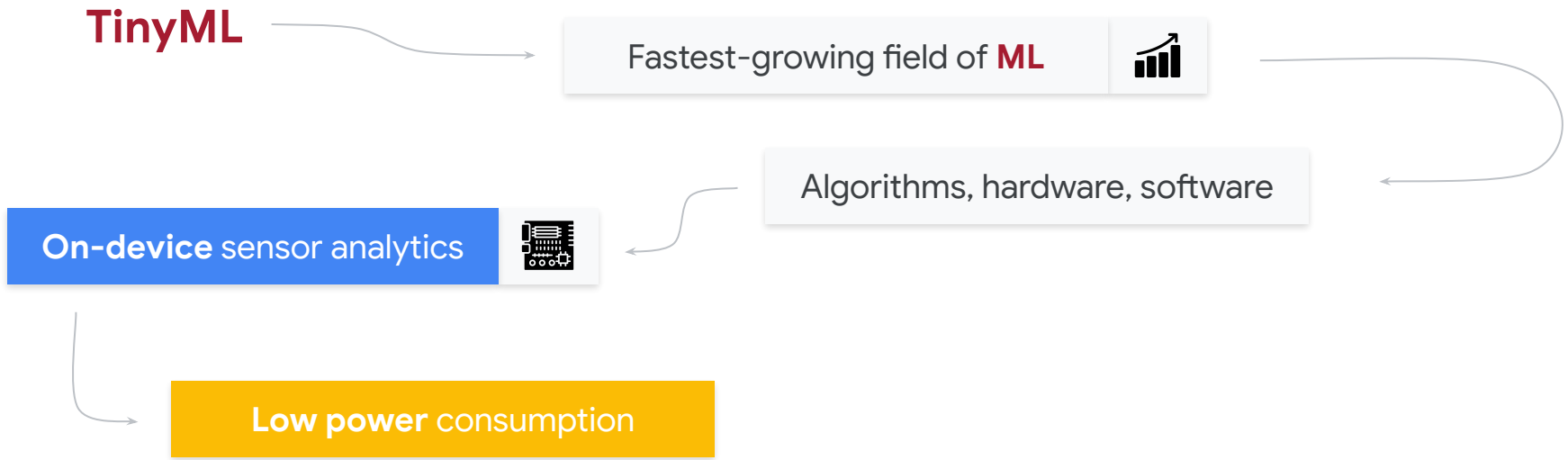
TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics

What is Tiny Machine Learning (**TinyML**)?



What is Tiny Machine Learning (**TinyML**)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics

Low power consumption

Always-on ML

What is Tiny Machine Learning (**TinyML**)?

TinyML

Fastest-growing field of **ML**

Algorithms, hardware, software

On-device sensor analytics

Low power consumption

Always-on ML

Battery-operated

Kicking

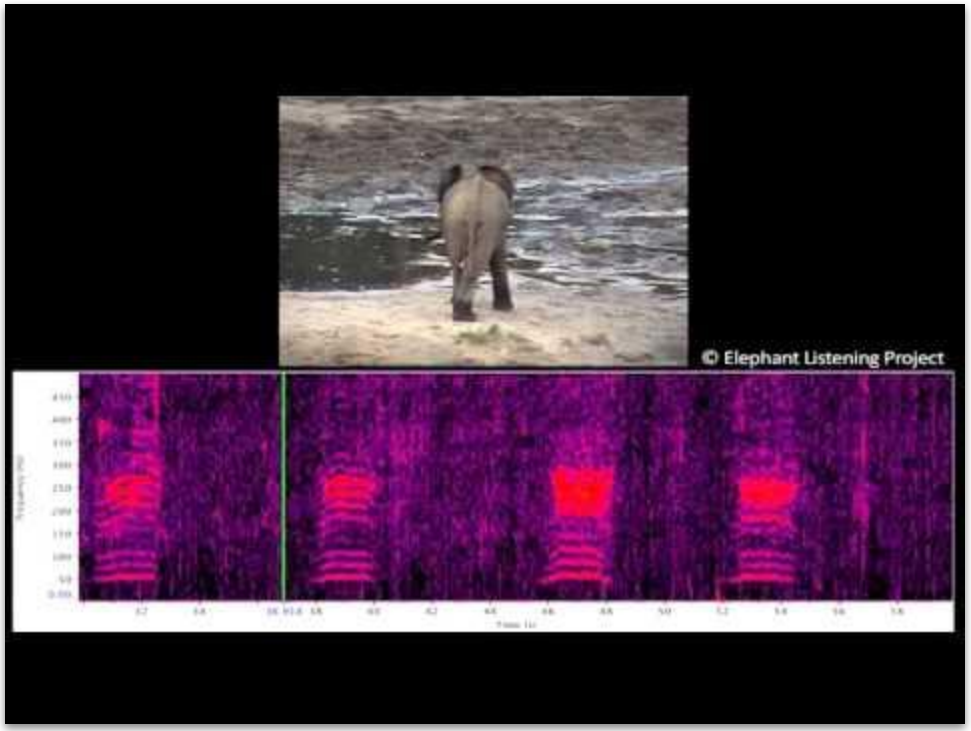
Penalty kicking

Passing

Dribbling

...

Wildlife Conservation



ElephantEdge

Risk Monitoring

"Know when an elephant is moving into a high-risk area and send real-time notifications to park rangers."

Conflict Monitoring

"Sense and alert when an elephant is heading into an area where farmers live."

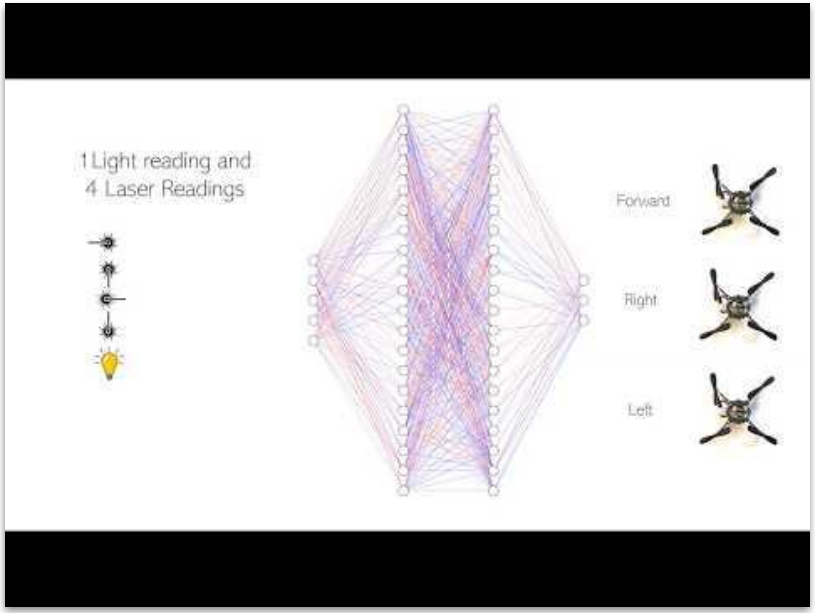
Activity Monitoring

"Classify the general behavior of the elephant, such as when it is drinking, eating, sleeping, etc."

Communication Monitoring

"Listen for vocal communications between elephants via the onboard microphone."

TinyRL: Autonomous Navigation on Nano Drone



[ICRA'21]

[IROS'21]

Meet TinyML: The Latest Mach

forbes.com/sites/sap/2021/11/08/meet-tinyml-the-latest-machine-learning-tech-having-a...

tinyML Google MLC Research TimeBuddy - CESMII – The S... AI Measurement a... Data Centric AI W... Other Bookmarks Reading List

Forbes

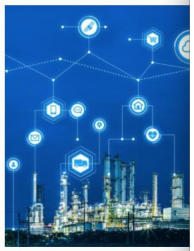
Subscribe Sign In

Nov 8, 2021, 08:30am EST | 15,696 views

Meet TinyML: The Latest Machine Learning Tech Having An Outsize Business Impact

Dr. Nicholas Nicoloudis Brand Contributor
SAP BRANDVOICE | Paid Program
Innovation

As device sensors proliferate across
product development through insp
surfacing to provide actionable insi
There are sound economic reasons
researchers predict IoT will have a
trillion by 2025, identifying manufa
trillion).



The rise of tinyML to collect data from edge de
explosion of sensors in pretty much every indu

The tinyML community was establi
learning architectures, techniques,
on-device analytics for a variety of
chemical, and others) at low power
devices. One of the tinyML founder

*"..we are in the midst of the digital
ultimate benefits of extreme energ
intelligence and analytics at low c
features...".*

EETimes - TinyML Sees Big H

eetimes.com/tinyml-se...

tinyML Google MLC Other Bookmarks

EETimes

Machine learning at the edge:

zdnet.com/article/machine-learning-at-the-edge-tinyml-is-getting-big/

tinyML Google MLC Research TimeBuddy - CESMII – The S... AI Measurement a... Data Centric AI W...

ZDNet Windows 11 5G Best VPNs Cloud Security AI Newsletters Blogs More Edition: US

MUST READ: Log4j flaw: Now state-backed hackers are using bug as part of attacks,

Machine learning at the edge: Tin
getting big

Being able to deploy machine learning applications at the edge is the key to unlocking
TinyML is the art and science of producing machine learning models frugal enough to
rapid growth.

Written by **George Anadiotis**, Contributing Writer
Posted in Big on Data on June 7, 2021 | Topic: Big Data

Is it $61 billion and 38.4% CAGR by 2028 or $43 billion and 37.4% CAGR by 2027? Depends on which report outlining the growth of edge computing you choose to go by, but in the end it's not that different.

What matters is that edge computing is booming. There is growing interest by vendors, and ample coverage, for good reason. Although the definition of what constitutes edge computing is a bit fuzzy, the idea is simple. It's about taking compute out of the data center, and bringing it as close to where the action is as possible.

Whether it's stand-alone IoT sensors, devices of all kinds, drones, or autonomous vehicles, there's one thing in common. Increasingly, data generated at the edge are used to feed applications powered by machine learning models. There's just one problem: machine learning models were never designed to be deployed at the edge. Not until now, at least. Enter TinyML.

EXECUTIVE GUIDE

What is machine learning? Everything you need to

Tiny machine learning (TinyML) is broadly defined as a fast growing

semiengineering.com...

tinyML Google MLC Research TimeBuddy Other Bookmarks Reading List

SEMICONDUCTOR ENGINEERING

How TinyML is powering big id

cio.com/article/188995/how-tinyml-is-powering-big-ideas-across-critical-industries.html

tinyML Google MLC Research TimeBuddy - CESMII – The S... AI Measurement a... Data Centric AI W... Other Bookmarks Reading List

CIO DIGITAL ISSUE AWARDS EVENTS CIO THINK TANK NEWSLETTERS RESOURCES INSIDER SIGN IN REGISTER

Hot Topics IT Leadership Digital Transformation Innovation Data Analytics & AI Enterprise Applications Diversity and Inclusion

Home

SAP **NEXT EVOLUTION OF MACHINE LEARNING IS UPON US**

SPONSORED

How TinyML is powering big ideas across critical industries

BrandPost Sponsored by SAP | Learn More | JUL 18, 2021 4:31 PM PDT

shutterstock

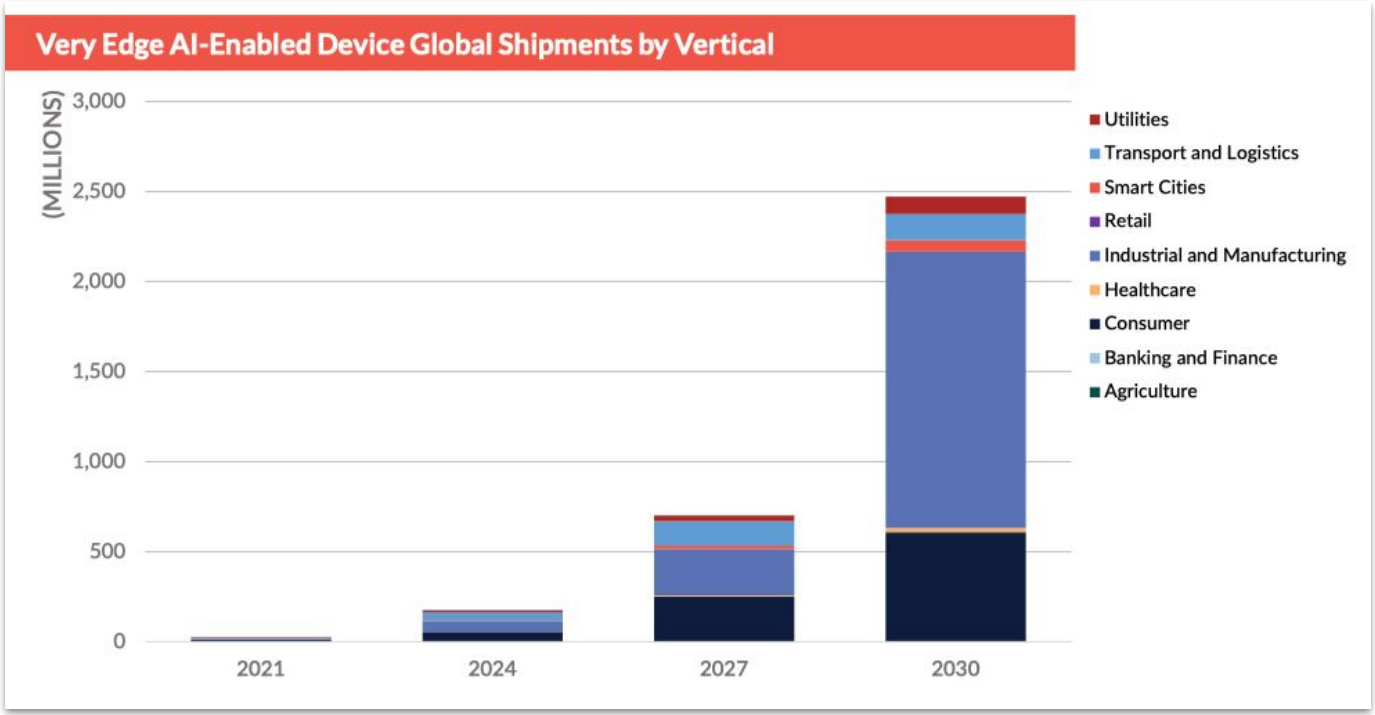
From cars and TVs to lightbulbs and doorbells. So many of the objects in everyday life have 'smart' functionality because the manufacturers have built chips into them.

But what if you could also run machine learning models in something as small as a golf ball dimple? That's the reality that's being enabled by TinyML, a broad movement to run tiny machine learning algorithms on embedded devices, or those with

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/

Market Forecast



Source: ABI Research: TinyML

TinyML Is All About **Sensor Data Intelligence**

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

...

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

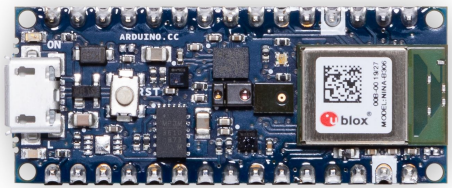
Source: Harvard Business Review, What's Your Data Strategy?, April 18, 2017
Cisco, Internet of Things (IoT) Data Continues to Explode Exponentially. Who Is Using That Data and How?, Feb 5, 2018

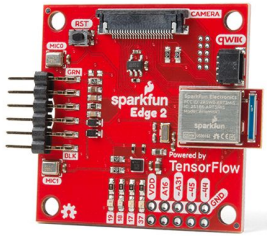
TinyML Challenges

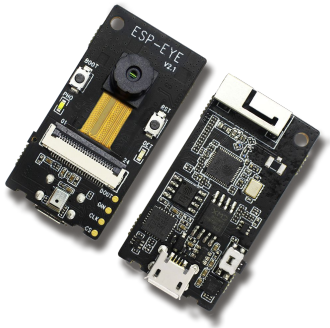
250 Billion

MCUs today


Himax

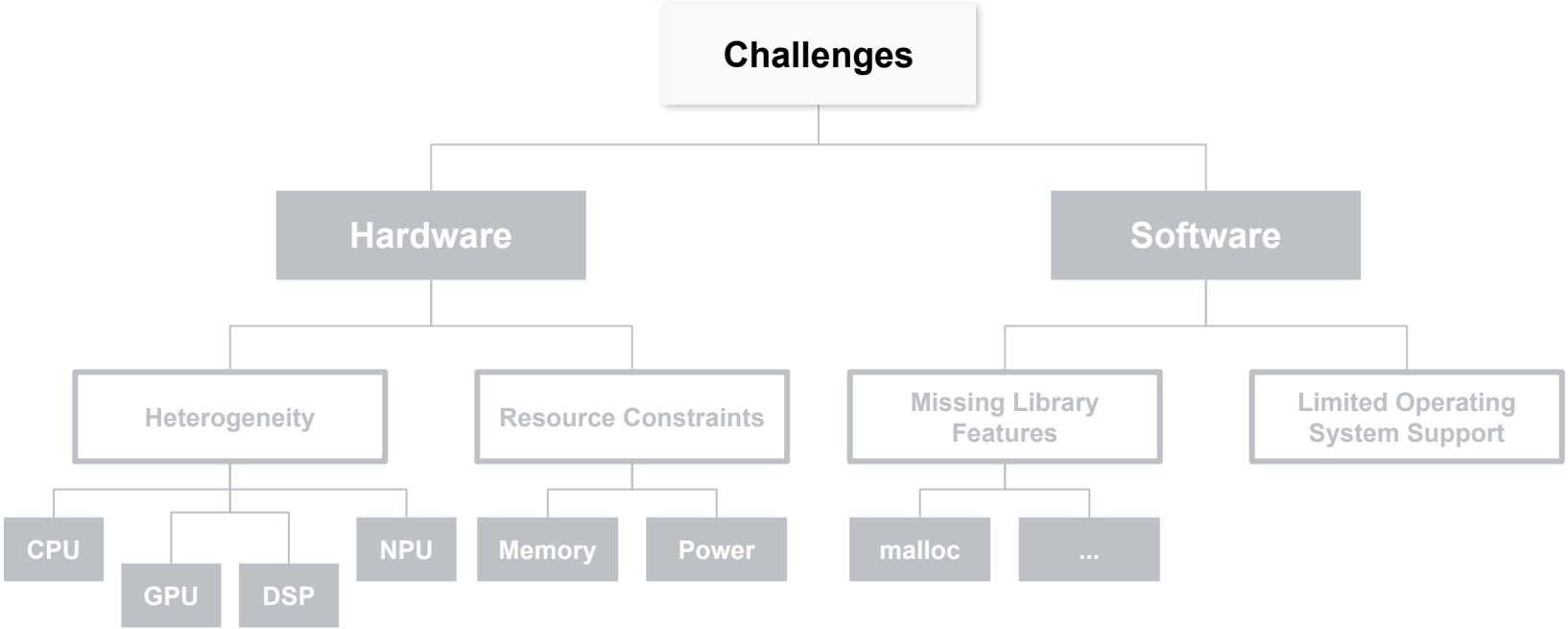

ARDUINO.CC
ON
RST
u-blox

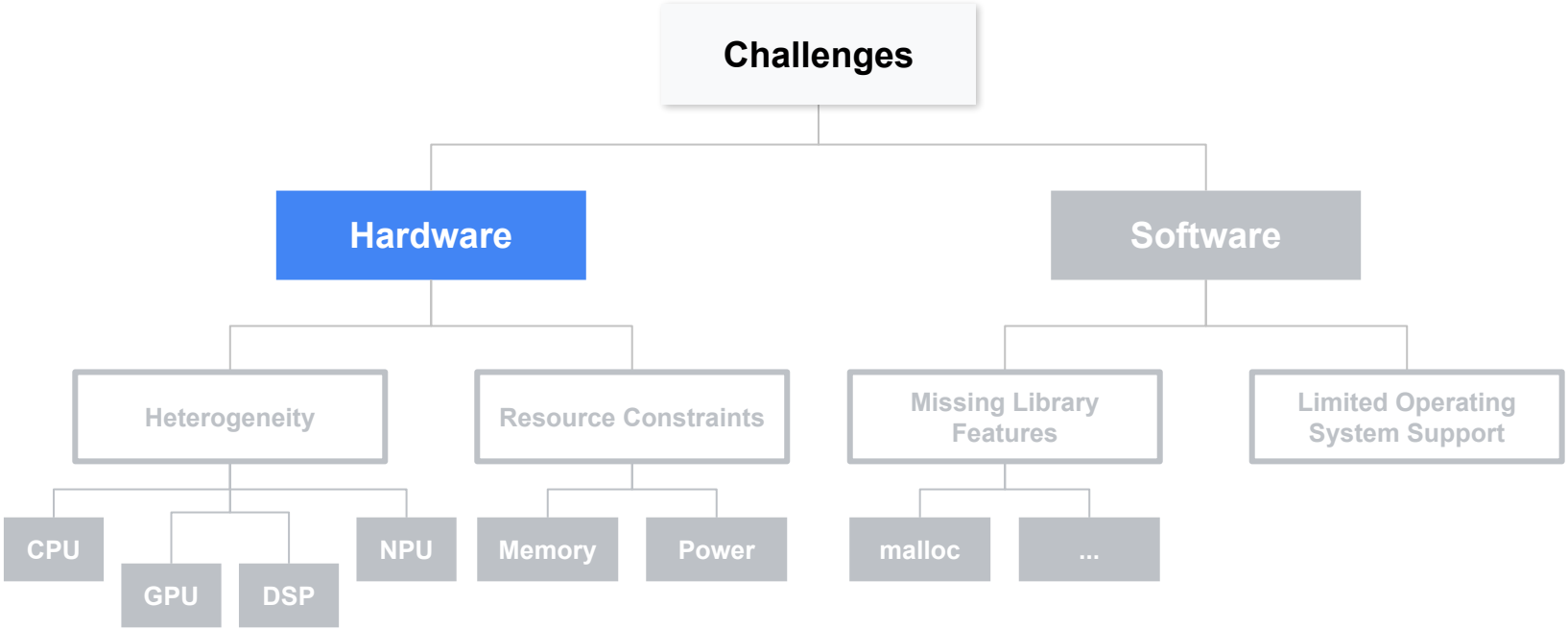

RST
CAMERA
QWIIC
sparkfun
Edge 2
Powered by
TensorFlow

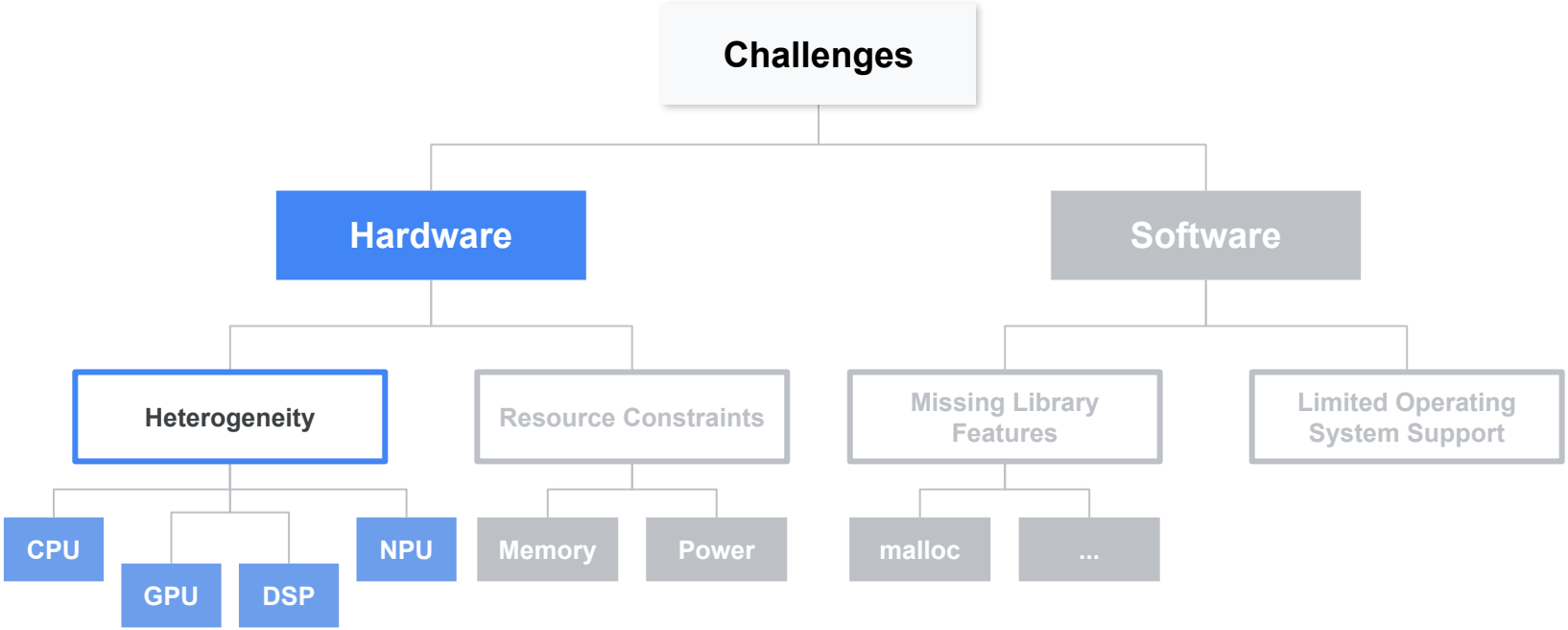

ESP-EYE
v2.1

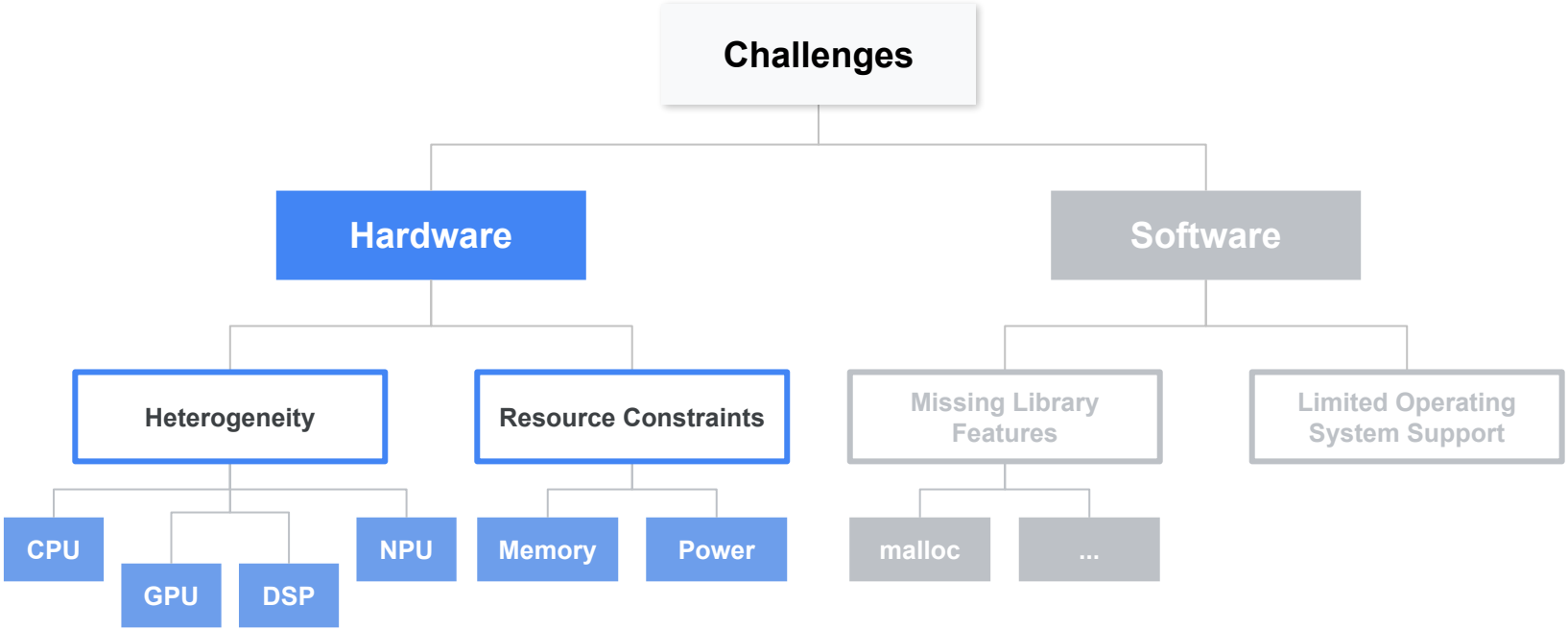
	Board	MCU / ASIC	Clock	Memory	Sensors	Radio
	Himax WE-I Plus EVB	HX6537-A 32-bit EM9D DSP	400 MHz	2MB flash 2MB RAM	Accelerometer, Mic, Camera	None
	Arduino Nano 33 BLE Sense	32-bit nRF52840	64 MHz	1MB flash 256kB RAM	Mic, IMU, Temp, Humidity, Gesture, Pressure, Proximity, Brightness, Color	BLE
	SparkFun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

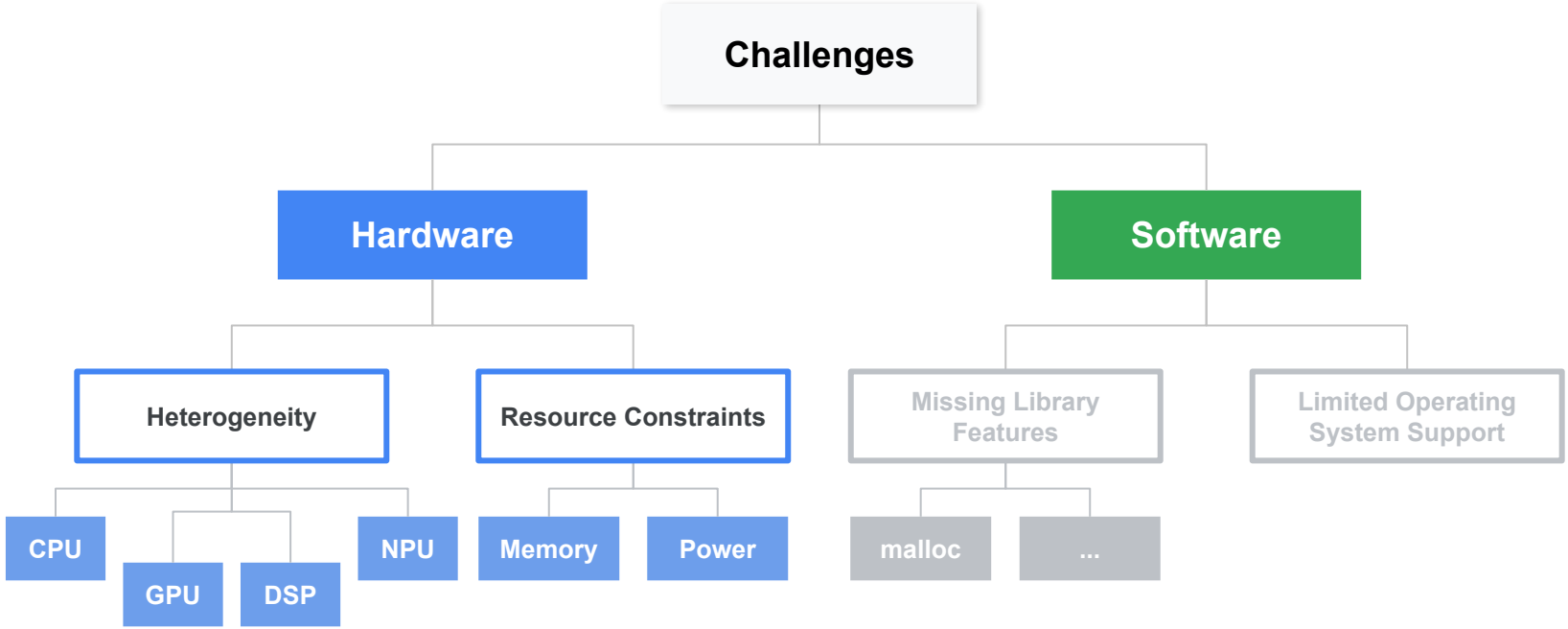
Challenges

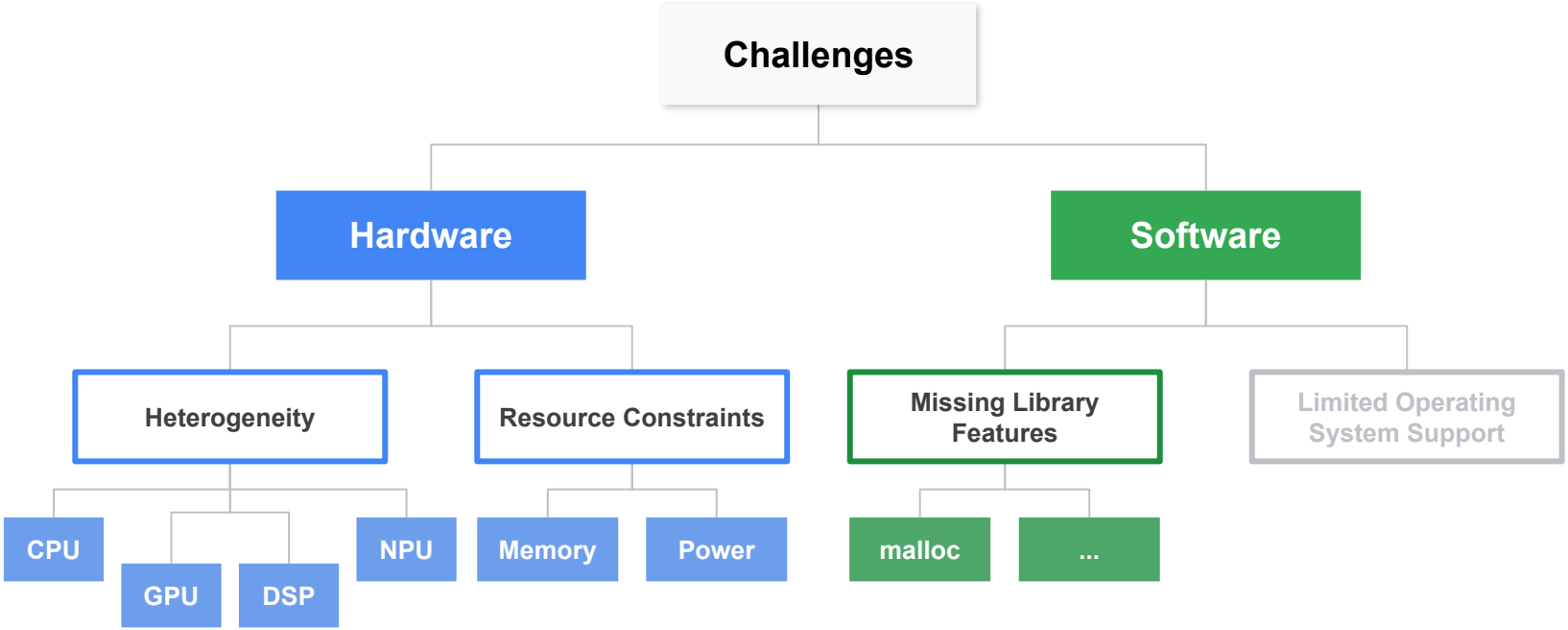

Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Power
malloc
...

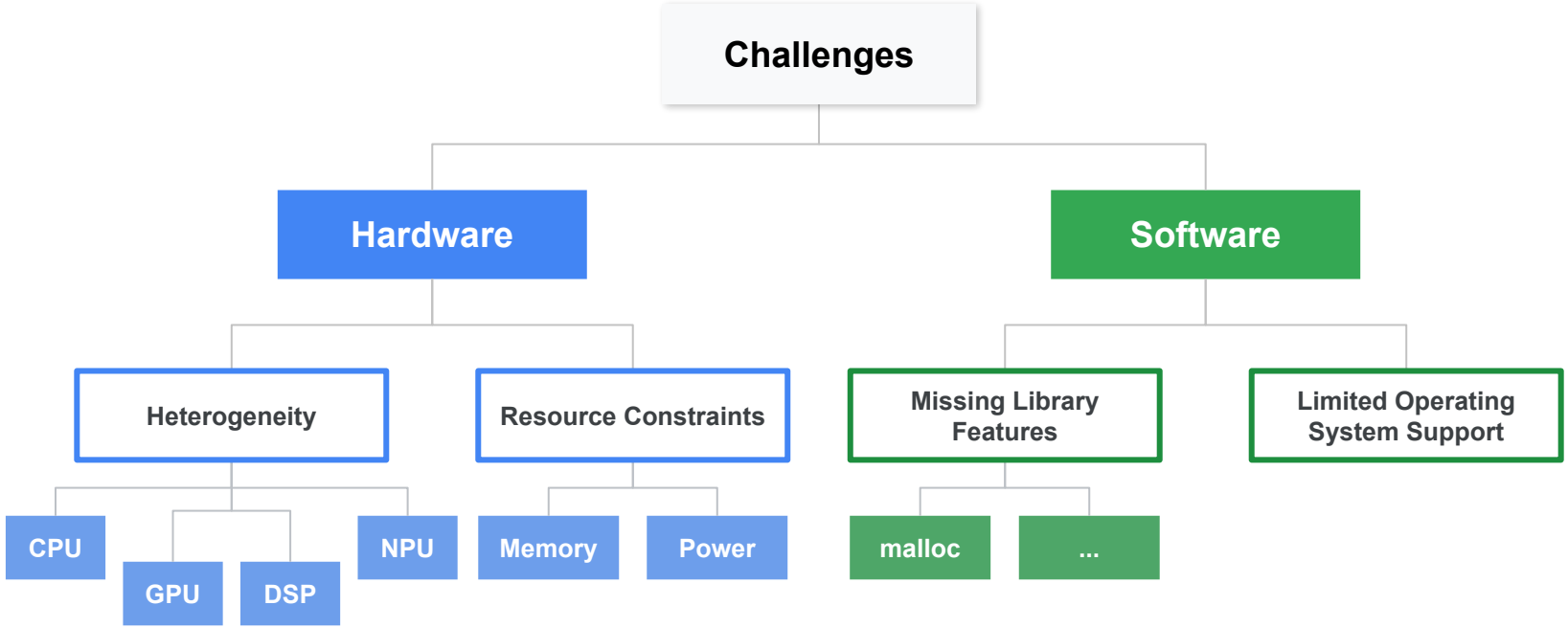

Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Power
malloc
...

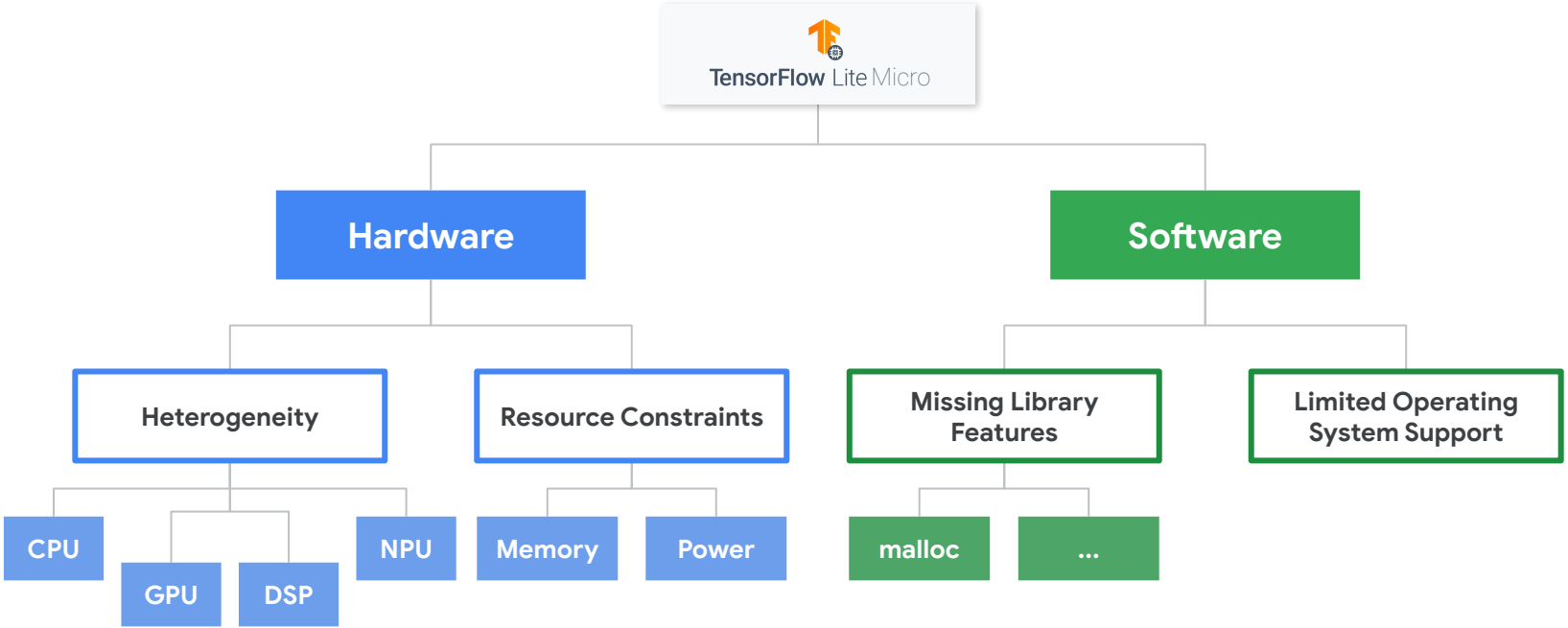

Challenges
Hardware
Software
Heterogeneity
Resource Constraints
Missing Library Features
Limited Operating System Support
CPU
GPU
DSP
NPU
Memory
Power
malloc
...


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


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


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


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


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

TensorFlow Lite Micro

...

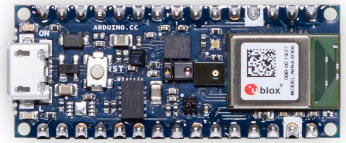
Arduino
BLE Sense 33

Himax
WE-I Plus EVB

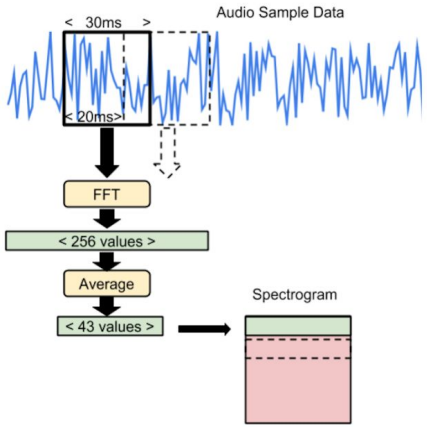
SparkFun
Edge 2

Espressif
EYE

...



Keyword Spotting



Warden, Pete. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209* (2018).

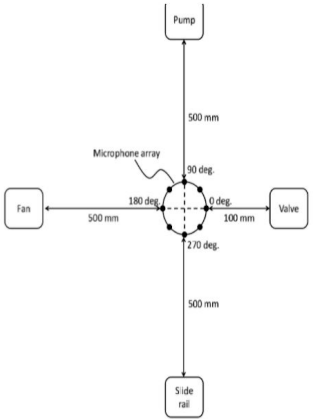
Visual Wake Words

(a) 'Person'

(b) 'Not-person'

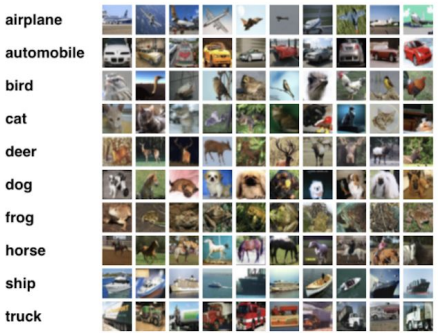
Chowdhery, Aakanksha, et al. "Visual wake words dataset." *arXiv preprint arXiv:1906.05721* (2019).

Anomaly Detection



Purohit, Harsh, et al. "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection." *arXiv preprint arXiv:1909.09347* (2019).

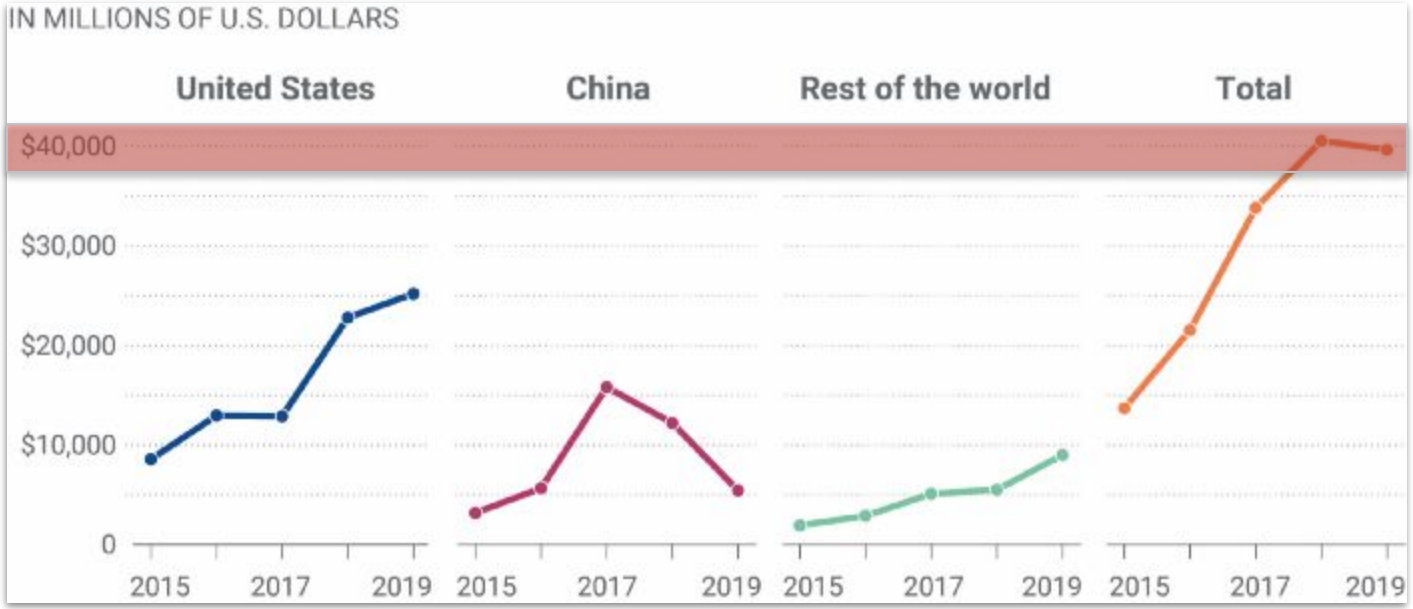
Image Classification



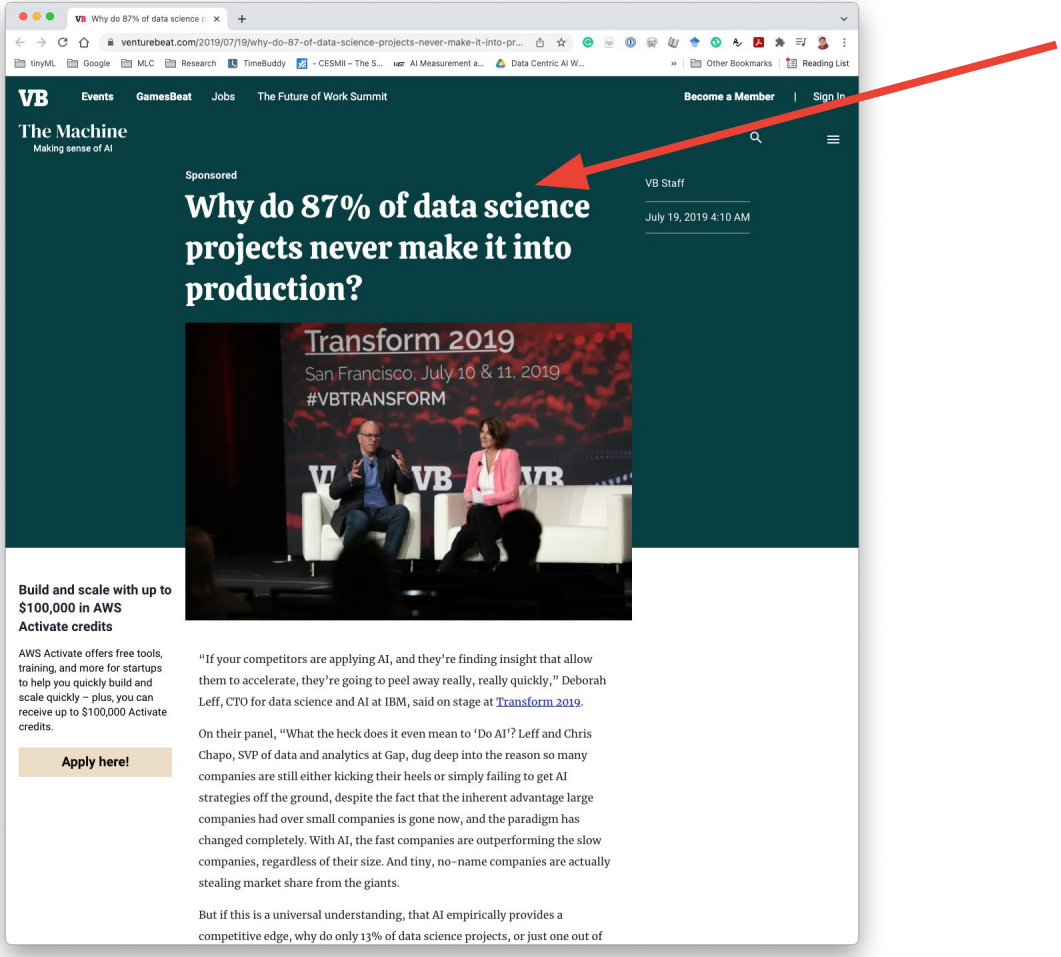
Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

Scaling TinyML

AI Investments



Source: Brookings Tech Stream

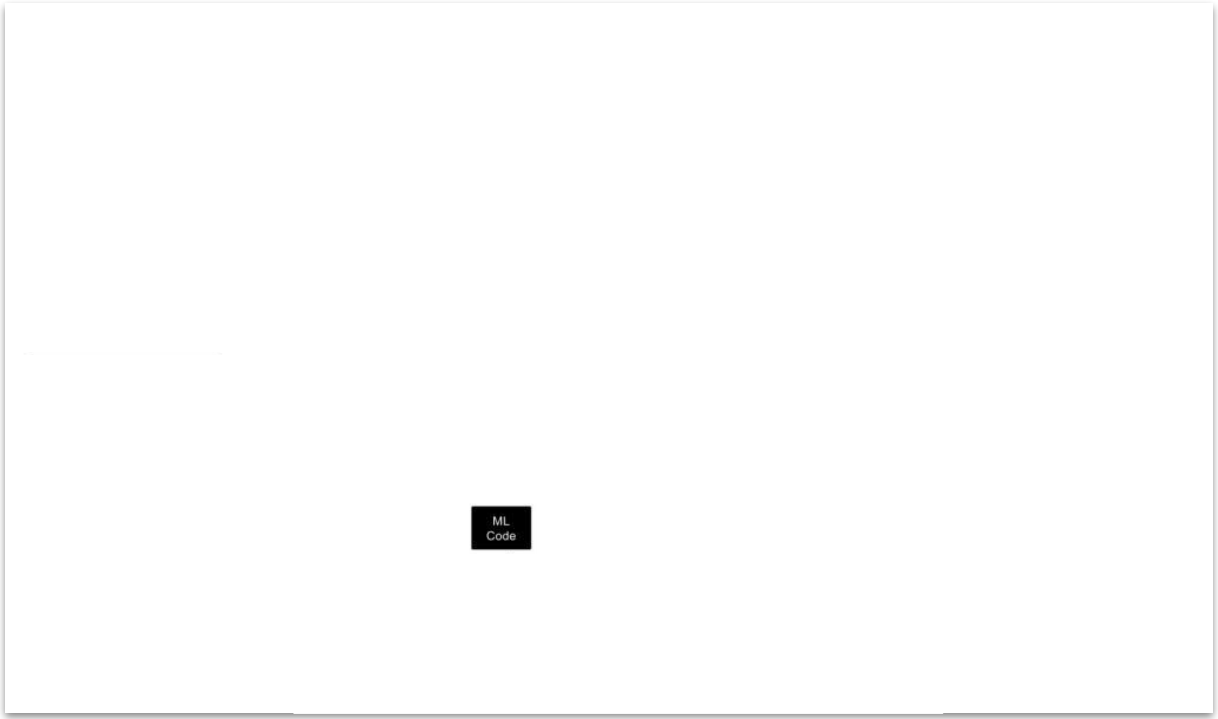

Why do 87% of data science p
venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-pr...
tinyML
Google
MLC
Research
TimeBuddy
- CESMII – The S...
AI Measurement a...
Data Centric AI W...
Other Bookmarks
Reading List
VB
Events
GamesBeat
Jobs
The Future of Work Summit
Become a Member
Sign In
The Machine
Making sense of AI
Sponsored
Why do 87% of data science projects never make it into production?
VB Staff
July 19, 2019 4:10 AM
Transform 2019
San Francisco, July 10 & 11, 2019
#VBTRANSFORM
Build and scale with up to $100,000 in AWS Activate credits
AWS Activate offers free tools, training, and more for startups to help you quickly build and scale quickly – plus, you can receive up to $100,000 Activate credits.
Apply here!
"If your competitors are applying AI, and they're finding insight that allow them to accelerate, they're going to peel away really, really quickly," Deborah Leff, CTO for data science and AI at IBM, said on stage at Transform 2019.
On their panel, "What the heck does it even mean to 'Do AI'? Leff and Chris Chapo, SVP of data and analytics at Gap, dug deep into the reason so many companies are still either kicking their heels or simply failing to get AI strategies off the ground, despite the fact that the inherent advantage large companies had over small companies is gone now, and the paradigm has changed completely. With AI, the fast companies are outperforming the slow companies, regardless of their size. And tiny, no-name companies are actually stealing market share from the giants.
But if this is a universal understanding, that AI empirically provides a competitive edge, why do only 13% of data science projects, or just one out of

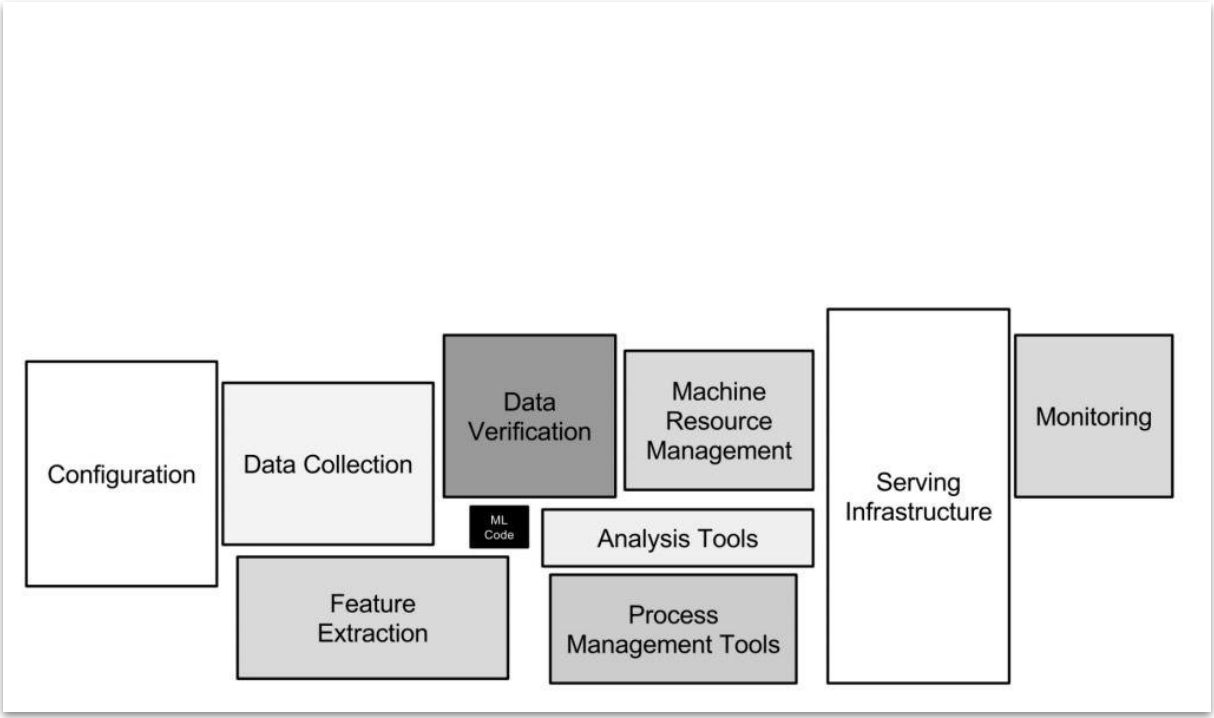
Predicts 2019: Analytics and BI Solutions

- Through 2020, 80% of AI projects will remain alchemy, run by wizards whose talents will not scale in the organization.
- Through 2022, only 20% of analytic insights will deliver business outcomes.
- By 2021, proof-of-concept analytic projects using quantum computing infrastructure will have outperformed traditional analytic approaches in multiple domains by at least a factor of 10

Source: https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

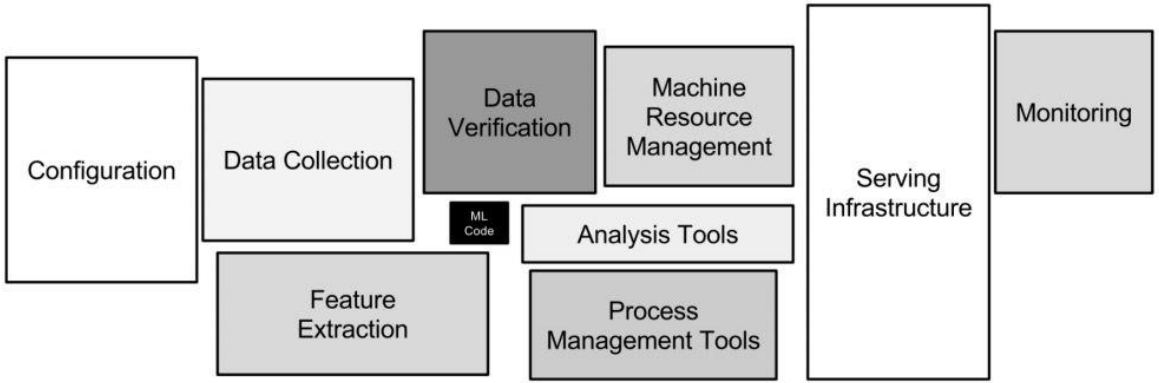
Let's quantify this a bit. In 2019 alone, approximately **USD 40 billions** were invested into privately held AI companies. If we extrapolate this and throw the approximated success rate of AI projects into these figures (and completely exclude intracompany ML investments), we reach the conclusion that in 2019, around **USD 38 billions were wasted due to unsuccessful Machine Learning projects.**


ML
Code

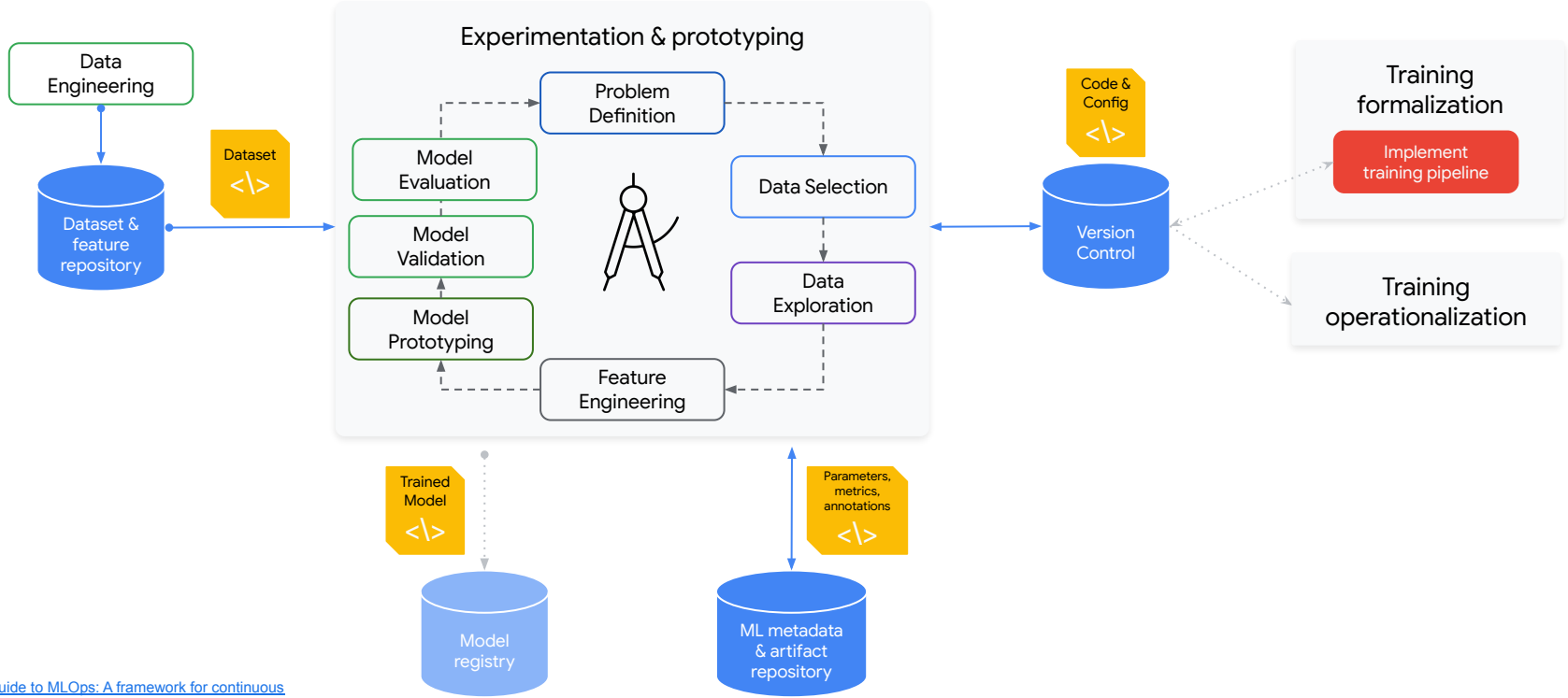

Configuration
Data Collection
Data Verification
Machine Resource Management
Serving Infrastructure
Monitoring
ML Code
Analysis Tools
Feature Extraction
Process Management Tools

Hidden Technical Debt in Machine Learning Systems

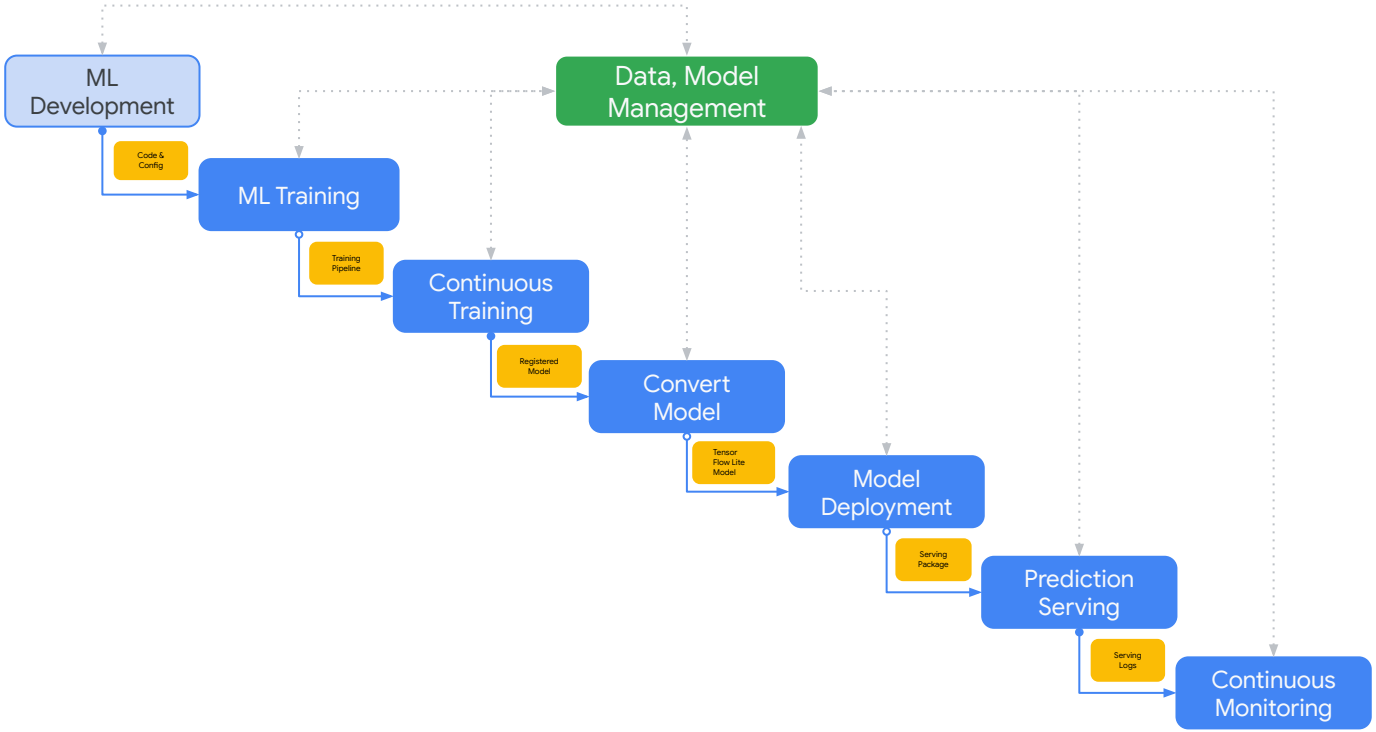
D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.



MLCode



Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning


ML Development
Code & Config
ML Training
Training Pipeline
Continuous Training
Registered Model
Convert Model
Tensor Flow Lite Model
Model Deployment
Serving Package
Prediction Serving
Serving Logs
Continuous Monitoring
Data, Model Management

ML Engineer

ML Researcher

Data Scientist

Data Engineer

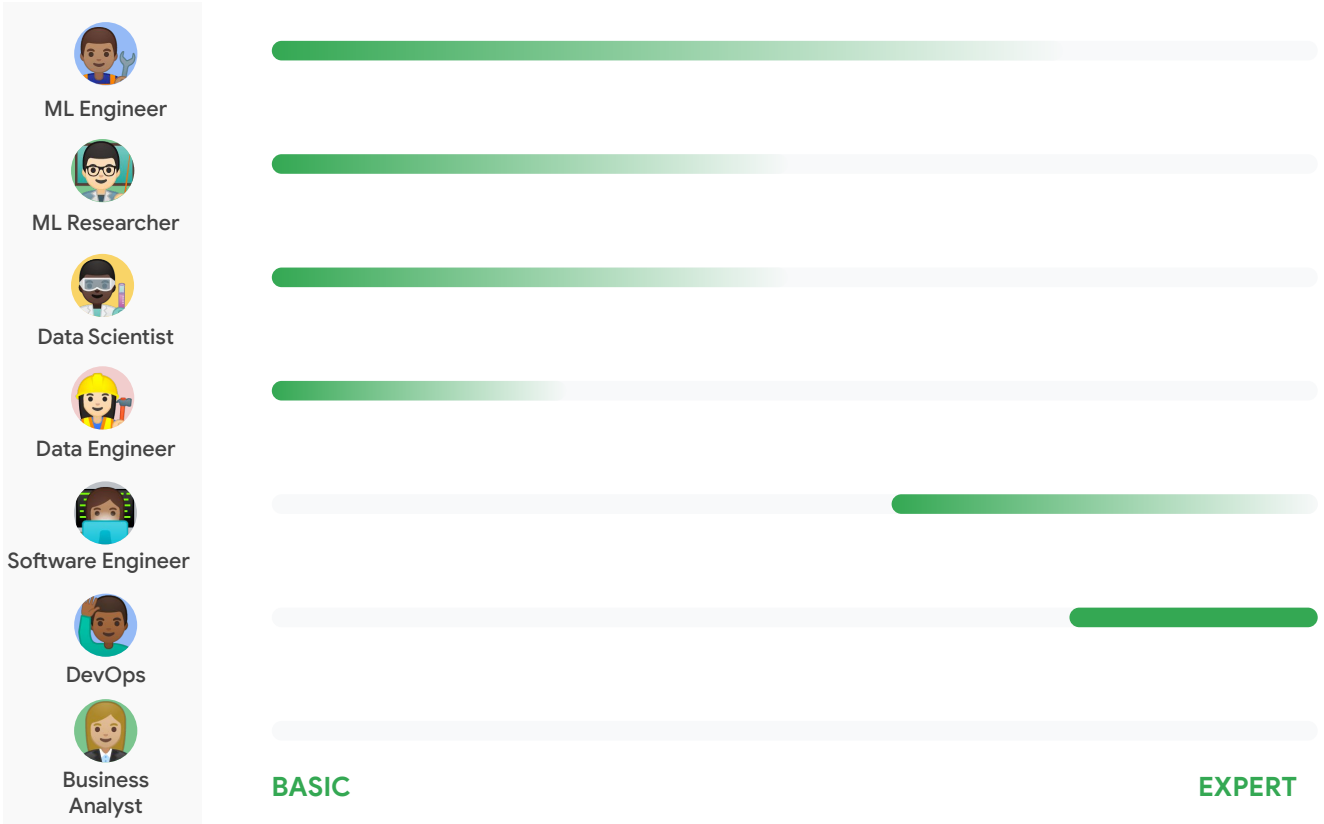
Software Engineer

DevOps

Business Analyst

ML Expertise

Deployment Expertise



ML Engineer

ML Researcher

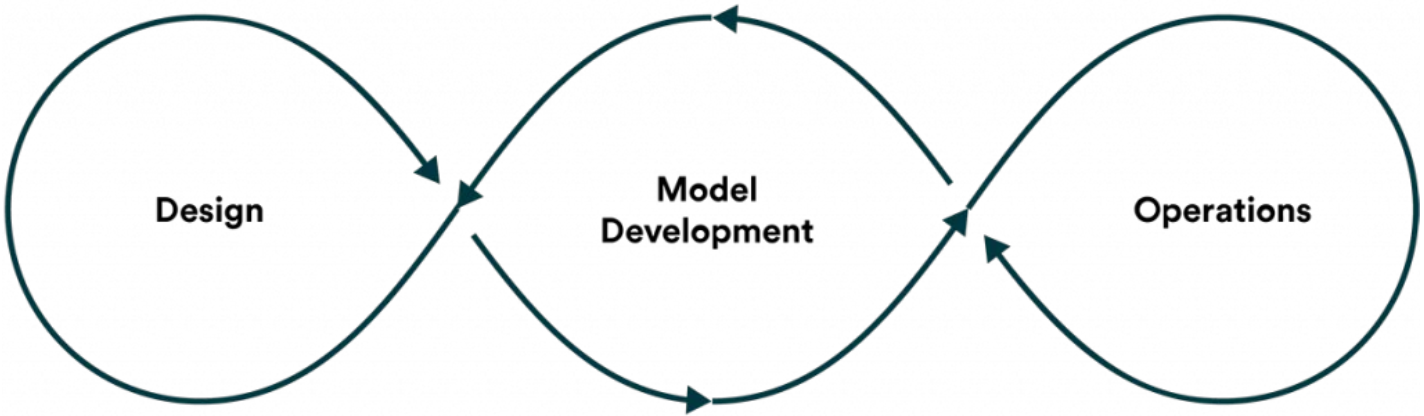
Data Scientist

Data Engineer

Software Engineer

DevOps

Business Analyst



ML Engineer

ML Researcher

Data Scientist

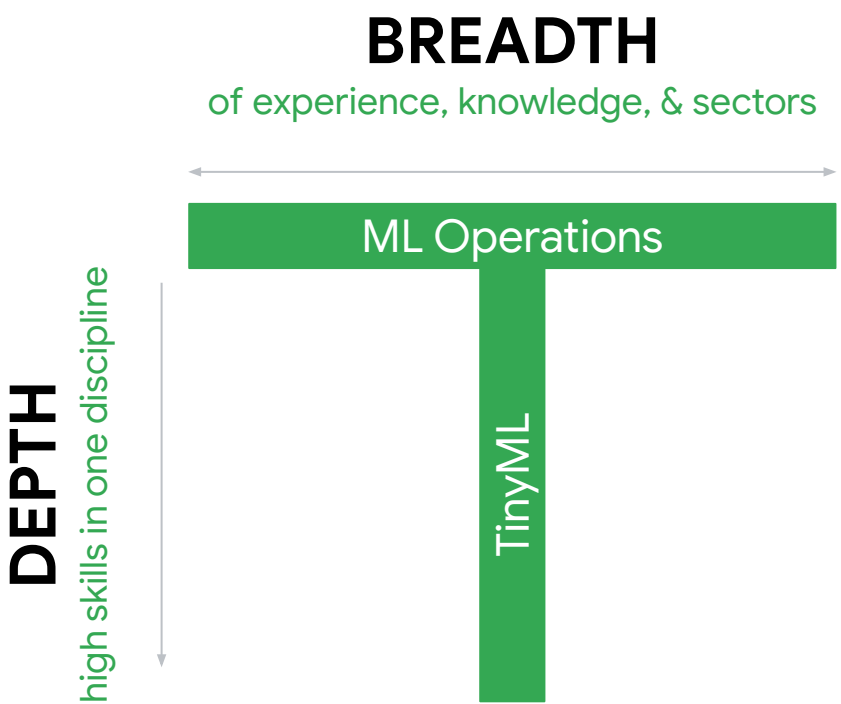
Data Engineer

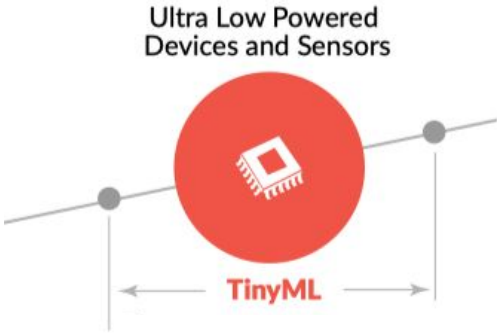
Software Engineer

DevOps

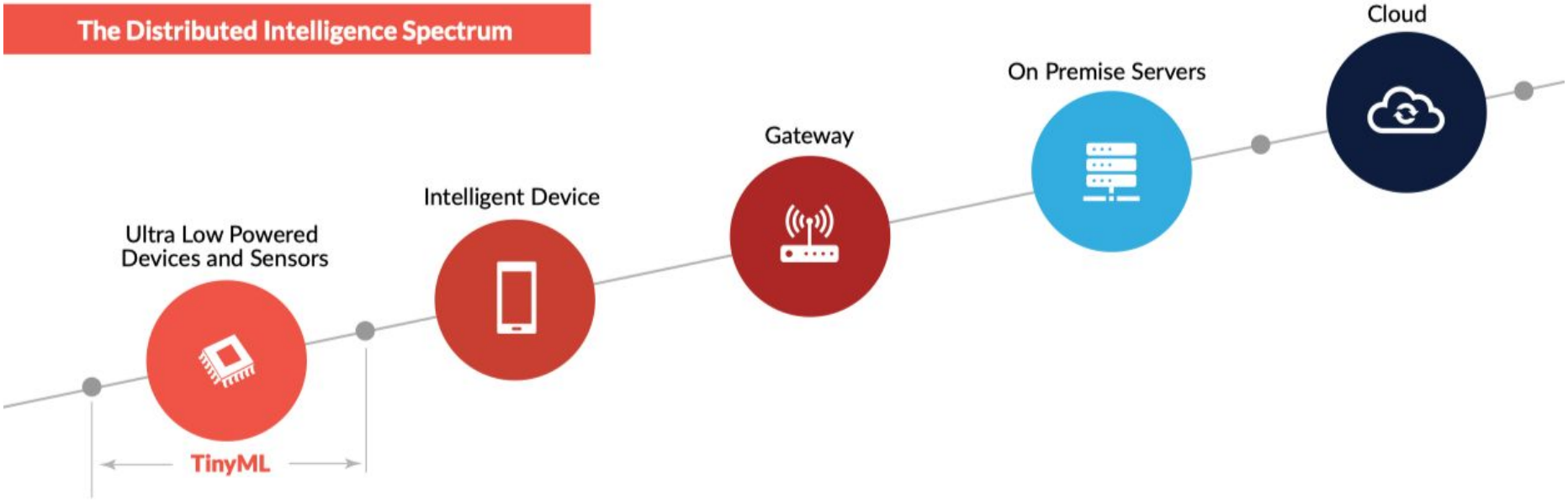
Business Analyst

“Applied ML Xpertise”

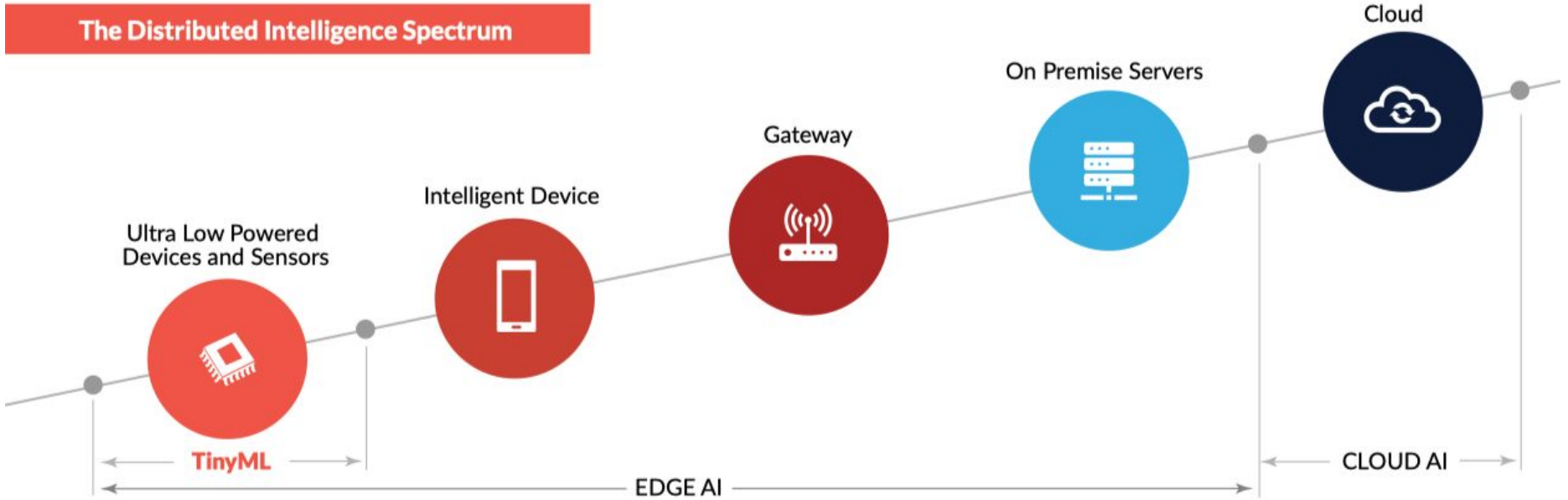

BREADTH
of experience, knowledge, & sectors
ML Operations
TinyML
DEPTH
high skills in one discipline



Source: ABI Research: TinyML

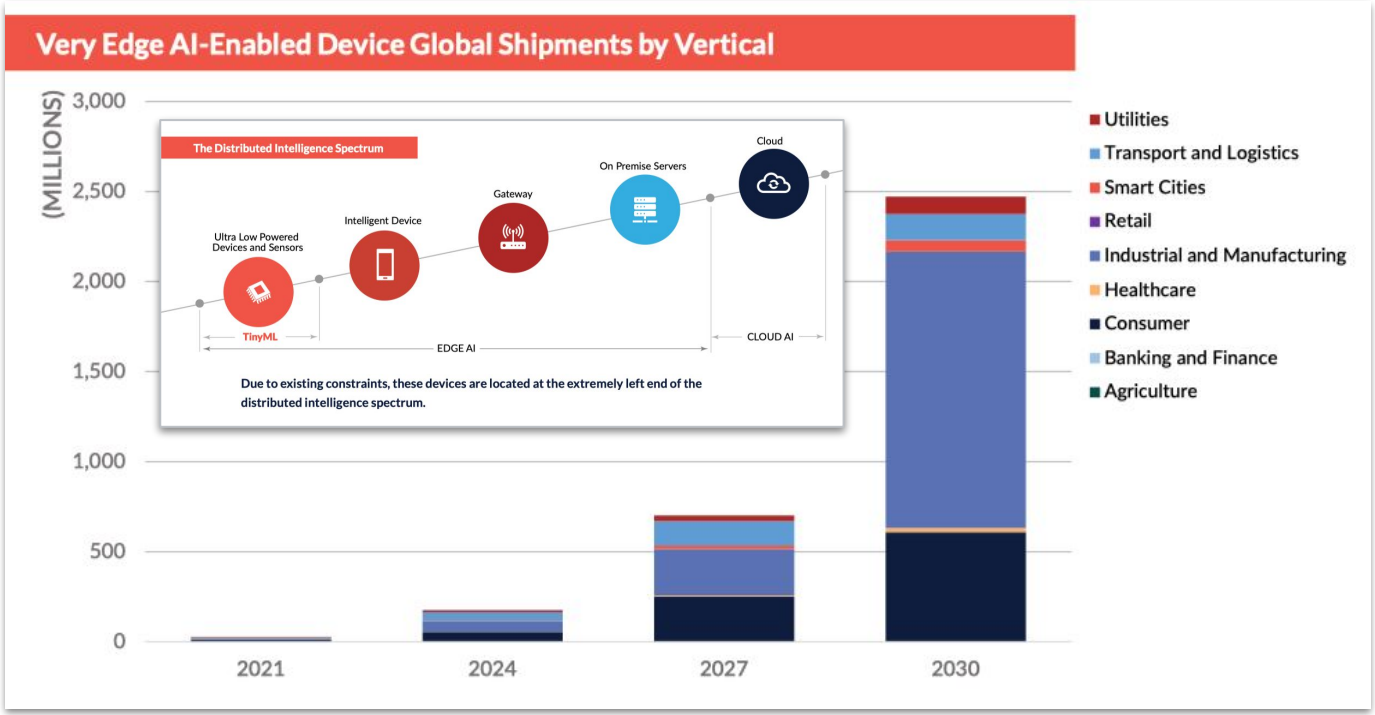


Source: ABI Research: TinyML



Source: ABI Research: TinyML

Massive Potential for Impact

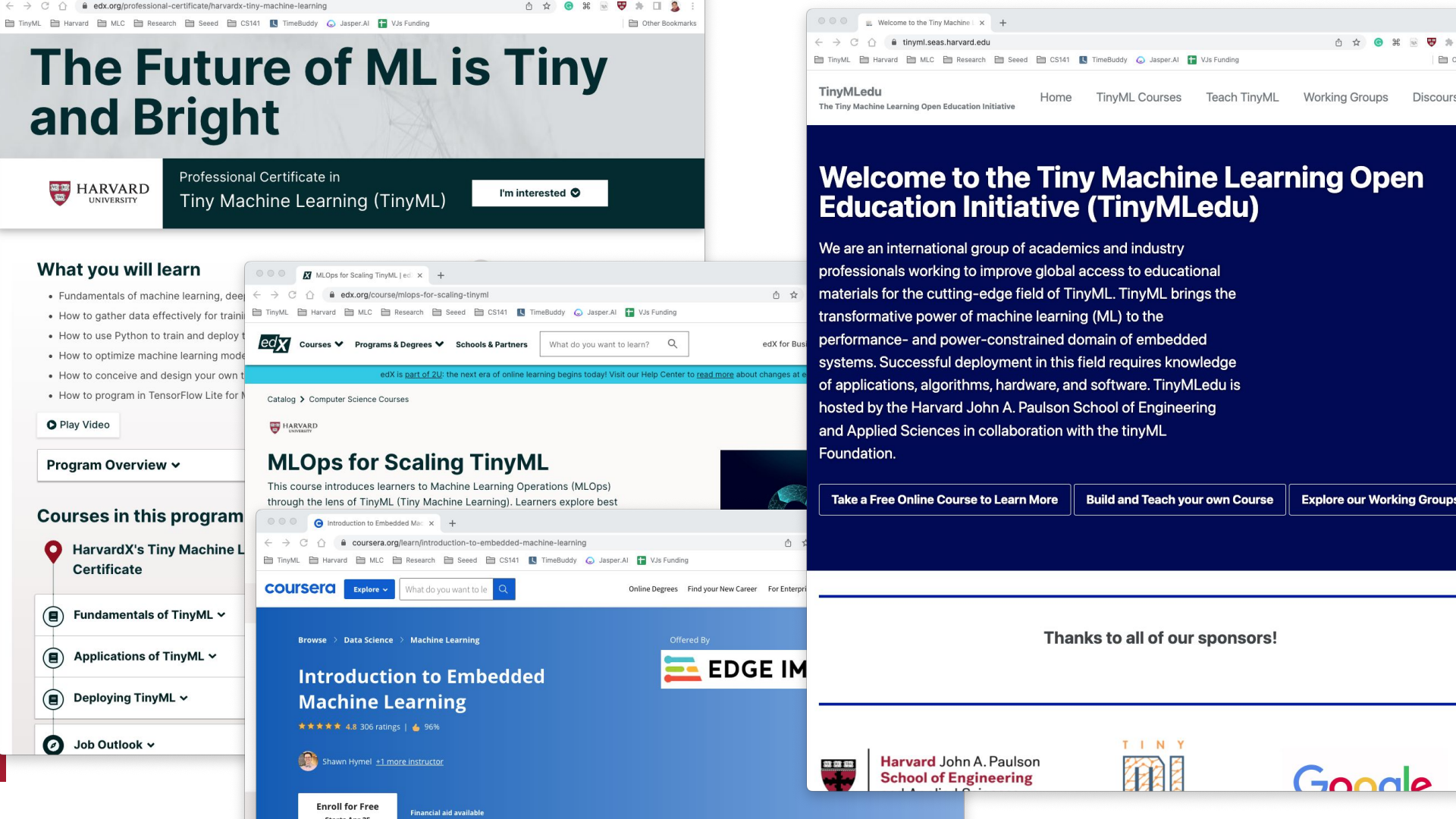


Source: ABI Research: TinyML

Promising **Social Applications** of TinyML

Wildlife conservation

Agriculture


edx.org/professional-certificate/harvardx-tiny-machine-learning
The Future of ML is Tiny and Bright
HARVARD UNIVERSITY
Professional Certificate in
Tiny Machine Learning (TinyML)
I'm interested
What you will learn
Fundamentals of machine learning, deep
How to gather data effectively for traini
How to use Python to train and deploy t
How to optimize machine learning mode
How to conceive and design your own t
How to program in TensorFlow Lite for M
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 | ed
edx.org/course/mlops-for-scaling-tinyml
Courses
Programs & Degrees
Schools & Partners
What do you want to learn?
edX for Bus
edX is part of 2U: the next era of online learning begins today! Visit our Help Center to read more about changes at e
Catalog > Computer Science Courses
HARVARD UNIVERSITY
MLOps for Scaling TinyML
This course introduces learners to Machine Learning Operations (MLOps) through the lens of TinyML (Tiny Machine Learning). Learners explore best
Introduction to Embedded Mac
coursera.org/learn/introduction-to-embedded-machine-learning
coursera
Explore
What do you want to le
Online Degrees
Find your New Career
For Enterpr
Browse > Data Science > Machine Learning
Offered By
EDGE IM
Introduction to Embedded Machine Learning
4.8 306 ratings | 96%
Shawn Hymel +1 more instructor
Enroll for Free
Financial aid available
Welcome to the Tiny Machine
tinyml.seas.harvard.edu
TinyMLedu
The Tiny Machine Learning Open Education Initiative
Home
TinyML Courses
Teach TinyML
Working Groups
Discours
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 Groups
Thanks to all of our sponsors!
Harvard John A. Paulson School of Engineering
TINY
Google

Conclusion

The Future of ML is Tiny and Bright