The Sustainability of

IoT, Edge Computing and Embedded ML

sebastian büttrich

sebastian@itu.dk

Who

Sebastian Büttrich

Background physics /
atomic ..., quantum ..., optics, RF spectroscopy

30+ years of internet/IT at large
25+ years of wireless networking
20+ years of IoT

At ITU, Research Lab Manager dasya.itu.dk
IoT course
DISCO Cubesat project

Globally, with
Network Startup Resource Center (https://nsrc.org) a.o.

sebastian@itu.dk

This talk

1/ General: remarks on the
concept of sustainability in general IT

2/ An introduction to
IoT, Edge Computing, Embedded ML/TinyML, …

3/ Specific: A paper on the
Sustainablity of Edge IoT and Embedded ML

Critical reading

Sustainability – Attempts at a definition

There are many possible definitions. One is:

Sustainability is the **ability to exist constantly.**

Development that
**”meets the needs of the present
without compromising the ability of
future generations to meet their own needs.”**

["Our Common Future, From One Earth to One World" - UN - Brundtland Commission Report 1987]

Sources: https://en.wikipedia.org/wiki/Sustainability
http://www.un-documents.net/ocf-ov.htm#1.2

The concept of sustainability, or *Nachhaltigkeit* in German, can be traced back to Hans Carl von Carlowitz (1645–1714), and was applied to **forestry.**

Sources: http://www.environmentandsociety.org/tools/keywords/hans-carl-von-carlowitz-and-sustainability
https://commons.wikimedia.org/wiki/File:Spruce_forest_at_Holma.jpg

Sustainability – Domains or Pillars

largely agreed on:

environmental, economic and social

with subdomains

cultural, technological and political

(our focus domains underlined)

Sources: https://en.wikipedia.org/wiki/Sustainability

This talk focuses on

environmental – ecological & technological aspects

and to some extent, the **economic** -

Social – political -
cultural

is another (long) talk …

”Social” media & society

Facebook took decisive action on QAnon Wednesday.
Getty Images, Facebook

POLICY \ TECH \ FACEBOOK

Facebook takes down ‘call to arms’ event after two shot dead in Kenosha

‘Any patriots willing to take up arms and defend our city tonight from the evil thugs?’ asked the Facebook group

Sources: Getty images
https://www.theverge.com/2020/8/26/21402571/kenosha-guard-shooting-facebook-deplatforming-militia-violence

The Sustainable Development Goals,
adopted on 25 September 2015 as a part of the UN 2030 Agenda.

Sources:
https://en.wikipedia.org/wiki/Sustainable_Development_Goals

Sustainability – Contradictions

Already at this point – there are **contradictions**:

the **”ability to exist constantly”** implies
lack of change, or at least **a circle.**

How does this concept relate to
”development” or **”growth”?**

In any finite system, constant growth per definition can not be sustainable.

A possible issue

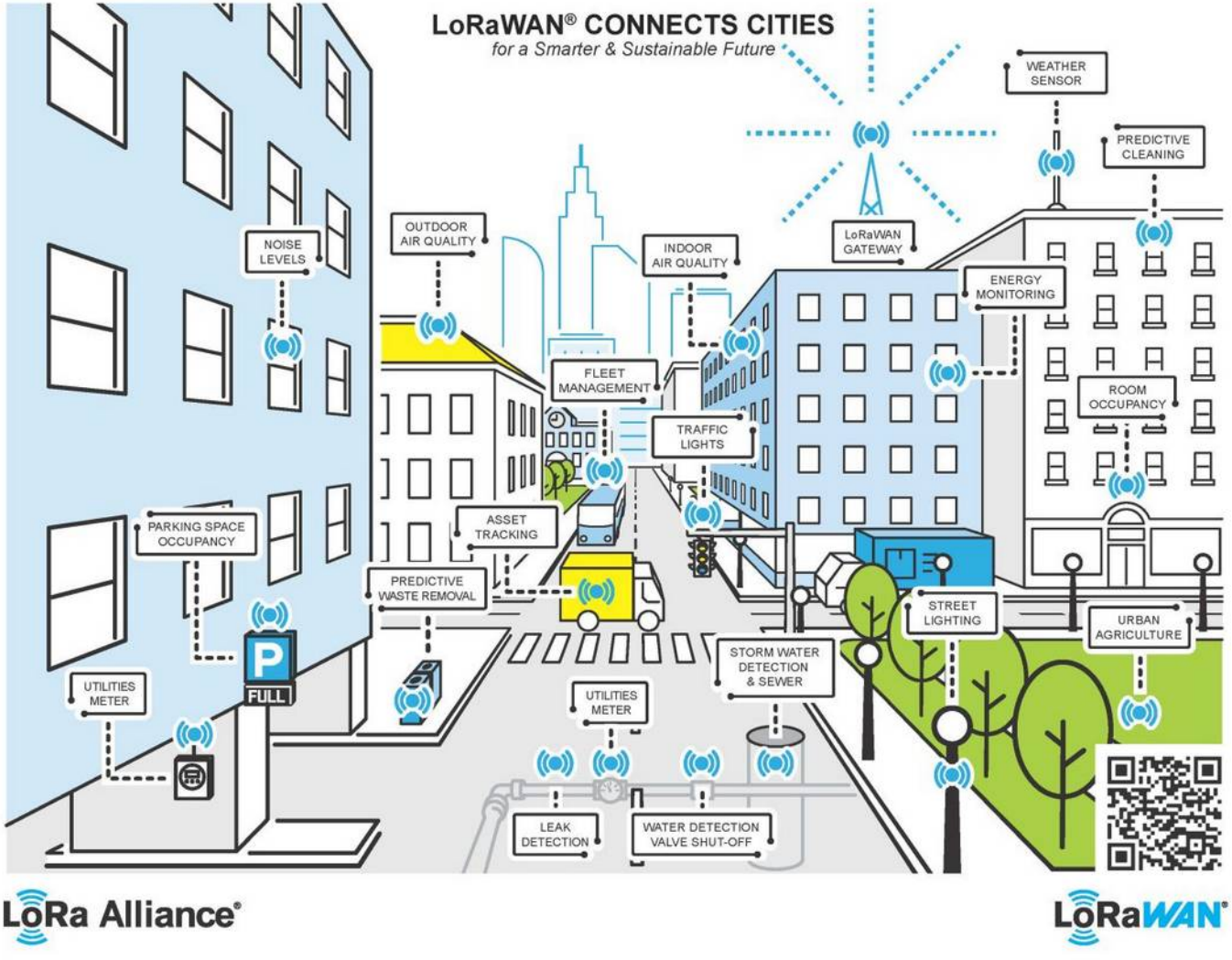
A well accepted and repeated commonplace that
Digitalization, IoT, sensors, data, some form of "smartness" are a part of the "green transition".

However, the IT sector is using energy and natural resources at a fast growing rate.

The very tool we claim will help us tackle climate change is responsible for further driving it.

Can we quantify which direction is stronger?

Smartness


LoRaWAN® CONNECTS CITIES
for a Smarter & Sustainable Future
WEATHER SENSOR
PREDICTIVE CLEANING
LoRaWAN GATEWAY
ENERGY MONITORING
OUTDOOR AIR QUALITY
NOISE LEVELS
INDOOR AIR QUALITY
FLEET MANAGEMENT
ROOM OCCUPANCY
TRAFFIC LIGHTS
PARKING SPACE OCCUPANCY
ASSET TRACKING
PREDICTIVE WASTE REMOVAL
STREET LIGHTING
URBAN AGRICULTURE
STORM WATER DETECTION & SEWER
UTILITIES METER
FULL
UTILITIES METER
LEAK DETECTION
WATER DETECTION VALVE SHUT-OFF
LoRa Alliance®
LoRaWAN®

Exact description or estimation of the IT environmental footprint is hard

A personal experience:

someone asked me about **energy footprint of keeping their pictures and videos in the cloud** -
Is it like ***a light bulb? A fridge? A car?***

I couldn't answer initially
and neither could my colleagues.
Can you? Take 1 TB of images –
what does it compare to?

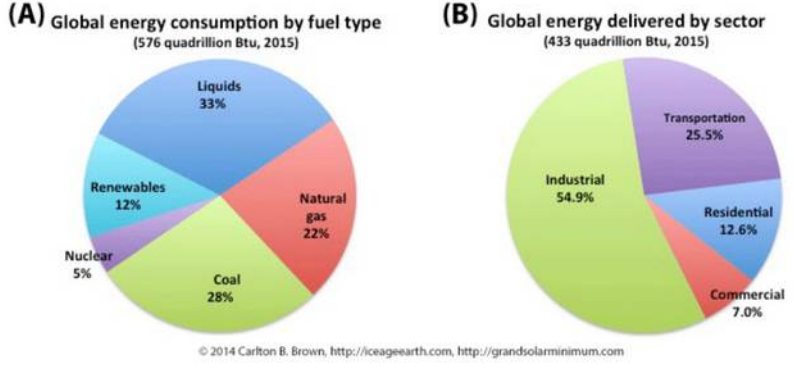
Exact description or estimation of the IT environmental footprint is hard

What to include?

Traditonal view of sectors:
Industrial, commercial, residential, transport
==> IT is in all of these

IT Industry is secretive:
Those who know won't tell

Data says what one wants it to say



Sources: https://www.bp.com/en/global/corporate/energy-economics/energy-outlook/demand-by-sector.html

ICTs use ~10%* of global electricity

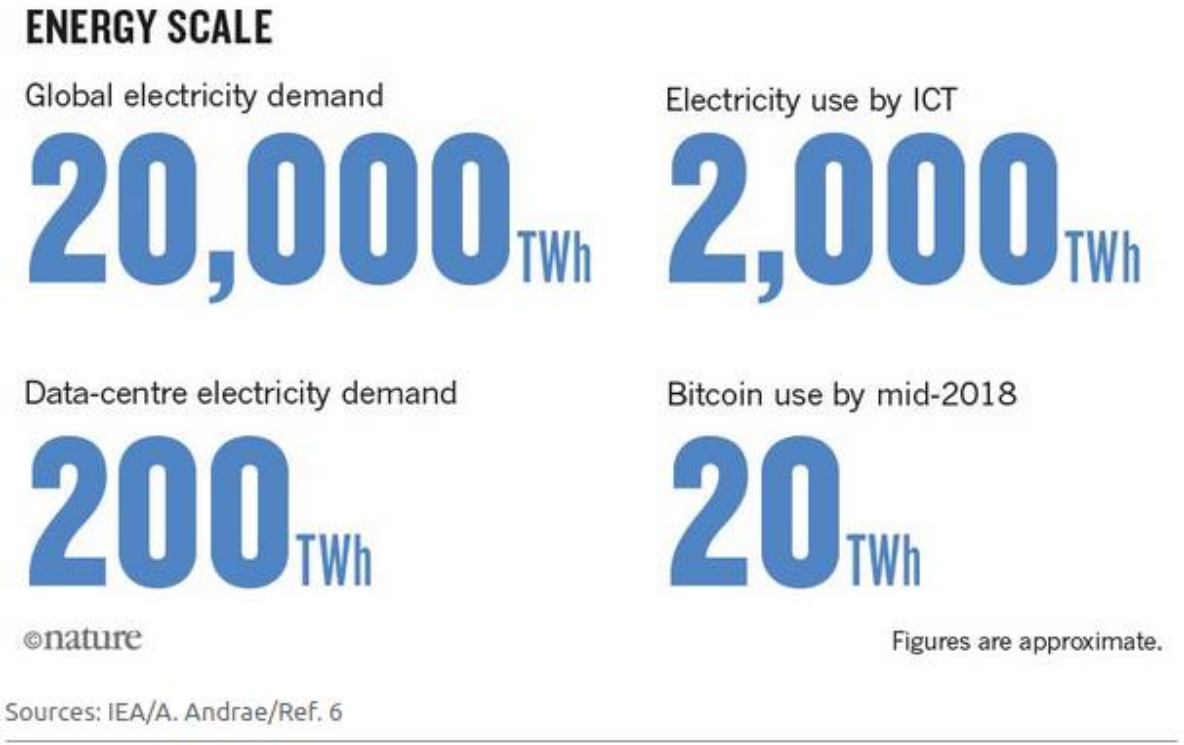
NEWS FEATURE · 12 SEPTEMBER 2018 · CORRECTION 13 SEPTEMBER 2018

How to stop data centres from gobbling up the world's electricity

The energy-efficiency drive at the information factories that serve us Facebook, Google and Bitcoin.

Nicola Jones

A Facebook data centre in Luleå, Sweden. Credit: Jonathan Nackstrand/AFP/Getty



** update 2024 – global demand approx 25,000 Twh

Excellent start point for reading: **https://www.nature.com/articles/d41586-018-06610-y**

*** with huge uncertainty**

Sources: https://www.iea.org/ https://eia.gov
https://yearbook.enerdata.net/electricity/electricity-domestic-consumption-data.html
https://ember-climate.org/wp-content/uploads/2020/03/Ember-2020GlobalElectricityReview-Web.pdf
https://www.nature.com/articles/d41586-018-06610-y

On uncertainty – data tells the story you want it to tell

Guardian analysis (Seot 2024):

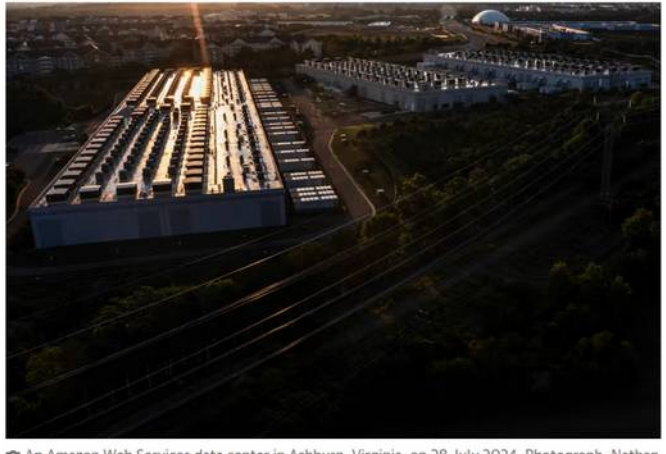
Company official data on gas emissions

off by

a factor of 7

Data center emissions probably 662% higher than big tech claims. Can it keep up the ruse?

Emissions from in-house data centers of Google, Microsoft, Meta and Apple may be 7.62 times higher than official tally



An Amazon Web Services data center in Ashburn, Virginia, on 28 July 2024. Photograph: Nathan Howard/Bloomberg via Getty Images

Big tech has made some big claims about greenhouse gas emissions in recent years. But as the rise of artificial intelligence creates ever bigger energy demands, it's getting hard for the industry to hide the true costs of the data centers powering the tech revolution.

According to a Guardian analysis, from 2020 to 2022 the real emissions from the "in-house" or company-owned data centers of Google, Microsoft, Meta and Apple are probably about 662% - or 7.62 times - higher than officially reported.

Source: https://www.theguardian.com/technology/2024/sep/15/data-center-gas-emissions-tech

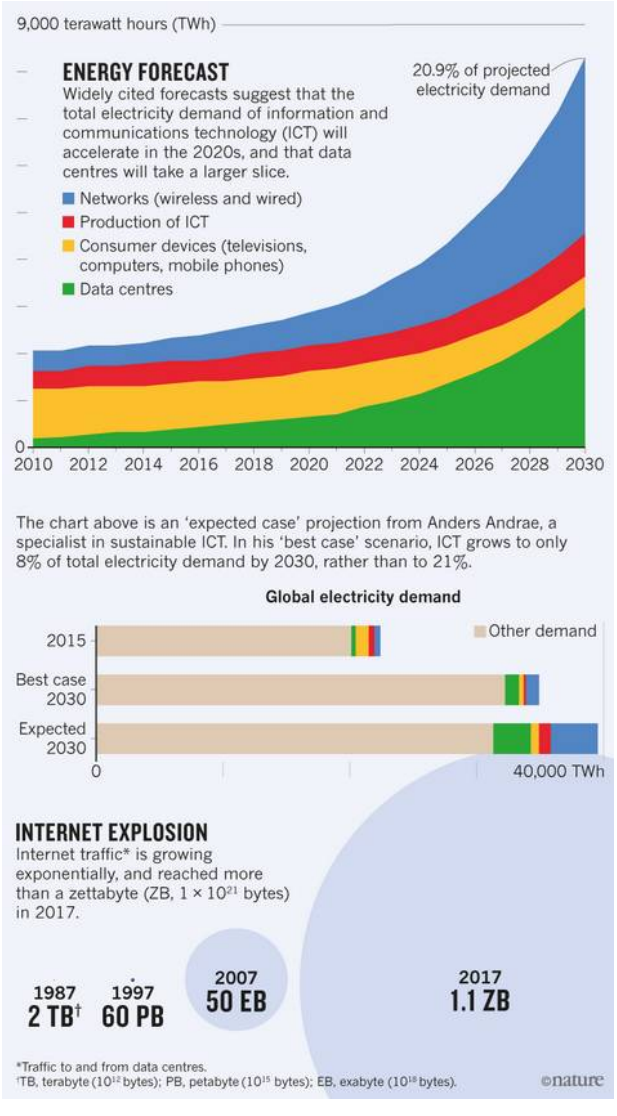
Some key sources

Andrae et al. (Huawei) have supplied data which is widely used and discussed

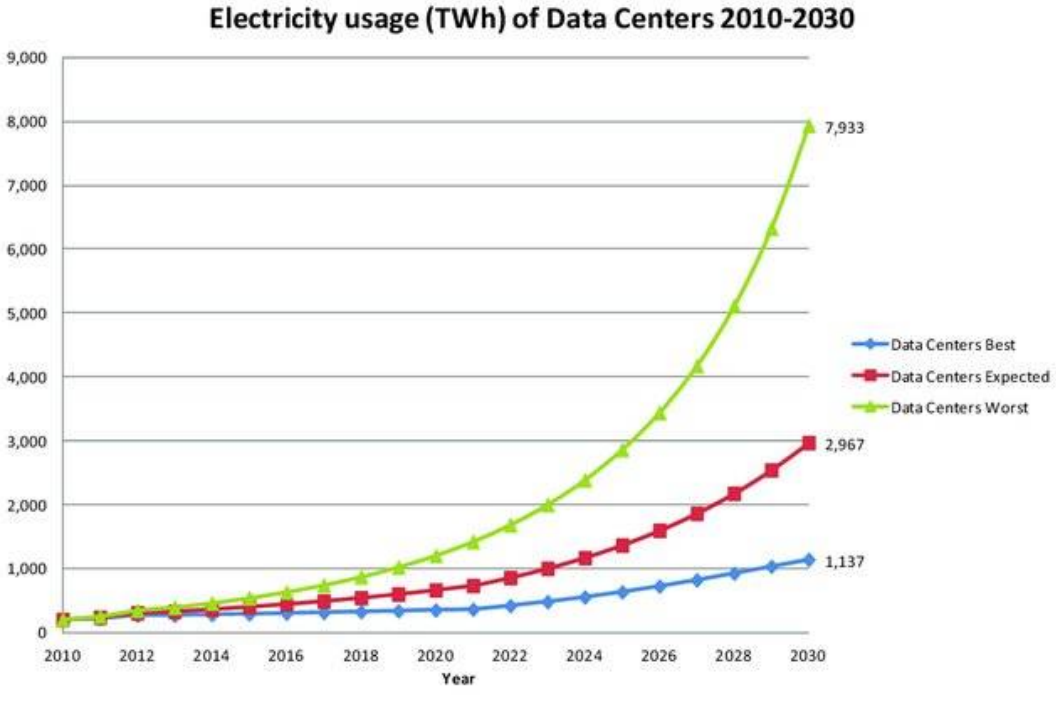
Andrae, A.S.G.; Edler, T.
On Global Electricity Usage of Communication Technology Trends to 2030.
Challenges 2015, 6, 117–157.
https://www.mdpi.com/2078-1547/6/1/117

Total Consumer Power Consumption Forecast
October 2017
Conference: Nordic Digital Business Summit
Project: Global Forecasting of ICT footprints
Anders S.G. Andrae
https://www.researchgate.net/publication/320225452_Total_Consumer_Power_Consumption_Forecast

key message:
centralized data & networks are drivers, not consumer devices and production



Central role of Data Centers: growing 10% per year *(possibly faster 2024 and on?)*



Environment

Global warming: Data centres to consume three times as much energy in next decade, experts warn

416.2 terawatt hours of electricity world's data centres used last year was far higher than UK's total consumption

Tom Bawden Environment Editor | @BawdenTom |

Saturday 23 January 2016 22:37 |

Data centres of the world will consume 1/5 of Earth's power by 2025

Joao Lima

12 Dec 2017 (Updated: 25 Jun 2020)
1 minute read

Alarming new research suggests that failure to source renewable energy could make data centres one of the biggest polluters in just seven years.

The rapid adoption of data-hungry machines and services is driving the need for more power to keep the lights on in the data centres of the world. As analysts estimate as many as 50 billion devices to be connected by 2020, with some statistics pointing to more than 100 billion a further five years down the line, new alarming research suggests that data centres will be one of the biggest energy consumers on the planet, beating many countries' energy consumption levels. According to a paper to be published by US researchers before the end of the year, the ICT industry is posed to be responsible for up to 3.5% of global emissions by 2020, with this value potentially escalating to 14% by 2040, according to Climate Change News. Researchers say this will be directly related to the fact that the data centre sector could be using 20% of all available electricity in the world by 2025 on the back of the large amounts of data being created at a fastest speed than ever before seen. The figures meet those published by Swedish researcher and Senior Expert Life Cycle Assessment at Huawei, Anders Andrae in 2016 in his "Total Consumer Power Consumption Forecast". Andrae predicts that by 2025, data centres will amount to ICT's largest share of global electricity production at 33%, followed by smartphones (15%) networks (10%) and TV (9%). As for the wider global usage, Andrae also expects data centres to use 20% of the world's energy, however, he places their carbon footprint at 5.5% of the global value, should adoption of more efficient energy sources not evolve at speed. The exponential utilisation of energy by data centres is not new, with the amount of power consumed increasing 9% between 2010 and 2015, according to KPN Integrated. On the global scale, data centres are poised to be the largest global energy users by 2025 at 4.5%, an increase from just 0.9% in 2015, according to Andrae's report. In comparison, consumer devices, fixed access wired services, wireless networks and production are all set to lag behind data centres in terms of energy usage. Globally, data centres were in 2014 responsible for around 1.62% of the world's utilised energy that year, according to Yole Développement. That has increased today to more than 3% of the world's energy (around 420 terawatts) and data centres are also responsible for 2% of total greenhouse gas emissions.

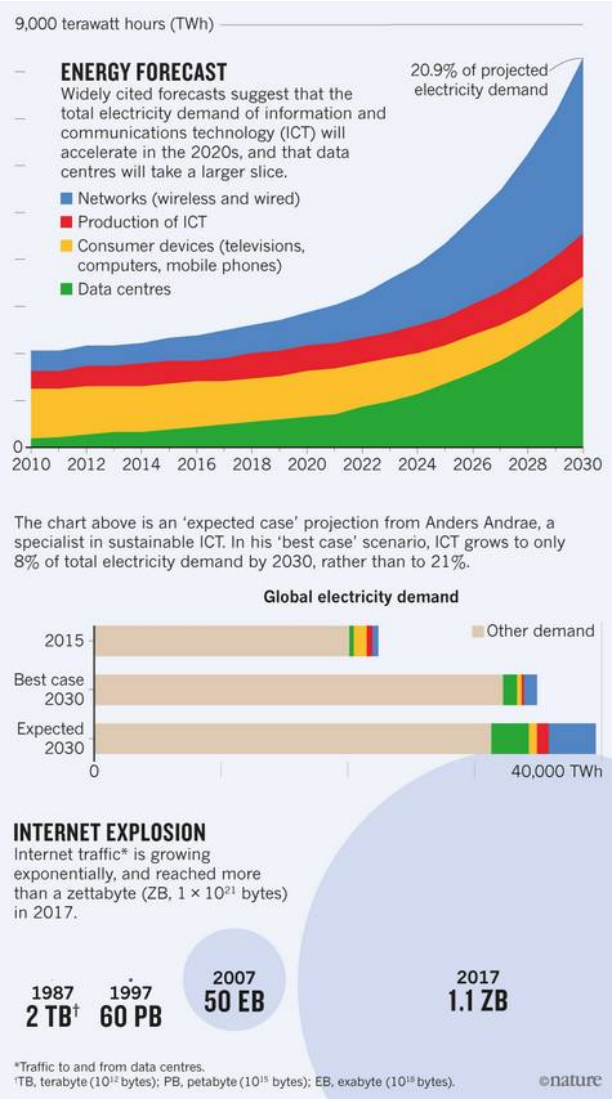
Sources:
https://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html
https://www.broad-group.com/data/news/documents/b1m2y6q1x5dv5t
https://www.researchgate.net/publication/320225452_Total_Consumer_Power_Consumption_Forecast

Difficult predictions

Some important studies are from 2017-2020.
We did not even see ChatGPT coming at that point.
Video and Bitcoin seemed the biggest contributions.

Koot, Martijn, and Fons Wijnhoven.
"Usage impact on data center electricity needs: A system dynamic forecasting model."
Applied Energy 291 (2021): 116798.
https://www.sciencedirect.com/science/article/pii/S0306261921003019

We have a very rough idea that
IT will be a 10% (or more?) part of global electricity consumption.



Update 2024 – post-LLMs & “AI”

The large scale adoption of “AI” has caused a significant rise in data center energy consumption.

“Electricity consumption from data centres, artificial intelligence (AI) and the cryptocurrency sector could double by 2026 [from 2022]. Data centres are significant drivers of growth in electricity demand in many regions. After globally consuming an estimated 460 terawatt-hours (TWh) in 2022, data centres’ total electricity consumption cAfter globally consuming an estimated 460 terawatt-hours (TWh) in 2022, data centres’ total electricity consumption could reach more than 1 000 TWh in 2026.ould reach more than 1 000 TWh in 2026.”

source:
The International Energy Agency (IEA), https://www.iea.org/reports/electricity-2024/executive-summary

Update 2024 – post-LLMs & “AI”

Bloomberg

Live TV | Markets | Economics | Industries | Tech | Politics | Businessweek | Opinion | More

The AI Race: Startups to Watch | AI Glossary | Model Collapse Risk | AI's Real Carbon Footprint | Small AI Models

Green

Big Oil Sees AI Boom Driving ‘Crazy Demand’ for US Natural Gas

By Will Mathis

October 29, 2024 at 5:02 PM GMT+1

Why Microsoft made a deal to help restart Three Mile Island

A once-shuttered nuclear plant could soon return to the grid.

FINANCIAL TIMES

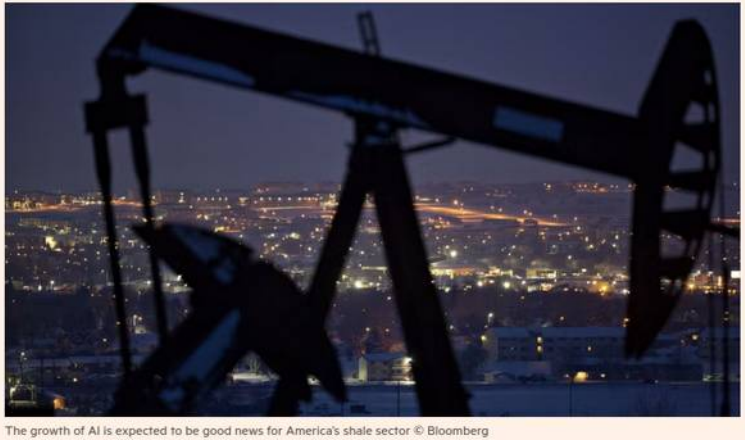
HOME WORLD US COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI

Natural gas | + Add to myFT

AI revolution will be boon for natural gas, say fossil fuel bosses

Data centres' need for reliable power supply set to soar

Save



The growth of AI is expected to be good news for America's shale sector © Bloomberg

more sources:
The International Energy Agency (IEA), https://www.iea.org/reports/electricity-2024/executive-summary
https://iea.blob.core.windows.net/assets/5e9122fc-9d5b-4f18-8438-dac8b39b702a/WorldEnergyOutlook2024.pdf

Update 2024 – post-LLMs & “AI”, sources

more sources:
The International Energy Agency (IEA), https://www.iea.org/reports/electricity-2024/executive-summary
https://iea.blob.core.windows.net/assets/5e9122fc-9d5b-4f18-8438-dac8b39b702a/WorldEnergyOutlook2024.pdf

AI is an energy hog. This is what it means for climate change.
MIT Tech review
https://www.technologyreview.com/2024/05/23/1092777/ai-is-an-energy-hog-this-is-what-it-means-for-climate-change/

How AI Is Fueling a Boom in Data Centers and Energy Demand
Time, June 2024
https://time.com/6987773/ai-data-centers-energy-usage-climate-change/

Big Oil Sees AI Boom Driving 'Crazy Demand' for US Natural Gas
Bloomberg Oct 2024
https://www.bloomberg.com/news/articles/2024-10-29/big-oil-sees-ai-boom-driving-crazy-demand-for-us-natural-gas

AI revolution will be boon for natural gas, say fossil fuel bosses
FT April 2024
https://www.ft.com/content/1f93b9b2-b264-44e2-87cc-83c04d8f1e2b

Update 2024 – Europe

Energy Consumption in Data Centres and Broadband Communication Networks in the EU

sources:
Europen Commisiion
https://interactdc.com/static/images/documents/JRC135926_01.pdf

Executive Summary

Demand for digital services is rising rapidly, raising concerns about the energy use and environmental impacts of data centres and telecommunication networks. Despite the increasing public and policy interest in addressing these impacts, there is a lack of official statistics on the energy use of digital infrastructure. This study reviews and uses existing literature and public data sources to estimate the energy consumption of data centres and telecommunication networks in the European Union (EU-27) in 2022.

Data centres in the EU used an estimated 45–65 TWh of electricity in 2022, equivalent to 1.8–2.6% of total regional electricity consumption. The top four data centre markets – Germany, France, the Netherlands, and Ireland – accounted for nearly two-thirds of the region's data centre energy use, despite having less than 40% of the population. Data centres represent over 2% of national electricity use in Ireland (18%), the Netherlands (5.2%), Luxembourg (4.8%), Denmark (4.5%), and Germany (3%), Sweden (2.3%), and France (2.2%).

Telecommunication networks used an estimated 25–30 TWh of electricity, equivalent to 1–1.2% of total EU electricity use. The four largest Member States by population and GDP (Germany, France, Italy, and Spain) were also the four largest users of energy for telecommunication networks, accounting for 65% of the total. Network energy use as a share of national electricity use was both lower and more uniform compared with data centres, ranging from 0.5% to 1.5%. In contrast, data centres as a share of national electricity use range from as low as 0.4% in some countries to as high as 18% in Ireland.

The combined energy use of data centres and telecommunication networks in the EU was 70–95 TWh in 2022, equivalent to 2.8–3.8% of total regional electricity use. The four largest Member States – Germany, France, Italy, and Spain – accounted for about 60% of total digital infrastructure energy use in the region. Digital infrastructure accounts for more than 5% of national electricity use in four countries, each with major data centre markets: Ireland (19%), the Netherlands (6%), Luxembourg (5.5%), and Denmark (5%).

Policymakers and companies must work together to improve data collection, quality and availability. While the estimates of this study represent a likely range of figures, it is critical to develop more robust estimates to better understand trends and make informed policy decisions to manage the energy and environmental impacts of digital infrastructure. Governments and statistical agencies should develop standardised definitions and classifications for data centres and networks, such as providing criteria and guidance on classifying different data centre types. Governments and companies should work together to improve data quality and availability regarding data centre energy consumption (by size and type), telecommunication network energy use (by type), as well as relevant activity indicators (e.g. connections, data traffic, data centre workloads). Data collection efforts should also seek to better understand energy use characteristics and implications of specific services and tasks such as artificial intelligence.

Update 2024 – Danmark

Datacentre er en bombe under den grønne omstil-ling

I 2040 ventes datacentre at suge 22 procent af det samlede elforbrug.

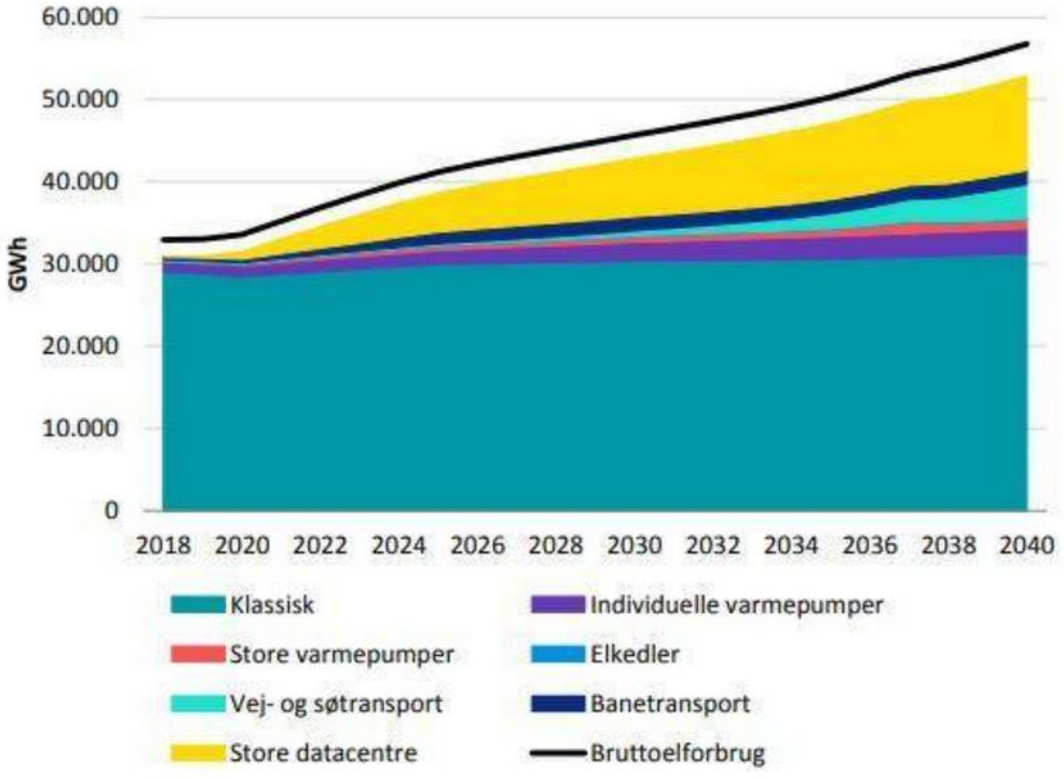
not easy to find updates –
but the numbers/forecasts are
somewhat consistent with
The European level study
by the European Commission -
Data Centers using 20% of all electricity

sources:
DR (2018!)
https://www.dr.dk/nyheder/penge/datacentre-er-en-bombe-under-den-groenne
Klimaraadet 2019

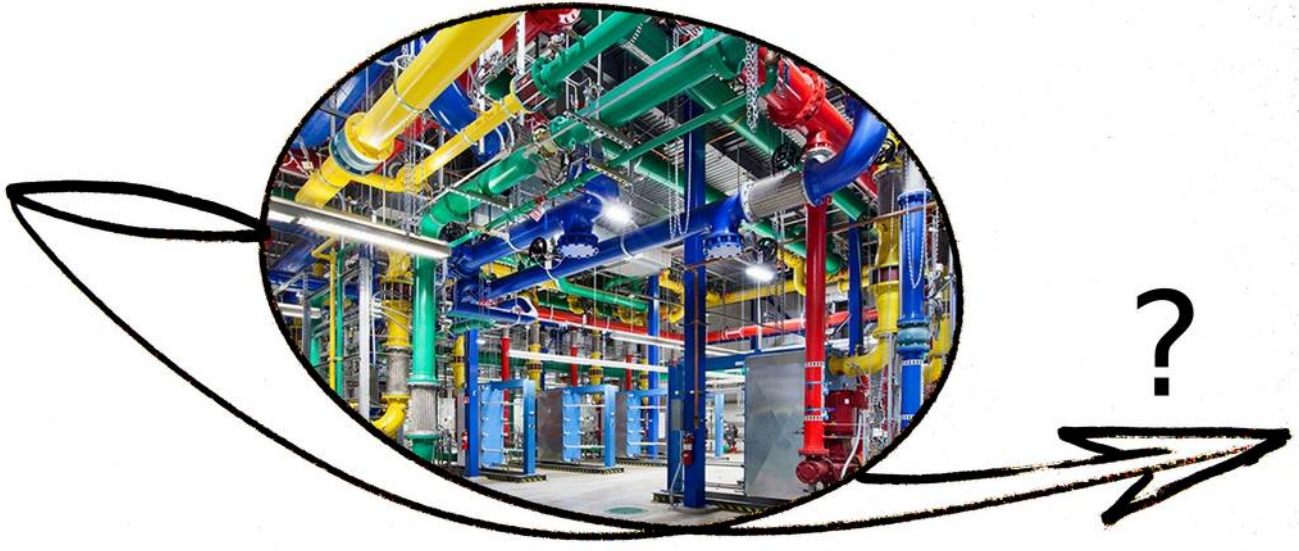
Internal remark:

Powering the
High Performance Computing (HPC) cluster
also a growing issue at ...

Figur 8: Forventet udvikling i det samlede danske elforbrug i fremskrivningsperioden

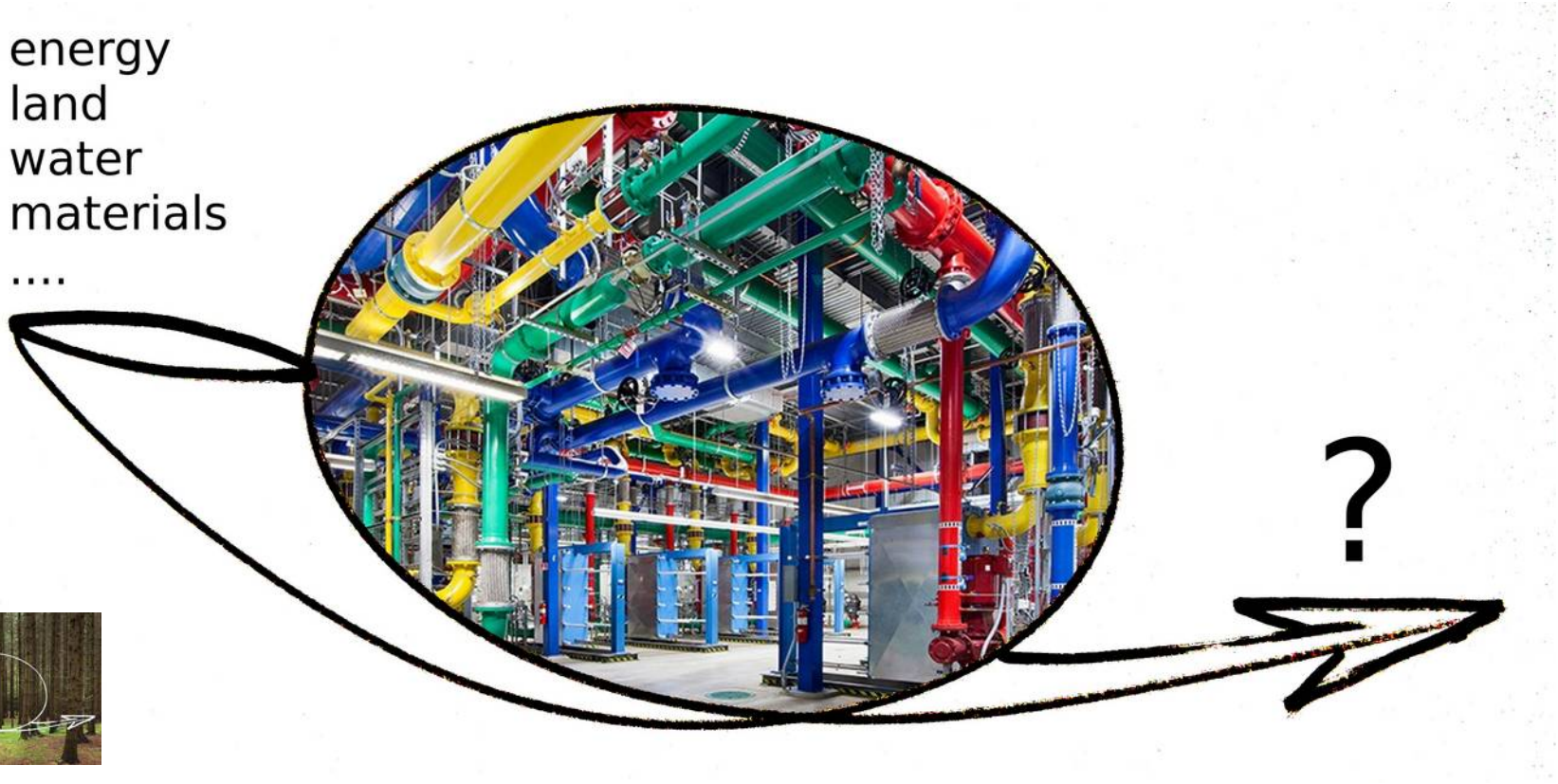


Perhaps the most fundamental problem: What are the INs and OUTs of IT?



Sources: https://en.wikipedia.org/wiki/Sustainability
http://www.un-documents.net/ocf-ov.htm#1.2

Perhaps the most fundamental problem



Unlike e.g. agriculture, forestry, energy sectors,

IT can not produce any of its own input resources

Sources: https://en.wikipedia.org/wiki/Sustainability
http://www.un-documents.net/ocf-ov.htm#1.2

The concept of footprint and handprint

”In environmental management and sustainability there is an increasing interest in measurement and **accounting of beneficial impact**—as an incentive to action, as a communication tool, and to move toward a positive, constructive approach focused on opportunities rather than problems. One approach uses **the metaphor of a “handprint,**” complementing the notion of environmental footprints, which have been widely adopted for impact measurement and accounting.”

Source: Guillaume, Joseph HA, et al. "Giving legs to handprint thinking: Foundations for evaluating the good we do." Earth's Future 8.6 (2020): e2019EF001422.

The concept of footprint and handprint

”**The “handprint” has been suggested as a way of looking at the good we do**, to complement the negative impacts captured by environmental “footprints.” There are many ways we could try to assess a handprint, which capture different perspectives on the world, and the potential role of the handprint assessment in moving toward sustainability.”

Source: Guillaume, Joseph HA, et al. "Giving legs to handprint thinking: Foundations for evaluating the good we do." Earth's Future 8.6 (2020): e2019EF001422.

Life Cycle Assessment (LCA)

Definition:
Life cycle assessment or LCA (also known as life cycle analysis) is a methodology for assessing
environmental impacts associated with all the
stages of the life cycle of a commercial product,
process, or service. [wikipedia]

Term: **”Cradle to Grave”**

How to include ”handprints”?

→ Paper we will be reading

2/ IoT,

Embedded ML / TinyML,

Edge Computing

**(with a LOT of slides that i include for reference for the TinyML-minded …
it s not all crucial for this lecture)**

Terms / Things

ITU-T Y.2060 says:

3.2.3 thing:

With regard to the Internet of things, this is an object of the physical world (physical things) or the information world (virtual things), which is capable of being identified and integrated into communication networks.

This is a very very wide definition.

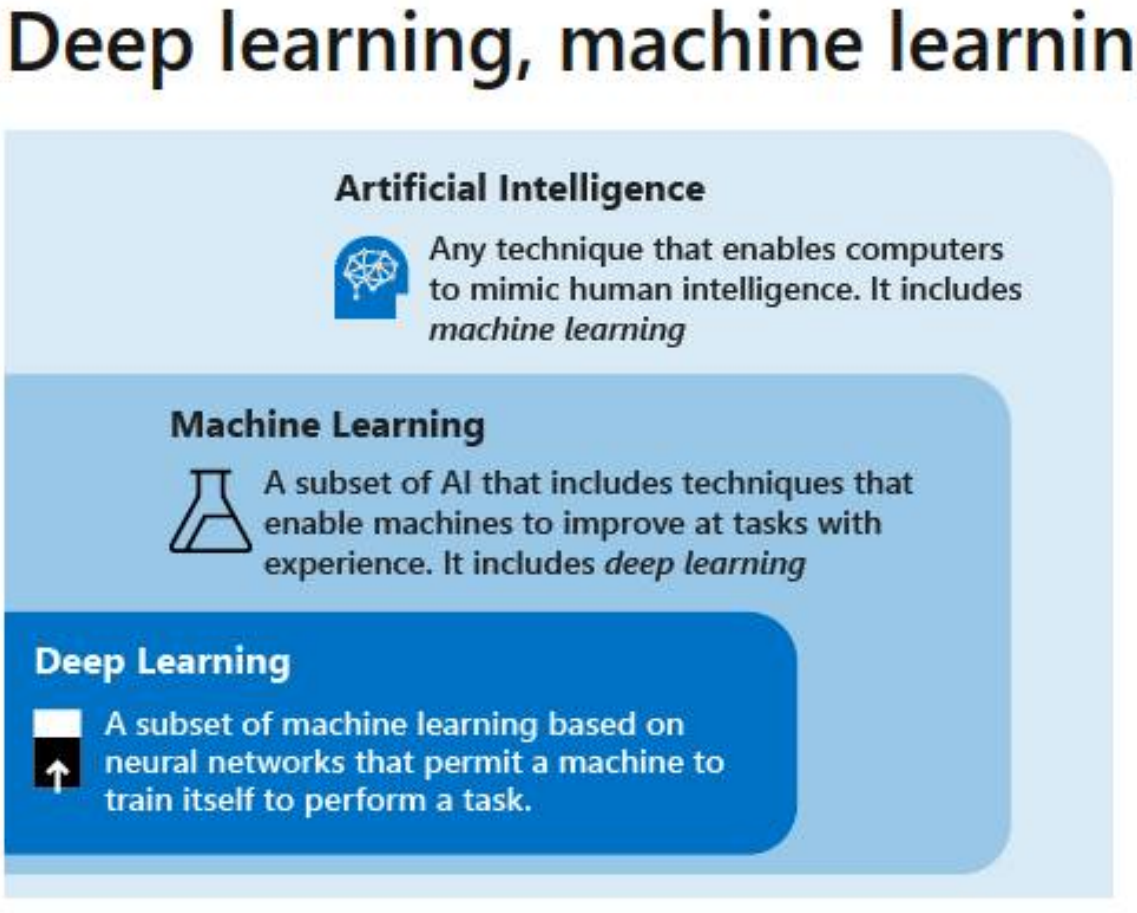
Terms / Devices

ITU-T Y.2060 says:

3.2.1 device: With regard to the Internet of things, this is a piece of equipment with the mandatory capabilities of communication and the optional capabilities of sensing, actuation, datacapture, data storage and data processing.

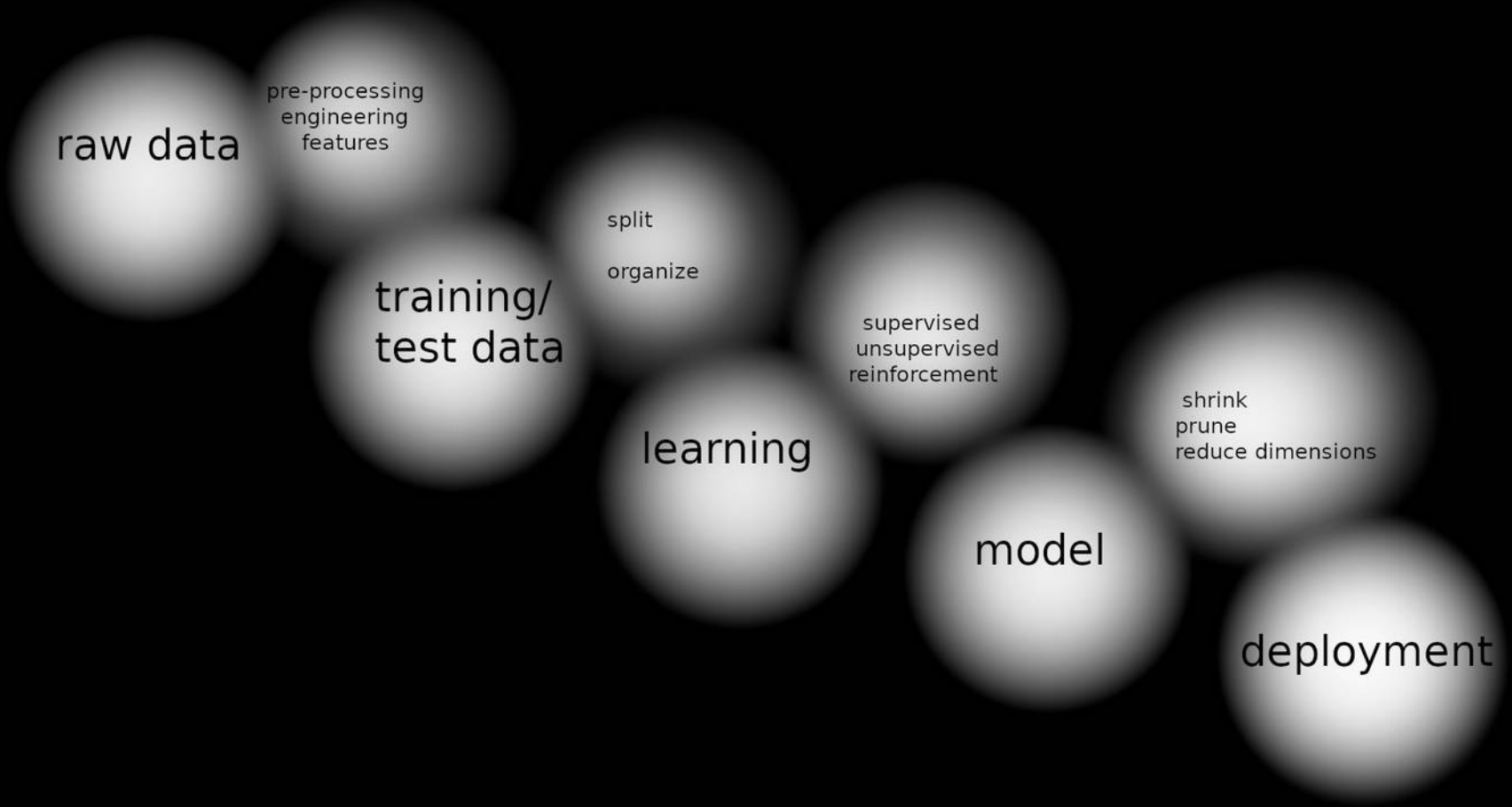
This, too, is a rather wide definition.

Machine Learning – what is it?

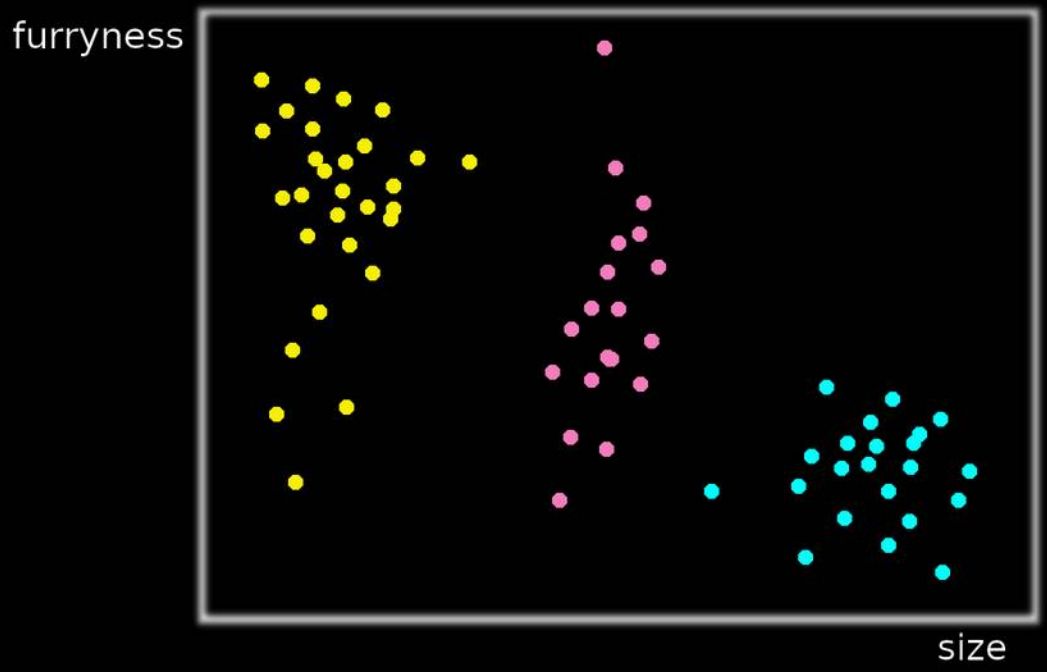


source: https://learn.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning

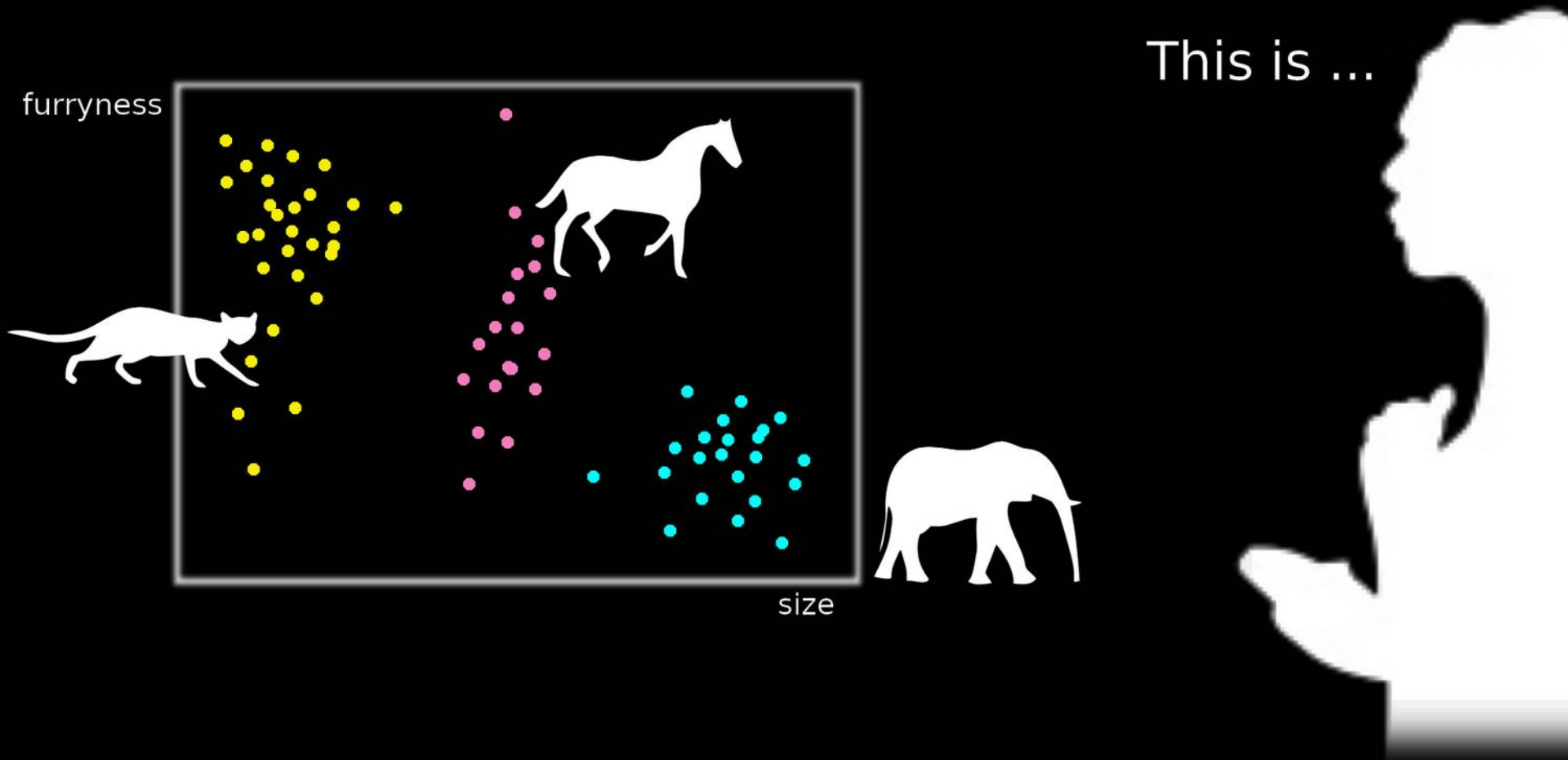
(Intro) … to Machine Learning



Unsupervised ...



...and supervised learning



Machine Learning – what is it for?

Named-entity recognition: This is a serial number! This is an IP number!

Object detection / classification: This is a cat! These are five cats!

Image caption generation: In this picture you see a cat and a dog.

Machine translation: Das ist eine Katze.

Text analytics: This text mentions cats a lot.

Generative AI: Draw cats or answer questions about cats.

e.g. Midjourney, ChatGPT

A cat is a small, carnivorous mammal that is often kept as a domestic pet. It belongs to the Felidae family and is known scientifically as Felis catus. Cats are characterized by their slender bodies, sharp retractable claws, and highly flexible movements. They have a variety of coat colors and patterns.

Machine Learning – our focus here

Computer vision

- Object detection / classification / counting
- Earth/space/mission observation from satellite

Audio

Networks - fiber

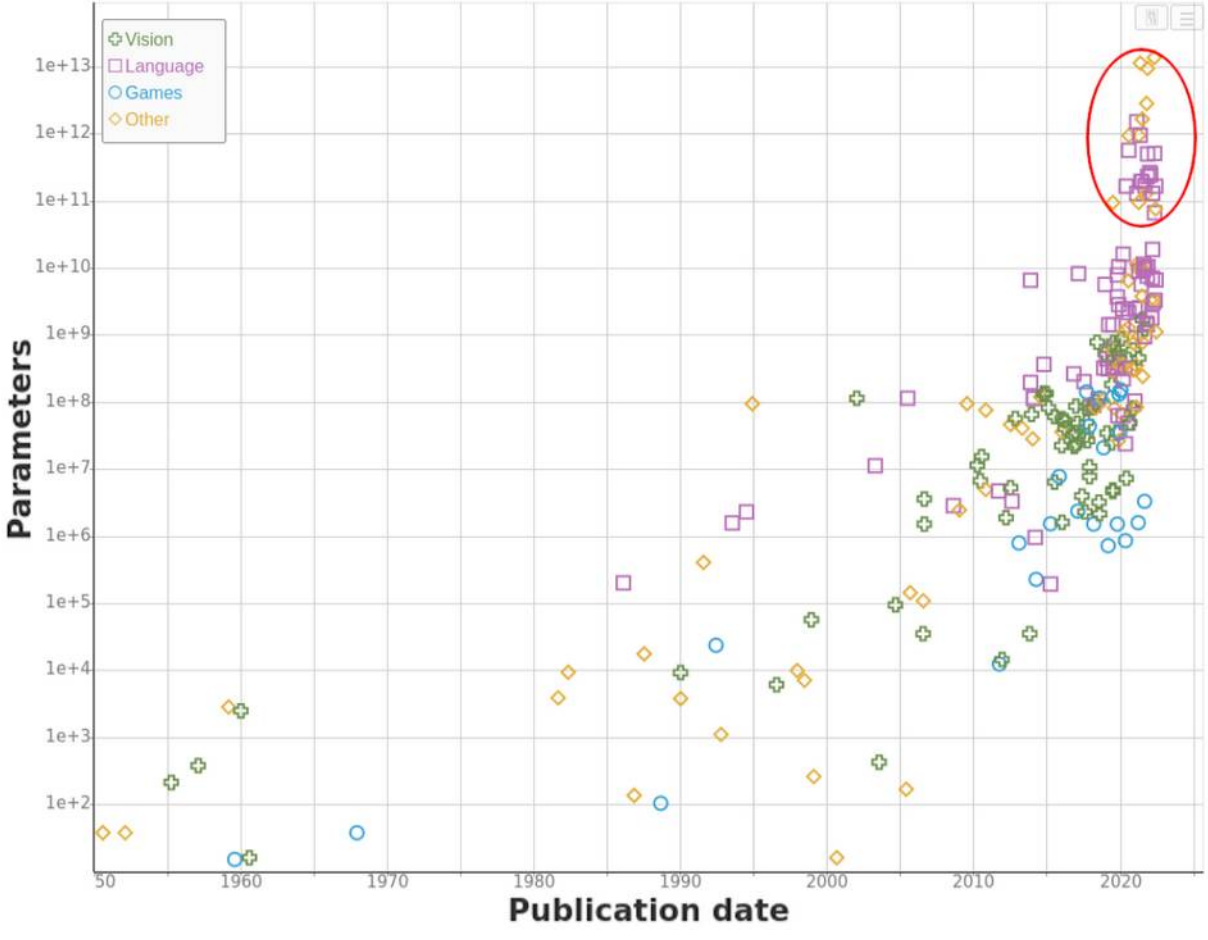
Other sensors … e.g. environmental, radio, tracking, physical, ...

Less relevant for us: Text, Social Media, ***Generative AI*** *etc. - Size is an issue!*

Machine Learning – size is an issue

ChatGPT says:

As of my knowledge cutoff in September 2021, some of the largest language models available were in the range of hundreds of billions of parameters. For example, OpenAI's GPT-3 model, which is one of the largest known models, has 175 billion parameters. However, it's worth noting that newer and larger models may have been developed since then.



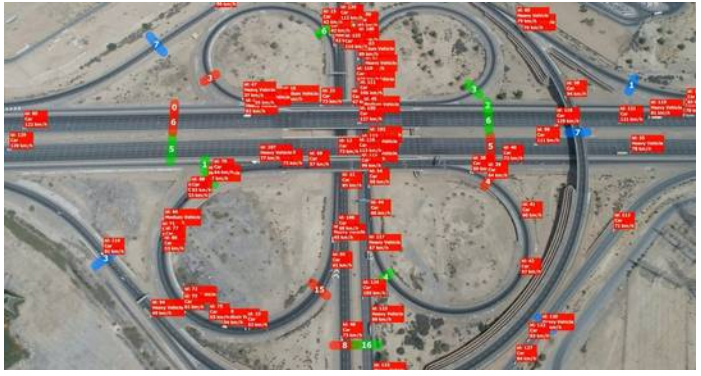
"Machine Learning Model Sizes and the Parameter Gap." arXiv preprint

Machine Learning – computer vision

Object detection / classification, e.g. wildlife

Counting – e.g. traffic, products

Translating – e.g. retrofitting analog meters



(a) Single class FOMO

(b) Multi class FOMO

Figure 25: Comparison of single class vs multiclass FOMO

source: https://blogs.nvidia.com/blog/2019/05/14/rce-systems-ai-improve-traffic-flow/
Energinet Assets

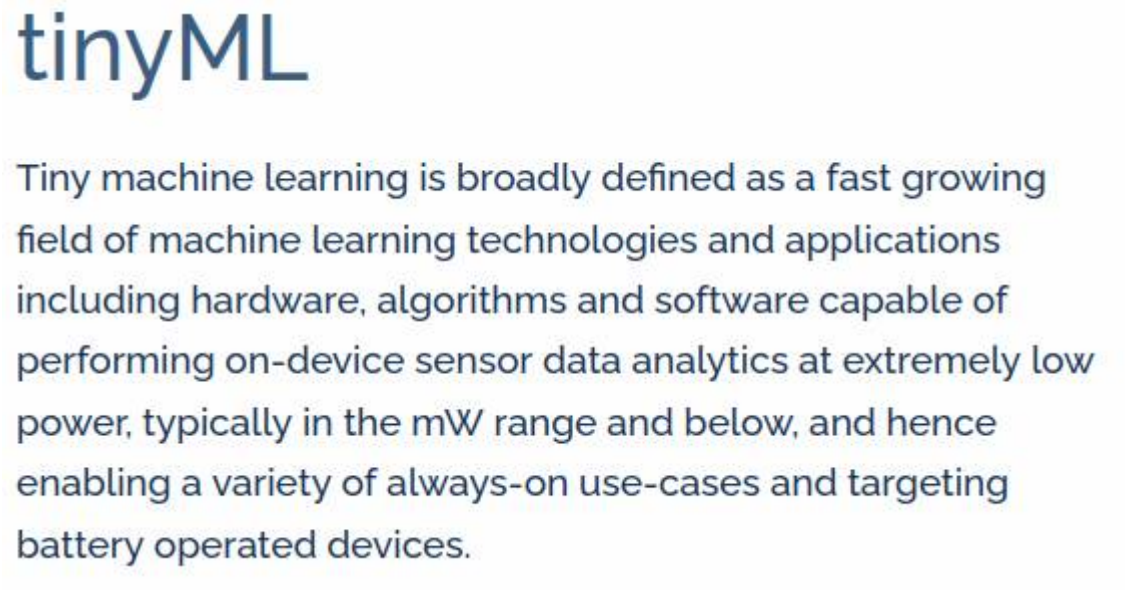
TinyML - background

Source: DASYA Lab / Seeedstudio, Arduino, Google Coral

What is TinyML?

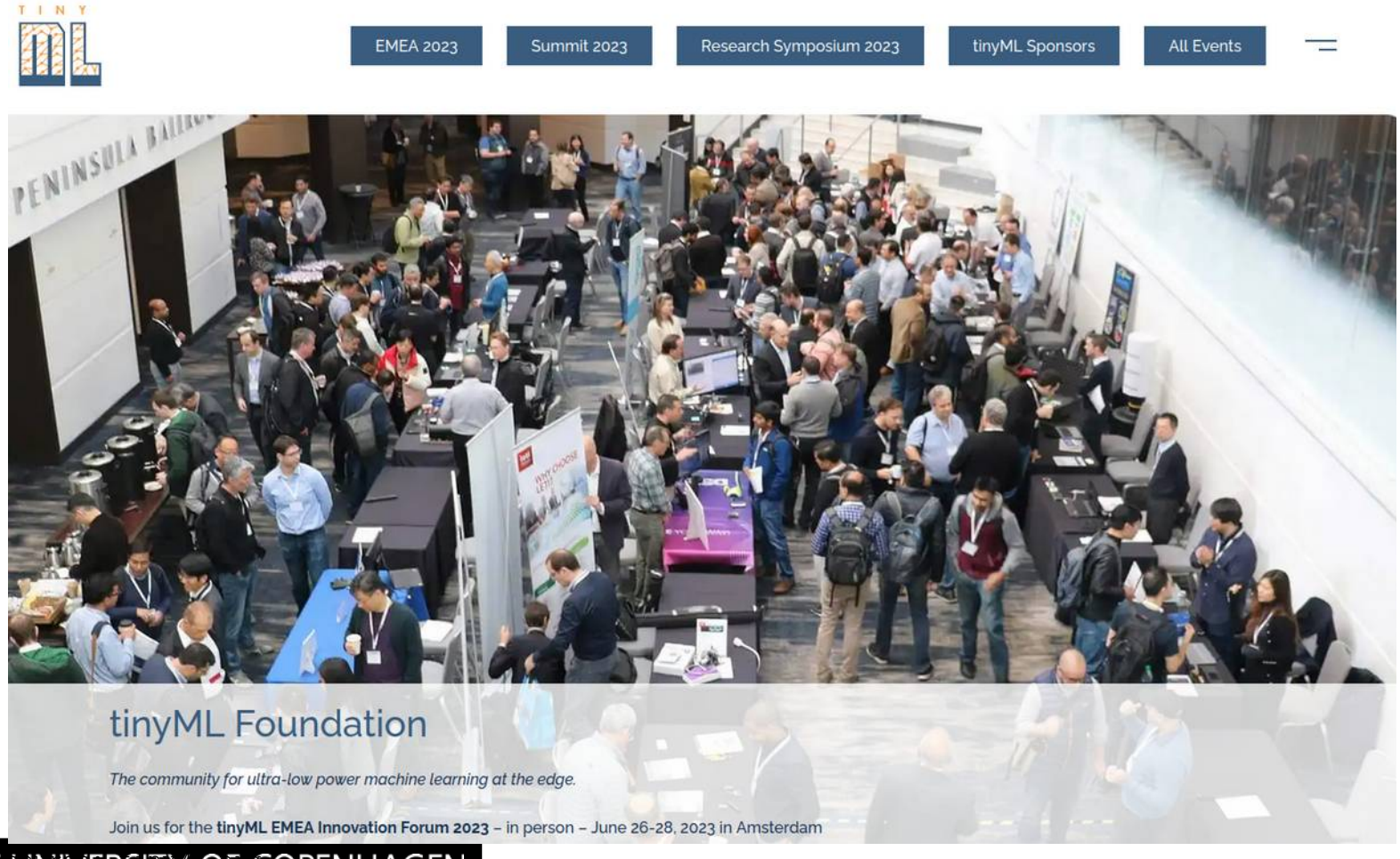
A loosely defined form of embedded ML

embedded systems – single purpose, constrained, “small” – not general purpose



Note the power range ... **mW** … battery (and below)

What is TinyML.org?



Why TinyML?

(when you could have full ML?)

TinyML makes sense wherever …

… you want fast local decisions

… moving data is costly

… resources are **constrained**

these resources might be: space, power, network, budgets, ...

Another Motivation for going Tiny

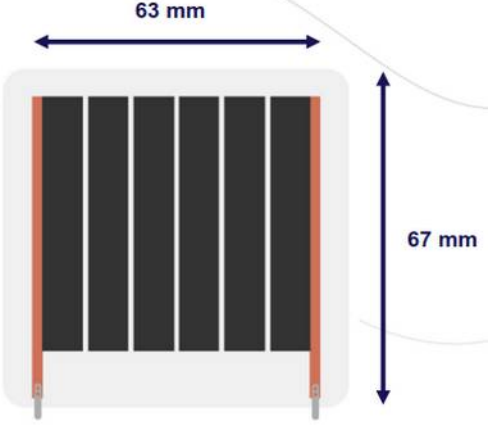
Low Power IoT
Sustainability
TinyML is (soon) becoming an option for battery-less, energy-harvesting approaches to IoT

DRACULA TECHNOLOGIES

Standard demokit performance

DEMOKIT #6 PERFORMANCES BETWEEN 50 - 1000 LUX

Illumination (lux)	Voc (V)	Isc (µA)	Vmax (V)	Imax (µA)	Pmax (µW)
50	3	17	2,35	13	31
200	3,4	72	2,7	59	160
500	3,65	171	2,8	138	386
1000	3,7	322	2,85	263	750



source: Dracula OPV

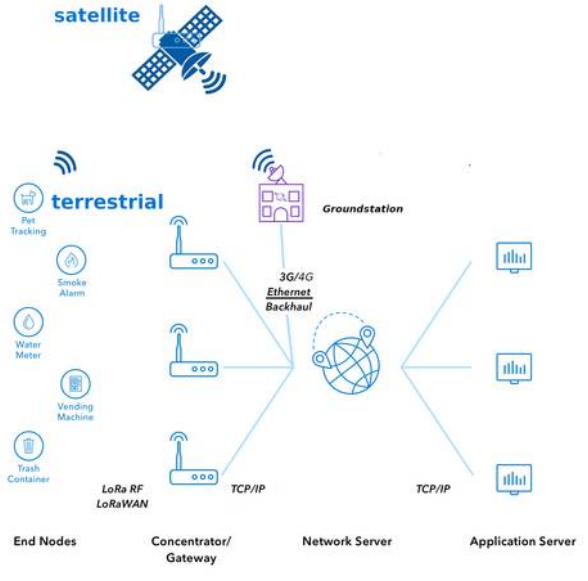
TinyML and Edge Computing

TinyML is one ingredient of the larger concept of **Edge Computing.**

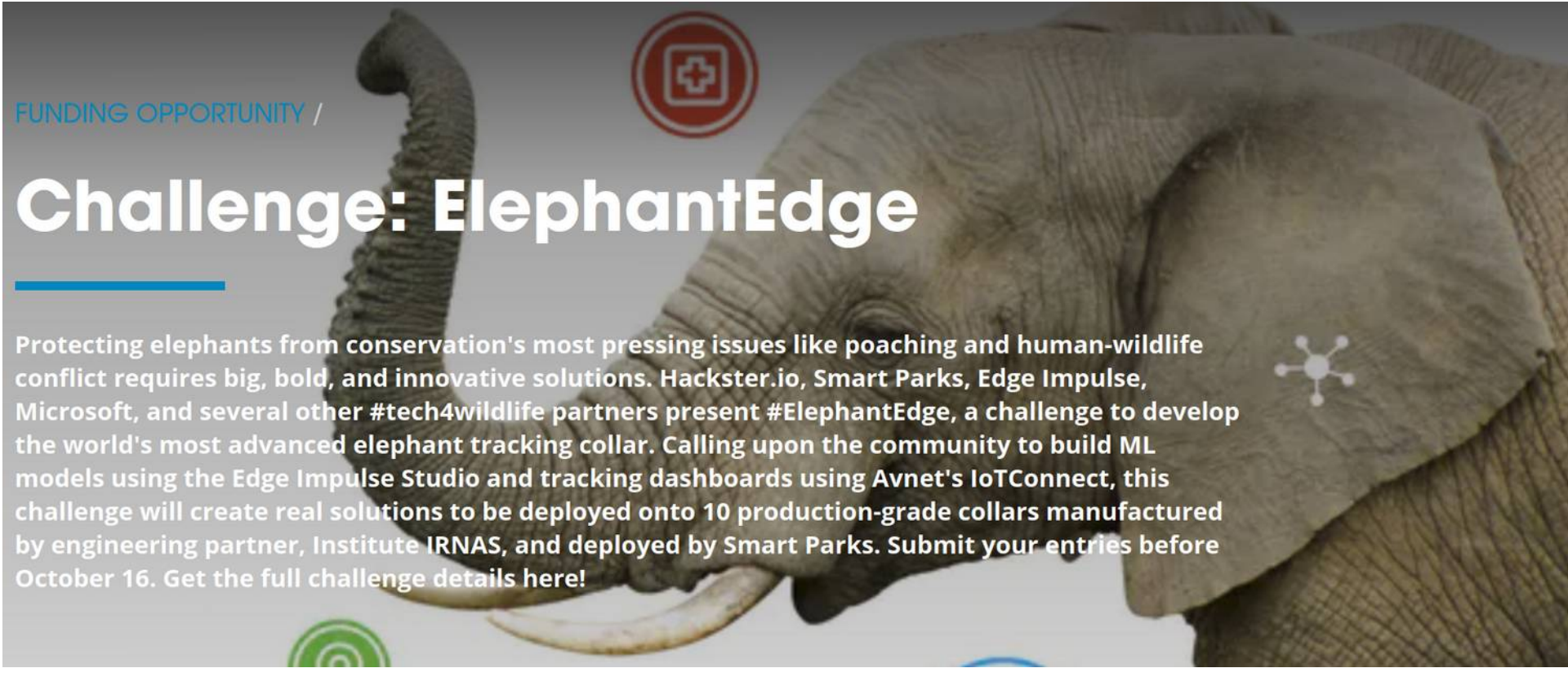
Careful - depending on who you ask,
there are many definitions of Edge.

Some mean
Edge of privileged network (base station, gateway, …)

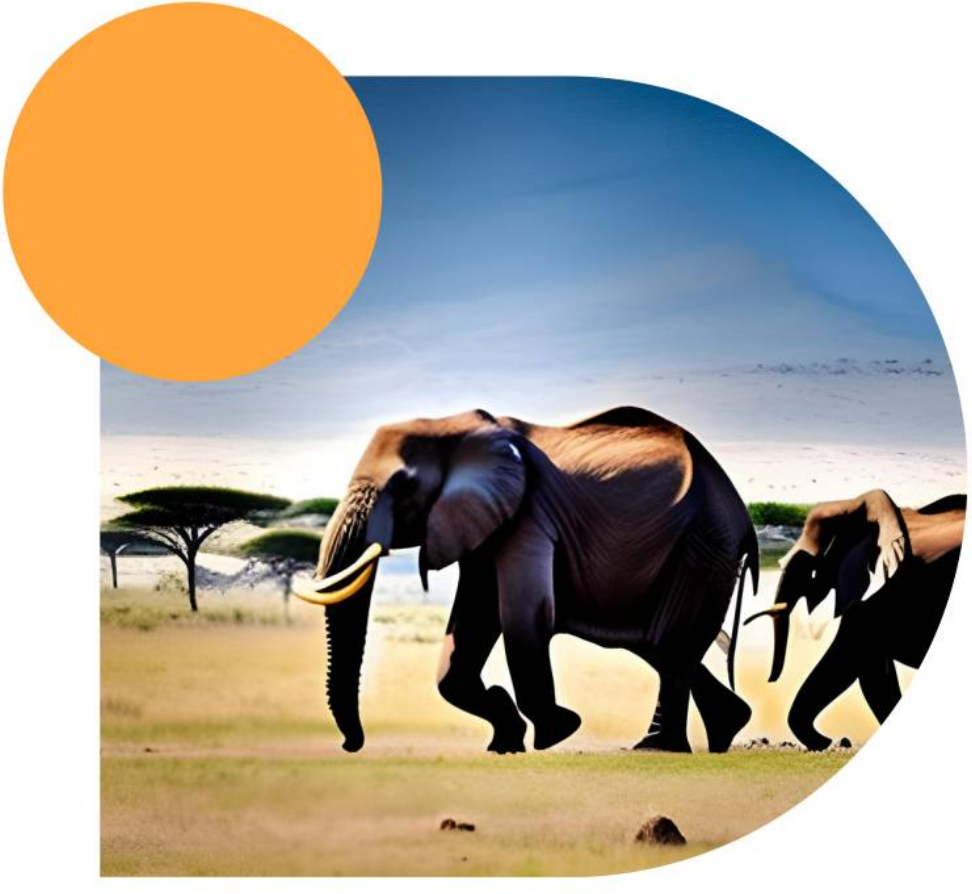
Others mean
Out on the node



Where and for what?



Where and for what?



Poaching

- ElephantEdge Collar
- Using IMUs to determine the state of activity of animals
- Using microphones to determine trumpeting

source: https://tinyml.seas.harvard.edu/SciTinyML-23/ , Robert Thas John

Where and for what?

Illegal Logging

- Dedicated MCU listens for chainsaw sounds

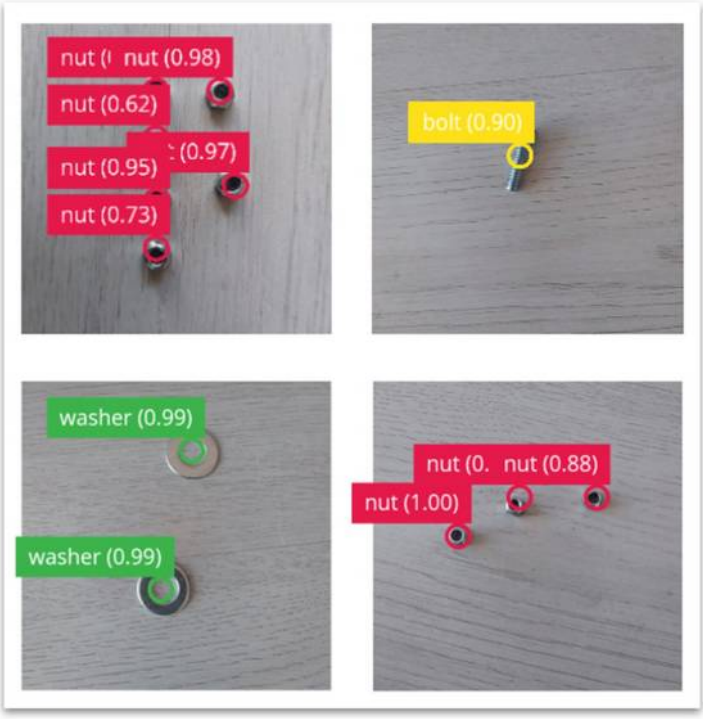
source: https://tinyml.seas.harvard.edu/SciTinyML-23/ , Robert Thas John

Where and for what?

there are a lot os less spectacular use cases ...

Where and for what?

there are a lot os less spectacular use cases ...



source: edgeimpulse.com / Garcia, Manuel B., Shaneth Ambat, and Rossana T. Adao. "Tomayto, tomahto: A machine learning approach for tomato ripening stage identification using pixel-based color image classification." 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM). IEEE, 2019. / https://www.kaggle.com/code/kaustubhb999/tomato-leaf-disease-detection-using-cnn/notebook

Hardware for TinyML

(non-systematic list)

Arduino

ARM

NVIDIA Jetson / Orin

(Mobile Phones)

(Raspberry Pi)

Officially supported MCU targets

- Alif Ensemble E7
- Arduino Nano 33 BLE Sense
- Arduino Nicla Sense ME
- Arduino Nicla Vision
- Arduino Nicla Voice
- Arduino Portenta H7 + Vision Shield
- Espressif ESP32
- Himax WE-I Plus
- Infineon CY8CKIT-062-BLE Pioneer Kit
- Infineon PSoC 62S2 Wi-Fi BT Pioneer Kit
- Nordic Semi nRF52840 DK
- Nordic Semi nRF5340 DK
- Nordic Semi nRF9160 DK
- Nordic Semi Thingy:53
- Nordic Semi Thingy:91
- OpenMV Cam H7 Plus
- Renesas CK-RA6M5 Cloud Kit
- Seeed Grove Vision AI Module
- Seeed SenseCAP A1101
- Silicon Labs xG24 Dev Kit
- Silicon Labs Thunderboard Sense 2
- Sony's Spresense
- ST B-L475E-IOT01A
- Synaptics Katana EVK
- Syntiant TinyML Board
- TI CC1352P Launchpad
- Raspberry Pi RP2040

Officially supported CPU/GPU targets

- Intel Based Macs
- Linux x86_64
- NVIDIA Jetson Nano
- Raspberry Pi 4
- Renesas RZ/V2L
- Texas Instruments SK-TDA4VM

Officially supported AI accelerators

- BrainChip AKD1000

Production targets

- Advantech ICAM-500
- Advantech MIC AI Series
- MCS AI Gateway 4434S

Community targets

- Arducam Pico4ML TinyML Dev Kit
- Blues Wireless Swan
- RAKwireless WisBlock
- Seeed Wio Terminal
- Seeed reComputer Jetson
- Seeed Studio XIAO nRF52840 Sense
- Texas Instruments SK-AM62

com/docs/development-platforms/fully-supported-development-boards

Arduinos for TinyML

Nano 33 BLE Sense

Nicla
- Sense ME
- Vision
- Voice

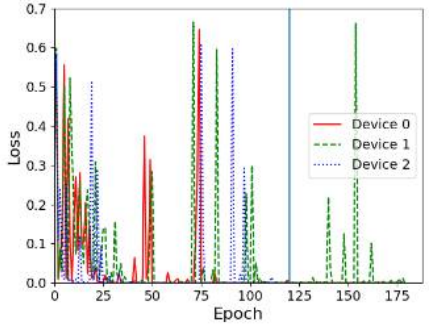
Portenta H7

Arduinos for TinyML

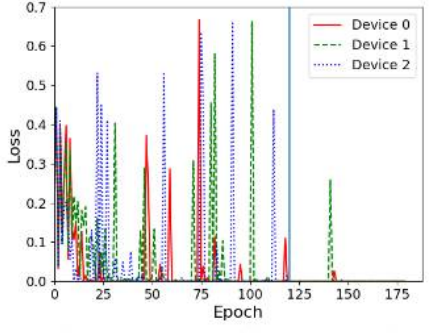
Nano

vs

Portenta H7



(a) Arduino Nano board: training on three boards a neural network with hidden layer of 25 neurons.



(b) Arduino Portenta board: training on three boards a neural network with hidden layer of 70 neurons.

Figure 2: Loss vs. epochs during training on both types of board the neural network with different hidden layer size.

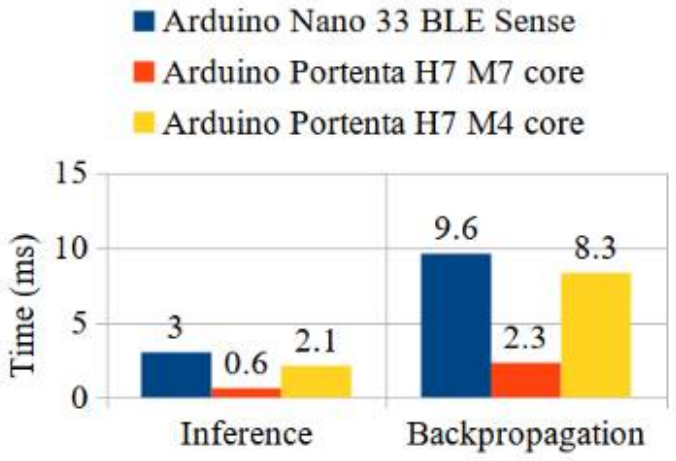


Figure 3: Inference and backpropagation times for both boards and M7 and M4 cores in Adruino Portenta, respectively.

t al. "Comparison of Two Microcontroller Boards for On-Device Model Training in a Mediterranean Conference on Embedded Computing (MECO). IEEE, 2022.

Hardware for TinyML

Hardware

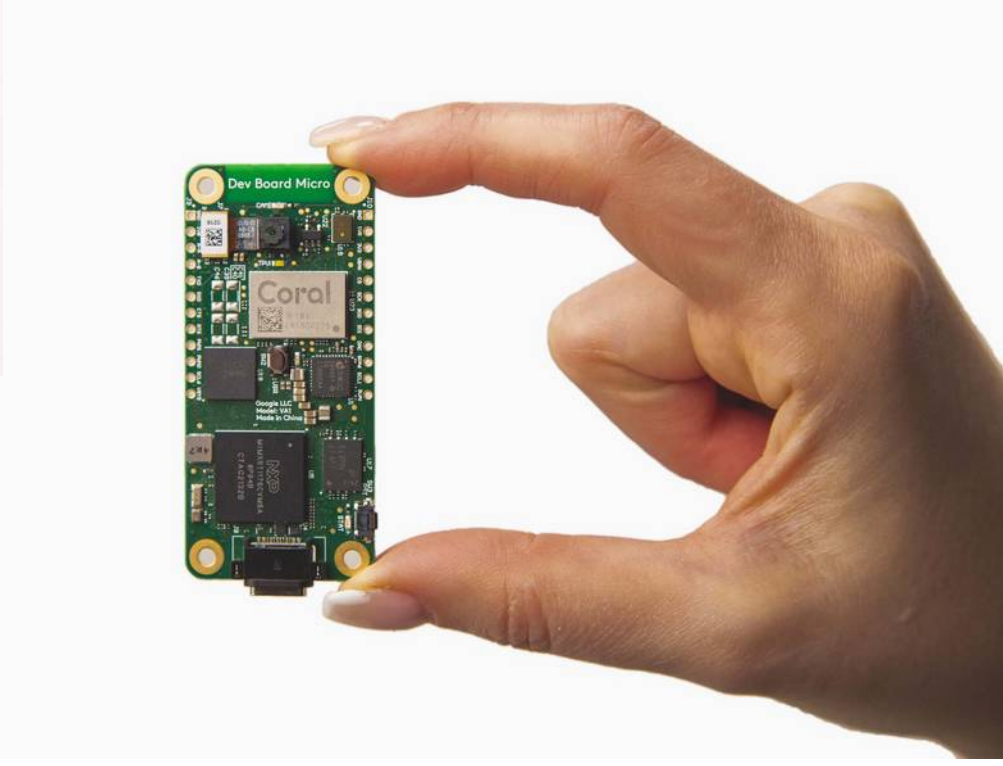



	Raspberry Pico (W)	Arduino Nano Sense	ESP 32	Seeed XIAO Sense / ESP32S3	Arduino Pro
32Bits CPU	Dual-core Arm Cortex-M0+	Arm Cortex-M4F	Xtensa LX6 Dual Core	Arm Cortex-M4F (BLE) Xtensa LX7 Dual Core	Dual Core Arm Cortex M7/M4
CLOCK	133MHz	64MHz	240MHz	64 / 240MHz	480/240MHz
RAM	264KB	256KB	520KB (part available)	256KB / 8MB	1MB
ROM	2MB	1MB	2MB	2MB / 8MB	2MB
Radio	(Yes for W)	BLE	BLE/WiFi	BLE / WiFi (ESP32S3)	BLE/WiFi
Sensors	No	Yes	No	Yes (Sense)	Yes (Nicla)
Bat. Power Manag.	No	No	No	Yes	Yes
Price	$	$$$	$	$$	$$$$$

https://media.digikey.com/Resources/Maker/the-original-guide-to-boards-2022.pdf

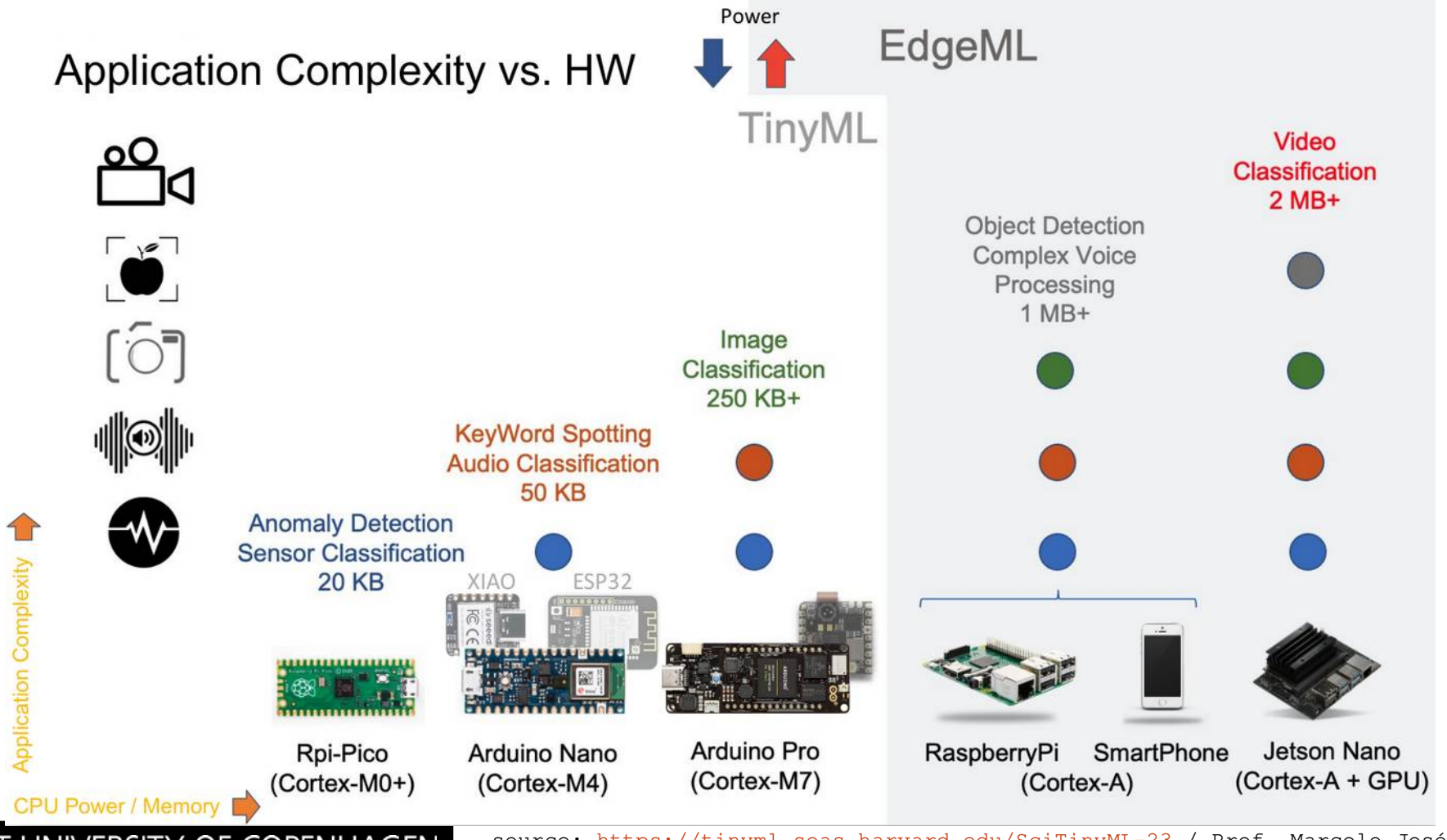
source: https://tinyml.seas.harvard.edu/SciTinyML-23 / Prof. Marcelo José Rovai

Hardware for TinyML: Google Coral TPU



source: https://coral.ai/docs

Hardware for TinyML



source: https://tinyml.seas.harvard.edu/SciTinyML-23 / Prof. Marcelo José Rovai

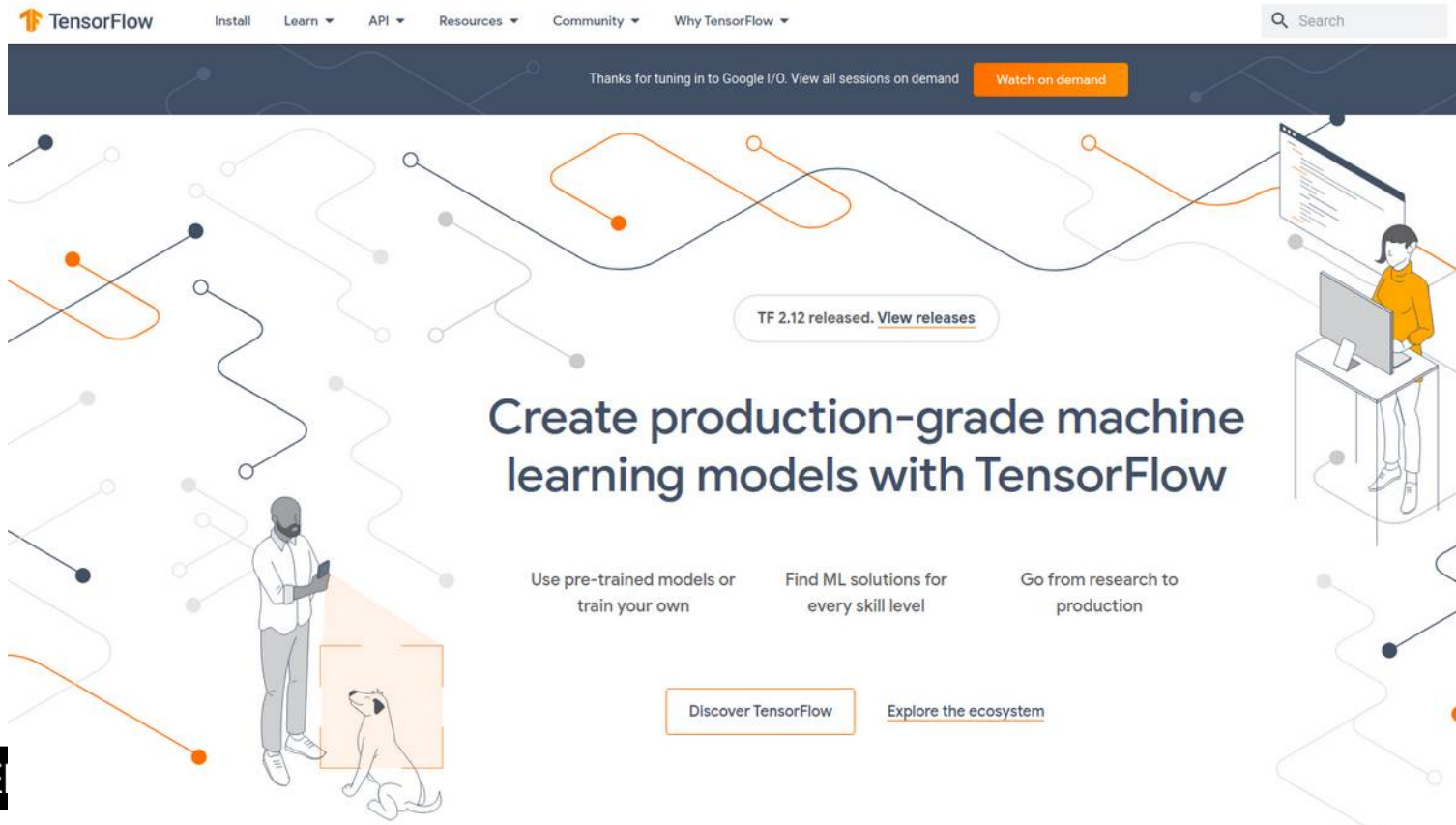
Software for TinyML

Machine learning is to a large extent

python based,

e.g. using

TensorFlow



Software for TinyML: Tensorflow Lite

for MCUs, in particular ARM Cortex, ESP32

TensorFlow Lite for Microcontrollers

TensorFlow Lite for Microcontrollers is designed to run machine learning models on microcontrollers and other devices with only a few kilobytes of memory. The core runtime just fits in 16 KB on an Arm Cortex M3 and can run many basic models. It doesn't require operating system support, any standard C or C++ libraries, or dynamic memory allocation.

★ **Note:** The TensorFlow Lite for Microcontrollers Experiments features work by developers combining Arduino and TensorFlow to create awesome experiences and tools. Check out the site for inspiration to create your own TinyML projects.

source: https://www.tensorflow.org/lite/microcontrollers

Software for TinyML: Tensorflow Lite, hardware

Supported platforms

TensorFlow Lite for Microcontrollers is written in C++ 17 and requires a 32-bit platform. It has been tested extensively with many processors based on the Arm Cortex-M Series architecture, and has been ported to other architectures including ESP32. The framework is available as an Arduino library. It can also generate projects for development environments such as Mbed. It is open source and can be included in any C++ 17 project.

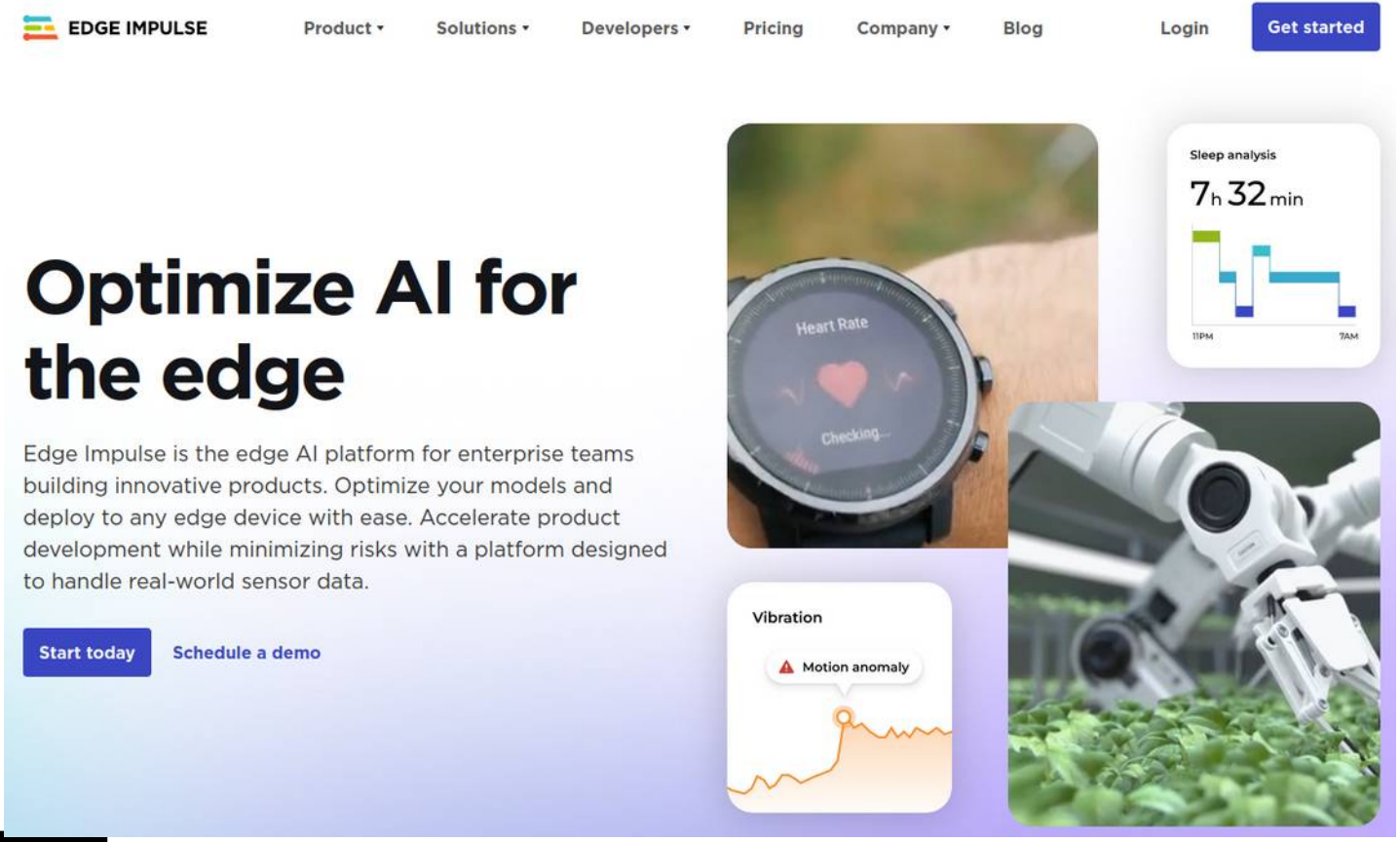
The following development boards are supported:

- Arduino Nano 33 BLE Sense
- SparkFun Edge
- STM32F746 Discovery kit
- Adafruit EdgeBadge
- Adafruit TensorFlow Lite for Microcontrollers Kit
- Adafruit Circuit Playground Bluefruit
- Espressif ESP32-DevKitC
- Espressif ESP-EYE
- Wio Terminal: ATSAMD51
- Himax WE-I Plus EVB Endpoint AI Development Board
- Synopsys DesignWare ARC EM Software Development Platform
- Sony Spresense

source: https://www.tensorflow.org/lite/microcontrollers

Software and platforms for TinyML: Edgeimpulse

Edgeimpulse is a platform offering an easy-to-use entry into Edge ML



Software and platforms for TinyML: Edgeimpulse

EDGE IMPULSE: AN MLOPS PLATFORM FOR TINY MACHINE LEARNING

Shawn Hymel * **Colby Banbury** * **Daniel Situnayake** **Alex Elium** **Carl Ward** **Mat Kelcey** **Mathijs Baaijens** **Mateusz Majchrzycki** **Jenny Plunkett** **David Tischler** **Alessandro Grande** **Louis Moreau** **Dmitry Maslov** **Artie Beavis** **Jan Jongboom** **Vijay Janapa Reddi**

ABSTRACT

Edge Impulse is a cloud-based machine learning operations (MLOps) platform for developing embedded and edge ML (TinyML) systems that can be deployed to a wide range of hardware targets. Current TinyML workflows are plagued by fragmented software stacks and heterogeneous deployment hardware, making ML model optimizations difficult and unportable. We present Edge Impulse, a practical MLOps platform for developing TinyML systems at scale. Edge Impulse addresses these challenges and streamlines the TinyML design cycle by supporting various software and hardware optimizations to create an extensible and portable software stack for a multitude of embedded systems. As of Oct. 2022, Edge Impulse hosts 118,185 projects from 50,953 developers.

source: Hymel, S., Banbury, C., Situnayake, D., Elium, A., Ward, C., Kelcey, M., ... & Reddi, V. J. (2022). Edge Impulse: An MLOps Platform for Tiny Machine Learning. arXiv preprint arXiv:2212.03332.

Software and platforms for TinyML: Edgeimpulse

TinyML for all developers

C++ library

ARDUINO

Arduino library

WA

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time
Open source SDK

Impulse

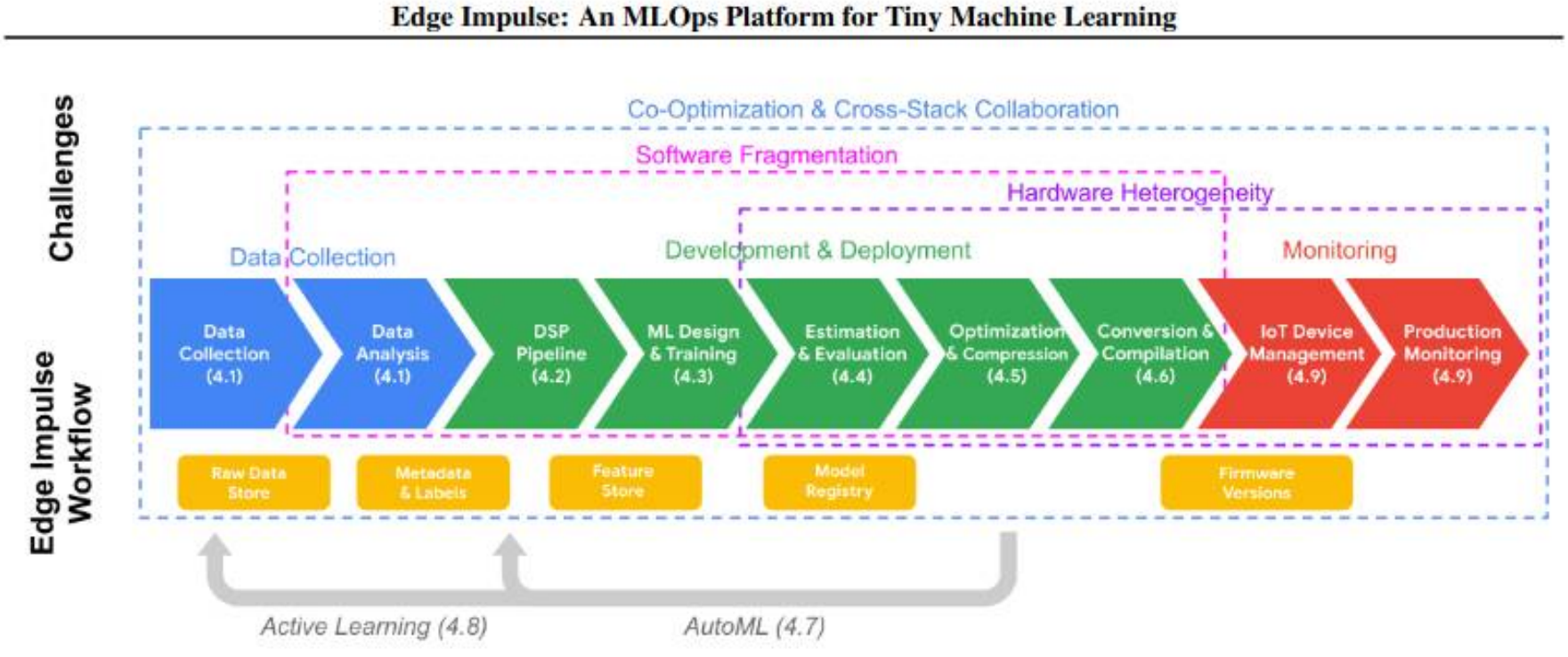
Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test

www.edgeimpulse.com

Software and platforms for TinyML: Edgeimpulse



source: Hymel, S., Banbury, C., Situnayake, D., Elium, A., Ward, C., Kelcey, M., ... & Reddi, V. J. (2022). Edge Impulse: An MLOps Platform for Tiny Machine Learning. arXiv preprint arXiv:2212.03332.

Software and platforms for TinyML: Edgeimpulse

Edge Impulse: An MLOps Platform for Tiny Machine Learning

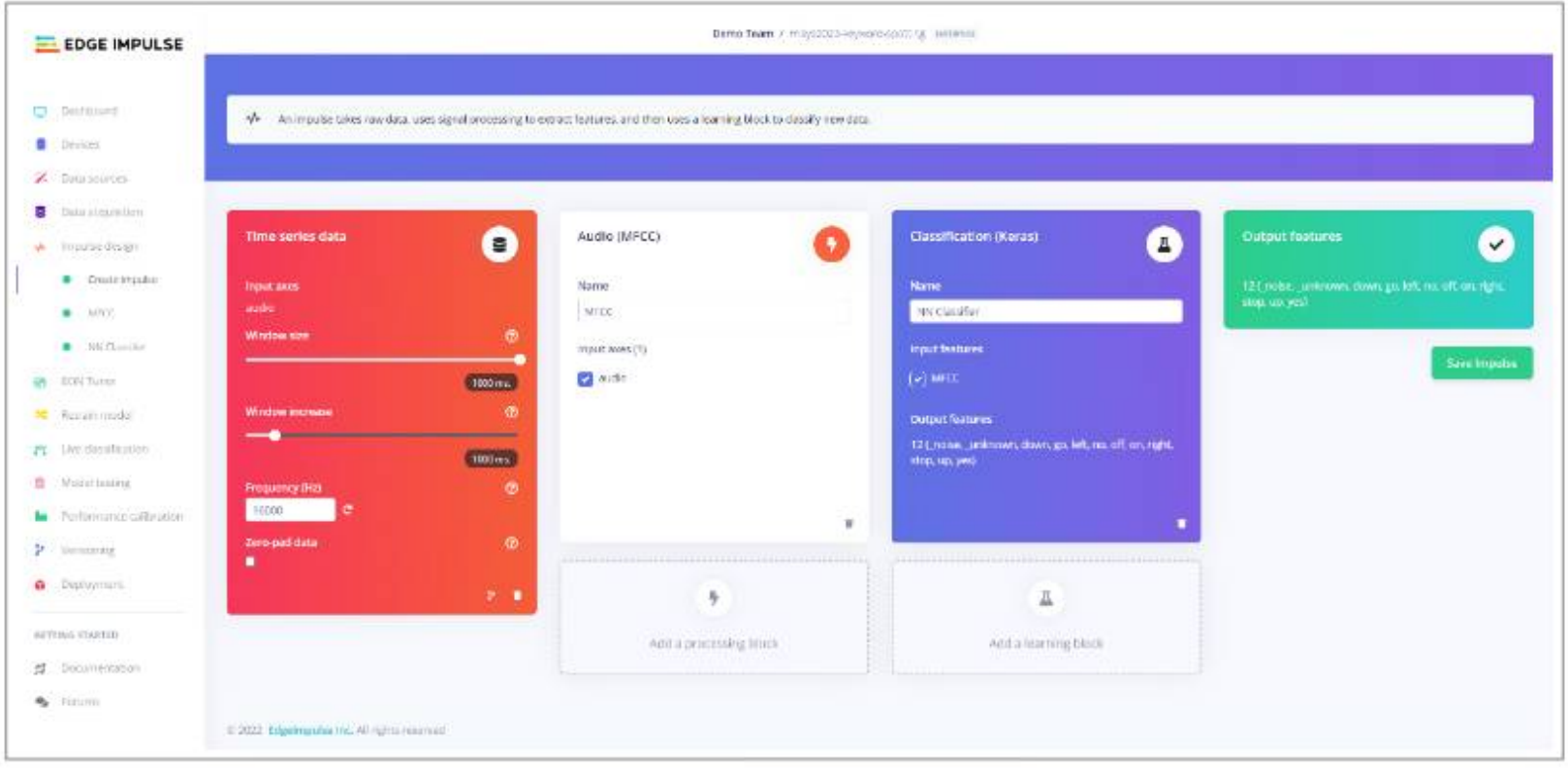


Figure 2. Screenshot showing the user's view inside an Edge Impulse project where the *blocks* are connected depicting the dataflow.

source: Hymel, S., Banbury, C., Situnayake, D., Elium, A., Ward, C., Kelcey, M., ... & Reddi, V. J. (2022). Edge Impulse: An MLOps Platform for Tiny Machine Learning. arXiv preprint arXiv:2212.03332.

Edgeimpulse - features

FOMO

Faster Objects, More Objects – a model for embedded object detection

python SDK

allows for export/import and interfacing with Tensorflow
and all kinds of python based ML

Fun projects: guitars (product - identification)

Inferencing...

gibson

Time per inference: 6 ms.

GIBSON

0.94

FENDER

0.06

Inferencing...

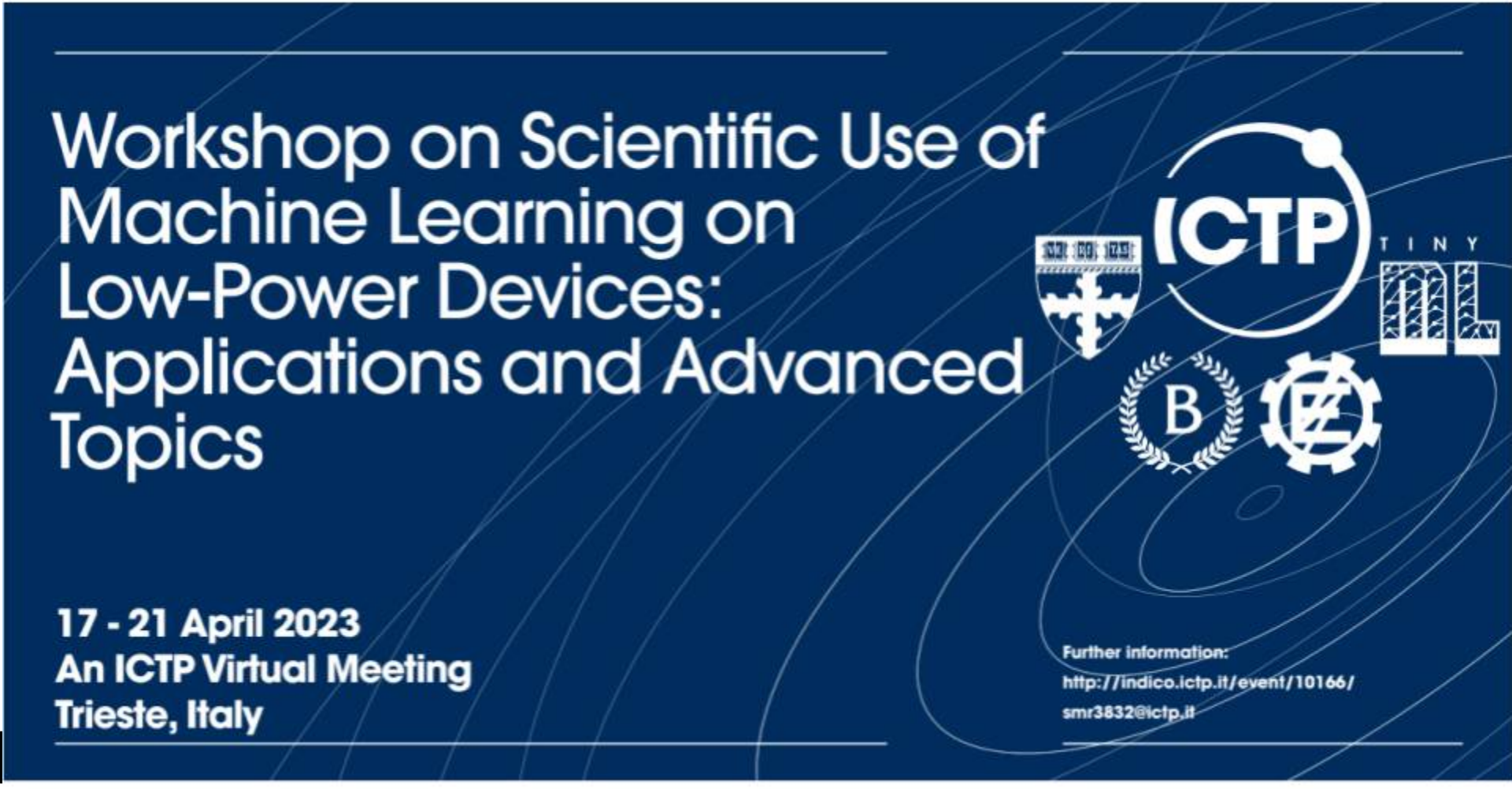
fender

Time per inference: 6 ms.

FENDER

0.85

TinyML in research and education, globally



TinyML @ ITU: some projects and lessons learned

Retrofitting analog meters (Bachelor project, ongoing)

Birdcams (Master thesis)

Fishcam (ongoing)

DISCOSAT (satellite based ML)

Project MOTH

source: edgeimpulse.com

TinyML @ ITU: Retrofitting analog meters (ongoing)

Recognize the meter
Fix perspectives, angles
Identify the pointer

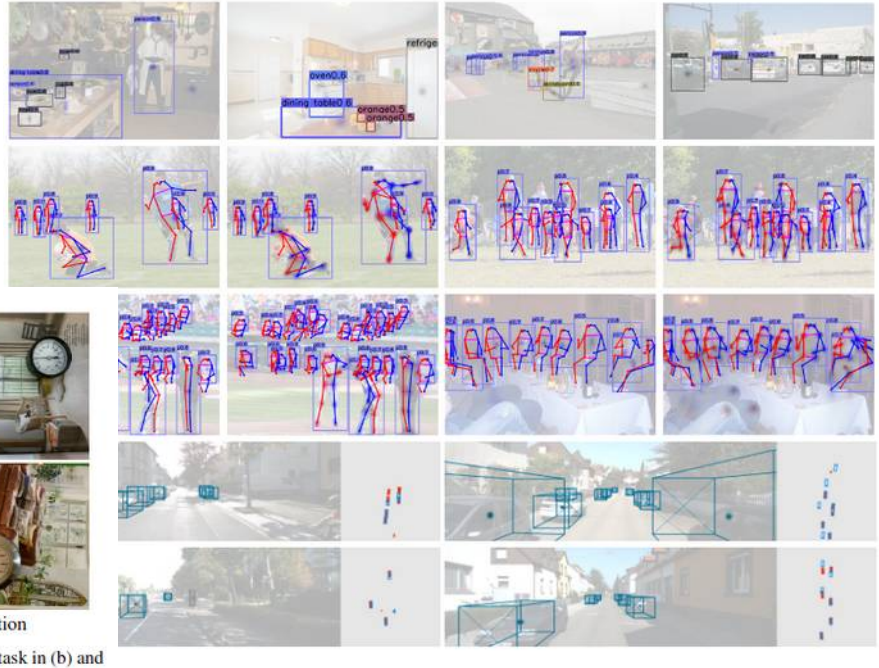


Figure 5. The 6 meters used for the Real-Gauges dataset are shown in (a), example data collected for the perspective recover task in (b) and examples of images used for gauge detection in (c).

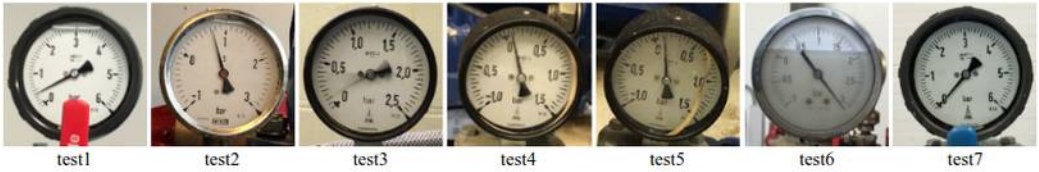
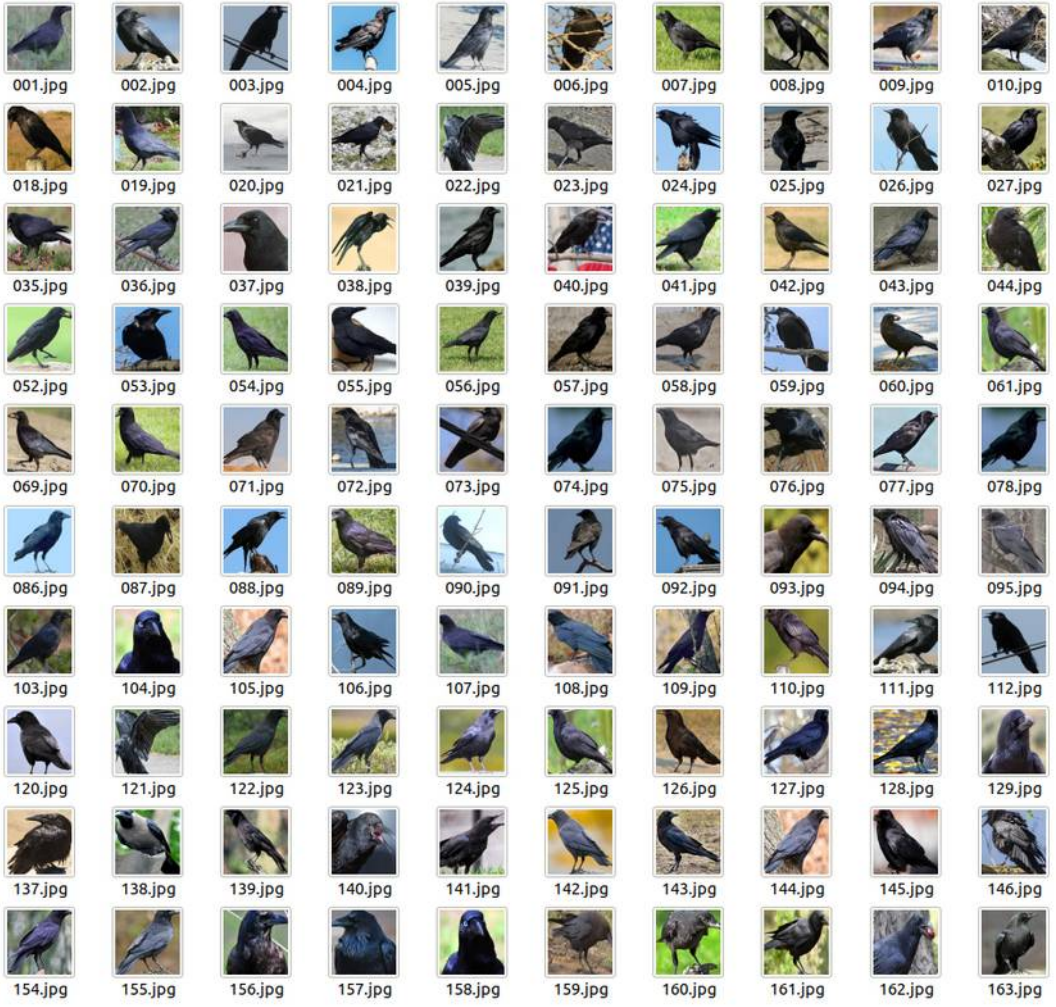


Figure 6. Example meter crops from the 7 test videos of the Kaggle-Dataset.

source: Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." arXiv preprint arXiv:1904.07850 (2019). / Howells, Ben, James Charles, and Roberto Cipolla. "Real-time analogue gauge transcription on mobile phone." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

TinyML @ ITU: Birdcams & Scarecrows

Data science is easy –
when your data comes from Kaggle.
Else, **not so much** …



TinyML @ ITU: Birdcams & Scarecrows

Real life data is hard!

Also, remember twigs ...

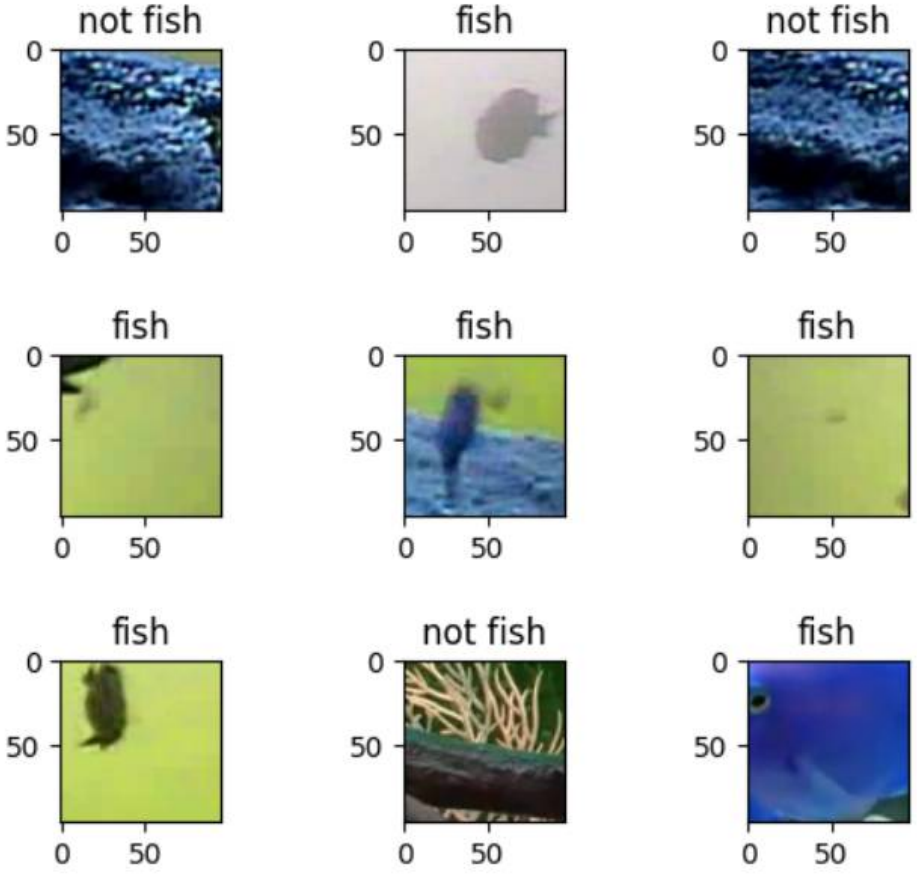
source: Steen Heesche Petersen and Tin Marecic

TinyML @ ITU: Fishcam, Coral Reefs

Project (to be reactivated)
with Den Blå Planet, Partners in Zanzibar, Kenya, California (MBARI), others …

(a) Bounding box (b) All windows (c) Windows labeled "fish"

TinyML @ ITU: Fishcam, Coral Reefs



source: DASYA, Jens Joergensen

Figure 15: Sample of the training data for the classifier

TinyML @ ITU: Fishcam, Coral Reefs

(a) Input image (b) Bounding box result

Figure 16: Bounding box object detection result

TinyML @ ITU: Fishcam, Coral Reefs

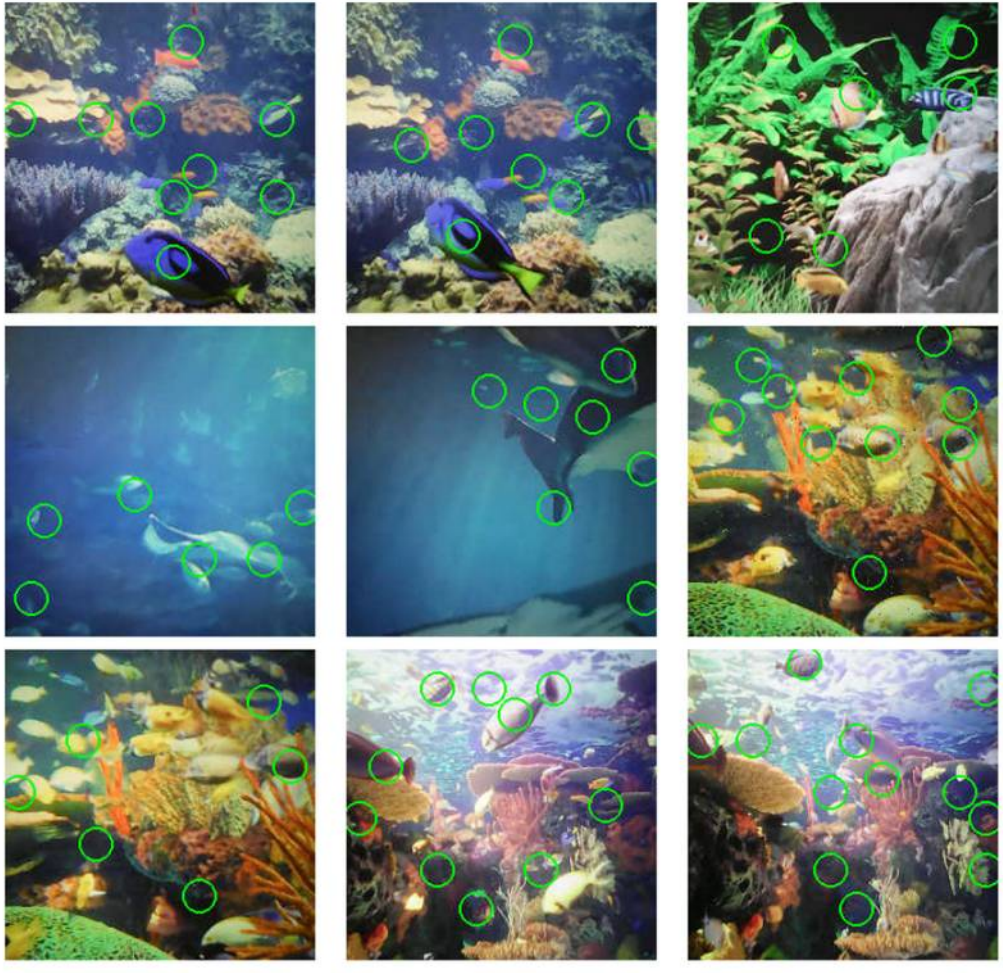


Figure 24: Examples of images with marked fish

source: DASYA, Jens Joergensen

TinyML @ ITU: Fishcam, Coral Reefs

(a) Single class FOMO

(b) Multi class FOMO

Figure 25: Comparison of single class vs multiclass FOMO

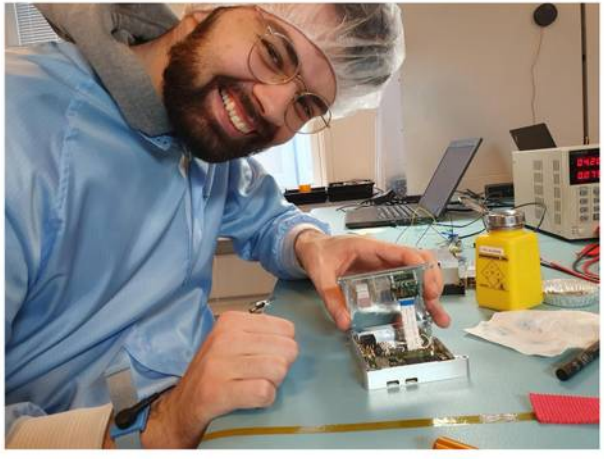
source: DASYA, Jens Joergensen, Sebastian Büttrich

TinyML on satellite: power

source: https://news.lockheedmartin.com/news-releases?item=128962, Render of La Jument nanosatellite.
Courtesy: University of Southern California – Satellite flying a Nvidia Jetson system

TinyML @ ITU: DISCOSAT (satellite based ML)

DISCO1 satellite



Students launch a satellite to test artificial intelligence in space

On April 14, students from ITU will contribute to writing space history. The satellite, DISCO-1, is launched into space and it carries a microcomputer to test artificial intelligence outside the atmosphere. The satellite is developed by the space program, DISCO, which is a collaboration between students from four Danish universities.

IT-Universitetet i København
April 15 ·

Så lykkedes det! 🚀🛰️✨

Satellitten DISCO-1, udviklet af danske studerende fra bl.a. ITU, blev her til morgen sendt ud i rummet med SpaceX' raket fra Californien.

Satellitten indeholder en mikrocomputer, der skal teste kunstig intelligens i rummet. 🤗

Læs mere om projektet her 👉 https://www.itu.dk/.../Studerende-opsender-satellit-der...

Julian Priest (CC BY-NC 3.0)

source: discosat.dk

TinyML @ ITU: DISCOSAT (satellite based ML)

source: discosat.dk

3/ Critical reading:

Is TinyML sustainable?

Is TinyML Sustainable?

Assessing the Environmental Impacts of Machine Learning on Microcontrollers

Shvetank Prakash
Harvard University
sprakash@g.harvard.edu, USA

Matthew Stewart
Harvard University
matthew_stewart@g.harvard.edu
USA

Colby Banbury
Harvard University
cbanbury@g.harvard.edu, USA

Mark Mazumder
Harvard University
markmazumder@g.harvard.edu, USA

Pete Warden
Stanford University
petewarden@stanford.edu, USA

Brian Plancher
Barnard College, Columbia University
bplancher@barnard.edu, USA

Vijay Janapa Reddi
Harvard University
vj@eecs.harvard.edu, USA

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

TinyML4D: Scaling Embedded Machine Learning Education in the Developing World

Brian Plancher , Sebastian Büttrich , Jeremey Ellis , Neena Goveas , Laila Kazimierski , Jesus Lopez Sotelo , Milan Lukic , Diego Mendez Chaves , Rosdiadee Nordin , Andres Oliva Trevisan , Massimo Pavan , Manuel Roveri , Marcus Rüb , Jackline Tum , Marian Verhelst , Salah Abdeljabar , Segun Adebayo , Thomas Amberg , Halleluyah Aworinde , José Antonio Bagur Nájera , Gregg Barrett , Nabil Benamar , Bharat Chaudhari , Ronald Criollo , David Cuartielles , JoseA Iberto Ferreira Filho , Solomon Gizaw , Evgeni Gousev , Alessandro Grande , Shawn Hymel , Peter Ing , Prashant Manandhar , Pietro Manzoni , Boris Murmann , Eric Pan , Rytis Paskauskas , Ermanno Pietrosemoli , Tales Pimenta , Marcelo Rovai , Marco Zennaro , Vijay Janapa Reddi

March 2024

Source: Plancher, Brian, et al. "TinyML4D: Scaling Embedded Machine Learning Education in the Developing World." Proceedings of the AAAI Symposium Series. Vol. 3. No. 1. 2024.

Is TinyML sustainable?

Approach:

Start statement: **IoT might make things worse**

“the question we must ask ourselves is
do we run the risk of producing an
Internet of Trash
over the course of TinyML devices’ lifetime?”

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

The reality of IoT

Is TinyML sustainable?

TinyML is seen as bearing a promise
of positive impact
with regards to SDGs. Why?

Platform	Freq.	Memory	Storage	Power	Price	CO_2-eq Footprint
Cloud	GHz	10+GB	TBs-PBs	~1 kW	$1000+	Hundreds of kgs
Mobile	GHz	Few GB	GBs	~1 W	$100+	Tens of kgs
Tiny	**MHz**	**KBs**	**Few MB**	**~1 mW**	**$10**	**Single kgs**

Table 1. Cloud and mobile ML systems compared with TinyML across frequency, memory, storage, power, price, and footprint. The footprint of TinyML systems is far less.

If TinyML has significant negative footprint,
let us look at offsetting it by
positive handprint.

→ Lifecycle Assessment

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

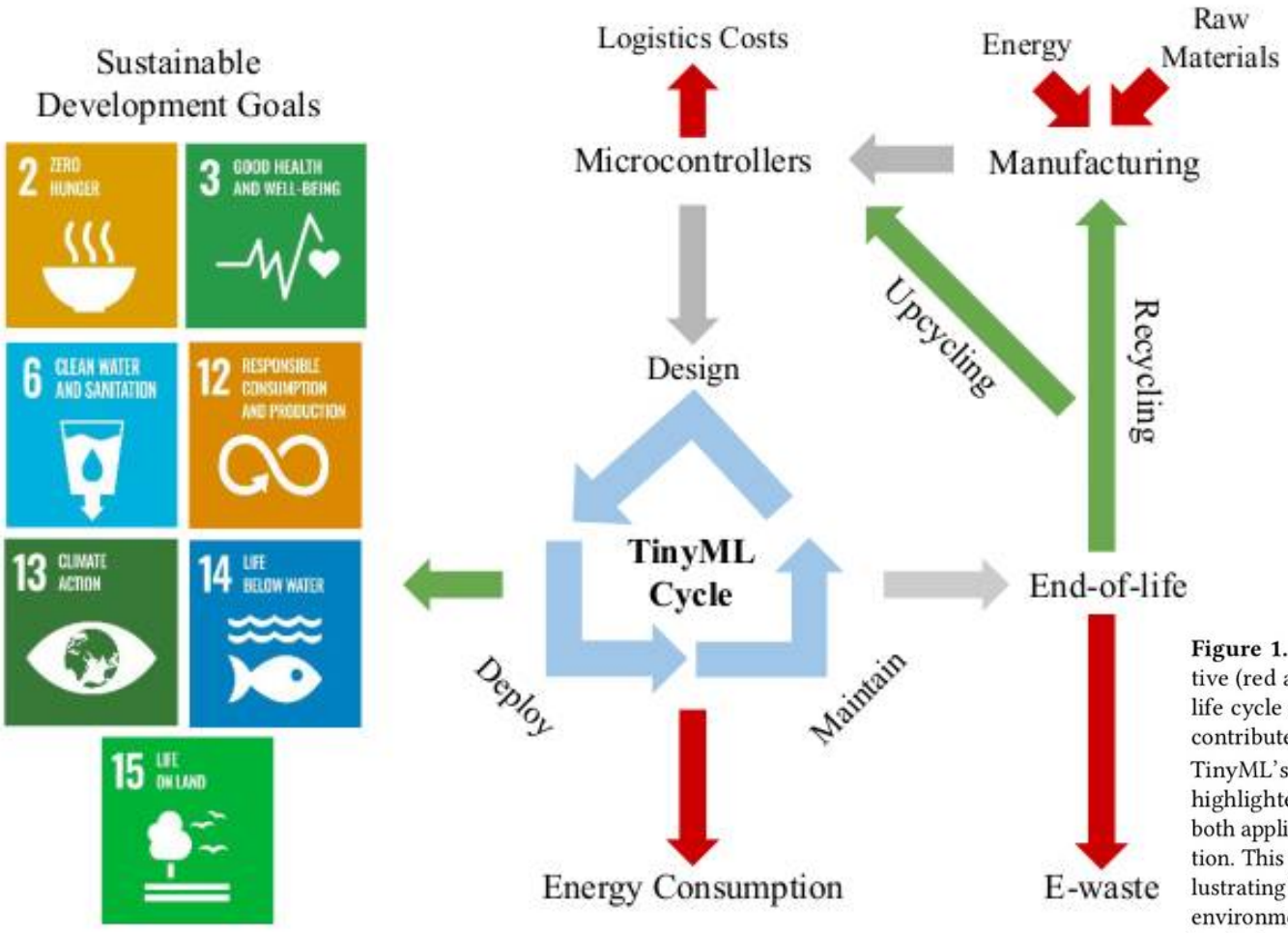


Figure 1. We show the positive (green arrows) and negative (red arrows) environmental footprint of the complete life cycle of TinyML systems as well as how TinyML can contribute to the UN's environmental sustainability goals. TinyML's operational benefits for sustainability are often highlighted, it is crucial to consider the entire life cycle of both applications and hardware to ensure a net carbon reduction. This paper contributes by (1) presenting case studies illustrating TinyML's sustainability benefits, (2) examining the environmental impacts of TinyML at both MCU and system levels through a life cycle analysis (LCA), and (3) identifying future research directions for sustainable TinyML.

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Lifecycle Assessment

1. Identify positive handprint

Zero Hunger & Good Health and Well-Being (SDG #2 & #3)

Precision agriculture, plant & animal diseases, harvest efficiency,
Malaria prevention (insect identification)

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Lifecycle Assessment

1. Identify positive handprint

Life on Land & Below Water (SDG #14 & #15)

Wildlife Conservation, conflict mitigation, poaching detection, deforestation detection

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Lifecycle Assessment

1. Identify positive handprint

Climate Action (SDG #13)

side remark:
collision of aspects of sustainability?

we just replaced 150-160 employees …

3.3 Climate Action (SDG #13)

Take urgent action to combat climate change and its impacts.

TinyML is well-suited to efforts aimed at combating climate change and its impacts through environmental monitoring applications. For example, Ribbit Network recently launched an effort to crowdsource the world's largest greenhouse gas emissions dataset through distributed intelligent sensors which enabled cheap, accurate, and actionable local data on emissions. Similarly, the SmartForest project utilizes a remote monitoring system to provide information on tree growth. This replaced the need for 150 – 160 employees to regularly go into the field with a single trip to install the sensors [13], significantly reducing human impact on the ecosystem while increasing data quality.

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Lifecycle Assessment

2. Identify negative footprint

LCA with 5 phases

Raw Materials
Manufacturing
Transportation
Operating
End of Life / E-waste

Focus on the **MCU (Micro Controller Unit)**

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Lifecycle Assessment / MCU

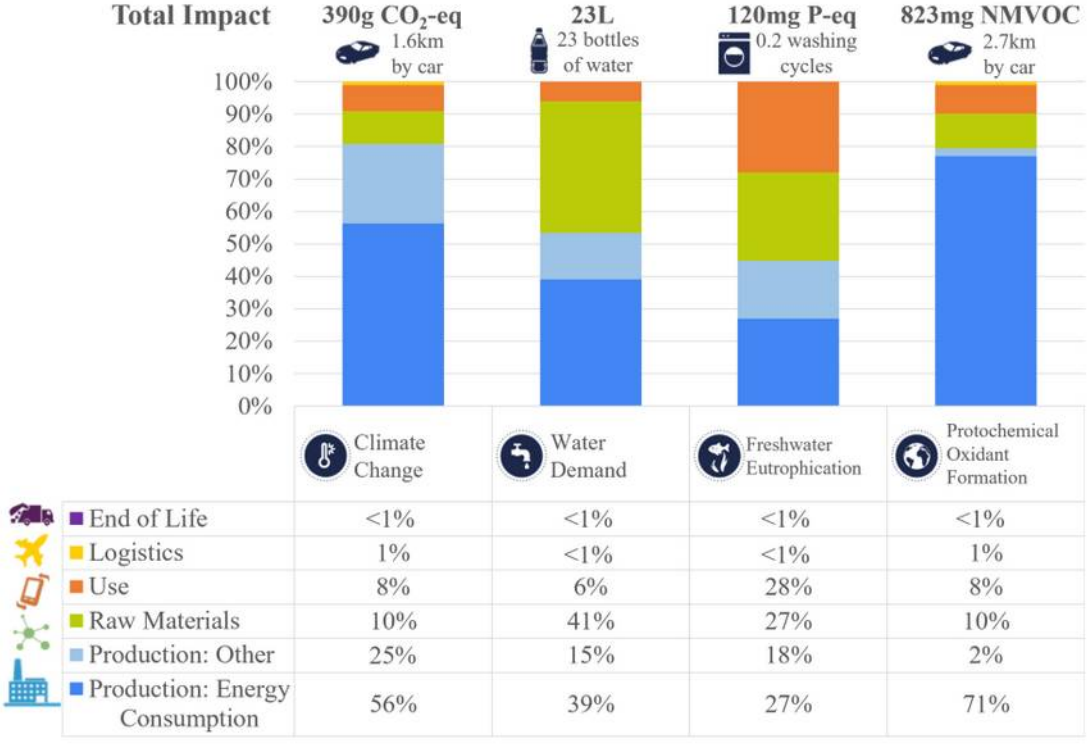


Figure 3. Four different environmental indicators measuring the impact of MCUs on our environment. Each footprint contains both the operational and embodied footprint of the device, including the five-stage life cycle of an MCU. Data courtesy of STMicroelectronics [44]. The data from other MCU providers follow the same operational and embodied footprint trends.

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Raw materials and production outweighs use

This is the same for consumer IT, e.g. a laptop:

The manufacture of a laptop is between 75% – 85% of the overall carbon footprint.

(Attention: lots of diverging data on this!)

Sources: https://circularcomputing.com/news/carbon-footprint-laptop/

A short discourse

Putting 390 grams of CO2 for an MCU in perspective:

The average kWh of electricity in Denmark? *
(Carbon Intensity)

Fossil fuel car per 100 km?

Electric car per 100 km?

* what's your household then?

A short discourse

Putting grams of CO2 in perspective:

The average kWh of electricity in Denmark? (Carbon Intensity)	180 g*
Fossil fuel car per 100 km?	10 kg
Electric car per 100 km?	2.5 kg

* what's your household then?

Sources: https://ourworldindata.org/grapher/carbon-intensity-electricity , https://www.europarl.europa.eu/news/en/headlines/society/20190313STO31218/co2-emissions-from-cars-facts-and-figures-infographics

The MCU is not all

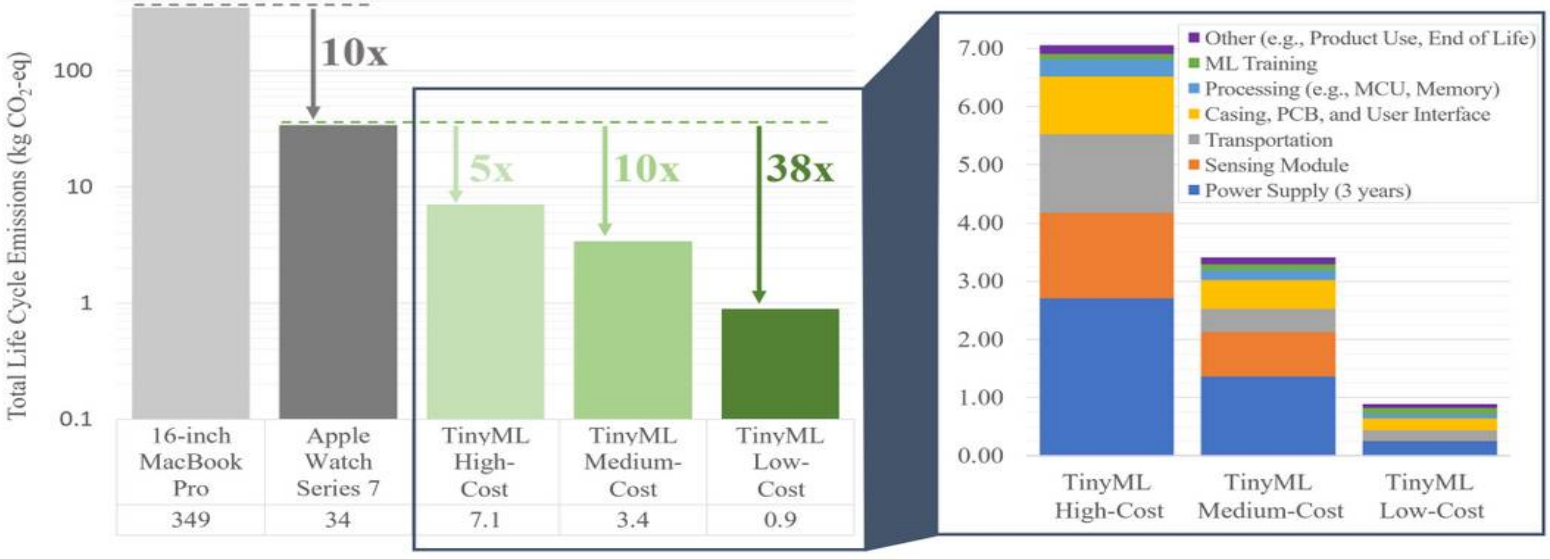


Figure 4. A breakdown of different TinyML system footprints highlights that the footprint is largely attributable to the embodied footprint of the power supply, onboard sensors, and transportation. Note that actuator and connectivity blocks from Pirson and Bol [32] are encapsulated in "Other" and "Processing", respectively, while "Product Use" captures the operational footprint. The carbon footprint of Apple's Series 7 Watch [22] and 16-inch MacBook Pro [21] are also provided for reference. For more details and to compute the footprint of your own TinyML system, see https://github.com/harvard-edge/TinyML-Footprint.

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Total lifecycle results for CO2 emissions

TinyML devices estimated to produce

1 .. 10 kg CO2-equivalent over lifetime
(depending on size of system)

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Balance footprint vs handprint

Figure 5. If all 250 billion MCUs were TinyML systems with three-year lifespans, their worst-case footprint would be 1765 million metric tons of CO_2. If these systems enabled a 20% emissions reduction for the residential sector and only a 0.6% reduction for all other sectors (Figure 2), the total footprint would be net-zero. Anything larger (e.g., 20%) results in more carbon savings from TinyML than emissions.

Source: Prakash, Shvetank, et al. "Is TinyML Sustainable? Assessing the Environmental Impacts of Machine Learning on Microcontrollers." arXiv preprint arXiv:2301.11899 (2023).

Discussion / Limitations of study

Lack of reliable LCA data for digital devices

Jevons' paradox

In economics, the Jevons paradox occurs when technological progress or government policy increases the efficiency with which a resource is used (reducing the amount necessary for any one use), but the falling cost of use induces increases in demand enough that resource use is increased, rather than reduced. [wikipedia]

Comparison is made beween

TinyML intervention vs *No intervention*

ignoring possibility of other interventions

Source: Guillaume, Joseph HA, et al. "Giving legs to handprint thinking: Foundations for evaluating the good we do." Earth's Future 8.6 (2020): e2019EF001422.

Discussion / Limitations of study

Assumption of best use case (*the Elephant* …)

Assumption of best practice (e.g. on recycling, waste handling)

Source: Guillaume, Joseph HA, et al. "Giving legs to handprint thinking: Foundations for evaluating the good we do." Earth's Future 8.6 (2020): e2019EF001422.

Discussion / Other limitations

Discuss!

Source: Guillaume, Joseph HA, et al. "Giving legs to handprint thinking: Foundations for evaluating the good we do." Earth's Future 8.6 (2020): e2019EF001422.

Discussion / Other limitations

Footprint is de-facto – Handprint is mere potential

”if it were used in this way …”

Realization of benefits is not achieved by mere deployment of TinyML – in fact, it might only be the first initial step, e.g. in building sector

Optimal/correct usage of technology is assumed. This is rarely the case.

Source: Guillaume, Joseph HA, et al. "Giving legs to handprint thinking: Foundations for evaluating the good we do." Earth's Future 8.6 (2020): e2019EF001422.

Discussion / Other limitations

While TinyML devices have some degree of autonomy, they still depend on a lot of infrastructure that is not taken into account.

Examples given actually mention the use of EdgeImpulse.

Generally, models will not be developed and trained on tiny devices, but in conventional datacenters.

Discussion / Other limitations

Personal view, Sebastian:

While the paper makes the important attempt to quantify the somewhat diffuse hope that IoT & TinyML might be beneficial,
the work remains too
tech-centric and likely overly optimistic (?).

Where to, from here?

1. “Greener” devices? IoT without batteries? Energy harvesting?

2. Frugal computing – “Run when there’s sun” ?

3. Or a disappearance of IT (where it is not needed)?

4. or can we _____ ?

All sources of this talk:

https://github.com/ITU-DASYALab/IT-sustainability/blob/master/IT-sustainability-sources.md

grateful for input!

sebastian@itu.dk