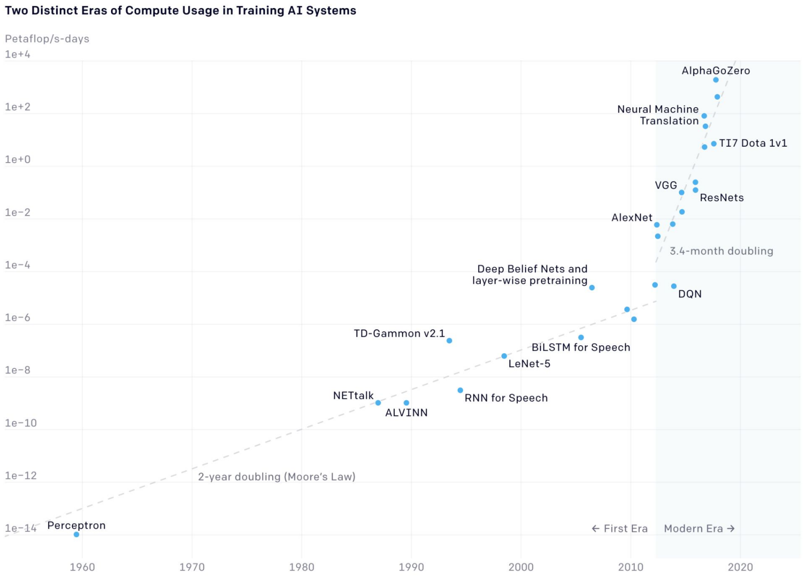
SciTinyML: Scientific Use of Machine Learning on Low-Power Devices

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

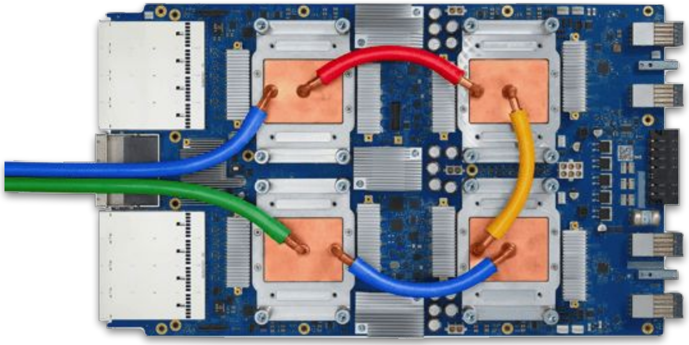
Two Eras of Computing

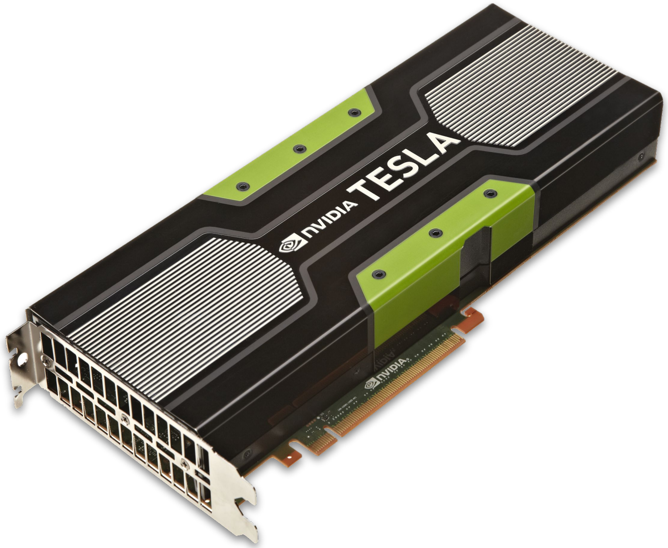
“... **since 2012 the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time** (by comparison, Moore’s Law had an 18-month doubling period). Since 2012, this metric has **grown by more than 300,000x (an 18-month [Moore’s Law] doubling period would yield only a 12x increase)**. Improvements in compute have been a key component of AI progress, so as long as this trend continues, it’s worth preparing for the implications of systems far outside today’s capabilities.”



Source: https://blog.openai.com/ai-and-compute/

TPUs/GPUs

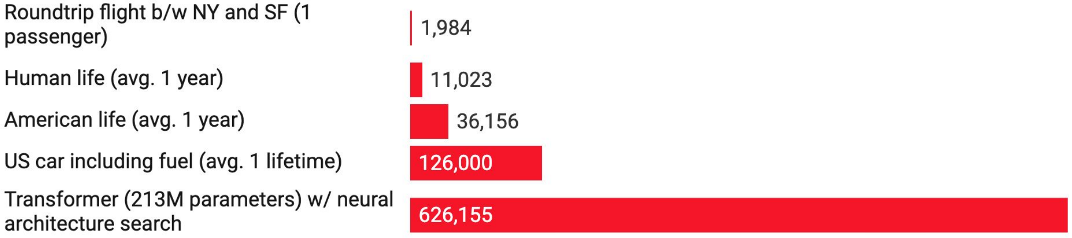




Impact on Climate

Common carbon footprint benchmarks

in lbs of CO2 equivalent



Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

arXiv:1906.02243v1 [cs.CL] 5 Jun 2019

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware. In this paper we bring this issue to the attention of NLP researchers by quantifying the approximate financial and environmental costs of training a variety of recently successful neural network models for NLP. Based on these findings, we propose actionable recommendations to reduce costs and improve equity in NLP research and practice.

1 Introduction

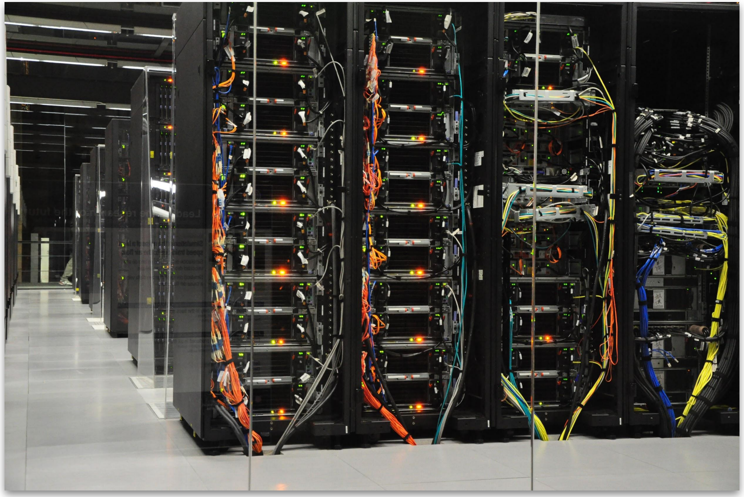
Advances in techniques and hardware for training deep neural networks have recently enabled impressive accuracy improvements across many fundamental NLP tasks (Bahdanau et al., 2015; Luong et al., 2015; Dozat and Manning, 2017; Vaswani et al., 2017), with the most computationally-hungry models obtaining the highest scores (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; So et al., 2019). As a result, training a state-of-the-art model now requires substantial computational resources which demand considerable energy, along with the associated financial and environmental costs. Research and development of new models multiplies these costs by thousands of times by requiring retraining to experiment with model architectures and hyperparameters. Whereas a decade ago most NLP models could be trained and developed on a commodity laptop or server, many now require multiple instances of specialized hardware such as GPUs or TPUs, therefore limiting access to these highly accurate models on the basis of finances.

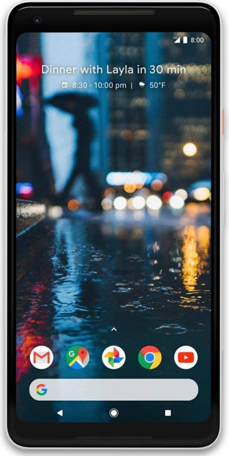
Even when these expensive computational resources are available, model training also incurs a substantial cost to the environment due to the energy required to power this hardware for weeks or months at a time. Though some of this energy may come from renewable or carbon credit-offset resources, the high energy demands of these models are still a concern since (1) energy is not currently derived from carbon-neural sources in many locations, and (2) when renewable energy is available, it is still limited to the equipment we have to produce and store it, and energy spent training a neural network might better be allocated to heating a family's home. It is estimated that we must cut carbon emissions by half over the next decade to deter escalating rates of natural disaster, and based on the estimated CO_2 emissions listed in Table 1,

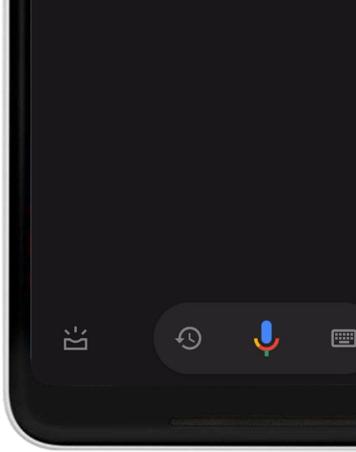
Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

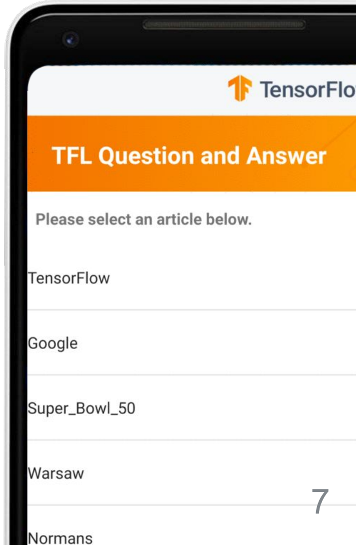
Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

[1]Sources: (1) Air travel and per-capita consumption: https://bit.ly/2Hw0xWc; (2) car lifetime: https://bit.ly/2Qbr0w1.

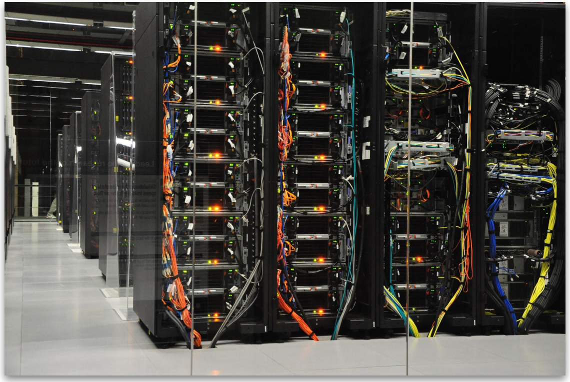


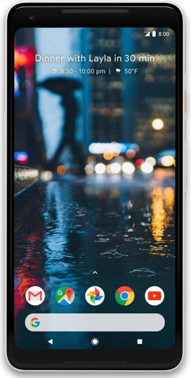

8:00
Dinner with Layla in 30 min
8:30 - 10:00 pm | 50°F




TensorFlo
TFL Question and Answer
Please select an article below.
TensorFlow
Google
Super_Bowl_50
Warsaw
Normans

https://plantvillage.psu.edu/





Google Assistant

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

Endpoint Devices

Bandwidth
Reliability
Latency
Privacy
Energy



ElephantEdge

Building The World's Most Advanced **Wildlife Tracker**.

Source: https://wildlabs.net/resources/competition/challenge-elephantedge

Dr. Iain Douglas-Hamilton

ElephantEdge

Risk Monitoring

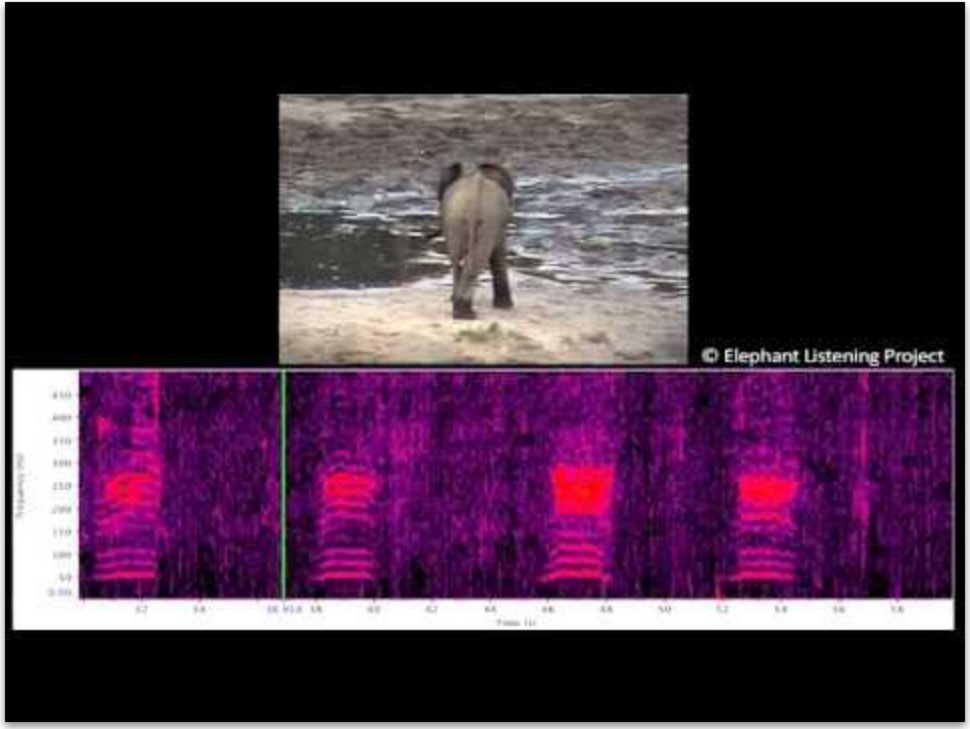
"Know when an elephant is moving into a high-risk area and send real-time notifications to park rangers."

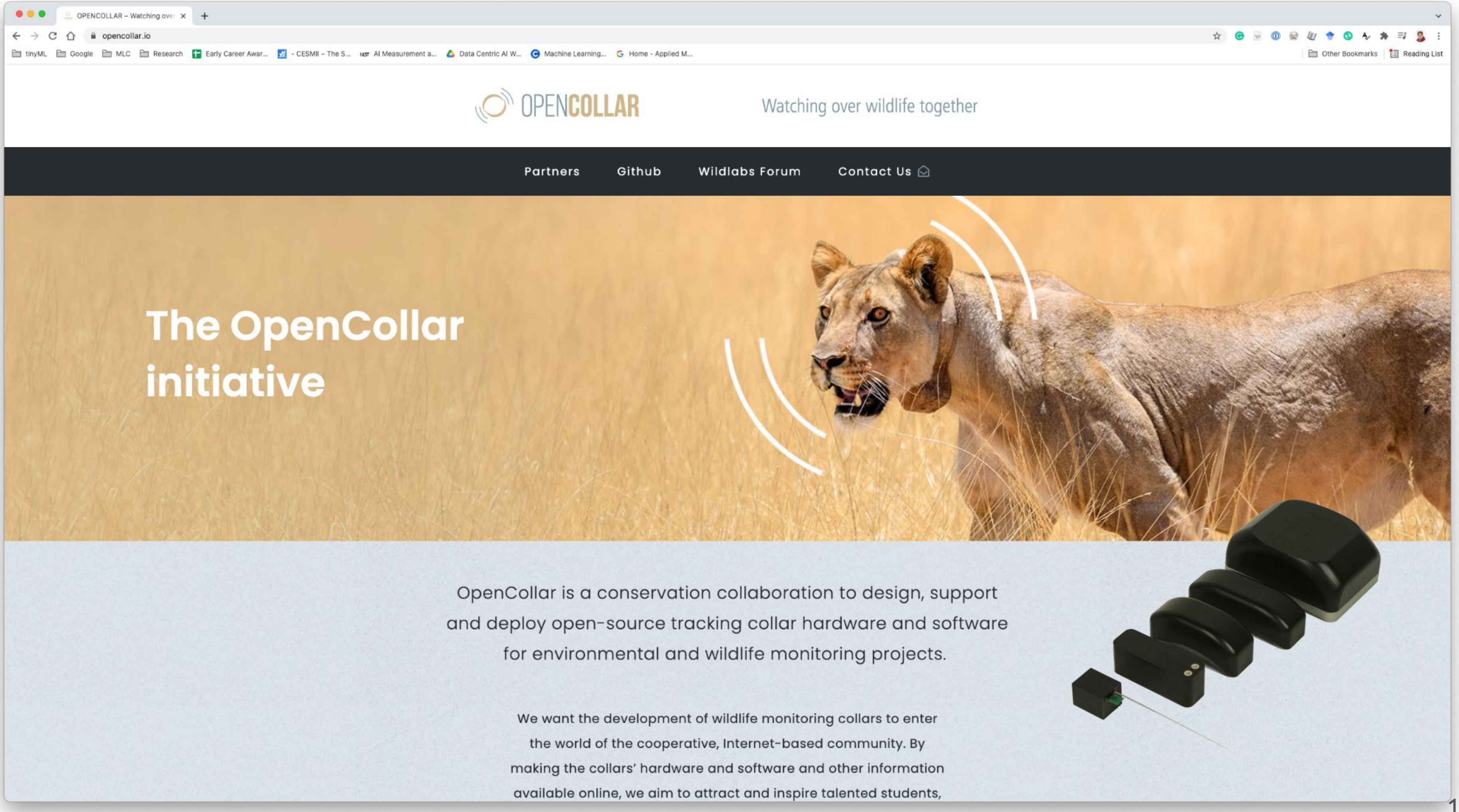
Conflict Monitoring

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

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.”


© Elephant Listening Project


OPENCOLLAR – Watching over
opencollar.io
OPENCOLLAR
Watching over wildlife together
Partners
Github
Wildlabs Forum
Contact Us
The OpenCollar initiative
OpenCollar is a conservation collaboration to design, support and deploy open-source tracking collar hardware and software for environmental and wildlife monitoring projects.
We want the development of wildlife monitoring collars to enter the world of the cooperative, Internet-based community. By making the collars' hardware and software and other information available online, we aim to attract and inspire talented students,

Read the latest updates on coronavirus from Harvard University. For SEAS specific-updates, please visit SEAS & FAS Division of Science: Coronavirus FAQs

Above
Female sperm whale (Image courtesy of Amanda Cotton)

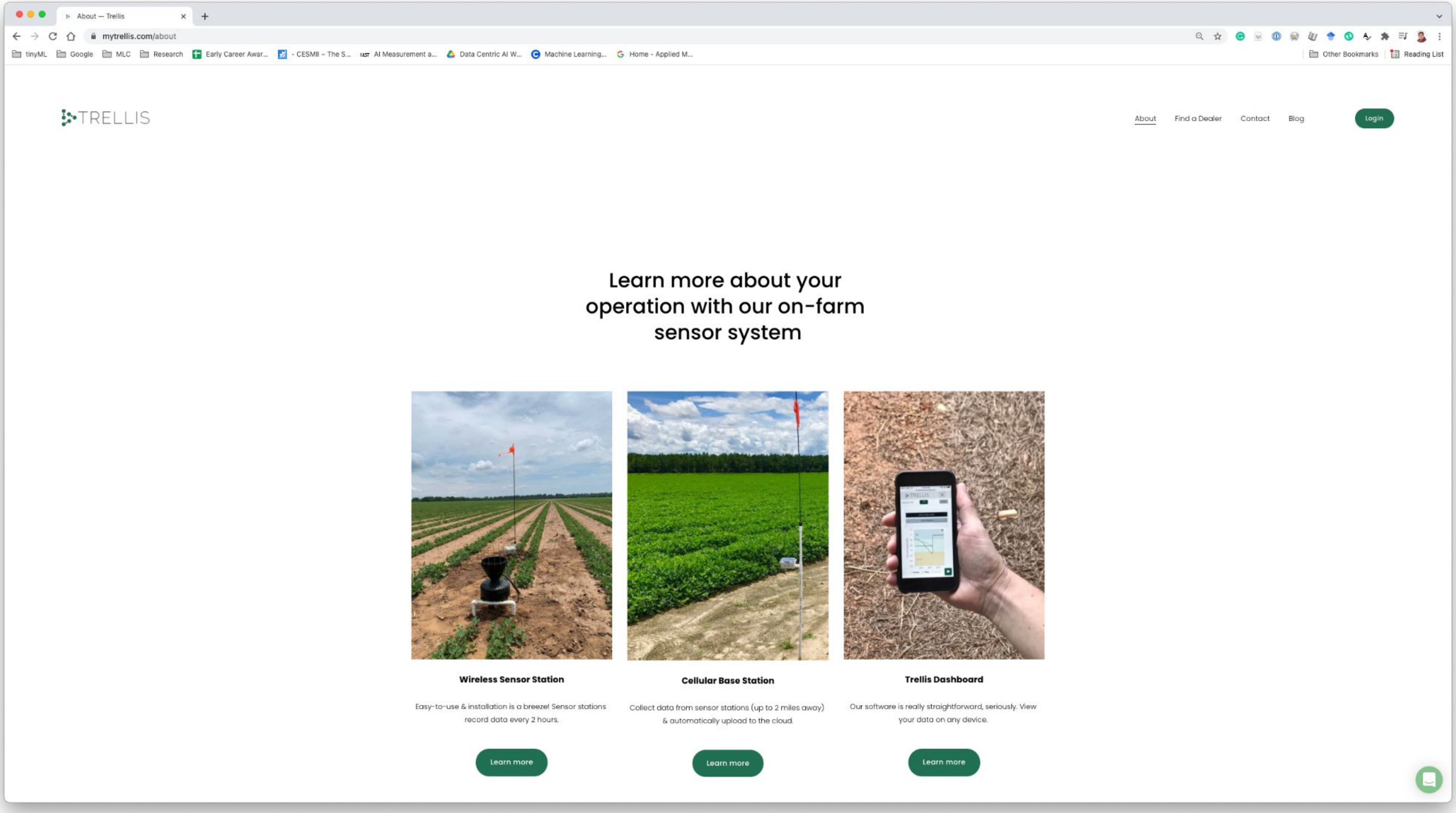
Talking with whales

Project aims to translate sperm whale calls

By Leah Burrows | Press contact
April 22, 2021

This week, a team of scientists in partnership with the Government of Dominica and the National Geographic Society, officially launched an ambitious, interdisciplinary research initiative to listen to, contextualize, and translate the communication of sperm whales.

Project CETI (Cetacean Translation Initiative) will bring together leading cryptographers,


About — Trellis
mytrellis.com/about
tinyML
Google
MLC
Research
Early Career Awar...
- CESMII – The S...
AI Measurement a...
Data Centric AI W...
Machine Learning...
Home - Applied M...
Other Bookmarks
Reading List
TRELLIS
About
Find a Dealer
Contact
Blog
Login
Learn more about your operation with our on-farm sensor system
Wireless Sensor Station
Easy-to-use & installation is a breeze! Sensor stations record data every 2 hours.
Learn more
Cellular Base Station
Collect data from sensor stations (up to 2 miles away) & automatically upload to the cloud.
Learn more
Trellis Dashboard
Our software is really straightforward, seriously. View your data on any device.
Learn more

No Good Data Left Behind

5 Quintillion

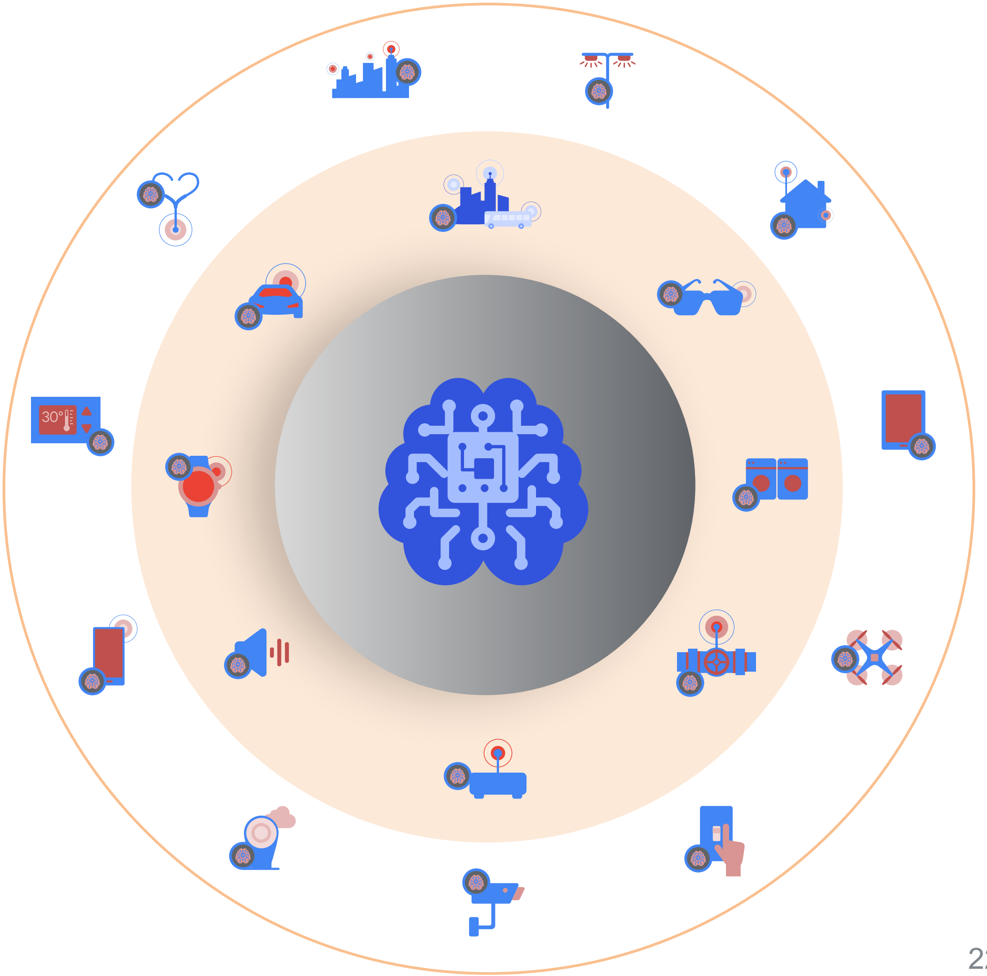
bytes of data produced every day by IoT

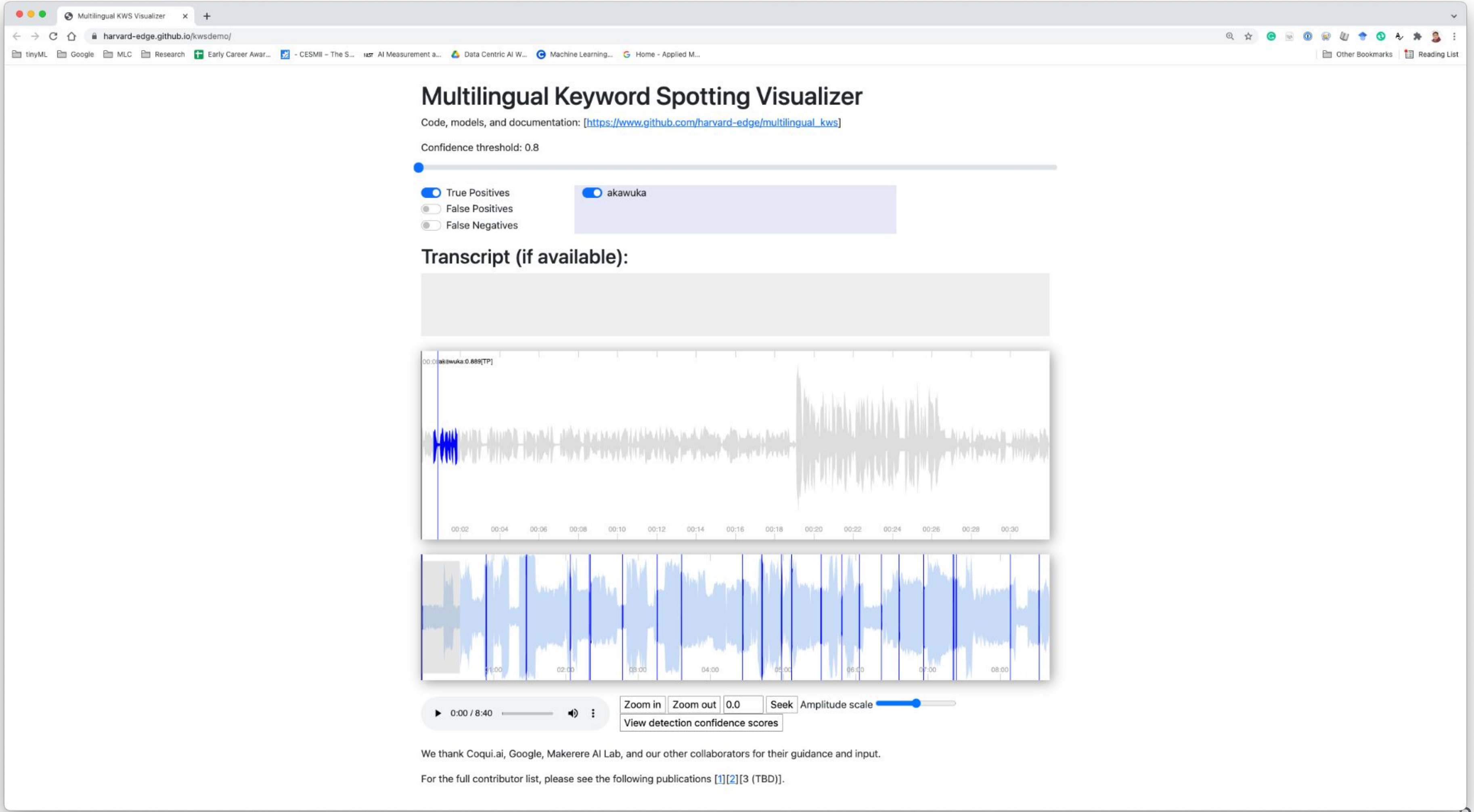
<1%

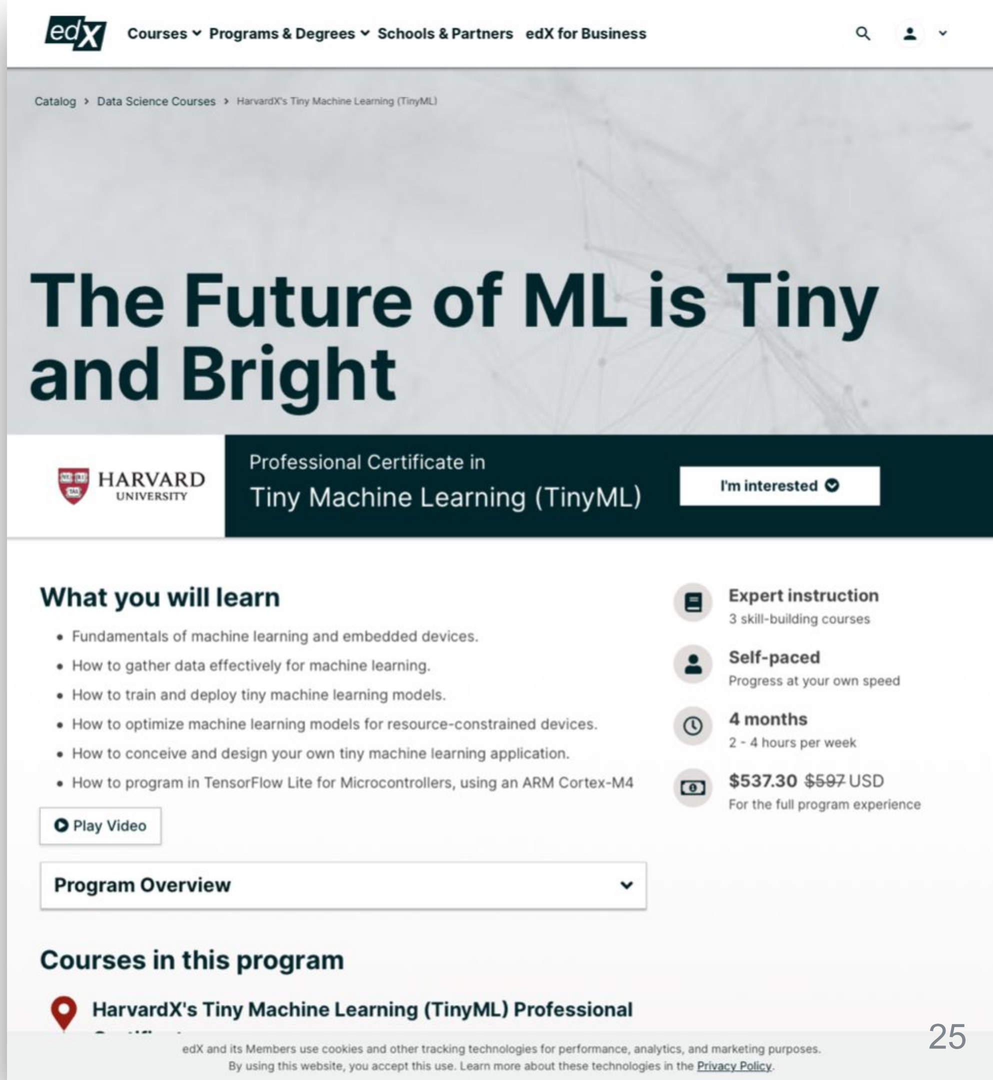
of unstructured data is analyzed or used at all

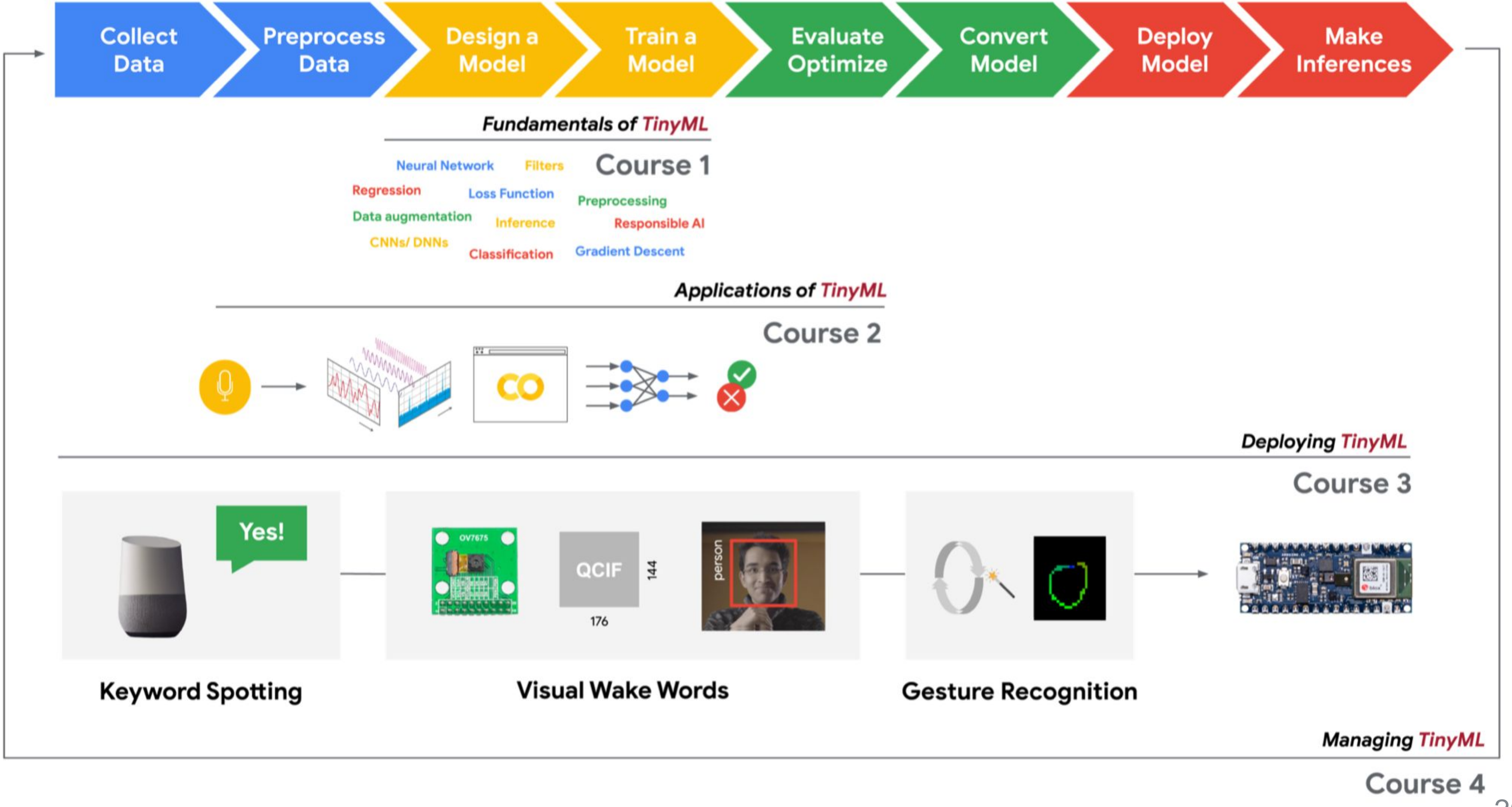
Source: Harvard Business Review, What's Your Data Strategy?, April 18, 2017
Cisco, Internet of Things (IoT) Data Continues to Explode Exponentially. Who Is Using That Data and How?, Feb 5, 2018

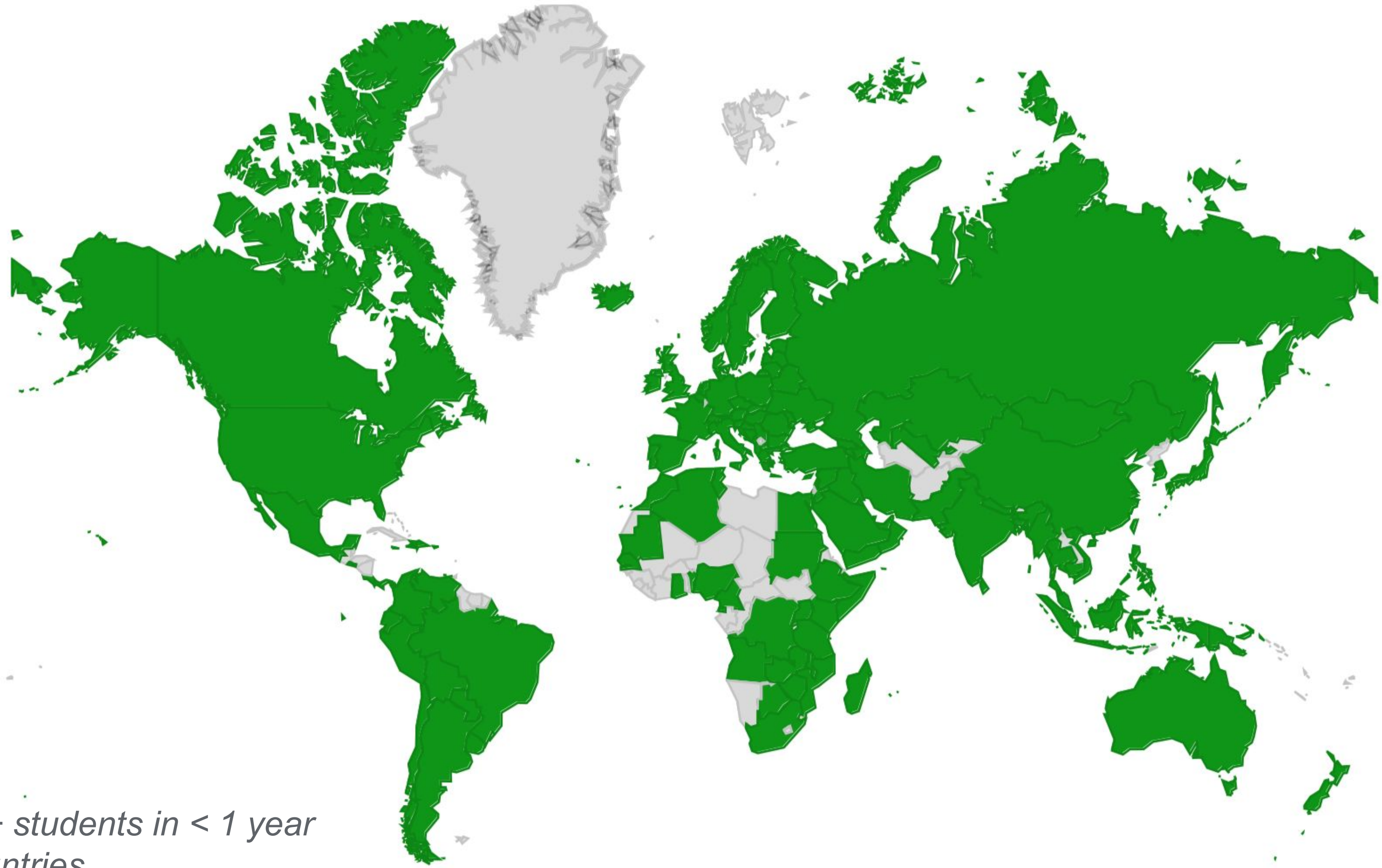
Massive tinyML opportunities in all verticals where machine intelligence meets physical world of billions of sensors



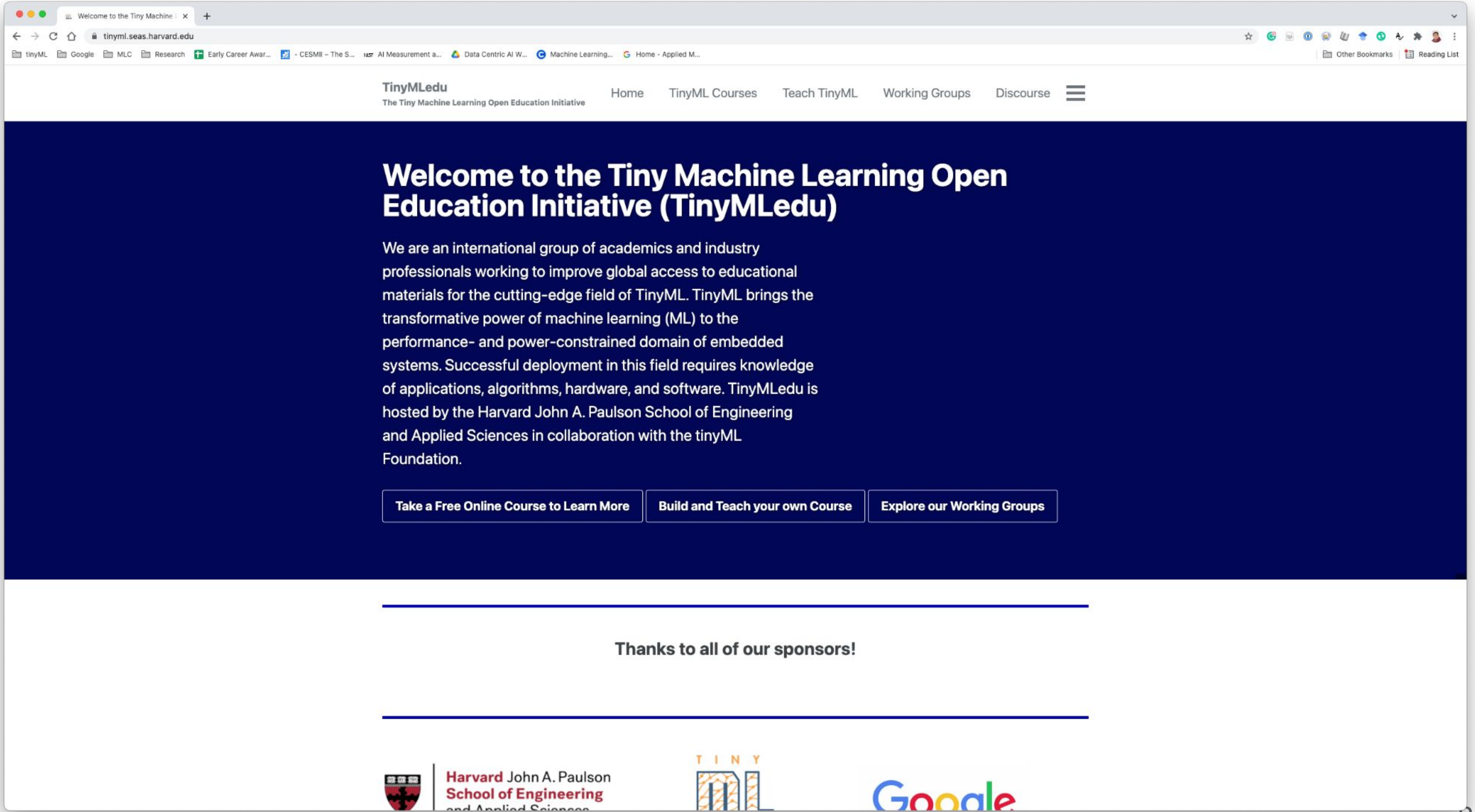

Multilingual KWS Visualizer
harvard-edge.github.io/kwsdemo/
Multilingual Keyword Spotting Visualizer
Code, models, and documentation: [https://www.github.com/harvard-edge/multilingual_kws]
Confidence threshold: 0.8
True Positives
False Positives
False Negatives
akawuka
Transcript (if available):
akawuka:0.889[TP]
0:00 / 8:40
Zoom in
Zoom out
0.0
Seek
Amplitude scale
View detection confidence scores
We thank Coqui.ai, Google, Makerere AI Lab, and our other collaborators for their guidance and input.
For the full contributor list, please see the following publications [1][2][3 (TBD)].


edX
Courses Programs & Degrees Schools & Partners edX for Business
Catalog > Data Science Courses > HarvardX's Tiny Machine Learning (TinyML)
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 and embedded devices.
• How to gather data effectively for machine learning.
• How to train and deploy tiny machine learning models.
• How to optimize machine learning models for resource-constrained devices.
• How to conceive and design your own tiny machine learning application.
• How to program in TensorFlow Lite for Microcontrollers, using an ARM Cortex-M4
Play Video
Program Overview
Expert instruction
3 skill-building courses
Self-paced
Progress at your own speed
4 months
2 - 4 hours per week
$537.30 ~~$597~~ USD
For the full program experience
Courses in this program
HarvardX's Tiny Machine Learning (TinyML) Professional
edX and its Members use cookies and other tracking technologies for performance, analytics, and marketing purposes.
By using this website, you accept this use. Learn more about these technologies in the Privacy Policy.


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
Fundamentals of TinyML
Course 1
Neural Network
Filters
Regression
Loss Function
Preprocessing
Data augmentation
Inference
Responsible AI
CNNs/ DNNs
Classification
Gradient Descent
Applications of TinyML
Course 2
CO
Deploying TinyML
Course 3
Yes!
OV7675
QCIF
144
176
person
Keyword Spotting
Visual Wake Words
Gesture Recognition
Managing TinyML
Course 4



50,000+ students in < 1 year
177 countries


Welcome to the Tiny Machine
tinyml.seas.harvard.edu
tinyML
Google
MLC
Research
Early Career Awar...
- CESMII – The S...
AI Measurement a...
Data Centric AI W...
Machine Learning...
Home - Applied M...
Other Bookmarks
Reading List
TinyMLedu
The Tiny Machine Learning Open Education Initiative
Home
TinyML Courses
Teach TinyML
Working Groups
Discourse
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

Explore our Working Groups -

tinyml.seas.harvard.edu/workingGroups

tinyML Google MLC Research Early Career Awar... - CESMII – The S... AI Measurement a... Data Centric AI W... Machine Learning... Home - Applied M... Other Bookmarks Reading List

Explore our Working Groups

Widening access to applied machine learning by establishing best practices in education.

If you want to be more involved with our effort to help improve access to TinyML educational materials and hardware resources worldwide reach out to us at edu@tinyML.org!

TinyML4D

The TinyML4D working group is building a network of academic institutions, based in Developing Countries, interested in expanding access to Applied Machine Learning by establishing best practices in education. We aim to ultimately develop a community of researchers and practitioners focused on both improving access to TinyML education and enabling innovative solutions for the unique challenges faced by Developing Countries. TinyML4D is co-hosted by the Abdus Salam International Centre for Theoretical Physics (ICTP).

Learn More

TinyML4K12

Expanding TinyML education into primary and secondary schools (K-12) requires the development of an end-to-end pipeline that is appropriate for school-aged children. We are working with education and industry partners to combine computer science education software and the physical computing ecosystem to enable an easy learning experience for creating, deploying, and using TinyML models. This pipeline will enable the creation of additional materials that can be used across the globe for students of all ages.

Learn More

TinyMLTranslations

Our mission is to enable all learners, regardless of their preferred language of learning, to be able to access and learn TinyML. As such, we work to translate and support material and course development in languages other than English.

Learn More

Widening Access to Applied ML with TinyML

- Tiny machine learning
- Embedded systems are the future of machine learning
- Focus on widening the reach of TinyML by democratization of ML data and education

arXiv:2106.04008v2 [cs.LG] 9 Jun 2021

Widening Access to Applied Machine Learning with TinyML

Vijay Janapa Reddi[*] Brian Plancher[*] Susan Kennedy[*] Laurence Moroney[†]
Pete Warden[†] Anant Agarwal[*‡] Colby Banbury[*] Massimo Banzi[§] Matthew Bennett[*]
Benjamin Brown[*] Sharad Chitlangia[¶] Radhika Ghosal[*] Sarah Grafman[*] Rupert Jaeger[||]
Srivatsan Krishnan[*] Maximilian Lam[*] Daniel Leiker[||] Cara Mann[*] Mark Mazumder[*]
Dominic Pajak[§] Dhilan Ramaprasad[*] J. Evan Smith[*] Matthew Stewart[*] Dustin Tingley[*]

[*]Harvard University
[†]Google

ABSTRACT

Broadening access to both computational and educational resources is critical to diffusing machine-learning (ML) innovation. However, today, most ML resources and experts are siloed in a few countries and organizations. In this paper, we describe our pedagogical approach to increasing access to applied ML through a massive open online course (MOOC) on Tiny Machine Learning (TinyML). We suggest that TinyML, ML on resource-constrained embedded devices, is an attractive means to widen access because TinyML both leverages low-cost and globally accessible hardware, and encourages the development of complete, self-contained applications, from data collection to deployment. To this end, a collaboration between academia (Harvard University) and industry (Google) produced a four-part MOOC that provides application-oriented instruction on how to develop solutions using TinyML. The series is openly available on the edX MOOC platform, has no prerequisites beyond basic programming, and is designed for learners from a global variety of backgrounds. It introduces pupils to real-world applications, ML algorithms, data-set engineering, and the ethical considerations of these technologies via hands-on programming and deployment of TinyML applications in both the cloud and their own microcontrollers. To facilitate continued learning, community building, and collaboration beyond the courses, we launched a standalone website, a forum, a chat, and an optional course-project competition. We also released the course materials publicly, hoping they will inspire the next generation of ML practitioners and educators and further broaden access to cutting-edge ML technologies.

Figure 1: We designed a new applied-ML course motivated by real-world applications, covering not only the software (algorithms) and hardware (embedded systems) but also the product life cycle and responsible AI considerations needed to deploy these applications. To make it globally accessible and scalable, we focused on the emerging TinyML domain and released the course as a MOOC on edX.

1 INTRODUCTION

The past two decades have seen machine learning (ML) progress dramatically from a purely academic discipline to a widespread commercial technology that serves a range of sectors. ML allows developers to improve business processes and human productivity through data-driven automation. Given applied ML's ubiquity and success, its commercial use should only increase. Existing ML applications cover a wide spectrum that includes digital assistants [1, 2], autonomous vehicles [3, 4], robotics [5], health care [6], transportation [7, 8], security [9], and education [10, 11], with new application use cases continuously emerging every few days.

The proliferation of this technology and associated jobs have great potential to improve society and uncover new opportunities for technological innovation, societal prosperity, and individual growth. But it all rests on the assumption that everyone, globally, has unfettered access to ML technologies, which isn't the case.

Expanding access to applied ML faces three challenges. First is a shortage of ML educators at all levels [12, 13]. Second is insufficient resources, as training and running ML models often requires costly, high-performance hardware, especially as data sets continue to balloon. Third is a growing gap between industry and academia, as even the best academic institutions and research labs struggle to keep pace with industry's rapid progress. Addressing these critical issues requires innovative education and workforce training to prepare the next generation of applied-ML engineers.

This paper presents a pedagogical approach, developed as an academic and industry collaboration led by Harvard University and Google, to address these challenges and thereby increase global access to applied ML. The resulting course, TinyML on edX, focuses not only on teaching the topic by exploring real-world TinyML

[§]Arduino, [¶]BITS Pilani, work done as Harvard intern, [||]CreativeClass.ai, [*]edX, [‡]MIT

1