tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

TINY ML

www.tinyML.org

TINYMLPERF

Development of a Benchmark Suite for tinyML Systems

Csaba Kiraly, Digital Catapult

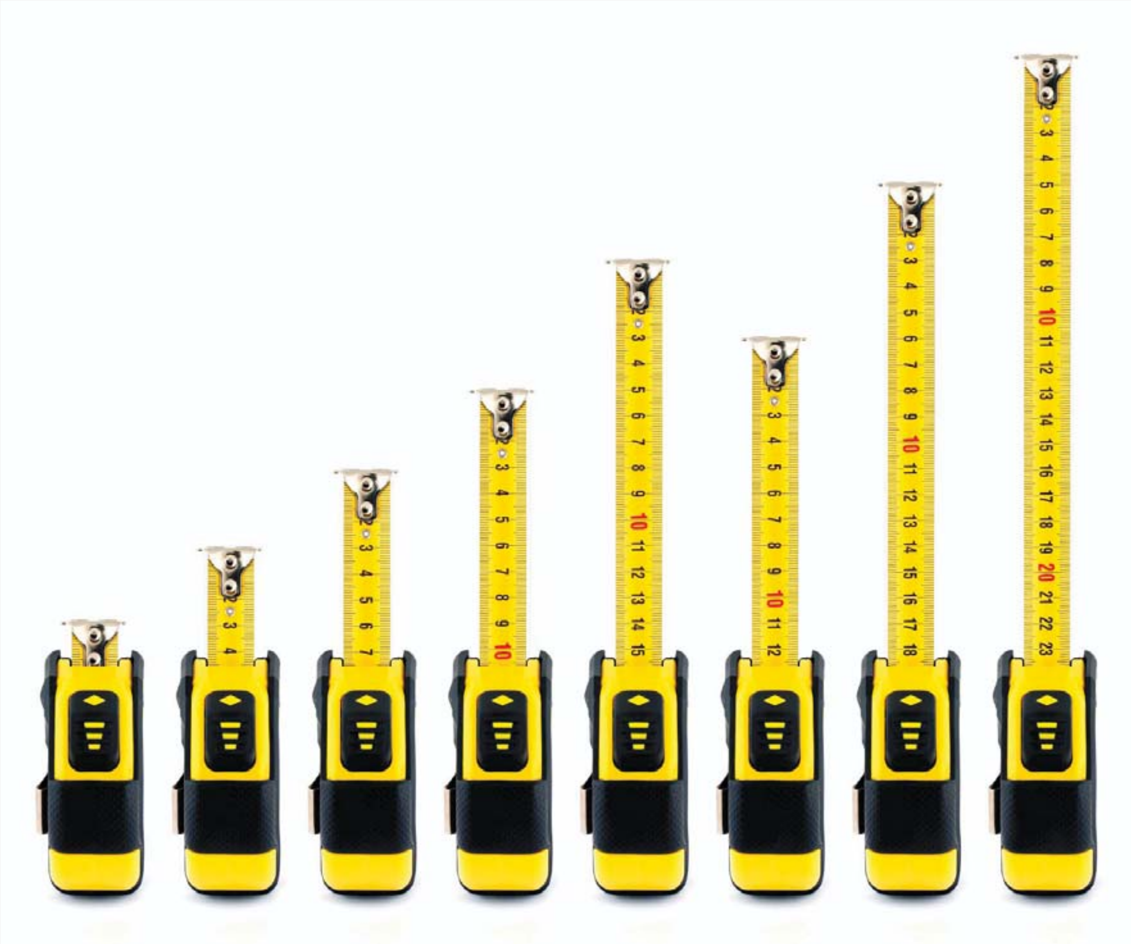
Colby Banbury, Harvard University

IMPORTANCE OF BENCHMARKING

GIVEN A TASK AT HAND HOW DO WE KNOW WHICH IS THE **RIGHT SYSTEM**?

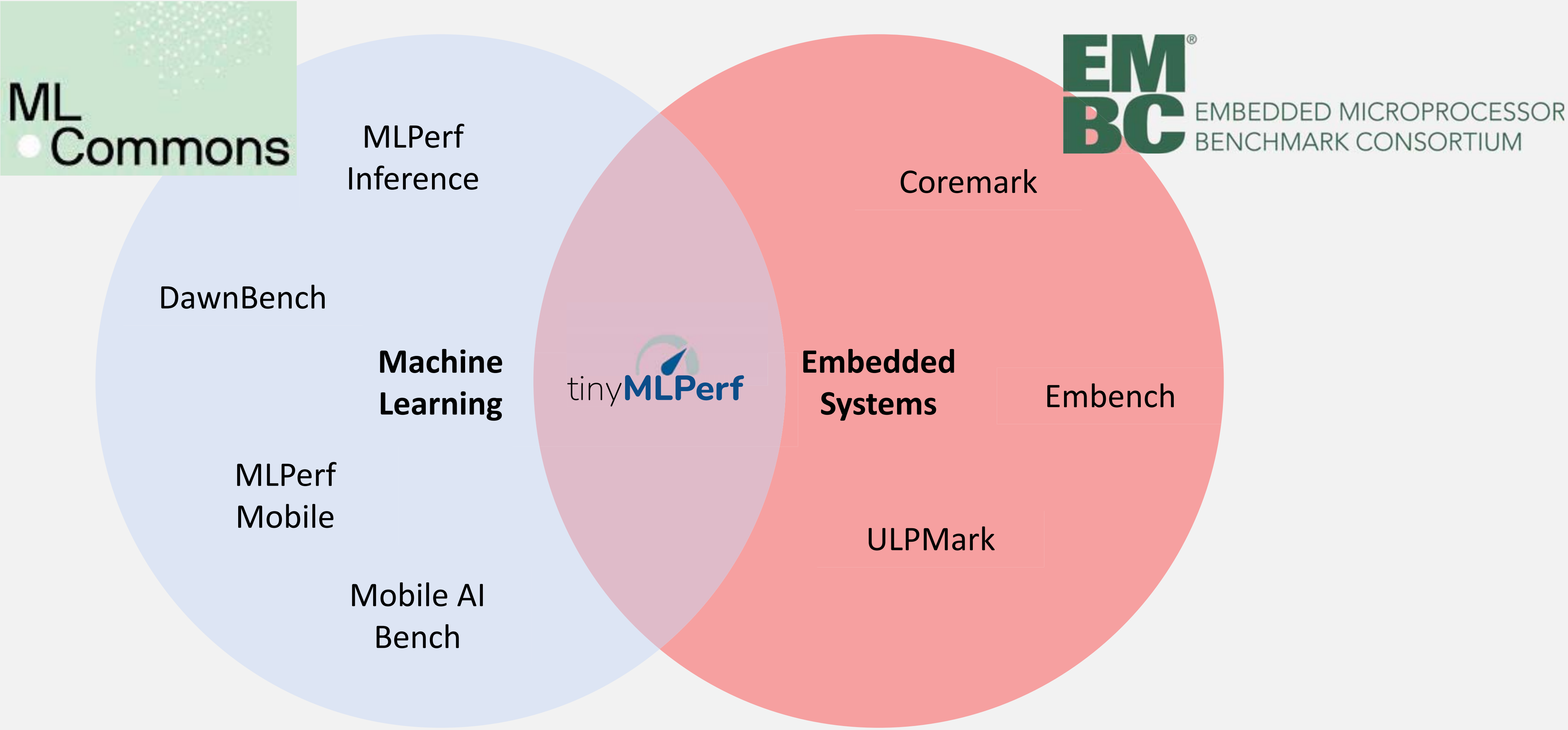
HOW DO WE **COMPARE** THE DIFFERENT SOLUTIONS?

BENCHMARKING IS ALL ABOUT …



- **Comparability** across hardware and software
- **Standardization** of use cases and workloads
- **Measuring** progress via rigorous methodology
- **Community** building and consensus generation

NEED FOR A TINYML SPECIFIC BENCHMARK



BENCHMARK DESIGN CHALLENGES

- Number and diversity of use cases for tinyML
  - Which use cases should we focus on?
- TinyML innovation is in HW, in SW, in tooling, in algorithms, etc.
  - How do we support fair comparison and innovation?
- Embedded system design is always a compromise
  - What metrics to choose to fairly evaluate systems?

BENCHMARK DESIGN CHALLENGES

- **Number and diversity of use cases for tinyML**
  - **Which use cases should we focus on?**
- TinyML innovation is in HW, in SW, in tooling, in algorithms, etc.
  - How do we support fair comparison and innovation?
- Embedded system design is always a compromise
  - What metrics to choose to fairly evaluate systems?

RELEVANT USE CASES

Task Category	Use Case	Model Type
Audio	Audio Wake Words Context Recognition Control Words Keyword Detection	DNN CNN RNN LSTM
Image	Visual Wake Words Object Detection Gesture Recognition Object Counting Text Recognition	DNN CNN SVM Decision Tree KNN Linear
Industry / Telemetry	Segmentation Anomaly Detection Forecasting Activity Detection	DNN Decision Tree SVM Linear

RELEVANT USE CASES & COMMUNITY

Task Category	Use Case	Model Type
Audio	**Audio Wake Words** Context Recognition Control Words Keyword Detection	DNN CNN RNN LSTM
Image	**Visual Wake Words** **Object Detection** Gesture Recognition Object Counting Text Recognition	DNN CNN SVM Decision Tree KNN Linear
Industry / Telemetry	Segmentation **Anomaly Detection** Forecasting Activity Detection	DNN Decision Tree SVM Linear

WG lead: Vijay Janapa Reddi + Colby Banbury @ Harvard

Benchmark harness: Peter Torelly @ EEMBC + Nat Jeffries @ Google

Jeremy Holleman @ Syntiant

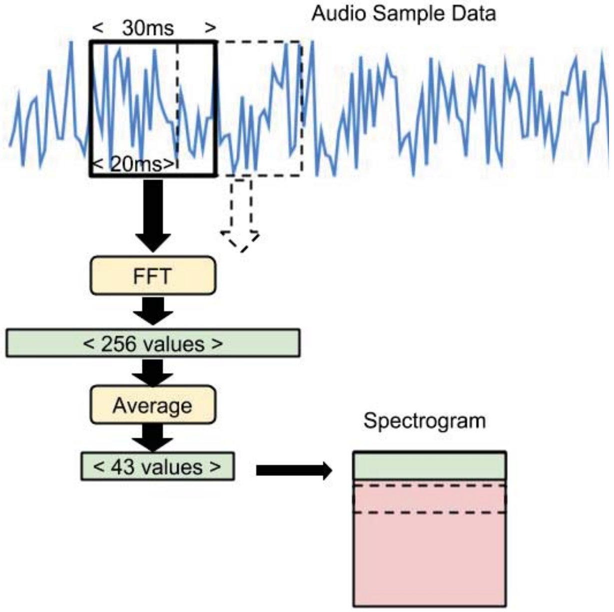
Nat Jeffries @ Google

Pietro Montino

Csaba Kiraly @ DC

SELECTED BENCHMARKS

Keyword Spotting



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

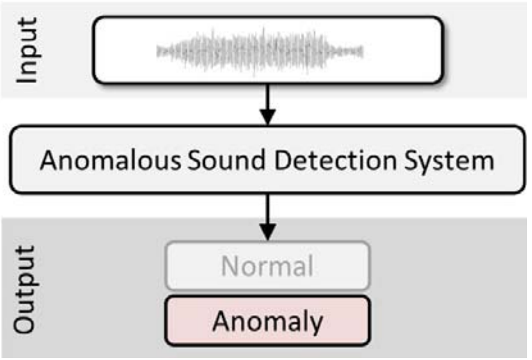
Visual Wake Words

(a) 'Person'

(b) 'Not-person'

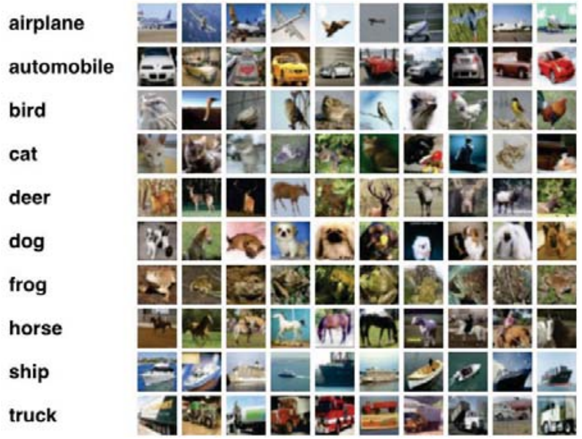
Chowdhery, Aakanksha, et al. "Visual wake words dataset." *arXiv preprint arXiv:1906.05721* (2019).

Anomaly Detection



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

Tiny Image Classification



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

BENCHMARK DESIGN CHALLENGES

- Number and diversity of use cases for tinyML
    - Which use cases should we focus on?

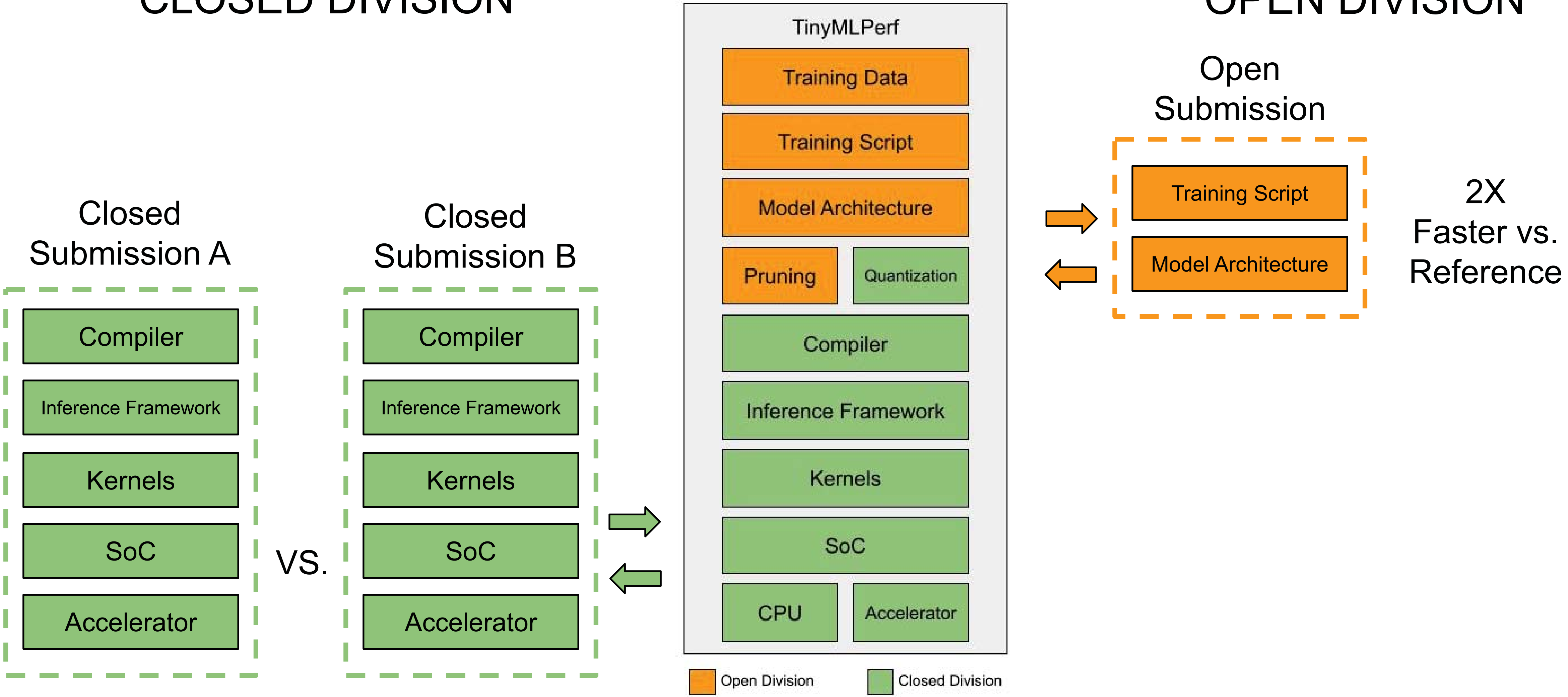
- **TinyML innovation is in HW, in SW, in tooling, in algorithms, etc.**
    - **How do we support fair comparison and innovation?**

- Embedded system design is always a compromise
    - What metrics to choose to fairly evaluate systems?

Modular Design

Direct Comparison
CLOSED DIVISION

Demonstrate Improvement
OPEN DIVISION

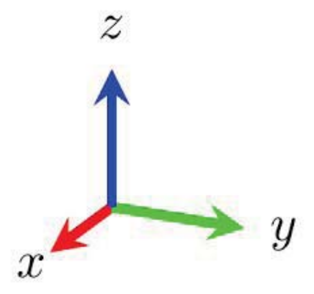


BENCHMARK DESIGN CHALLENGES

- Number and diversity of use cases for tinyML
    - Which use cases should we focus on?
- TinyML innovation is in HW, in SW, in tooling, in algorithms, etc.
    - How do we support fair comparison and innovation?
- **Embedded system design is always a compromise**
    - **What metrics to choose to fairly evaluate systems?**

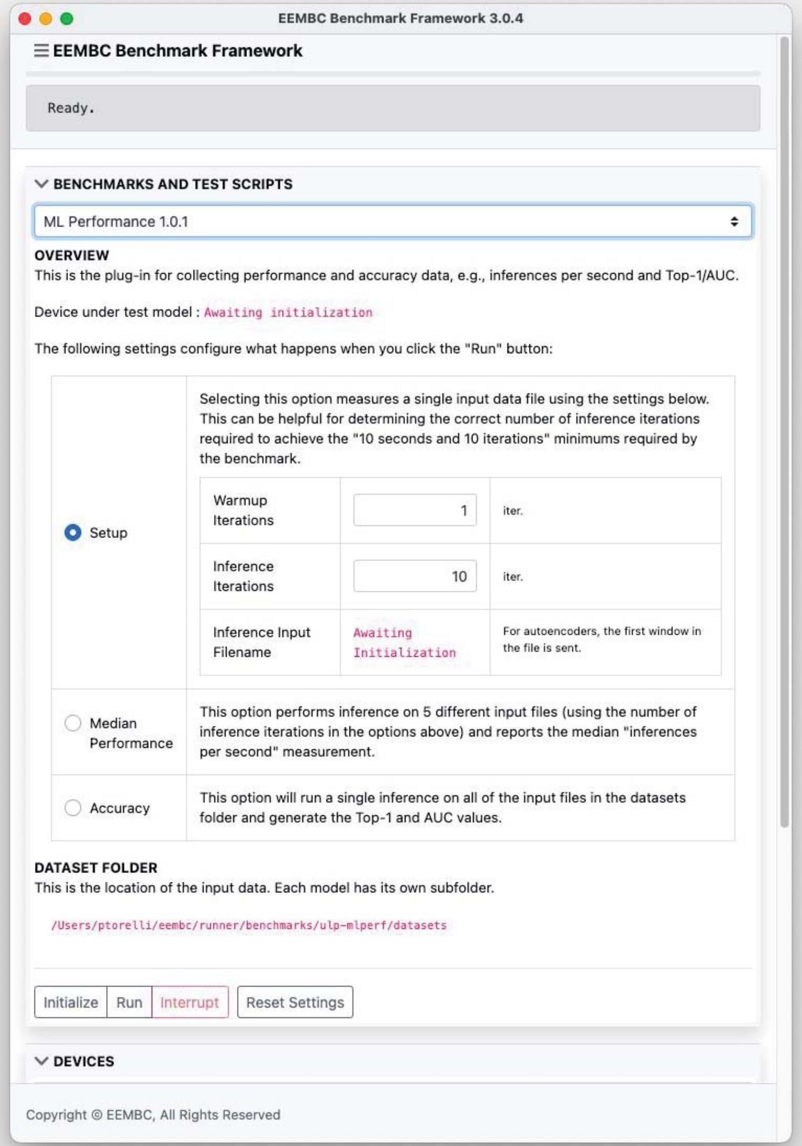
METRICS OF INTEREST

- What metrics to choose, how to define
  - latency: ms/inference, or ms/task?
  - accuracy: which metric?
  - energy: what should be included?



- (Some) measurement difficulties
  - pre/post processing
  - single inference vs. more complex processing
  - accurate power measurement
  - accurate timestamping

- EEMBC Benchmark Runner



MEASURING THOSE METRICS

Latency

Accurate timestamping
Evaluated on series of inferences
Data-dependant execution?

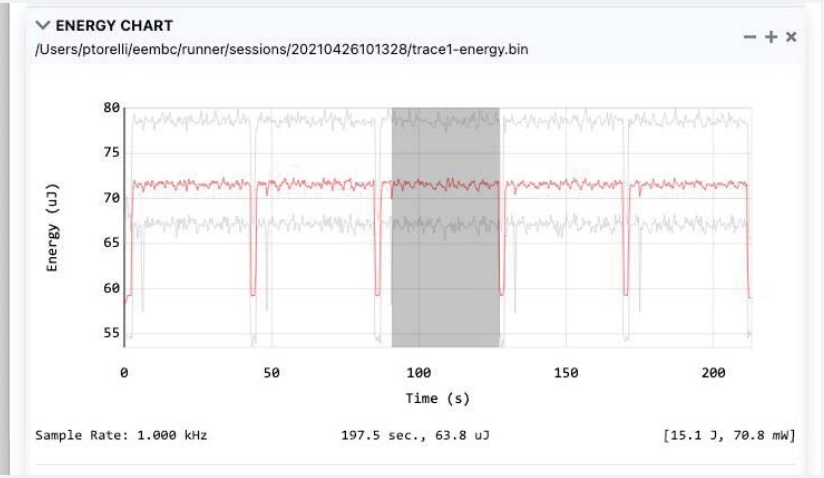
Accuracy

Top-1 accuracy & AUC
CLOSED: meet threshold
OPEN: part of the metrics

Energy

Power Monitor integration
No "cherry-picking"
Median result

```
Runtime requirements have been met.
Performance results for window 10:
  # Inferences :         1000
  Runtime      :       10.524 sec.
  Throughput   :       95.020 inf./sec.
Runtime requirements have been met.
------------------------------------------
Median throughput is 95.019 inf./sec.
------------------------------------------
```



ALL TOGETHER: THE FULL BENCHMARK FLOW

Workgroup

Submitter

Benchmark definition

- Problem definition
- Dataset selection (public domain)
- Model selection / training

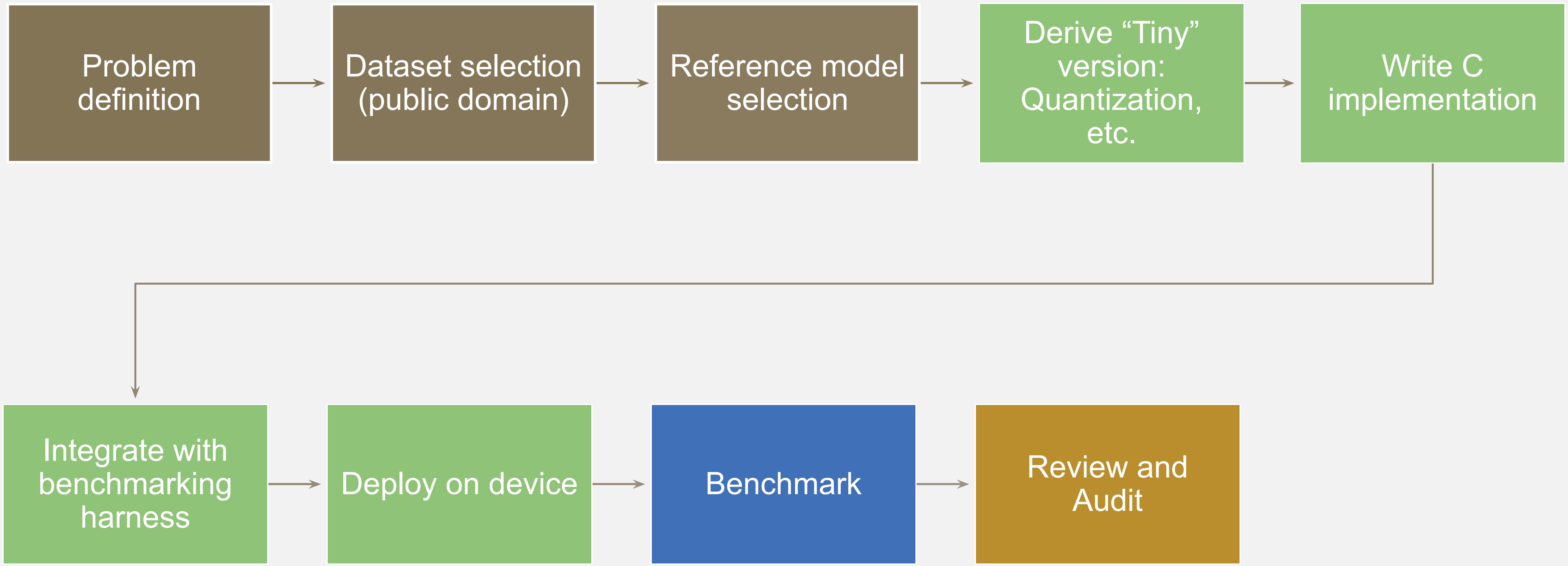
Reference implementation

- Derive "Tiny" version: Quantization, etc.
- Write C implementation
- Integrate with benchmarking harness
- Deploy on device
- Benchmark
  - ML performance
  - Power consumption
- Review and Audit

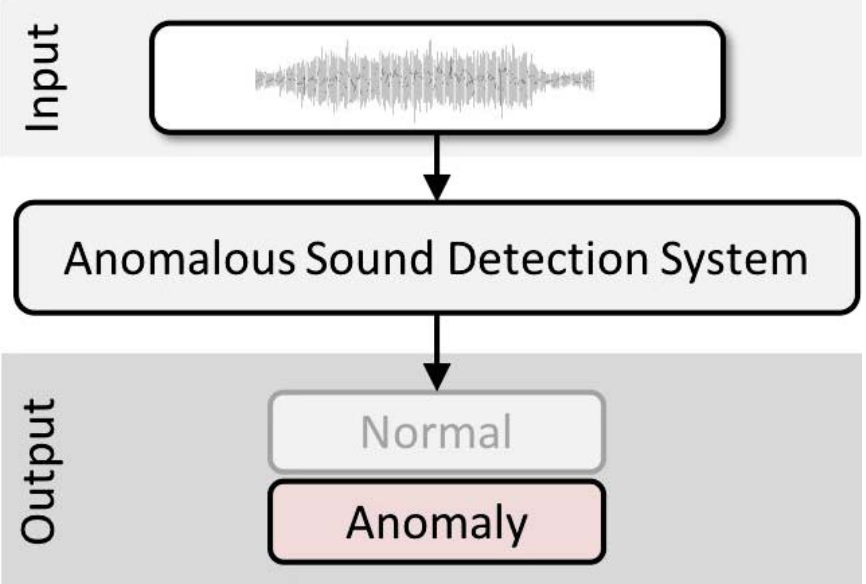
Open submission

Closed submission

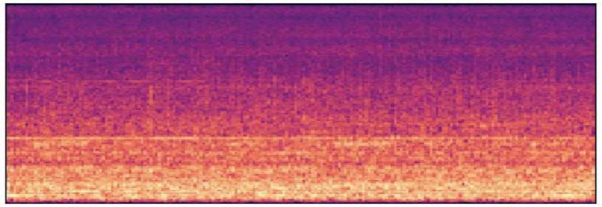
ALL TOGETHER: ANOMALY DETECTION EXAMPLE



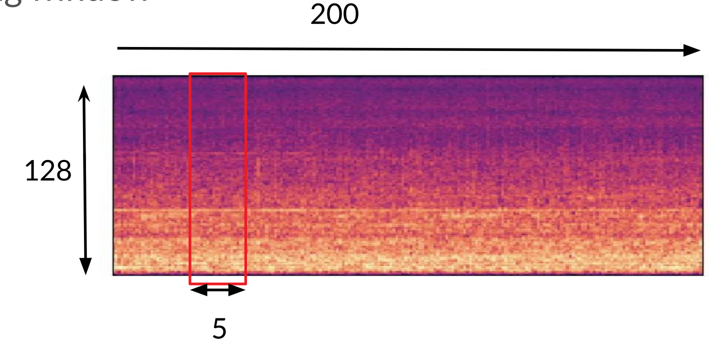
EXAMPLE: ANOMALY DETECTION



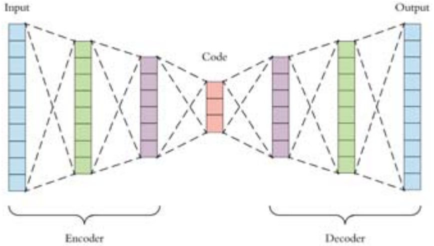
- Pre-processing: Spectrogram



- Sliding window



- Anomaly score on individual window
  AutoEncoder



- Post-processing:
  average score

REFERENCE IMPLEMENTATION

ST NUCLEO-L4R5ZI

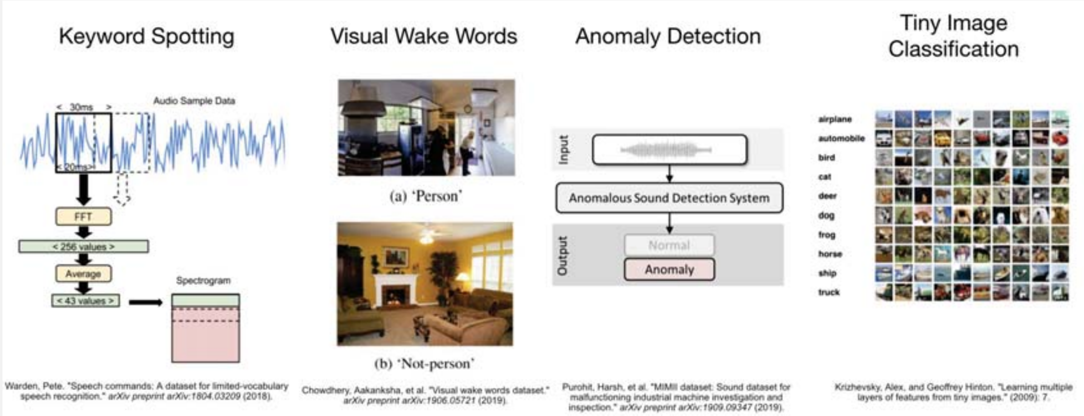
STM32 Nucleo + arm MBED Enabled + TensorFlow Lite

- Open platform
- Widely known, available, and affordable
- ARM Cortex-M4 with FPU
- 2 MB Flash
- 640KB SRAM

- Open source
- Portability of reference implementations

- Open source
- Full toolchain with everything we need
- Improving quantization support

TINYMLPERF: WHERE NEXT?



V0.1.0 is out
bootstrapping the process

Just the beginning …
Grow the community
Refine and define use cases
towards V0.2.0

Premier Sponsor

Intelligent Agent Neuton

Automated TinyML

Zero-code SaaS solution

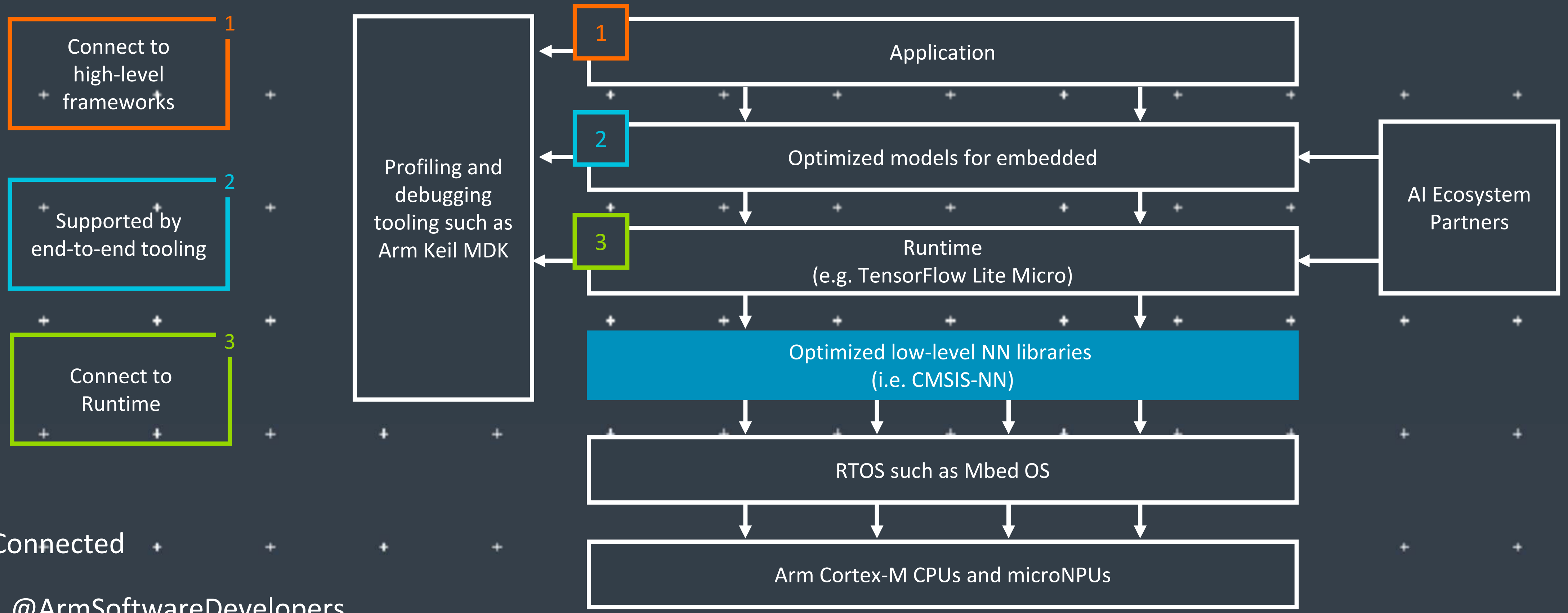
Create tiny models, ready for embedding, in just a few clicks!

Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

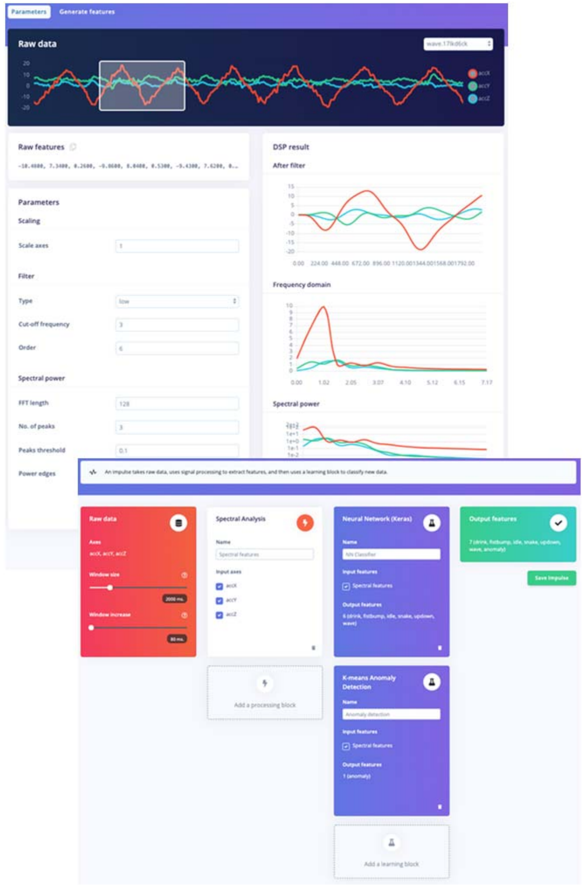
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



www.edgeimpulse.com

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

SYNTIANT

Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp

Platinum Sponsors

Part of your life. Part of tomorrow.

www.infineon.com

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai info@reality.ai @SensorAI Reality AI

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions

- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars "see with sound"

Reality AI Tools® software

- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

TINY ML EMEA
Virtual
JUNE 7-10,2021

Silver Sponsors

emza visual sense

GREENWAVES TECHNOLOGIES

HOTG

imagimob

Qeexo

seeed studio
The IoT Hardware Enabler

ST life.augmented

Copyright Notice

www.tinyML.org