# tinyML® EMEA

*Enabling Ultra-low Power Machine Learning at the Edge*

# tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

www.tinyML.org

# tinyML EMEA Technical Forum 2021
# June 7-10, 2021

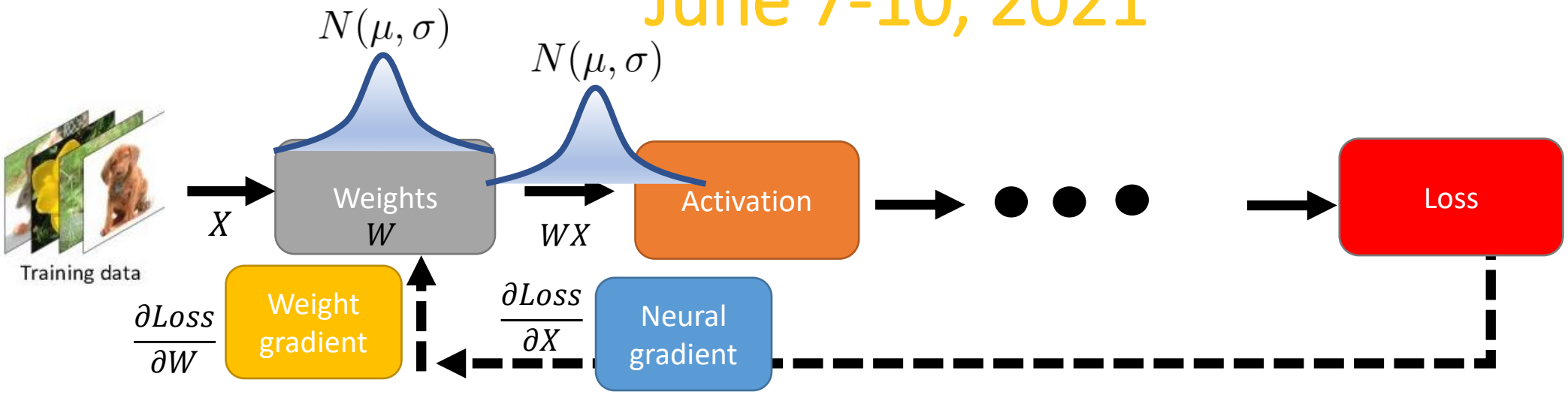
**Neural gradients are near-lognormal:**
**Improved quantized and sparse training**
**ICLR 2021**

Presented by: Brian Chmiel , Habana labs & Technion, Israel

habana
An Intel Company

TECHNION
Israel Institute
of Technology

June 10, 2021




- Most previous works focus on quantization or pruning of the weights and activations - Approximating them with Normal distribution.
- We find that neural gradients have very different statistics - approximately lognormal.
- We use Kolmogorov-Smirnov test to estimate the goodness of fit of neural gradients to lognormal distribution:

| Distribution | Model (Dataset) | | | | | |
|---|---|---|---|---|---|---|
| | BERT (CoLa) | BERT (MRPC) | ResNet18 (ImageNet) | MobileNetV2 (ImageNet) | VGG16 (ImageNet) | DenseNet121 (ImageNet) |
| Normal | 0.46±0.02 ($2 \cdot 10^{-4}$) | 0.39±0.04 ($5 \cdot 10^{-5}$) | 0.38±0.1 ($3 \cdot 10^{-6}$) | 0.22±0.09 ($5 \cdot 10^{-6}$) | 0.35±0.08 ($3 \cdot 10^{-6}$) | 0.33±0.1 ($5 \cdot 10^{-5}$) |
| Lognormal | **0.05±0.002** (0.28) | **0.04±0.002** (0.23) | **0.02±0.002** (0.26) | **0.07±0.003** (0.18) | **0.06±0.002** (0.31) | **0.05±0.001** (0.29) |

- We can decompose any positive real value:

$$x = 2^{\ln x} = \overbrace{2^{\ln x - \lfloor \ln x \rfloor}}^{M \in [1,2)} . 2^{\overbrace{\lfloor \ln x \rfloor}^{E \in \mathbb{Z}}}$$

- In FP quantization we seek the optimal bits allocation $n_1, n_2$ for the mantissa and exponent, respectively.

$$x_q = \begin{cases} 2^{E_{\max}} & E \geq E_{\max} \\ M_q \cdot 2^E & -E_{\max} \leq E \leq E_{\max} \\ 0 & E \leq -E_{\max} \end{cases}$$

- The relative error between a FP number $x_q$ and a real number $x$ is:

$$\eta(n_1, n_2) = \left| \frac{x_q - x}{x} \right|$$

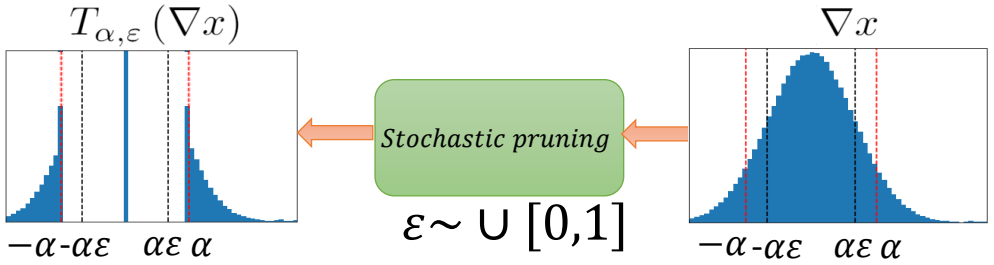
- Assuming $x \sim \text{Lognormal}(\mu, \sigma^2)$ we can obtain a closed formula for the relative error:

$$E\left[\eta(n_1, n_2)\right] = \frac{2\Phi\left(\frac{E_{\max}}{\sigma}\right) - 1}{8 \cdot \ln(2) \cdot (2^{n_1} - 1)} + 2^{E_{\max}-1} e^{\frac{\sigma^2 \ln^2(2)}{2}} \left( \text{erf}\left( \frac{\sigma \ln 2}{\sqrt{2}} + \frac{E_{\max}}{\sqrt{2}\sigma} \right) - 1 \right) - \frac{1}{2} \text{erf}\left( \frac{E_{\max}}{\sqrt{2}\sigma} \right) + \frac{3}{2} - \Phi\left( \frac{E_{\max}}{\sigma} \right)$$

| Dataset | Model | $\sigma$ Range | Baseline | FP | $E^*$ | $E^*$+1 | $E^*$-1 | $E^*$-2 |
|---|---|---|---|---|---|---|---|---|
| Cifar100 | ResNet18 | 2.5 - 4.5 | 64.9% | FP5 | **64.0%** | N/A | 58.9%† | 26.6% |
| | | | | FP6 | **64.9%** | 64.6% | 59.7%† | 28.6% |
| | ResNet101 | 2.5-4.5 | 71.3% | FP5 | **70.4%** | N/A | 66.5%† | 35% |
| | | | | FP6 | **70.97%** | 70.82% | 67.5%† | 42.7% |
| ImageNet | ResNet18 | 3 - 5.5 | 70.4% | FP6 | **70.0%** | N/A | 67.1%† | 30.8% |
| | | | | FP7 | **70.4%** | 70.1% | 66.7% | 47.5%† |
| | SqueezeNet | 3 - 5.5 | 58.19 % | FP5 | **55.2%** | N/A | 47.3%† | 33.2% |
| | | | | FP6 | **57.8%** | N/A | 56.1%† | 54.3% |

# tinyML EMEA Technical Forum 2021
# June 7-10, 2021

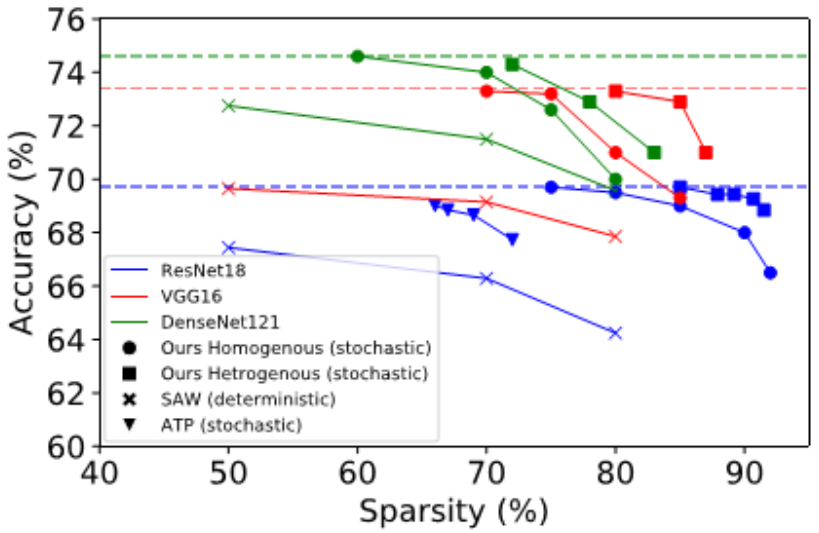
- Given a threshold $\alpha$ and a uniform random variable $\varepsilon \sim U[0,1]$ stochastic pruning is defined:

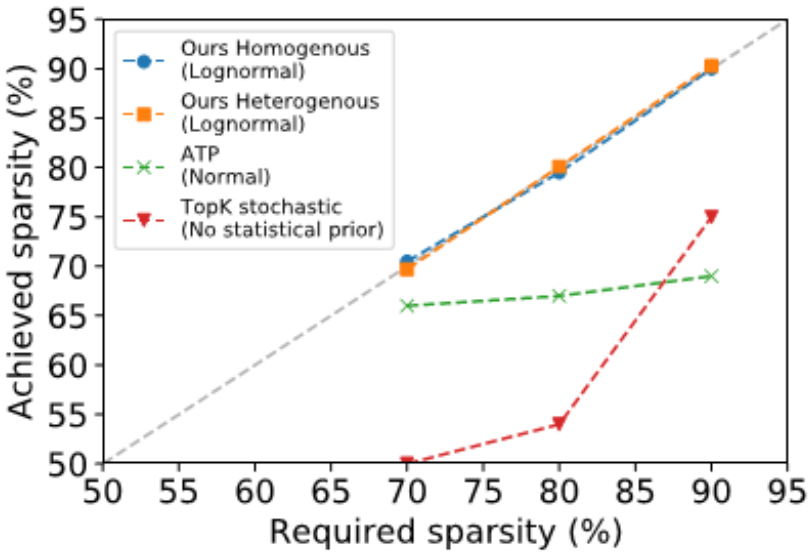



$$T_{\alpha,\varepsilon}(x) = \begin{cases} x & |x| > \alpha \\ \text{sign}(x) \cdot \alpha & \alpha \cdot \varepsilon \le |x| \le \alpha \\ 0 & |x| < \alpha \cdot \varepsilon \end{cases}$$

- Assuming that $x \sim \text{Lognormal}(\mu, \sigma^2)$ we can obtain a closed formula to obtain threshold $\alpha$ which induces a required sparsity ratio S.

$$S = \frac{1}{2} + \frac{e^{\mu}}{2\alpha}\left[e^{\frac{\sigma^2}{2}}\,\text{erf}\left(\frac{\sigma}{\sqrt{2}} - \frac{\ln\left(\frac{\alpha}{e^{\mu}}\right)}{\sqrt{2}\sigma}\right) + \frac{\alpha}{e^{\mu}} \cdot \text{erf}\left(\frac{\ln\left(\frac{\alpha}{e^{\mu}}\right)}{\sqrt{2}\sigma}\right) - e^{\frac{\sigma^2}{2}}\right]$$



TINY ML EMEA
Virtual
JUNE 7-10,2021

# Premier Sponsor

Intelligent Agent
Neuton

# Automated TinyML

Zero-code SaaS solution

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

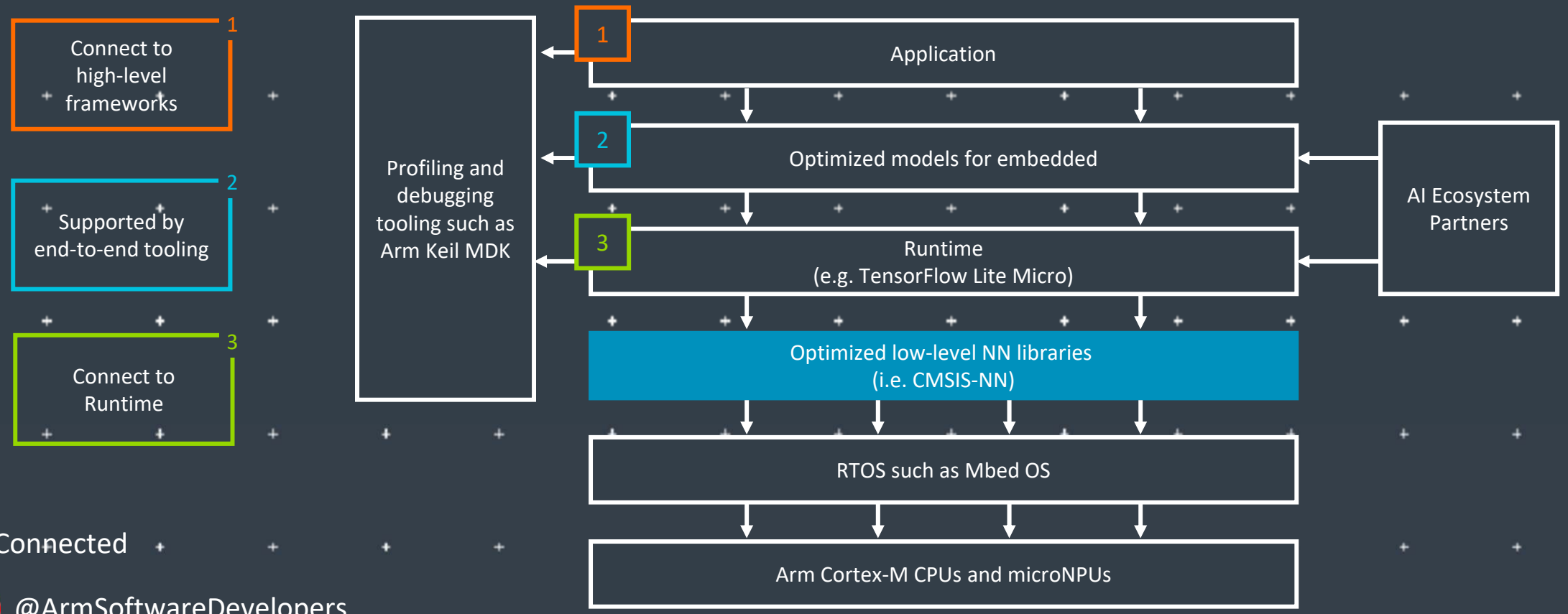
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

***Build Fast. Build Once. Never Compromise.***

TINY ML EMEA
Virtual
JUNE 7-10,2021

# Executive Sponsors

# Arm: The Software and Hardware Foundation for tinyML




Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

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




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

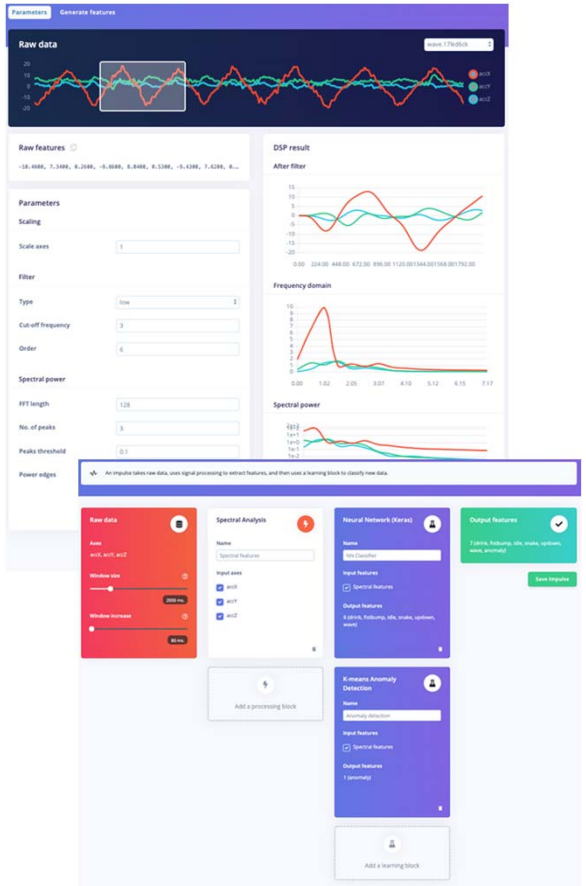
Open source SDK

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

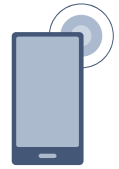
Reinforcement learning for decision making

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

TINY ML EMEA
Virtual
JUNE 7-10,2021

# 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

TINY ML EMEA
Virtual
JUNE 7-10,2021

# Gold Sponsors

LatentAI

# Adaptive AI for the Intelligent Edge

Latentai.com

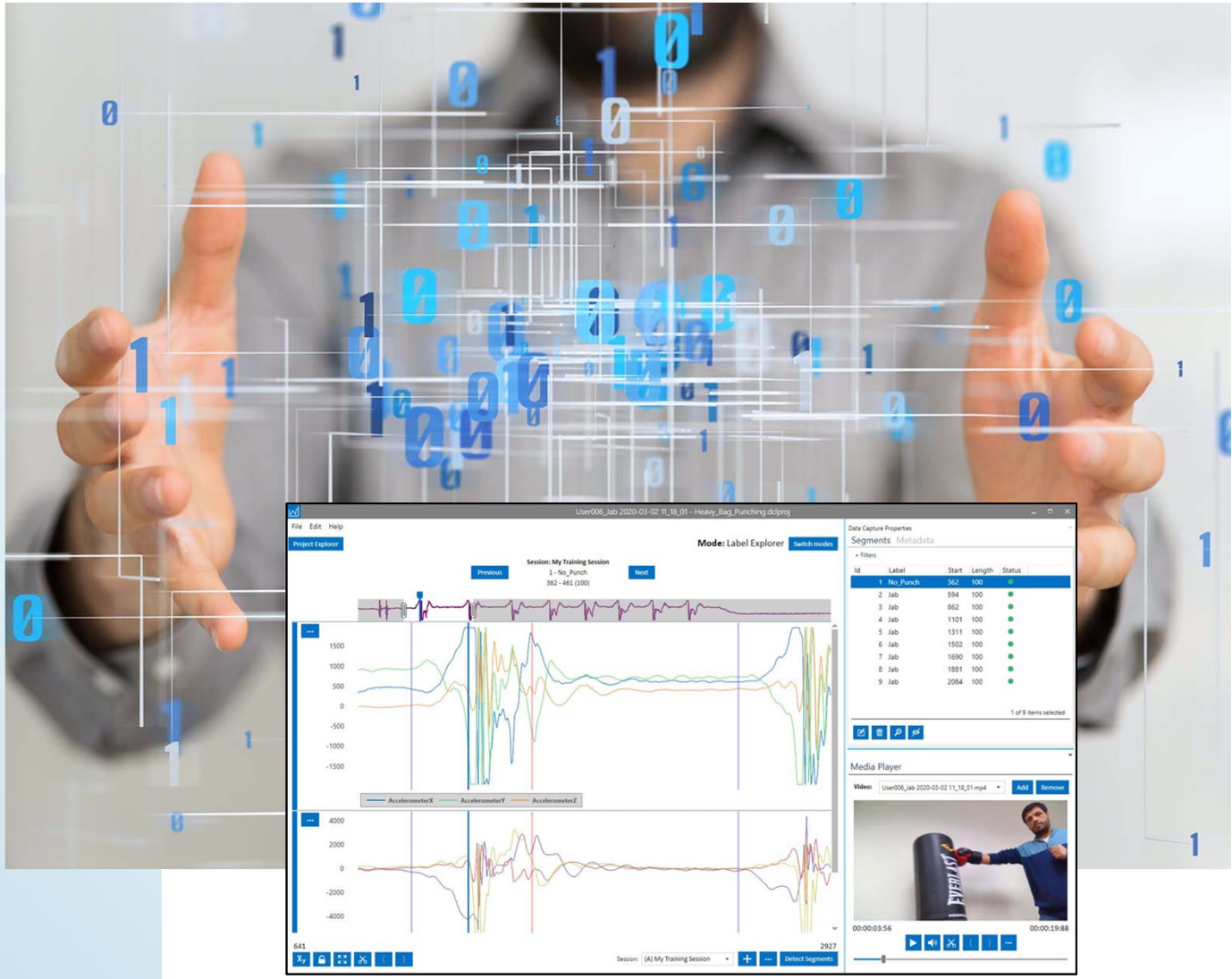
SensiML™

# Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

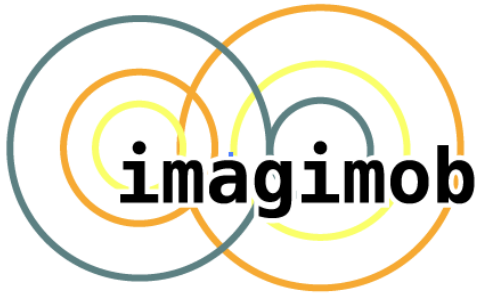



sensiml.com

TINY ML EMEA Virtual JUNE 7-10,2021

# Silver Sponsors






# Copyright Notice