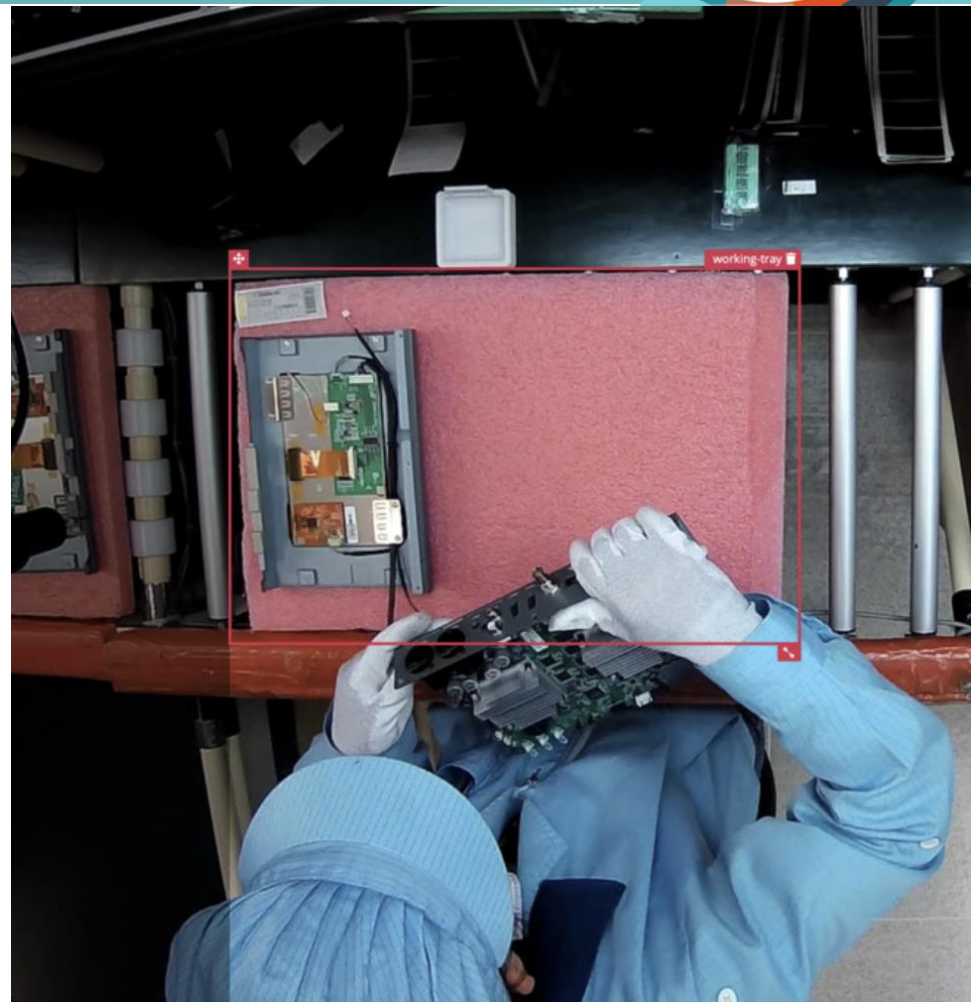
How Do We Enable Edge ML Everywhere? Data, Reliability, and Silicon Flexibility

Zach Shelby
Co-founder and CEO
Edge Impulse

Advantech increases manufacturing productivity by 15%

Visual inspection system to monitor worker safety and flag delays on the production line in real-time.

- A reported 15% overall increase in production line efficiency
- Faster detection of idle time raises assembly-line productivity
- Managers free up time to focus on production planning and operations



EDGE IMPULSE

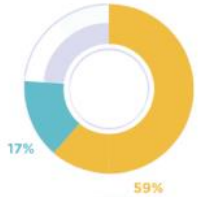
State of the edge ML 2022

User population

1K+ People polled

91 Countries represented

Frameworks



59% of participants with ML experience use **TensorFlow** and **17%** use **PyTorch** as their ML framework.

Top 3 Data Types

Participants with ML Experience

Professional Experience

37% of survey participants have 1+ years of professional ML experience.

53% of survey participants have 1+ years of professional embedded systems experience.

Top 5 Favourite Boards for Edge ML Projects

1. Raspberry Pi 4 
2. Arduino Nano 33 BLE Sense 
3. ESP32 
4. Raspberry Pi Pico 
5. NVIDIA Jetson Nano 

Common ML Tasks

Dataset Source

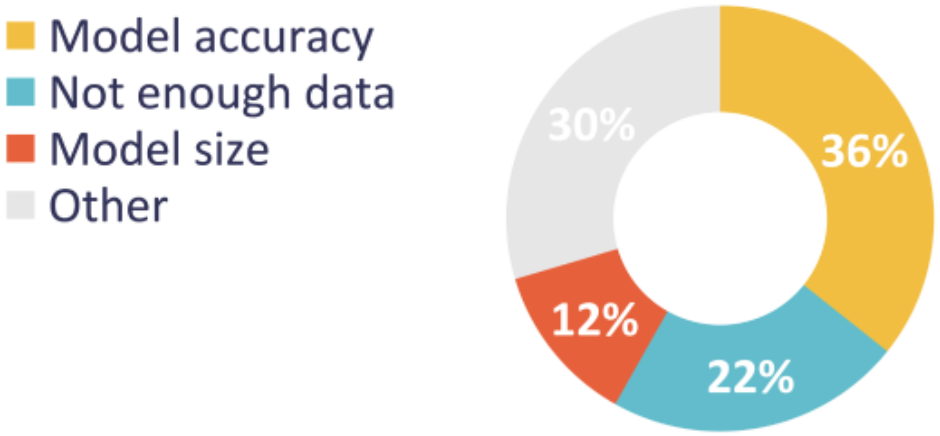
Participants with ML Experience

EDGE IMPULSE

What are the barriers to edge ML scale?

Biggest Concerns

Users with ML experience



Challenges with Data

Users with ML experience

1. Extracting meaningful features from data
2. Cleaning and preparing data
3. Obtaining data
4. Understanding what data to focus on
5. Managing large datasets

Top 5 Challenges

Preventing Users from Getting Involved in Edge ML

1. Lack of time
2. Lack of expertise in Data Science
3. Lack of resources
4. Difficulty in collecting the right data
5. Challenge with developing models optimized for the edge

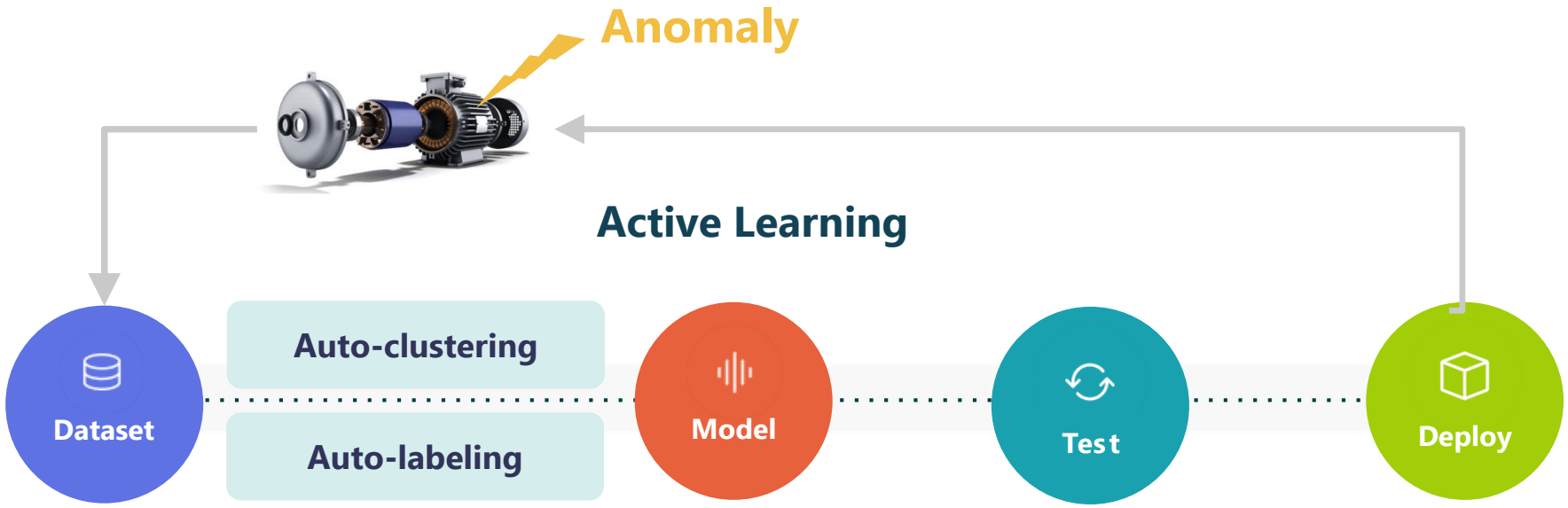
Unleashing the data

EDGE IMPULSE

Big data, big model is a problem

- State of the art: Monolithic batch, from big data to big model
- Developed for ML research, unsuitable for most edge ML applications

Data-Centric ML – from zero to hero



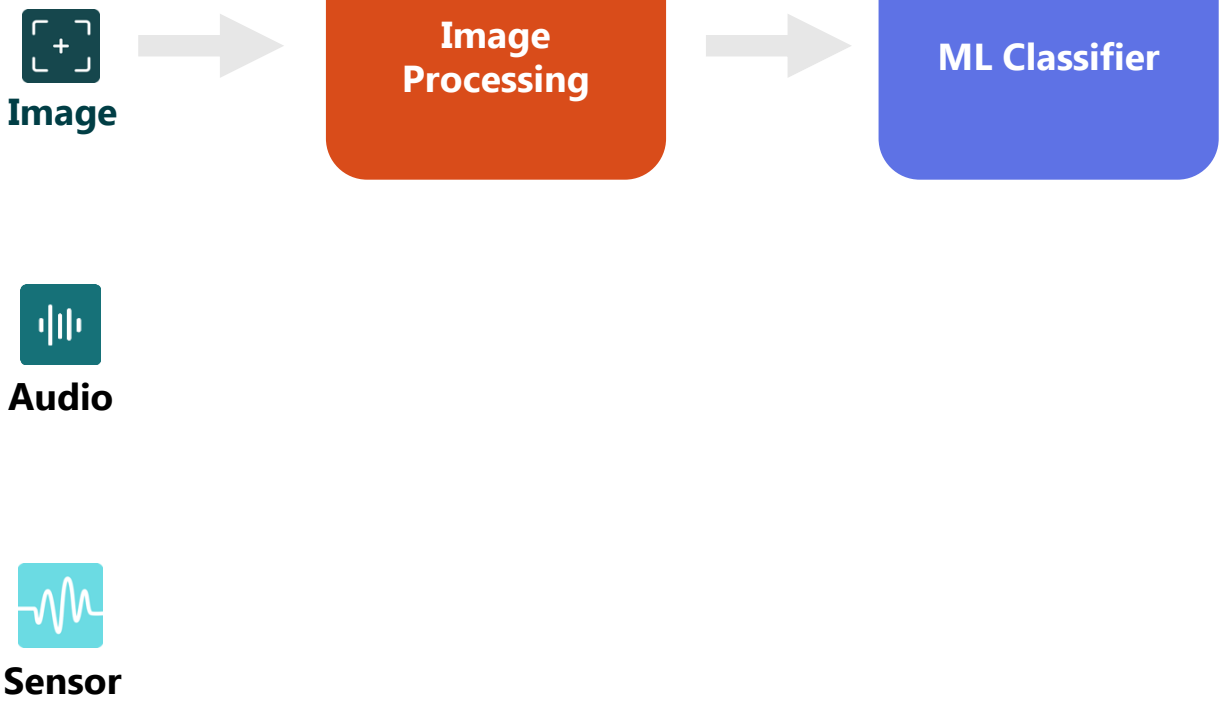
- Auto-clustering, using feature analysis to help experts quickly label data
- Auto-labeling, using pre-trained expert models to suggest labels
- Active learning, using inference to drive the data collection process

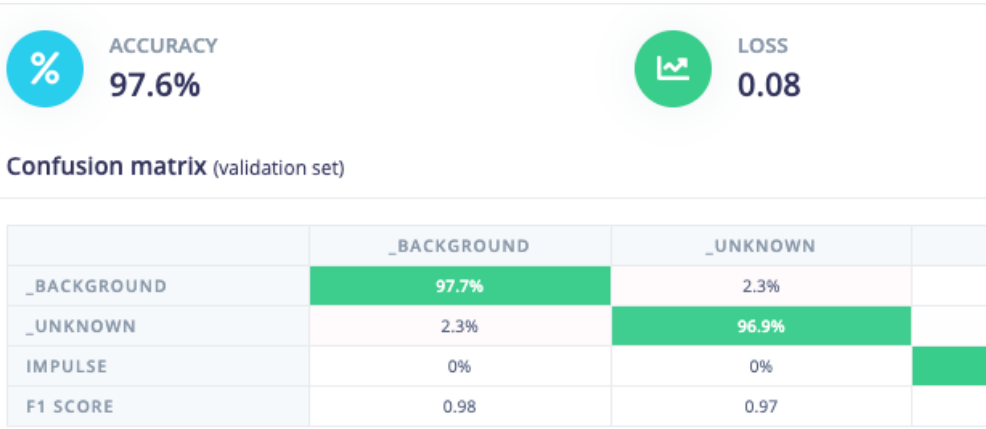
Auto-clustering in action

Making ML industrial grade

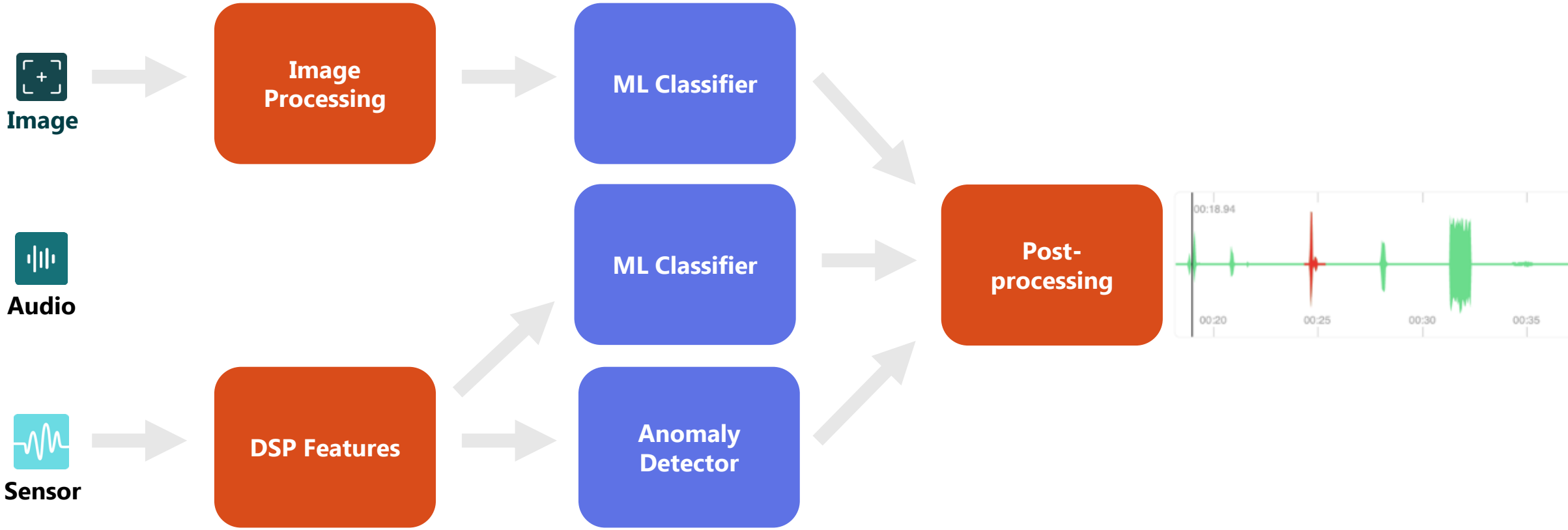
EDGE IMPULSE

Sensor fusion for industrial-grade ML

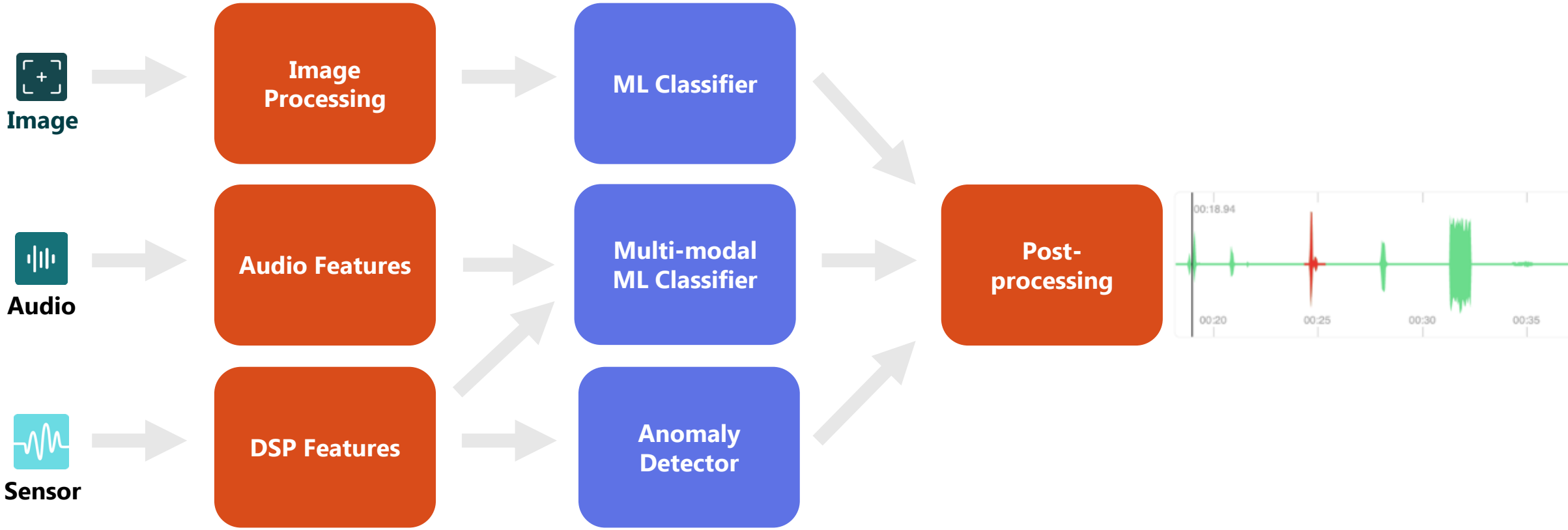




Sensor fusion for industrial-grade ML

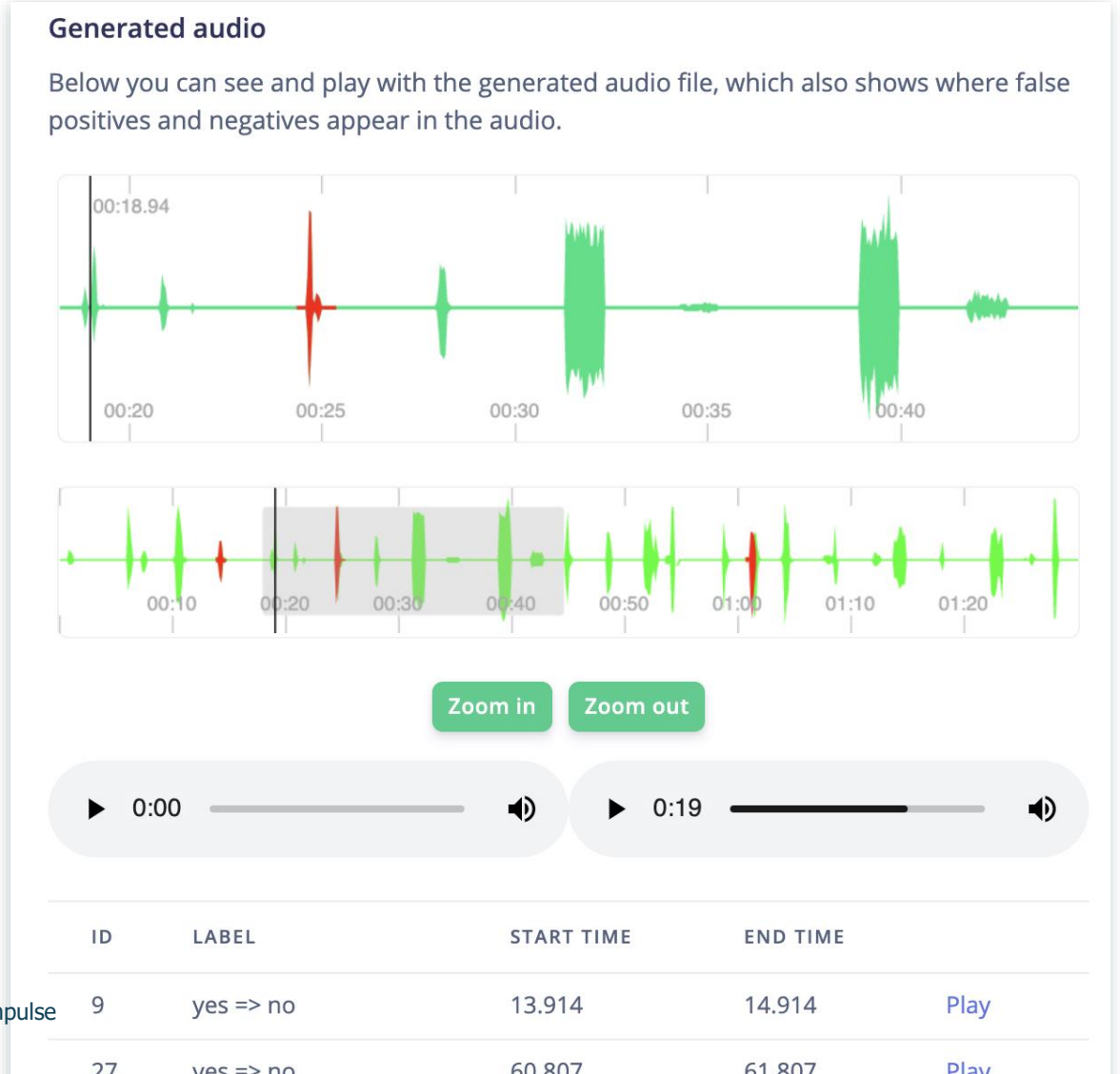


Sensor fusion for industrial-grade ML



Calibrating performance at scale

- Model training validation != performance
- Requires testing on the entire algorithm with real-world data for realistic performance
- Understand the impact of post-processing while accounting for device constraints and latency
- Choose the ideal balance between false activations and false rejections
- Leverage genetic algorithms to design optimal post-processing configuration

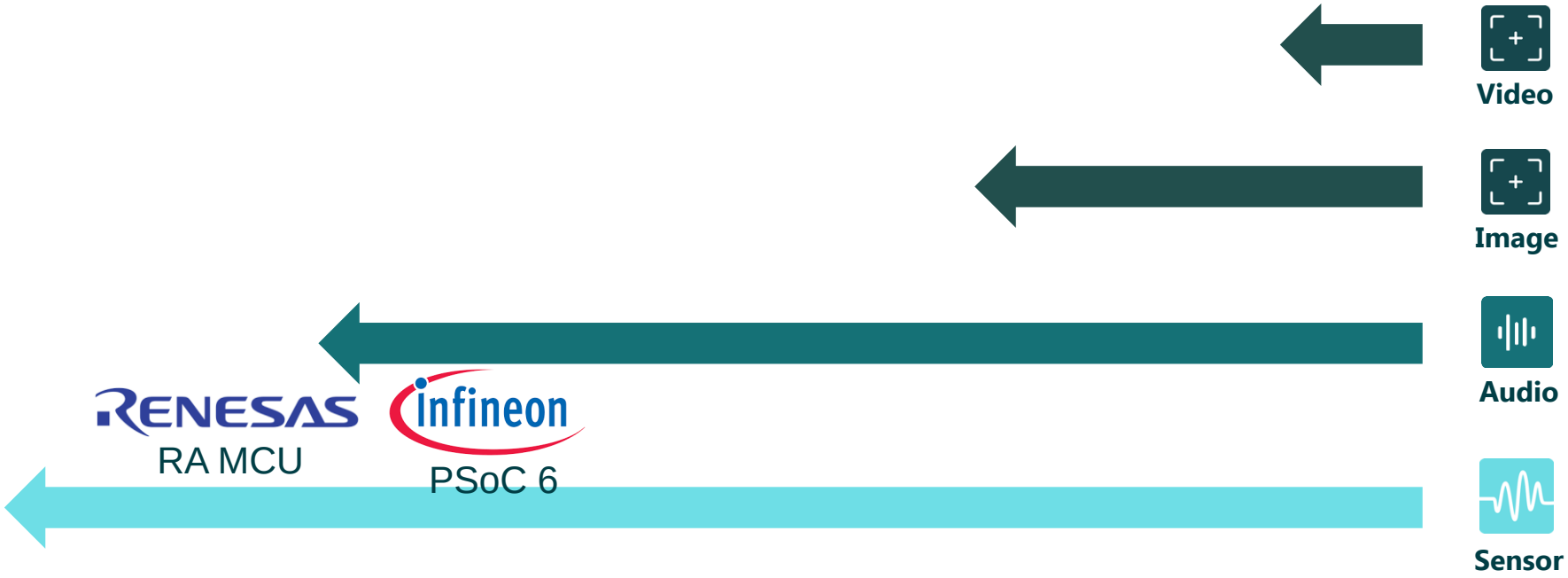


EDGE IMPULSE

ML on today and tomorrow's silicon

EDGE IMPULSE

Hardware is sexy again!



Hardware is sexy again!

Hardware is sexy again!

Synaptics™ Katana

ALIF SEMICONDUCTOR M55/U55

brainchip™ Akida

SYNTIANT NDP200

SILICON LABS xG24

SYNTIANT NDP100

RENESAS RA MCU

Infineon PSoC 6

Video

Image

Audio

Sensor

Wakeup NPU

Low-end MCU

MCU + FPU

MCU + NPU

CPU

GPU

Hardware is sexy again!

RENESAS RZ MPU

TEXAS INSTRUMENTS TD4VA

Synaptics Katana

ALIF SEMICONDUCTOR M55/U55

brainchip Akida

Video

SYNTIANT NDP200

SILICON LABS xG24

Image

Audio

SYNTIANT NDP100

RENESAS RA MCU

Infineon PSoC 6

Sensor

Wakeup NPU

Low-end MCU

MCU + FPU

MCU + NPU

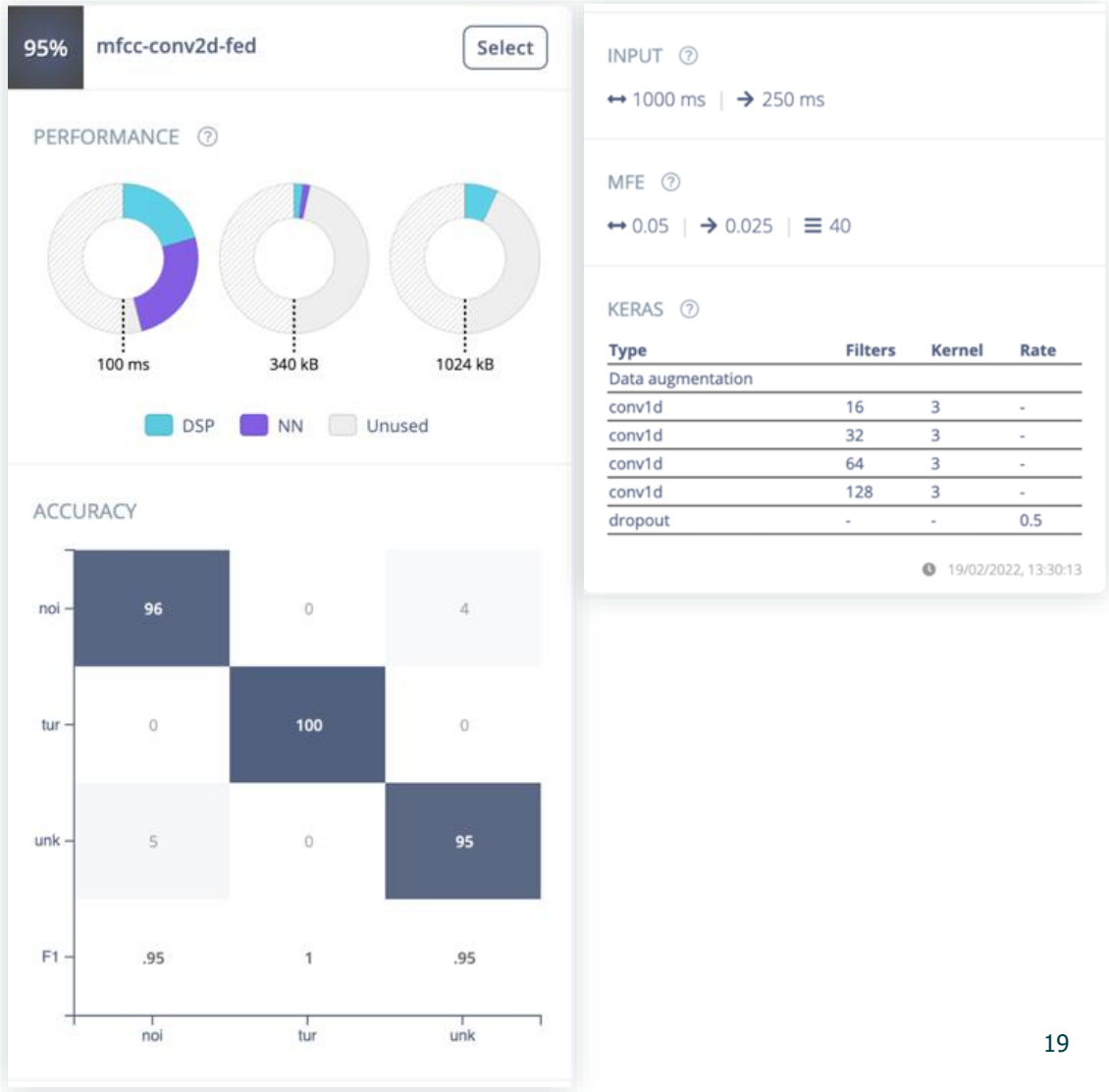
CPU + NPU

NPU + GPU

EDGE IMPULSE

Hardware profiling & tuning

- The power of hardware profiling
- Digital twin of ML on hardware
- We are combining hardware profiling with hyperparameter search – **EON Tuner**
- Hardware-aware AutoML across data, pre-processing and ML blocks



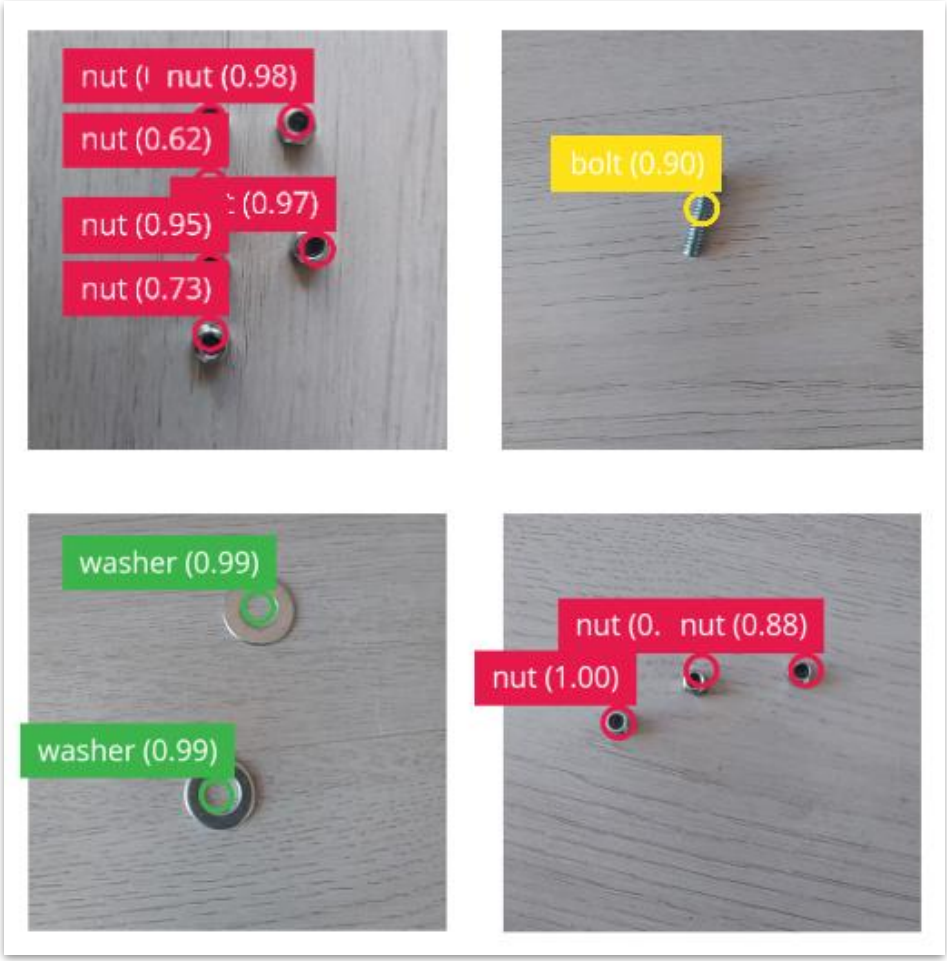
On-device performance

INFERENCING TIME	PEAK RAM USAGE	FLASH USAGE
1 ms.	1.7K	19.4K

FOMO: Faster objects, more objects

- Object detection on constrained compute
- **20x** performance improvement
- Scales down to 100k RAM to enable MCUs
- Better at detecting smaller and more numerous objects
- Capable of segmentation and counting objects

	Cortex-M4	Cortex-M7	Cortex-A	Nvidia
FOMO	2 fps	15-30 fps	60+ fps	150+ fps
SSD	NA	NA	3 fps	20 fps



Resources

Increasing Manufacturing Efficiency

https://casestudies.edgeimpulse.com/industrial-iot-with-advantech

Constrained object detection: FOMO Blog

docs.edgeimpulse.com/docs/tutorials/fomo-object-detection-for-constrained-devices

Request a demo

edgeimpulse.com/schedule-a-demo

Don’t miss these talks!

FOMO: Real-time Object Detection on Microcontrollers

Jan Jongboom, Co-founder and CTO, Edge Impulse

- **Date:** Wednesday, May 18
- **Start Time:** 10:50 am

Deep Dive: Develop and Deploy Advanced Edge Computer Vision—Fast!

Jenny Plunkett and Shawn Hymel, Snr. DevRel engineers

- **Date:** Thursday, May 19
- **Time:** 9 am – 12:00 pm