



tinyML[®] Talks

Enabling Ultra-low Power Machine Learning at the Edge

DNN based AI application “Everywhere and Anywhere”

Amit Roy - AigenEdge Private Limited

August 3, 2021



www.tinyML.org

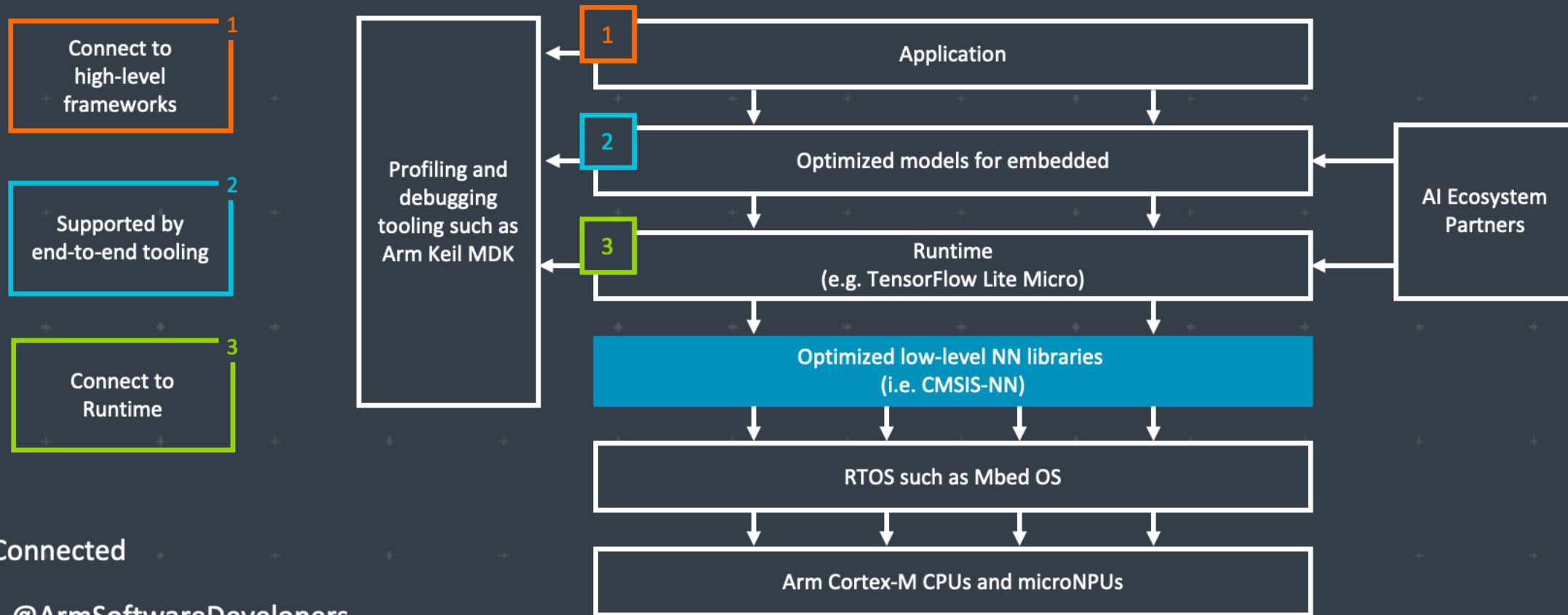


tinyML Talks Sponsors



Additional Sponsorships available – contact Olga@tinyML.org for info

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

 @ArmSoftwareDevelopers

 @ArmSoftwareDev

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



WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT



Automatically compress SOTA models like MobileNet to <200KB with **little to no drop in accuracy** for inference on resource-limited MCUs



Reduce model optimization trial & error from weeks to days using Deeplite's **design space exploration**



Deploy more models to your device without sacrificing performance or battery life with our **easy-to-use software**

BECOME BETA USER bit.ly/testdeeplite

TinyML for all developers



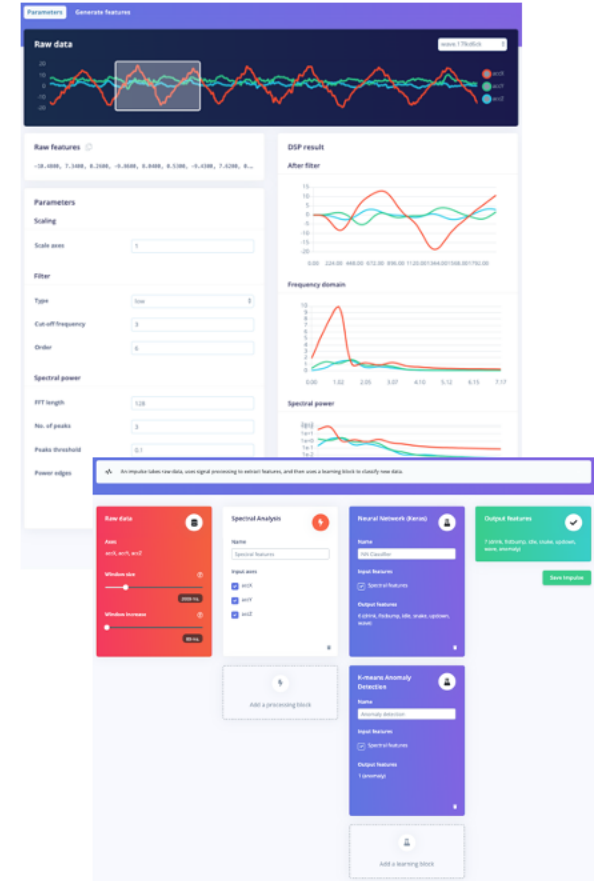
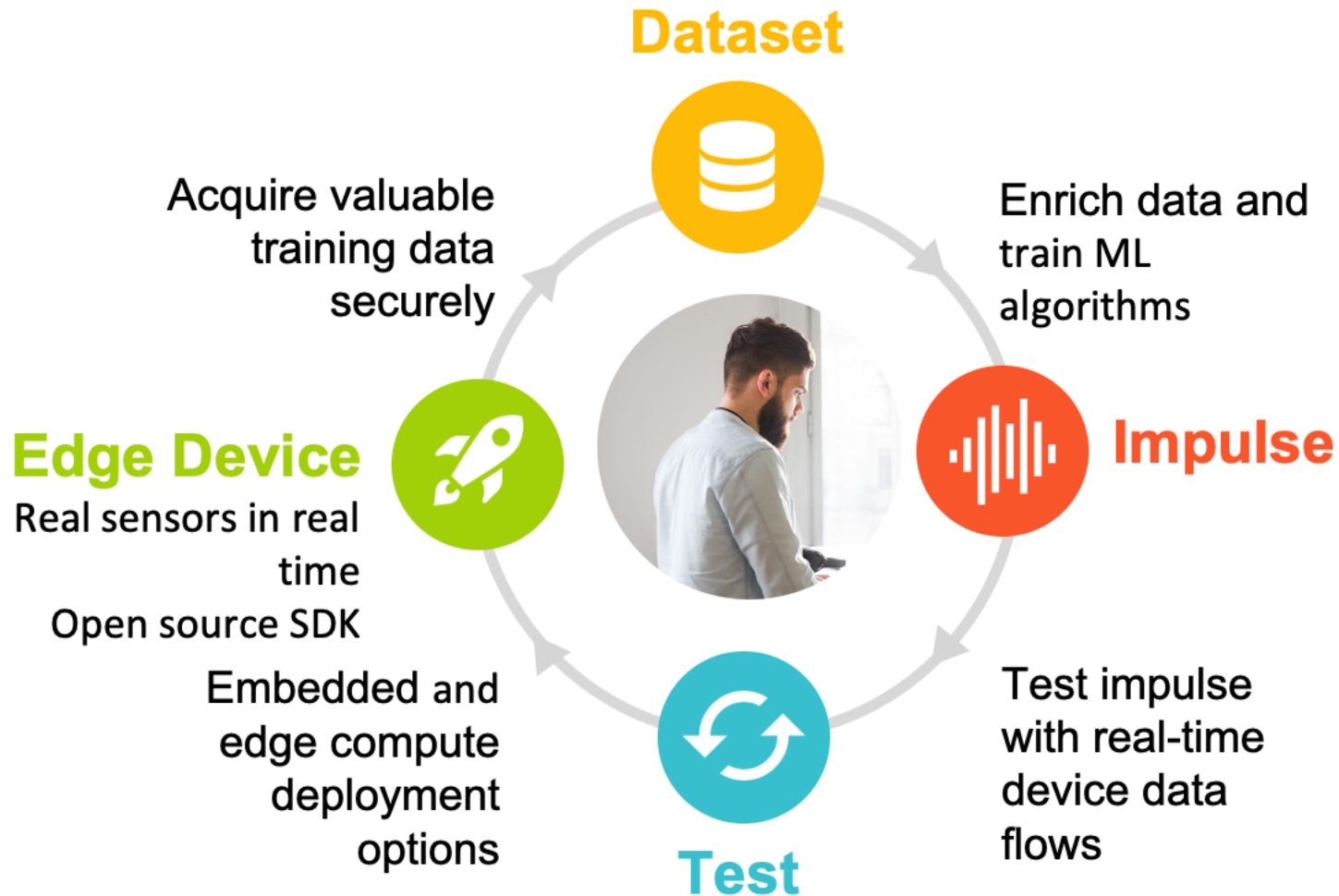
C++ library



Arduino library



WebAssembly

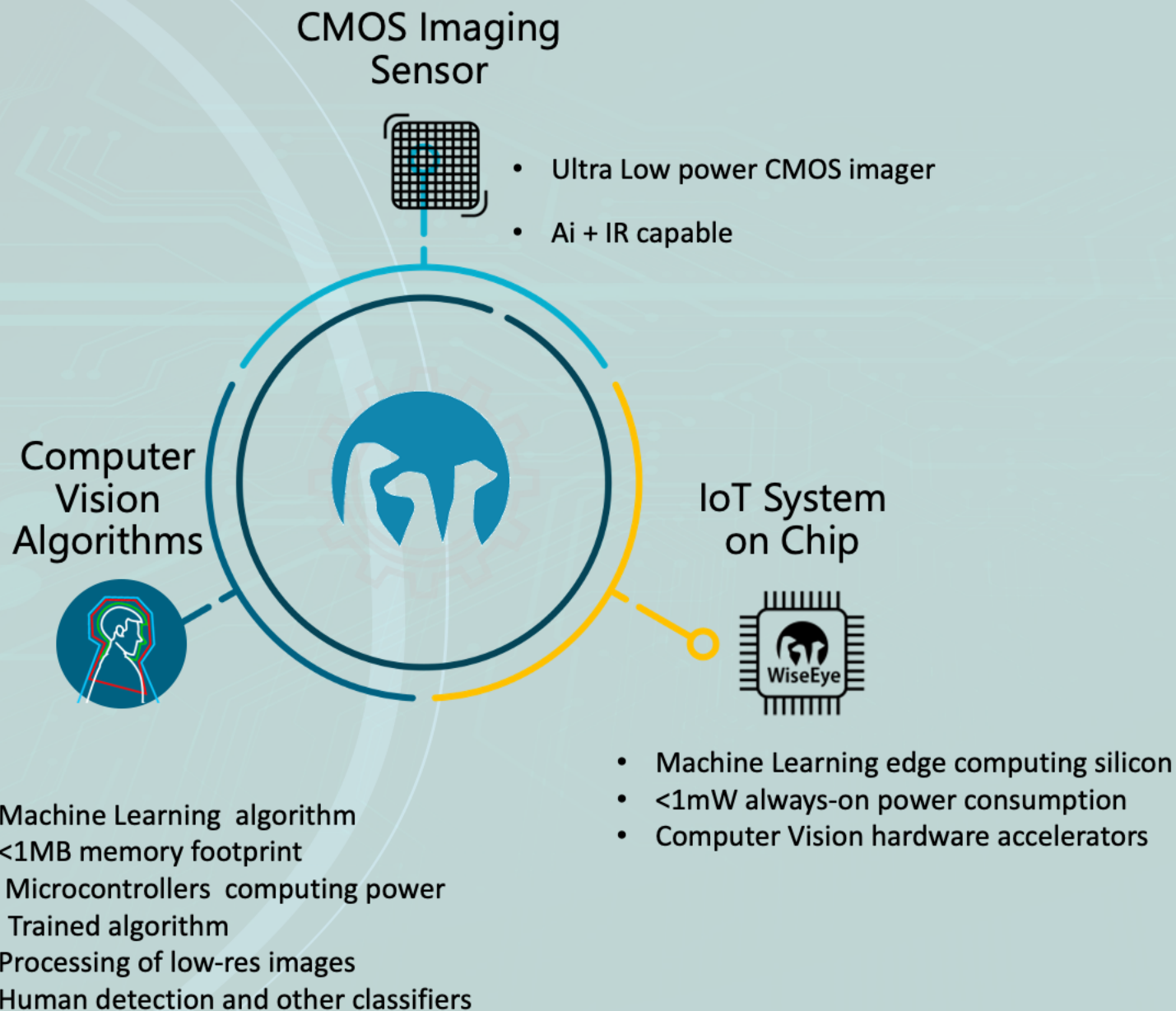


www.edgeimpulse.com



The Eye in IoT

Edge AI Visual Sensors



info@emza-vs.com



Enabling the next generation of **Sensor and Hearable products** to process rich data with energy efficiency

Visible Image



Sound



IR Image



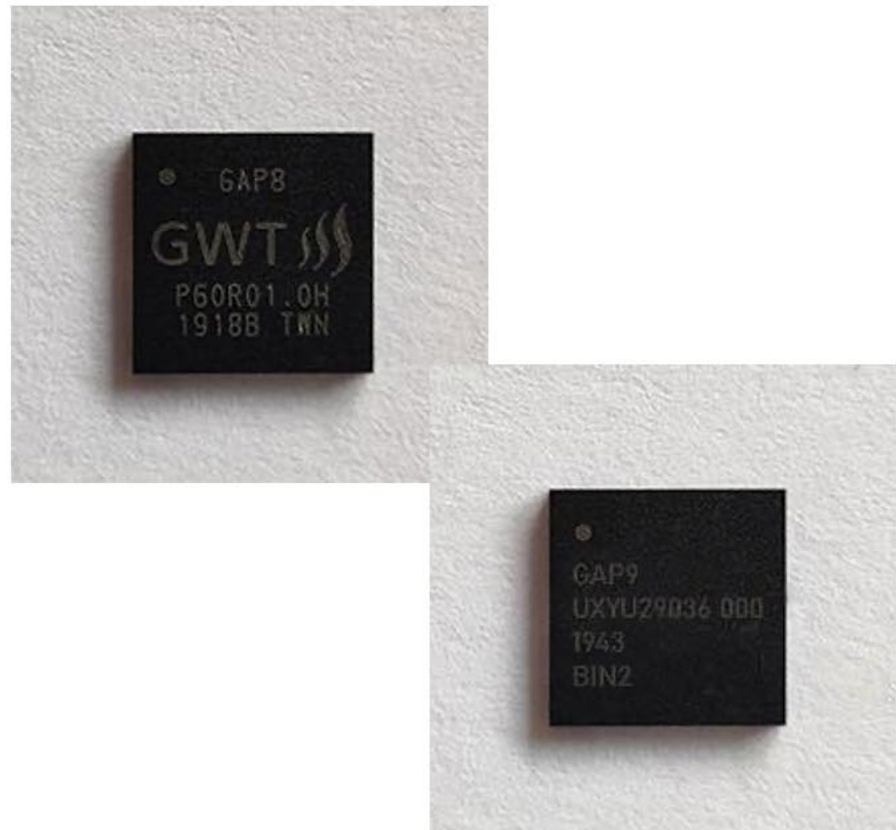
Radar



Bio-sensor



Gyro/Accel



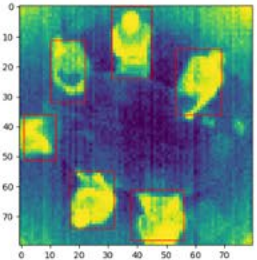
Wearables / Hearables



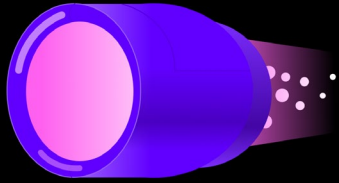
Battery-powered consumer electronics



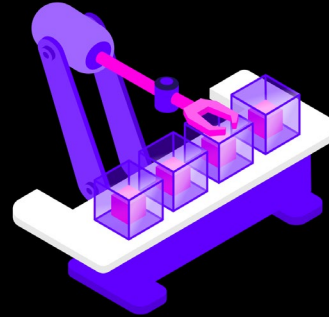
IoT Sensors



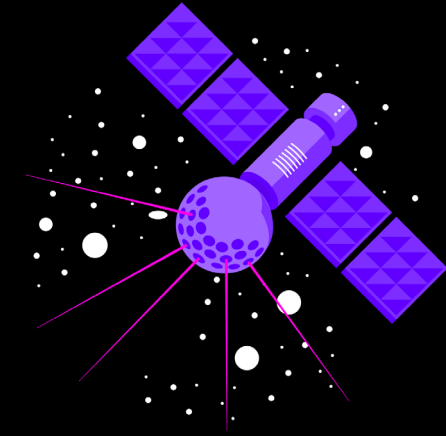
Distributed infrastructure for TinyML apps



Develop at warp speed



Automate deployments



Device orchestration

HOTG is building the **distributed infrastructure** to pave the way for **AI enabled edge applications**



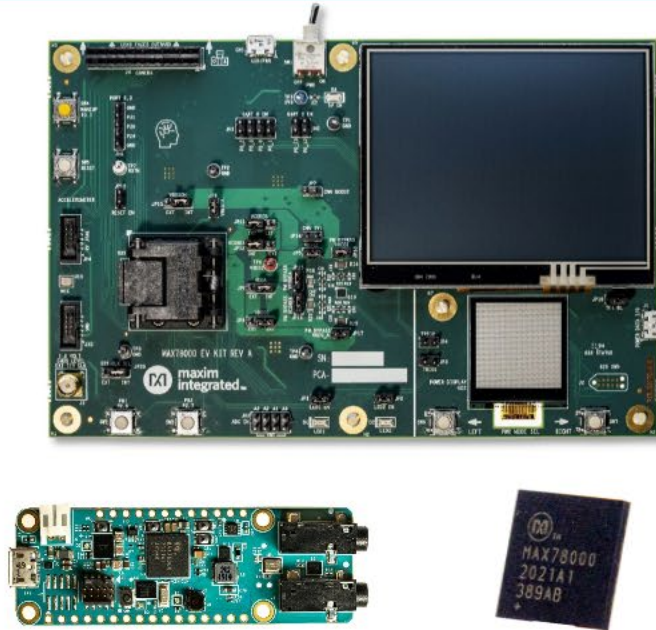
Latent AI

Adaptive AI for the Intelligent Edge

latent.ai

Maxim Integrated: Enabling Edge Intelligence

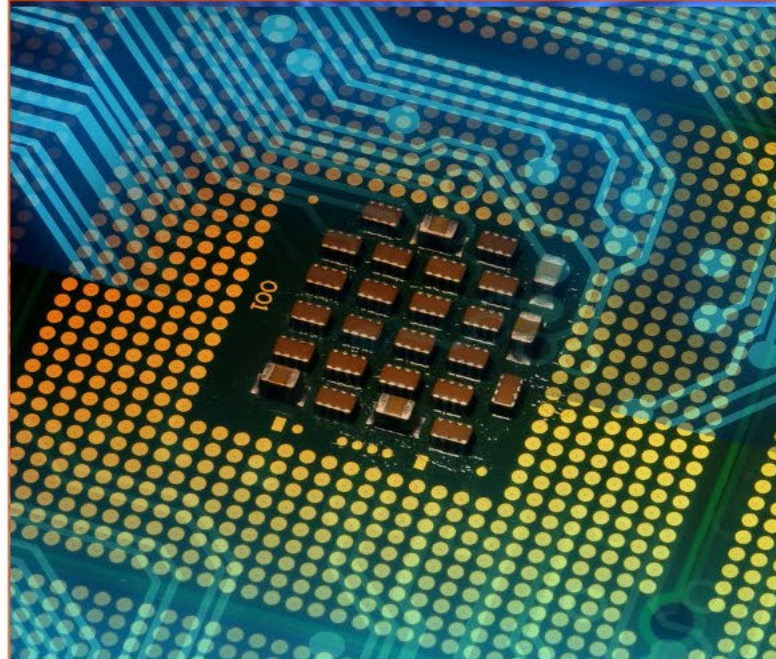
Advanced AI Acceleration IC



The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

www.maximintegrated.com/MAX78000

Low Power Cortex M4 Micros



Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

www.maximintegrated.com/microcontrollers

Sensors and Signal Conditioning



Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

www.maximintegrated.com/sensors

Qeexo AutoML

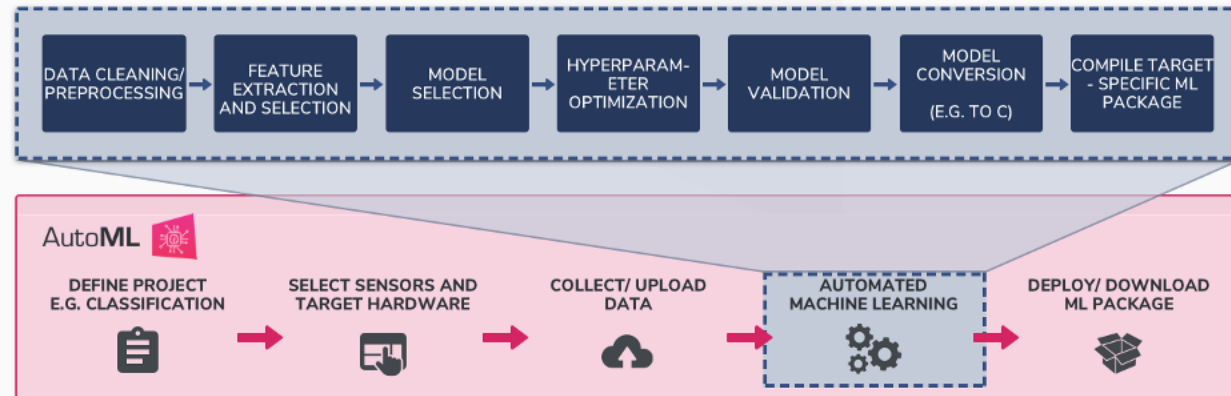


Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
 - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
 - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™ - M0 to M4 class MCUs

End-to-End Machine Learning Platform



For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

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



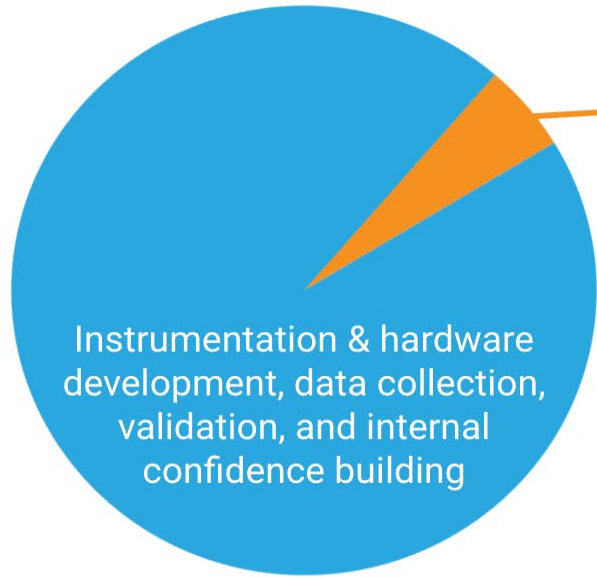
IoT/IIoT



Automotive



Mobile



Only 5% of your project costs will be spent on model building



software helps with the other 95% too

More information at <https://reality.ai/>

Join us at:

Sensors Converge
(San Jose + Virtual, 9/21-24)

Autotech Council
(Silicon Valley + Virtual, 10/14-15)

Infineon Oktobertech
(Silicon Valley, 10/21)

 **Reality AI** Tools[®] software

AI Explore™
(AutoML)

Sensor Selection
and BOM
Optimization

Data Readiness

Edge AI / TinyML
Code Optimization

Optional Add-ons
Reality AI for MATLAB
Reality AI for Radar

Algorithmically-Driven Feature Discovery



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.



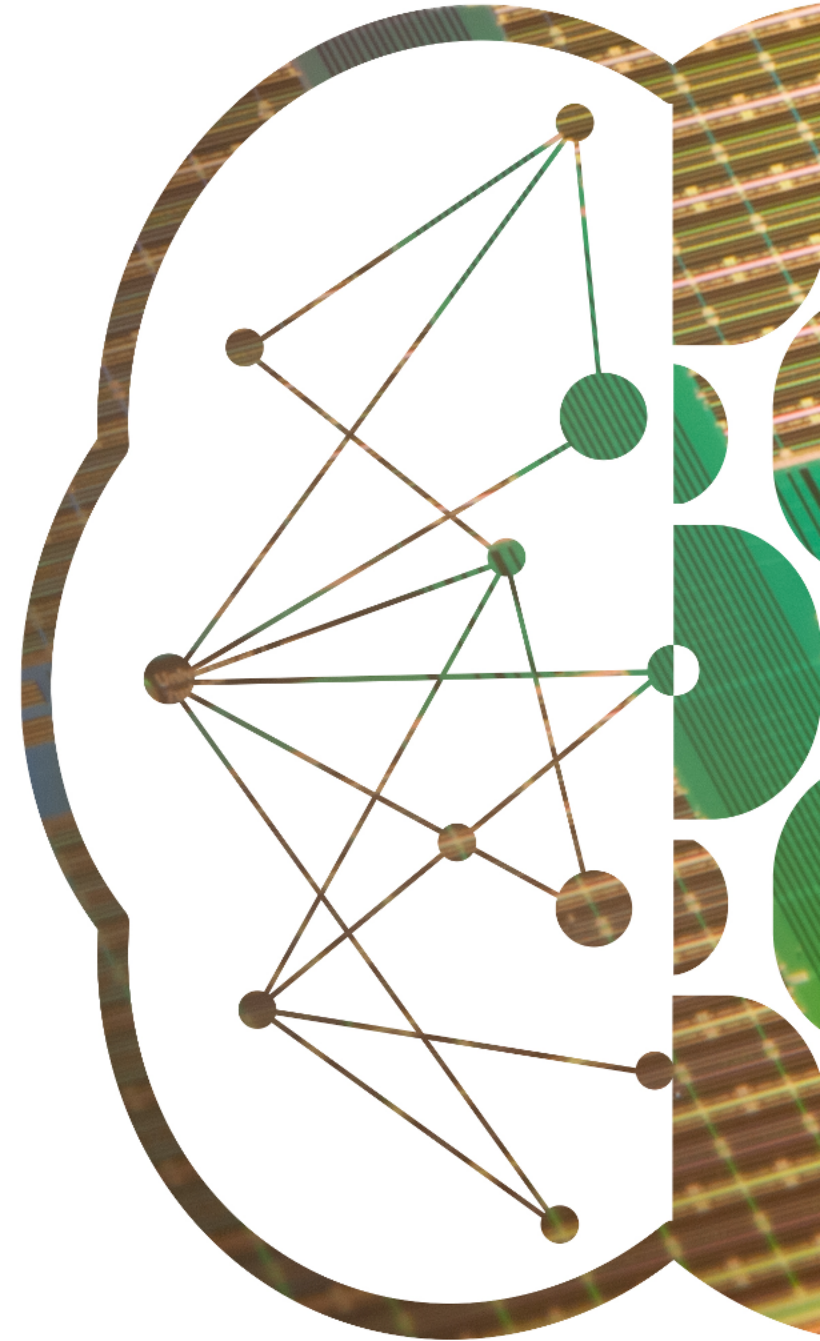
sensiml.com



SynSense

SynSense builds **sensing and inference** hardware for **ultra-low-power** (sub-mW) **embedded, mobile and edge** devices. We design systems for **real-time always-on smart sensing**, for audio, vision, IMUs, bio-signals and more.

<https://SynSense.ai>



SYNTIANT

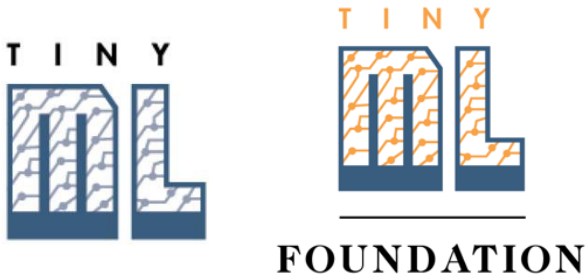
[Syntiant Corp.](#) is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications 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

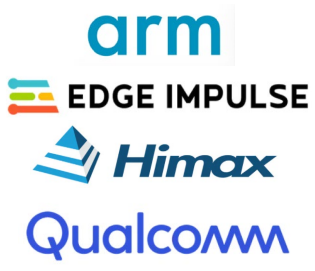
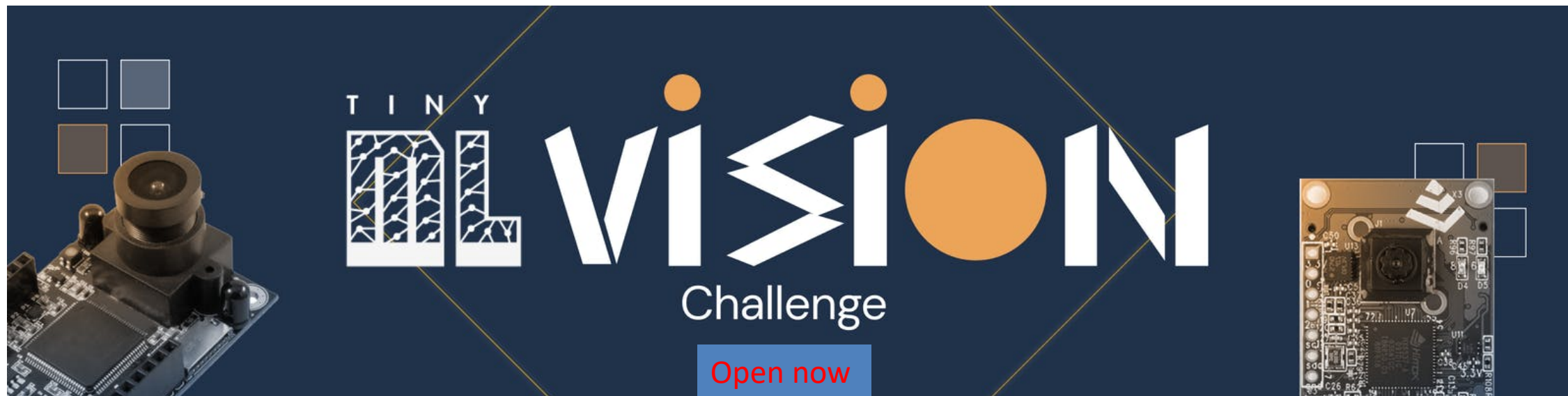


collaboration with



Focus on:

(i) developing new use cases/apps for tinyML vision; and (ii) promoting tinyML tech & companies in the developer community



Submissions accepted until August 20th, 2021
Winners announced on September 1, 2021 (\$6k value)
Sponsorships available: sponsorships@tinyML.org



<https://www.hackster.io/contests/tinyml-vision>



Successful tinyML EMEA 2021



- Videos are available on www.youtube.com/tinyML

- **4** days of tinyML excitement

- **2** tutorials
- **5** keynotes
- **15** tinyTalks
- **7** lightning talks
- **3** panel discussions & networking
- **16** papers in the Student Forum
- **4** partner sessions
- **16** sponsoring companies

- **58** speakers, **1687** registered attendees!



250 videos with 121k views
as of July 10, 2021





Next tinyML Talks

Date	Presenter	Topic / Title
Tuesday, August 3	Vikram Shrivastava, Sr. Director, IoT Marketing, Knowles Corporate	Dedicated Audio Processors at the Edge are the Future of AI

Webcast start time is 8 am Pacific time

Please contact talks@tinymml.org if you are interested in presenting



Local Committee in India



Chetan Singh Thakur, PhD

Assistant Professor at the Indian Institute of Science (IISc), Bangalore. He is a Ph.D. in neuromorphic engineering. Dr. Thakur's research interest spans VLSI Design, Edge Computing, Neuromorphic Engineering.



Anup Rajput

Co-founder at Envir AI, trying to bring ML into the real world. Anup has a background in semiconductor design and applied ML from edge to cloud.



Sandipan Chatterjee

Sandipan is Lead Data scientist at DXC Technology where he develops and implements vision-based automation in manufacturing, automotive and healthcare. He has a background in image and statistical analysis.



Abhishek Nair

Abhishek is a PhD student at IISc Neuronics lab. His research area includes exploring low power ML algorithms for digital hardware implementation.



Arijit Das

Arijit is a 15-year-old high-schooler. He is the youngest Ambassador for Edge Impulse and has been in the AIoT field since 2017. His interests include Edge Computing, EdgeAI, and Low-Power Wide Area Networks.

Follow us for more updates at:

<https://www.linkedin.com/company/tinyml-india>

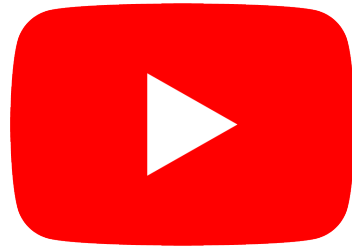


Reminders

Slides & Videos will be posted tomorrow

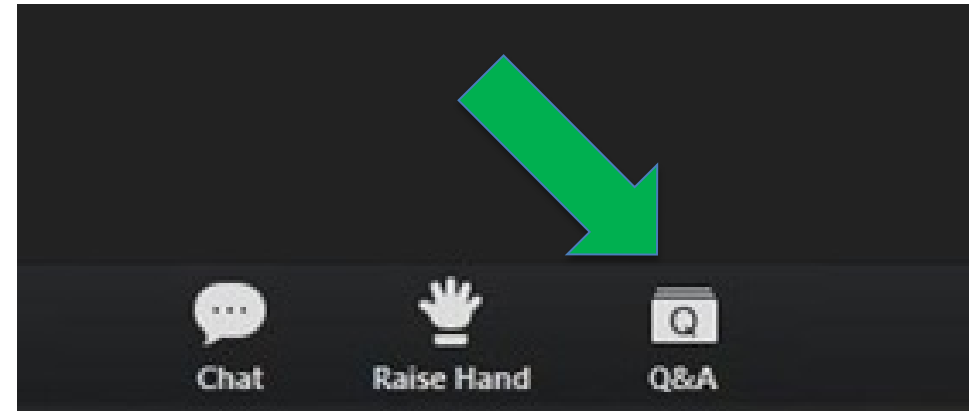


tinyml.org/forums



youtube.com/tinyml

Please use the Q&A window for your questions





Amit Roy



Dr. Amit Roy received his undergraduate degree in E&C from Delhi College of Engineering and his master's and doctoral degrees from the University of California-Berkeley. He spent over 17 years in a chip design business before founding AigenEdge, a start-up focused on developing technologies for the tiniest, lowest power, fastest DNN with built-in explainability. He has been awarded sixteen patents by USPTO.

DNN-AI: Everywhere And Anywhere

TinML Talk



AigenEdge



Table of Contents



01

Introduction

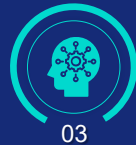
- Why AI is Booming
- Industry 4.0
- AI market
- Hardware and software market



02

TinyML

- Introduction to TinyML
- Oppurtunities
- Challanges



03

AigenEdge

- AigenEdge NAS platform
- Result



04

Q&A



AI is booming



AI-ML: The rise of data-based AI, advances in deep learning, and the necessity for robotic autonomy to remain competitive in a global market are predicted to accelerate AI adoption

USD 58.3 billion
Y'21

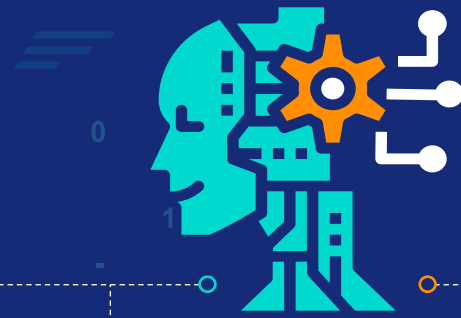


CAGR-39.7%



USD 309.6 billion
Y'26

AI is Everywhere but NOT Anywhere



Energy Feedstock & Utilities

- Power Usage Analytics
- Seismic Data Processing
- Smart Grid Management
- Energy Demand & Supply Optimization



Financial Services

- Risk Analytics & Regulation
- Customer Segmentation
- Credit Worthiness Evaluation



Travel & Hospitality

- Aircraft Scheduling
- Dynamic Pricing
- Traffic Patterns & Congestion Management



Manufacturing

- Predictive Maintenance or Condition Monitoring
- Demand Forecasting
- Process Optimization
- Telematics



Retail

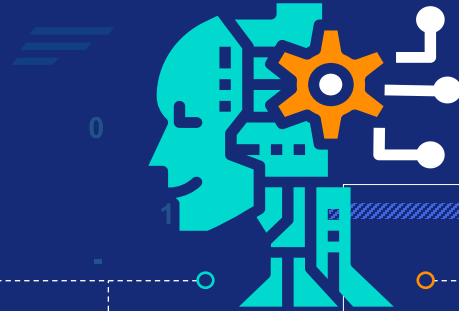
- Predictive Inventory Planning
- Recommendation Engines
- Your Text Here
- Customer ROI & Lifetime Value



Healthcare & Life Sciences

- Alerts & Diagnostics from Real-time Patient Data
- Predictive Health Management
- Healthcare Provider Sentiment Analysis

Manufacturing Industry is Adopting AI Fast and Furious



Energy Feedstock & Utilities

- Power Usage Analytics
- Seismic Data Processing
- Smart Grid Management
- Energy Demand & Supply Optimization



Financial Services

- Risk Analytics & Regulation
- Customer Segmentation
- Credit Worthiness Evaluation



Travel & Hospitality

- Aircraft Scheduling
- Dynamic Pricing
- Traffic Patterns & Congestion Management



Manufacturing

- Predictive Maintenance or Condition Monitoring
- Demand Forecasting
- Process Optimization
- Telematics



Retail

- Predictive Inventory Planning
- Recommendation Engines
- Your Text Here
- Customer ROI & Lifetime Value



Healthcare & Life Sciences

- Alerts & Diagnostics from Real-time Patient Data
- Predictive Health Management
- Healthcare Provider Sentiment Analysis

Advantage of AI in Industries



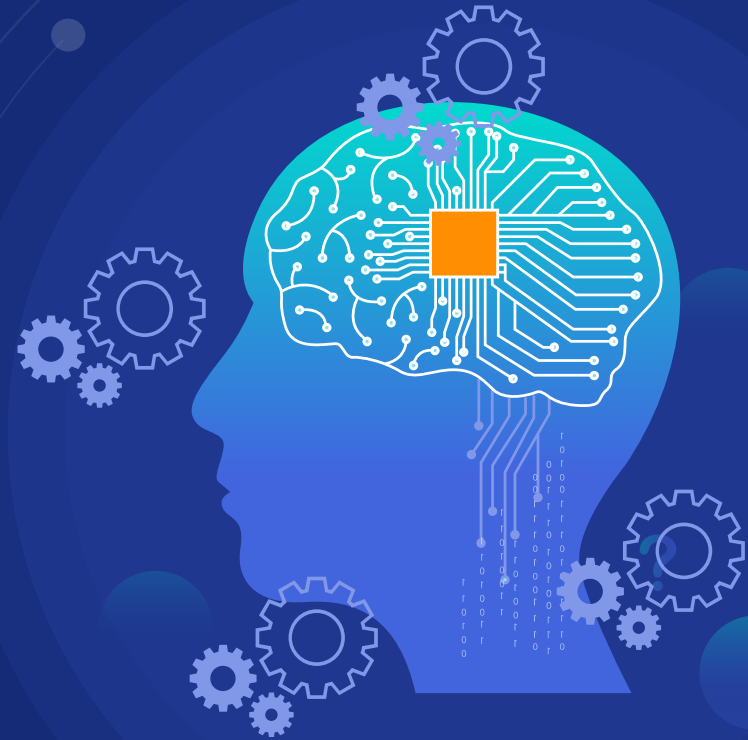
Deep Learning

»»»» Operation expenditure can be reduced >20%

»»»» With reduction of OPEx , workforce can be increase resulting in 70% more output

To put the value of predictive maintenance in context, SAP estimates globally that a two per cent saving in maintenance costs across the top 40 miners would yield a AUD \$18 billion (USD \$13.4 billion) saving to the mining sector.

Source: <https://www.cio.com/article/3625829/how-tinyml-is-powering-big-ideas-across-critical-industries.html>



Rise of Industry 4.0



The key to Industry 4.0 is adoption of AI-ML

USD 101 billion
Y'20.



CAGR-16.4%



USD 337 billion
Y'26.

Top Two AI driven Use-Cases



Industry 4.0: ~50% of Industry4.0 revenue is expected to come from the application with the lowest hanging fruit.



Intelligent Maintenance

- Sound/Vibration .. sensor based early
- Supervised/Unsupervised Learning
- Retrofit: Non intrusive
- OEM: Integrated intelligent maintenance feature
- >90% accuracy even in domain shift
- Extreme low cost solution



Product Quality Control

- Image based
- Supervised/Unsupervised learning
- Low latency
- >90% accuracy even in domain shift
- Extremely low cost solution



Rise of AI Market due to Industry 4.0



AI Market: The key to unlocking a AI multi-billion dollar market is being able to deploy machine learning applications at the edge for manufacturing

USD 1 billion in Y'19.



CAGR-39.7%



USD 27 billion in Y'27.

Intelligent Maintenance and Quality Inspection:
90%(USD 25 billion) of AI in manufacturing will be captured by this segment



Component of AI and Its Share



Industry 4.0: ~50% of Industry4.0 revenue is expected to come from the application with the lowest hanging fruit.



Hardware

- 32bit MCU accounts of 40% of global MCU sell.
- 2.5 billion low cost AI chip will be shipped by 2030
- Requirement: Low latency, Low Power
- Spec: Upto 256KB RAM, <200MHz CPU, No external memory with extremely well fed compute unit using intelligent data pipeline

**Intelligent Main/
Qual. Inspection
~USD 25 billion**



Software

- Preferred DL due to high accuracy
- Extremely low false alarm rate
- Should work under domain shift
- Low compute and memory requirement without sacrificing accuracy
- Homogeneous data for efficient compute
- Unsupervised learning preferred due to lack of anomalous data

Rise of Mighty TinyML



TinyML, the science and art of machine learning, is seeing significant growth as edge devices begin to harness ML models that are both cheap and accurate.



Hardware

- 32bit MCU accounts of 40% of global MCU sell.
- 2.5 billion low cost AI chip will be shipped by 2030
- Requirement: Low latency, Low Power
- Spec: Upto 256KB RAM, <200MHz CPU, No external memory with extremely well fed compute unit using intelligent data pipeline

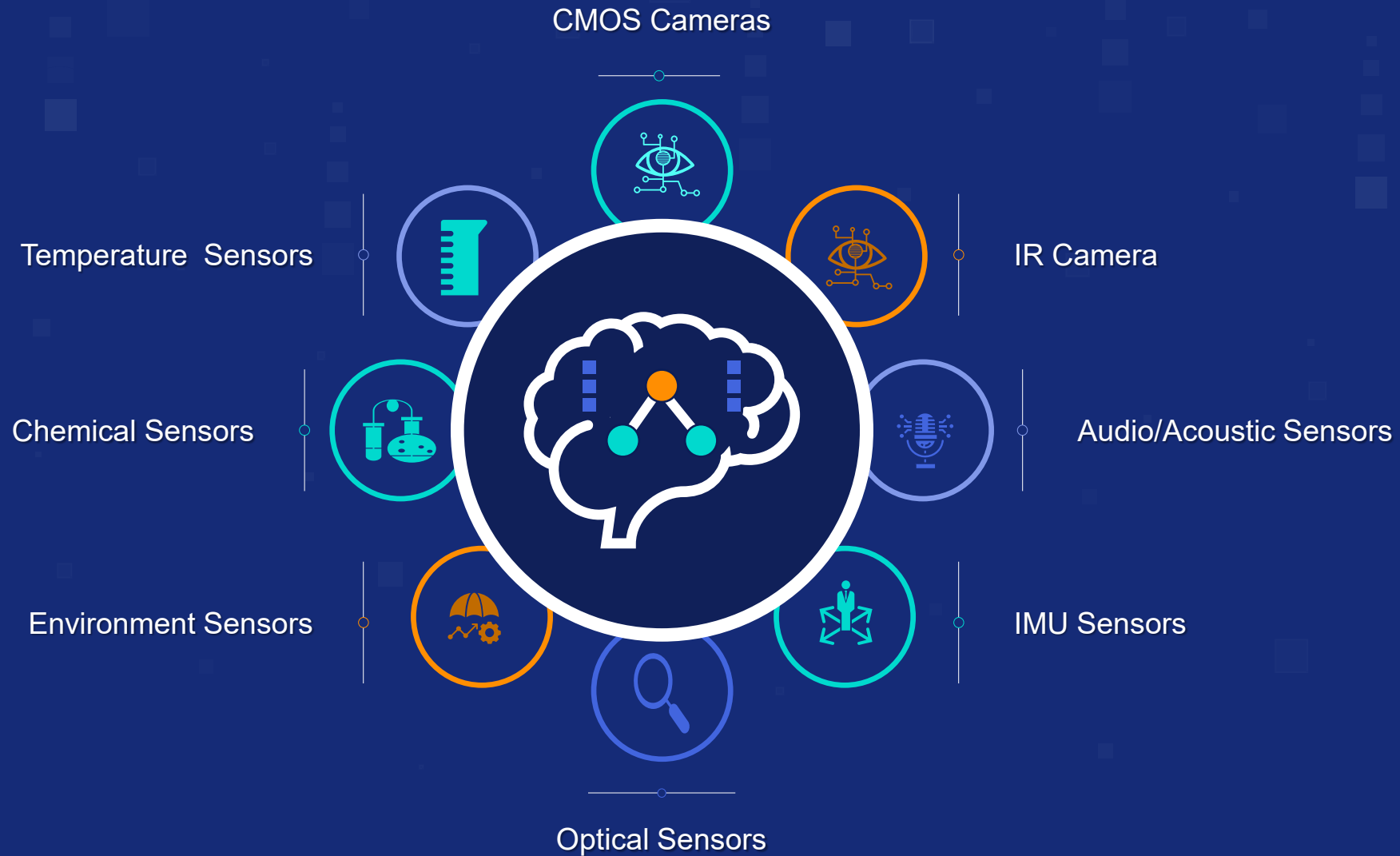


Software

- Preferred DL due to high accuracy
- Extremely low false alarm rate
- Should work under domain shift
- Low compute and memory requirement without sacrificing accuracy
- Homogeneous data for efficient compute
- Unsupervised learning preferred due to lack of anomalous data



Endless Application of TinyML



Challenges of TinyML in fields deployment

○ Robustness/Reliable

1. Extremely low false alarm rate
2. Robust across domain shift/data shift

○ Skill set and Infrastructure

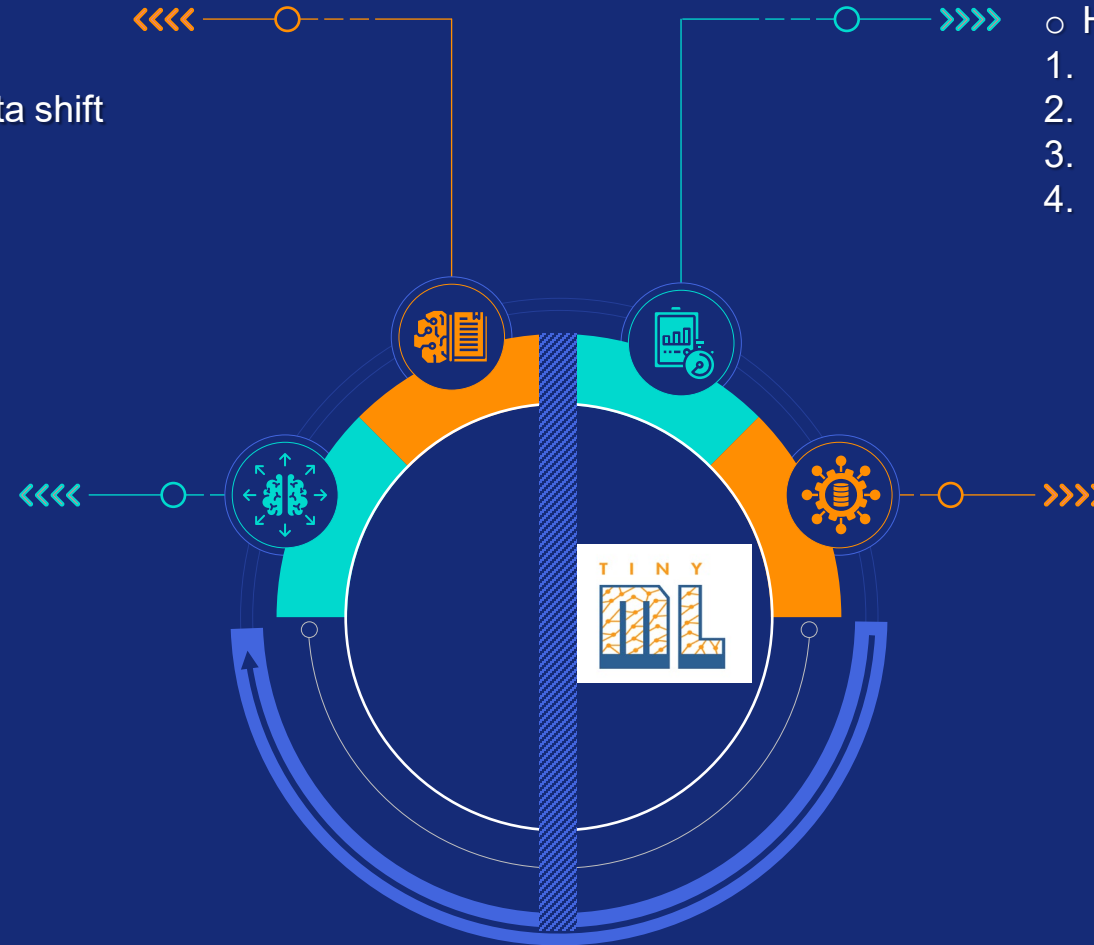
1. Lack of domain expertise
2. Turn around time for deployment/maintenance should be low
3. Lack of dataset. Specifically anomalous dataset

○ Hardware

1. Low cost <USD 1-USD 2
2. Low Power
3. High throughput
4. Reusability

○ Algorithm

1. Low compute Low Ram
2. High Accuracy
3. Low latency
4. Supervised and/or Unsupervised



Challenges

On-Field Challenges

Intelligent Maintenance and Quality Inspection:
USD 25 Billion Potential Revenue

Challenges of field deployment

Ease of Use



1. Mass User don't have the required skill set to tackle machine learning problems
2. An end to end flow need to be deployed

Robust/Reliable AI application



1. DNN has structural defect
2. Fails miserably and silently in domain/data shift
3. This makes high false alarm rate
4. Need a system where DNN can distinguish domain/data shift
5. Acceptability will be high with this system

Lack of Dataset



1. Anomaly dataset are rare events and can come at any shape and size
2. Unsupervised learning without anomaly data for intelligent maintenance and quality inspection

Challenge#1:

The demand for End to End, Easily Deployable and Custom DL solution



Dedicated team of AI experts is required. Factory owners are reluctant to create the AI team



Too much of complexity in creating AI based applications. Which algorithm to use, which framework, DL vs ML, training, overfitting ...



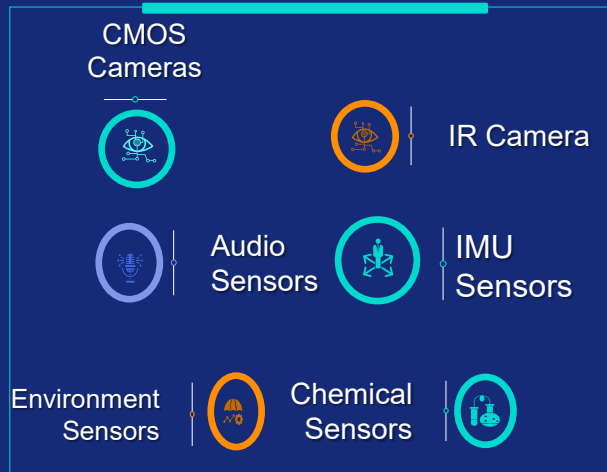
Creating a hand-crafted model to fit in constrained hardware is extremely time consuming job and the model created is not generic, reusable, and scalable. Every new application requires similar efforts



The industries owner calls for solutions that are scalable and reusable, preferably on-premises and built around inexpensive off-the-shelf infrastructure, such as a basic i7-based desktop with simply a push button end-to-end application oriented model

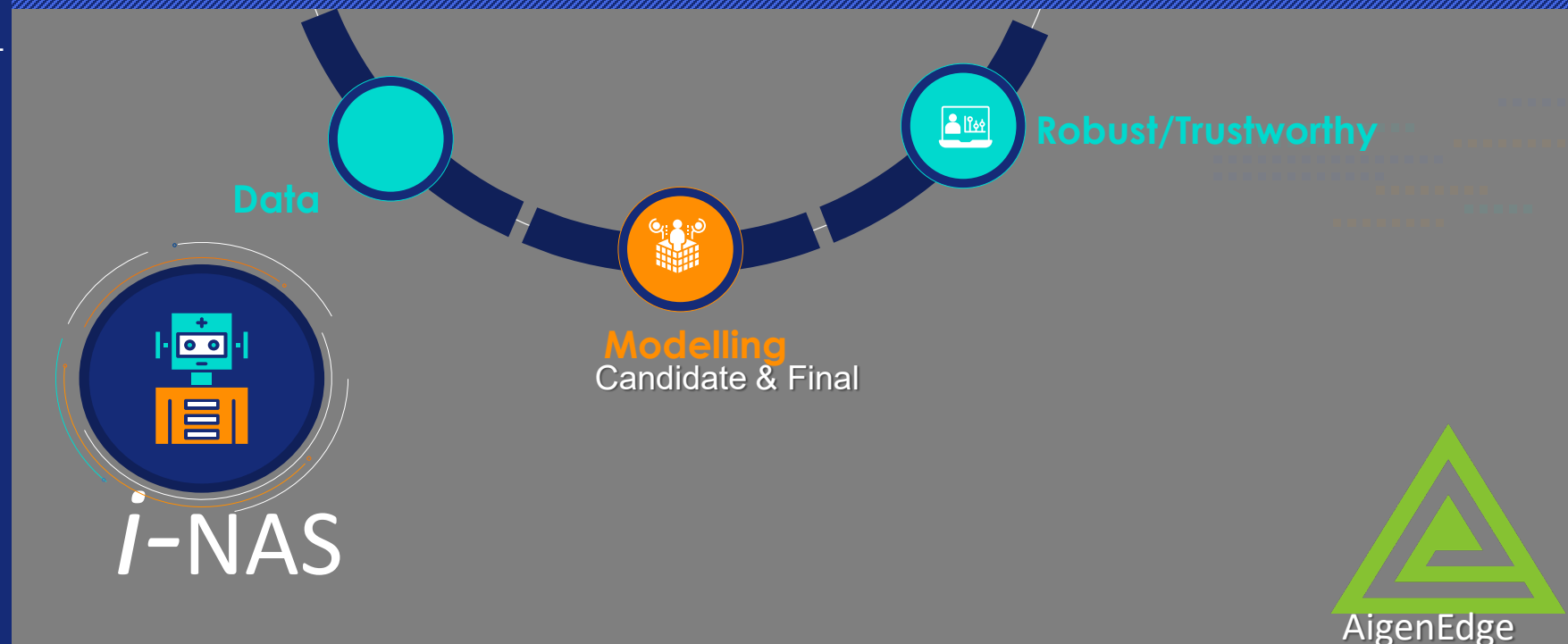
Solution# 1: Ease of Use

AigenEdge's Iota-NAS: Push Button NO GPU AutoML

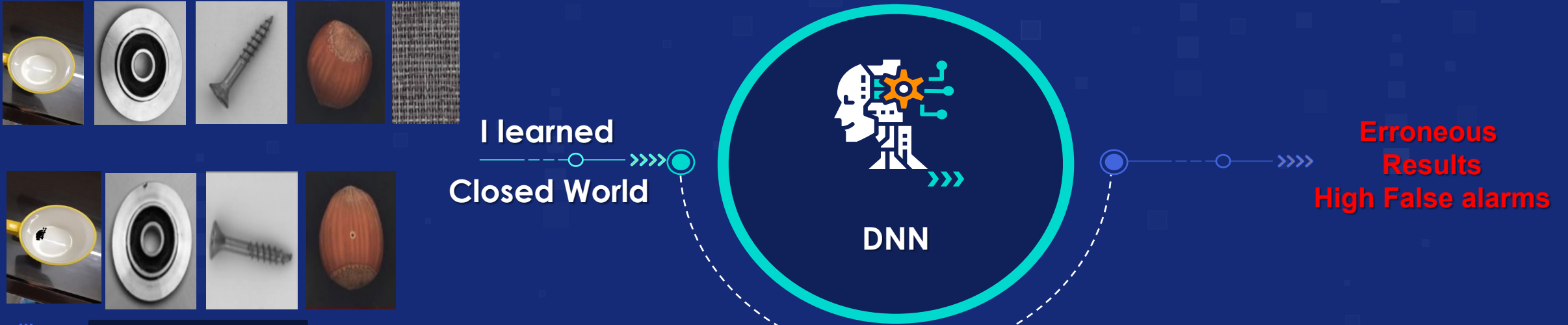


»»» No-AI-Expertise, Application oriented, scalable and reusable AutoML framework

»»» Smallest, fastest, lowest power yet highly accurate, robust/reliable /trustworthy Deep Learning based application adhering to the hardware specifications



Challenge# 2: Unreliable/Untrustworthy- Supervised Classification based Quality Inspection

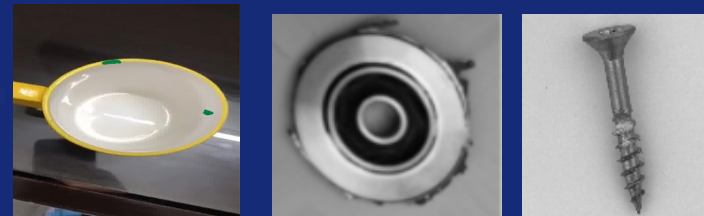


I learned
Closed World



Erroneous
Results
High False alarms

What-if data shown
is not inside the training set



- »»» Deep Neural Network assume closed world
- »»» Works exceedingly well in this assumption
- »»» But in reality this assumption is wrong
- »»» The world is open world
- »»» In open world DNN fails silently and confidently
- »»» This makes the real-life usage of DNN limited

Challenge# 2:

Unreliable/Untrustworthy- Supervised Classification based Acoustic Intelligent Maintenance

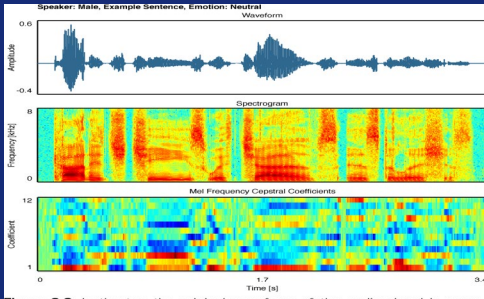
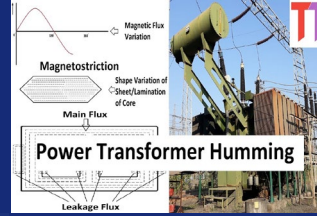
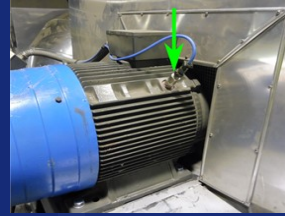


Figure 3.2 In the top, the original waveform of the audio signal is represented

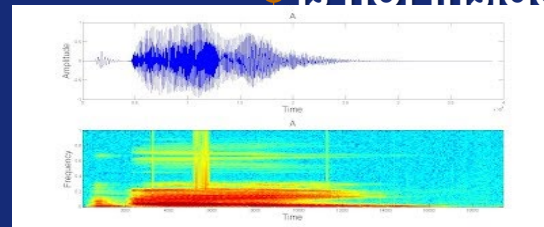
I learned
Closed World



DNN

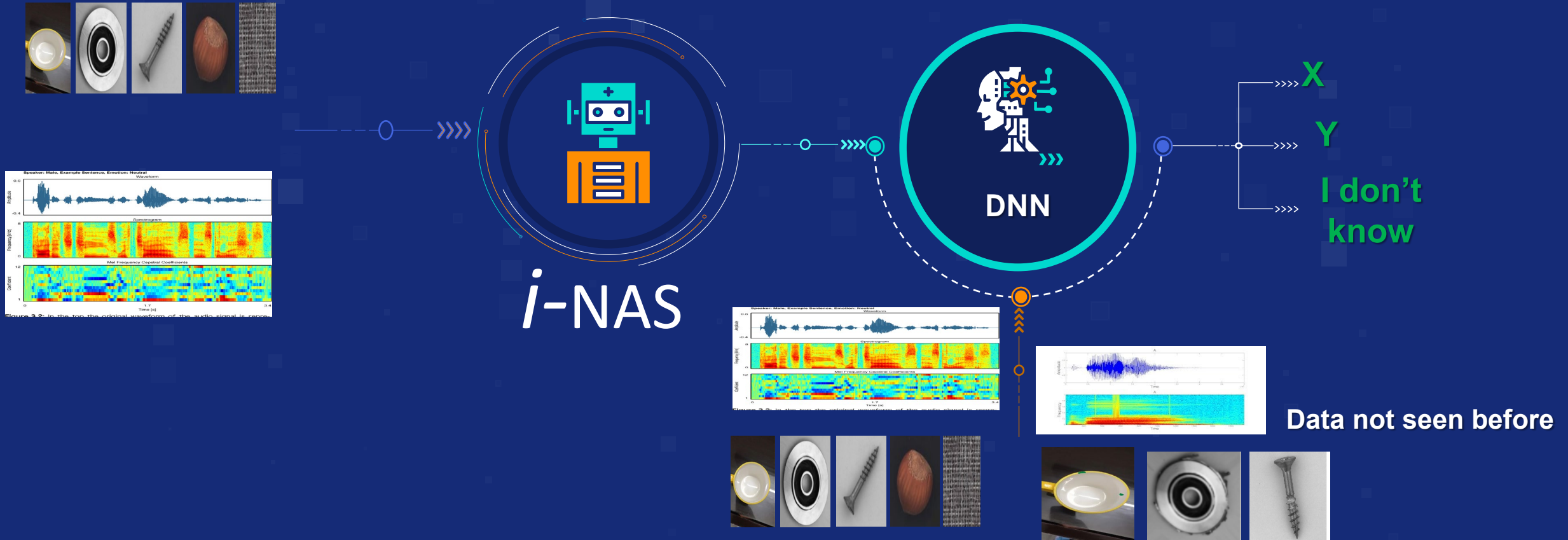
Erroneous
Results
High False Alarms

What-if data shown
is not inside the training set



Challenge# 2: Solution

AigenEdge Reliable and Trustworthy DNN



AutoIDK can take any neural network and automatically converts the DNN with “I don't know” addition label

AutoIDK don't require any additional dataset

Reliable and Trustworthy DNN for Quality Inspection under 1\$ MCU



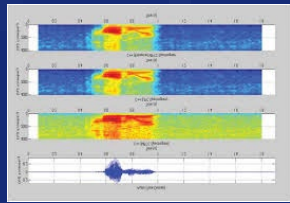
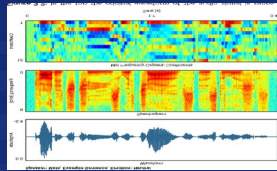
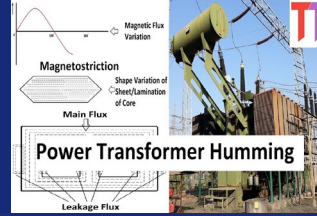
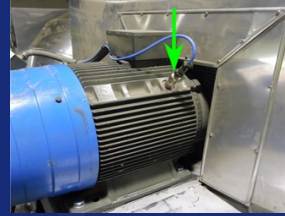
- AigenEdge IDK(I Don't Know) output makes DNN robust/trustworthy with much lesser false alarm rate

What-if data shown is not inside the training set

MACC	Peak Memory(INT8)	Parameter(INT8)	Accuracy(INT8)
1M	55KB	17KB	98%



Reliable/Trustworthy DNN: Audio based Intelligent Maintenance under 1\$ MCU



I learned
Closed World

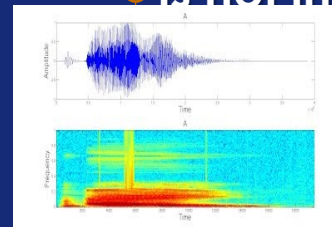


Good

Need Maintenance

I don't know

What-if data shown
is not inside the training set



- AigenEdge IDK(I Don't Know) output makes DNN robust/trustworthy with much lesser false alarm rate

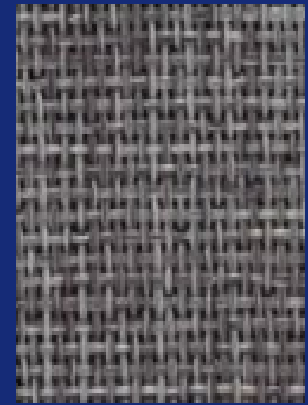
MACC	Peak Memory (INT8)	Parameter (INT8)	Accuracy (INT8)
1M	55KB	17KB	98%

Challenge# 3: Lack of Dataset for Quality Inspection and Intelligent Maintenance

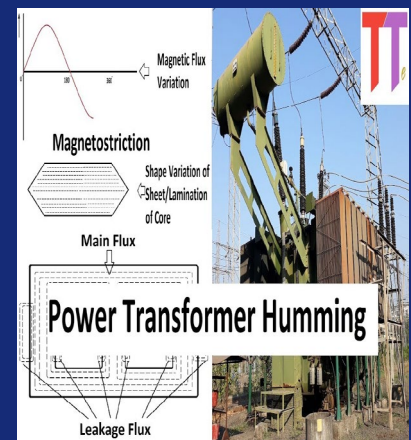
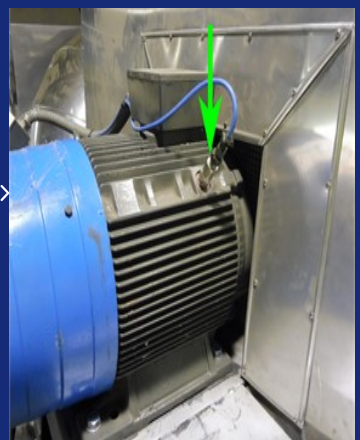
Anomalous Data are rare event and like finding needle in haystack
Don't expect industries to provide these rare data, it doesn't exist



Good Data

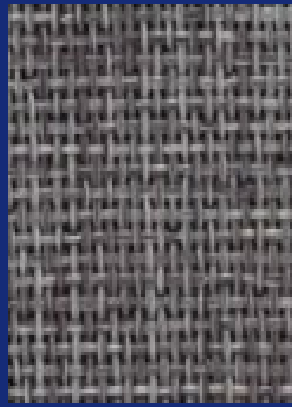


Good Data

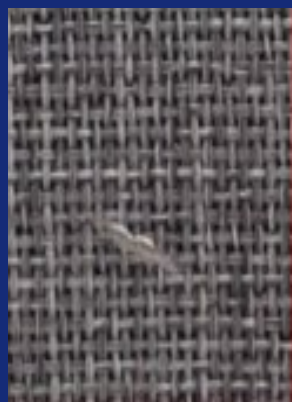
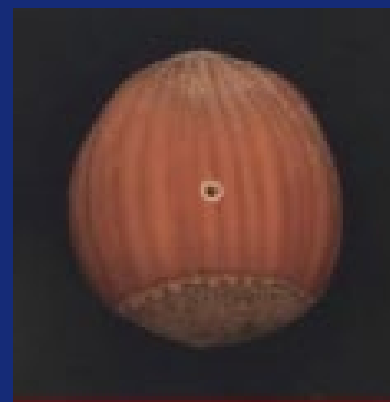
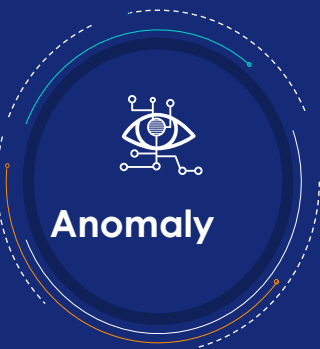


Challenge# 3: Lack of Dataset for Quality Inspection and Intelligent Maintenance

In absence of Anomalous Data, the job is still to detect anomaly



Good Data



Anomaly

Challenge# 3: Solution

Good Data based Unsupervised Deep Learning Solutions



- AigenEdge TinyNAS based AutoML generates state-of-art Unsupervised Deep Learning network
- Highly resilient towards domain shift and data shift

Audio/Acoustic based Quality Inspection under 1\$ MCU: Only Good Dataset

MACC	Peak Memory(INT8)	Parameter(INT8)	Accuracy(INT8)
5M	200KB	50KB	95%

Robust towards domain/data shift such as sudden changes in factory environment, noise ...

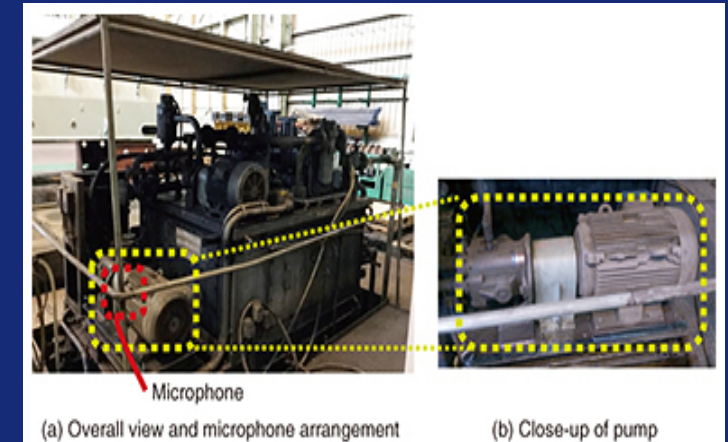
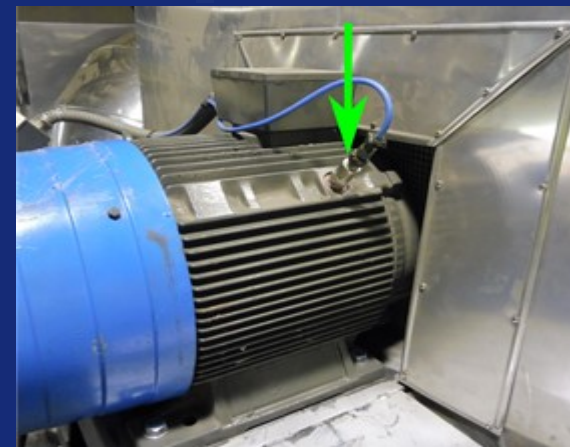
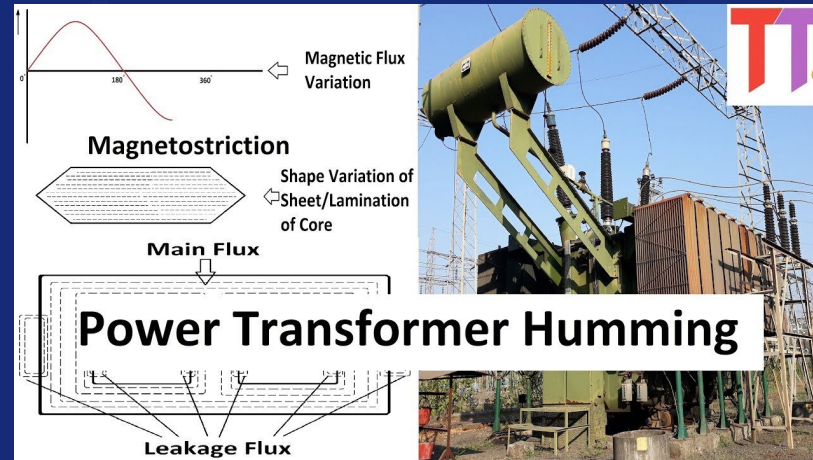
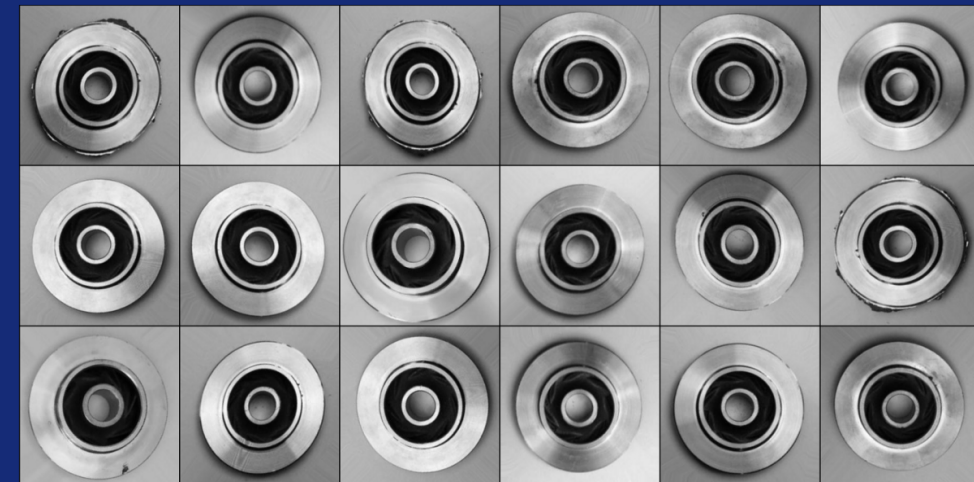
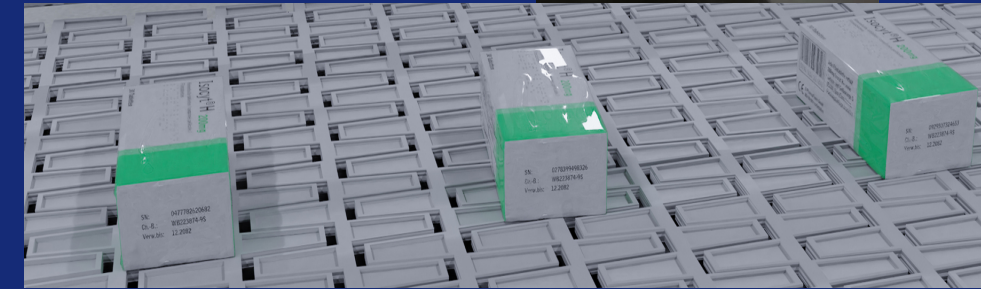
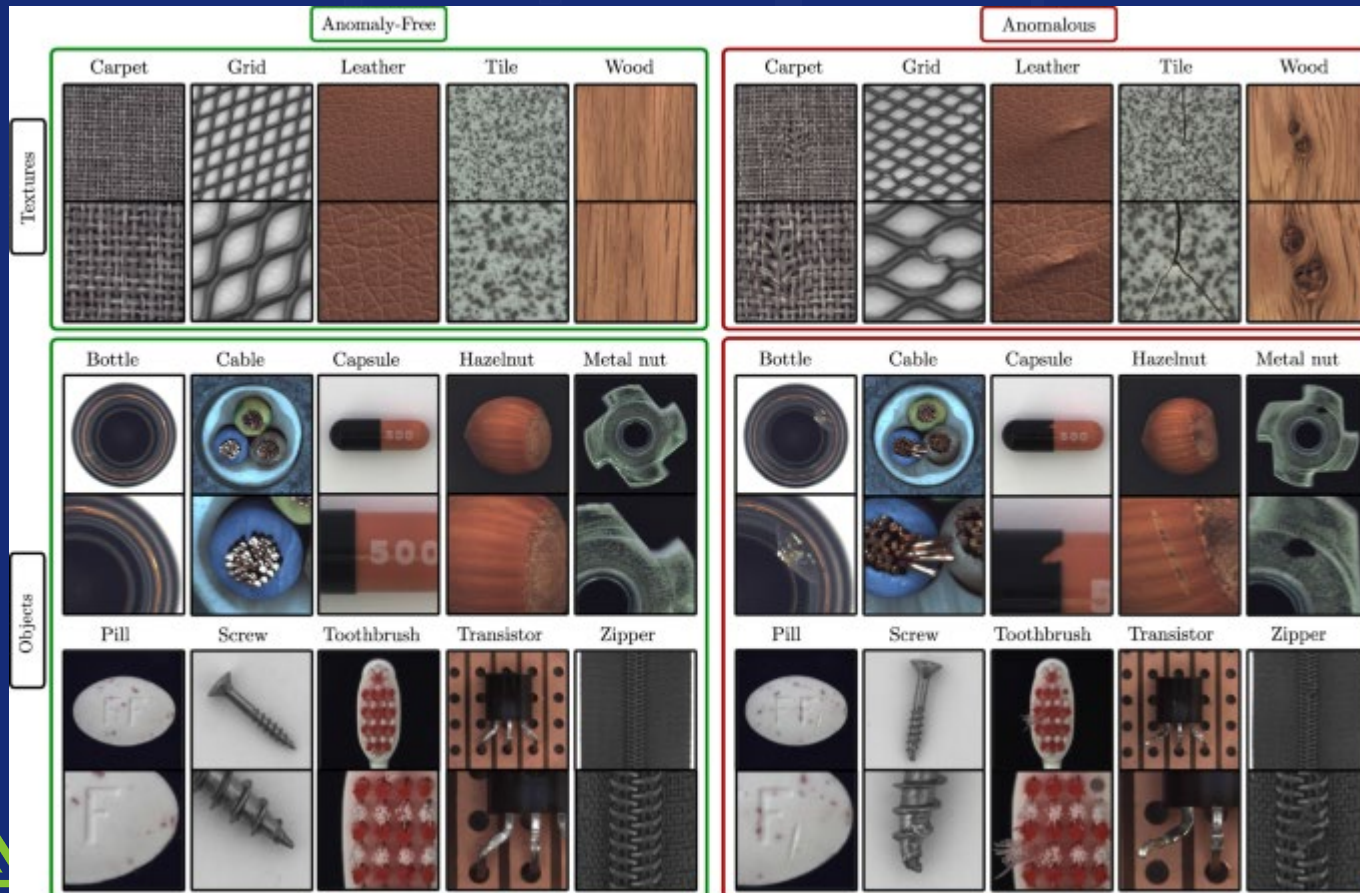


Image based Quality Inspection under 1\$ MCU: Only Good Dataset

MACC	Peak Memory(INT8)	Parameter(INT8)	Accuracy(INT8)
5M	200KB	50KB	95%



AigenEdge-iota-NAS






Automatic Neural Architecture Search for DNN based AI application Everywhere and Anywhere



Advantages



AigenEdge

-  Hardware and Application Oriented AutoNAS with NO GPU search time max 12hrs
-  Search and Generate DL applications 5x smaller than the State-Of-Art DL network such as MobileNetV2/3, Efficientnet, FDMobilenet ... without losing any accuracy from baseline
-  User need to provide just raw data and choose the hardware specification. Rest leave it to our AutoNAS
-  Can generate both supervised and unsupervised DL applications. In case of unsupervised algorithm only good data is needed.
-  In supervised DL application

Advantages of AigenEdge's Iota-NAS Platform



All the application is DL based with more than 90% accuracy, on sub \$ 1 MCU with real time performance

supervised

Unsupervised

Images

Other sensors(Audio)

Images

Other sensors(Audio)

- VWW/Person-No Person
- Quality Inspection
- Fire/Smoke detection
- Digital Pathology
- Medical diagnostic
- Driver Monitoring
- Digital cockpit

- Audio based intelligent maintenance
- Vibration based intelligent maintenance
- Activity recognition

- Anomaly detection such as quality inspection

- Sound and vibration based intelligent maintenance



Use-Case under 1\$ device: Image Based Classification- Supervised

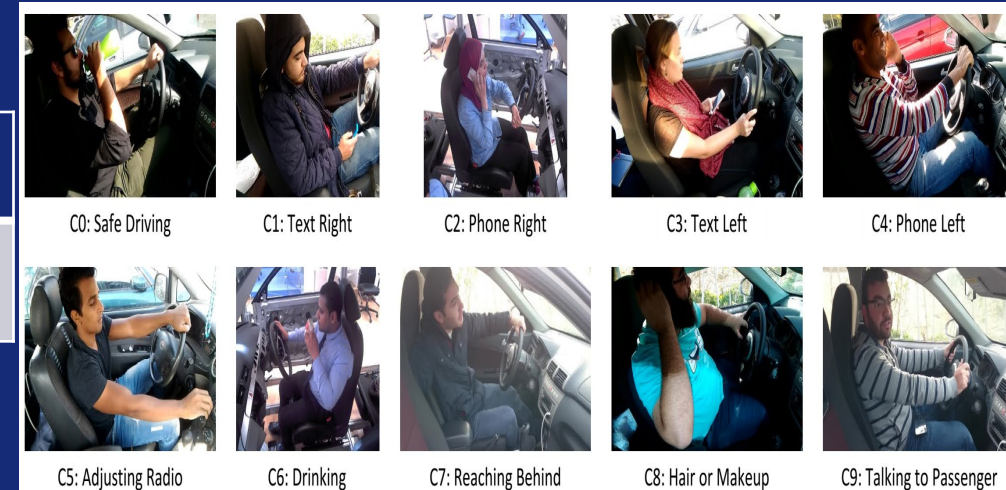
Visual Wake Word- Person-No Person classification

Model	MACC	Peak Memory(INT8)	Parameter(INT8)	Accuracy(INT8)
Model#1	3.3M	55KB	40KB	85%
Model#2	4.2M	95KB	38KB	87%
Model#3	5.8M	118KB	90KB	88.3%



Driver Distraction Monitoring System

Model	MACC	Peak Memory(INT8)	Parameter(INT8)	Accuracy(INT8)
Model	5M	226K	17K	98%



Summary of AigenEdge-Iota-NAS Platform



Easy to Develop and Modify the System

Fast Response



Low Cost

**DL model:
High accuracy**



Robust/Reliable System

Post It Notes



Booming AI

One of the fastest growth rate.



AI Assisted Industry 4.0

Industry 4.0 only possible with AI



TinyML Enabler

AigenEdge's Iota-NAS is well poised to enable this AI-ML market

**We collaborate with ambitious brands and people;
let's build something great together.**

- >>> **First level : info@aigenedge.com**
- >>> **Second Level : aroy@aigenedge.com**
- >>> **<https://www.aigenedge.com/>**



Copyright Notice

This multimedia file is copyright © 2021 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML[®] is a registered trademark of the tinyML Foundation.

www.tinyml.org



Copyright Notice

This presentation in this publication was presented as a tinyML[®] Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org