tinyML® Summit

Miniature dreams can come true…

March 28-30, 2022 | San Francisco Bay Area

TINY ML

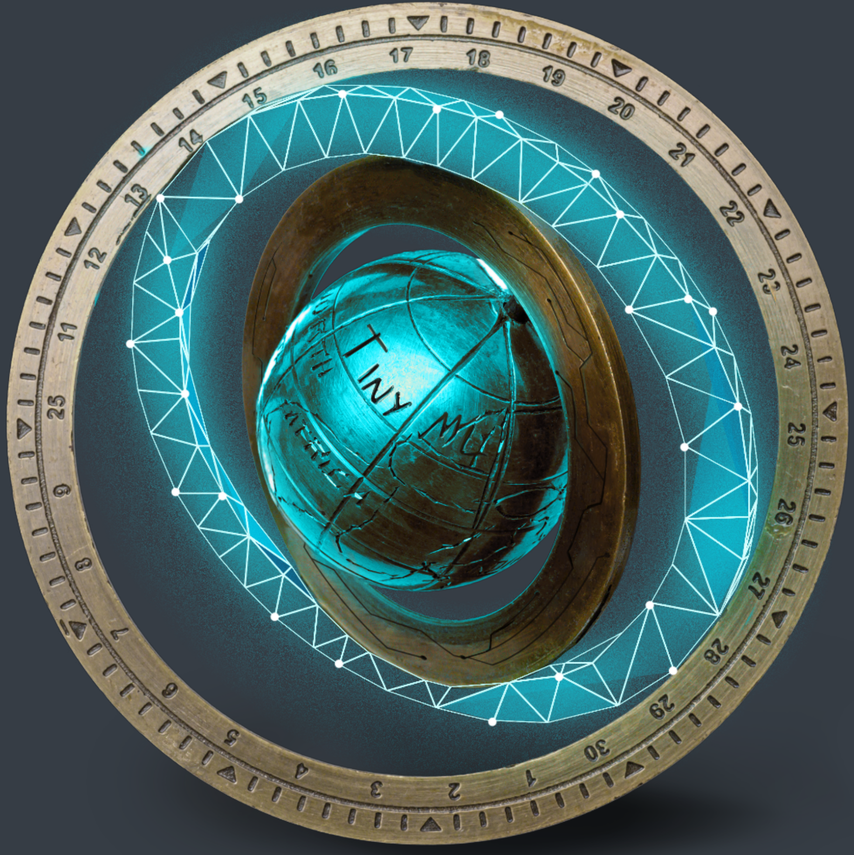
www.tinyML.org

Intelligent Agent

Neuton

NEUTON.AI

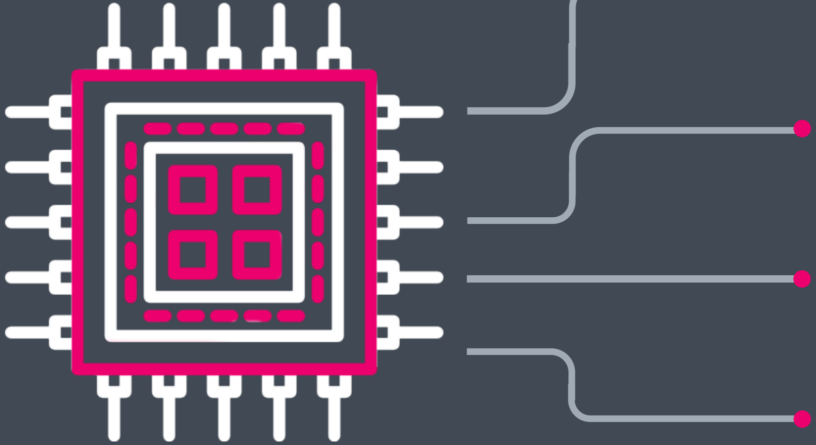
A Novel Approach to Building Exceptionally Tiny Models without Loss of Accuracy



Tiny ML Defined – 2019

Dr. Evgeni Gousev
Qualcomm

Pete Warden
Google



Total memory - often **< 100 kB**

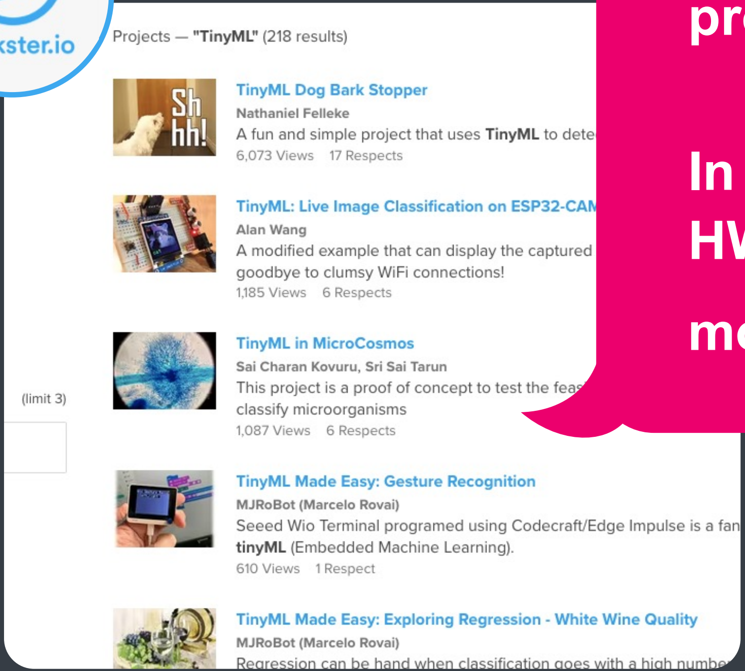
Energy - µW scale,
battery to last for years

Processor –
10s - 100s MHz, at most

Cost - very low cost to enable
massive deployment

Intelligent Agent
Neuton

TinyML projects – What do we see today



Projects — "TinyML" (218 results)

(limit 3)

TinyML Dog Bark Stopper
Nathaniel Felleke
A fun and simple project that uses **TinyML** to dete
6,073 Views 17 Respects

TinyML: Live Image Classification on ESP32-CAM
Alan Wang
A modified example that can display the captured
goodbye to clumsy WiFi connections!
1,185 Views 6 Respects

TinyML in MicroCosmos
Sai Charan Kovuru, Sri Sai Tarun
This project is a proof of concept to test the feas
classify microorganisms
1,087 Views 6 Respects

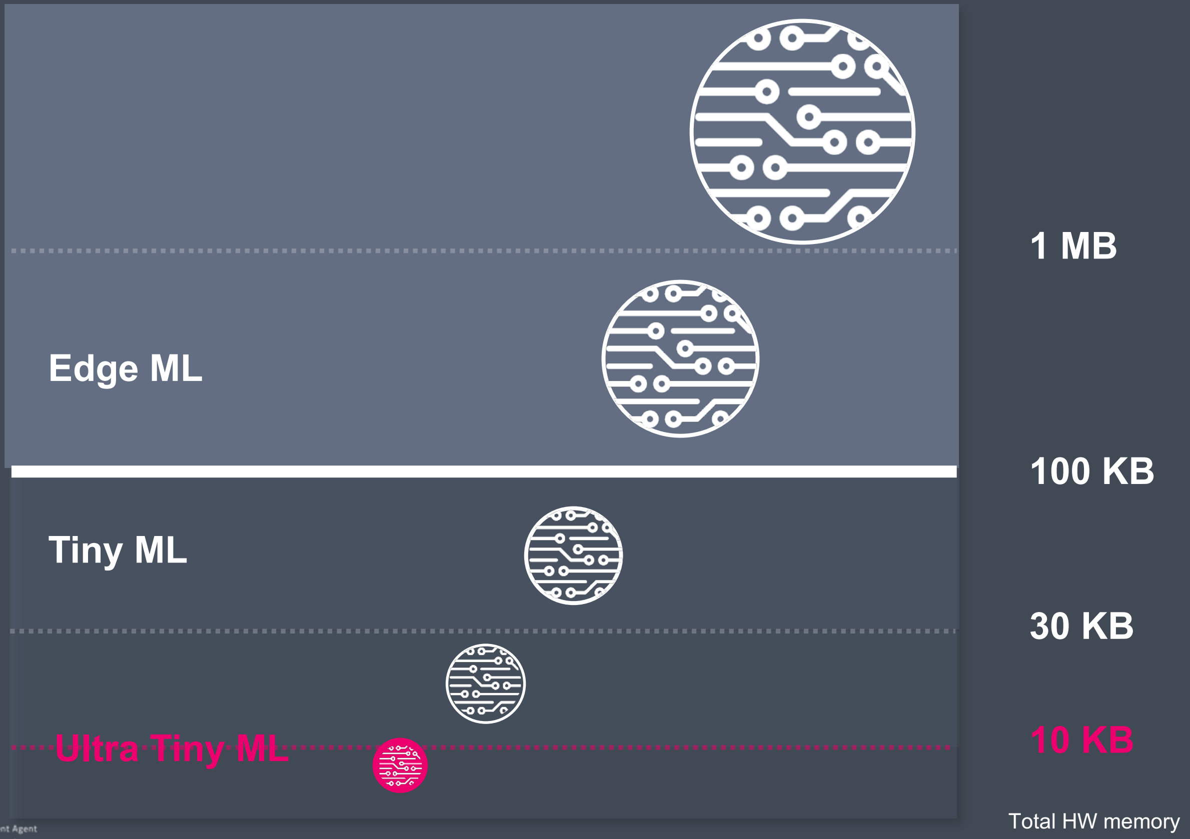
TinyML Made Easy: Gesture Recognition
MJRoBot (Marcelo Rovai)
Seeed Wio Terminal programed using Codecraft/Edge Impulse is a fan
tinyML (Embedded Machine Learning).
610 Views 1 Respect

TinyML Made Easy: Exploring Regression - White Wine Quality
MJRoBot (Marcelo Rovai)
Regression can be hand when classification goes with a high numbe

There are 218 'TinyML' projects on hackster.io

In 96% of cases are used HW with a total memory of more than 100 KB

Where are you in TinyML journey?



Edge ML

Tiny ML

Ultra Tiny ML

1 MB

100 KB

30 KB

10 KB

Total HW memory

96% of todays cases

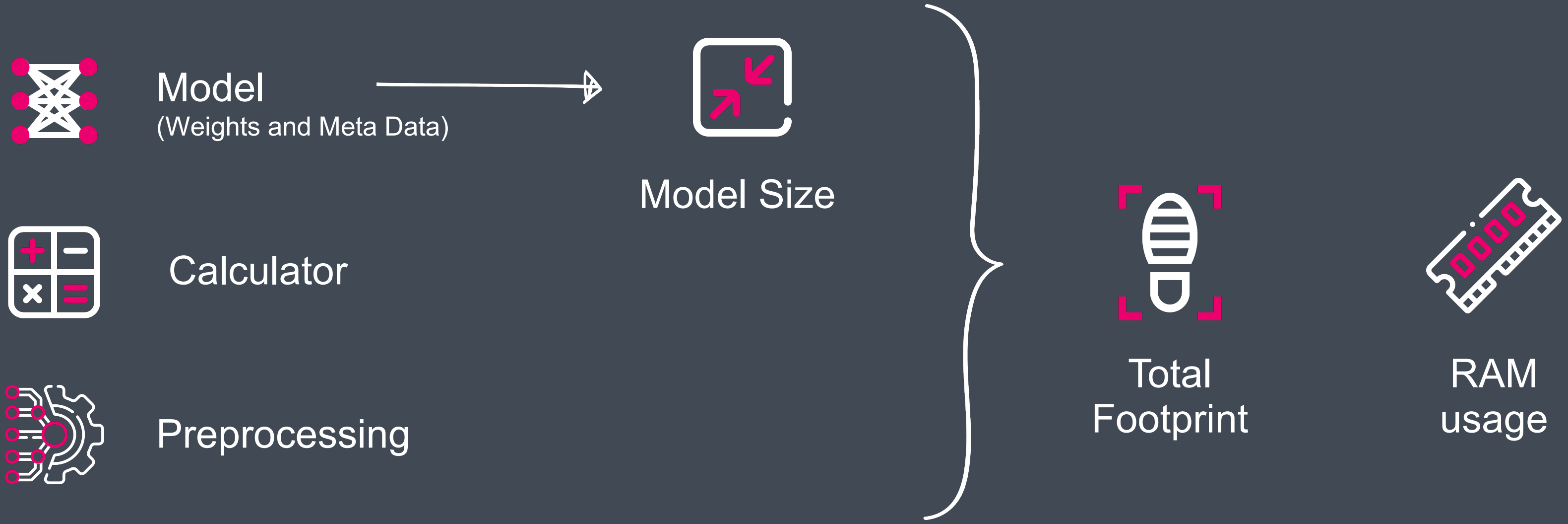
4% really TinyML cases

New opportunities!

Intelligent Agent Neuton

Moving TinyML Forward!

Embedded model consideration



Model
(Weights and Meta Data)

Model Size

Calculator

Preprocessing

Total Footprint

RAM usage

Moving TinyML Forward!

10 kB	Total memory for HW
< 5 kB	The Ideal Weight for Total Footprint
< 1 kB	The Ideal Weight for a TinyML Models

One is not enough!

BEST METRIC

There are many Neural Architecture Search methods, Auto ML tools and Frameworks (TensorFlow, Keras and PyTorch).
However, most of them are focused on finding the **best metric.**

MINIMAL SIZE

There are many technics reducing size of a model: quantization, pruning, nor distillation. All of them effect to the accuracy.

BEST METRIC + MINIMAL SIZE

While TinyML tasks require building models with **best metric and minimal size**

Taking the next step!

Neuton – The First Neural Network Framework that empowers you to build models with minimal size and without loss of accuracy

- automatically
- in one iteration
- without compression

No Model Size & Quality Trade Off

Neuton's models are extremely compact:

up to

1000

times

- Fewer coefficients and neurons
- Smaller in size (Kb)
- Faster inference

in comparison to TensorFlow and other algorithms

- No compression techniques (quantization, pruning, etc.)
- Accuracy is not affected

Small scale – huge opportunities!

If your model is **1 KB** your

8, 16, 32, 64 bit HW can:

Have many models in one MCU

Embed model into really tiny pieces of HW:

- sensors
- 8, 16 bit MCUs
- ASICs

Spend less energy on calculation

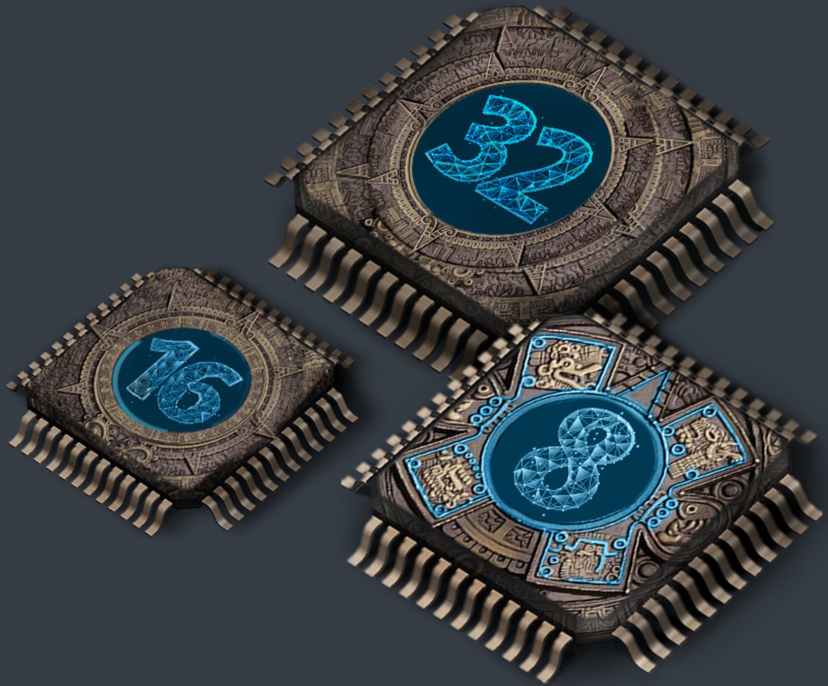
Have more business logic in one MCU

Intelligent Agent Neuton

Bring Intelligence to the tiniest MCUs

Even 8-bit MCU can now be AI Driven

Bit depth	Neuton	TensorFlow
8-bit	✓	✗
16-bit	✓	✗
32-bit	✓	✓



Neuton vs. TensorFlow Lite Benchmarks

		NEUTON (8 bit)				TENSORFLOW LITE (Quantized)				
DATASET	METRIC	METRIC VALUE	MODEL SIZE, KB Metadata + Weights in Flash Memory	TOTAL FOOTPRINT, KB Model Size+ Calculator + Preprocessing in Flash Memory	RAM USAGE, KB Preprocessing + Calculator	METRIC VALUE	MODEL SIZE, KB Model in Flash Memory	TOTAL FOOTPRINT, KB Model Size + Interpreter + Preprocessing in Flash Memory	RAM USAGE, KB Preprocessing + Interpreter	NEUTON`S MODEL IS IN X TIMES SMALLER
Abnormal Heartbeat Detection	AUC	0,98	2,56	3,73	0,8	0,97	14,22	166,19	6,7	5,6
Hole Drilling Deviation Prediction	Accuracy	0,98	0,21	1,38	0,06	0,96	18,5	170,47	7,42	88,1
Air Pressure System Failures	Accuracy	0,99	1,6	2,77	0,7	0,97	10,66	162,63	6,88	6,7
Detection of storage condition violations	AUC	0,95	0,13	2,17	0,03	0,93	4,9	156,88	6,88	37,7
IoT based Gesture Recognition	Accuracy	0,99	5,03	15,2	5,4	0,97	97,06	249,04	11,33	19,3
Food Quality Monitoring	Accuracy	0,99	0,1	1,27	0,04	0,98	3,47	155,44	6,37	34,7
Air Quality Prediction	MAE	0,21	0,16	1,2	0,05	0,22	7,14	159,11	6,83	44,6
Energy Output Definition	MAE	3,23	0,33	1,37	0,04	3,35	4,88	156,91	6,58	14,8
Electric Grid Prediction	Accuracy	0,93	0,66	1,84	0,09	0,93	3,72	155,69	6,51	5,6
Room Occupancy Detection	Accuracy	0,98	0,18	1,36	0,04	0,97	10,72	162,69	6,73	59,6
MNIST	Accuracy	0,94	13,33	14,51	3,38	0,91	17,39	169,36	9,87	1,3
Gearbox Fault Diagnosis	Accuracy	0,92	1,93	12,52	2,52	0,91	30,75	186,19	9,23	15,9
Air Writing Digits Recognition	Accuracy	0,94	0,86	11,45	2,55	0,93	24,6	179,96	9,13	28,6
"Flex" or "Punch" Recognition	Accuracy	0,97	0,65	7,18	3,07	0,96	4,13	159,76	9,48	6,4
Snowfall prediction	Accuracy	0,88	0,34	1,52	0,05	0,87	2,27	154,23	6,01	6,7

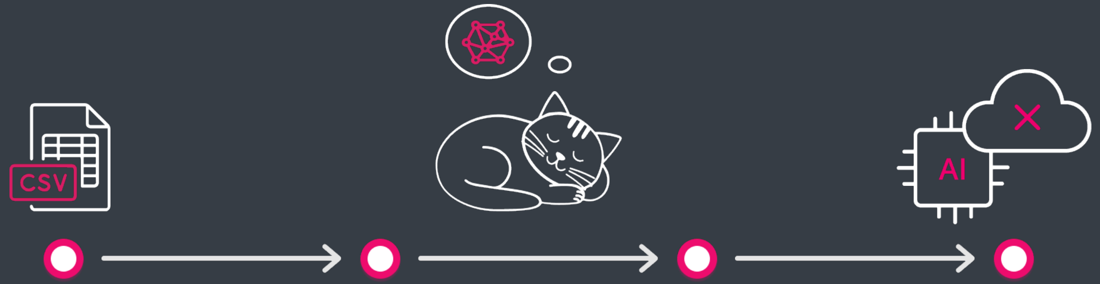
All benchmarks were made on 32-bit MCU (Nordic nRF52840) as TensorFlow Lite for Microcontrollers requires a 32-bit platform. 8-bit post-training quantization was implemented for TF models. Neuton models do not require any compression techniques.

How Do We Create Compact Models without Comprising Accuracy?

- Selective approach to the connected features
- Automatic neuron-by-neuron network structure growth
- No manual search for neural network parameters
- Unique patented global optimization algorithm
- Permanent cross-validation

Neuton as an AutoML

Automatically build extremely tiny models and embed them into any microcontroller



Upload your data

Neuton will automatically build a model of optimal size and accuracy

Download your model and embed it into a microcontroller

Make inferences right on the device

- ☑ No-Code SaaS Solution
- ☑ No Data Science experience required
- ☑ Fully automated pipeline

Bring Intelligence to the sensor edge

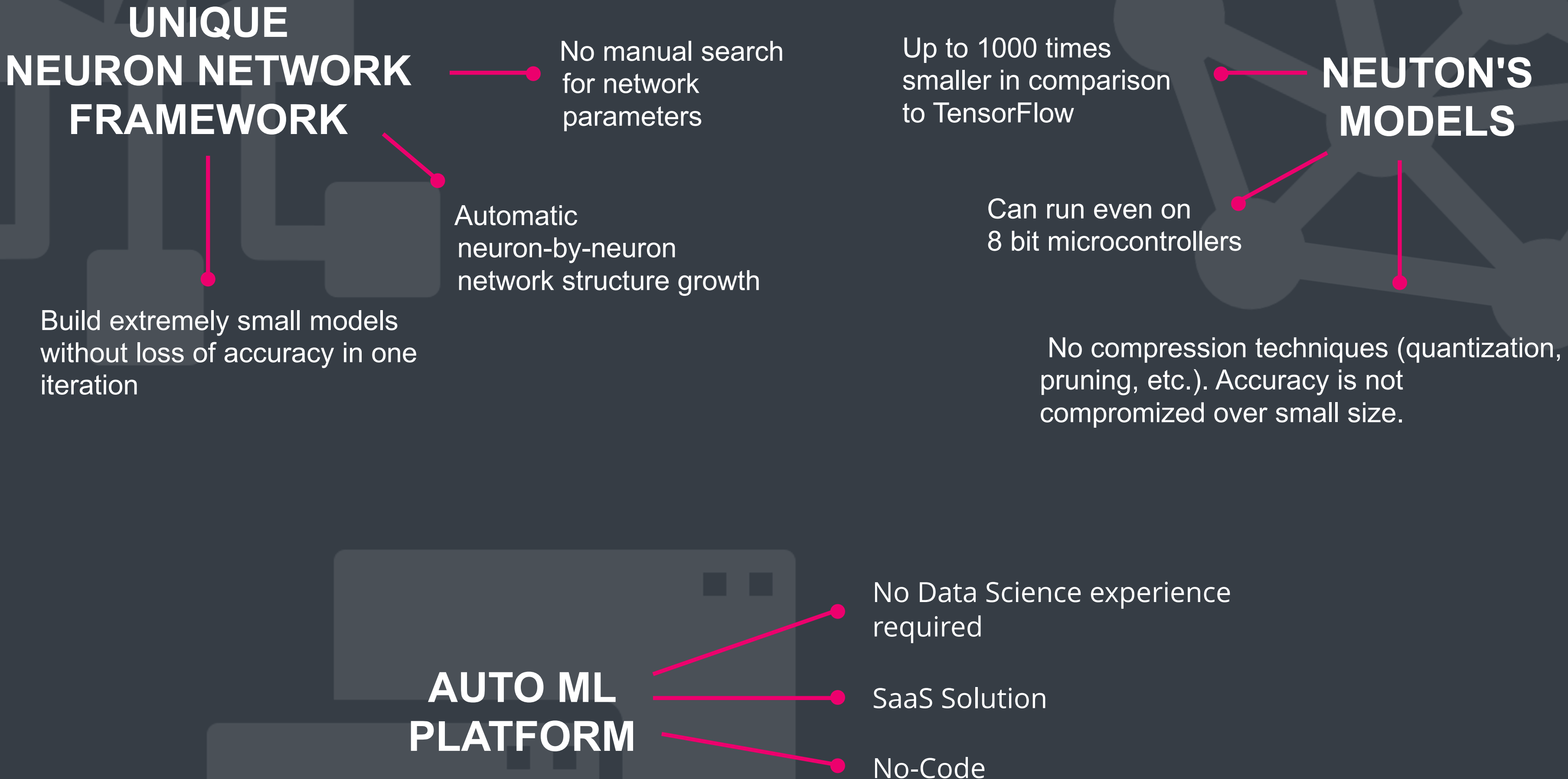
life.augmented

The STM LSM6DSO16IS it supported real-time applications that rely on sensor data.

ISPU (intelligent sensor processing unit) RAM:
32 kb - of program
8 kb - for data

‘Flex’ or ‘punch’ movement recognition based on an accelerometer.

Model Size – 0,65 kB
Total footprint 7,18 kB
RAM usage - 3,07 kB
Accuracy – 97%



UNIQUE NEURON NETWORK FRAMEWORK

No manual search for network parameters

Automatic neuron-by-neuron network structure growth

Build extremely small models without loss of accuracy in one iteration

NEUTON'S MODELS

Up to 1000 times smaller in comparison to TensorFlow

Can run even on 8 bit microcontrollers

No compression techniques (quantization, pruning, etc.). Accuracy is not compromized over small size.

AUTO ML PLATFORM

No Data Science experience required

SaaS Solution

No-Code

Free unlimited plan for developers

$0/mo

Start to build tinyML models today!

https://neuton.ai/start

Intelligent Agent
Neuton

Thank you!

TINY ML SUMMIT

tinyML Summit 2022 Sponsors

ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

AONdevices

arm

ASPINITY

brainchip
The Neuromorphic Computing Company

CEVA®

Deeplite

EDGE IMPULSE

emza
visual sense

FotaHub

GREENWAVES
TECHNOLOGIES

Grovety Inc.

Himax

HOTG

imagimob

Infineon

itemis

KLIKA TECH
GLOBAL IOT SOLUTIONS

LatentAI

LATTICE
SEMICONDUCTOR

Micro.ai

OmniML

NXP

PX:I

Plumerai

PROPHESEE

Qeexo

Qualcomm

Rackner

Reality AI®
Engineering Solutions for the Edge

REEXEN
technology

RENESAS

SAP

seeed studio
The IoT Hardware Enabler

SensiML™

Sony
Semiconductor
Solutions
Corporation

ST
life.augmented

SA STREAM ANALYZE

synaptics®

SynSense

SYNTIANT

Tensil.ai

TensorFlow

XMOS

TINY ML SUMMIT

Copyright Notice

www.tinyml.org