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

Intelligent Agent

Neuton

Neuton.ai

A Novel Approach to Building Exceptionally Tiny Models

Our TinyML Community

TECH RESOURCES

New to Edge Computing / Tiny ML
Experienced Tech Resources

DATA SCIENCE

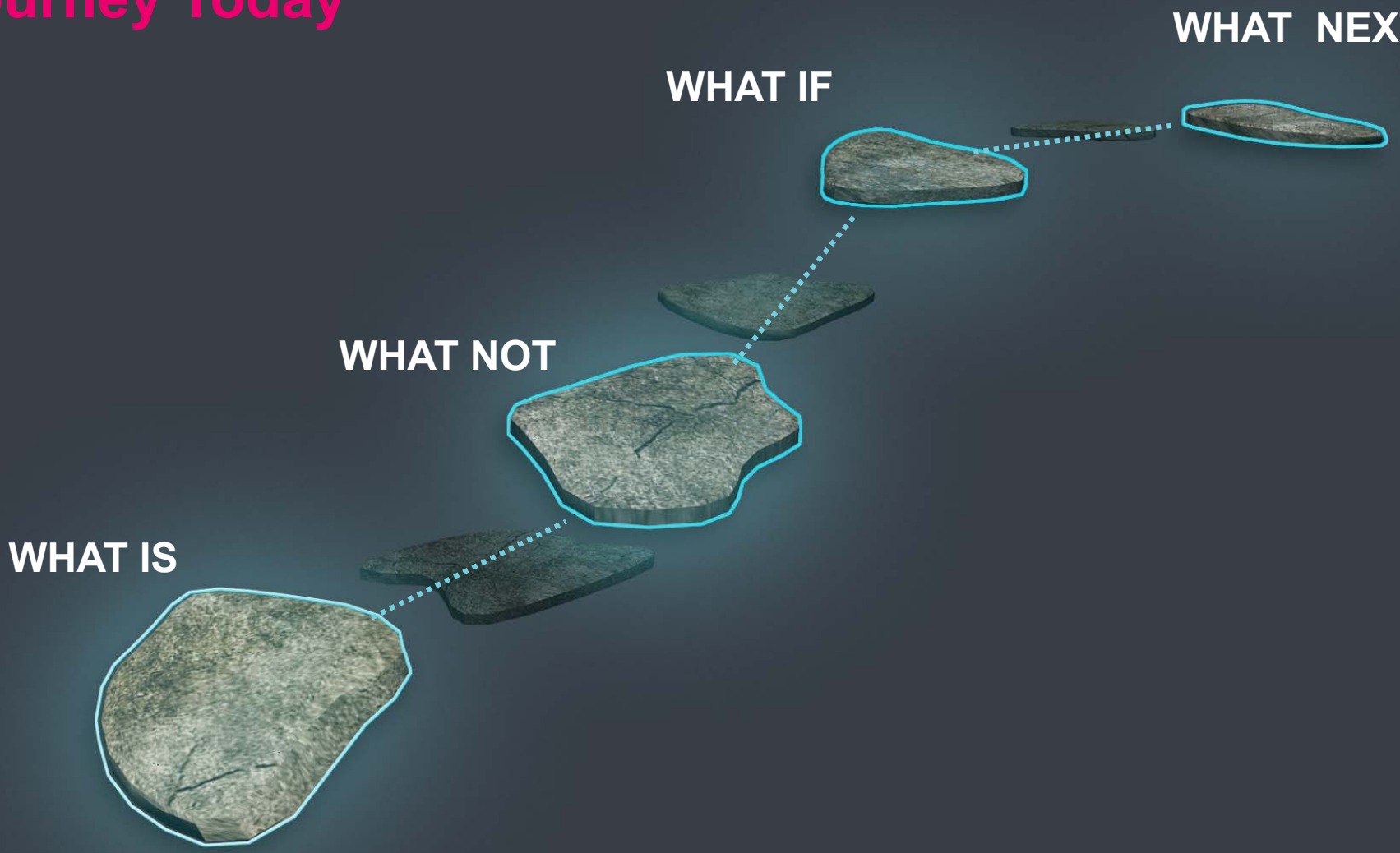
New to AI /ML
Experienced Data Scientists

BUSINESS

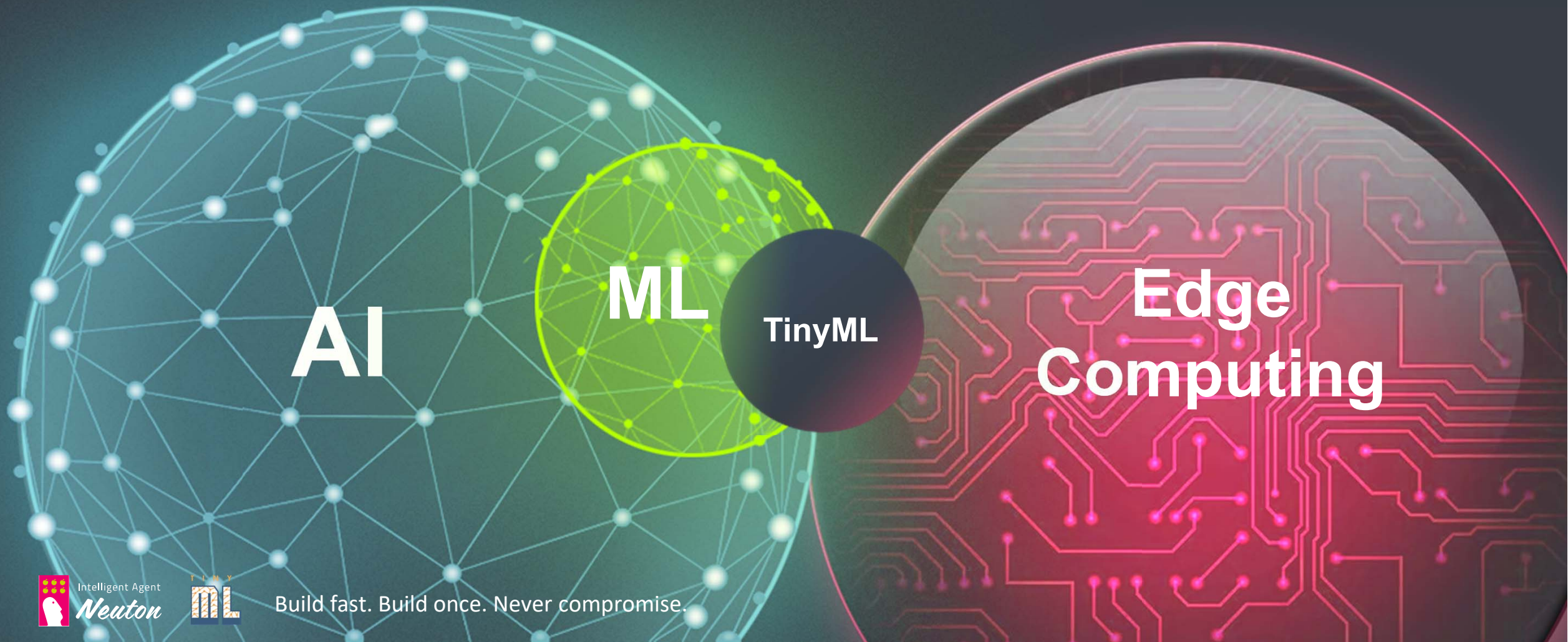
Business Users
Executives


Physical World
Digital World
Intelligent Agent
Neuton
The Missing Link
Between physical world
and device intelligence
Intelligent Agent
Neuton
TINY ML
Build fast. Build once. Never compromise.

Our Journey Today



What Is…



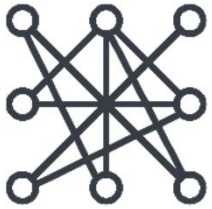
Intelligent Agent Neuton

Build fast. Build once. Never compromise.

What Not…

Main Barriers and Challenges for TinyML

Limited knowledge and availability of resources in Machine Learning and software development



The challenge of integrating **large ML models** into edge devices

Barriers of evaluating the quality and understanding the logic behind the model

What If You Could…

- build a model without having any technical expertise?
- create a model in 3 clicks?
- find the most optimal model in one iteration?
- produce models up to 1,000 times smaller than TensorFlow lite?
- run inferences up to 60% faster?
- eliminate the need to perform compression and not compromise accuracy?
- explain why your model makes every single prediction?
- accelerate your time to market by 85%?
- build fast, build once and never compromise accuracy or your business requirements?

What Next?

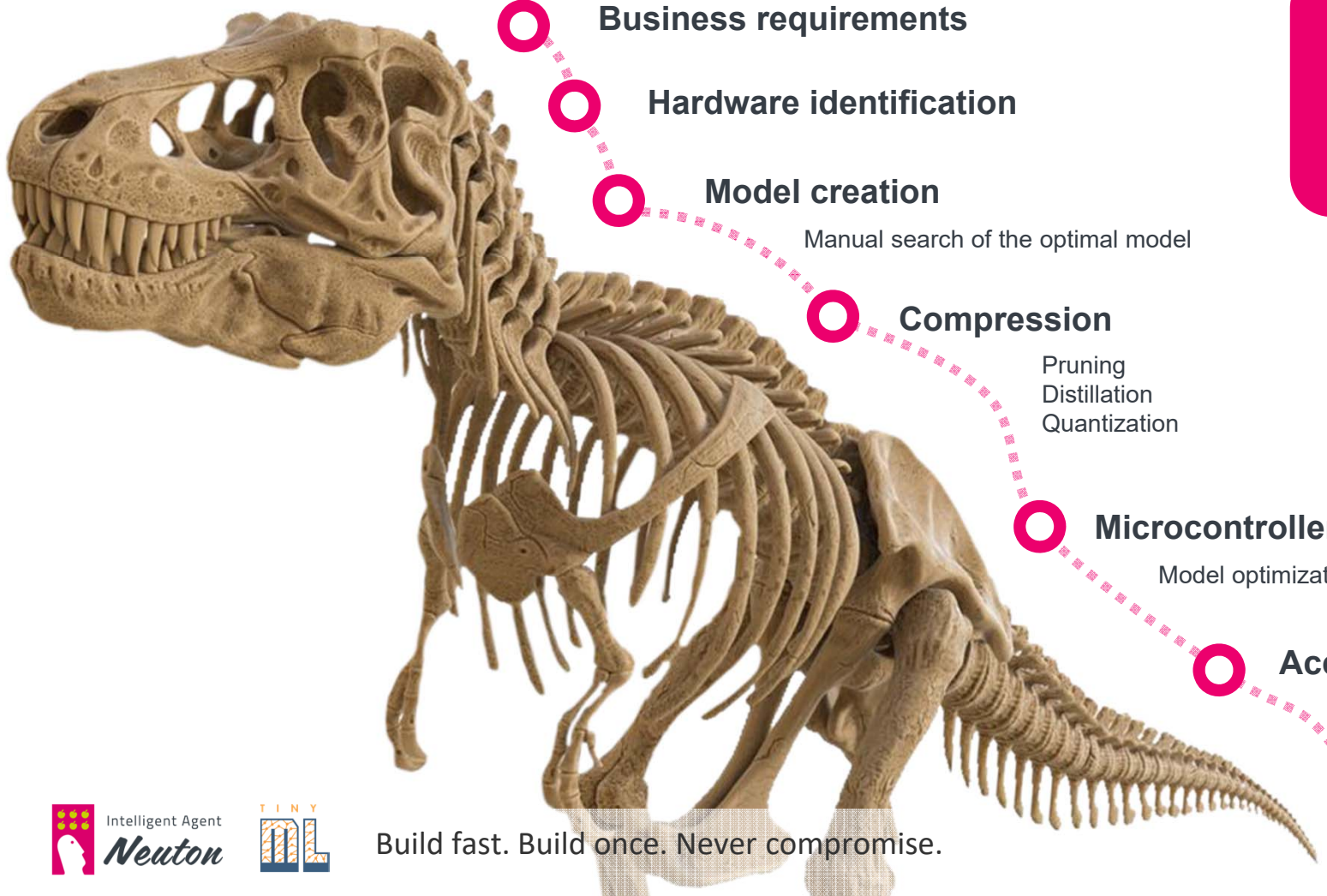
Intelligent Agent Neuton

TINY ML

Build fast. Build once. Never compromise.

Traditional Approach

Embedding Models to Edge Devices



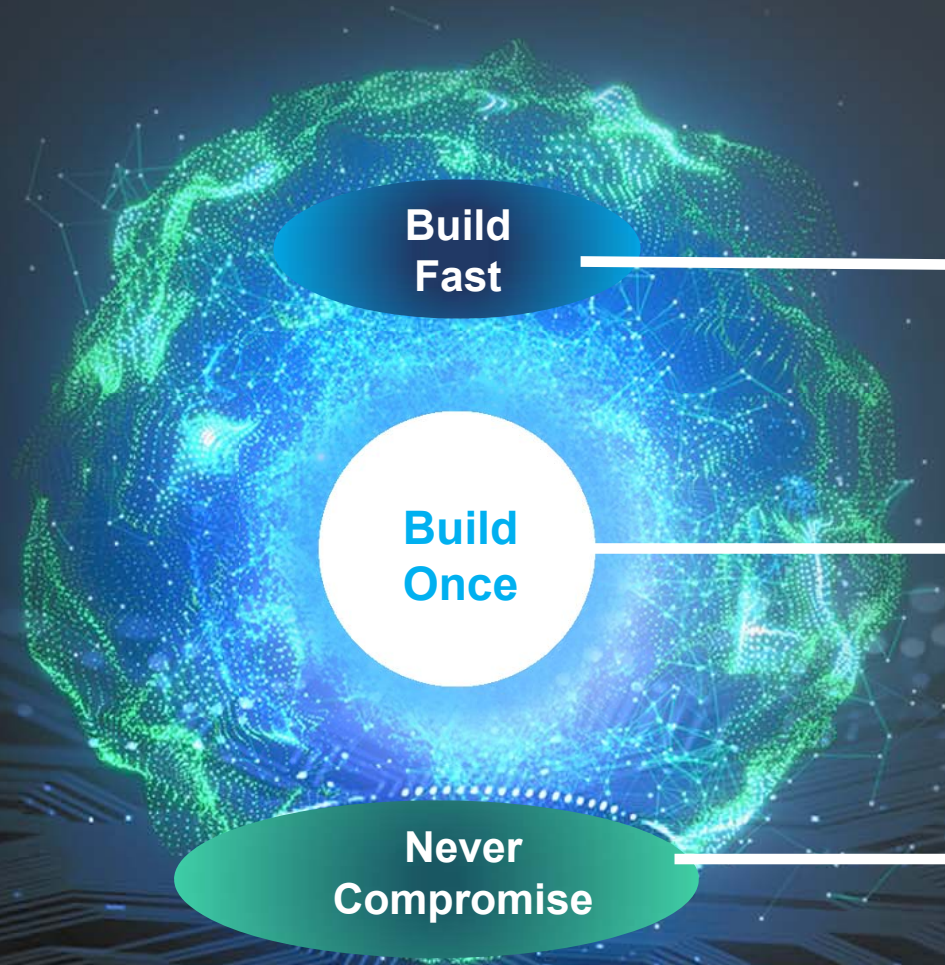
- **Business requirements**
- **Hardware identification**
- **Model creation**
  - Manual search of the optimal model
- **Compression**
  - Pruning
  - Distillation
  - Quantization
- **Microcontroller embedding**
  - Model optimization
- **Accuracy evaluation**
- **Can I trust this model?**
- **Test and repeat**

- **100% Manual efforts**
- **Takes months**
- **Minimum 4 different skill profiles**

Neuton Approach

Build Fast. Build Once. Never Compromise

True democratization of TinyML



Zero-Code AutoML platform **for All!**

Neural Network Framework that builds **extremely compact models**

Explainability tools for comprehensive model evaluation

+

Model Size or Model Quality

NEUTON'S MODELS ARE:

up to

X's

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

In comparison to TensorFlow and other algorithms

Our unique framework allows creation of a neural network structure of:

THE MOST OPTIMAL SIZE & ACCURACY

How Do We Create Compact Models without Comprising Accuracy?

Intelligent Agent Neuton

TINY ML

Build fast. Build once. Never compromise.

- **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**

How We Got to Today

Building Neural Networks

- **Manual random search of too many variables:**
  - Seed
  - Number on Neurons
  - Number of Layers
  - Activation Function (Sigmoid, ReLU etc)
  - Learning Rate
  - Number of epoches
  - Cross Validation Folds
  - Dropout
- Predetermined architecture (structure) defined by the researcher and the method of stochastic gradient descent
- Only neuron parameters undergo optimization the architecture remains predetermined
- Unnecessary growth of network size

Our Future Today

Efficient Global Optimization Algorithm

Our Framework uses a new efficient global optimization algorithm

- It is not based on back propagation of errors or stochastic gradient descent
- No problems of local extremes and plateaus
- Helps significantly improve each neuron’s efficiency and to reduce the network’s volume as a result
- Has enormous potential for parallelizing
- Allows for permanent cross-validation
- Automatic neuron-by-neuron network growth with overfitting control
- Dynamic growth of the network until it achieves its maximum generalization ability
- Learning the parameters of each neuron also allows for a significant reduction in the volume of the network

Our Future Today

Neuron by Neuron Structure Growth

Neuton uses neuron-by-neuron network structure growth

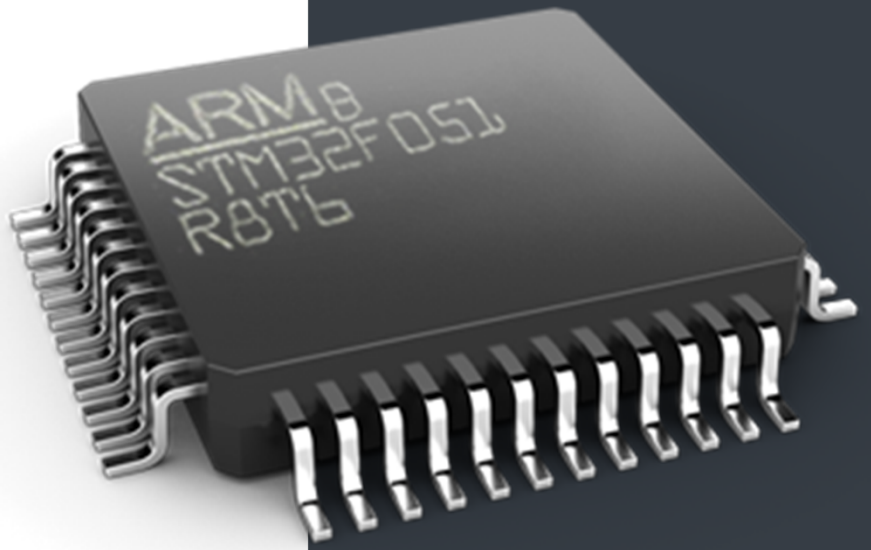
- Optimization occurs neuron by neuron at each iteration
- Eliminating the necessity of a Neural Network architecture search
- Creates minimum-size neural networks, with no loss of accuracy
- Neuton does not reduce the model size after creation (quantization, pruning, distillation)
- Accuracy is not compromised by compactness

Embedding into Edge Devices

Neuton Meets the World

Neuton's models can be built into microcontrollers even with the following characteristics

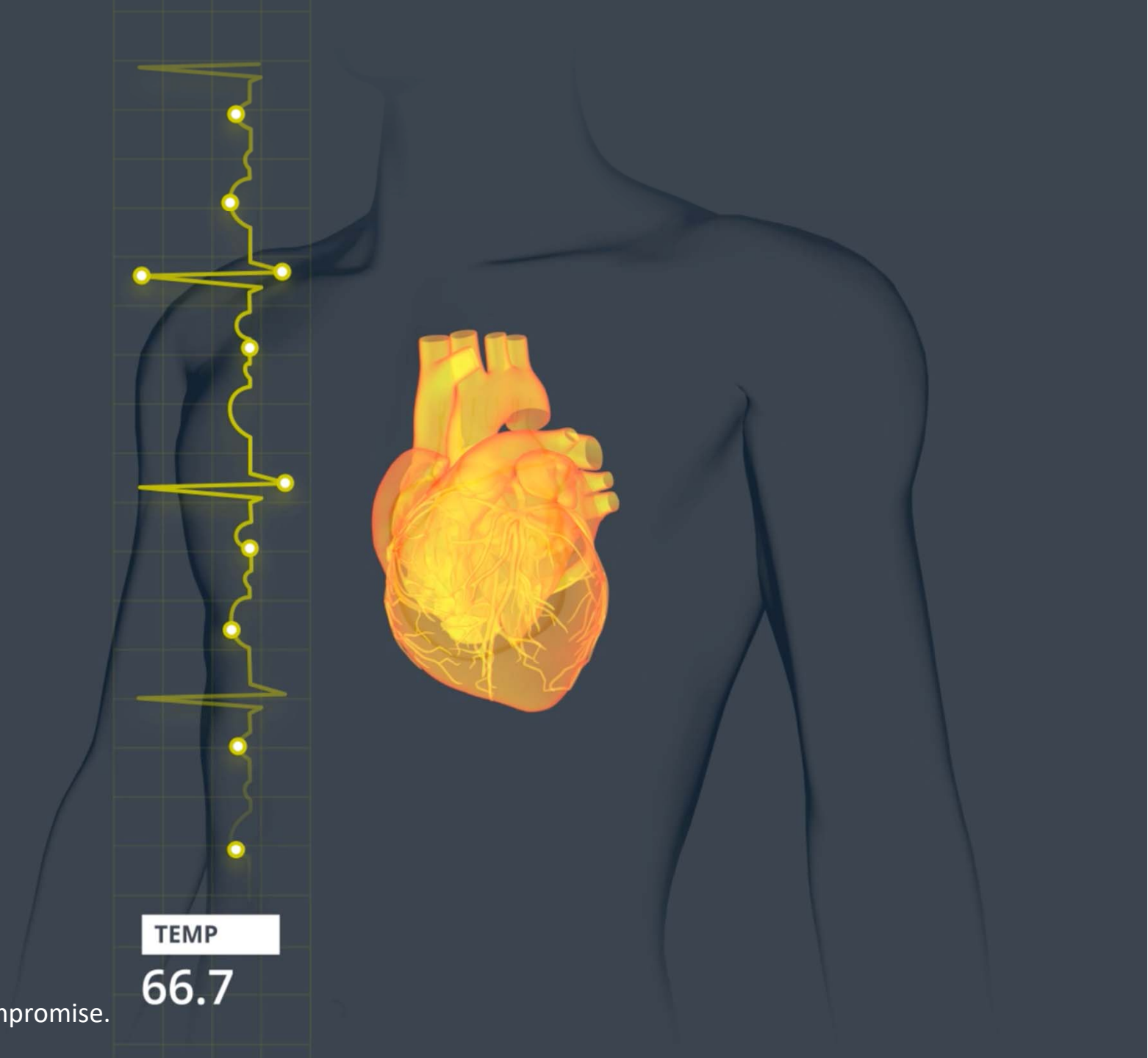
- Energy - 10s-100s mAh
- Processor < 100 MHz
- Memory < 100 Kb



Embedding into a Microcontroller

Neuton in Action

Determine cardiac arrhythmias case study



Intelligent Agent Neuton

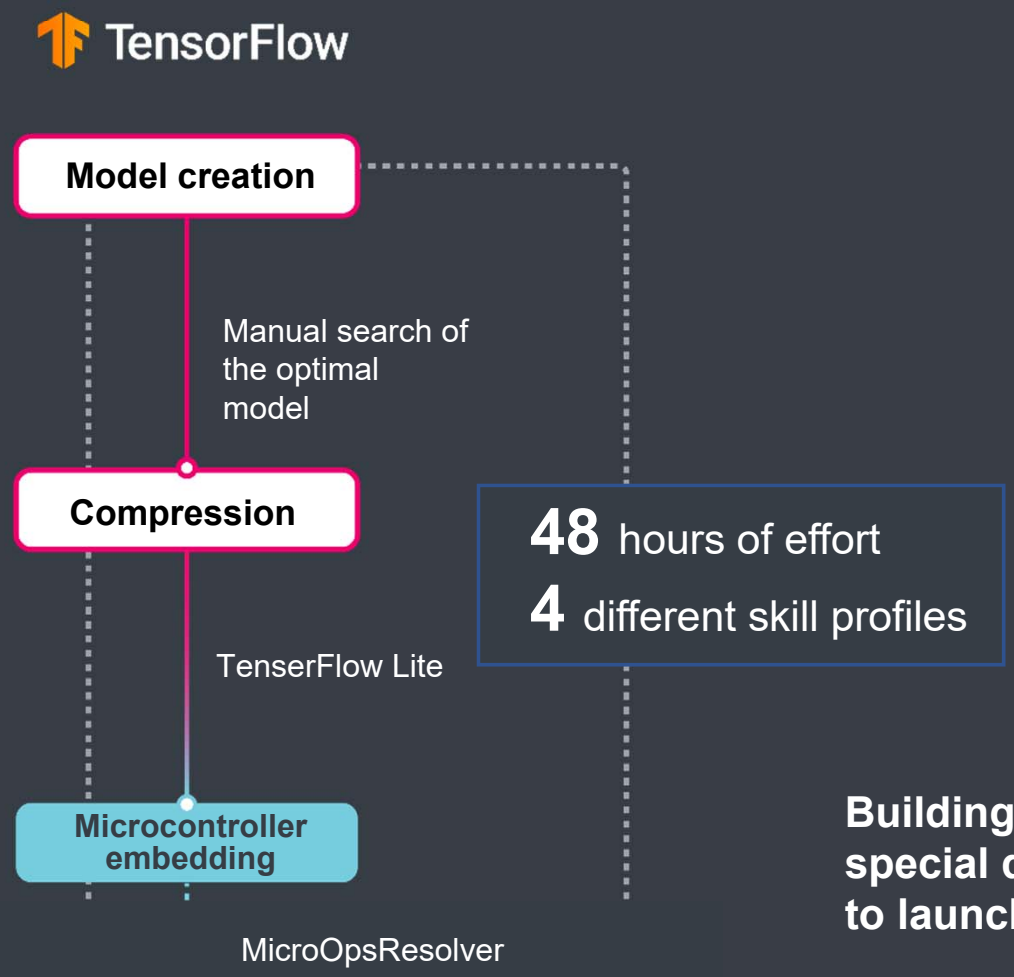
TINY ML

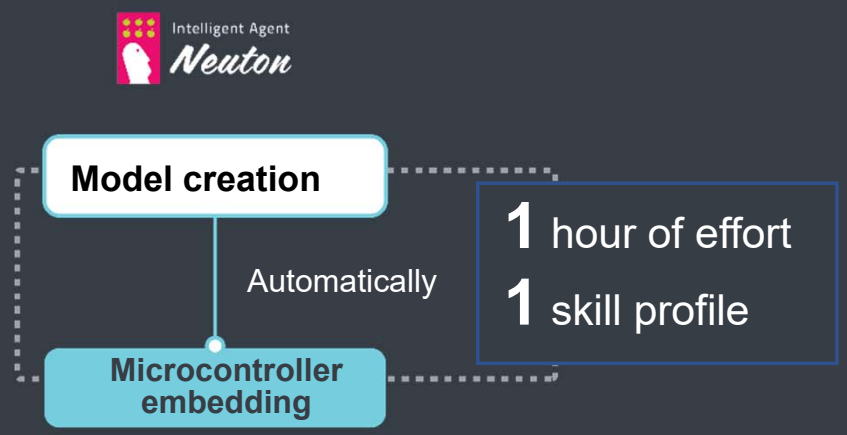
Build fast. Build once. Never compromise.

Neuton vs. TensorFlow Lite

	Neuton	TensorFlow Lite	TensorFlow	Edge Impulse	
AUC	0.966	0.958	0.965	0.852	
Size, Kbytes	0.7	6.93	33.04	-	**10 times smaller**
Coefficients	253	N/A	2,414	N/A	
RAM Usage, Bytes	970	3,326		2,203	**3.5 times less RAM utilization**
Flash Usage, Bytes	24,532	56,628		47,448	
Inference time, Microseconds	1,987	4,979		6,556	**2.5 times faster inference time**

Bringing the Future Today

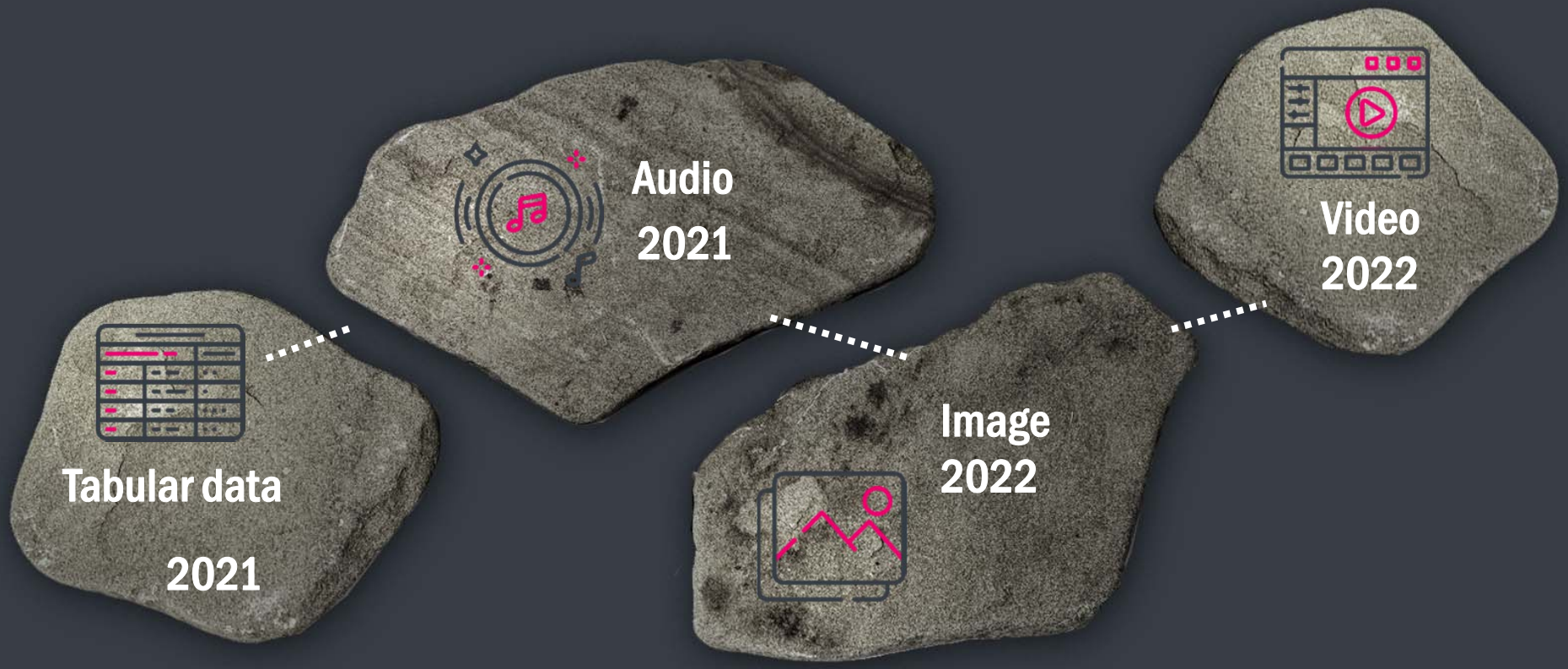




Building models with Neuton does not require any special qualifications and requires minimal effort to launch similar projects

Neuton Today

Positioned for the Future



Regression, Time Series, binomial and multinomial classification including NLP

A Model is Embedded into a Device, but What's Next?

In the context of making our devices AI driven, Explainability is essential

The 4 W's We Should Ask Ourselves:

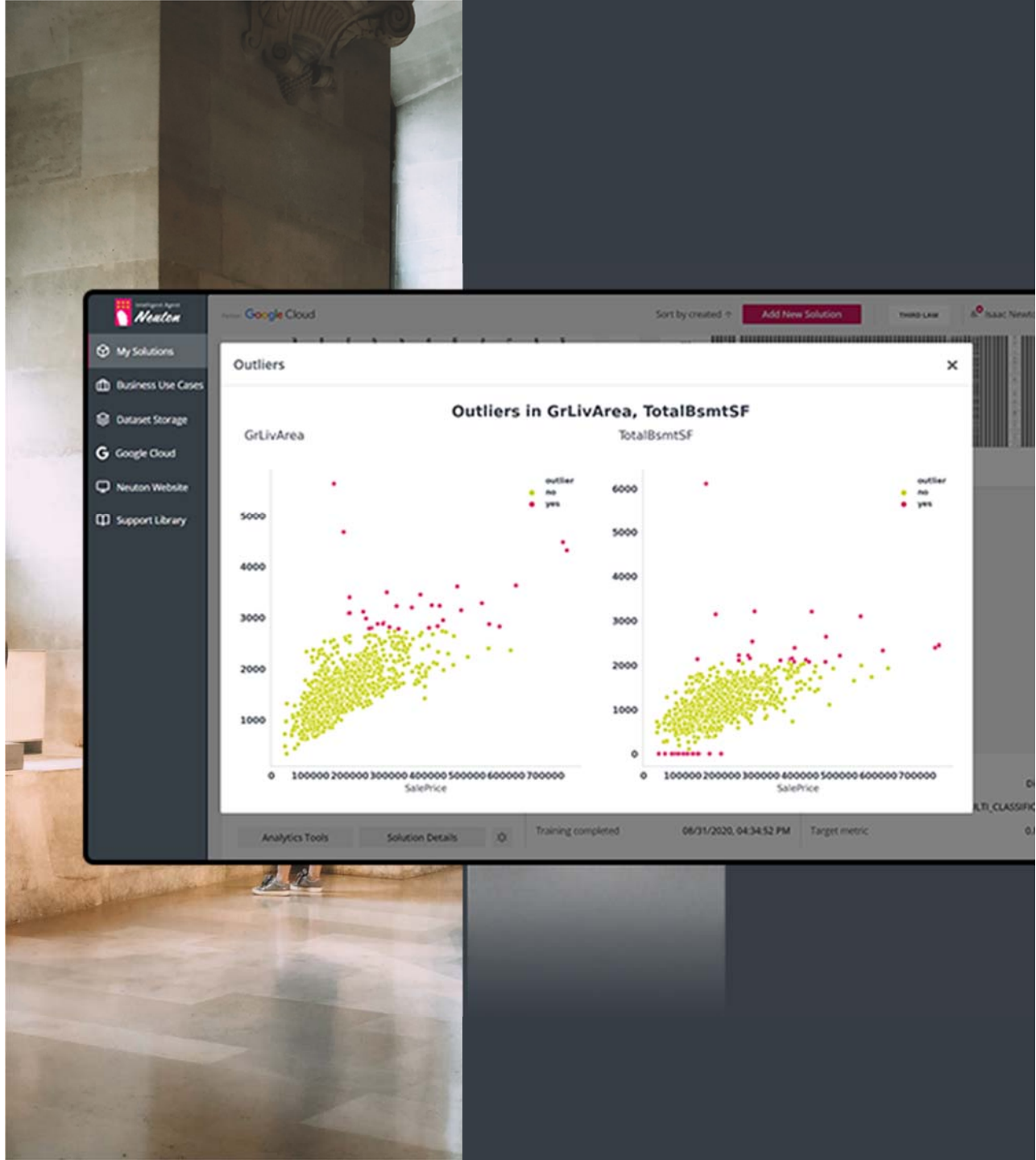
1. **What** does my data consist of?
2. **Where** are the important features?
3. **Why** does my model make certain Predictions?
4. **When** should I consider retraining my model?

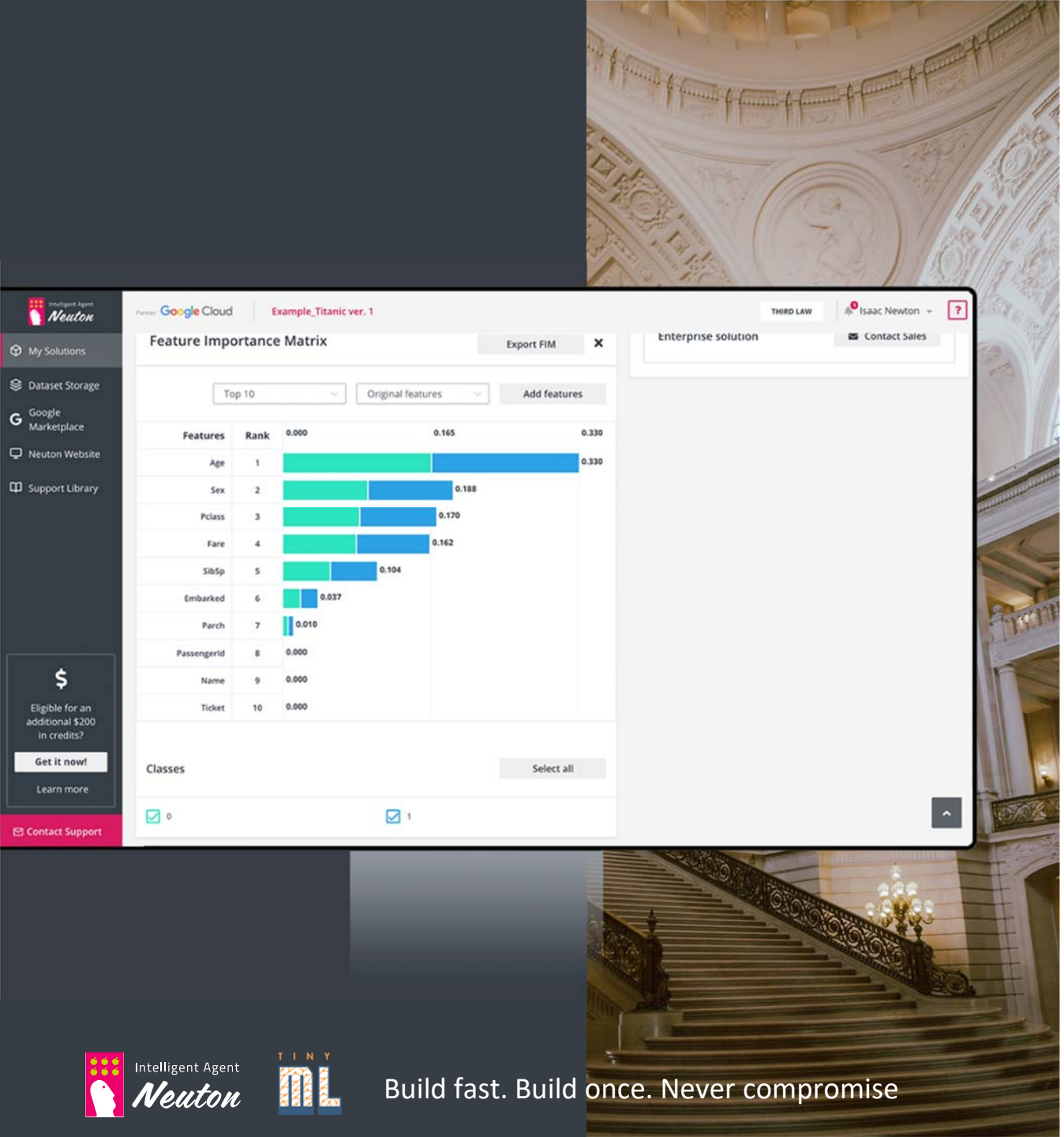
+

What

Does my Data Consist of?

Exploratory Data Analysis - graphical data analysis and the most important statistics





+

Where

are the Important Features in my Dataset?

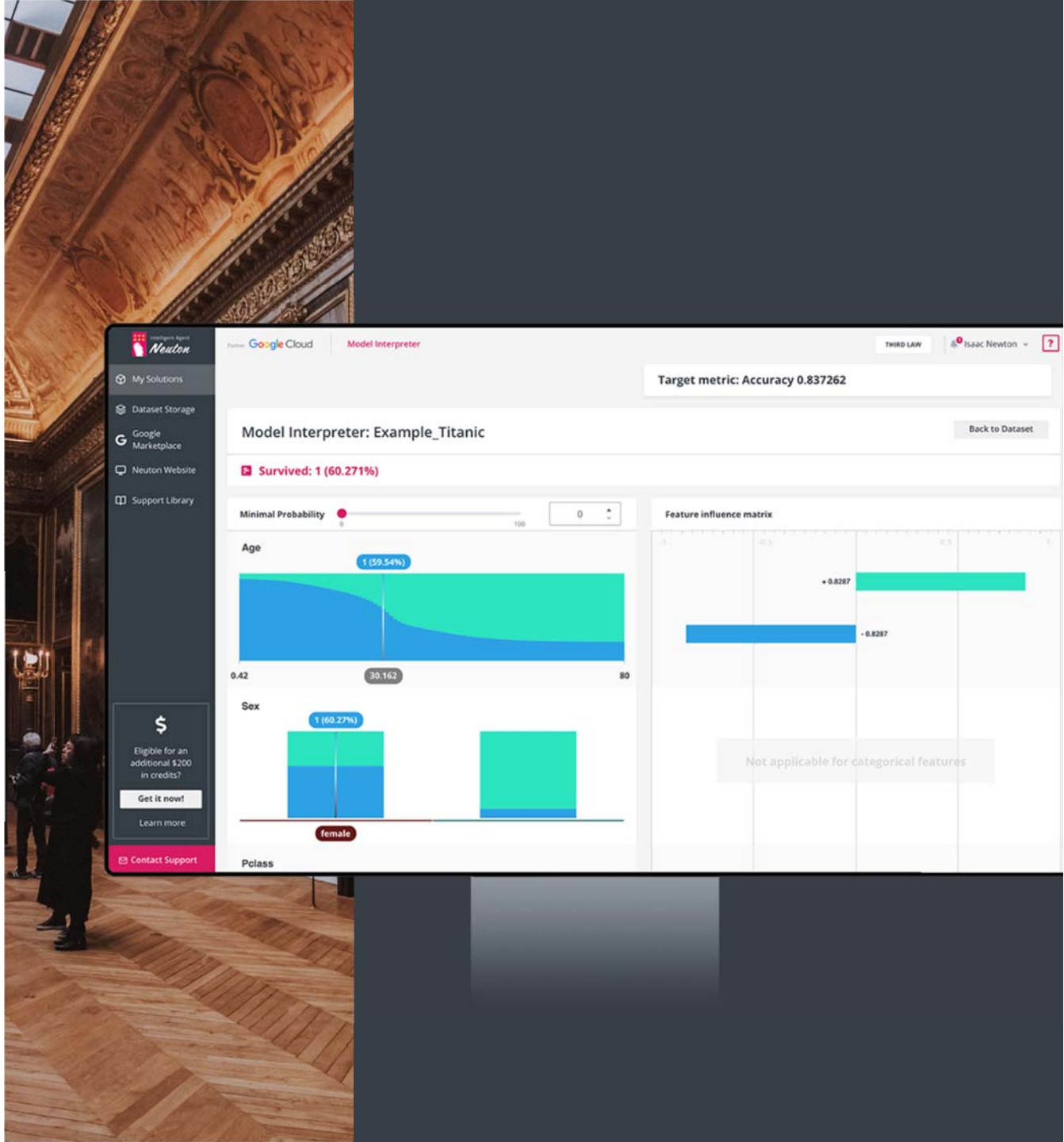
Feature Importance Matrix is a chart with features that had the most and least significant impact on the model prediction power.

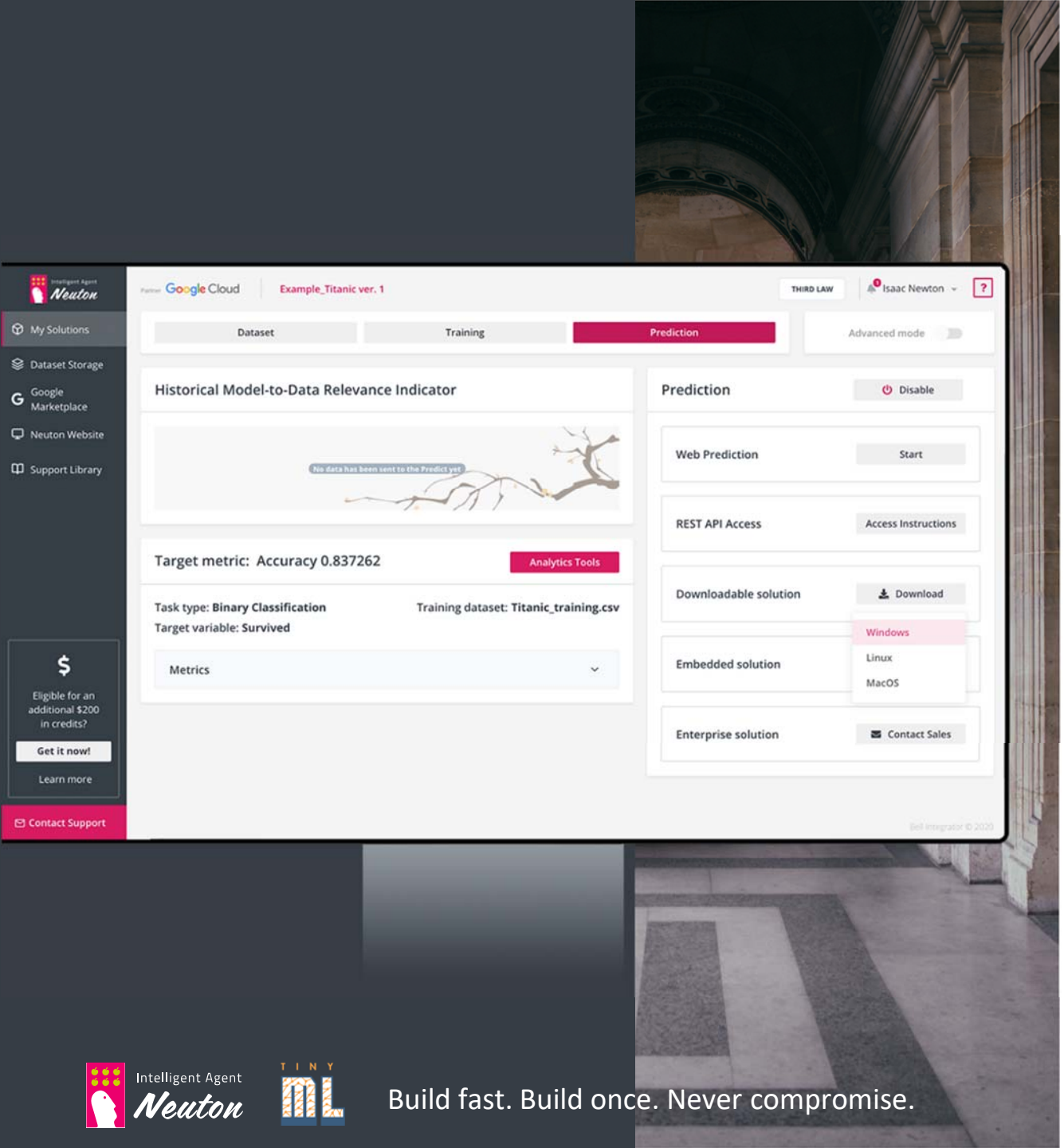
+

Why

Does my Model Make Certain Predictions?

The Model Interpreter allows you to see the logic, direction and the effects of changes in individual variables in the model.





+

When

Should I Consider Retraining my Model?

Historical Model-to-data Relevance Indicator allows to manage a model lifecycle by signaling for models to retrain.

TinyML EMEA Technical Forum 2021

Explainability Office:

the next chapter for tinyML

Tuesday June 8, 2021
6:30 pm UTC +1

Join our Explainability office demo

Intelligent Agent Neuton

TINY ML

Build fast. Build once. Never compromise.

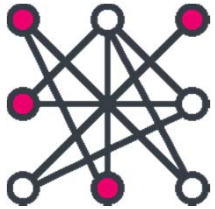
Your Future Today

- build a model without having any technical expertise?
- create a model in 3 clicks?
- find the most optimal model in one iteration?
- produce models up to 1,000 times smaller than TensorFlow lite?
- run inferences up to 60% faster?
- eliminate the need to perform compression and compromise accuracy?
- explain why your model makes every single prediction?
- accelerate your time to market by 85%?
- build fast, build once and never compromise accuracy or your business requirements?

Build Fast. Build Once. Never Compromise.

Limited knowledge and availability of resources
in Machine Learning and software development

Zero-Code AutoML platform **for non-Data Scientists**



The challenge of integrating **large ML models** into edge devices

Neural Network Framework that builds **extremely compact models in one iteration**

Barriers of evaluating the quality
evaluating the quality and understanding the logic behind the model

Explainability tools for comprehensive model evaluation

Thank you!

www.linkedin.com/company/neuton/

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!

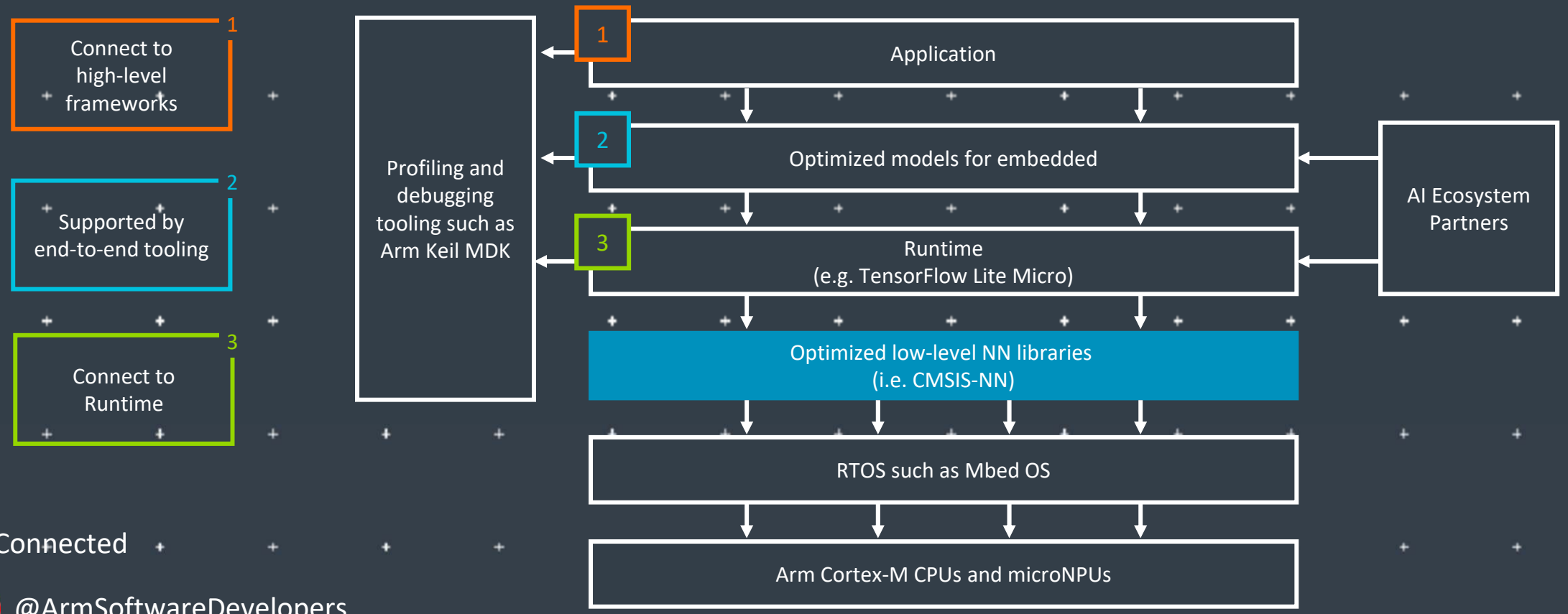
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

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

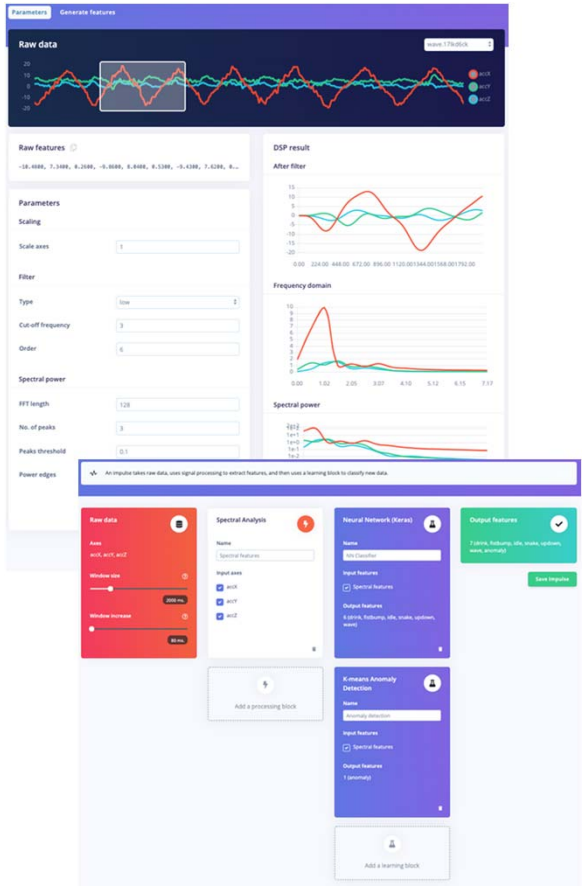
Open source SDK

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

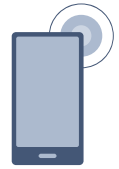
Reinforcement learning for decision making

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

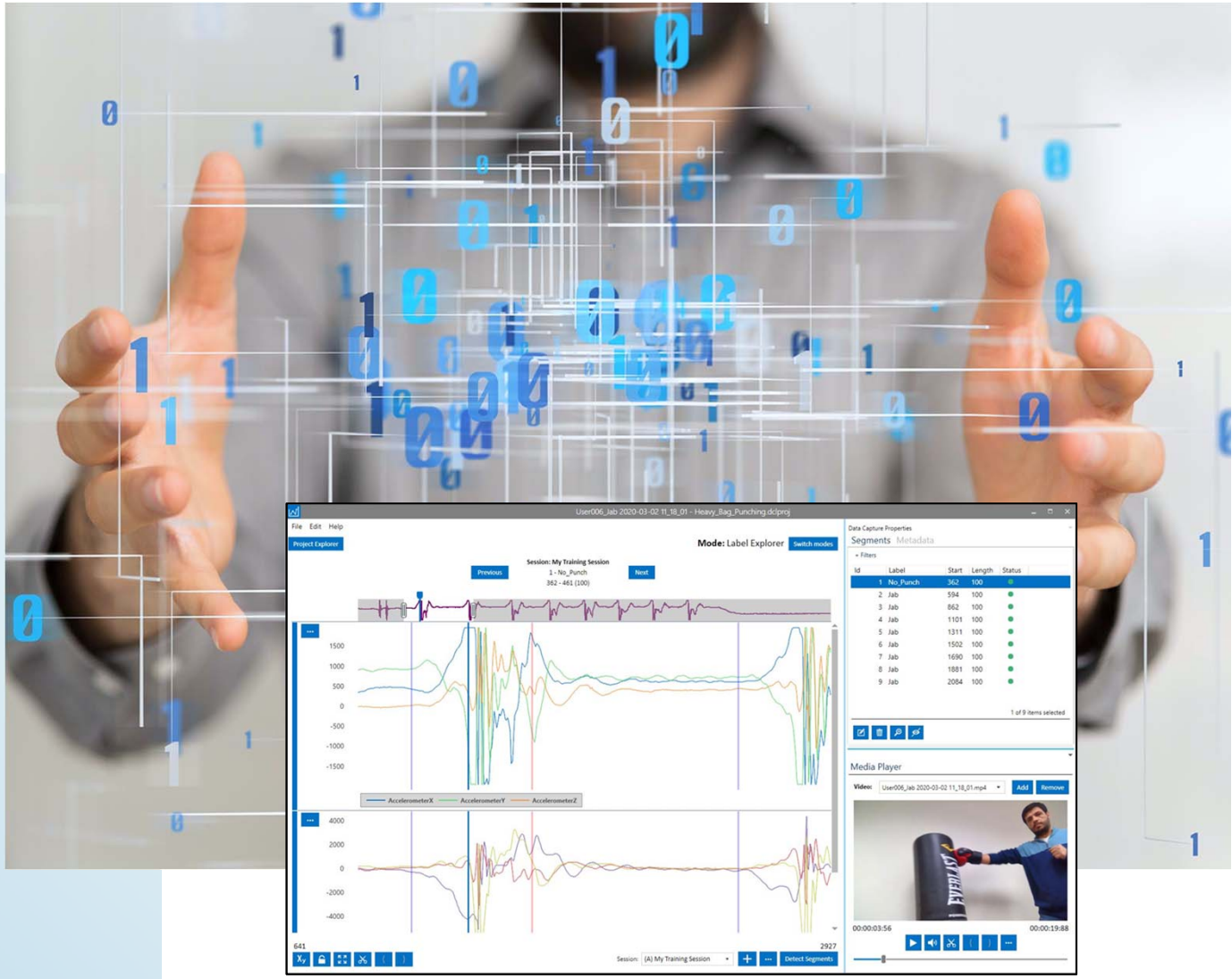
SensiML™

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

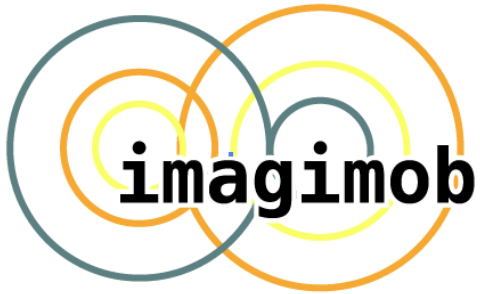
We enable the creation of production-grade smart sensor devices.



sensiml.com

TINY ML EMEA Virtual JUNE 7-10,2021

Silver Sponsors



Copyright Notice