tinyML® Summit

Enabling Ultra-low Power Machine Learning at the Edge

Products and applications enabled by tinyML

March 28 – 29, 2023

www.tinyML.org

Tiny spiking AI for the sensor-edge

Petrut Bogdan

Neuromorphic Architect

petrut.bogdan@innatera.com

Outline

- Innatera
- Spiking Neural Processor
- Talamo SDK
- Ultra-low power edge applications

Made in Delft

innatera

- **Ultra-low power intelligence for sensors**
- Spun out of the Delft University of Technology in 2018
- 57 employees, offices in the Netherlands and India
- Funded by deep-tech investors Matterwave Ventures and MIG Capital

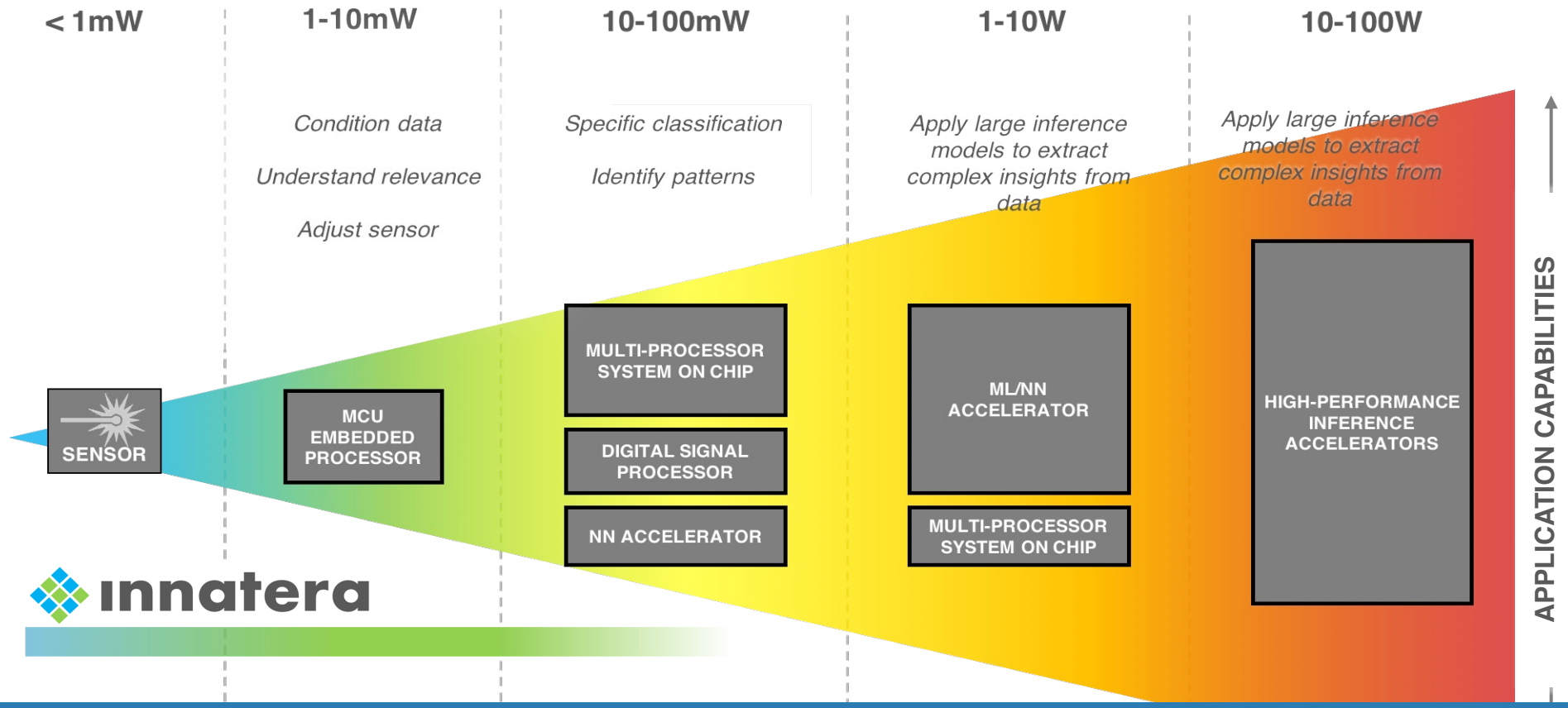
matter wave ventures

MIG CAPITAL

TU Delft

Bringing intelligence to the sensor edge

innatera

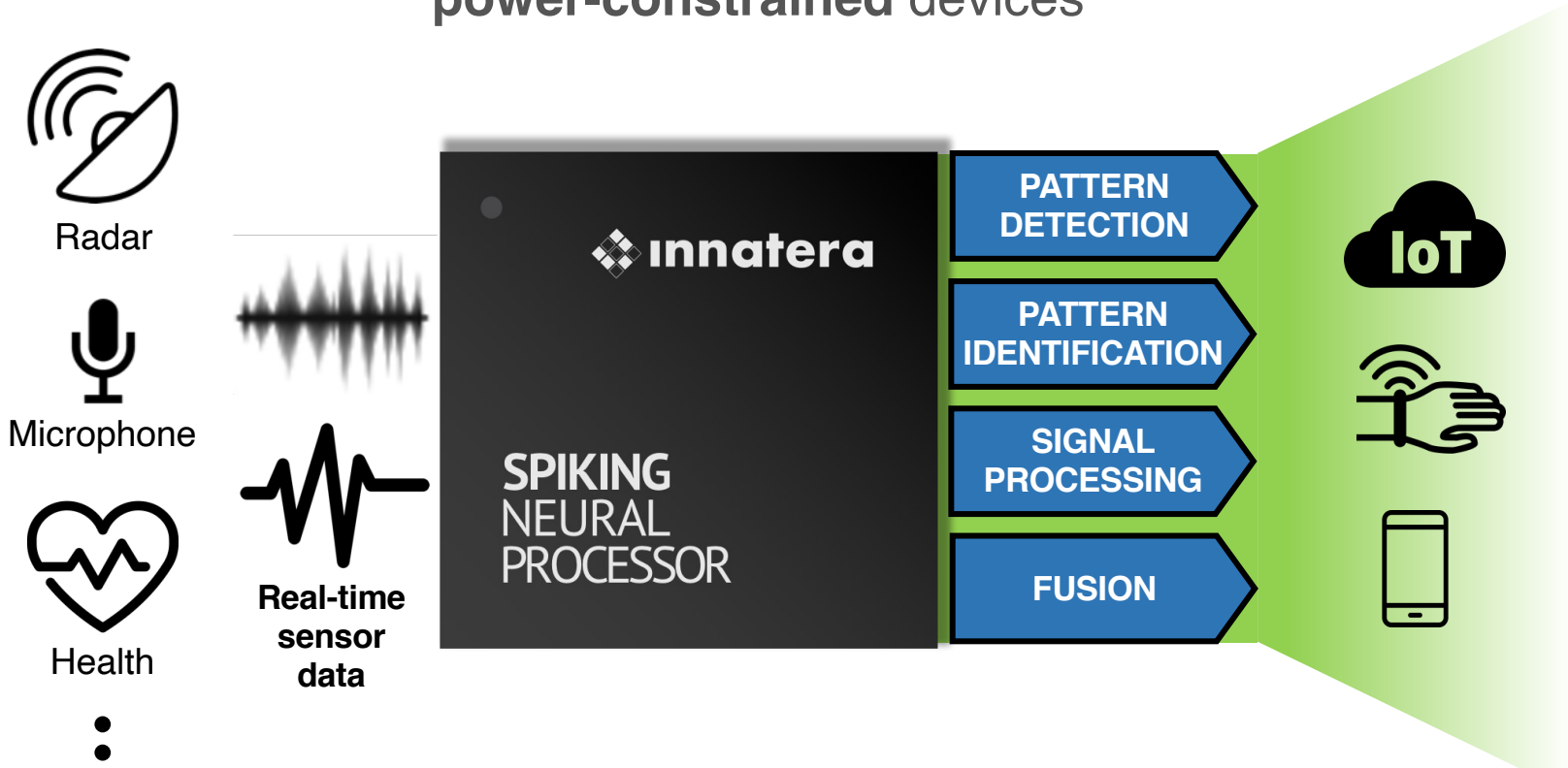


Sensor-edge constraints:

- average power < 1mW
- code size < 10 KB
- latency < 100 ms
- bill of materials

Spiking Neural Processor

Brain-inspired processor for **turn-key intelligence** in **power-constrained** devices

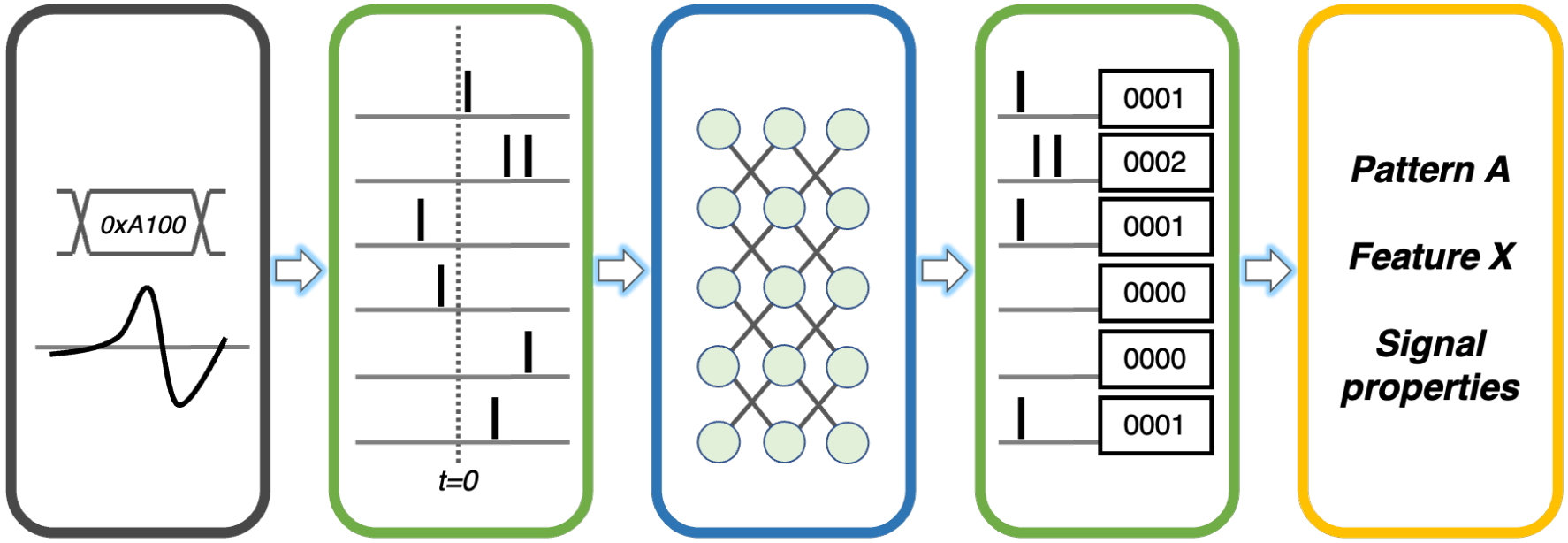


Always-on sensing applications

Millisecond-scale processing latency envelope

Milli- and sub-milliwatt power envelope

Processing sensor data with Spiking Neural Networks



Spiking Neural Network models can be up to 100x smaller than conventional Artificial / Deep Neural Networks

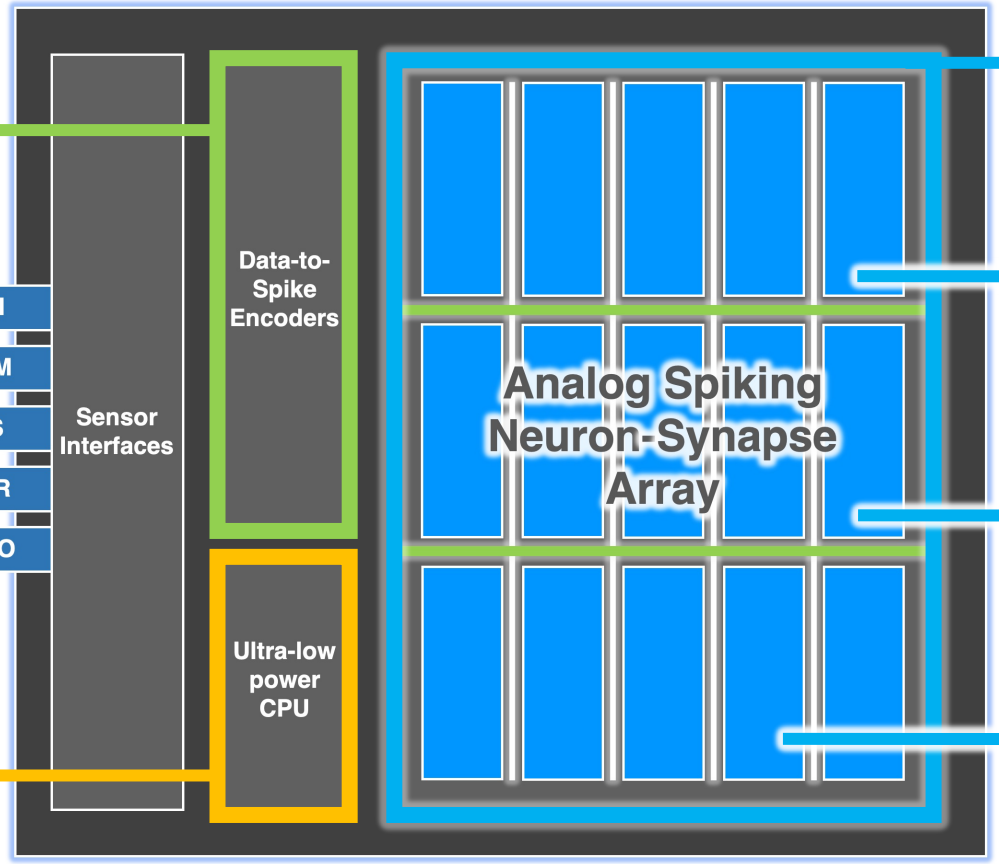
Spiking Neural Processor (SNP)

innatera

On-chip encoders automatically convert sensor data into spikes

Standard sensor interfaces

Ultra-low power CPU for standalone operation



Real-time, massively parallel inference of spiking neural networks

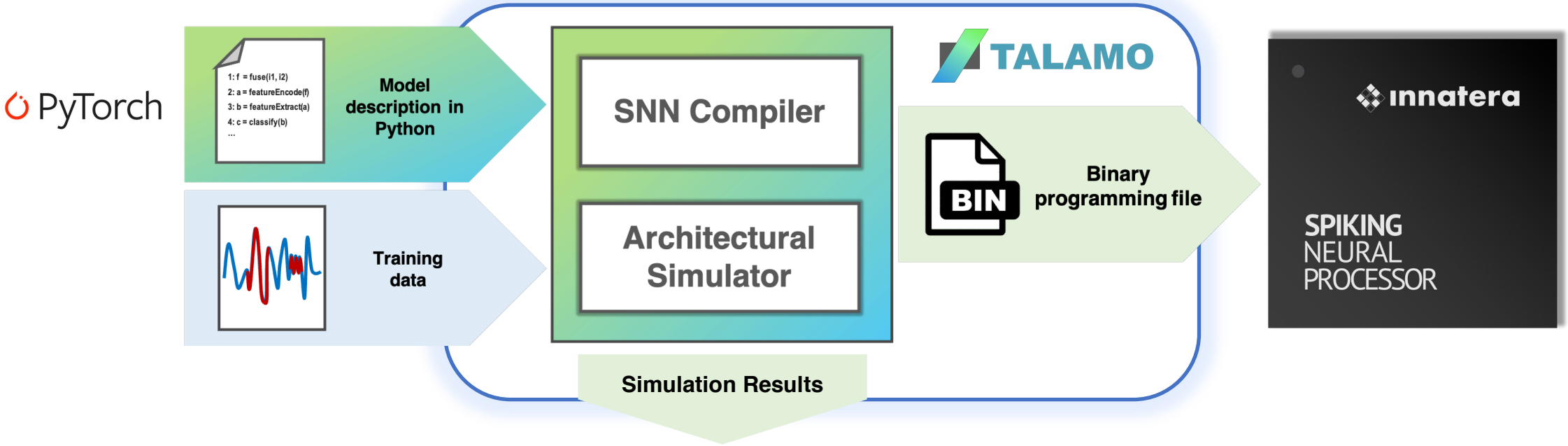
Ultra-low power, ultra-low latency computing elements

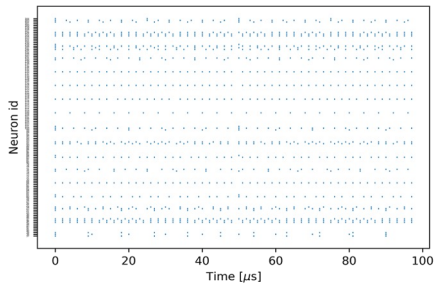
Programmable array supports complex network topologies

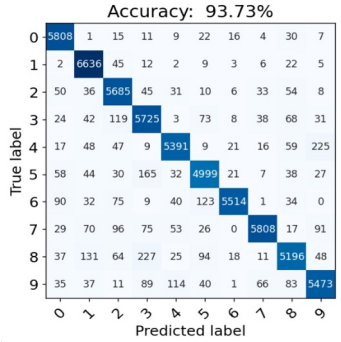
Scalable segmented architecture

The only processor needed for always-on sensing applications

Innatera's powerful Software Development Kit - Talamo



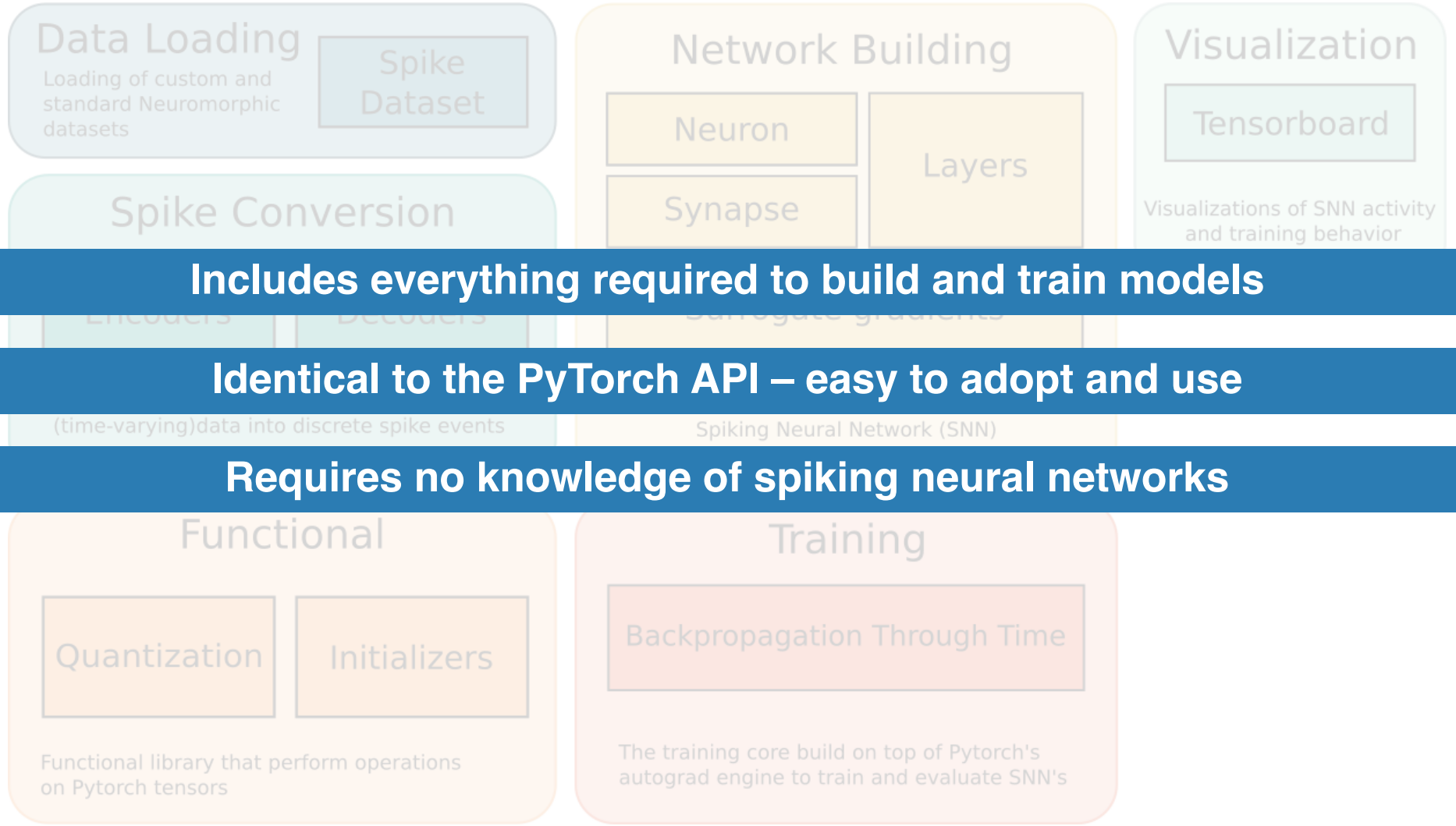




Simple	Turn-key	Standard	Fast
Easy to use, familiar workflow	Easily build and deploy models to hardware	Native integration with PyTorch, TensorBoard	Rapid simulation and deployment

Talamo's PyTorch API - spiking neural networks made easy

innatera

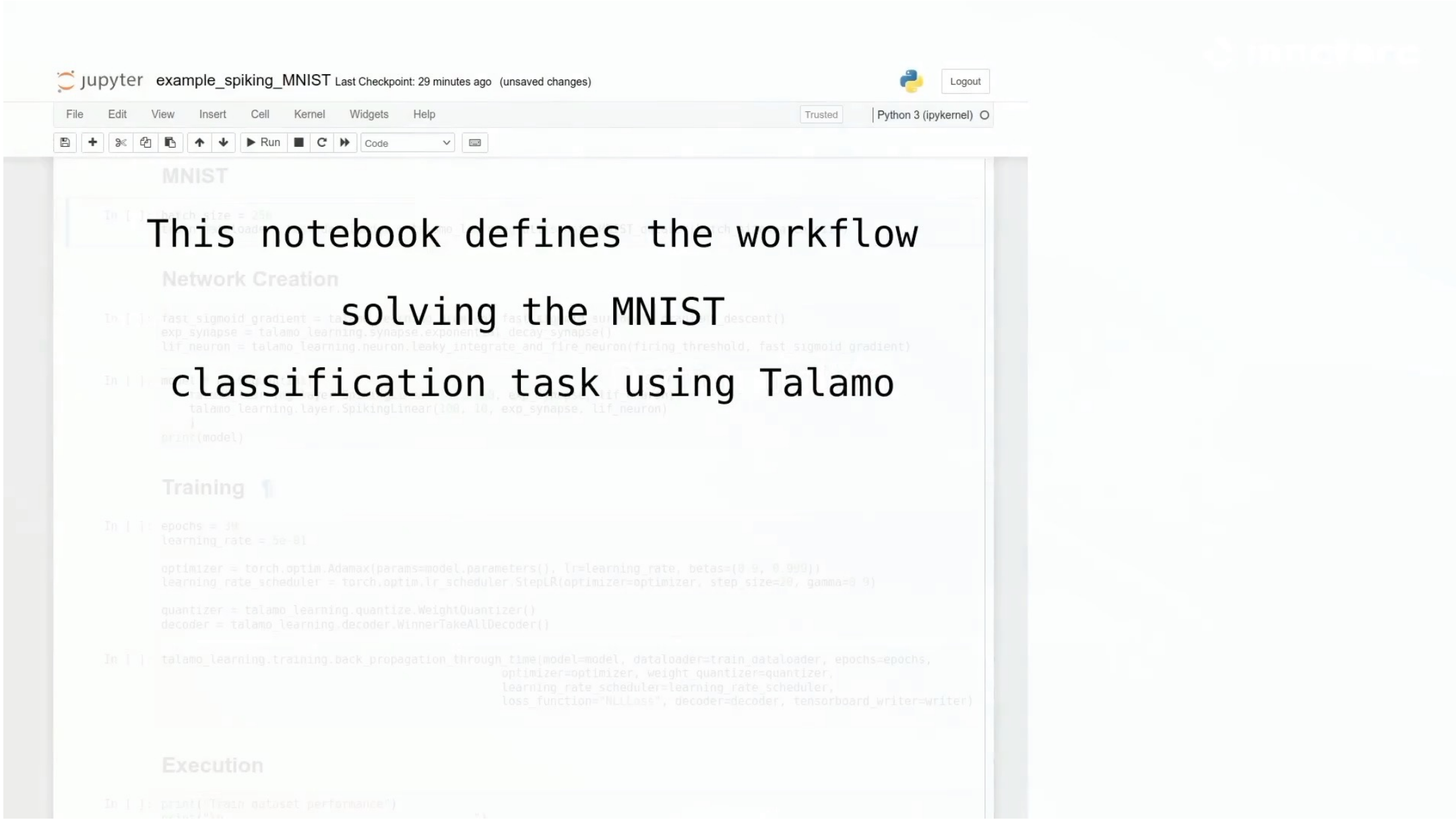


Includes everything required to build and train models

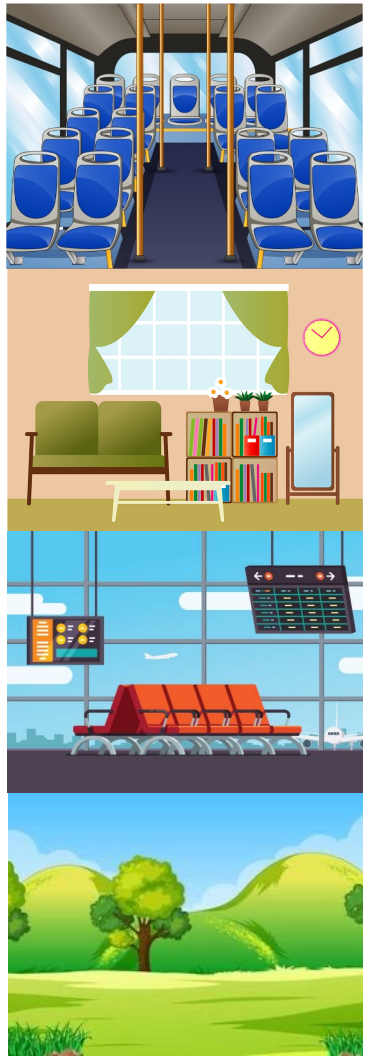
Identical to the PyTorch API – easy to adopt and use

Requires no knowledge of spiking neural networks

Building a spiking neural network in Talamo



Always-on audio scene classification in hearables



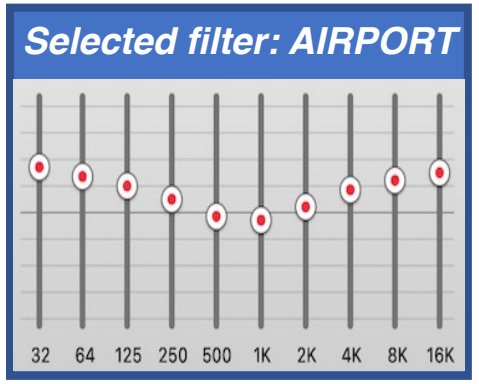
Audio scenes

Audio data

SPIKING NEURAL PROCESSOR

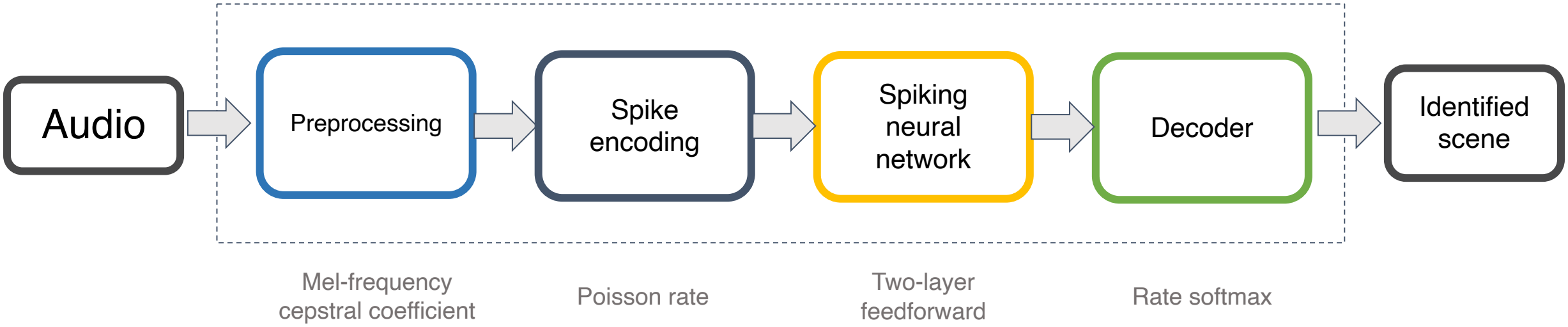
"Airport"

Identified scene



Filter selection for optimal sound quality

Solution pipeline



Audio scene classification (DCASE 2020)

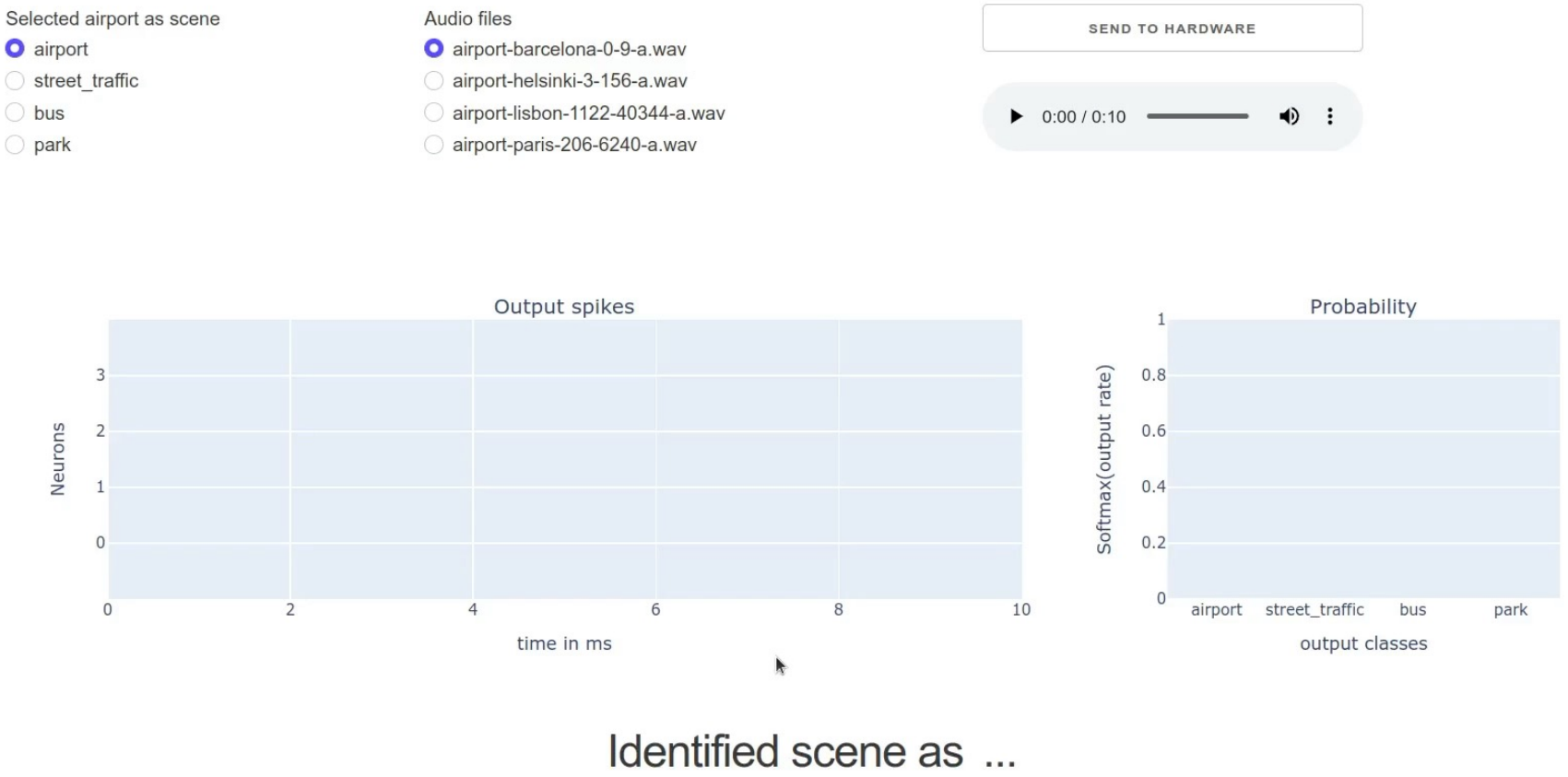
innatera

Power
(total peak)
1.06 mW

Accuracy
~85%

Inference latency
~1 ms / 1 s

Model size
~3kB



Summary

- Edge AI is:
  - Power constrained
  - Latency constrained
  - Code-size constrained
  - BOM constrained
- The Spiking Neural Processor delivers audio scene classification in
  - a power budget of ~1mW
  - an inference latency of ~1ms
  - with a code size of ~3kB
- Talamo SDK – simplifies model development with a standard, well-understood work flow
- Easy to adopt, build, and deploy sensor-edge solutions with unprecedented power-performance

The future of TinyML is Neuromorphic!

Let’s make sense together.

Innatera Nanosystems BV
Patrijsweg 20
Rijswijk 2289EX
The Netherlands

info@innatera.com

www.innatera.com

Copyright Notice

www.tinyml.org