

tinyML[®] Summit

Enabling Ultra-low Power Machine Learning at the Edge

Products and applications enabled by tinyML

March 28 – 29, 2023



www.tinyML.org

T I N Y



Tiny spiking AI for the sensor-edge

Petrut Bogdan

Neuromorphic Architect

petrut.bogdan@innatera.com

Outline

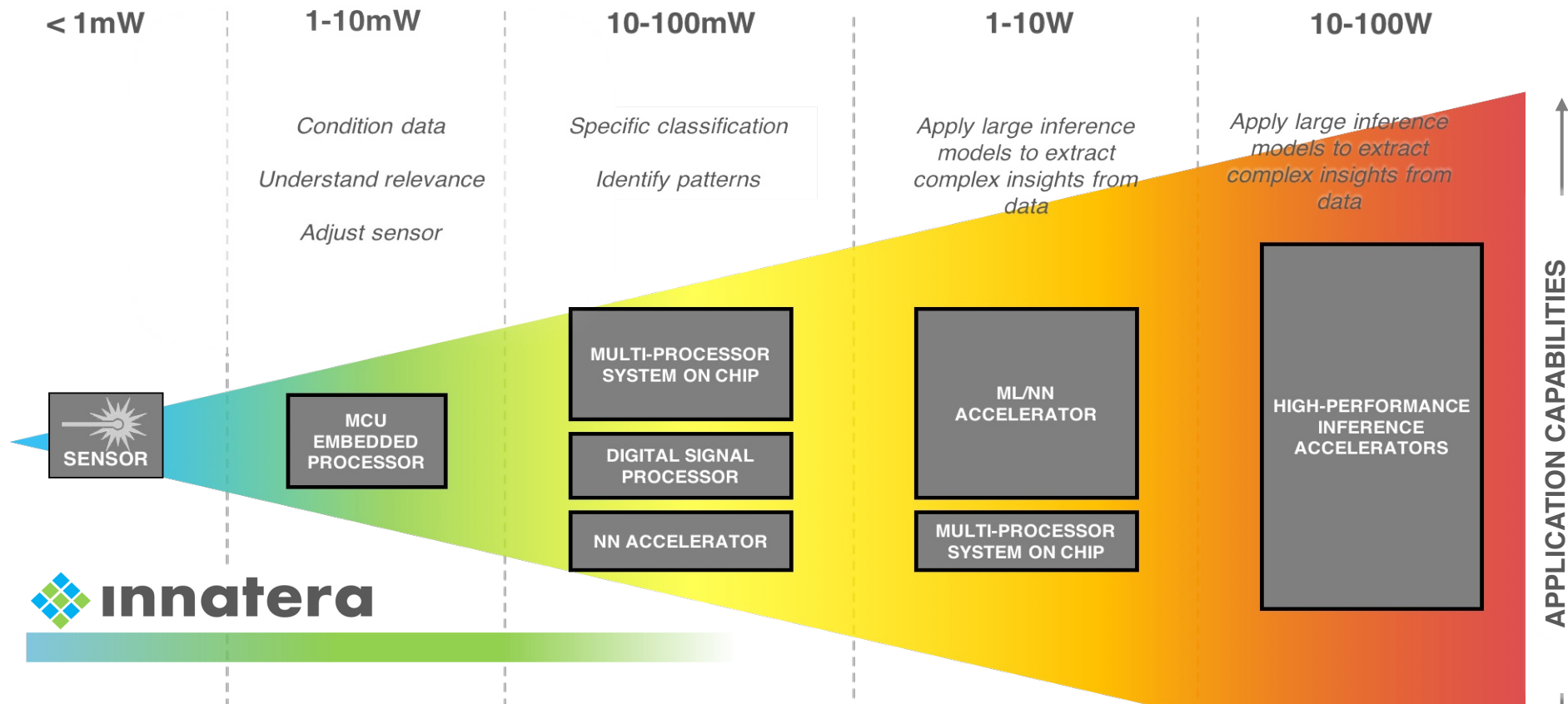
- Innatera
- Spiking Neural Processor
- Talamo SDK
- Ultra-low power edge applications

Made in Delft

- 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](#)



Bringing intelligence to the sensor edge

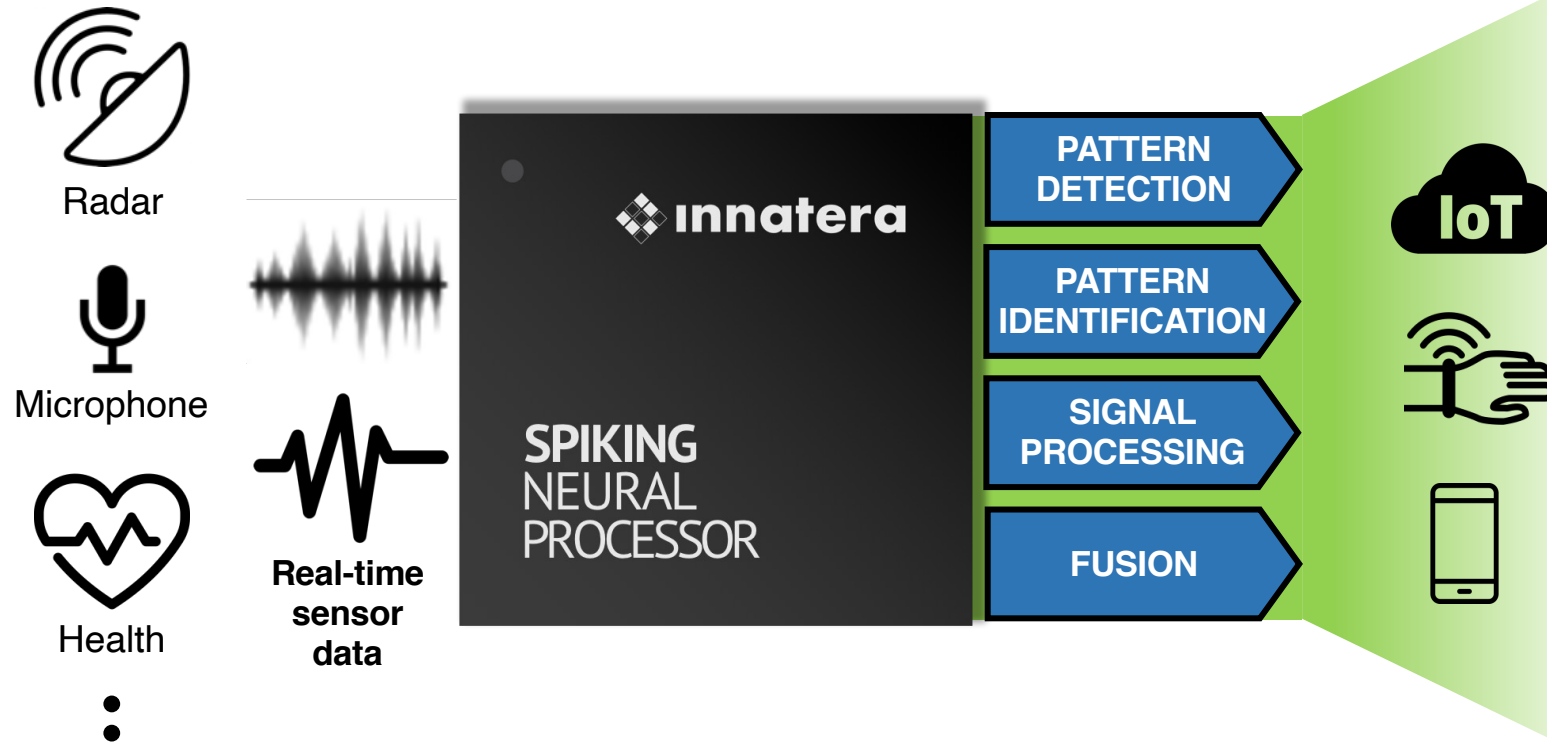


Sensor-edge constraints:

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

Spiking Neural Processor

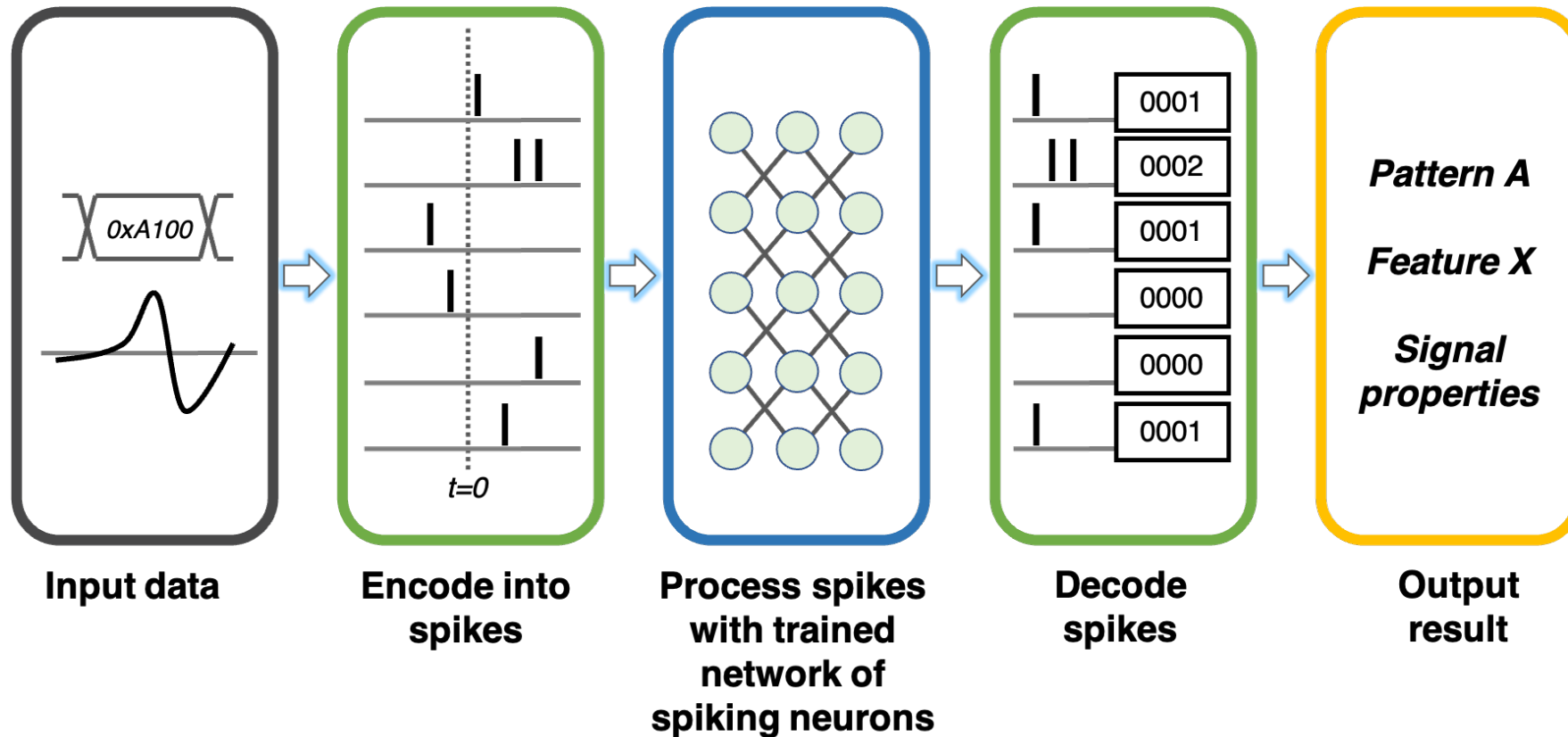
Brain-inspired processor for **turn-key intelligence** in
power-constrained devices



Always-on sensing applications

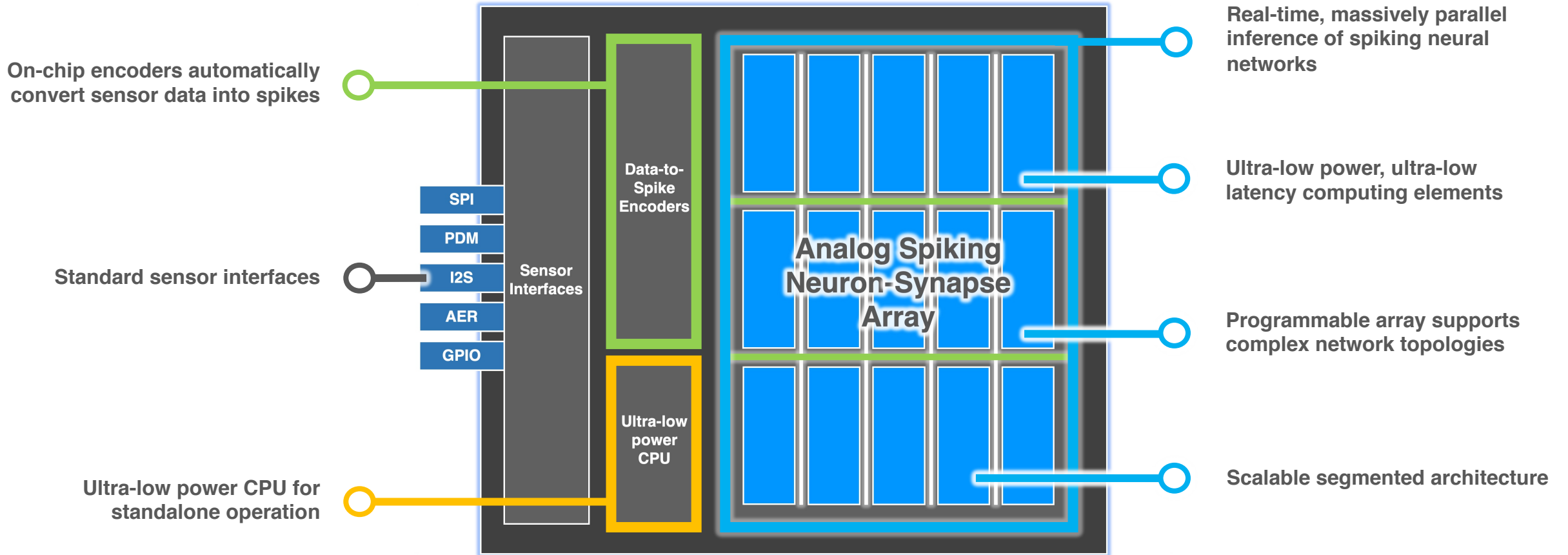
Millisecond-scale processing latency envelope

Milli- and sub-milliwatt power envelope



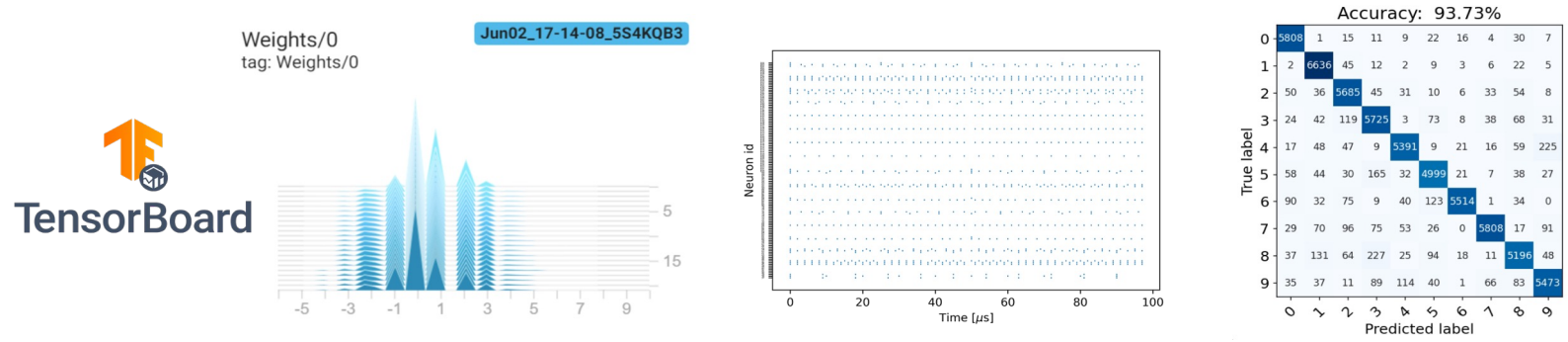
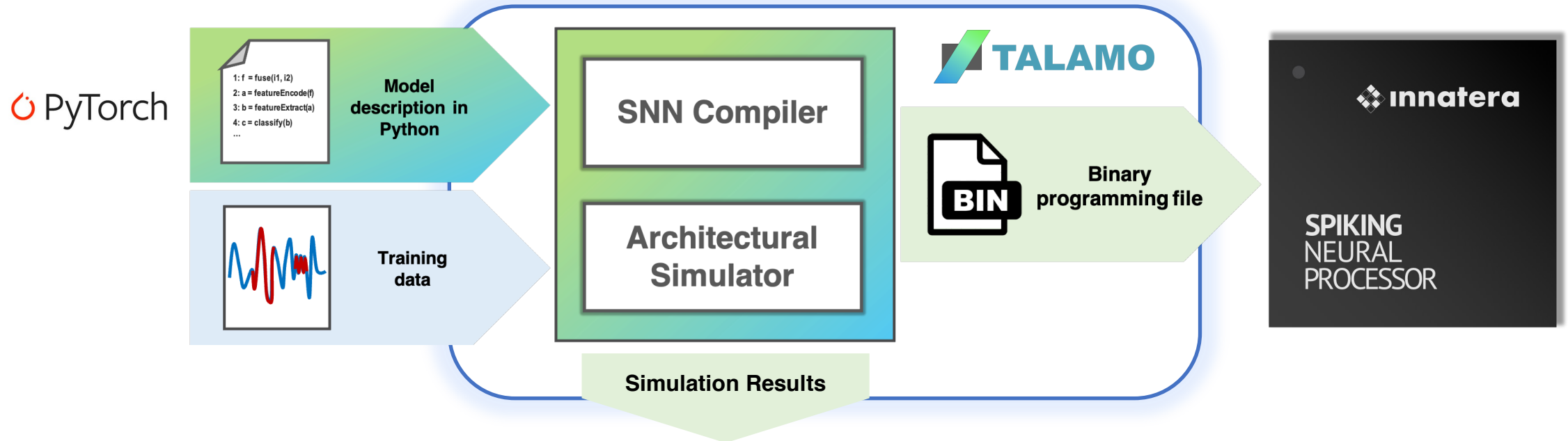
Spiking Neural Network models can be up to 100x smaller than conventional Artificial / Deep Neural Networks

Spiking Neural Processor (SNP)



The only processor needed for always-on sensing applications

Innatera's powerful Software Development Kit - Talamo



Simple
Easy to use, familiar workflow

Turn-key
Easily build and deploy models to hardware

Standard
Native integration with PyTorch, TensorBoard

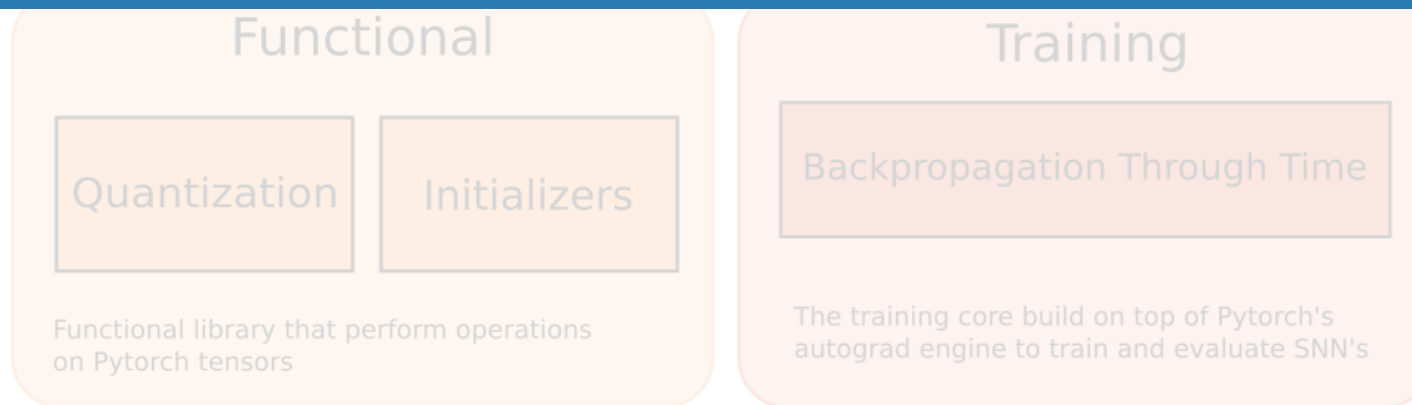
Fast
Rapid simulation and deployment



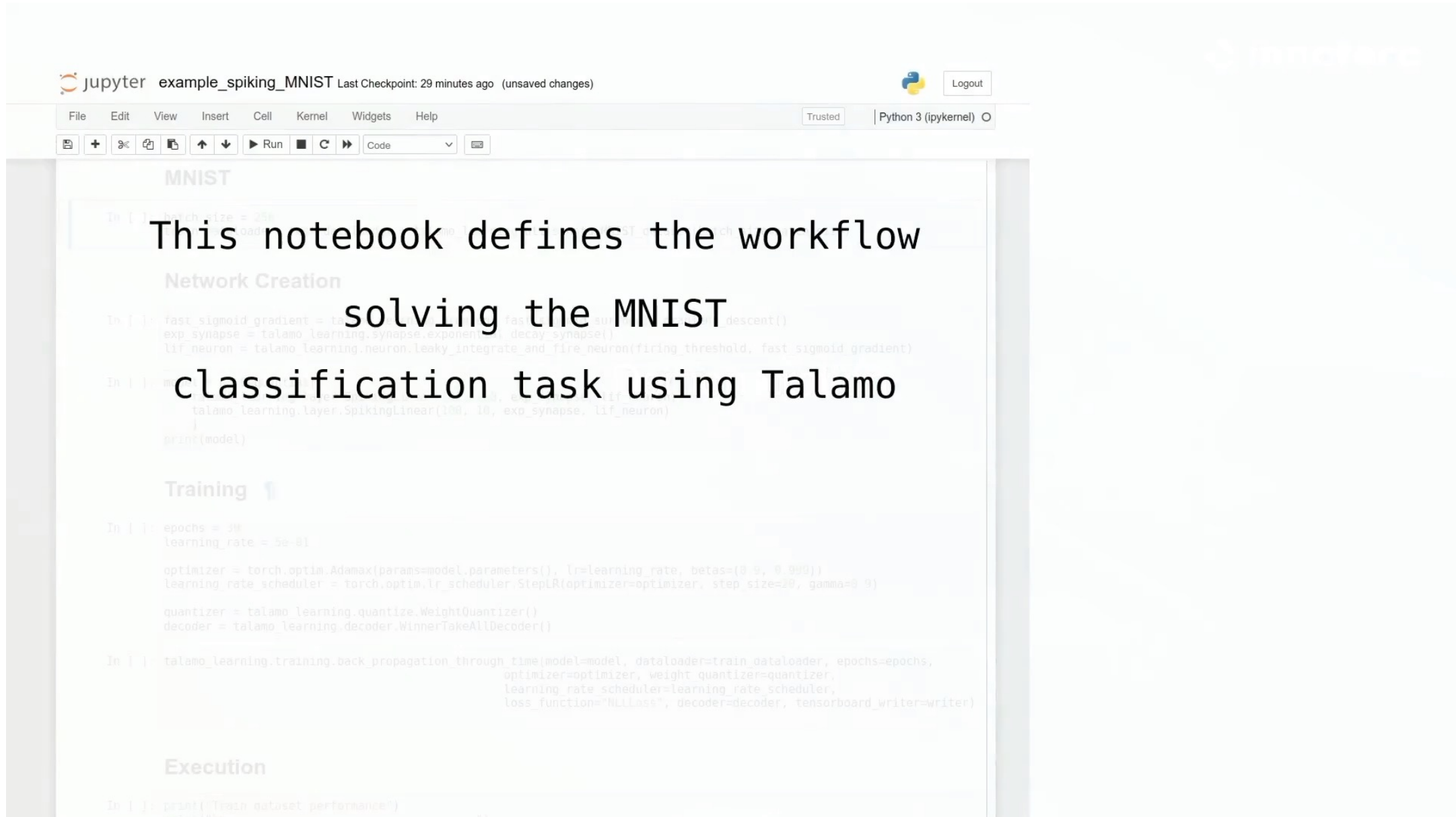
Includes everything required to build and train models

Identical to the PyTorch API – easy to adopt and use

Requires no knowledge of spiking neural networks



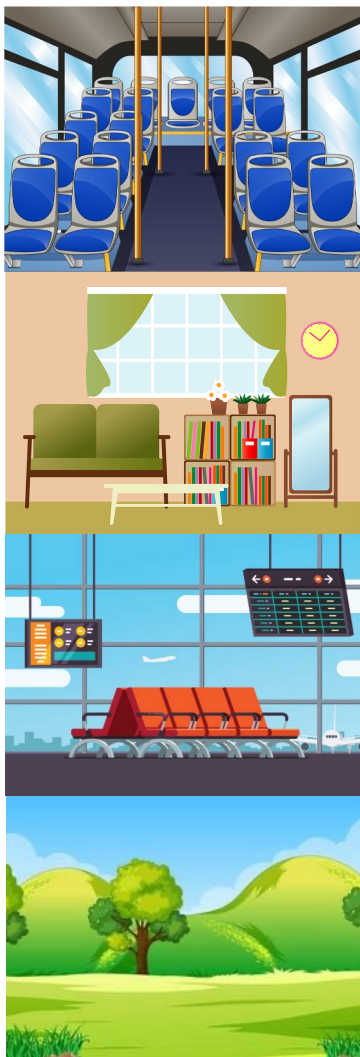
Building a spiking neural network in Talamo



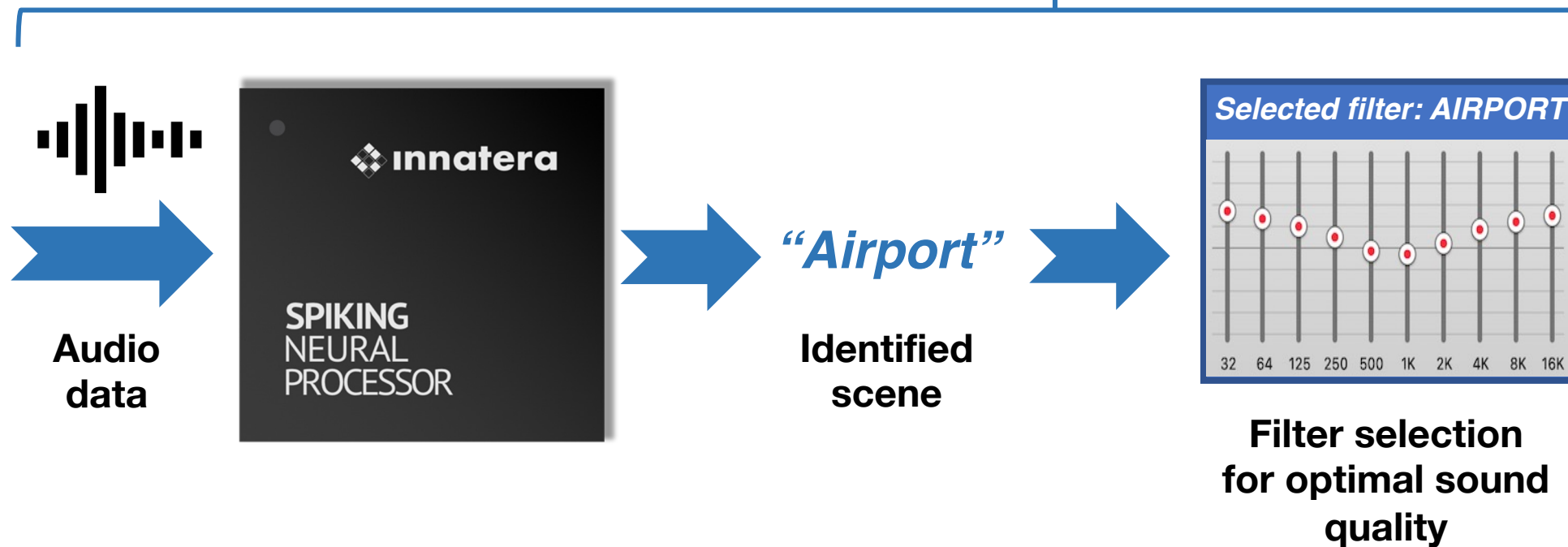
The screenshot shows a Jupyter Notebook interface with the following elements:

- Header:** "jupyter example_spiking_MNIST Last Checkpoint: 29 minutes ago (unsaved changes)" and a "Logout" button.
- Menu:** File, Edit, View, Insert, Cell, Kernel, Widgets, Help.
- Toolbar:** Includes icons for file operations, a "Run" button, and a dropdown menu set to "Code".
- Code Cell:** Contains Python code for MNIST classification using Talamo. The code is organized into sections:
 - Network Creation:** Defines a spiking neural network model using Talamo's classes: `fast sigmoid gradient`, `exp_synapse`, `lif_neuron`, and `talamo_learning.layer.SpikingLinear`.
 - Training:** Sets `epochs = 30` and `learning_rate = 5e-01`. It uses `torch.optim.Adamax` as the optimizer and `torch.optim.lr_scheduler.StepLR` for the learning rate scheduler. It also defines a `quantizer` and a `decoder`.
 - Execution:** A final cell with `print("Train dataset performance")`.

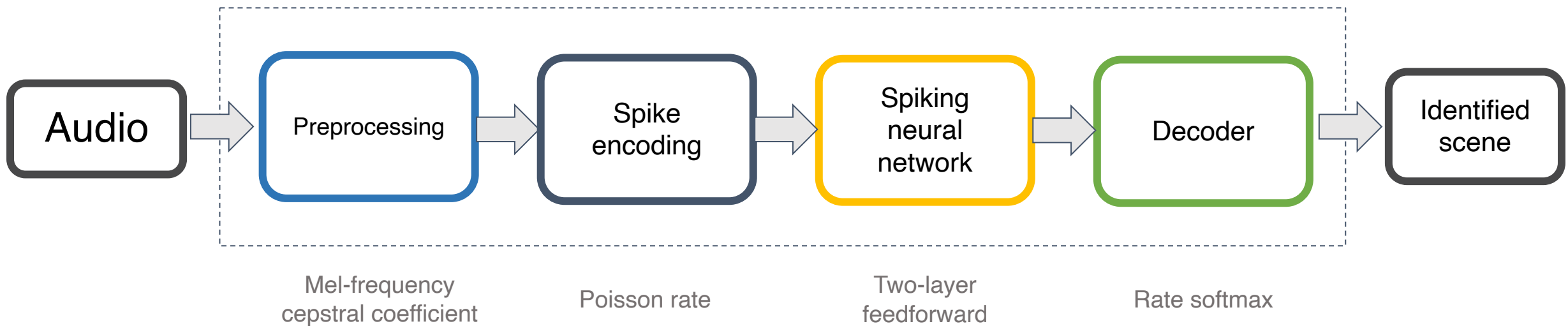
Always-on audio scene classification in hearables



Audio scenes



Solution pipeline



Audio scene classification (DCASE 2020)



Power
(total peak)
1.06 mW

Accuracy
~85%

Inference latency
~1 ms / 1 s

Model size
~3kB

Selected airport as scene

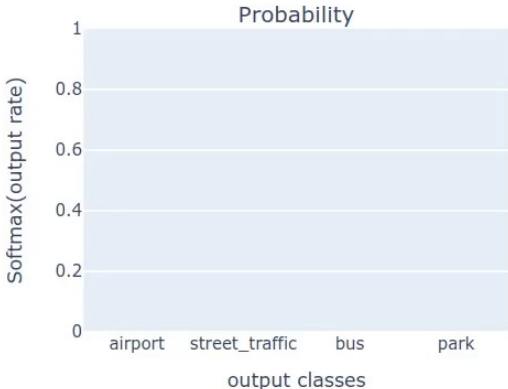
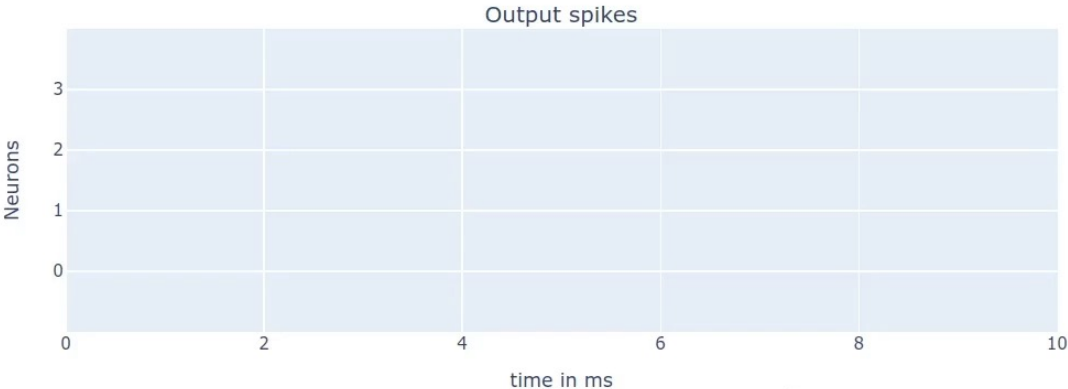
- airport
- street_traffic
- bus
- park

Audio files

- airport-barcelona-0-9-a.wav
- airport-helsinki-3-156-a.wav
- airport-lisbon-1122-40344-a.wav
- airport-paris-206-6240-a.wav

SEND TO HARDWARE

0:00 / 0:10



Identified scene as ...

- 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 $\sim 1\text{mW}$
 - an inference latency of $\sim 1\text{ms}$
 - with a code size of $\sim 3\text{kB}$
- 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

This presentation in this publication was presented at the tinyML[®] Summit (March 28 - 29, 2023). 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