AI4EO: from big to small architecture for deployment at the edge

Φ-lab Explore office @EOP/ESRIN
by Nicolas LONGEPE

in collaboration with Pierre Philippe Mathieu, Andrzej Kucik, Bertrand le Saux (Φ-lab), Massimiliano Pastena (ESTEC)
And industries: KP Labs (BEETLES), Agenium (CORTEX), Cosine and Univ. Pisa (Phisat-1), OpenCosmos, CGI, Ubotica, Simera CH Innovative, CEiiA, GEO-K, and KP Labs (Phisat-2)


.0 Stack to query our planet
EO
AI
ICE MELTING
ANOMALY IN PERMAFROST
SHIP DETECTED
FOREST LOSS
OCEAN BLOOM
OIL SPILL
EL NIÑO IMPACT
INSIGHT
FIRE ALERT
15

Observing System

and now AI@edge...

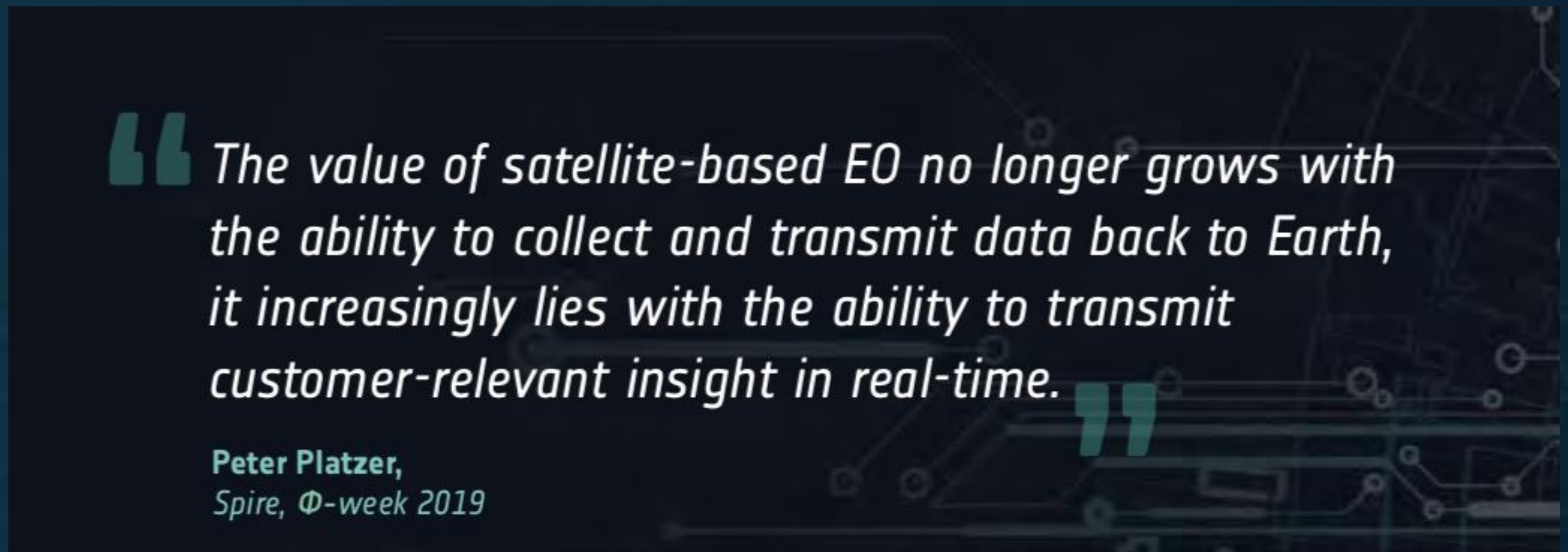
AI @ Edge: why?

Versatility

Enhanced security

High responsiveness

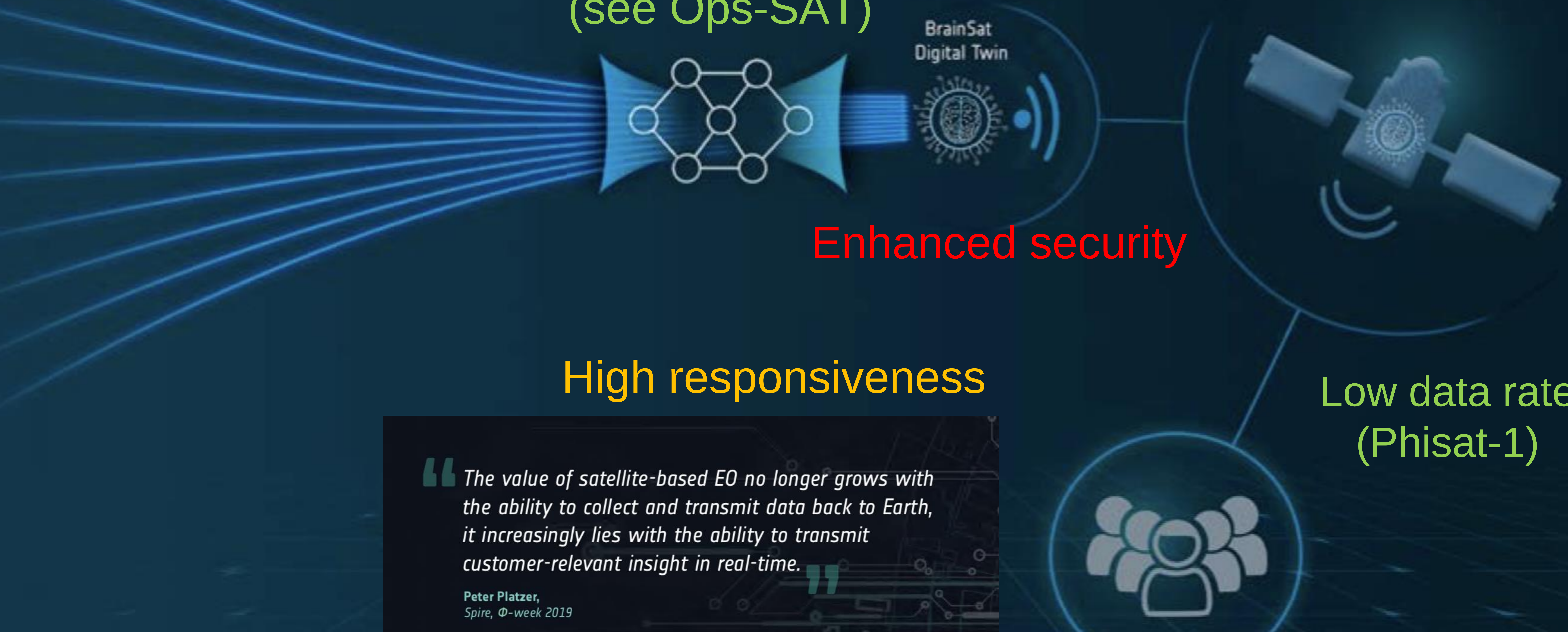
Low data rate


The value of satellite-based EO no longer grows with the ability to collect and transmit data back to Earth, it increasingly lies with the ability to transmit customer-relevant insight in real-time.

Peter Platzer,
Spire, Φ-week 2019

AI @ Edge: status as of 2021

Versatility
(see Ops-SAT)



Enhanced security

High responsiveness

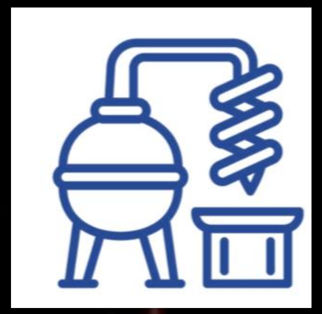
Low data rate
(Phisat-1)

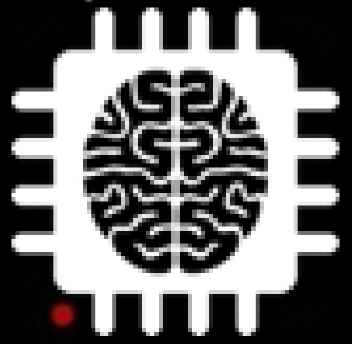
“The value of satellite-based EO no longer grows with the ability to collect and transmit data back to Earth, it increasingly lies with the ability to transmit customer-relevant insight in real-time.”

Peter Platzer,
Spire, Φ-week 2019

Φ-sat-1

Training on Ground
(S-2 mimicking)
Cloud detection





Inference in Space

Hardware

This project is **technology driven:**

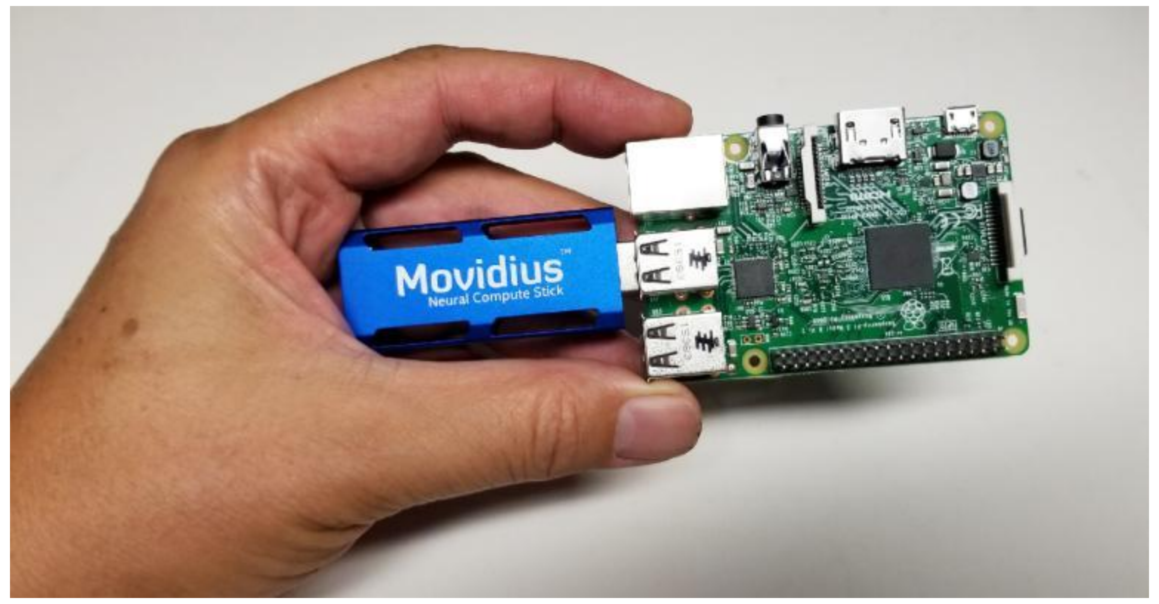
Visual Processing Unit (VPU)
Myriad-2

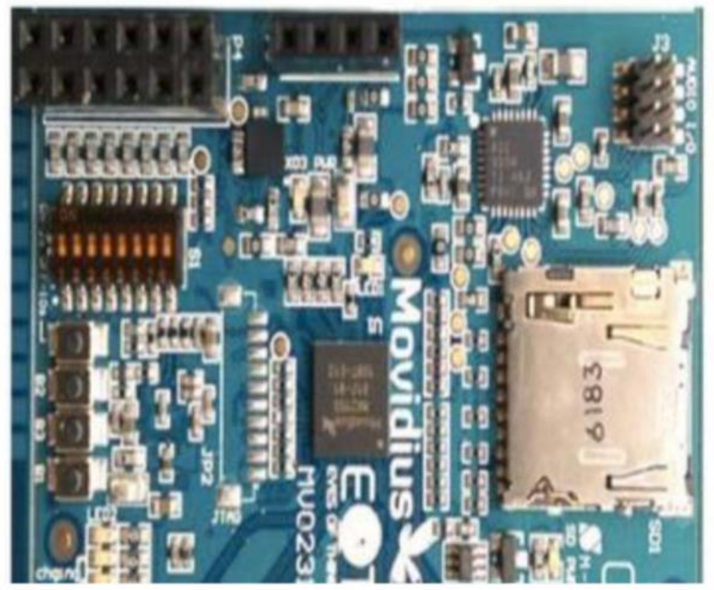
Hardware accelerator for Convolution Neural Networks (CNNs)

Fast: 1 TOPS

Low power: O(1 Watt)

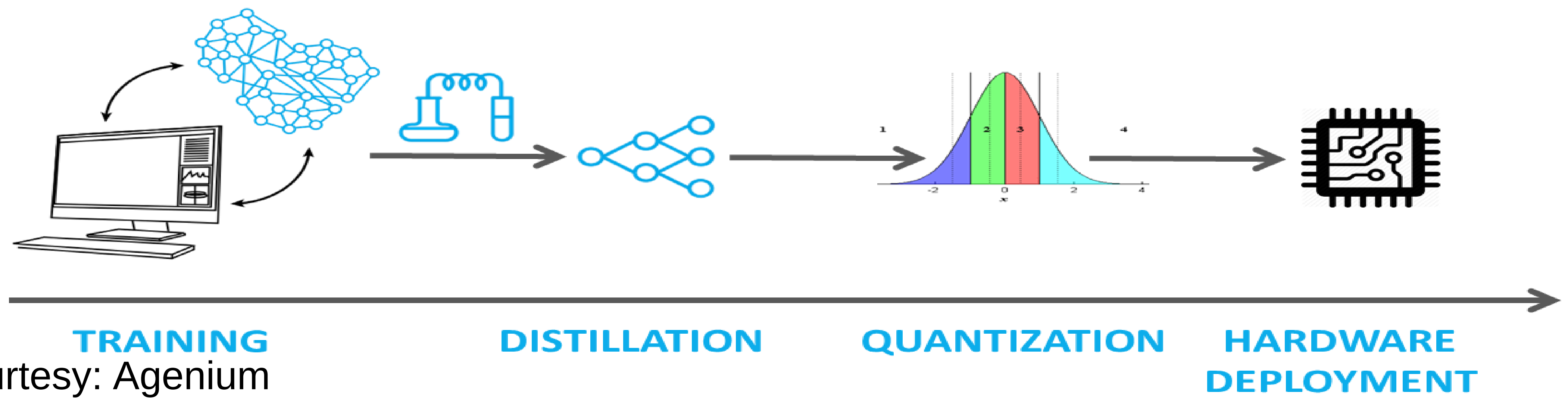
Tiny



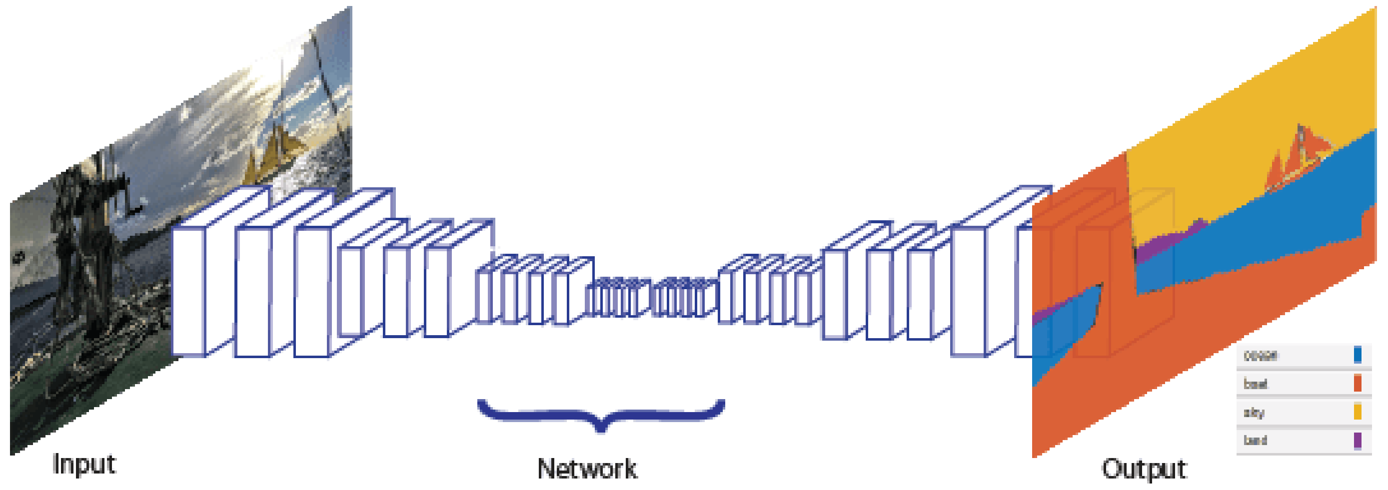


Process for DNN on-board deployment

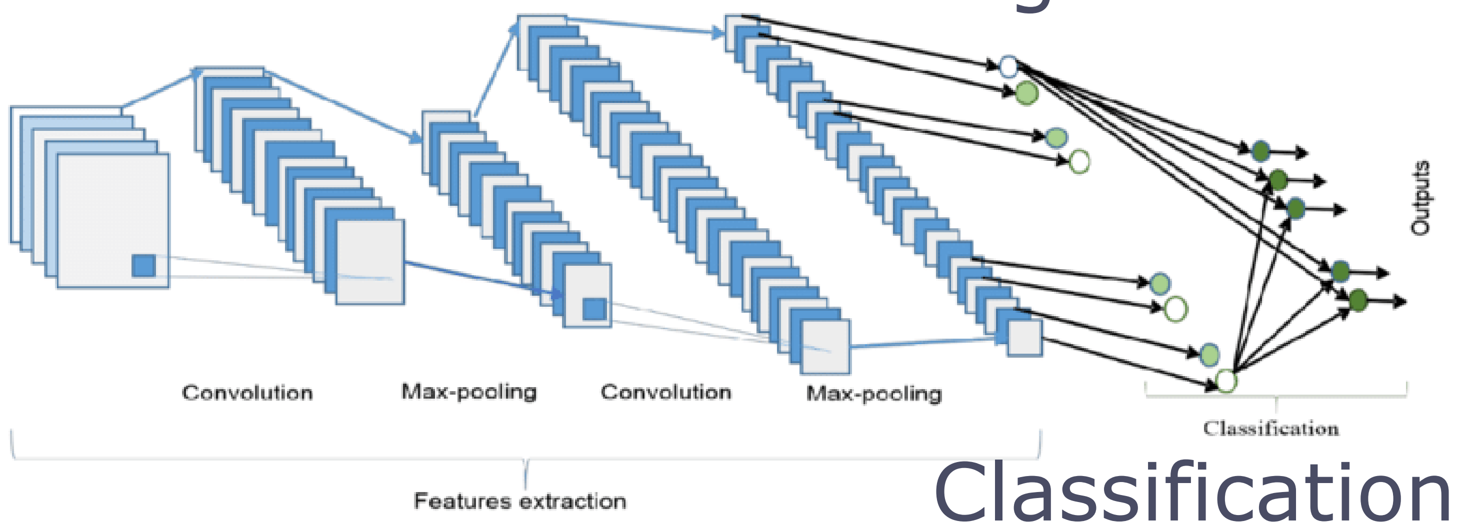
- **Training (in the cloud … in future maybe directly @edge):**
  - Database with Ground truth
  - How to synthetize EO data mimicking the performance of the upcoming payload?
  - Ensure robustness to all uncertainties (e.g. data augmentation)
- **Fitting to onboard requirements (throughputs, power consumption…)**
  - Which NN architecture? Network Architecture Search (NAS), Knowledge Distillation
  - Network Optimization: pruining, quantization
- **Hardware deployment**



Courtesy: Agenium



Segmentation DNN



Classification DNN

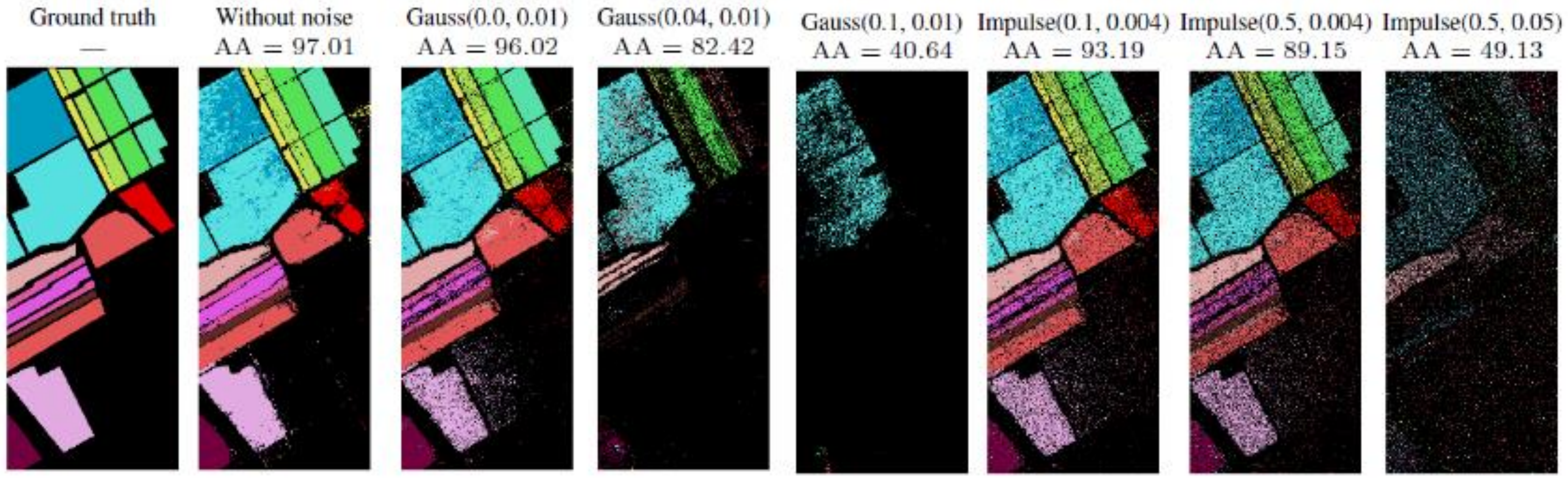
DNN training: robustness to noise

Hyperspectrometers:

- Presence of various kinds of noise (thermal noise, sensor saturation)

⇒ Degrade classification accuracy, a potential obstacle for deploying onboard satellite

⇒ Denoising and/or better regularization may help obtain high-quality classification.



J. Nalepa, M. Stanek, ``Segmenting hyperspectral images using spectra convolutional neural networks in the presence of noise,,' Proc. IGARSS 2020, pp 1-4, 2020.

⇒Require robustness to SNR, but also data mis-calibration, inter-band displacement, band-to-band registration

Resource-frugal quantized CNN: simplification

Knowledge Distillation :

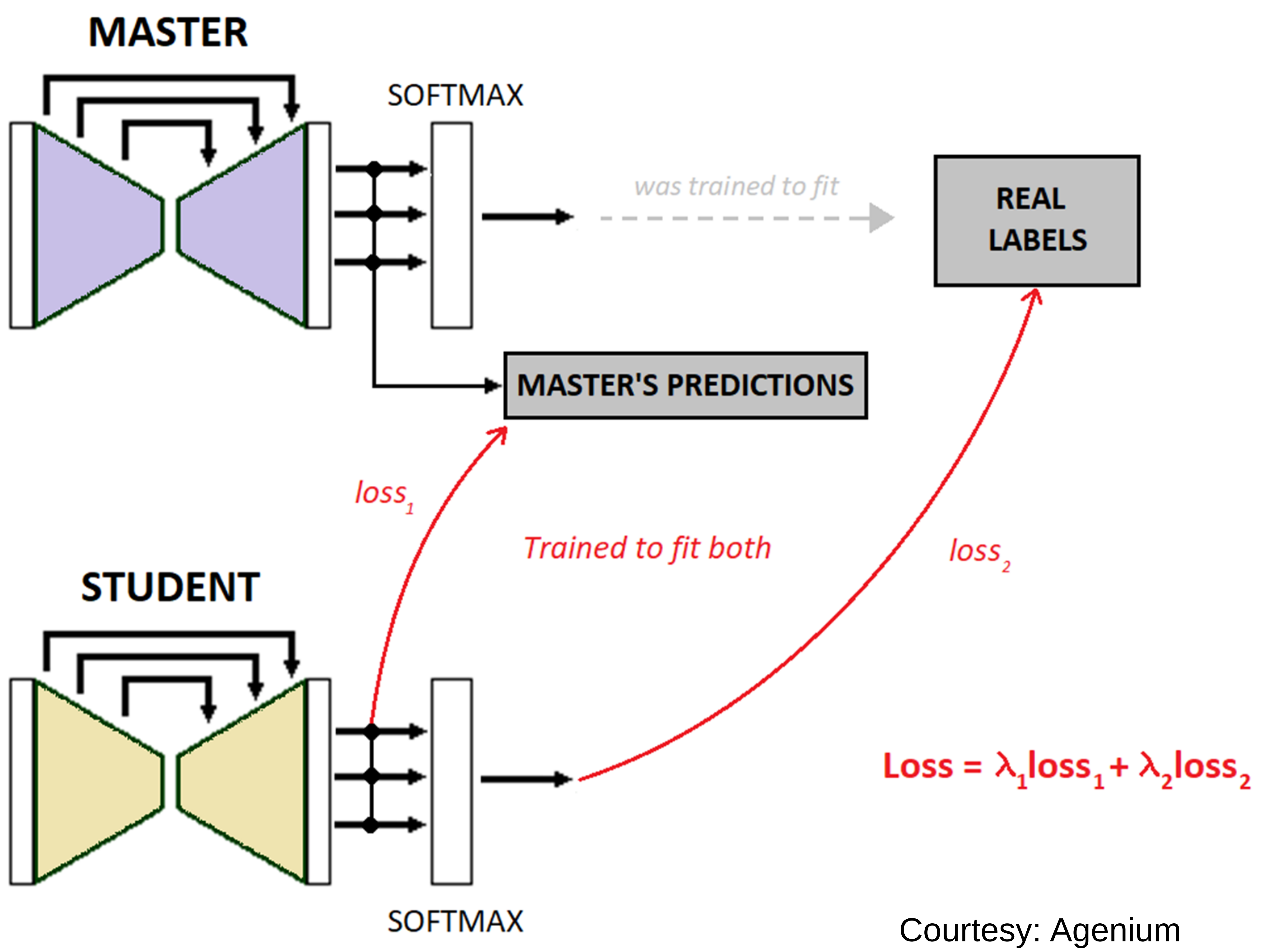
Process of transferring knowledge from a large model to a smaller one, without loss of perormance

Input: high-performance DNN

Output: DNN with small target architecture

Aim:

- Reduced number of parameters for the distilled DNN



Courtesy: Agenium

⇒ Reduction of x52 shown (136M -> 2,6M) without performance loss

⇒ Another process: Network Architecture Search (NAS)

Resource-frugal quantized CNN: quantization

Network quantization:

Reduce float32 (Model trained) -> Int8 (Quantized Model)

-> Better memory footprint (div. by 4)

-> Better throughput

-> Better energy efficiency

Table 11 : Comparison between the best distilled model for the Boat Detection DB, before quantization (32bits float) and after quantization of weights and activations with the same number of bits, spanning from 2 to 8. Bold values measure a loss smaller than 1% of the result obtained with the distilled model.

Courtesy: Agenium

Model 379	32bits Float	8bits Int	7bits Int	6bits Int	5bits Int	4bits Int	3bits Int	2bits Int
F1 class 0	0.977100	**0.973312**	**0.967102**	0.926616	0.839211	0.805933	0.000000	0.000000
F1 class 1	0.923300	0.913118	0.891938	0.800595	0.673939	0.643182	0.375813	0.375813

Vitis AI (Xilinx) Quantization

- Symmetric (Signed Int)
- Static (weights) & Dynamic (Activations)
- Required ~100-1000 images to calibrate activations

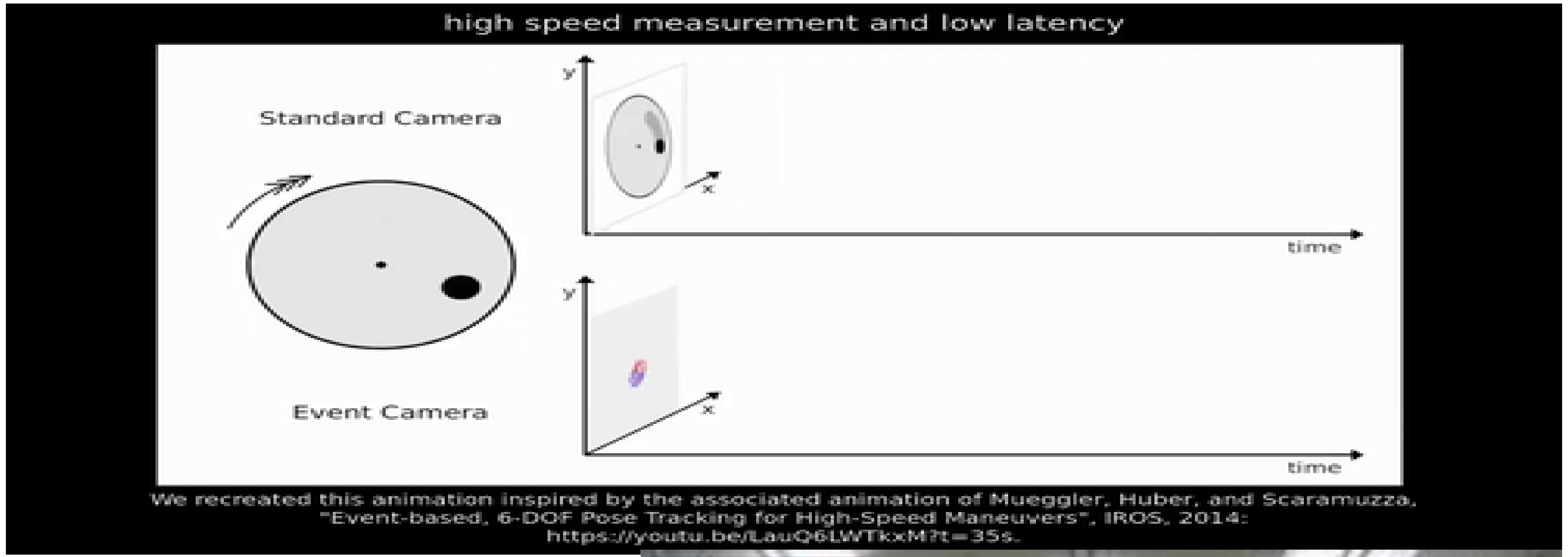
⇒ Reduction of x4 without performance loss

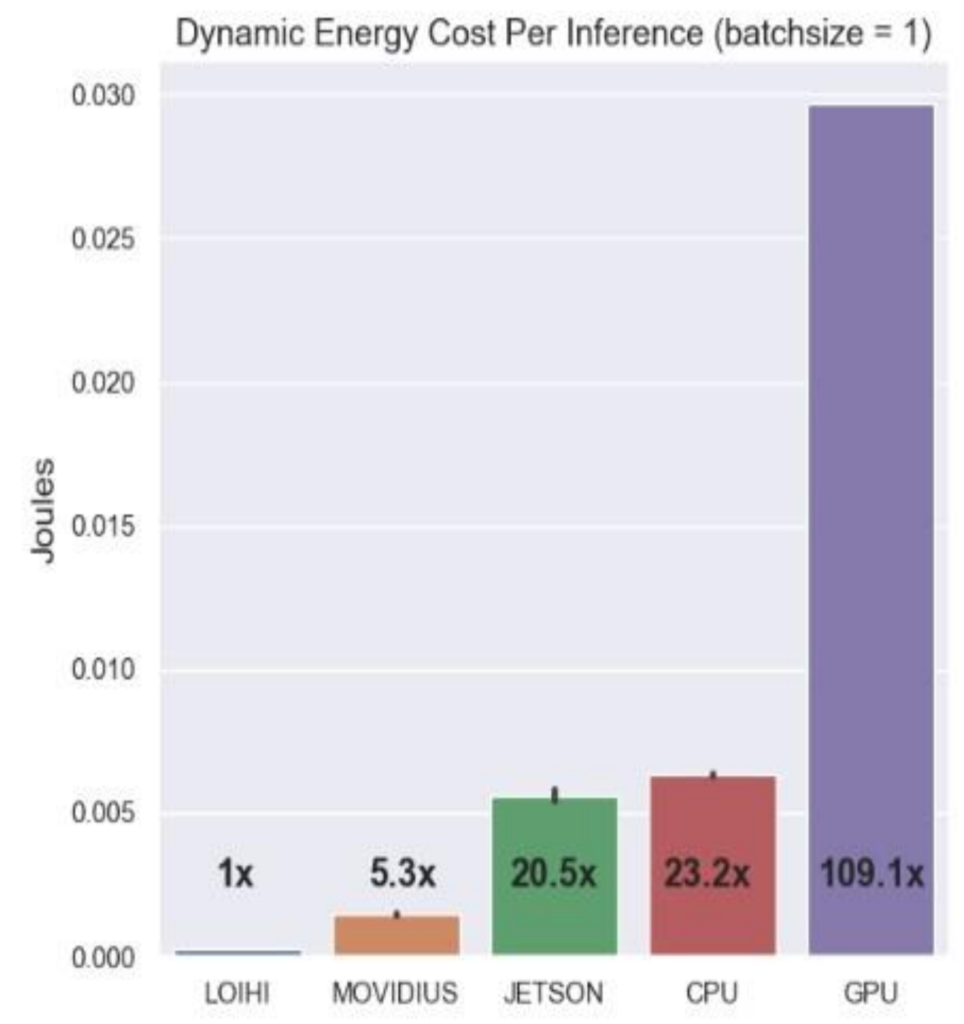
A short term prospect:
Φ-sat-2

A game-changing Earth Observation CubeSat platform in space capable of running AI Apps that can be developed by its users, then easily deployed in the spacecraft, and operated from ground.

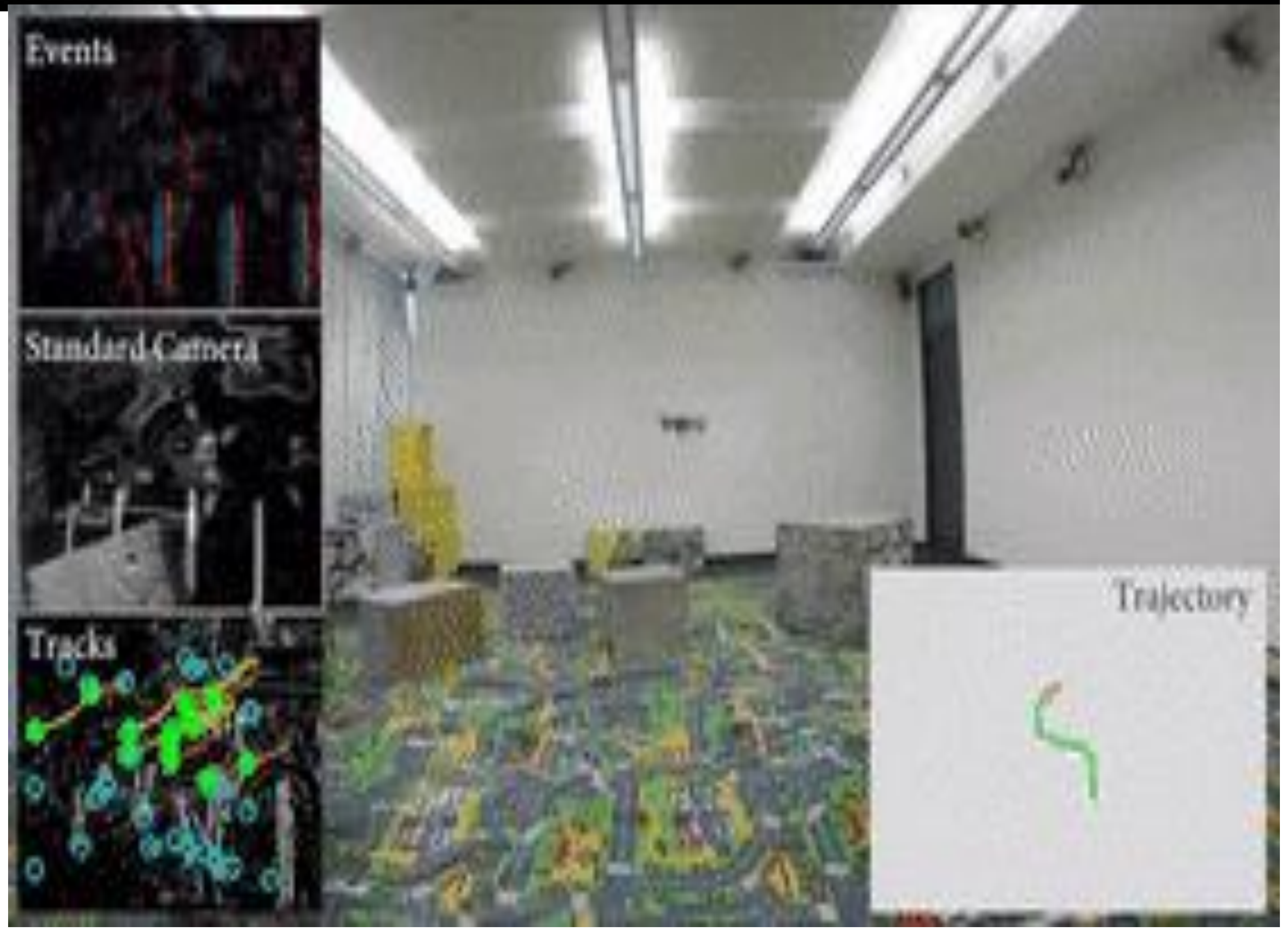
An example of longer term perspective: Neuromorphic computing for EO

Andrzej Kucik (Φ-Lab), Gabriele Meoni (ACT)





P. Blouw, X. Choo, E. Hunsberger Ch. Eliasmith, *Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware* (2019)



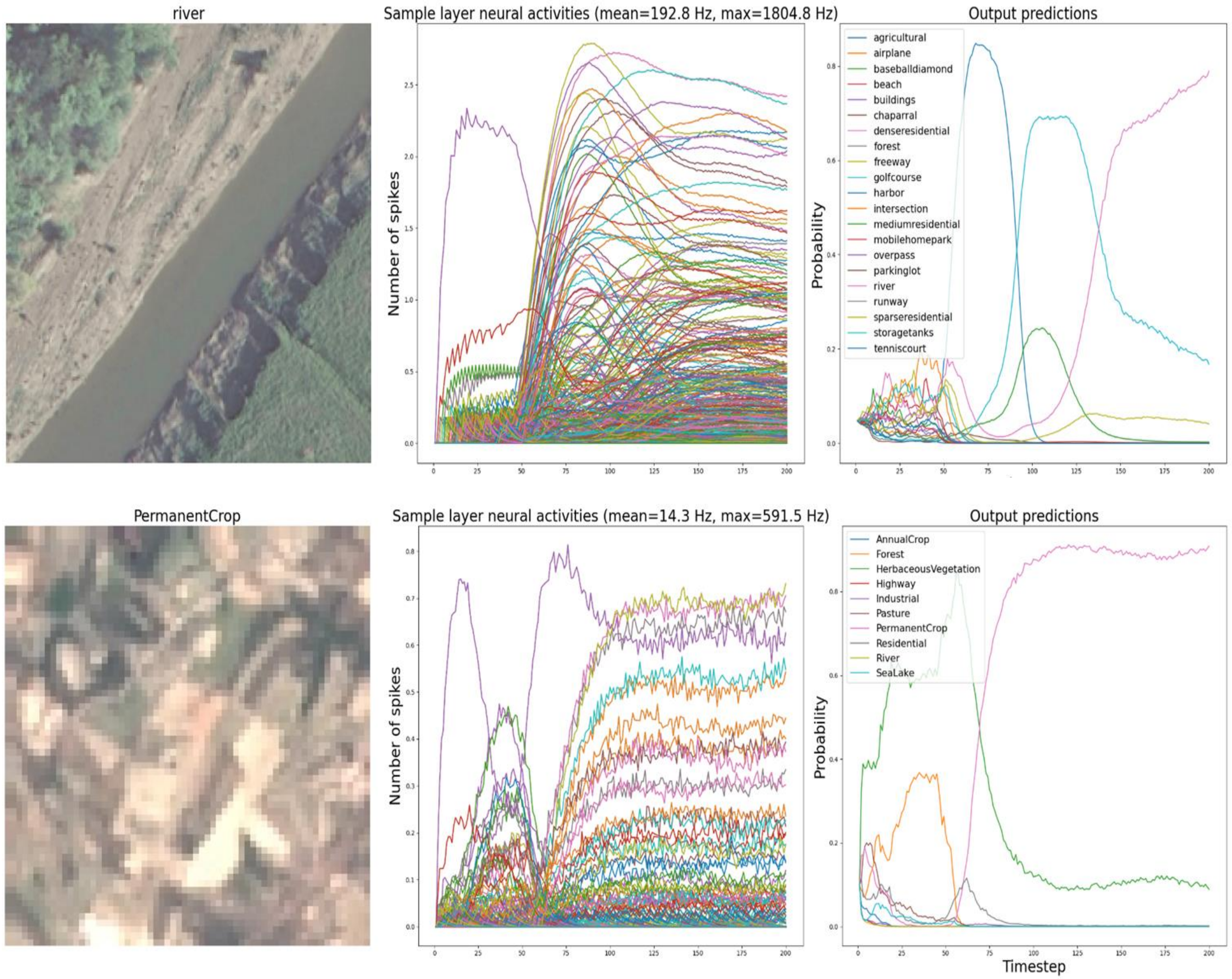
ETH Robotics and Perception Group

Objectives

1) Establish the feasibility of deploying the dynamic vision sensors to collect geospatial data:
   - Low latency
   - High dynamic range
   - Low power
2) Establish the feasibility of processing the geospatial data on neuromorphic chips/hardware platforms:
   - Asynchronous and massively parallel
   - Neuronal state
   - Low power
3) Design neuromorphic algorithms (spiking neural networks) for solving the Earth observation problems

An example of longer term perspective: Neuromorphic computing for EO

Andrzej Kucik (Φ-Lab), Gabriele Meoni (ACT)



Data

- UC Merced (21 classes, 100 256×256 examples each).
- EuroSAT RGB (10 classes, 27,000 64×64 images).

Methods

- Take one of the standard networks (possibly pre-trained on ImageNet).
- Train on the geospatial data.
- Convert to SNN.
- Replace elements incompatible with SNN.
- Optimise wrt accuracy and energy usage.

Results

- With the right parameters, the drop in accuracy is not very significant (~2%).
- 16x less energy for SNN on neuromorphic hardware than for ANN on a GPU.

"**Technology**… is only a **magnifier** of **human intent** and capacity. It is not a **substitute**."

Kentaro Toyama
Geek Heresy: Rescuing Social Change from the Cult of Technology

Φ-lab team
Nicolas LONGEPE
Nicolas.Longepe@esa.int