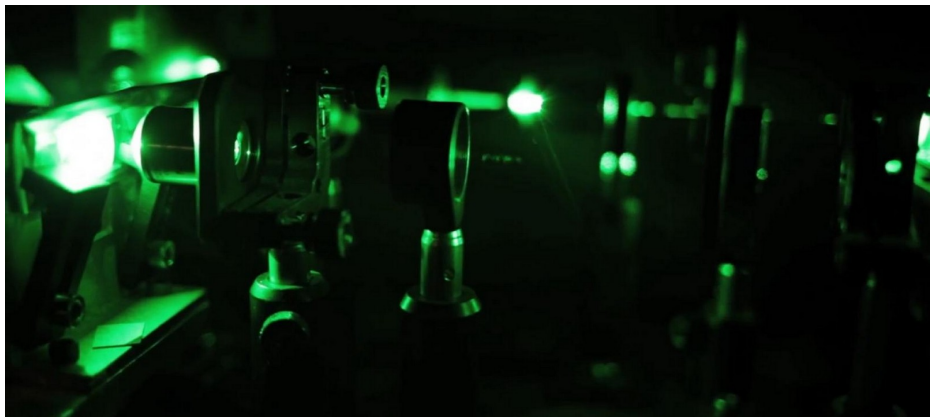
End to End Learning with an Optical Processing Unit

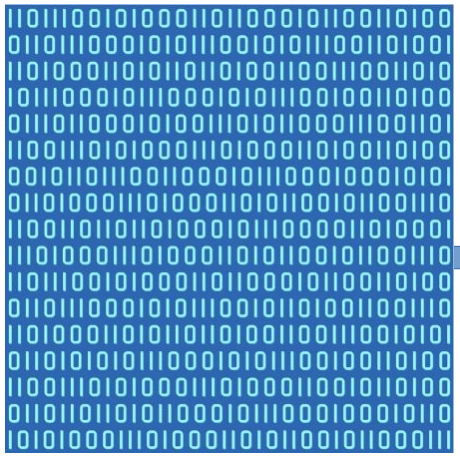
Laurent Basara (LRI-Orsay)
Biswajit Biswas, Aishik Ghosh, Max Marly, **David Rousseau** (IJCLab-Orsay)
Amélie Chatelain (LightOn)

ACAT 2021, Daejeon, South Corea

Optical Processing Unit

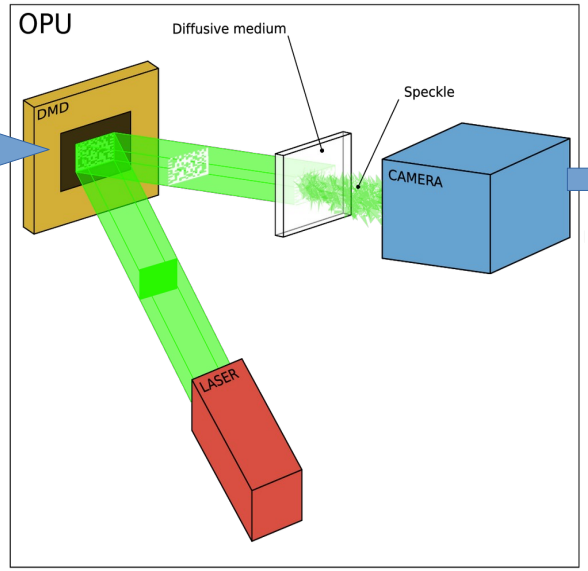


https://docs.lighton.ai/notes/opu.html

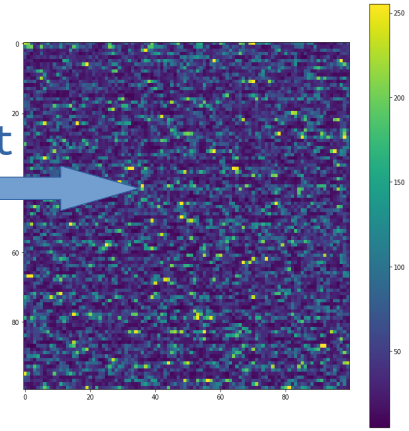


1 M bits vector X

OPU in

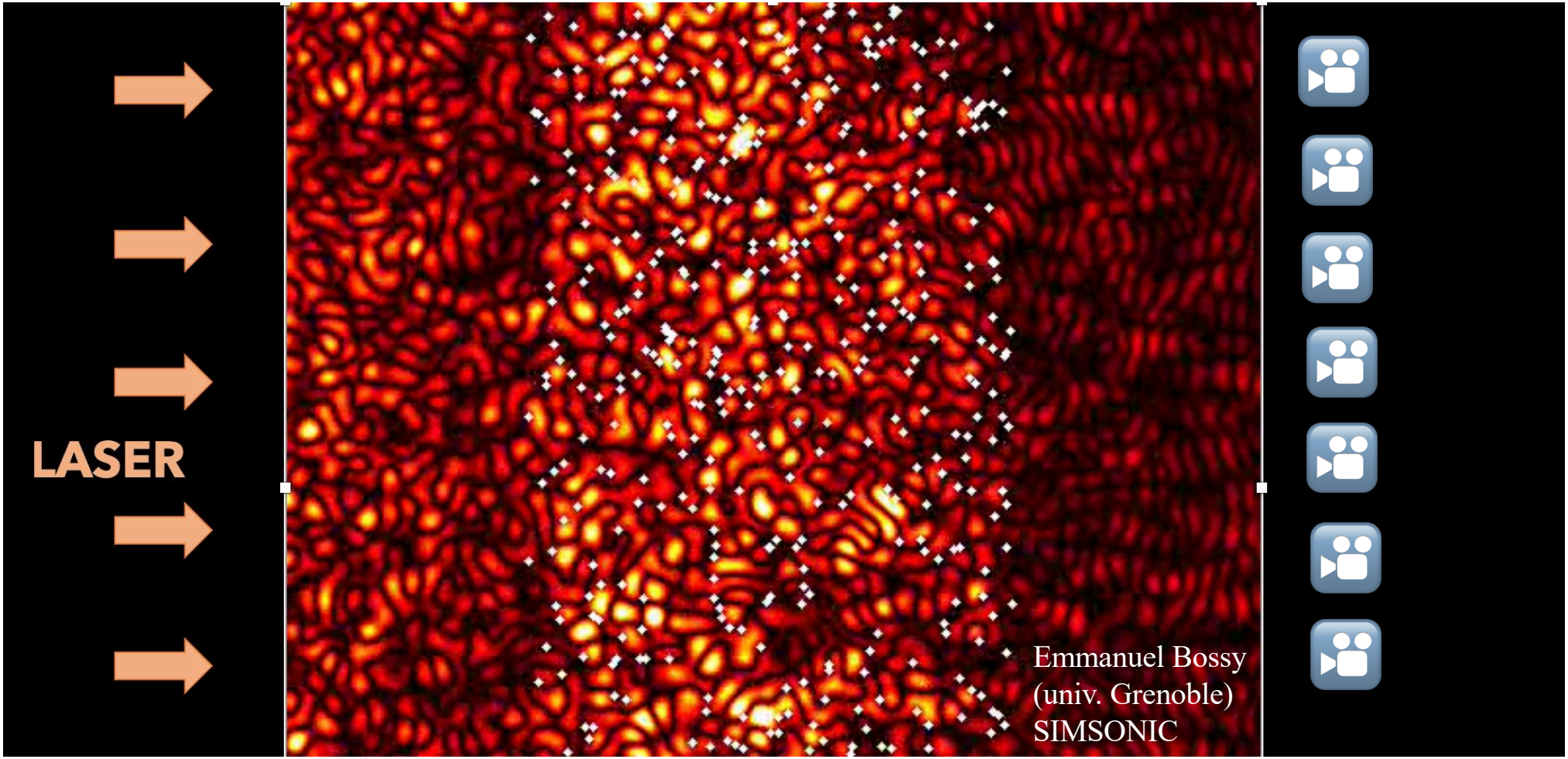


OPU out

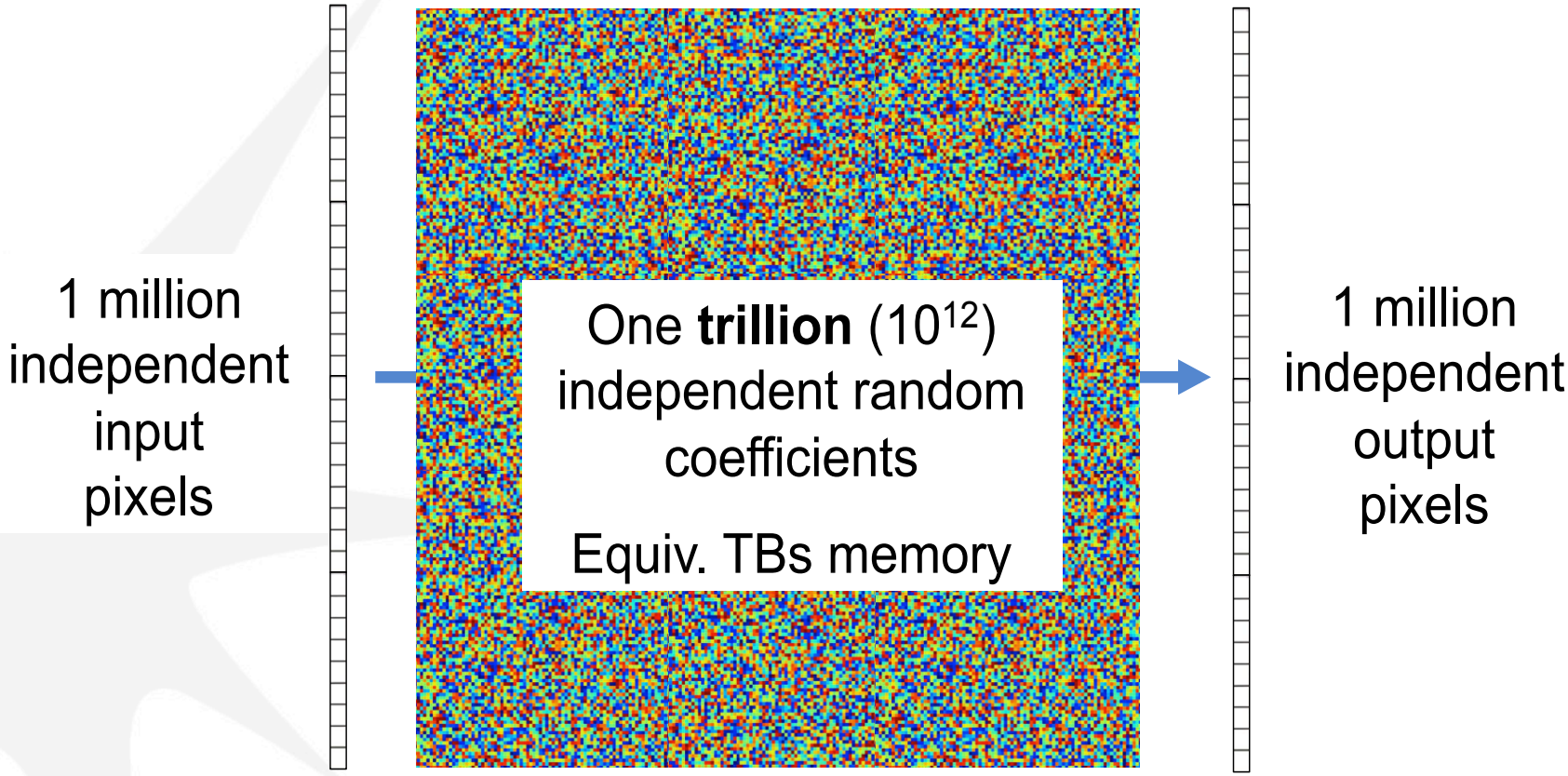


1 M
Random features

Physics principle



Random Matrix mult. Through light scattering



The OPU performs **Random Projections** in the analog domain

input vector x → *output vector* y = $|Hx|^2$

element-wise $|.|^2$ nonlinearity

with H a complex random iid matrix

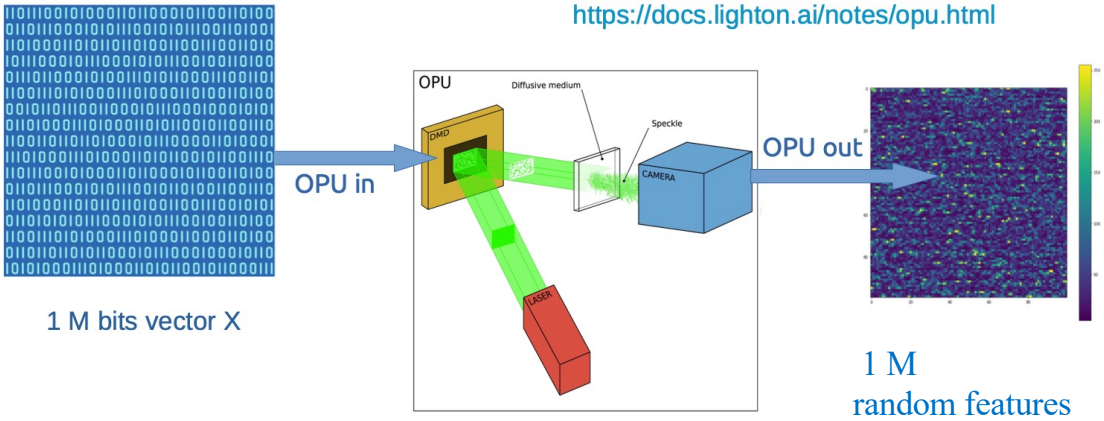
LARGE	&	FAST
H of size higher than $10^6 \times 10^6$ (TBs of memory)		kHz operation → 4.10^3 such multiplies / s

Equivalent **4,000 TOPS * @ 30 W**

* Analog - non Von Neumann - OPS not directly comparable to flops

More OPU factoids

https://docs.lighton.ai/notes/opu.html



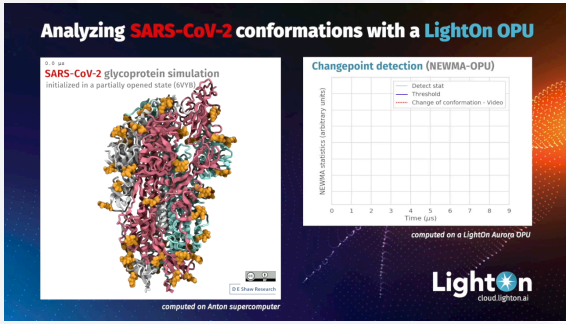
- OPU equivalent to a random matrix multiplication. Matrix is fixed. Unique to an OPU.
- Any input pixel is likely to be correlated to any output random feature
- Neighbouring information is lost
- All 1M output random features are equivalent. In practice we use 10k, 20k, 50K,…
- More : LightOn White paper , arXiv:1910.09880 « *Kernel computations from large-scale random features obtained by Optical Processing Units* »

The physical device

- Fits in standard 2U rack. Associated to regular Intel CPU as front-end
- Low consumption : 30 W
- Some units installed at french computing centers (OVH, Jean Zay @ IDRIS), free **(**shared) access to researchers. Also for sale
- pip install lightonml (on github). Helper library (sk-learn aware) to talk to OPU and much more (e.g. can simulate a small OPU) Very good support.

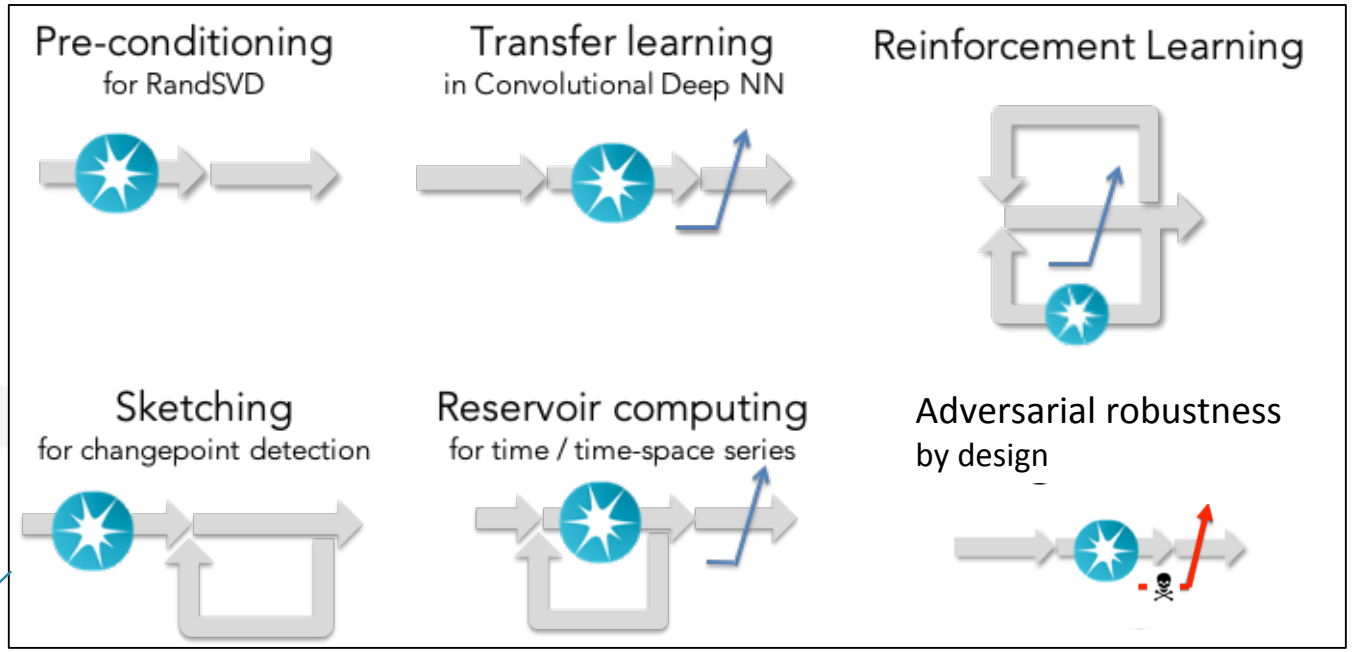
What to do with an OPU ?

Different kind of large scale Random Algorithms

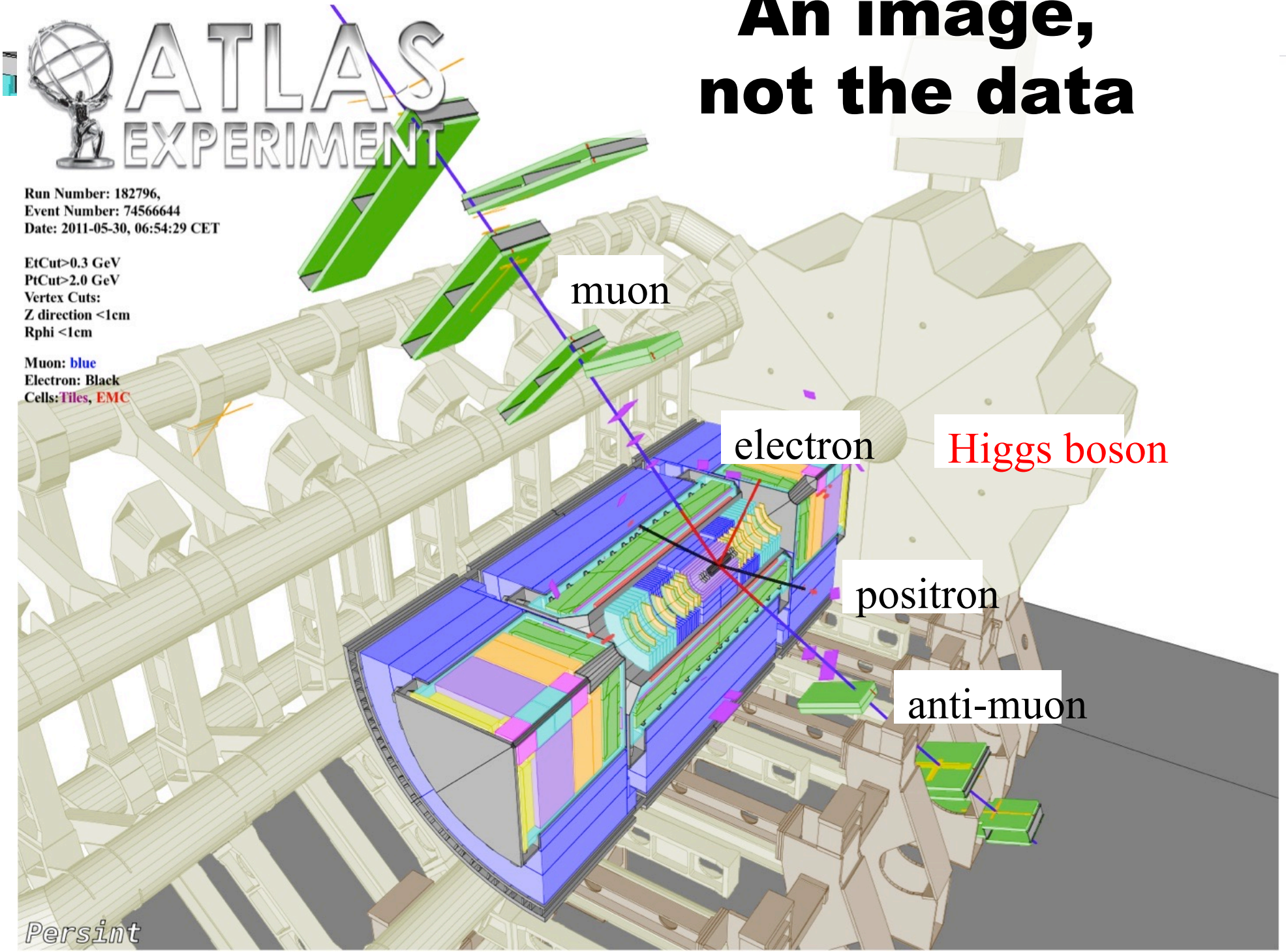


Accelerating SARS-COv2 Molecular Dynamics Studies with Optical Random Features

Amélie Chatelain
LightOn ML R&D engineer



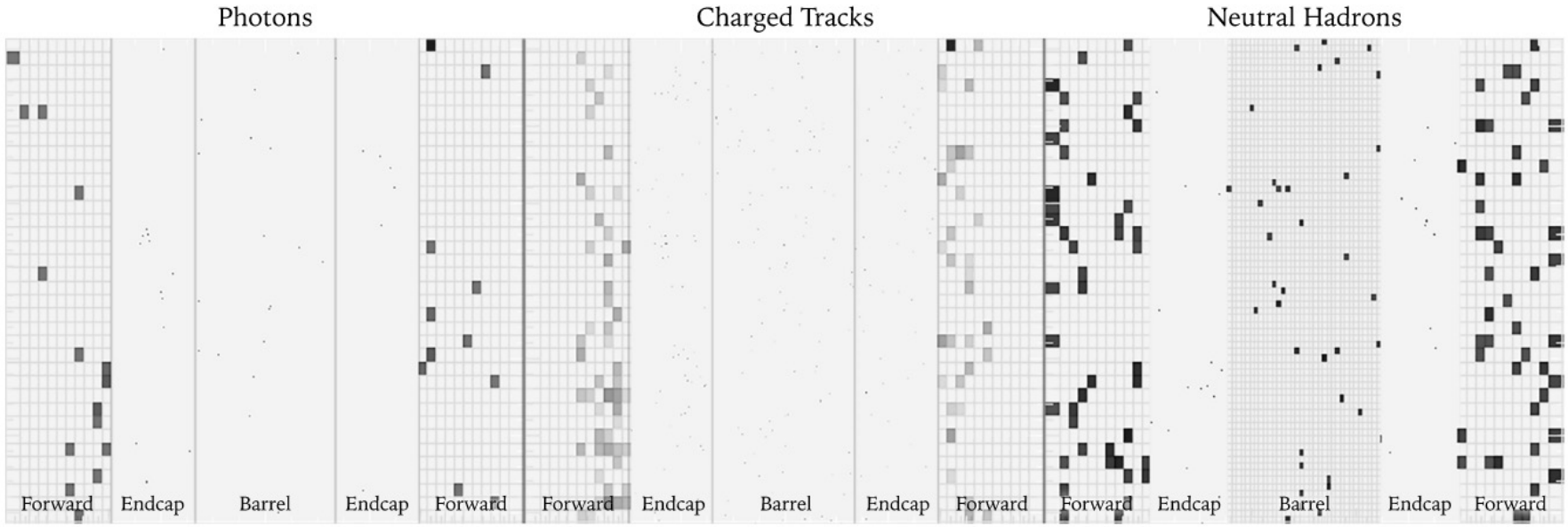
An image, not the data



Calorimeter end-to-end classification

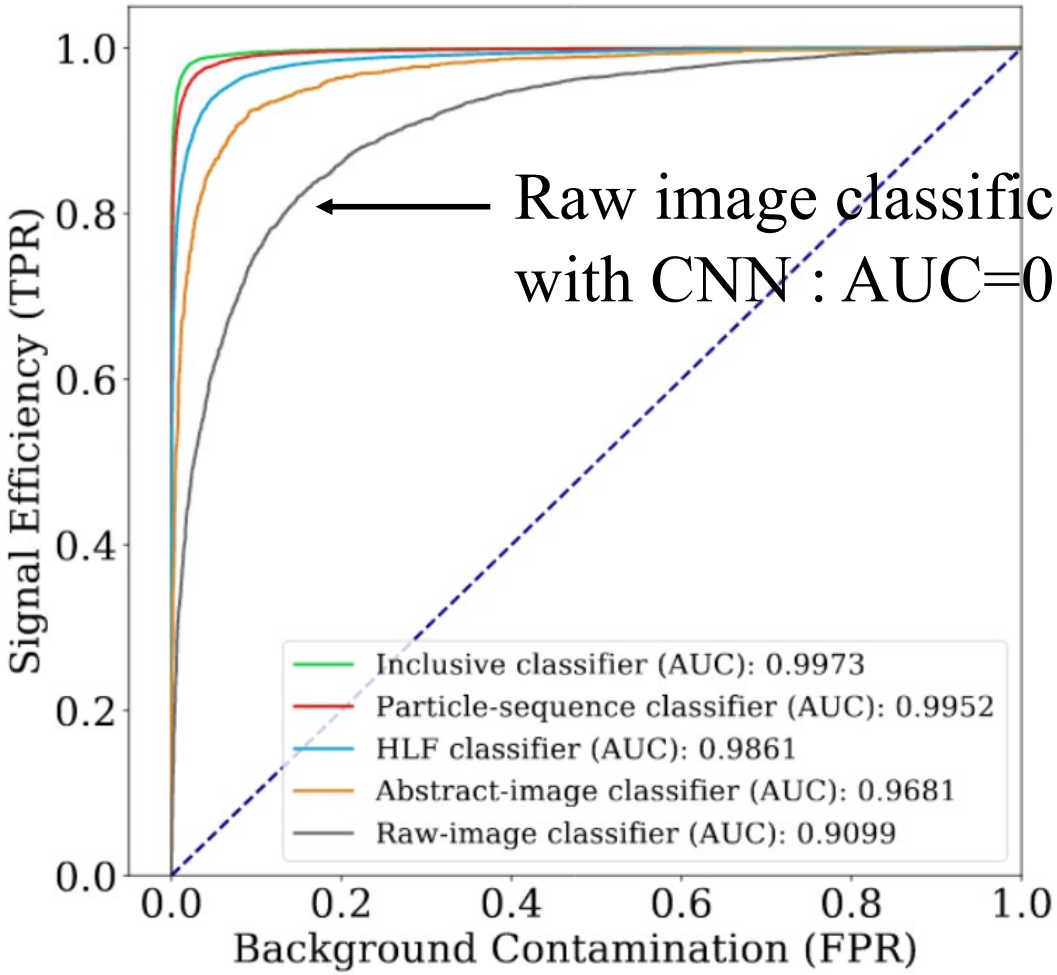
Calorimetry classification

- Dataset from Comput Softw Big Sci (2019) 3: 12 1807.00083 T. Nguyen et al *Topology classification with deep learning to improve real-time event selection at the LHC*
- Classification of 3 type of (Delphes with CMS card) simulated events : top pairs, W, and QCD (main background)
- Data: 3 channels, different granularity➔much more image-like (40.000 cells)
- Preselection with one high-Pt lepton : W 63.5%, QCD 36.2%, ttbar 0.3%
- ➔task is to select ttbar vs W+QCD



Reference

- ROC curve from reference paper

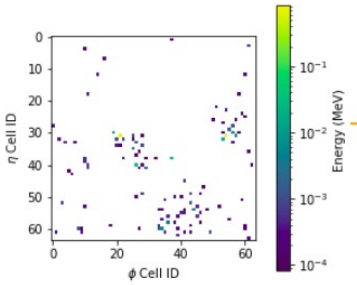


Raw image classification with CNN : AUC=0.91

- (other curves obtained using MET, jet…., meaning significant processing)

Goal is now to see whether this end to end classification can be done with OPU

OPU Workflow



Data

STEP 1

Convert to Binary

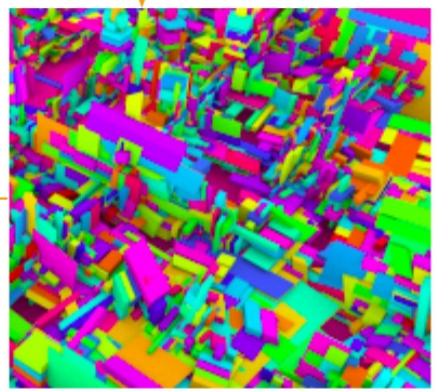
0
1
1
.
.
.
1
0
0

Send to LightOn's server

LightOn
We bring Light to AI

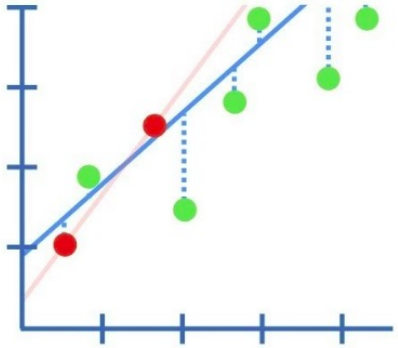
STEP 2

Get random features



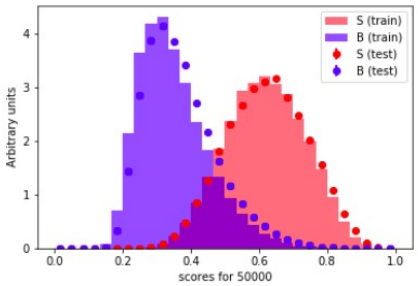
STEP 3

Train Linear classifier



STEP 4

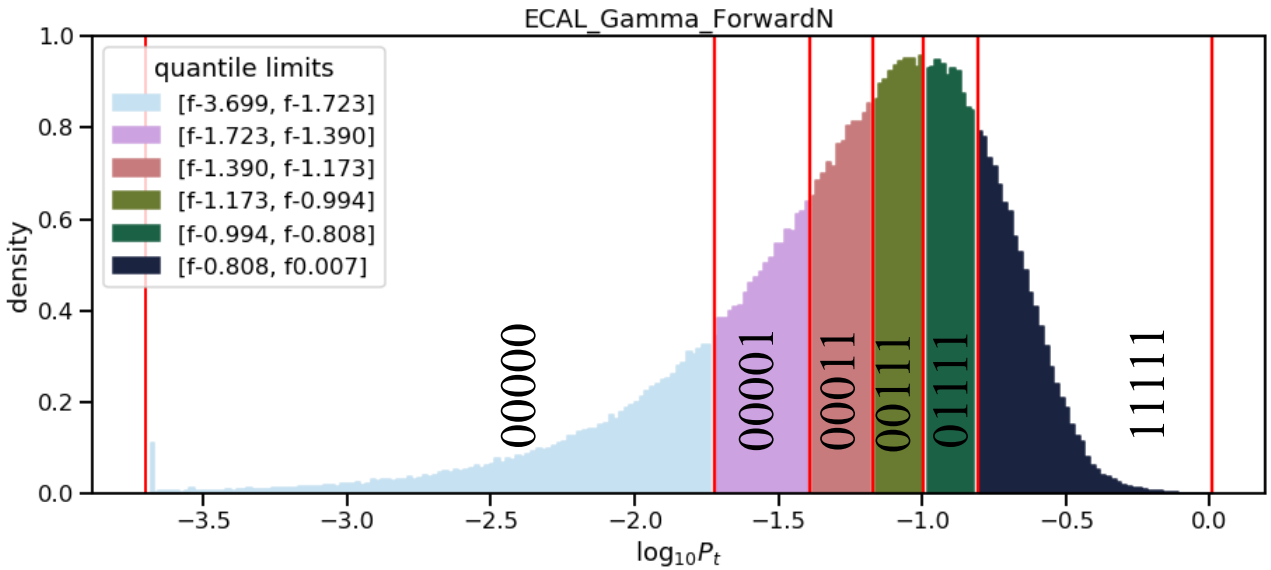
Make predictions for new data



Score

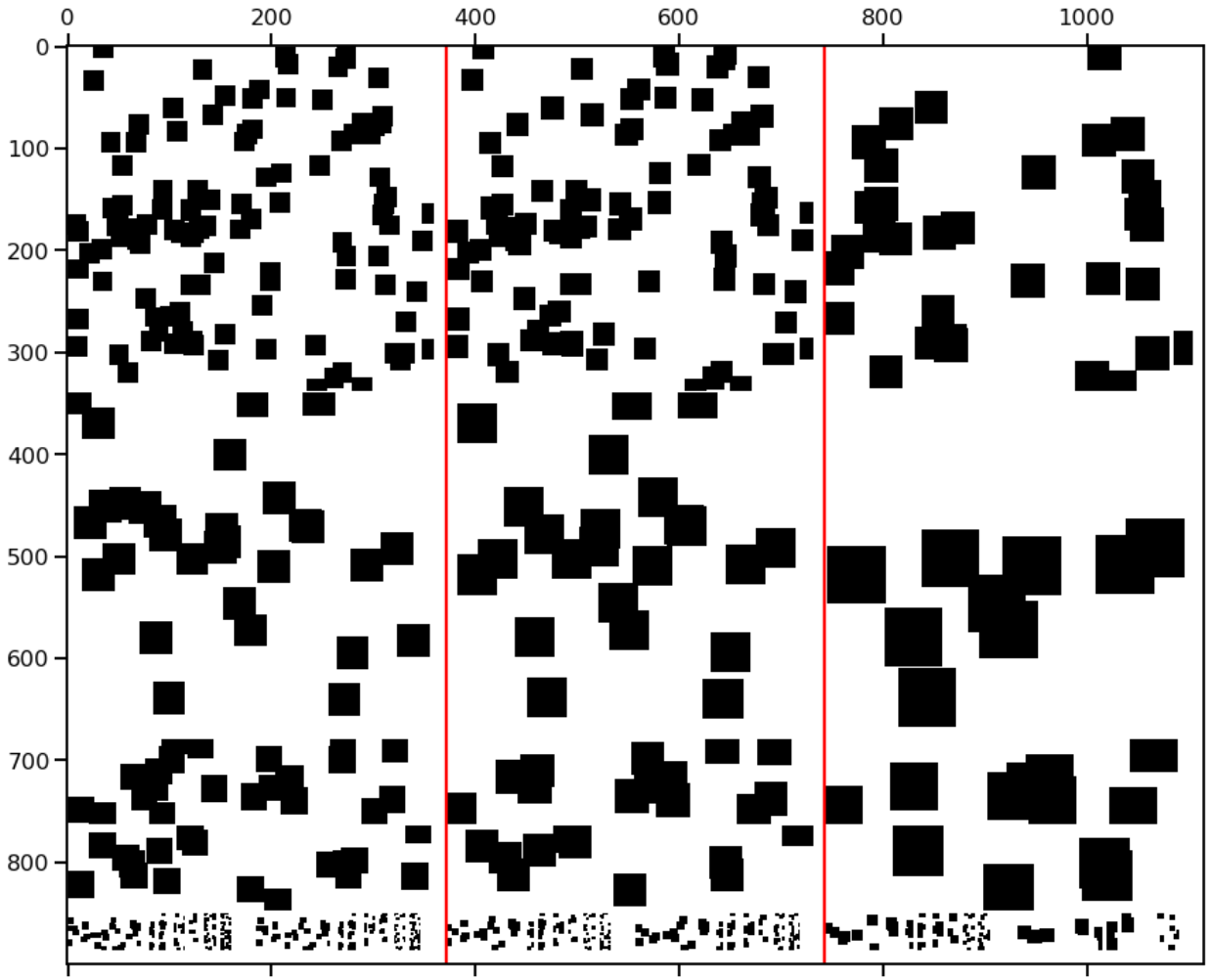
Quantization

- Quantization of the logarithm of energy, separately for each region / channel
- Each cell/channel represented by 5 «bits», one for each quantile : 00000,00001,00011,00111,01111,11111
- Reminder : each bit treated independently through the OPU



Binarisation

- Image fed to the OPU



End to End Learning with an Optical Processor Unit, David Rousseau, ACAT 2021

Linear model

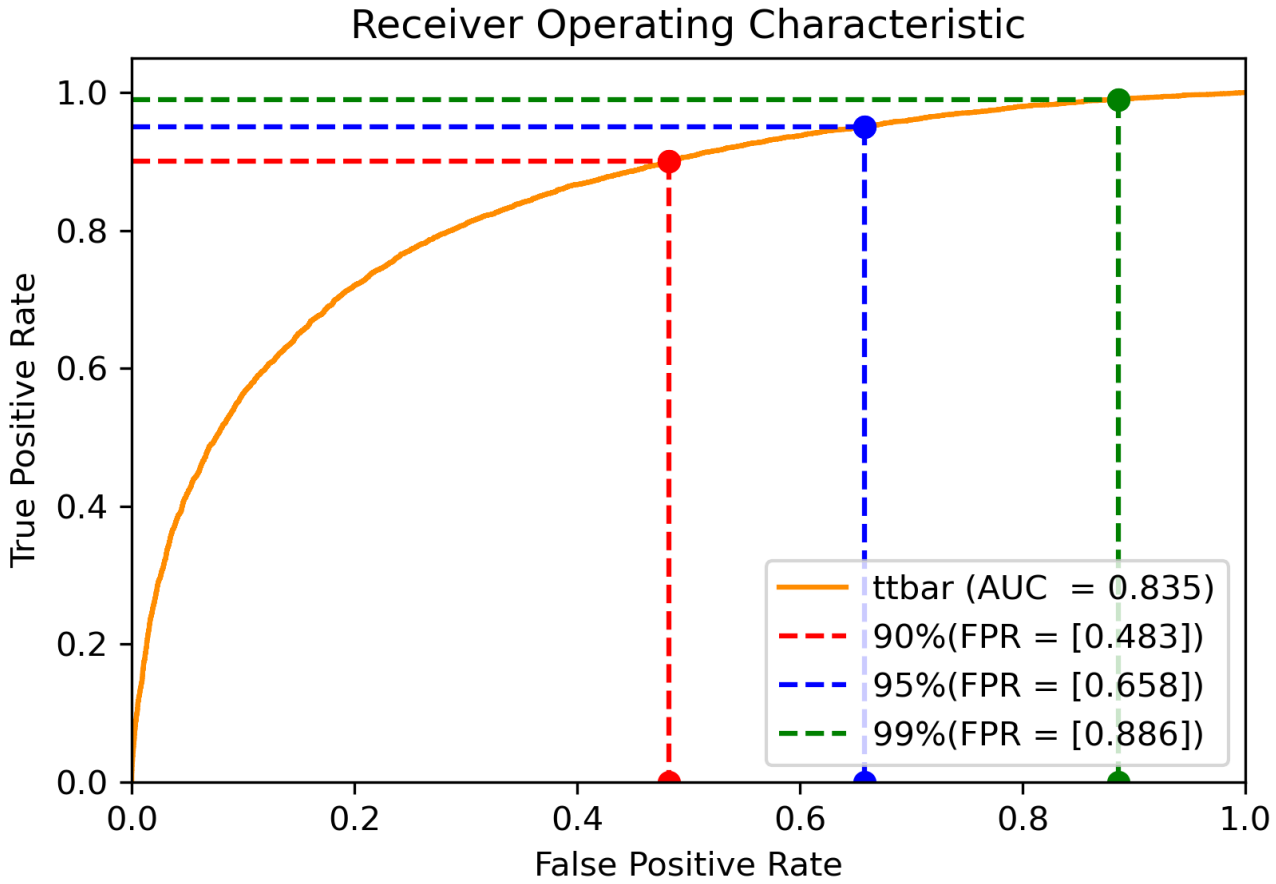
target:0 or 1 regularisation

$$\min_{w} \sum_{i} ||X_i w - y||^2 + \alpha ||w||^2$$

- Solve the linear system with regularisation term
- Sklearn RidgeClassifier : build matrix N_{events} x $N_{features}$, fine for small values
- Sklearn SGD : Stochastic Gradient Descent , much better scaling (not surprising…)

Performance

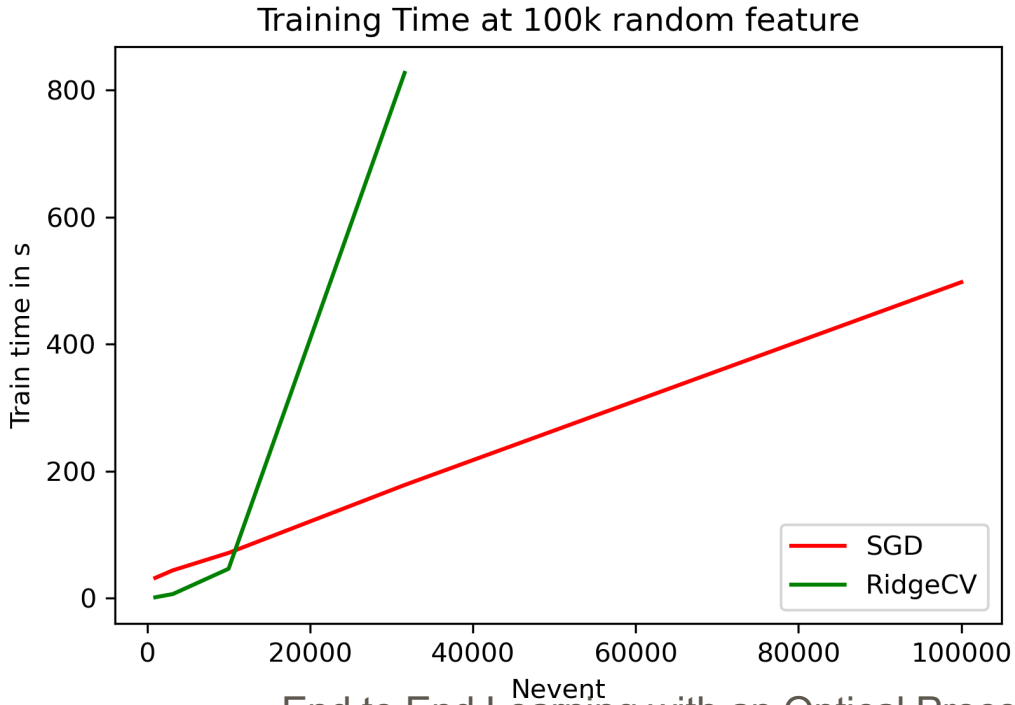
$N_{features}$ =100.000 N_{events} =100 000



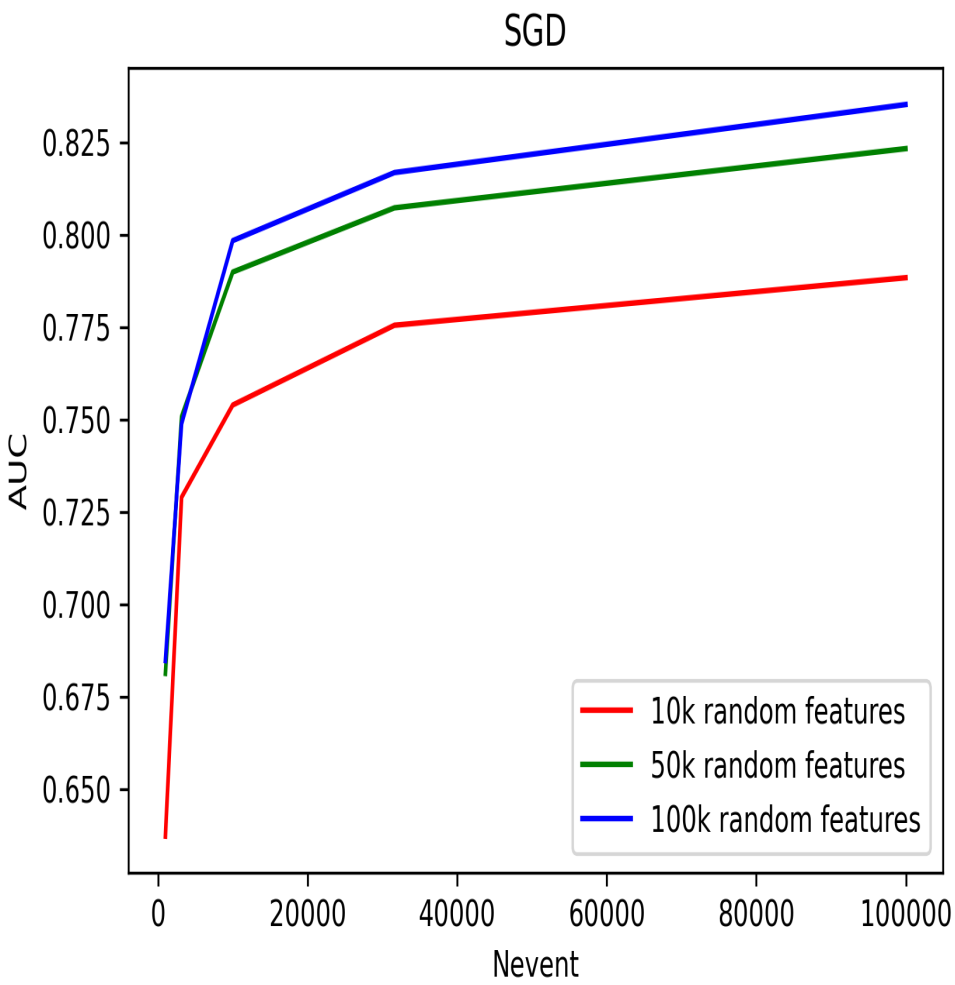
Reference CNN
AUC=0.91

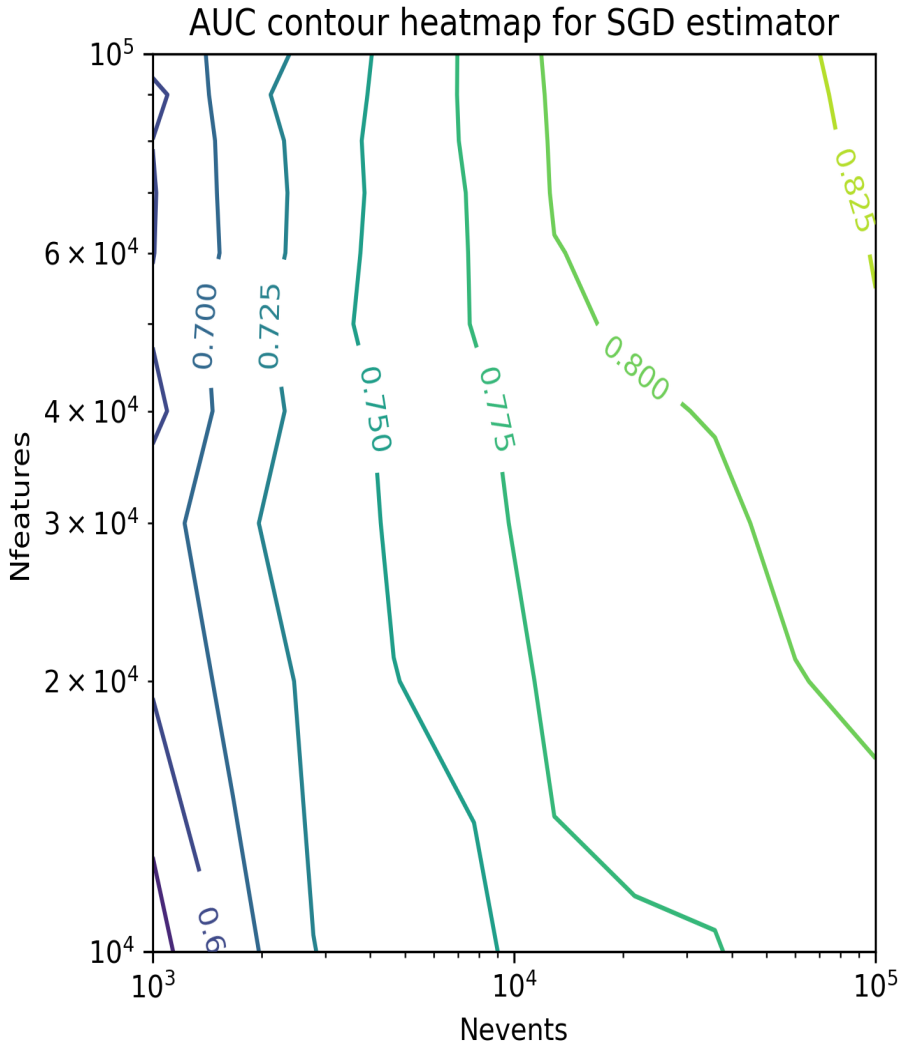
Training time

- SGD training time ~10 minutes on multi core CPU (no GPU)
- Ridge chokes (memory)
- 100k event mapping to OPU : 2000s for 1 CPU core (can be parallelised)
- 100k event through OPU : 300s (can be parallelised), independent of number of random features up to 1 million
- To be compared to ~day for reference CNN training



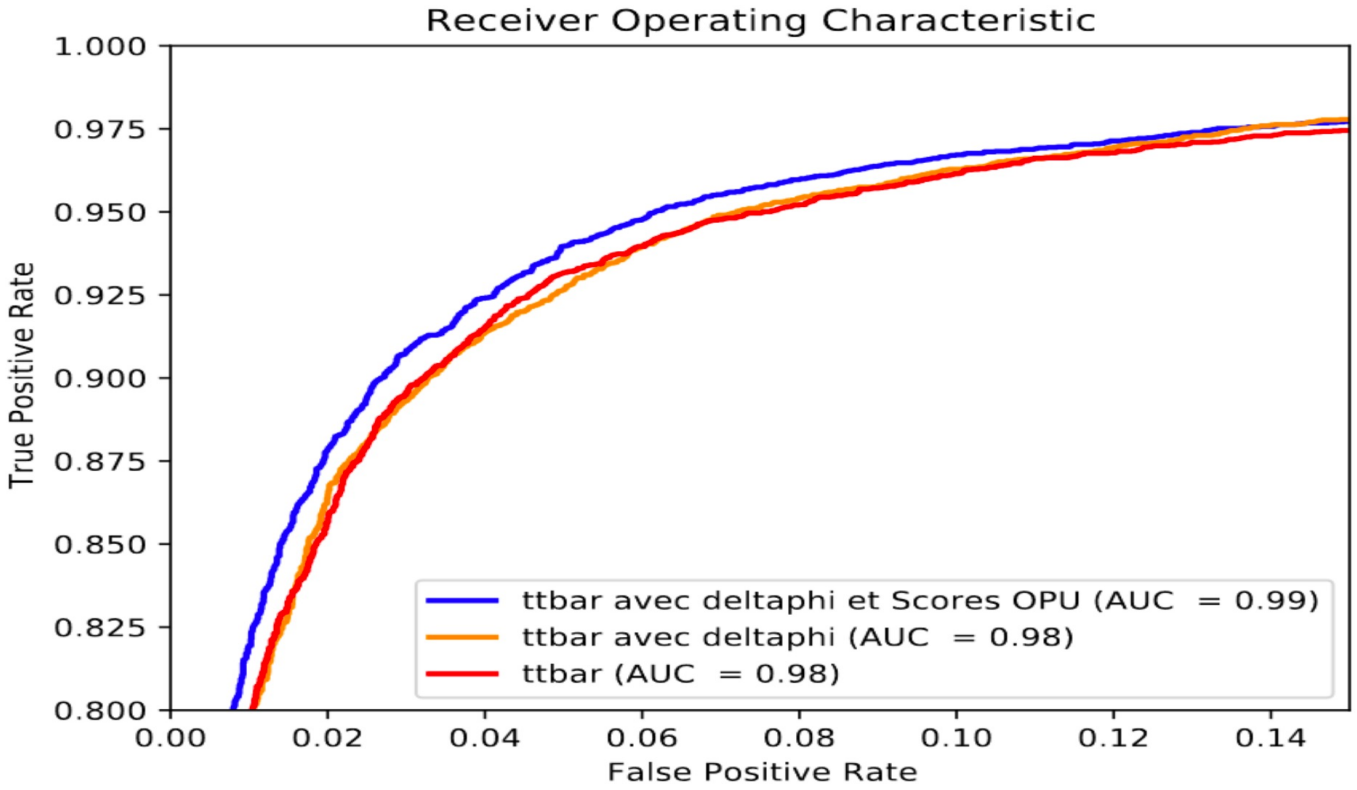
Nfeatures x Nevents





OPU + High Level Feature

- High Level features derived from raw data + high pT lepton info Sum Pt, MET and Met Phi, Phi MET, Mt (lep, MET), Njet, Nbjet, LepPT, LepEta, LepPhi, Lep Iso Ch, Lep Iso Gam, Lep Iso Neut
- BDT trained ➔ results comparable to reference paper
- Add OPU score as an additional feature ➔ improvement in the classification



Conclusion

- We manage to do End to End LHC classification with an OPU

- Significant classifying power (but lower AUC than CNN)
- Training time much faster than CNN (hour vs days)
- Inference speed would be limited by current OPU throughput ~kHz, while CNN inference ~MHz
- Outlook: explore HEP applications where frequent retraining is needed. E.g. Data Quality Monitoring in unstable conditions

- Many thanks to :
  - (In addition to Amélie Chatelain) Iacopo Poli, Laurent Daudet at LightOn for OPU access and great support
  - Thong Nguyen, Maurizio Pierini, Jean-Roch Vlimant et al for giving us access to the dataset of their paper arXiv:1807.00083