The Challenges of Machine Learning at the Edge

Audrey Corbeil Therrien, Xavier Groleau, Berthié Gouin-Ferland, Ryan Coffee

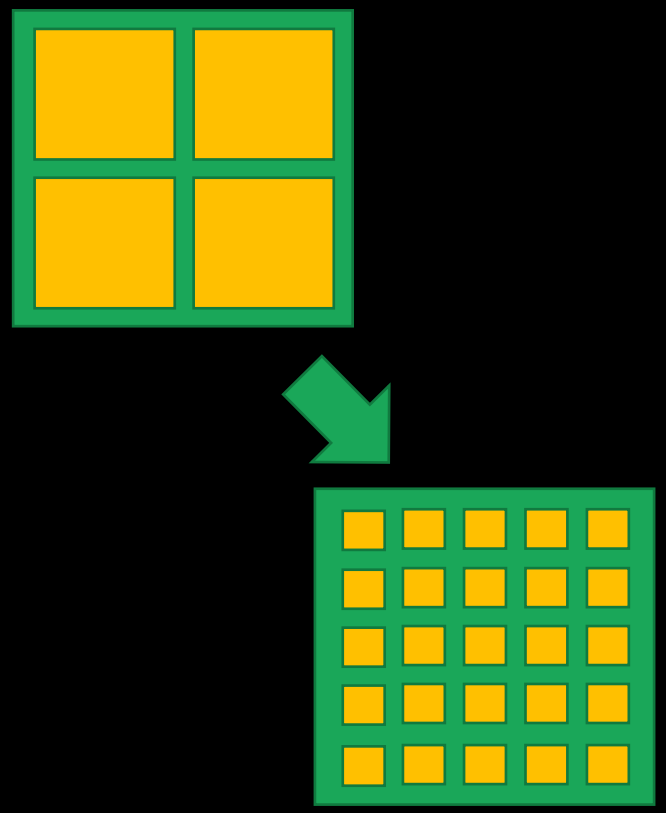
CPAD Instrumentation Frontier Workshop March 2021

3!T Institut interdisciplinaire d'innovation technologique
UNIVERSITÉ DE SHERBROOKE
SLAC NATIONAL ACCELERATOR LABORATORY

Why do we need Real-Time ML ?

New generation sensors and detectors have :

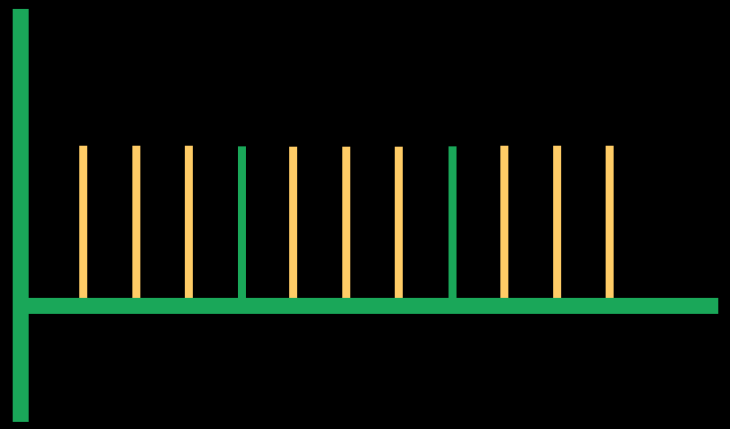
- More pixels;



Why do we need Real-Time ML ?

New generation sensors and detectors have :

- More pixels;
- Faster sampling rates;



Why do we need Real-Time ML ?

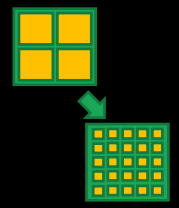
New generation sensors and detectors have :

- More pixels;
- Faster sampling rates;
- Better sensitivity and dynamic range;

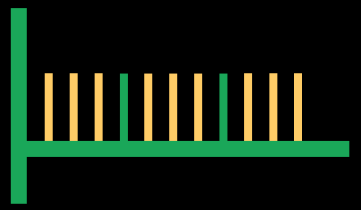
© Dan Plucinski

Why do we need Real-Time ML ?

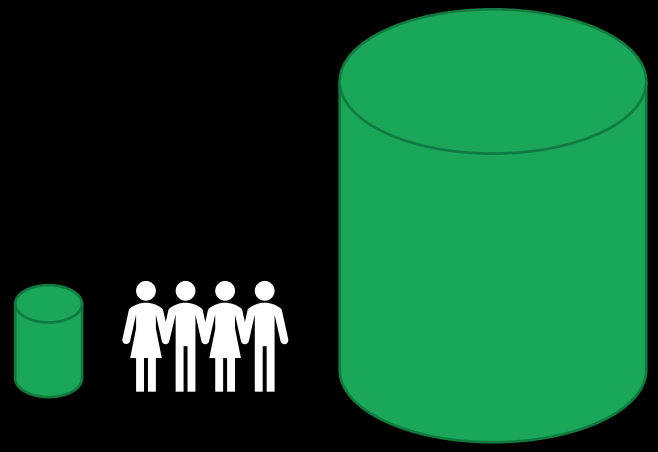
New generation sensors and detectors have :



- More pixels;
- Faster sampling rates;

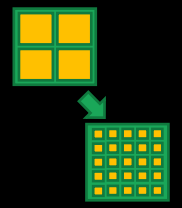


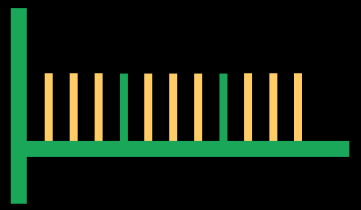
- Better sensitivity and dynamic range;
- Larger surface area;



Why do we need Real-Time ML ?

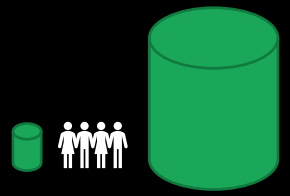
New generation sensors and detectors have :





- More pixels;
- Faster sampling rates;
- Better sensitivity and dynamic range;
- Larger surface area;

© Dan Plucinski



- Too much data!

Why do we need Real-Time ML ?

Collecting all raw data

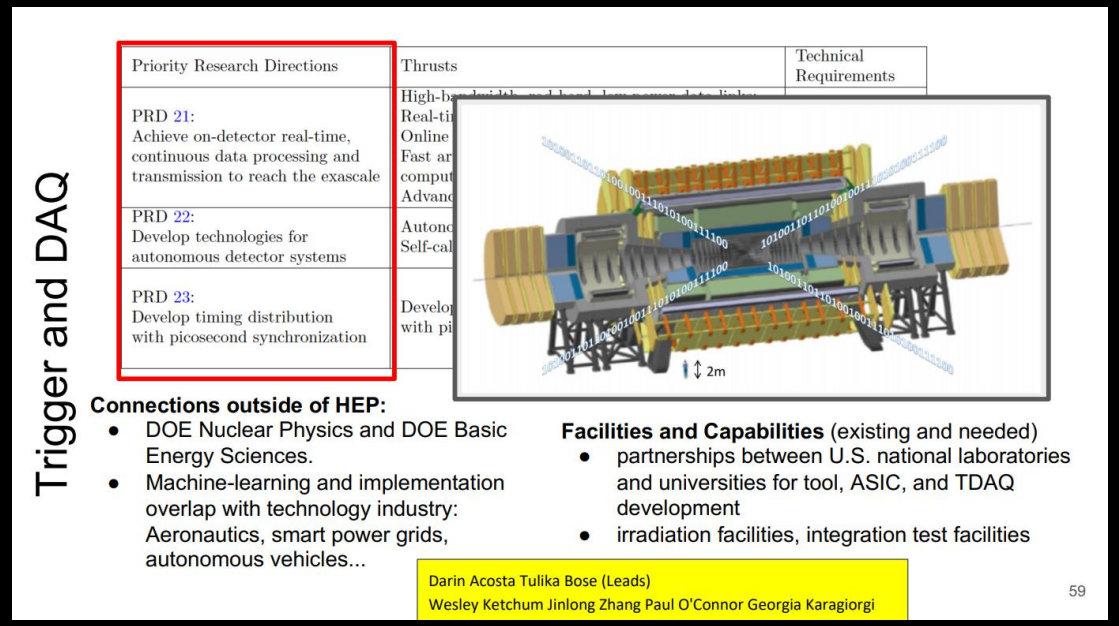
- Requires more wiring, more transmission power
  - kW to MW of power just for transmission
- Requires sophisticated short term storage
  - Parallel SSD array
- Generates cost for long term storage
  - M$ per month for maintenance of facilities
- Adds to transfer times and compute times
  - Often repeating the same initial data preparation

Why do we need Real-Time ML ?

According to the BRN Panel

Priority Research Direction 21 is:

Achieve on-detector real-time, continuous data processing and transmission to reach the exascale.

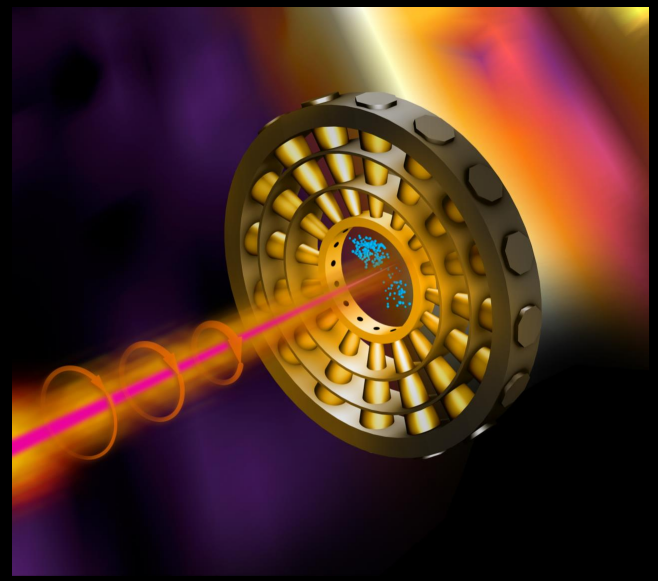


Ian Shipsey and Bonnie Fleming
Thursday 10:10
CPAD workshop

Applications

Applications

- Why EdgeML?
- **Applications**
- What is ML?
- Hardware
- Challenges
- Conclusion



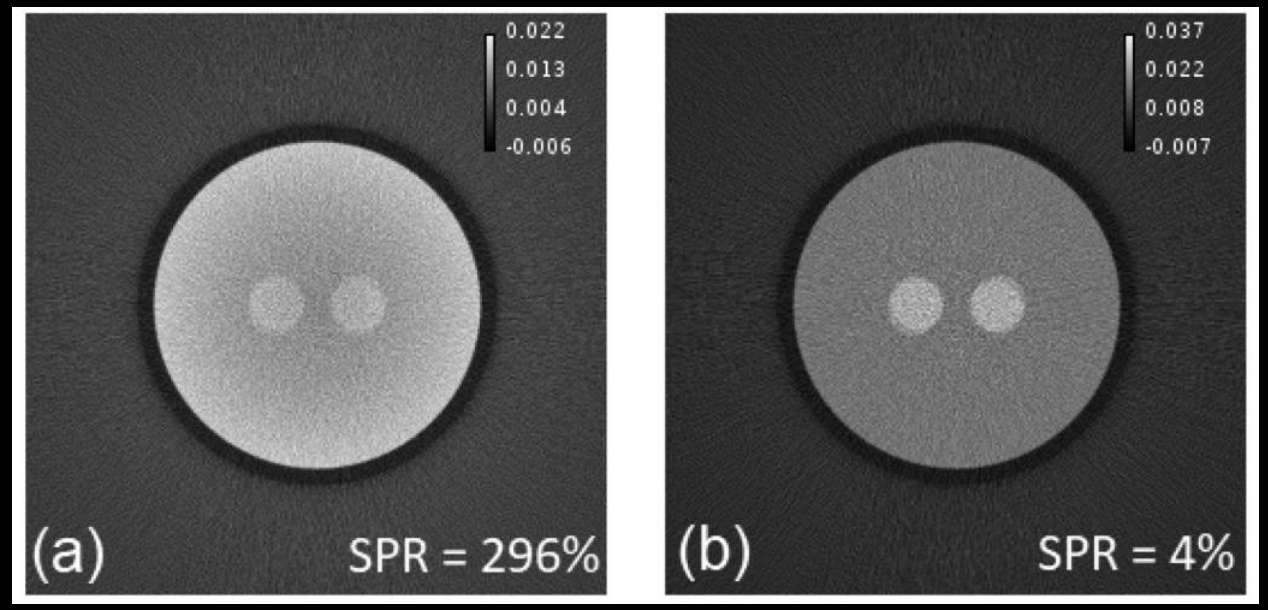
Angular Streaking Detector (LCLS-II)

N. Hartmann et al., Nature Photonics, 2018
https://doi.org/10.1038/s41566-018-0107-6

168 GB/s
Low latency diagnostic detector

Applications

- Why EdgeML?
- **Applications**
- What is ML?
- Hardware
- Challenges
- Conclusion



Time of Flight Computed Tomography

Rossignol, J. et al. 2020

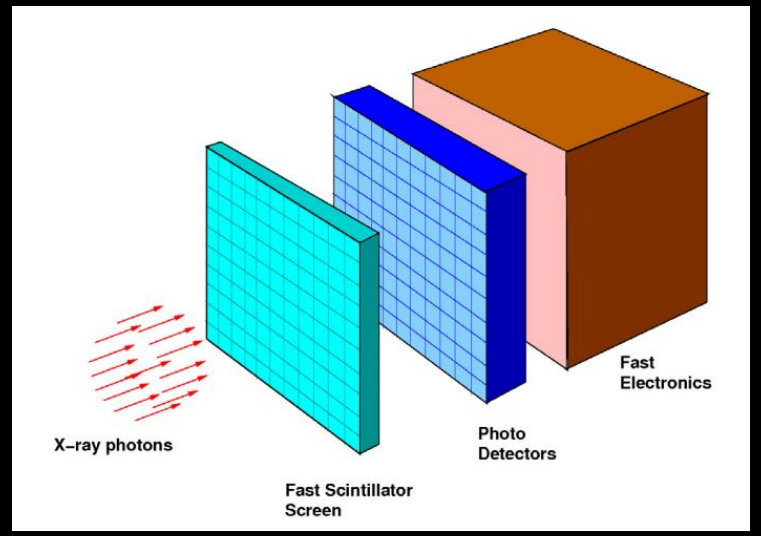
doi.org/10.1088/1361-6560/ab78bf

~120 TB/s

14x14 in^2

Applications

- Why EdgeML?
- **Applications**
- What is ML?
- Hardware
- Challenges
- Conclusion



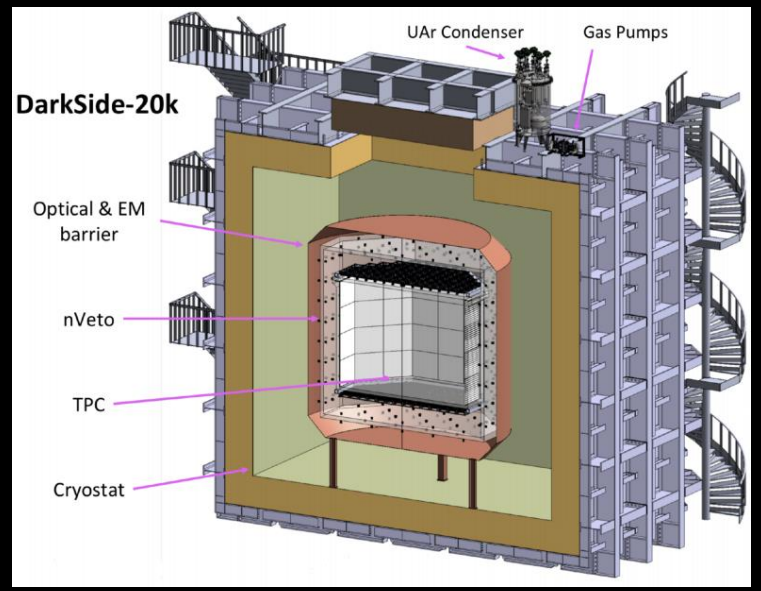
Billion-pixel camera for X-ray applications

Hu, C. et al, 2019

doi.org/10.1016/j.nima.2019.06.011

Los Alamos NATIONAL LABORATORY EST. 1943

~2GB per image



Liquid Argon detectors for dark matter search

Global Argon Dark Matter Collaboration

CPAD2019

~1.2 GB/s

DS20K Veto system

Future ARGO DAQ system

What is Machine Learning?

What is Machine Learning

- Why EdgeML?
- Applications
- **What is ML?**
- Hardware
- Challenges
- Conclusion

Traditional programming

Input + Program = Output

Machine Learning

Input + Output = Program

Benefits of ML

Recognizing patterns

Recognizing anomalies

Non linear regression → reconstruction

Benefits of ML

Faster more flexible programming

Lower computational burden

Fast inference
Low latency decision

Hardware

ML Hardware - GPU

- Why EdgeML?
- Applications
- What is ML?
- **Hardware**
- Challenges
- Conclusion

Thousands of ALU

Highly parallel

Batch oriented

Host CPU

Power

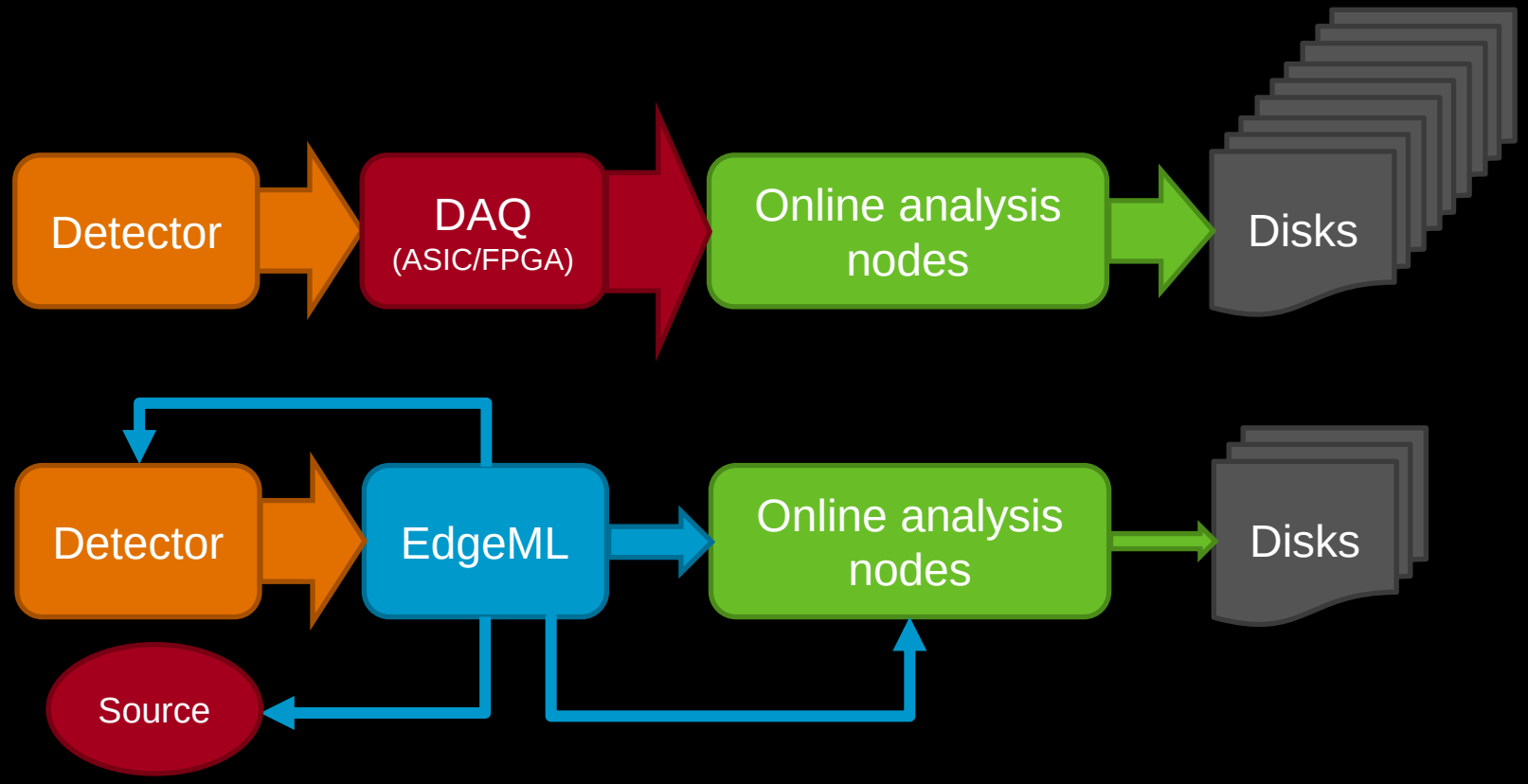
Size

https://nyu-cds.github.io/python-gpu/01-introduction/

ML Hardware – the Edge

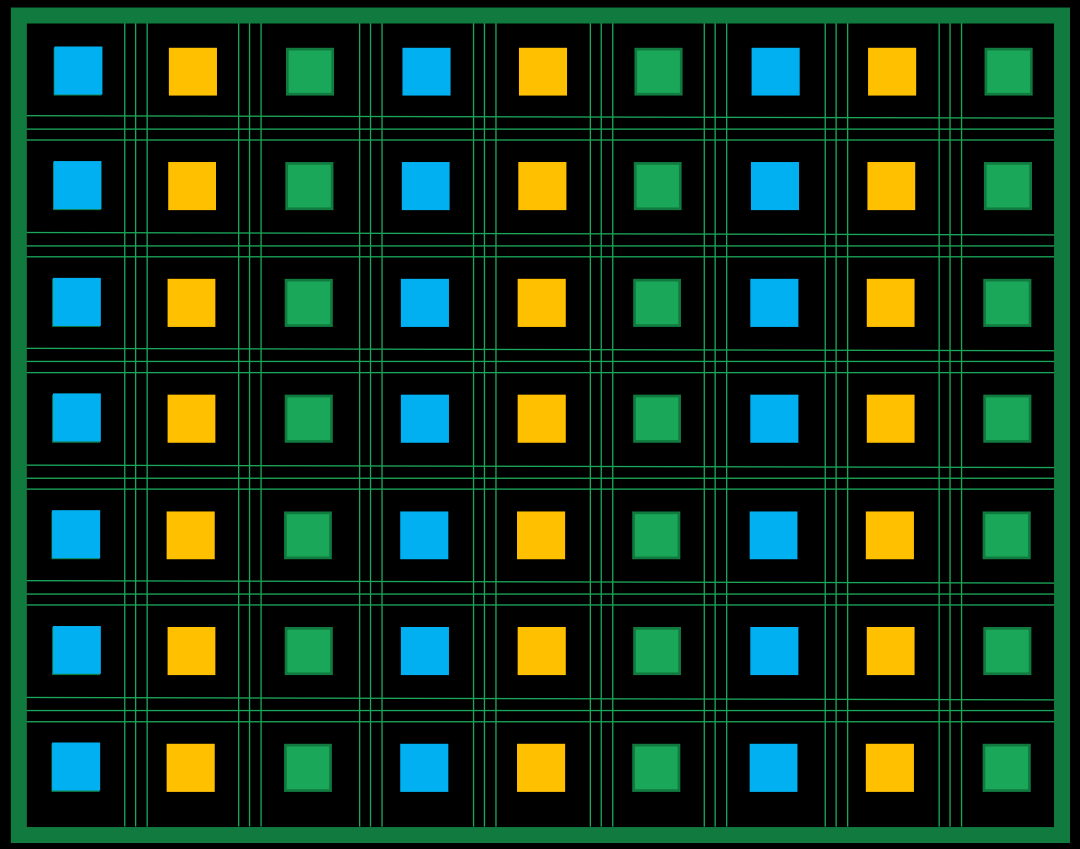
CERN-ATLAS

- Why EdgeML?
- Applications
- What is ML?
- **Hardware**
- Challenges
- Conclusion



ML Hardware - FPGA

- Why EdgeML?
- Applications
- What is ML?
- **Hardware**
- Challenges
- Conclusion



Logic
Memory
Digital signal processing slices

Reconfigurable

Efficient

I/O capacity

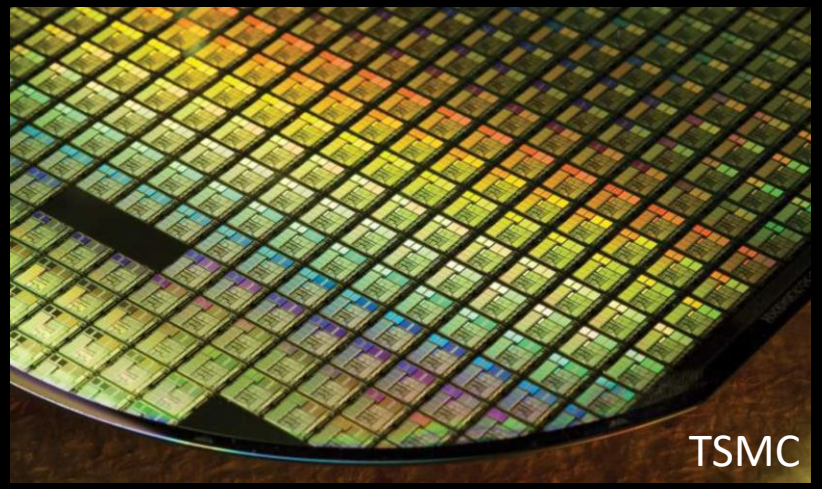
Programming

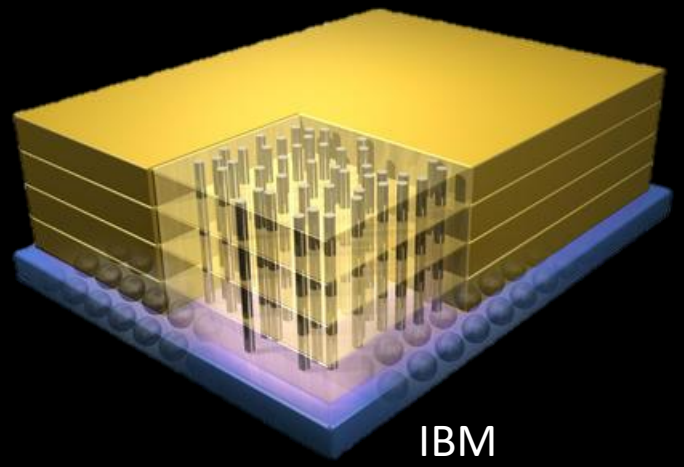
Limited clock

Limited resources

ML Hardware - ASIC

- Why EdgeML?
- Applications
- What is ML?
- **Hardware**
- Challenges
- Conclusion


TSMC


IBM

Efficient++

Custom I/O

3DIC

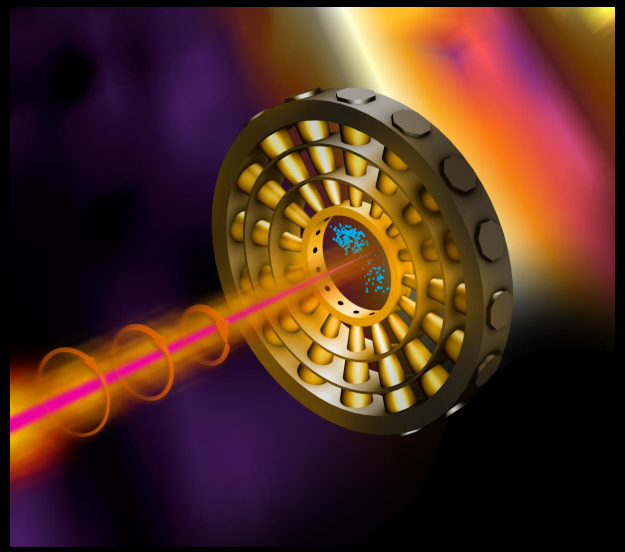
Reconfigurable

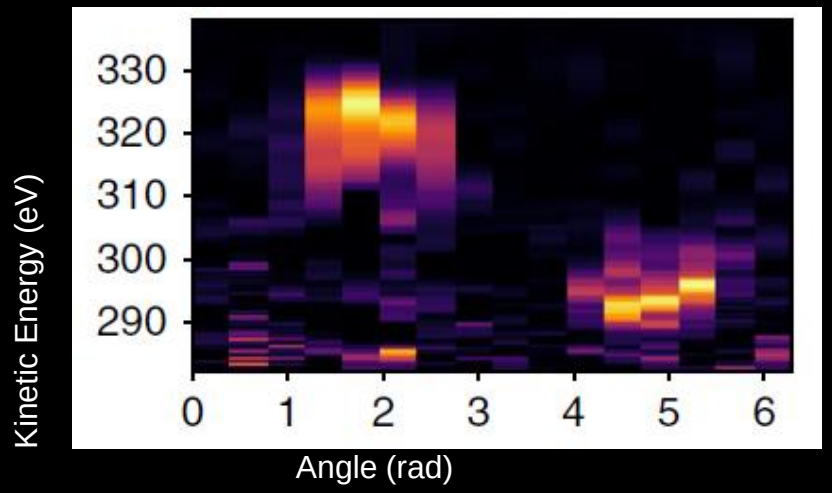
Expensive

Long design cycle

Challenges

Selecting models for hardware

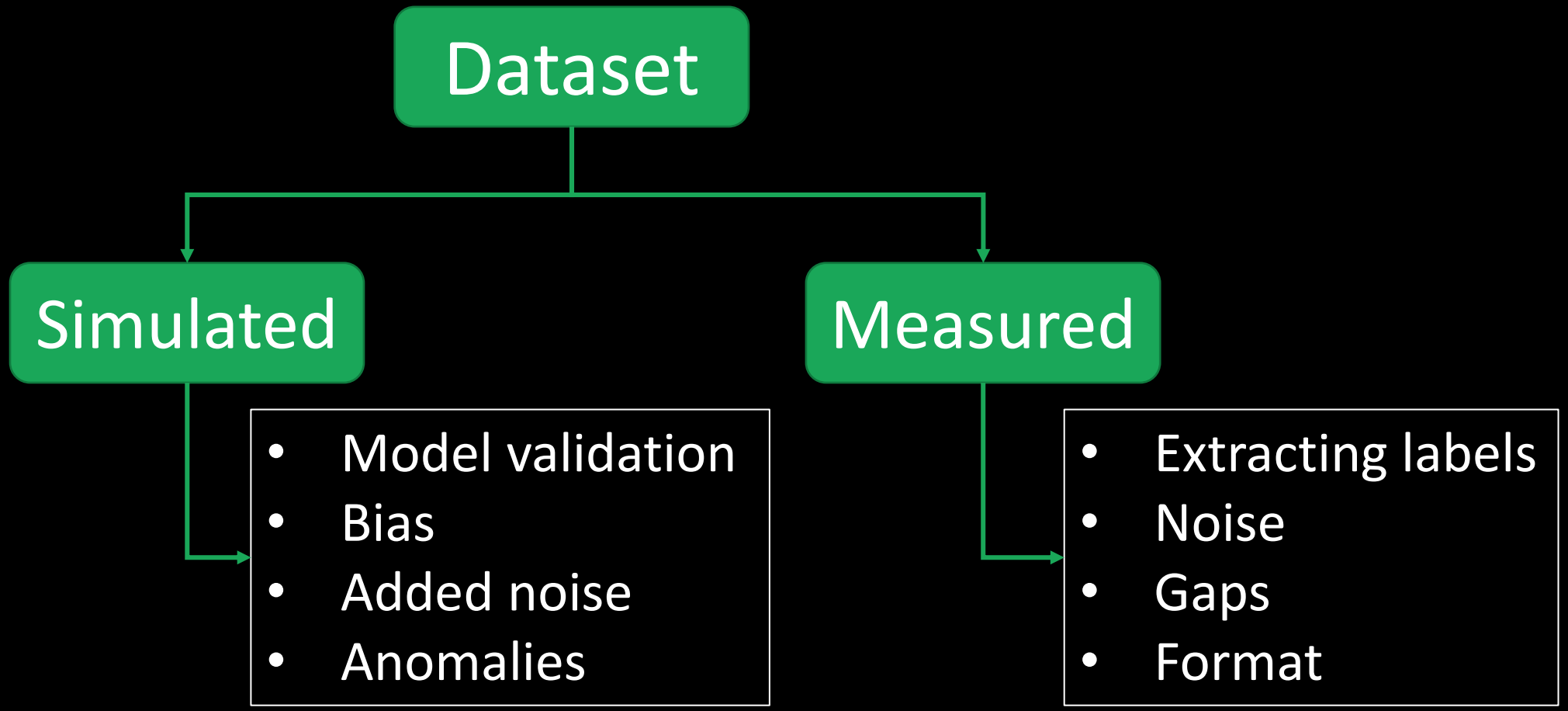




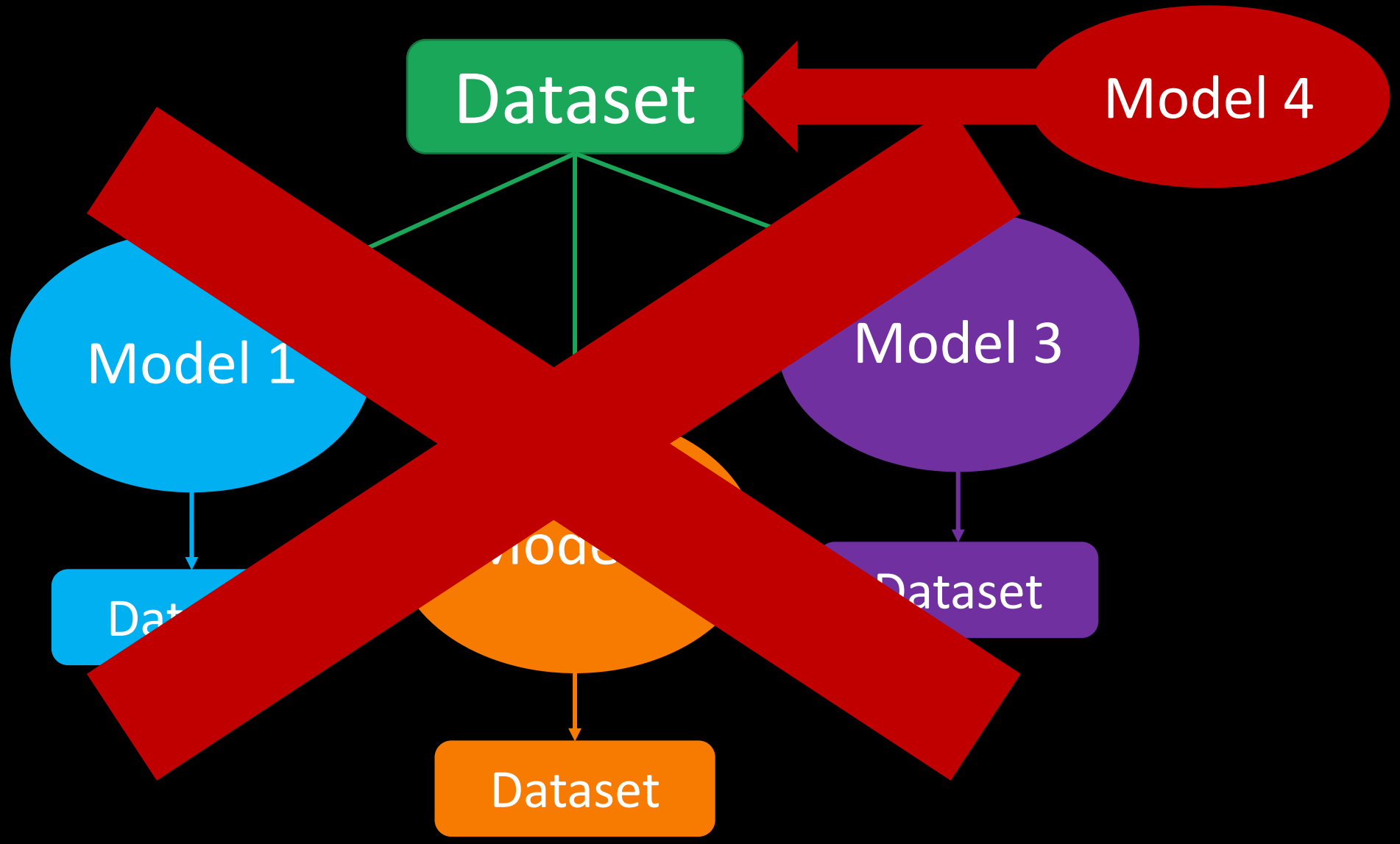
Architecture	Loss (test set)	Processing time (s)	# parameters
Fully connected	0.435	6.970^{e-4}	1 729 964
Convolutional	0.778	4.959^{e-4}	615 748
Recurrent	0.582	2.698^{e-3}	197 120

Work : Xavier Groleau

Training Datasets



Data and model provenance



User endorsement

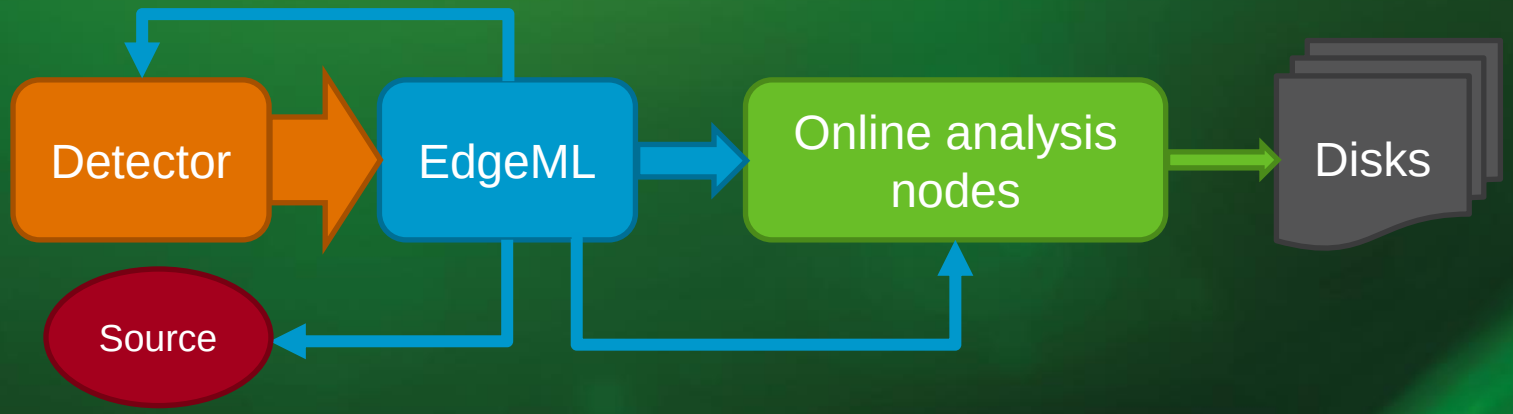
How do we convince the users of the instruments that the machine learning inference gives them accurate information?

- Validation
- Interpretation
- Uncertainty measurement
- Raw data sampling
- Others?

Conclusion

Edge Machine Learning is key to exploit the full potential of new high rate detectors and will accelerate critical discoveries…

…but we have a lot of work to do!



3!T
Institut interdisciplinaire d'innovation technologique

UNIVERSITÉ DE SHERBROOKE

SLAC NATIONAL ACCELERATOR LABORATORY

Bourses postdoctorales Banting Postdoctoral Fellowships

NSERC CRSNG