Why do we need Real-Time ML

200 GB/s

2000 compute nodes

2.5 GB/s

Source: Hegeman, J. (2018)

LARGE HADRON COLLIDER

Credit: CERN
Why do we need Real-Time ML

5500 GB/s
TBD
compute nodes

61 GB/s

Source: Hegeman, J. (2018)

Hi-Luminosity Upgrade

Credit: CERN
Why do we need Real-Time ML

2 000 000 GB/s

2 MW

TBD

Compute nodes

Why do we need Real-Time ML
LINAC Coherent Light Source

20 to 1200 GB/s

• Assuming 1 TB/s, 12 hour shift, nonstop
• 43 200 TB per shift – 56 years of 4K movies
• 1.3 M$/month of storage costs created every shift
What is Machine Learning?
What is Machine Learning

Traditional programming

Input + Program = Output

Machine Learning

Input + Output = Program
What is Machine Learning

- Decision Trees
- Random Forest
- Perceptron
- FeedForward

https://www.asimovinstitute.org/neural-network-zoo/
What is Machine Learning

Convolutional neural network

Image recognition

Phung V. H. and Rhee, J. R., DOI: 10.3390/app9214500
What is Machine Learning

Recurrent Neural Network

Time series data

Time series data
What is Machine Learning

Autoencoders

Data compression
Feature extraction

Phung V. H. and Rhee, J. R., DOI: 10.3390/app9214500
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**

- Thousands of ALU
- Highly parallel
- Batch oriented
- Host CPU
- Power
- Size

---

[https://nyu-cds.github.io/python-gpu/01-introduction/](https://nyu-cds.github.io/python-gpu/01-introduction/)
ML Hardware - FPGA

Reconfigurable
Efficient
I/O capacity
Programming
Limited clock
Limited resources

What is ML?
Hardware
CookieBox
Pitfalls
Conclusion
ML Hardware - ASIC

- Efficient++
- Custom I/O
- 3DIC
- Reconfigurable
- Expensive
- Long design cycle
Cookiebox
Proof of concept
CookieBox

What is ML?
Hardware
CookieBox
Pitfalls
Conclusion

Complex reconstruction

N. Hartmann et al., Nature Photonics, 2018
What is ML?
Hardware
CookieBox
Pitfalls
Conclusion

CookieBox

Detector → DAQ (ASIC/FPGA) → Online analysis nodes → Disks

Detector → EdgeML → Online analysis nodes → Disks

Source
CookieBox

What is ML?
Hardware
CookieBox
Pitfalls
Conclusion
Layer 1: 800 inputs
Layer 2: 200 inputs
Output Layer: 100 inputs

Maximum theoretical throughput $R$:

$$R = \frac{1}{\text{MAX (layer latency)}}$$

$R$ (62.5 MHz) = 77 kHz
$R$ (250 MHz) = 308 kHz
Other projects

Billion-pixel camera for X-ray applications
Hu, C. et al, 2019
doi.org/10.1016/j.nima.2019.06.011

~2GB per image

Liquid Argon detectors for dark matter search
Global Argon Dark Matter Collaboration
CPAD2019

~1.2 GB/s
DS20K Veto system
Other projects

Time of Flight Computed Tomography
Rossignol, J. et al. 2020
doi.org/10.1088/1361-6560/ab78bf

~120 TB/s
14x14 cm$^2$
Pitfalls
Training Datasets

Dataset

Simulated
- Model validation
- Bias
- Added noise

Measured
- Extracting labels
- Anomalies
- Format
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
Data and model provenance

What is ML?
Hardware
CookieBox
Pitfalls
Conclusion
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!