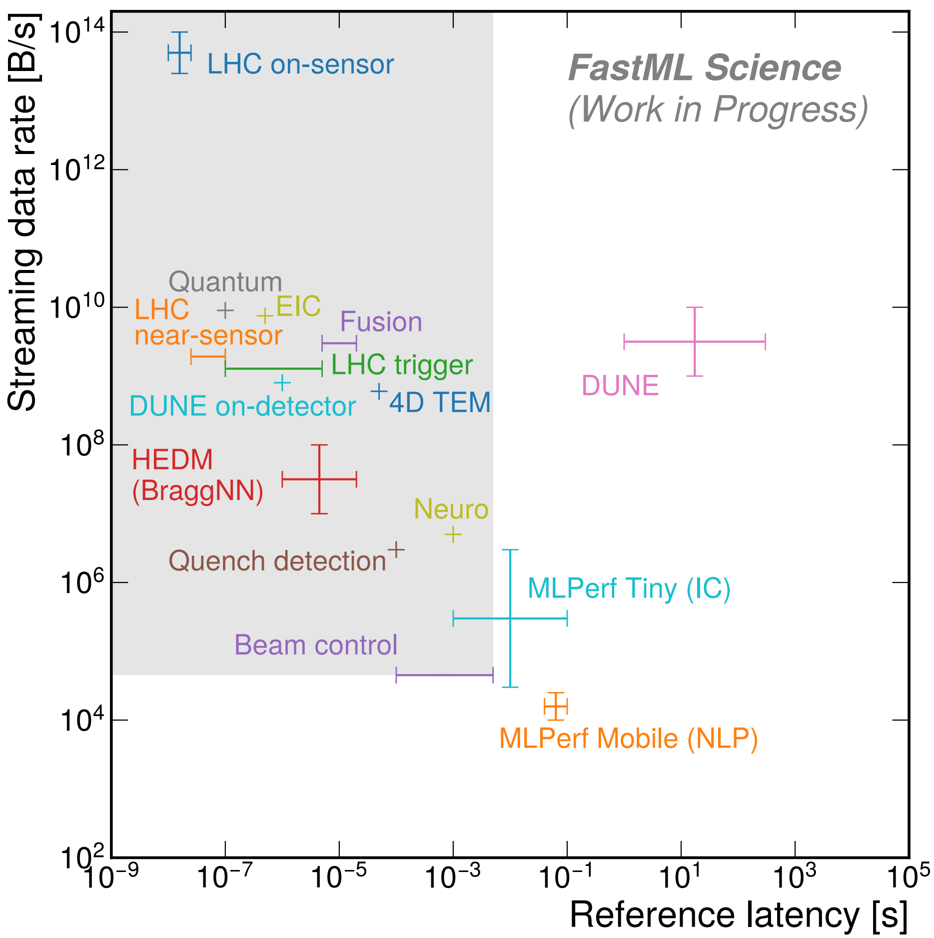
FastML Science Benchmarks: Accelerating Real-Time Scientific Edge ML

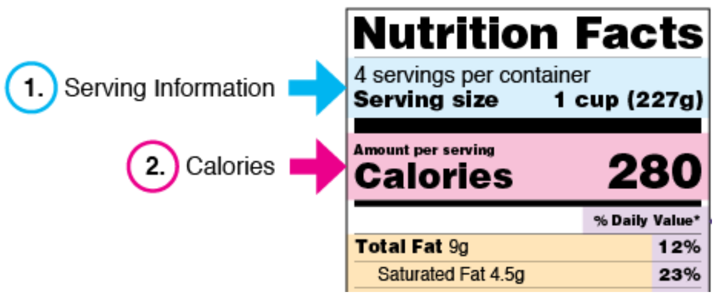


Aim to provide datasets, objectives, constraints, a reference solution, and metrics for each task.

- Current material on github.

Eager to scale this effort to incorporate new scientific domains and new kinds of technical challenges.

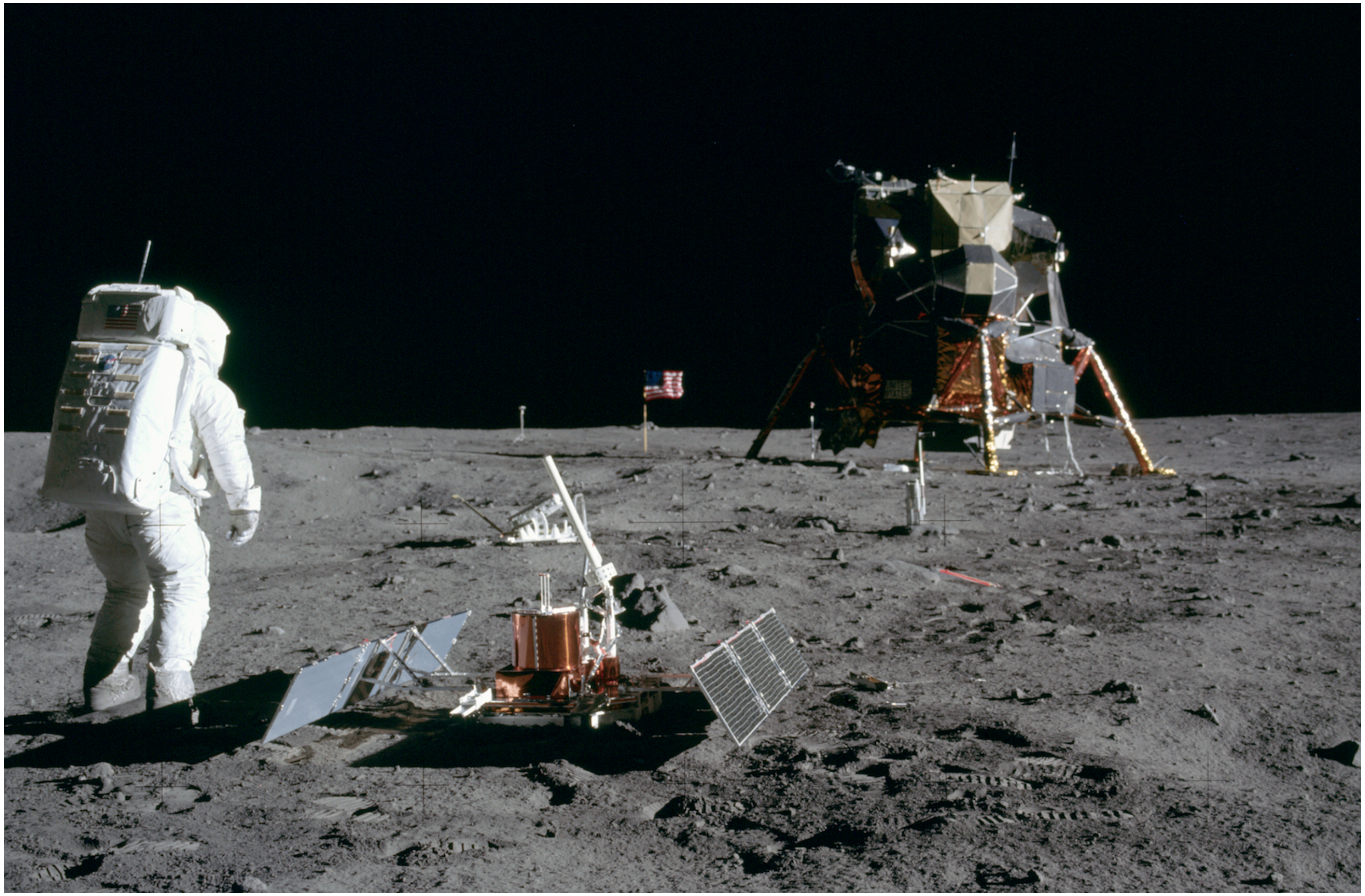
- How to effectively *curate* these tasks?
  - Aim to highlight the unique challenges.
  - “Representative” versus “cutting-edge”?
- “Nutrition facts” summary for each task?



Grand challenges, moonshots

Benchmarks are great but not super glamorous, instead...

Capture people's imagination of what's possible from the science side



Fermilab

<u>Scientific Grand Challenges</u> spur innovation

LHC: all sub-detectors analyzing data at 40 MHz

DUNE: expansive (non-)accelerator ν program (solar, supernova, proton decay, ββ decay)

Accelerator controls with adaptive online agents and digital twin

Science:

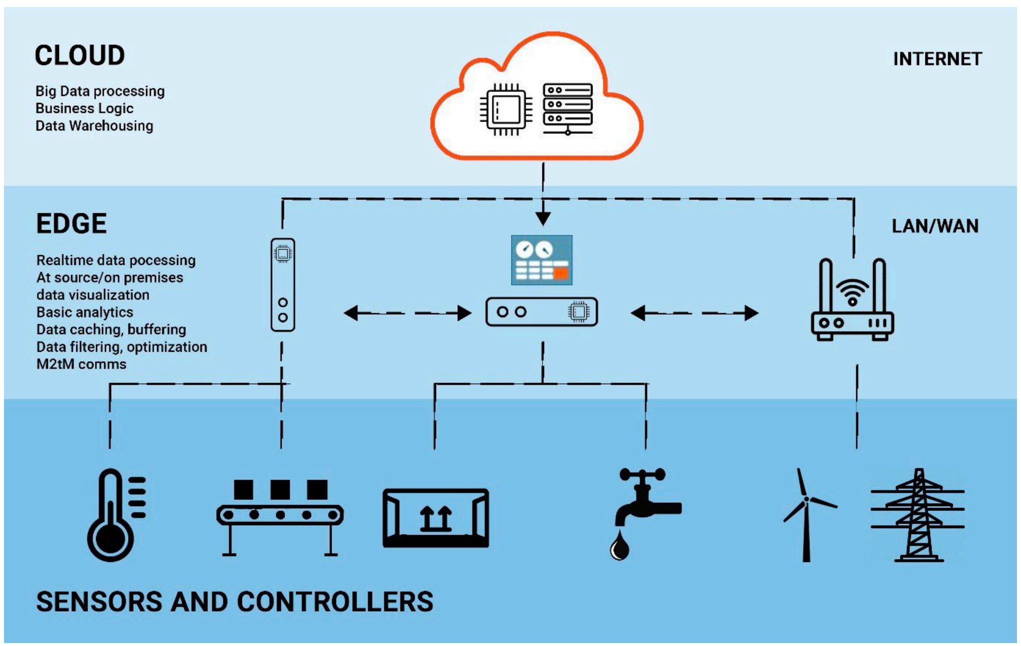
- Quantum,
- Magnet development,
- Fusion,
- Neuroscience,
- Nuclear,
- Material sciences, etc.

Industry: Internet-of-Things, AVs, manufacturing

Grand challenges, moonshots

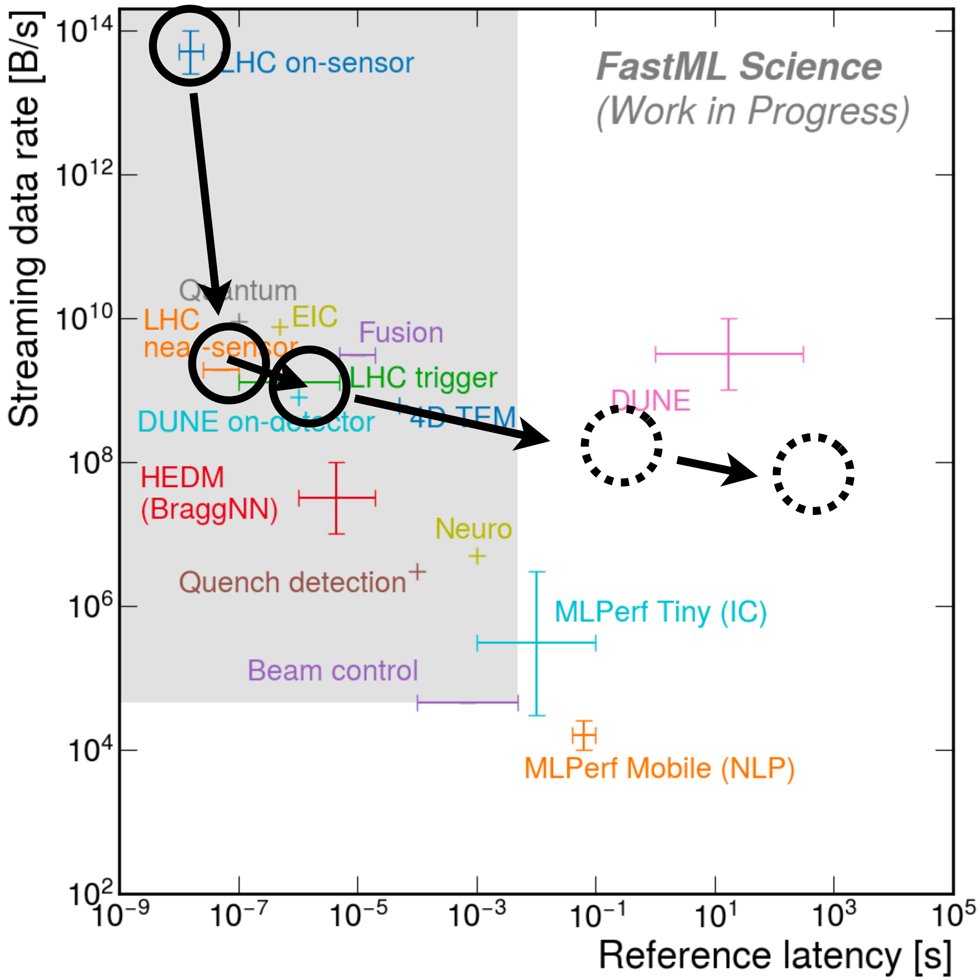
Grand challenges are **multi-scale problems**:

Real-time processing, control, reconstruction, simulation, analysis
→ Edge AI, Efficient AI, Foundation models, Digital Twins, Robustness, UQ



Scientific Grand Challenge = $\Sigma_{1...N}$ {benchmarks}$_i$

- **Compile *benchmarks as a critical path* to grand challenges**



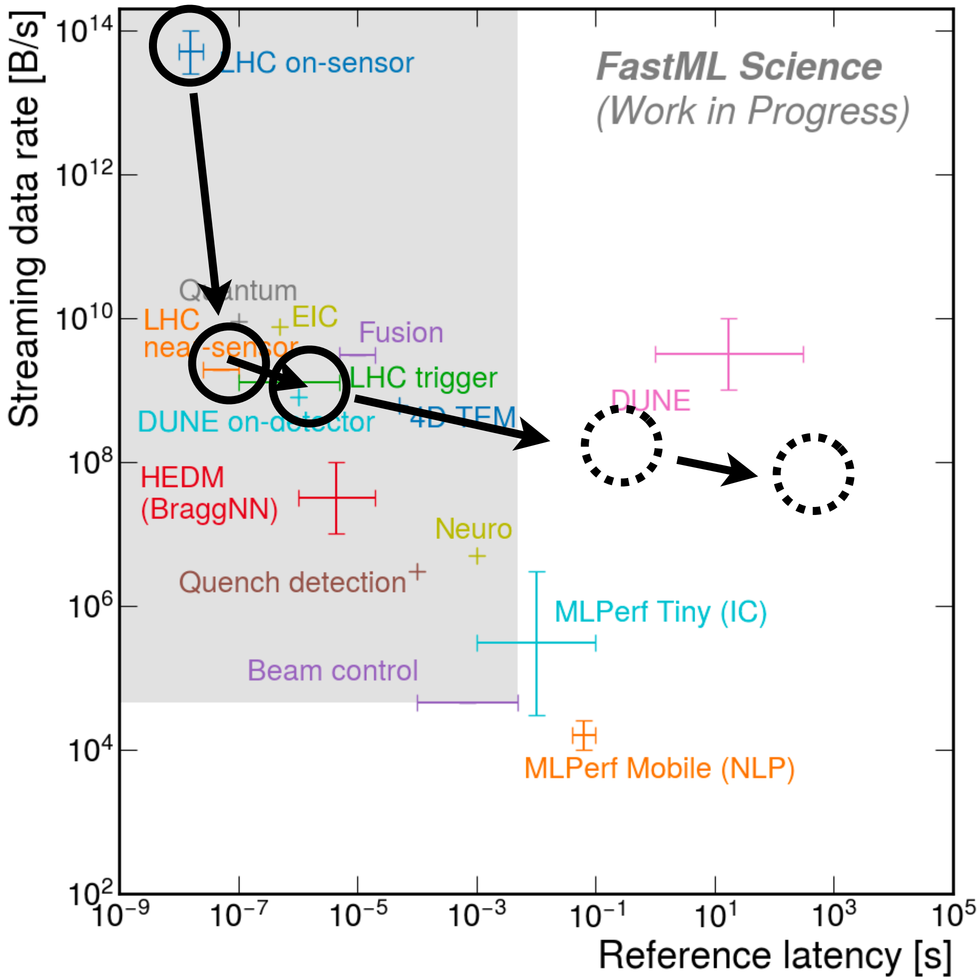
Example Moonshot: analyze all data from LHC

Benchmarks:

- Read out pixel detector at 40 MHz with full fidelity, e.g. smart pixels
- Enable real-time streaming processing (40 MHz scouting), e.g. dark machines
- Offline tracking at 200 PU with same performance as 60 PU with 10x speedup/event, e.g. ExaTrkX challenge
- Accelerate fast simulations by 100x with same fidelity as full simulation, e.g. Calo Challenge
- Enable automized calibrations (no more re-reconstruction), .e.g ??

Scientific Grand Challenge = $\Sigma_{1...N}$ {benchmarks}$_i$

- **Compile *benchmarks as a critical path* to grand challenges**



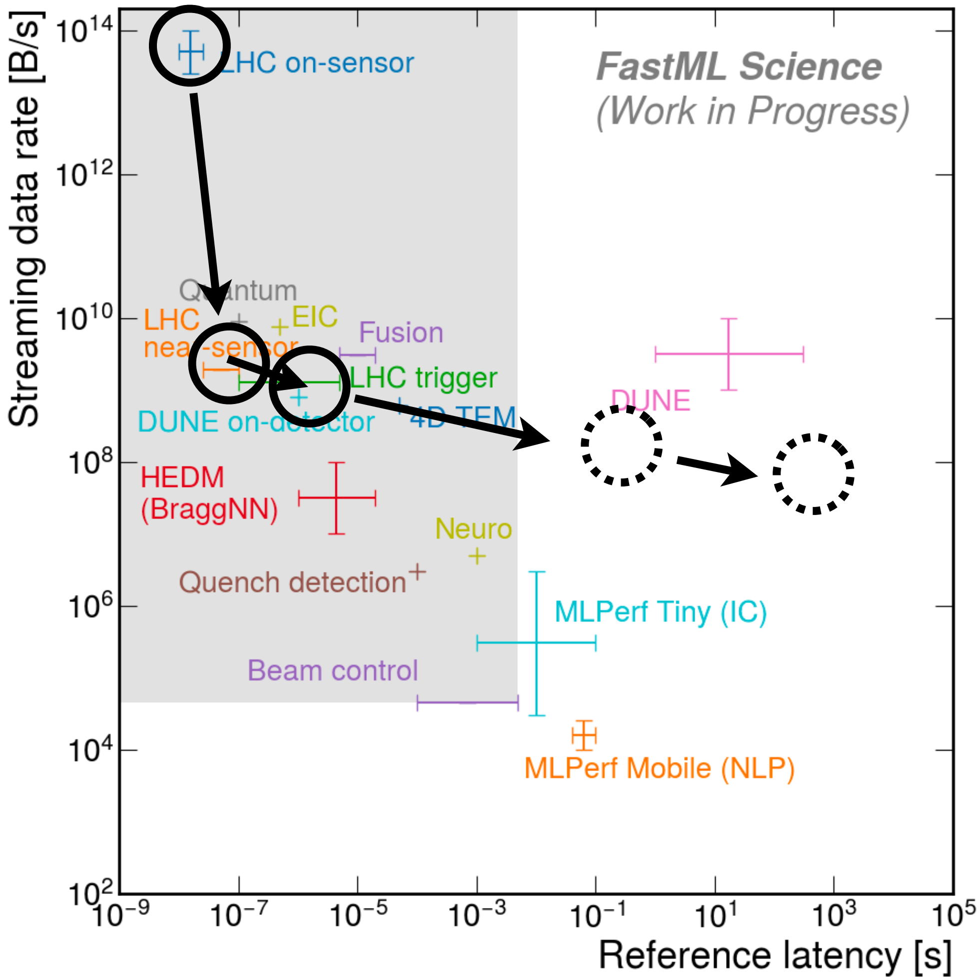
Example Moonshot: neutrino physics

Benchmarks:

- Improve SuperNova Burst streaming signal acceptance by XX% with data reduction rate of YY
- Enable ZZ sensitivity to neutrinoless double beta decay with low energy reconstruction task A
- Charge current reconstruction performance with resolution of σ
- ???

Scientific Grand Challenge = $\Sigma_{1...N}$ {benchmarks}$_i$

- **Compile *benchmarks as a critical path* to grand challenges**



Moonshot = $\Sigma_{1...N}$ {benchmarks}$_i$

Can we do this for other fields?

Another axis for benchmark metrics: Robustness and UQ

- What happens when the detector, environment, conditions change?
- Develop new benchmarks for robust science use-cases
  - harder to develop benchmarks

A nice example:

RobustBench Leaderboards Paper FAQ Contribute Model Zoo 🚀

RobustBench

A standardized benchmark for adversarial robustness

The goal of **RobustBench** is to systematically track the *real* progress in adversarial robustness. There are already more than 3'000 papers on this topic, but it is still unclear which approaches really work and which only lead to overestimated robustness. We start from benchmarking common corruptions, ℓ_∞- and ℓ_2-robustness since these are the most studied settings in the literature. We use AutoAttack, an ensemble of white-box and black-box attacks, to standardize the evaluation (for details see our paper) of the ℓ_p robustness and CIFAR-10-C for the evaluation of robustness to common corruptions. Additionally, we open source the RobustBench library that contains models used for the leaderboard to facilitate their usage for downstream applications.

To prevent potential overadaptation of new defenses to AutoAttack, we also welcome external evaluations based on *adaptive attacks*, especially where AutoAttack flags a potential overestimation of robustness. For each model, we are interested in the best known robust accuracy and see AutoAttack and adaptive attacks as complementary.

Vision

		Scientific Moonshots		
		Domain A	…	Domain N
AI thrusts	AI - 1: Real-time	Benchmark 1A		Benchmark 1N
	AI - 2: Control			
	AI - 3: Autonomous			
	AI - 4: Foundation			
	AI - 5: Generative	Benchmark 5A		Benchmark 5N

Extra

FastML Science Benchmarks: Accelerating Real-Time Scientific Edge ML

Experiments continuously evolve to probe ***shorter distances + timescales***.

- Powerful detectors → large data volumes
- Require some shrewd selection or distillation ***at the edge***.
  - Increasingly with ML, deep neural nets

Benchmarks guide development of next-gen edge ML hardware + software.

- Science tasks eclipse current standards!

We've developed a **set of science benchmarks**, including representative tasks across a range of domains.

- Eager to collaborate, expand to a wider range of benchmarks!

https://arxiv.org/abs/2207.07958

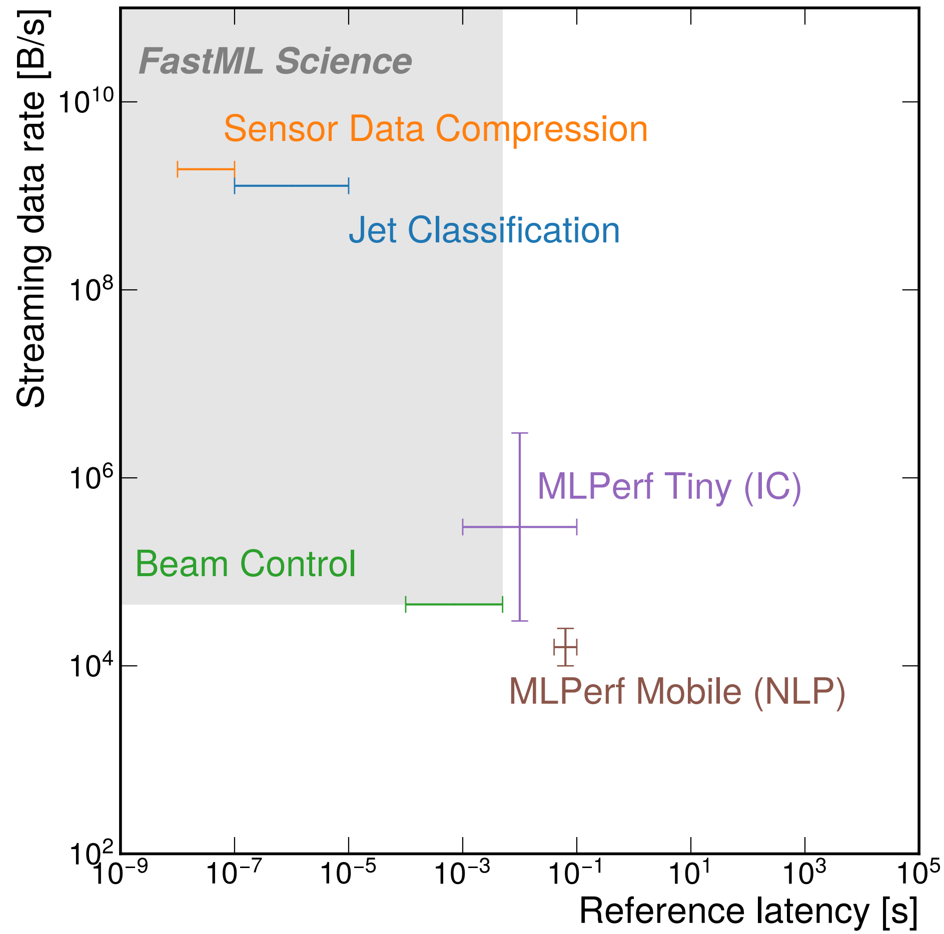
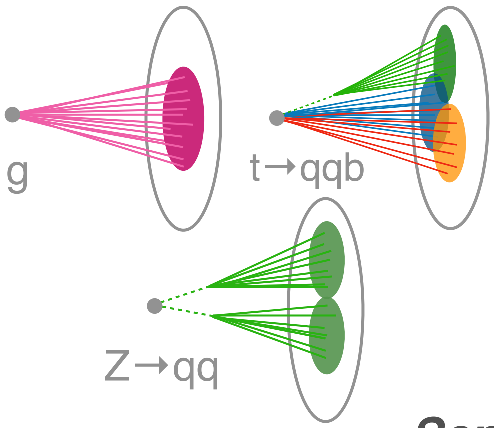


Figure: Reference latencies and streaming input data rates for common industry benchmarks and FastML Science.

Fermilab

FastML Science Benchmarks: Accelerating Real-Time Scientific Edge ML

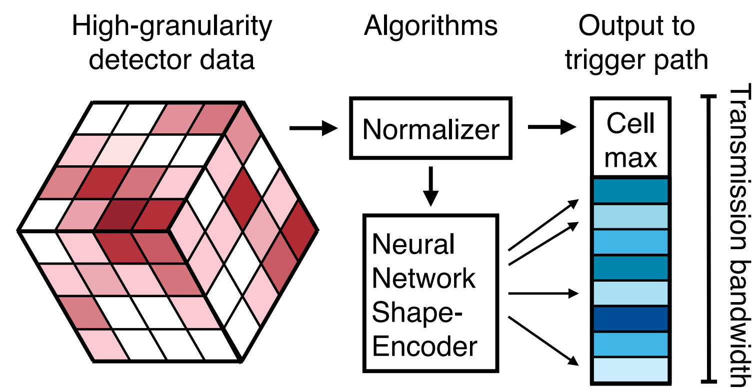


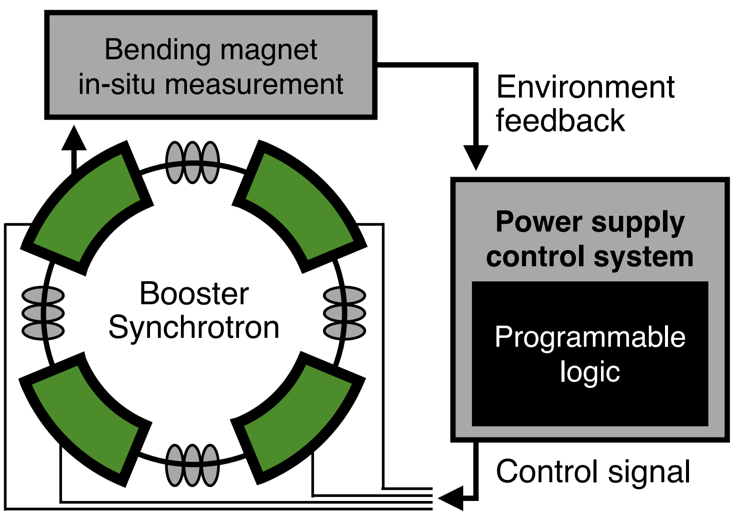
Supervised classification of particle jets

- LHC Trigger: 100 TB/s (Virtex Ultrascale+)
- 1 µs latency, 150 ns pipeline

Sensor data compression

- Next-gen “imaging calorimeter”
- ASIC → area, power constraints





Accelerator beam controls

- FPGA controller interacts with ‘environment’
  - Reinforcement learning
- Inputs from 50 devices across accel. complex.

FastML Science Benchmarks: Accelerating Real-Time Scientific Edge ML

Benchmark task	Data representation	ML type	Latency	Throughput	Target platform	Material status? (Data/model/	…
LHC Sensor	image	unsupervised	10-100ns	1-10 TBps	ASIC	✅	
LHC Trigger	Point cloud	supervised	0.1-100us	1-10 TBps	FPGA	✅	
Neutrino	Image/spectral	unsupervised	1s - 5 min	1-50 GBps	ASIC / GPU(?)		
Accelerator control	spectral	control	1ms	100 kBps	FPGA	✅	
4D TEM	image	unsupervised	50us	1 GBps	FPGA		
Qubit readout	RF spectral	supervised	0.1-1ns	50 GBps	FPGA		
Fusion	image	supervised	5-20us	1-10 GBps	FPGA		
Neuro	spectral	unsupervised	1ms	5 MBps	ASIC (?)		

(And many more…)

What is the most impactful set of material to highlight?

- Power budget, number of operations, …