Fast Machine Learning

Allison McCarn Deiana, SMU
Nhan Tran, Fermilab
On behalf of the whole writing team

July 19, 2022
IF04 - Seattle Snowmass Summer Meeting 2022

Snowmass 2021

Background

- Fast Machine Learning for Science Workshop was held 30 November – 3 December, hosted virtually by Southern Methodist University
  - Website available here: https://indico.cern.ch/e/fml2020
  - Workshop was interdisciplinary and attracted over 500 participants, talks on a wide variety of scientific applications.
  - Workshop also included a hands-on tutorial session, to get people started on applications of fast machine learning.
- After the workshop, a community white paper has been prepared, and was accepted to a special issue of Frontiers in AI

Shameless plug: Oct 3-6 in Dallas!
https://indico.cern.ch/e/fml2022

frontiers | Frontiers in Big Data

REVIEW
published: 12 April 2022
doi: 10.3389/fdata.2022.787421

Check for updates

Applications and Techniques for Fast Machine Learning in Science

Allison McCarn Deiana [1]*, Nhan Tran [2,3]*, Joshua Agar [4], Michaela Blott [5], Giuseppe Di Guglielmo [6], Javier Duarte [7], Philip Harris [8], Scott Hauck [9], Mia Liu [10], Mark S. Neubauer [11], Jennifer Ngadiuba [2], Seda Ogrenci-Memik [3], Maurizio Pierini [12], Thea Aarrestad [12], Steffen Bähr [13], Jürgen Becker [13], Anne-Sophie Berthold [14], Richard J. Bonventre [15], Tomás E. Müller Bravo [16], Markus Diefenthaler [17], Zhen Dong [18], Nick Fritzsche [19], Amir Gholami [18], Ekaterina Govorkova [12], Dongning Guo [3], Kyle J. Hazelwood [2], Christian Herwig [2], Babar Khan [20], Sehoon Kim [18], Thomas Klijnsma [2], Yaling Liu [21], Kin Ho Lo [22], Tri Nguyen [8], Gianantonio Pezzullo [23], Seyedramin Rasoulinezhad [24], Ryan A. Rivera [2], Kate Scholberg [25], Justin Selig [14], Sougata Sen [26], Dmitri Strukov [27], William Tang [28], Savannah Thais [28], Kai Lukas Unger [13], Ricardo Vilalta [29], Belina von Krosigk [13,30], Shen Wang [21] and Thomas K. Warburton [31]

[1] Department of Physics, Southern Methodist University, Dallas, TX, United States, [2] Fermi National Accelerator Laboratory, Batavia, IL, United States, [3] Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, United States, [4] Department of Materials Science and Engineering, Lehigh University, Bethlehem, PA, United States, [5] Xilinx Research, Dublin, Ireland, [6] Department of Computer Science, Columbia University, New York, NY, United States, [7] Department of Physics, University of California, San Diego, San Diego, CA, United States, [8] Massachusetts Institute of Technology, Cambridge, MA, United States, [9] Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, United States, [10] Department of Physics and Astronomy, Purdue University, West Lafayette, IN, United States, [11] Department of Physics, University of Illinois Urbana-Champaign, Champaign, IL, United States, [12] European Organization for Nuclear Research (CERN), Meyrin, Switzerland, [13] Karlsruhe Institute of Technology, Karlsruhe, Germany, [14] Cerebras Systems, Sunnyvale, CA, United States, [15] Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [16] Department of Physics and Astronomy, University of Southampton, Southampton, United Kingdom, [17] Thomas Jefferson National Accelerator Facility, Newport News, VA, United States, [18] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, United States, [19] Institute of Nuclear and Particle Physics, Technische Universität Dresden, Dresden, Germany, [20] Department of Computer Science, Technical University Darmstadt, Darmstadt, Germany, [21] Department of Bioengineering, Lehigh University, Bethlehem, PA, United States, [22] Department of Physics, University of Florida, Gainesville, FL, United States, [23] Department of Physics, Yale University, New Haven, CT, United States, [24] Department of Engineering and IT, University of Sydney, Camperdown, NSW, Australia, [25] Department of Physics, Duke University, Durham, NC, United States, [26] Birla Institute of Technology and Science, Pilani, India, [27] Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, CA, United States, [28] Department of Physics, Princeton University, Princeton, NJ, United States, [29] Department of Computer Science, University of Houston, Houston, TX, United States, [30] Department of Physics, Universität Hamburg, Hamburg, Germany, [31] Department of Physics and Astronomy, Iowa State University, Ames, IA, United States

OPEN ACCESS

Edited by:
Elena Cuoco,
European Gravitational Observatory,
Italy

Reviewed by:
Andreea Anghel,
IBM Research-Zurich, Switzerland
Rudy Raymond,
IBM, Japan

***Correspondence:**
Allison McCarn Deiana
adeiana@smu.edu
Nhan Tran
ntran@fnal.gov

Specialty section:
This article was submitted to
Big Data and AI in High Energy
Physics,
a section of the journal
Frontiers in Big Data

Received: 30 September 2021
Accepted: 31 January 2020
Published: 12 April 2022

Citation:

Contents

Vision

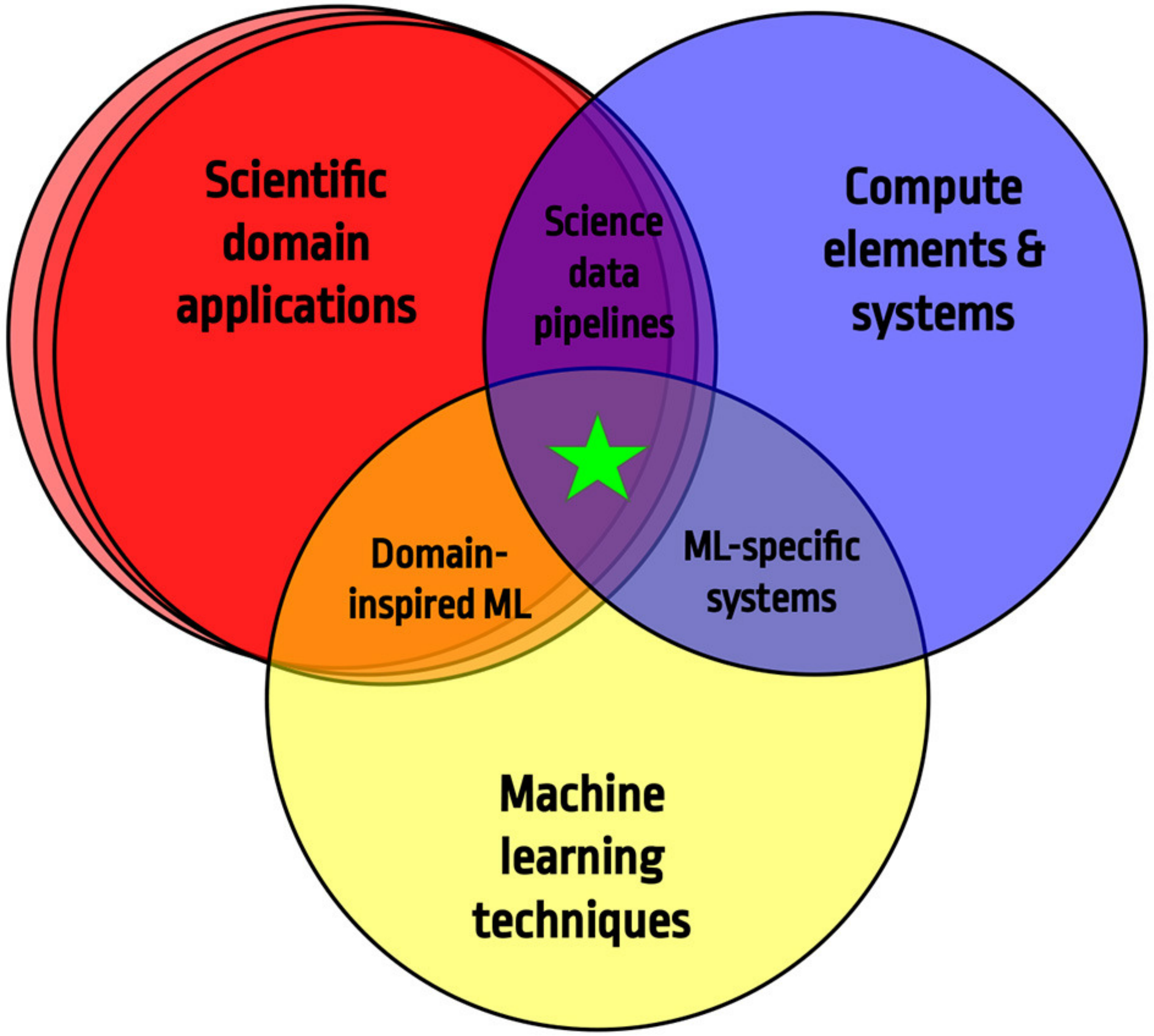
Snowmass 2021

BOX 1 | Fast machine learning in science.

Within this review paper, we refer to the concept of ***Fast Machine Learning in Science*** as the integration of ML into the experimental data processing infrastructure to enable and accelerate scientific discovery. Fusing powerful ML techniques with experimental design decreases the “time to science" and can range from embedding real-time feature extraction to be as close as possible to the sensor all the way to large-scale ML acceleration across distributed grid computing datacenters. The overarching theme is to lower the barrier to advanced ML techniques and implementations to make large strides in experimental capabilities across many seemingly different scientific applications. Efficient solutions require collaboration between domain experts, machine learning researchers, and computer architecture designers.

Vision

Snowmass 2021



“Necessarily, such a broad scope of topics *cannot* be comprehensive. For the scientific domains, we note that the contributions are *examples* of how ML methods are currently being or planned to be deployed. We hope that giving a glimpse into specific applications will inspire readers to find more novel use-cases and potential overlaps. The summaries of state-of-the- art techniques we provide relate to rapidly developing fields and, as such, may become out of date relatively quickly. The goal is to give non-experts an overview and taxonomy of the different techniques and a starting point for further investigation. To be succinct, we rely heavily on providing references to studies and other overviews while describing most modern methods.”

Sec 2: Domain Exemplars

- Large section on Large Hadron Collider because it is a technology driver for this community:
  - Event Reconstruction
  - Event Simulation
  - Heterogeneous Computing
  - Real-Time Analysis at 40 MHz
  - Bringing ML to Detector Front-End

Example use cases are not comprehensive, but representative (unique physics challenge)

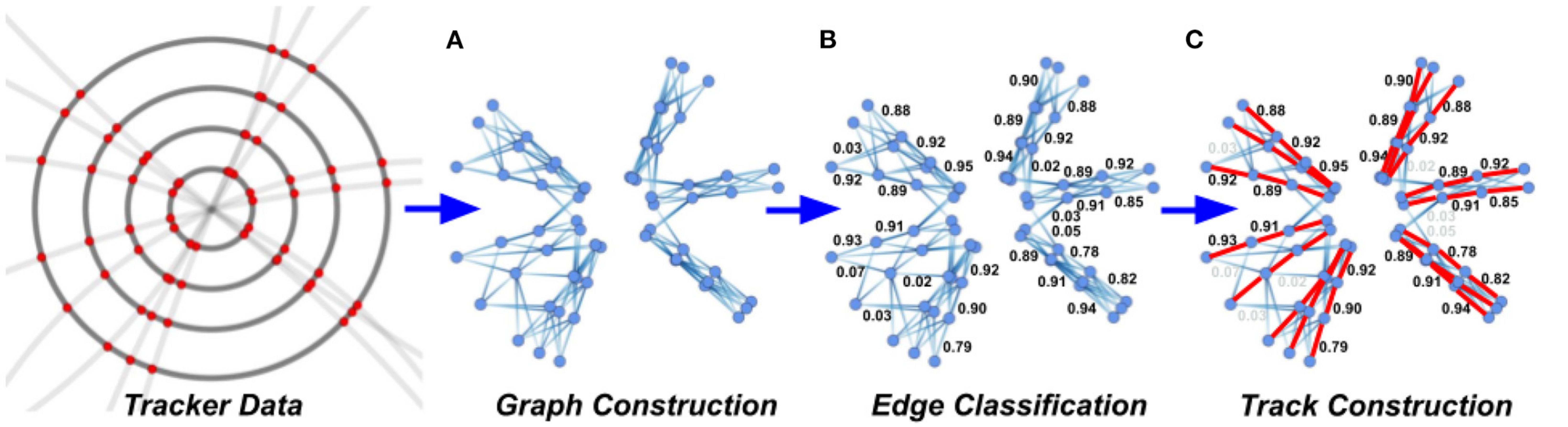


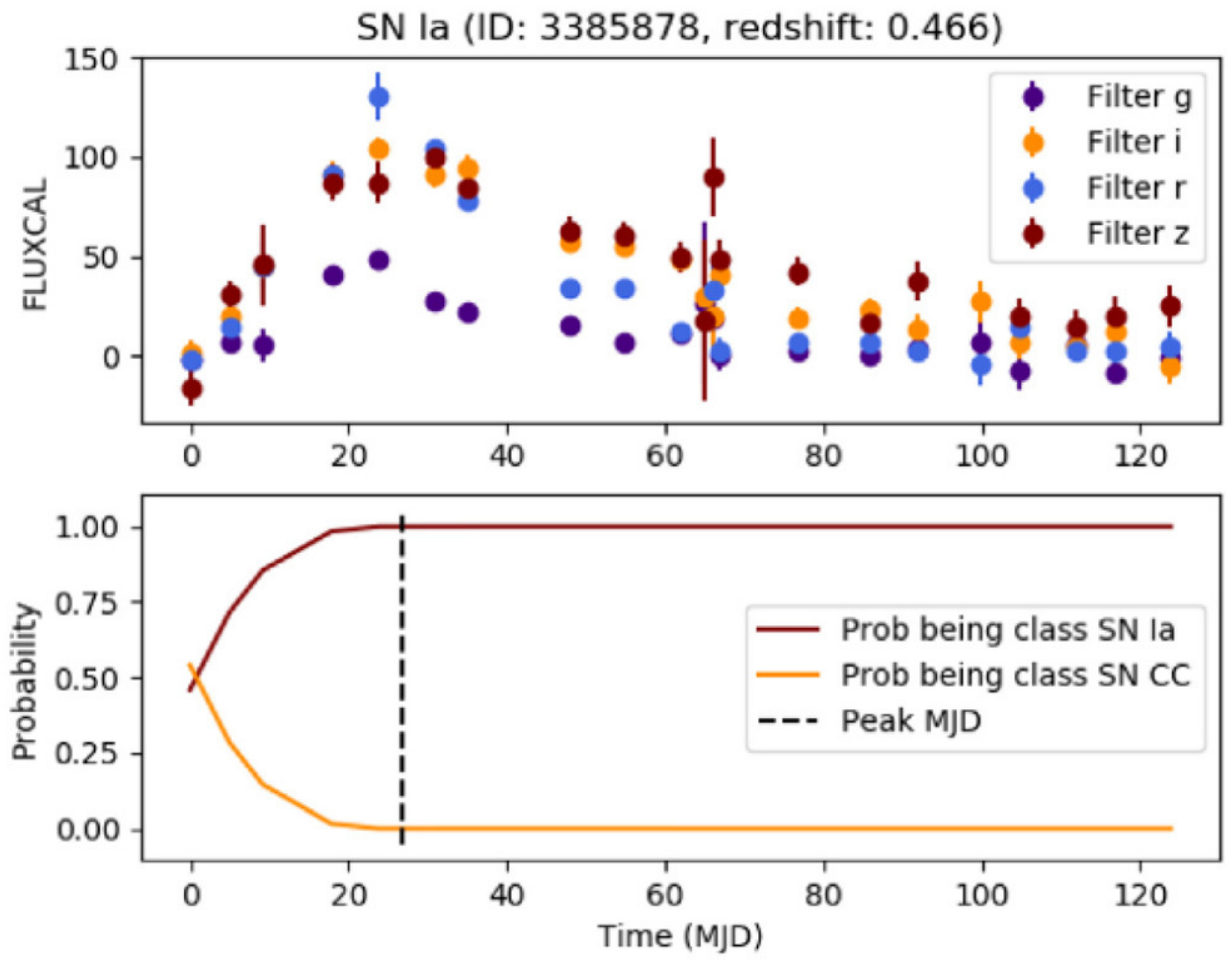
FIGURE 2 | High-level overview of the stages in a GNN-based tracking pipeline. Only a subset of the typical edge weights are shown for illustration purposes. **(A)** Graph construction, **(B)** edge classification, and **(C)** track construction.

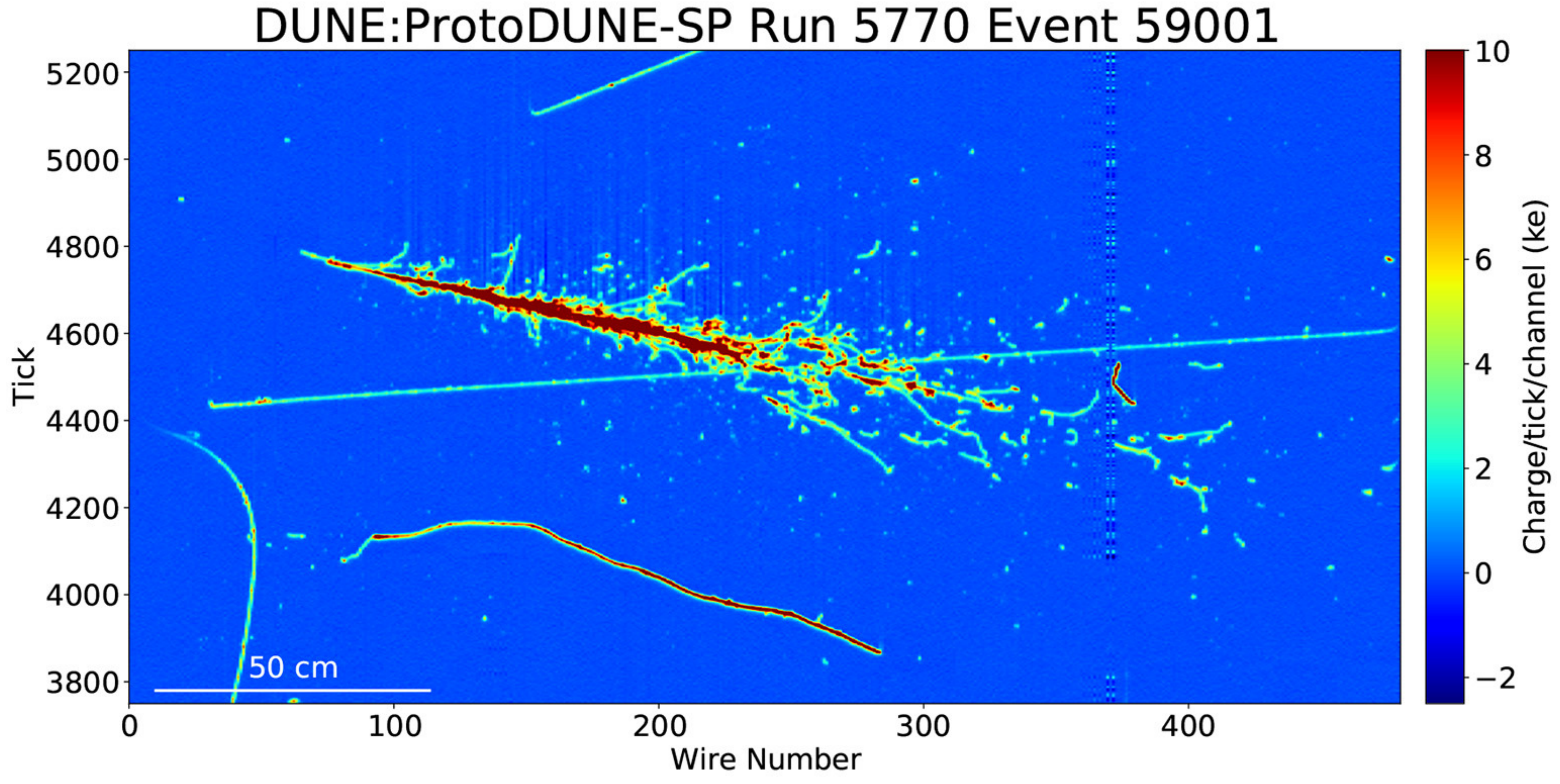
Exemplars of domain applications

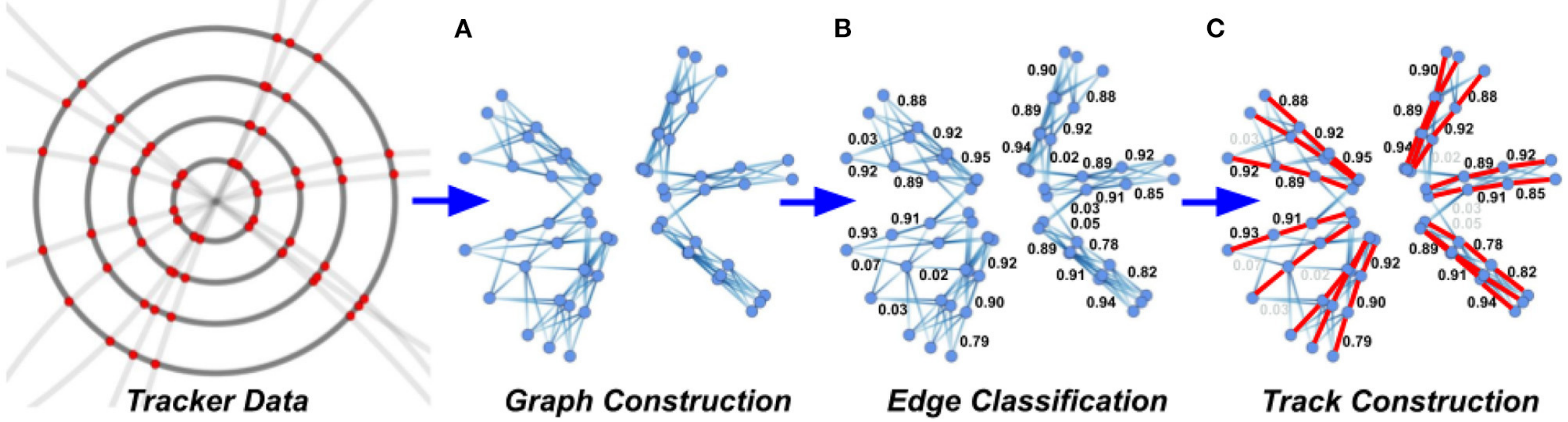
- High-intensity Accelerators: Belle II, Mu2e
- Materials Discovery: Materials Synthesis, Scanning Probe Microscopy
- Fermilab Accelerator Controls
- Neutrino/Dark Matter Experiments: e.g. DUNE, MINERvA, Direct Detection Dark Matter
- Electron-Ion Collider
- Gravitational Waves
- Health: Biomedical Engineering and Health Monitoring
- Cosmology
- Plasma Physics
- Wireless Networking and Edge Computing

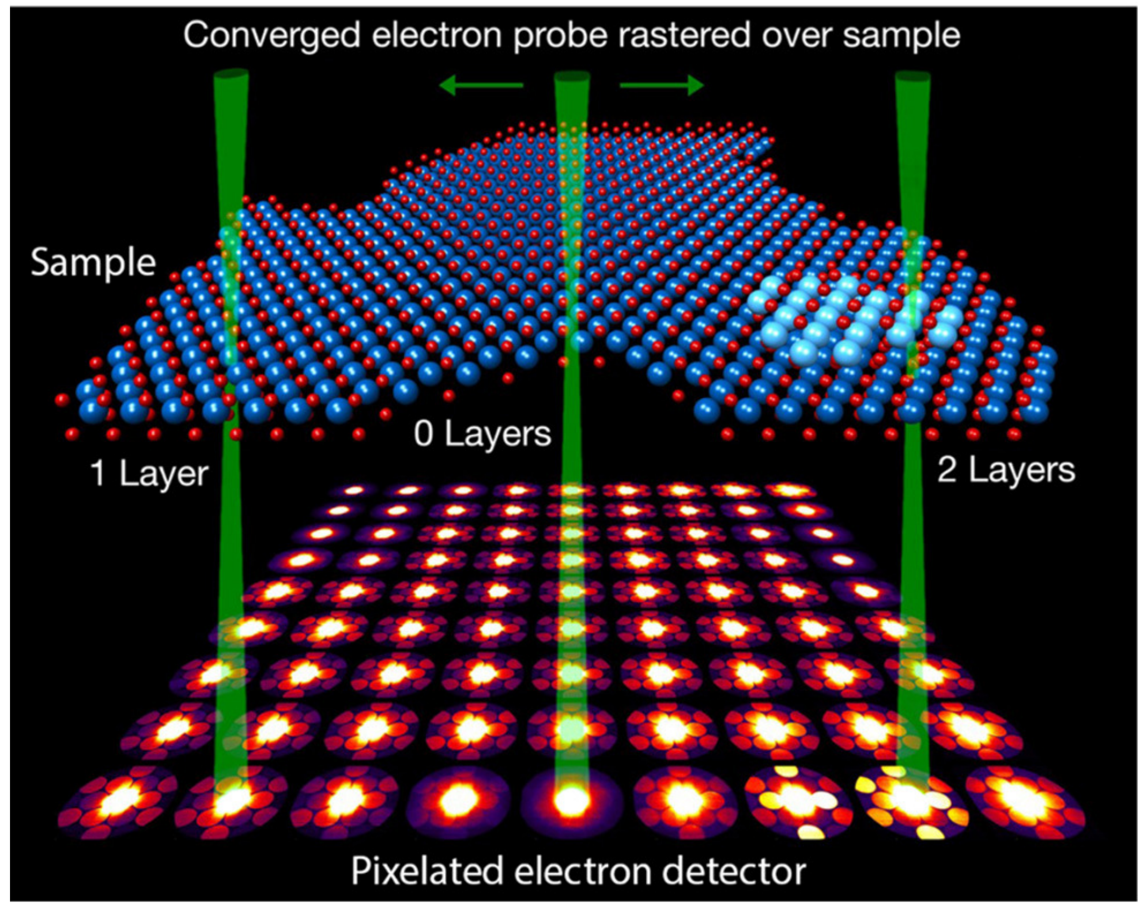
Post publishing - new domains called out including neuroscience and x-ray spectroscopy

Sec 3: Areas of overlap - representations









Areas of overlap - representations

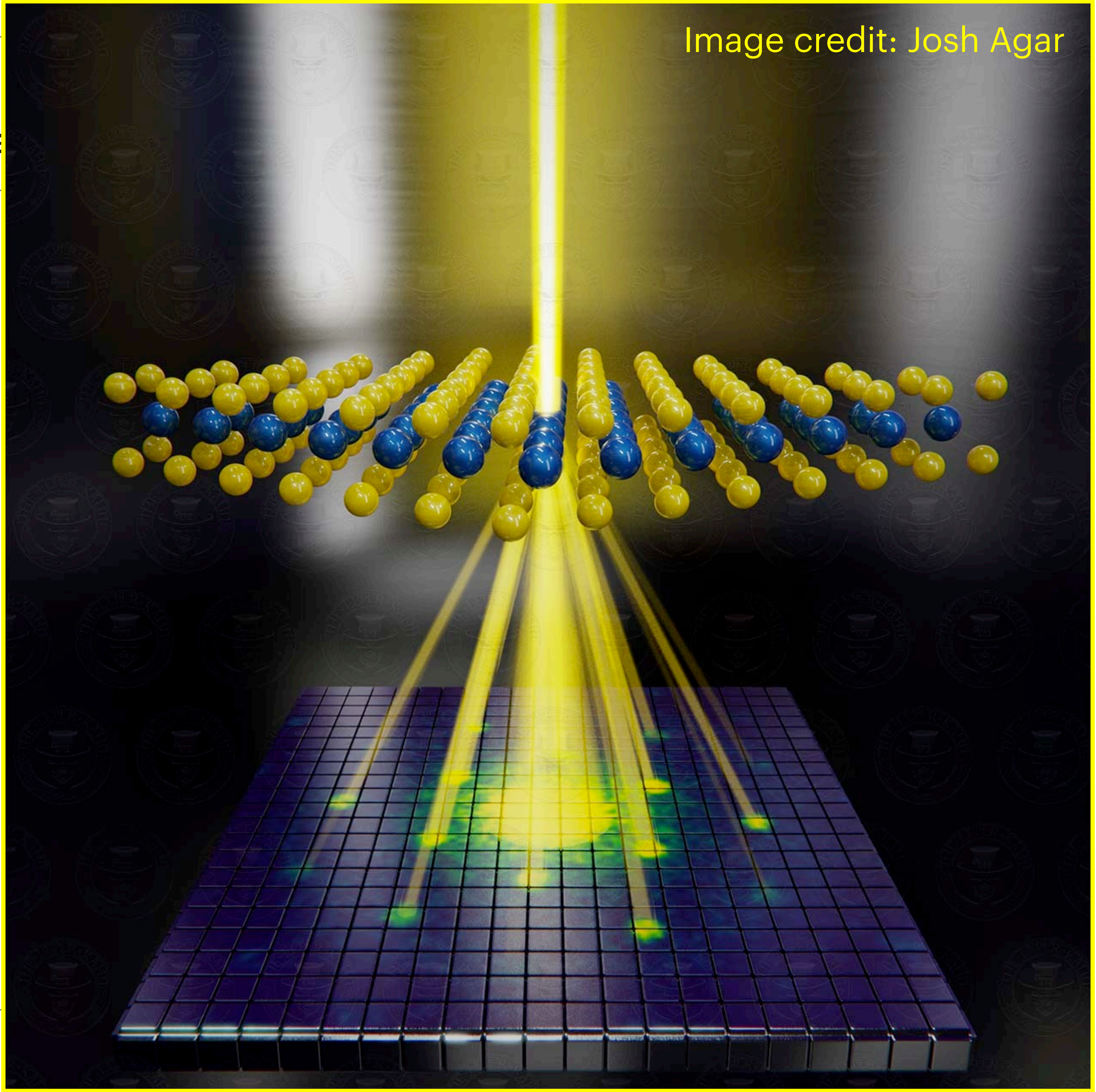
TABLE 1 | Types of data representations and their relevance for the scientific domains discussed in this paper; ✓✓= Particularly important for domain, ✓= Relevant for domain.

Domain	Spatial	Point cloud	Temporal	Spatio-Temporal	Multi/Hyper-spectral	Examples
LHC	✓✓	✓✓	✓	✓	–	Detector reconstruction
Belle-II/Mu2e	✓✓	✓✓	–	–	–	Track reconstruction
Material Synthesis	✓	–	✓	✓✓	✓✓	High-speed plasma imaging
Accelerator Controls	✓	–	✓✓	–	–	Beam sensors
Accelerator neutrino	✓✓	✓✓	✓	✓	–	Detector reconstruction
Direct detection DM	✓✓	✓✓	✓	✓	–	Energy signatures
EIC	✓✓	✓✓	✓	✓	–	Detector reconstruction
Gravitational Waves	✓	–	✓✓	–	–	Laser inference patterns
Biomedical engineering	✓✓	–	–	✓✓	–	Cell and tissue images
Health Monitoring	✓	–	✓✓	✓	✓	Physiological sensor data
Cosmology	✓✓	✓✓	✓✓	✓	✓✓	Lensing/radiation maps
Plasma Physics	✓	–	✓✓	✓	–	Detector actuator signals
Wireless networking	–	–	✓✓	–	–	Electromagnetic spectrum

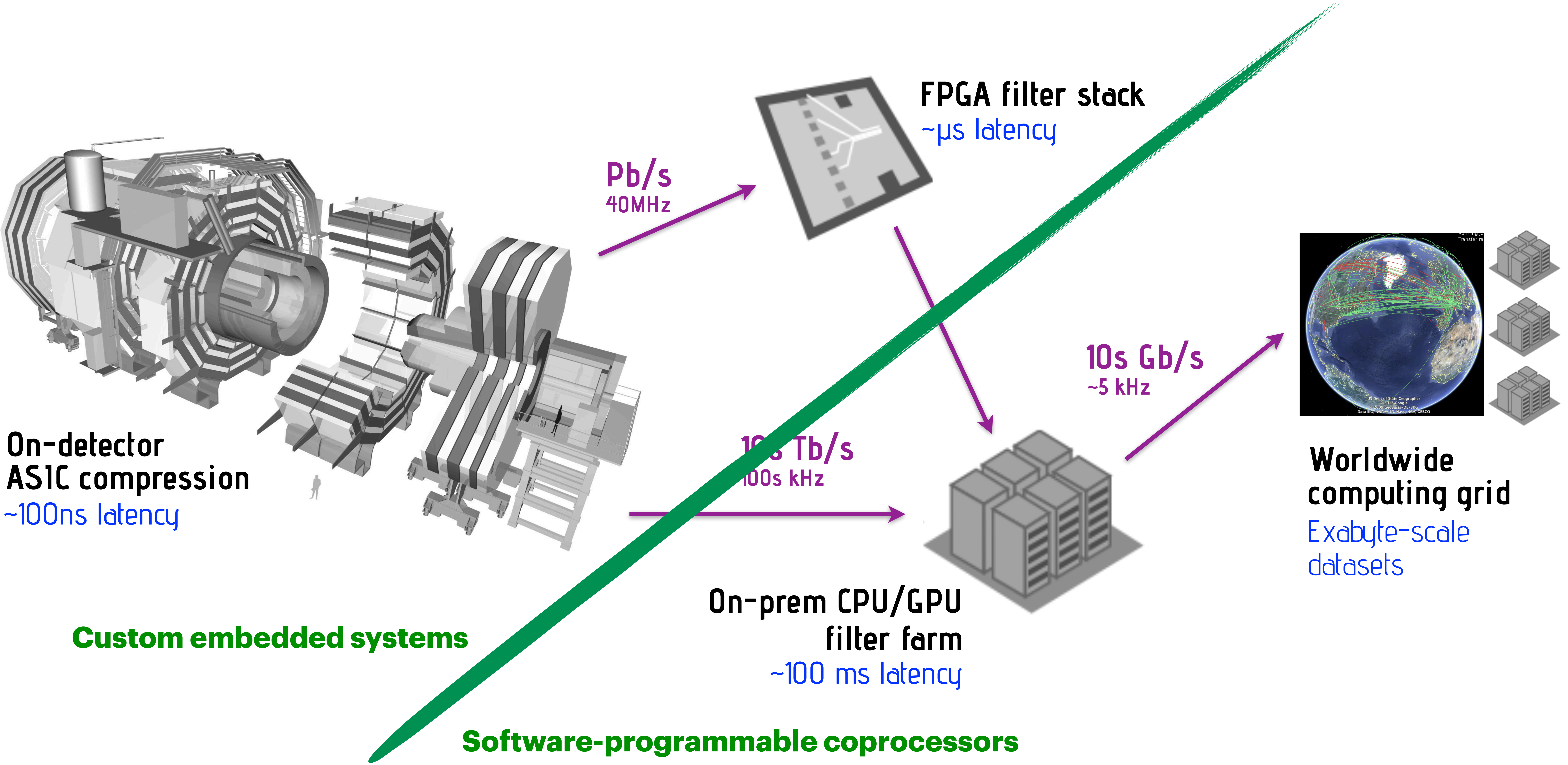
Areas of overlap - representations

TABLE 1 | Types of data representations and their relevance for the scientific domains discussed in this paper; ✓✓= Particularly important for domain, ✓= Relevant for domain.

Domain	Spatial	Point cloud	Temporal	Spatio- Tempora
LHC	✓✓	✓✓	✓	✓
Belle-II/Mu2e	✓✓	✓✓	–	–
Material Synthesis	✓	–	✓	✓✓
Accelerator Controls	✓	–	✓✓	–
Accelerator neutrino	✓✓	✓✓	✓	✓
Direct detection DM	✓✓	✓✓	✓	✓
EIC	✓✓	✓✓	✓	✓
Gravitational Waves	✓	–	✓✓	–
Biomedical engineering	✓✓	–	–	✓✓
Health Monitoring	✓	–	✓✓	✓
Cosmology	✓✓	✓✓	✓✓	✓
Plasma Physics	✓	–	✓✓	✓
Wireless networking	–	–	✓✓	–



Areas of overlaps - systems



Areas of overlaps - systems

Snowmass 2021

TABLE 2 | Domains and practical constraints: systems are broadly classified as soft (software-programmable computing devices: CPUs, GPUs, and TPUs) and custom (custom embedded computing devices: FPGAs and ASICs).

Domain	Event rate	Latency	Systems	Energy-constrained
Detection and event reconstruction				**No**
LHC and intensity frontier HEP	10s Mhz	ns-ms	Soft/custom	
Nuclear physics	10s kHz	ms	Soft	
Dark matter and neutrino physics	10s MHz	μs	Soft/custom	
Image processing				
Material synthesis	10s kHz	ms	Soft/custom	
Scanning probe microscopy	kHz	ms	Soft/custom	
Electron microscopy	MHz	μs	Soft/custom	
Biomedical engineering	kHz	ms	Soft/custom	Yes (mobile settings)
Cosmology	Hz	s	Soft	
Astrophysics	kHz–MHz	ms-us	Soft	Yes (remote locations)
Signal processing				
Gravitational waves	kHz	ms	Soft	
Health monitoring	kHz	ms	Custom	Yes
Communications	kHz	ms	Soft	Yes (mobile settings)
Control systems				
Accelerator controls	kHz	ms–μs	Soft/custom	
Plasma physics	kHz	ms	Soft	

Areas of overlaps - feedback

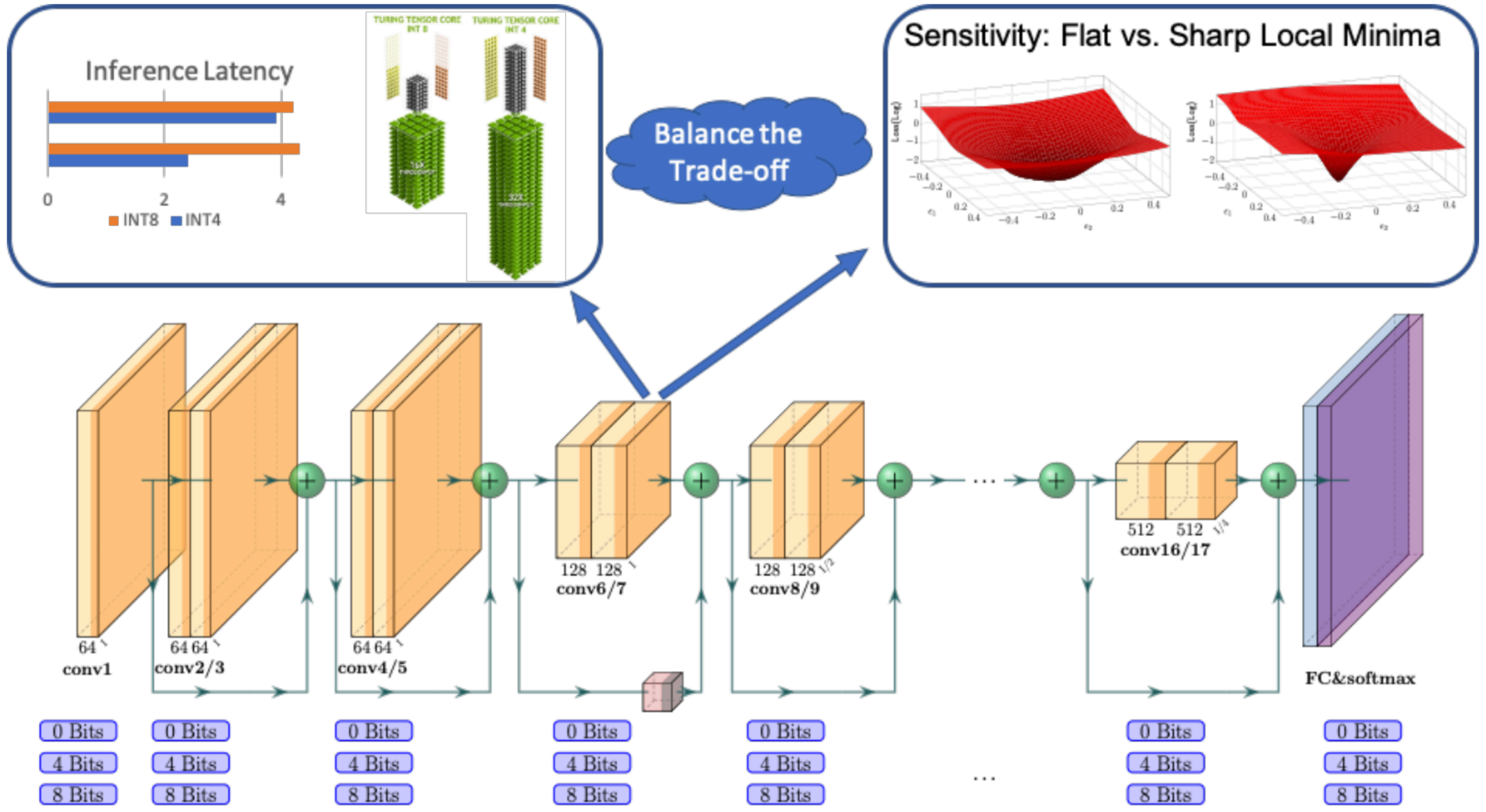
Snowmass 2021

TABLE 3 | Classification of domains and their system requirements with respect to real-time needs.

Domain	Real-time data reduction	Real-time analysis	Closed-loop control
Detection/Event reconstruction			
LHC	Yes	Yes	No
Nuclear physics	Yes	No	No
Dark matter-neutrino	Yes	No	No
Image processing			
Material synthesis	Yes	Yes	Yes
Scanning probe microscopy	Yes		
Electron microscopy	Yes		
Biomedical engineering	Yes		
Cosmology	Yes	No	No
Astrophysics	Yes	No	No
Signal processing			
Gravitational waves	Yes	No	No
Health monitoring	Yes	Yes	Yes
Communications	Yes	Yes	Yes
Control systems			
Accelerator controls	Yes	Yes	Yes
Plasma physics	Yes	Yes	Yes

Sect 4: Efficient ML

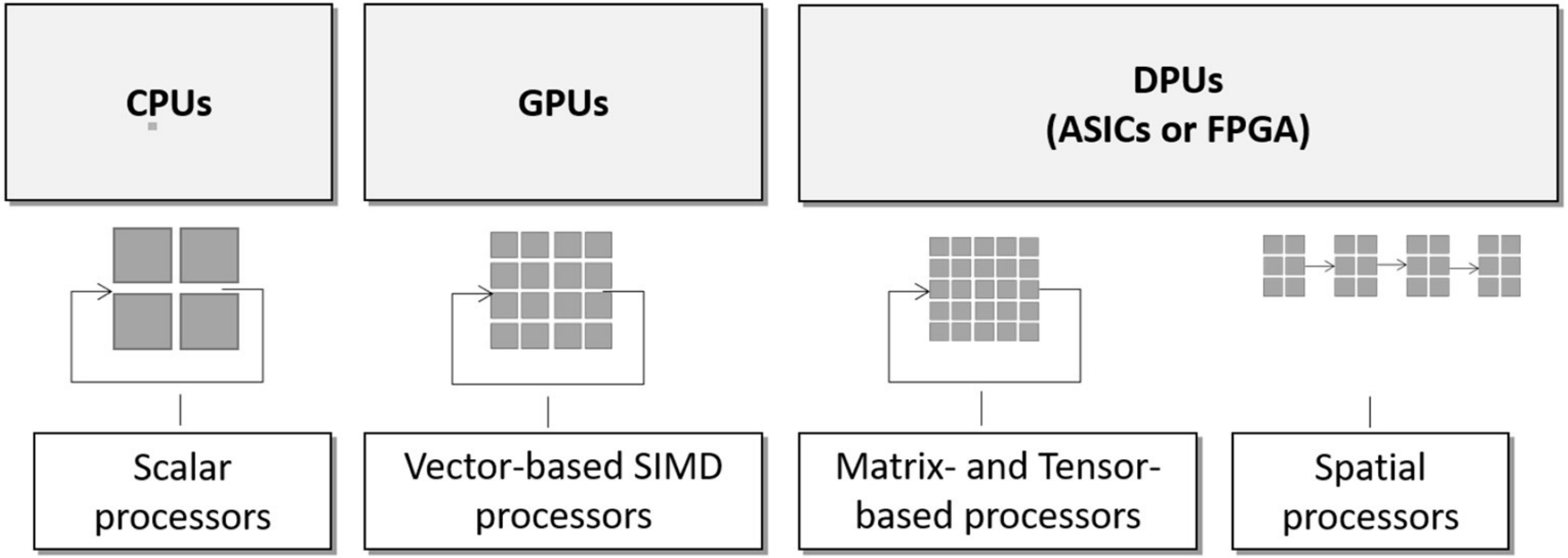
- **A discussion of strategies for improving ML efficiency to enable lower latency**
  - Designing new efficient ML architectures
  - NN & hardware co-design
  - Quantization
  - Pruning and sparse inference
  - Knowledge distillation
- Discussion of automation of the NN architecture design process (Neural Architecture Search).



- **Not mentioned in WP but important!**
  - **Fault-tolerant, reliable ML**

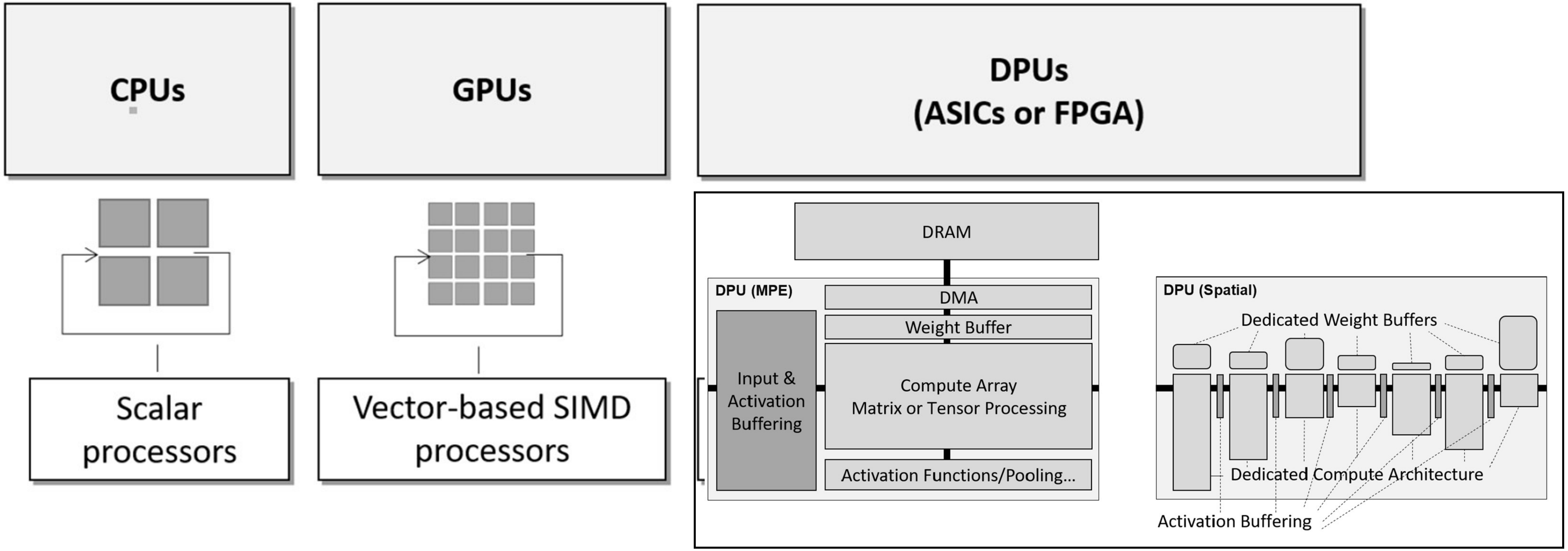
Section 4: Hardware Architecture

- Discussion of different computing architectures: CPU, GPU, FPGA/ASIC
- DPU: Deep learning processing unit, customized for CNNs. These can be implemented on FPGAs or ASICs.



Section 4: Hardware Architecture

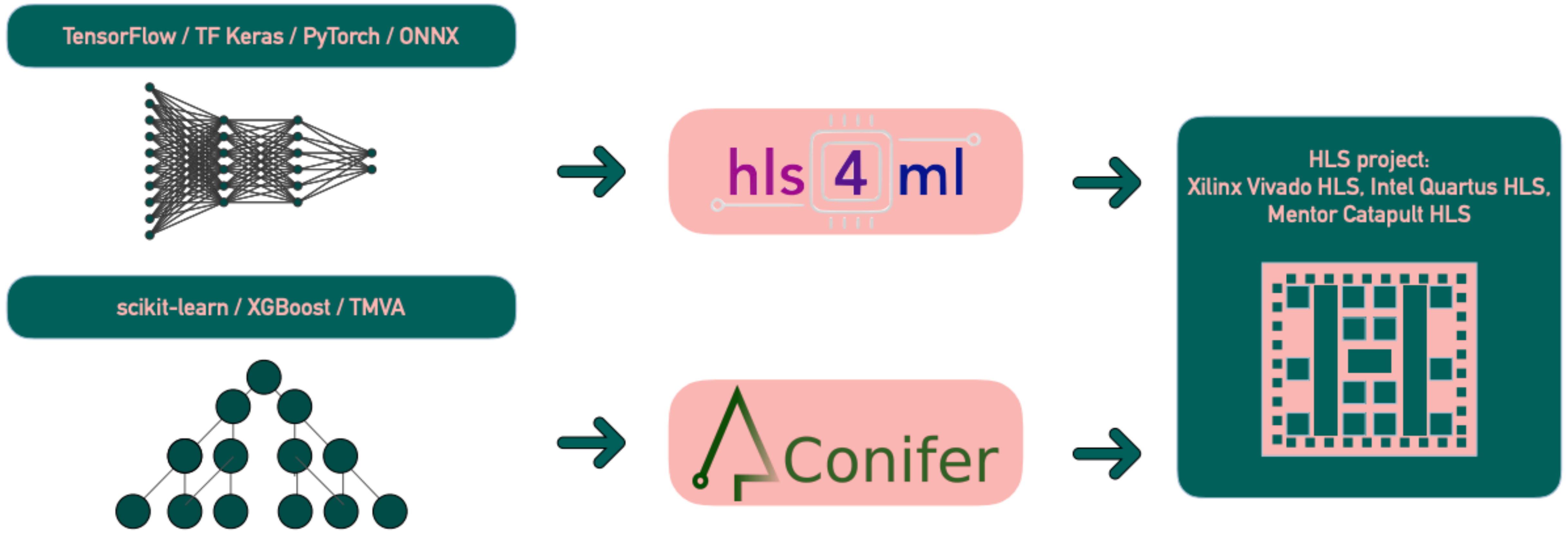
- Discussion of different computing architectures: CPU, GPU, FPGA/ASIC
- DPU: Deep learning processing unit, customized for CNNs. These can be implemented on FPGAs or ASICs.

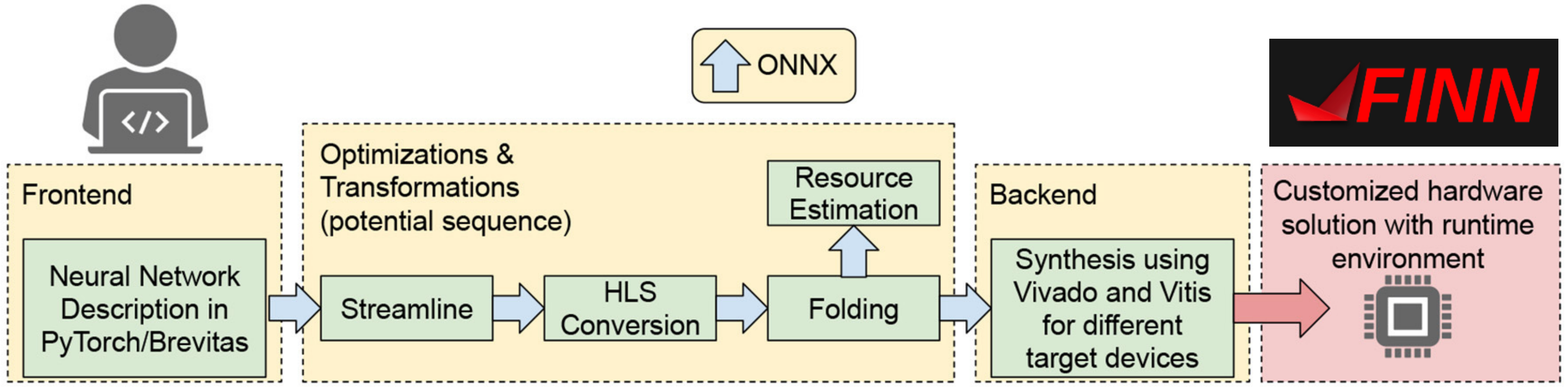


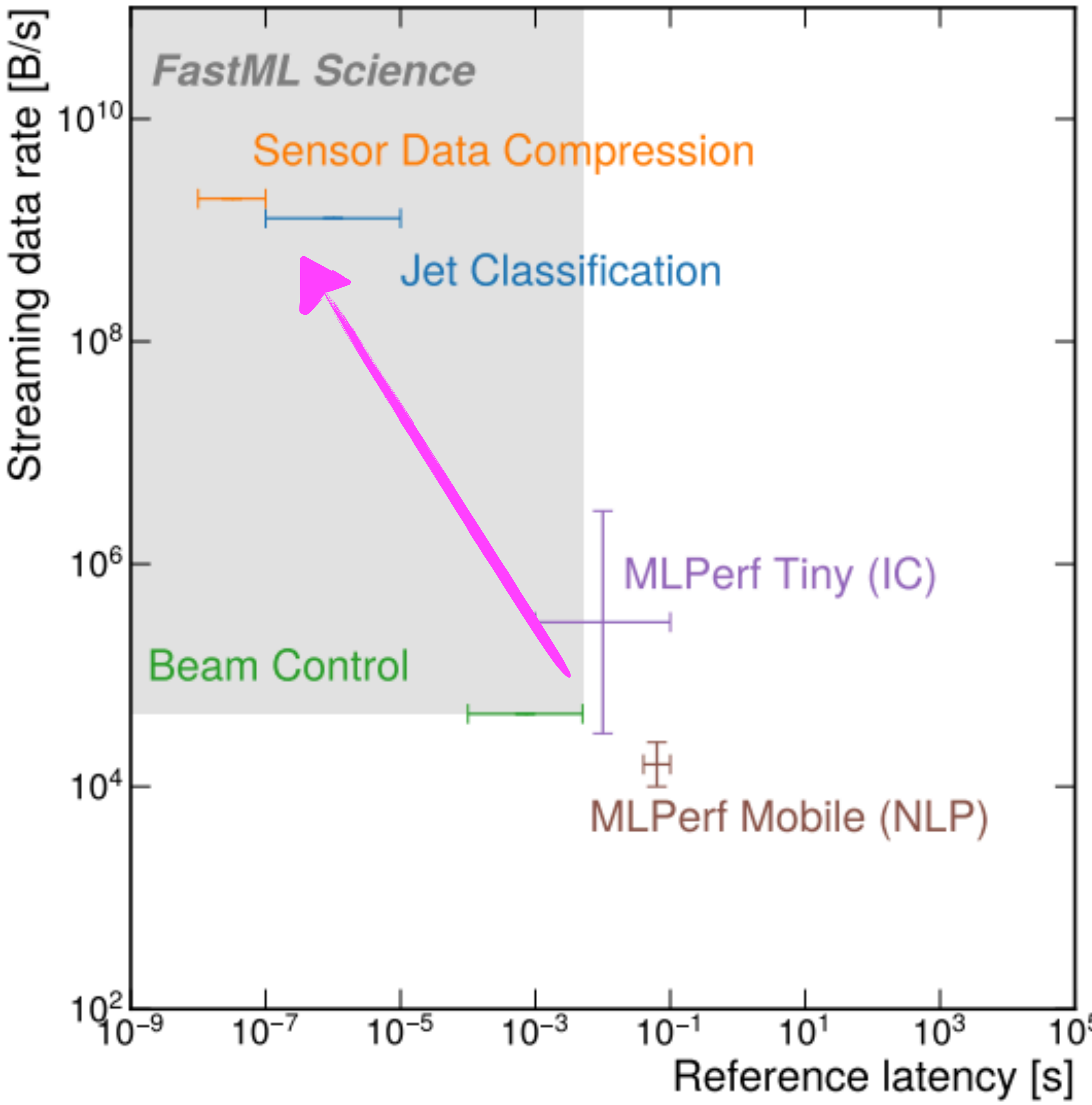
Sec 4: codesign

Snowmass 2021

https://arxiv.org/abs/2207.07958







MLIR

Multi-Level IR Compiler Framework

CIRCT

Circuit IR Compilers and Tools

Beyond-CMOS Neuromorphic Hardware

- In this section, the most prominent emerging technology proposals, including those based on emerging dense analog memory device circuits, are grouped according to the targeted low-level neuromorphic functionality.
  - Analog Vector-by-Matrix Multiplication
  - Stochastic Vector-by-Matrix Multiplication
  - Spiking Neuron and Synaptic Plasticity
  - Reservoir Computing
  - Hyperdimensional Computing / Associative Memory

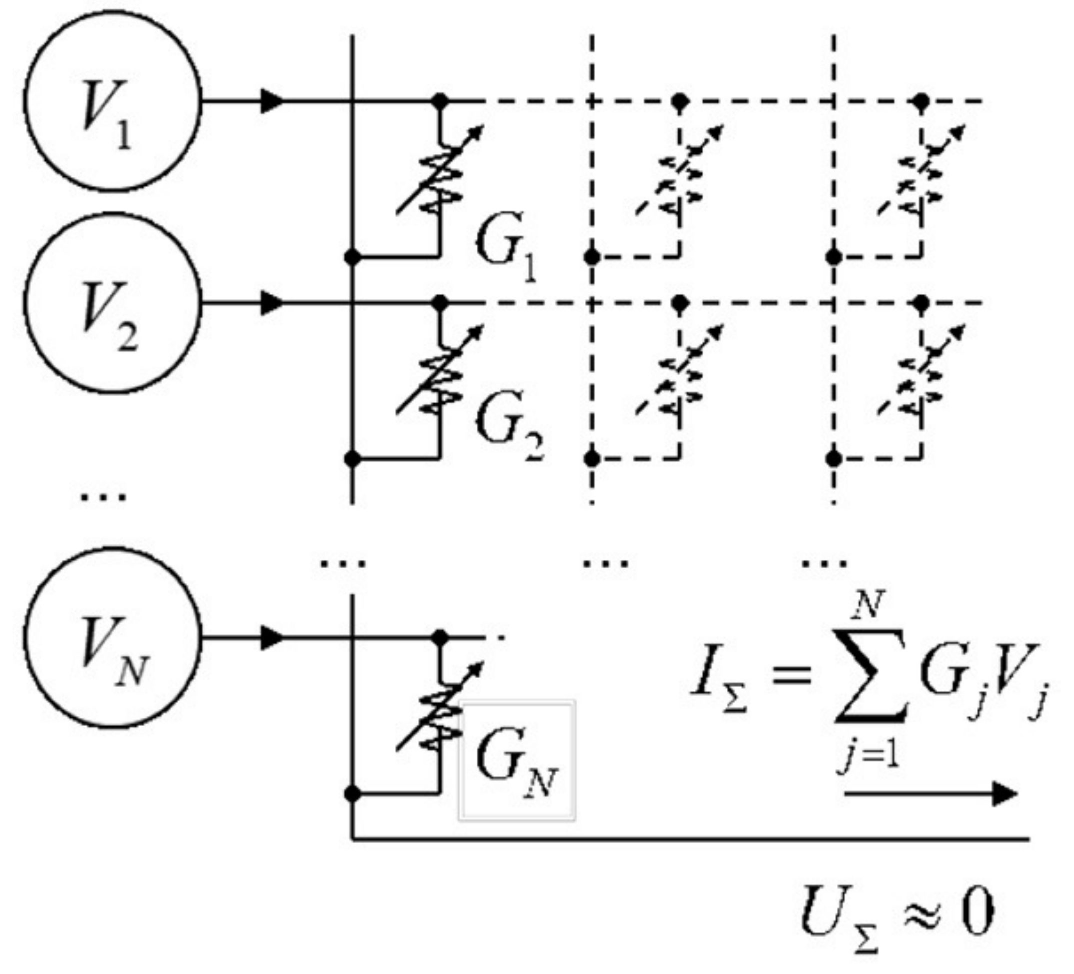
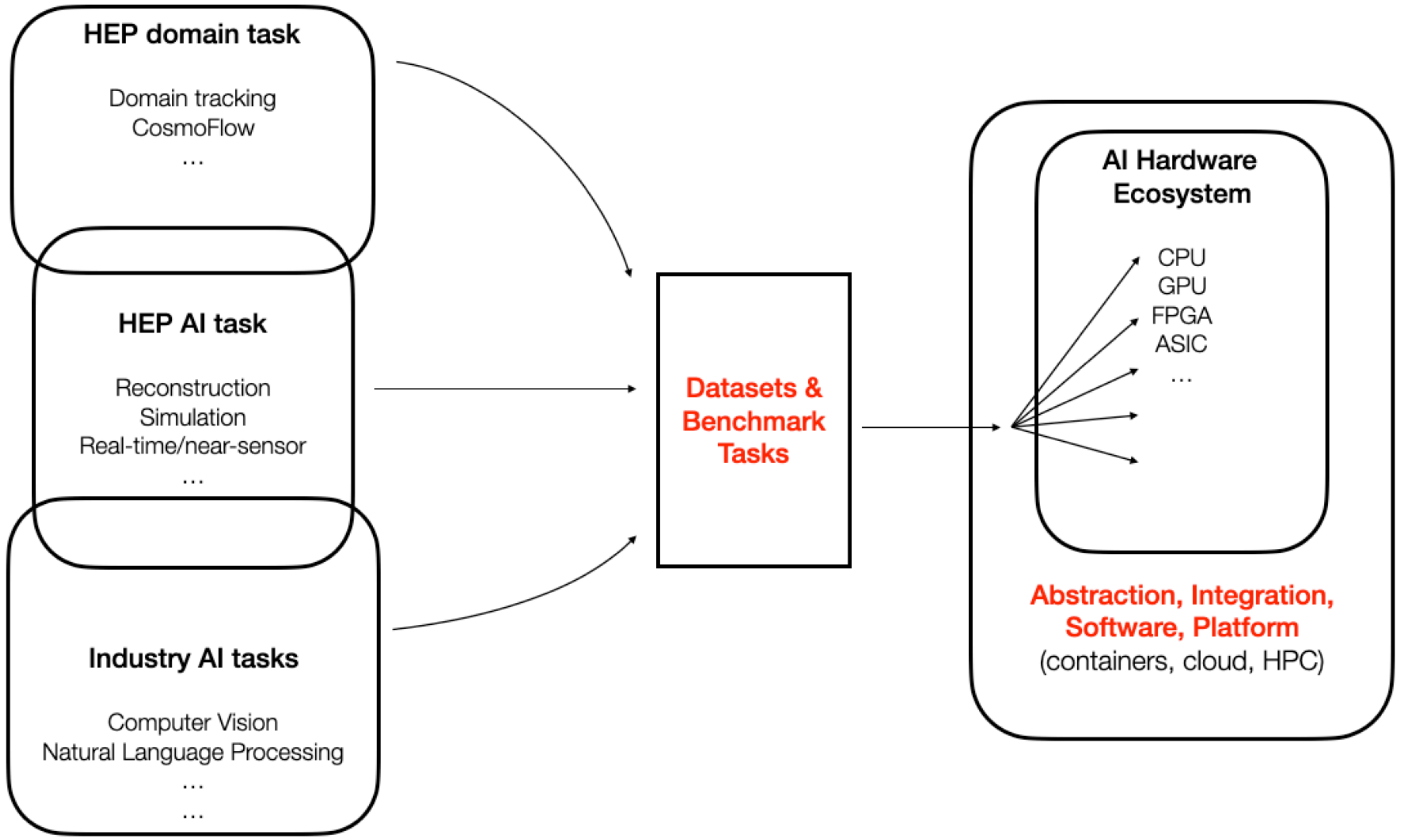


FIGURE 10 | Analog vector-by-matrix multiplication (VMM) in a crossbar circuit with adjustable crosspoint devices. For clarity, the output signal is shown for just one column of the array, while sense amplifier circuitry is not shown. Note that other VMM designs, e.g., utilizing duration of applied voltage pulses, rather than their amplitudes, for encoding inputs/outputs, are now being actively explored see, e.g., their brief review in Bavandpour et al. (2018).

Connections

- **CompF3: ML**
  - WP on "Physics Community Needs, Tools, and Resources for Machine Learning", arXiv: 2203.16255
  - Related talk
- **CompF4: Storage & Processing**
  - Subsection on AI hardware
  - AI Hardware talk
- **IF07: Electronics/ASICS**
  - WP on "Smart sensors using artificial intelligence for on-detector electronics and ASICs", arXiv:2204.13223



Parting thoughts

Promote interdisciplinary collaborations
physicists, computer scientists, electrical and computer engineers, software engineers

Custom embedded systems

Off-the-shelf coprocessors

Build open-source, multi-technology codesign workflows

Be nimble: abstraction, portability, containerization

Novel ML research concepts: efficient, fault-tolerant, reliable

Open data, task-based, and data-based benchmarks

Support ecosystem integration and operation

Extra

Parting thoughts

Snowmass 2021

Promote interdisciplinary collaborations
physicists, computer scientists, electrical and computer engineers, software engineers

Extremely valuable to learn from non-domain expertise; challenge is to find common goals

Custom embedded systems

Off-the-shelf coprocessors

here, our problems surpass industry and there are no OTS solutions

Build open-source, multi-technology codesign workflows

Be nimble: abstraction, portability, containerization

We are at the whim of industry! Adapt to new technologies

Our problems can inspire new technologies and techniques!

Novel ML research concepts: efficient, fault-tolerant, reliable

Open data, task-based, and data-based benchmarks

Catalyze and consolidate progress

Support ecosystem integration and operation

Projectization makes longevity and support very hard, need avenues for this