

# How can ML go beyond traditional project/frontier boundaries?

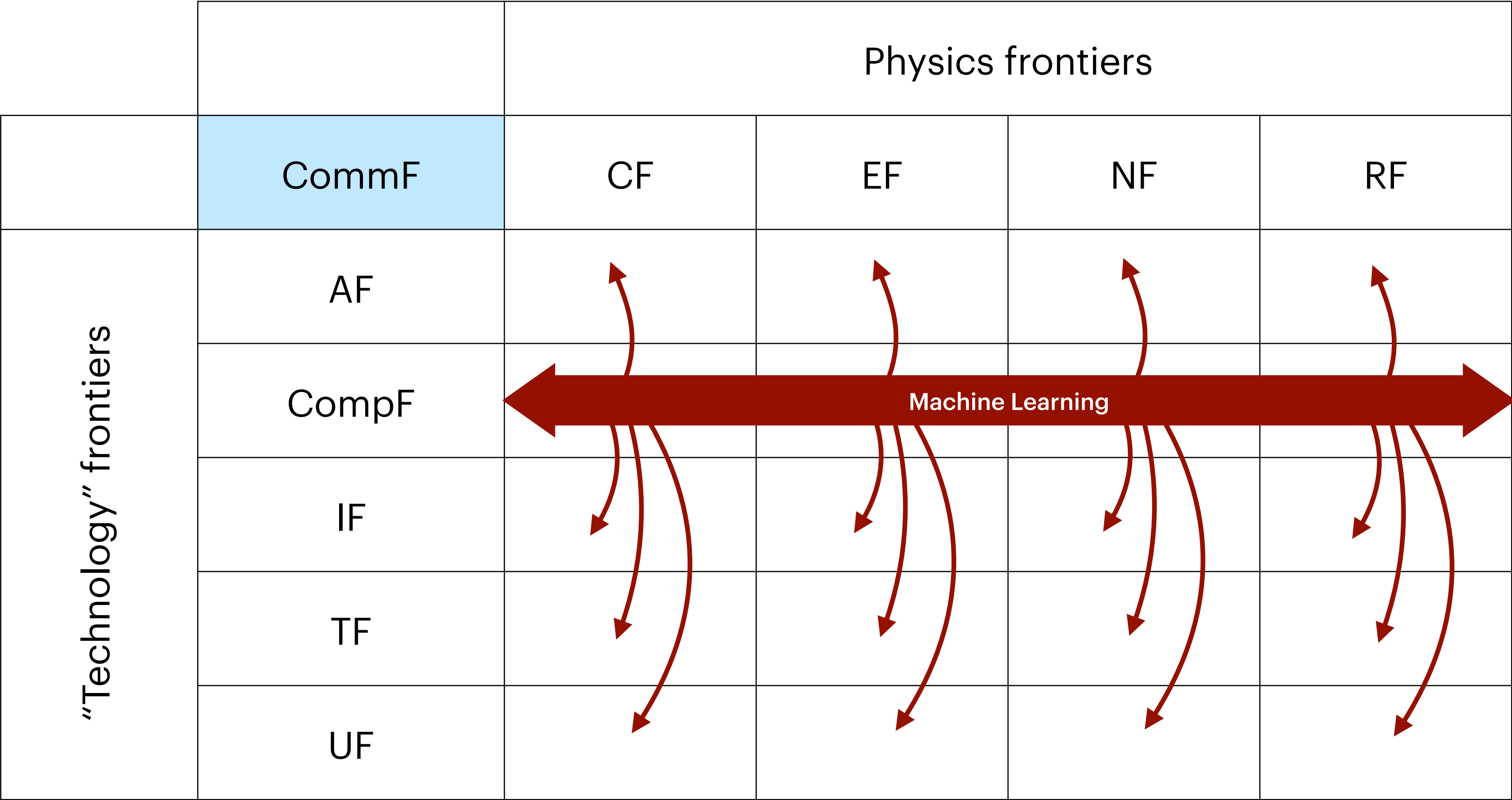
Nhan Tran, Fermilab

July 18, 2022

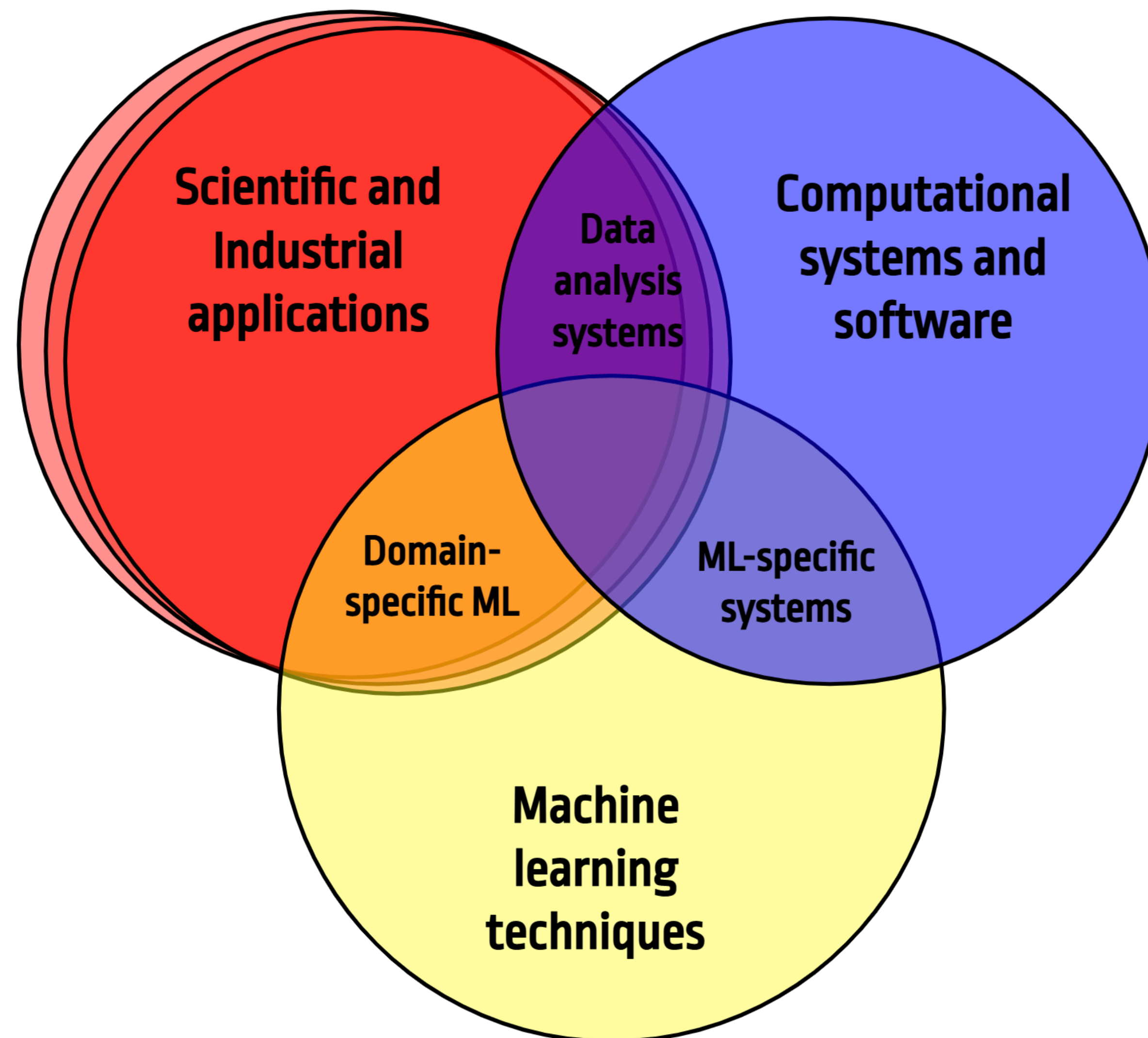
Seattle Snowmass Summer Meeting 2022



# Across projects, frontiers,...



# ... and beyond



# Message

- **ML can, is, and will improve (the way we do) physics**
- ML spans traditional boundaries
  - We should not stovepipe in traditional silos
  - Seemingly unrelated topics closely related and benefit from crosstalk
- Promote interdisciplinary exploration and teams
  - Inside **and** outside our particle physics community
  - ML techniques and research growing rapidly from many sources

# Rest of talk outline

- Machine learning for particle physics
- Particle physics for machine learning

*Disclaimer:*

*Examples throughout the talk based primarily on personal familiarity.*

*There are many (many!) other instances of exciting work.*

**<https://iml-wg.github.io/HEPML-LivingReview/>**

# Machine learning for particle physics

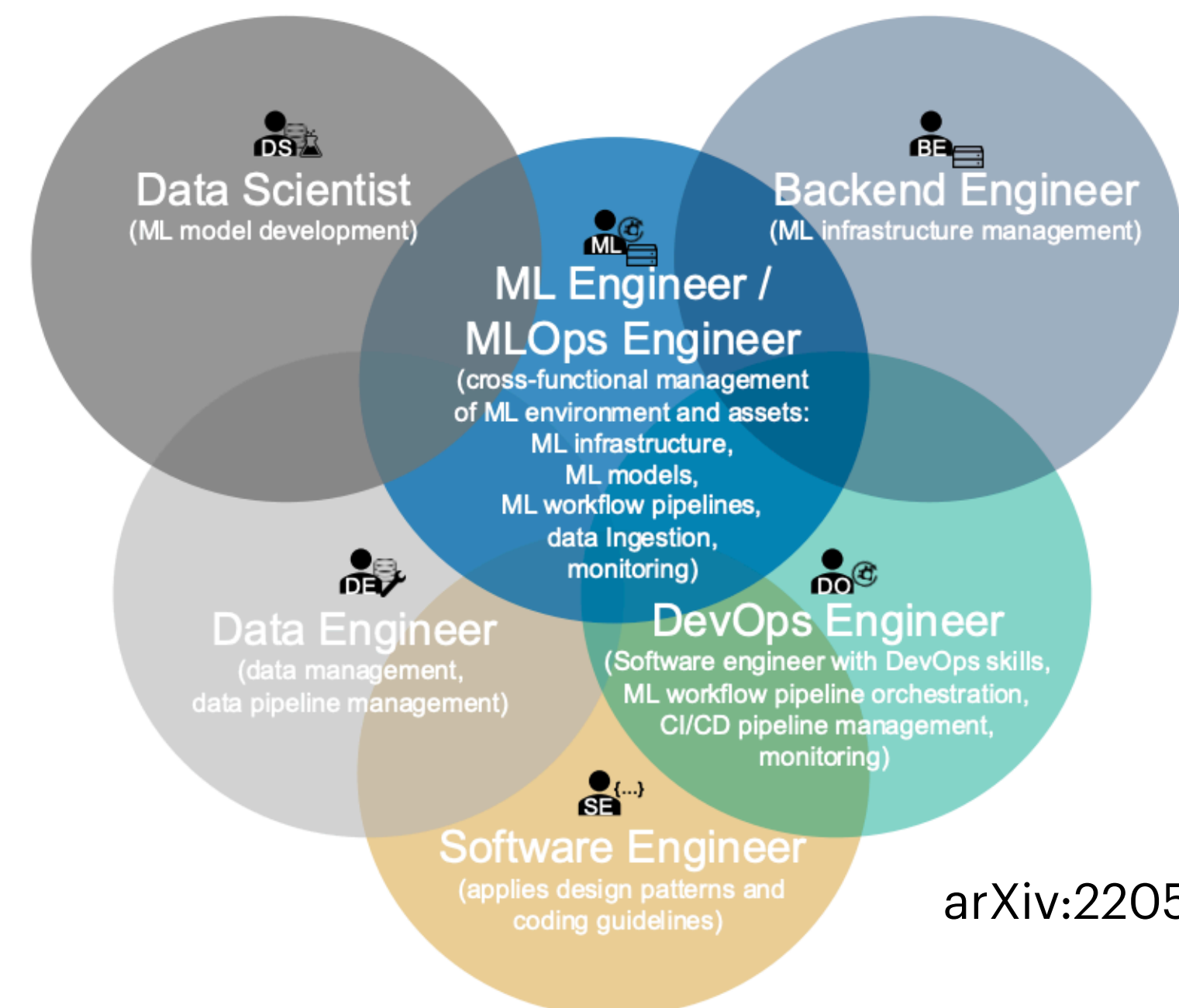
# Why should we care about (deep) ML?

- **Improves our science**

- See Daniel Whiteson's talk on physics and ML in the deep learning era
- See David Shih's talk on areas of physics opportunities for ML

- **We are not alone in deploying ML**

- **Training:** it can be a valuable skill to develop for early career scientists
- Conversely, many early career scientists are enthusiastic about developing machine learning for physics - it is **pervasive**



MLOps  
arXiv:2205.02302



# Traversing traditional boundaries

- **Algorithm-external:**

Domain cross-over

- Task-based
- Data representations
- Experimental system and data processing constraints
- Software, tools, education, training

- **Algorithm-internal:**

Cross-cutting ML themes



- Physics-constraints, interpretability
- Domain adaptation, fault tolerance, uncertainty quantification
- Efficient, resource-constrained



# Traversing traditional boundaries

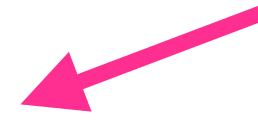
- **Algorithm-external:**

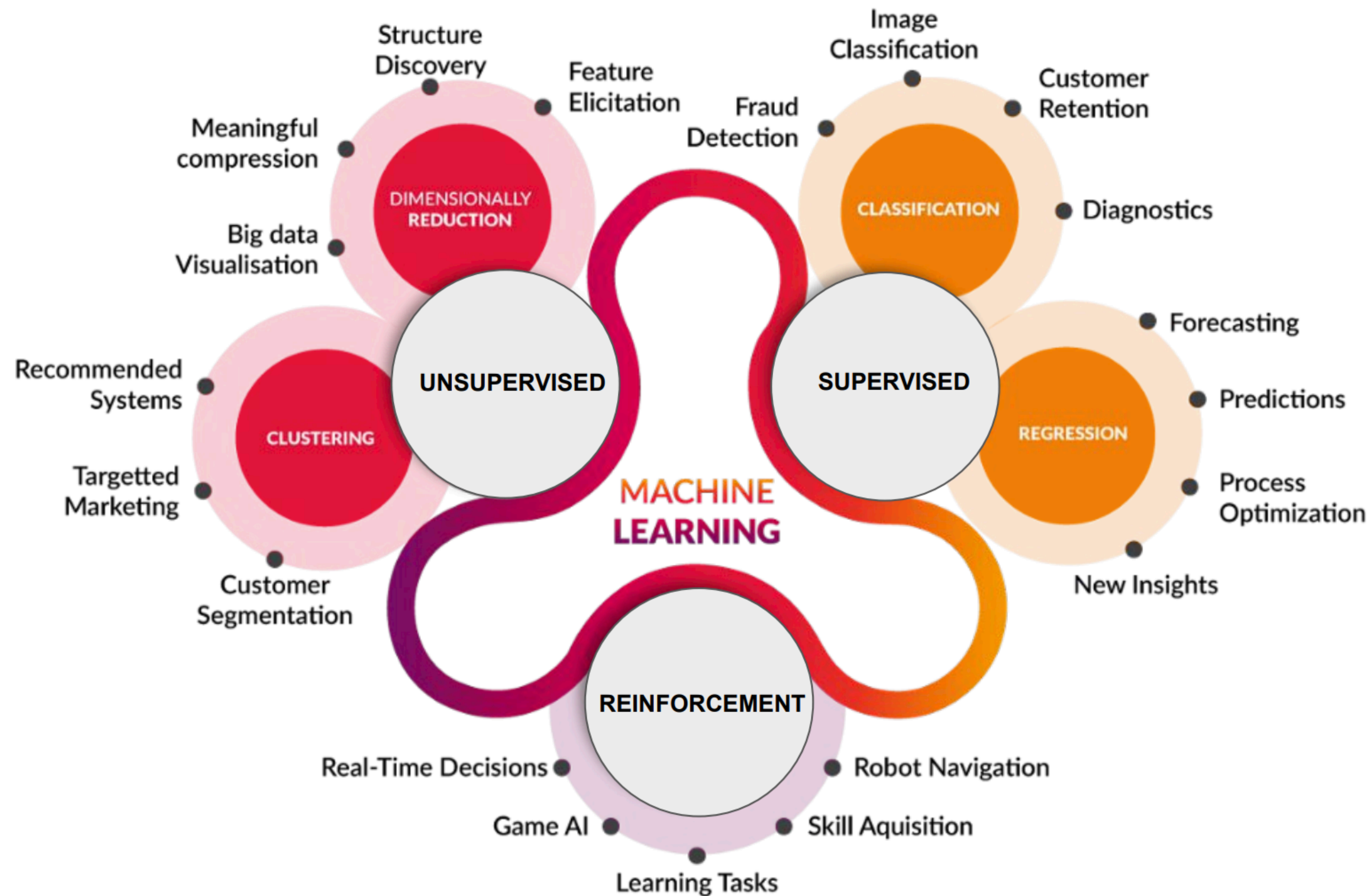
Domain cross-over

- Task-based 
- Data representations 
- Experimental system and data processing constraints
- Software, tools, education, training

- **Algorithm-internal:**

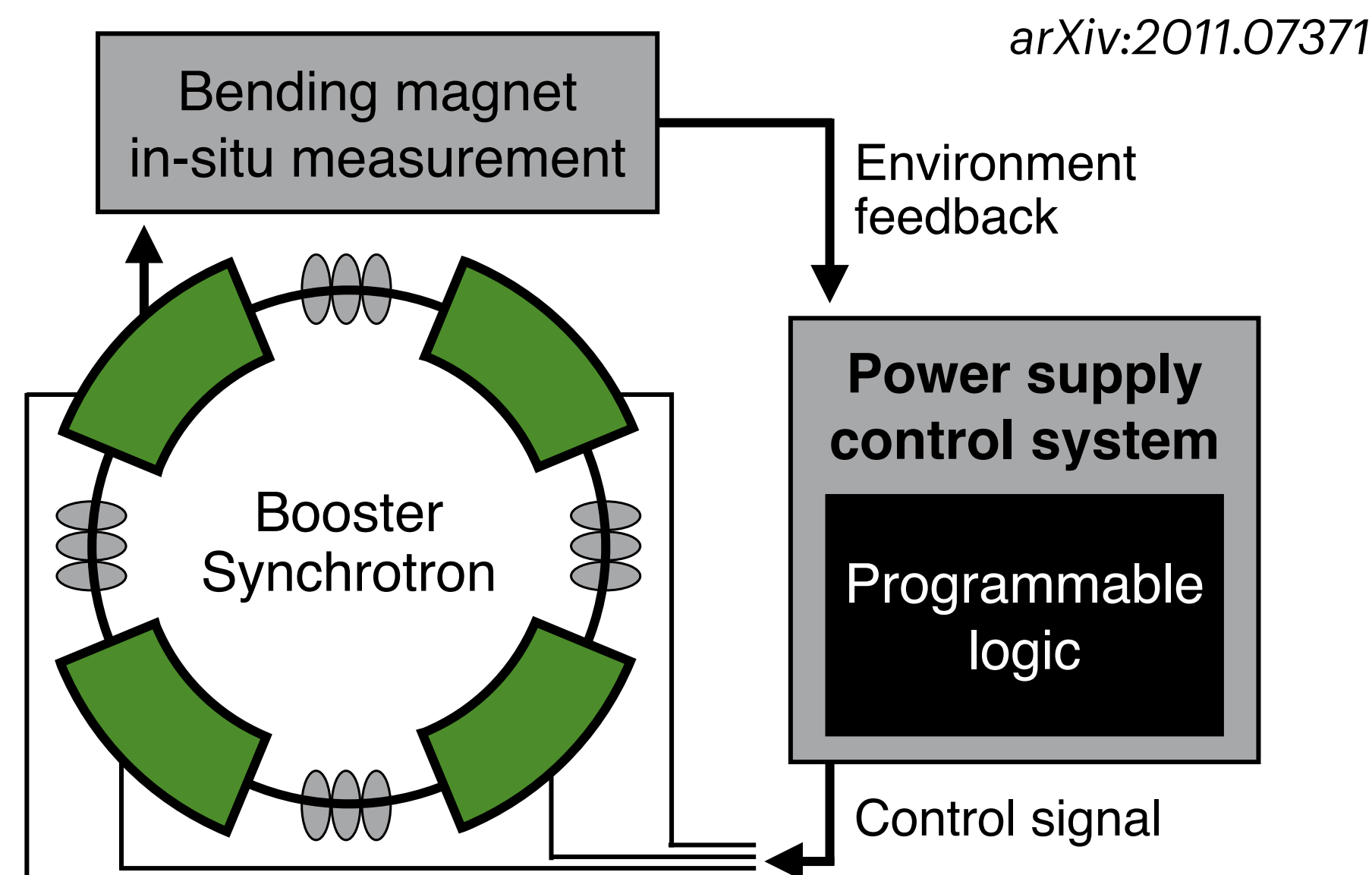
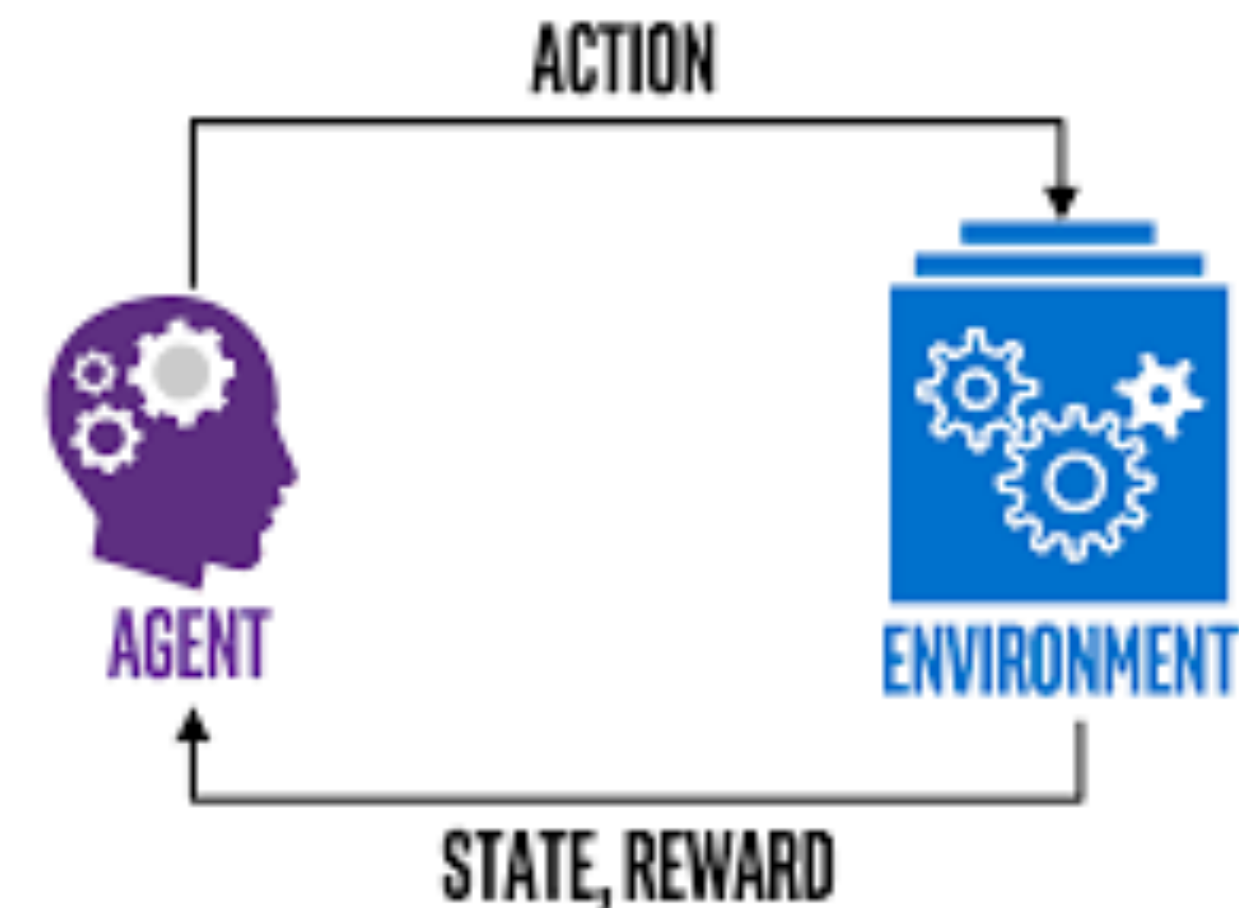
Cross-cutting ML themes

- Physics-constraints, interpretability
- Domain adaptation, fault tolerance, uncertainty quantification 
- Efficient, resource-constrained



# Reinforcement , active learning

- *RL less widely-used than (un)supervised*
  - *Surrogate modeling, digital twins important*
- Applications studied for accelerator control - beyond standard PID loops
- Similar techniques are being explored for:
  - Real-time adaptive collider triggers
  - Self-driving telescopes
  - Automated sensor/detector construction
  - Gravitational wave sensor denoising
  - ...

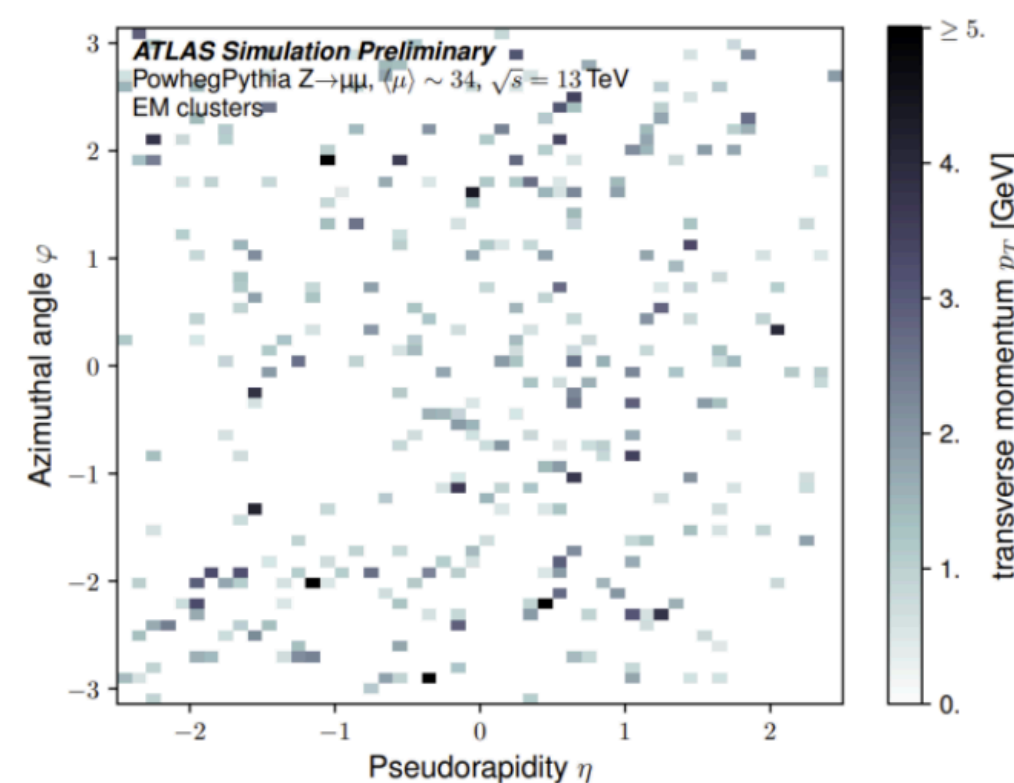




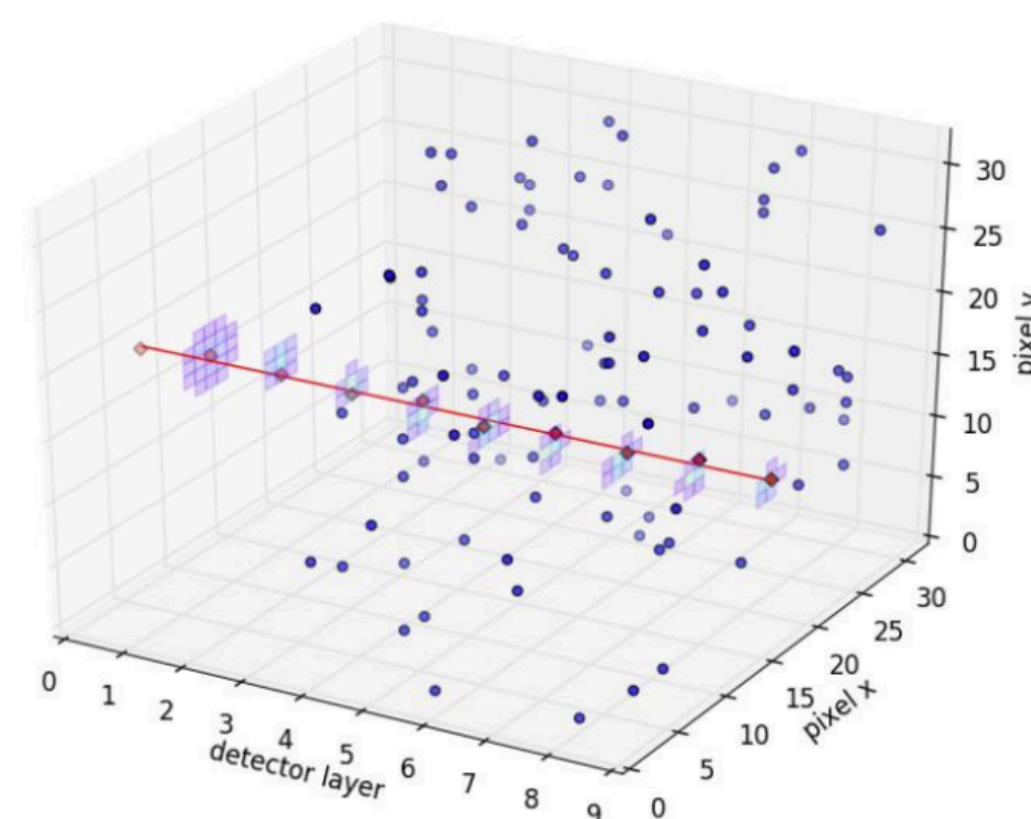
# Data representations - graphs

slides, Daniel Murnane  
ExaTrkX

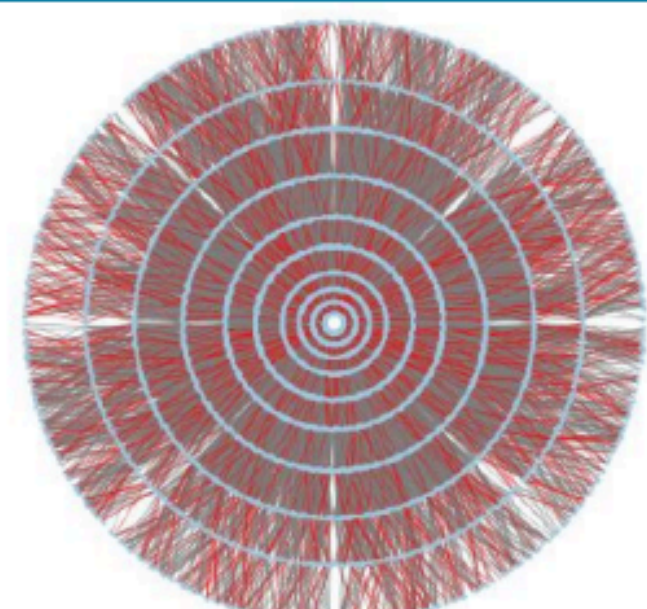
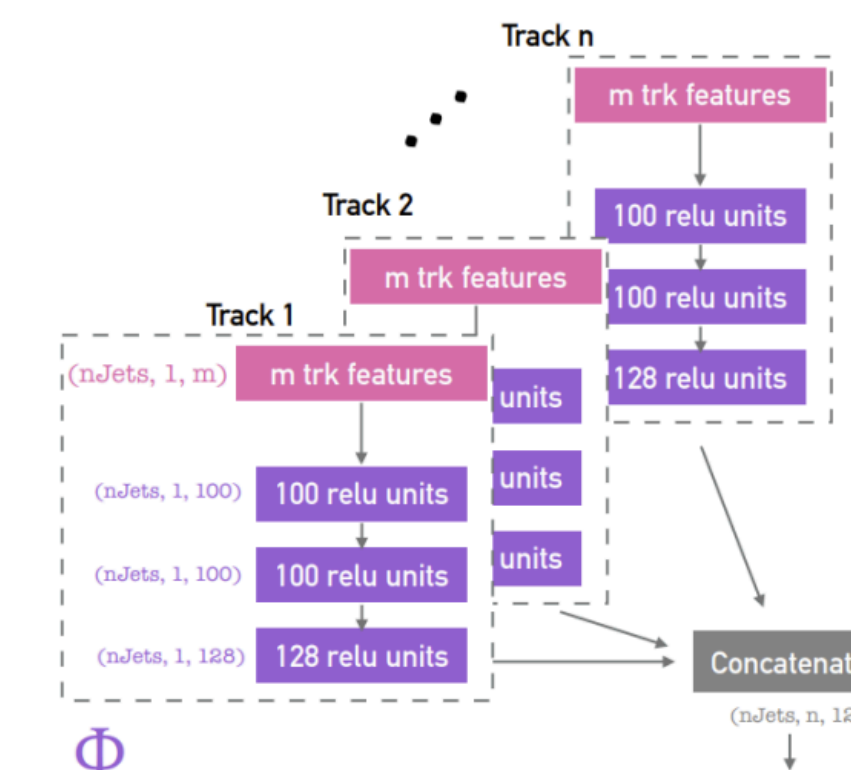
## Image?



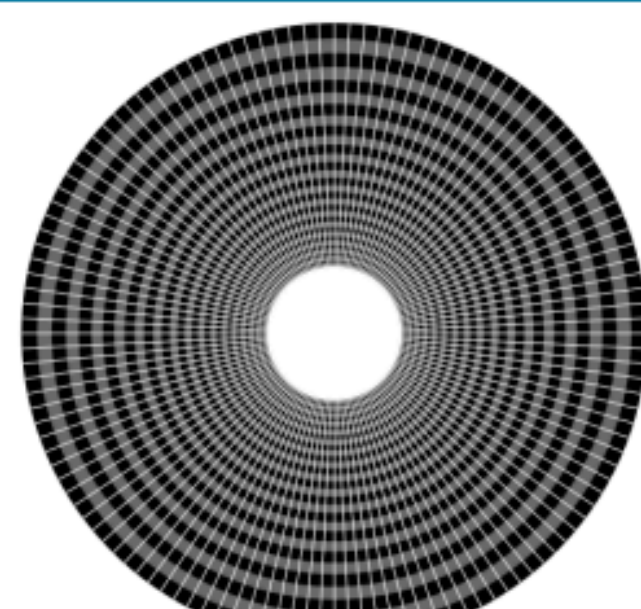
## Sequence?



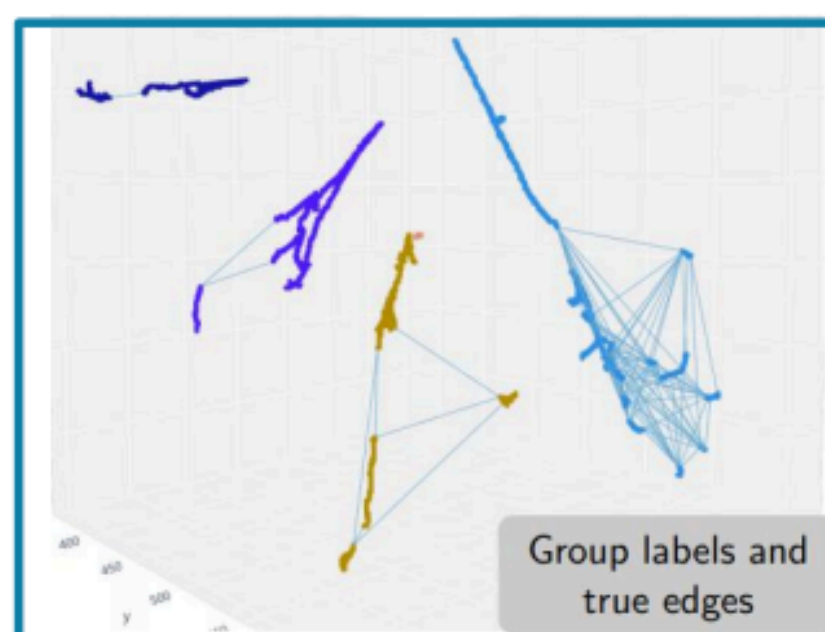
## Set/Point Cloud?



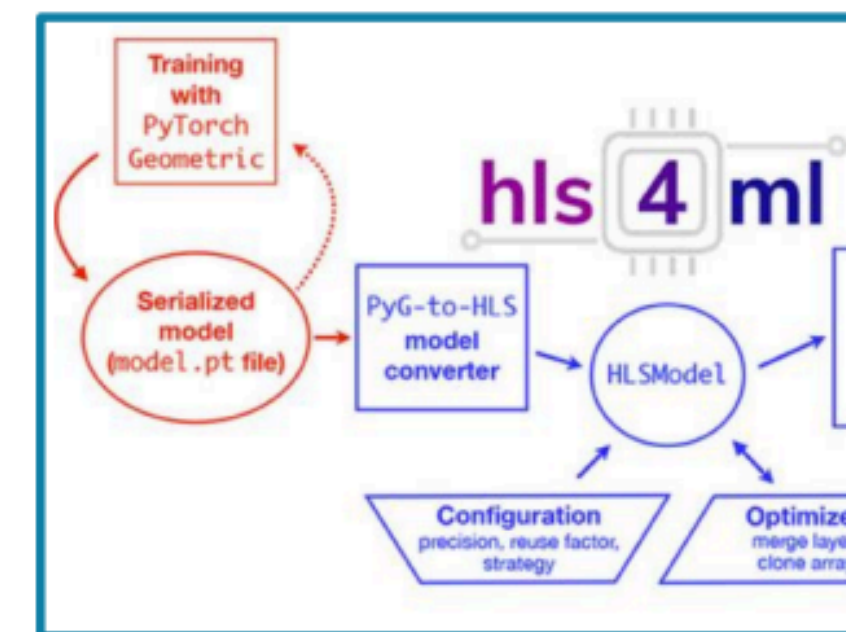
High Lumi Generic  
Tracking



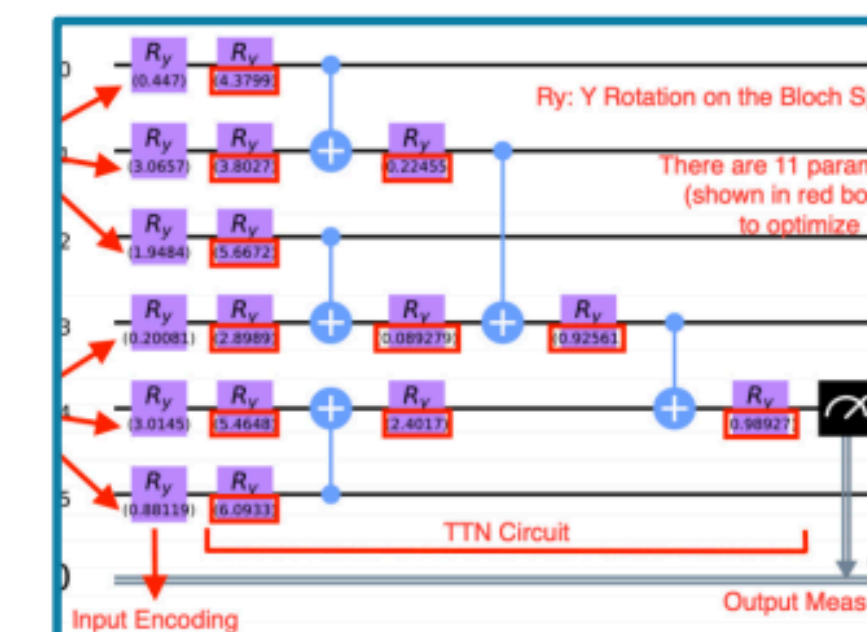
High Lumi CMS  
Calorimetry



LArTPC Particle  
Reconstruction



FPGA-based Track  
Reconstruction

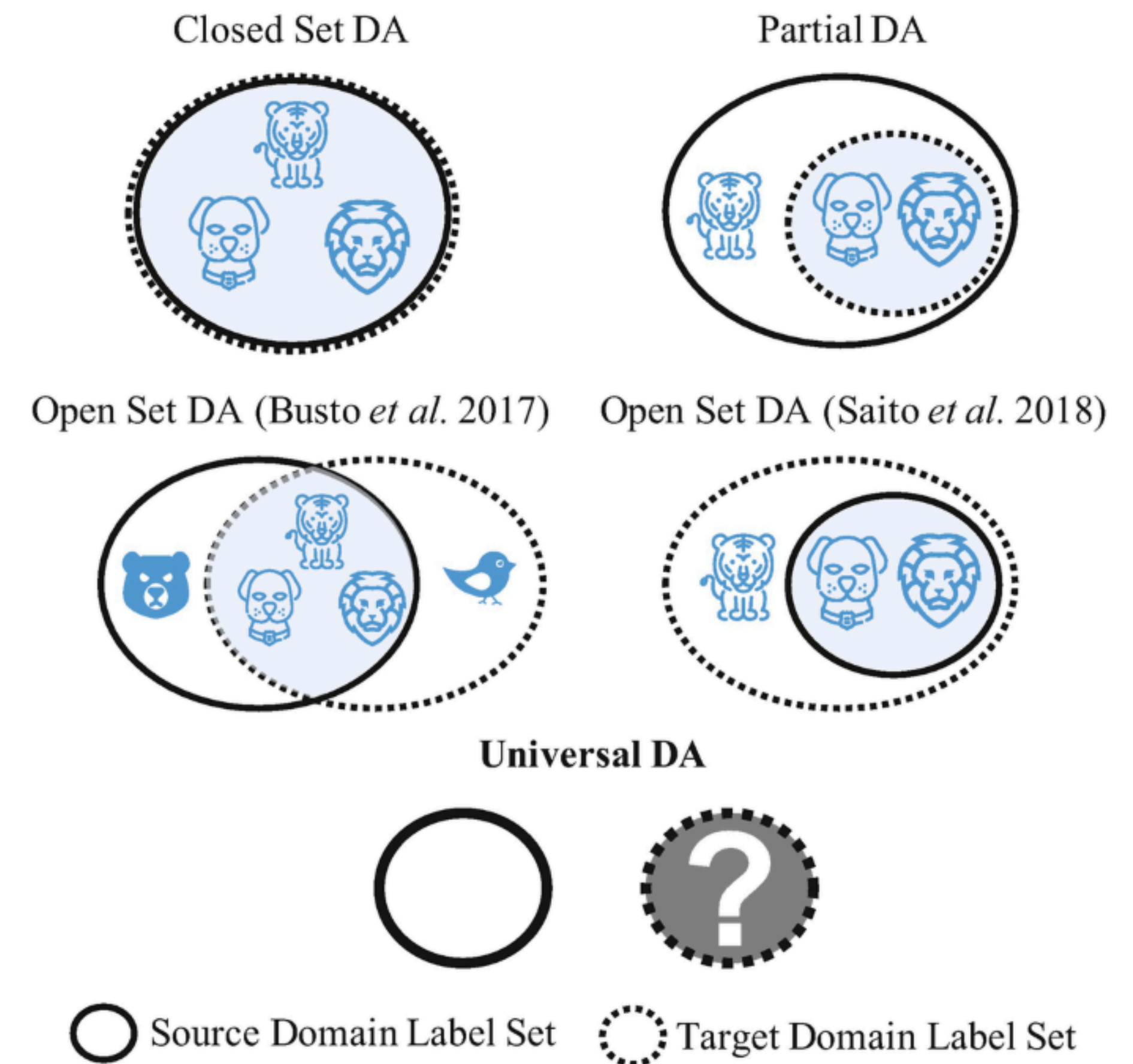


Quantum Track  
Reconstruction



# Domain adaptation, fault tolerance, etc.

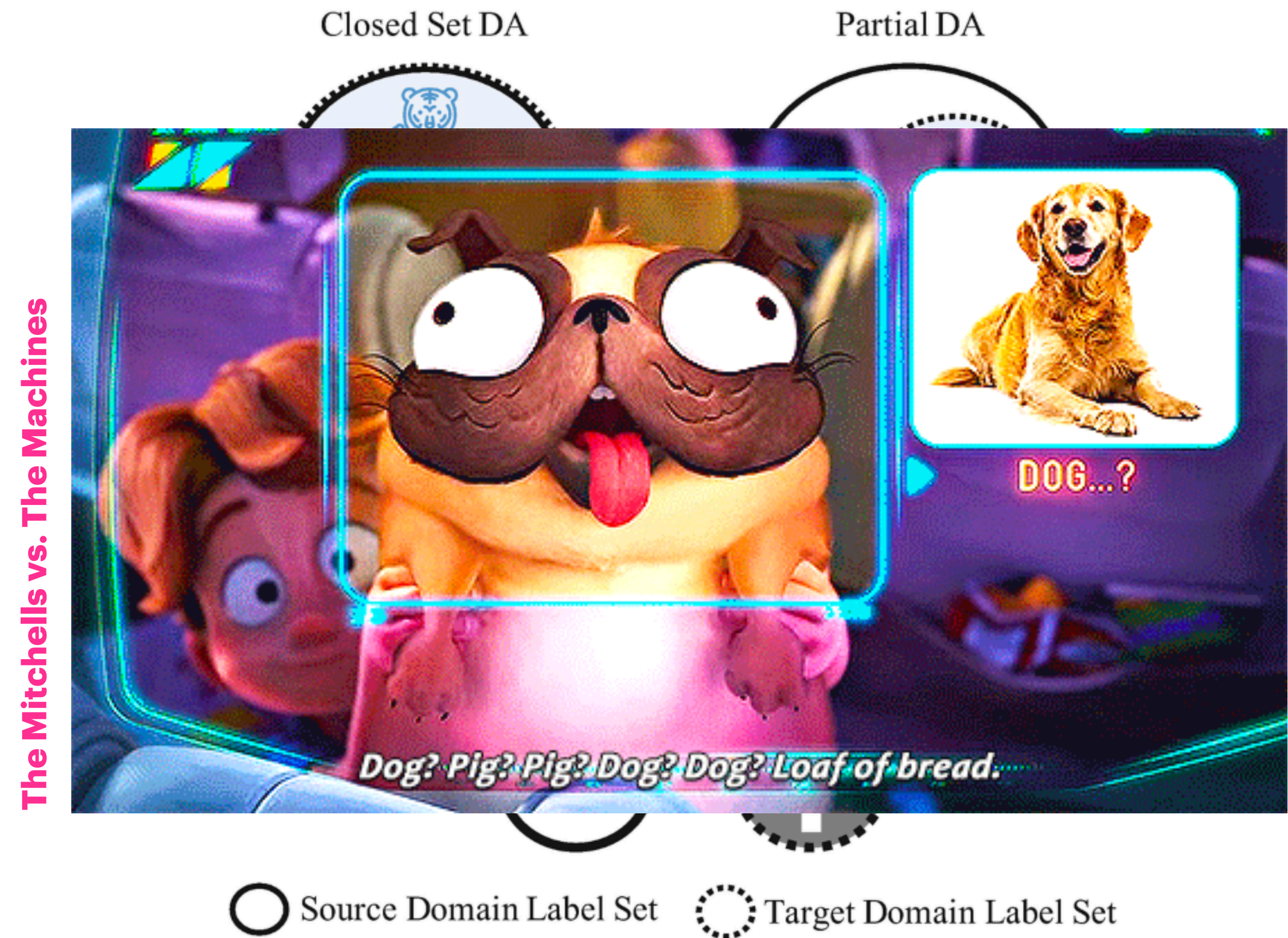
- Connected to persistent set of challenges when applying ML, e.g.
  - *How do ML model X works at a different energy, mass, region, etc.?*
- Physics-based simulation is a uniquely powerful tool, but...
  - *Do we understand data vs MC*
- *Can we learn directly from the data?*
  - semi-/self-supervised learning
- *What happens when you see something you don't understand?*





# Domain adaptation, fault tolerance, etc.

- Connected to persistent set of challenges when applying ML, e.g.
  - *How do ML model X works at a different energy, mass, region, etc.?*
  - Physics-based simulation is a uniquely powerful tool, but...
    - *Do we understand data vs MC*
    - *Can we learn directly from the data?*
    - semi-/self-supervised learning
  - *What happens when you see something you don't understand?*





# Anomaly detection

- Large topic area with broad-ranging applications
  - monitoring/validation to new physics
- E.g. LHC olympics (ML4Jets) & Dark Machines challenges
- Rich **inter-experiment effort** with **extensive theory-experiment collaborations**
- Highlights the power of **public datasets and benchmarks** as a way to catalyze progress

## The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics



Gregor Kasieczka (ed),<sup>1</sup> Benjamin Nachman (ed),<sup>2,3</sup> David Shih (ed),<sup>4</sup> Oz Amram,<sup>5</sup> Anders Andreassen,<sup>6</sup> Kees Benkendorfer,<sup>2,7</sup> Blaz Bortolato,<sup>8</sup> Gustaaf Brooijmans,<sup>9</sup> Florencia Canelli,<sup>10</sup> Jack H. Collins,<sup>11</sup> Biwei Dai,<sup>12</sup> Felipe F. De Freitas,<sup>13</sup> Barry M. Dillon,<sup>8,14</sup> Ioan-Mihail Dinu,<sup>5</sup> Zhongtian Dong,<sup>15</sup> Julien Donini,<sup>16</sup> Javier Duarte,<sup>17</sup> D. A. Faroughy,<sup>10</sup> Julia Gonski,<sup>9</sup> Philip Harris,<sup>18</sup> Alan Kahn,<sup>9</sup> Jernej F. Kamenik,<sup>8,19</sup> Charanjit K. Khosa,<sup>20,30</sup> Patrick Komiske,<sup>21</sup> Luc Le Pottier,<sup>2,22</sup> Pablo Martín-Ramiro,<sup>2,23</sup> Andrej Matevc,<sup>8,19</sup> Eric Metodiev,<sup>21</sup> Vinicius Mikuni,<sup>10</sup> Inês Ochoa,<sup>24</sup> Sang Eon Park,<sup>18</sup> Maurizio Pierini,<sup>25</sup> Dylan Rankin,<sup>18</sup> Veronica Sanz,<sup>20,26</sup> Nilai Sarda,<sup>27</sup> Uroš Seljak,<sup>2,3,12</sup> Aleks Smolkovic,<sup>8</sup> George Stein,<sup>2,12</sup> Cristina Mantilla Suarez,<sup>5</sup> Manuel Szwec,<sup>28</sup> Jesse Thaler,<sup>21</sup> Steven Tsan,<sup>17</sup> Silviu-Marian Udrescu,<sup>18</sup> Louis Vaslin,<sup>16</sup> Jean-Roch Vlimant,<sup>29</sup> Daniel Williams,<sup>9</sup> Mikaeel Yunus<sup>18</sup>



# Particle physics for machine learning

# ML out in the world

- **ML is growing rapidly everywhere**
  - Development driven by academia *and industry*, largely outside of particle physics
  - We cannot (and should not) ignore this!
- Why?
  - Bring **new expertise, knowledge, and resources** to bear on our challenges
  - Contribute to the advancement of machine learning itself and other related applications



# Interdisciplinary collaboration

- *How do we capture the interest of non-HEP collaborators?*
- Straightforward way: **the physics mission is beautiful and engaging!**
- Find **unique aspects for our science** that could push the bounds of ML research
- Because sometimes a computer vision problem is a computer vision problem whether its in industry or physics (and that's ok!!)

[nvidia link](#)



## TECHNICAL BLOG

SUBSCRIBE

TECHNICAL WALKTHROUGH

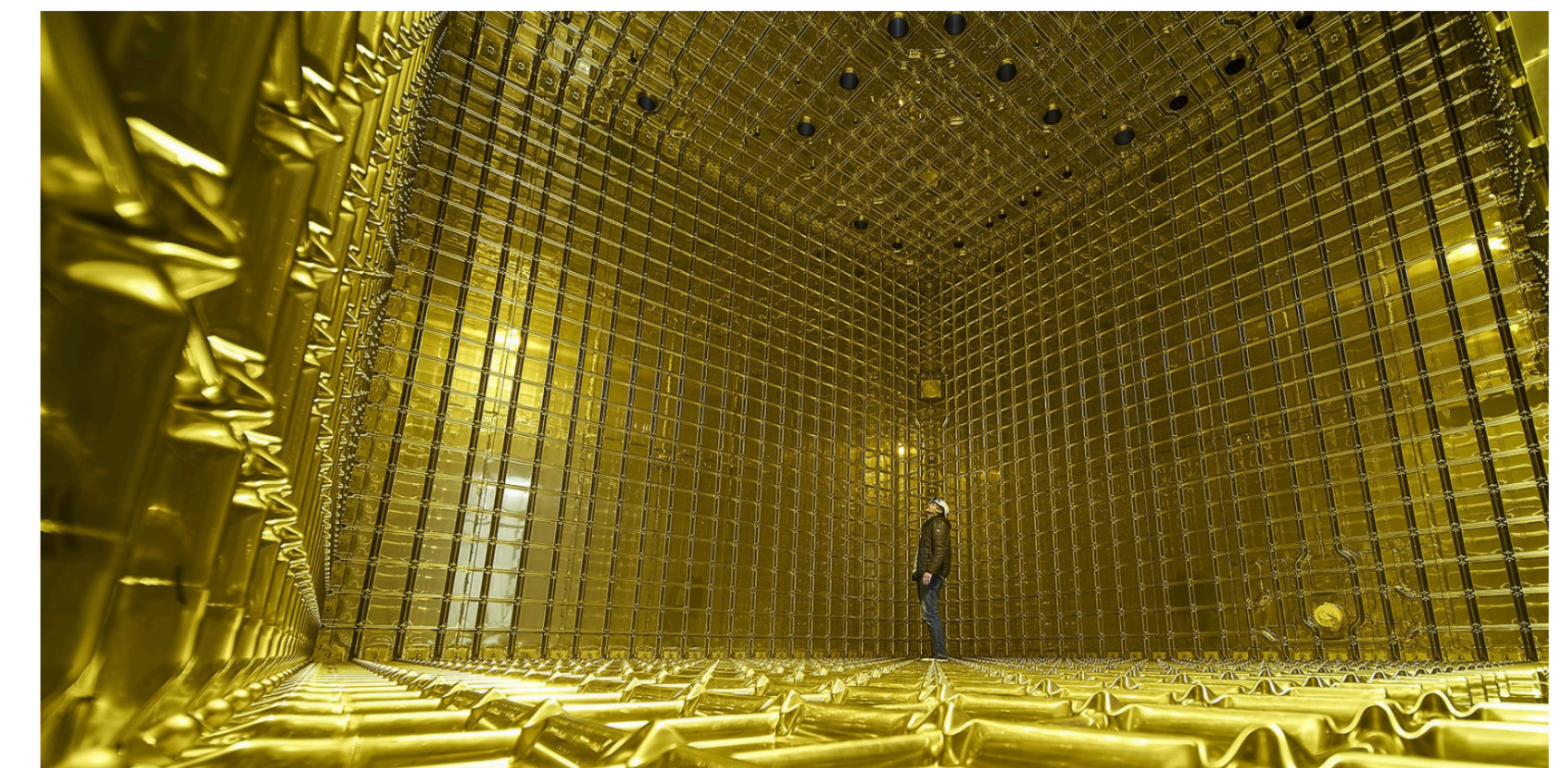
Apr 30, 2021

### Scaling Inference in High Energy Particle Physics at Fermilab Using NVIDIA Triton Inference Server

By Shankar Chandrasekaran, Lindsey Gray, Farah Hariri, Kevin Pedro, Vartika Singh, Nhan Tran, Mike Wang and Tingjun Yang

[Discuss \(0\)](#) [Share](#) [0 Like](#)

Tags: [featured](#), [Kubernetes](#), [NGC](#), [physics](#), [Triton](#)



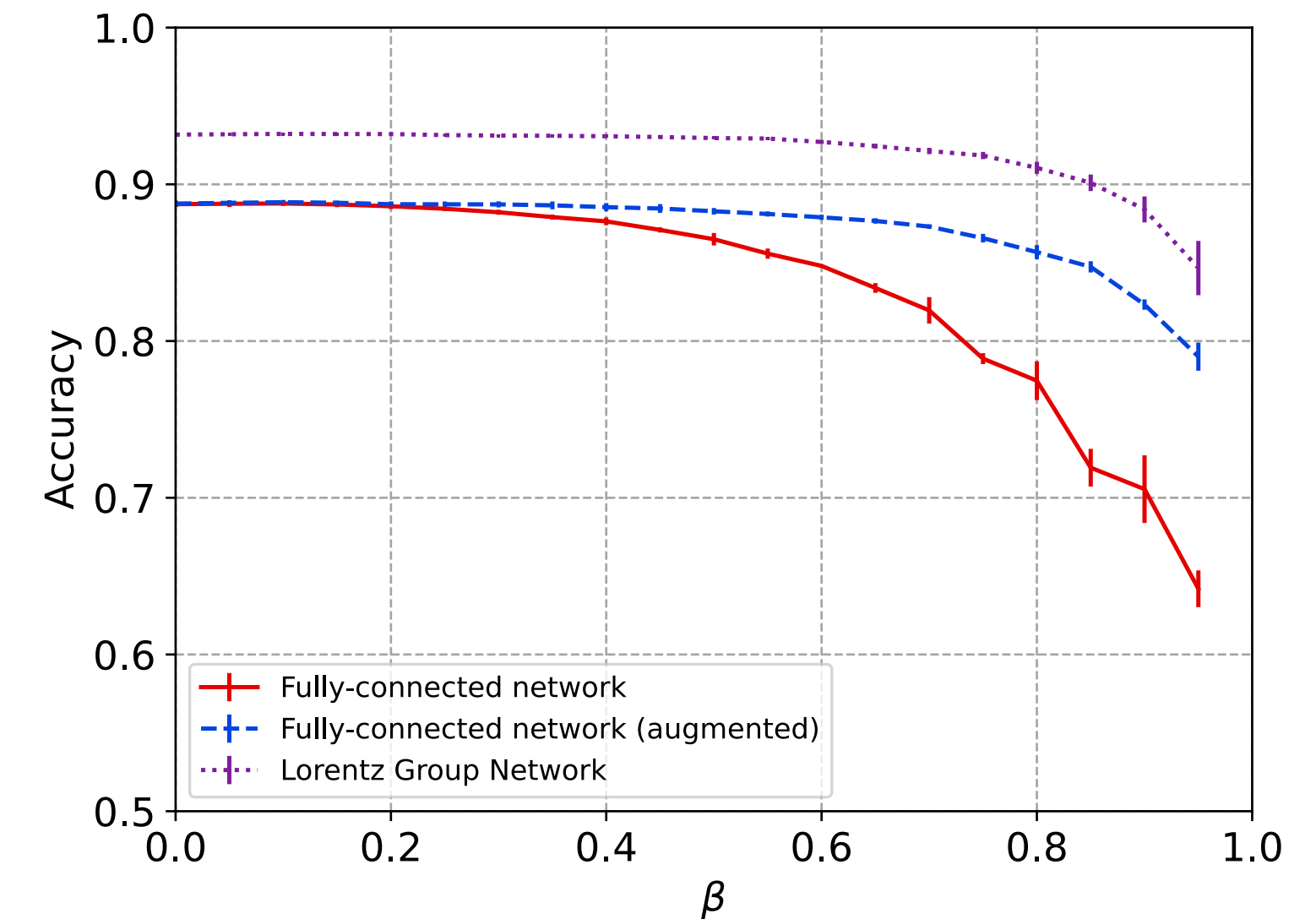


# Physics for machine learning

- Some thrusts and themes we've found that resonate *we'd love to hear others' experiences!*
- **Uncertainty quantification** - scientific rigor requires more formal and quantitative understand of uncertainties than other (industry) applications
- **Physics-informed/constrained ML** - underlying physical laws, symmetries, or constraints to improve models (next slide) or infer physical parameters (simulation-based inference) - under larger umbrella of inductive bias
- **Fast/efficient ML** - dataset sizes and data rates are in physics experiments are uniquely massive w.r.t. industry and other scientific domains

# Physics-constrained ML


- *Convolutional neural networks* were a paradigm-shifting concept for deep learning and computer vision
  - Leverages spatial symmetries
- Deep understanding of physical laws and constraints including sophisticated simulations that encode our physics knowledge
  - See for example, white paper on “Symmetry Group Equivariant Architectures for Physics” [arXiv: 2203.06153]
  - **Potential benefits:** model size/complexity, interpretability, sample efficiency, generalizability, faithfulness to physical laws



# Fast, efficient ML

doi.org:10.3389/fdata.2022.787421

<https://a3d3.ai/>




Frontiers in

Big Data

REVIEW

published: 12 April 2022

doi: 10.3389/fdata.2022.787421

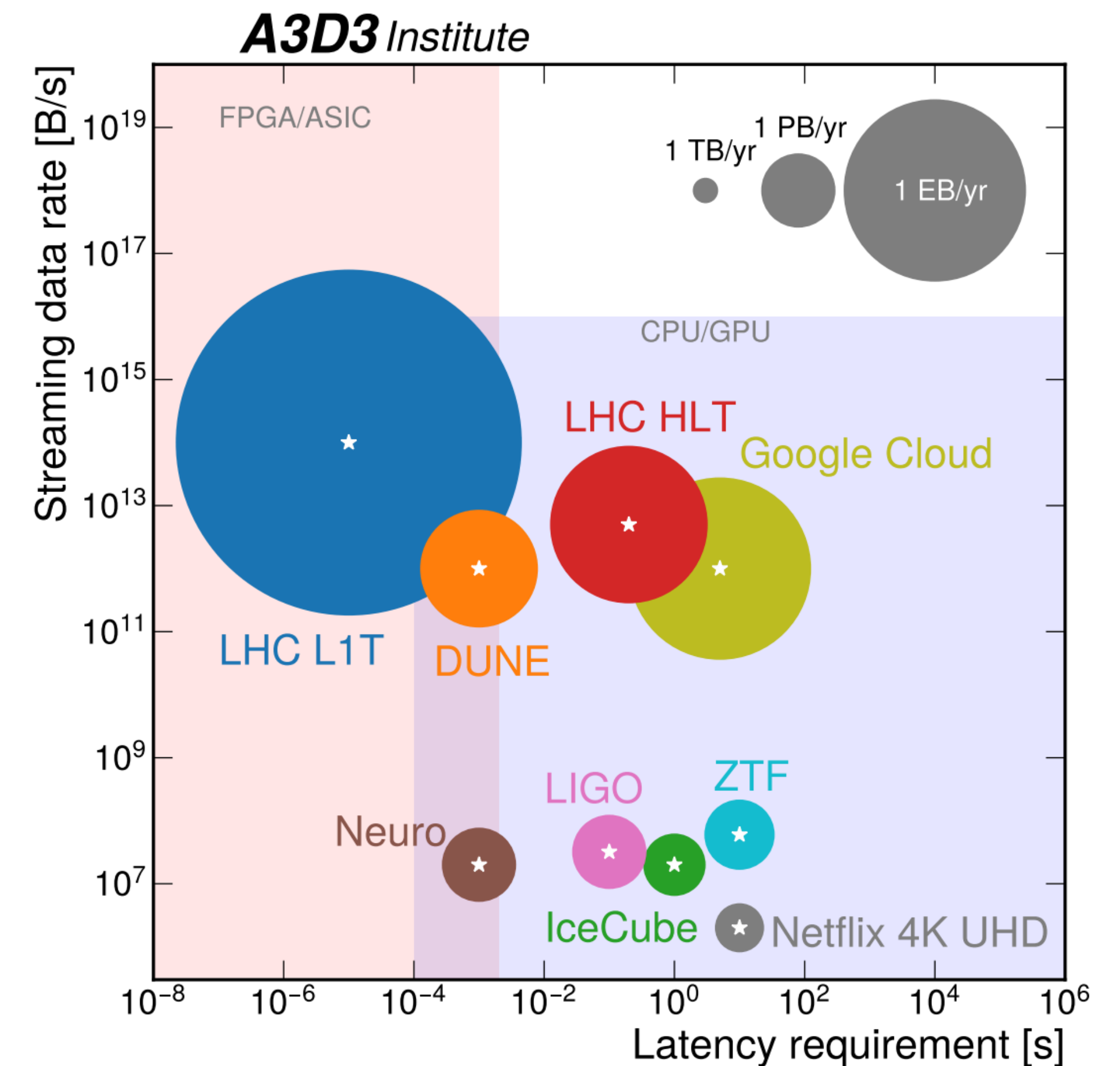


## Applications and Techniques for Fast Machine Learning in Science

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

**OPEN ACCESS**  
**Edited by:**  
Elena Cuoco,  
European Gravitational Observatory,  
Italy

More discussion on this in IF04, IF07,  
CompF03, and CompF04 sessions!



# Fast, efficient ML

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



# Fast, efficient ML

[fastmachinelarning.org](https://fastmachinelarning.org)

arXiv:2006.10159

arXiv:2102.11289

arXiv:2206.11791

arXiv:2206.07527

*Unique particle physics challenges necessitates novel solutions, techniques, and tools*

E.g. **community collaboration** with computer scientists, engineers in academia & industry (Google, AMD/Xilinx, MLCommons,...), among others on **open-source tools beyond physics**

## Ultra Low-latency, Low-area Inference Accelerators using Heterogeneous Deep Quantization with QKeras and hls4ml

Claudionor N. Coelho Jr.  
Aki Kuusela, Hao Zhuang  
Google LLC  
Mountain View, California, USA

Thea Aarrestad, Vladimir Loncar\*  
Jennifer Ngadiuba, Maurizio Pierini  
Sioni Summers  
European Organization for Nuclear Research (CERN)  
Geneva, Switzerland

## QONNX: Representing Arbitrary-Precision Quantized Neural Networks

Alessandro Pappalardo, Yaman Umuroglu, Michaela Blott  
AMD Adaptive and Embedded Computing Group (AECG) Labs  
Dublin, Ireland

Jovan Mitrevski<sup>1</sup>, Ben Hawks, Nhan Tran  
Fermi National Accelerator Laboratory  
Batavia, IL, USA

Vladimir Loncar\*  
Massachusetts Institute of Technology  
Cambridge, MA, USA

Sioni Summers  
European Organization for Nuclear Research (CERN)  
Geneva, Switzerland

Hendrik Borras  
Heidelberg University  
Heidelberg, Germany

Jules Muhizi  
Harvard University  
Cambridge, MA, USA

Matthew Trahms, Shih-Chieh Hsu, Scott Hauck  
University of Washington  
Seattle, WA, USA

Javier Duarte<sup>1</sup>  
University of California San Diego  
La Jolla, CA, USA

## Ps and Qs: Quantization-Aware Pruning for Efficient Low Latency Neural Network Inference

Benjamin Hawks<sup>1</sup>, Javier Duarte<sup>2</sup>, Nicholas J. Fraser<sup>3</sup>, Alessandro Pappalardo<sup>3</sup>, Nhan Tran<sup>1,4\*</sup> and Yaman Umuroglu<sup>3</sup>

<sup>1</sup>Fermi National Accelerator Laboratory, Batavia, IL, United States, <sup>2</sup>University of California San Diego, La Jolla, CA, United States, <sup>3</sup>Xilinx Research, Dublin, Ireland, <sup>4</sup>Northwestern University, Evanston, IL, United States

## OPEN-SOURCE FPGA-ML CODESIGN FOR THE MLPERF™ TINY BENCHMARK

Hendrik Borras<sup>1</sup> Giuseppe Di Guglielmo<sup>2</sup> Javier Duarte<sup>3</sup> Nicolò Ghielmetti<sup>4</sup> Ben Hawks<sup>5</sup> Scott Hauck<sup>6</sup>  
Shih-Chieh Hsu<sup>6</sup> Ryan Kastner<sup>3</sup> Jason Liang<sup>3</sup> Andres Meza<sup>3</sup> Jules Muhizi<sup>5,7</sup> Tai Nguyen<sup>3</sup> Rushil Roy<sup>3</sup>  
Nhan Tran<sup>5</sup> Yaman Umuroglu<sup>8</sup> Olivia Weng<sup>3</sup> Aidan Yokuda<sup>6</sup> Michaela Blott<sup>8</sup>

ML  
Commons

## MLCommons launches machine learning benchmark for devices like smartwatches and voice assistants

by Ben Wodecki 6/16/2021



With experts from Qualcomm, Fermilab, and Google aiding in its development

MLCommons, the open engineering consortium behind the MLPerf benchmark test, has launched a new measurement suite aimed at 'tiny' devices like smartwatches and voice assistants.

MLPerf Tiny Inference is designed to compare performance of embedded devices and models with a footprint of 100kB or less, by measuring



[link](#)

# Outlook

- **Machine learning runs as common thread through nearly everything we do**
- A rising tide; we are not alone
  - Naturally traverses our traditional project and frontier boundaries
  - Engaging the broader ML community can be **challenging but high impact**
- Thus far, have only scratched the surface but potential is high
  - Many collaborations started from grassroots efforts, others supported from project funding
  - **Support needed to build more connections and collaborations at different scales**