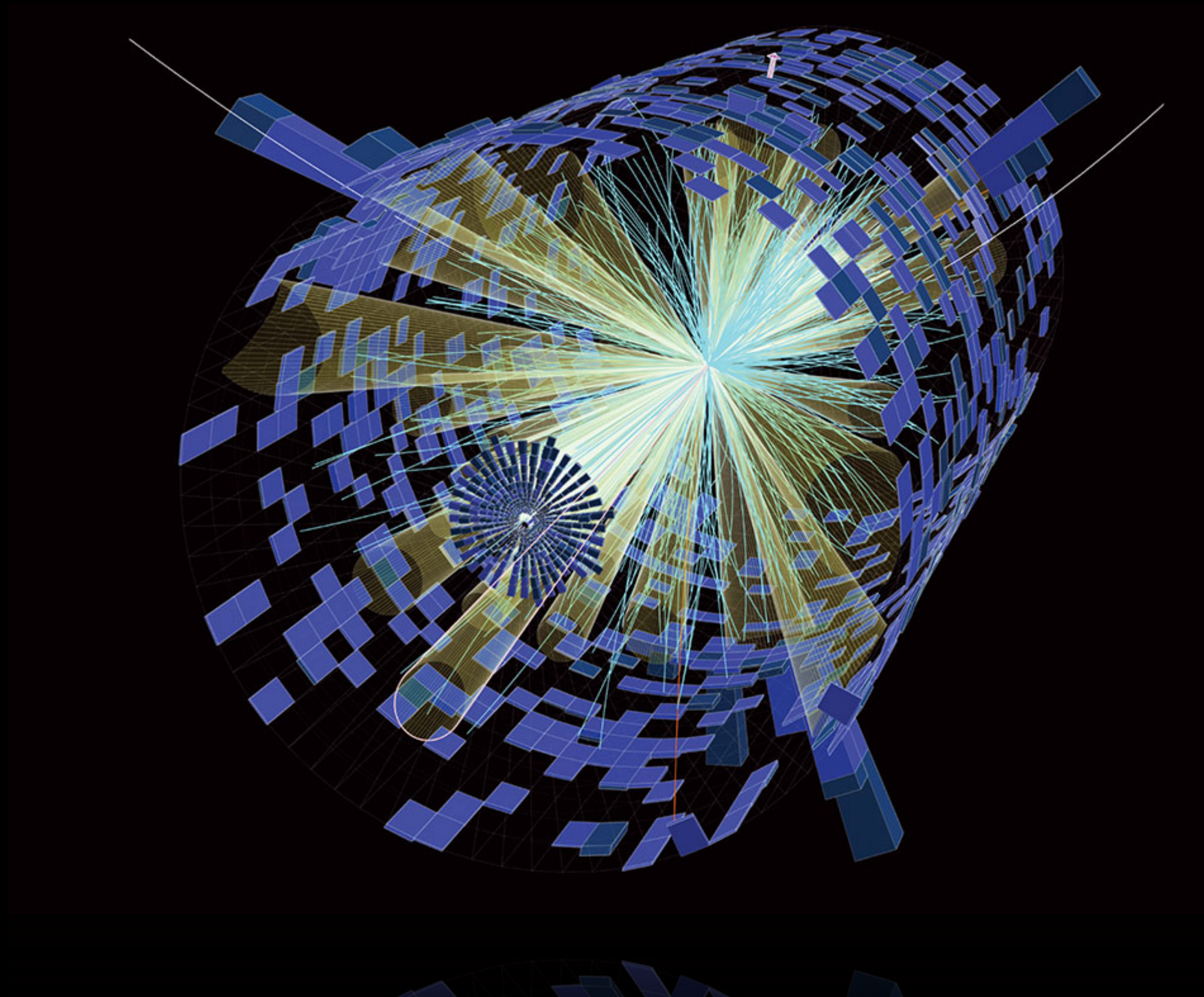
Accelerating discovery in particle physics with AI

Jennifer Ngadiuba (Fermilab)

QTML conference
CERN
November 19–24, 2023

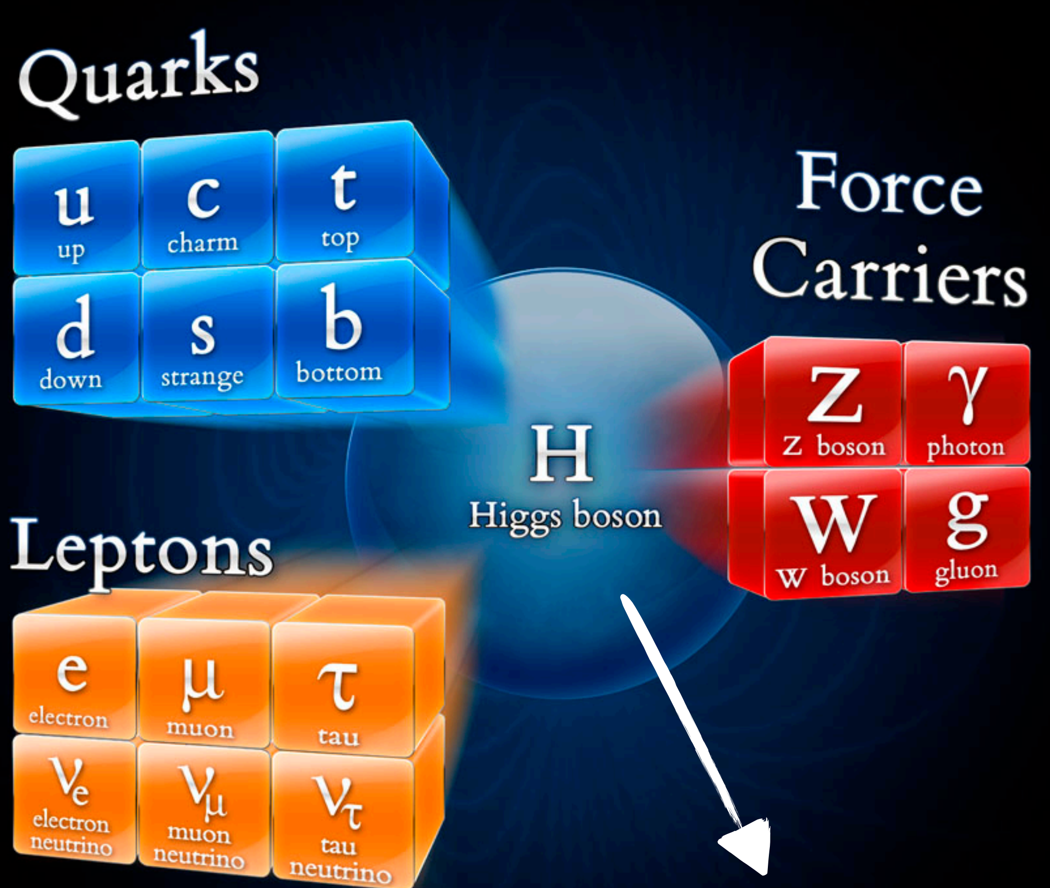


Fermilab

What we know

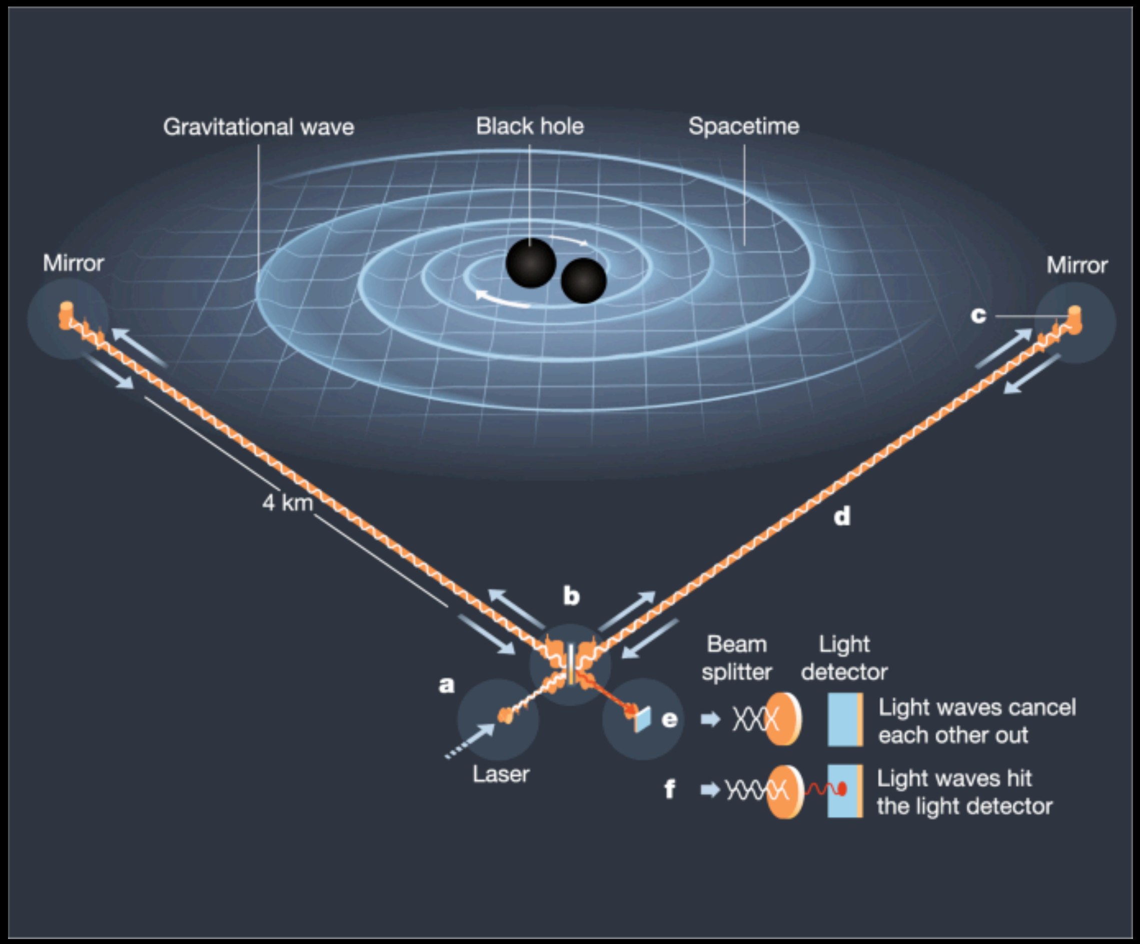
The beginning of this century marked a big expansion of our knowledge of the Universe, from the very small to the very large scale

The Standard Model



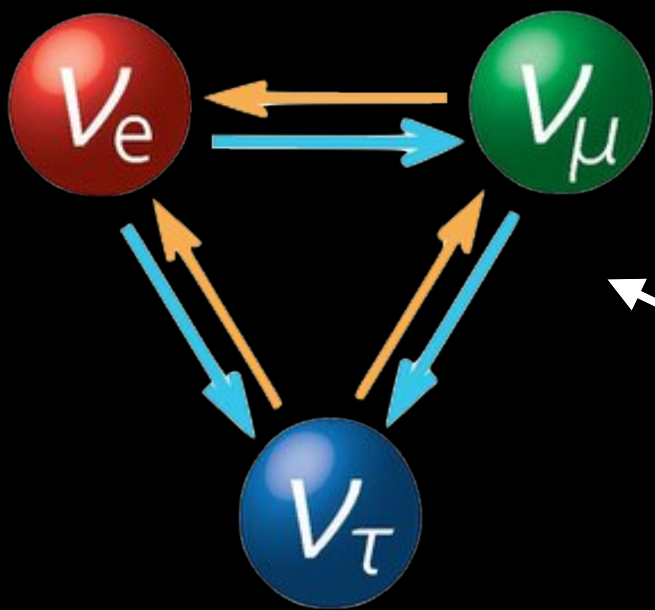
The last missing piece, discovered in 2012 at the LHC!

General relativity

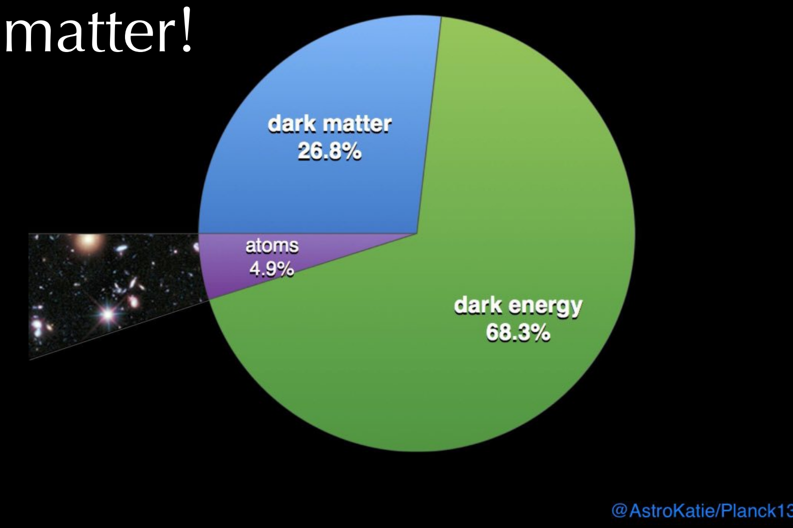


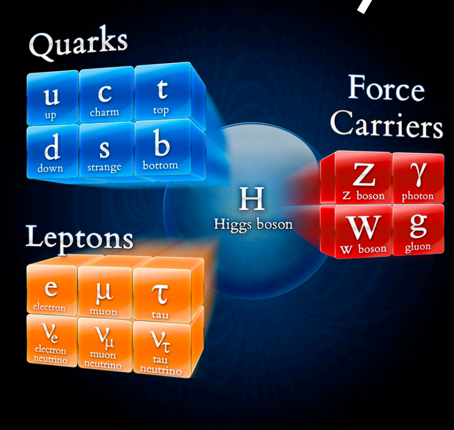
What we don't know

Neutrino masses

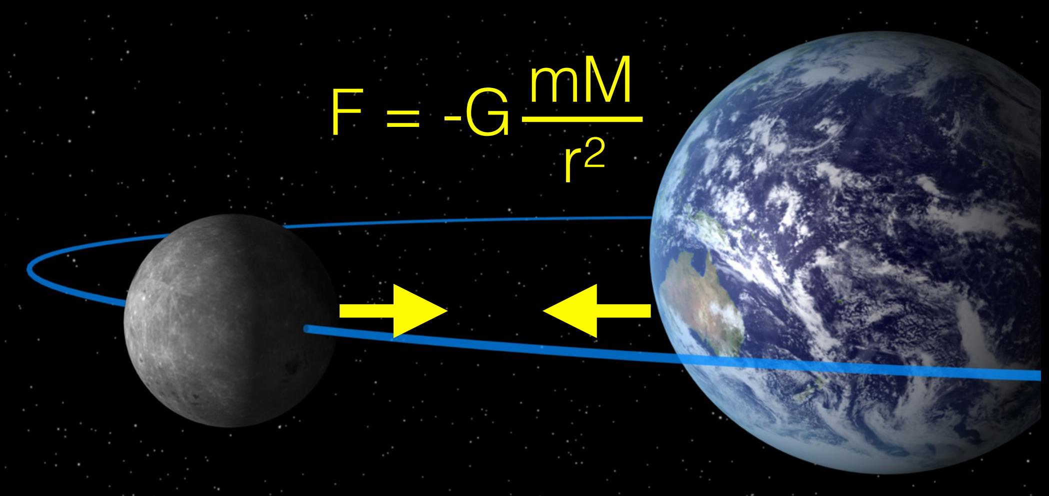


Explains only 5% of the universe.
No SM candidate
for dark matter!

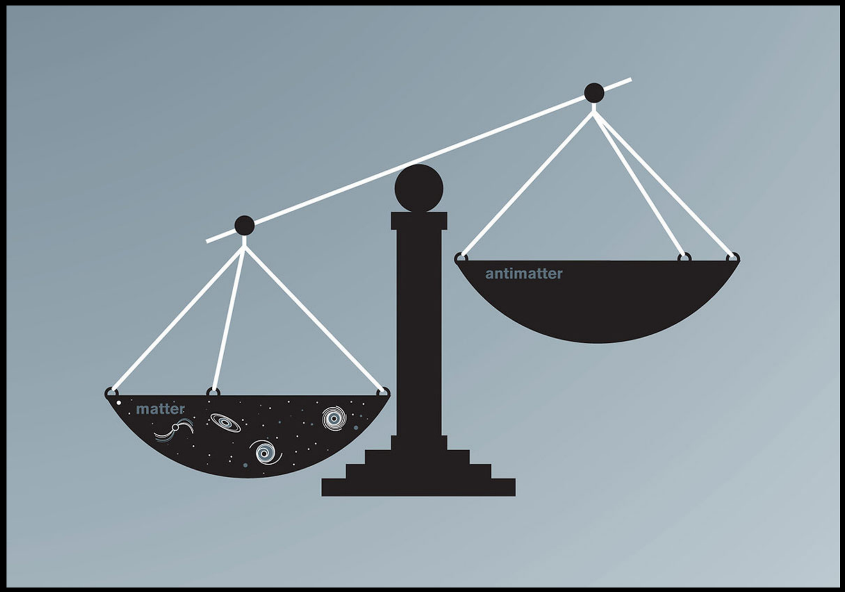




Gravity not included
in the theory

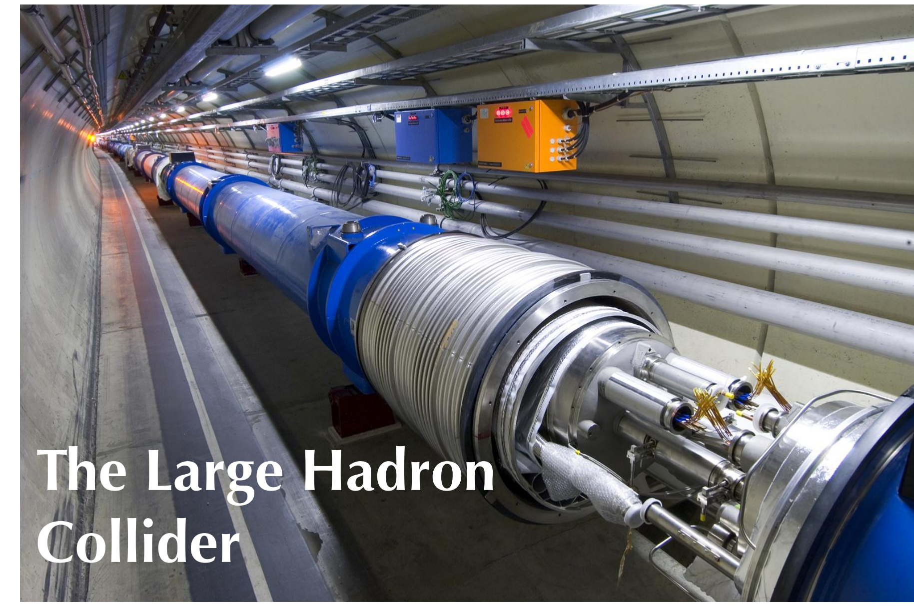


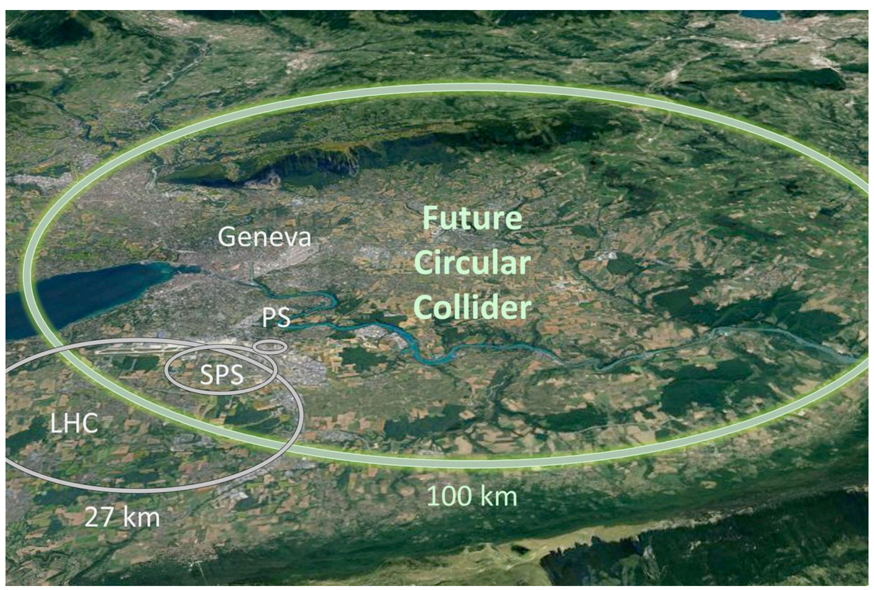
Baryon-antibaryon
asymmetry?



Big Science in 21st century

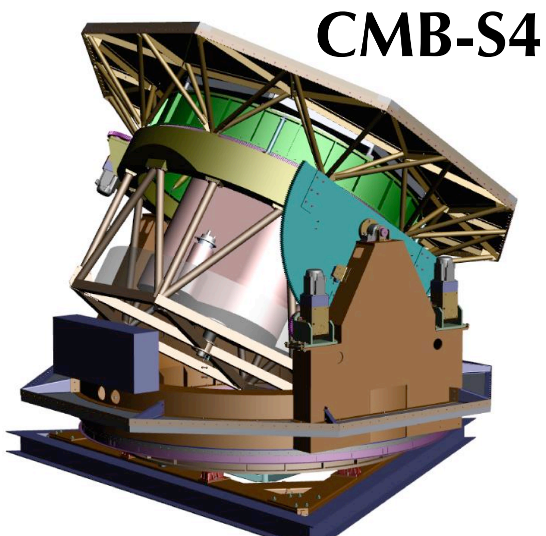
Probing the **fundamental structure of nature** requires complex experimental devices, large infrastructures and big collaborations.



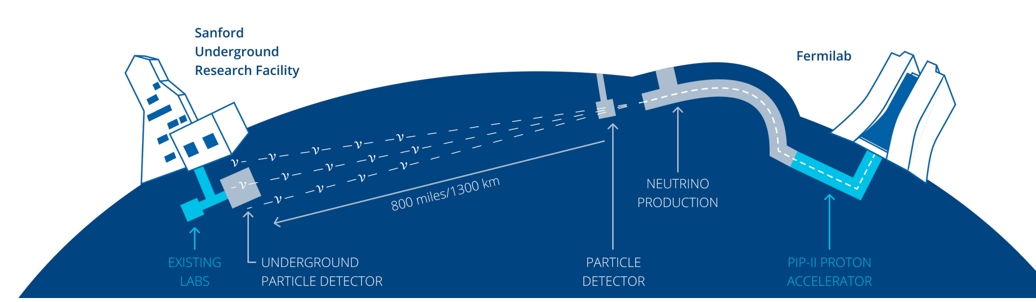


International MUON Collider Collaboration

CMB-S4



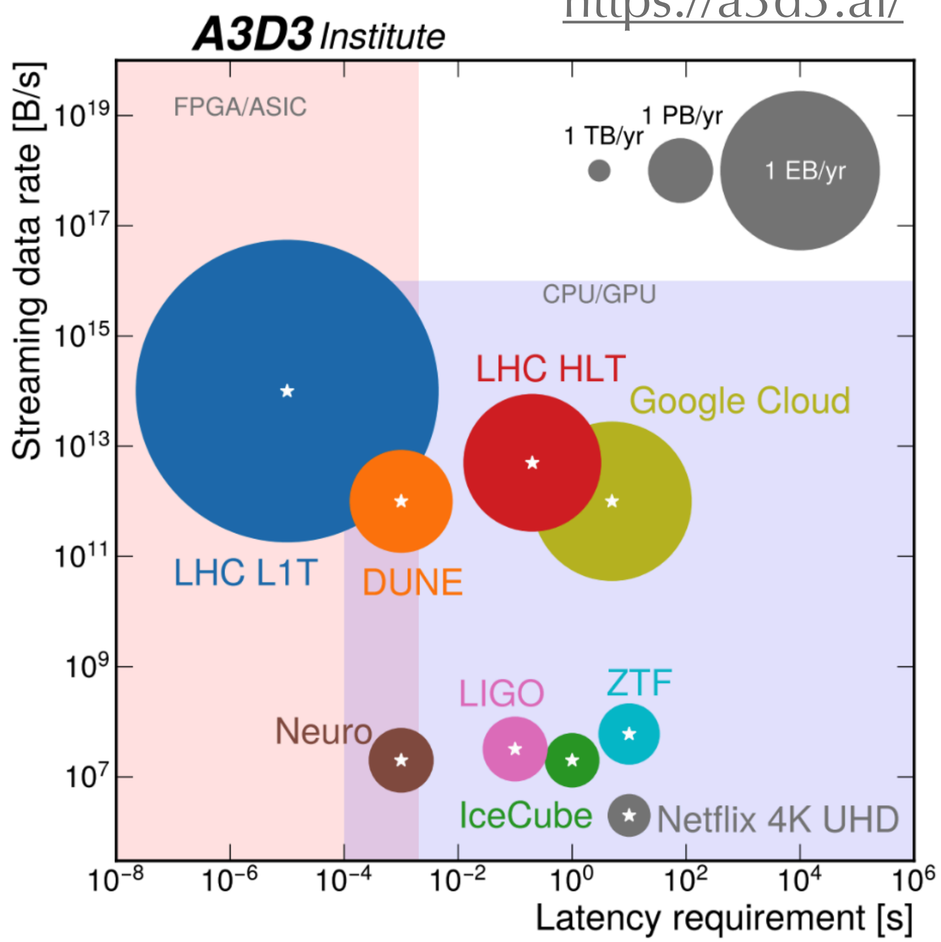
The DUNE neutrino experiment

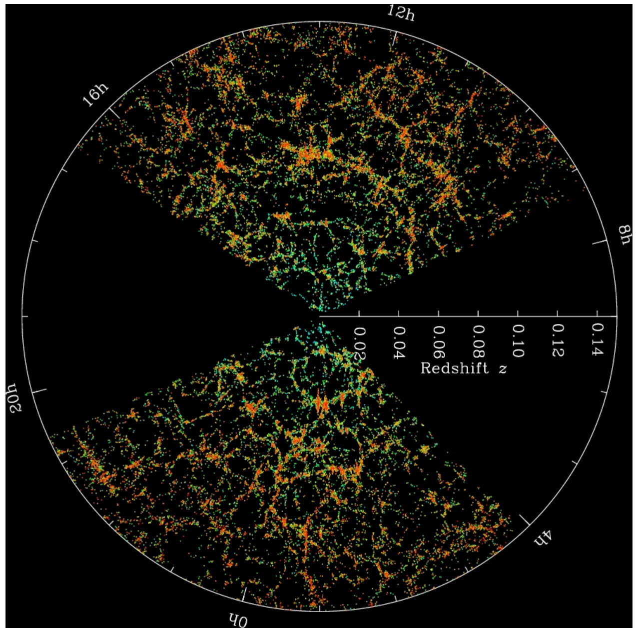


Big Science = Big Data

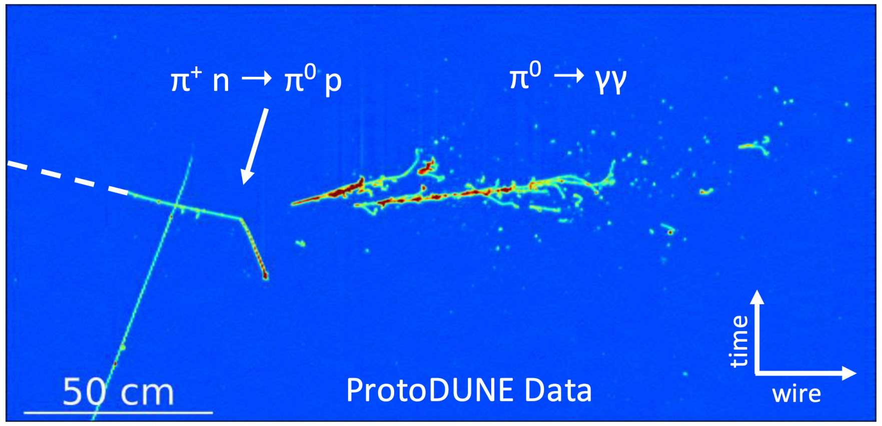
- Increasingly complex data both in **volume and dimensionality**
- Increasing need for **efficient and accurate data processing** for high-throughput applications
- Challenge in **simulating expectations** for what experiments may observe
- But also need for innovative **data & discovery driven** physics analyses approaches

https://a3d3.ai/

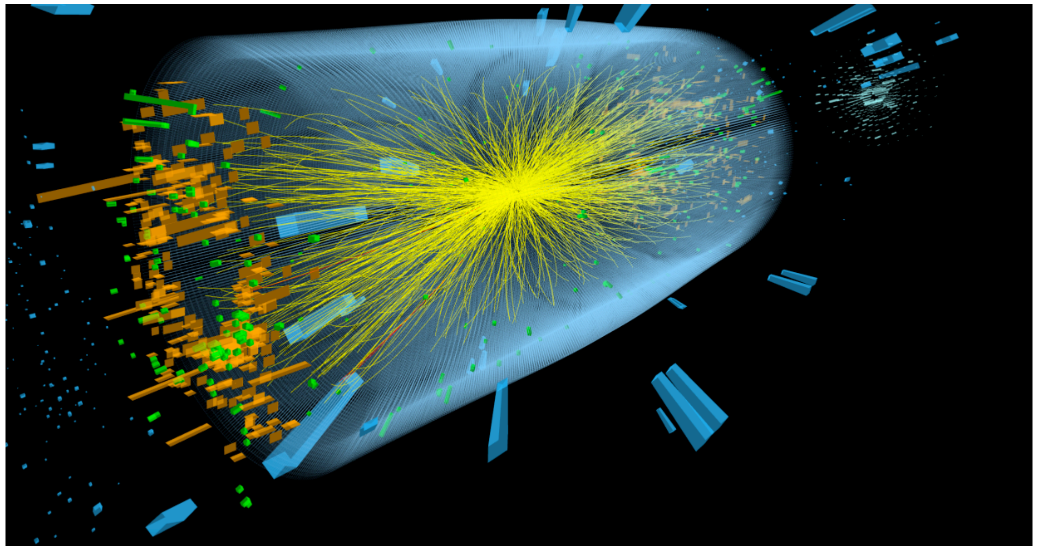




Sloan Digital Sky Survey



Interactions in LArTPC

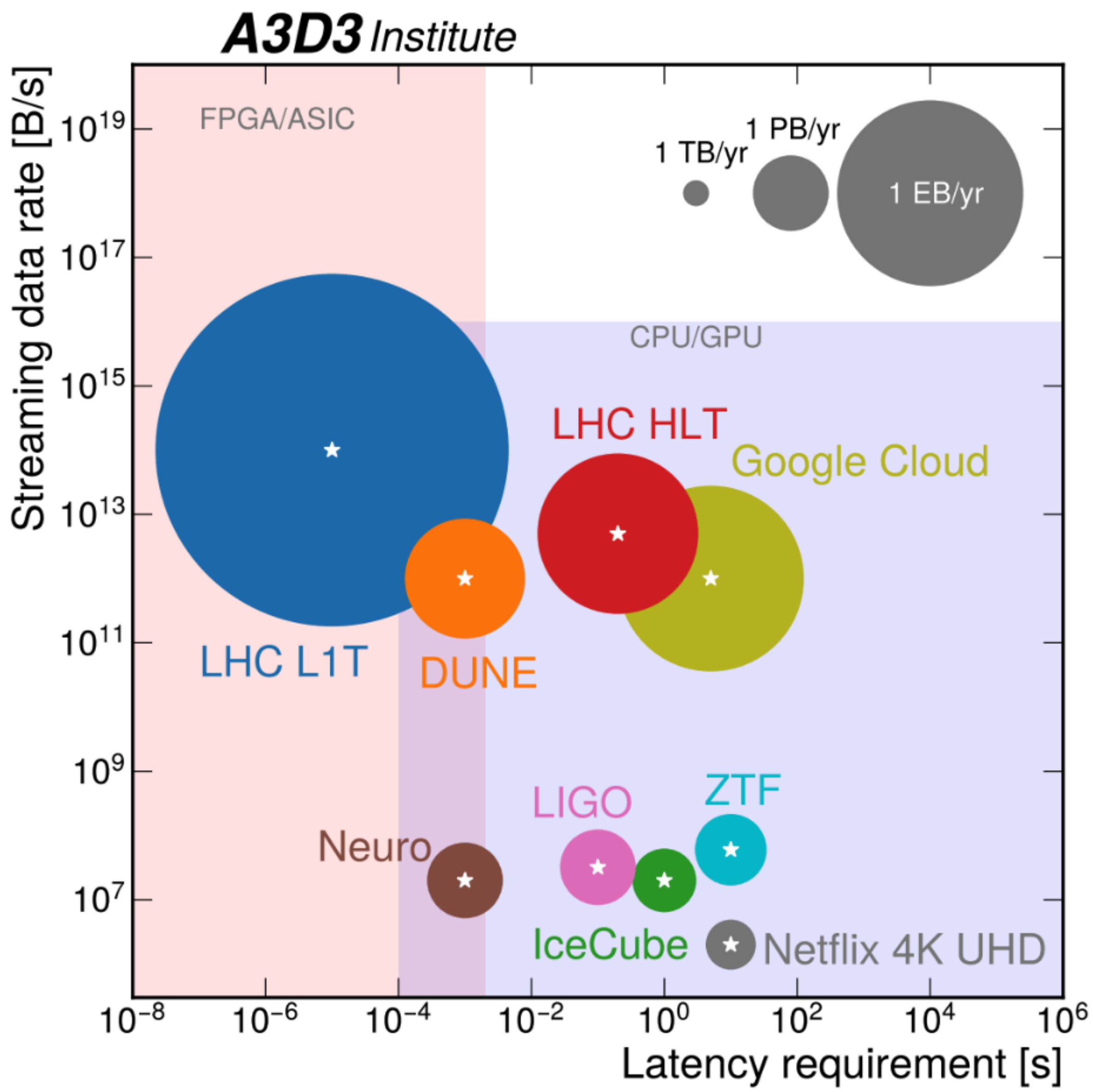


A LHC collision

This talk

- In this era of science **Artificial Intelligence can greatly accelerate time to discovery** as well-suited for **efficient analysis of large amounts of highly-dimensional data to find subtle patterns**

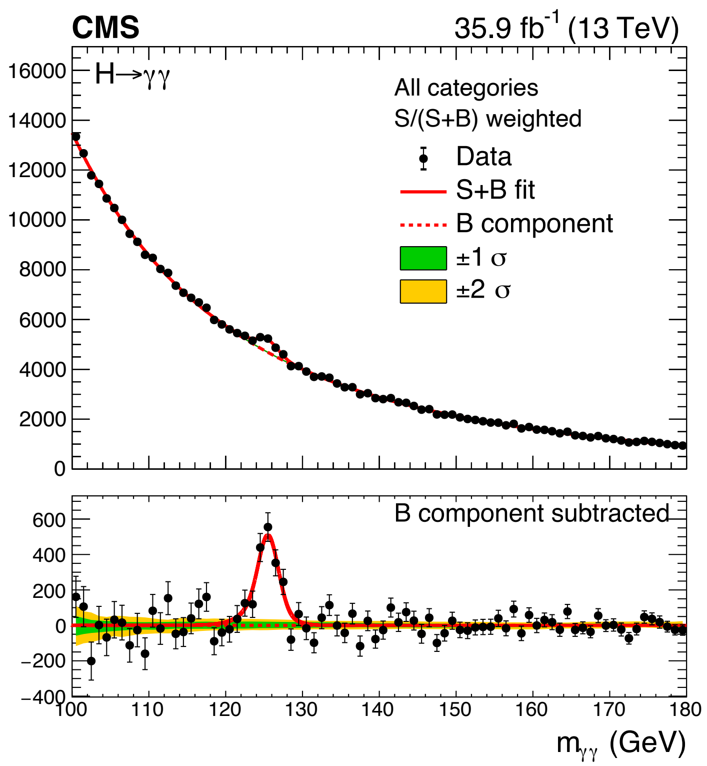
- With such capability it will allow us
  - enhance control and operations of detectors and accelerators
  - automate online and offline experimental workflows
  - save and maximize potentially lost data
  - accelerate detector R&D
  - and therefore, test hypotheses significantly faster



Machine Learning in HEP

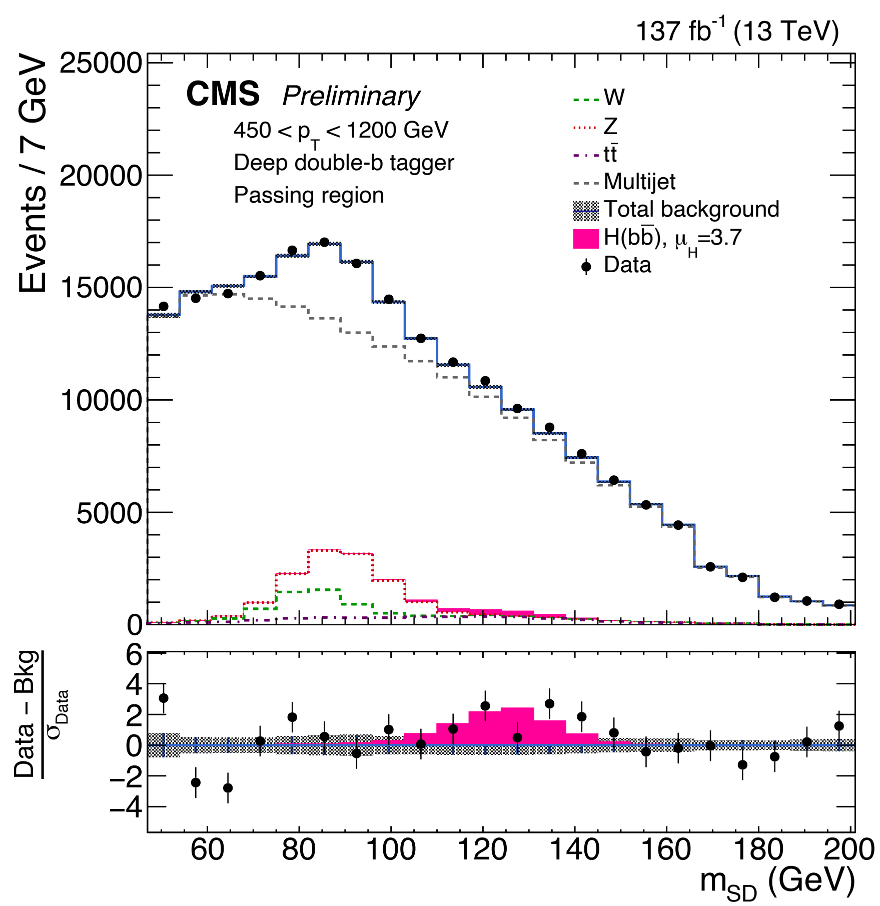
- **ML is used in particle physics since the ‘80s**
  Shallow networks back then, mostly BDTs since ~ 2004 (e.g., Higgs boson discovery)

- **Over the last decade a rapid progress has led to a revolution in this area**
  Take advantage of industry breakthrough in deep learning and computing hardware



Higgs → photons

Phys. Lett. B 805 (2020) 135425

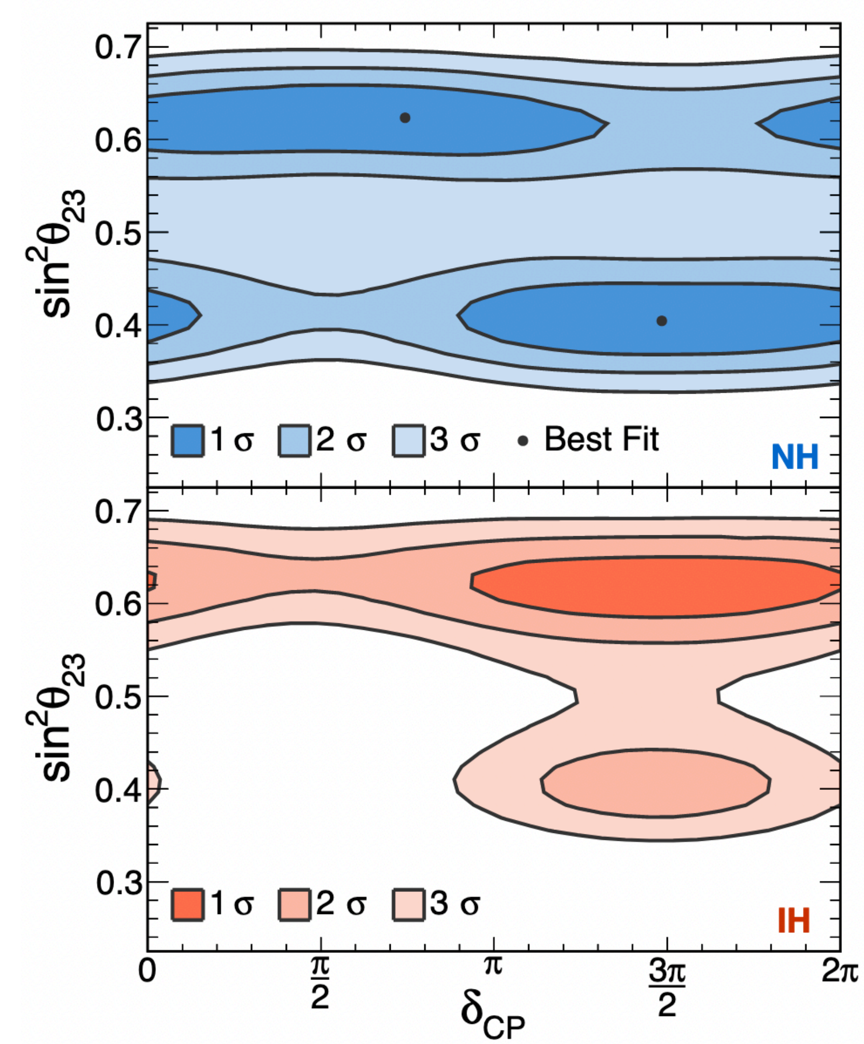


Higgs → bottom quarks

JHEP 12 (2020) 085

Measurement of neutrino oscillation parameters @ NovA

Phys. Rev. Lett. 118, 231801 (2017)



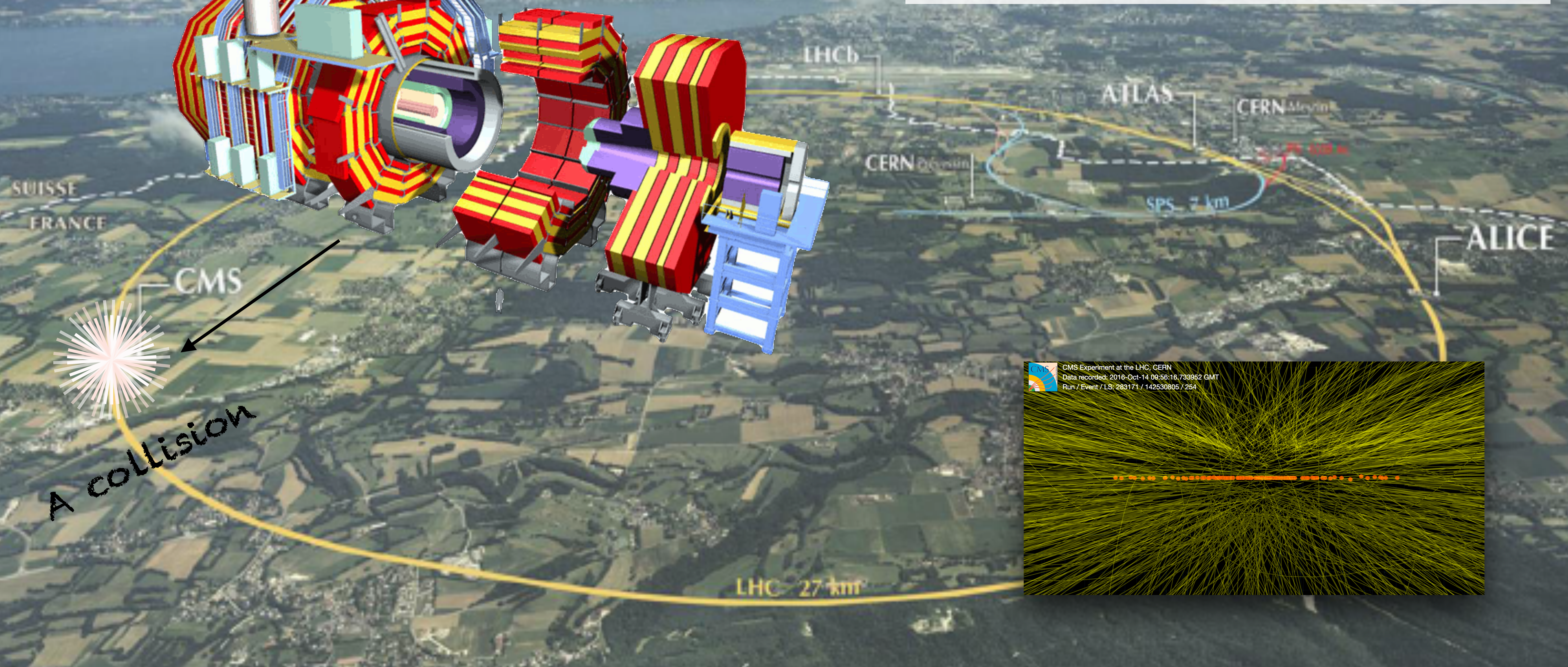
What changed?

- Consider a typical data reduction workflow in place to deal with a high volume of data
- Let's take LHC as an example…

Big Data @ the Energy Frontier

The Large Hadron Collider (LHC)

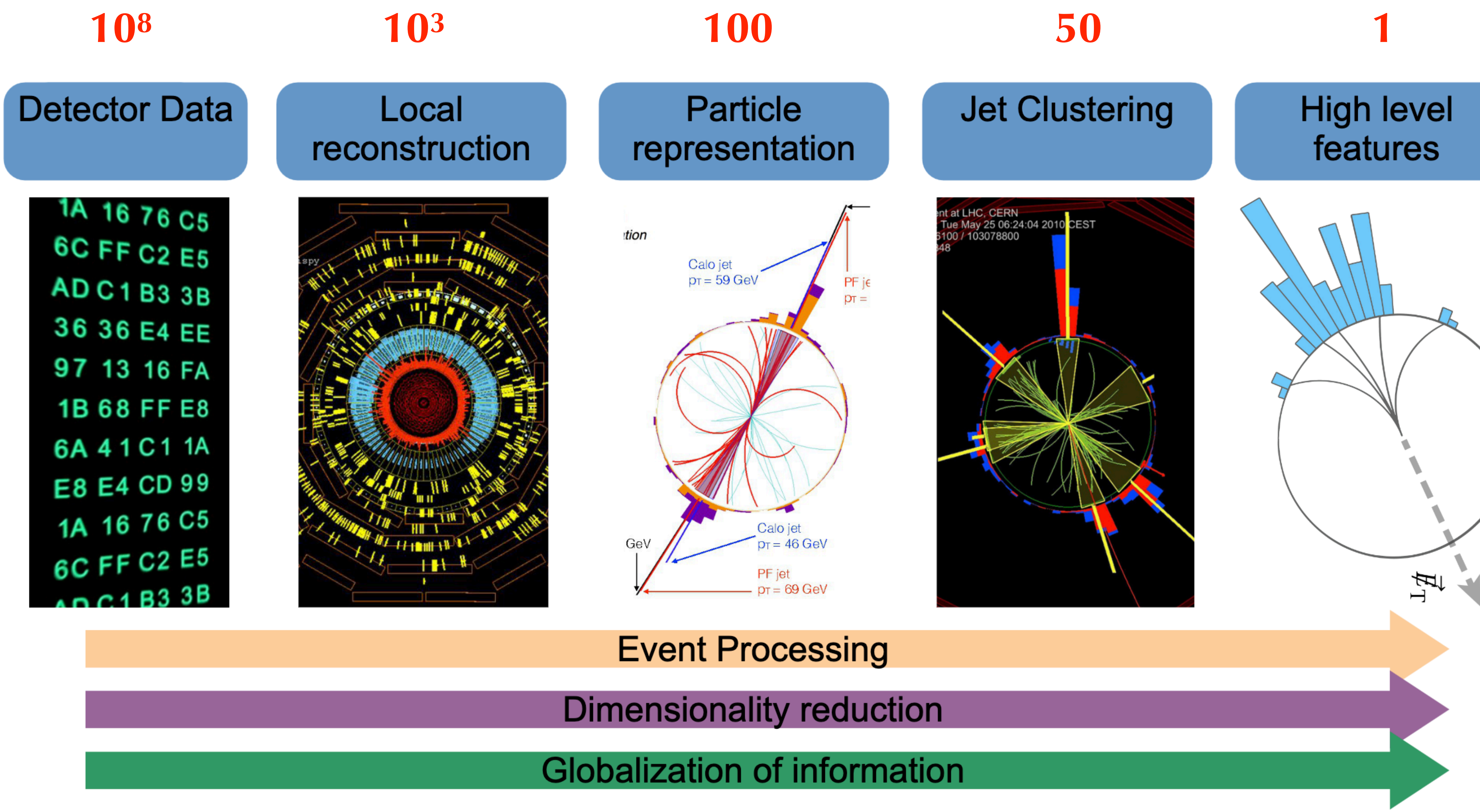
Collision frequency: 40 MHz
Particles per collision: O(10³)
Detector resolution: O(10⁸) channels



Extreme data rates of ~PB/s!

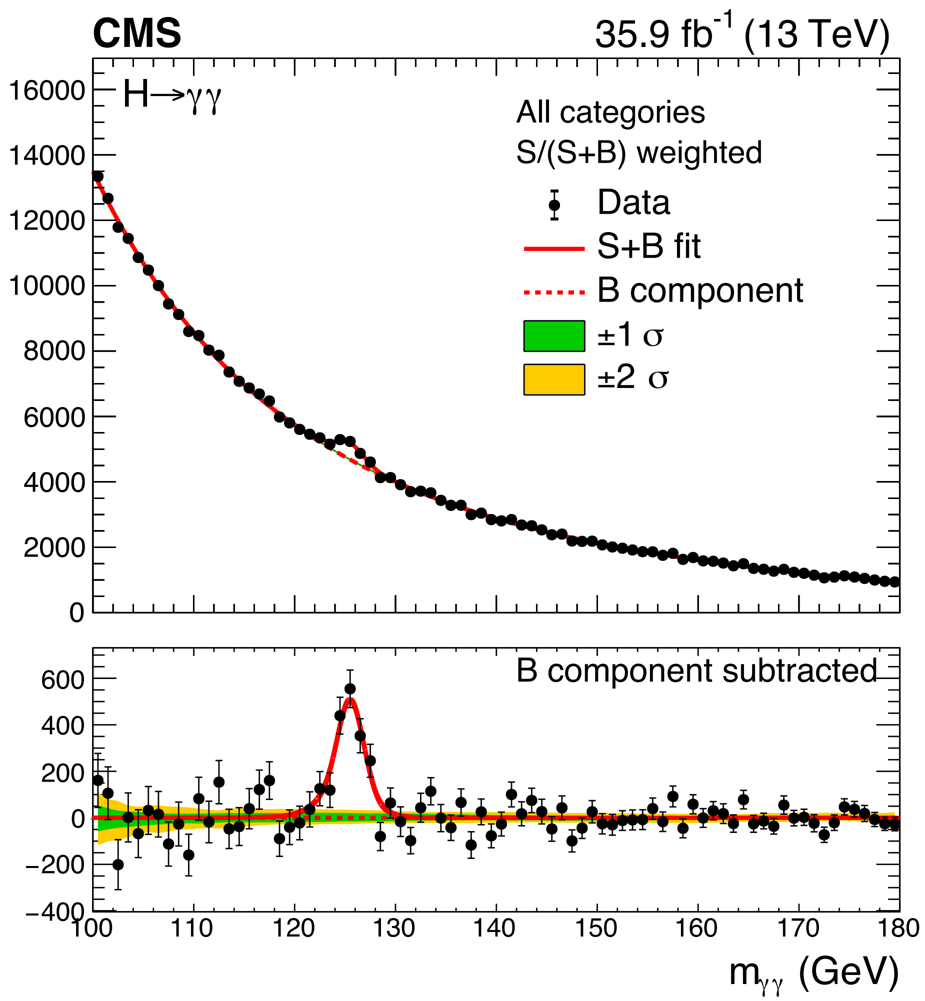
What changed?

- Consider a typical data reduction workflow in place to deal with a high volume of data
- Let's take LHC as an example…

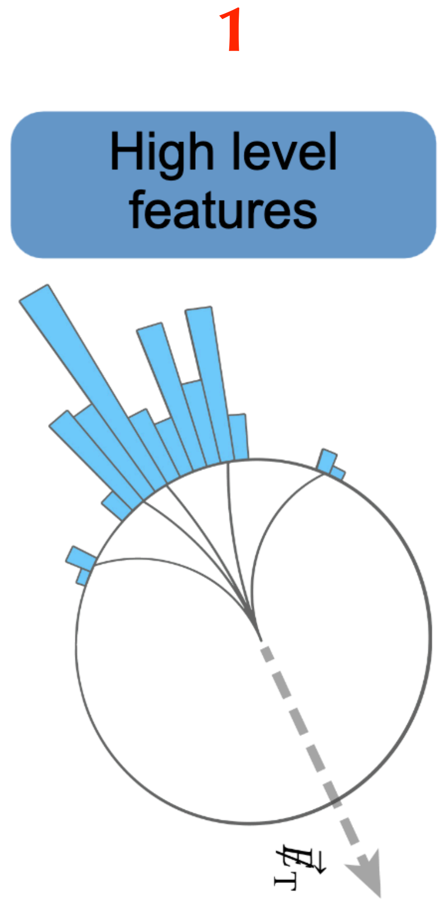


What changed?

- Consider a typical data reduction workflow in place to deal with complex data
- Let's take LHC as an example…
- **This worked well… but can fail when patterns are even more subtle and rare**

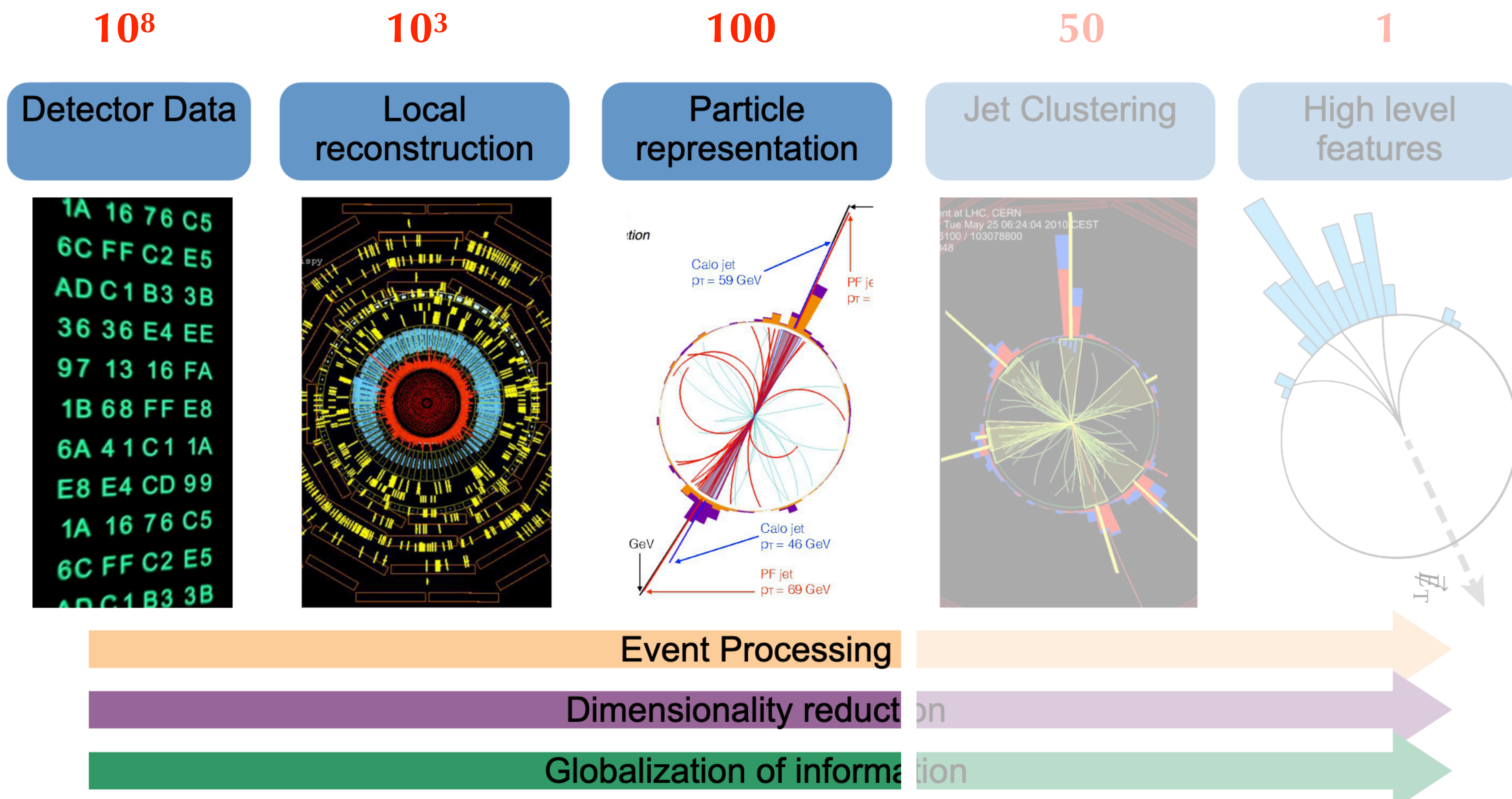


invariant mass of two photons



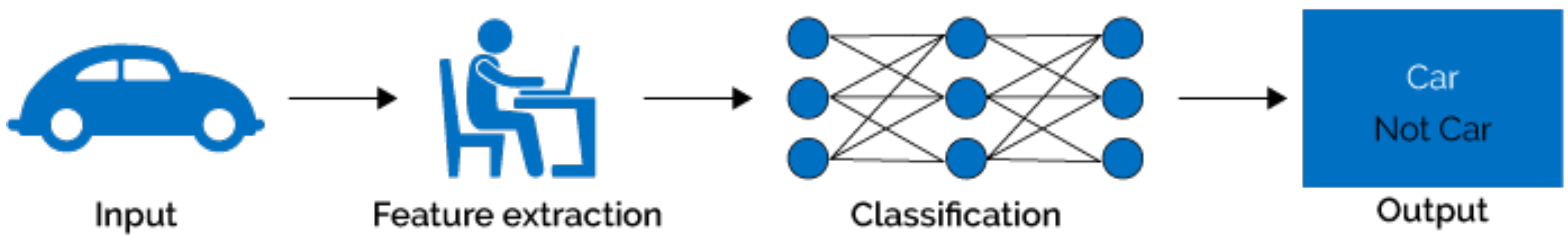
What patterns are we washing out here?

- Go back looking at the source highly dimensional data: ***did we miss something through expert-level data reduction algorithms?***
- An AI could efficiently analyze these data and we can check, as experts, what the answer is: ***does AI match our expectations and/or does it teach us something new?***

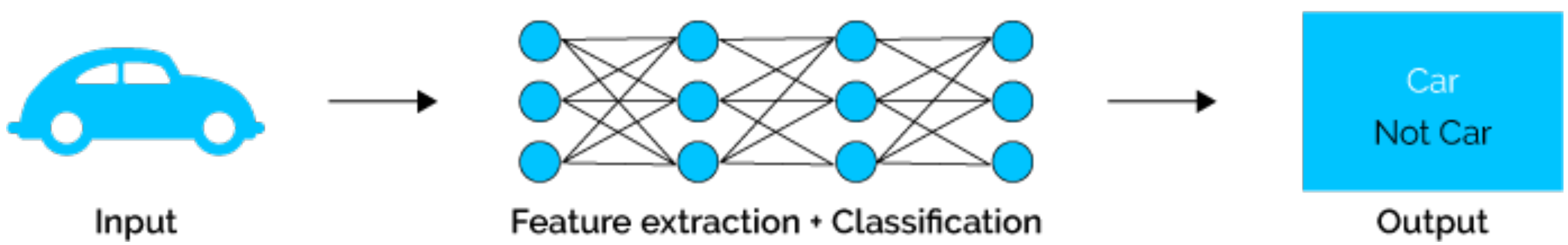


From expert-level features to raw data

A shallow neural network

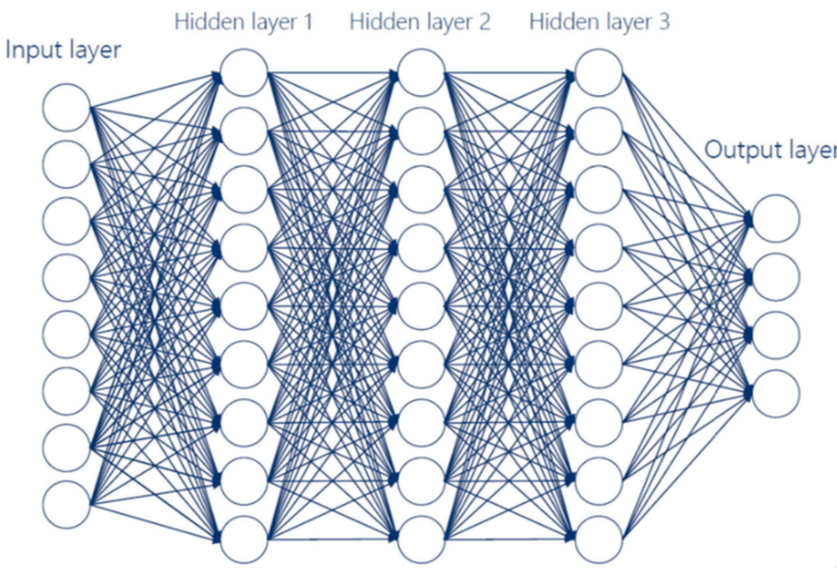


A deep neural network

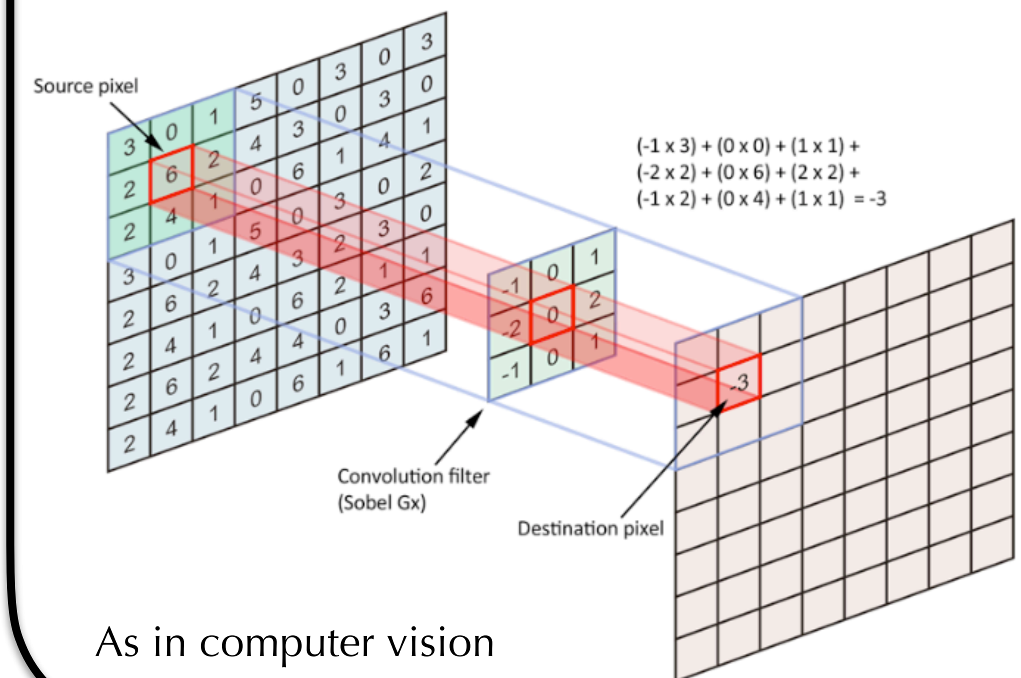


Data representation: which one?

High level/no structure: fully connected NN

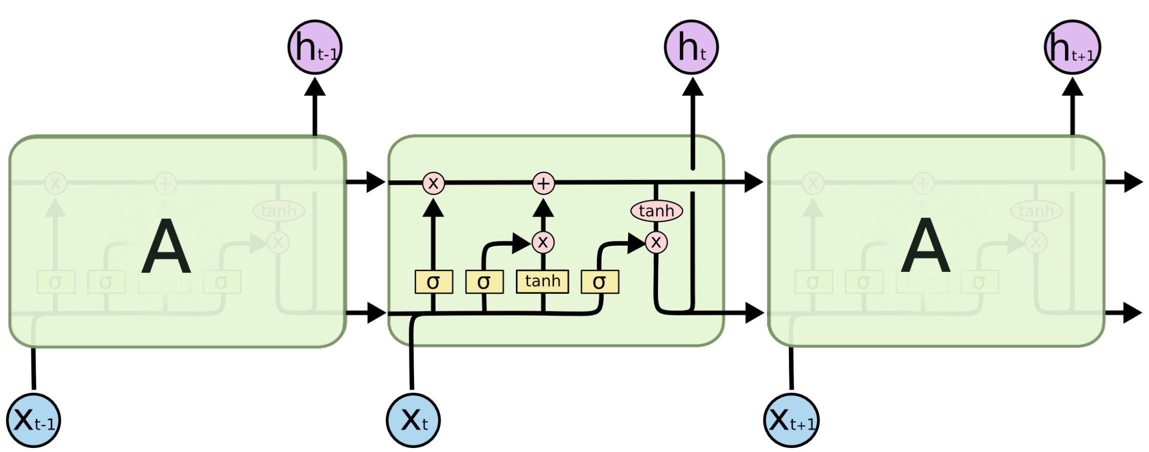


Regular grid: convolutional NN



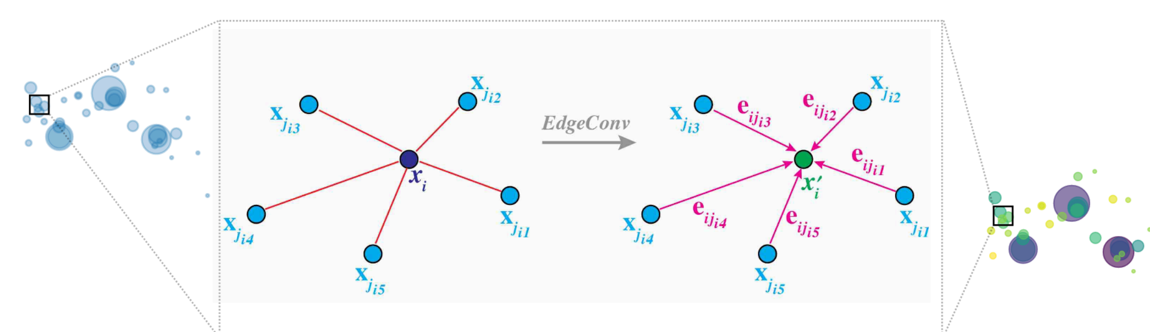
As in computer vision

Ordered sequence/time series: recurrent NN



As in natural language processing

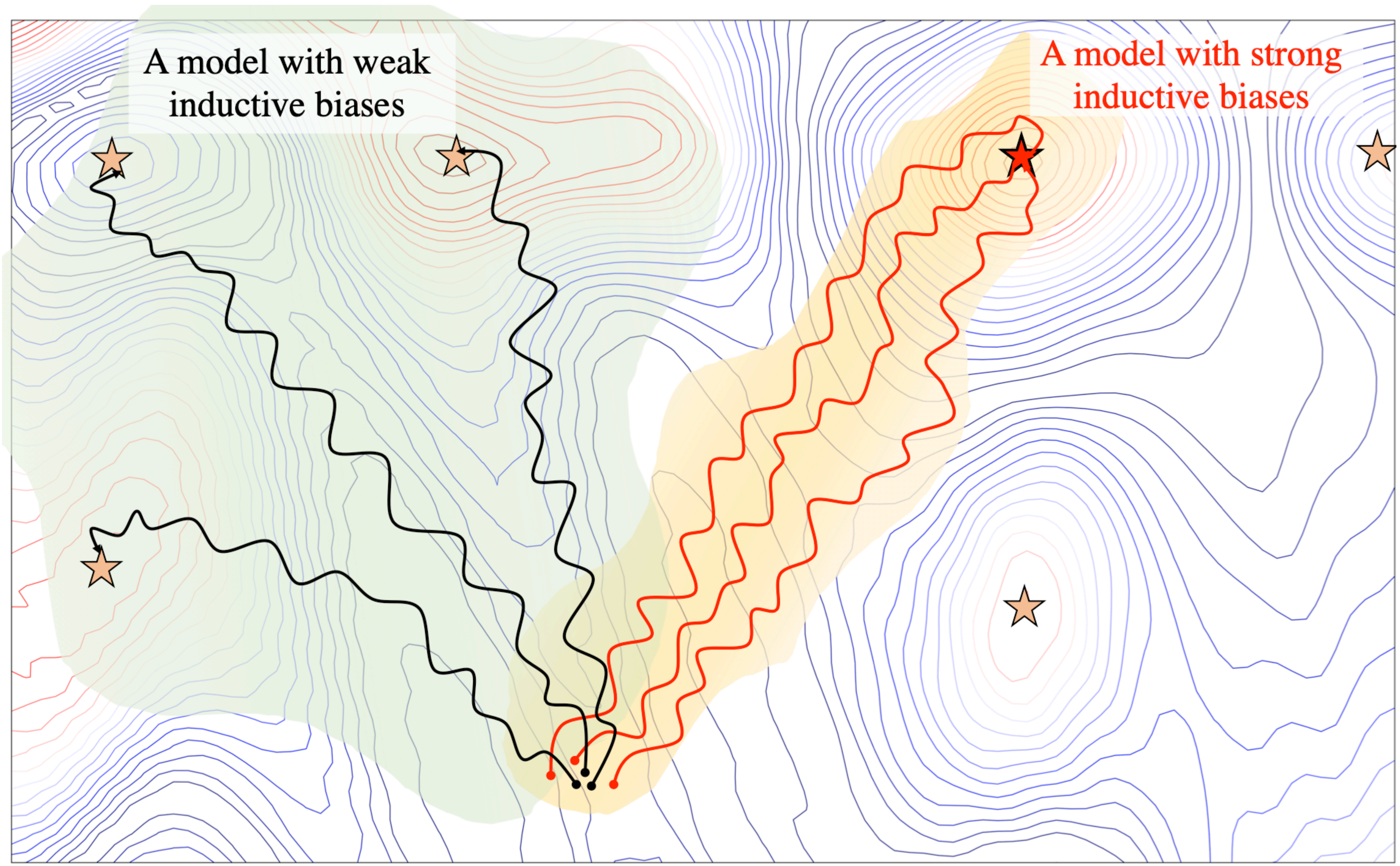
Point cloud: Deep Sets & Graph NN



As in social media analysis

The role of inductive bias

- Incorporating **domain knowledge** into ML *(inductive bias)* can provide **better accuracy, training/inference efficiency, smaller model size, interpretability and robustness**

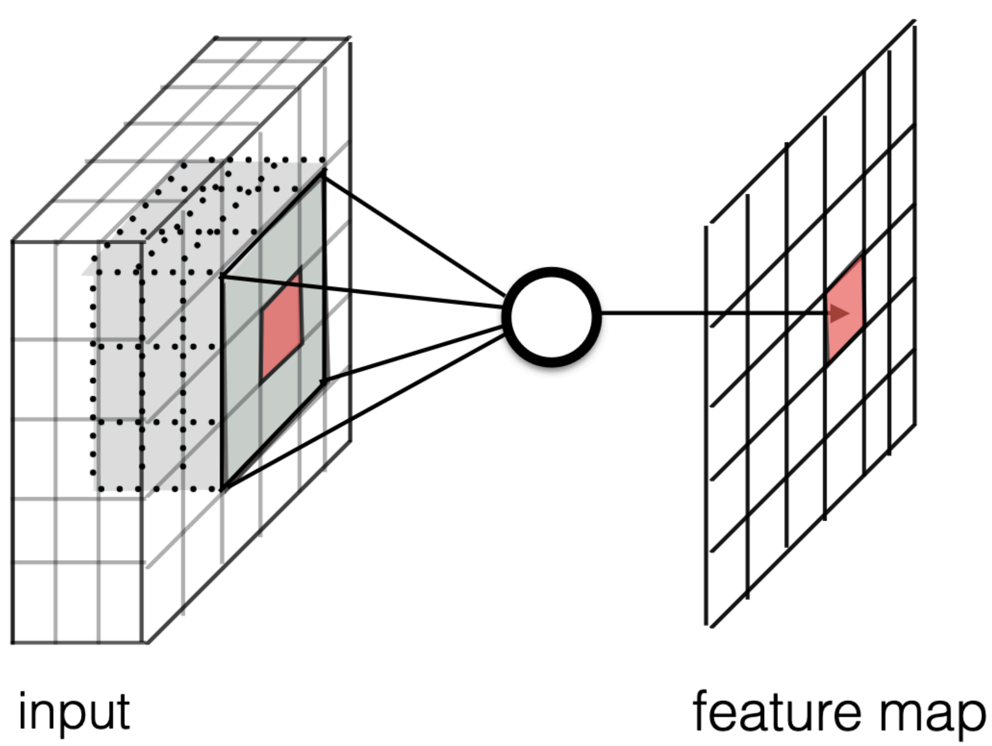


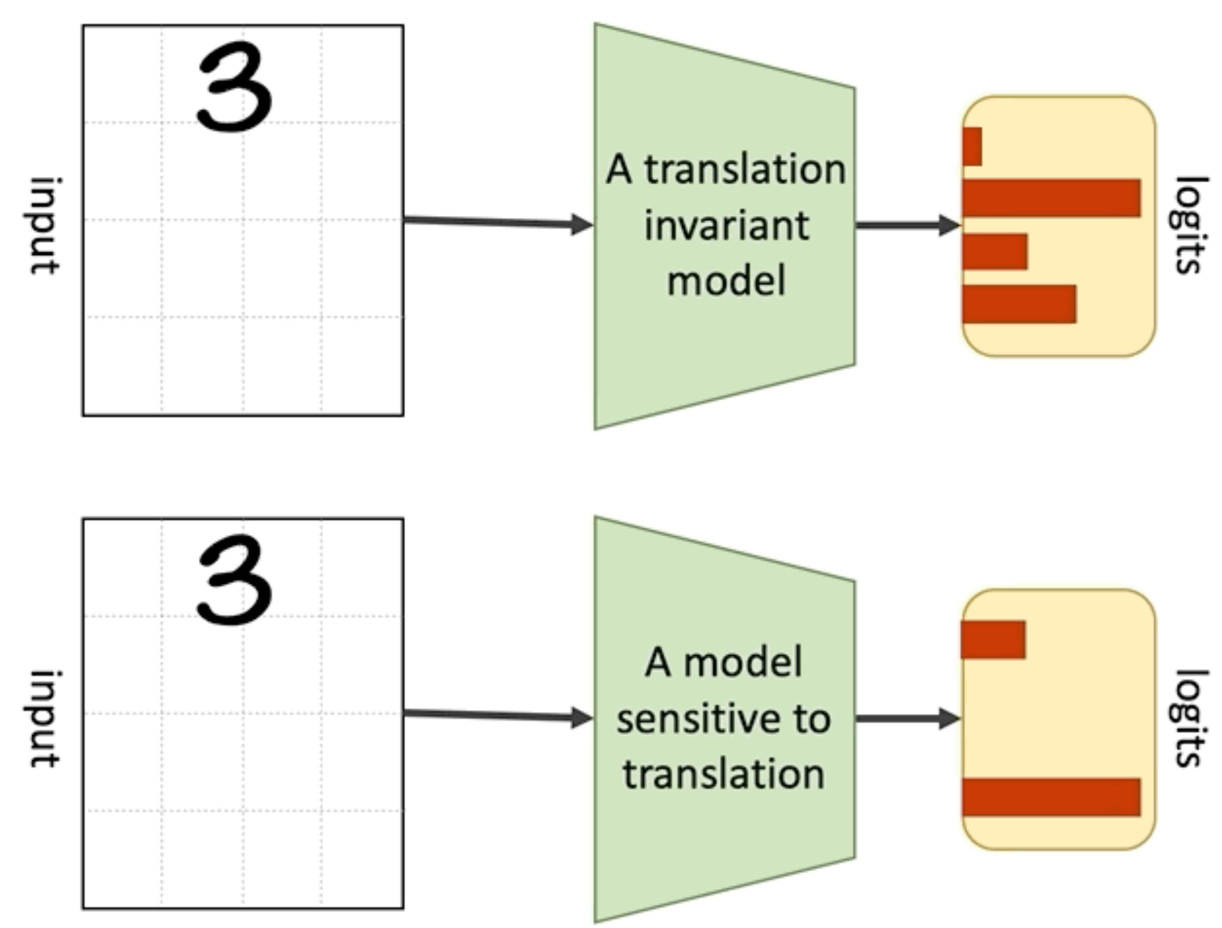
https://samiraabnar.github.io/articles/2020-05/indist

Example: Convolutional NN

- CNNs was a breakthrough: tailored algorithms to the structure (and symmetries) of the image data in computer vision tasks

- **Leverage spatial symmetries** (translation invariance and equivariance) to achieve higher accuracy at lower computational cost wrt fully connected NNs
    - intelligent feature (patterns) extraction from raw pixel-level high-dimensional data
    - dramatic reduction in number of parameters

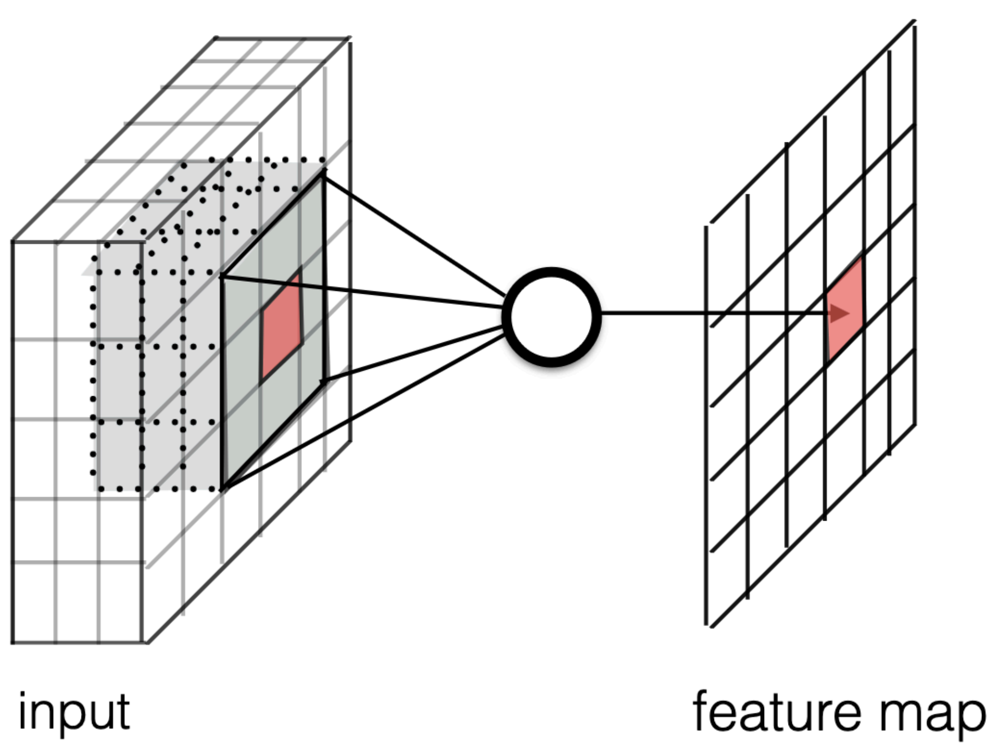


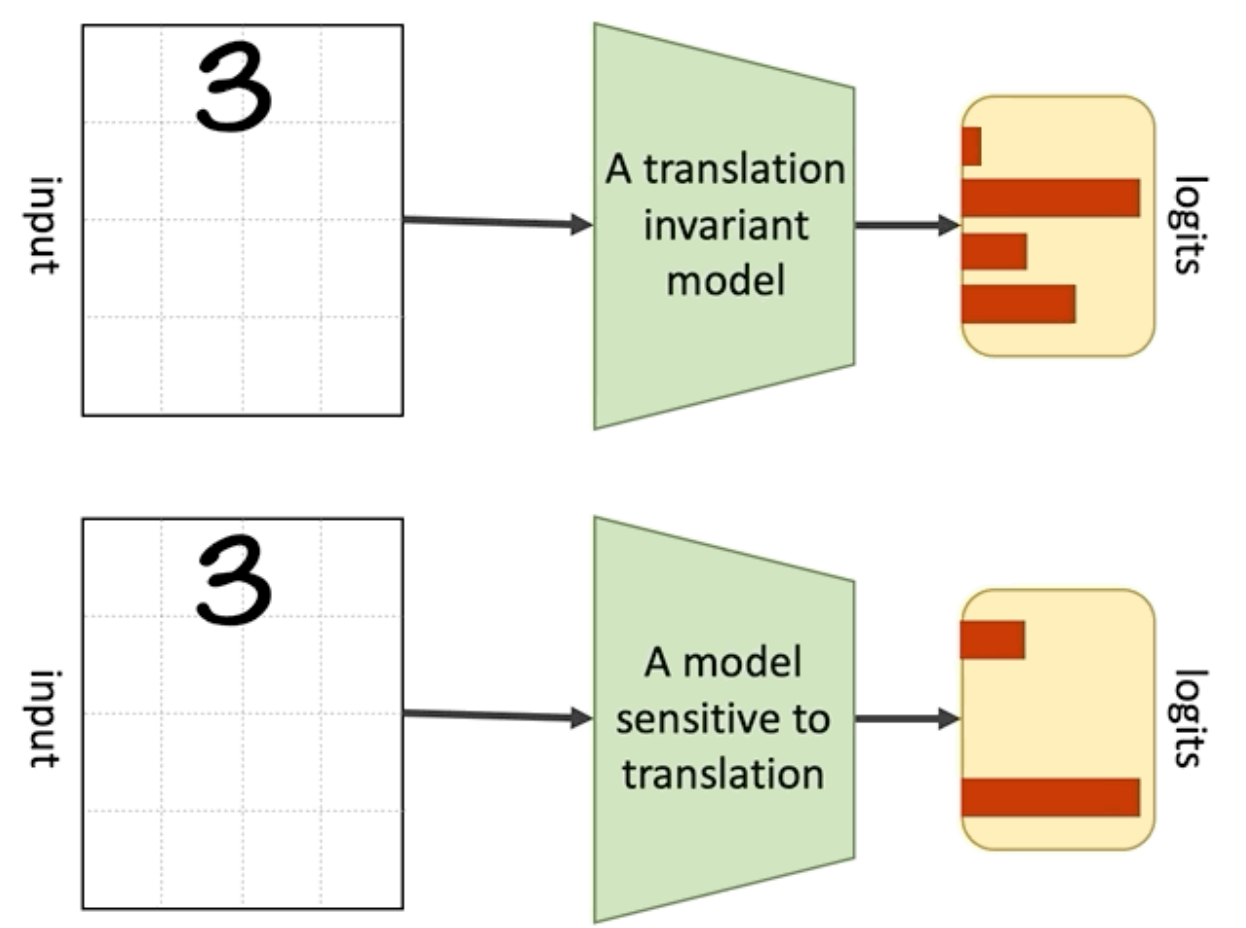


Example: Convolutional NN

- CNNs was a breakthrough: tailored algorithms to the structure (and symmetries) of the image data in computer vision tasks

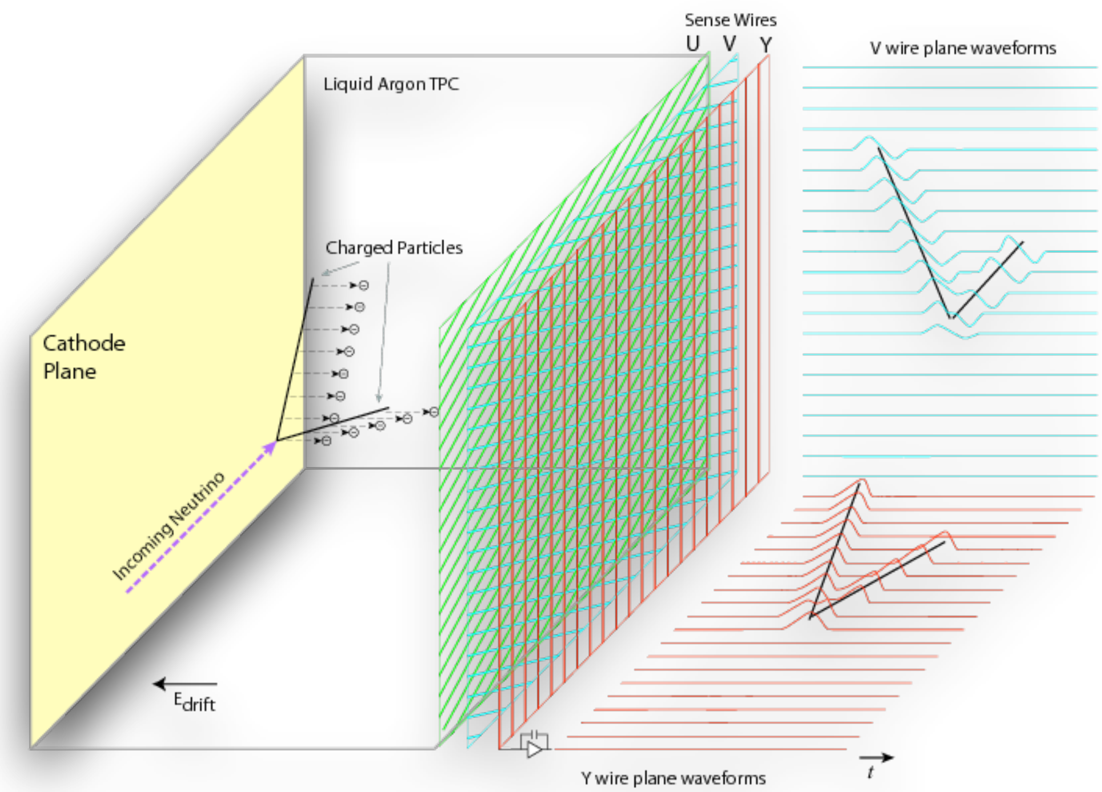
- **Leverage spatial symmetries** (translation invariance and equivariance) to achieve higher accuracy at lower computational cost wrt fully connected NNs
    - intelligent feature (patterns) extraction from raw pixel-level high-dimensional data
    - dramatic reduction in number of parameters





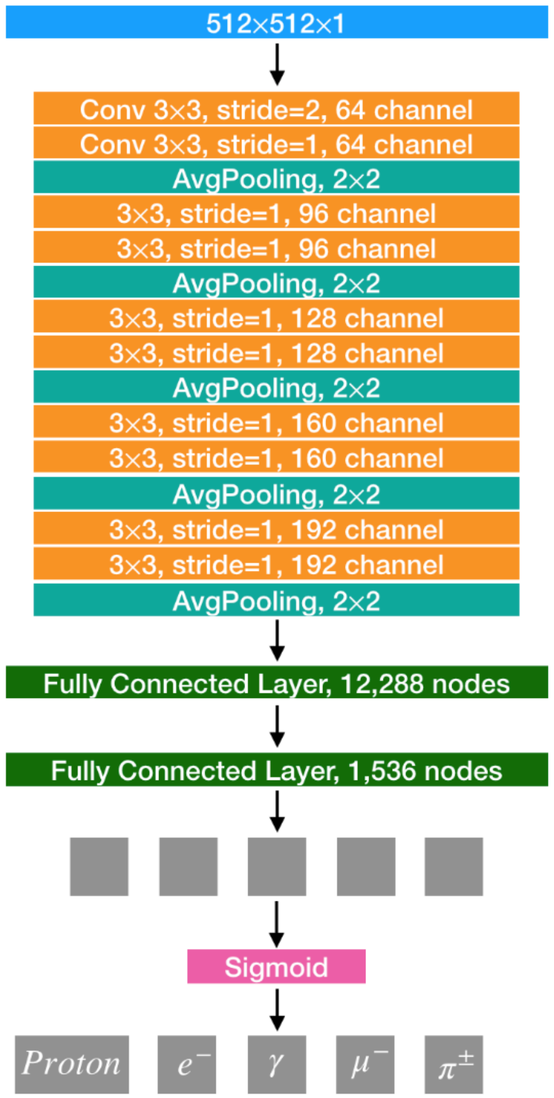
Example: Convolutional NN

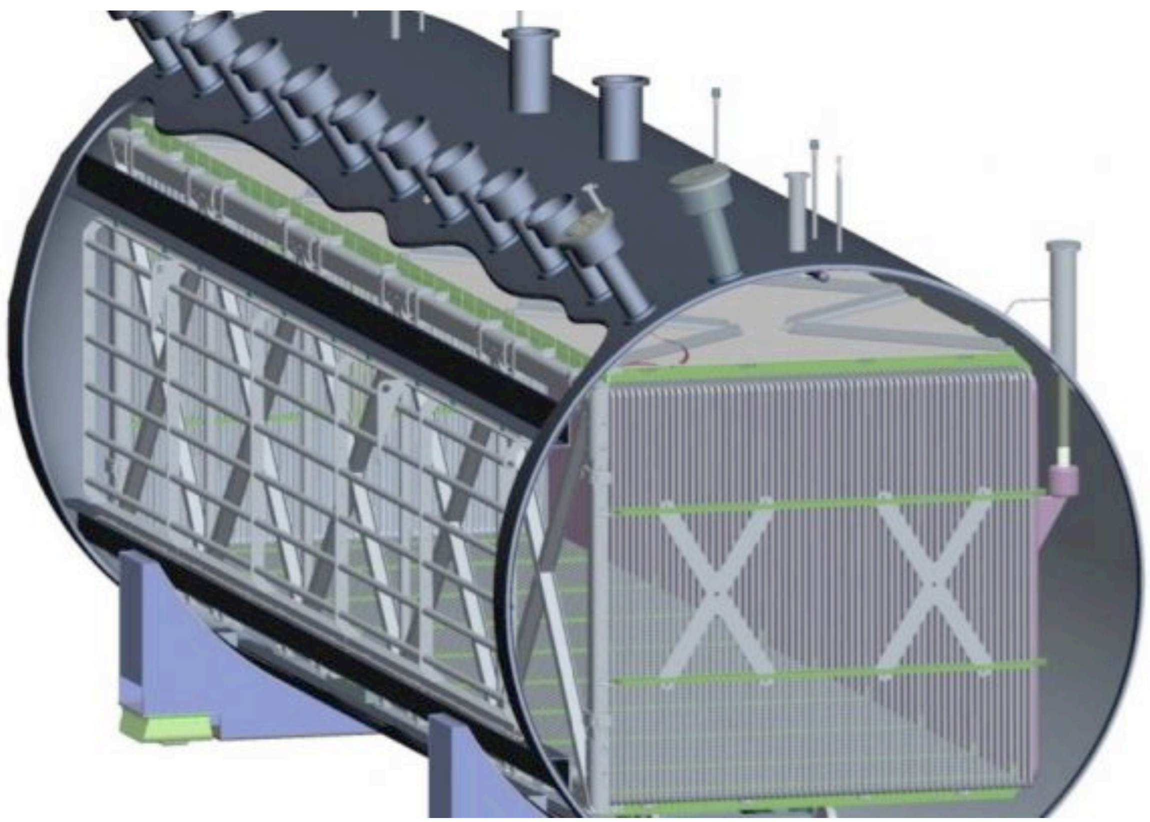
- Well suited for experimental data that come naturally as regular grid like images from **LArTPC detectors**
  - electrons produced by charged particles interacting with a large multiple-cubic meters volume of LAr
  - continuous stream of 3D images of detector volume yielding a **high-resolution "video"**

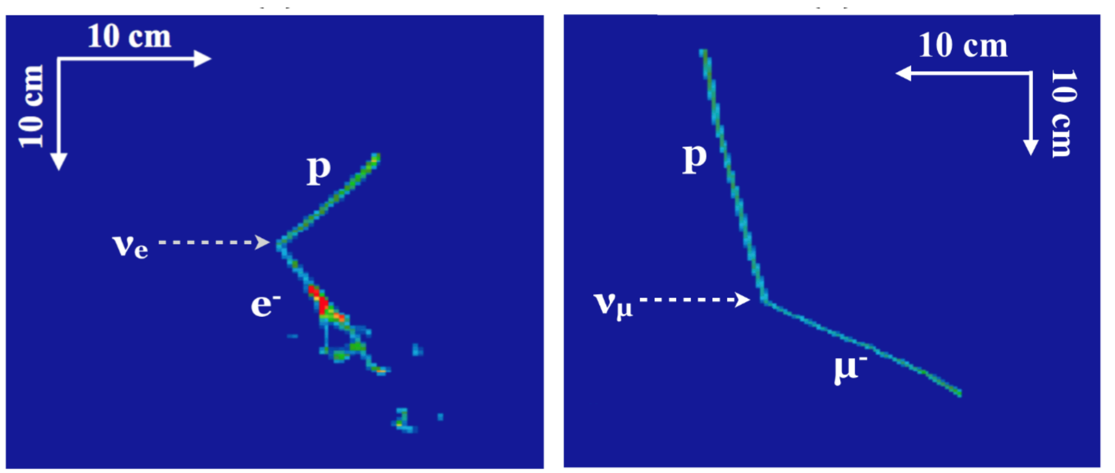


Example: Convolutional NN

- **MicroBooNE was the first LArTPC (170 ton) to deploy CNNs to solve otherwise technically challenging tasks:**
  - event classification [1611.05531]
  - particle ID [1611.05531, 1808.07269, 2010.08653, 2012.08513]
  - region-of-interest detection [1611.05531]

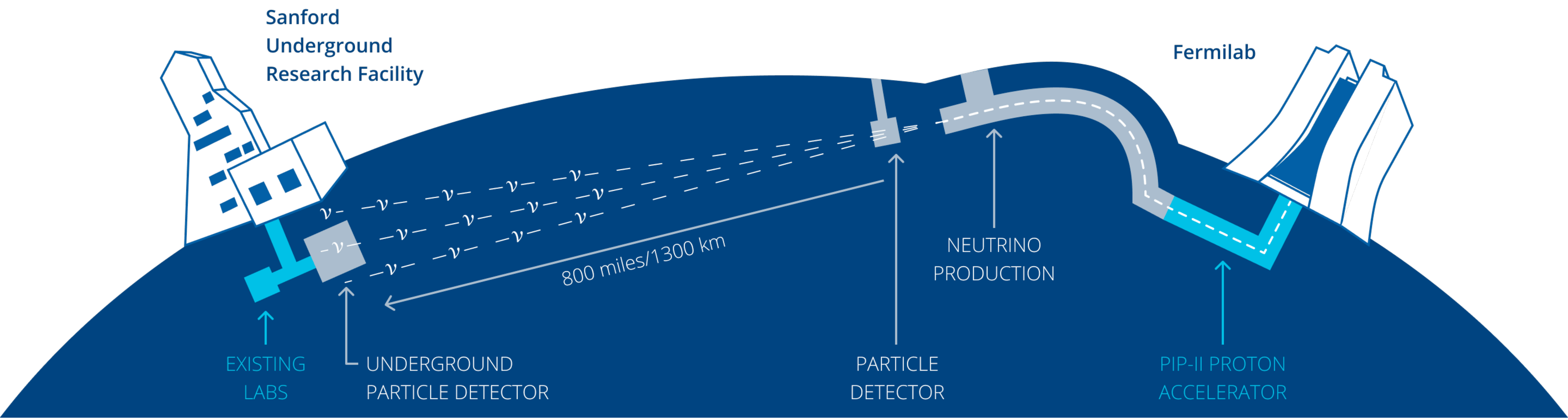






Big data @ the Intensity Frontier

The Deep Underground Neutrino Experiment (DUNE)

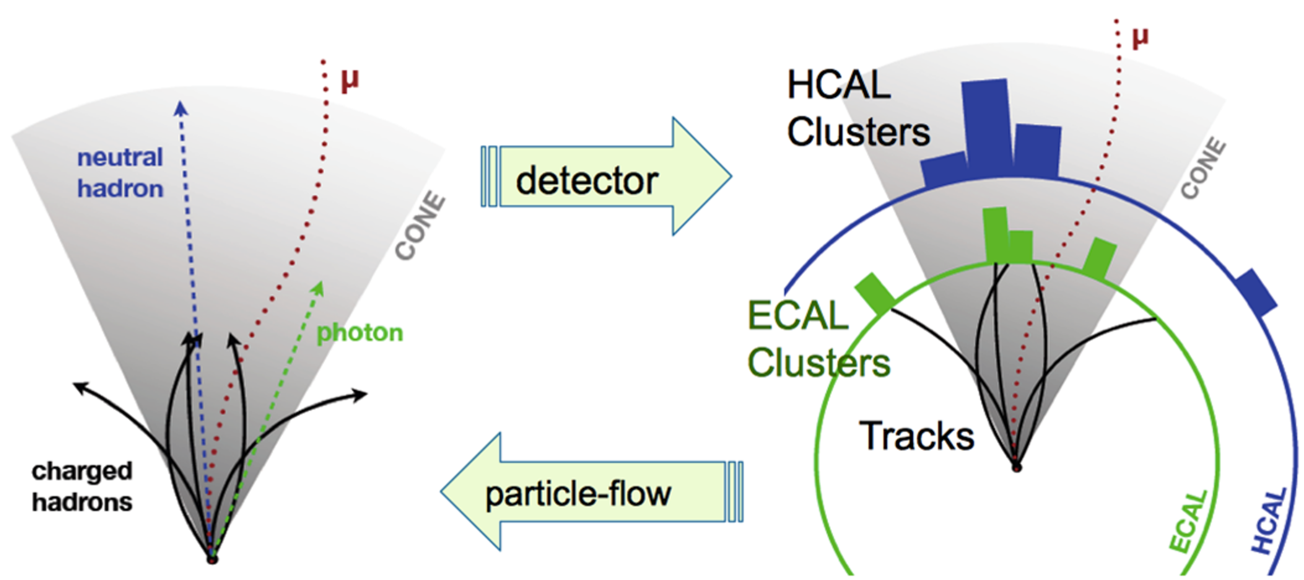


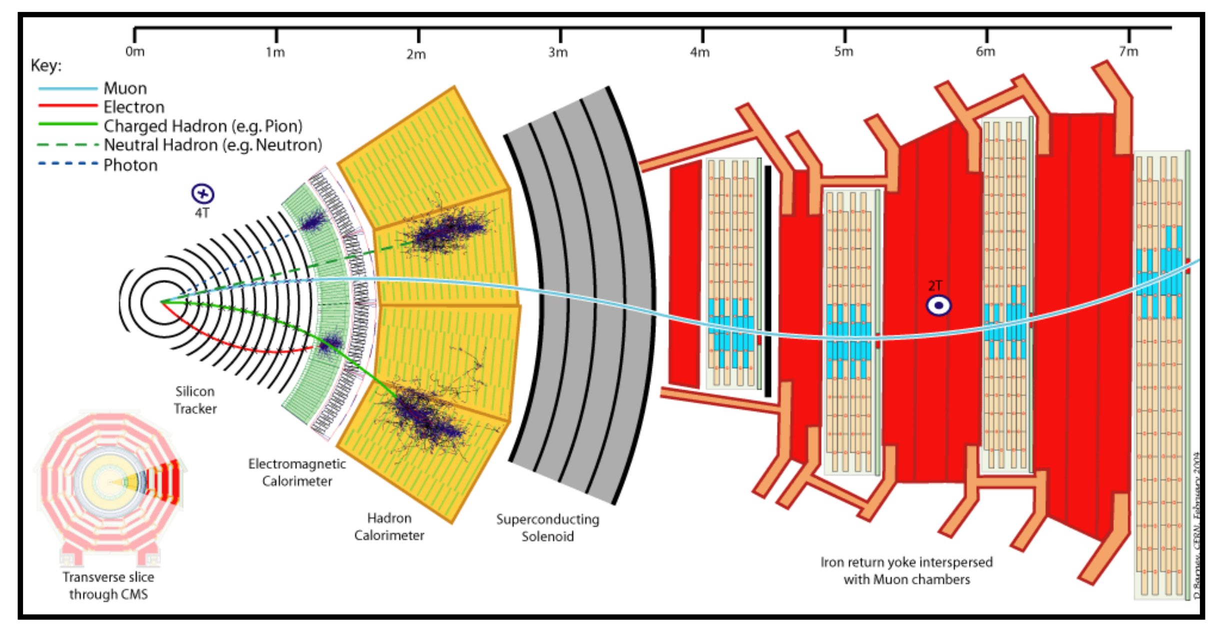
- Next generation neutrinos oscillation experiment now under construction and R&D to start operations in late 2020s
- Massive far detector 1 mile underground comprising **70k tons of LAr** and advanced technology to record neutrino interactions with extraordinary precision
- Uncompressed continuous readout of modules will yield **O(10 Tb/s)** → unprecedented for this type of experiment!

From images to point cloud

- Experimental data are not always arranged as a regular grid-like structure
    - a heterogenous detector can provide high-resolution data on different information types

Eg, the CMS detector





Detector	p_T-resolution	η/Φ-segmentation
Tracker	0.6% (0.2 GeV) – 5% (500 GeV)	0.002 x 0.003 (first pixel layer)
ECAL	1% (20 GeV) – 0.4% (500 GeV)	0.017 x 0.017 (barrel)
HCAL	30% (30 GeV) – 5% (500 GeV)	0.087 x 0.087 (barrel)

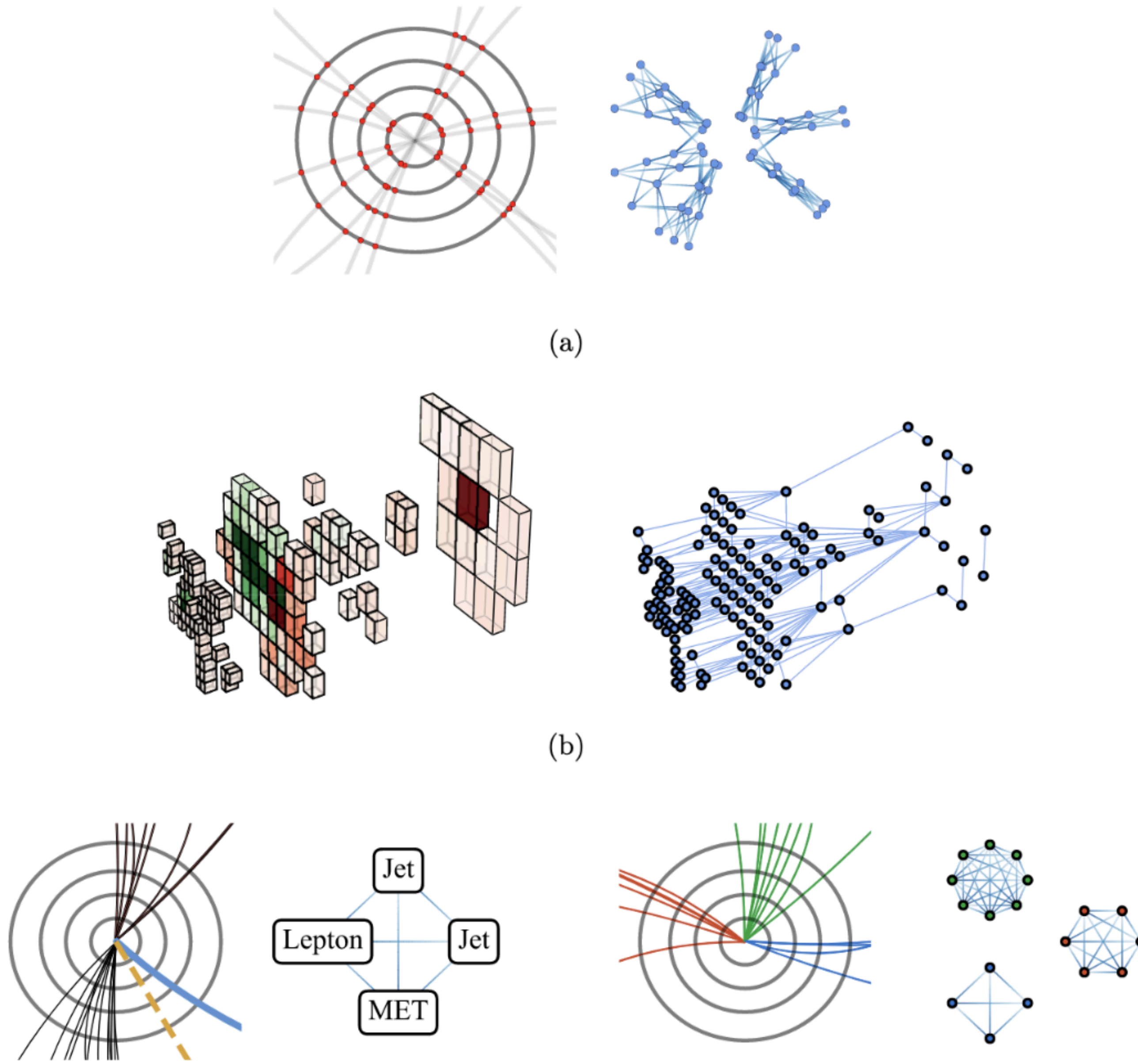
From images to point cloud

- Experimental data are not always arranged as a regular grid-like structure

- **How do these data look like?**
  - Distributed unevenly in space
  - Sparse
  - Heterogenous
  - Variable size
  - No defined order
  - Interconnections

- A **point cloud representation** provides the required flexibility

- **Graph Neural Networks** architectures can be designed that leverage physics laws → **inductive bias**
  - permutation invariance/equivariance
  - symmetry group equivariance



arXiv.2203.12852

Graph NNs in HEP

- Represent objects as points with pairwise relationships
- Effectively capture complex relationships and dependencies between objects of many different kinds in HEP
    - energy deposits, individual physics objects, individual particles, heterogenous information
- **Applications and architectures keep successfully growing!**

Static isotropic
- E.g. GCN

Static anisotropic
- E.g. Interaction Network

Dynamic (An)isotropic
- E.g. GravNet

Node prediction
- E.g. Node regression or classification

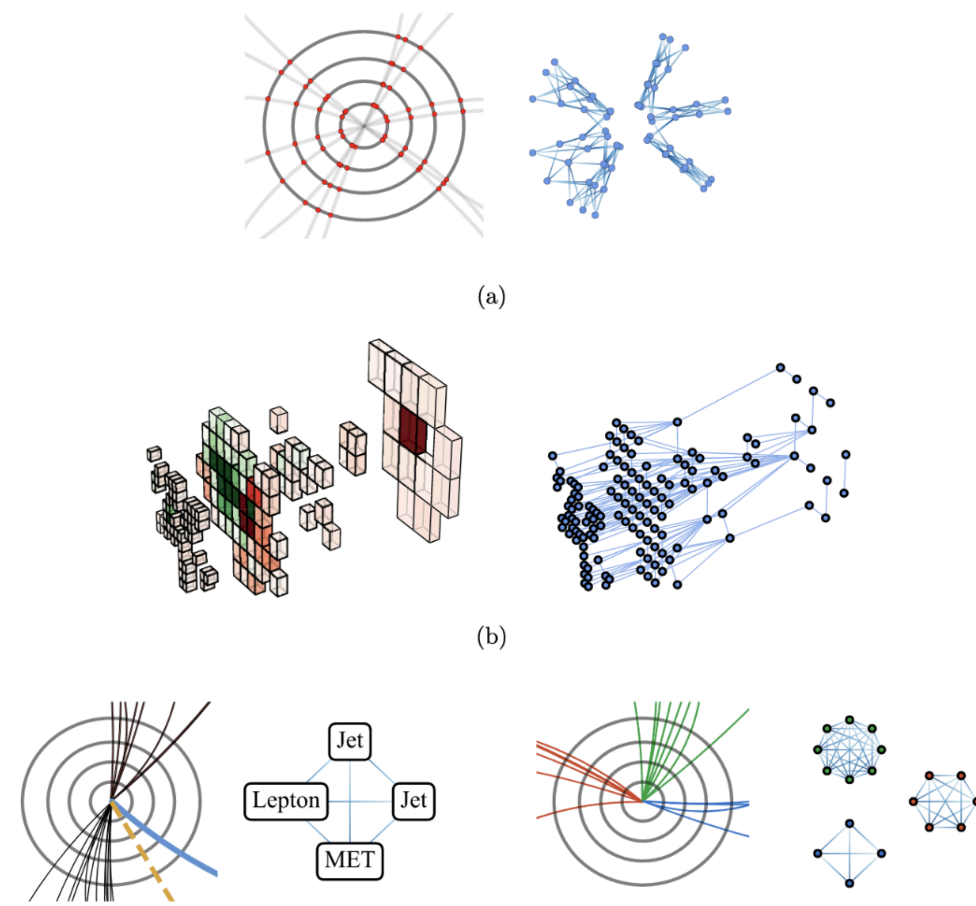
Edge prediction
- E.g. Social network link prediction

Graph prediction
- E.g. Molecular property regression

Object segmentation
- E.g. Find all hydrogen in graph

Instance segmentation
- E.g. Find *each* hydrogen in graph

Spatio-Temporal
- E.g. STGCN (Graph conv. + temporal conv.)



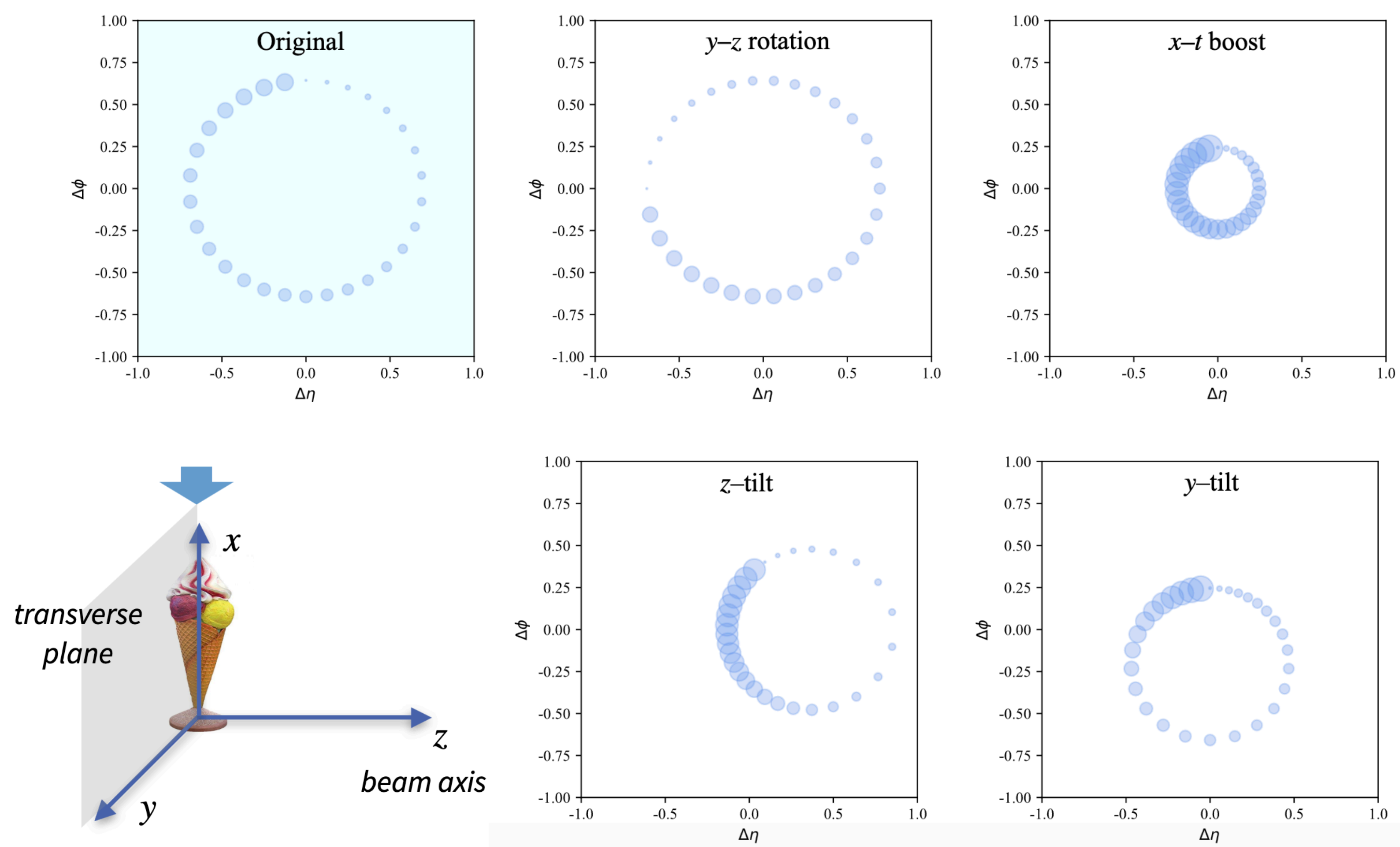
(a)

(b)

arXiv.2203.12852 arXiv.2007.13681

Physics-informed ML

- Target applications of ML in HEP often have specific features such as symmetries, invariances/equivariances unique to our field
- Developing **physics-specific solution can lead to improved performance**
- A growing effort to design and study architectures with **injected symmetries**
- **Dedicated NNs have been proposed such as to be invariant/equivariant to certain symmetries, e.g.:**
  - permutations
  - boost on z-axis,
    rotation on x-y plane
  - rotation on the η–ϕ plain
  - boost along the "jet axis"
  - full Lorentz transformations

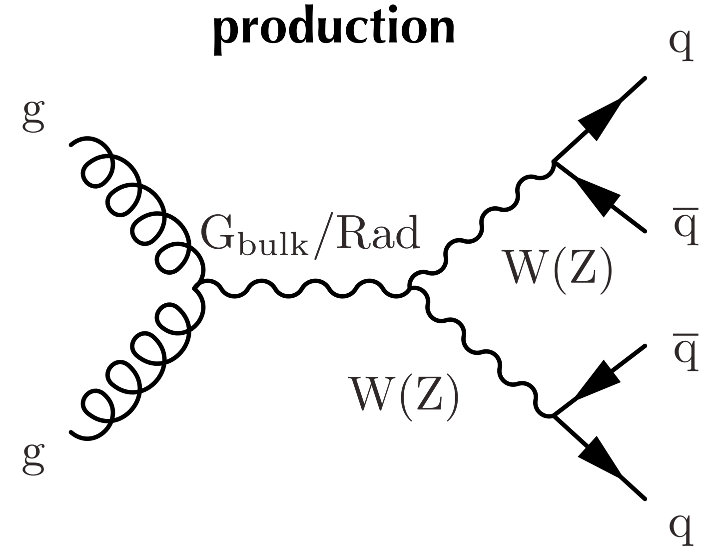


arXiv.2208.07814

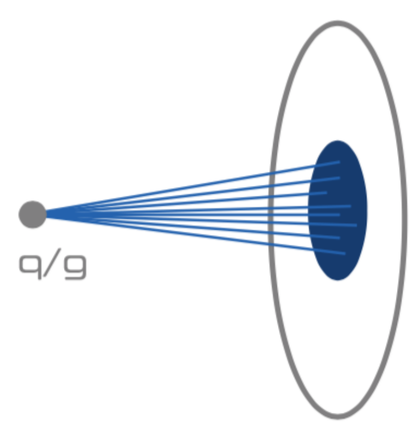
Example: jet tagging

- **Identification of jets arising from hadronization of boosted W/Z/H/top is a key task in LHC physics:**
  - new physics searches, standard model measurements, higgs sector
  - **unique signature** from hadrons merging in single jet with substructure
  - exploit to suppress **overwhelming background from multijet** processes in most sensitive all-hadronic and semi-leptonic channels
- A topic of interest in both theory and experiment communities since ~ 30 years
- Recent years advancement in ML enabled more powerful algorithms (graph NNs, transformers, …)

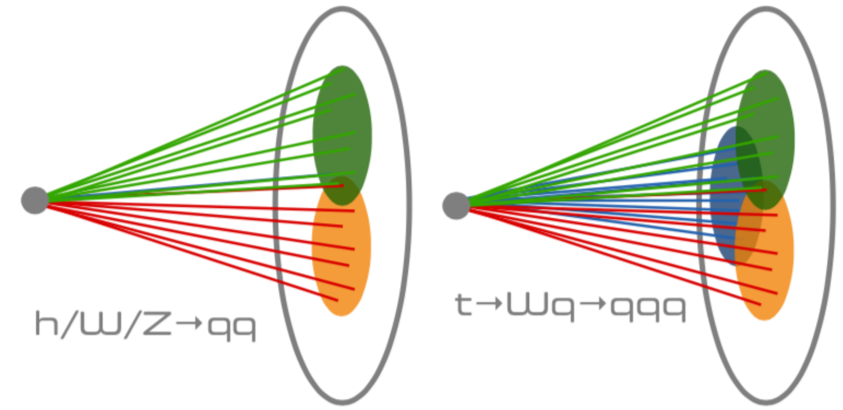
ex: Graviton or Radion production



BACKGROUND JET (single q/g)



SIGNAL JETS

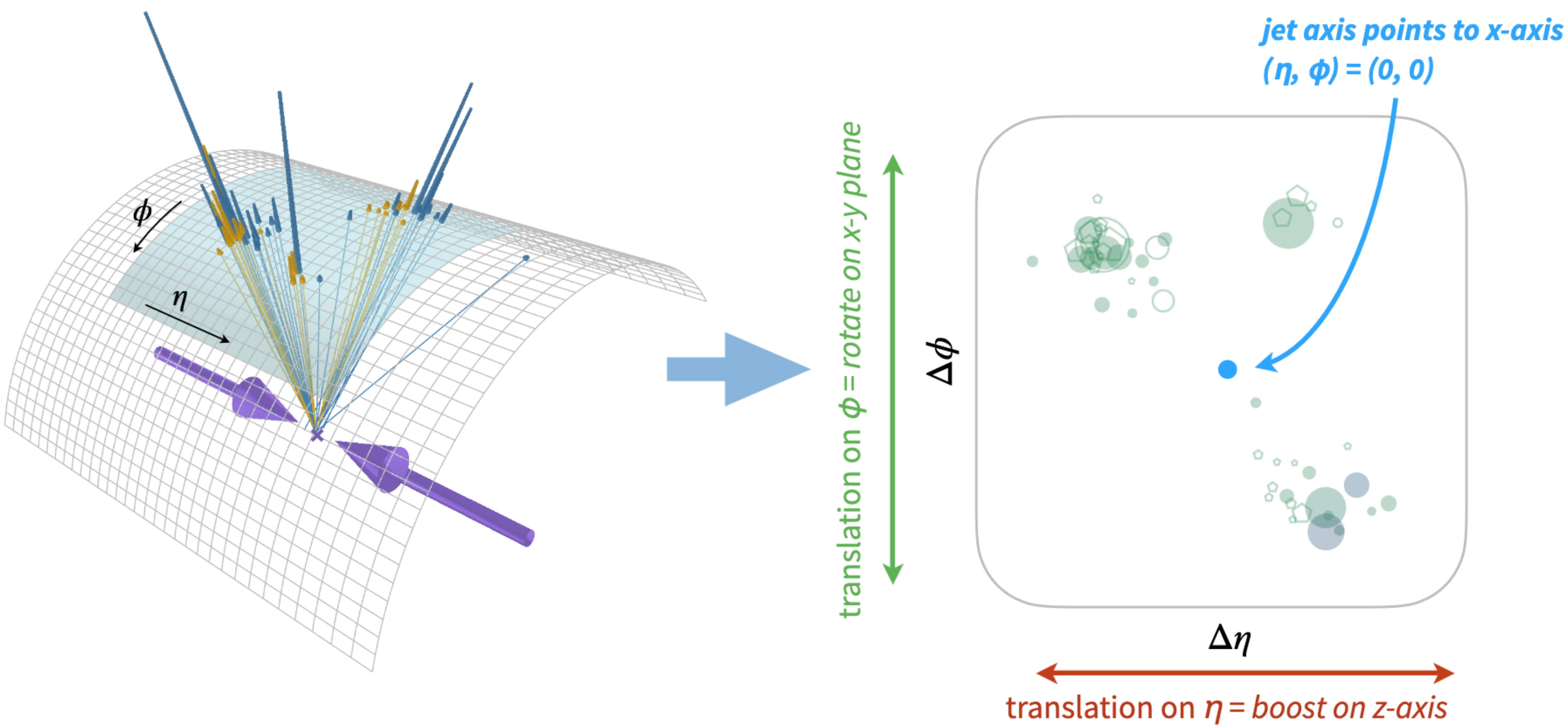


Example: jet tagging

- **Invariance under a Lorentz boost along the beam axis** (z-boost) has been obtained in the past through input preprocessing

- In this case the jet is conventionally represented as a set of particles with relative coordinates $\Delta\eta/\Delta\phi$ w.r.t. jet axis

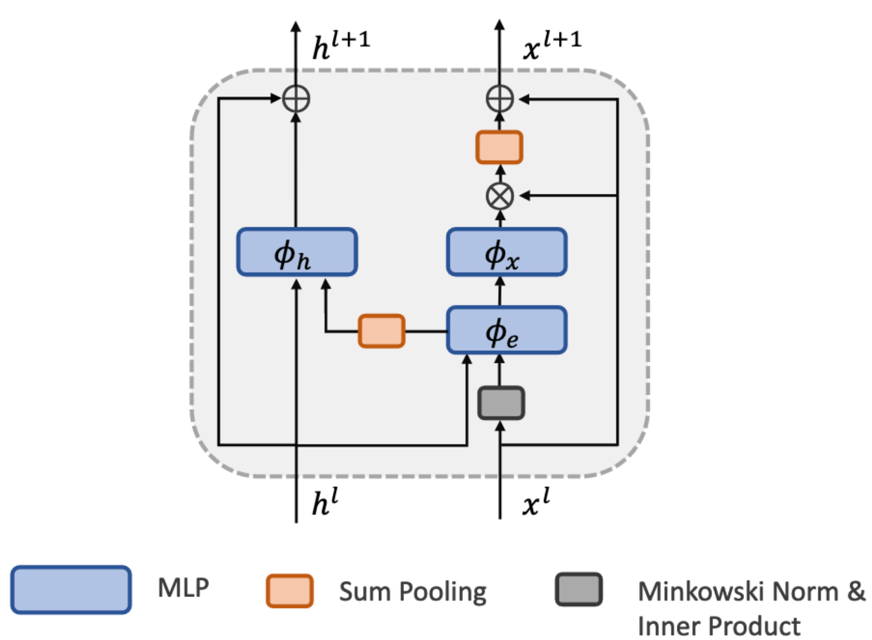
  - this pre-processing step is equivalent as: ***apply a boost on z-axis*** → ***then a rotation on x-y plane*** (transverse plane) → **now jet points to the *x*-axis**, *i.e.* $(\eta, \phi) = (0, 0)$



Example: jet tagging

- Jet is represented as a set of particles with relative coordinates Δη/Δϕ w.r.t. jet axis
  - after pre-processing, we still have four additional DoFs for Lorentz transformation!
- A NN that respects Lorentz symmetry outputs a score that is invariant under any Lorentz transformation of the input jet

- **A solution: design a dedicated structure to maintain invariant/equivariant**



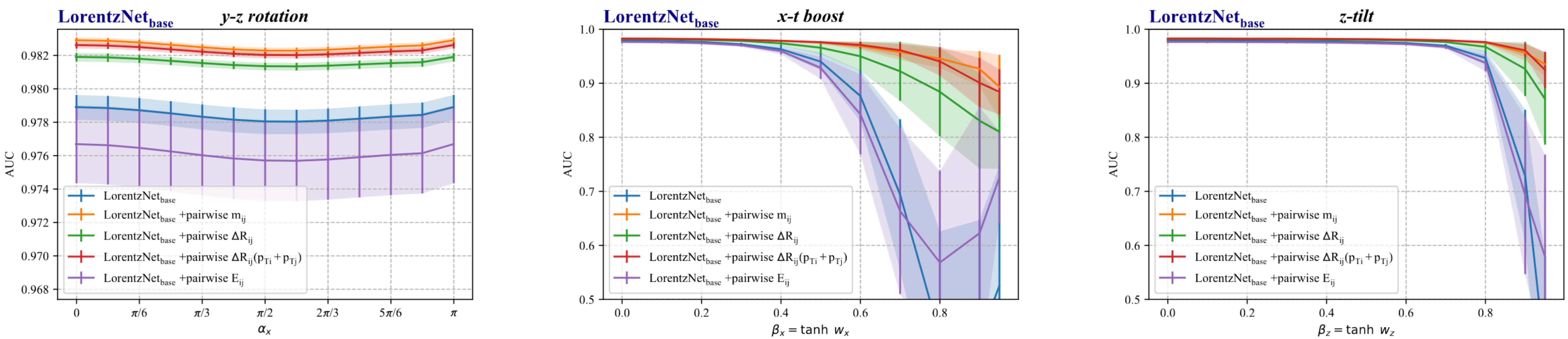
Lorentz Group Equivariant Block (LGEB)

LorentzNet:
Physics-informed graph edge features given by the Minkowski inner product of two 4-vectors per each particle pair + Lorenz invariant particle interactions

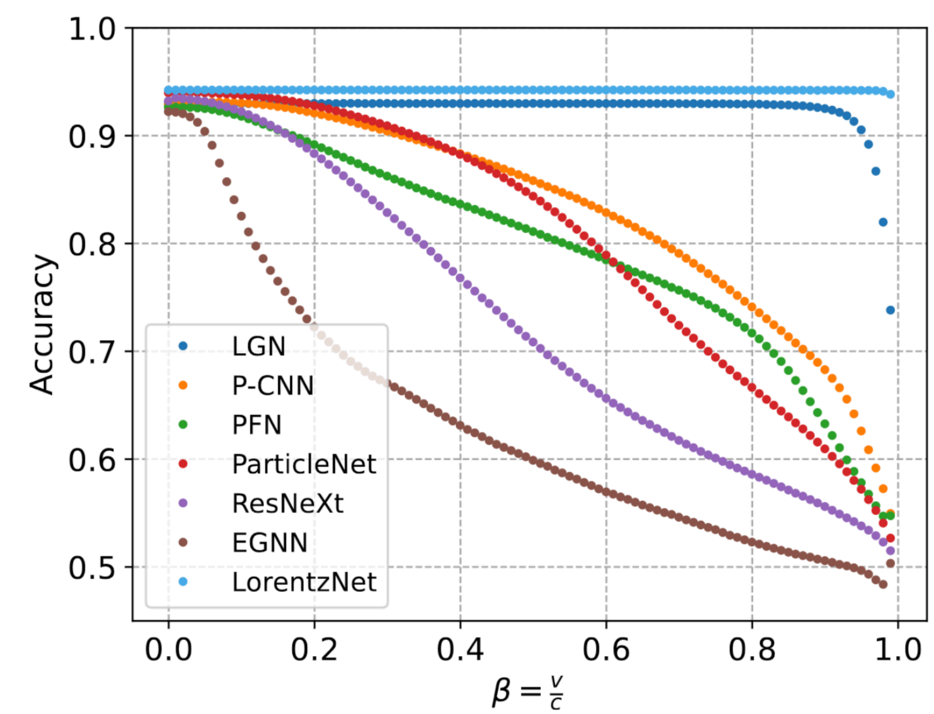
$$m_{ij}^l = \phi_e\left(h_i^l, h_j^l, \psi(\|x_i^l - x_j^l\|^2), \psi(\langle x_i^l, x_j^l\rangle)\right)$$

JHEP 07, 30 (2022)

Example: jet tagging



Model	Equivariance	Time on CPU (ms/batch)	Time on GPU (ms/batch)	#Params
ResNeXt	✗	5.5	0.34	1.46M
P-CNN	✗	**0.6**	**0.11**	348k
PFN	✗	**0.6**	0.12	82k
ParticleNet	✗	11.0	0.19	366k
EGNN	E(4)	30.0	0.30	222k
LGN	$SO^+(1,3)$	51.4	1.66	4.5k
LorentzNet	$SO^+(1,3)$	32.9	0.34	224k



Data efficiency

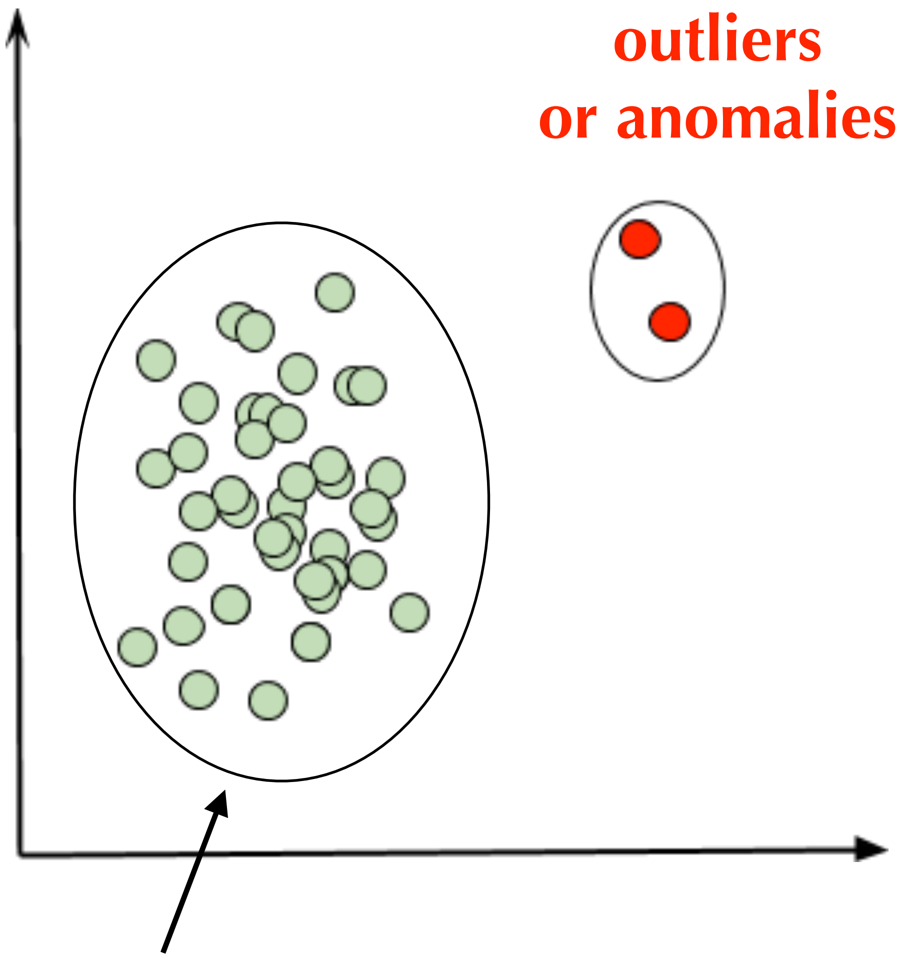
Training Fraction	Model	Accuracy	AUC	$1/\varepsilon_B$ ($\varepsilon_S = 0.5$)	$1/\varepsilon_B$ ($\varepsilon_S = 0.3$)
0.5%	ParticleNet	0.913	0.9687	77 ± 4	199 ± 14
	LorentzNet	**0.929**	**0.9793**	$\mathbf{176 \pm 14}$	$\mathbf{562 \pm 72}$
1%	ParticleNet	0.919	0.9734	103 ± 5	287 ± 19
	LorentzNet	**0.932**	**0.9812**	$\mathbf{209 \pm 5}$	$\mathbf{697 \pm 58}$
5%	ParticleNet	0.931	0.9807	195 ± 4	609 ± 35
	LorentzNet	**0.937**	**0.9839**	$\mathbf{293 \pm 12}$	$\mathbf{1108 \pm 84}$

arXiv.2208.07814

JHEP 07, 30 (2022)

Away from supervision

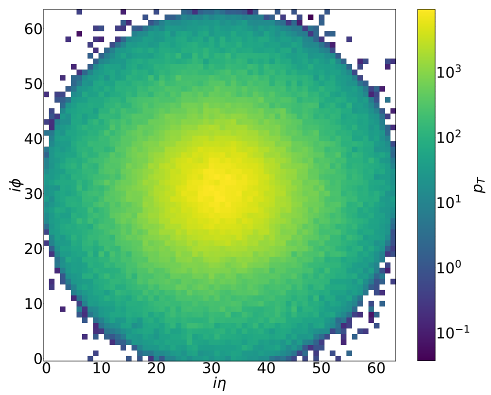
- **Most of the tasks in HEP are supervised,** i.e. ground truth labels or values are given to guide the learning
    - signal = 1 vs. background = 0 → classification
    - target = observable (e.g. the Higgs mass) → regression
- Novel **unsupervised** approaches being explored for new physics searches
    - fully data driven
    - no signal prior
- **The anomaly detection approach:**
    - identifying rare events in data sets which deviate significantly from the majority of the data and do not conform to "normal" behaviour
    - normal behaviour can be learnt through a NN

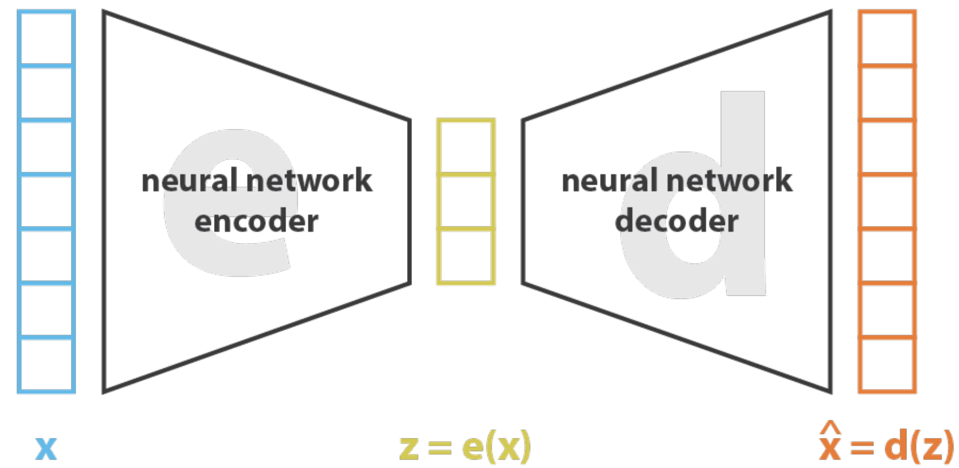


How do we learn the normal behaviour?

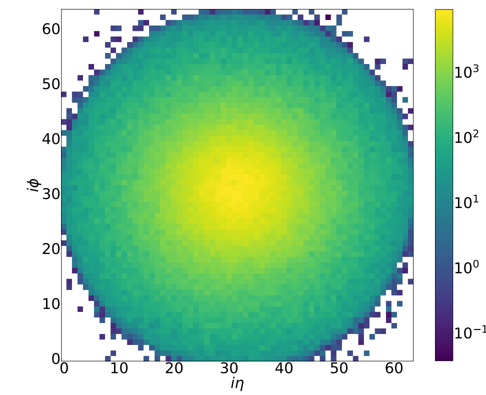
Anomaly detection for jets

e.g, jet images



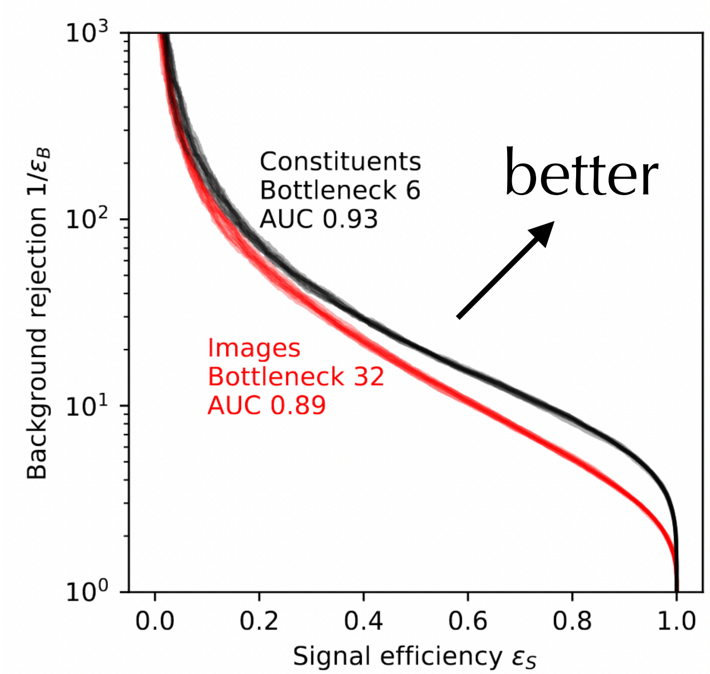


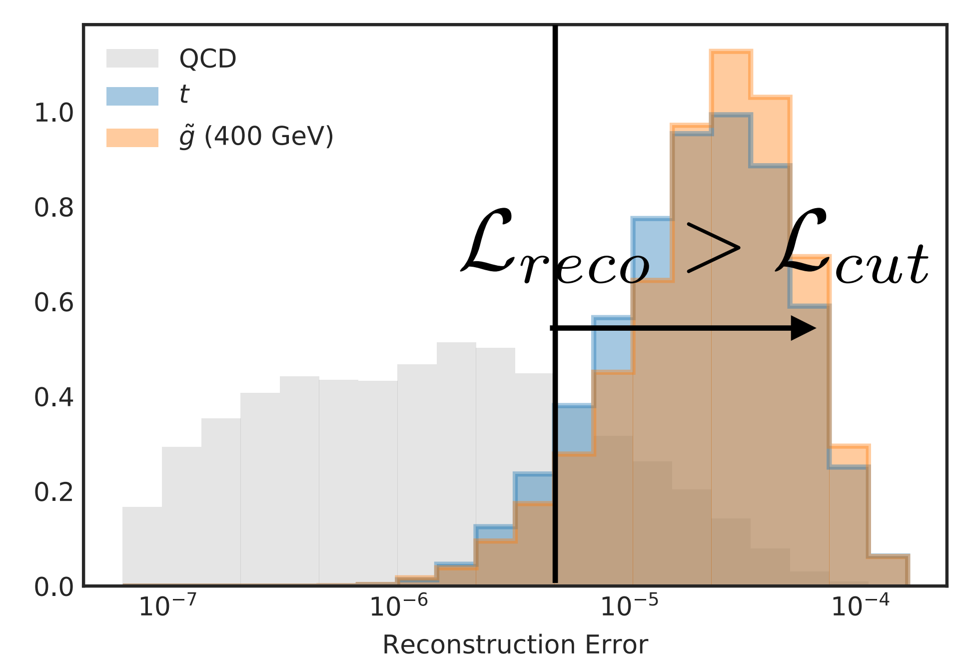
e.g, jet images



$$\text{loss} = ||x - \hat{x}||^2 = ||x - d(z)||^2 = ||x - d(e(x))||^2$$

- One of the first applications was for signal-independent jet tagging using images
- Recently also point-cloud AE architectures were studied [see 2212.07347]





[*] Heimel et al.: SciPost Phys. 6, 030 (2019) , Farina et al.: Phys. Rev. D 101, 075021 (2020)

Apply to the analysis

- Many different strategies studied but no single strategy has emerged as the most universally powerful → **big community effort!**

- CMS and ATLAS analyses using these techniques are now emerging and growing in number [2306.03637, 2005.02983, CERN-EP-2023-112]

- **Many challenges still remain, e.g.:**
  - training setup (e.g., avoid spurious correlations resulting in "fake" anomalies)
  - incorporate physics knowledge w/o loosing generalizability to unknown physics
  - extension to more complex final states
  - hard to find a control region to test robustness
  - anomalies interpretations

The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

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

arXiv.2101.08320

More AI = less interpretability?

- When studying the universe, physicists do not just want the facts, but want to understand why things work the way they do
- Similarly AI demands explanation, not just accurate predictions
  - how did it solve the problem
  - does the solution make sense
- Major challenge for interpretability of ML models comes their strength: **the ability to non-parametrically describe non-linear functions of high-dimensional data**
  - with unconstrained functional form ability to discover unexpected strategies but also cloaks the learned strategy within a black box
- One could open the box but **challenging to extract insights from thousands of nodes and their millions connections**

More AI = less interpretability?

- In simple cases (using expert level features), drop an input feature or decorrelate the model from feature to find feature importance

- For unstructured data like images, new approaches being developed to tackle this crucial problem although still a relatively limited efforts expected to grow

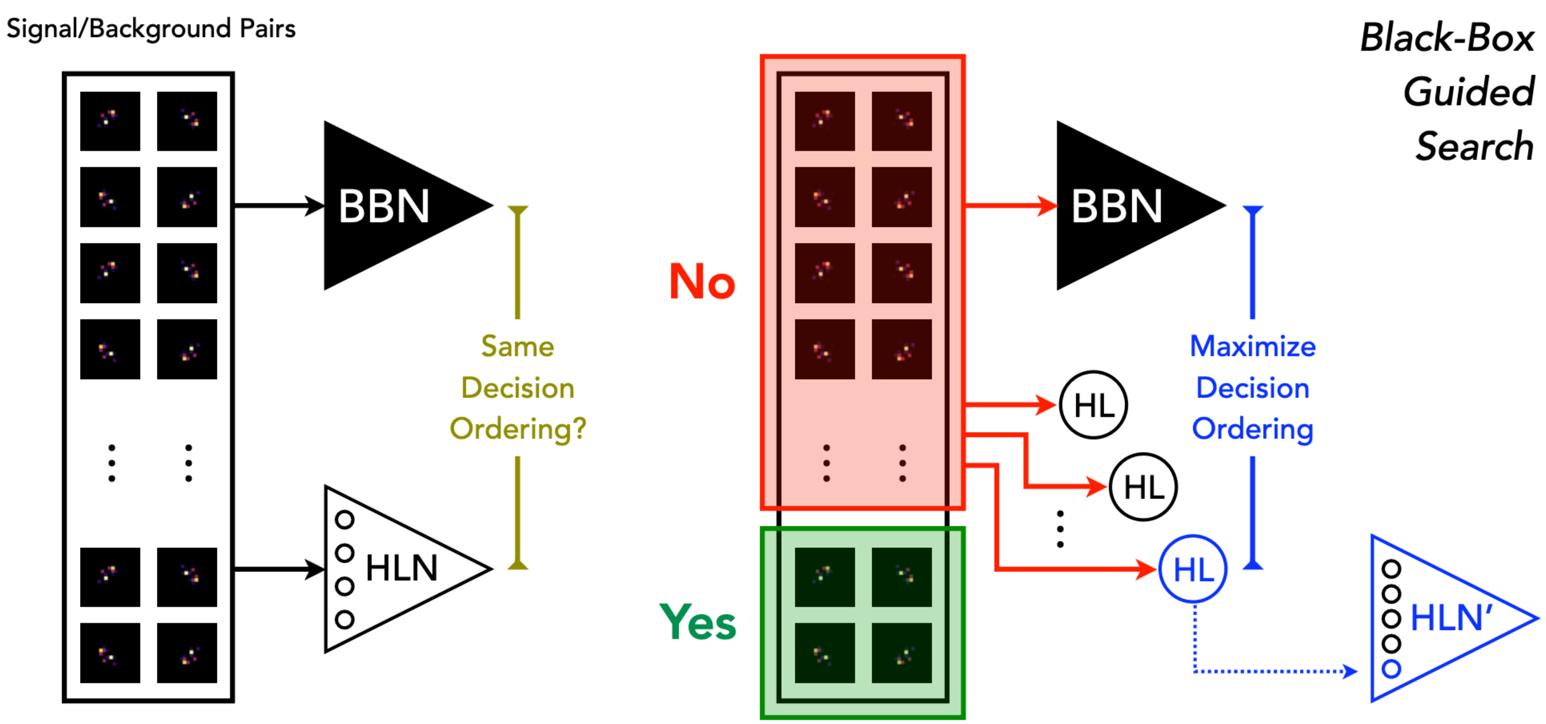
  - embedding physics laws (symmetries and/or theoretical constraints) [previous slides]

  - construct provably monotonic w.r.t. some features [2112.00038]

  - project metrics on already-identified physical observables

  - assemble a complete basis of interpretable observables and map the black box into that space [2010.11998]

Example [2010.11998]: Automated way to find a set of high level features [EFPs] that include all the information a CNN is implicitly learning from raw calorimeter images



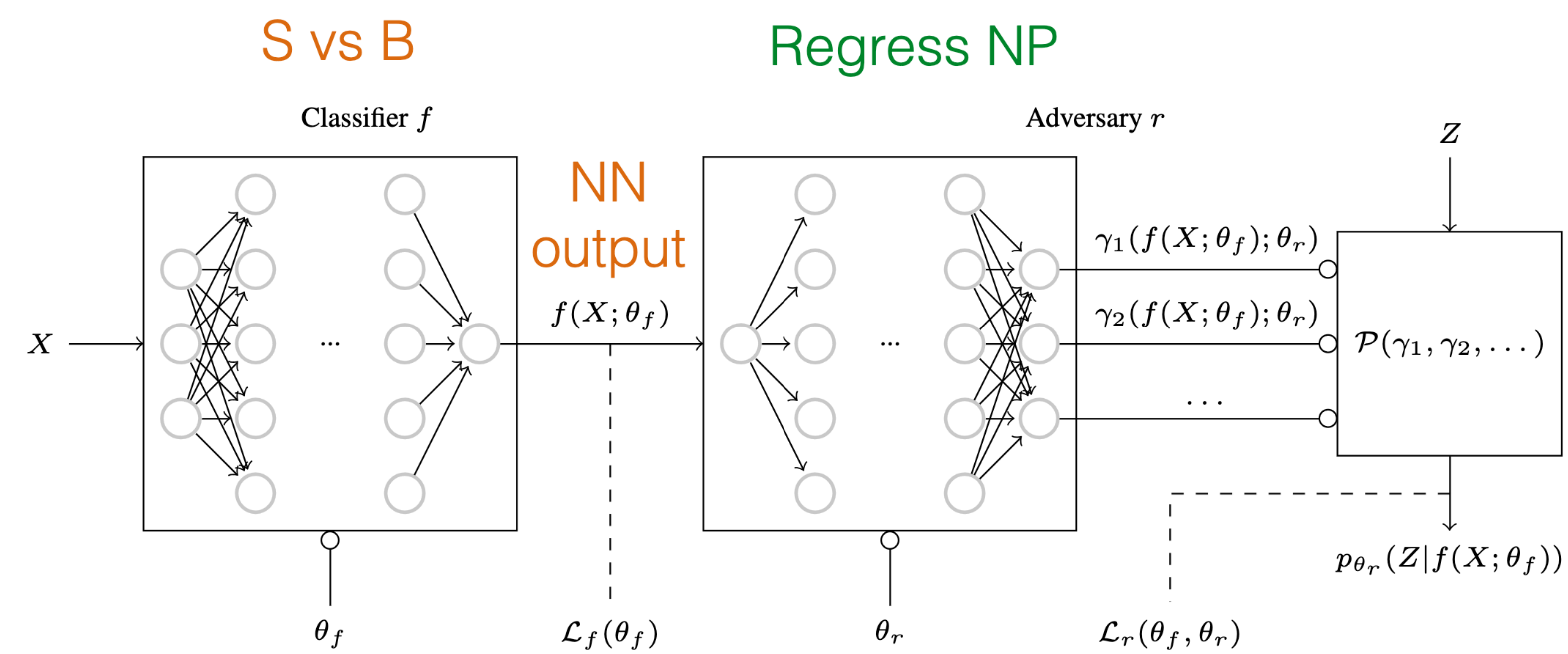
How to deal with uncertainties?

- Power of AI comes from finding subtle non-linear patterns in training data — this also makes it more susceptible to discrepancies between simulation and data
- A wide variety of techniques have been developed to quantify uncertainties as propagated through machine learning models
  - decorrelation
  - data augmentation
  - conditioning
- **Opportunity: ML unlocks completely new methods to tackle uncertainties in a way classical methods could not → back port to traditional algorithms?**
  - eg., more robust ML-based background estimation

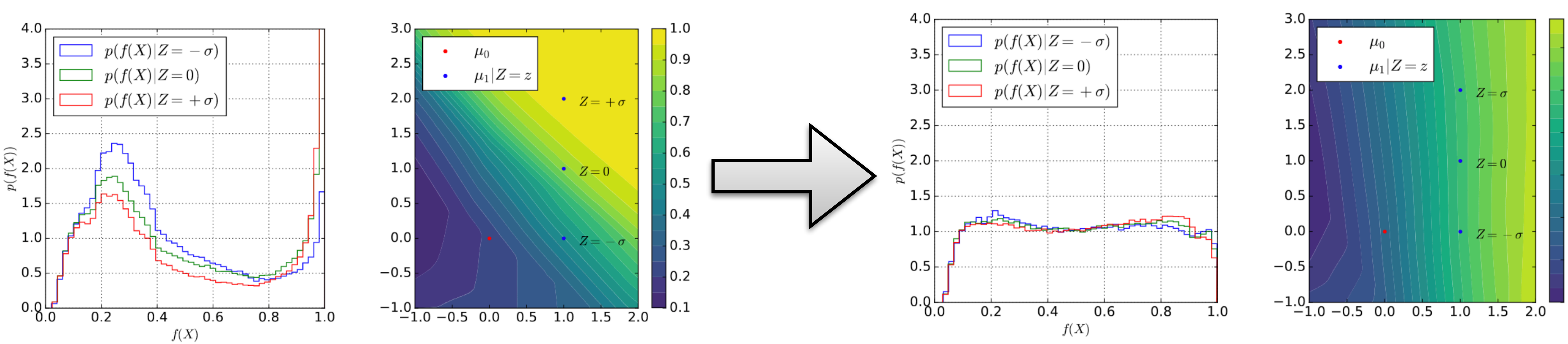
Adversarial decorrelation

Popular approach based on the idea of **decorrelating** from specific nuisance parameters *Z:*

e.g., through adversarial training



$$L_{Classifier} = L_{Classification} - \lambda \cdot L_{Adversary}$$



arXiv.1611.01046

ML for fast simulation

CERN-LHCC-2020-015

- HEP experiments rely heavily on simulations from experimental design all the way to data analysis
- Detector simulation (GEANT4) and event generation (MG5, Pythia, Herwig, …) are major and growing bottlenecks at LHC and other experiments

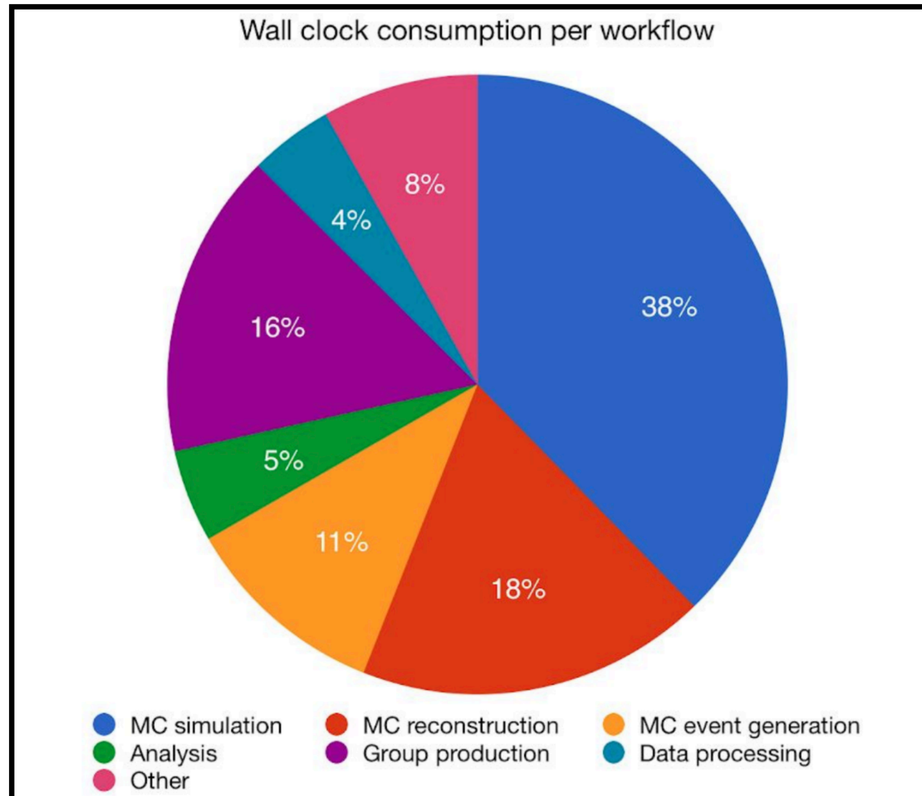


Figure 1: ATLAS CPU hours used by various activities in 2018

GEANT4 → 10^{10} events **SLOW but ACCURATE**

GEANT4 → 10^{5} events → Surrogate model → 10^{10} events

(GAN, VAE, Normalizing Flow, …)
Learn underlying distribution of GEANT4 events

FAST and ACCURATE?

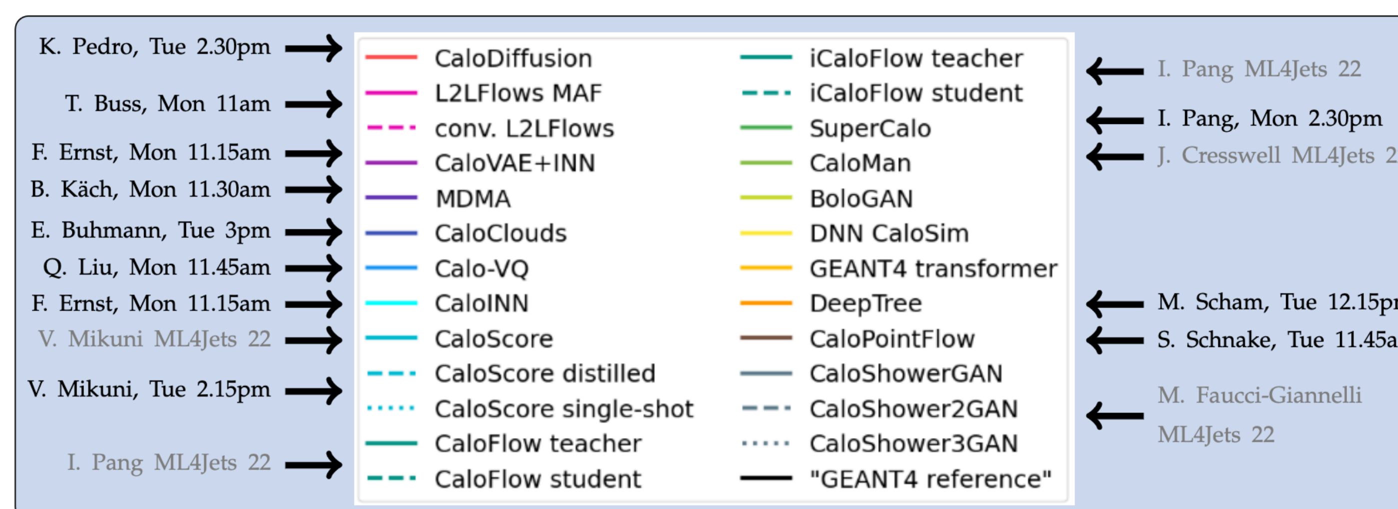
ML methods can provide fast and accurate “surrogate models” for GEANT4 etc

from D. Shih at Snowmass 2021 (Seattle)

Accelerating simulation with ML

- Many different approaches being explored but no one has emerged yet as the final solution:
  - Variational Autoencoders
  - Generative Adversarial Networks
  - Normalizing Flows
  - Diffusion models

- **Impressive effort to tackle major challenges of obtaining high-fidelity and computationally efficient models**

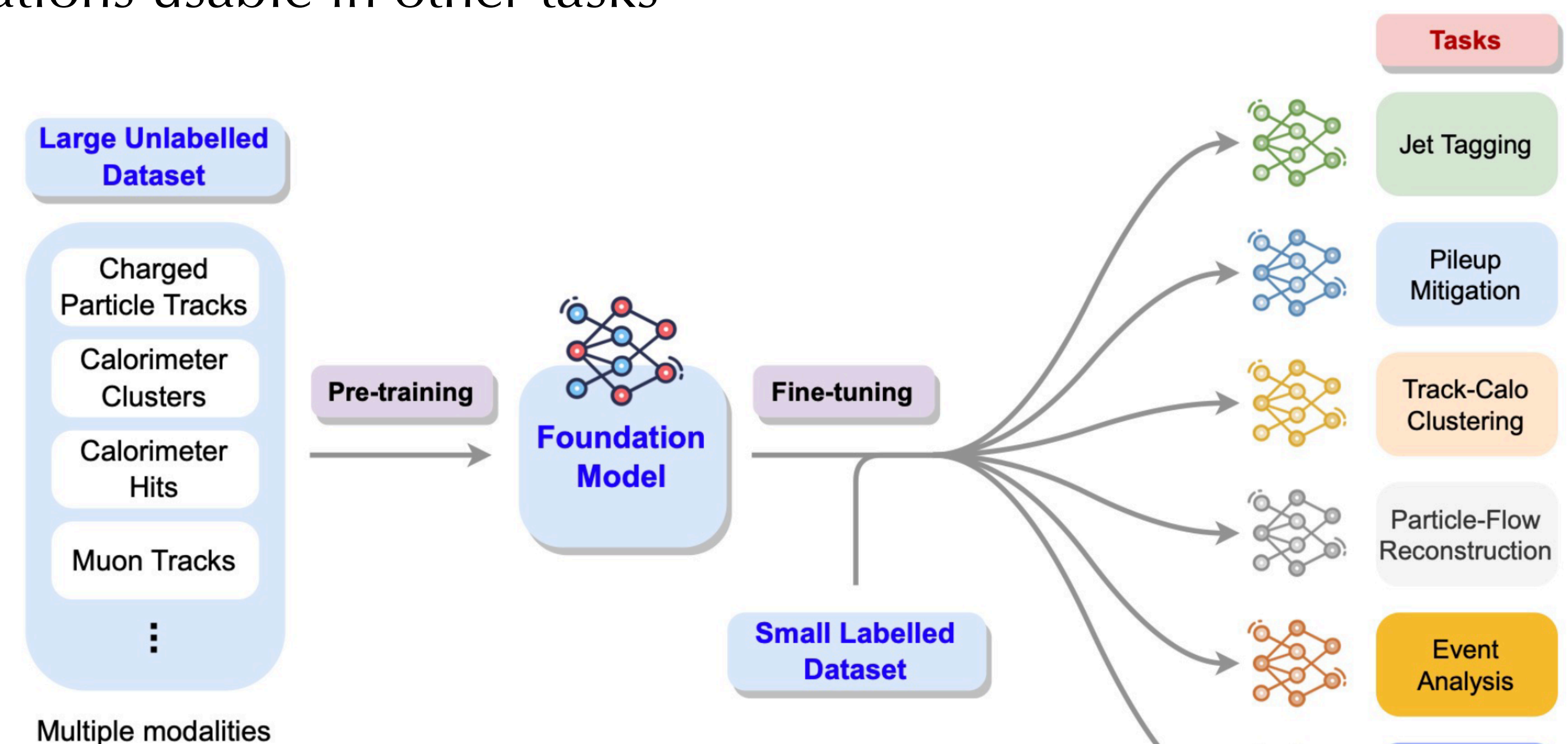


https://calochallenge.github.io/

See summary at ML4Jets 2023

Towards foundation models

- A foundation model is a large ML model trained on a vast quantity of data such that it can be adapted to a wide range of downstream tasks (e.g., BERT, GPT, …)
  - **self-supervised learning:** use the data itself to create training objective
  - **outputs:** powerful representations usable in other tasks



from M. Kagan at H&N 2023

reusable — one backbone used for several tasks

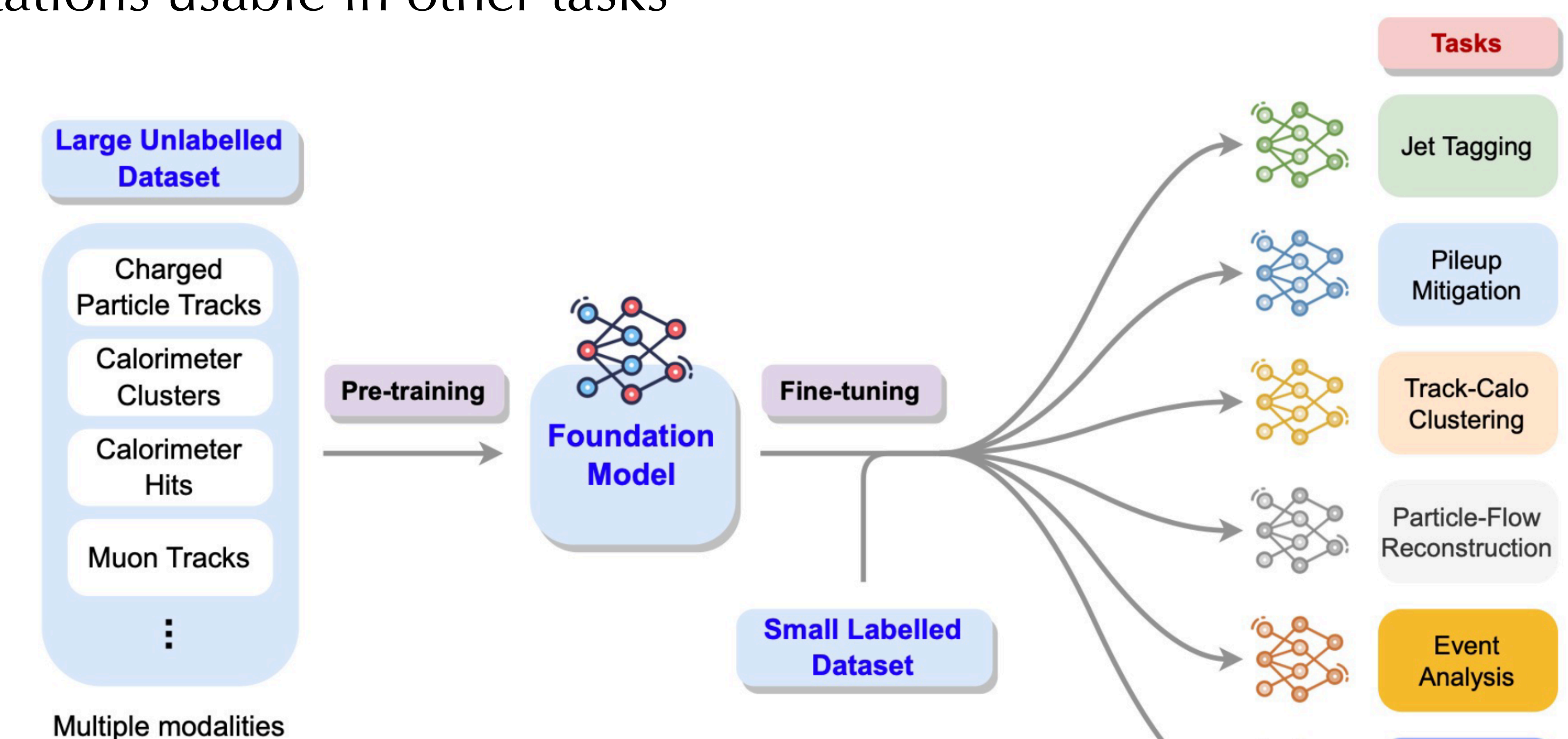
train on huge real data – leverage experimental data

leverage multi-modal methods – combine data from different detectors to address more complex tasks

uncertainty reduction – reduce dependence on simulation-based training

Towards foundation models

- A foundation model is a large ML model trained on a vast quantity of data such that it can be adapted to a wide range of downstream tasks (e.g., BERT, GPT, …)
  - **self-supervised learning:** use the data itself to create training objective
  - **outputs:** powerful representations usable in other tasks

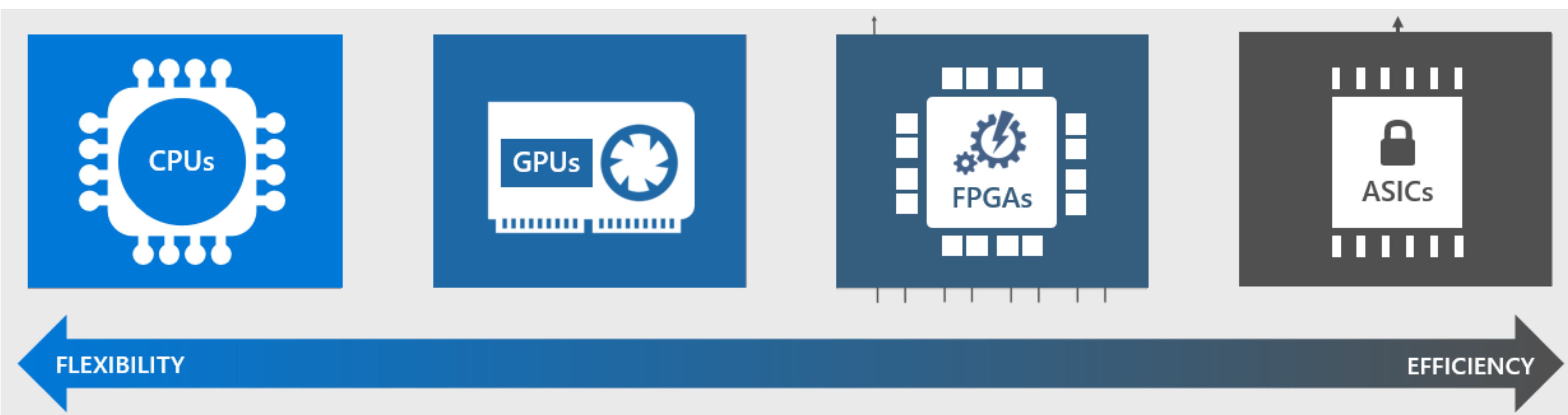


from M. Kagan at H&N 2023

Brings new challenges,
adaptation from language & vision not always direct,
lost of space to explore for development in HEP

Computing aspects

- Deploying AI in production for experiments requires increasingly computing resources
- CPU-based computing suboptimal with the growth of models and data complexity
- Need specialized hardware as well as AI-hardware/software empowered infrastructures and tools for computing (edge, local, and cloud)
  - the ability to exploit new generation AI hardware can really boost the exploration and adoption of ML-based solutions



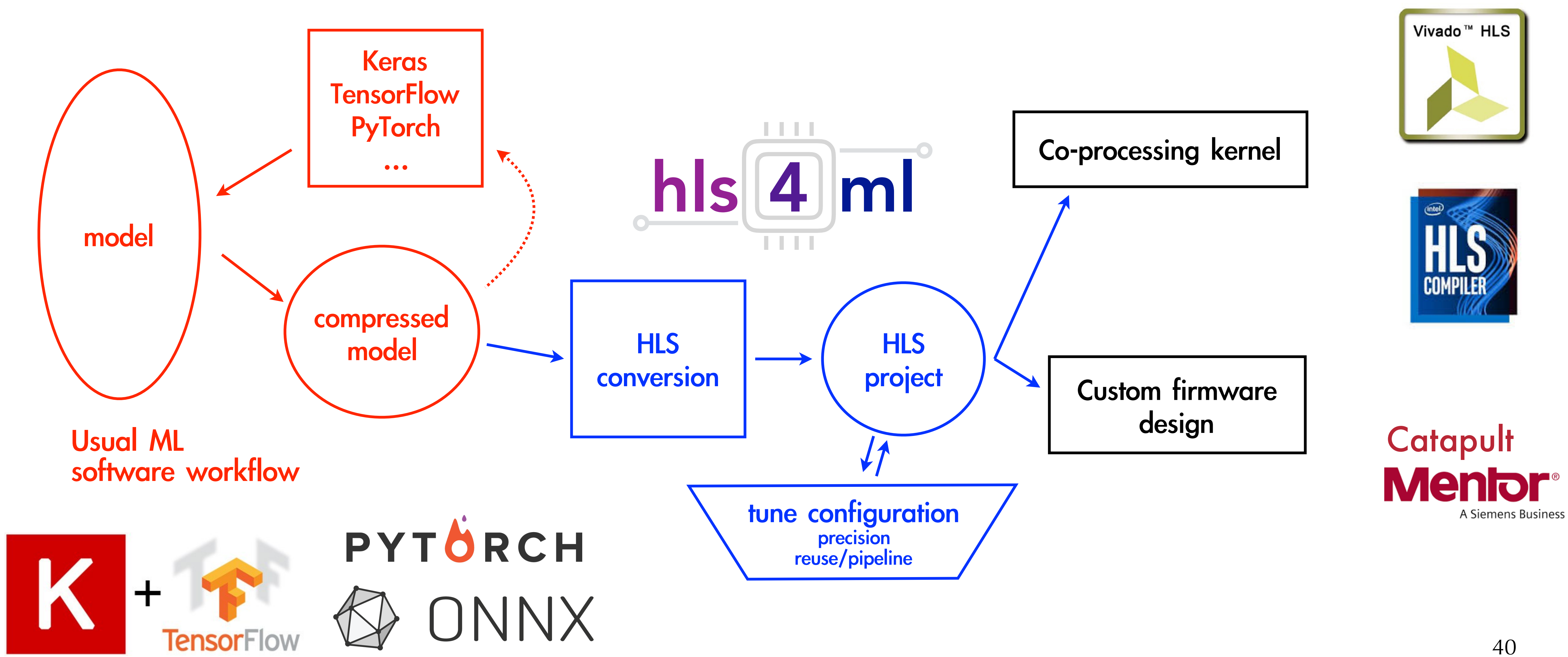
offline computing **real-time processing**

Bring ML models to hardware for real-time AI

high level synthesis for machine learning

A codesign tool to build ML models with hardware in mind and providing efficient platforms for programming the hardware.

Many use cases in HEP and beyond… and still growing!
(see Fast Machine Learning For Science Workshop Sep '23)



Real-time AI

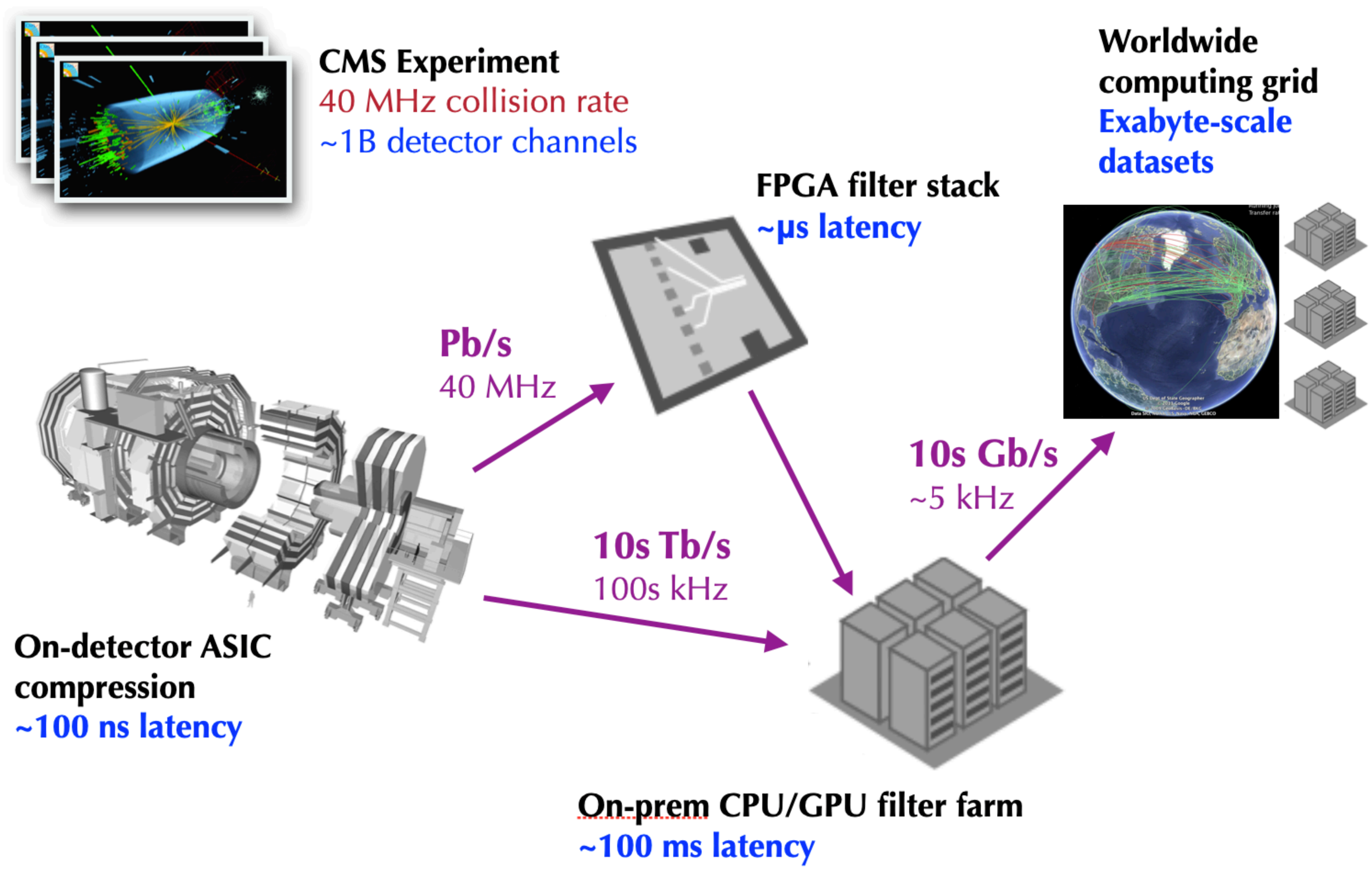
- Port ML to real-time data processing systems for **smarter data reduction**
  - eg., can do sophisticated event-level or particle-level reconstruction and identification already on hardware in first level of trigger systems
- Port ML to detector front-end electronics for **smarter data compression**

Huge input data rates up to PB/s

Latencies down to the ns

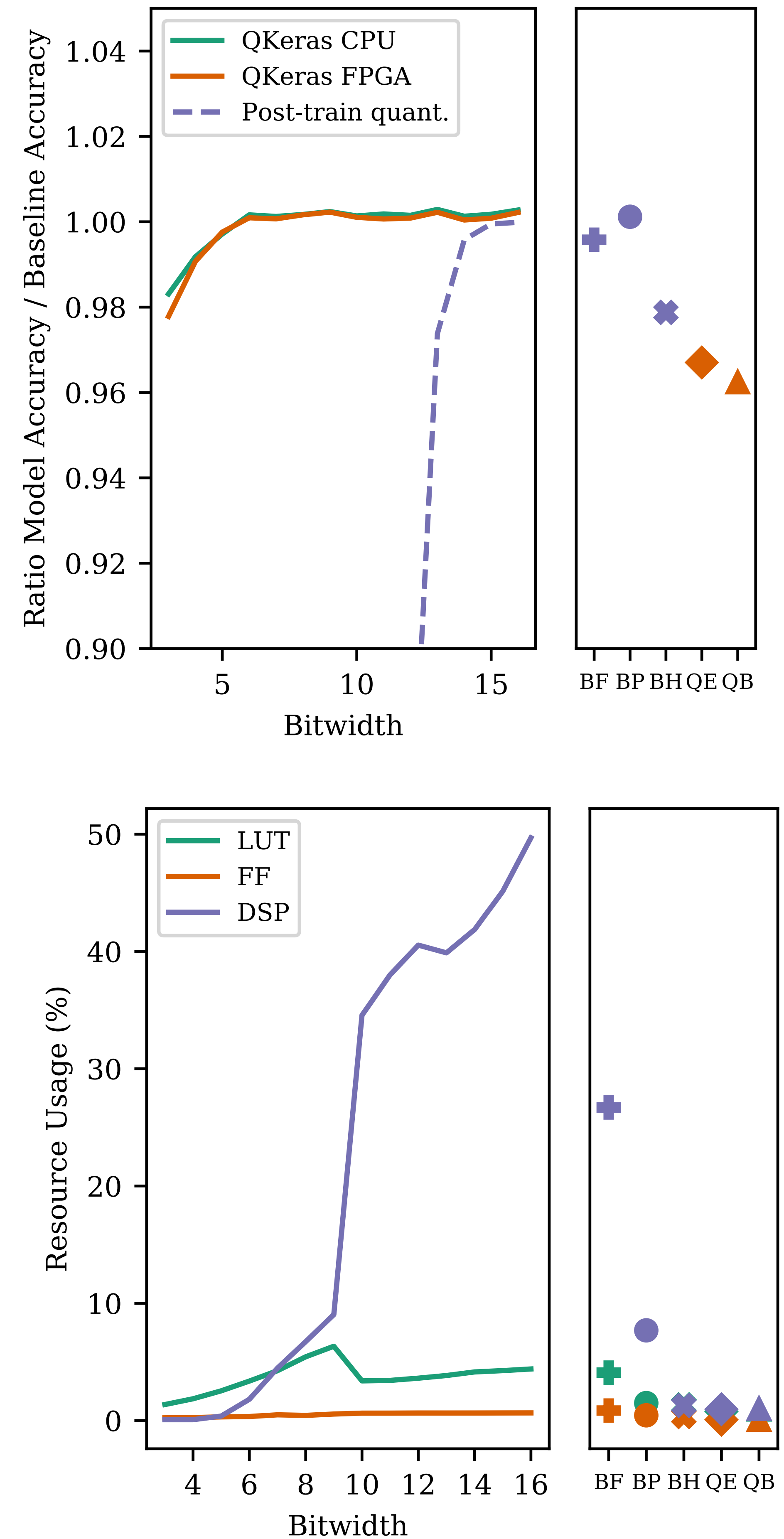
Low resources

Low power



Quantization-aware training

- Efficient hardware implementation uses reduced precision wrt floating point
- Post-training quantization can affect accuracy
  - for a given bit allocation, the loss minimum at floating-point precision might not be the minimum anymore
- One could specify quantization while look for the minimum
  - maximize accuracy for minimal FPGA resources
- Workflow: quantization-aware training with [Google QKeras](#) and firmware design with [hls4ml](#) for best NN inference on FPGA performance



C. N. Coelho et al.: Nature Machine Intelligence, Volume 3 (2021)

Summary and outlook

- **Machine Learning has developed into a powerful set of statistical methods** that have influenced nearly every aspect of high-energy physics
  - while covering only a subset of applications in this talk the main message is that **with ML we can do more with less**
  - new approaches and applications, new human-unknown patterns, more automatization and thus accelerated time to discovery!
- **Our unique challenges need new expertise, knowledge, and resources**
- ***How do we capture 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