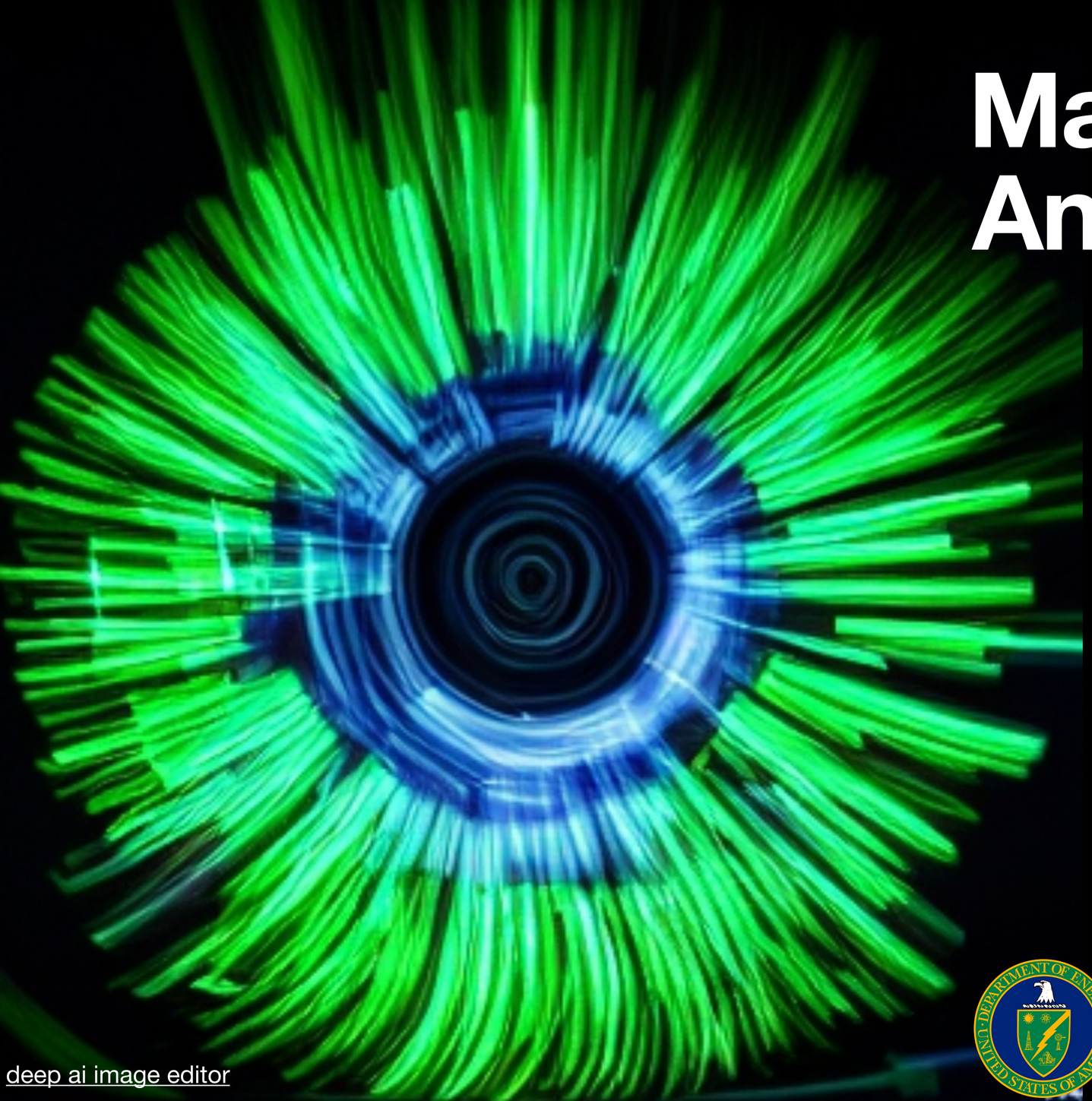


Machine Learning in Analysis Techniques for future physics

Hannah Bossi (MIT)

Quark Matter 2025
Frankfurt, Germany
April 12th, 2025

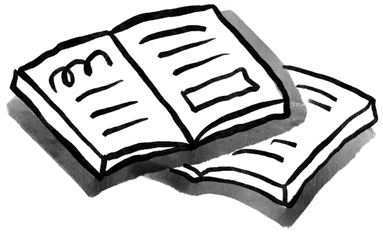
MIT HIG's work was supported by US DOE-NP

Laboratory for Nuclear Science

deep ai image editor

ROADMAP

WHAT IS ML AND WHY IS IT USEFUL FOR PHYSICS ANALYSIS?

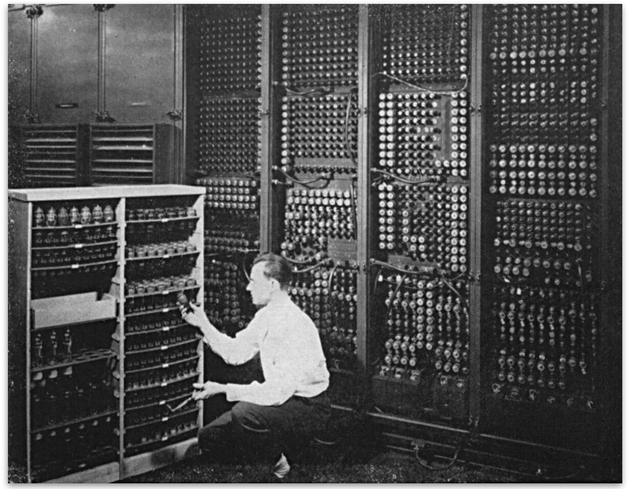


HOW IS ML CURRENTLY BEING USED AS AN ANALYSIS TECHNIQUE?

WHAT DOES THE FUTURE HOLD?

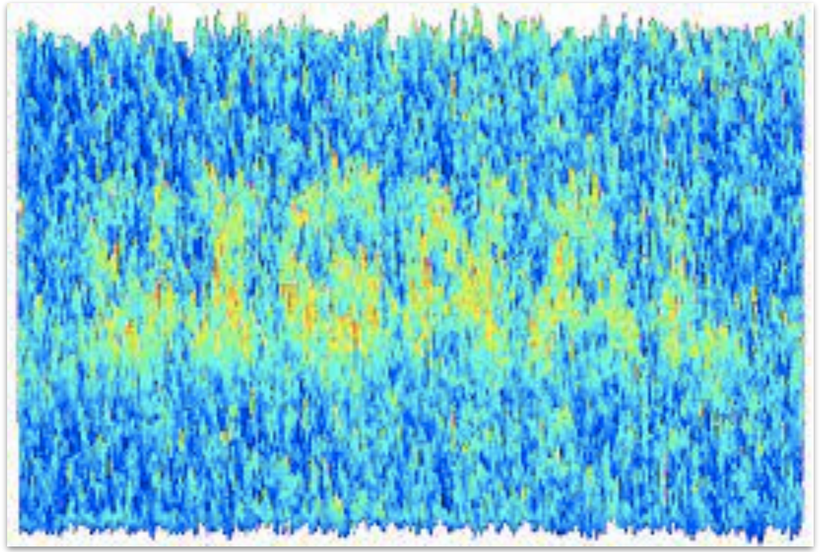
**Focus on new items since QM 2023 or items presented at QM 2025*

ORIGINS OF ML



[Electronic Numerical Integrator and Computer, 1945]

- First electronic computers developed with the purpose of performing **computations that were difficult and time-consuming for humans** to perform accurately.

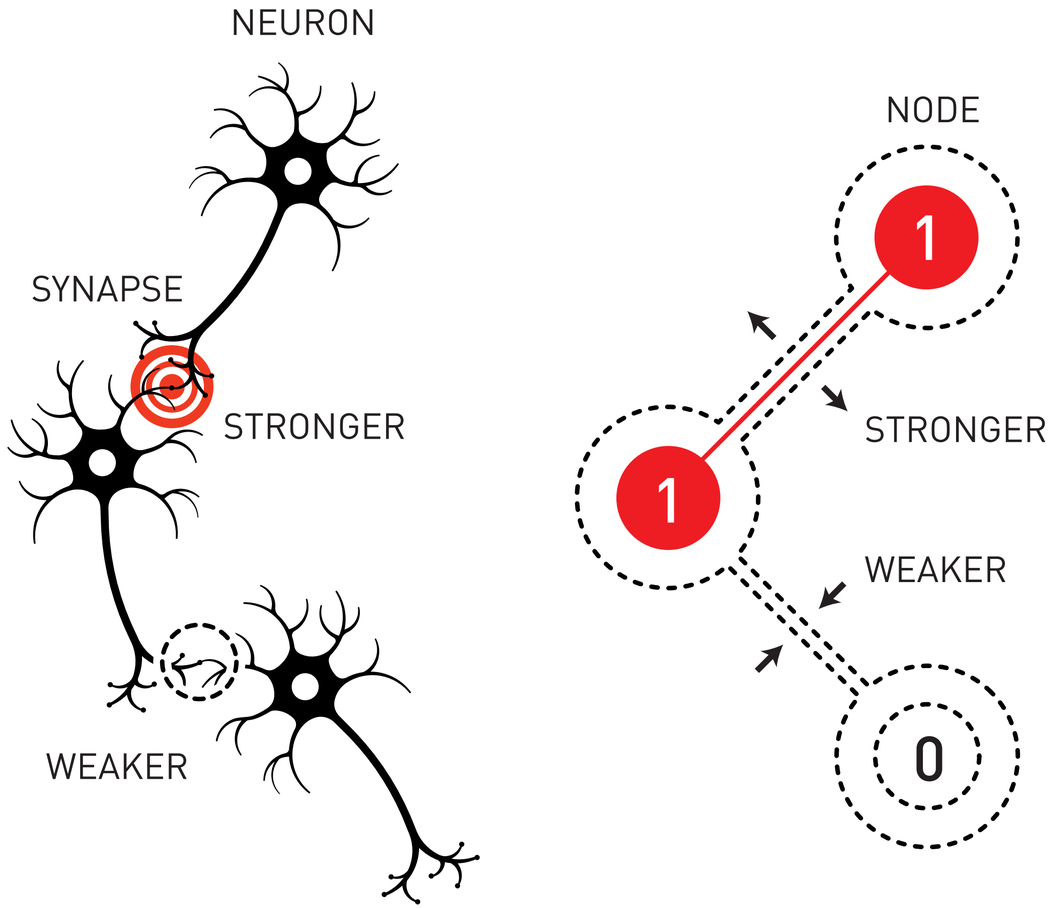


- Later push for computers to **replicate computations humans are naturally good at, i.e. pattern recognition**, but on larger datasets.

As a result, created algorithms that imitate how humans learn, i.e. learning from examples in a way that improves over time → **machine learning**

FOUNDATIONAL TOOLS

- Fundamental building blocks of machine learning directly mimic the brain!



Neurons → Nodes

Synapses → Weighted couplings

Learning → training

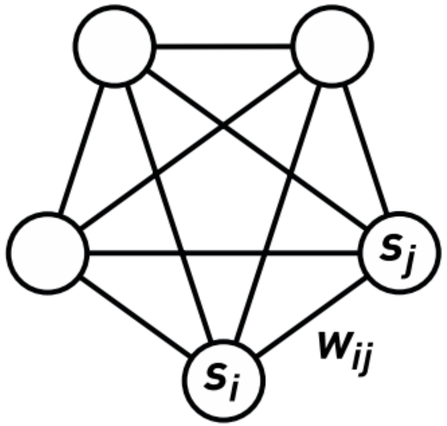
- Connections between some neurons (synapses) become stronger and others weaker as we learn.

[D.O Hebb, "The organization of behavior", 1949]

[Press release Nobel Prize in Physics 2024]

FOUNDATIONAL TOOLS

Collections of nodes give rise to computational collective behavior. - artificial neural networks

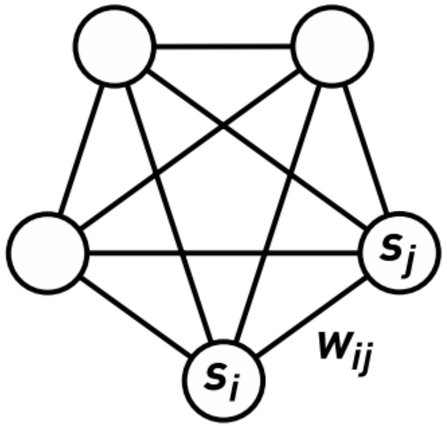


- Hopfield network: binary nodes ($s_i = 0/1$), information read in/read out from all nodes.

[Press release Nobel Prize in Physics 2024]

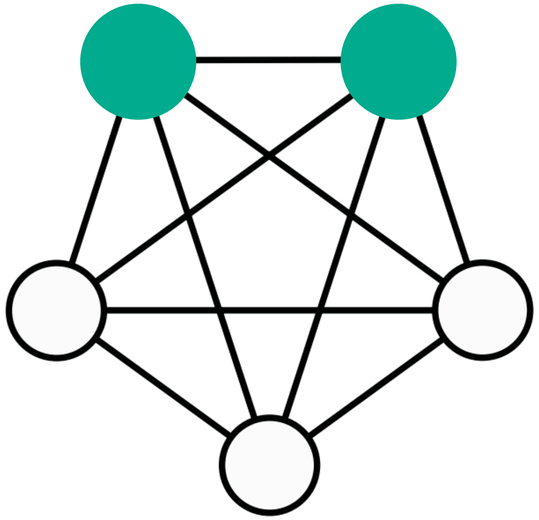
FOUNDATIONAL TOOLS

Collections of nodes give rise to computational collective behavior. - artificial neural networks



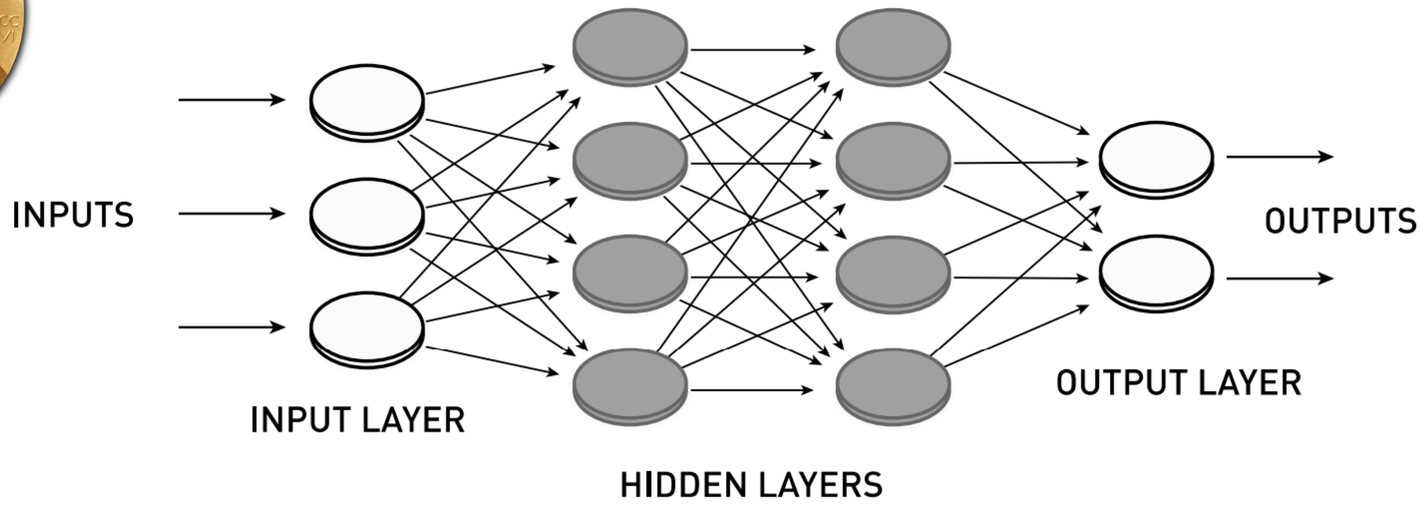
- Hopfield network: binary nodes ($s_i = 0/1$), information read in/read out from all nodes.

- Boltzmann machine: Nodes are subdivided into visible nodes and hidden nodes.



[Press release Nobel Prize in Physics 2024]

FOUNDATIONAL TOOLS



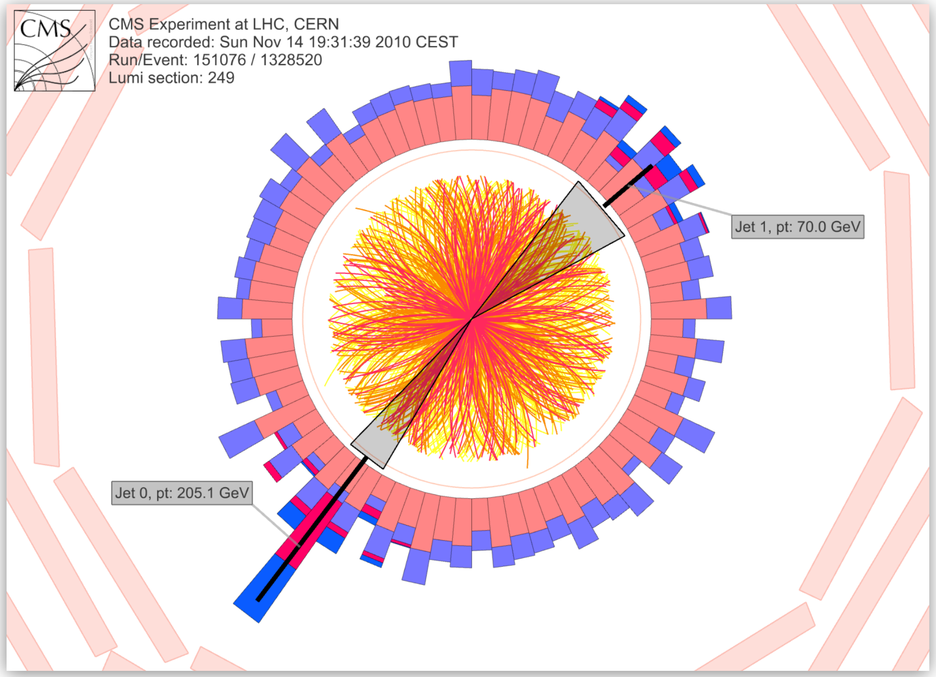
John Hopfield
(Hopfield Network)

Geoffrey Hinton
(Boltzmann Machine)

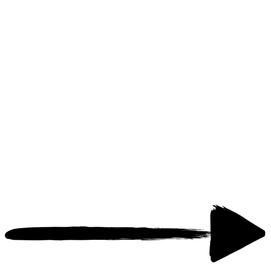
[Press release Nobel Prize in Physics 2024]

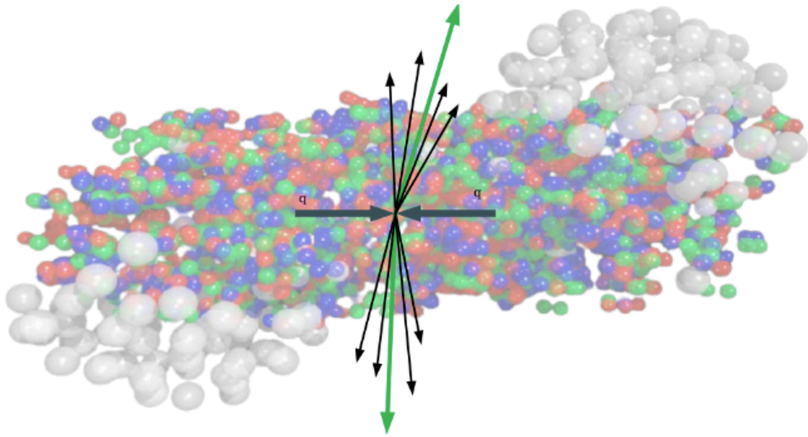
ML AND PHYSICS

Goal of measurements: *To extract relevant physics information from available data!*



[CMS observes jet quenching]



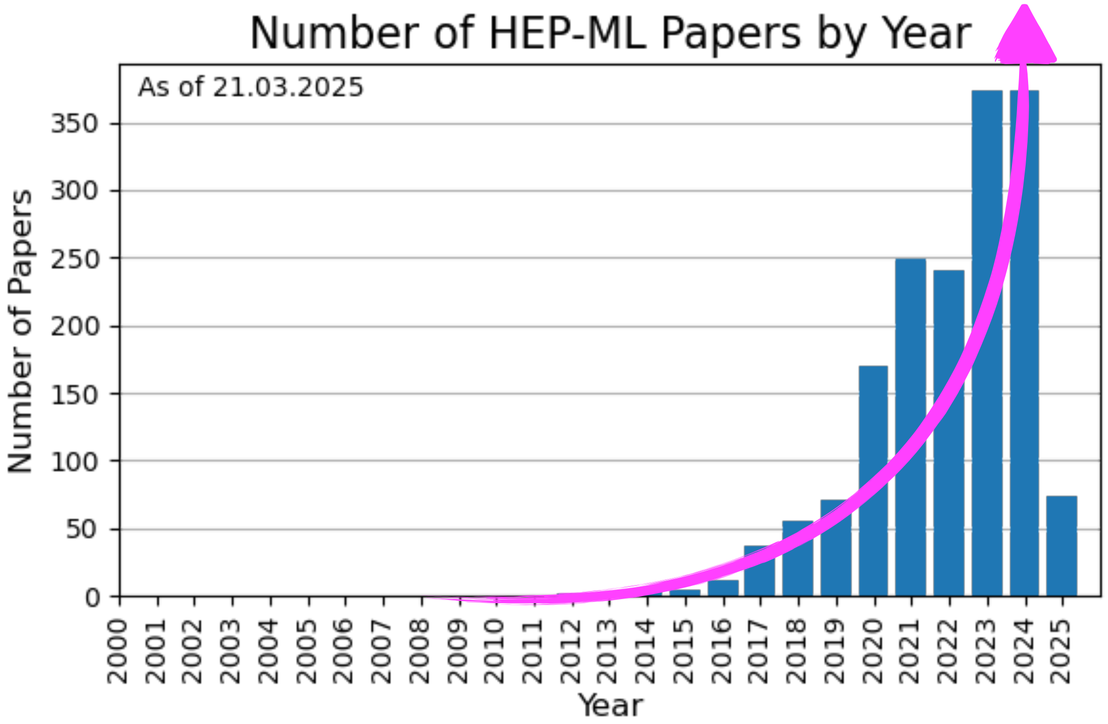


Quark-Gluon Plasma (QGP)

Historical approach: (1) make selection using a series of boolean decisions motivated by physics/experimental constraints (2) perform a statistical analysis on selected data.

What do we do when decision becomes difficult to derive from expert-knowledge alone?

ML ON THE RISE



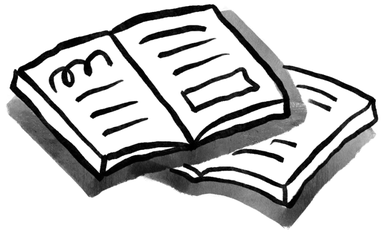
[HEPML-Living Review]

- Increase in computing power and collection of large datasets at modern facilities have lead to increased use of ML for physics.

In the coming decades, ML will help usher high energy physics into it's next era! (More on this later)

ROADMAP

WHAT IS ML AND WHY IS IT USEFUL FOR PHYSICS ANALYSIS?



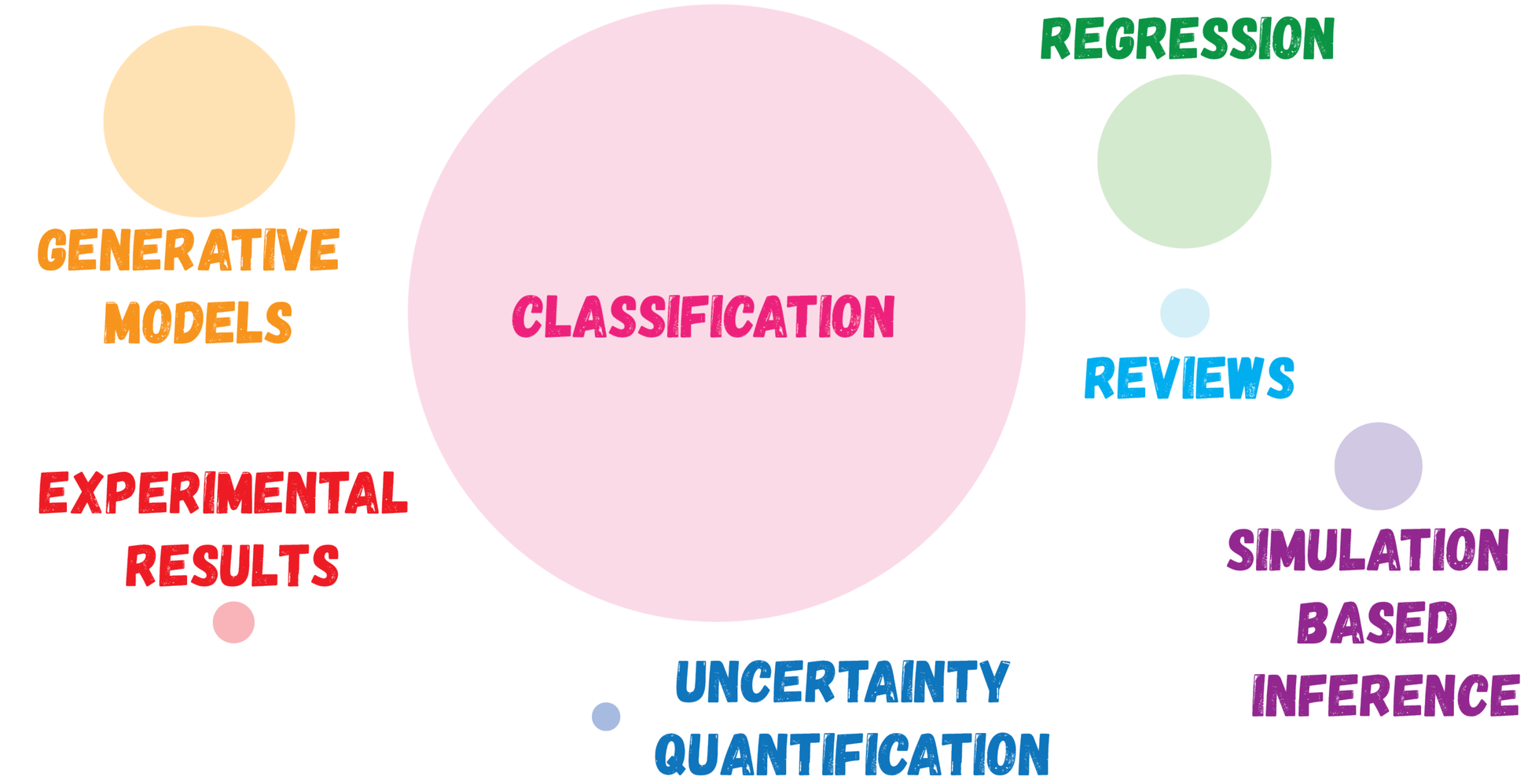
HOW IS ML CURRENTLY BEING USED AS AN ANALYSIS TECHNIQUE?

WHAT DOES THE FUTURE HOLD?

*Focus on new items since QM 2023 or items being presented at QM 2025

HOW IS ML USED FOR PHYSICS?

[HEPML-Living Review]

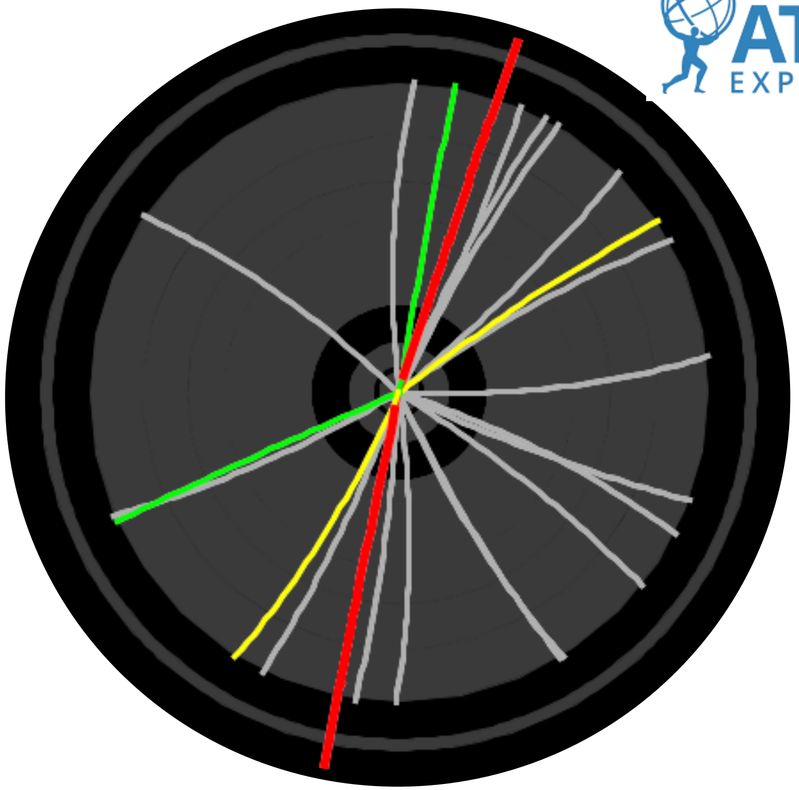


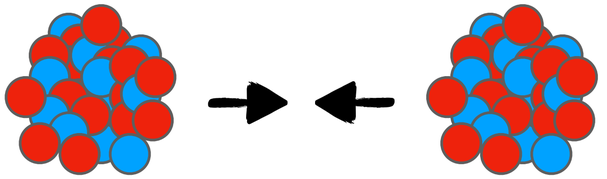
Radius of circle = # of papers/4

** Note in this scheme papers are allowed to count in more than one category.*

ML IN HEAVY IONS

- Subset ML applications for heavy-ion collisions has some additional challenges.

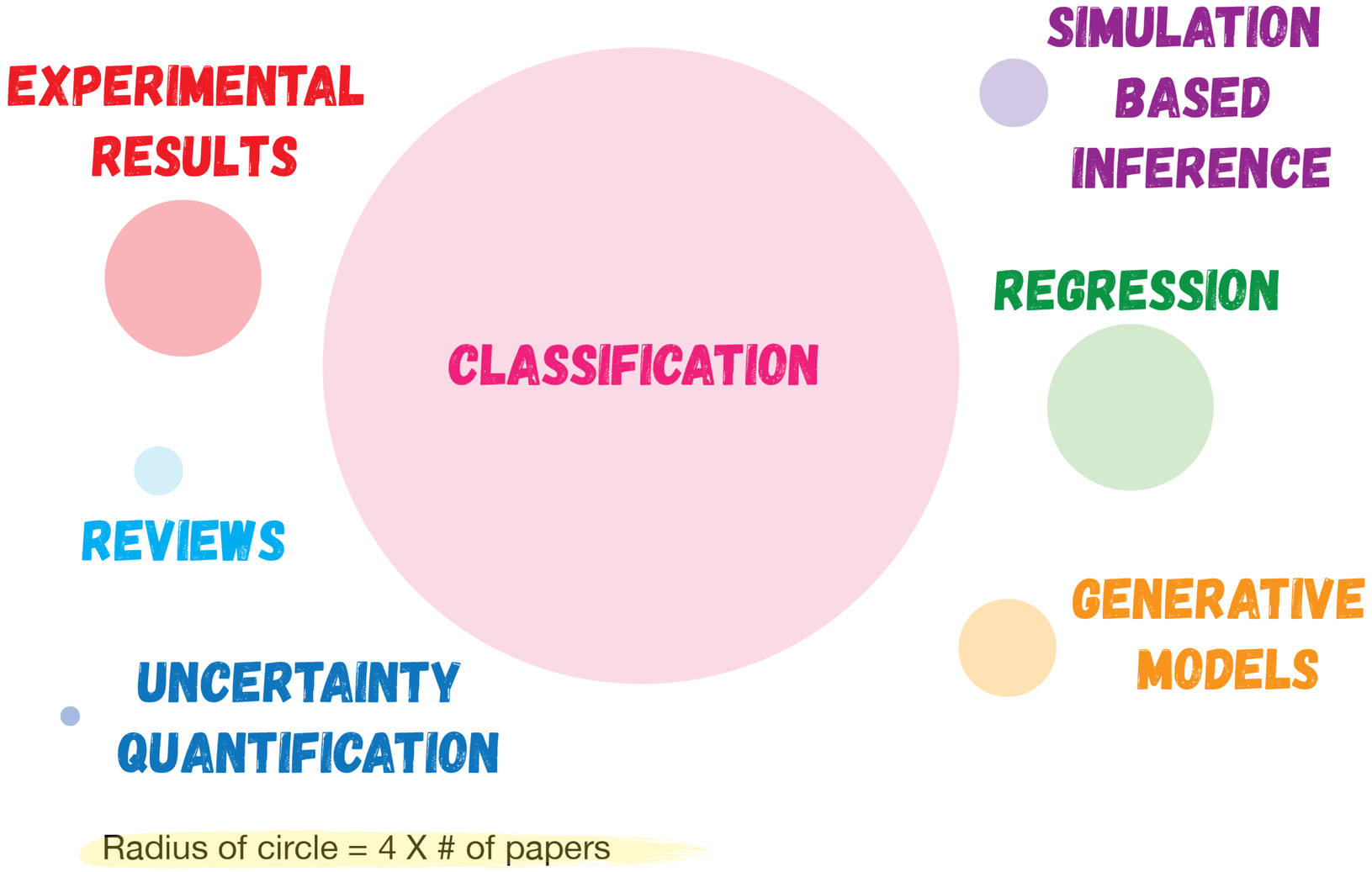




- Higher particle multiplicities, higher degree of complexity.
- Large variance in simulations of heavy-ion collisions causes high dependence on sample used in training.

ML IN HEAVY IONS

[HEPML-Living Review]



- Investigate paper subset within HEP that specifically relates to HIs or is a technique readily employed in HIs.
- Similar distribution between the different subtopics as HEP overall.

** Note experimental results include those performed with TMVA*
** Note in this scheme papers are allowed to count in more than one category.*

FIRST: WHAT CAN ML NOT DO?

Garbage In

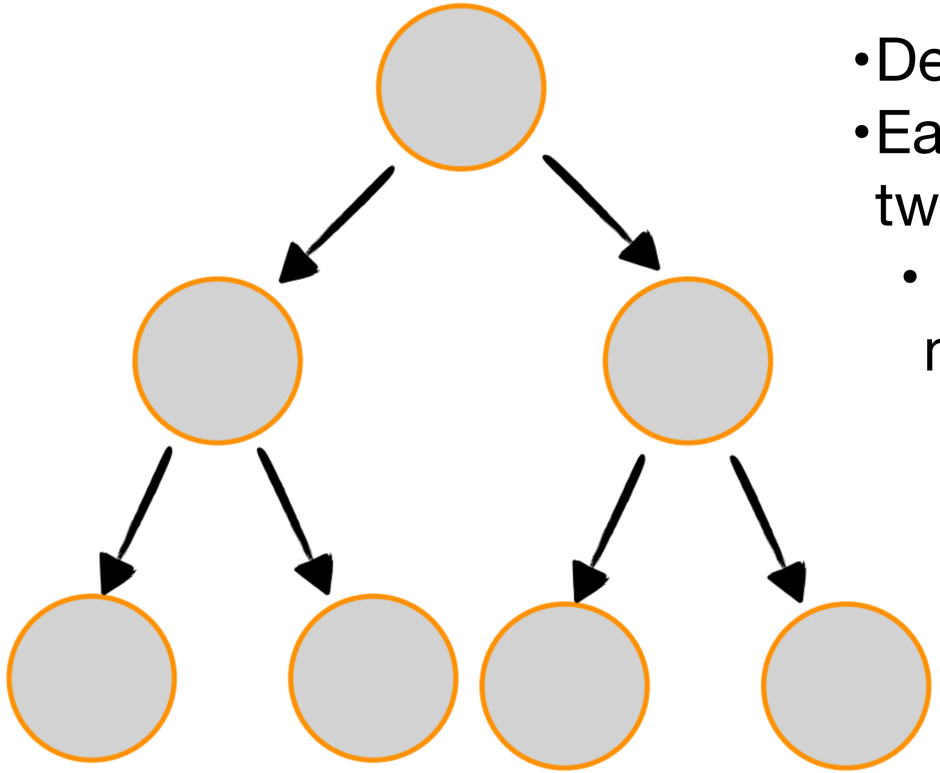
Garbage Out

- ML cannot replace domain knowledge.
- ML is not a causation tool.
- **ML is not a magic fix!**

SIGNAL/BACKGROUND CLASSIFICATION

CLASSIFICATION

- Large amount of experimental applications of machine learning rely on signal vs. background classification.



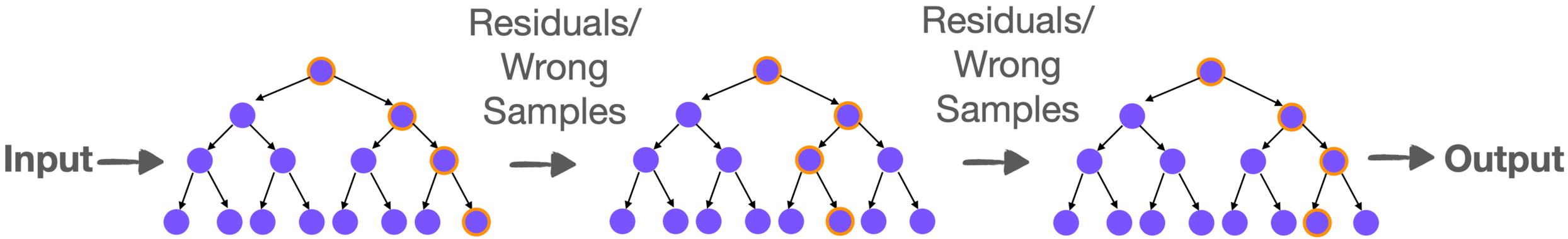
- Decision trees are commonly used for signal classification
- Each node is a classification rule that splits the data into two or more parts.
  - In training you determine the proper rules that maximizes the information gain and minimize entropy

$$E(x) = \sum -p(x)log_2(p(x))$$

SIGNAL/BACKGROUND DISCRIMINATION

CLASSIFICATION

- Boosted decision trees are used when multiple weaker learners are combined in a series where each additional component seeks to minimize error of previous one.



Yann Coadou, arXiv: 2206.09645

HEAVY FLAVOR JET TAGGING

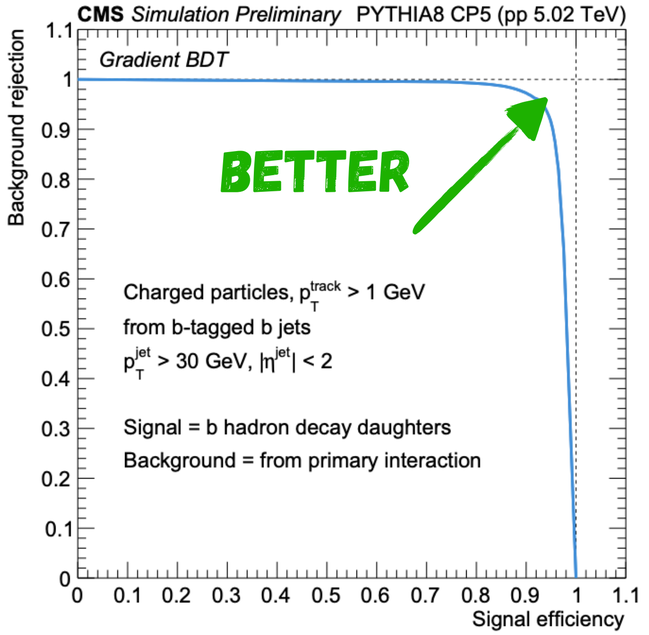
CLASSIFICATION

Goal: identify jets initiated by a heavy-quark

Conventional approach: Apply cuts to select jets with displaced decay vertices and large impact parameter tracks.

ML approach: Learn from low-level features in a supervised approach using BDT or a GNN

CMS-PAS-HIN-24-005



QUARK MATTER 2025 — L. KALIPOLITI POSTER

QUARK MATTER 2025 — C. CHOI POSTER

QUARK MATTER 2025 — Z. XU POSTER

[JINST 13 (2018) 05, P05011]

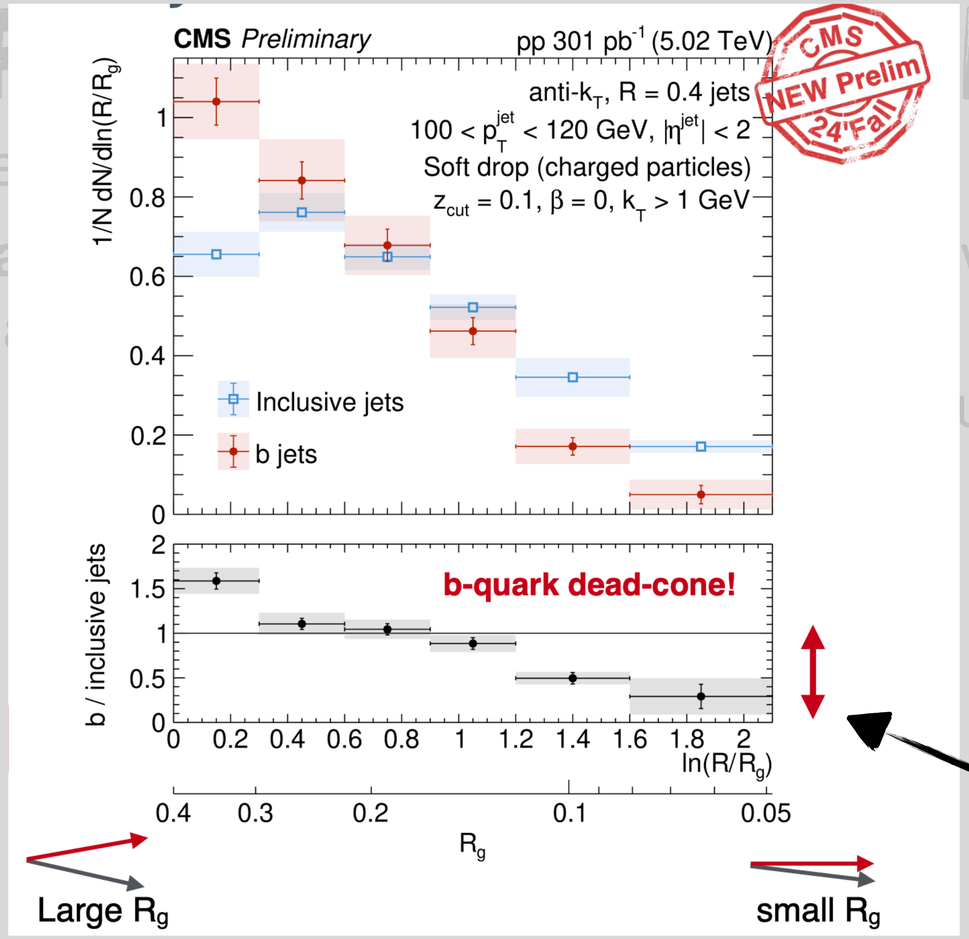
HEAVY F… …ING

Goal: identify je…

Conventional a… …with displaced decay vertices …

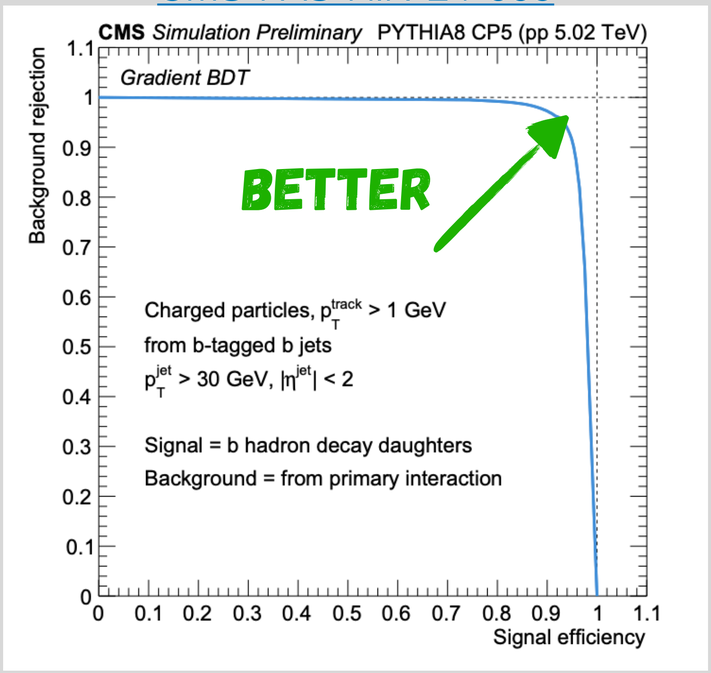
ML approach: …upervised approach using BDT or a GNN

[JINST 13 (2018) 05, P05011]



ENABLES THE FIRST OBSERVATION OF B-QUARK DEAD CONE!

CMS-PAS-HIN-24-005



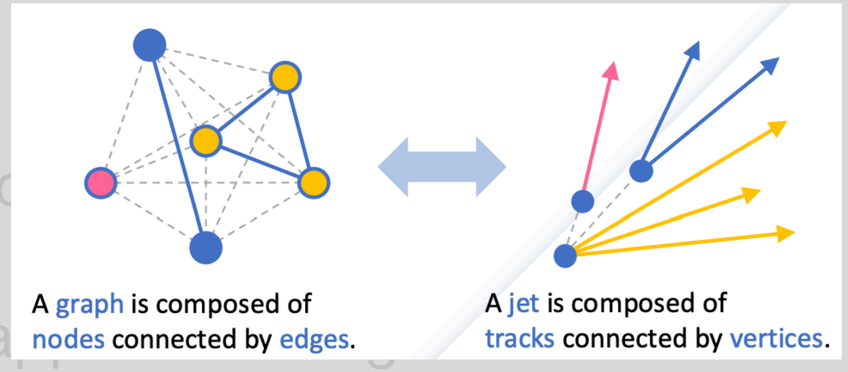
QUARK MATTER 2025
L. KALIPOLITI POSTER

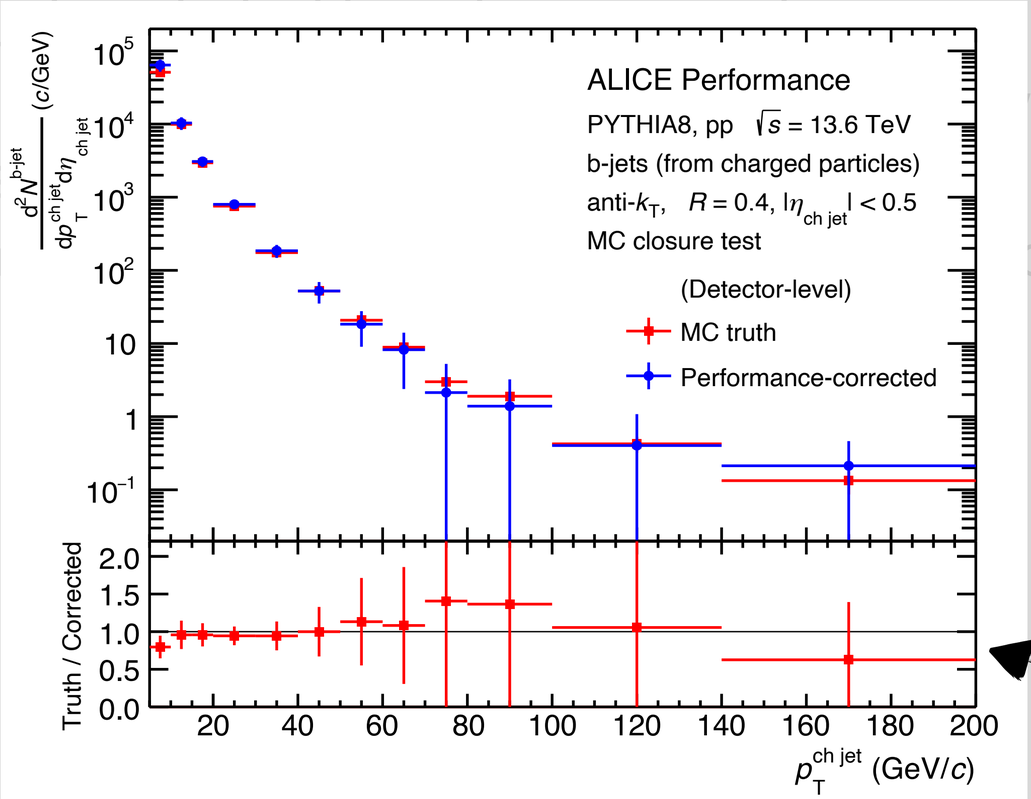
QUARK MATTER 2025
C. CHOI POSTER

QUARK MATTER 2025
Z. XU POSTER

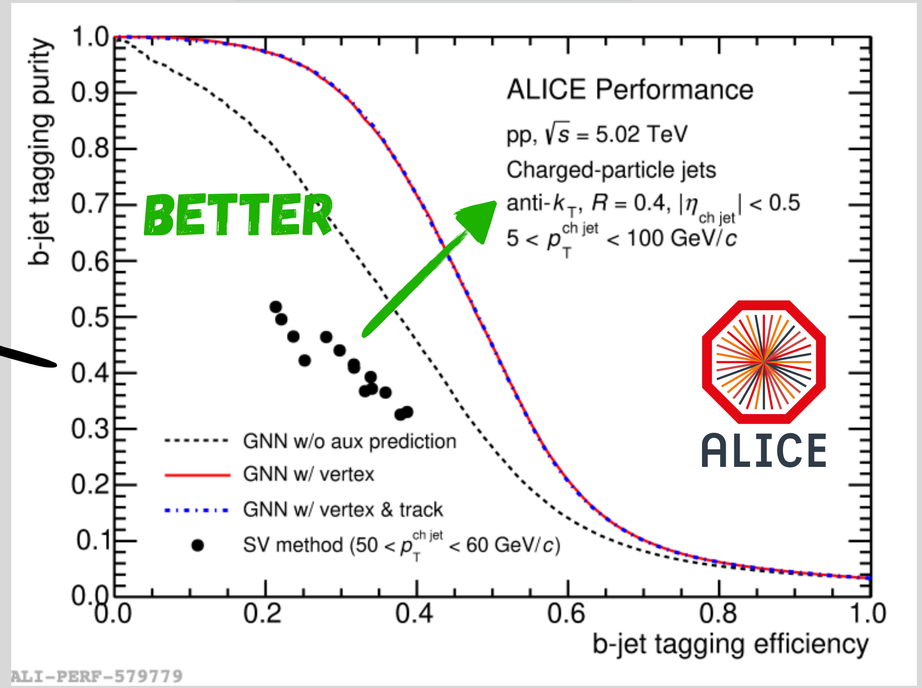
ENABLES IMPROVED MEASUREMENT OF THE B-JET CROSS SECTION IN ALICE!!

[JINST 13 (2018) 05, P05011]





CMS-PAS-HIN-24-005



QUARK MATTER 2025
L. KALIPOLITI POSTER

QUARK MATTER 2025
C. CHOI POSTER

QUARK MATTER 2025
Z. XU POSTER

HEAVY FLAVOR JET TAGGING

[JINST 13 (2018) 05, P05011]

Goal: identify jets initiated by a heavy-quark

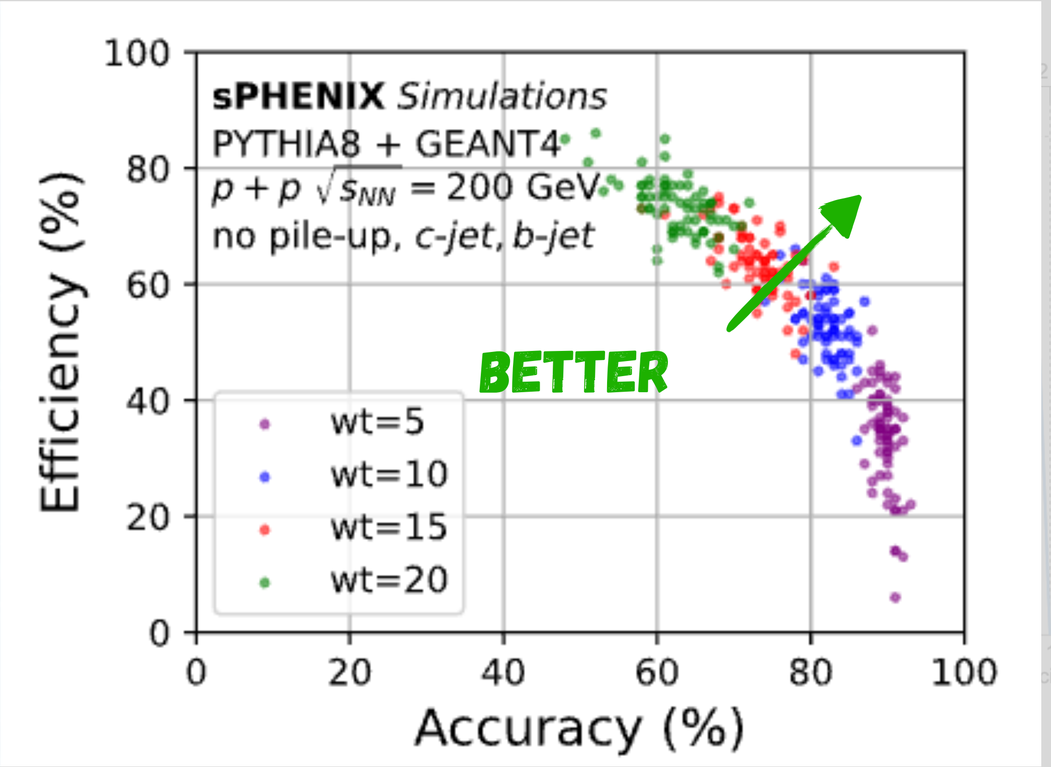
Conventional approach: select jets with displaced decay vertices and large number of charged tracks

ML approach: Learn from low-level features in a supervised approach using BDT or a GNN

- Use "long short-term memory (LTSM)" network that allows important information to be retained over long sequences.

- Training variables:
  - Jet variables
  - Jet constitutent variables
  - Event variables

ML METHOD SHOWS IMPROVED HF JET TAGGING EFFICIENCY OVER TRADITIONAL APPROACH!



QUARK MATTER 2025
L. KALIPOLITI
POSTER

QUARK MATTER 2025
C. CHOI
POSTER

QUARK MATTER 2025
Z. XU
POSTER

REGRESSION OF JET PT

- ***Differential measurements of jets are key to understanding jet quenching!***
- These often involve pushing to large *R* and/or low p_{T}, where background contribution is difficult to subtract.

By now many methods in which ML can be used to solve this problem! We will discuss one.

See also…..

- Shallow NN trained on PYTHIA + HI Background

  [PRC 99, 064904 (2019)] ALICE, [PLB 849 (2024) 138412]

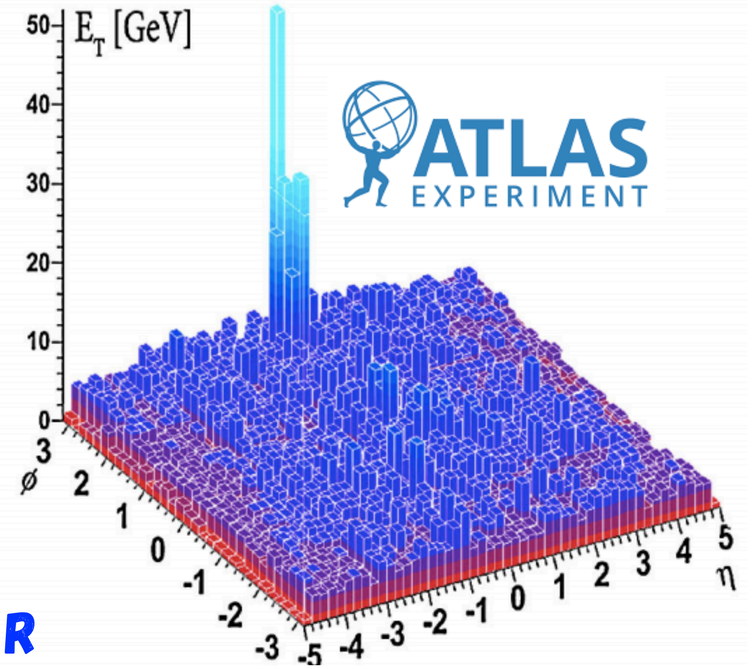
- CNN for regression of energy loss fraction

  [JHEP 03 (2021) 206]

- Interpretable Deep NN

  [Phys.Rev.C 108 (2023) 2, L021901]

ENABLED LARGE R JET MEASUREMENT AT ALICE!

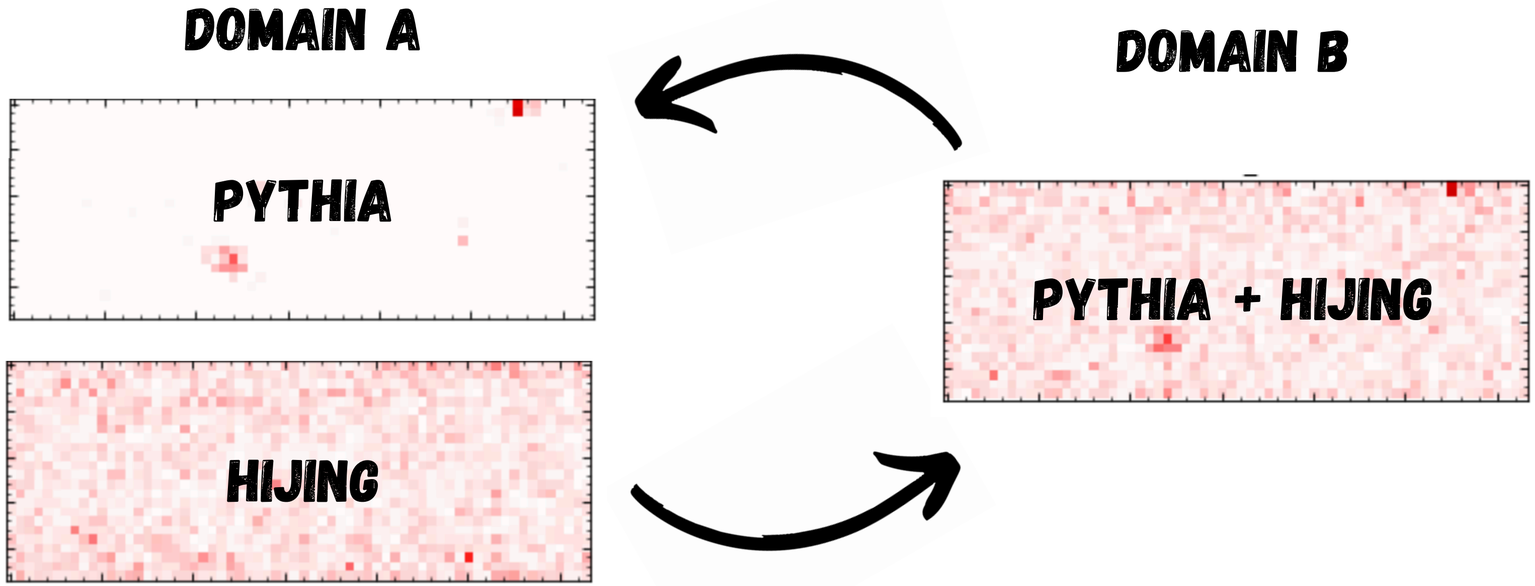


[ATLAS, PRL 105, 252303 (2010)]

JET PT RECONSTRUCTION

REGRESSION

- Use generative AI (unpaired image-to-image translation, cycleGANs) to subtract jet background in an ***unsupervised*** way. [UVCGAN: arXiv:2203.02557]
  - Composed of two generator-discriminator pairs w/ cyclic closure (i.e. $A \rightarrow B \rightarrow A \sim A$)
    - One to translate from domain A $\rightarrow$ B
    - One to translate from domain B $\rightarrow A$



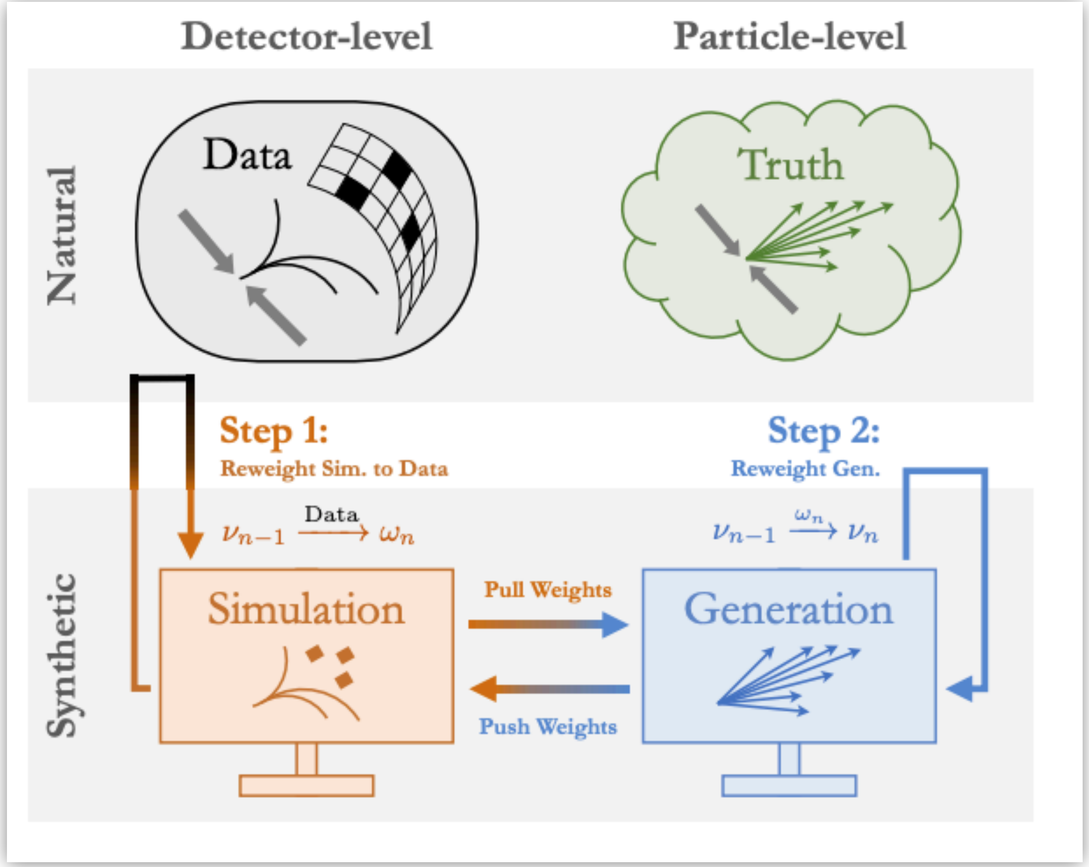
QUARK MATTER 2025

Y. GO

POSTER

UNFOLDING WITH ML

SIMULATION BASED INFERENCE



[PRL 124, 182001 (2020)]

Goal of unfolding: Correct measured distributions for smearing.

Conventional Approach: Apply unfolding procedure on a binned distribution and repeat for each observable.

ML-based Approach: Use ML to calculate weighting factors and unfold the phase space all at once, before the choice of binning or observable!

QUARK MATTER 2025

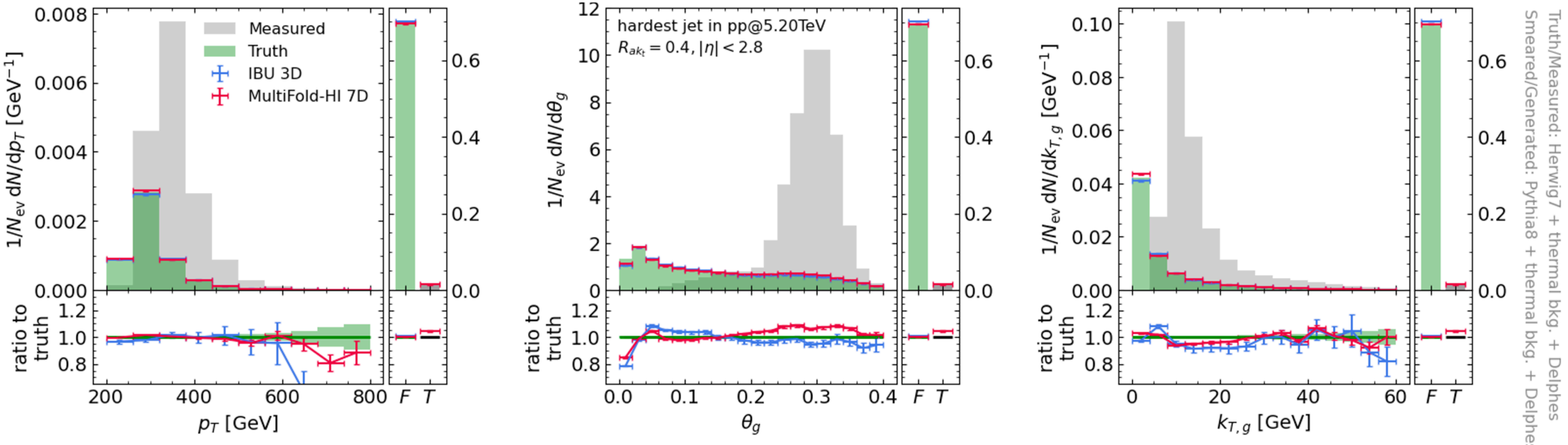
T. PANI POSTER

UNFOLDING WITH ML

SIMULATION BASED INFERENCE

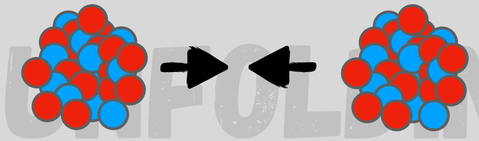
- **Tested for the first time on simulation in a HI environment** (PYTHIA/HERWIG + thermal background), similar or better performance to Bayesian unfolding in 3D.

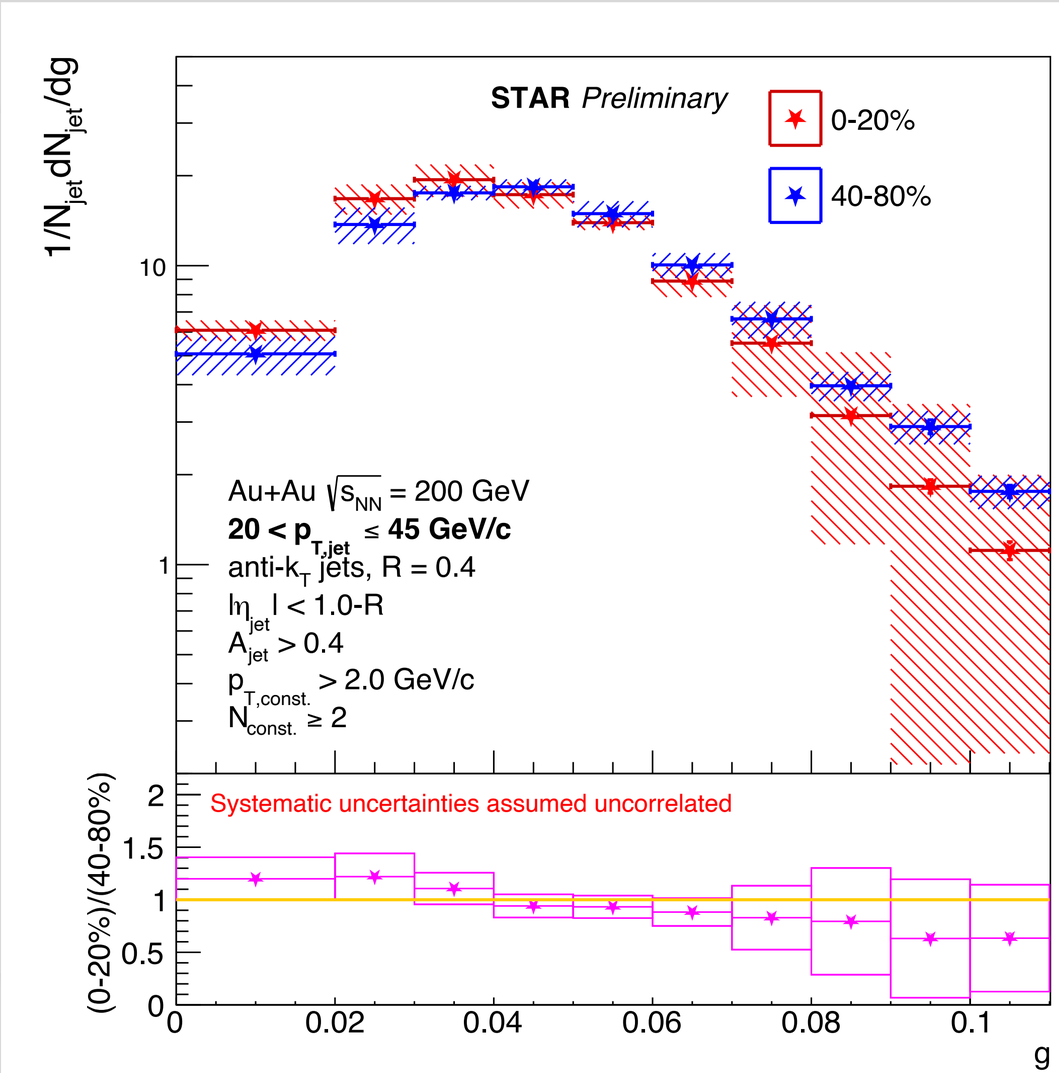
- ***No explicit background subtraction, built into MultiFold-HI!***



[A. Falcao, A. Tackas, to appear on arXiv]

UNFOLDING WITH ML





MULTIFOLD ENABLES MULTI-DIFFERENTIAL MEASUREMENTS OF JET SUBSTRUCTURE IN AU+AU.

Conventional Approach: Apply unfolding procedure on a binned distribution and repeat for each observable.

ML-based Approach: Use OmniFold to calculate weighting factors and unfold the phase space all at once, before the choice of binning or observable!

TREATMENT OF SIMULATION DEPENDENCE SHOWN FOR THE FIRST TIME!!

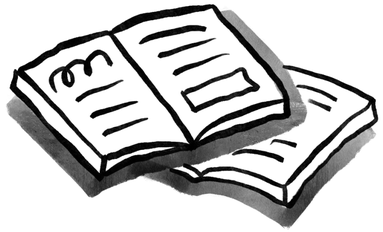
[Phys. Rev. D 102, 076004]

QUARK MATTER 2025

T. PANI POSTER

ROADMAP

WHAT IS ML AND WHY IS IT USEFUL FOR PHYSICS ANALYSIS?



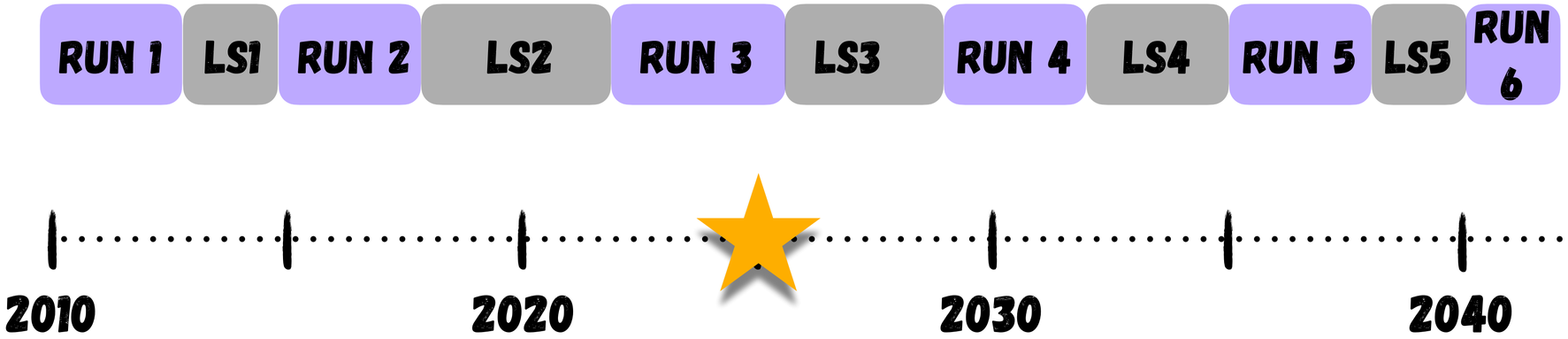
HOW IS ML CURRENTLY BEING USED AS AN ANALYSIS TECHNIQUE?

WHAT DOES THE FUTURE HOLD?

*Focus on new items since QM 2023 or items presented at QM 2025

WHERE ARE WE GOING?

LARGE HADRON COLLIDER



Brookhaven National Laboratory

STAR

PHENIX

SPHENIX

EPIC

RELATIVISTIC HEAVY ION COLLIDER

ELECTRON ION COLLIDER

Very large volumes of will be taken and analyzed in the decades to come - new tools will be *increasingly important*!

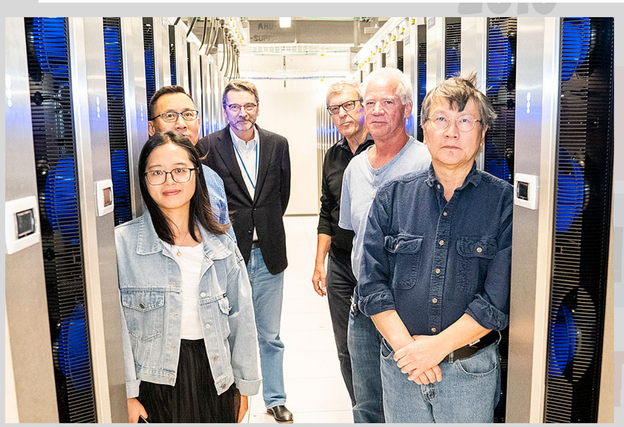
Brookhaven National Laboratory Newsroom Media & Communications Office

Newsroom Photos Videos Fact Sheets Lab History News Categories

Contact: Karen McNulty Walsh, (631) 344-8350, or Peter Genzer, (631) 344-3174 share:

Brookhaven's Computing Center Reaches 300 Petabytes of Stored Data

Largest compilation of nuclear and particle physics data in the U.S., all easily accessible — with plans for much more



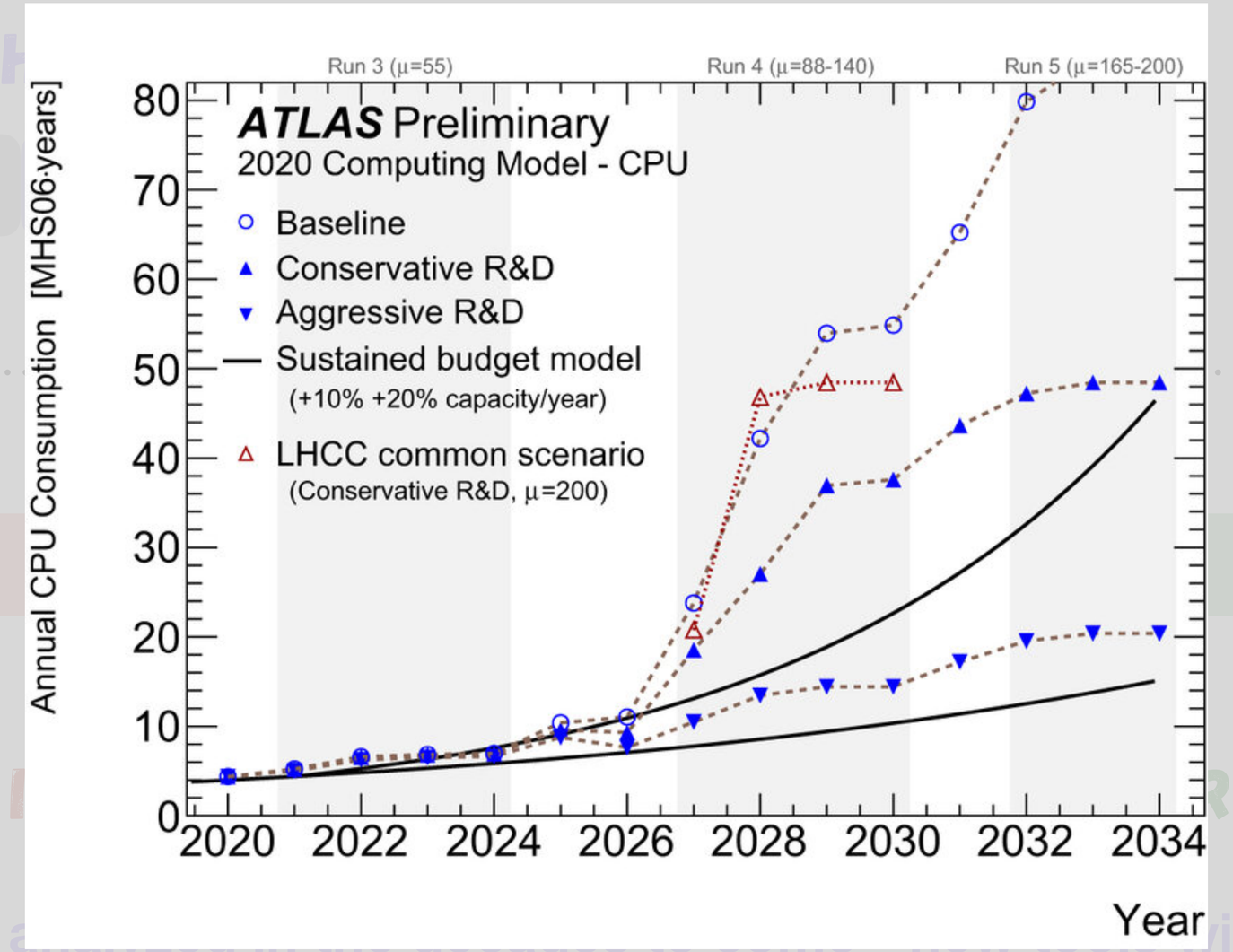
"RHIC's newest detector, sPHENIX, with a readout rate of 15,000 events per second, is projected to more than double the data we have now," said Chou.

- Total expected output from sPHENIX ~565 Petabytes

DRAMATIC IMPROVEMENTS IN READOUT RATES MAKING SPHENIX A LARGE FRACTION OF TOTAL DATA STORED!

Very large volumes of will be taken and analyzed in the decades to come - new tools will be *increasingly important*!

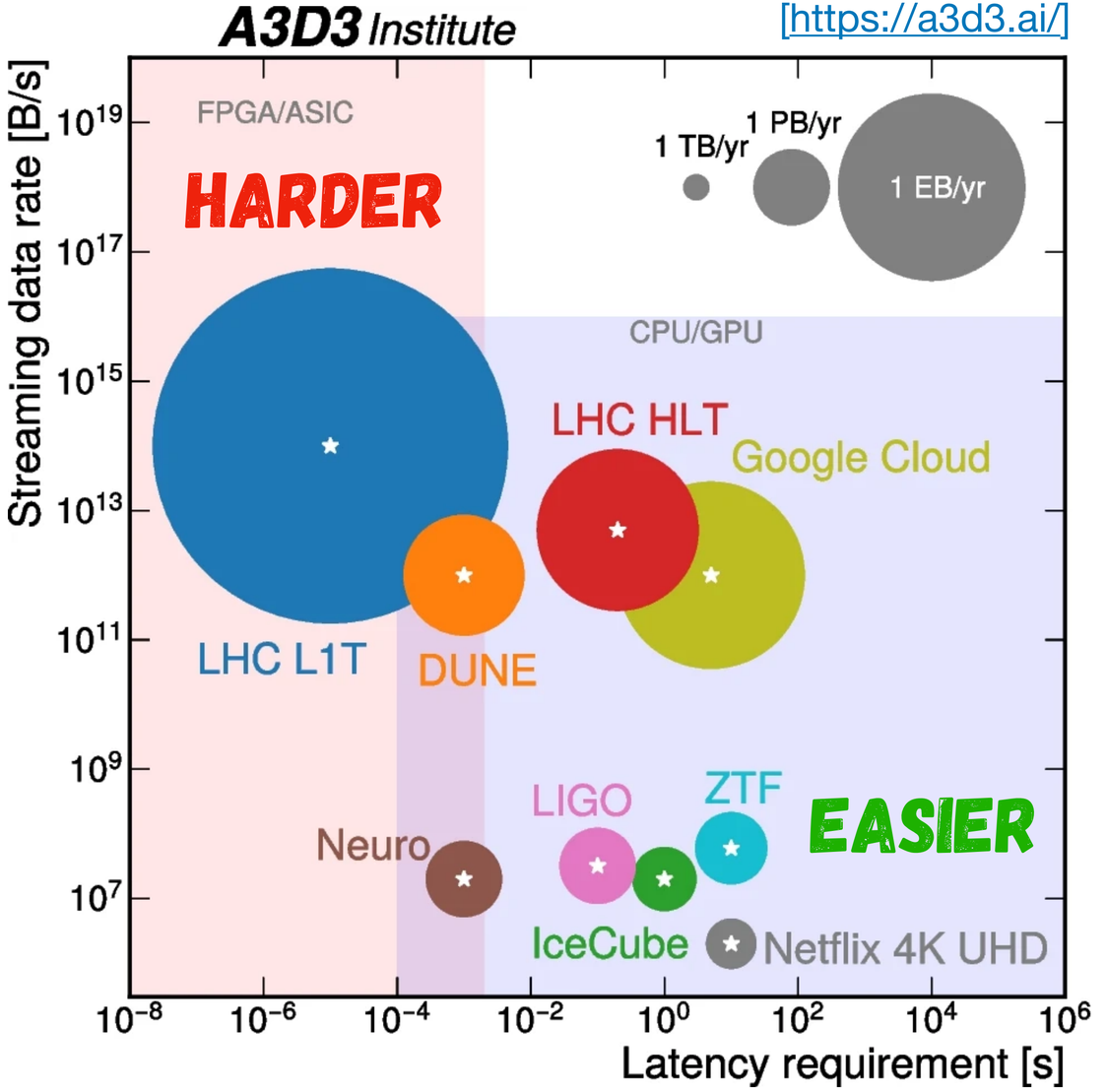
WHERE ARE WE GOING?



SAME TRENDS TRUE AT THE LHC!

EVENT FILTERING

- Data volume is increasing at a fast rate, need solutions for limited computing resources.
  - If we took all raw data, would easily exceed storage capabilities.



EVENT FILTERING

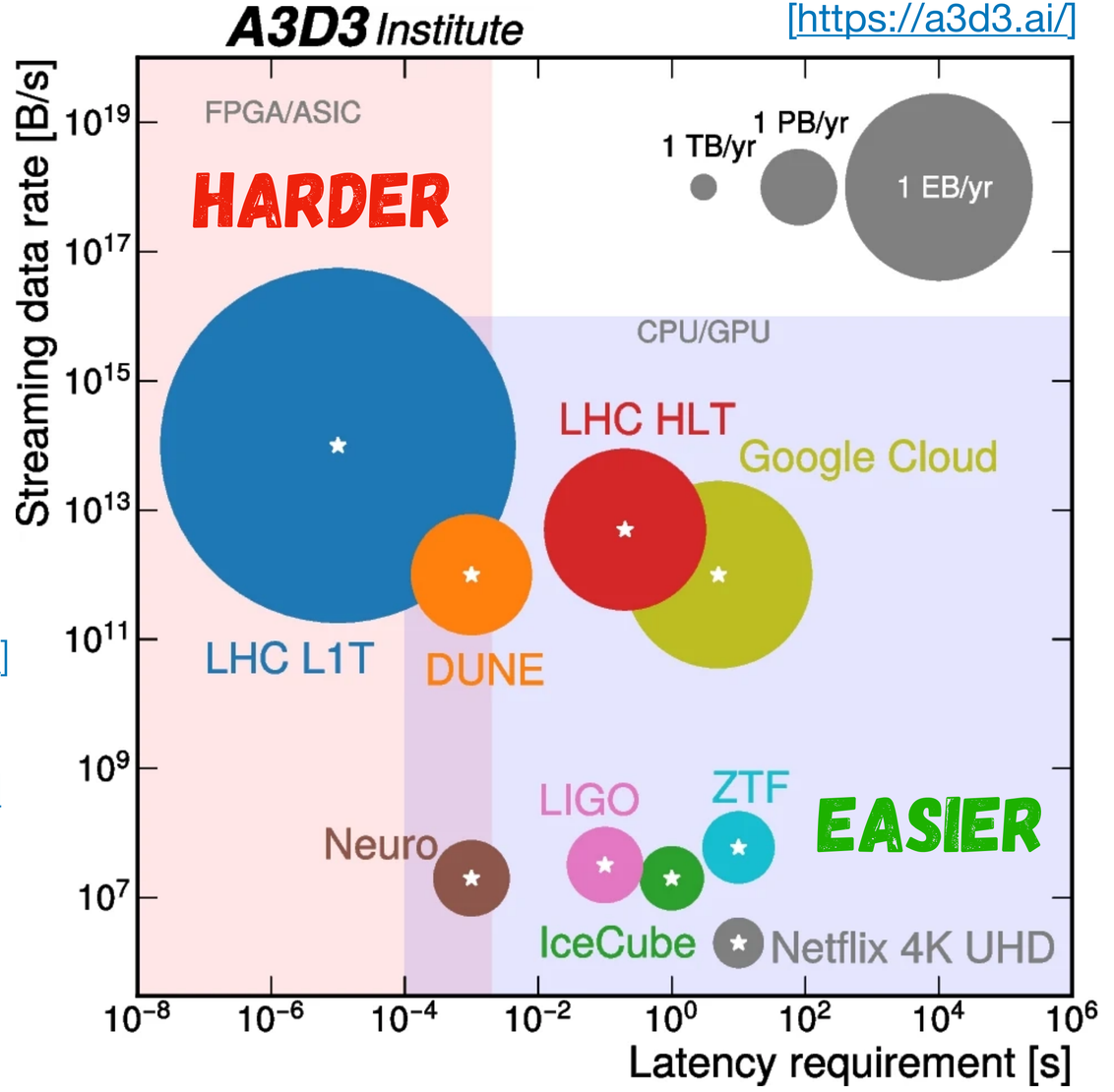
- Data volume is increasing at a fast rate, need solutions for limited computing resources.
  - If we took all raw data, would easily exceed storage capabilities.

- Perform fast selection/rejection of data with ML integrated into the firmware (FPGAs)
  - Use high level synthesis packages ex: hls4ml

CMS L1 Trigger [CMS-TDR-021] sPHENIX HF Trigger [JINST 19 C02066]

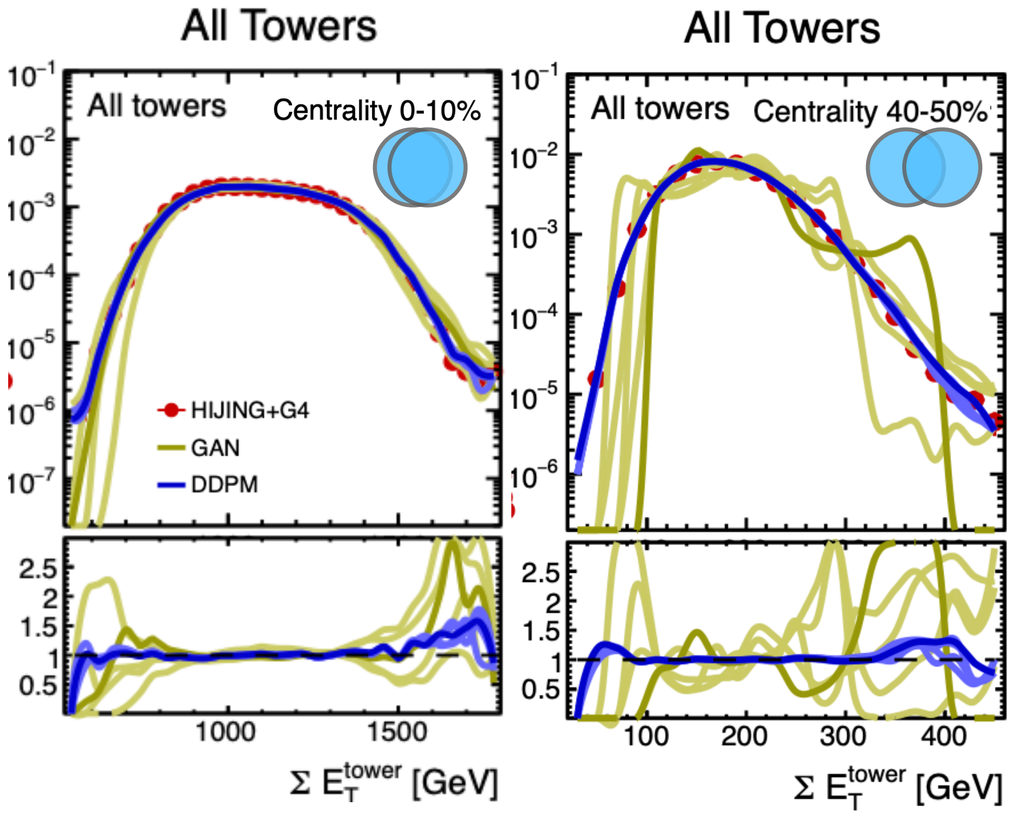
ATLAS Fake Track Rejection in Event Filter [ATLAS-TDR-029-ADD-1]

LHCb track reconstruction for HLT system [See website here]



FULL EVENT SIMULATION

- Full event simulations are highly complex and can be prohibitively computationally expensive (especially in HIs)
  - Ex: using denoising diffusion probabilistic model (DDPM) and Generative Adversarial Networks (GAN) to generate heavy-ion events in sPHENIX designed to mimic HIJING



Method	Time per event
HIJING + GEANT4	40 mins
DDPM	1.34s
GAN	0.42 ms

- Tradeoff between time and accuracy!

QUARK MATTER 2025

Y. GO POSTER

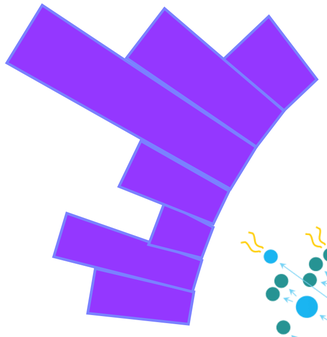
[Phys. Rev. C 110, 034912 (2024)]

PARTIAL EVENT SIMULATION

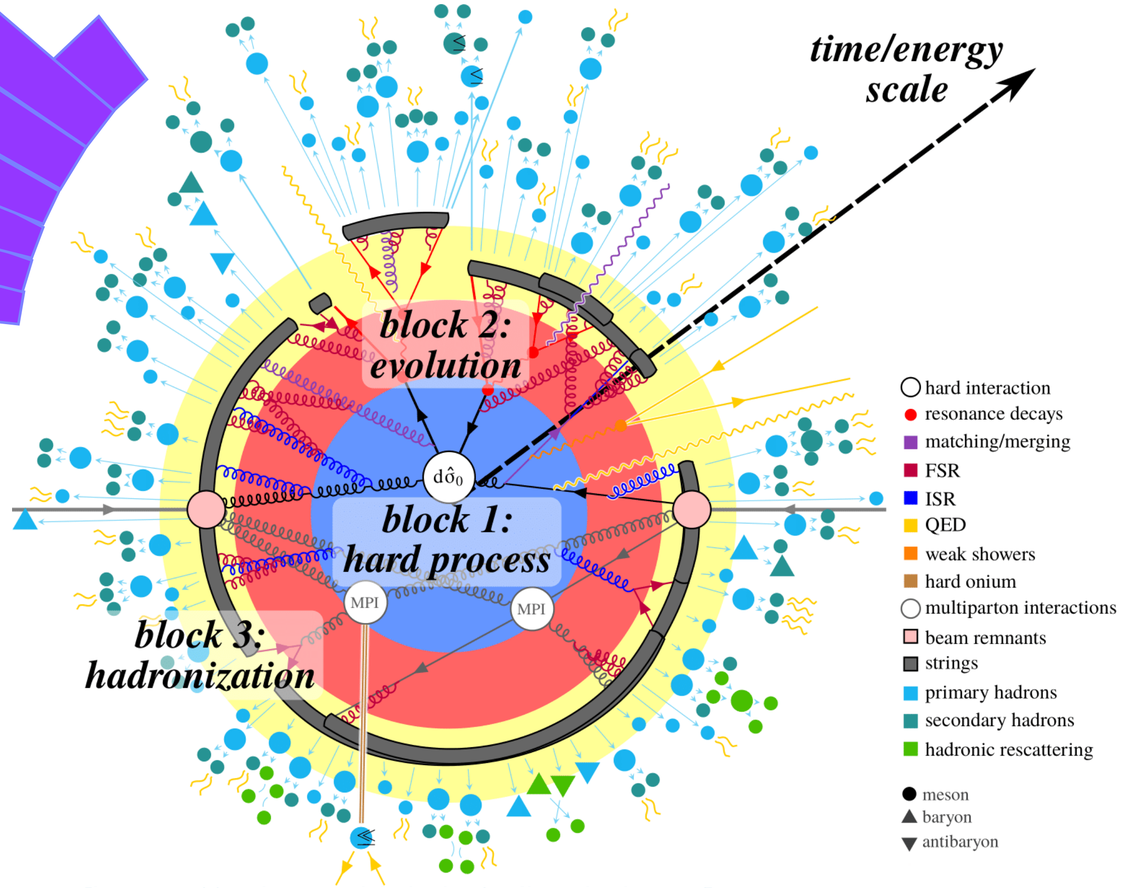
- Can also take components of simulation that are computationally expensive and replace that with machine learning.

(If Needed) Block 4: Detector sim

MLHAD



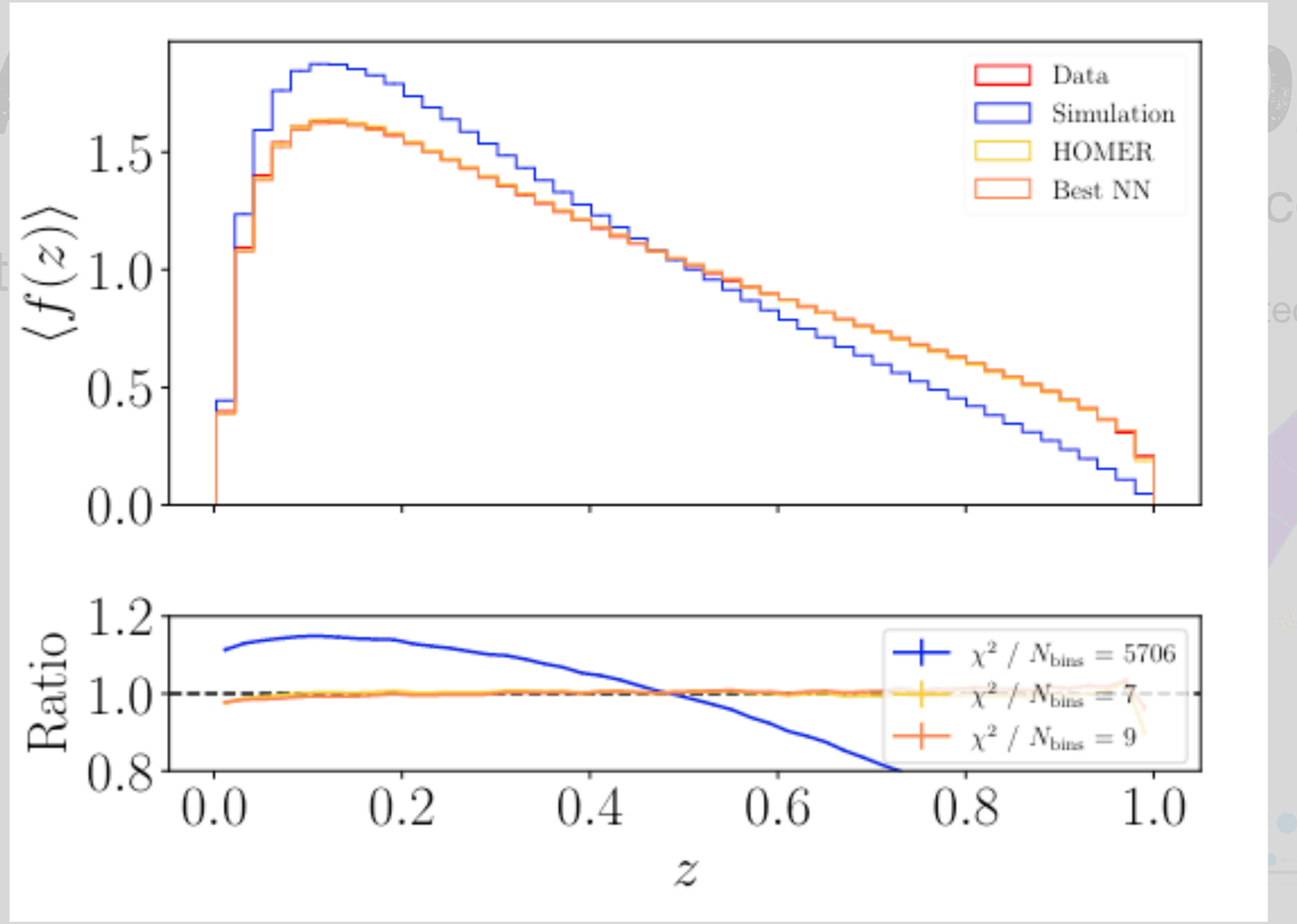
- Use ML to evaluate the collision kernel



[https://uchep.gitlab.io/mlhad-docs/]

QUARK MATTER 2025
S. CABODEVILA
TUES. 15:30

QUARK MATTER 2025
C. BIERLICH
POSTER



- **Goal:** learn data-driven fragmentation function by estimating likelihood ratios for each string break.
- Use *event weights* derived from base simulation in order to *train two neural networks* that are used to predict string break weights.

ABLE TO REPRODUCE FRAGMENTATION FUNCTION FROM DATA!

[SciPost Phys. 18 (2025) 2, 054]

QUARK MATTER 2025
S. CABODEVILA
TUES. 15:30

QUARK MATTER 2025
C. BIERLICH
POSTER

[https://uchep.gitlab.io/mlhad-docs/]

OPEN QUESTIONS FOR NEXT ~5 YEARS

HOW DO WE ASSIGN A SYSTEMATIC UNCERTAINTY FOR THE ML?

HOW DO WE CONSTRUCT MORE INTERPRETABLE MODELS?

HOW CAN WE MAKE ML-BASED APPLICATIONS REPRODUCIBLE?

DO WE NEED TO STANDARDIZE ML APPLICATIONS ACROSS EXPERIMENTS?

SYSTEMATIC UNCERTAINTY FOR ML

Bias is not inherently bad!

- Just like any analysis technique, you need to study and quantify the bias on your results.
- ***Should*** correspond to a systematic uncertainty
  - No standard way to do this for physics applications, but there has been progress!

SYSTEMATIC UNCERTAINTY FOR ML

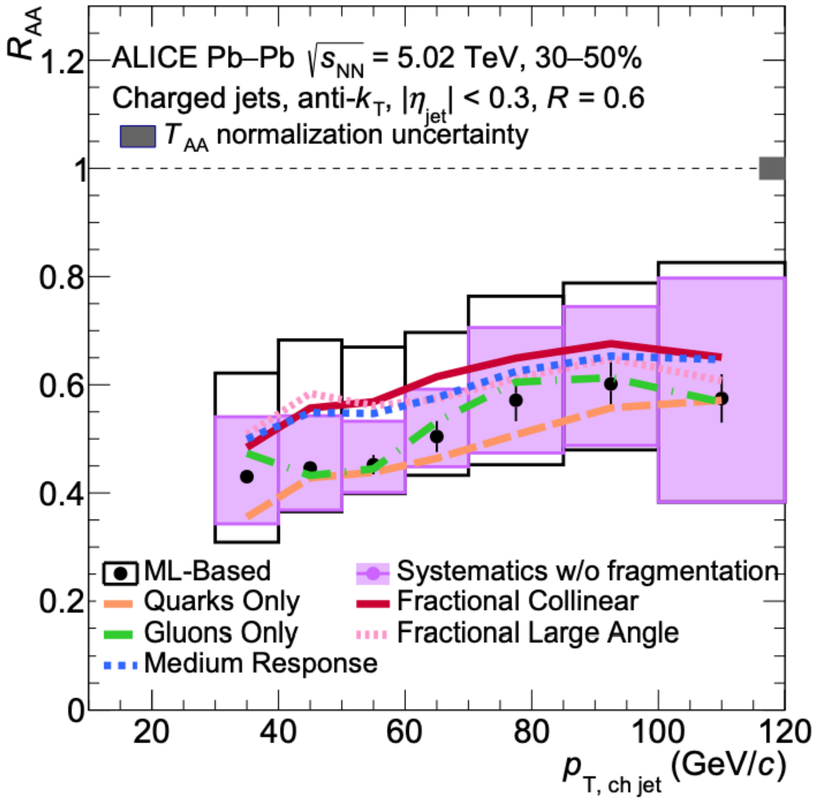
Bias is not inherently bad!

- Just like any analysis technique, you need to study and quantify the bias on your results.
- ***Should*** correspond to a systematic uncertainty
  - No standard way to do this for physics applications, but there has been progress!

Example case: jet p_{T} reconstruction

- Initial ALICE approach is to train on five different fragmentation models, taking difference in final result as systematic uncertainty

ALICE, [PLB 849 (2024) 138412]



SYSTEMATIC UNCERTAINTY FOR ML

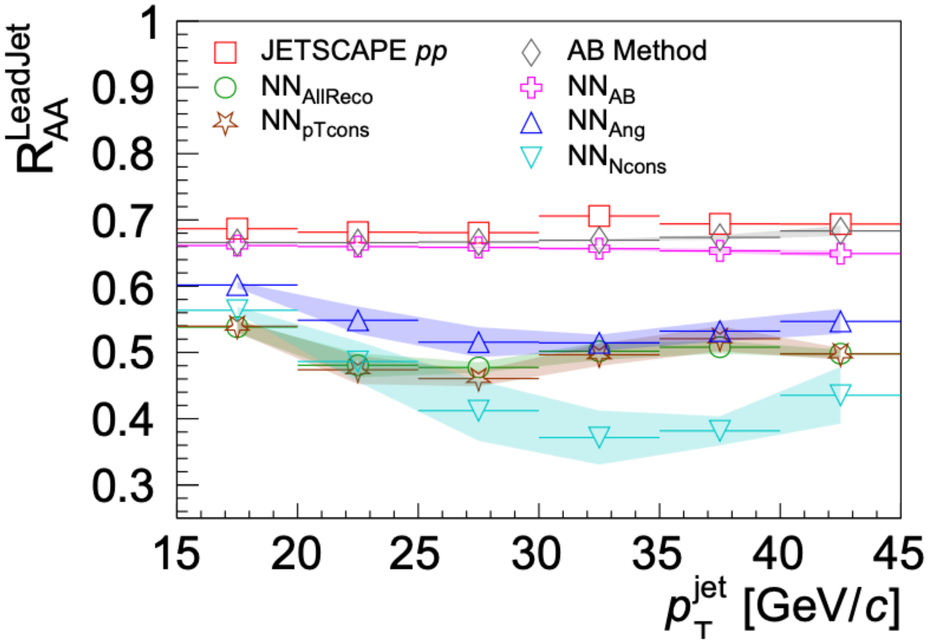
Bias is not inherently bad!

- Just like any analysis technique, you need to study and quantify the bias on your results.
- ***Should*** correspond to a systematic uncertainty
  - No standard way to do this for physics applications, but there has been progress!

- If the bias is too large, traditional techniques or ML w/ reduced complexity is preferable.

Example case: jet p_{T} reconstruction

- Indications that corresponding systematic uncertainty may be larger at lower p_{T}



[arXiv:2412.15440]

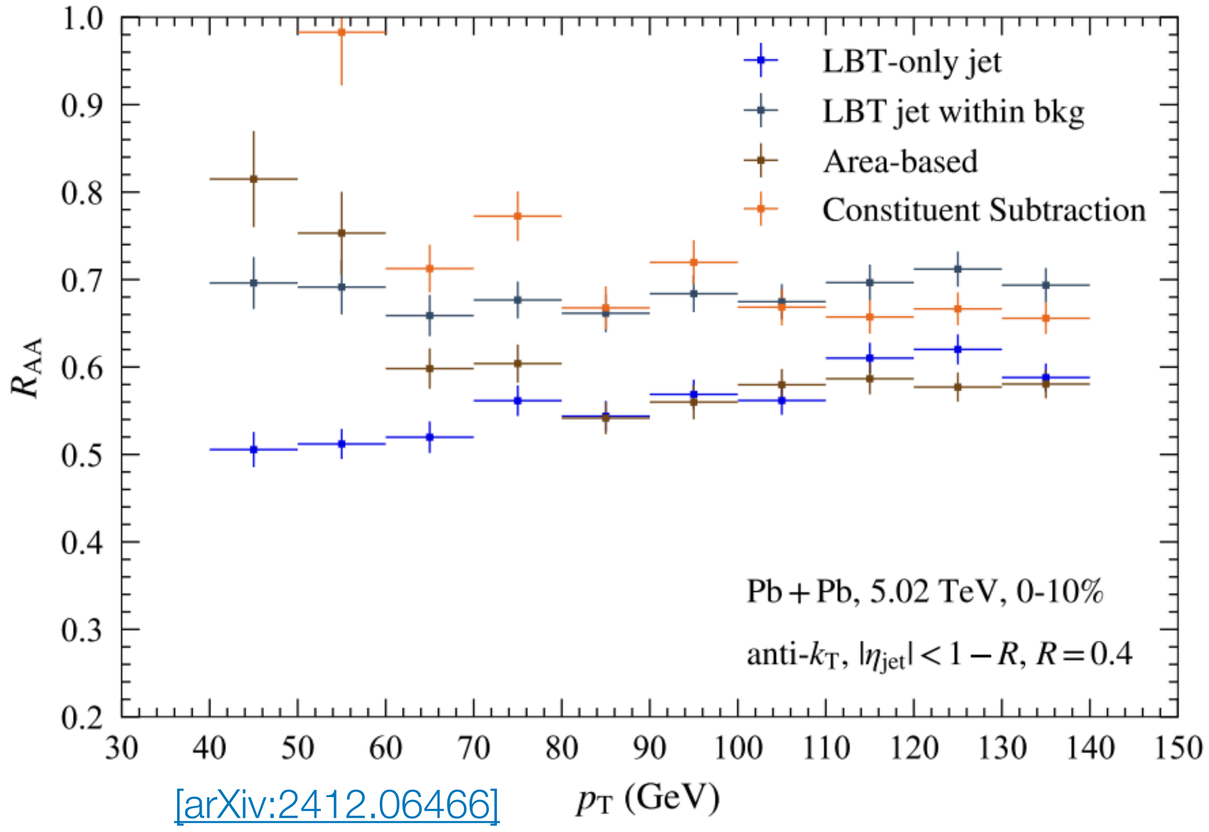
SYSTEMATIC UNCERTAINTY FOR ML

Bias is not inherently bad!

- Just like any analysis technique, you need to study and quantify the bias on your results.
- ***Should*** correspond to a systematic uncertainty
  - No standard way to do this for physics applications, but there has been progress!

Example case: jet p_{T} reconstruction

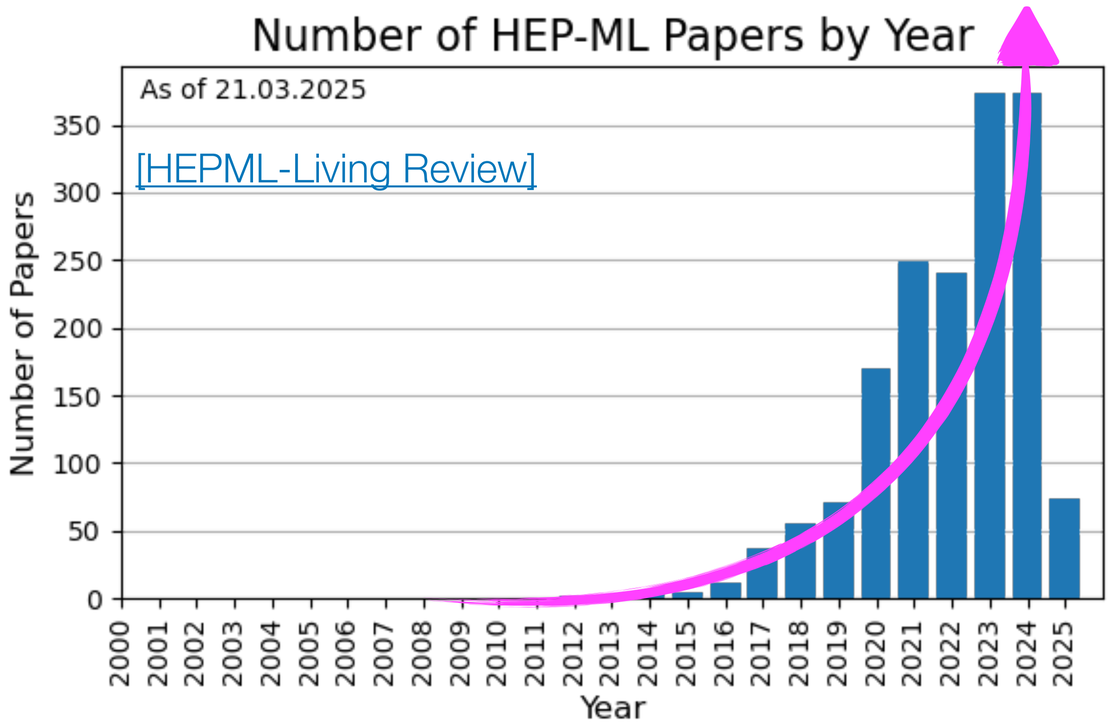
- As simulations improve, so will our treatment of the systematic uncertainties.
- Training on LBT model appears to better reproduce results from traditional approaches than when trained on PYTHIA.
  - Applied to LBT + thermal background.



[arXiv:2412.06466]

CONCLUSIONS

- We are taking more data and making more complex measurements than ever before!
- Machine learning has led to new physics insights and can be used throughout the whole analysis pipeline!
  - **Many great examples at this conference!**



- Will be crucial at future facilities such as FAIR, HL-LHC, and the EIC!
- **Stay tuned for many interesting results in the future!!**

deep ai image editor

Thank you!!

Special thanks to Lee Barnby, Xuan Li, Changwhan Choi, and the MIT Heavy Ion group for useful discussions and feedback!

deep ai image editor

Backup

DATA REDUCTION

- One other solution is to reduce the data size, for example, with auto-encoders.

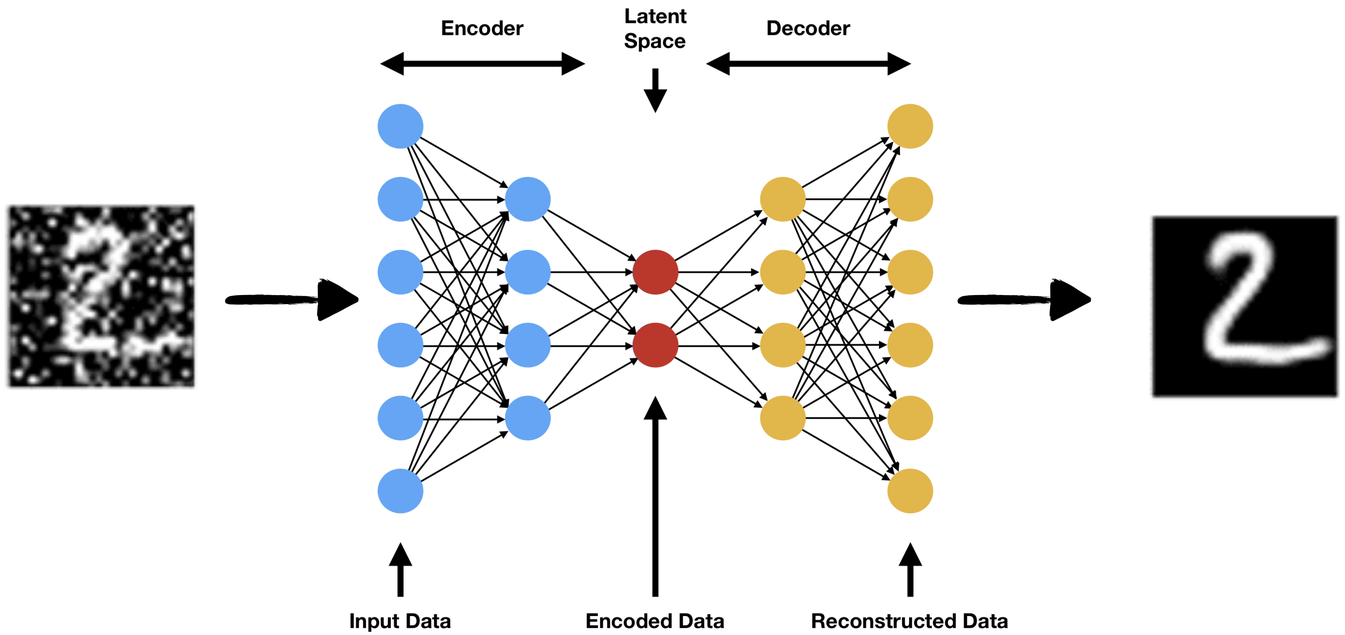
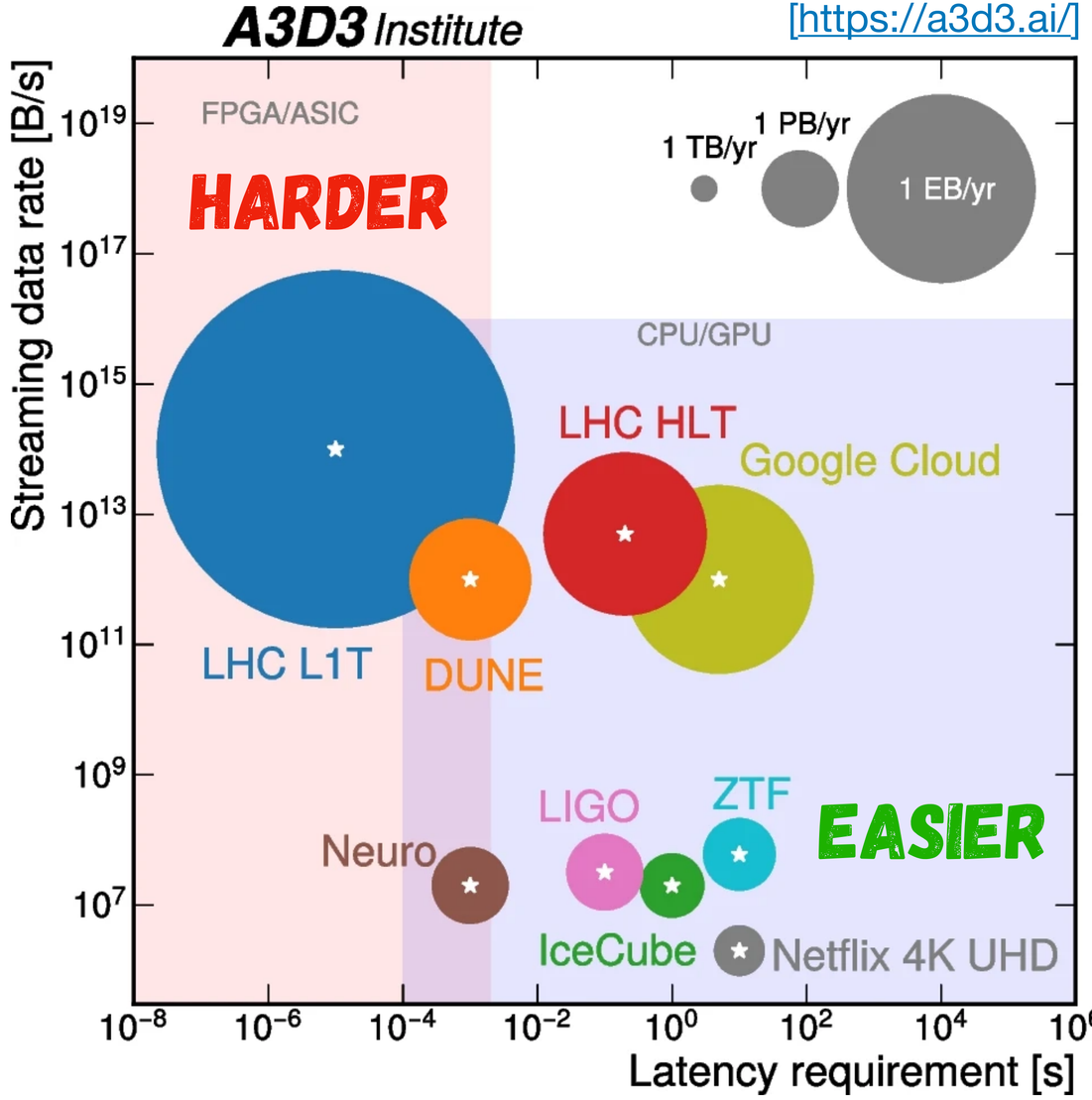


Image Credit: https://www.compthree.com/blog/autoencoder/

sPHENIX TPC [arXiv:2310.15026]

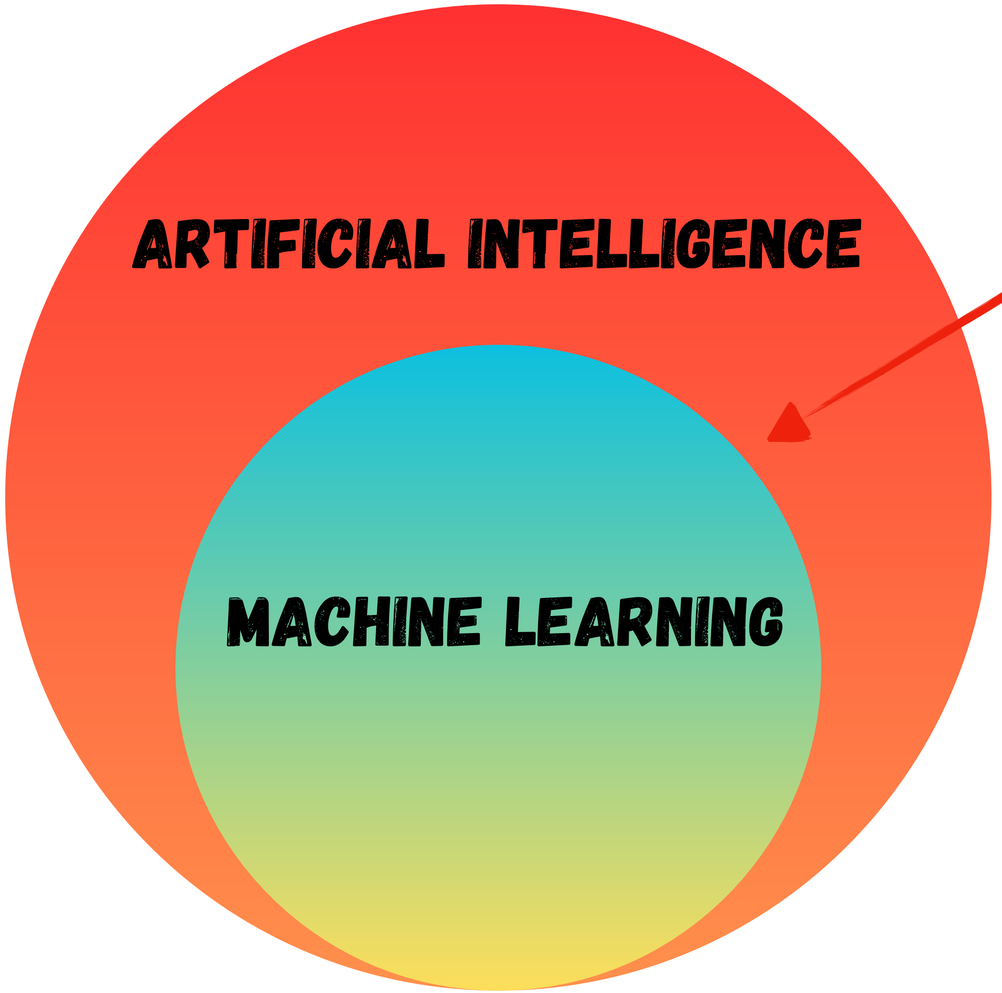
CMS HGCAL upgrade [See talk at Fast ML 2022]

LHCb trigger system [See website here]



WHAT IS AI/ML?

Artificial Intelligence: Programs with the ability to acquire and apply knowledge and skills.



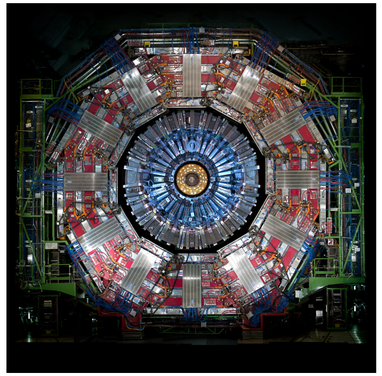
Ex: Chatbots (humans give rules)

Machine Learning: computational algorithms that imitate how humans learn, i.e. learning from examples in a way that improves over time

ANALYSIS PIPELINE

MACHINE LEARNING CAN BE USED THROUGHOUT THE ANALYSIS PIPELINE!

ACCELERATOR COMPLEX



EXPERIMENTS

EVENT FILTERING

RECONSTRUCTION

RECONSTRUCTED DATA

SIMULATION

EVENT/ANALYSIS SELECTIONS

RESULTS

ML FOR UNDERLYING PHYSICS

"Data"-based learning complements simulation-based inference.

~ Given an answer ~ "White Box" ML ~ Underlying physics	~ Domain knowledge ~ "Black Box" ML ~ Answer

- Learning from data is difficult due to systematic experimental biases.
- Helpful in understanding uncertainties or shortcomings of models!

Proof of concept identifying the AP splitting function exists [PLB 829 (2022) 137055]

THIS IS A LONG TERM EFFORT!

HOW DOES THE MACHINE LEARN?

SUPERVISED LEARNING

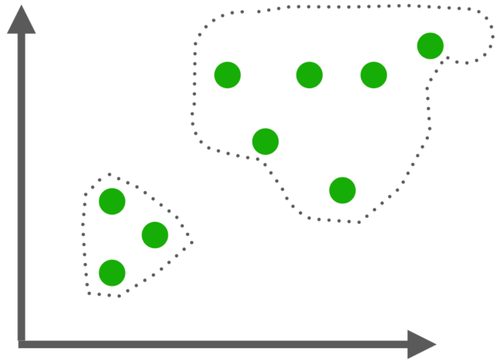
Algorithm learns from a labeled set of “true values”.

Driven by the Task

Analogy: Taking a test

UNSUPERVISED LEARNING

Algorithm finds structure in the data without knowing the desired outcome.



Driven by the Data

Analogy: Clustering

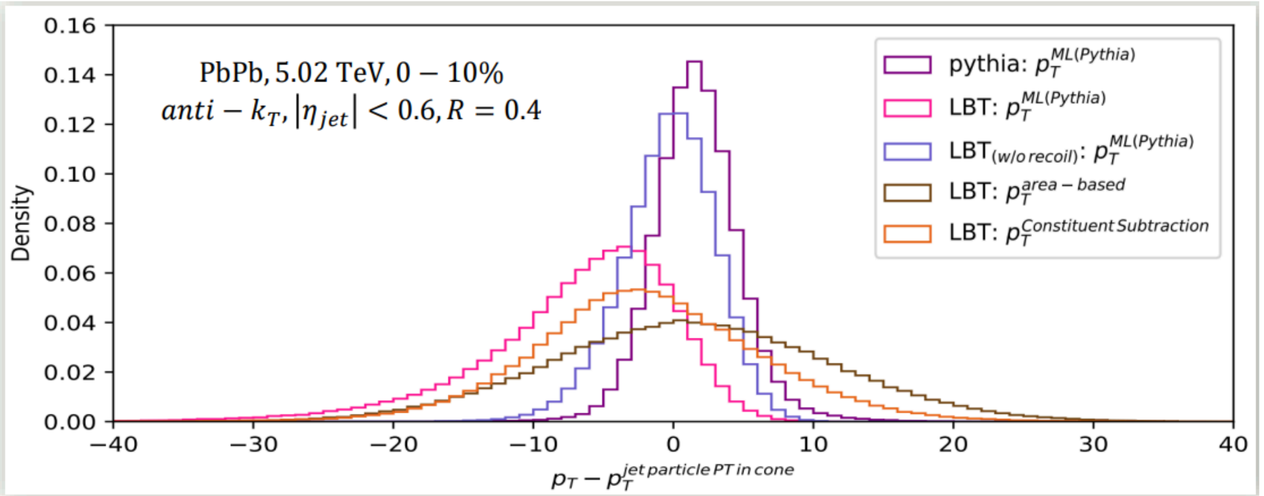
REINFORCEMENT LEARNING

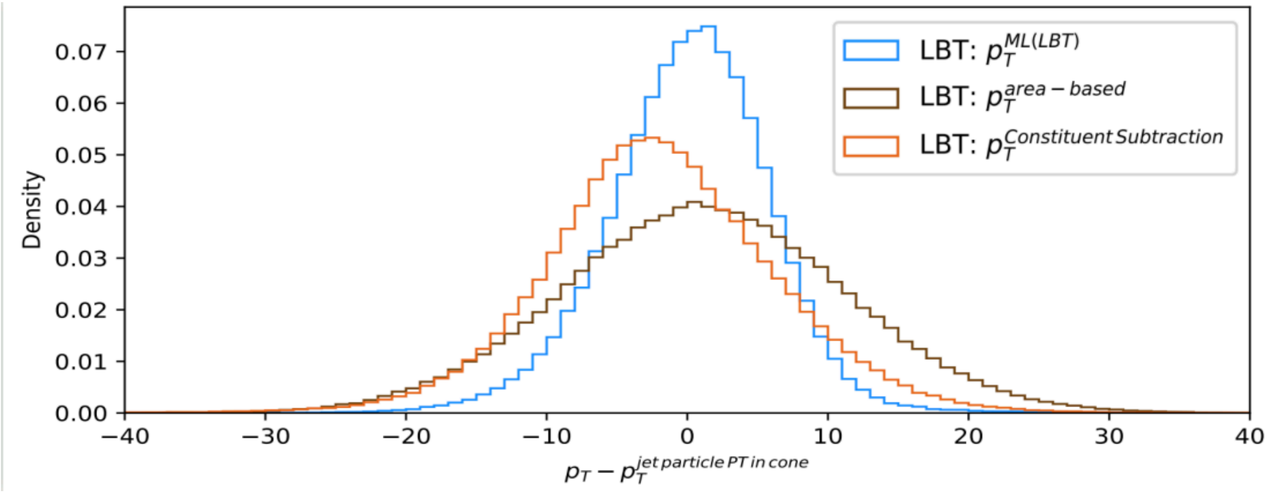
Algorithm learns in a reward based system to determine a series of actions.

Driven by the Reward

Analogy: Dog training

STUDIES WITH NN JET PT RECONSTRUCTION





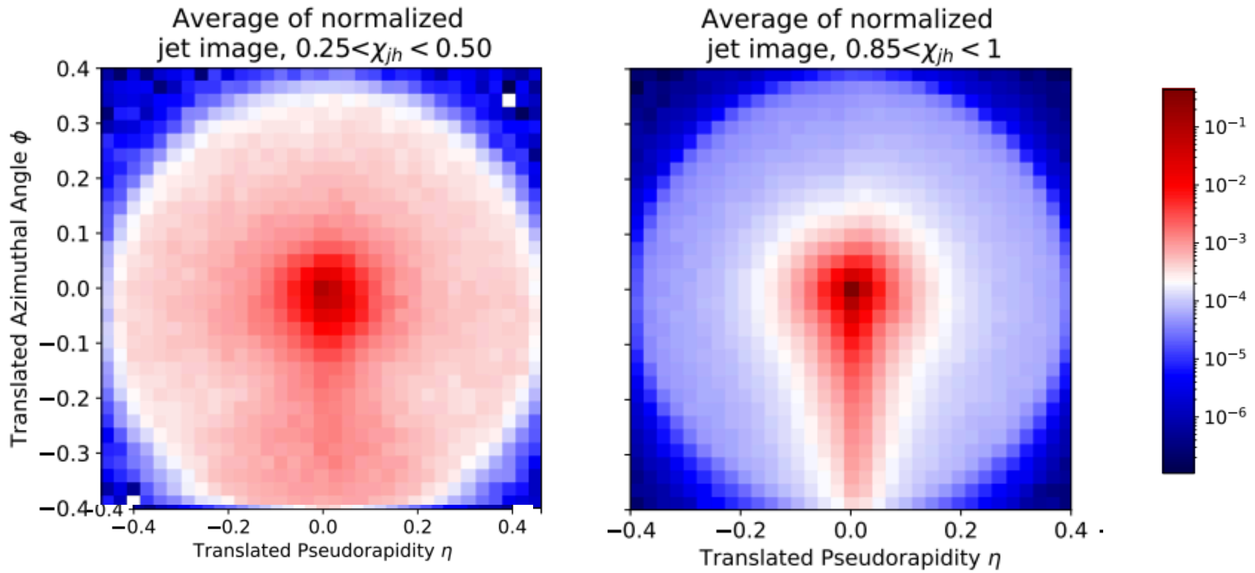
- See offset (bias) in δp_{T} when ML is trained in PYTHIA vs. LBT.
- Crucial for applications in data to correct for this bias in an unfolding procedure.
  - Apply same model on your data and the response matrix.

HARD PROBES 2024

SEE R. LI POSTER

DEEP LEARNING JET MODIFICATIONS

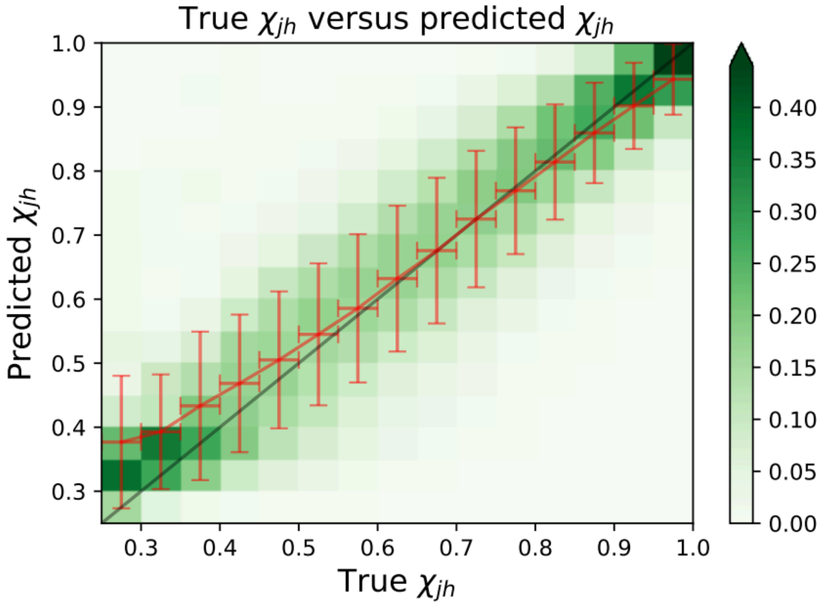
Use supervised learning on jet images with a CNN to perform the regression task of predicting the energy loss ratio in HI collisions (hybrid model).



quenched **unquenched**

$$\chi_{jh} = \frac{E_f^h}{E_i^h}$$

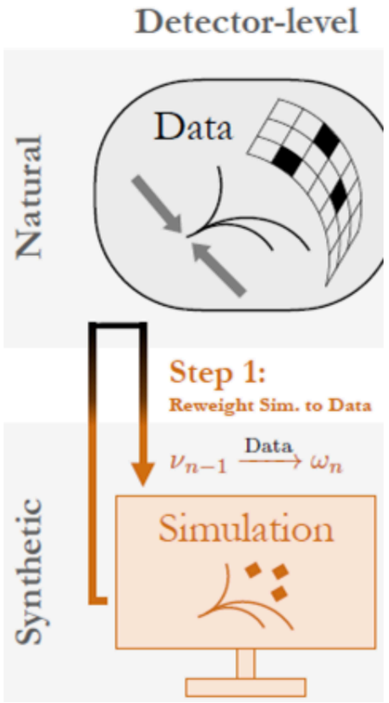
[JHEP 2021, 206 (2021)]

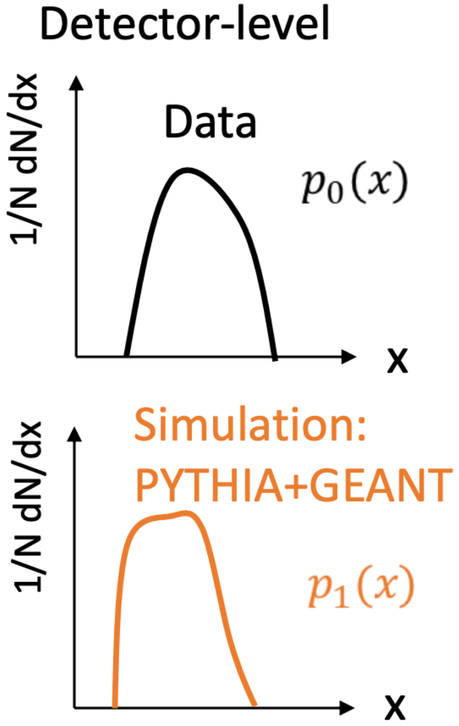


Shows good performance!

- Very useful to separate and study quenched vs. unquenched jets as well as extracting the initial energy of the jet. (Ideal probe of selection bias!)

OMNIFOLD/MULTIFOLD





Where does the machine learning part come in?

E.g., Iteration 1, step 1:

Weights: $w(x) = p_0(x)/p_1(x)$ Ok for 1D

$\approx f(x)/(1 - f(x))$ (Andreassen and Nachman PRD 101, 091901 (2020))

where f(x) is a neural network and trained with the binary cross-entropy loss function

to distinguish jets coming from data vs from simulation

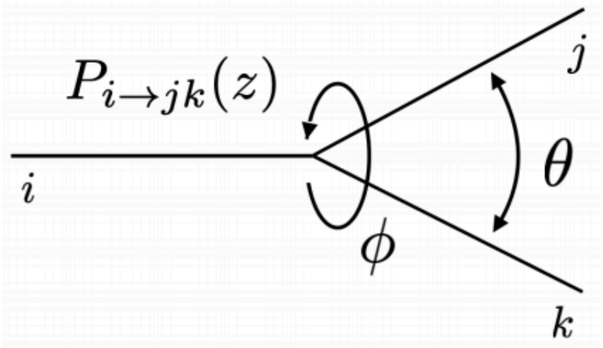
Unfolding → Reweighting histograms → Classification → Neural network

Slide from Youqi Song

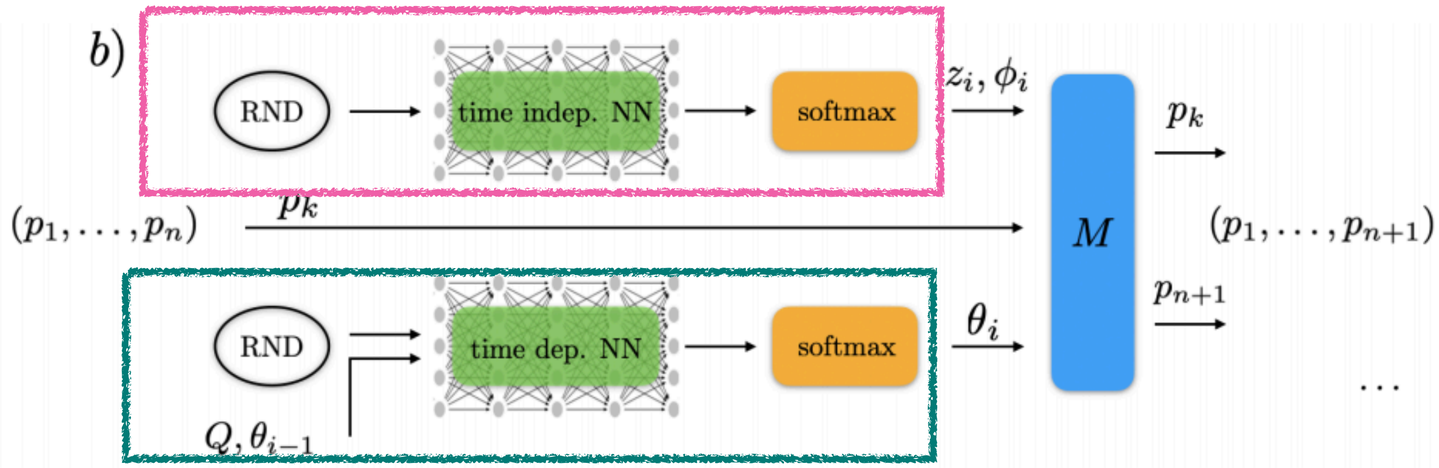
PROOF OF CONCEPT

- Extract splitting function from the network in white-box ML.

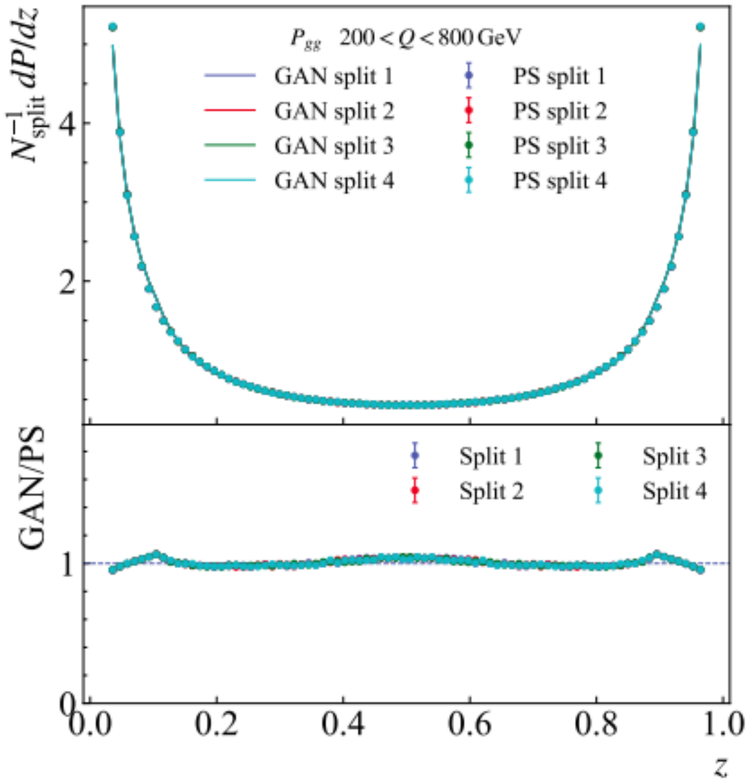
Done with a GAN split into two components.



1. Time independent learns the z, ϕ



2. Time dependent learns the θ



Was able to reproduce AP splitting function.

EVENT CLASSIFICATION AT THE EIC

[JHEP 03 (2023) 085]

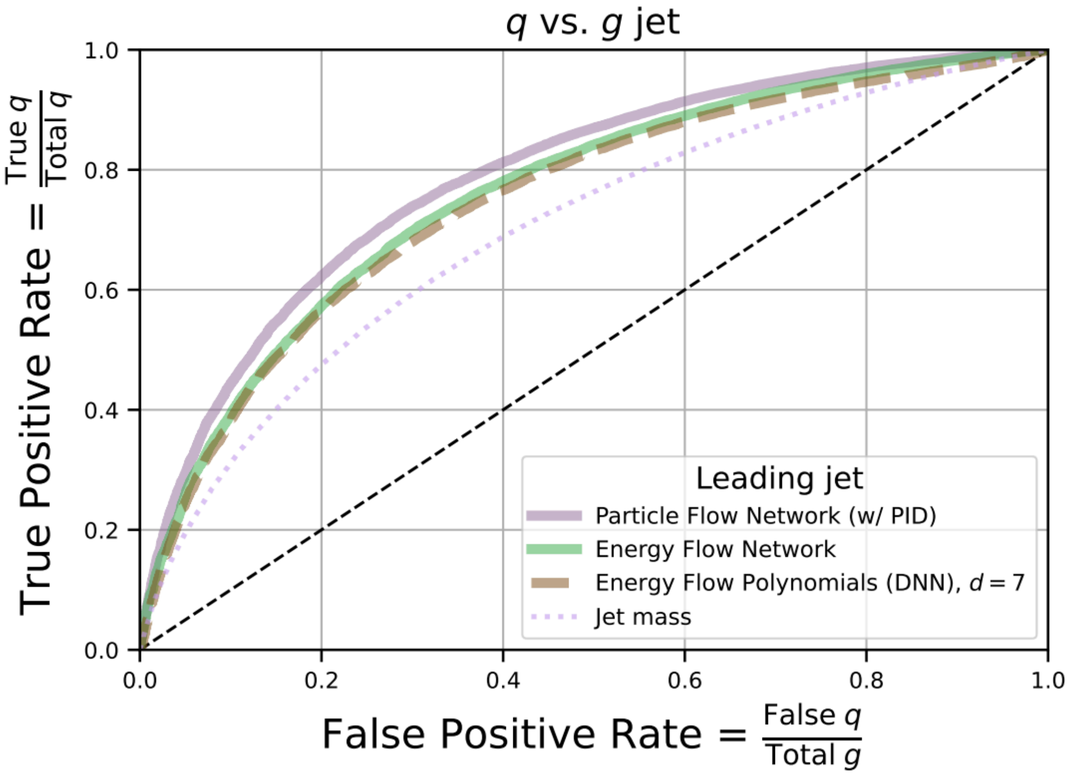
- Study the effectiveness of ML-based classifiers to
  - Identify the flavor of the jet
  - Identify the underlying hard process of the collision

- Additionally study the effectiveness of different ways of representing information
  - Particle Flow Networks [JHEP 01 (2019) 121]

$$f(p_1, \ldots, p_N) = F\left(\sum_{i=1}^{N} \Phi(p_i)\right) \qquad p_i = (z_i, \eta_i, \phi_i, \mathrm{PID}_i)$$

  - Energy Flow Polynomials [JHEP 04 (2018) 013]

$$\mathrm{EFP}_G = \sum_{i_1} \cdots \sum_{i_V} z_{i_1} \cdots z_{i_V} \prod_{(k,l)\in E} \theta_{i_k i_l}$$

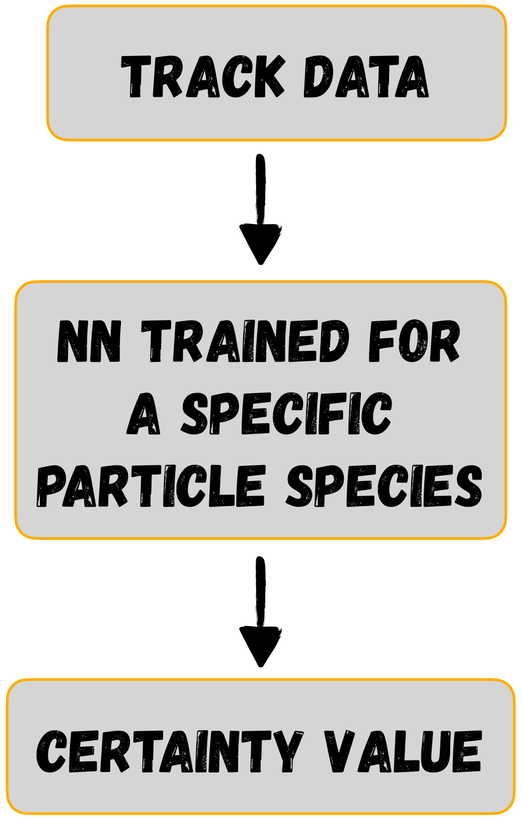


Indications that ML-based methods will have an improved performance over traditional techniques!

See also event classification with large language models, [arXiv:2404.05752]

ML FOR PARTICLE IDENTIFICATION

[Eur. Phys. J.C 84 (2024) 7, 691]



- Use NN trained on a specific particle type to predict a certainty value that is then compared to a pre-set threshold.
- Decide threshold based on efficiency/purity tradeoff.
- Takes into account particles from different sub-detectors (here TPC, TOF, TRD of ALICE) , robust against missing data.

Model	Precision	Recall
Standard	**99.99 ± 0.01**	78.37 ± 0.01
Ensemble	97.47 ± 0.25	99.46 ± 0.21
Mean	97.31 ± 0.07	99.52 ± 0.07
Proposed	97.49 ± 0.06	**99.54 ± 0.05**
Regression	97.33 ± 0.06	99.49 ± 0.07

Pion

Model	Precision	Recall
Standard	**99.40 ± 0.01**	59.72 ± 0.03
Ensemble	97.16 ± 0.46	93.74 ± 0.30
Mean	97.85 ± 0.41	93.34 ± 0.32
Proposed	97.80 ± 0.44	**93.86 ± 0.27**
Regression	97.38 ± 0.40	93.67 ± 0.38

Proton

Model	Precision	Recall
Standard	**92.87 ± 0.01**	60.37 ± 0.05
Ensemble	91.18 ± 02.00	82.72 ± 01.42
Mean	90.83 ± 01.71	82.32 ± 0.96
Proposed	91.55 ± 0.71	**83.68 ± 0.82**
Regression	91.17 ± 01.00	81.78 ± 0.21

Kaon

- When comparing the standard method to the proposed method, proposed method has better balance of precision (purity) and recall (efficiency)!

See also LHCb NN to identify calo hits [Int. J. Mod. Phys. A 30, 1530022 (2015)], ATLAS Electron PID w/ CNN [ATL-PHYS-PUB-2023-001]
CMS Deep NN to identify hadronic τ-lepton decays [JINST 17 (2022) P07023]

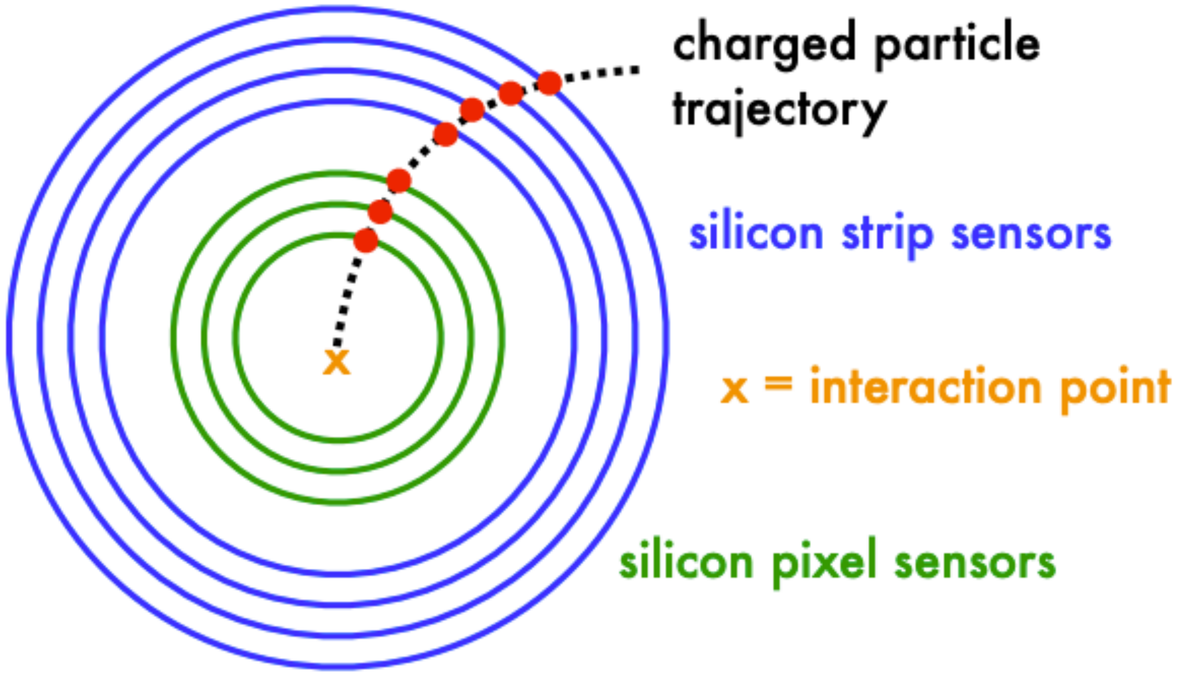
TRACK RECONSTRUCTION AT THE HL-LHC

- Data volume and reconstruction will also be a problem for the HL-LHC
  - Reconstructing charged particle trajectory is computationally expensive - increases with the power of the multiplicity.

Standard approach: Kalman Filter used to locate hits in charged particle trajectory

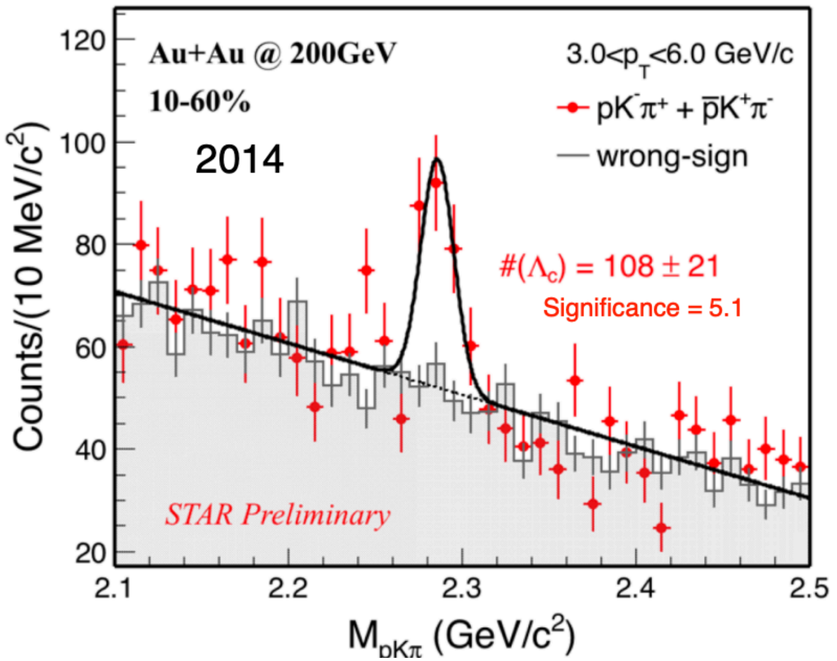
ML-based approach: Use ML tools to speed this up such as…

- Recurrent Neural Network [arXiv:2212.02348]
- Convolutional neural network [See Here]

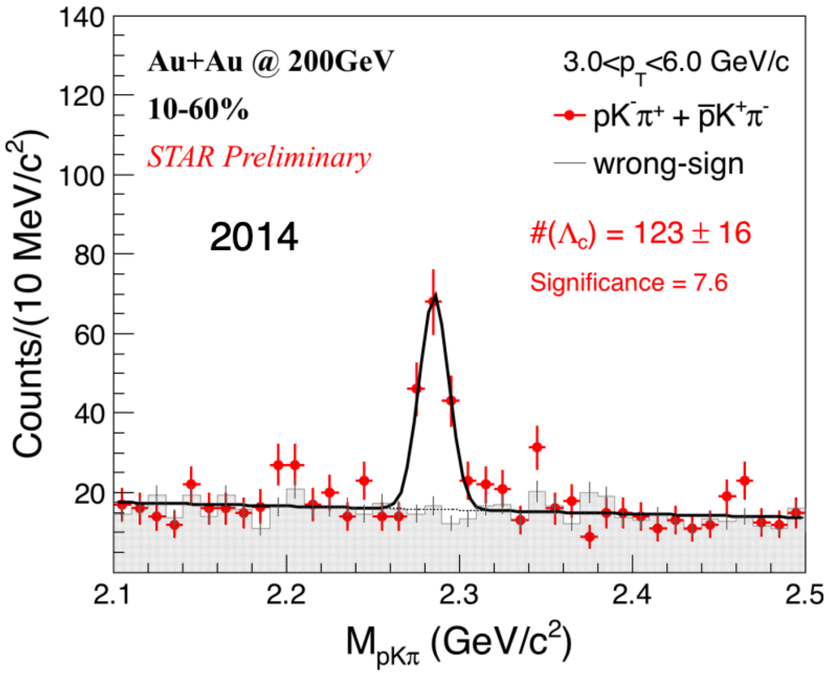


SIGNAL/BACKGROUND DISCRIMINATION

Traditional Techniques



With BDT

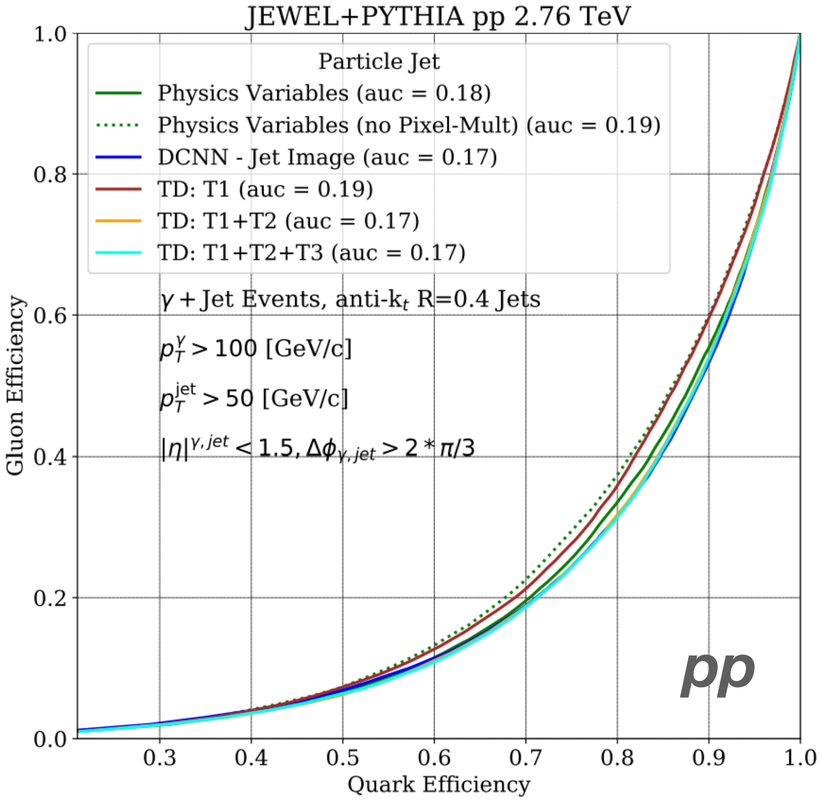


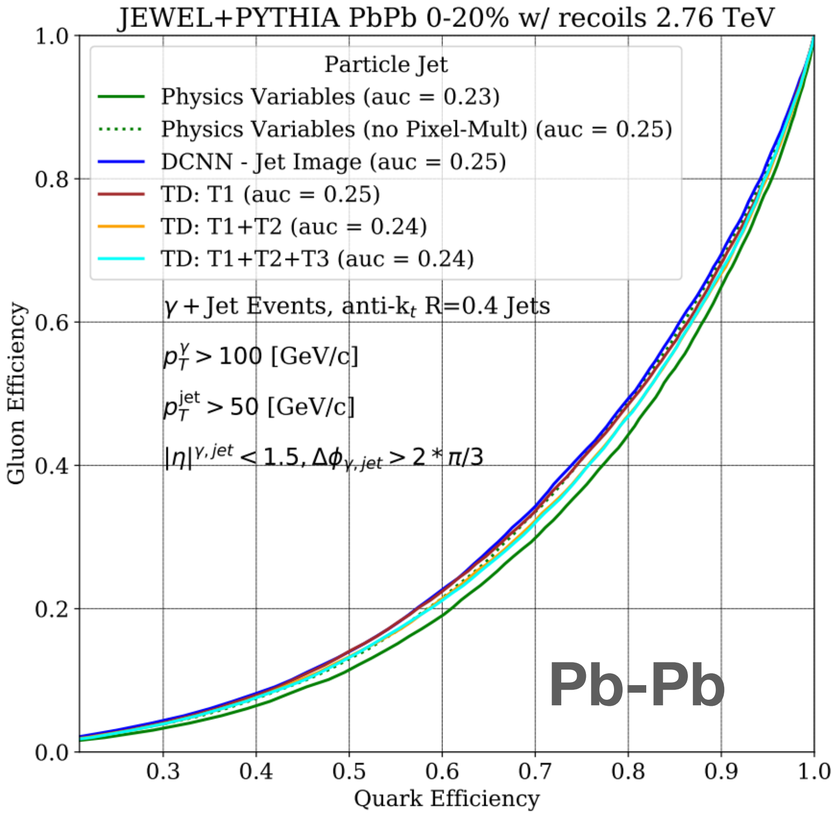
[PRL 124, 172301 (2020)]

- Boosted Decision Tree implemented in ROOT TMVA to optimize signal for Λ_c baryon production.
- Trained in a supervised manner with EvtGen
- 50% increase in signal significance with ML!

QUARK VS. GLUON JETS WITH ML

Y. Chien, R. Elayavalli: 1803.3589





→ *Lower the curve, the better the performance.*

→ All methods explored tend to perform consistently, indicating that they may be picking up on similar features.

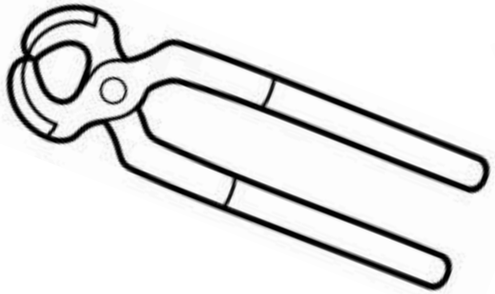
→ The performance worsens for Pb—Pb, due to the large UE.

Quark and gluon discrimination is a difficult and ongoing effort in HIs!

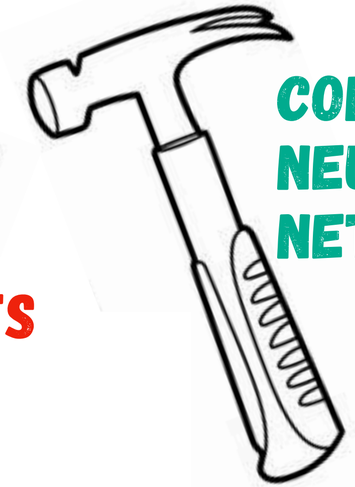
Future: Apply these methods to data in pp and Pb—Pb!

MODERN ML TOOLS

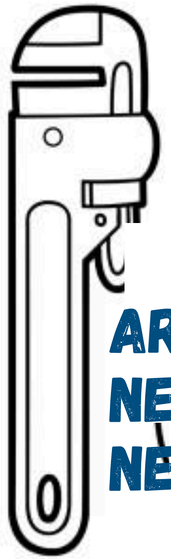
BOOSTED DECISION TREES (BDTS)



RANDOM FORESTS

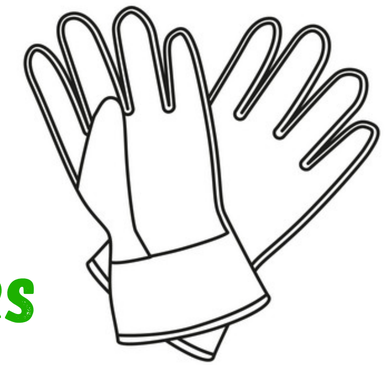


CONVOLUTIONAL NEURAL NETWORKS (CNN)



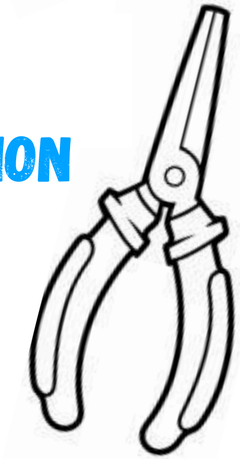
ARTIFICIAL NEURAL NETWORKS (ANN)

NORMALIZING FLOWS

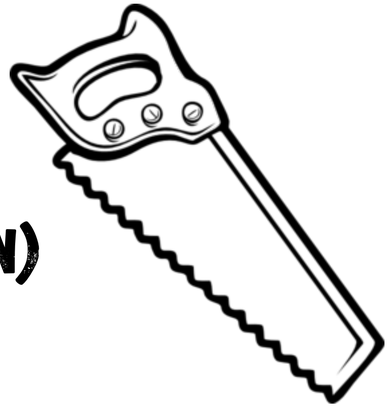


AUTOENCODERS

LINEAR REGRESSION



GENERATIVE ADVERSARIAL NETWORKS (GAN)



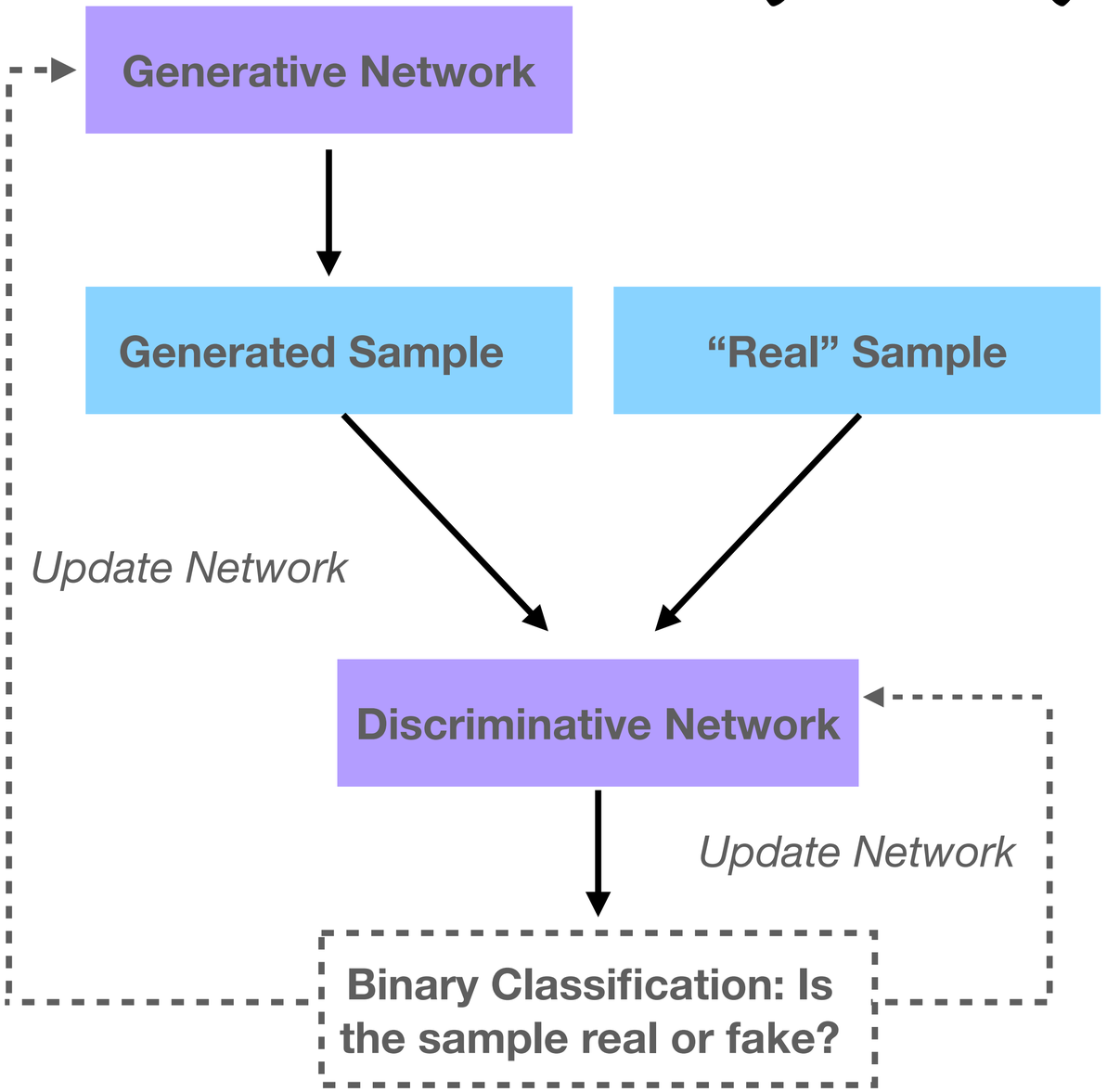
GENERATIVE ADVERSARIAL NETWORKS (GANS)

Two networks compete with one another in a game.

The generative network seeks to fool the discriminative network.

The discriminative network seeks to find the real sample from the generated samples.

Indirect training → generative network never sees the true distribution!

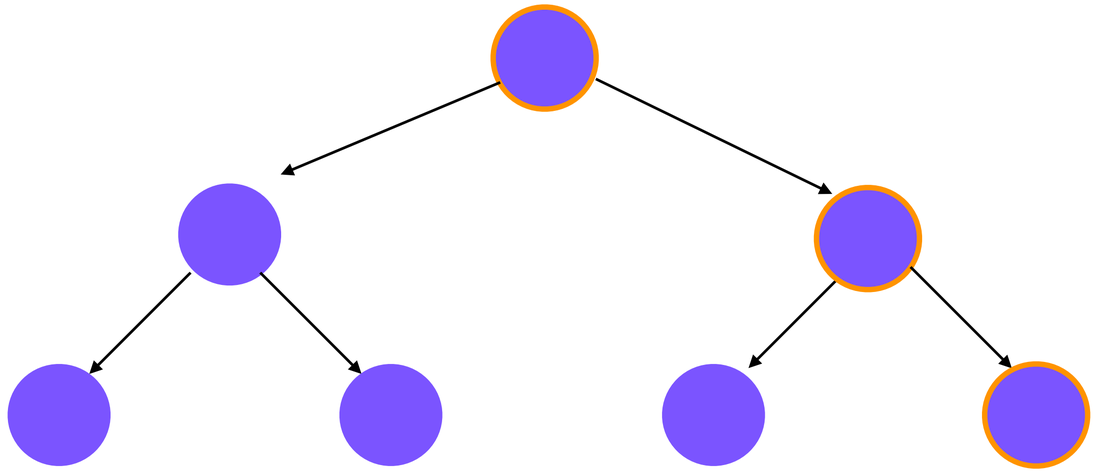


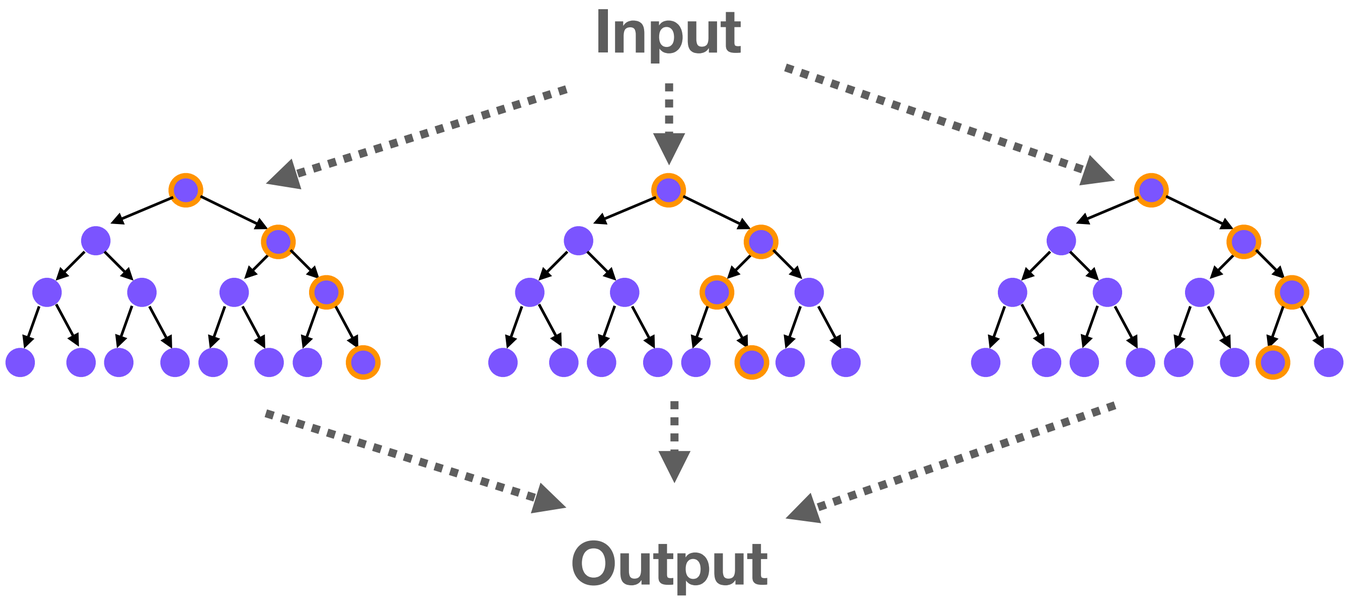
INTRO TO RANDOM FOREST

Random forests are composed of decision trees.

Decision trees are a set of rules organized in a tree structure.

Each node is a rule which subdivides the dataset into two or more parts (think 20 questions).





Output of the random forest is a combination of the output of each of the decision trees.

In training, the algorithm sets up the rules of each decision tree.

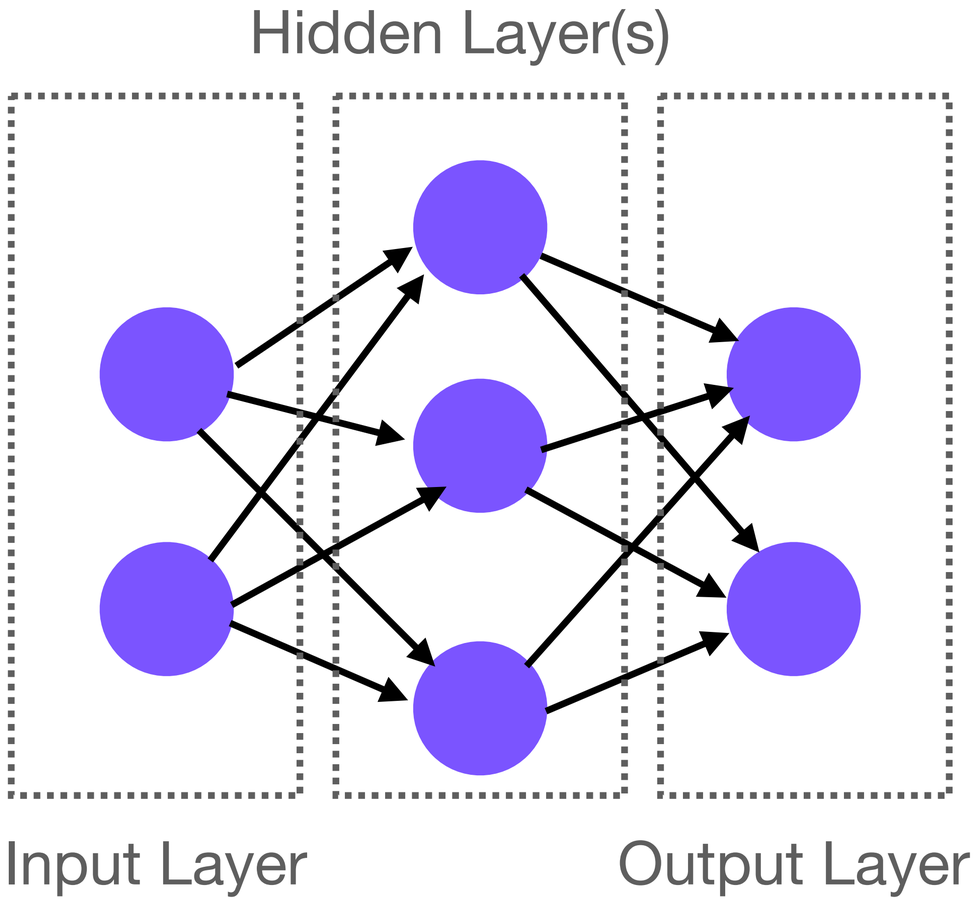
NEURAL NETWORKS

Flow of information happens between nodes.

A weight is associated with each input to a given node.

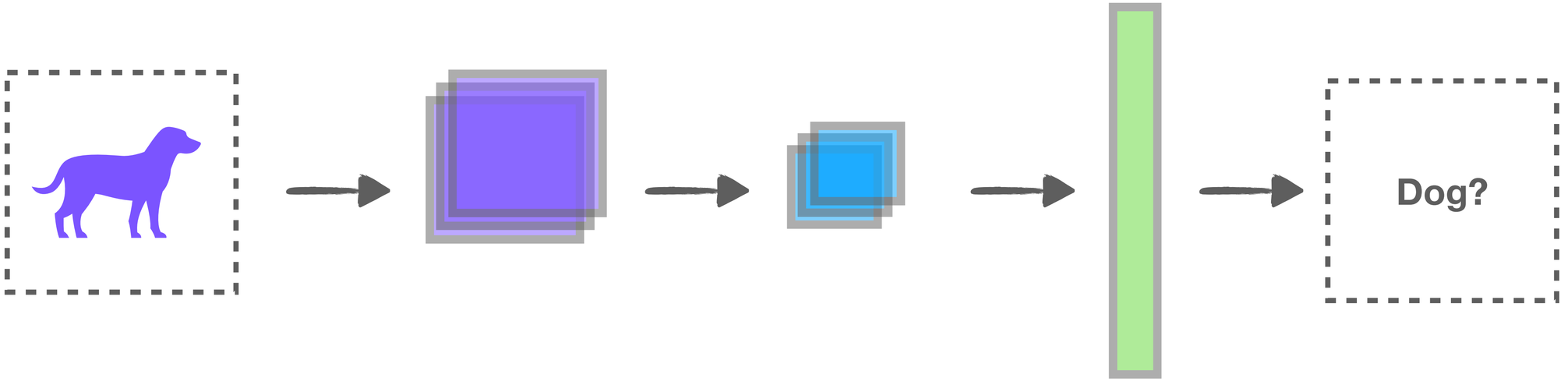
The output of each node is a function of the weighted inputs. The output of a node j, is generally written something like

$$O_j = \sum_{i=0}^{N-1} w_{ij} O_i$$



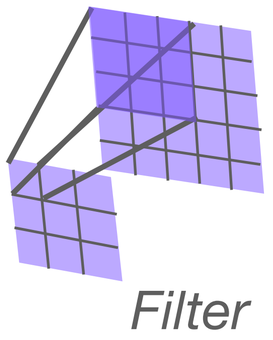
In training we seek to learn the set of weights which minimize the total error of the network.

CONVOLUTIONAL NEURAL NETWORKS (CNNS)



Input Layer | Convolution Layer | Pooling Layer | Fully Connected Layer | Output Layer

→ Key component of a CNN is the **convolution layer**, which (with the help of a filter) will determine if a feature/pattern is present.



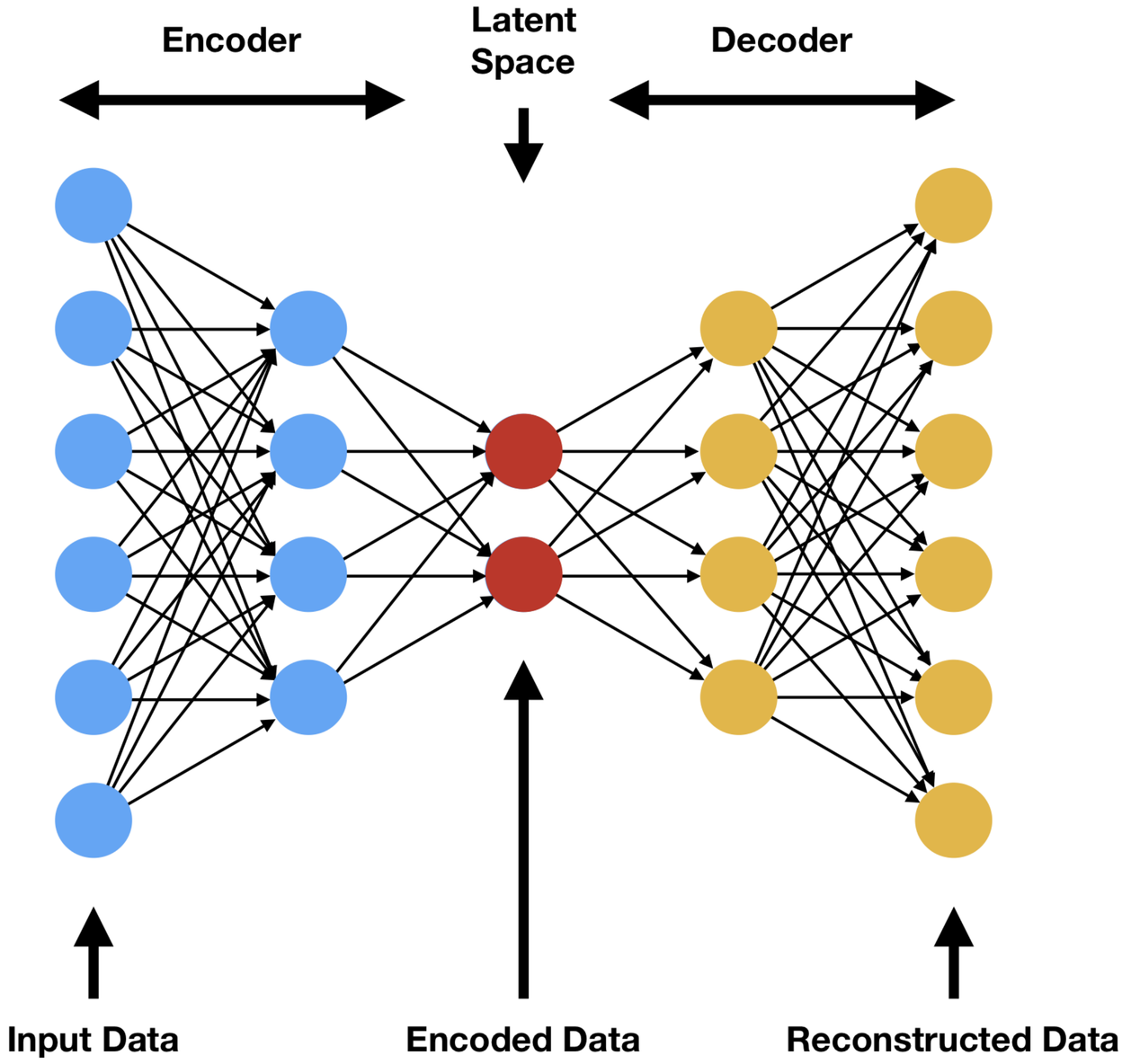
AUTO-ENCODERS

Simple task: NN architecture trained to copy inputs to outputs!

Encoder takes the input and dramatically reduces its complexity via a NN.

Decoder takes the encoded data and reconstructs outputs like the data.

Does not require labeled data as input!

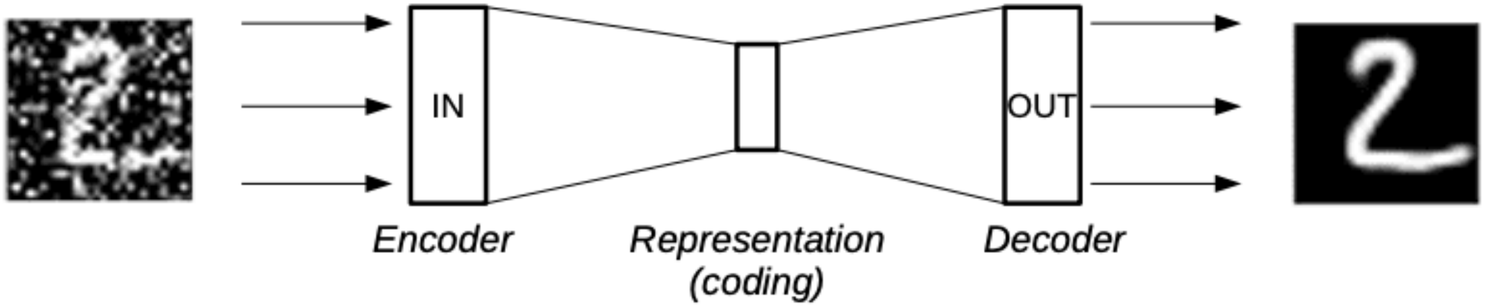


https://www.compthree.com/blog/autoencoder/

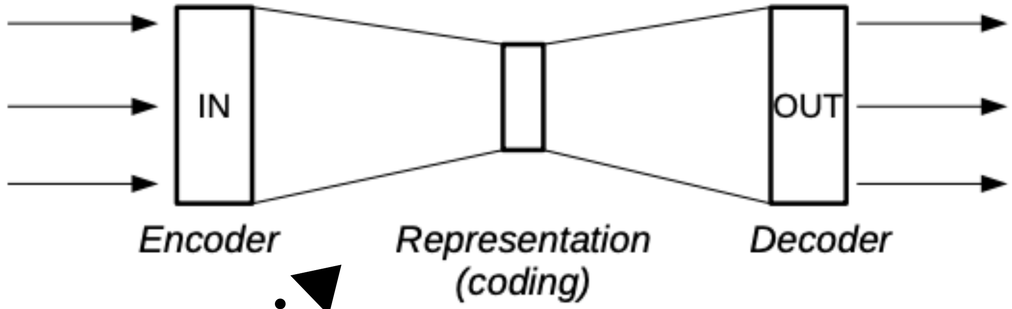
USES OF AUTO-ENCODERS

Used to learn efficient representations of some input data.

1. De-noising inputs



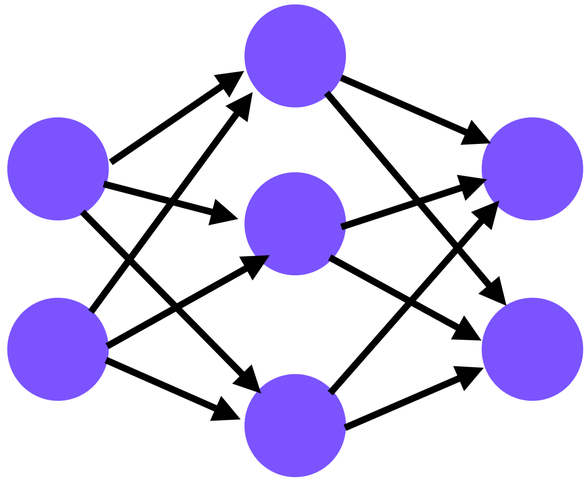
2. Unsupervised learning



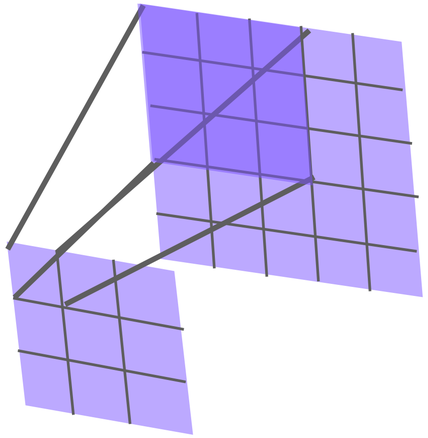
Sort items into classes here

3. Anomaly detections: If you fail to reconstruct data in the decoding step you have an anomaly!

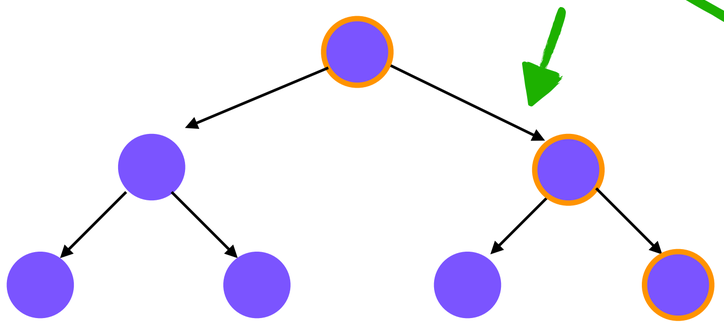
DIFFERENT ALGORITHMS FOR DIFFERENT PROBLEMS!



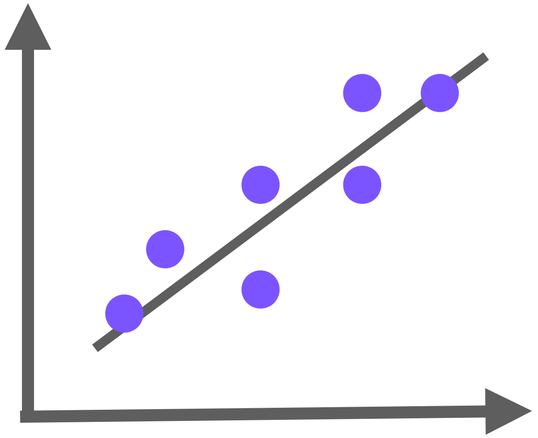
(Shallow or Deep) Neural Networks → *Great for making predictions!*



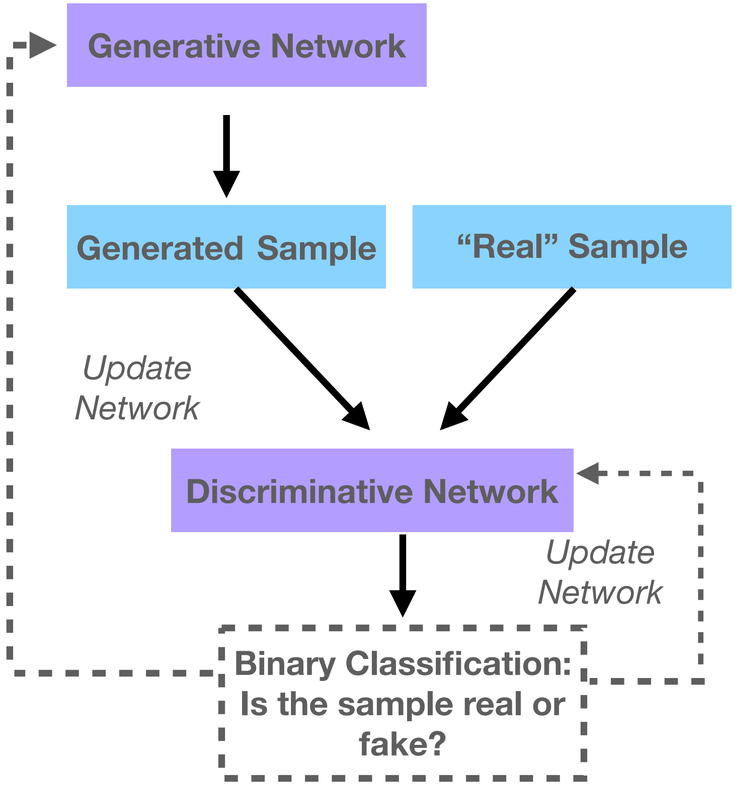
Convolutional Neural Networks (CNNs)→ *Great for image processing!*



Random Forest (Decision Trees)



Linear Regression



Generative Adversarial Networks (GANs) → *Powerful tool for generating samples!*