Summary of the Weizmann workshop: Hammers & Nails - Machine Learning & HEP

Cécile Germain

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

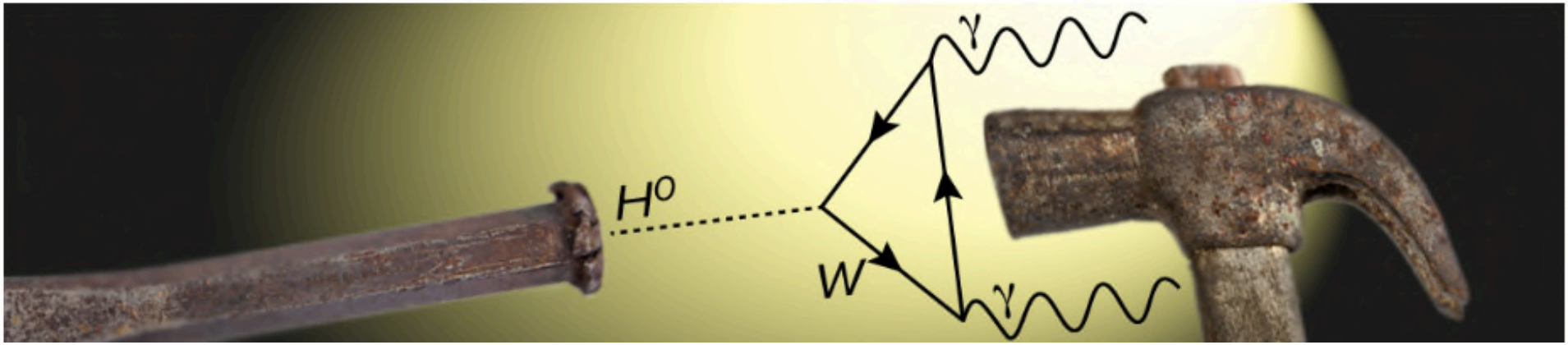
Schwartz/Reisman
Institute for Theoretical Physics

SRitp Workshop

Hammers & Nails - Machine Learning & HEP

July 19-28, 2017

Weizmann Institute of Science, Israel



SUMMARY

Cécile Germain − Université Paris Sud

Disclaimer

Most slides here are extracted from the workshop presentations

Introductions to HEP-ML

- Kyle Cranmer - Broad HEP/ML Framing
- Maurizio Pierini - DSHEP Review
- Amir Farbin - Example Data Sets & Challenges
- Mike Wiliams - Machine Learning in the trigger of LHCb

Reductionist

Maybe AI should start with problems where causal structure is clear and mechanistic models are available?

Emergent

mechanistic models
clear causal structure

descriptive models
unclear causal structure

ecology

nuclear & particle physics

astrophysics

health

climate

epidemiology

language

cosmology

perception

connectome

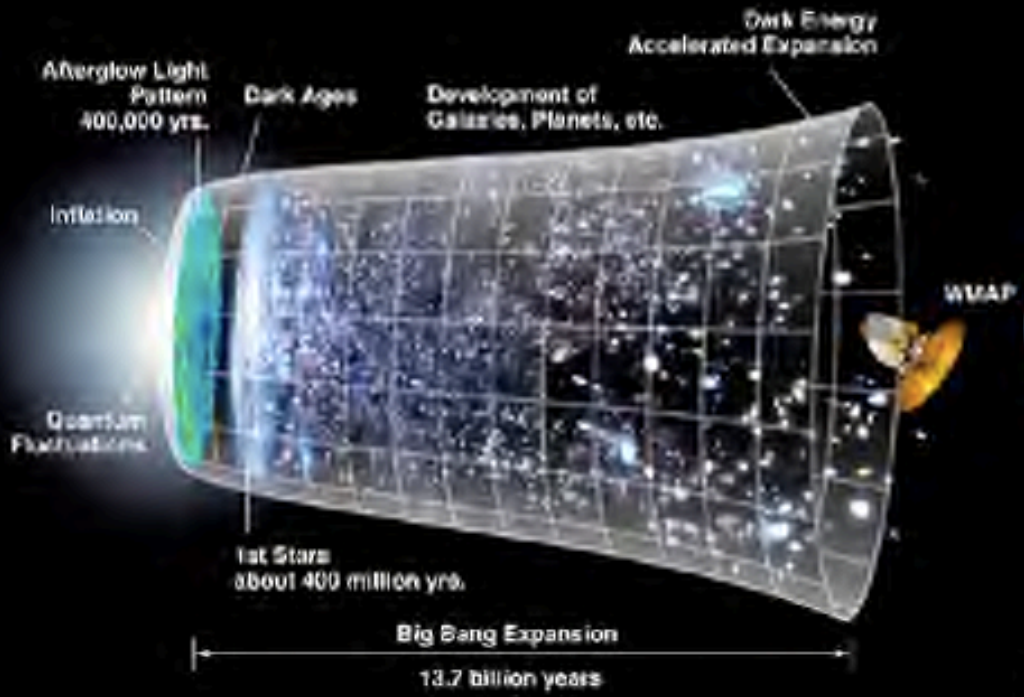
lattice simulations

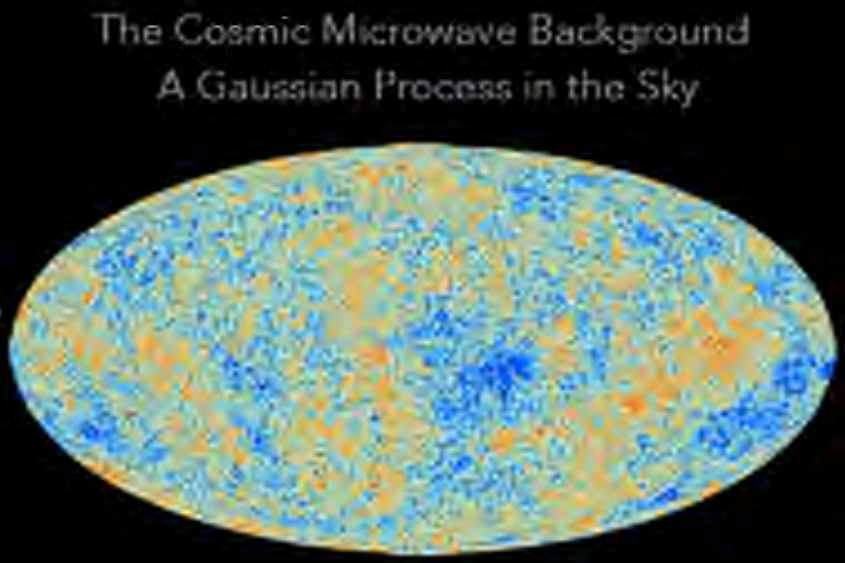
protein folding

psychology

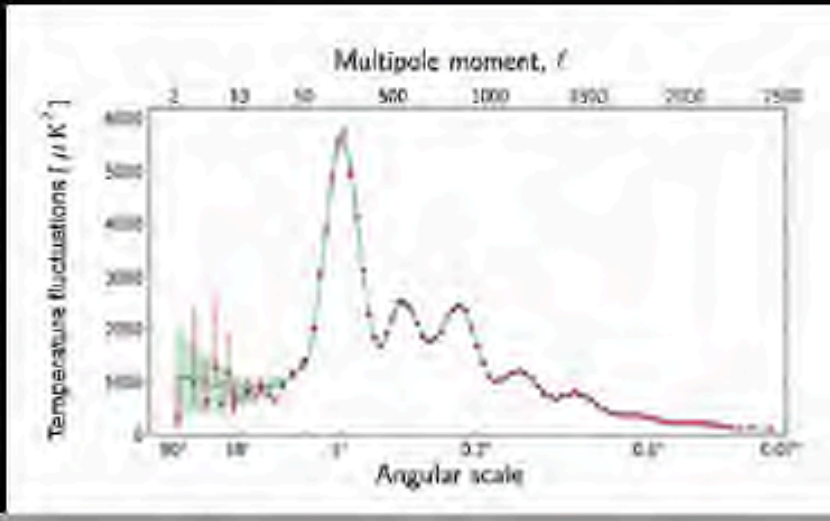
systems biology

COSMOLOGY: 6 PARAMETERS

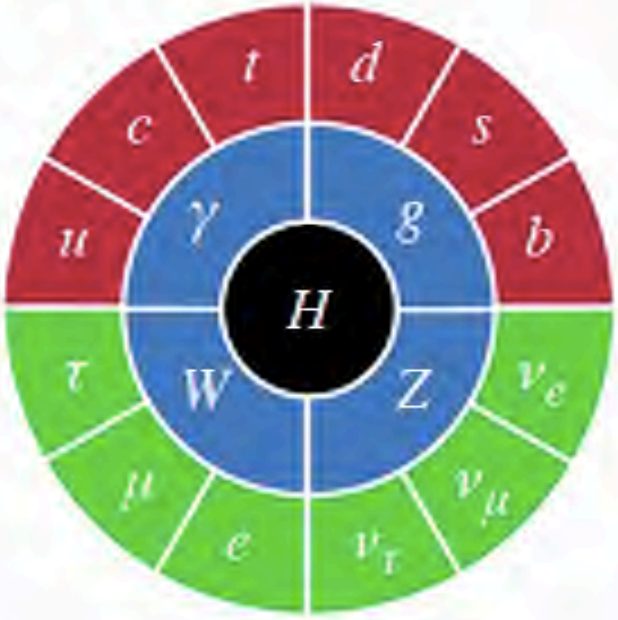




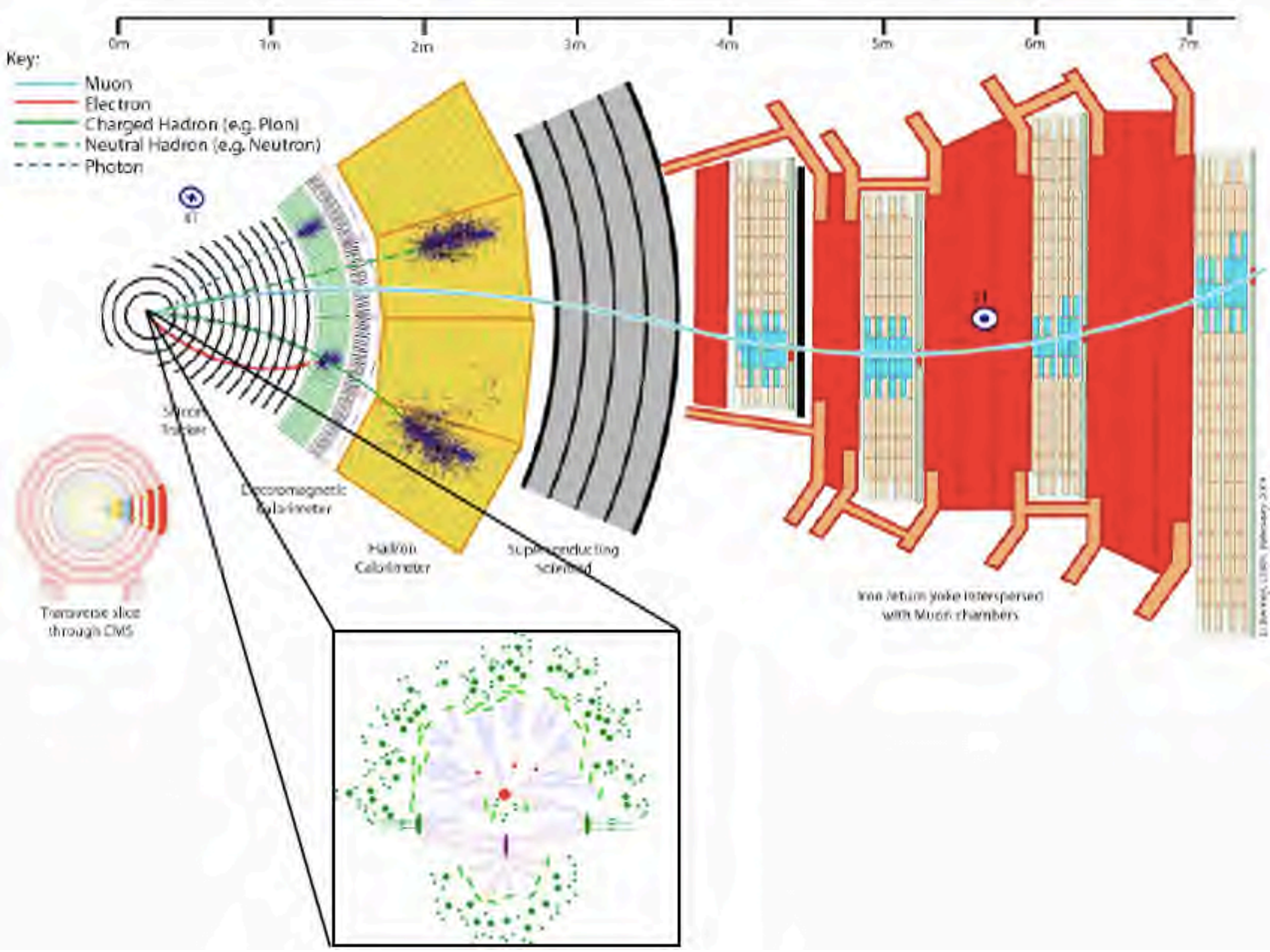
Symbol	Description	Value
$\Omega_b h^2$	Physical Baryon Density Parameter	0.02230 ± 0.00014
$\Omega_c h^2$	Physical Dark Matter Density Parameter	0.1188 ± 0.0010
T_0	Age Of The Universe	13.799 ± 0.021 × 10^9 Years
N_s	Scalar Spectral Index	0.9667 ± 0.0040
Δ_2	Curvature Fluctuation Amplitude	2.441 ± 0.09 × 10^{-9}
T	Reionization Optical Depth	0.066 ± 0.012



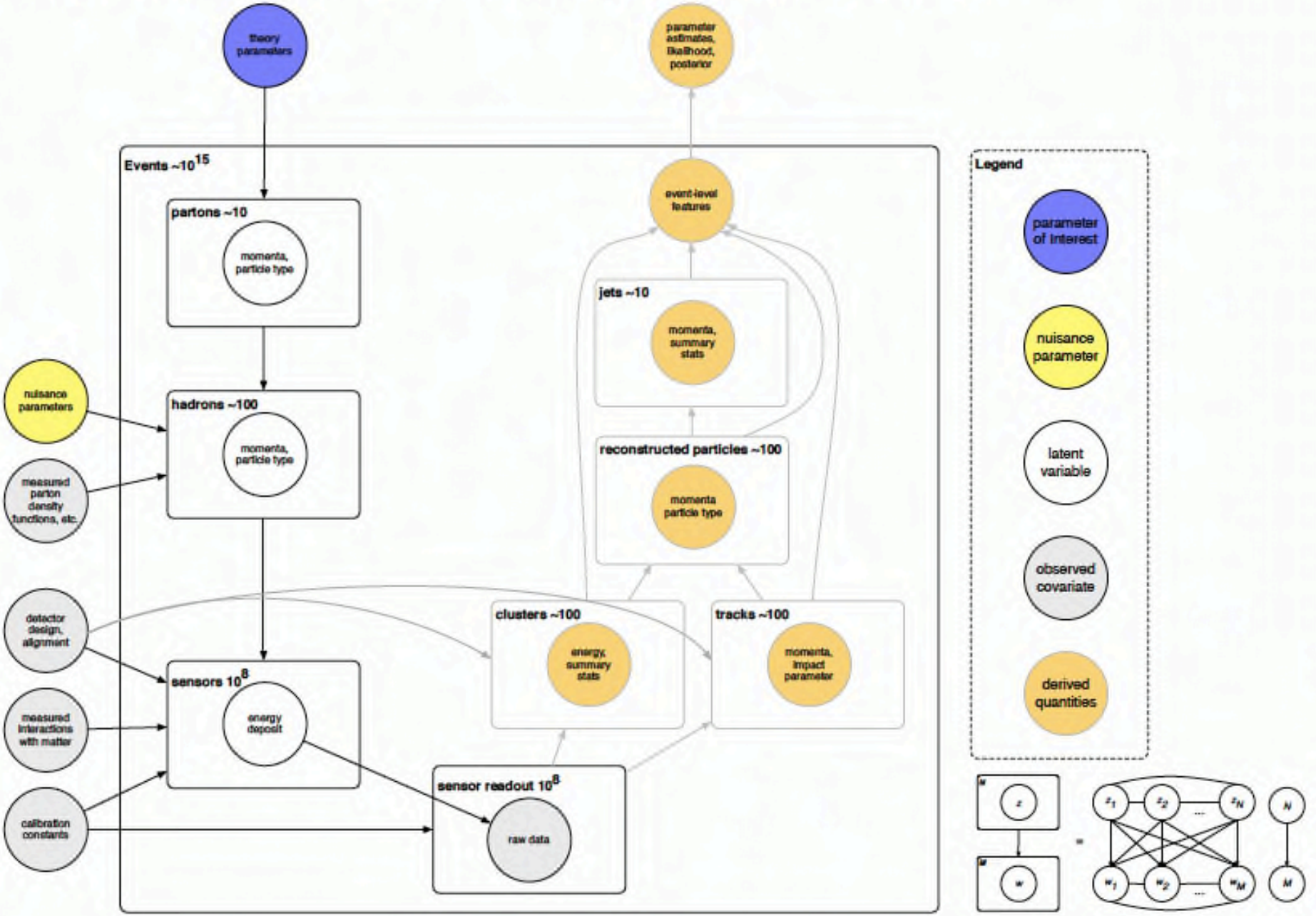
PARTICLE PHYSICS: 19 PARAMETERS



Symbol	Description	Value
m_e	Electron mass	511 keV
m_μ	Muon mass	105.7 MeV
m_τ	Tau mass	1.78 GeV
m_u	Up quark mass	1.9 MeV
m_d	Down quark mass	4.4 MeV
m_s	Strange quark mass	87 MeV
m_c	Charm quark mass	1.32 GeV
m_b	Bottom quark mass	4.24 GeV
m_t	Top quark mass	172.7 GeV
θ_{12}	CKM 12-mixing angle	13.1°
θ_{23}	CKM 23-mixing angle	2.4°
θ_{13}	CKM 13-mixing angle	0.2°
δ	CKM CP-violating Phase	0.995
g_1	U(1) gauge coupling	0.357
g_2	SU(2) gauge coupling	0.652
g_3	SU(3) gauge coupling	1.221
θ_{QCD}	QCD vacuum angle	~0
v	Higgs vacuum expectation value	246 GeV
m_H	Higgs mass	125 GeV


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS

FULL SIMULATION + RECONSTRUCTION



Physics Measurements & Searches for new particles

=

Likelihood-free Inference with

Simulation-based Implicit Models

FIXED VS. VARIABLE LENGTH, SPARSE VS. DENSE

Low-level detector-level sensor data:

- fixed length ~10^8 real-valued sensor measurements
- raw sensor data is often very sparse (no energy deposited in many of the sensors)

Mid-level "reconstructed objects":

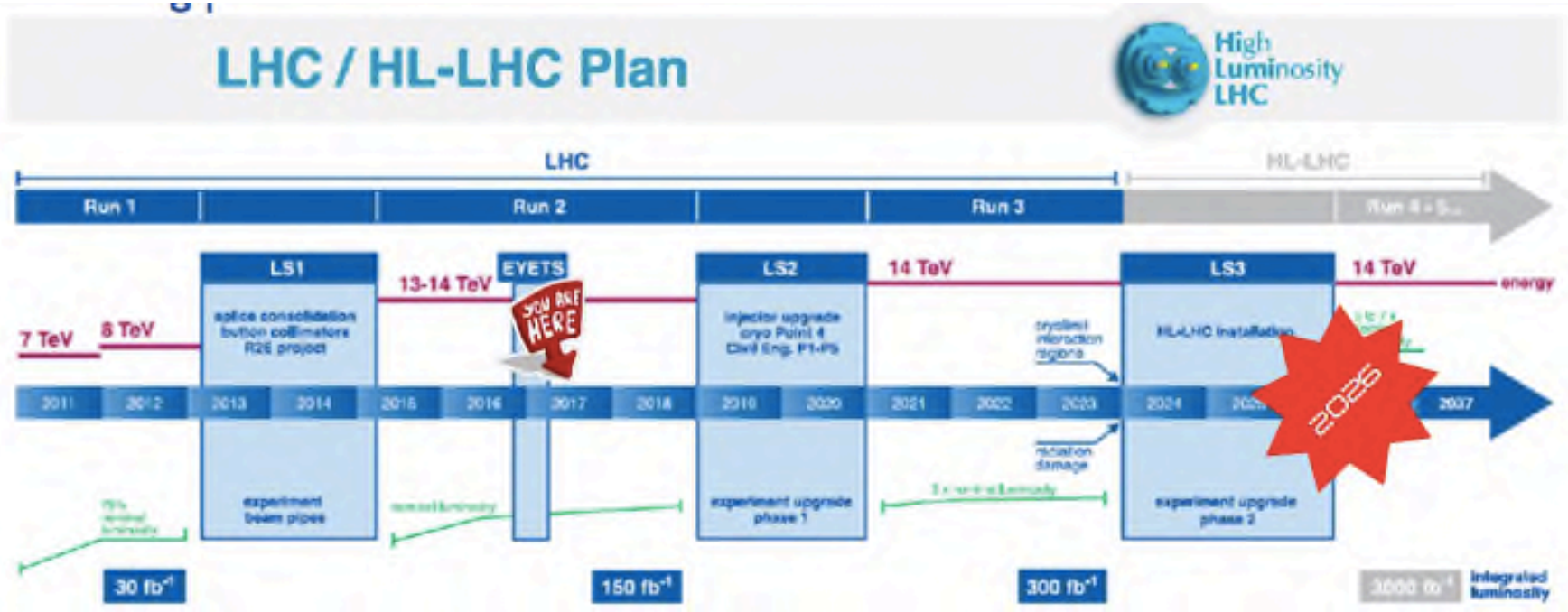
- variable number of them, typically sorted by "pT" / "transverse momentum"
- like attention mechanism / segmentation + engineered features
- relatively modest data associated to each object

High-level event quantities

- typically fixed length summary statistics / engineered features
- dimensionality of data manifold at this level is often same as dimensionality of features (eg. "dense and full rank")

The nails

- Data Analysis: better
- Simulation: faster
- Reconstruction: better and faster
- Trigger: increase throughput
- Data quality monitoring: automate



Data Analysis

- Objectives:
  - ***Searches*** (hypothesis testing): Likelihood Ratio Test (Neyman-Pearson lemma)
  - ***Measurements***: Maximum Likelihood Estimate
  - ***Limits*** (confidence intervals): Also based on Likelihood

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

- ***Likelihood***

$$p(\{x\}|\theta) = \mathrm{Pois}(n|\nu(\theta)) \prod_{e=1}^{n} p(x_e|\theta)$$

  - *n* Independent Events (*e*) with Identically Distributed Observables ({*x*})
  - Significant part of Data Analysis is ***approximating the likelihood*** as best as we can.

Hammer

Traditional supervised learning − classification at moderate scale

- Glen Cowan - Statistical Treatment of DATA in HEP
- Clayton Scott - Weakly Supervised Learning
- Tobias Golling - Flavour Tagging and ML
- Rita Osadchy - Hinge-Minimax Classifiers

Q. Is jet tagging (classification?) a potential Challenge? well defined, interesting testbed for end-to-end vs hiérarchique.

Optimal test for discovery

Likelihood function is:

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \prod_{i=1}^{n} \left[\frac{\mu s}{\mu s + b} p(\mathbf{x}_i|s) + \frac{b}{\mu s + b} p(\mathbf{x}_i|b) \right]$$

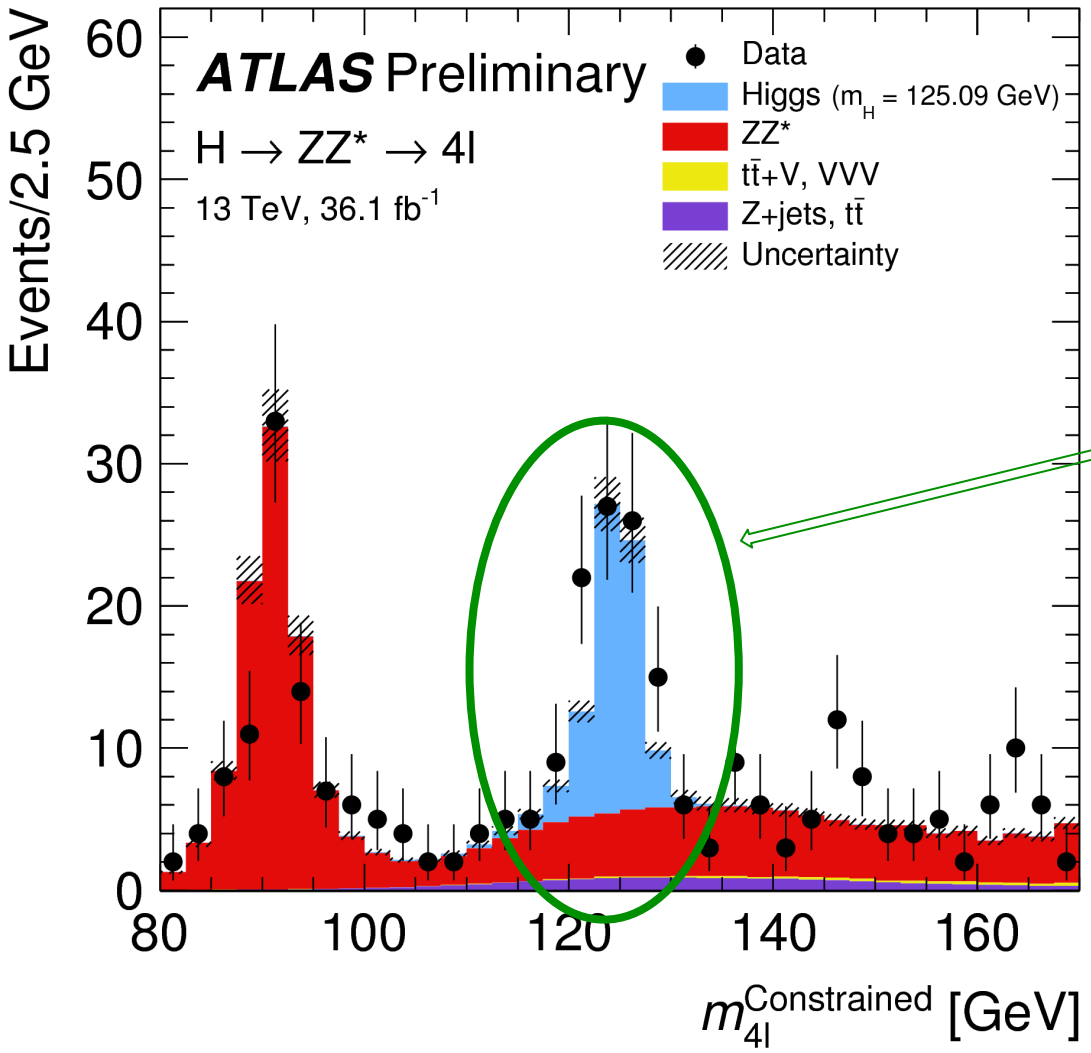
Neyman-Pearson say optimal statistic for test of $\mu = 0$ versus alternative of nonzero μ is

$$\frac{L(\mu)}{L(0)} = e^{-\mu s} \prod_{i=1}^{n} \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i|s)}{p(\mathbf{x}_i|b)} \right)$$

or take log and drop constant term $-\mu s$,

$$Q = \sum_{i=1}^{n} \ln \left(1 + \frac{\mu s}{b} \frac{p(\mathbf{x}_i|s)}{p(\mathbf{x}_i|b)} \right)$$

The Goal



The Higgs Boson!

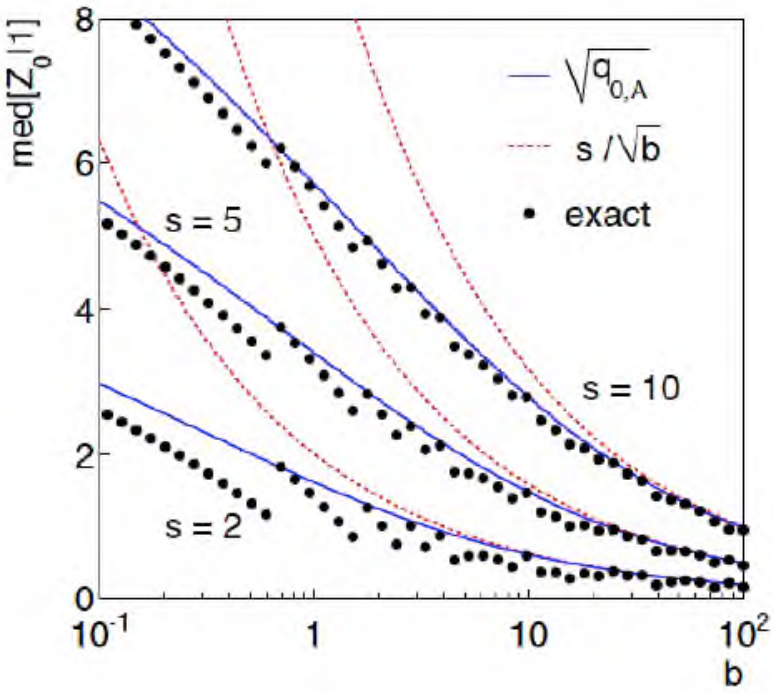
- Want to reject the "No Higgs" hypothesis (red + purple)
- Hopefully measure the size of the Signal

Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

(a) $\dfrac{s}{\sqrt{b}}$

(b) Profile likelihood ratio test & Asimov: $\sqrt{2\left((s+b)\ln\left(1+\dfrac{s}{b}\right)-s\right)}$



uncertainty in b, σ_b:

atio test & Asimov:

$$\left.\left.\frac{)}{\frac{2}{b}}\right]-\frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)\right]^{1/2}$$

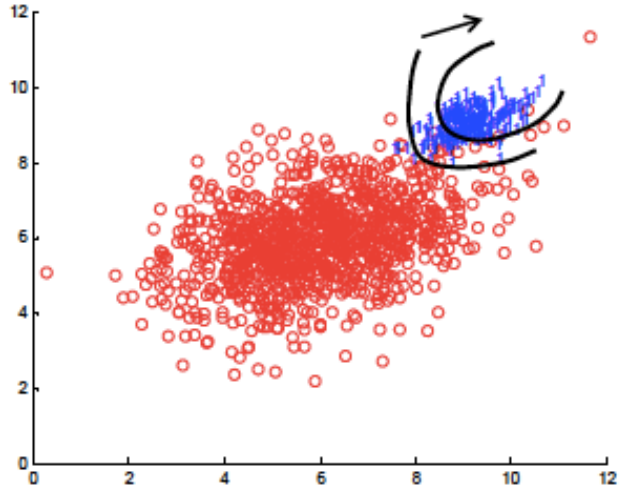
Neyman-Pearson

- False positive/negative rates:

$$R_0(f) = P(f(X) = 1 \,|\, Y = 0)$$
$$R_1(f) = P(f(X) = 0 \,|\, Y = 1)$$

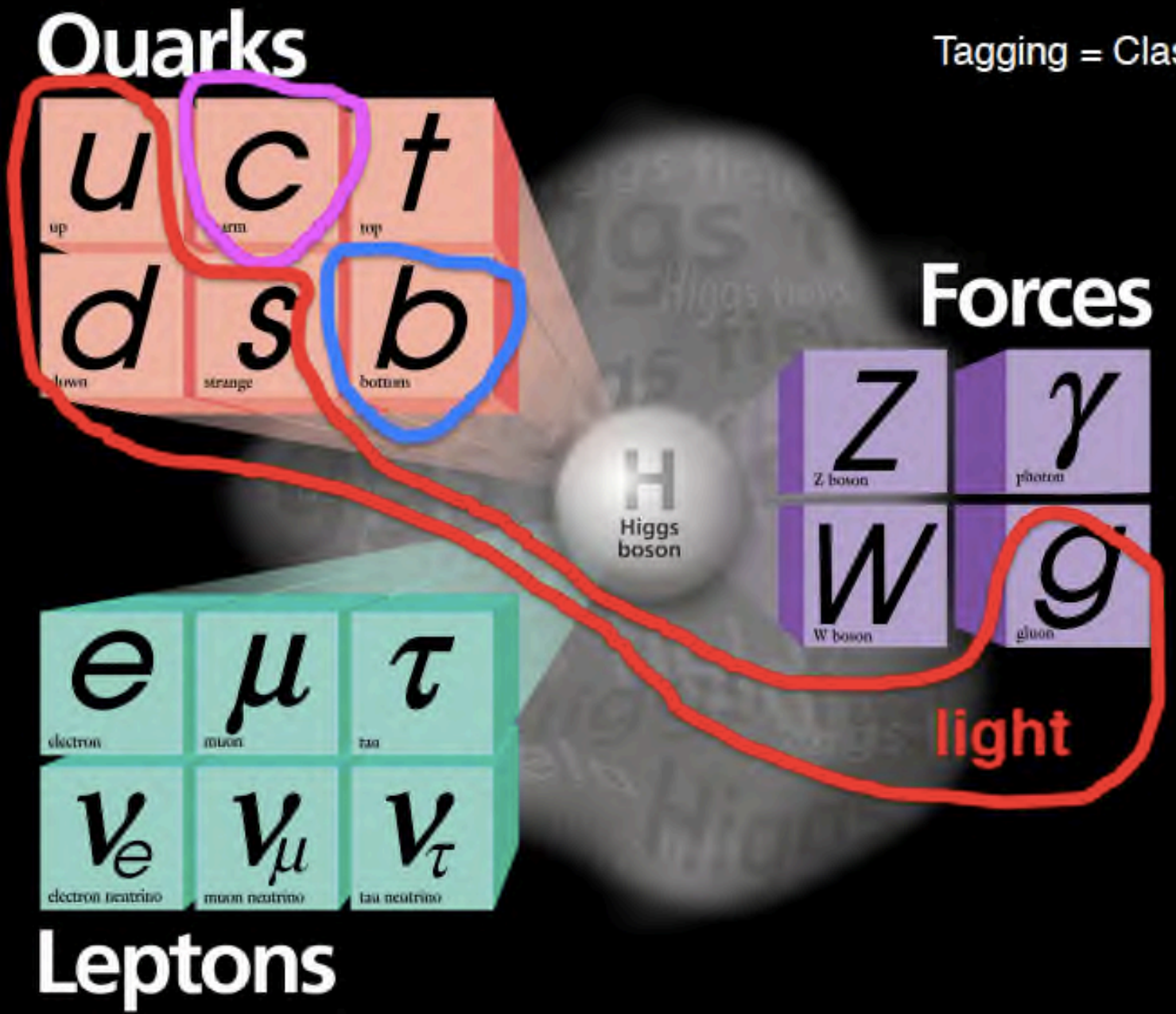
- Given $\alpha \in (0, 1)$, the **Neyman-Pearson classifier** solves

$$\begin{aligned} \min \quad & R_1(f) \\ \text{s.t.} \quad & R_0(f) \leq \alpha \end{aligned}$$



- Solution also a likelihood ratio test
- Advantages:
  - Class proportions in test and training data need not be the same
  - Imbalanced data

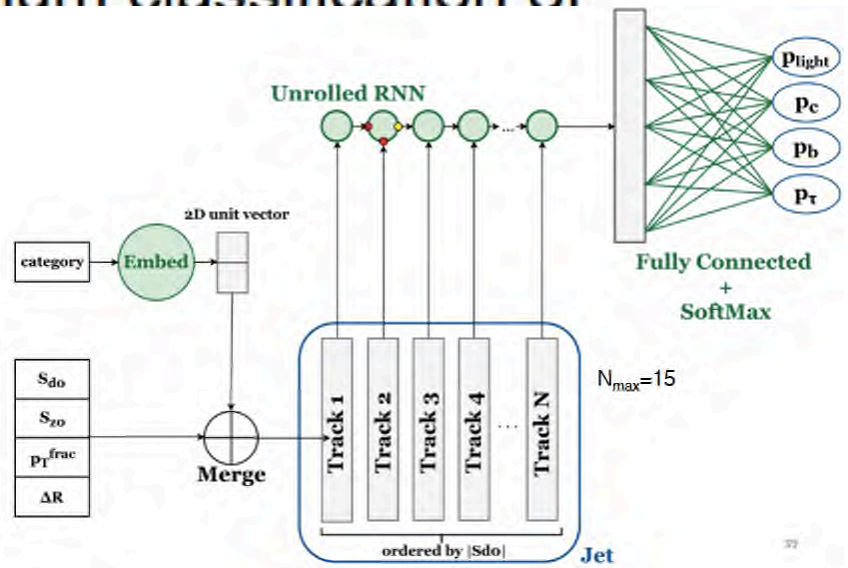
Flavor tagging in a nutshell: classify jets in b, c, light

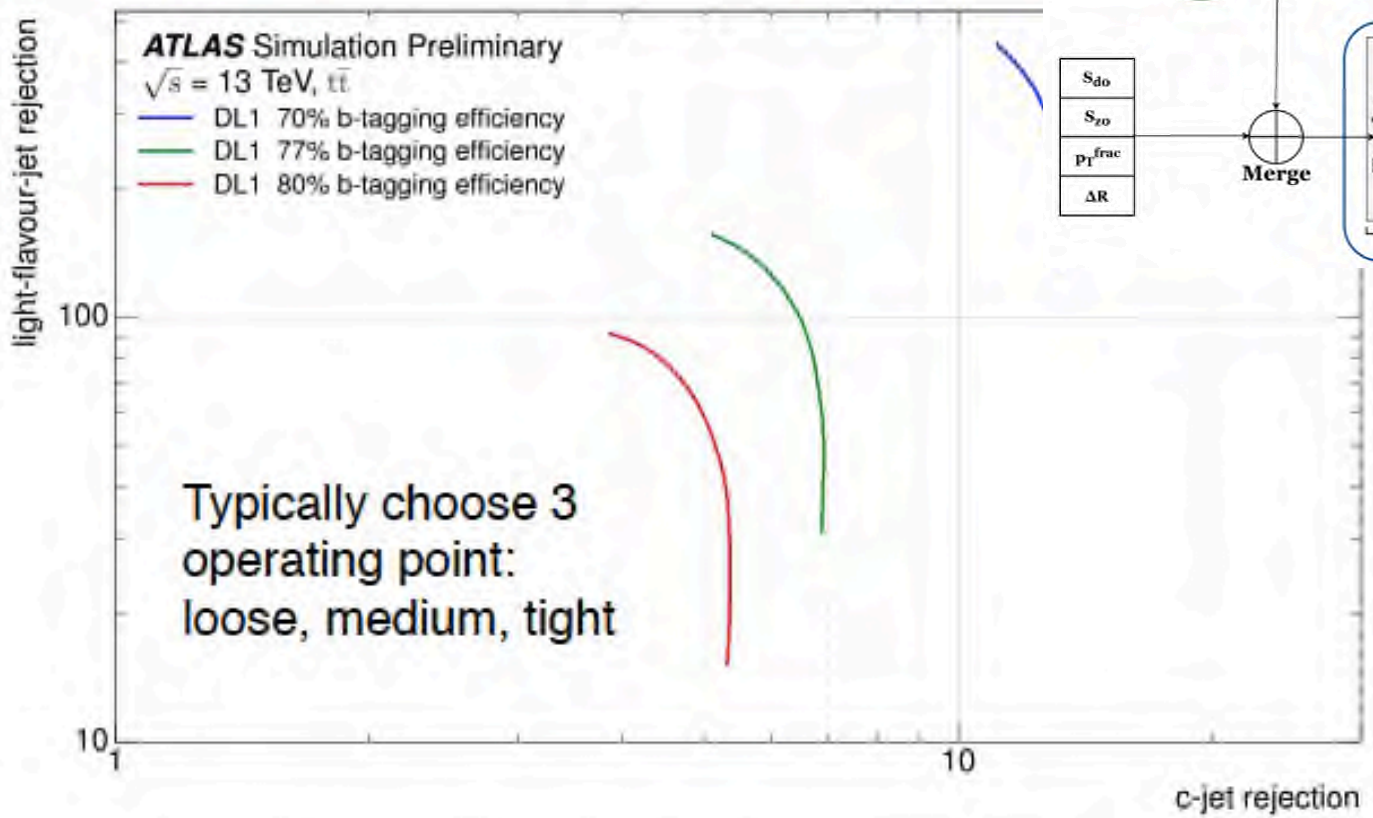


DL1 b-tagging

- 3 outputs ⇒ how to get optimum classification of signal and background?

b-tagging: $DL1b f_{\text{c-jets}} = \ln\left(\frac{p_b}{f_{\text{c-jets}} \cdot p_c + (1 - f_c}\right)$





Vary $f_{\text{c-jets}}$ parameter and adjust cut on discriminant to hold b-tagging efficiency constant

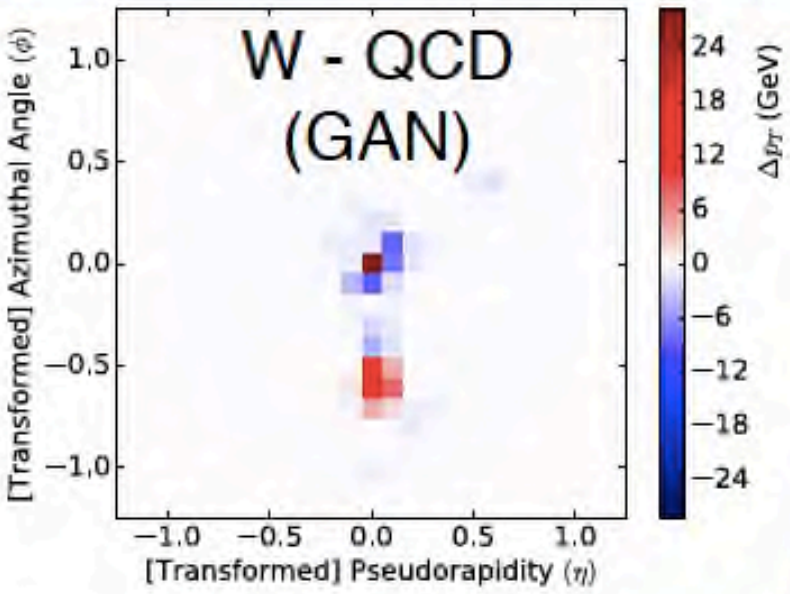
Why go Deep?

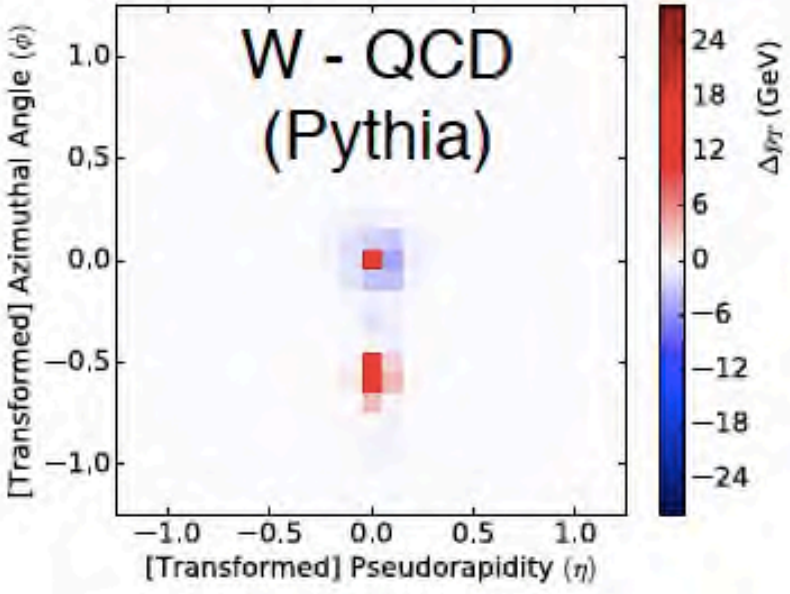
- ***Better Algorithms***
  - DNN-based classification/regression generally **out perform** hand crafted algorithms.
  - In some cases, it may provide a ***solution*** where ***algorithm approach doesn't exist or fails***.
  - ***Unsupervised learning***: make sense of complicated data that we don't understand or expect.
- ***Easier Algorithm Development***: ***Feature Learning*** instead of *Feature Engineering*
  - Reduce time physicists spend writing developing algorithms, ***saving time and cost***. (e.g. ATLAS > $250M spent software)
  - Quickly perform performance ***optimization*** or ***systematic studies***.
- ***Faster Algorithms***
  - After training, DNN inference is often *faster* than sophisticated algorithmic approach.
  - DNN can ***encapsulate expensive computations***, e.g. Matrix Element Method.
  - ***Generative Models*** enable fast simulations.
  - ***Already parallelized*** and optimized for GPUs/HPCs.
  - ***Neuromorphic*** processors.

ML = DNN?

- Convolutional NN: high dimensional
  - Classification, regression : tracking, flavour tagging, neutrinos
    - Amir Farbin, Example Data Sets & Challenges
  - Generative adversarial: simulation
    - Ben Nachman, GAN applications (in particular calorimeter simulation)

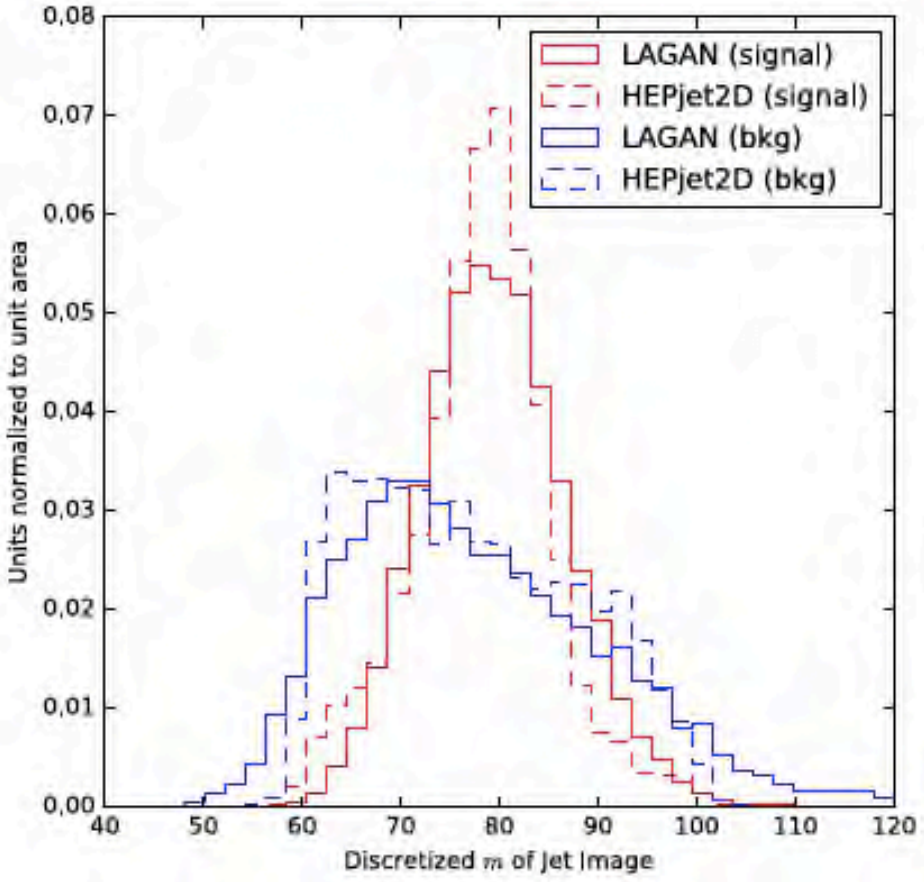
Locally Aware GAN (LAGAN)



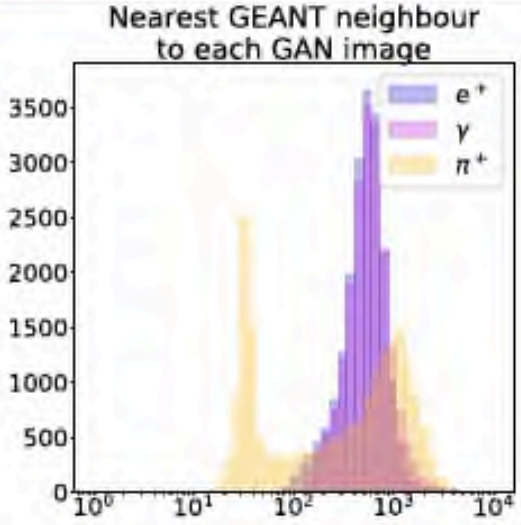


Unlike `natural images', we have physically meaningful 1D manifolds (here, jet mass)

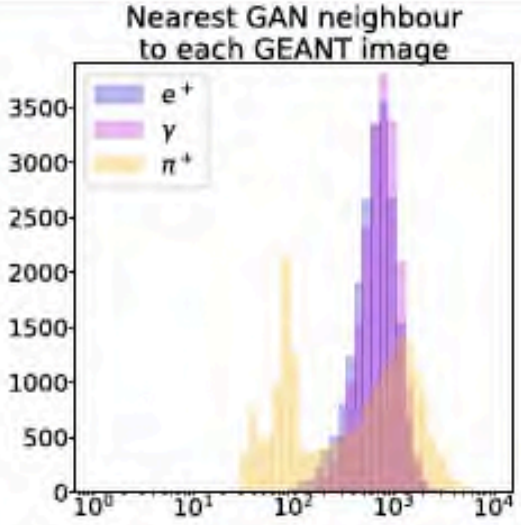
And no translational invariance



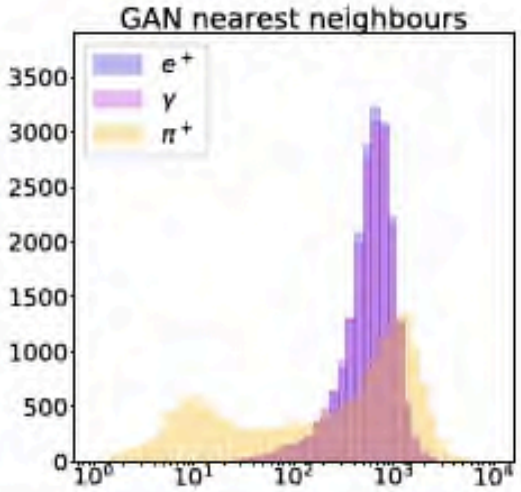
"Overtraining"



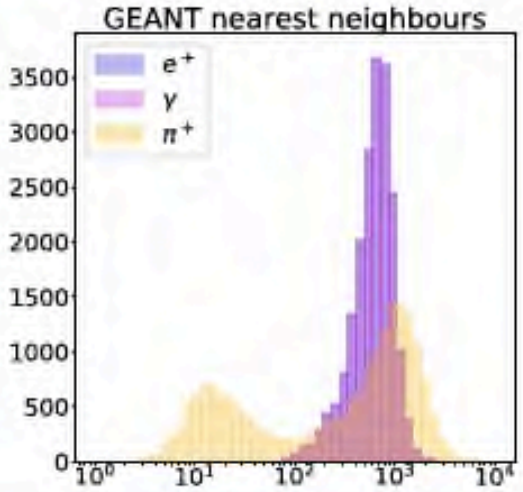
not memorizing



A key challenge in training GANs is the diversity of generated images. This does not seem to be a problem for CaloGAN.



no mode collapse

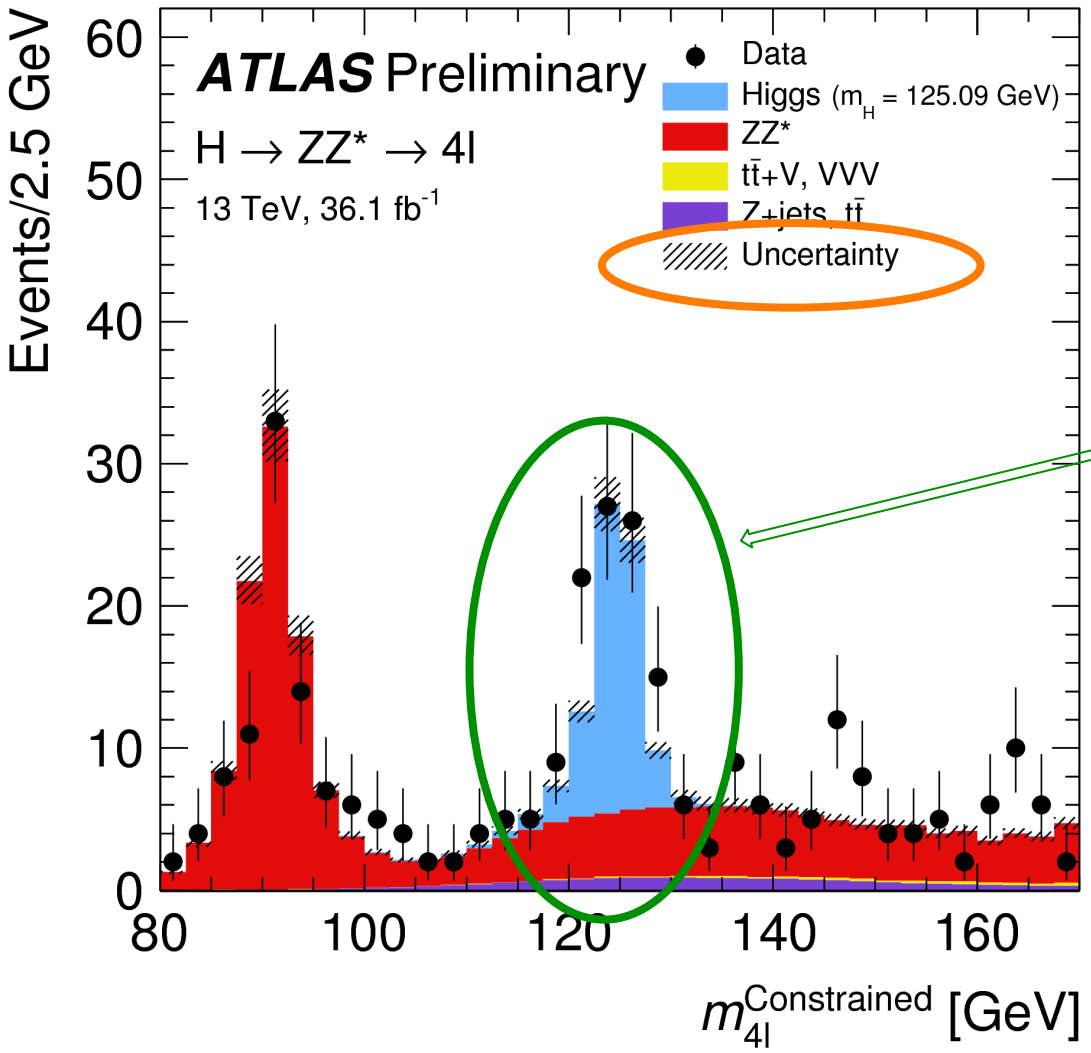


ML = DNN?

- Adversarial MLP networks: model-free profiling
  - Daniel Whiteson, Gaining physics insight from ML
  - Michael Kagan Domain Adaptation & Systematics in HEP

- Recurrent: integrating some domain knowledge
  - Kyunghyun Cho. Sequential and Recursive Learning, part 2
  - Jean Roch Vlimant, ML Techniques in Tracking

The Goal

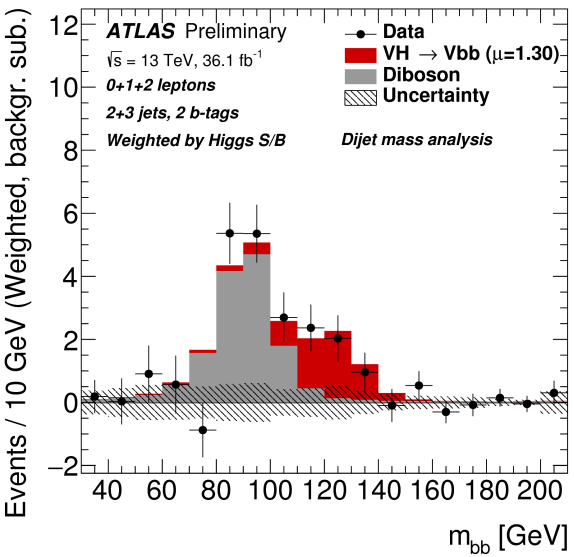


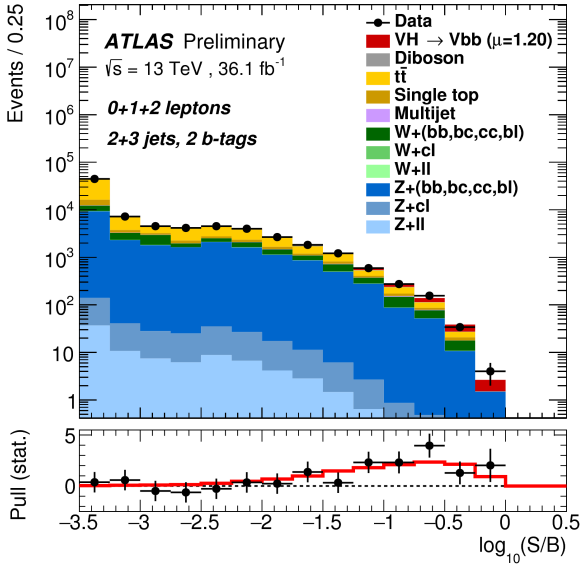
The Higgs Boson!

- And we need to know how wrong these predictions could be… Systematic Uncertainties

Sources of Systematic Uncertainty

- Experimental
  - How well we understand the detector and our ability to identify / measure properties of particles given that they are present
- Theoretical
  - How well we understand the underlying theoretical model of the physical processes
- Modeling uncertainties
  - How well can I constrain backgrounds in alternate data measurements / control regions

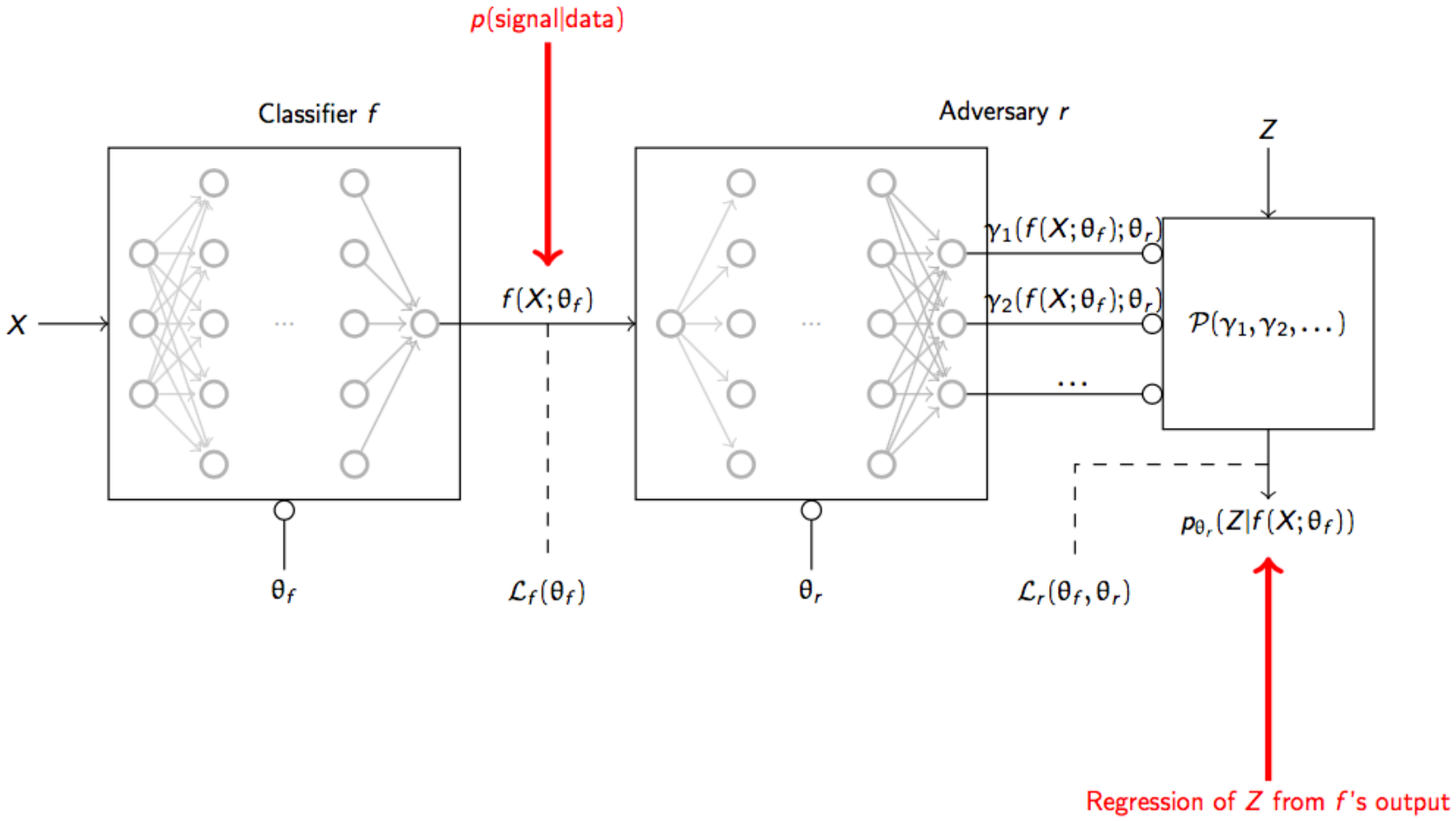




Source of uncertainty		σ_μ
Total		0.39
Statistical		0.24
Systematic		0.31
Experimental uncertainties		
Jets		0.03
$E_{\mathrm{T}}^{\mathrm{miss}}$		0.03
Leptons		0.01
b-tagging	b-jets	0.09
	c-jets	0.04
	light jets	0.04
	extrapolation	0.01
Pile-up		0.01
Luminosity		0.04
Theoretical and modelling uncertainties		
Signal		0.17
Floating normalisations		0.07
Z+jets		0.07
W+jets		0.07
$t\bar{t}$		0.07
Single top-quark		0.08
Diboson		0.02
Multijet		0.02
MC statistical		0.13

Adversarial Networks

G. Louppe, M. K., K. Cranmer, arXiv:1611.01046



- Adversary is built to predict the value of Z given the classifier output
  - If adversary can predict Z, than there is information in f(…) about nuisance parameter, i.e. f(…) and Z are correlated in some way

DNN = “Universal” hammer? Of course not

- No free lunch results
  - We cannot learn solely from examples without prior assumptions (inductive bias).
  - Universal learning is impossible (no poly-time learning rule)

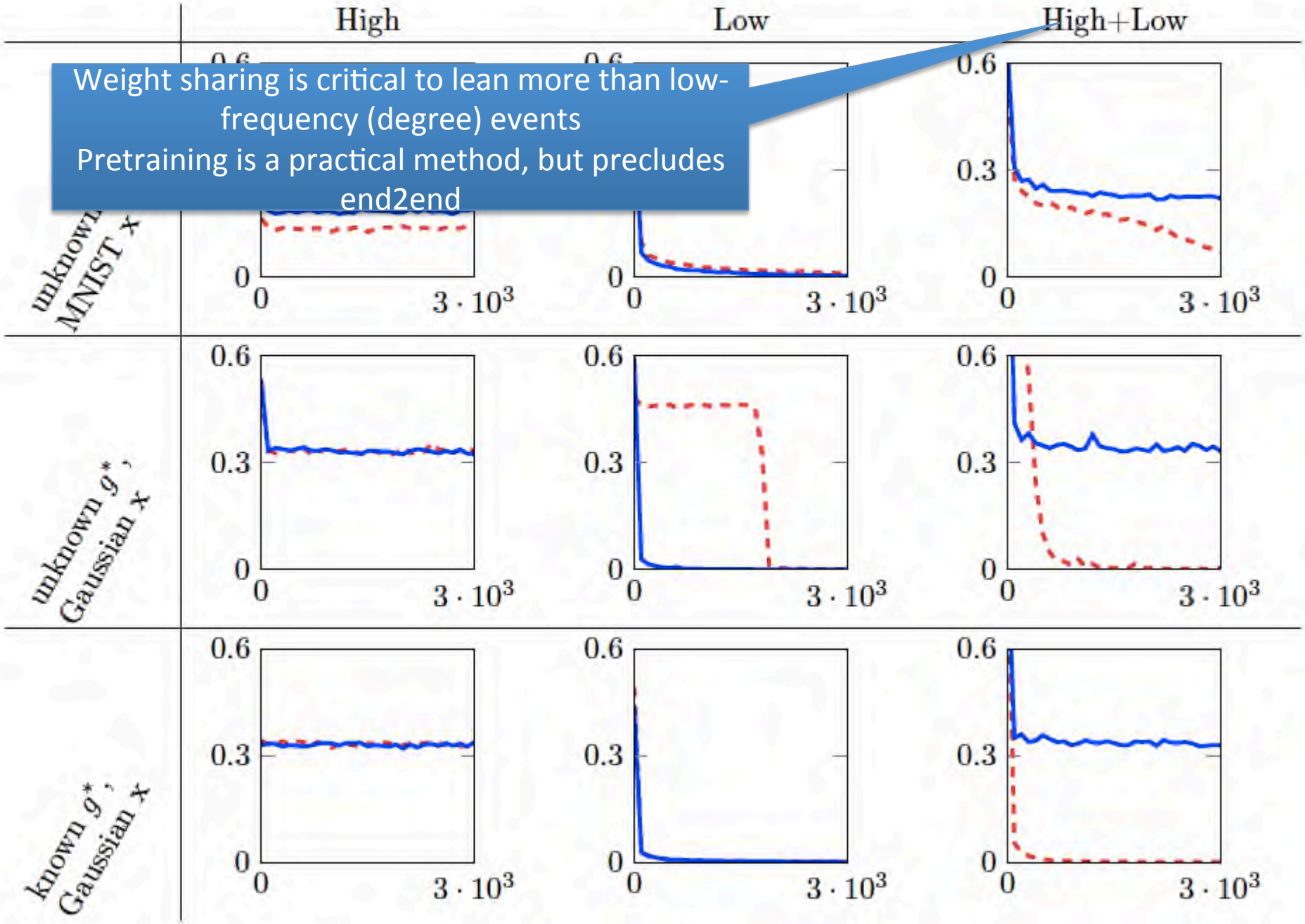
DNN = versatile hammer

- Theoretical frameworks for analyzing
  - The architecture-application interaction for CNN
    - Amnon Shashua, Expressive Efficiency and Inductive Bias of Convolutional Networks: Analysis & Design via Hierarchical Tensor Decompositions.
  - The relations between optimization and generalization error: stochastic optimization as generalized learning, and rescaling-invariant gradient descent Path-SGD - More directly dependent on the **functions computed by the network**, not the vector of weights
    - Nati Srebro Universality, Implicit and explicit Biases in Learning, including Deep Learning
    - Shai Shalev Deep Learning: Successes and Failures
  - The learning process as information compression
    - Naftali TishbyNaftali Tishby, The statistical physics of Deep Learning

Questions about Efficiency and Inductive Bias

- Depth Efficiency: deep ConvNets are (exponentially) Efficient compared to shallow networks
- Pooling scheme affects inductive bias in an Efficient manner
- ConvNets with Overlapping convolution are Efficient compared to non-overlapping ones.
- Modern connectivity schemes (split/merge/skip) are Efficient compared to standard feed-forward (LeNet, AlexNet,..).
- Layer width distribution affects inductive bias in an Efficient manner.

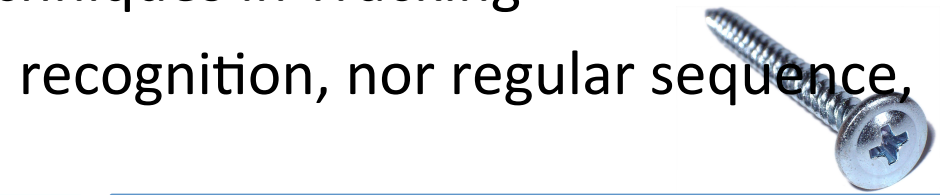
Shai Shalev Deep Learning: Successes and Failures

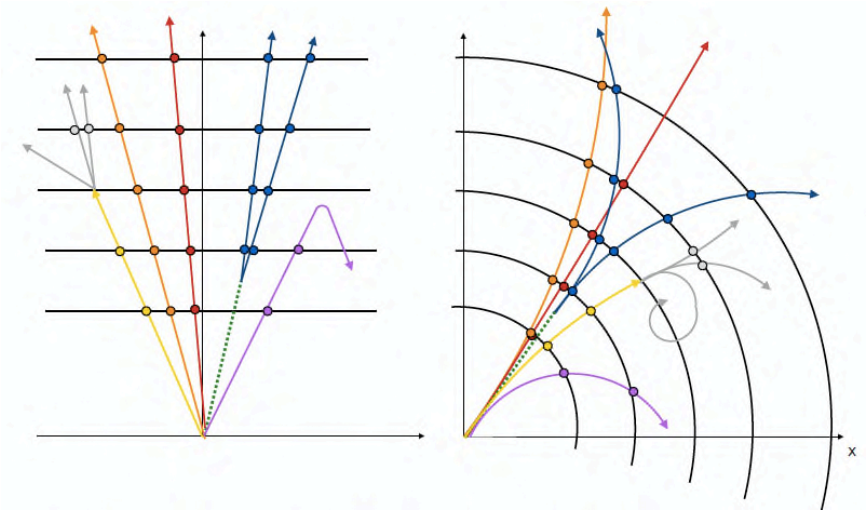


Tracking

- Andreas Salzburger - Tracking Challenge
- Jean Roch Vlimant - ML Techniques in Tracking

Tracking is not typical pattern recognition, nor regular sequence, nor usual clustering.



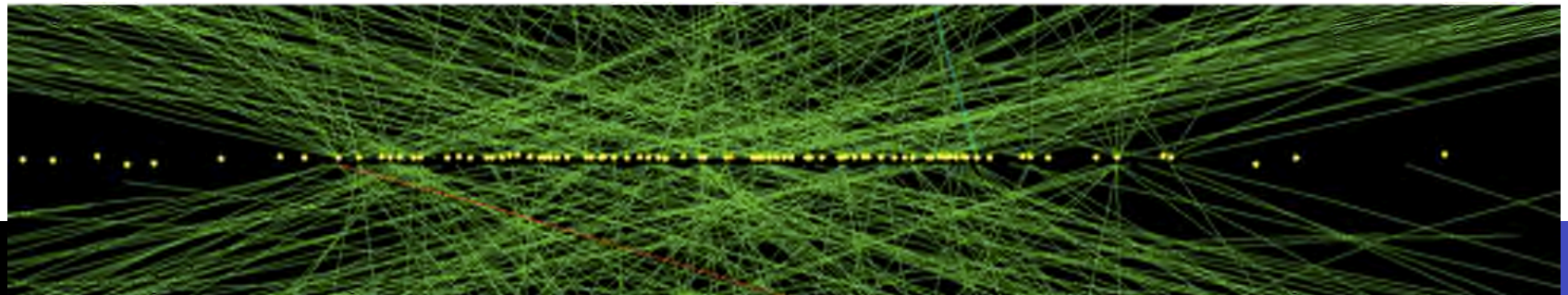


HEP.TrkX Project

➢ Pilot project funded by DOE ASCR and COMP HEP, Part of HEP CCE

➢ Mission

Explore deep learning techniques for track formation



Best hammer so far

Shai Ben David - How Far Are We From Having a Satisfactory Theory of Clustering?

- model selection - how should a user pick an appropriate clustering tool for a given clustering problem, and how should the parameters of such an algorithmic tool be tuned?
- In contrast with other common computational tasks, in clustering, different algorithms often yield drastically different outcomes.

Figure of Merit

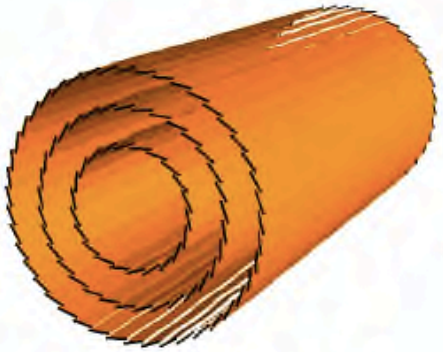
- A combination of resolution, fake rate, efficiency, …
  - → Tracking has been improved within a given a method (CKF+CTF) and within processing time constraints

- Not all tracks are equal. Not all features matter
  - → High dimensional cost function

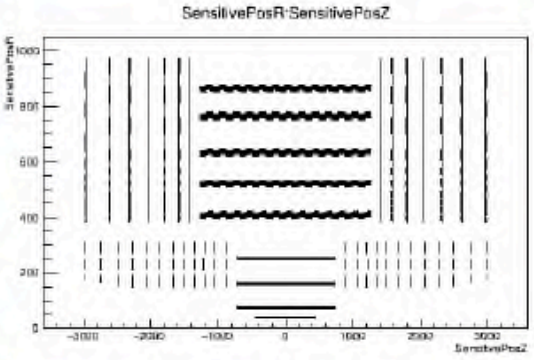
- No golden metric for "tracking" in a general purpose detector
  - → Things would be done differently, if the purpose was different

- ~~Remember the breaking point is computation requirement~~
  - → ~~Not something that folds in a cost function~~ …

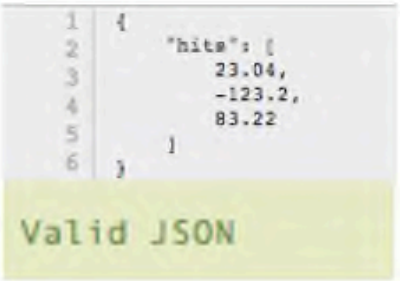
Summary Tracking ML challenge



detector geometry
planar barrel/EC type detector
pixel/strip system

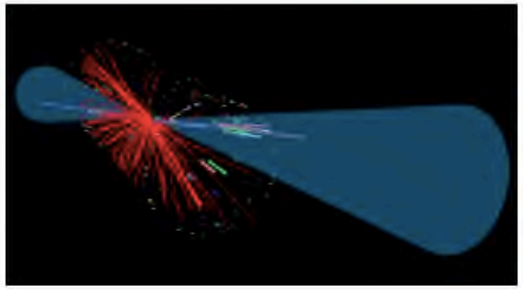


simulation
with the possibility to
simplify where possible



event data
easily readable,
platform independent

well defined goal
what is success
and how we measure it

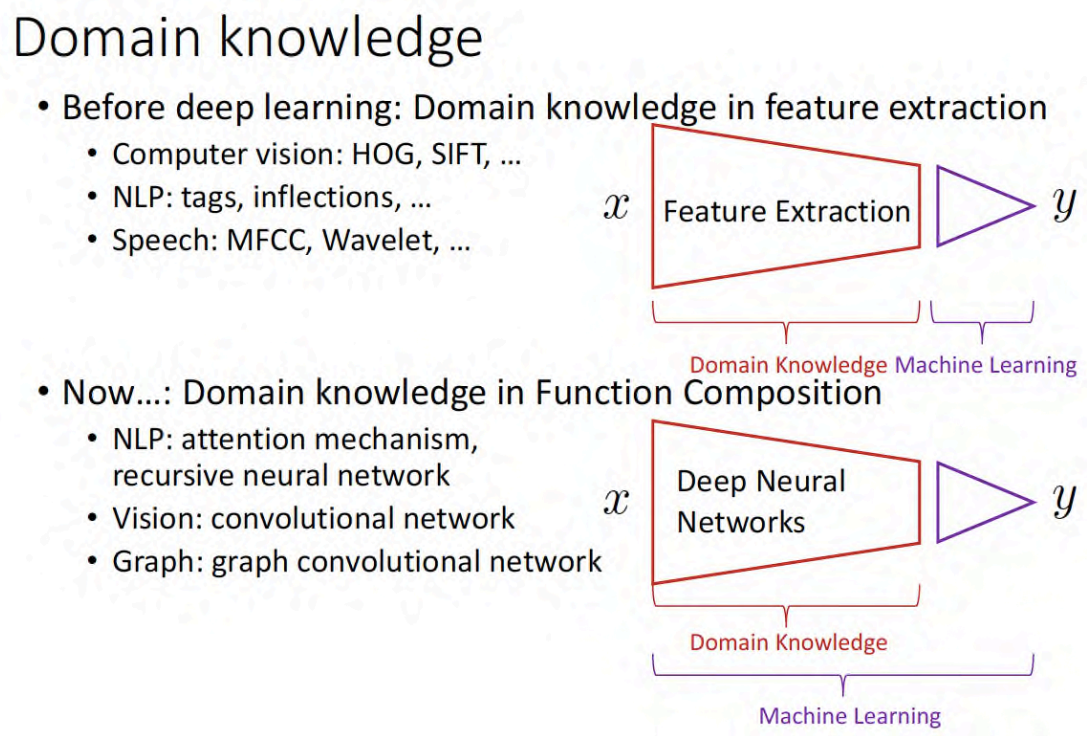


visualisation
of geometry,
hits & found tracks

different categories
for different
solutions

Integrating domain knowledge

- Kyunghyun Cho. Sequential and Recursive Learning



- Relation with physics-aware ML?

Trainable Decoding

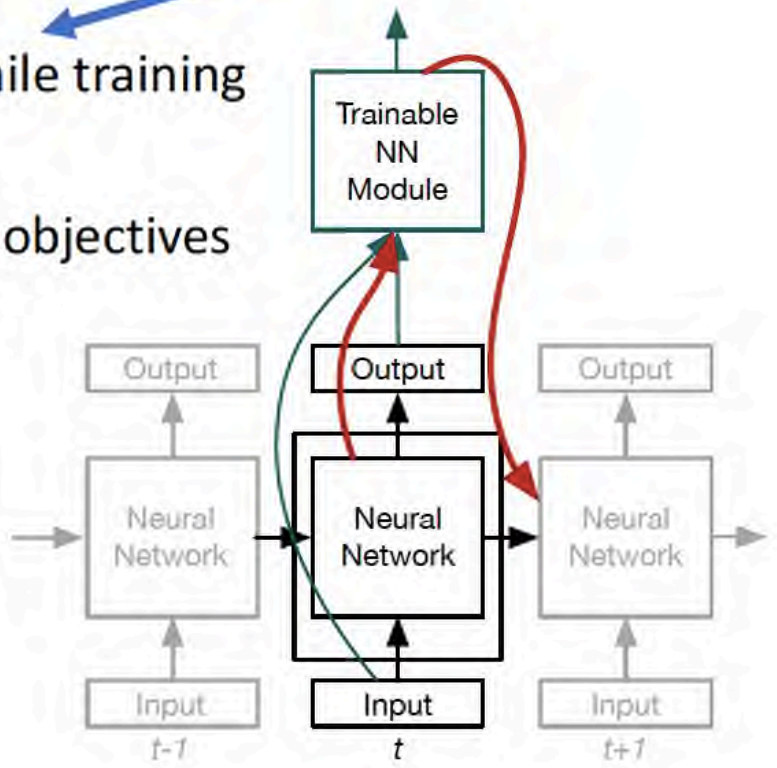
*As **Yossi Keshet** said earlier, we want to maximize task loss not log-likliehood*

Motivation

- Many decoding objectives unknown while training
- Lack of target training examples
- Arbitrary (non-differentiable) decoding objectives
- Sample-"**in**"efficiency of RL algorithms

Our Approach

- Train NMT with supervised learning
- Train a decoding module on top



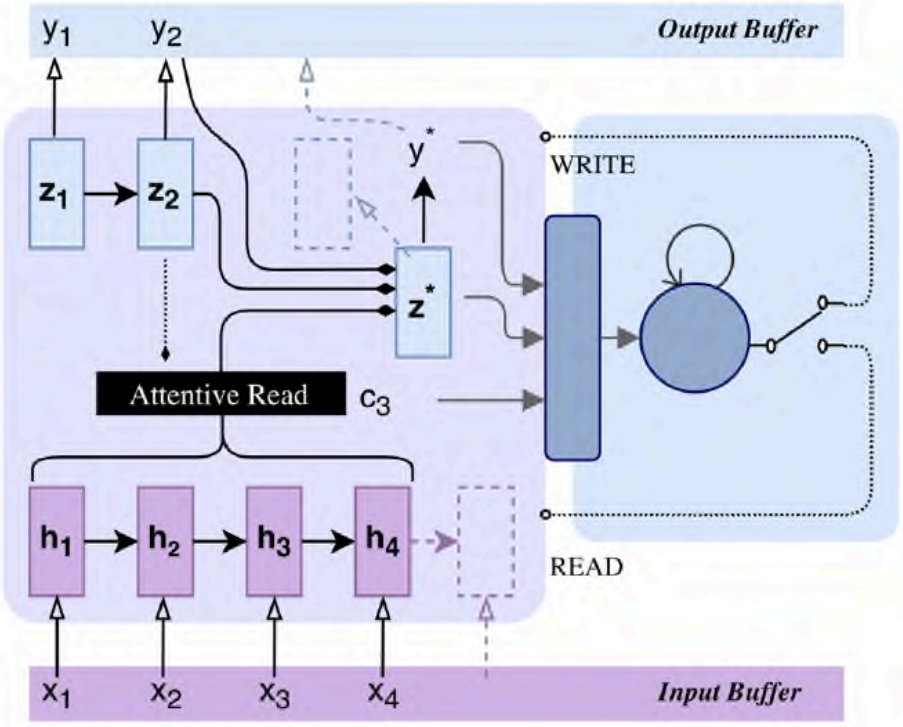
(1) Real-Time Translation

Decoding

1. Start with a pretrained NMT
2. A simultaenous decoder intercepts and interprets the incoming signal
3. The simultaneous decoder forces the pretrained model to either
   1. output a target symbol, or
   2. wait for a next source symbol

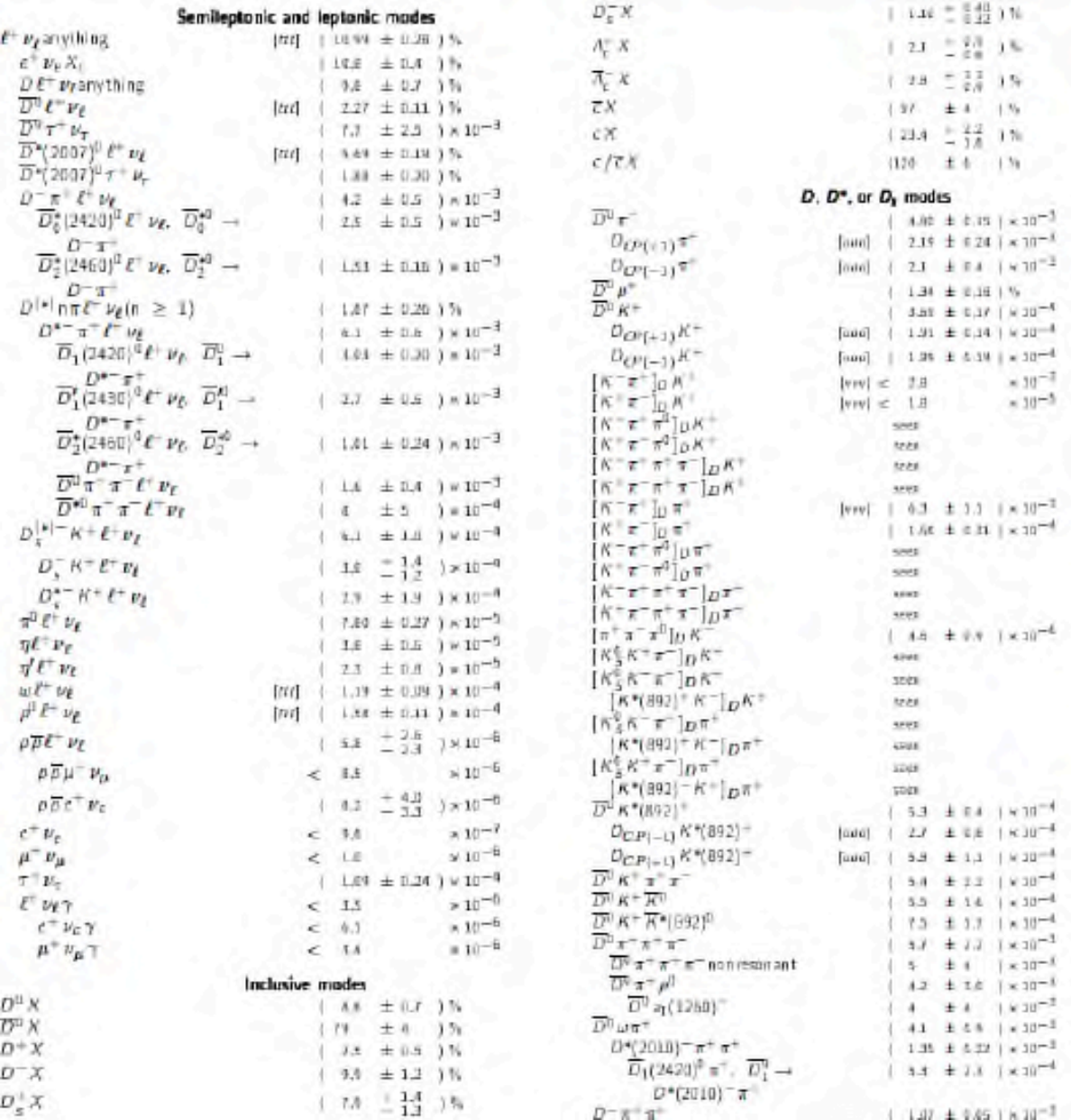
Learning

1. Trade-off between delay and quality
2. Stochastic policy gradient (REINFORCE)



(Gu, Neubig, Cho & Li, EACL 2017)

Physics-aware ML



Semileptonic and leptonic modes

$\ell^+ \nu_\ell$ anything [nz] (10.99 ± 0.28) %
$e^+ \nu_e X_c$ (10.2 ± 0.4) %
$D \ell^+ \nu_\ell$ anything (9.8 ± 0.7) %
$\overline{D}^0 \ell^+ \nu_\ell$ [nz] (2.27 ± 0.11) %
$\overline{D}^0 \tau^+ \nu_\tau$ (7.7 ± 2.5) × 10^{-3}
$\overline{D}^*(2007)^0 \ell^+ \nu_\ell$ [nz] (5.69 ± 0.19) %
$\overline{D}^*(2007)^0 \tau^+ \nu_\tau$ (1.88 ± 0.20) %
$D^- \pi^+ \ell^+ \nu_\ell$ (4.2 ± 0.5) × 10^{-3}
$\overline{D}_0^*(2420)^0 \ell^+ \nu_\ell$, $\overline{D}_0^{*0} \to D^- \pi^+$ (2.5 ± 0.5) × 10^{-3}
$\overline{D}_2^*(2460)^0 \ell^+ \nu_\ell$, $\overline{D}_2^{*0} \to D^- \pi^+$ (1.51 ± 0.16) × 10^{-3}
$D^{(*)} n\pi \ell^- \nu_\ell (n \geq 1)$ (1.87 ± 0.20) %
$D^{*-} \pi^+ \ell^+ \nu_\ell$ (6.1 ± 0.6) × 10^{-3}
$\overline{D}_1(2420)^0 \ell^+ \nu_\ell$, $\overline{D}_1^0 \to D^{*-} \pi^+$ (4.04 ± 0.30) × 10^{-3}
$\overline{D}_1'(2430)^0 \ell^+ \nu_\ell$, $\overline{D}_1^{\prime 0} \to D^{*-} \pi^+$ (2.7 ± 0.5) × 10^{-3}
$\overline{D}_2^*(2460)^0 \ell^- \nu_\ell$, $\overline{D}_2^{*0} \to D^{*-} \pi^+$ (1.81 ± 0.24) × 10^{-3}
$\overline{D}^0 \pi^- \pi^+ \ell^+ \nu_\ell$ (1.6 ± 0.4) × 10^{-3}
$\overline{D}^{*0} \pi^- \pi^+ \ell^+ \nu_\ell$ (4 ± 5) × 10^{-4}
$D_s^{(*)-} K^+ \ell^+ \nu_\ell$ (6.1 ± 1.0) × 10^{-3}
$D_s^- K^+ \ell^+ \nu_\ell$ (3.5 $^{+1.4}_{-1.2}$) × 10^{-3}
$D_s^{*-} K^+ \ell^+ \nu_\ell$ (2.9 ± 1.3) × 10^{-3}
$\pi^0 \ell^+ \nu_\ell$ (7.80 ± 0.27) × 10^{-5}
$\eta \ell^+ \nu_\ell$ (3.8 ± 0.5) × 10^{-5}
$\eta' \ell^+ \nu_\ell$ (2.3 ± 0.8) × 10^{-5}
$\omega \ell^+ \nu_\ell$ [nz] (1.19 ± 0.09) × 10^{-4}
$\mu^+ \nu_\mu$ [nz] (3.88 ± 0.11) × 10^{-4}
$p \overline{p} \ell^+ \nu_\ell$ (5.8 $^{+2.5}_{-2.3}$) × 10^{-6}
$p \overline{p} \mu^+ \nu_\mu$ < 8.5 × 10^{-6}
$p \overline{p} e^+ \nu_e$ (8.2 $^{+4.0}_{-3.3}$) × 10^{-6}
$e^+ \nu_e$ < 9.8 × 10^{-7}
$\mu^+ \nu_\mu$ < 1.0 × 10^{-6}
$\tau^+ \nu_\tau$ (1.09 ± 0.24) × 10^{-3}
$\ell^+ \nu_\ell \gamma$ < 3.5 × 10^{-6}
$e^+ \nu_e \gamma$ < 6.1 × 10^{-6}
$\mu^+ \nu_\mu \gamma$ < 5.4 × 10^{-6}

Inclusive modes

$D^0 X$ (8.8 ± 0.7) %
$\overline{D}^0 X$ (79 ± 6) %
$D^+ X$ (3.5 ± 0.5) %
$D^- X$ (9.8 ± 1.2) %
$D_s^+ X$ (7.8 $^{+1.4}_{-1.3}$) %
$D_s^- X$ (1.46 $^{+0.40}_{-0.32}$) %
$\Lambda_c^+ X$ (2.1 $^{+0.9}_{-0.6}$) %
$\overline{\Lambda}_c^- X$ (2.8 $^{+2.2}_{-0.9}$) %
$\overline{c} X$ (97 ± 4) %
$c X$ (23.4 $^{+2.2}_{-1.8}$) %
$c/\overline{c} X$ (120 ± 5) %

D, D^*, or D_s modes

$\overline{D}^0 \pi^-$ (4.82 ± 0.15) × 10^{-3}
$D_{CP(+1)} \pi^-$ [uuu] (2.15 ± 0.24) × 10^{-3}
$D_{CP(-1)} \pi^-$ [uuu] (2.1 ± 0.4) × 10^{-3}
$\overline{D}^0 \rho^-$ (1.34 ± 0.16) %
$\overline{D}^0 K^-$ (3.83 ± 0.17) × 10^{-4}
$D_{CP(+1)} K^-$ [uuu] (1.91 ± 0.14) × 10^{-4}
$D_{CP(-1)} K^-$ [uuu] (1.96 ± 0.19) × 10^{-4}
$[K^- \pi^+]_D K^-$ [vvv] < 2.8 × 10^{-7}
$[K^+ \pi^-]_D K^-$ [vvv] < 1.8 × 10^{-5}
$[K^- \pi^+ \pi^0]_D K^-$ seen
$[K^+ \pi^- \pi^0]_D K^-$ seen
$[K^- \pi^+ \pi^+ \pi^-]_D K^-$ seen
$[K^+ \pi^- \pi^+ \pi^-]_D K^-$ seen
$[K^- \pi^+]_D \pi^-$ [vvv] 0.3 ± 1.1) × 10^{-7}
$[K^+ \pi^-]_D \pi^-$ (1.46 ± 0.31) × 10^{-4}
$[K^- \pi^+ \pi^0]_D \pi^-$ seen
$[K^+ \pi^- \pi^0]_D \pi^-$ seen
$[K^- \pi^+ \pi^+ \pi^-]_D \pi^-$ seen
$[K^+ \pi^- \pi^+ \pi^-]_D \pi^-$ seen
$[\pi^+ \pi^- \pi^0]_D K^-$ (4.6 ± 0.9) × 10^{-4}
$[K^0_S K^+ \pi^-]_D K^-$ seen
$[K^0_S K^- \pi^+]_D K^-$ seen
$[K^*(892)^+ K^-]_D K^+$ seen
$[K^0_S K^- \pi^+]_D \pi^+$ seen
$[K^*(892)^+ K^-]_D \pi^+$ seen
$[K^0_S K^+ \pi^-]_D \pi^-$ seen
$[K^*(892)^- K^+]_D \pi^+$ seen
$\overline{D}^0 K^*(892)^+$ (5.3 ± 0.4) × 10^{-4}
$D_{CP(-1)} K^*(892)^+$ [uuu] (2.7 ± 0.6) × 10^{-4}
$D_{CP(+1)} K^*(892)^-$ [uuu] (5.8 ± 1.1) × 10^{-4}
$\overline{D}^0 K^+ \pi^-$ (5.8 ± 2.2) × 10^{-4}
$\overline{D}^0 K^+ \overline{K}^0$ (5.5 ± 1.6) × 10^{-4}
$\overline{D}^0 K^+ \overline{K}^*(892)^0$ (7.5 ± 3.7) × 10^{-4}
$\overline{D}^0 \pi^+ \pi^+ \pi^-$ (5.7 ± 2.2) × 10^{-3}
$\overline{D}^0 \pi^+ \pi^+ \pi^-$ nonresonant (5 ± 4) × 10^{-3}
$\overline{D}^0 \pi^+ \rho^0$ (4.2 ± 1.6) × 10^{-3}
$\overline{D}^0 a_1(1260)^+$ (4 ± 4) × 10^{-2}
$\overline{D}^0 \omega \pi^+$ (4.1 ± 0.5) × 10^{-3}
$D^*(2010)^- \pi^+ \pi^+$ (1.35 ± 0.22) × 10^{-3}
$\overline{D}_1(2420)^0 \pi^+$, $\overline{D}_1^0 \to D^*(2010)^- \pi^+$ (5.3 ± 2.3) × 10^{-4}
$D^- K^+ \pi^+$ (1.07 ± 0.09) × 10^{-3}

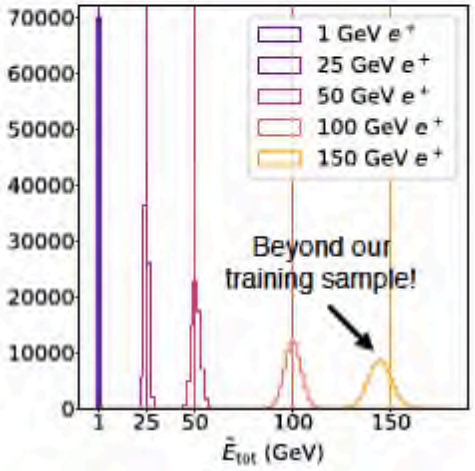
- We have a model
- We know the exclusive decay modes to high fidelity
- This list goes on for ~10 pages
- Can we use this model in a smart way?

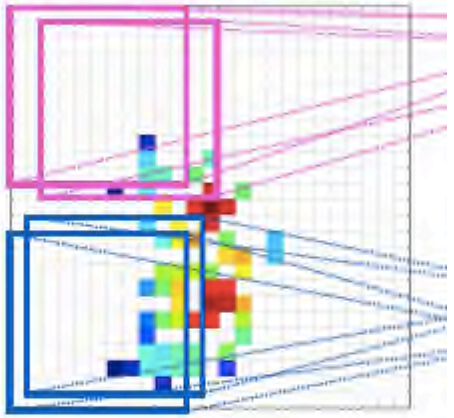
Interpretability

. Understanding what algorithm learns
– This is a different question not relevant for our motivation
• Ok to replace current black box with new black box
– We can calibrate with data

Conclusions 39

(Jet) image-based NN classification, regression, and generation are powerful tools for fully exploiting the physics program at the LHC





The key to robustness is to study what is being learned; this may even help us to learn something new!

Danilo Rezende,
Approximate Inference and Generative Models

From Marumi Kado

- A very efficient way to communicate with the TH/MC community has been to keep a wishlist from the HEP/EXP community.
- Wishlist: the experiments could provide a specific formulation of the question, a clear definition of the figure(s) of merit and realistic datasets to further solve the issues?

- The essentials: game changers for the physics
  - Trigger: classification and clustering (Nice e.g. from LHCb).
  - Tracking: clustering (reconstruction time in very dense pile-up challenge)
  - Simulation: Generative Networks (simulation time of the ATLAS calorimeter showers).

Wishlist

- Features improvements (very interesting but not essential).
  - ME generation: Integration (for higher order MC programs).
  - Calorimeter clustering: clustering.
  - Jet clustering and identification: clustering and classification.
  - Pile-up mitigation: classification.
  - Electron and photon identification: clustering and classification.
  - B and C tagging: classification and clustering (could substantial improvement be gained from DNNs starting from low level variables?).
  - Analysis and systematics: classification and adversarial networks.
- Prospective and perhaps too ambitious: can data and ML teach us about physics?
  - GEANT tuning with data?
  - Finding the right PS parametrisation for the best split function?

Systematics

- I would do the following : - I would pick a physics measurement which is systematics dominated (or will be at Run 2, by end 2018) - part of the PhD would be to exercise different systematics mitigation technique (« traditional », tangent prop, pivot, parameterised learning...) , first with a simplified toy set up, then with the full analysis, aiming at a final Run 2 ATLAS publication using one of these thechnique

Projects of Interest (Rita Osadchy)

- Calorimeter classification using data from a stack of 3 detectors (including improvements of clustering).
- More principled (end-to-end) approach to b-tagging (using graph networks)
- Tracking challenge
  - Exploring methods from computer vision
  - Formulating the problem as a clustering problem where trajectories are clusters (manifolds)
- Domain adaptation
- Understanding/exploring deep networks robustness (in HEP)

ML applications for fast local reconstruction

Next generation calorimeters will be extremely granular and difficult to handle with rule-based algorithms, while Deep Learning might be more suitable (see experience in neutrinos). In order to transition to a full ML approach, one would need

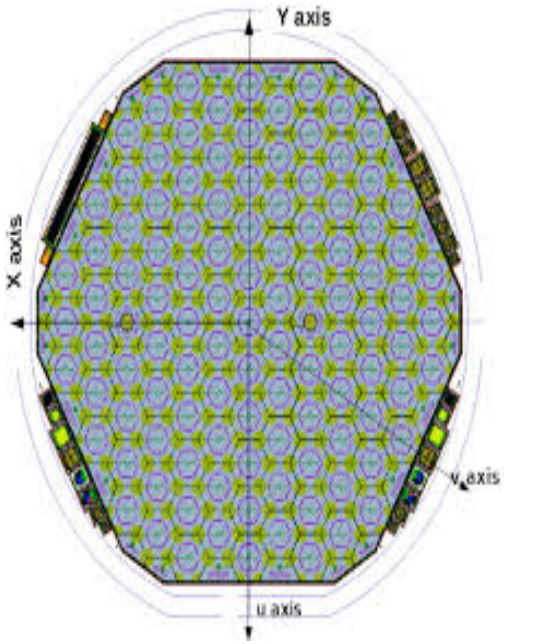
A clustering algorithm that makes particle candidates (see the discussion at the workshop on how to optimize similarity given a clustering algorithm)

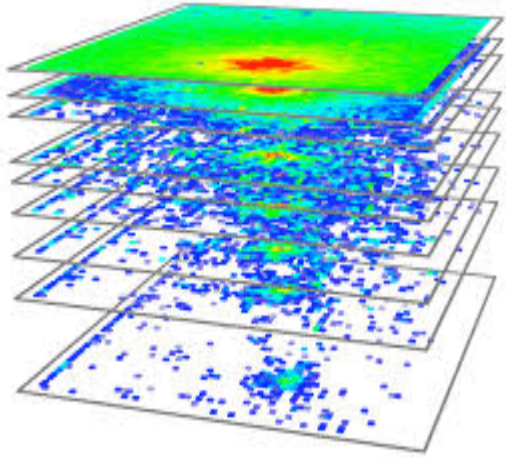
A set on classifier and regression algorithms that associate to each cluster an hypothesis (electron, pion, etc) and a measurement of energy and direction

A classifier that distinguish “intereting” particles from those originating from pileup collisions

An algorithm (trained in adversarial) that would morph the simulated events into real data (e.g. those obtained in testbeams)

The CMS HGCAL for High-Luminosity LHC could be used as a benchmark for this project

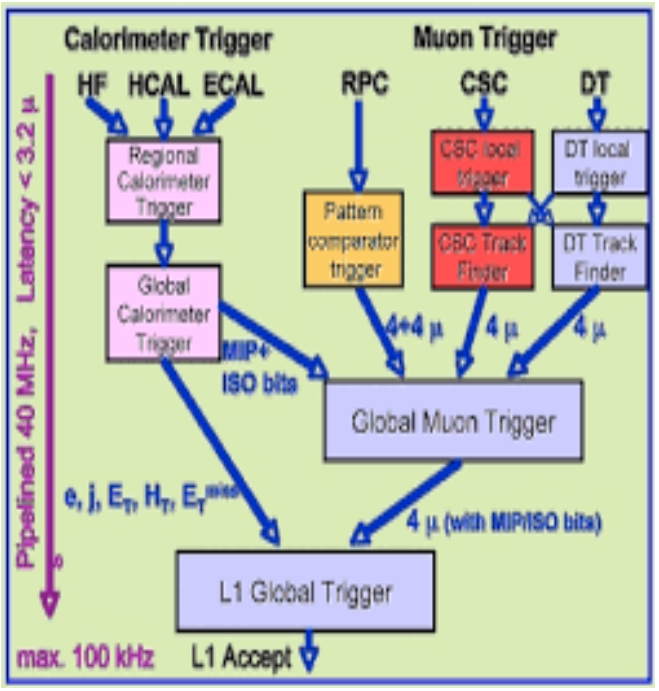




Fast DL Inference for trigger applications

The project consists in developing tools to optimize the inference of given NN models on parallel architectures (e.g., FPGAs and GPUs), targeting a latency of O(10 ms), typical of the L1 hardware-based layer of a trigger system for the LHC experiments. Connecting to the “ML applications for fast local reconstruction” project, this project would allow to deploy the models developed in that context in the early stages of the data acquisition system. The main task will be in finding the optimal solution between the model complexity, the inference time, and the hardware platform used at this purpose.

The project could focus on specific local-reconstruction tasks (i.e., muons in the muon stations, clusters in the calorimeter), on the high-level steps of the reconstruction, which put together these information (so-called trigger primitives) into a coherent view of the event, or on the selection algorithms themselves (e.g., regression models to improve resolution)



Machine Learning Questions – One Physicist's Perspective

- Non-flat Geometries in Neural Networks
  - Nati Srebro (TTIC) made a compelling case for alternative distance measures – perhaps a problem has been the implicit assumption of a flat space in which a Euclidean distance is sufficient. What if one considered curved or highly warped hyperspaces where non-Euclidean methods are more appropriate?
  - It's possible that where we have seen statistical mechanics/ information theory as useful for understanding learning, a kind of "General Relativity" - movement in non-flat space – would be equally useful for machine learning.
- "Path Integral Formulation"
  - As was noted by K. Cranmer and others, Srebro's distance measure resembles the Feynman Path Integral Formulation – is there more to be learned there?