粒子物理中的机器学习

梅华林

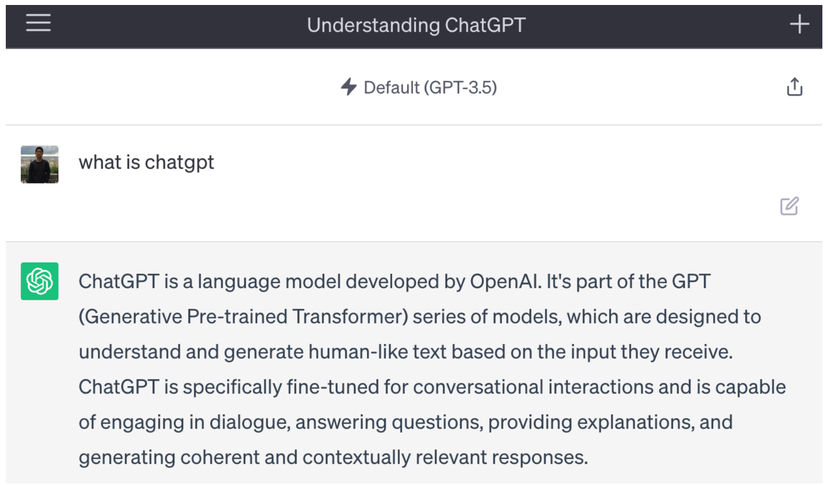
上海交通大学李政道研究所

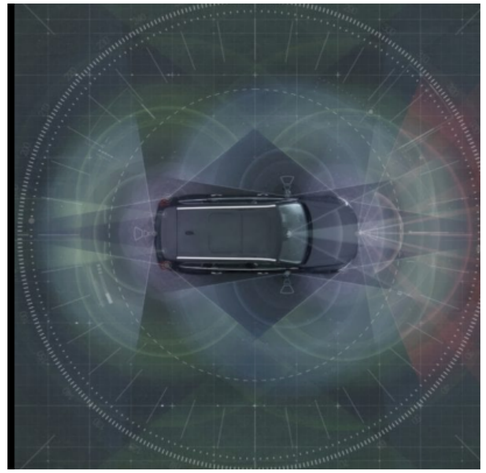
Disclaimer

- Not to teach what is machine learning (ML)
- Try to show examples of ML used in experimental particle physics
- Focus on collider physics
- More material from CMS experiment in this talk

什么是机器学习

- 通常是使用大量的数据来训练模型，这些模型一旦被训练，可以被应用到其他相似数据上做出预测或决策
- “监督学习”：从我们知道正确答案的数据中训练的算法
- “无监督学习”：不知道正确的答案，算法在数据中自行探索

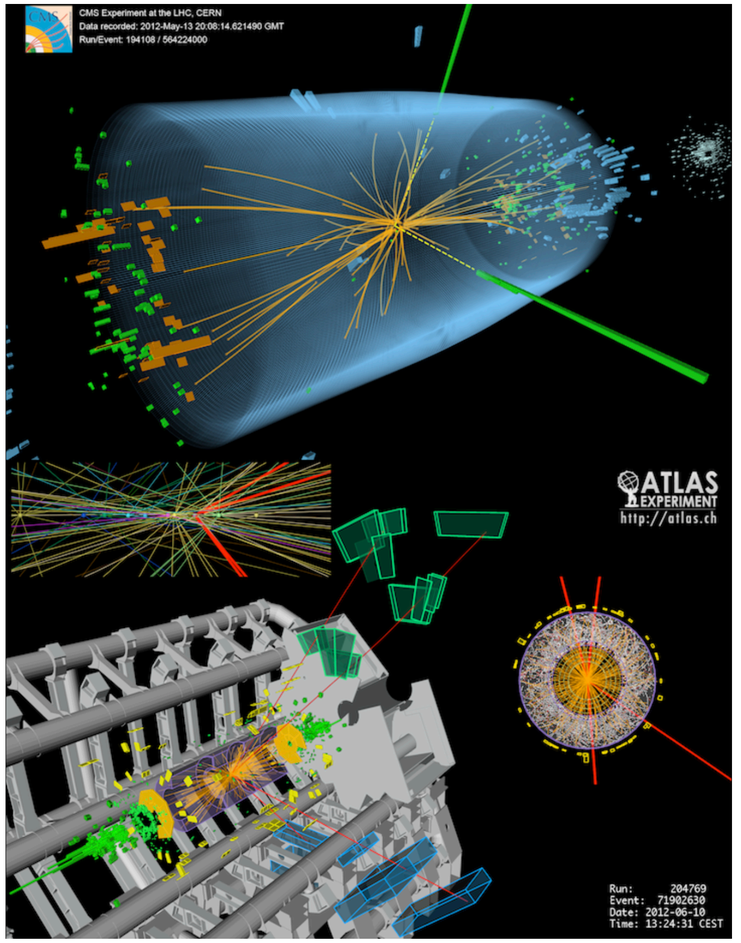


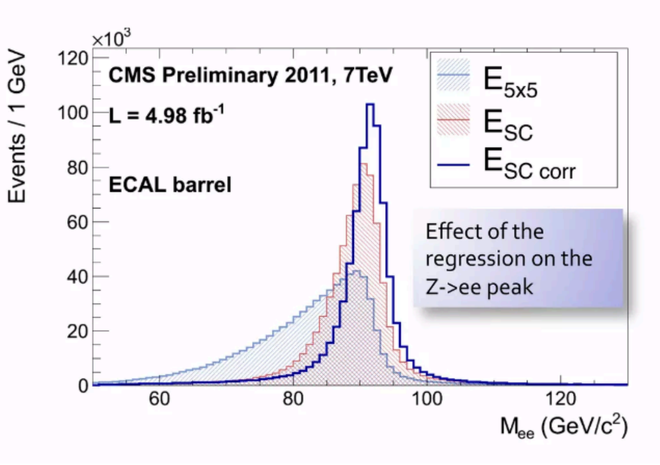


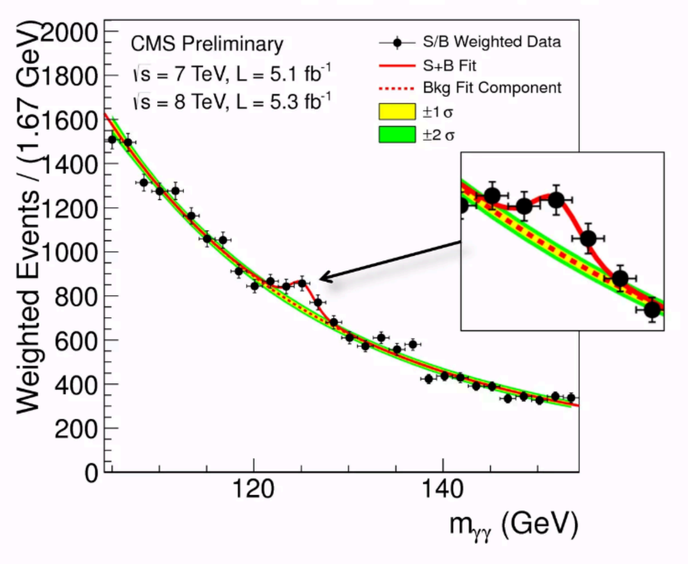
粒子物理中的机器学习

- 有着广泛运用
  - 提高物理量的测量精度
  - 区别不同的过程（信号 VS 本底）
  - 降低模拟成本
  - 重建复杂过程
  - 快速在线选择事例

。 。 。

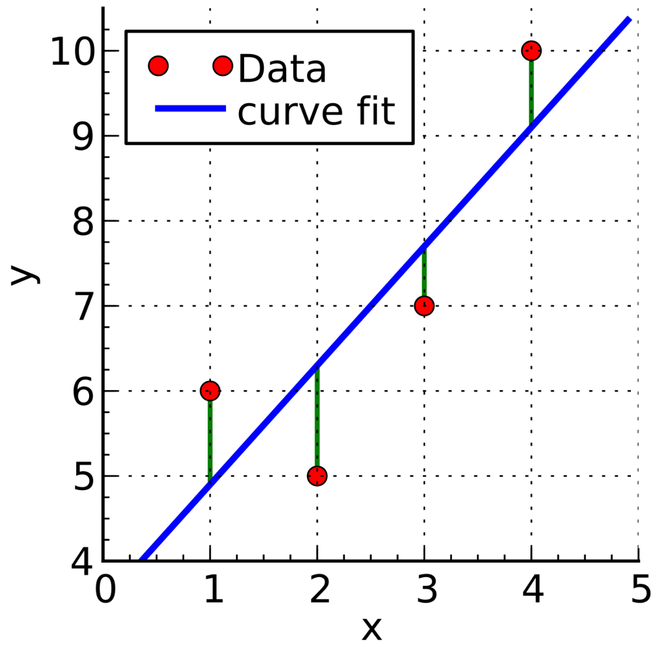






在2012年希格斯玻色子的发现中起到重要作用！

最小二乘回归



如何找到一个线性函数描述这四个红色的数据点？

最简单的“机器学习”

某次实验得到了四个数据点 (x, y)：$(1, 6)$、$(2, 5)$、$(3, 7)$、$(4, 10)$（右图红色的点）。我们希望找出一条和这四个点最匹配的直线 $y = \beta_2 x + \beta_1$，即找出在某种“最佳情况”下能够大致符合如下超定线性方程组的 β_1 和 β_2：

$$\begin{aligned}
\beta_1 + 1\beta_2 &= 6 \\
\beta_1 + 2\beta_2 &= 5 \\
\beta_1 + 3\beta_2 &= 7 \\
\beta_1 + 4\beta_2 &= 10
\end{aligned}$$

最小平方法采用的方法是尽量使得等号两边差的平方最小，也就是找出这个函数的最小值：

$$\begin{aligned}
S(\beta_1, \beta_2) = &[6 - (\beta_1 + 1\beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 \\
&+ [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2.
\end{aligned}$$

最小值可以通过对 $S(\beta_1, \beta_2)$ 分别求 β_1 和 β_2 的偏导数，然后使他们等于零得到。

$$\begin{aligned}
\frac{\partial S}{\partial \beta_1} &= 0 = 8\beta_1 + 20\beta_2 - 56 \\
\frac{\partial S}{\partial \beta_2} &= 0 = 20\beta_1 + 60\beta_2 - 154.
\end{aligned}$$

如此就得到了一个只有两个未知数的方程组，很容易就可以解出：

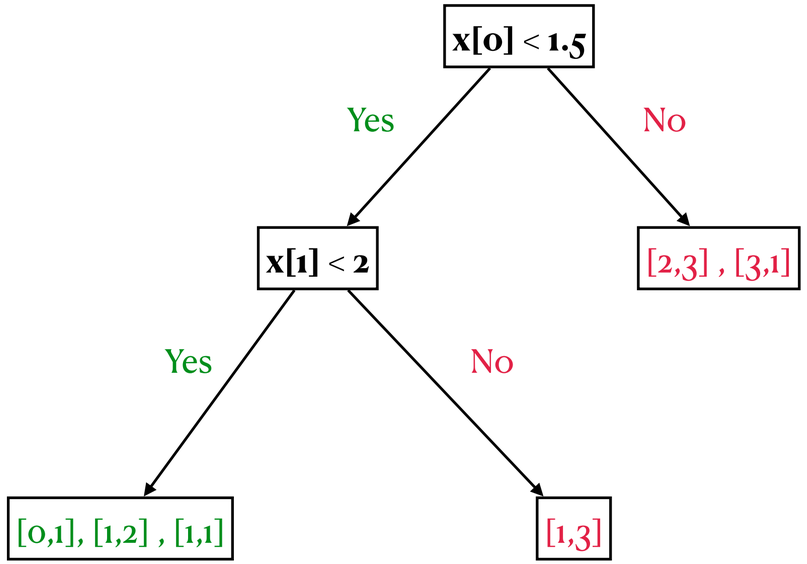
$$\begin{aligned}
\beta_1 &= 3.5 \\
\beta_2 &= 1.4
\end{aligned}$$

也就是说直线 $y = 3.5 + 1.4x$ 是最佳的。

From wikipedia

决策树

- 可用于解决分类和回归问题
- 基于输入数据的特征，通过一系列决策进行分类或预测
- 例子，假设有6个事例用于训练，每个事例有两个特征
  - 信号：[0, 1], [1, 2], [1, 1]
  - 本底：[1, 3], [2, 3], [3, 1]
  - 训练得到的决策树包含两次决策：x[0] < 1.5, x[1] < 2
  - 运用到新的事例 [0, 1.2], [2, 1] .... 可判断是信号还是本底
- 在粒子物理实验中，则根据需求选择合适的特征
  - 轻子数，缪子动量，量能器单元能量....



神经网络

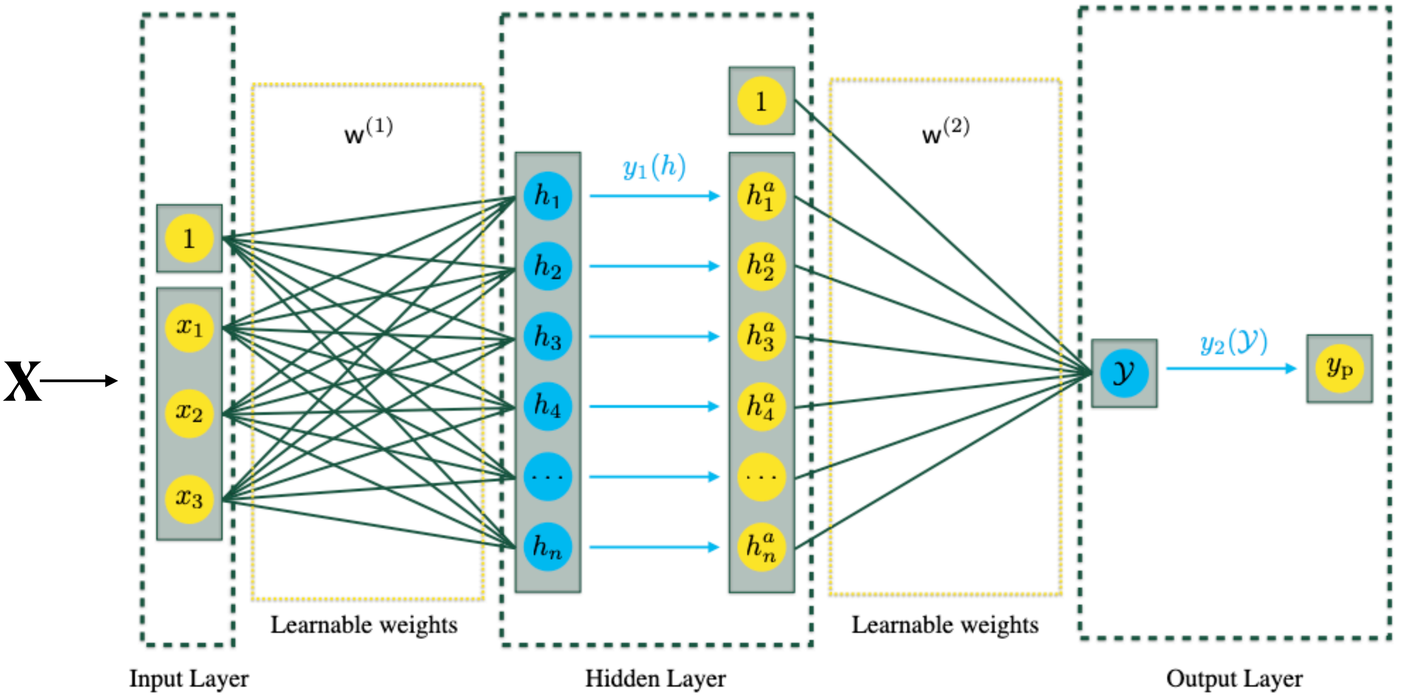
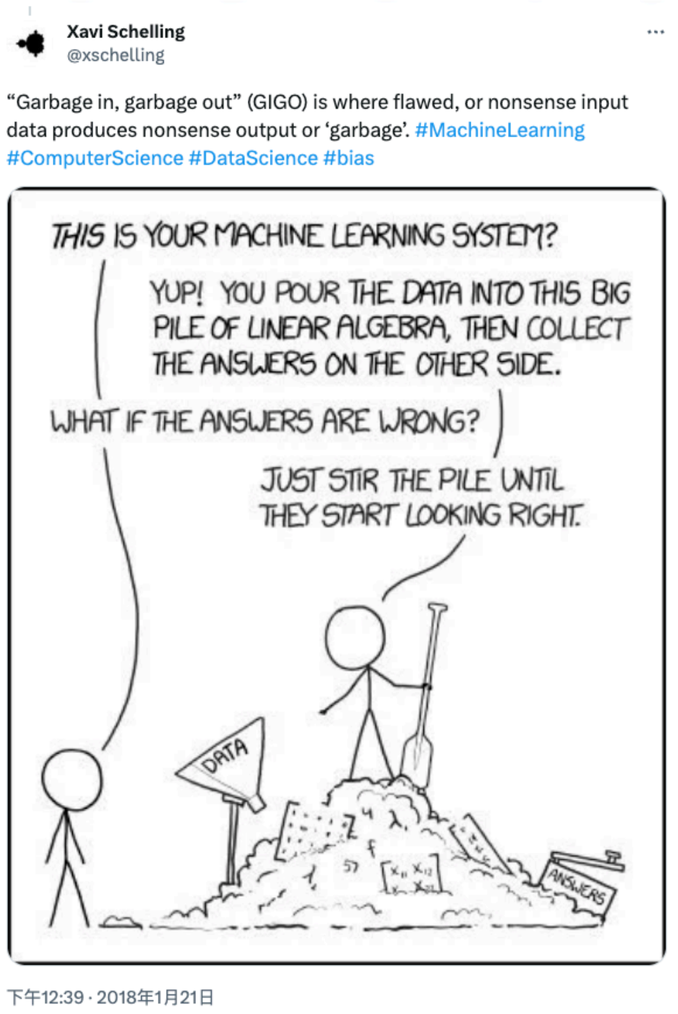


Diagram from Cohen, Freytsis, Ostdiek, 2017

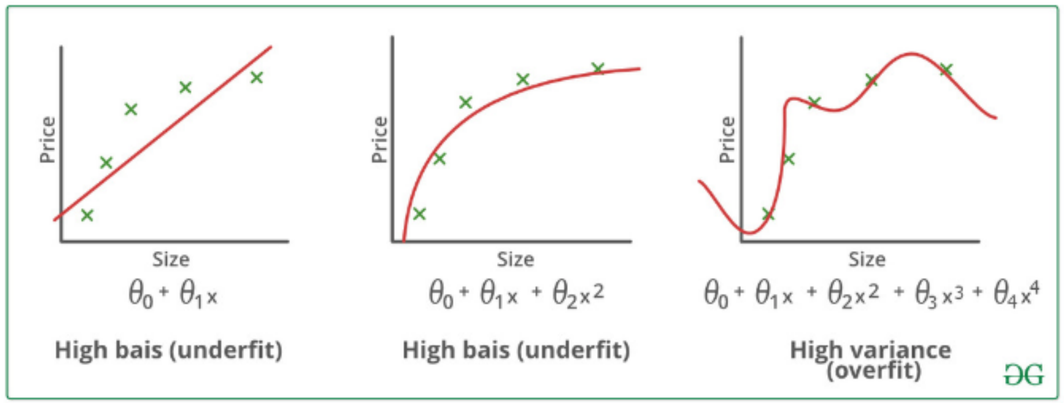
- 受到人类神经系统启发的计算模型
- $y_p = f_{model}(x_i, \boldsymbol{\theta})$
- 需要一系列的输入特征：x_i
- $\boldsymbol{\theta}$为模型的参数，需通过训练得到
- 隐含层，用于学习数据中的模式与特征，常见的例子：
  - $h1 = w10 + w11*x1 + w12*x2 + w13*x3$
  - $y(h) = 1/(1 + e^{-x})$ or $y(h) = \tanh(h)$
- 训练神经网络通常需要反向传播(链式法则)来更新模型中的参数

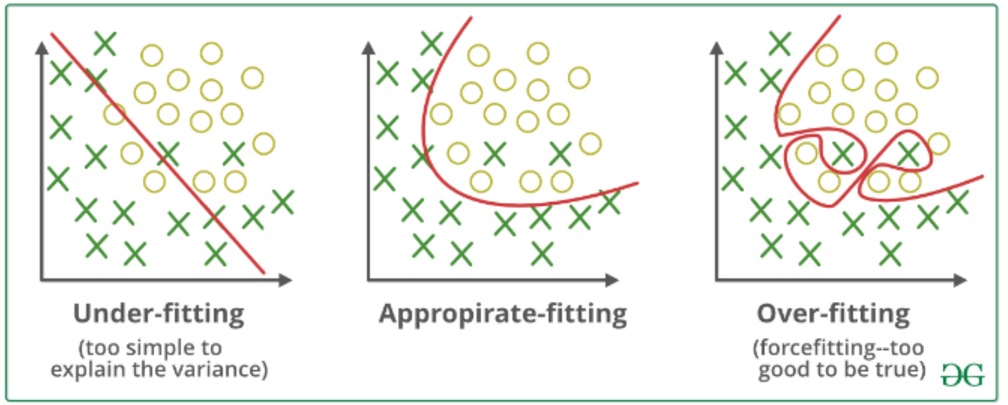
常见需要注意的问题（1）


Garbage in
Garbage OUT


Xavi Schelling
@xschelling
“Garbage in, garbage out” (GIGO) is where flawed, or nonsense input data produces nonsense output or ‘garbage’. #MachineLearning #ComputerScience #DataScience #bias
THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS
下午12:39 · 2018年1月21日

常见需要注意的问题（2）



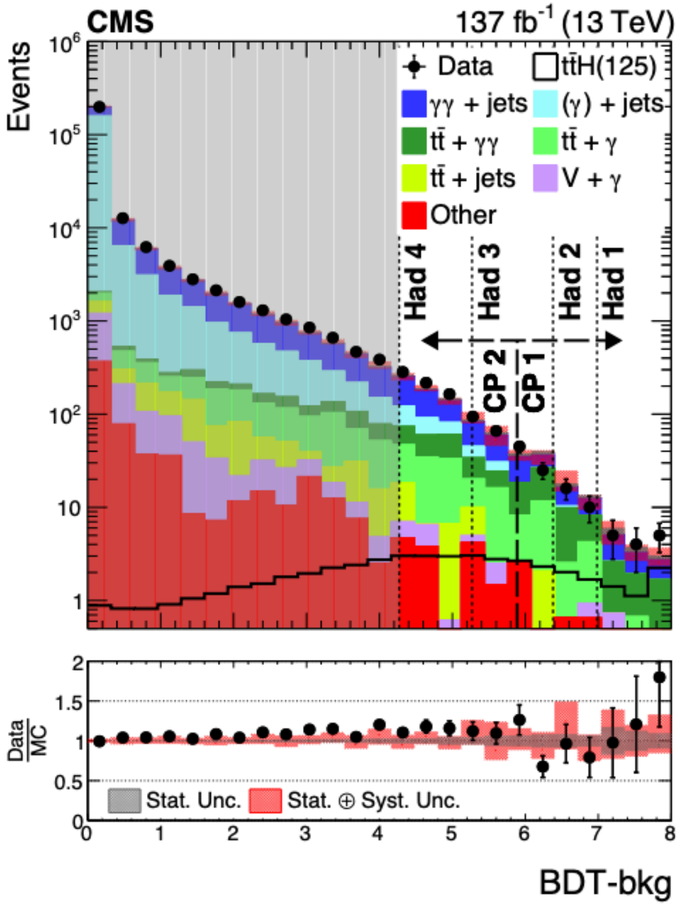


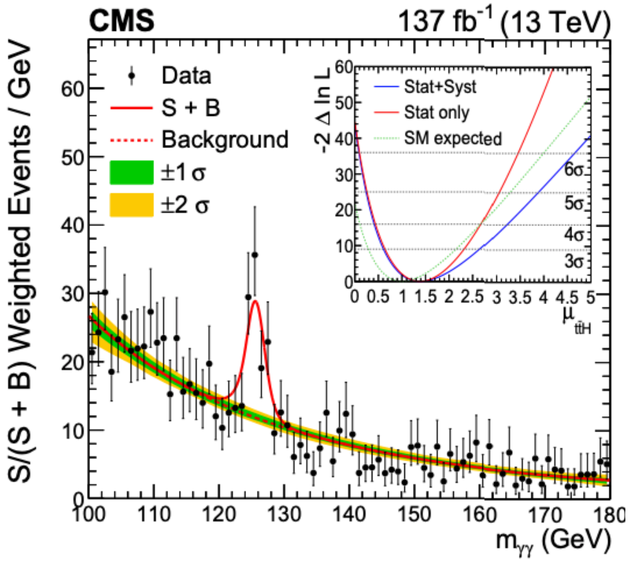
source

- 现有的机器学习模型通常会有很多参数，容易造成过度拟合
- 可适当优化模型，引入一定的约束 （正则化，or regularization）
- 常见的检查方法是把数据集分为几个部分
  - 训练集(training set)：用于通过优化目标函数获得参数
  - 验证集(validation set): 用于测试各种参数，确保我们在训练数据上没有过拟合
  - 测试集(test set): 查看我们最终选择的参数和超参数在我们没有拟合的数据上的表现如何

提升决策树的应用

- Boost decision tree (BDT)
- 多个比较弱的决策树结合成一个强学习器
- 常用于分析中的分类任务 (Signal vs bkg)
- CMS Run 2 ttH, H→γγ analysis
  - 输入：光子，强子喷注，轻子的动量，光子，b-jet 鉴别的质量等
  - 基于BDT的事例选择
    - S/B: O(1/100000) → O(1/1)

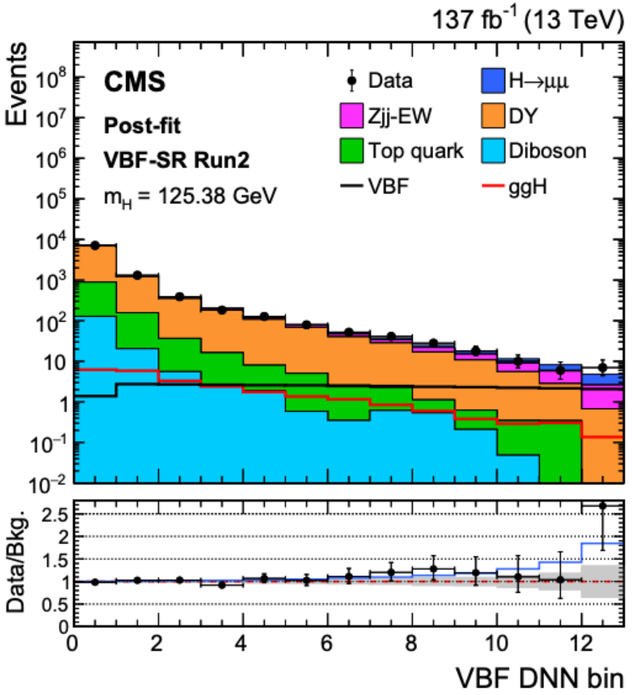


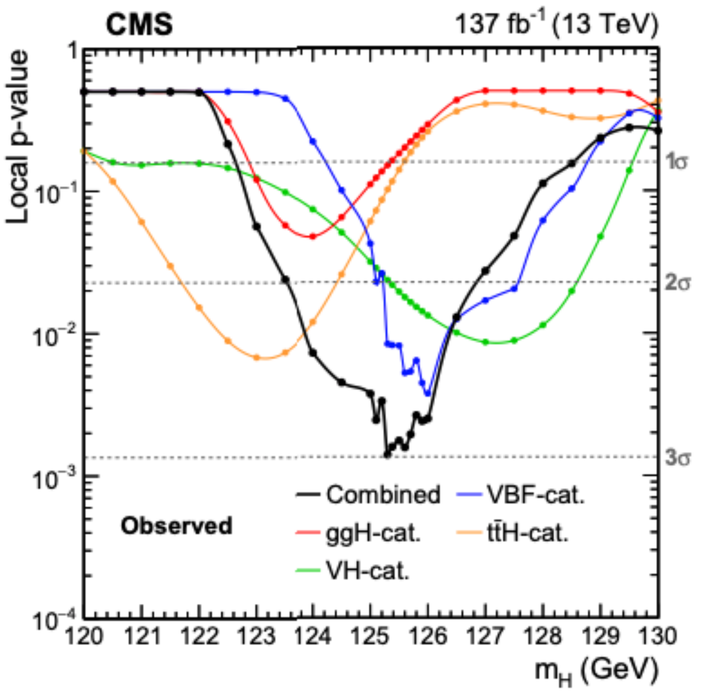


CMS Run 2 first ttH observation using a single Higgs decay channel (H→γγ), and test CP properties between the Higgs boson and fermion (top quark)
Phys. Rev. Lett. 125, 061801 (2020)

深度神经网络的应用

- 比起神经网络，深度神经网络(DNN)通常有更多的隐藏层，学习能力更强
- 比起BDT，DNN更适合利用更多的输入特征用来训练，更容易提炼出数据中隐藏的复杂特征
- CMS Run 2 H→uu search, VBF channel
  - 输入：muon, jet kinematics, angular variables of di-muon, di-jet system etc

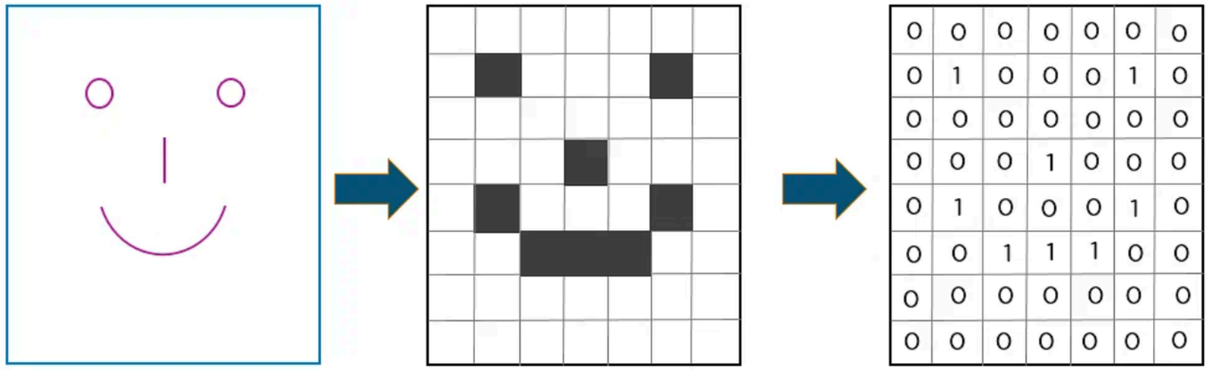


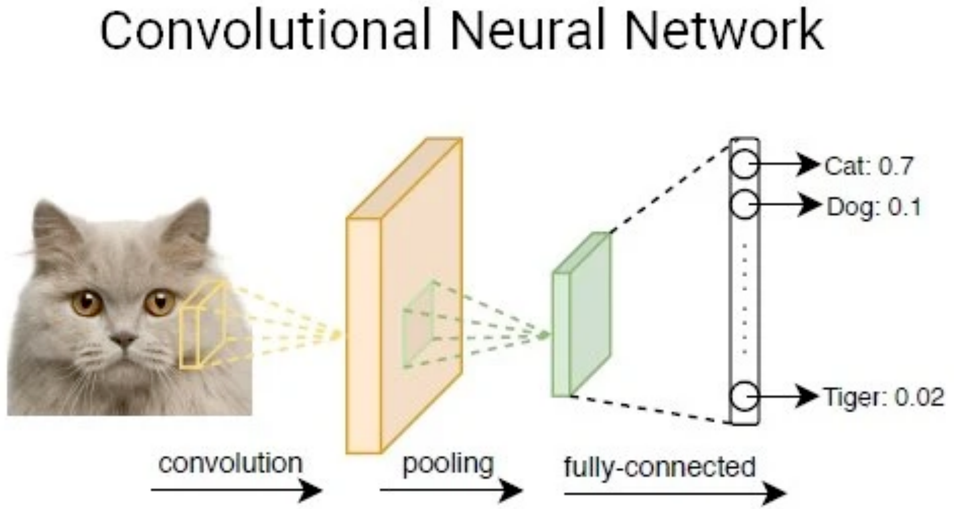


Evidence of Higgs boson couples to 2nd generation fermions, VBF channel dominates sensitivity
JHEP 01 (2021) 148

卷积神经网络

- 是一种常用于计算机视觉的深度学习神经网络架构
- 会用特殊的“卷积层”学习图像中区域性的特殊细节
- 在图像识别中有着非常成功的应用



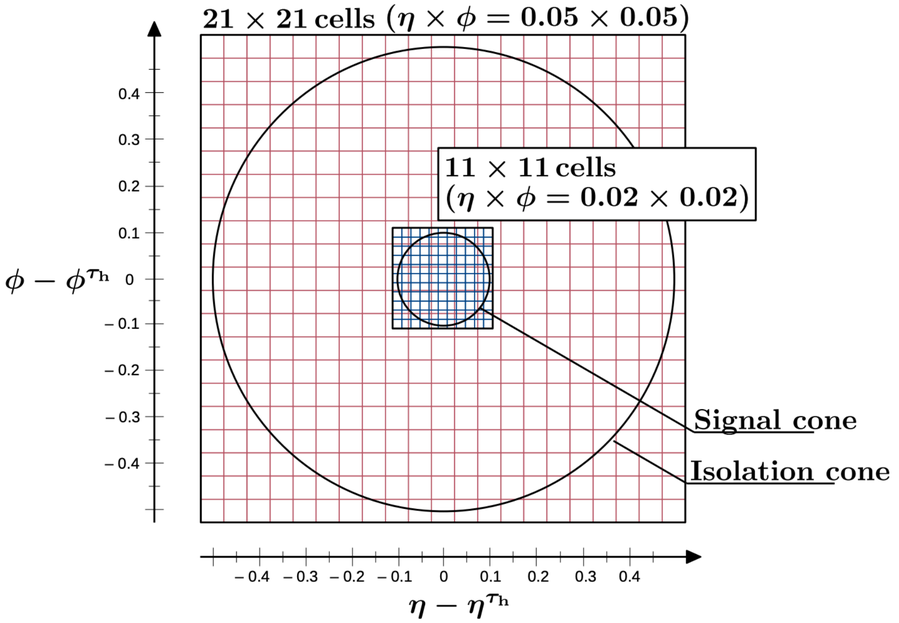


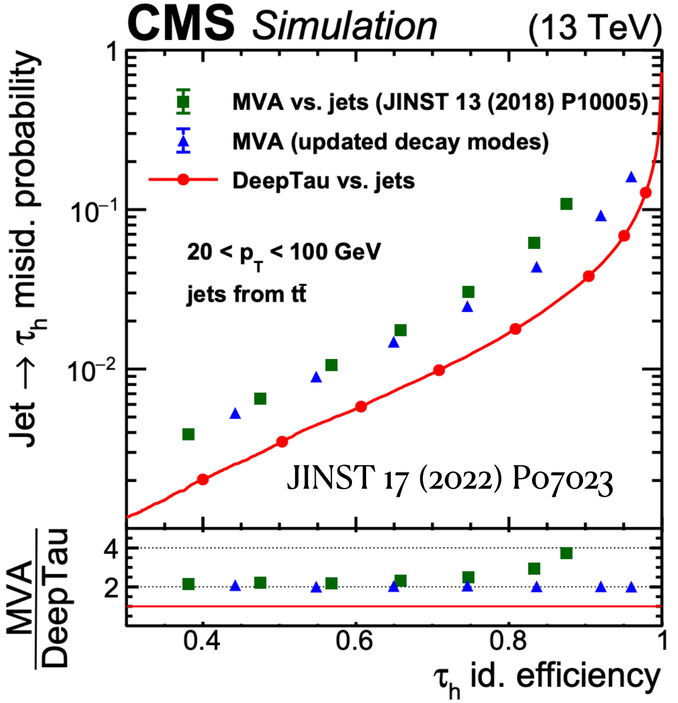
As you can see from the above diagram, only those values are lit that have a value of 1.

reference

卷积神经网络

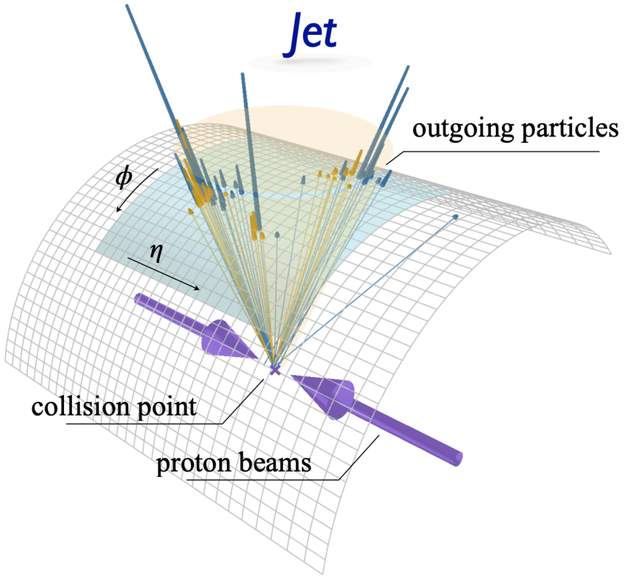
Decay mode	Resonance	$\mathcal{B}$ (%)
Leptonic decays		35.2
$\tau^- \to e^- \overline{\nu}_e \nu_\tau$		17.8
$\tau^- \to \mu^- \overline{\nu}_\mu \nu_\tau$		17.4
Hadronic decays		64.8
$\tau^- \to h^- \nu_\tau$		11.5
$\tau^- \to h^- \pi^0 \nu_\tau$	$\rho(770)$	25.9
$\tau^- \to h^- \pi^0 \pi^0 \nu_\tau$	$a_1(1260)$	9.5
$\tau^- \to h^- h^+ h^- \nu_\tau$	$a_1(1260)$	9.8
$\tau^- \to h^- h^+ h^- \pi^0 \nu_\tau$		4.8
Other		3.3

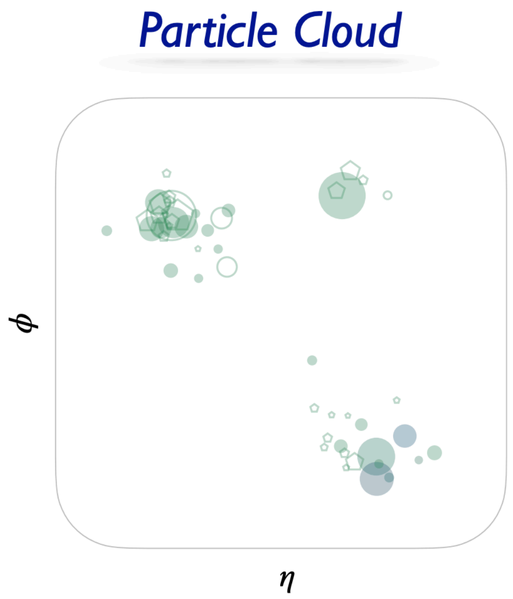




- 探测器本身就是一台大型照相机，CMS探测器每25ns就对发生的对撞照相一次
- 适合利用CNN来实现粒子鉴别，例如τ轻子，衰变末态多，在探测器中痕迹复杂
  - 对于其强子衰变道，满足~80%的效率同时只有~1%的误判率

图神经网络



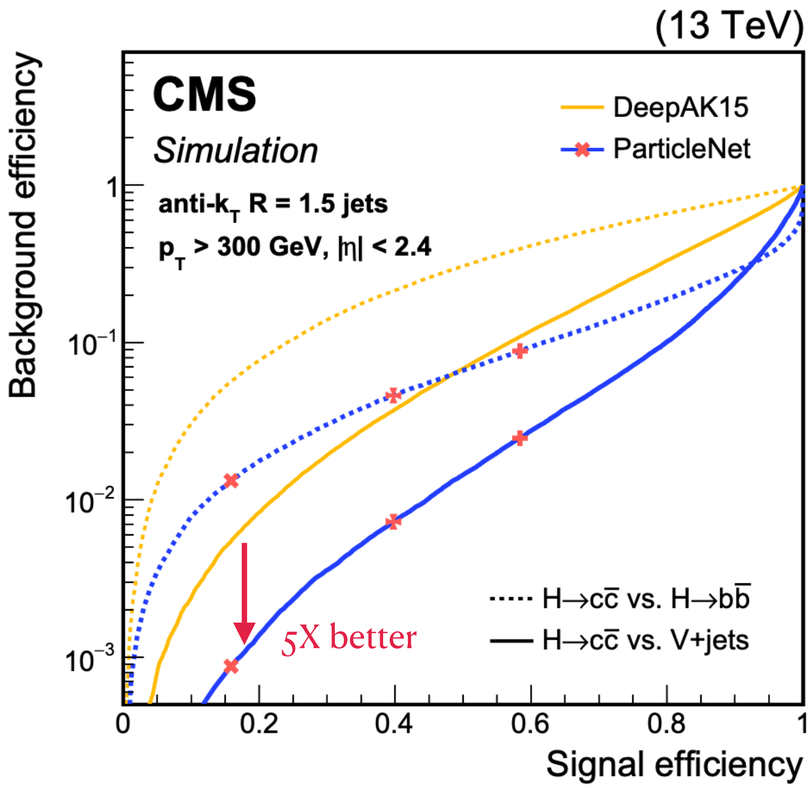


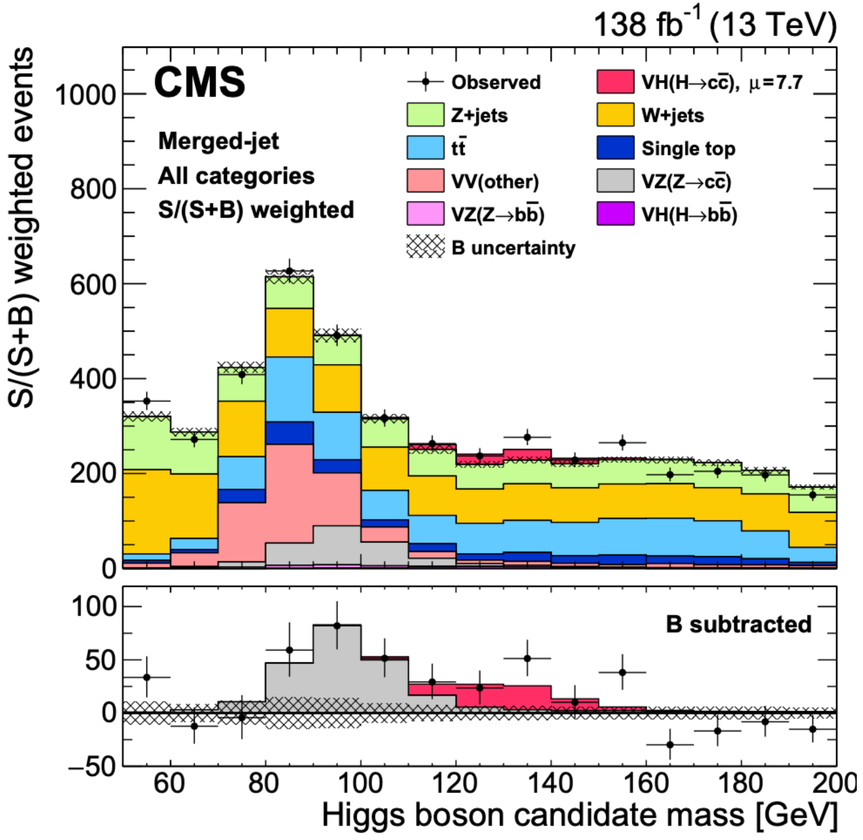
	Accuracy	AUC	$1/\varepsilon_b$ at $\varepsilon_s = 50\%$	$1/\varepsilon_b$ at $\varepsilon_s = 30\%$
ResNeXt-50	0.821	0.8960	30.9	80.8
P-CNN	0.818	0.8915	31.0	82.3
PFN	...	0.8911	30.8 ± 0.4	...
ParticleNet-Lite	0.826	0.8993	32.8	84.6
ParticleNet	0.828	0.9014	33.7	85.4
P-CNN (w/ PID)	0.827	0.9002	34.7	91.0
PFN-Ex (w/ PID)	...	0.9005	34.7 ± 0.4	...
ParticleNet-Lite (w/ PID)	0.835	0.9079	37.1	94.5
ParticleNet (w/ PID)	**0.840**	**0.9116**	$\mathbf{39.8 \pm 0.2}$	$\mathbf{98.6 \pm 1.3}$

Phys. Rev. D 101, 056019 (2020)

- 处理图数据的深度学习模型
- 图数据由节点（nodes）和连接节点的边（edges）构成
- 在节点及其相邻节点之间传递信息，从而捕获图数据中的结构和特征
- 强子喷注（jet）在探测器中的信号可表征为点云
- 用GNN来对jet进行分类在近几年得到非常成功的应用

图神经网络



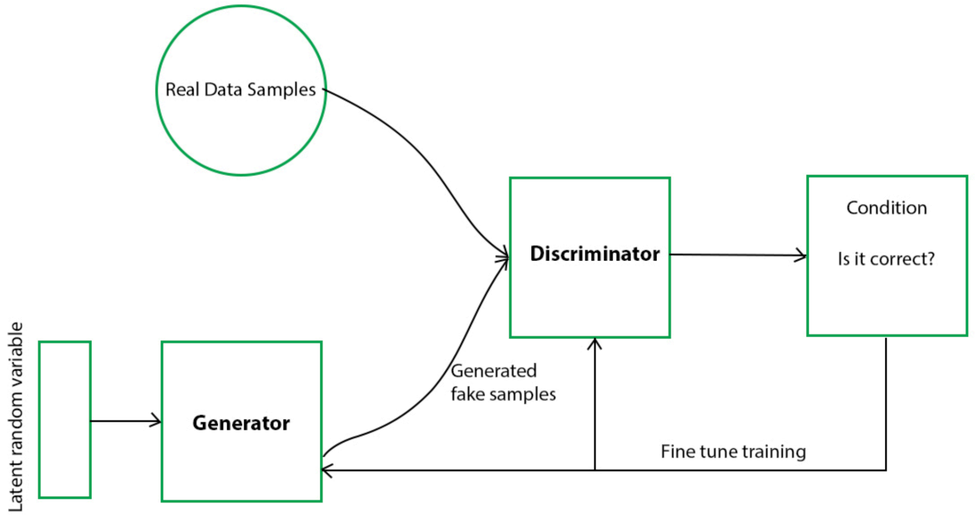


Best limit on XS(VH,H→cc) to-date, 14 X SM
Phys. Rev. Lett. 131 (2023) 061801

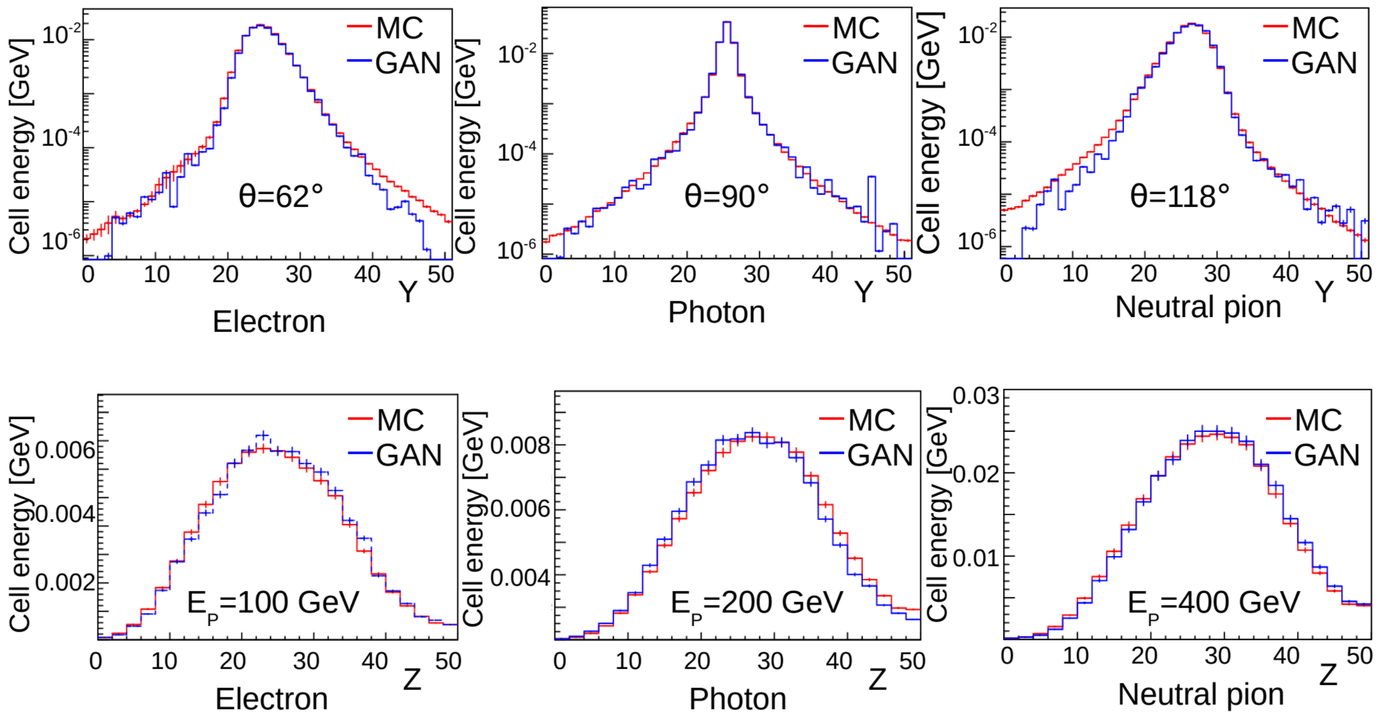
- CMS Run 2 H→cc search
- 研究希格斯玻色子和第二代夸克的耦合
- 之前被认为在HL-LHC也无法实现
  - charm quark induced jet重建难度极大，难以与本底区分
- 使用GNN大幅提高灵敏度，使得研究该耦合在未来成为可能

生成对抗网络

- Generative neural network (GAN)
- 由生成器（Generator）和判别器（Discriminator）两部分组成，它们通过对抗的方式相互训练，用于生成具有逼真度的新数据样本
  - 生成器和判别器都由神经网络组成



生成对抗网络

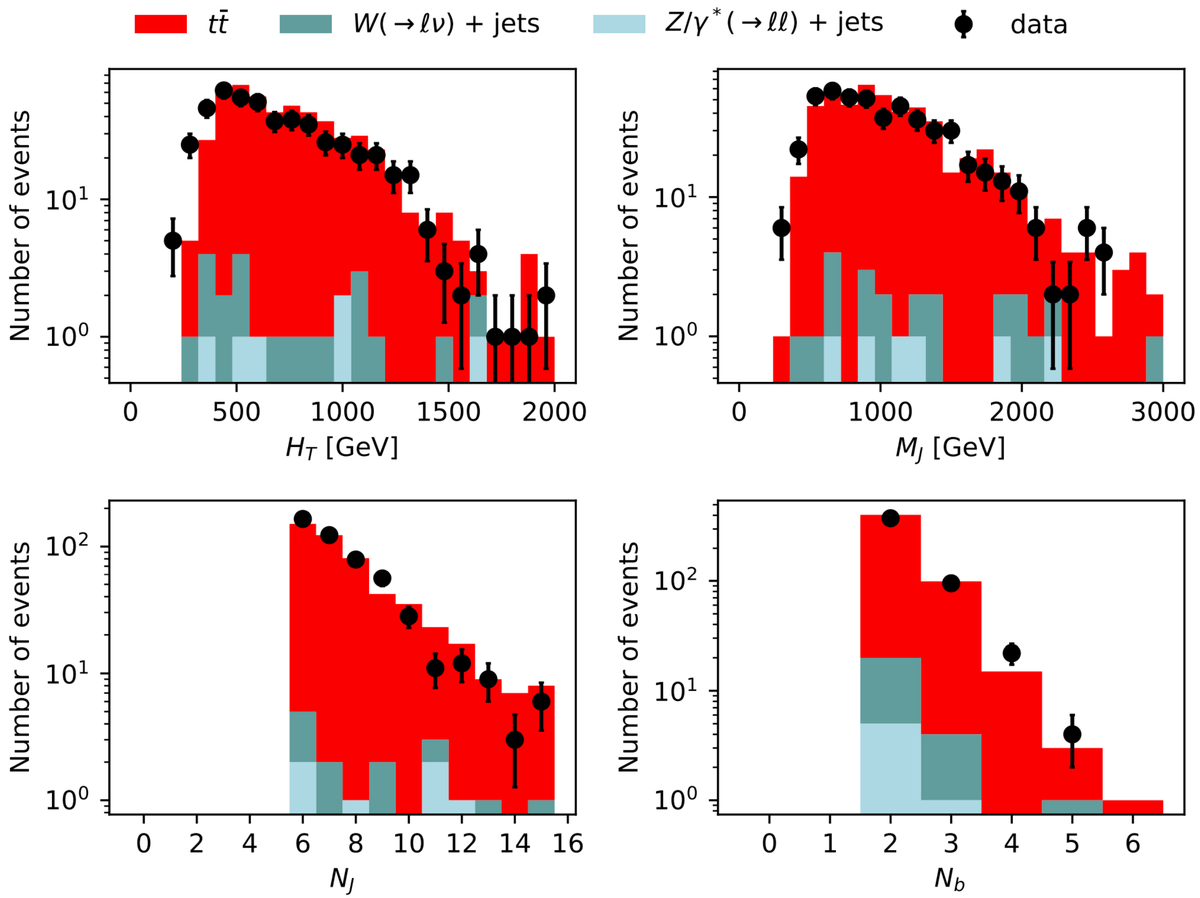


The showers generated by GAN present accuracy within 10% of Monte Carlo for a diverse range of physics features, with three orders of magnitude speedup

arXiv:2109.07388

- 模拟是粒子物理实验非常重要的一环
- 常常耗费巨大的计算资源 (>50%)
- 算力的需求也会随着对撞机亮度提升和探测器颗粒度增加而进一步扩大
- 探索用GAN进行快速模拟在近几年得到快速发展

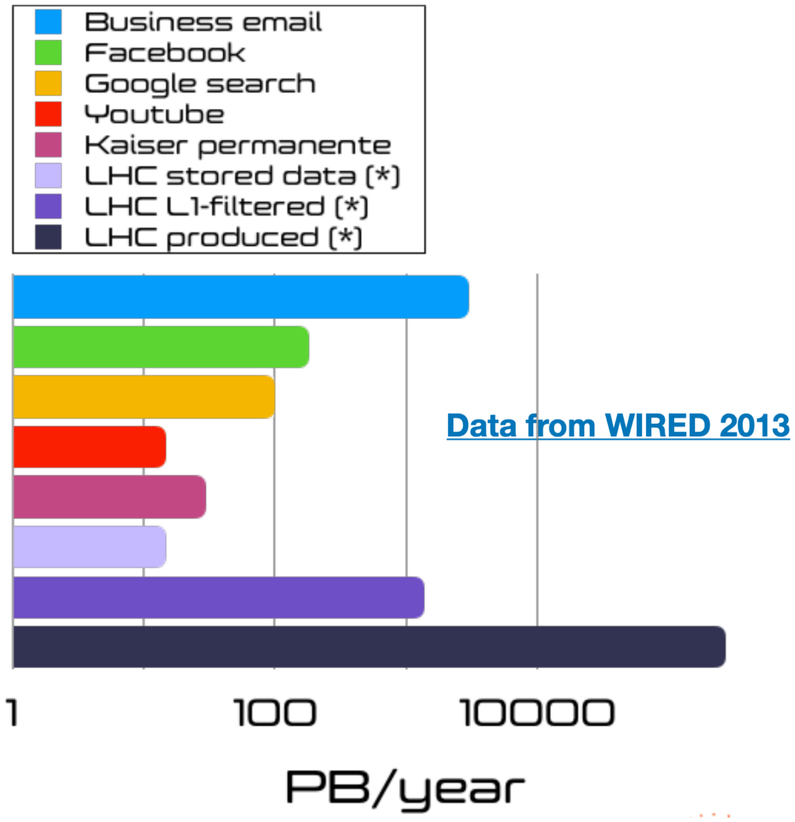
对抗性学习异常检测



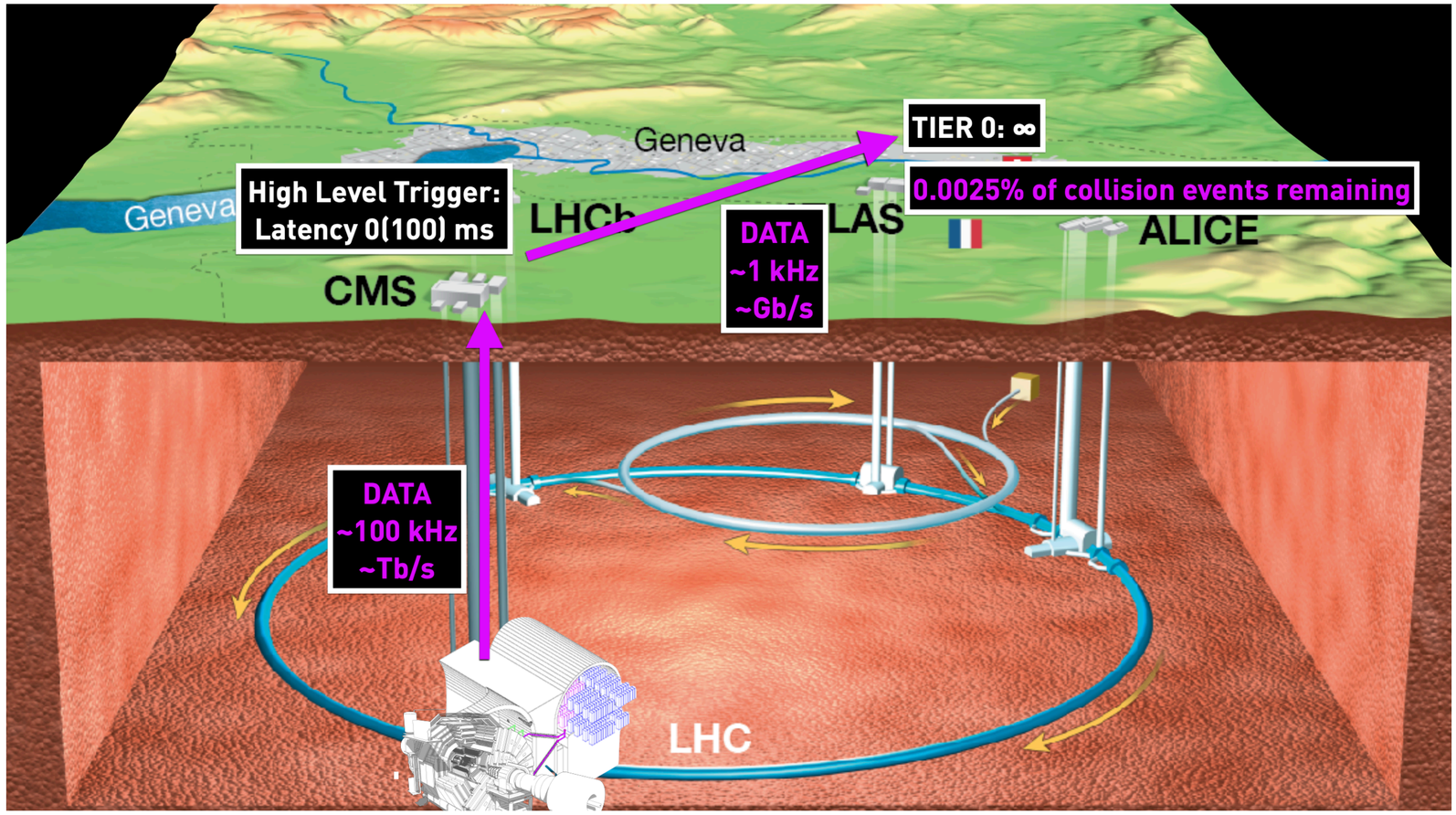
- 用Adversarially Learned Anomaly Detection (ALAD) 算法寻找数据中的异常
  - 基于GAN
- 用CMS公布的Run 1 open data
- 只对数据做最基本的筛选
  - 1 isolated lepton
  - 2 high pT jets
- 假设只知道有W+jets，Z这些标准模型的过程
- 数据中剩下的就是“新物理”
  - 重新“发现”顶夸克！

Rediscover top quark from real data
arXiv:2005.01598

LHC实验上的大数据挑战

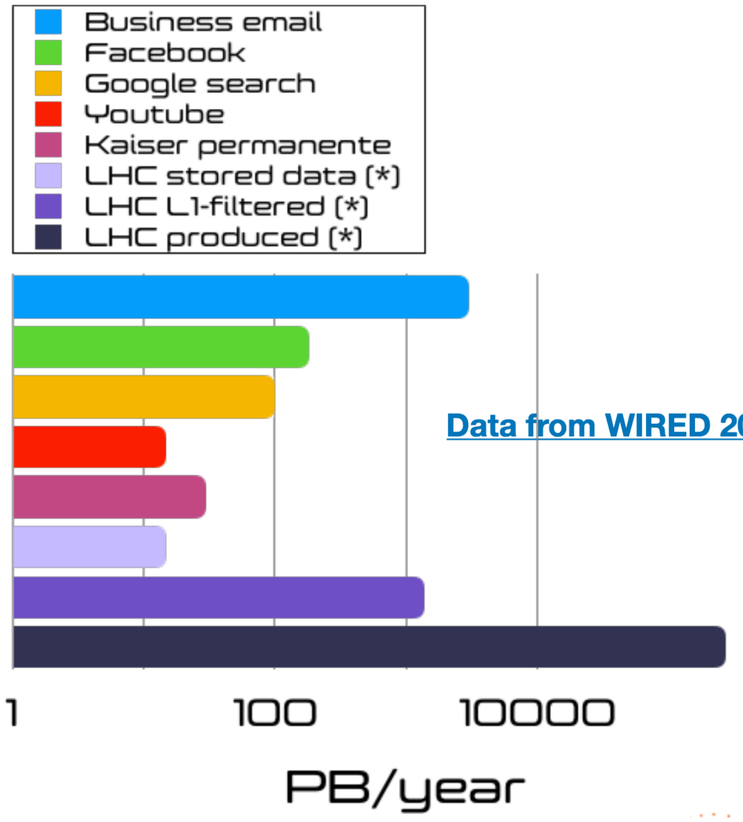


Data from WIRED 2013

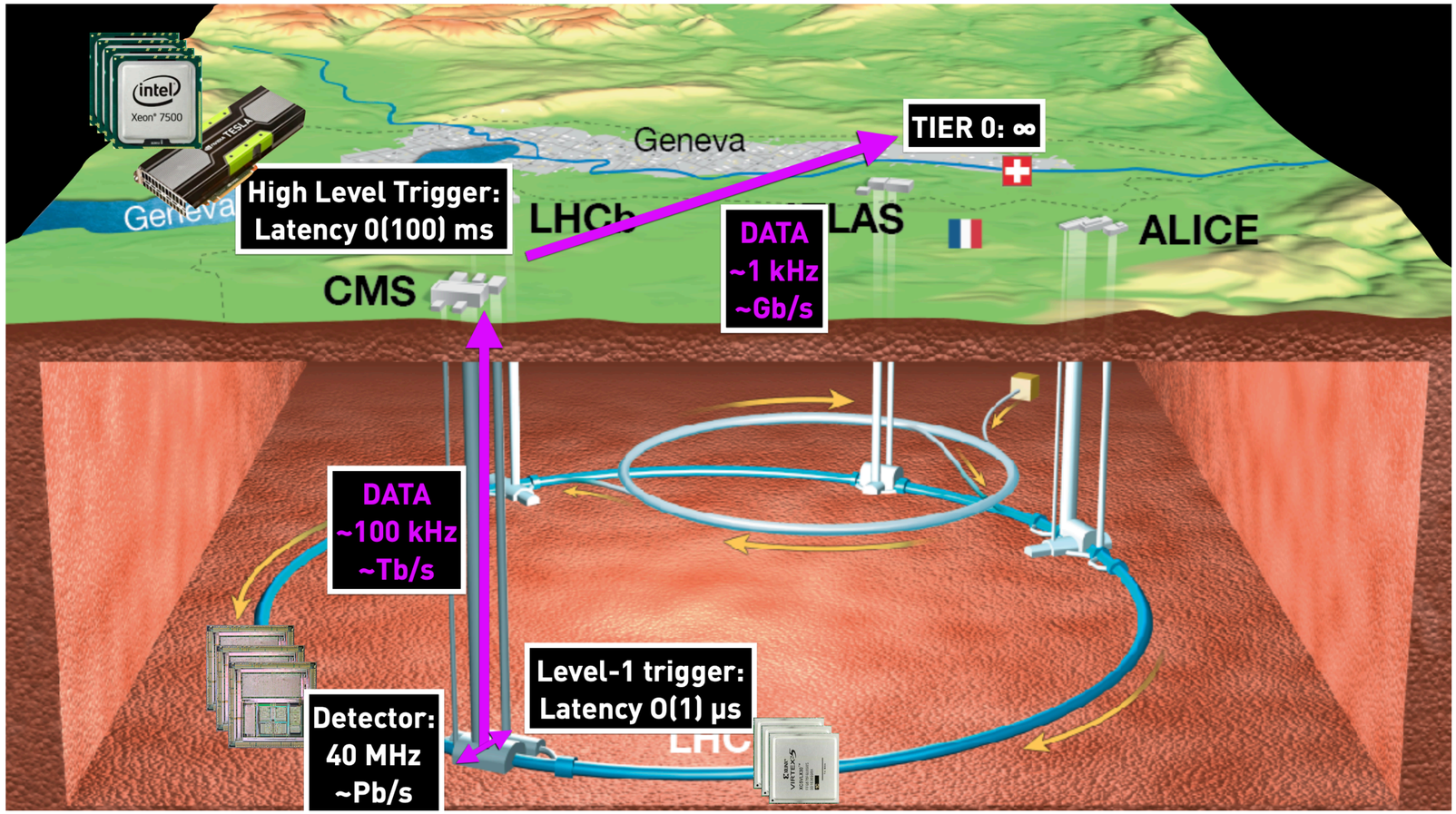


Taken from this talk

LHC实验上的大数据挑战



Data from WIRED 2013

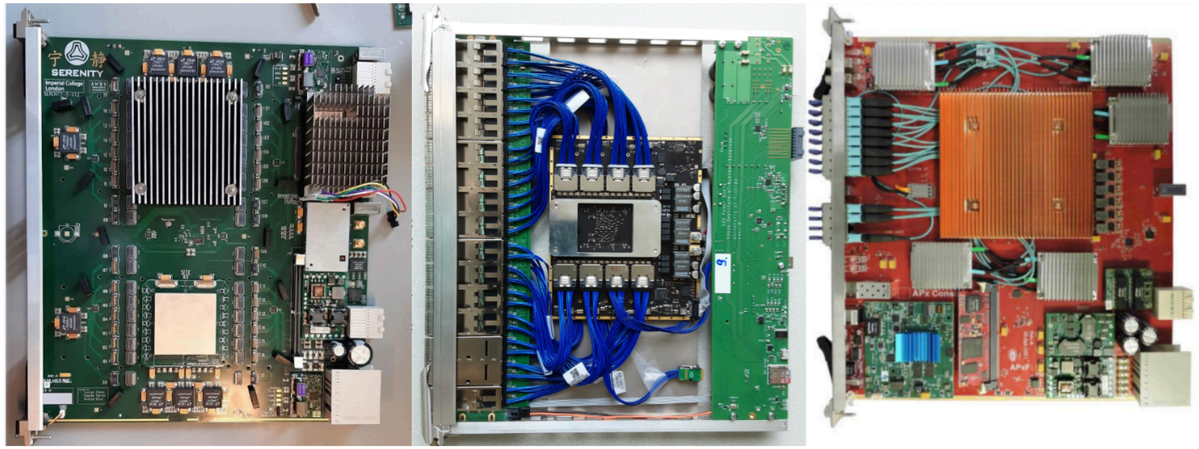


我们需要对数据进行实时的筛选

Taken from this talk

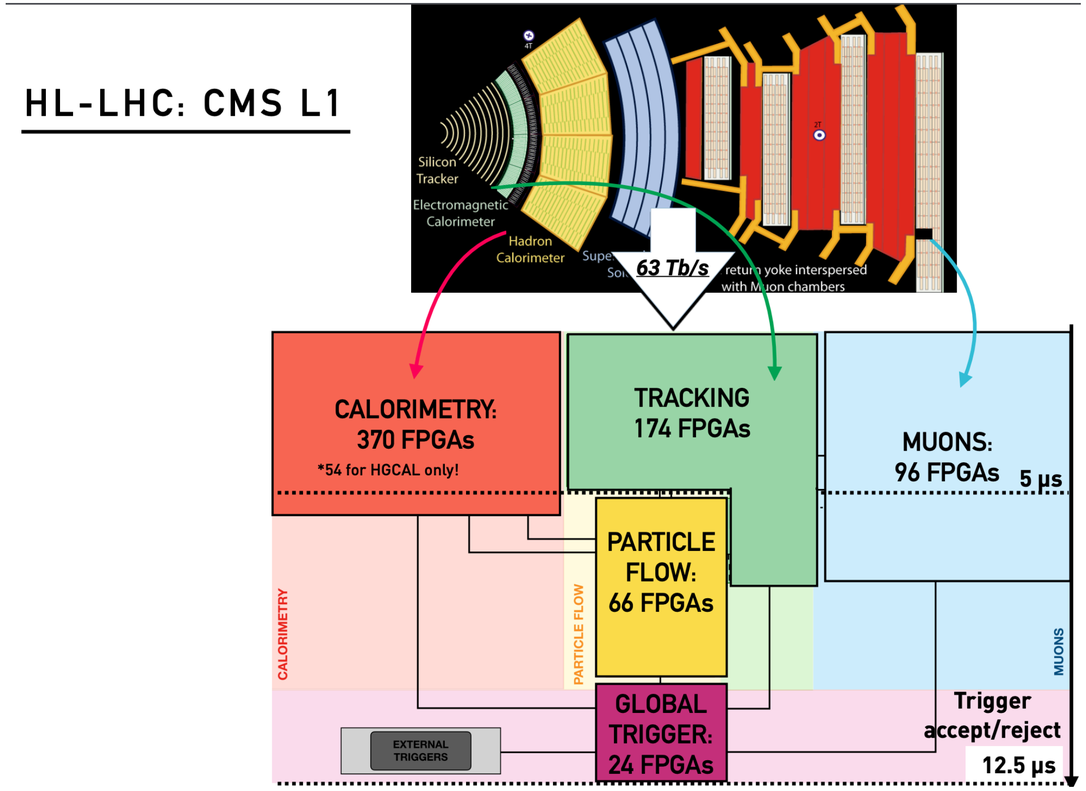
触发与机器学习

- 目前的触发分为L1-trigger (hardware based) 和High level trigger (HLT, software based)
- 在HLT上可利用前面提到的机器学习算法获得更好的在线事例选取
- 在可编程逻辑芯片(FPGA)上实现机器学习算法也将成为未来的趋势



Processing boards currently under development for the Phase-2 Level-1 Trigger upgrade project. These prototypes feature large Xilinx FPGAs hosting the trigger algorithms and more than 100 Input/Output high-speed optical links (28 Gb/s) for receiving/transmitting the data. From left to right: Serenity, X2O and APx boards.

Credits: Michalis Bachtis



机器学习与粒子物理学家

Artificial Intelligence Tool Predicts Life Expectancy in Heart Failure Patients

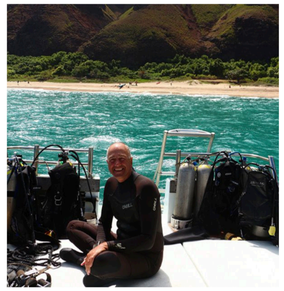
Algorithm developed by physicists and cardiologists achieved 88 percent success rate

November 13, 2019

By:
Michelle Brubaker

Share This:

When Avi Yagil, PhD, Distinguished Professor of Physics at University of California San Diego flew home from Europe in 2012, he thought he had caught a cold from his travels. When a "collection of pills" did not improve his symptoms, his wife encouraged him to see a doctor.



Avi Yagil, PhD, Distinguished Professor of Physics at University of California San Diego, back to his hobbies after a heart transplant.

Further tests revealed something far more life-threatening to Yagil than the common cold. "A chest X-Ray showed my lungs were flooded with fluid, and a subsequent echocardiogram found I had damage to my heart."

Yagil was diagnosed with heart failure. "UC San Diego Health cardiologists tried to manage my condition with medication, but all systems were failing as my heart struggled to keep me alive."

In June 2016, Yagil received a heart transplant. "I consider June 17 my second birthday."

While Yagil recovered from surgery, he began thinking about how he could improve the process for patients like him.

"In my day job, I use machine learning to understand a vast amount of information and measurements of particles and how they interact," he said. "The human body is even more complex, but the medical profession isn't utilizing the technologies that are needed to capture the multi-dimensional correlations between the measurements, such as lab tests and vital signs, and the outcomes. We hypothesized that such methodology and techniques could contribute to improving the prognosis and treatment of heart patients with heart failure."

Link to web

MARKER-HF™ Calculator

Enter values to calculate the MARKER-HF™ score based on *Improving risk prediction in heart failure using machine learning* Eric D. Adler *et al.*, published in *European Journal of Heart Failure*

Diastolic pressure (mm Hg):		(20-120)
Creatinine (mg/dL):		(0-25)
Blood Urea Nitrogen (mg/dL):		(0-160)
Hemoglobin (g/dL):		(2-20)
White Blood Cell Count (10^3µL):		(0-40)
Platelets (10^3µL):		(0-1500)
Albumin (g/dL):		(0-6)
Red Blood Cell Distribution Width (%):		(10-30)

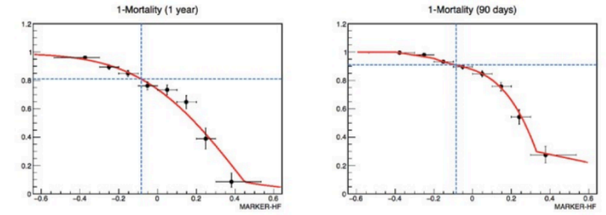
Calculate MARKER-HF | Clear fields

UC San Diego
Altman Clinical and Translational Research Institute

Results:

Marker-HF™:	-0.085
One-year Survival Probability (i.e., 1-Mortality):	0.81
90-day Survival Probability (i.e., 1-Mortality):	0.91

The values of Survival Probability (i.e., 1-Mortality) are calculated from the value of MARKER-HF and the red curves shown



Back

总结

- 机器学习作为一个有效的工具已经被广泛的运用在粒子物理领域里
  - 分类、提高测量精度、重建、模拟、触发。。。
- 可根据需要选取合适的工具和模型，关注工业界前沿和其他领域，避免闭门造车
- 确保理解输入的数据，避免过度拟合, machine learning is not god
- Have fun！

Backup

- Useful materials or links
  - https://iml-wg.github.io/HEPML-LivingReview/
  - https://jduarte.physics.ucsd.edu/capstone-particle-physics-domain/README.html
  - https://cms.cern/news/real-time-analysis-cms-level-1-trigger
  - https://indico.in2p3.fr/event/20424/contributions/92502/attachments/62853/86263/IWAPP_Feb2021.pdf
  - https://indico.cern.ch/event/1223711/attachments/2562341/4416719/CERNDS_fastml.pdf