非监督学习：生成式对抗网络
Generative Adversarial Networks

张 瑞

rui.zhang@cern.ch

TeV高能实验物理暑期学校

2025.08.15

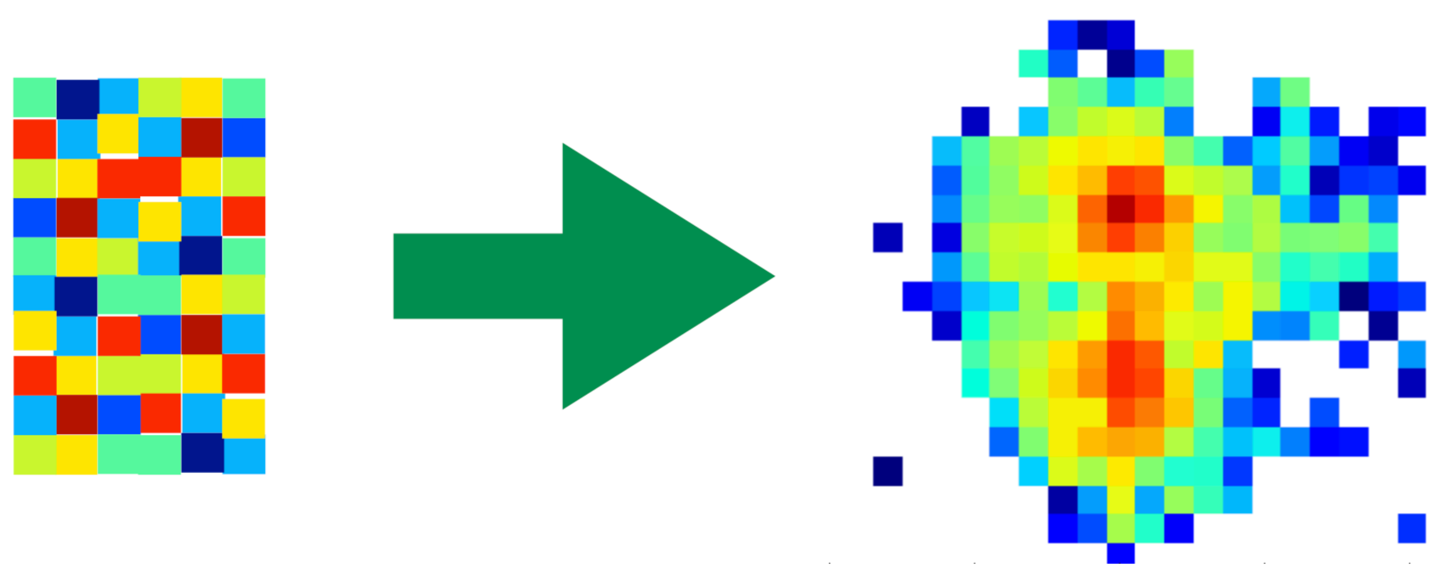
目 录

Generative network

Input → Network → Output

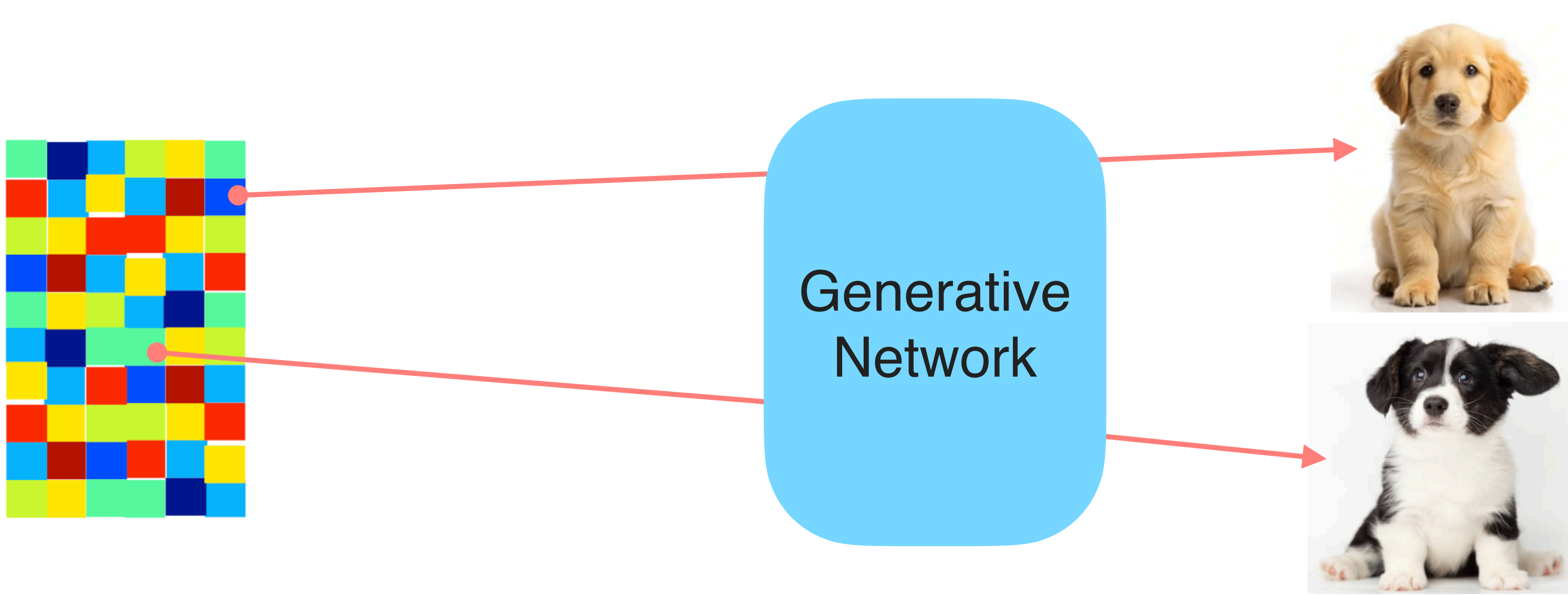
Input: complex → Classifiers → Output: simple

Input: simple → Generative networks → Output: complex



Generative networks

- 任务举例：随机生成若干张小狗的图片（图片大小16×16)
  - 每张图片都可以将所有像素点的RGB数值或灰度值构成的向量表示
  - 所有图片构成向量所在的一个空间，空间中的每个点可以唯一地表示一张图片
  - （假设：）内容相似的图片具有较小的距离，内容无关图片具有较大的距离
  - 因此，如果能在“小狗”区域内随机采集若干个点，即可生成若干张不同的小狗图片
- 构造一个生成模型，使得：
  - 输入为：高维高斯随机采样的点（高斯也可以改为均匀）
  - 输出为：16×16维的灰度值，经过重新排列得到图片



Generative network可以做什么

- 生成数据
  - 随机生成图片、视频，Sora
  - 聊天机器人，chatGPT
  - 生成稀缺的数据样本（如医学影像、少见物理过程），为监督学习提供数据支撑
- 重建与去噪

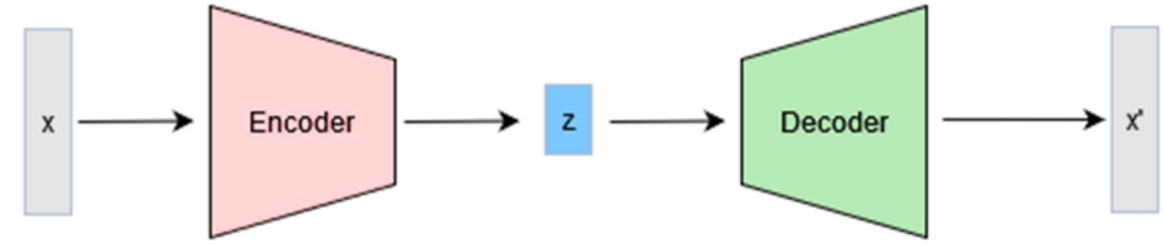
Input Image.

技术路线

- VAE

arXiv
https://arxiv.org › stat
[1312.6114] Auto-Encoding Variational Bayes
by DP Kingma · 2013 · Cited by 41189 — We introduce a stochastic v
algorithm that scales to large datasets and, under some mild different

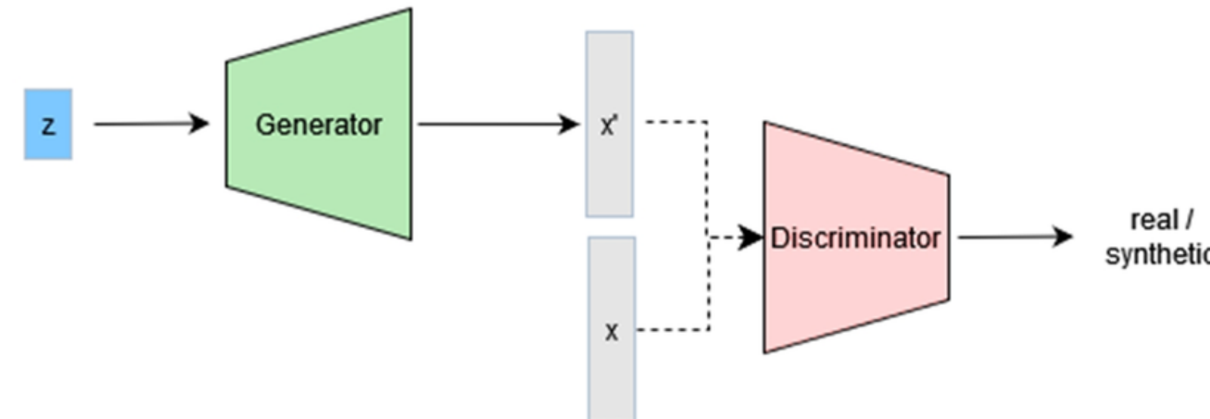
Variational Autoencoder



- GAN

arXiv
https://arxiv.org › stat
[1406.2661] Generative Adversarial Networks
by IJ Goodfellow · 2014 · Cited by 76074 — We propose a new fra
models via an adversarial process, in which we simultaneously trai

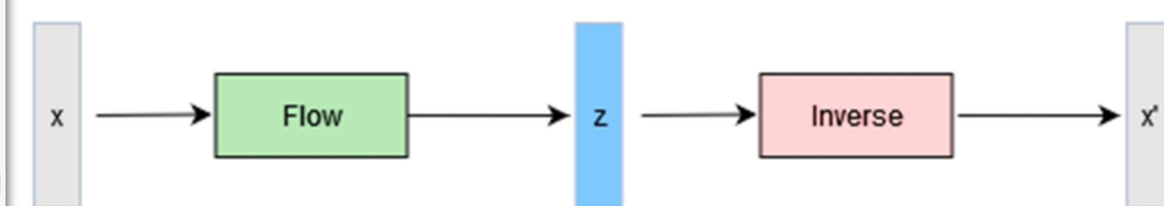
Generative Adversarial Network



- Normalising flow

arXiv
https://arxiv.org › stat
[1505.05770] Variational Inference with Normalizing Flows
by DJ Rezende · 2015 · Cited by 4633 — We introduce a new approach for specifyi
complex and scalable approximate posterior distributions.

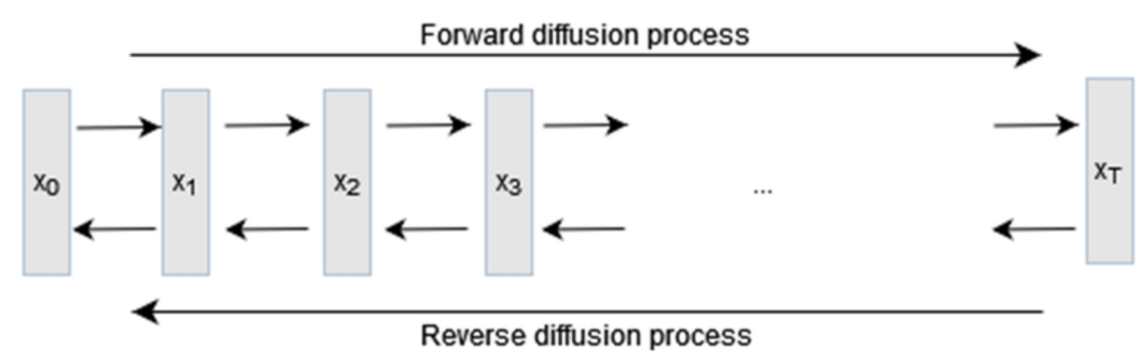
Normalising Flows



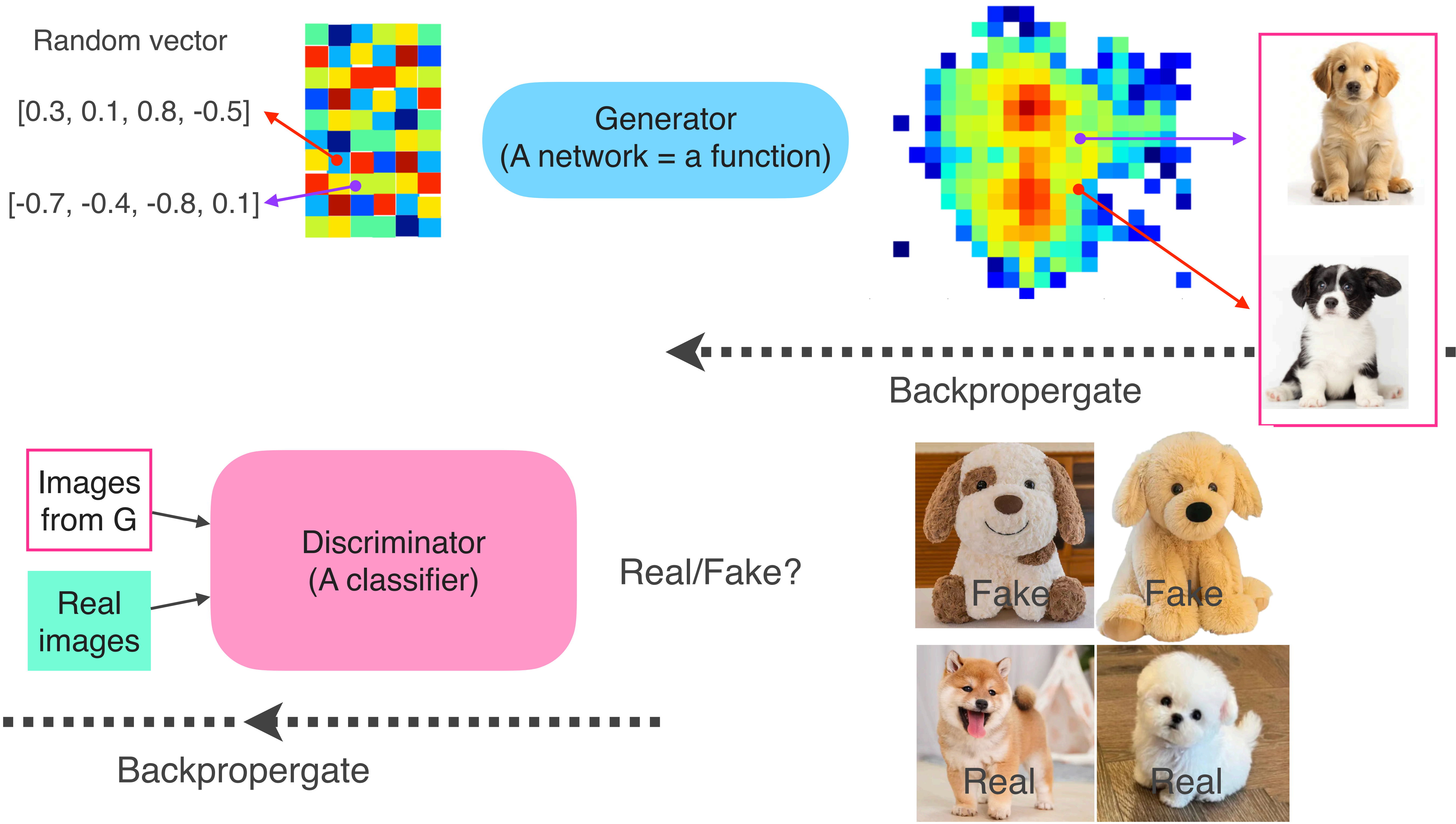
- Diffusion model

arXiv
https://arxiv.org › cs
[2006.11239] Denoising Diffusion Probabilistic Models
by J Ho · 2020 · Cited by 15959 — We present high quality image synthesis res
probabilistic models, a class of latent variable models inspired by consideration

Diffusion Model

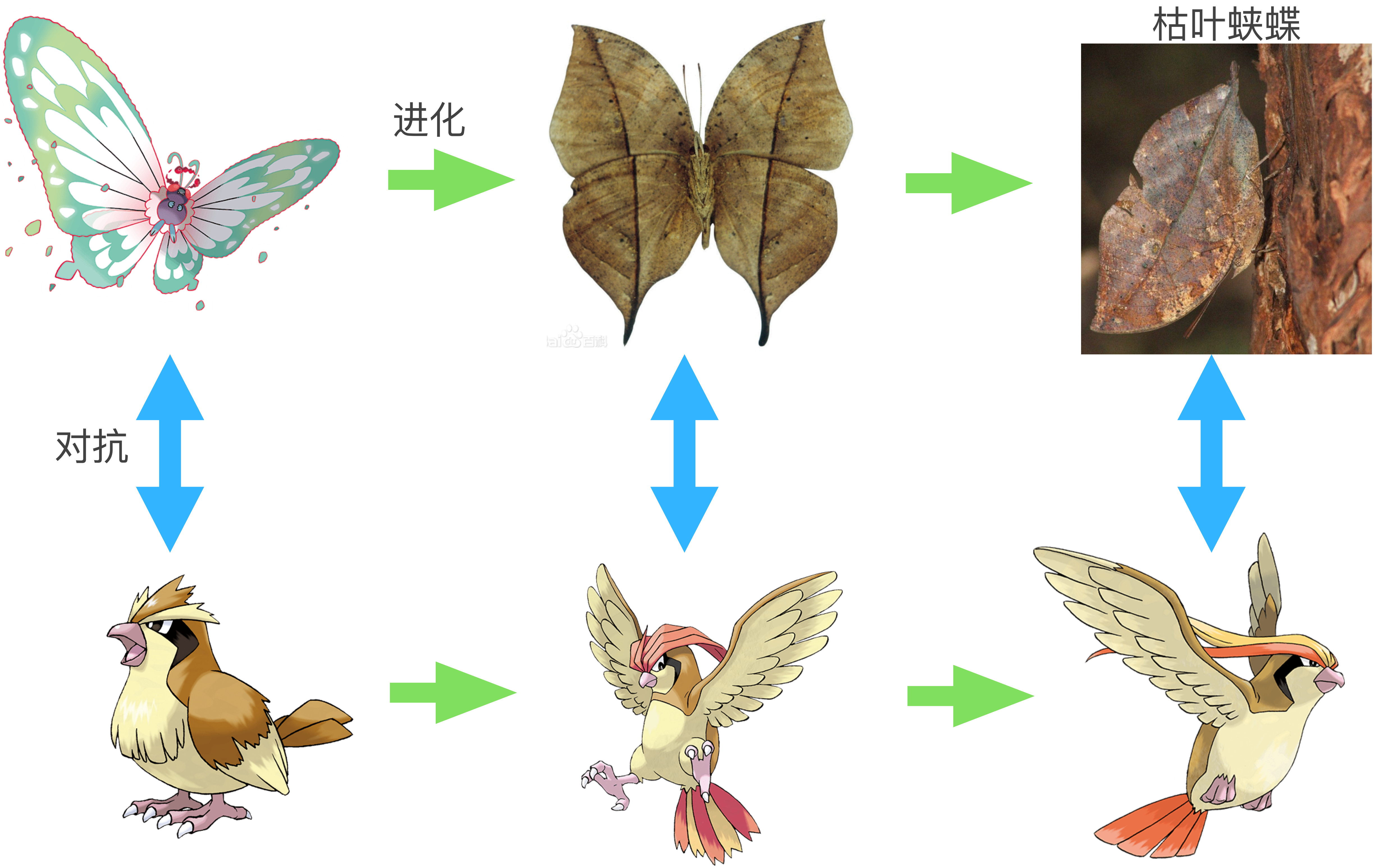


Generative Adversarial Network



- Generator和discriminator的架构可以自己设计，random vector的维度和分布也可以自行定义

Basic idea of "adversarial"

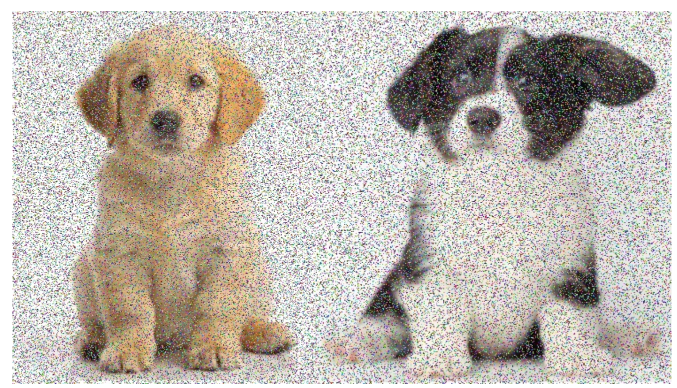


Basic idea of "adversarial"

Generator (t_0)

Generator (t_{10})

Generator (t_n)

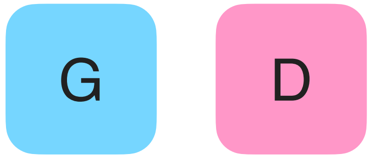


Discriminator (t_0)

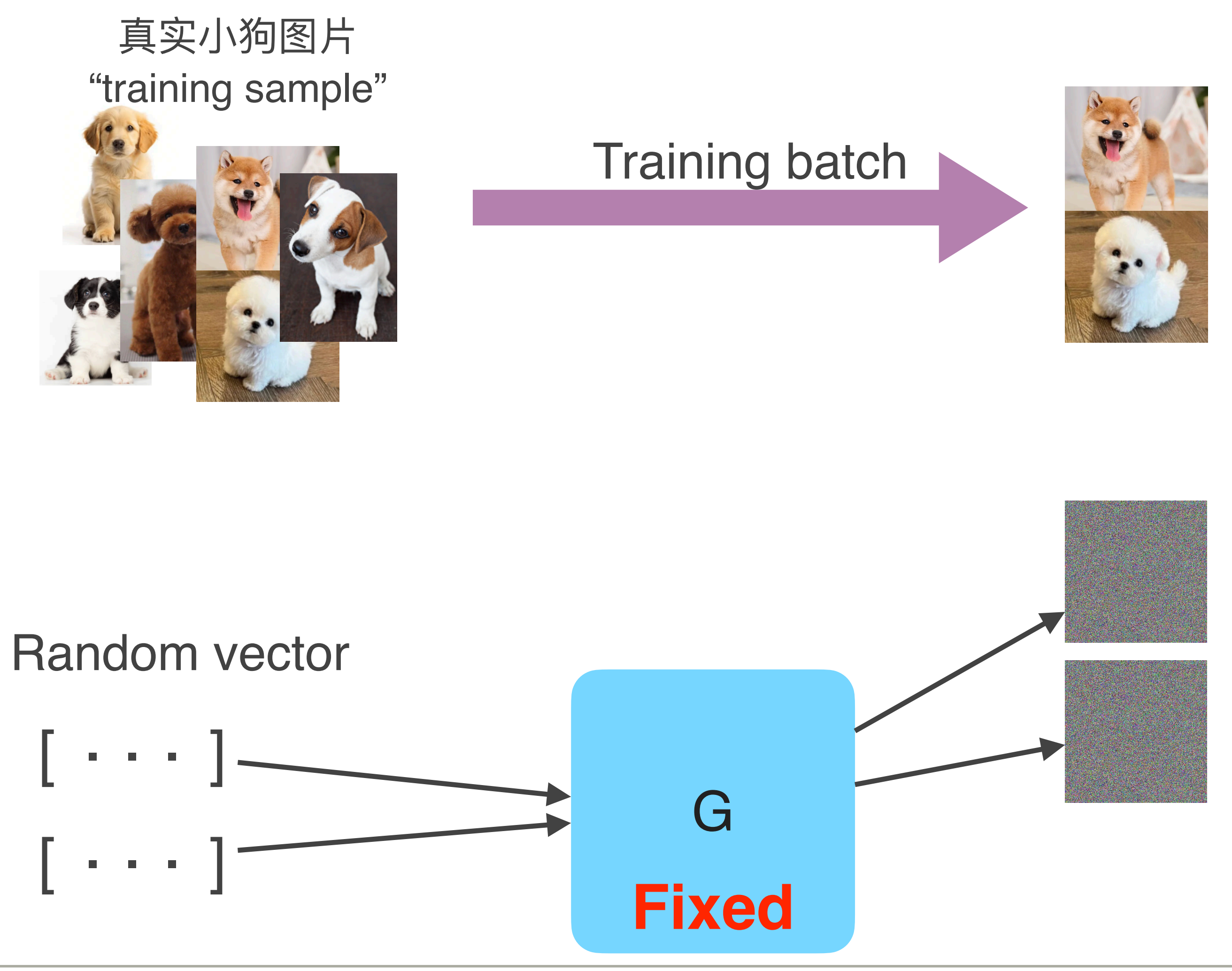
Discriminator (t_{10})

Discriminator (t_n)

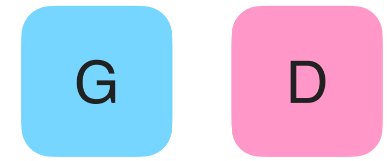
Algorithm of GAN



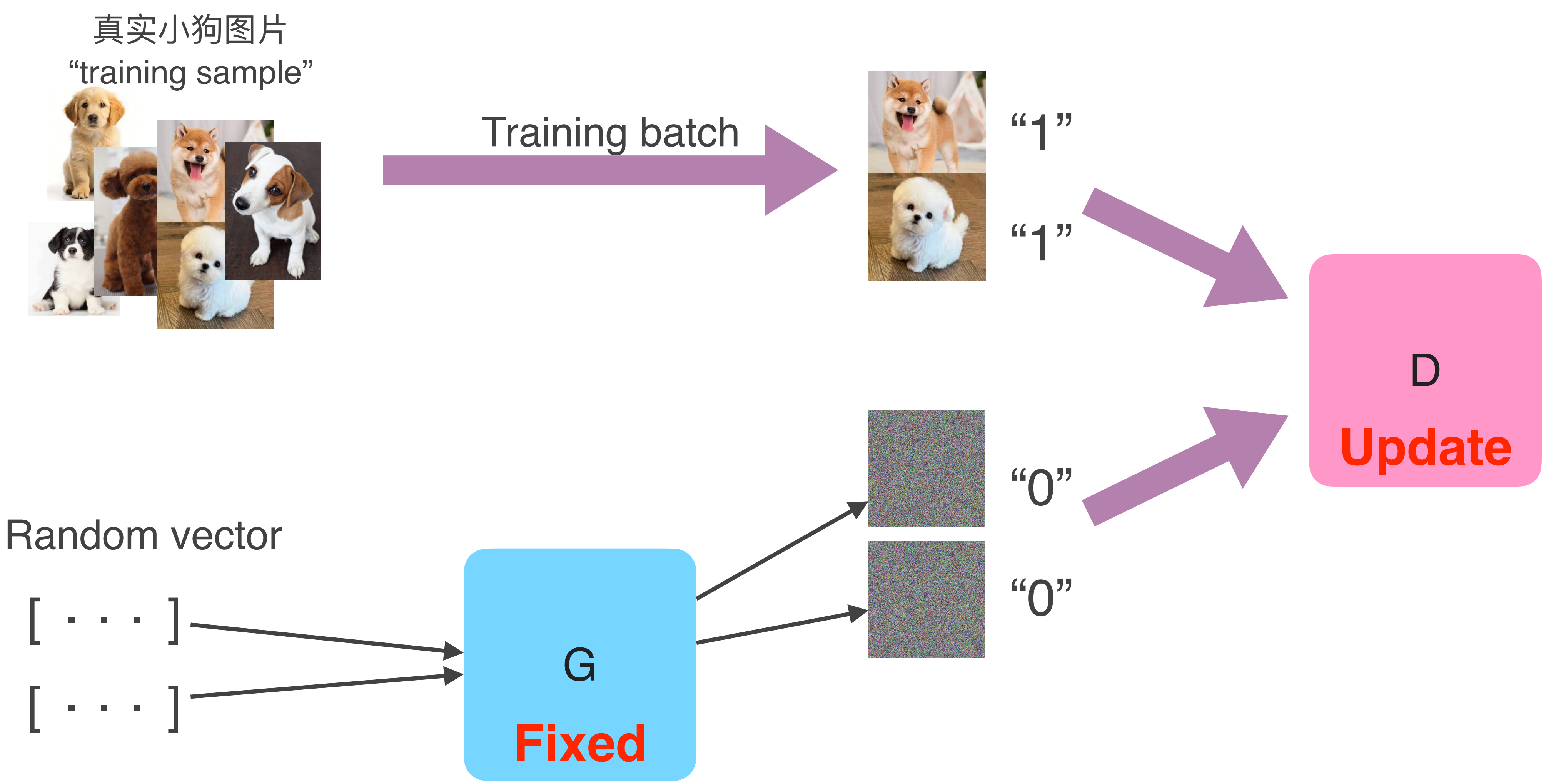
- Initialise generator and discriminator
- In each training iteration:
  - Fix generator G, and update discriminator D



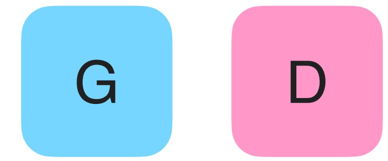
Algorithm of GAN



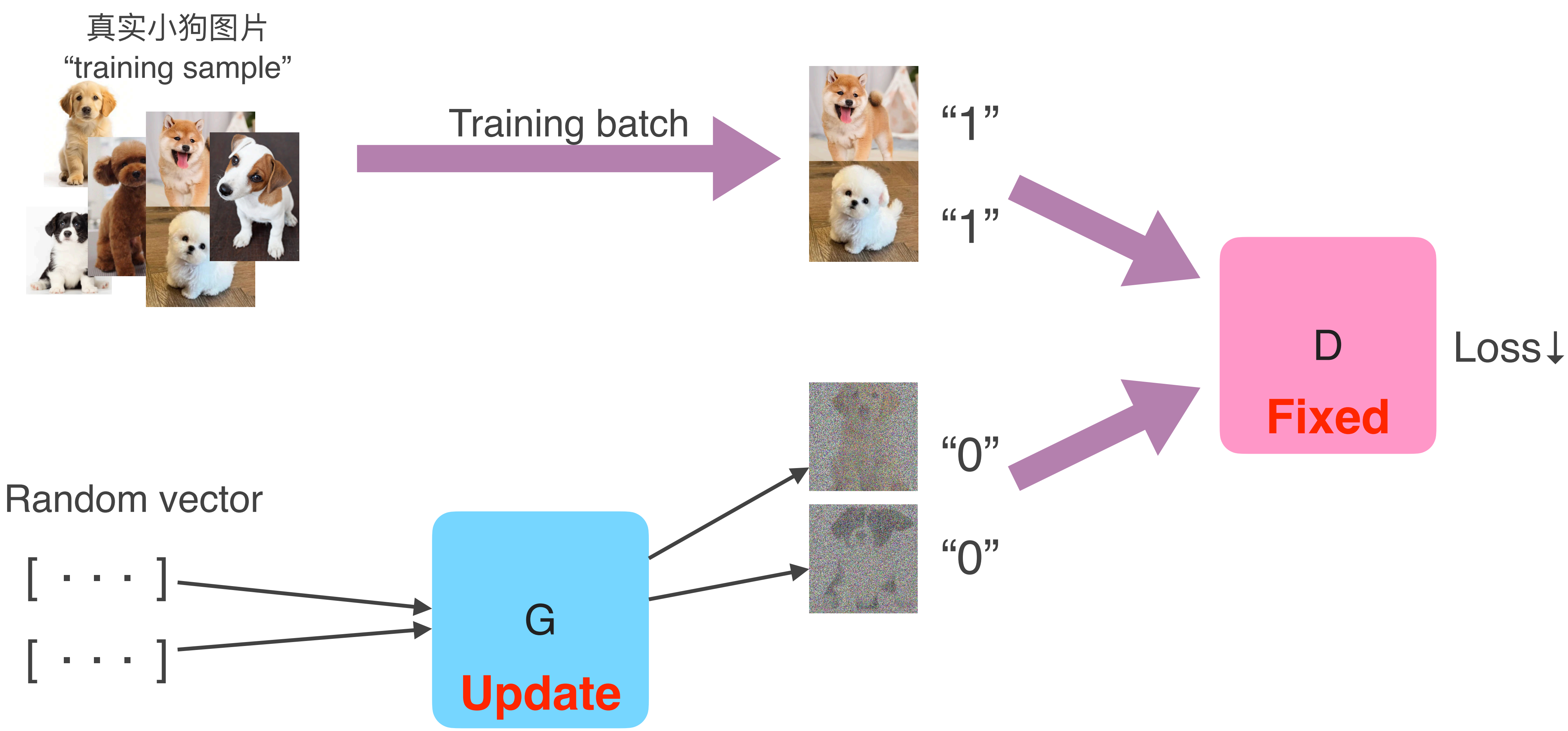
- Initialise generator and discriminator
- In each training iteration:
  - Fix generator G, and update discriminator D



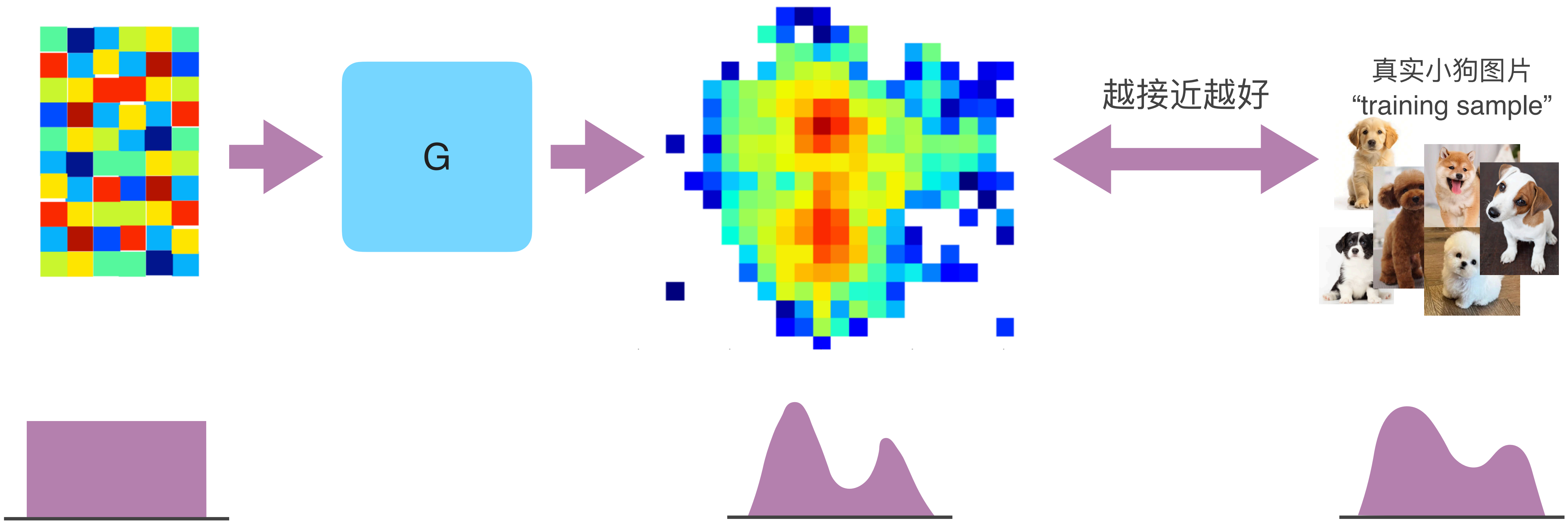
Algorithm of GAN



- Initialise generator and discriminator
- In each training iteration:
  - Fix generator G, and update discriminator D
  - Fix discriminator D, and update generator G (G learns to "fool" D)



Math behind GAN



$$G^* = \arg\min_G \underline{Div(P_G, P_{data})}$$

Divergence between distributions P_G and P_{data}

- 寻找到一个G（或一组G对应的神经元参数），使得divergence越小越好
- 困难：P_G and P_{data}未知，divergence难以计算

Divergence: a measure of distance

Math behind GAN

$$G^* = \arg\min_G Div(P_G, P_{data})$$

- GAN allows to calculate divergence with sampling from P_G and P_{data}

$$D^* = \arg\max_D V(D, G)$$

★ : data sampled from P_G

★ : data sampled from P_{data}

$$V(D, G) = E_{y \sim P_{data}}[\log D(y)] + E_{y \sim P_G}[1 - \log D(y)]$$

训练完之后得到的$\max_D V(D, G)$就是对Div的一个估计

Math behind GAN

$$G^* = \arg\min_G \boxed{Div(P_G, P_{data})}$$

- GAN allows to calculate divergence with sampling from P_G and P_{data}

$$D^* = \arg \boxed{\max_D V(D, G)}$$

★ : data sampled from P_G

★ : data sampled from P_{data}

$$V(D, G) = E_{y \sim P_{data}}[\log D(y)] + E_{y \sim P_G}[1 - \log D(y)]$$

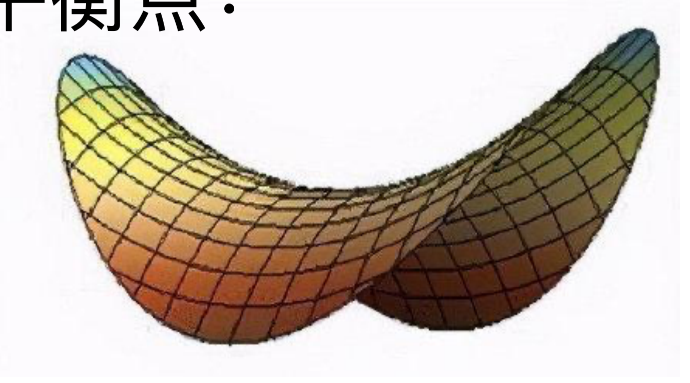
训练完之后得到的$\max_D V(D, G)$就是对Div的一个估计

$$\boxed{G^* = \arg\min_G \max_D E_{y \sim P_{data}}[\log D(y)] + E_{y \sim P_G}[1 - \log D(y)]}$$

MinMax problem

$$G^* = \arg\min_G \max_D E_{y\sim P_{data}}[\log D(y)] + E_{y\sim P_G}[1 - \log D(y)]$$

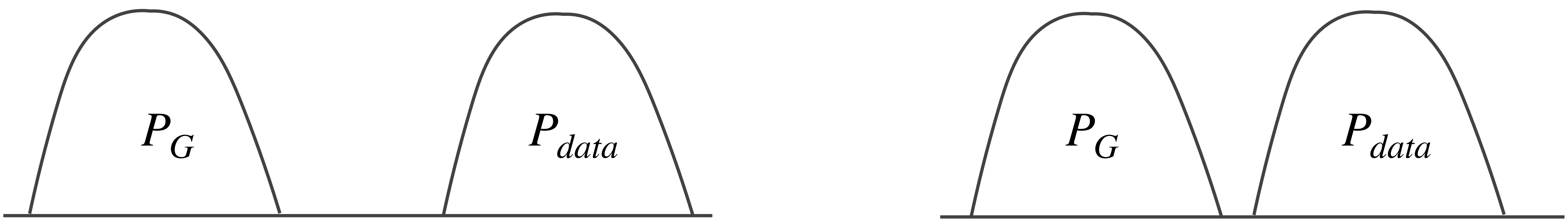
- GAN的训练过程可以看作是在马鞍形曲面上寻找平衡点：
  - 生成器的优化方向是“向下滑坡”（寻找最小值）
  - 判别器的优化方向是“向上爬坡”（寻找最大值）



- GAN是很难训练的，许多研究工作致力于优化其训练过程

GAN的改进

- 当D足够强大时，可以完美地区分P_G and P_{data}，此时无法为G提供正确的梯度



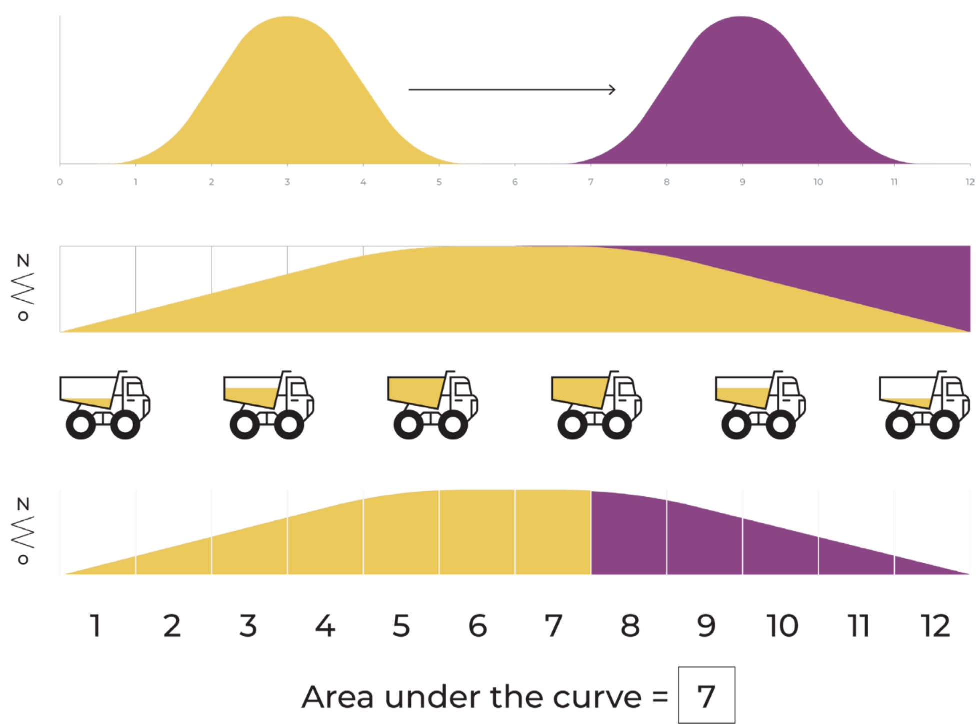
- 但我们希望右图D的loss小于左图，因为两个分布的距离变近了
- 为解决该问题，Wasserstein GAN（WGAN）对原始的GAN的loss函数做了改进，提出了使用Wasserstein distance（也称为Earth Mover's Distance）作为距离的测度

arXiv
https://arxiv.org › stat

[1701.07875] Wasserstein GAN

by M Arjovsky · 2017 · Cited by 17925 — We introduce a new algorithm named WGAN, an alternative to traditional GAN training. In this new model, we show that we can improve the stability of learning.

WGAN的改进



- Wasserstein距离的优势在于，即使两个分布没有重叠，它仍然能够提供有意义的梯度，从而缓解模式崩溃和梯度消失问题。
- 为了实现Wasserstein距离，WGAN要求discriminator满足Lipschitz连续性，即D的梯度范数必须被限制在一个常数范围内。$||\Delta_x D(x)|| \leq K$
- Gradient Penalty的loss函数形式为$L_{GP} = \lambda E_{x \sim P_x}[(||\Delta_x D(x)||_2 - 1)^2]$，通过$\lambda$调节相对大小
- 将这一项加到WGAN的loss函数中，能够使得WGAN更好地收敛
  - WGAN-GP

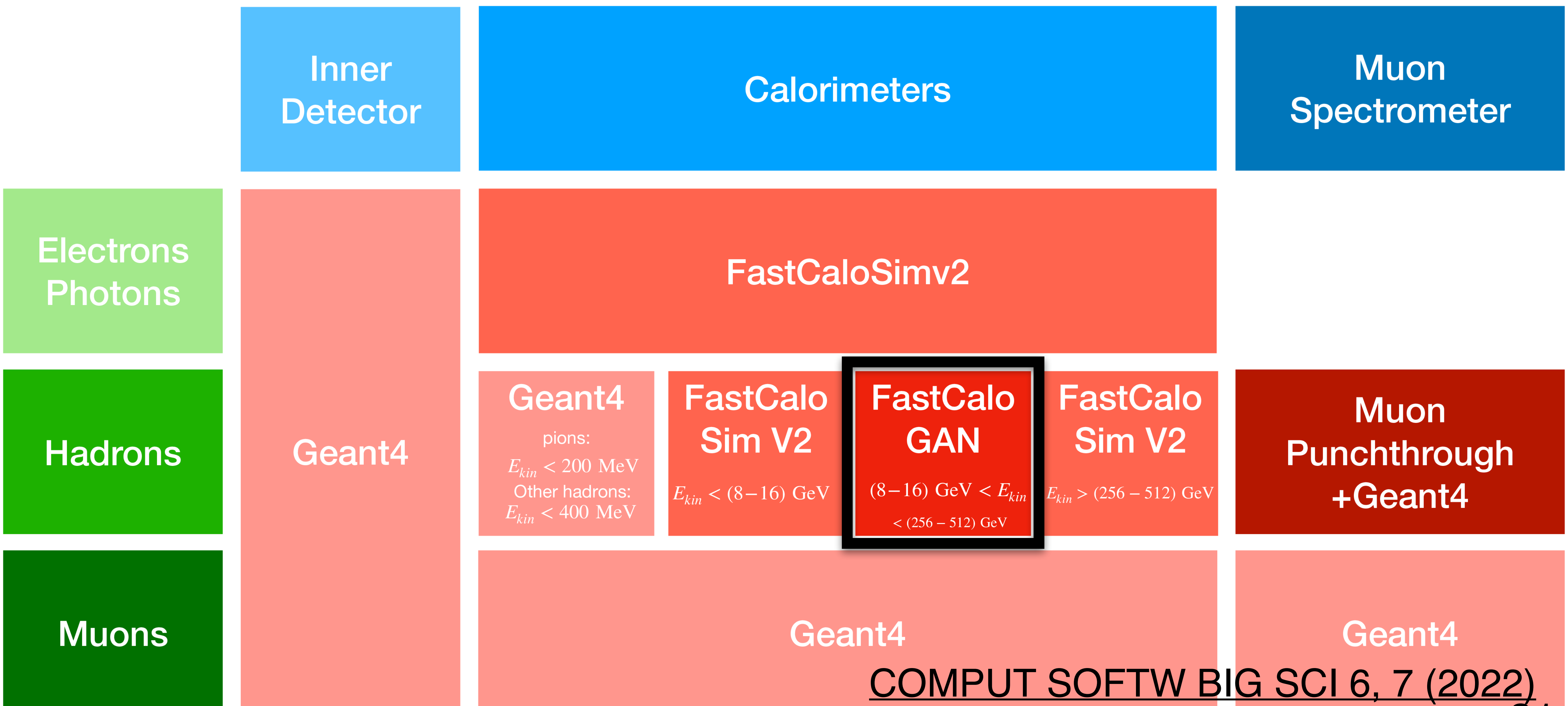
Tutorial notebook

- https://gitlab.cern.ch/zhangruiPhysics/tutorials

目 录

应用

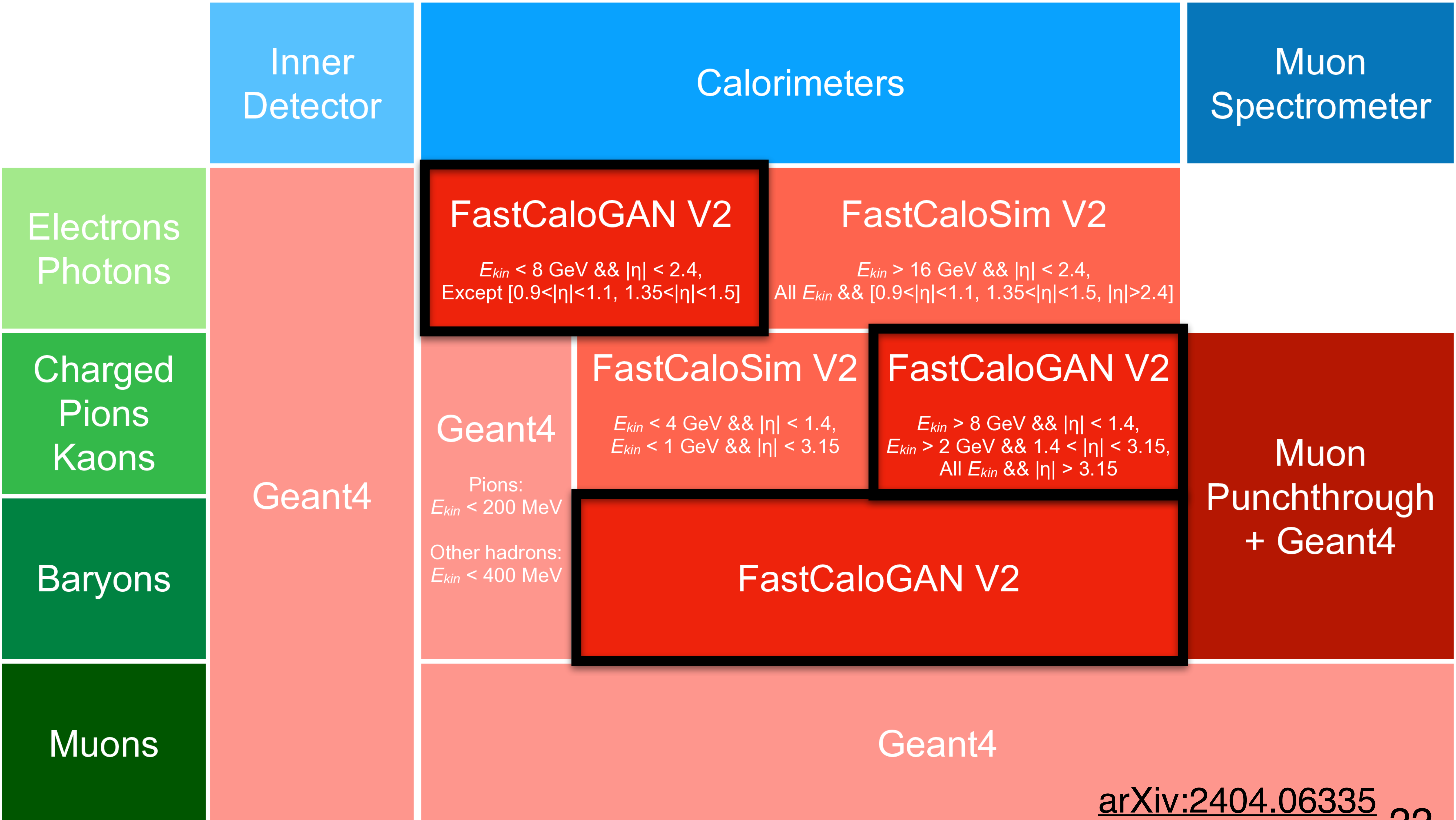
- 2022年，ATLAS率先在量能器快速模拟中使用了机器学习算法



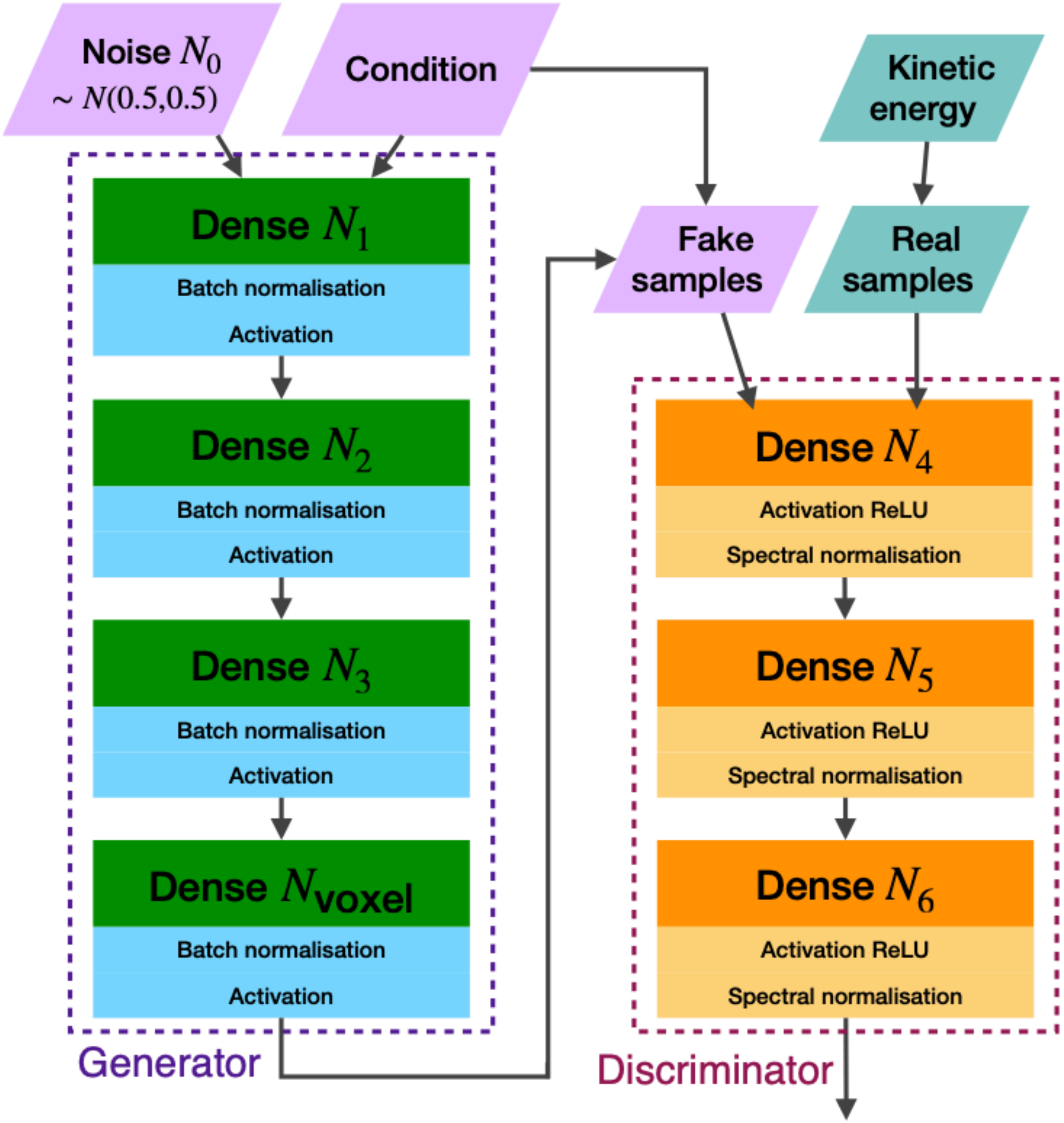
COMPUT SOFTW BIG SCI 6, 7 (2022)

应用

- 2024年，FastCaloGAN性能进一步优化，在探测器中应用的范围更广



FastCaloGAN模型



In the tutorial, less number of layers are used.

从prototype到production

Outline

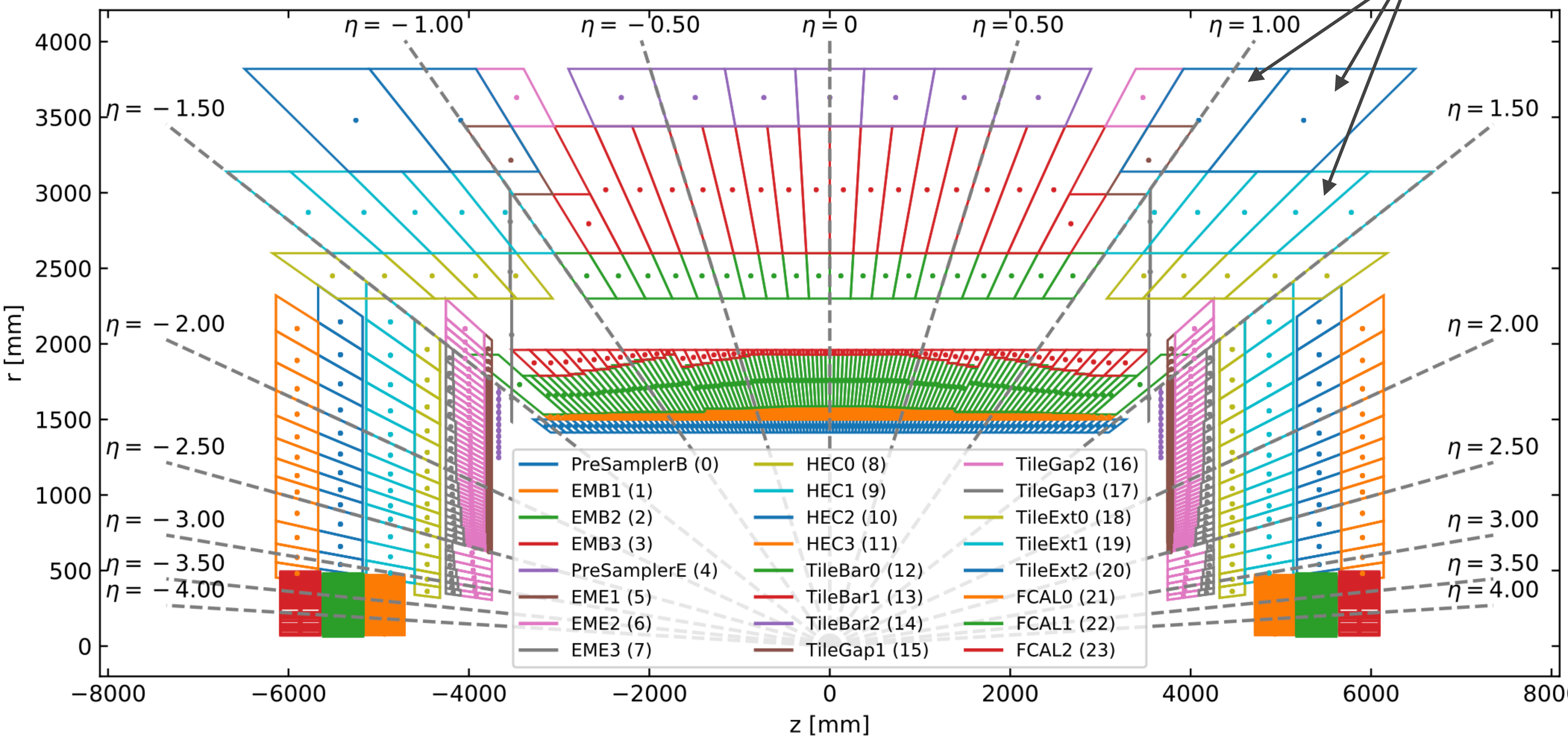
- Developing a good model is not enough, you need to know your input samples and your detector
- Many (not so) small effects need to be considered to produce objects that match G4
  - Energy corrections 能量修正
  - Geometry corrections 量能器几何修正
  - Assign energy from voxels to cells 能量分配
- Other parameters and effects that are relevant in a production tool that you can neglect in a ML project
  - Scale up issues 拓展到全空间
  - Time and memory requirements 时间和内存挑战
  - Integration with other models 软件嵌入

Source

ATLAS几何结构

- 几何结构极其复杂，最大限度地提高探测器的覆盖范围和性能
- 导致在η方向物质分布不均匀

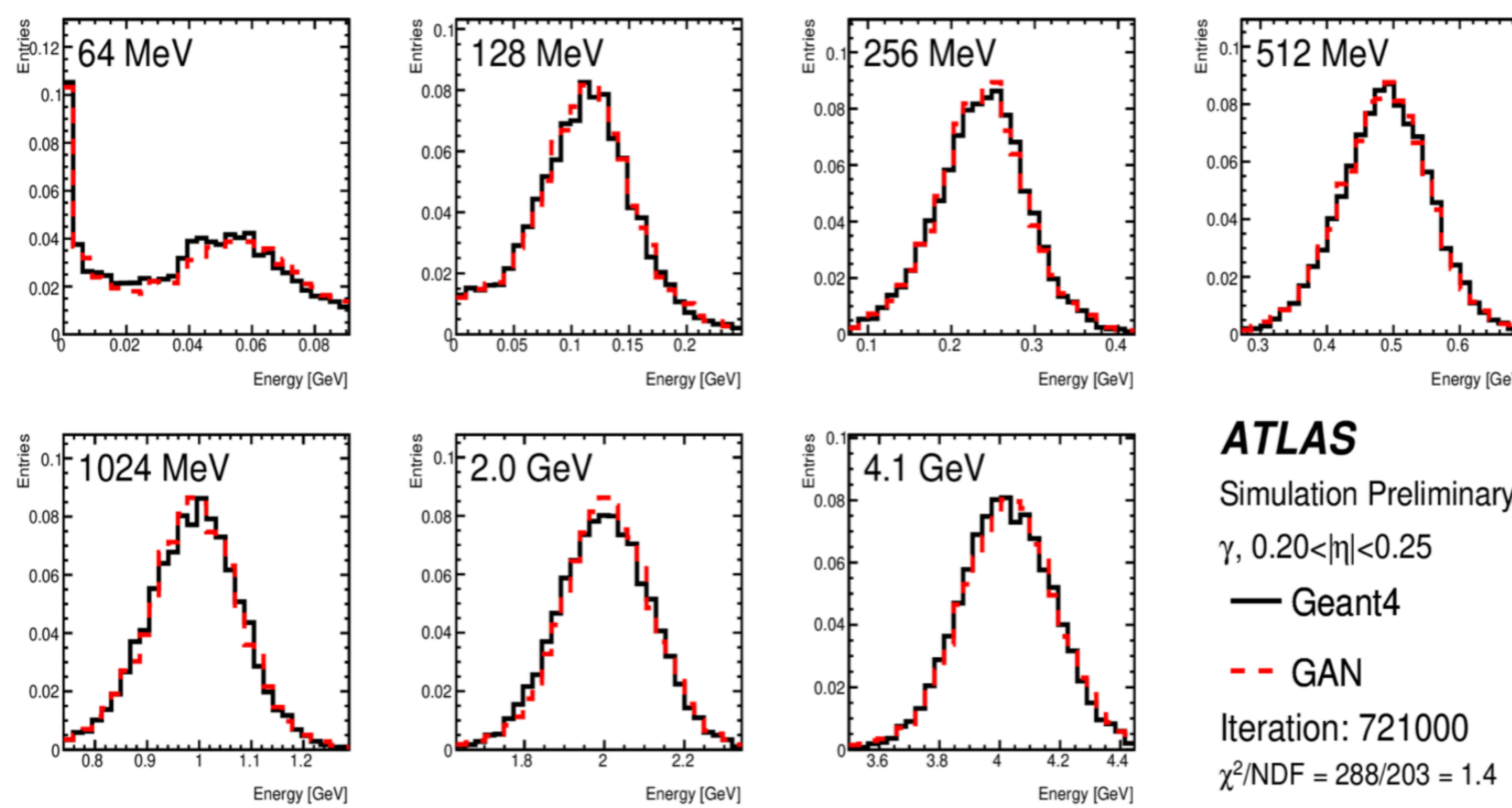
量能器晶体块



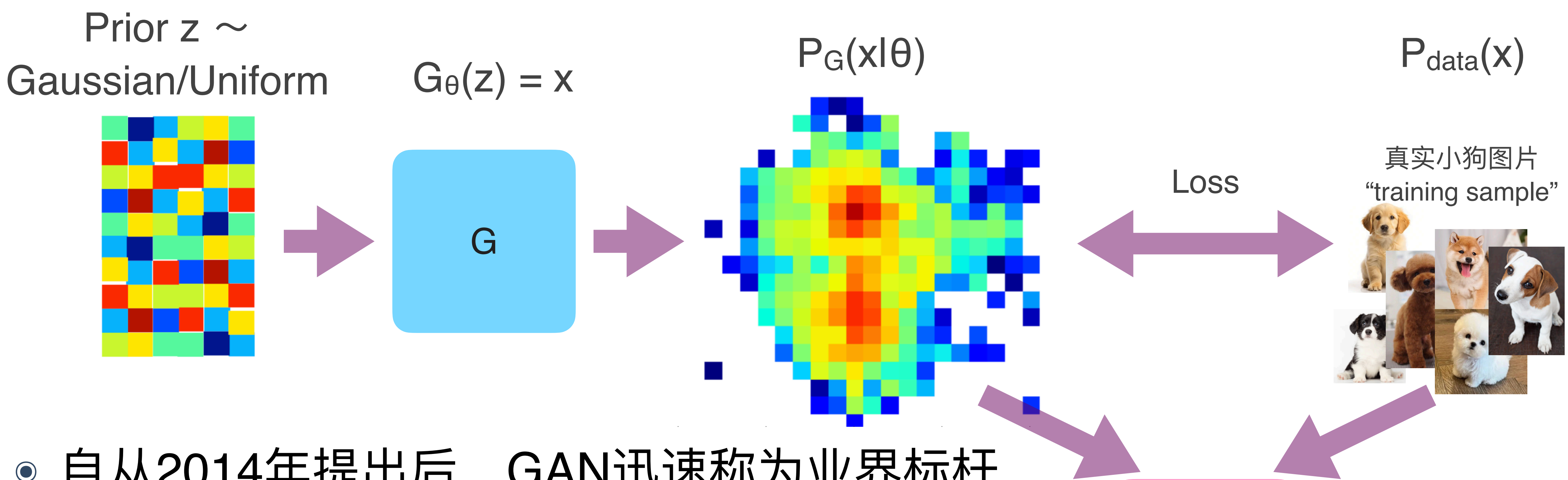
对撞点

ATLAS策略

- 当前ATLAS策略是将探测器在|η|方向划分为100等份，每份训练一个GAN模型
  - 由于一些特殊区域的响应差异显著，以|η|做条件输入的方式不足以捕捉这些差异
- 不同粒子的簇射差异显著，针对不同粒子分别训练一个GAN模型
  - e, γ 主要通过电磁相互作用产生簇射
  - K, π , p 既涉及电磁相互作用，又包括强相互作用
- 针对不同能量的粒子，将能量作为GAN的条件输入进行训练，以更好地捕捉能量依赖的簇射特性。



总结



- 自从2014年提出后，GAN迅速称为业界标杆
- 最大的特点是训练困难
- GAN可能是第一个变种特别多的出圈模型，因此有个戏称是“GAN动物园”https://github.com/hindupuravinash/the-gan-zoo
  - 今天只涉及到GAN的基本知识以及最有名的WGAN+GP
- GAN没有过时，但生成模型的新宠是Diffusion model
  - DALL·E 2、Stable Diffusion以及Imagen等业界产品均采用了扩散模型技术

报 告
提 纲

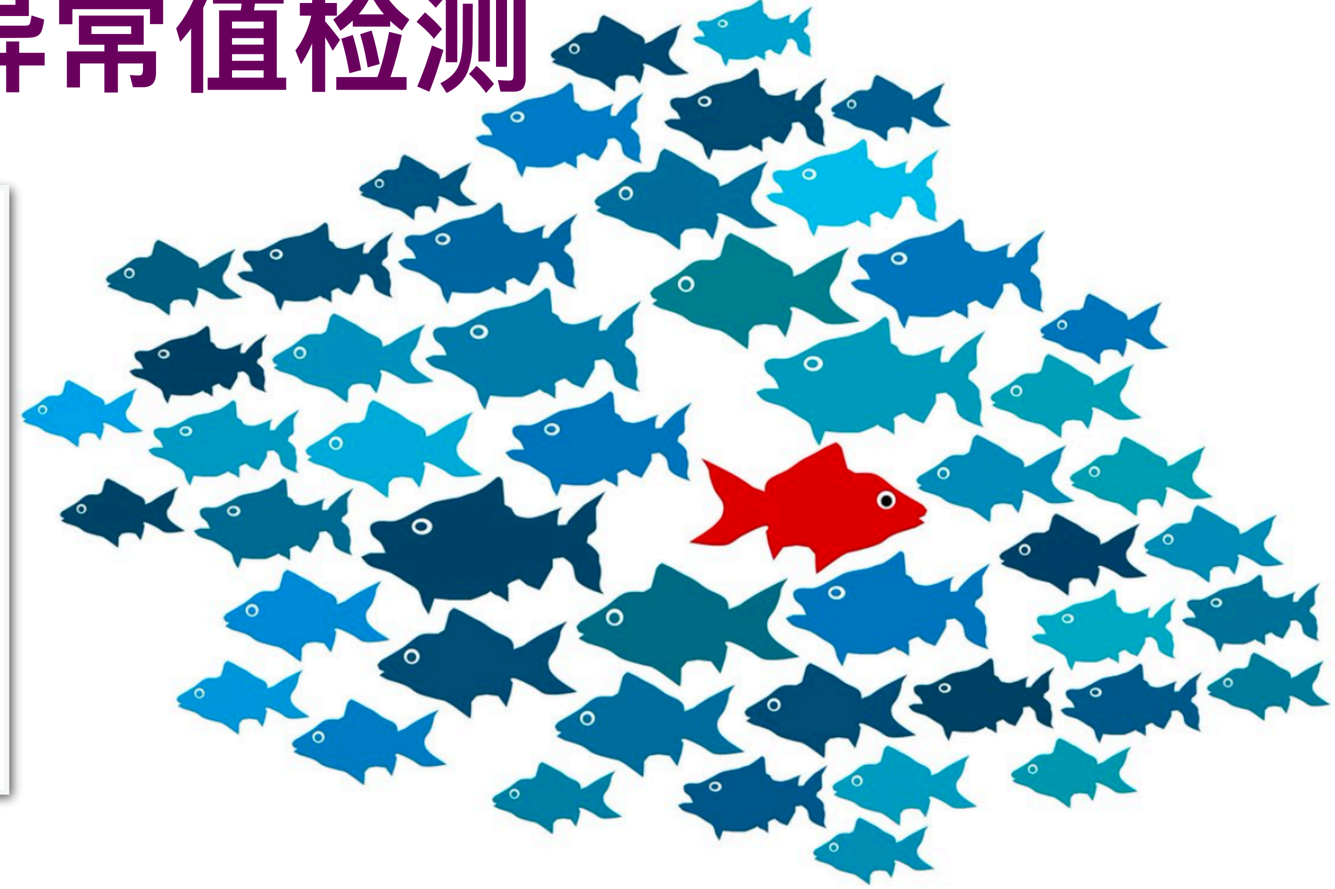
1. GAN介绍

2. GAN应用

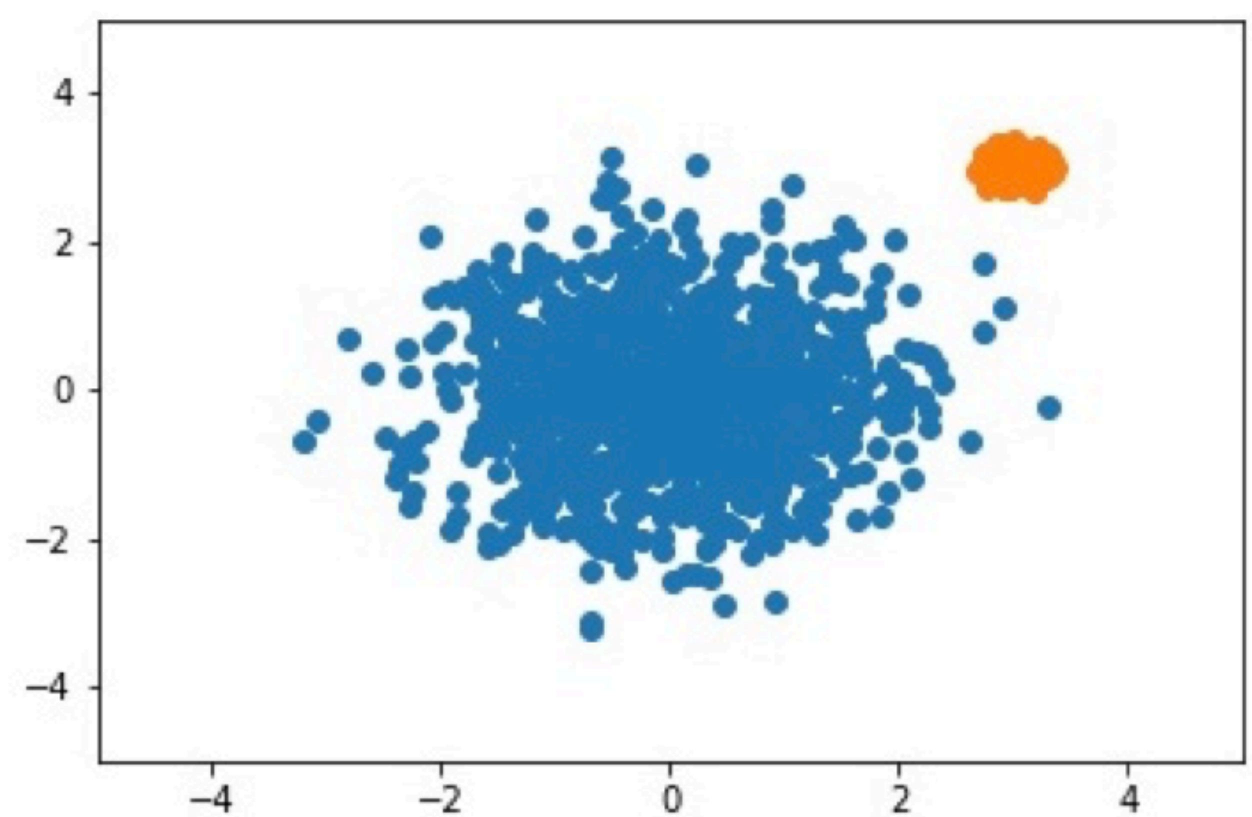
3. 异常值检测

深度学习异常值检测

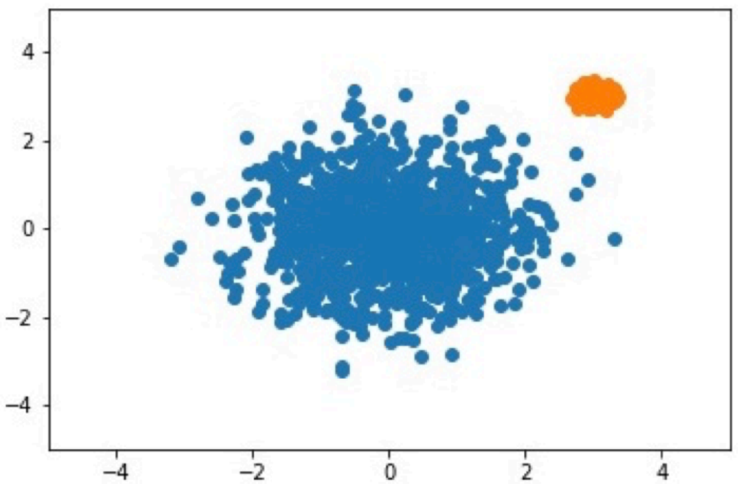
在数据挖掘中，**异常检测**（英語：Anomaly detection）对不符合预期模式或数据集中其他项目的项目、事件或观测值的识别。[1]通常异常项目会转变成银行欺诈、结构缺陷、医疗问题、文本错误等类型的问题。异常也被称为离群值、新奇、噪声、偏差和例外。[2]



- 高能物理实验的主要目标之一是探索“超出标准模型”的新物理现象
- 相较于标准模型的过程，这些超标准模型现象可能表现出不同的动力学特征
    - 可以尝试利用“异常值检测”技术来识别这些潜在的信号

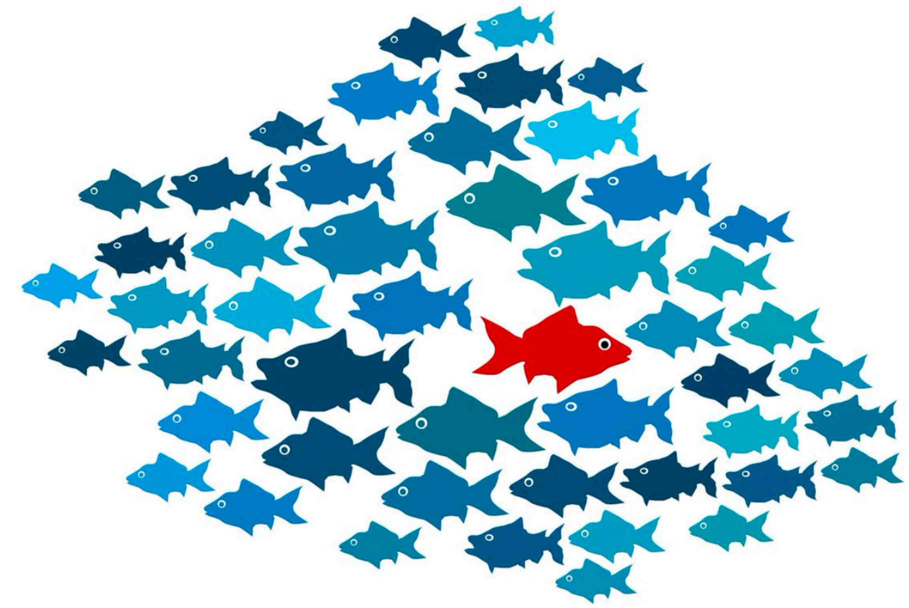


异常值检测的分类

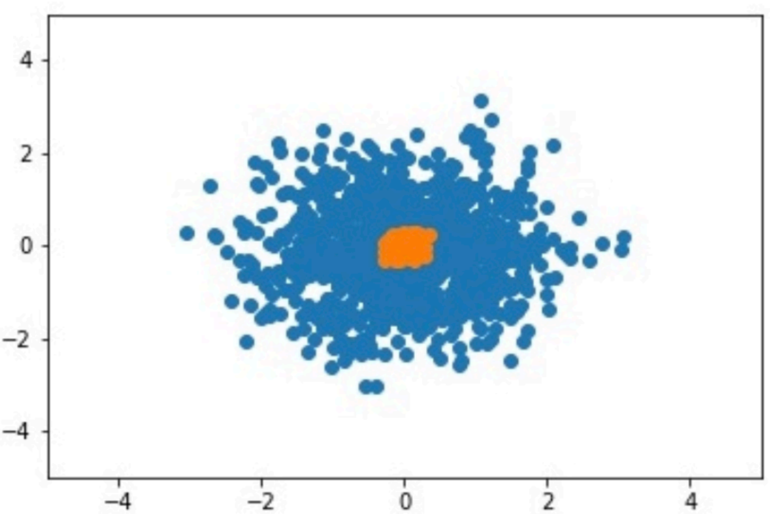


点异常

- 寻找“独特”的对撞事例
- 例如某个运动学变量“尾巴”附近的事例

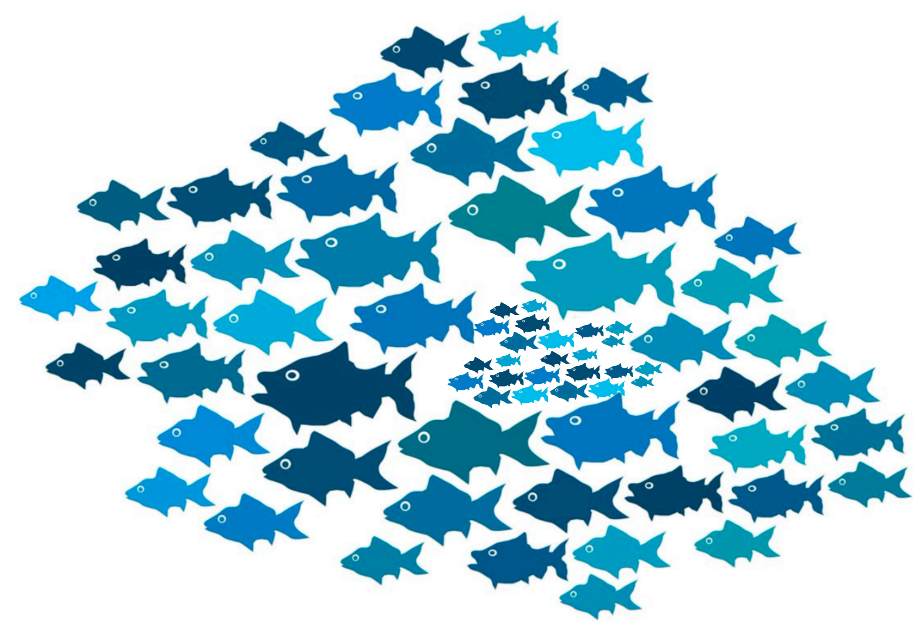


[1807.10261, 1808.08979, 1808.08992, 1811.10276, 1903.02032, 1912.10625, 2004.09360, 2006.05432, 2007.01850, 2007.15830, 2010.07940, 2102.08390, 2104.09051, 2105.07988, 2105.10427, 2105.09274, 2106.10164, 2108,03986, 2109.10919, 2110.06948, 2112.04958, 2203.01343,2206.14225, 2303.14134, 2304.03836, 2306.03637, 2308.02671, 2309.10157, 2309.13111, ...]



密度异常

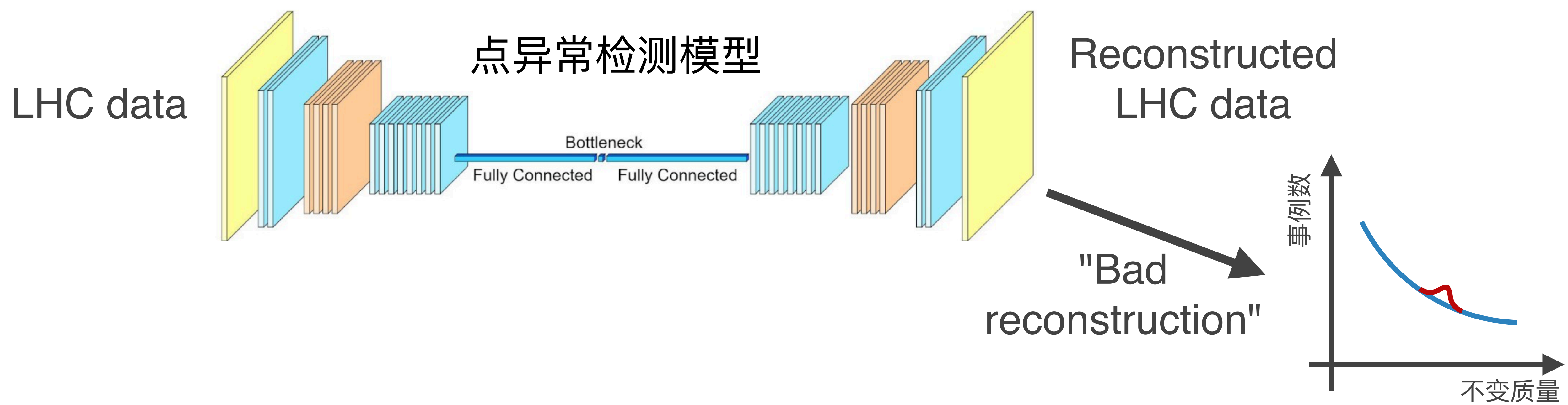
- 寻找特定区域的“聚集”效应

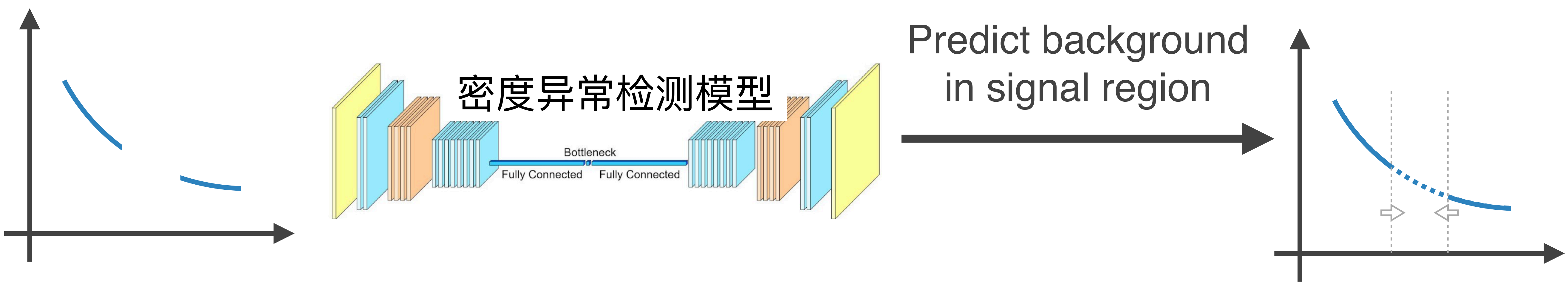


[1805.02664, 1806.02350, 1902.02634, 1912.12155, 2001.05001, 2001.04990, 2012.11638, 2106.10164, 2109.00546, 2202.00686, 2203.09470, 2208.05484, 2210.14924, 2212.11285, 2305.04646, 2305.15179, 2306.03933, 2307.11157, 2309.12918, 2310.06897, 2310.13057, ...]

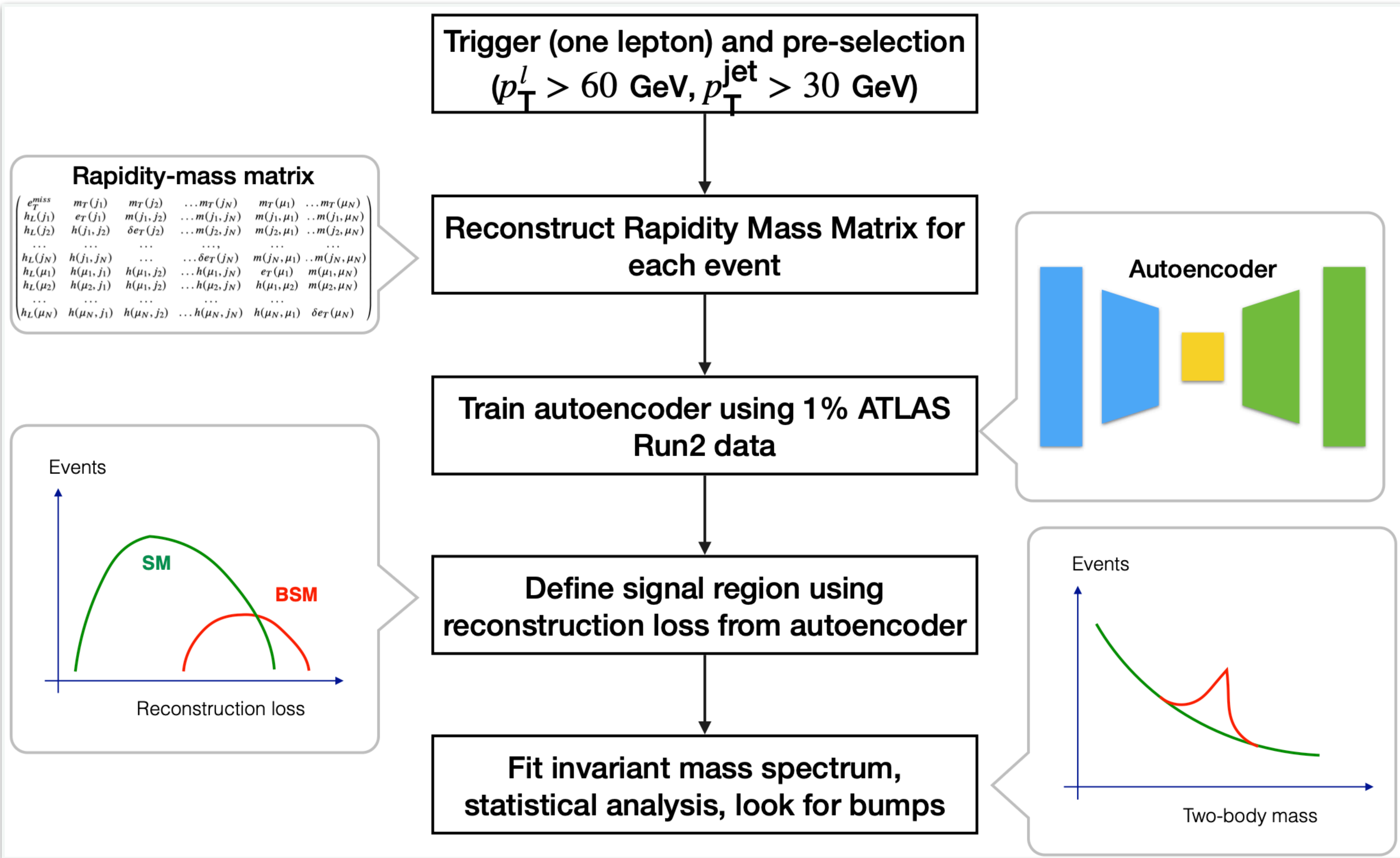
论文只统计到2023年

异常值检测应用思路


LHC data
点异常检测模型
Bottleneck
Fully Connected
Fully Connected
Reconstructed
LHC data
"Bad
reconstruction"
事例数
不变质量


密度异常检测模型
Bottleneck
Fully Connected
Fully Connected
Predict background
in signal region

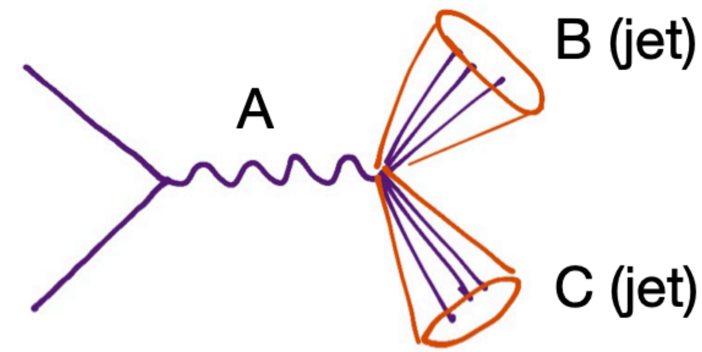
ATLAS实验上“点异常”的应用



Phys. Rev. Lett. 132 (2024) 081801

ATLAS实验上“密度异常”的应用

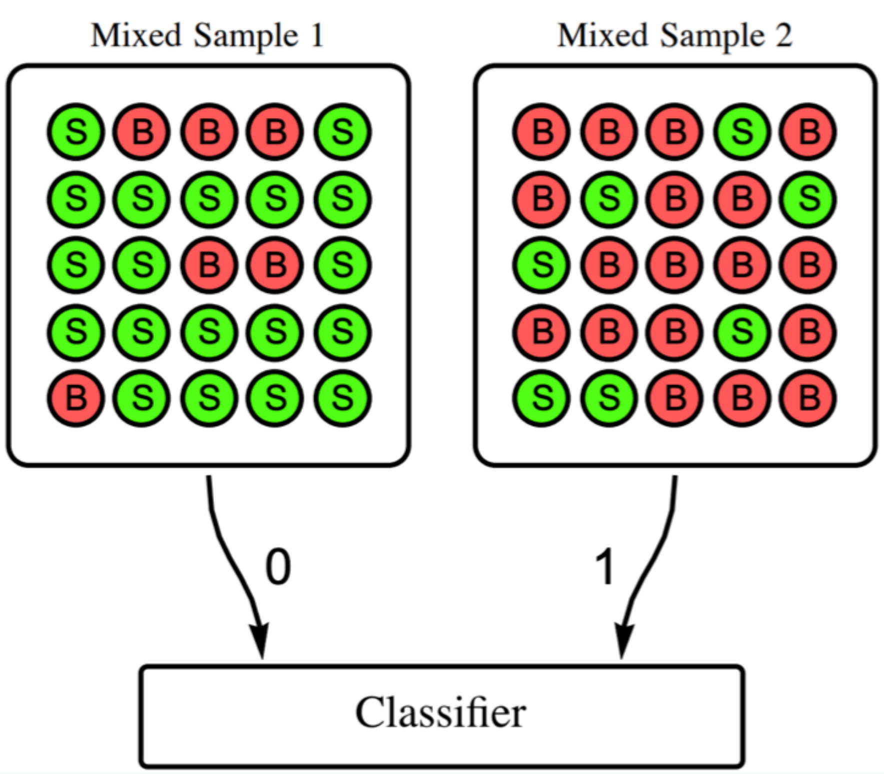
- Di-jet (large-R jets) resonance search
  - ➡ $pp \to A \to BC \to JJ$
  - ➡ Training classifiers on **data**, with **no labels**

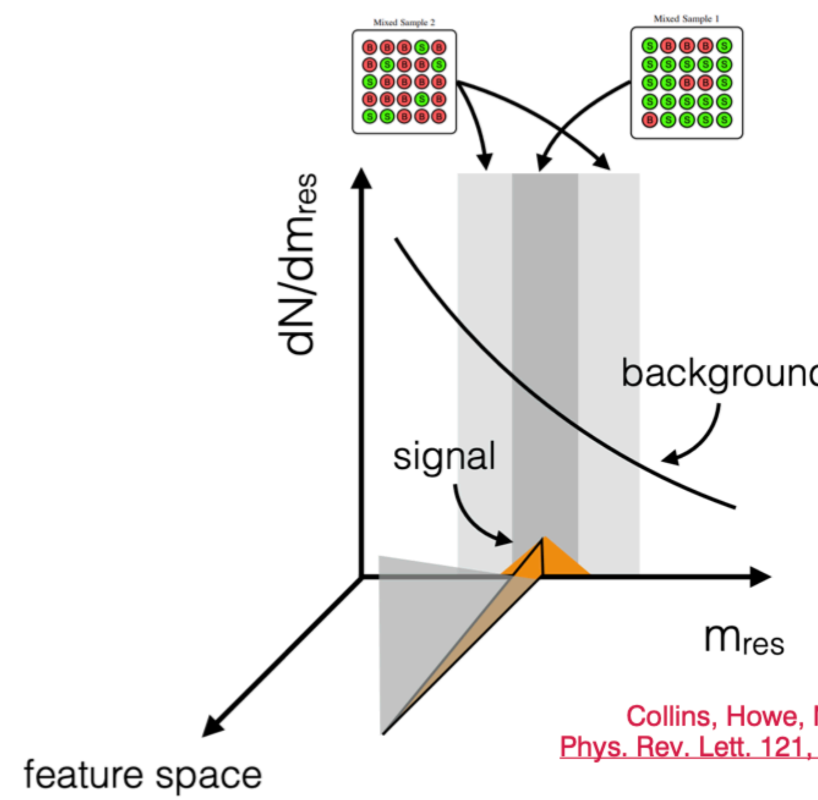


CWoLa

Classification Without Labels

Metodiev, Nachman, Thaler
JHEP 10 (2017) 174





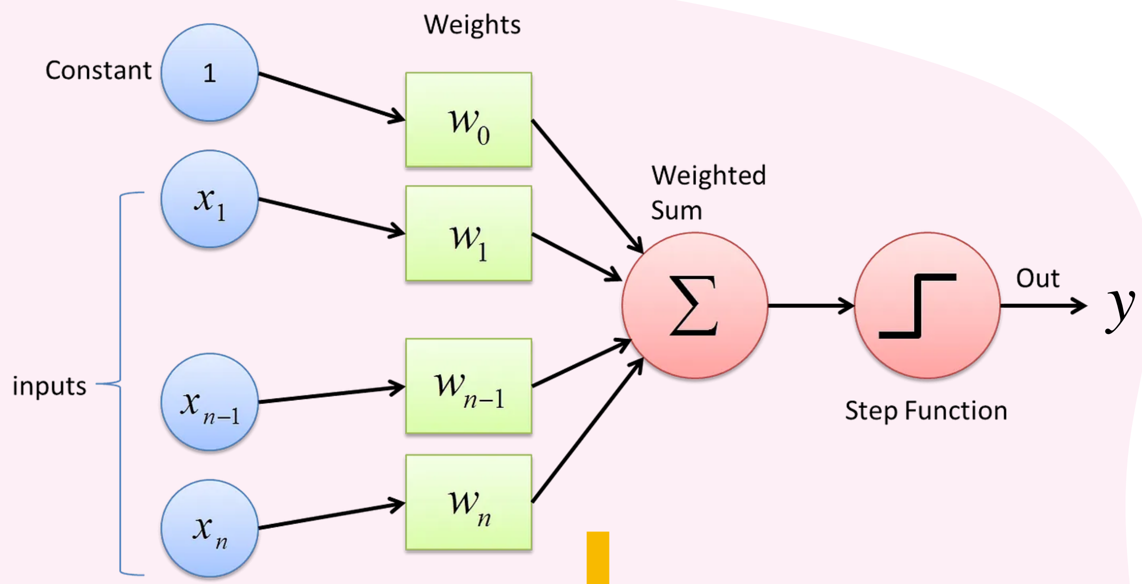
Collins, Howe, Nachman
Phys. Rev. Lett. 121, 241803 (2018)

Phys. Rev. Lett. 125 (2020) 131801

Backup

ML is not a magic

It's built upon linear algebra and information theory



$$\sigma\left(\Sigma \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}\right) = y$$

$$y = f_W(x)$$

Neural network is a function that maps input to output; "universal approximation theorem"

Learning procedure is to compress the input to output.

$$y_1 = f_1(x)$$
$$y_2 = f_2(x)$$
$$\vdots$$
$$y_n = f_n(x)$$

Which function is close to truth?

Need to quantify "similarity" between y_i and y_{truth}.

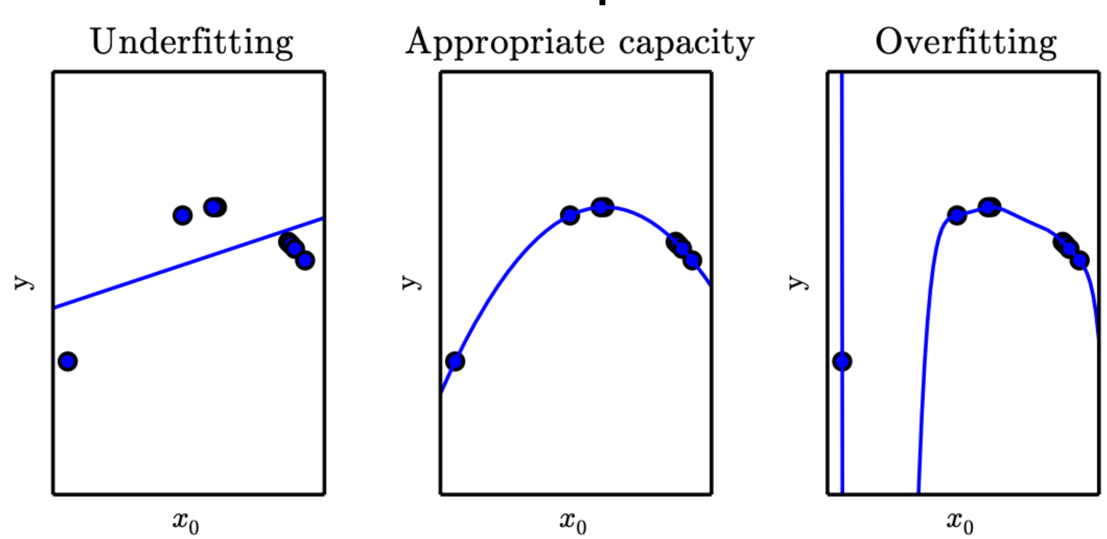
- Both are distributions (PDF)
- Also known as "loss" => min(Loss(y_i, y_{truth}))

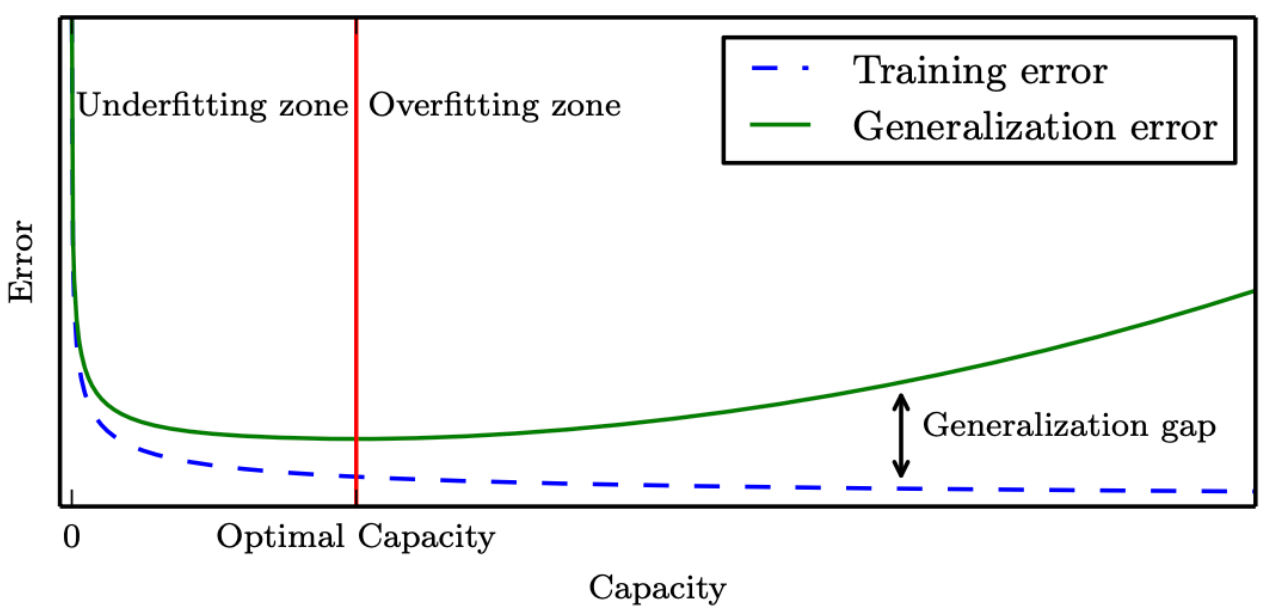
Information theory offers measures for quantifying similarity

- Entropy: disorder of 1 PDF
- Divergence: disorder between 2 PDFs

ML is not a magic: capacity and regularisation

- f_W has too few free parameters ➜ underfit $y = f_W(x)$
- …verfit
- test sample (generalisation error)





- Model's effective capacity can be affected by some factors:
  - Optimisation is very difficult and may not find the global optimum
  - Adding regularisation can limit the capacity
- An example regularisation: L2-norm

$$J(w) = \text{MeanSquaredError} + \lambda ||w||_2^2$$

$$||w||_2 = \sqrt{\sum_{k=1}^{n} |w_k|^2}$$

  - Large weight are suppressed