Генеративные модели

И применение в физике частиц

Денис Деркач

Лаборатория методов анализа больших данных

Факультет компьютерных наук

Высшая школа экономики

Летняя научная школа "Супер с-тау фабрика"

LAMBDA • HSE

Дисклеймер

Обширная тема, пересечение машинного обучения и физики частиц.

Быстроменяющаяся тема: основные результаты получены в последние 5 лет.

Более полный обзор генеративных моделей:

https://github.com/HSE-LAMBDA/DeepGenerativeModels

https://www.youtube.com/playlist?list=PLFbo11UoF5zSsRuNLxVmrNtVY7lnaLit7

https://www.youtube.com/playlist?list=PLEwK9wdS5g0pjDfggLYVUfPCZNnU7Ttdd

Generative Modeling

This X Does Not Exist!

This Person Does Not Exist

The site that started it all, with the name that says it all. Created using a style-based generative adversarial network (StyleGAN), this website had the tech community buzzing with excitement and intrigue and inspired many more sites.

Created by Phillip Wang.

This Cat Does Not Exist

These purr-fect GAN-made cats will freshen your feeline-gs and make you wish you could reach through your screen and cuddle them. Once in a while the cats have visual deformities due to imperfections in the model – beware, they can cause nightmares.

Created by Ryan Hoover.

This Rental Does Not Exist

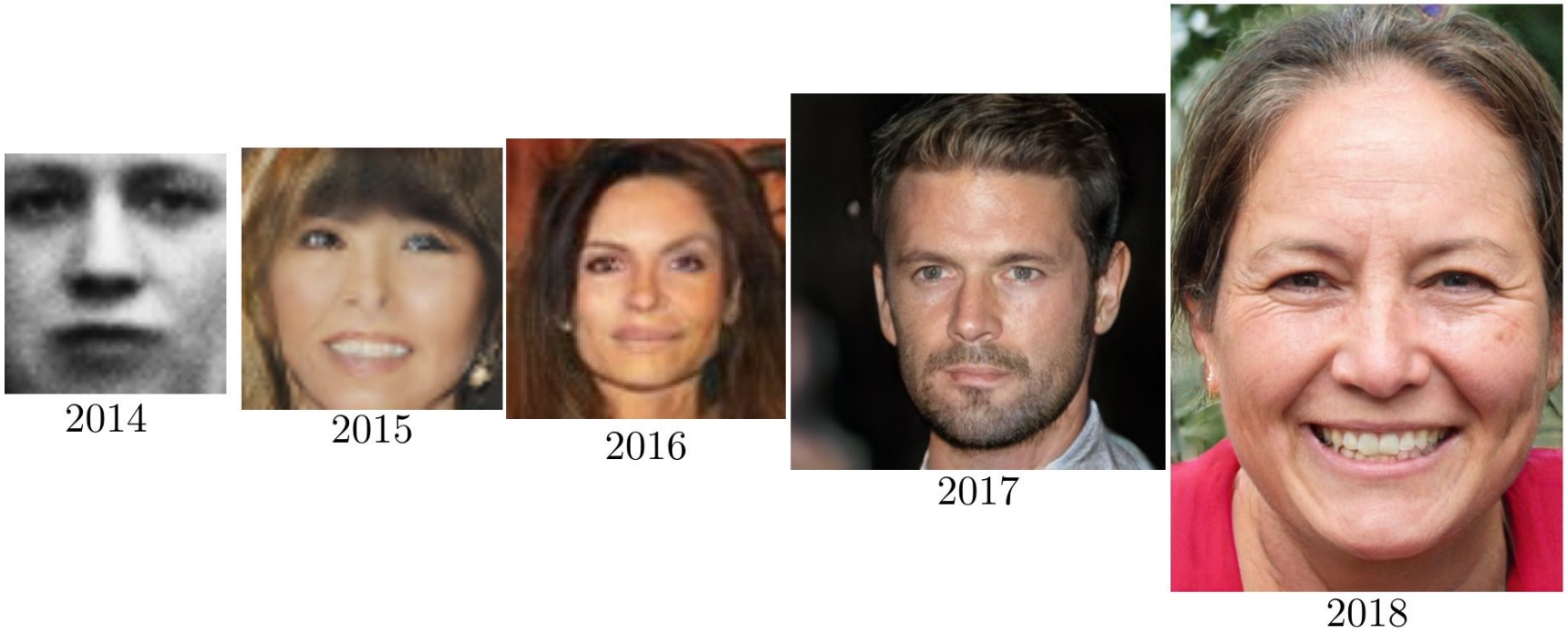
Why bother trying to look for the perfect home when you can create one instead? Just find a listing you like, buy some land, build it, and then enjoy the rest of your life.

Created by Christopher Schmidt.

https://thisxdoesnotexist.com/

Generative Models Progress

The news are well motivated.



- Enormous progress in recent years.
- Technology is ready for new tasks.

https://twitter.com/goodfellow_ian/status/1084973596236144640

More Tricks for Your Brain

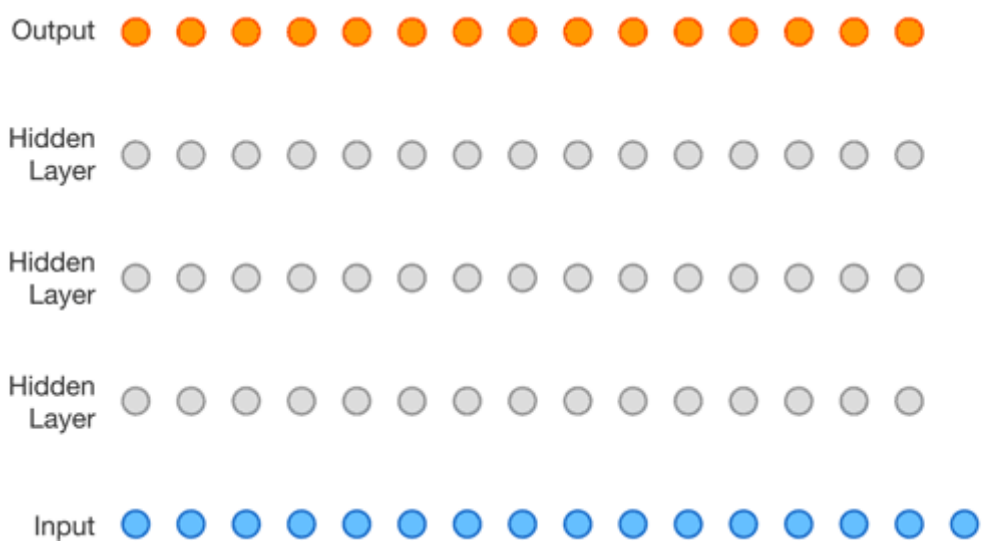
- Text generation.

Two men happily working on a plastic computer.
The toilet in the bathroom is filled with a bunch of ice.
A bottle of wine near stacks of dishes and food.
A large airplane is taking off from a runway.
Little girl wearing blue clothing carrying purple bag sit
SeqGAN (Baseline)
A baked mother cake sits on a street with a rear of it.
A tennis player who is in the ocean.
A highly many fried scissors sits next to the older.
A person that is sitting next to a desk.
Child jumped next to each other.
RankGAN (Ours)
Three people standing in front of some kind of boats.
A bedroom has silver photograph desk.
The bears standing in front of a palm state park.
This bathroom has brown bench.
Three bus in a road in front of a ramp.

More Tricks for Your Brain

- Text generation.

- Voice from text generation.



More Tricks for Your Brain

- Text generation.
- Voice from text generation.
- Style transfer.

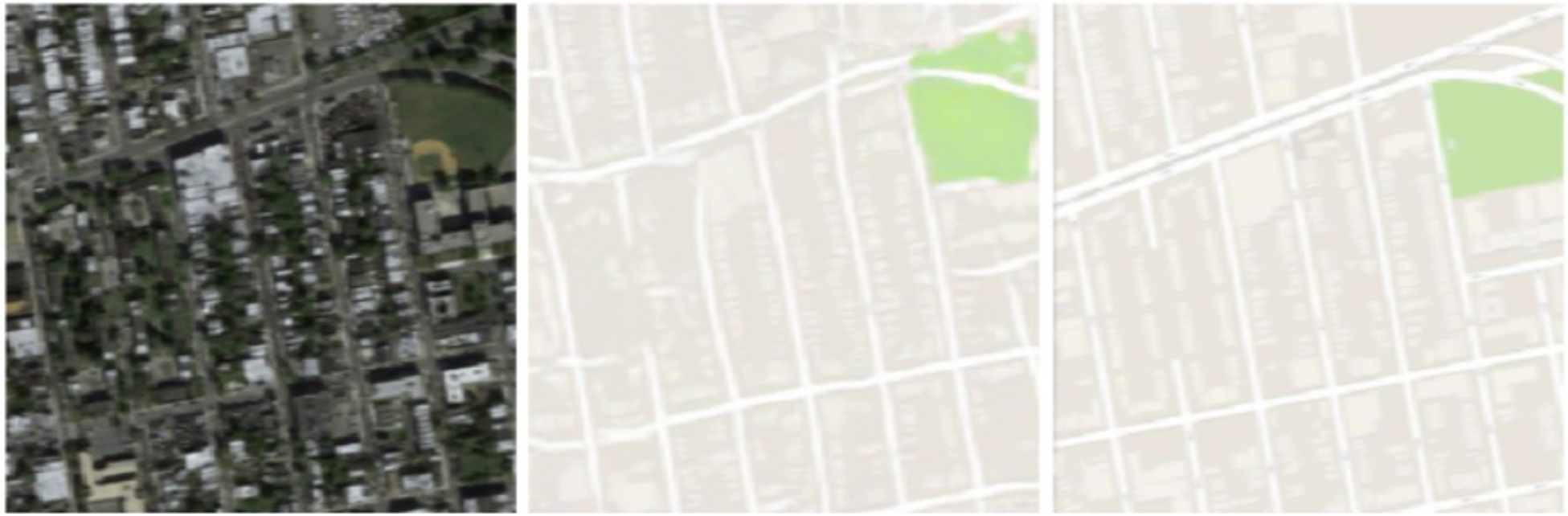
Generative Models Failures

- Image is created as **interpolation** between existing ones.

https://www.fastcompany.com/90303908/this-ai-dreams-about-cats-and-theyll-haunt-your-nightmares

Dealing with Maps: generating map

- Image-to-image style transfer.
- Creates map on-the-fly from satellite image.



Input **Generated** **True**

https://github.com/ChengBinJin/pix2pix-tensorflow

Dealing with Maps: generating satellite image

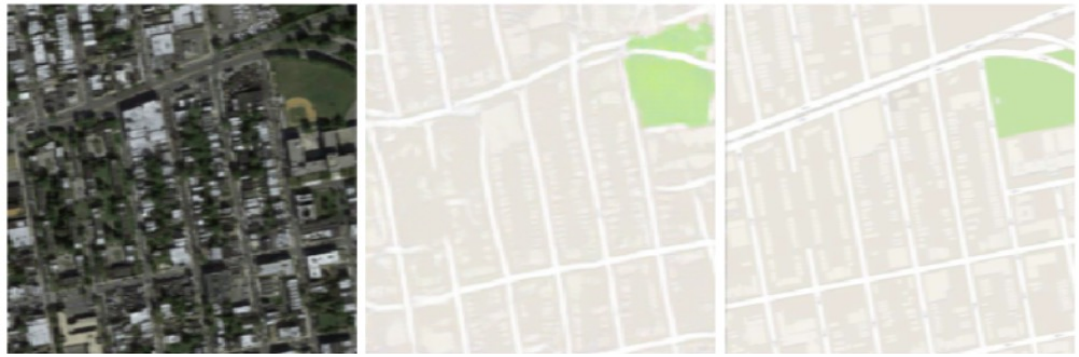
- Image-to-image style transfer
- Creates map on-the-fly from satellite image and vice versa.

Input **Generated** **True**

https://github.com/ChengBinJin/pix2pix-tensorflow

Dealing with Maps: generating satellite image

- Image-to-image style transfer
- Creates map on-the-fly from satellite image and vice versa.
- The technology is the same as for “Monet” painting. Just need good representation.



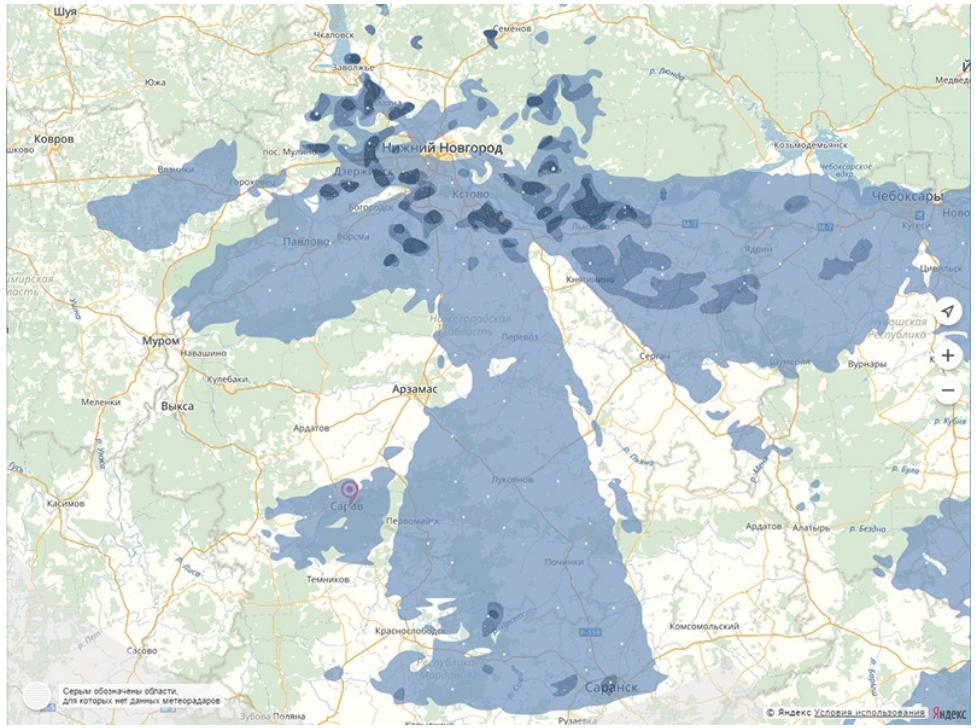
=

Dealing with Satellite Images: Super-resolution

- We can “create” a more appropriate map quality.
- This later can be used in segmentation task.

https://omdena.com/blog/super-resolution/

Weather prediction: nowcast



- Video prediction for precipitation.
- Generation of future state, based on the previous one.

https://www.kdd.org/kdd2019/accepted-papers/view/precipitation-nowcasting-with-satellite-imagery

Dirty Road Signs Generation

- Road signs from the book are too clean.
- Need to put mud and shadows on the signs.

https://arxiv.org/abs/1907.12902
https://www.hse.ru/sci/diss/426009543

What Generative Models **Do not** Produce

- No new information is created.
- All interpolations are done in some representation space.

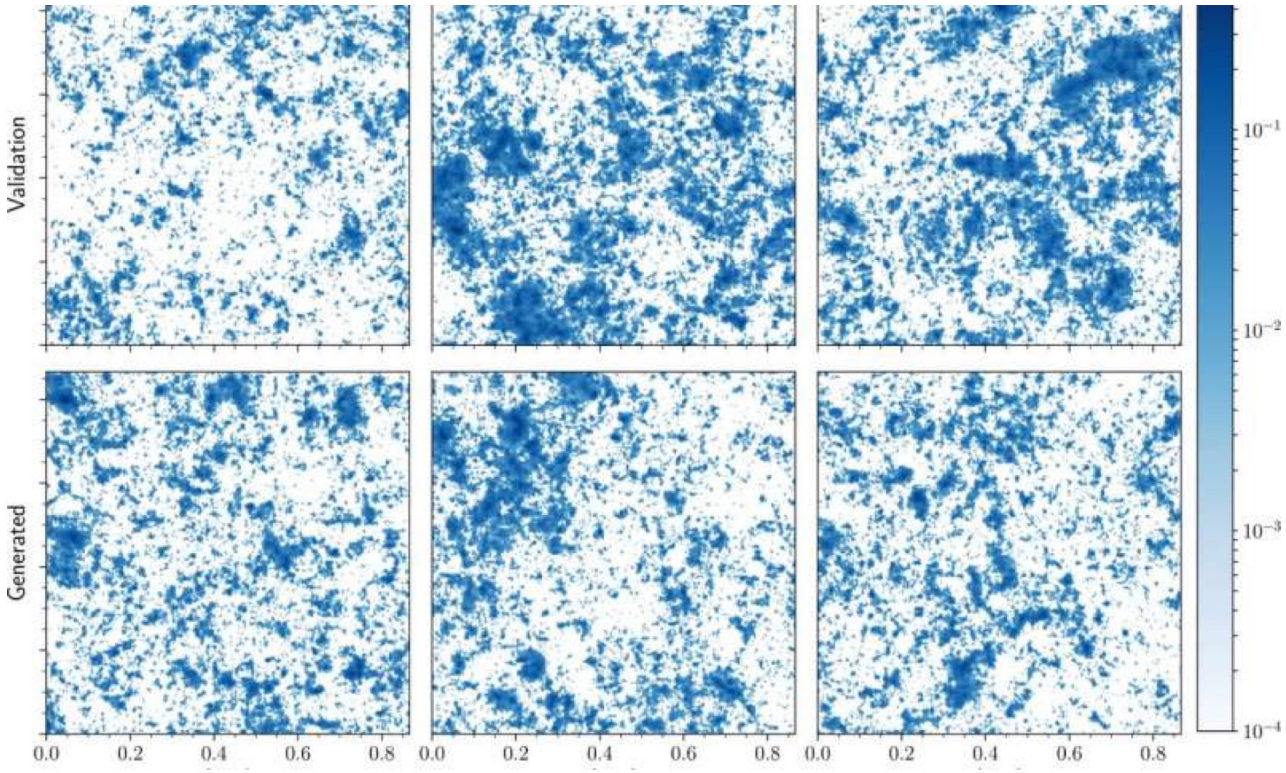
Chapter outcome

- Generative models in machine learning were developing quickly in the last years.
- Current state-of-the-art allows to implement generative models in more serious tasks than deceiving non-expert human.

Generative Models for Science

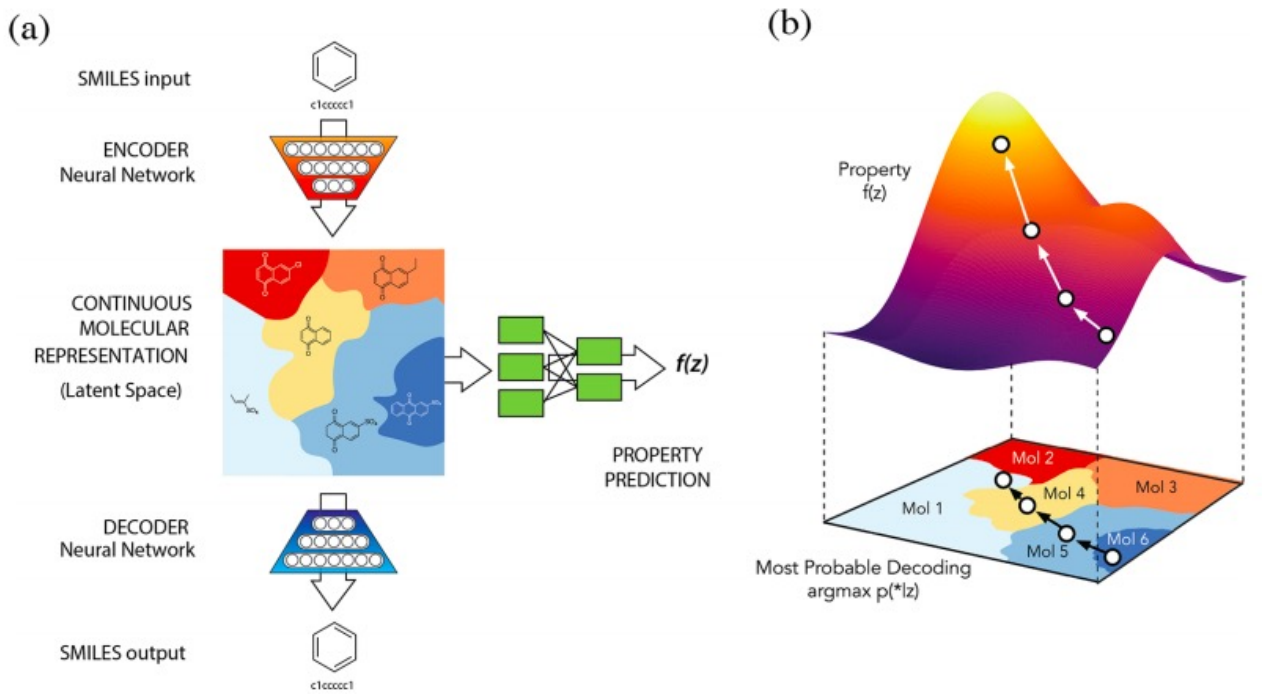
Astronomy Example

- Generate weak lensing convergence maps.
- "Visually, an **expert** cannot distinguish the generated maps from the full simulation ones"



Mustafa, M., et al.. Comput. Astrophys. 6, 1 (2019).

Medicine Example



- generates a new candidate molecule with best property;
- to be tested by **engineers and have some further medical tests**.

High-Energy Physics

- Event can be considered as a photo.
- The event is than passed through the pipeline.

Chapter outcome

- Many scientific applications.
- High-energy physics has got specific requirements to the properties (no brain tricks assumed).

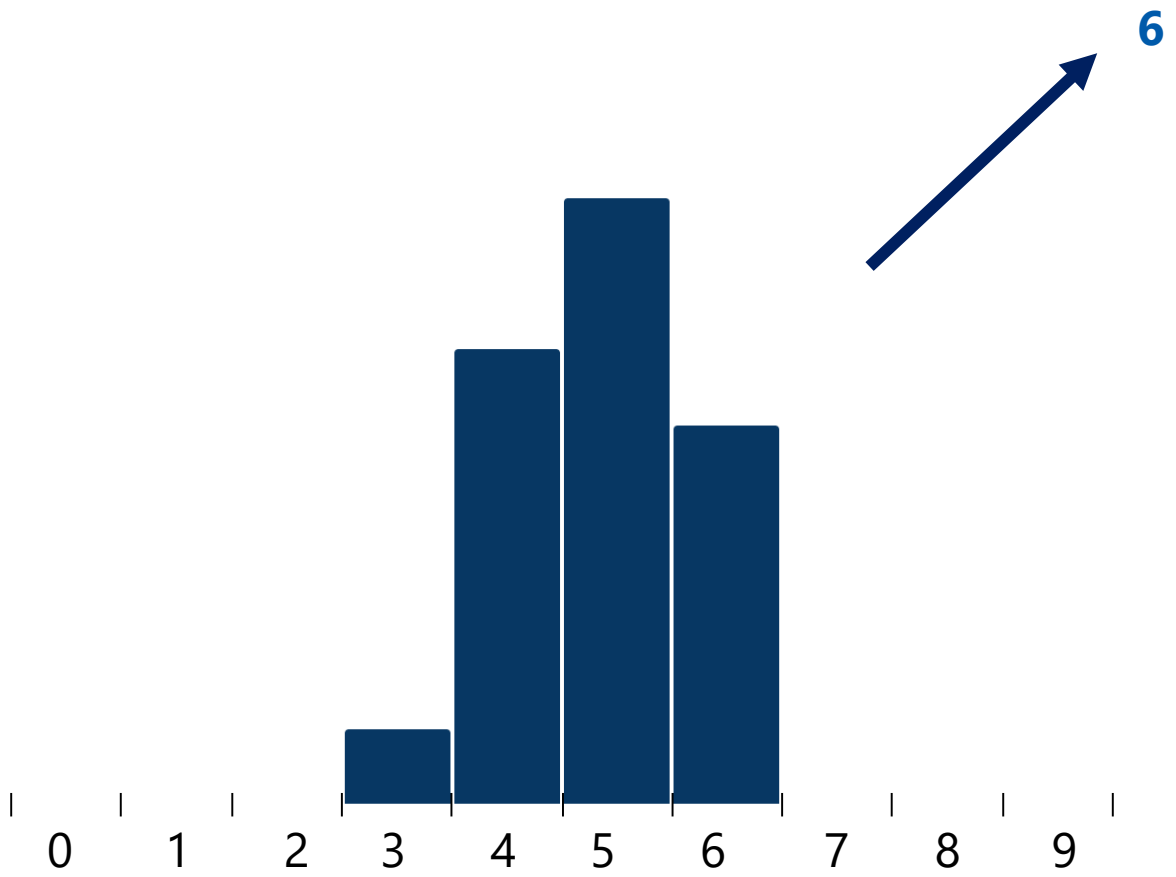
What is Generative Modeling

Random Number Generation

- We have sample with numbers:

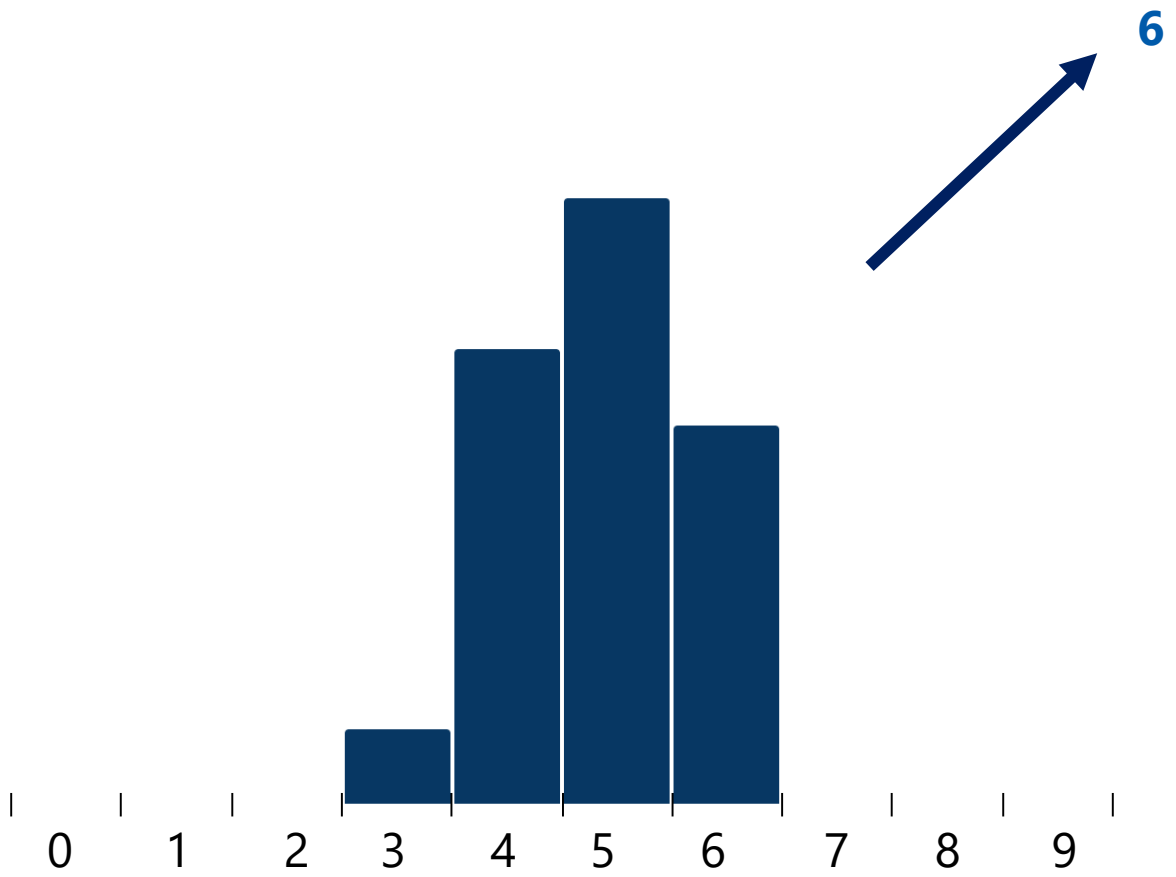
  3; 5; 4; 4; 4; 4; 5 ; 6 ; 5 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 6

- Want to create a new number alike.



How we did it?

- Assume there is a probability density $p_{true}(x)$.
- Try to estimate $p_{true}(x)$ using data and obtain $p_{data}(x)$.
- Sample from $p_{data}(x)$.



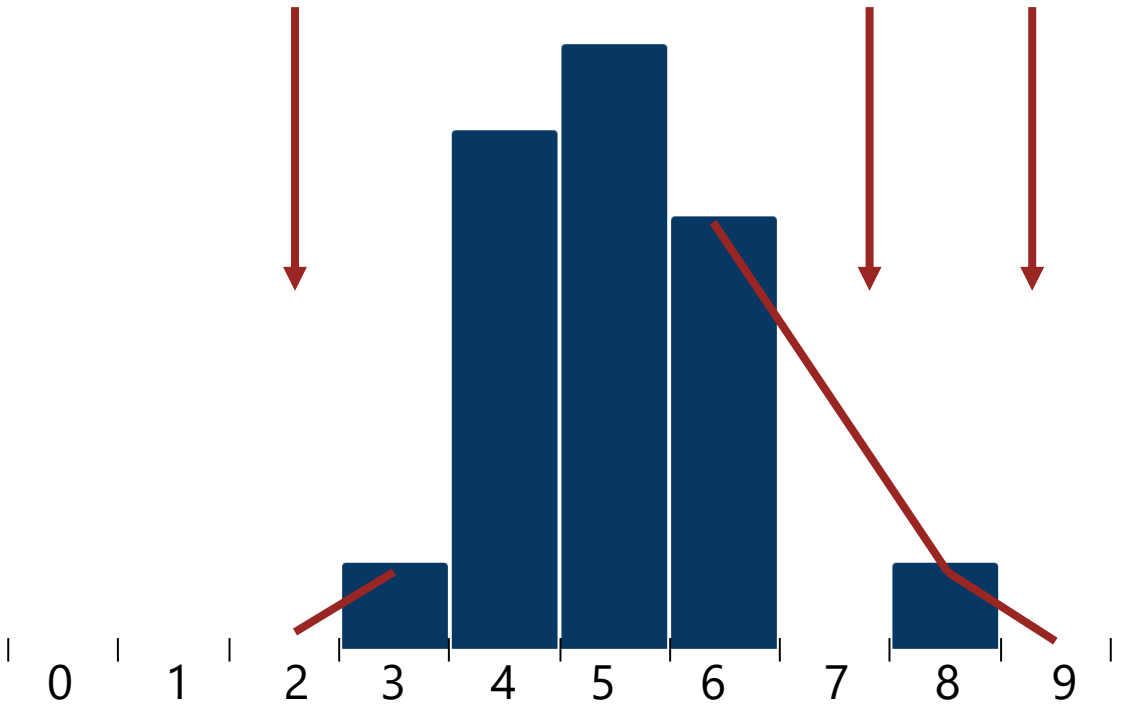
Random Number Generation

- We have **different** sample with numbers:

  3; 5; 4; 4; 4; 4; 5 ; 6 ; 8 ; 4 ; 5;

  4; 5; 6; 5; 6; 5; 5; 6; 5

- Want to create a new number alike.

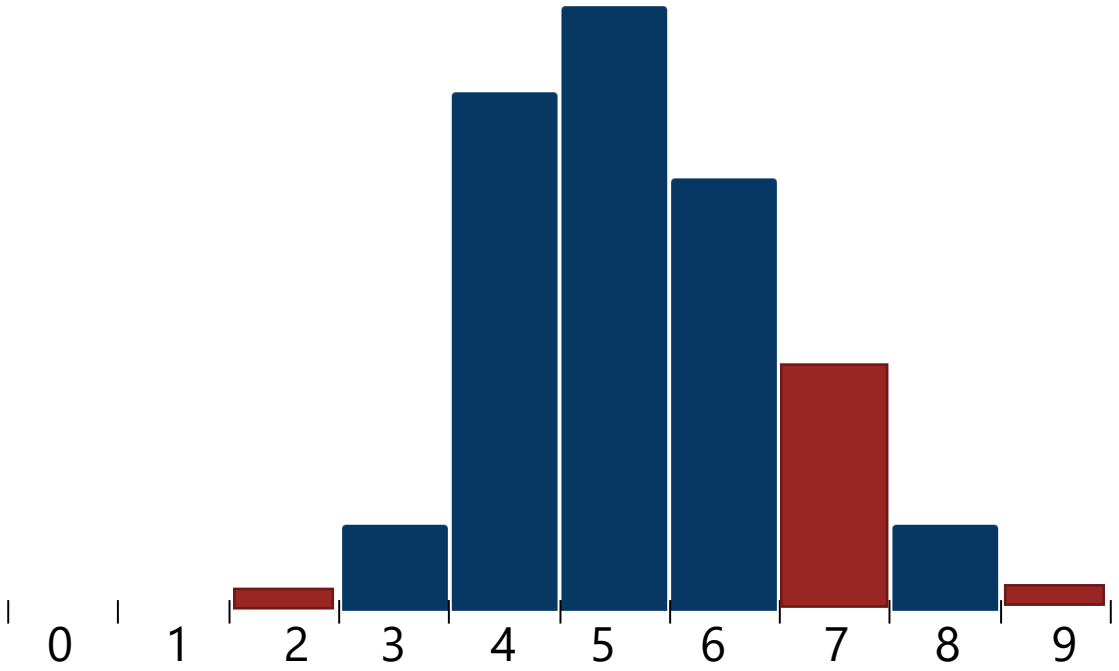


Random Number Generation

- We have **different** sample with numbers:

  3; 5; 4; 4; 4; 4; 5 ; 6 ; 8 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 5

- Want to create a new number alike.

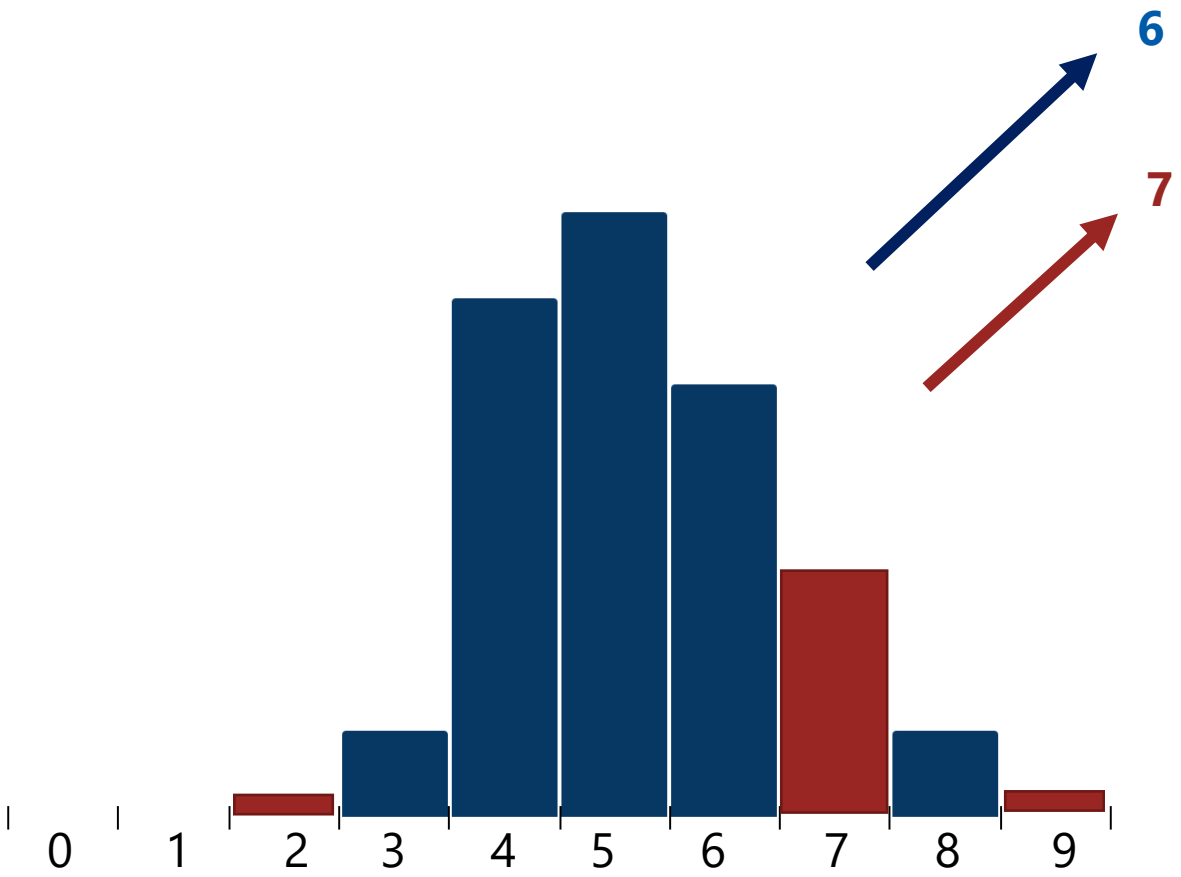


Random Number Generation

- We have **different** sample with numbers:

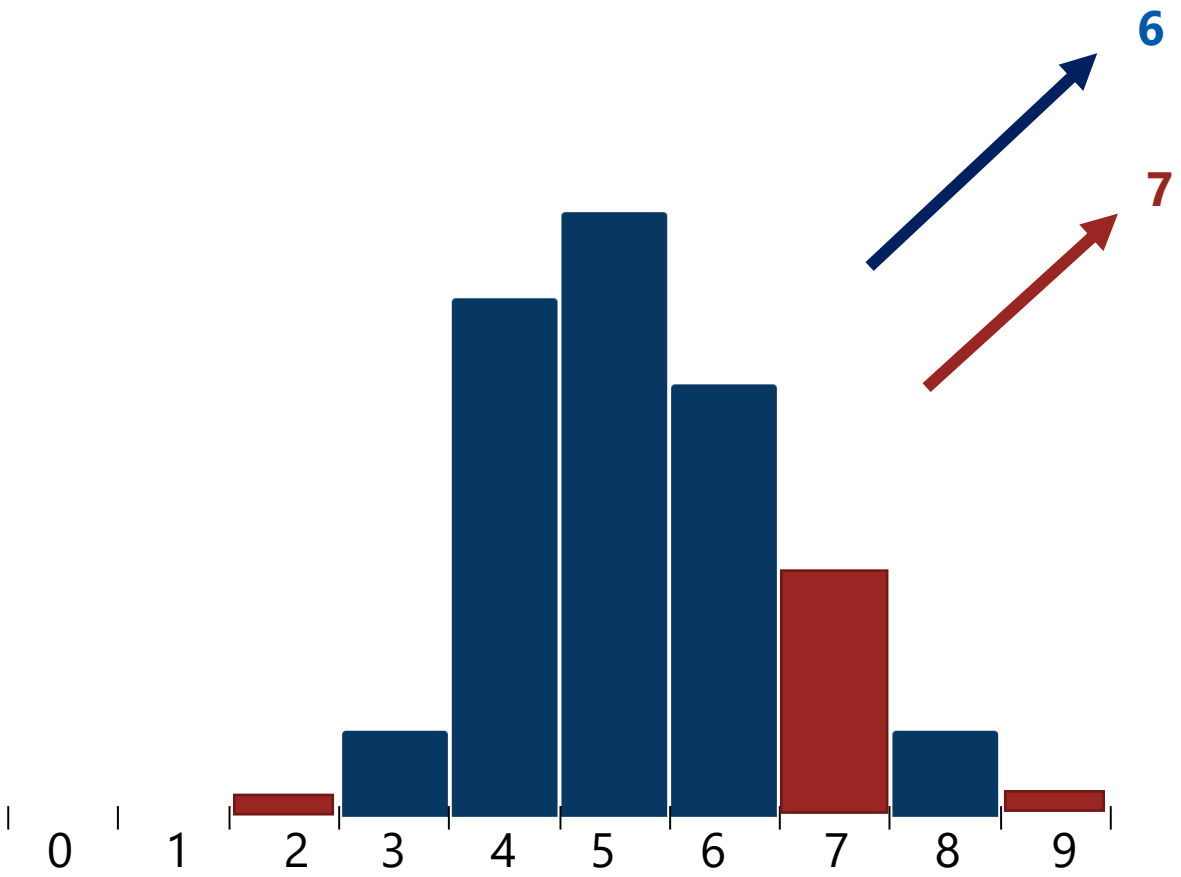
  3; 5; 4; 4; 4; 4; 5 ; 6 ; 8 ; 4 ; 5; 4; 5; 6; 5; 6; 5; 5; 6; 5

- Want to create a new number alike.



Random Number Generation

- Assume there is a probability density $p_{true}(x)$.
- **Choose interpolation model.**
- Try to estimate $p_{true}(x)$ using data and obtain $p_{data}(x)$.
- Sample from $p_{data}(x)$.



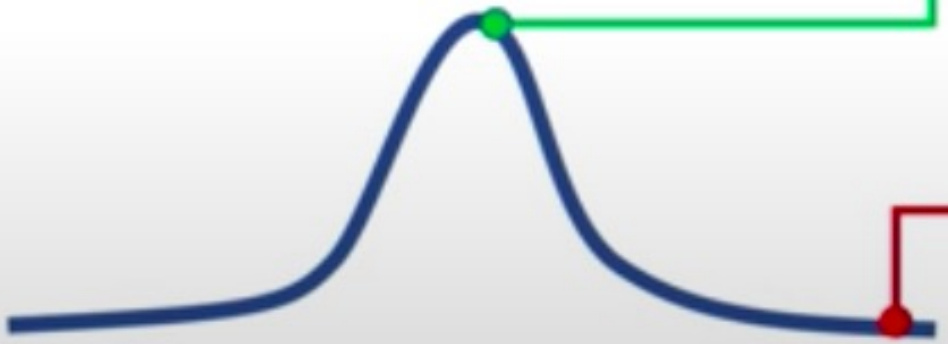
Case Study: Anomaly Detection

- **Problem:** How can we detect when we encounter something new or rare?
- **Strategy:** Leverage generative models, detect outliers in the distribution
- Use outliers during training to improve even more!

95% of Driving Data:

(1) sunny, (2) highway, (3) straight road

Detect outliers to avoid unpredictable behavior when training



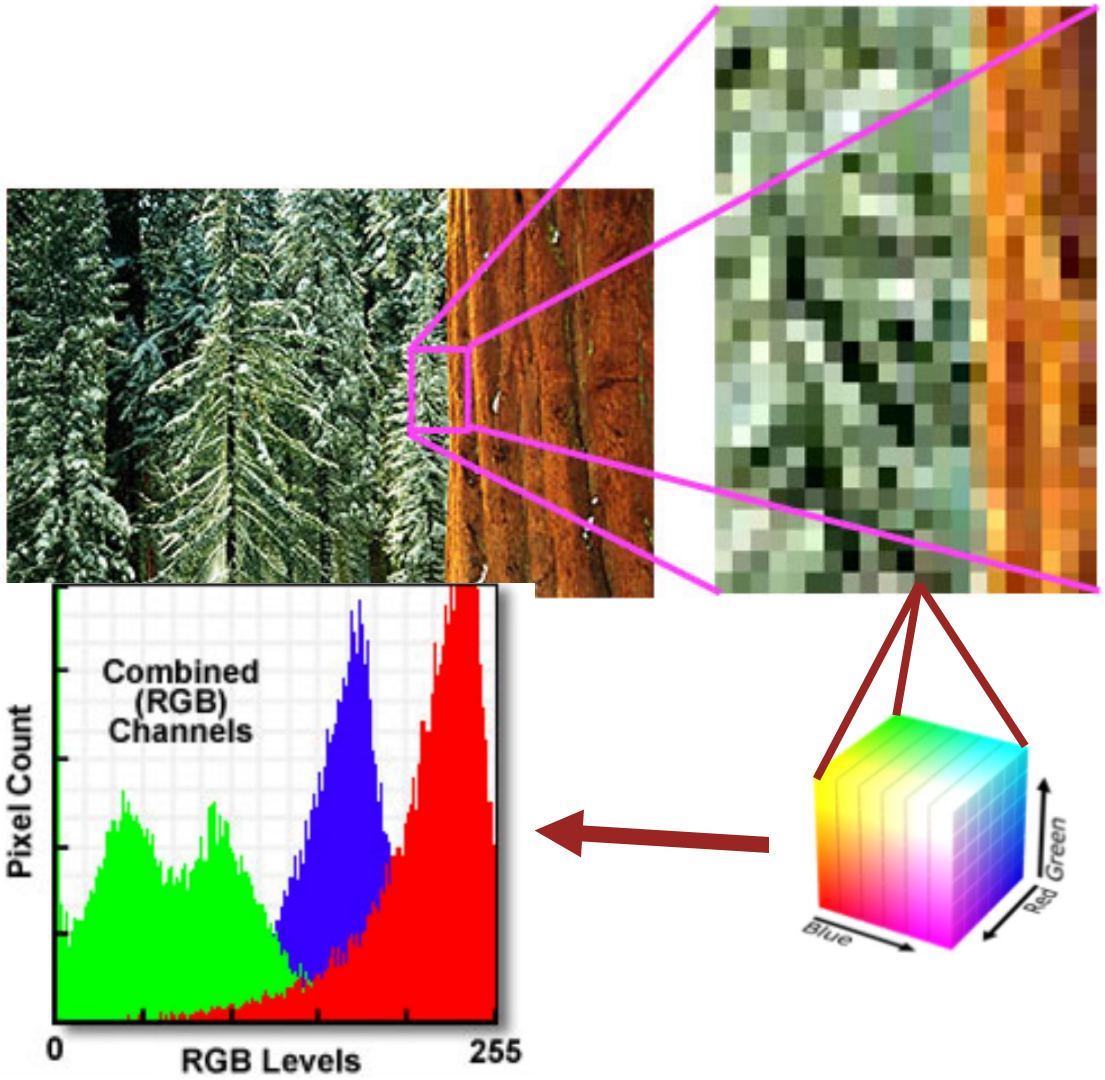
Edge Cases

Harsh Weather

Pedestrians

http://introtodeeplearning.com/

More Complicated Case: Figures



- Figure consists of pixels.
- One can use this representation.
- Each pixel is encoded by 3 colours.
- **Multi-modal distribution**.
- **Multidimensional problem**.

Number of Parameters

- Handwritten digits dataset.
- Only black and white pixels.
- Number of pixels 28X28.
- Number of possible states:

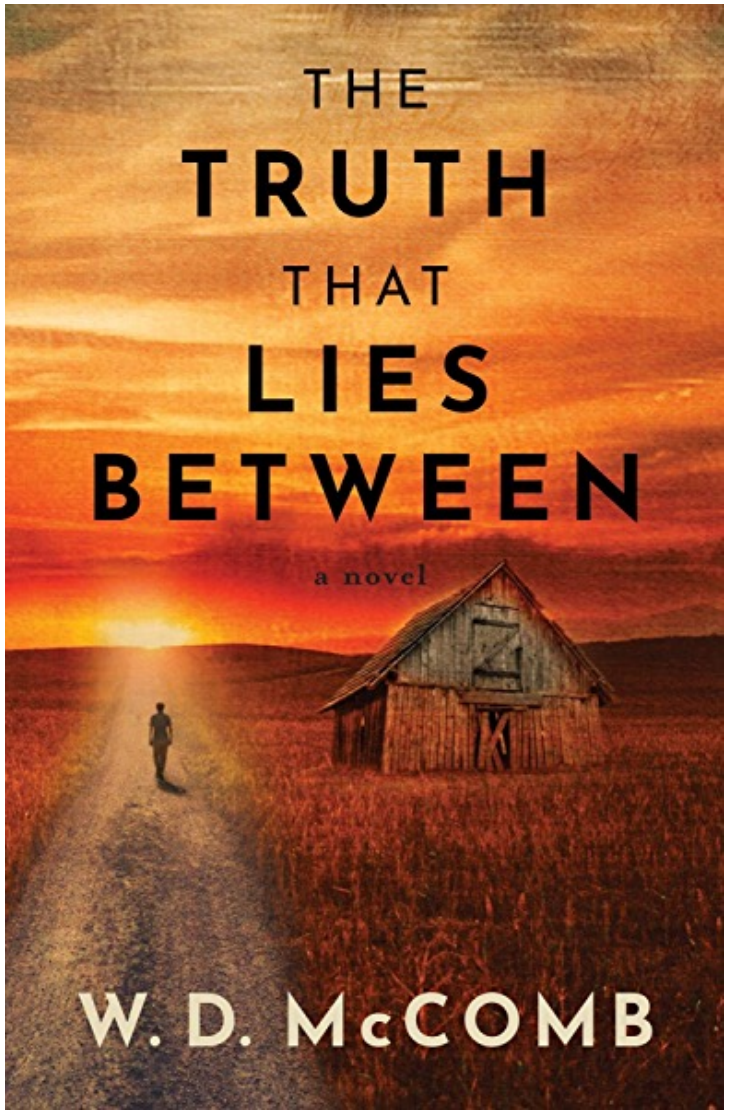
$$2x2x2x....x2 = 2^n.$$

- **Number of parameters:**

$$\mathbf{2^n - 1.}$$

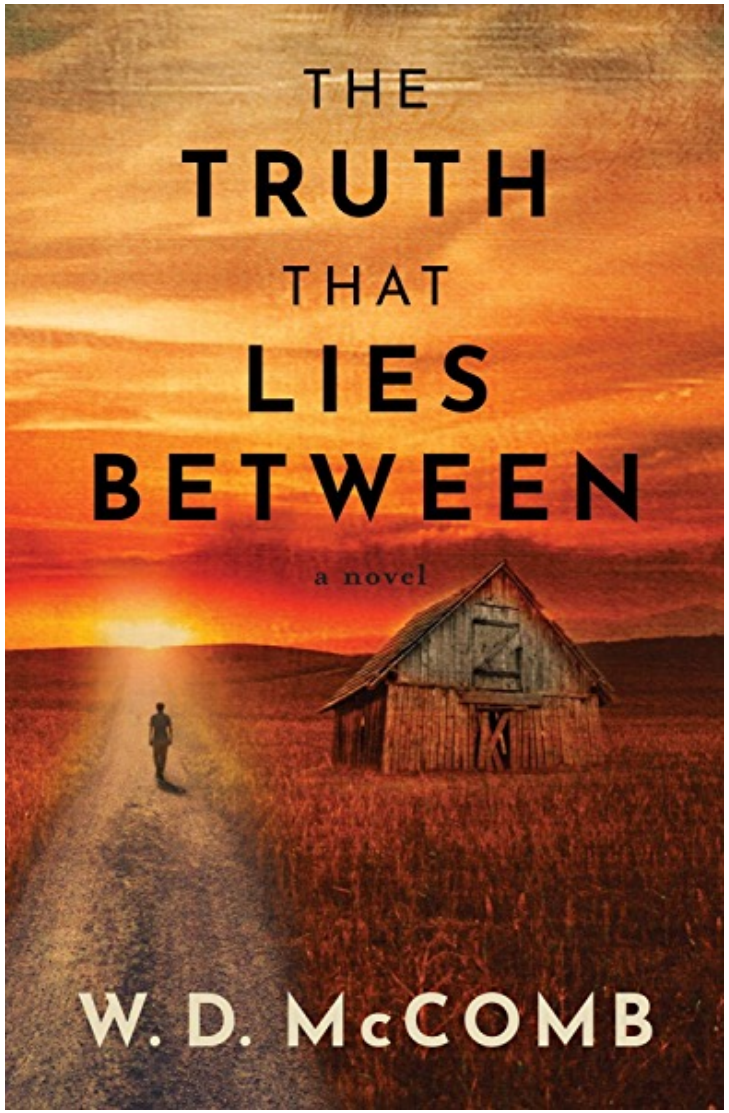
- **For Independent pixels:**

$$\mathbf{n.}$$



Generative model: Final Touch

- Assume there is a probability density $p_{true}(x)$.
- Choose interpolation model.
- **Reduce number of dimensions.**
- Try to estimate $p_{true}(x)$ using data and obtain $p_{data}(x)$.
- Sample from $p_{data}(x)$.



Amazon

Generative model: Problem Statement

Three major tasks, given a generative model f from a class of models $\mathcal{F}$:

- **Estimation**: find the f in $\mathcal{F}$ that best matches observed data.
- **Evaluate Likelihood**: compute $f(z)$ for a given z.
- **Sampling**: drawing from f.

S. Nowozin et al. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization

Generative model vs Discriminative model

Discriminative models

- learn $\mathbb{P}(y|x)$
- Directly characterizes the decision boundary between classes only
- Examples: Logistic Regression, SVM, etc

Generative models

- learn $\mathbb{P}(x|y)$ (and eventually $\mathbb{P}(y, x)$)
- Characterize how data is generated (distribution of individual class)
- Examples: Naive Bayes, HMM, etc.

https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf

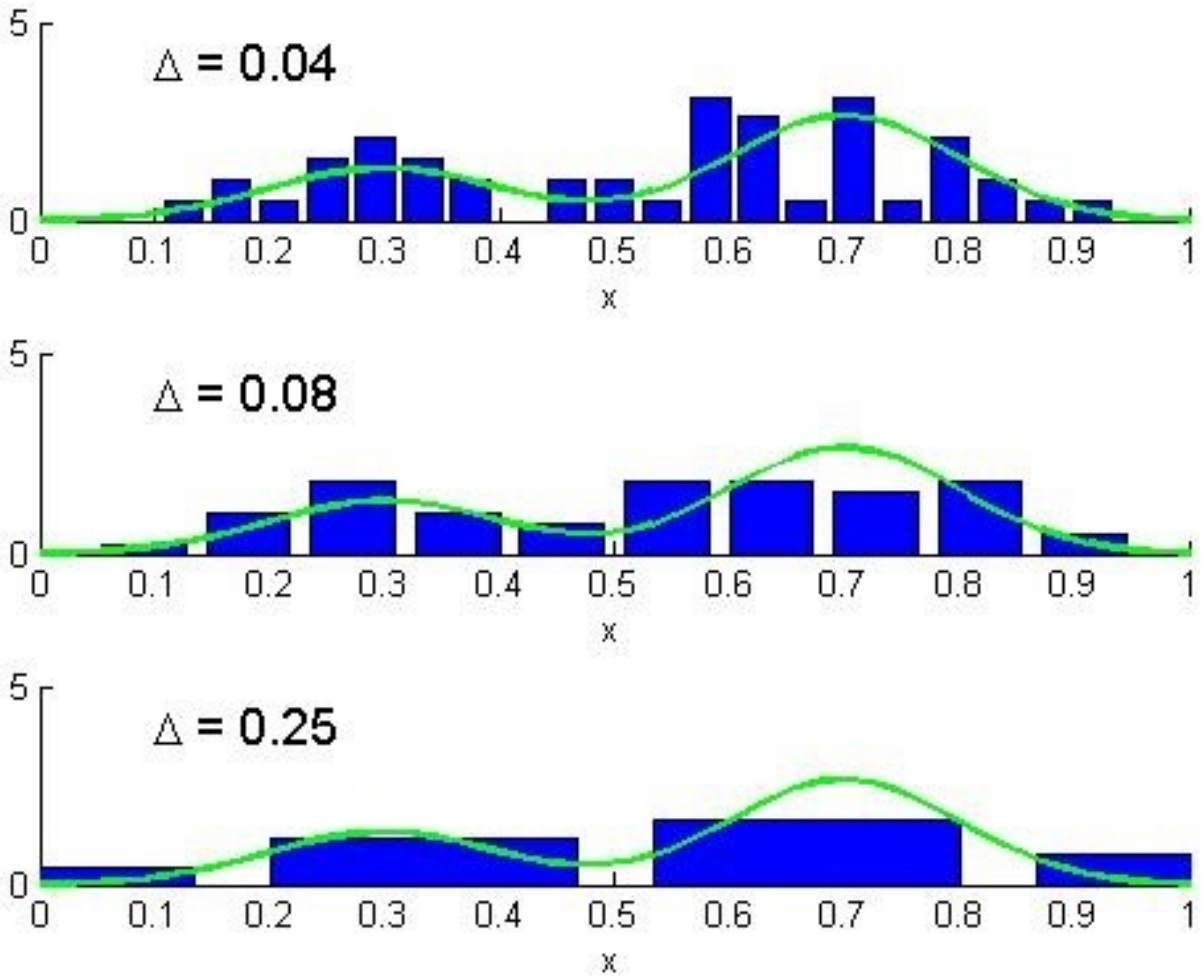
Chapter outcome

- Generative modeling is a distinct task in machine learning.
- Mathematically, it aims to reconstruct the probability density, from which the given dataset was sampled.

“Early” Generative Models

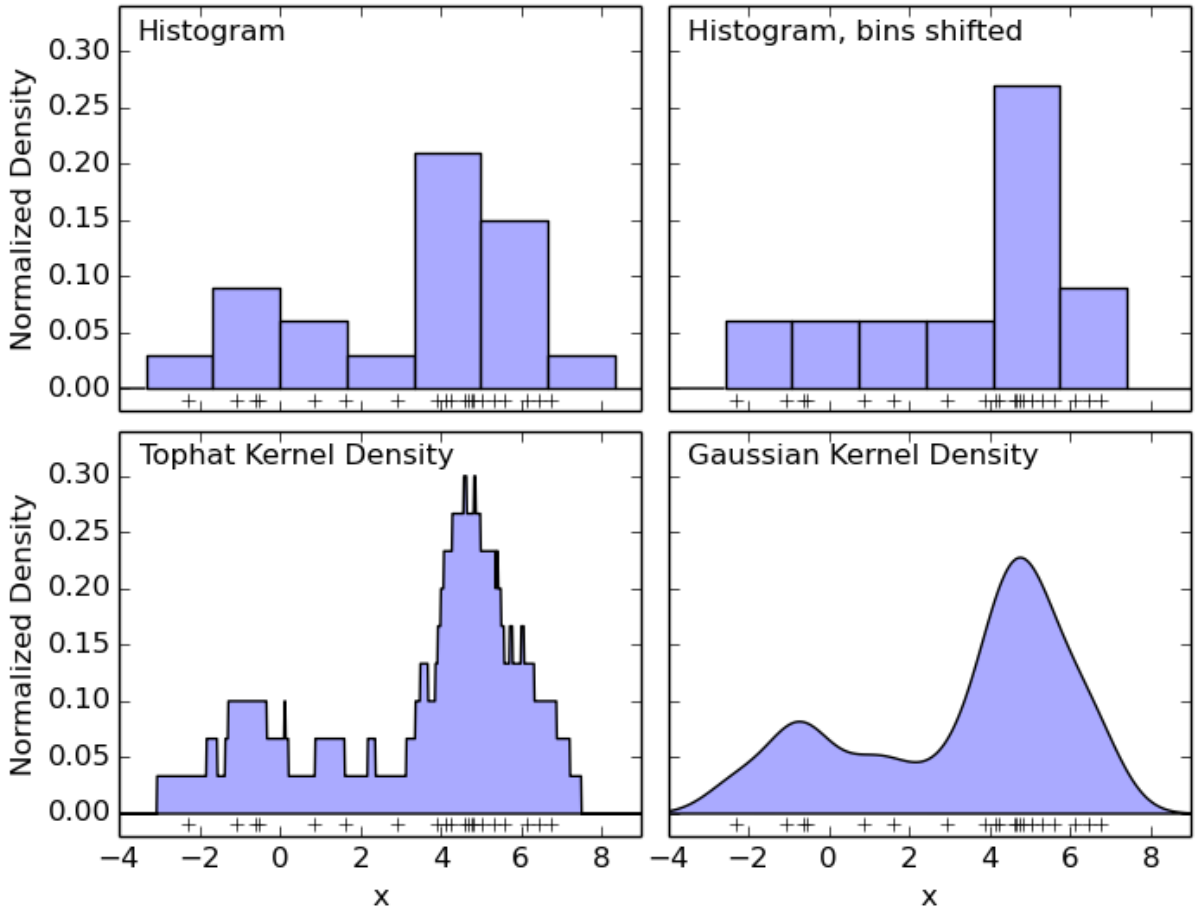
“Non-parametric” Approaches

- Histograms can be used.
- Need to choose optimal bin size.
- Smaller bins for approximate constant estimate.
- Larger bins for less fluctuations.
- Can be chosen using empirical risk.



SunLei's Linear Space

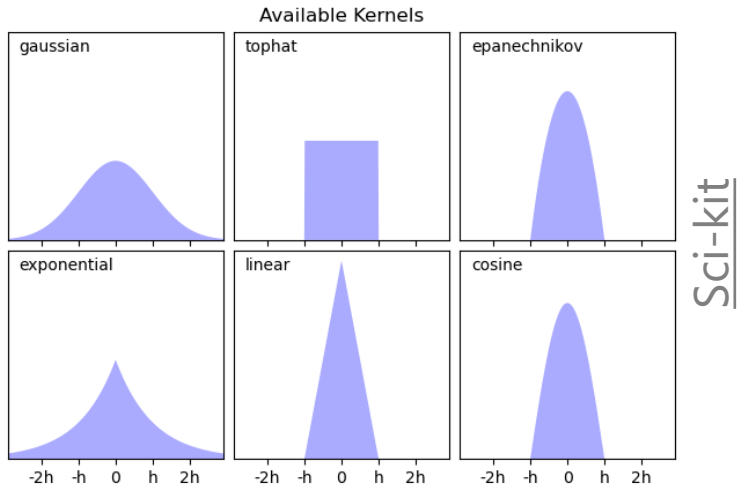
Kernel-density estimation



- Assign **every** event a weight.
- Smooth between events.
- Kernel Density Estimation:

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

K – some kernel, h – bandwidth.



Sci-kit

KDE Summary

- Efficient in low dimensional estimation.
- Controllable convergence rate for bias or variance but the overall rate is similar.
- To speed up the convergence, once can attempt to find manifolds in the d-dimension.
- Fairly hard to sample and keep the model in memory.

Type	Method	Convergence rate	Tuning parameter	Limitation
Parametric	Parametric model	$O\left(\frac{1}{\sqrt{n}}\right)$	None	Unavoidable bias
	Mixture model	$O\left(\frac{1}{\sqrt{n}}\right)$	K, number of mixture	Hard to compute
Nonparametric	Histogram	$O\left(\frac{1}{n^{1/3}}\right)$	b, bin size	Lower convergence rate
	Kernel density estimator	$O\left(\frac{1}{n^{2/5}}\right)$	h, smoothing bandwidth	
	K-nearest neighbor	$O\left(\frac{1}{n^{2/5}}\right)$	k, number of neighbor	
	Basis approach	$O\left(\frac{1}{n^{2/5}}\right)$	M, number of basis	

see for example Yen Chi Chen, Learning Theory, Lec 8.

Chapter Summary

- Generative modeling is a distinct task of machine learning.
- Several pre-deep learning algorithms can produce reasonable results in the low dimensional data.

HEP Simulation

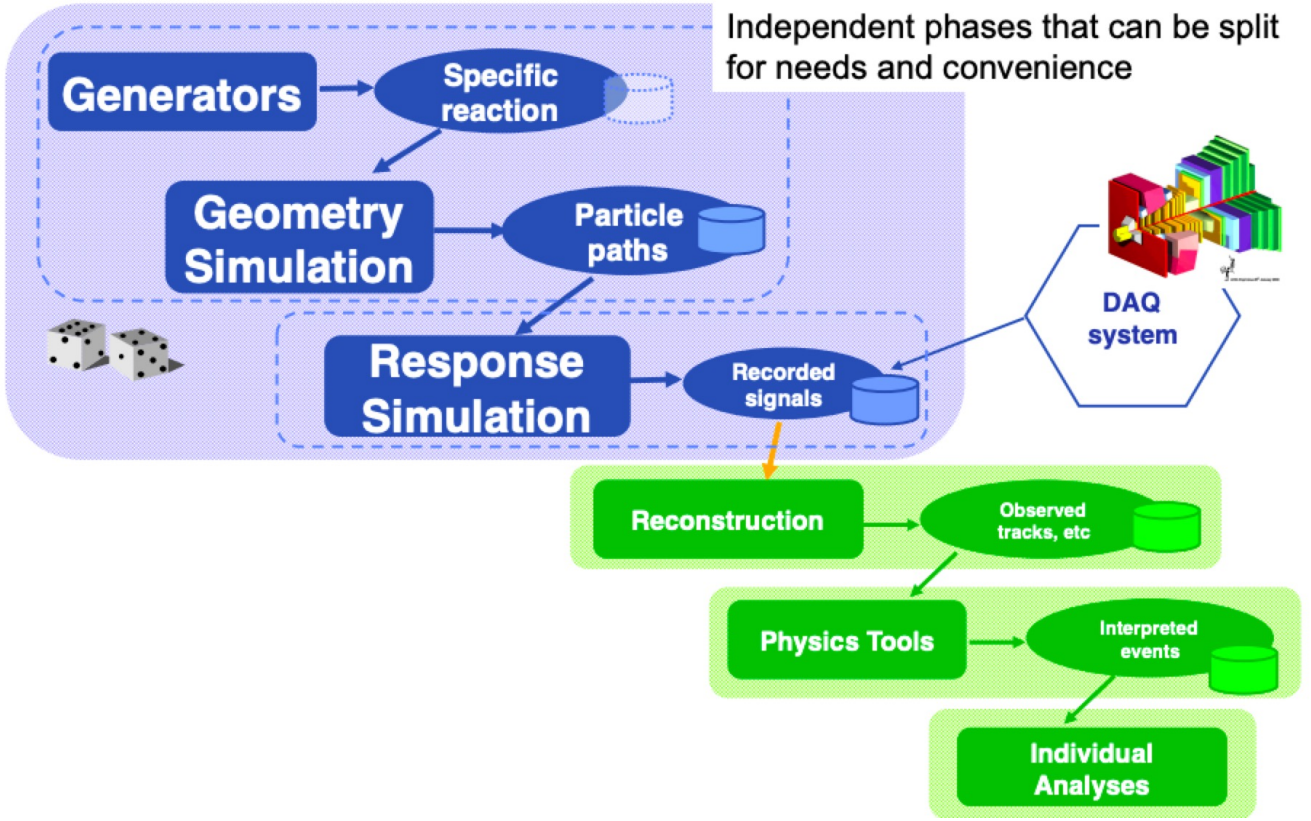
Simulation

Several model-motivated transitions.

Sequence:

- collision;
- decay;
- matter interaction;
- digitisation;
- reconstruction.

Each event takes 1 minute to generate (real world data is “generated” at several MHz).



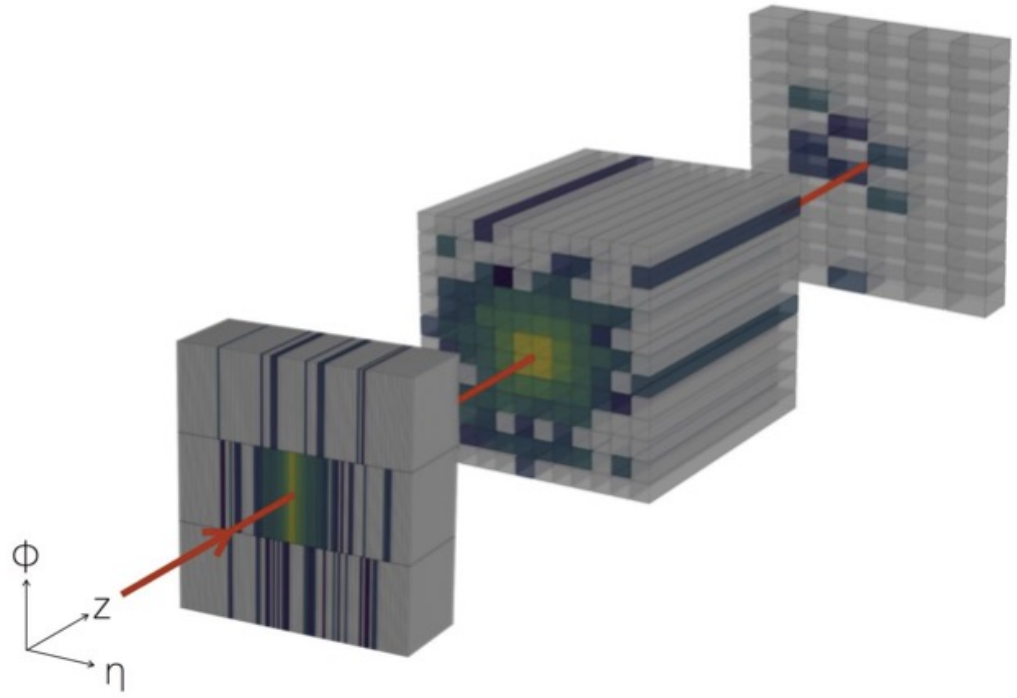
Simulate “simulation” using effective parameterization.

What is the event.

The calorimeter consists of many cells that reads out the energy deposit of a single particle.

A single particle deposits energy to several cells. An event is a sum of all particles and some noise.

We are normally using some reconstructed parameters of the event.



Paganini, M. et al. "CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks." *Physical Review D* 97.1 (2018): 014021.

Ideas for Simulation

Since we know all processes in the subdetector, we can fully simulate an event using precise physics-motivated rules.

For calorimeters this means taking into account the structure of response that consists of many secondary particles.

This is done using Geant toolkit.

Pro: physics behind the simulation is controlled
Cons: slow, needs fine tuning.

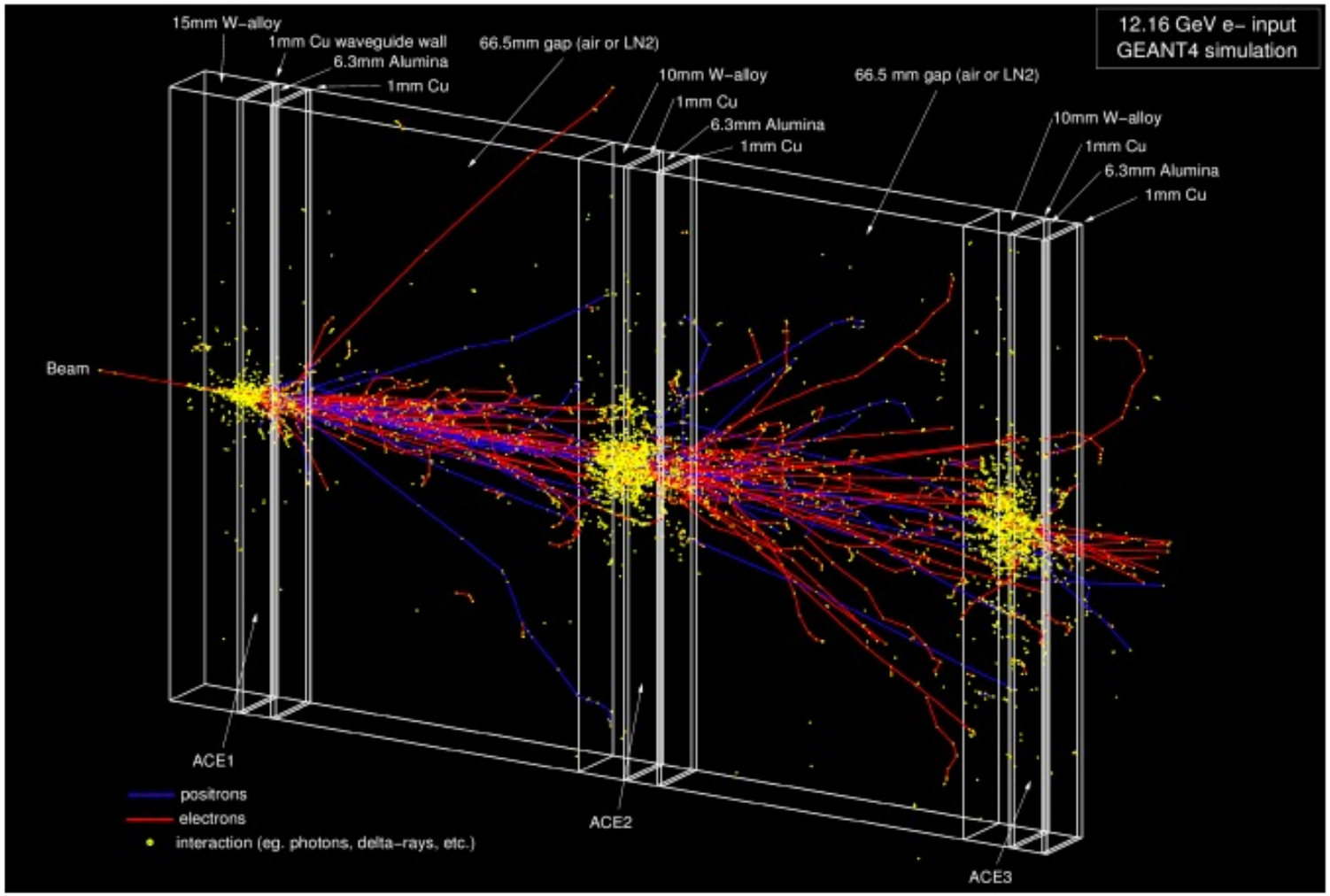
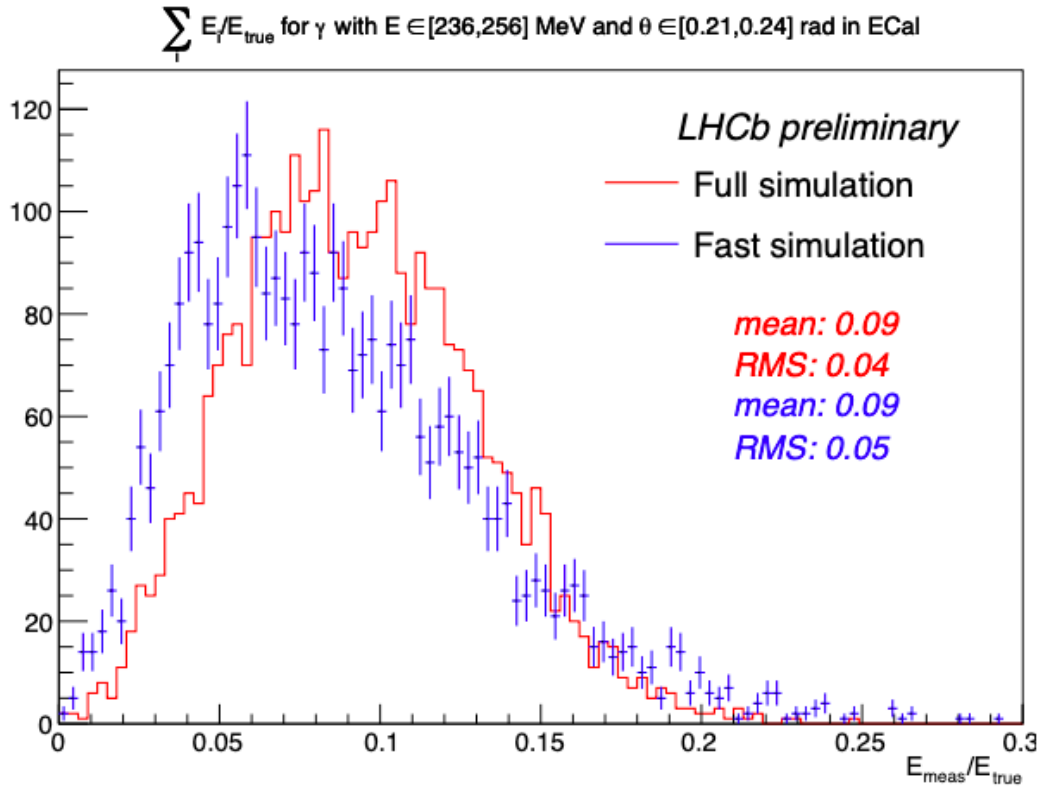


FIG. 2: Layout diagram, and GEANT4 simulation of a single 12.16 GeV electron event in our ACE detector system; in this case liquid nitrogen occupies the interelement spaces.

Ideas for Tabular Methods



Build a library of calorimeter responses to impact particle in corresponding 5D phase space using detailed simulation («frozen showers»).

5D = 3D momentum + 2D coordinate for every particle type.

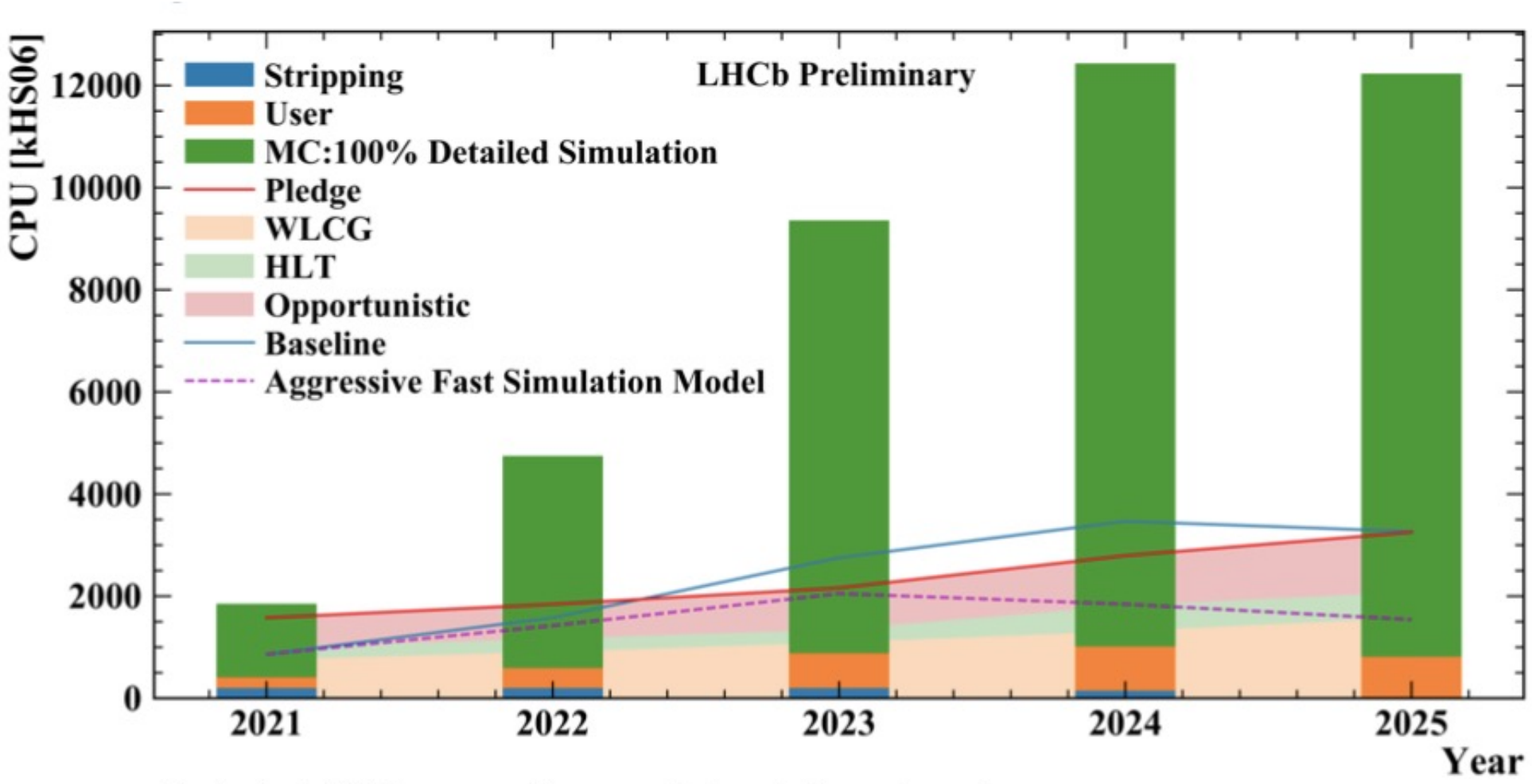
The whole phase space is split into bins, the exact observable is obtained interpolating between the bins.

One can also construct full interpolation (without using bins).

Pros: easy to interpret, quality is controlled by the number of samples.

Cons: curse of dimensionality, memory consumption, full interpolation takes huge efforts.

Upcoming Needs



Projected LHCb computing needs breakdown by category

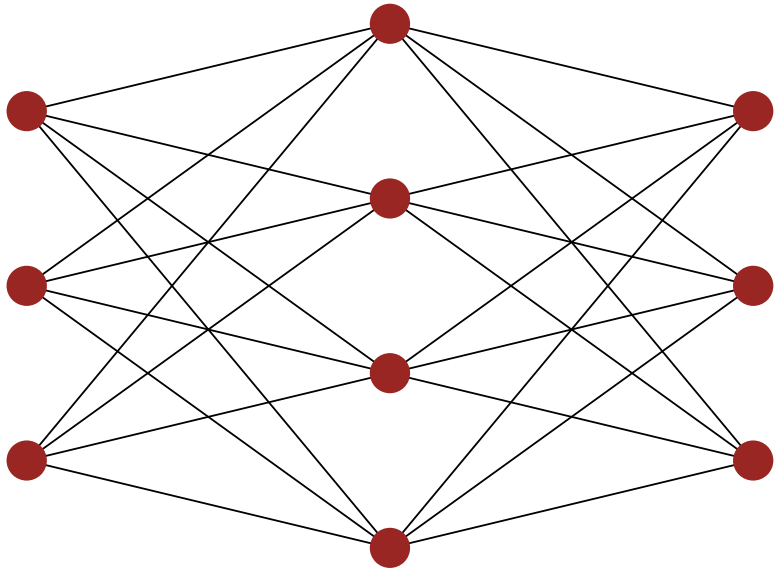
https://indico.cern.ch/event/773049/contributions/3474742/

Chapter Outcome

- Detailed Simulation of high-energy physics experiments is based on physics modeling.
- Speed up can be obtained by storing detector responses.
- The responses can be parameterized.

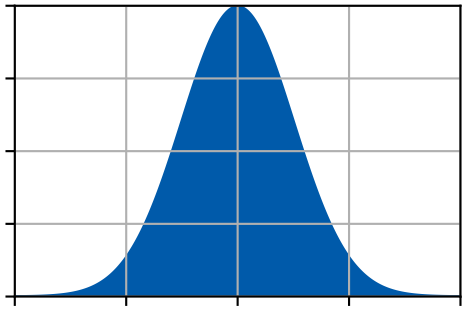
Machine Learning to Help

How can a neural network generate data?

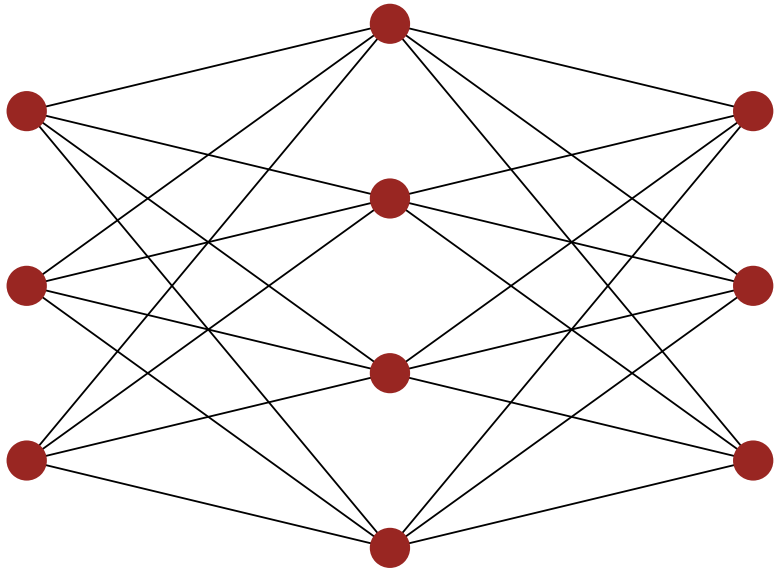


Neural network

How can a neural network generate data?



Random noize
e.g. multivariate normal

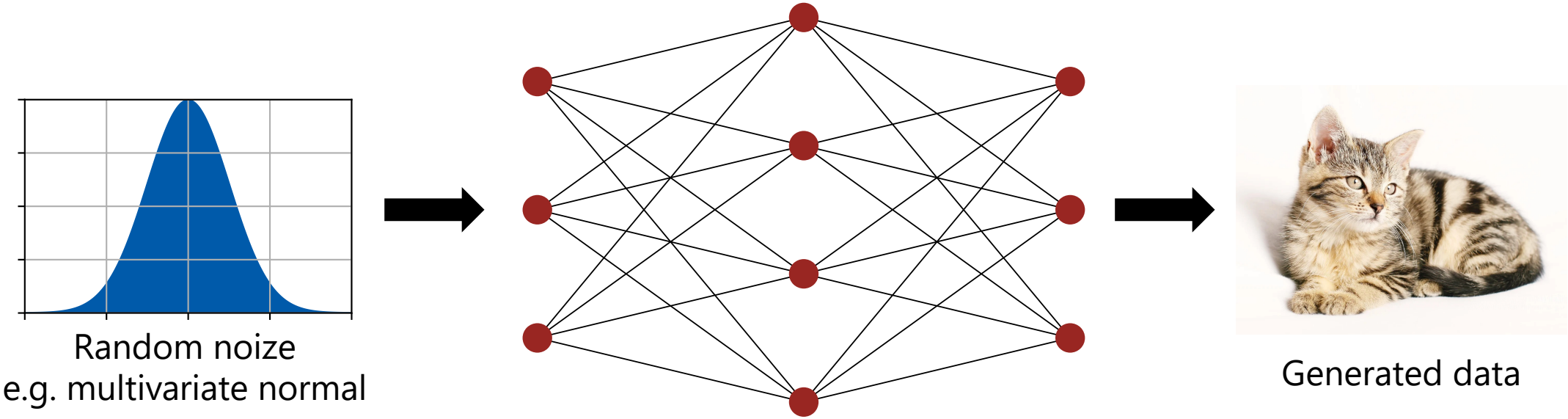


Neural network

Generated data

Cat image attribution: https://pixabay.com/users/chiemsee2016-1892688/

How can a neural network generate data?



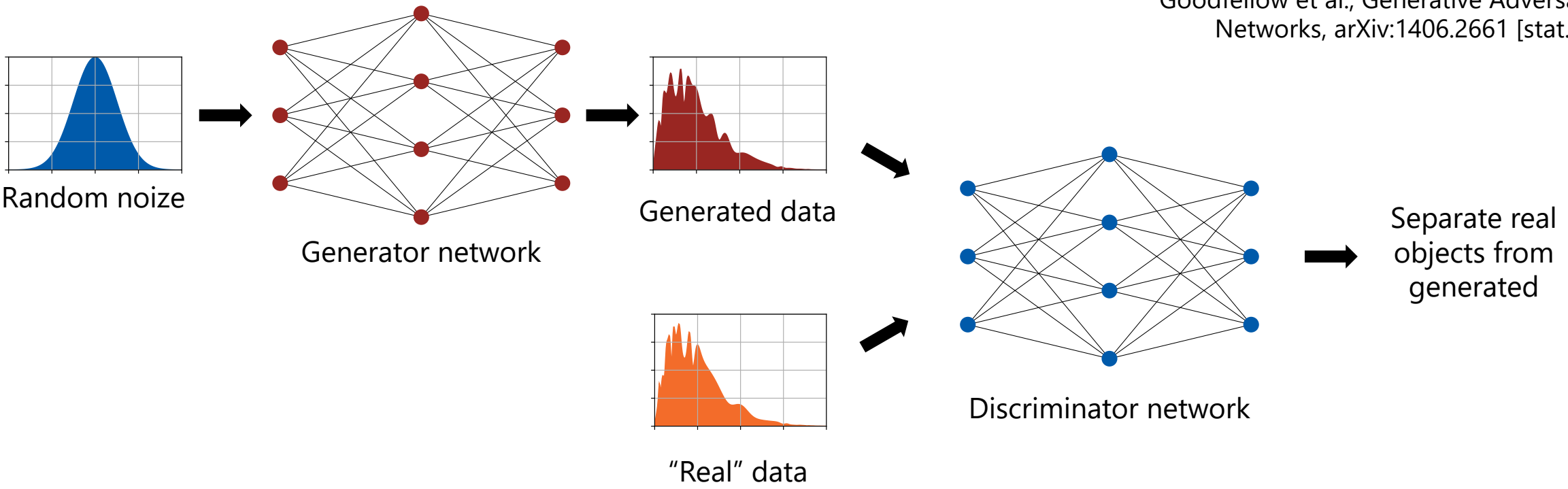
- This makes the generated object being a **differentiable function** of the network parameters

Cat image attribution: https://pixabay.com/users/chiemsee2016-1892688/

How to train such a generator?

- Generated object is a differentiable function of the network parameters
- Need a differentiable **measure of similarity** between the sets of generated objects and real ones
  - Can optimize with gradient descent
- How to find such a measure?

Adversarial approach



Goodfellow et al., Generative Adversarial Networks, arXiv:1406.2661 [stat.ML]

- Measure of similarity: how well can another neural network (discriminator) tell the generated objects apart from the real ones

Generator

- G_θ is a **generator**. It should sample from a random noise:

$$z_j \sim N(0;1);$$

$$x_j = \mathrm{G}_\theta(z_j).$$

- Our aim is G_θ as a neural network.
- We thus have a sample:

$$\{x_j\} \sim q_\theta(x)$$

- G_θ can be defined in many ways. For example, physics generator.

Borisyak M et al. Adaptive divergence for rapid adversarial optimization. *PeerJ Computer Science* 6:e274 (2020)

Discriminator

- Add a classifying neural network, **discriminator D_ϕ**, to distinguish between the real and generated samples.
- Optimize:

$$\max_{\phi} \left(\mathbb{E}_{x \sim p(x)} (\log(D_\phi(x)) + \mathbb{E}_{\tilde{x} \sim q_\theta(x)} (1 - \log(D_\phi(\tilde{x}))) \right)$$

Real samples

Generated samples

G+D recap

We can now put together generator and discriminator.

- objective of discriminator:

$$\max_{\phi} \left(\mathbb{E}_{x \sim p(x)} (\log(D_\phi(x)) + \mathbb{E}_{z \sim \mathcal{N}(0;1)} (1 - \log(D_\phi(G_\theta(z))) \right).$$

- objective of generator:

$$\min_{\theta} \mathbb{E}_{z \sim \mathcal{N}(0;1)} (1 - \log(D_\phi(G_\theta(z))$$

We thus defined a minimax game:

$$\min_{\theta} \max_{\phi} \mathbb{E}_{x \sim p, \tilde{x} \sim q_\theta} V(f_\phi(x), f_\phi(\tilde{x})).$$

In exactly the way we wanted.

GAN results

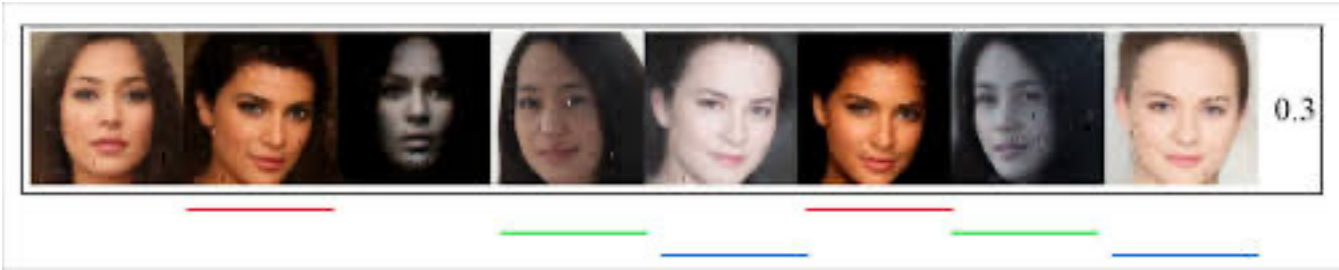
Figure 2: Visualization of samples from the model. Rightmost column shows the nearest training example of the neighboring sample, in order to demonstrate that the model has not memorized the training set. Samples are fair random draws, not cherry-picked. Unlike most other visualizations of deep generative models, these images show actual samples from the model distributions, not conditional means given samples of hidden units. Moreover, these samples are uncorrelated because the sampling process does not depend on Markov chain mixing. a) MNIST b) TFD c) CIFAR-10 (fully connected model) d) CIFAR-10 (convolutional discriminator and "deconvolutional" generator)

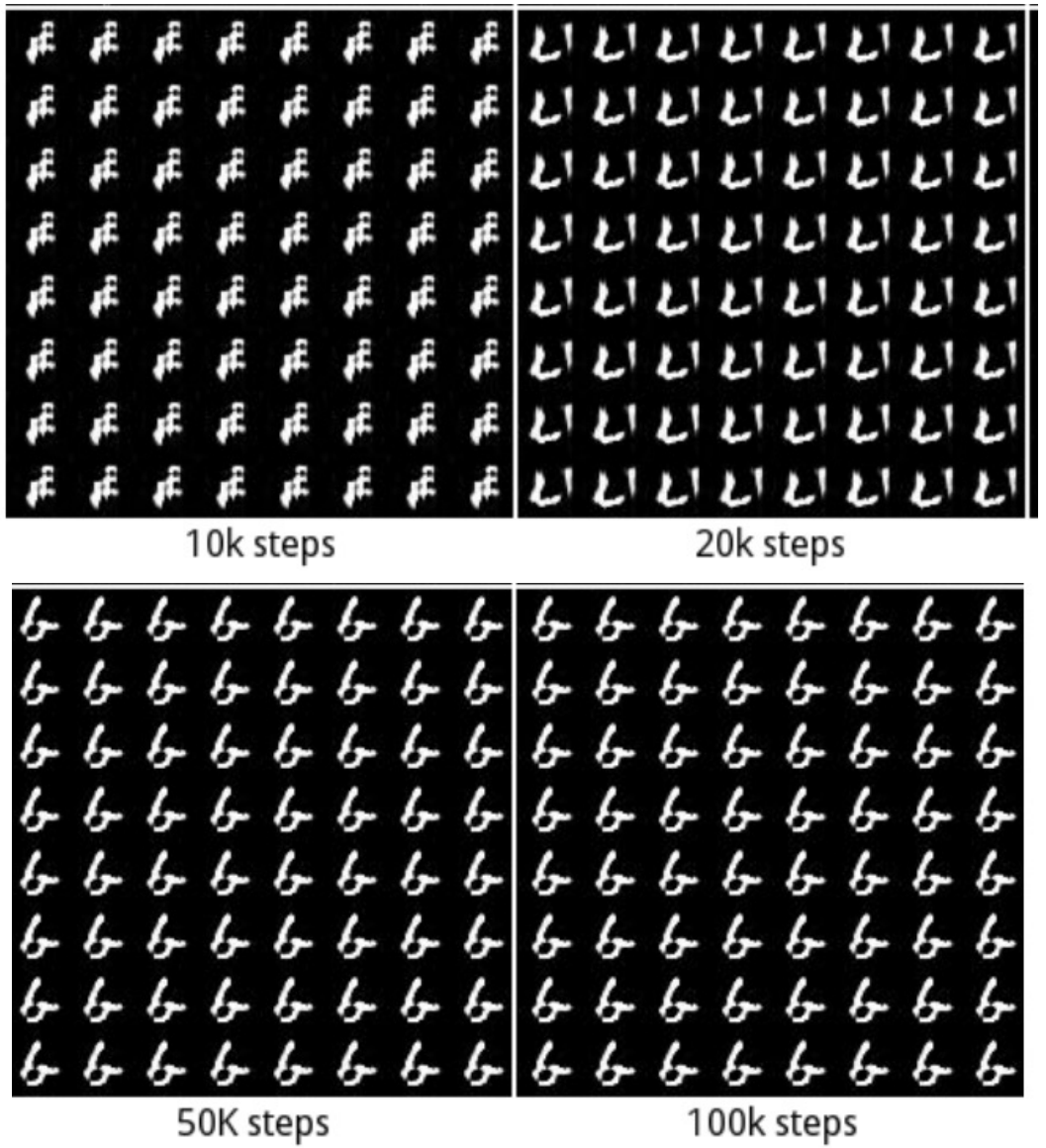
I. Goodfellow,et al. Generative Adversarial Networks, NIPS 2014

Mode Collapse

- GANs choose to generate a small number of modes due to a defect in the training procedure, rather than due to the divergence they aim to minimize.

I. Goodfellow NIPS 2016 Tutorial: Generative Adversarial Network





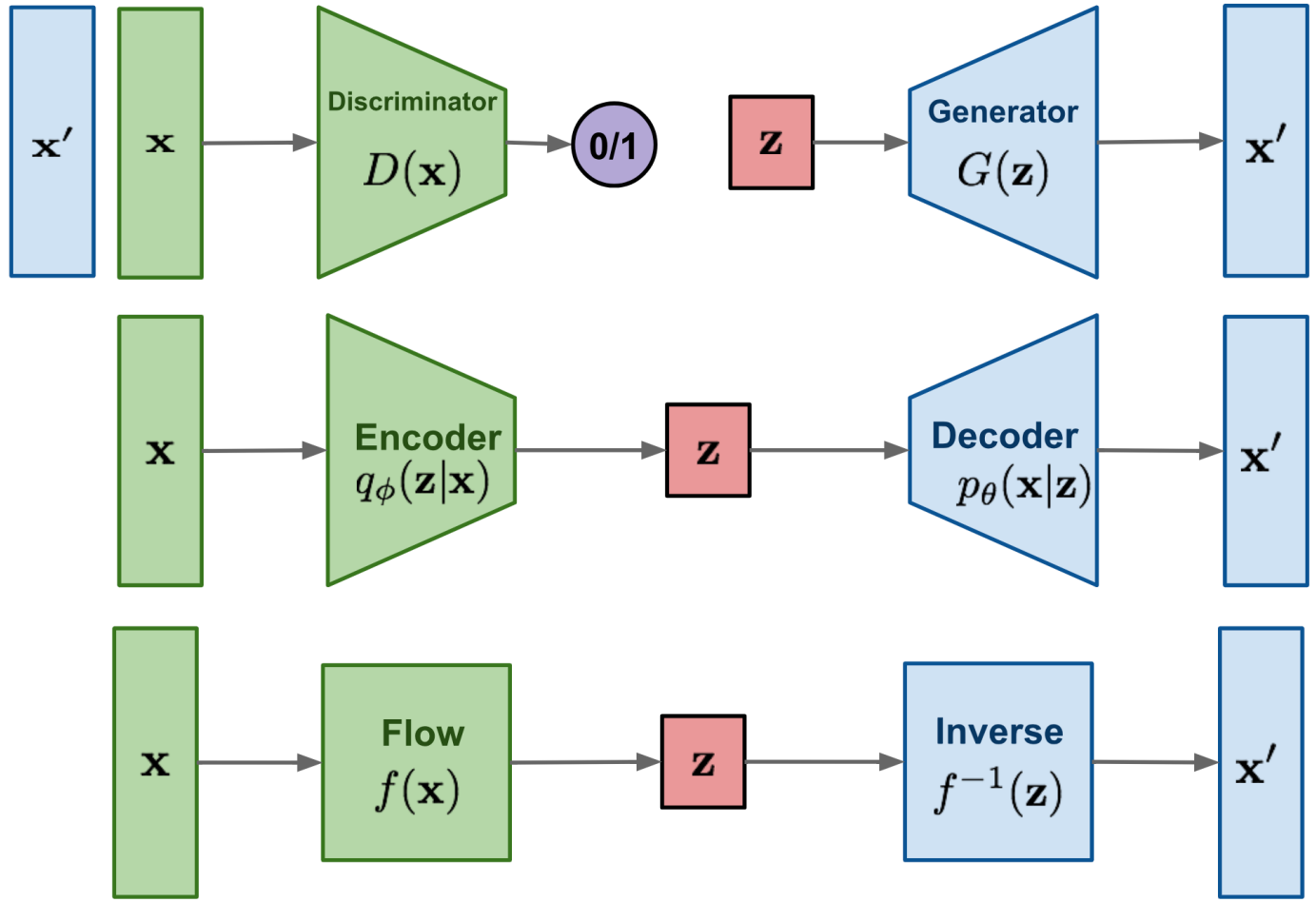
Luke Metz et al Unrolled Generative Adversarial Networks ICLR 2017

Generative zoo

GAN: minimax the classification error loss.

VAE: maximize ELBO.

Flow-based generative models: minimize the negative log-likelihood



Generative zoo choice

Generative adversarial Networks (GAN)

Pros:

- good likely objects generation quality;
- fast sampling.

Cons:

- hard to control tails;
- mode collapse.

Variational Autoencoders (VAE)

Pros:

- explorable representation space;
- average sampling.

Cons:

- blurred image;

Flow models (normalizing flows)

Pros:

- easy to evaluate likelihood;
- explicit modeling;

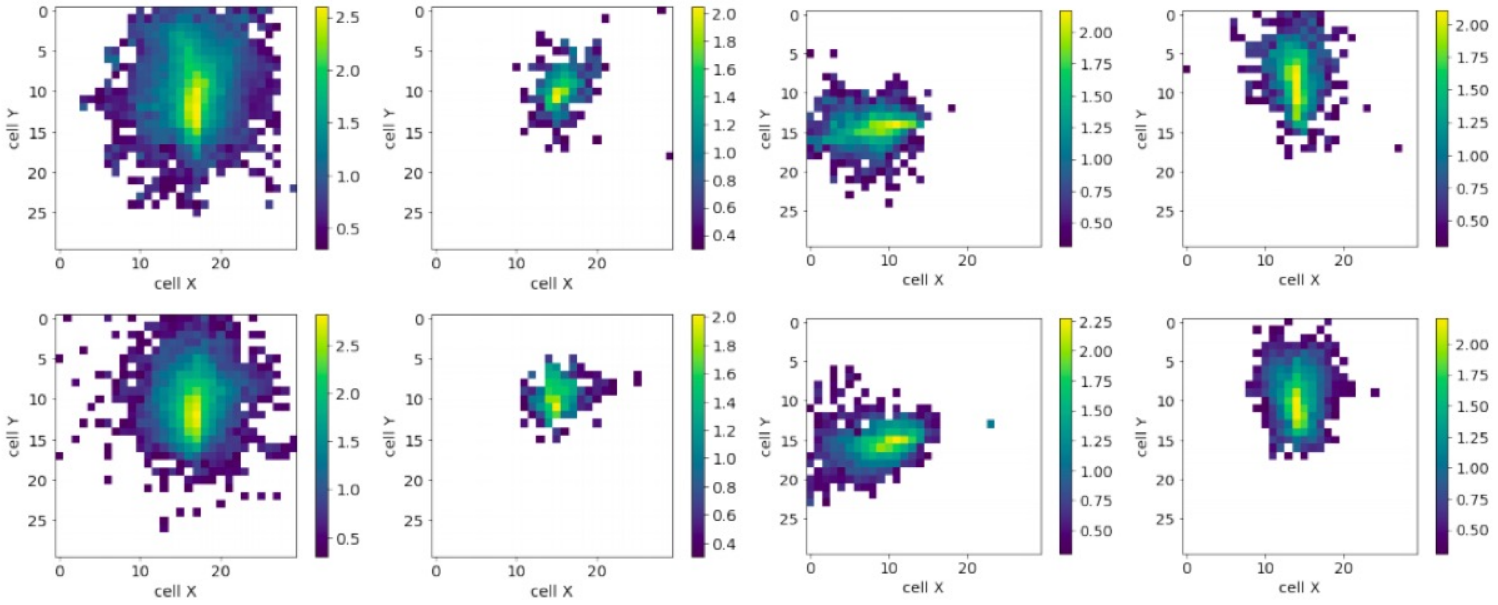
Cons:

- very slow sampling.

Generative modeling for HEP

- **Conditional** dependence on detector parameters and incident particle information.

Need to be
- Tunable.
- Robust.
- Fast for sampling.



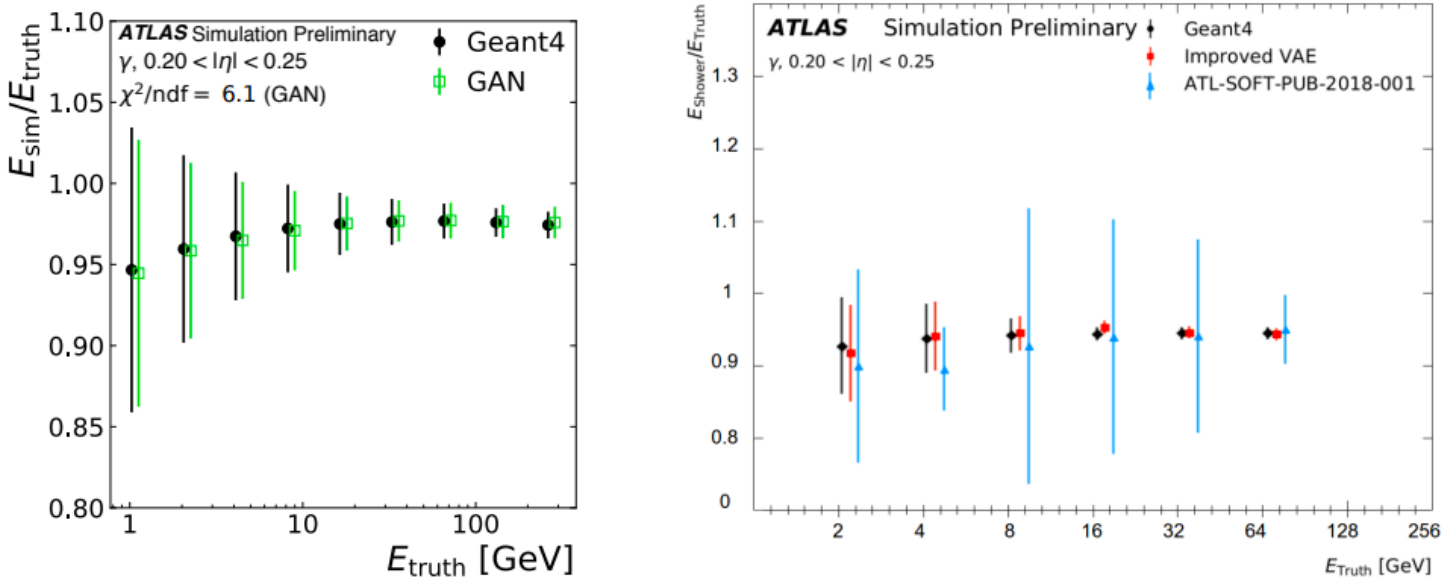
(a) $E_0 = 63.7$ GeV (b) $E_0 = 6.5$ GeV (c) $E_0 = 15.6$ GeV (d) $E_0 = 15.9$ GeV

V. Chekalina et al. *EPJ Web Conf.* 214 (2019) 02034

Generative Models for Fast Simulation

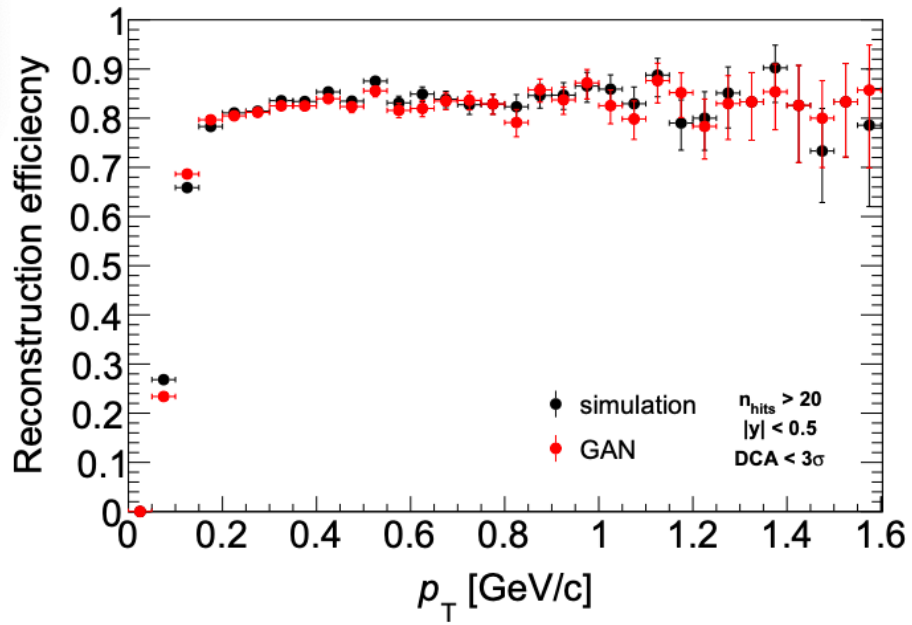
- Many neural based generative description attempted in recent years

ATLAS: VAE and GAN for Calorimeter



Chapman et al., EPJ Web of Conferences **245**, 02035 (2020)

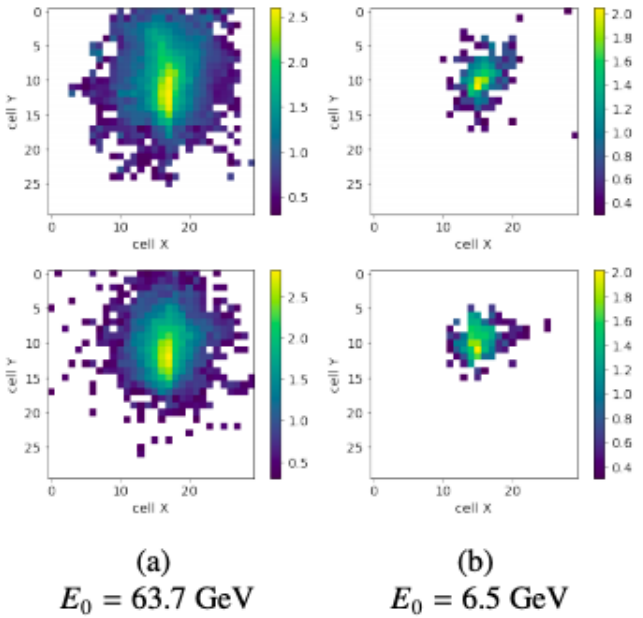
MPD: GAN for TPC



A. Maevskiy et al. Eur.Phys.J.C 81 (2021) 7, 599

Generative Models

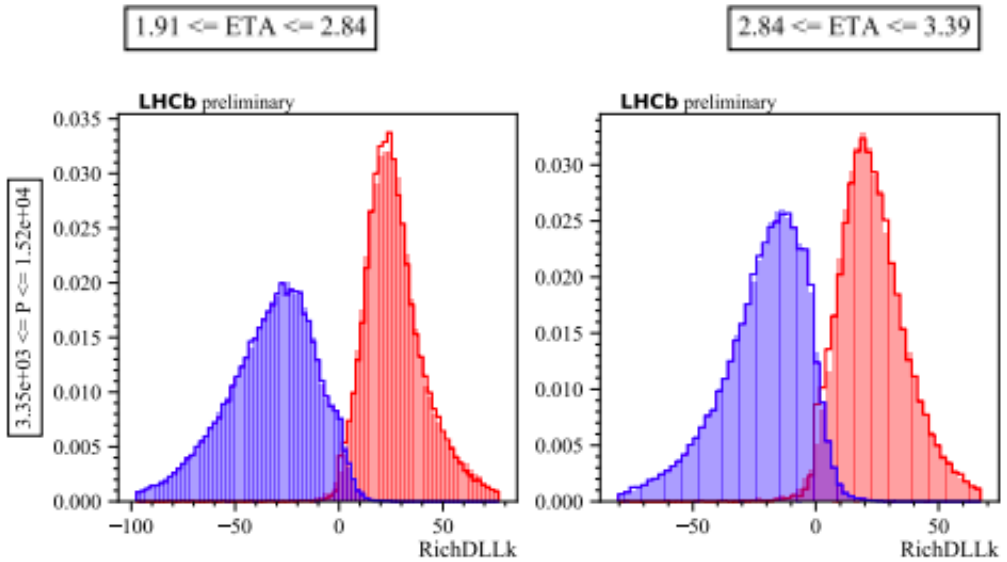
Direct simulation of calorimeter responses



(a) $E_0 = 63.7$ GeV (b) $E_0 = 6.5$ GeV

V. Chekalina et al. EPJ WoC: 214, 02034 (2019)

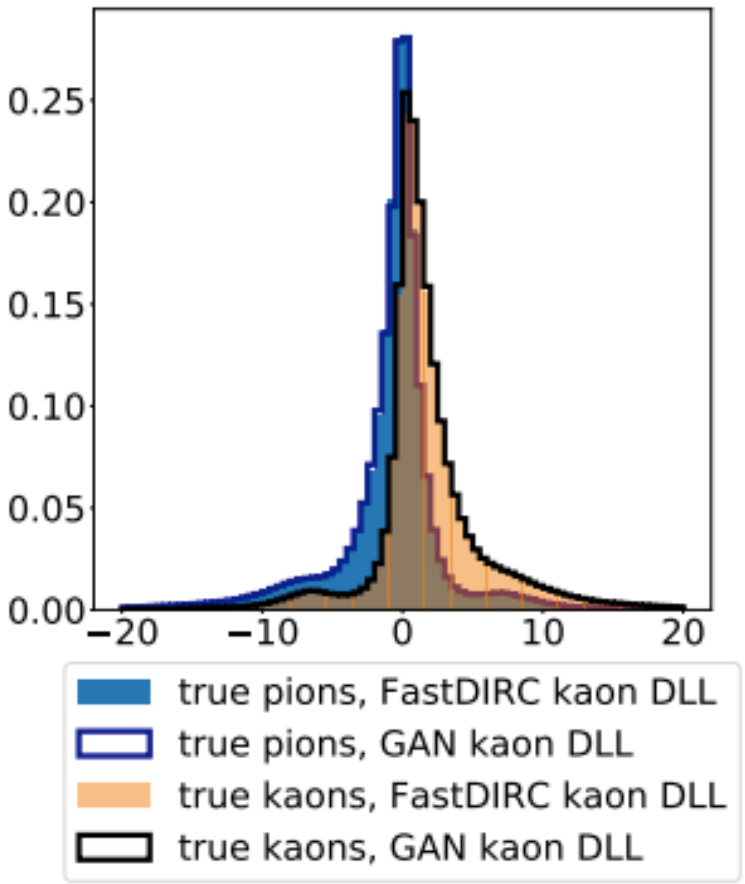
Simulation of reconstruction output for RICH and Muon



A. Maevskiy et al., ML4PHYS@Neurips 2019

- NB: KDE can also be considered as a generative model.

DIRC Example

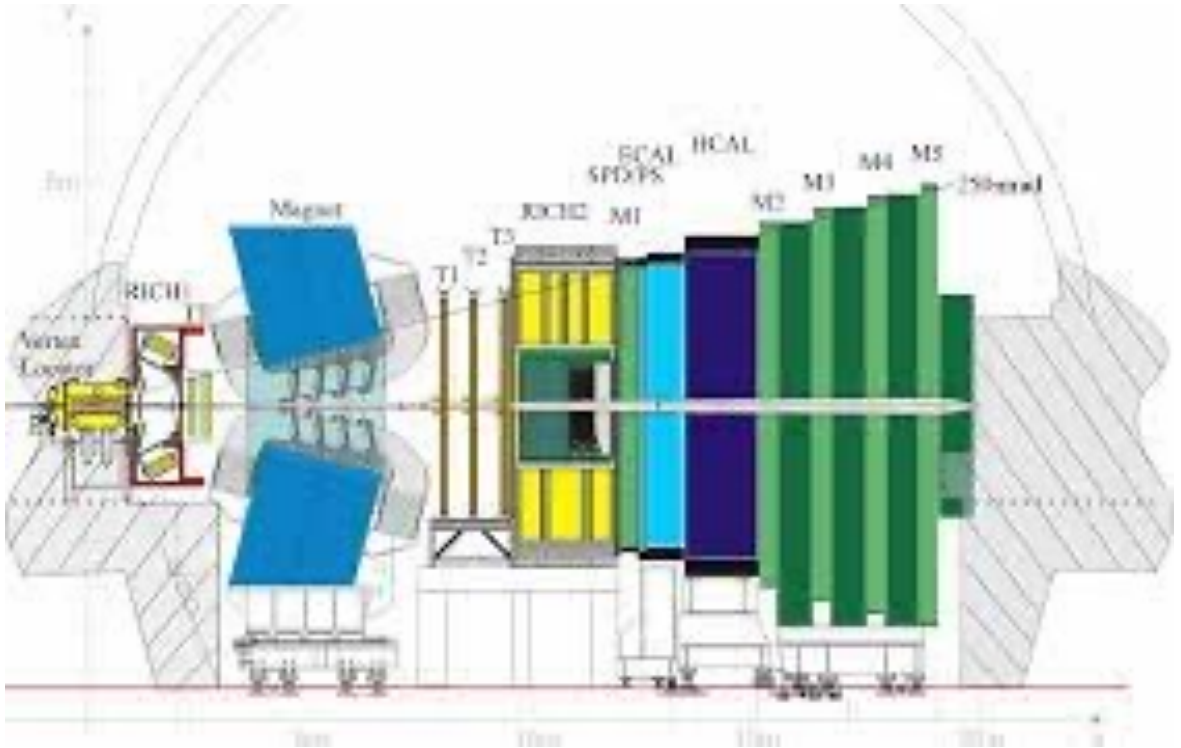


Derkach et al,Nucl.Instrum.Meth.A 952 (2020) 161804

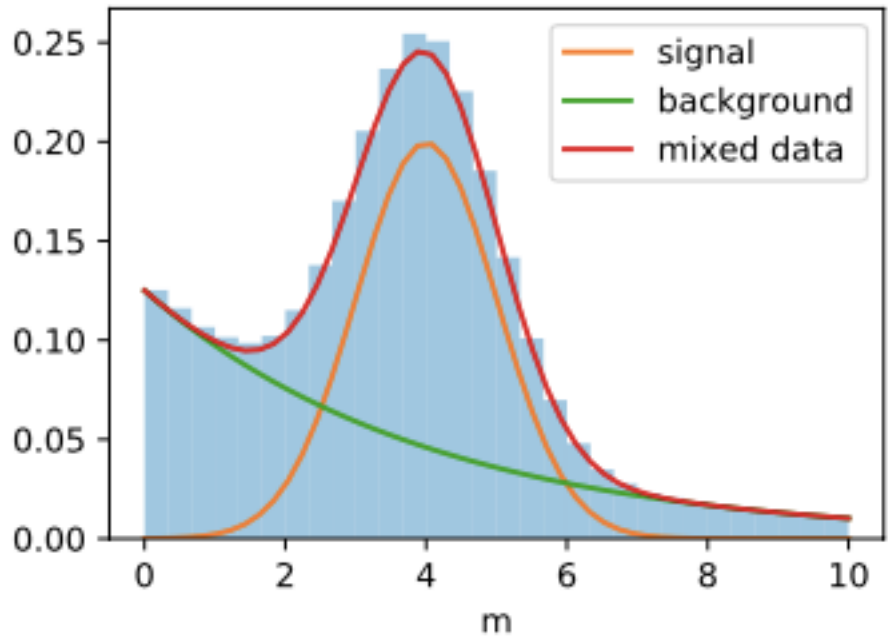
https://github.com/yandexdataschool/mlhep2019/blob/master/notebooks/day-6/06_DIRC_GAN_solution.ipynb

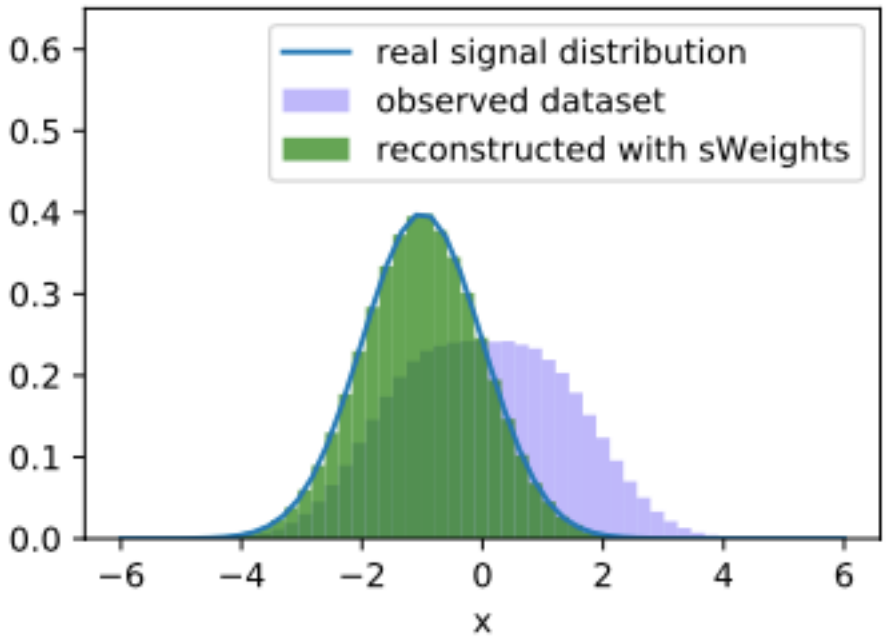
Why it works/should work?

- Treatment of physics data as pictures.
- Expressivity of NN solutions.
- Decomposition of data.



Challenges: Training Samples

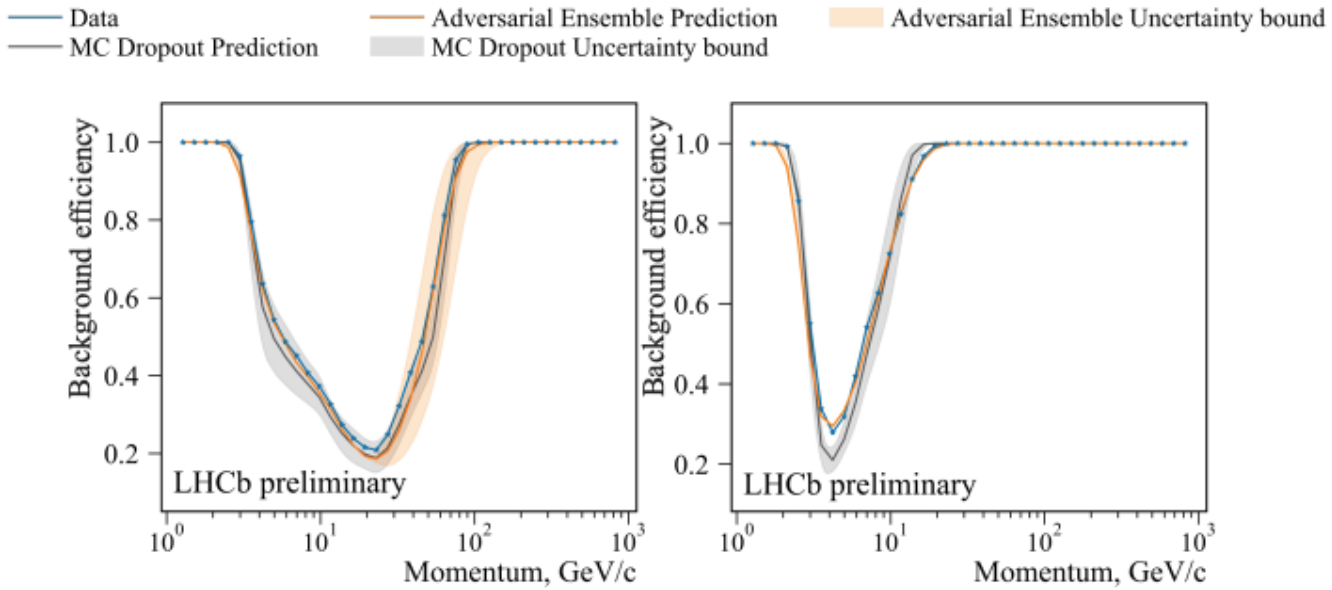




- Use real data sample, but we need to reduce noise from it.
- Model information introduced in the training procedure using maximum likelihood fit.

A. Maevskiy et al., Neurips 2019 Workshop

Challenges: uncertainty



Uncertainty is a key for the use of generative modeling in natural sciences.

An ensemble of several GANs with the following loss functions and training procedure:

$$f(\boldsymbol{y}) = \left\| D(\boldsymbol{y}) - D(\boldsymbol{y}'_g) \right\|_2 - \|D(\boldsymbol{y})\|_2$$

$$L_G = f(\boldsymbol{y}_r) - f(\boldsymbol{y}_g) - \alpha \left\| D(\boldsymbol{y}_g) - D(\boldsymbol{y}_{\cup_g}) \right\|_2$$

$\boldsymbol{y}_{\cup_g}$ is a concatenation of the predictions of the ensemble, corresponding to a model with averaged probability density

N. Kazeev et al., ACAT-2021

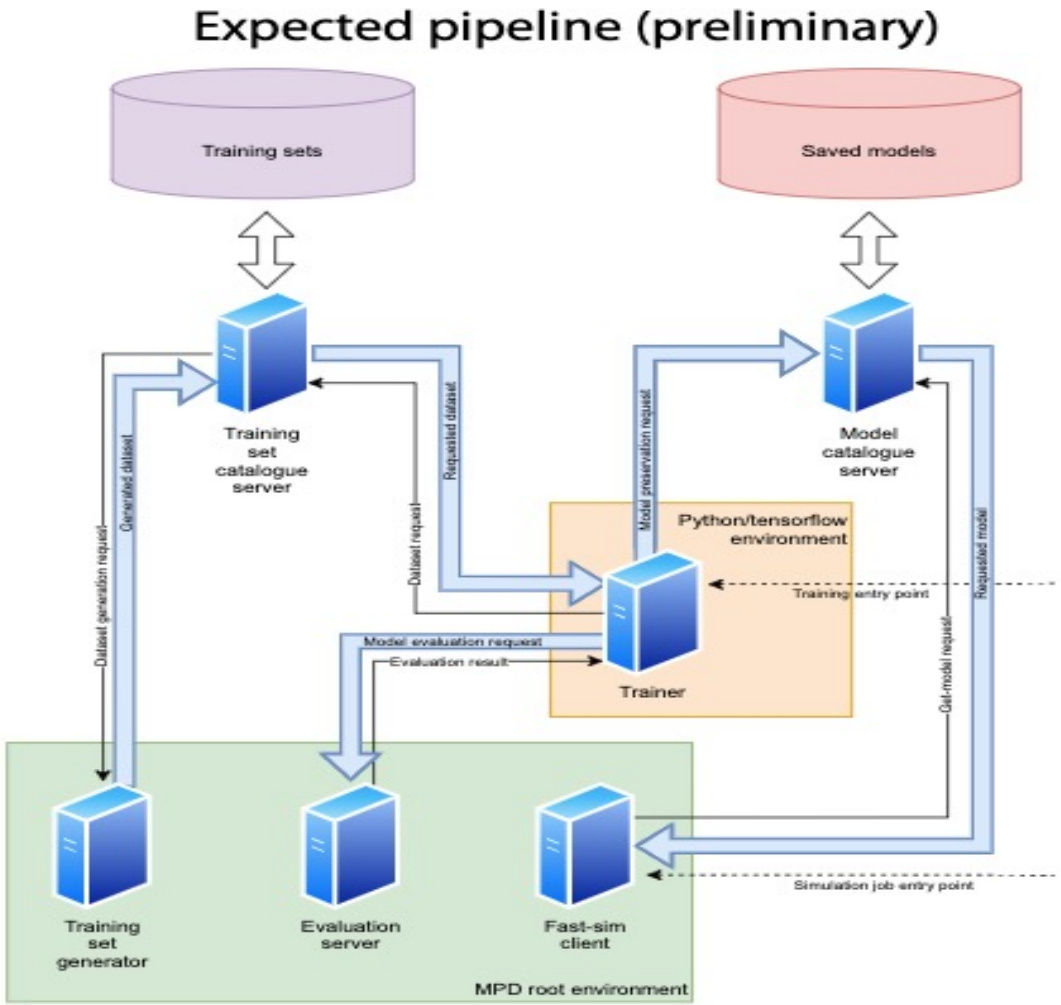
Challenges: Implementations

More challenges:

- **Distilling the generators.**
  Aim: beyond 100ms/event.

- **Testing the generator quality in the limit of small data samples.**
  Aim: on-the-fly algorithms.

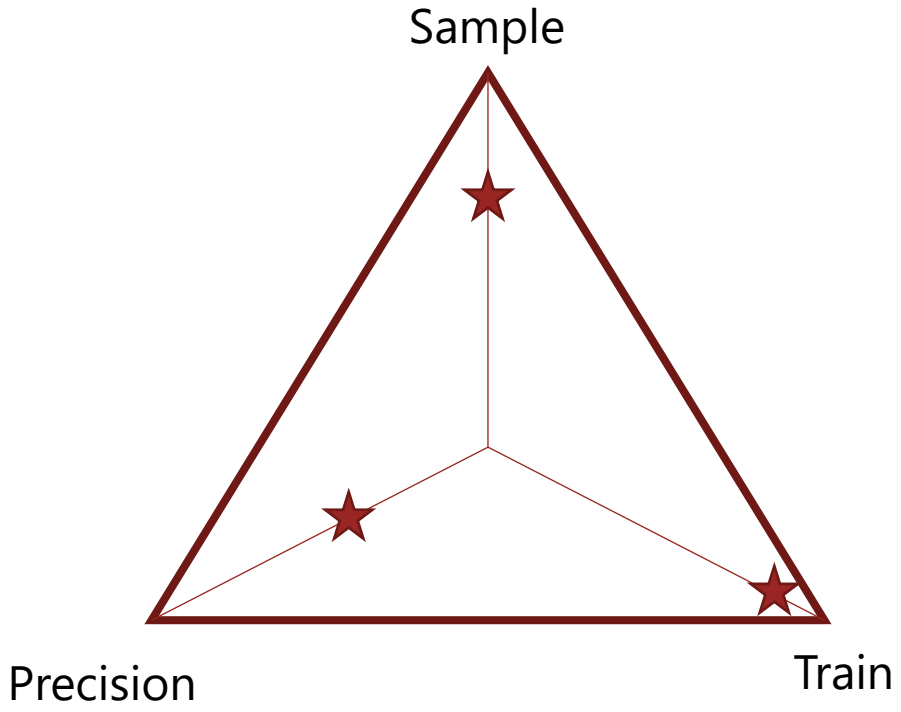
- **Implementing pipeline in the online environment (200xNVidia RTX A5000 from LHCb).**
  Aim: Efficient architecture and Scheduling given resources.



Sukhorosov BSc Diploma

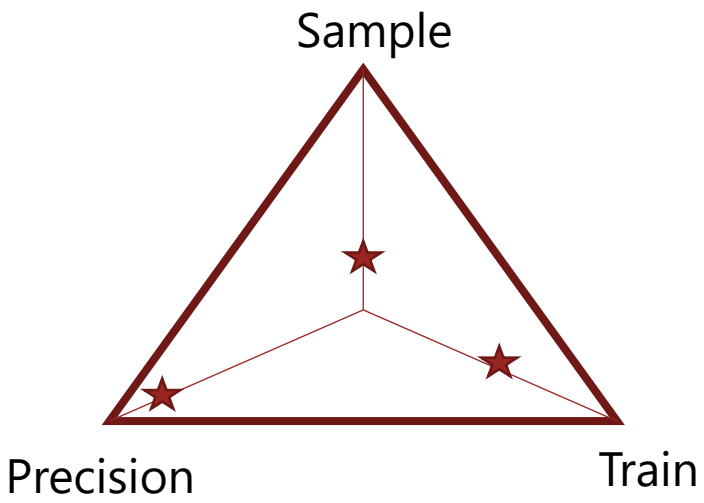
Generative Models Characteristics

- Fast Sampling:
  - much faster than detailed MC;
  - models can get complicated;
  - current RICH-LHCb simulation speed ~70 ms.
- Very Fast training:
  - retrain can be done very fast;
  - train process still should be periodically controlled;
  - current RICH-LHCb model trains ~2 days using GPU.
- Good Precision:
  - complicated models can be quite precise;
  - precision is controlled by train sample statistics;
  - need to understand influence on the final systematics.

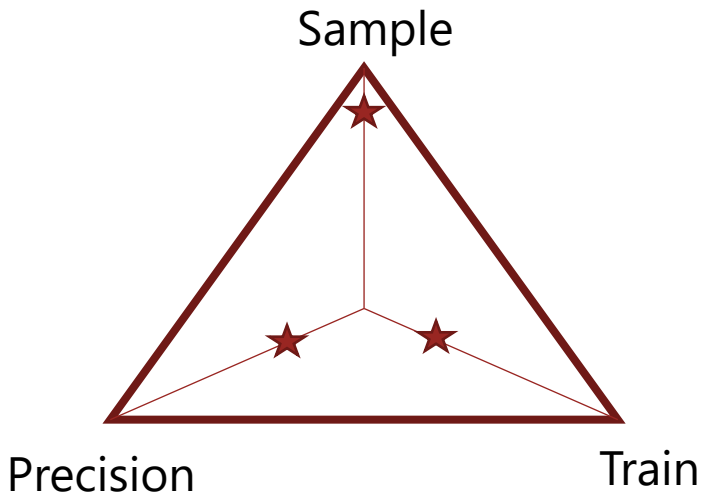


Simulation Picture

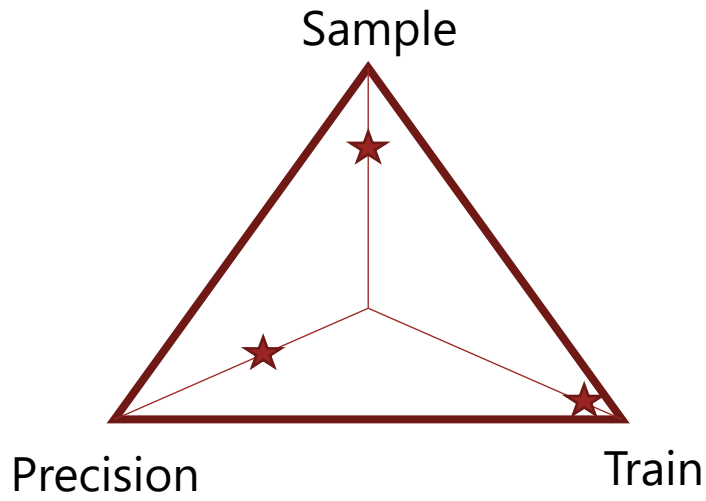
Detailed simulation



Parametric simulation



Machine learning simulation



Each approach has vices and virtues.

A possibility to have easily retrainable model can give several benefits in case of using machine learning.

(*) are my opinion

More Developments

- Generative models at microscopic detailed simulation.
- Generative models for fast integration (for example, lattice calculations).
- Generative models for detector optimization studies.
- Likelihood-free inference for New physics scenarios.
- Generative based data compression.

Final Outcome

- Generative modeling had a great boost in physics uses with arrival of advanced machine learning techniques.
- With more precise machine learning results we expect more implementation scenarios.
- Look for results at:
  - https://cs.hse.ru/lambda/