CERN openlab

Estimating Support Size of the 3DGAN

CERN openlab students meeting

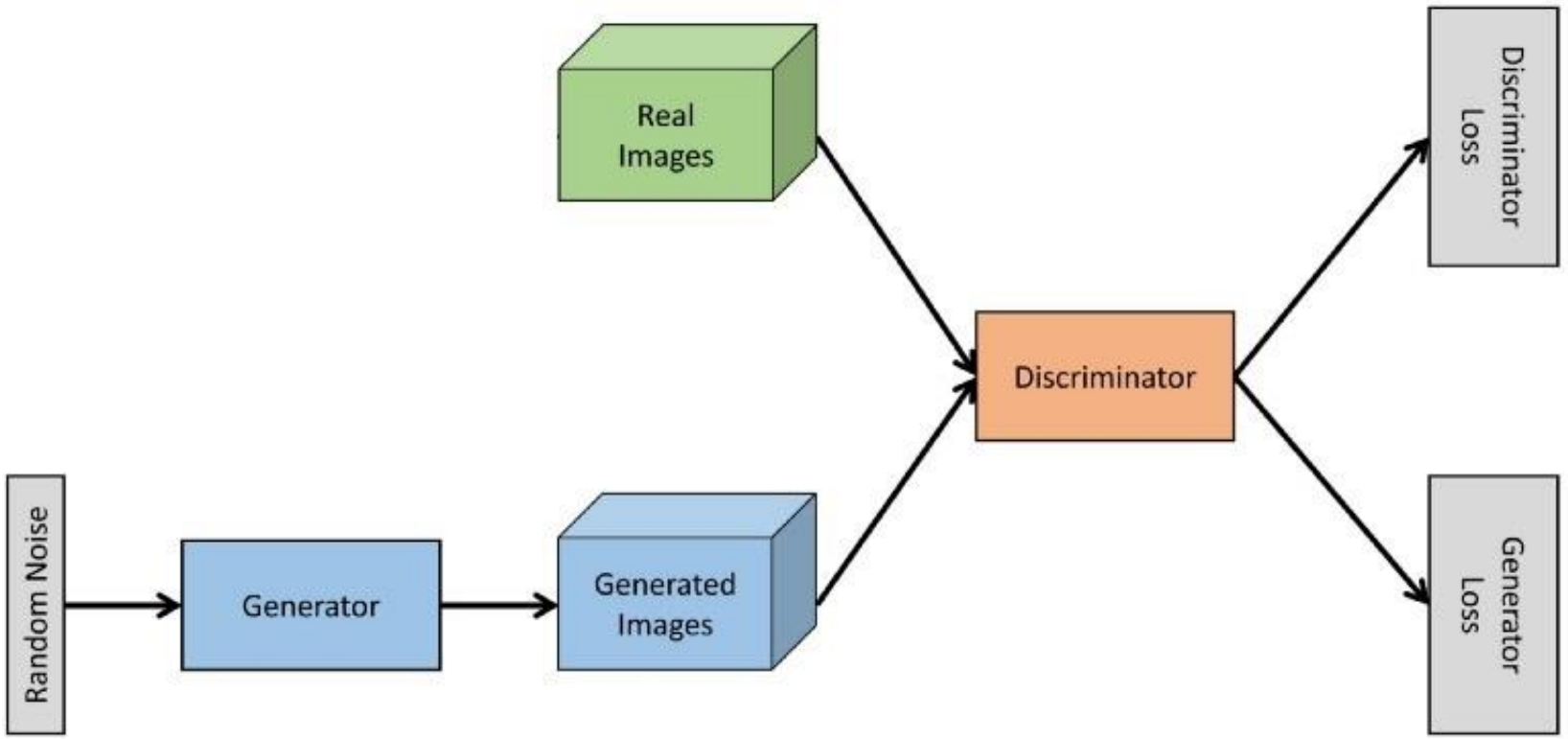
Kristina Jaruskova, Czech Technical University in Prague

CERN Supervisor: Sofia Vallecorsa

September 2020

What is GAN?

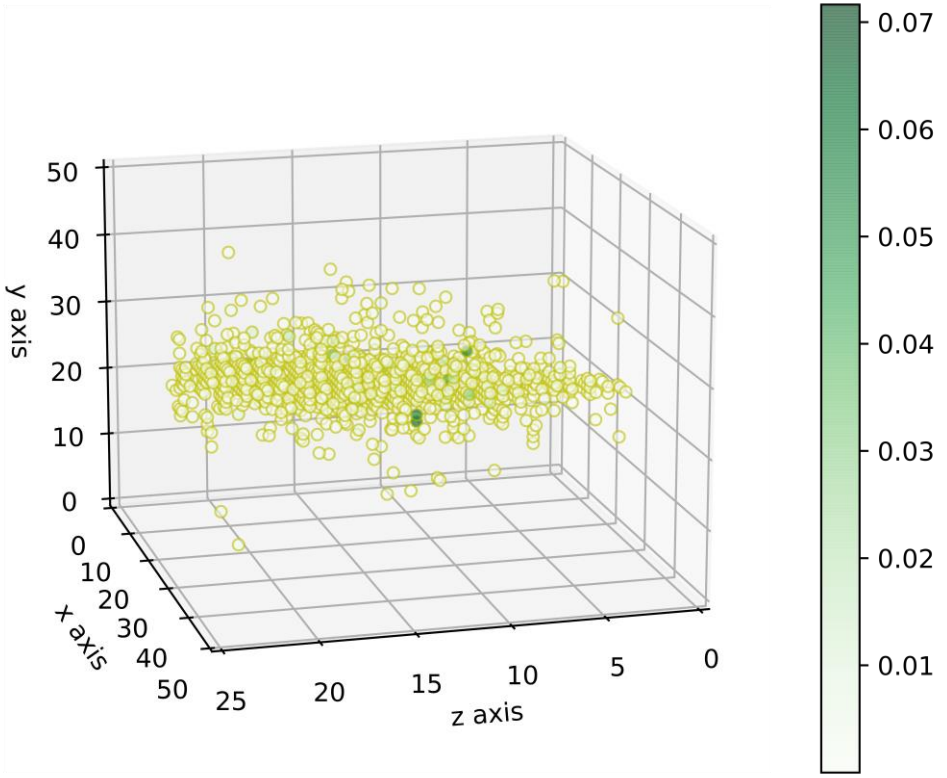
Generative Adversarial Network



3DGAN

- Convolutional GAN architecture
- Calorimeter's energy response
- 3D convolutions
- Alternative approach to the Monte Carlo simulations

- Output: 3D image (51x51x25) representing the deposited energy



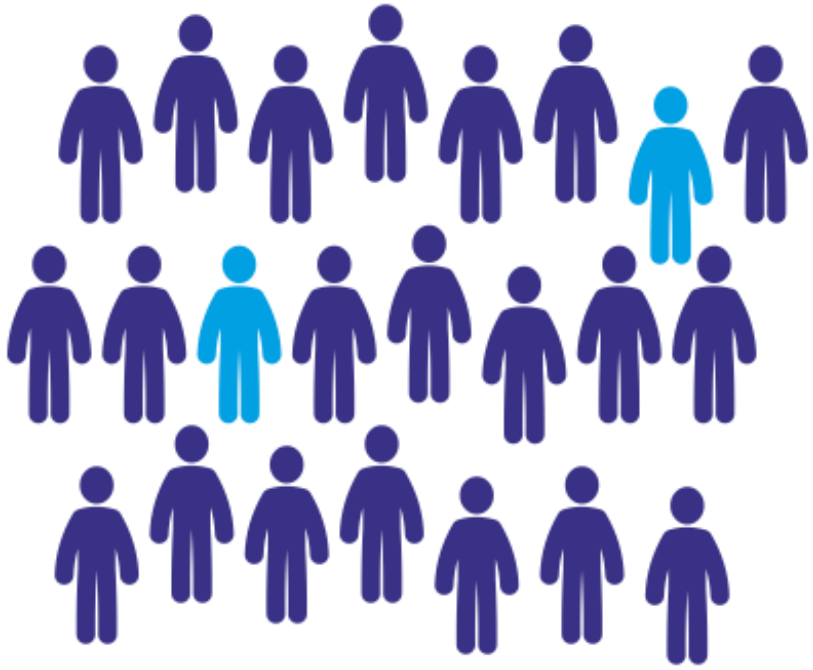
Validation based on Birthday Paradox

Birthday paradox

- How many people need to be in one room so that
  P(*at least two people were born on the same day of the year*) > 0.5 ?

- 365 (366) days in a year
  -> 23 people is enough

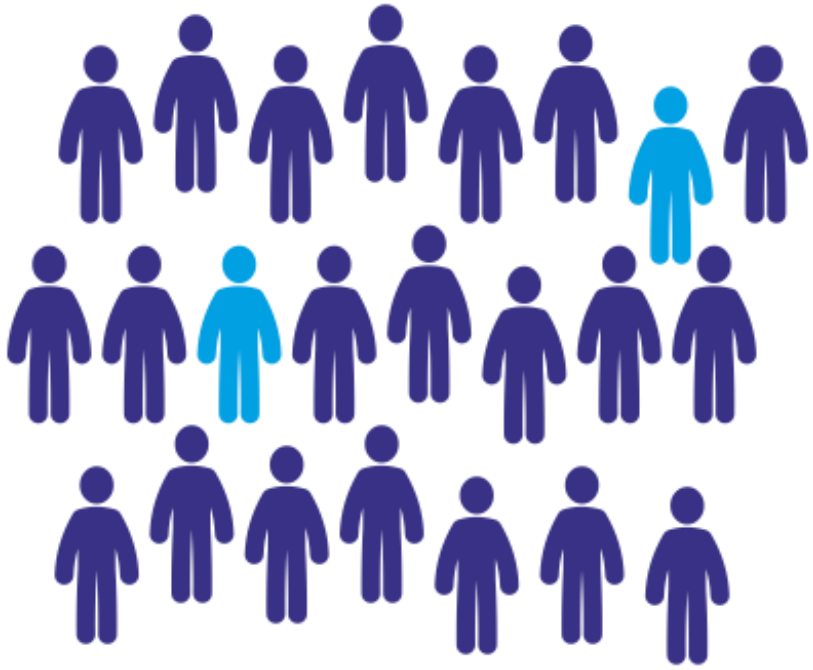
- For a year with d days, approx. $\sqrt{d}$ people are needed.



Validation based on Birthday Paradox

For 3DGAN:

- How many samples do I need to generate at least one pair of duplicate samples with the probability of 50 %?
  - (The answer)^2 = estimate of the support size
- How many training data do I need to take to encounter duplicates?
- Goal:
  - Support size of GAN ≈ support size of training data
- How to define the duplicate?



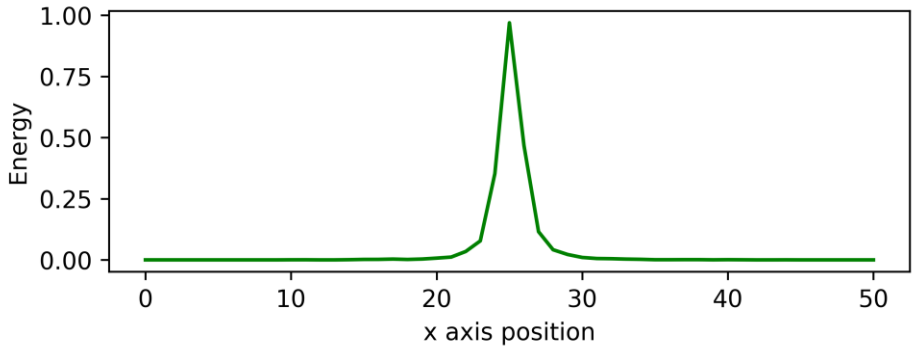
Birthday Paradox for 3DGAN

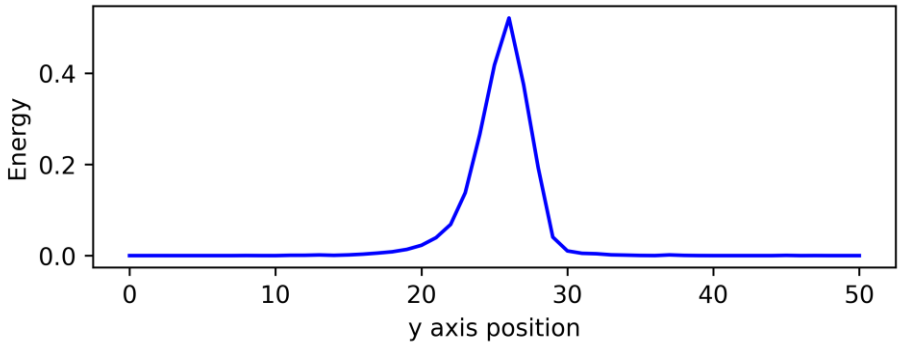
Definition of duplicates

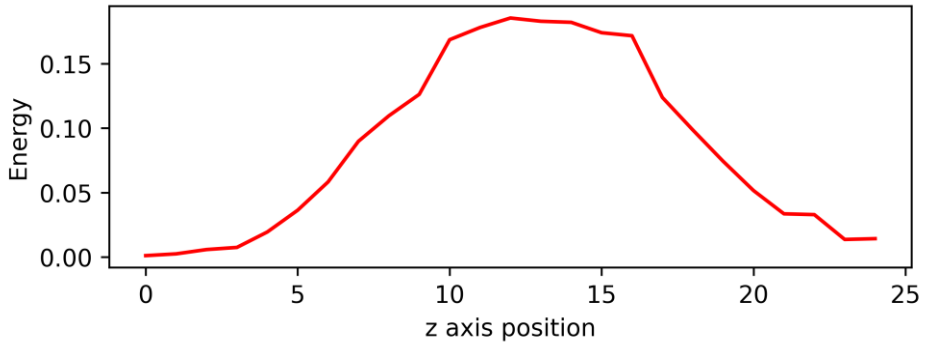
- Quantities to asses:
  - Energy distributions along the main axes (*x*, *y*, *z*)
- Metric of the distance:
  - Jensen-Shannon divergence

$$D_{JS}(P,Q) = \frac{1}{2} D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

- Threshold for the distance
  - Quantile value computed on GEANT4 data (0.05 and 0.02-quantile)
- Joint condition (distance below threshold for all three axes)



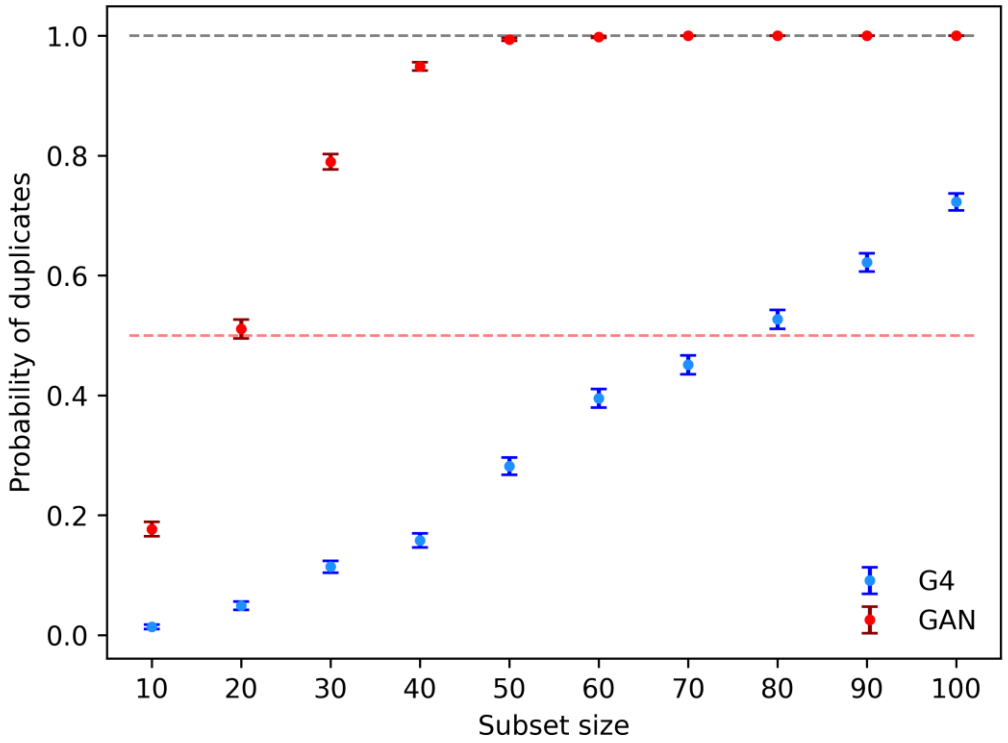




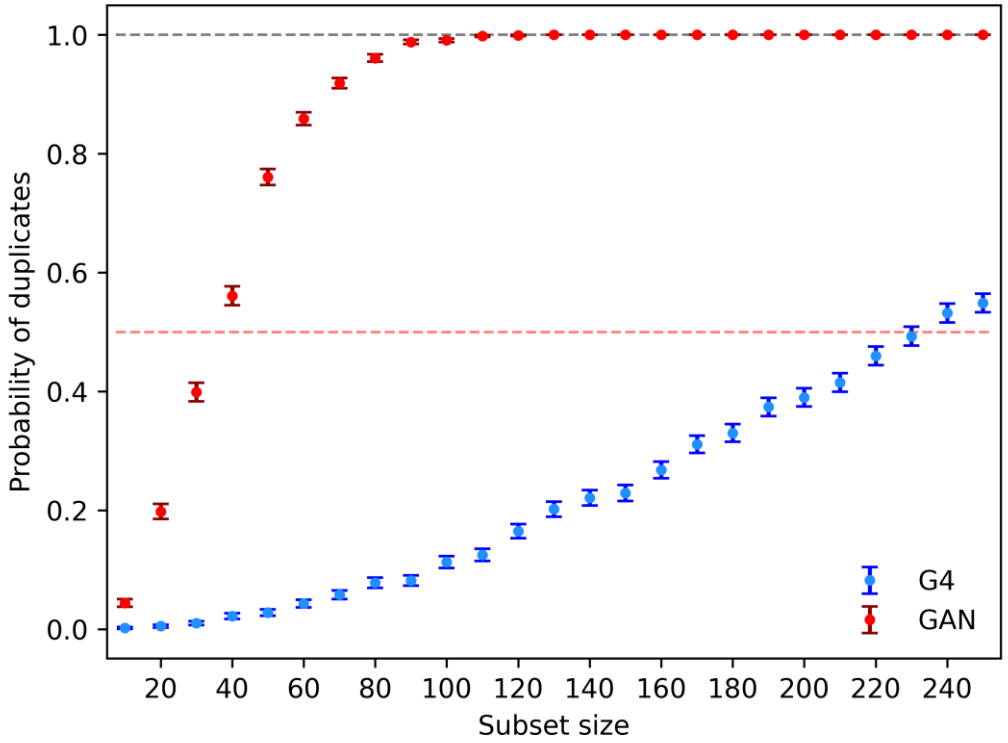
CERN openlab

Estimates of the Support Space

For a year with d days, approx. $\sqrt{d}$ people are needed.



a) 0.05-quantile threshold



b) 0.02-quantile threshold

Probability of getting at least one duplicate in a set of '*subset size*' samples.

(1 000 replications)

Where to go next?

- The current results depend strongly on the definition of the duplicates.

- What can be changed?
  - Use different metric on the energy distributions.
  - Use different features for the definition of the duplicate.
  - Use different rule for the choice of the threshold.

Thank you !
Any questions?

kristinajaruskova@gmail.com

Kristina Jaruskova