What can machine learning do for event reconstruction, fast simulations, new physics search

Olena Linnyk

21.02.2020

seriously creative

FIAS Frankfurt Institute for Advanced Studies

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

creatively serious

ML is revitalized, also in science

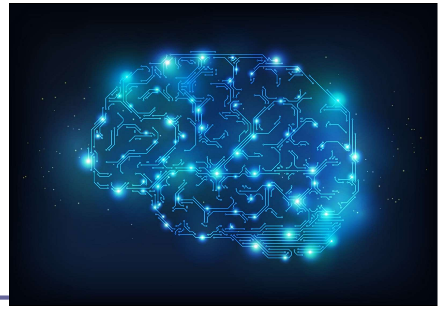
"Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we'll augment our intelligence." Ginni Rometty

OpenAI

Requests for Research 2.0

Scientific Papers using GAN

Who am I



Lecturer (PD) at the Justus Liebig University

Regulärer Kurs „KI Anwenden und Verstehen“ für Naturwissenshaftler

Researcher, Frankfurt Institute for Advanced Studies

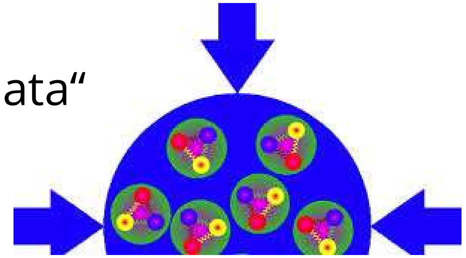
Team Leader “Deep Learning 4 Critical Point” of the NA61/SHINE Collaboration,
PI “Large Sensor Systems” (with K. Zhou, J. Steinheimer, H. Stöcker)

Artificial intelligence specialist, software company „milch&zucker“

PI “AI for competence oriented matching in HR”

Our partners:

CBM Collaboration, FAIR
Project „Erforschung von Universum und Materie ErUM - Data“



NA61/SHINE Collaboration, CERN
Project “Deep Learning 4 Critical Point”

Goethe University Frankfurt am Main
Computer Science, Physics, Law Departments

Xidian-FIAS International Joint Research Centre

Long term center founded in 2019 to research for AI in natural and life sciences

FIAS-SCNU Guangzhou and FIAS-Huzhou University planned

Justus Liebig University of Giessen
Medical Education, Psychology, Law Departments

European Commission, Brussels, Belgium
Project „ESCO“

SPONSORED BY THE

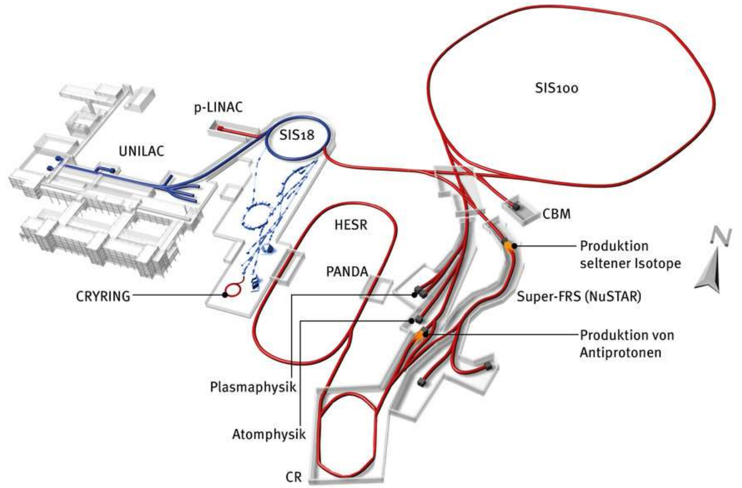
Federal Ministry of Education and Research

ErUM-Data IDT

ErUM Data Pilot project

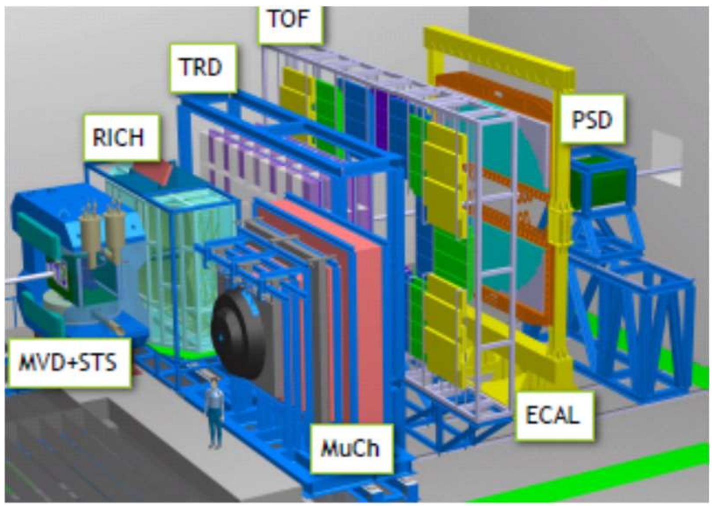
- Verbundsprojekt funded by BMBF
- Aim is to exploit modern technologies for developing **experiment overarching** solutions.
- Our area C: applications of DL
    - Specifically for CBM (our task)
- 17 partner institutes

AI for CBM



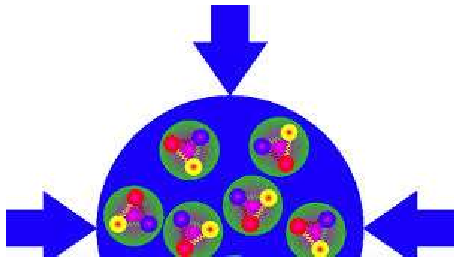
Challenges at the new FAIR collider:

- 10^5-10^7 collisions per second, high data flux
- High radiation load, aging
- Many particles/tracks per collision

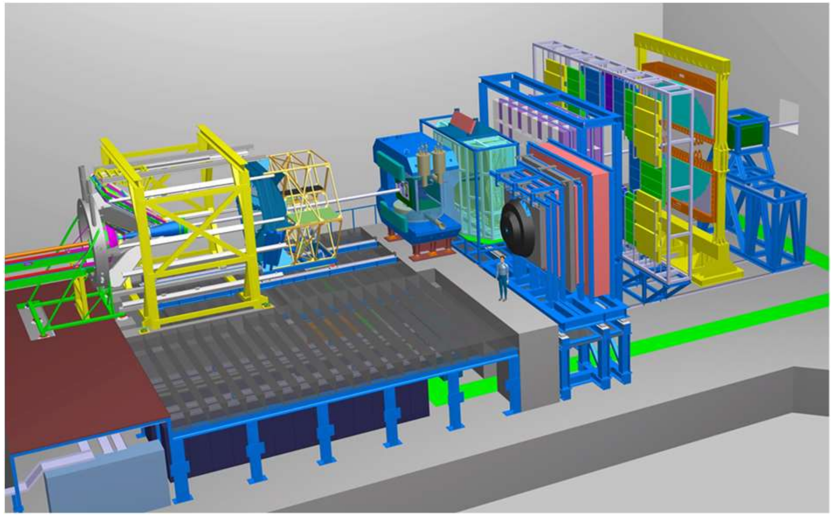


Ideas:

1. AI allows online decoding of underlying physics for the event selection, with controlled accuracy
2. AI-algorithms for frequent recalibration and quality control of detector sub-systems
3. Fast Simulations (efficiency correction!)

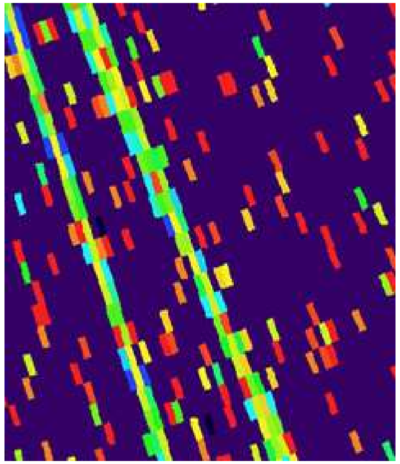


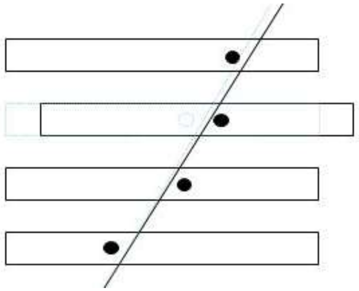
AI for CBM

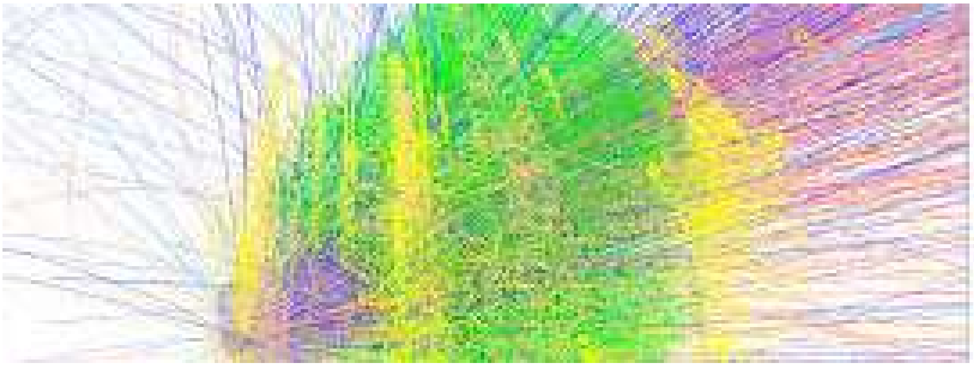


Our group contributes in 3 directions:

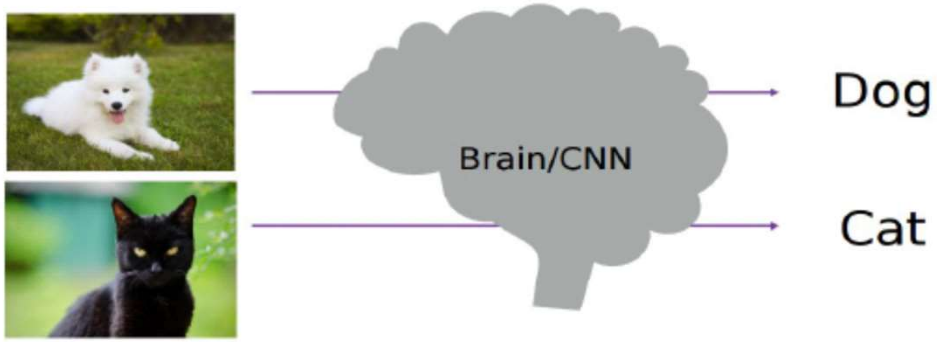
- **fast** event reconstruction (noise reduction, tracking optimization and speed)
- **fast** simulations (efficiency),
- "**physics** filters": centrality determination, new vs known physics







Classification



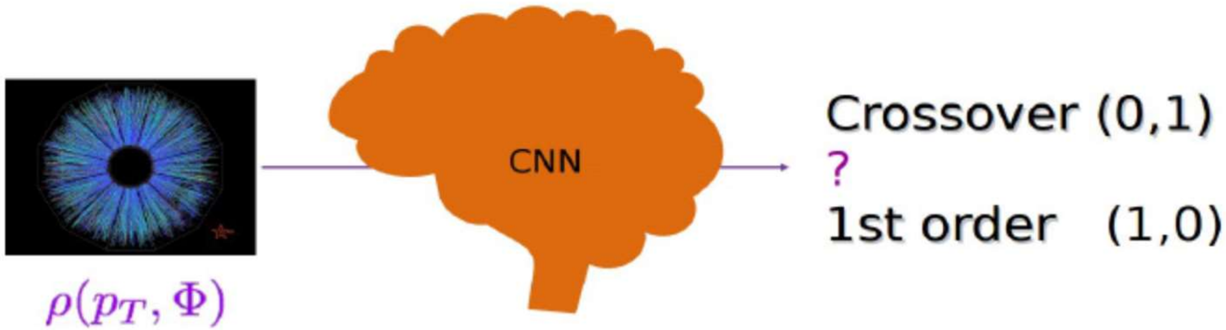
nature COMMUNICATIONS

Article | OPEN | Published: 15 January 2018

An equation-of-state-meter of quantum chromodynamics transition from deep learning

Long-Gang Pang, Kai Zhou, Nan Su, Hannah Petersen, Horst Stöcker & Xin-Nian Wang

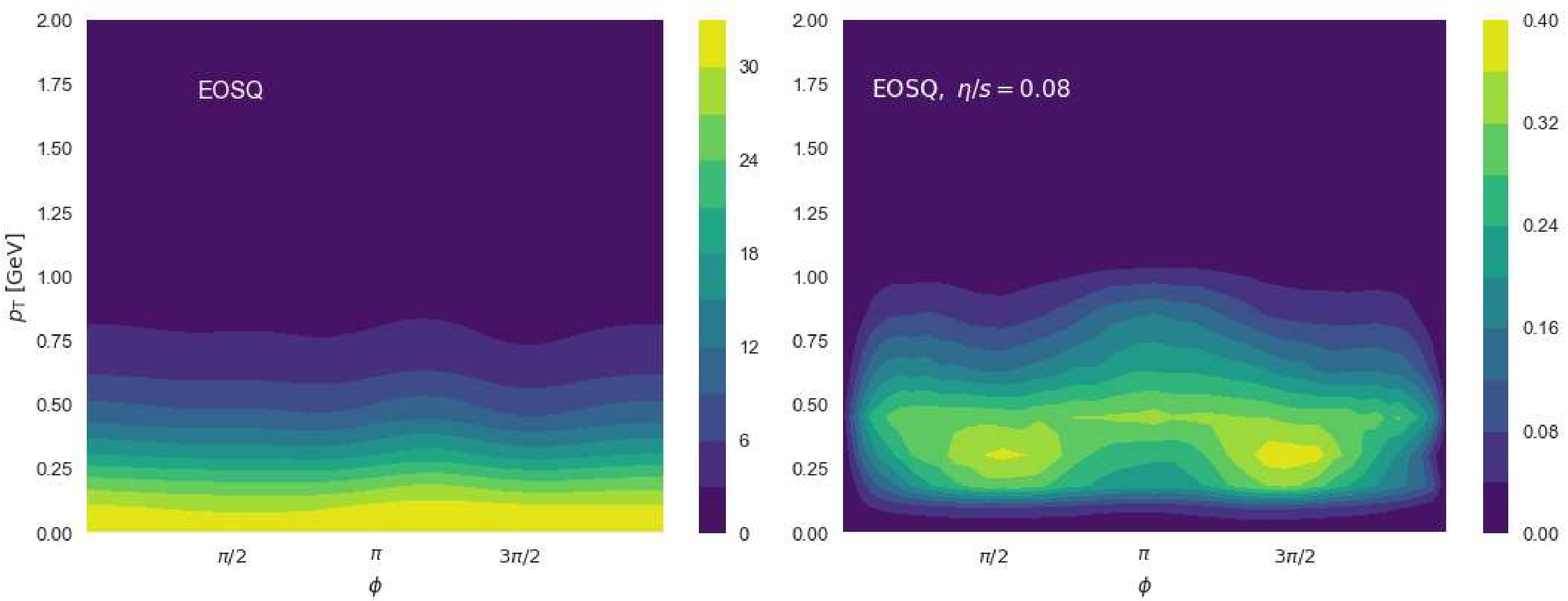
Nature Communications **9**, Article number: 210 (2018) | Download Citation



$\rho(p_T, \Phi)$

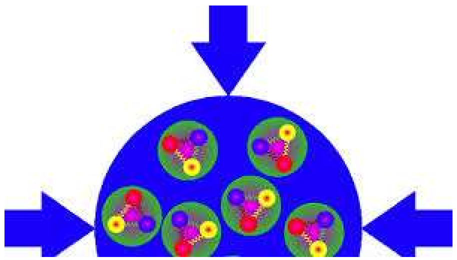
Nature Communication 9, 210(2018)

Classification, importance map



Effects of the CBM detector acceptance?

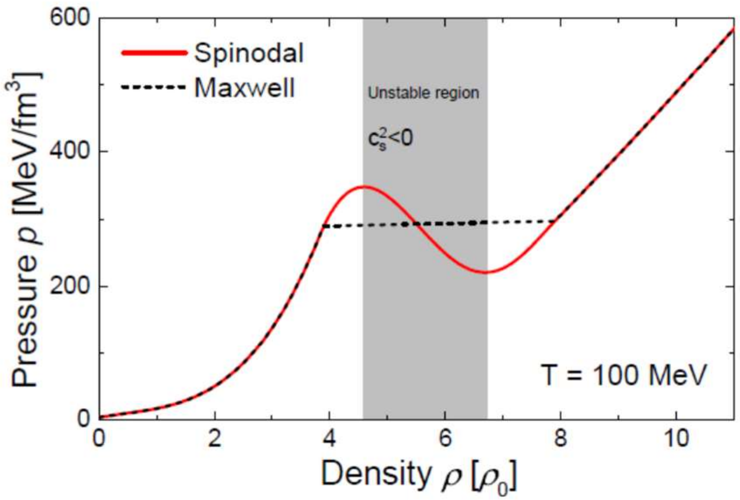
Event by event?



Equation of state event by event?

A machine learning study to identify spinodal clumping in high energy nuclear collisions

Jan Steinheimer[1], LongGang Pang[2,3], Kai Zhou[1], Volker Koch[3], Jørgen Randrup[3] and Horst Stoecker[1,4,5]
[1] *Frankfurt Institute for Advanced Studies, Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany*
[2] *Physics Department, University of California, Berkeley, CA 94720, USA*
[3] *Nuclear Science Division, Lawrence Berkeley National Laboratory, Berkeley,CA 94720, USA*
[4] *Institut für Theoretische Physik, Goethe Universität Frankfurt, D-60438 Frankfurt am Main, Germany and*
[5] *GSI Helmholtzzentrum für Schwerionenforschung GmbH*



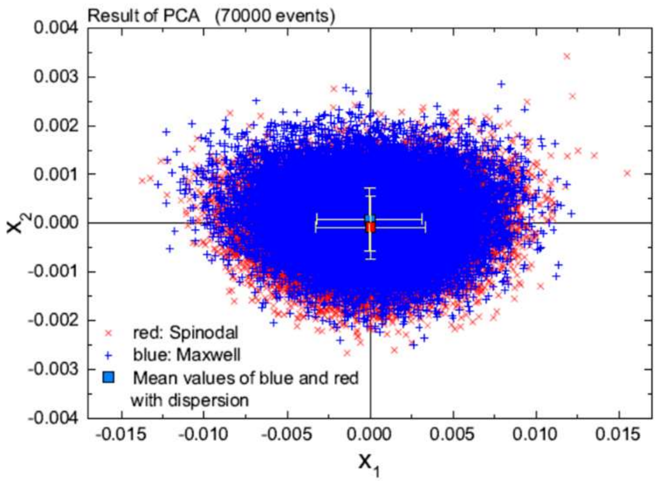


FIG. 8 (color online). Scatter plot of the first two components of a PCA of the distribution of the momentum difference for baryon pairs. The red crosses indicate the events with the spinodal instabilities, which dominate the southern hemisphere, while the blue pluses indicate events with a Maxwell construction. The large symbols with error bars near the center indicate the mean values with their dispersions. The spinodal EoS creates a clear crescent of crosses in the southern hemisphere, $x_2 < 0$. Also the mean value is shifted downwards slightly.

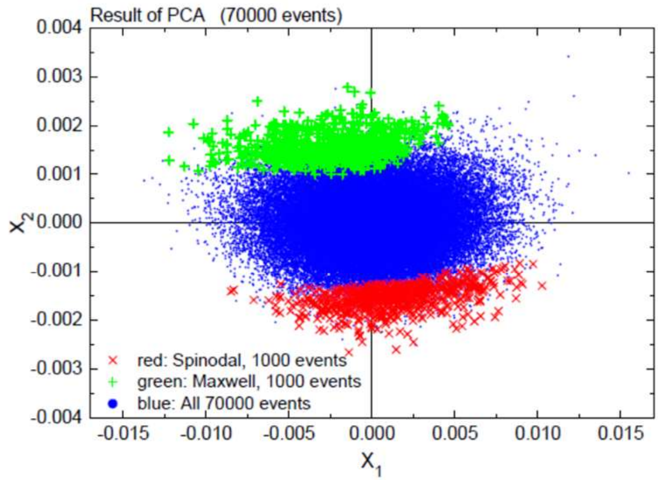


FIG. 9 (color online). Scatter plot of the first two components of a PCA of the distributions of the momentum difference for baryon pairs. The blue points correspond to all 70 000 events, Maxwell and spinodal. The red crosses correspond to those events that were identified correctly as being in the spinodal class from among those 1000 events that had the highest probability for being spinodal events, according to the neural network. Similarly, the green pluses show the correctly identified Maxwell events among the 1000 event having largest probability of belonging to that class. According

Fast simulations

- Interpolation – prediction of forward neural networks
- GAN – generative adversary networks to produce „events“
- Time series (RNN, LSTM)

FIAS results for relativistic hydro are coming – stay tuned

Data-driven Fluid Simulations using Regression Forests

Ľubor Ladický*† ETH Zurich
SoHyeon Jeong*† ETH Zurich
Barbara Solenthaler† ETH Zurich
Marc Pollefeys† ETH Zurich
Markus Gross† ETH Zurich, Disney Research Zurich

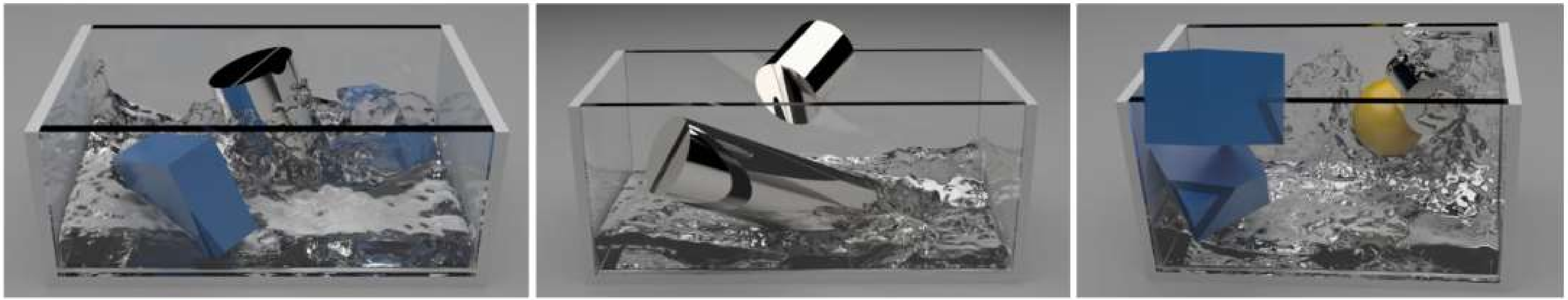
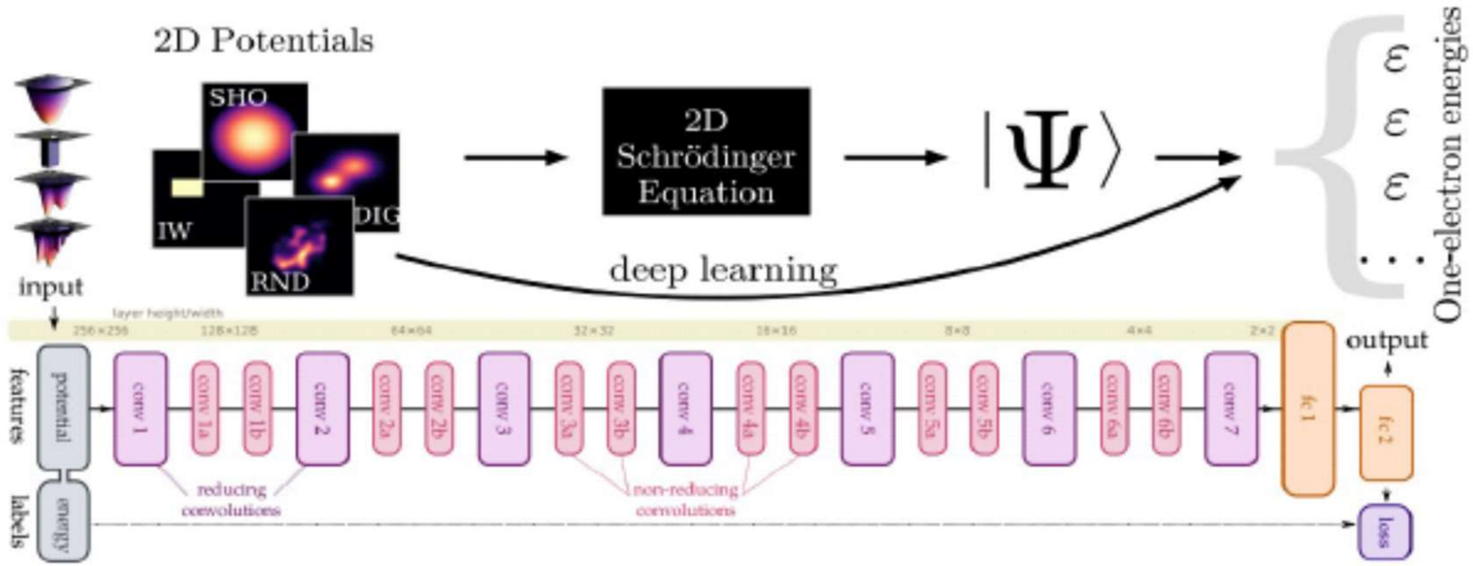
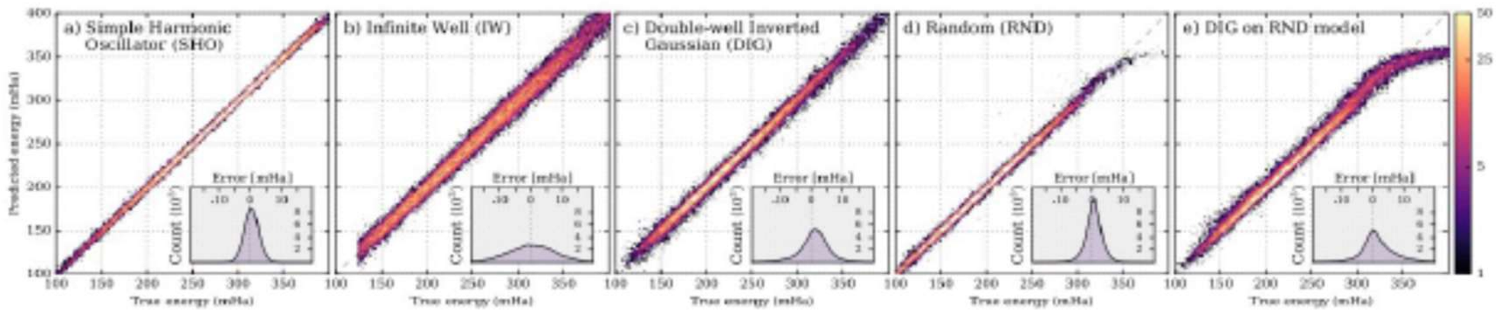


Figure 1: *The obtained results using our regression forest method, capable of simulating millions of particles in realtime. Our promising results suggest the applicability of machine learning techniques to physics-based simulations in time-critical settings, where running time matters more than the physical exactness.*

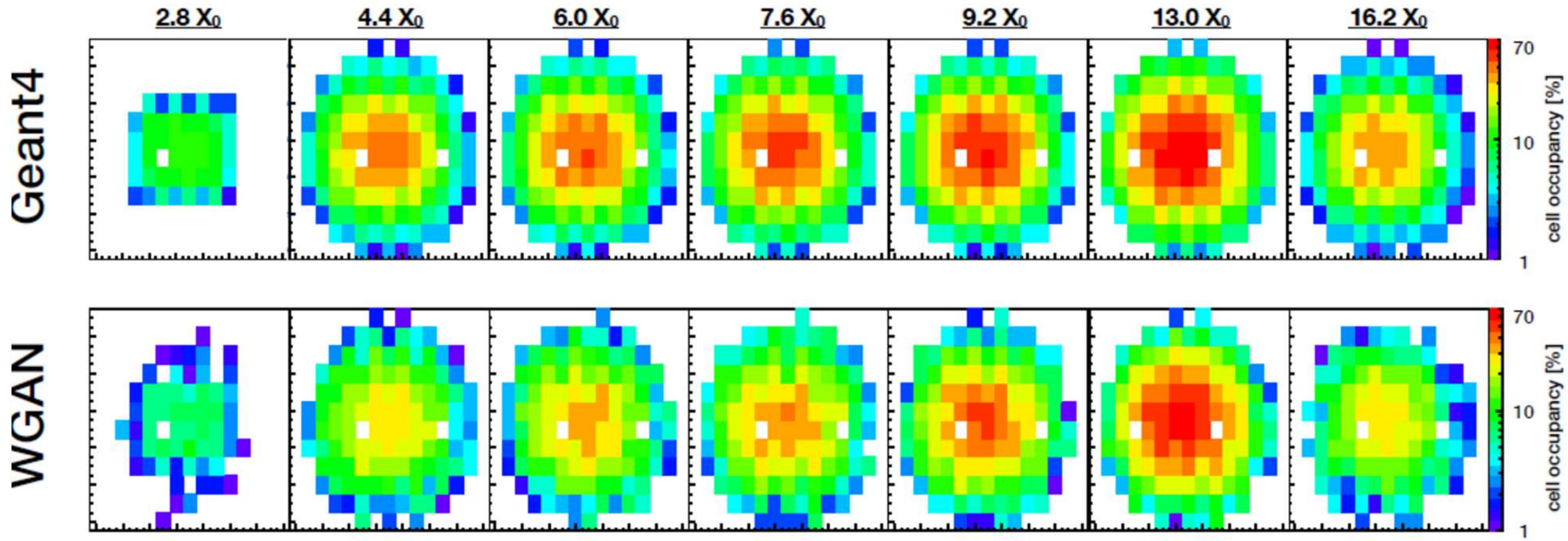
Fast solvers





Deep learning and the Schrodinger equation, by K. Mills, M. Spanner, Tamblyn (February 7, 2017)

Calorimeter response

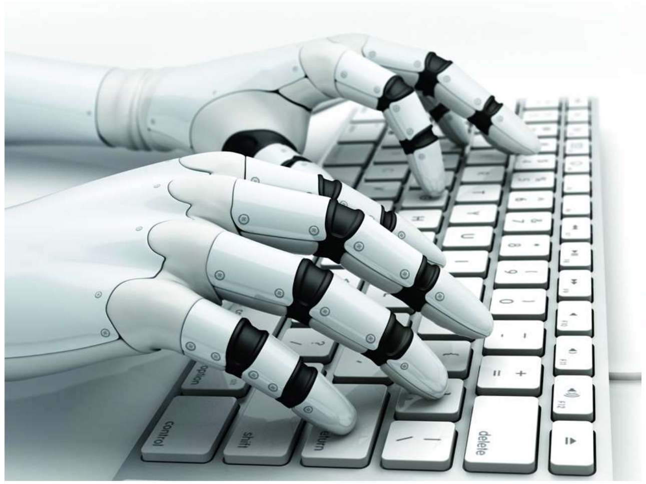


✓Radial development.

xWGAN: Overall scale slightly underestimated.

Martin Erdmann, Jonas Glombitza, Thorben Quast for CMS and ErumData
DPG 2019

Generating texts



GANS WITH REINFORCEMENT LEARNING

Chinese Poetry Generation with Planning based Neural Network

Zhe Wang[†], Wei He[‡], Hua Wu[‡], Haiyang Wu[‡], Wei Li[‡], Haifeng Wang[‡], Enhong Chen[†]
[†]University of Science and Technology of China, Hefei, China
[‡]Baidu Inc., Beijing, China
xiaose@mail.ustc.edu.cn, cheneh@ustc.edu.cn
{hewei06, wu_hua, wuhaiyang, liwei08, wanghaifeng}@baidu.com

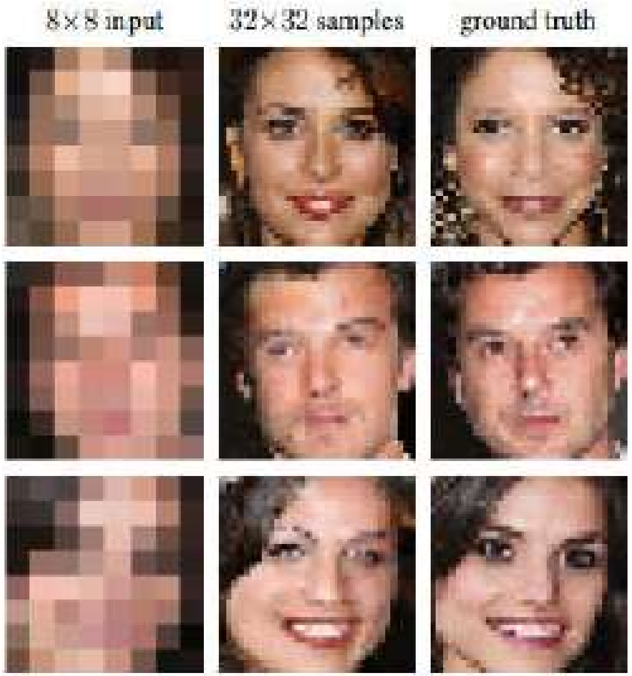
秋夕湖上	秋夕湖上
By a Lake at Autumn Sunset	By a Lake at Autumn Sunset
一夜秋凉雨湿衣，	荻花风里桂花浮，
A cold autumn rain wetted my clothes last night,	The wind blows reeds with osmanthus flying,
西窗独坐对夕晖。	恨竹生云翠欲流。
And I sit alone by the window and enjoy the sunset.	And the bamboos under clouds are so green as if to flow down.
湖波荡漾千山色，	谁拂半湖新镜面，
With mountain scenery mirrored on the rippling lake,	The misty rain ripples the smooth surface of lake,
山鸟徘徊万籁微。	飞来烟雨暮天愁。
A silence prevails over all except the hovering birds.	And I feel blue at sunset .

Table 6: A pair of poems selected from the blind test. The left one is a machine-generated poem, and the right one is written by Shaoti Ge, a poet lived in the Song Dynasty.

Generation

Image Super Resolution

Conditional generative model P(high res image | low res image)



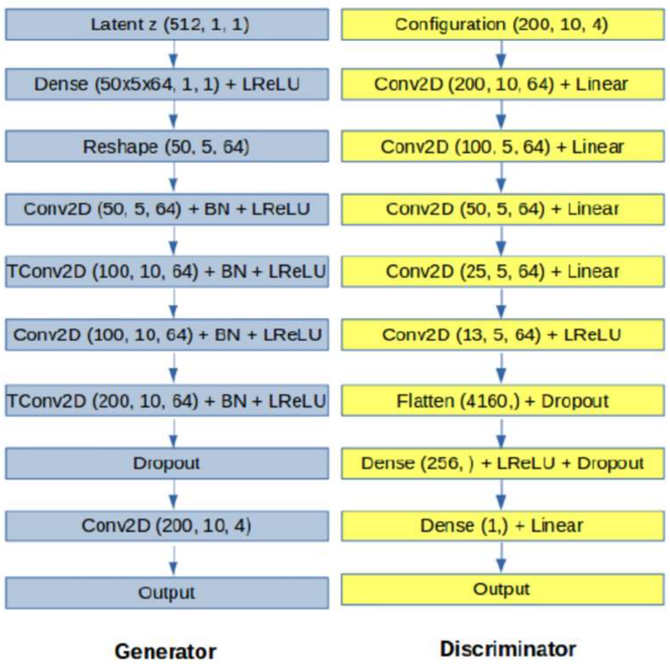
Ledig et al., 2017

GAN-generating configurations

Regressive and generative neural networks for scalar field theory

Kai Zhou,[1,2,*] Gergely Endrődi,[2] Long-Gang Pang,[1,3,4] and Horst Stöcker[1,2,5]

[1] *Frankfurt Institute for Advanced Studies, 60438 Frankfurt am Main, Germany*
[2] *Institut für Theoretische Physik, Goethe Universität, 60438 Frankfurt am Main, Germany*
[3] *Department of Physics, University of California, Berkeley, CA 94720, USA*
[4] *Nuclear Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[5] *GSI Helmholtzzentrum für Schwerionenforschung, 64291 Darmstadt, Germany*



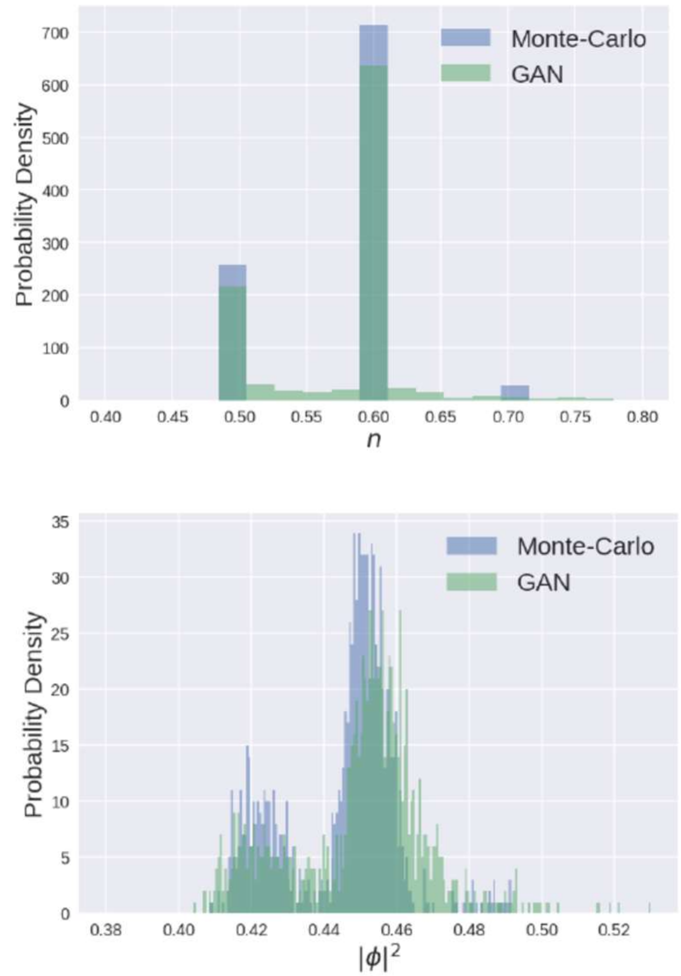
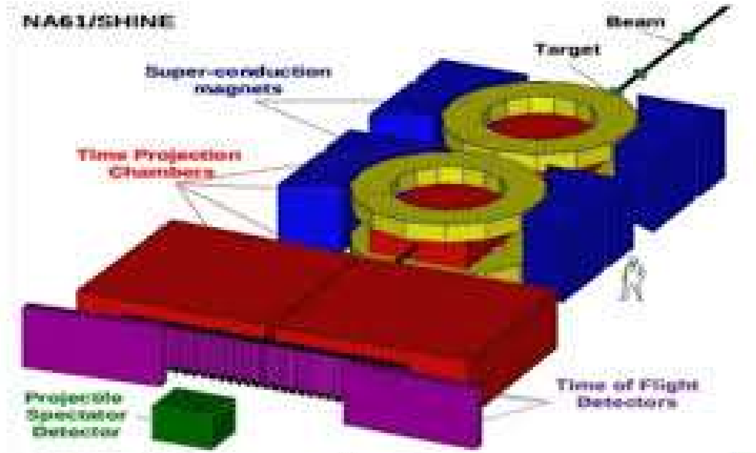
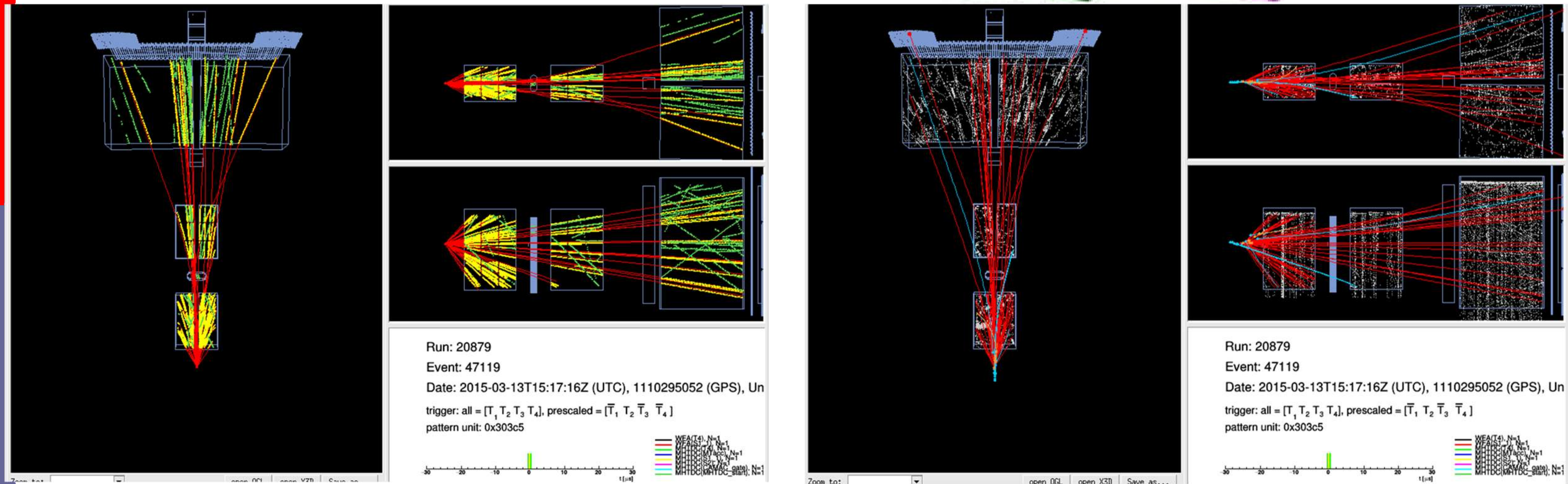


FIG. 10. The probability density distribution of the number density n (top panel) and of the squared field $|\phi|^2$ (bottom panel) from the GAN (green) along with training data distribution obtained from the Monte-Carlo simulation (blue) for fixed chemical potential $\mu = 1.05$ with 1000 samples.

Experiment NA61/SHINE at CERN

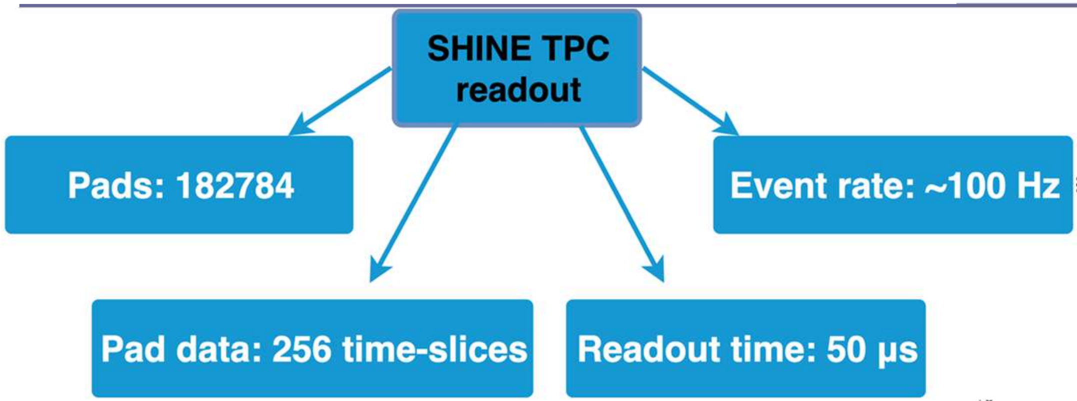


- Noisy clusters:
  - ~40-50% of clusters for Pb+Pb
  - Up to 70% for small systems



Work in progress: O. Linnyk, W. Bryliński, M. Gaździcki, N. Davis, A. Rybicki

Task at NA61



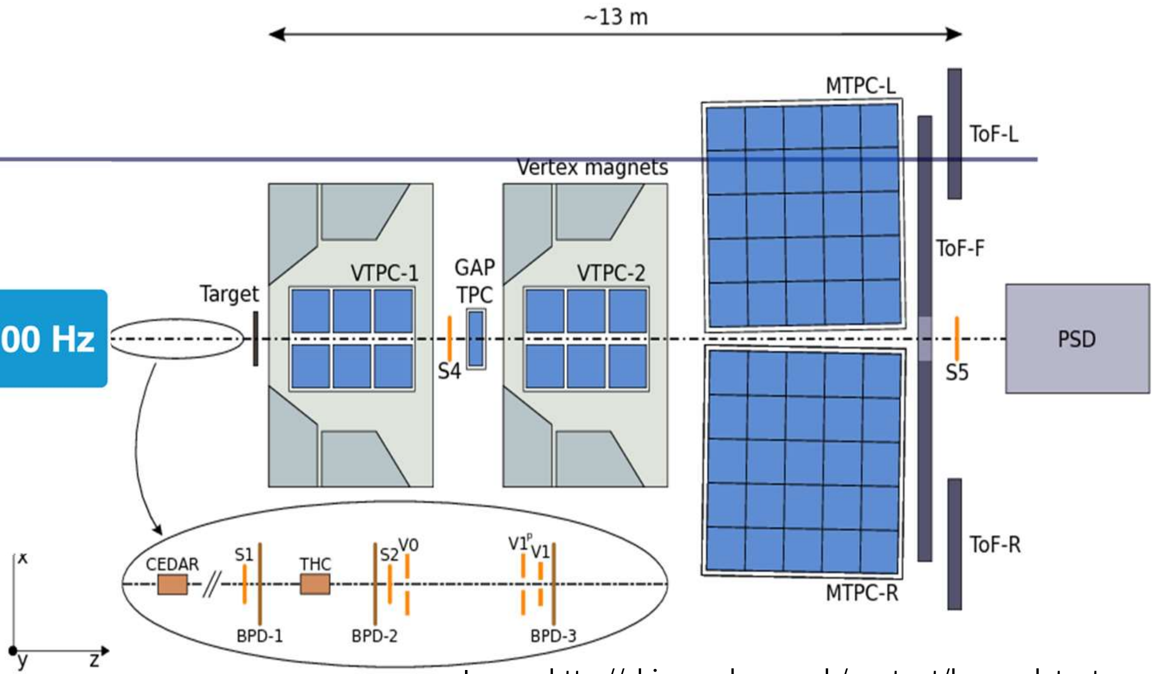


Image: http://shine.web.cern.ch/content/beam-detectors

- Noisy pad signals reduces speed and efficiency of event reconstruction. Identifying the noise
  - improves the event reconstruction
  - reduces data storage requirements
- We use machine learning techniques to classify the pad signals as recorded during the data taking*

*courtesy A.Rybicki, N.Davis for the idea to use machine learning for the reduction of noise in the TPC.

Machine learning for cluster classification

- We have tested for our task three powerful approaches:
  - Decision tree (Xgboost)
  - Convolutional neural network (ResNet)
  - Unsupervised learning (AE+K-means)

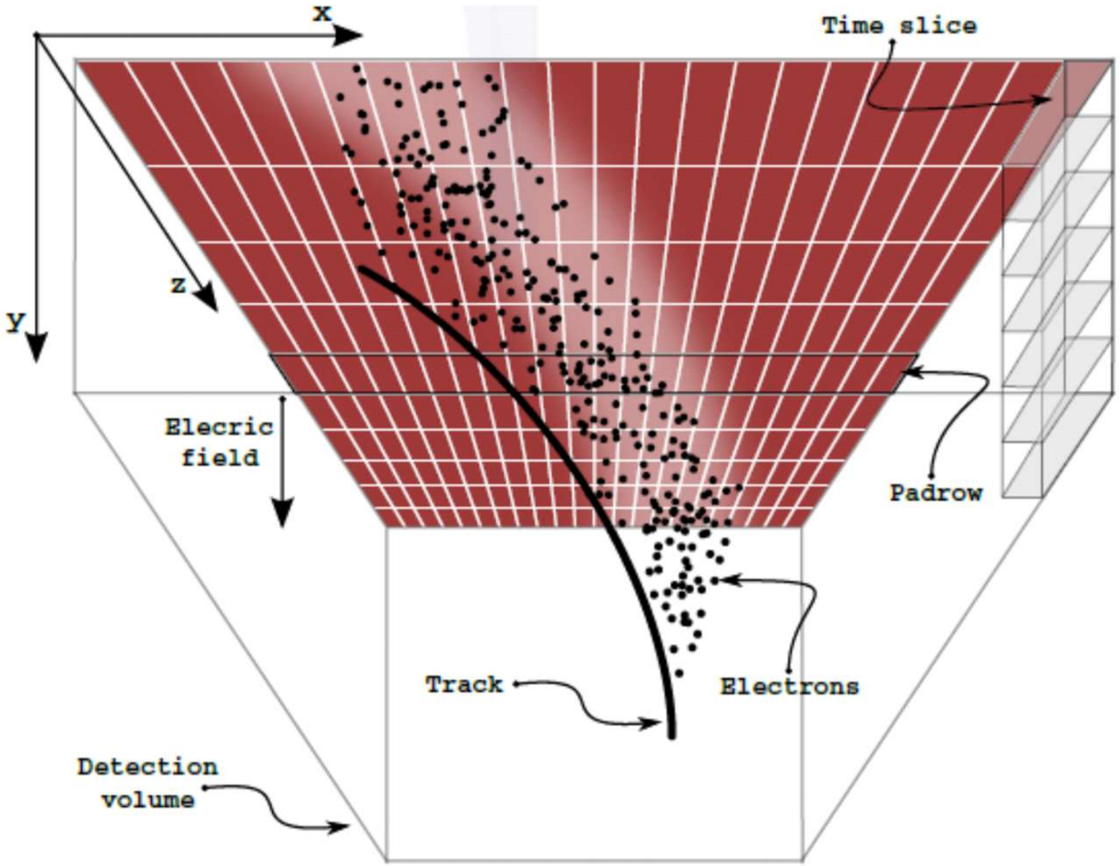
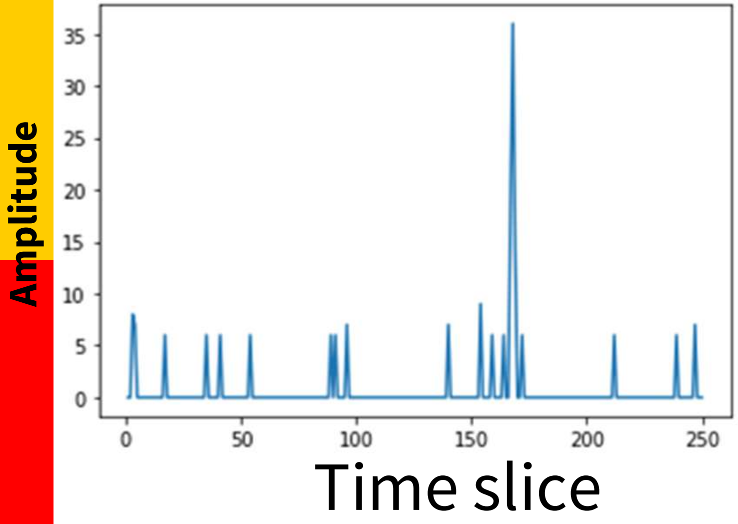
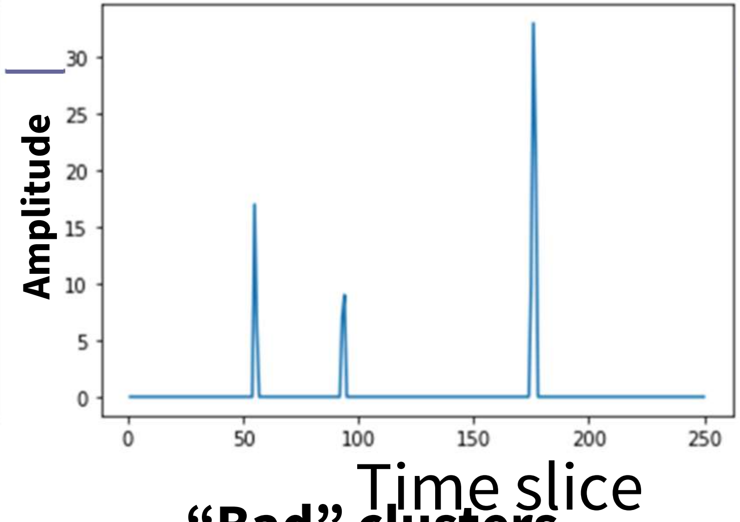


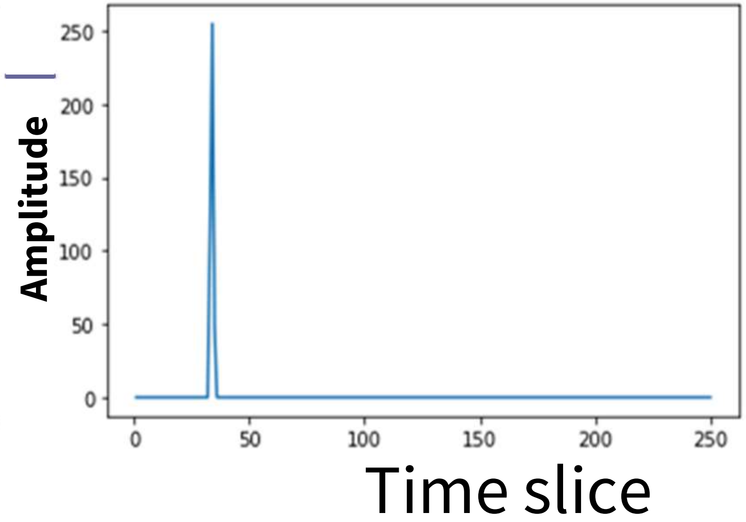
Figure 2. Simplified illustration of TPC working principle.

Sample data

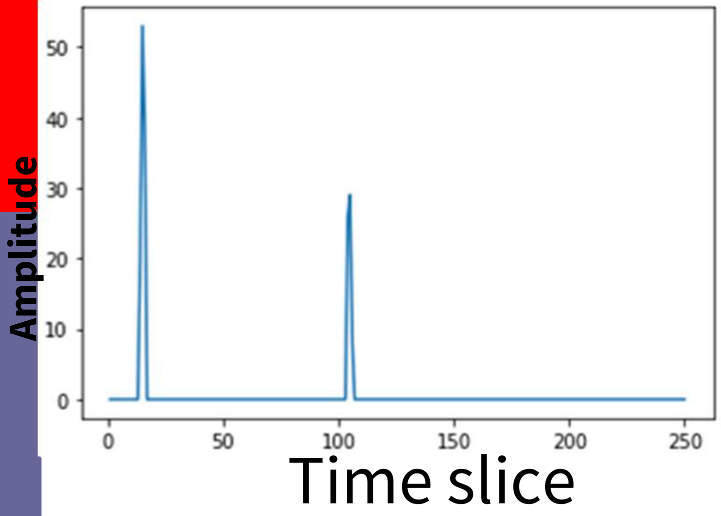
“Good” clusters

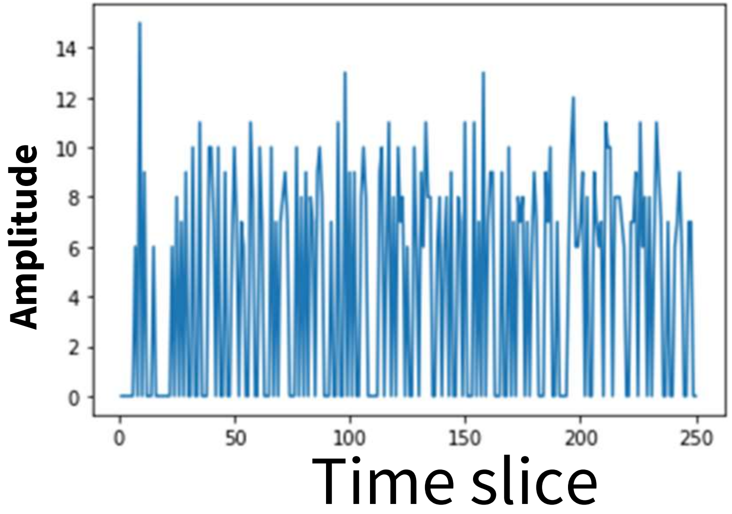


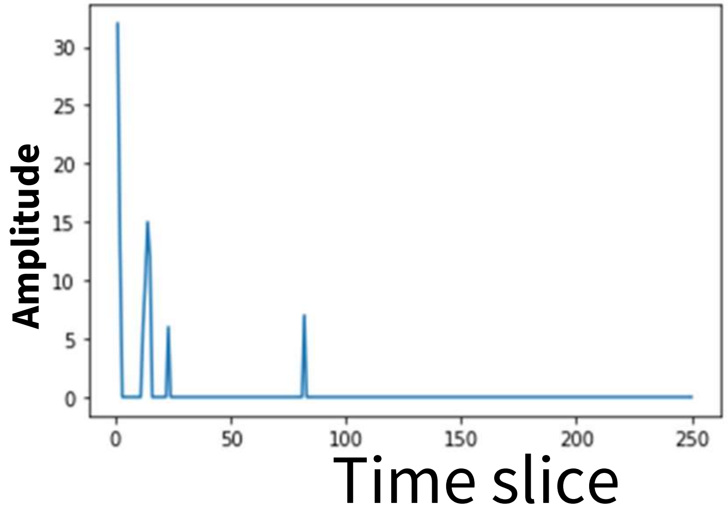




“Bad” clusters







1. XGBoost model

- Ensemble of "learned" decision trees

- Overall Accuracy : 84%
- 4% good are mistaken for bad
- 28% bad are mistaken for good

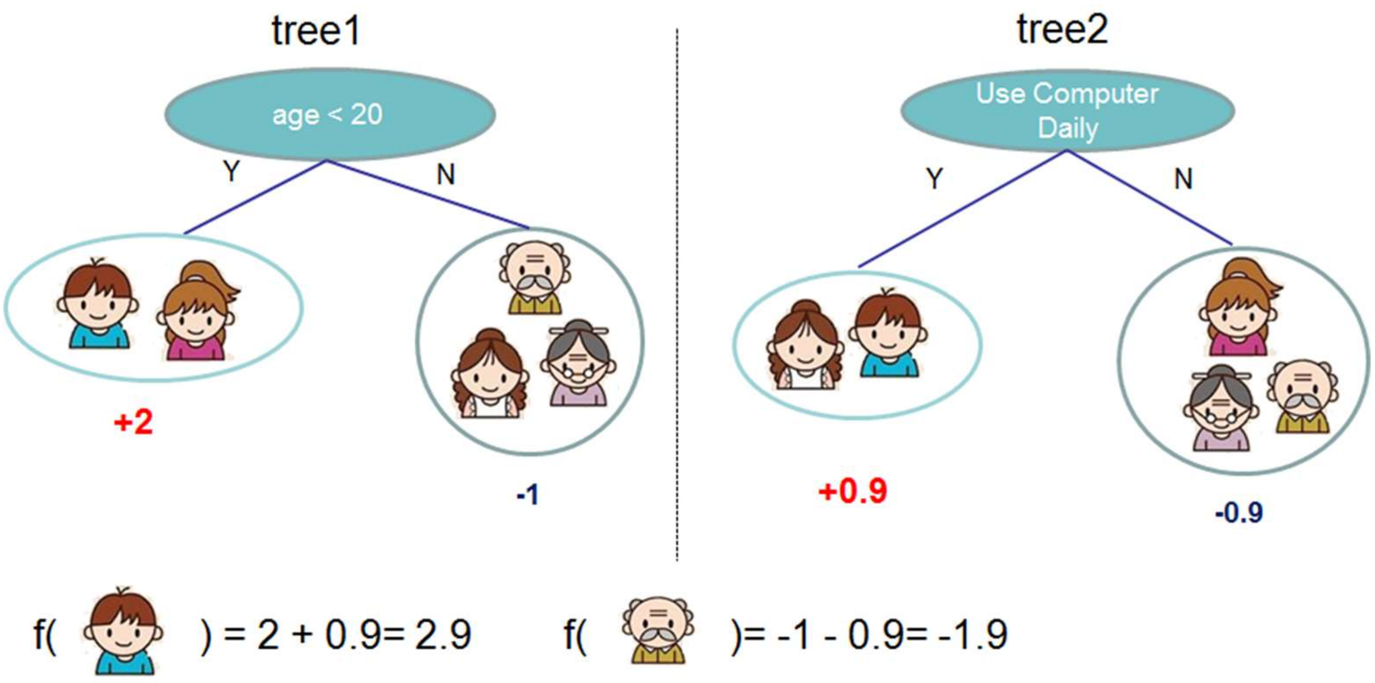


Image: https://xgboost.readthedocs.io

Can Deep learning do better?

2. Convolutional neural networks

- Convolution Neural network (CNN):
  - better feature extractor
  - Learns on this extracted features
- Resnet:
  - Variant of CNN
  - Learns better than regular CNN

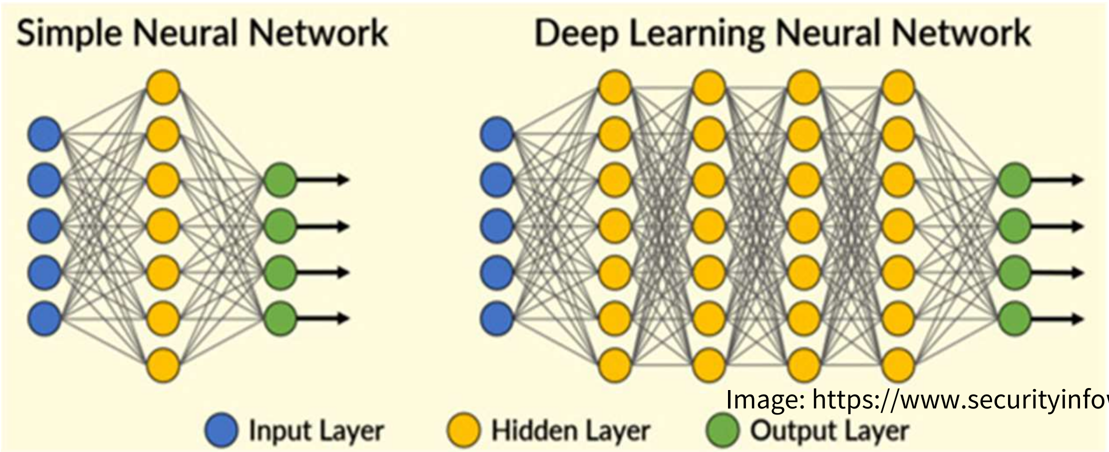
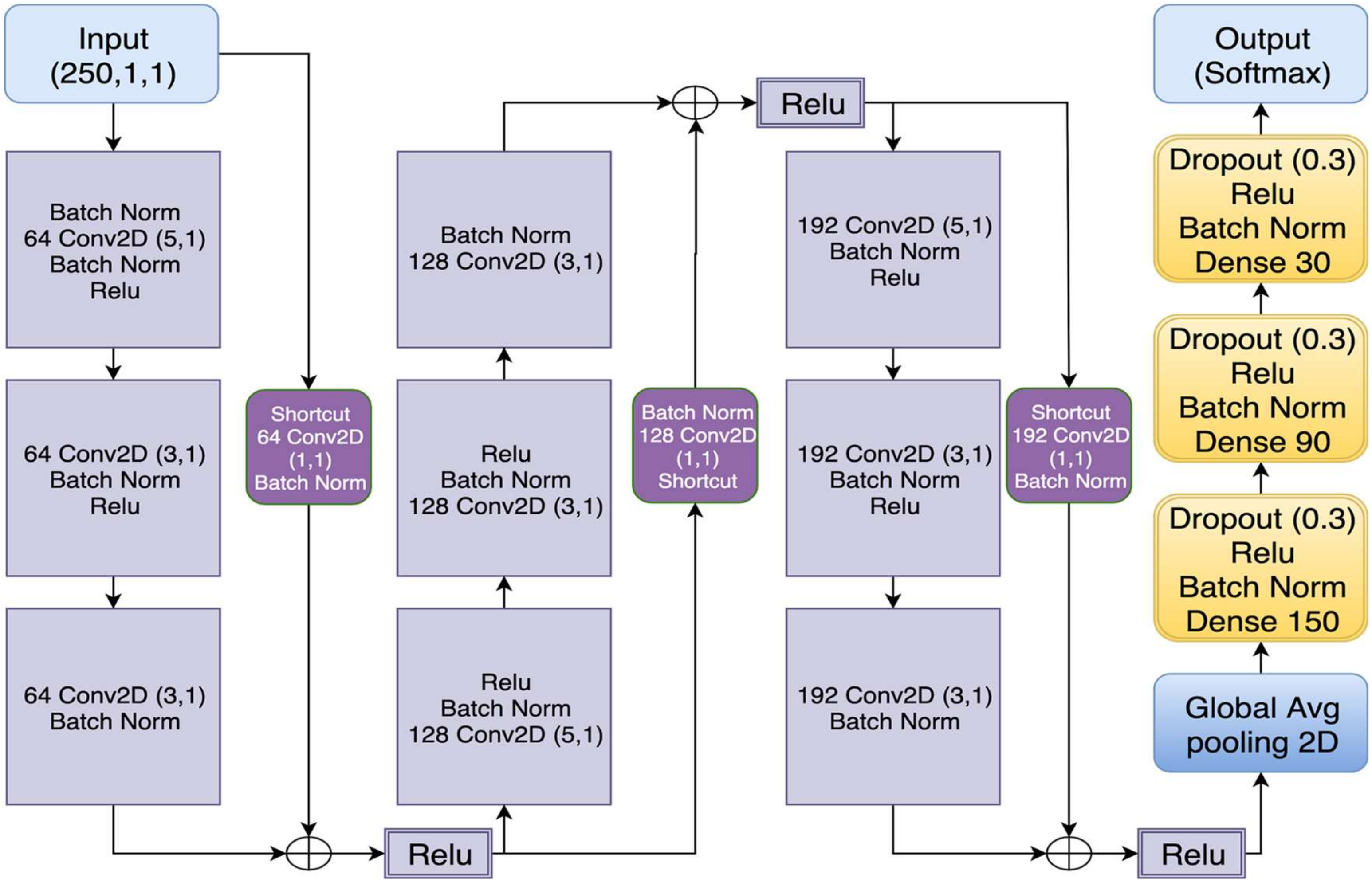
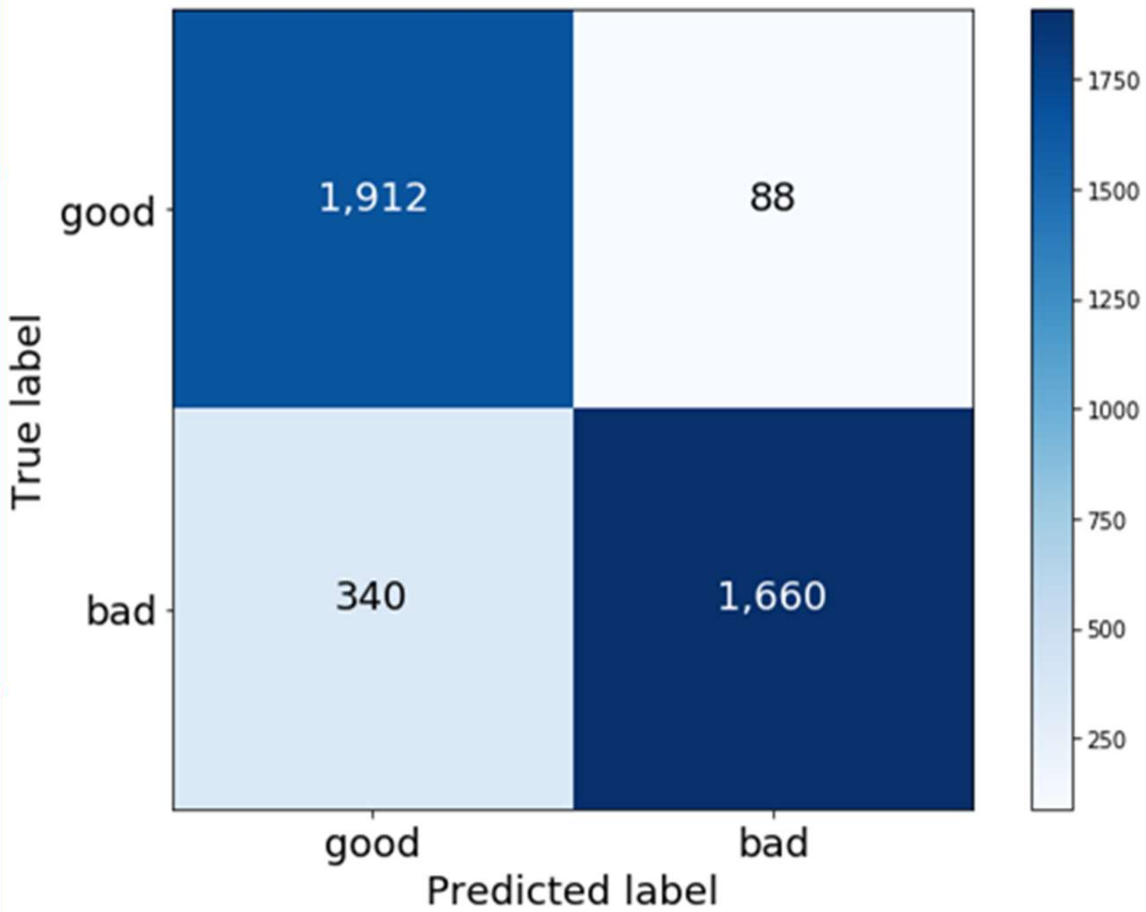


Image: https://www.securityinfowatch.com

The ResNet architecture



Results (confusion matrix)



- 89% overall accuracy
- 96% of good clusters and 83% of bad clusters correctly identified
- 4% good are mistaken for bad
- 17% bad are mistaken for good -> “safety first”
- Removing 83% of “bad clusters” at the expense of losing 4% of useful data

Unsupervised clustering:IT security, finances, physics

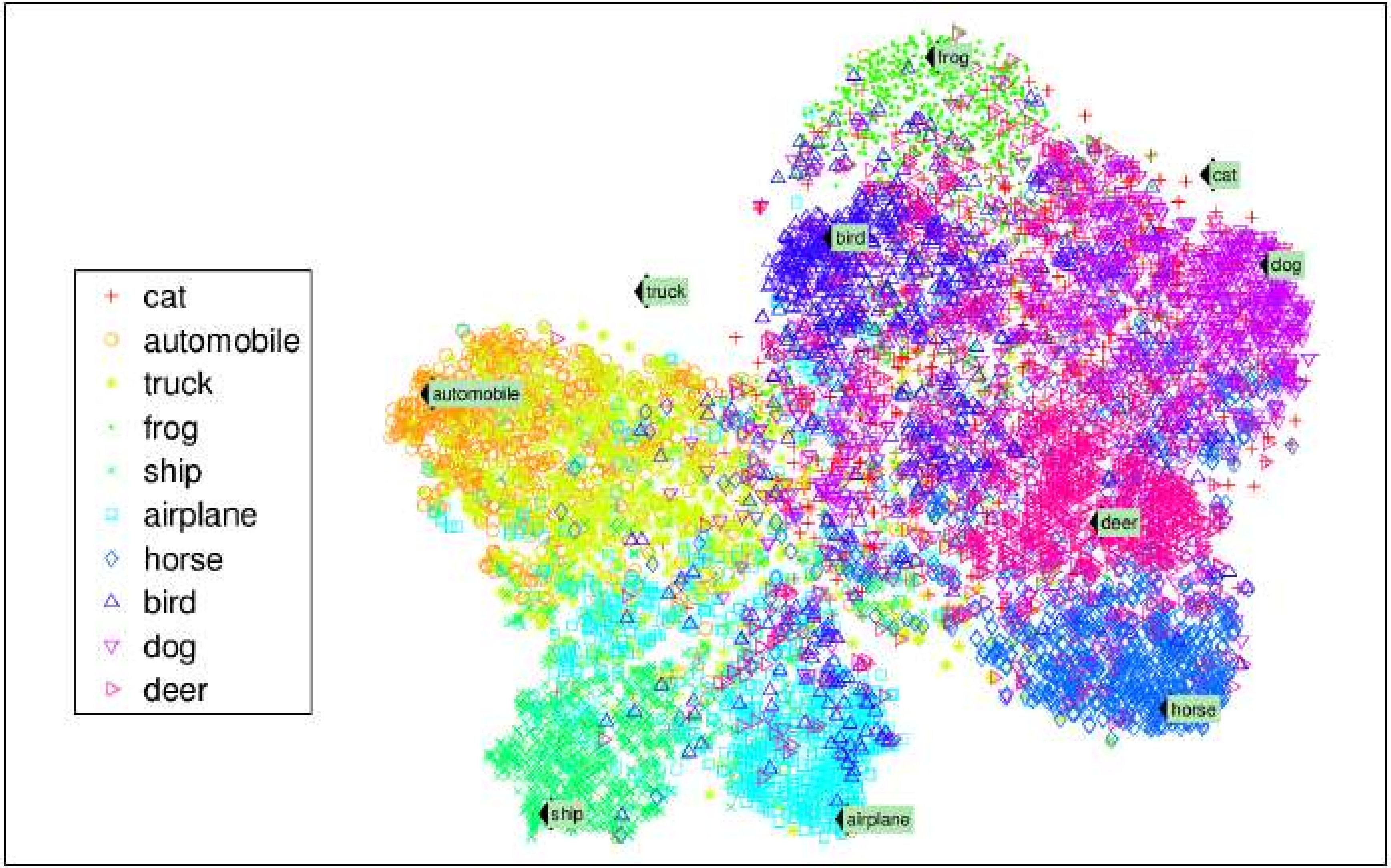
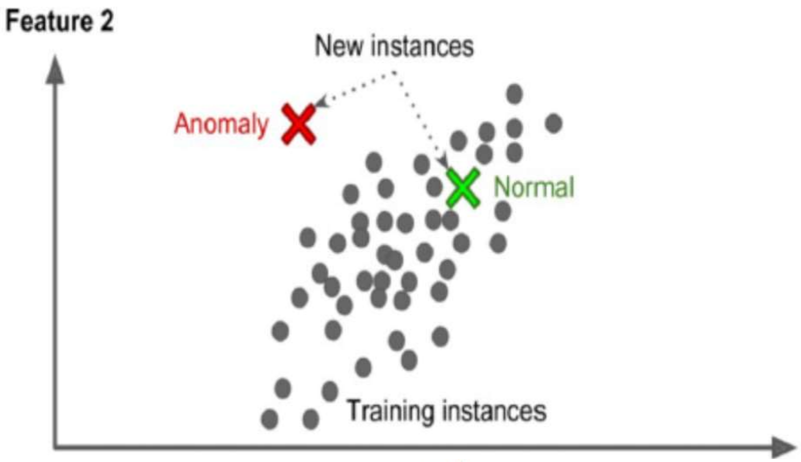


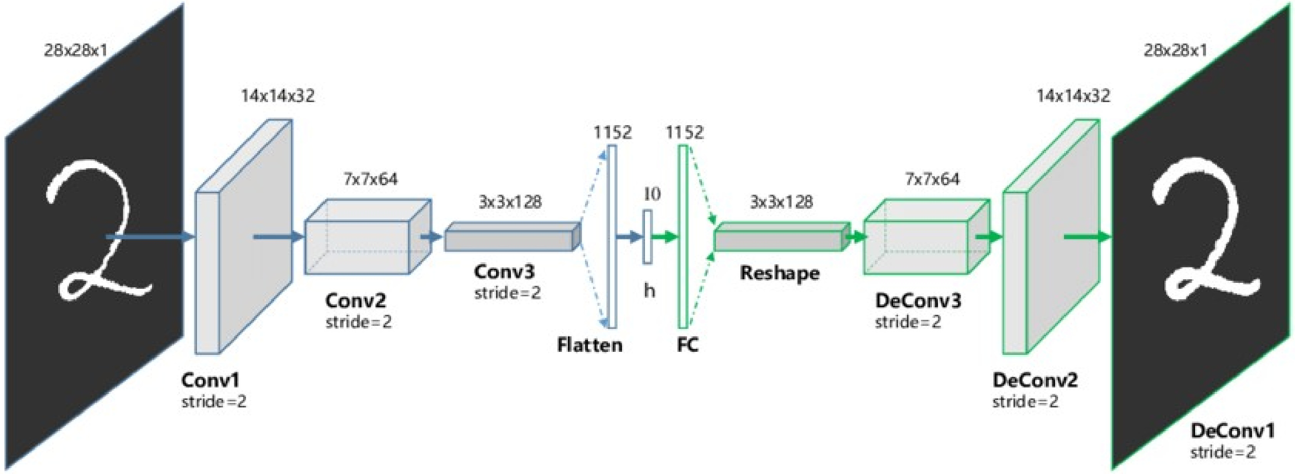
Figure 1-9. Example of a t-SNE visualization highlighting semantic clusters[3]

Anomaly detection

Anomaly detection (find out outliers)

- Autoencoder
- Density estimation
- KNN, k-means





Unsupervised learning for classification

- Autoencoder encodes the information in a low dimensional space
- Clustering this data could reveal its underlying structure

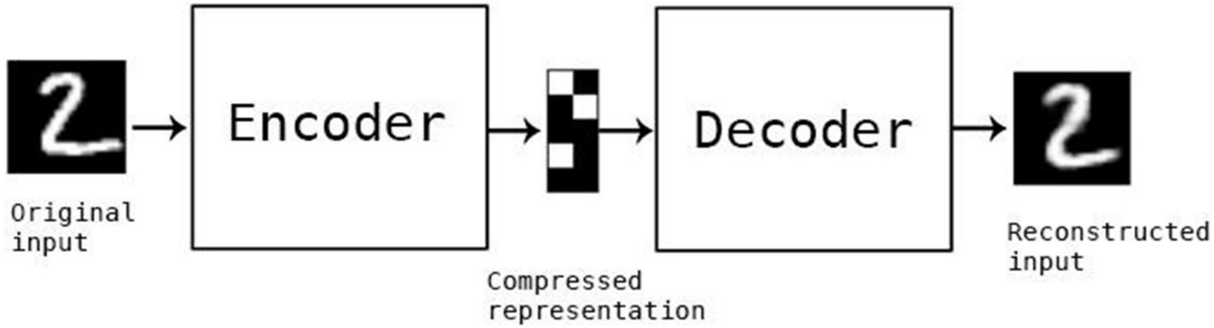
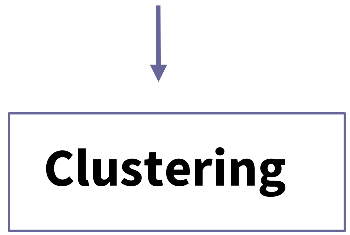


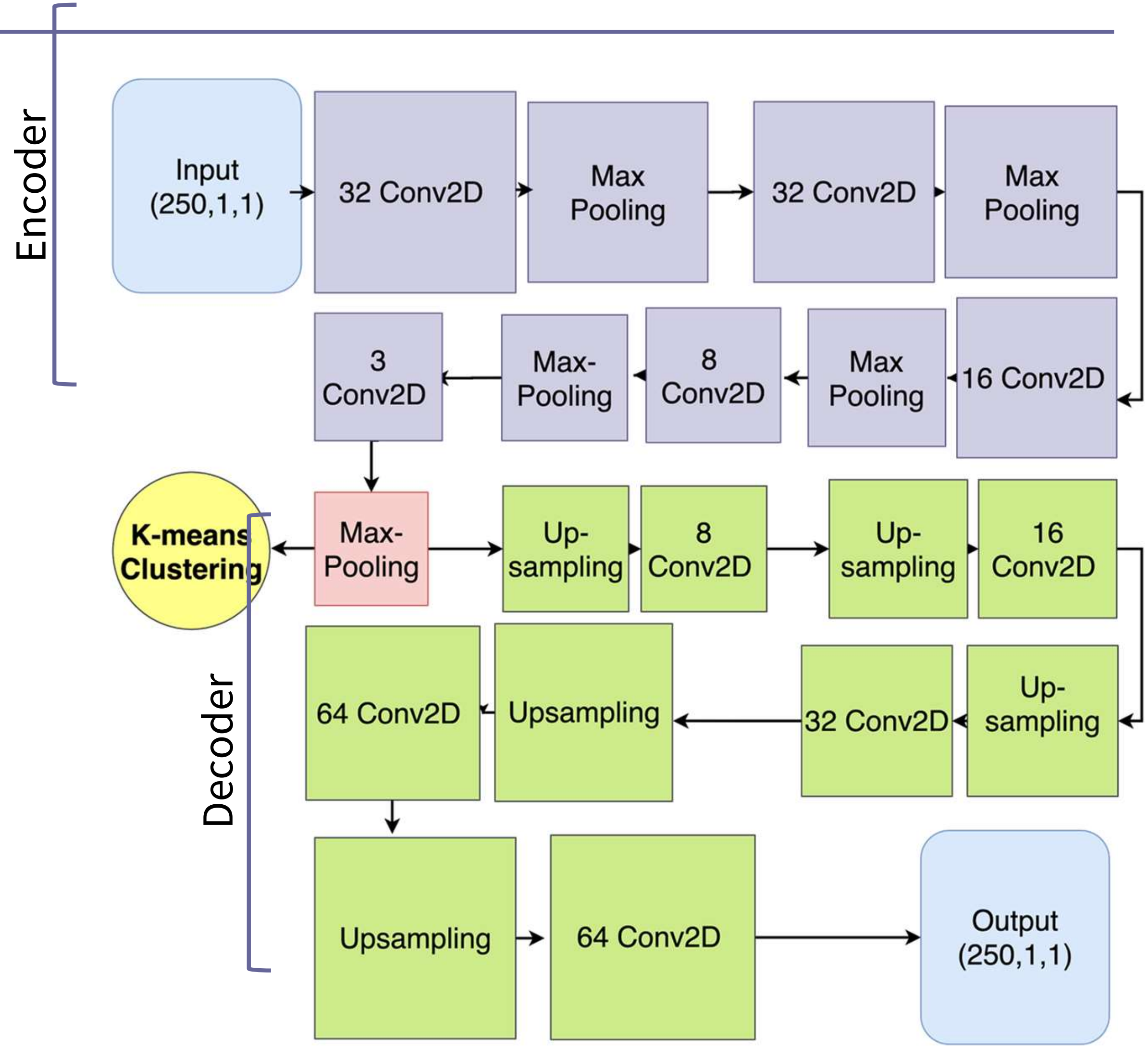
Image: https://blog.keras.io



Autoencoder + K-means clustering = Unsupervised classification

Autoencoder architecture

- The 250x1 input is projected to 3x8 dimensions by the encoder
- This encoded data is clustered via K-means clustering



Clustering: results

Case 1: No of clusters=2

Input data
Good: 2000
Bad: 2000

Cluster 1
Good:1997
Bad: 1638

Cluster 2
Good:3
Bad: 362

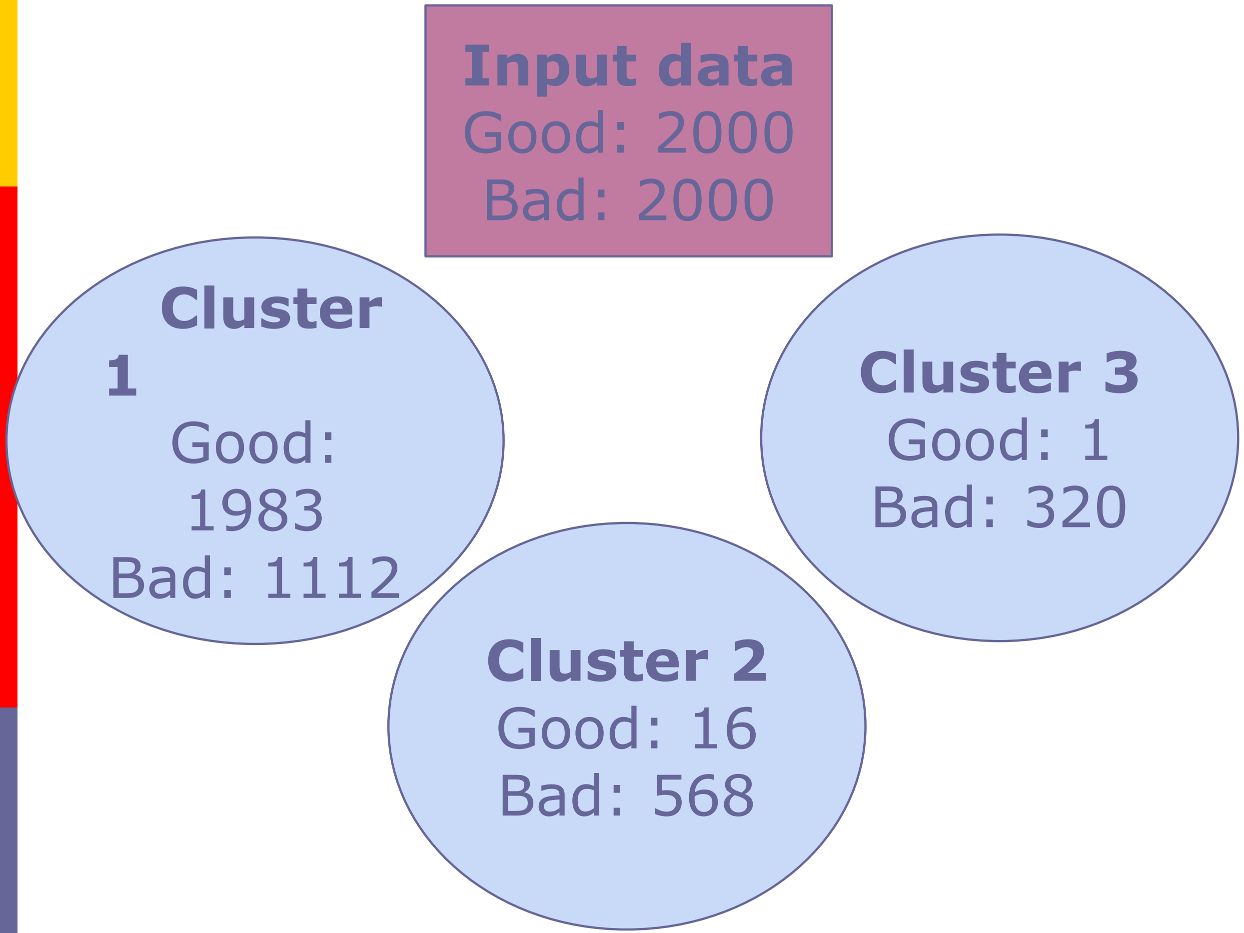
- **99.8%** “good hits” were grouped into cluster 1
- **82%** “bad hits” were also grouped to same cluster

- **18%** “bad hits” that were grouped to cluster 2 had many “obvious bad hits” which had multiple peaks with maximum amplitude

Is it safe to remove all hits which doesn’t contribute to a track?

Clustering: results

Case 2: No of clusters=3



- 99% “good hits” and 56% “bad hits were grouped together
- The remaining “bad hits” were distributed among 2 other clusters

Computation time (ResNet) & BG reduction

The study was conducted on an Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz with 8GB physical memory and Nvidia GeForce RTX 2080 Ti GPU with 10 GB graphics processing memory

- Training time: ~4 s/ epoch ~ 30 minutes
- Testing time: ~130 µs/ sample

- The ResNet model with 89% accuracy and can remove 83% of data labelled as “bad”
- The unsupervised technique groups together 99% hits contributing to a track along with 56 % of hits which didn’t contribute to any track

Summary and outlook

Signal-to-noise improvement AND online physics analysis for the heavy-ion-collision experiments (CBM, NA61/SHINE) can be done effectively and fast by ML/DL

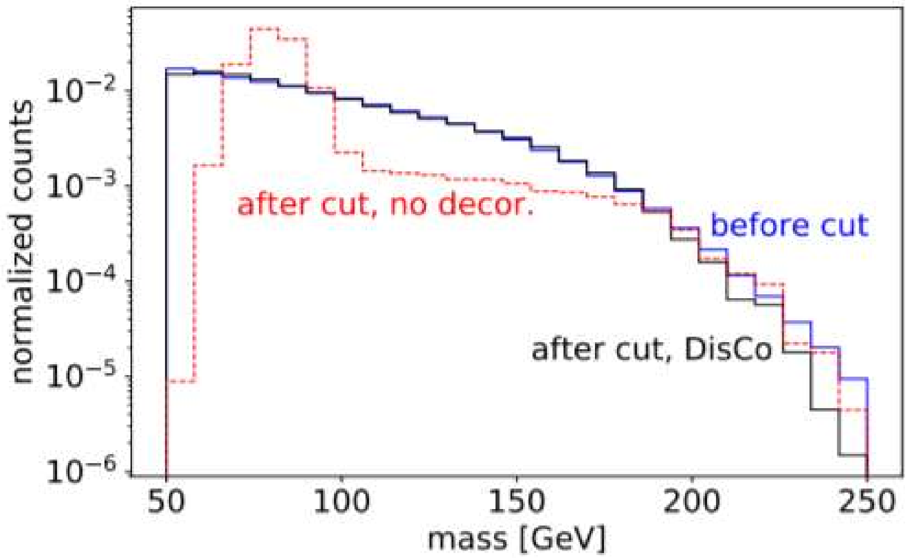
->Unsupervised clustering, fast simulations

Bias detection and removal: high energy detectors AND social science applications („AI Judge", „AI recruiter")

->De-correlation method

(DisCo Fever: Robust Networks Through Distance Correlation, by Gregor Kasieczka, David Shih, arXiv:2001.05310 [hep-ph])

Let's study this together!



Thank you for your attention!

Great thanks to:

Frankfurt: H. Stöcker, K. Zhou, J. Steinheimer, M. Omana Kuttan

Giessen: I. Teetz, J. Kohl, L. Novakovskij, D. Donges, T. Weber

CERN: N. Davis, A. Rybicki, W. Bryliński, M. Gaździcki