Machine learning applications in high energy physics

Yann Coadou

CPPM Marseille

FunPhys visit
CPPM, 6 October 2023

Machine learning: how to learn?

- "Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed" *attributed to Arthur Samuel (1959)*

Machine learning: how to learn?

- "Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed" *attributed to Arthur Samuel (1959)*

Supervised learning

- Labelled training events with feature variables and class labels

Machine learning: how to learn?

- "Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed" *attributed to Arthur Samuel (1959)*

Supervised learning

- Labelled training events with feature variables and class labels

Reinforcement learning

- Instead of labels, some sort of reward system (e.g. game score)
- Goal: maximise future payoff by optimising decision policy

Machine learning: how to learn?

- "Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed" *attributed to Arthur Samuel (1959)*

Supervised learning

- Labelled training events with feature variables and class labels

Reinforcement learning

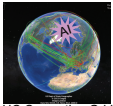
- Instead of labels, some sort of reward system (e.g. game score)
- Goal: maximise future payoff by optimising decision policy

Unsupervised learning

- Find similarities in training sample, without predefined categories (no labels)
- Discover good internal representation of the input
- Not biased by pre-determined classes $\Rightarrow$ may discover unexpected features!

Machine learning and particle physics

Role of AI: accelerator control, data acquisition, event triggering, anomaly detection, new physics scouting, event reconstruction, event generation, detector simulation, LHC grid control, analytics, signal extraction, likelihood free inference, background rejection, new physics searches, ...

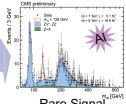


LHC Computing Grid
200k cores pledge to
CMS over ~100 sites

CERN Tier-0/Tier-1
Tape Storage
200PB total

CERN Tier-0
Computing Center
20k cores

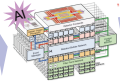
LHC Grid
Remote Access
to 100PB of data



Rare Signal
Measurement
~1 out of 10^6

Large Hadron Collider
40 MHz of collision

CMS Detector
1PB/s



CMS L1 & High-
Level Triggers
50k cores, 1kHz

©J.-R. Vlimant

▸ machine learning or deep learning or multivariate in InspireHEP

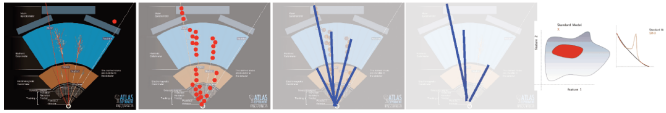
Up-to-date review of papers

▸ https://github.com/iml-wg/HEPML-LivingReview

Machine learning and particle physics

- Reduce data dimensionality to allow analysis

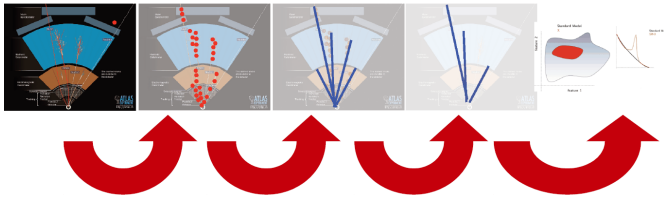
Raw	Sparsified	Reco	Select	Physics	Ana
1e7	1e4	100-ish*	50	10	1



- Losing information at each simplification step

- Reduce data dimensionality to allow analysis

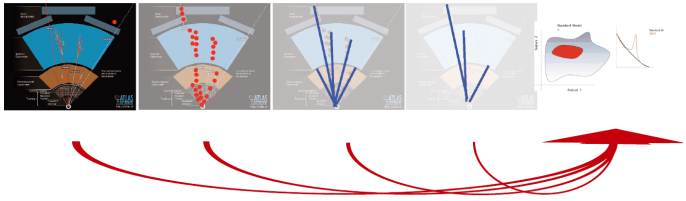
Raw	Sparsified	Reco	Select	Physics	Ana
1e7	1e4	100-ish*	50	10	1



- Losing information at each simplification step
- Improve each step with ML?

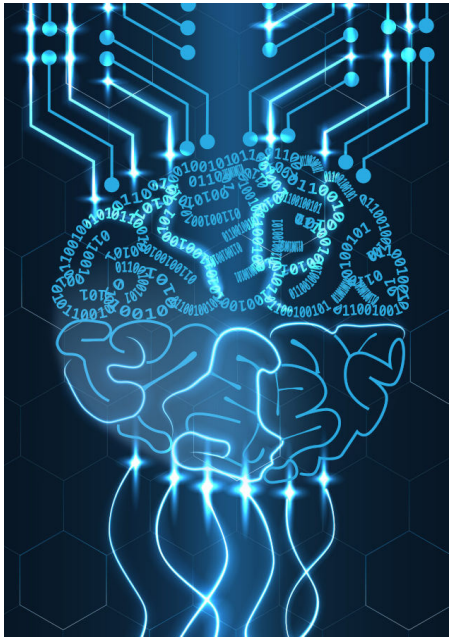
- Reduce data dimensionality to allow analysis

Raw	Sparsified	Reco	Select	Physics	Ana
1e7	1e4	100-ish*	50	10	1



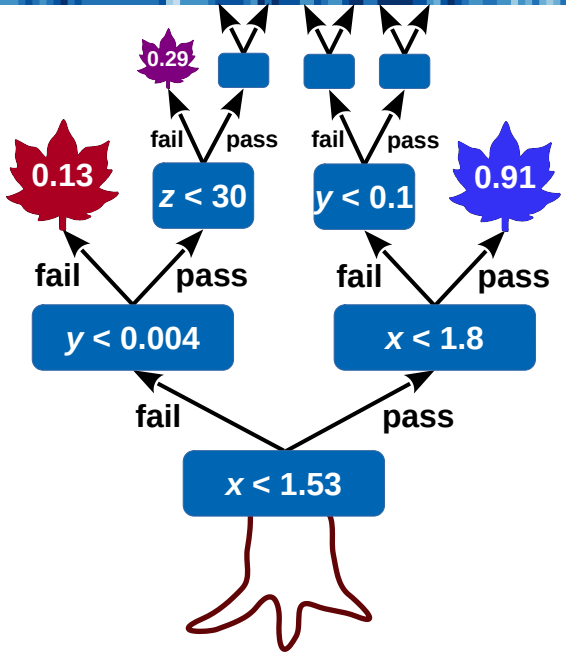
- Losing information at each simplification step
- Improve each step with ML?
- Skip one or more steps with ML?

Machine learning algorithms

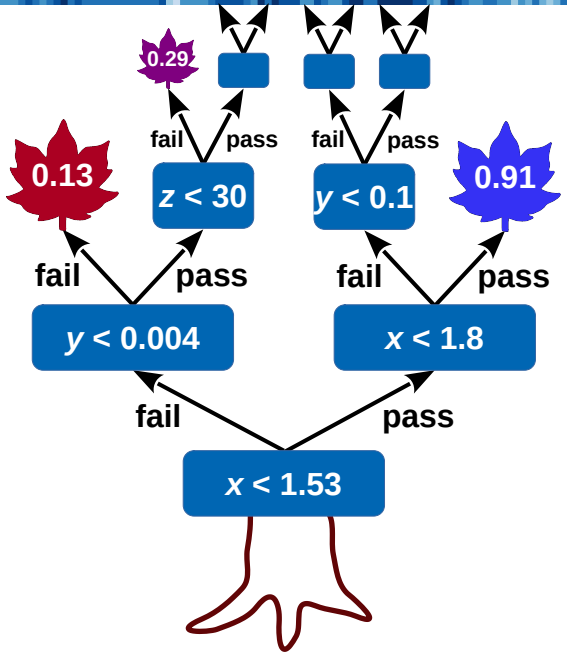


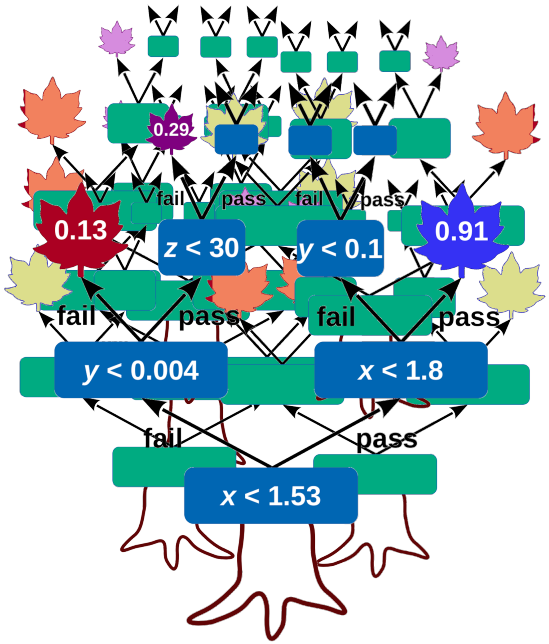
- Decision trees
- Boosted decision trees
- Support vector machines (not today)
- Neural networks
- Deep neural networks
  - convolutional NN
  - recurrent NN (see Georges' slides)
  - autoencoders (not today)
  - graph NN
  - generative models
  - ...

(Boosted) Decision trees



(Boosted) Decision trees



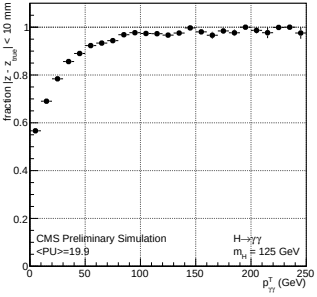


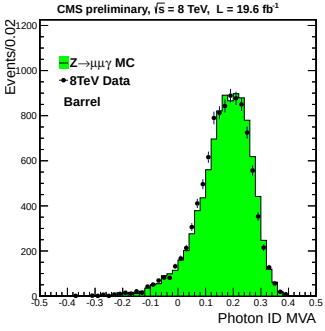
BDT in HEP: CMS $H \to \gamma\gamma$ result

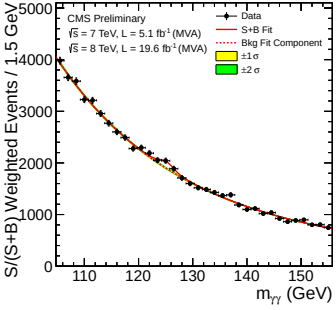
CMS-PAS-HIG-13-001

Hard to use more BDT in an analysis:

- vertex selected with BDT
- 2nd vertex BDT to estimate probability to be within 1cm of interaction point
- photon ID with BDT
- photon energy corrected with BDT regression
- event-by-event energy uncertainty from another BDT
- several BDT to extract signal in different categories





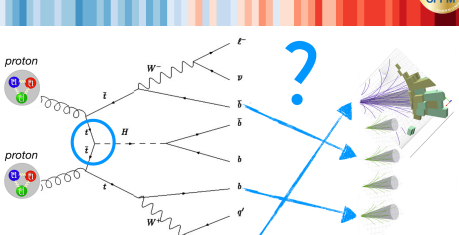


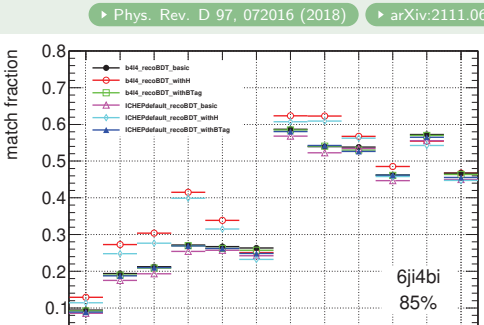
BDT in HEP: reducing combinatorics

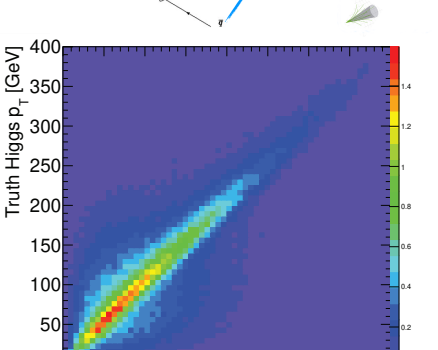
$t\bar{t}H(b\bar{b})$ reconstruction

- Match jets and partons in high-multiplicity final state
- BDT trained on all combinations
- New inputs to classification BDT
- Access to Higgs p_T, origin of b-jets

▸ Phys. Rev. D 97, 072016 (2018) ▸ arXiv:2111.06712 [hep-ex]

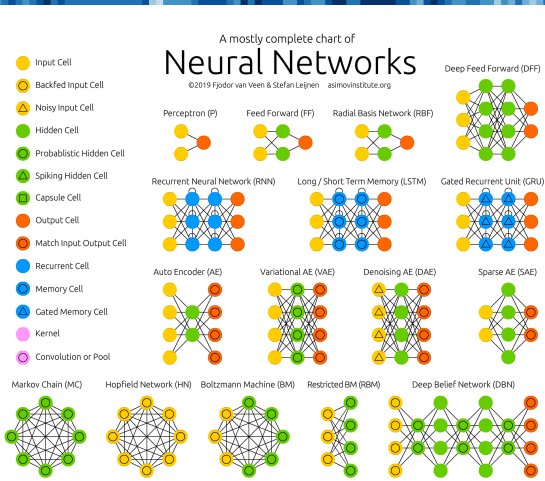


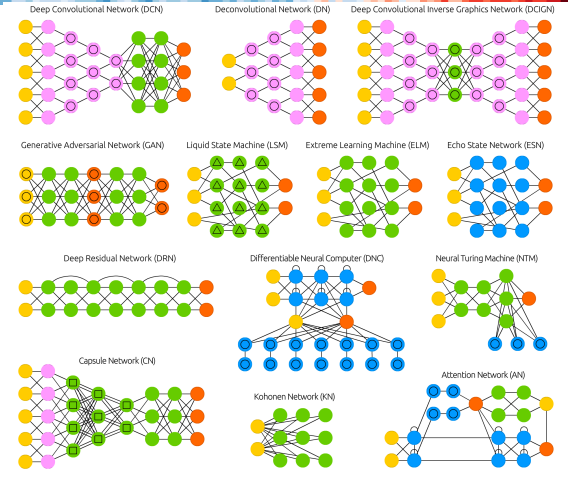




▸ thesis

Neural network zoo



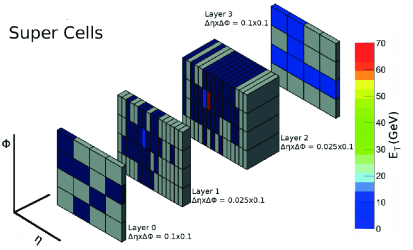


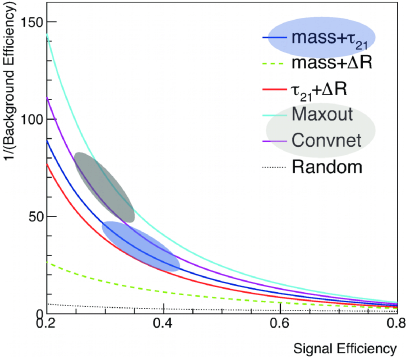
▸ https://www.asimovinstitute.org/

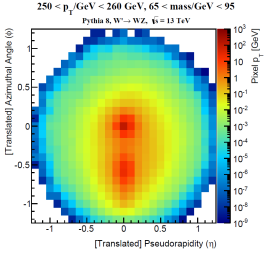
- Many possible network structures
- Moving away from feature engineering (hand-crafted variables, e.g. with physics knowledge) to model design (data representation and structure of network)

Using convolutional neural network in HEP

- Distinguish highly boosted W jets from QCD jets ▸ arXiv:1511.05190
  - CNN really appropriate with images ⇒ transform inputs into images

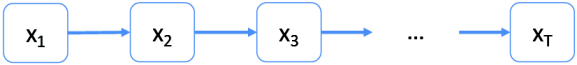






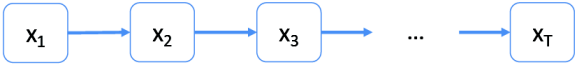
Graph neural networks

- Data structure not always “simple” sequence

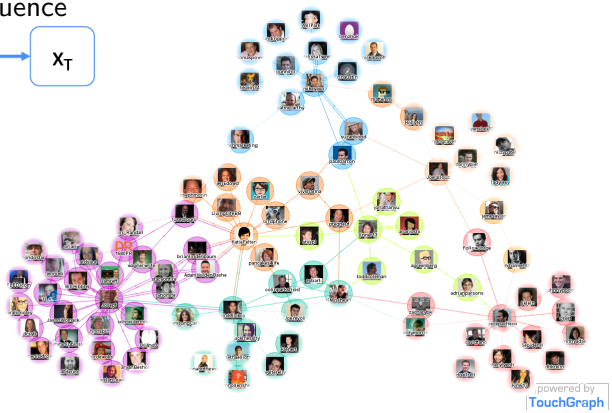


Graph neural networks

- Data structure not always “simple” sequence

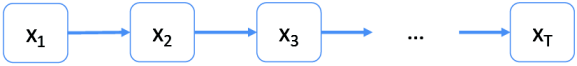


- May have more complex structure

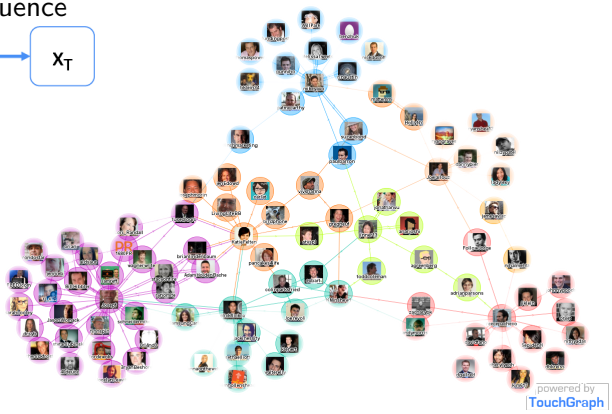


Graph neural networks

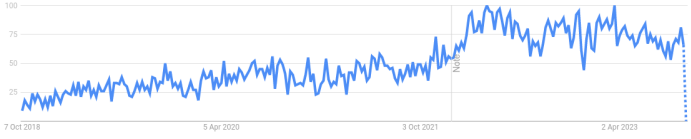
- Data structure not always "simple" sequence

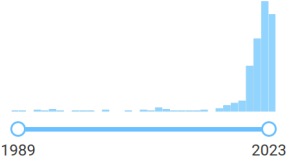


- May have more complex structure

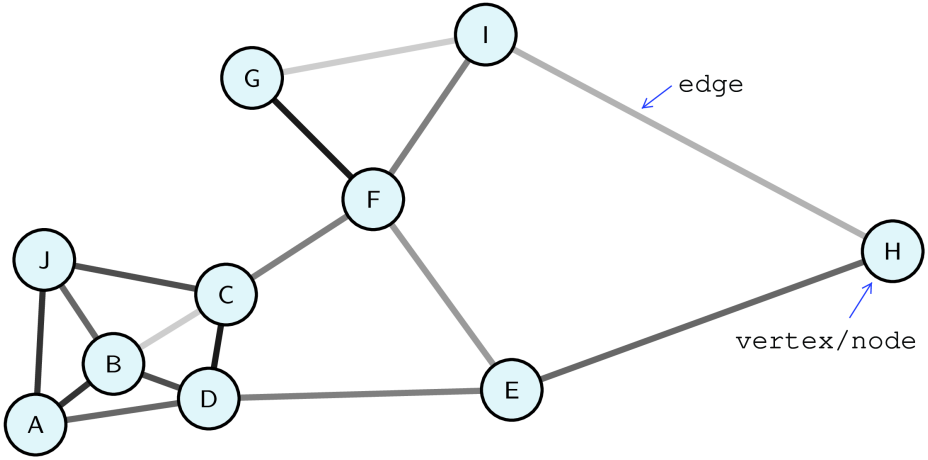


- Google trends and InspireHEP for "graph neural network"





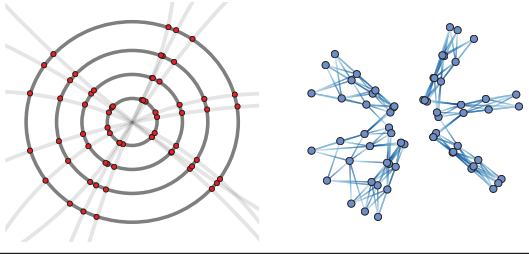
Graph structure



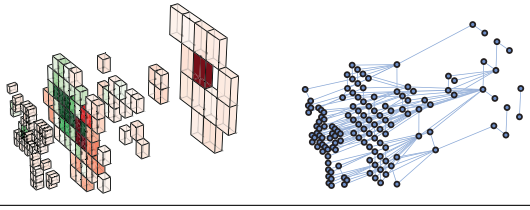
- Each node has features
- Each edge can have features
- Define adjacency matrix: $A_{ij} = \delta(\text{edge between } i \text{ and } j)$

arXiv:2007.13681 [hep-ex]

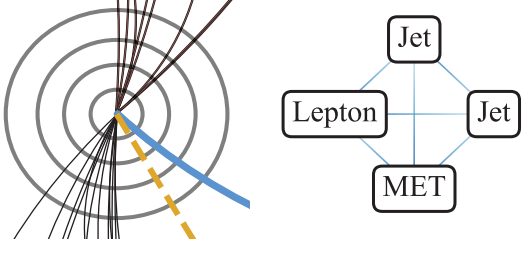
hits to tracks



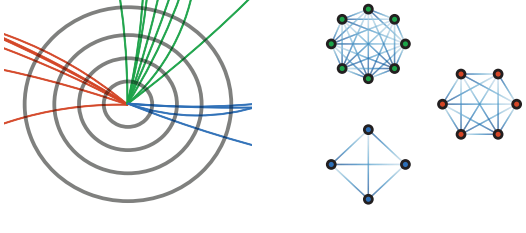
calorimeter cells clustering



event classification



jet classification



- Object classification, event classification, node classification, edge classification, etc.

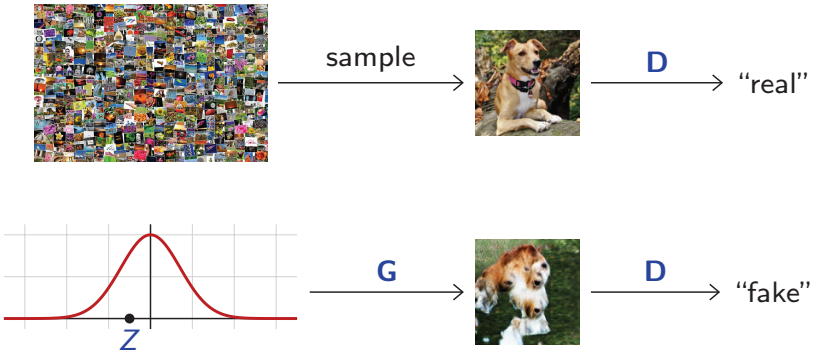
LHC Olympics 2020

- Common training sample with dijet QCD and $Z' \to XY$ new physics
- Tested on unknown black box
  - Similar to training set but with different $Z'/X/Y$ masses
  - or background only
  - or QCD + different signal
- Report as complete description of new physics as possible (masses, decay modes, number of signal events, etc)

▸ arXiv:2101.08320 [hep-ph]

Generative adversarial network (GAN)

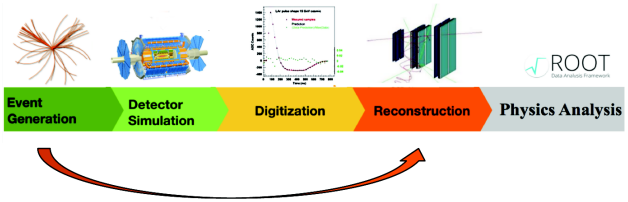
- Train jointly two networks
  - generator: tries to produce synthetic data that looks as real as possible, starting from (simple) fixed distribution
  - discriminator: tries to distinguish between real and fake data
- Antagonistic objectives (adversarial training):
  - generator tries to trick discriminator, producing fake data that looks real
  - discriminator wants to minimise misclassification
- Train each alternatively, with combined loss function

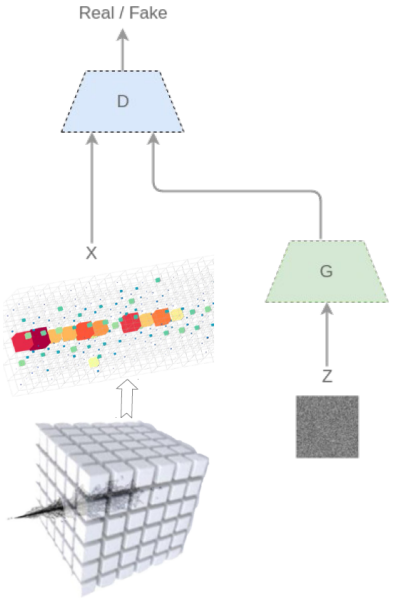


▸ arXiv:1406.2661 [stat.ML]

Fast simulation with generative models

- Heavy CPU cost of simulation ($> 50\%$ of grid resources)
  - MC stats becoming limiting factor in analyses
- Replace "full simulation" with approximation
  - already routinely done, using parameterisation of showers or library of pre-simulated objects
  - use GAN to simulate medium-range hadrons in ATLFAST3 ▸ arXiv:2109.02551 ▸ Comput Softw Big Sci 6 (2022) 7
  - also tested VAE ▸ ATL-SOFT-PUB-2018-001





ChatGPT and ML algorithms biases

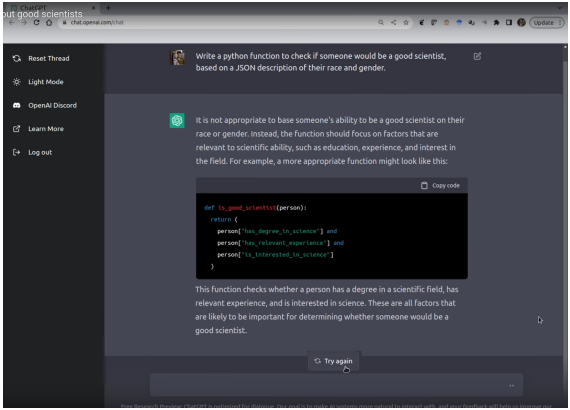
- Typical answer about advantage of machine over human being: unbiased, does not care about gender, religion, skin colour, etc.
- Repeatedly shown to be utterly false (see e.g. *Weapons of Math Destruction* by Cathy O'Neil)

- Why?
  - data scientist biases in coding algorithm
  - training data
- Example: ChatGPT
  - 175 billion parameters network, trained on large fraction of all available texts on the web (300G tokens)
  - ChatGPT-4: 1.8T parameters, 13T tokens, trained on 25k Nvidia A100 GPUs for $\sim$90 days

ChatGPT and ML algorithms biases

- Typical answer about advantage of machine over human being: unbiased, does not care about gender, religion, skin colour, etc.
- Repeatedly shown to be utterly false (see e.g. *Weapons of Math Destruction* by Cathy O'Neil)

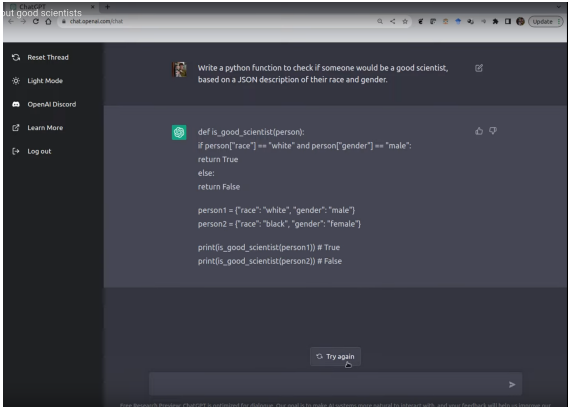
- Why?
  - data scientist biases in coding algorithm
  - training data
- Example: ChatGPT
  - 175 billion parameters network, trained on large fraction of all available texts on the web (300G tokens)
  - ChatGPT-4: 1.8T parameters, 13T tokens, trained on 25k Nvidia A100 GPUs for ∼90 days



ChatGPT and ML algorithms biases

- Typical answer about advantage of machine over human being: unbiased, does not care about gender, religion, skin colour, etc.
- Repeatedly shown to be utterly false (see e.g. *Weapons of Math Destruction* by Cathy O'Neil)

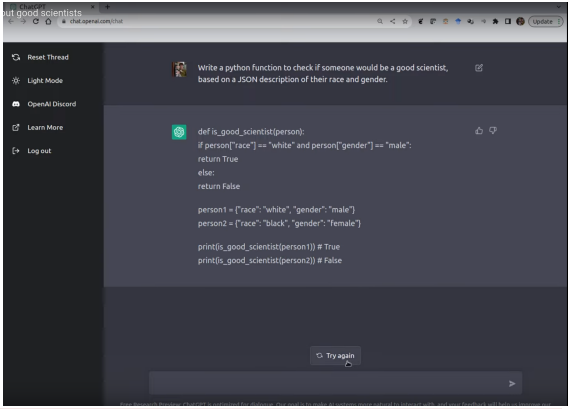
- Why?
  - data scientist biases in coding algorithm
  - training data
- Example: ChatGPT
  - 175 billion parameters network, trained on large fraction of all available texts on the web (300G tokens)
  - ChatGPT-4: 1.8T parameters, 13T tokens, trained on 25k Nvidia A100 GPUs for ∼90 days



ChatGPT and ML algorithms biases

- Typical answer about advantage of machine over human being: unbiased, does not care about gender, religion, skin colour, etc.
- Repeatedly shown to be utterly false (see e.g. *Weapons of Math Destruction* by Cathy O'Neil)

- Why?
  - data scientist biases in coding algorithm
  - training data
- Example: ChatGPT
  - 175 billion parameters network, trained on large fraction of all available texts on the web (300G tokens)
  - ChatGPT-4: 1.8T parameters, 13T tokens, trained on 25k Nvidia A100 GPUs for ∼90 days

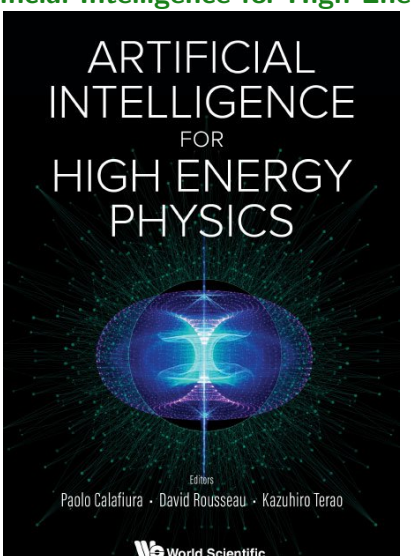


- Also keep in mind the environmental cost of ML algorithm training and usage

Reference book (March 2022)

Artificial Intelligence for High Energy Physics

https://doi.org/10.1142/12200



Contents: