



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

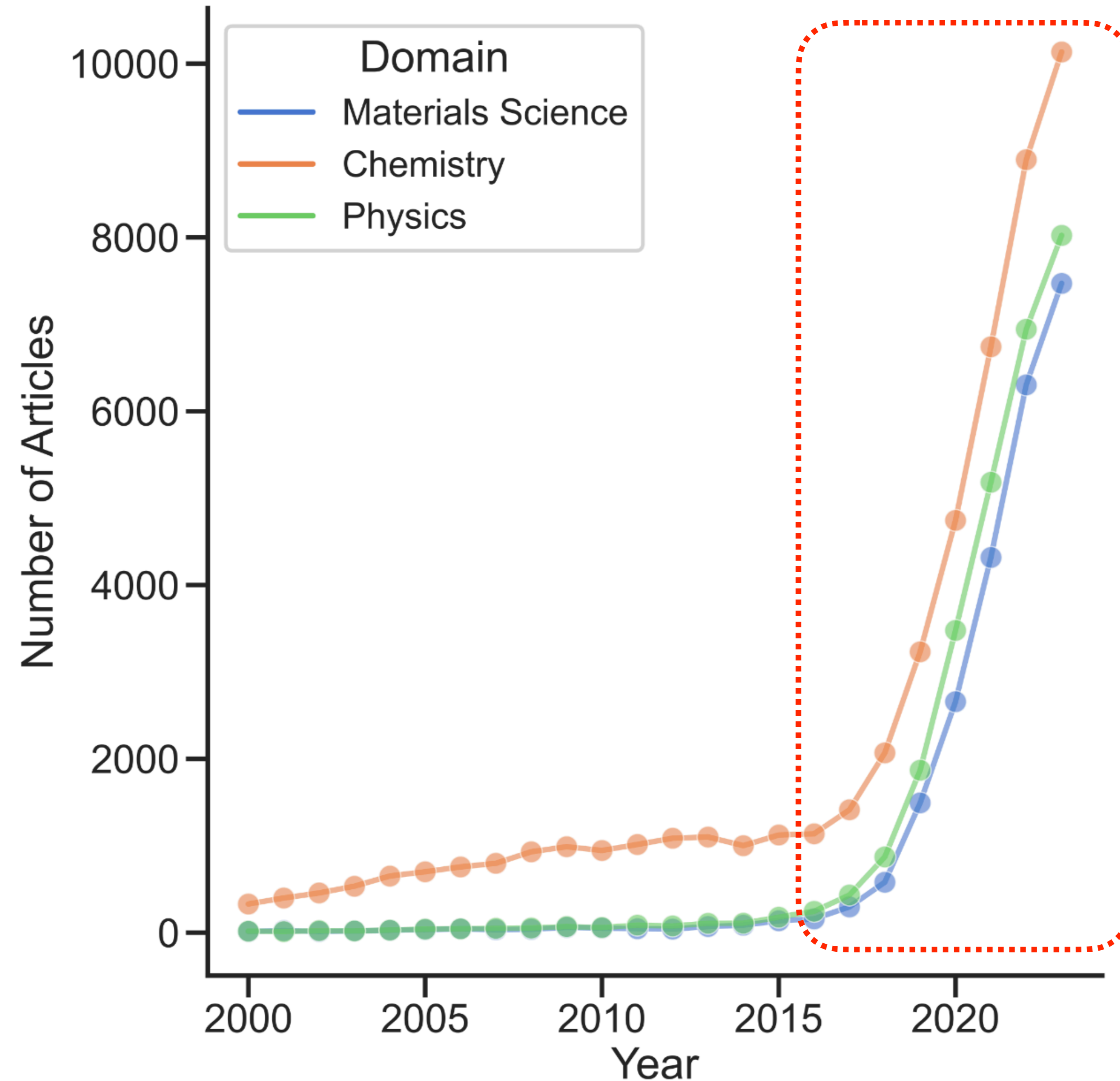
# AI-powered science: advancing High-Energy Physics and beyond

Daniele Bonacorsi (University of Bologna / INFN)



ISGC 2025 - Taipei, Taiwan

# ML-based publications in science



**What** is this ramp-up?!

And **why** at that time?!

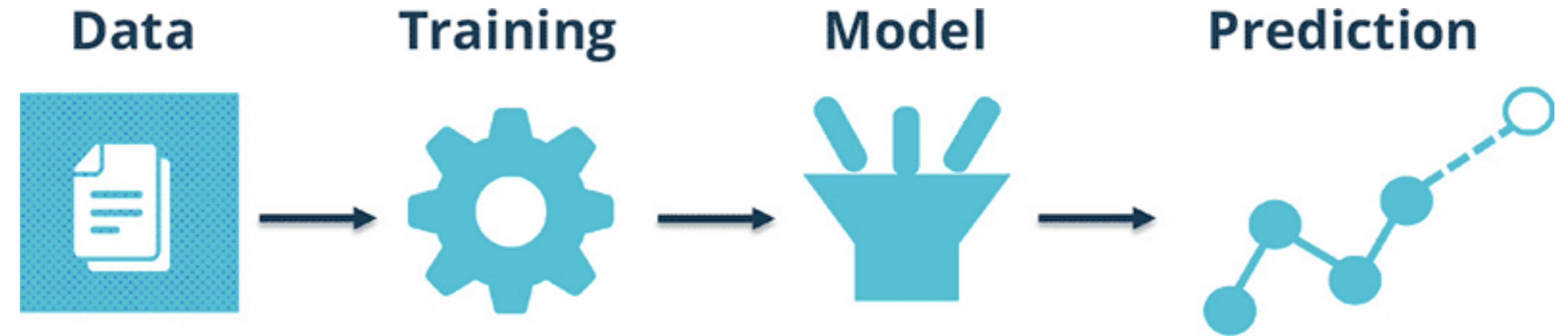
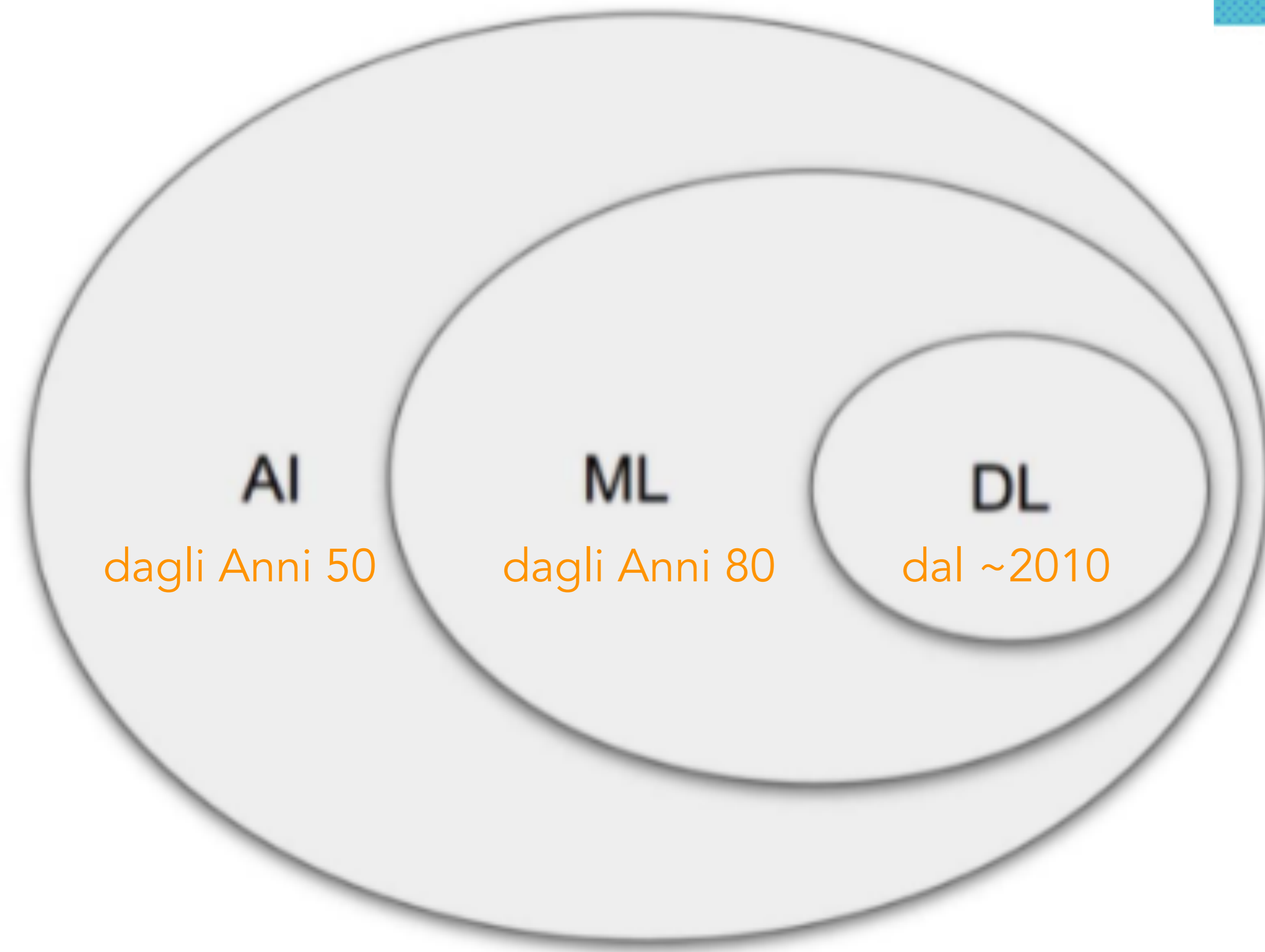
# AI vs ML vs DL

AI = Artificial Intelligence

ML = Machine Learning

NN = Neural Network

DL = Deep Learning



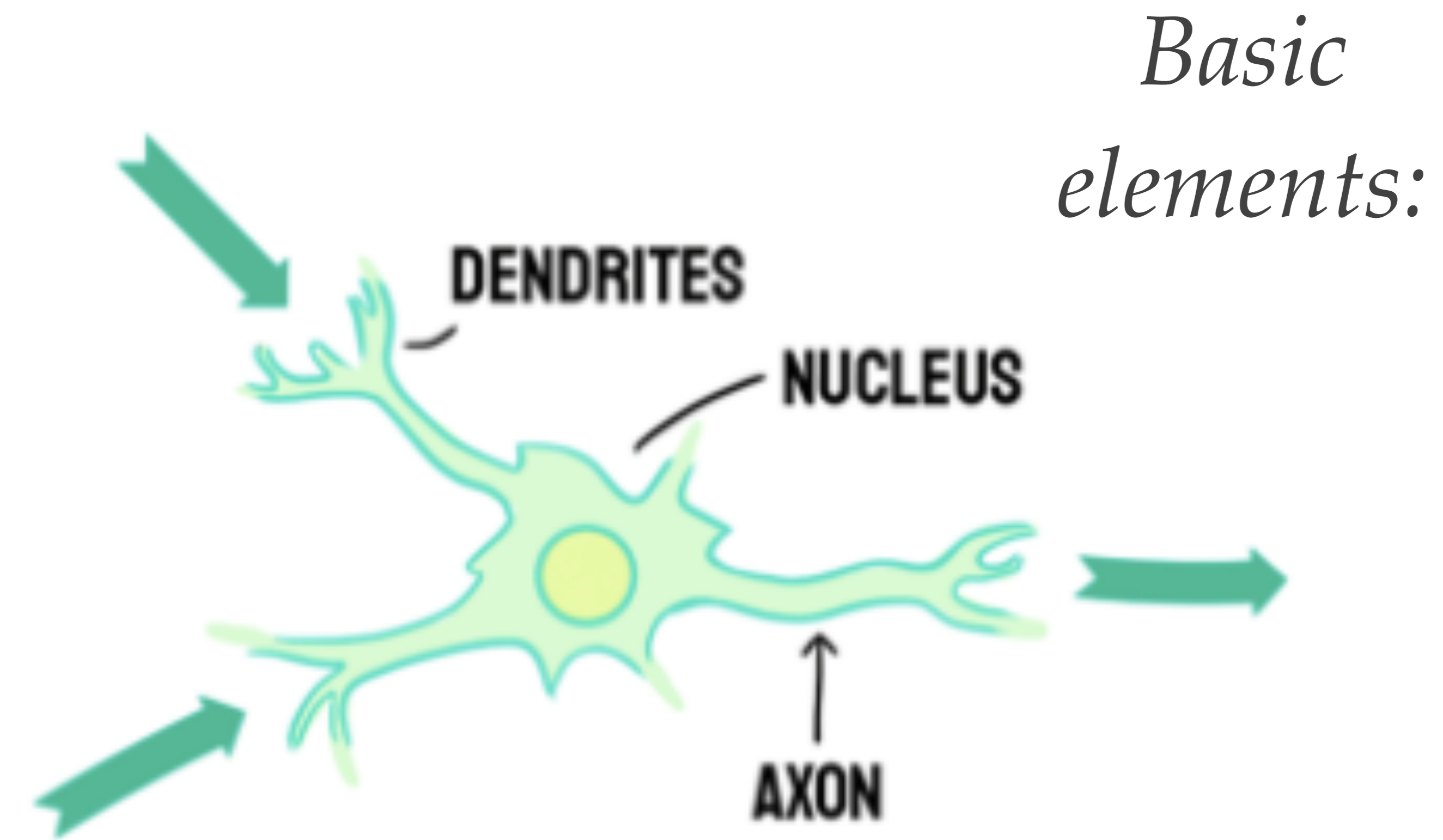
Objective: extract “actionable insight” from (big) data

Choose an **algorithm**, perform its “**training**” on data (“attributes” vs “**features**”) to extract “**parameters**” with optimisation techniques (e.g. “**gradient descent**”) that minimise the errors of the model on the observations (“**cost function**”), in a process governed by “**hyper-parameters**” tuning

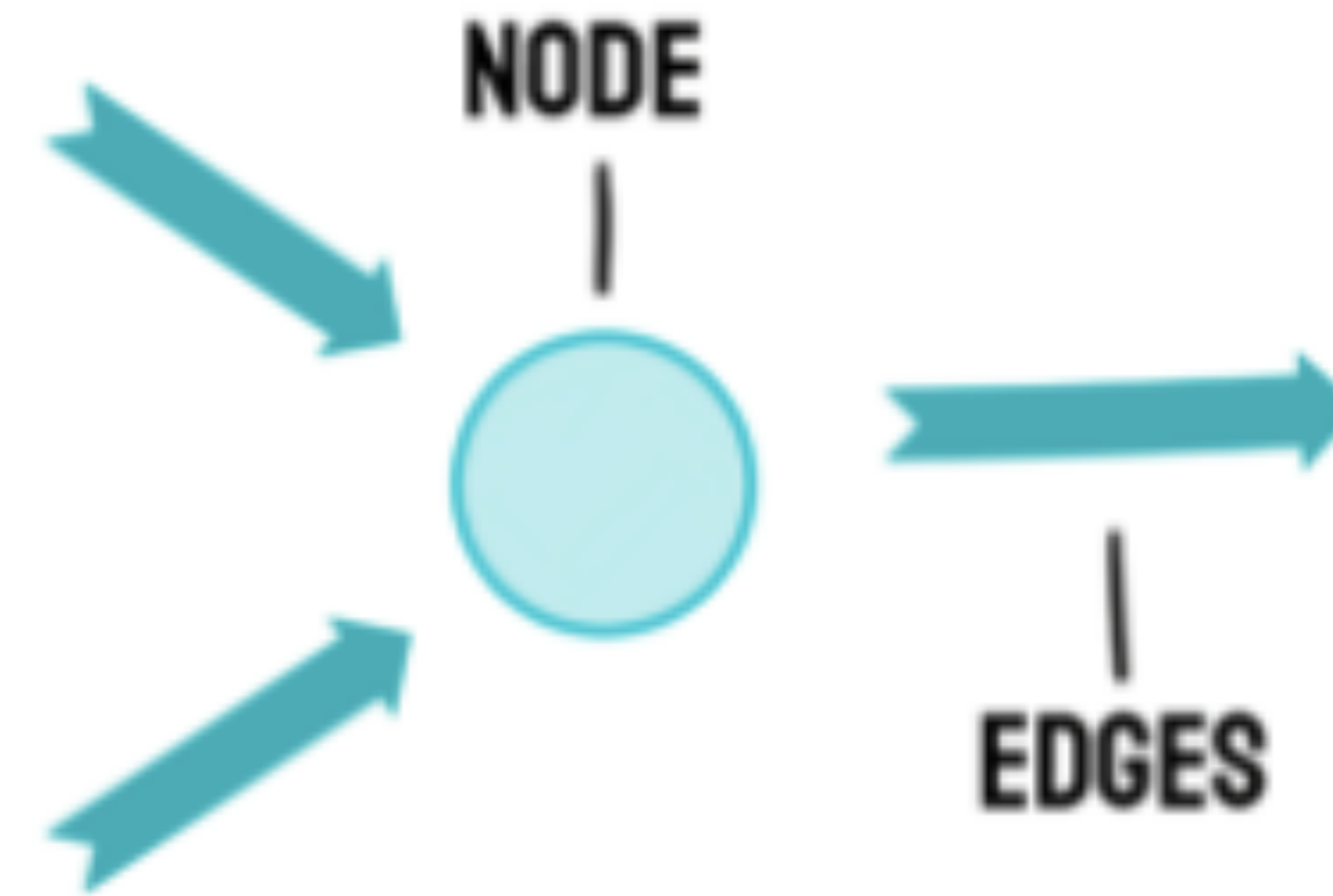
Result: a **ML model** to be applied to previously unseen data

→ “**data-driven modelling**”

Biological Neural Networks

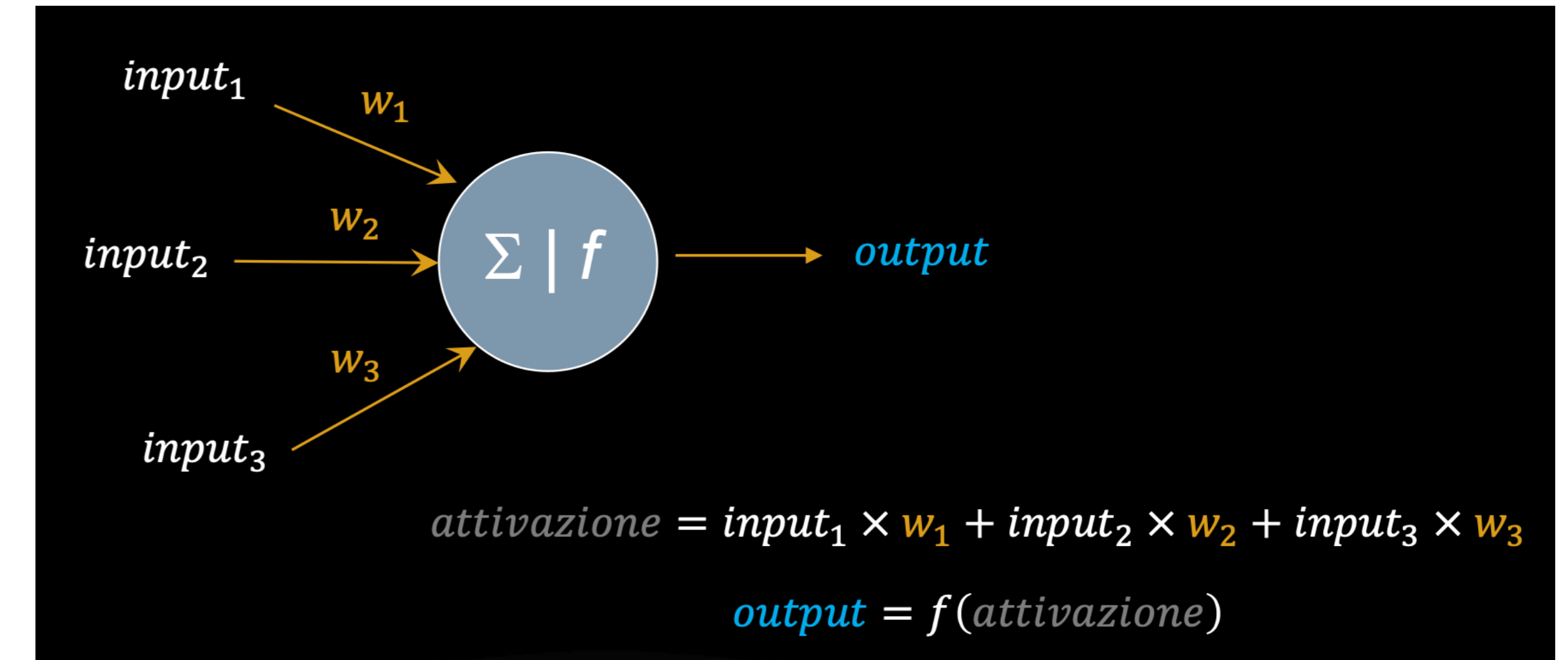


Artificial Neural Networks



## First models of **artificial neuron** (McCulloch, Pitts, 1943)

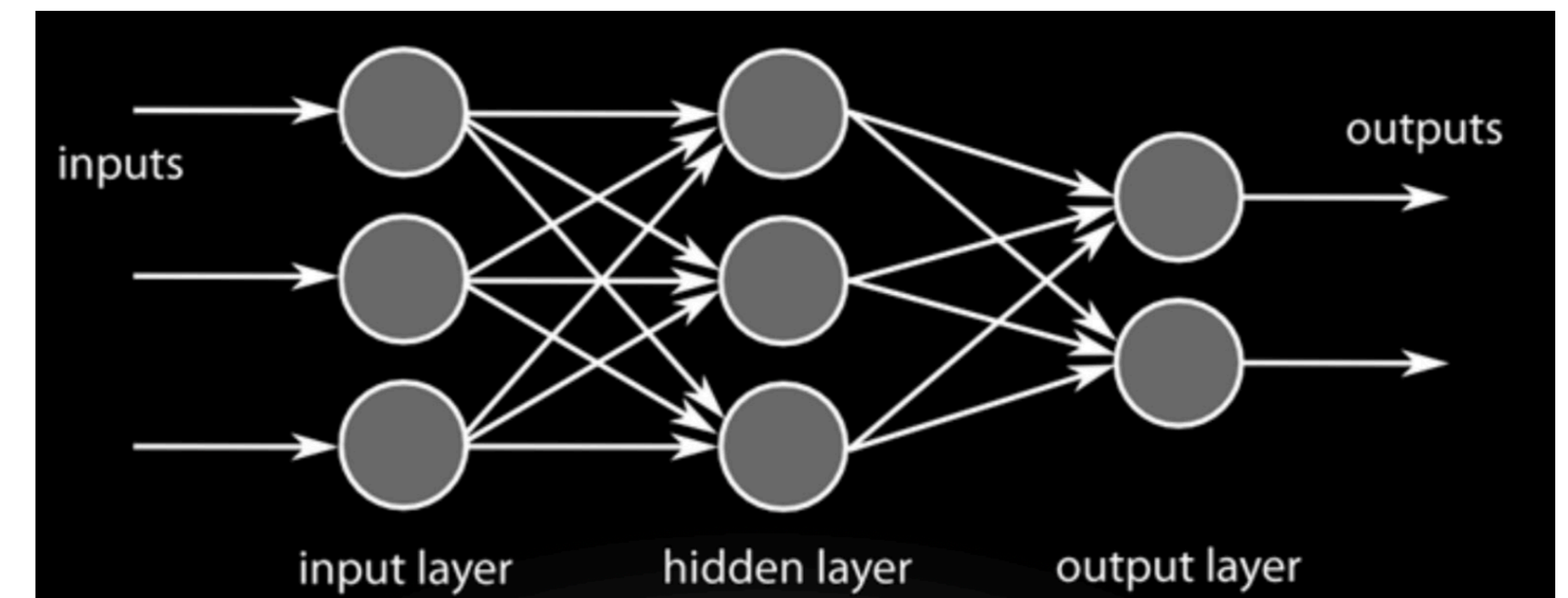
- Neuron behaviour (i.e. info passing) are determined by “weights”: initially randomly set, they are modified during “training”



## Towards **neural networks**: “Perceptron” (Rosenblatt, 1958)

### Multi-Layer Perceptron (MLP)

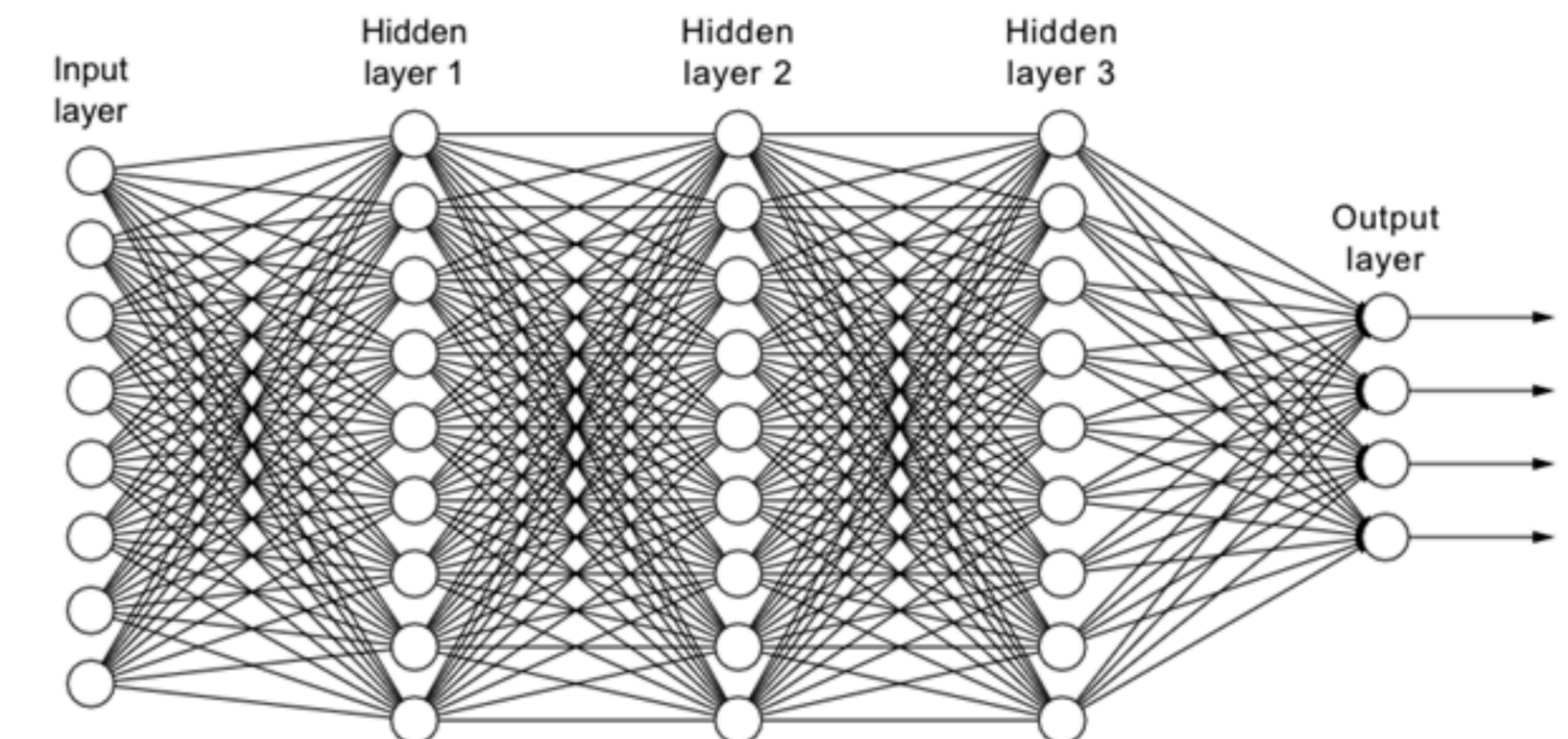
- Groups of neurons organized in **layers** (**input** layer, **output** layer, and one or more intermediate “**hidden**” layers, each with one or more neurons)



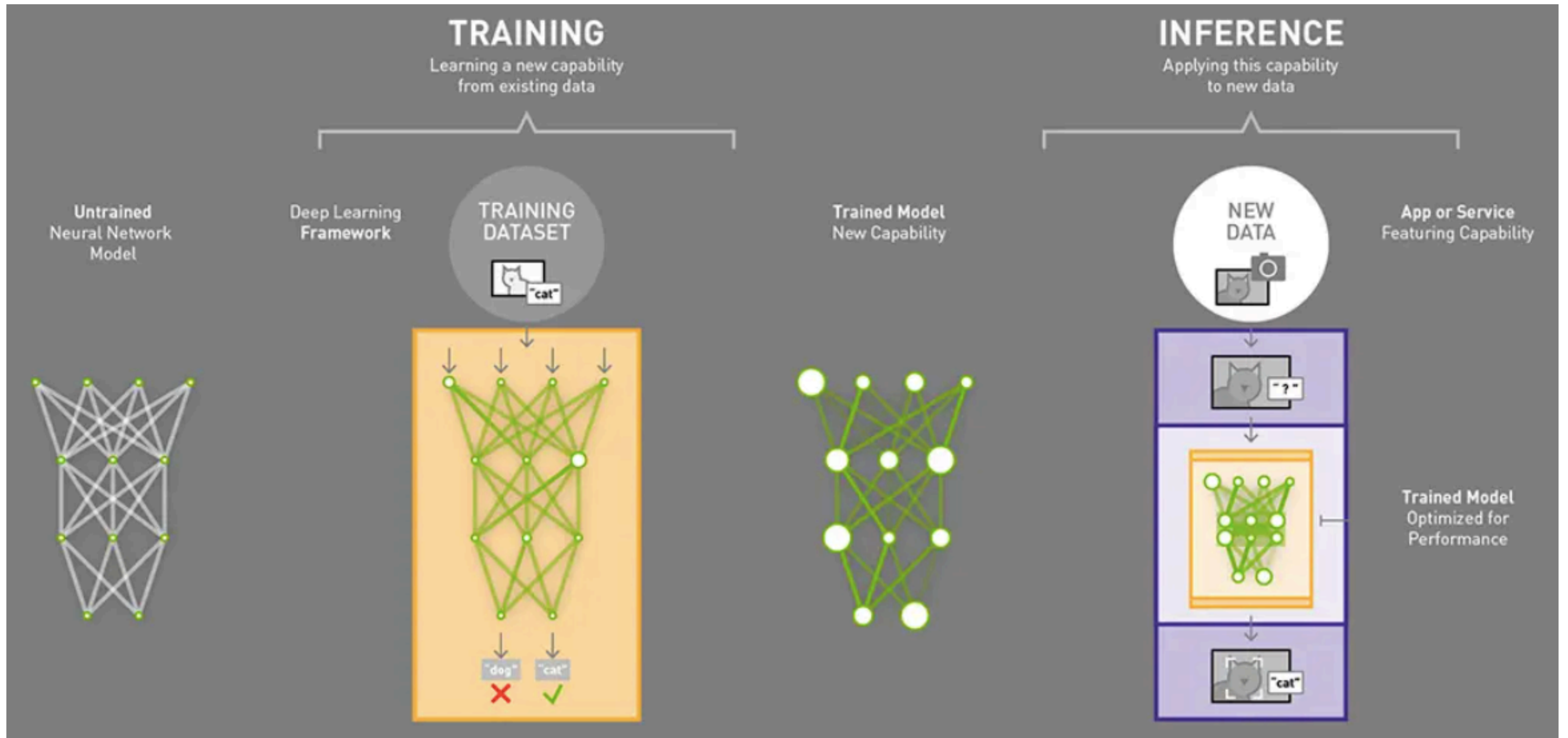
## **Back-propagation** algo (Rumelhart, Hinton, Williams, 1986)

- an effective technique to train them (i.e. setting weights’ values)

## From shallow NN to Deep NN → **Deep Learning**



# E.g. "supervised learning": ML training vs inference



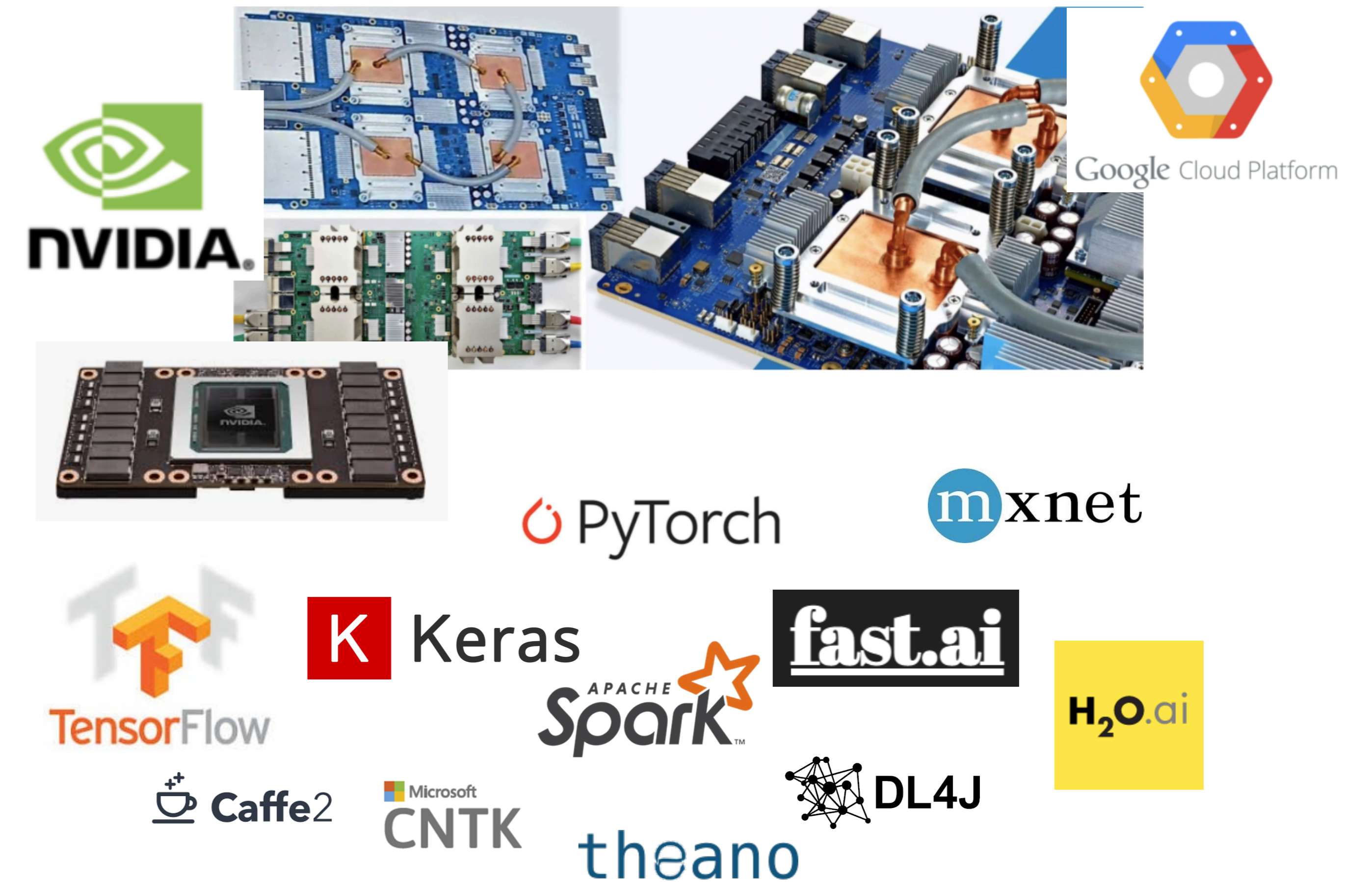
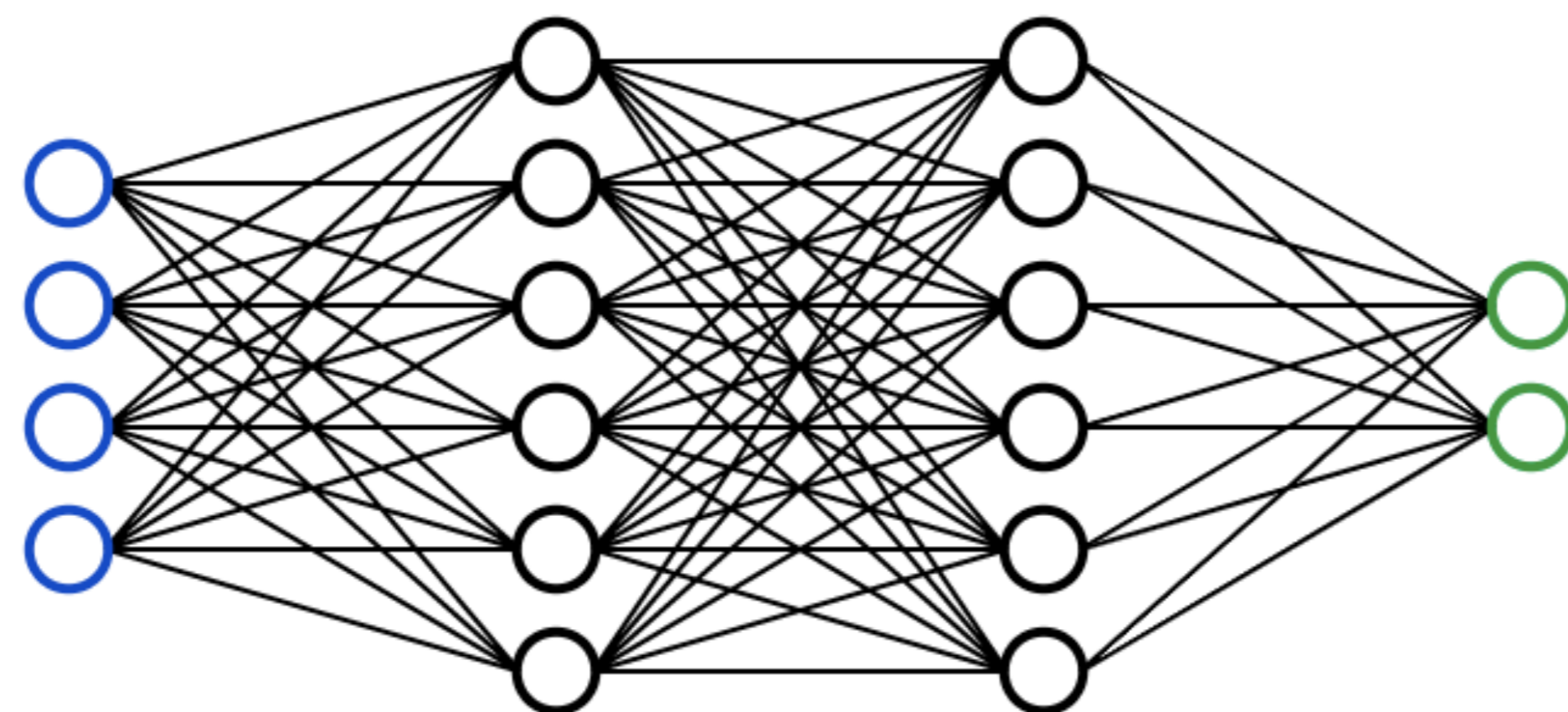
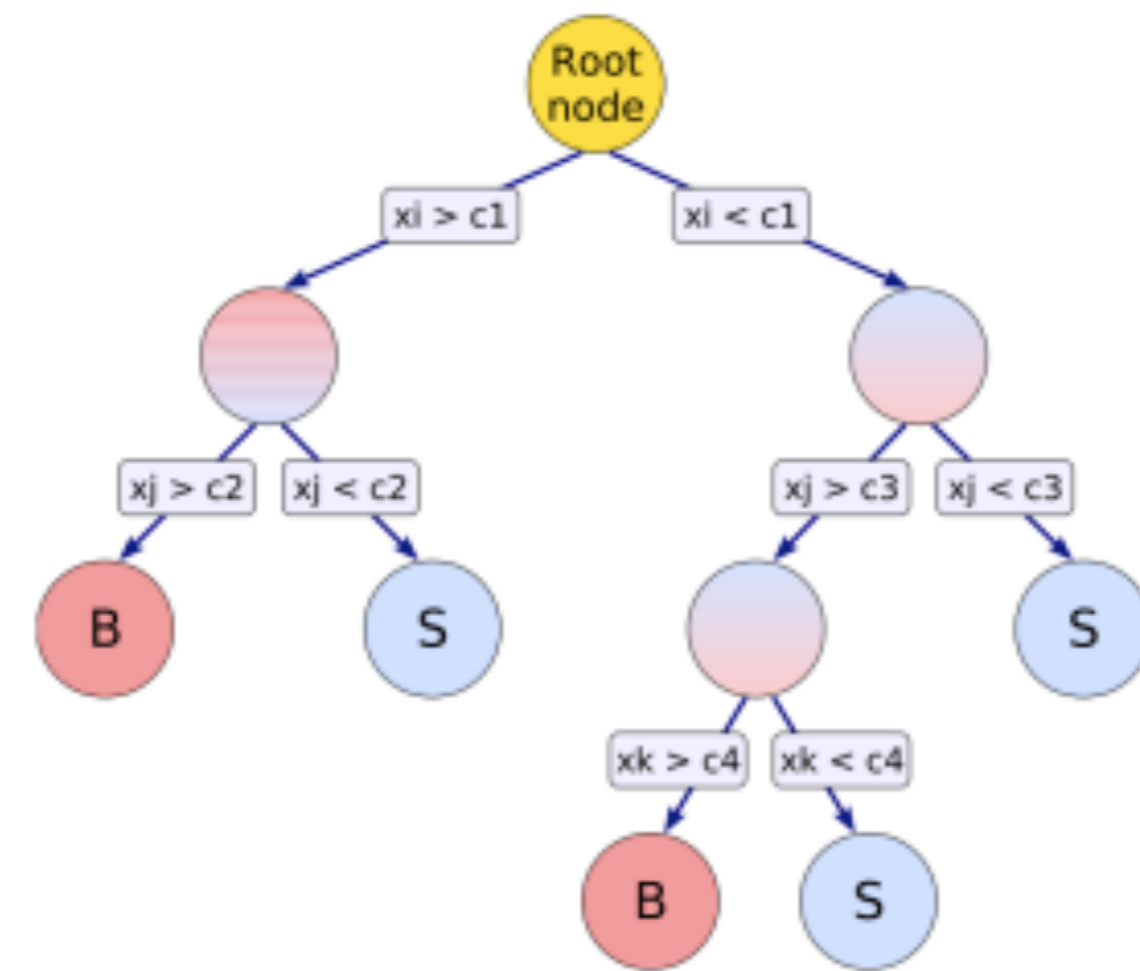
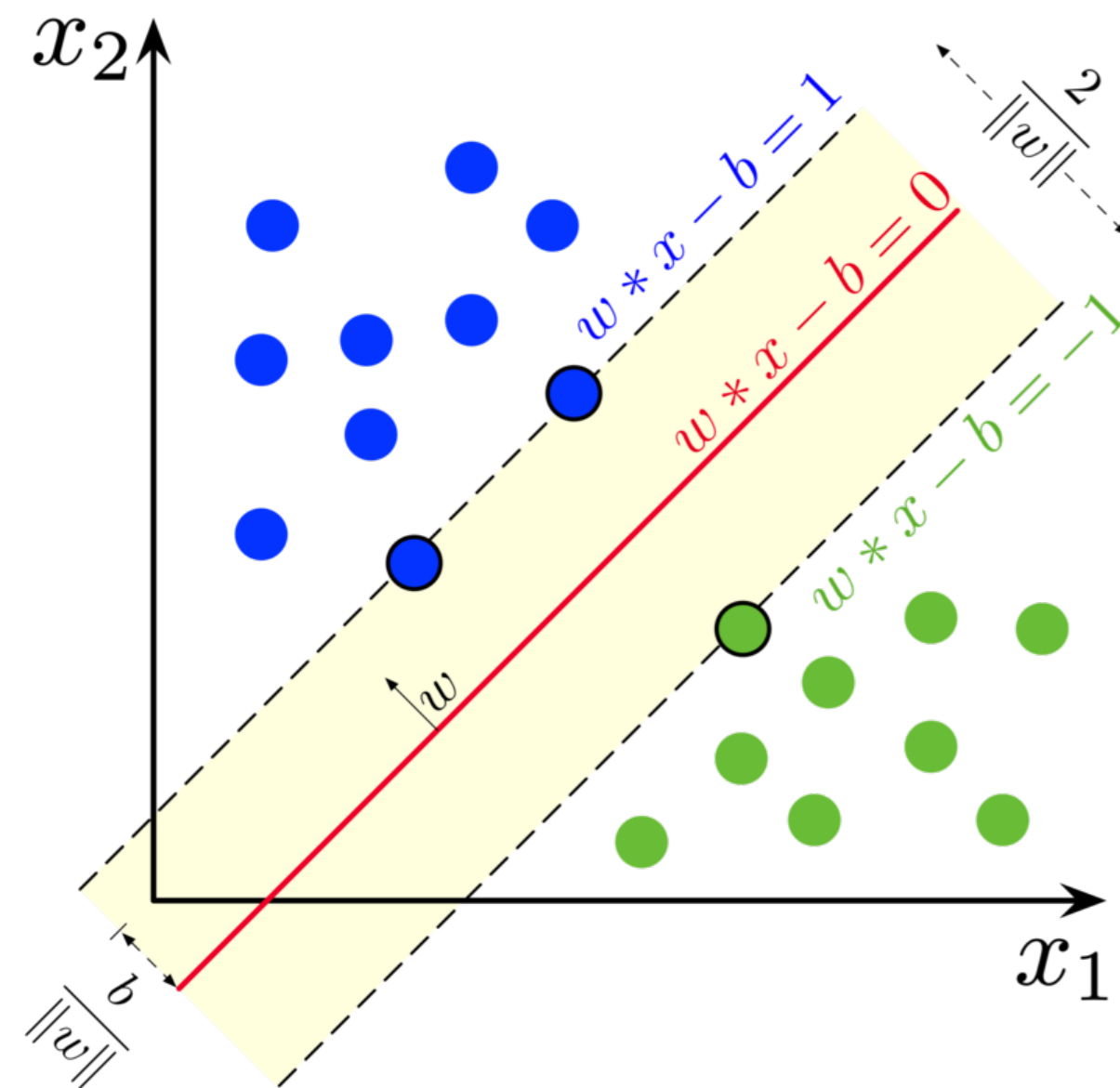
# AI's past winters and current spring

## Past "winters" of AI

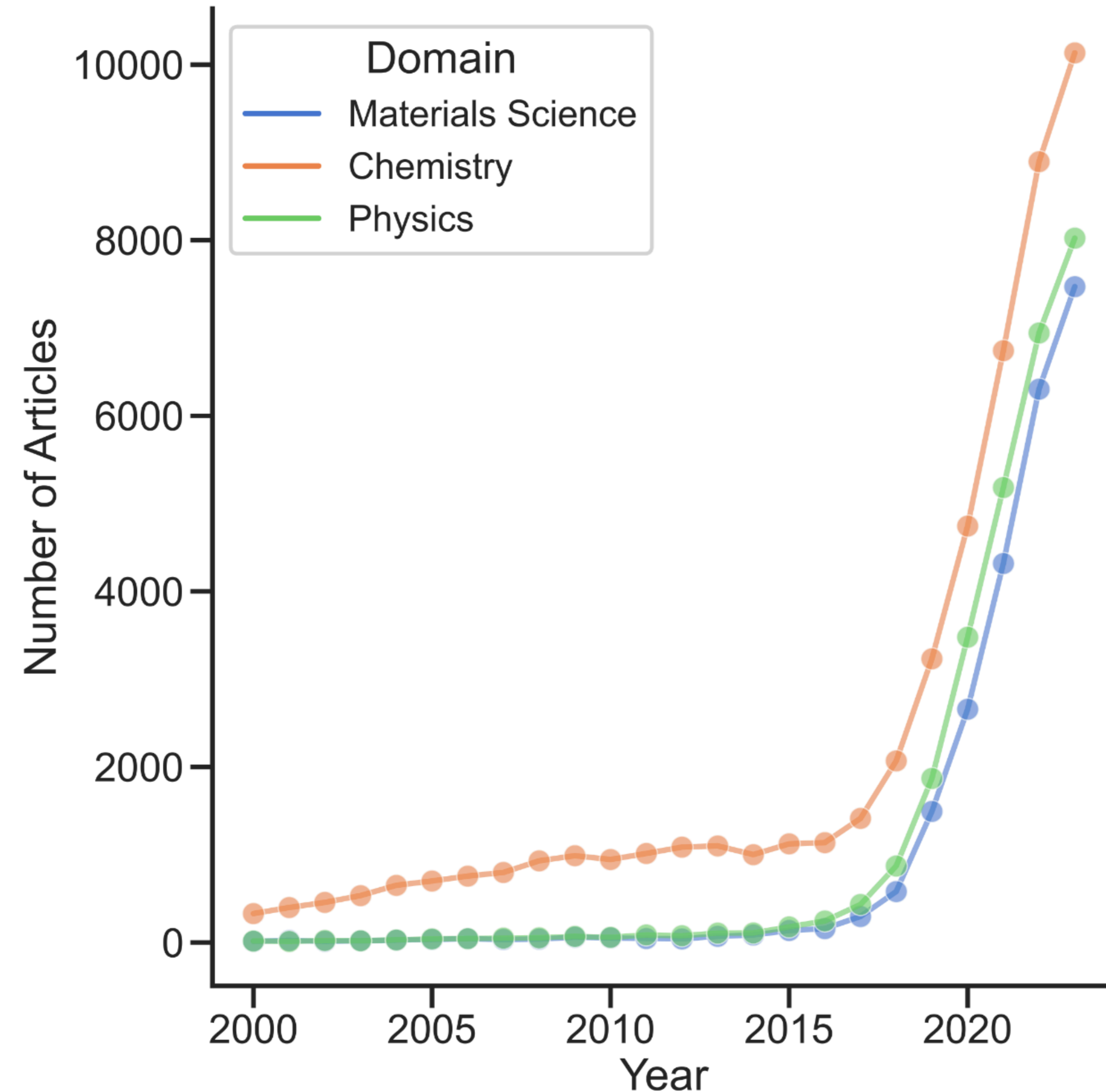
- '60: **shallow NN** hard to train
- '90: Support Vector Machines (**SVM**), Boosted Decision Trees (**BDT**), ..
- 2000+: advanced **deep NN** architectures

Current "**spring**" explainable by:

- "Big data"
- **Technology + ML research**
- (cloud and accessibility)



# ML-based publications in science



All this may explain the rise as from the initial plot..

Extremely large adoption in incredibly short times, towards high level of **pervasiveness**



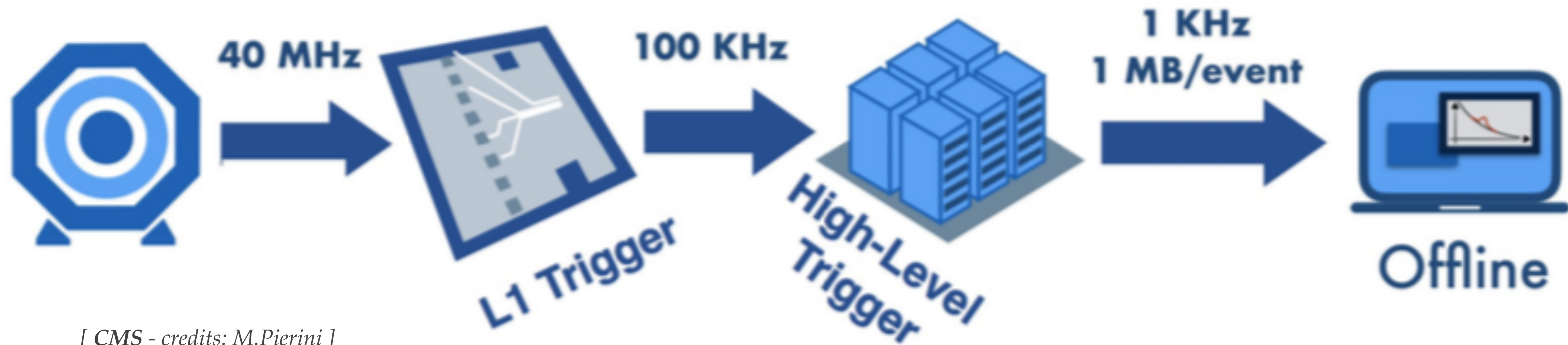
## ML in **data acquisition** and **trigger**

- Bkg and trigger rate reduction
- Signal specific trigger paths
- Anomaly detection in data taking
- Unsupervised new physics mining

E.g. **LHC experiments' trigger** is a strong “driver” for **high-performances ML applications**

- Next-gen trigger systems → real-time reconstruction → real time analysis

Challenge is the trade-off between **algorithmic complexity** and the **performances** achievable under severe time constraints in inference

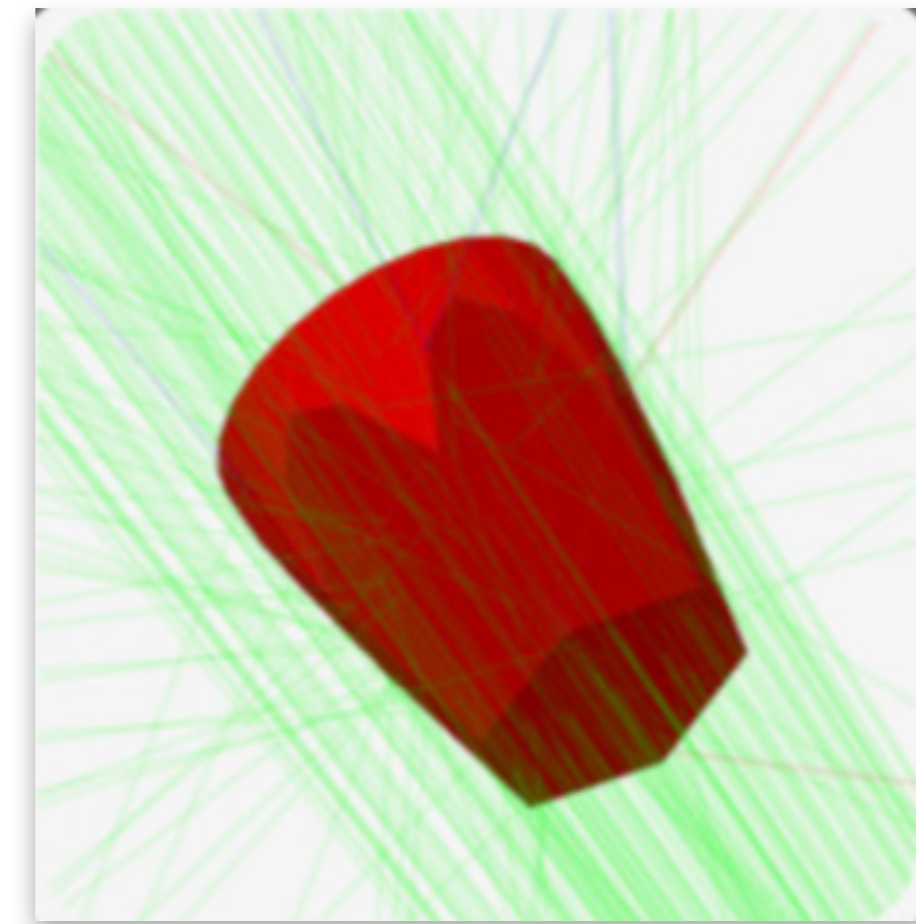


[ CMS - credits: M.Pierini ]

## ML in Event **Simulation**

The production of simulated events (full/fast simulation) is extremely intense from the computation standpoint (up to the point it might impact the physics reach of the experiments). **ML can help to reduce such load**

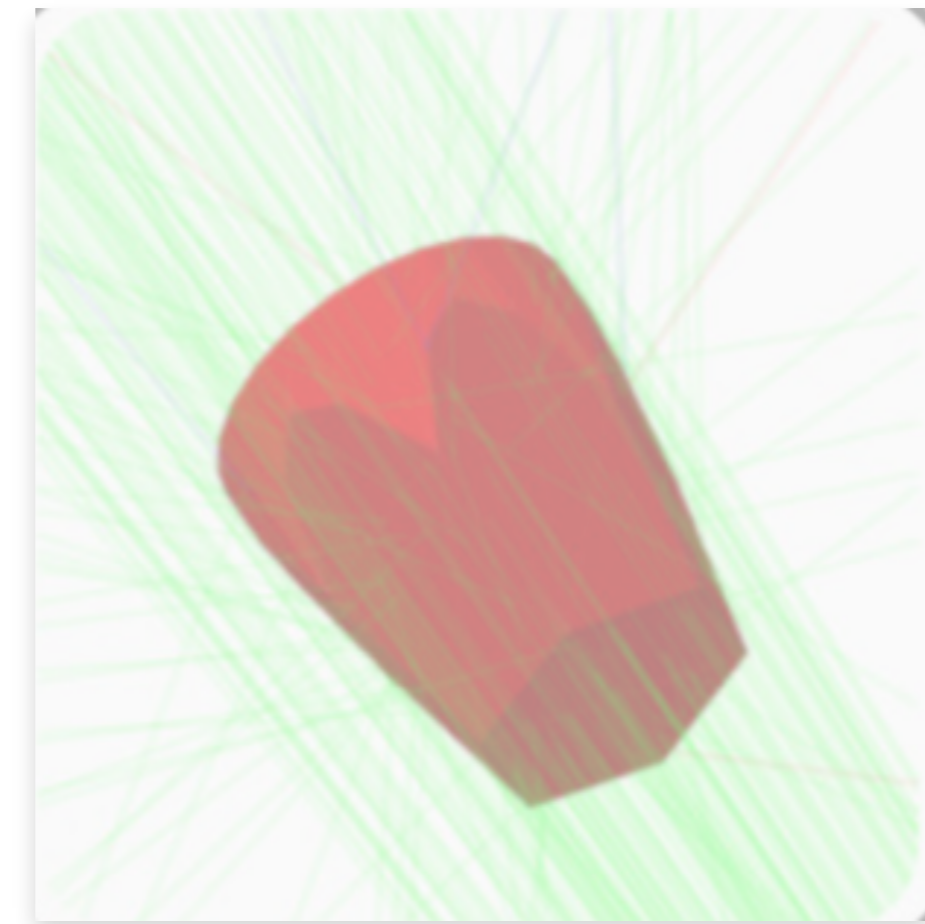
- Calorimeter shower surrogate simulator
- Analysis level simulator
- Pile-up overlay generator
- Monte Carlo integration
- ML-enabled fast-simulation
- Invertible full-simulation (probabilistic programming, ...)
- ...



## ML in Event **Simulation**

The production of simulated events (full/fast simulation) is extremely intense from the computation standpoint (up to the point it might impact the physics reach of the experiments). **ML can help to reduce such load**

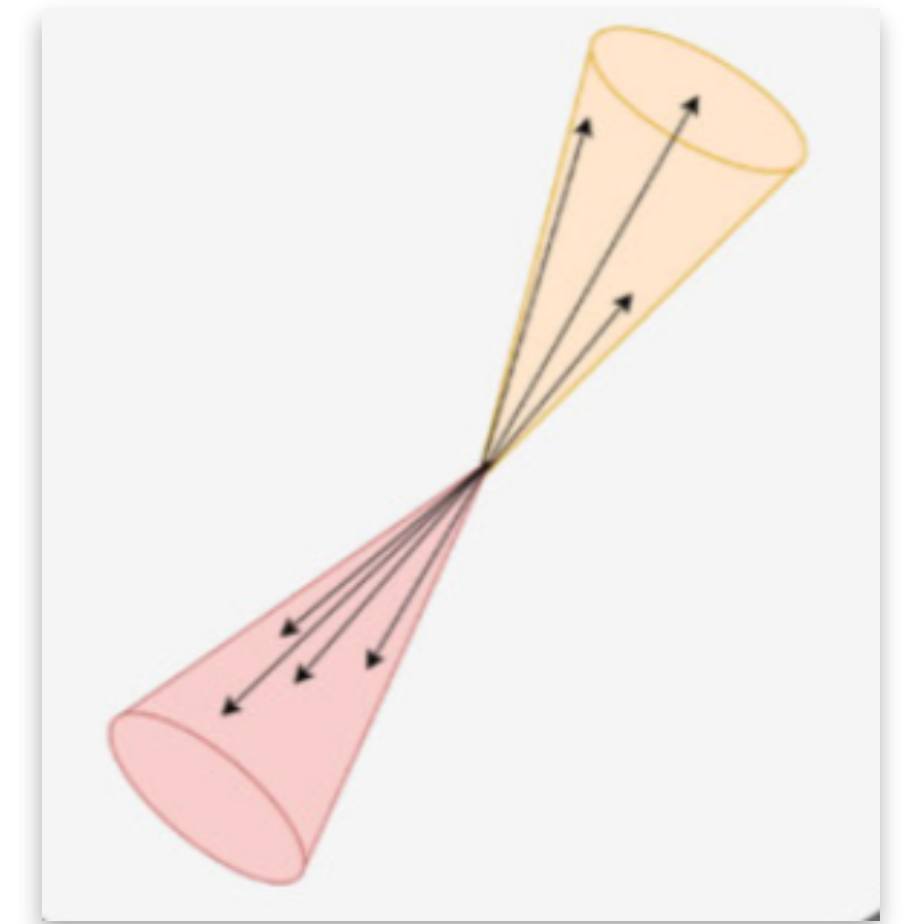
- Calorimeter shower surrogate simulator
- Analysis level simulator
- Pile-up overlay generator
- Monte Carlo integration
- ML-enabled fast-simulation
- Invertible full-simulation (probabilistic programming, ...)
- ...



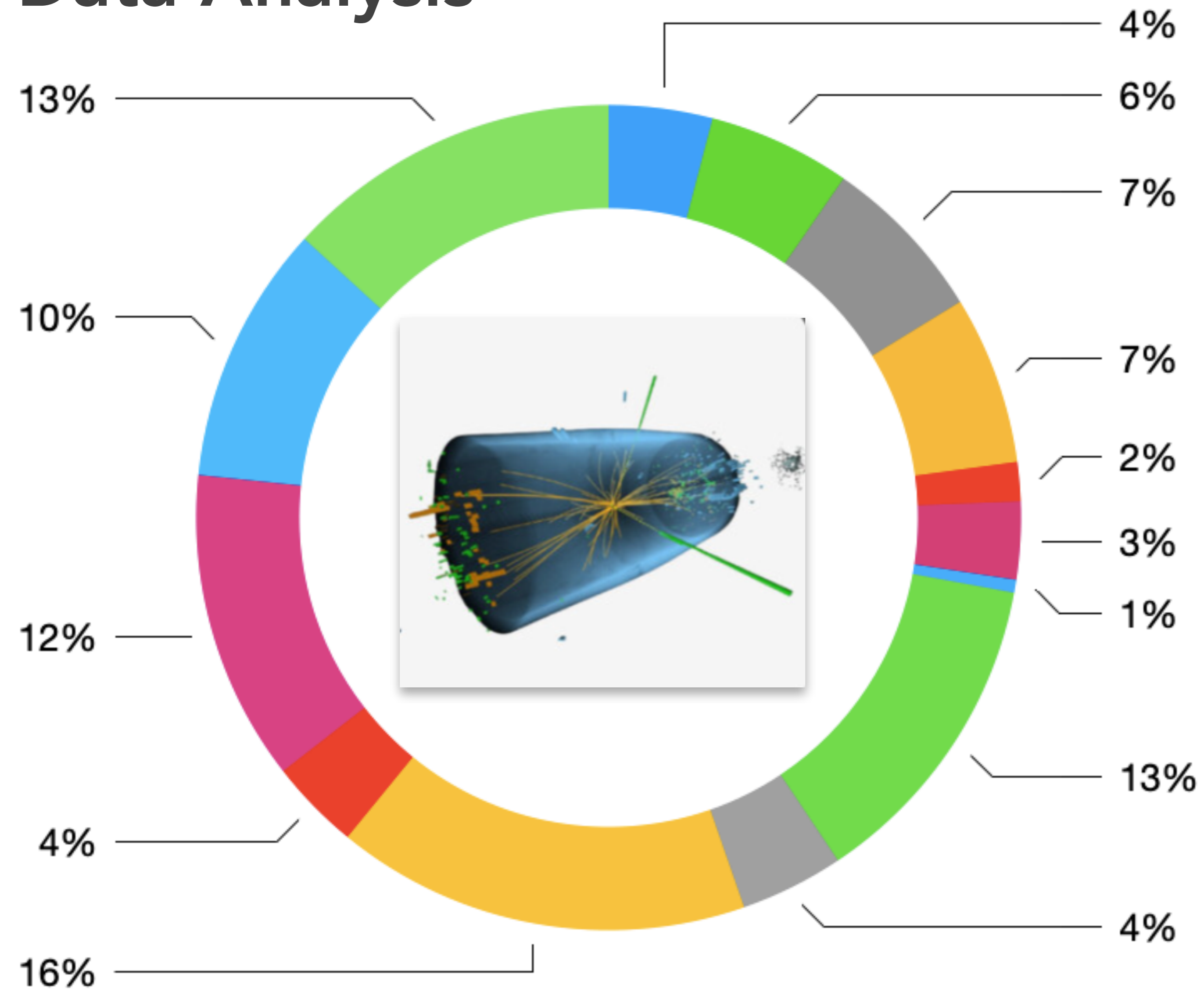
## ML in Event **Reconstruction**

Online/offline reconstruction might be partially **replaced by surrogate models** (approximate → faster) or by **new algorithms** (that might offer unprecedented performances)

- Charged particle tracking (GraphNN, vertexing, ...)
- Calorimeter reconstruction (local, clustering, ...)
- Particle flow (GraphNN, ...)
- Particle identification (boosted jets, isolation, ...)
- Pileup mitigation
- Energy regression (end-2-end, ...)
- ...



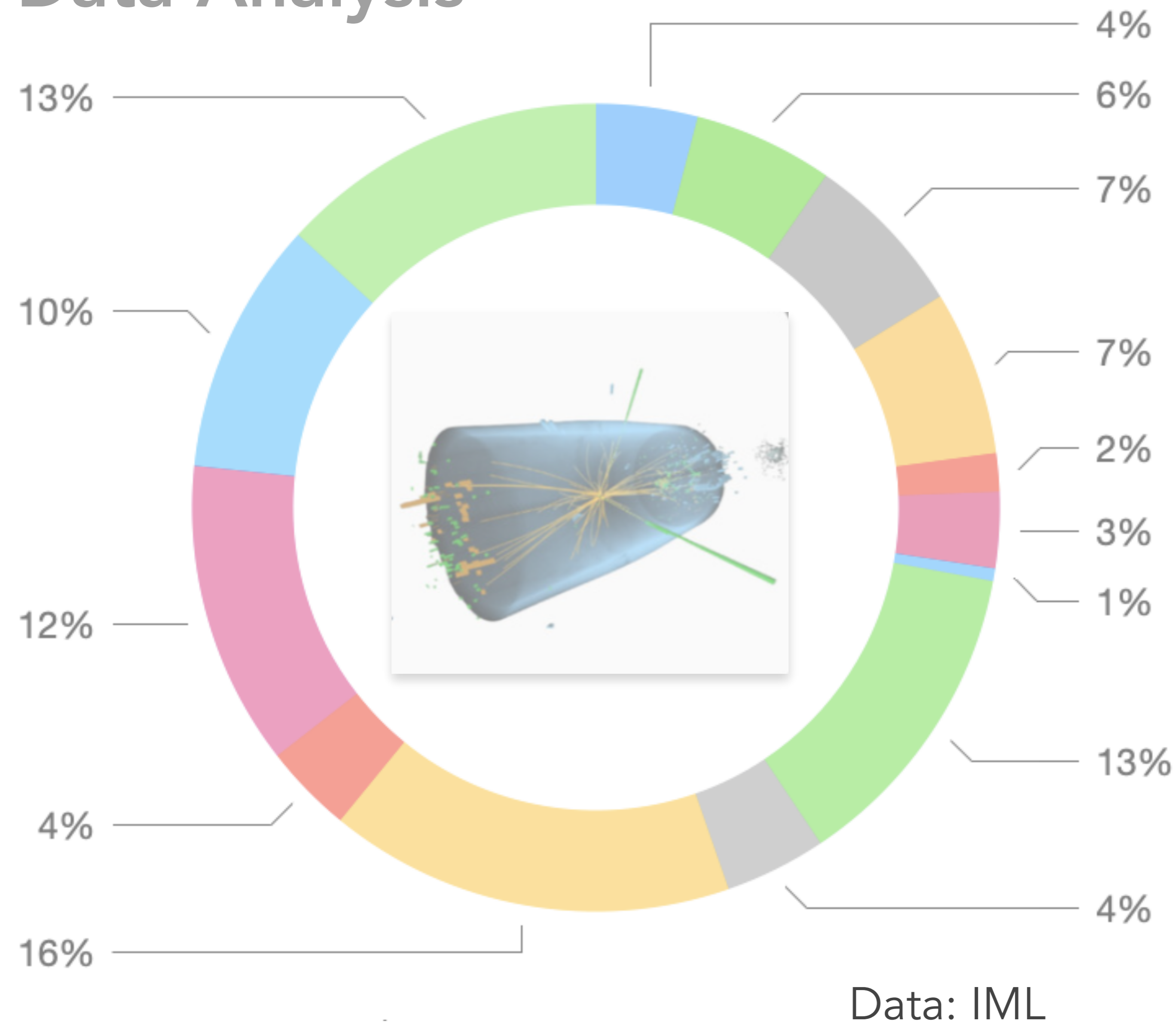
## ML in Data Analysis



Data: IML

- W/Z tagging
- H→b bar
- quarks and gluons
- top quark tagging
- strange jets
- b-tagging
- flavor physics
- BSM particles and models
- particle identification
- neutrino detectors
- direct Dark Matter detectors
- cosmology, astroparticle, and cosmic ray physics
- tracking
- heavy ions / nuclear physics

## ML in Data Analysis



- W/Z tagging
- H→b bar
- quarks and gluons
- top quark tagging
- strange jets
- b-tagging
- flavor physics
- BSM particles and models
- particle identification
- neutrino detectors
- direct Dark Matter detectors
- cosmology, astroparticle, and cosmic ray physics
- tracking
- heavy ions / nuclear physics

## ML in Computing Operations

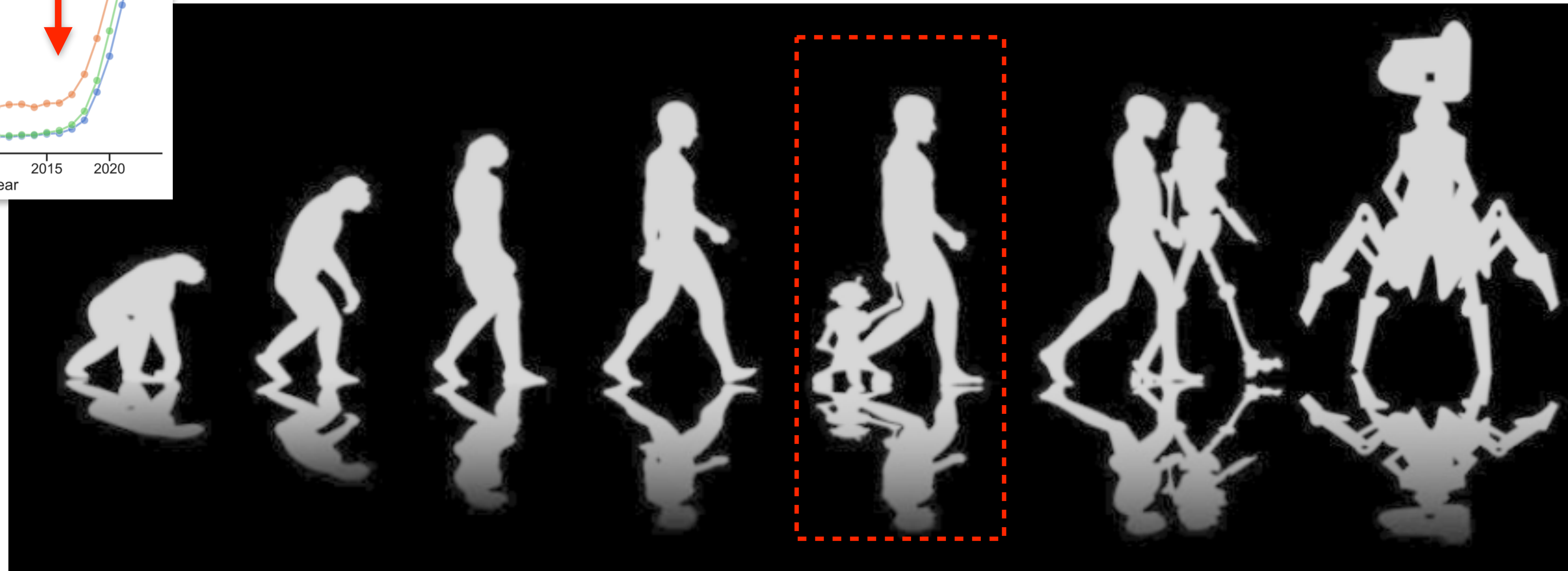
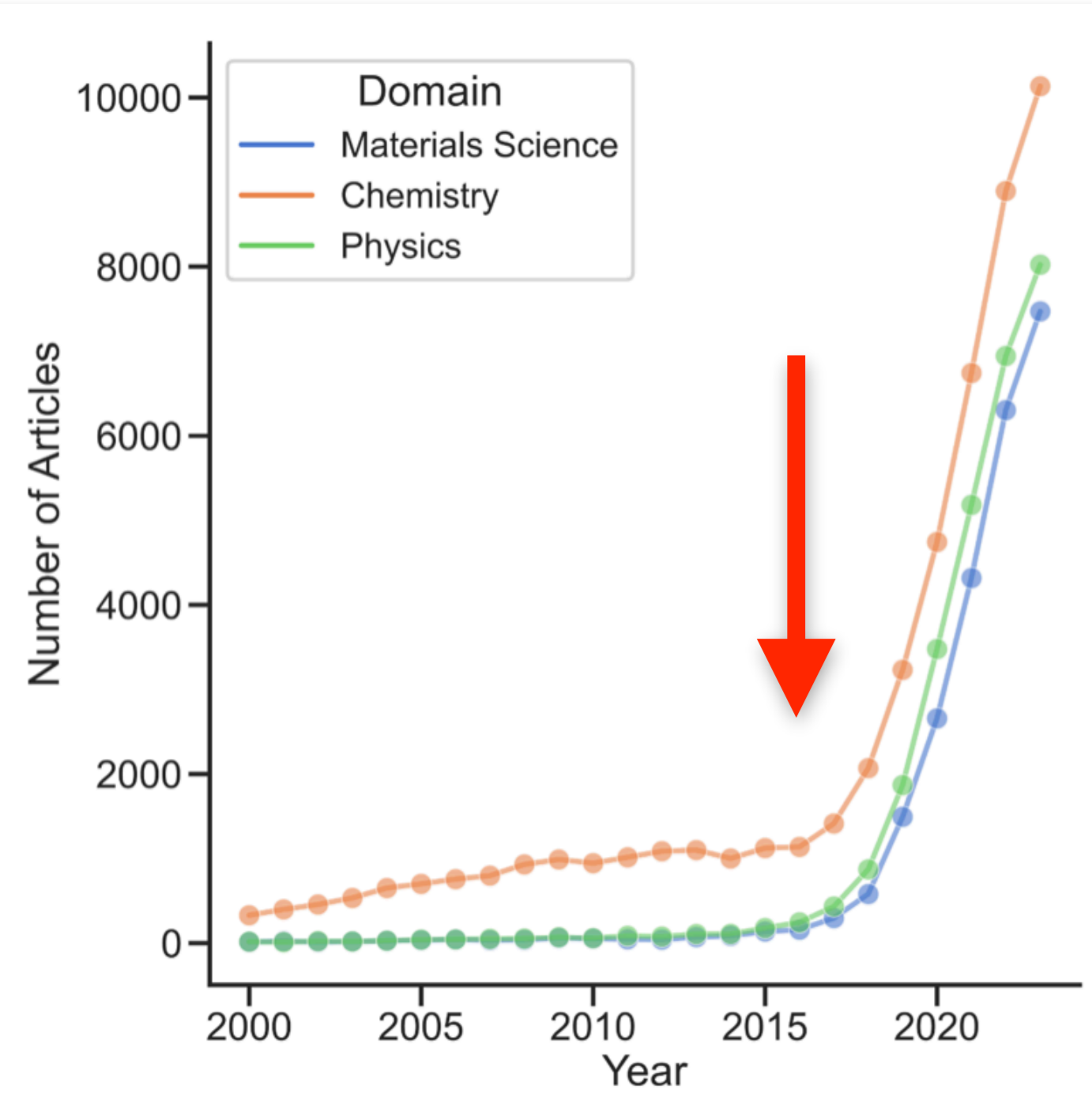
Application of ML to **non-event (meta-)data** might help to **increase efficiency** and **reduce the need of manpower** in Ops, e.g. automating specific tasks, developing intelligent/adaptive systems, ultimately acting on the full chain - from data collection to data analysis - and make it more agile

- Detector control
- Data quality monitoring
- Operational intelligence
- Predictive maintenance
- ...



# ML/DL in HEP

“Traditional” ML



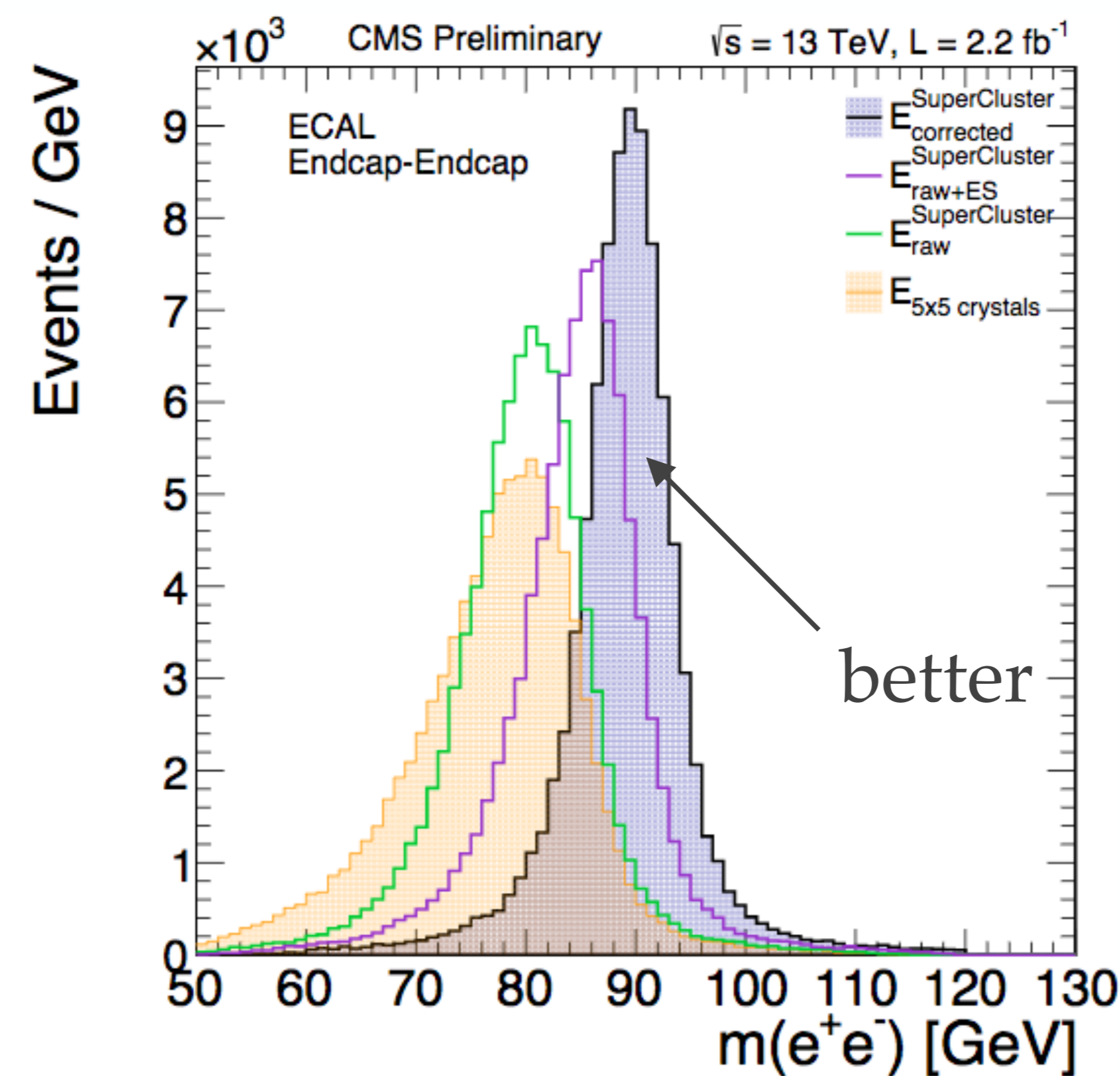
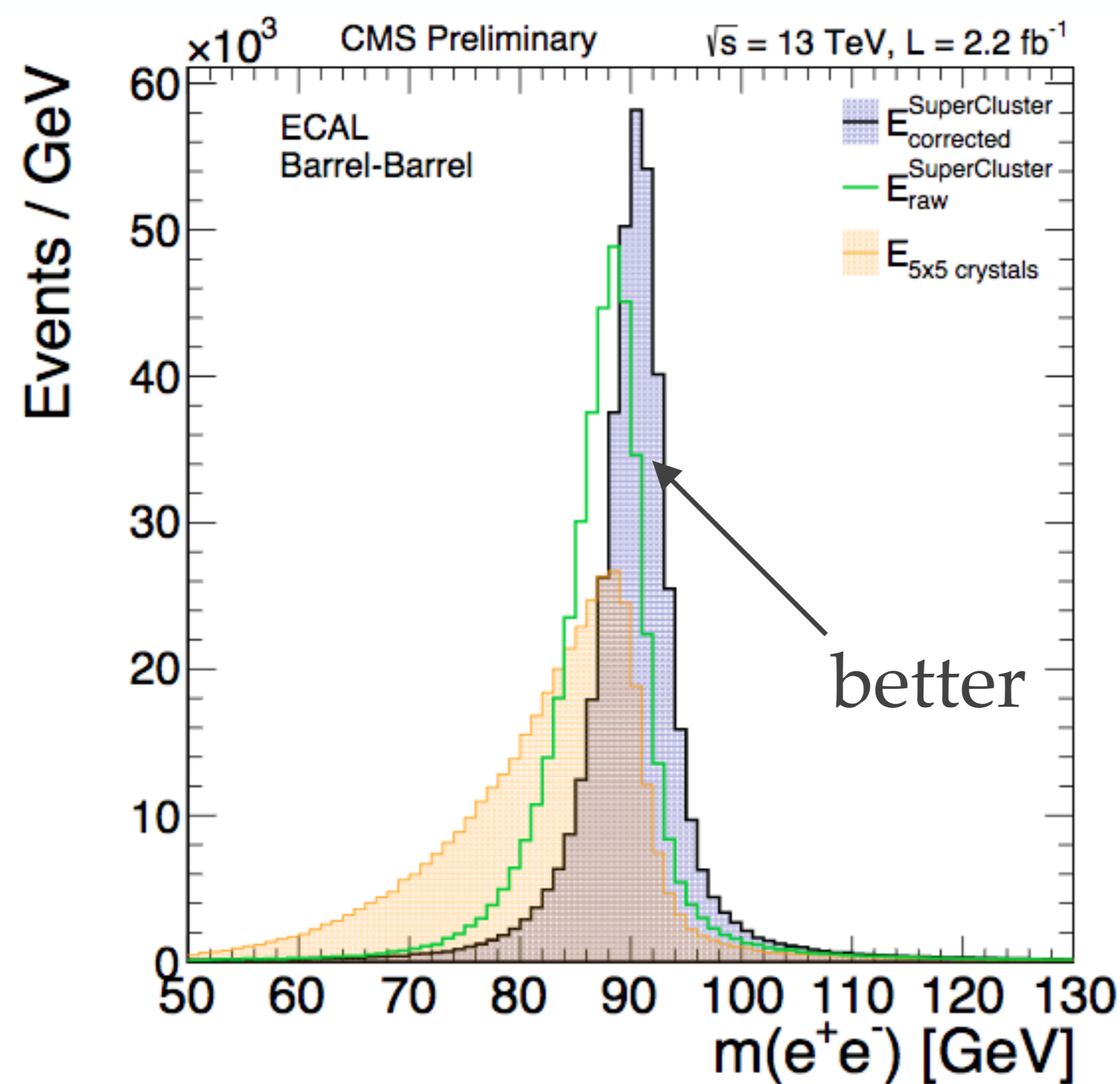
ML in HEP started by using domain knowledge to perform feature extraction/engineering

- HEP physicists design high-level features, and send them as input to traditional ML “shallow” algorithms

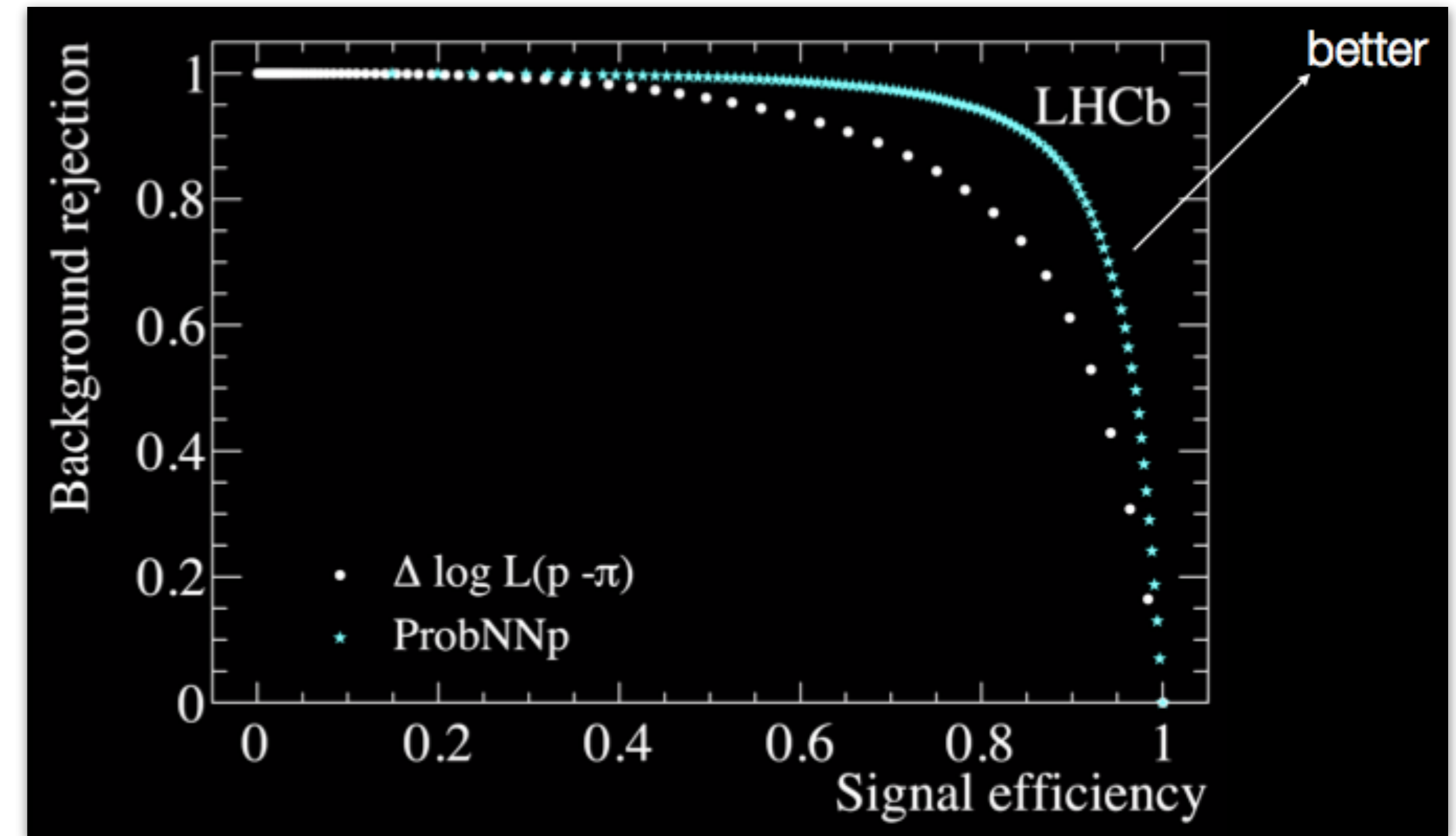
Using ML to improve the determination of particle properties is now commonplace in **all LHC experiments**

E.g. energy deposited in calorimeters is recorded by many sensors, which are clustered to **reconstruct the energy** of the original particle

- e.g. **CMS** trained **BDTs** to learn corrections using all information available in the various calorimeter sensors - thus resulting in a **sizeable improvement in resolution**



[ 2015 ECAL detector performance plots, CMS-DP-2015-057. Copyright CERN, reused with permission ]



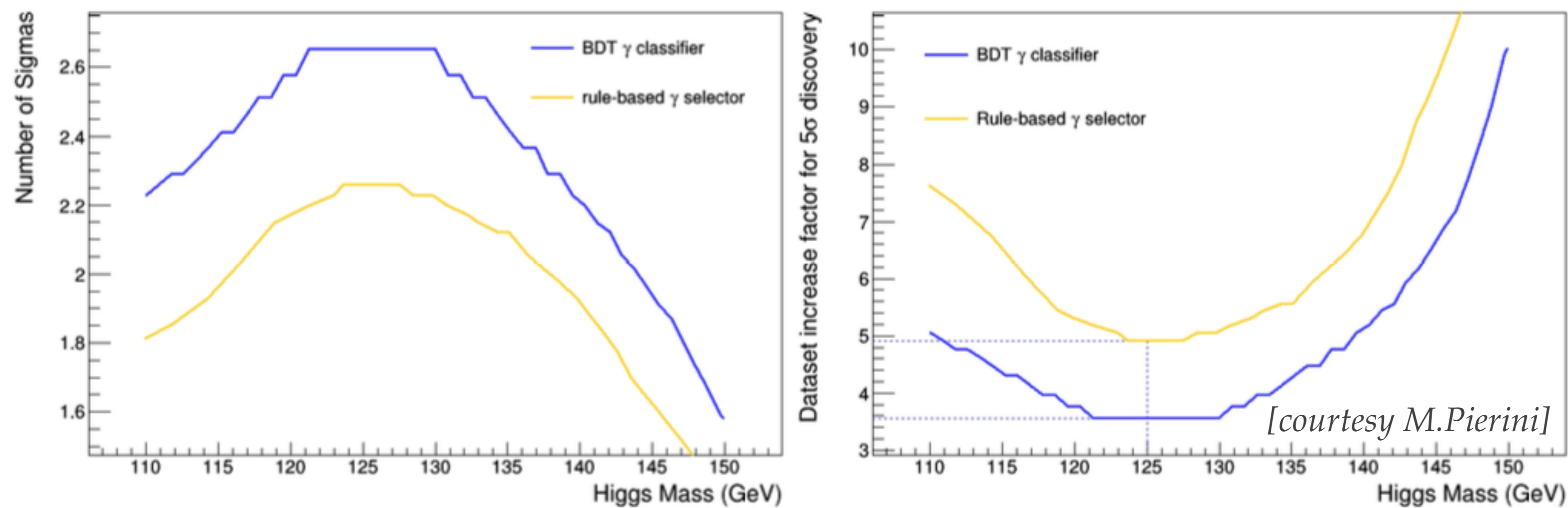
[courtesy: M.Williams]

Similarly, ML is commonly used to **identify particle types**

- e.g. **LHCb** used **NNs** trained on  $O(30)$  features from all its subsystems, each of which is trained to identify a specific particle type
- $\sim 3x$  less mis-ID bkg /particle**. Further estimates indicated already that **more advanced algorithms may reduce bkg by another  $\sim 50\%$**

## ML played a key role in the **discovery of the Higgs boson**

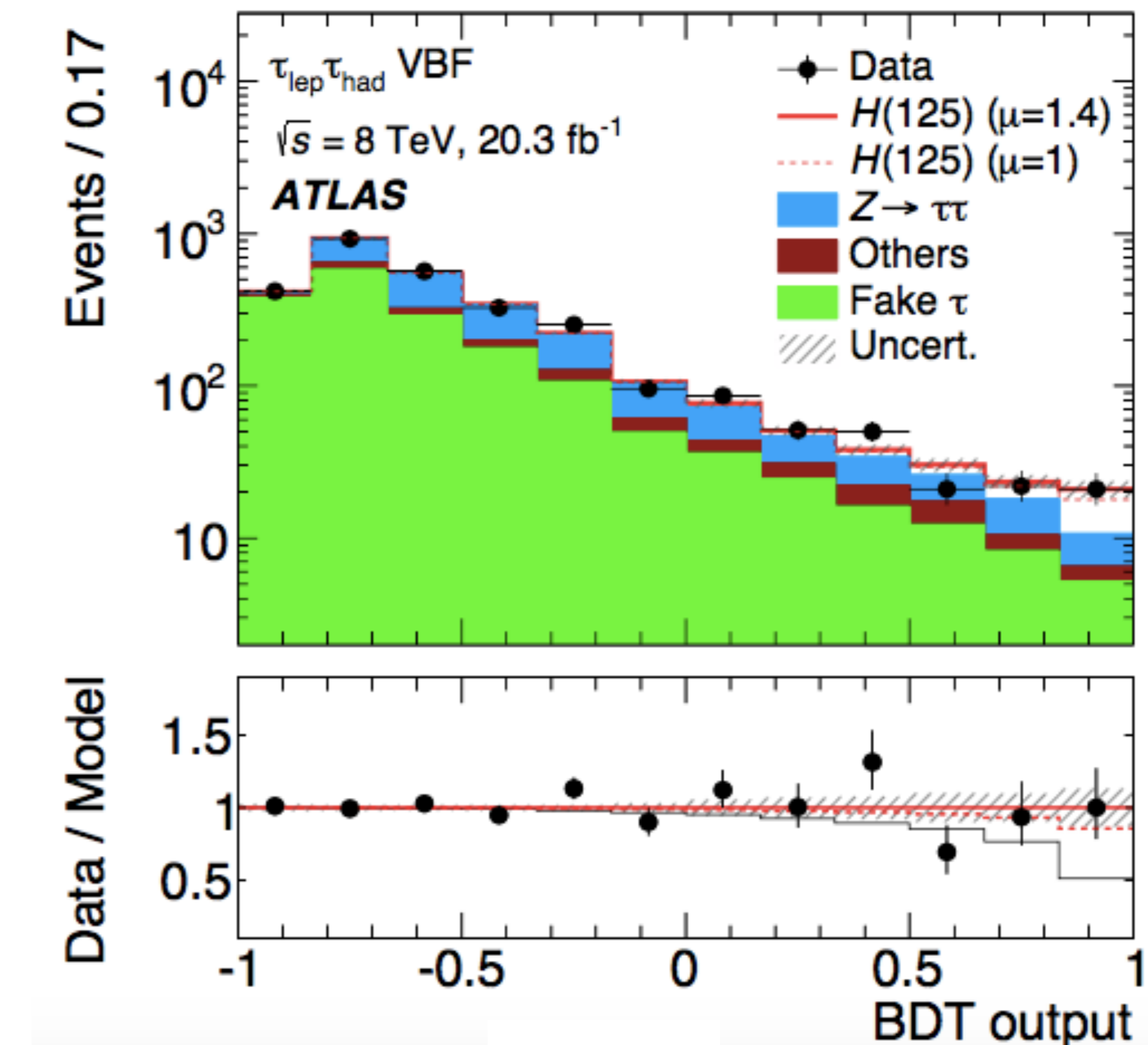
- especially in the diphoton analysis by **CMS**, where **BDTs** (used to improve the resolution and to select/categorise events)
- → **sensitivity increased by roughly the equivalent of collecting ~50% more data**



We were not supposed to discover the Higgs boson **as early as 2012**

- Given how machine progressed, we expected discovery by end 2015 / mid 2016

We made it earlier thanks (also) to ML



ML impact also on the **study of Higgs properties**

→ e.g. analysis of  $\tau$  leptons at LHC complex, as they decay before detection + loss of subsequently produced neutrinos + bkg from Z decays

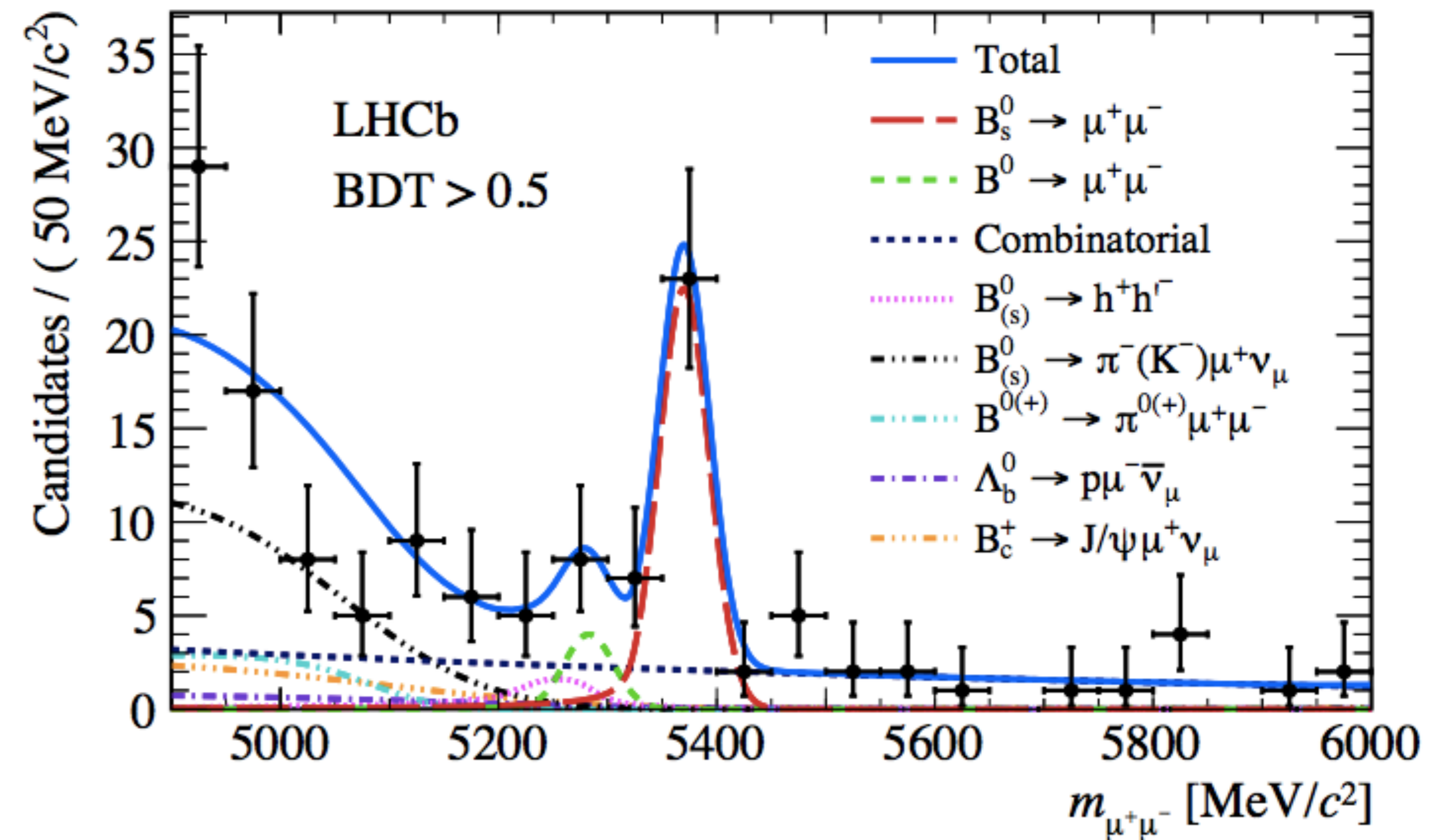
- e.g. **ATLAS** divided the data sample into 6 distinct kinematic regions, and in each a **BDT** was trained using 12 weakly discriminating features [1] → **improved sensitivity by ~40% vs a non-ML approach**



**CMS** and **LHCb** were the first to find evidence for the  $B^0_s \rightarrow \mu^+ \mu^-$  decay with a combined analysis [1] (as rare as  $\sim 1 / 300$  billion pp collisions..)

- **BDTs** used to reduce the dimensionality of the feature space - excluding the mass - to 1 dimension, then an analysis was performed of the mass spectra across bins of BDT response
- decay rate observed is consistent with SM prediction with a precision of  $\sim 25\%$ , placing stringent constraints on many proposed extensions to the SM

Mass distribution of the selected  $B^0 \rightarrow \mu^+ \mu^-$  candidates with  $BDT > 0.5$  [2]



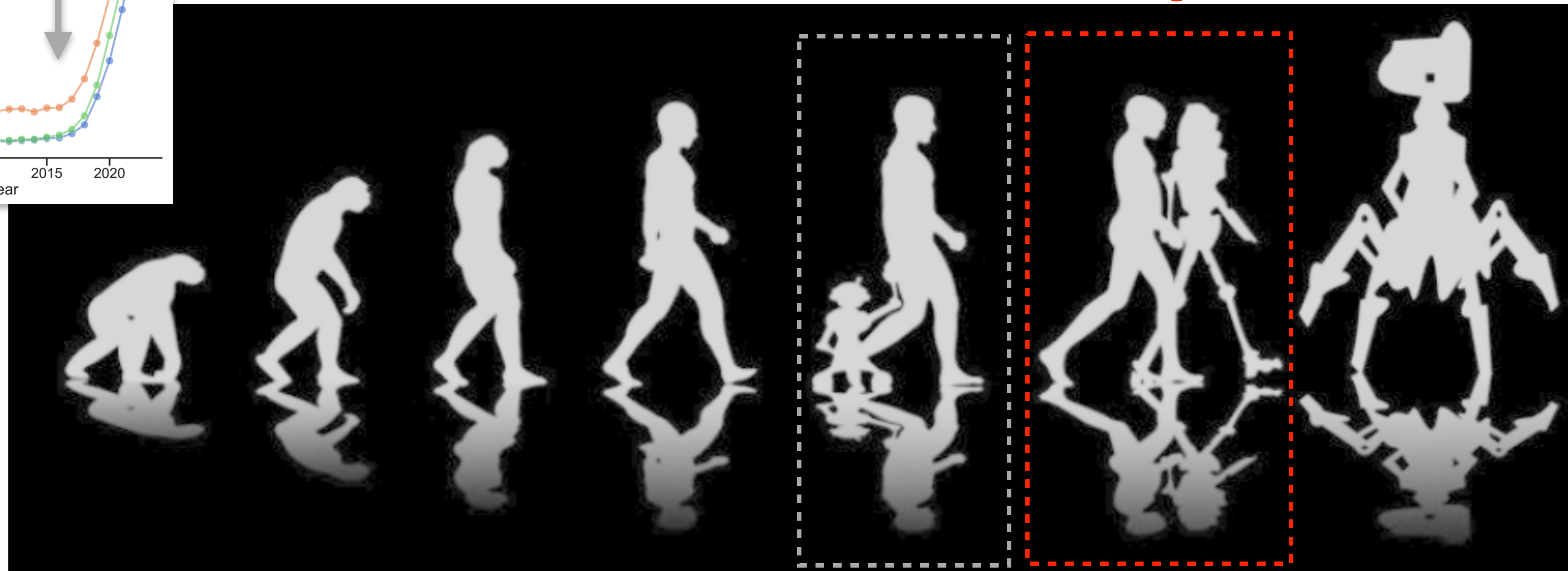
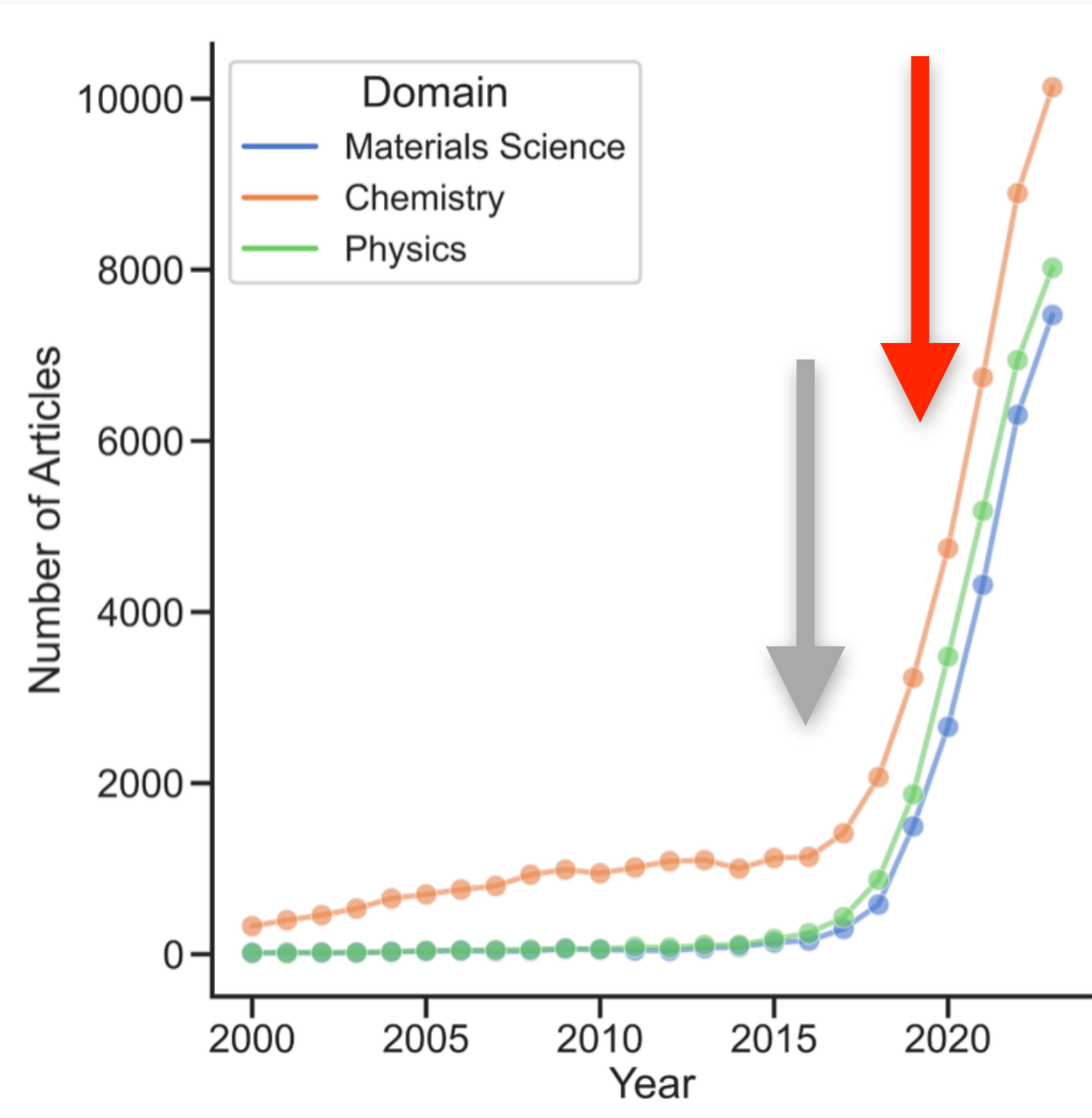
To obtain the same sensitivity without ML by LHCb as a single experiment would have required  $\sim 4x$  more data

[1] Nature 522 68–72 (2015)

[2] Phys.Rev.Lett. 118 (2017) 19, 191801

# ML/DL in HEP

“Traditional” ML      Seeking DL solutions



Soon, ML (DL) in HEP started to seek for more advanced techniques, e.g. [deep NNs](#)

- **Use all the features space at its full dimensionality to train deep NN - no more manual feature engineering**

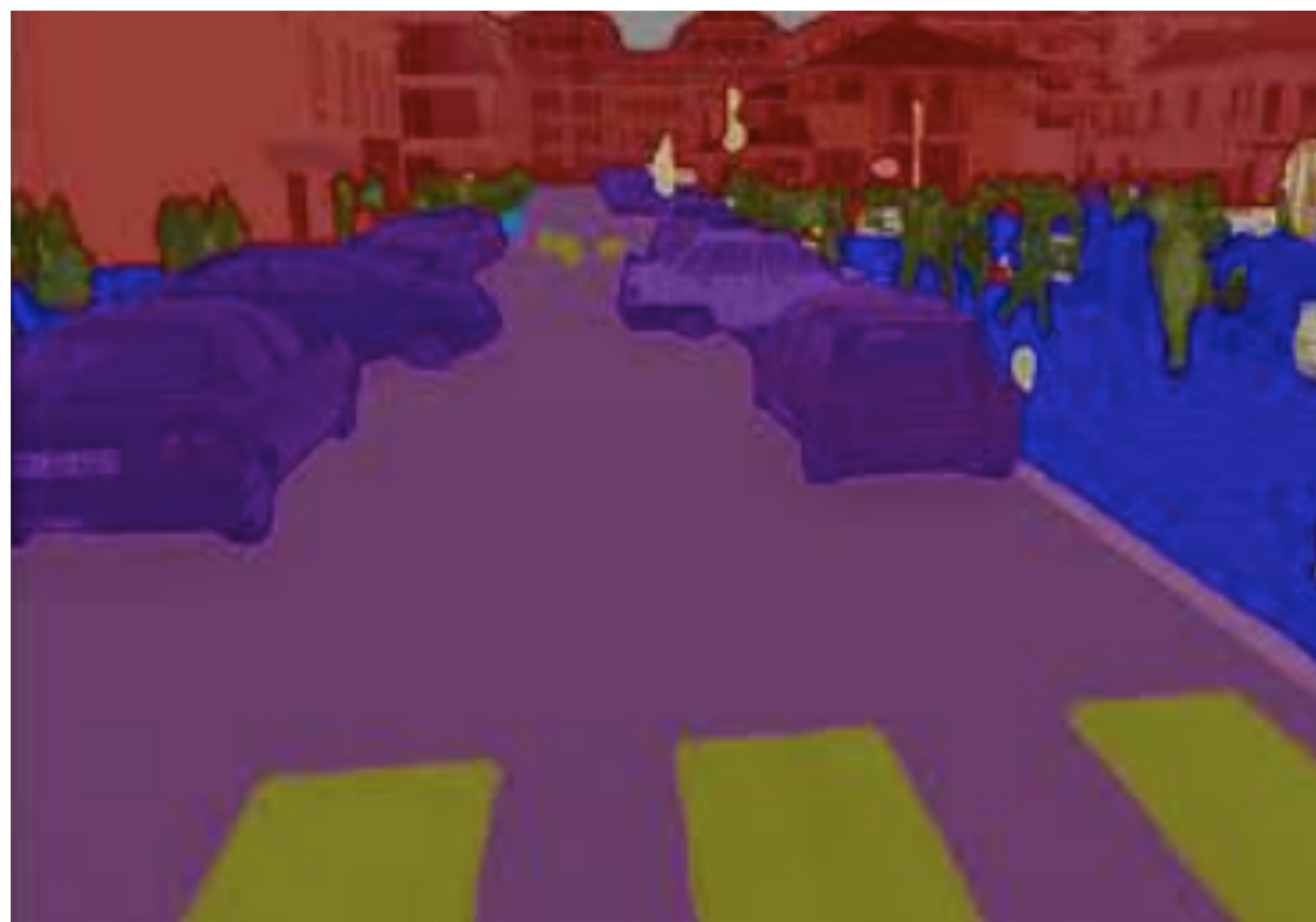
→ extract best from data, and do so by exploiting any architecture that might work for a given use-case (e.g. input from CV and NLP solutions..)

**CNNs** offer translational-invariant feature learning, robustness against noise, versatility in application to a variety of domains

- extremely vast zoo of architectures! Primary target: computer vision
- They are based on sequences of convolutional and pooling layers, and additional tricks

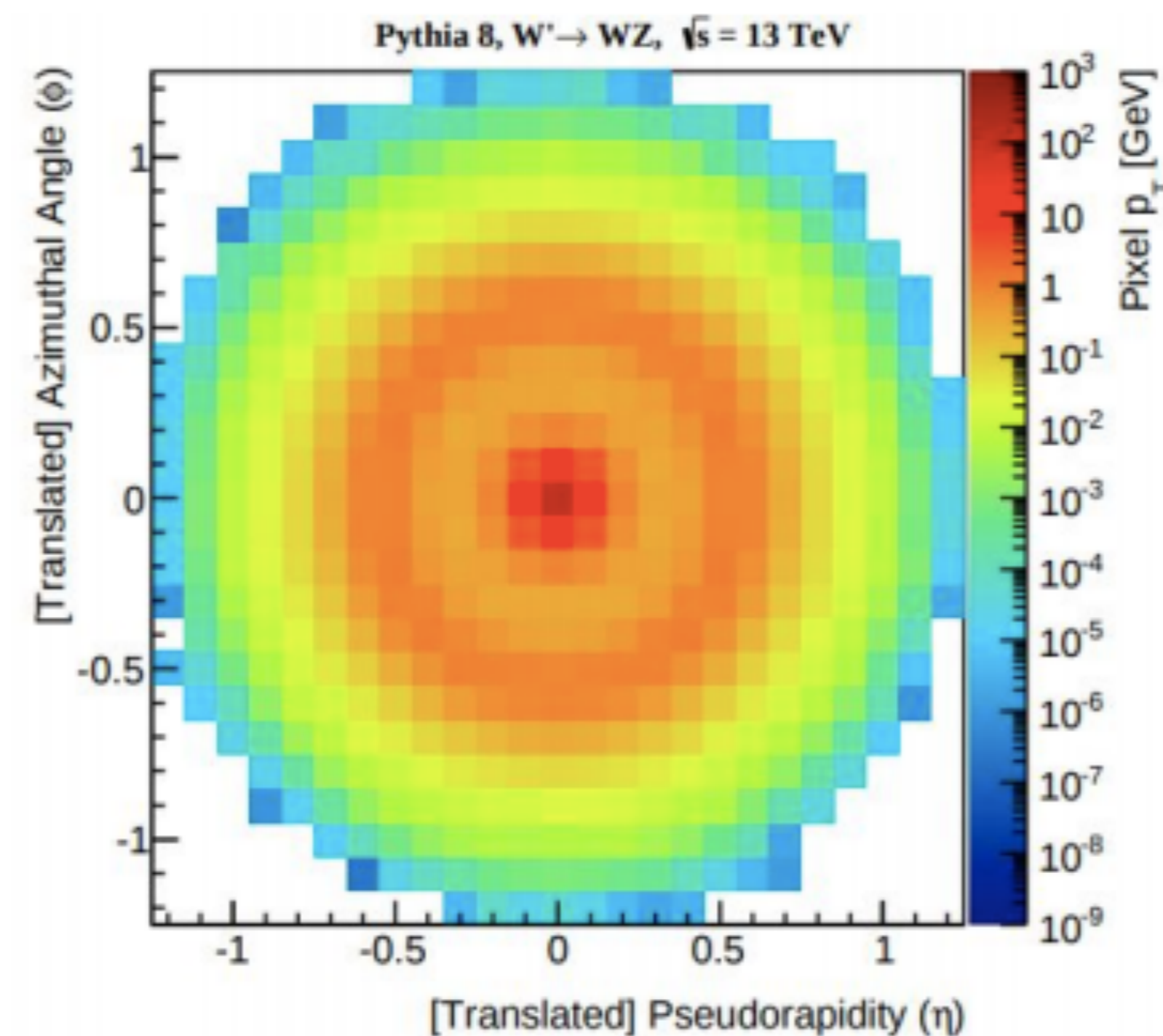
### Industry:

Large adoption in computer vision applications (e.g., self-driving cars, ..)



### HEP:

3D imaging in detectors, event classification, automation of hist checking (e.g. data quality), ...



General tactics: (TPCs, CALOs..): represent your data as 2D/3D images (even 4D w/ timing info)

→ **problem casting into a computer vision task**

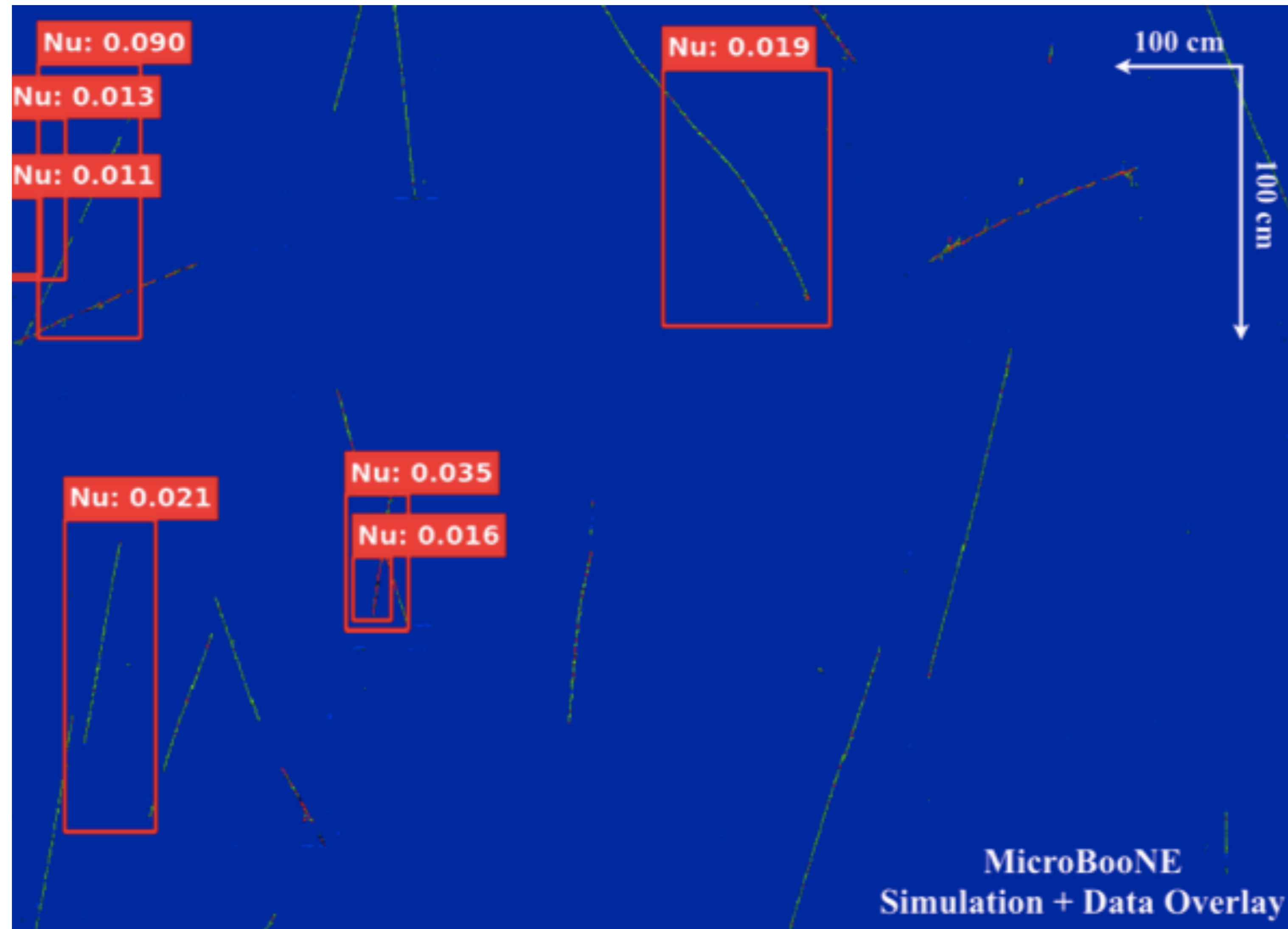
**Example**

Heavy-flavour (b/c) jets in LHC collisions @13 TeV → **CMS** DeepCSV algorithm: **immediate +15% improvement in efficiency w.r.t likelihood-based methods**

[1] JINST 13 (2018) P05011

[2] JINST 15 (2020) 12, P12012

# "HEP is so different from other applications". Is it?



Detection of **neutrinos** on cosmic background event (method: **CNN**)



Detection of **airports** from satellite images (method: **CNN**)

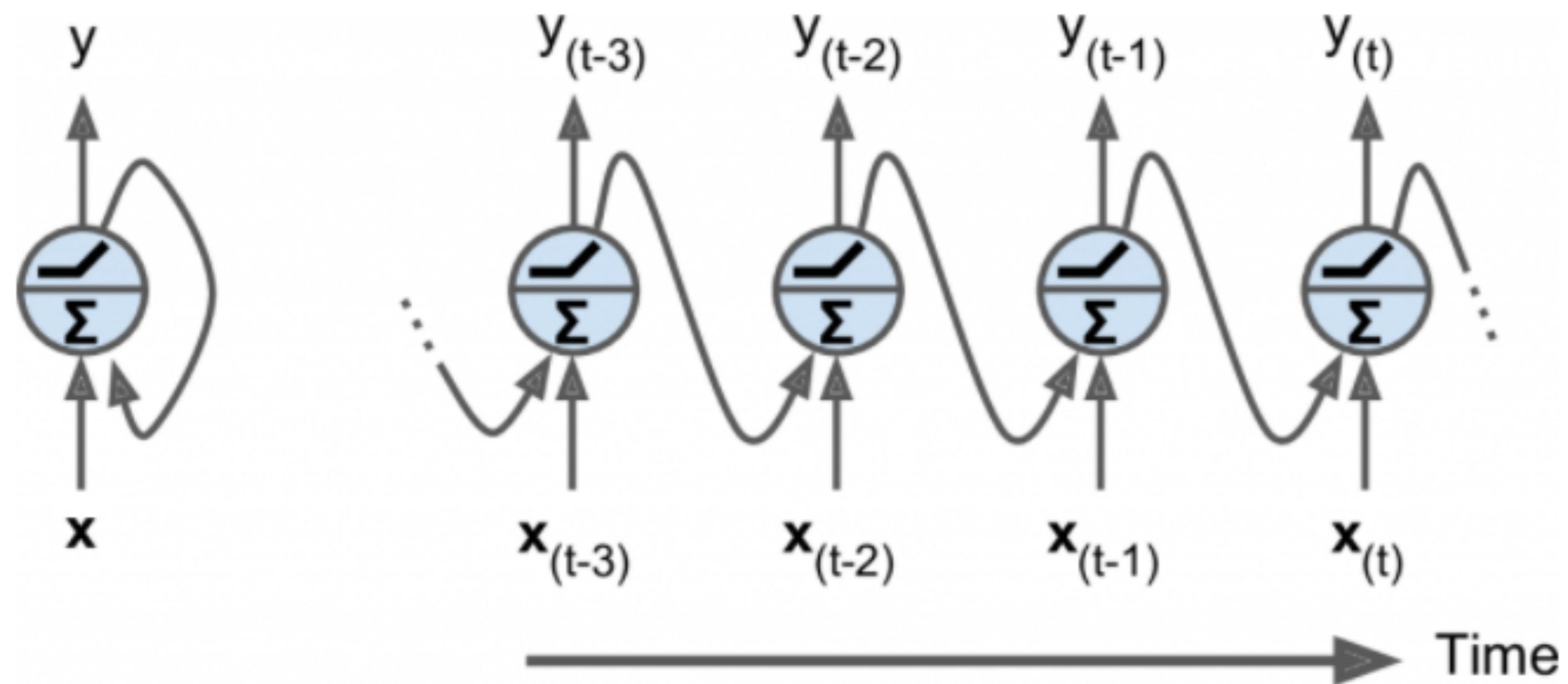
**RNN** allow to handle variable-length inputs and process time-series, accumulating and using together info at various times in the sequence

- Based on “recurrent neurons” (backward-pointing connections)
- A variety of application in time-series of all kinds, e.g. language translation, ..

**Industry:**

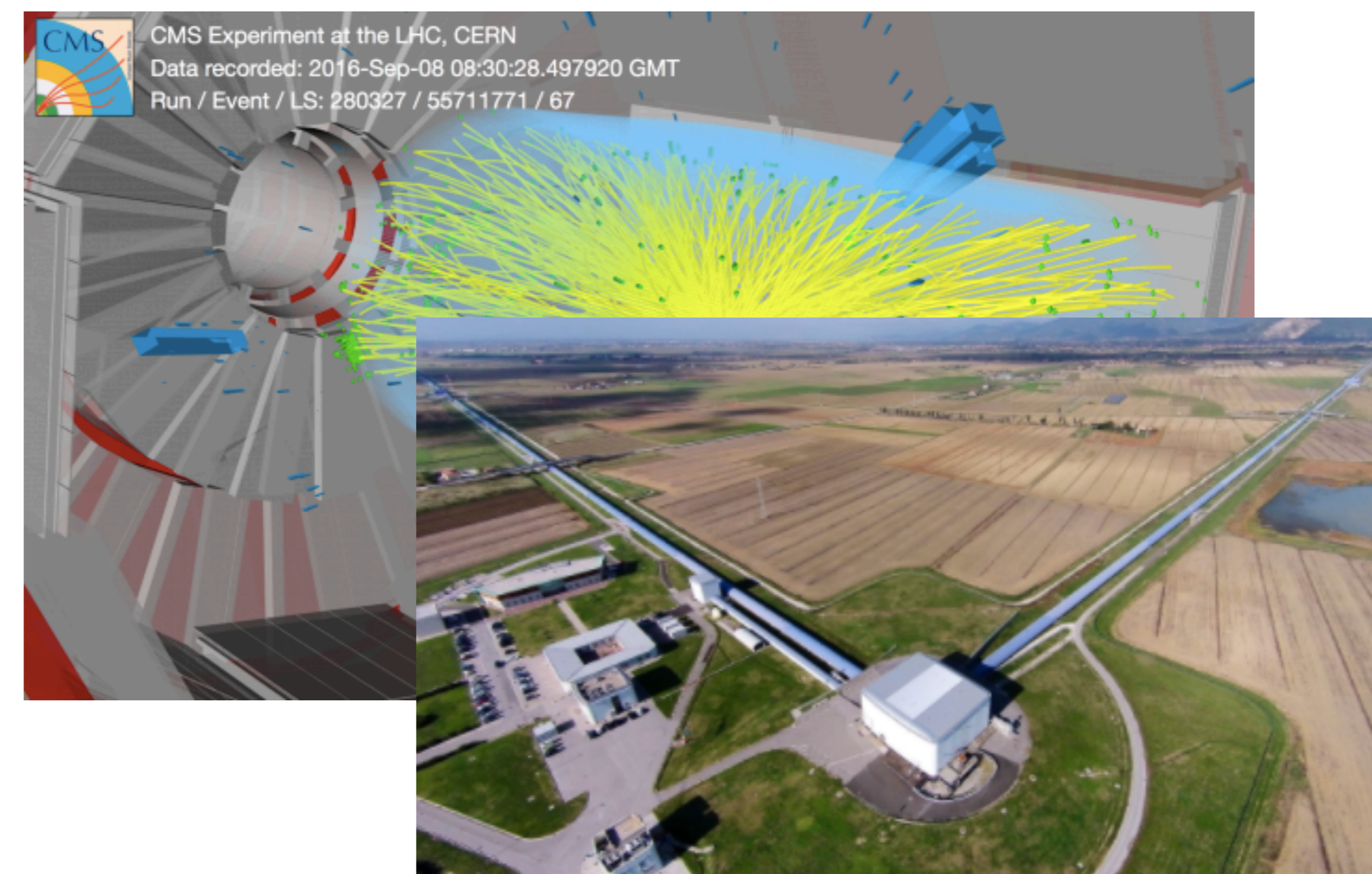
handling “time series”

(audio, video, natural language processing)



**HEP:**

Classifiers capable to process variable-length signals of different nature (tracks, particles in jets, etc) - ample application in astro-particle physics



**Example**

RNN-based b-tagging able to exploit low-level features of particles in jets → **ATLAS mis-identification rate reduced by ~4x** w.r.t non-ML algo

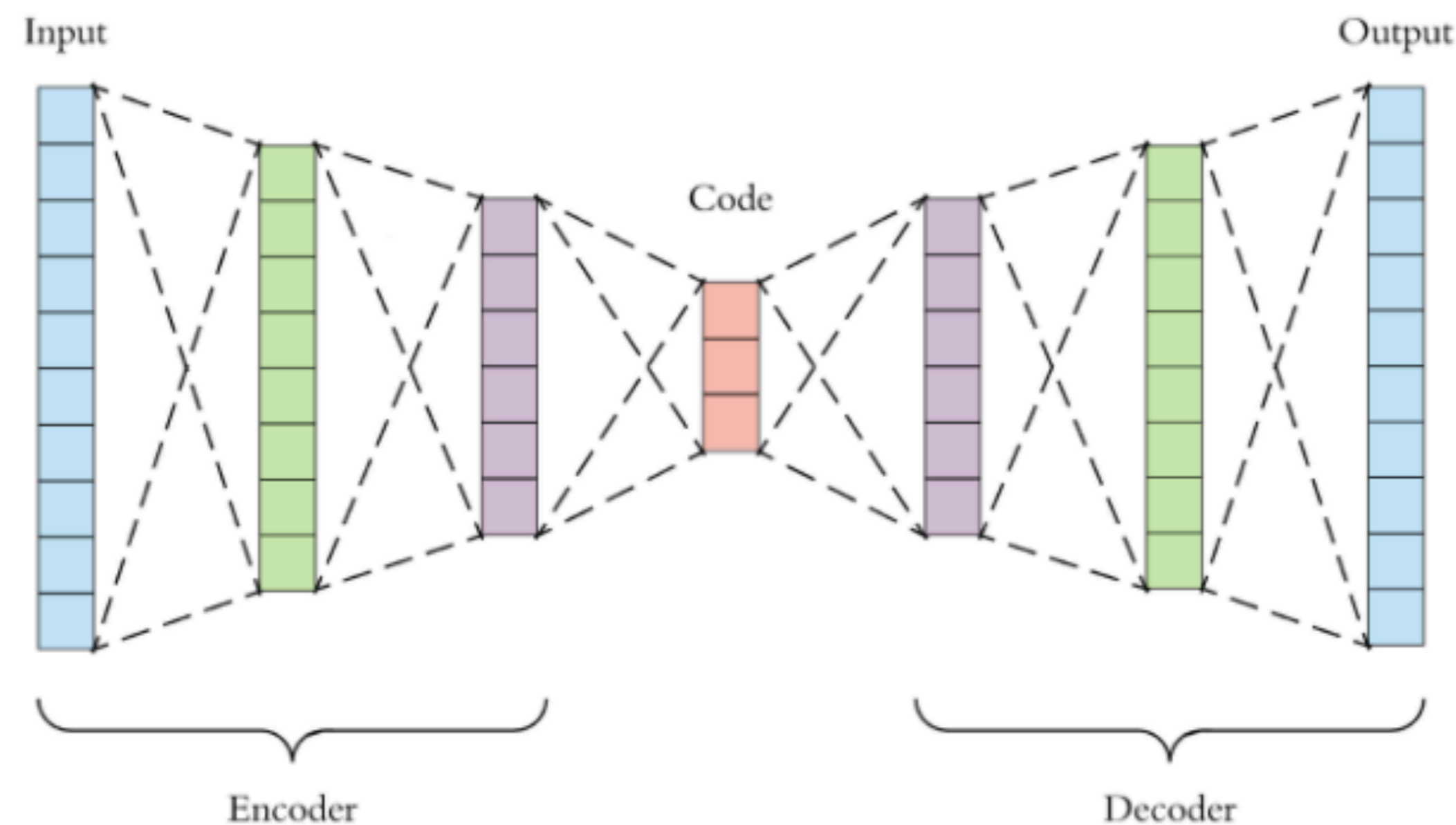
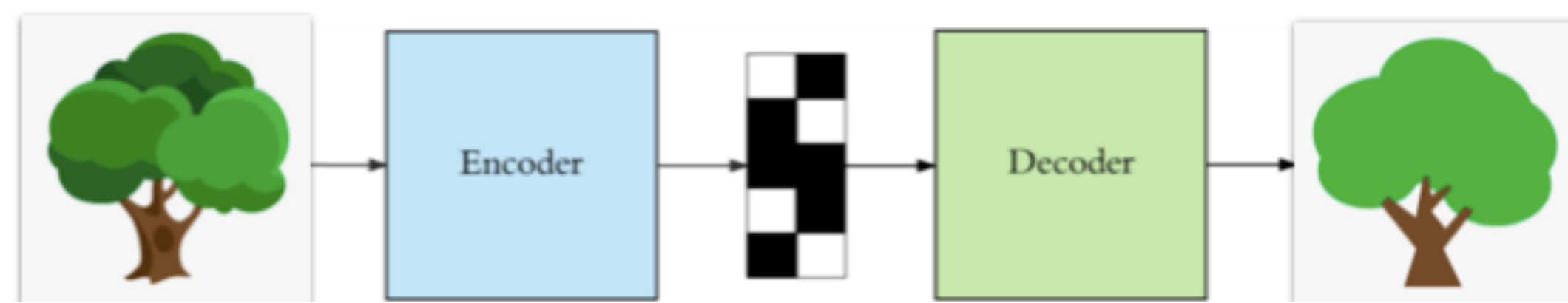
[1] ATL-PHYS- PUB-2017-003 and 013  
 [2] CMS-DP-2017-005  
 [3] Phys. Rev. D 94, 112002 (2016)

**AE** is a “data-specific” compression algorithm, able to reduce dimensionality and extract “the juice” of an input

- a feed-forward (**un/self-supervised**) NN capable to encode the input into a reduced-dimensionally representation (“latent space”) and decode it in output

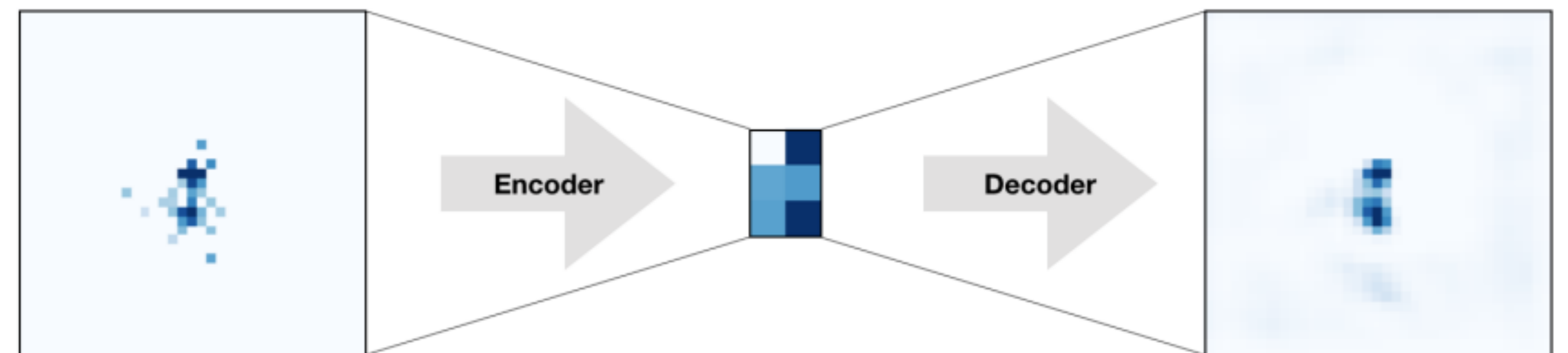
### AEs in Industry:

dimensionality reduction (like PCA),  
clustering, denoising, ...



### AEs in HEP:

anomaly detection (interesting events are those whose decoding in output is distant from the input, according to a given metric)



Potentially, a powerful tool to discover new physics in a “unsupervised” manner

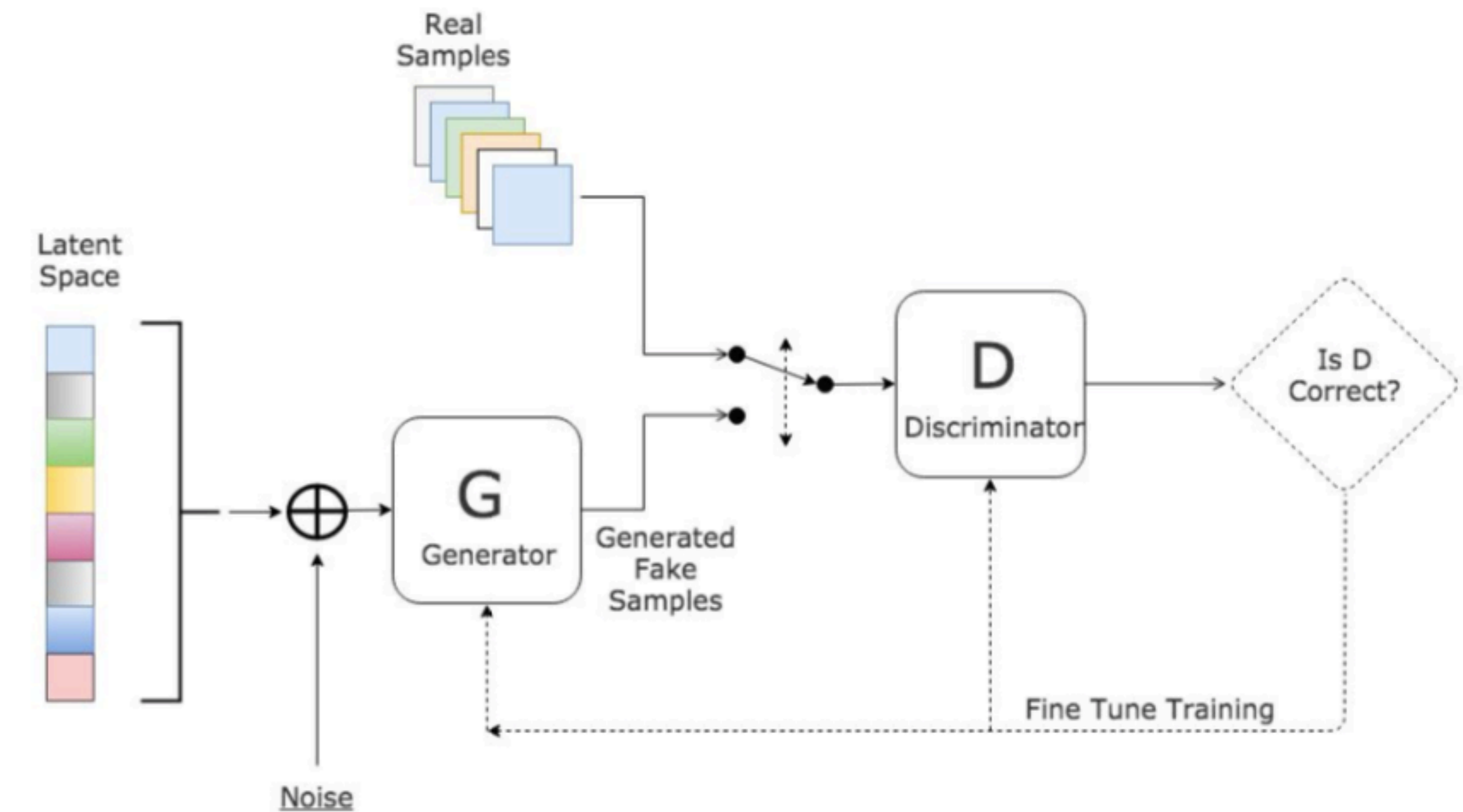
A **Variational AE** has been introduced (**CMS**) for “new physics mining” [1]

- Training on known SM processes, build threshold to identify “anomalous” (i.e. interesting: BSM?) events
- Treat them as outliers, save them (no trigger kill!), build a catalog for further inspection
- Model-independence: training not dependent from specific new physics signatures → assumptions-free
- Might be complementary to classical methods, which are i.e. model-dependent hypothesis testing
- Recurrent topologies in the catalog might inspired focussed searches, as well as standing as input towards building new theoretical models
- Target (not easy): up to the trigger level.. and in production!

DISCLAIMER: at the threshold between **discriminative AI** and **generative AI** ...

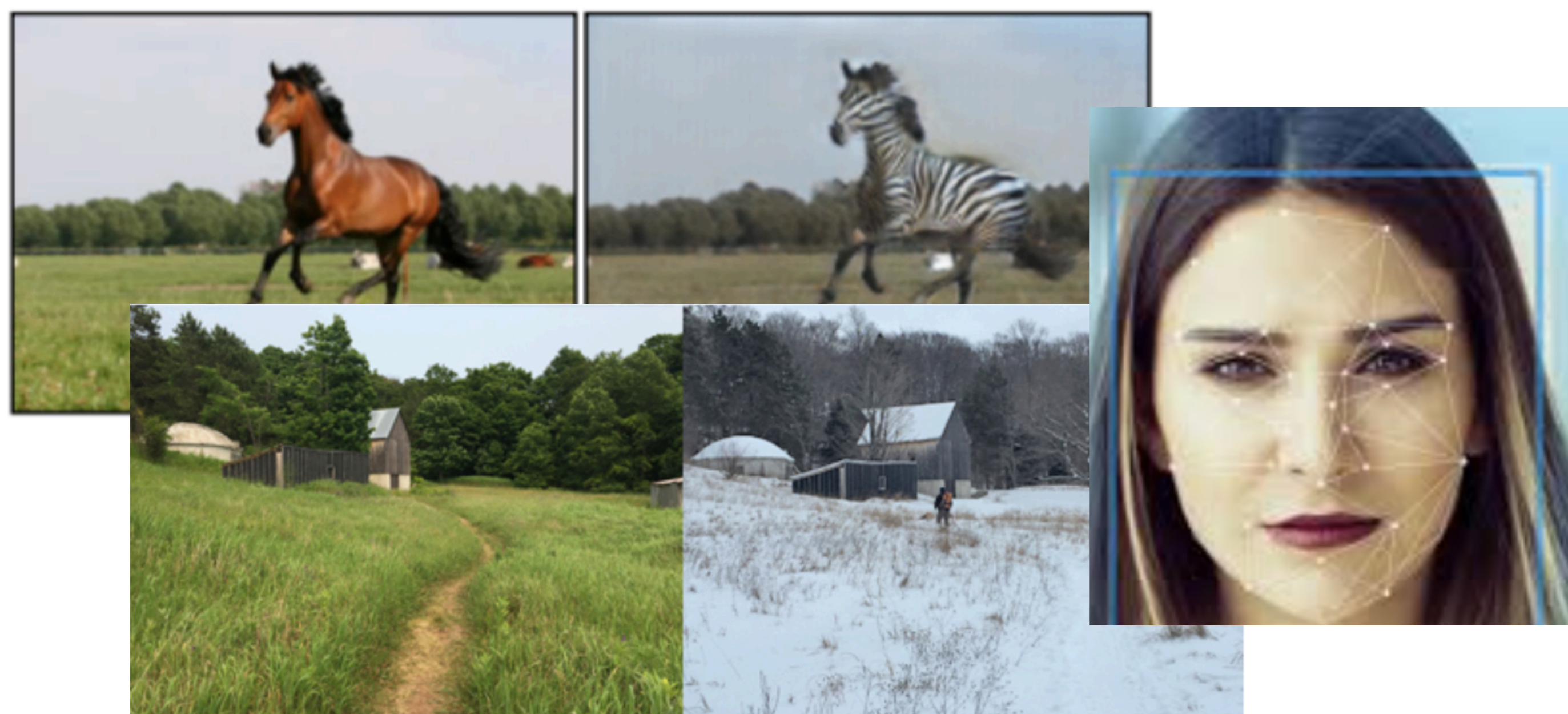
A generative algorithm, based on an architecture with 2 NNs, a generator G and a discriminator D, which compete

- G creates images from noise, D classifies them real vs fake
- Once trained one against the other, G pursues its goal which is to confuse D, and in the process it learns how to create fake but very realistic images



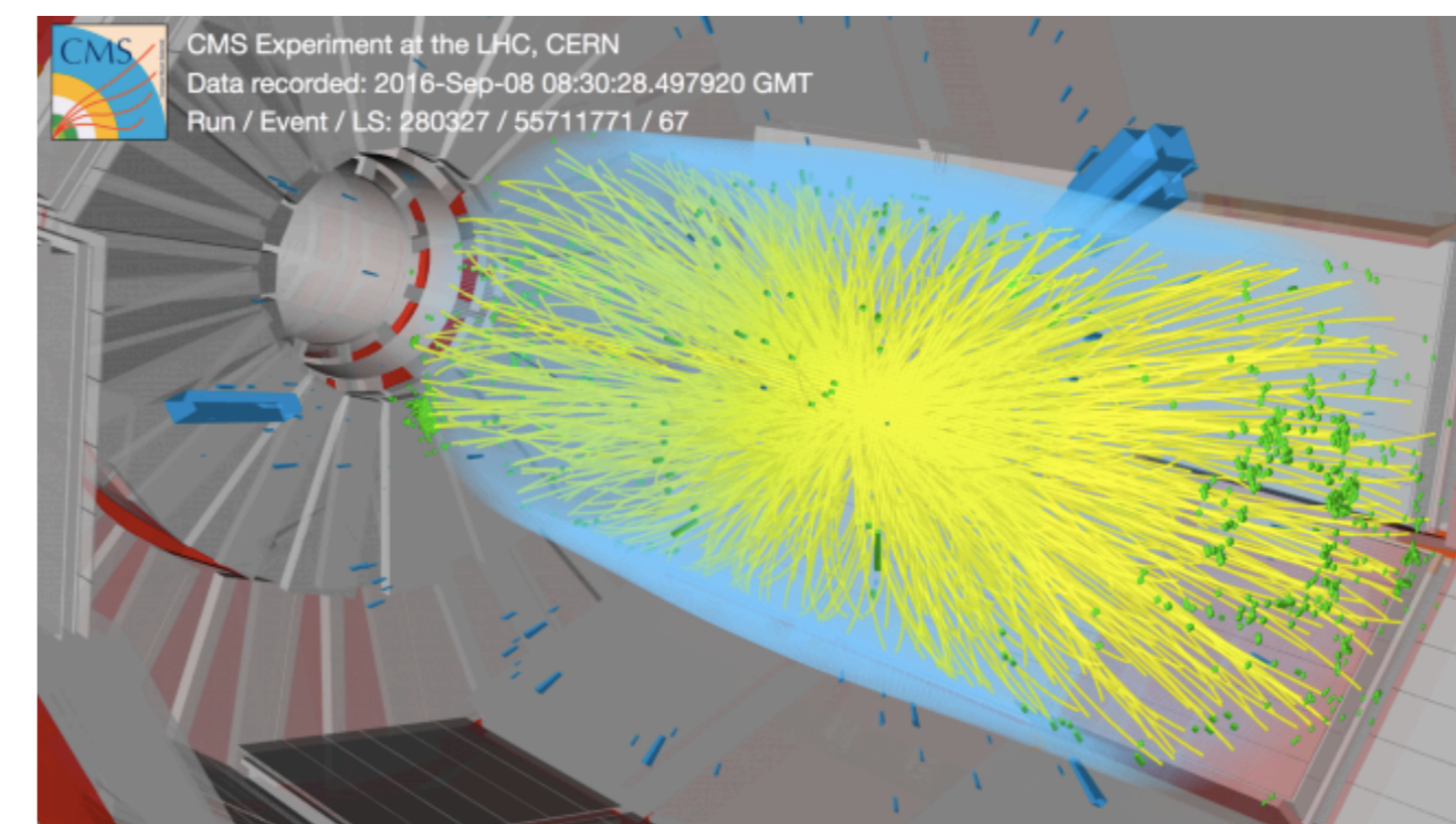
### Industry:

image editing, data generation, security, ..



### HEP:

Simulate the detector response at reduced computational costs



### Example

**CaloGAN**, a FastSim techniques to simulate 3D showers in multi-layer ECAL with GANs: can learn and generate the reconstructed calo images with no need to use expensive GEANT and RECO expensive cycles..

→ **10k faster** (still reliability is an issue, though)



# Data sparsity and point clouds

HEP handles **high sparsity** datasets (not a HEP-only issue..)

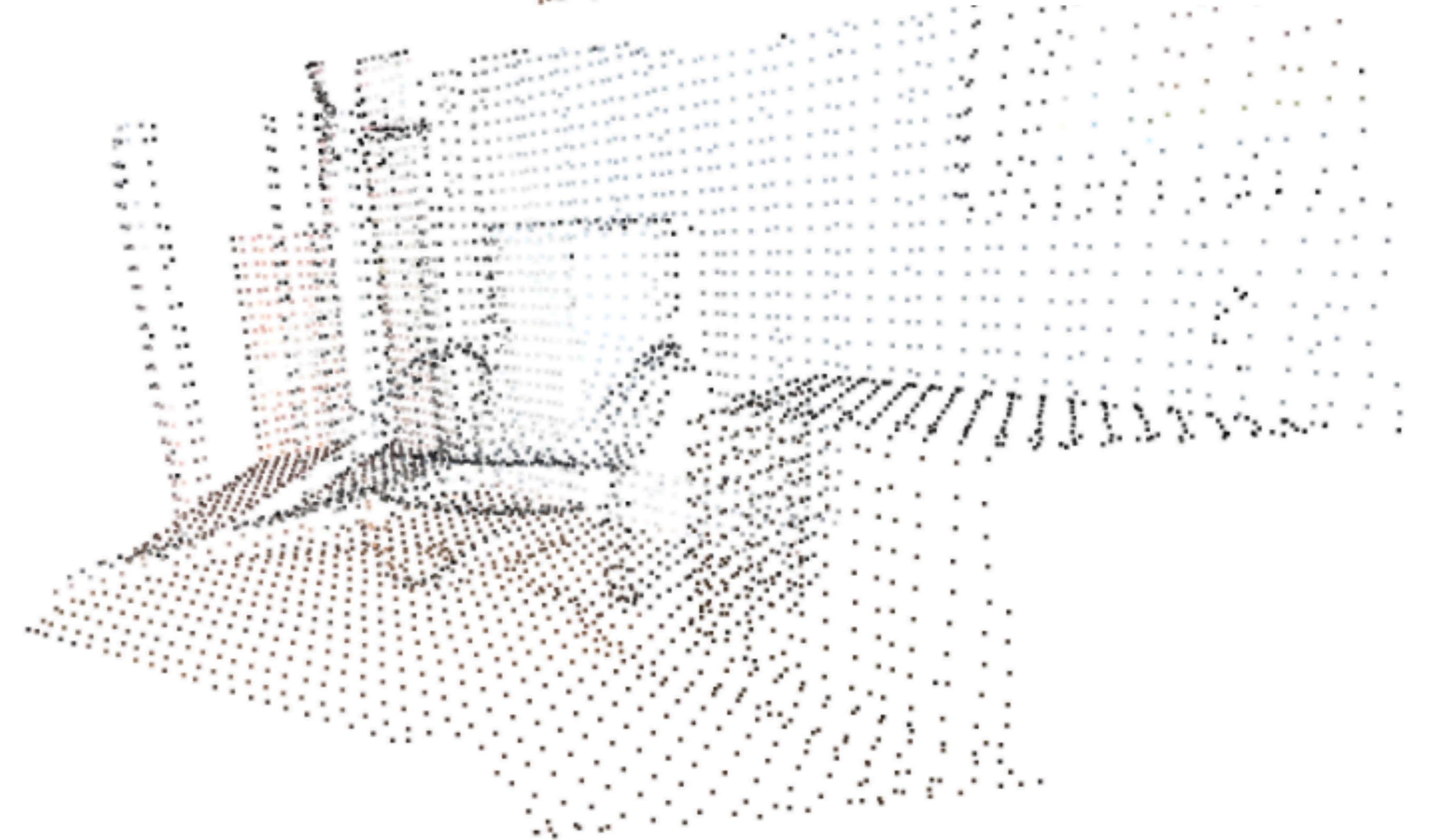
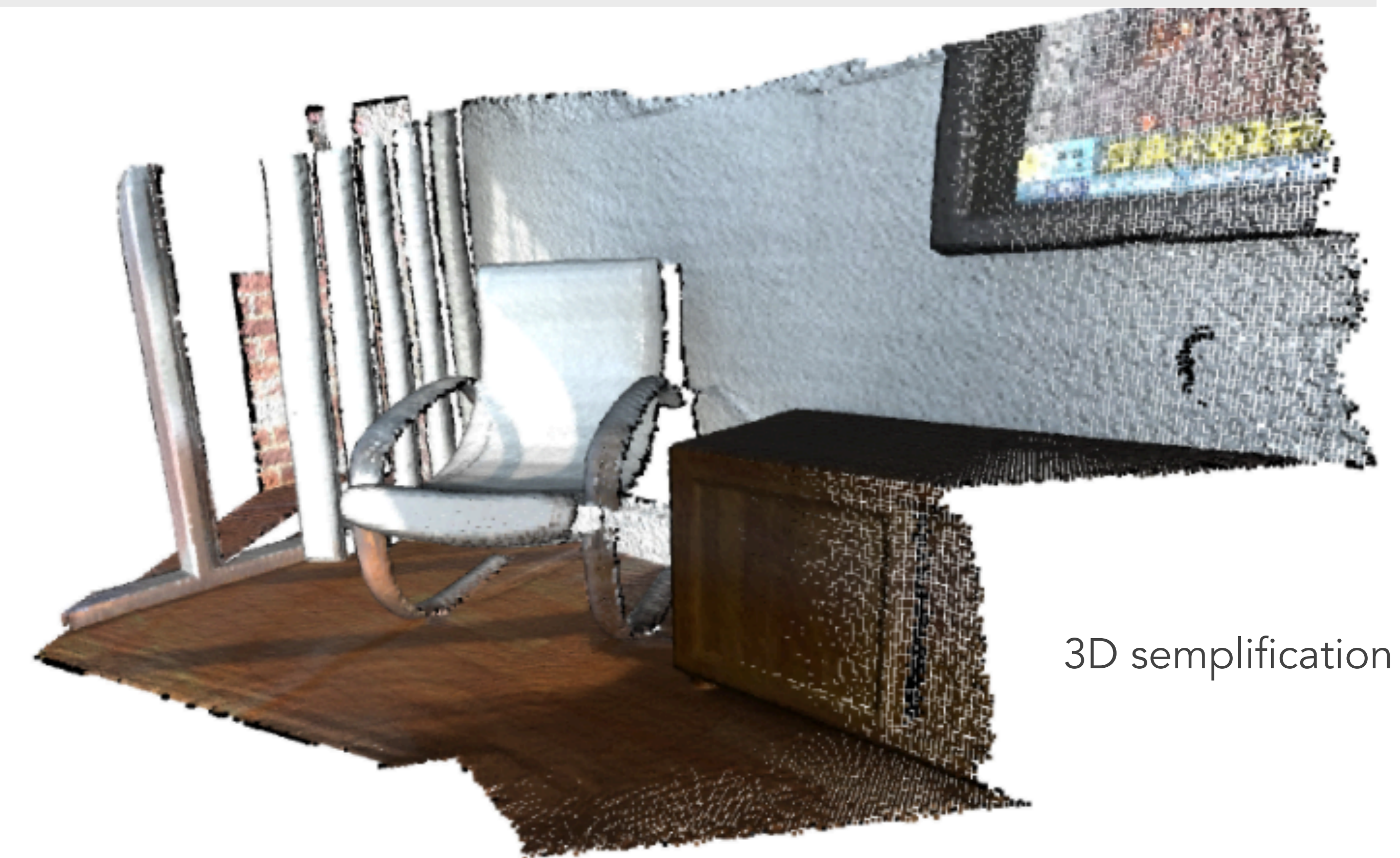
- Granularity and occupancy in HEP sensors
- Population of stars and galaxies in the cosmos
- Molecular description in computational chemistry

Abstract space with coordinate of sparse elements, each characterised by an array of features, a set of arrays as a function of event/run, ...

- e.g. EM shower  $\rightarrow$  E deposited in active volumes of an ECAL

More adequate representation would be a “**point cloud**”, and best approach might be not (HEP-)traditional

- Need to be open-minded towards methods not familiar (so far) to the HEP community
- e.g. problems configurable in extraordinarily similar ways as how social media datasets are treated (!)



# Graph Neural Network (GNN)

Think of a CNN acting on its input features (pixels). Its power resides also on a “regular-array dataset” paradigm

- Data represented as sets of dense arrays/tensors, with intrinsic metrics

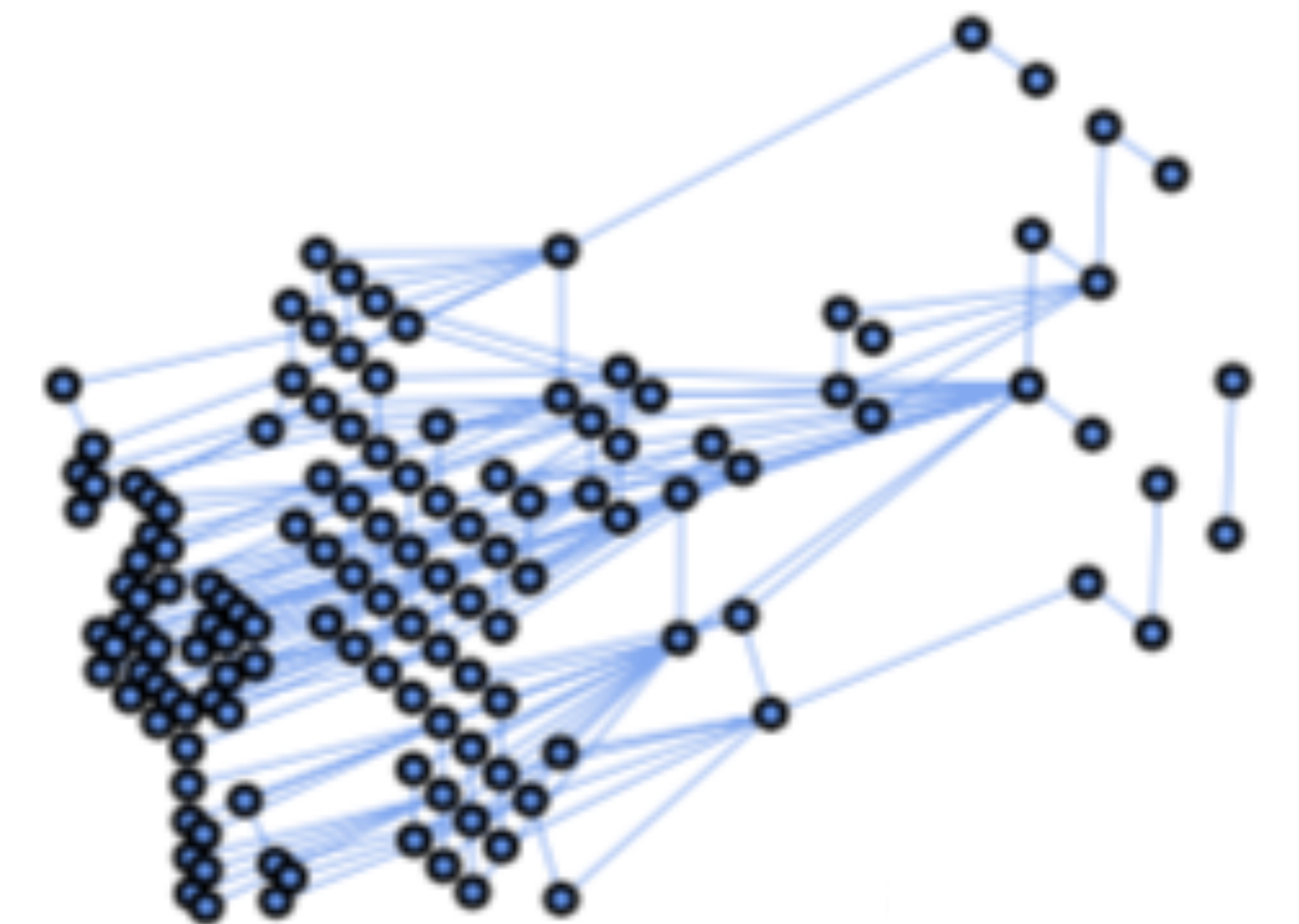
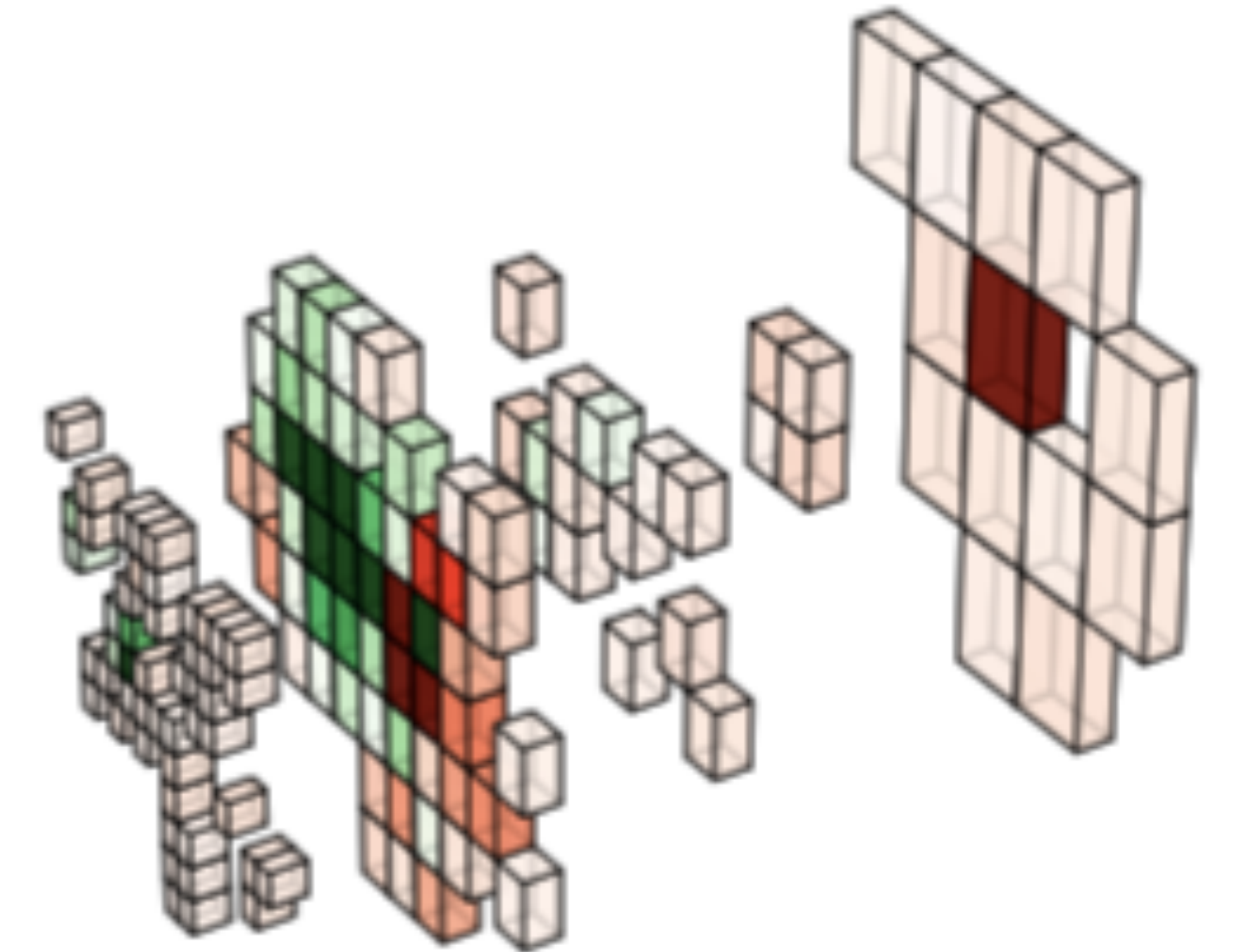
In a sparse representation, we need a metric that defines proximity in an abstract space of features

How? Migrating from “**datasets**” to “**graphs**”

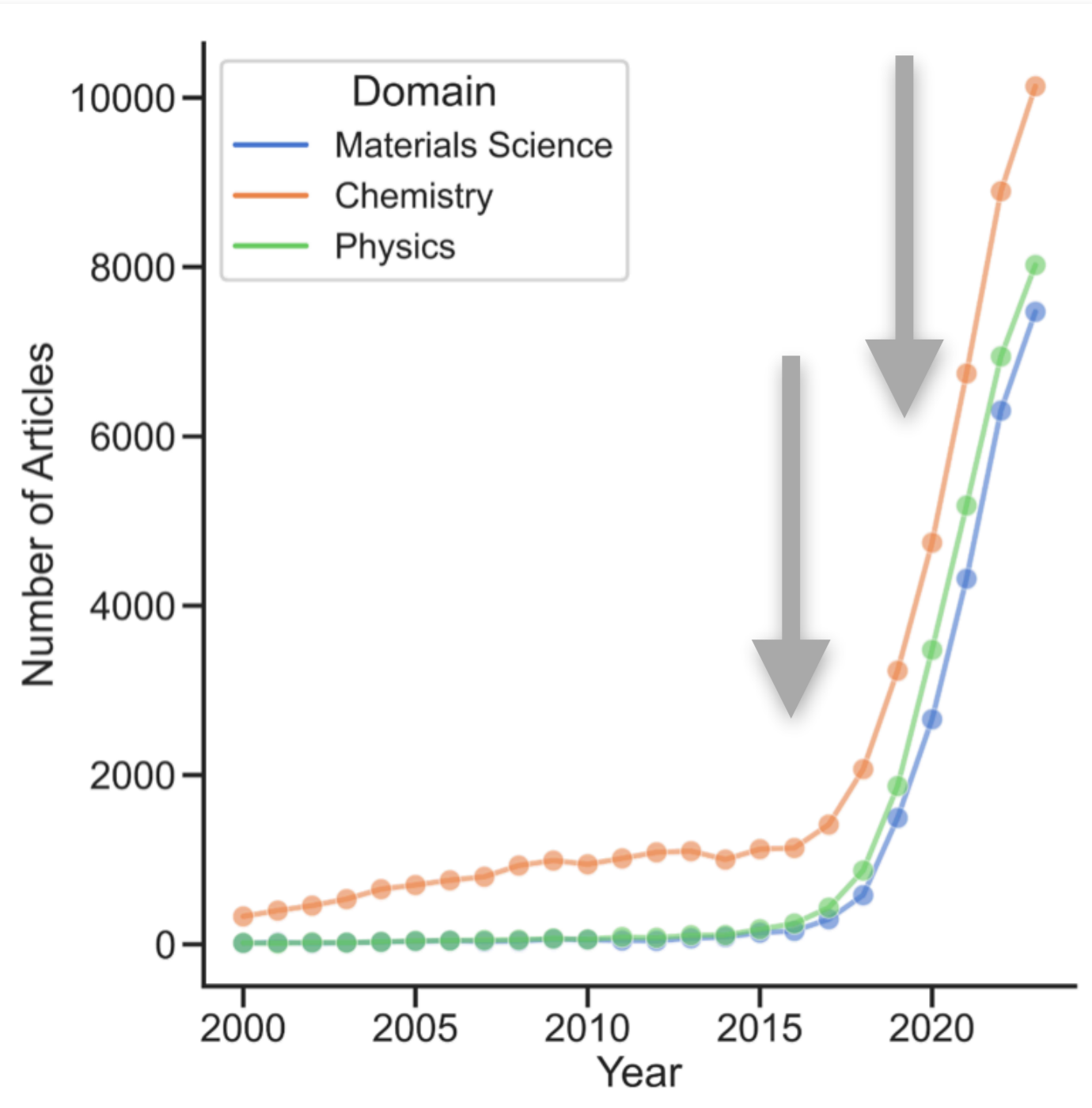
- Connect elements of a dataset and train a NN to learn which are the relevant connections

GNN → build a data structure,  $(V,E)$  with  $V$ =vertex and  $E$ =edges, choose possible types of vertices (if no prior, one builds a fully-connected graph), etc

**GNNs may well grow as an actor in DL applications in HEP**

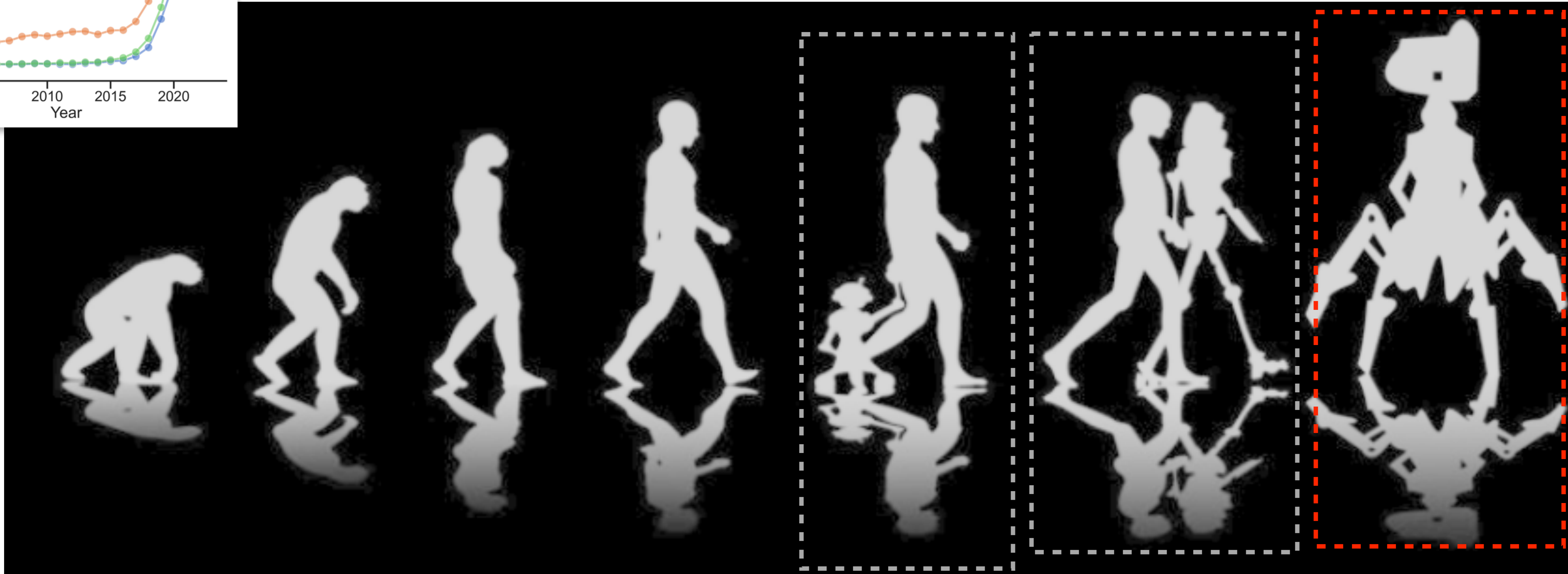


# ML/DL in HEP



“Traditional” ML Seeking DL solutions

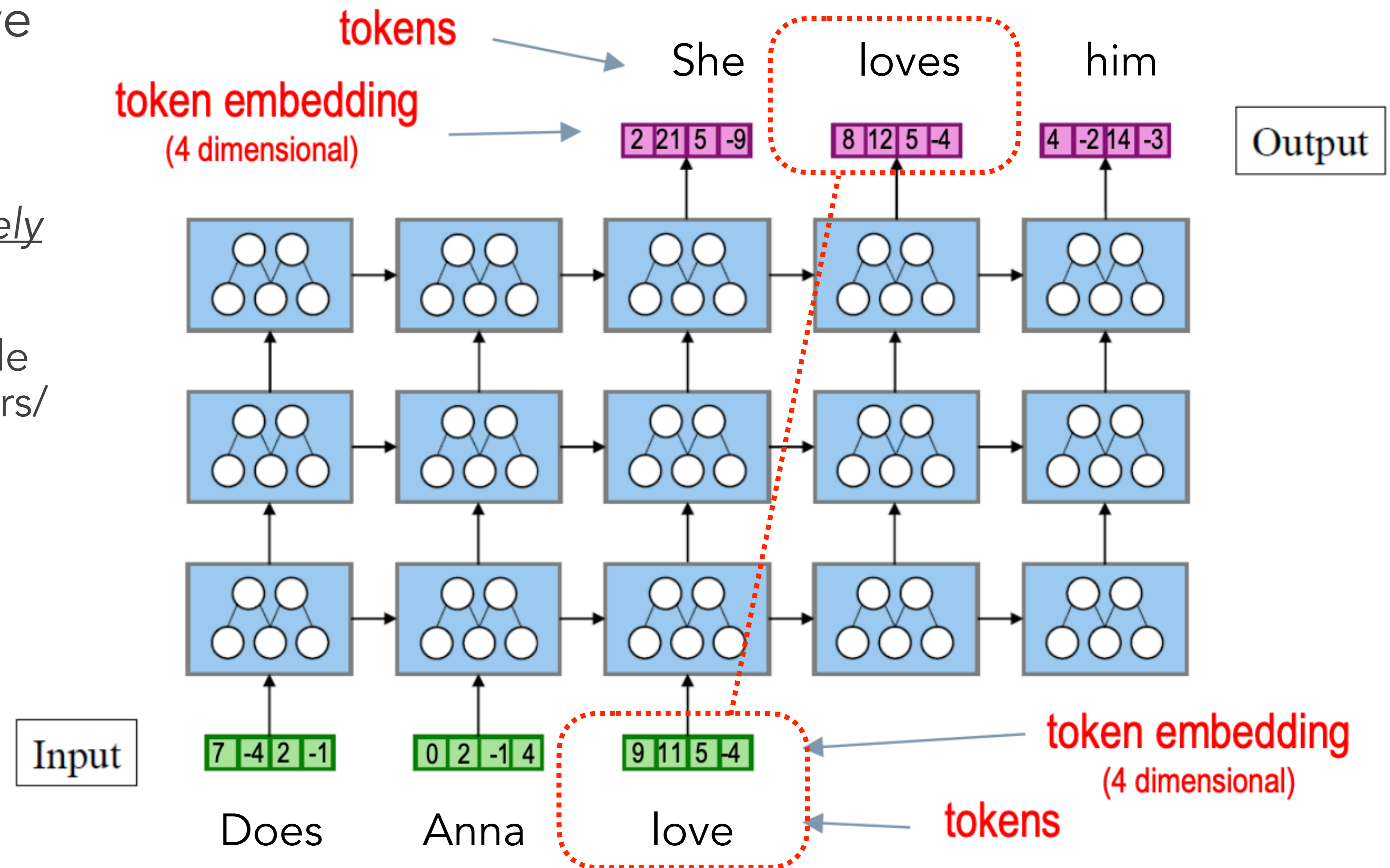
What next?



**Language Models** (LM) are generative (in the way we use them) models that specialise in handling **text**

- NNs are good with data that are *natively* numbers and vectors
- .. but to handle text, one needs to code characters and words, text into numbers/vectors..
- .. and in a semantically relevant way

LMs handle **text** as a sequence of symbols ("**token**"), mapped to multidimensional vectors ("**embedding**") and process them through a hierarchy of levels



Similar words need to be close-by in the vector space of their representation

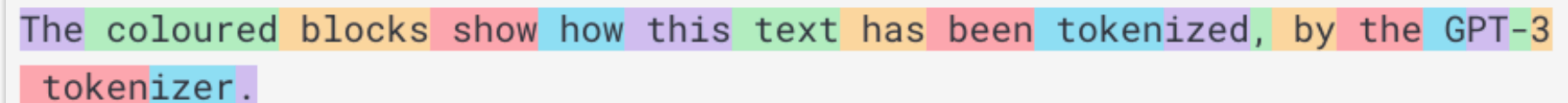


**First Generation** (2010-2015) of LMs: **Recurrent Neural Networks (RNNs)**, such as **LSTMs**, trained on a text corpus to understand language structure

- they process text as a sequence of symbols (tokens)
- tokens are mapped into multidimensional vectors (embeddings) and processed through a hierarchical structure of layers.

**Current Generation** (2018-today): so-called “modern” Large LMs (**LLMs**) - moving from RNN architecture to **large-scale transformers**

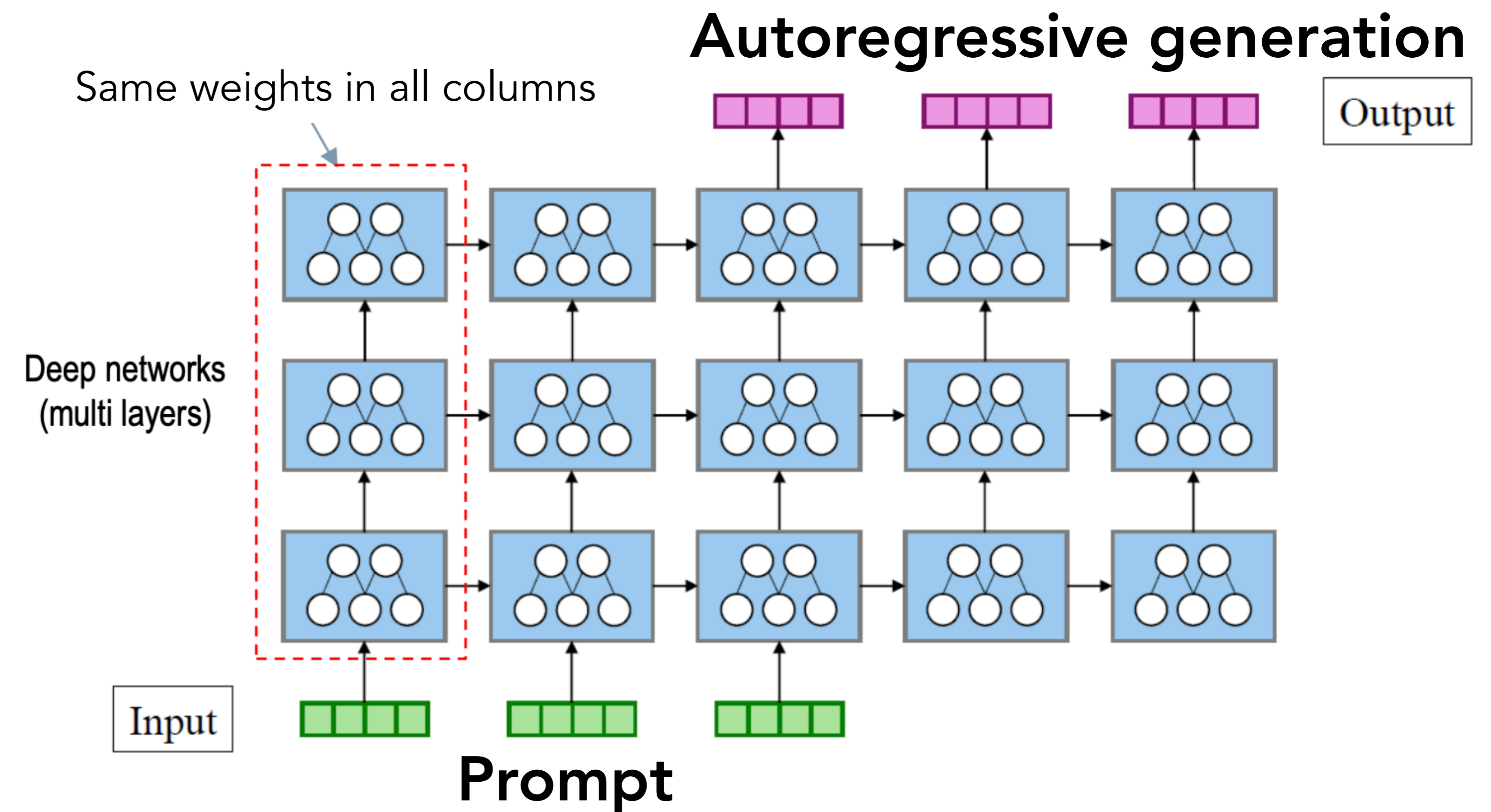
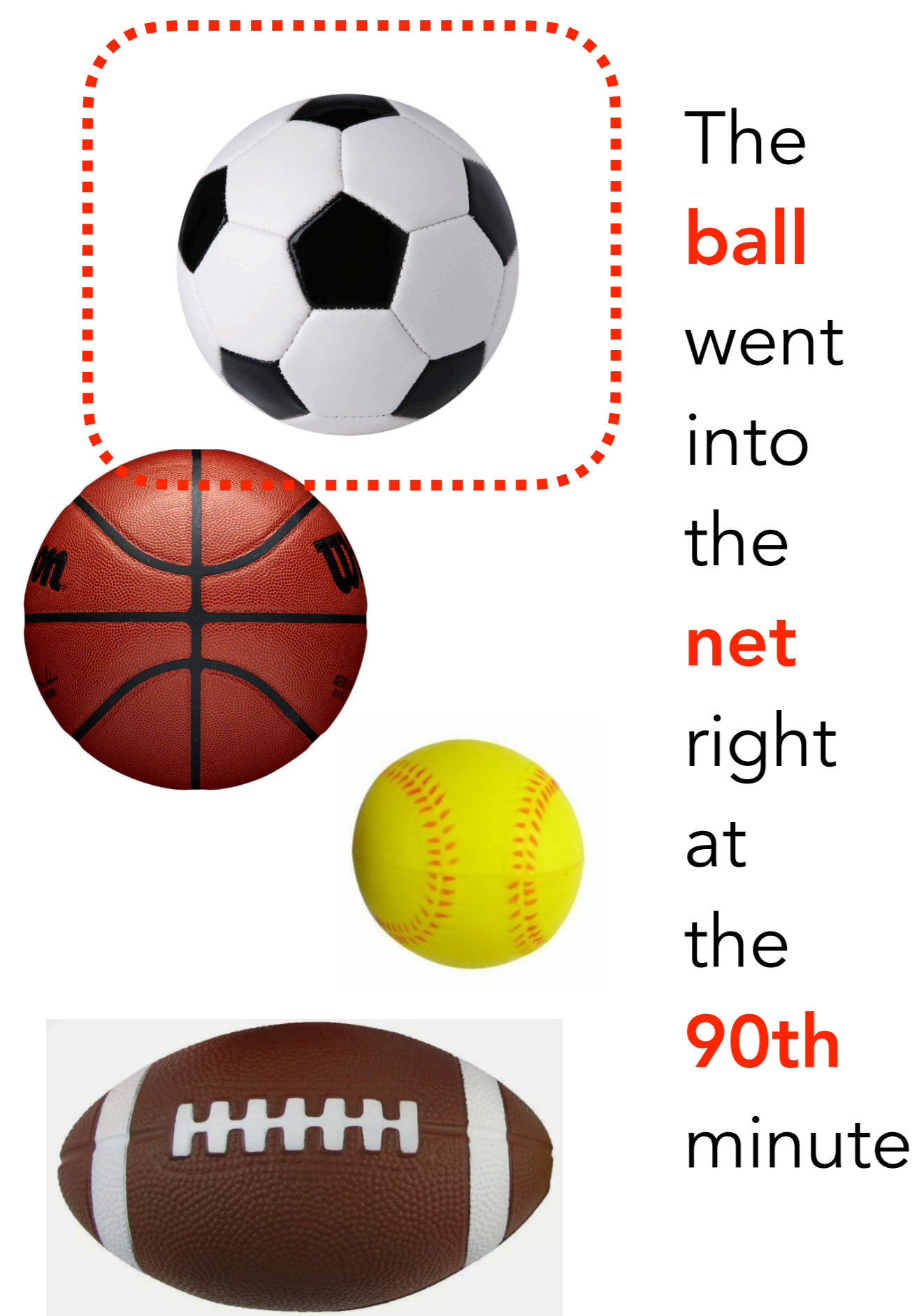
- improved tokenisation (see example below + multi-language + multi-domain, ..)
- “large-scale” → training on increasingly (massively!) larger text corpora (including e.g. code bases..)



The coloured blocks show how this text has been tokenized, by the GPT-3 tokenizer.

Transformers are the reference model for LLMs (the **T** in GPT)

- a DeepNN architecture introduced by Google (Brain) in 2017
- key element is the **self-attention** layers, which relate words within a sentence to better capture their semantics



Lower part, here comes the prompt. Pass through layers, including attention. Upper part: here comes the generation of the tokens of the answer. All in a autoregressive mode.

Transformers are not inherently generative models, but they become so when used with **sequence-to-sequence architectures** for prompt completion through autoregressive generation (one token at a time)

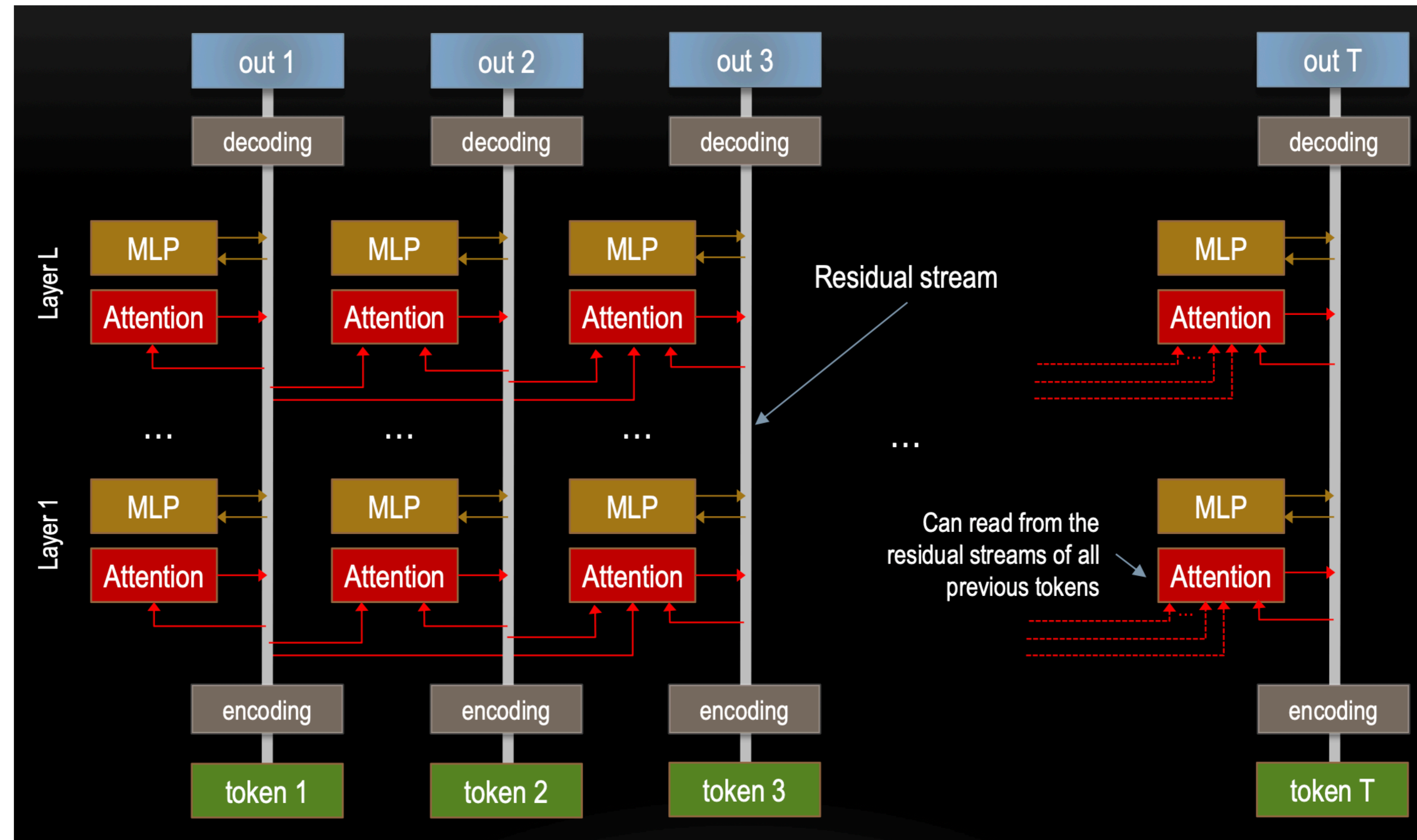
# Transformers

Every  
MLP+Attention  
contains millions  
of neurons!

Recent models:  
>100 layers

Incredible horizontal  
and vertical scalability !

Decoder only (es. GPT series)



The **ball** went into the **net** right at the **90th** minute

Recent models: >100k tokens



## GPT-3 (OpenAI, 2020)..

- GPT = Generative Pre-trained Transformer

.. from which **ChatGPT** (OpenAI, 2022) was derived

- trained on approx. 45 TB of text (equivalent to over 2'000x Wikipedia)
- estimated training cost: \$4.6M (initially.. now, for most recent models → O(100M\$))

## 2023: OpenAI releases **GPT-4**

- significantly superior to ChatGPT, it is considered state-of-the-art
- technical details unclear, but estimated to be an order of magnitude larger in terms of parameters and depth wrt previous models
- one data is known, though: in 2022, OpenAI reported an operating loss of \$540M (\$416M in computing costs)

Other major ICT players release/integrate LLMs, including:

- **Gemini** (Google), **Claude** (Anthropic), **LLaMA** (Meta), **DeepSeek**, .. – some appears as open source

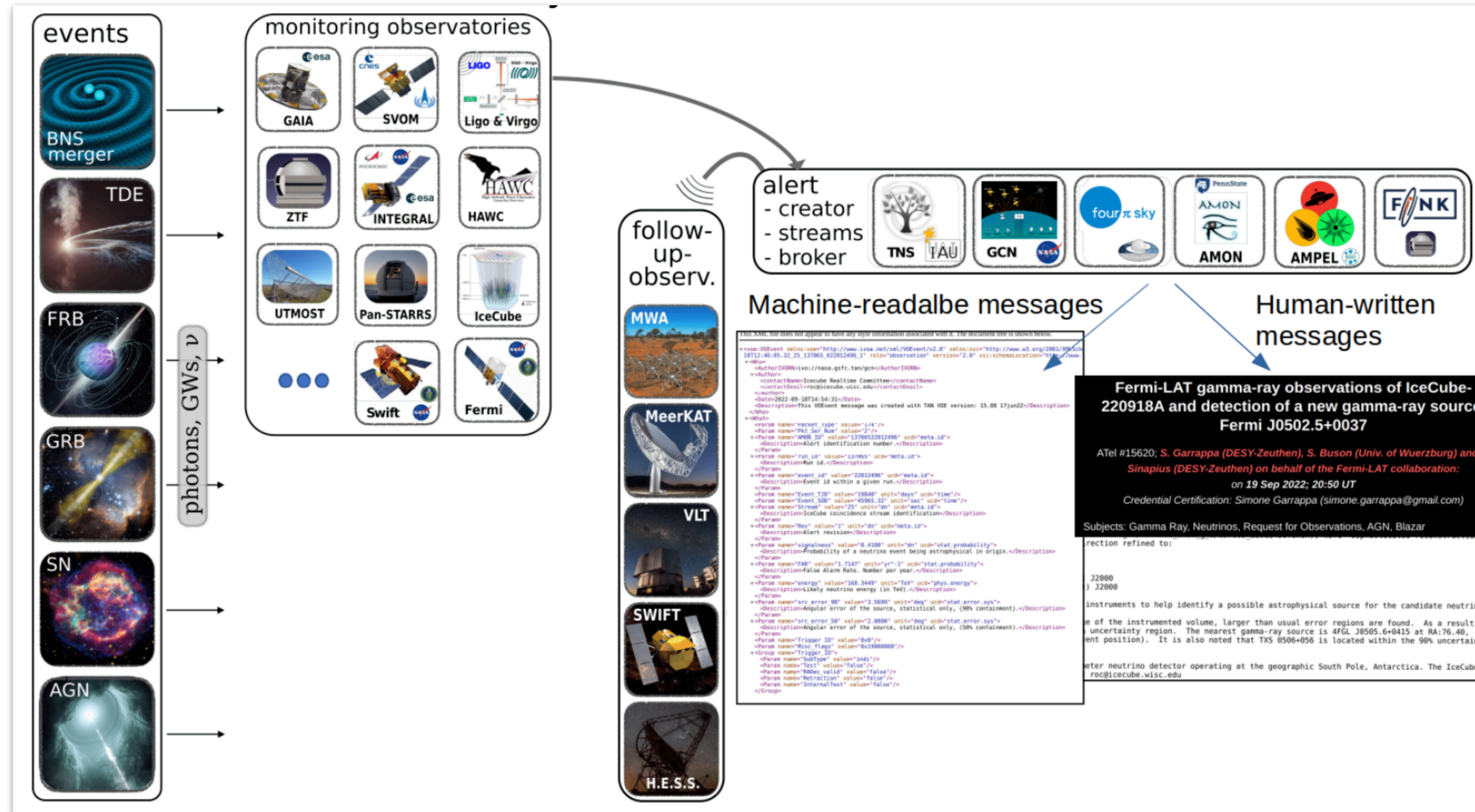
Dec. 2023: **Google Gemini Ultra's training cost: approx. \$191M**

Significant improvements thanks to the use of LLMs:

- **Machine translation:** e.g. Italian to English
- **Text classification:** e.g. sentiment analysis
- **Named Entity Recognition (NER):** extracts relevant information from unstructured texts (e.g. vital signs from medical records)
- **Summarisation:** e.g. generating document summaries
- **Question-Answering:** multi-domain and multilingual factual knowledge
- **Digital assistants:** e.g. Amazon Alexa, Google Assistant, Apple Siri

And in science? Well, sky is the limit!

Observations in astronomy:



# LLMs for Multimessenger Astronomy

*Description: Referred to by ATel #: 8706, 8718, 8783, 8789 On Jan 14, 2016, the Large Area Telescope (LAT) observed strong **gamma-ray emission** from a new source. The best-fit location of this **gamma-ray source** (RA=8.91 deg, Dec=61.52 deg, J2000.0) has a 95% containment radius of 0.08 deg. This source is not in any published LAT atalog and in the past has not been detected by AGILE or EGRET. The closest candidate counterpart is the **radio source** **87GB 003232.7+611352**.*

An example text based on  
**ATel** messages  
([astronomerstelegam.org](http://astronomerstelegam.org))

■ object name; ■ the type of the object or physical phenomena; ■ event type

Goal is to build an information extraction system, i.e. recognise a list of predefined concepts (celestial objects, astronomical facilities, physical properties, people, organisation etc.) from a text and produce LLM-generated event summaries based on the parameters of each event

A Cycle 6 ALMA **obs** proposal for 4 epochs of stand-alone **Telescope** Atacama Compact Array **wlght** Band 7 observations of variables in **obj** Serpens identified by the **Survey** JCMT Transient Survey has been accepted (PI: **Person** Logan Francis, project code **Grant** 2018.1.00917.S ). These observations will complement results from the contemporaneous Transient Survey by observing at **Wavelength** 850  $\mu$ m with a resolution of 38 (compared to the 146 resolution of the **tel** JCMT), sufficient to reach the scale of the inner envelopes ( $\approx 1500$  au) of protostars in **obj** Serpens.

From entity recognition...

... to semantic relationship extraction

**Celestial Object** MASTER OT J105440.86-391319.0 - PSN in PGC600519 ( 2.9°E,7°N ) MASTER-SAAO auto-detection system ( Lipunov et al. , " MASTER Global Robotic Net " , Advances in Astronomy , 2010 , 30L )  
 discovered OT source at ( RA , Dec ) = 10h 54 m 40.86s , -39d 13 m 19s.0 on 2016 - 10 - 08 02:56:46.019UT with m\_OT=16.5 . This PSN is in 2.9°E,7°N from the center of PGC600519 and is seen on 8 images . We have reference image without OT on 2015 - 02 - 27 22:35:30UT with m\_lim=19.8 Spectral observations are required The discovery and reference images are  
 http://master.sai.msu.ru/static/OT/MASTEROTJ105440.86-391319.0.jpg. **Celestial Object** MASTER OT J160918.30 - 333456.7 discovery - bright possible dwarf nova outburst , amplitude 5.4 m MASTER-SAAO

“Cosmosage”, a general-purpose AI-assistant specialised in answering questions about cosmology (based on Mistral-7B-v0.1)

- training dataset: arXiv papers, astro textbooks, physics textbooks, wikipedia

**user:**  
What came before the Big Bang?

**cosmosage:**  
The concept of what came before the Big Bang is a topic of debate and speculation in the realm of cosmology. The Big Bang theory states that our universe began approximately 13.8 billion years ago from a singularity, an infinitely dense and hot point. At this point, space, time, matter, and energy were all created simultaneously. Scientists cannot explain with any certainty why the Big Bang happened, so any speculation about what came "before" is pure theoretical musings. The Big Bang is widely accepted as accounting for everything we have been able to observe about the universe, but it doesn't explain itself. There are many theories and models that attempt to explain the pre-Big Bang universe, such as the cyclic model, the ekpyrotic model, and the string theory landscape. However, these theories are still under development and have not been proven to be accurate descriptions of the actual events that took place before the Big Bang.  
**ChatGPT 3.5 gives a similar quality answer.**

Popular science

**user:**  
Why does a cross-spectrum not suffer from noise bias?

**cosmosage:**  
An auto spectrum suffers from noise bias because the noise is correlated with itself, leading to an overestimation of the signal. In contrast, a cross spectrum does not have this noise bias because the noise in different maps is uncorrelated. Therefore, cross spectra are often preferred for cosmological analysis because they provide an unbiased estimate of the signal.  
**ChatGPT 3.5 says [... cross-spectra are less prone to noise due to cancellation effects ...] which is incorrect.**

Data analysis principles

**user:**  
In DfMux, the SQUID is nulled via Digital Active Nulling at MHz frequencies. How is the SQUID kept nulled at low frequency?

**cosmosage:**  
In DfMux, the SQUID is kept nulled at low frequency by using a low-frequency flux-locked loop (FLL). The FLL is a feedback system that maintains the SQUID in its nulled state by adjusting the input current to cancel out the flux signal. This is necessary because the SQUID is sensitive to magnetic fields, and the magnetic field generated by the input current can change the bias flux, causing the SQUID to lose its null state. The FLL helps to maintain the null state by continuously adjusting the input current to counteract any changes in the bias flux.  
**ChatGPT 3.5 says [... passive nulling ... without requiring active intervention...] which is exactly the wrong answer.**

Detailed Instrumentation

**user:**  
What is the current best measure of the optical depth to reionization?

**cosmosage:**  
The current best measure of the optical depth to reionization is  $\tau=0.054\pm0.007$ , as determined by the Planck Collaboration et al. (2018). This value provides strong evidence for reionization at high redshift and helps constrain models of cosmic reionization.  
**ChatGPT 3.5 talks about future measurements but does not mention a numerical value or even mentions Planck.**

Current cosmological parameter constraints



# A LLM-based AI-assistant for a CERN experiment



“ChATLAS” a **prototype LLM project in a LHC experiment (ATLAS) at CERN** (as of end 2023)

Data **gathering** part is interesting (data **chunking** and data **retrieval** not described here)

- Docs: twiki (>2k), sw docs (>500), e-groups/emails archive (>10k), indico meetings' agendas incl. attached slides and minutes (>440k), Mattermost, Jira tickets, experiment' papers and internal notes (>66k)
- Either HTML or scraped into markdown

Many open challenges:

- highly heterogeneous data
- ensure that collaboration DBs are accessible and exportable; websites should live on a git repo; pubs should be saved as latex, and compiled separately; discussion forums should have anonymisation options... Estimates indicate that this would have saved ~1 yr of data wrangling
- Hallucinations are a real problem
- Not many gpu-hrs, but many expert-hrs, needed for any high-quality fine-tuned AI assistant



# Educational Outreach with AI-Assisted CERN Open Data



opendata  
CERN

Help About ▾

Explore more than **five petabytes** of open data from particle physics!

Search

search examples: [collision datasets](#), [keywords:education](#), [energy:7TeV](#)

**Explore**

- [datasets](#)
- [software](#)
- [environments](#)
- [documentation](#)

**Focus on**

- [ATLAS](#)
- [ALICE](#)
- [CMS](#)
- [LHCb](#)
- [OPERA](#)
- [PHENIX](#)
- [Data Science](#)

<https://opendata.cern.ch/>

# An example: a Higgs analysis guide

**Define a training goal**

ATLAS Open Data Higgs analysis guide  
By Philipp Gadow  
Guides on Higgs boson data analysis with ATLAS Open Data.

What is the Higgs boson and how can I measure its m... Show me how to plot the invariant mass distribution ...  
How do I read ROOT files for my analysis? How can I fit the mass peak of the Higgs and how can...

Message ATLAS Open Data Higgs analysis guide...

ChatGPT can make mistakes. Consider checking important information.

**Get draft code for it**

```
sorted_indices = np.argsort(event_E[::-1])  
if len(sorted_indices) < 2: # Check if there are at least two photons  
    continue # Skip events with fewer than two photons  
idx1, idx2 = sorted_indices[:2] # Indices of the two photons with largest energy  
  
# Extract the leading and subleading photon four-momentum components  
E1, E2 = event_E[idx1], event_E[idx2]  
eta1, eta2 = event_eta[idx1], event_eta[idx2]  
phi1, phi2 = event_phi[idx1], event_phi[idx2]  
  
# Convert eta, phi, E to px, py, pz  
px1,
```

Message ATLAS Open Data Higgs analysis guide...

ChatGPT can make mistakes. Consider checking important information.

**Get explanations, refine, learn**

you are calculating the invariant mass using these ene  
GeV (gigaelectronvolts), this could be the source of discrepancy leading to an unexpected range or  
empty plots.

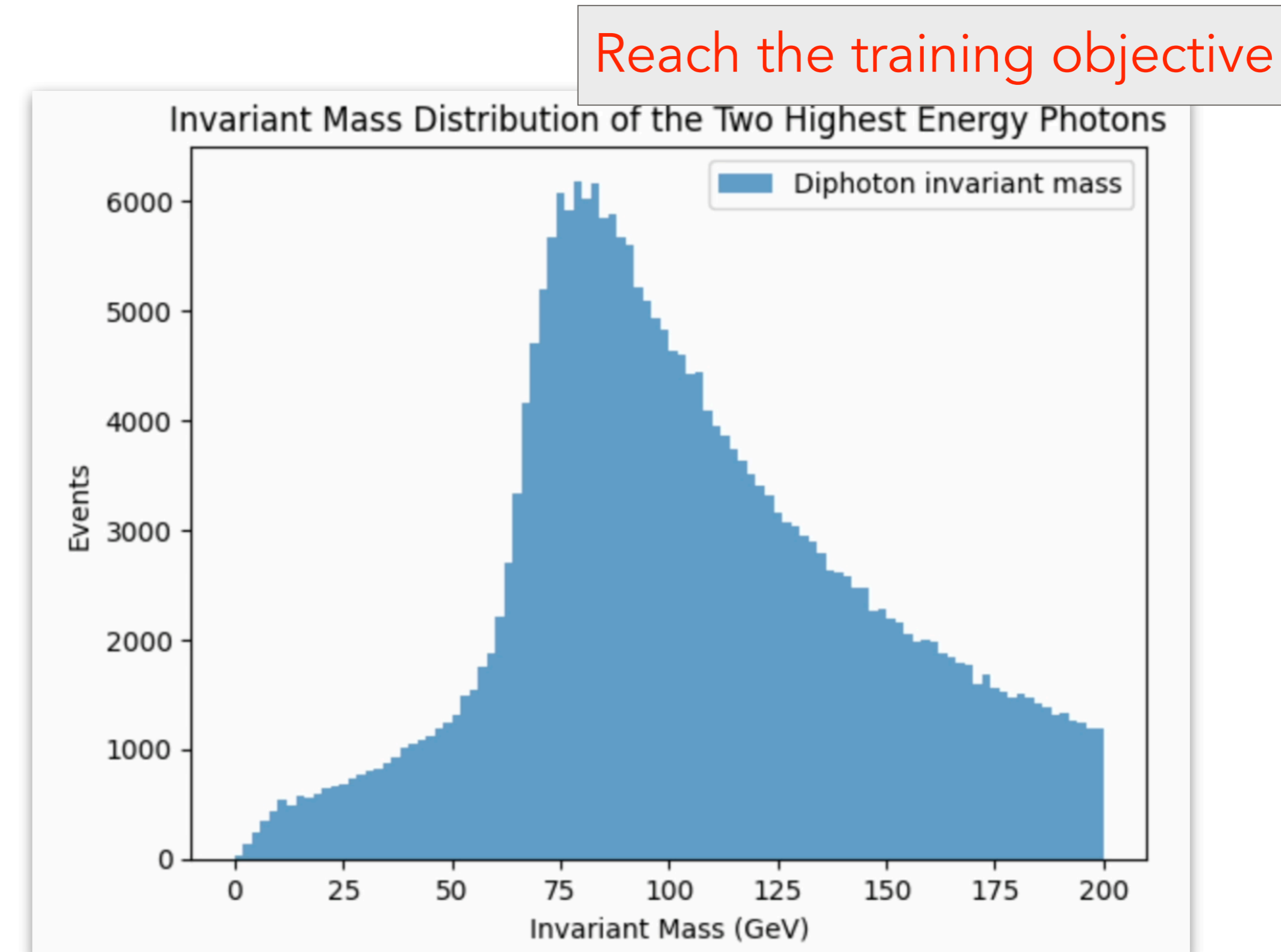
To fix this, you should convert the energies from MeV to GeV before performing the invariant mass  
calculation. This can be done by dividing the energy values by 1000:

```
python  
E1, E2 = event_E[idx1] / 1000., event_E[idx2] / 1000. # Convert from MeV to GeV
```

Incorporate this conversion into your code where you define `E1` and `E2` and then plot the  
invariant mass distribution again. This adjustment should help ●

Message ATLAS Open Data Higgs analysis guide...

ChatGPT can make mistakes. Consider checking important information.

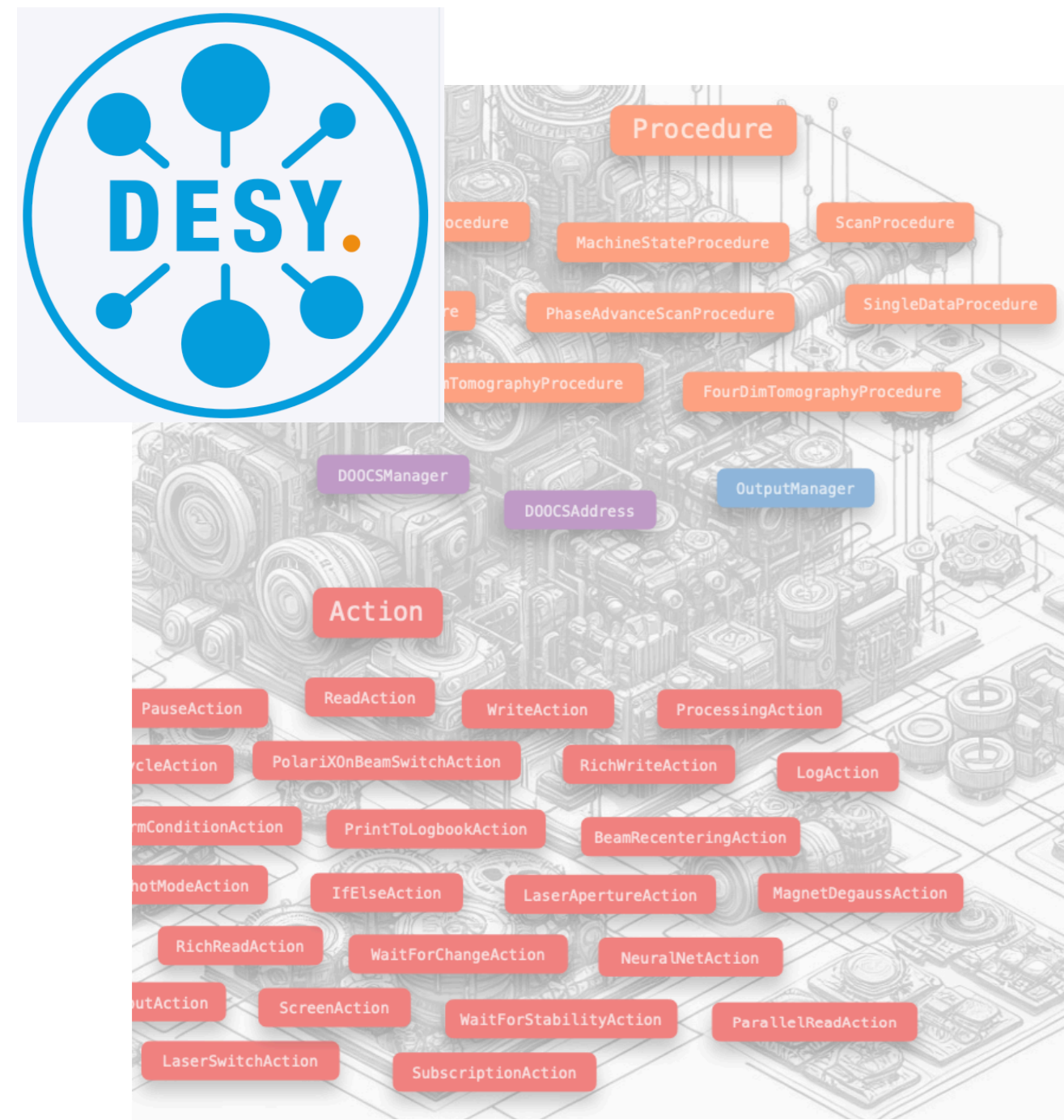




A variety of projects..

Plenty of work in progress on LLMs, showing potential towards natural language driven autonomous particle accelerators

- Attempts with GPT 3.5 Turbo, Megadolphin, Vicuna 7B 16K, Mistral 7B, Mixtral 7x8B, Starling-LM, GPT 4 Turbo, GPT4, Orca 2 7B, Orca 2 13B , Llama 2 70B, Falcon 180B, ..
- Constant seek for (and tests with) better models, better prompting, ...



GAIA (@**DESY**): a General AI-assistant for Intelligent Accelerator Ops

- Experimental “procedures” defined as a collection of high-level “actions” in a Control System e.g. for managing machine pre-sets
- Exploring a LLM (mixtral:8x7b-instruct-v0.1-q8\_0 with 32k context size), agent implemented in Python using the langchain module, prompting based on ReAct (as a combination of chain-of-thought prompting and information injection via “actions”)

EPA project (@**CERN**), AccGPT, etc..



- EPA = Efficient Particle Accelerator project
- AccGPT = accelerating science via a chatbot for knowledge retrieval for CERN specific content

## Different from supervised learning

- No need to label anything! I have plenty of text, so..

**Self-supervised learning:** masked portion of text and sliding windows as training tactics

- take a large text (easy to find..) as training set, and assume (for simplicity) tokens = words
- fixed-length sequences (e.g. 10 words) are extracted from the long text
- the model is trained to predict the 11th word given the first 10 (**no labelling required**)

1	Nel	Nel	Nel	Nel	Nel
2	mezzo	mezzo	mezzo	mezzo	mezzo
3	del	del	del	del	del
4	cammin	cammin	cammin	cammin	cammin
5	di	di	di	di	di
6	nostra	nostra	nostra	nostra	nostra
7	vita	vita	vita	vita	vita
8	mi	mi	mi	mi	mi
9	ritrovai	ritrovai	ritrovai	ritrovai	ritrovai
10	per	per	per	per	per
11	una	una	una	una	una
12	selva	selva	selva	selva	selva
13	oscura	oscura	oscura	oscura	oscura
14	ché	ché	ché	ché	ché
15	la	la	la	la	la

**ORANGE:** words given in input

**RED:** word to predict

## Different from supervised learning

- No need to label anything! I have plenty of text, so..

## Self-supervised learning: masked portion of text and sliding windows as training tactics

- take a large text (easy to find..) as training set, and assume (for simplicity) tokens = words
- fixed-length sequences (e.g. 10 words) are extracted from the long text
- the model is trained to predict the 11th word given the first 10 (no labelling required)

After training, the model can be used for **autoregressive text generation**

→ “next token prediction” mechanism

1	nostra	nostra	nostra	nostra
2	vita	vita	vita	vita
3	mi	mi	mi	mi
4	ritrovai	ritrovai	ritrovai	ritrovai
5	per	per	per	per
6	una	una	una	una
7	selva	selva	selva	selva
8	oscura	oscura	oscura	oscura
9	ché	ché	ché	ché
10	la	la	la	la
11	diritta	diritta	diritta	diritta
12		via	via	via
13			era	era
14				smarrita

**BLU** becomes the new **GREEN**, and the window slides..

Self-supervised training is based on this “next token prediction”:

Criticised as being **too simplistic**

- It does nothing really sophisticated, indeed: it just relies on the statistics that I have in languages..

Actually, this is one of its main key strengths (e.g. I.Sutskever [1], 2023)

- Evidences that it enables the model to learn, and not based only on statistical properties of language...
- ... and it also forces the model to make accurate predictions even in mathematics, logic, coding, and common sense reasoning, where - in the absence of an oracle or exhaustive memorisation of all cases - correct predictions are only possible **by learning an underlying model of the problem**

Additionally, the simplicity of the approach **does not impose constraints on the learning/modelling**, hence leaving the LLM **free to choose the most suitable strategy for different problems**

$$3 + 2 = ?$$

I have seen this pattern plenty of times in training

→ 5

$$333 + 2 = ?$$

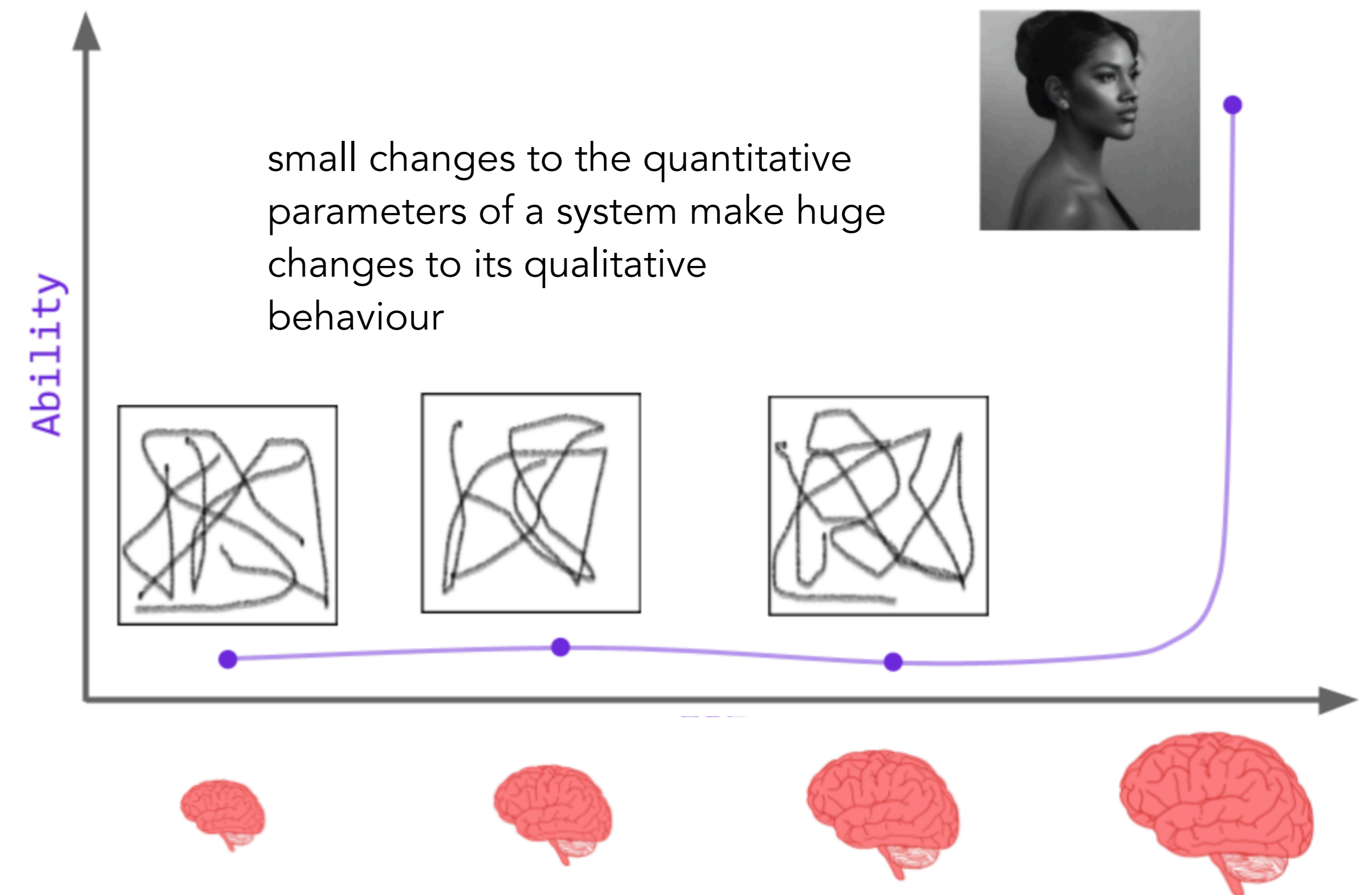
I have NOT seen this pattern many times.. nevertheless..

→ 335

Impossible to show all cases in training → it just extrapolates!

“**Emergence**”: a sudden appearance of a novel behaviour (often referred to as a “phase transition”)

- Scaling LMs → hit a series of critical scales at which new abilities are suddenly “unlocked”
- not directly trained to gain such abilities: they just manifest rapidly and in unpredictable ways



Examples:

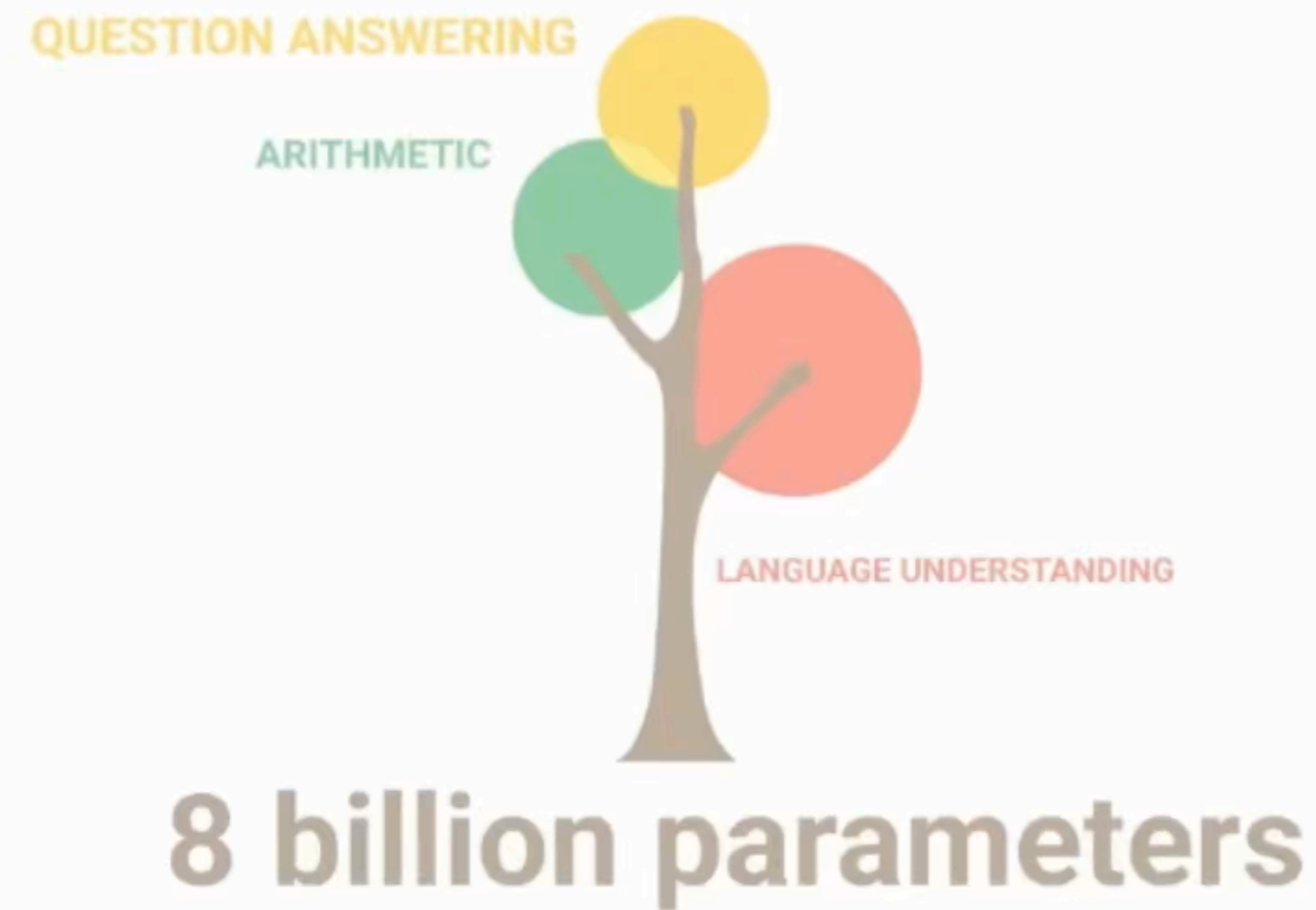
- problem solving (math, logic, quantitative reasoning), common sense and social behaviour, (controlled) generation of texts, images, sounds, .. ; ability to write, correct, and execute (pseudo)-code



# Large Language Models and “emergent abilities”



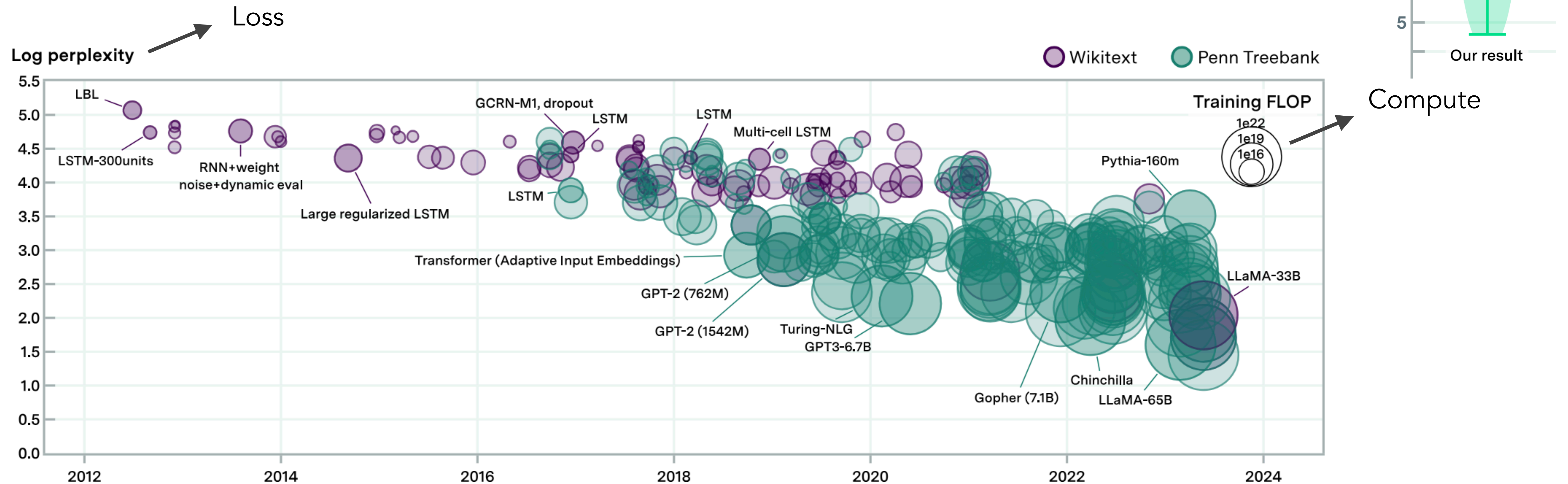
Animation here - will not be visible in the PDF



# Algorithmic progress in LLMs

The compute required to reach a set performance threshold has **halved approx. every 8 months**, with a 95% confidence interval of around 5 to 14 months

- Algorithmic improvements **faster** than hardware gains per Moore's Law!



Will we run out of data? compute? networks? ... or **energy**?

.. and algorithms will continue to get better → also if/when **written by AI** itself..

## Mammalian biological brains

Cat brain



0.760 billion neurons  
10 trillion synapses

Human brain



80 billion neurons  
150 trillion synapses



size of GPT 3.5

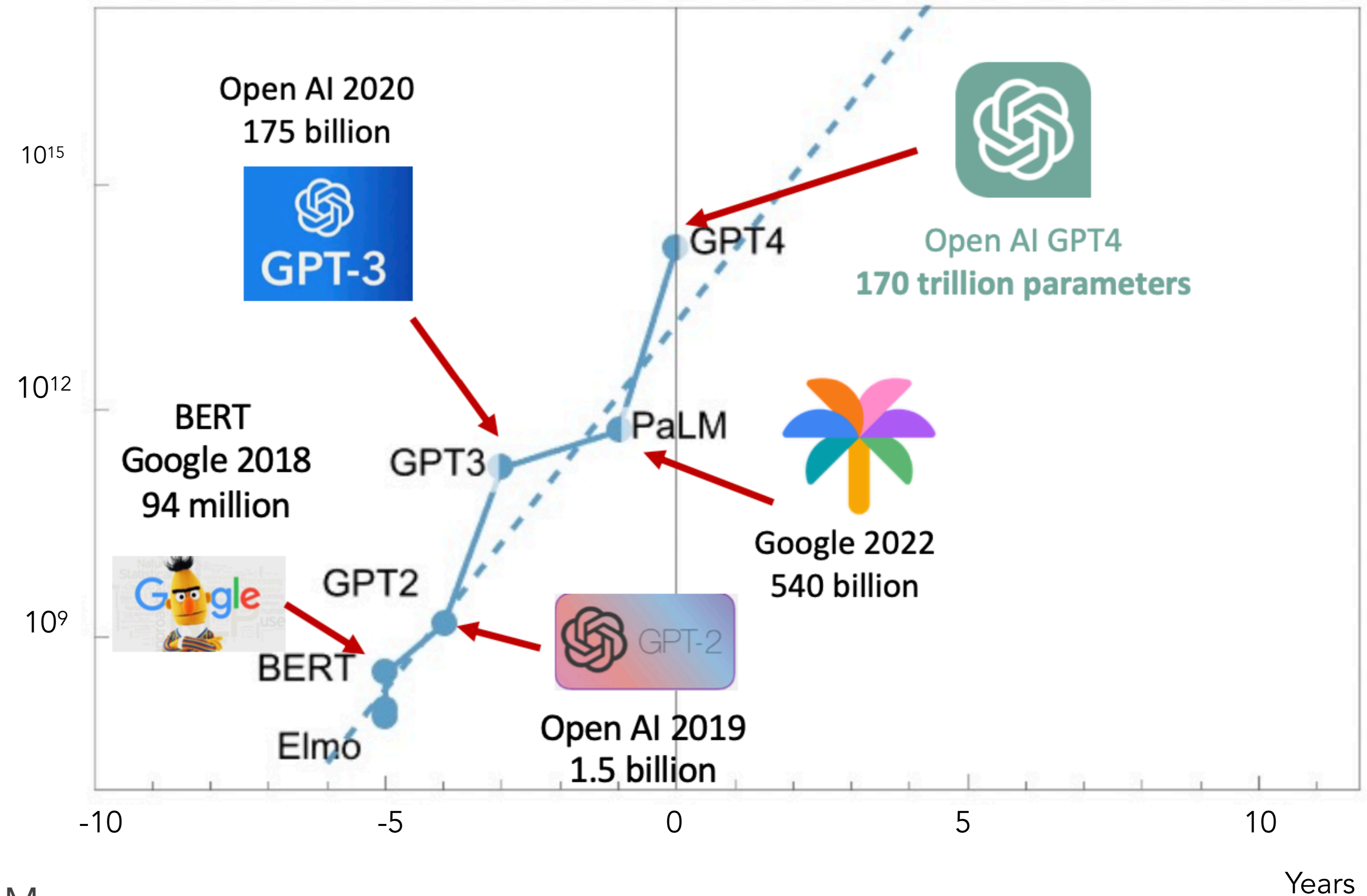


size of GPT 4



Parameters

LLMs



Current LLMs:

- Parameters: roughly **the same nb** ( $10^{14}$ ) as the human brain
- .. but **more compute**: brain ( $10^{16}$  FLOPS) over a lifetime (100 years)  $\rightarrow 10^{22}$  ops, to be compared with LLM training time, around  $10^{25}$  ops
  - ❖ And it consumes more.. **Red AI** is a serious issue!

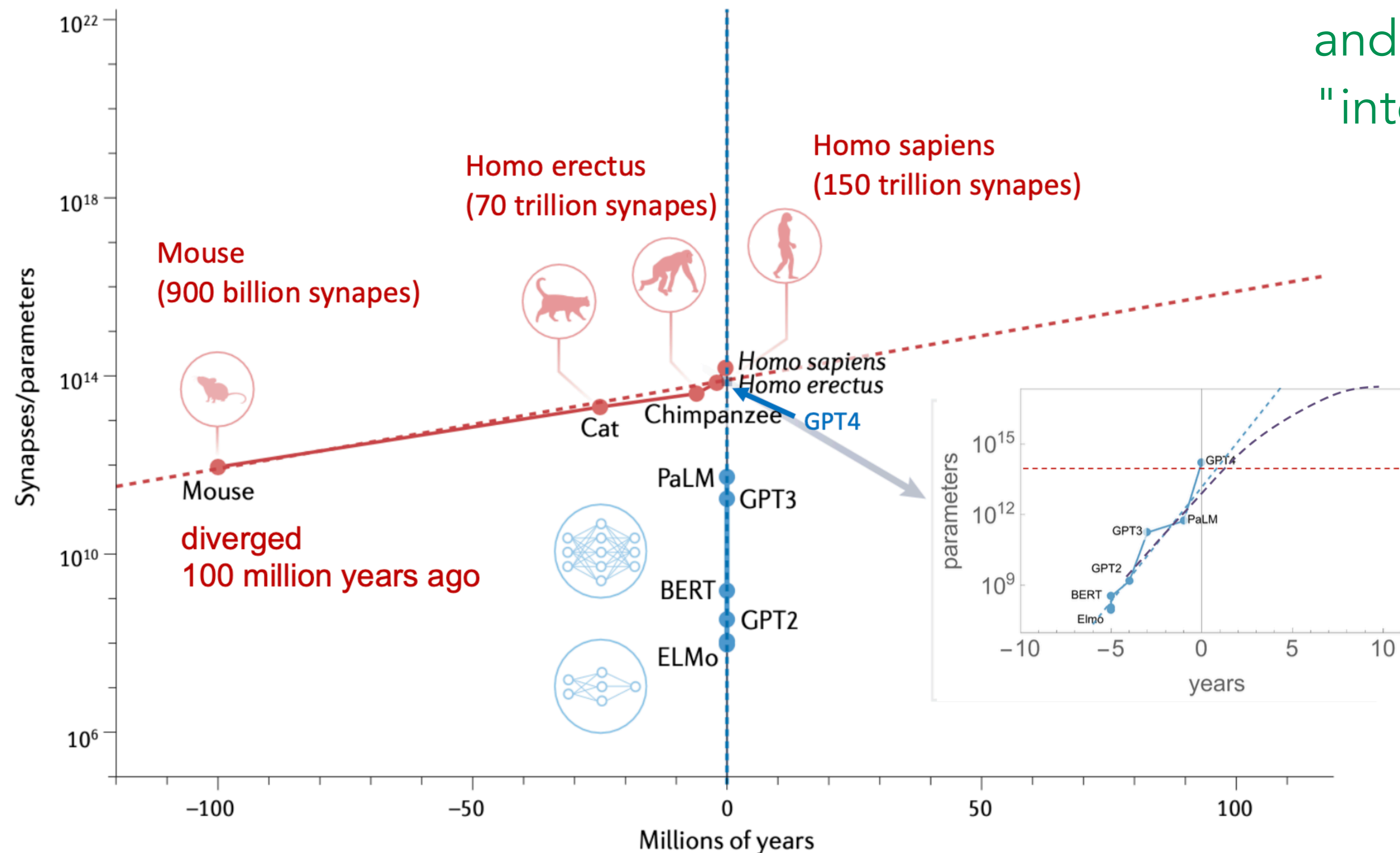


# Biological vs Machine "intelligence"

Measuring "intelligence" by number of neurons (or computational units):

- biological → growth by a factor 2x in 1 million years
- machine → growth by a factor 10x in 1 year

The intersection - when machines and biology have comparable "intelligence" - is **~now**



even sub-exponential growth will soon be superhuman !

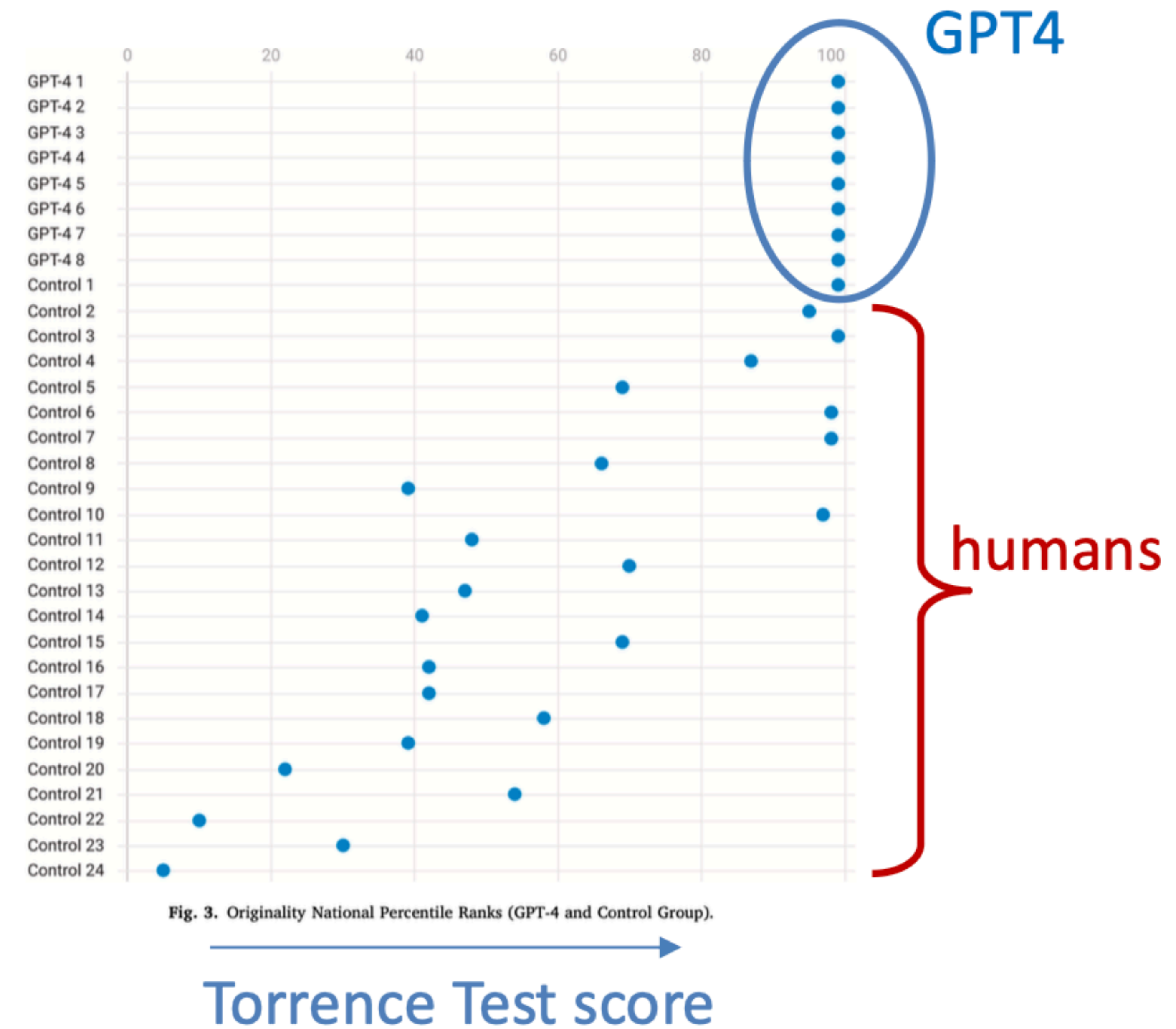
→ **AGI around the corner??**

Article | [Open access](#) | Published: 10 February 2024

## The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks

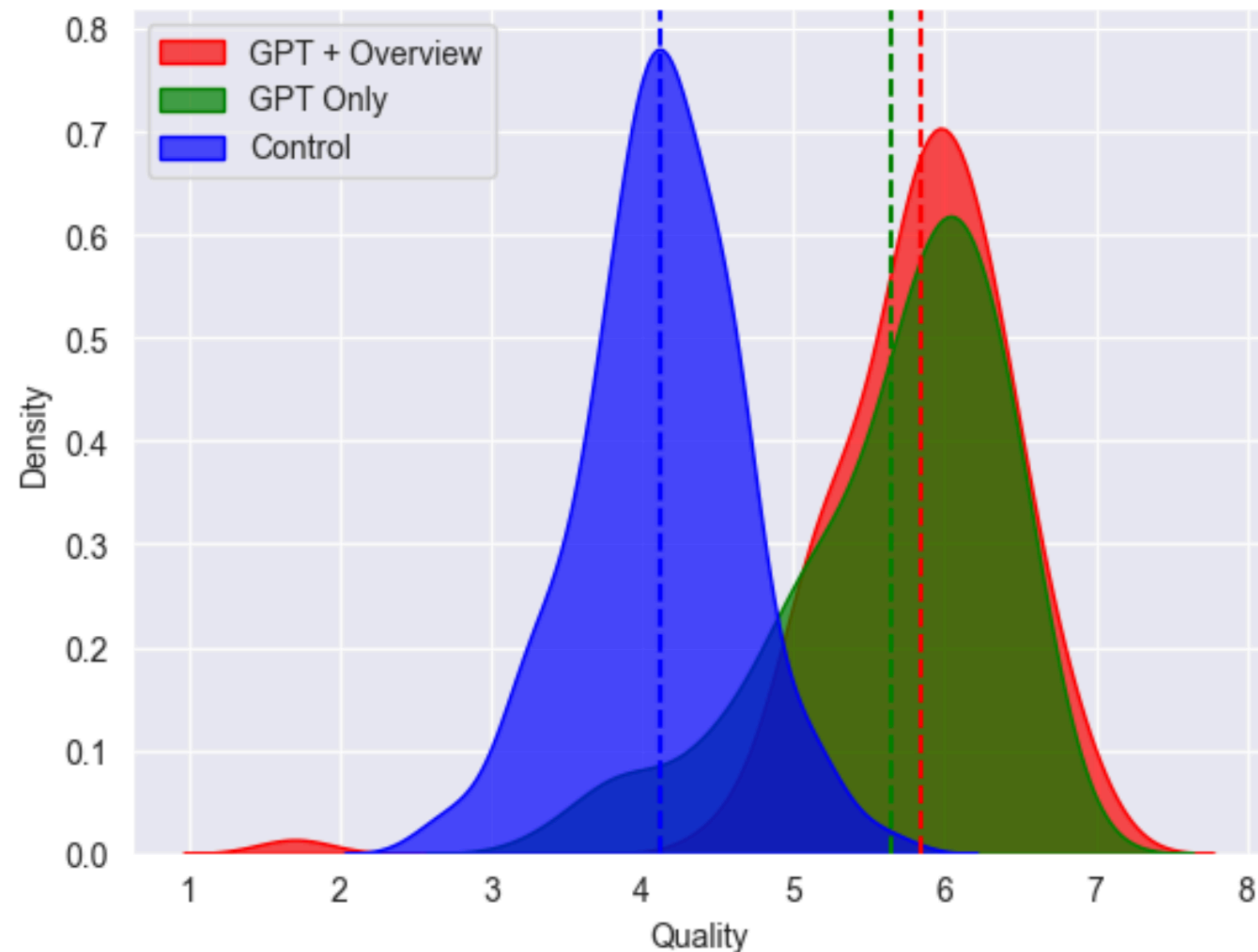
[Kent F. Hubert](#) , [Kim N. Awa](#) & [Darya L. Zabelina](#)

The emergence of publicly accessible artificial intelligence (AI) large language models such as ChatGPT has given rise to global conversations on the implications of AI capabilities. Emergent research on AI has challenged the assumption that creative potential is a uniquely human trait thus, there seems to be a disconnect between human perception versus what AI is objectively capable of creating. Here, we aimed to assess the creative potential of humans in comparison to AI. In the present study, human participants (N = 151) and GPT-4 provided responses for the Alternative Uses Task, Consequences Task, and Divergent Associations Task. We found that AI was robustly more creative along each divergent thinking measurement in comparison to the human counterparts. Specifically, when controlling for fluency of responses, AI was more original and elaborate. The present findings suggest that the current state of AI language models demonstrate higher creative potential than human respondents.



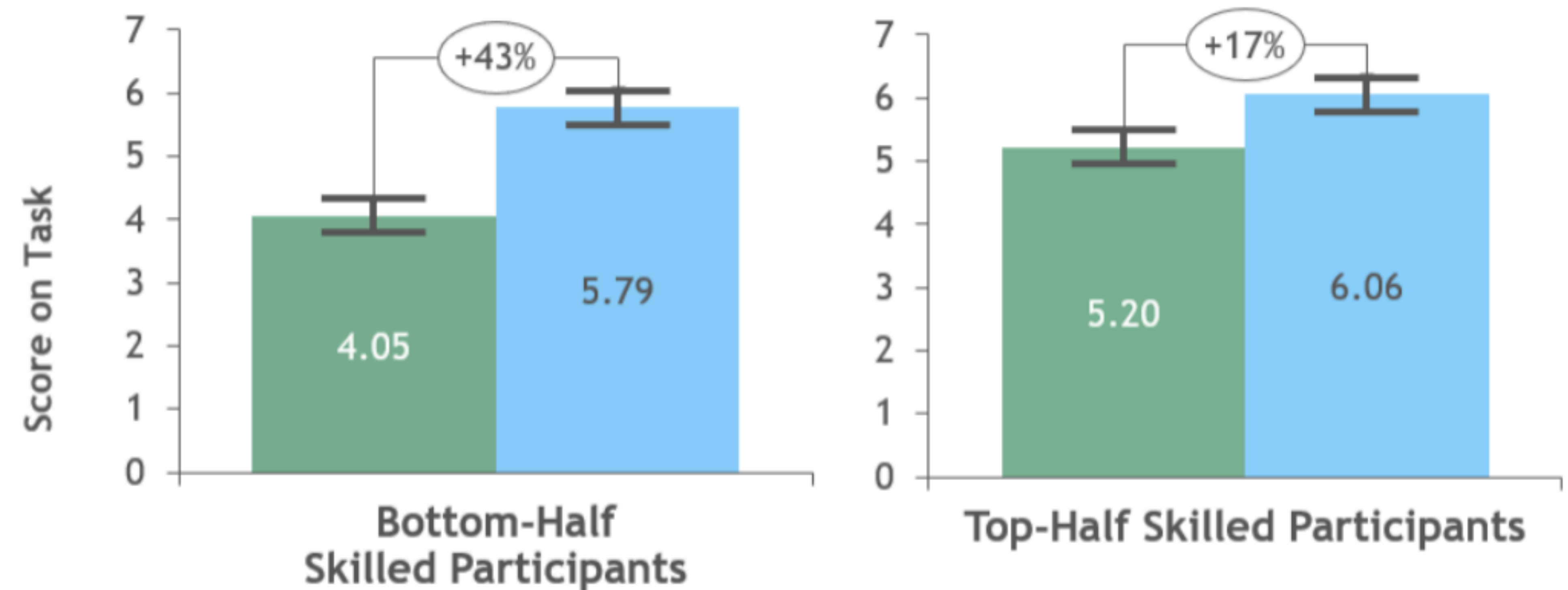
GPT4 more creative than 99% of humans..

## Can LLM help “consultants”?



Yes, and by a lot!

## AI-assistants as skill-levellers



Average consultants using AI perform as good as best consultants

→ towards “augmented intelligence”?

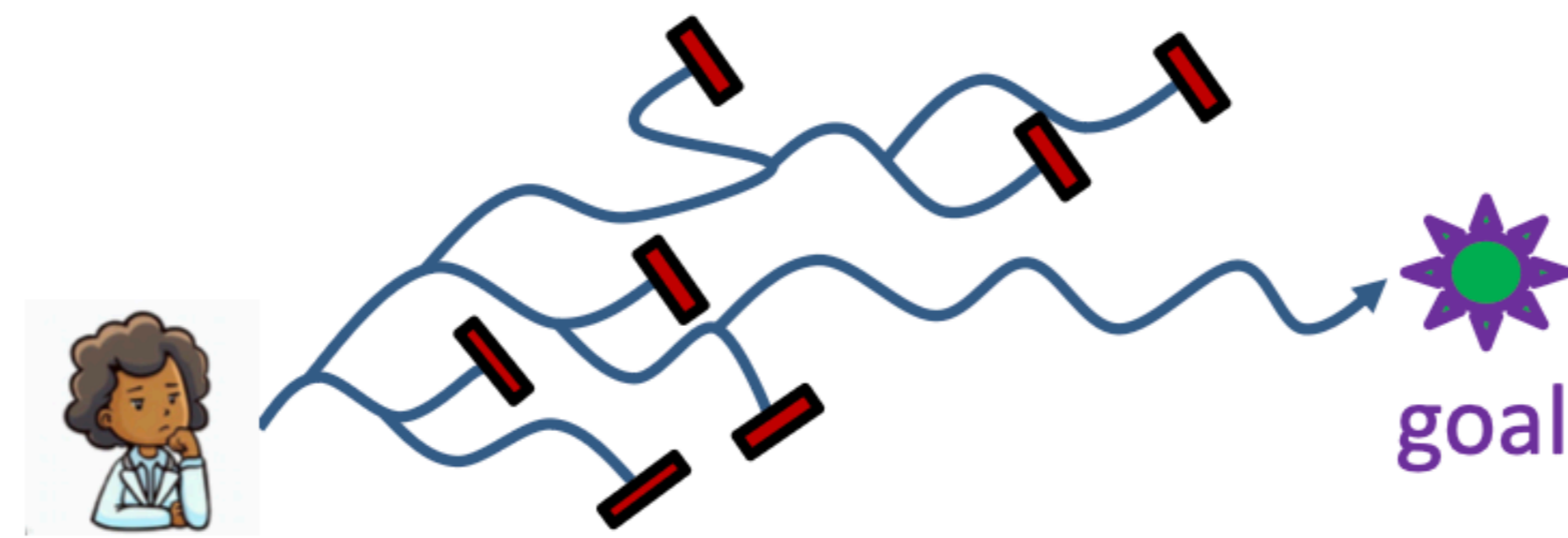
# Skill-leveller, indeed? What about HEP theory?

Can AI be a skill-leveller (or more) for e.g. HEP theory?

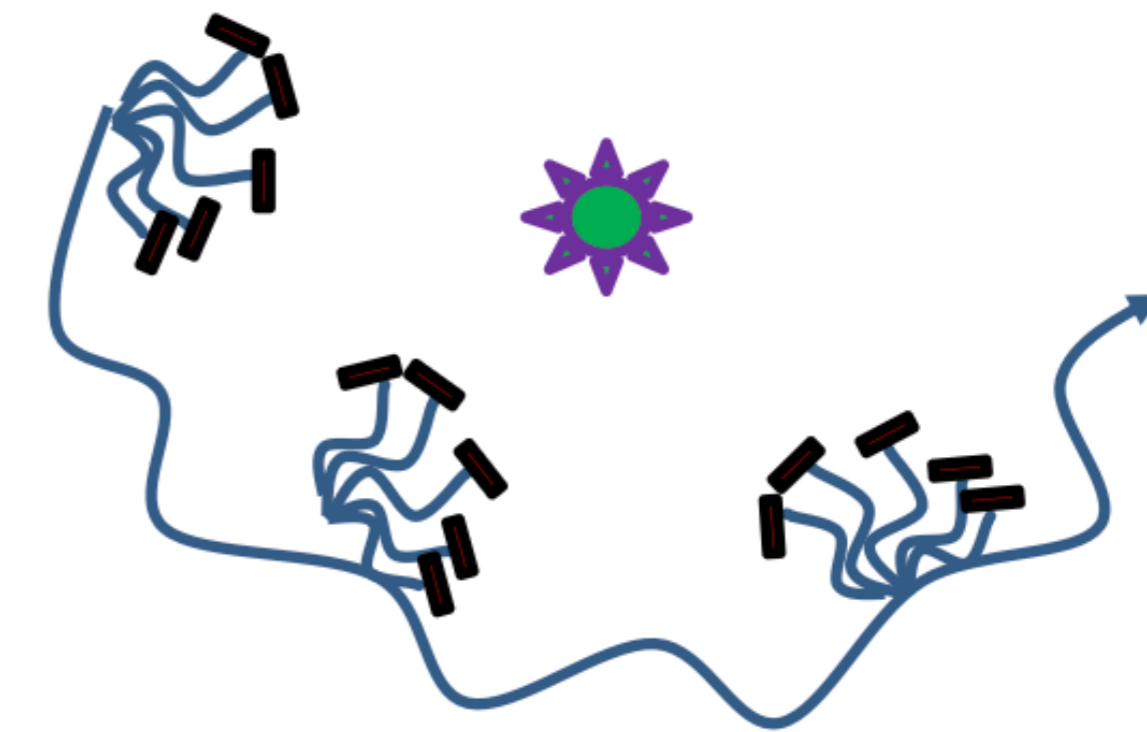


Credits: M. Schwartz, Harvard and NSF IAIFI, elaborating on hard theoretical physics problems and AI, at EuCAIFCon (Amsterdam, 2024)

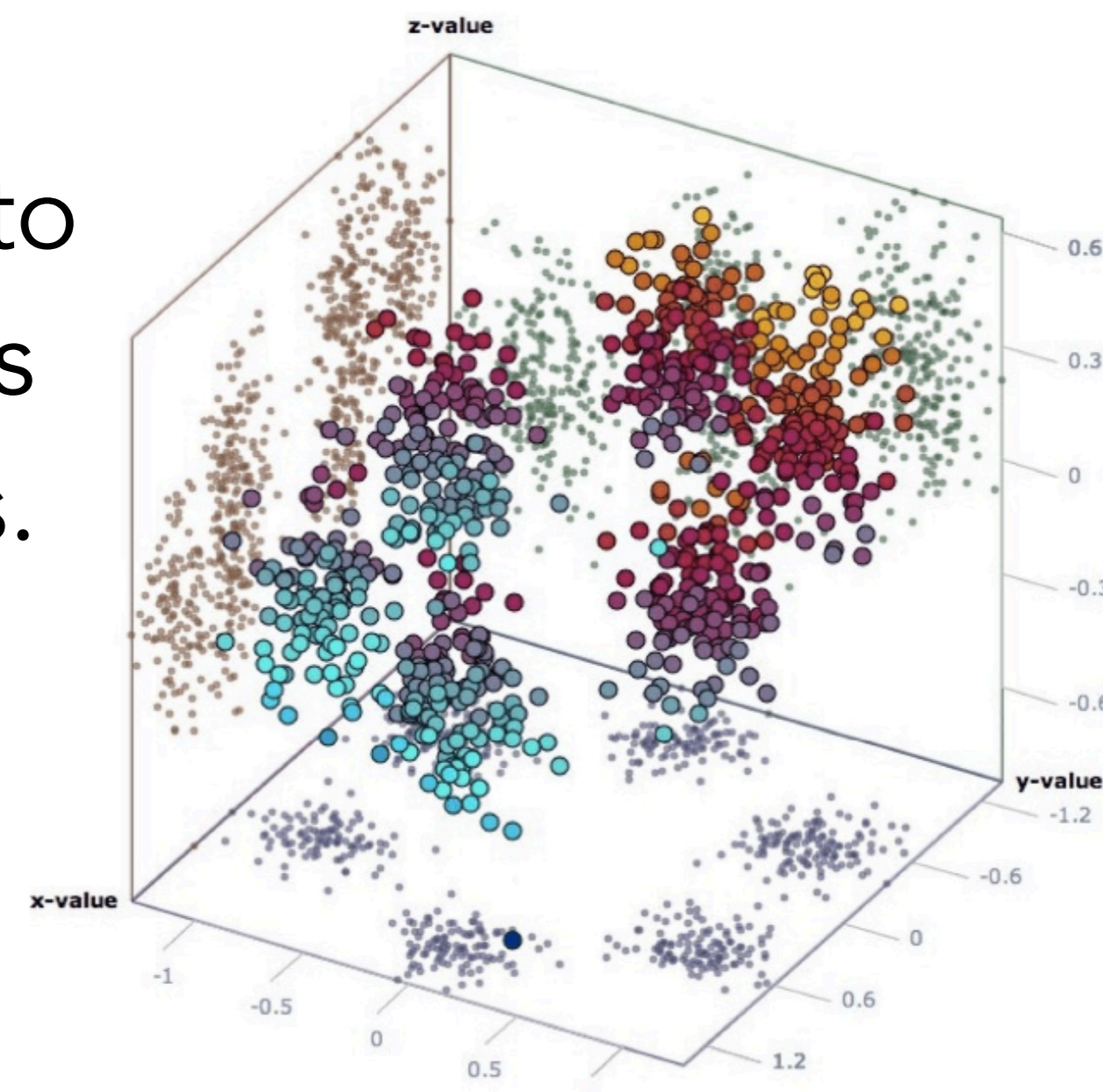
In the past, we made progress despite many dead ends



Are we even making forward progress anymore?



Humans like to "visualise", as we have eyes.



→ project in 2D

For a machine, 2D is not special: it can easily visualise in d dimensions

Humans hold few concepts in working memory at once, and like "simple and elegant" equations

$$i\partial_t\psi = H\psi$$

$$G_{\mu\nu} = \kappa T_{\mu\nu}$$

A computer memory can handle much more concepts at once, and can understand systems not governed by simple equations

# We are a training set for machines

Current state-of-the-art AI can answer questions / (~) solve textbook problems

How? → via training on huge datasets of answered questions / solved problems

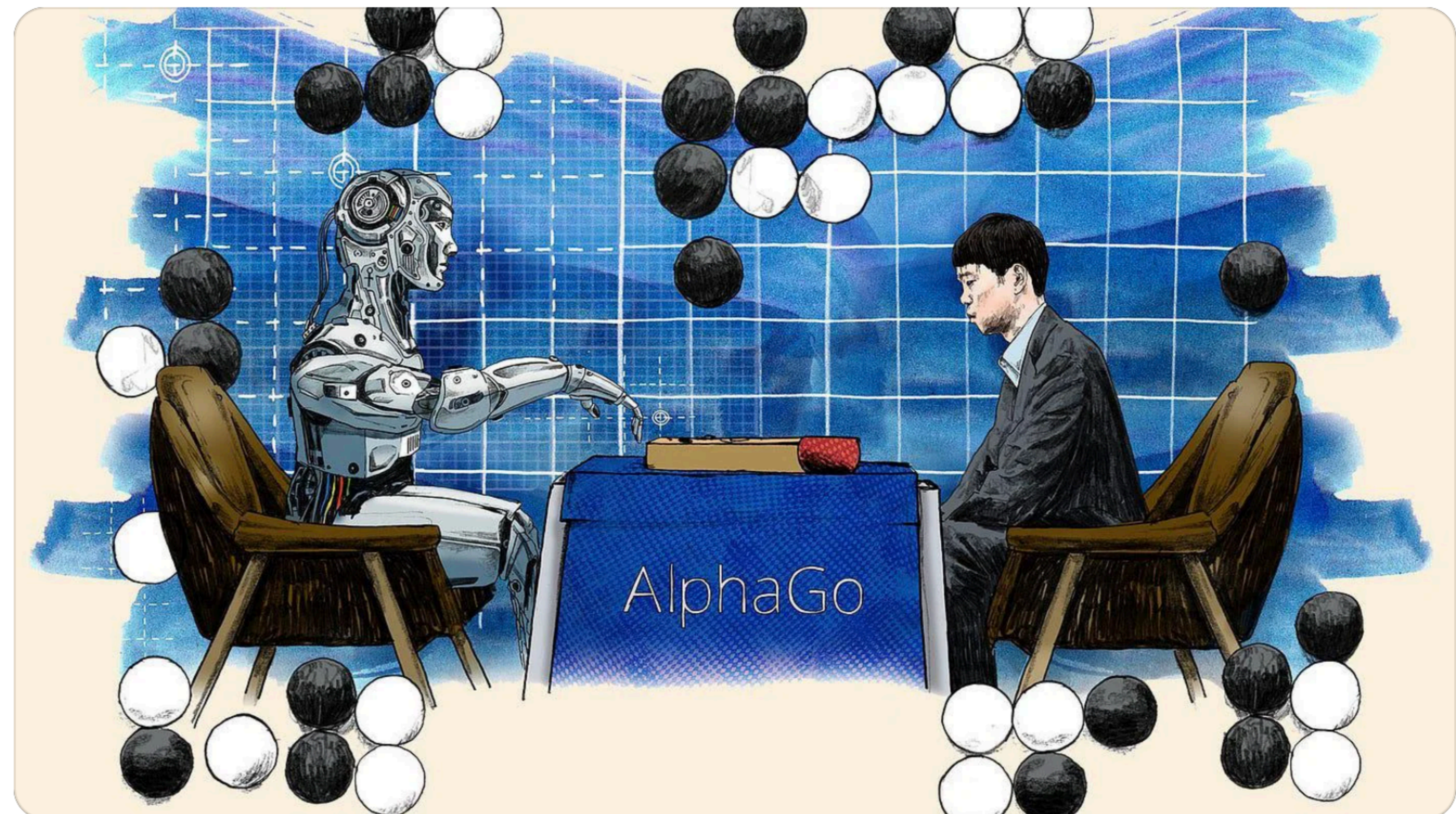
By whom? → Us! we answered and solved all that, we actually generated its training set

- (and we do the same for ourselves)

E.g. LLMs:

- learn from our training set
- Human feedback helps refine the models
- Machines generates data and refines its models

Humans and machines seem very close to be not so different..



# (even beyond) augmented intelligence

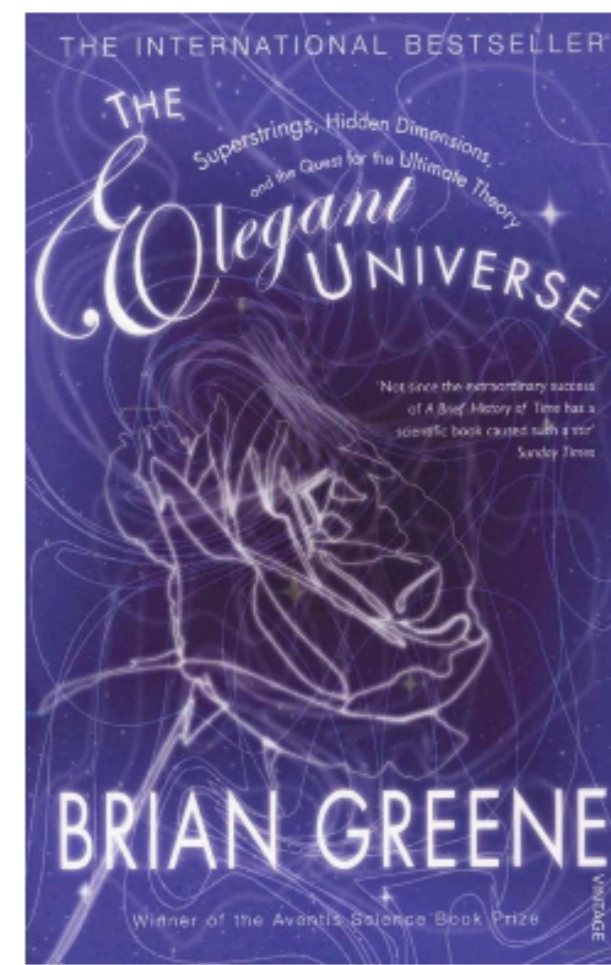
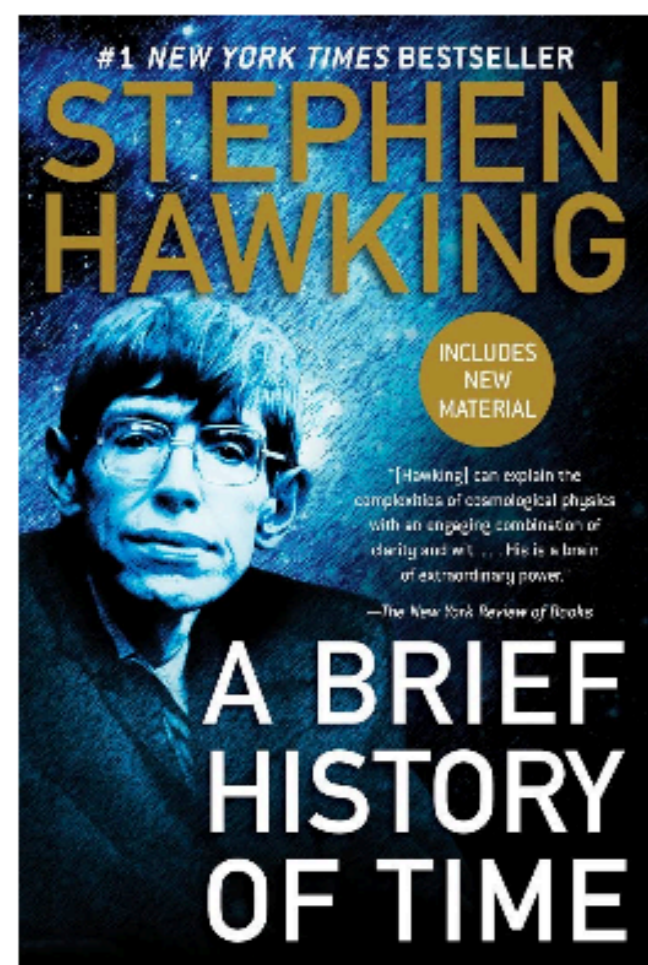
Maybe the problems are just too difficult (for us)?

- Example: could a cat ever learn to play chess? Humans have limits too.. e.g. biology

Suppose a machine understands the theory of everything but we don't

- e.g. can calculate the fine-structure constant from scratch
- e.g. can predict the endpoint of black-hole evaporation

Is this enough or do we need to understand it too?



- The authors of **Popular science books** understand the details; we just get the general idea

I don't understand the proof of Fermat's last theorem

- I'm glad that somebody does
- Does it matter that the person is human?

If a machine understands fundamental physics it can

1. Dumb it down so we can get the general idea
2. Find practical applications

Is this what we want? **I guess not.**

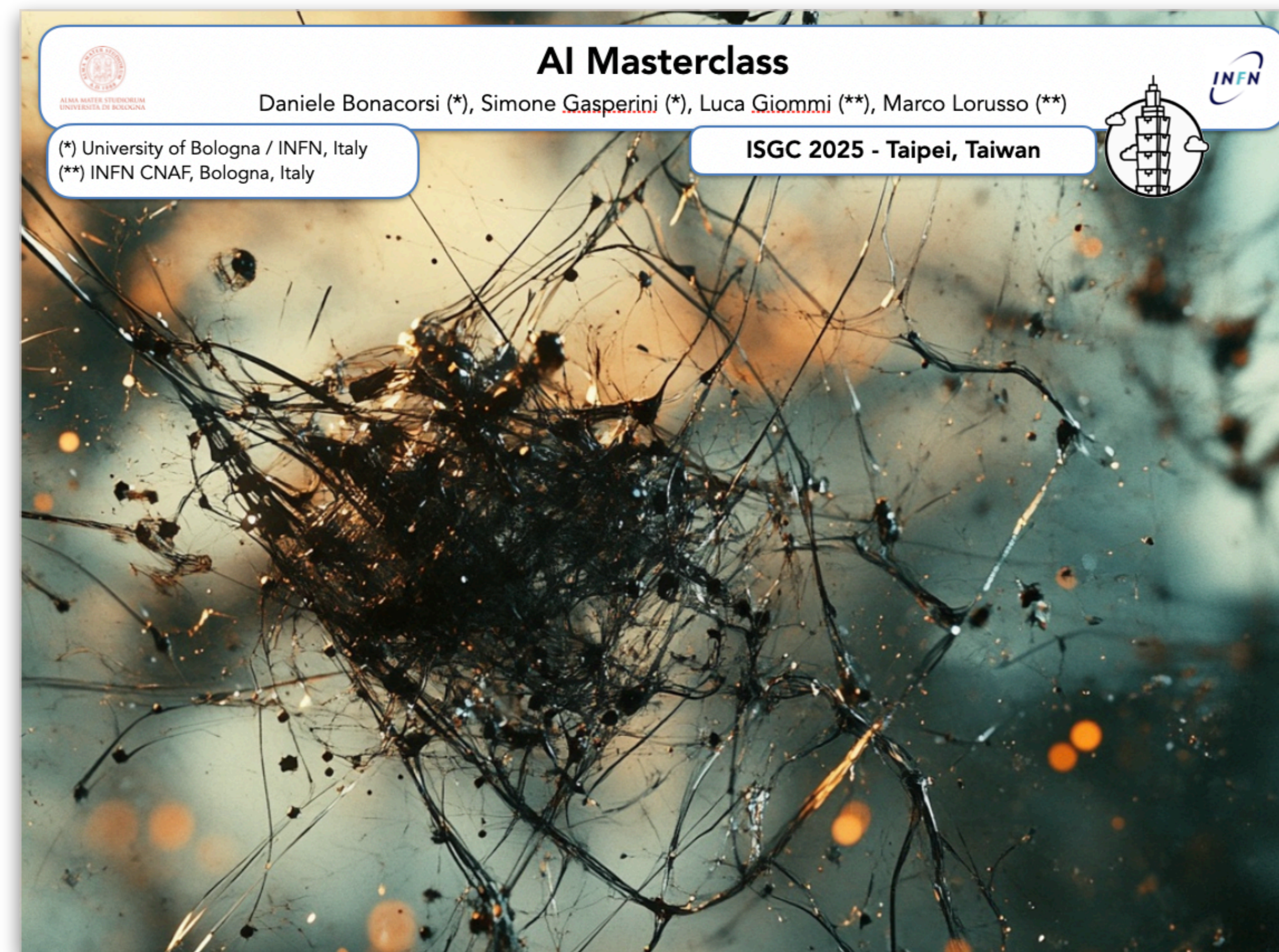
But..

What if this is the best we will get?  
What if AI could make us optimistic for a substantial progress in HEP theory **in our lifetime**, while it would be largely unprobable without AI?

**Would you give AI the keys of HEP research? And of ALL scientific research??**

If you have a definitive answer to the last question..

.. drop me a mail before I run the **AI masterclass** later today! → [daniele.bonacorsi@unibo.it](mailto:daniele.bonacorsi@unibo.it)



Today starting at **11am**,  
**three 1h30 slots**,  
this same **auditorium**

# Thanks for the attention