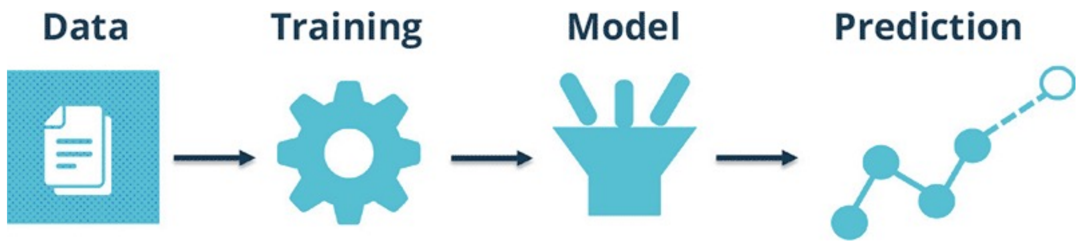
Ia Artificial Intelligence all we need?

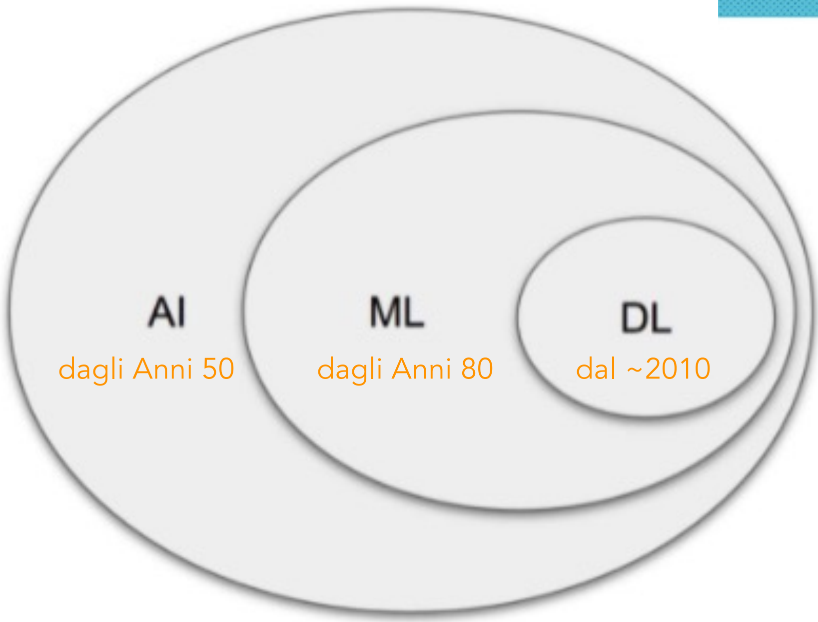
Daniele Bonacorsi (INFN e Università di Bologna)

INFN BOLOGNA
Istituto Nazionale di Fisica Nucleare
Sezione di Bologna

AI vs ML vs DL

AI = **Artificial Intelligence**
ML = **Machine Learning**
NN = **Neural Network**
DL = **Deep Learning**





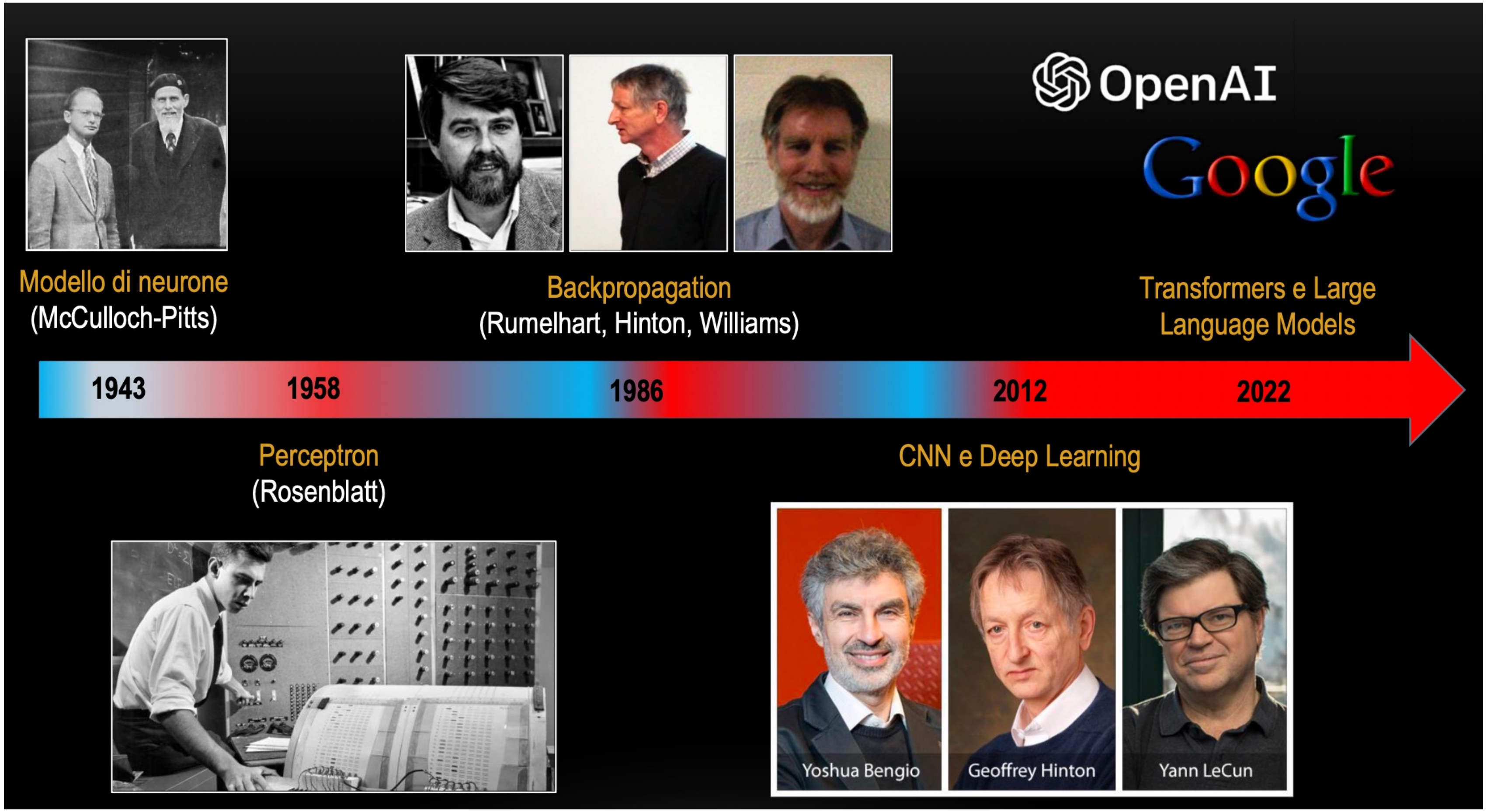
Objective: extract "actionable insight" from (big) data

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**"

Neural Networks: milestones

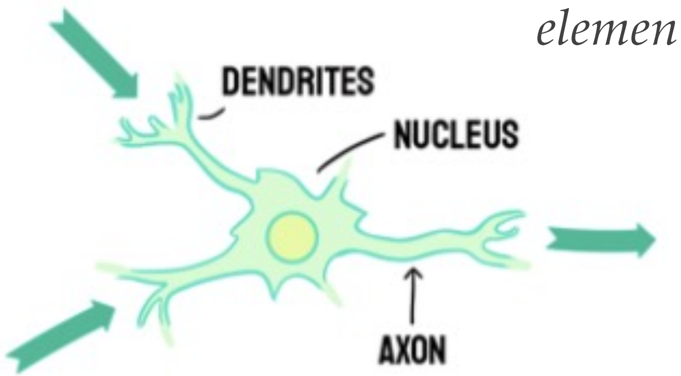


Credits: D. Maltoni

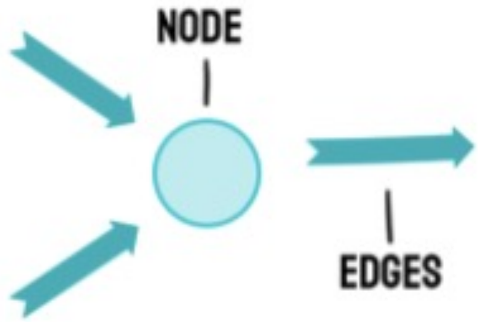
Neural Networks

Basic elements:

Biological Neural Networks



<u>Artificial</u> Neural Networks



A brief history of NNs

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)

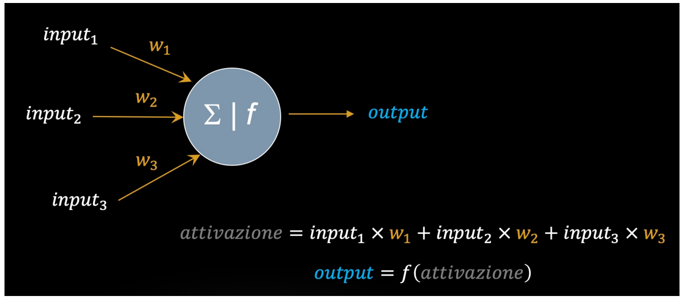
Multi Layer Perceptron (MLP)

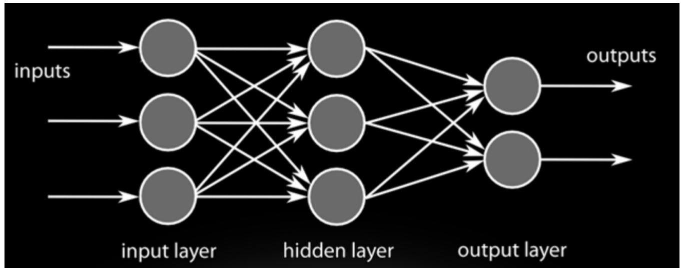
- Groups of neurons organized in **layers** (**input** layer, **output** layer, and one or more intermediate "**hidden**" layers, each with one or more neurons)

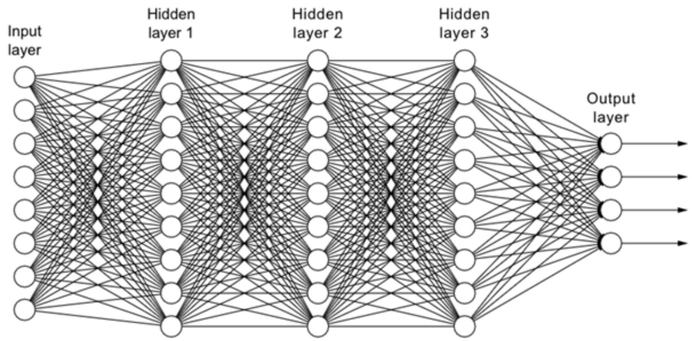
Back-propagation algo (Rumelhart, Hinton, Williams, 1986)

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

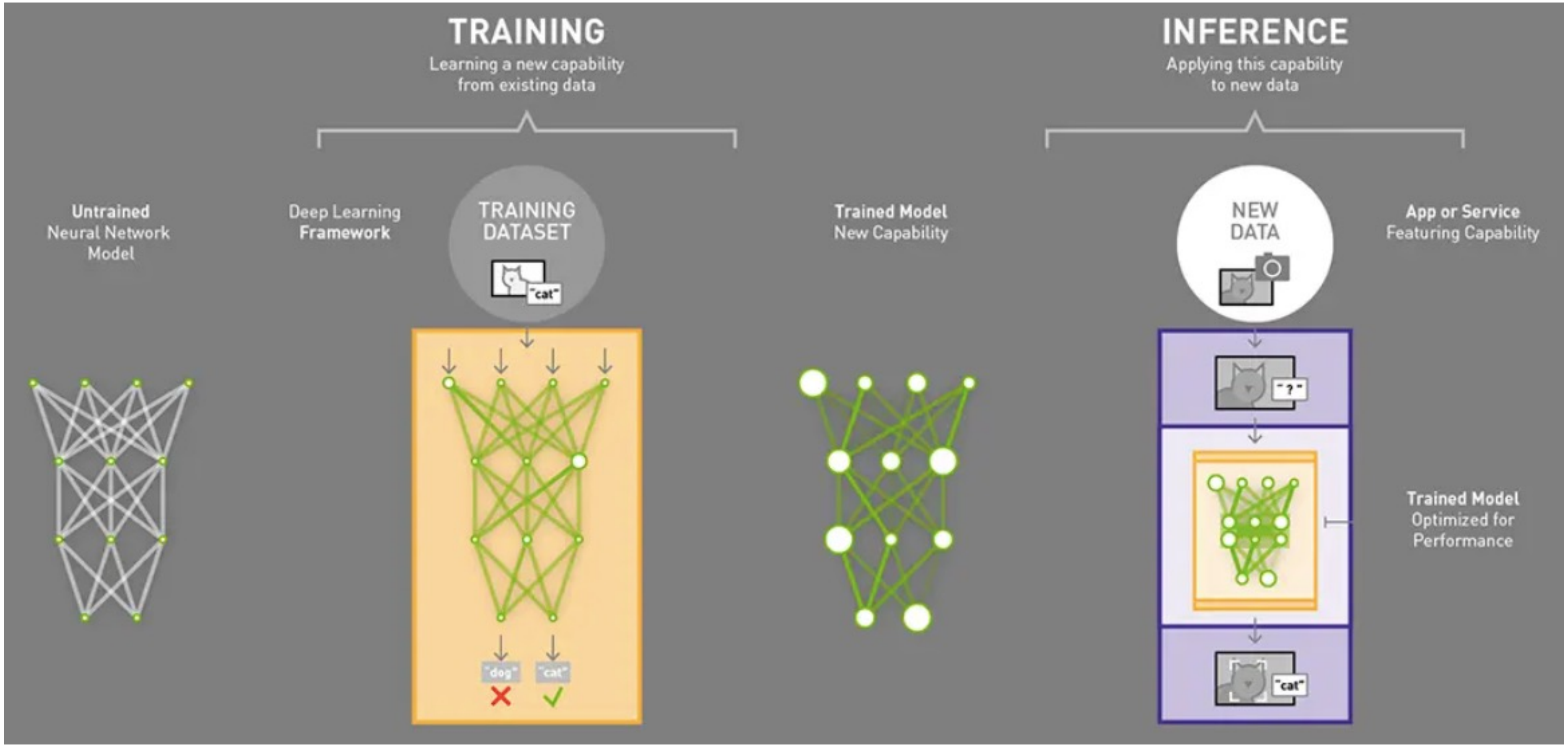
From shallow NN to Deep NN → **Deep Learning**







E.g. “supervised learning”: ML training vs inference



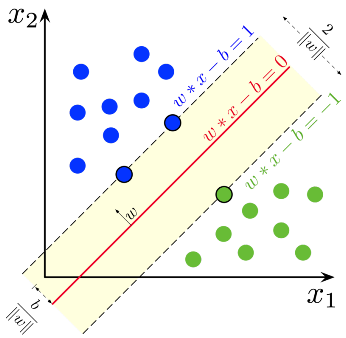
AI's past winters and current spring

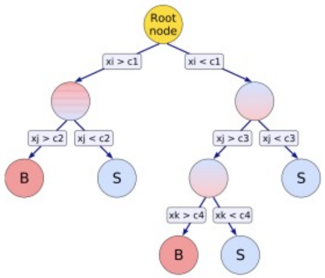
Past "**winters**" of AI

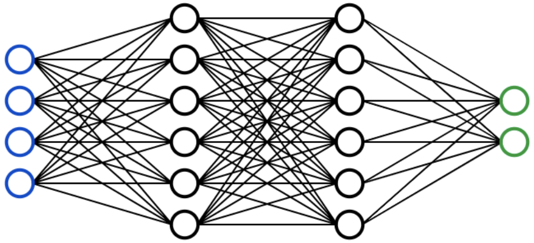
- '60: **shallow NN** hard to train
- '90: Support Vector Machines (**SVM**), Boosted Decision Trees (**BDT**), ..
- 2000+: advanced **deep NN** architectures

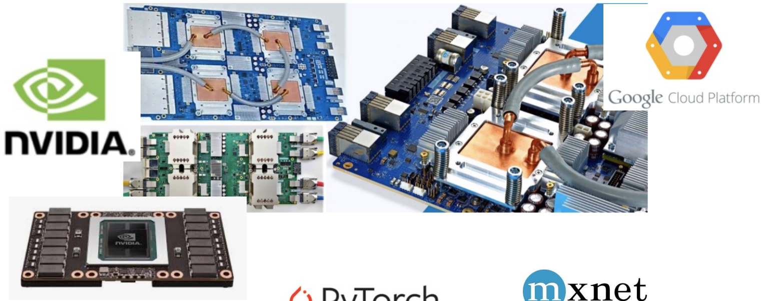
Current "**spring**" explainable by:

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



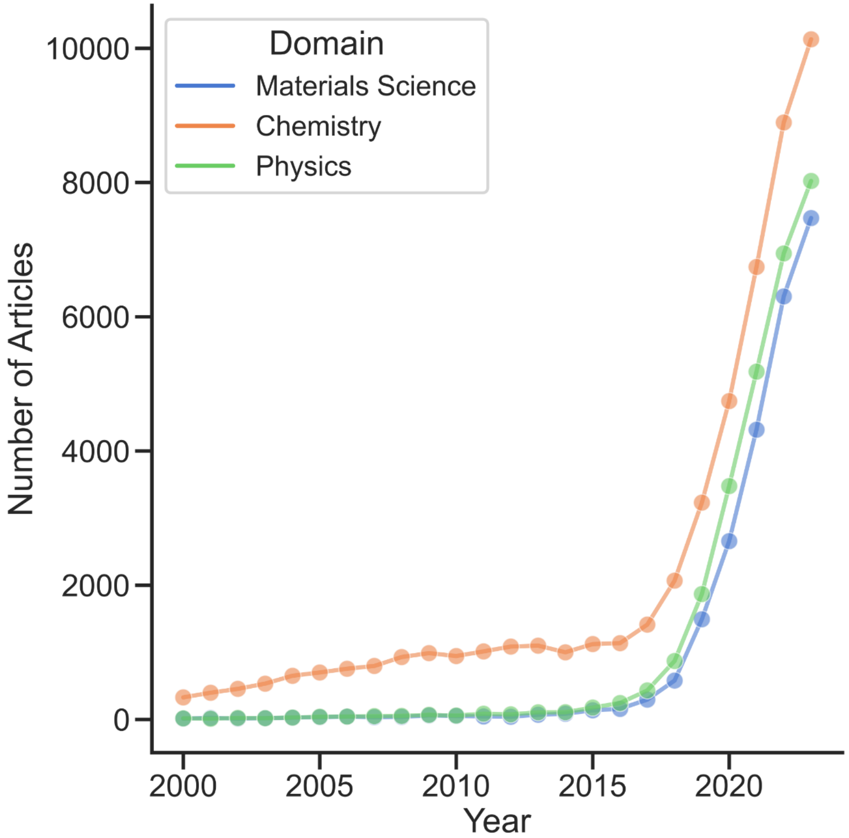






PyTorch mxnet TensorFlow Keras Spark fast.ai H2O.ai Caffe2 Microsoft CNTK theano DL4J

ML-based publications in science



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

B. Blaiszik, "2021 AI/ML Publication Statistics and Charts", 10.5281/zenodo.7057437

Pervasive ML in HEP [1/3]

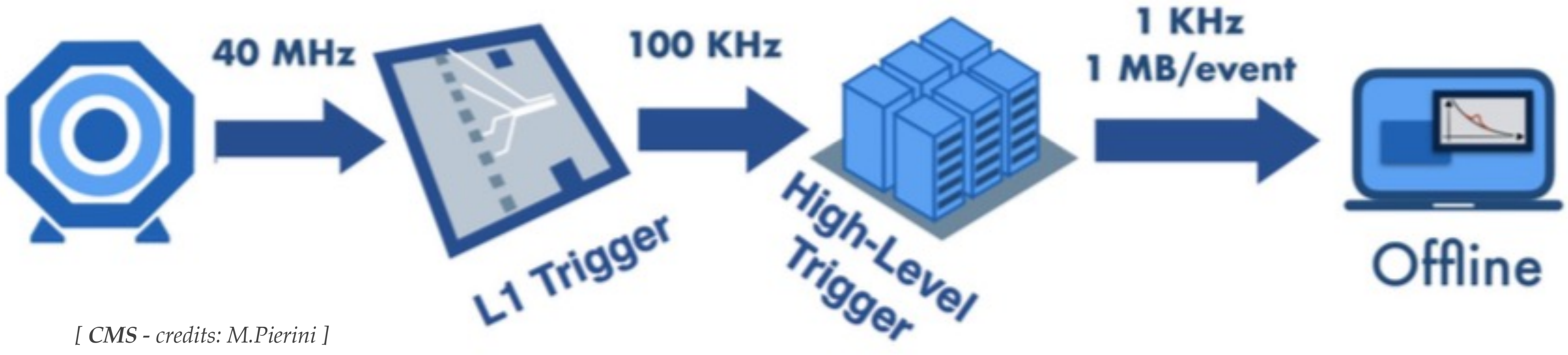
ML in **data acquisition** and **trigger**

- 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]*

Pervasive ML in HEP [2/3]

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, …)
- …

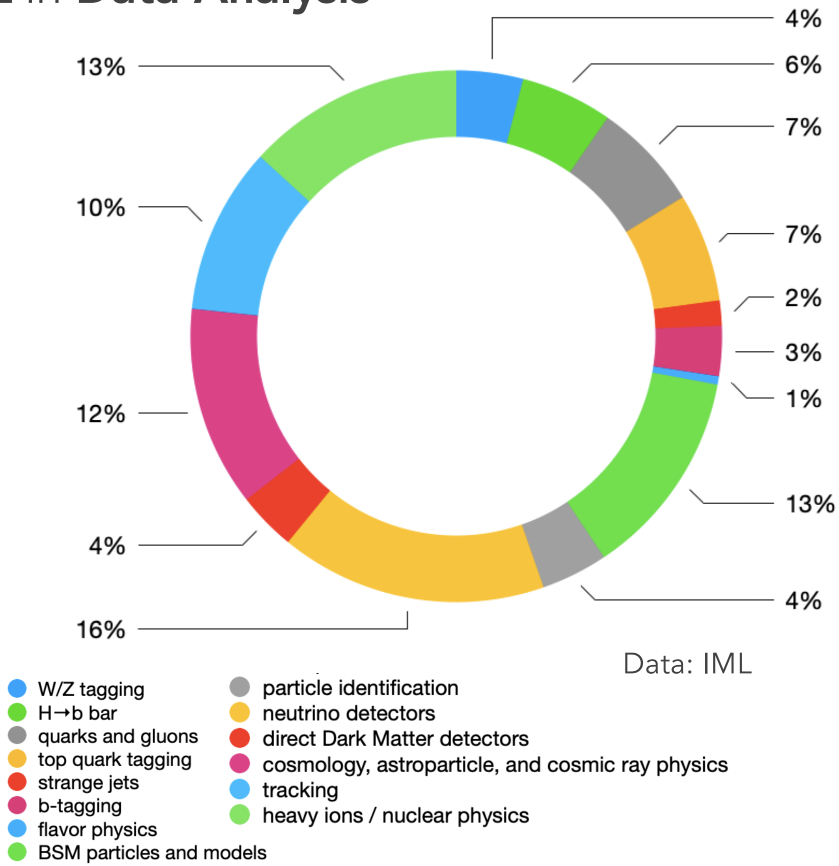
ML in Event **Reconstruction**

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, …)
- …

Pervasive ML in HEP [3/3]

ML in **Data Analysis**



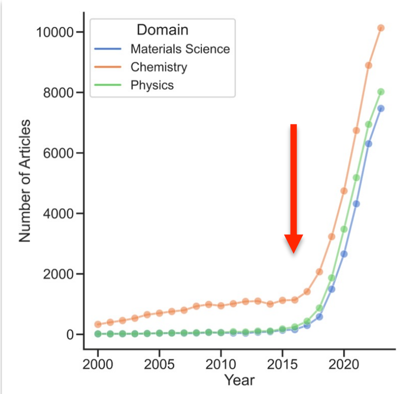
Data: IML

ML in **Computing Operations**

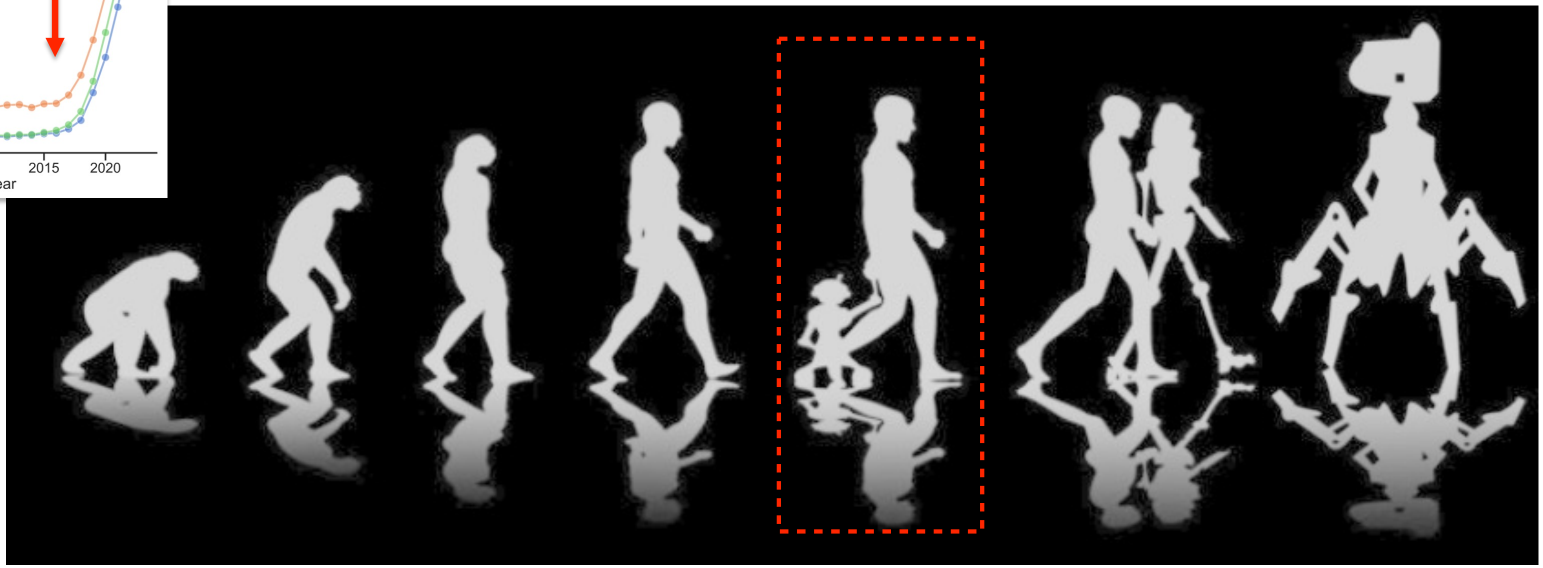
Application of ML to **non-collision (meta-)data** might help to increase efficiency and reduce the need omg personpower in Ops, e.g. automating specific tasks, developing intelligent/adaptive systems, ultimately acting on the full chain - from data collection to data analysis - and make it more agile

- 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**

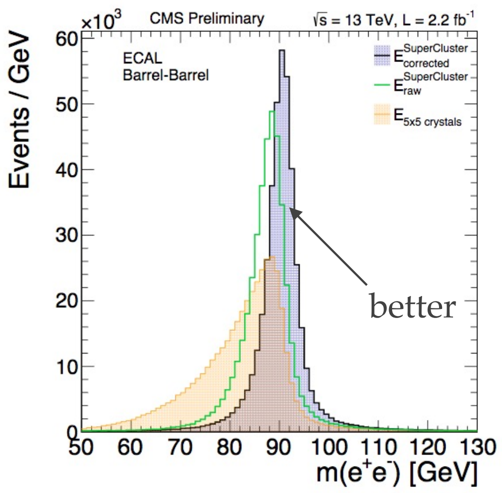
- **HEP physicists design high-level features, and send them in punt to traditional ML "shall" algorithms**

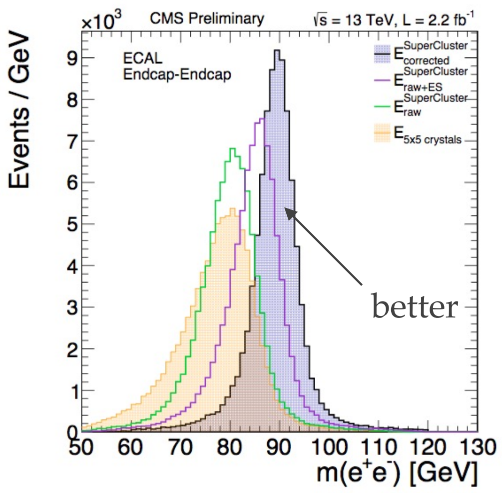
Esempio

Particle id, energy resolution, e oltre..

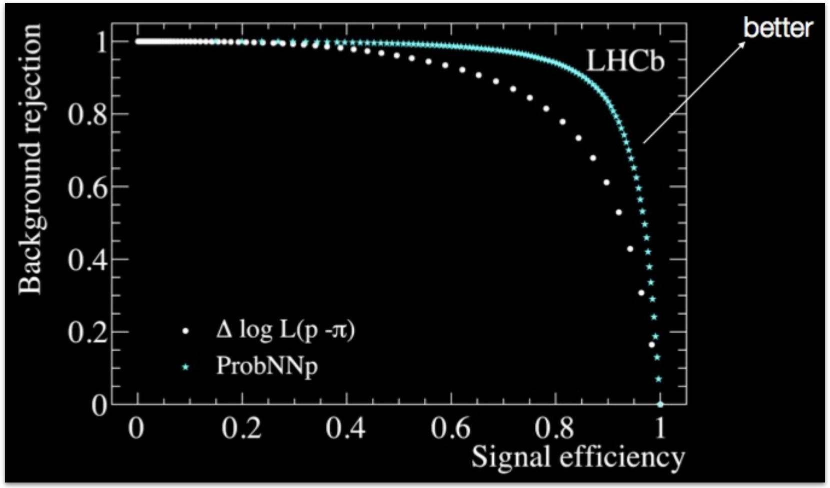
BDT usati per il learning delle energy corrections usando tutte le info disponibili nei vari sensori calorimetri

- es. **CMS**: energy sum, recupero bremsstrahlung con supercluster, inclusione della pre-shower energy, poi **energy correction con algo multivariato**





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



LHCb usa delle **NN** allenate su $O(30)$ features da tutti i sottosistemi, ciascuna allenata a identificare uno specifico tipo di particella

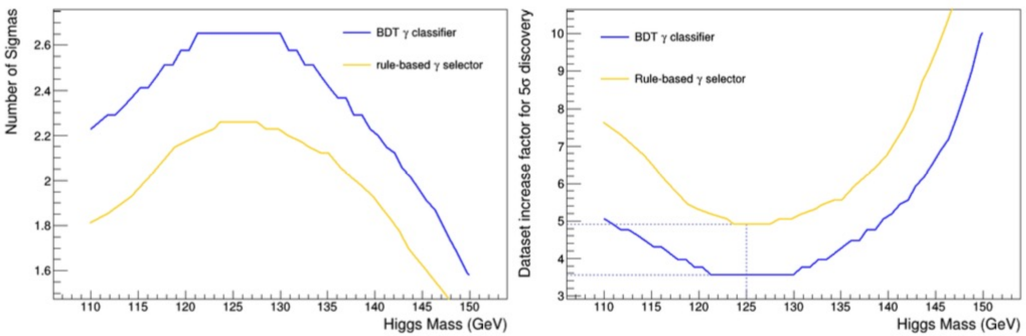
- → **mis-ID bkg/particle ridotta di ~3x (e oltre..)**

Esempio

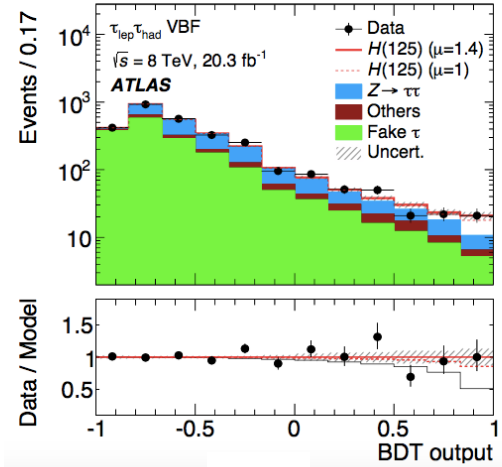
Uso di ML in scoperta e studio delle proprietà Higgs

Ruolo chiave del ML nella scoperta del bosone di Higgs prima delle attese

- es. diphoton analysis di **CMS**, **BDT** usati per migliorare la risoluzione e selezionare/categorizzare eventi
- → **sensitività aumentata di un ammontare equivalente al ~50% di dati raccolti in più**



[courtesy M.Pierini]



Studio delle proprietà H: es. τ leptons, **ATLAS** su 6 regioni cinematiche distinte, training di una BDT in ciascuna, con 12 features

→ **sensitività migliorata del ~40% rispetto a un approccio non-ML**

[1] JHEP 04 (2015) 117

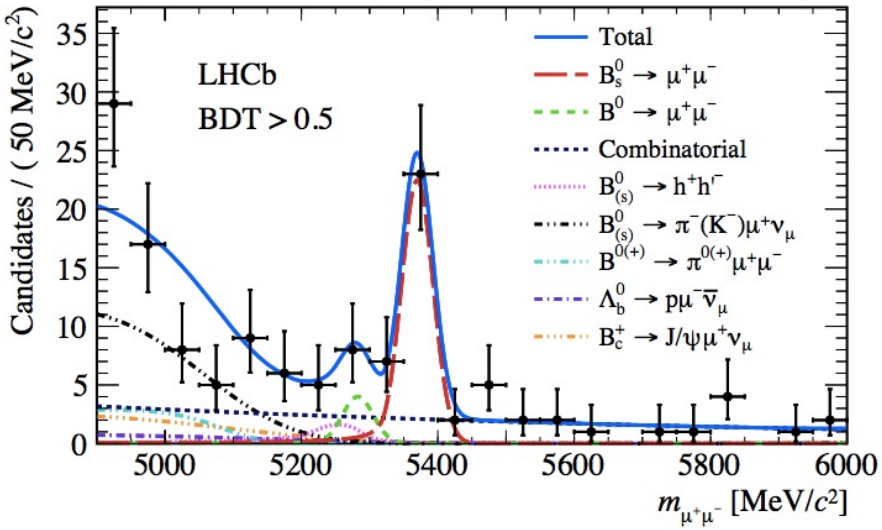
Esempio

Test di alta precisione del MS

CMS e **LHCb**: evidenza per il decadimento raro $B^0_s \rightarrow \mu^+\mu^-$ con analisi combinata [1]

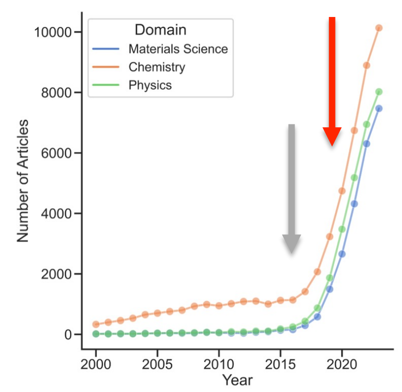
- es. **BDT** usati per ridurre la dimensionality dello spazio delle features, poi analisi spettro massa in BDT response bins → decay rate consistente con predizioni SM con precisione ~25% → constraints a estensioni SM

- → es. **per avere la stessa sensitività senza ML, LHCb come singolo esperimento avrebbe dovuto raccogliere ~4x più dati**



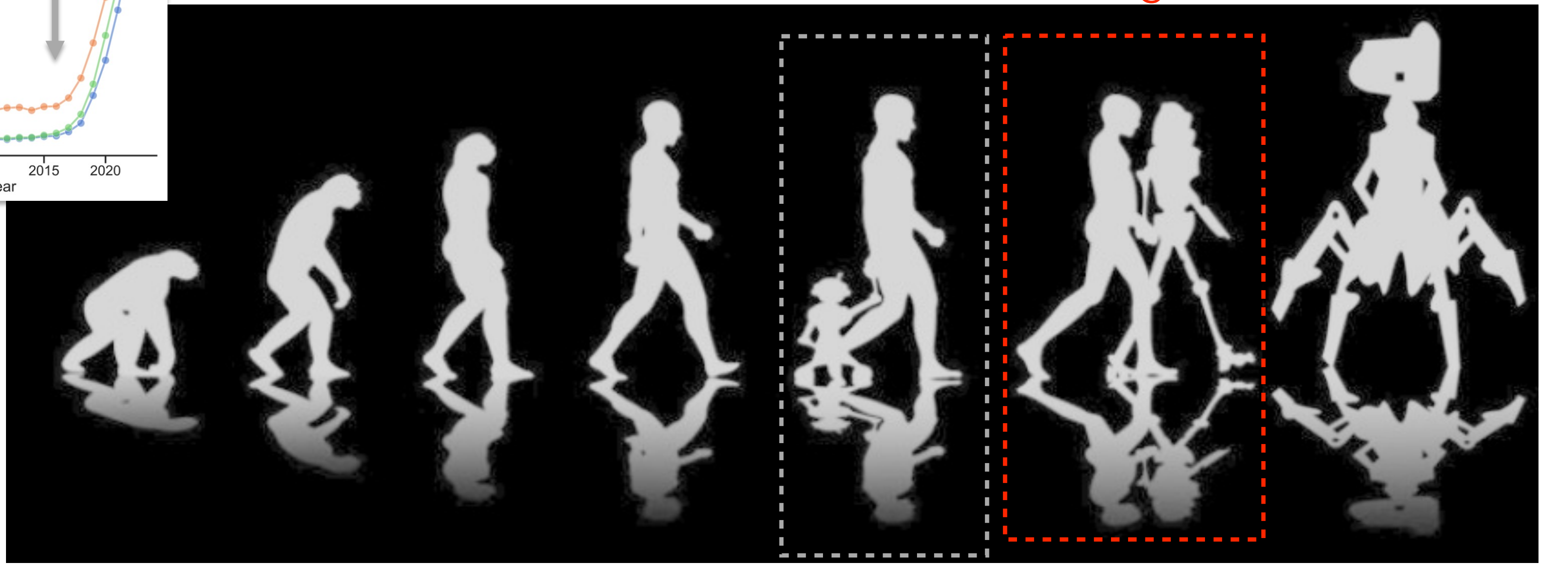
Mass distribution of the selected $B^0 \rightarrow \mu^+\mu^-$ candidates with BDT > 0.5 [2]

[1] Nature 522 68–72 (2015)
[2] *Phys.Rev.Lett.* 118 (2017) 19, 191801



ML/DL in HEP

"Traditional" ML Seeking DL solutions



Since some years, ML (DL) in HEP seeks 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**

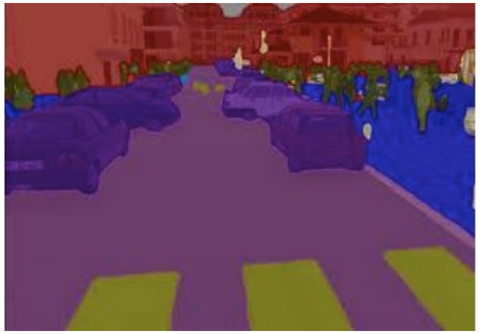
→ estract best from data, and do so by exploiting any architecture that might work for a given use-case (e.g. input from CV and NLP solutions..)

Convolutional Neural Networks (**CNN**)

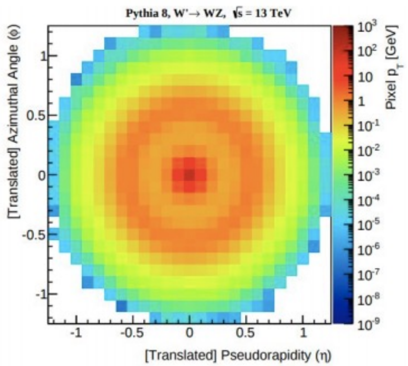
CNNs offer translational-invariant feature learning, robustness against noise, versatility in application to a variety of domains

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

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



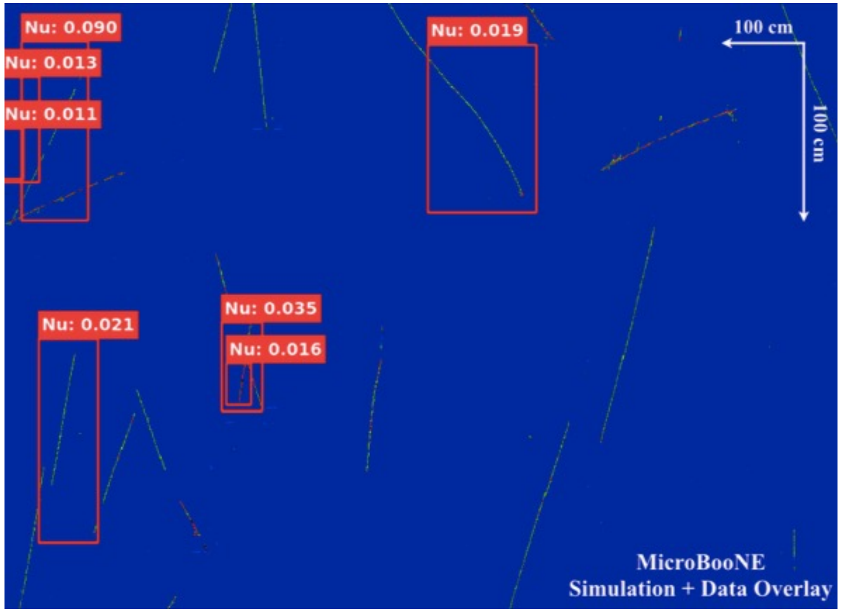
HEP:
3D imaging in detectors, event classification,
automation of hist checking (e.g. data quality), …



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

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

“HEP is so different from other applications..”. Are we?



Rivelazione di tracce di **neutrini** su cosmic background events (metodo: **CNN**)

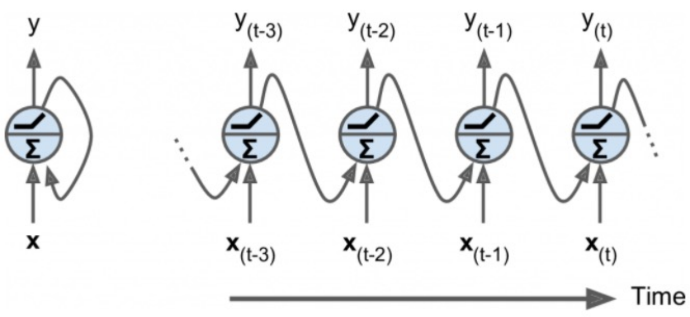
Rivelazione di **aeroporti** da immagini satellitari (metodo: **CNN**)

Recurrent Neural Networks (**RNN**)

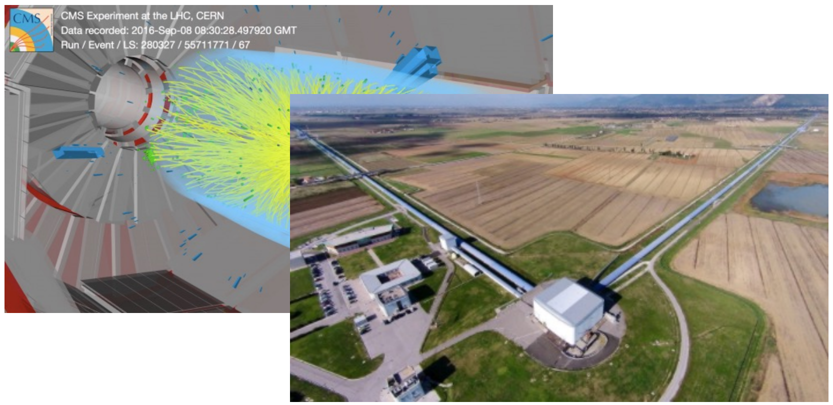
RNN allow to handle variable-length inputs and process time-series, accumulating and using together info at various times in the sequence

- Based on “recurrent neurons” (backward-pointing connections)
- A variety of application in time-series of all kins, 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

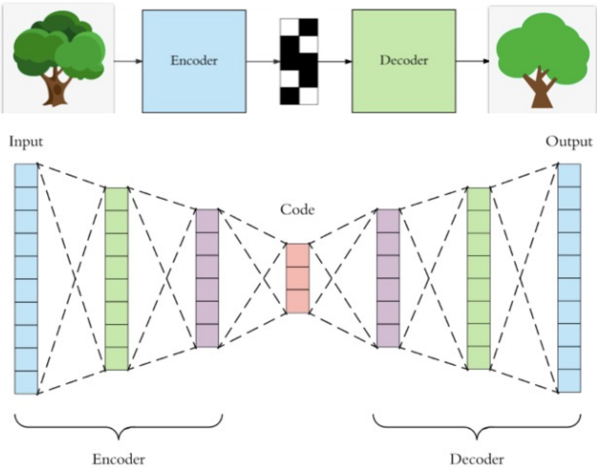


Autoencoder (**AE**)

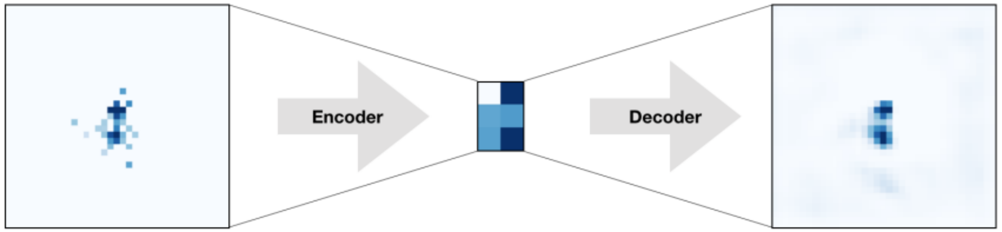
AE is a"data-specific" compression algorithm, able to reduce dimensionality and extract "the juice" of an input

- 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 (intestìni events are those whose decoding in output is dtstant from the input, according to a given metric



Potenziale strumento per scoprire nuova fisica in modo "<u>unsupervised</u>"

Esempio

(V)AE per "new physics mining" - at LHC and beyond

A **Variational AE** has been proposed (CMS) for "new physics mining"

- Traing on known SM processes, build threshold to identity "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, i.e. model-dependent hypothesis testing
- topologie ricorrenti nel catalogo possono ispirare search focalizzate, e anche costruzione di modelli teorici
- Target: up to the trigger level..

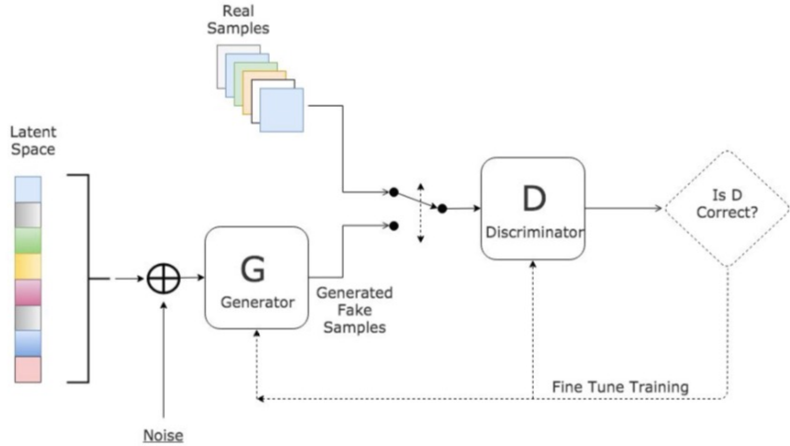
Note: going from **discriminative AI** to **generative AI** …

[1] JHEP05 (2019) 036

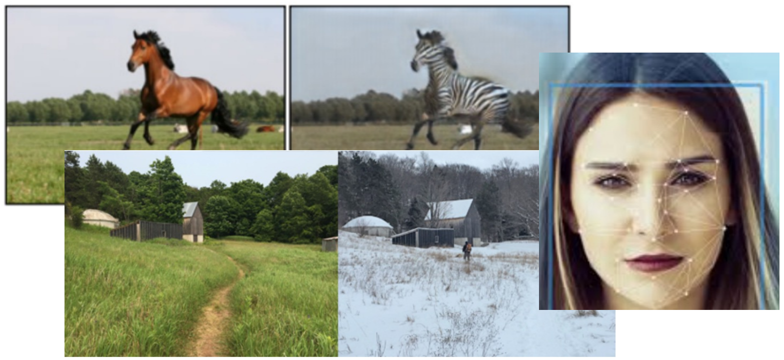
Generative Adversarial Networks (**GAN**)

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

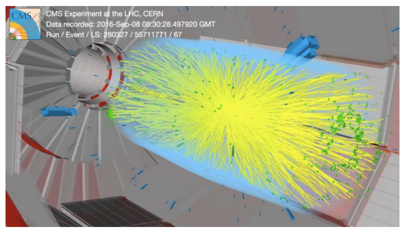
- G creates images from noise, D classifies them real vs fake
- Once trained one against the other, G pursues its goal which is to confuse D, and in the process it learns how to creare fake but very realistic images



Industry:
image editing, data generation, security, ..



HEP:
Simulate the detector response at reduced computational costs



Data sparsity and point clouds

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

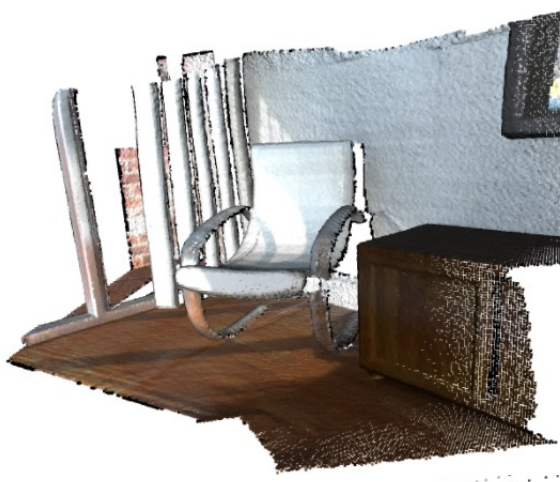
- Granularity and occupancy in HEP sensors
- Popolation of stars and galaxies in the cosmos
- Molecular description in computational chemistry

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

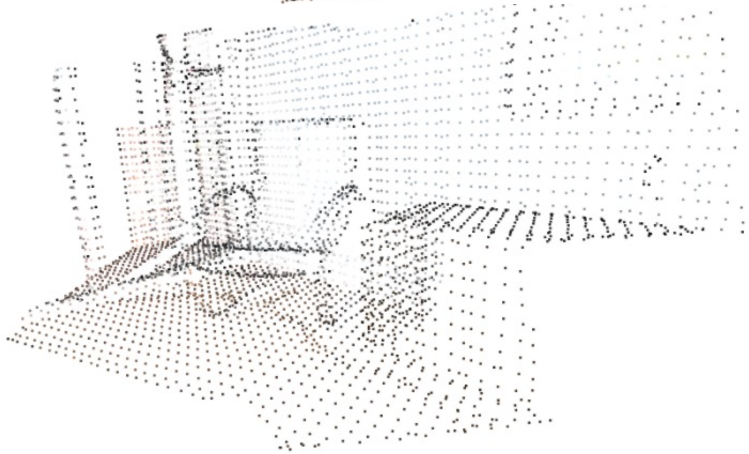
- e.g. EM shower → E deposited in active volumes of an ECAL

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 (co far) to the HEP community
- e.g. problems configurable in extraordinarily similar way to how social media datasets are treated (!)



3D semplification



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

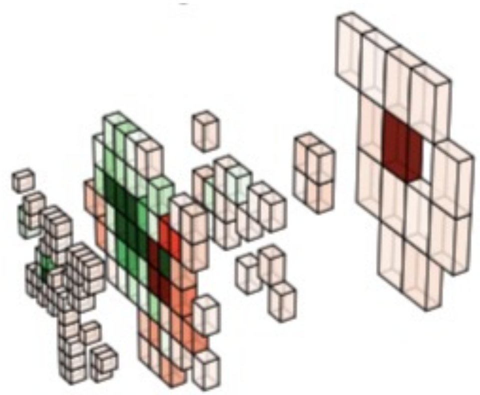
In a sparse representation, we need a metric that defines proximity in the abstract space of features

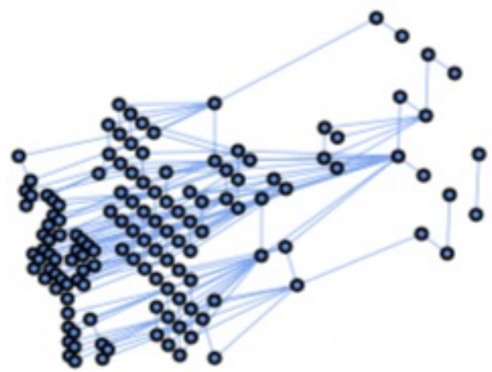
How? Migrating from“**datasets**” to “**graphs**”

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

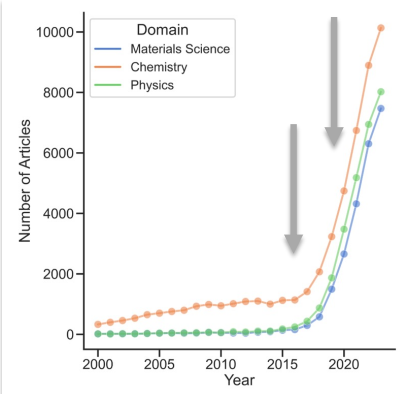
GNN → build a data structure, (V,E) with V=vertex and E=edges, choose possible types of vertices (if no prior one builds a fully-connected graph), etc

GNNs start to be part of present/future of DL in HEP



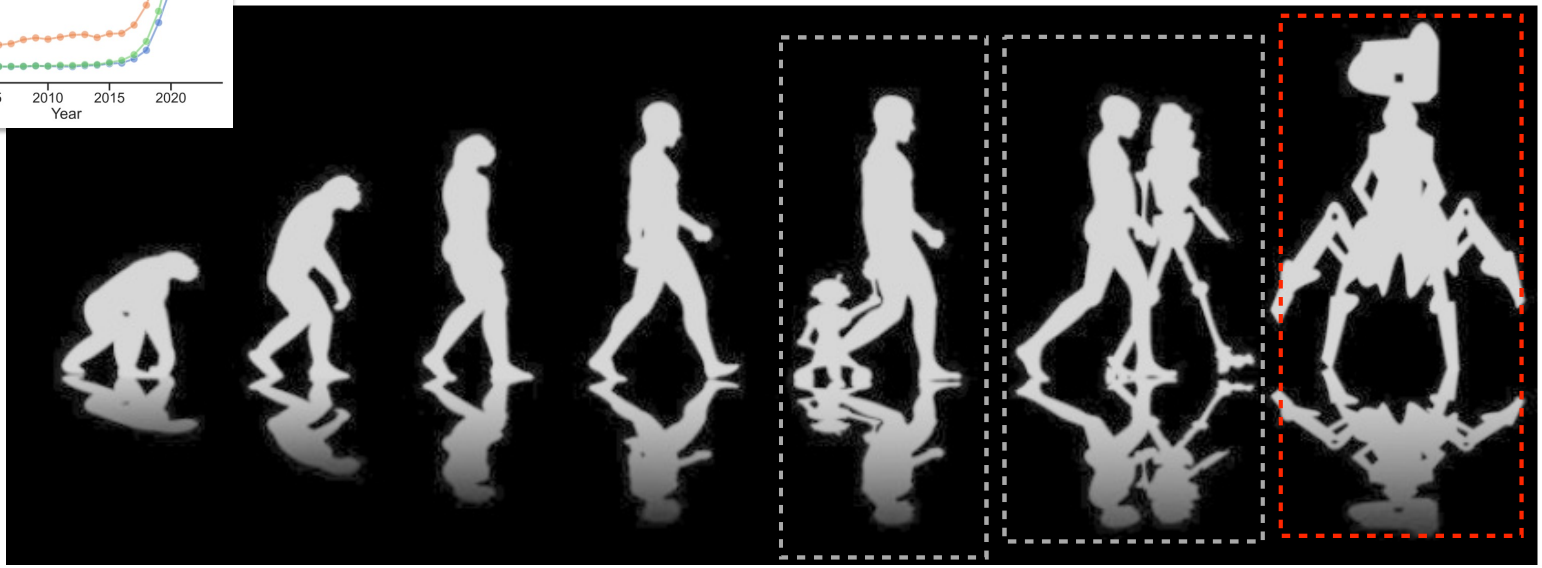


ML/DL in HEP



?

"Traditional" ML   Seeking DL solutions   What next?



?

Natural Language Processing (NLP)

Important advances in the decade 2010-2020, thanks to deepNN ("Deep Learning")

- **Machine Translation** → e.g. language A - Language B translation
- **Text/Document Classification** → e.g. doc clustering, sentiment analysis, ..
- **Entity Extraction** (a.k.a. Named Entity Recognition) → extracting relevant information from unstructured text (e.g. vital parameters of a patient from her/his medical record)
- **Summarisation** → e.g. generating concise summaries of docs
- **Question-Answering** → multi-domain and multi-language factual knowledge
- **Digital Assistants** → e.g., Amazon Alexa, Google Assistant, Apple Siri, ..

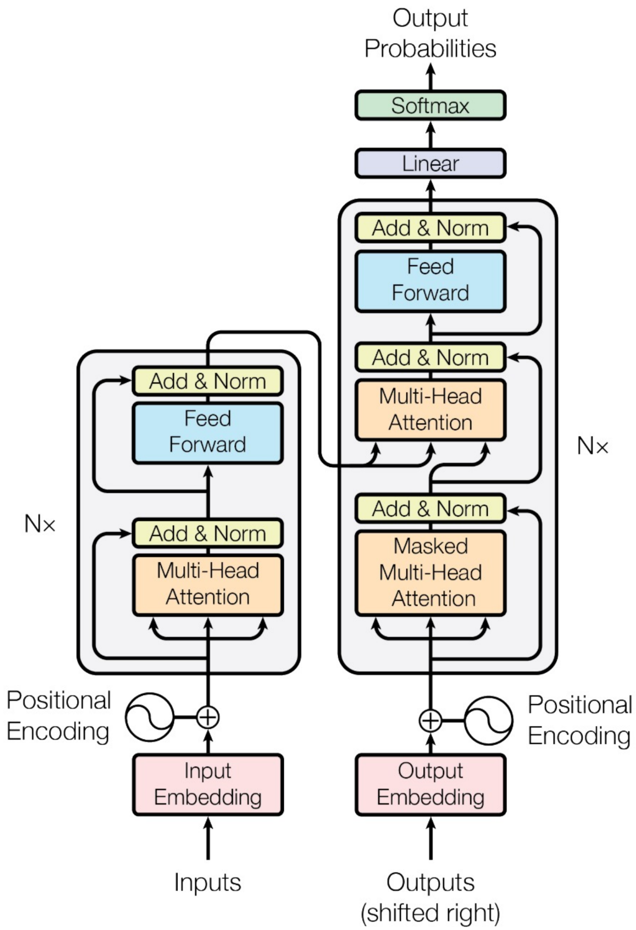
NLP "first generation" (2010 - ~2017):

- RNN, or Long Short-Term Memory (LSTM), trained on text corpus to learn the structure of language
  - Handle text as a sequence of symbola ("token"), mapped to multidimensional vectors ("embedding") and processed through a hierarchy of levels

But then.. a "second generation" came in..

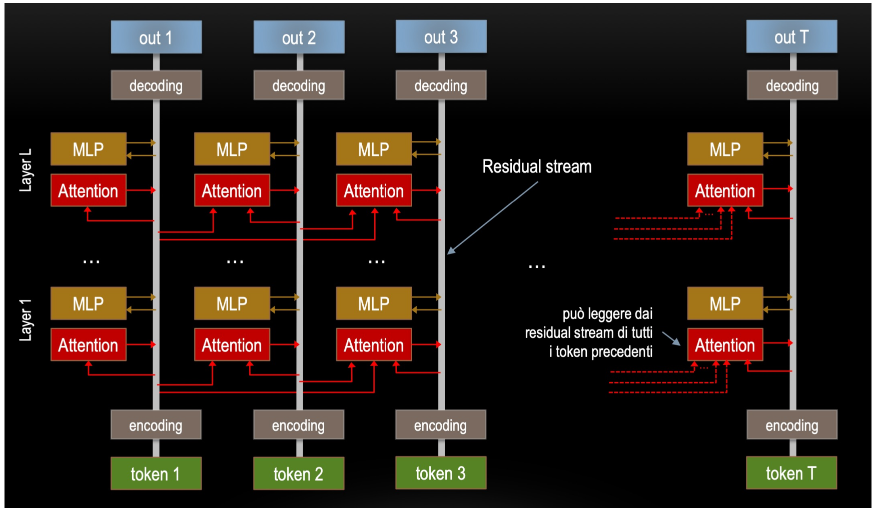
BTW: is NLP relevant for HEP? → ***Yes.***

"Attention is all you need"

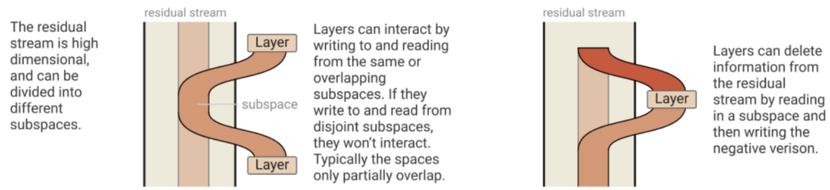


Google (Brain), 2017: a new deepNN architecture called **Transformer**

- A native seq-2-seq, with a key element: the "**attention**" mechanism, that allows to pass the meaning of a token in the context of other tokens in the same text
- Soon became the reference model for language processing (see next)



Residual streams: info channel at high capacity (d=12288 in GPT-3). Attention and MLP modules can read/write from/to subspaces of residual streams w/o interference with message passing.



Language Models: from LM to **L**LM

"Second generation" → modern **Language Models** (**LM**): a class of probabilistic models that learn patterns in NL via more advanced methods

- Architetture change: from RNN to **Transformer**
- Tokenization → multi-language and multi-domain
- Training on extremely large text corpuses (incl. source code!)

Major players:

- **GPT-3** (OpenAI, 2020) from which **ChatGPT** derived (OpenAI, 2022): trained on 45 TB of text (equiv. 2000x Wikipedia), Estimated training cost: 4.6 M$. **GPT-4** (OpenAI, 2023). Estimate: 1 order of magnitude more params
- **Bard**/**Gemini** (Google), **Claude** (Anthropic), **LLaMA** (Meta)

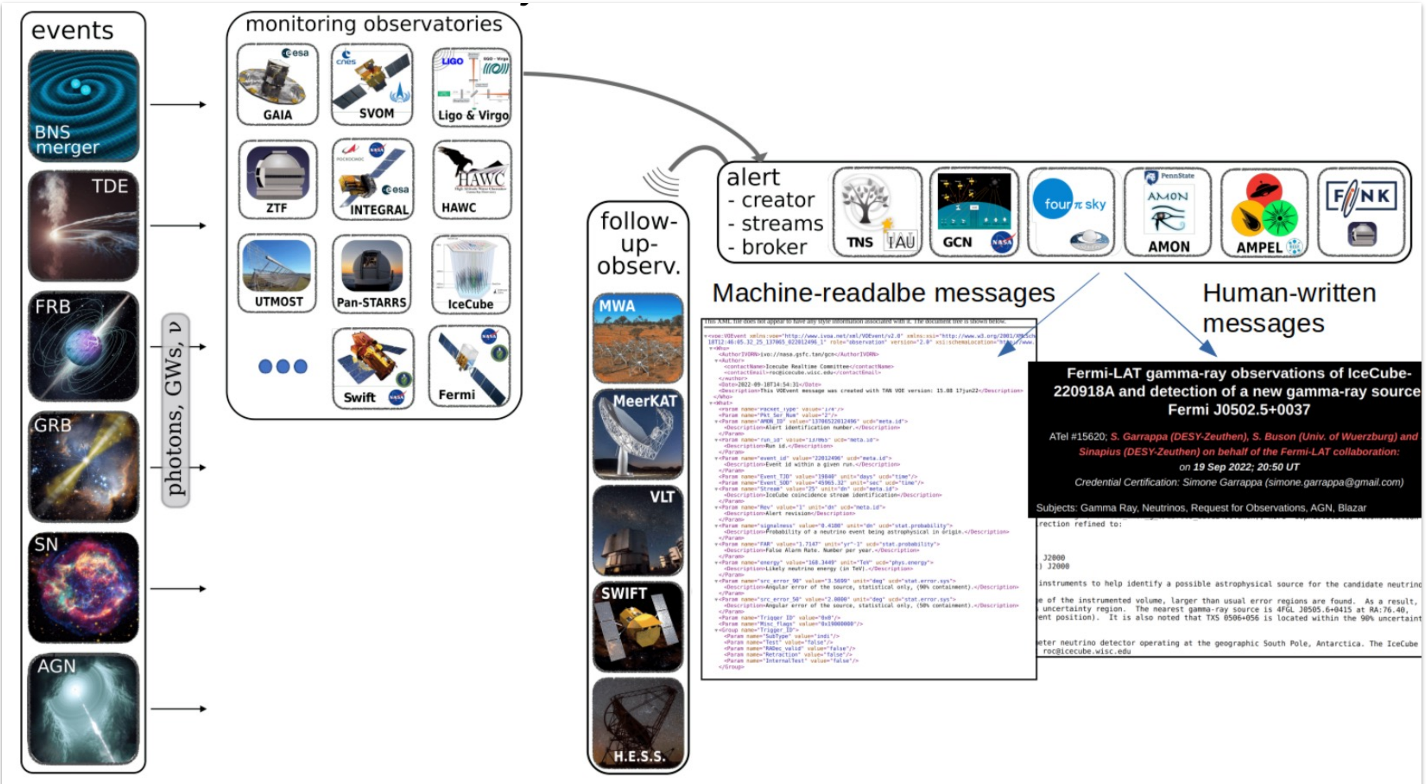
Significant efforts into scaling LMs into **Large LMs** (**LLM**)

- → training bigger models on more data with greater compute
- → steady+predictable improvements in their ability to learn patterns
- This could be observed in improvements to *quantitative* metrics.. but also *qualitative* (!)

Before getting to this, some examples..

LLMs for **Multimessenger Astronomy**

Observations in astronomy:



33

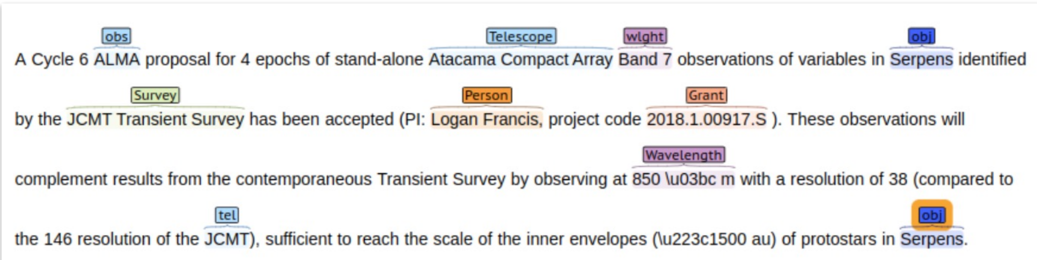
LLMs for **Multimessenger Astronomy**

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.

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

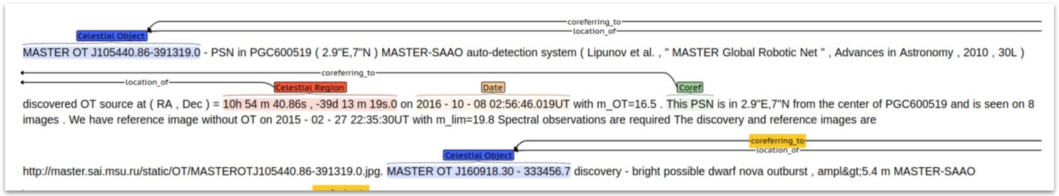
An example text based on **ATel** messages (astronomerstelegram.org)

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



From entity recognition…

… to semantic relationship extraction



[credits: D. Kostunin, A. Alkan, A. Chaikova, V. Sotnikov et al.]

A 7B LLM fine-tuned on **Cosmology** papers and textbooks

"*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

Popular science

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.

Detailed Instrumentation

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.

Data analysis principles

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.

Current cosmological parameter constraints

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 τ=0.054±0.007, 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.

[credits: Tijmen de Haan]

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)

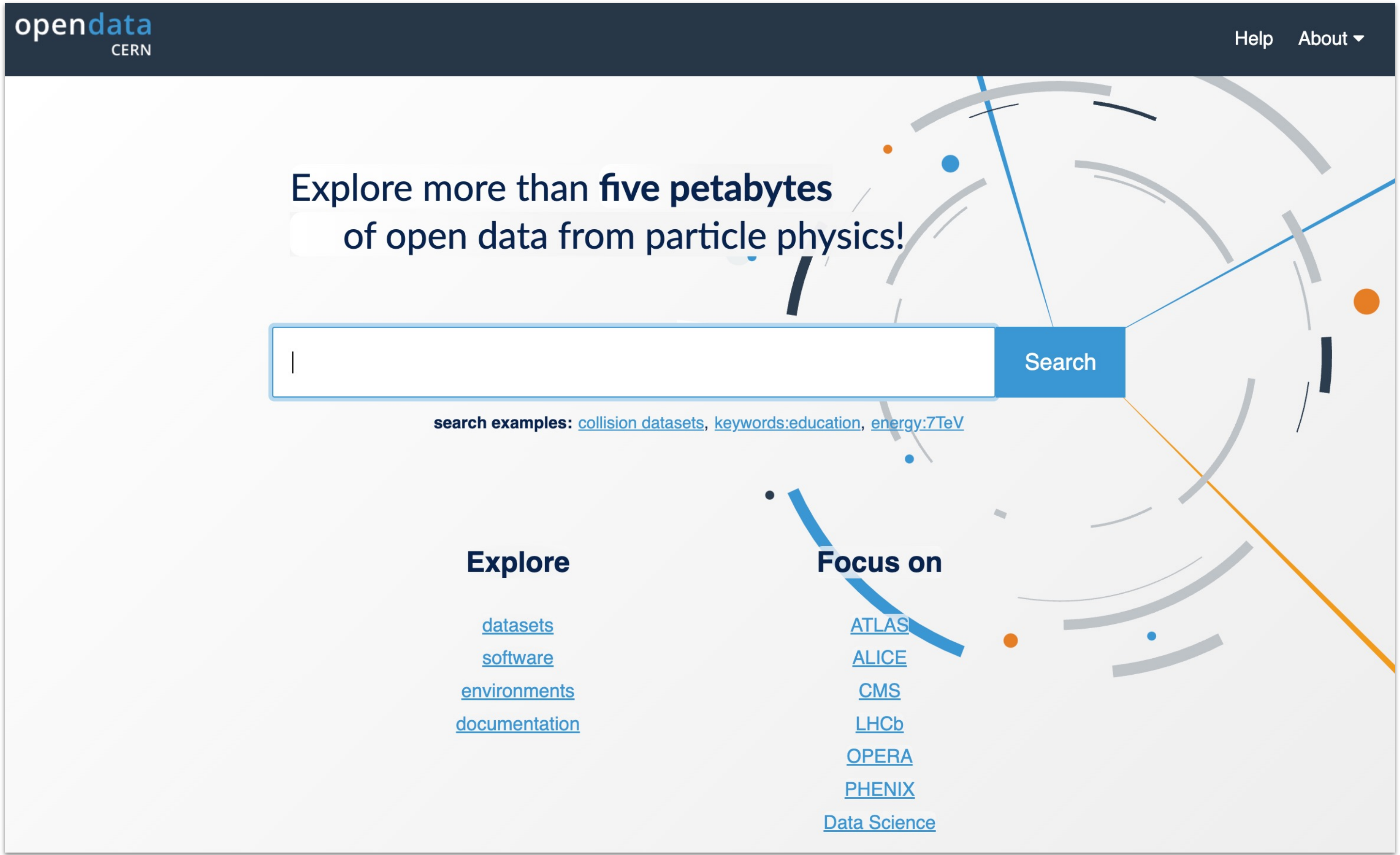
- Docs: twiki (>2k), sw docs (>500), e-groups/mails archive (>10k), indico meetings' agendas incl. attached slides and minutes (>440k), Mattermost, Jira tickets, experiment' papers and internal notes (>66k)
- Either HTML or scraped into markdown

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

[credits: Cary Randazzo]

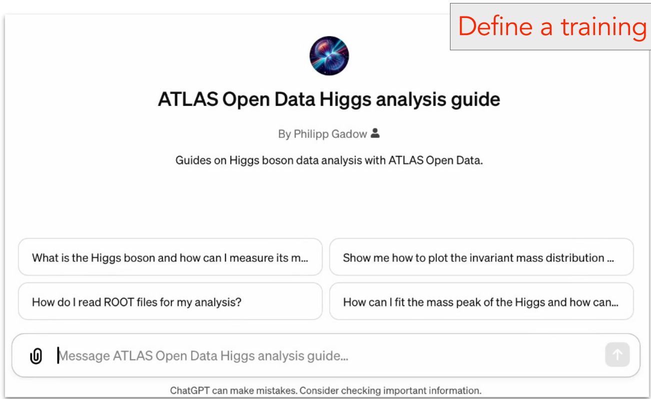
Educational Outreach with **AI-Assisted CERN Open Data**



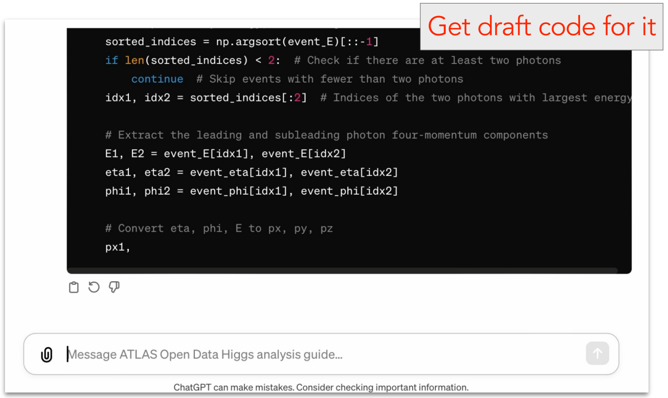
https://opendata.cern.ch/

An example: a Higgs analysis guide

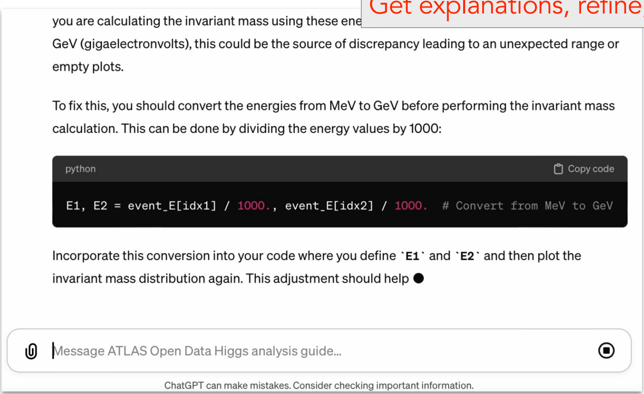
Define a training goal



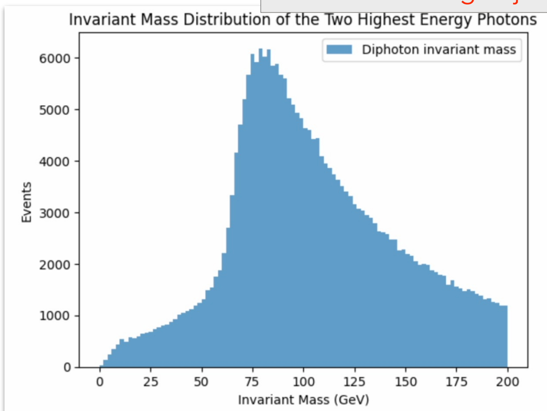
Get draft code for it



Get explanations, refine, learn



Reach the training objective



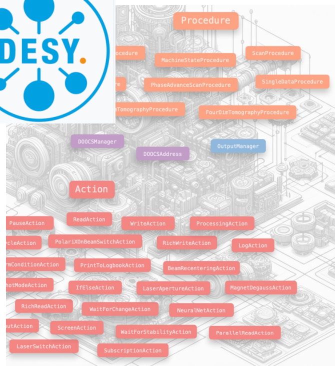
[credits: Paul Philipp Gadow et al]

LLM for **particle accelerators**

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-though 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

[credits: F. Mayet, J. Kaiser. F. Rehm et al]

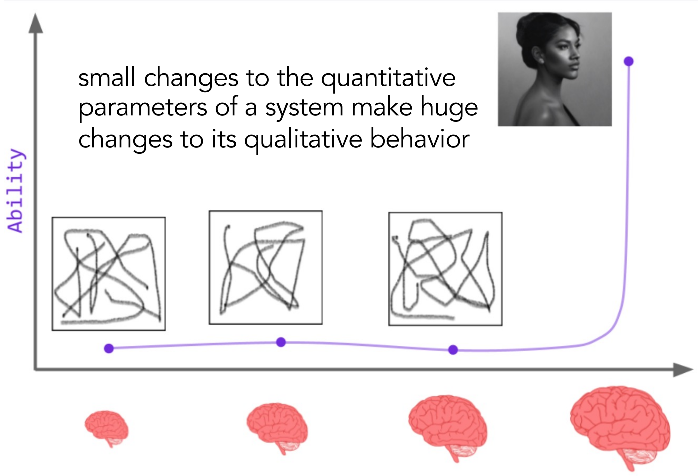
Large Language Models and “emergent abilities”

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

- Scaling LLMs → 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

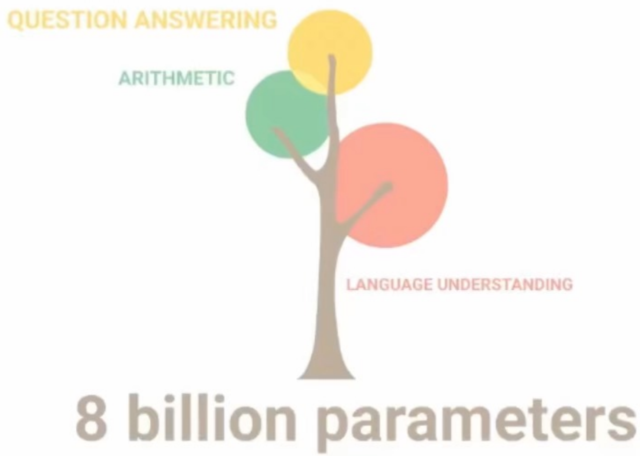
Examples:

- 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



More → Sparks of AGI: early experiments with GPT-4, Mar 2023, arXiv:2303.12712

Large Language Models and "emergent abilities"



Emergence and prompting

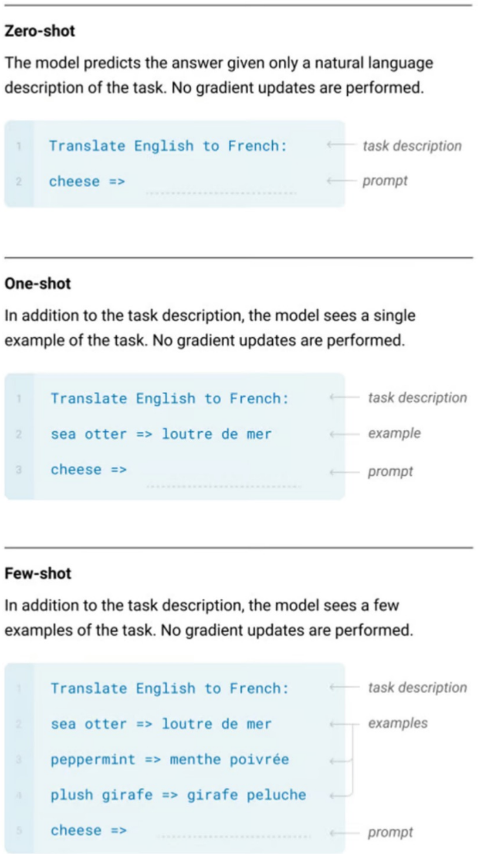
LLMs can be prompted: no fine-tuning of a model, but give it NL instructions

in-context learning: after training (frozen weights), a model is prompted a set of "gold-standard" examples to illustrate how to complete a task for which it was not trained, and it learns

Few-shot prompting

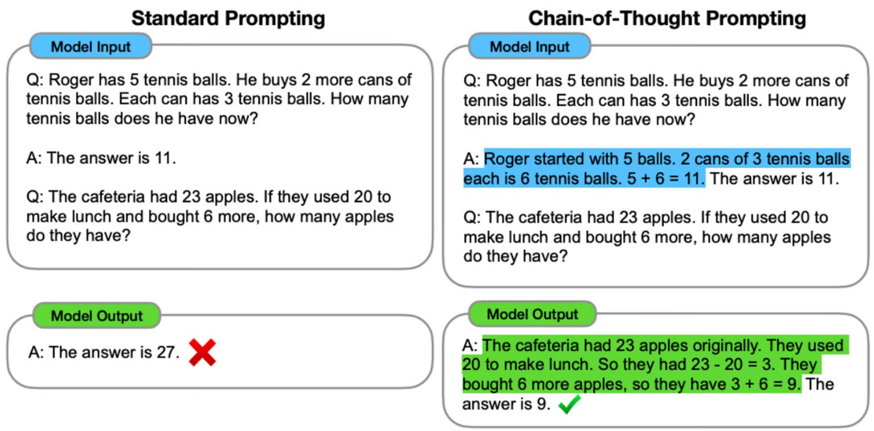
emergence at ~100B params for a wide range of cases

→ *prompt engineering, Langchain, ..*



In addition to in-context learning, another set of interesting abilities stem from **prompt augmentation**

Chain-of-thought (CoT) reasoning



E.g. it has been demonstrated (in some cases) that just adding "*let's think step by step*" may trigger multi-step reasoning and lead to impressively increased accuracy in arithmetic tasks

Emergent Abilities of LLMs, Jun 2022, arXiv:2206.07682

Chain-of-Thought Prompting Elicits Reasoning in LLMs, Jan 2023, arXiv:2201.11903v6

Biological vs Machine "intelligence"

Mammalian biological brains

Cat brain

0.760 billion neurons
10 trillion synapses

Human brain

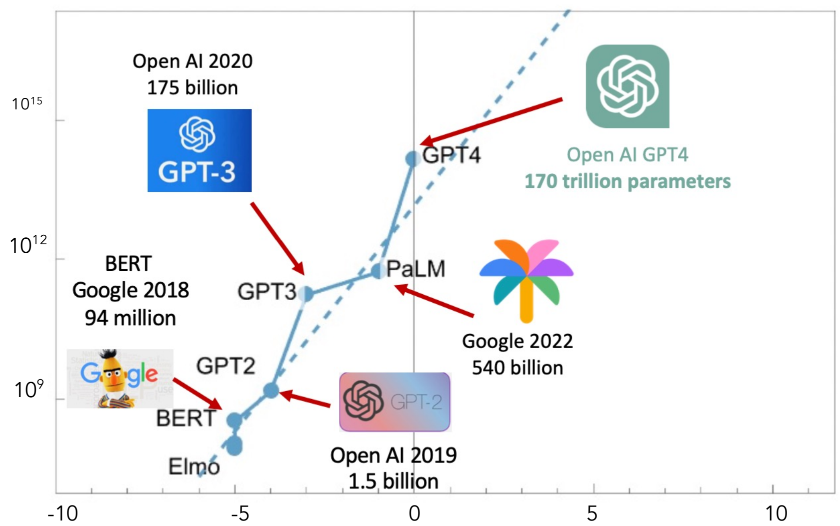
80 billion neurons
150 trillion synapses

size of GPT 3.5

GPT-3

size of GPT 4

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) → 10^{22} ops, to be compared with LLM training time, around 10^{25} ops
  - And it consumes more.. **Red AI** is a serious issue!

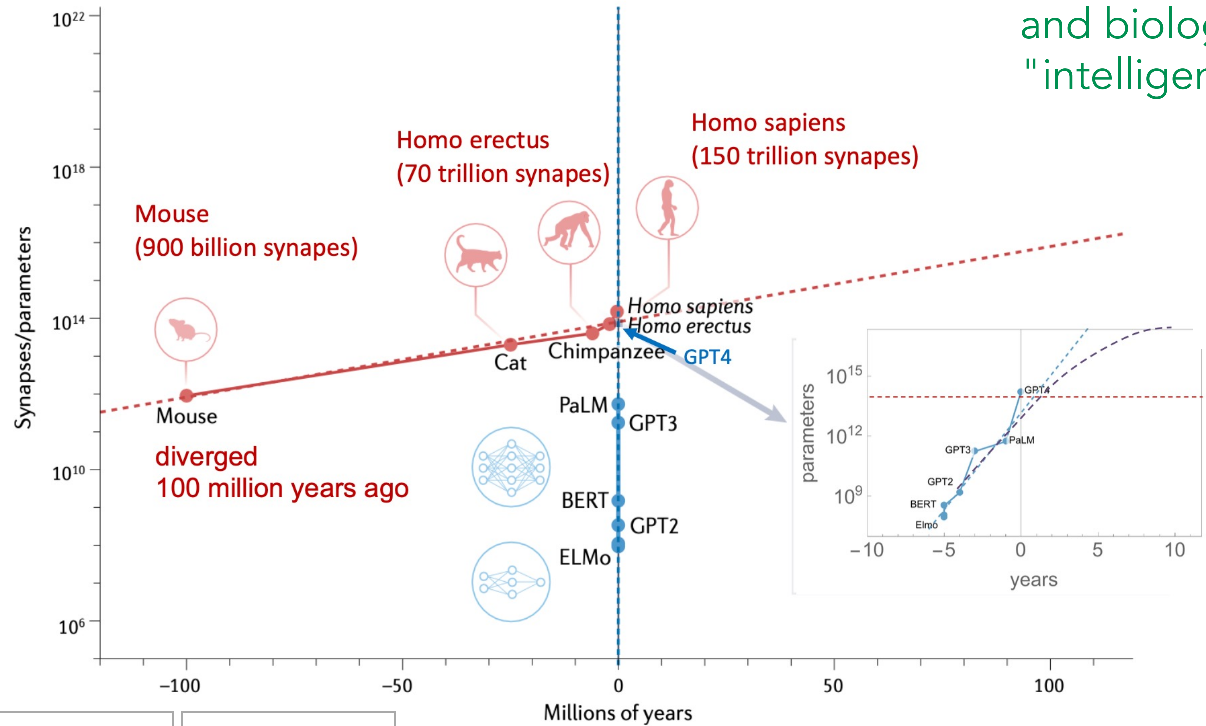
Credits: M. Schwartz

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 grown will soon be superhuman !

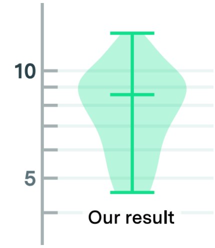
Nature Reviews Physics, 4, 741–742 (2022)

Credits: M. Schwartz

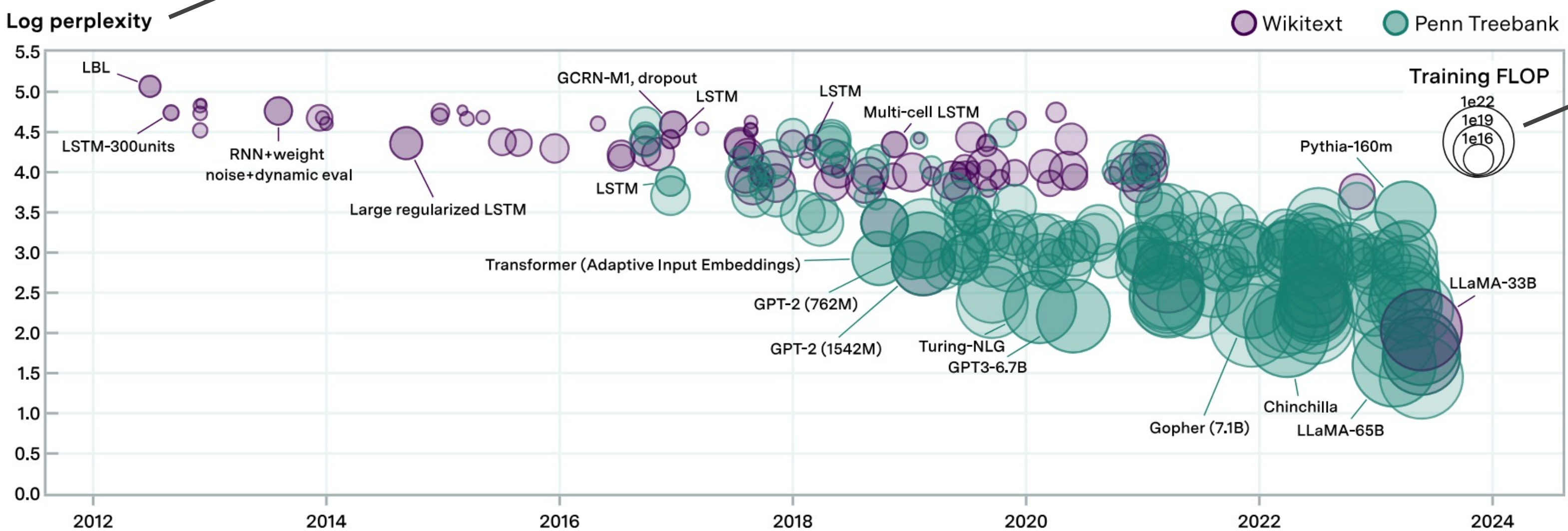
Algorithmic progress in LLMs

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!



Loss



Compute

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

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

Algorithmic progress in LM, Mar 2024, arXiv:2403.05812

Credits: M. Schwartz

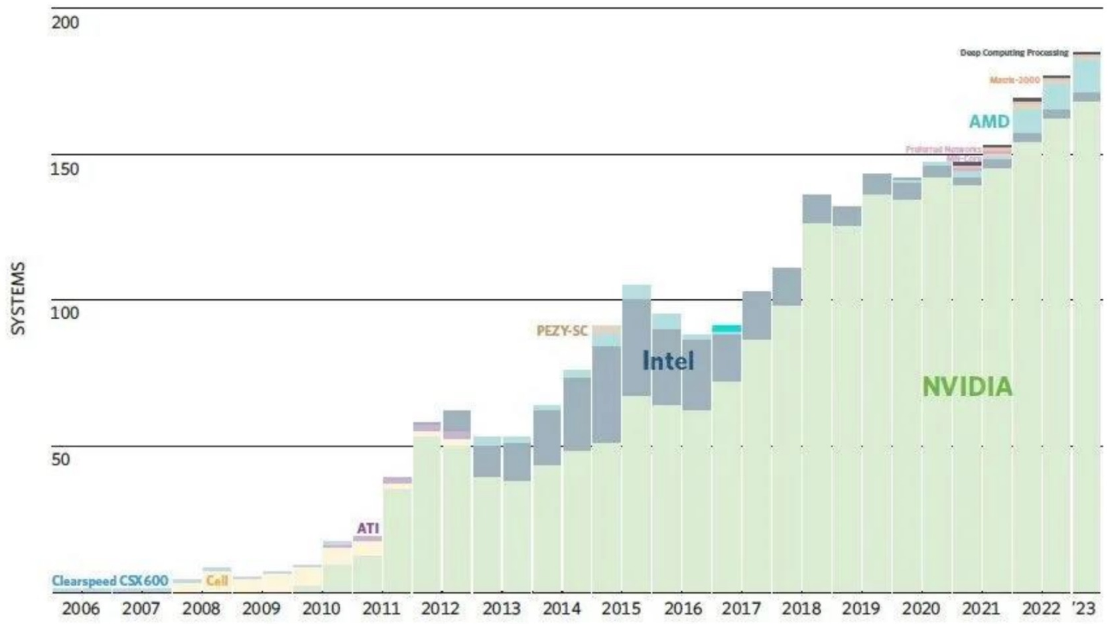
A glance to hardware: the NVIDIA “gravity”

Of the 184 accelerated machines on the TOP500 (June 2023) list, 167 have Nvidia GPUs

- 11 of them have AMD GPUs
- 6 have other kinds of accelerators like Intel Knights coprocessors.

AMD has 5.9 percent of accelerated system share in GPUs, compared to 90.8 percent for Nvidia (95% in AI)

Nvidia is <u>not</u> slowing down in 2024..



“but physics requires creativity..” (?!)

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.

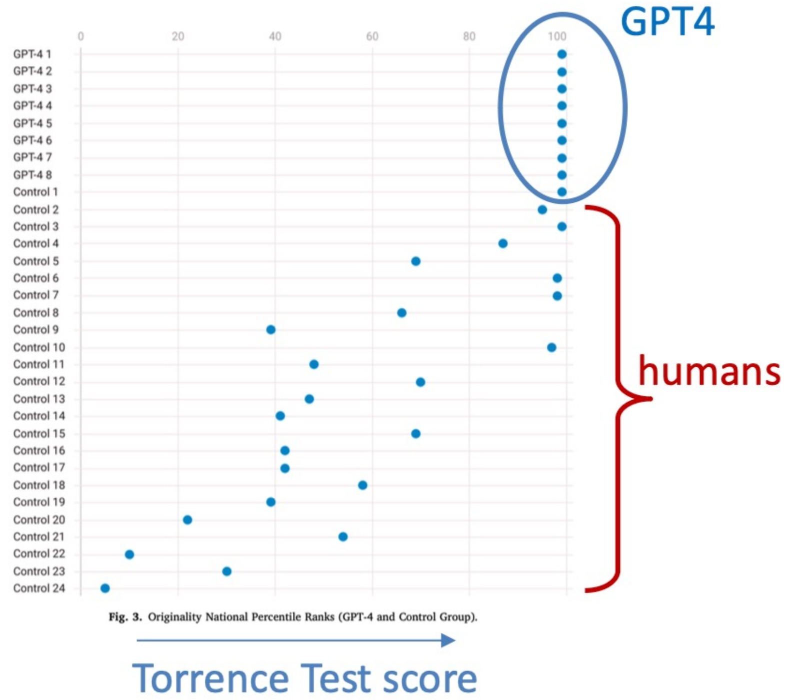


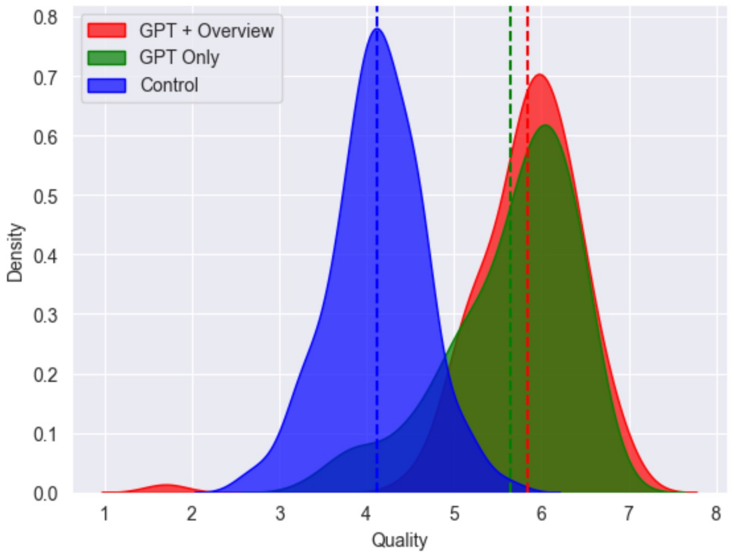
Fig. 3. Originality National Percentile Ranks (GPT-4 and Control Group).

GPT4 more creative than 99% of humans..

Nature *Sci Rep* **14**, 3440 (2024)

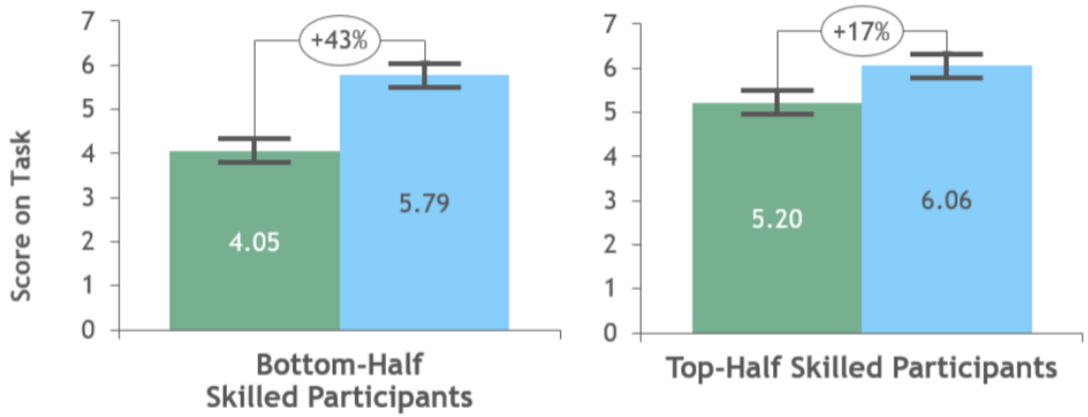
Is this “augmented" intelligence?

Can LLM help “consultants”?



Yes, and by a lot!

AI-assistants as skill-levellers



Average consultants using AI perform as good as best consultants

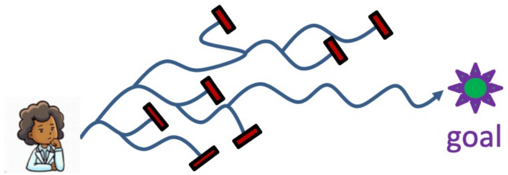
Navigating the Jagged Technological Frontier (…), Harvard Business School, 2023)

Machines not limited by biology

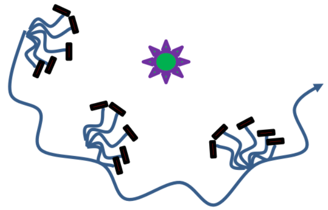
Can AI be a skill-leveller (or more) for e.g. HEP theory?

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 depsite many dead ends



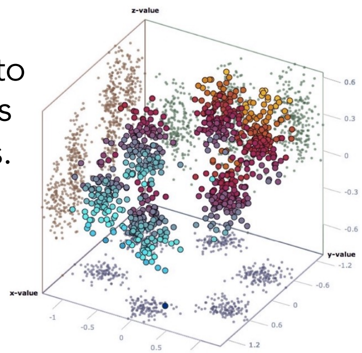
Are we even making forward progress anymore?



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

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

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

Credits: M. Schwartz

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

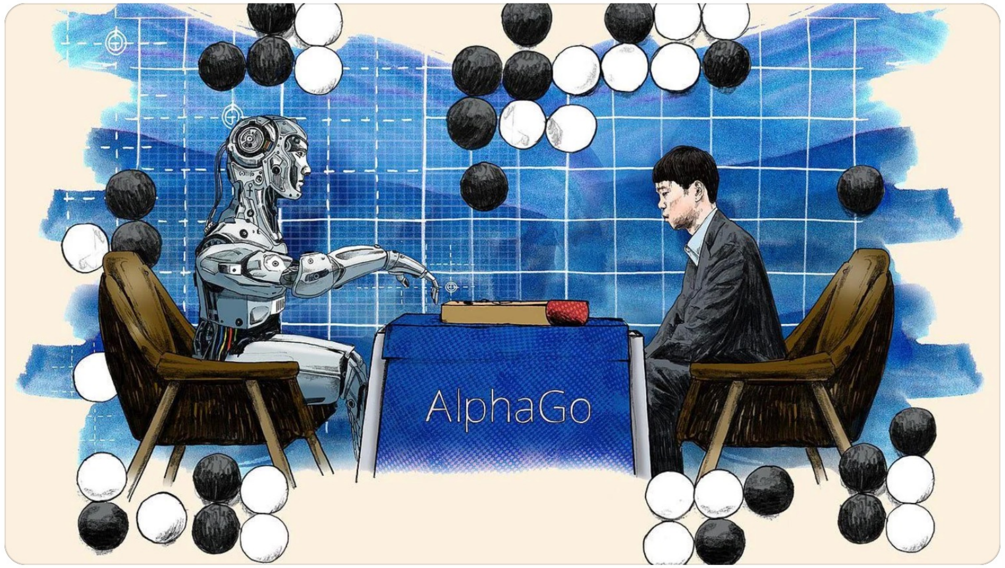
By whom? → **Us**! **we** answered and solved all that, **we actually generated its training set**

- (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

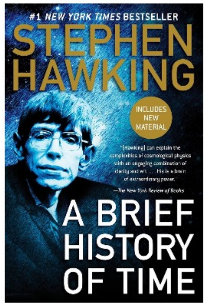
Humans and machines seem very close to be not so different..

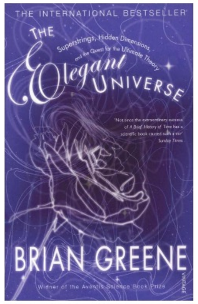


(Beyond) augmented intelligence

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 preduct 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 makes us optimistic for substantive progress in HEP theory in our lifetime?

So, in a nutshell: **is AI all we need?**

AI is not “all” we need, but “something” that we need for sure, at some level.

- Software/Computing challenges to keep up with HEP goals are tough: AI as “**part of the solution**”

Nevertheless, we had better start envisioning ML/DL not as mere “tools”

- Think of this as a **discipline** that showed up to be impressively useful, but whose theory is “under construction”

Our **best**:

- **Early career researchers**
- **Domain knowledge** and scientific rigour, ability to explore what’s solid and ignore the hype, direct efforts.

Our **risk**:

- **Hard to keep the pace**, but reluctant (time-wise, funding-wise, ..) to admit that this is not a part-time job
- R&D vs production: need more efforts to bring approaches to **production quality** (e.g. DL pipelines on AI infrastructures)

Given the AI recent past, in a 5-years timescale something radically new may probably come up.

If we fail to prepare for it, let’s **prepare to fail** in exploiting its value