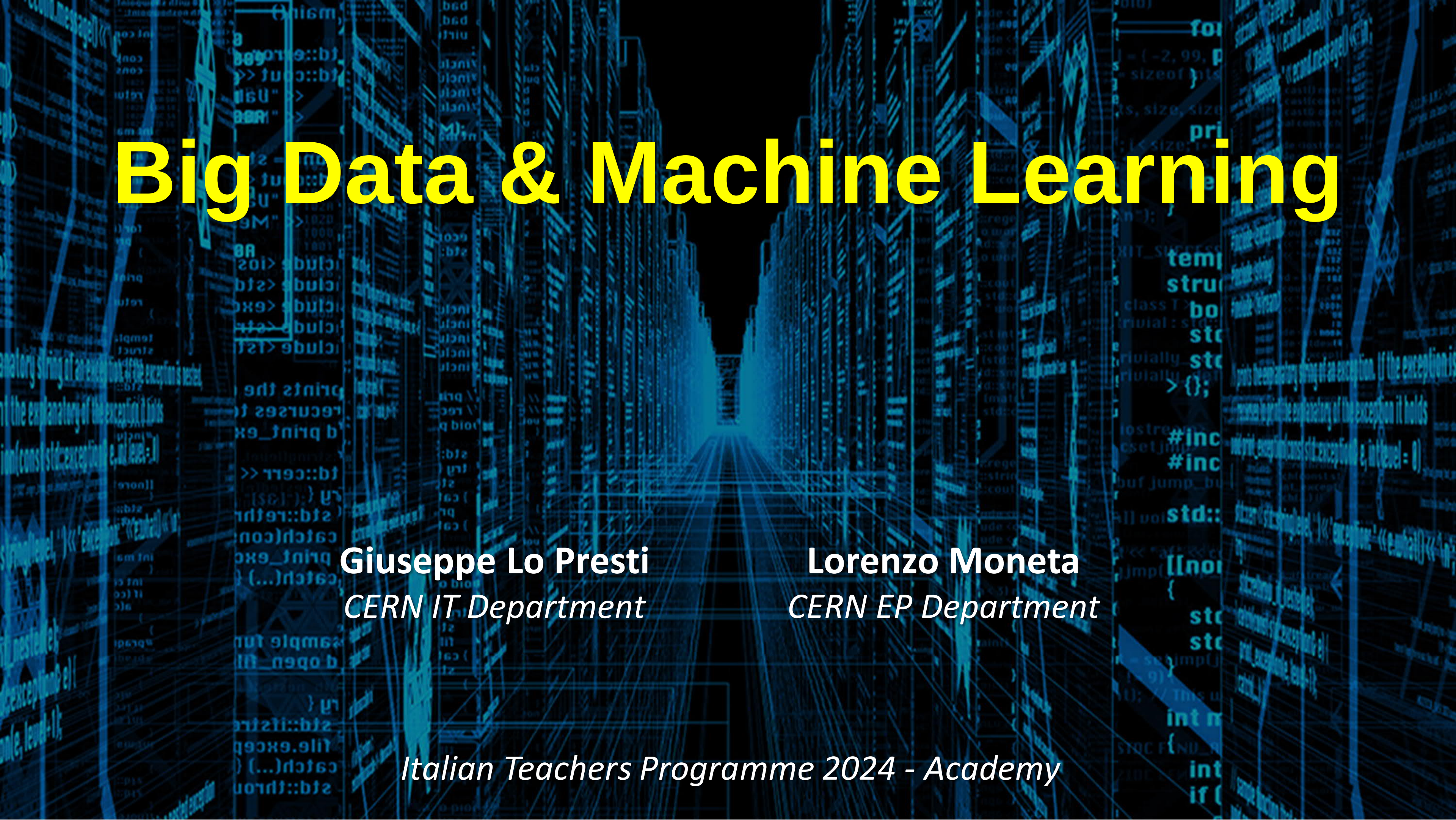


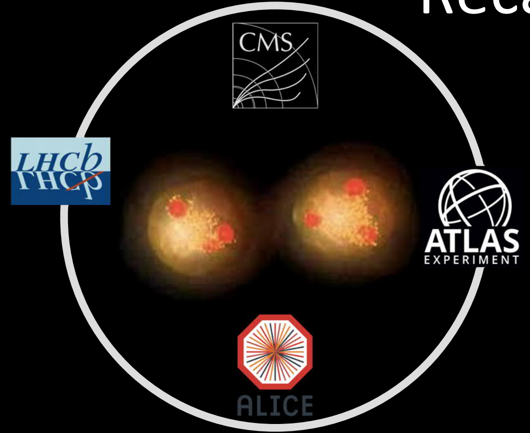
Big Data & Machine Learning

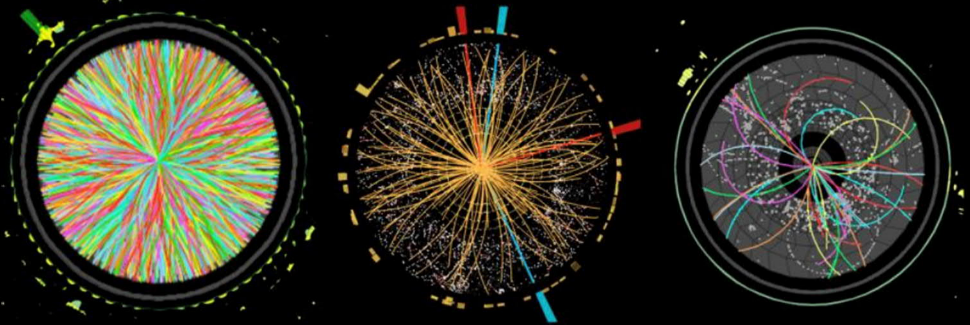
Giuseppe Lo Presti
CERN IT Department

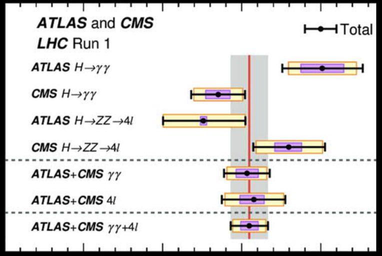
Lorenzo Moneta
CERN EP Department

Italian Teachers Programme 2024 - Academy

Recap on computing at CERN: The Big Picture



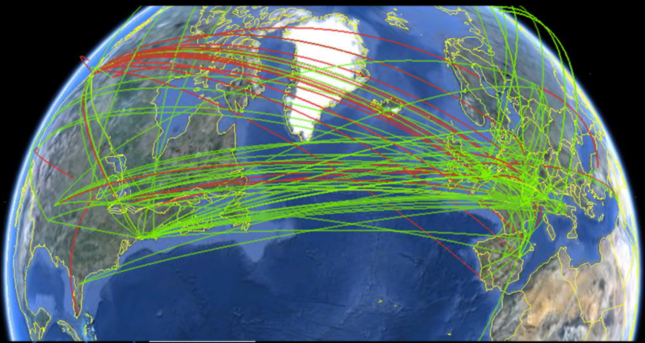




Data Storage - Data Processing - Event generation - Detector simulation - Event reconstruction - Resource accounting

Machine Learning

Distributed computing - Middleware - Workload management - Data management - Monitoring



GAUDI-LHCb

ATHENA-ATLAS

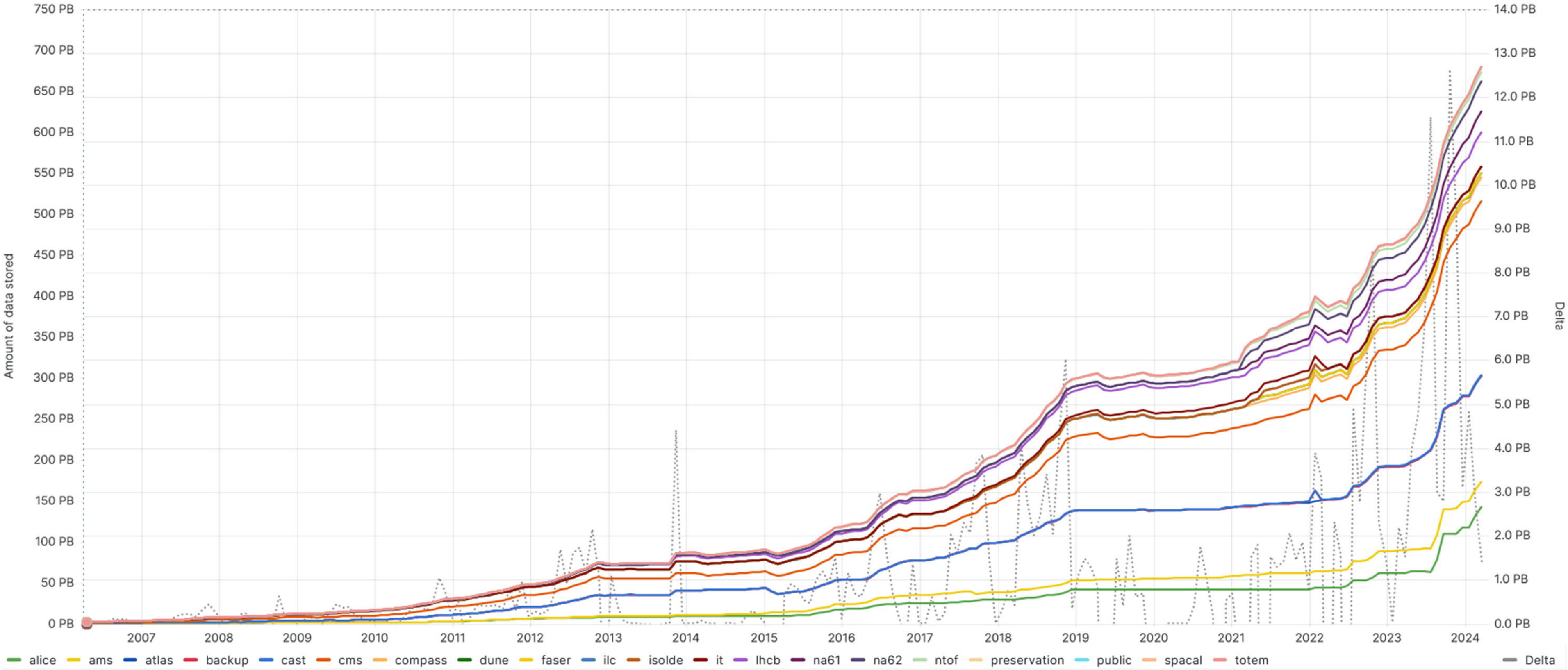
CMSSW-CMS

GEANT4 A SIMULATION TOOLKIT

SHERPA

PYTHIA

The CERN Data Archive

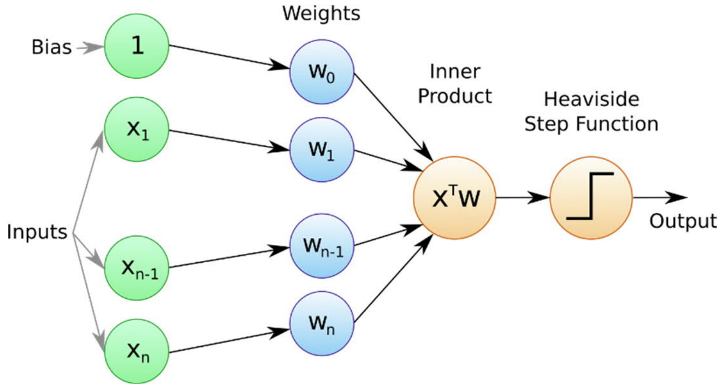


Big Data?

- *Big data* is a field that treats of ways to analyse […] or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software (*Wikipedia*)
  - **Moving target** by definition!
- From structured data, relational DBs, centralized processing…
- To ***unstructured*** data and decentralized (i.e. parallel and loosely-coupled) processing, more adapted to the Cloud
  - E.g. trend analysis, pattern recognition, **image segmentation**, natural language interpretation/translation (ChatGPT!), …

The Power of Data

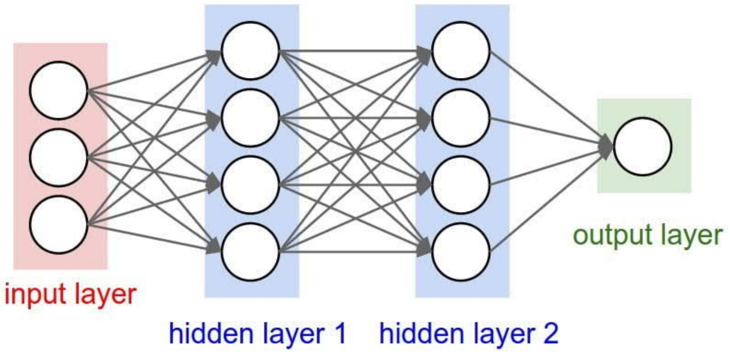
- Neural Networks are well known since the 1960s, but it's only now with **very large** and **easily accessible** data sets that they become effective!
- They are all based on a very simple "unit", the ***perceptron*** *[Rosenblatt, 1958]*
  - The weights w_i can be iteratively estimated (the ***learning*** phase) by imposing the outputs for several given inputs (*backpropagation*)
  - We may also have ***unsupervised learning***, where the learning phase is partly automated



$$y = S(w_0 + \sum_i x_i w_i)$$

Diving Deeper

- Perceptrons are connected in multiple layers



- Software frameworks are readily available to implement many configurations for ***Deep Machine Learning***

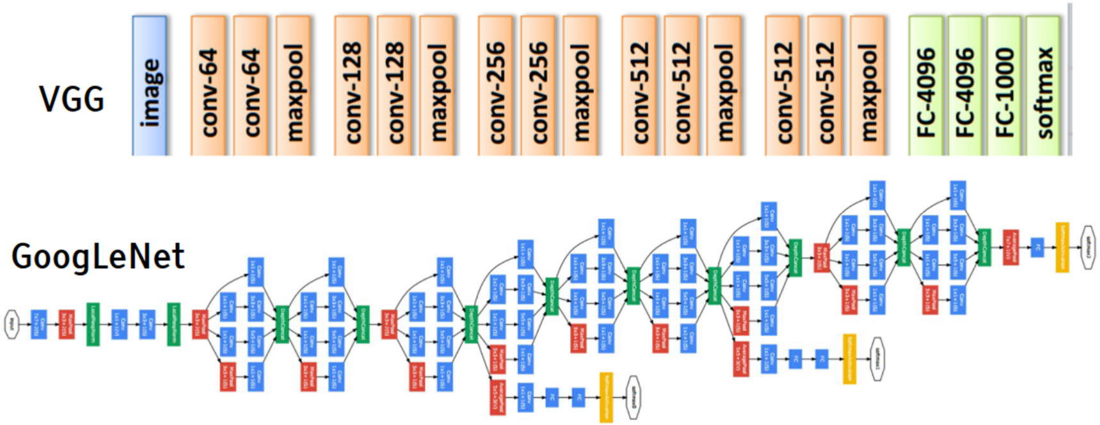
TensorFlow

PYTORCH

Deep Learning with PyTorch

How Deep?

- Example: image classification/tagging
  - Thousands of layers, **millions** of parameters
  - Facebook: a billion pictures per day goes through such networks, which delivers its result within ~2 seconds



How Deep?

- Example: natural language generation
  - Use of ***Generative Pre-trained Transformers*** to speed up the training phase
    - Transformers were proposed by Google in 2017
  - 2023: ChatGPT-4 estimated at a **trillion** parameters!
  - **Large Language Models** (LLMs) for encapsulating domain-specific knowledge
    - Being prototyped at CERN-IT to help Support and Service Desk

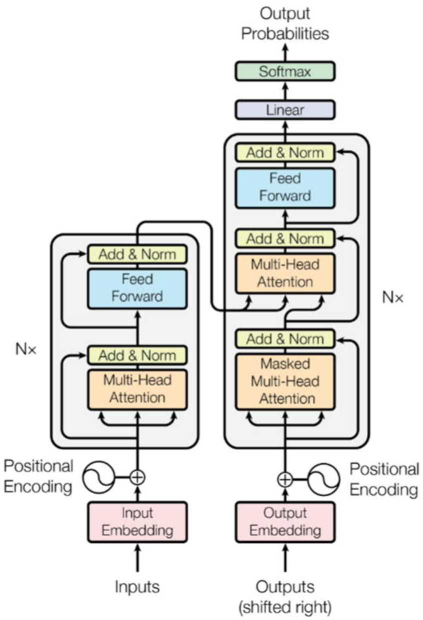


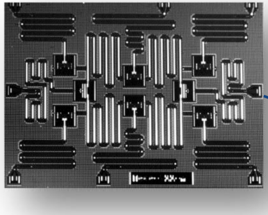
Figure 1: The Transformer - model architecture.

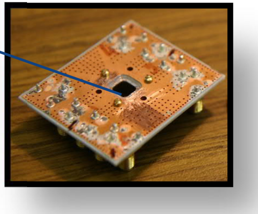
New frontiers: Heterogeneous Computing

- (Deep) Machine Learning is so **crucial** that industry has long invested into hardware acceleration
  - **GPUs** (Graphical Processing Units) for videogames (!) are being used on top of CPUs for faster matrix computations
  - **TPUs** (Tensor Processing Units), developed by Google, are offered in the Google Cloud Platform

New frontiers: Heterogeneous Computing

- A potential game changer: **Quantum Computing**
  - Quantum Computers can only execute a very limited set of “programs”, but with exponential parallelism (on paper)
  - *Quantum Machine Learning* is being demonstrated – also at CERN – as one of those programs, which can be executed by such hardware





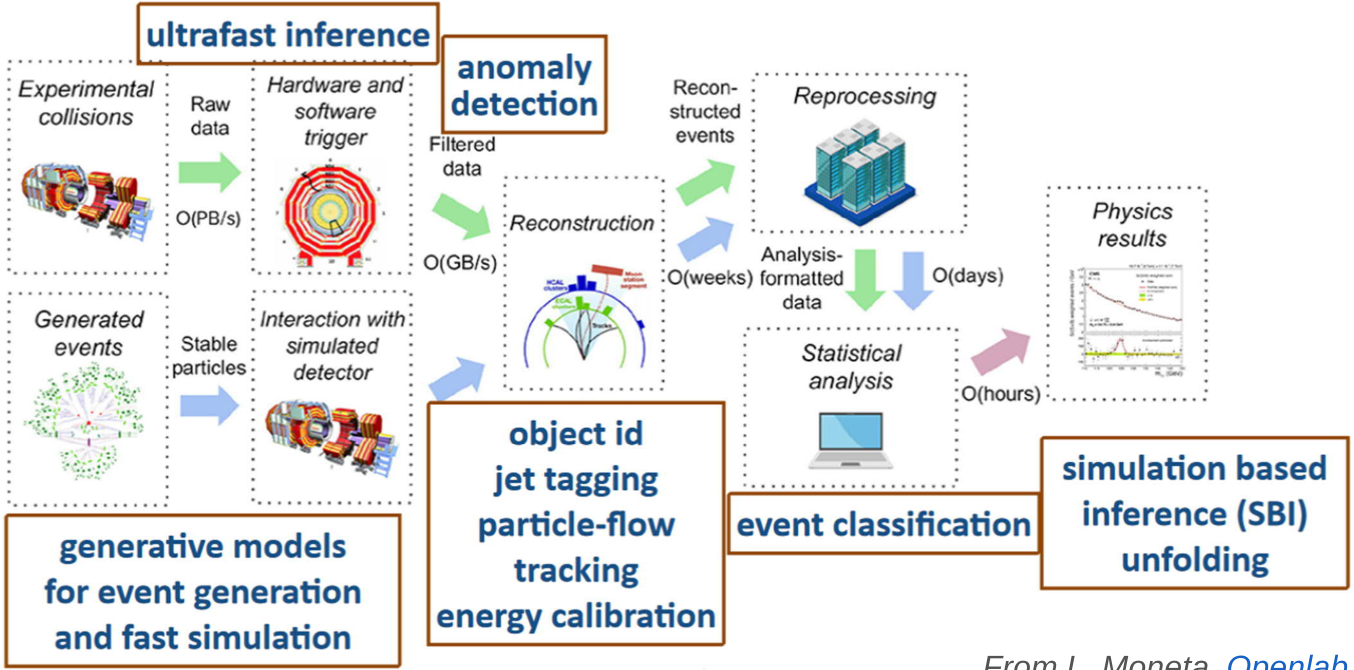
Courtesy M. Grossi

Machine Learning at CERN and beyond

- ML applied to extract trends, detect or predict failures, detect anomalies (new Physics?), …
  - Astronomy: galaxies' morphology classification
  - Gravitational Waves: real-time detection
  - Control Systems: LHC Beams Control Logging
  - Security forensics, system analysis/profiling, etc.
- In general, ML techniques implemented where analytical approaches are inapplicable/unpractical

Machine Learning for Particle Physics

Inter-experiment ML working group to coordinate such activities



ultrafast inference

anomaly detection

generative models for event generation and fast simulation

object id
jet tagging
particle-flow
tracking
energy calibration

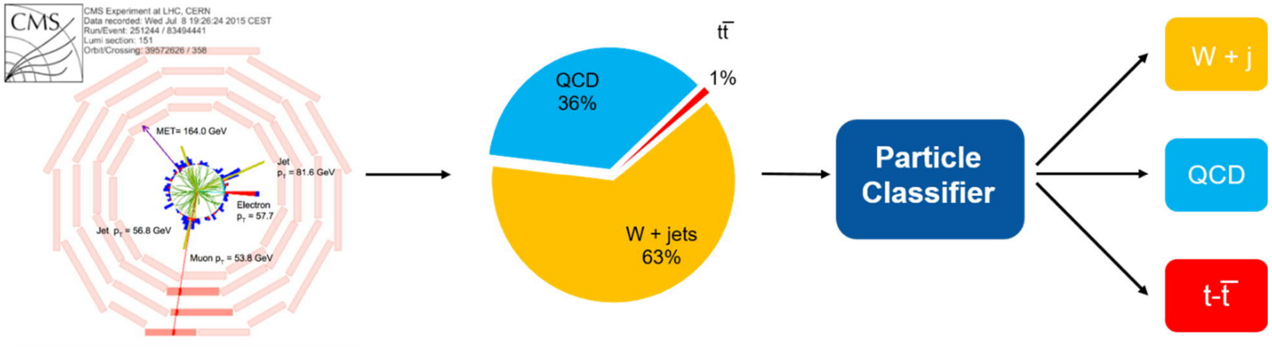
event classification

simulation based inference (SBI)
unfolding

From L. Moneta, Openlab workshop 2024

Machine Learning for Particle Physics

- Example: particles classification with Deep Learning, using TensorFlow on Spark for cluster orchestration



- References:
  - https://github.com/cerndb/SparkDLTrigger
  - https://db-blog.web.cern.ch/blog/luca-canali/2020-03-distributed-deep-learning-physics-tensorflow-and-kubernetes
  - Credits: Luca Canali, Maurizio Pierini et al.

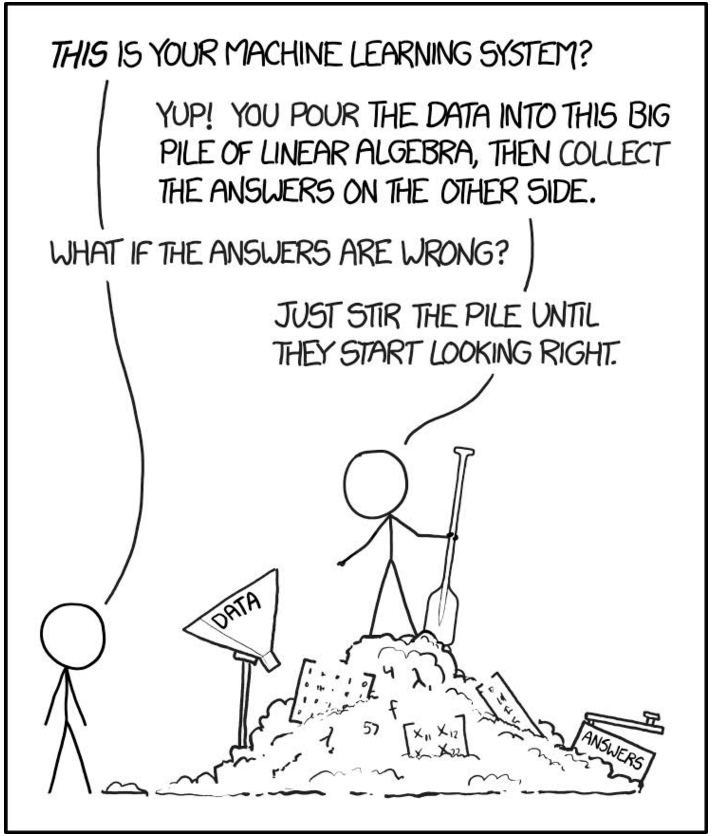
Machine Learning for Particle Physics

- A simpler case: particles classification with Deep Learning using TensorFlow
- Runs with GPUs on SWAN at CERN, or with Google Colab

- References: https://github.com/pierinim/tutorials/tree/master/HiggsSchool
  - Credits to Maurizio Pierini

Machine Learning Traps…

This was quoted at the CERN Academic Training on Machine Learning…


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838, May 2017

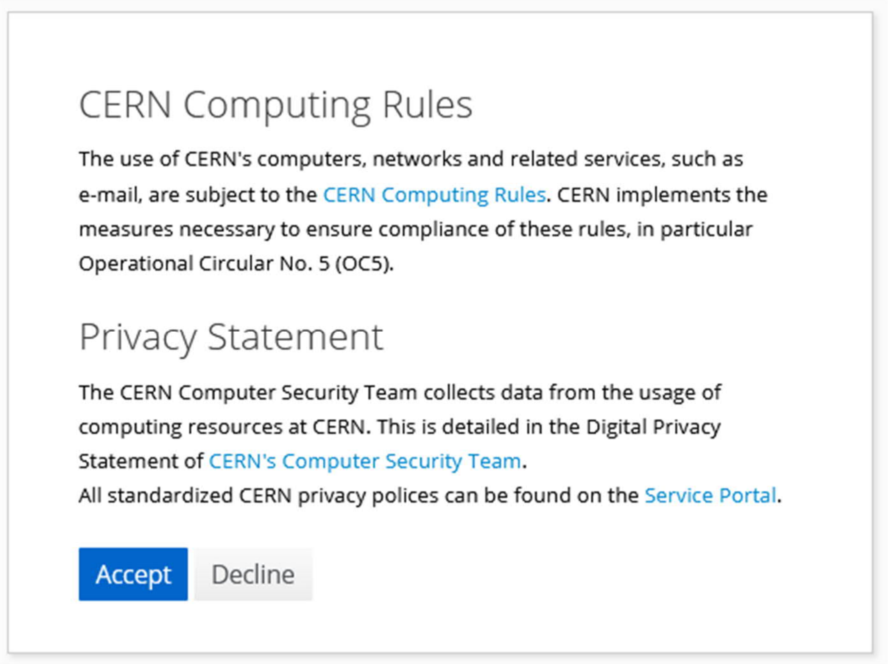
Opportunities and Risks…

- **Data Science** is a popular career path, crossing the boundaries between Computer Science, Physics and Statistics
- Fundamental science and engineering remain the pillars to understand technology!
- Big Data and Machine Learning demonstrate **data's ever-growing value**, especially when dealing with personal data
  - In **2023, 7** out of the **top 10** world-largest companies by capitalization (including the GAMAM) are entirely **based on the Data economy**
  - At **13.7 T$**, they compare with the **GDP of Germany + UK + France + Italy**!

What’s next

- You will try some ML techniques in Python, using the CERN IT infrastructure
  - In the same way as a CERN staff, you will use CERNBox and SWAN
  - Only a web browser is required
  - You will form pairs; each pair will get a CERN account, login = `itpswan1, 2, 3...`
    - More details (including password) in a moment, with Lorenzo
- The Physics goal is to work with CMS data
- The “Educational” goal is to get dirty with a hands-on, real machine learning activity!

The small print



CERN Computing Rules

The use of CERN's computers, networks and related services, such as e-mail, are subject to the CERN Computing Rules. CERN implements the measures necessary to ensure compliance of these rules, in particular Operational Circular No. 5 (OC5).

Privacy Statement

The CERN Computer Security Team collects data from the usage of computing resources at CERN. This is detailed in the Digital Privacy Statement of CERN's Computer Security Team.
All standardized CERN privacy polices can be found on the Service Portal.

https://home.cern/news/news/computing/computer-security-rules-whats-allowed-and-what-isnt

Thanks for your attention! Questions so far?

Giuseppe.LoPresti@cern.ch
www.linkedin.com/in/giuseppelopresti