



Artificial Intelligence at CERN and how the IT Department supports it

Alberto Di Meglio
Head of Innovation
IT Department

A few words about Artificial Intelligence and what it means

So, What is AI?

“The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.” **(Oxford Dictionary)**

“Artificial intelligence (AI), in its broadest sense, is intelligence exhibited by machines, particularly computer systems. It is a field of research in computer science that develops and studies methods and software that enable machines to perceive their environment and use learning and intelligence to take actions that maximize their chances of achieving defined goals” **(Russel and Norvig, Artificial Intelligence: A Modern Approach, 2021)**

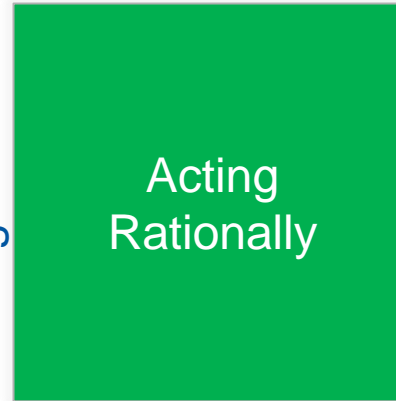
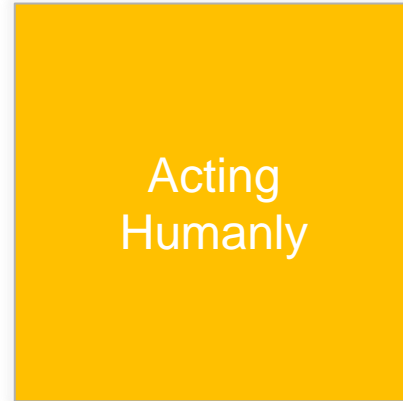
“The term refers indistinctly to systems that are pure science fiction (so-called "strong" AIs with a self-aware form) and systems that are already operational and capable of performing very complex tasks (face or voice recognition, vehicle driving - these systems are described as "weak" or "moderate" AIs).” **(The Council of Europe)**

“The term ‘artificial intelligence’ means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments.” **(The National AI Initiative Act of 2020, USA)**

“AI is the ability of a machine to display human-like capabilities such as reasoning, learning, planning and creativity.” **(The EU AI Act, 2023, European Union)**

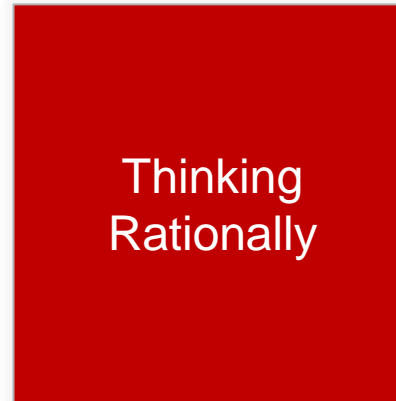
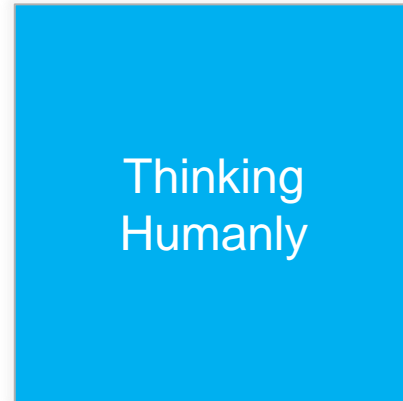
The Four Goals of AI

Replication of intelligent human behaviour (Alan Turing's approach): by behaving indistinguishably from a human being, the computer has exhibited intelligence. This goal represents most of the capabilities which AI has been focusing on since its conception (knowledge, reasoning, language understanding, and learning)



Acting so as to achieve what one believes to be the **best outcome**. Russell and Norvig themselves favour this approach to building so-called **rational agents**, pointing out that it in fact includes many of the other approaches above. Consider that acting rationally is a matter of doing what is 'right', given the situation you are in. This is a source of concern because a system acting rationally might not act humanly and implies the concepts of ethics and morality.

The idea that modelling human thought processes could enable us to somehow replicate such processes in computer systems. This has been one of the goals of **Cognitive Science**, an interdisciplinary pursuit made up of Psychology, Computer Science, Philosophy, Linguistics and Anthropology.



Attempts to formulate so-called "**laws of thought**", often expressed using special systems of symbols deriving from mathematical logic, and thereby build computer systems which are able to reason similarly to humans (assuming logic adequately models human thought). Since humans do not necessarily reason according to specific rational laws, this approach is **not a good match for actual thinking**.

Thinking Vs. Acting

Human Vs. Rational

Weak and Strong AI

Narrow or weak AI

focuses on specific tasks, operating under stringent constraints in order to perfect that task and perform it even better than humans. Its limited functionality allows it to automate a specific task with ease, and its narrow focus has allowed it to power many technological breakthroughs in just the last few years

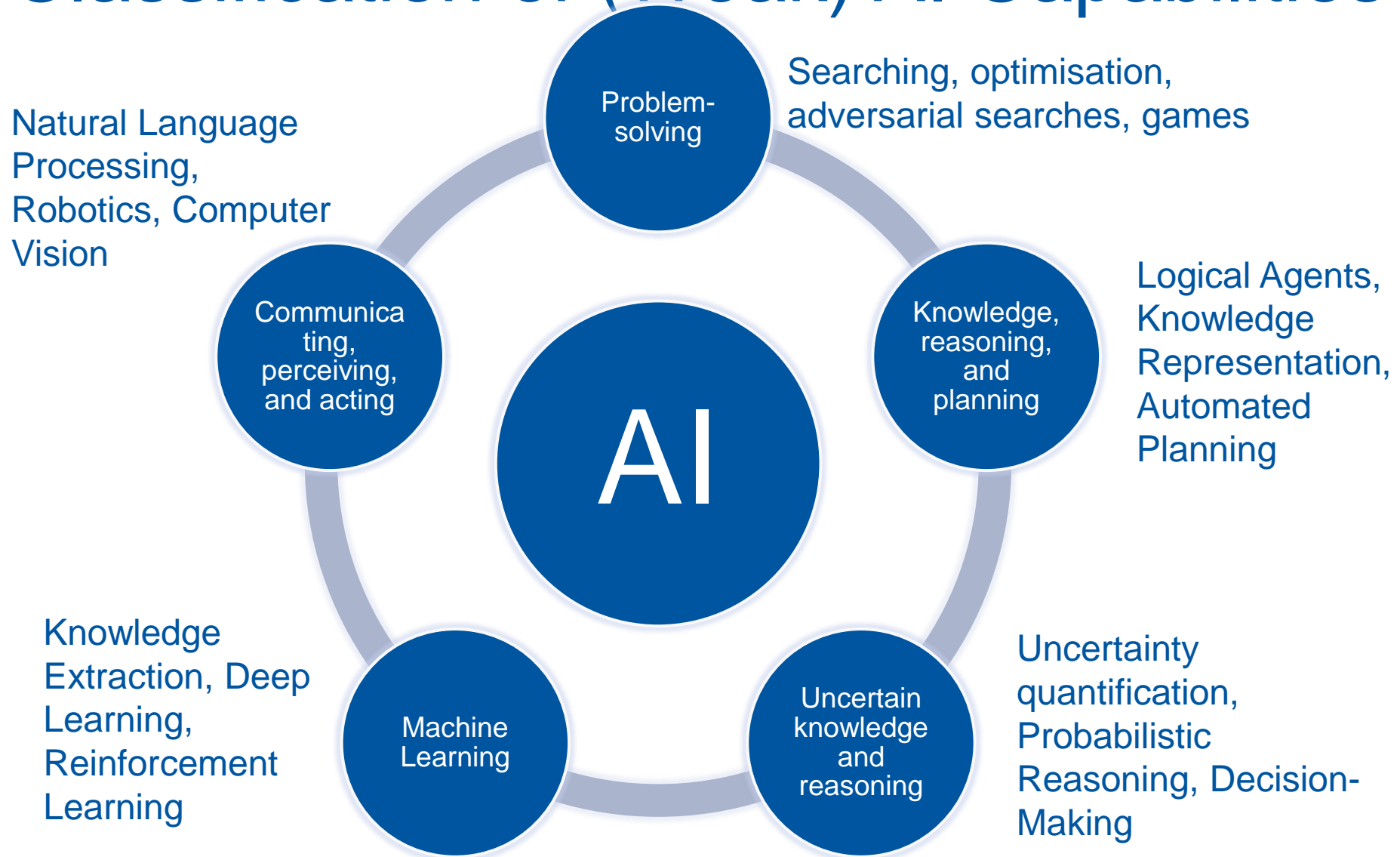


General or strong AI

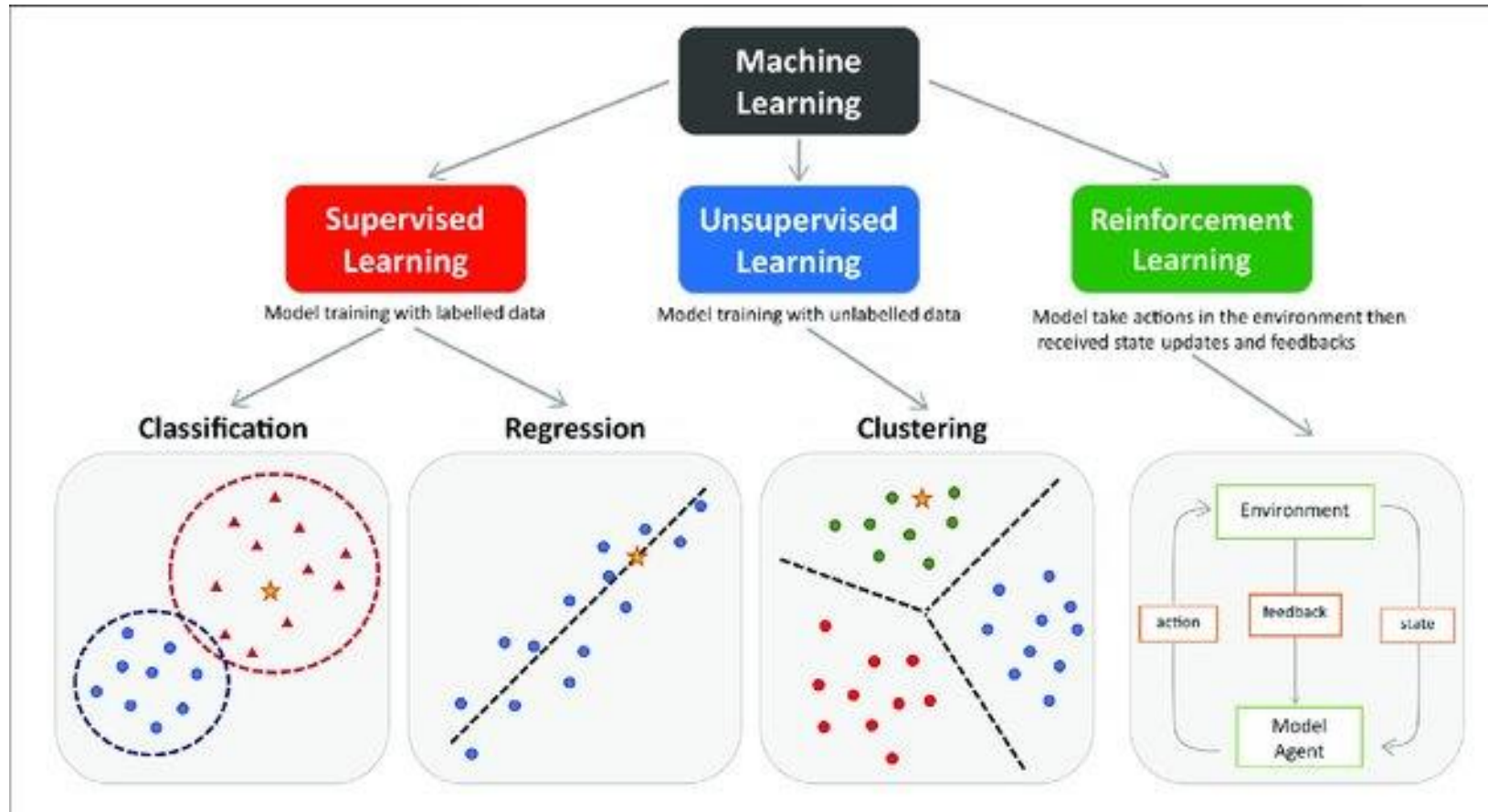
is capable of thinking and performing actions in the same ways human beings can and is able to solve problems, plan, and learn new skills in ways similar to our own. “The more an AI system approaches the abilities of a human being, with all the intelligence, emotion, and broad applicability of knowledge, the more ‘strong’ the AI system is considered”

Definitions by Kathleen Walch, managing partner at [Cognilytica](https://cognilytica.com)'s Cognitive Project Management for AI certification and co-host of popular podcast called [AI Today](https://builtin.com/artificial-intelligence/strong-ai-weak-ai), <https://builtin.com/artificial-intelligence/strong-ai-weak-ai>

Short Classification of (Weak) AI Capabilities

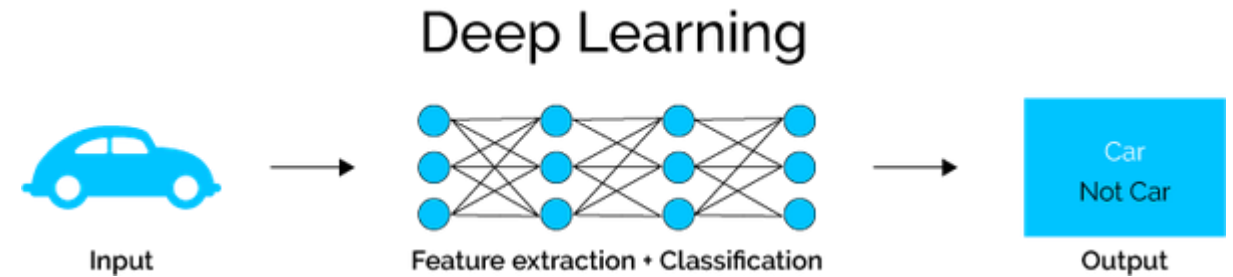
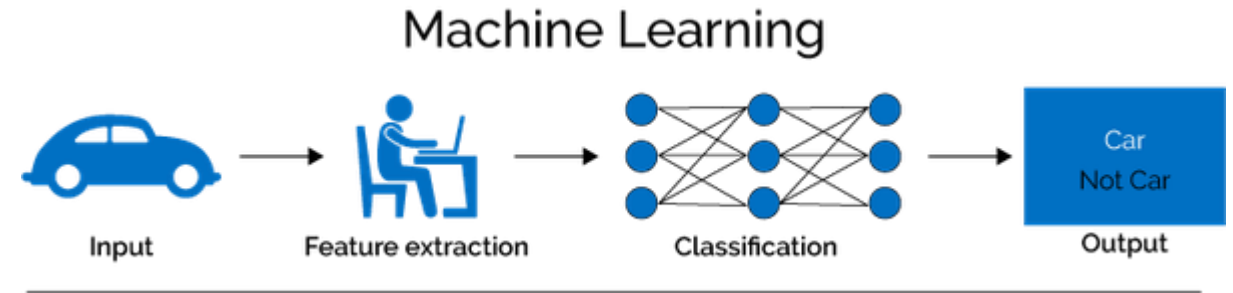
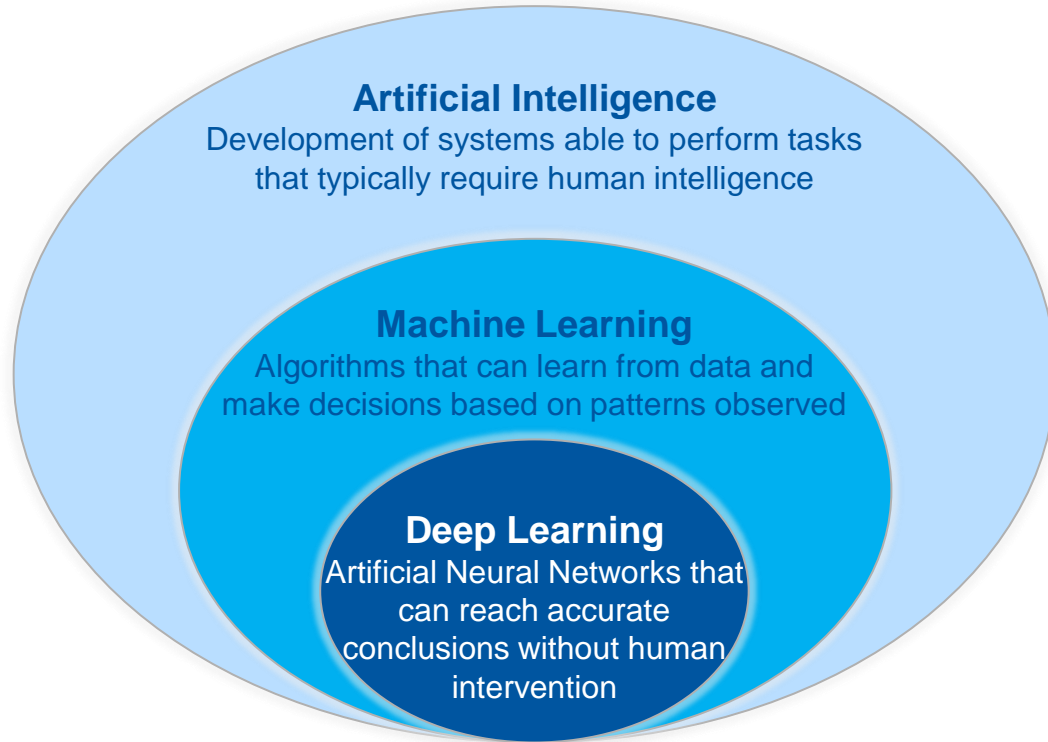


Machine Learning



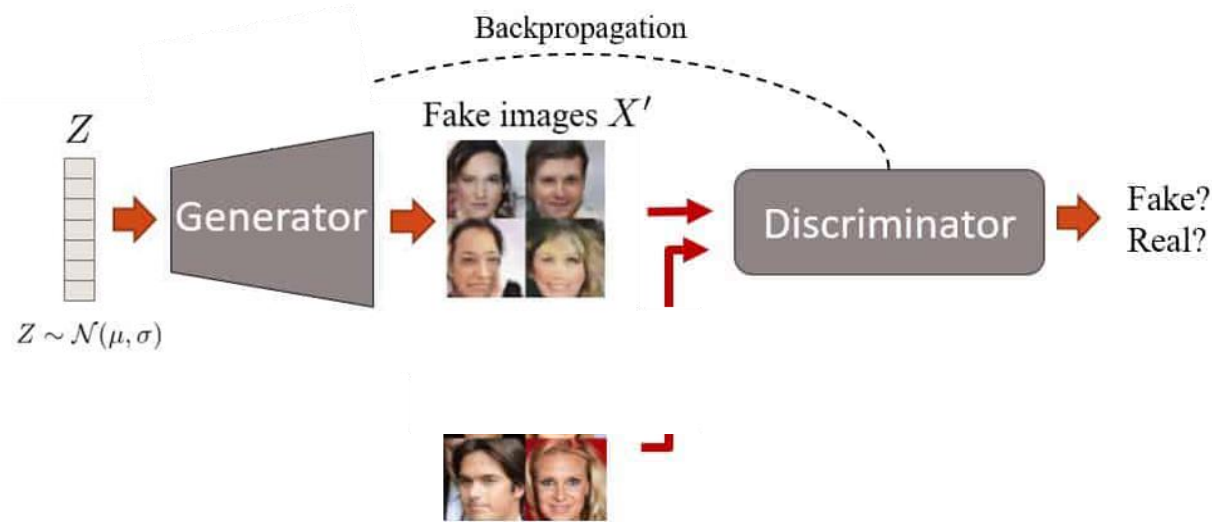
<http://dx.doi.org/10.3389/fphar.2021.720694>

Deep Learning



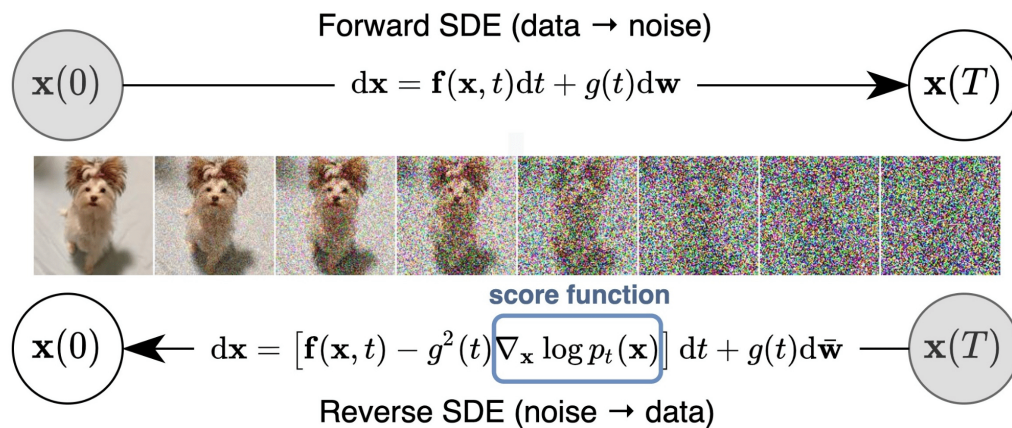
Source: <https://blog.dataiku.com/hs-fs/>

Generative/Adversarial/Diffusion Models



Generative Models allow to create artificial data sets with the same probability distribution as a training data set (for example images)

To improve accuracy a combination of **Generative** and **Adversarial** models can be used, where the adversarial network estimates "how close" the "fake" is from the "real" data



A recent approach called **Diffusion Models** allows to generate extremely realistic images from noise (with or without guiding conditions). The forward process (Markov chains) adds noise to a given image to produce Gaussian noise, the reverse process inverts the steps to extract an image from noise (NN)

Attention and Transformers

The **Attention** mechanism in Machine Learning (2014) mimics the process of human cognitive attention.

It has been proposed as a solution to situations (especially time-dependent or sequential problems (like NPL) where fixed-weight NN have a bias towards later input compared to earlier input.

It is based on the idea that the weights in a NN are “soft”, that is they can be changed and adapted during the continuous training process, in the same way as human beings can “focus” on specific parts of a sentence and “adapt” the relative importance of previous words as they keep reading.

If the weights are processed sequentially we have **Recursive Neural Networks (RNN)**, if they are processed in parallel we have **Transformers**.

Transformers were proposed in 2017 (<https://arxiv.org/abs/1706.03762>) and are becoming one of the most interesting mechanism for training Large Language Models because they require less training time. They are at the base of frameworks like BERT and GPT

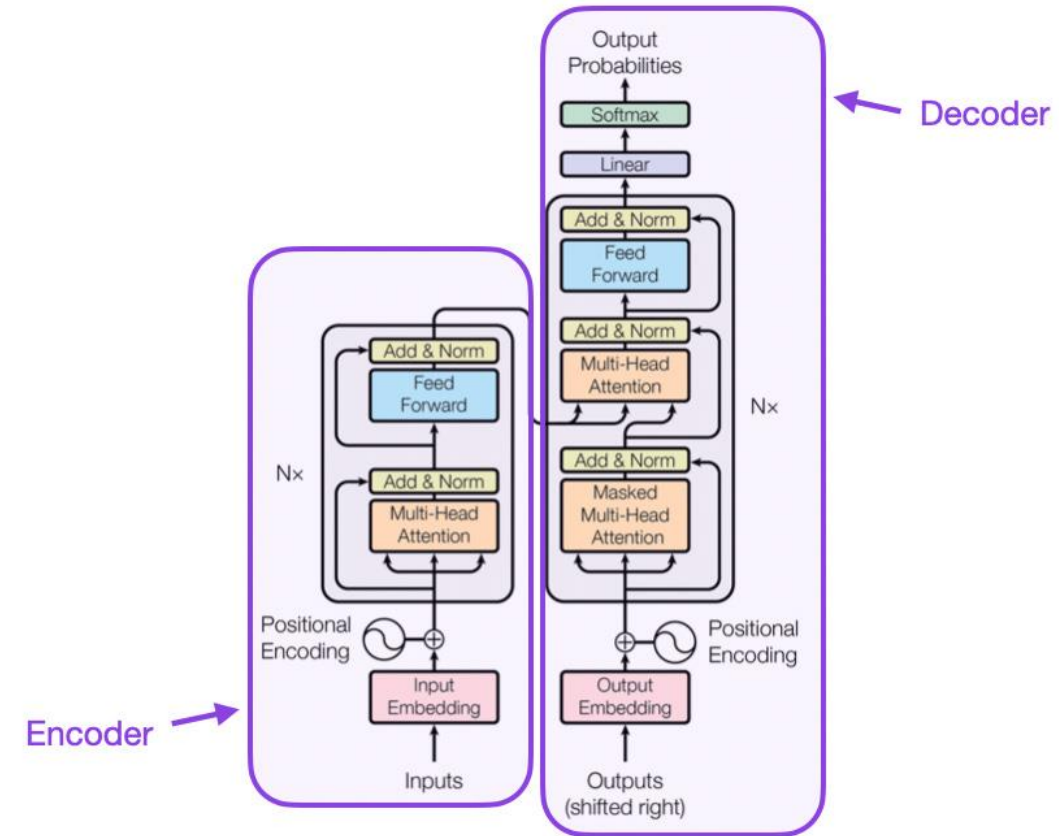


Figure 1: The Transformer - model architecture.

Source: <https://arxiv.org/abs/1706.03762>

Foundation Models

Foundation Models are the closest we have ever been to “General Purpose” AI.

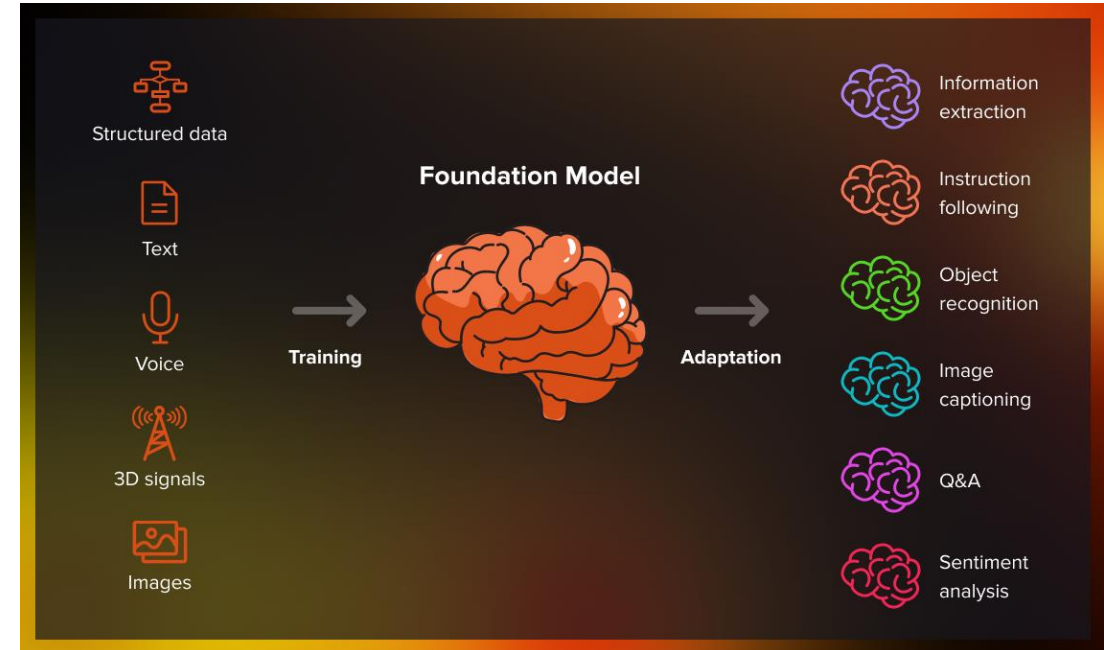
FMs were introduced in 2020 by researchers at Stanford University as “any model that is trained on **broad data** (generally using self-supervision at scale) that can be **adapted** (e.g., fine-tuned) to a **wide range of downstream tasks**”¹.

They are similar to LLMs but are not developed for specific tasks, but rather to consume many different types of data (multi-modality) and producing adaptable output for different tasks.

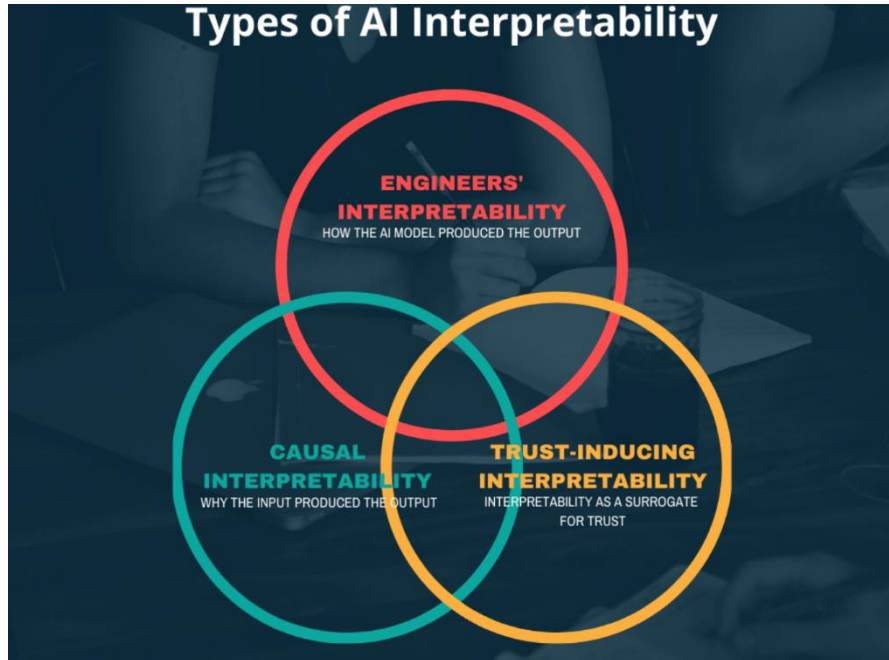
Key characteristics of foundation models are “**emergence**” and “**homogenization**”¹, that is the ability to discover models from data and to be used in different domains.

GPT3.5 and GPT4 are considered to be Foundation Models already as they are able to use different types of input (not just text) and be used as building blocks for more specific applications

¹<https://arxiv.org/abs/2108.07258>



Interpretability and explainability



Source: [Stanford HAI](#)

Because of the way Deep Learning neural networks work, the values of the internal hidden layers are not visible and it is difficult to ascertain the specific impact of an input value (**cause**) on the output (**effect**) or the exact path that produced the output from the input (**black box effect**).

Considerable research is devoted to the so-called “**interpretability**” and “**explainability**” of deep learning models. In many fields, the fact of not being able to consistently reproduce or explain how the output was generated or the possible confusion between **causation** and **correlation** is a major problem (*the priest effect*). A great exposition about this is Judea Pearl’s “The Book of Why”.

Typical examples:

If a decision-support system suggest a treatment for a certain medical condition that conflicts with the human doctor’s opinion, what decision should the doctor take?

If an autonomous vehicle has an accident as consequence of a decision of the software, who is responsible?

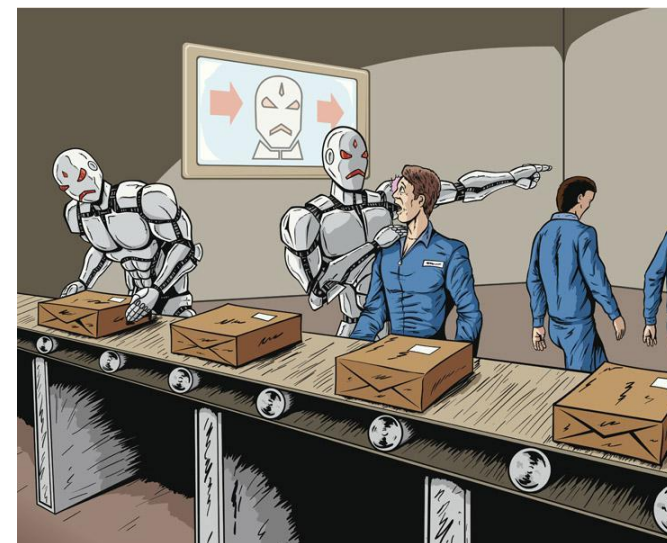
Will AI disrupt the job market?

There is intense debate around the fact that AI might/will make lots of people redundant. This is a statement made every time a new disruptive technology comes around, suggesting that history must repeat itself. It has been said about steam engines and horses, and it didn't go well for the horses...

A common complementary statement is made more and more often “AI Won't Replace Humans — But Humans With AI Will Replace Humans Without AI”. There are many variations of this statement, but all to the same effect.

Removing the hype, certainly technologies with the type of impact AI are showing require an adaptation of the **education and skills development systems** and especially a careful look at possible “digital divide” effects.

Human ingenuity and expertise will not be replaced by AI anytime soon and are actually an important part of end-to-end AI pipelines



AI at CERN

A (really) non-exhaustive list of examples

AI in Physics Research

Great interest to explore the opportunities provided by new AI algorithms/models to physics research

The screenshot shows the top part of the IML website. At the top left is the CERN logo and the text 'Accelerating science'. To the right are links for 'Sign in' and 'Directory'. Below this is a large banner image with the text 'IML' and 'Inter-Experimental LHC Machine Learning Working Group'. Underneath the banner is a blue bar with the text '6th Inter-experiment Machine Learning Workshop'. Below that is a date range 'January 29, 2024 to February 2, 2024' and 'CERN Europe/Zurich timezone'. There is also a search bar with the placeholder text 'Enter your search term'.

- Overview
- Scientific Program
- Call for Abstracts
- Timetable
- Contribution List
- Poster sessions
- Book of Abstracts
- Registration
- Participant List
- Videoconference
- Code of conduct

This is the sixth annual workshop of the LPCC inter-experimental machine learning working group.

The workshop will be held on 29Jan-2Feb 2024 at CERN in a hybrid format, with remote participation made possible.

Confirmed invited speakers

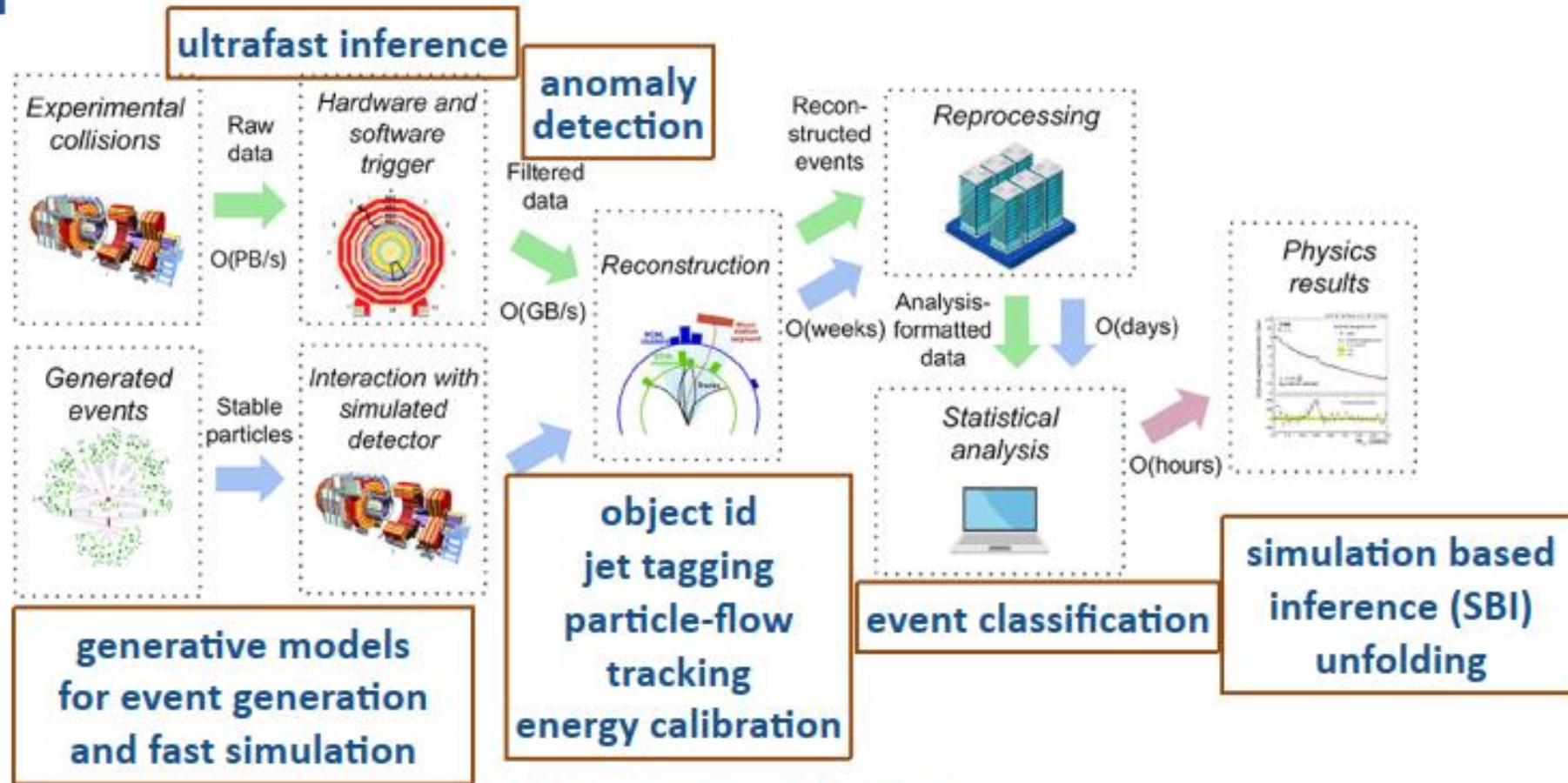
- Jürgen Schmidhuber (IPSI and KAUST), opening talk
- François Charton (META AI) on symbolic learning
- Casey Fitzpatrick (Contextual AI), on LLMs
- Eliska Greplova (TU Delft), on quantum optics
- Gregor Kasieczka (UHamburg), on particle physics
- Jonas Kohler (Microsoft Research), on molecular physics
- Gaël Varoquaux (INRIA), on scientific inference
- Francesco Maria Velotti (CERN), on accelerators physics
- Gail Weiss (EPFL), full tutorial on transformers

The screenshot shows the agenda for the 'IML Machine Learning Working Group: LLMs for HEP' videoconference. The header includes the title, date 'Tuesday Apr 9, 2024, 3:00 PM → 6:00 PM', location 'Europe/Zurich', and address '500/1-001 - Main Auditorium (CERN)'. Below the header is a 'Description' section with the topic 'LLMs for HEP' and a 'Videoconference' section with the name 'IML Machine Learning Working Group' and a 'Join' button. The agenda items are as follows:

- 3:00 PM → 3:05 PM News** (5m)
Speakers: Anja Butter (Centre National de la Recherche Scientifique (FR)), Daniel Whiteson (University of California Irvine (US)), Fabio Catalano (CERN), Julian Garcia Pardinas (CERN), Lorenzo Moneta (CERN), Dr Pietro Vischia (Universidad de Oviedo and Instituto de Ciencias y Tecnologías Espaciales de Asturias (ICTEA)), Stefano Carrazza (CERN)
Attachment: News_090424.pdf
- 3:05 PM → 3:20 PM AccGPT – A CERN Chatbot for Internal Knowledge Retrieval** (15m)
Speaker: Dr Florian Rehm (CERN)
Attachment: AccGPT-IML_v2.pdf
- 3:20 PM → 3:25 PM Discussion** (5m)
- 3:25 PM → 3:40 PM Learning the language of QCD Jets with transformers** (15m)
Speakers: Dr Alexander Mück (RWTH Aachen University), Michael Kramer (Rheinisch Westfälische Tech. Hoch. (DE))
Attachment: mueck.pdf
- 3:40 PM → 3:45 PM Discussion** (5m)
- 3:45 PM → 4:00 PM OmniJet- α : The first cross-task foundation model for particle physics** (15m)
Foundation models are multi-dataset and multi-task machine learning methods that once pre-trained can be fine-tuned for a large variety of downstream applications. The successful development of such general-purpose models for physics data would be a major breakthrough as they could improve the achievable physics performance while at the same time drastically reduce the required amount of training time and data. We report significant progress on this challenge on several fronts. First, a comprehensive set of evaluation methods is introduced to judge the quality of an encoding from physics data into a representation suitable for the autoregressive generation of particle jets with transformer architecture (the common backbone of foundation models). These measures motivate the choice of a higher-fidelity tokenization compared to



AI in Experiment Data Analysis



AI is everywhere !

pic from [fdata.2021.661501](https://indico.cern.ch/event/1356148/contributions/5815418/attachments/2827180/4939205/fdata.2021.661501)

L. Moneta / CERN EP-SFT

Lorenzo Moneta – CERN openlab Technical Workshop, March 26th, 2024

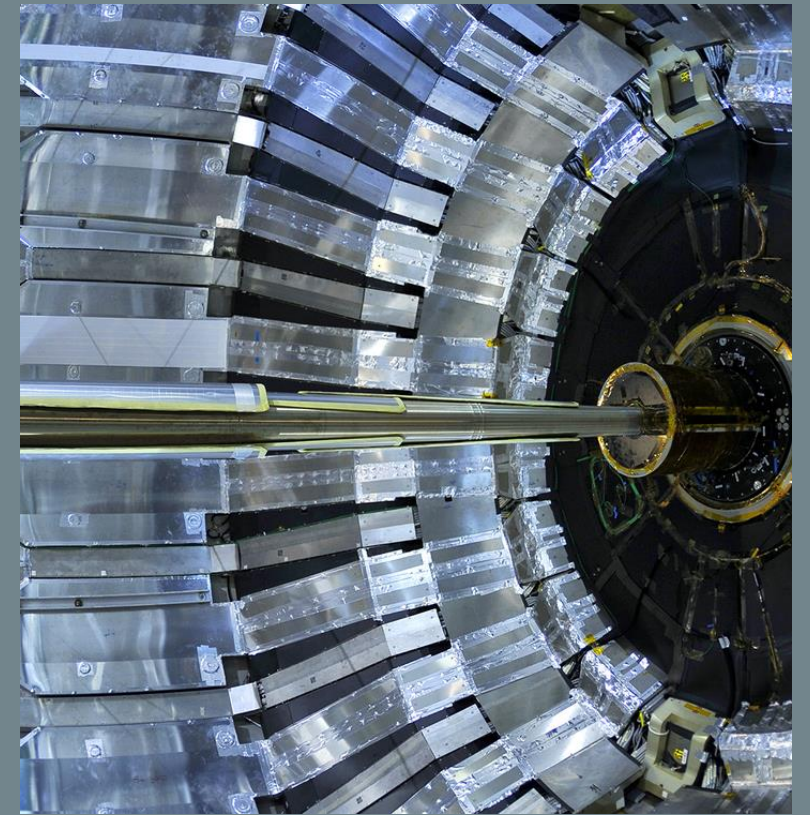
https://indico.cern.ch/event/1356148/contributions/5815418/attachments/2827180/4939205/AI_CERN_Overview_OpenlabWS.pdf

WHAT IS THE NEXTGEN TRIGGERS Project?



The Next Generation Triggers project started in January 2024 as a collaboration between CERN (the Experimental Physics, Theoretical Physics and Information Technology Departments) and the ATLAS and CMS experiments.

The key objective of the five-year NextGen project is to get more physics information out of the HL-LHC data to uncover as-yet-unseen phenomena by more efficiently selecting interesting physics events while rejecting background noise



NextGen explores the use of Artificial Intelligence, quantum-inspired algorithms, and high-performance computing to improve theoretical modelling and optimise methods and tools in the search for ultra-rare events.



AI in Accelerators Operations

RL @ CERN - a selection



PS

- Correct RF phase & voltage for uniform bunch splitting (LHC beams)
- Successful sim2real & fully operational
- Multi-agent (SAC) & CNN for initial guess
- Next: continuous controller (UCAP)

A. Lasheen, J. Wulff

PS to SPS

- Adjust fine delays of SPS injection kicker
- RL agent (PPO) trained on data-driven dynamics model
- Ready for sim2real test

M. Remta, F. Velotti

LINAC3 / LEIR

- PhD project (B. Rodriguez): control LINAC3 cavities for optimal injection efficiency into LEIR
- RL state based on VAE-encoded Schottky spectra
- Agent trained on data-driven dynamics model

V. Kain, N. Madysa

SPS

- Steer DC beams in TT20 TL using split-foil secondary emission monitors
- Works well in simulations, with noise and varying emittances
- Ready for sim2real test

N. Bruchon, V. Kain

Courtesy M. Schenk

Data Science Seminar - Liverpool University, V. Kain, 11-June-2024

45:42

14th International Particle Accelerator Conference, Venice, Italy
 ISBN: 978-3-95450-231-8 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2023-THPL038

ULTRA FAST REINFORCEMENT LEARNING DEMONSTRATED AT CERN AWAKE

S. Hirlander*, L. Lamminger, Paris Lodron University Salzburg, Austria
 Z. Della Porta, V. Kain¹, CERN, Geneva, Switzerland

Abstract

Reinforcement learning (RL) is a promising direction in machine learning for the control and optimisation of particle accelerators since it learns directly from experience without needing a model a-priori. However, RL generally suffers from low sample efficiency, and thus training from scratch on the machine is often not an option. RL agents are usually trained or pre-tuned on simulators and then transferred to the real environment. In this work, we propose a model-based RL approach based on Gaussian processes (GPs) to overcome the sample efficiency limitation. Our RL agent was able to learn to control the trajectory at the CERN AWAKE (Advanced Wakefield Experiment) facility, a problem of 10 degrees of freedom, within a few interactions only. To date, numerical optimisers are used to restore or increase and stabilise the performance of accelerators. A major drawback is that they must explore the optimisation space each time they are applied. Our RL approach learns as quickly as numerical optimisers for one optimisation run, but can be used afterwards as single-shot or few-shot controllers. Furthermore,

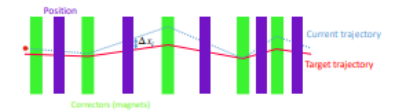


Figure 1: Illustration of a beam steering problem as in the AWAKE electron line.

line of the AWAKE experiment at CERN, as described in the following section.

PROBLEM DEFINITION - AWAKE ELECTRON LINE TRAJECTORY STEERING

The electron line of AWAKE (see Fig. 1) served in the past as an excellent environment to test optimisation and control algorithms, as also an accurate simulation of the

This work must maintain attribution to the author(s), title of the work, publisher, and DOI.

Reinforcement Learning applications

AI in Control Systems, Robotics, Infrastructures

19th Int. Conf. Accel. Large Exp. Phys. Control Syst. ICALEPCS2023, Cape Town, South Africa JACoW Publishing
 ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-H03A007

CONTROL DESIGN OPTIMIZATIONS OF ROBOTS FOR THE MAINTENANCE AND INSPECTION OF PARTICLE ACCELERATORS

A. Díaz Rosales^{*1,2}, H. Gamper^{*1,3}, M. Di Castro¹

¹ European Organization for Nuclear Research (CERN), 1211 Meyrin, Switzerland
² Department of Cognitive Robotics, Delft University of Technology, 2628 Delft, The Netherlands
³ Johannes Kepler University, Linz, 4040 Linz, Austria

Abstract

Automated maintenance and inspection systems have become increasingly important over the last decade for the availability of the accelerators at CERN. This is mainly due to improvements in robotic perception, control, and cognition and especially because of the rapid advancement in artificial intelligence. The robotic service at CERN performed the first interventions in 2014 with robotic solutions from external companies. However, it soon became clear that a customized platform needed to be developed in order to satisfy the needs and in order to efficiently navigate through the cluttered, semi-structured environment. This led to the formation of a robotic fleet of about 20 different robotic systems that are currently active at CERN. In order to increase the efficiency and robustness of robotic platforms for future accelerators it is necessary to consider

and boosting machines availability [3]. The advancements in robotic perception, control, and cognition, particularly in artificial intelligence, have contributed to this development. The CERN robotic service initially used external company solutions for interventions but later had to create customized platforms to meet their specific requirements and navigate the cluttered and semi-structured environment efficiently. This led to a robotic fleet of about 20 different robotic systems [4,5]. In order to increase the efficiency and robustness of robotic platforms for future accelerators it is necessary to consider robotic interventions in the early design phase of new machines. Task-specific solutions tailored to particular needs can then be designed, which in general show higher efficiency than universal robotic systems [6, p. 284].

This approach is currently applied to the design of the new robotic manipulators at CERN. This paper presents the latest

work must maintain attribution to the author(s), title of the work, publisher, and DOI

19th Int. Conf. Accel. Large Exp. Phys. Control Syst. ICALEPCS2023, Cape Town, South Africa JACoW Publishing
 ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-TUPDP102

LEVERAGING LOCAL INTELLIGENCE TO CERN INDUSTRIAL CONTROL SYSTEMS THROUGH EDGE TECHNOLOGIES

A. Patil, F. Varela, F. Ghawash, B. Schofield, CERN, Geneva, Switzerland,
 T. Kaufmann, A. Sundermann, D. Schall, Siemens AT - T DAI DAS, Austria,
 C. Kern, Siemens DE - T CED SES, Germany

Abstract

Industrial processes often use advanced control algorithms such as Model Predictive Control (MPC) and Machine Learning (ML) to improve performance and efficiency. However, deploying these algorithms can be challenging, particularly when they require significant computational resources and involve complex communication protocols between different control system components. To address these challenges, we showcase an approach leveraging industrial edge technologies to deploy such algorithms. An edge device is a compact and powerful computing device placed at the network's edge, close to the process control. It executes the algorithms without extensive communication with other control system components, thus reducing latency and load on the central control system. We also employ an analytics function platform to manage the life cycle of the algorithms.

However, new control hardware, such as multi-processor PLCs and AI expansion cards, have emerged, making this deployment possible.

Another challenge is the notable disparities between the focus areas of control engineers and data scientists when devising control systems. Control engineers primarily concentrate on industrial communication protocols, control devices, PLC programming, and SCADA development. In contrast, data scientists and software engineers focus on creating new control strategies using Python or C++ and utilize software development tools like package managers and containers. New computing paradigms tailored to industrial control systems have been developed that bridge this divide and integrate information technology (IT) tools into operational technology (OT). Examples include integrating control systems with Cloud computing, High-Performance

work must maintain attribution to the author(s), title of the work, publisher, and DOI

Network Traffic Prediction with Deep Learning-Based Encoder-Decoder Algorithms to Improve the Network Controller NOTED

Elisabetta Schneider
 Supervisor: Carmen Misa Moreira
 Co-Supervisor: Edoardo Martelli

A Report Presented as Part of the
 CERN Summer Student Program 2024

Operational Intelligence for Distributed Computing Systems for Exascale Science

Alessandro Di Girolamo¹, Federica Legger², Panos Paparrigopoulos¹, Alexei Klimentov⁶, Jaroslava Schovancová¹, Valentin Kuznetsov³, Mario Lassnig¹, Luca Clissa^{8,9}, Lorenzo Rinaldi^{8,9}, Mayank Sharma¹, Hamed Bakhshiansohi⁵, Marian Zvada⁷, Daniele Bonacorsi^{8,9}, Simone Rossi Tisbeni¹⁰, Luca Giommi^{8,9}, Leticia Decker de Sousa^{8,9}, Tommaso Diotallevi^{8,9}, Maria Grigorieva⁴, and Sergey Padolski⁶

¹CERN, Geneva, Switzerland
²INFN Turin, Italy
³Cornell University, USA
⁴Moscow State University, Moscow, Russia
⁵DESY
⁶Brookhaven National Laboratory (BNL), USA
⁷University of Nebraska-Lincoln, Lincoln, NE, USA
⁸University of Bologna, Bologna, Italy
⁹INFN Bologna, Italy
¹⁰INFN-CNAF Bologna, Italy


Applications of Large Language Models (LLM)

Rapidly growing interest in the applications of Large Language Models and intelligent conversational agents (AKA ChatGPT-like frameworks)

We have so far inventoried 26 projects and small-scale evaluation activities at CERN to use or develop LLMs and variations/customisations of generative AI models for knowledge discovery, information retrieval, documentation management, user support, software coding assistants, etc. etc.

Most of them comes from ATS Sector Departments

<https://indico.cern.ch/event/1423858/>



AccGPT
A Chatbot for CERN Internal Knowledge

Florian Rehm, Verena Kain, Juan Manuel Guijarro, Sofia Vallecora
28.06.2024

**From an initiative in the BE Dep., today a joint ATS/IT project
(GenAI Pilot Feasibility Study)**

Digital Twins Applications

A digital twin is a digital replica of a physical object, person, system, or process, contextualized in a digital version of its environment. Digital twins can help simulate real situations and their outcomes as part of decision support systems. Developed initially for industrial applications (e.g jet engines or manufacturing plants, today they are being considered for many more types of applications from physics, to climate and medical research. **AI models provide a powerful way of simulating physical process in DT engines.**

Digital Twin Applications plm.cern.ch

Outreach

Training

Design

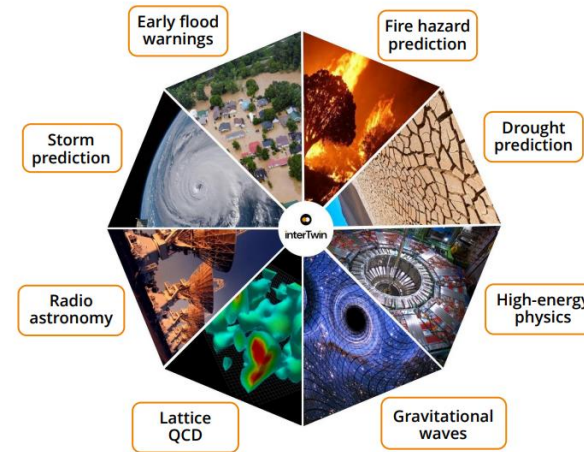
Manufacturing

Maintenance

Operation/Control



interTwin: a Digital Twin Engine for science



29 Participants, including sciences, technology and resource providers.

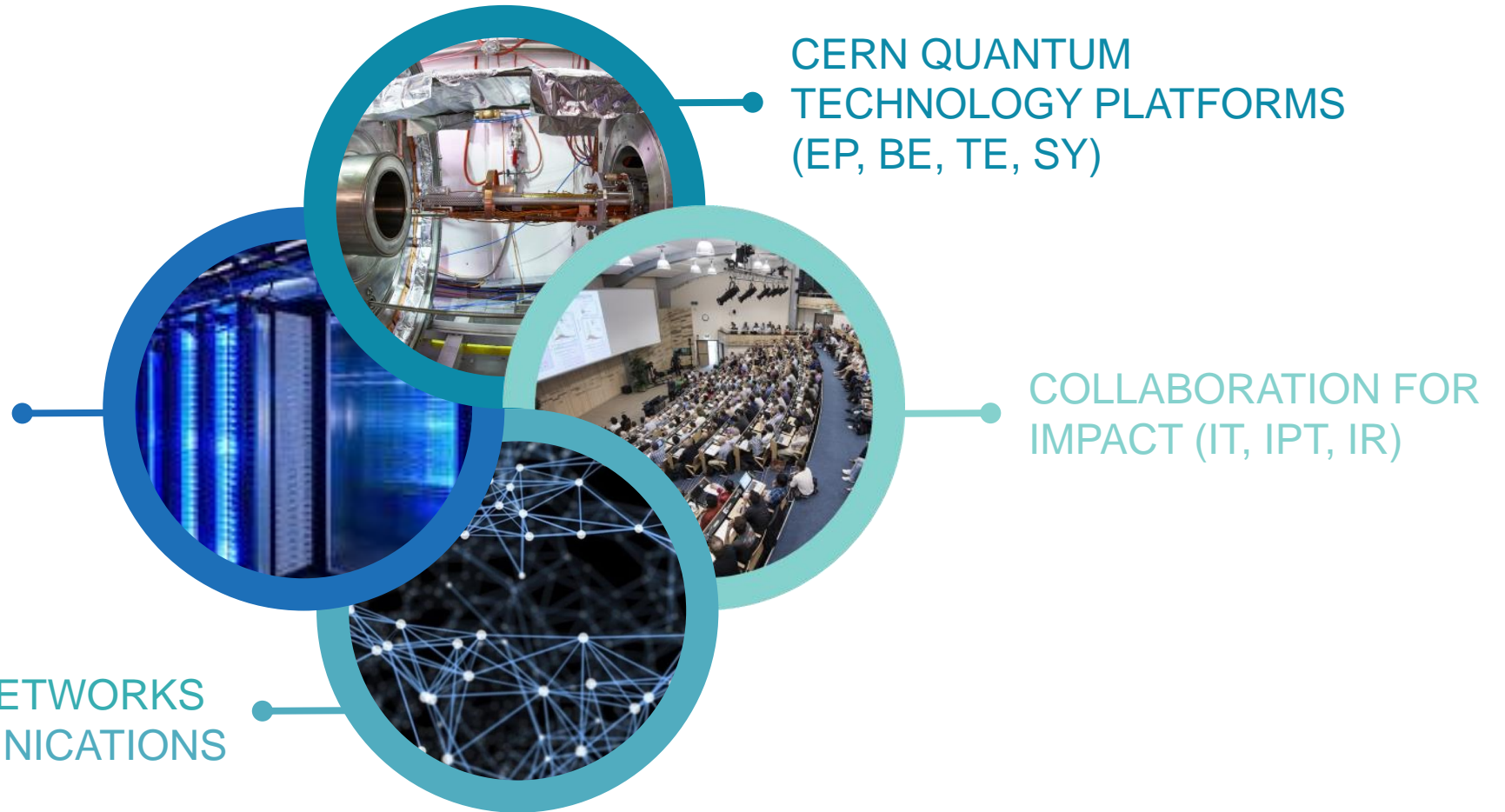


CERN QTI Phase 2 – Centres of Competence

Quantum Machine Learning



HYBRID QUANTUM COMPUTING AND ALGORITHMS (IT, EP, TH)



09/09/2024

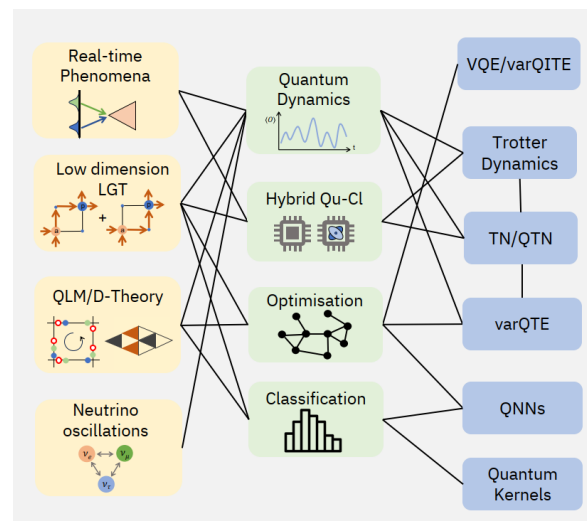
IT Innovation

Foster a expert community studying usability of Quantum Computing for HEP



- Lead the creation of a new community of experts from the Member States and beyond (about 40 researchers worldwide)
- Focus on concrete challenges of QC for HEP
- White Paper on a realistic roadmap in experimental and theoretical physics.
- Growing impact through increasing links with Snowmass initiatives

Di Meglio, A. , et al. **Quantum Computing for High-Energy Physics: State of the Art and Challenges**. *PRX Quantum* 5.3 (2024): 037001.



PRX QUANTUM
a Physical Review journal

Highlights Recent Accepted Authors Referees Search About Scope Editorial Team

Roadmap Open Access

Quantum Computing for High-Energy Physics: State of the Art and Challenges

Alberto Di Meglio *et al.*
PRX Quantum 5, 037001 – Published 5 August 2024

Article References No Citing Articles PDF HTML Export Citation

ABSTRACT

Quantum computers offer an intriguing path for a paradigmatic change of computing in the natural sciences and beyond, with the potential for achieving a so-called quantum advantage—namely, a significant (in some cases exponential) speedup of numerical simulations. The rapid development of hardware devices with various realizations of qubits enables the execution of small-scale but representative applications on quantum computers. In particular, the high-energy physics community plays a pivotal role in accessing the power of quantum computing, since the field is a driving source for challenging computational problems. This concerns, on the theoretical side, the exploration of models that are very hard or even impossible to address with classical techniques and, on the experimental side, the enormous data challenge of newly emerging experiments, such as the upgrade of the Large Hadron Collider. In this Roadmap paper, led by CERN, DESY, and IBM, we provide the status of high-energy physics quantum computations and give examples of theoretical and experimental target benchmark applications, which can be addressed in the near future. Having in mind hardware with about 100 qubits capable of executing several thousand two-qubit gates, where possible, we also provide resource estimates for the examples given using error-mitigated quantum computing. The ultimate declared goal of this task force is therefore to trigger further research in the high-energy physics community to develop interesting use cases for demonstrations on near-term quantum computers.

Received 25 August 2023 Revised 29 March 2024 Accepted 25 June 2024

DOI: <https://doi.org/10.1103/PRXQuantum.5.037001>



AI, Collaborations and Knowledge Sharing

- CERN initiatives
 - ATS Strategy Paper on AI (Verena Kain)
 - EP-SFT Initiative on AI/ML (Lorenzo Moneta)
 - Several IT initiatives (more on this later)
- AI workshop in Dec 2023 moderated by IPT-KT
 - Brainstorming on how to create critical mass on AI at CERN
 - <https://indico.cern.ch/event/1352021/>
- KT Fund projects
 - Several KT projects are based on AI/ML (a few examples later)
- International initiatives
 - CLAIRE
 - ELLIS
 - AI Alliance
 - WEF AI Governance Alliance
 - ITU AI for Good
 - Governmental/Institutional Acts/Policies/Boards
 - Many more...

CERN for AI?

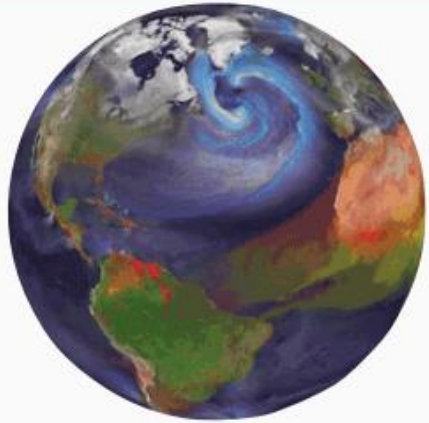
EMP²:

Environmental Modelling and Prediction Platform

A foundation model for the atmosphere



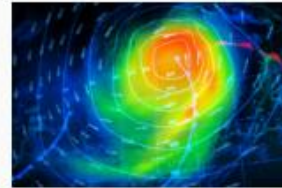
Spatio-temporal representation of atmospheric dynamics



Model
given by the trained neural network

Task dependent
Adaptable smaller
networks

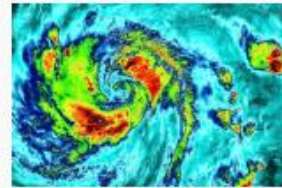
Adaptation



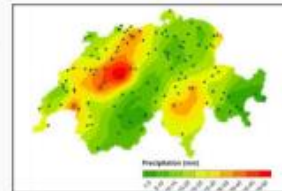
Weather predictions ✓



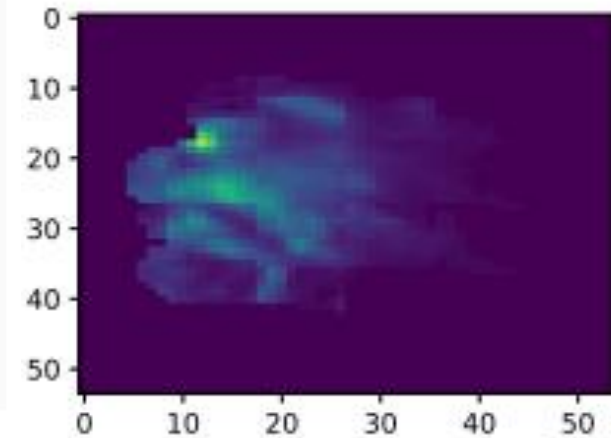
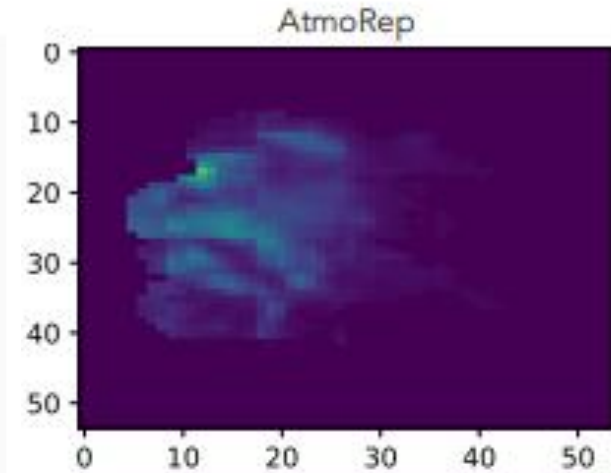
Downscaling ✓



Bias corrections ✓



Spatio-temporal Interpolations (WIP) 🔄



Federated Learning

CAFEIN - Federated network platform for the development and deployment of AI based analysis and prediction models

KEY FACTS

CAFEIN - Federated network platform for the development and deployment of AI based analysis and prediction models

Submission Year


2019

CONTACT PERSON



Luigi Serio


✉ luigi.serio@cern.ch



CERN
CAFEIN

**FEDERATED LEARNING
PLATFORM FOR
COLLABORATIVE AI TRAINING**

Current application:
Multi-institute brain MRI anomaly screening



AI in the IT Department

What we do and plan to do about it

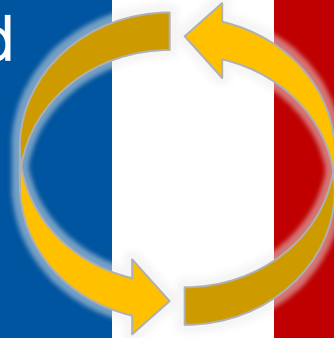
IT Innovation and Engagement Channels

Innovation

A means to discuss co-development projects and investigate new technologies and ways of working

Medium to long-term initiatives lower on the maturity scale

A mix of informal and formal discussions



Engagement

Formal technical and steering committees dedicated to each CERN Sector to discuss requirements for IT services and operations

Short to medium-term initiatives with higher quality of service and resources expectations

Formal documentation and decision tracking

Choose what's best for your project

Innovation Areas

Open Science and Impact

Technology and services
Scale-up, collaborations

Artificial Intelligence

Algorithms, Platforms and Services

Computing

Heterogeneous
Infrastructures and
Software

Data Storage and Management

Hierarchical storage and
data distribution

Long-term investigation

SDIs

Digital Twins

Foundation
Models

Quantum
Technologies

IT AI Working Group

The IT Department is currently planning to consolidate the different activities around AI development and infrastructure under a common “AI Working Group” (name not contractual, still to be defined...)

The objective is to look at AI requirements from all the necessary points of view (computing resources, storage, software tools, skills, etc.) and work with the user community to co-develop future services.

The details will be discussed at the annual Programme of Work and the WG put in place in early 2025

Complementary to the other AI initiatives in other Departments/Sectors, focused on IT infrastructures and Computing Science aspects necessary to support user applications from small (on-premise) to large-scale (Cloud, HPC) operations

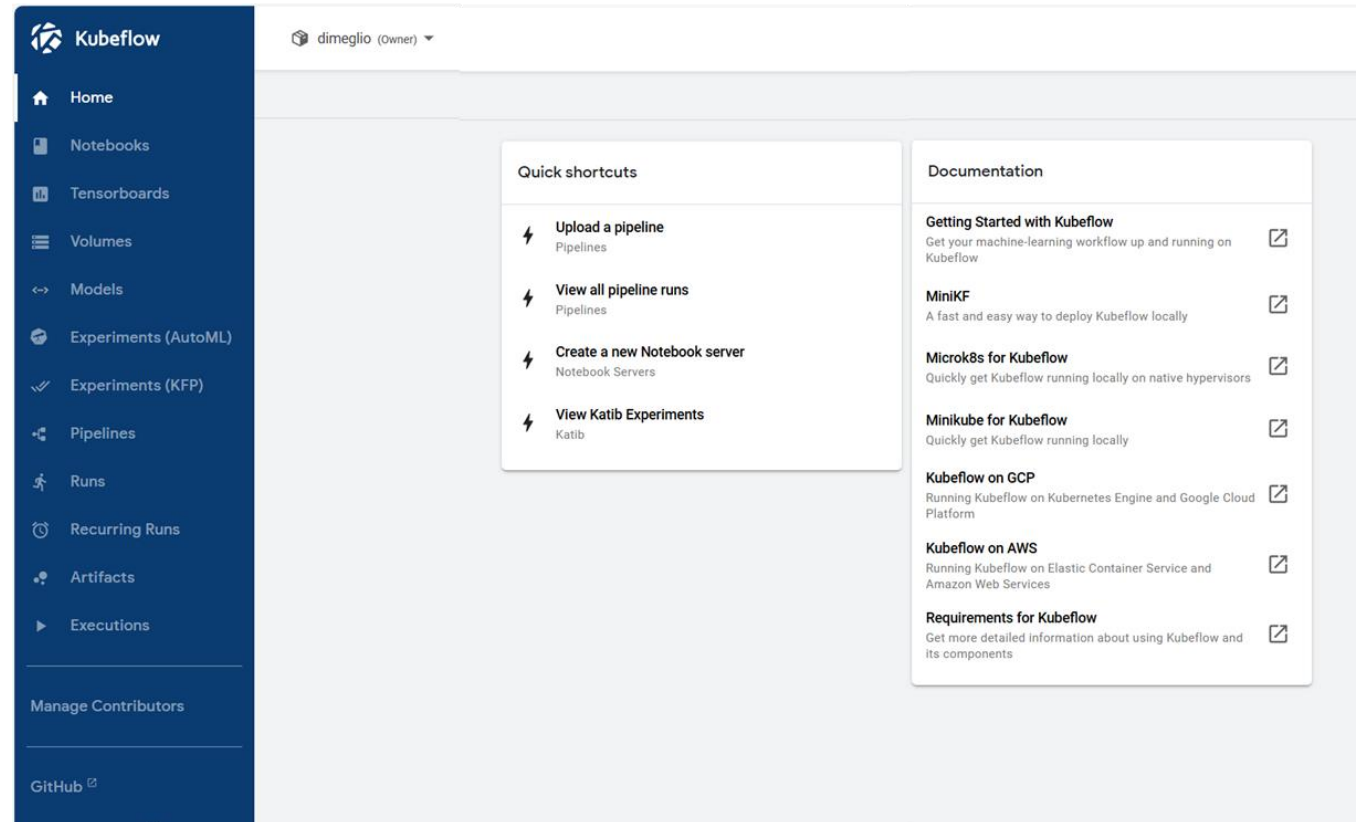
More information: **Sofia Vallecorsa, Ricardo Rocha**

Machine Learning Service

Scalable Machine Learning at CERN with Kubeflow

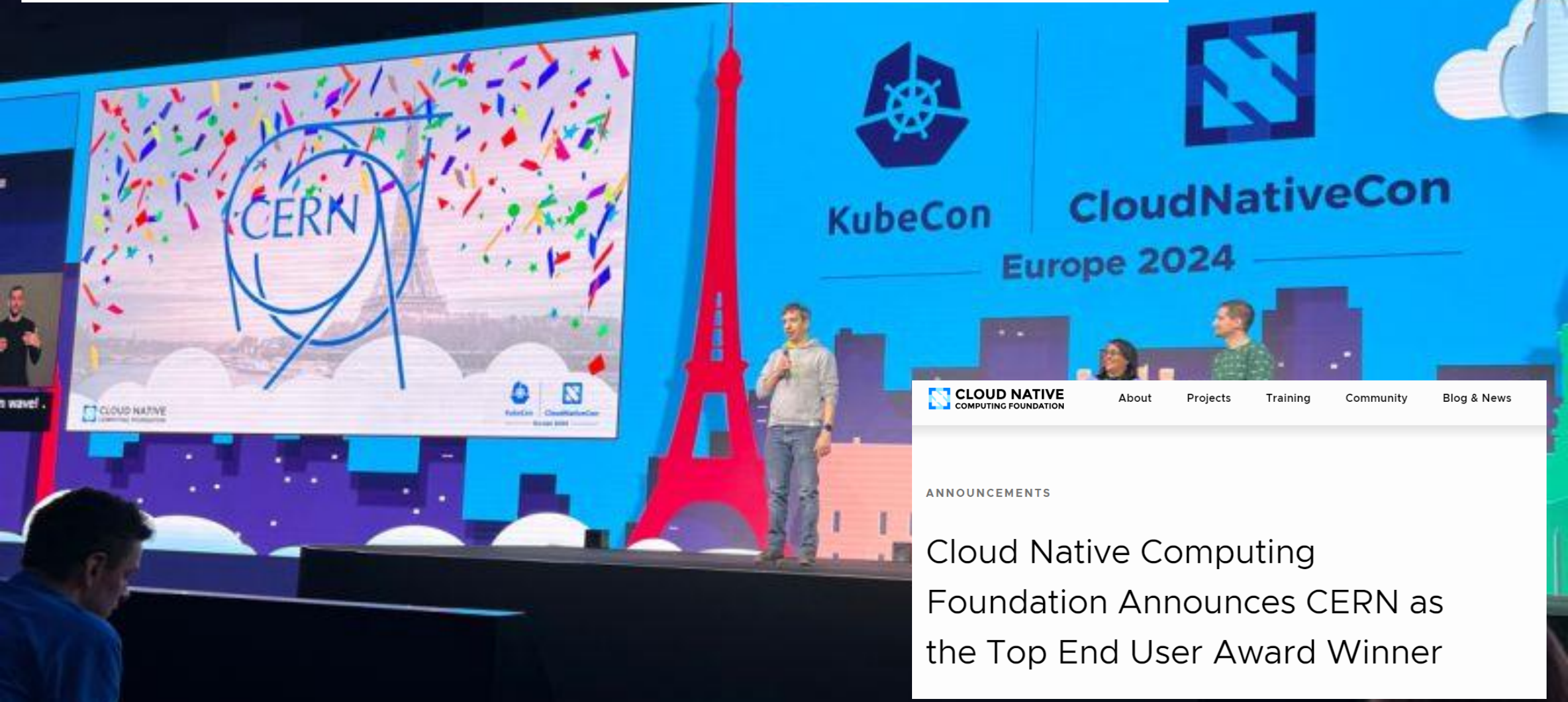
Preparation, Training and Model Serving

Dejan Golubovic, Ricardo Rocha
CERN IT-PW-PI

A screenshot of the Kubeflow web dashboard. The interface is dark-themed with a blue sidebar on the left containing navigation links: Home, Notebooks, Tensorboards, Volumes, Models, Experiments (AutoML), Experiments (KFP), Pipelines, Runs, Recurring Runs, Artifacts, and Executions. Below these are links for Manage Contributors and GitHub. The main content area is light gray and shows a user profile "dimeglio (Owner)" at the top right. It features two columns of content: "Quick shortcuts" with links for "Upload a pipeline", "View all pipeline runs", "Create a new Notebook server", and "View Katib Experiments"; and "Documentation" with links for "Getting Started with Kubeflow", "MinikF", "Microk8s for Kubeflow", "Minikube for Kubeflow", "Kubeflow on GCP", "Kubeflow on AWS", and "Requirements for Kubeflow".

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

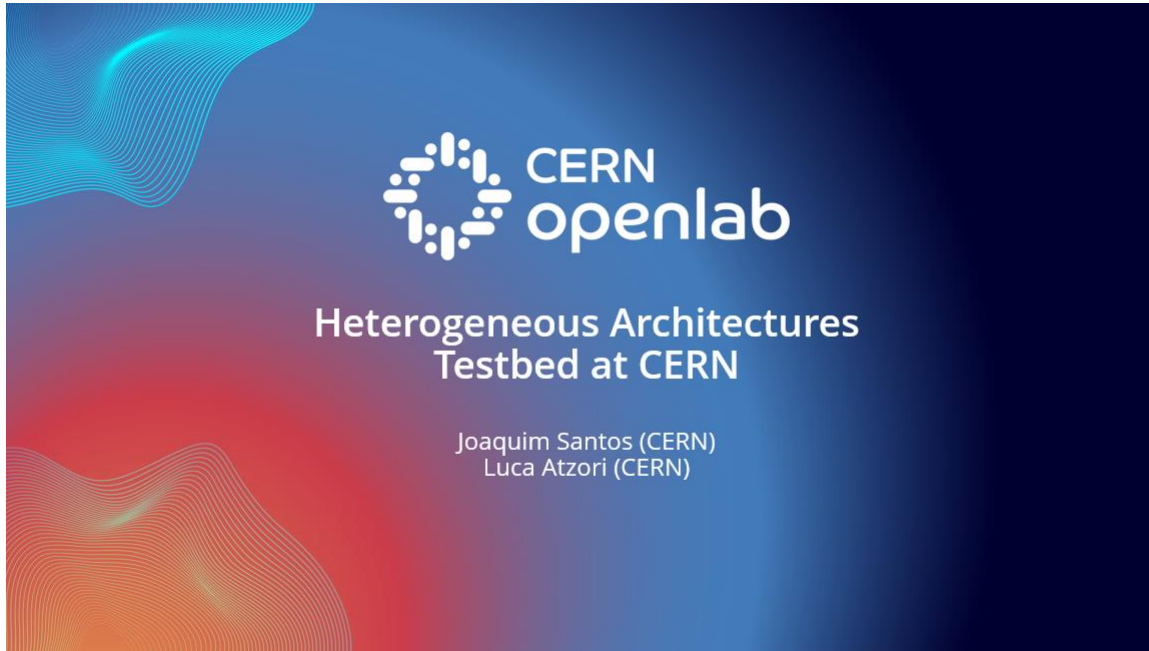
“CERN’s innovative use of cloud native technologies is a shining example of how open source and collaboration can drive cutting-edge research,” said Taylor Dolezal, head of ecosystem, CNCF. “By leveraging Kubernetes and other CNCF projects at an immense scale, CERN demonstrates the power of cloud native to tackle the world’s most complex challenges. We are thrilled to recognize their outstanding contributions with the Top End User Award.”



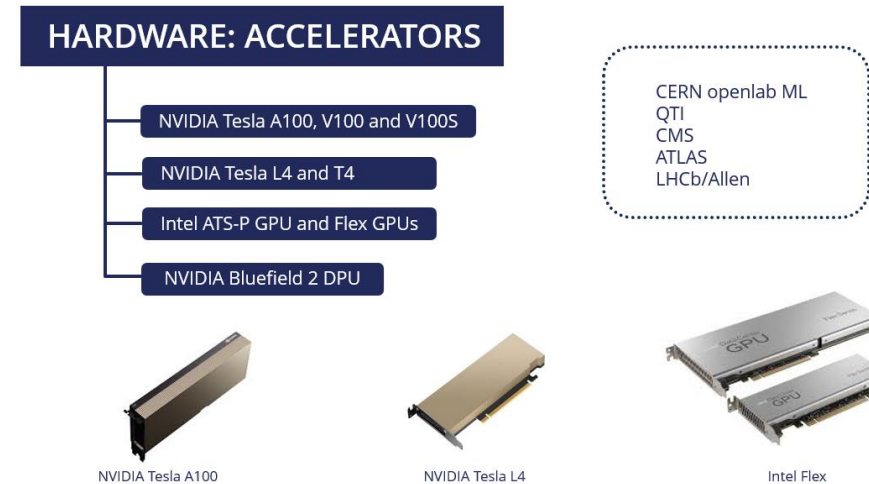
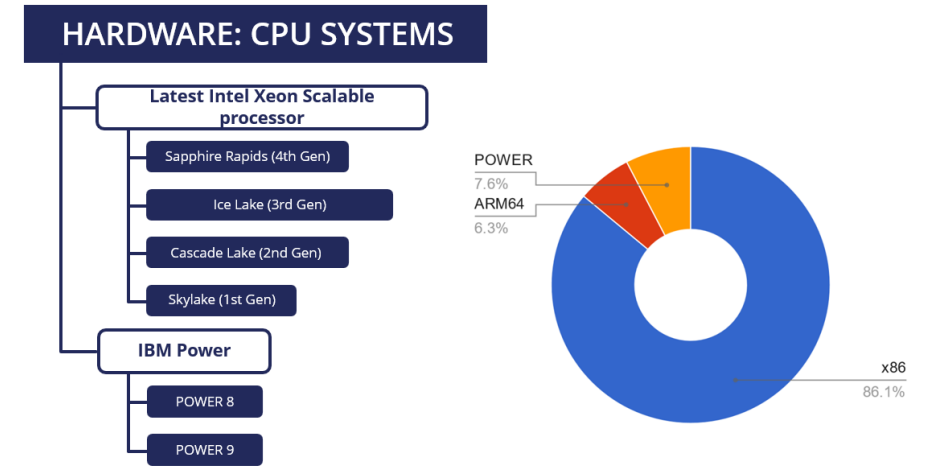
ANNOUNCEMENTS

Cloud Native Computing Foundation Announces CERN as the Top End User Award Winner

Heterogeneous Computing Testbed



Working on providing access to test resources in Cloud and HPC environments



IT Energy & Carbon Aware Computing Programme

Initial Lines of Action

Carbon aware HEP data processing: energy benchmarking of HEP simulation software applications (e.g. MadGraph and AdePT) as a model to expand to other HEP applications.



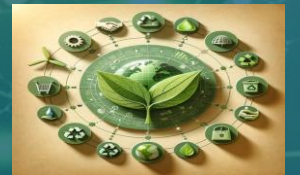
Sustainable AI: Assessment of the environmental impact of IT ML services to include energy-efficiency aspects by design (models training and reuse, communication patterns, data formats).



Promote sustainable computing and green software patterns in the existing educational programmes such as the CERN School of Computing.



Green Procurement: mainly on-premises but also gradually leveraging the public cloud, developing strategies for low carbon intensity deployments.



CERN Data Centres: further develop strategies to increase low carbon energy consumption and continuous improvement of infrastructure lifecycle



Learning and Education Opportunities

Thematic CERN School of Computing on Machine Learning 2024

Overview

The 15th **Thematic** CERN School of Computing (tCSC **Machine Learning 2024**) will take place on **October 13-19, 2024**.

Scientific Programme

Timetable

Application

Privacy Information

School guide

Lecturers

Organisers

Practical Information

- Terms and Conditions
- Fees and Payment
- Sport/spare time
- Laptop configuration/CERN services activation

Visit Split


CERN School of Computing Contact
Computing.School@cer...

The school will focus on the theme of **Machine Learning** and **Artificial Intelligence** applied to **Data Analysis** and **Accelerator Technology**. The programme will offer 22 hours of lectures and hands-on exercises, and student presentation sessions.

This school is organized by **CERN** in collaboration with **the Faculty of Science, University of Split**. The school will take place in **Split, Croatia**, and be hosted at the Mediterranean Institute For Life Sciences (**MEDILS**) Conference Centre. The Centre is a historical renovated building situated in a wooded and landscaped park located on the Adriatic Sea coast, a few kilometers from the centre of Split.

Important dates 2024

- 8 May application opens
- 19 June application close
- 3 July invitations sent to selected students
- 4 September participation fee deadline



NextGen

About Us Activities News & Events Resources Jobs

> Activities > WP4

Task 4.2: The STEAM Programme (Software Training, Education, and Advanced Modules)



Task lead: Felice Pantaleo (CERN)

The CERN-STEAM Programme is an initiative designed to equip postgraduate students, Ph.D. scholars, and researchers with cutting-edge computing and data science skills, ensuring a vibrant future for the field of research. This comprehensive and immersive educational program focuses on critical areas such as algorithms design, AI, trigger systems, heterogeneous computing, and quantum computing as applied to HEP. Renowned professors and experts from academia and industry will give lectures, seminars, hands-on training, and hackathons to bridge the skills-gap between academic proficiency and autonomy in developing cutting-edge technologies within the NGT project. The Programme aims to provide an enriching learning experience complementing and building upon the courses taught in established schools and events in the field, through the practical application to CERN experiments' realistic use cases. We will investigate how to make the Programme courses eligible for European Credit Transfer and Accumulation (ECTS) credits.

CERN openlab

Summer Student Lectures

Below you will find the list of the lectures presented by CERN openlab, to give CERN summer students an introduction to the main topics and challenges we address in openlab projects. In these lectures we are going to cover topics such as Artificial Intelligence (AI), High-performance computing (HPC), Quantum Computing, heterogeneous computing and algorithms, digital twins, advanced storage solutions... and their applications to real use cases at CERN!

July 2024

- 31 Jul [Quantum Computing Applications and Use-cases](#)
- 30 Jul [Introduction to Quantum Computing, Quantum Machine Learning and Optimization](#)
- 29 Jul [Alexander Zochbauer, Kalliopi Tsolaki, "Digital Twins: introduction and use cases"](#)
- 26 Jul [Ilaria Luise, Sofia Vallecorsa, "Foundation models"](#)
- 24 Jul [Danilo Piparo, Marta Czurylo, Vincenzo Eduardo Padulano, "ROOT Summer Student Workshop"](#)
- 23 Jul [Stephan Hageboeck, "GPU programming"](#)
- 22 Jul [Axel Naumann, "Best practices: the theoretical and practical underpinnings of writing code that is less bad"](#)
- 18 Jul [David Southwick, "High Performance Computing"](#)
- 17 Jul [Abhishek Lekshmanan, "Storage"](#)
- 17 Jul [Luca Atzori, "Data Centre Hardware"](#)

Thanks!

alberto.di.meglio@cern.ch

[@AlbertoDiMeglio](#)