AI developments at IN2P3

David Rousseau (IJCLab), Alexandre Boucaud (APC), with inputs from the IN2P3 AI community

IN2P3 Conseil Scientifique June 2022 ML report v1.0

Summary

The use of AI across all scientific fields covered by IN2P3 has been growing exponentially in the last few years. This report gives a brief – necessarily incomplete – overview of the different techniques applied in the various physics domains, laying out along the way the specificities of AI in HEP w.r.t. other application domains of AI.

The content of this report is based on an IN2P3-wide survey carried out in May 2022. It gathers answers to about 60 AI projects (a project being a successful application to calls from ANR or Labexes, a PhD thesis or an informal study undertaken by a few people). It has allowed us to sketch in some detail the environment of people working with AI at IN2P3.

The main conclusions of the report can be summarised as follows

- the AI computing resources are currently diverse and adequate at IN2P3, but the computing needs for AI in physics are expected to grow in the near future,
- human resources are a vital ingredient for the sustainable growth of these developments, with more and more PhD with a significant AI component and a recent influx of postdocs funded by different projects.
- now that the pandemic is over, there is an expressed need for high-level training and networking opportunities, also travel to AI conferences and exchanges with AI specialists.
- opportunities for publication are numerous but sometimes require an extra effort to publish papers in AI journals in addition to the physics paper with only AI applications.

Introduction

Artificial Intelligence has been developing quickly in recent years at IN2P3. This report gives a brief overview of the different methods being developed in different fields, before focussing on a few relevant issues: resources, training and workshops, and publications, before concluding with a SWOT analysis. The document was built from work done for the IN2P3 Prospectives at Giens in fall 2021, updated using a broad survey conducted in May 2022 on AI projects at IN2P3, which received 60 individual answers.
The document has been written as follows: first, a scientific overview briefly lays out the different activities where AI plays a role, with examples from IN2P3 activities, and the specificities of AI in HEP are detailed. Second, the computing, financial and human resources issues are outlined. A SWOT analysis precedes the conclusion.

1 Scientific overview

Experimental High Energy Physics relies on two steps : (i) collect data from a particular interaction (e.g., high energy proton-proton collision, cosmic ray) with an appropriately designed detector (ii) infer from the data collected a measurement (a confidence level) on a nature parameter. The data collected comes in generally as a collection of ”events”, where an event refers to an astrophysical image or to the data collected about a particular event, e.g., a proton collision, a neutrino interaction, a gravitational wave signal, or an atmospheric shower.

The parameters of the visible particles or astrophysical objects are inferred algorithmically in what can be called **detector-level inference** (or classically reconstruction). A first step is to group energy deposits corresponding to the same particle (clustering, tracking; or deblending), then to identify them – possibly using a *classifier* – and/or determine their parameters (e.g. 4-momentum vector and origin typically; shape, orientation and flux for galaxies) – possibly using a *regressor*. There may be additional steps (like identification and calibration of jet).

The properties of the particles of a single event are then combined to infer what has happened in this particular event in what can be called **event-level inference**. The same happens when combining multi-messenger observations of the same astrophysical source or multi-band flux of galaxies to obtain their properties (e.g. redshift).

Finally, the study of all the recorded data (often billions of events), with the help of very accurate simulators, allows inferring fundamental laws of nature, more specifically a confidence interval on a specific parameter, in what can be called **experiment-level inference.**

When **detector-level** or **event-level** inference are made online, for preprocessing, triggering or filtering, there is the additional constraint of inference speed, for which specific hardware may need to be designed, which can be grouped under the **fast AI** label**,** for example, to compute calorimeter energy from time samples (see details in the report about real-time data processing in particle physics).

HEP has a several decade-long history of developing complex models to compute and simulate particle collisions, the interaction of particles in detectors (Geant4), lattice QCD, accelerator orbits, N-body simulations, etc. These models have grown in complexity, precision, and resource requirements (memory, computing time). As a properly trained sufficiently large neural network can emulate any multi-dimensional function, **Surrogate Models** have been developed**.** They are trained to emulate the original models within a limited region of the parameter space, providing several orders of magnitude faster throughput under tight accuracy constraints. A class of Surrogate Models are generator models (using e.g. Generative Adversarial Network or Variational Auto Encoder) which can emulate Geant4 to simulate the interaction of particles in a detector. Another is to emulate a multidimensional function, for example, to estimate the stability of an accelerator orbit.

In most cases, AI at HEP falls under the category of *supervised learning*, where an algorithm is trained (the **training** stage) and on labelled data (typically simulated events for which the label, e.g. “signal” vs “background”, is known) and validated and finally applied (the **inference** stage) to real data.

Another AI field separate from supervised learning is **Optimisation** with techniques like Bayesian optimisation or reinforcement learning, possibly using differentiable programming, which can be applied to accelerator tuning or to the design of future experiments.

A separate topic is Quantum Machine Learning, where the use of quantum devices to learn a model is investigated, see details in the report about quantum computing.

2 Specificities of AI in High Energy Physics

Probably the most common use of AI in particle physics is the use of Boosted Decision Trees on tabular data (like classifying events from a dozen of high-level features). Most recent developments attempt to introduce AI earlier in the processing chain, where data is usually only **semi-structured data**. This is the first specificity of AI in HEP, which is that in most cases (with the notable exception of astrophysics), data are not tabular data, nor images, nor time series, which are the three types of data for which there exists abundant literature and wealth of tools and methods. Graph Neural Networks have been gaining popularity as a versatile architecture suitable for many HEP-specific problems.

A second specificity is that in HEP, we have developed for decades **very good simulators** (meaning describing accurately measured data) to deliver labelled simulated data on which to train algorithms (for some domains like cosmology, this might not be so true). The remaining imperfection of simulators must be dealt with, in particular when evaluating the systematic uncertainties.

The third specificity of that High Energy Physics is that we have been doing **data science for half a century** since the invention (in 1968) of the multiwire proportional chamber, which allowed us to fully automatise the data processing. The community has been building tools and infrastructures to deal with millions, billions, trillions of instances, and gigabytes, terabytes, petabytes of data (and more). Boosted Decision Trees have gained popularity in HEP since ~2010 and (modern) Neural Networks since ~2014[1]. These AI algorithms developed are inserted in these time-tested pipelines and benchmarked against “classical” but sophisticated algorithms. In other scientific fields (e.g. genomic), the data avalanche is much more recent.

Finally, the fourth specificity is that our primary output is scientific papers, in which the main message is a measurement with a confidence interval. The **confidence interval** covers statistical uncertainties and all sources of known hypothetical systematic uncertainties. The confidence interval also carries an unquantifiable element of trust in the techniques used. This has to remain true when AI takes a growing role in the production of the measurement.

With these specificities, AI in HEP is not a matter of just applying off-the-shelf AI techniques but requires specific developments.

3 AI at IN2P3

This section analyses the answers to the IN2P3-wide survey carried out in May 2022. It gathers answers to about 60 AI projects. Although not guaranteed to be exhaustive, this survey allows broad brush cartography of AI at IN2P3, which will need to be detailed in the future. Given the space constraints of this report, a choice was made to limit the analysis of the survey to statistical analysis, followed by tables of ANR JCJC and PRC projects and PhD theses.

The different types of projects in which AI is developed at IN2P3 are reported in Figure 1. PhDs are clearly the main structuring channel (with a total of 29 ongoing or recent PhD, including four Computer Science PhD and two physics PhD with a CS co-supervisor). ANR JCJC and collaborative, which usually fund at least one post-doc, are also significant.

The main physics domain of these projects is reported in Figure 2. We can see that the main IN2P3 domains are present.

The type of input data is reported in Figure 3. As already mentioned it is very diverse and spans the diversity of data used in the AI world. One sees that “image” and “time series” contributions are large, despite the statement in the overview that they are not so frequent at IN2P3. This is most likely a survivor bias: applying AI to images or time series is so natural that physicists dealing with these data will more naturally switch to AI compared to others.

[1]A.Radovic et al., Nature 2018 Aug;560(7716):41

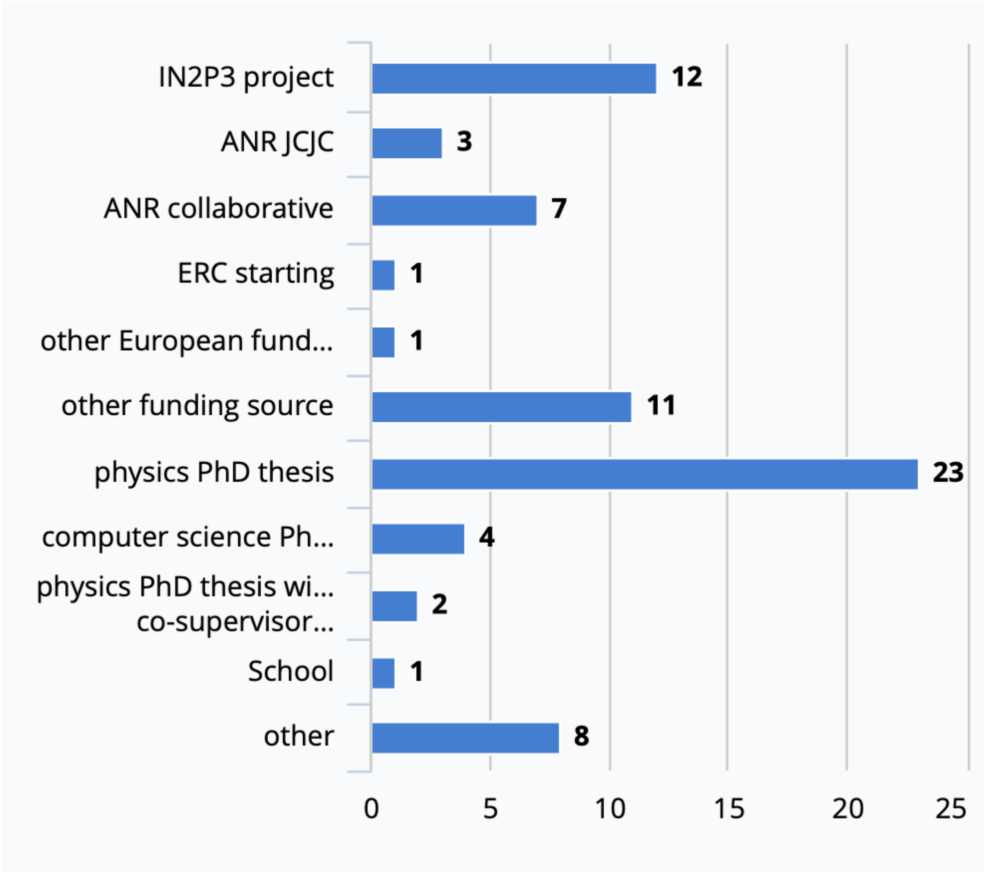

***Fig 1**: Type of funding used or obtained by IN2P3 members. Several answers were possible.*

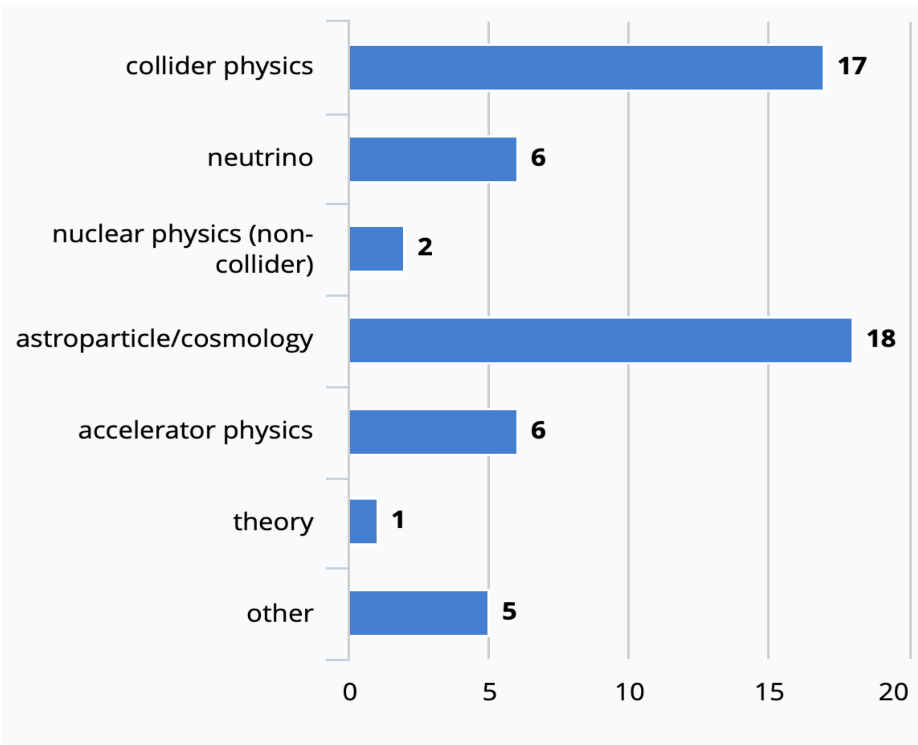

***Fig 2**: Main physics topics of IN2P3 members practicing AI*

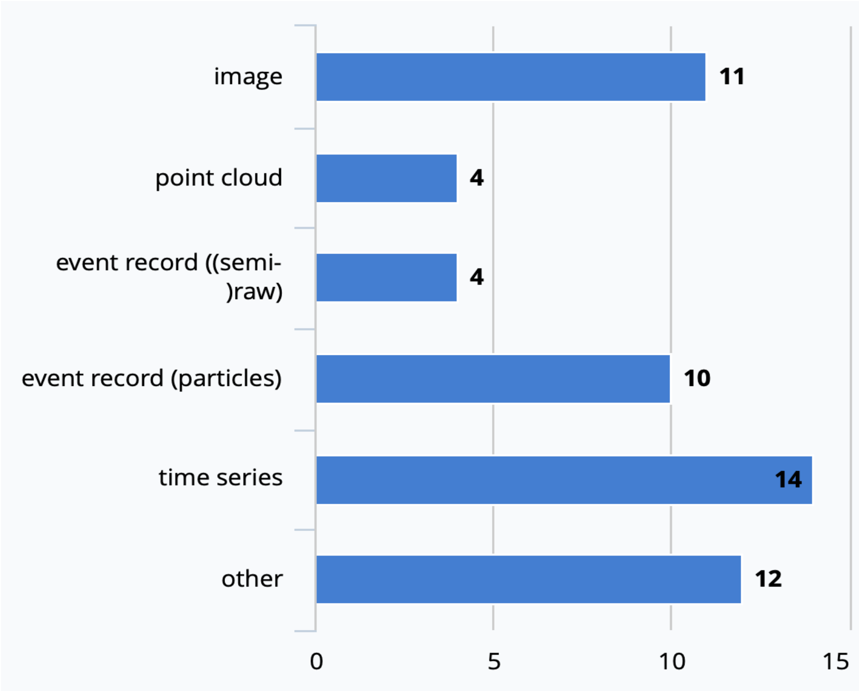

Fig 3: *Type of input data. Several answers were possible.*

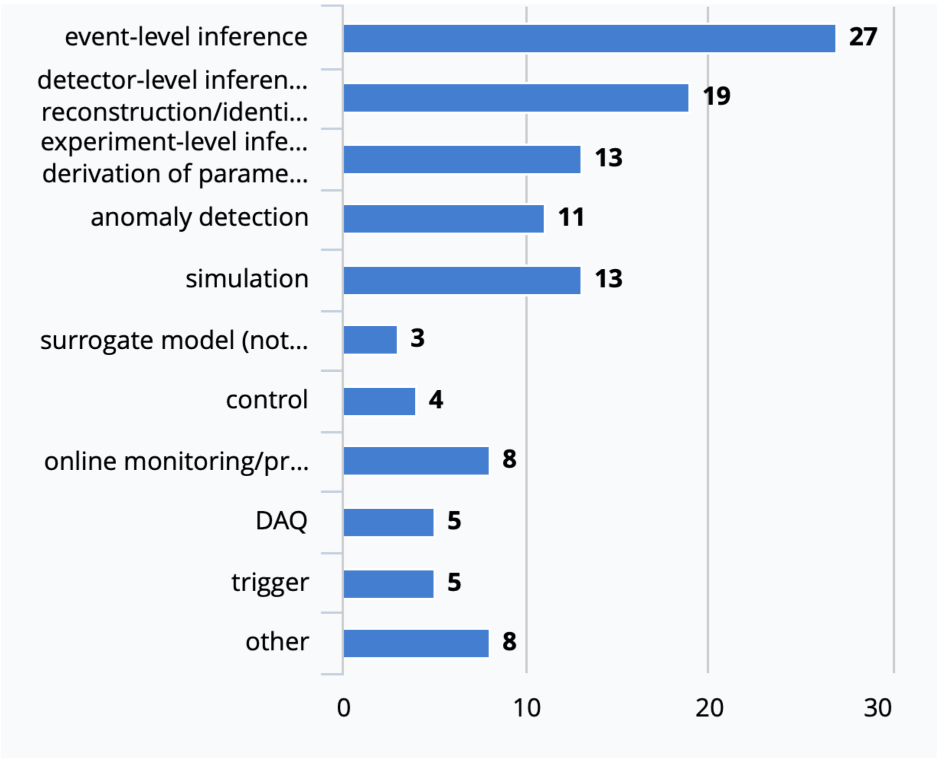

Fig 4: *Main AI themes*

A variety of AI techniques are used at IN2P3 (see Figure 5). Boosted Decision Trees (BDT) contribution is modest, but this is primarily due to a bias in the poll: BDT is so commonly used that many physicists using them are simply not reporting about them.

Finally, to give a flavour of the diversity of AI topics at IN2P3, this section ends with two tables listing ANR projects and PhDs where AI plays a significant role. Before this, it should be noted there is an ERC Starting Grant USNAC started in 2021 at IP2I on cosmology which will build a full forward modelling going from cosmology parameters to SN Ia distances and redshift through local Universe density and velocity fields. There are no other ERC grants we know of.

Table 1 below is reporting the list of ANR JCJC and PRC projects with a significant AI role, ordered by starting year.

Type	Acronym	Start	IN2P3 labs	Non IN2P3 labs	Description
ANR JCJC	HIGRANTS	2018	LLR		Optimization of ML models for the HGCAL reconstruction at the Level 1 trigger
ANR JCJC	AIDAQ	2019	CPPM		Artificial Intelligence on FPGAs: for Data AcQuisition in hep experiments
ANR PRC	DEEPDIP	2019	CPPM	LAM, IAP, LIRMM	DEEP learning for Deep Imaging Projects
ANR PRC	ASTRODEEP	2019	APC	LORIA, DAP	Analysis of massive astronomical data with Machine Learning
ANR PRC	DMwithLLPatLHC	2021	LPNHE, LPSC		Search for Dark Matter with Long-Lived Particles at the LHC
ANR JCJC	FIDDLE	2022	IPHC		Full Event Interpretation using Graph Neural Networks at Belle II
ANR PRC	ATRAPP	2022	LAPP,IJCLAB		Advanced particle tracking algorithms for particle
ANR JCJC	OGCID	2022	LLR		Optimal Graph convolution for efficient particle identification
ANR PRC	RICOCHET	2022	APC	GIPSA, CRISTAL,CRAN	Bivariate signal processing : a geometrical approach to decypher polarisation

***Table 1**: AI ANR JCJC and PRC at IN2P3*

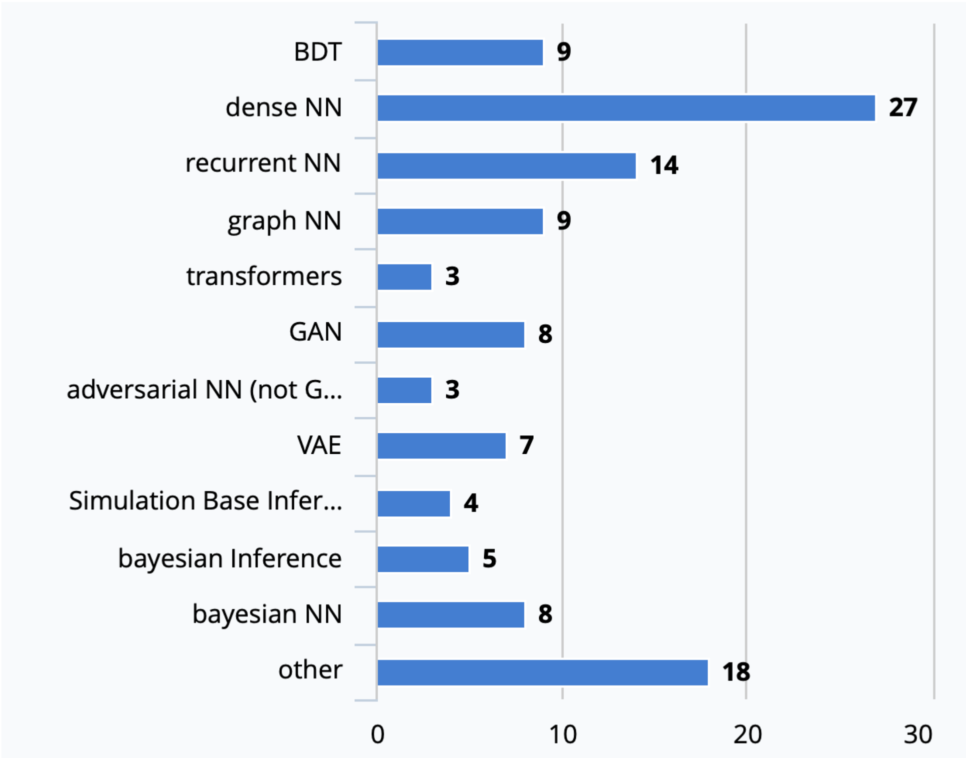

***Fig 5**: AI techniques. Several answers were possible.*

Table 2 below shows the list of recently finished or ongoing PhD at IN2P3, where AI plays a significant role. The table is ordered by starting year and field.

Field	Experiment	Lab	Start	Subject
collider physics	ATLAS	IJCLAB	2017	Simulation of the ATLAS electromagnetic calorimeter using generative adversarial networks and likelihood-free inference
accelerator physics	MYRRHA	LPSC	2018	Study of machine learning methods for optimization and reliability improvements of high power linacs
accelerator physics	MYRRHA & IPHI	LPSC	2018	Study of machine learning methods for optimization and reliability improvements of high power linacs
astroparticle/ cosmology	CTA	LAPP	2018	GammaLearn: DEEP LEARNING FOR IMAGING CHERENKOV TELESCOPES DATA ANALYSIS
astroparticle/ cosmology	ALTO/CoMET	APC	2019	Deep Learning for signal/background separation in ALTO/CoMET (DeepCoMET)
astroparticle/ cosmology	large photometric surveys	CPPM	2019	Deep learning applied to cosmology

collider physics	ATLAS	CPPM	2019	ATLAS ttH(H->bb) Run 2 legacy analysis
collider physics	ATLAS	L2IT	2019	Track reconstruction at HL-LHC with Graph Neural Networks
accelerator physics	THOMX	IJCLAB	2020	use of machine learning to tune and control particle accelerator
astroparticle/ cosmology	Euclid	APC	2020	Probabilistic segmentation of overlapping galaxies
astroparticle/ cosmology	Euclid	APC	2020	Generative models : Variational Autoencoders (VAE) with CNN and normalizing flows
collider physics	ATLAS	IJCLAB	2020	Analysis of off-shell Higgs into 4 leptons at ATLAS using simulation-based inference
collider physics	ATLAS	LPSC	2020	Hadronic jet energy and mass calibration with DNN
astroparticle/ cosmology	IceCube & Fermi	APC	2021	Cosmic Neutrino Investigation via source Classification (CoNIC)
astroparticle/ cosmology	Rubin	APC	2021	Galaxy scene deblending
astroparticle/ cosmology		APC	2021	Accelerating SBI with score-matching
astroparticle/ cosmology		CPPM	2021	Constraints on gravity by tomographic galaxy clustering with Euclid data
collider physics	ATLAS	CPPM	2021	DIPS with ITk
collider physics	CMS	IP2I	2021	BSM(Tprime to tH hadronic) search
neutrino	JUNO	Subatec h	2021	Neutrino events reconstruction and identification in a liquid scintillator detector
neutrino	Spherical detector s at Lp2ib	LP2IB	2021	R2D2: neutrino double bêta decay R&D
accelerator physics	SPIRAL2	GANIL LPSC	2022	AI for cryogenics and RF of superconducting accelerators (ACRAS)
astroparticle/ cosmology	HESS, CTA	APC	2022	Deep Learning for event classification in Imaging Atmospheric Cherenkov Arrays (FiBER)
collider physics	ATLAS	IJCLAB	2022	Measurement of SMEFT parameters using Simulation Based Inference in the off-resonance Higgs to 4 leptons channel in ATLAS
neutrino	KM3Ne T	APC	2022	OrcaNet Reconstruction
neutrino		LP2IB	2022	AI development for the R2D2 experiment

nuclear physics (non-collider)	AGATA	IP2I	2022	Machine Leaning technics for gamma ray tracking
nuclear physics (non-collider)	FALSTAFF	GANIL	2022	Fission-fragment charge-identification using neural-network analysis of ionisation-chamber tracks
theory		IP2I	2022	Unsupervised sampling NN for high dimensional unconstrained parameter spaces

Table 2*: Recent and on-going AI PhD at IN2P3*

4 Means and Resources

4.1 Computing resources

Developers of AI techniques require computing resources to train their models. The simplest ones can be trained in a few minutes on a laptop. More complex models can take days on GPUs, sometimes even more. It should be noted that the total computing time used by a project may be one thousand times larger than the training time of the final model, taking into account the many iterations to optimise the architecture and other *hyper-parameters*.

IN2P3 AI developers report (see Figure 6) using mostly laptop and campus-level clusters, a growing number of them also using the GPU farm at CC IN2P3, especially through the interactive Jupyter notebook platform, a handful (the more involved) the Jean-Zay super calculator at Idris or the cloud resources for specific needs (AWS, AzureML or GoogleCloud). The general trend in the AI world is to use larger and larger models, and the trend at IN2P3 is to apply AI earlier in the processing chain (at detector-level rather than event-level), hence on more features. Overall this should lead to an increasing need for computing resources in the coming years.

There is a growing worry worldwide about the resource-frugality of training large models. However, computing resources used for training at IN2P3 are many orders of magnitude smaller than resources allocated to simulate and process large experiments. In addition, many AI projects are about speeding up tasks, hence reducing overall resource needs.

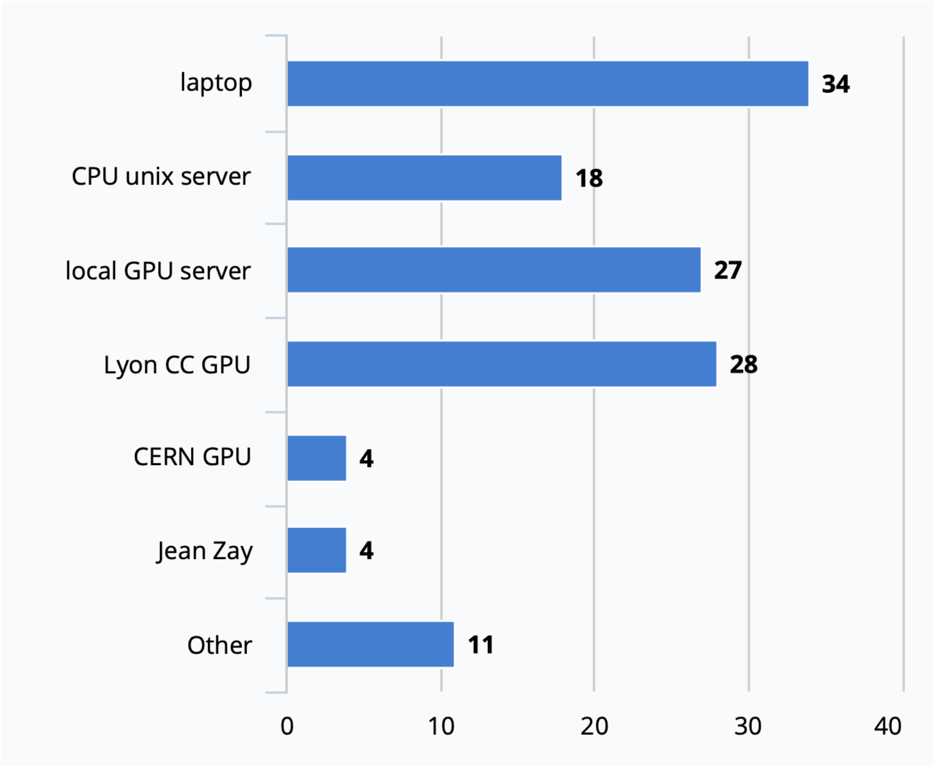

Fig 6: *Type of computing resources*

ML software tools used by IN2P3 members are reported in Figure 7 (with the under-reported contribution of TMVA, which is the ML suite in ROOT). Overall this is mostly scikit-learn associated with one NN library, with twice more users of Tensorflow than Pytorch.

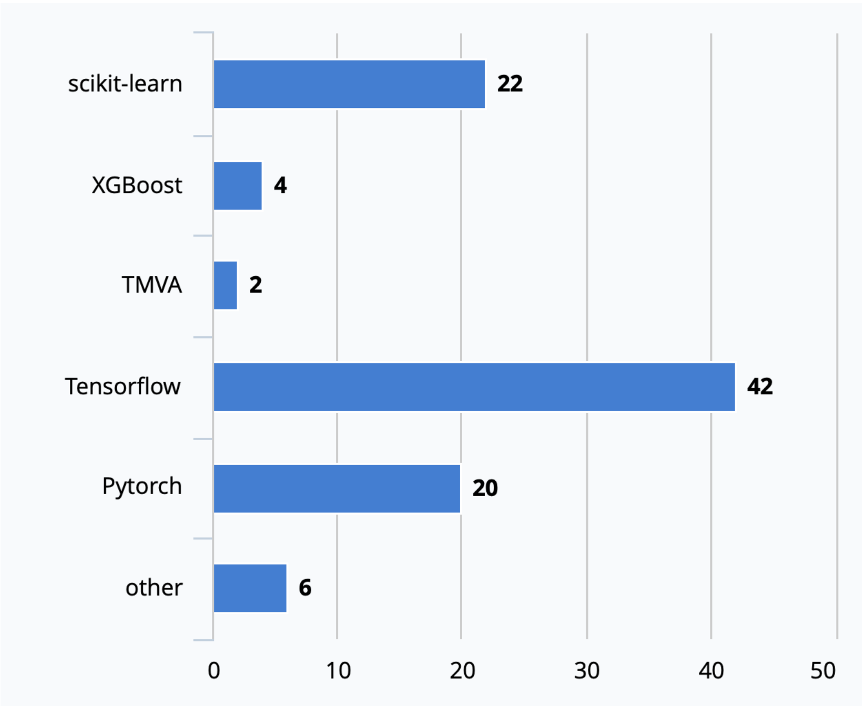

Fig 7: *ML frameworks*

4.2 Human resources

A key ingredient for the development of AI techniques is trained manpower. These are engineers and physicists that have been involved in multiple AI projects in recent years. The majority of the trained manpower at IN2P3 currently relies on non-permanent people (PhDs, postdocs, CDD engineers).

There are more and more PhDs in which AI plays a major role: about 30 are very recent or ongoing at IN2P3, of which five are co-supervised by a computer scientist. After such PhDs, they can be hired as post-docs (funded by ANR or other grants), although the recruitment is made difficult both by the large salary gap with industry with such skills in their portfolio and the other big gap between the competitiveness of recruitment in academia vs headhunters reaching out with very competitive job offers.
This makes the recruitment for permanent positions of AI and physics specialists (physics PhD with in-depth experience in AI, as the opposite, is extremely rare) who can promote AI locally and through IN2P3 very precious.

Finally, given the complexity and specificities of AI development for HEP, close collaboration with computer scientists – from both academia and industry – should be established, as it has been done at many IN2P3 labs. Such collaborations appear to be fostered by the co-supervision of PhD students. However, the physicists involved should be aware of the specific CS field (see later).

4.3 Financial resources

All cutting-edge AI software (Python AI ecosystem, scikit-learn, deep learning frameworks Tensorflow (Google) or PyTorch (Facebook)...) is open-source and, as such available at no cost. Then, provided that the computing and human resources are granted, as mentioned above, the only financial resources needed are for training and networking (see later).

A 15k€ Machine Learning IN2P3 Project “CompStat” exists, which can fund actions like IN2P3 engineers and physicists going to “pure” AI conferences or invite of Computer Scientists for short stays at IN2P3 laboratories. These activities were on hold during the pandemics however there have been about 50 (evenly shared between both actions, see Figure below) expressions of interest for the near future. Clearly, such actions should also be funded through other means (laboratory/team budget, local call).

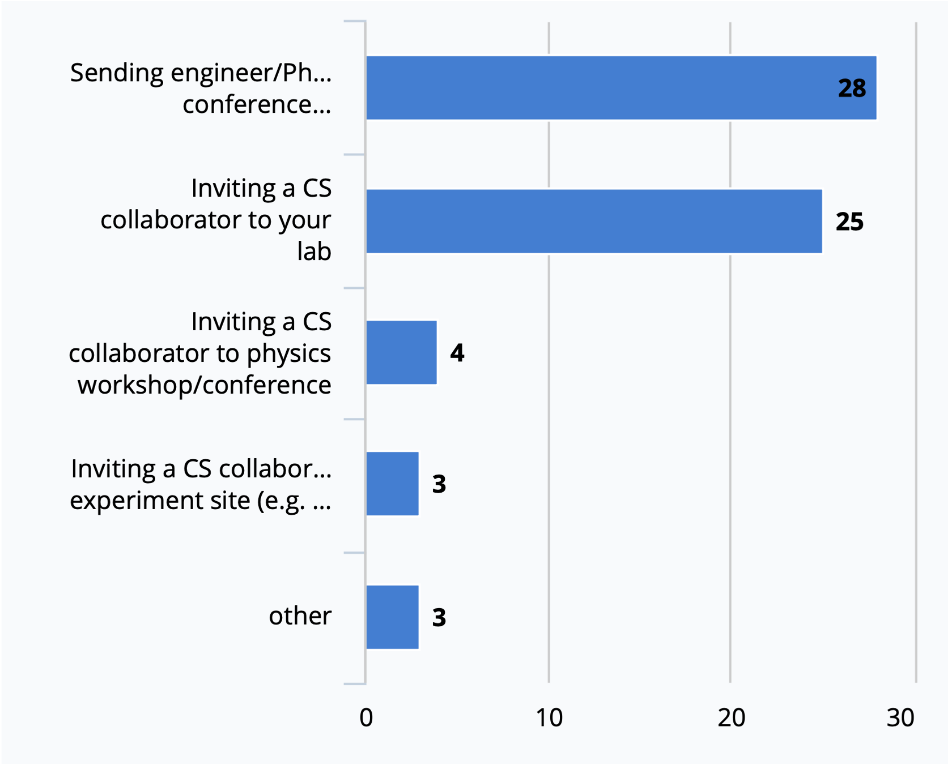

Fig 8: *Preferred support request. Several answers were possible.*

4.4 Training and workshops

For the training of IN2P3 PhD students (and permanent), bi-yearly week-long summer schools have been organised, notably School of Statistics[2] (particle physics-oriented) and AstroInfo[3] (cosmology/astroparticle oriented). In both cases, AI was not the only topic but an important one. In both cases, the recent sessions have been recorded, and all material (slides and tutorials) remain available for future reference. Nevertheless, developing a database of self-tutorials specifically oriented toward the direct application in everyday IN2P3-like physics problems (e.g. not generic AI tutorials, which are numerous) would be very beneficial.

A week-long school[4], specifically designed for IN2P3 engineers (ANF Machine Learning pour informaticiens de l'IN2P3 et de l'IRFU), was organised in September 2020 and was a huge success for kickstarting interaction and collaboration between IN2P3 computer scientists on ML projects.

National (IN2P3 and CEA) AI and HEP workshops are held yearly. The last one (2 days) remotely in March 2021[5] and the next one (3 days) in person in Sep 2022. Their main purpose is the exchange of experience tackling HEP problems with AI. With the growing scope and number of participants, dedicated (typically one-day) topical workshops will be organised in the future, for example, AI and accelerator physics.

[2] Dernière instance en mai 2022 à Carry-le-Rouet https://www.in2p3.cnrs.fr/fr/evenement/school-statistics-sos-2022
[3] Dernière instance en décembre 2021 à Barcelonnette https://astroinfo2021.sciencesconf.org
[4] https://gitlab.in2p3.fr/ri3/ecole-info/2020/anf-machine-learning
[5] https://indico.in2p3.fr/event/22938/

IN2P3 researchers have organised international workshops in France, for example, recently Learning To Discover in April 2022 at Institut Pascal Paris-Saclay by IJClab[6] and Bayesian Deep Learning workshop for Cosmology and Time Domain Astrophysics in March 2020 and June 2022 at APC[7].

International data science competitions on well-known platforms like Kaggle attract data scientists' attention on HEP problems, like TrackML[8] in 2018-2019 on particle tracking at the LHC or PlastiCC[9] in 2018 on Supernova light curve classification. More generally, releasing well-documented data with a clear figure of merit is beneficial for the development and benchmarking of innovative algorithms and to ease the collaboration with Computer Scientists.

The IN2P3 poll on the training and networking actions has revealed that the ones deemed most useful were (in this order): 1) online tutorials with HEP-specific data set, 2) a website cataloguing all available resources, 3) several days of in-person workshop, 4) one-day topical discussions in hybrid mode (which was still deemed "very useful" or "almost mandatory" by 40% answers. Online seminars were not favoured, probably given the already available offer. As far as 1) is concerned, there are already many on-line tutorials available worldwide, but they are spread on many web sites. Item 2) should definitely be provided at IN2P3 level, even if, in many cases, it is just a matter of pointing to resources available elsewhere. Item 3) is already regular, and item 4) is just starting and should be amplified. A spontaneous suggestion from a number of participants is to have a pool of experts to whom one could submit specific problems.

4.5 Publications

The role of AI in HEP is (mostly) to improve the sensitivity of measurement published in typical physics journals. Nowadays, one sees and expects many more scientific publications with a paragraph or a section describing the AI technique.

Dedicated algorithmic papers describing at length AI techniques applied to HEP problems as Proof-of-Concept appear in physics journals, as well as dedicated journals like Computing and Software for Big Science[10], Big Data and AI in High Energy Physics[11], Computing parallel sessions (and proceedings) of big physics conferences like ICHEP or EPS-HEP, dedicated conferences like ACAT or Connecting The Dots. Such publications, using a simulation or data from a big experiment, require approval by the said experiment. For faster publication, toy simulations are often used, which requires an additional effort not always

[6] https://indico.ijclab.in2p3.fr/event/5999/
[7] https://indico.in2p3.fr/event/26887/
[8] https://sites.google.com/site/trackmlparticle/
[9] https://www.kaggle.com/c/PLAsTiCC-2018
[10] https://www.springer.com/journal/41781

[11] https://www.frontiersin.org/journals/big-data/sections/big-data-and-ai-in-high-energy-physics

available. A more open-data policy by big experiments would create an easier framework for collaboration with AI involved.

More difficult are publications in "pure" AI venues, which are typically proceedings of major conferences like NeurIPS or ICML. Such publications require close collaboration with Computer Scientists as the article format is rather different from the one of physics or physics and AI journals. We should note that these major conferences are starting to be so huge that they are creating dedicated tracks specifically aimed at interdisciplinary research, such as the ML for astrophysics workshop at ICML 2022[12] or Machine Learning and the Physical Sciences at NeurIPS 2020 and 2021[13].

In general, it appears that IN2P3 engineers and researchers AI contributions could be more visible across experiments/domains in international workshops and conferences. This is difficult to evaluate systematically, but for example, some clues can be obtained for LHC physics by participating in the CERN IML workshops with three and one speakers from IN2P3 in 2021 and 2022, respectively of 50 speakers in each case.

A possible explanation is the lack of incentive for IN2P3 personnel to advertise their developments outside their experiments compared to other countries where this is more the norm for PhD students and post-docs.

5 SWOT

- Strength
  - data science is at the core of IN2P3
  - large labelled datasets
  - accurate simulators
  - relatively easy access to computing resources for training
- Weakness
  - AI competence is not easy to acquire
  - specific semi-structured data not suitable for off-the-shelf tools
  - difficulty to recruit AI-capable post-docs or engineers (salary)
- Opportunities
  - HEP specificities mean opportunity for specific AI developments, potentially interesting for other science
  - HEP « prestige » helps to attract Computer Science collaborators (e.g. Google developers)
- Threat
  - publication pace slow in HEP compared to AI world
  - lack of incentive at IN2P3 for dedicated publication/workshop contribution

[12] https://ml4astro.github.io/icml2022/
[13] https://ml4physicalsciences.github.io/

Conclusion

IN2P3 contributions are diverse, both in terms of application domain and techniques. They are already visible within each experiment or physics domain, and this visibility could be increased through AI workshops and dedicated publications.

However, the trend is positive. In recent years, we have seen that AI has become an important aspect of scientific life for many people. AI (beyond Boosted Decision Tree) has become a hot topic in HEP in 2014 (by then, very few physicists had ever heard of “Machine Learning”). The first generation (a handful) of PhD students with a significant AI contribution in their PhD topic defended their PhD in 2019-2020. The new generation is more numerous (about thirty ongoing PhDs) and benefits from growing local expertise, budding collaborations with Computer Scientists, and relatively easy access to computing resources. In parallel, IN2P3 scientists have successfully bid to ANR and other calls to fund a wide variety of projects, funding postdocs in particular.

Training resources exist, but specificities of AI in HEP are such that “you need to talk to someone”. S/he could be someone more expert on AI techniques (e.g. Graph Neural Network, Differentiable Programming, etc.), but less or not at all on the physics domain. S/he could be a peer trying to use a similar technique on similar data. Structuring and encouraging such exchanges is important.
The recruitment on permanent positions of physicists with deep expertise in AI is to be encouraged, given the important role they would have in spreading AI expertise at IN2P3.