AKPIK 1: Data Integration & Processing

Time: Monday 16:15–18:30 Location: AKPIK-H13

AKPIK 1.1 Mon 16:15 AKPIK-H13

The PUNCH4NFDI Consortium in the NFDI - status, first results and outlook — •Thomas Schörner for the PUNCH4NFDI-Collaboration — Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg

With the "Nationale Forschungsdateninfrastruktur" (NFDI, national research data infrastructure), a massive effort is undertaken in Germany to provide a coherent research data management, to make research data sustainably utilisable and to implement the FAIR data principles. PUNCH4NFDI is the consortium of particle, astro- and astroparticle, as well as hadron and nuclear physics within the NFDI. It aims for a FAIR future of the data management of its community and at harnessing its massive experience not least in "big data" and "open data" for the benefit of "PUNCH" sciences (Particles, Universe, NuClei and Hadrons) as well as for physics in general and the entire NFDI. In this presentation, we will introduce the work programme of PUNCH4NFDI, its connection to everyday work in the physical sciences and beyond, and in particular the idea of digital research products and the PUNCH science data platform.

AKPIK 1.2 Mon 16:30 AKPIK-H13

Community Initiative for a VHE Open Data Format — •Maximilian Nöthe[1] and Lars Mohrmann[2] — [1]Astroparticle Physics WG Elsässer, TU Dortmund University, Germany — [2]Max-Planck-Institut für Kernphysik, Heidelberg, Germany

The operation of the next-generation gamma-ray telescopes as observatories, the wish of currently operating instruments to archive and publish their data in an accessible format, and enabling multi-instrument analyses are strong reasons for developing an open, software independent format for gamma-ray data.

A first attempt of a common specification has been developed by members of different Imaging Atmospheric Cherenkov Telescopes (IACT) within the "Data formats for gamma-ray astronomy" initiative. The current version defines formats for high-level gamma-ray data, including event lists of candidate photons and instrument response functions, serialized as FITS files.

Open-source software for gamma-ray analyses, including gammapy and ctools, have recently developed support for this format and, as a result, a series of publications relying on standardized datasets and software have been issued.

Currently, an effort to formalize the endeavor is underway, creating a Coordination Committee formed from representatives of the participating instruments to steer the future development of the specification.

In this talk, current developments and future plans will be presented, including the already implemented extension to ground-based wide-field experiments and possible extension to other messengers.

AKPIK 1.3 Mon 16:45 AKPIK-H13

From sample management to workflow integration: Semantic research data management with CaosDB — •Daniel Hornung[1], Florian Spreckelsen[1], and Johannes Freitag[2] — [1]IndiScale Gmbh, Göttingen — [2]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven

Organizing data from a diversity of sources, from acquisition to publication, can be a tough challenge. We present research data management implementations using the flexible open-source toolkit CaosDB at the Alfred Wegener Institute. CaosDB is used in a diversity of fields such as turbulence physics, legal research, animal behavior and glaciology. CaosDB links research data, makes it findable and retrievable, and keeps data consistent, even if the data model changes.

In the presented example, CaosDB keeps track of ice core samples and to whom samples are loaned for analyses. It made possible additional features such as: A revision system to track all changes to the data and the sample state at the time of analysis. Automated gathering of information for the publication in FAIR-DO meta-data repositories, e.g. Pangaea. Tools for storing, displaying and querying geospatial information and graphical summaries of all analyses performed on each ice core. Automatic data extraction and refinement into data records in CaosDB to minimize manual users interaction. A state machine which guarantees certain workflows, simplifies development and can be extended to trigger additional actions upon transitions.

We demonstrate how CaosDB simplifies semantic data in science and enables advanced data processing and understanding.

AKPIK 1.4 Mon 17:00 AKPIK-H13

CaosDB – a scientific research data management toolkit — •Daniel Hornung[1], Florian Spreckelsen[1], and Johannes Freitag[2] — [1]IndiScale Gmbh, Göttingen — [2]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven

Processing interconnected, multi-modal data poses a challenge in many fields, especially when the data model, i.e. the way how data is organized, changes over time or when its structure is poorly documented. The open-source software **CaosDB** is a toolkit for research data management which was originally developed at the Max Planck Institute for Dynamics and Self-Organization (Göttingen) because existing software could not fulfill the needs of the scientists.

We present examples where CaosDB helped make data FAIR (Findable, Accessible, Interoperable, Retrievable) and how it can simplify the workflows for researchers: Automated data collection and integration, export to data repositories, API libraries for third-party programs, integrated revisioning and workflow state machines. If the data model needs to change, existing data can remain as-is and future search queries will return matching results containing "old" and "new" data. We demonstrate how raw and processed data, analysis settings and results, and even labnotebooks and publications can be linked against each other, to improve long-term usability of data and reproducibility of results.

AKPIK 1.5 Mon 17:15 AKPIK-H13

Data curation in astroparticle physics data centers on example of KCDC and GRADLCI — •Victoria Tokareva[1], Andreas Haungs[1], Doris Wochele[1], Jürgen Wochele[1], Frank Polgart[1], Alexander Kryukov[2], Minh-Duc Nguyen[2], Andrey Mikhailov[3], and Alexey Shigarov[3] — [1]Karlsruhe Institute of Technology, IAP, 76021 Karlsruhe, Germany — [2]Moscow State University, SINP, Moscow 119991, Russia — [3]Matrosov Institute for System Dynamics and Control Theory, Irkutsk 664033, Russia

The KASCADE Cosmic Ray Data Center (KCDC), introduced in 2013, is a multi-functional public data center for high-energy astroparticle physics. Its distinctive features include use of open standards and technologies, providing materials and service both for professional scientists and a broad outreach audience and furnishing open access to scientific data. The GRADLCI (German-Russian Astroparticle Data Life cycle Initiative), which spawned from KCDC in 2018, proposed an alternative approach to metadata management and utilized the optimized models and algorithms for processing requests. Today, the work on organizing flexible cross-collaboration data sharing is going on in various areas of science within the framework of the EOSC project and others, such as PUNCH4NFDI. A big share of this work includes collection and analysis of the data curation practices in order to reach a more abstract and complex understanding of the challenges of data curation committing into new advanced solutions ready for further extension. In this report the use cases of the KCDC and GRADLCI data centers will be considered.

AKPIK 1.6 Mon 17:30 AKPIK-H13

Optimizing Computer Vision for Radiosource Detection — •Janis Sowa and Kevin Schmidt — Astroparticle Physics AG Elsässer, TU Dortmund University, Germany

Earthbound radio astronomy utilizes interferometric arrays to achieve the highest possible resolution by combining the measurements of multiple telescopes. The resolution then depends on the distance between telescopes as opposed to the diameter of a single dish. Modern improvements in computing performance and telescope design are allowing radio astronomers to collect increasing amounts of data. In sky surveys, information about hundreds of thousands of astronomical sources are obtained. On this scale, a manual analysis is a time-consuming task. Deep Learning-based source detection thus naturally comes to mind as a candidate for identifying these individual objects. In a previous work, a Convolutional Neural Network architecture was shown to be faster but less accurate in comparison to the state-of-the-art source detection tool PyBDSF, when tested on simulated data. This talk will showcase how the existing model can be further improved and fine-tuned for application on real data.

AKPIK 1.7 Mon 17:45 AKPIK-H13

Evaluation of deep learning accelerators for the usage in the cosmic ray simulation CORSIKA˜8 — •Dominik Baack and Jean-Marco Alameddine for the CORSIKA 8-Collaboration — Astroparticle Physics, WG Elsässer, TU Dortmund University, D-44227 Dortmund, Germany

The proliferation of neural networks has led to the acquisition of specialized hardware to accelerate training and application at an increasing number of scientific sites.

To take advantage of this growth, we investigated the extent to which this hardware can be used to accelerate the complex simulation of cosmic particle showers and which parts of the simulation benefit most.

A number of examples based on CORSIKA˜8 are presented to illustrate advantages, disadvantages, and limitations in the choice of methods. In particular, the widely used Nvidia accelerators that was used very successfully for ray tracing of optical photons (e.g. Cherenkov light) will be discussed.

AKPIK 1.8 Mon 18:00 AKPIK-H13

Structured Sparsity for CNNs on Reconfigurable Hardware — •Hendrik Borras, Günther Schindler, and Holger Fröning — Institute of Computer Engineering; Heidelberg University; Heidelberg (Germany)

While Convolutional Neural Networks (CNNs) are gaining crucial importance for various applications, including modern analysis and trigger systems, their memory and compute requirements are increasing steadily and the requirements of many CNNs impose serious challenges for achieving high inference throughput and low latency on edge devices, situated close to an experiment. To improve the performance of CNNs on such resource-constrained devices, model compression through quantization and pruning has been proposed and evaluated as a possible solution in the past. Field Programmable Gate Arrays (FPGAs) are a prime example of low-power devices and suitable for a pervasive deployment. Here, FINN is one of the most widely used frameworks for deploying highly quantized CNN models on edge devices. In this work, we extend FINN for pruning by introducing two methods for column pruning, enabling further compression of CNN-based models. The two techniques vary in their granularity and implementation complexity. The coarse-grain method only prunes blocks of columns, while the fine-grained method is able to prune single columns. Both approaches are then evaluated on the CIFAR10 image classification task. We demonstrate significant throughput improvements of on average 83% while keeping the accuracy degradation within reasonable bounds (4.2%) at 50% sparsity.

AKPIK 1.9 Mon 18:15 AKPIK-H13

IEA-GAN: Intra-Event Aware GAN for the Fast Simulation of PXD Background at Belle II — •Hosein Hashemi, Nikolai Hartmann, Thomas Kuhr, and Martin Ritter — Faculty of Physics, Ludwig Maximilians University of Munich, Germany

The pixel vertex detector (PXD) is the newest and the most sensitive sub-detector at the Belle˜II. Data from the PXD and other sensors allow us to reconstruct particle tracks and decay vertices. The effect of background processes on track reconstruction is simulated by adding measured or simulated background hit patterns to the hits produced by simulated signal particles that originate from the processes of interest. This model requires a large set of statistically independent PXD background noise samples to avoid a systematic bias of reconstructed tracks. However, the fine-grained PXD data requires a substantial amount of storage. As an efficient way of producing background information for fast simulation, we introduce the idea of an on-demand PXD background generator with Intra-Event Aware GAN (IEA-GAN), conditioned over the number of PXD sensors in order to produce sensor-dependent PXD images by approximating the concept of an "event" in the detector as these PXD images share both semantic and statistical features that makes it extremely hard for even the State of the Art GANs to mimic these exact properties. As a result, we developed the IEA-GAN model which captures these dependencies by imposing relational inductive bias over the batch dimension.