PUNCH4NFDI Task Area 3

Methods for maximizing the scientific value of data

Marcus Brüggen, Gregor Kasieczka, Thomas Kuhr, Joe Mohr

punch4nfdi-ta3@desy.de

Work Package Categories

Statistical methods → Kevin, Joe

- Fits of complex models with many parameters to huge datasets in a resource efficient way
- Extraction of tiny signals in the presence of large backgrounds
- ➔ Examples of expertise: Bayesian Analysis Toolkit, Origins Data Science Lab

Numerical methods and simulations → Susanne, Frithjof

- Solvers for large matrices
- Optimizations for specific hardware architectures
- ➔ Examples of expertise: Lattice QCD experience with optimised multi-grid and deflated matrix solver for new exascale hardware and performance portability across many architecture

Work Package Categories

Machine learning methods → Gregor, Marcus

- Automated optimization of model hyperparameters for a large variety of problems
- Federated learning on very large (petabyte), partitioned datasets
- Interpretation, reliability and visualization of machine learning results

➔ Examples of expertise: Inter-Experimental LHC Machine Learning Working Group, Coordination of ML exercise in CMS experiments, Platform for Challenges in Data Science

Methods for analyses across datasets → Joe, Thomas

- Interface for extracting physical properties or likelihoods from datasets
- Combination of likelihoods and comparison with theoretical predictions

➔ Examples of expertise: Fittino, Heavy Flavor Averaging Group, Multiwavelength astro/cosmo analyses (DES, SPT, eROSITA, MeerKAT)

Draft for LoI

- https://docs.google.com/document/d/1jcQ7Ak412VtaXDhuZrfKjujLj7xQLk_zkKHry3at5GU/edit
- Data have scientific value if they lead to new insights. Algorithms play a crucial role here, and indeed are an established and integral part of data-driven science. As with the construction of new instruments or devices, the development of new algorithms is a key element of progress in research. We focus on methods common to many astro, particle, and hadron and nuclear physics data analyses which have clear application outside their original domain. Those methods that could be applied even beyond PUNCH are the primary emphasis of TA3.

  One category of generally applicable methods is statistical tools. Fitting complex models with many parameters to huge datasets in a resource efficient way is a technical challenge. Numerical methods and simulations, especially on large, heterogeneous compute grids, are based on another category of algorithms. A particular issue here is the optimization of algorithms for specific hardware architectures. While many people work on Machine Learning methods, in our program we focus on the issue of training on very large, partitioned datasets. Furthermore, we plan to exploit the variety of problems in our science fields to develop robust methods for the automated design and optimization of machine learning models. The last category of methods we address in TA3 is the joint analysis across multiple, large datasets. These methods allow the scientists to exploit the full potential of the data by combining information from different sources.

  All methods developed in TA3 will be made available to the scientific community as plugins on the Science Platform.