Karlsruhe Institute of Technology

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

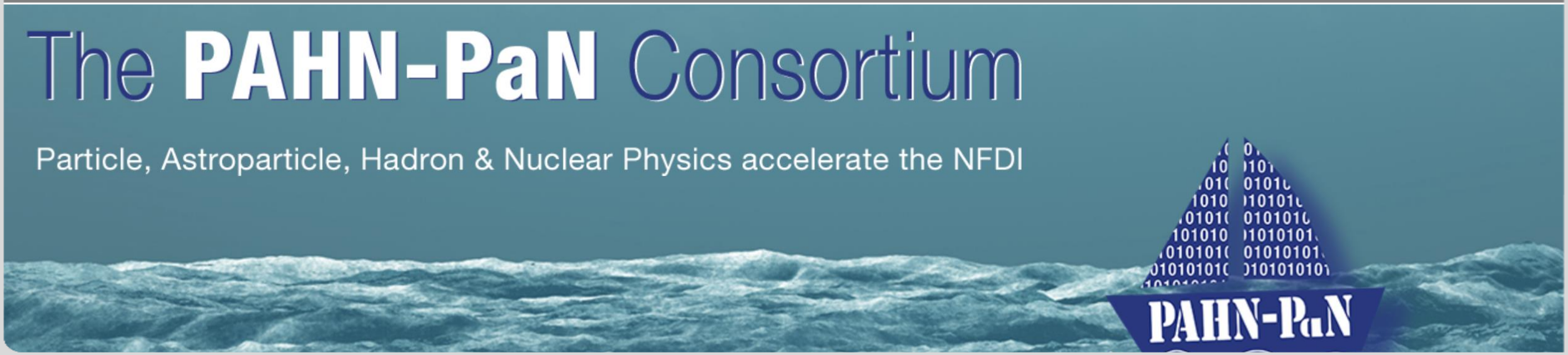
HELMHOLTZ GEMEINSCHAFT
Allianz für Astroteilchenphysik

KAT. Komitee für Astro.Teilchen.Physik

The National Research Data Infrastructure – NFDI

HAP / AKPIK workshop | Big Data Science in Astroparticle Physics
Aachen, 17-19 February 2020

Andreas Haungs



NFDI: The Goal and Realization

https://www.dfg.de/foerderung/programme/nfdi/

Dr. Katerbow (DFG)

- **Objective:** Data stocks of science and research are to be systematically indexed, sustainably secured and made accessible as well as (inter-) nationally networked.
- Organized by **NFDI Consortia**
- The **DFG** conducts the procedure for **reviewing and evaluating NFDI consortia**.
- **Funding recommendations** are made by the NFDI Expert Committee of the DFG.
- **Financial volume** of around 85 million euros per year to support the consortia
- Directorate in Karlsruhe (KIT and FIZ)
- Funding of approximately 30 consortia from three rounds of calls
- Total project funding for 10 years; decision on further design and financing of the NFDI in 2026
- Funding of consortia by 2-5 million € / year; initial period 5 years

NFDI: Requirements to a succesful consortium

https://www.dfg.de/foerderung/programme/nfdi/

Prof. Gehring

Rat für Informations Infrastrukturen

1. What consortia and NFDI as a whole should achieve

- Comprehensive research data management and increased efficiency throughout the scientific system
- Linking of research-oriented data services, improving interoperability
- Accepted, standardised processes and procedures in line with methodological requirements of (very) different disciplines
- A common voice for data concerns in the science-policy arena

But not

- Merely accumulate „Data“
- Collect local solutions (or repositories) waiting for future users
- On-size-fits-all
- Overly strict reglementation („juridification“)

NFDI: Requirements to a succesful consortium

https://www.dfg.de/foerderung/programme/nfdi/

Prof. Gehring

Rat für Informations Infrastrukturen

2. Role of consortia in NFDI & NFDI Consortia Assembly

- Help building the NFDI as a whole
  - Control question: What is the added value a consortium brings to the overall structure?
- Creating a common knowledge base and organising horizontal structures between the consortia
- Agreeing on common elements and standards for a federated data landscape in Germany
- Contributing and sharing IT services as well as common concepts for training, consulting, software maintenance
- Providing gateways to international networks

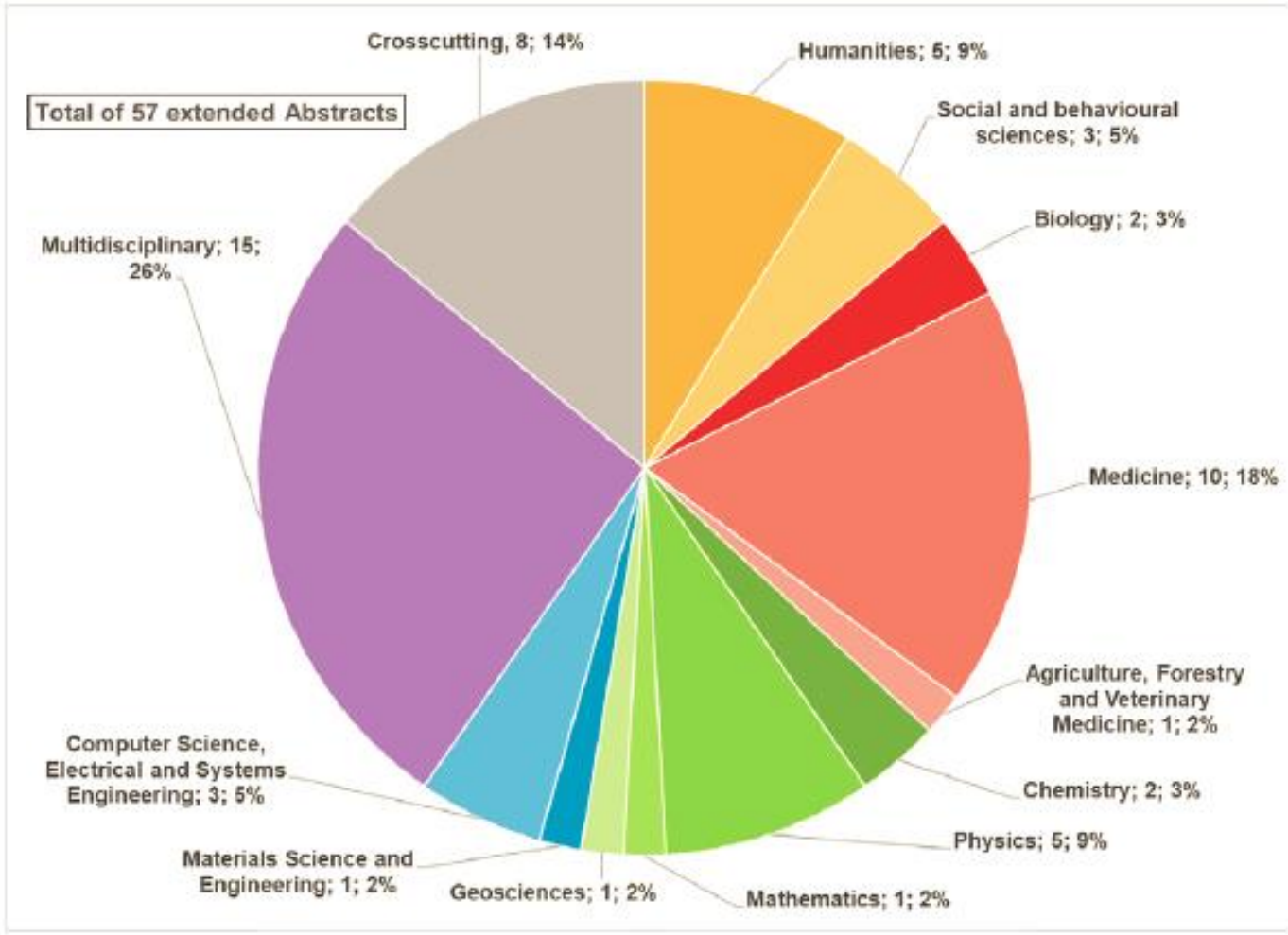
But not

- Outsourcing to additional service entities which apply separately for NFDI-Funding
- Meta-Consortia (i.e. „small NFDI's" within NFDI)
- Debate Clubs waiting for top-down initiatives

NFDI: Received abstracts for consortia

https://www.dfg.de/foerderung/programme/nfdi/

Dr. Katerbow (DFG)



➔ **Indeed covering nearly all scientific research fields**

NFDI: Timeline

https://www.dfg.de/foerderung/programme/nfdi/

Dr. Katerbow (DFG)

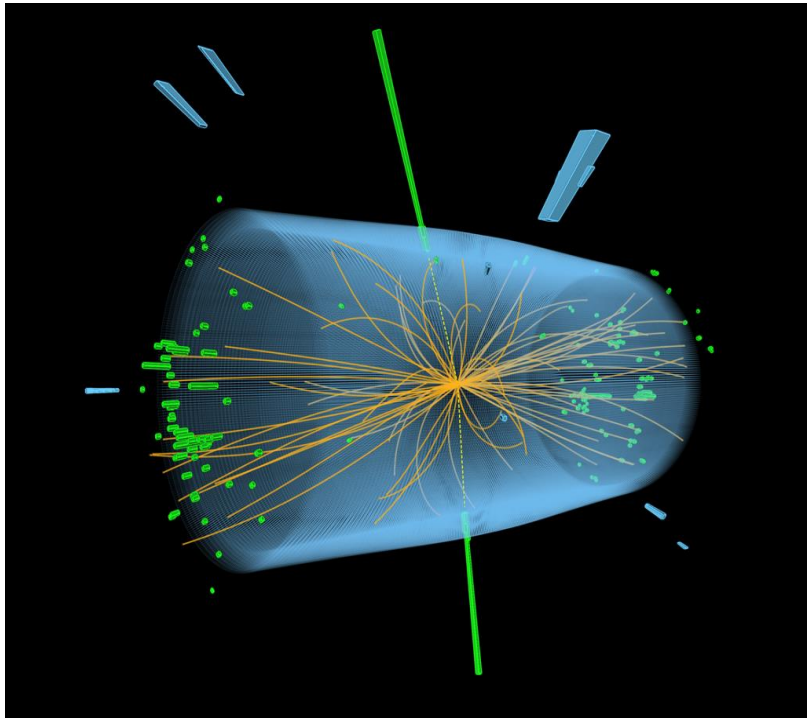
29 March 2019	Deadline submission of Extended Abstracts for the NFDI Conference
13-14 May 2019	NFDI Conference
June 2019	Publication of first call for consortia
July 2019	Deadline for submission of LOI for 2019 proposals (57 LoI)
October 2019	Deadline for submission of proposals
Nov 2019 – Jan 2020	Review of the proposals (22 in first round)
February 2020	Communication of review results with consortia with possibility of statements
April 2020	Discussion of board of experts with funding recommendations to the GWK (decides!)
June 2020	Communication of decisions (GWK)

➔ **Period of waiting and hoping**

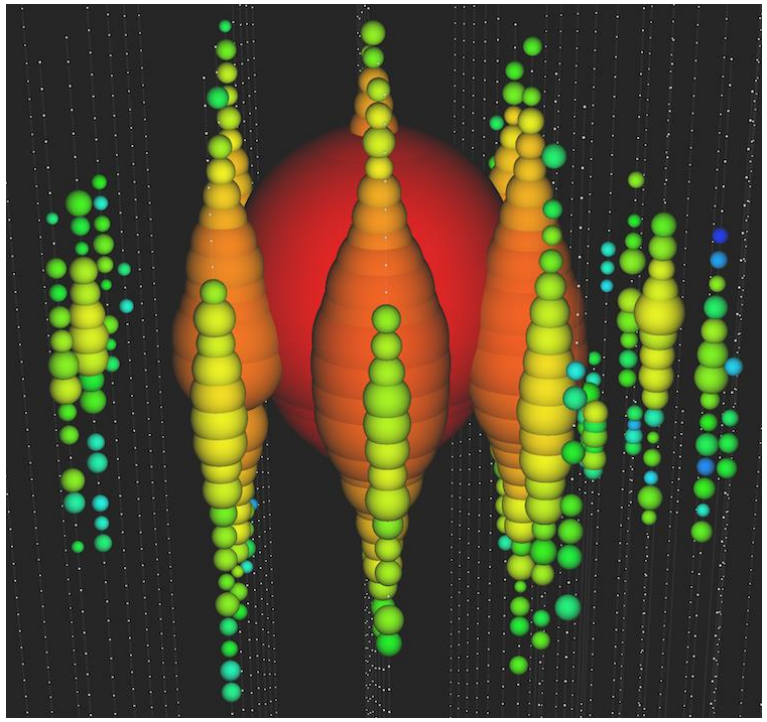
PAHN-PaN: Introduction

Particle, Astroparticle, Hadron & Nuclear Physics

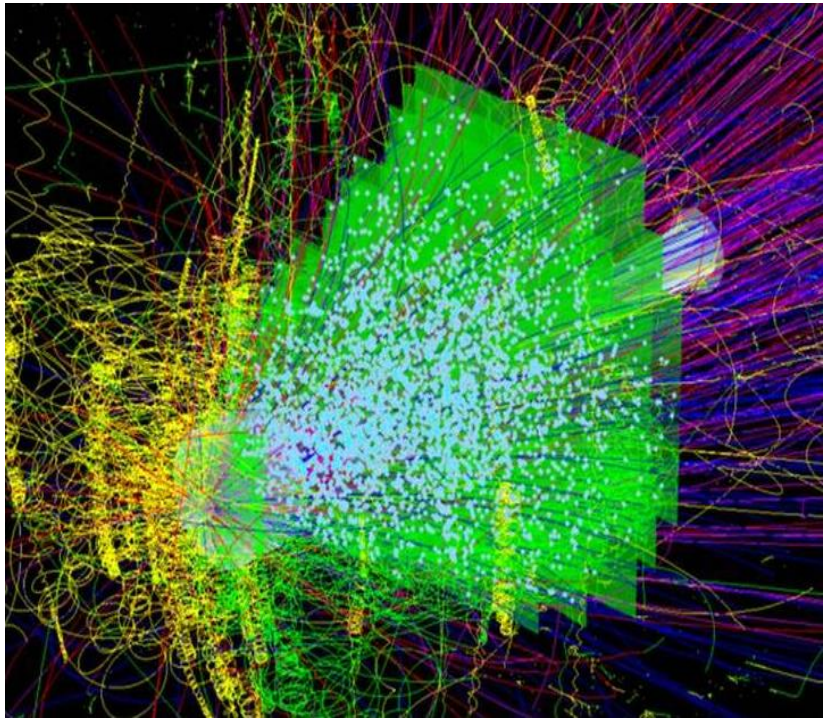
The three PAHN-PaN communities: science



Particle physics
Visualisation of a proton-proton collision in the LHC



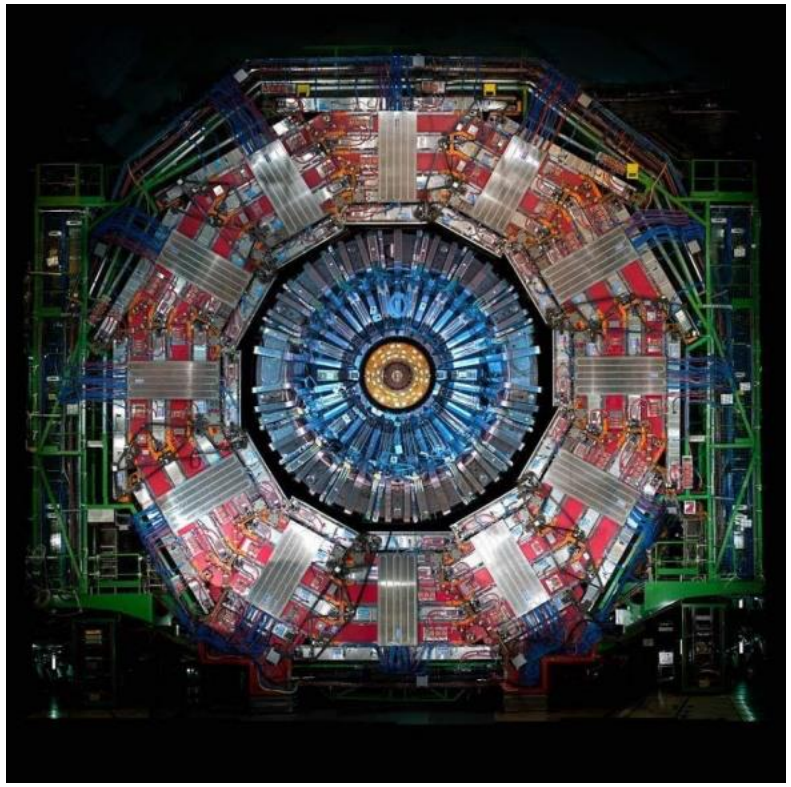
Astroparticle physics
Visualisation of a neutrino event in IceCube



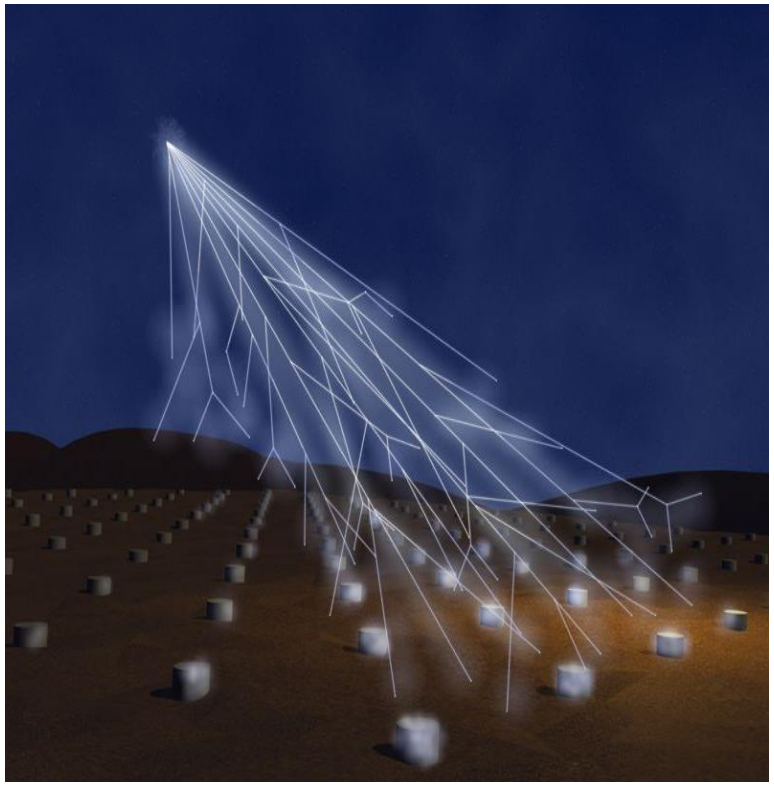
Hadron&nuclear physics
Simulated collision in the CBM experiment at FAIR

Particle, Astroparticle, Hadron & Nuclear Physics

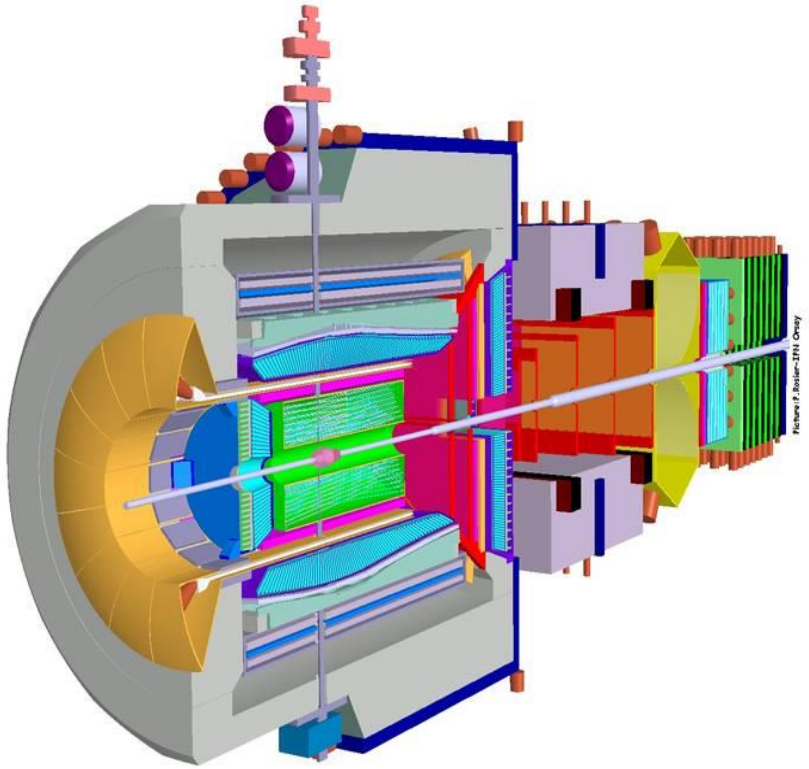
The three PAHN-PaN communities: data providers



Particle physics
The CMS experiment at CERN‘s LHC



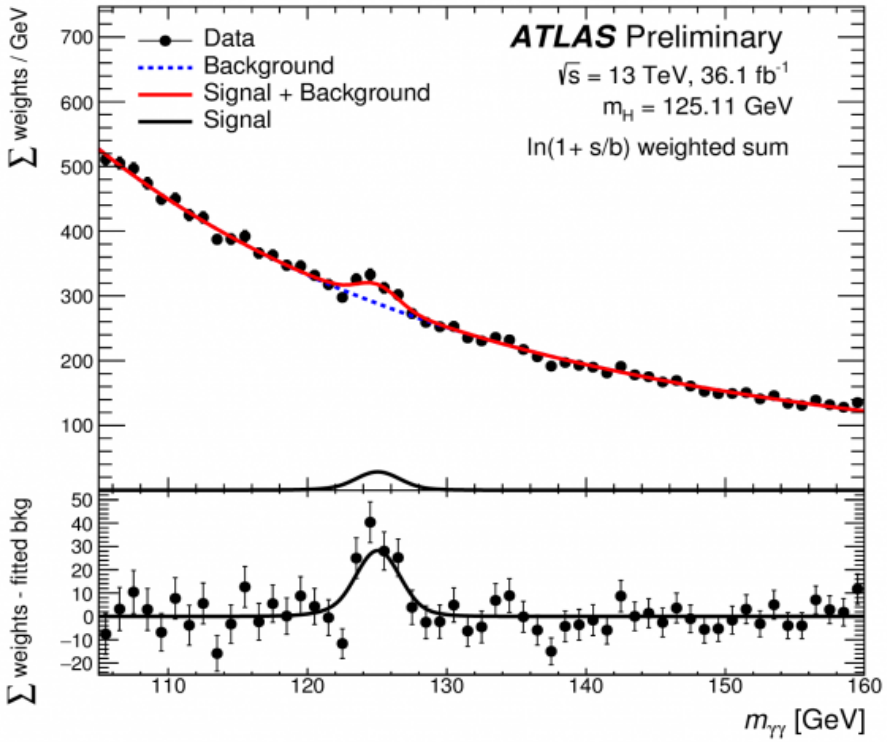
Astroparticle physics
Air shower over the Auger experiment



Hadron & nuclear physics
The PANDA experiment at FAIR

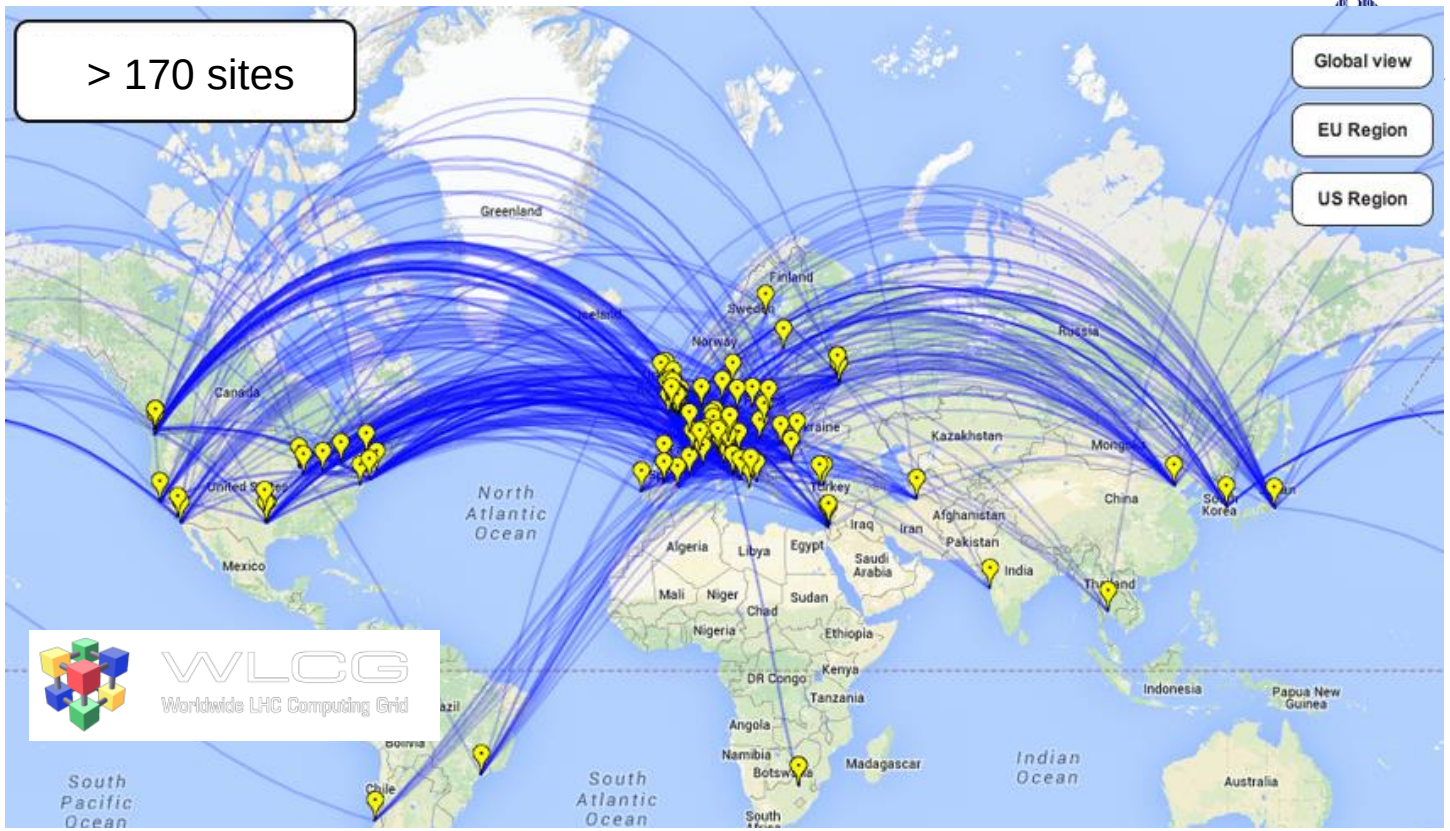
Data Management Today

Example LHC and WLCG



Necessary for analysis by O(10000 users):

- 1 PB/s non-zero-suppressed raw input data / experiment
- > 500.000 grid jobs running at any time
- Example ATLAS storage: 223+315 PB



NFDI Nationale Forschungsdateninfrastruktur

EUROPEAN OPEN SCIENCE CLOUD

Close connection to national and international bodies

- PAHN-PaN as one interface between different levels
- Involvement in all relevant decision structures

PAHN-PaN: The Mission

Particle, astroparticle, and hadron & nuclear physics

- Decade-long experience in operating self-developed global big data management infrastructure (WLCG, the world‘s largest grid).

PAHN-PaN goals

- Innovative, industry-standard solutions for *FAIR* Exabyte data management and scientific services.
- Foster data management at small sites: accelerators like MAMI, S-DALINAC, on-site experiments like KATRIN, theory efforts, and elsewhere

Synergies, solutions and services

- Using NFDI synergies for common developments
- Knowlegde and technology transfer to entire NFDI.
- Accessible to PAHN-PaN and the entire NFDI.

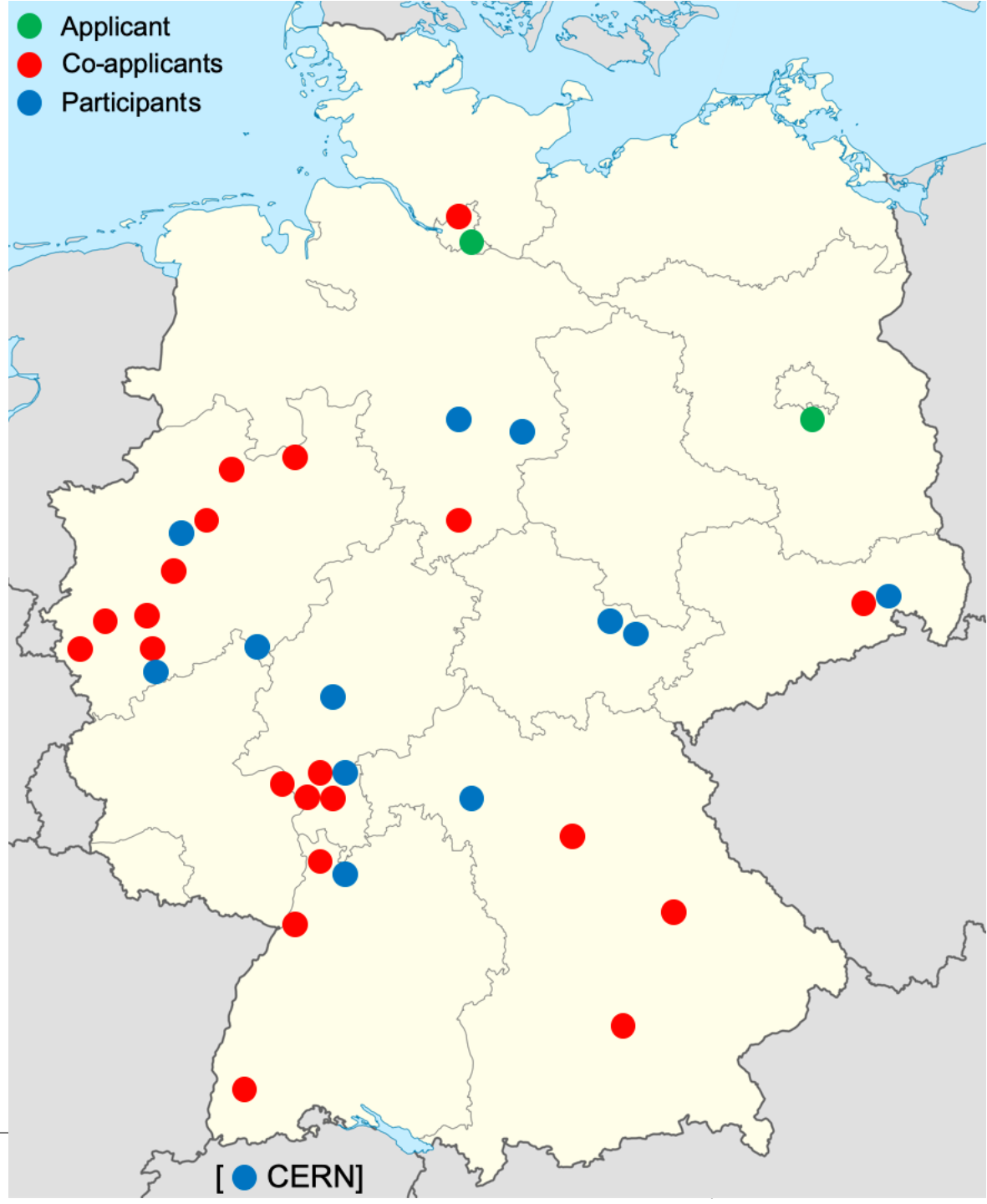
Unprecented scale & complexity of data!

But: facing massive challenges: volumes, rates – order of magnitude in next decade: HL-LHC upgrade, CTA, FAIR facility etc.!

Solutions for

- **the entire data lifecycle (including data publication and archiving) and**
- **long-term re-usability of data.**

PAHN-PaN Institutions



Our Communities

KET Komitee für Elementarteilchenphysik

PHYSICS AT THE TERA SCALE Helmholtz Alliance

ATLAS EXPERIMENT

KHuK

EMMI

ALICE

CMS

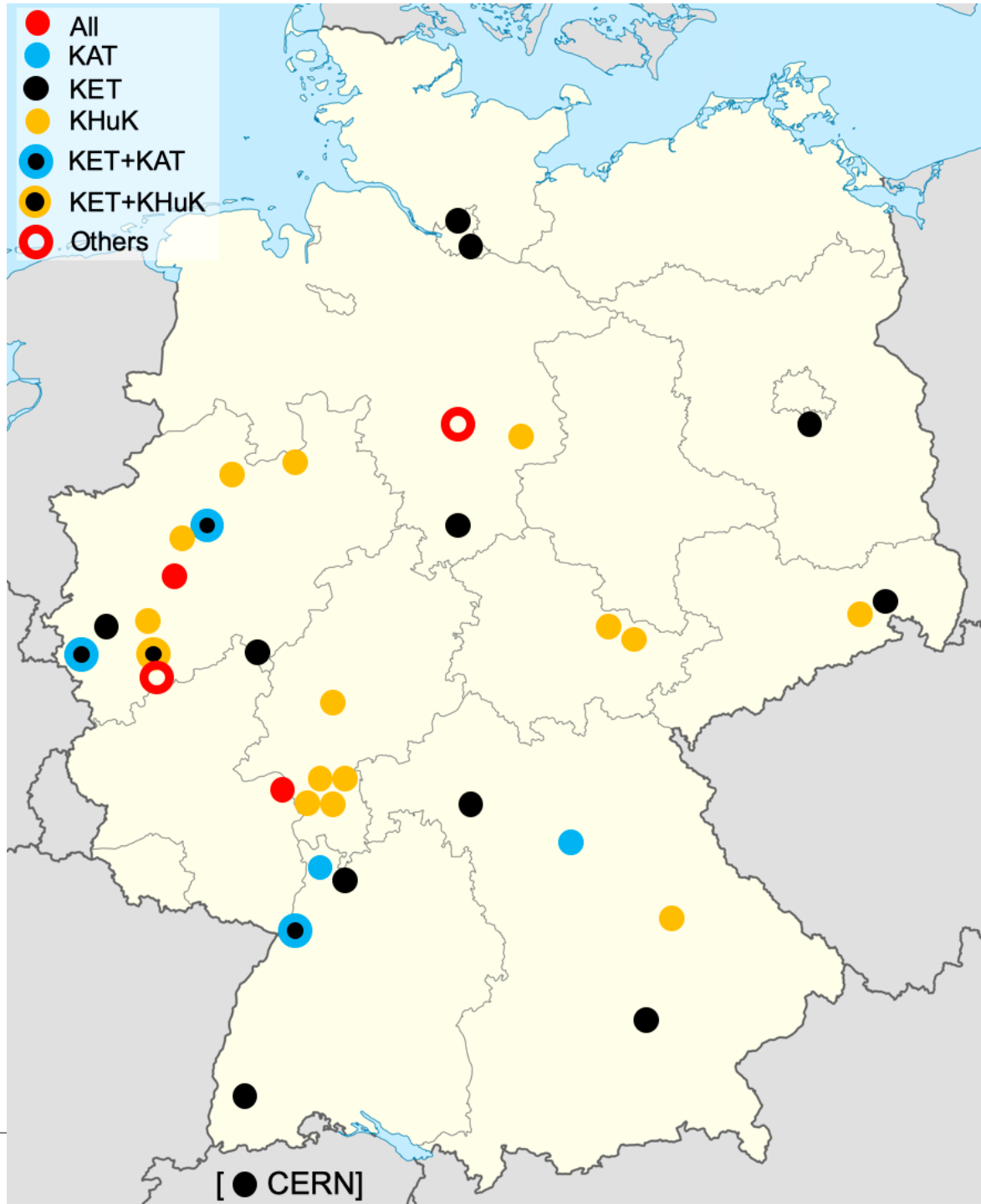
KAT. Komitee für Astro.Teilchen.Physik

Belle II

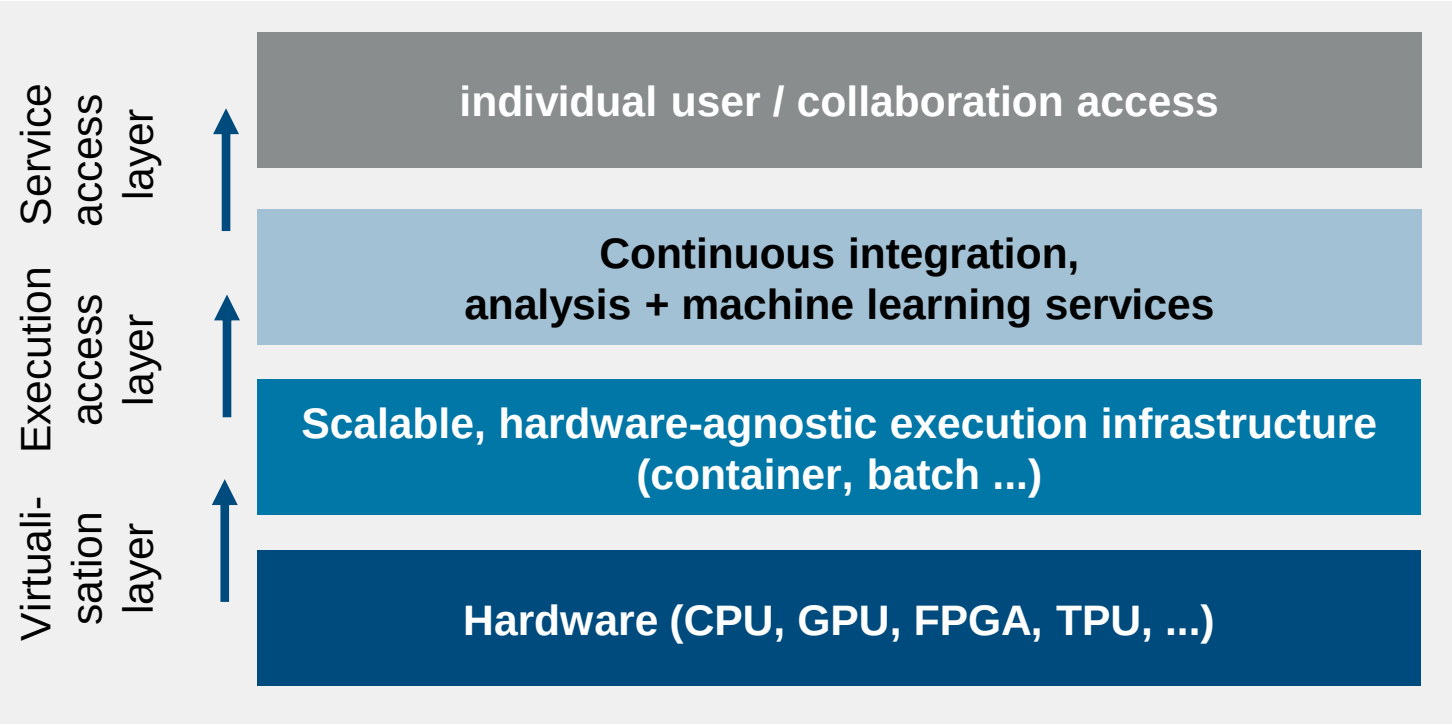
HELMHOLTZ GEMEINSCHAFT Allianz für Astroteilchenphysik

USQCD

Community-wide national and international structures give support to PAHN-PaN



The structure of PAHN-PaN



Task area 1: **Developing workflows and tools for data management**
Task area 2: **FAIR data lifecycle concepts and open data**
Task area 3: **Data analysis procedures and services**
Task area 4: **Real-time data analysis and selection**

Cross-cutting topic A: **Synergies**
Cross-cutting topic B: **Services**
Cross-cutting topic C: **Professional training, education, and outreach**

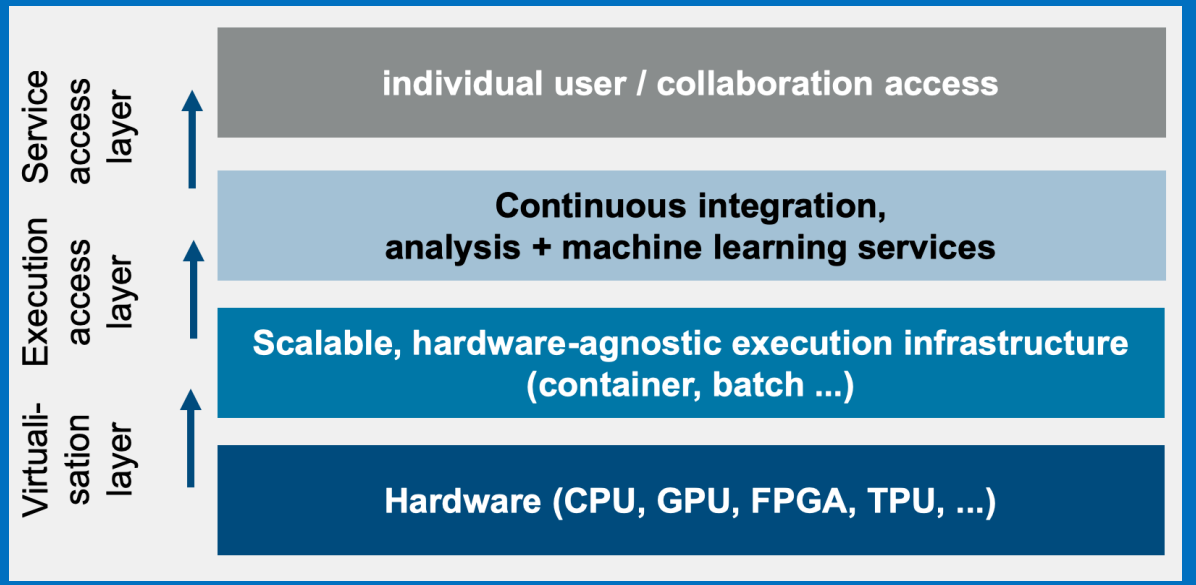
Layered model: scalability and easy replacement of modules!

For the next 10 years: implement and use generic interfaces – irrespective of hardware.

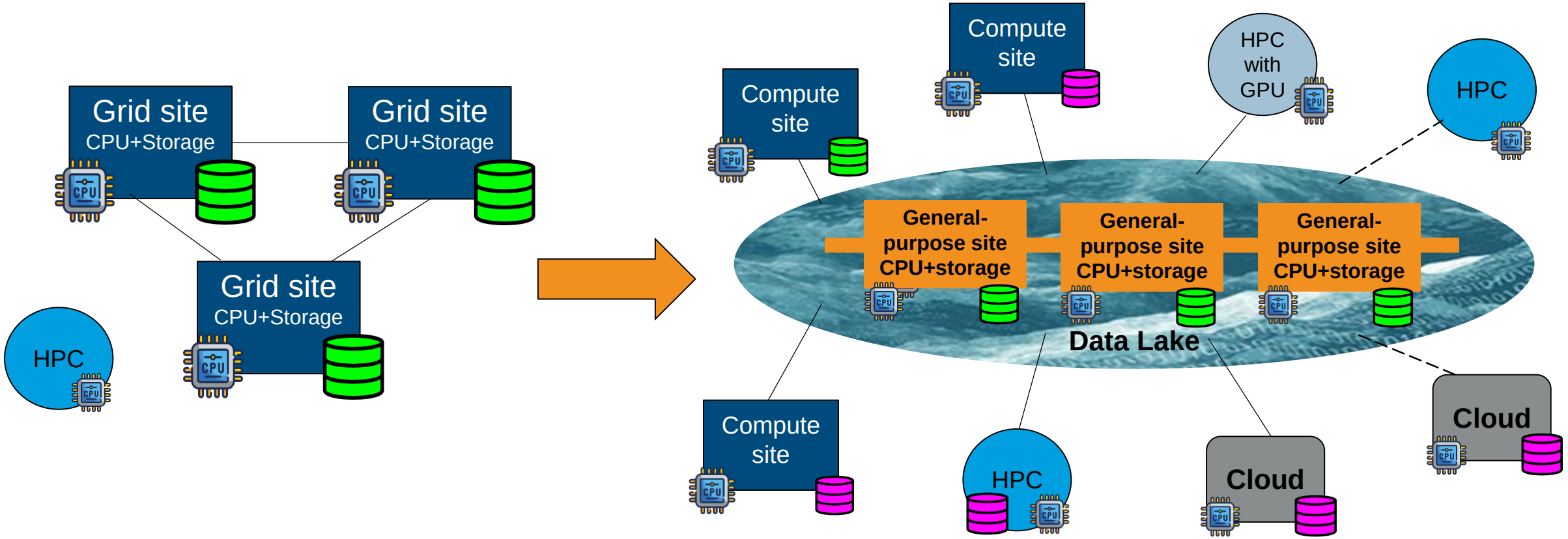
Adaption + further development of existing open source cloud middleware

Task Area 1
“Developing Workflows and Tools for Data Management”

Facilitate the transition from a static grid environment to a dynamic and distributed environment



TA1: Developing Workflows and Tools for Data Management



TODAY

- >170 dedicated grid sites
- Based on high-throughput computing (HTC) architectures
- Connected via dedicated networks
- Data storage at the sites

FUTURE

- Globally distributed data lakes with remote access
- Additional compute resources at clouds and high-performance computing (HPC) centres
- More complex storage architecture (cache)

TA1: Developing Workflows and Tools for Data Management

Example for work packages and deliverables in this task area:

Access to highly distributed federated storage

HEP experiments get towards Exabyte scale, demand for multi-Petabyte storage solutions

- optimal capacity/performance/data security per invest
- aim for low operations effort

Approach well aligned with international efforts; applicable for other storage-intense sciences

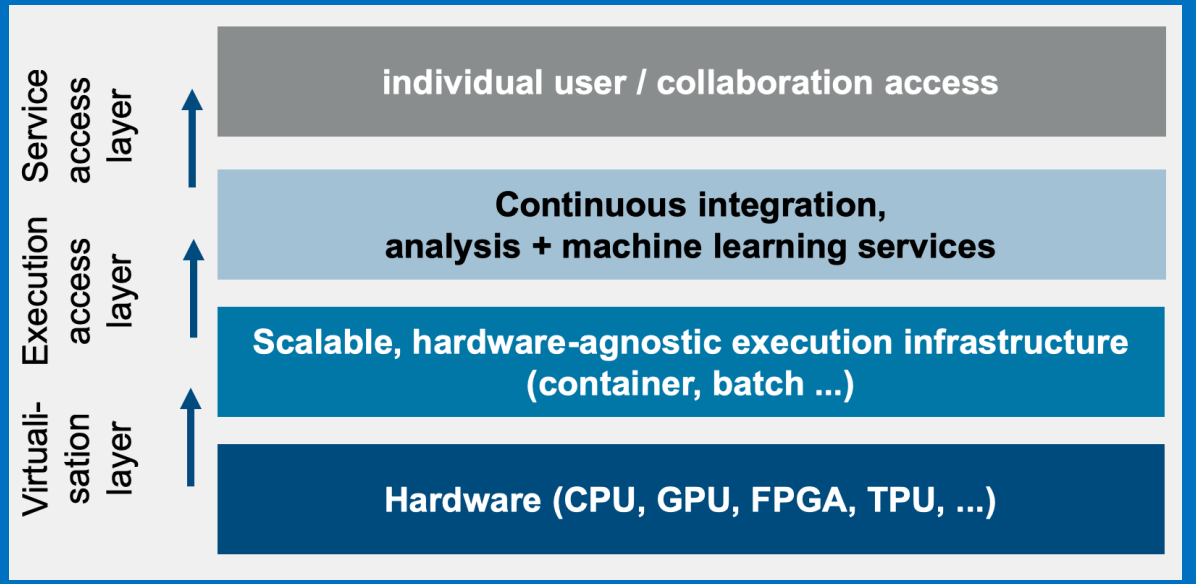
Example for services and synergies:

Data Lakes

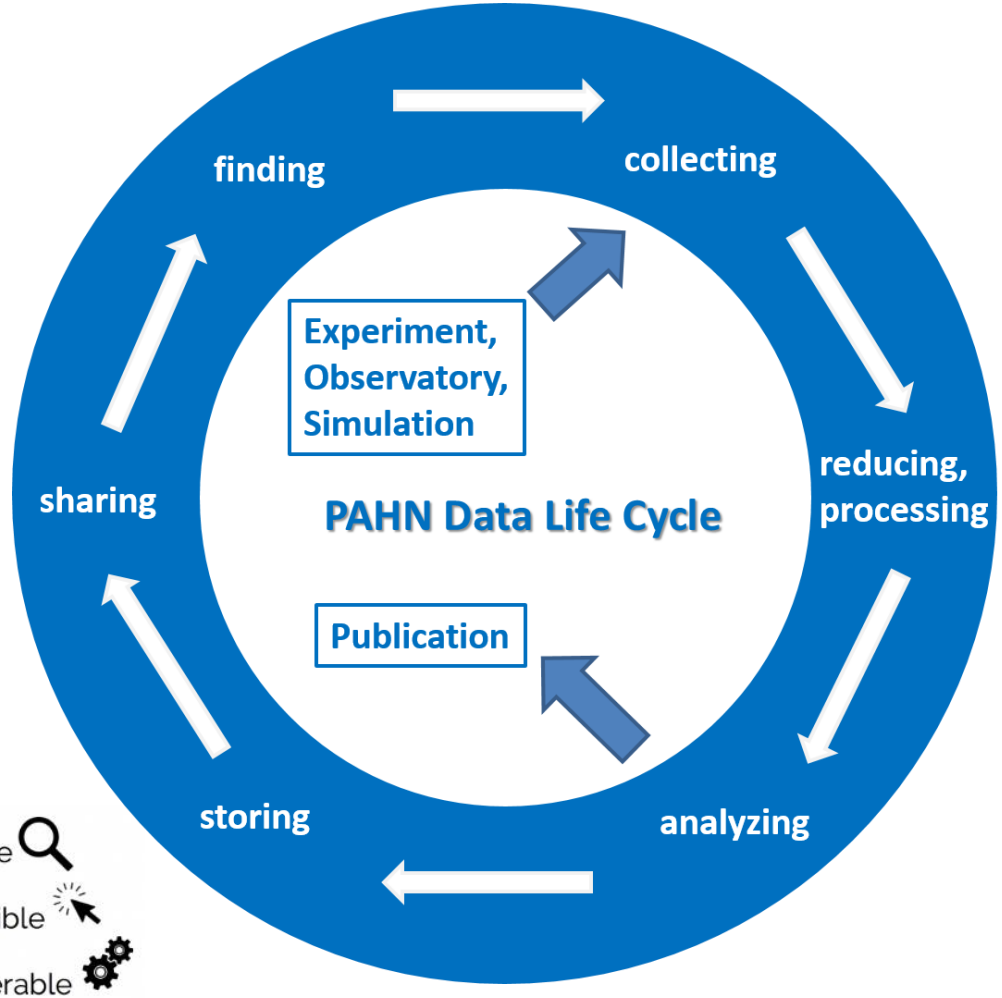
Make the concept of data lakes usable for a broader scientific community!

Task Area 2
“*FAIR* data lifecycle concepts and open data”

Foster/build/provide concepts and demonstrators (infrastructure) for a PAHN data management plan



TA2: *FAIR* Data Lifecycle Concepts and Open Data



Findable
Accessible
Interoperable
Reusable

Where possible, establish common standards to foster interoperability

Importance of "data stewards" as data lifecycle managers and metadata curators

The lifecycle has to provide a FAIR environment for

(i) data availability
(ii) method development
(iii) data analysis
(iv) big data education
(v) open access
(vi) data archiving
(vii) data mining

- **Each arrow requires *FAIR* data management**
- **Each step needs appropriate metadata**
- **The cycle includes data, metadata and workflows**

TA2: *FAIR* Data Lifecycle Concepts and Open Data

Example for work packages and deliverables in this task area:

Public Workflows

To use open data, the complex workflows with which the results are achieved must also be made available.

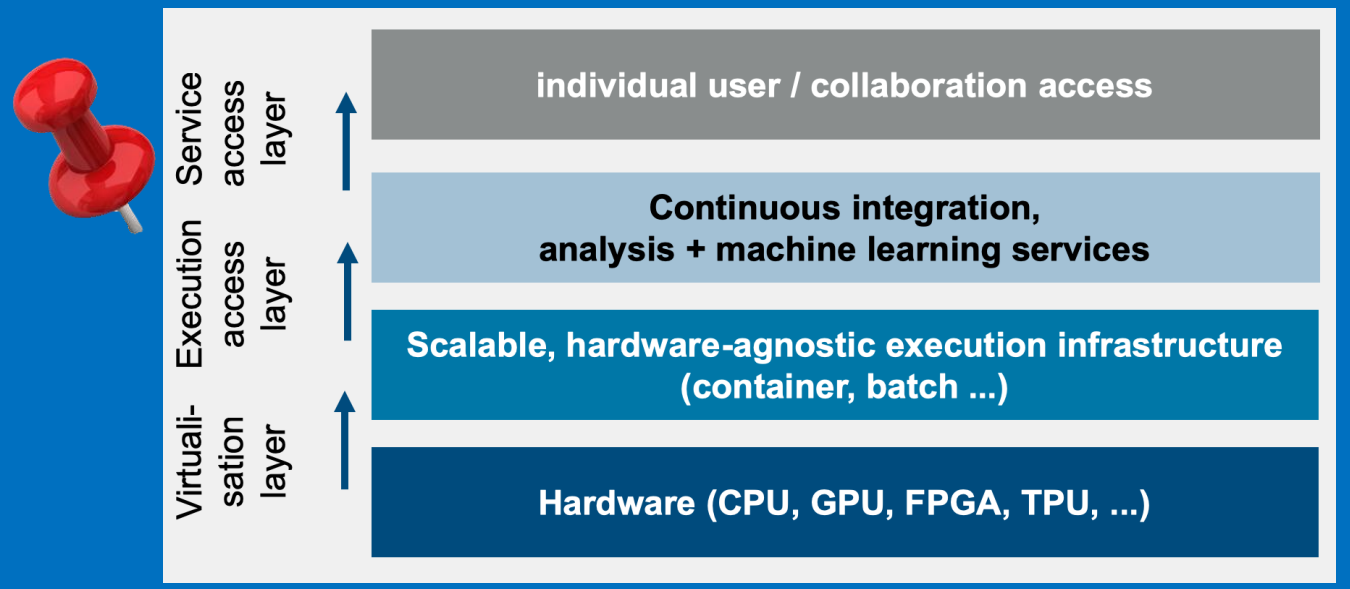
Example for services and synergies:

Metadata

NFDI-wide concept on metadata definition

Task Area 3
“Data Analysis Procedures and Services”

Data provided by an infrastructure only have scientific value if services and tools exist to analyse them



TA3: Data Analysis Procedures and Services

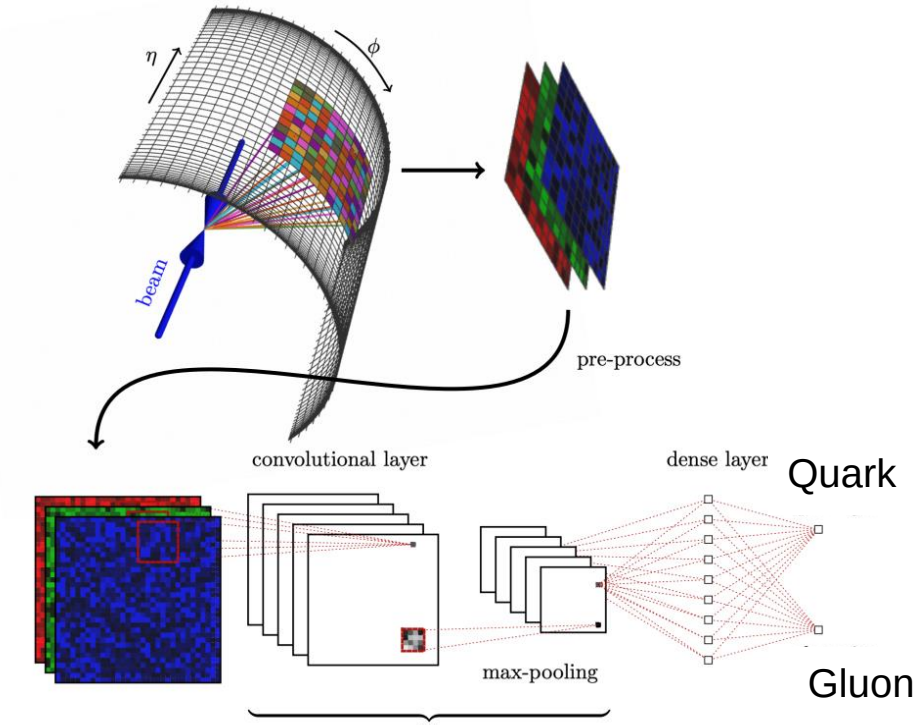
Maximise scientific output via:

- **Fast interactive workflows for large datasets**
- **Efficient inference methods for large datasets**
- **Automated training of machine learning methods (AutoML)**
- **Numerical method and hardware-specific optimisations**

➔ **Combine into a simple web-based service accelerating analysis for PAHN-PaN research and other domains**

➔ **Experience of PAHN-PaN community enables efficient and generic tools**

Neural network for particle identification



from DOI:10.1007/JHEP01(2017)110

TA3: Data Analysis Procedures and Services

Example for work packages and deliverables in this task area:

Automated Machine Learning

Build a tool to automatically build decision algorithms on scientific data

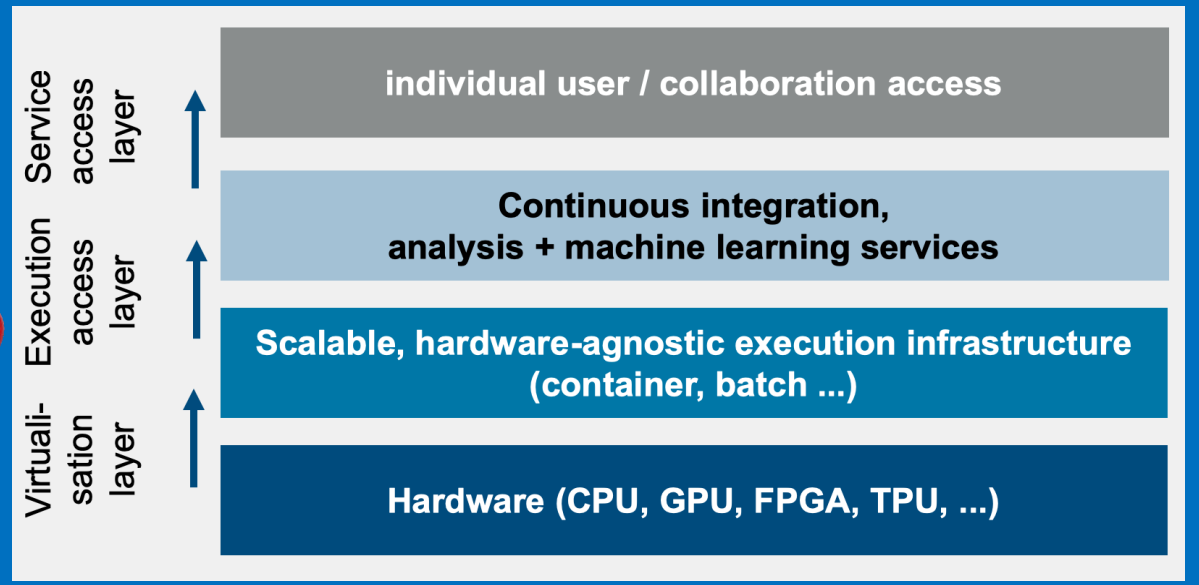
Example for services and synergies:

Service Tools

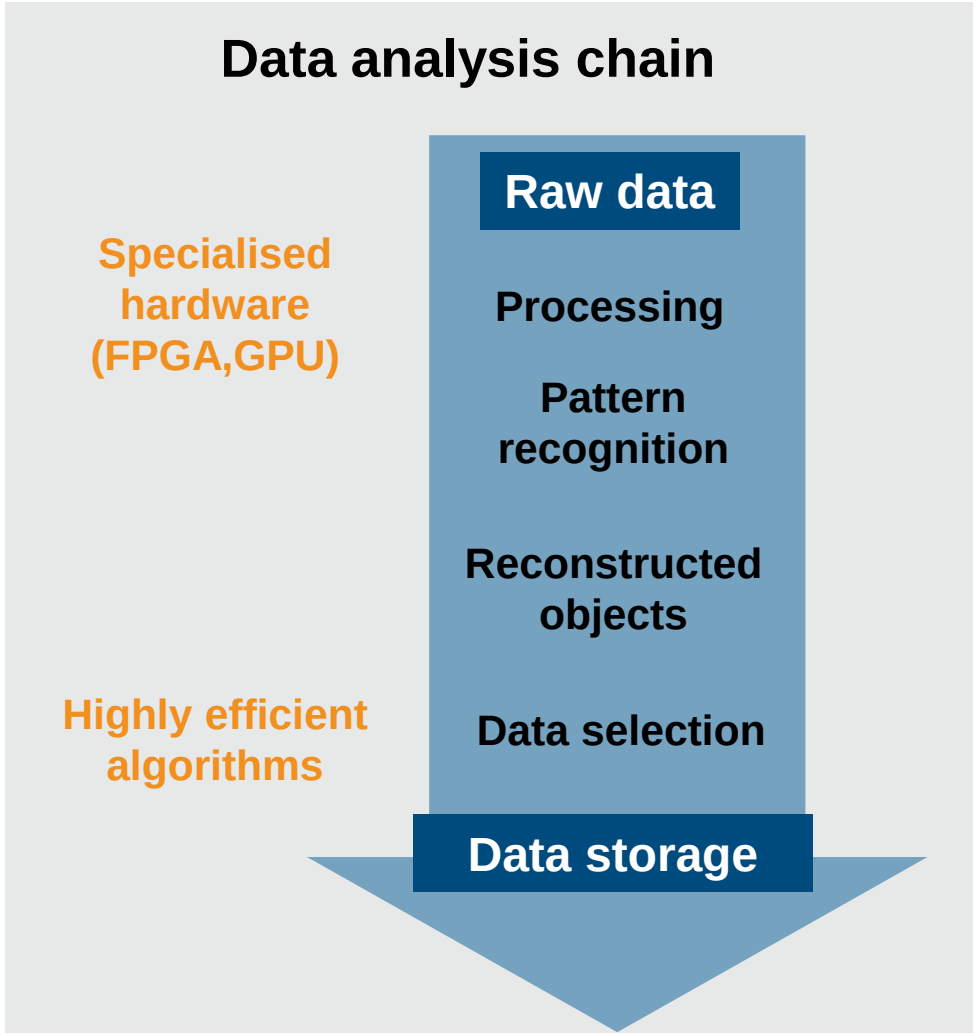
Due to the diverse nature of scientific data in the PAHN domain, developed analysis services will be generic and robust for wider use

Task Area 4
“Real-Time Data Analysis and Reduction”

Reproducible fast selection and storage of massive amounts of data



TA4: Real-Time Data Analysis and Selection

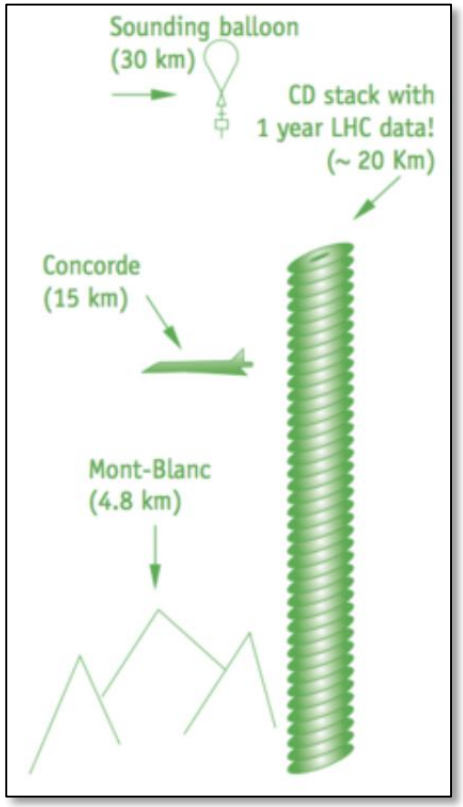


Huge data reduction:

- Irreversibility
- Noise suppression
- Reproducibility
- Optimised with respect to resources and physics reach

Current status:

- LHC data reduction: ~ factor 10^6
- Special solutions for each experiment
- Extra challenges from higher rates and more complex signatures
- Development of efficient software triggers



Experience from PAHN experiments: ALICE, ATLAS, Auger, Belle II, CBM, CTA, IceCube, ILC, LHCb, ...

TA4: Real-Time Data Analysis and Selection

Example for work packages and deliverables in this task area:

Real time feature extraction

- Efficient and reliable algorithms
- Generalised interfaces
- Set of use cases

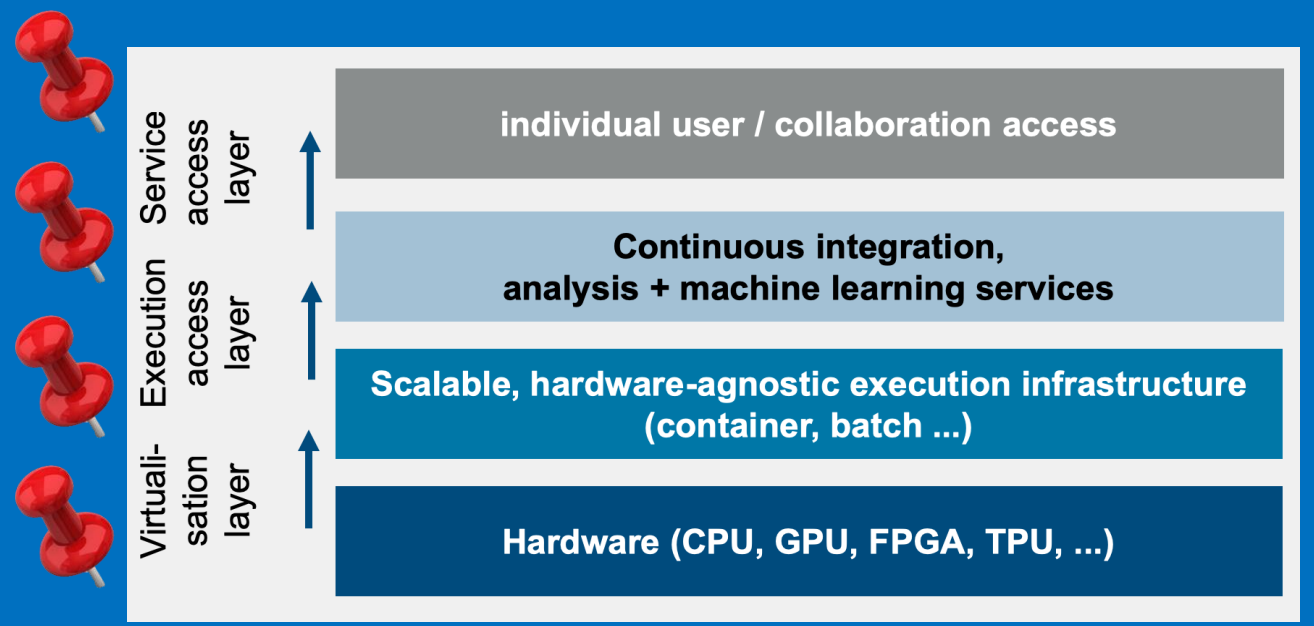
Example for services and synergies:

Real time software tools

First focus on fast real-time pattern recognition

Cross-Cutting Topics A and B “Synergies” and “Services”

„Synergy - the bonus that is achieved when things work together harmoniously"
(Mark Twain)

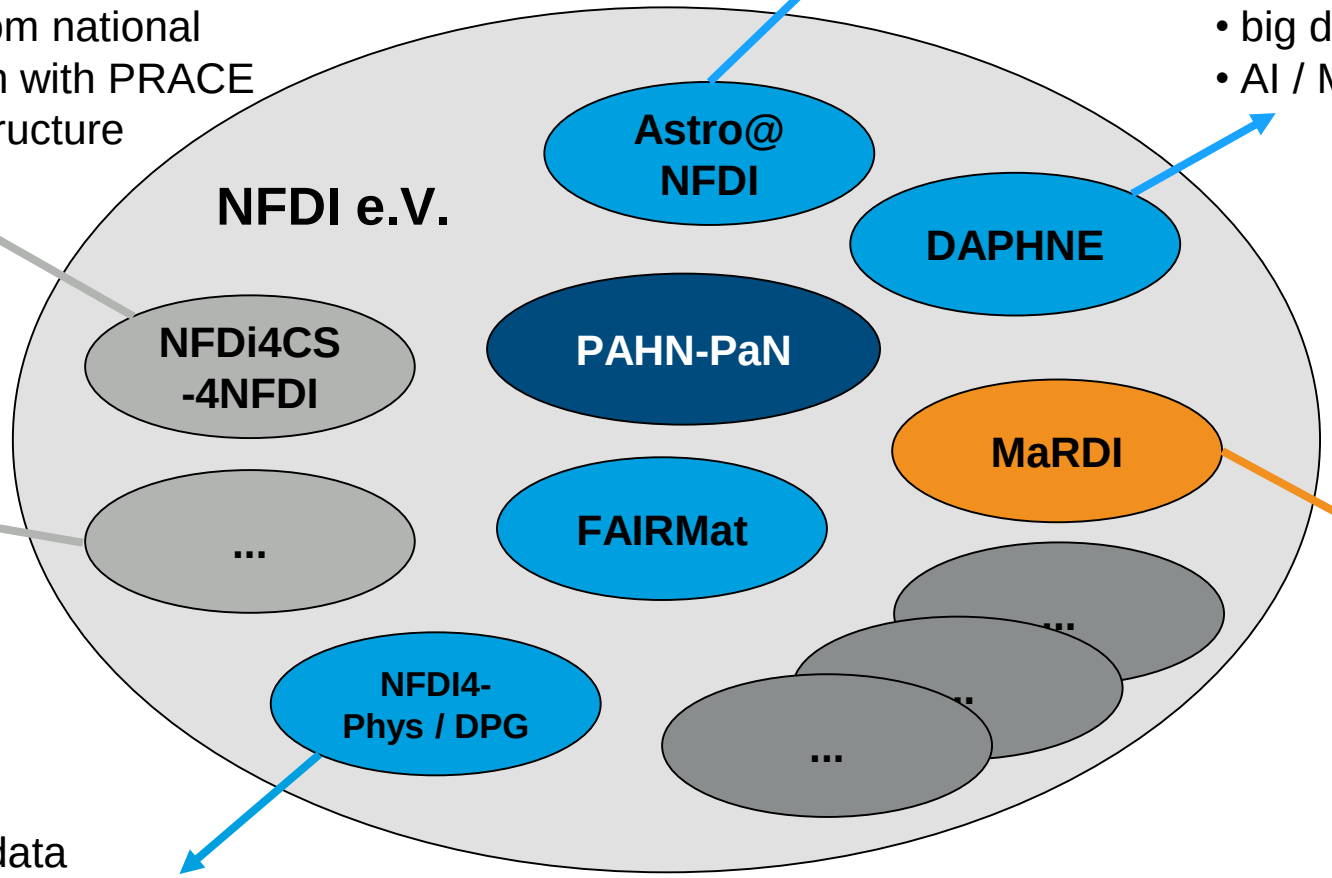


CCT A: Synergies

- mastering data irreversibility challenge
- open data and metadata definitions
- data lake / layer concept

- opportunistic resources from national HPC centres, collaboration with PRACE
- provisioning of DOI infrastructure

- big data management models
- AI / ML in data analysis / curation

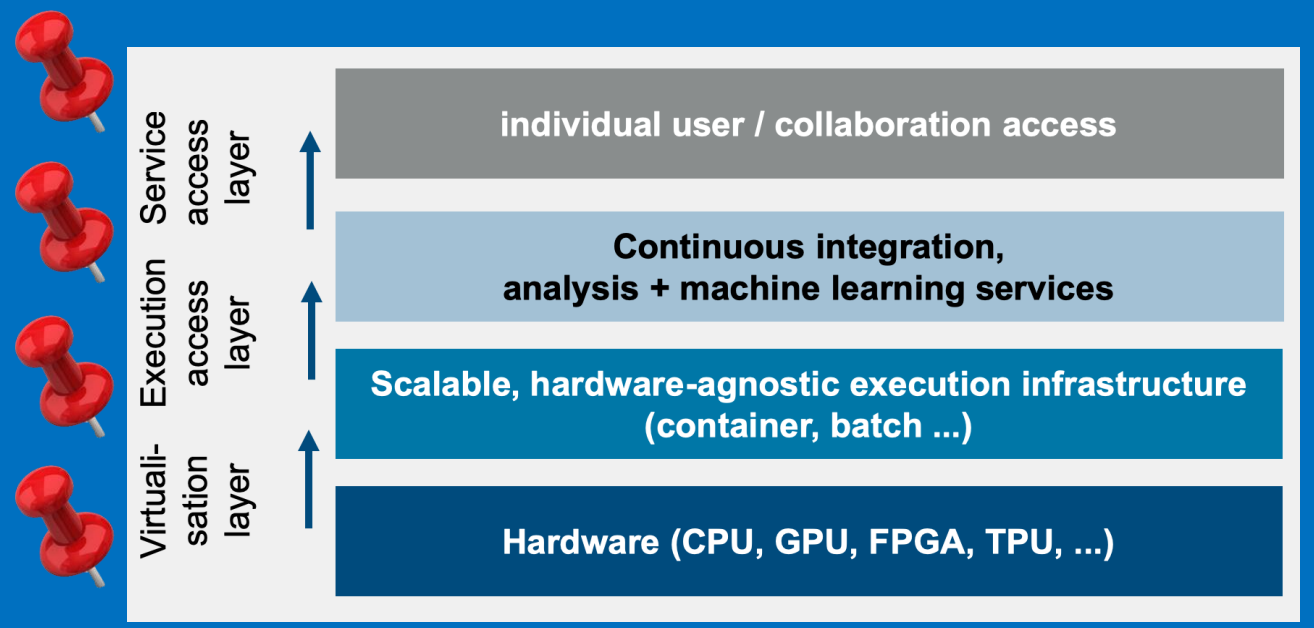


Medicine
- neural networks
- apply AI/ML methods

- multiplier for data management questions
- synergies with physics community

- data analysis with mathematical tools / machine learning
- benchmarking optimisation algorithms
- high-performance computer algebra

Cross-Cutting Topic C “Professional Training, Education, Outreach”



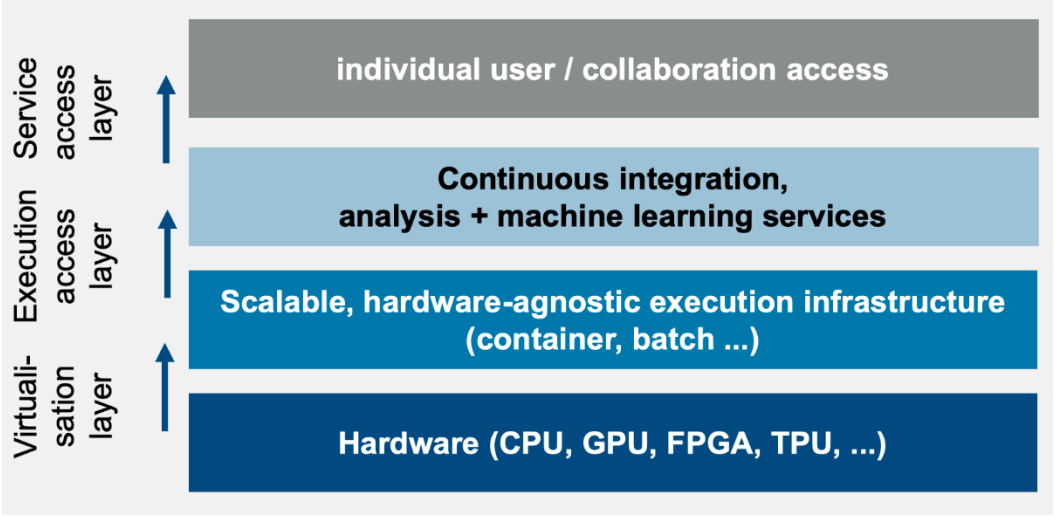
Summary

PAHN-PaN: three established international communities

- decade-long experience in large-scale data management
- existing infrastructures
- successful support of thousands of users

A pioneering consortium – exciting challenges ahead in terms of data volumes and rates. Solutions:

- New technologies and concepts
- NFDI as incubator for future ideas



Already now, PAHN-PaN can

- assist the NFDI to tackle medium- and long-term challenges in data management,
- help to keep the German science system competitive and
- strengthen the industry location via technology development and training.

PAHN-PaN contributes to a quick and successful start of the NFDI.

PAHN-PaN Synergies with ErUM-Data

…are complementary initiatives

ErUM-Data is a proposed action plan of the BMBF for **digitization in ErUM („Research on Universe and Matter“).**

- Input of 8 physics communities

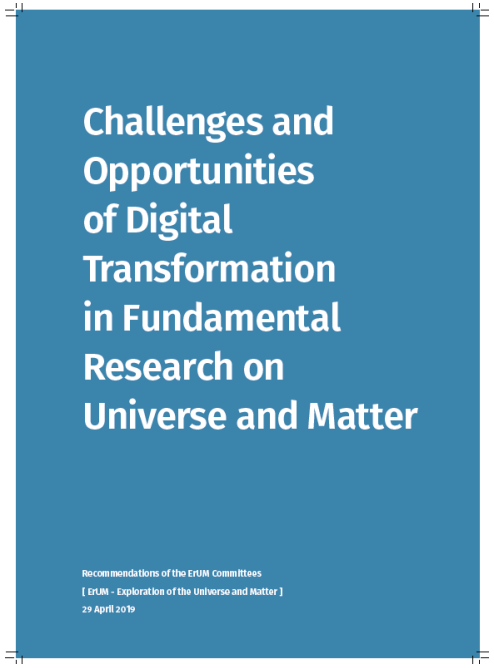
ErUM-Data focuses to **ErUM- and community specific**:

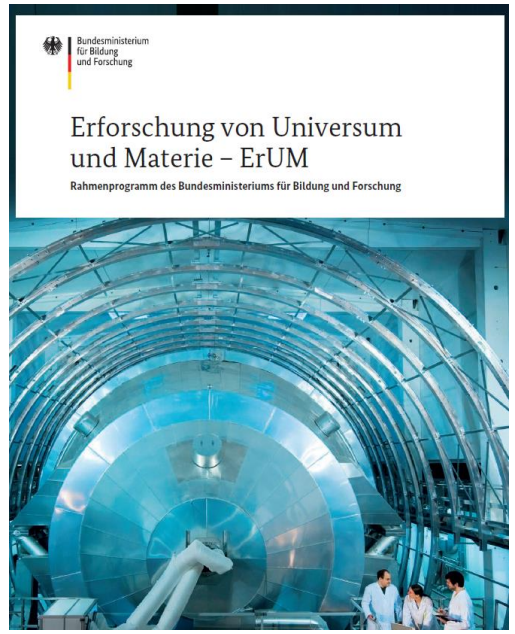
- Federated infrastructures (Compute power, Utilization, Workflows, ….)
- Big Data Analytics (Algorithms, Autonomization, Results, …)
- Research Data (Data models, Management, Curation)

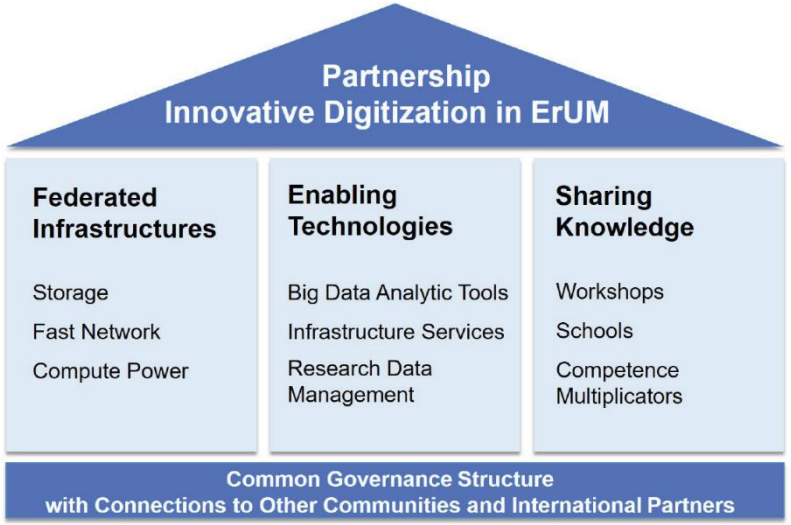
PAHN-PaN in NFDI focuses on **science-wide**:

- Curation of scientific data across all disciplines.
- Open Data and Workflows
- Development and provision of data management infrastructures and generalized services within and beyond the consortium.

➔ ErUM-Data will give valuable complementary input to PAHN-PaN and entire NFDI







WILHELM UND ELSE HERAEUS-STIFTUNG

The Science Cloud – Towards a Research Data Ecosystem for the next Generation of Data-intensive Experiments and Observatories

711. WE-Heraeus-Seminar

https://www.we-heraeus-stiftung.de/veranstaltungen/seminare/2020/the-science-cloud-towards-a-research-data-ecosystem-for-the-next-generation-of-data-intensive-experiments-and-observatories/

Physics ABOUT BROWSE PRESS COLLECTIONS

Facing a Downpour of Data, Scientists Look to the Cloud

February 3, 2020 • *Physics* 13, 14

To improve access to large data sets, scientists are looking to cloud-based solutions for data management.

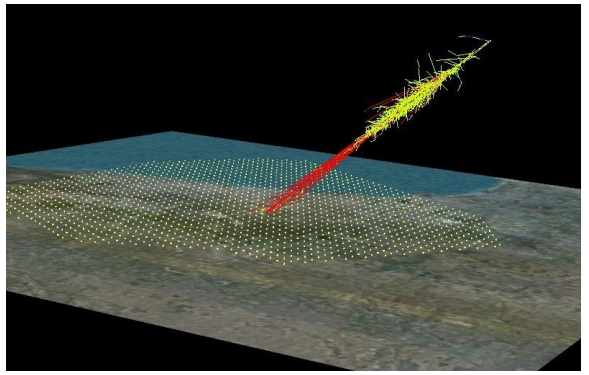
iStock.com/JuSun

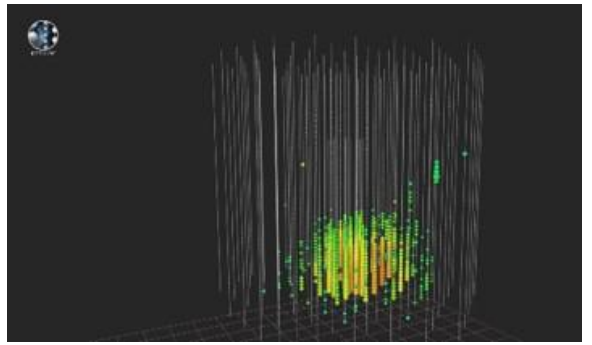
Storing experimental data in a "science cloud" has some advantages, such as making information more accessible to a wider scientific community.

"We all have to work on better recognition and visibility for people working on the interface between information technology and science"

Analysis and Data Centre for Multi-messenger Astroparticle Physics ADC-MAPP

- **Basics**
  - ADC-MAPP project period 2019-2020
  - funded by Helmholtz
- **Main targets of the Project**
  - Provide sustainable access to scientific data
  - Archiving of Data and Meta-Data
  - Providing analysis tools
  - Education in Big Data Science
  - Development area for multi-messenger analyses (e.g. Deep Learning)
  - Platform for communication and exchange within Astroparticle Physics





PIERRE AUGER OBSERVATORY

cta cherenkov telescope array

IceCube

Analysis and Data Center in Astroparticle Physics

- **Data availability:**

All researchers of the individual experiments or facilities require quick and easy access to the relevant data.

- **Analysis:**

Fast access to the generally distributed data from measurements and simulations is required. Corresponding computing capacities should also be available.

- **Simulations and methods development:**

Researchers need an environment for simulations and the development of new methods (machine learning).

- **Real-time analysis center:**

The multi-messenger ansatz requires a framework to develop and apply methods for joint data stream analysis.

- **Open access:**

More and more it is necessary to make the scientific data available also to the interested public: public data for public money!

- **Education in data science:**

Not only data analysis itself, but also the efficient use of central data and computing infrastructures requires special training.

- **Data archive:**

The valuable scientific data and metadata must be preserved and remain interpretable for later use (data preservation).

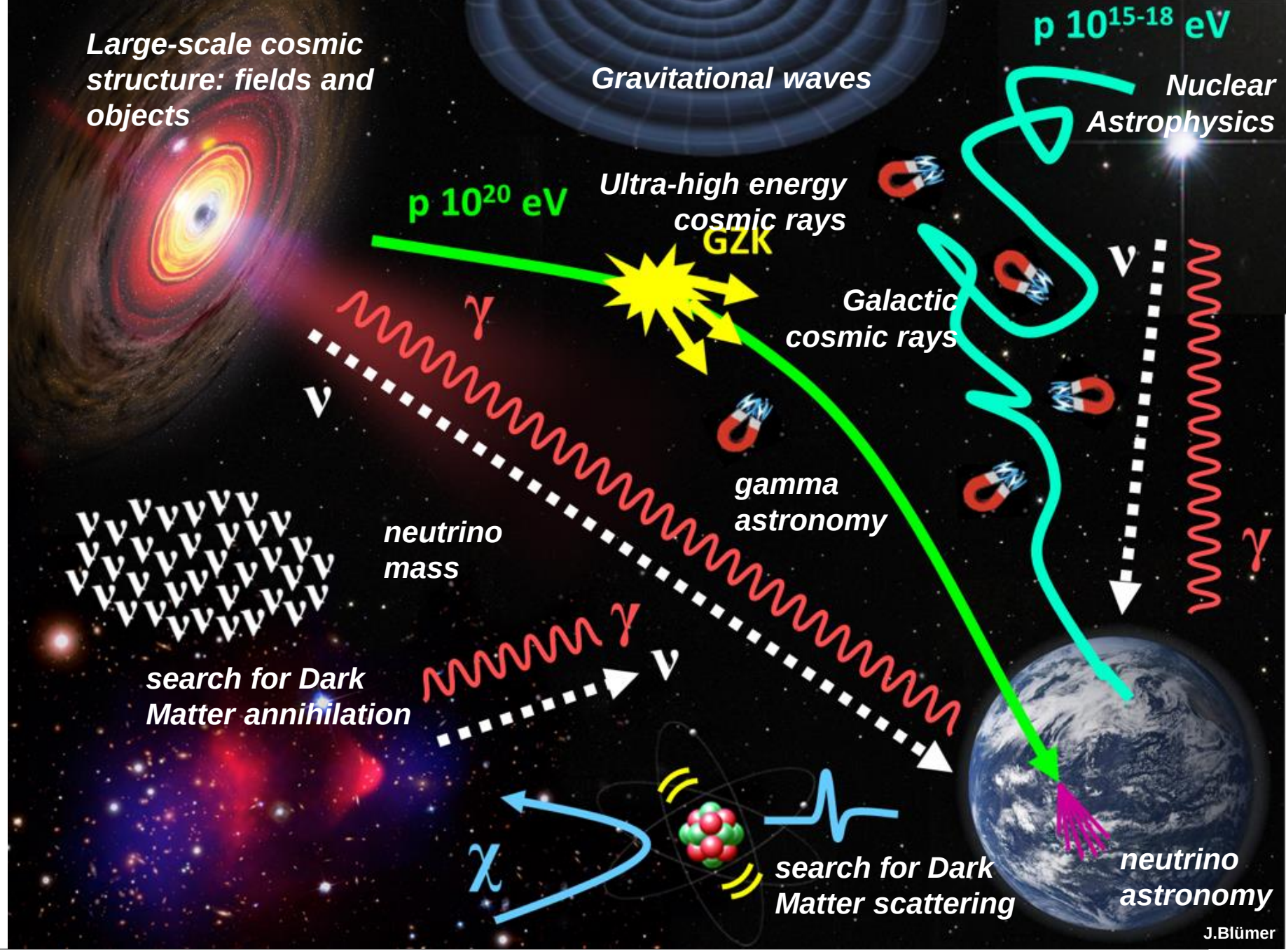
Partly realized in individual experiments

….everything for the benefit of Astroparticle Physics!

Astroparticle Physics = Understanding the

- **Multi-Messenger Universe**
- **Dark Universe**

needs an experiment-overarching platform!



END