Karlsruhe Institute of Technology

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

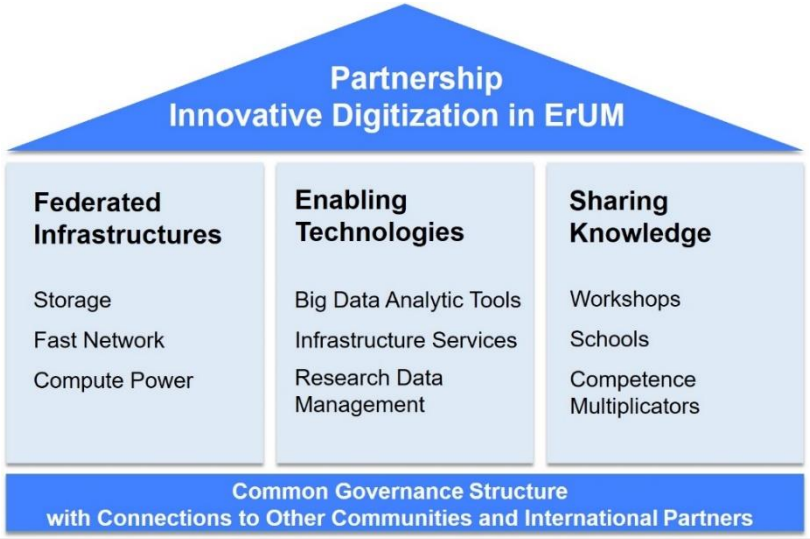
HELMHOLTZ GEMEINSCHAFT
Allianz für Astroteilchenphysik

KAT. Komitee für Astro. Teilchen. Physik

Challenges and Opportunities of Digital Transformation in Fundamental Research on Universe and Matter

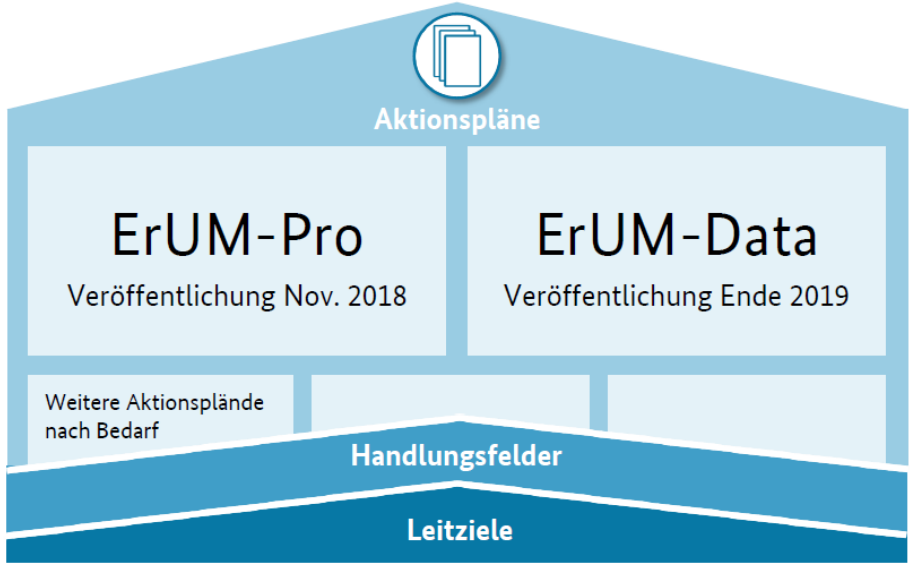
HAP / AKPIK workshop | Big Data Science in Astroparticle Physics
Aachen, 18-20 February 2019

Andreas Haungs



ErUM

action plan: 2017-2027



one plan of action:
ErUM-Data: Contributions to the digital agenda

Committees related to ErUM in Germany

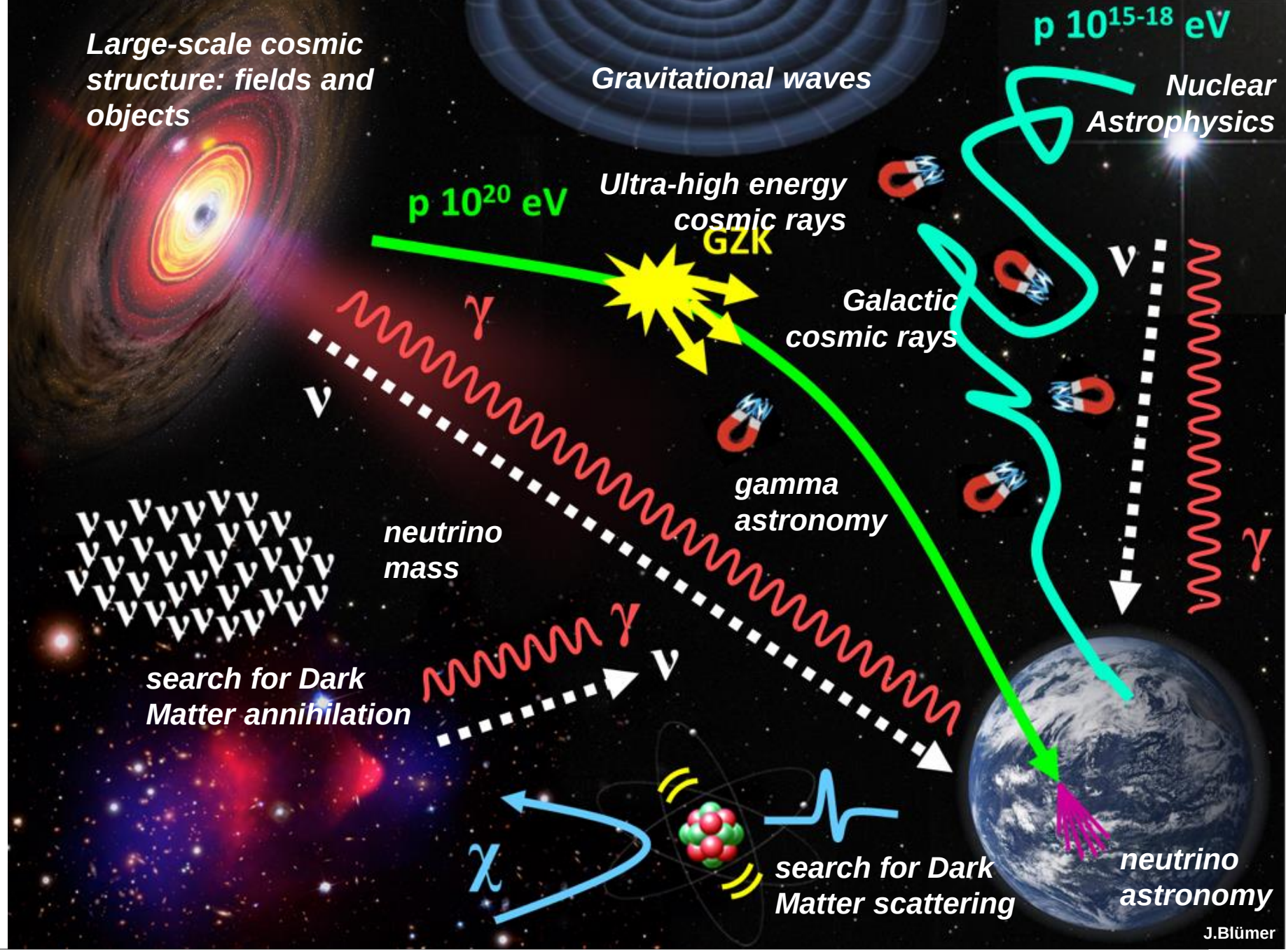
	Scientists with doctoral degree
KFS	4.000
RDS	1.500
KHuK	1.500
KET	1.300
KFN	1.000
KAT	500
KfB	200
KFSI	100
	10.100

Initiative for a (global) Analysis & Data Center in Astroparticle Physics

Astroparticle Physics = Understanding the

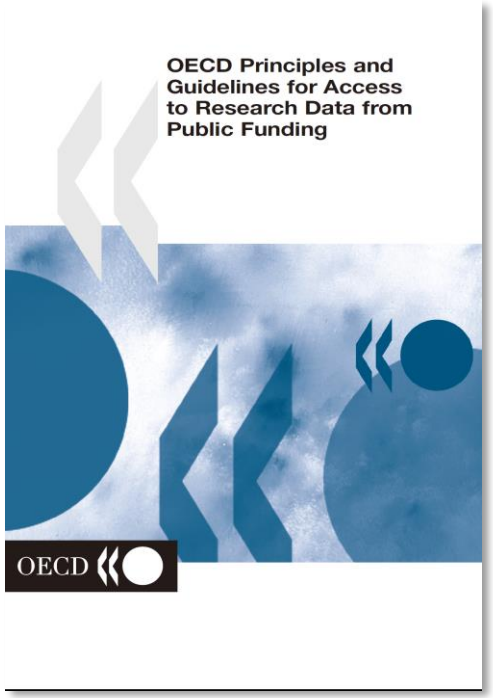
- **Multi-Messenger Universe**
- **Dark Universe**

needs an experiment-overarching platform!



Initiative for a (global) Analysis & Data Center in Astroparticle Physics

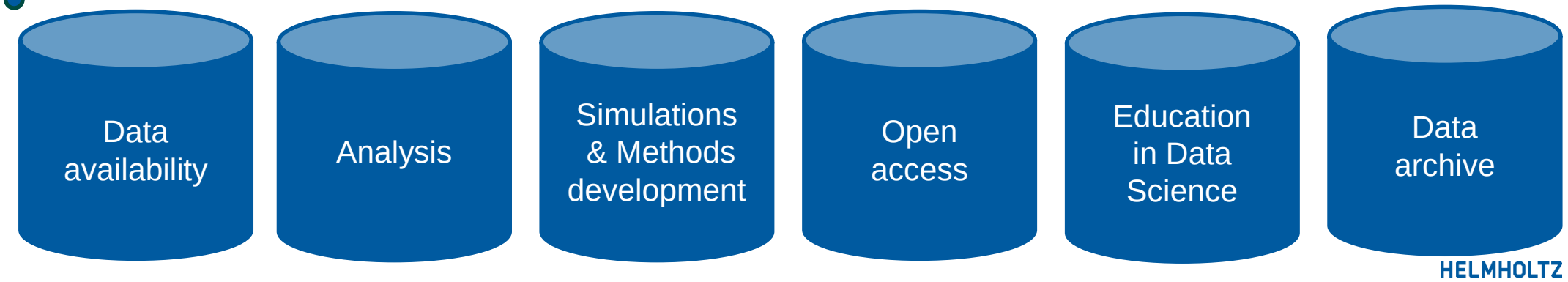
- Astroparticle Physics requests for multi-messenger analyses - this needs an **experiment-overarching** platform!



- **Tasks**
  - 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
- **Elements**
  - Advancement, generalization of existing structures (like KCDC and others)
  - In direction of a virtual Observatory (like in astronomy)
  - In direction of Tier-systems and DPHEP (like in particle physics)
  - „Digitale Agenda der Bundesregierung“
  - OECD Principles and Guidelines for Access to Research Data from Public Funding
  - Follow the FAIR principles of data handling

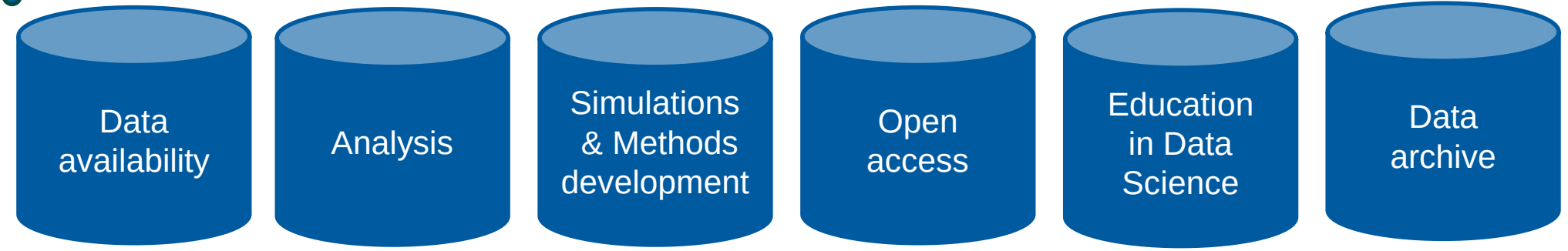
FINDABLE-ACCESSIBLE-INTEROPERABLE-REUSABLE

Analysis and Data Center in Astroparticle Physics



- **Develop a global analysis & data centre as user facility for multi-messenger studies in astroparticle physics**
- Motivation:
  - Needed, as experiments globally distributed and no worldwide centre like CERN exist
  - Implementation of ‘Digital Agenda’ and ‘Big Data Science’ in Astroparticle Physics
  - Apply ‘FAIR’ data handling in Astroparticle Physics
- Elements:
  - Data Preservation; virtual Observatory; distributed resources, data provider; outreach;
  - Based on experience of KCDC, GridKa, CTA data center, IceCube-Tier1, VISPA, NIFTY5
  - User-led facility (in Germany: 2 Helmholtz, 3 Max-Planck, 15 Universities)
- Realization as sustainable User Facility

Analysis and Data Center in Astroparticle Physics



Partly realized in individual experiments

- **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:**

The researchers need an environment for the production of relevant simulations and the development of new methods (machine learning).

- **Open access:**

More and more it is necessary to make the scientific data available not only to the internal research community, but 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).

KASCADE Cosmic ray Data Centre

- **Motivation and Idea of KCDC:**
  - public access to the data
  - data has to be preserved for future generations
- **Web portal:**
  - modern software solution
  - release the software as Open Source
  - educational courses
- **Data access:**
  - new release (Feb. 2017) with $4.3 \cdot 10^8$ EAS
  - simulation data
  - spectra
- **Pioneering work in publishing research data in astroparticle physics**

#KCDC_KIT

https://kcdc.ikp.kit.edu/

[J.Phys.Conf.Ser. 632 (2015) 012011]
[EPJ C78 (2018) no.9, 741]

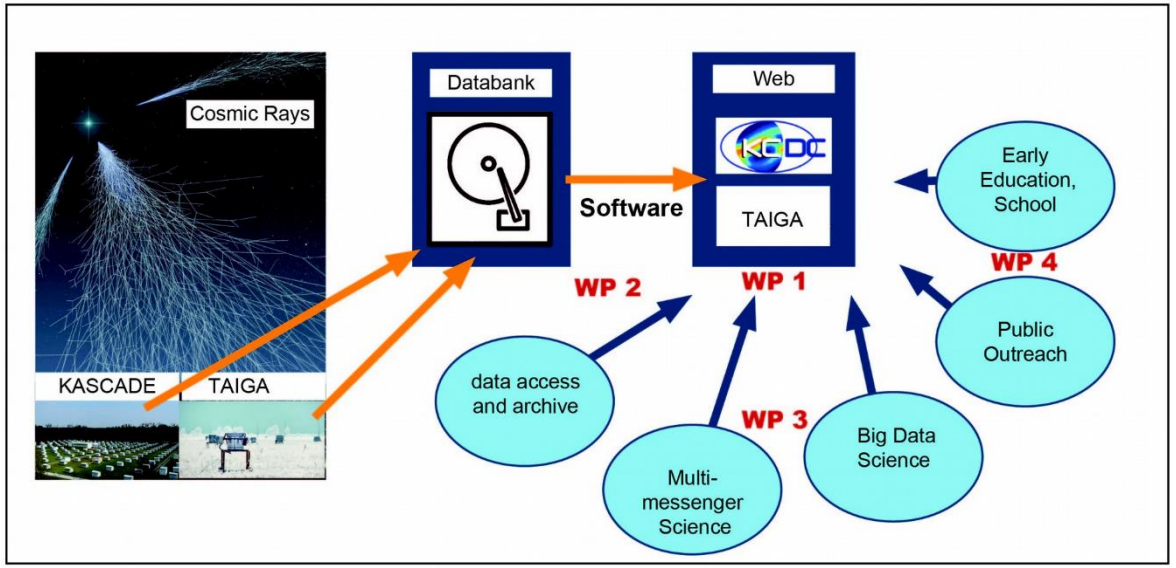
Astroparticle Data Life Cycle Initiative

see talk Victoria Tokareva

- **Basics**
  - project period 2018-2020
  - funded by Helmholtz and RSF
  - Russia: SINP MSU, ISU, ISDCT SB RAS
    Germany: KIT, DESY

- **Main targets of the Project**
  - Extension: data from Tunka/TAIGA and KASCADE-Grande
  - Developing solutions of distributed data storage techniques with a common metadata catalog
  - Development of appropriate machine-learning workflows
  - Perform experiment overarching multi-messenger astroparticle physics
  - Learn to use GridKa environment
  - Creation of an educational subsystem

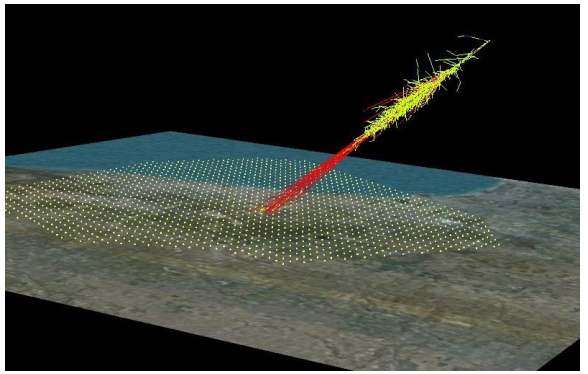
    http://astroparticle.online

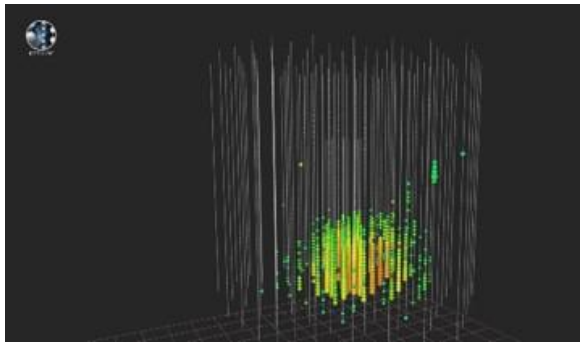


Project is a first step in extension and generalization of KCDC

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





Open positions for data scientists - Contact me…. ☺

PIERRE AUGER OBSERVATORY

cta cherenkov telescope array

IceCube

Support by the BMBF:

Innovative Digital Technologies for the Erforschung von Universum und Materie

Consortium KAT-KET-KHuK (pilot project ErUM-Data)

Proposal (11 Universities + 6 associated partners; coordinator T. Kuhr of LMU)

- Topic A: Development work for the provision of technologies to leverage heterogeneous computing resources
- Topic B: Application and testing of virtualized software components in the environment of heterogeneous computing resources
- Topic C: Deep learning, gaining knowledge through well-founded data-driven methods
- Topic D: Event reconstruction: cost and energy efficient use of computing resources

Approved for period 10/2018-9/2021

one plan of action:
ErUM-Data: Contributions to the digital agenda

How do we organize our digitization era

Education/Schools

Deep Learning in Physics

- Master course, RWTH 17,18,19
- GRIDKA summer school, KIT 17,18
- Grad. Kolleg, Freiburg 18
- Belgium Dutch German summer school, Berlin 18
- …

Workshops

Big Data Science

- made in Germany, Berlin 17
- Astroparticle Research, RWTH 17,18, 18.-20.2.19
- Machine Learning, CERN 17,18,19
- …

White papers

- KAT: Astroteilchenphysik im Licht der Digitalen Agenda
- KAT, KET, KfB, KFN, KFS, KHuK, RDS: Gemeinsames Strategiepapier
- A Roadmap for HEP Software and Computing R&D for the 2020s
- …

Grants

Funded

- 12 Universities, 5 Research Centers: Innovative Digitale Technologien für die Erforschung von Universum & Materie
- …

Whitepaper ErUM-Data

"Digitalisation in ErUM":
BMBF-Workshop 4-5 October 2018

- Federated Infrastructures
  - Efficient usage
  - Services
- Research Data Management
  - Data life cycle;
  - Networking (NFDI, EOSC);
- Big Data Analytics
  - Deep Learning;
  - Provide sustainable algorithms and tools;

Challenges and Opportunities of Digital Transformation in Fundamental Research on *Universe and Matter*

Martin Erdmann[1], Christian Gutt[2], Andreas Haungs[3], Klaudia Hradil[4], Thomas Kuhr[5], Marcel Kunze[6], Anke-Susanne Müller[7], Günter Quast[8], and Matthias Steinmetz[9]

[1]RWTH Aachen University, KAT
[2]University of Siegen, KFS
[3]Karlsruhe Institute of Technology, KAT
[4]Technische Universität Wien, KFN
[5]Ludwig Maximilians University Munich, KET
[6]Universität Heidelberg, KHuK
[7]Karlsruhe Institute of Technology, KfB
[8]Karlsruhe Institute of Technology, KET
[9]Leibniz-Institut für Astrophysik Potsdam, RDS

Our charge: write down concrete portfolio of measures → BMBF action plan → calls

Federated Infrastructures

- **Increasingly heterogeneous computing infrastructures available and needed (HTC vs. HPC)**
- **Huge Storage: Multiple Exabytes**
- **Fast Networks: >100 Gb/s for entire ErUM**
- **Substantial large-scale experiences in all related aspects and connected to computer science, multiple domain specific aspects**
- **need large scale federated infrastructures from experienced providers (including commercial providers)**
- **Utilization needs sustained software development thus sustained positions**
- **Infrastructure in ErUM as building block of national (NFDI) and international (EOSC) initiatives**

GridKa KIT

DESY Grid Centre & NAF

GreenCube GSI

ICT infrastructures at Univs

BERLIN BIG DATA CENTER

Gauß-Allianz

SuperMUC Garching

etc…

CernVM Software Appliance

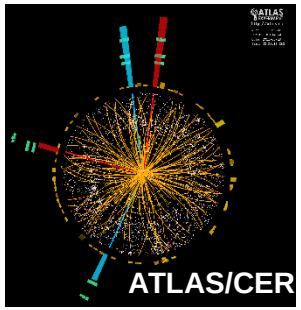
CernVM File system

HTCondor High Throughput Computing

XRootD

ROOT

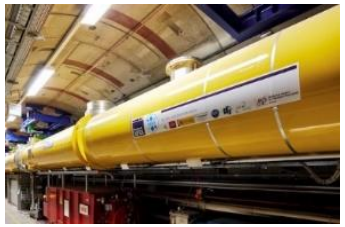
High luminosity LHC



Square Kilometre Array

300 PetaBytes per year

Experiments at XFEL



Big Data Analytics

- **Utilization Big Data Analytics in national and international contexts:**
  - **Development and implementation of tools for Big Data Analytics**
  - **Need for a collaborative effort in terms of Big Data Analytics including users, facilities, mathematics and computer science**
  - **A platform for sharing Big Data Analytics solutions (inside or even across communities).**
  - **Integration with data management (e.g. for efficient data access or mining archived data)**
  - **Integration with federated infrastructure (e.g. for utilizing resources optimized for Big Data Analytics tasks).**
  - **Training and education of the next generation of scientists in Big Data Analytics;**
  - **Ensure sustainable development and curation of algorithms and tools.**

Scoogle: Scientist‘s data & algorithms

Scientists:
Questions

Web Interface

Big Data Analytic Tools

Algorithms
Visualization
Machine Learning

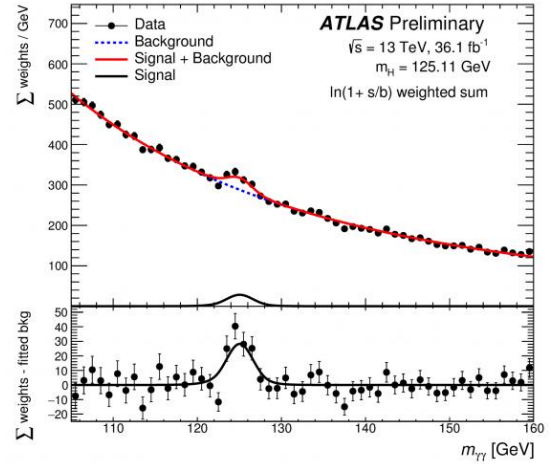
Data

Experiment data
Metadata
Simulation

Web vision

Access to
- Metadata & experimental data
- Computing resources
- Software libraries
- Algorithm development
- Workflow management system

Web vision



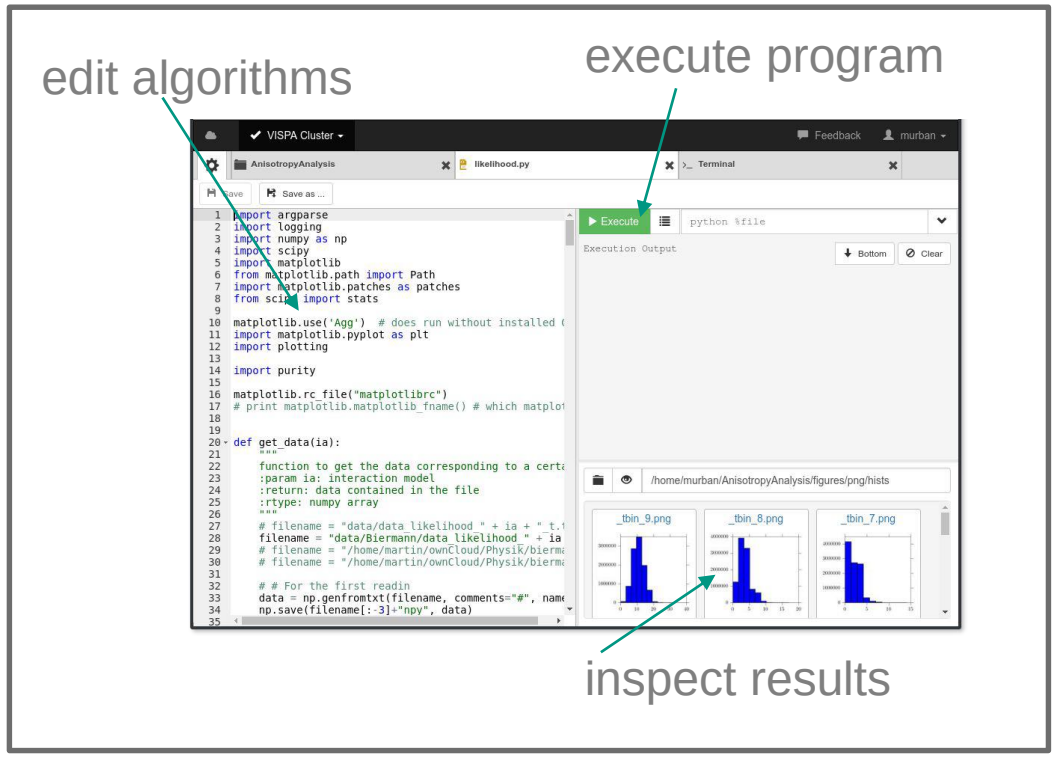
Higgs Boson






Scoogle: Scientist‘s data & algorithms

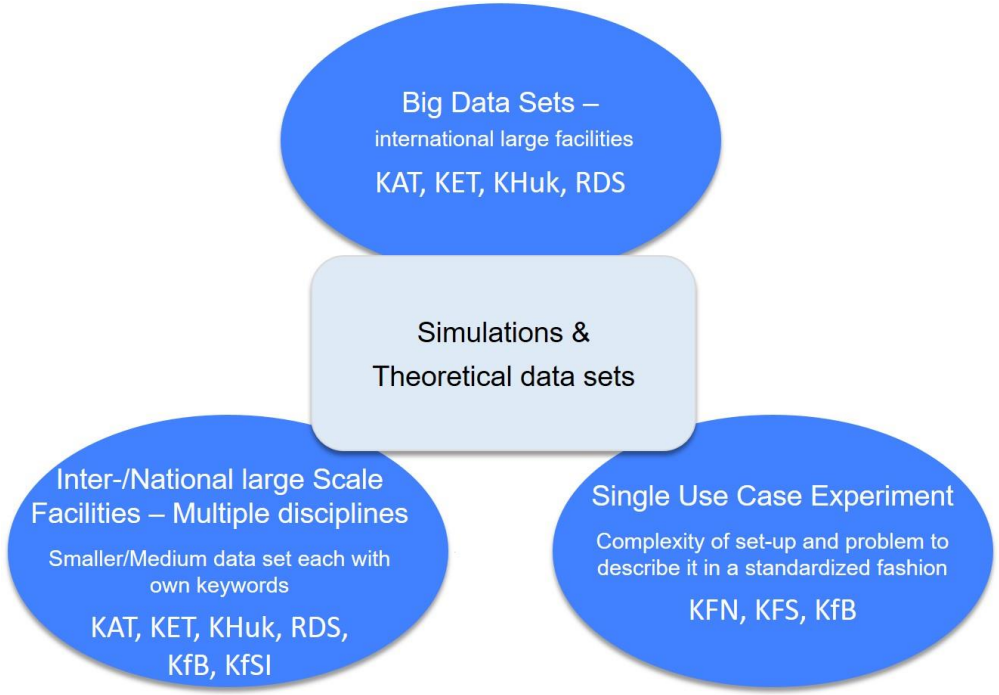
Medium-scale prototypes exist, developed in our community (~5 years experience)

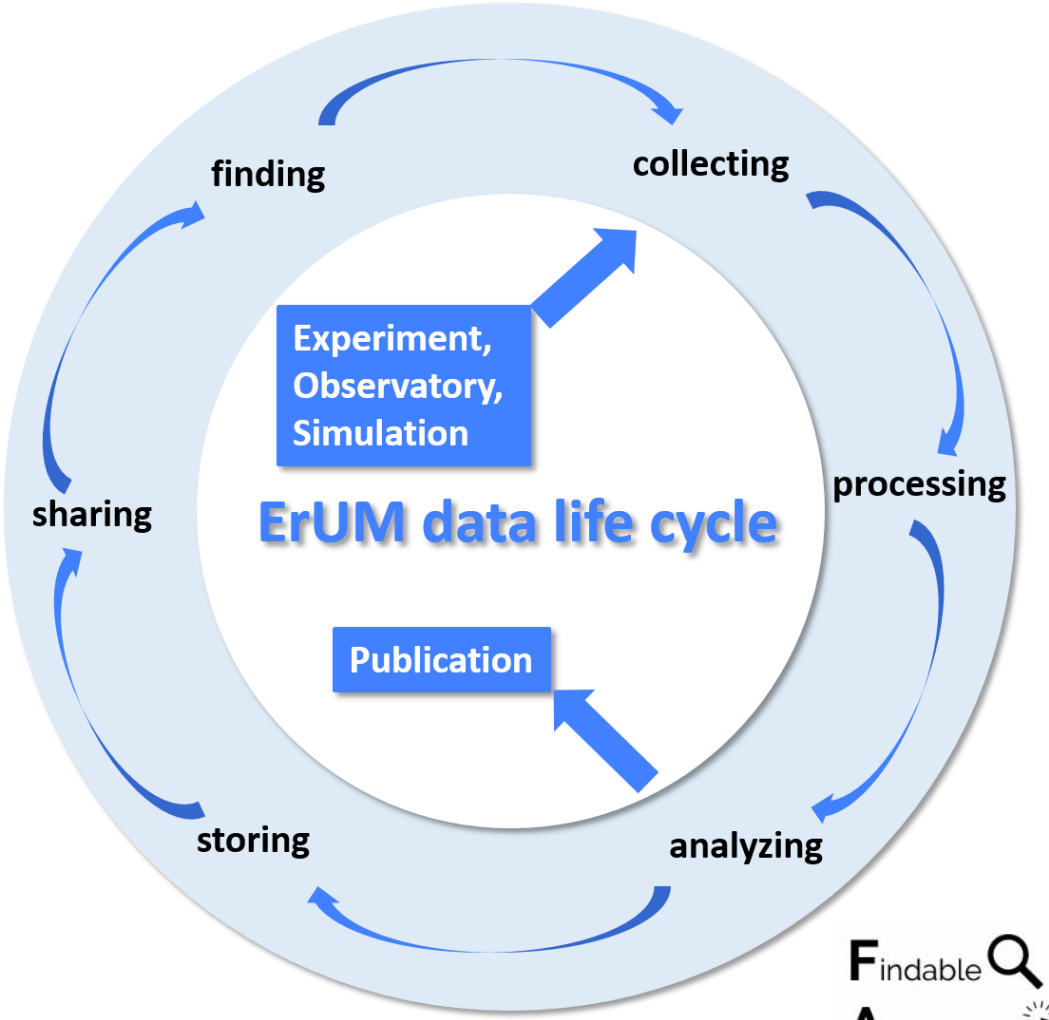


In ErUM: Substantial experiences in all related aspects, partly complemetary, international context, also connected to mathematics, computer science, economy.

Research Data Management

- **Where possible, common standards should be established to foster interoperability**
- **Importance of “data stewards” to manage the data life cycle and to act as a curator for metadata**

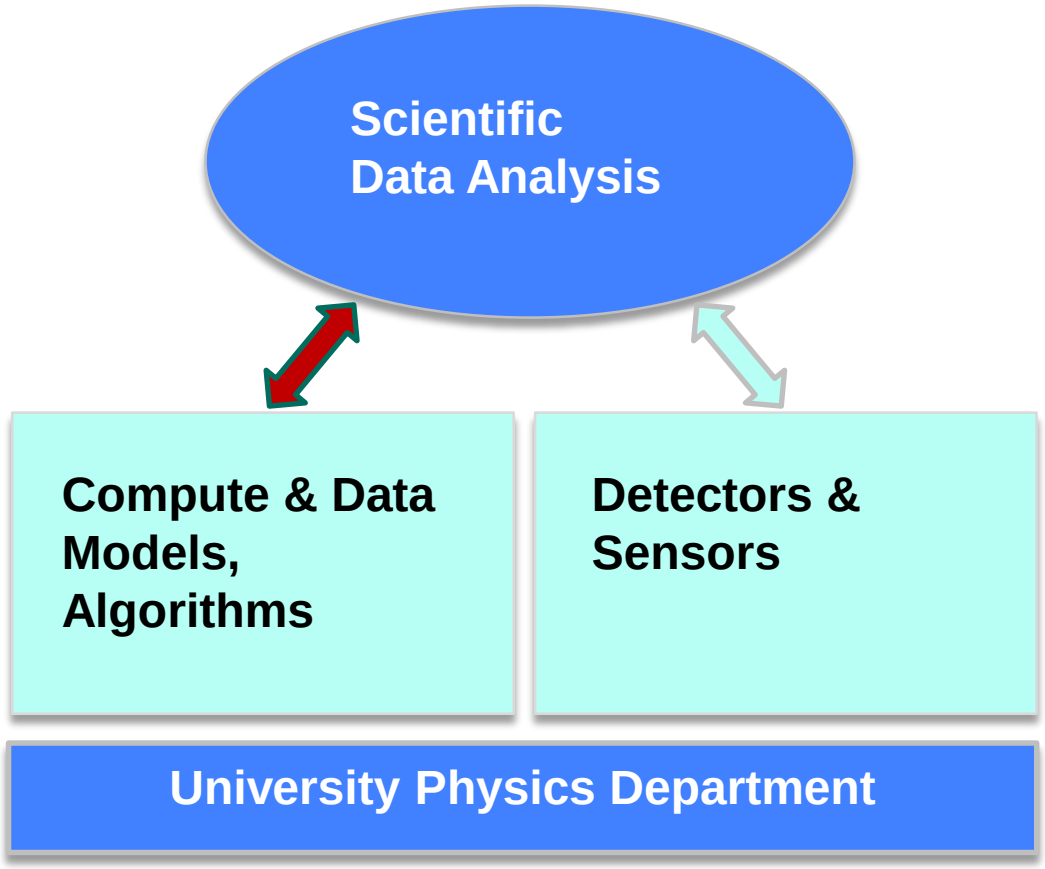




Findable
Accessible
Interoperable
Reusable

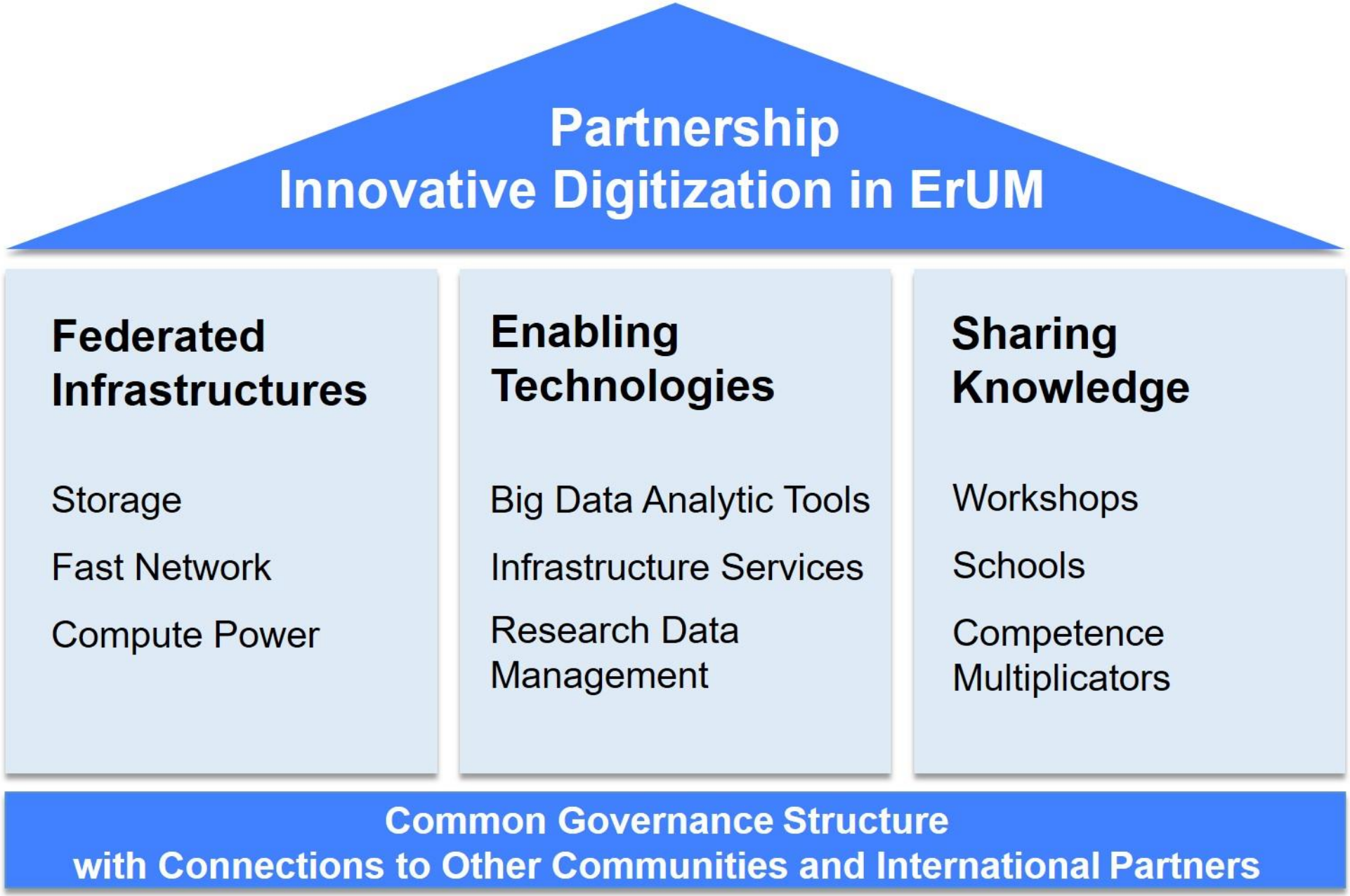
Tenure Track Programm

- **Education by scientific leaders: distribute and deepen knowledge in digitization**
- **Large tenure track programme for:**
  - **Development of compute models for online & offline reconstruction, simulation, analytics**
  - **New algorithmic concepts, machine learning**
  - **Access to heterogeneous computing resources**
  - **New chairs will advance curricula**



The ErUM House:

- **user-led home to bundle and steer the activities**



The needs ands costs for digitization of the research area ErUM (estimate for a period of 10 years)

Manpower:

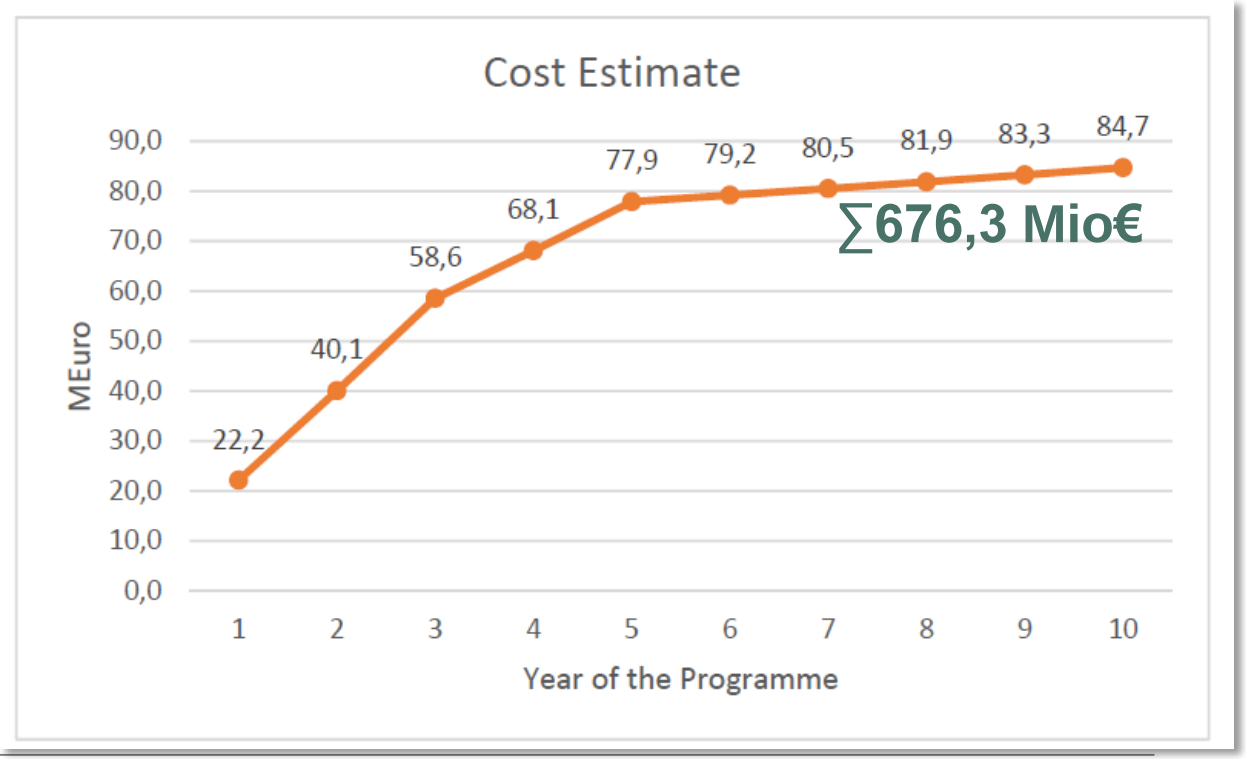
1. Workflows to exploit infrastructures: from 40/a to 100/a
2. Management of research data: from 40/a to 100/a
3. Big Data Analytics in physics research: from 40/a to 200/a
4. Scientist's web working environment: from 40/a to 100/a
5. Tenure track ErUM programme + 1 postdoc; from 40/a to 100/a

No numbers → No money

Costs:

- Full Time Equivalents:
  from 16M€/a to 59M€/a
- Large-scale federated infrastructures:
  from 5M€/a to 25M€/a
- Partnership for innovative digitization:
  for 1M€/a

→ **10,100 ErUM scientists:**
0.6% FTE *increase*/year:
6.8k€/scientist/year!



DPG-AKPIK, Arbeitskreis Physik, moderne Informationstechnologie und Künstliche Intelligenz

Broad representation of interests for the relevant topics in physics

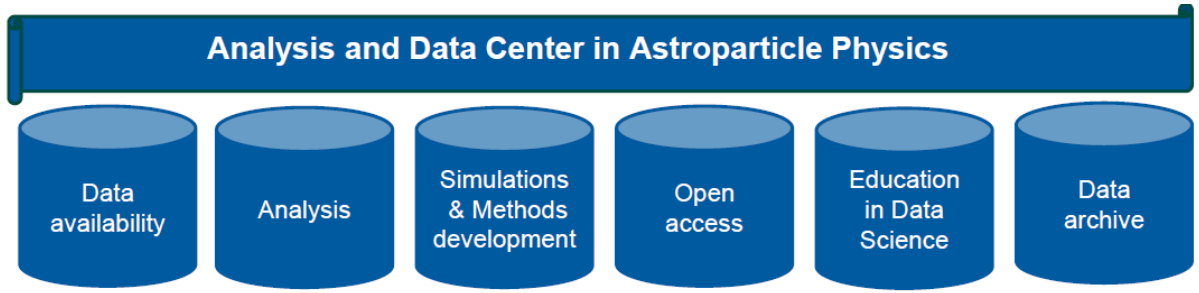
www.dpg-physik.de/dpg/gliederung/ak/akpik/index.html

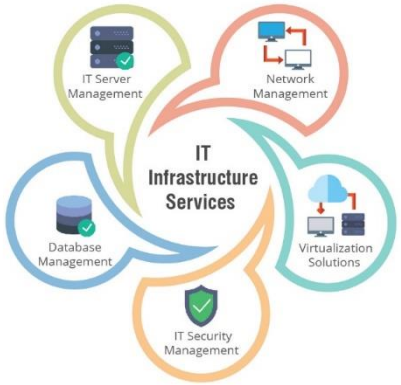
Ask Karl Mannheim, Martin Erdmann, Wolfgang Rhode

Topics:

1. **BIG DATA:** archiving, processing, management, analysis and simulation of complex data streams, HPC, information theory, statistical methods
2. **IT:** high-performance data readout systems and mass storage, visualization, smart sensors, bridge technologies for the next level of big data
3. **KI & ROBOTIK:** Data Driven Algorithms & Software, Autonomous Devices, Remote Control, Innovative Applications, Algorithms for Quantum Computers
4. **UNIVERSITY:** curricula and multi-disciplinary research centres, cooperation with the GI Task Force "Data Scientist", IT infrastructure
5. **INDUSTRY and SOCIETY:** Ethics, Technology Assessment, Sustainability, Business, Law, Start-Ups, Public

Era of Digitization in Astroparticle Physics





Lots of activities in

- **Infrastructures**
- **Big Data Analytics**
- **Research Data Management**

} **Data Life Cycle**

Next steps:

- **ErUM-Data is a big chance!!**
- **NFDI = Nationale ForschungsDatenInfrastruktur**
- **EOSC = European Open Science Cloud**