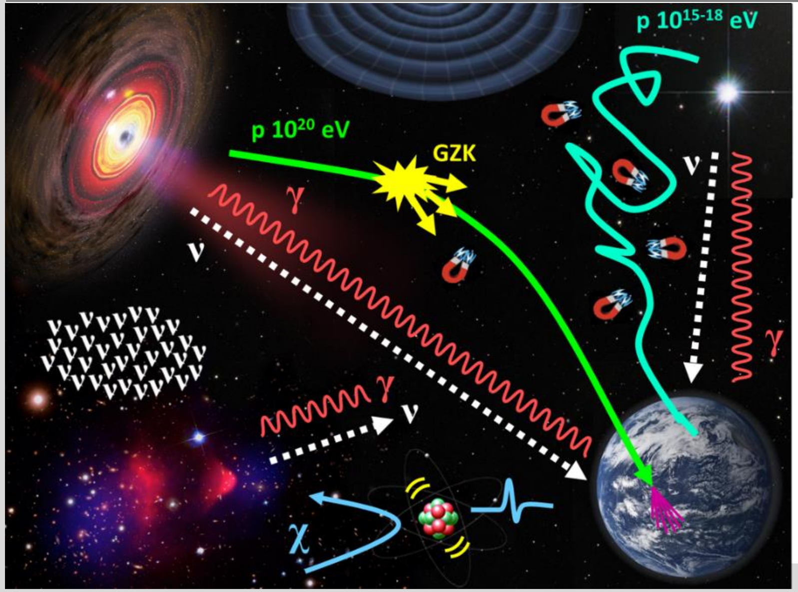
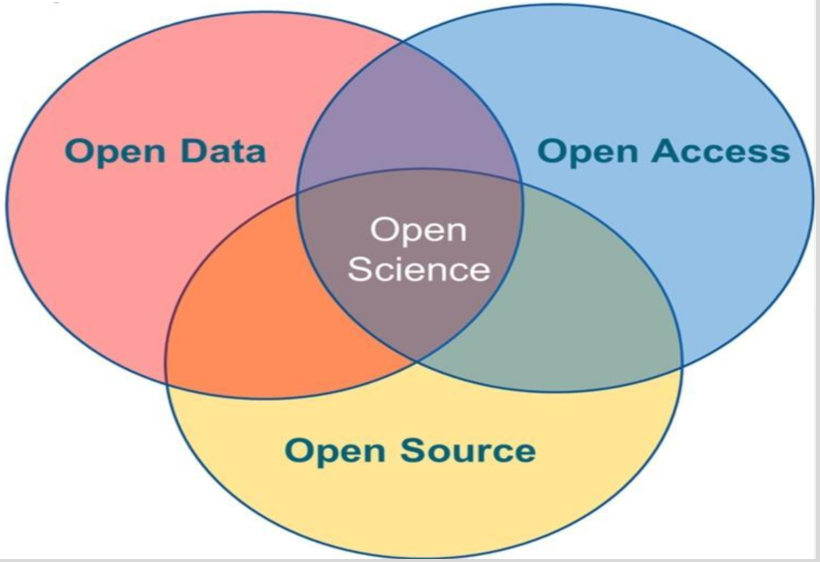
KRAD-APPDS joint meeting
Irkutsk, April 2019

Towards the GRADLC Initiative

1. General slides on motivation on future of a global data centre
2. National Initiative in Germany
3. GRADLCI
4. Objectives of this meeting

Andreas Haungs

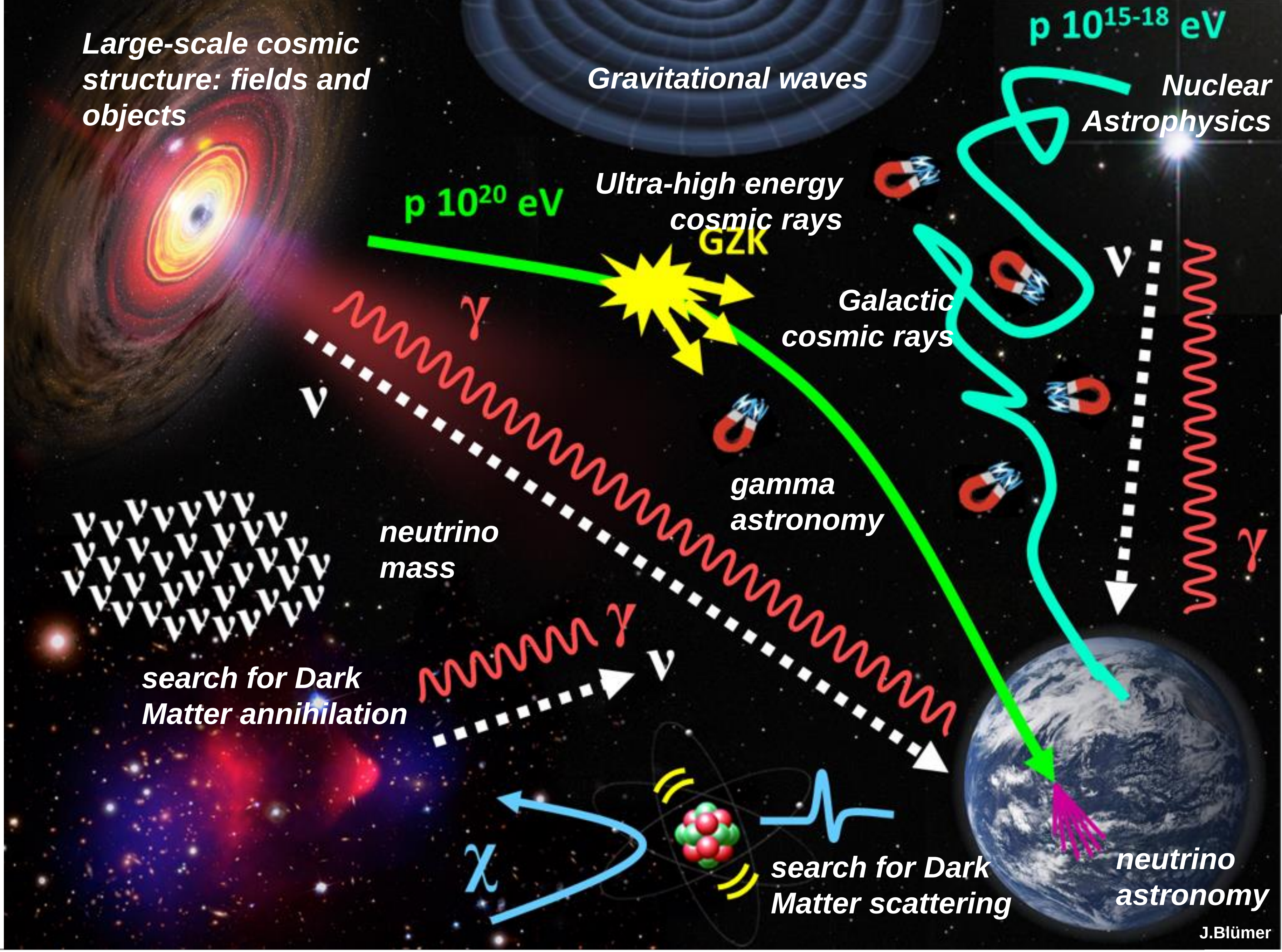




Initiative for a (global) Analysis & Data Center in Astroparticle Physics

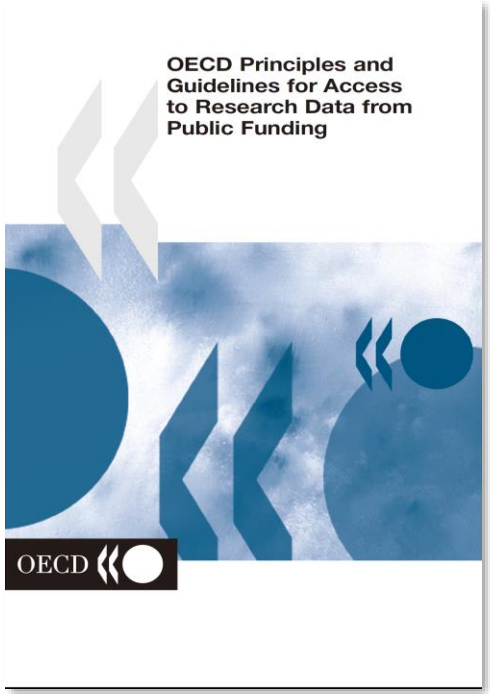
Astroparticle Physics = Understanding the
- **Multi-Messenger Universe**
- **Dark Universe**

needs an experiment-overarching platform!



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

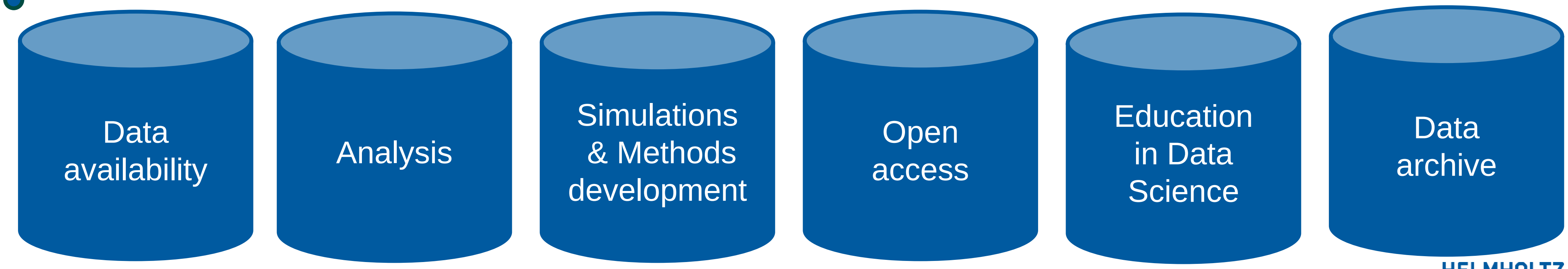
- Astroparticle Physics requests for multi-messenger analyses - this needs an experiment-overarching platform!



- 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

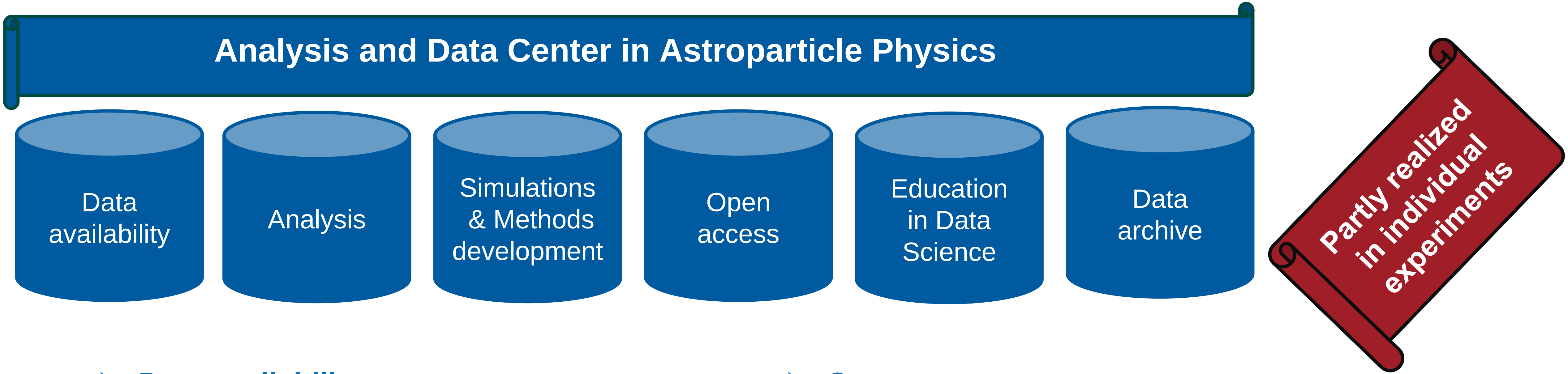
FINDABLE-ACCESSIBLE-INTEROPERABLE-REUSABLE

Analysis and Data Center in Astroparticle Physics



HELMHOLTZ

- **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



- **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

#KCDC_KIT

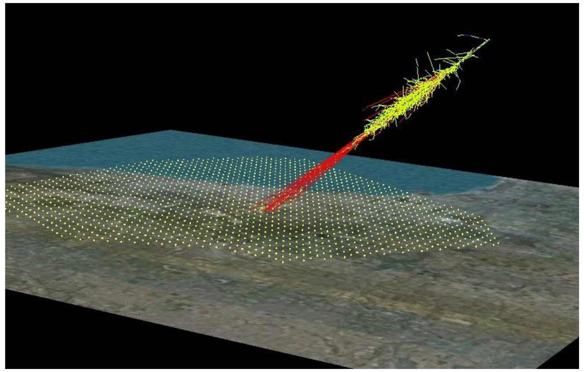
- **Pioneering work in publishing research data in astroparticle physics**

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

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

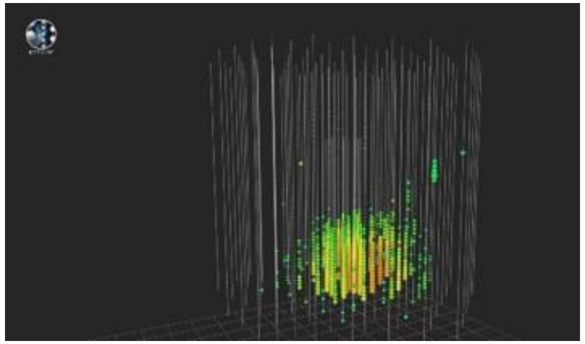
Analysis and Data Centre for Multi-messenger Astroparticle Physics ADC-MAPP

- **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

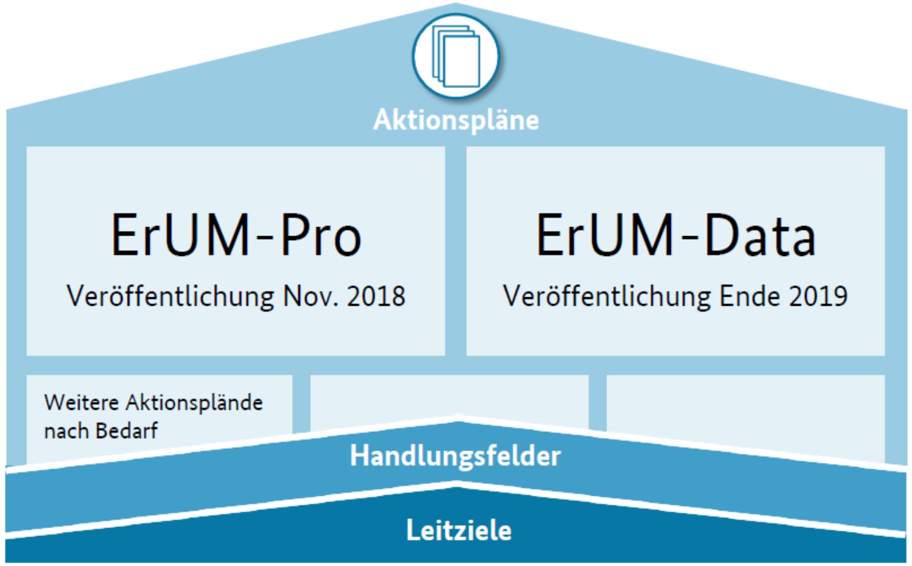


ICECUBE

Open positions for data scientists - Contact me…. ☺

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

Whitepaper ErUM-Data

Bundesministerium für Bildung und Forschung

"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

BBDC BERLIN BIG DATA CENTER

GA Gauß-Allianz


SuperMUC Garching

etc…

CernVM Software Appliance

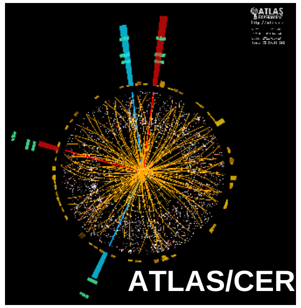
CernVM File system

HTCondor High Throughput Computing

XRootD

ROCED

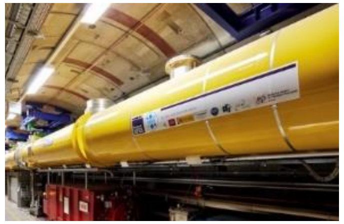
High luminosity LHC



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

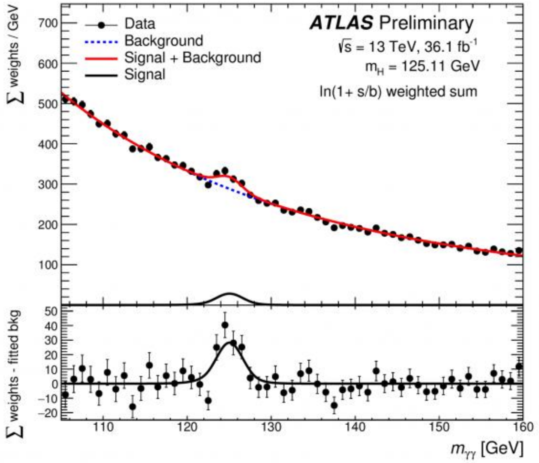
Data

Experiment data
Metadata
Simulation

Web vision

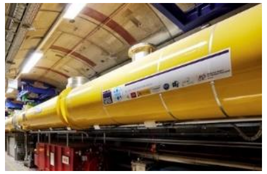
Access to
- Metadata & experimental data
- Computing resources
- Software libraries
- Algorithm development
- Workflow management system

Web vision



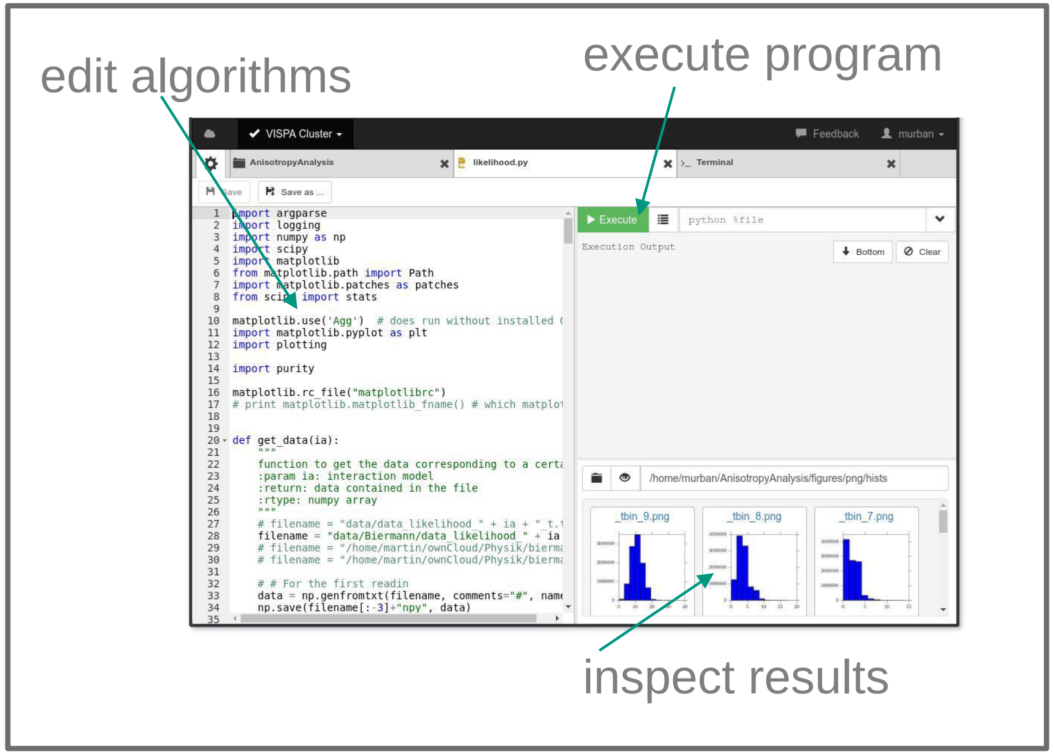
Higgs
Boson






Scoogle: Scientist‘s data & algorithms

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

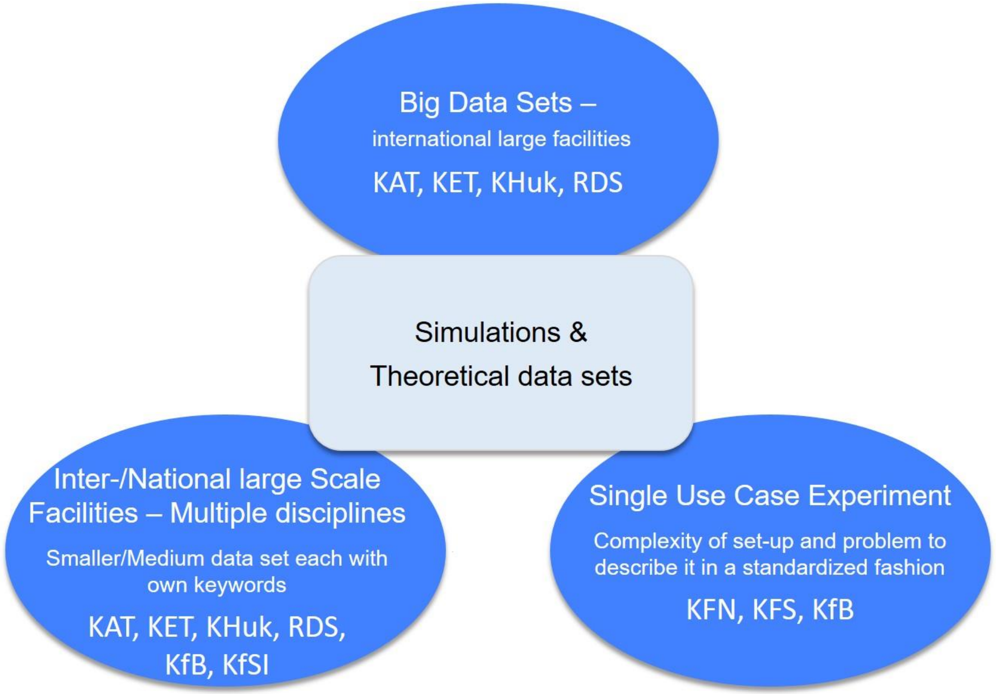


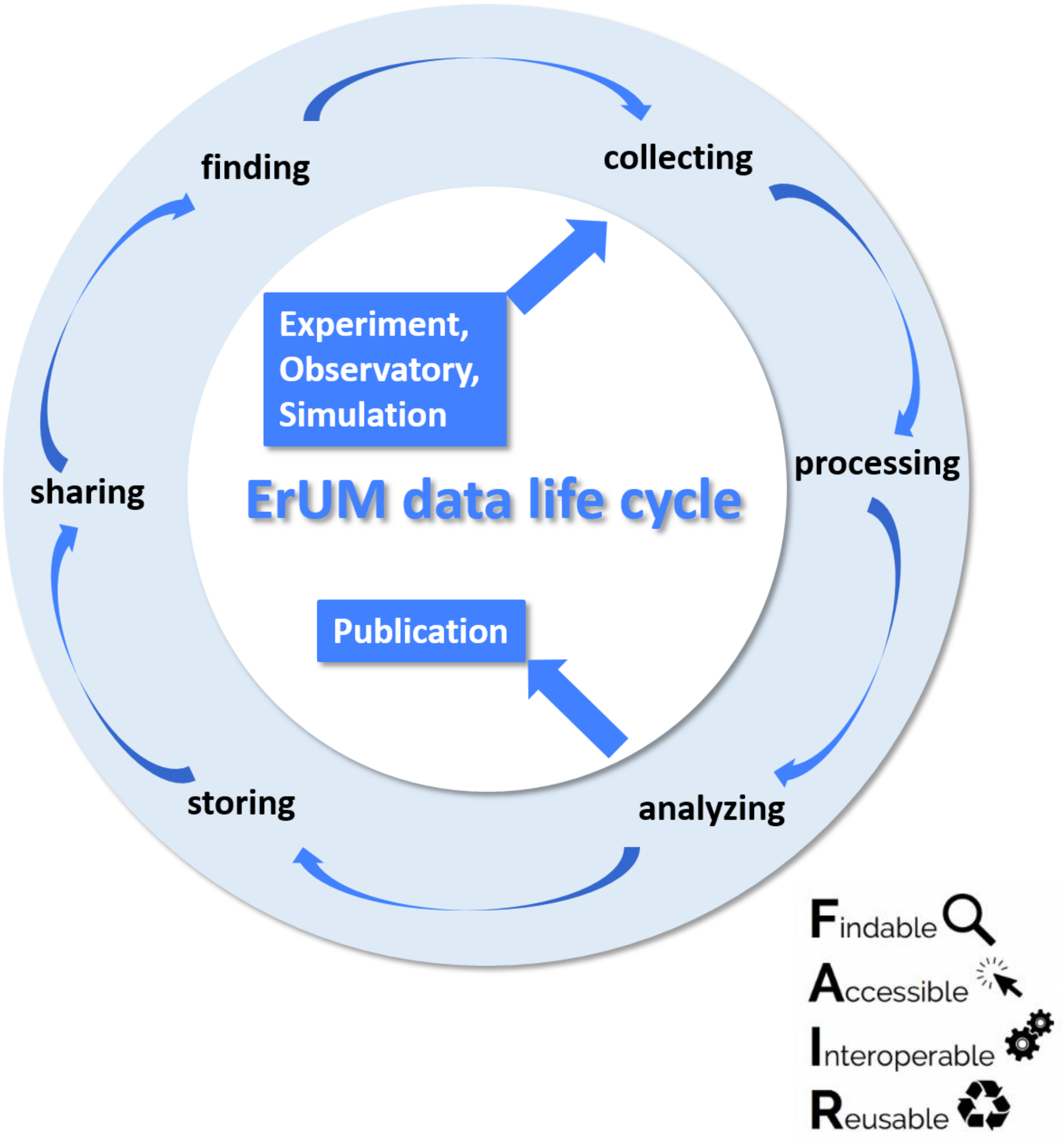
https://swan.cern.ch
https://vispa.physik.rwth-aachen.de
https://stash.desy.de/projects/B2T/repos/b2-starterkit

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**

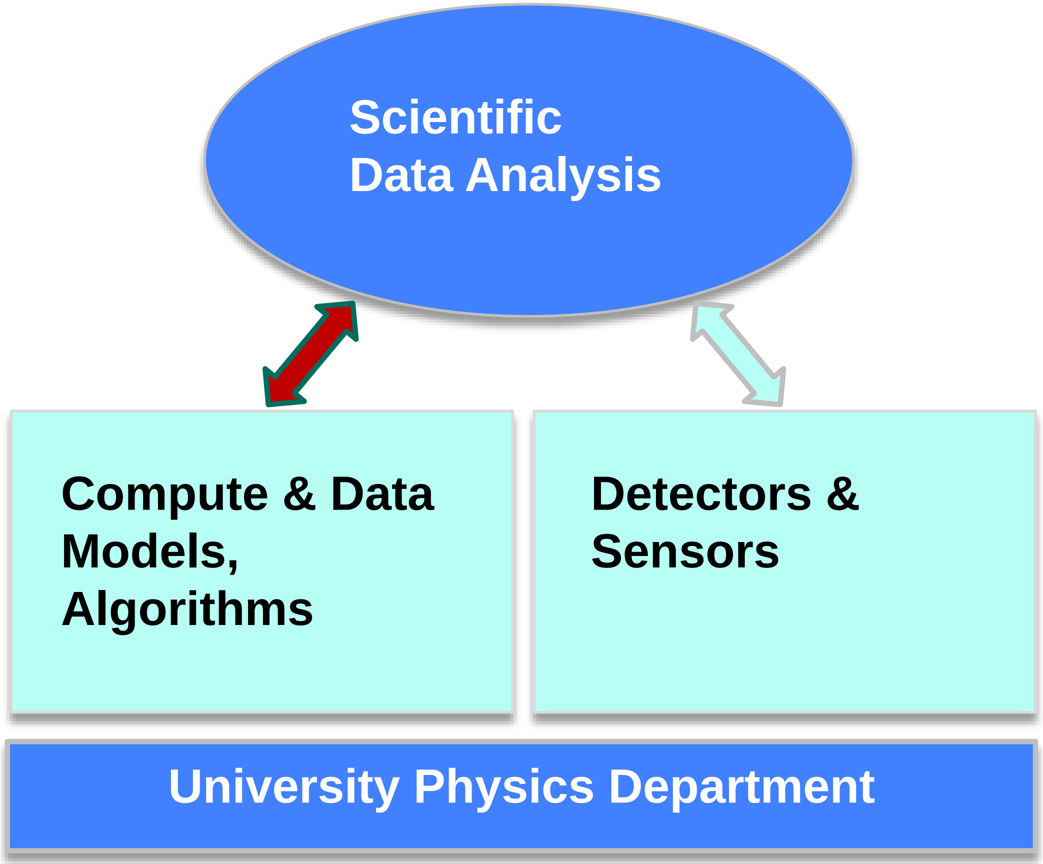




Tenure Track Programm

- **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:

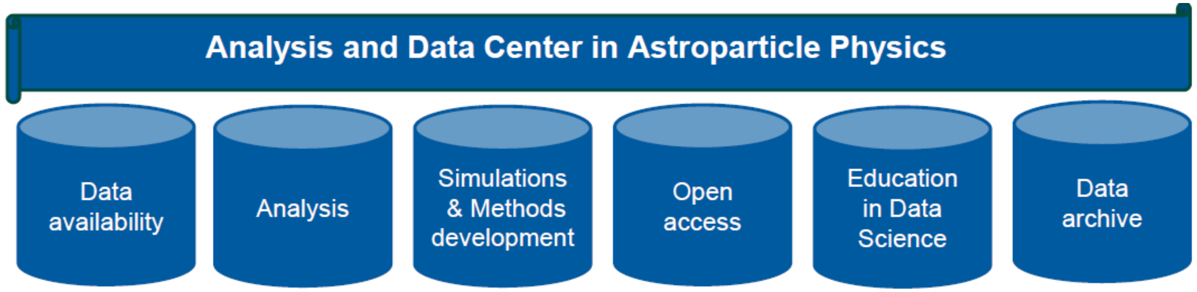
- user-led home to bundle and steer the activities

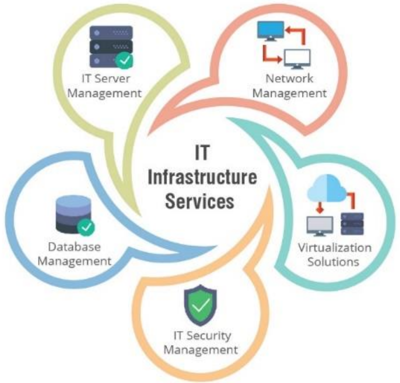
Partnership
Innovative Digitization in ErUM

Federated Infrastructures	Enabling Technologies	Sharing Knowledge
Storage	Big Data Analytic Tools	Workshops
Fast Network	Infrastructure Services	Schools
Compute Power	Research Data Management	Competence Multiplicators

Common Governance Structure
with Connections to Other Communities and International Partners

Era of Digitization in Astroparticle Physics





Lots of activities in

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

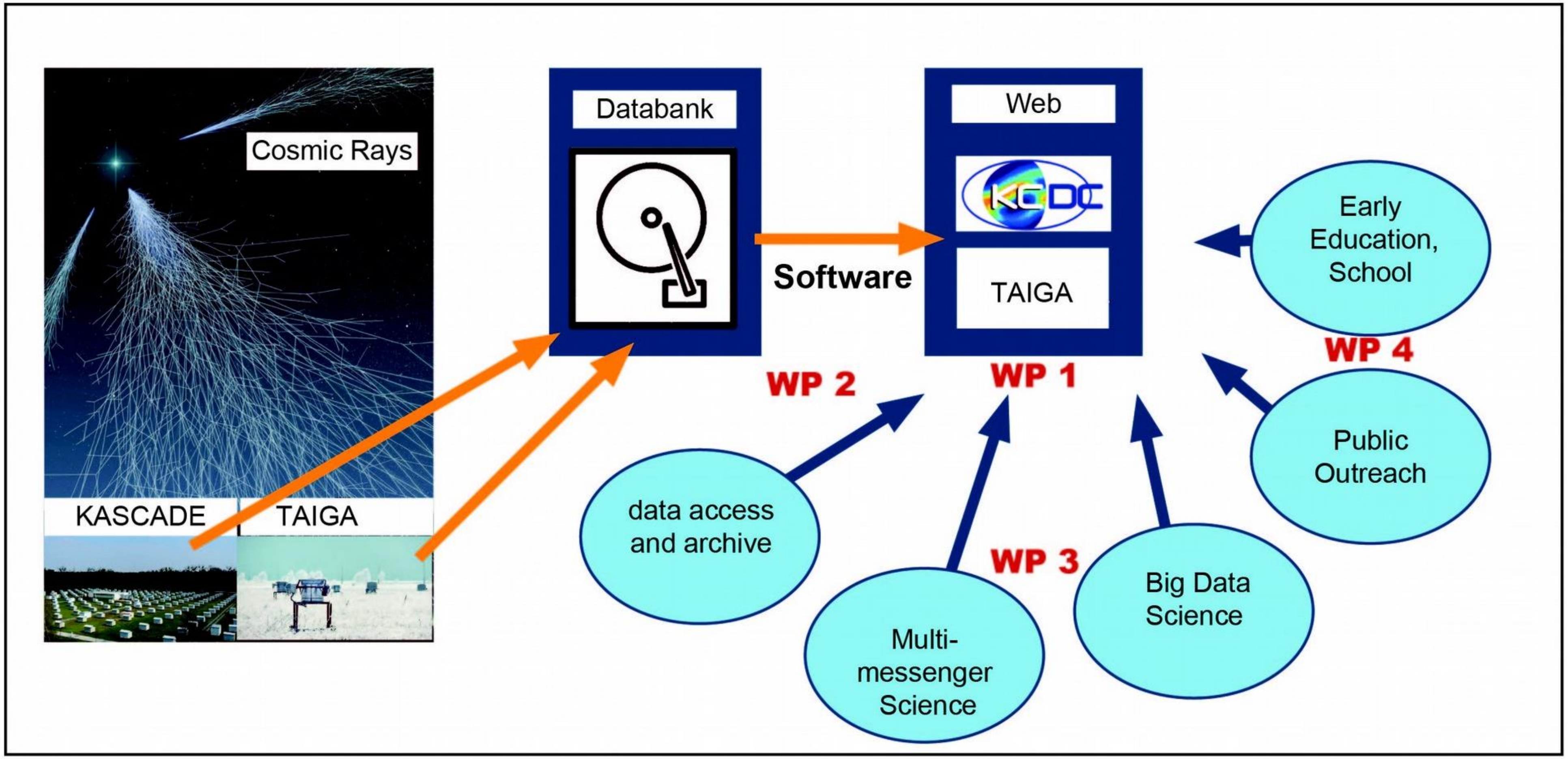
Data Life Cycle

Next steps:

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

GRADLCI: work packages

1. WP1: KCDC extension
2. WP2: Big Data Science Software
3. WP3: Multi-Messenger Data Analysis
4. WP4: Go for the public



Karlsruhe Institute of Technology

GRADLCI: work packages

1. **WP1: KCDC extension**

We will extend KCDC by scientific data from the TAIGA (Tunka) experiment. Further goal is to improve KCDC and make it more attractive to a broader user community.

2. **WP2: Big Data Science Software**

We will develop specific analysis methods and corresponding simulations in the new environment which needs a move to most modern computing, storage and data access concepts (“Data Life Cycle Lab”).

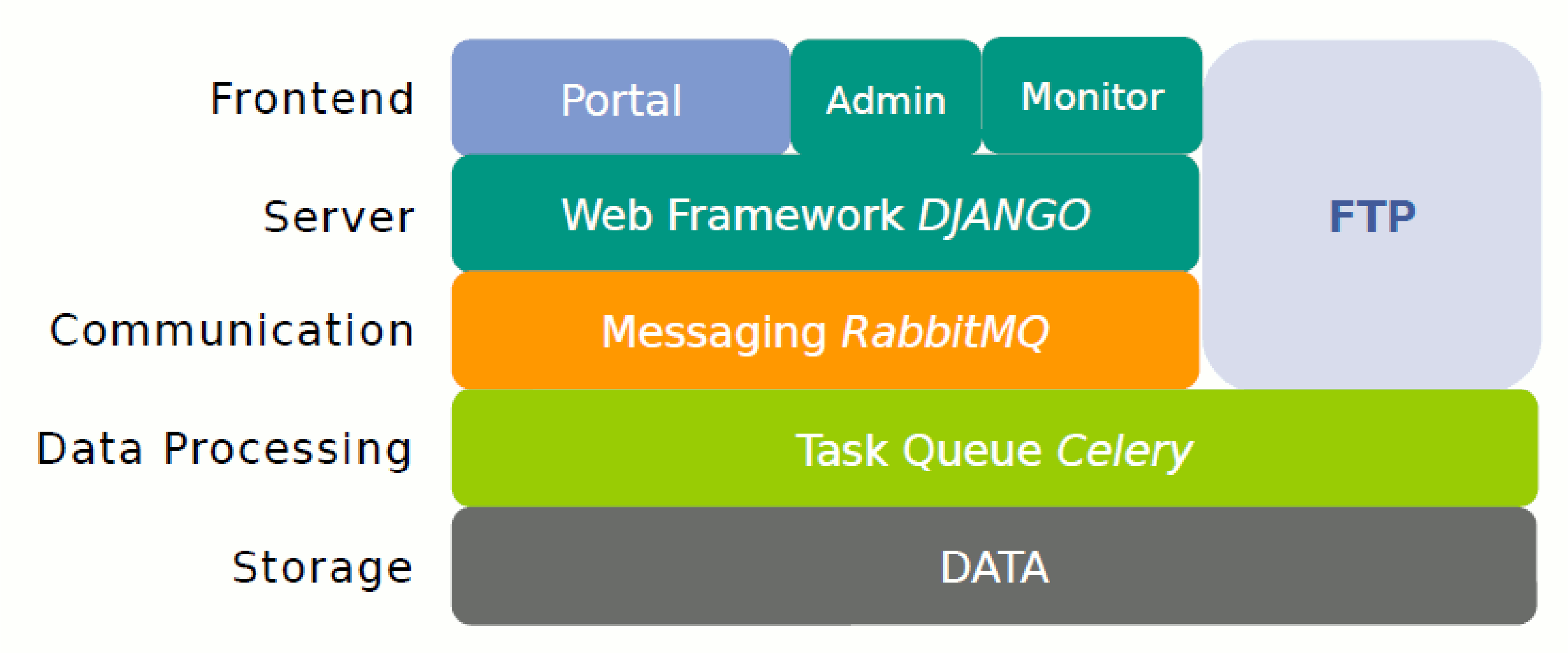
3. **WP3: Multi-Messenger Data Analysis**

We will perform specific analyses using the new data centre to test the concept. This will give confidence to the facility as a valuable scientific tool.

4. **WP4: Go for the public**

A comprehensive outreach is part of the project for all level of users - from pupils to the directly involved scientists to theoreticians.

WP1: KCDC extension



Frank, Jürgen, Doris are working on this…..

- Software extension of KCDC to allow for a new databank and data shop (KIT-IKP, KIT-SCC, ISDCT)
- Preparing and providing the new data for inclusion into KCDC (ISU, MSU)
- Putting the new data into KCDC (KIT-IKP)

WP2: Big Data Science Software

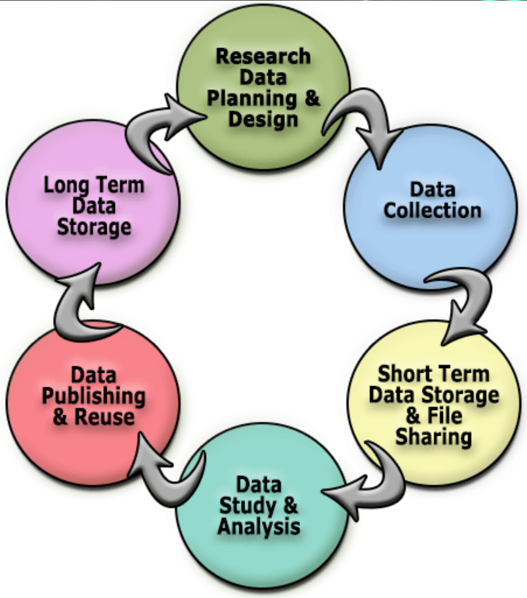
- Movement of KCDC to large-scale computing facility and adapting the new environment (KIT-SCC, KIT-IKP, MSU-SINP)
- Optimizing data bank and access interfaces (MSU-SINP, ISDCT, KIT-SCC)

Frank, Vika

- A distributed system of storage and archiving the data is developed (MSU-SINP, KIT-SCC)

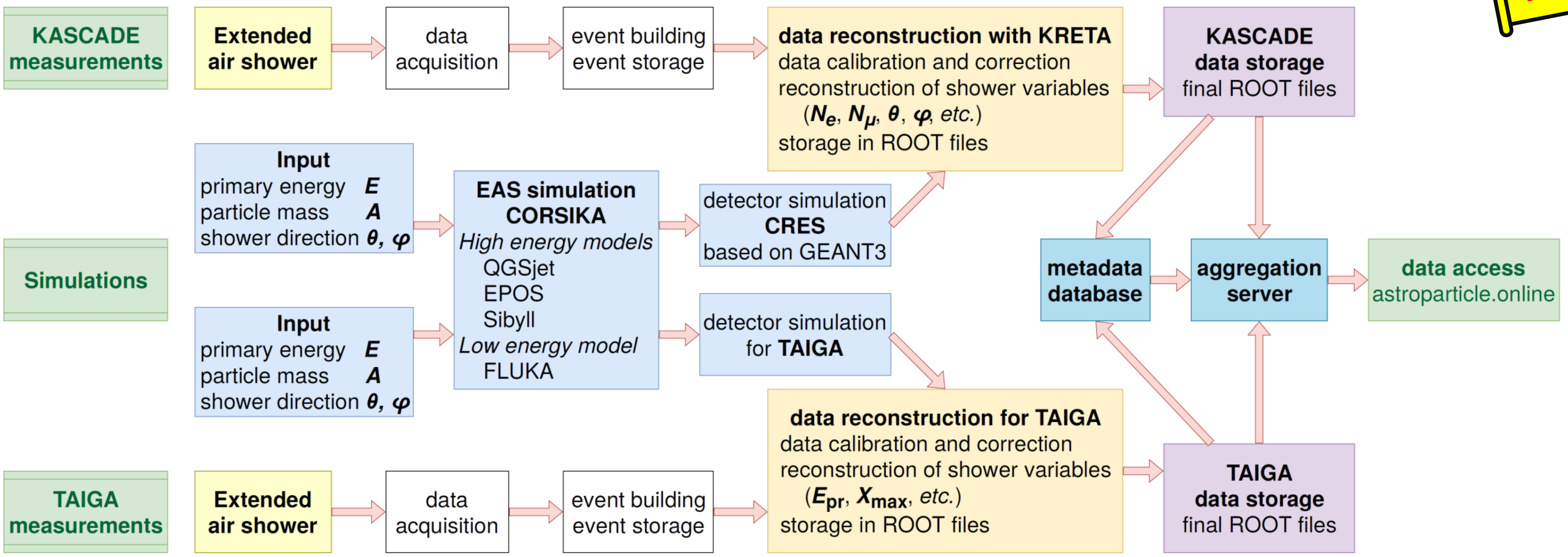
MSU

- Installation of appropriate hardware (KIT-SCC)
- Installing the Data Life Cycle Lab" (KIT-SCC)



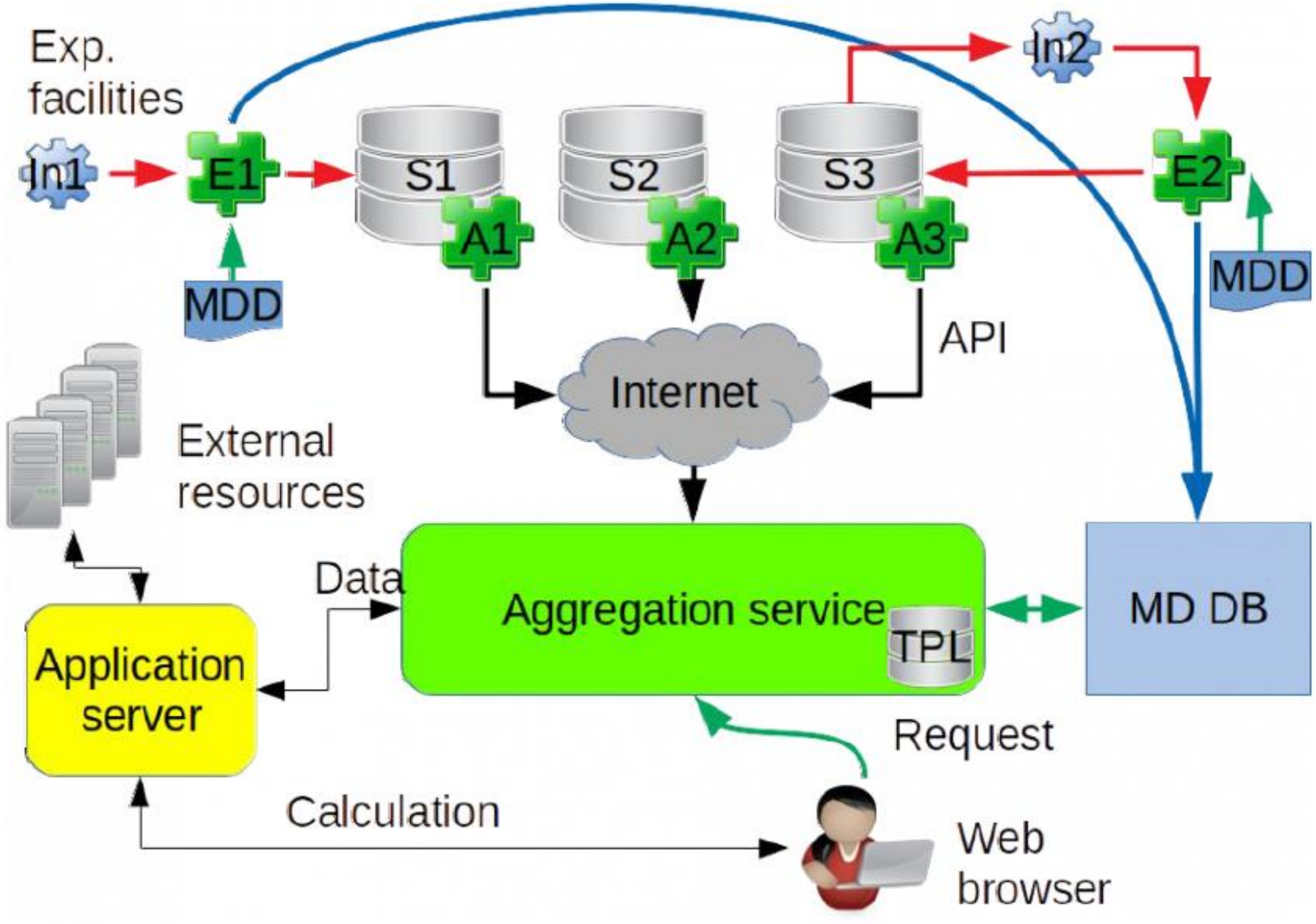
Proposed joint data workflow

Vika



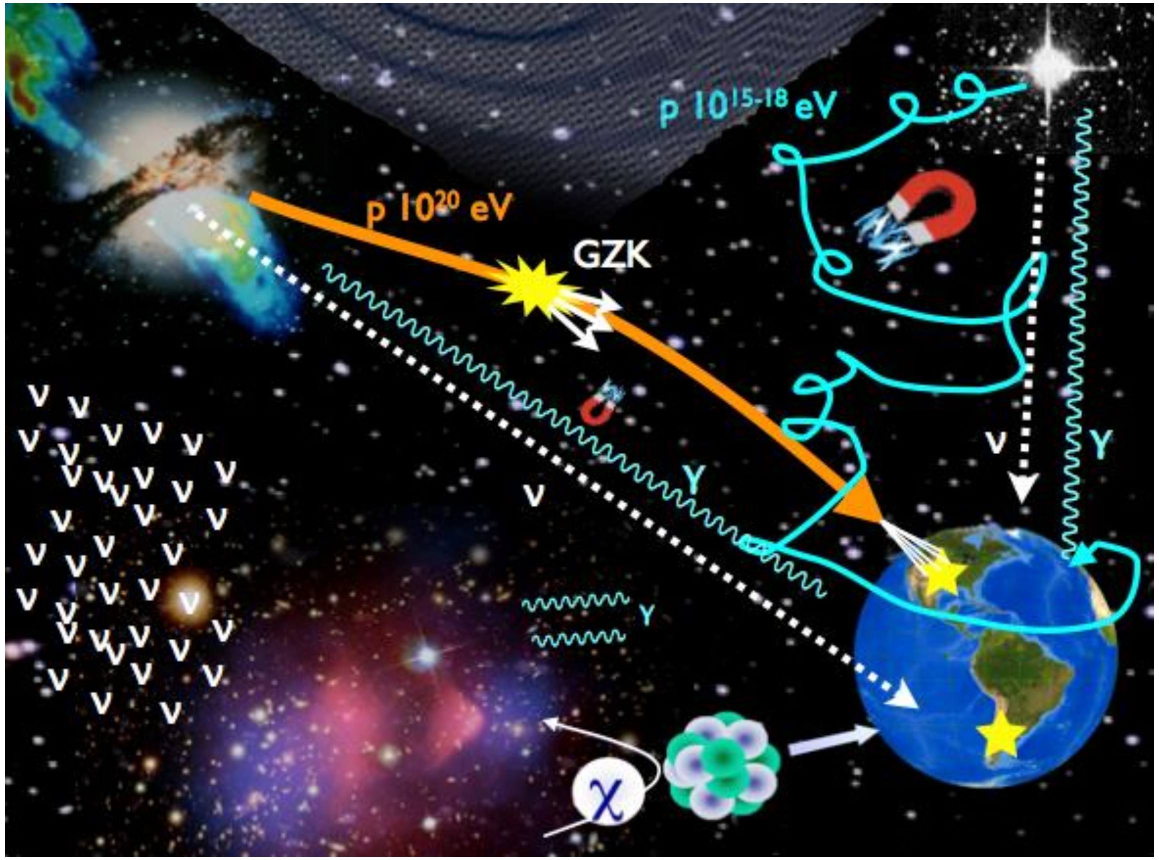
* Granted by RSF-Helmholtz Joint Research Groups. The work is supported by KSETA.

Data Workflow



Vika, MSU…

WP3: Multi-Messenger Analysis



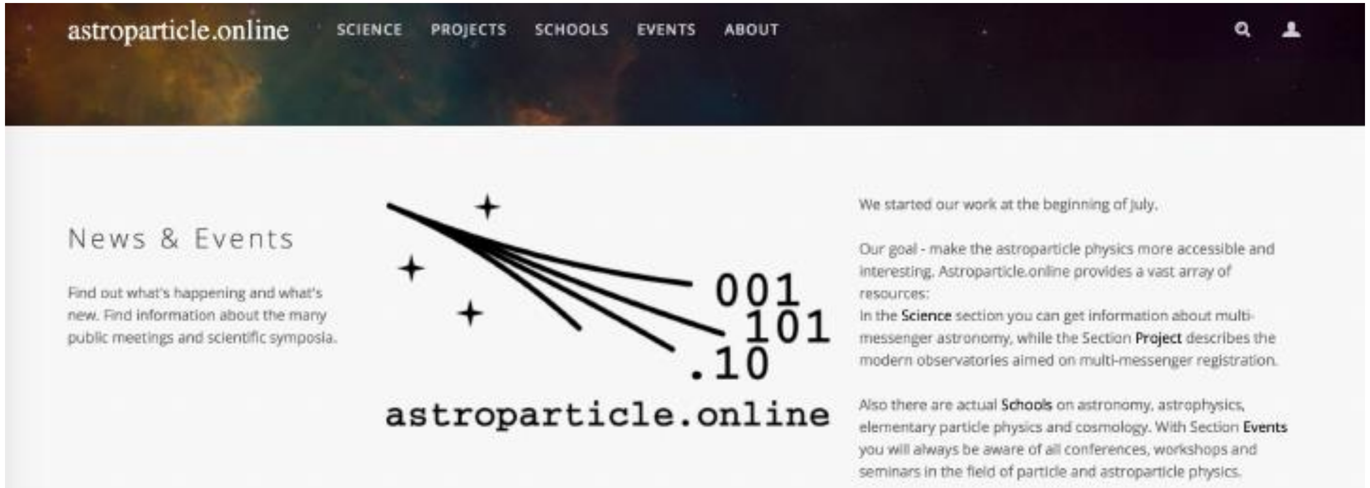
- Defining appropriate physics questions, where the data centre is used (KIT-IKP, MSU-SINP, ISU)
- developing new methods like machine learning, etc. (all)
- Cross-checks of the reliability of all the specific user functions (KIT-IKP, MSU-SINP, ISU)
- Performing the combined Tunka – KASCADE data analysis (KIT-IKP, ISU)
- Performing the multi-messenger data analysis (ISU, MSU-SINP)

Progress by many…too slow?

- Examples:
  - Gamma-ray search
  - Hadronic interaction models
  - Radio cross-calibration

WP4: Go for the public

- comprehensive outreach (KIT-IKP, MSU-SINP, ISU)
  - KCDC public relations
  - publications (papers)
  - conference attendance
  - astroparticle.online
  - workshops



Progress ….
Should we start an outreach campaign?

This Meeting:

- Status of present activities
- Alignment of activities
- Organisationals
- Next steps

Dear Andreas Haungs,

The below submission has been selected as a POSTER PRESENTATION at the upcoming International Cosmic Ray Conference.

Please let us know by email to icrc2019@wisc.edu if you are unable to present this poster.

Please note that papers for the proceedings will be due ahead of the meeting. Directions will be sent via email in a couple weeks.

You may complete your registration at ICRC Registration site.

Speaker Presentations

Title	German-Russian Astroparticle Data Life Cycle Initiative
Paper Status	Accepted
Presentation Type	Poster Presentation
Theme	Cosmic Ray Indirect
Presenting Author	Andreas Haungs Affiliations: Karlsruhe Institute of Technology