KRAD: Karlsruhe-Russian Astroparticle Data Life Cycle Initiative – the KIT view

KRAD/APPDS Kick-off Meeting
Moscow, 5 - 6 March 2018

Andreas Haungs

KRAD: Karlsruhe-Russian Astroparticle Data Life Cycle Initiative

[v] Information and Data Science

Project Partners

	Name and affiliation of the German Principal Investigator	Name and Affiliation of the Russian Principal Investigator
Name, title:	Dr. Andreas Haungs	Prof. Kryukov, Aleksander Pavlovich
Host Institution:	KIT	SINP MSU
Address	Institut für Kernphysik Hermann-von-Helmholtzplatz 1 76344 Eggenstein-Leopoldshafen Germany	M.V.Lomonosov Moscow State University, Skobeltsyn Institute of Nuclear Physics 1(2), Leninskie gory GSP-1, Moscow 119991 Russian Federation
Contact telephone number and E-Mail address:	0049-721-60823310 andreas.haungs@kit.edu	+7(495) 939 1818 kryukov@theory.sinp.msu.ru

Recommendations of the KAT (white paper)

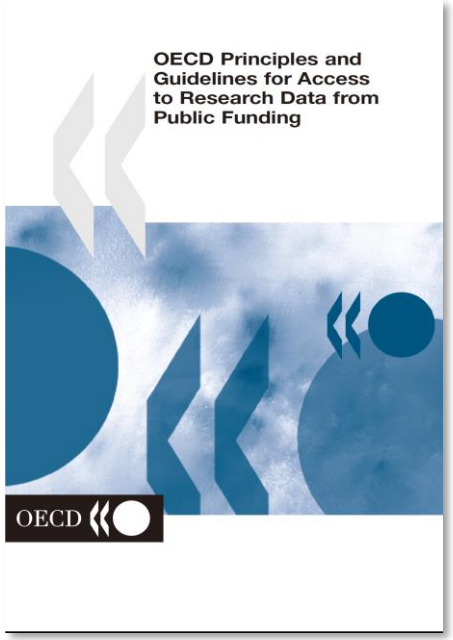
„Astroparticle Physics in the Light of the Digitalen Agenda der Bundesregierung*“

Recommendations of the KAT
The KAT emphatically emphasises the importance of setting up and developing centres for data storage, the provision of data and the necessary computing resources as a basic digital service for German scientists and, moreover, for public participation in scientific data.
The KAT supports the establishment of a structure that facilitates communication between scientists as users of scientific data and modern data analysis methods on the one hand, and continues to implement expert advice within the framework of user support.
The KAT draws attention to the central importance of externally funded and sustainably invested human resources positions, which are absolutely necessary for the support of users.

* https://www.bmbf.de/de/die-digitale-agenda-relevant-auch-fuer-bildung-wissenschaft-und-forschung-206.html

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

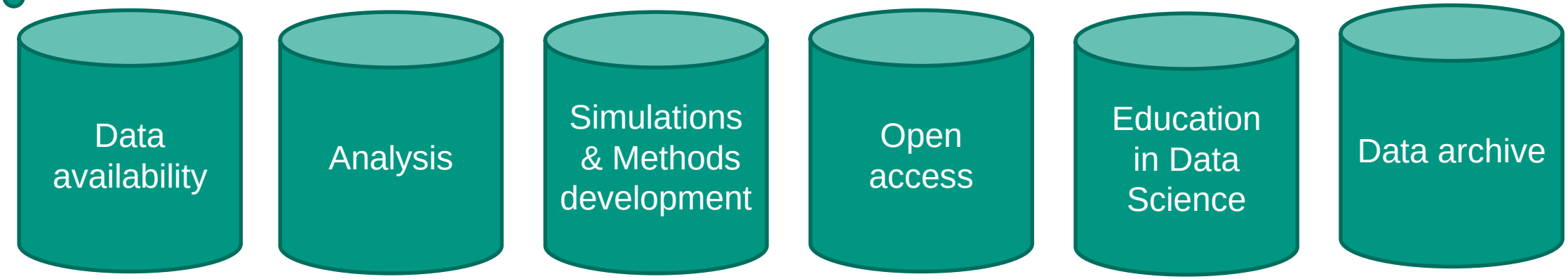
- Astroparticle Physics requests for multi-messenger analyses - this needs an **experiment-overarching** platform!



- **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 Centre 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:**

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!

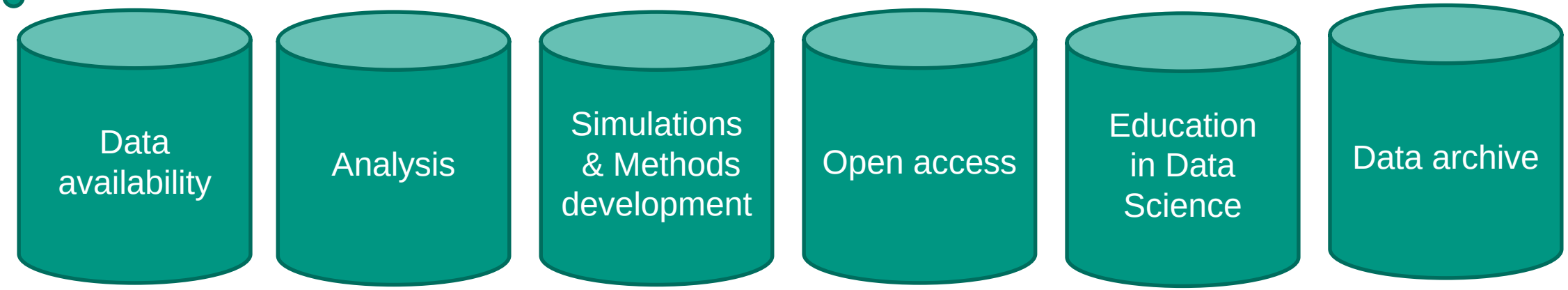
- **Education in data science:**

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

- **Data archive:**

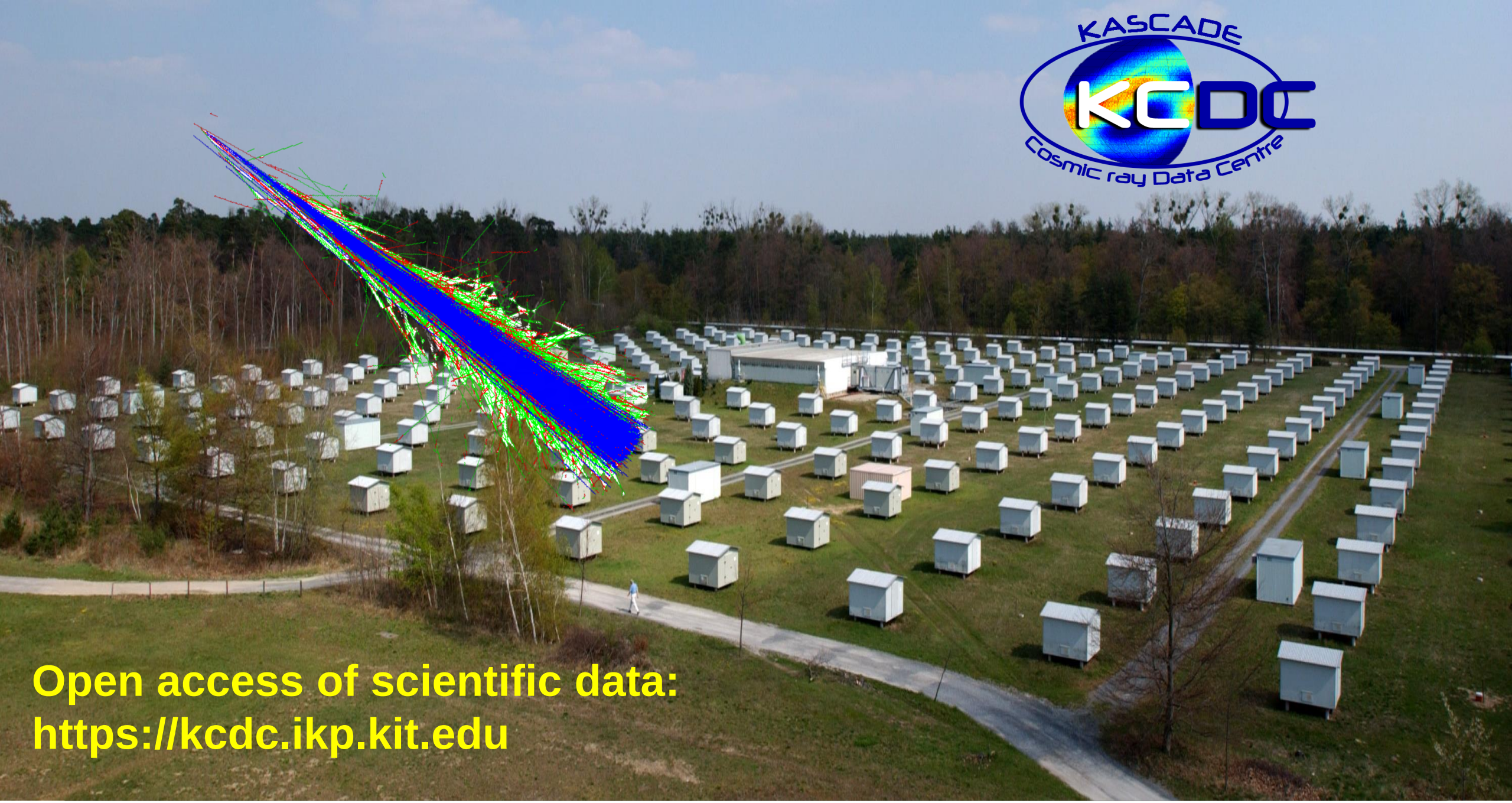
The valuable scientific data and metadata must be preserved and remain interpretable for later use (data preservation).

Analysis and Data Centre in Astroparticle Physics



- **Data preservation** ----
  like DPHEP, KCDC
- **Metadata preservation** ----
  like KCDC
- **Data storage (archive)** ----
  like DPHEP, GridKa
- **Computing services (Grid vs. Cloud)** ---
  like CERN Tier-centres
- **Data access (policy, technology, rate)** ---
  like GridKa, KCDC
- **Training on Data use (maintenance, tutorials)** ---
  like KCDC, VISPA, CDS
- **Data analysis, Simulation, modeling** ---
  like GridKa, advanced VISPA?
- **Data science, workflows** (tools, e.g. deep learning, tutorials) ---
  like VISPA
- **Data publication / Outreach** ---
  like KCDC, masterclasses
- **Data education** ---
  like KCDC, GridKa-school
- **Data exchange** ---
  like AMON, GAVO
- **Data catalogues** ---
  like Re3Data
- **.....**

Partly realized in individual experiments


KASCADE
KCDC
Cosmic ray Data Centre
Open access of scientific data:
https://kcdc.ikp.kit.edu

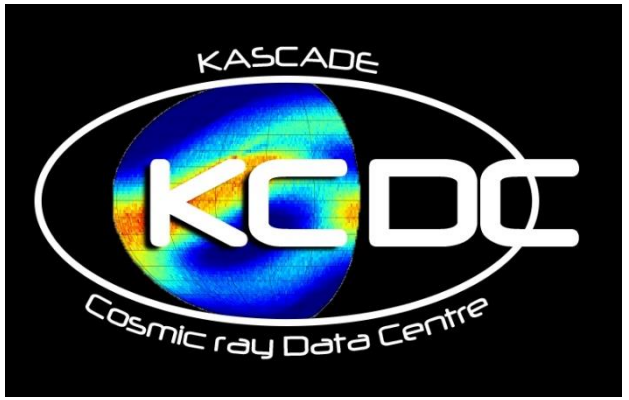
KCDC in a nutshell

- providing open access to astroparticle physics research data as required by funding agencies

#KCDC_KIT

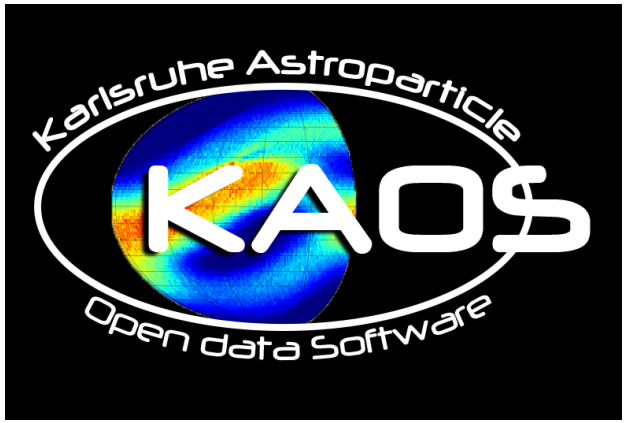
- **data provider**
  - follows the “Berlin Declaration on Open Data and Open Access”
  - free, unlimited, open access to KASCADE cosmic ray data
  - selection of fully calibrated quantities and detector signals
  - reliable data source
  - guaranteed data quality

The web portal



- **information platform**
  - experiment description
  - meta information for data analysis
  - physics background
  - use of modern and open source web technologies
  - tutorials (focused on teachers and pupils)

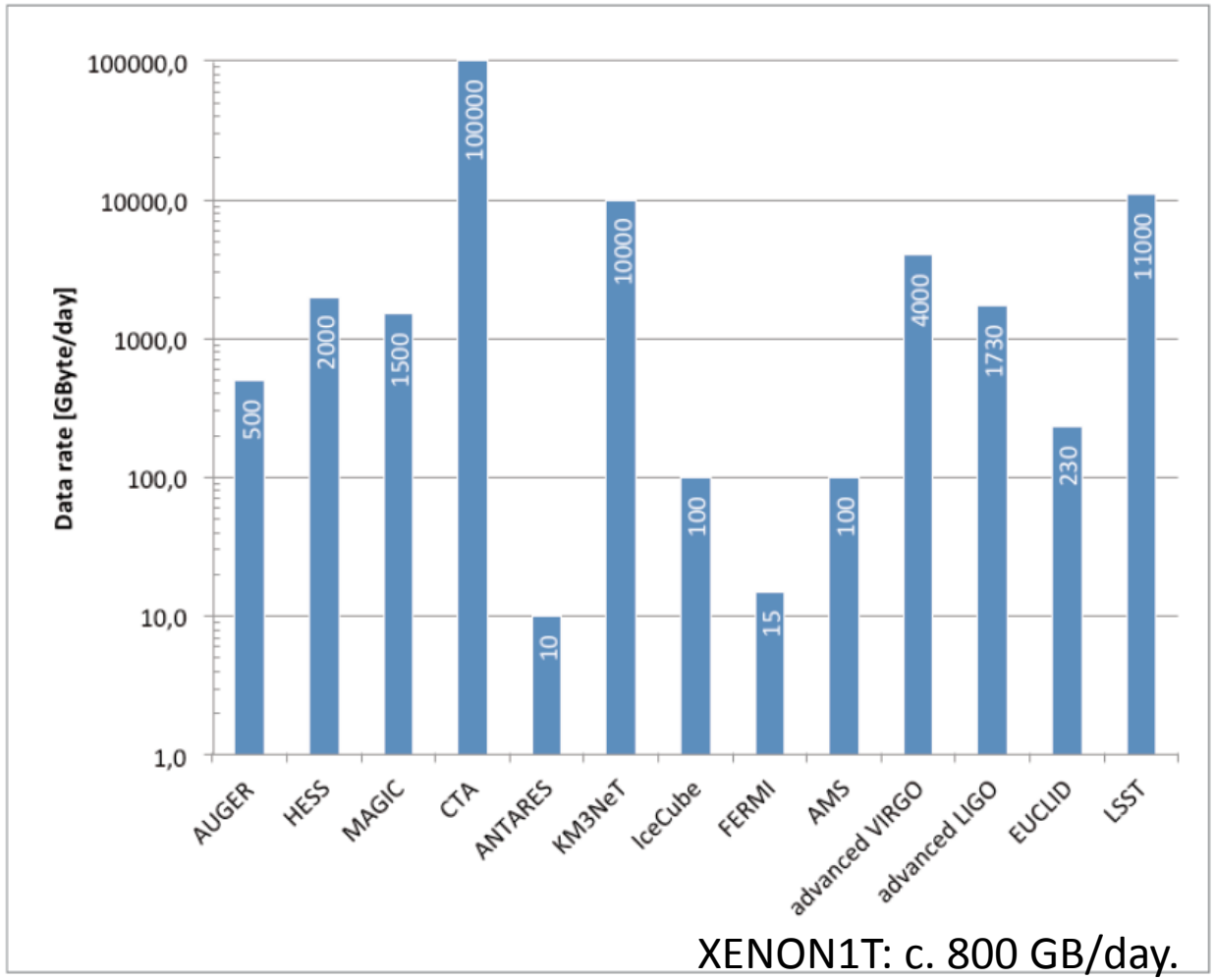
KCDC software



- **as long-term digital data archive**
  - archive of software and data
  - for the collaboration
  - for the public

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

Computing in Astroparticle Physics (Astro-Grid / Astro-Cloud)



Source: APPEC brochure on Computing, 2016

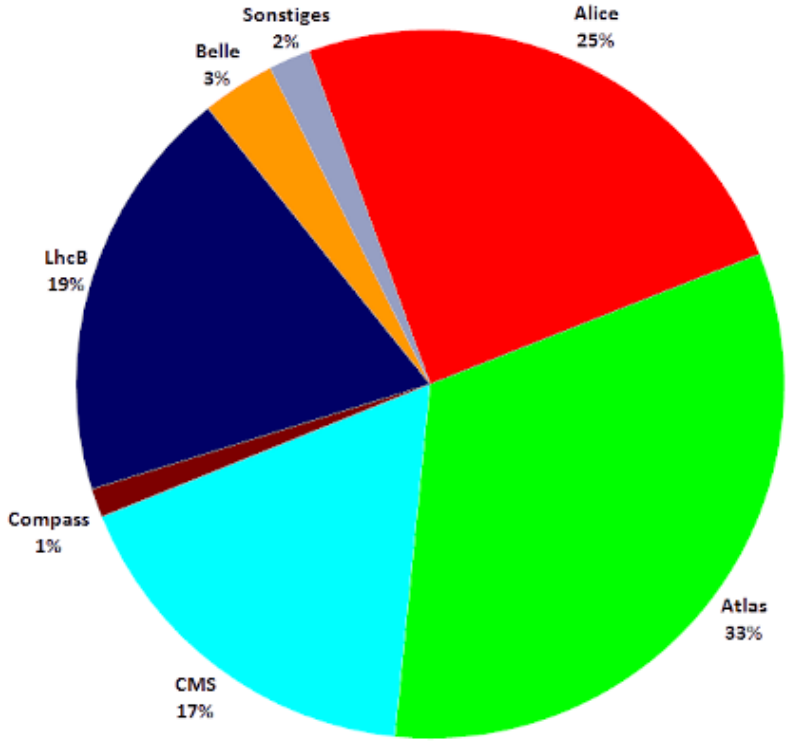
➔Do we need an own Astroparticle Physics computing infrastructure?

- **independent of particle physics?**
- **Grid or Cloud or other technology?**
- **Use of commercial provider (amazon, google, …)?**

Example Particle Physics: GridKa

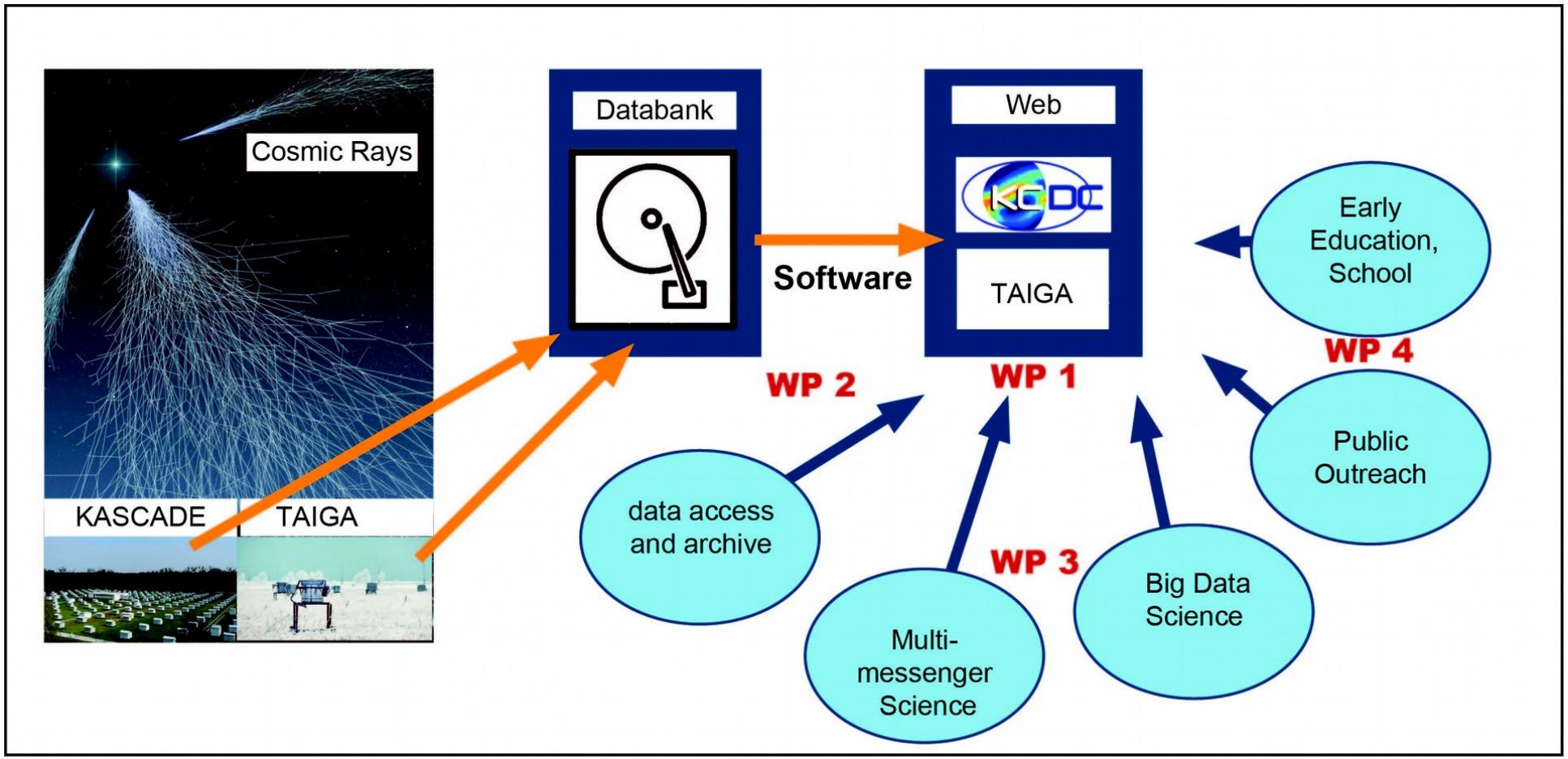
GridKa:

- Central German data and computing centre for particle (and astroparticle) physics
- Tier1-centre in the world wide LHC Computing Grid
- Provides essential part of the German contribution to the LHC-Computing
- Supports non-LHC-experiments with German participation (e.g. Belle-II, Compass und AUGER).
- Is part of the Helmholtz Data Federation
- Governance via the GridKa-Overviewboard (OB) and Technical Advisory Board (TAB)
- LK-II in Program MU, FB Matter
- At Helmholtz/SCC: Concept of Data Life Cycle Labs to initiate Big Data facilities to other science fields.



KRAD: work packages

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



KRAD: 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”).

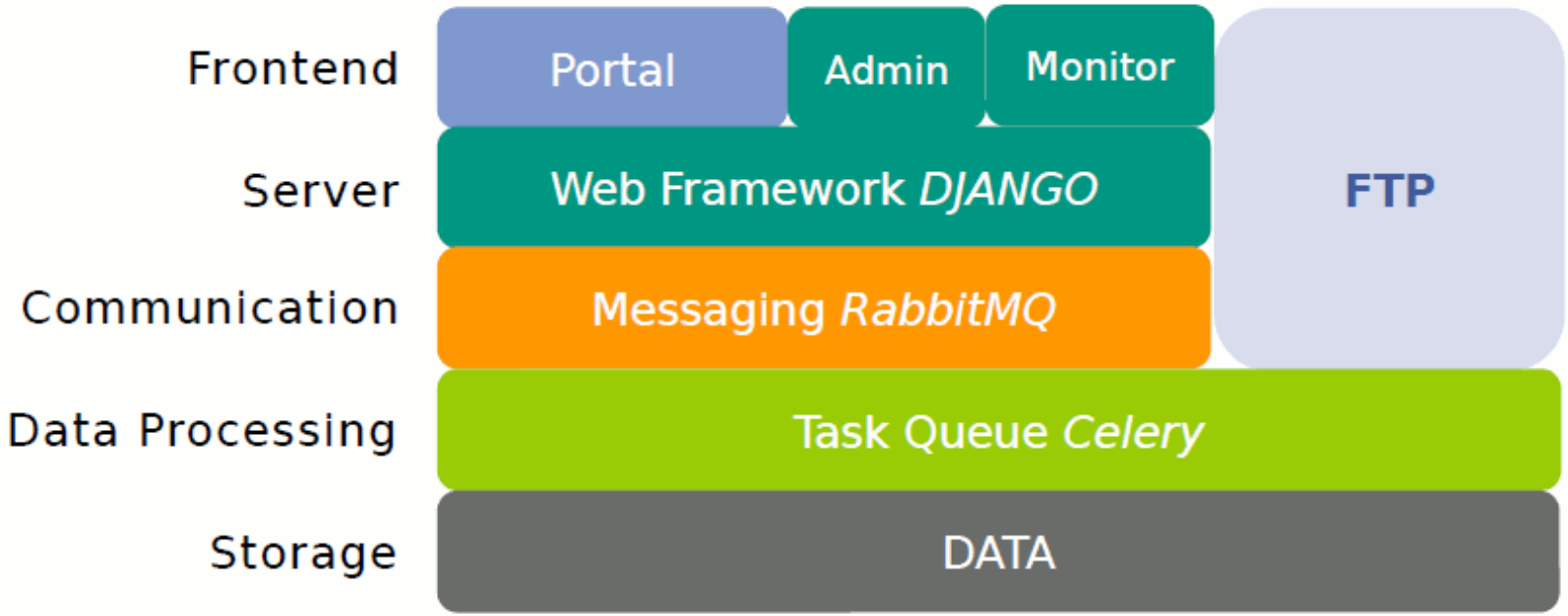
3. WP3: Multi-Messenger Data Analysis

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.

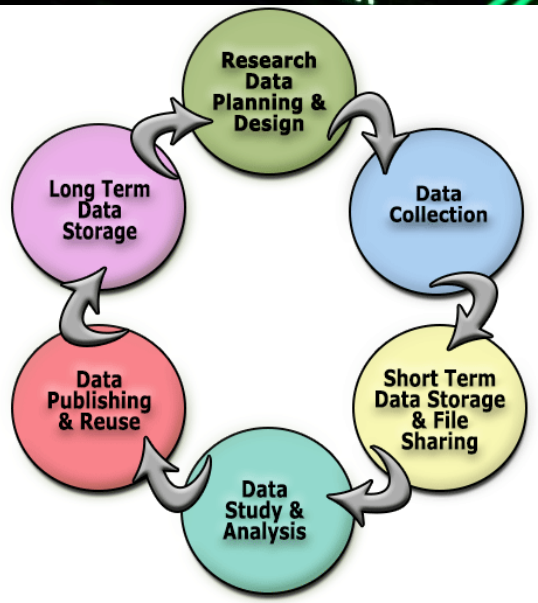
WP1: KCDC extension



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

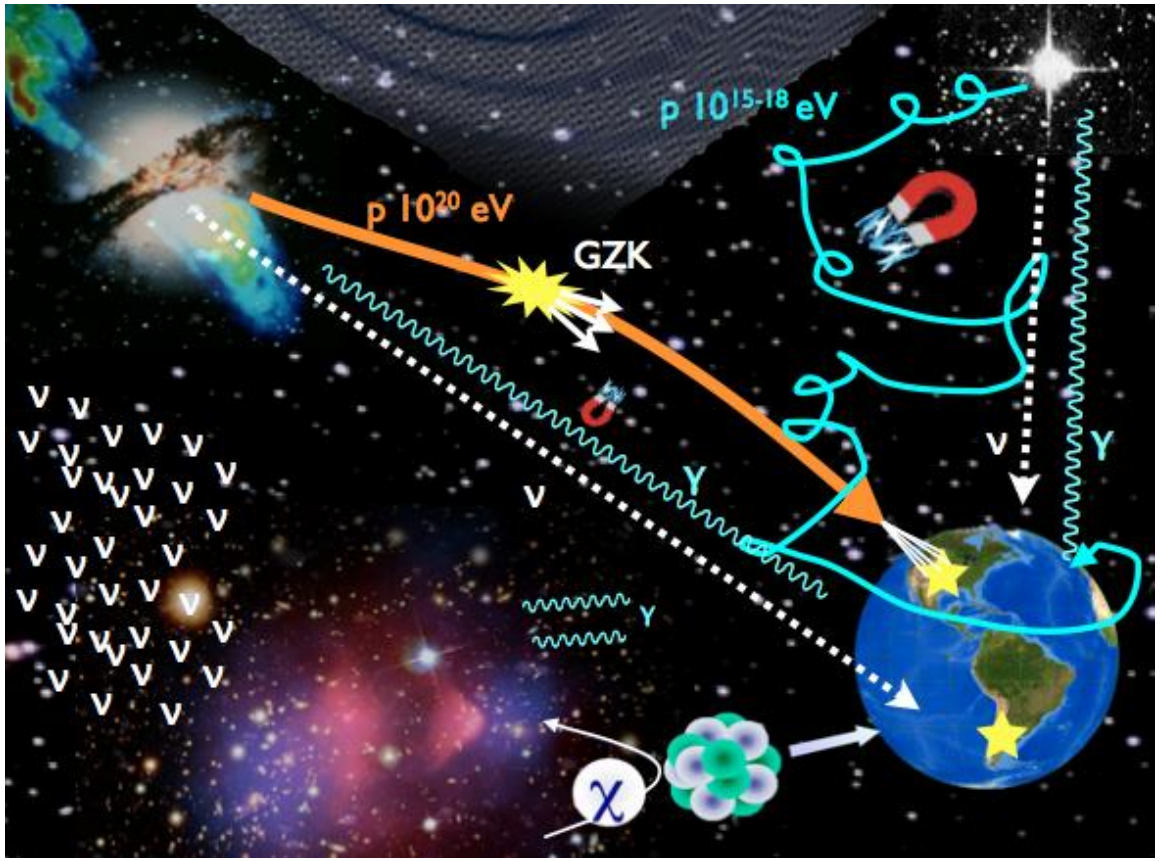
WP2: Big Data Science Software

- 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)
- A distributed system of storage and archiving the data is developed (MSU-SINP, KIT-SCC)
- Installation of appropriate hardware (KIT-SCC)
- Installing the Data Life Cycle Lab” (KIT-SCC)



WP3: Multi-Messenger Analysis

- Defining appropriate physics questions, where the data centre is used (KIT-IKP, MSU-SINP, ISU)
- Cross-checks of the reliability of all the specific user functions (KIT-IKP, MSU-SINP, ISU)
- Performing the combined TAIGA – KASCADE data analysis (KIT-IKP, ISU)
- Performing the multi-messenger data analysis (ISU, MSU-SINP)

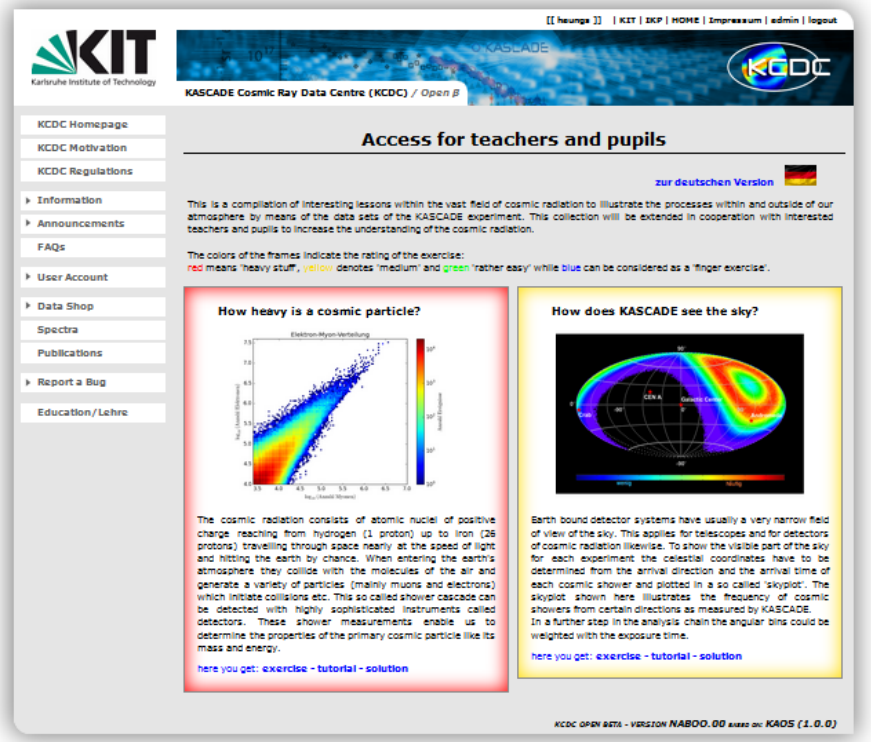


- Examples:
  - Gamma-ray search
  - Hadronic interaction models
  - Radio cross-calibration

WP4: Go for the public

#KCDC_KIT

- Creating adequate web pages for the project, the data centre and the new Data Life Cycle Lab (KIT-IKP)
- Other outreach activities: e.g. press releases related to the project, roll-up posters, brochures etc. for adverting the project in public rooms and events (all)
- Place the project's progress in a variety of social media (KIT-IKP)
- Monitoring of user statistics of the new data centre and the social media (KIT-IKP)

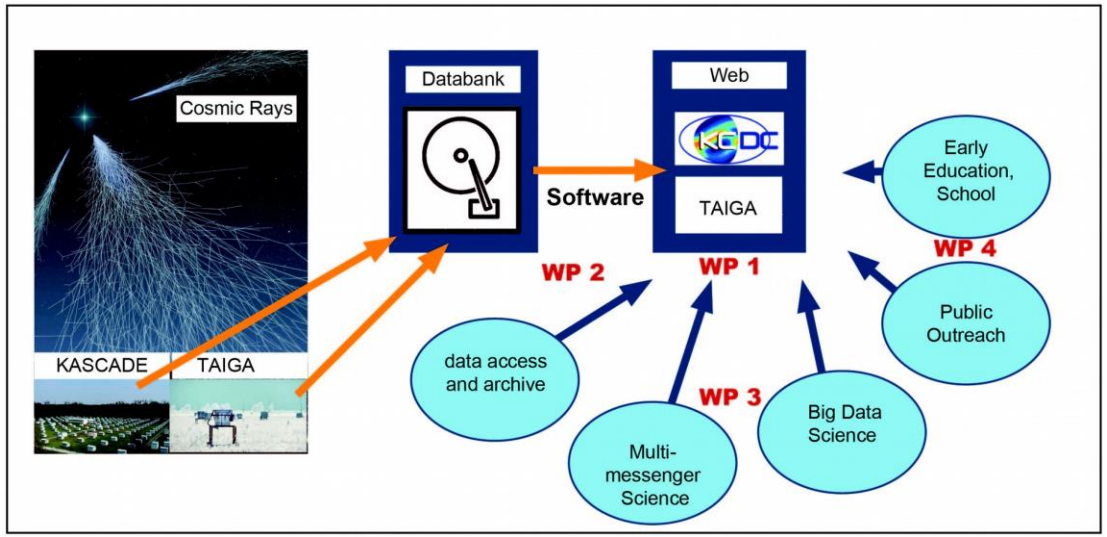


Karlsruhe-Russian Astroparticle Data Life Cycle Initiative

Open PhD positions!
Contact me…. ☺

- **Basics**
  - project period (2018-2020)
  - Russia: SINP MSU, ISU, ISDCT SB RAS
    Germany: KIT, DESY
  - Team leaders: A. Kryukov (SINP MSU)
    and A. Haungs + A. Streit (KIT)

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



KRAD is a first step in extension and generalization of KCDC to a global analysis and data centre for astroparticle physics