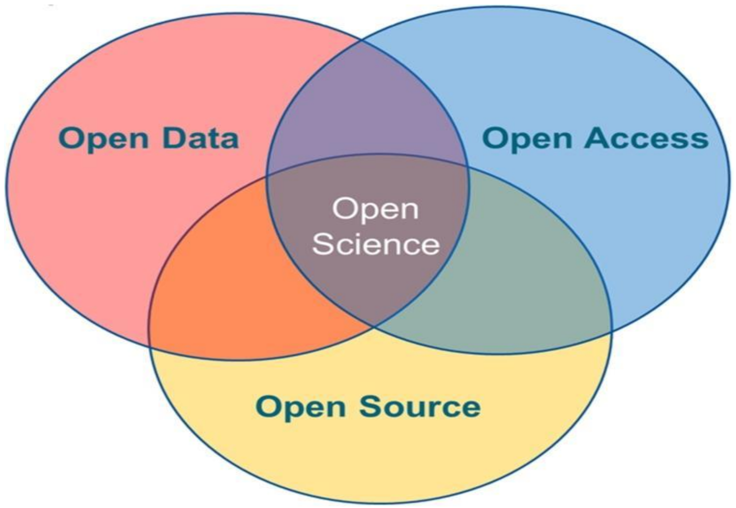
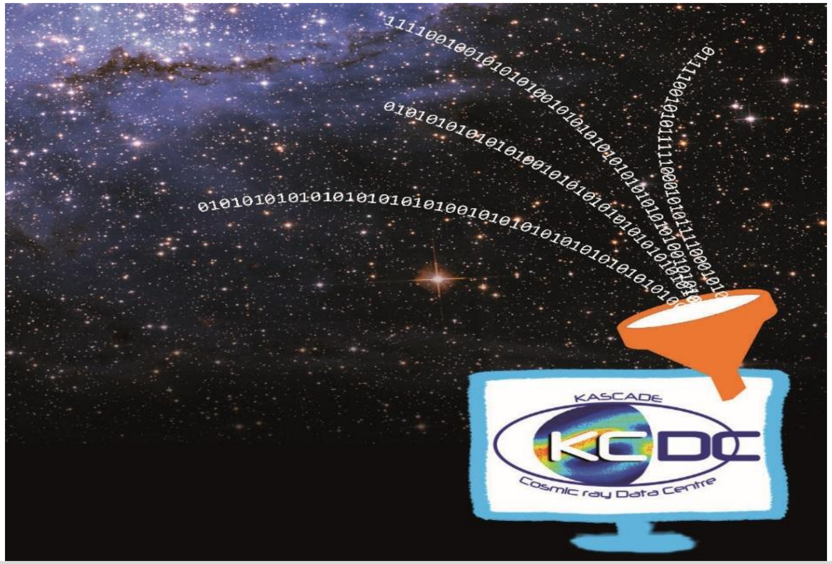
Towards a Global Analysis and Data Centre for Multi-Messenger Astroparticle Physics

IV International Workshop on *Data Life Cycle in Physics*
Zoom, 08-10 June 2020

Andreas Haungs



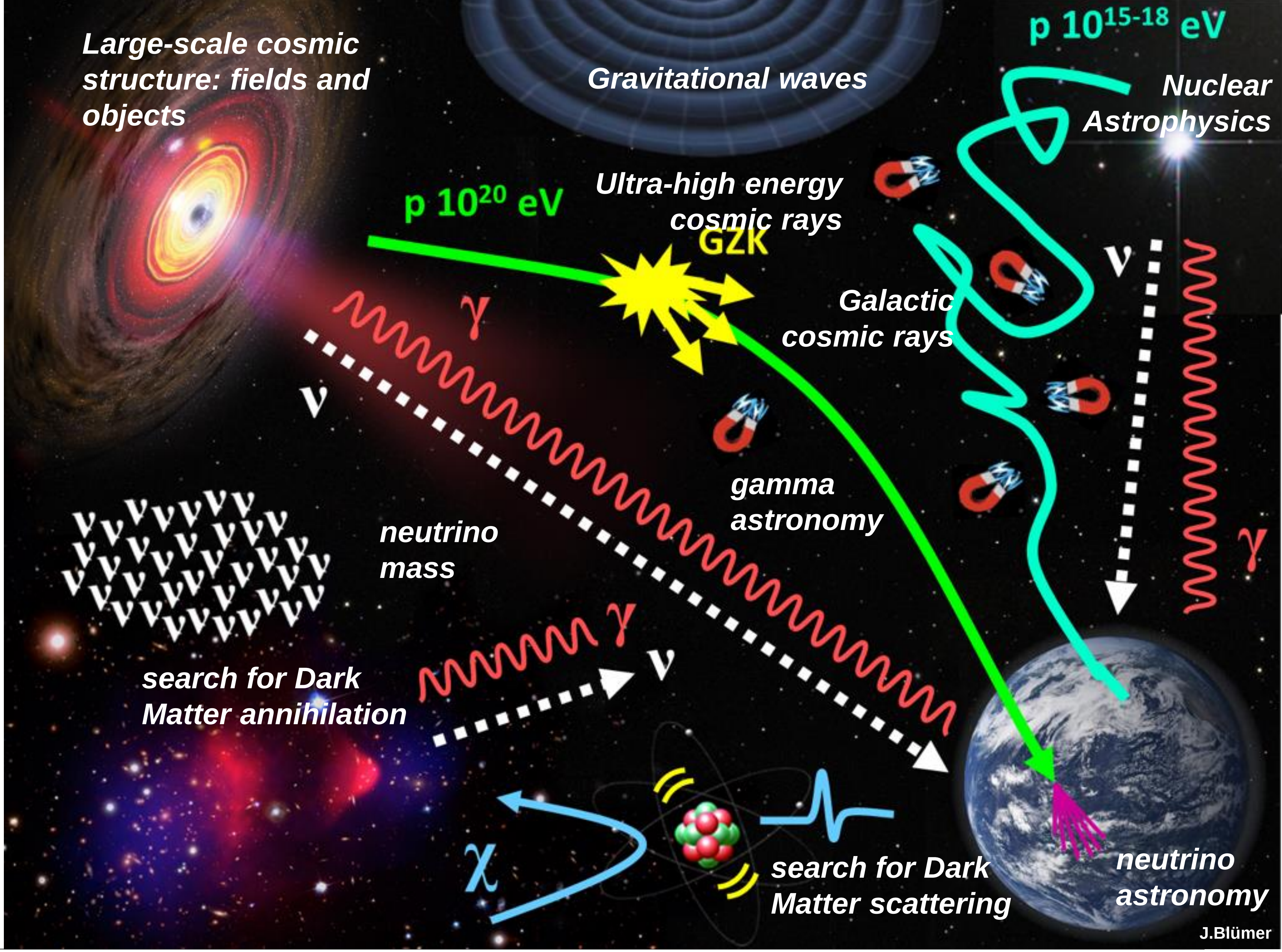


Initiative for a (global) Analysis & Data Center in Astroparticle Physics

Astroparticle Physics = Understanding the

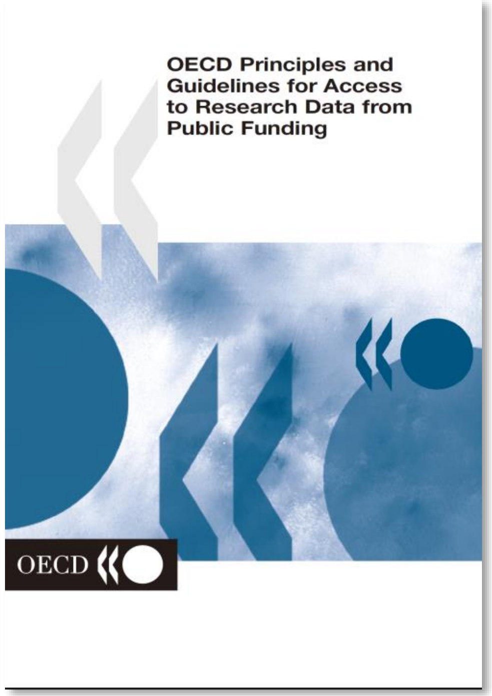
- Multi-Messenger Universe
- Dark Universe

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

Data availability	Analysis	Simulations & Methods development	Real-time analysis center	Open access	Education in Data Science	Data archive

- **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:**

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

Status Infrastructures in Astroparticle Physics

Computing:

- (Co-use of) Institutional resources (partly WLCG resources)
- GridKa: Tier1-centre in the world wide LHC Computing Grid (e.g. Auger@GridKa)
- Experiment-oriented resources (e.g. CTA@DESY)
- Co-use of facility infrastructures (e.g. IceCube at DESY)
- Moderate use of HPC cluster (Gauß Alliance)

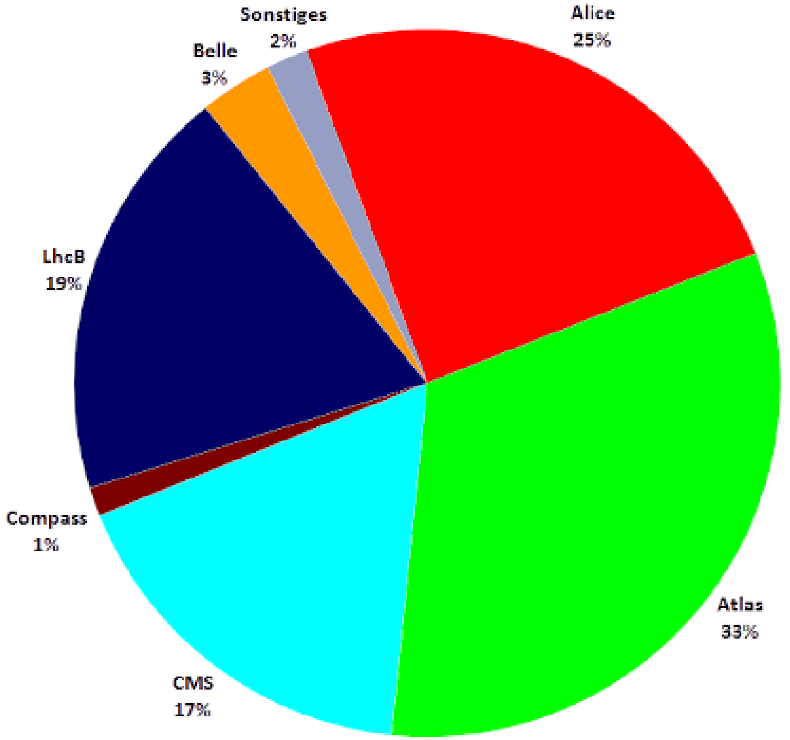
Resarch Data Management:

- KCDC: KASCADE Cosmic ray Data Centre (data access)
- VISPA: to analyze data (Learning Deep Learning)
- GAVO (German Astrophysical Virtual Observatory)
- CERN Open Data Portal (not yet used by APP)

Particle Physics: GridKa (and other Tier-centres)

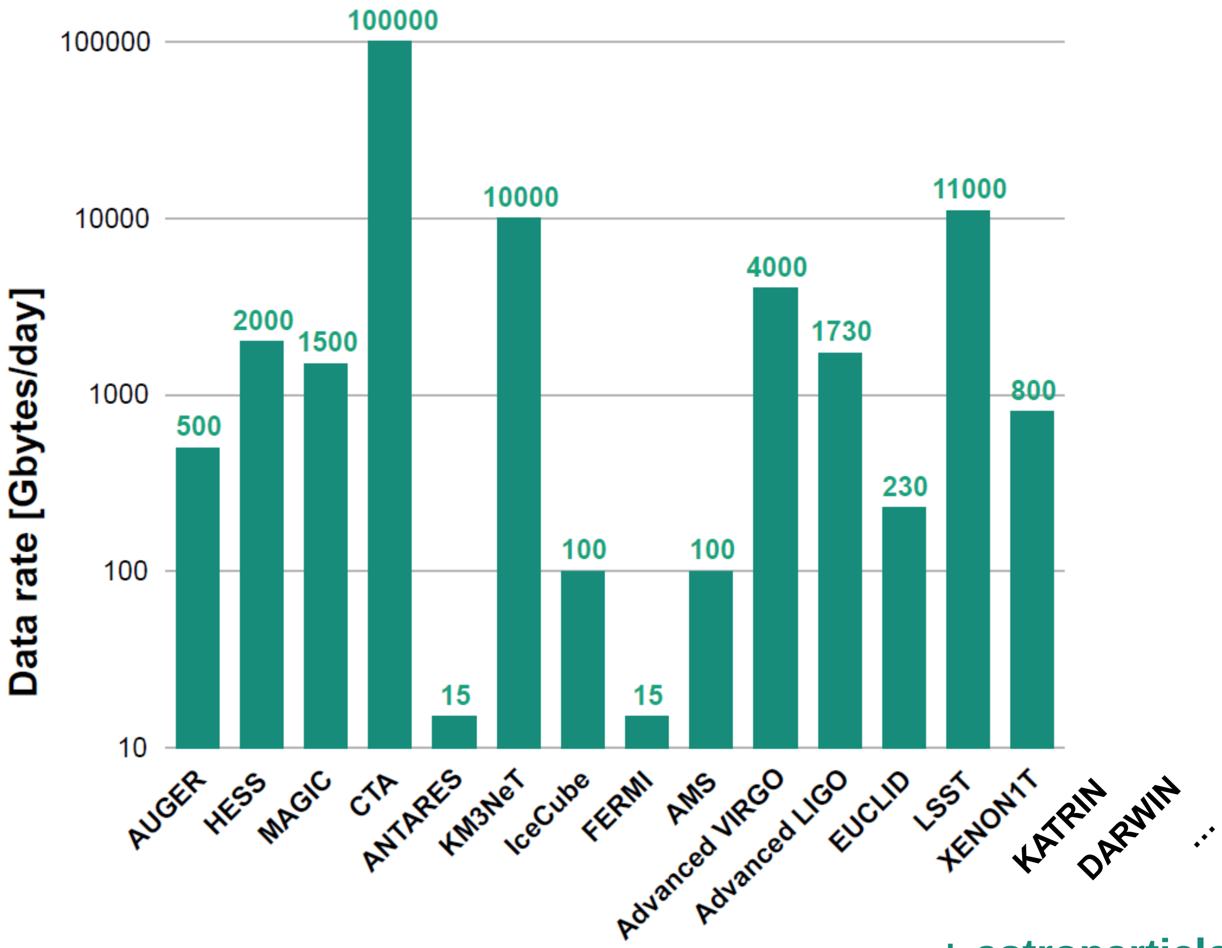
- Central German data and computing centre for particle (and astroparticle) physics
- Tier1-centre in the world wide LHC Computing Grid
- Provides essential part of the German contribution to the LHC-Computing
- Supports non-LHC-experiments with German participation (e.g. Belle-II, Compass and Auger).

Number of cores	28000
Number of compute jobs (last 12 months)	23 million
Number of CPU-hours delivered (last 12 months)	212 million
Disk space	34 PB
Tape space (used)	53 PB



Includes Pierre Auger Observatory
Since 2020: IceCube

Computing in Astroparticle Physics



Source: 2016 APPEC brochure on Computing: Towards a model for computing in European astroparticle physics

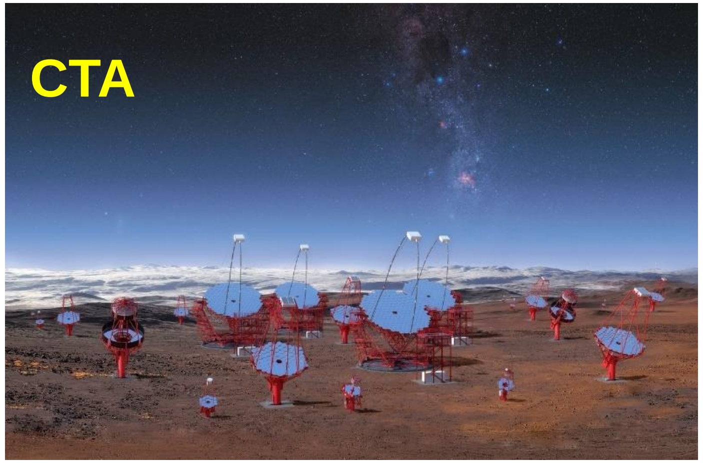
+ astroparticle part of SKA?
+ Einstein Telescope
+ enhanced request from simulations

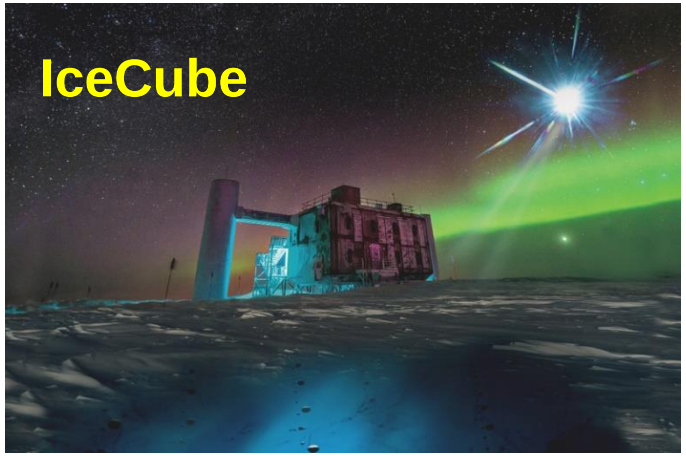
➔Do we need an own Astroparticle Physics computing infrastructure?

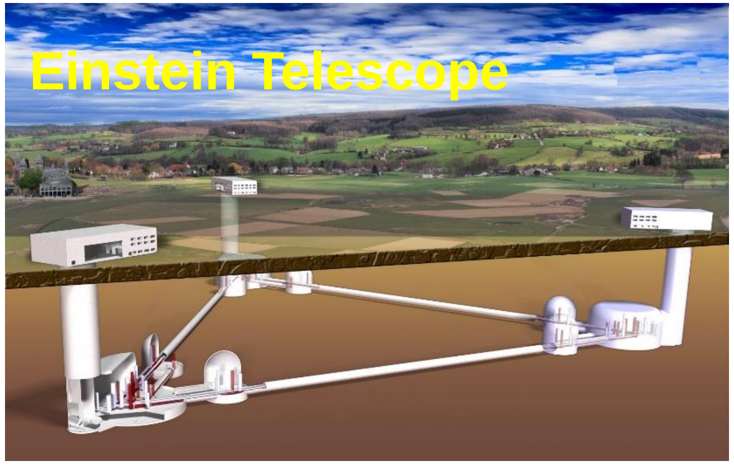
- Synergy with particle physics?
- Grid or Cloud or Lake or other technology?
- Use of commercial providers (amazon, google, …)?
- Is there a relation to NFDI, ErUM-Data, EOSC?

partly organized in the new facilites of astroparticle physics

2020+: Flagship Experiments of German Astroparticle Physics (ErUM-Pro)







Example Computing Model:
CTA Science Data Management Centre

cherenkov telescope array

The Science Data Management Centre will coordinate science operations and make CTA's science products available to the worldwide community.

- **~20 personnel will manage CTA's science coordination including software maintenance and data processing for the Observatory.**
- **CTA will generate approximately 100 petabytes (PB) of data by the year 2030.**
- **The SDMC will be located in a new building complex at DESY in Zeuthen.**
- **Provides well-established infrastructure and a powerful computing centre.**

@ DESY in Zeuthen

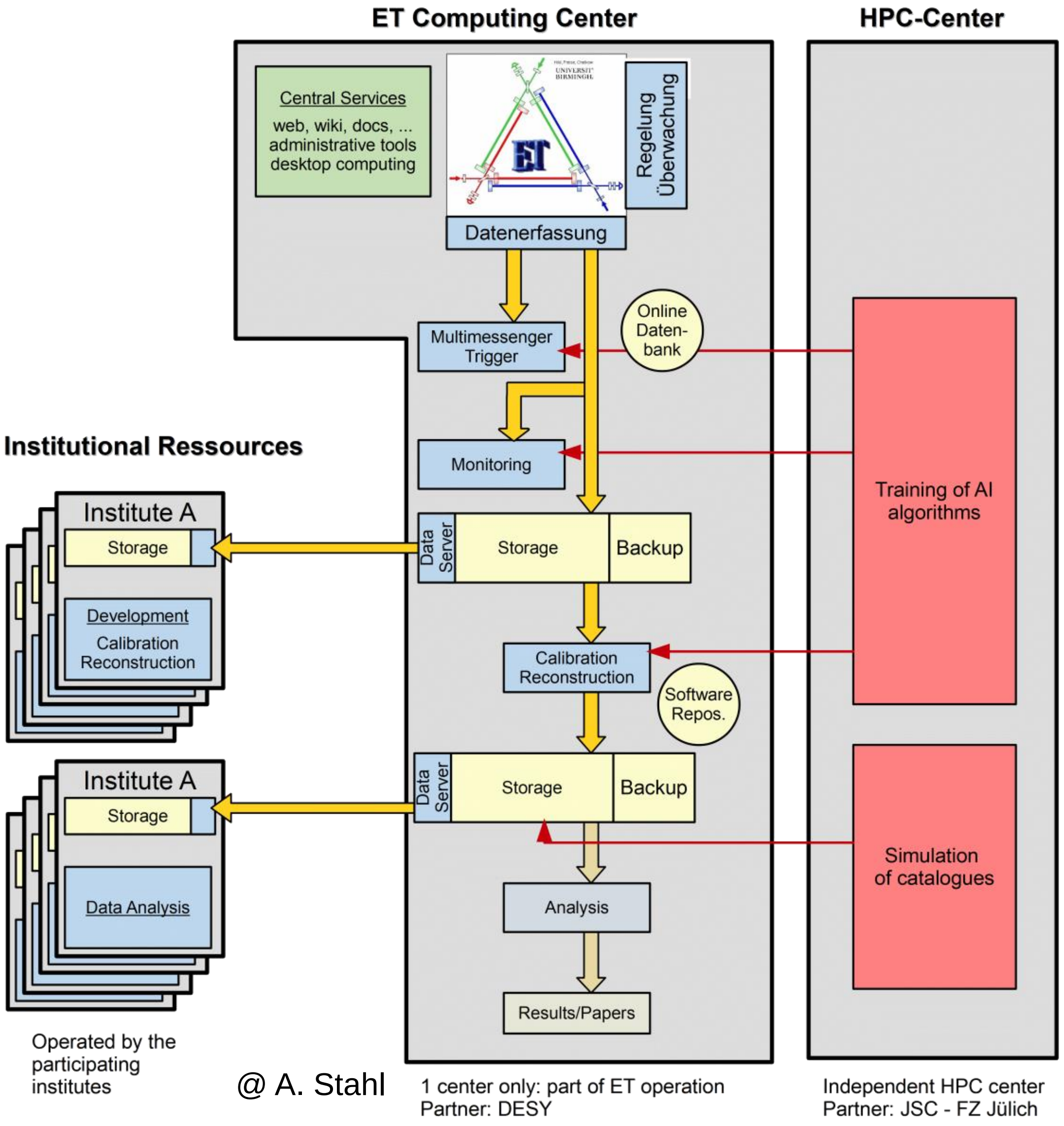
Example Computing Model: Computing Challenges of Einstein Telescope

Computing Model:

- ET Computing Center, only low latency (= operation costs)
- HPC-Center (= member country costs)
- Institutional Resources (= institutional costs)

Challenge:

- LIGO/Virgo analysis path does not work, since:
  - Many more signals / events
  - Longer signal traces at low frequencies (hours)
  - Parameter set per event much higher (better fit and comparison to template)
  - More parameters available (e.g. polarisation)
  - More types of events, i.e. more template catalogues.
  - Huge amount of (online) monitoring data
- Requests large resources (HPC) for generating and training of catalogues as well as the development of smart algorithms



@ A. Stahl

Andreas Haungs
08-10.06.2020

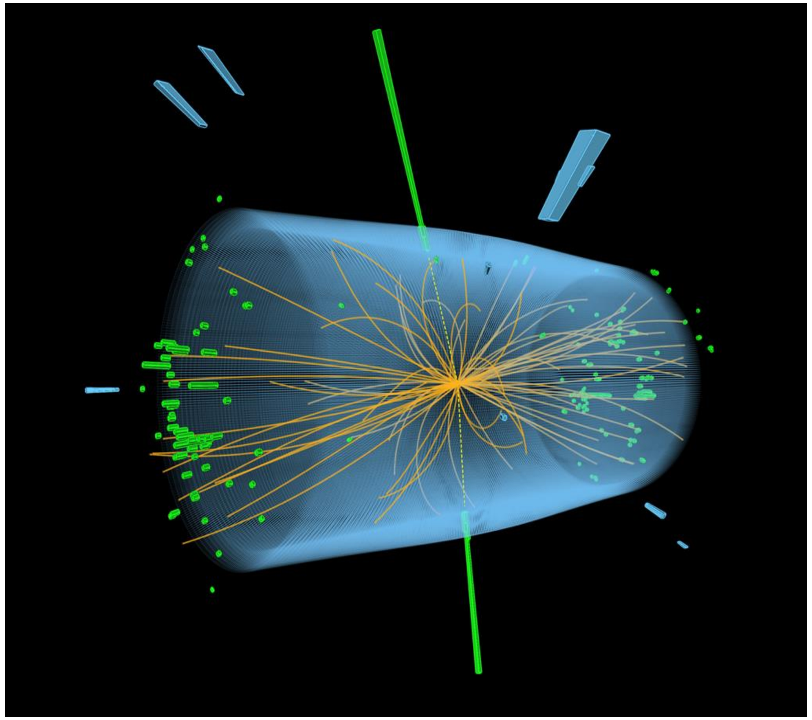
Census of Computing Requests of German Astroparticle Physics:

The demand for computing resources for astroparticle physics in Germany will increase considerably in the coming years. In 2020, the computing for the German flagship experiments (Auger, CTA, IceCube, ET, KATRIN, Gerda/Legend, DARWIN, Multi-Messenger, Theory) will mainly be carried out via institutional, experiment-specific or, as in the case of theory, federated supercomputer resources and only to a small extent via the German WLCG network. An estimation of the 2021 requirements for the German fair-share of the computing of the international experiments resulted in a sum of 2,000 CPU years, 300 GPU years, 2.5 PB disk space and 3 TB tape capacity, which are already largely covered by the WLCG (Tier-1 and Tier-2). A projection into the year 2028 showed an increased demand of about factor 8 in CPU years, factor 20 in GPU years, factor 5 in disk space and factor 10 in tape capacity.

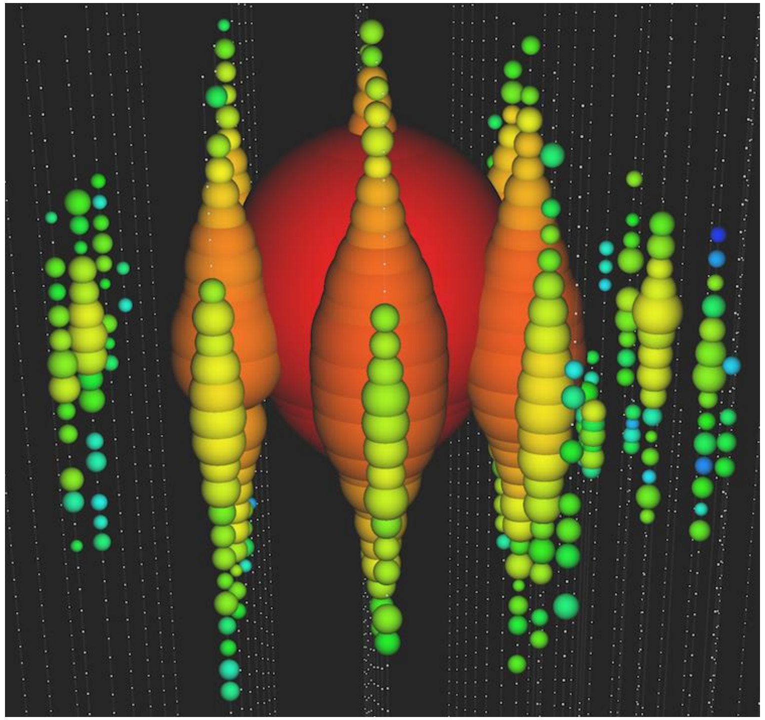
KAT. Komitee für Astro.Teilchen.Physik

PAHN-PaN: a broad German initiative

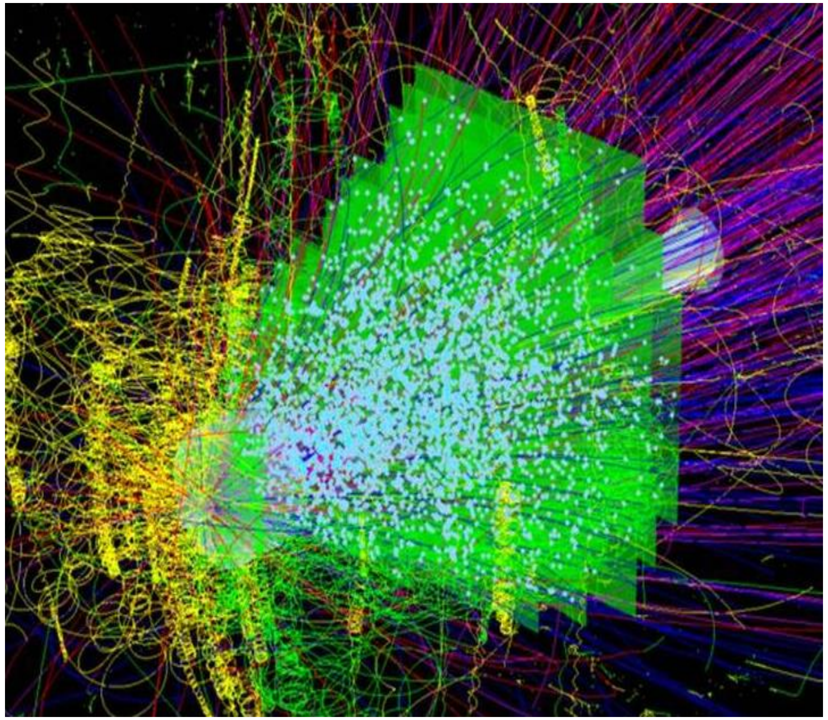
Particle, Astroparticle, Hadron & Nuclear Physics and Astronomy/Astrophysics (future)



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

The Computing Model

Service access layer
Execution access layer
Virtualisation layer

individual user / collaboration access

Continuous integration,
analysis + machine learning services

Scalable, hardware-agnostic execution infrastructure
(container, batch ...)

Hardware (CPU, GPU, FPGA, TPU, ...)

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

Layered model: scalability and easy replacement of modules!

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

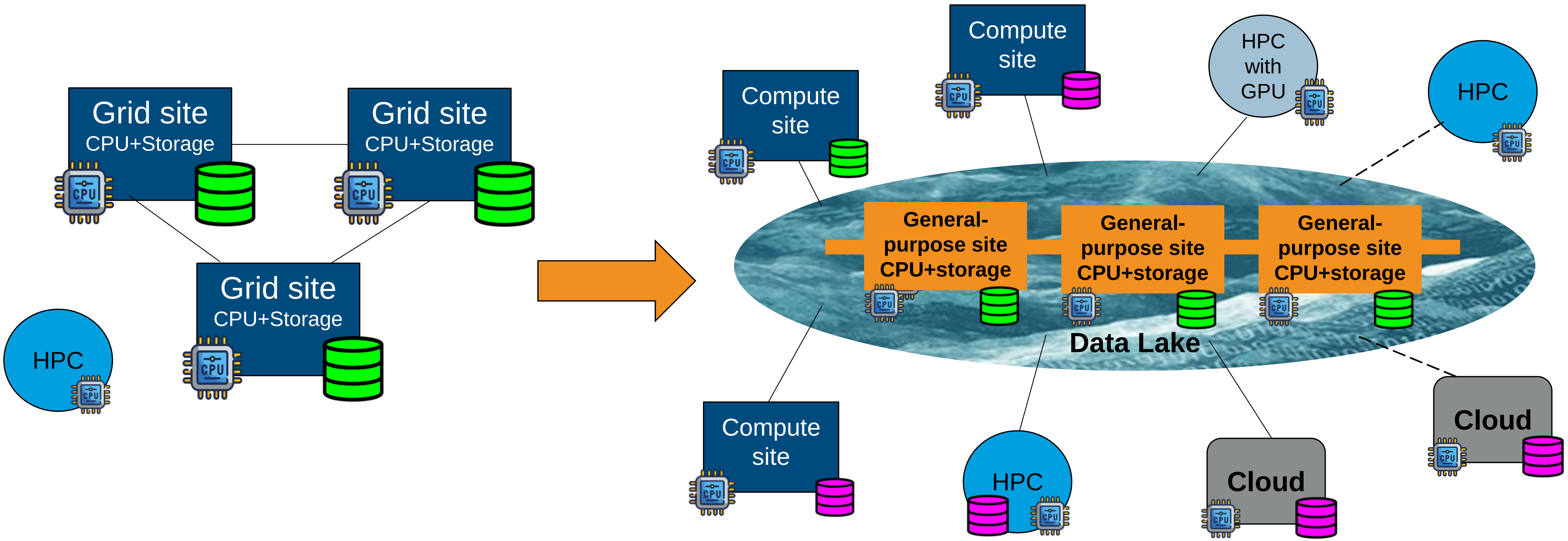
Adaption + further development of existing open source cloud middleware

Cross-cutting topic A: Synergies

Cross-cutting topic B: Services

Cross-cutting topic C: Professional training, education, and outreach

Developing Workflows and Tools for Data Management



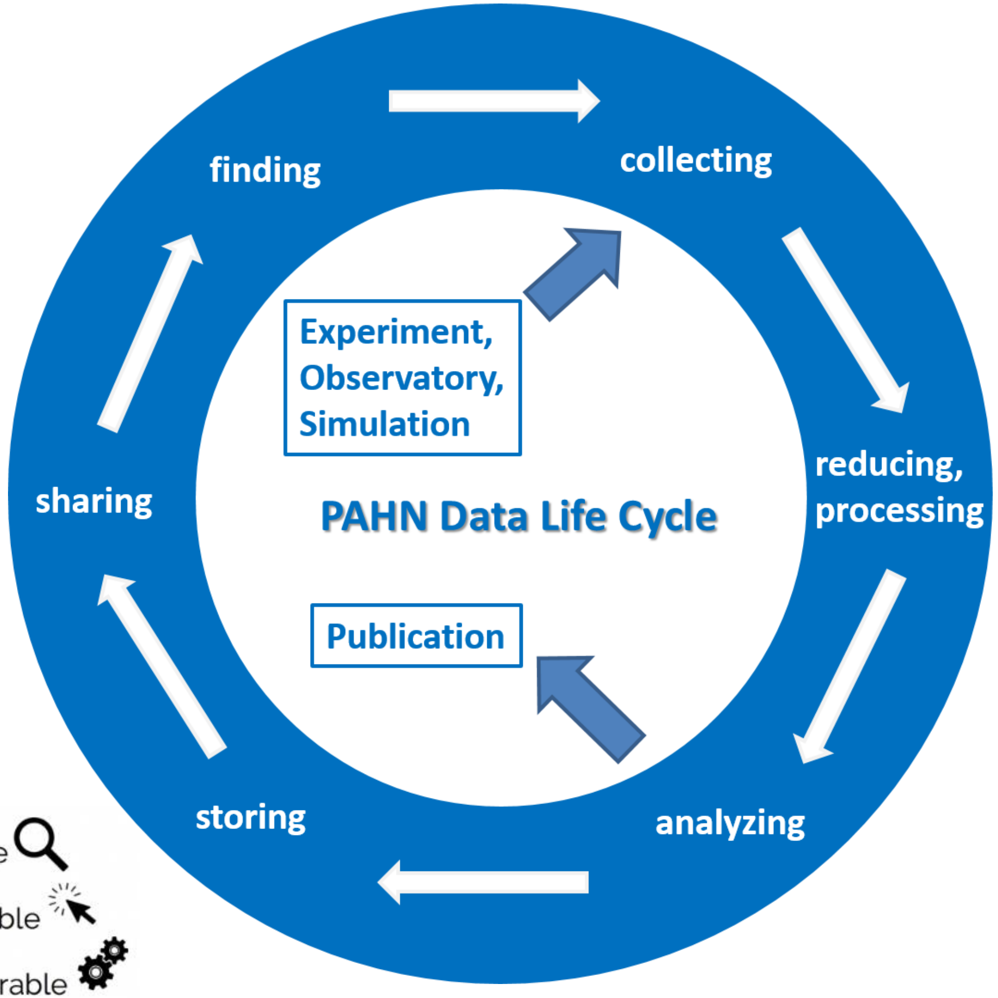
TODAY

- >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)

FAIR Data Lifecycle Concepts and Open Data



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

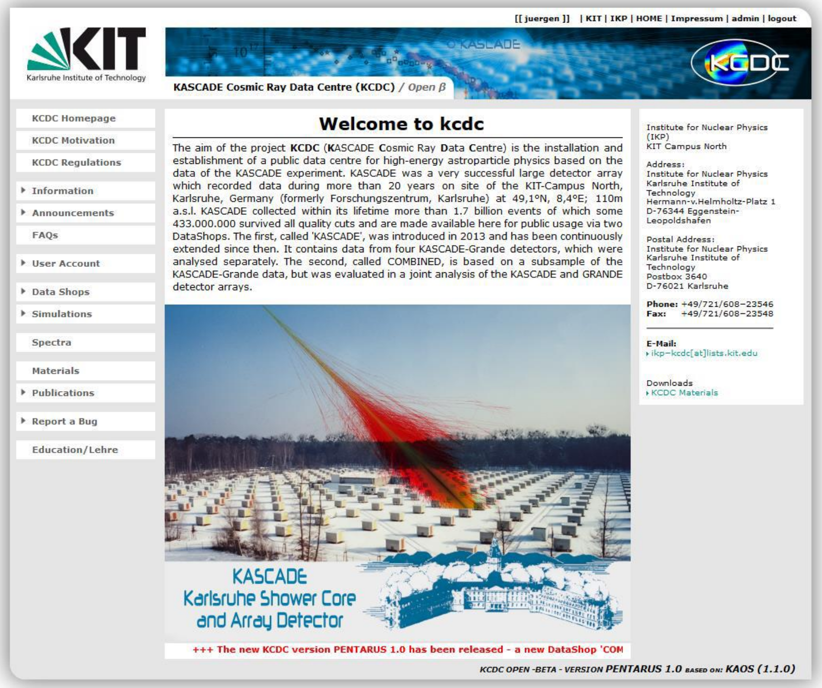
- (i) data availability
- (ii) method development
- (iii) data analysis
- (iv) big data education
- (v) open access
- (vi) data archiving
- (vii) data mining

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

KASCADE Cosmic ray Data Centre

#KCDC_KIT

- **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:**
  - release (Feb. 2017) with $4.3 \cdot 10^8$ EAS
  - simulation data
  - spectra
- **Pioneering work in publishing research data in astroparticle physics**

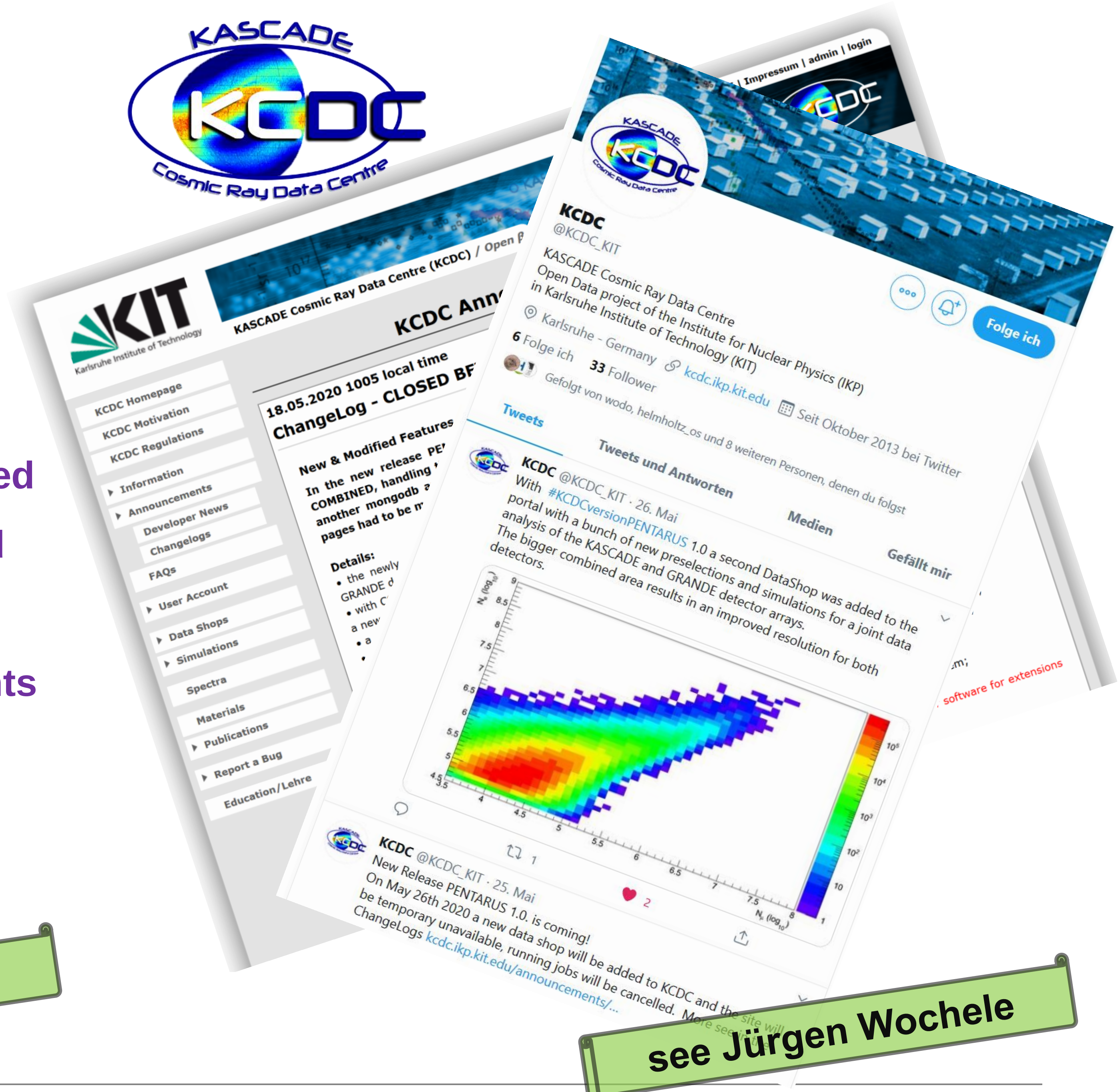


https://kcdc.ikp.kit.edu/

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

PENTARUS 1.0 is released!

26.05.2020



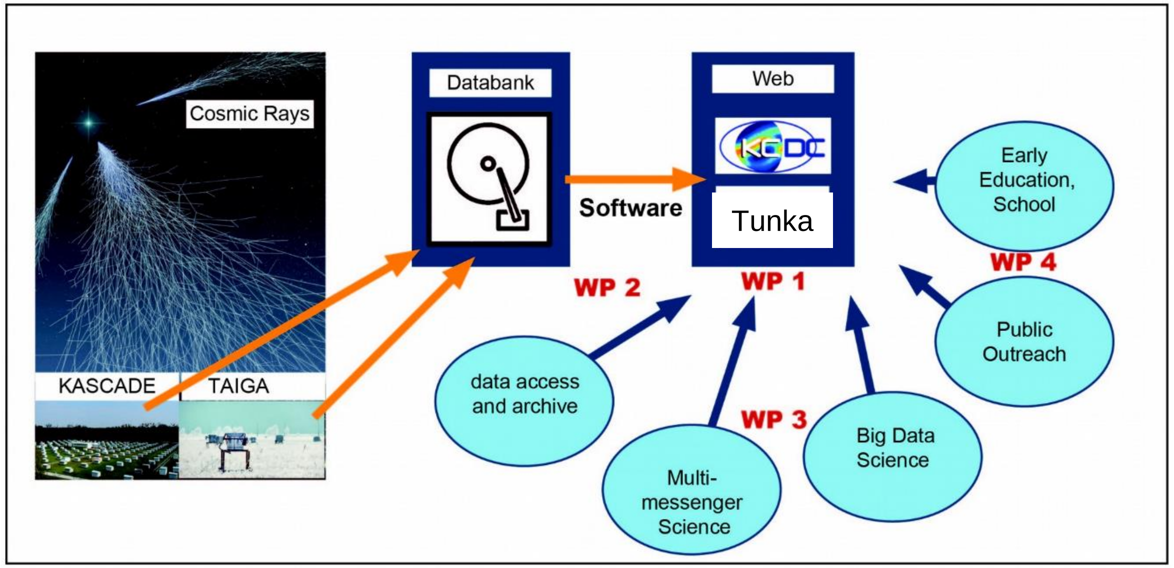
- **Now:**
  - UUID included
  - number of simulations increased
  - increase in processing and download speed
  - KCDC based publications & KCDC related publications included
  - new Data shop for independent experiments (KASCADE+Grande combined)
- **Next:**
  - open for more data shops
  - analysis platform

see Frank Polgart

see Jürgen Wochele

Astroparticle Data Life Cycle Initiative

- **Basics**
  - project period 2018-2020
  - funded by Helmholtz and RSF
  - Team leaders: A. Kryukov (SINP MSU) and A. Haungs + A. Streit (KIT)

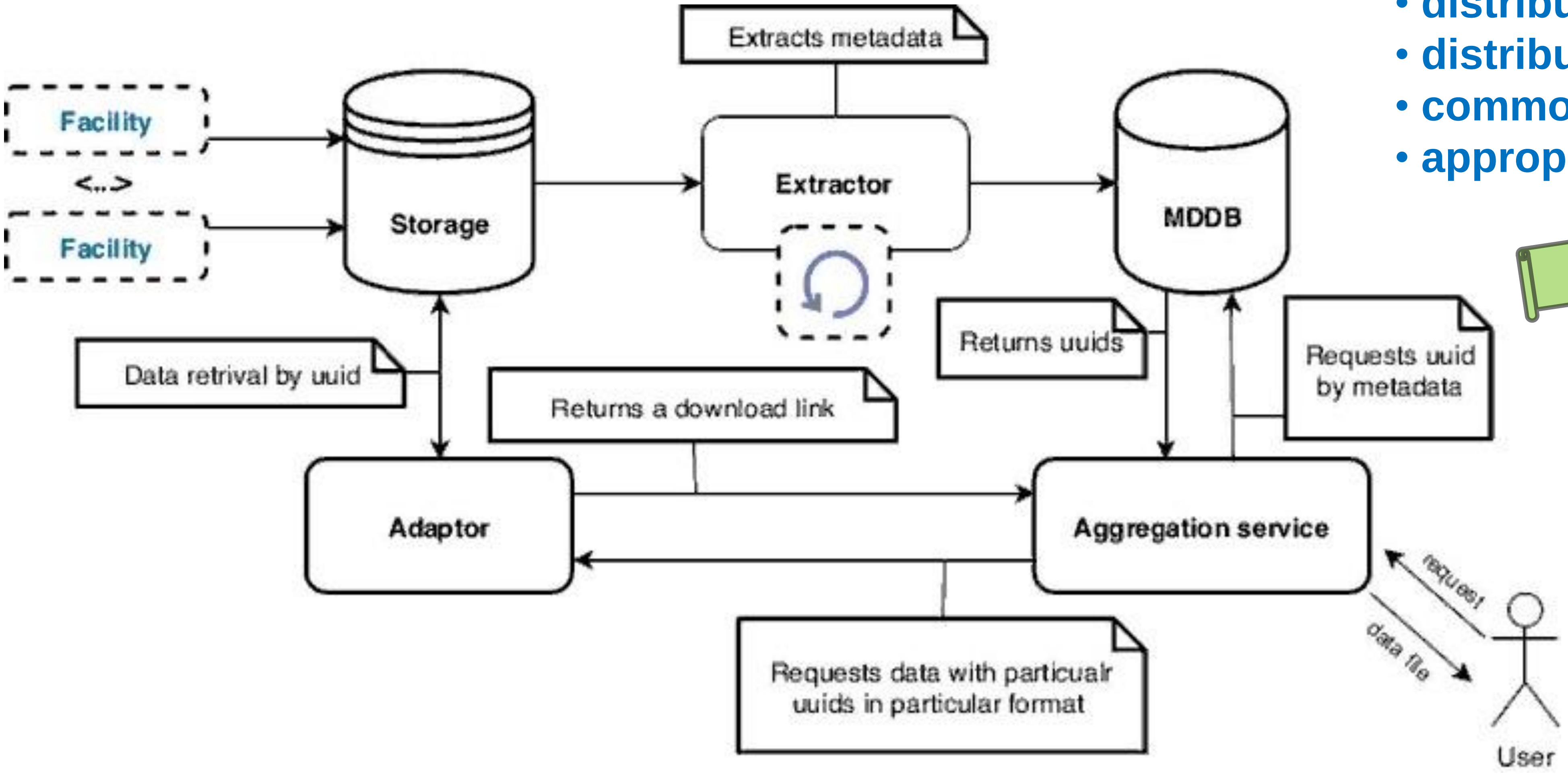


- **Main targets of the Project**
  - Extension example: data from Tunka and KASCADE-Grande
  - Developing solutions of distributed data storage techniques with a common meta-catalog
  - Development of appropriate machine-learning techniques
  - Perform experiment overarching multi-messenger astroparticle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

Project is an important step in extension and generalization of KCDC

http://astroparticle.online

Astroparticle Data Life Cycle Initiative: Data Aggregation



- distributed data provider
- distributed data storage
- common Metadata definition
- appropriate software

see A. Kryukov et al.

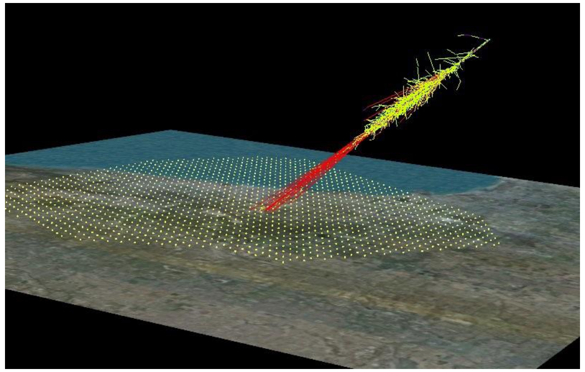
from Victoria Tokareva

from: KIT-Russia project on Astroparticle DLC (Tokareva)

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

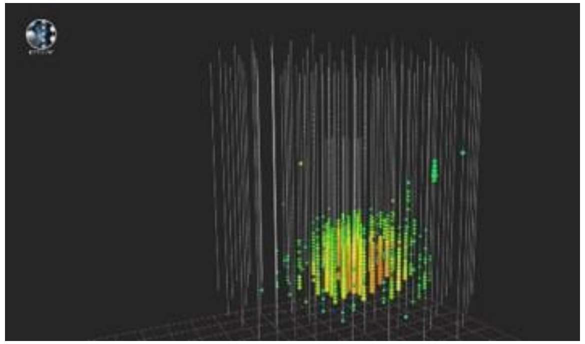
KIT Karlsruher Institut für Technologie DESY.

- **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**
  - **Foster real-time analysis**
  - **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

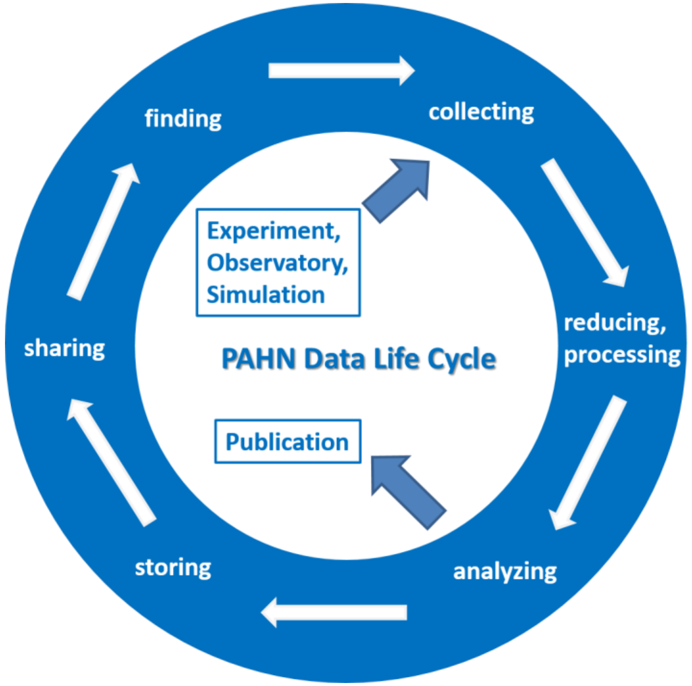
cta cherenkov telescope array



ICECUBE

Current work topics:

- **Data Management:**
  - Completion of the FAIR data cycle for major infrastructures (CTA, IceCube, Auger, CORSIKA)
  - Format and quality of data and metadata from these different observatories (and related simulations)
- **Big Data Analysis:**
  - Method development (e.g. deep learning)
  - Efficient simulations (CORSIKA)
  - Software tools (e.g. Gammapy, CTA simulation chain)
- **Multi-Messenger Analysis:**
  - Real-time services (e.g. AMPEL)
  - Access to archives (including interface questions)
  - Common (astronomical) data formats
  - Development of workflows
- **Hardware and Services:**
  - Access to HTC and HPC (GPU) in local and distributed clusters
  - Interface software (container, docker, ...)
  - Building a common Tier-1 infrastructure for IceCube;
- **Networking and Training:**
  - Activities accompanying the cooperation with MT(DMA); NDFI; EOSC; users (universities),…
  - Outreach



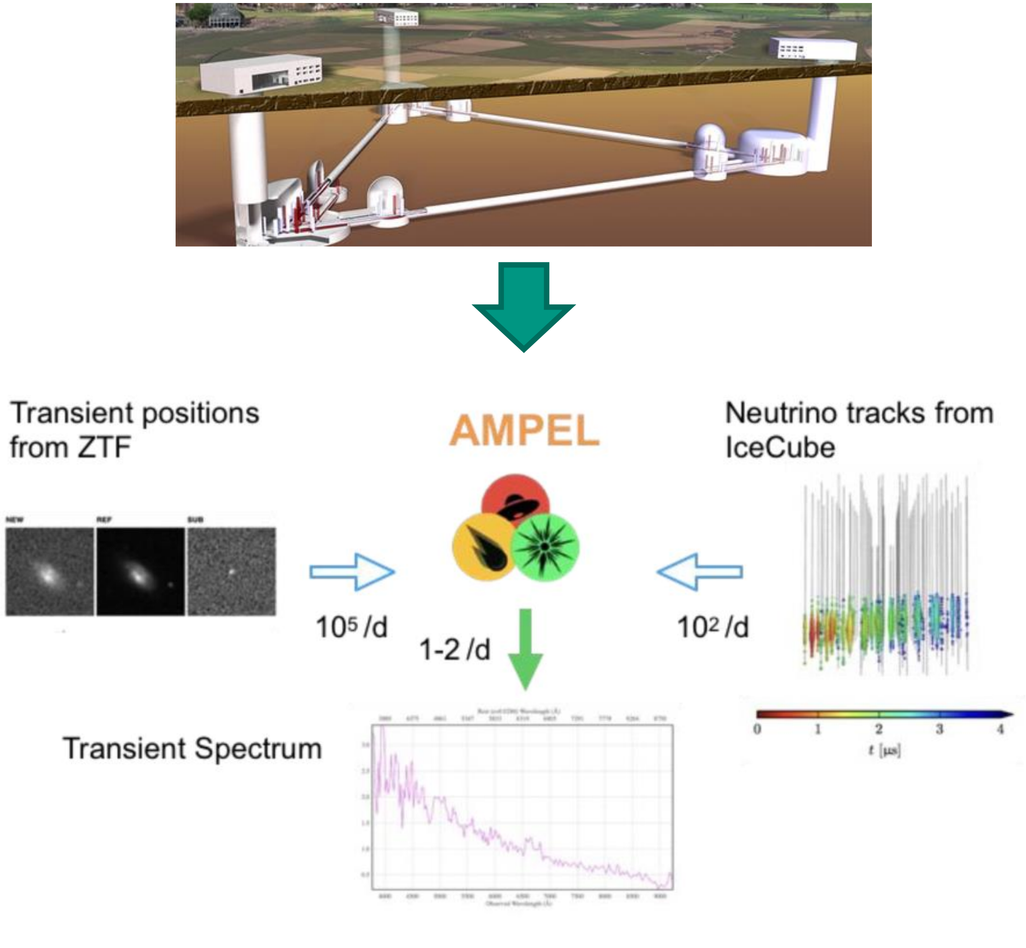
(aus PAHN-PaN NFDI Proposal, ©A.Haungs)

Example: Multi-Messenger with Gravitational Waves

Application of Helmholtz-IN2P3 bilateral project

Preparation of multi-messenger follow-up studies of gravitational wave events

- Objectives
  - Prepare extended multi-messenger follow-up studies for ET
  - Cover wide science range of astrophysics, cosmology, element synthesis, Lorenz-violation, …
  - Perform a messenger-overarching FAIR data management
- Milestones
  - Provide improved search pipeline for BNS candidates
  - Provide software for automatic scheduling of follow-up observations of robotic telescopes
  - Automated search for sub-threshold counterparts of GW by optical/UV/gamma/neutrino telescopes



AMPEL: Astron.Astrophys. 631 (2019) A147

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

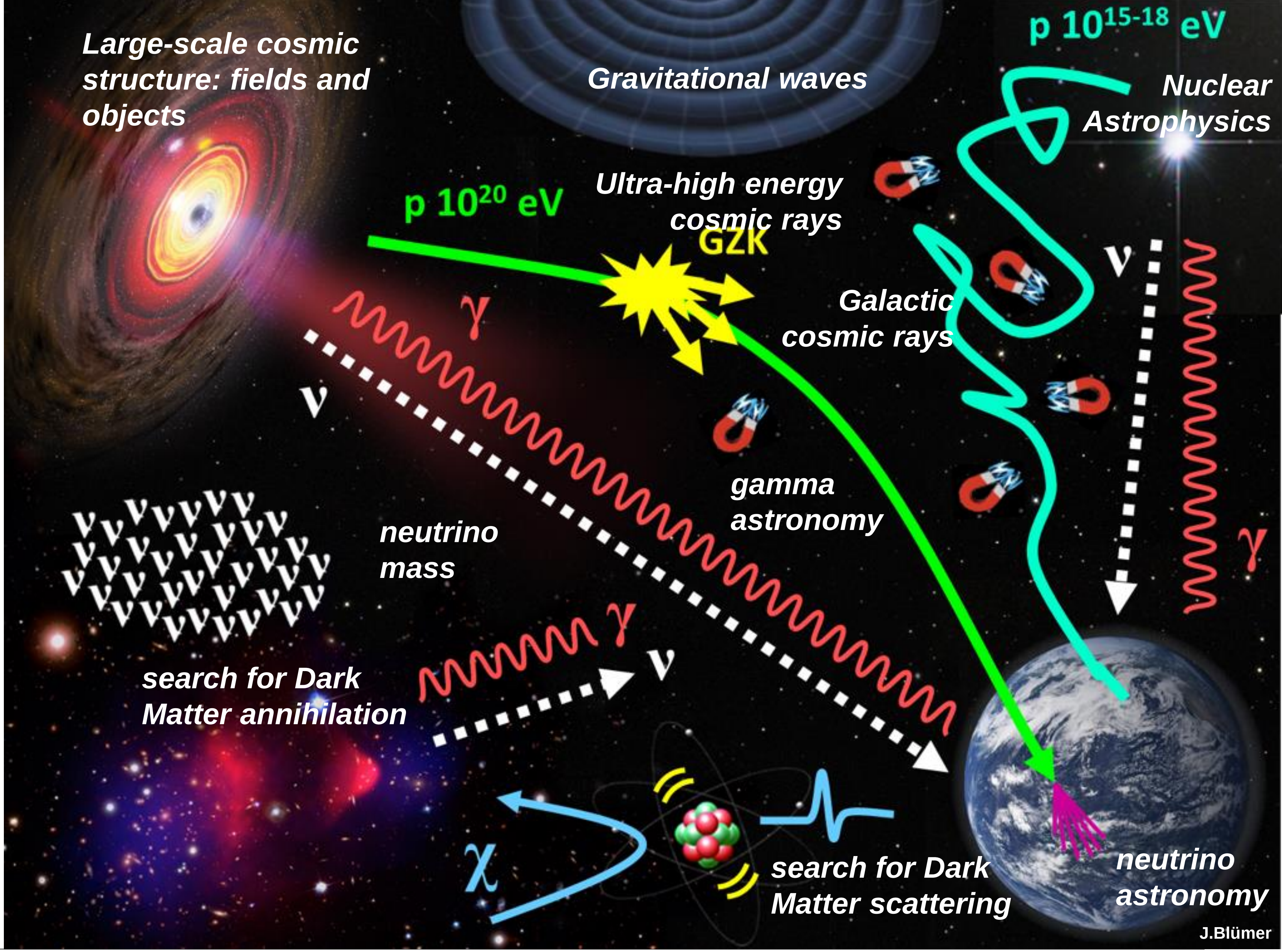
To improve access to large data sets, scientists are looking to cloud-based solutions for data management.

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"

….everything for the benefit of Astroparticle Physics!



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

needs an **experiment-overarching** platform!

END