Big Data Science in Astroparticle Physics:

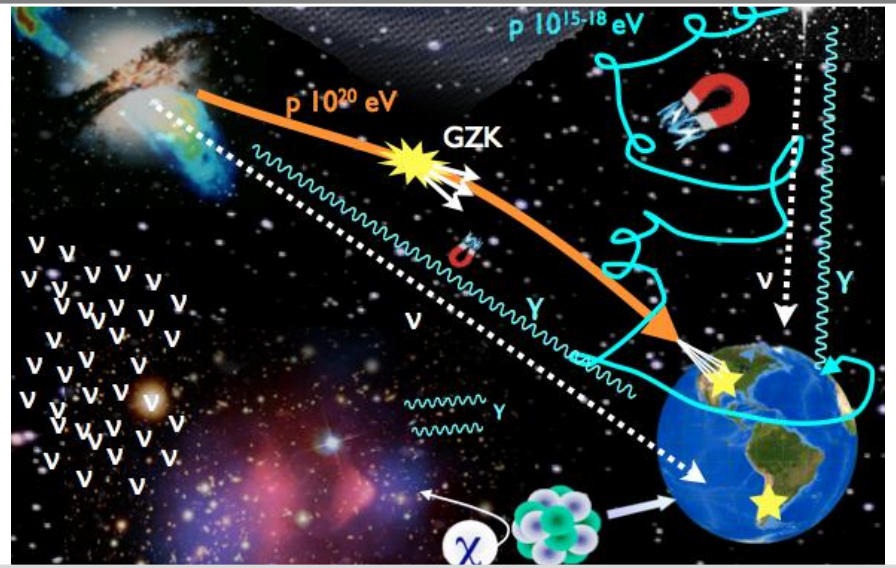
Initiative for a dedicated facility

Workshop @ KIT 2/11/2017

Andreas Haungs

HELMHOLTZ
RESEARCH FOR
GRAND CHALLENGES

KAT. Komitee für
Astro. Teilchen. Physik



Recommendations of the KAT (white paper)

„Astroteilchenphysik im Licht der Digitalen Agenda der Bundesregierung*“

KAT. Komitee für Astro. Teilchen. Physik

Empfehlungen des KAT
Das KAT unterstreicht mit Nachdruck die Bedeutung der Einrichtung bzw. den Ausbau von Zentren zur Datenspeicherung, Zurverfügungstellung der Daten und der erforderlichen Rechenressourcen als digitale Grundversorgung der deutschen Wissenschaftlerinnen und Wissenschaftler und darüber hinaus für die öffentliche Teilhabe an den wissenschaftlichen Daten.
Das KAT unterstützt den Aufbau einer Struktur, die die Kommunikation zwischen Wissenschaftlerinnen und Wissenschaftlern als Nutzer wissenschaftlicher Daten und moderner Datenanalysemethoden einerseits ermöglicht und weiterhin eine Expertenberatung im Rahmen von User Support umsetzt.
Das KAT weist auf die zentrale Bedeutung drittmittelfinanzierter sowie nachhaltig angelegter Personalstellen hin, die für die Unterstützung der Nutzer zwingend erforderlich sind.

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

Motivation

- Setting up a first **national computing and data center** to ensure the digital basic care of the scientists in astroparticle physics.
- Developing methods for the **preservation of the measured data** in a form which ensures (i) the long-term use of the data for both the scientific work as well as for educational purposes and (ii) that data of **different experiments can be combined.**
- Development of **applications of modern methods** in data analysis via a dedicated astroparticle and interdisciplinary platform.
- Build-up an environment for the **training of young academics** in modern analytical methods.

We aim to develop and implement a concept that meets the needs of the digitization of the research field, and which is also attractive to society. Goal is to achieve an efficient analysis of the data obtained in various observatories scattered around the world (multi-messenger analyses), as well as a modern solution for the synergy of basic research and the information society.

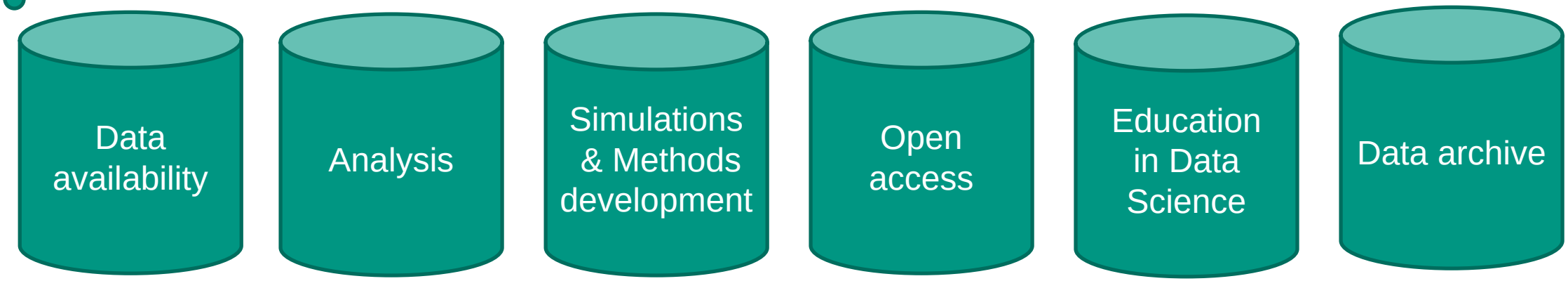
Initiative for a (global) Analysis & Data Center in Astroparticle Physics

- 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

- **High demand in community (German and international)**
  ← **white paper**
- **Astroparticle Physics experiments are globally distributed (no CERN or ESA)**
- **Needs dedicated efforts and resources, i.e. concerted action**

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!

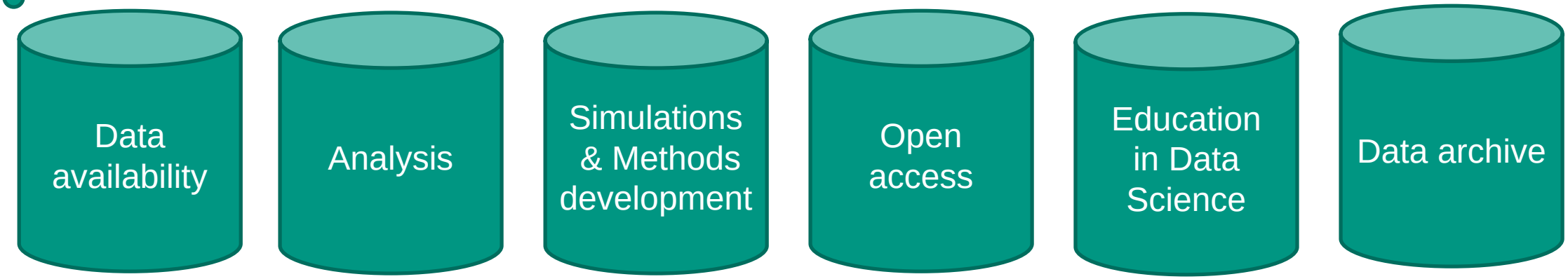
- **Education in data science:**

  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

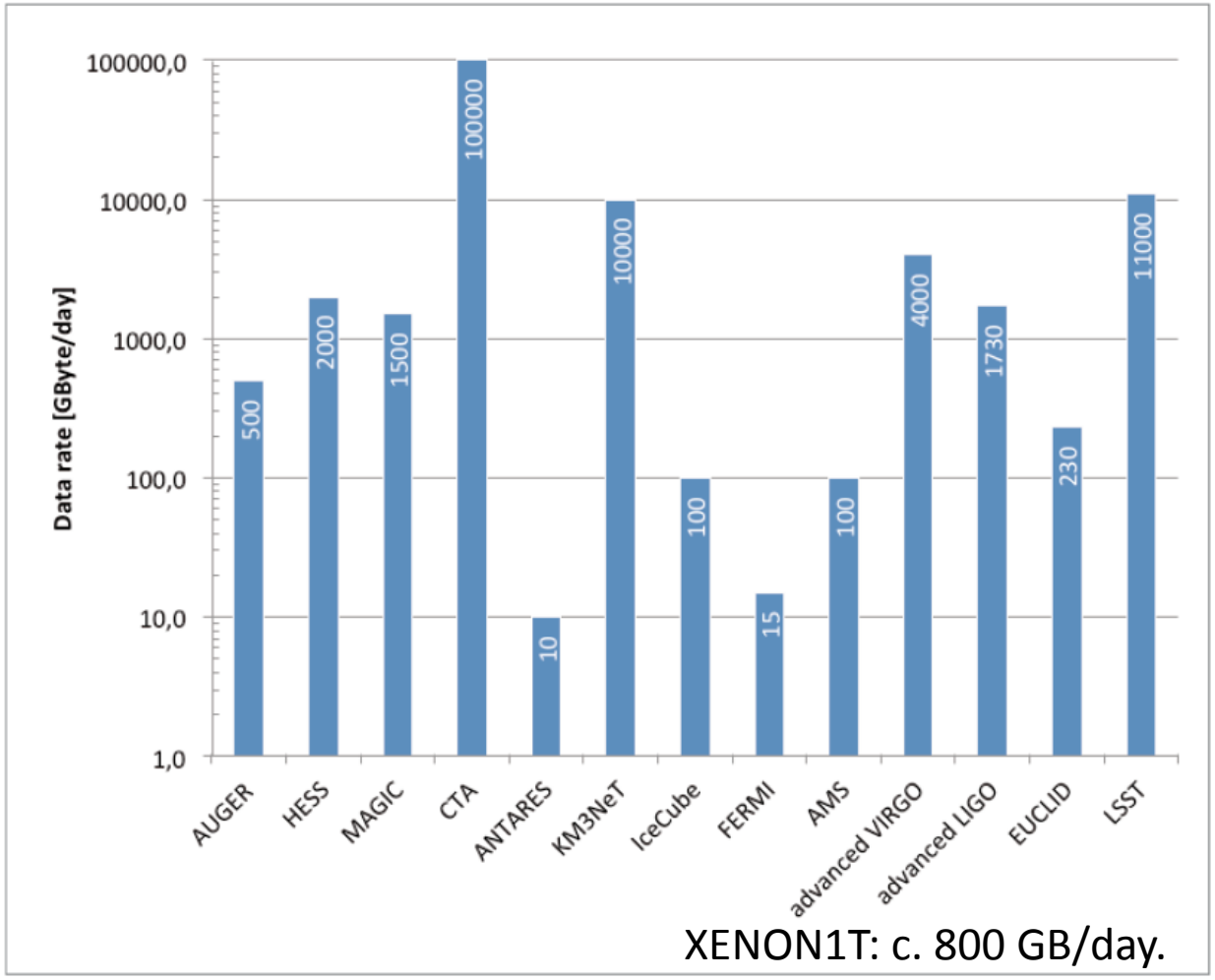


- **Data preservation** ----
  like DPHEP, KCDC
- **Metadata preservation** ----
  like KCDC
- **Data storage (archive)** ----
  like DPHEP, GridKa
- **Computing services (Grid vs. Cloud)** ---
  like CERN Tier-centres
- **Data access (policy, technology, rate)** ---
  like GridKa, KCDC
- **Training on Data use (maintenance, tutorials)** ---
  like KCDC, VISPA, CDS

- **Data analysis, simulation, visualization, 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

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



XENON1T: c. 800 GB/day.

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, …)?**

Two step process?:
first high-energy astroparticle physics
later low-energy astroparticle physics

Data Catalogues

http://www.re3data.org/

- Sample and links to repositories of scientific data, mostly results, not the “data”.

e.g., search for “Cosmic Rays”:

Found 7 result(s):

1. World Data Center for Cosmic Rays WDCCR
2. KASCADE Cosmic Ray Data Centre KCDC
3. Spitzer Science Archive SHA
4. World Data Center for Solar-Terrestrial Physics, Moscow
5. Virtual Space Science Observatory VSSO
6. LAADS Web
7. High Energy Astrophysics Science Archive Research Center

Search for “gamma rays”: found 5 results

Search for “neutrino”: found 0 results

Search for “sun” : found 28 results

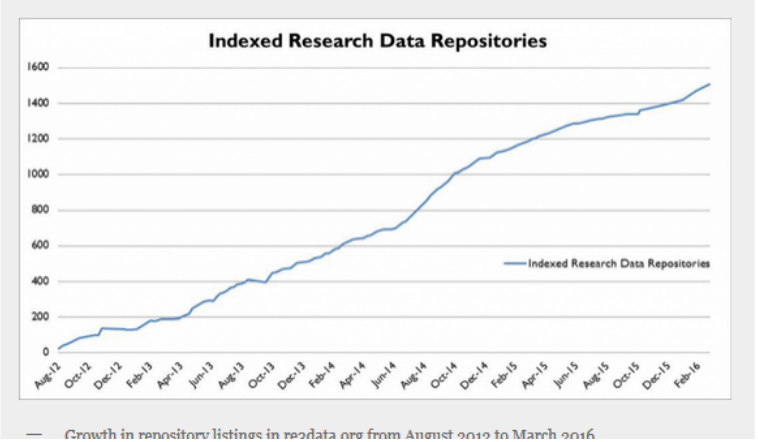
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Home | Search | Browse | Suggest | FAQ | About | Schema | API | Contact | Legal notice / Impressum

re3data.org Reaches a Milestone & Begins Offering Badges

Posted on April 13, 2016 by re3data.org team

re3data.org has reached a milestone of identifying and listing 1,500 research data repositories, making it the largest and most comprehensive registry of data repositories available on the web. It has grown steadily since its launch four years ago to cover a wide range of disciplines from around the world.



Growth in repository listings in re3data.org from August 2012 to March 2016

SEARCH

re3data.org search

A SERVICE BY

DataCite

PARTNERS

GFZ Helmholtz Centre Potsdam

HUMBOLDT-UNIVERSITÄT ZU BERLIN

KIT

Scopes of such an Analysis & Data Centre, some examples

- Platform for (public available) data access (enhancement of KCDC)
- Covering gamma-rays, neutrinos, cosmic rays, …. (multi-messenger astroparticle physics)
- Distribution of data and alerts (advanced AMON)
- Data in 'Matter and the Universe' (interface/connection to GAVO, CERN, …)
- Development of methods and algorithms (deep learning for EAS reconstruction)
- Simulations (efficient use of CORSIKA)
- Data and simulation visualization
- Providing computing infrastructure (CPU, storage, GPU,…)
- Access to distributed infrastructure (heterogeneous resources)
- Development of efficient interfaces to the infrastructure (container, docker)
- Higher level data products (e.g. data compression algorithms)
- ….

← Definition / concept / identity of the centre

← Requirements to the centre

Environment

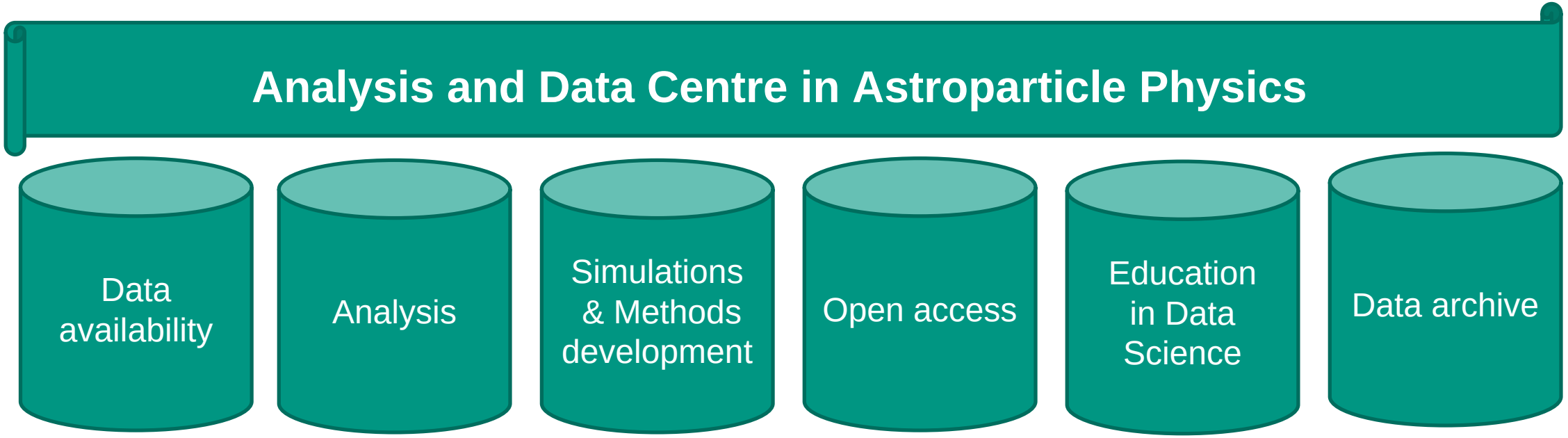
- Tier Centers
  - GridKa, DESY (Zeuthen), …
  - GAVO, ….
  - CERN ?
- EOSC / Go FAIR Initiatives
- H2020 / ASTERICS
- Querschnittsthema Verbundforschung
- Heraeus Semiarreihe
- Helmholtz funding schemes
  - PoF
  - IVF
  - Inkubator
- …
- …

➔ Embedding / Funding sources / Start-up / …

Important dates:

- Strategy meeting of KAT
  7-8 December 2017
- Big Data Science in Astroparticle Physics
  19-21 February 2018
- Helmholtz Evaluation DESY
  5-9 February 2018
- Helmholtz Evaluation KIT
  12-16 February 2018

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



Next:

- **Definition of the needs (today)**
- **Secure funding**
- **Implementation + Internationalization**
- **Specific adaption for lower energy astroparticle physics**