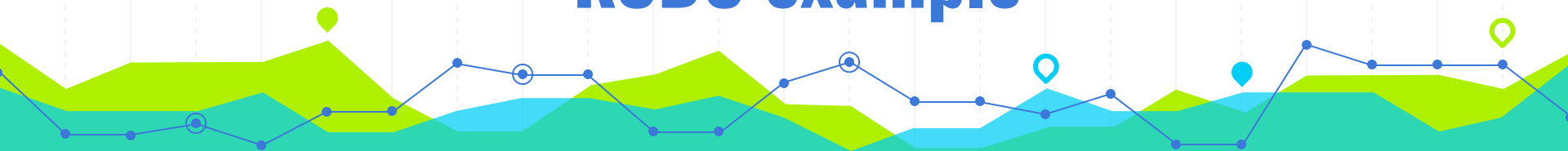


Open data and open-source tools throughout research data life cycle: KCDC example



V. Tokareva for KCDC Team

Karlsruhe Institute of Technology, Institute for Astroparticle Physics

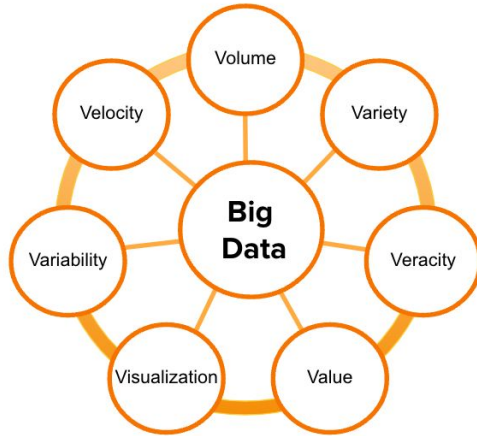
DPG Spring Meeting of the Matter and Cosmos Section (SMuK)

Hacky Hour (joint session AKjDPG/AGI)

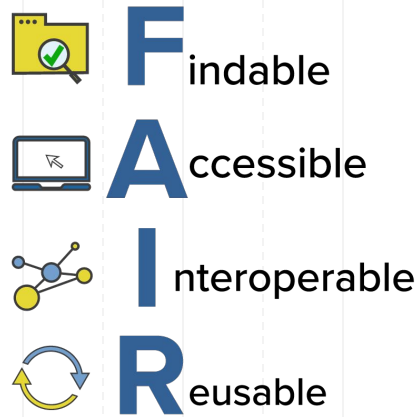
Dresden, 23 March 2023

Data-trends in particle, astro-, astroparticle, hadron and nuclear physics

Big data



FAIR data & software



Open science

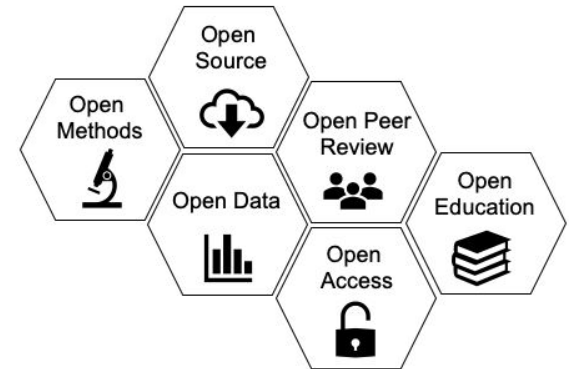
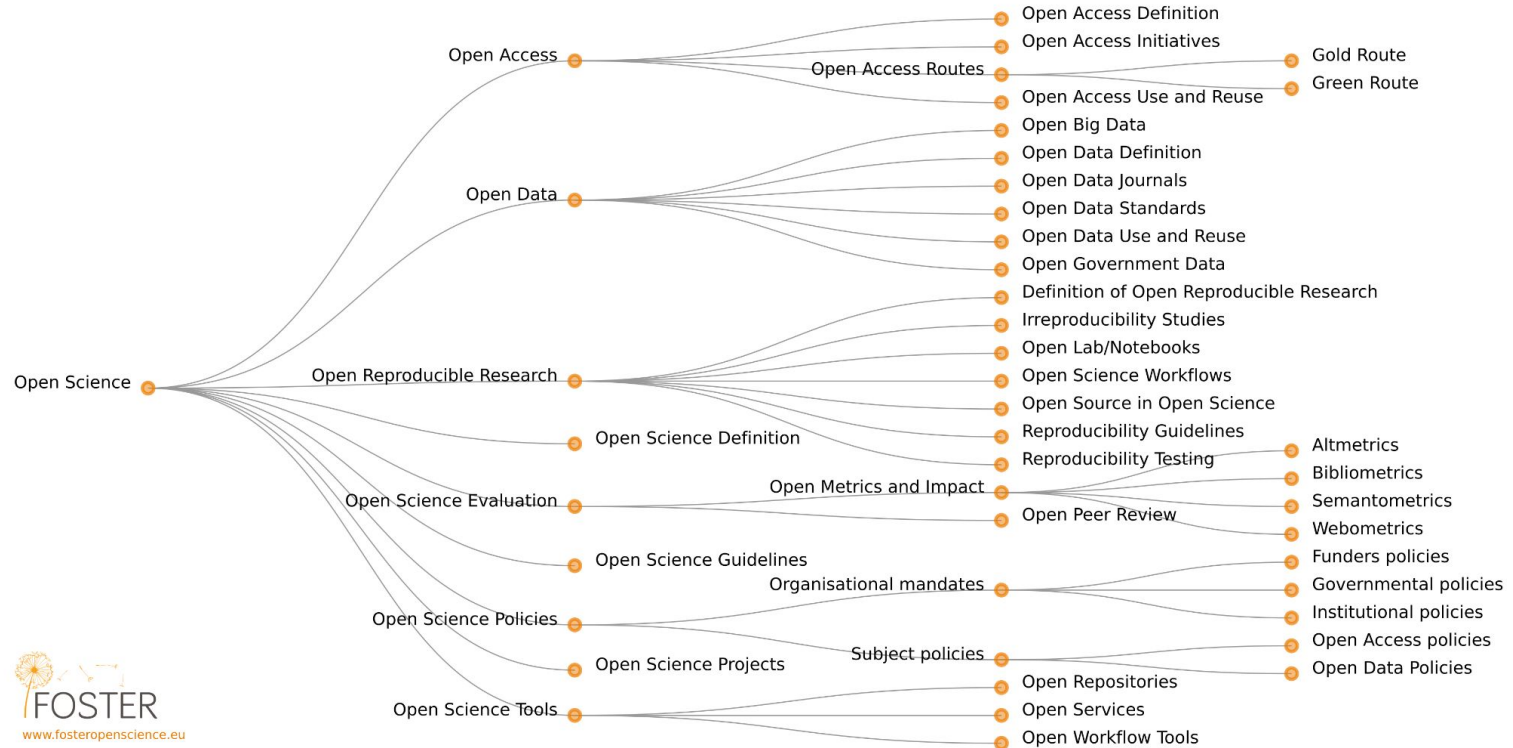


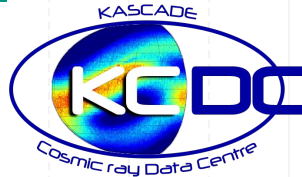
Image sources: <https://shorturl.at/nzFN8>, <https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html>, <https://opensocialwork.org/research/open-science/>

Open Science Taxonomy



KCDC - KASCADE Cosmic Ray Data Centre

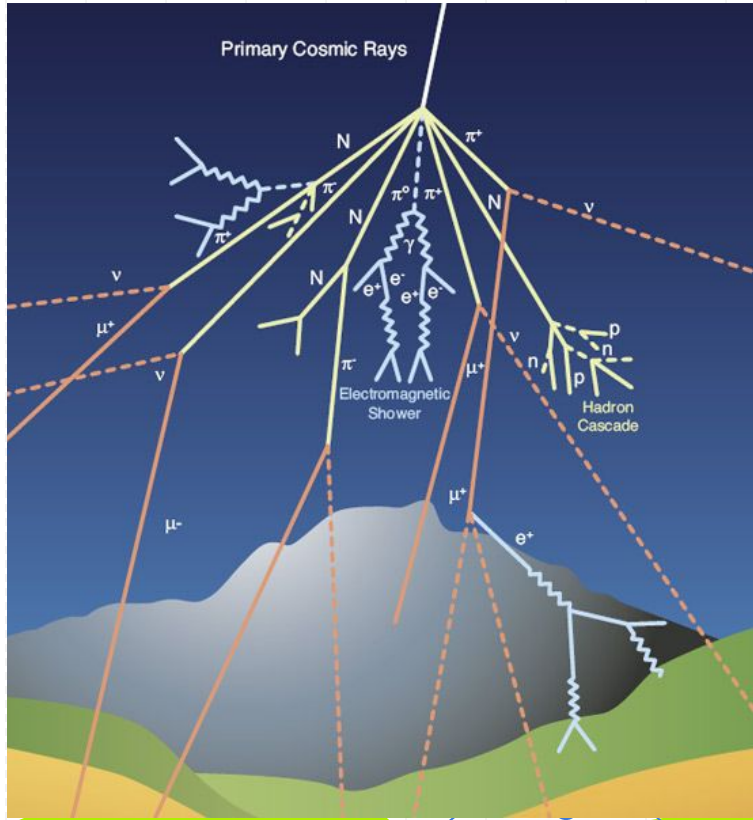
- KCDC is the public data centre for high-energy astroparticle physics
- Based on the data of the KASCADE experiment, contains as well data by KASCADE-Grande, LOPES, Maked-Ani, allows further extensions
- More than 433.000.000 events
- Established in 2013
- <https://kcdc.iap.kit.edu/>



Functionality

- Archive of KASCADE software and data
- Provides free, unlimited, reliable open access to data of the KASCADE cosmic ray experiment
 - Selection of fully calibrated quantities and detector signals
 - Custom user data cuts
 - Allows data selection using both GUI and RESTless API
- Allows interactive analysis with integrated Jupyter Notebooks
- Information platform: physics and experiment backgrounds, tutorials, reference information

KASCADE - Karlsruhe Shower Core and Array DEtector



- Location: 110 m a.s.l., 49° N, 8° E, KIT-Campus North, Karlsruhe, Germany
- Operation time: 1996 October – 2010 May $\Rightarrow$ e/ γ detector liquid scintillator effective time ~ 4223.6 days
- Area: $200 \times 200 \text{ m}^2$, $E = 100 \text{ TeV} - 80 \text{ PeV}$
- 252 scintillator detectors



Other open datasets at KCDC

GRANDE (KARlsruhe Shower Core and Array DEtector-Grande)

is an extension of the KASCADE experiment. By this the energy range KASCADE was extended to 10^{14} – 10^{18} eV. 35 310 393 events are available

LOPES (LOfar PrototypE Station)

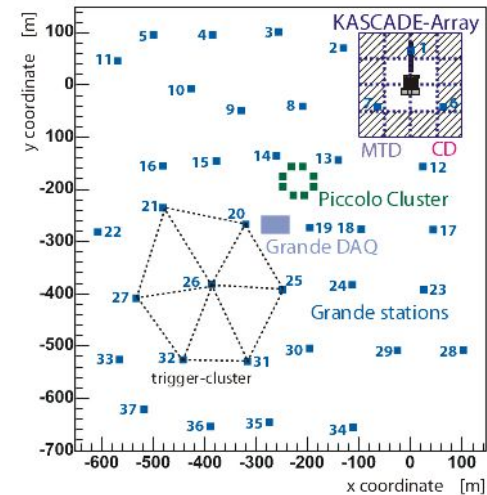
is an experiment, which measures the radio emission of cosmic ray air showers in the frequency range from 40 to 80 MHz. 3 058 events

COMBINED

is a combined dataset, made of data by 'KASCADE' and 'GRANDE' detector components for multi-messenger analysis. 15 635 550 events

MAKET-ANI

is an extensive air shower experiment placed on Mt. Aragats (Aragats Cosmic Ray Observatory, Armenia). 2 682 264 events



Datasets and formats at KCDC

The following **data formats*** are supported:

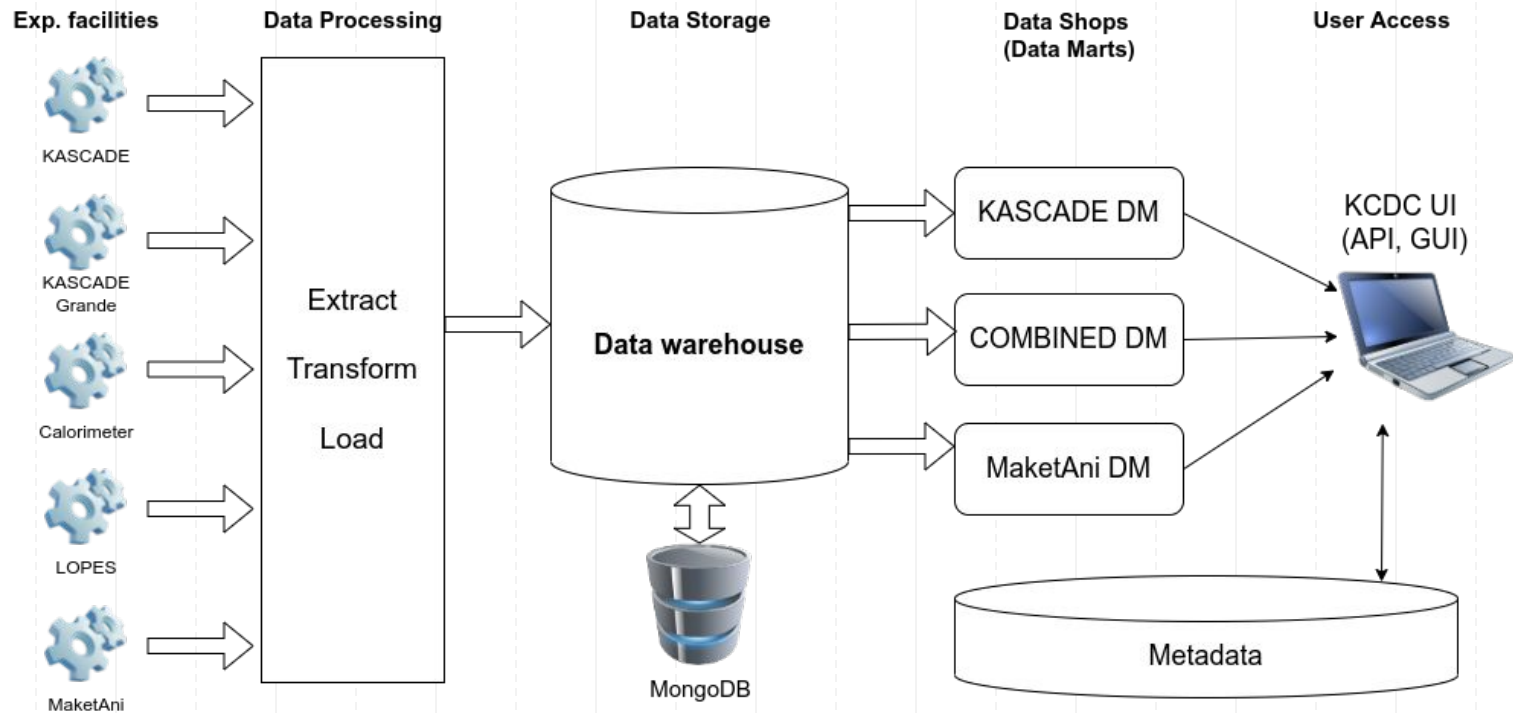
- ASCII - plain text format
- ROOT - object oriented framework developed by CERN
- HDF5 - hierarchical data format

* Selectable by the user and depending on the quantities chosen

Kinds of datasets, available at KCDC:

- Through Web GUI:
 - Preselected datasets
 - Simulations, based of the following models:
 - High energy models from three different model families: QGSjet, EPOS, SIBYLL
 - Low energy model FLUKA
- Through Web GUI + API:
 - Custom data selections

KCDC's software architecture



Metadata handling at KCDC

JSON metadata schema, example of a record from KCDC

```
392 {
393   "model": "kaos_datashop.quantity",
394   "fields": {
395     "quant_type": "num",
396     "allow_cuts": true,
397     "head_description": "<p class=dcInfoBoxHeaderDS>Zenith Angle Info</p>",
398     "descr_type": "HTML",
399     "composite_data_handler": "",
400     "unit": "\\u00B0",
401     "detector": [
402       "",
403       "grande"
404     ],
405     "quant_sub_type": "f64",
406     "display_format": "default",
407     "min_value": "0.0",
408     "display_name": "Zenith Angle",
409     "description": "<div>\\r\\n<span class=dcInfoBoxDetailsDS>\\r\\nThe reconstructed Zenith Angle of the
KASCADE showers is derived from the arrival time distribution of the of the particles at the detector
stations. The range is from <span class=dcMathFunc>0&deg;</span> to <span class=dcMathFunc>60&deg;</
span> where <span class=math>0&deg;</span> corresponds to a vertical shower. The angular resolution is
between <span class=dcMathFunc>0.4&deg;</span> and <span class=dcMathFunc>0.1&deg;</span> depending
on the energy.\\r\\n<br>\\r\\n<b>We recommend to use data only up to 42&deg;</b><br><br></span>\\r\\n <span
class=dcInfoBoxReference> [details see <b>KDCDC-Manual</b></span> \\r\\n</div>\\r\\n",
410     "name": "Ze",
411     "max_value": "60.0",
412     "selection_mode": "D",
413     "order": 2,
414     "descr_head_html": "<p class=dcInfoBoxHeaderDS>Zenith Angle Info"
415   }
416 }
```

KCDC Preselections



Karlsruhe Institute of Technology

[[guest_user001]] | KIT | IAP | HOME | Data Privacy | Impressum | admin | logout



KASCADE Cosmic Ray Data Centre (KCDC) / Open β

- KCDC Homepage
- KCDC Motivation
- KCDC Regulations
- Information
- Announcements
- FAQs
- User Account
- Data Shops
 - KASCADE
 - COMBINED
 - Maket-Ani
- Review Requests
- Preselections
- Simulations
- Analyses

Data Preselections from 'KASCADE' and 'COMBINED'

- Data Preselections from the 'KASCADE' DataShop
- Data Preselections from the 'COMBINED' DataShop

Data Preselections from 'KASCADE'

Data sets	Download Info		
	ASCII	ROOT	HDF5
KASCADE_ReducedData_runs_0877-4683_xxx all KASCADE-Array quantities in the run range 877-4683, no cuts applied, no data arrays Full data sample from the first releases VULCAN and MERIDIAN	download 7.3 GB details	download 7.1 GB details	download 6.2 GB details
KASCADE_ReducedData_runs_4685-7417_xxx all KASCADE quantities in the run range 4685-7417, no cuts applied, no data arrays	download 13.0 GB details	download 13.0 GB details	download 1 GB details
KASCADE-GRANDE_ReducedData_runs_4775-7398_xxx	download	download	download

Data Preselections

This menu offers the option to download preselected data samples for the detectors 'KASCADE' and 'COMBINED' directly without selecting quantities and applying cuts in the DataShops. 'details' provides a more detailed information page of the respective data set. The 'Small Data Samples' offer the opportunity to check your own requirements on a small data sample statistically evenly distributed. The 'download' button connects to our download server.

[details -> [KCDC Manual](#) and [KCDC-Combined Manual](#) respectively]

Simulations

KCDC Homepage

KCDC Motivation

KCDC Regulations

► Information

► Announcements

FAQs

► User Account

► Data Shops

► Simulations

► Analyses

Spectra

Event Display

Materials

► Publications

► Report a Bug

► Education/Lehre

KCDC Partners

Air Shower Simulations with SIBYLL 2.1 and FLUKA

Technical data for SIBYLL 2.1 simulations

High energy model	SIBYLL 2.1 ($E_{\text{lab}} > 200\text{GeV}$)
Low energy model	FLUKA 2008.3 ($E_{\text{lab}} < 200\text{GeV}$)
Energy range KASCADE	$1.0 \times 10^{14} < E_0 < 1.0 \times 10^{18} \text{ eV}$ (HE: $< 3.16 \times 10^{18} \text{ eV}$)
Energy range GRANDE	$1.0 \times 10^{14} < E_0 < 1.0 \times 10^{18} \text{ eV}$ (HE: $< 3.16 \times 10^{18} \text{ eV}$)
Energy range CALORIMETER	$1.0 \times 10^{14} < E_0 < 3.16 \times 10^{17} \text{ eV}$

Detector component	Primary particle	Events simul*reused	Events processed	download size	download root files
KASCADE	proton	256,948 * 3	331,292	416 MB	download
	helium	256,948 * 3	338,234	408 MB	download
	carbon	256,948 * 3	328,929	383 MB	download
	silicon	256,948 * 3	312,520	355 MB	download
	iron	256,948 * 3	286,327	321 MB	download
GRANDE	proton	256,948 * 10	62,775	60 MB	download
	helium	256,948 * 10	57,277	57 MB	download
	carbon	256,948 * 10	51,412	53 MB	download
	silicon	256,948 * 10	47,276	51 MB	download
	iron	256,948 * 10	43,456	48 MB	download

SIBYLL 2.1 Information

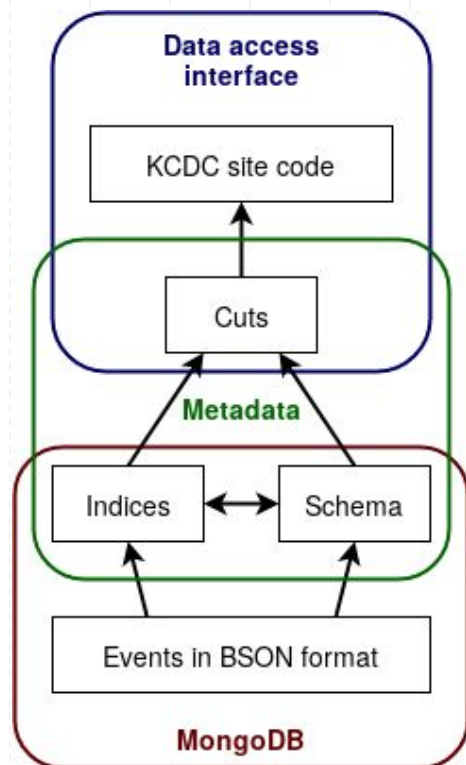
Events simul * reused denotes the number of events simulated and how often they have been reused for the respective detector component, while **Events processed** specifies the number of events surviving all quality cuts which are included in the respective root file.

Clicking on the respective primary **download** link will open a dialog box to start the download.

download ALL will download the simulations of all 5 primaries in one zipped file.

For a more detailed description of the simulation sets, applied cuts and download procedure check the [KCDC Simulations Manual](#)

KCDC data acquisition



KASCADE Data Shop

Components Available	Components Selected	Quantities and Cuts	
☐ GRANDE	☐ General Info	☐ Toggle all	<i>KASCADE</i>
☐ Calorimeter	☒ KASCADE	☐ Energy	range: 13 to 19 eV [log10] Add Cut
☐ LOPES		☐ X Core Position	range: -91 to 91 m Add Cut
		☐ Y Core Position	range: -91 to 91 m Add Cut
		☐ Zenith Angle	range: 0 to 60 ° Add Cut
		☐ Azimuth Angle	range: 0 to 360 ° Add Cut
		☐ Electron Number	range: 2 to 8.7 [log10] Add Cut
		☐ Muon Number	range: 2 to 7.7 [log10] Add Cut
		☐ Shower Age	range: 0.1 to 1.48 Add Cut

KCDC Application Programming Interface (API)

Shell example: Extraction of the all data with an energy range from 17-19eV[log10]

Request:

```
curl --insecure --request POST 'https://kcdc-  
dev.iap.kit.edu/datashop/api/submit' \  
--header 'Authorization: Basic cG92dGVyOmhhcnJ5Kytxb3R0ZXI=' \  
--header 'Content-Type: application/json' \  
--data-raw '{  
  "reconstruction": "",  
  "output_format": "ascii",  
  "datasets": [  
    {  
      "name": "array",  
      "quantities": [  
        {  
          "name": "E",  
          "cuts": [[17, 19]]  
        }  
      ]  
    }  
  ]  
'
```

Response:

job id:

```
{"id":"dbf1e608b6044223afe472125c020  
d88"}
```

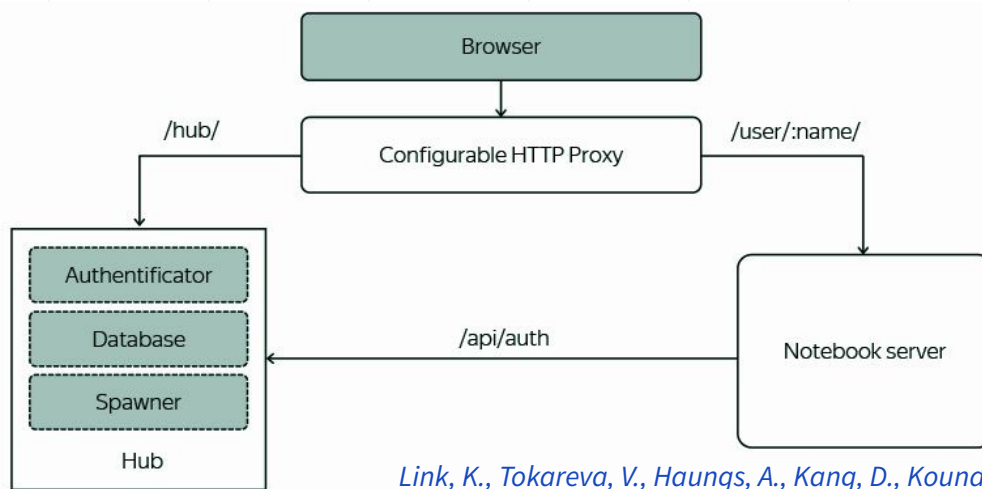
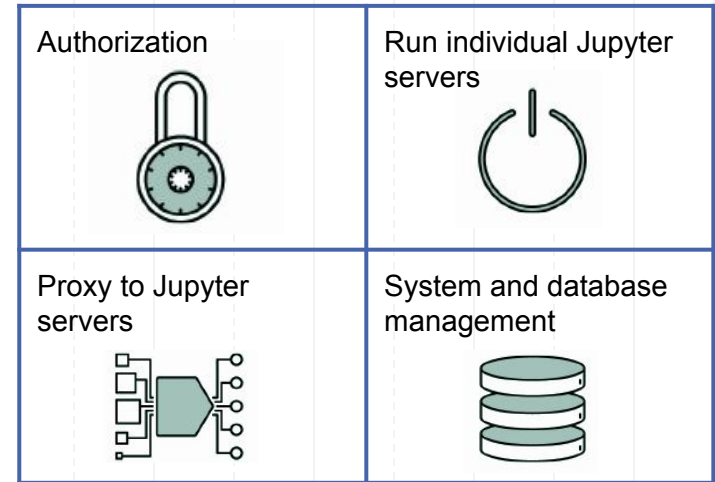
or error message:

```
{"detail":"Invalid basic header.  
Credentials not correctly base64  
encoded."}
```

- [4] Online API documentation: <https://kcdc.iap.kit.edu/datashop/api/docs/index.html>
[5] Wochele J. et al. KCDC User Manual: https://kcdc.iap.kit.edu/static/pdf/kcdc_mainpage/kcdc-Manual.pdf

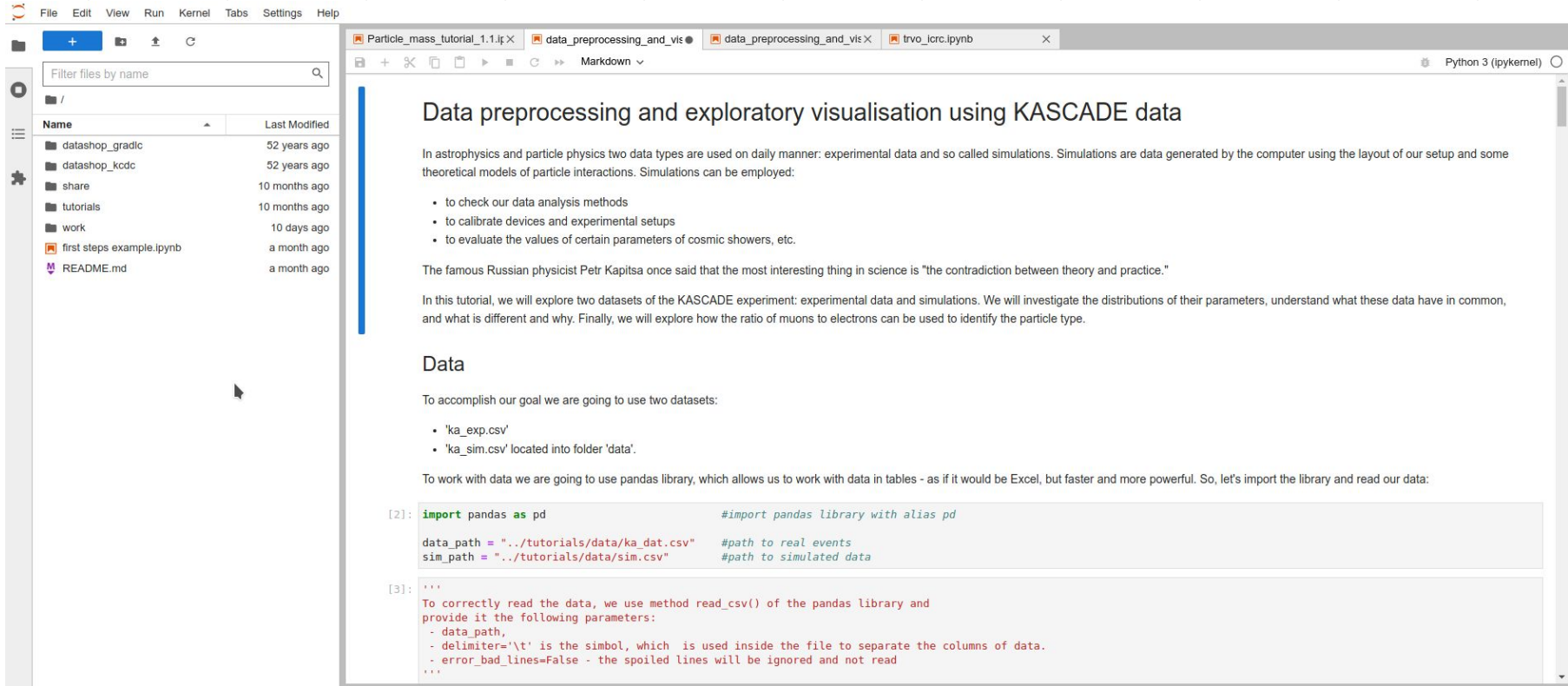
JupyterHub for data analysis

- Login via KCDC credentials
- Administration using Docker Swarm
- Tutorials by: KASCADE, IceCube, TRVO



Link, K., Tokareva, V., Haungs, A., Kang, D., Koundal, P., Polgart, F., Tkachenko, O., Wochele, D., Wochele, J. Online masterclass built on the KASCADE cosmic ray data centre. In 37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07.

Usage of KCDC's JupyterHub



The screenshot displays a JupyterHub environment. On the left is a file browser with a search bar and a list of files and folders. The main area shows a Jupyter notebook with the following content:

Data preprocessing and exploratory visualisation using KASCADE data

In astrophysics and particle physics two data types are used on daily manner: experimental data and so called simulations. Simulations are data generated by the computer using the layout of our setup and some theoretical models of particle interactions. Simulations can be employed:

- to check our data analysis methods
- to calibrate devices and experimental setups
- to evaluate the values of certain parameters of cosmic showers, etc.

The famous Russian physicist Petr Kapitsa once said that the most interesting thing in science is "the contradiction between theory and practice."

In this tutorial, we will explore two datasets of the KASCADE experiment: experimental data and simulations. We will investigate the distributions of their parameters, understand what these data have in common, and what is different and why. Finally, we will explore how the ratio of muons to electrons can be used to identify the particle type.

Data

To accomplish our goal we are going to use two datasets:

- 'ka_exp.csv'
- 'ka_sim.csv' located into folder 'data'.

To work with data we are going to use pandas library, which allows us to work with data in tables - as if it would be Excel, but faster and more powerful. So, let's import the library and read our data:

```
[2]: import pandas as pd                                #import pandas library with alias pd

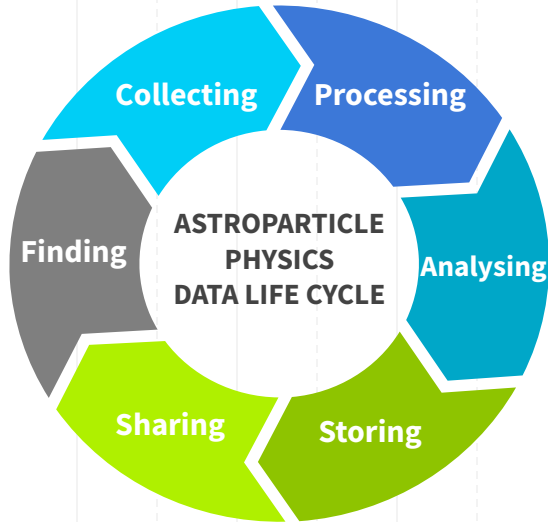
data_path = "../tutorials/data/ka_dat.csv"              #path to real events
sim_path = "../tutorials/data/sim.csv"                  #path to simulated data

[3]: '''
To correctly read the data, we use method read_csv() of the pandas library and
provide it the following parameters:
- data_path,
- delimiter='\t' is the simbol, which is used inside the file to separate the columns of data.
- error_bad_lines=False - the spoiled lines will be ignored and not read
'''
```

Open technologies in use:

- Django Web Framework
- Messaging RabbitMQ
- Celery Task Queue
- NoSQL (MongoDB) database
- JupyterHub for Jupyter Notebooks
- Docker/Singularity
- RESTful API
- Python, bash, JSON, HTML

Next steps: data life cycle and good research practice



Good Research Practice

According to DFG Code of conduct:

- Guideline 7: Quality Assurance
- Guideline 10: Usage Rights
- Guideline 12: Documentation
- Guideline 13: Public Access
- Guideline 17: Archiving

https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp_en.pdf

Data Life Cycle - the sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion

Takeaways

- KCDC is the public data centre for high-energy astroparticle physics
- All-in-one: data center, archive, information and educational platform
- Uses only open-source technologies
- Allows custom data selection using both GUI and RESTless API and interactive analysis with integrated Jupyter Notebooks
- <http://kcdc.iqp.kit.edu>
- Provides a good set of tools to be used on most of steps of research life cycle
- Promotes open science and good research practices

THANKS!

Hands-on time!

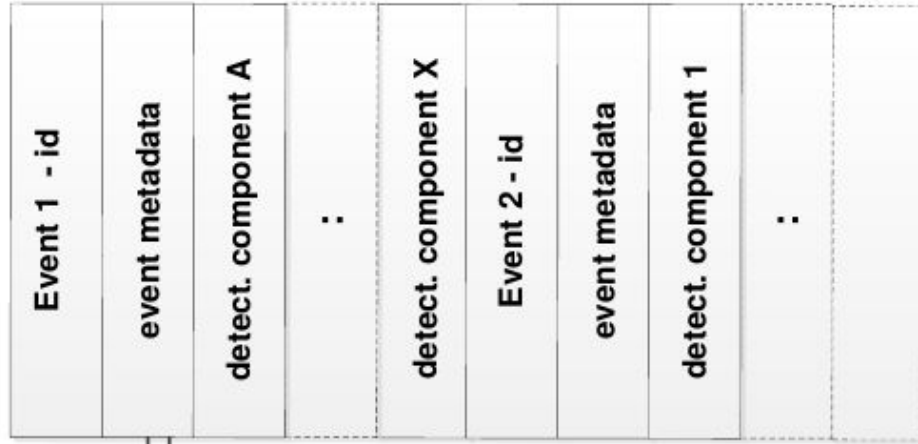
The materials are at <https://uni-muenster.sciebo.de/s/VNenNLMwxopn1kR>

Contact me: victoria.tokareva@kit.edu

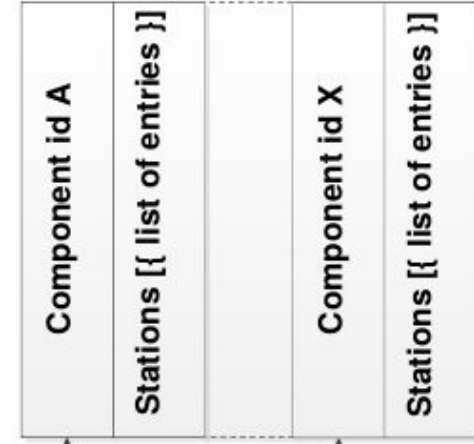


MongoDB data storage structure

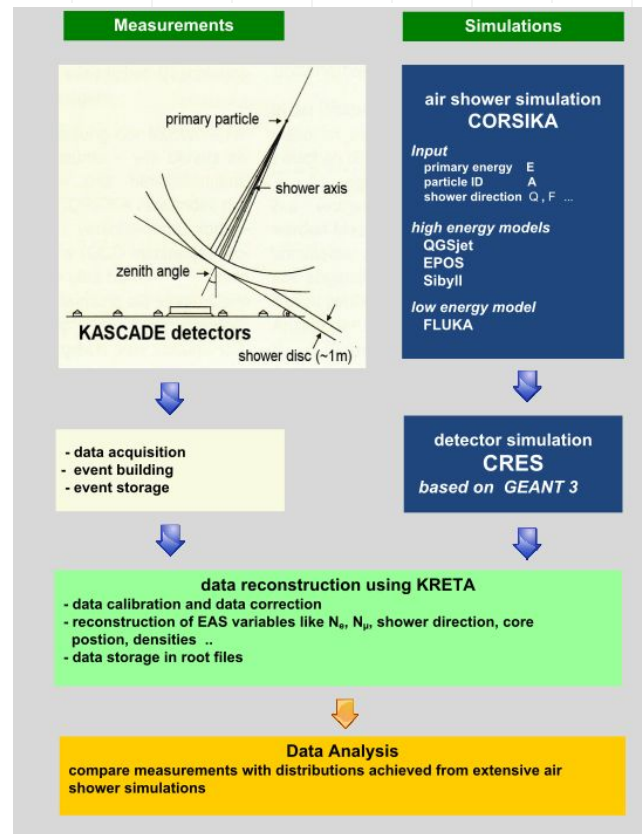
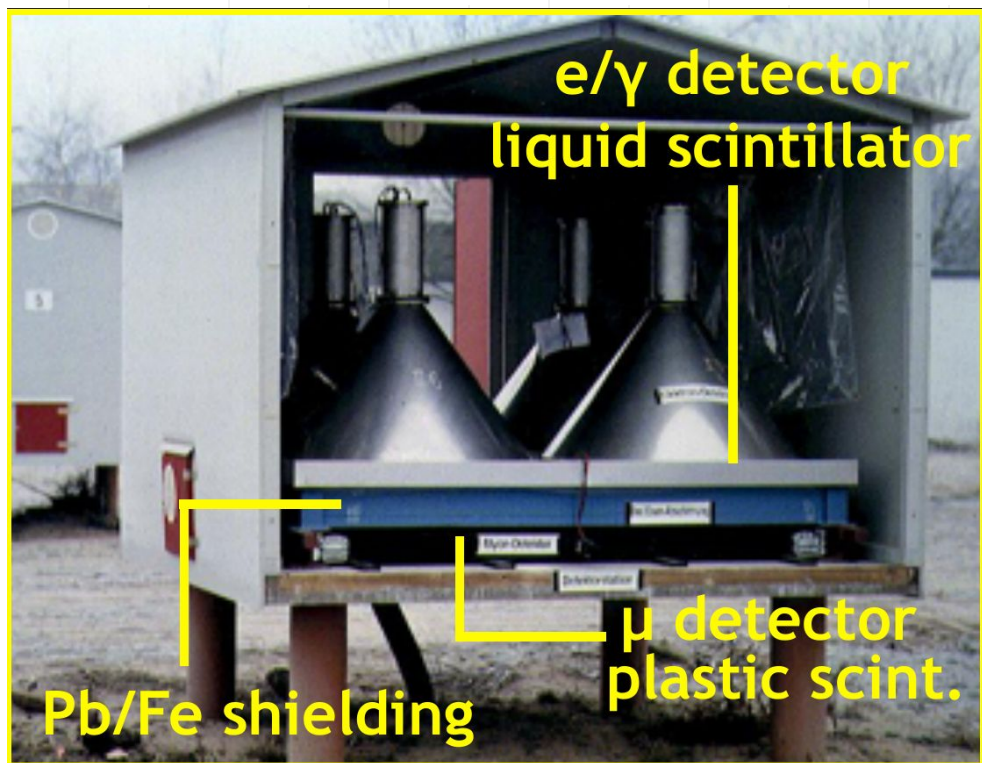
,DATA' Collection



,ARRAYS' Collection



Wochele, D., Wochele, J., Polgart, F., Tokareva, V., Kang, D., & Haungs, A. Data Structure Adaption from Large-Scale Experiment for Public Re-Use. CEUR-WS (2019) 2406, 114



Data quantities example

CALORIMETER Quantities

Var	Name	Available Data Range	Unit	Representation
Nhad	Nr of Hadrons	0. - 511.		
Ehad	Hadron Energy Sum	0.; 1.e10 - 1.e16	eV	log10 -> 10.0 - 16.0

GRANDE Quantities

Var	Name	Available Data Range	Unit	Representation
Xc	X-Core Position	-500.0 - +100.0	m	
Yc	Y-Core Position	-600.0 - +100.0	m	
Ze	Zenith Angle	0.0 - 40.0	°	
Az	Azimuth Angle G	0.0 - 360.0	°	
Nch	Number of charged part	11111. - 1,000,000,000.		log10 -> 4.0 - 9.0
Nmu	Number of Muons	1500. - 100,000,000.		log10 -> 3.2 - 8.0
Age	Shower Age G	-0.385 - +1.485		
GDeposit	Energy Deposit charged	0.0 - 100,000.0	MeV	/station
GArrival	Arrival Time	1000. - 10,000.0	ns	/station