PROCEEDINGS OF SCIENCE

Online Masterclass built on the KASCADE Cosmic ray Data Centre

Katrin Link,[a,*] **Victoria Tokareva,**[a] **Andreas Haungs,**[a] **Donghwa Kang,**[a] **Paras Koundal,**[a] **Frank Polgart,**[a] **Olena Tkachenko,**[a] **Doris Wochele**[a] **and Jürgen Wochele**[a]

[a] *Karlsruhe Institute of Technology, Institute for Astroparticle Physics, 76021 Karlsruhe, Germany*
E-mail: Katrin.Link@kit.edu

During the ongoing Covid-19 pandemic, people all over the world were forced to think about new ways of interacting with each other and this has especially challenged academics in their outreach activities with pupils. New online formats needed to be developed, and we used this opportunity to design and implement an (not only) online Masterclass using data from the KASCADE experiment. The masterclass is built on the KASCADE Cosmic Ray Data Centre and uses Jupyterhub and Notebooks for data analysis. We gained first practical experience during the International Cosmic Day with students at the age of 14-19 years. The Masterclass includes lectures on cosmic ray physics and data analysis, which are then consolidated in a hands-on part. By performing a cosmic-ray composition analysis on KASCADE data, the participants gain experience in using the KCDC open data web platform, working in the Jupyter environment, preprocessing data from a real astroparticle physics experiment, programming Python and performing exploratory data analysis.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)1378

1. Introduction

In November 2020 we developed an online Masterclass built on the KASCADE Cosmic ray Data Centre (KCDC) [1]. Continuous enhancement of BigData infrastructure at the Institute for Astroparticle Physics at KIT [1, 2] made it possible to provide online computation opportunities for both, scientific use and education. That came particularly handy in 2020 due to the Covid-19 situation. In order to continue with our outreach activities under such unusual circumstances, we developed this online masterclass based on an analysis of the data collected by the KASCADE-Grande [3] experiment.

KASCADE-Grande was an extensive air shower experiment to study the cosmic ray primary composition and the hadronic interactions in the energy range of $10^{15} - 10^{18}$ eV. The experiment was located at the Campus North of KIT and was running in different configurations from 1996 until the dismantling in 2013. Measuring simultaneously the electromagnetic, muonic and hadronic components of an air shower, KASCADE-Grande was able to determine the composition dependant energy spectra of cosmic rays in the transition region from galactic to extra-galactic origin. The discovery of the knee in the heavy component at around 80 PeV was one of the major results obtained with the KASCADE-Grande experiment.

The KASCADE-Grande collaboration followed the idea of open data and made the scientific data available through the KASCADE Cosmic ray Data Centre (KCDC) [1]. This is a web-based interface where initially the data from KASCADE-Grande was made available not only for the astroparticle community but as well for the interested public. Although the experiment is not running anymore, the data shop was extended with various releases in the last years and both the number of detector components from the KASCADE-Grande experiment and the data sets and corresponding simulations were increased.

2. Educational Goals

With this KCDC Masterclass, we want to address two main aspects. On the one hand the students should learn about the physics of cosmic rays: How they are produced and how we can detect them. On the other hand, they should learn how scientists handle a large amount of data: How it is processed and what kind of programming methods are used to analyze it.

This allows the students to experience the excitement of fundamental physics and answer questions about our universe while also providing insights into how astroparticle physicists, or scientists in general, are working and what challenges they are facing.

Additionally we work with data and simulations. On the basis of this we explain to the students, why we need simulations, how we use them and what this means for the interpretation of our results.

The implementation as an online course is of particular benefit during the still ongoing pandemic situation and brings worldwide equal access to education.

3. Requirements to the tutorial environment

In order to make the masterclass easily available for students with different experiences in programming and to reduce organizational costs, we formulated the following requirements:

- Identical workspace for all participants and tutors
- Save as much time as possible on installation and setup of the working environment
- Fast processing of rather significant amount of data
- Use of a data format, well-known for students
- Interactivity
- Minimal programming knowledge required

We considered the following possibilities for organizing online analysis (Table 1).

Approach	Purpose	Pros	Cons
Excel	general	+ Intuitive interface + no programming	- Rather slow - Not good for big datasets - Many things are either done manually or require knowledge of special functions and formulas
Python	general	+ allows to process large amount of data + Easy to learn	- Requires some introduction
C++ & ROOT	[Astro]particle physics	+ allows to process large amount of data + Highly optimized for [astro]particle physics + Gives a full professional experience	- May be quite cumbersome for beginners

Table 1: Considered solutions for online analysis

Taking into account the possibilities considered, we have arrived at the following technical solution: We use Jupyter Notebooks (see section 4) with Python3 kernel. For handling the data we have chosen the libraries pandas and matplotlib. Pandas is a popular library for data manipulation and analysis, which operates with so-called dataframes - tables of data, which our students could associate with familia Excel tables. Matplotlib is a widely used python library for data visualisation. Among the available data formats we have chosen .txt as it is well known for our students.

4. Jupyter Interactive Environment

Jupyter Notebook [4] is an environment, which allows novice developers, data analysts, and students to get started coding in Python faster. It is a well-known solution, which supports usage of different programming languages, data visualisations and interactive coding.

For this tutorial we needed to run several Jupyter Notebooks at the same time. For this purpose we used JupyterHub [5].
JupyterHub is a multi-user server application with the ability to launch multiple Jupyter Notebooks at once. It is great for organizing online tutorials as well as for practical analysis. The user only needs a browser, so we can eliminate problems with software installation, packages compatibility and make sure all students work in the same environment, where all the tutorials exercises are reproducible. The JupyterHub instance at KIT/IAP [6] supports a login via KCDC credentials and supports different kernels, pyRoot and C++ ROOT

5. Data analysis

We developed two tutorials for different prior experience, one on "Data pre-processing and exploratory visualisation using KASCADE data" and a more advanced one on "Determining the mass of a cosmic ray particle with the KASCADE detector". The implementation using Jupyter notebooks allows to guide the students through all steps of the analysis, starting with reading the data, over producing simple plots for receiving a physics result. The first tutorial focuses on understanding and plotting the data, starting with simple example codes and minimal prior knowledge on Python is needed. The second tutorial includes less introduction and the students will have to code several parts by themselves. For this tutorial prior programming knowledge is mandatory.
For both tutorials it is necessary to introduce the Experiment and the basic concepts of air shower reconstruction in advance. This should be done by either a scientist or an experienced teacher. In the following we explain the two tutorials in more detail.

5.1 Data preprocessing and exploratory visualisation using KASCADE data

The tutorial starts with reading a data file with measured and reconstructed KASCADE data and the students learn about the given parameters such as particle ID, energy, estimated zenith and azimuth angle of the particle's arrival direction, shower core coordinates and also additional parameters like temperature and pressure. Then they start with preproccesing the data by e.g. removing some unphysical values. This is repeated using simulations.
As next step the first 1d-histograms are produced to compare the zenith angle of data and simulations. The results are shown in figure 1. Here one can start the first discussion about the differences between data and simulations.
The students are requested to also plot further parameters and discuss the results. They are also shown, how to fit a normal distribution. Depending on the prior knowledge of the students one could go into deeper discussions at this point.
Using the shower core parameters x and y, the 2-dimensional histograms are explained and the difference between data and simulations can be further discussed.
In the simulated data the particle type is given and the students can plot the number of particles for each simulated primary, as shown in figure 2. Finally they are explained the relation between primary particle mass and muon number to electron number ratio and draw the muon to electron distribution for different simulated primary particles and finally for the measured data, see figure 2.
Using this final plot one can discuss, how we use this ratio to determine the primary mass.

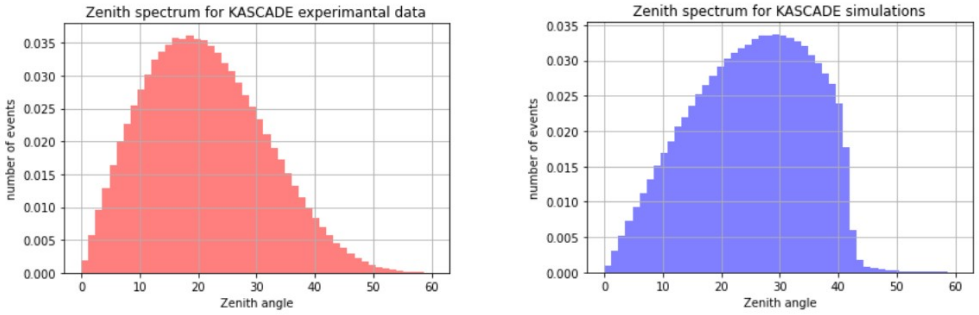

Figure 1: Zenith angle distribution for data (left) and simulations (right). For simulations only showers up to 42° are simulated and a spectral index of -2 is used.

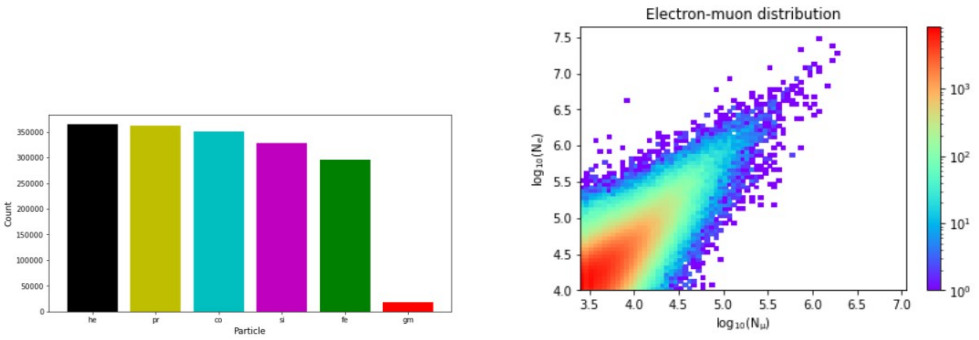

Figure 2: Left: Number of particles simulated for each primary mass. Right: Muon to electron distribution of the KASCADE data sample.

5.2 Determining the mass of a cosmic ray particle with the KASCADE detector

For this tutorial a different data set is used, with less parameters, only energy, muon number, electron number and zenith angle, but more statistics, i.e. we use around 4 million measured KASCADE events. After reading the data and plotting all parameters individually, the moun to electron ratio is plotted and further analysed.The students draw an upper and lower line associated with the expected primary mass. The resulting plot is shown in figure 3.

As additional exercise the students are advised to use simulated data and to plot the muon to electron ratio for two different primary masses.

6. First Outreach event

In November 2020 the International Cosmic Day [7] took place during the "Woche der Teilchenwelt", a week full of events celebrating the 10^{th} Birthday of the network "Netzwerk Teilchenwelt"[1]. At this occasion we hold the KCDC Masterclass for the first time. Eight students at the age of 14-19 years from a bilingual school in Austria took part. All of them were interested in physics and curious to learn about particle and astroparticle physics. One student had already gained experience in (Astro-)Particle Physics and was part of the organizational team, including doctoral students and

[1]Netzwerk Teilchenwelt www.teilchenwelt.de

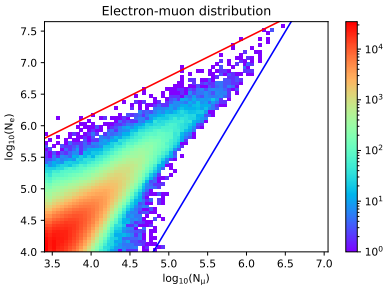

Figure 3: Muon to electron distribution of the KASCADE data sample with an upper and lower line associated with the expected primary mass. The upper line (red) represents light particles and the lower line (blue) represents heavy particles.

PostDocs from KIT. Due to the pandemic situation and the far distance the masterclass was held online via Zoom. The masterclass started in the afternoon with a welcome by the organizers.
An overview of the program is shown in Fig. 4.

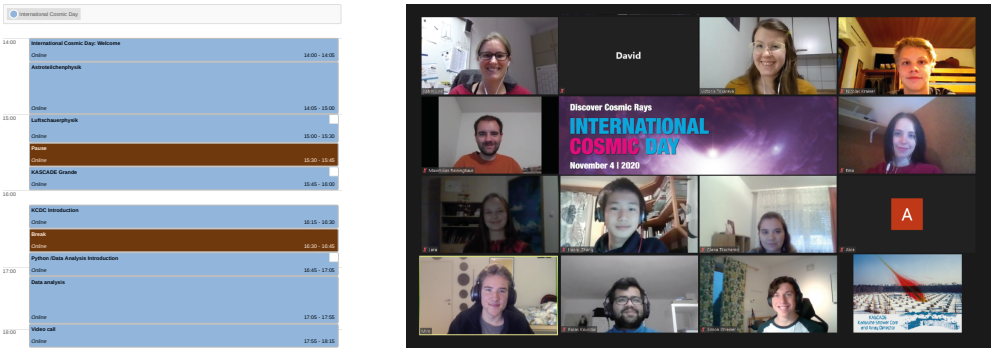

Figure 4: Timetable and participants of the first event based on the KCDC masterclass during the International Cosmic Day in November 2020.

The experienced student gave an overview of Astroparticle Physics. This was followed by a more detailed description of extensive air shower physics and how they are measured, and an introduction to the KASCADE-Grande experiment.
After these physics-related talks, the students got a short introduction to the usage of the KCDC portal. It was also explained how to request and download data from there. In principle, the students were asked to create an account in advance, which was only partly done. During the following break, the students could still create an account if necessary and try their first data request.
Since most of the students did not have any experience in programming, they got an introduction on using JupyterHub environment and Jupyter Notebooks and some basic Python programming was explained.
Now the analysis part started, for which the students were divided into smaller groups of two or three people. Each group got one breakout-Session and was supported by a Ph.D. student. The analysis part is divided into two tutorials, a more basic one on data processing and visualization and an advanced one on reconstructing the primary mass. Both were realized in individual notebooks. After they login into the JupyterHub environment, they could start the first notebook. Two different

data sets were already provided on the server, one with measured KASCADE data and the other with corresponding simulations. After reading the files, the two data sets were explored in more detail, and the students should investigate the distributions of their parameters to understand what these data have in common, and what is different and why.
The students learned about the parameters given in the data file, how to manipulate the data using pandas, and plotting the data in different ways, e.g., 1d, 2d histograms and bar plots.
The last part of this tutorial discusses the dependence between the muon to electron distribution and the mass of the primary particle.
The advanced tutorial starts with similar discussion on the muon to electron distribution and its dependence on the primary particle's mass. It was shown how the mass of the primary particle can be determined from the ratio, and simulated data is used to show this dependence.
With the help of their tutors, all students worked on the first tutorial and those students with prior knowledge in Python could complete also the second tutorial. After the analysis session, the whole group met again and discussed the results with each other.
The last part of the day was a call organized by the ICD team with students from the UK, where both groups presented their results. As follow-up action, the students prepared an article for the ICD booklet which was distributed to all participants.
In this booklet, the student's conclusion was: "*Apart from the extremely interesting theoretical input regarding cosmic rays, the event gave us an insight into how data analysis works and how meaningful conclusions can be drawn from a large amount of data. This enabled us a sneak peek into how science is conducted in real life, which we greatly appreciated.*"
After this first experience, we improved the masterclass further and want to make it easier for others to adopt the KCDC masterclass for their outreach events.

7. Summary

The big data infrastructure of IAP at KIT and the experience of our team in the organisation of outreach events enabled us to organize an online data analysis masterclass. We implemented an educational program and a masterclass, we organized preliminary data selection and preprocessing and instructed the students about the work with KCDC and Jupyter Notebook in JupyterHub environment.
A first outreach event was used for beta testing our outreach methodology and analysis environment at the same time. The students acquired new knowledge and skills while working with open data and presented their results to other participants of the ICD. This first application was a great success and we are looking forward to further events.
With the KCDC Masterclass we want to offer students, teachers and other interested people the possibility to do a basic analysis using KASCADE data either on their own, or in cooperation with KASCADE-Grande scientists. The Masterclass is also available for scientific organisations which can make use of the Jupyter environment.
One advantage of this kind of masterclass is that the students learn to work with a large amount of data and directly get to know the open data principles. Additionally they can work with state-of-the-art programming language without the need of prior knowledge.
From the physics perspective the masterclass allows to reproduce a scientific result from KAS-

CADE. The combination of data and simulations shows the principles of air shower reconstruction and the students learn about the challenges of reconstructing primary particle's properties.
Thanks to the implementation in KCDC it is also possible to further expand the analysis for interested students. After the masterclass student should be able to download data from the web shop and start a Jupyter notebook. With programming knowledge and advise from a tutor, they could write their own codes and think of their own analysis.
As an extension we plan to expand our available materials with masterclasses that include machine learning, enrich our methodology with automated code validation capabilities, and further development of our big data infrastructure.
A detailed manual with all necessary instruction and the Jupyter environment will be, or partly already is, available at the KCDC website https://kcdc.iap.kit.edu. This is open for use to institutes, schools and individuals and will be used in the future for further outreach activities at KIT.

References

[1] A. Haungs, D. Kang, K. Link, F. Polgart, V. Tokareva, D. Wochele et al., *Status and future prospects of the KASCADE Cosmic-ray Data Centre KCDC*, in *37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07. 2021*, 2021.

[2] V. Tokareva and e. al, *German-Russian Astroparticle Data Life Cycle Initiative to foster Big Data Infrastructure for Multi-Messenger Astronomy*, in *37th International Cosmic Ray Conference (ICRC 2021), Online, 12.07. 2021–23.07. 2021*, 2021.

[3] W. Apel, J. Arteaga, A. Badea, K. Bekk, M. Bertaina, J. Blümer et al., *The KASCADE-Grande experiment*, *Nuclear Instruments and Methods in Physics Research Section A: accelerators, spectrometers, detectors and associated equipment* **620** (2010) 202.

[4] T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic et al., *Jupyter Notebooks-a publishing format for reproducible computational workflows.*, vol. 2016 (2016).

[5] L. Fernández, R. Andersson, H. Hagenrud, T. Korhonen, E. Laface, B. Zupanc et al., *Jupyterhub at the ess. an interactive python computing environment for scientists and engineers*, in *This conference*, 2016.

[6] F. Polgart, A. Haungs, D. Kang, D. Wochele, J. Wochele and V. Tokareva, *An analysis framework for KCDC*, in *DLC 2020: Data Life Cycle in Physics: Proceedings of the 4th International Workshop on Data Life Cycle in Physics, Moscow, Russia, June 8-10, 2020*, p. 111, 2020.

[7] M. Hütten, T. Karg, C. Schwerdt, C. Steppa and M. Walter, *The International Cosmic Day - an outreach event for astroparticle physics*, *arXiv preprint arXiv:1711.01441* (2017) .