APPDS: AstroParticle Physics Distributed Storage

A.Kryukov (SINP MSU)

kryukov@theory.sinp.msu.ru

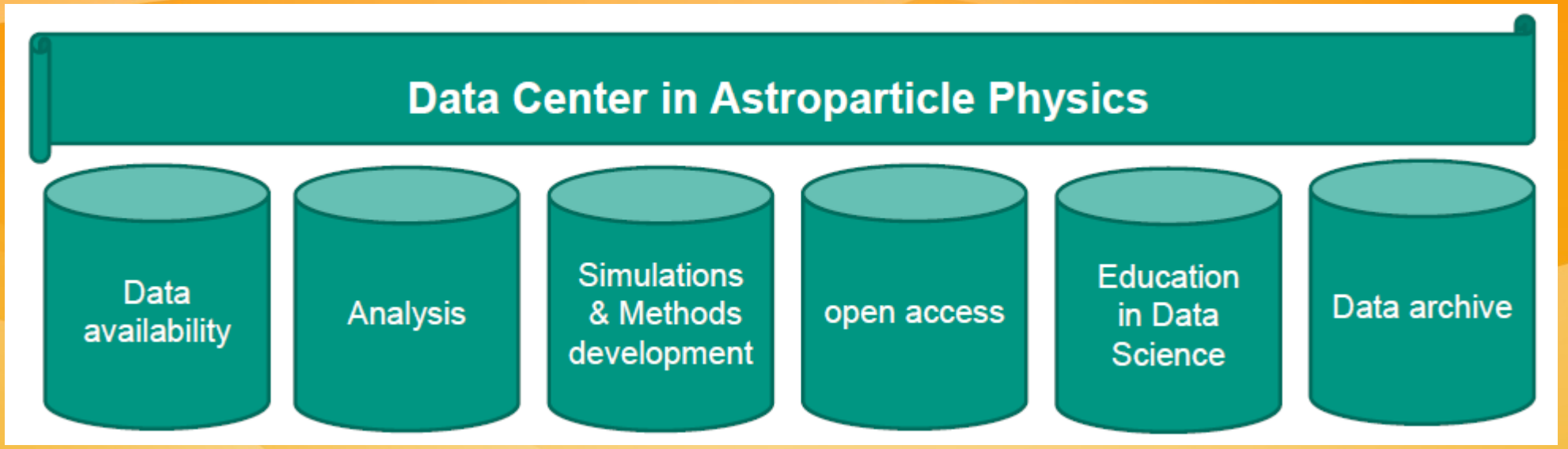
Content

- Introduction
- Data Life Cycle
- Main targets and approaches
- Architecture
- Software stack
- Conclusions

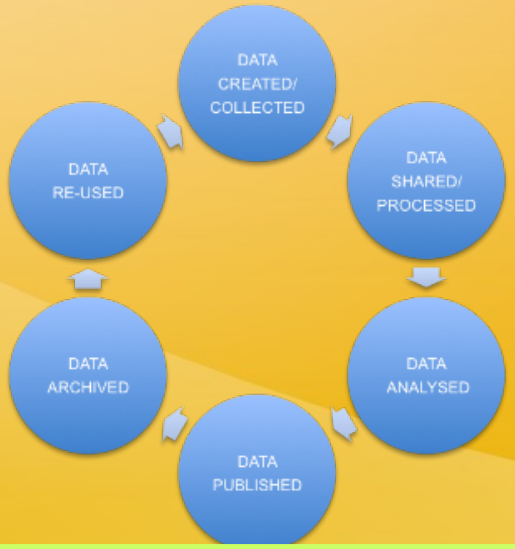
Karlsruhe-Russian Astroparticle Data Life Cycle Initiative

- Joint project is supported by RSF and Helmholtz
- Main participants:
  - Russia: SINP MSU, ISU, ISDCT SB RAS
  - Germany: KIT
  - Team leaders: A. Kryukov (SINP MSU) and A. Haungs (KIT)
- Duration: 2018-2020
- Financial request
  - RSF – 18*10^6 Rub.
  - Helmholtz – 390*10^3 Euro
- **Start of the project – Jan. 2018**

Data life cycle



- Data created/collected
- Data shared/processed
- Data analyzed
- Data published
- Data archived
- Data re-used



Main targets of the project

- The project will strive to develop an open science system to be able to collect, store astrophysical data having the TAIGA and KASCADE experiments as an example and provide it for analysis.
- The novelty of the proposed approach can be seen in developing integrated solutions including:
  - distributed data storage with a common meta-catalog to provide a common information space of the distributed repository;
  - data transmission from several data repositories and aggregating data “on fly”. Thus significantly reducing load time

Main targets of the project (cont.)

- development of machine-learning techniques for identifying initial particles and their properties;
- installation of the KCDC-based prototype system of Big Data analysis and exporting the experimental data from KASKADE and TAIGA for testing technology of data life cycle management.

- We will also create an educational system on the HubZero platform dedicated to astroparticle physics.

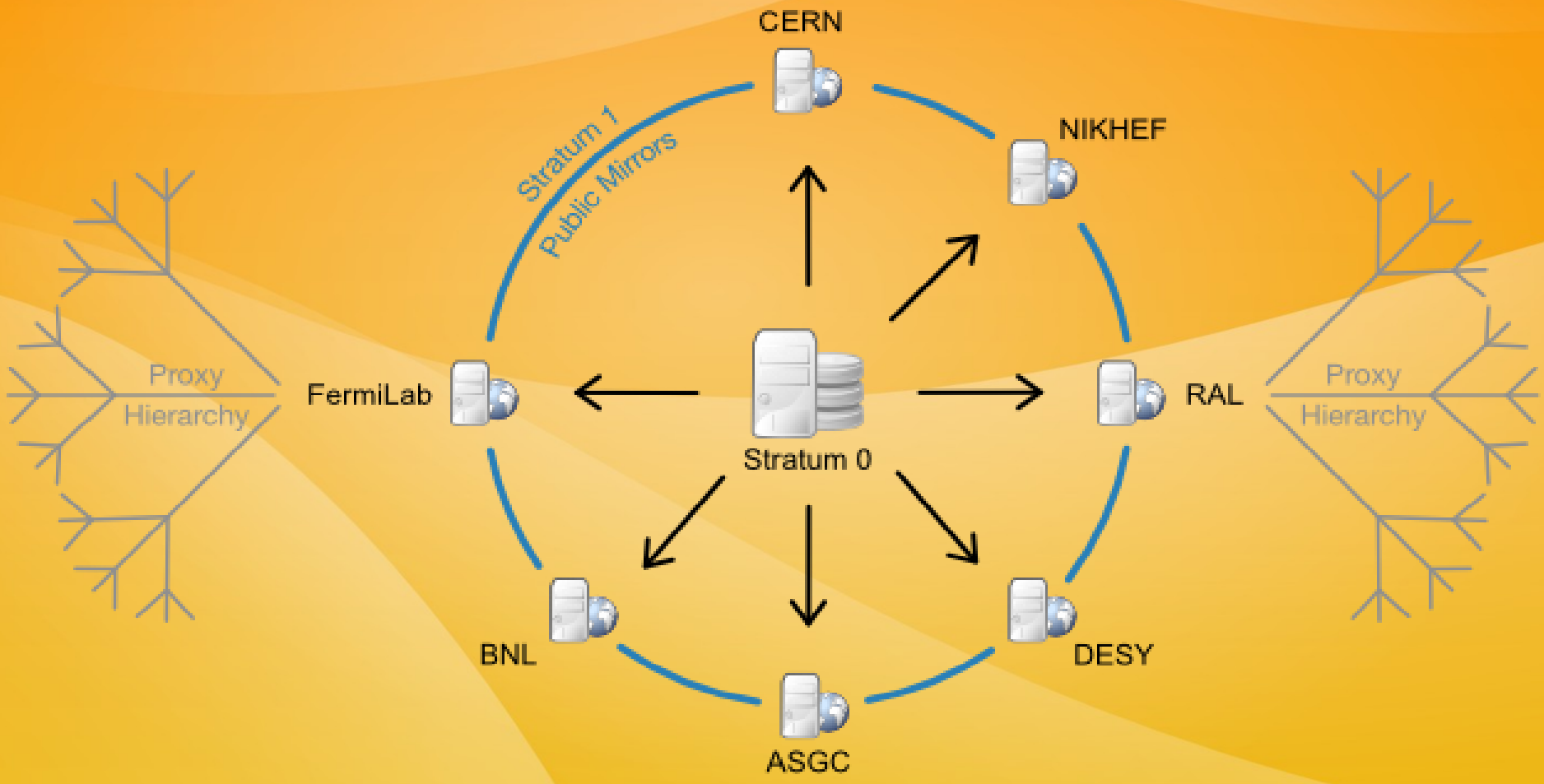
Data Availability

- The main goal of the project is to provide scientists with data on requests. The request is a set of conditions and logical operations on them which define what kind of the data the user want to obtain.
- All requests proceeds by using the metadata (MD) information only via special MD servers. A search within the data will not be available. If one needs to carry out more sophisticated requests, the appropriate information must be extracted from the data and inserted into the MD registry.

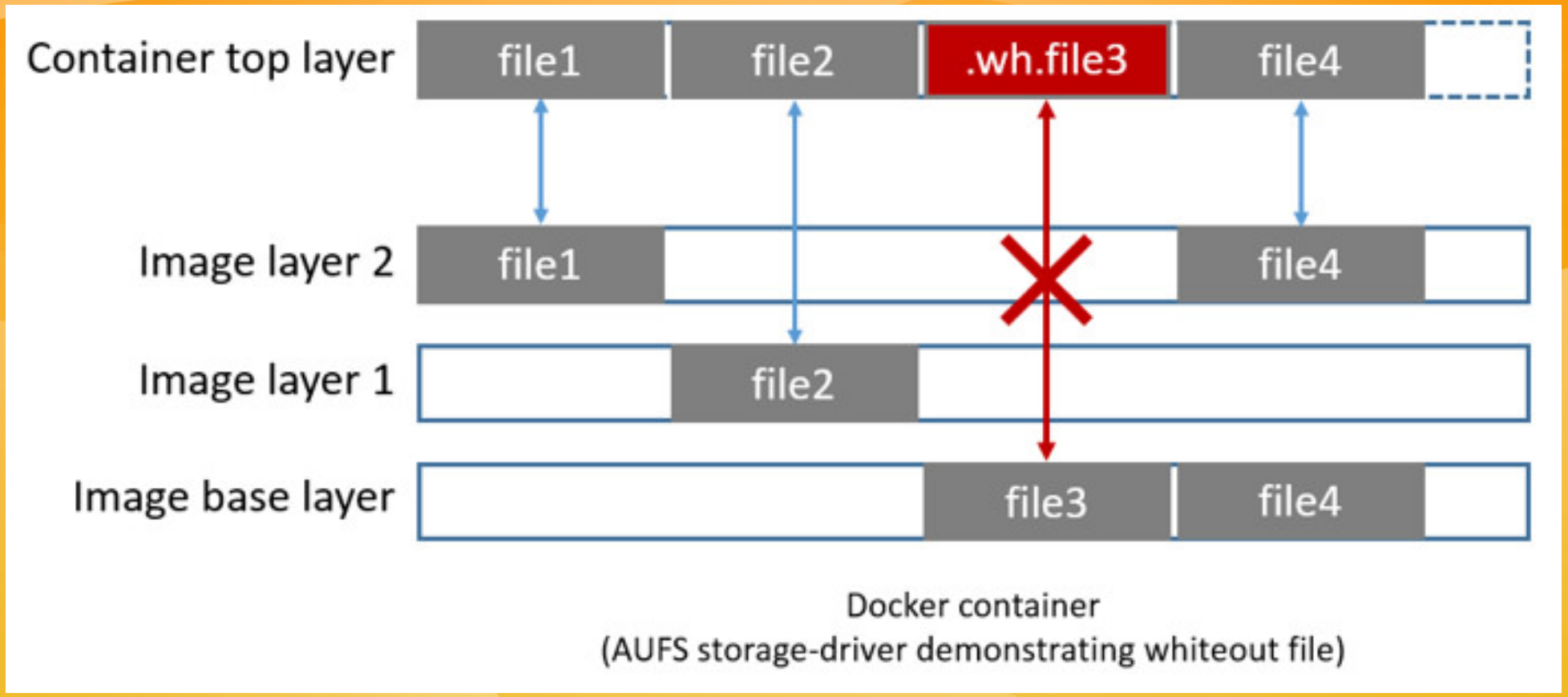
Data Availability

- The data itself is stored on some local data storages which collecting raw data from the astroparticle facilities such as TIAGA, KASCADE, etc. Each storage has its own format of data storing, directory structure and policy. We do not touch the internal structure of storage and traditions of physics community. Therefore, to provide access to the data storage, it will be necessary to deploy a (RESTful) services which will unify the external interface for all the storages. We will call such a service an adapter.
- One of candidates for the adapter is the CERNVM-FS service[https://cernvm.cern.ch/portal/filesystem]. This service provides export of a local file system over Internet in read-only mode. For tasks of the data analysis this mode will be enough.
- From the end-user point of view the set of data storages will look like a single virtual storage.

CERN VM-FS


CERN
NIKHEF
Stratum 1
Public Mirrors
Proxy
Hierarchy
FermiLab
Stratum 0
RAL
Proxy
Hierarchy
BNL
DESY
ASGC

Update files (Union-fs)



Data Analysis

1. Conventional analysis. In this case a user writes some algorithm which inspired by physical model of the phenomena under consideration.

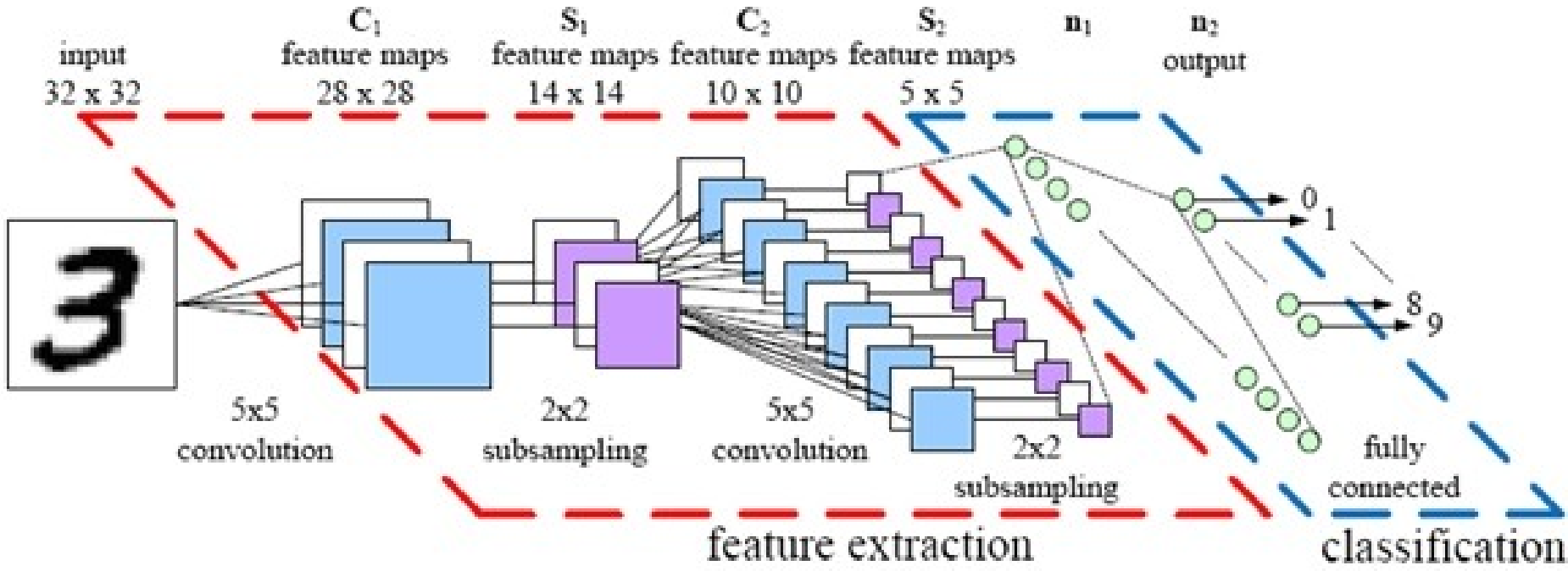
2. Machine learning (ML). In this case one uses an artificial neural network (ANN) technique with supervised or unsupervised learning.

Let us consider the problem of gamma identification in events registered by Cherenkov telescope.

The traditional approach assumes calculation of the main axes of the ellipse (event image). The identification of gammas is based on the statistical relation of the axes length ratio for gamma and non-gamma events.

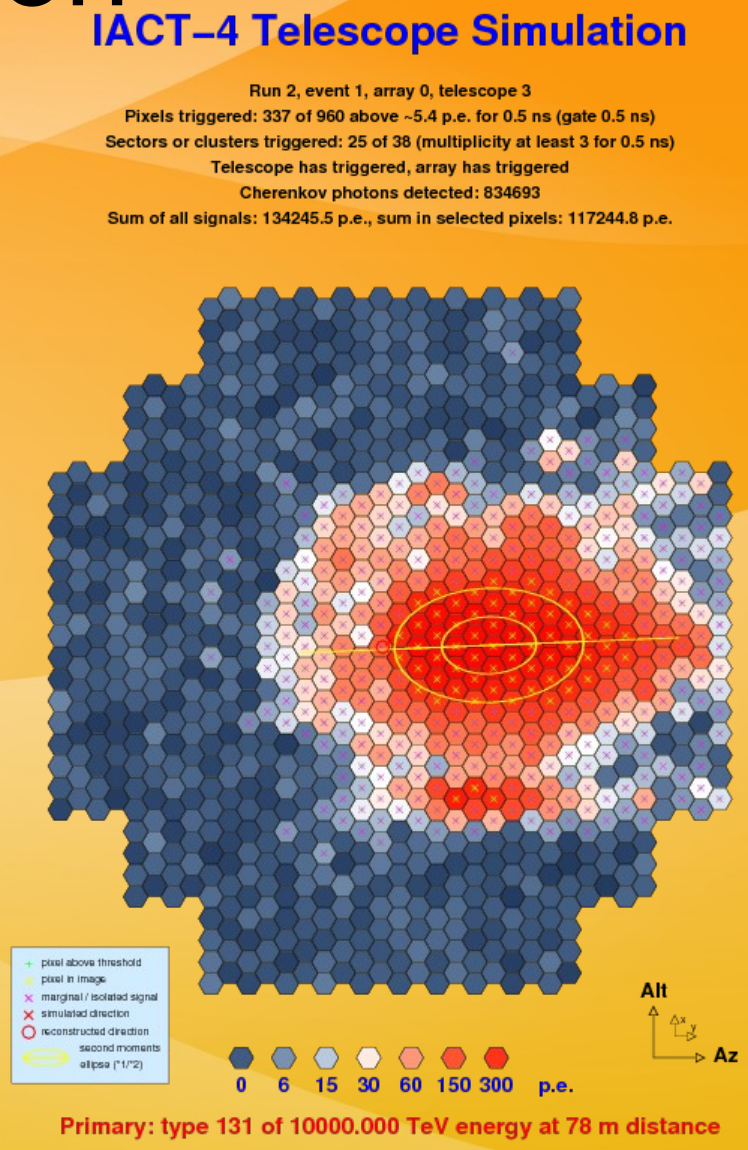
The ML approach supposes ANN learning on a huge set of known events and classification of unknown events after that.

Convolutional NN



Simulation

Simulations is one of the important stages in modern experiments. They require a lot of computer resources and produce data volume which is comparable with the volume of raw data. Usually simulations proves to be “data factory” similar to the experimental facilities. So, we propose to consider the simulations as a specific source of data (like experimental facilities). Thus simulation data should be uploaded to the particular storage by special service.



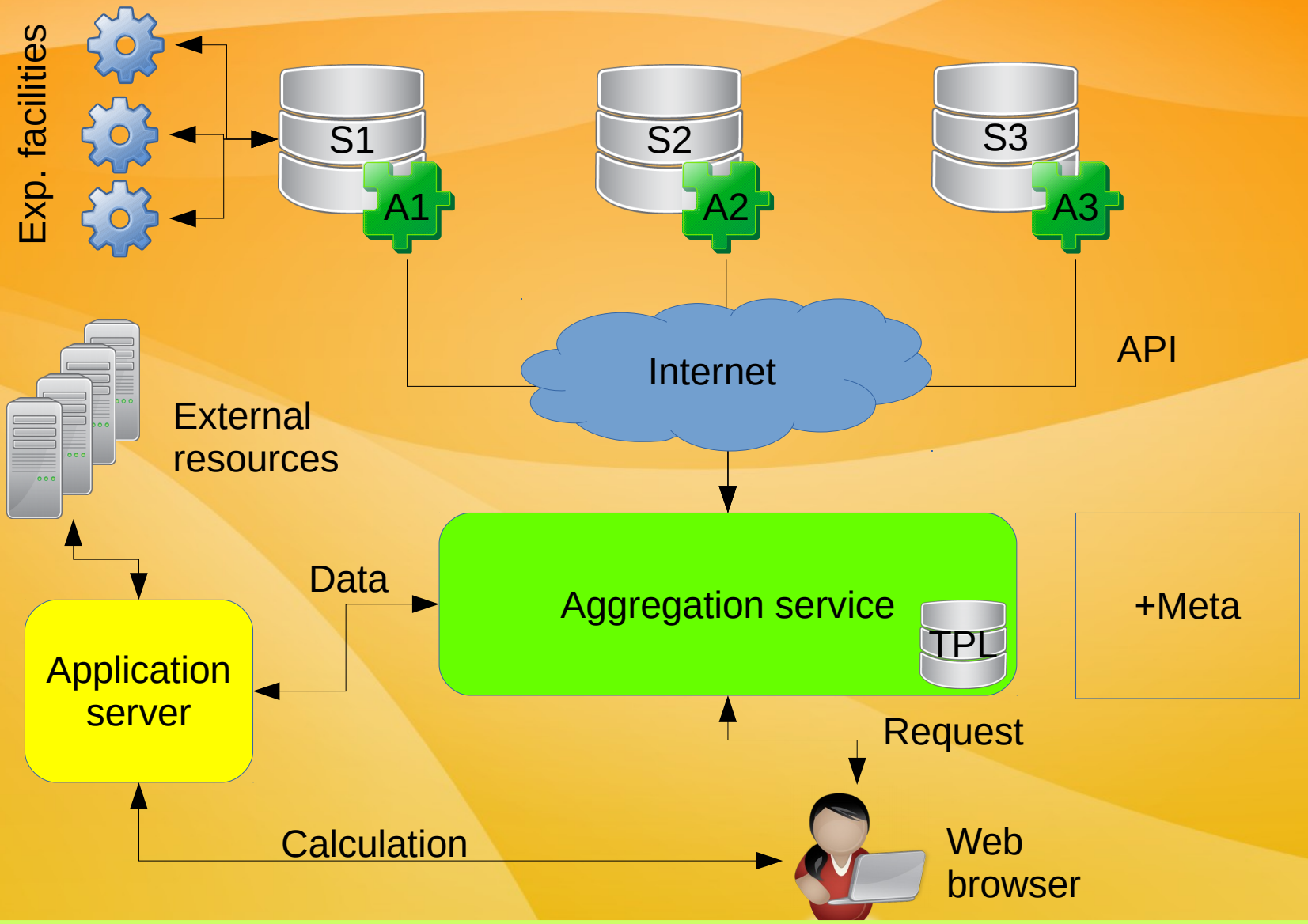
Open Data

- The open access to the data is provided by the standard way under specially formulated access policy. The policy depends on the local policy of integrated storage which are data owners.

Education

- We are going to achieve this target by using special service (Web portal) based on the HubZERO [https://hubzero.org] platform. The platform will supply users with education courses, documentation and exercises on MC simulation, examples of data analysis and so on.

Architecture of APPDS (draft)



THANK YOU!