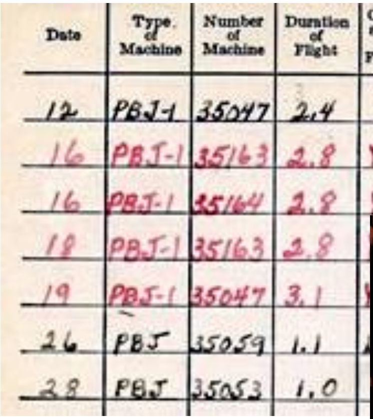
LUND
UNIVERSITY

Principles of scientific data management, big data workflows

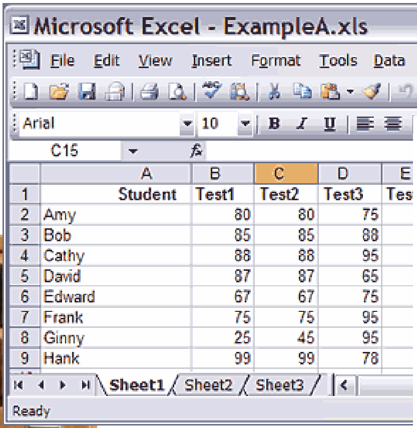
COMPUTE RESEARCH SCHOOL COURSE NTF004F

Research starts with data

- The ultimate goal of science is to understand natural phenomenae
  - Understanding leads to anticipation, reproduction, prevention, utilization etc
- Information is key to understanding
  - **Data** is information organised in a structured manner
    - » There are very many ways of structuring information

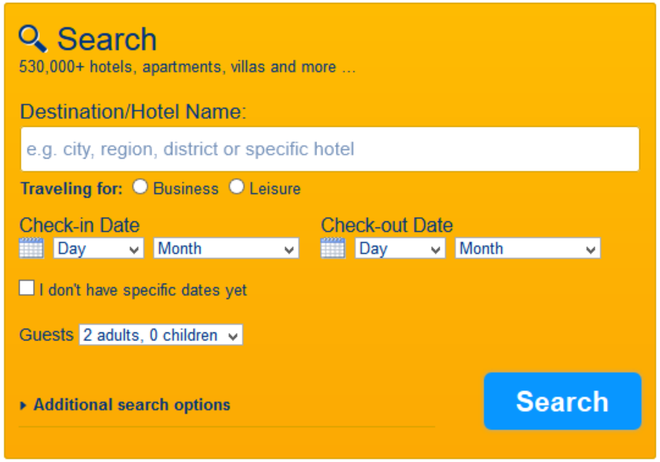


Date	Type of Machine	Number of Machine	Duration of Flight	Character of Flight
12	PBJ-1	35047	2.4	X
16	PBJ-1	35163	2.8	YX
16	PBJ-1	35164	2.8	YX
18	PBJ-1	35163	2.8	
19	PBJ-1	35047	3.1	
26	PBJ	35059	1.1	
28	PBJ	35053	1.0	



Microsoft Excel - ExampleA.xls
File Edit View Insert Format Tools Data Window Help
100%
Arial 10 B I U
C15 fx
Sheet1 Sheet2 Sheet3
Ready

	A	B	C	D	E
1	Student	Test1	Test2	Test3	Tes
2	Amy	80	80	75	
3	Bob	85	85	88	
4	Cathy	88	88	95	
5	David	87	87	65	
6	Edward	67	67	75	
7	Frank	75	75	95	
8	Ginny	25	45	95	
9	Hank	99	99	78	



Search

530,000+ hotels, apartments, villas and more ...

Destination/Hotel Name: e.g. city, region, district or specific hotel

Traveling for: ○ Business ○ Leisure

Check-in Date: Day Month

Check-out Date: Day Month

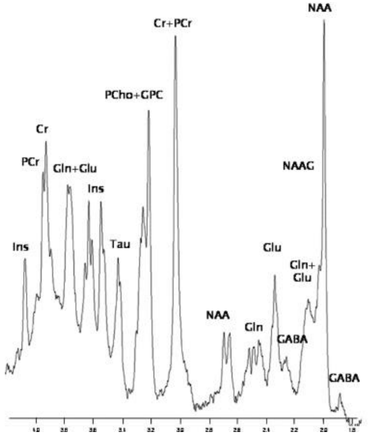
□ I don't have specific dates yet

Guests: 2 adults, 0 children

Research data differ by: size

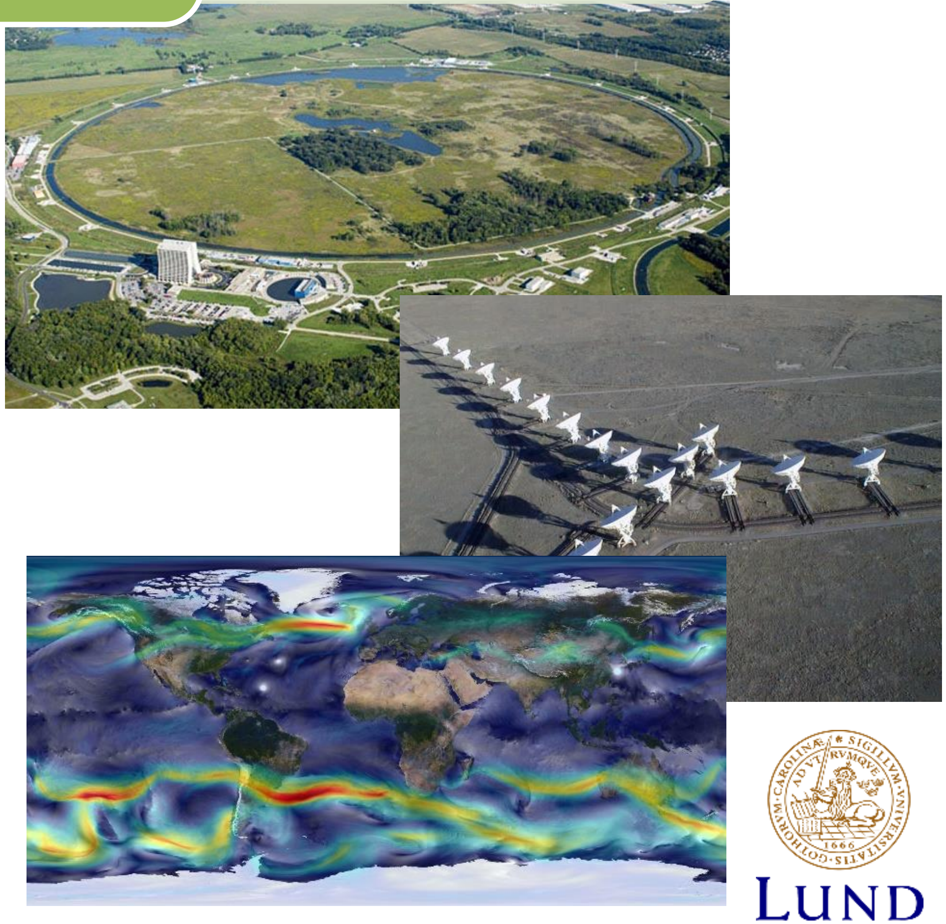
Small data

- Small devices
- Portable USB drives
- Personal computers



Large data

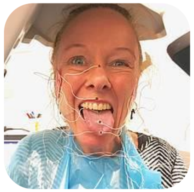
- Large devices
- Storage servers
- Supercomputers



Research data differ by: source

Observation:
often non-reproducible

Imaging



Recording of signals

Collection of samples

Measurements

Generation:
reproducible statistically

In vitro

In vivo

Synthesis

Simulation

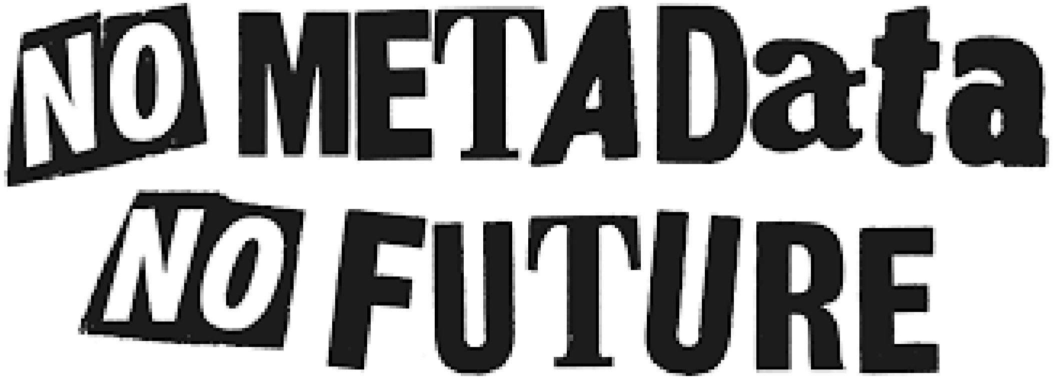
Natural sciences rely on measurements

- Exclusive measurement: focussed on one particular object or phenomenon, excluding all others
  - Example: register all photons emitted at a particular angle
  - Simpler experimental setup
  - Little data, simple analysis
- Inclusive measurement: register all the processes, objects etc
  - Example: digital sky survey (could produce 1 Exabyte a day, 1 EB = 109 GB)
  - More complex experimental setup
  - Lots of data, complicated analysis (“needle in a haystack” problem)
- Inclusive measurements can be filtered to exclude unwanted information
  - On-line threshold: minimal value of the measurement to be recorded
  - On-line trigger: a set of conditions that must be satisfied in order to record measurements
  - Off-line post-processing: derived data

Raw data, derived data, metadata, datasets

- **Raw data**: data as acquired by an experimental device or method
  - Examples: filled questionnaires, unprocessed satellite images, electronic hits in a detector
  - Raw data often contain unnecessary or excessive information, have large volume, and are recorded in different method-specific ways
- **Derived data**: data derived from raw data by applying various algorithms: filtering, compression, enhancement etc
  - There can be a chain of derived data
  - Derived data usually contain less information, but can also contain additional information as a result of processing
- **Dataset**: a collection of data characterised by common data taking conditions
  - Examples: same year, same object, same device settings etc
  - Data and data sets can be mutable (can be changed) or immutable (never change once recorded)
- **Metadata**: data about data, such as time stamps, data ownership, quick summary etc
  - Metadata often are stored together with data

Metadata is as important as data

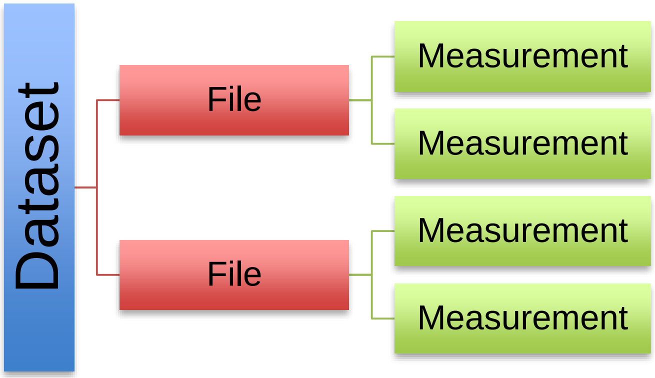


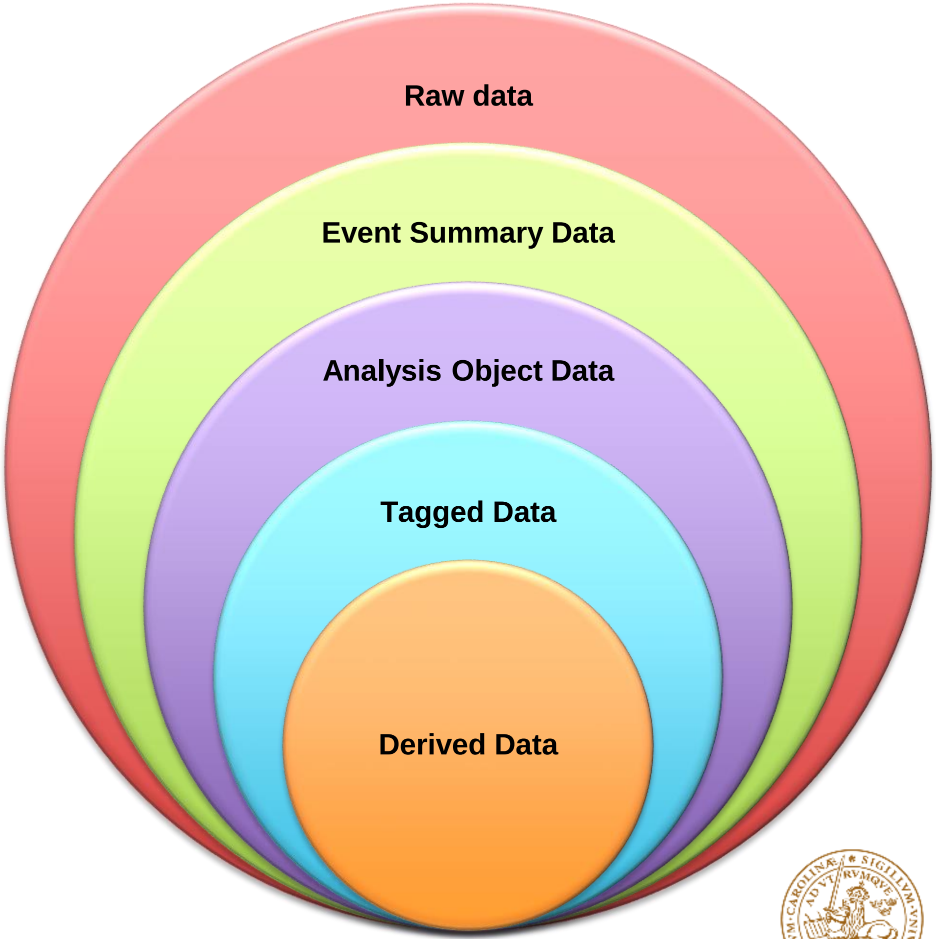
nometadata.org

- When data are many, it is easy to lose track
  - We can even end up with Dark Data
- Metadata are needed to:
  - Group and aggregate
  - Browse and search
  - Annotate and curate
  - Research: metadata is data, too

Example of data hierarchy: particle physics

- Different sciences use different data models
- Data are often recorded in structured files
- Each file contains many measurements
- Many files recorded in identical conditions constitute a data set
- Data sets are derived from each other: from raw data to analysis objects





Where are the data?

- Research data are often stored as **<u>files</u>**
- A dataset_may consist of a large number of files
  - Such files would typically have similar names
  - File names often contain metadata, e.g. `data14ver8nocalib.dat`

Data (not simulation)

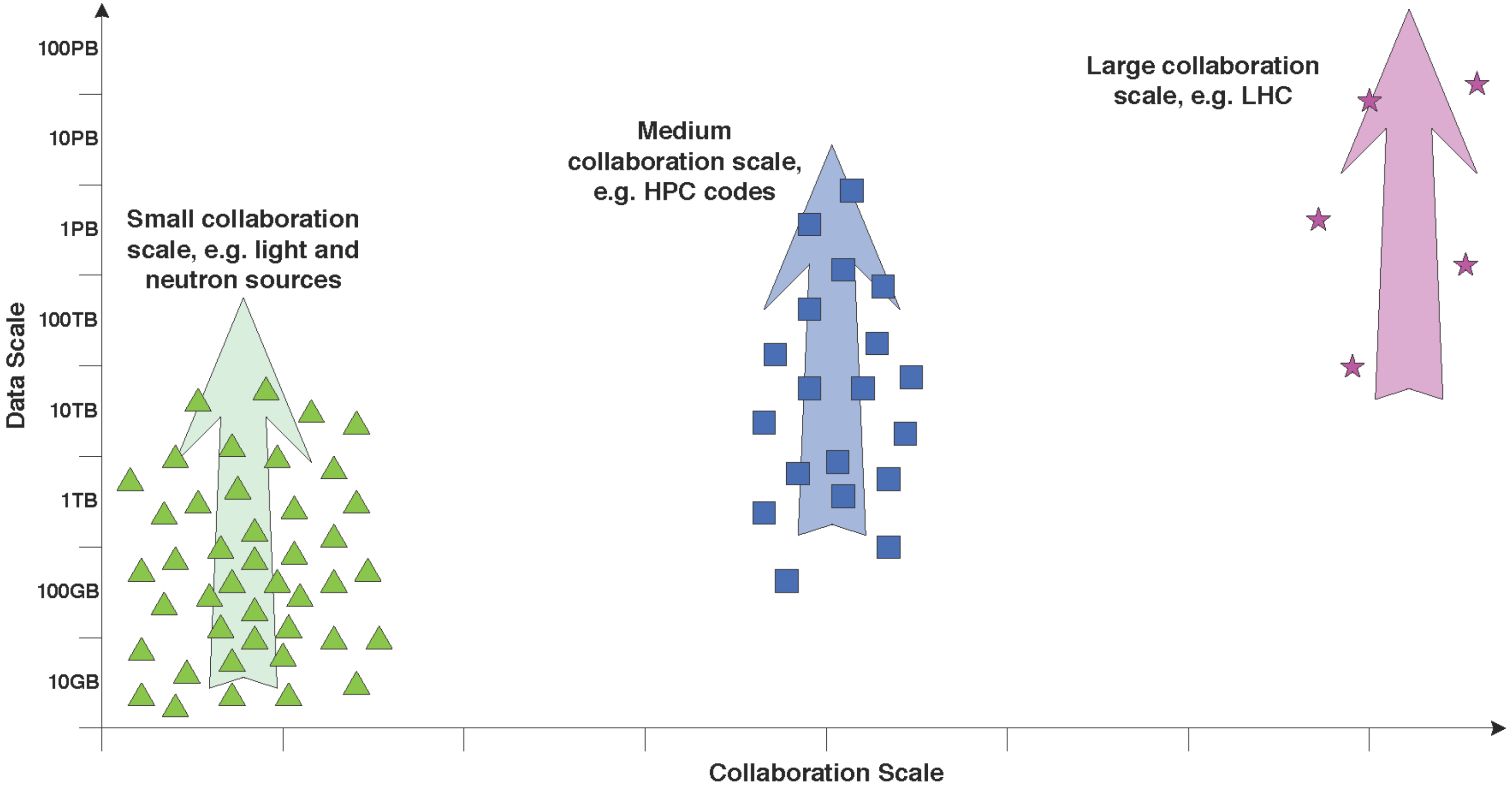
year

Software version

Non-calibrated

- There are many different ways of writing data to a file
  - Alphanumeric text files: strings or arrays of data and keywords, readable by any document processing utility
  - Binary files: packaged information to be read by a dedicated software
    - » Examples: JPEG pictures, Excel spreadsheets, ROOT files
- Data are also stored in **<u>databases</u>**
  - A database is a structured file (or set of files), interpreted by a specialized software
    - » Data from a database are read <u>directly</u>, from files – <u>sequentially</u>
  - Databases can establish <u>relations</u> between data objects
  - Databases are needed to enable quick access to large amounts of data
  - Typically, databases are hosted by specialised servers, and are accessed (*queried*) remotely, using special *query languages*
    - » Files are easy to copy and transfer, databases are not

Sizes of research data sets and scientists teams



Graph by Eli Dart, ESnet/LBNL

- Larger is data set, more scientists work on collecting and analyzing it
  - Need to follow common rules, have common software etc
- Petabytes and Exabytes of data are a reality today

So what is Big Data everybody is talking about?

- Generated data samples can be as big as we can afford
  - Run accelerator for 25 years
  - Run simulation on a supercomputer
- Observational data samples can be as big as our devices can handle
  - Telescopes with huge resolution
  - DNA sequencers with high rates
- Truly Big Data are harvested *in vivo* by social networks
  - Also smart devices, CCTV etc
  - Similarly to ethnographic data, are unstructured

Research Data Management

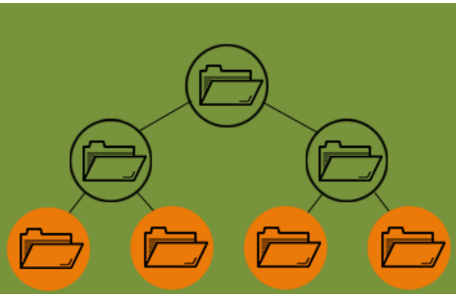
Storage considerations

- Large enough storage is needed
  - From Gigabytes for some to Petabytes for some others
- Large enough bandwidth connecting the instrument, storage facilities, data processing facility and the user
  - Terabit per second is possible, 10 Gbps is more likely
- Controlled write, read and list access
  - From password to certificates and access tokens
- Encryption: for data, or transfer, or both, or none
- Adequate space management
  - quotas, space recovery utilities
- Transfer, replication and migration tools
  - Per file, per collection, per database

Storage considerations (continued)

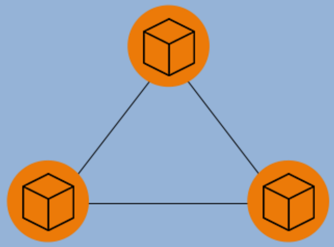
- Backup
  - a large range of requirements: from basic RAID to multiple replicas, local and remote, tapes or other media, with recovery of older versions in case of accidental modifications etc
- Indexing of what is stored
  - From basic POSIX information listing to metadata catalogs, adequately protected from unauthorised access
- Logical organization in terms of metadata-based grouping
  - possibly reflected in physical grouping for optimization
- Monitoring and statistics collection (accounting)
  - Various usage parameters as a function of time, access monitoring and such – all protected from unauthorised access

A note on different kinds of storage



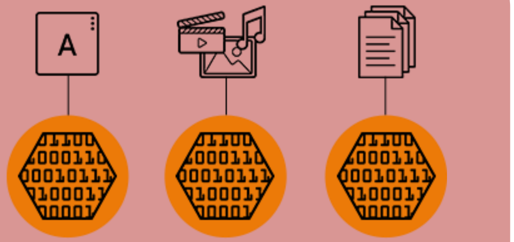
File storage

- Essentially, your regular disk, with hierarchy of directories and associated file-system metadata.



Block storage

- Data stored in volumes called “blocks”, blocks have labels but no metadata assigned. Fast, expensive, good for structured data (databases), RAID arrays, VM file systems.



Object storage

- No directories or blocks, data (files) and any metadata jointly constitute objects – excellent scalability but complicates modification of data. Good for distributed storage, Big Data.

Illustrations by RedHat

Data processing considerations

- Accessibility of data from a computing resource
  - streaming and/or caching
  - copying for a direct access: single files, collections, per job or per site
  - querying in case of databases
- Persistent and unique identifiers
  - per file, per collection, per database and such
- Mechanisms to match/resolve identifiers to physical addresses
  - For streaming, copying, querying and such
- Well-defined (formalized) data formats/structures and tools for conversion between (at least some of) them
  - Even analytics relies on certain structure conventions

Preservation considerations

- Continuous storage upgrade and media migration
  - Modern hardware has very limited life time
- Identifiable authorship and ownership of the data
  - per file, collection, database etc
- Provenance information
  - original data taking conditions, possible modifications, changes of ownership, changes of access rights etc
- Preservation of data format description
  - possibly encapsulated in data
- Experiment documentation
  - Gallileo was probably not the first to discover Jupiter moons, but he was the first to document the process and the result

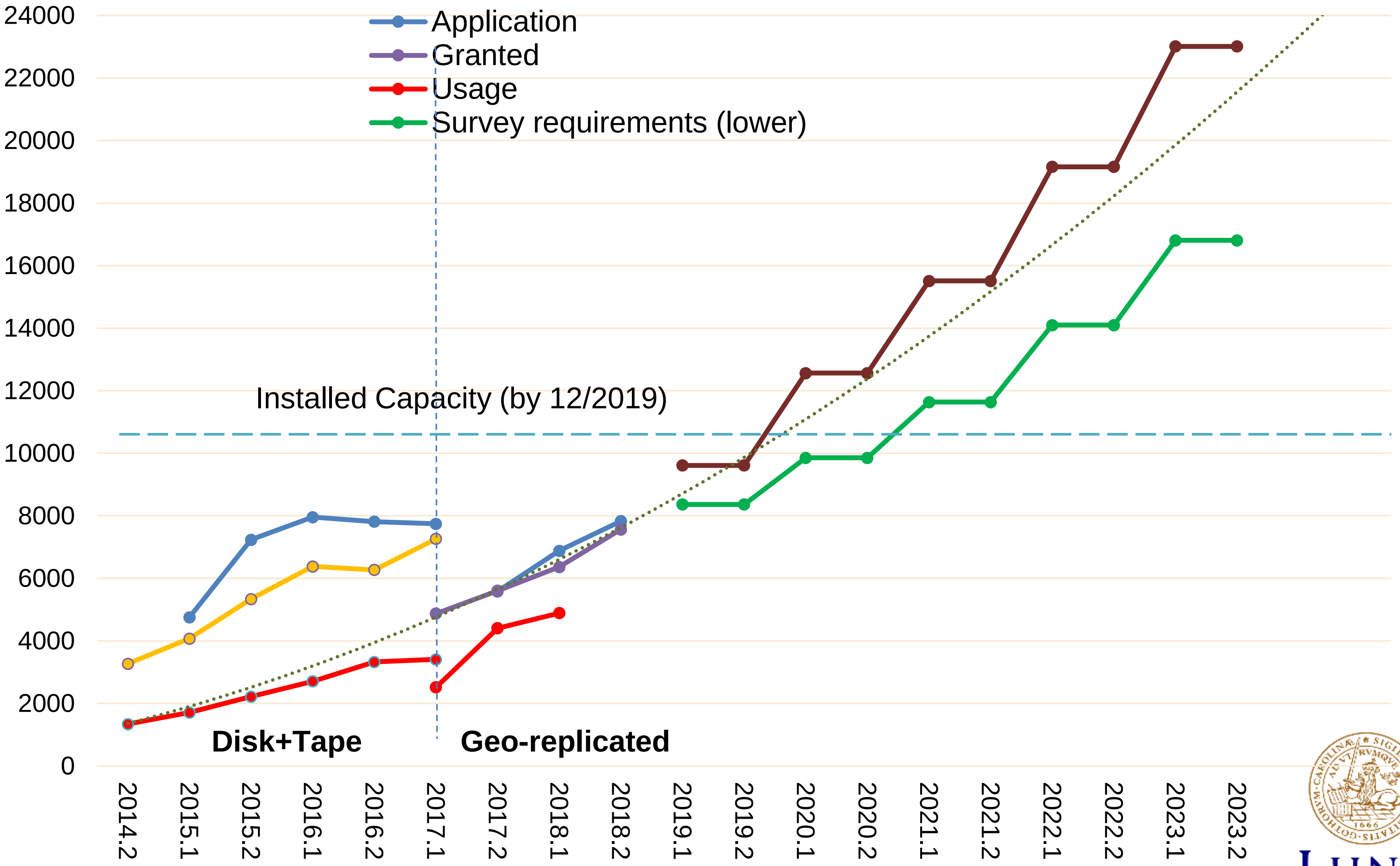
Preservation considerations (continued)

- Preservation of processing algorithms and/or workflows
  - Study shows that odds of data being analysable reduce by 17% a year
- Preservation of computing environments used to produce or process the data
- Preservation of metadata
- Preservation of accessibility:
  - if possible, preservation of protocols
  - when protocols change, consistent re-mapping of data identifiers to new protocols
  - long-term access rights management: granting write access to curators, migration to new security technologies, revoking access rights of non-authorized individuals, opening up for public read and list access

Sharing considerations

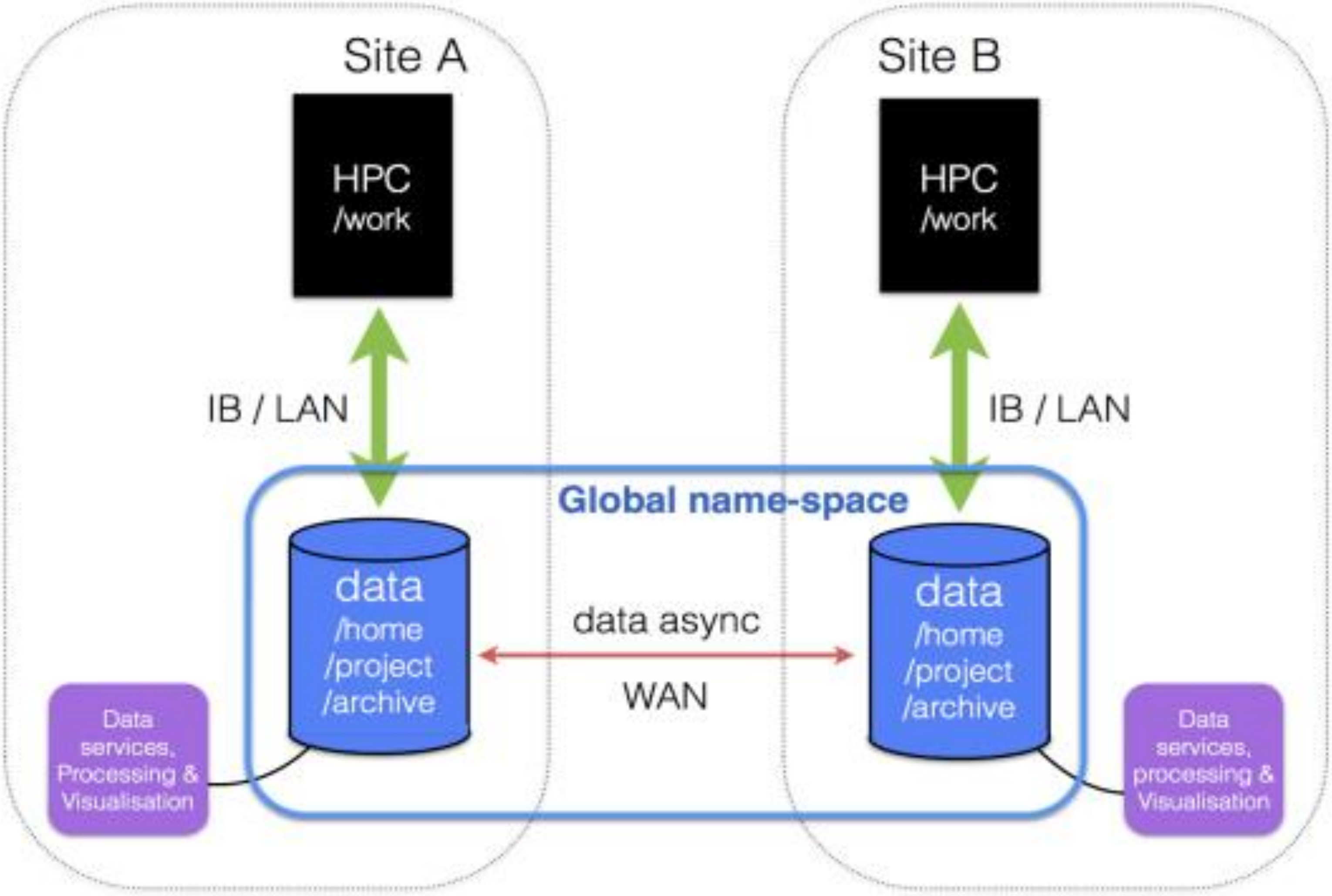
- Access control managed by authorized individuals
  - From dedicated managers for large research groups, to individual researchers who produce the data
- Networked access via industry-standard protocols and means
  - like http/webdav today, remotely mounted or synched file systems, etc
- Discovery tools relying on metadata
  - Including authorship, provenance and other info
- Upload tools
  - From simple copy or file-by-file transfer, to portals and other utilities dealing with logical collections, databases etc

Example: Research storage in Norway (NIRD)



Adapted from Maria Francesca Iozzi

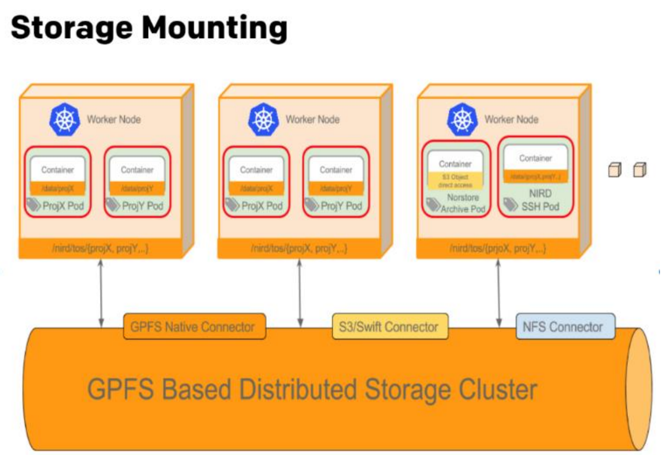
A NIRD distributed data centre model

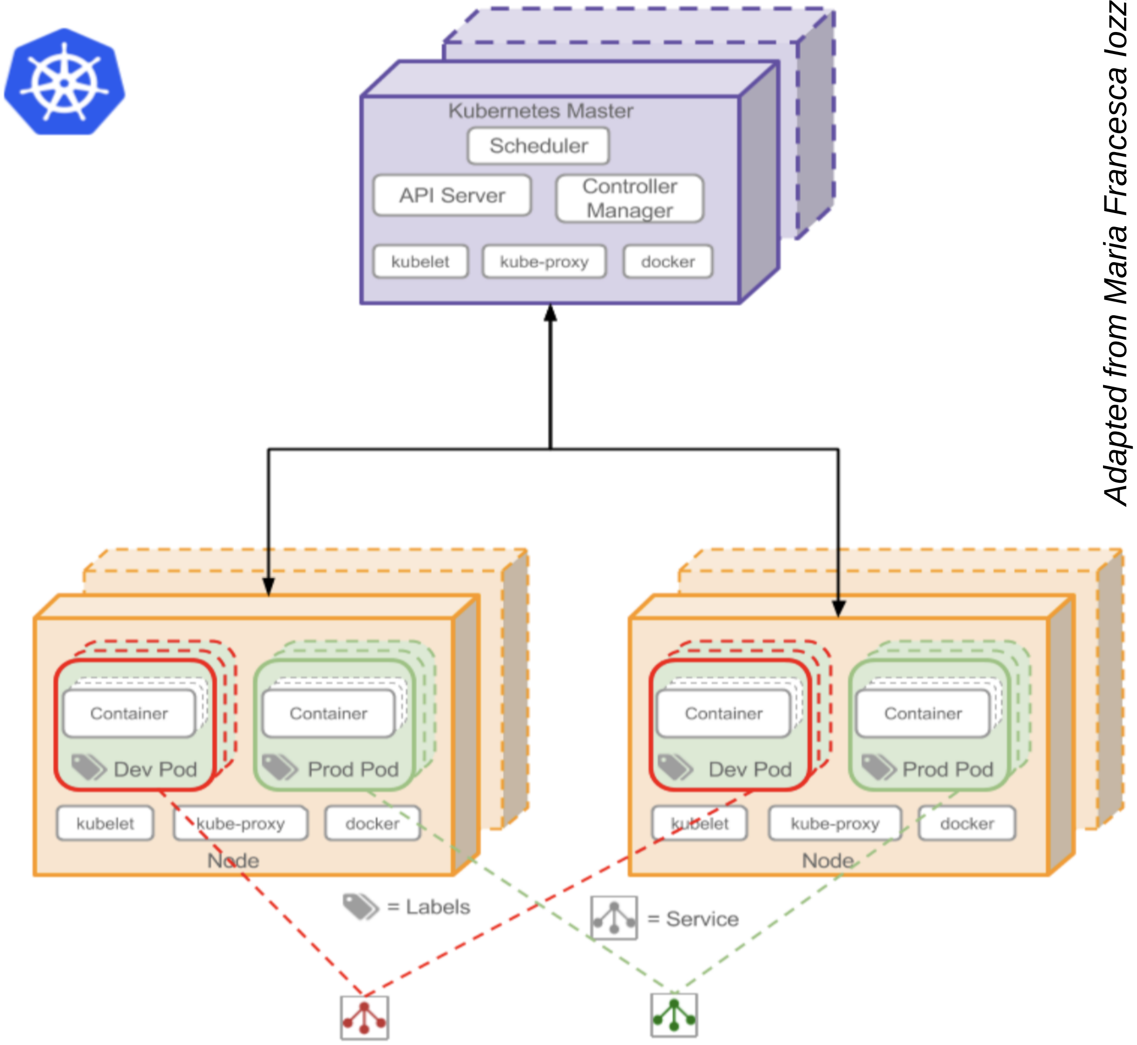


Adapted from Maria Francesca Iozzi

Data Processing: NIRD Service Platform

A **Kubernetes-platform** running on computing nodes that mount the GPFS distributed **storage**. Services run inside **Docker containers**.





Adapted from Maria Francesca Iozzi

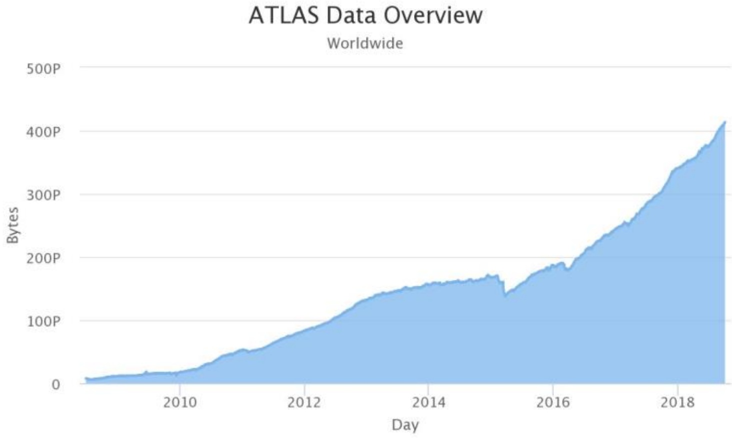
Rucio: a data manager

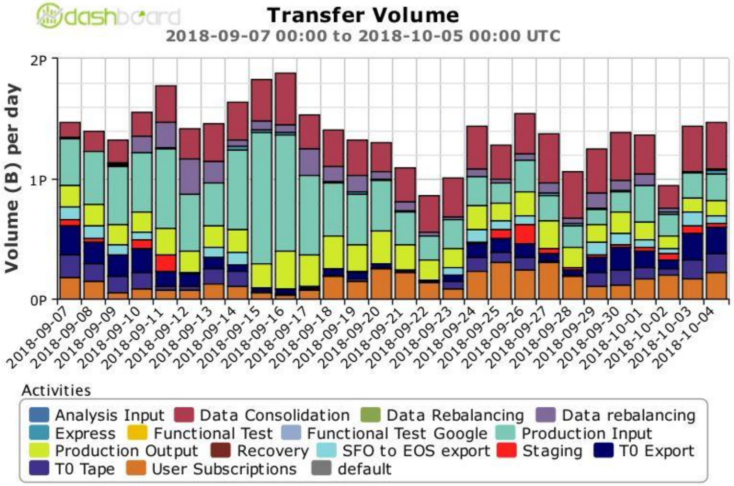
RUCIO
SCIENTIFIC DATA MANAGEMENT

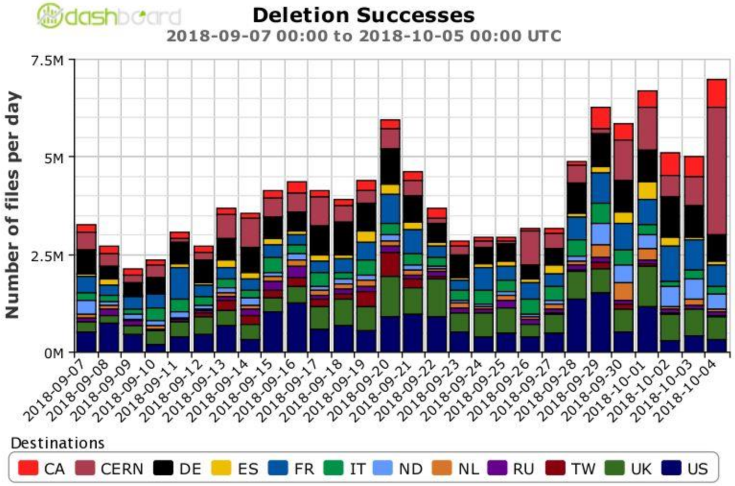
- A system to manage files stored in different places
  - Keeps track of copies – replicas
  - Can create replicas, move files around and delete them
    - All in a consistent manner
  - Handles access control, lifetime of data and other policies
- An Open Source software used by CERN, LIGO, LSST, CTA, AMS, DUNE and others

Rucio workload in ATLAS experiment

- Approx. 400 PB of data in more than a billion files
- Every day 2.5 PB is moved around
- More than 1000 users





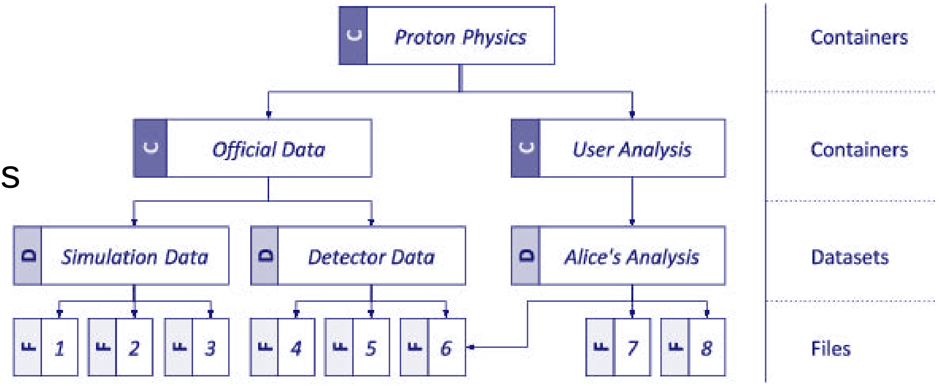


Adapted from Martin Barisits

Rucio Concept

- All data are assigned a Data Identifier – DID
- DIDs can belong to the following objects:
  - Single file
  - A dataset – a **collection** of files
  - A container – a collection of datasets
    - » not to be confused with Docker containers
    - » A collection of containers is also a container
- DIDs are meant to be globally unique
  - Are composed of a scope and a name:

```
user.martin:test.file.001
```



Metadata in Rucio

- Metadata are defined as attributes of DIDs
  - Custom attributes, but conventions can be enforced
- Various metadata are foreseen:
  - File system-specific: time stamp, size etc
  - Workflow-specific: what produced the object
  - Science-specific: device status, run number etc
  - Internal data management details, e.g. replication factor
- Metadata are used to organise data
  - Search by metadata
  - Aggregation of information by metadata

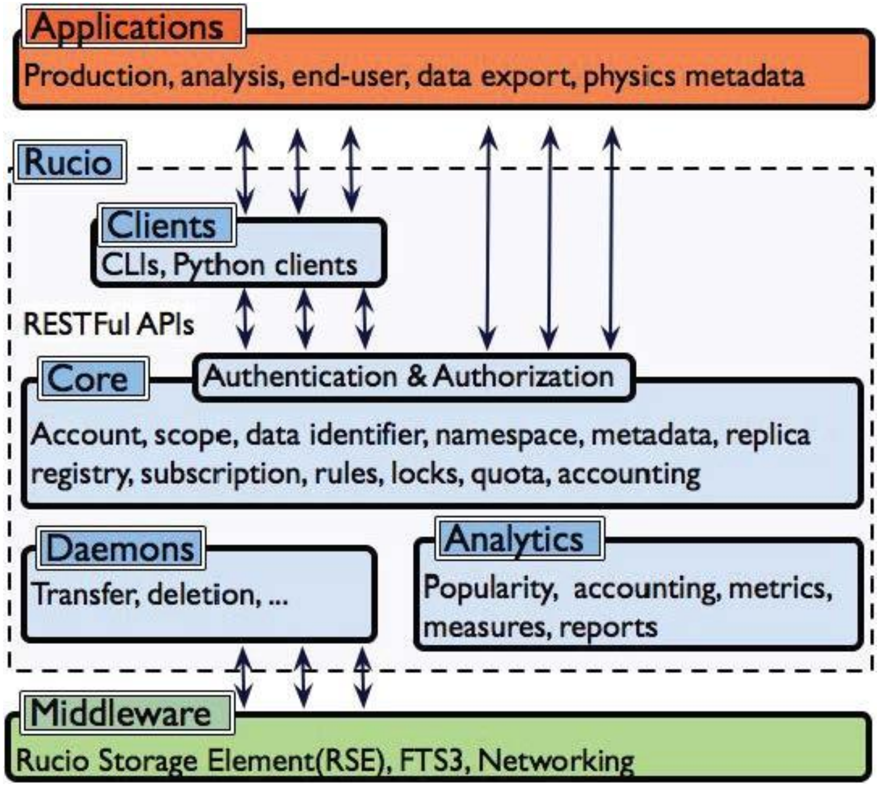
Storage handling in Rucio

- Each storage system is a logical entity – Rucio Storage Element (RSE)
  - No need to run RUCIO-specific software on your storage
  - Storage entities names are arbitrary labels
    - » Expects different labels for e.g. disk and tape storage and for different sites
- Each RSE has attributes recorded by Rucio (metadata): host name, protocols, ports, access priorities etc
  - Can be a country, a support address, storage type
  - One can list RSEs by a combination of attributes

Rucio rules

- Like some other data management systems, relies on declarative rules
  - *"Make two replicas on disk and one on tape, all in different countries"*
- Rules can be added and removed by users (some need authorisation)
  - If a dataset has a given lifetime, it can not be completely deleted before expiration
    - » Master copies have indefinite lifetime rules
- Copies can be created dynamically based on traced popularity data
  - Assumes sufficient storage space
- Users can create subscription rules
  - *"Send newly collected data my way"*

Rucio architecture



Applications
Production, analysis, end-user, data export, physics metadata
Rucio
Clients
CLIs, Python clients
RESTFul APIs
Core
Authentication & Authorization
Account, scope, data identifier, namespace, metadata, replica registry, subscription, rules, locks, quota, accounting
Daemons
Transfer, deletion, ...
Analytics
Popularity, accounting, metrics, measures, reports
Middleware
Rucio Storage Element(RSE), FTS3, Networking

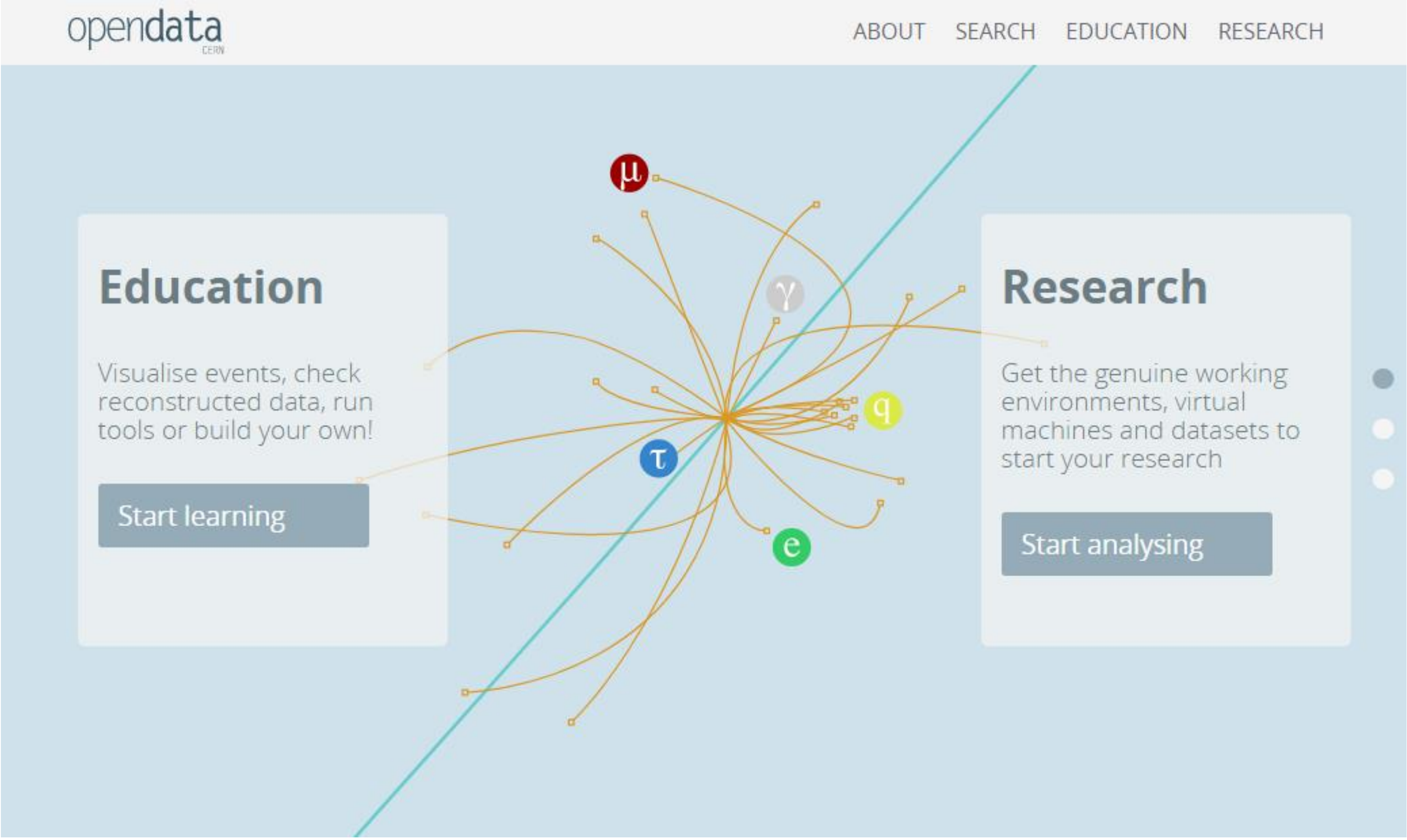
- Relies on HTTP and messaging
- Uses relational databases for book-keeping
  - Does not handle database replications itself
- Interfaces to all known storage systems and some data transfer tools
- Has user interfaces for operations and their monitoring
- Written in Python
  - Needs PIP, Docker and Kubernetes to deploy

Policies

Open Data

- Concept: data obtained using taxpayers’ money should be openly available to the taxpayers
  - Primarily to other scientists
- Issues:
  - How to protect data from free-riders?
  - How to make data usable by non-experts?
    - » A lot of data need specialist software to interpret
- Solutions:
  - Impose an embargo period
  - Convince people with permanent positions to do data curation for common good
    - » Some funding is also available
  - Use Jupyter notebooks

opendata.cern.ch



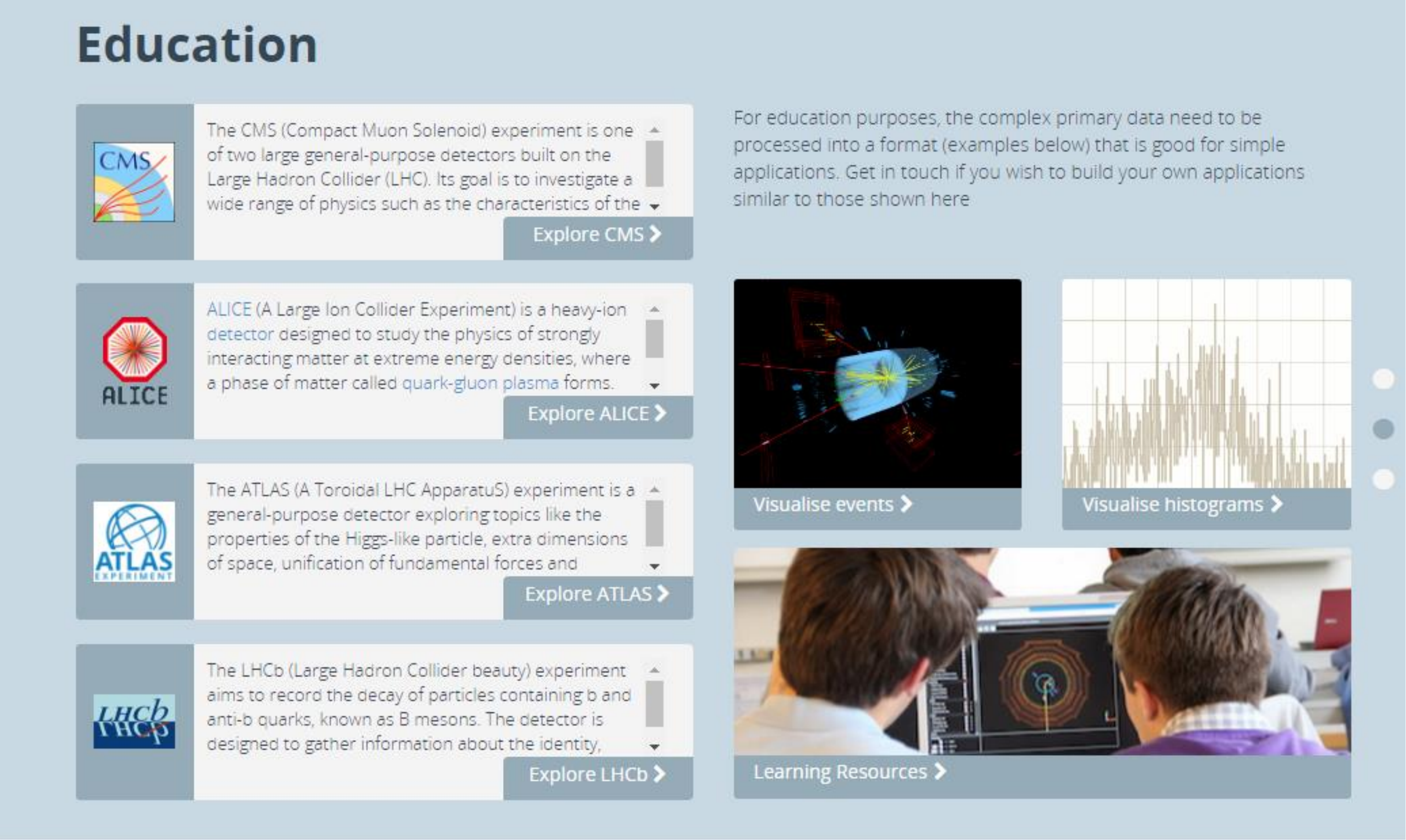
Education

Visualise events, check reconstructed data, run tools or build your own!

Research

Get the genuine working environments, virtual machines and datasets to start your research

opendata.cern.ch



Education

The CMS (Compact Muon Solenoid) experiment is one of two large general-purpose detectors built on the Large Hadron Collider (LHC). Its goal is to investigate a wide range of physics such as the characteristics of the

Explore CMS ›

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms.

Explore ALICE ›

The ATLAS (A Toroidal LHC ApparatuS) experiment is a general-purpose detector exploring topics like the properties of the Higgs-like particle, extra dimensions of space, unification of fundamental forces and

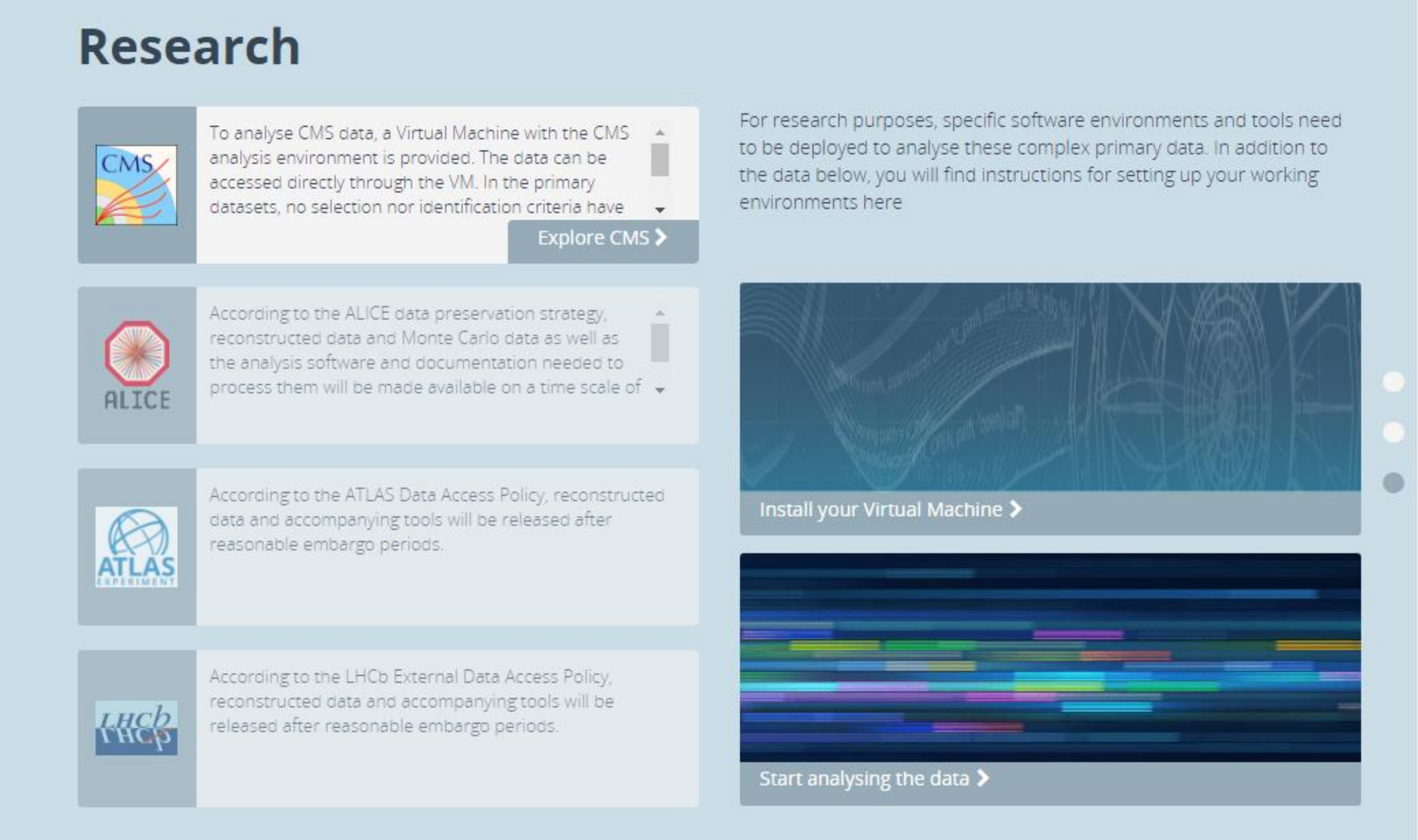
Explore ATLAS ›

The LHCb (Large Hadron Collider beauty) experiment aims to record the decay of particles containing b and anti-b quarks, known as B mesons. The detector is designed to gather information about the identity,

Explore LHCb ›

For education purposes, the complex primary data need to be processed into a format (examples below) that is good for simple applications. Get in touch if you wish to build your own applications similar to those shown here

opendata.cern.ch



Research

To analyse CMS data, a Virtual Machine with the CMS analysis environment is provided. The data can be accessed directly through the VM. In the primary datasets, no selection nor identification criteria have

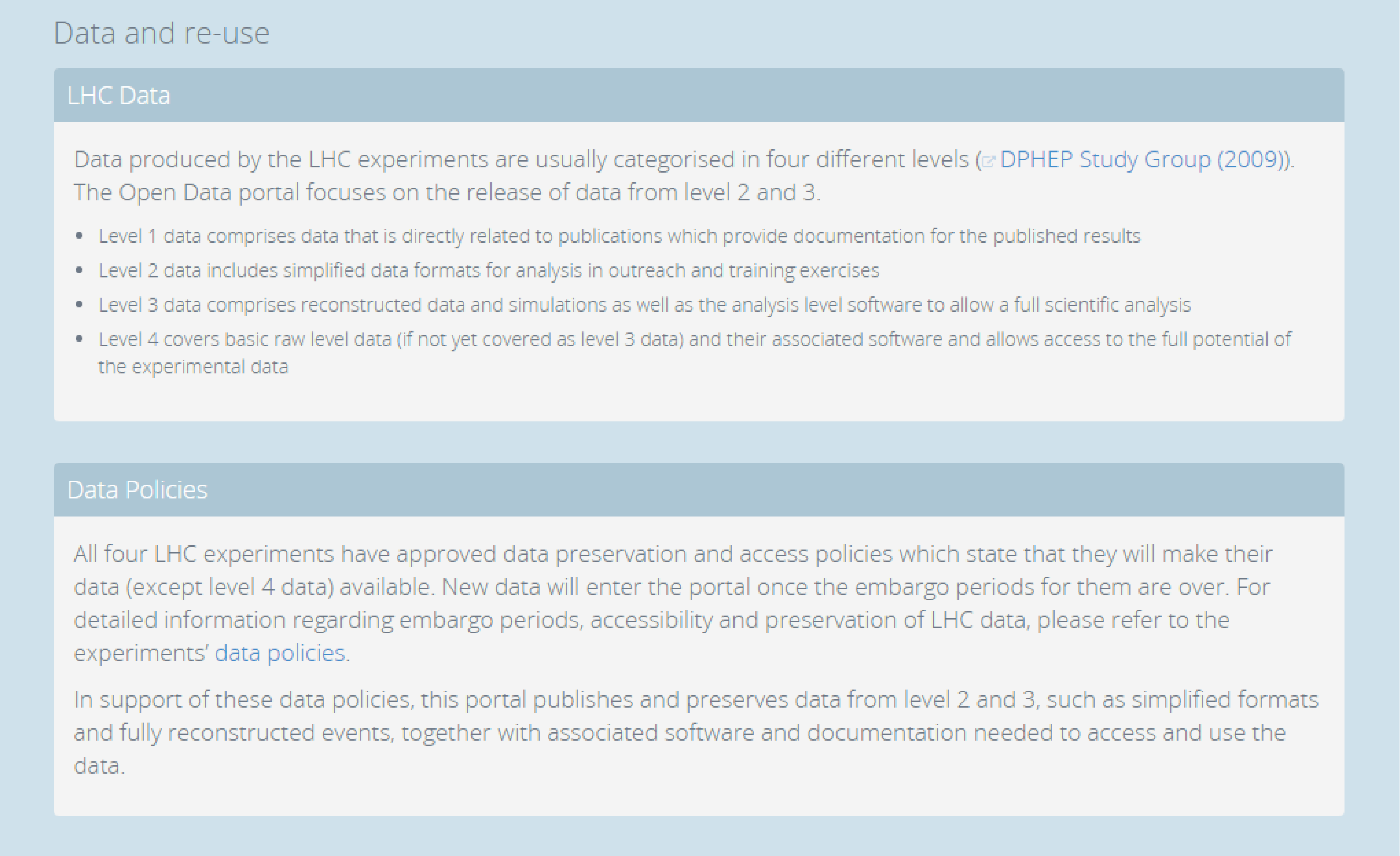
According to the ALICE data preservation strategy, reconstructed data and Monte Carlo data as well as the analysis software and documentation needed to process them will be made available on a time scale of

According to the ATLAS Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.

According to the LHCb External Data Access Policy, reconstructed data and accompanying tools will be released after reasonable embargo periods.

For research purposes, specific software environments and tools need to be deployed to analyse these complex primary data. In addition to the data below, you will find instructions for setting up your working environments here

opendata.cern.ch



Data and re-use

LHC Data

Data produced by the LHC experiments are usually categorised in four different levels (DPHEP Study Group (2009)). The Open Data portal focuses on the release of data from level 2 and 3.

- Level 1 data comprises data that is directly related to publications which provide documentation for the published results
- Level 2 data includes simplified data formats for analysis in outreach and training exercises
- Level 3 data comprises reconstructed data and simulations as well as the analysis level software to allow a full scientific analysis
- Level 4 covers basic raw level data (if not yet covered as level 3 data) and their associated software and allows access to the full potential of the experimental data

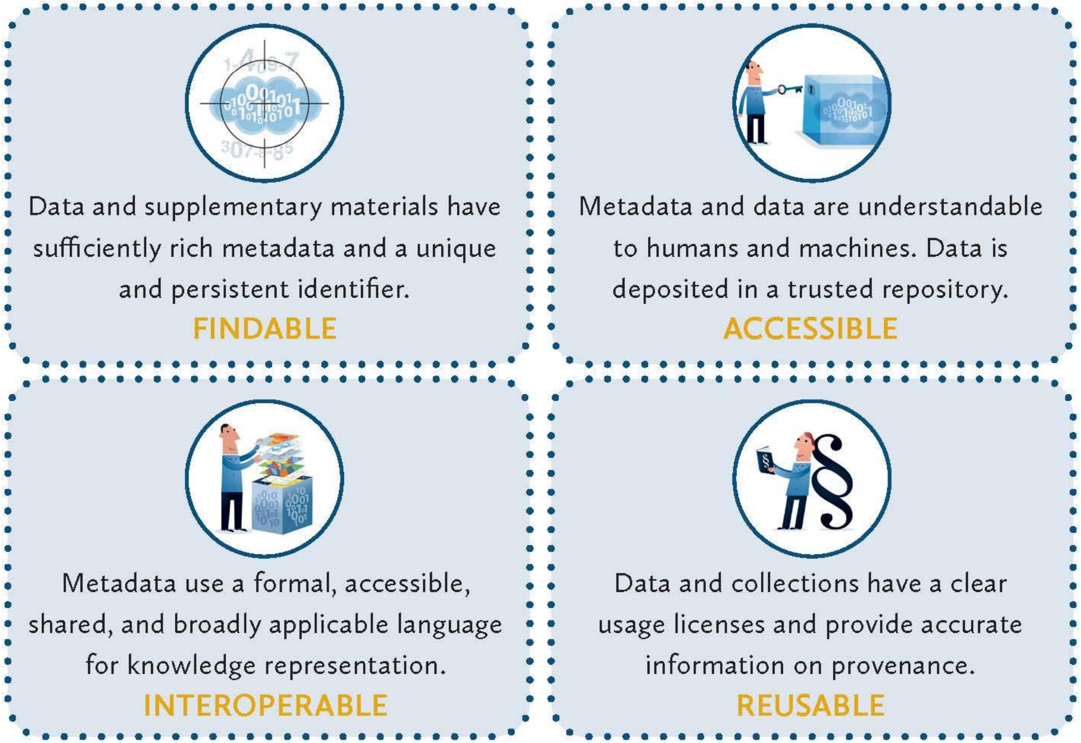
Data Policies

All four LHC experiments have approved data preservation and access policies which state that they will make their data (except level 4 data) available. New data will enter the portal once the embargo periods for them are over. For detailed information regarding embargo periods, accessibility and preservation of LHC data, please refer to the experiments' data policies.

In support of these data policies, this portal publishes and preserves data from level 2 and 3, such as simplified formats and fully reconstructed events, together with associated software and documentation needed to access and use the data.

FAIR Data Principles

- It is not always possible to have data fully open
- FAIR principles define key requirements to be met:



Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

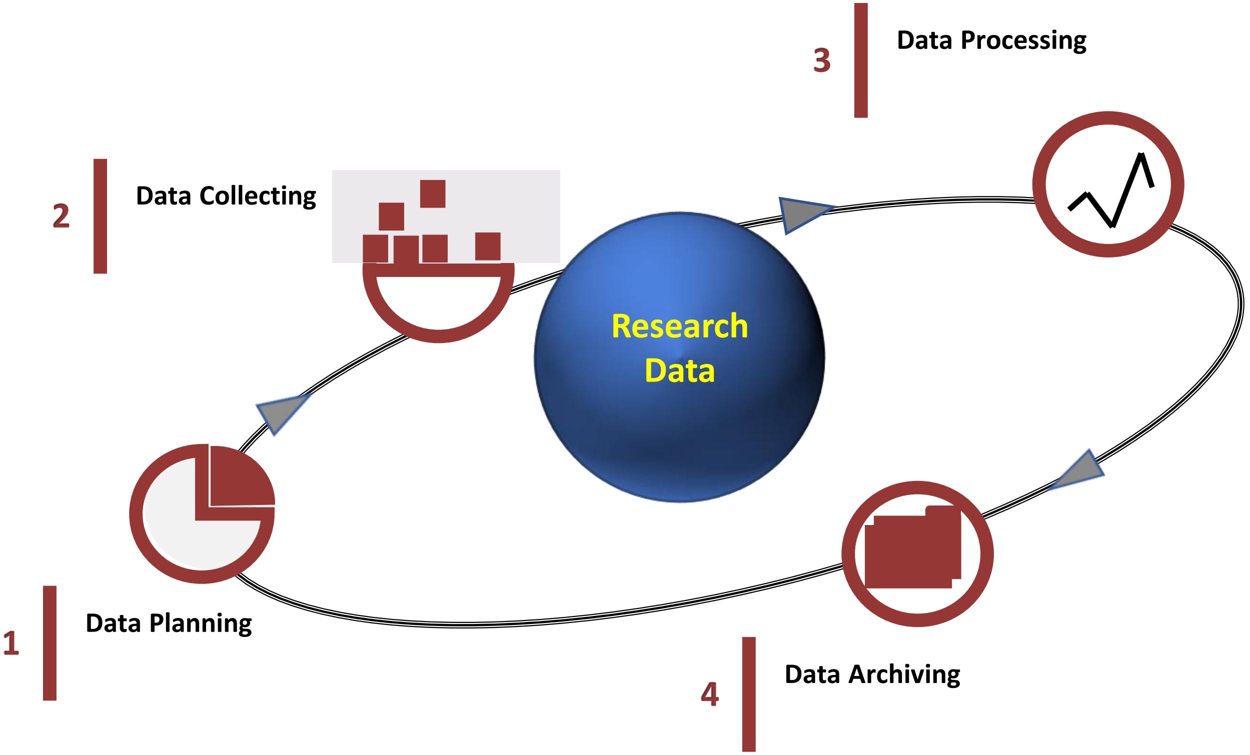
Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Graphics from digitalbevaring.dk

Research Data Life Cycle

- It starts with planning!

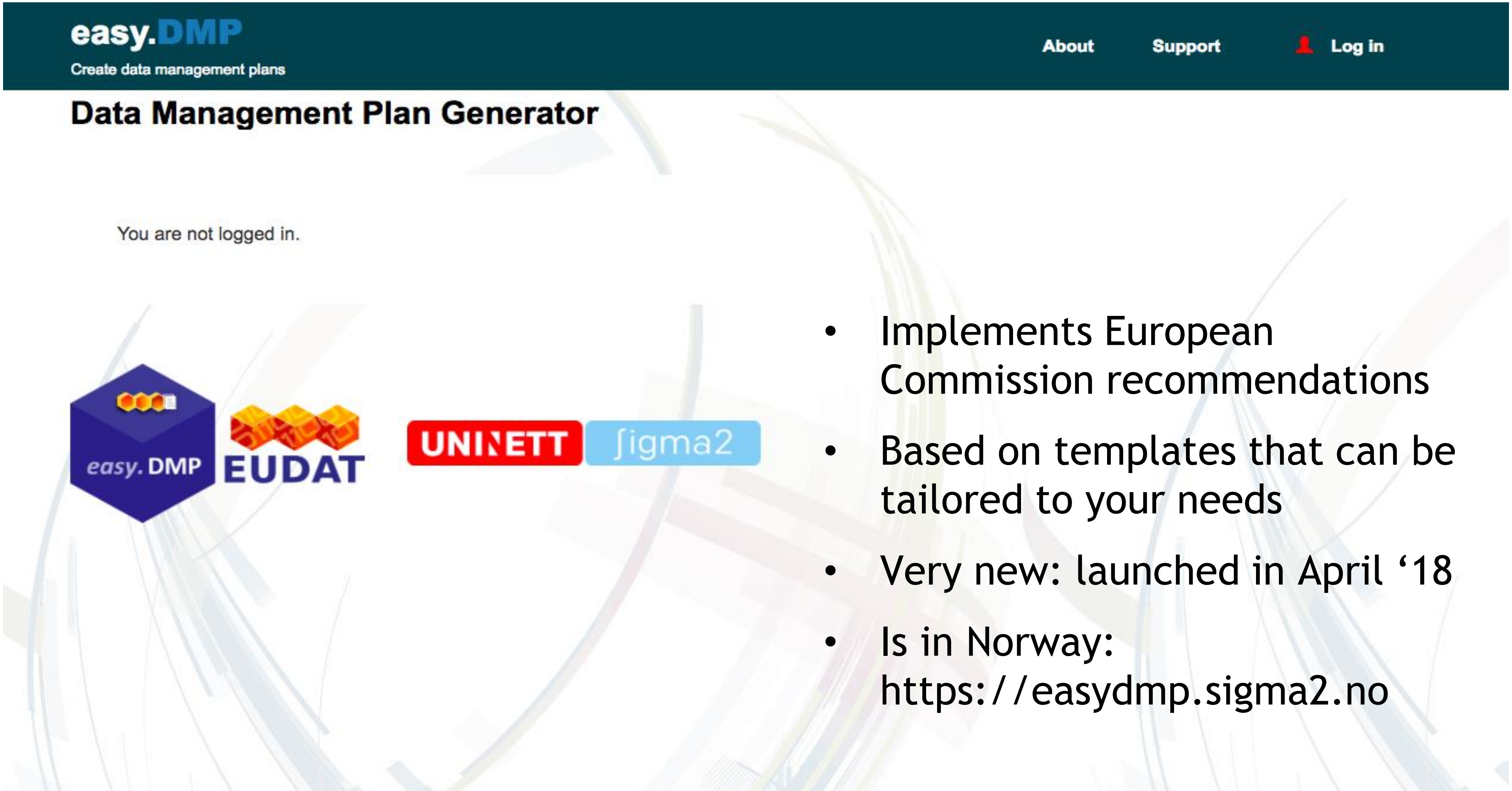


Adapted from Maria Francesca Iozzi

Data Management Plan – DMP

- Key points as recommended by Lund University:
  - **What** kind of data will you collect or generate?
  - **How** will the data be collected or generated?
  - Which documentation and what **metadata** will accompany the data?
  - Are there **ethical** issues that need consideration?
  - How will you manage **copyright** and Intellectual Property Rights (IPR) issues?
  - How will data be handled to ensure it is stored and transferred **securely**?
  - How will the data be **backed up** during the projects?
  - Which data should be retained, **shared**, and/or preserved?
  - What is the long term **preservation** plan?
  - How will you share the data? Are there any **restrictions** on data sharing?
  - Who will be **responsible** for data management?
  - Can the **resources** (people, hardware, software) required be specified and what costs are involved?
- See a 18-pages checklist at **Swedish National Data Service (SND)**: https://snd.gu.se

A handy new tool: easyDMP



Data Management Plan Generator

You are not logged in.

- Implements European Commission recommendations
- Based on templates that can be tailored to your needs
- Very new: launched in April ‘18
- Is in Norway: https://easydmp.sigma2.no

Significance of Open Data, FAIR and DMP

- Taxpayers via funding agencies require scientists to share data
  - European Commission, national research councils, but also some private funders
- If you pursue an academic career, you will have to seek academic grants
  - Increasingly, project proposals are required to include DMPs
  - Good DMPs are expected to ensure FAIRness and openness of data

Projects in Sweden: consult SND

SND
Swedish National Data Service

https://snd.gu.se

- Provides support to researchers in Sweden throughout the process of data management:
  - Guidance on how to write a data management plan
  - Advice on how to manage data
  - Information on legal aspects
- Offers some storage and persistent identifiers for stored data
  - Use your university account to log in
  - Homework: browse the SND Web site and see if you have data to upload