

# CDCI Online Analysis System

V. Savchenko  
for CDCI, ISDC

ASTERICS European Data Provider Forum and Training Event  
Heidelberg  
27-28/06/2018



**UNIVERSITÉ  
DE GENÈVE**

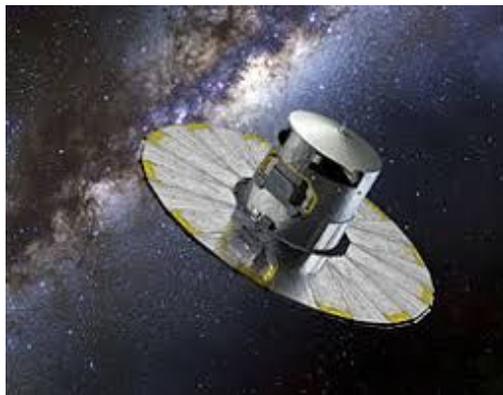


# Common Data Center Infrastructure (CDCI) in University of Geneva, Department of Astronomy

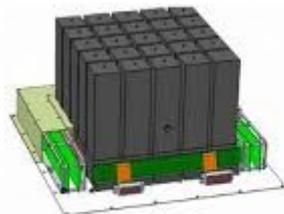
Among experiments supported by CDCI are



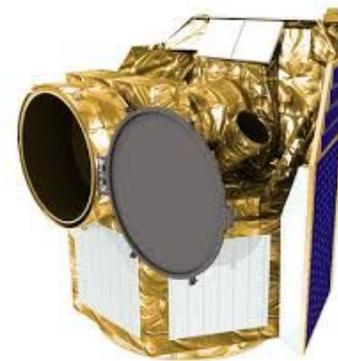
INTEGRAL



Gaia



POLAR



CHEOPS



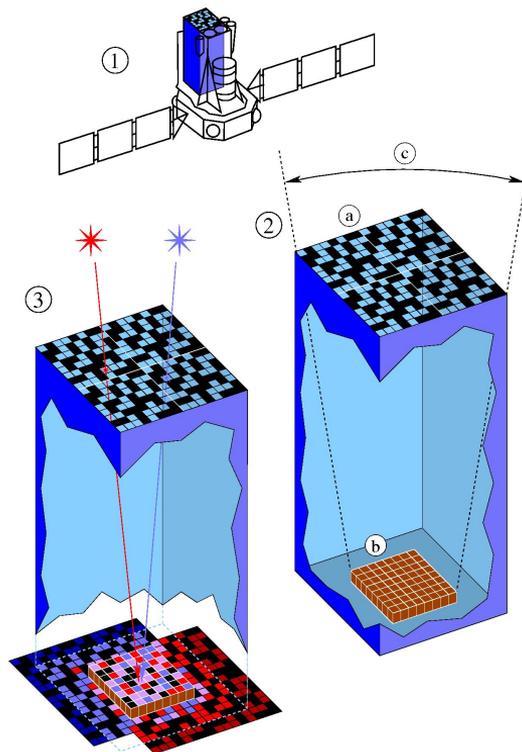
**UNIVERSITÉ  
DE GENÈVE**



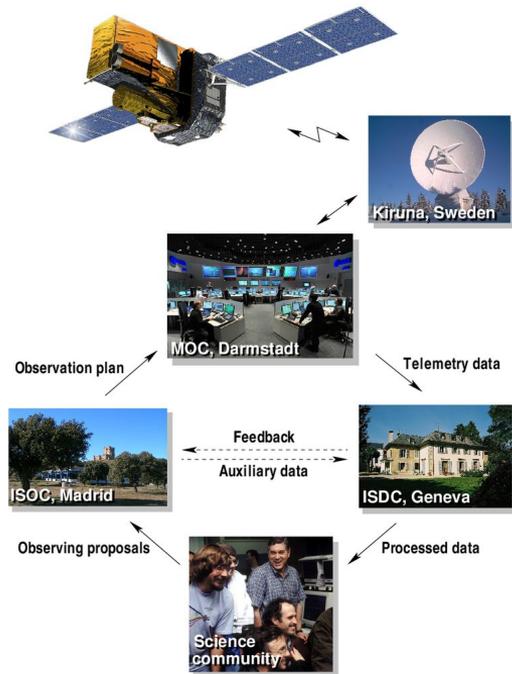
# Common Data Center Infrastructure (CDCI) and INTErnational Gamma-Ray Laboratory (INTEGRAL)

2002-2029

Sub-MeV Gamma-Ray Astronomy is hard: mirrors can not be used, trackers do not work, and the signal is encoded with mask projections. The **data analysis is a complex process of reconstructing source properties**. Scientific software is old and difficult to port.



# ISDC: INTEGRAL Data Center and INTEGRAL time domain astronomy driver

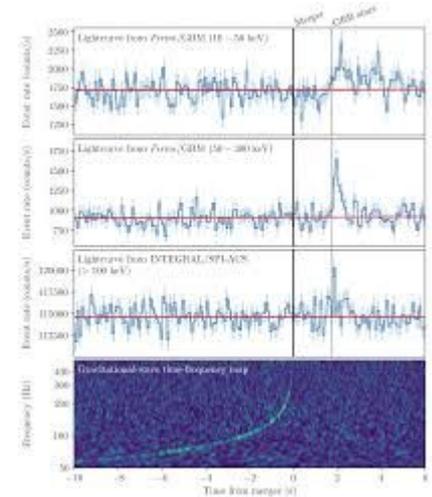


INTEGRAL Science Data Center  
(Versoix) is in charge of

- primary data processing
- data and software distribution
- quick-look analysis and
- prompt investigation of transient astronomical events (including GW, UH Neutrino, etc)

We receive public and private alerts,  
and distribute our own (GCN)

Large grasp yields good discovery  
potential: **need for efficient data  
exploration**



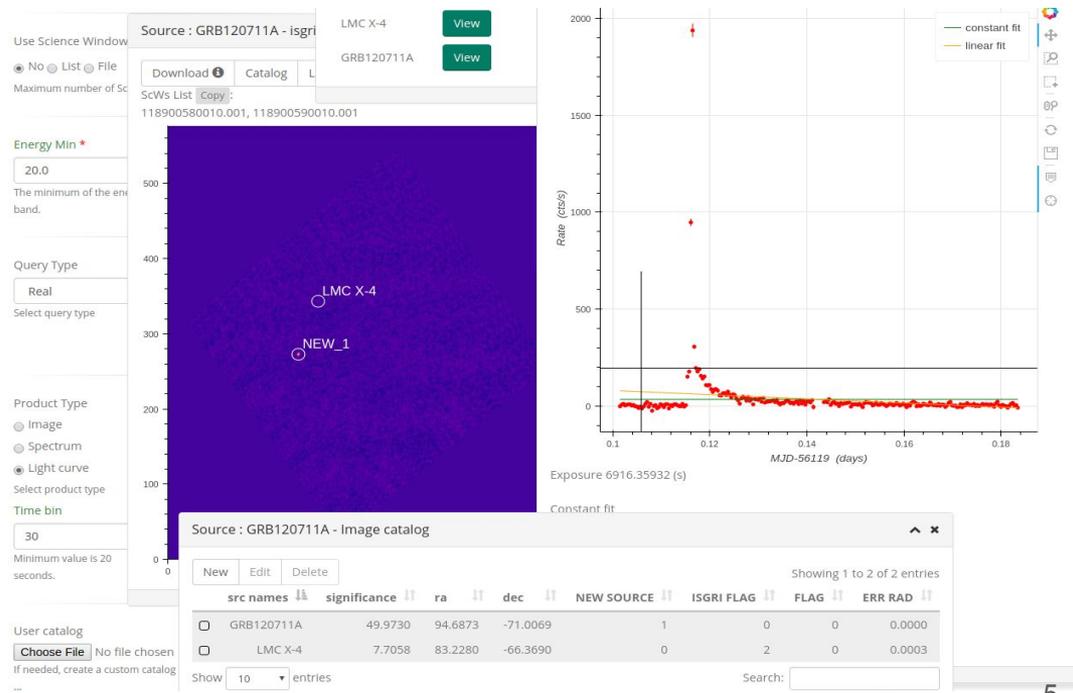
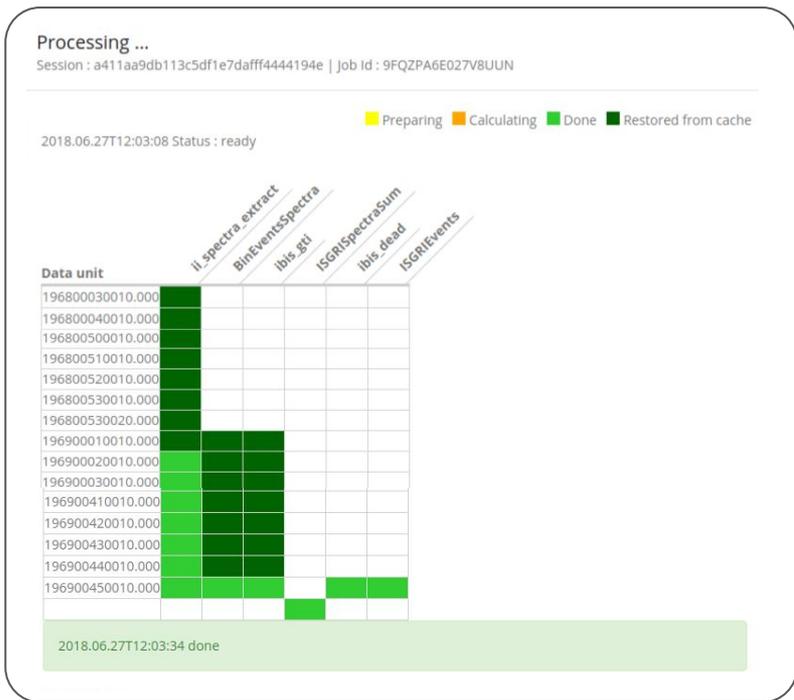
One of the transients  
detected at ISDC:

**GW170817/GRB170817A**

# Online Analysis System: frontend UI

Frontend for easy data presentation and exploration. Based on Drupal/AJAX

The results or their dependencies are reused when already available.

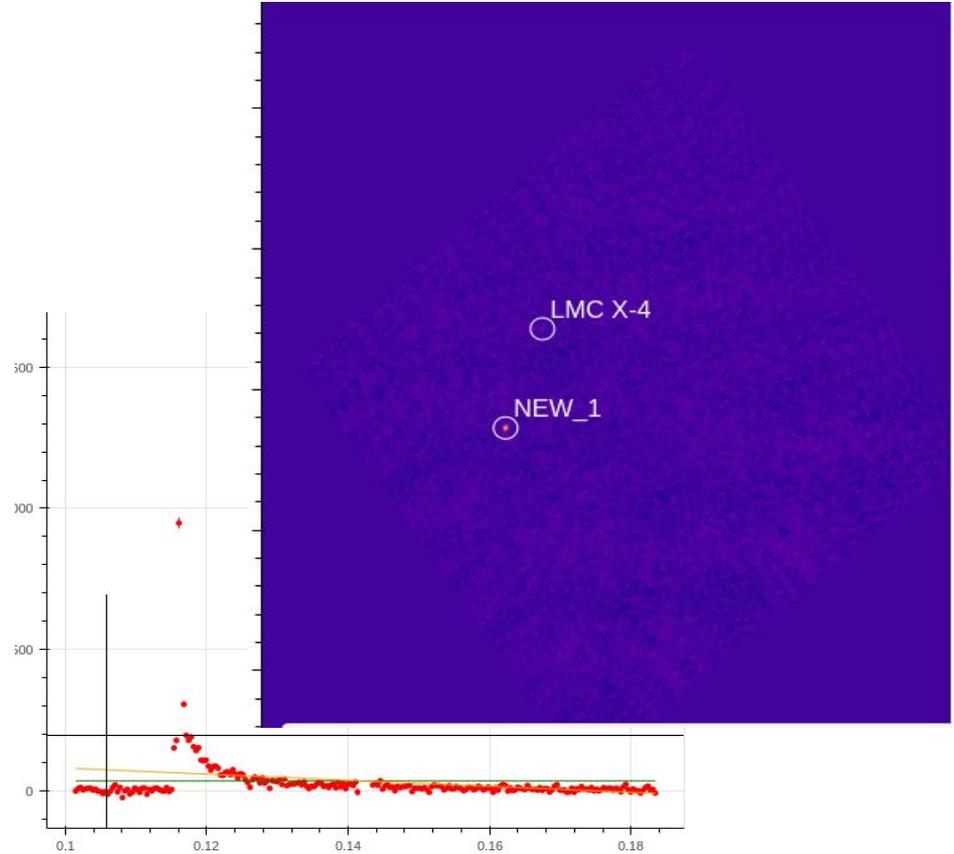


# Online Analysis System: product level

Provides **astronomical data products**:  
images, catalogs, spectra, light-curves

Can be queried through frontend, or  
directly with an **HTTP API**.

Reformulates the requests for the  
astronomical products received from the  
frontend to workflow requests to the  
backend.



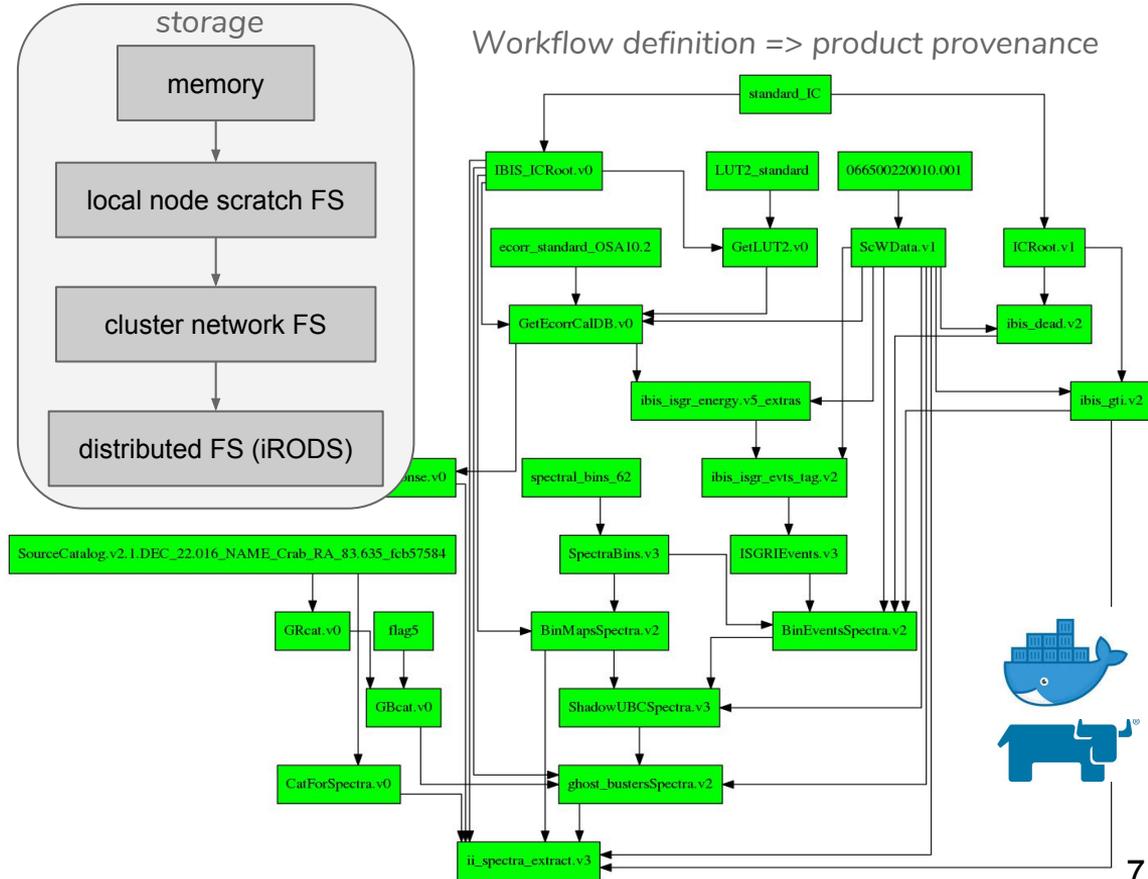
# Online Analysis System: backend

Declarative data analysis **definition** is separated from scheduling and storage.

The pipeline is composed of analysis nodes with no side effects. Pipeline execution consists in cascading resolution of node dependencies.

Dependency DAG is used for **distributed scheduling**.

Analysis definition openly stored on github/gitlab.



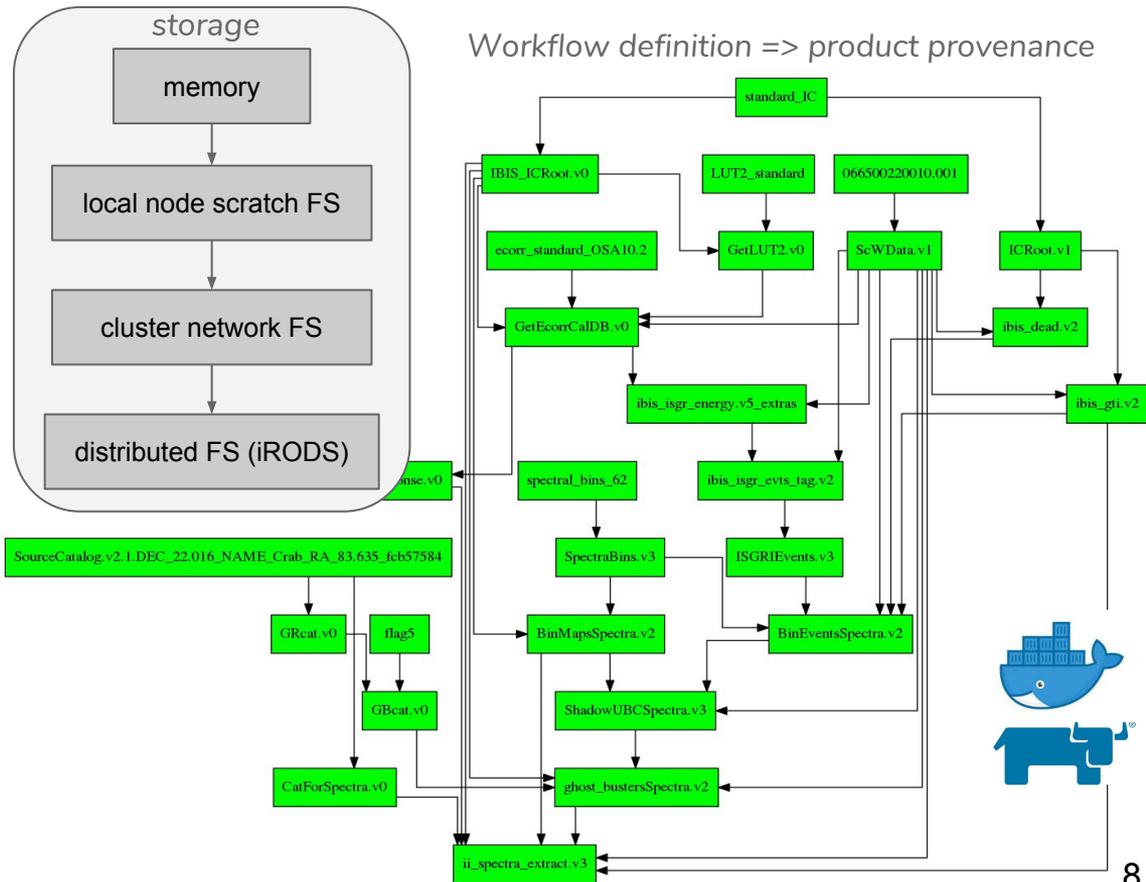
# Online Analysis System: backend storage

Storage is a hierarchical immutable cache of the pipeline results, **indexed with data provenance metadata** expressed as **directed acyclic graphs**.

Products are fairly heterogeneous and feature complex ontology

Can be queried with an **API** to execute **any compliant user-defined workflow**

The pipeline engine and analysis definition is open-source, typically stored on github, and **can be also executed offline (no black-box services)**

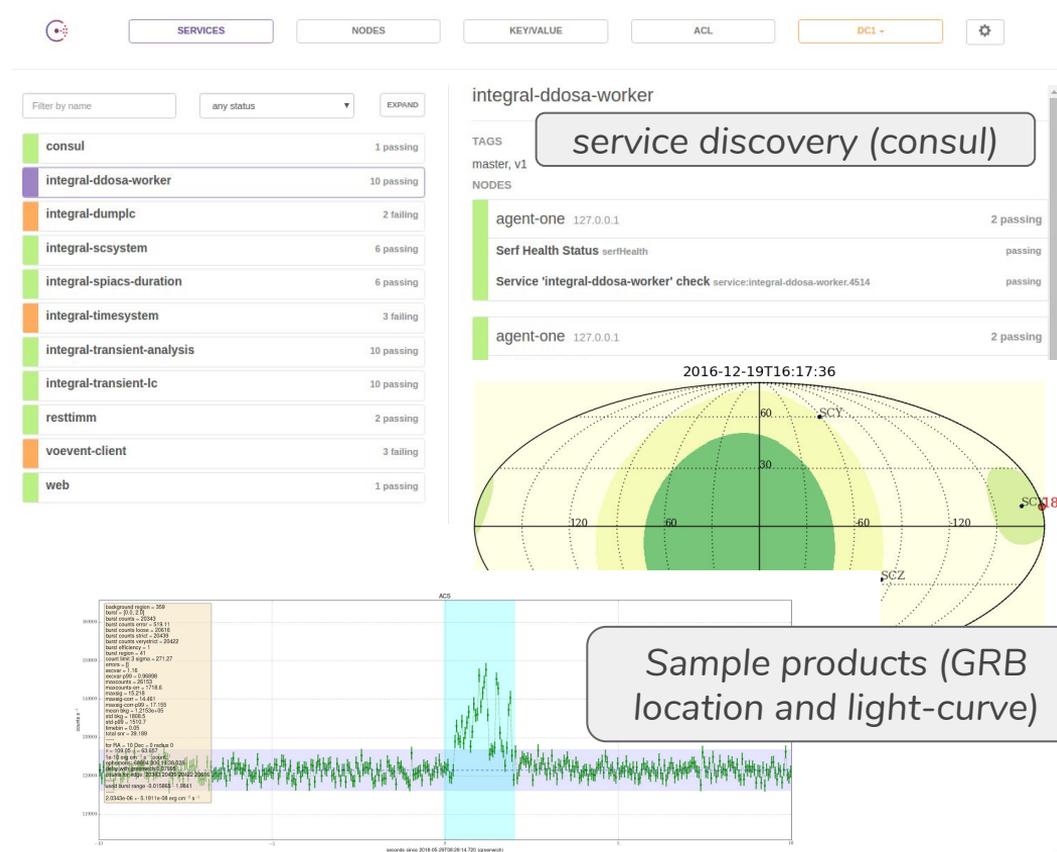


# Microservices for real-time analytics

Time-critical real-time scientific analysis is largely performed with a **distributed network of microservices** optimally performing primary data reduction where the data lives.

We publicly share direct access to a limit set of specific microservices for easy interoperability. API providing INTEGRAL data are routinely used by different teams in follow-up of multimessenger transients.

Will become progressively more public



# Exploring workflows and provenance

We collaborate with a multidisciplinary project at EPFL (Renku/SDSC) which helps to data scientists **collaboratively explore data provenance and analysis options**.

We also coordinate with CERN Analysis Preservation efforts: **REANA** (Reusable analysis platform), Zenodo.



## First steps

First steps with Renku

Updated 16 hours ago.

Overview Kus Files Settings

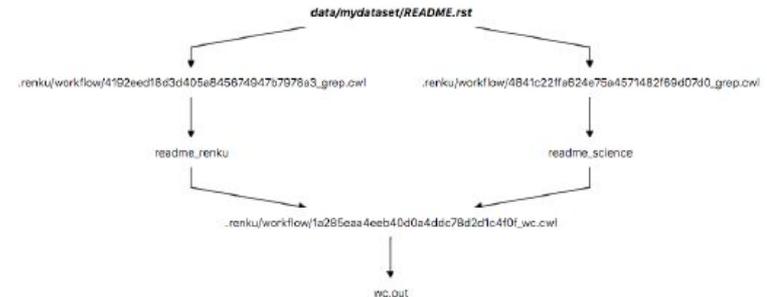
Notebooks

Data

Workflows

Other

### data/mydataset/README.rst — lineage and use



<https://datascience.ch/renku-platform>



CERN  
Analysis  
Preservation

REANA

<https://github.com/reanahub/reana> 10

# Prospects

- OAS is expected to be released publicly soon (before autumn 2018)
- We plan to include more astronomical experiments, of UniGe Department of Astronomy and open data repositories.
- Adopt workflow definition standards (CWL)
- Adopt W3C PROV-O
- UI will assist in assigning DOI to the products
- Provide VO-compliant interfaces

