ESA Datalabs

Multi Mission Science Exploitation and Preservation Platform

→ ADASS 2019

V. Navarro, R. Alvarez, S. del Rio, R. Guerra, V. Martín, M.A. Diego, N. Alexandre, A. Lousa, X. Rodrigues, A. Pereira, D. Monteiro, A. Héliou, M. Marseille

Groningen, the Netherlands

08/10/2019

→ COSMIC OBSERVERS

esa

CONCEPTS


spica


theseus

IN DEVELOPMENT


webb (2021)


ariel (2028)


euclid (2022)


cheops (2019)


plato (2026)


xrism (2021)


einstein probe (2022)


athena (2031)


lisa (2034)

OPERATIONAL


hubble (1990–)


gaia (2013–)


xmm-newton (1999–)


integral (2002–)

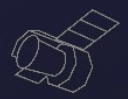
microwaves | sub-millimetre | infrared | optical | ultraviolet | x-rays | gamma rays | gravitational waves

LEGACY


planck (2009–2013)


herschel (2009–2013)


iso (1995–1998)


akari (2006–2011)


hipparcos (1989–1993)


corot (2006–2014)


iue (1978–1996)


exosat (1983–1986)


hitomi (2016)


suzaku (2005–2015)


cos-b (1975–1982)

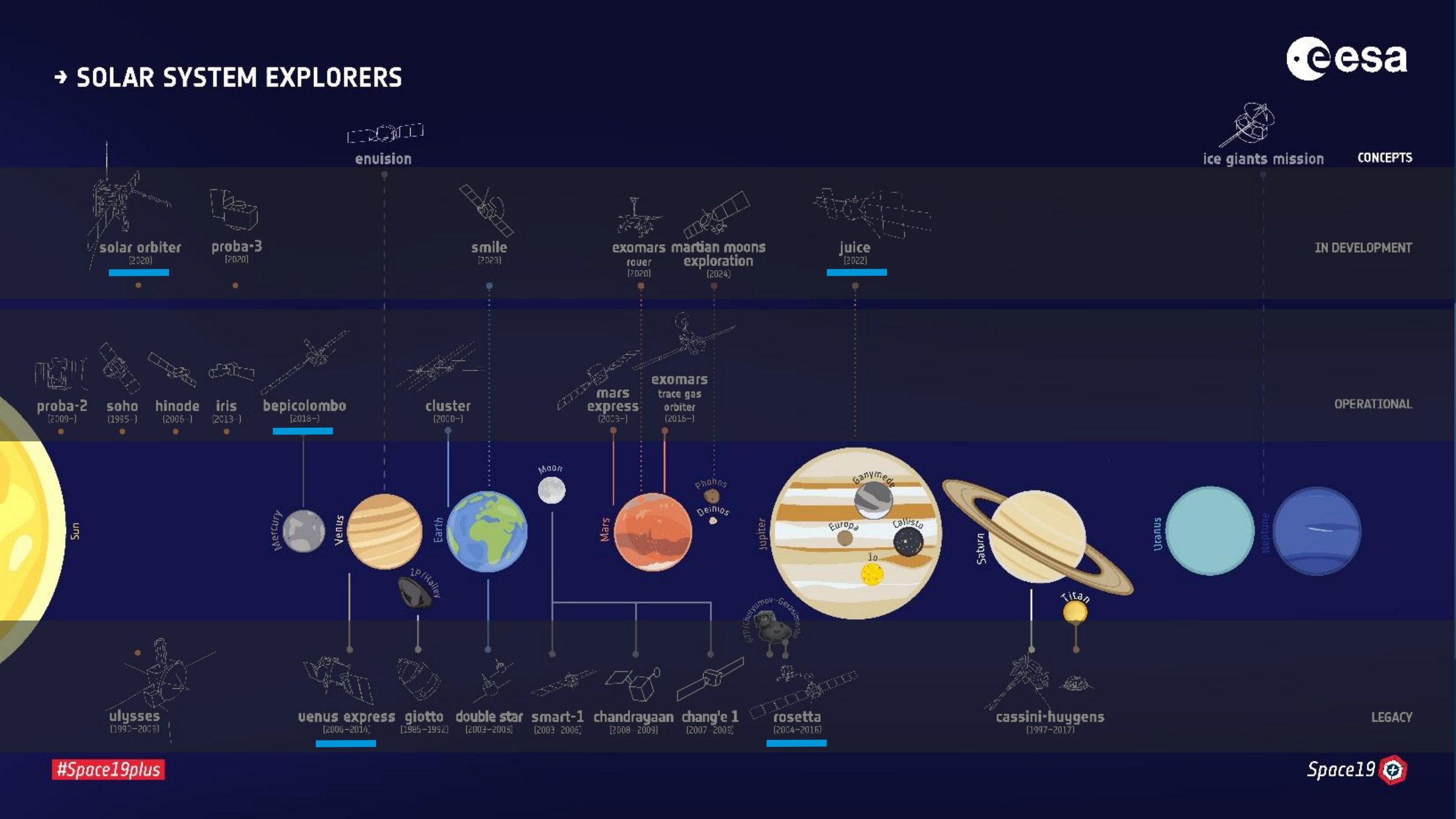

lisa pathfinder (2015–2017)


microscope (2016–2018)

#Space19plus

Space19

→ SOLAR SYSTEM EXPLORERS



#Space19plus

Space19

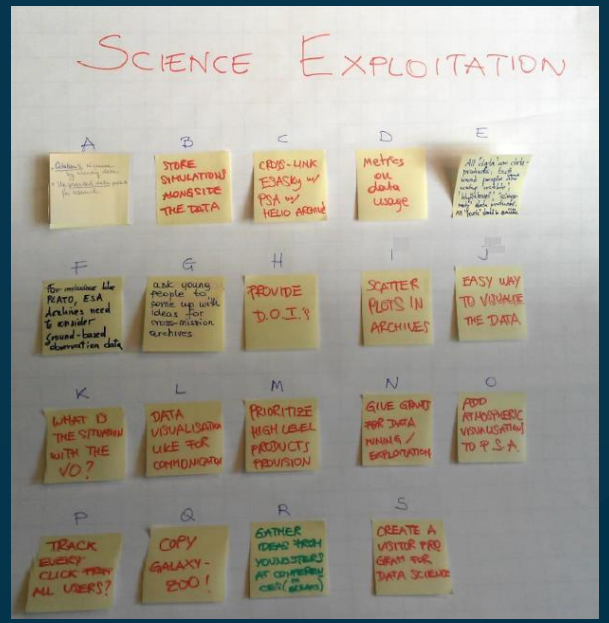
Strategy Definition

SEPP-013: ESDC Collaborative Research Lab

Primary actors: astronomers, planetary scientists and heliophysicists with knowledge of one or more of the data products in the ESDC science archives and wanting to data mine, visualise and/or analyse data from one or a combination of missions.

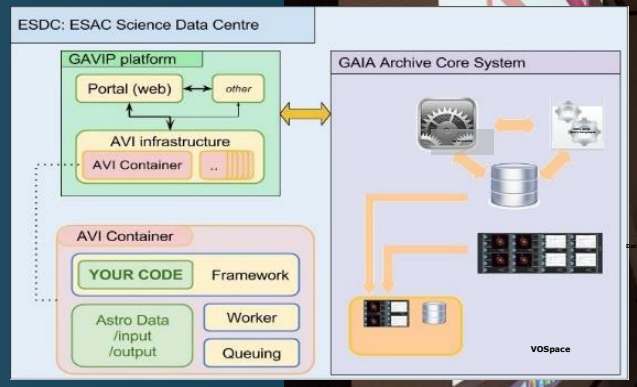
Secondary actors: members of the general public, educators, students, etc. wanting to explore and analyse archival data in the ESDC Science archives.

Efficient long-term preservation of data, software and knowledge

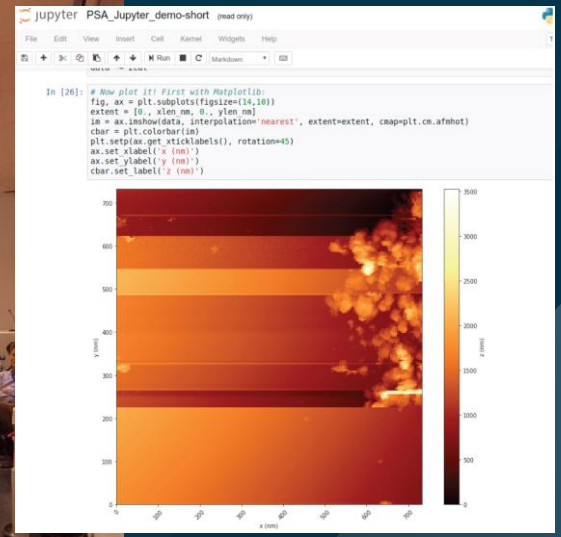


Enabling maximum scientific exploitation of datasets

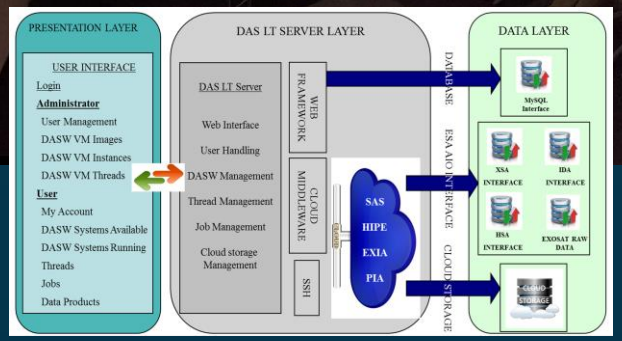
SCI Related Activities



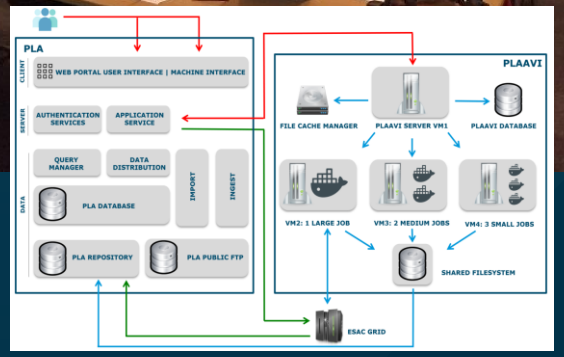
Gaia AVI



SCI-O JupyterHub



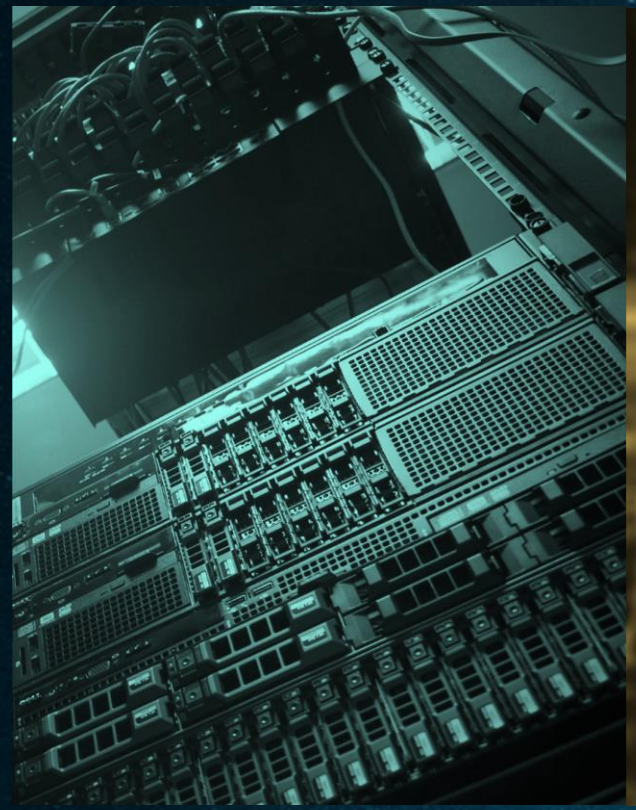
DAS-LT



Planck AVI

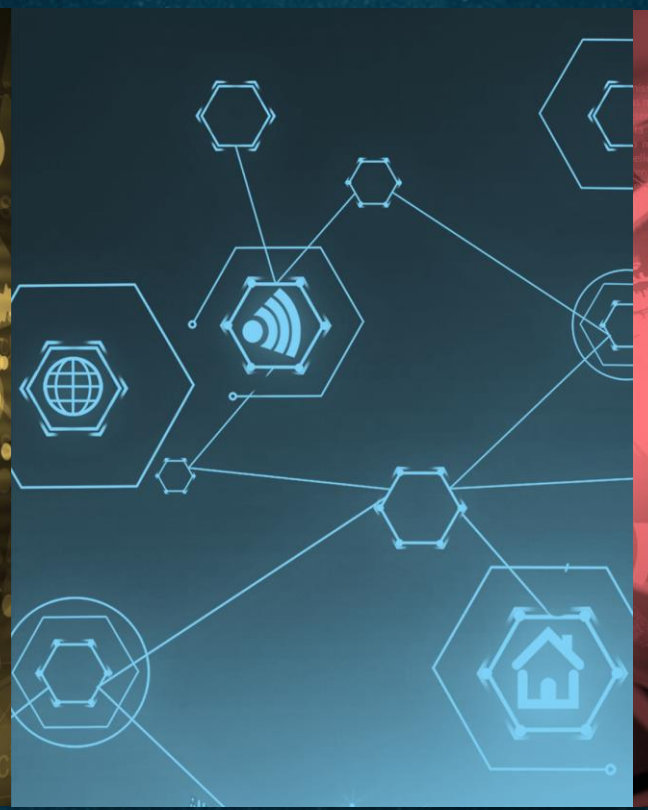
Key Goals

Preservation



Exploitation

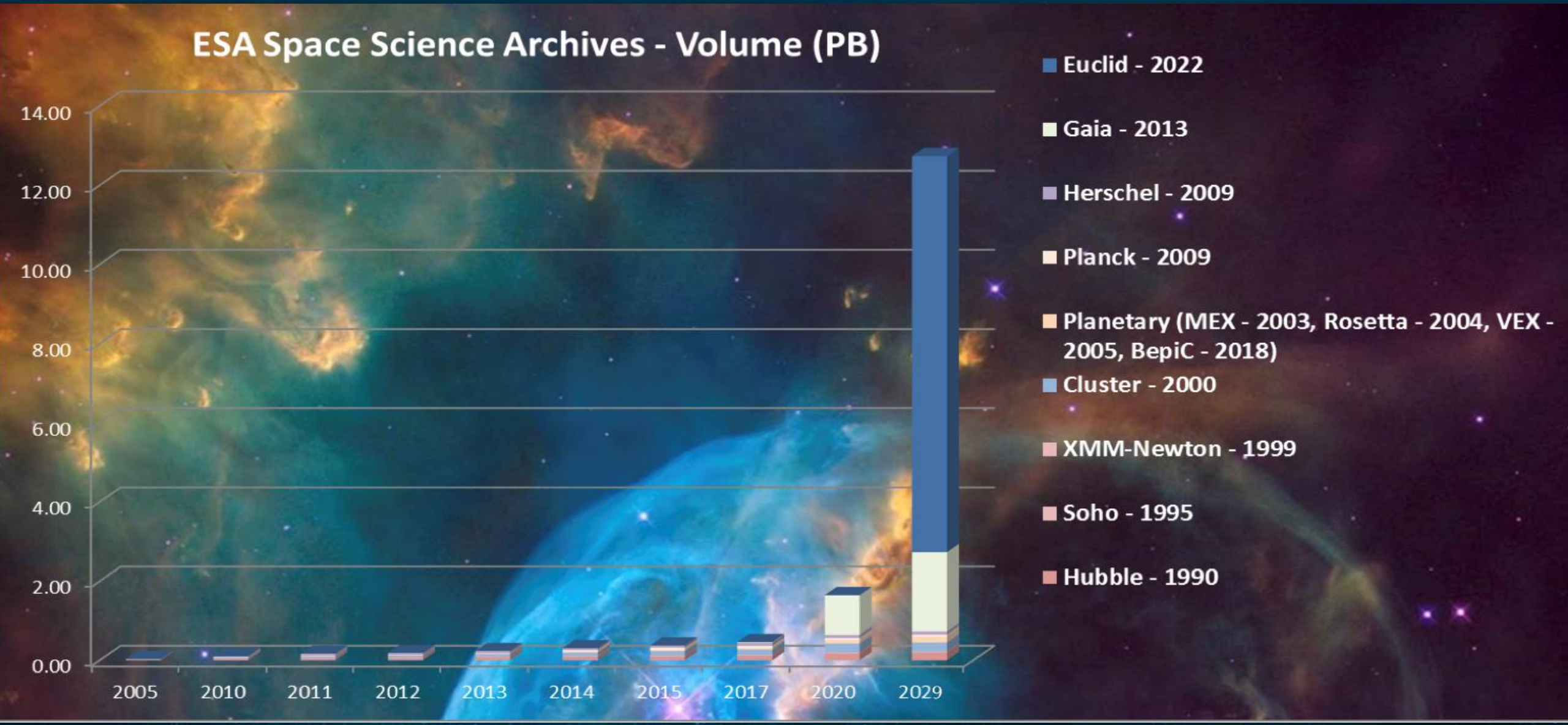
Pipeline Management



Collaboration

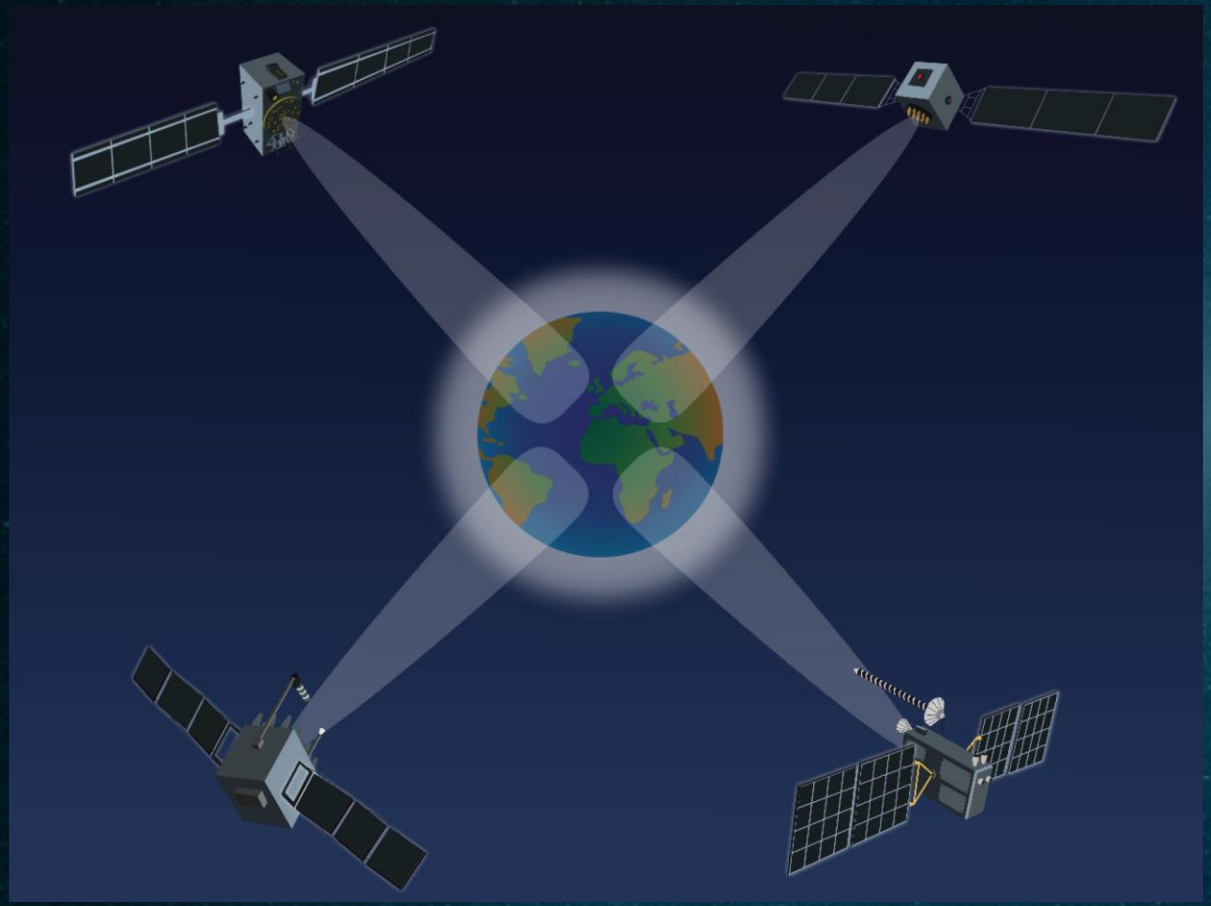

Big Data

The ESA Space Science Case



ESA Navigation Science Case

IF Data ▶ High Potential for Discoveries and Innovations



→ Need to manage
8.3 TB of data
per day at each station

→ **1 PB of data**
per day for 120 stations

Paradigm Shift

"Moue compute to the data, instead of the data to the compute

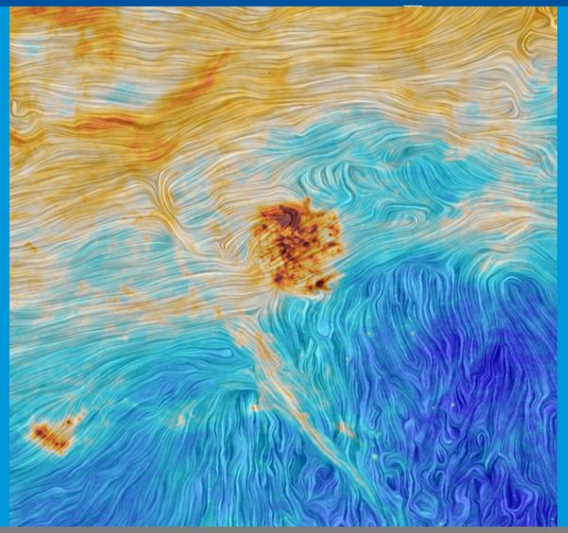
... keeping the focus on end-users

How does it look like?

esa | datalabs

«YOU CAN EITHER MOVE YOUR QUESTIONS OR THE DATA. [...] OFTEN IT TURNS OUT TO BE MORE EFFICIENT TO MOVE THE QUESTIONS THAN TO MOVE THE DATA.»

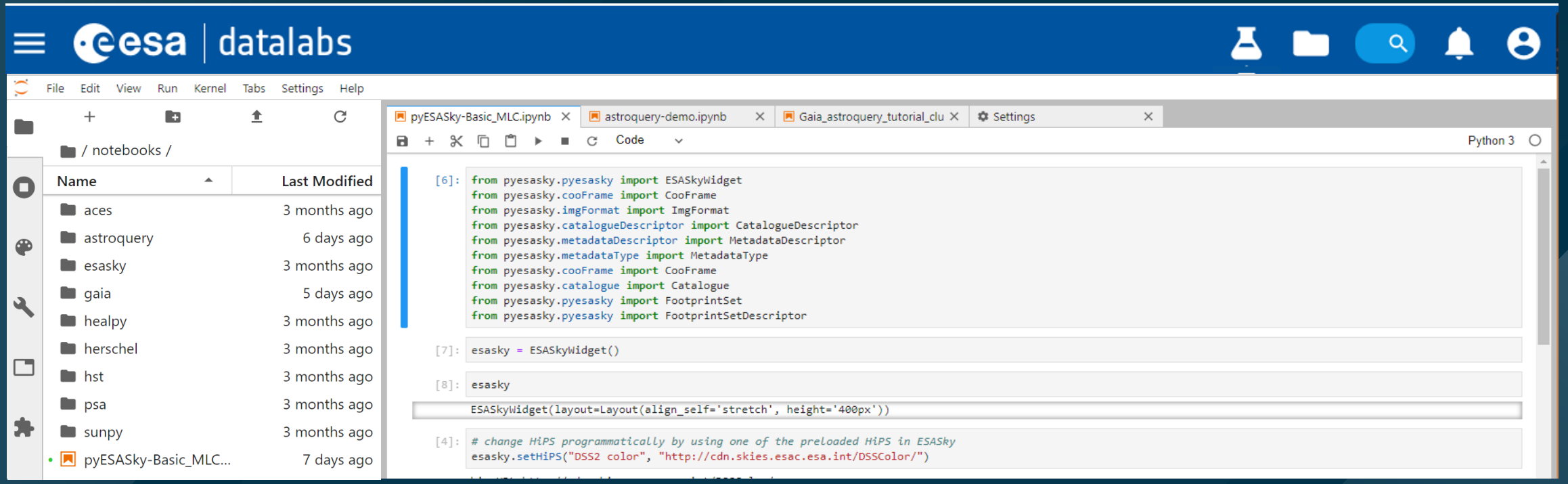
Jim Gray, eScience: A Transformed Scientific Method



→ THE ARCHIVES, A SCIENTIFIC TREASURE TROVE

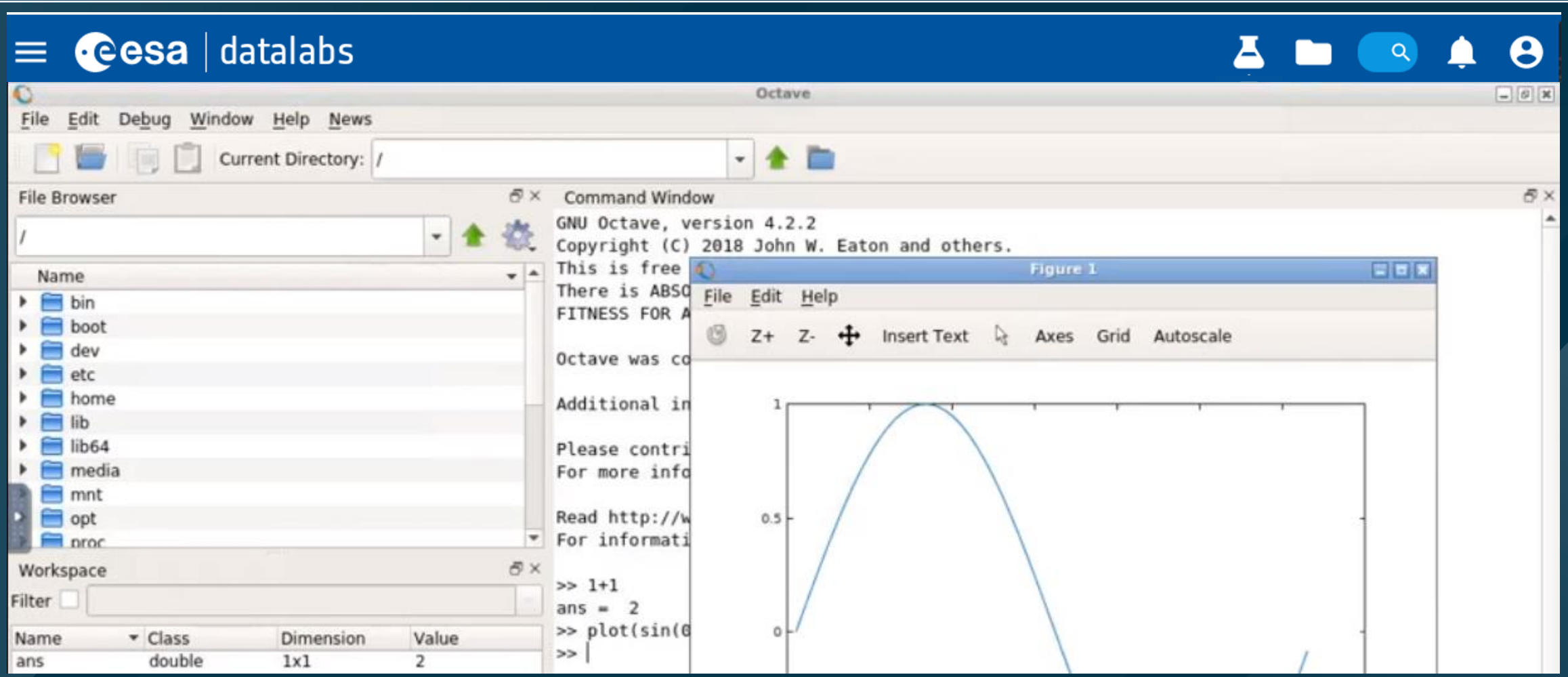
The vast amounts of scientific data obtained during a space science mission have a much longer lifetime than the satellite mission itself. The data are archived and made freely accessible online to the global scientific community, and these archives are frequently a mine of unexpected discoveries. They allow researchers to study, for instance, the evolution of a certain celestial object with time, or its appearance at different wavelengths as observed by different telescopes.

Analysis Services: JupyterLab



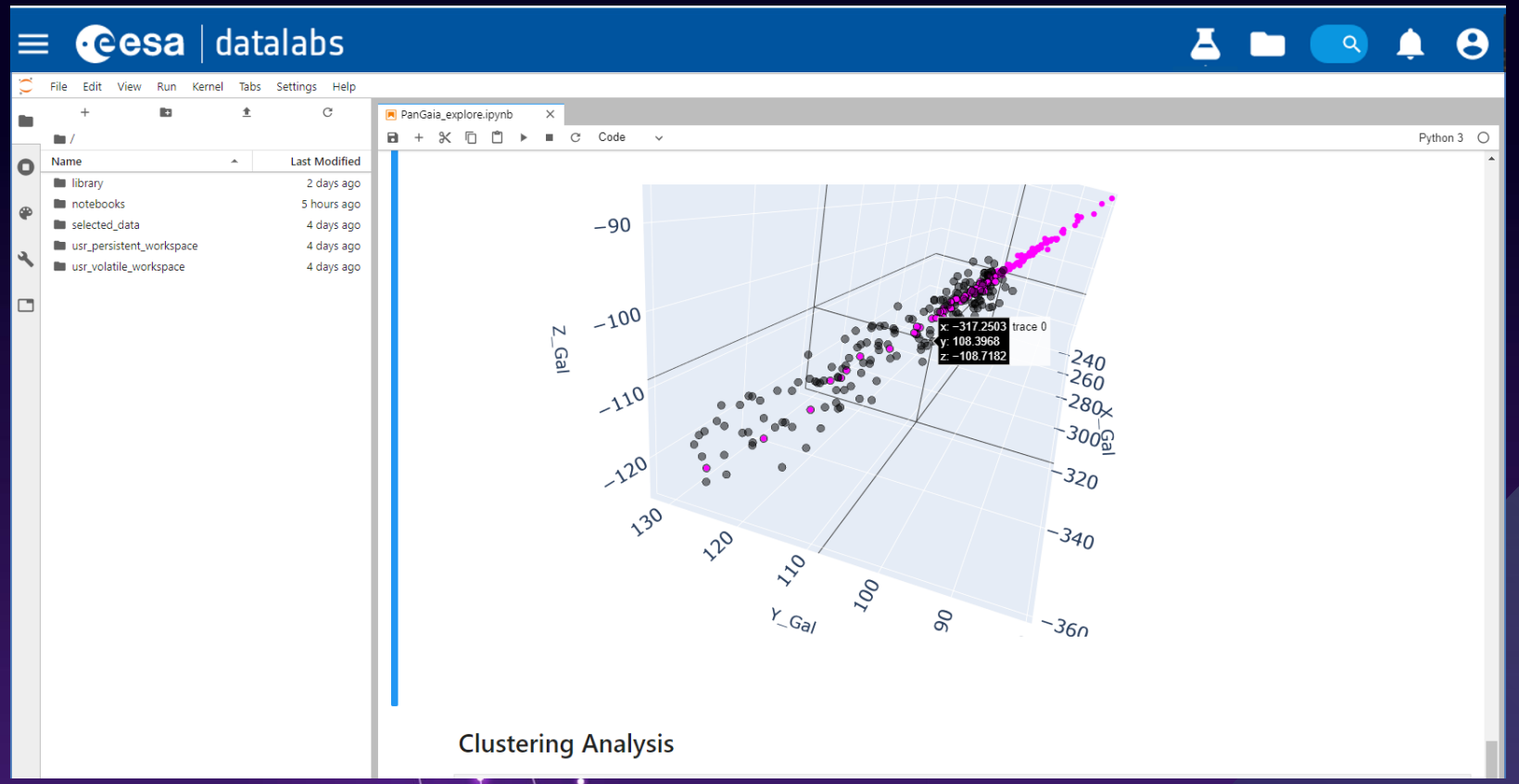
Customised, ready-to-use environment to maximise focus on scientific work

Desktop Services: Octave



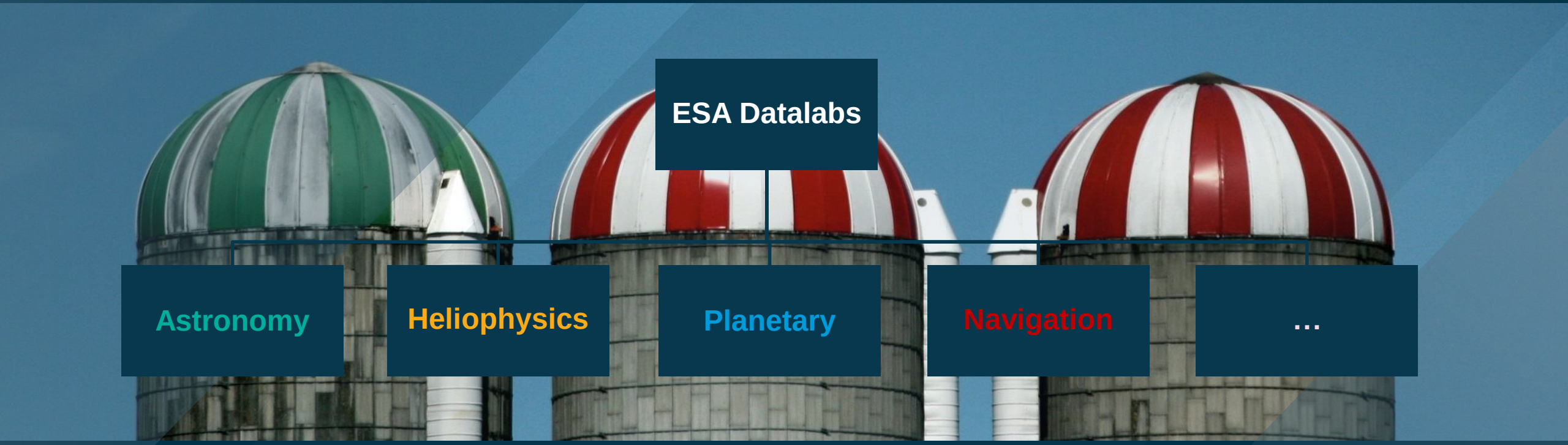
Full access via web browser to desktop based applications

HDBSCAN Clustering on Gaia DR2

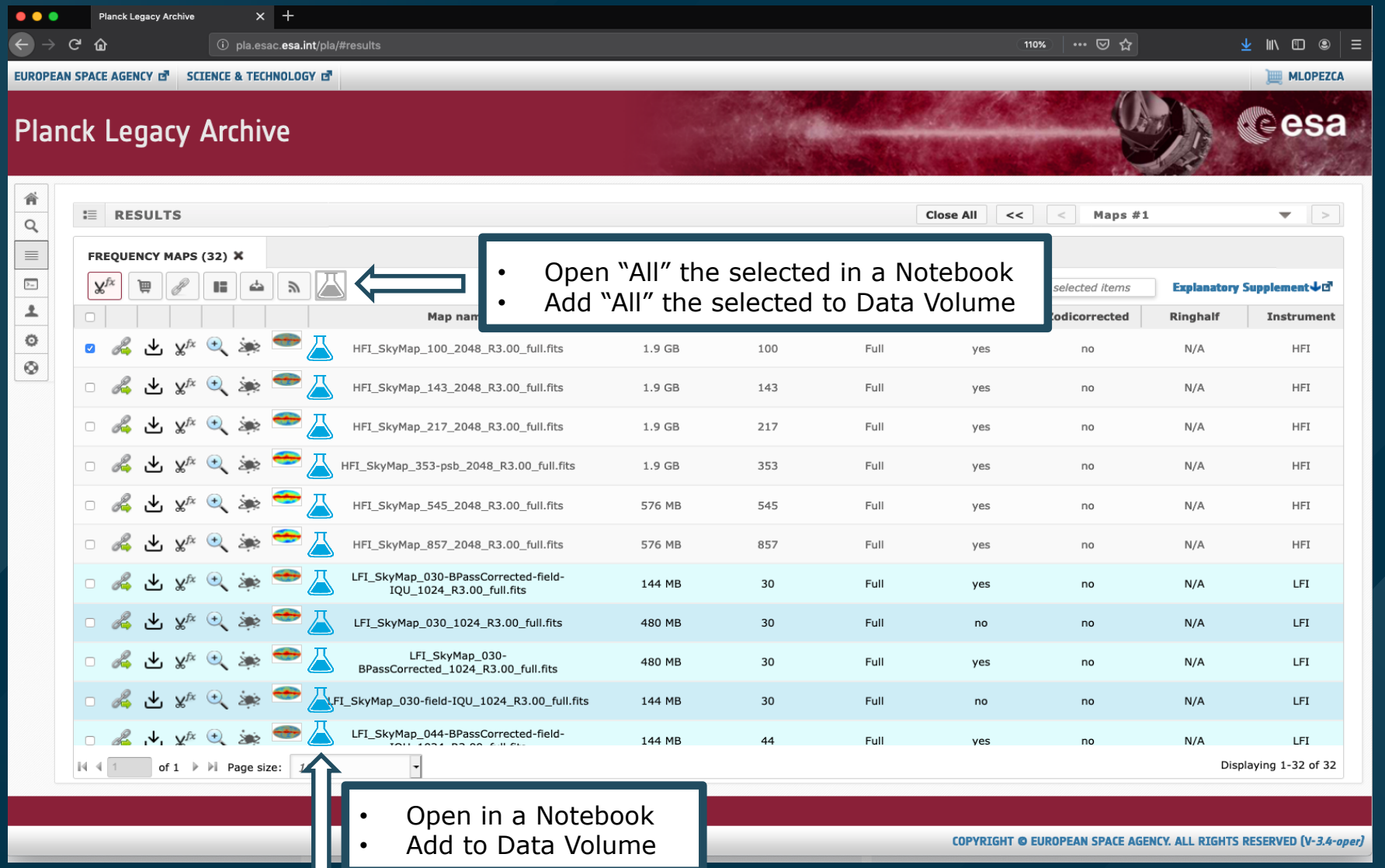


Canovas et al. 2019b in preparison
P 3.3 Merin, Canovas et al.

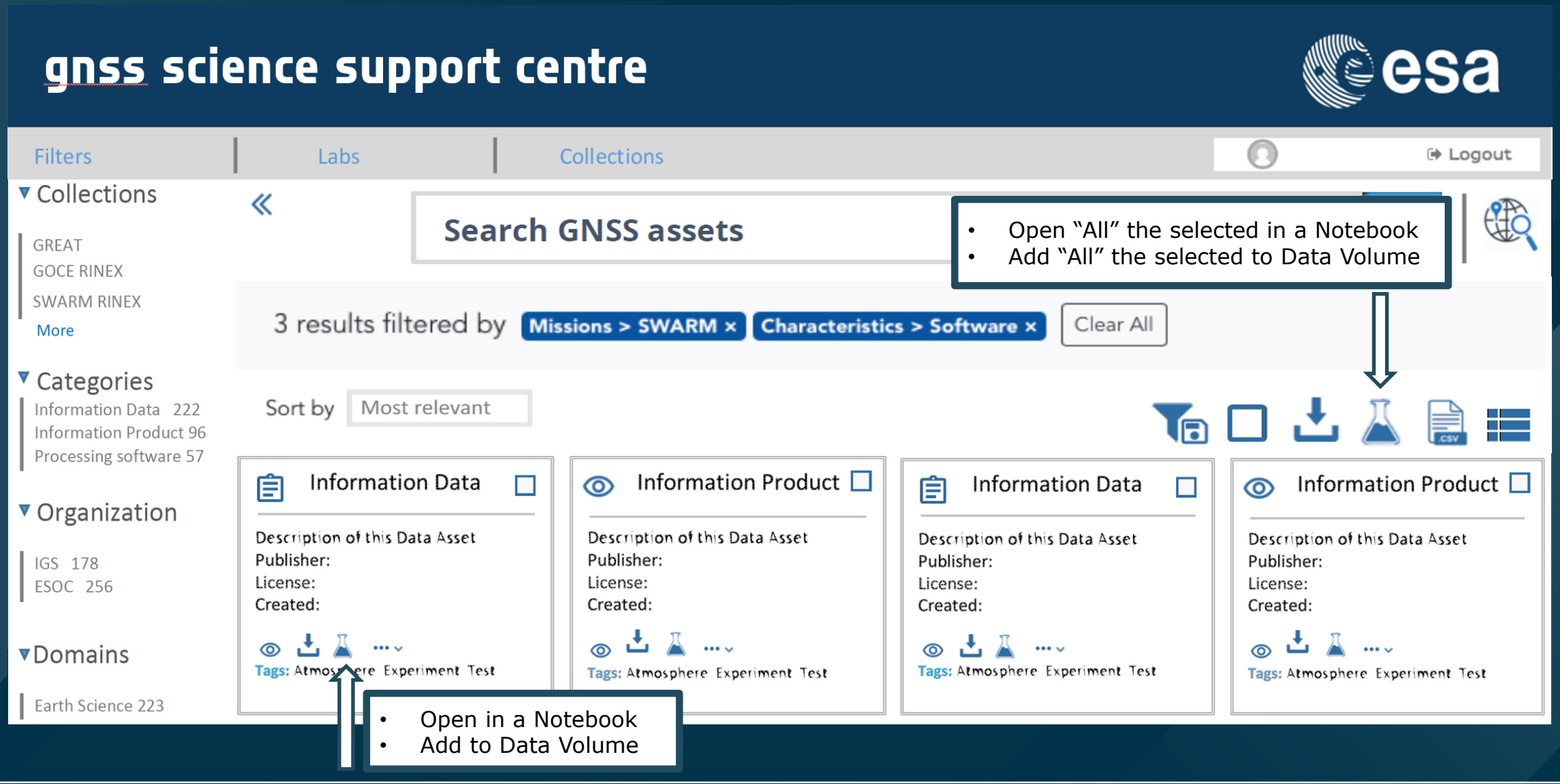
Multi Mission Platform



Science Archives Integration: Planck

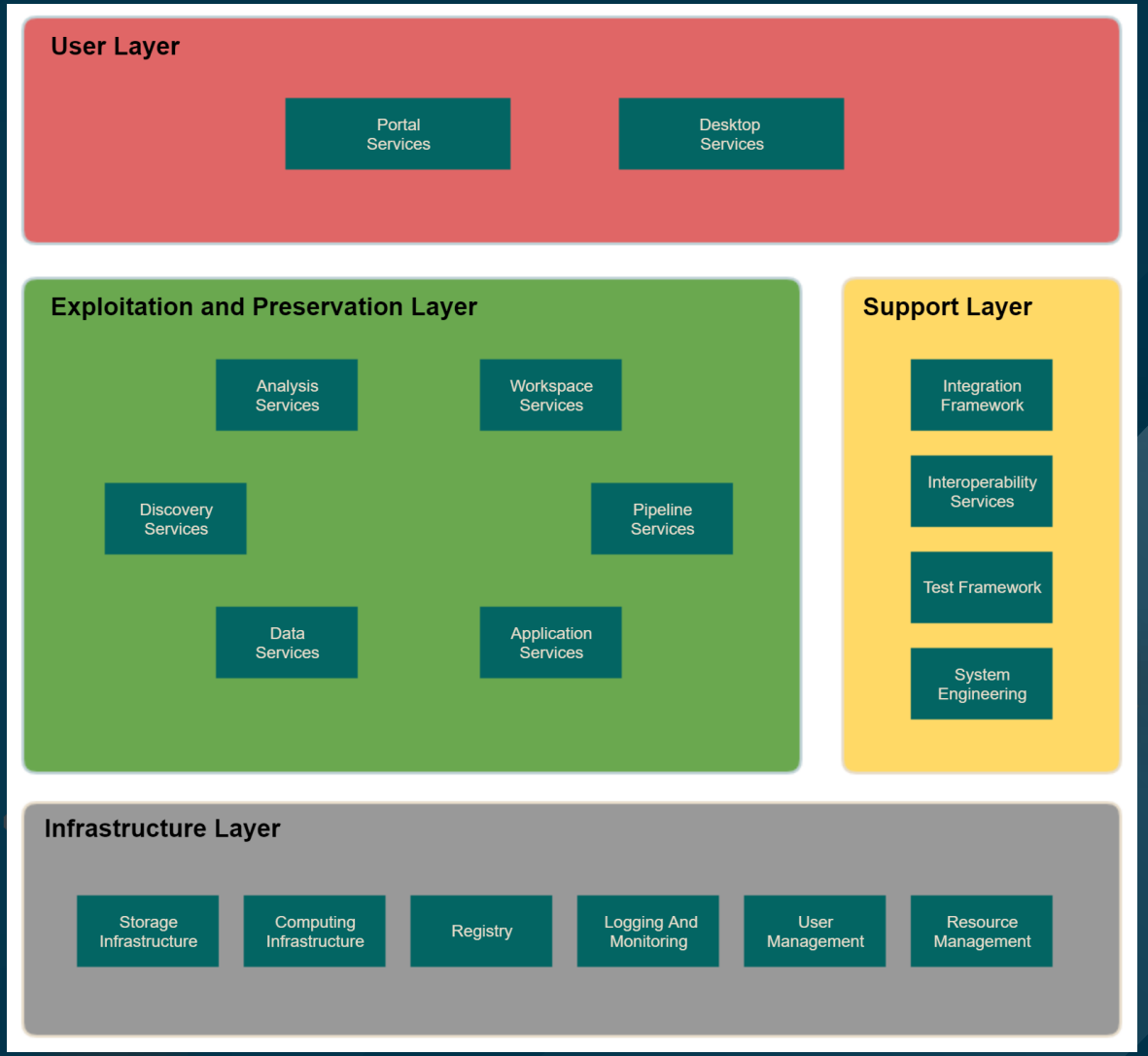


Science Archives Integration: GNSS



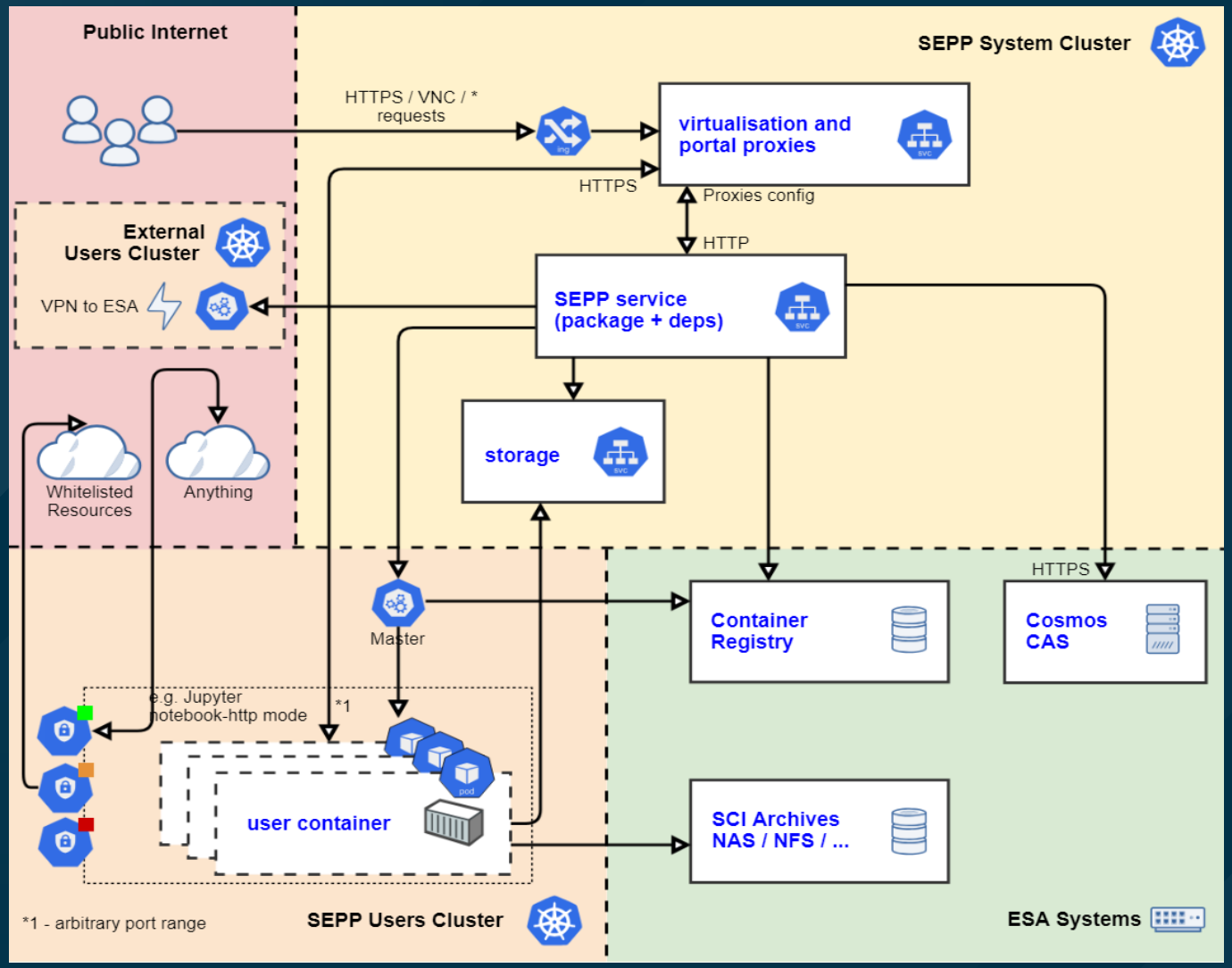
What's inside?

Reference Architecture leverages on mainstream technologies

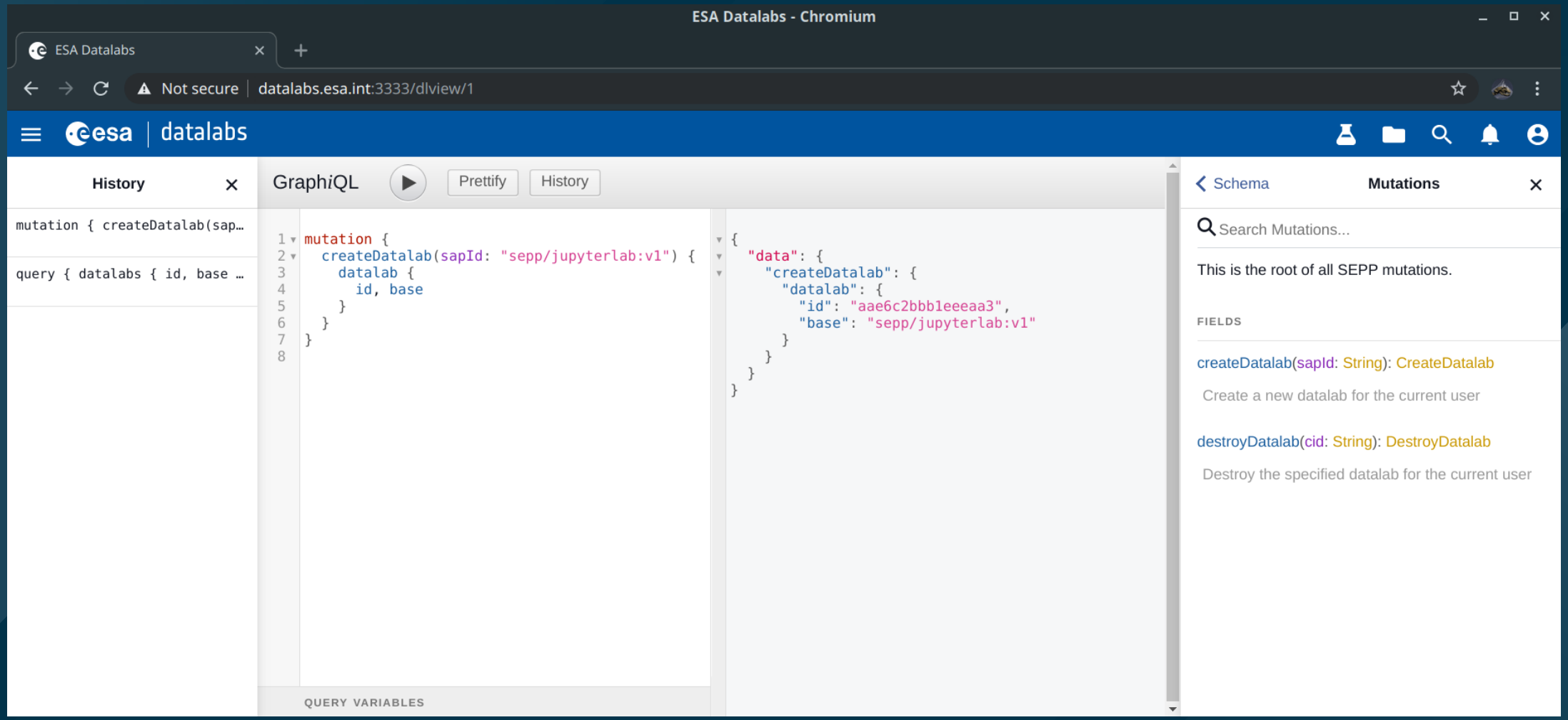


P 9.18 Del Rio, Navarro et al.

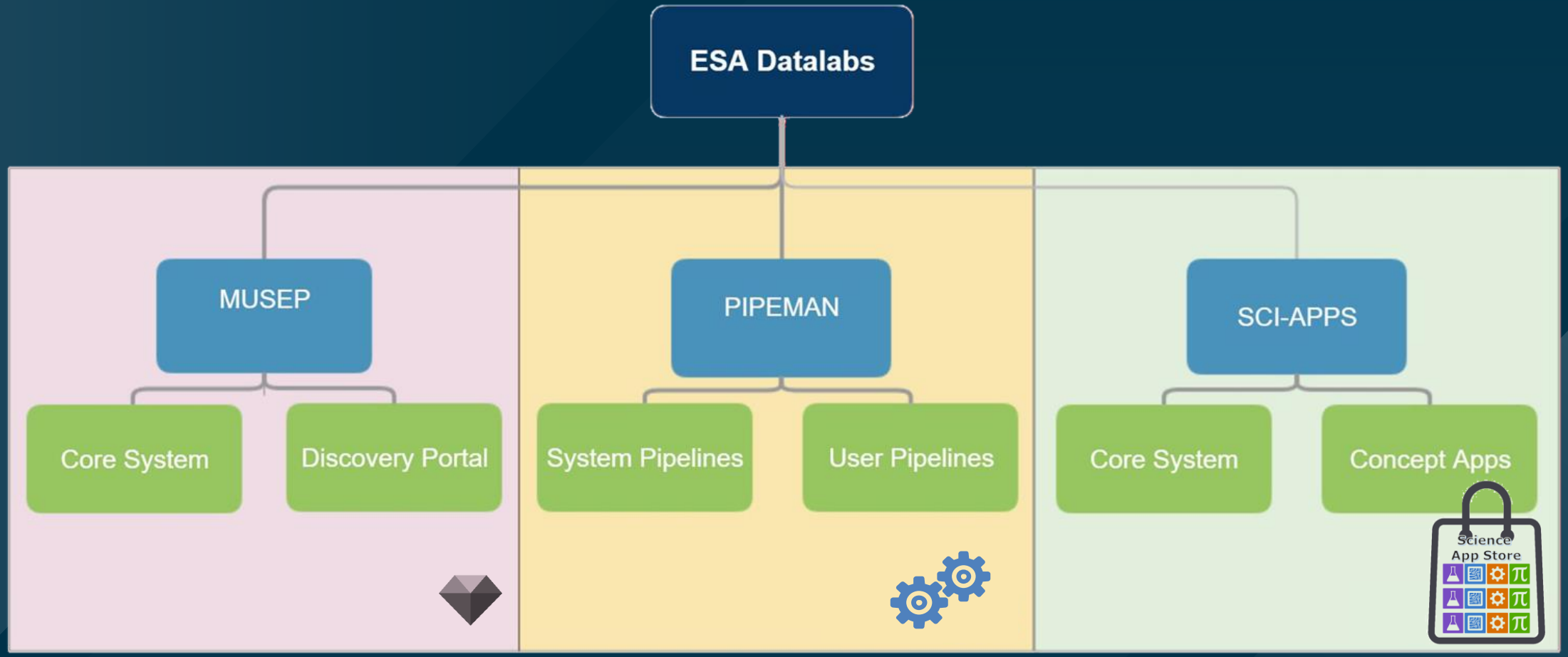
Container Management with Kubernetes



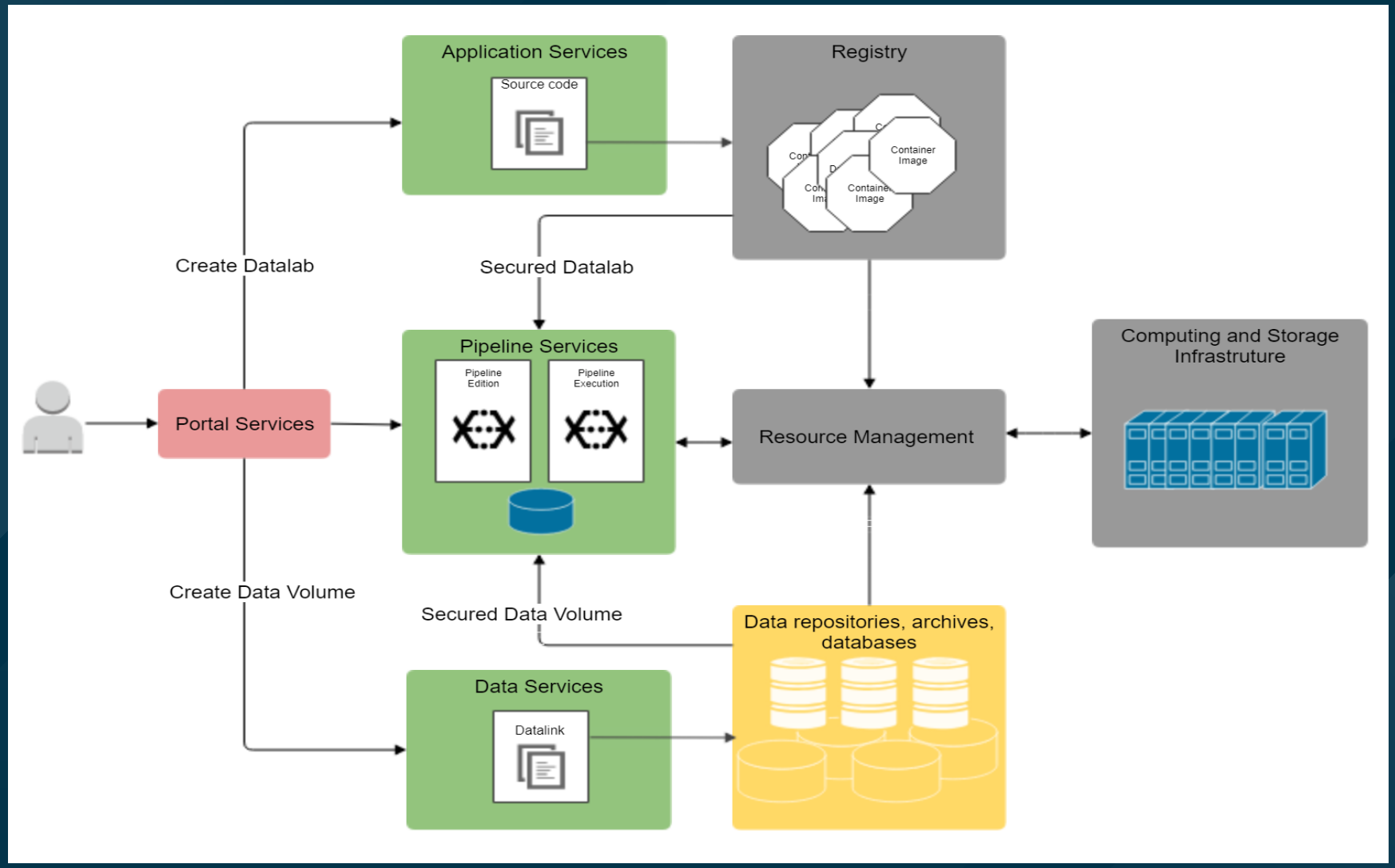
GraphQL Interface



Delivery Plan



Bringing all together



Timeline


05/2017
ESA Datalabs Workshop
03/2018
System Concept First Release
03/2019
Implementation Kick-Off
10/2019
MUSEP Seruices PDR
03/2020
MUSEP Seruices roll-out
Q3/2020
PIPEMAN Seruices roll-out
Q1/2021
SCI-APPS Seruices roll-out

Challenges

Security
Single-Sign-On, public, shared and restricted areas, code-injection …

1

Storage and Computing Dimensioning
Small evolution, quotas, reuse / evolution of existing infrastructure, pay-per-use hybrid models

2

Fast evolution of the technology:
Jupyter compatibility, Containers vs VMs, Kubernetes vs Swarm …

3

IPR
Define clear boundaries and recognition for user property rights

4

Open platform
Avoid vendor lock-in, develop and adopt OSS, standards and available ICT

5

Community driven
Close involvement of the scientific community to maximise science return

6

Conclusions

New missions call for:

- Paradigm shift from “bring the data to the user” to “bring the user to the data”
- Close interaction between archives and data processing services

Legacy missions call for data and software long term preservation

Scientists call for collaborative research environment

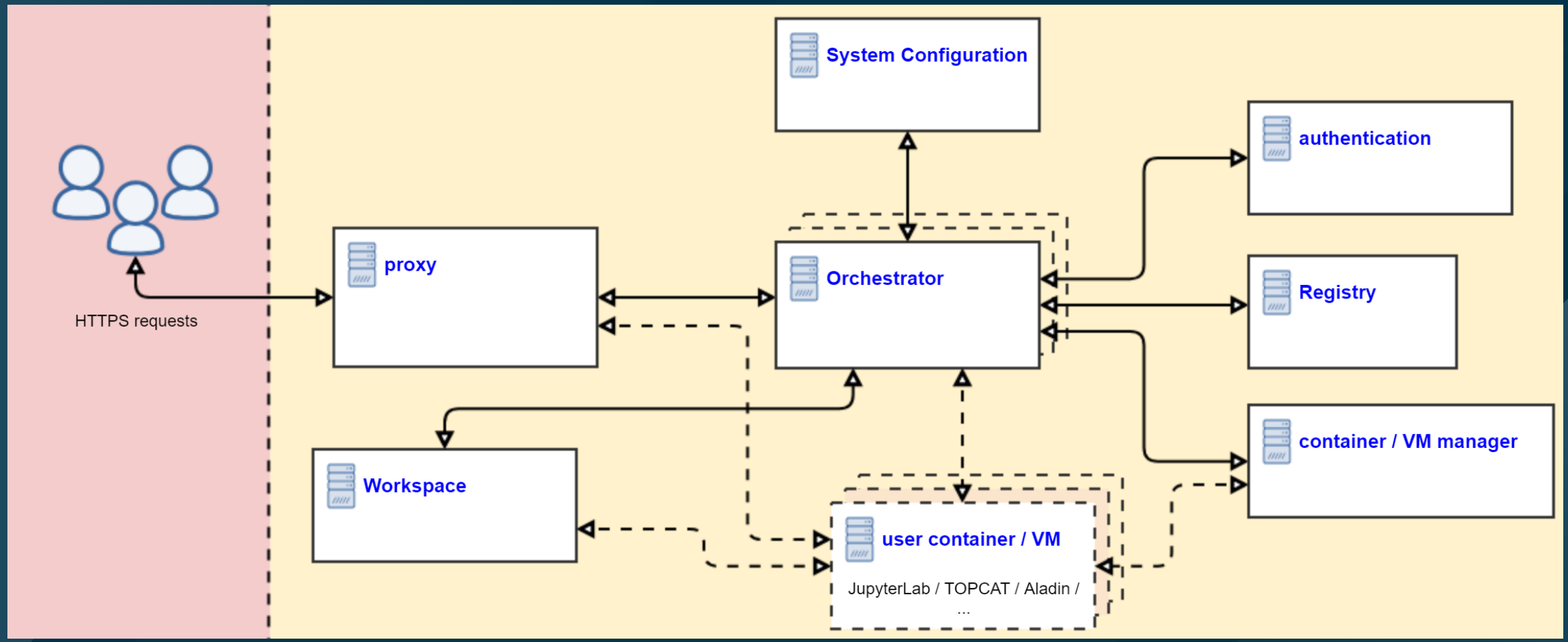
Leverage on existing VO standards, archives, mission and IT systems

Thank You!

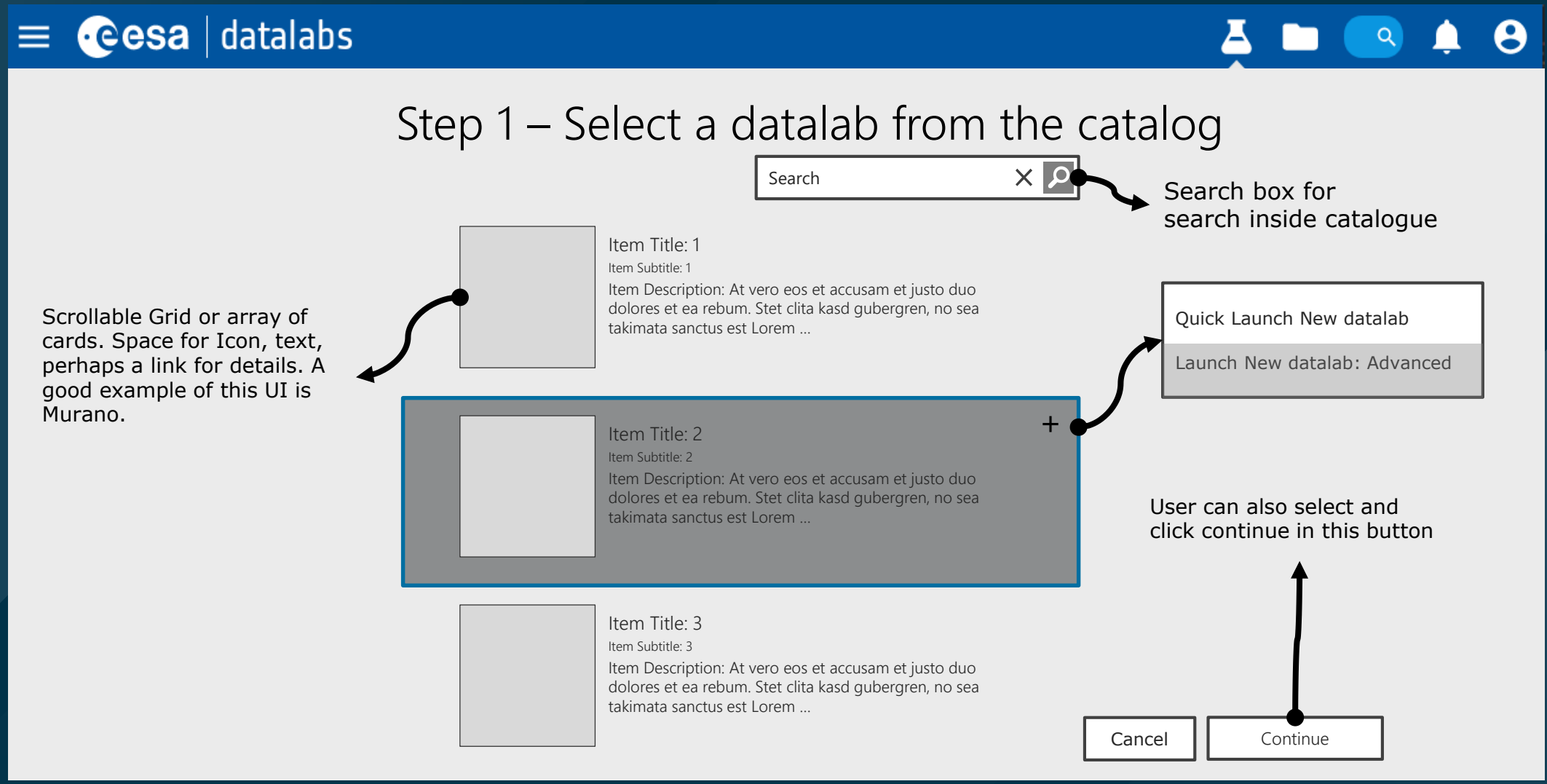
datalabs.esa.int

BackUp Slides

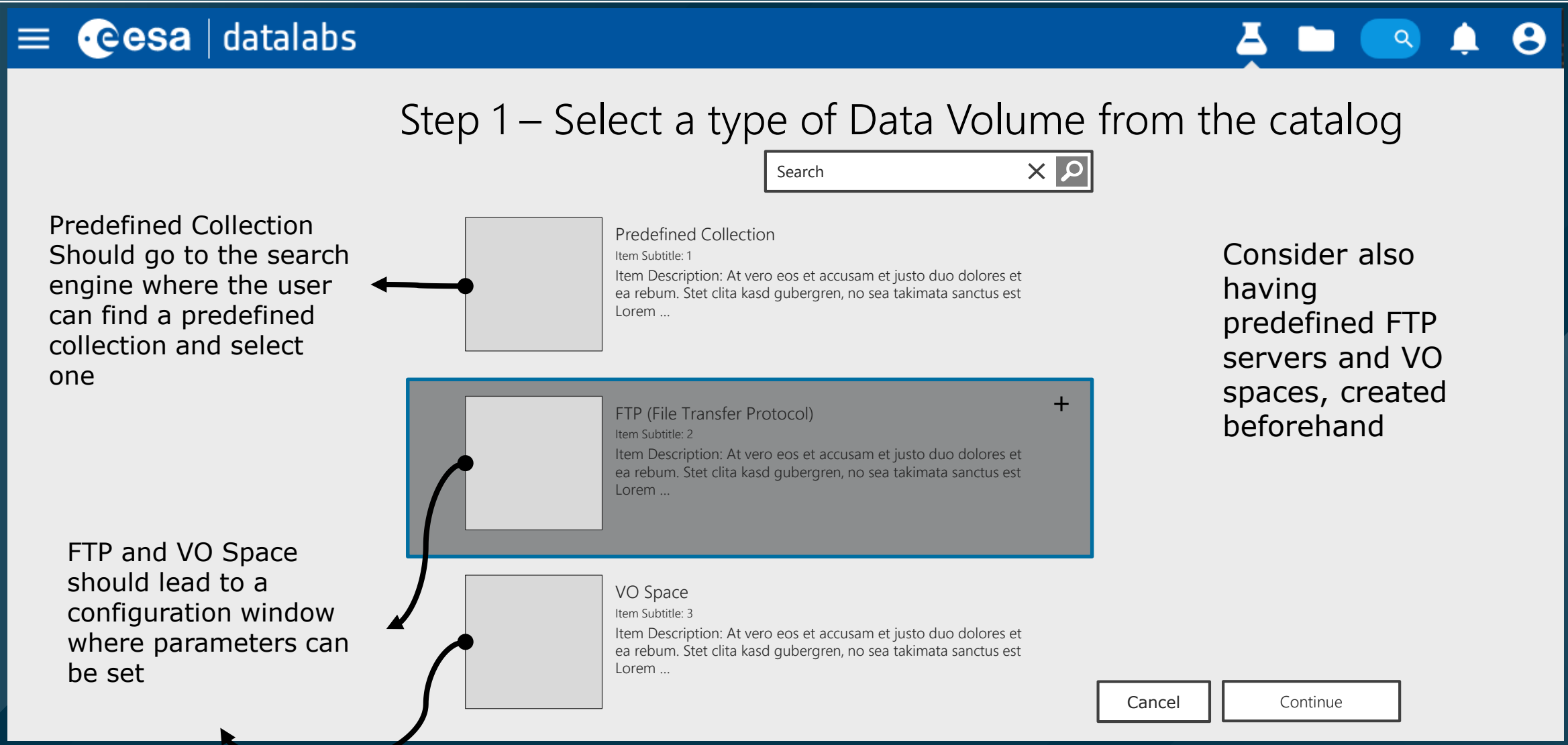
Container Management



Pipeline Services and Datalab Concept



Pipeline Services and Data Volume Concept

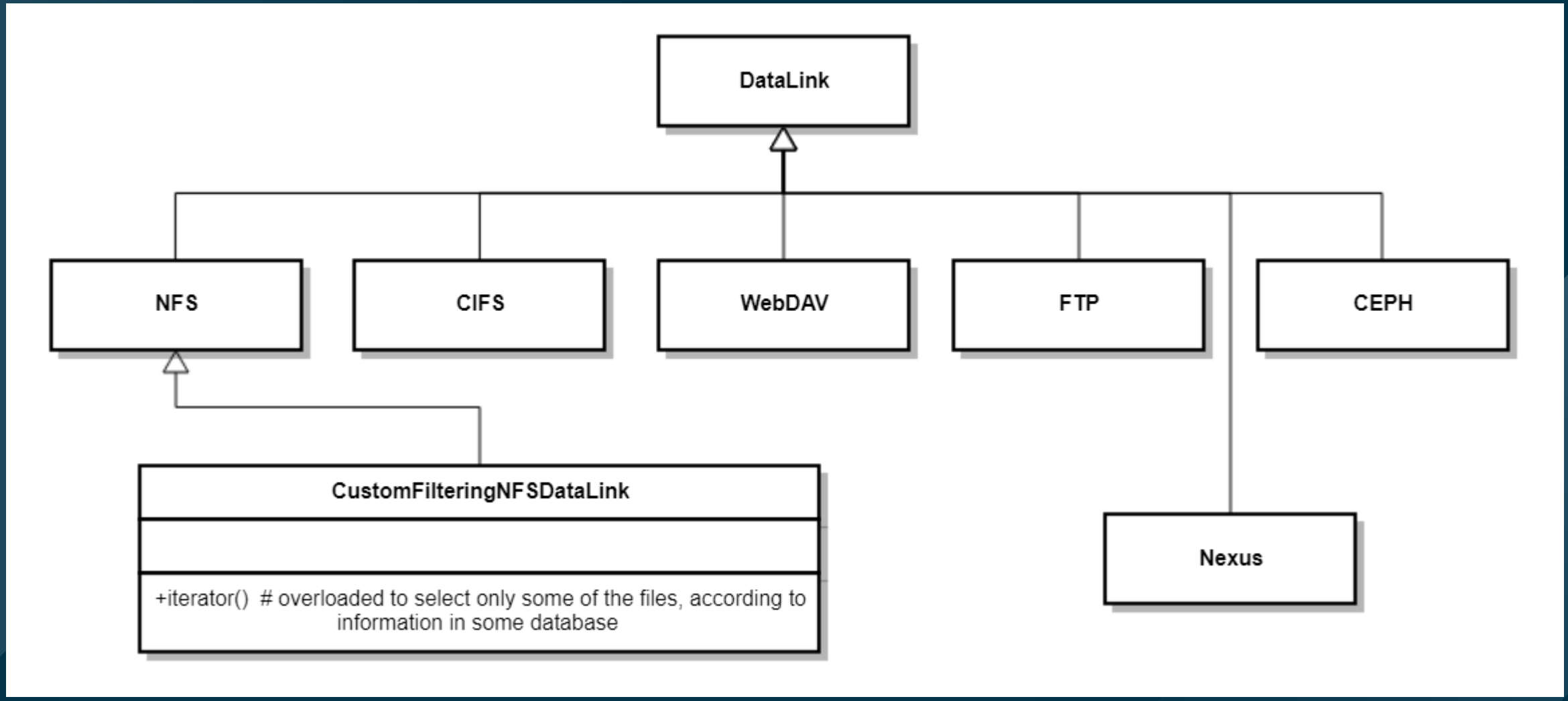


Predefined Collection Should go to the search engine where the user can find a predefined collection and select one

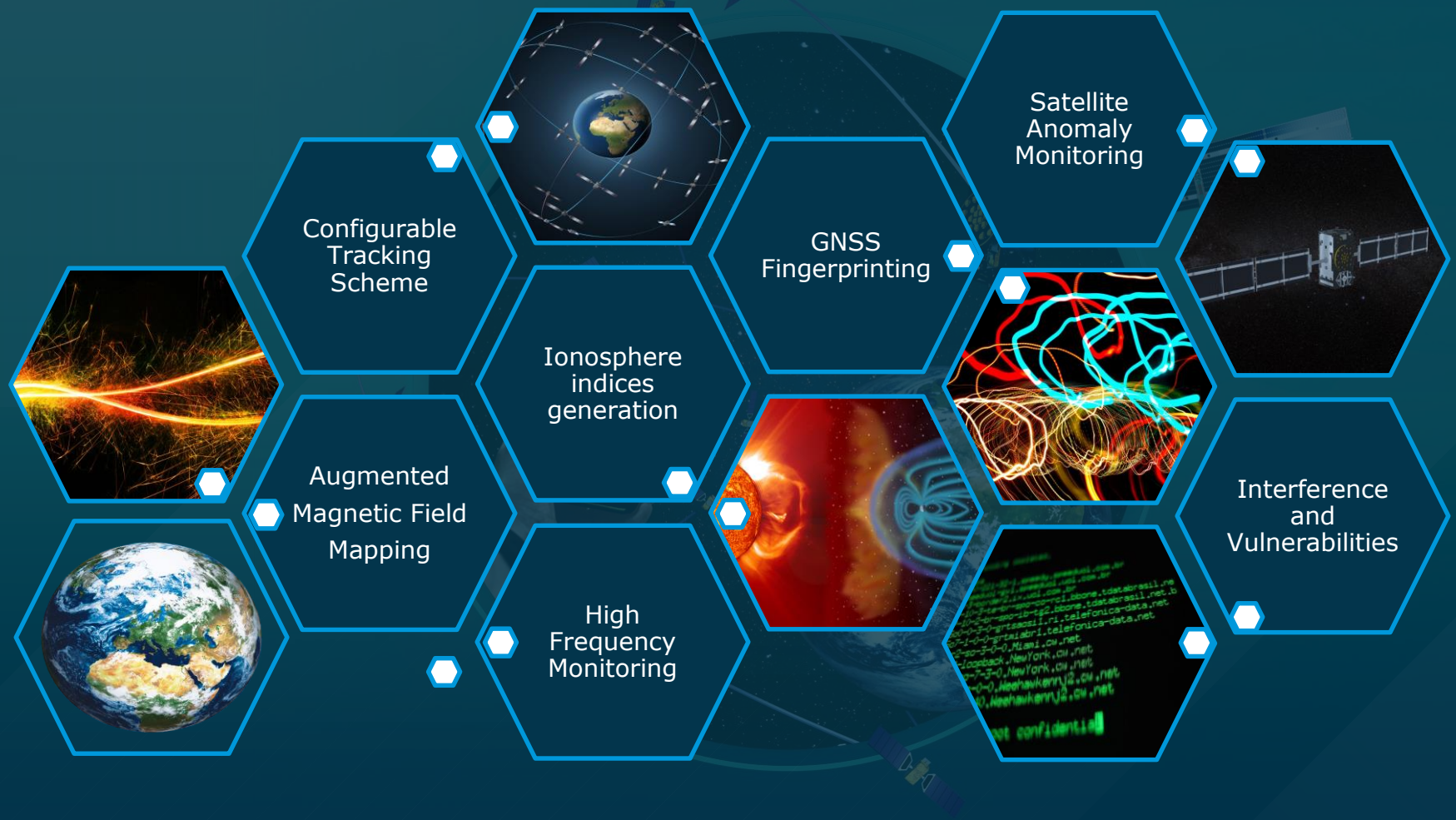
FTP and VO Space should lead to a configuration window where parameters can be set

Consider also having predefined FTP servers and VO spaces, created beforehand

Datalink


DataLink
NFS
CIFS
WebDAV
FTP
CEPH
CustomFilteringNFSDataLink
+iterator() # overloaded to select only some of the files, according to information in some database
Nexus

GSSC Exploitation Platform Innovative Use Cases



GSSC Jupyter Notebook - Configuration

Visualisation Layer	•Matplotlib •iPympl •Beakerx •iPyleaflet
GNSS Layer	•GSSC LoggerReader •Georinex: load Rinex files •Ephem: leap second calculation
Mathematical Layer	•Numpy:scientific computing Pandas:Data Analysis Library •PyAstronomy: calculation of orbital parameters •Pyproj: Reference Coordinate System Transformation
Core Layer	•NBconvert