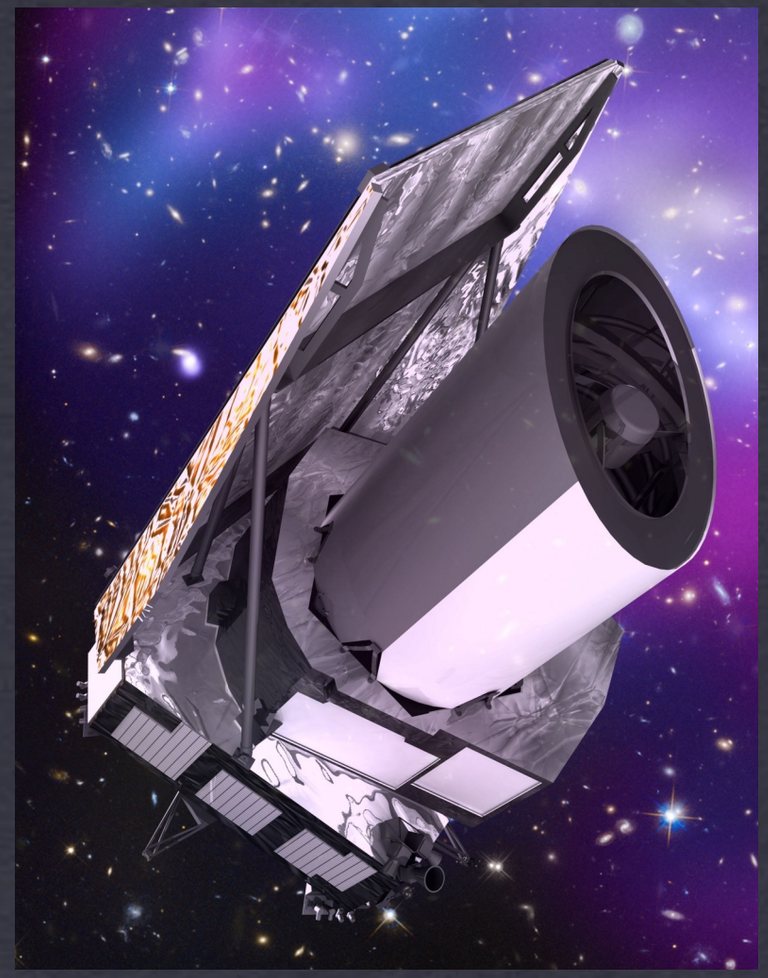
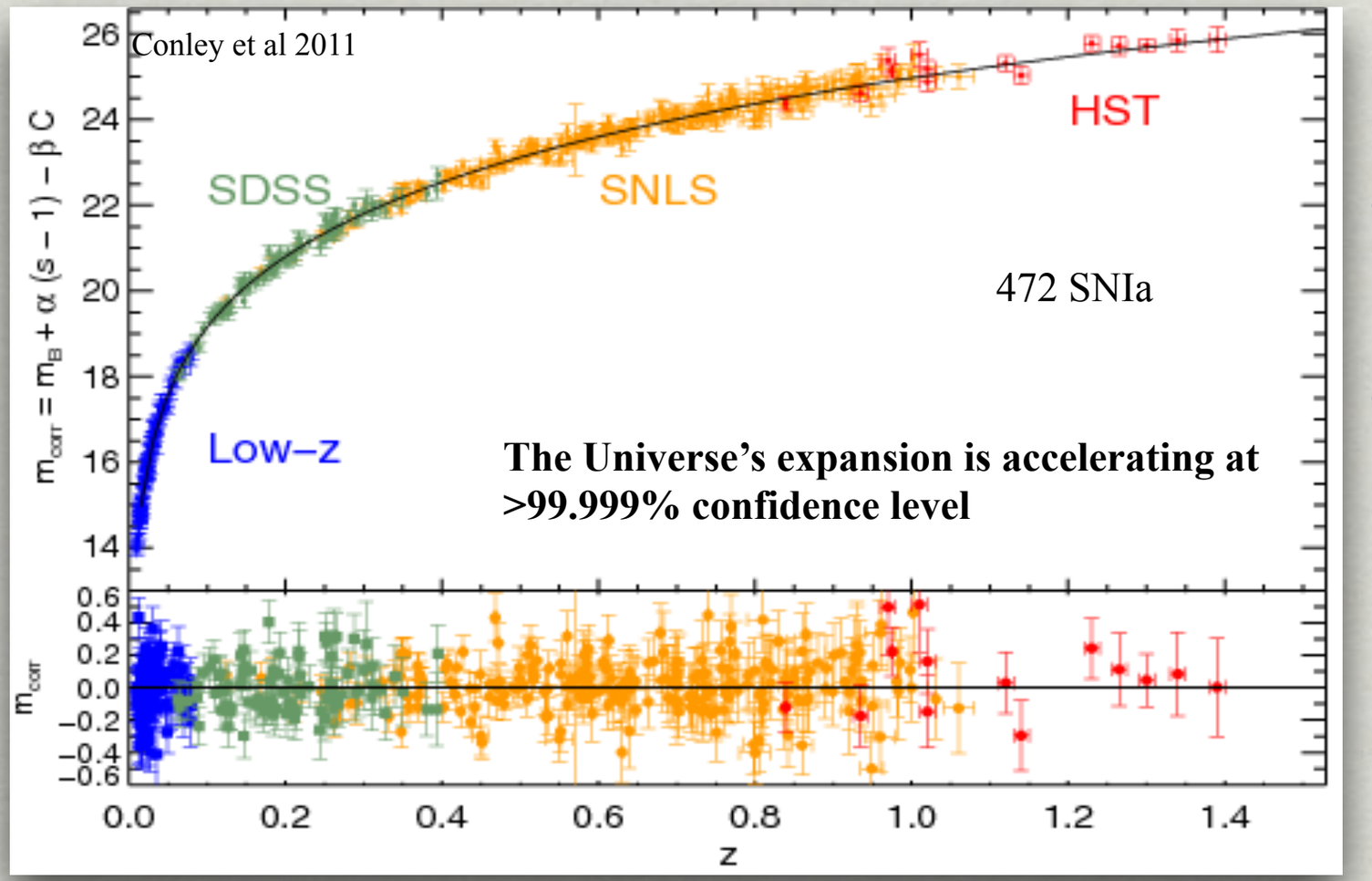
Data acquisition, processing and distribution for the Euclid Mission

M. Sauvage - Euclid Science Ground Segment
with Ch. Dabin, Y. Mellier, J. Hoar, M. Poncet, A. Zacchei
and F. Pasian

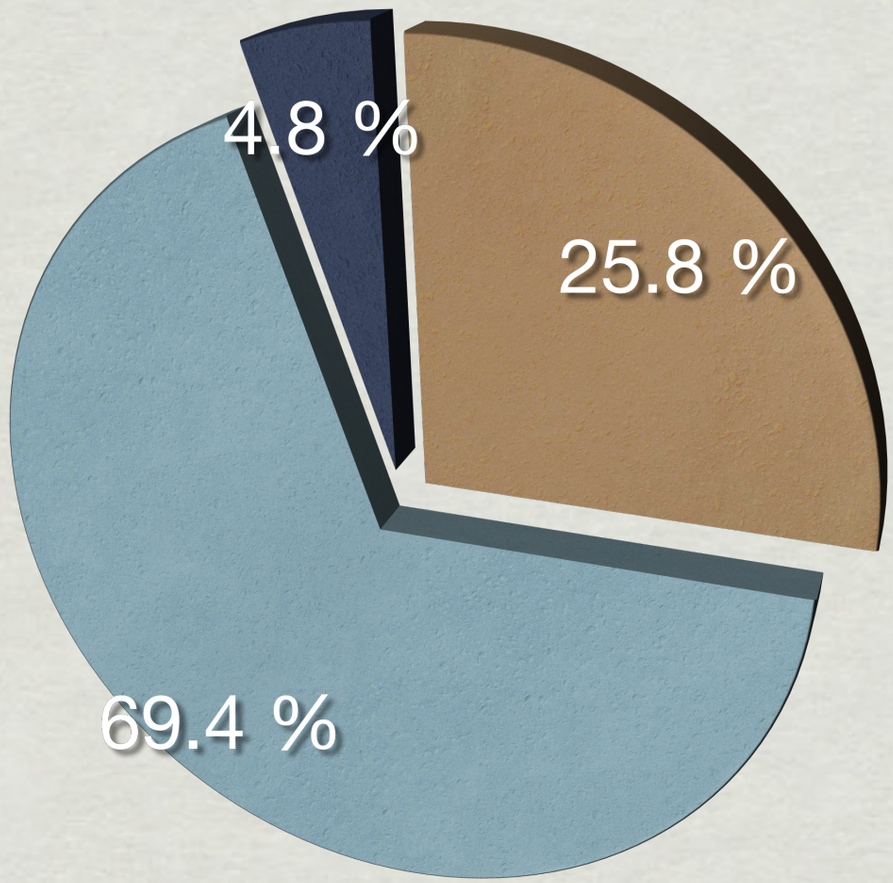


The Euclid Mission

The accelerating Universe



- This Hubble Diagram shows the relation between the corrected magnitude (equivalent to the distance derived from the luminosity) and the redshift.
- The fit to this relation requires a cosmological model, here incorporating dark matter and dark energy.

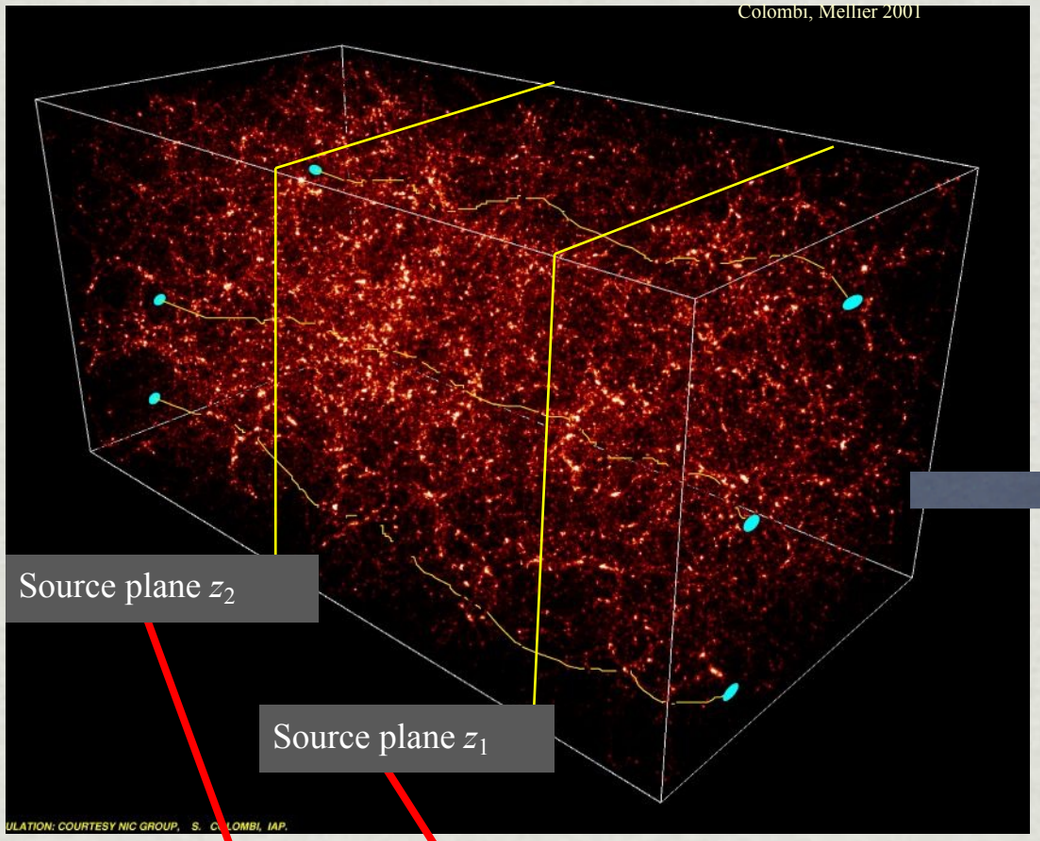


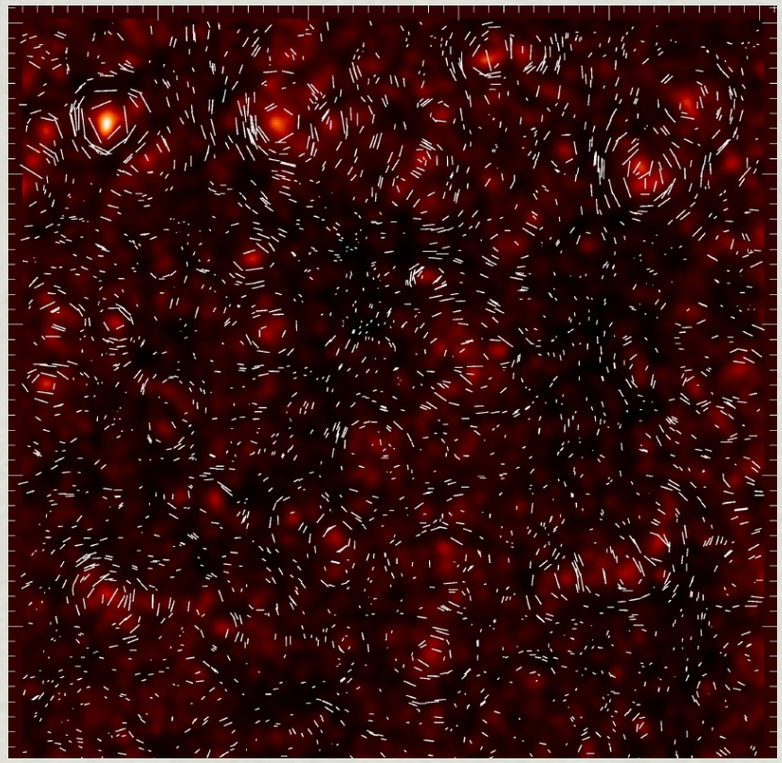
Baryons, photons, neutrinos
Dark Matter
Dark Energy

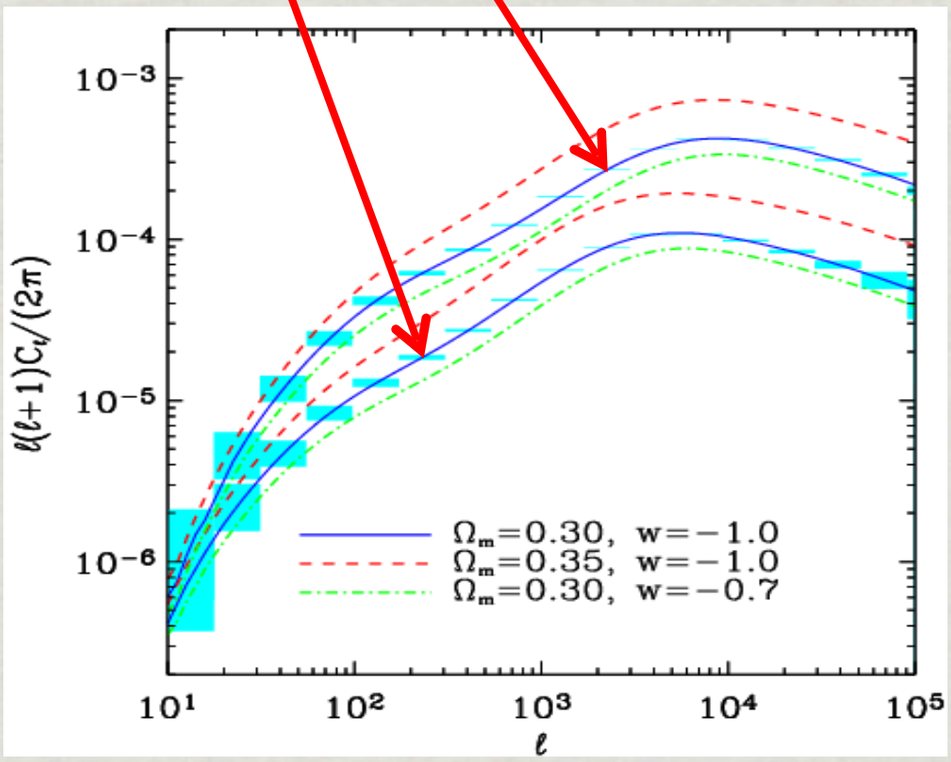
The Universe according to Planck (Planck collaboration. Ade et al 2013)

- Euclid's objectives: Probe the effects of DE, DM and Gravity by tracking their observational signatures on:
  - The geometry of the Universe.
  - The cosmic history of structure formation.

Weak Lensing tomography over 0<z<2

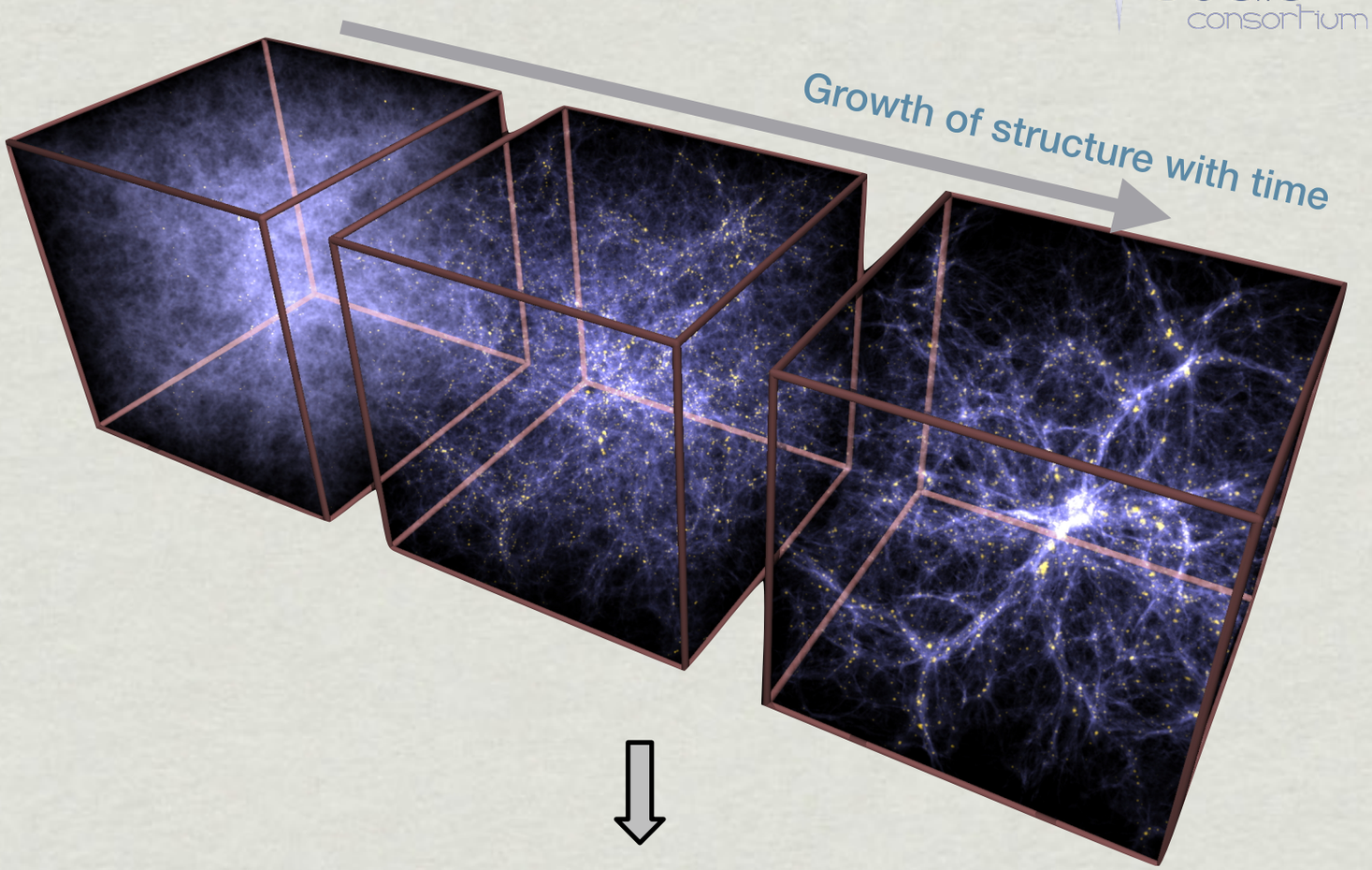




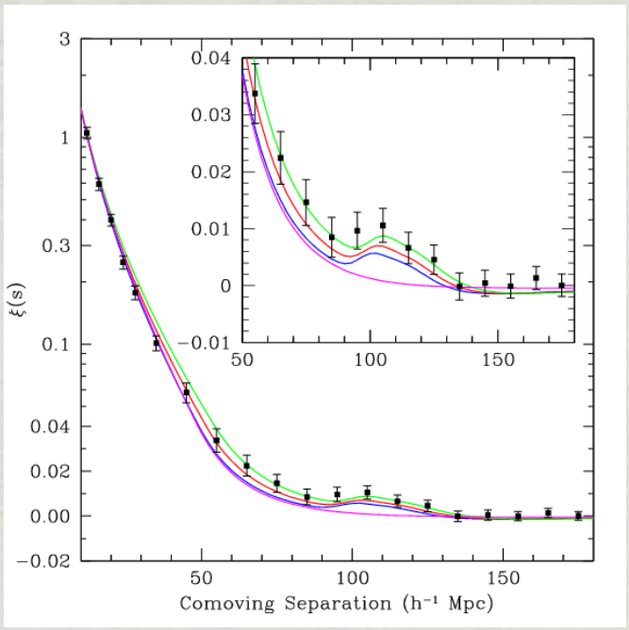


- Probes distribution of matter (Dark mostly): expansion history, growth rate of structure formation.
  - Shapes of galaxies: shear amplitude or mass map
  - “Photometric redshifts” to bin the maps in slices:
    - NIR data obtained from Euclid NISP instrument
    - Optical data requires ground-based coverage
- Euclid will measure Weak-Lensing with the shapes of 1.5 10^9 galaxies over 15,000 deg^2.
  - The catalog will contain photometric information on about 10^{10} galaxies and other objects.

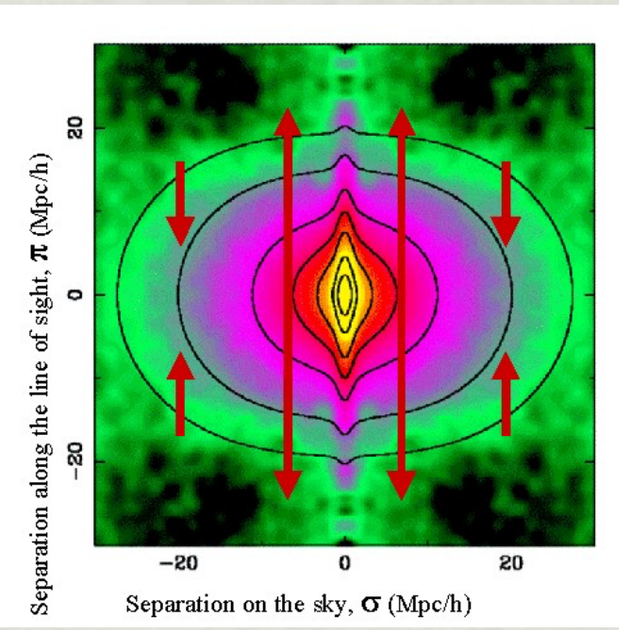
Galaxy Clustering



- Probes DE properties and expansion rate of the Universe (BAO) and clustering history of galaxies induced by gravity (RSD).
  - Need high precision 3-D distribution of galaxies with spectroscopic redshifts over $0.7<z<1.8$
- Euclid will obtain 30 million spectroscopic redshifts with 0.001 (1+z) accuracy over 15,000 deg^2.
  - Spectra of all photometrically detected objects can be extracted.

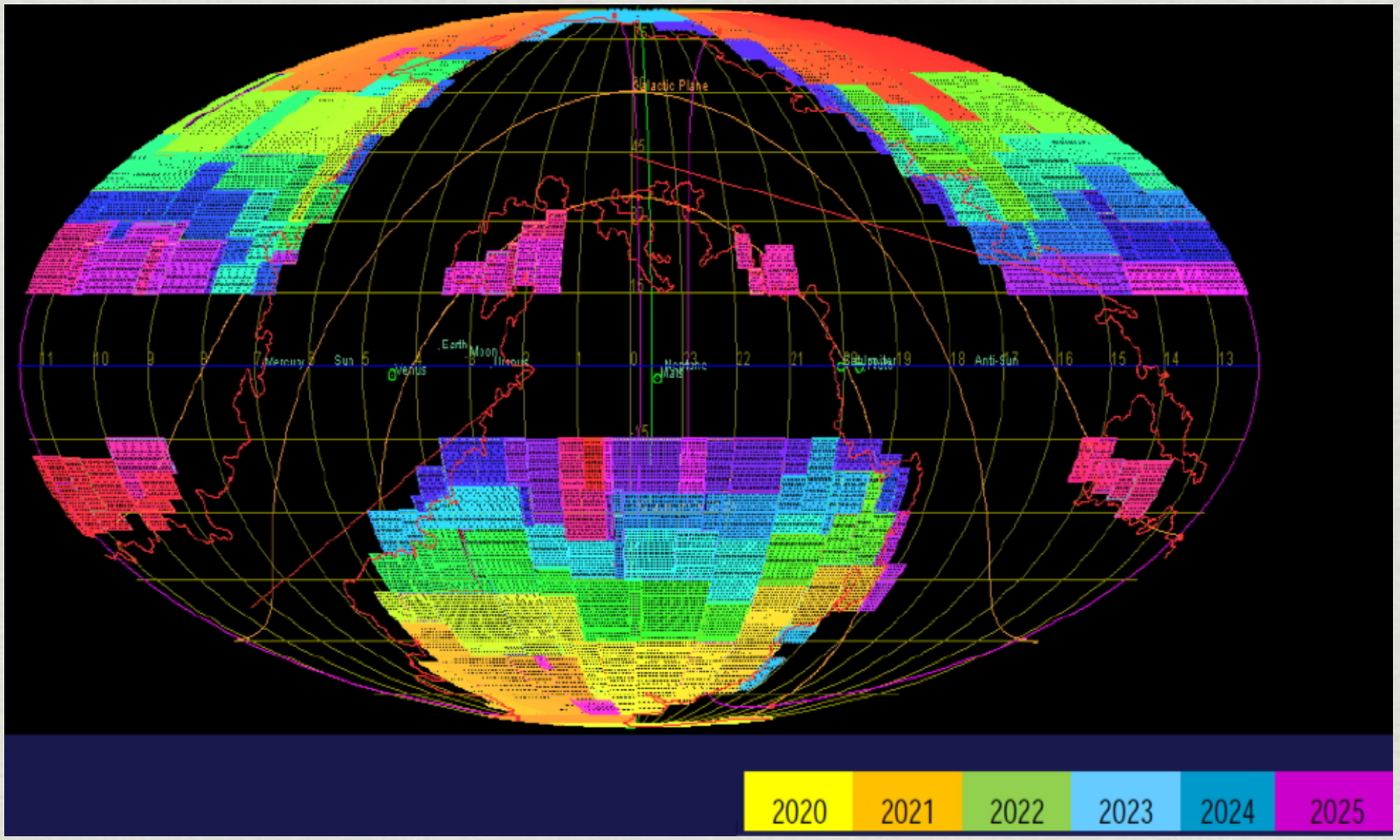


BAO



RSD

Euclid Survey and data



Euclid is a cosmological survey mission, but unlike CMB experiments, it will only do its survey once!

Survey strategy is constrained by the number of times we can point the satellite!

Euclid is not a time-domain experiment.

VIS:

- Imaging
- 36 4k x 4k CCD
- 0.54 deg^2 per field
- 0.1” pixels on the sky
- limiting magnitude: 24.5 AB @10σ
- 520 Gbit/day

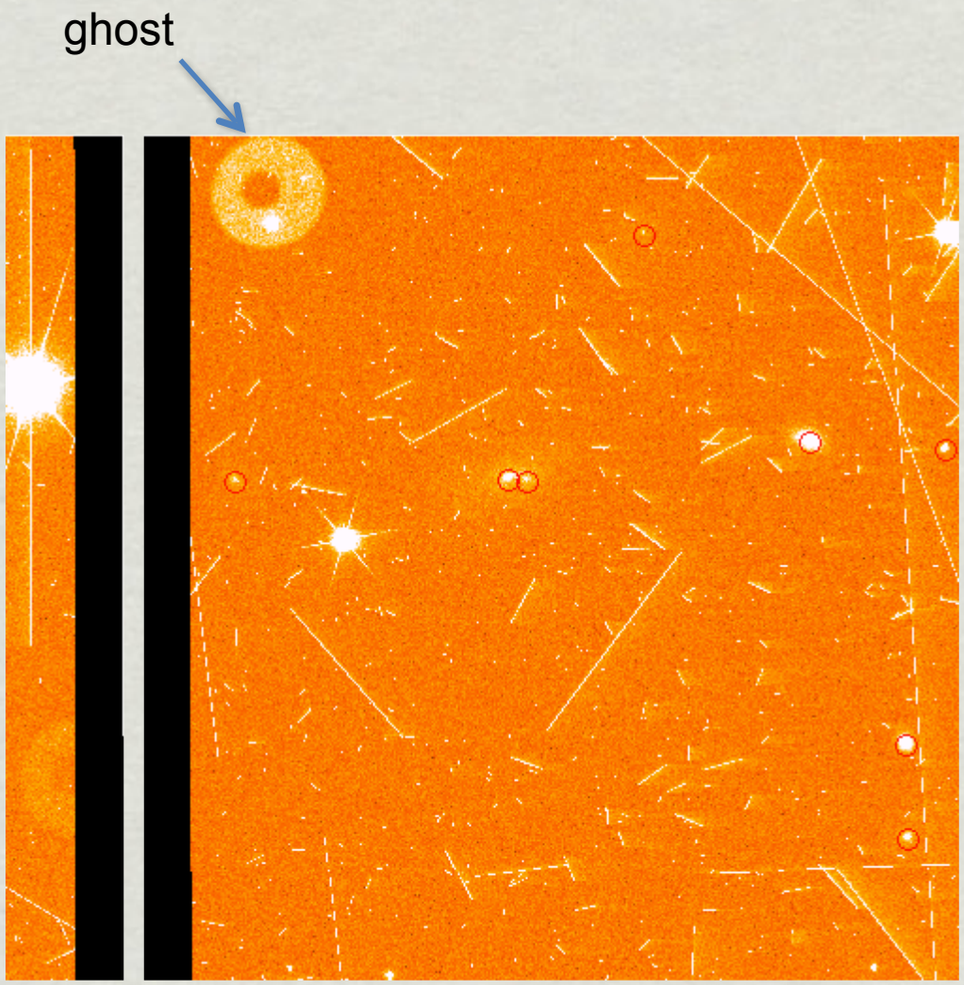
NISP:

- Imaging and slitless grism spectroscopy
- 16 2k x 2k NIR arrays
- 0.55 deg^2 per field
- 0.3” pixels on the sky
- limiting magnitude: 24 AB @5σ
- limiting flux: Hα 2 10^{-16} erg.cm^{-2}.s^{-1} @3.5σ
- 240 Gbit/day

→ **800 Gbit/day for 6 years**

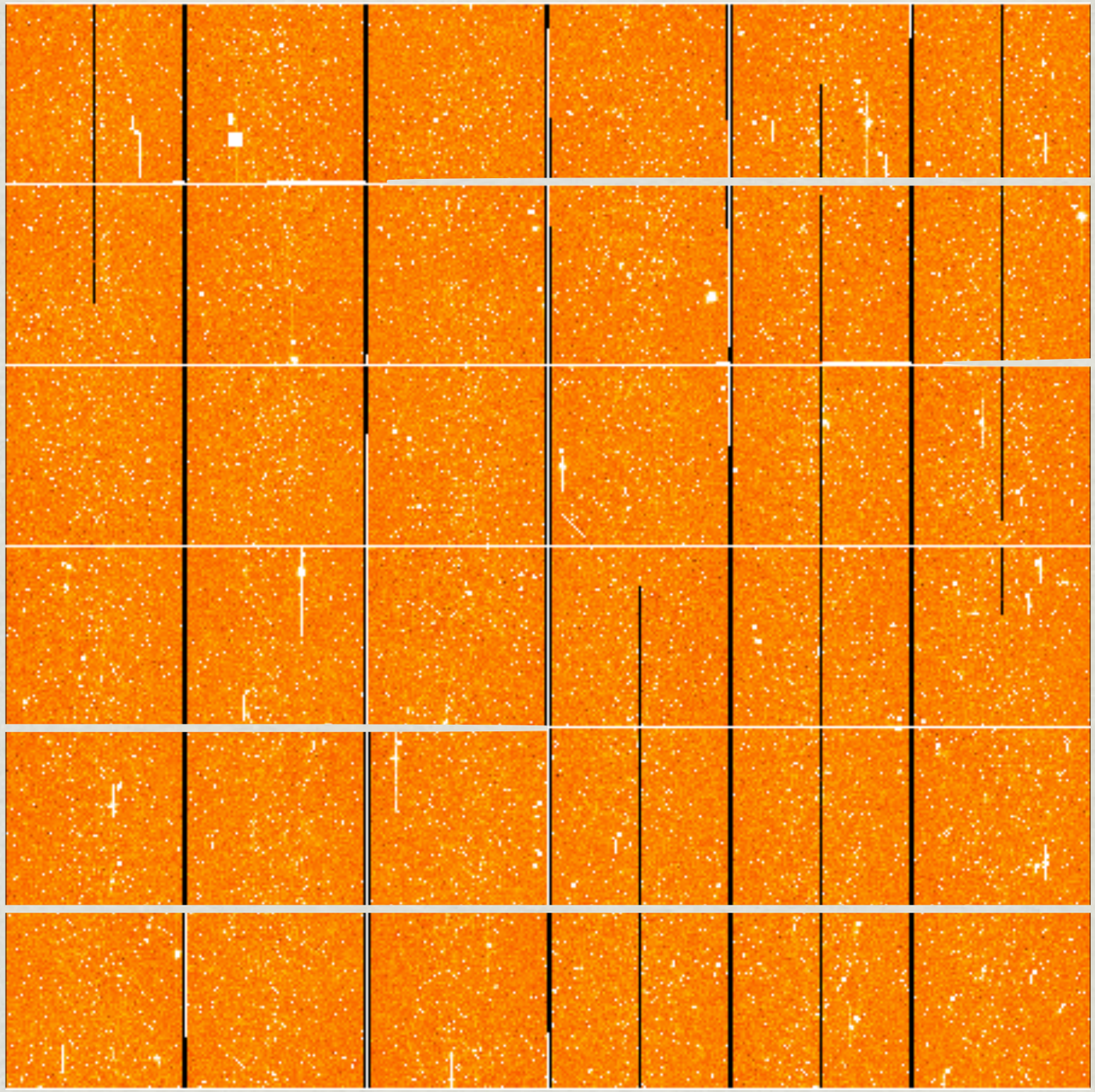
Visible and infrared imaging, as well as infrared spectroscopy are obtained “simultaneously”

VIS imaging



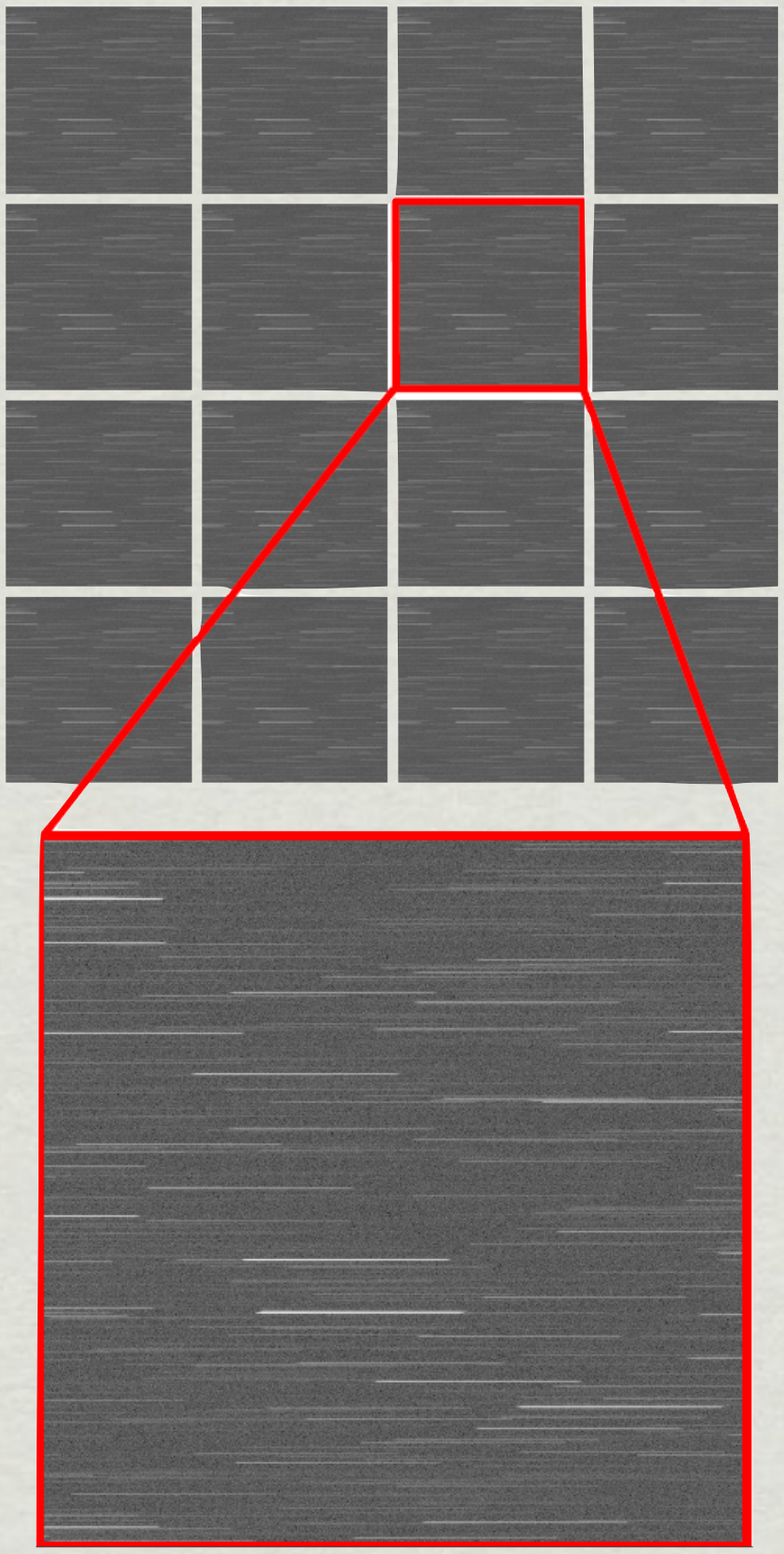
Simulations by OU SIM (provided by E. Jullo)

A simulated Full Focal Plane view of the Euclid sky

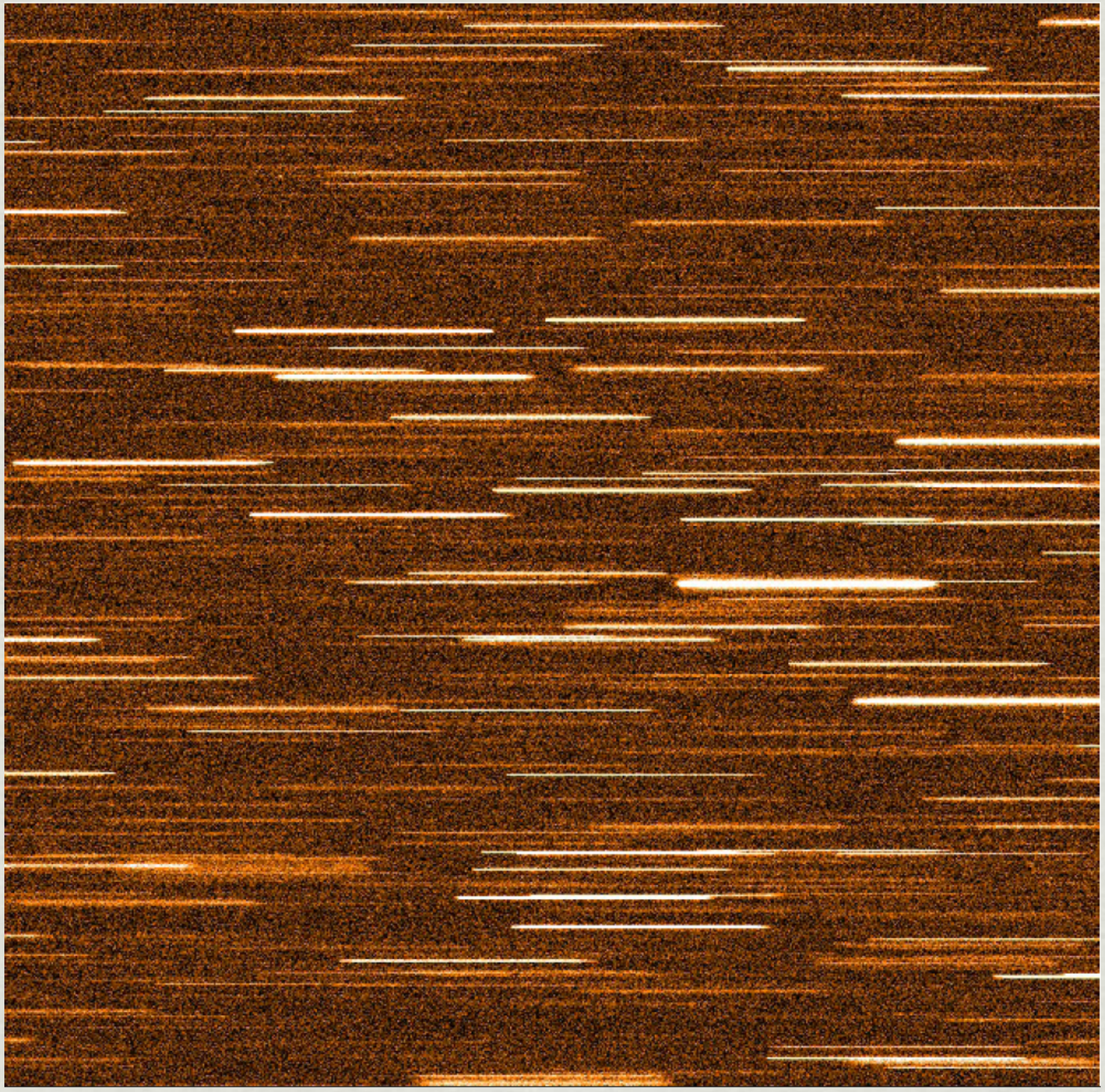


VIS FPA

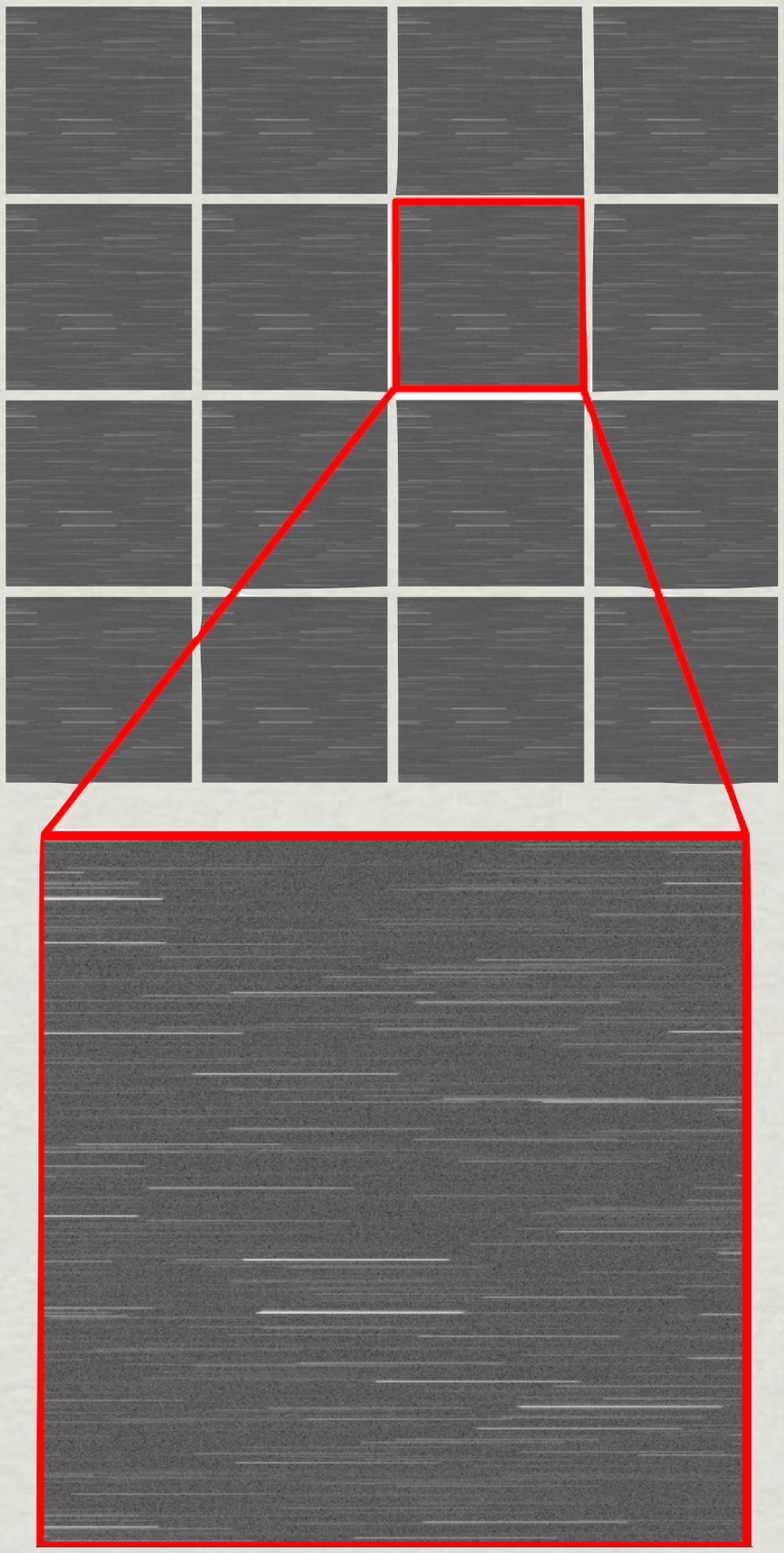
NISP spectroscopy



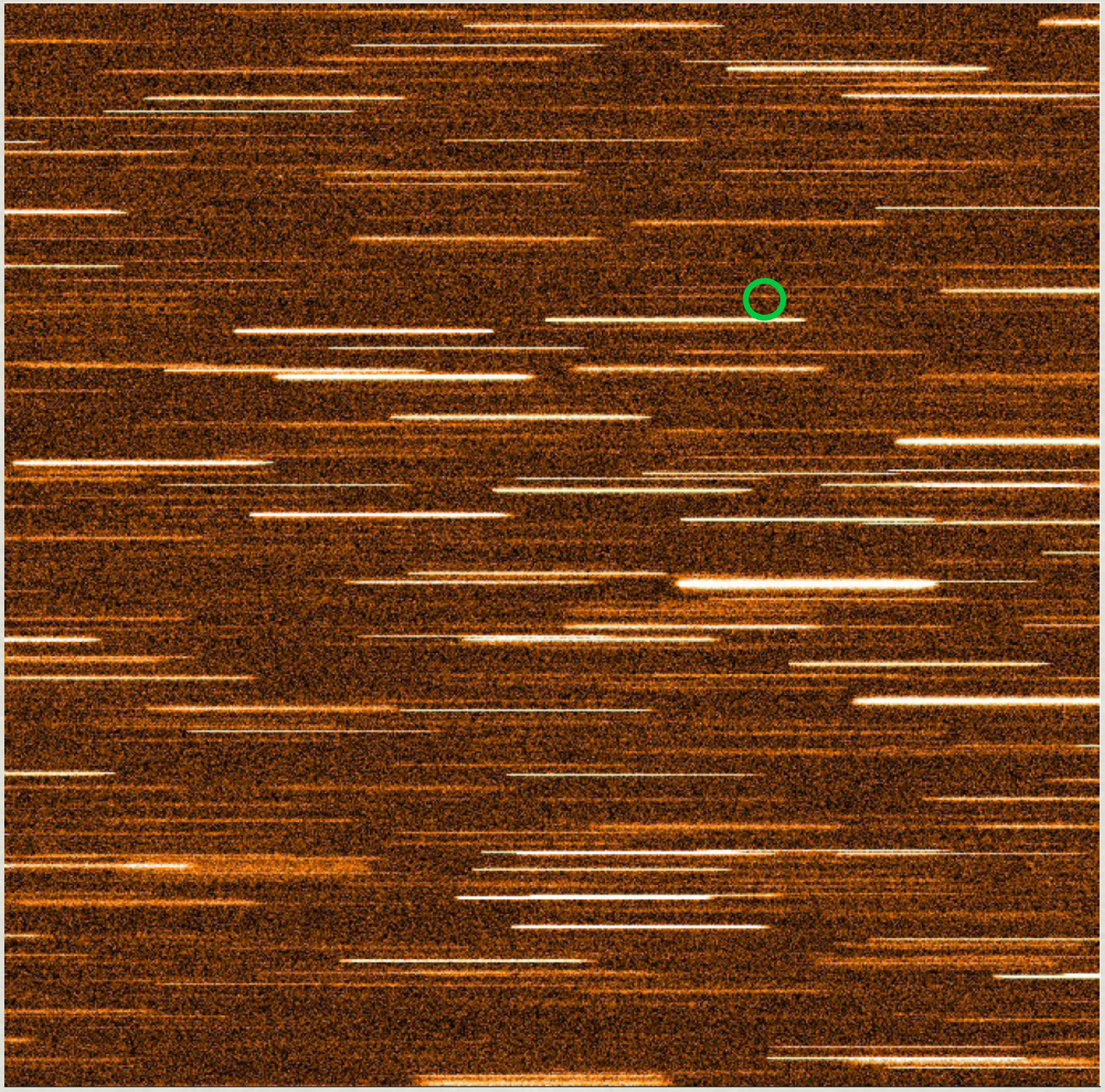
2015 simulations from P. Franzetti, B. Garilli, A. Ealet, N. Fourmanoit & J. Zoubian



NISP spectroscopy



2015 simulations from P. Franzetti, B. Garilli, A. Ealet, N. Fourmanoit & J. Zoubian



Challenges of the Euclid Mission

- Data volume is not the major challenge, although we have to think about it in planning the system.
  - Although the necessity to combine with ground-based data modifies that picture significantly (and create organisational challenges).
- Keeping the processing in synch with the data acquisition is a major challenge:
  - Euclid is a space mission, it is operating in a hostile environment and thus data acquisition can only be stopped for very serious reasons.
  - Pointing to a given part of the sky is subject to severe constraints, so deviation from the survey plan must be considered seriously.
- Tracing the data processing, the calibration sources, the systematics present in the final data products is by far the main challenge of Euclid.
  - The signal of cosmological interest is a very small fraction of the measured signal (e.g. in the case of weak-lensing), or is hidden is angular correlations that must be cleaned of survey artefacts (e.g. in the case of galaxy clustering).
  - This will have to rely on extensive sets of simulated data injected in the data processing pipeline at various stages.
    - “secondary” ressource challenge, which projects Euclid into the data challenging mission.

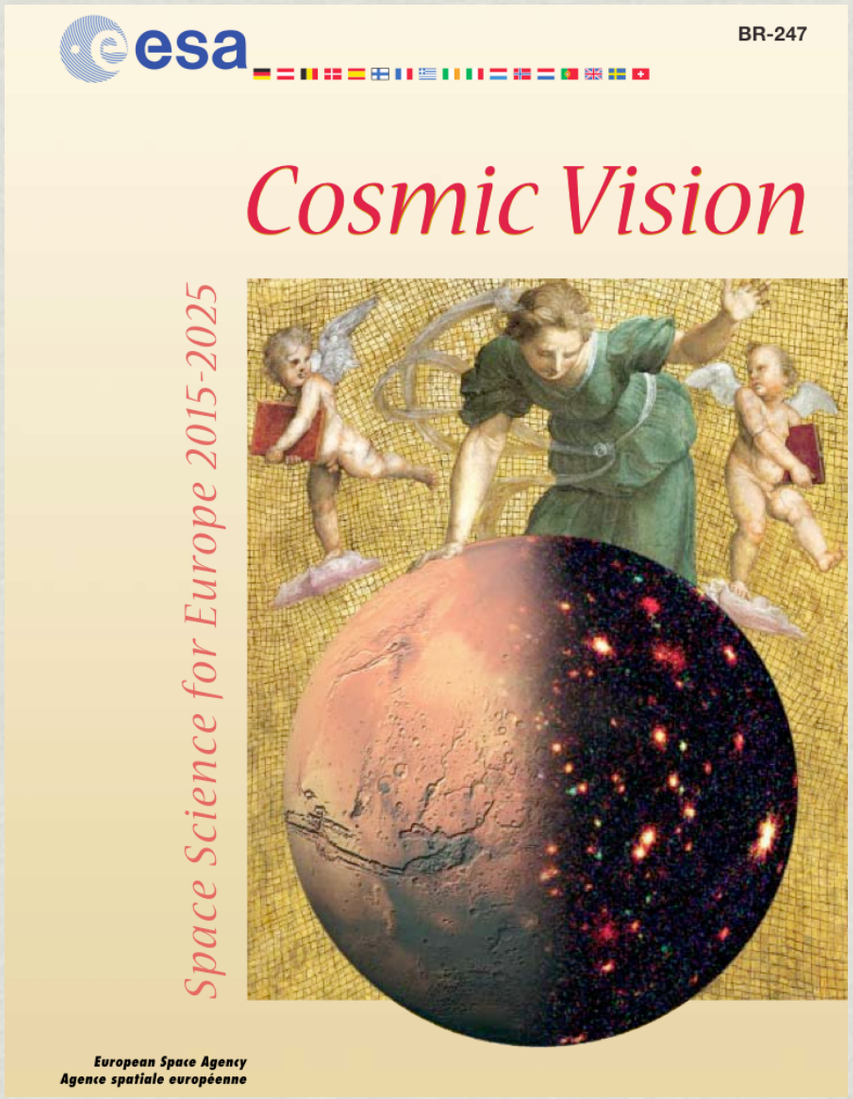
Particulars of an ESA mission

- Euclid is part of the Cosmic Vision program of ESA.
- This imposes constraints on the cost structure.
  - No fund transfer possible between countries, hence no possibility to jointly fund computing resources.
- ESA considers that its mandate is to deliver missions that have performances in line with the original expectations, and data products that will enable the best science.
  - It does not consider scientific exploitation of the data as part of its mandate.
- ESA however considers that distribution of public data is its responsibility.
  - The Euclid consortium has no requirement to provide a service to the general astrophysical community, only to its consortium.
  - We have a requirement to deliver data to ESA with a planned schedule.

Euclid consortium

esa

European Space Agency



Data processing by the Science Ground Segment

Foreword

- The SGS will not produce measurements of the Dark Energy EoS, or any statement regarding alternate theory of Gravitation.
- The SGS will not produce statements regarding Galaxy Evolution or the Primordial Universe.
- However, the SGS is tasked with turning the measurements made by Euclid (wide-field photometric exposures and slit-less grism exposures) into data products from which the above results can directly be extracted.
  - Correlation functions, power spectra (and associated “errors”) for shear and positions.
  - Source catalogs containing, photometry, spectroscopy (lines and fluxes), redshifts (photometric and spectroscopic), shapes (ellipticities, morphologies), *physical parameters* (for legacy studies).

This is the “science” part of the SGS task, data product interpretation is the task of the Science Working Groups in the Euclid Consortium

Working structures within the EC-SGS

- For the data processing implementation phase, the EC-SGS consists mainly of two series of entities:
  - The Organization Units (OUs)
  - The Science Data Centers (SDCs)

Working structures within the EC-SGS

- For the data processing implementation phase, the EC-SGS consists mainly of two series of entities:
  - The Organization Units (OUs)
  - The Science Data Centers (SDCs)

The SDCs are built around existing national computing facilities, they gather IT support as well as developper expertise.
SDCs are in charge of the pipeline development (software & support architecture).
Production software will run in the SDCs.
The SDCs and SOC are the operational sites of the SGS.

The OUs group EC members according to their data processing expertise.
OUs are in charge of analyzing the science data processing requirements, and of producing the pre-integration version of the pipeline modules.
OUs will be in charge of future evolution of the pipeline during operations.

Working structures within the EC-SGS

- For the data processing implementation phase, the EC-SGS consists mainly of two series of entities:
  - The Organization Units (OUs)
  - The Science Data Centers (SDCs)

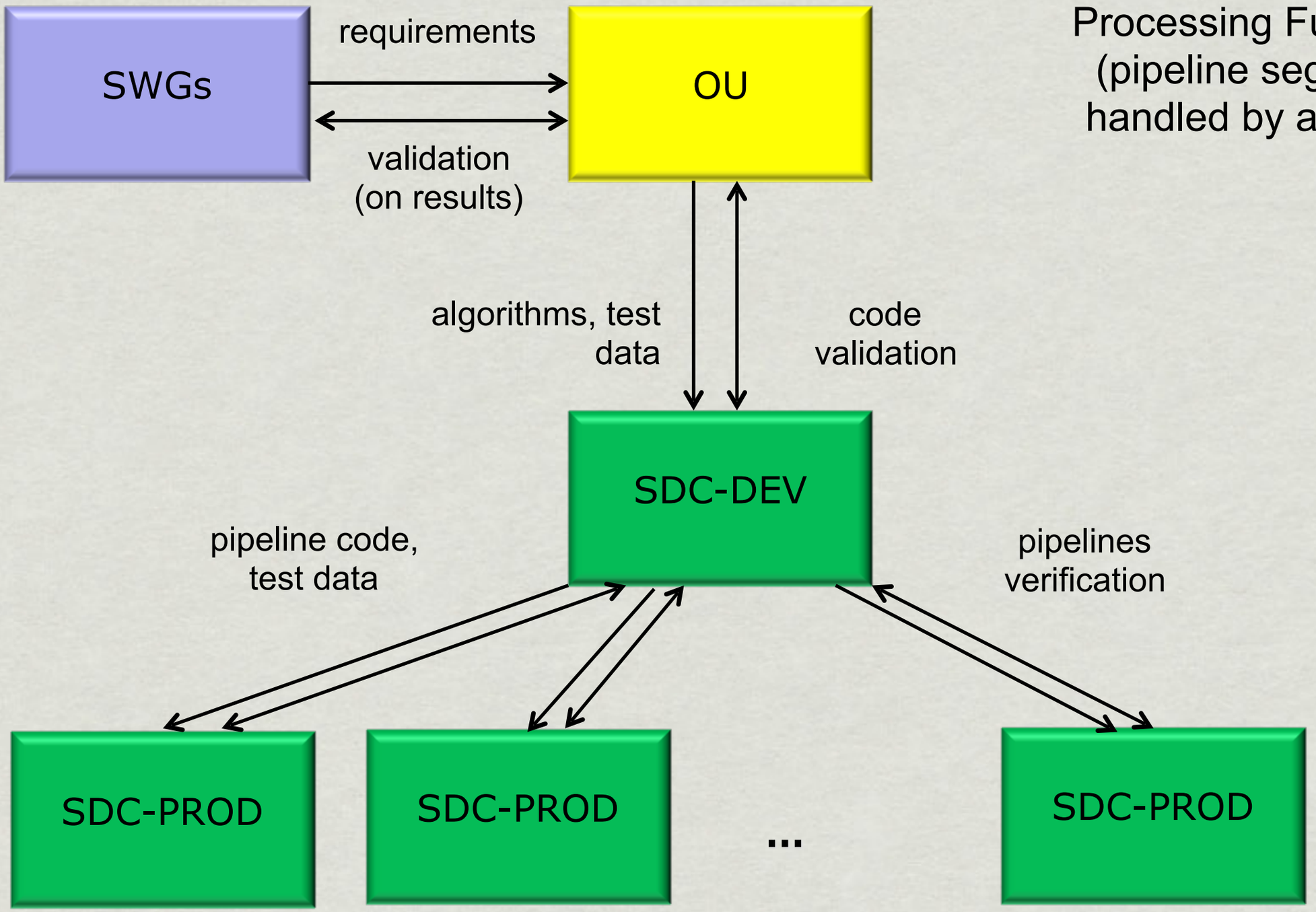
The SDCs are built around existing national computing facilities, they gather IT support as well as developper expertise.
SDCs are in charge of the pipeline development (software & support architecture).
Production software will run in the SDCs.
The SDCs and SOC are the operational sites of the SGS.

The OUs group EC members according to their data processing expertise.
OUs are in charge of analyzing the science data processing requirements, and of producing the pre-integration version of the pipeline modules.
OUs will be in charge of future evolution of the pipeline during operations.

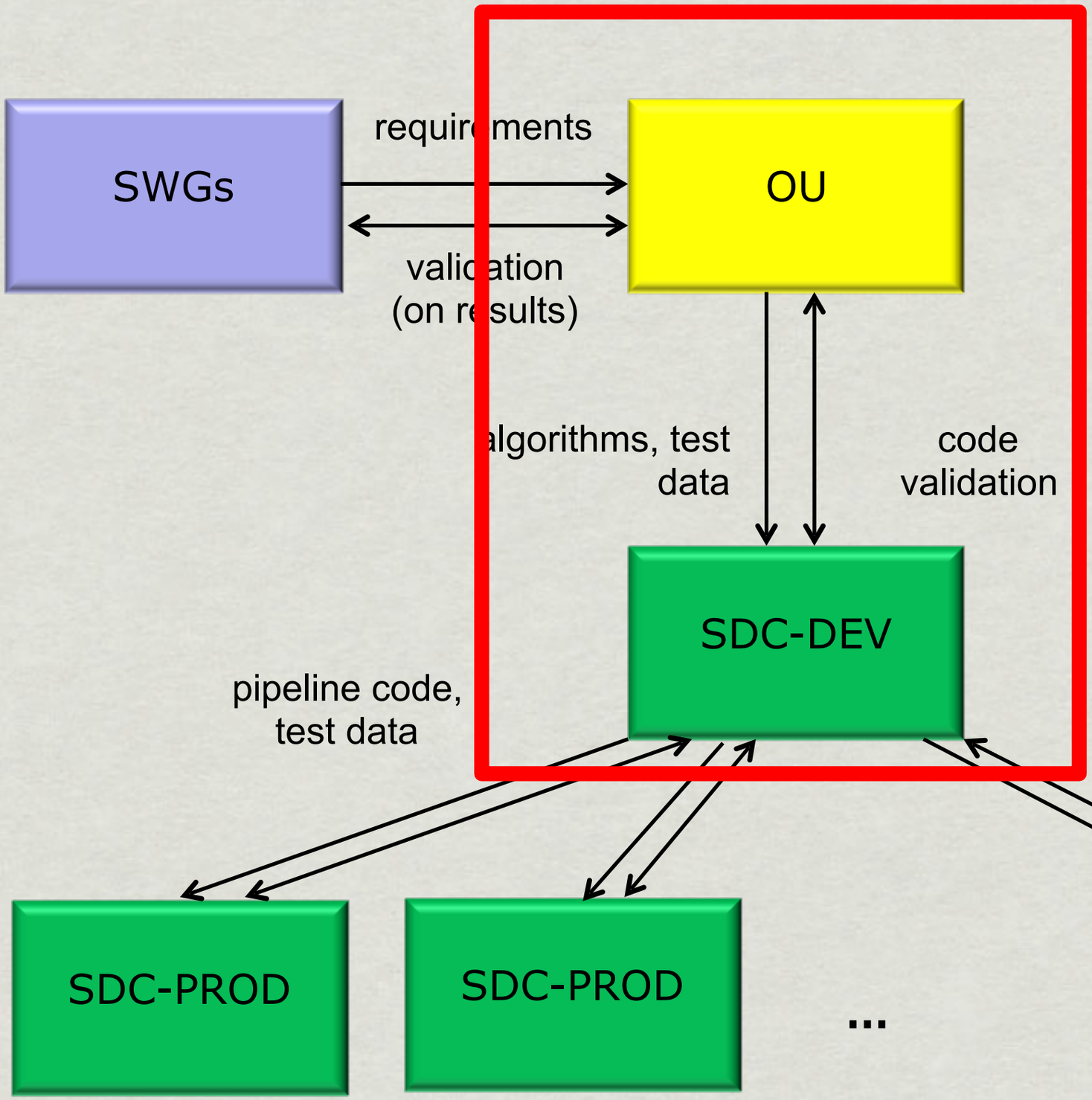
- **Notable group coming from the SDCs: the System Team in charge of building the software infrastructure.**

Development – Verification & Validation

For every Processing Function (pipeline segment handled by an OU)



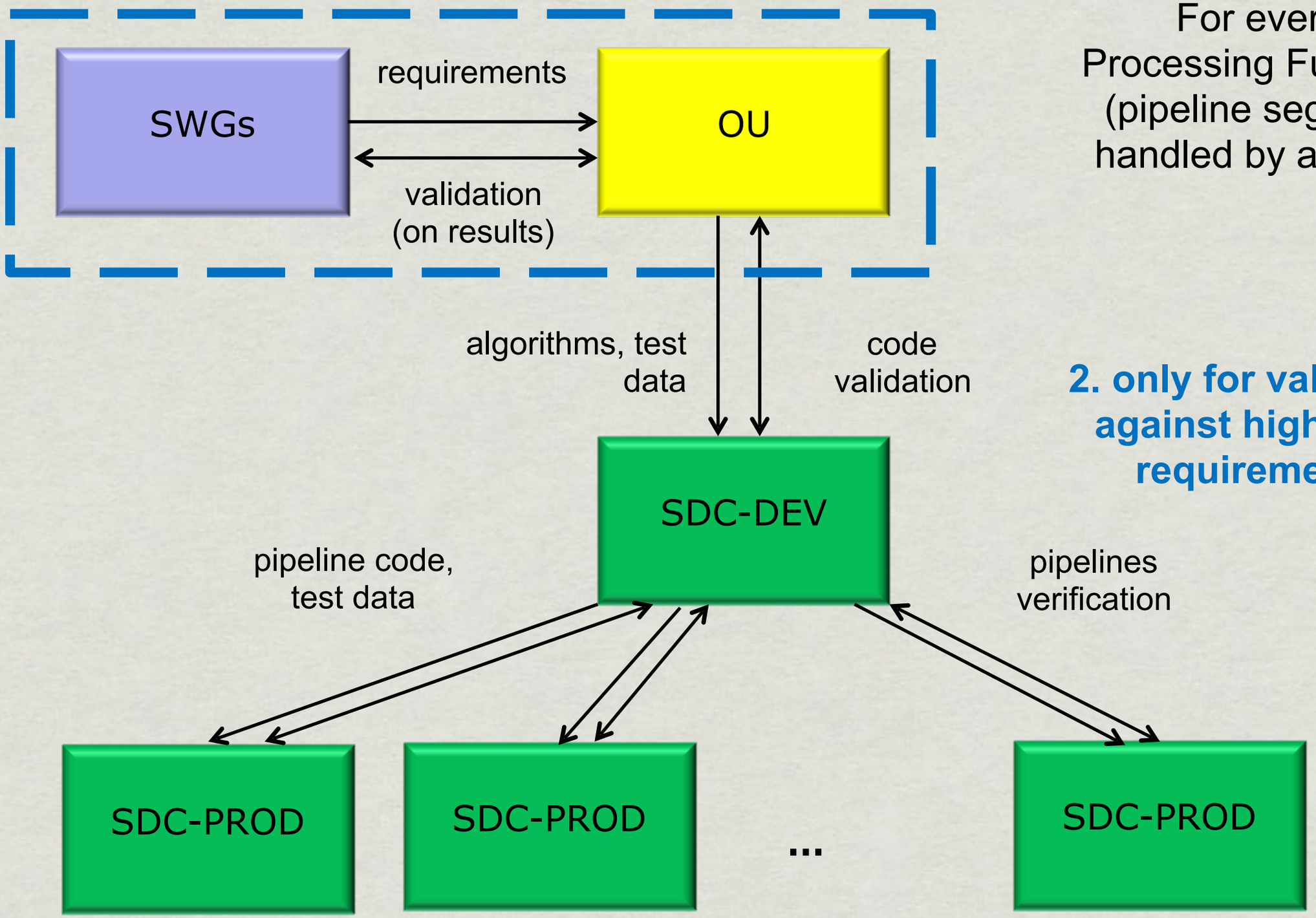
Development – Verification & Validation



For every Processing Function (pipeline segment handled by an OU)

1. in most cases joint development resting on code maturity roadmap

Development – Verification & Validation

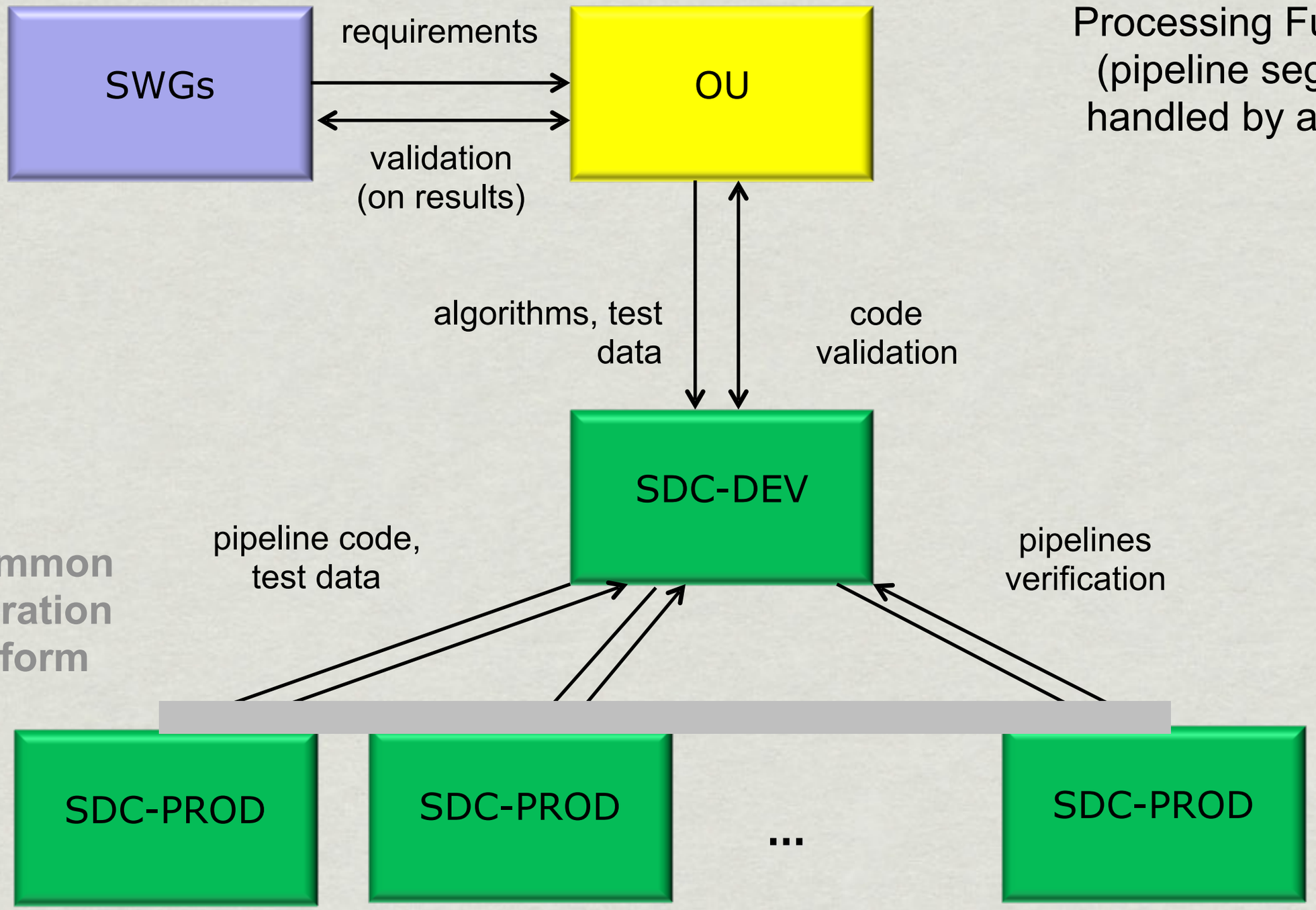


For every Processing Function (pipeline segment handled by an OU)

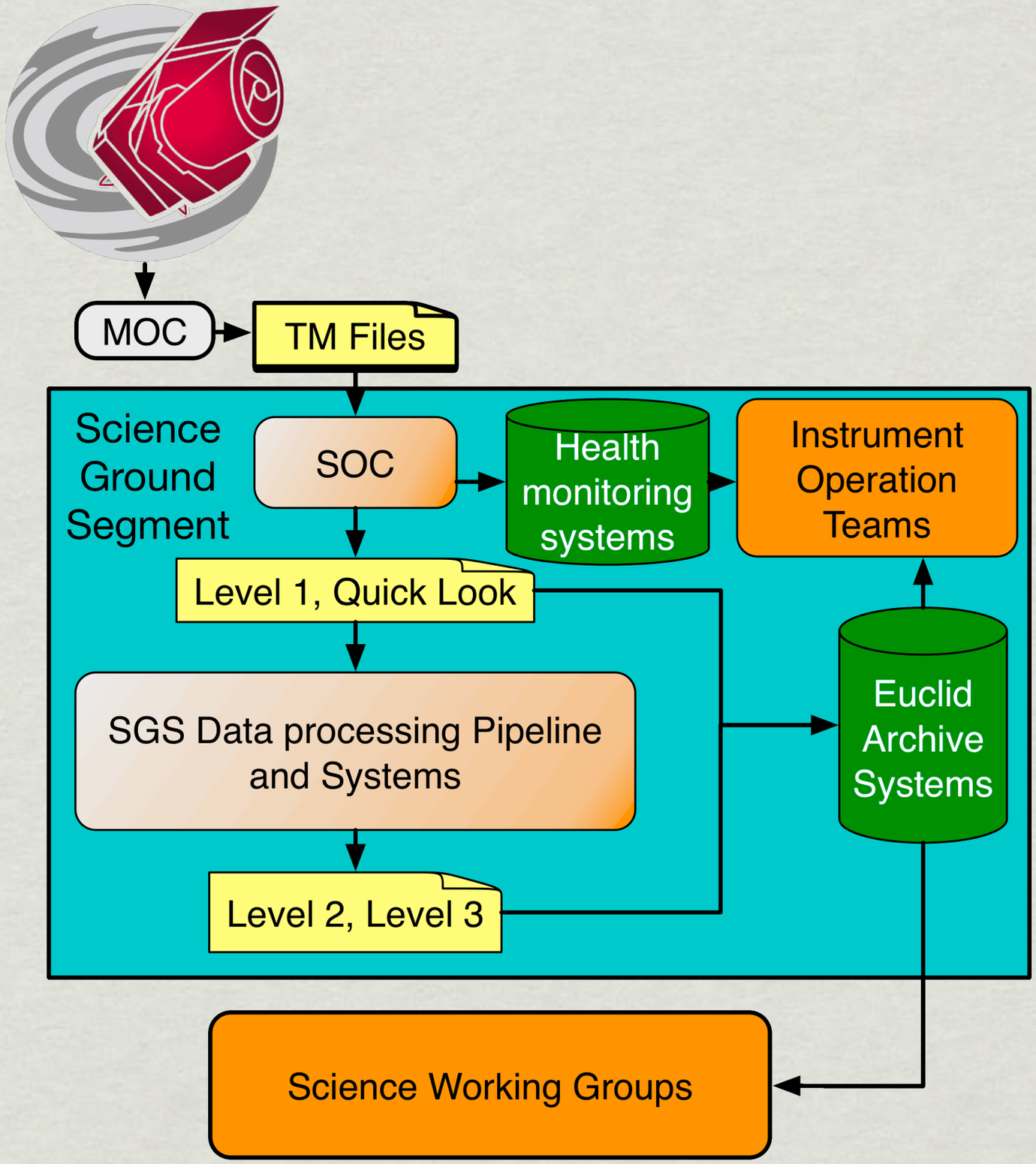
2. only for validation against high-level requirements

Development – Verification & Validation

For every Processing Function (pipeline segment handled by an OU)



The SGS perimeter and operational role



- This is a picture of what the systems we are working on are aiming at for **Euclid operations**.
- In **grey**: systems performing operations, in **orange** teams performing actions and taking decisions.
  - SGS and SOC are both!
- Not represented here are feedback actions:
  - IOTs feed the SOC team with information on instrument health to take decisions on survey execution.
  - SWGs feed the SGS with diagnostics on data science quality for pipeline improvement.
  - SOC feeds MOC with survey planning request (including re-scheduling of observation).
- **EAS is in fact an active system**, it provides data management and a transfer system for the SGS.

Euclid pipeline software development

Config. Mgmt

Euclid SVN

Ext. repos *(O/S, EPEL)*

Development Environment

Local Development

Central Continuous Integration

LODEEN

CODEEN

Euclid Pipeline Software Stack

3rd Party Software

PROCESSING FUNCTIONS

Common Scientific libs

Standard Technical and Scientific libraries

Core libs

O/S Kernel

EDEN referential

Distribution Systems

YUM

Delivery

CernVM-FS

App. Repository

Syst. Repo.

Docker Repository

Deployment

Different types of Processing Nodes

µCVM

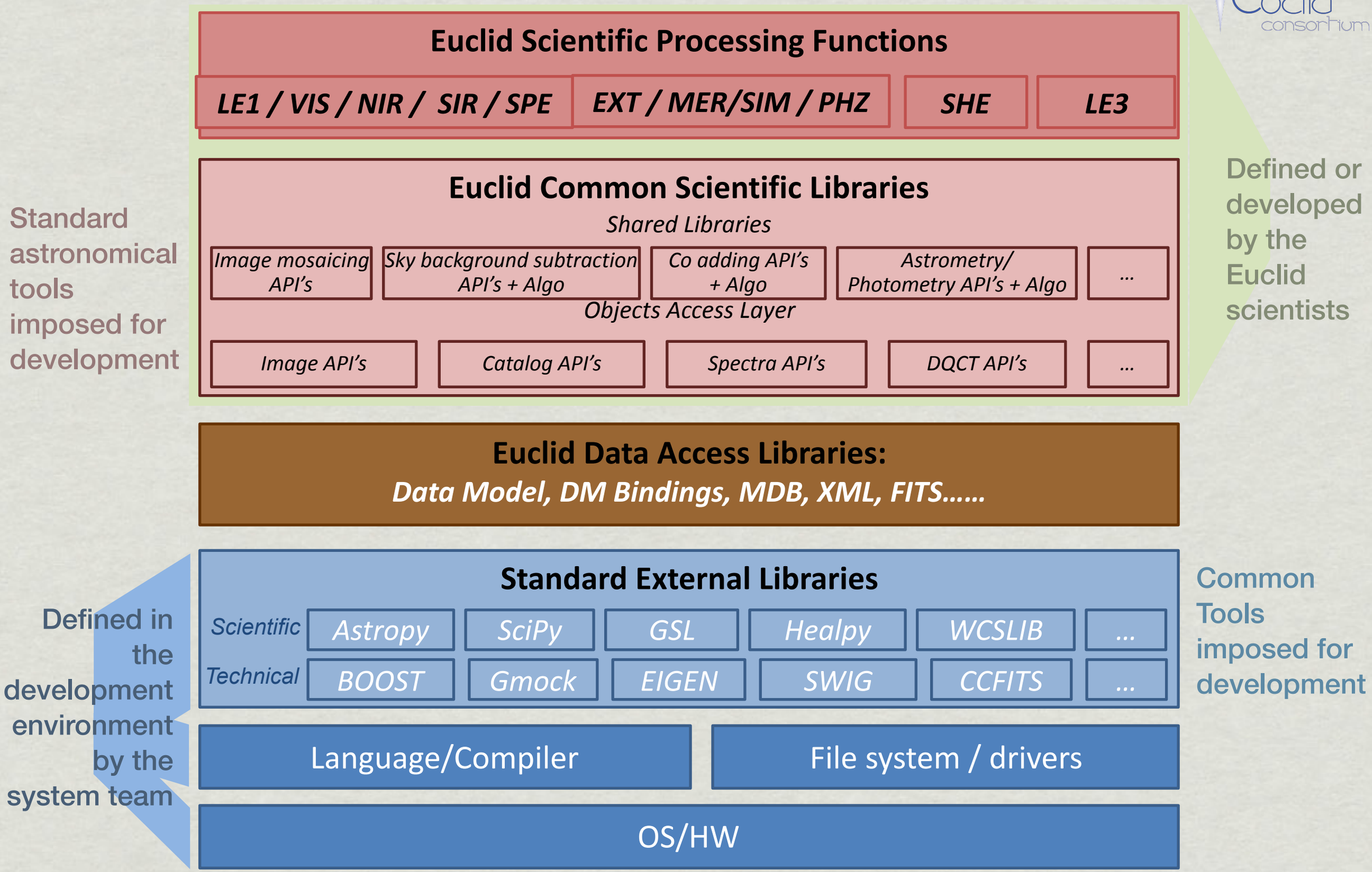
Euclid VM

Docker container

Physical PN

From Development to Production

Euclid Pipeline Software Stack



Standard astronomical tools imposed for development

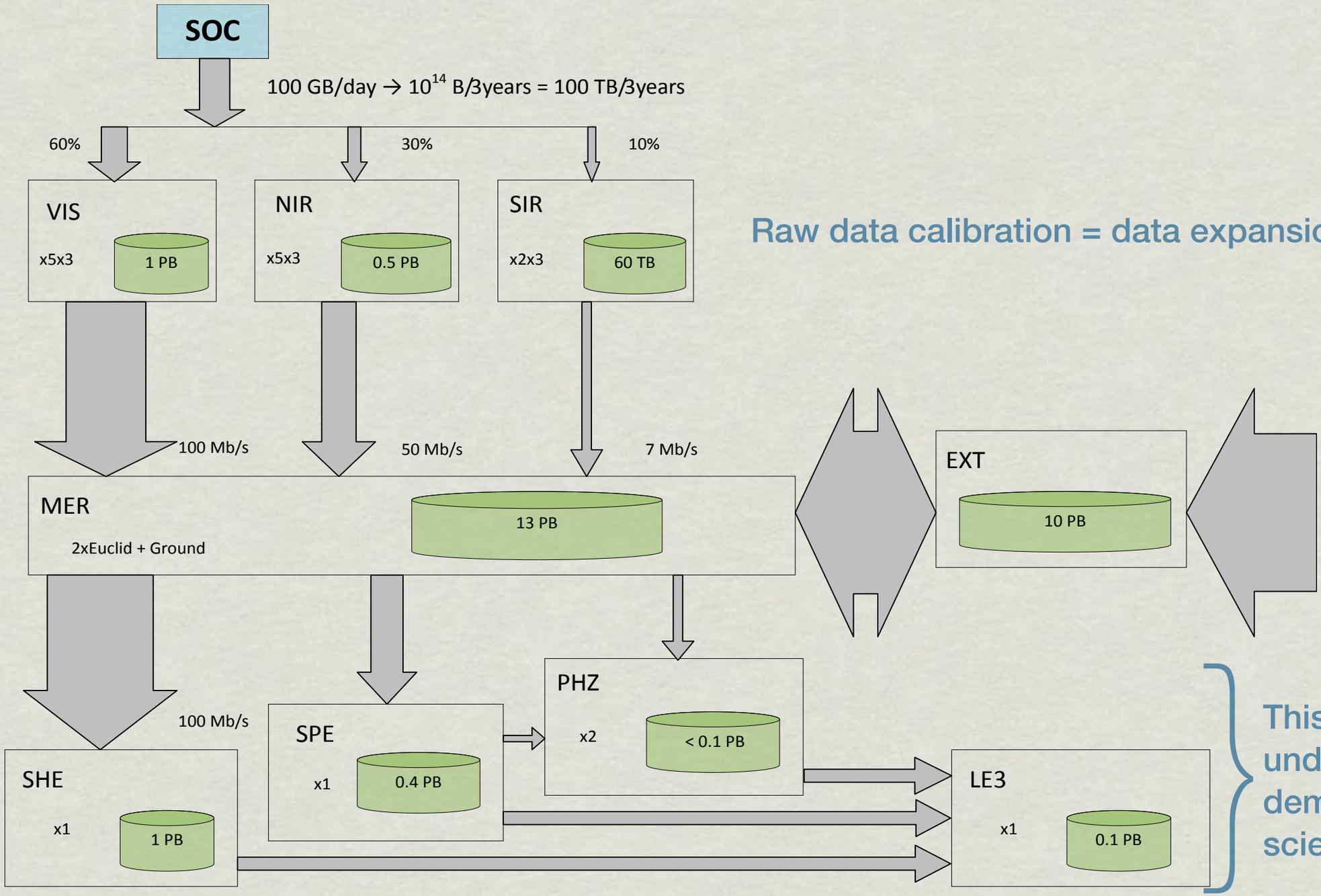
Defined or developed by the Euclid scientists

Common Tools imposed for development

Defined in the development environment by the system team

Orchestration of the data processing

Data flow estimation

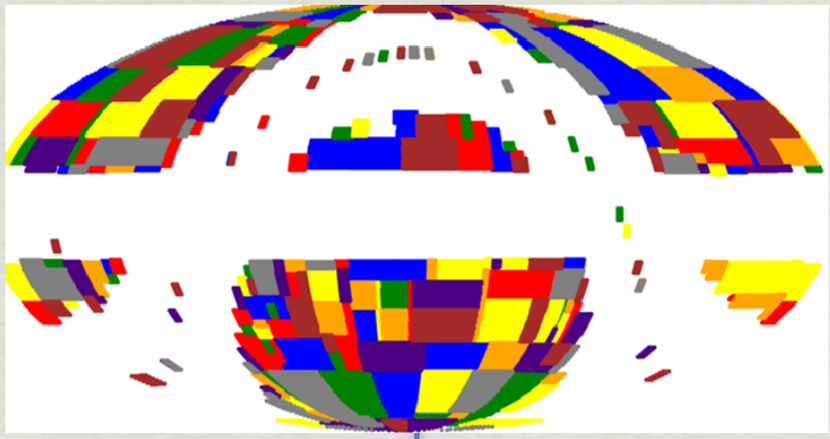


Raw data calibration = data expansion

Merging of all data source and catalog creation

This layer could be underestimated as demand from scientists increases

- "Brute force" SDC to SDC bulk data transfer is inefficient: we need to move the processing software, not the data.
- "One size does not fit all": part of the processing can be split by sky regions, but part of it require access to the whole sky (but not to every bit of data)

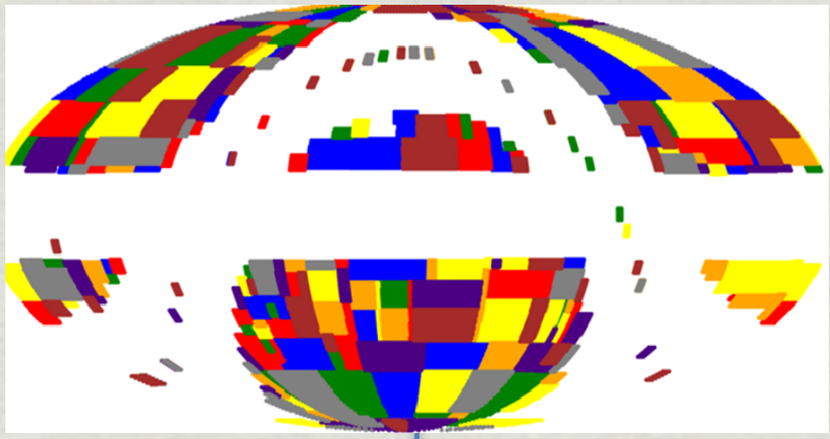


Mission Operations Centre

External Data Providers

Possible scenario: per-determined SDC-Sky allocation

Science Operations Centre	SDC-NL	SDC-DE	SDC-FR	SDC-CH	SDC-ES	SDC-US	SDC-UK	SDC-FI
Public Data	Raw EXT Data	Raw EXT Data	Raw EXT Data (TBC)					
Level 1 Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files
Metadata (prime)	Metadata (backup)							



Mission Operations Centre

External Data Providers

Possible scenario: per-determined SDC-Sky allocation

Science Operations Centre	SDC-NL	SDC-DE	SDC-FR	SDC-CH	SDC-ES	SDC-US	SDC-UK	SDC-FI
Public Data	Raw EXT Data	Raw EXT Data	Raw EXT Data (TBC)					
Level 1 Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files
Metadata (prime)	Metadata (backup)							

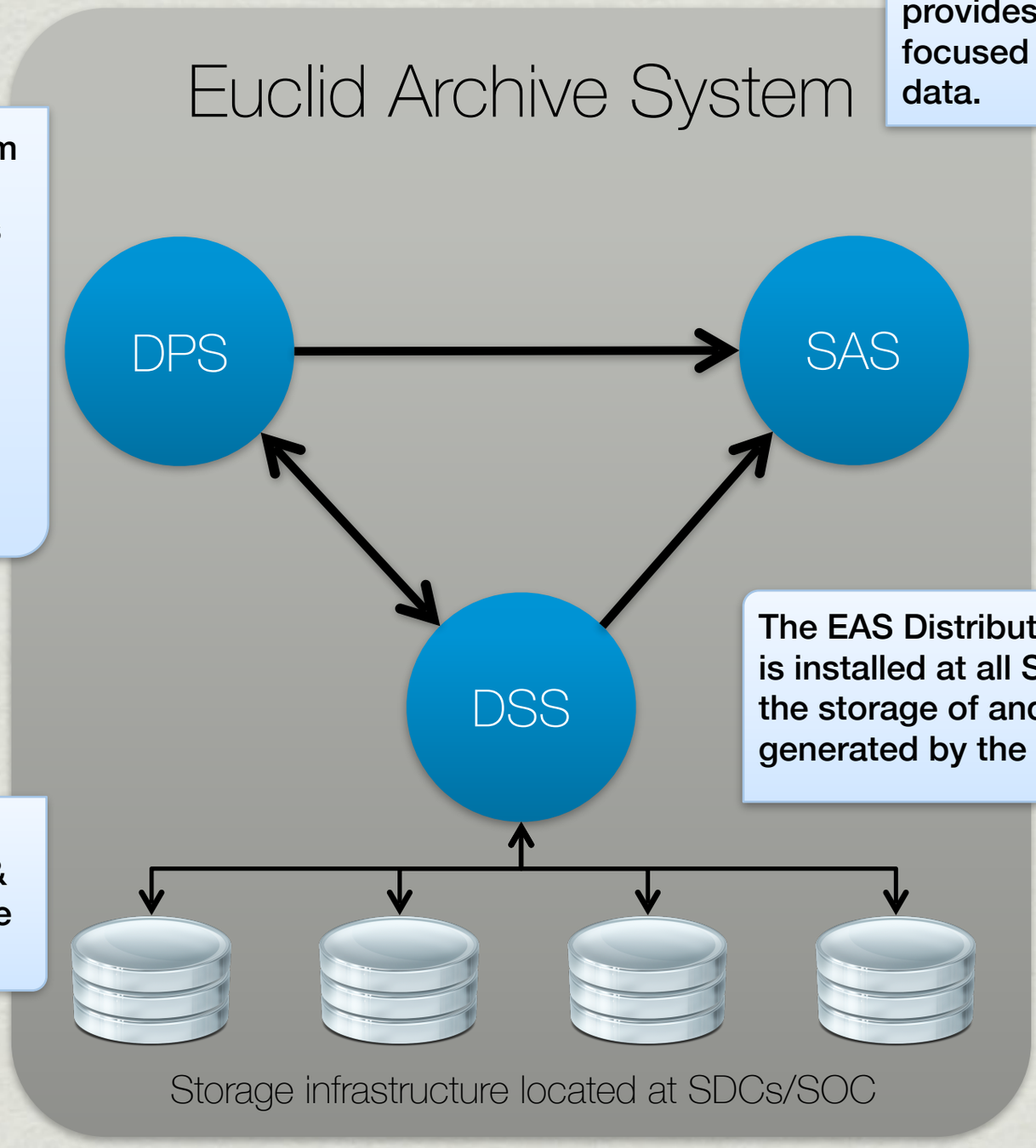
Redundant storage of the data

Archive overview

The EAS Science Archive System (SAS) provides access to the Euclid data focused on the scientific use of the data.

The EAS Data Processing System (DPS) stores the data product metadata including the locations of the data files.

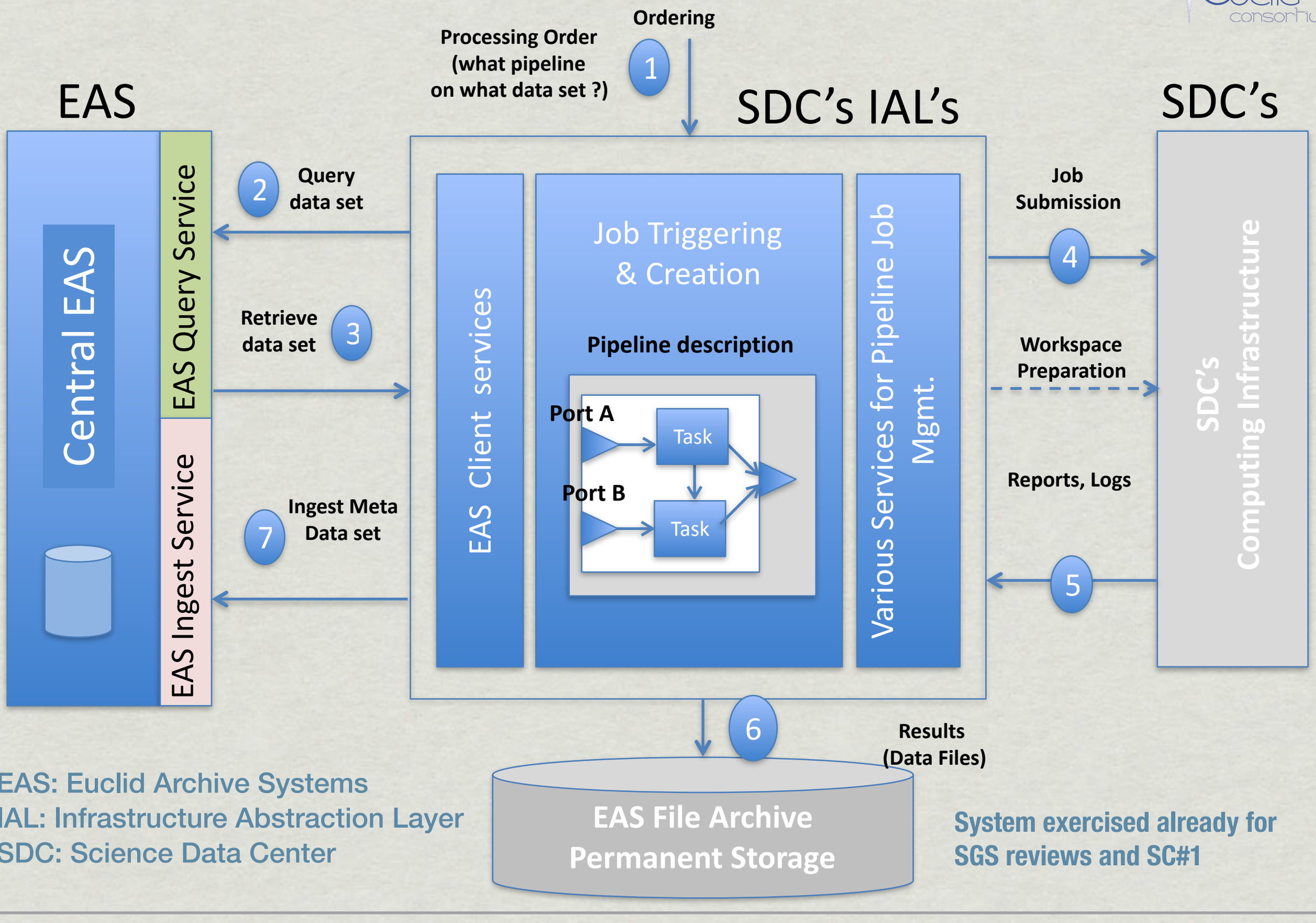
It provides access to the data products to the EC members, including processing coordination, quality control and processing history tools.



The EAS Distributed Storage System (DSS) is installed at all SDCs & SOC and manages the storage of and access to the data files generated by the pipelines.

Storage systems are provided by all the SDCs & SOC. All public data will be located at ESAC.

Data processing mechanism



EAS: Euclid Archive Systems
IAL: Infrastructure Abstraction Layer
SDC: Science Data Center

System exercised already for SGS reviews and SC#1

Conclusions (at least for now)

- The Euclid SGS has essentially passed its reality-based assessment (i.e. move from paper concepts to ones that start to be implemented).
- The SGS has now the elements (systems) to work with a distributed heterogenous infrastructure, as demonstrated in actual runs for reviews and internal challenges.
- Most elements of the pipeline will have a running prototype of significant maturity in 2017.

Thank you!