Science Ground Segment data processing

Insights from Euclid and LISA Missions

A.Fumagalli / D.Tavagnacco on behalf of "gruppo spazio" @OATs

Science Ground Segment role

Define, Organize and Maintain the data analysis infrastructure software and hardware

Manage and monitor the instrument and the sky survey progression

Process raw data into final science products

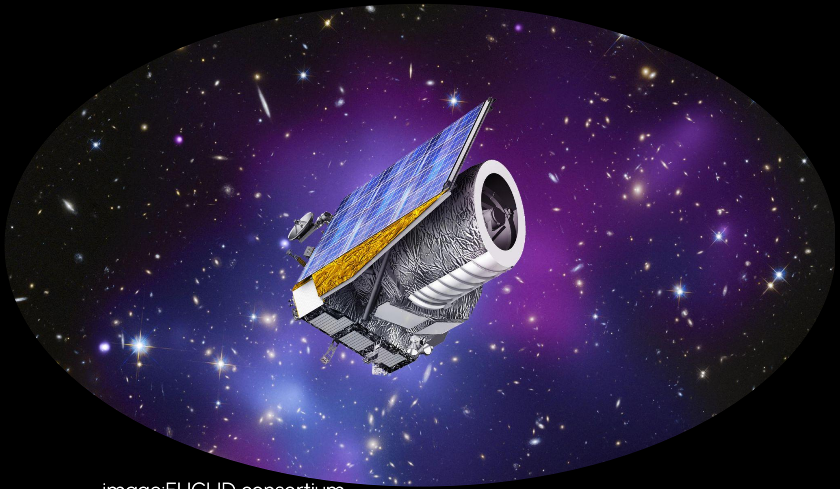


image:EUCLID consortium

1 Luglio 2023

Euclid

map of the large-scale structure of the Universe → Imaging

Euclid SGS as a distributed system

Science Data Centers (SDC)

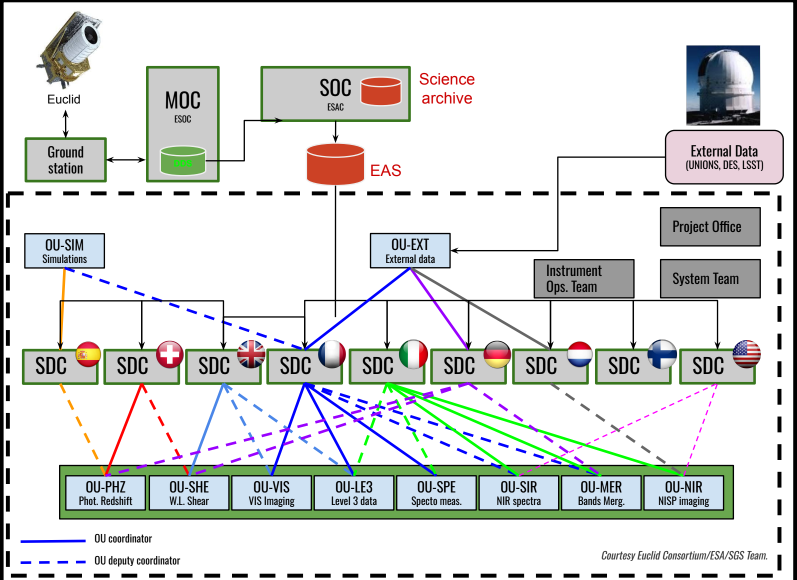
- national HPC facilities (run)
- develop. expertise (integration)

Organization Units (OU)

- processing definition (algorithms)
- science expertise (requirements)

SDC + OU

- pipelines development+test+integration+maintenance



Courtesy Euclid Consortium/ESA/SGS Team.

... 9 SDCs... 9 OUs... ~1500 persons

Euclid SGS constraints

Large Data Volume

- ❖ 500+ k raw images + external data

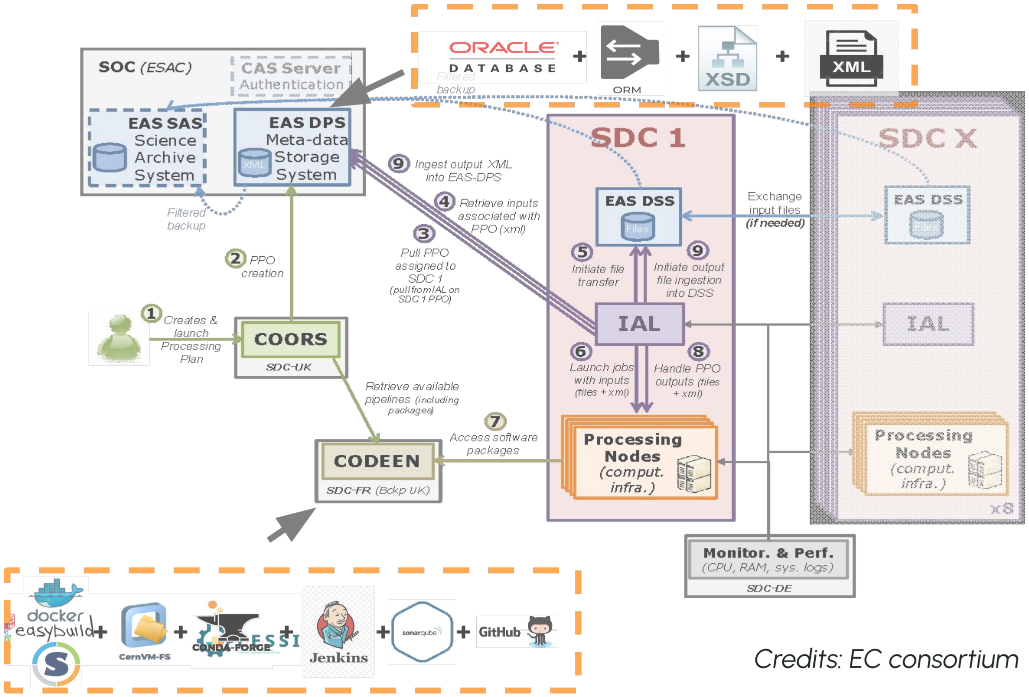
Varied dataset on the processing flow (type and granularity)

- ❖ images ..to.. catalogs ..to.. science data

Continuously evolving Software

- ❖ more data → more instrument knowledge → better algorithms

Euclid SGS Data circulation


ORACLE DATABASE
+
ORM
+
XSD
+
XML
Filtered backup
SOC (ESAC)
CAS Server Authentication
EAS SAS Science Archive System
EAS DPS Meta-data Storage System
XML
Filtered backup
SDC 1
SDC X
9 Ingest output XML into EAS-DPS
4 Retrieve inputs associated with PPO (xml)
3 Pull PPO assigned to SDC 1 (pull from IAL on SDC 1 PPO)
2 PPO creation
1 Creates & launch Processing Plan
COORS
SDC-UK
EAS DSS
Files
Exchange input files (if needed)
EAS DSS
Files
5 Initiate file transfer
9 Initiate output file ingestion into DSS
IAL
IAL
6 Launch jobs with inputs (files + xml)
8 Handle PPO outputs (files + xml)
Retrieve available pipelines (including packages)
7 Access software packages
CODEEN
SDC-FR (Bckp UK)
Processing Nodes (comput. infra.)
Processing Nodes (comput. infra.)
x8
Monitor. & Perf. (CPU, RAM, sys. logs)
SDC-DE
docker
easybuild
+
CernVM-FS
+
CONDA-FORGE
EESSI
+
Jenkins
+
sonarqube
+
GitHub

Credits: EC consortium

EAS - Distributed Storage System (DSS)

Custom Object Storage, SDC

http data transfer -transparent

data: mainly FITS and HDF5 formats

EAS - Data Processing System

central metadata archive

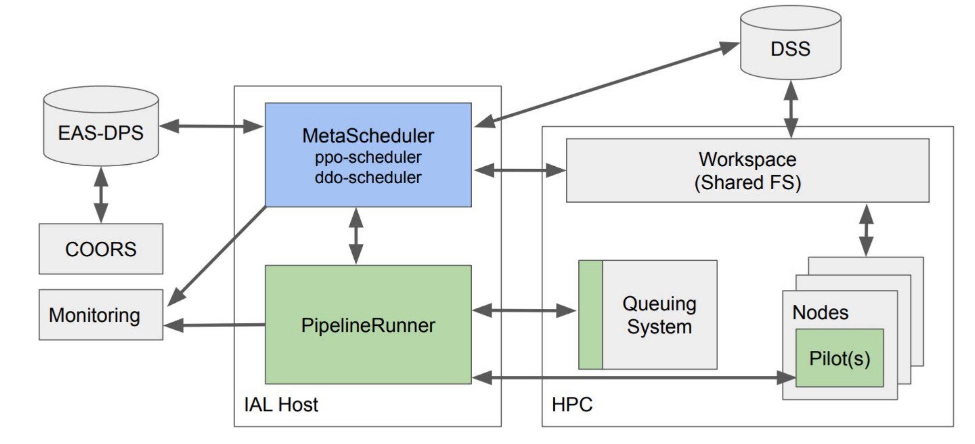
Object-to-relational mapping

Custom query language (REST)

data product oriented db, products as XML

Processing: Infrastructure Abstraction Layer

- uniform interface to HW and bridge over different queuing systems (national HPC facilities)
- common meta-scheduler
- Workflow manager with pipelines definitions as python scripts specifying relations and requirements of each step
- enforced PF I/O description + stateless processing
- Processing as "pilot jobs" payloads and profiling of jobs



Credits:M.Frailis - SAIT 2025

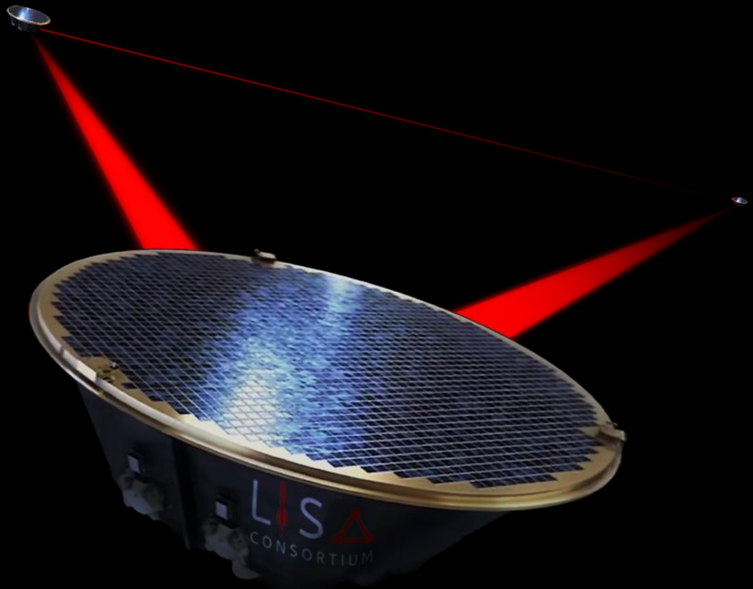


image:LISA consortium

2035

LISA

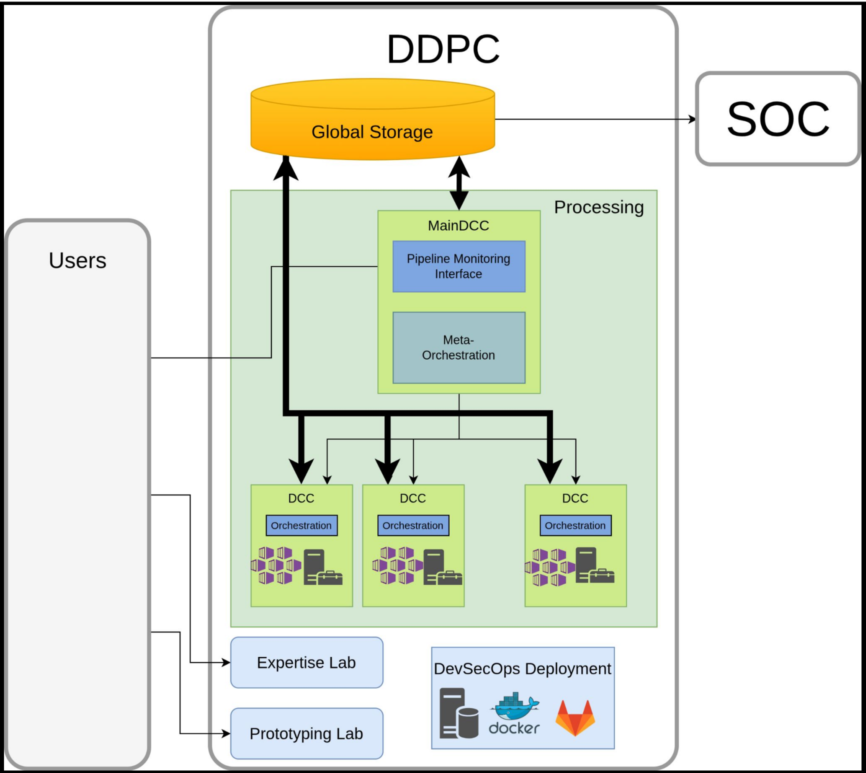
large scale GW detector → Timelines

LISA SGS as distributed system

Distributed
Data Processing Center

User interaction

- Main DDPC
  Operations
- Expertise Lab
  monitoring
  create&execute pipelines
- Prototype Lab
  develop code/processing
  simulation/challenges



LISA SGS constraints

Quick Processing

- for alerts “near-real-time” ~ 1h after data acquisition (LLAP pipeline) long before the Time-Delay Interferometry treams (L1)

Iterative processing pipeline

- data “signal-dominated”: more data → better parameter estimate (Global FIT)

Sensitivity requirementr

- ~ fm/(s2 Hz) → Extreme control of systematics → simulations

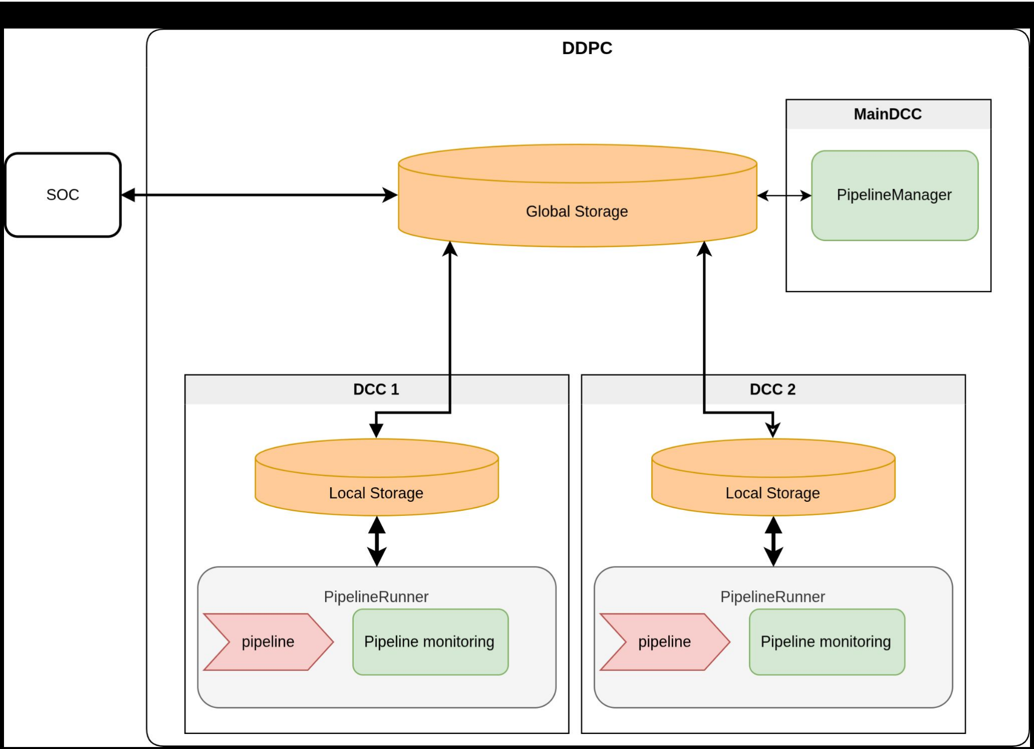
Data Circulation

Centralized Storage System

- Data and Metadata
- Based on NextCloud
- Object Storage
- data HDF5 mainly

Data Processing System

- local storage
- data products as JSON



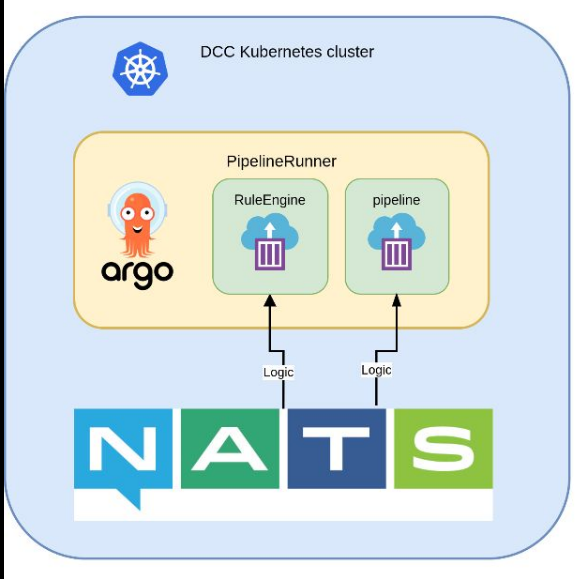
Processing: k8s + cloud

Distributed Data Centers

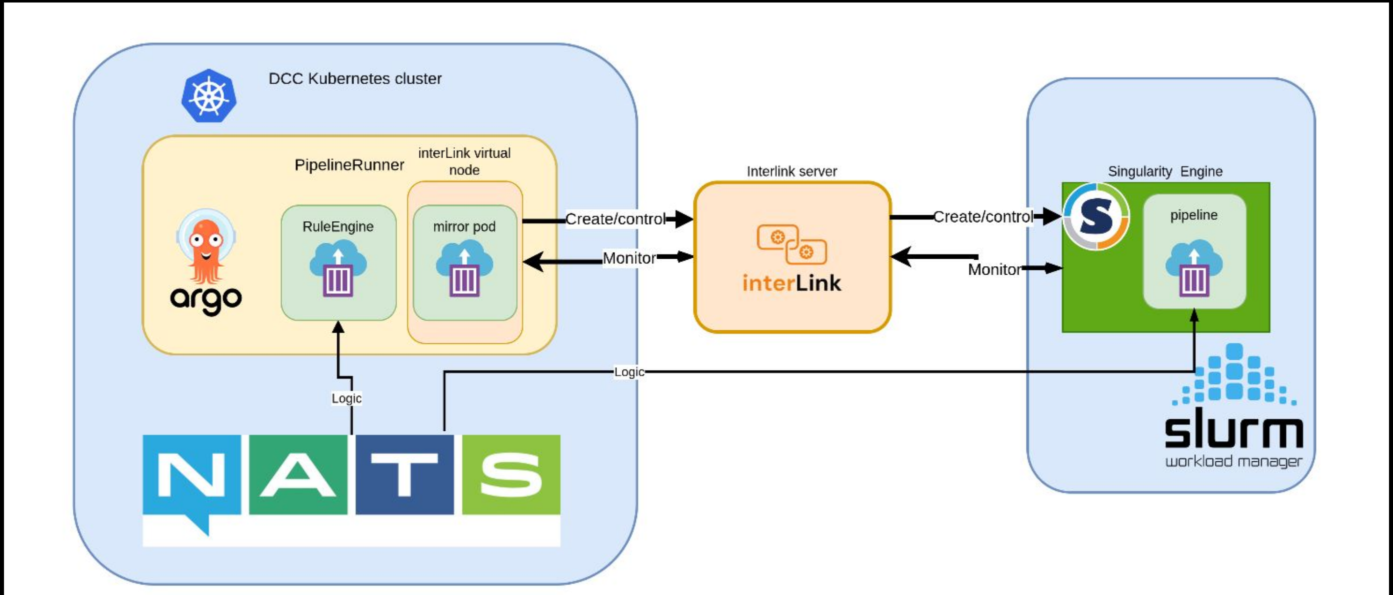
- k8s custer
- pipeline ad k8s pods
- argo orchestrator
- logic through rule engine
- NATS messaging system

Cloud approach

- global storage based on Nextcloud
- possibility to offload on slurm cluster



Processing: K8s + slurm



“AI” within the SGS

+ interfaces

Humanware

Management

SW development

Learning curve

Software

DevOps environment

Workflow management

Algorithms

Hardware

HPC/HTC infrastructure

Storage, archive

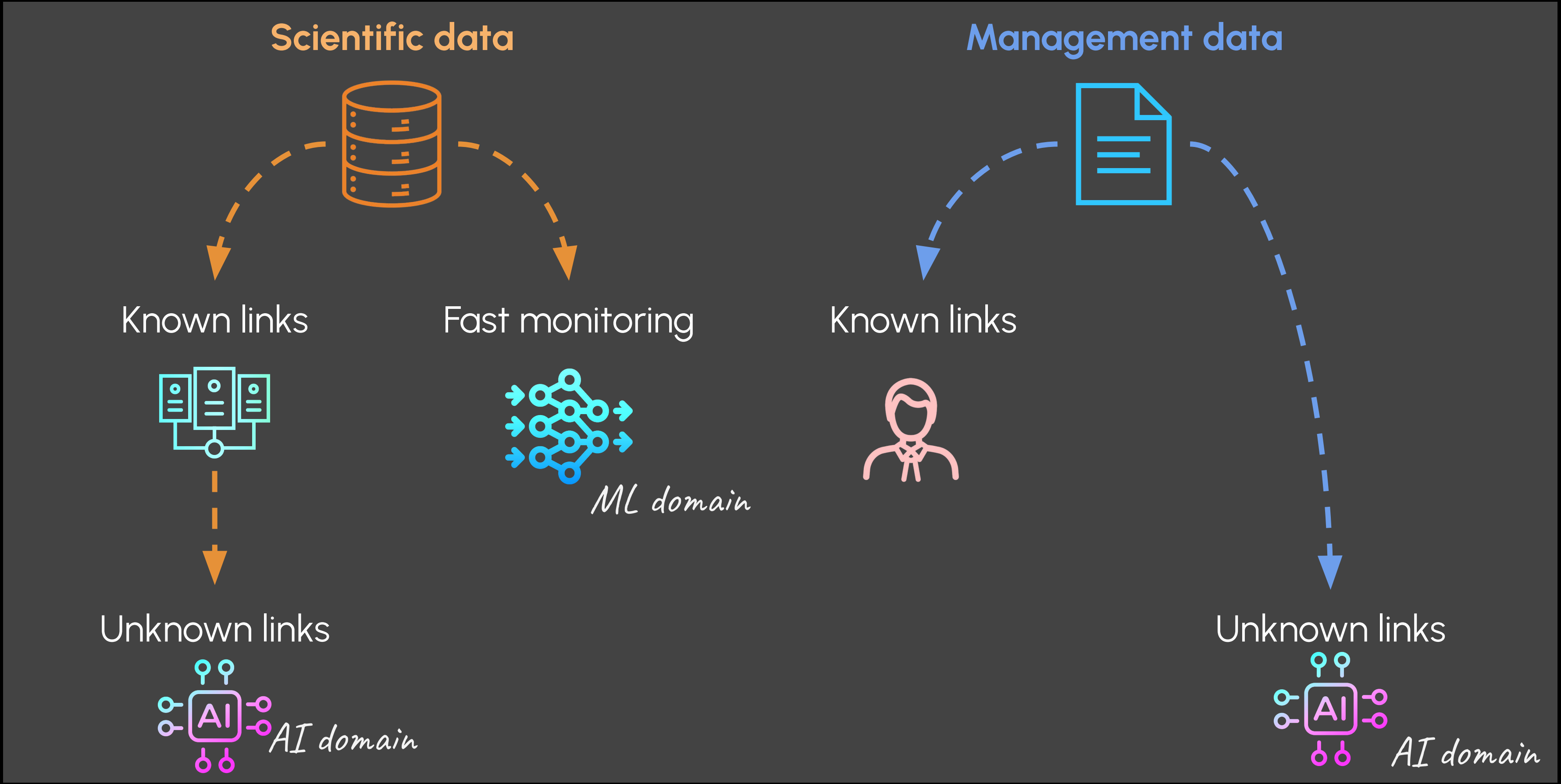
Technology

Scientific data:

- Structured
- SW+HW processing
- Interface: machine-to-machine

Management data:

- Unstructured
- Human processing
- Interface: machine-to-human



Scientific data

Known links

Fast monitoring

ML domain

Unknown links

AI domain

Management data

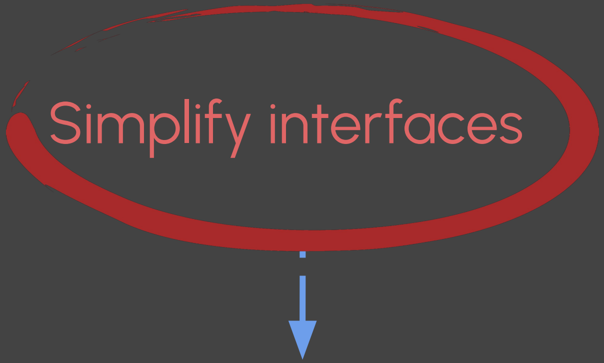
Known links

Unknown links

AI domain

Credits: Romelli, Tavagnacco, Gasparetto - Innovazioni e Tecnologie Emergenti nella Gestione dei Dati Scientifici per Missioni Spaziali

Focus: interfacce SW-Human


Simplify interfaces

SW-Human:

- Simplify user interfaces
- Operator assist
- Task automation
- Coding co-piloting

Applications:

- Repetitive tasks: doc, report, minutes
- Personalization
- Newcomers training

Benefits:

- Time saving
- Shallower learning curve
- Better team-work

Credits: Romelli, Tavagnacco, Gasparetto - Innovazioni e Tecnologie Emergenti nella Gestione dei Dati Scientifici per Missioni Spaziali

…few takeaway notes

Within a project, the SGS is one of the **longest-lived** key elements

SGS and project evolution are entangled and **technologies evolve** during SGS lifecycle

SGS activities require **various expertise** to cover all operational fields

Every SGS deal with different datasets: identify their nature and processing tool is crucial

Future projects will **collect and analyze larger datasets** using distributed and complex infrastructures

AI tools are powerful and **can improve several areas in the SGS**, and SGS data analysis but it's crucial to identify the nature of the dataset to apply the correct tool

Main focus is Interface simplification, large dataset handling, efficiency and flexibility

The team

Marco Frailis, **Alessandra Fumagalli**, Samuele Galeotta, **Roberta Giusteri**, **Marius Lepinzan**, Gianmarco Maggio, Oriana Mansutti, **Erik Romelli**, **Federico Rizzo**, Daniele Tavagnacco, **Antonino Troja**, **Thomas Vassallo**, Andrea Zacchei

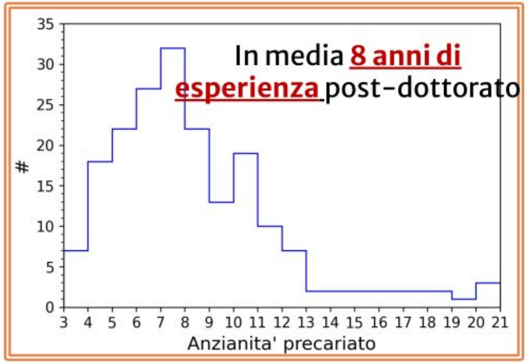
support:

Massimo Sponza, **Cristiano Urban**, Giuliano Taffoni, Claudio Vuerli

-- fixed term personnel

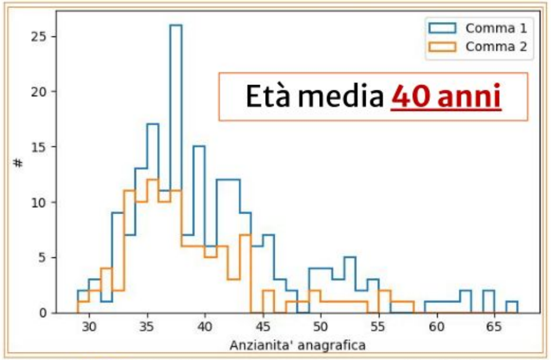
La situazione del personale precario in INAF è **INSOSTENIBILE!**

1.200 Tempo Indeterminato Vs **650** precari: più di 1 precario ogni 2 persone di ruolo



In media **8 anni di esperienza** post-dottorato

Plot di un campione rappresentativo dei precari INAF al 31/12/2024



Età media **40 anni**

Dei **650**, **287** possono essere stabilizzati:
173 tramite chiamata diretta (comma 1)
114 tramite concorsi riservati (comma 2)

Entro l'anno, l'attuale situazione determinerà l'esodo di > 100 lavoratori altamente qualificati e il MUR se ne lava le mani

È **URGENTE** che INAF **PROCEDA ORA** con le **STABILIZZAZIONI TRAMITE MADIA:** unica soluzione per questa emergenza

Molti colleghi (972) hanno già firmato, per sostenerci e aggiungere il nome alla lista del QR,
contattaci a retestabilizzandi1.inaf@gmail.com