# Data validation beyond Big Data

Edwin A. Valentijn

Kapteyn Astronomical Institute

6 June 2018 VST in the era of large sky surveys- Napoli

# STORY LINES

- processing/archiving/distribution:

  - AstroWISE- KiDs - Ou-Ext – Euclid

- data validation:

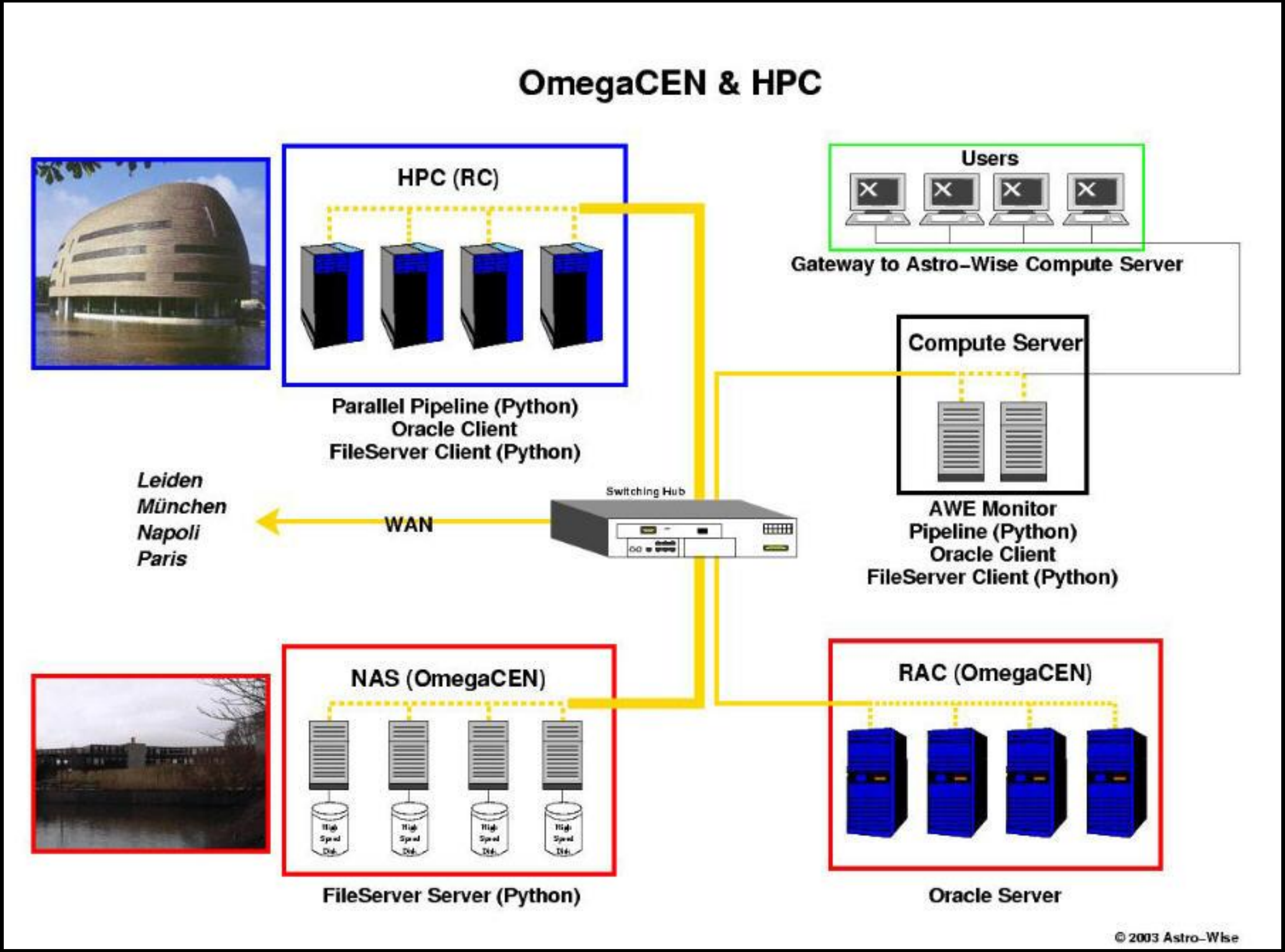
  - lineage - OU-Ext - Euclid- Facts and Fakes

Sequence of hypes:

GRID - Big Data - Machine learning -> data validation

# The Datacentric approach

## local networks and distributed



2003
RUG-CIT

# Astro-WISE – Data federations

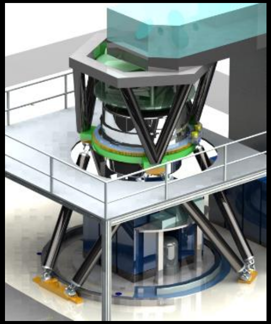
## Distributed Information Systems - handling surveys

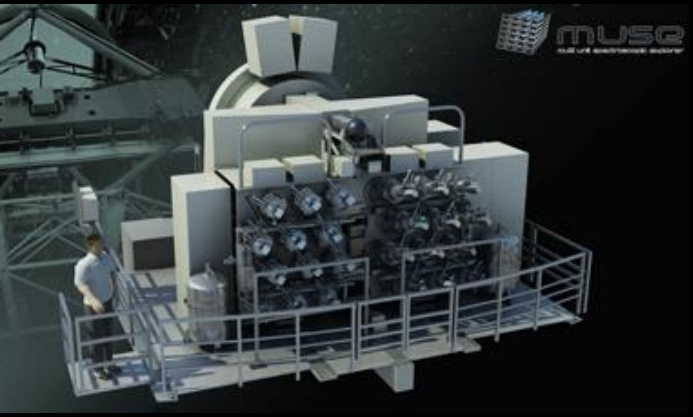
## since 2003 - it works

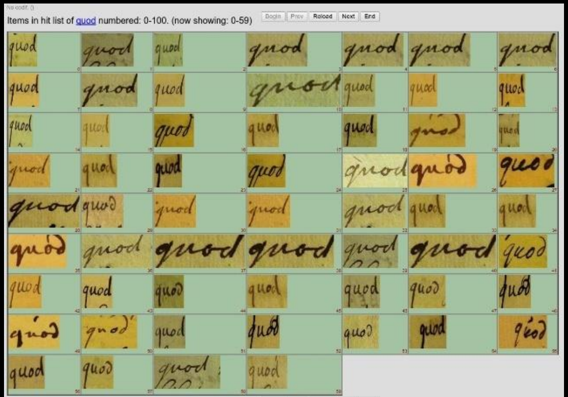
OmegaCEN@Kapteyn datacenter ~15-20 fte

KiDS - ESO – OmegaCAM@VST
MUSE - ESO - VLT
Lofar - LTA - Astron
Glimps - AI Handwritten text – Lifelines DNA
Target Holding

-> Euclid - ESA
-> Micado - ESO - ELT

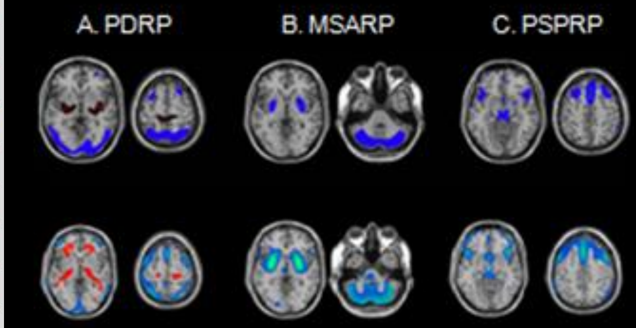






No codif. ()

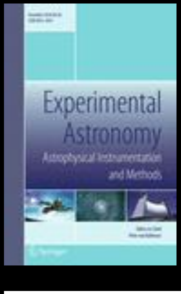
Items in hit list of quod numbered: 0-100. (now showing: 0-59)



# all published

http://www.astro-wise.org Manuals & tutorials

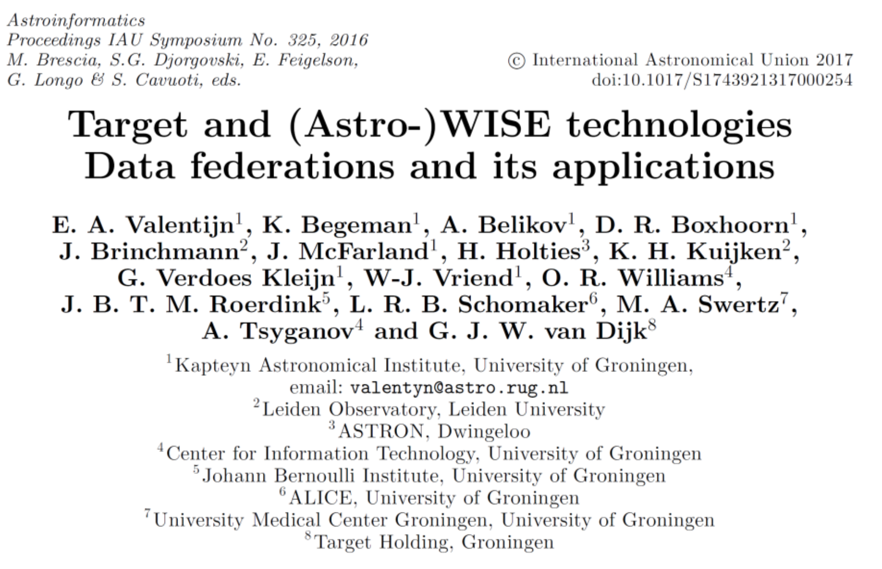
http://www.rug.nl/target Target Consortium



Experimental Astronomy Astrophysical Instrumentation and Methods

Experimental Astronomy - Vol. 35, 2013

All papers are online



*Astroinformatics*
*Proceedings IAU Symposium No. 325, 2016*
*M. Brescia, S.G. Djorgovski, E. Feigelson,*
*G. Longo & S. Cavuoti, eds.*

© International Astronomical Union 2017
doi:10.1017/S1743921317000254

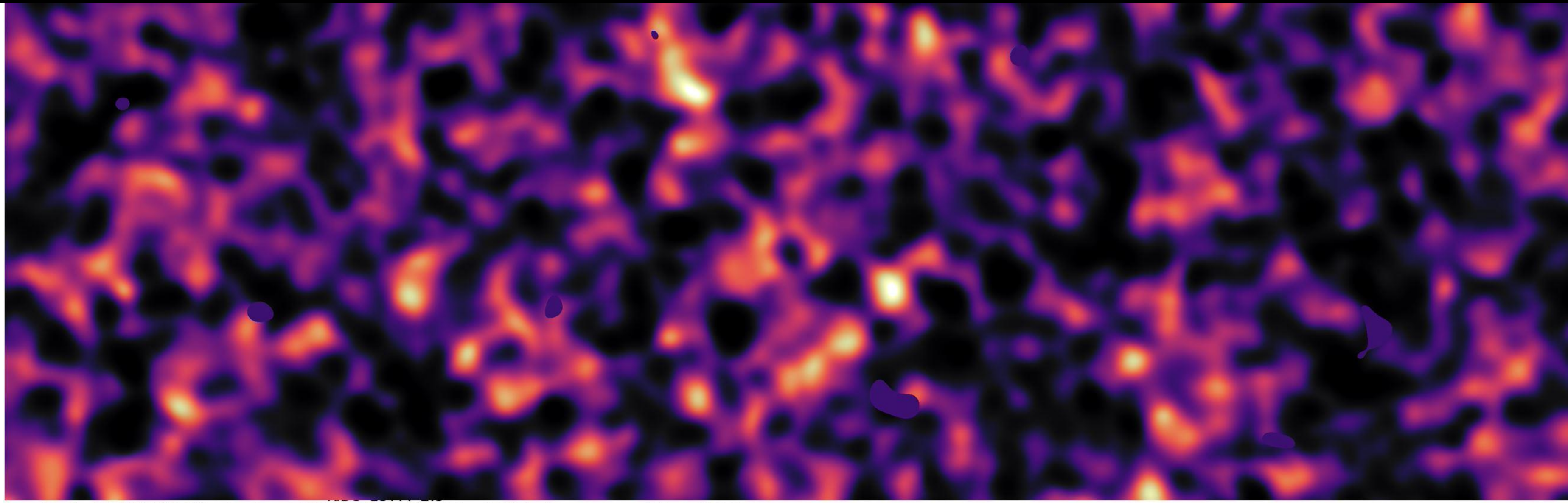
**Target and (Astro-)WISE technologies**
**Data federations and its applications**

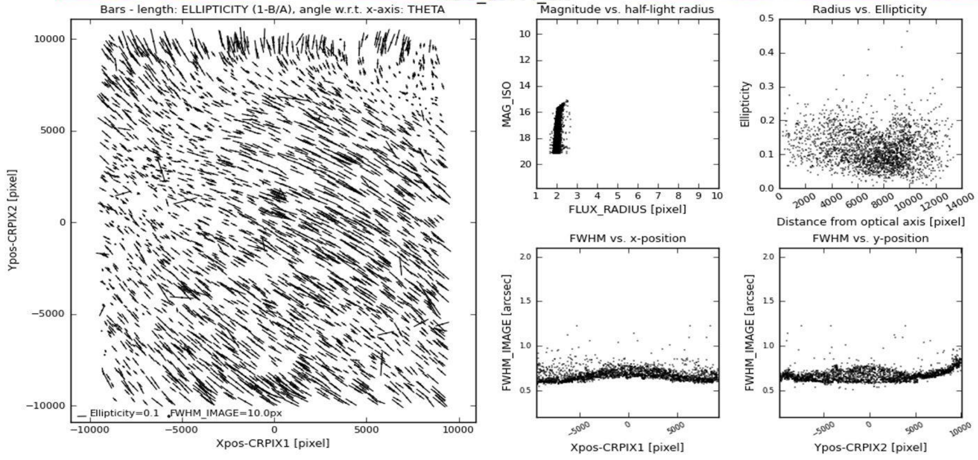
**E. A. Valentijn[1], K. Begeman[1], A. Belikov[1], D. R. Boxhoorn[1], J. Brinchmann[2], J. McFarland[1], H. Holties[3], K. H. Kuijken[2], G. Verdoes Kleijn[1], W-J. Vriend[1], O. R. Williams[4], J. B. T. M. Roerdink[5], L. R. B. Schomaker[6], M. A. Swertz[7], A. Tsyganov[4] and G. J. W. van Dijk[8]**

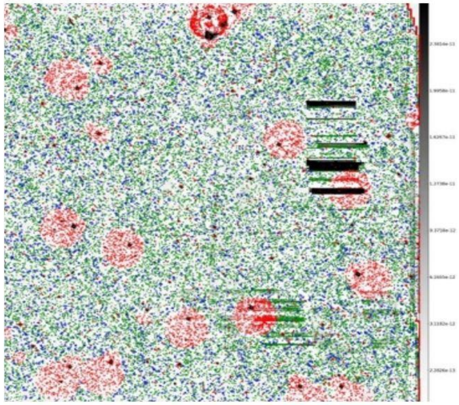
[1]Kapteyn Astronomical Institute, University of Groningen,
email: valentyn@astro.rug.nl
[2]Leiden Observatory, Leiden University
[3]ASTRON, Dwingeloo
[4]Center for Information Technology, University of Groningen
[5]Johann Bernoulli Institute, University of Groningen
[6]ALICE, University of Groningen
[7]University Medical Center Groningen, University of Groningen
[8]Target Holding, Groningen

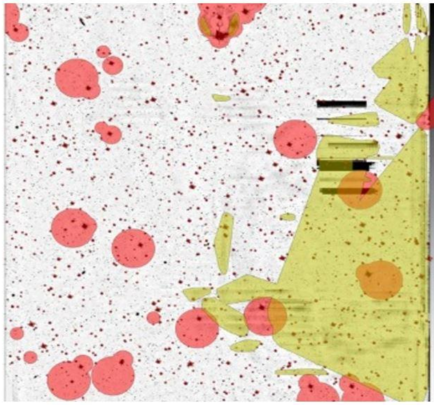
Astroinformatics 2016
IAU symposium 325
Datafederations
Valentijn et al. 2017

# KiDS Quality control DR1-DR2-DR3
# OmegaCAM@VST  740 sq deg









# Links as workhorse in data federations

*The Universe as a spreadsheet*

ERCIM News 2006

AstroWISE *Chaining to the Universe*

ADASS XVI ASP Conference Series,

15-18 October 2006 in Tucson, Arizona, USA.

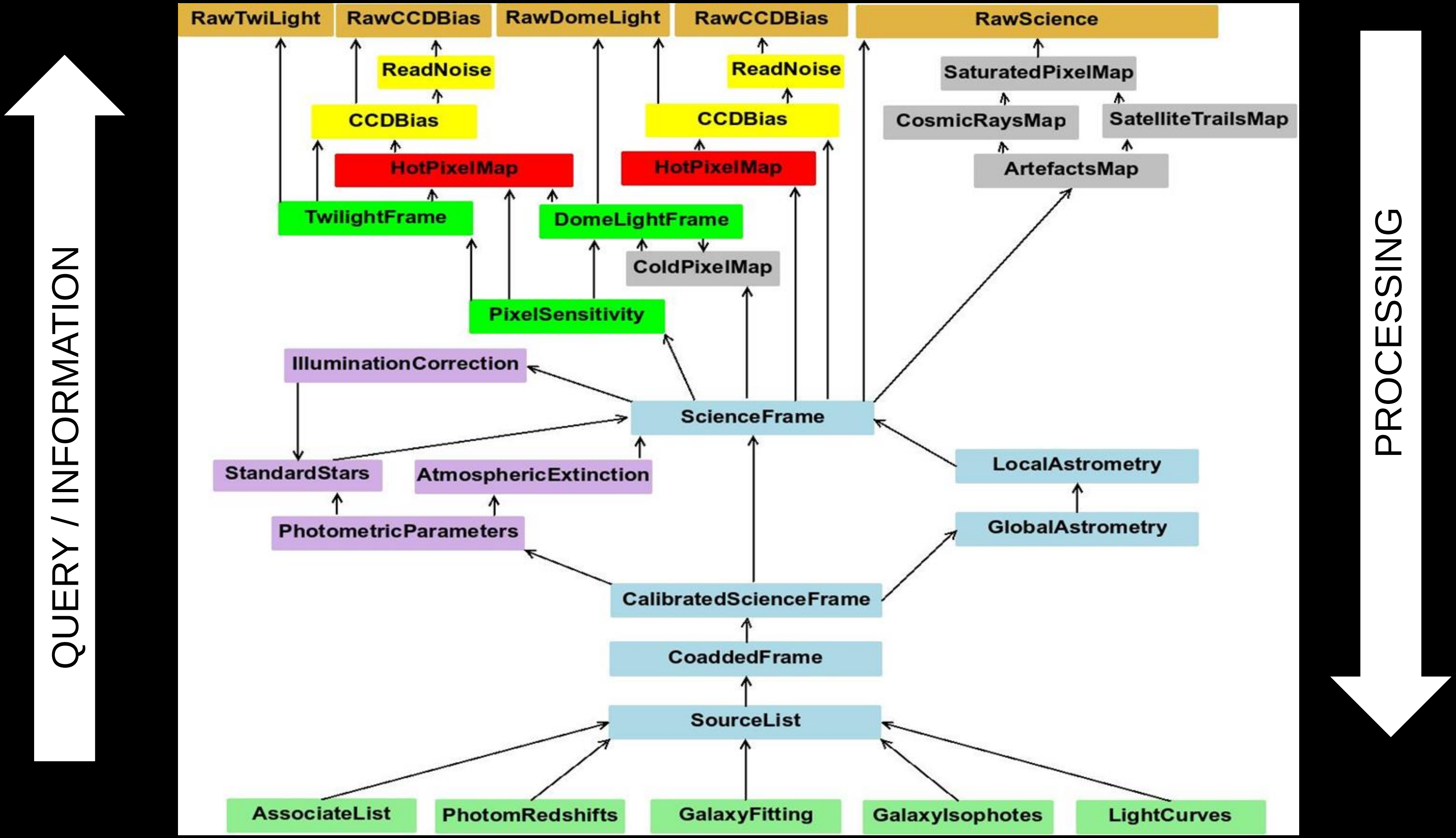
- Distributed Information Systems
  - Users, computers, storage
- Processing and Quality control
- Reproducable ( re-processing)

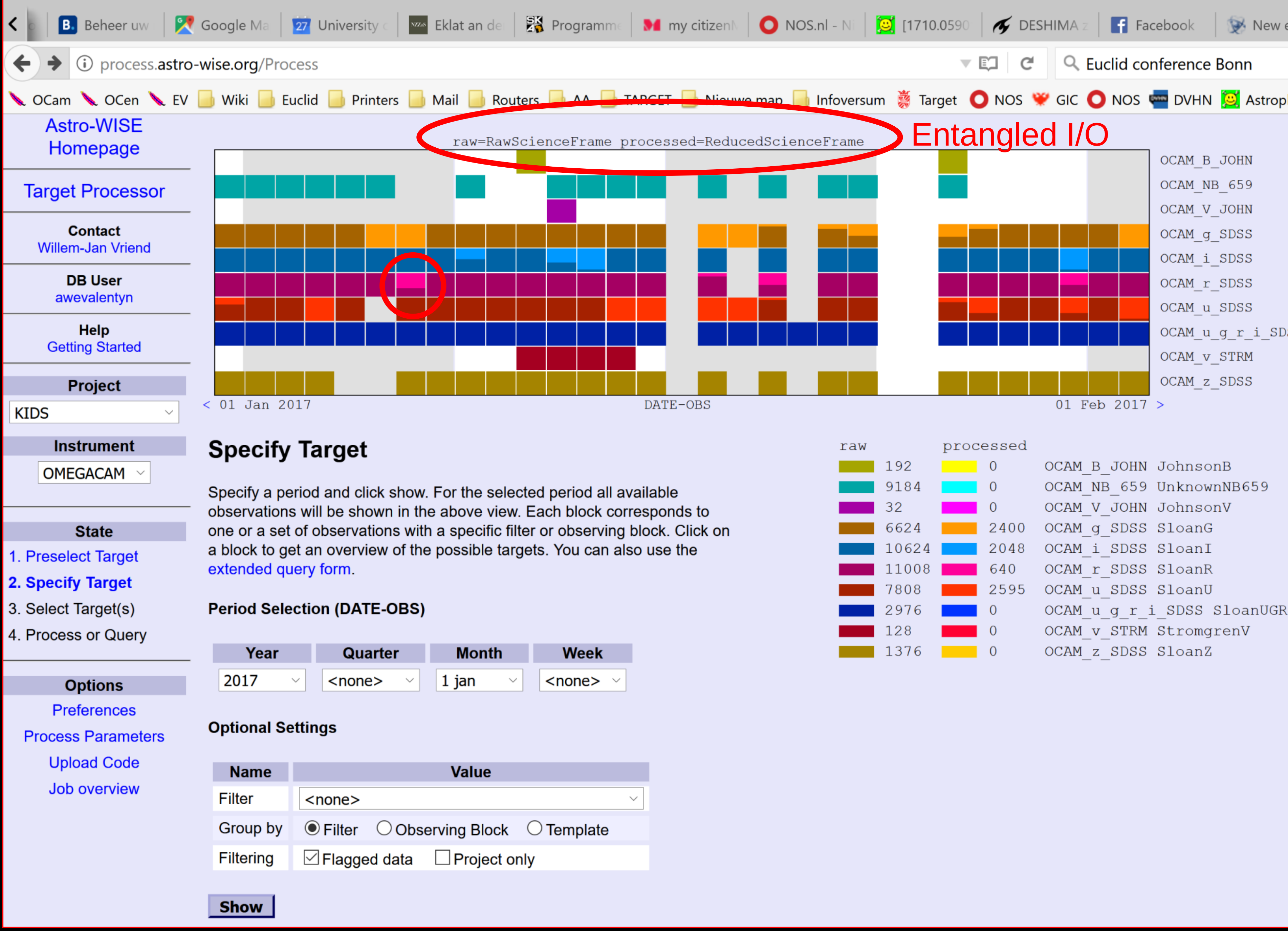
2018: Open Science - **FAIR** principles

**F**indable **A**ccessable **I**nteroperable **R**eproducable

# The universe as a spreadsheet

## Target Diagram/Data lineage /backward chaining ++ programming - dependencies





## Specify Target

Specify a period and click show. For the selected period all available observations will be shown in the above view. Each block corresponds to one or a set of observations with a specific filter or observing block. Click on a block to get an overview of the possible targets. You can also use the extended query form.

**Period Selection (DATE-OBS)**

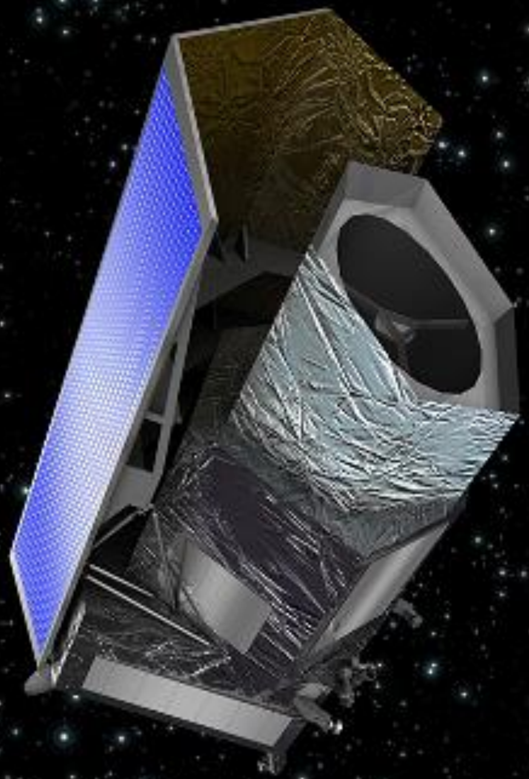
| Year | Quarter | Month | Week |
|---|---|---|---|
| 2017 | <none> | 1 jan | <none> |

**Optional Settings**

| Name | Value |
|---|---|
| Filter | <none> |
| Group by | ● Filter ○ Observing Block ○ Template |
| Filtering | ☑ Flagged data □ Project only |

| raw | | processed | | | |
|---|---|---|---|---|---|
| ■ | 192 | ■ | 0 | OCAM_B_JOHN | JohnsonB |
| ■ | 9184 | ■ | 0 | OCAM_NB_659 | UnknownNB659 |
| ■ | 32 | ■ | 0 | OCAM_V_JOHN | JohnsonV |
| ■ | 6624 | ■ | 2400 | OCAM_g_SDSS | SloanG |
| ■ | 10624 | ■ | 2048 | OCAM_i_SDSS | SloanI |
| ■ | 11008 | ■ | 640 | OCAM_r_SDSS | SloanR |
| ■ | 7808 | ■ | 2595 | OCAM_u_SDSS | SloanU |
| ■ | 2976 | ■ | 0 | OCAM_u_g_r_i_SDSS | SloanUGR |
| ■ | 128 | ■ | 0 | OCAM_v_STRM | StromgrenV |
| ■ | 1376 | ■ | 0 | OCAM_z_SDSS | SloanZ |

# Euclid

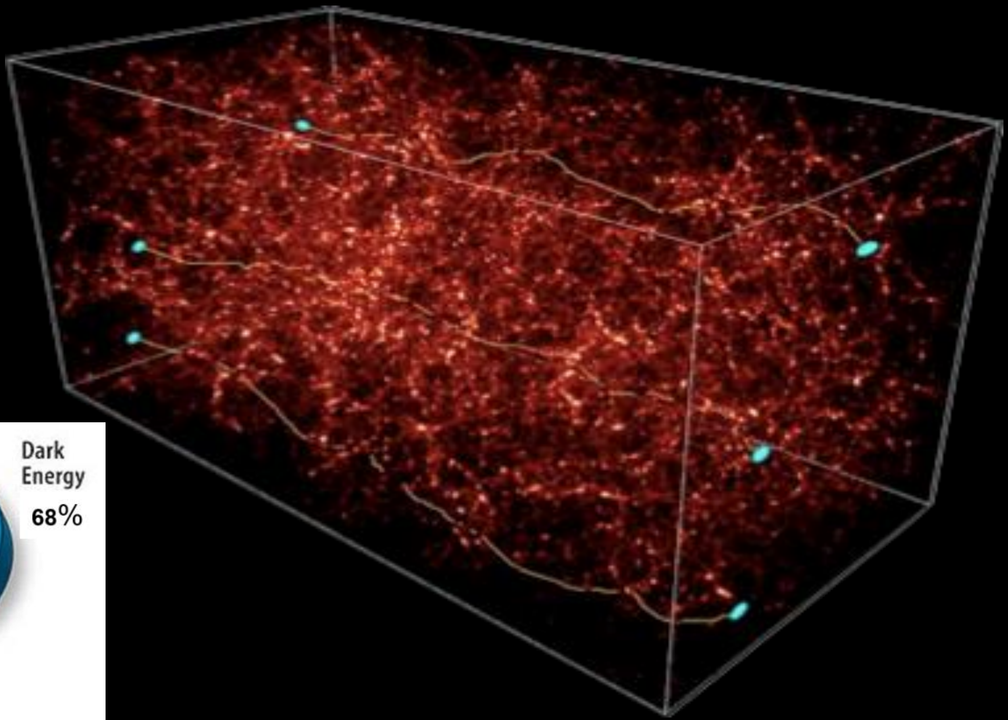


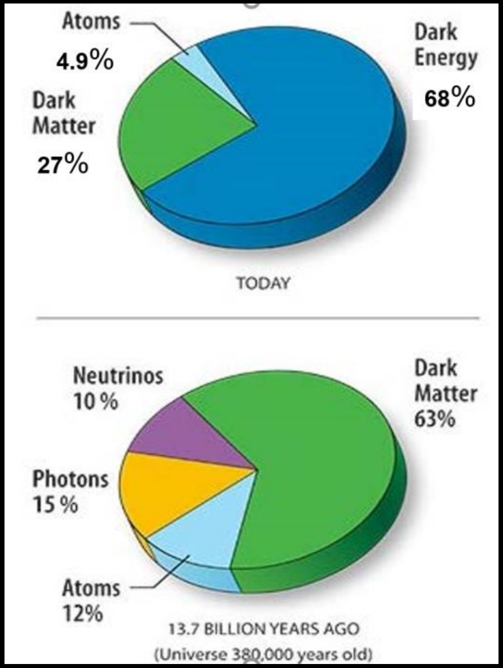
ESA launch in May 2021

Euclid Archive System (EAS)

- data centric information system
- many of the WISE concepts
- prototype uses Astro-WISE
- db hosted in the Euclid SDC-NL in Groningen

# Weak gravitational lensing as probe of dark matter





KiDS: < 100 $10^6$ redshifts

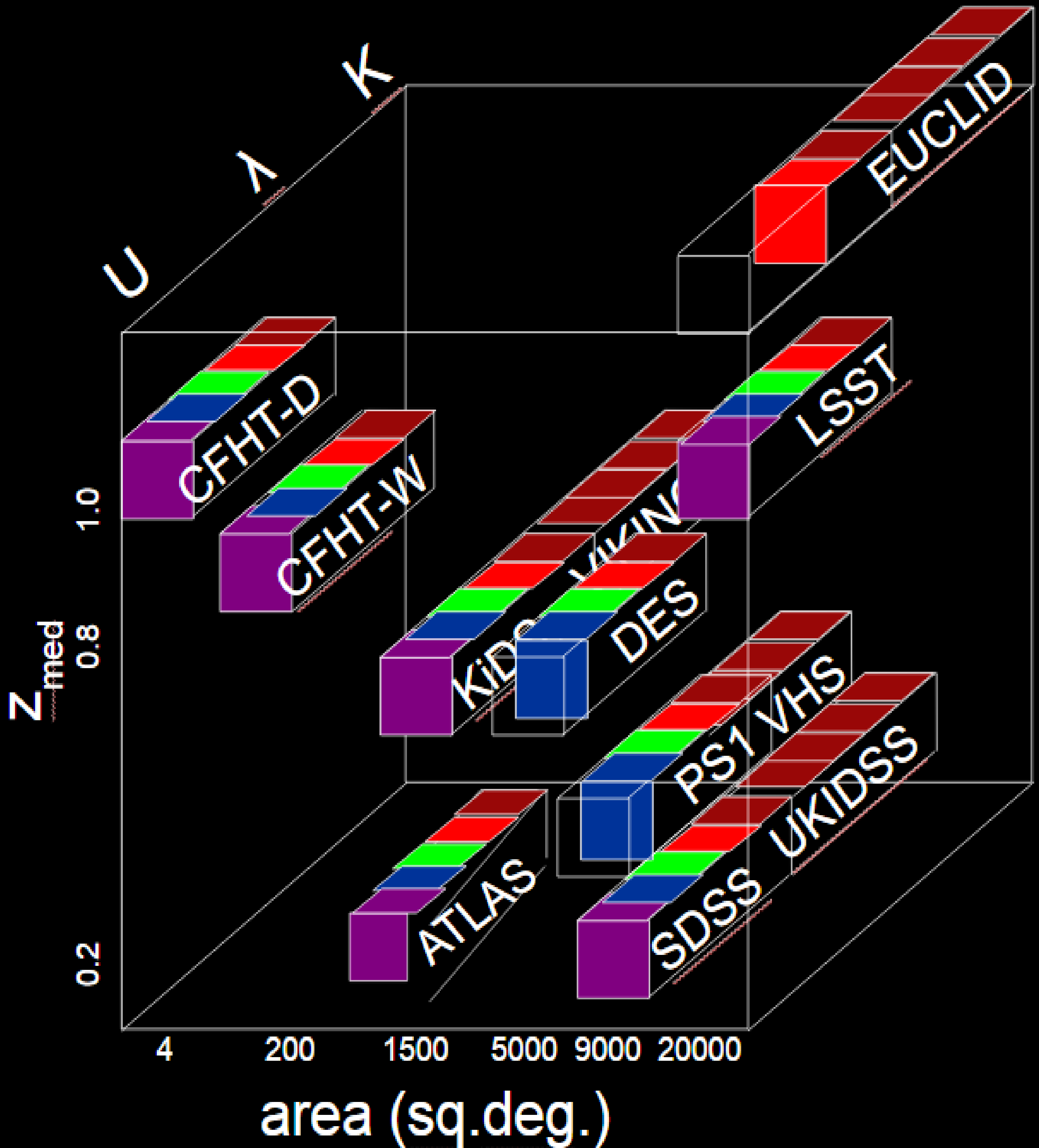
EUCLID: 1.5 $10^9$ redshifts - phot- z

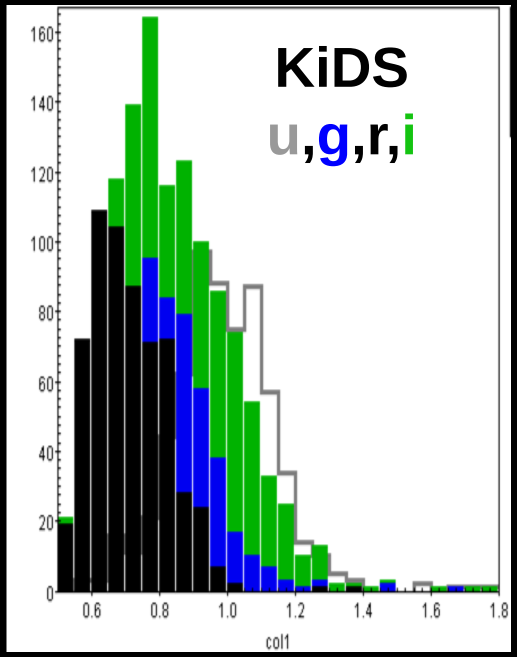
Ground based data – OU-Ext

Every galaxy has its own 4 PSFs

QC- bias – re-processing

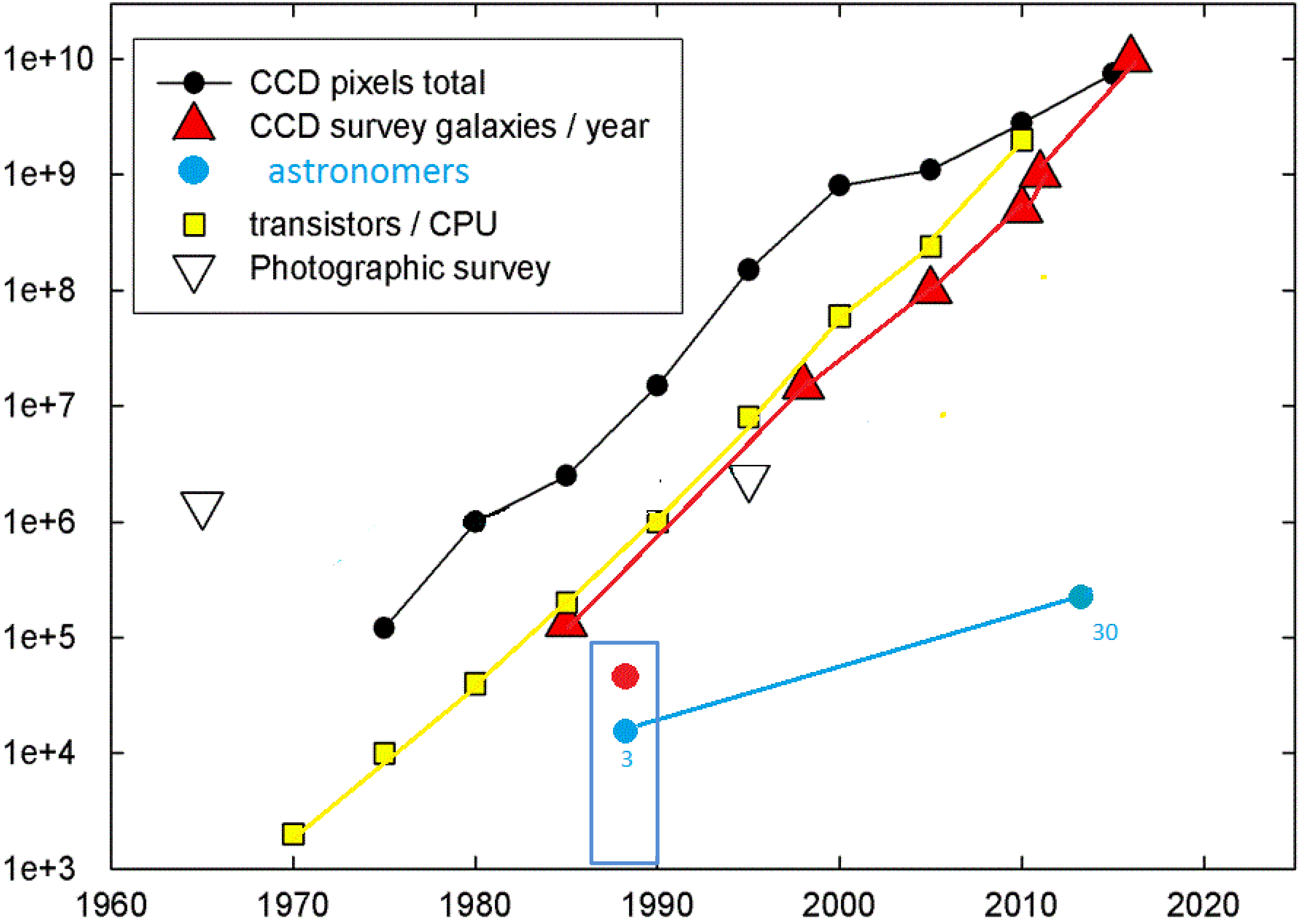
# KiDS/VIKING





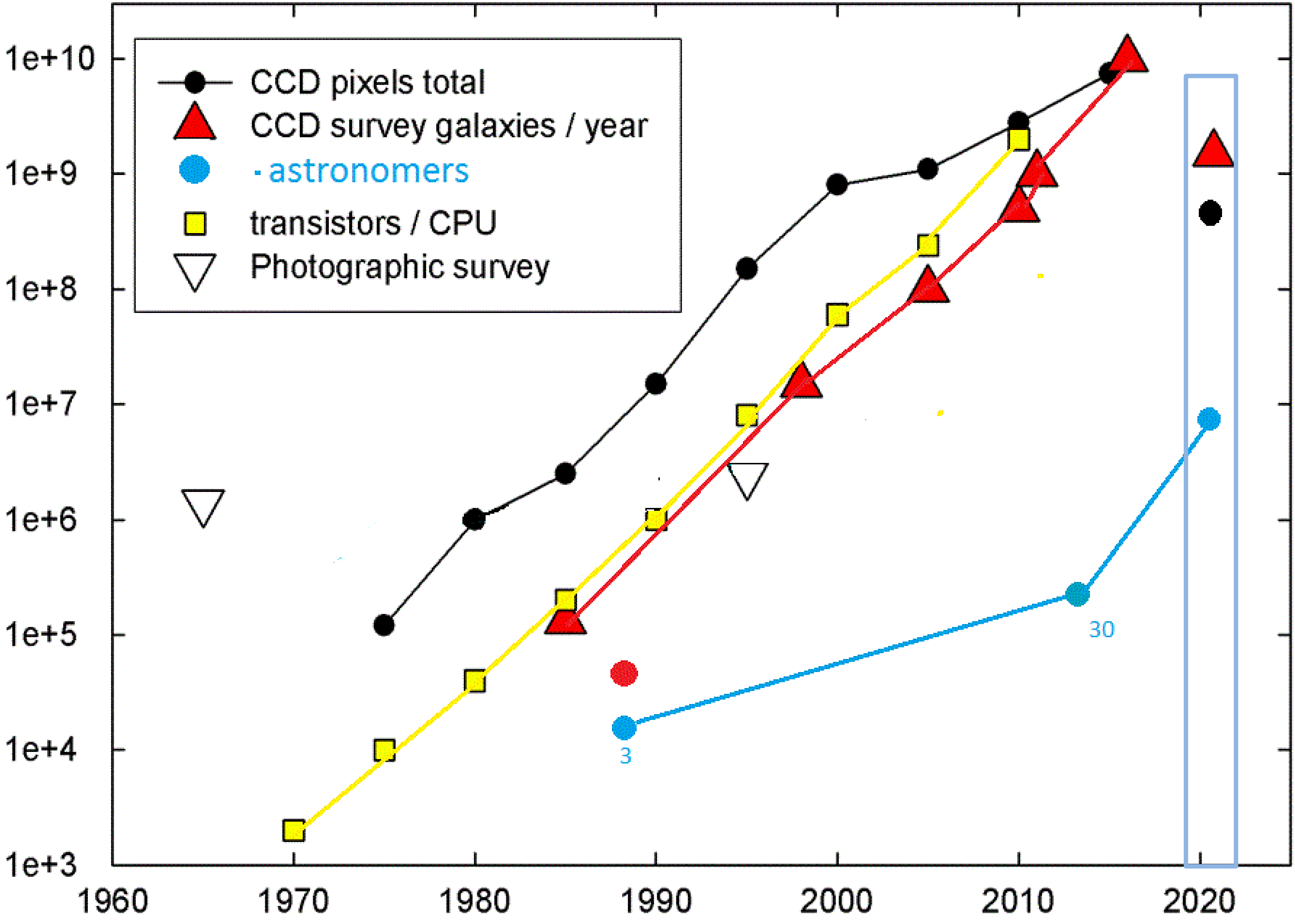
Seeing (“)

# Trends in Optical Astronomy Survey Data



after Tyson- Szalay

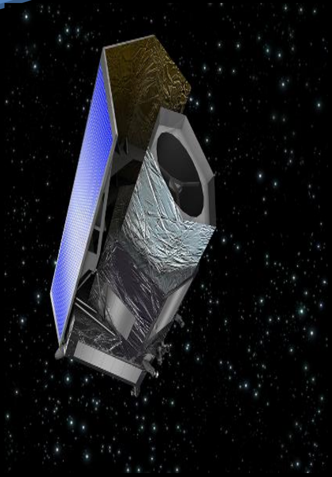
# Trends in Optical Astronomy Survey Data



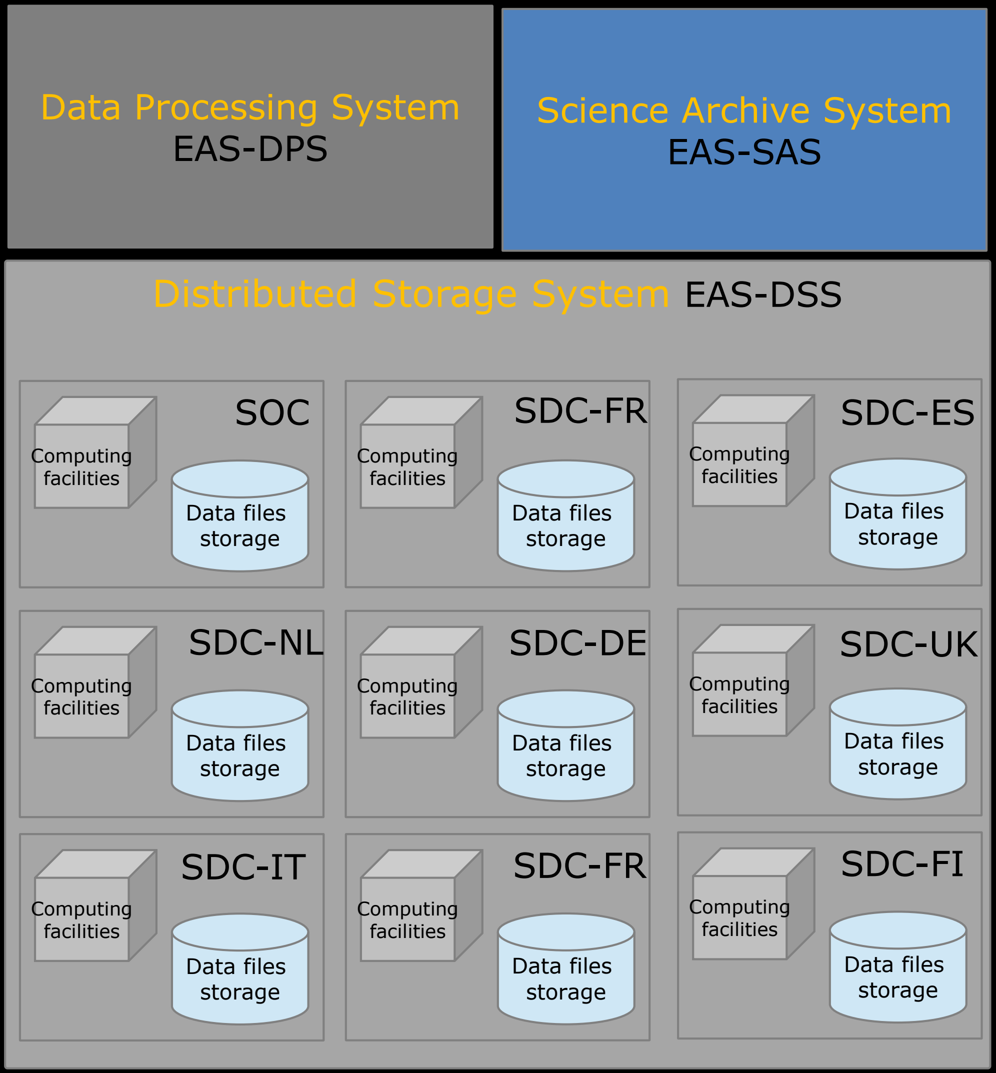
# Distributed communities acces-proces-calibrate-analyse publish

Euclid:

- 1500 registered members and growing
- 200 laboratories/departments
- 16 countries contributing
- NASA/US: provides the IR detectors.



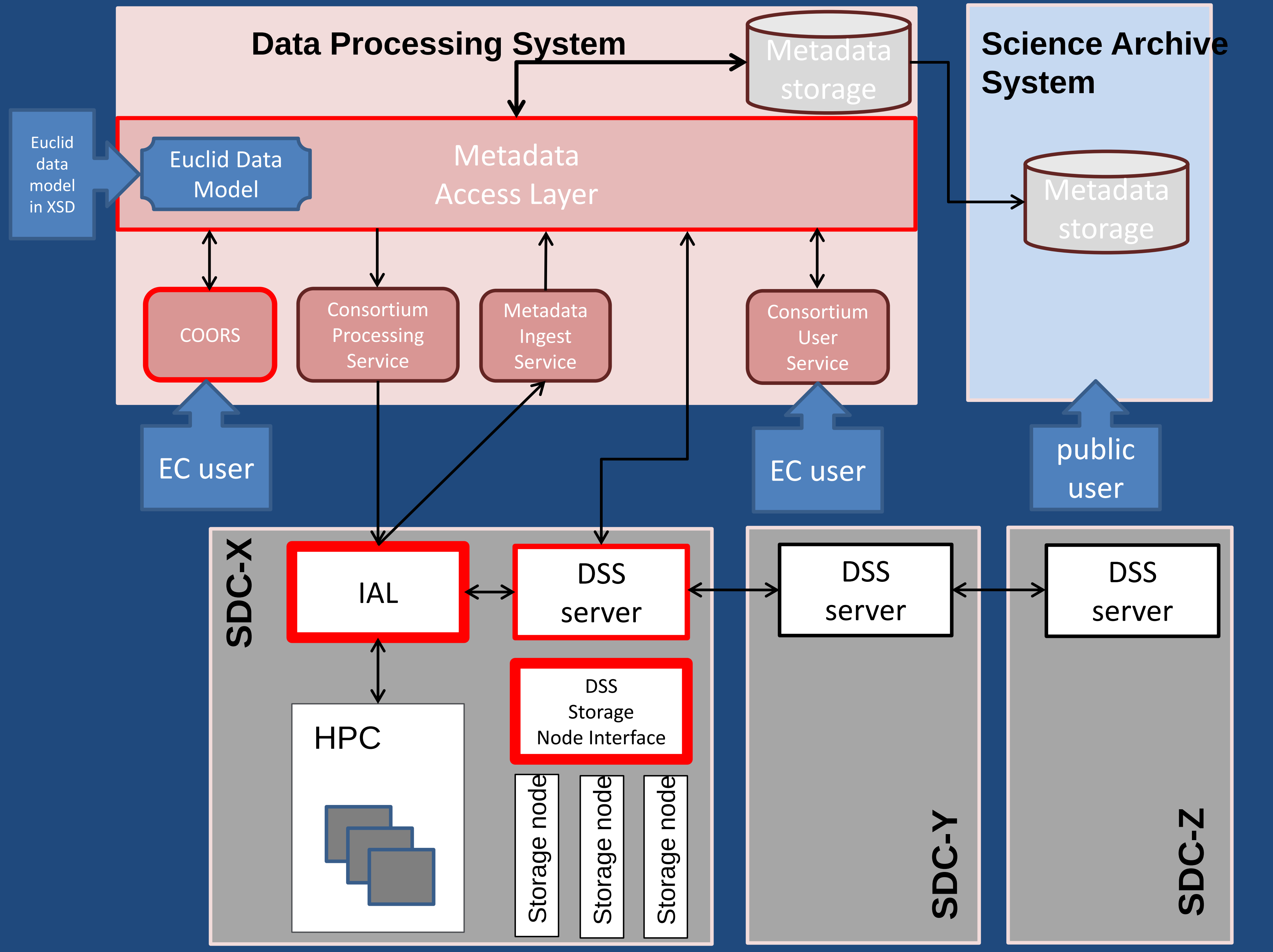
# Euclid Archive system – EAS – lay out



Data Processing System
EAS-DPS

Science Archive System
EAS-SAS

Distributed Storage System EAS-DSS



Data Processing System

Metadata storage

Science Archive System

Metadata storage

Euclid data model in XSD

Euclid Data Model

Metadata Access Layer

COORS

Consortium Processing Service

Metadata Ingest Service

Consortium User Service

EC user

EC user

public user

SDC-X

IAL

DSS server

HPC

DSS Storage Node Interface

Storage node

Storage node

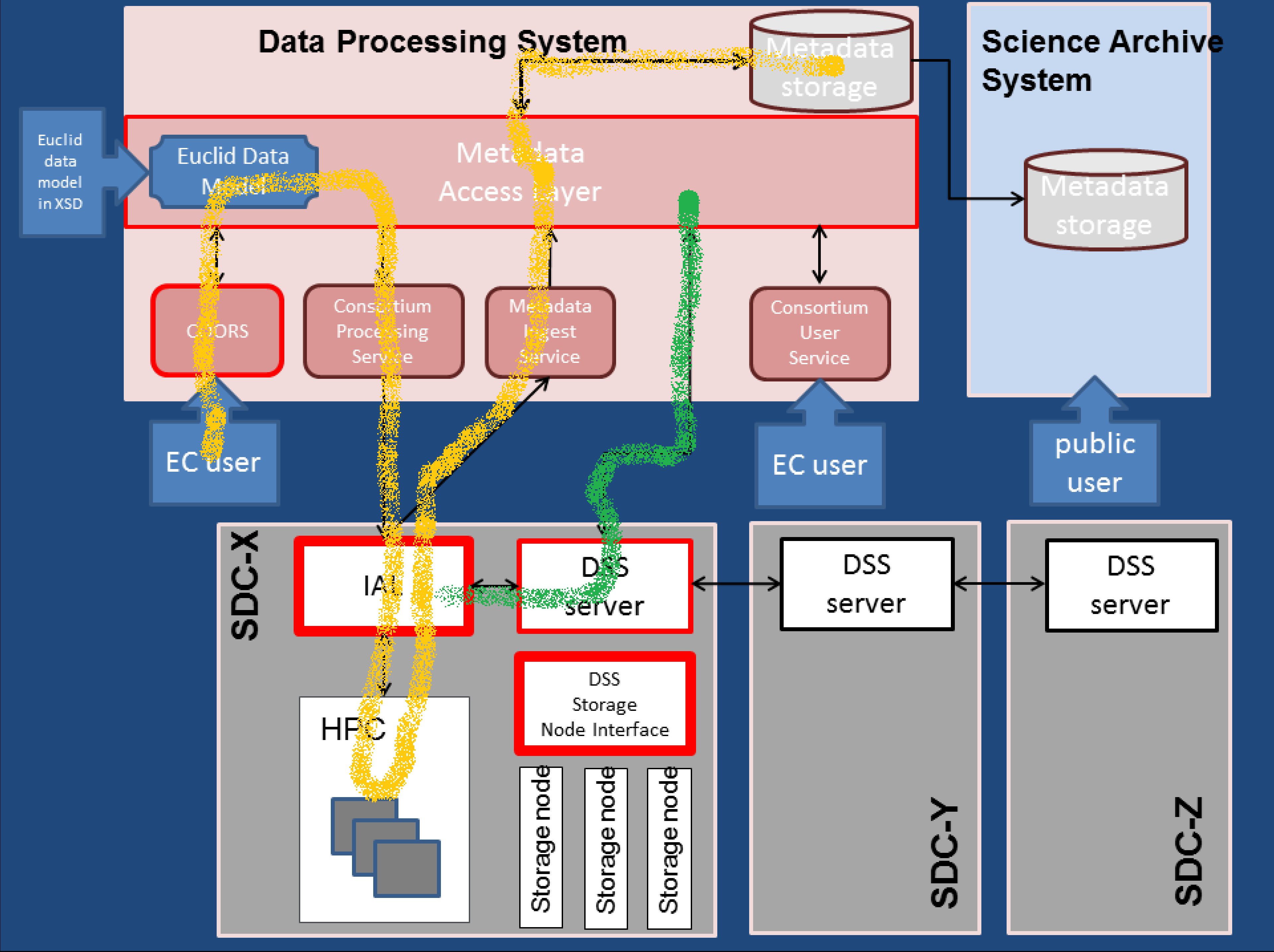
Storage node

DSS server

SDC-Y

DSS server

SDC-Z



Data Processing System

Metadata storage

Science Archive System

Metadata storage

Euclid data model in XSD

Euclid Data Model

Metadata Access Layer

COORS

Consortium Processing Service

Metadata Ingest Service

Consortium User Service

EC user

EC user

public user

SDC-X

IAL

DSS server

HPC

DSS Storage Node Interface

Storage node

Storage node

Storage node

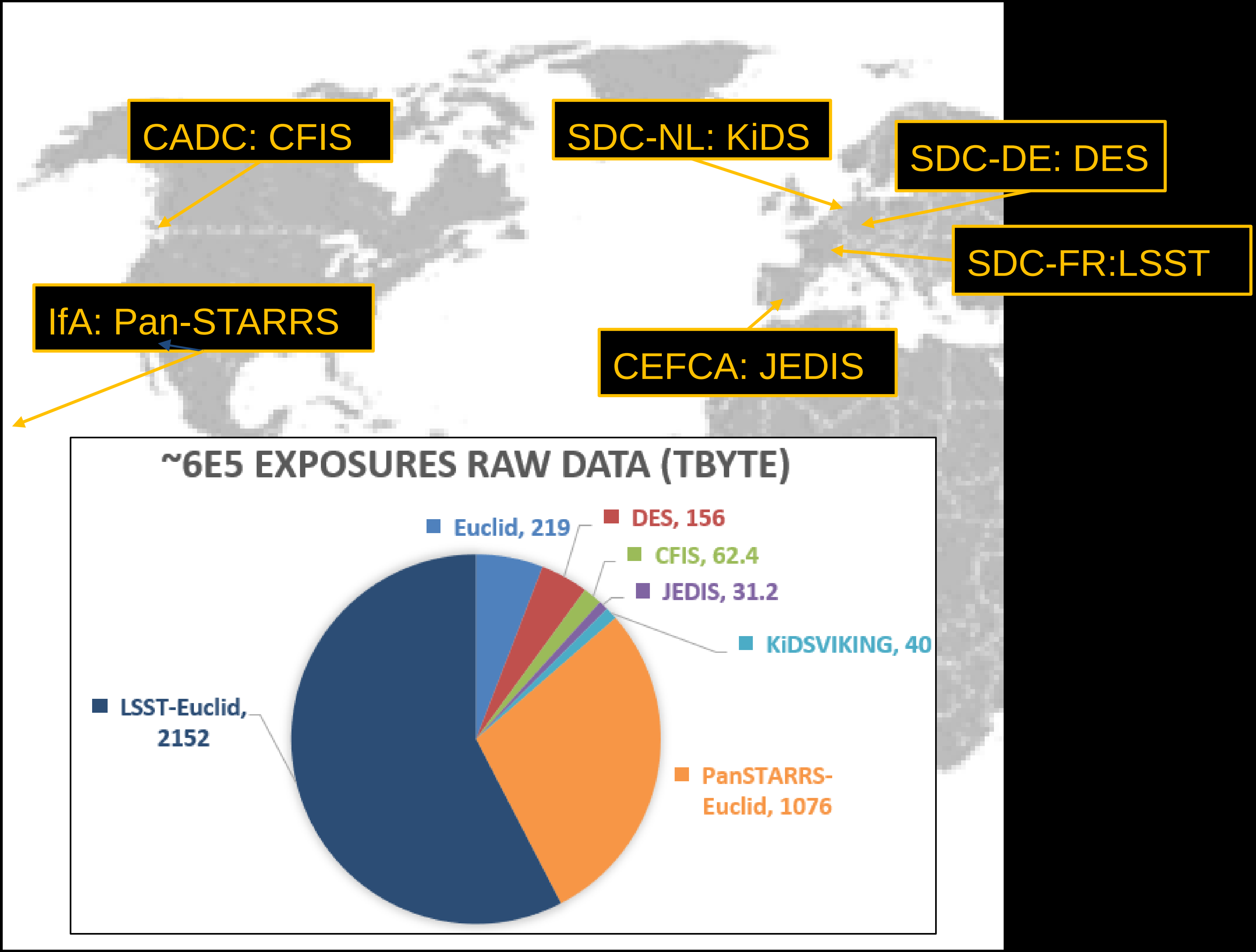
DSS server

SDC-Y

DSS server

SDC-Z

# Euclid-EXT: massive pixel volumes - distributed archives



CADC: CFIS

SDC-NL: KiDS

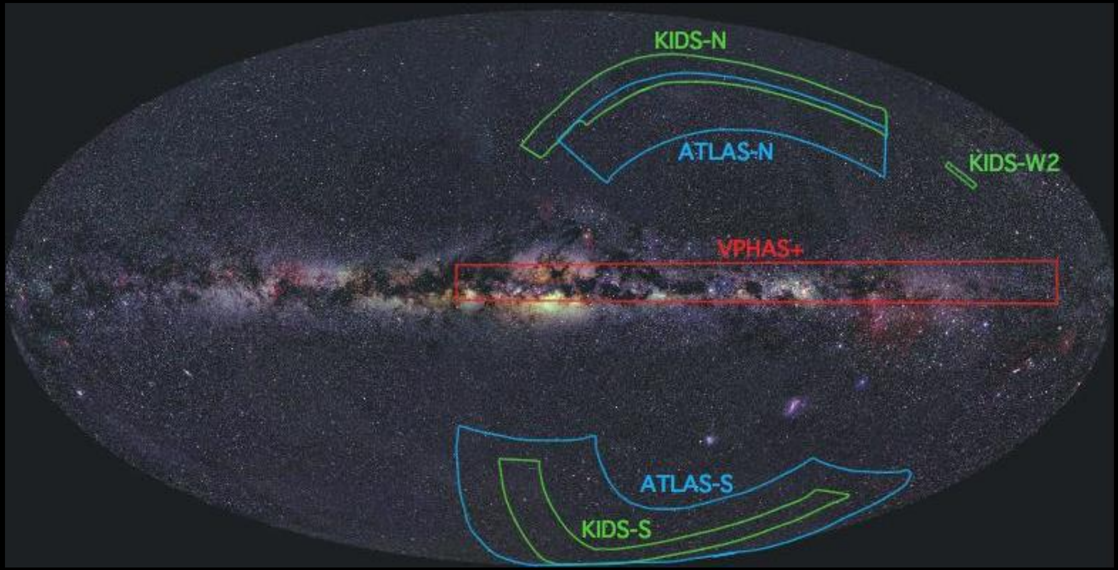
SDC-DE: DES

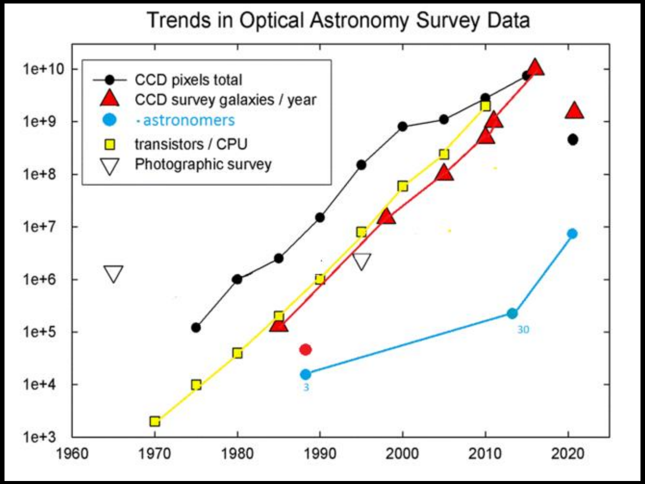
SDC-FR:LSST

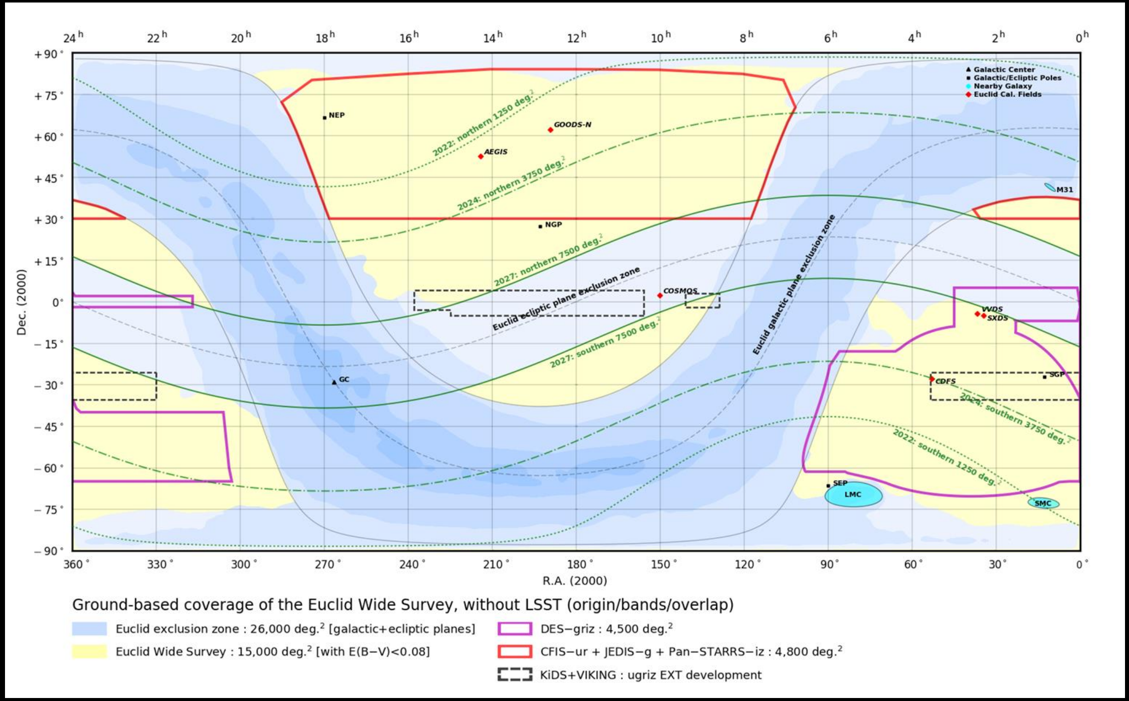
IfA: Pan-STARRS

CEFCA: JEDIS

# From KiDS to Euclid-EXT

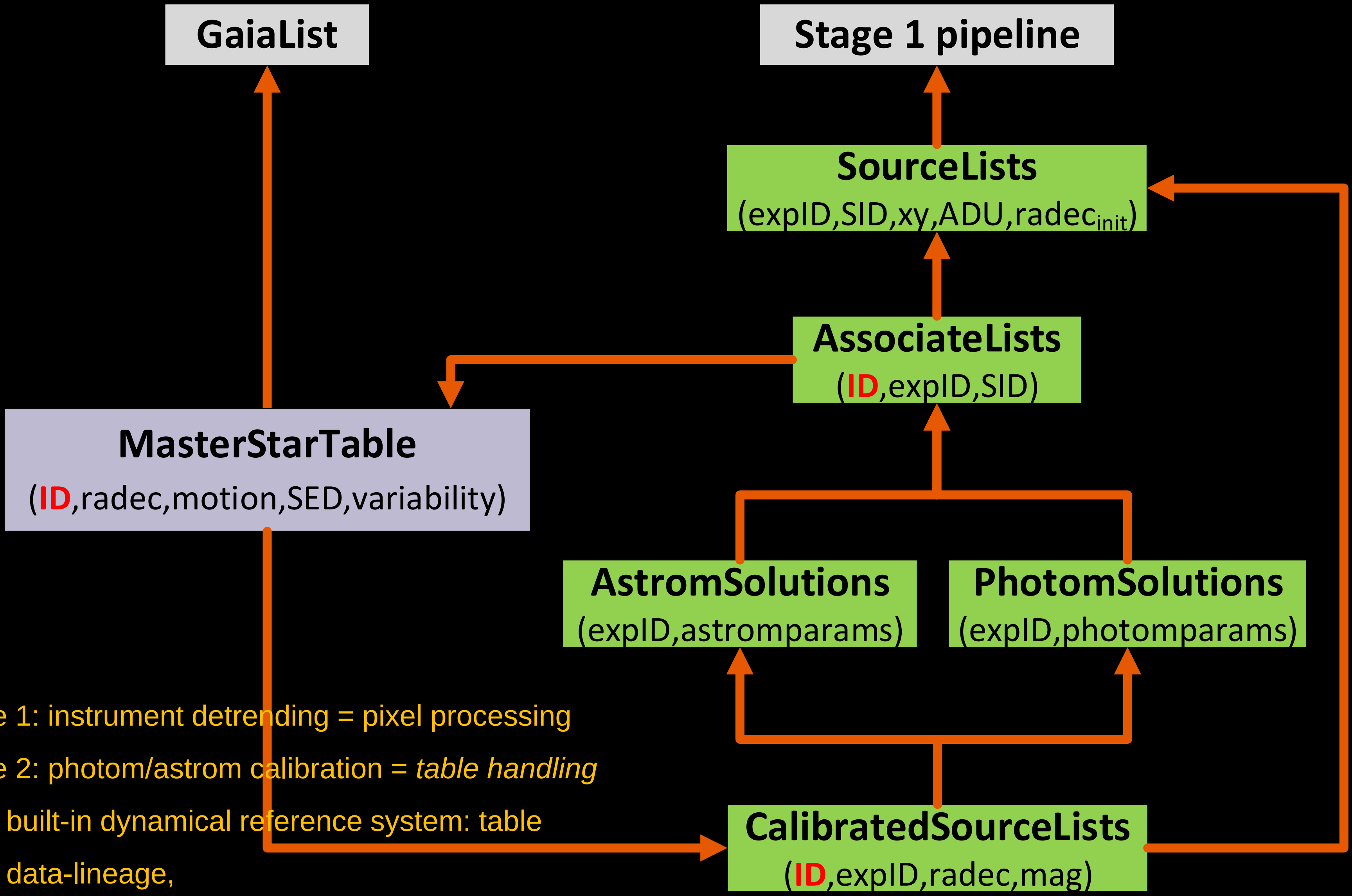





24h
22h
20h
18h
16h
14h
12h
10h
8h
6h
4h
2h
0h
+90°
+75°
+60°
+45°
+30°
+15°
0°
−15°
−30°
−45°
−60°
−75°
−90°
Dec. (2000)
360°
330°
300°
270°
240°
210°
180°
150°
120°
90°
60°
30°
0°
R.A. (2000)
Galactic Center
Galactic/Ecliptic Poles
Nearby Galaxy
Euclid Cal. Fields
NEP
GOODS-N
AEGIS
NGP
COSMOS
M31
VVDS
SXDS
CDFS
SGP
GC
SEP
LMC
SMC
2022: northern 1250 deg.²
2024: northern 3750 deg.²
2027: northern 7500 deg.²
2027: southern 7500 deg.²
2024: southern 3750 deg.²
2022: southern 1250 deg.²
Euclid ecliptic plane exclusion zone
Euclid galactic plane exclusion zone
Ground-based coverage of the Euclid Wide Survey, without LSST (origin/bands/overlap)
Euclid exclusion zone : 26,000 deg.² [galactic+ecliptic planes]
Euclid Wide Survey : 15,000 deg.² [with E(B−V)<0.08]
DES−griz : 4,500 deg.²
CFIS−ur + JEDIS−g + Pan−STARRS−iz : 4,800 deg.²
KiDS+VIKING : ugriz EXT development

Euclidization
Changing reference systems
Astrometry- photometry

# Target diagram ( ++ dependencies) for OU-EXT – Euclid external data - stage 2- dynamic Euclidization



GaiaList

Stage 1 pipeline

SourceLists

(expID,SID,xy,ADU,$radec_{init}$)

AssociateLists

(ID,expID,SID)

MasterStarTable

(ID,radec,motion,SED,variability)

AstromSolutions

(expID,astromparams)

PhotomSolutions

(expID,photomparams)

CalibratedSourceLists

(ID,expID,radec,mag)

Stage 1: instrument detrending = pixel processing

Stage 2: photom/astrom calibration = *table handling*

- built-in dynamical reference system: table data-lineage,
- QC, re-processing

# Beyond Big Data

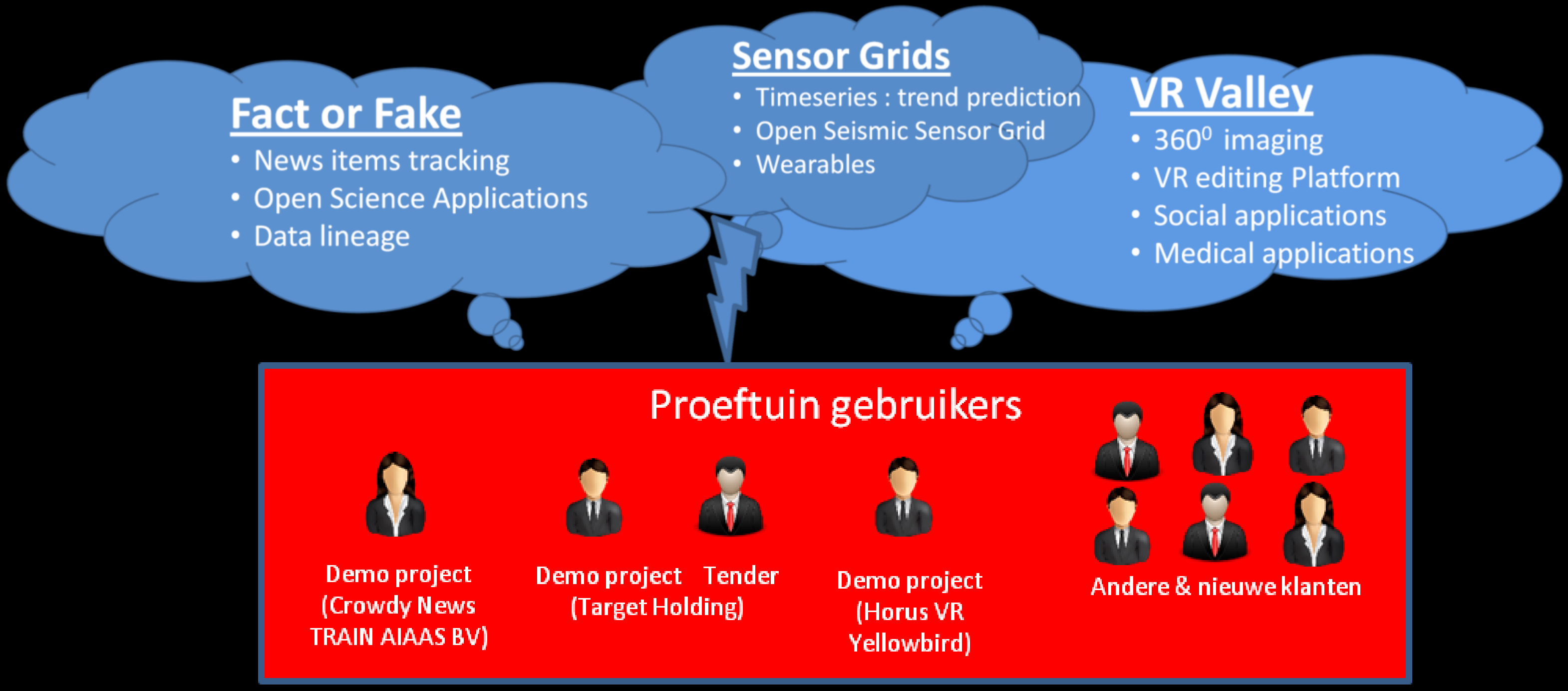
- QC and re-processing – Kids Euclid **FAIR**
- OU EXT > Billion – dynamic tables

All techniques go back to the source

Scientists and journalists- > Fact and Fakes

Structured data and unstructured data

# TARGET Fieldlab



**Fact or Fake**

- News items tracking
- Open Science Applications
- Data lineage

**Sensor Grids**

- Timeseries : trend prediction
- Open Seismic Sensor Grid
- Wearables

**VR Valley**

- $360^0$ imaging
- VR editing Platform
- Social applications
- Medical applications

Proeftuin gebruikers

Demo project (Crowdy News TRAIN AIAAS BV)

Demo project Tender (Target Holding)

Demo project (Horus VR Yellowbird)

Andere & nieuwe klanten

**Media scanner**

Focus on domains

**ML Lineage builder**

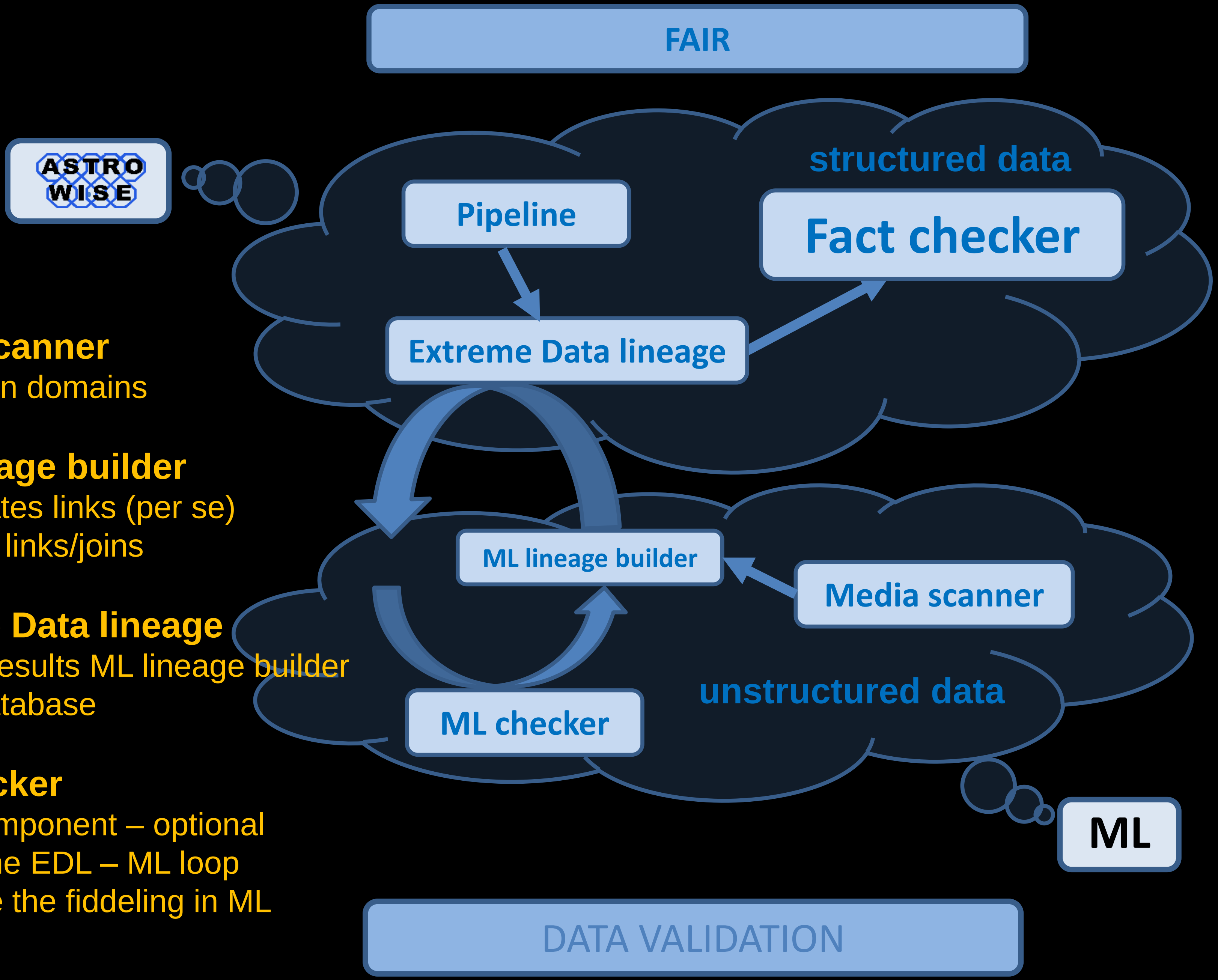
ML creates links (per se)
multiple links/joins

**Extreme Data lineage**

Import results ML lineage builder
AWE database

**ML Checker**

New component – optional
Close the EDL – ML loop
Replace the fiddeling in ML



FAIR

ASTRO WISE

structured data

Pipeline

Fact checker

Extreme Data lineage

ML lineage builder

Media scanner

unstructured data

ML checker

ML

DATA VALIDATION

# conclusions

Next level is all about Data validation

- check ML
- QC
- systematics  in data sets
- OU-ext  dynamic Euclidization
- unstructured data:  ML + lineage

Almost all about going back to the source

Facts and Fakes