Laboratório Interinstitucional de e-Astronomia

ASTRONOMY WORKSHOP 2016

BRICS

ASTRONOMICAL DATA AND COMPUTATION

# Analyzing large astronomical datasets: The Science Portal solution

Angelo Fausti Neto

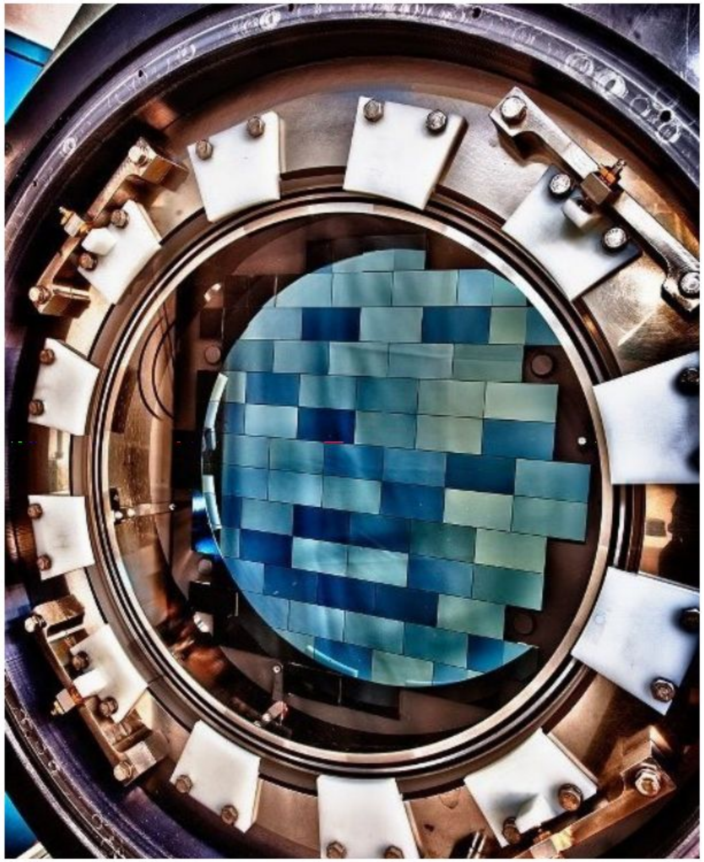
Laboratório Interinstitucional de e-Astronomia (LIneA)
Large Synoptic Survey Telescope (LSST)

# Outline

- Brazilian participation in Large Surveys
  - The role of LIneA  (see Ricardo Ogando's presentation)
  - DES-Brazil, BPG-SDSS, DESI, BPG-LSST
- Science Portal
  - Applications in DES
  - Challenges
- Conclusions and perspectives

# Dark Energy Survey (DES)

https://www.darkenergysurvey.org/



Dark Energy Camera, in operation since 2012 at Blanco (CTIO)

- Photometric Survey (grizY)
- DECam 570 Mpixels (62 CCDs)
- Blanco 4m  (CTIO)
- ~ 300 exposures each night (500GB)
- 5 years (~100 nights/year)
- 5.000 sq deg
- 4th year of operations (Ago 2016)
- First Public Data Release - DR1 (2017)
- Objects catalog ~1Billion
- DES-Brazil  (2007)

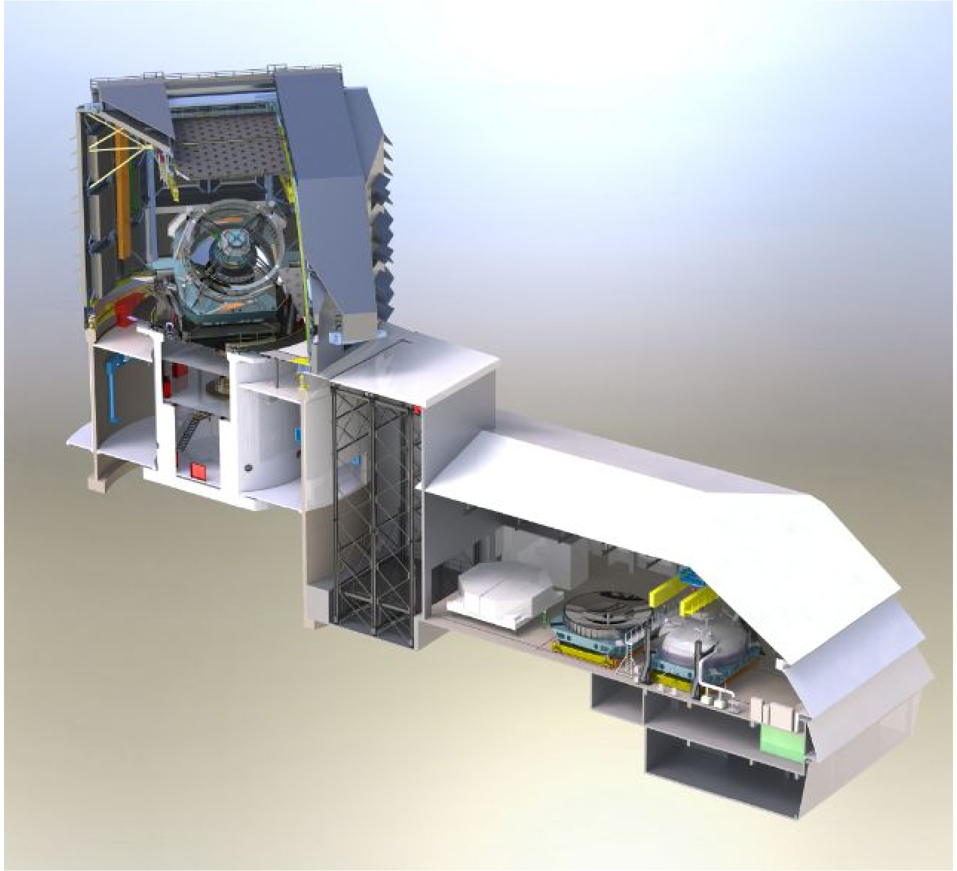
Começa o quarto ano de observações do Dark Energy Survey

19 de agosto de 2016

http://www.linea.gov.br/Noticias/

# Large Synoptic Survey Telescope (LSST)

https://www.lsst.org/scientists



- Photometric Survey (ugrizy)
- LSST Cam 3.2 Gpixels (189 CCDs)
- M1/M3 primary 8.4m  (CTIO)
- 1 "visit" = 2 exposures in 30s
- 1000 "visits" each night (~15 TB)
- 18.000 sq deg - twice a week for 10 years
- Start operations in 2022
- 10 Million alerts/night
- Objects Catalog ~37 Billion
- BPG-LSST (2016)

http://bpg-lsst.linea.gov.br/

# Common challenges

- Big science questions
  - Active participation in international collaborations
  - “M-shaped” person (astronomy, scientific computing, statistics)
- Big collaborations
  - Cultural change  (**collaborative tools)**
  - Communication challenges (**telecons, distributed information**)
  - Understand the projects and their opportunities
- Big data volume and variety
  - Infrastructure to support science (data transfer, storage and processing)
  - Efficient data preparation and analysis
  - Data science techniques

# The “new astronomy"

**IN THE PAST**

Data > Inference > Model

(astronomer)

**THESE DAYS**

Data > Processing > Catalog > Inference > Model

(“Big data” problem)

(“New astronomer”)

“M-shaped” person

Adapted from Andrew Connolly - LSST

# The role of LIneA in Brazil

http://www.linea.gov.br/

- Support science
  - DES-Brazil (2007), BPG-SDSS-III (2008), BPG-SDSS-IV (2014), DESI (2016) and BPG-LSST (2016)
  - 71 members (ON, UFRJ, USP, UNESP, UNICAMP, UFABC, UFRGS, UFSM)
  - Scientific and technical education
  - Public Outreach
- Data center operation
  - Hardware and services maintenance (external contract)
    - helpdesk, e-mail, twiki, git, doc-db, slack, etc
  - Data transfer (RNP/Brazil)
  - Database and processing consultant (LNCC/Brazil)
- Software Development
  - R&D
  - Software Development team (9 FTEs)
  - 3 PhDs in astronomy + external contributions

# What is the Science Portal?

- It is the collection of software tools developed by LIneA to **assess the data quality** and to **explore large astronomical datasets.**
- It is also a **web framework** that facilitates the integration of the **science analysis codes** into a hardware and software infrastructure that provides **efficient data preparation and analysis**.

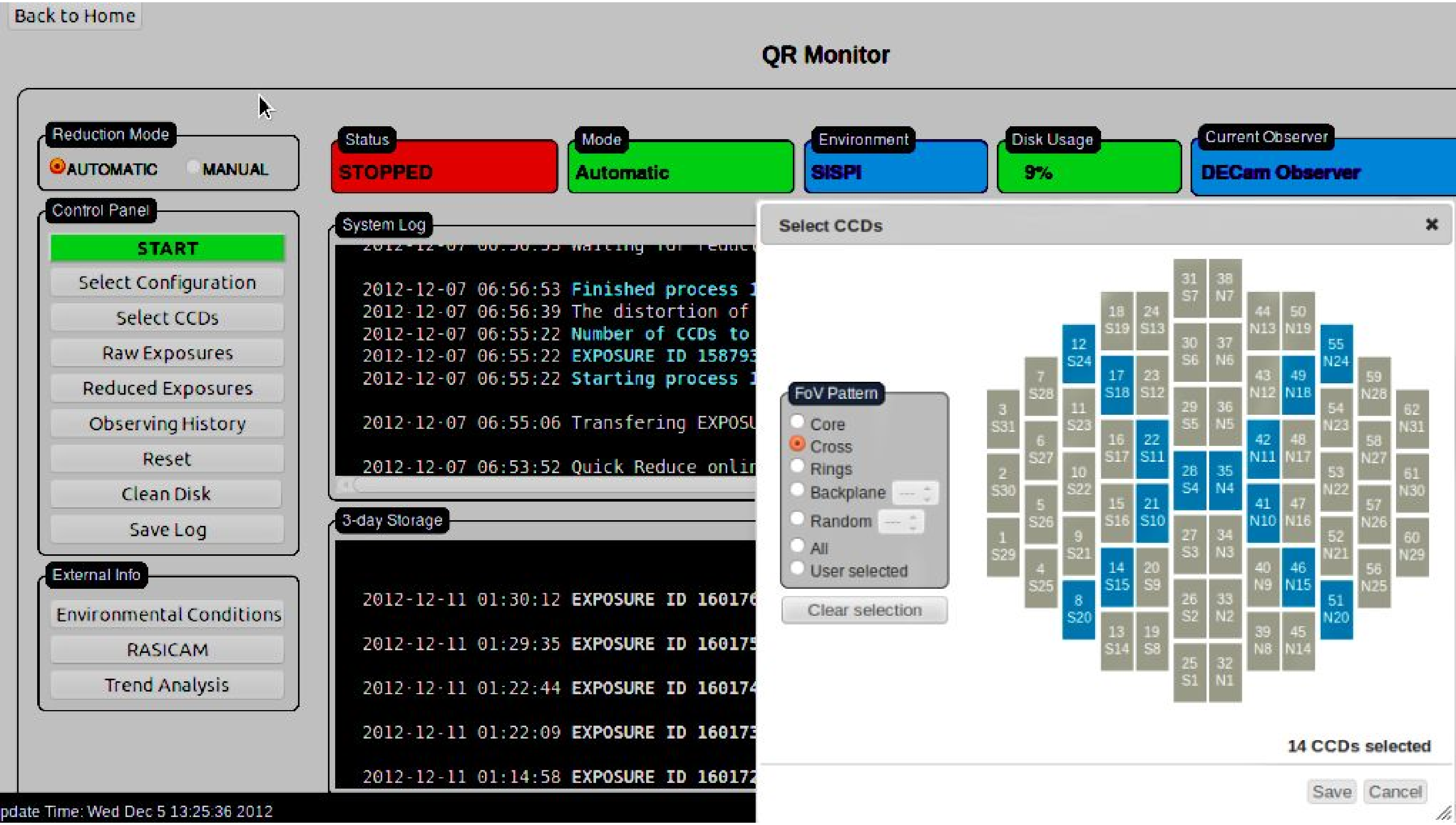
# Science Portal applications in DES

- Data Quality Assessment in real-time
  - Quick Reduce @ CTIO (2012)
  - http://quick1.ctio.noao.edu:8080/
- Data validation and exploration
  - Science Server @ Fermilab (2014)
  - Science Server @ NCSA (2016)
  - https://des-portal.fnal.gov/
  - http://desportal.cosmology.illinois.edu/
- Preparation of science-ready catalogs and science workflows
  - Science Portal @ LIneA (2016)
  - http://des-portal.linea.gov.br/

A new production system each 2 years

# DES Quick Reduce @ CTIO

Data reduction in real-time ~250k DECam exposures (~2.5M CCDs) since 2012
Developed for DES but also available for other programs



QR Monitor

Reduction Mode: ● AUTOMATIC ○ MANUAL

Control Panel

Status: STOPPED

Mode: Automatic

Environment: SISPI

Disk Usage: 9%

Current Observer: DECam Observer

System Log

```
2012-12-07 06:56:53 Finished process I
2012-12-07 06:56:39 The distortion of
2012-12-07 06:55:22 Number of CCDs to
2012-12-07 06:55:22 EXPOSURE ID 158793
2012-12-07 06:55:22 Starting process I

2012-12-07 06:55:06 Transfering EXPOSU

2012-12-07 06:53:52 Quick Reduce onlin
```

3-day Storage

```
2012-12-11 01:30:12 EXPOSURE ID 160176

2012-12-11 01:29:35 EXPOSURE ID 160175

2012-12-11 01:22:44 EXPOSURE ID 160174

2012-12-11 01:22:09 EXPOSURE ID 160173

2012-12-11 01:14:58 EXPOSURE ID 160172
```

External Info

Select CCDs

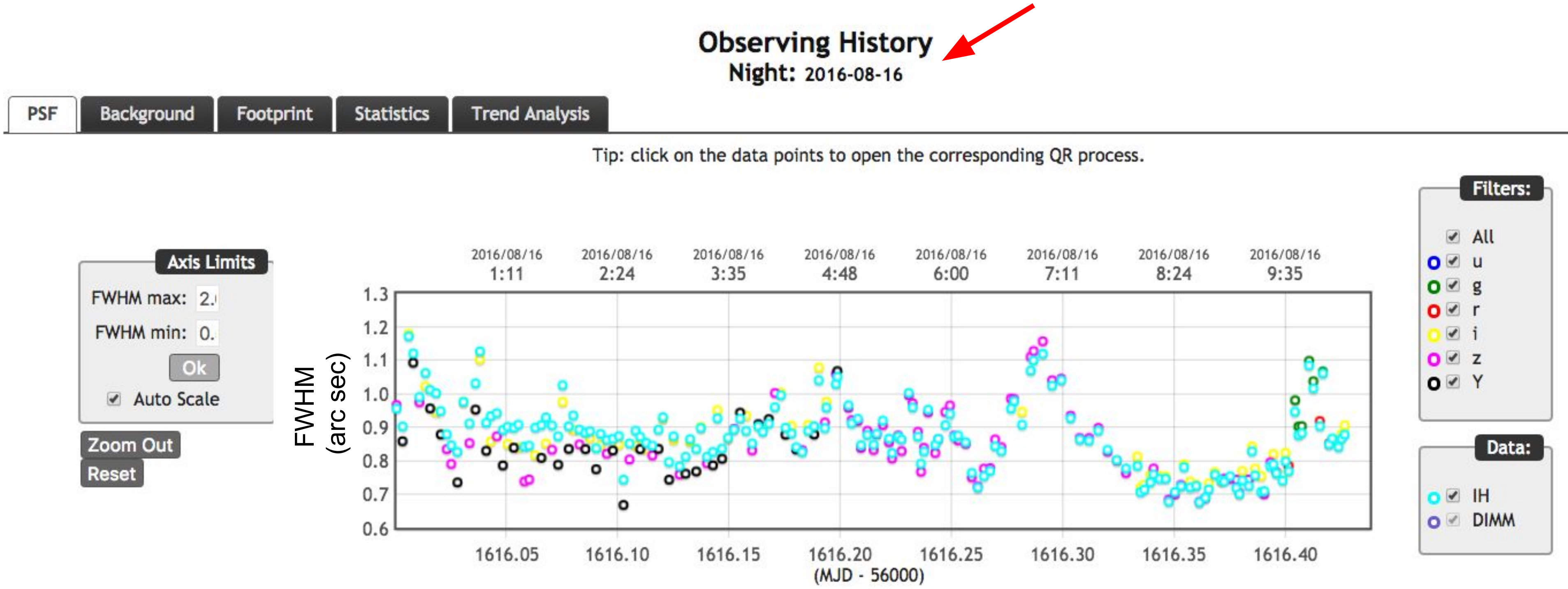
FoV Pattern
○ Core
● Cross
○ Rings
○ Backplane ---
○ Random ---
○ All
○ User selected

14 CCDs selected

pdate Time: Wed Dec 5 13:25:36 2012

# Monitoring DES observations

Daily transfer of QR results to Fermilab, access to DES collaboration



Observing History

Night: 2016-08-16

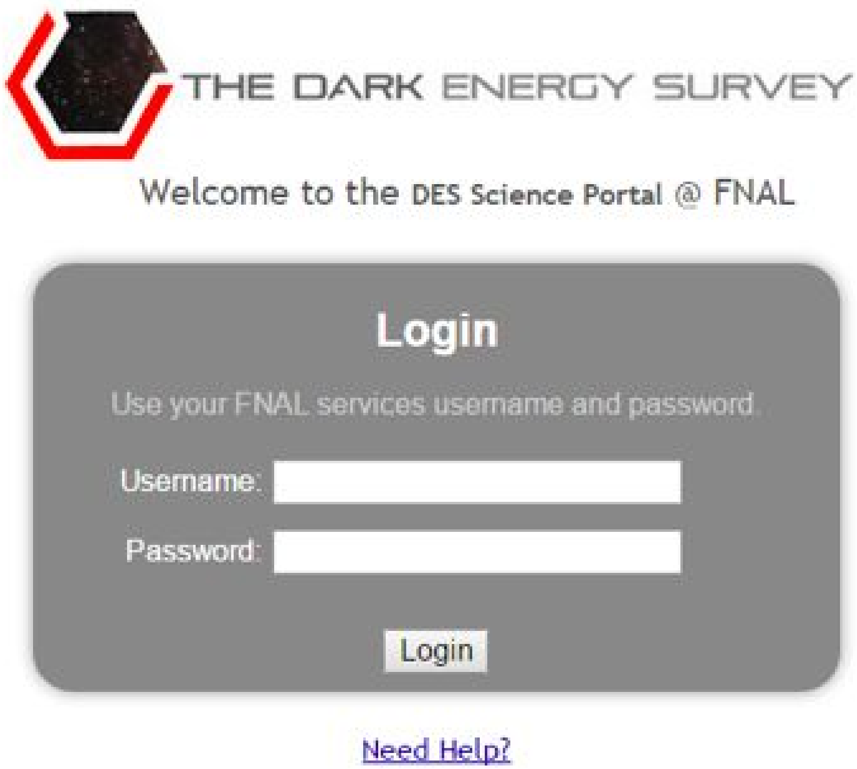
Tip: click on the data points to open the corresponding QR process.

Axis Limits

FWHM max: 2.(

FWHM min: 0.

☑ Auto Scale



THE DARK ENERGY SURVEY

Welcome to the DES Science Portal @ FNAL

Login

Use your FNAL services username and password

Username:

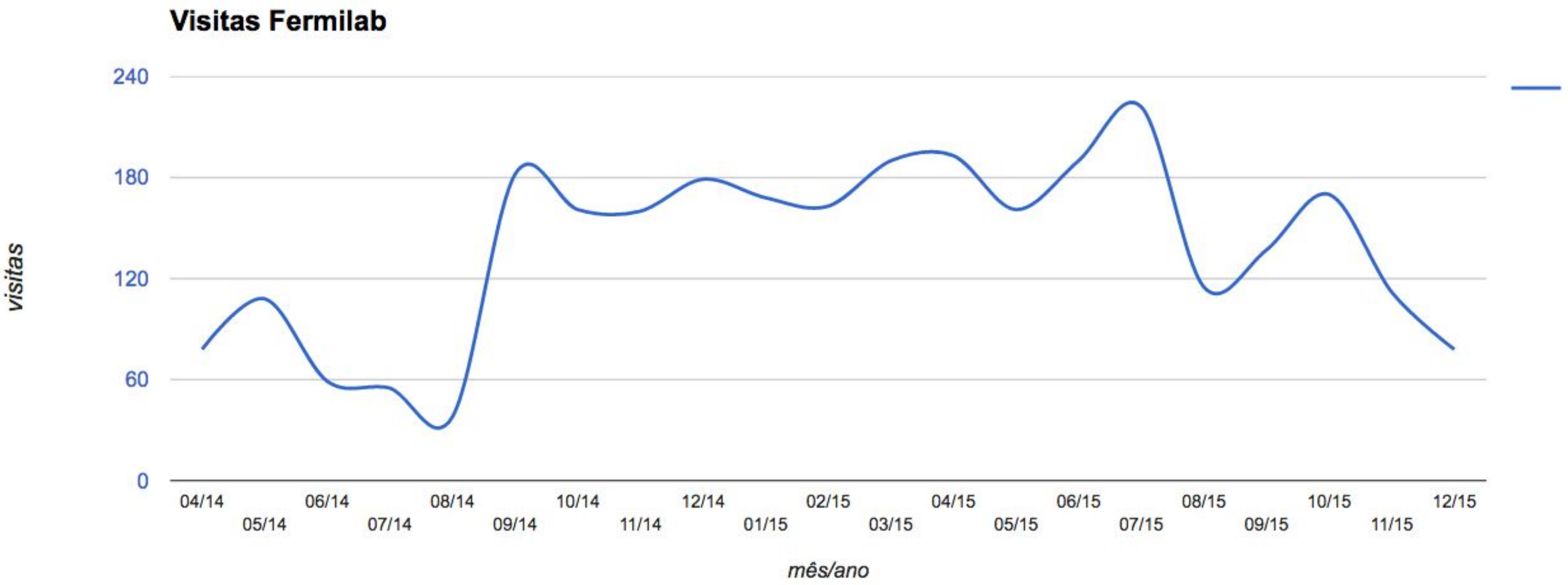
Password:

## Services available

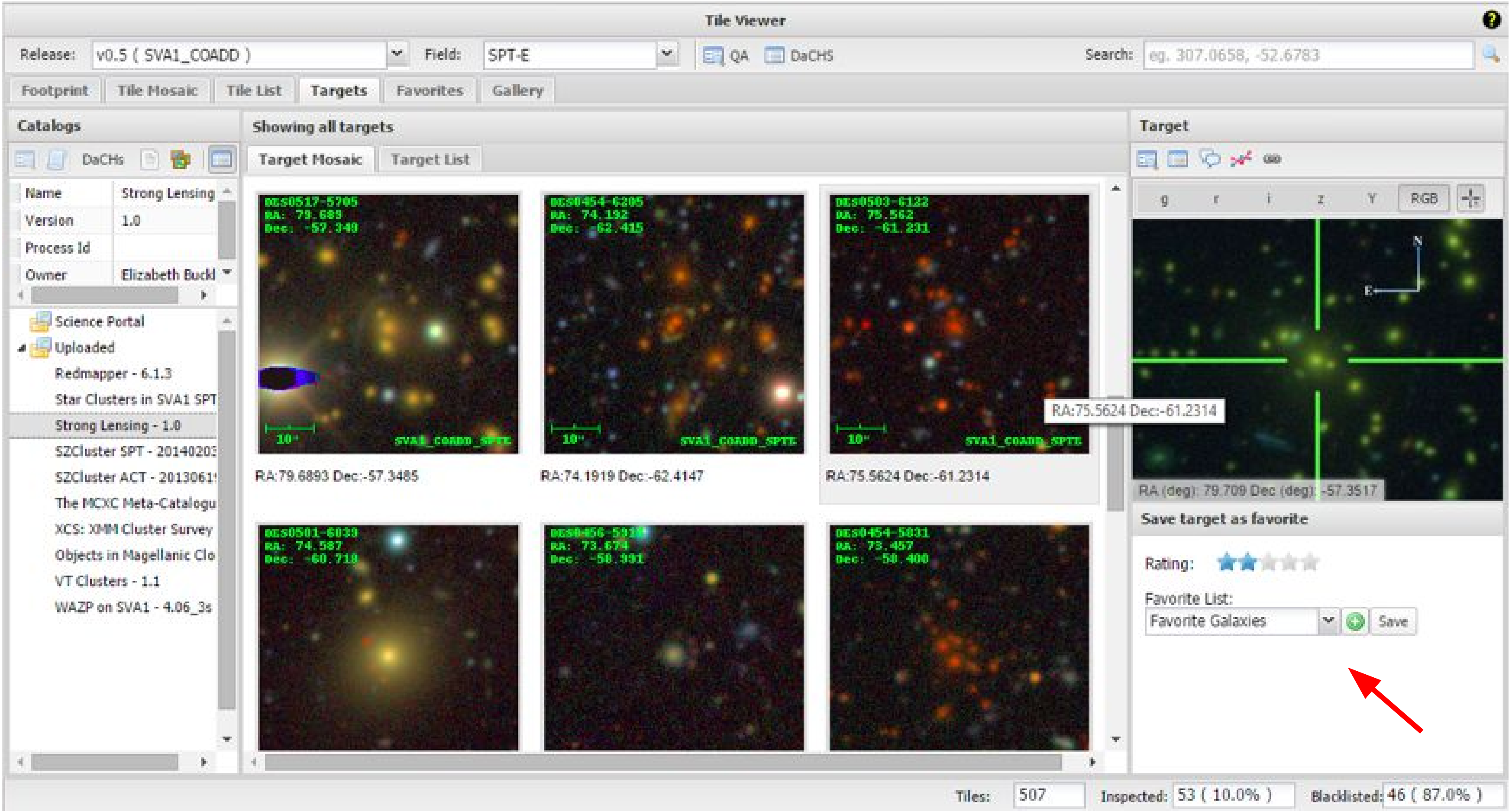
- Data Upload
- Visualization tools
- Cutout service
- User query

~180 DES collaboration members each month

“If you build they will come...”



# Upload, cutout and visual inspection
## e.g Strong Lensing, Galaxy Clusters



Release: v0.5 ( SVA1_COADD )

Field: SPT-E

Search: eg. 307.0658, -52.6783

| | |
|---|---|
| Name | Strong Lensing |
| Version | 1.0 |
| Process Id | |
| Owner | Elizabeth Buckl |

- Science Portal
- Uploaded
  - Redmapper - 6.1.3
  - Star Clusters in SVA1 SPT
  - Strong Lensing - 1.0
  - SZCluster SPT - 20140203
  - SZCluster ACT - 2013061
  - The MCXC Meta-Catalogu
  - XCS: XMM Cluster Survey
  - Objects in Magellanic Clo
  - VT Clusters - 1.1
  - WAZP on SVA1 - 4.06_3s

RA:79.6893 Dec:-57.3485

RA:74.1919 Dec:-62.4147

RA:75.5624 Dec:-61.2314

RA:75.5624 Dec:-61.2314

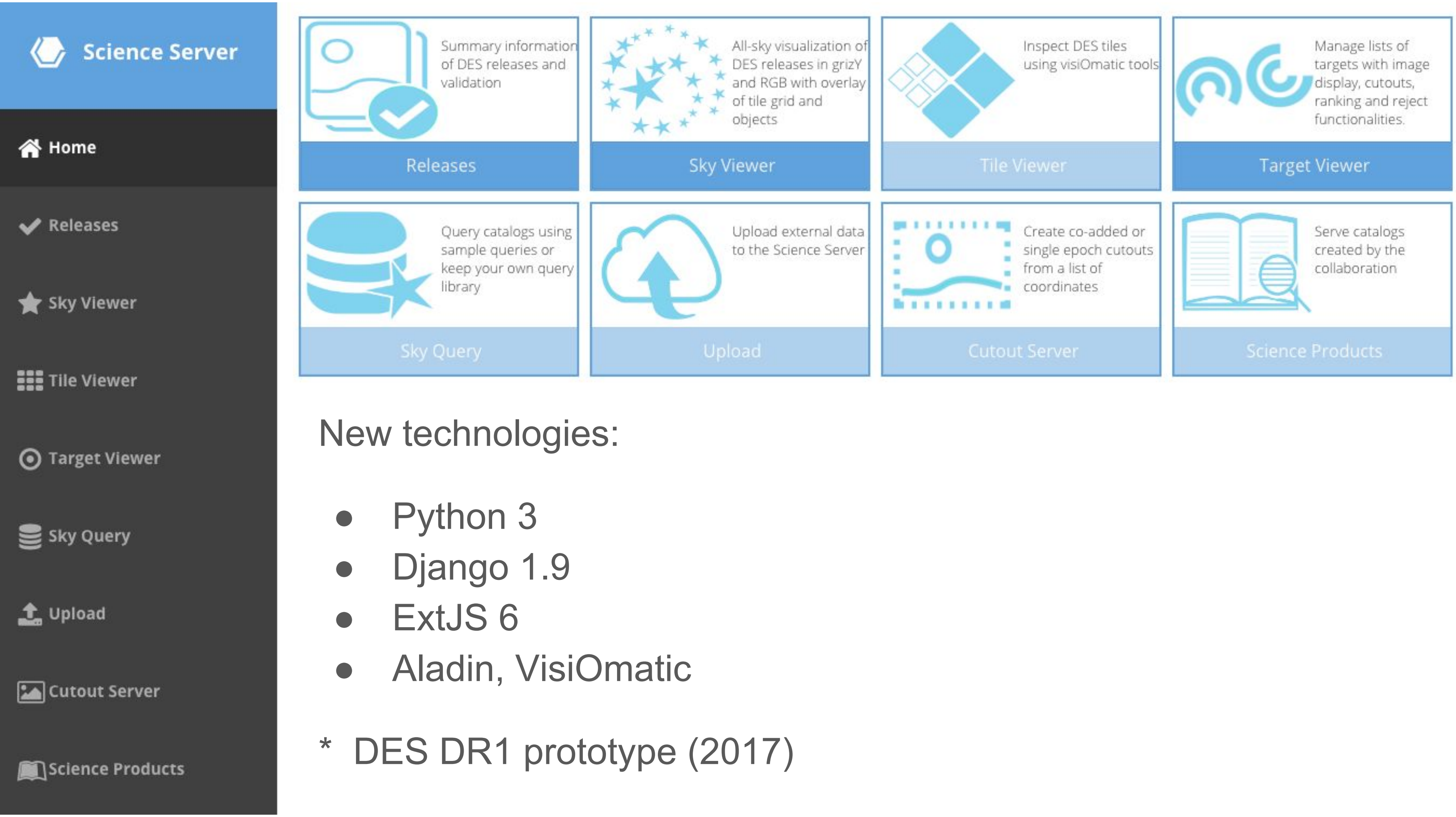
RA (deg): 79.709 Dec (deg): -57.3517

Rating: ★★☆☆☆

Favorite List: Favorite Galaxies

Tiles: 507 Inspected: 53 ( 10.0% ) Blacklisted: 46 ( 87.0% )

# New Science Server @ NCSA*



Summary information of DES releases and validation

Releases

All-sky visualization of DES releases in grizY and RGB with overlay of tile grid and objects

Sky Viewer

Inspect DES tiles using visiOmatic tools

Tile Viewer

Manage lists of targets with image display, cutouts, ranking and reject functionalities.

Target Viewer

Query catalogs using sample queries or keep your own query library

Sky Query

Upload external data to the Science Server

Upload

Create co-added or single epoch cutouts from a list of coordinates

Cutout Server

Serve catalogs created by the collaboration

Science Products

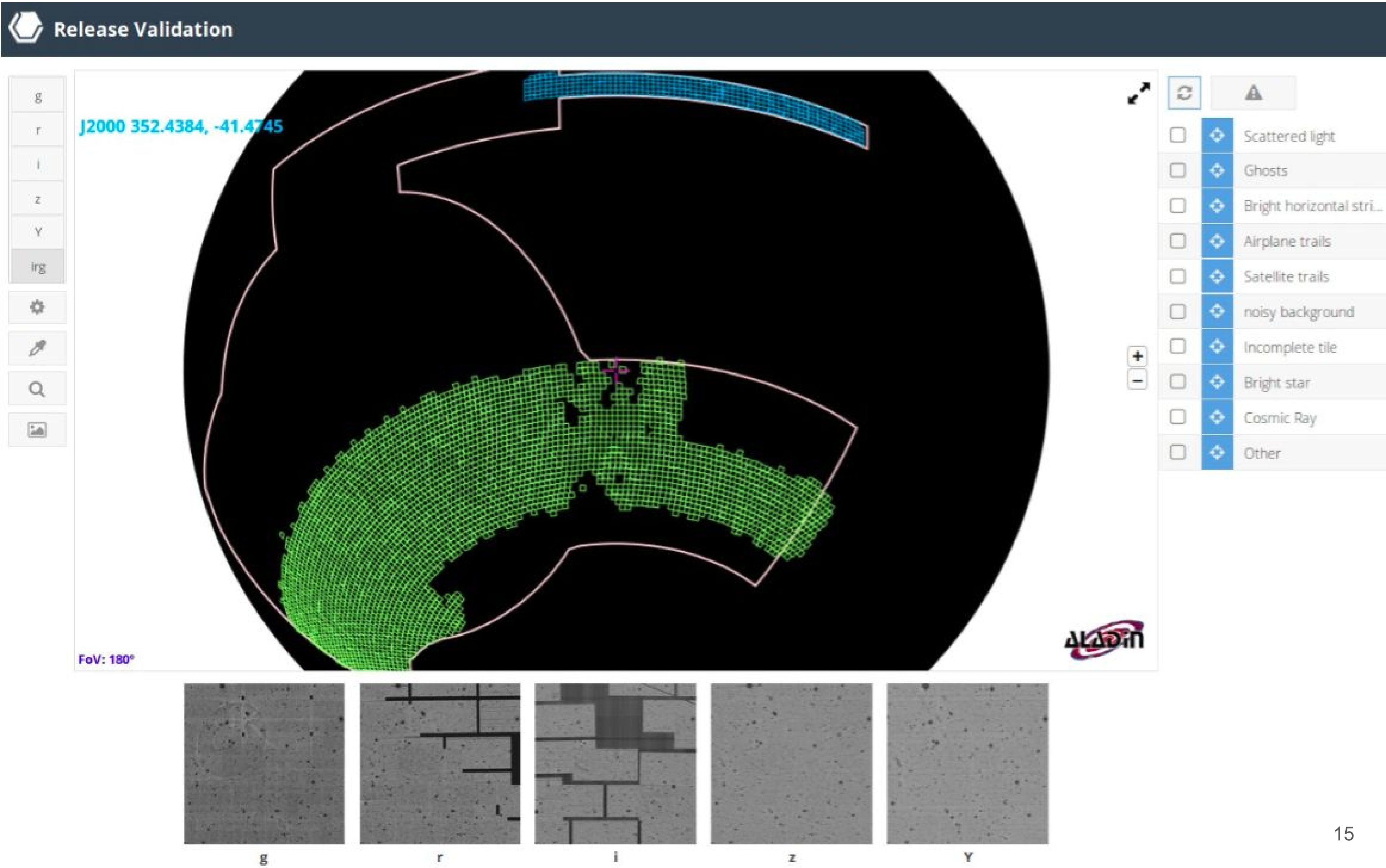
New technologies:

- Python 3
- Django 1.9
- ExtJS 6
- Aladin, VisiOmatic

* DES DR1 prototype (2017)

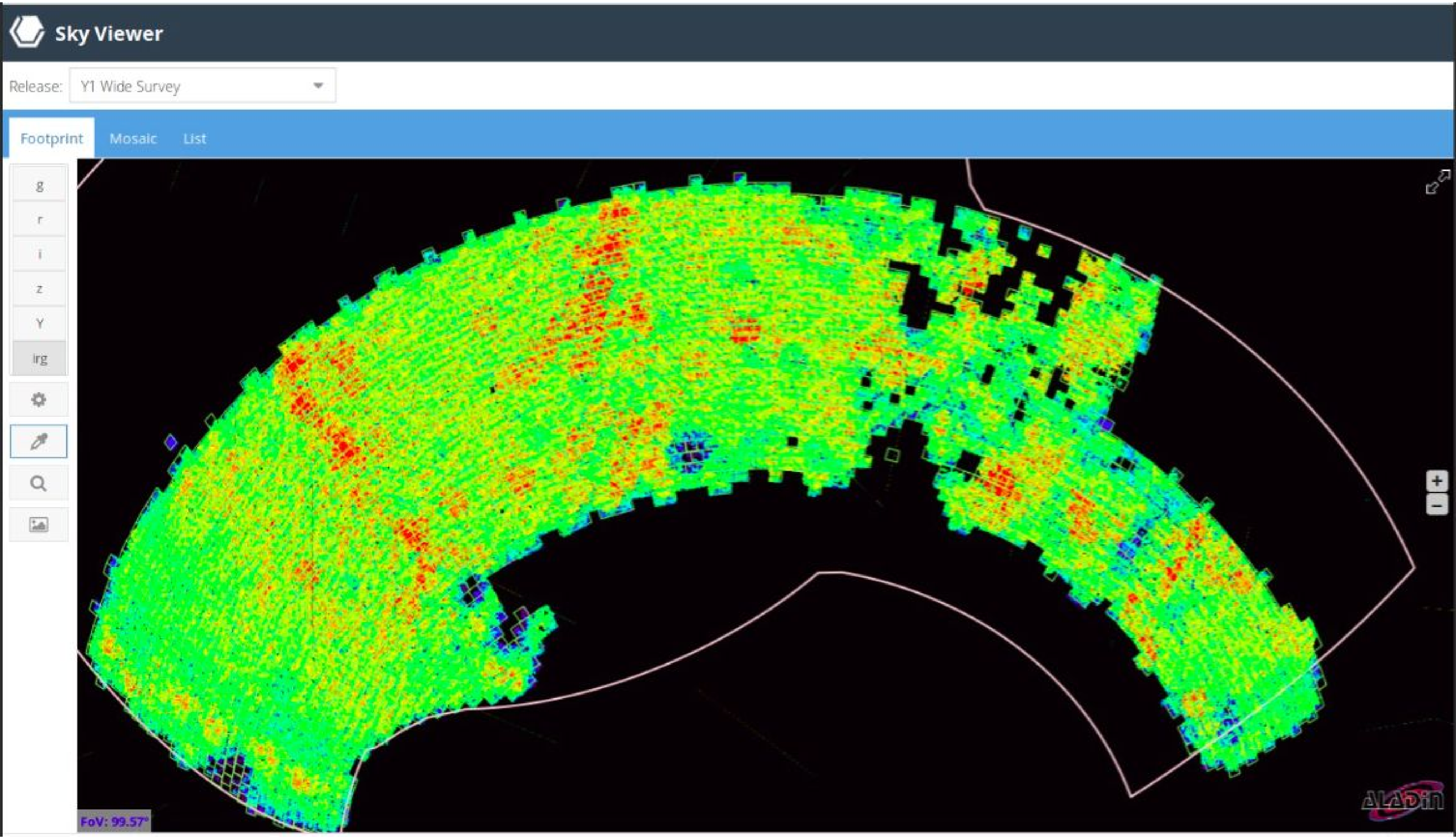
# Data Release Validation

Being used to evaluate DES internal data releases

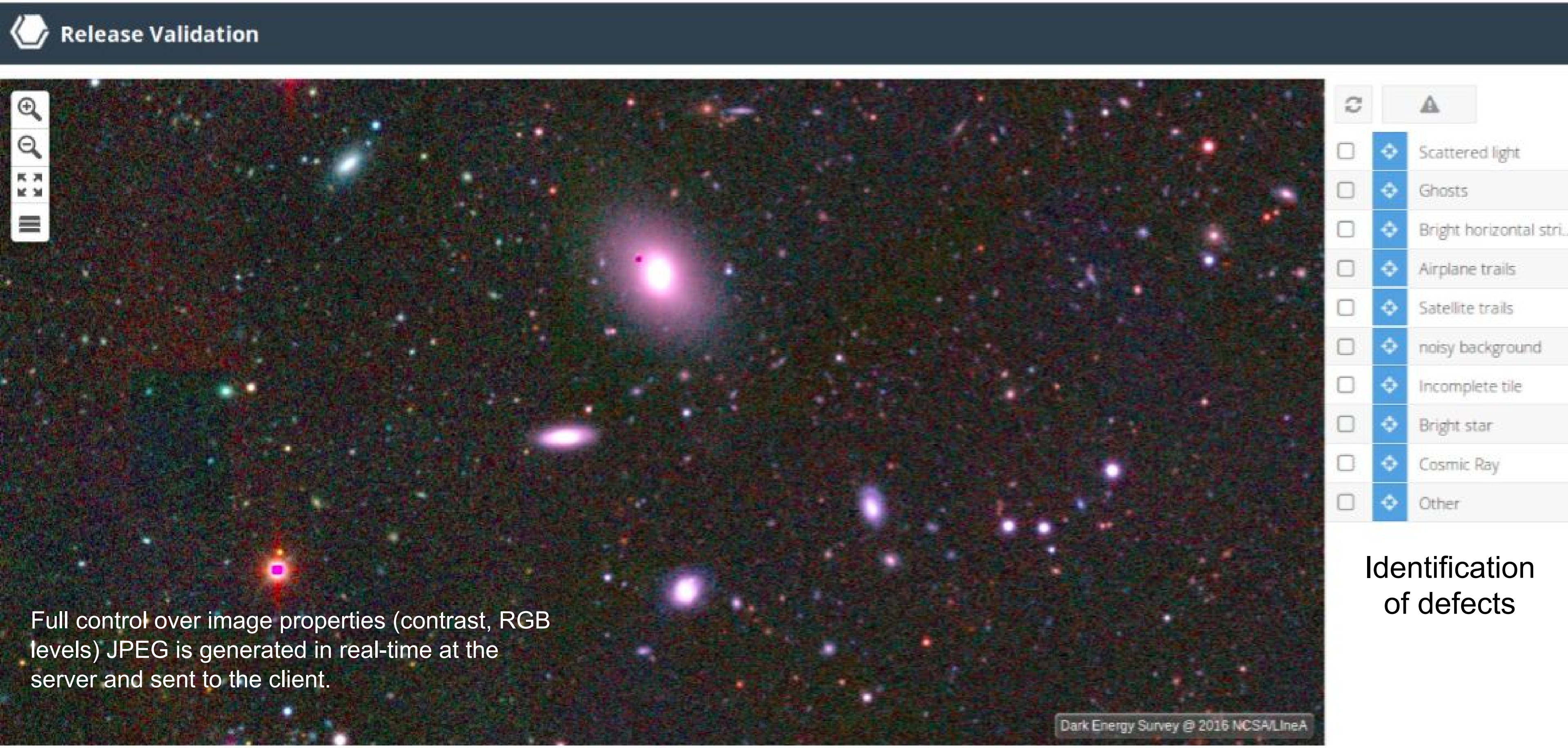


# Visualization of the survey properties

## HEALPix map of the survey magnitude limit

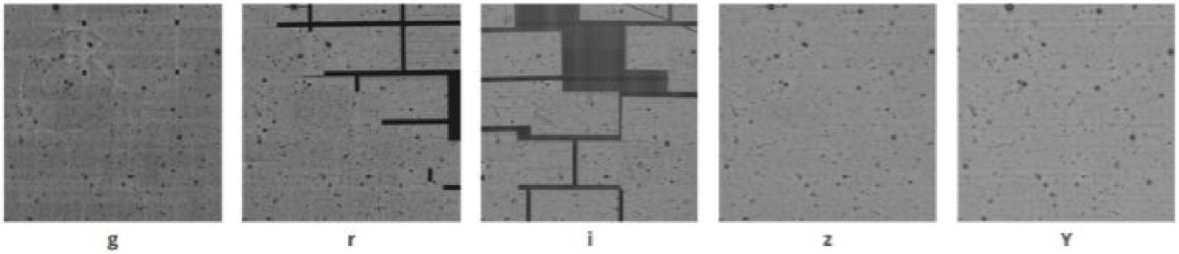


# Detailed image visualization in the browser (visiOmatic  E. Bertin)



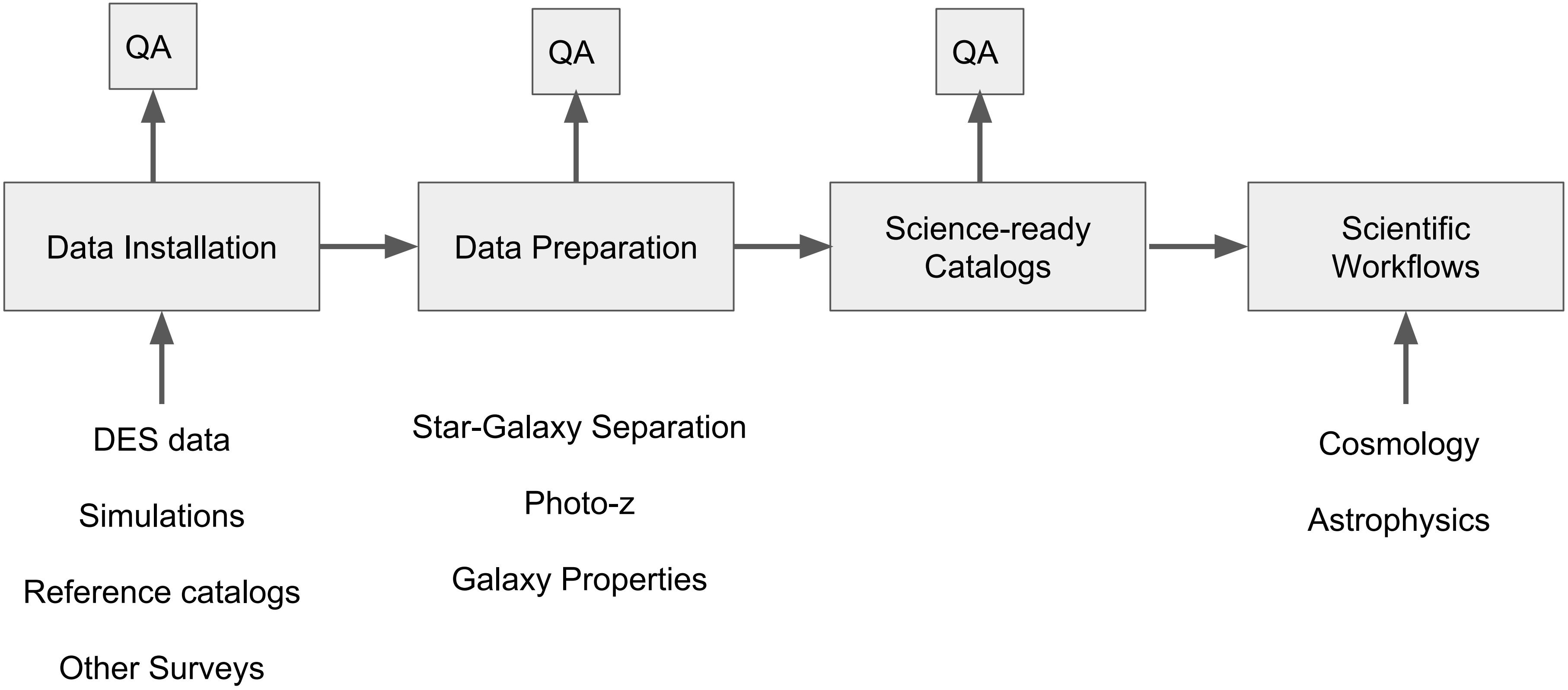
Full control over image properties (contrast, RGB levels) JPEG is generated in real-time at the server and sent to the client.

Identification of defects



grizY thumbnails

# Science Portal: Science Analysis Framework



QA

QA

QA

Data Installation

Data Preparation

Science-ready Catalogs

Scientific Workflows

DES data

Simulations

Reference catalogs

Other Surveys

Star-Galaxy Separation

Photo-z

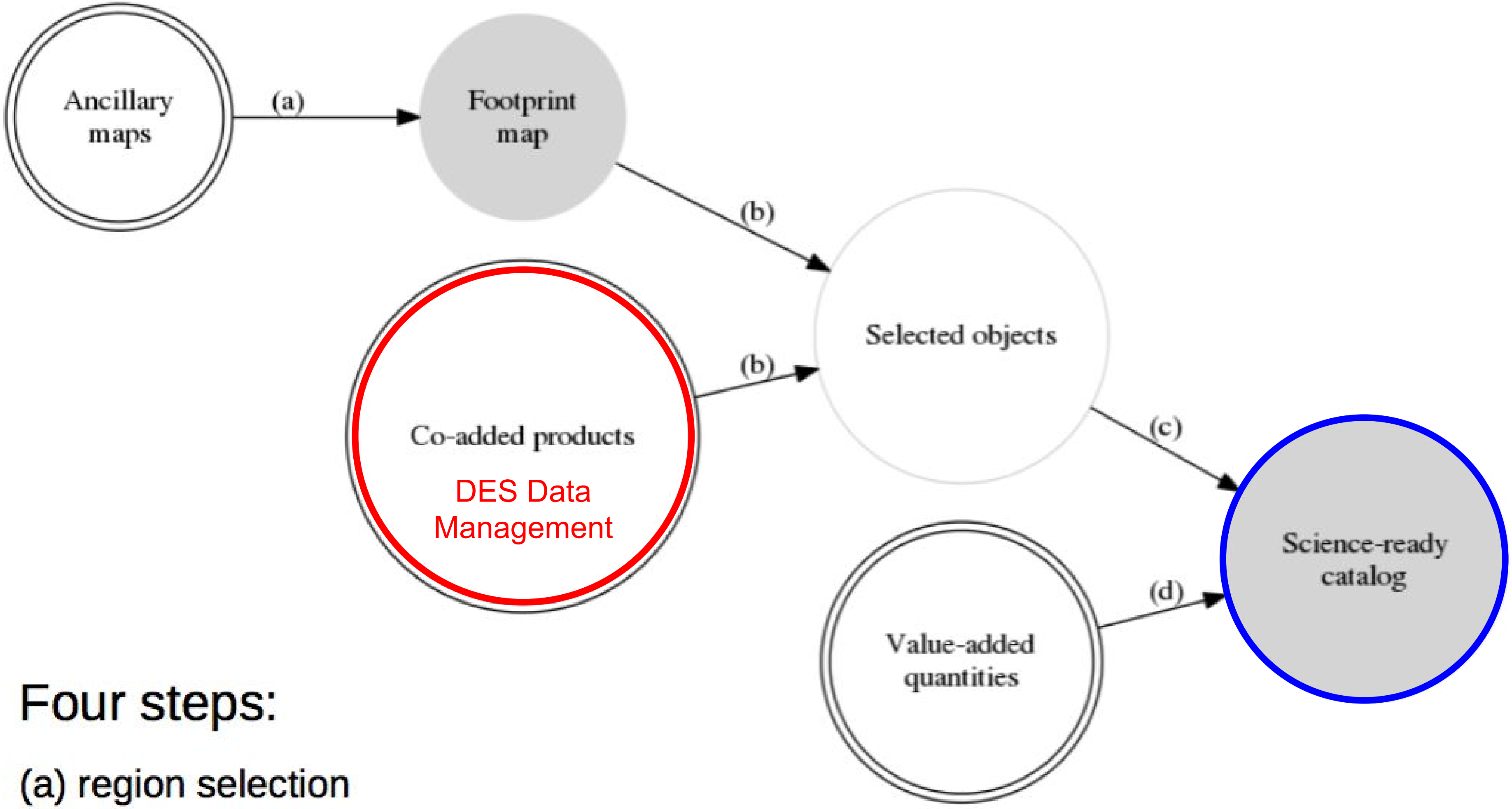
Galaxy Properties

Cosmology

Astrophysics

# Preparation of science-ready catalogs

(Fausti et. al 2016 in prep)



## Four steps:

(a) region selection
(b) object selection
(c) column selection
(d) addition of value-added quantities (sg separation, photo-z, galaxy properties)

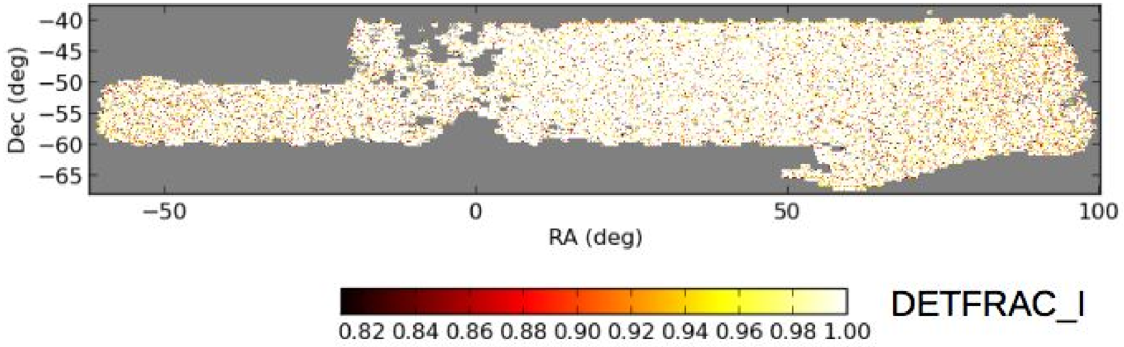
# Why building this infrastructure?

(Fausti et. al 2016 in prep)

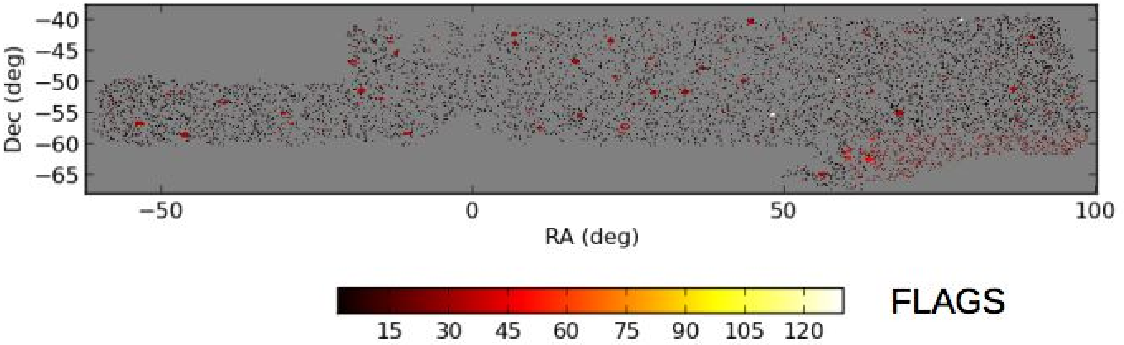
- Preservation of the codes developed by DES Collaboration to create "ancillary products"
- Reproducibility of the catalogs
- Control the parameters used in the creation of the catalogs
- Provenance of the input data products
- Documentation of the catalog and its properties

# Example: galaxy magnitude-limited catalog

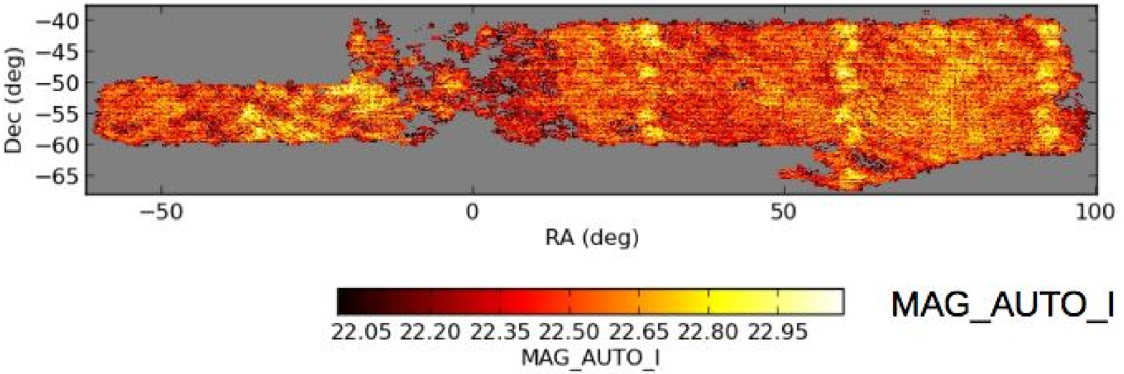
(Fausti et. al 2016 in prep)



"Good" regions for science



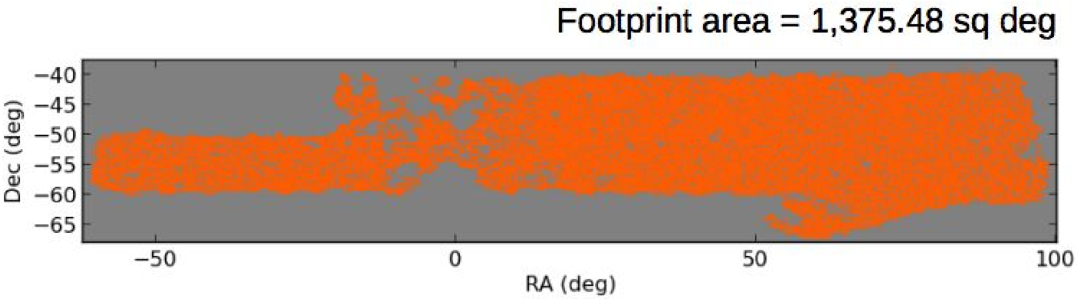
Foreground objects mask



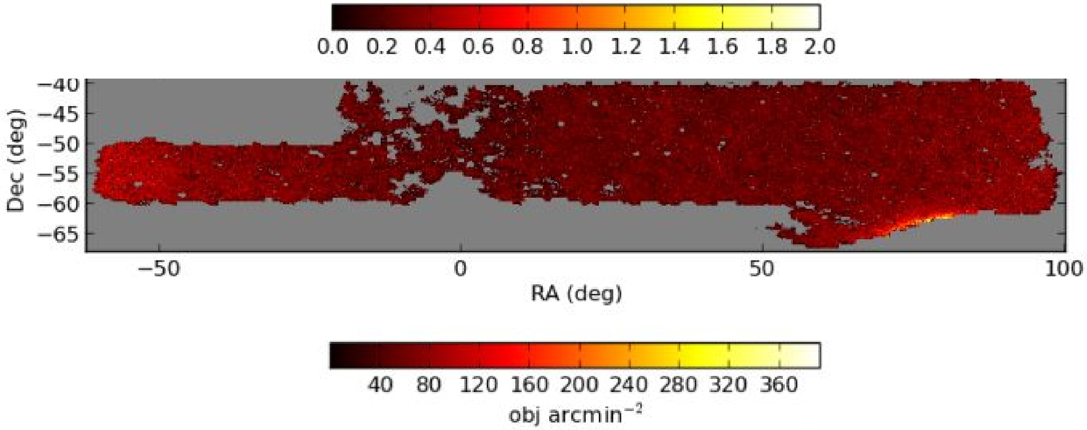
Magnitude limit map

# Preparation of science-ready catalogs

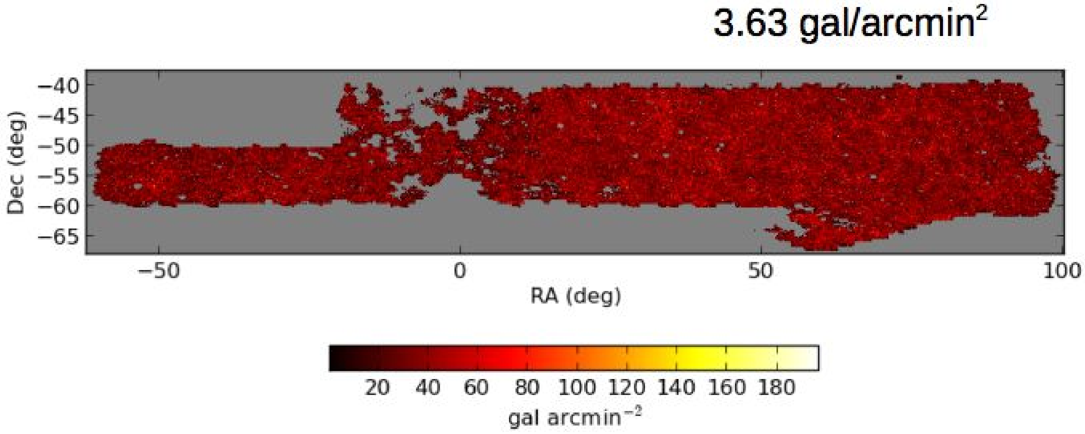
(Fausti et. al 2016 in prep)



Footprint map  (after region selection)



Object Selection

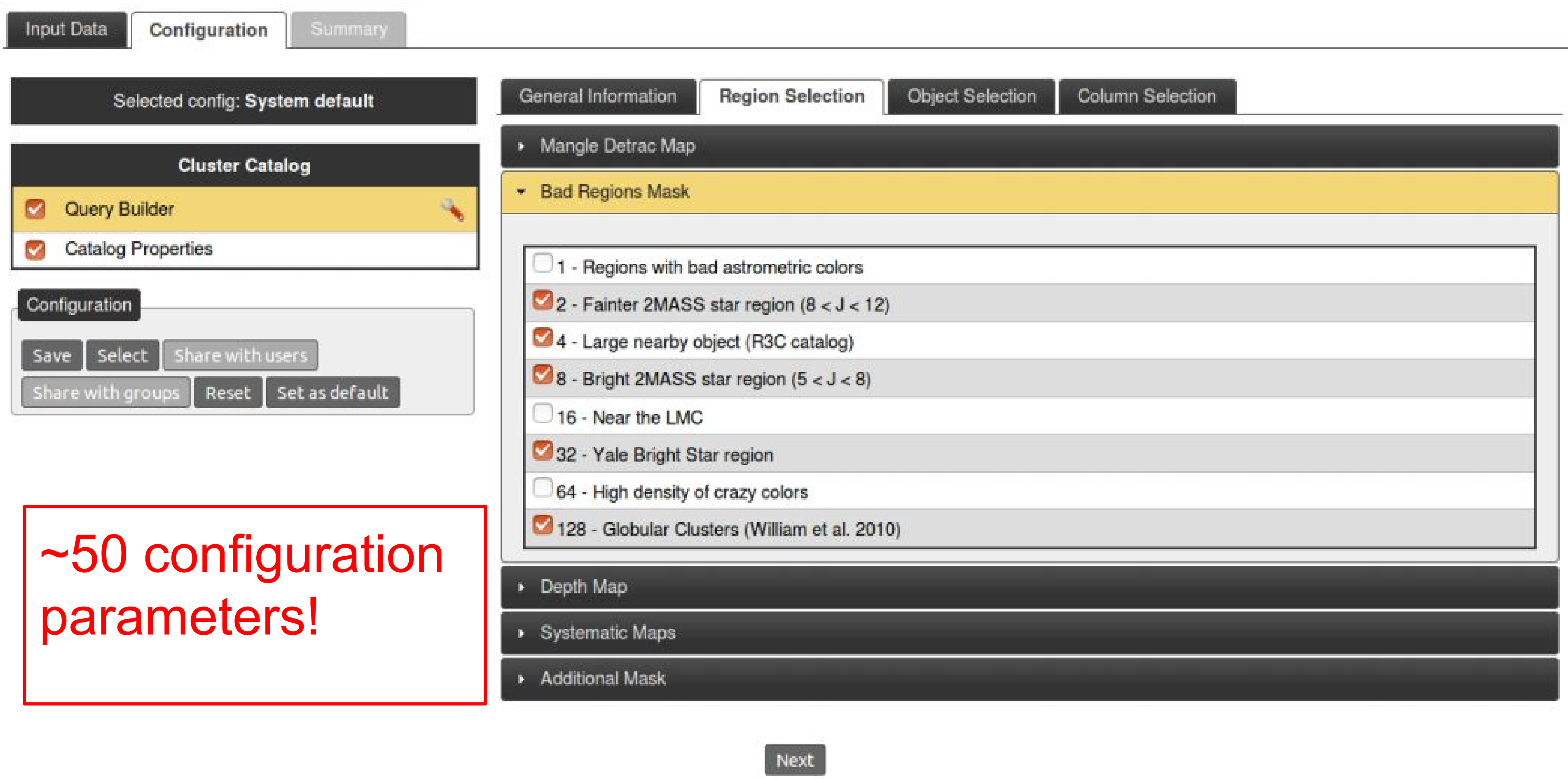


Star-galaxy separation, photo-z

**Science-ready catalog**

# Configuration Interface

## All decisions about the catalog content are made here



☑ Query Builder
☑ Catalog Properties

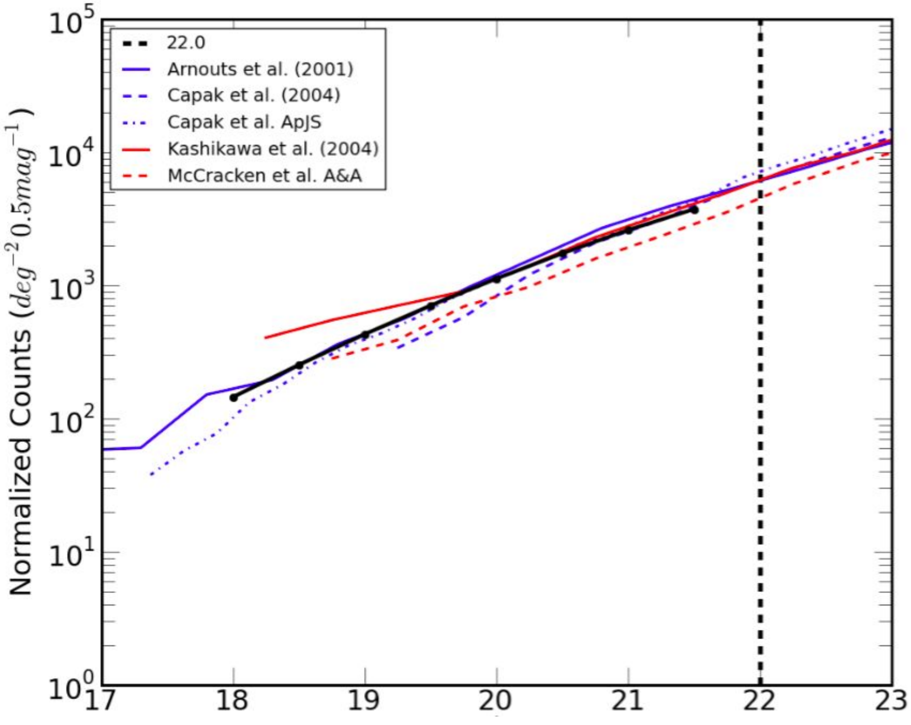
Bad Regions Mask

- □ 1 - Regions with bad astrometric colors
- ☑ 2 - Fainter 2MASS star region (8 < J < 12)
- ☑ 4 - Large nearby object (R3C catalog)
- ☑ 8 - Bright 2MASS star region (5 < J < 8)
- □ 16 - Near the LMC
- ☑ 32 - Yale Bright Star region
- □ 64 - High density of crazy colors
- ☑ 128 - Globular Clusters (William et al. 2010)

~50 configuration parameters!

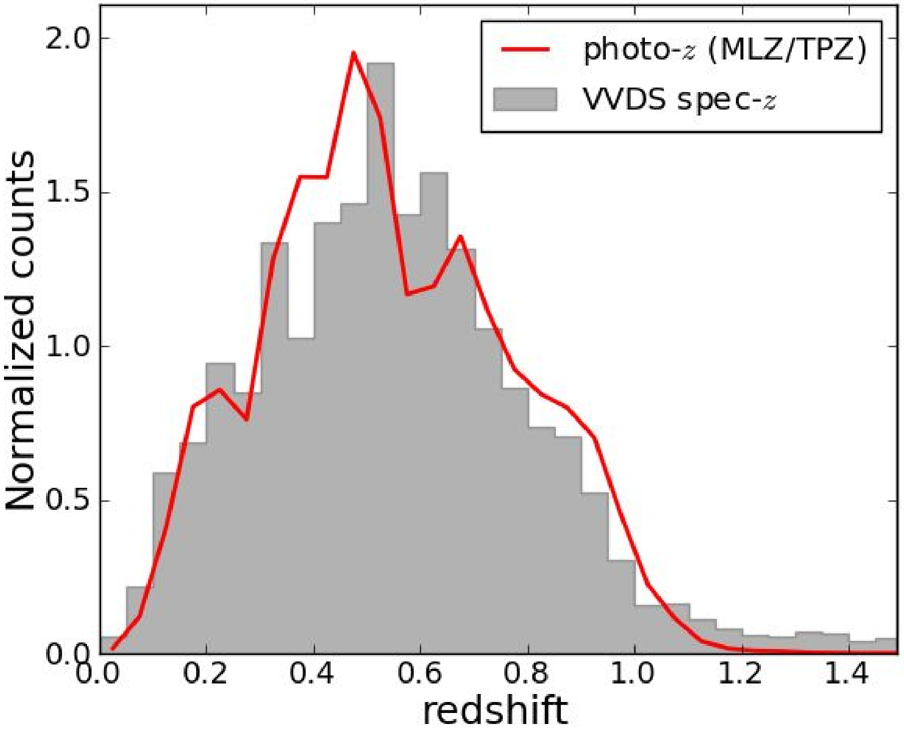
# Catalog properties

Number counts

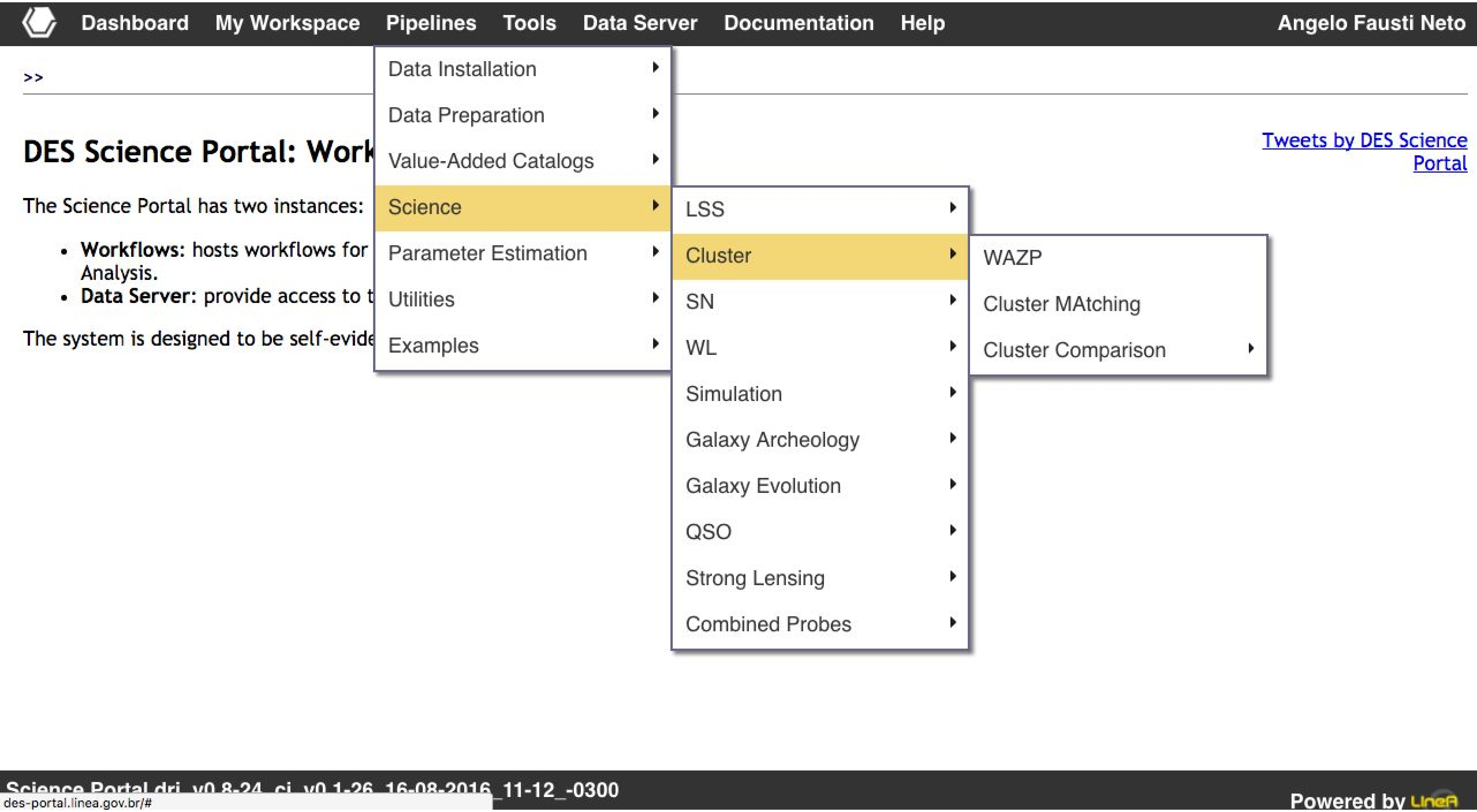


i-band magnitude

photo-z and spec-z distribution



# Scientific Workflows @ LIneA



## DES Science Portal: Work

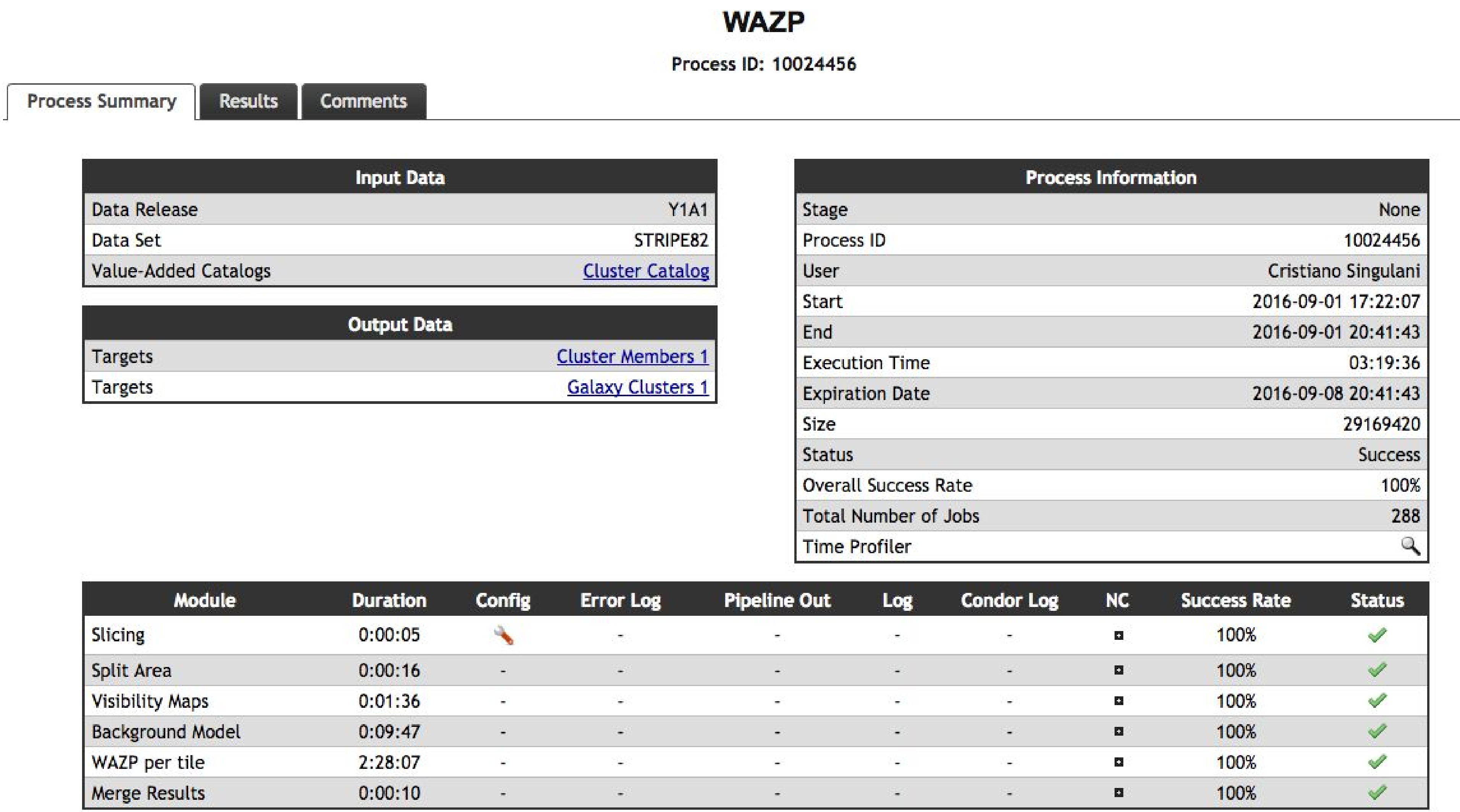
The Science Portal has two instances:

- **Workflows:** hosts workflows for Analysis.
- **Data Server:** provide access to t

The system is designed to be self-evide

Science Portal dri_v0.8-24_ci_v0.1-26_16-08-2016_11-12_-0300

# Example of a cluster finder workflow



## WAZP

**Process ID: 10024456**

| Input Data | |
|---|---|
| Data Release | Y1A1 |
| Data Set | STRIPE82 |
| Value-Added Catalogs | Cluster Catalog |

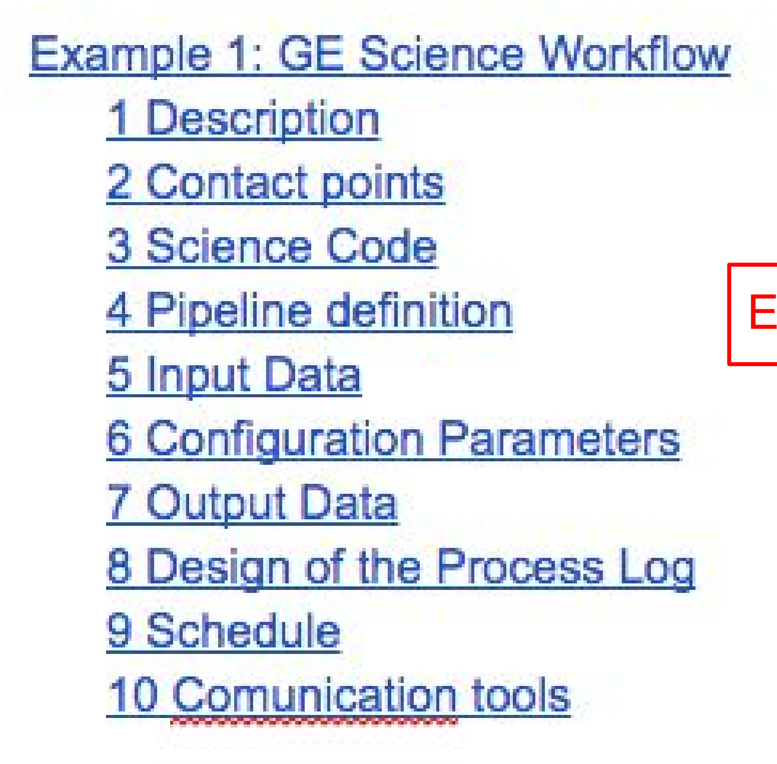
| Output Data | |
|---|---|
| Targets | Cluster Members 1 |
| Targets | Galaxy Clusters 1 |

| Process Information | |
|---|---|
| Stage | None |
| Process ID | 10024456 |
| User | Cristiano Singulani |
| Start | 2016-09-01 17:22:07 |
| End | 2016-09-01 20:41:43 |
| Execution Time | 03:19:36 |
| Expiration Date | 2016-09-08 20:41:43 |
| Size | 29169420 |
| Status | Success |
| Overall Success Rate | 100% |
| Total Number of Jobs | 288 |
| Time Profiler | 🔍 |

| Module | Duration | Config | Error Log | Pipeline Out | Log | Condor Log | NC | Success Rate | Status |
|---|---|---|---|---|---|---|---|---|---|
| Slicing | 0:00:05 | 🔧 | - | - | - | - | ⊞ | 100% | ✓ |
| Split Area | 0:00:16 | - | - | - | - | - | ⊞ | 100% | ✓ |
| Visibility Maps | 0:01:36 | - | - | - | - | - | ⊞ | 100% | ✓ |
| Background Model | 0:09:47 | - | - | - | - | - | ⊞ | 100% | ✓ |
| WAZP per tile | 2:28:07 | - | - | - | - | - | ⊞ | 100% | ✓ |
| Merge Results | 0:00:10 | - | - | - | - | - | ⊞ | 100% | ✓ |

# How  do scientists and developers interact?

Adding a new workflow to the Science Portal



Example 1: GE Science Workflow

1 Description

2 Contact points

3 Science Code

4 Pipeline definition

5 Input Data

6 Configuration Parameters

7 Output Data

8 Design of the Process Log

9 Schedule

10 Comunication tools

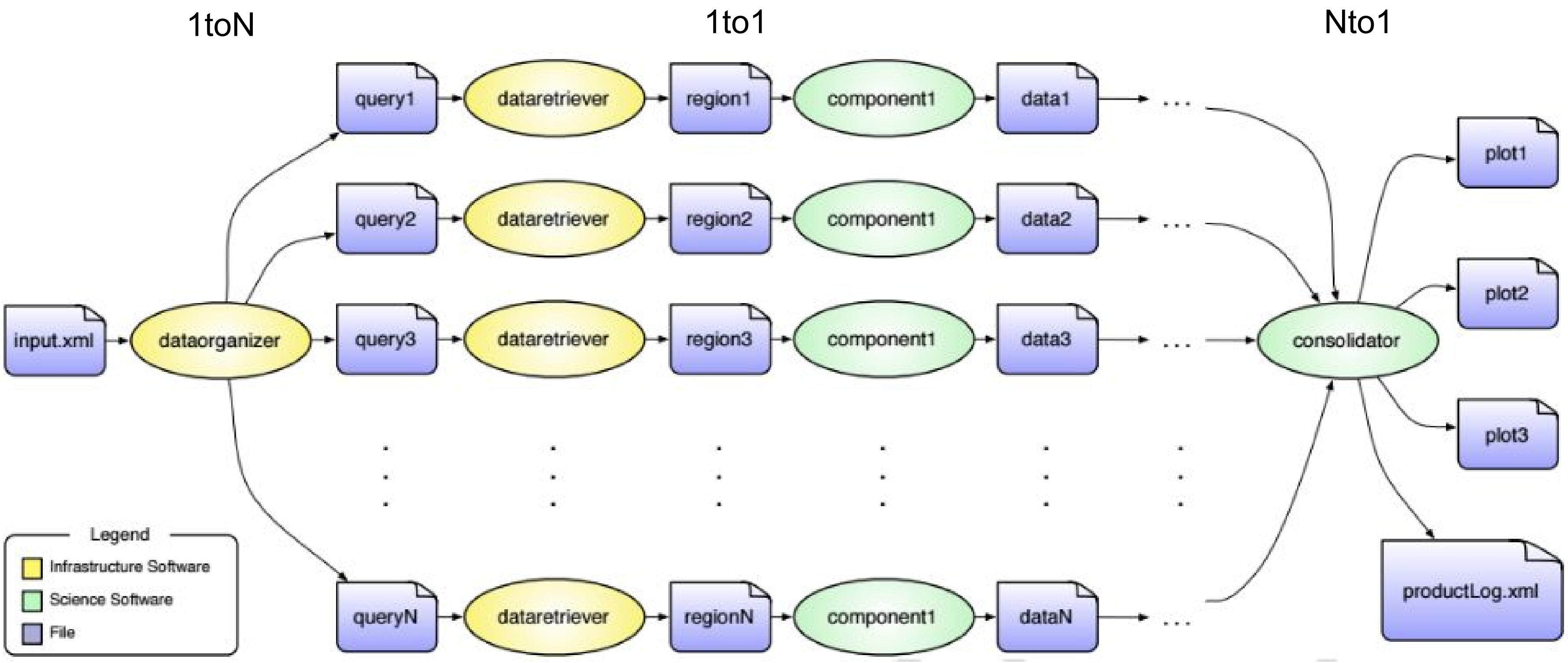
Example of specification document

# Challenges I - Data Processing

- DES Year 1 (~1500 sq deg) objects catalog ~400G
- ~140M objects e ~600 attributes
- Data partitioned in 3,703 files ~100-150M (DES tile)
- Data access during processing
  - PostgreSQL DB ✗
  - Lustre File System ✗
  - Hadoop File System (local data processing) ✓
- DES Year 3 release (Setembro 2016)
  - Objects catalog > 1TB partitioned in 10k files

# Challenges I - Data processing

## Parallel and distributed processing (similar to Map-Reduce)



Implementation problems

- Moving data during processing
- Consolidation of results in a single step

# Challenges II - How to use other resources available?

- SDumont (LNCC/Brazil), FermiGrid , Blue Waters, NERSC
  - Different environments:  PBS, Condor, SLURM, Condor-g
  - Data movement
  - "Big software" complex software and dependencies
- **Science-as-a-service**
  - Science APIs (iPlant/CyVerse)
  - Science Gateways (NERSC)
- **Portability**
  - cloud processing (private or public clouds)
- Federation of private clouds
  - UFCG Distributed systems lab (Francisco Brasileiro)
- OpenStack, AWS
  - LSST/SQuaRE (Frossie Economou)

# Challenges III - Data access and distribution

- Optimized data transfer (RNP/Brazil)
- Science portal integration with DES Science DB (NCSA)
    - Data access interface (large variety of products)
    - Documentation
- Science Server  DES DR1
- LSST/DM Data Acess Center (DAC) prototype at  LIneA

# Conclusions and perspectives

- LIneA: support brazilian participation in DES, SDSS, DESI and LSST
- Science Portal: necessary infrastructure for efficient science analysis
- DES Public DR1 (2017)
- LSST first light in 2020 start operations in 2022
- DES as a “prototype” for LSST

# Extra Slides

# The Science Portal and DES

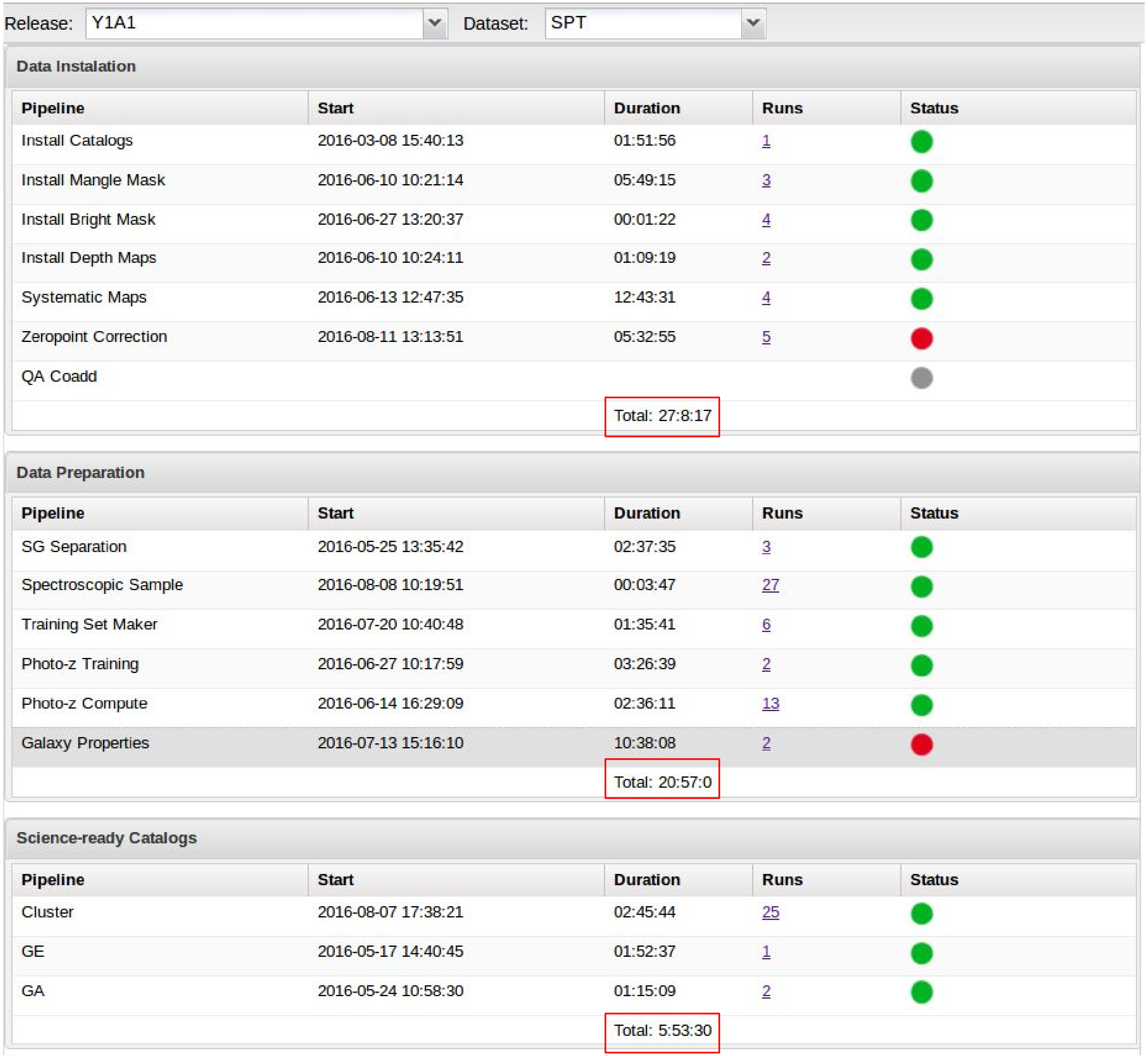
http://www.linea.gov.br/

- Software development started in 2007*
- 9 years! about 56 FTEs
- 8 international reviews

| Emphasis | When | Where |
|---|---|---|
| Introduction, Science Workflows | Oct 2010 | Fermilab |
| Precam, Quick Reduce, Science Workflows | Oct 2011 | UPenn |
| Quick Reduce | May 2012 | MPA |
| End-to-end vision, data preparation | Jul 2013 | Fermilab |
| Data validation | Nov 2013 | Fermilab |
| Data validation and exploration | Ago 2014 | Fermilab |
| Data validation, exploration and science-ready catalogs | Nov 2014 | NCSA |
| Science-ready catalogs | May 2015 | Fermilab |

* https://youtu.be/1Qv8HOoeUF4

# Monitoring the execution of all processes involved



Release: Y1A1 Dataset: SPT

**Data Instalation**

| Pipeline | Start | Duration | Runs | Status |
|---|---|---|---|---|
| Install Catalogs | 2016-03-08 15:40:13 | 01:51:56 | 1 | 🟢 |
| Install Mangle Mask | 2016-06-10 10:21:14 | 05:49:15 | 3 | 🟢 |
| Install Bright Mask | 2016-06-27 13:20:37 | 00:01:22 | 4 | 🟢 |
| Install Depth Maps | 2016-06-10 10:24:11 | 01:09:19 | 2 | 🟢 |
| Systematic Maps | 2016-06-13 12:47:35 | 12:43:31 | 4 | 🟢 |
| Zeropoint Correction | 2016-08-11 13:13:51 | 05:32:55 | 5 | 🔴 |
| QA Coadd | | | | ● |
| | | Total: 27:8:17 | | |

**Data Preparation**

| Pipeline | Start | Duration | Runs | Status |
|---|---|---|---|---|
| SG Separation | 2016-05-25 13:35:42 | 02:37:35 | 3 | 🟢 |
| Spectroscopic Sample | 2016-08-08 10:19:51 | 00:03:47 | 27 | 🟢 |
| Training Set Maker | 2016-07-20 10:40:48 | 01:35:41 | 6 | 🟢 |
| Photo-z Training | 2016-06-27 10:17:59 | 03:26:39 | 2 | 🟢 |
| Photo-z Compute | 2016-06-14 16:29:09 | 02:36:11 | 13 | 🟢 |
| Galaxy Properties | 2016-07-13 15:16:10 | 10:38:08 | 2 | 🔴 |
| | | Total: 20:57:0 | | |

**Science-ready Catalogs**

| Pipeline | Start | Duration | Runs | Status |
|---|---|---|---|---|
| Cluster | 2016-08-07 17:38:21 | 02:45:44 | 25 | 🟢 |
| GE | 2016-05-17 14:40:45 | 01:52:37 | 1 | 🟢 |
| GA | 2016-05-24 10:58:30 | 01:15:09 | 2 | 🟢 |
| | | Total: 5:53:30 | | |

## Stages

- Data Installation
- Data Preparation
- Science-ready catalogs

- 16 workflows
- 64 data products