

# O Portal Científico

Uma ferramenta para analisar dados de grandes levantamentos astronômicos

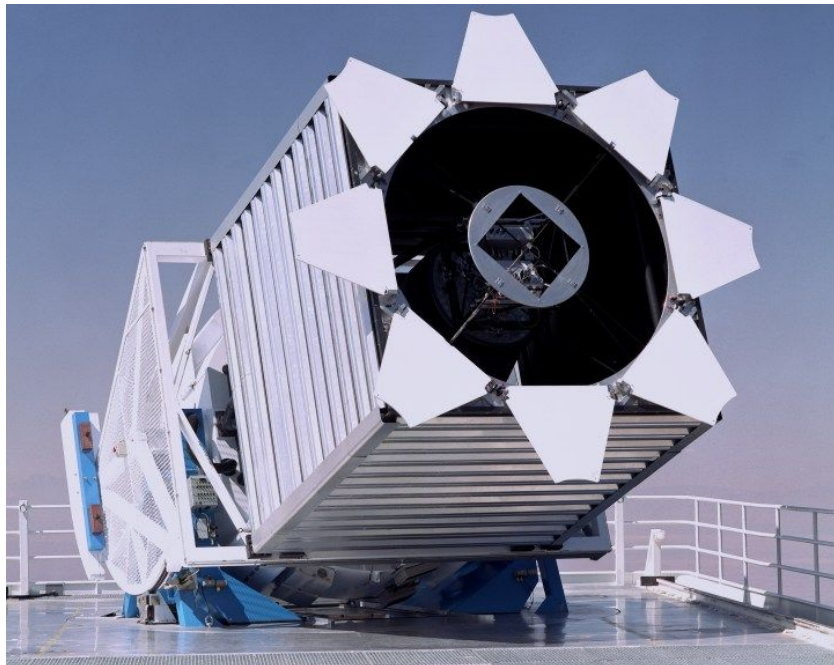
Angelo Fausti Neto  
LineA/LSST

# Sumário

- A era dos grandes levantamentos astronômicos
  - SDSS, DES, LSST
- O papel do LIneA
- A importância do Portal Científico
  - O que é?
  - Aplicações
  - Desafios
- Conclusão e perspectivas futuras

# Sloan Digital Sky Surveys (SDSS)

<http://www.sdss.org/>



SDSS 2.5m telescope at Apache Point Observatory

**O décimo terceiro lançamento de dados do SDSS**  
02 de agosto de 2016

**RNP bate recorde de transferência de dados do SDSS**  
29 de julho de 2016

## Levantamentos:

- SDSS I/II (2000-2008)
- SDSS III (2008-2014)
- SDSS IV (2014-2020)
  - APOGEE 2, eBOSS, MaNGA

## Áreas de estudo:

- Cosmologia, Quasares, Galaxias, Via-Láctea, Estrelas e Sistema Solar

## Brazilian Participation Group (2006)

## Impacto:

- 5.800 artigos
- 245.000 citações

## Site espelho do SDSS mantido pelo LInEA (DR8-DR13)

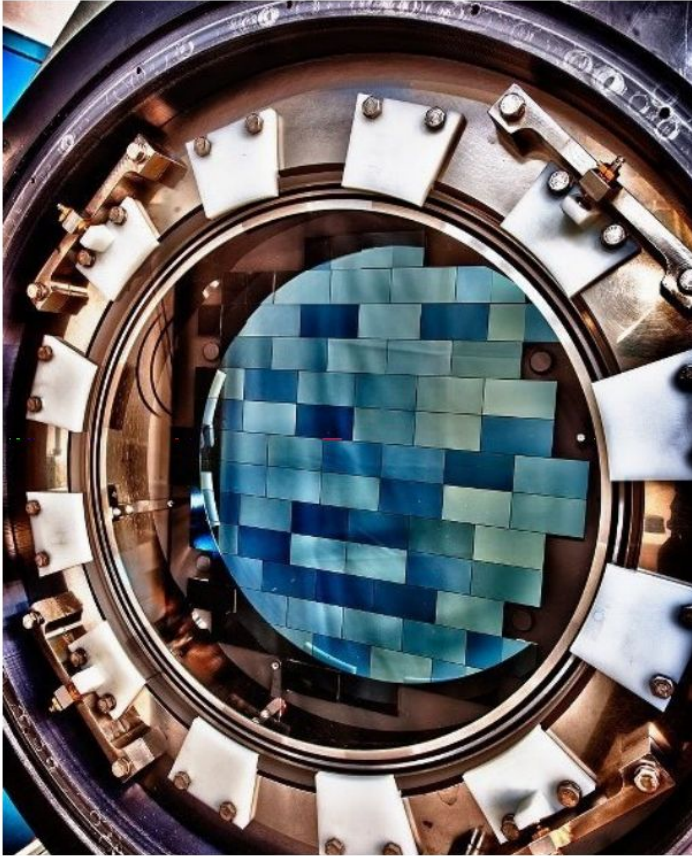
- Base de dados de 15TB
- Sky Server, CASJobs

## Suporte a operação do APOGEE 2 em Las Campanas, Chile

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

# Dark Energy Survey (DES)

<https://www.darkenergysurvey.org/>



Dark Energy Camera, em operacao desde 2012 no telescópio Blanco (CTIO)

- Levantamento fotométrico (grizY)
- DECam 570Mpixels (62 CCDs)
- Blanco 4m (Tololo)
- ~ 300 exposições por noite (500GB)
- ~100 noites/ano durante 5 anos
- 5.000 sq deg
- 4° ano de operação (Ago 2016)
- Data Management ~8 FTEs (13 pessoas)
- DR1 (público) em Set 2017



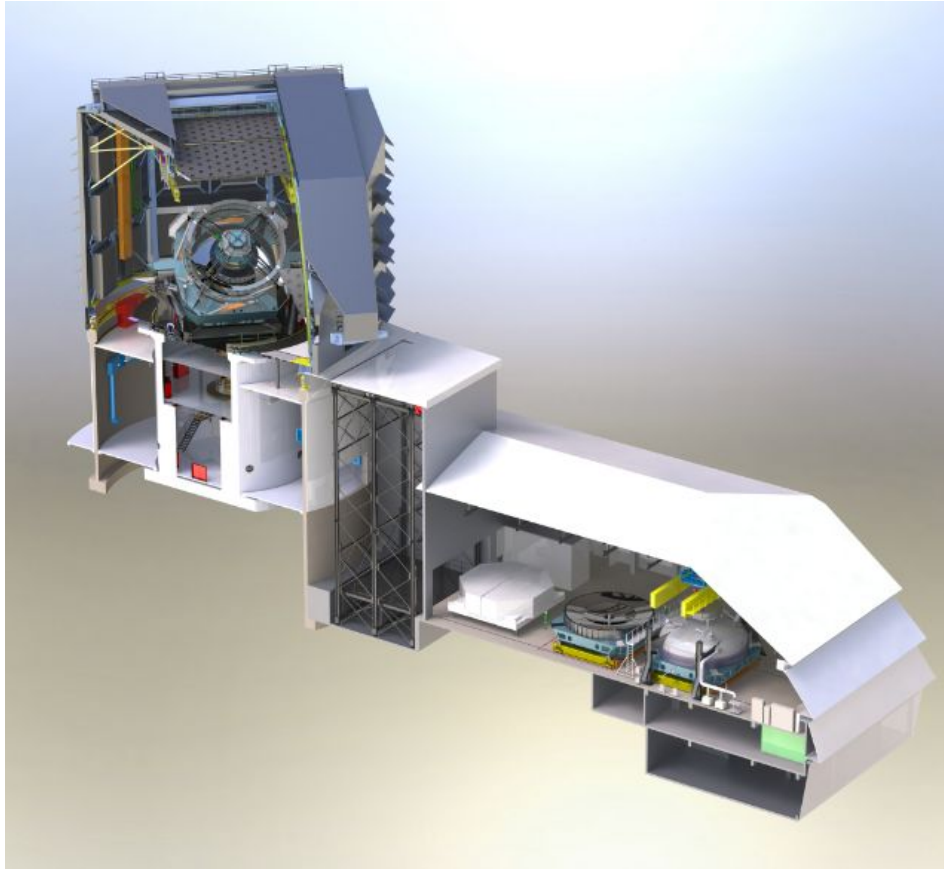
**Começa o quarto ano de observações do Dark Energy Survey**

19 de agosto de 2016

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

# Large Synoptic Survey Telescope (LSST)

<https://www.lsst.org/scientists>



- Levantamento fotométrico (ugrizy)
- Camera de 3.2 Gpixels
- Telescópio de 8.4m (Pachón)
- 1 "visit" = 2 exposicoes a cada 30s
- 1000 "visits" por noite
- ~15 TB por noite (SDSS a cada 2 noites)
- 18.000 sq deg (hemisfério sul celeste 2 vezes por semana por 10 anos)
- Difference imaging
- Deep coadds
- Data Management 54 FTEs (70 pessoas)
- Inicio das operações em 2022



Site do LSST Brazilian Participation Group entra no ar

05 de julho de 2016

# LSST "Data Products and Capabilities"

- Level 1: ~10 milhões de alertas por noite (detectados e transmitidos a cada 60s!). Um catálogo de parâmetros orbitais para ~6 milhões de corpos do Sistema Solar.
- Level 2: Os dados são reprocessados anualmente e os catálogos disponibilizados através de um sistema de banco de dados online. DR11 um catálogo com ~37 bilhões de objetos (20B de galáxias, 17B de estrelas) ~7 trilhões de fontes, 30 trilhões de "fontes forçadas" 5.5 milhões de exposições
- Level 3: "produtos" derivados de L1 e L2; serviços e infraestrutura computacional nos Centros de Acesso aos Dados (ou DACs) que permitam o processamento e análise de dados. Software e APIs para o desenvolvimento dos códigos de análise.

## Fontes:

- <https://www.lsst.org/scientists/keynumbers>
- LSST Data Products Definition Document (<http://ls.st/dpdd>) (In review)

# Os desafios em comum

- **Grandes questões científicas**
  - Pesquisadores com formação sólida e polivalente
  - Participação pró-ativa em colaborações internacionais
- **Grandes colaborações**
  - Mudança cultural na forma de trabalho (**ferramentas colaborativas**)
  - Comunicação (**informação distribuída**, inúmeras *telecons*)
  - Compreensão do projeto e das suas oportunidades
- **Grande volume e variedade de dados**
  - **Infraestrutura computacional** necessária para transferência, armazenamento e processamento de dados
  - **Ferramentas** para **preparação e análise dos dados** de forma eficiente
  - Domínio das técnicas da "**ciência de dados**"

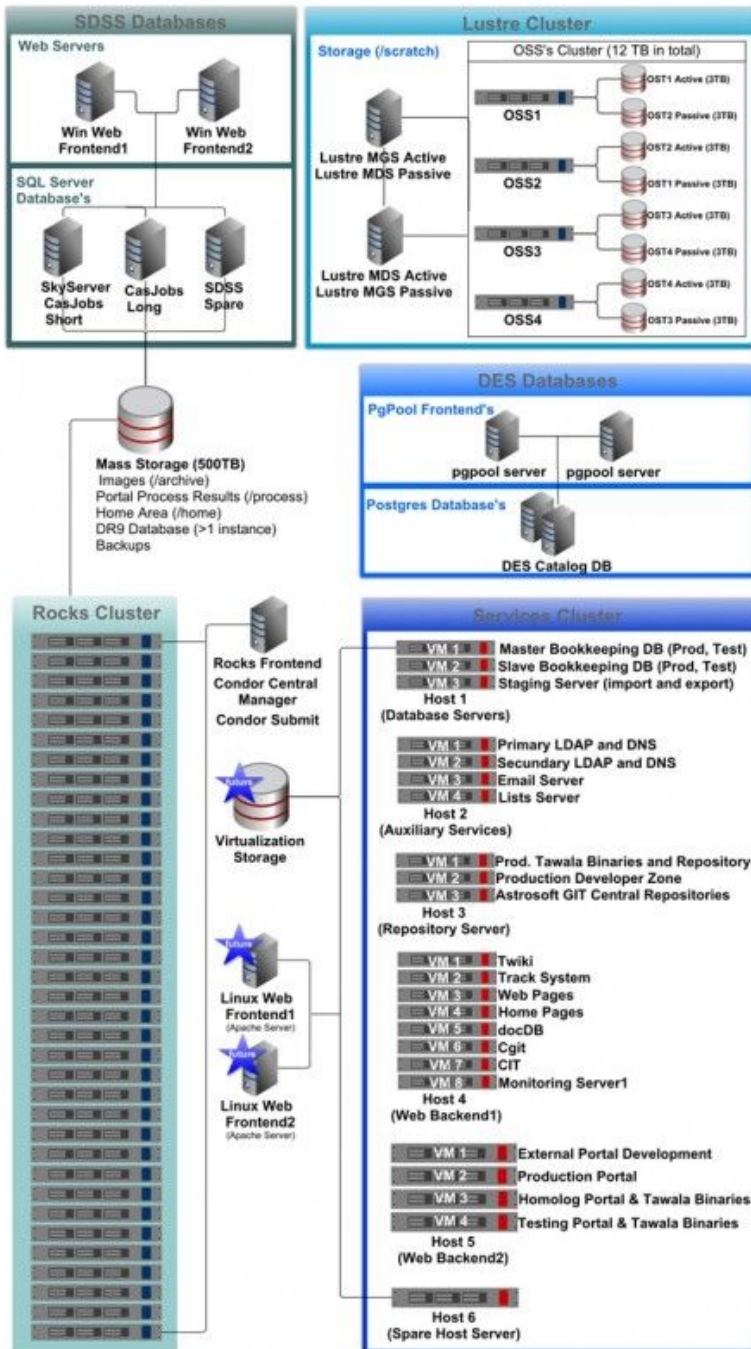
# O papel do LIneA

- Apoio a ciência
  - SDSS BPG, DES-Brazil, LSST BPG
  - 71 afiliados (ON, UFRJ, USP, UNESP, UNICAMP, UFABC, UFRGS, UFSM)
  - Formação de pesquisadores e tecnologistas
  - Divulgação científica
- Operação do centro de dados
  - Manutenção do hardware e serviços (SLACAM)
    - helpdesk, e-mail, twiki, git, doc-db, slack, etc
  - Transferência de dados (RNP)
  - Armazenamento e processamento de dados (LNCC)
- Desenvolvimento do Portal Científico
  - P&D
  - Manutenção do time de desenvolvimento 9 FTEs (R\$1.8M/ano)
  - 3 pesquisadores + contribuições externas (precisamos crescer)



# O centro de dados do LineA

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



- SGI Cluster
  - 912 cores
  - 24 cores per node, 4GB per core
  - 38 nodes
- INCT e-Universo e FINEP (US\$ 2M + R\$500k/ano manutenção)
  - Novo cluster de 2000 cores
  - "QServ" distribuição de dados
- Lustre cluster (10TB)
- Banco de dados SDSS (30TB) and DES (20TB)
- SGI Storage (500 + 350TB)
- VMs para os serviços do LineA

# O que é o Portal Científico?

- É um framework web que facilita a integração de códigos de análise na forma de "workflows científicos" com acesso a um cluster de processamento e a uma base de dados centralizada.
- É também o conjunto de ferramentas de software desenvolvido pelo LIneA para validação e exploração dos dados de grandes levantamentos astronômicos.

# O Portal Científico e o DES

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

- Início do desenvolvimento de software em 2007\*
- 9 anos! 56 FTEs
- Oito avaliações internacionais

<b>Ênfase</b>	<b>Data</b>	<b>Local</b>
Introdução, Workflows Científicos	Outubro 2010	Fermilab
Precam, Quick Reduce, Workflows Científicos	Outubro 2011	UPenn
Quick Reduce	Maio 2012	MPA
Visão end-to-end e validação de dados	Julho 2013	Fermilab
Validação de dados	Novembro 2013	Fermilab
Validação e exploração de dados	Agosto 2014	Fermilab
Validação, exploração de dados e catálogos científicos	Novembro 2014	NCSA
Preparação de catálogos científicos	Maio 2015	Fermilab

\* <https://youtu.be/1Qv8HOoeUF4>

# Atividades do Portal Científico

- Análise da qualidade dos dados do DES em tempo real
  - Quick Reduce (CTIO) 2012
  - <http://quick1.ctio.noao.edu:8080/>
- Validação e exploração de dados
  - Science Server (Fermilab em 2014, migração para o NCSA)
  - <https://des-portal.fnal.gov/>
  - <http://desportal.cosmology.illinois.edu/>
- Preparação de catálogos e workflows científicos
  - Science Portal (LIneA) 2016
  - <http://des-portal.linea.gov.br/>



# Monitoramento das observações do DES

Acesso a colaboração: transferência diária dos resultados do QR para o Fermilab

Observing History  
Night: 2016-08-16



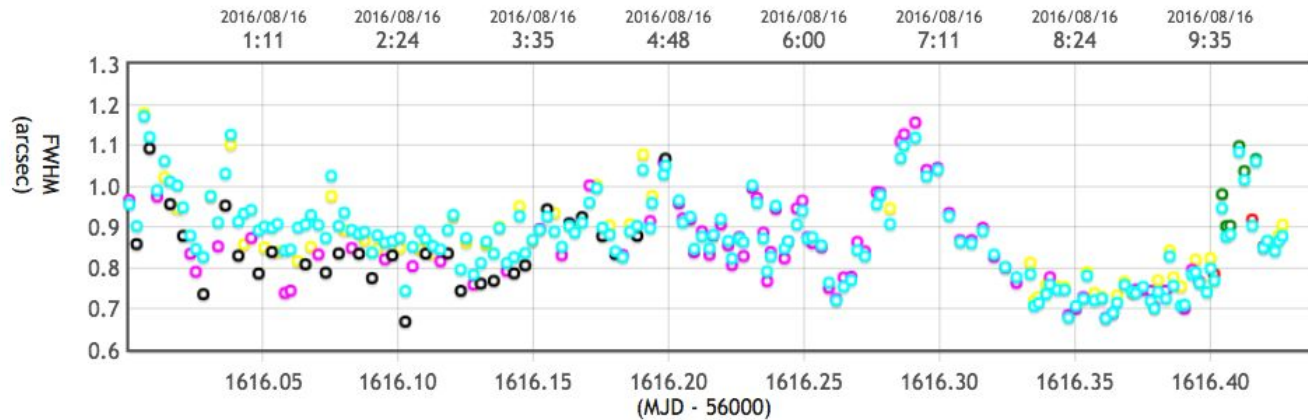
PSF Background Footprint Statistics Trend Analysis

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

**Axis Limits**

FWHM max: 2.1  
FWHM min: 0.1

Auto Scale



**Filters:**

- All
- u
- g
- r
- i
- z
- Y

**Data:**

- IH
- DIMM



## Serviços disponíveis

- Upload
- Visualização de imagens e catálogos
- Cutout
- User query
- Download

"CASJobs moderno"

### Login

Use your FNAL services username and password.

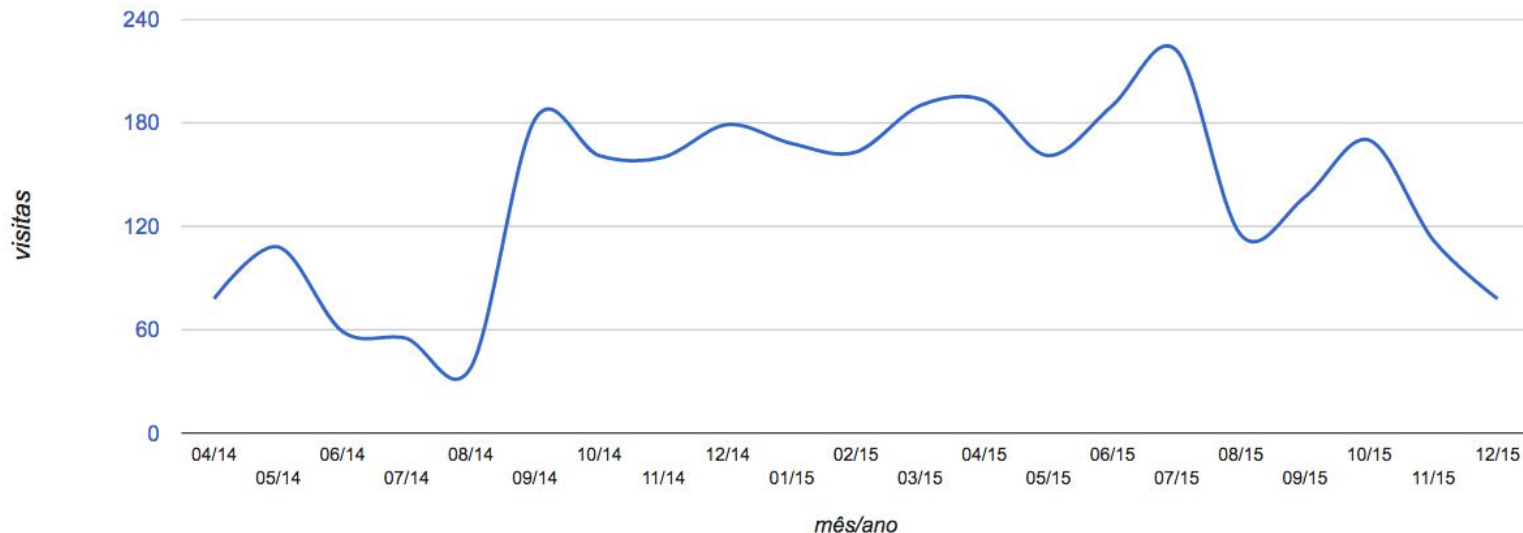
Username:

Password:

[Need Help?](#)

Atende em média 180 membros da colaboração do DES mensalmente.

### Visitas Fermilab



# Ferramenta de cutout e inspeção visual de objetos

## Exemplos: Strong Lening, Galaxy Clusters

The screenshot displays the 'Tile Viewer' application interface. At the top, the 'Release' is set to 'v0.5 ( SVA1\_COADD )' and the 'Field' is 'SPT-E'. A search bar contains 'eg. 307.0658, -52.6783'. Below the search bar are tabs for 'Footprint', 'Tile Mosaic', 'Tile List', 'Targets', 'Favorites', and 'Gallery'. The 'Targets' tab is active, showing a 'Target Mosaic' of six galaxy cluster cutouts. Each cutout is labeled with its ID, RA, and Dec, and includes a 10'' scale bar. The cutouts are arranged in a 2x3 grid. The top row cutouts are: DES0517-5705 (RA: 79.689, Dec: -57.349), DES0454-6205 (RA: 74.192, Dec: -62.415), and DES0503-6122 (RA: 75.562, Dec: -61.231). The bottom row cutouts are: DES0501-6039 (RA: 74.597, Dec: -60.719), DES0456-5928 (RA: 73.679, Dec: -59.931), and DES0454-5831 (RA: 73.457, Dec: -59.400). A tooltip for the top-right cutout shows 'RA:75.5624 Dec:-61.2314'. On the left, the 'Catalogs' panel shows a list of catalogs, with 'Strong Lensing - 1.0' selected. On the right, the 'Target' panel shows a detailed view of the selected target with a coordinate grid and a color palette (g, r, i, z, Y, RGB). At the bottom, a status bar indicates 'Tiles: 507', 'Inspected: 53 ( 10.0% )', and 'Blacklisted: 46 ( 87.0% )'.

Tile Viewer

Release: v0.5 ( SVA1\_COADD ) Field: SPT-E QA DaCHS Search: eg. 307.0658, -52.6783

Footprint Tile Mosaic Tile List Targets Favorites Gallery

Catalogs

Showing all targets

Target Mosaic Target List

Target

g r i z Y RGB

RA:75.5624 Dec:-61.2314

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

Save target as favorite

Rating: ★★★★★

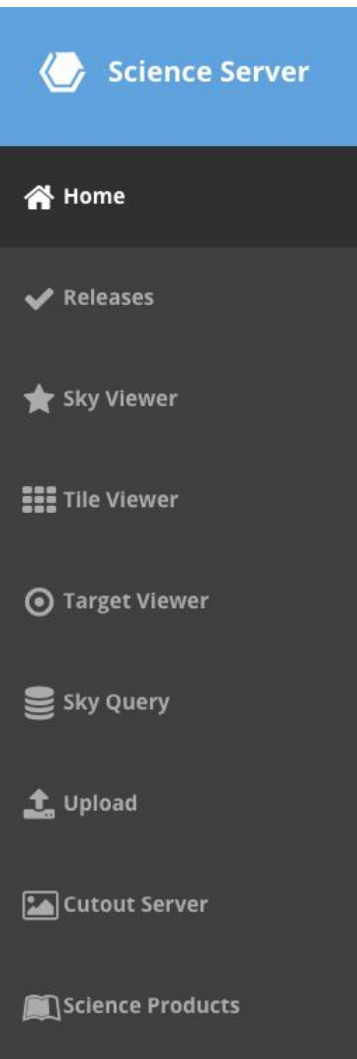
Favorite List:









Favorite Galaxies Save

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



# Validação e exploração de dados @ NCSA\*



 <p>Summary information of DES releases and validation</p> <p>Releases</p>	 <p>All-sky visualization of DES releases in grizY and RGB with overlay of tile grid and objects</p> <p>Sky Viewer</p>	 <p>Inspect DES tiles using visiOmatic tools</p> <p>Tile Viewer</p>	 <p>Manage lists of targets with image display, cutouts, ranking and reject functionalities.</p> <p>Target Viewer</p>
 <p>Query catalogs using sample queries or keep your own query library</p> <p>Sky Query</p>	 <p>Upload external data to the Science Server</p> <p>Upload</p>	 <p>Create co-added or single epoch cutouts from a list of coordinates</p> <p>Cutout Server</p>	 <p>Serve catalogs created by the collaboration</p> <p>Science Products</p>

\* Protótipo para o release público do DES DR1 (Setembro 2017)

Novas tecnologias

- Python 3
- Django 1.9
- Django Rest Framework
- ExtJS 6

Integração com a base de dados do DES no NCSA

Integração de ferramentas opensource

- Aladin, VisiOmatic, MarZ

# Validação de dados (Aladin Lite)

Sera usada na validação dos dados do Ano 3 do DES

Release Validation

J2000 352.4384, -41.4745

FoV: 180°

Aladin

- Scattered light
- Ghosts
- Bright horizontal stri...
- Airplane trails
- Satellite trails
- noisy background
- Incomplete tile
- Bright star
- Cosmic Ray
- Other

g r i z Y

# Validação de dados (Aladin Lite)

Imagens são convertidas para o formato PTIF e armazenadas no NCSA

Release Validation

SPT  
DES2339-4206  
J2000 354.7282, -42.0979

FoV: 1.99°

ALADIN

- Scattered light
- Ghosts
- Bright horizontal stri...
- Airplane trails
- Satellite trails
- noisy background
- Incomplete tile
- Bright star
- Cosmic Ray
- Other

g r i z Y

# Visualização das imagens co-adicionadas e propriedades dos objetos (visiOmatic)

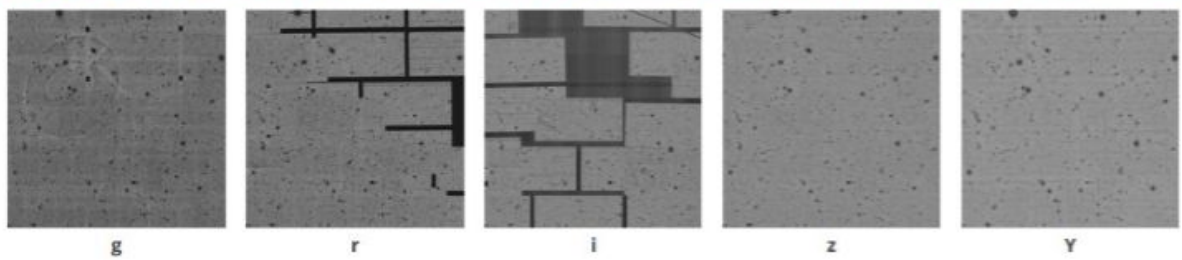
Release Validation



Total controle sobre as propriedades da imagem (contraste, níveis RGB) via browser. JPEG é gerado em tempo real pelo servidor e enviado para o cliente.

- Scattered light
- Ghosts
- Bright horizontal stri.
- Airplane trails
- Satellite trails
- noisy background
- Incomplete tile
- Bright star
- Cosmic Ray
- Other

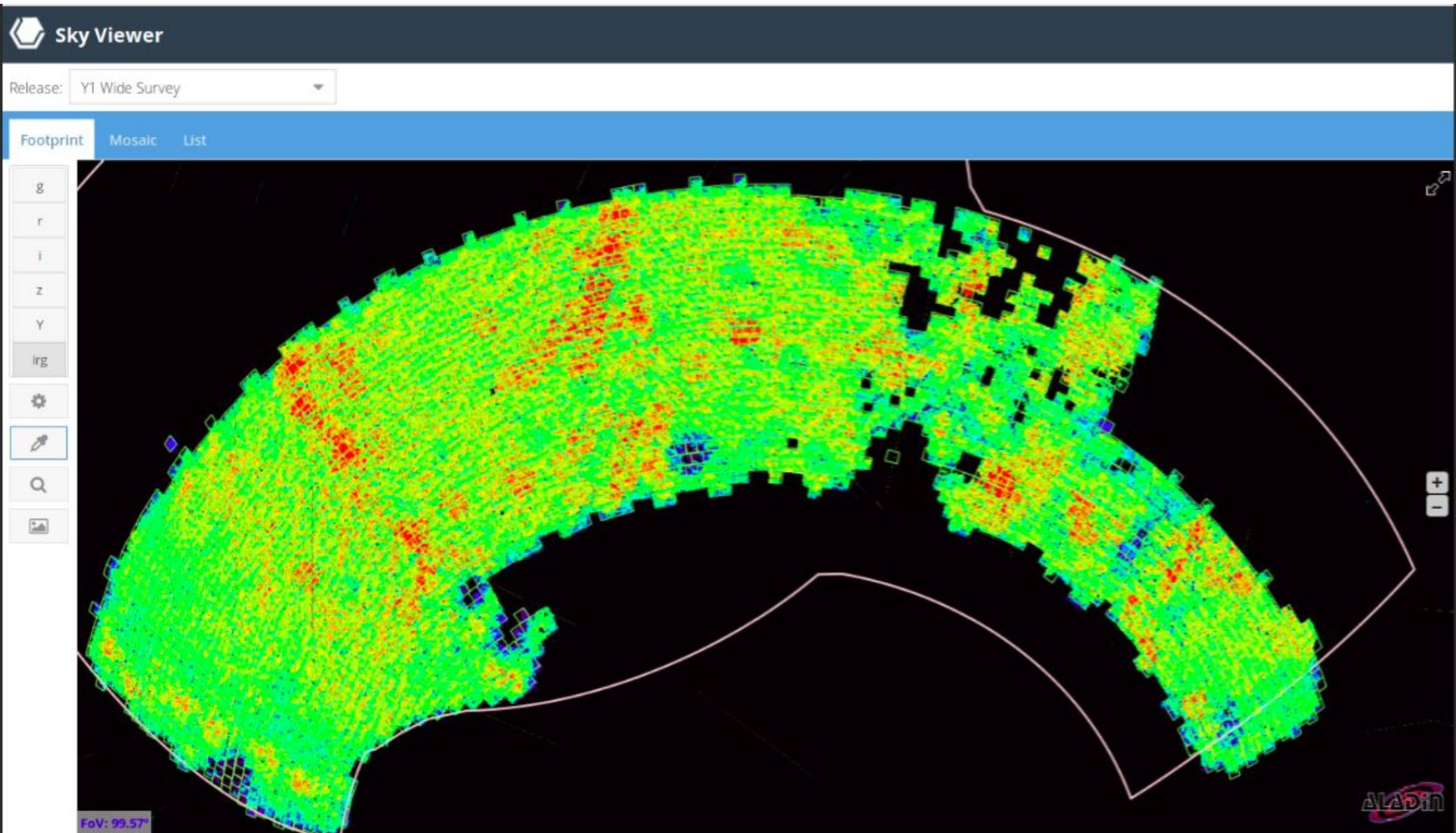
Identificação de defeitos nas imagens



Miniaturas em grizY

# Visualização das propriedades do levantamento

Mapa HEALPix da magnitude limite dos dados do Ano 1 do DES



# A nova "ciência de dados"

Dados > Inferência > Modelo

(Pesquisador)

"BIG DATA"

Dados > Processamento > Catálogo > Inferência > Modelo

(Nova infraestrutura para dar suporte a ciência)

(Novo Pesquisador)

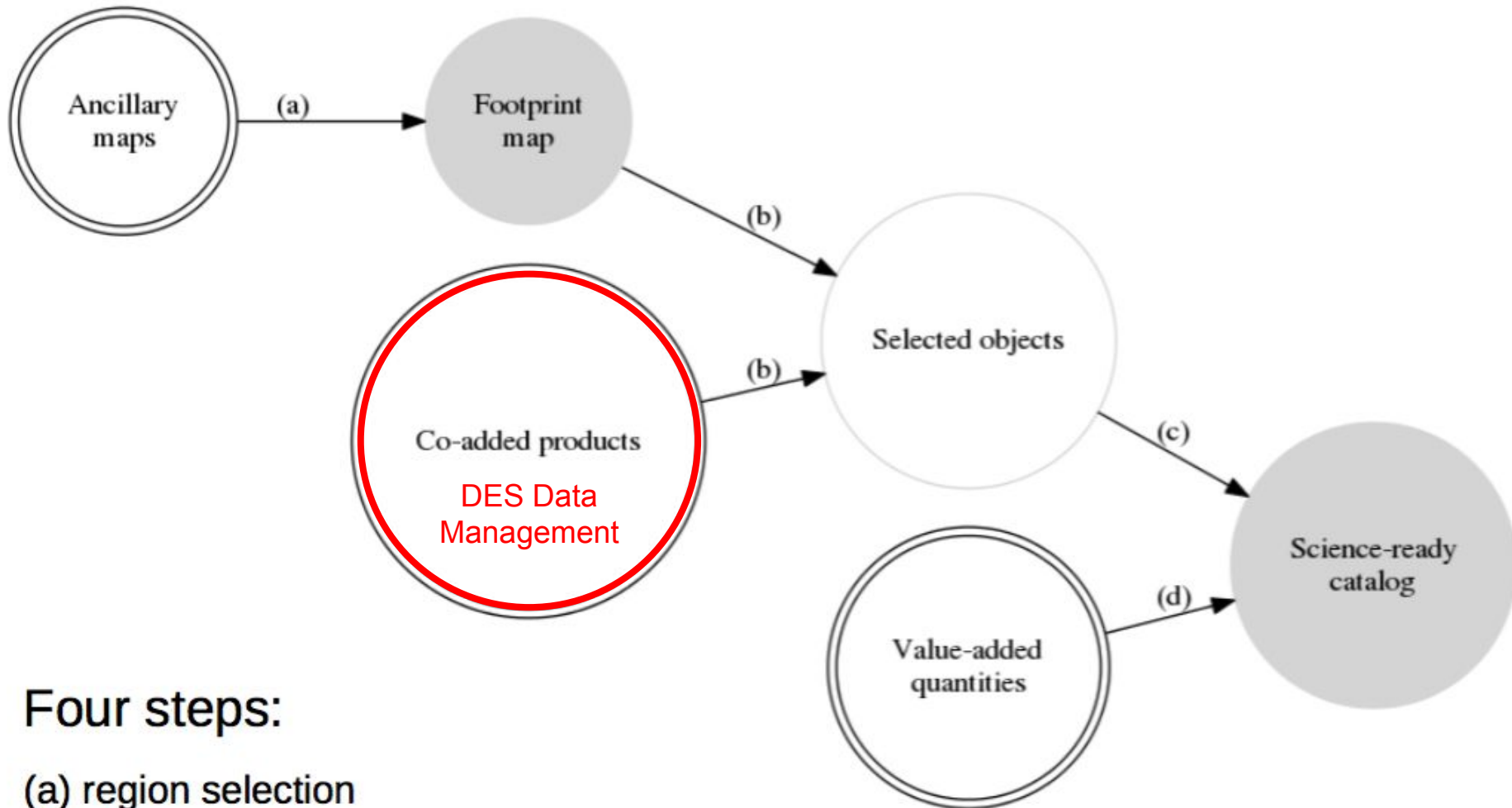
"Big software" para preparação dos dados

Novas habilidades:  
Estatística, computação,  
matemática aplicada

Credit: Andrew Connolly - LSST

# Preparação de catálogos científicos

(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)

# Dashboard para monitoramento dos processos

Release:  Dataset:

### Data Installation

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

### Data Preparation

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

### Science-ready Catalogs

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

- 16 pipelines
- 64 "data products"



# Interface de configuração

~45 parametros

Input Data Configuration Summary

Selected config: System default

Cluster Catalog

- Query Builder
- Catalog Properties

Configuration

Save Select Share with users

Share with groups Reset Set as default

General Information Region Selection Object Selection Column Selection

Mangle Detrac Map

Bad Regions Mask

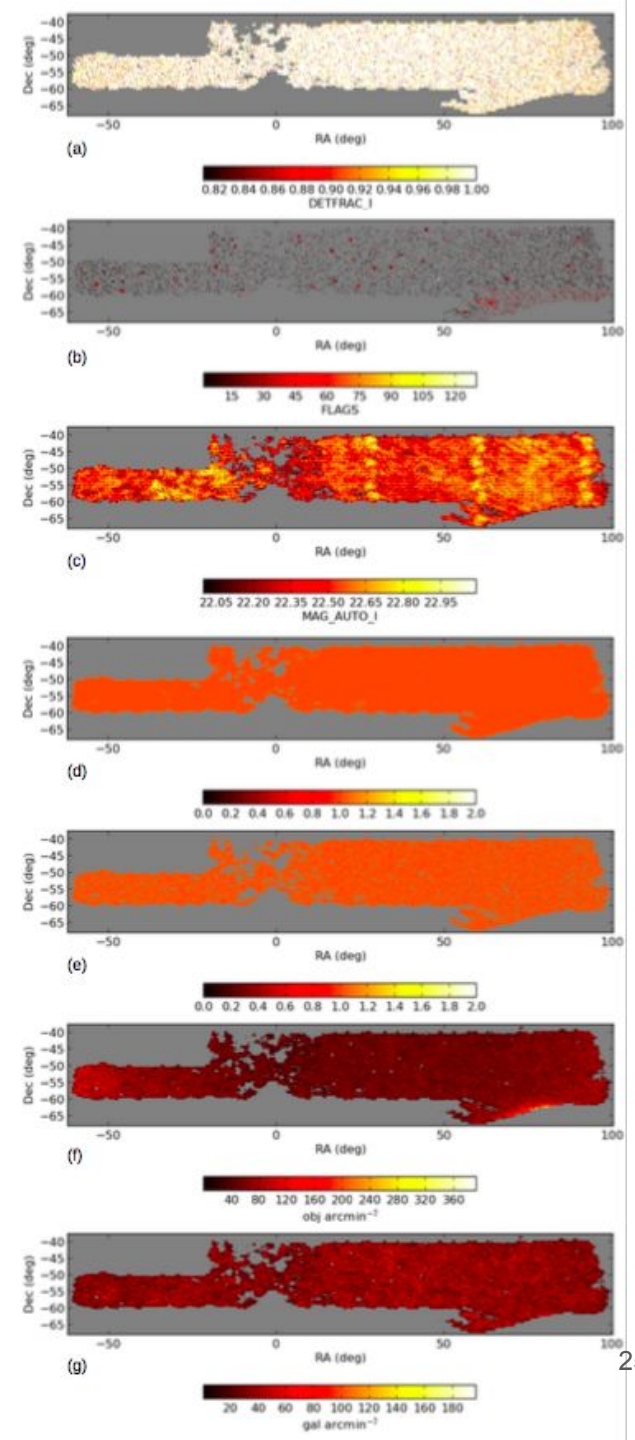
- 1 - Regions with bad astrometric colors
- 2 - Fainter 2MASS star region ( $8 < J < 12$ )
- 4 - Large nearby object (R3C catalog)
- 8 - Bright 2MASS star region ( $5 < J < 8$ )
- 16 - Near the LMC
- 32 - Yale Bright Star region
- 64 - High density of crazy colors
- 128 - Globular Clusters (William et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

Depth Map

Systematic Maps

Additional Mask

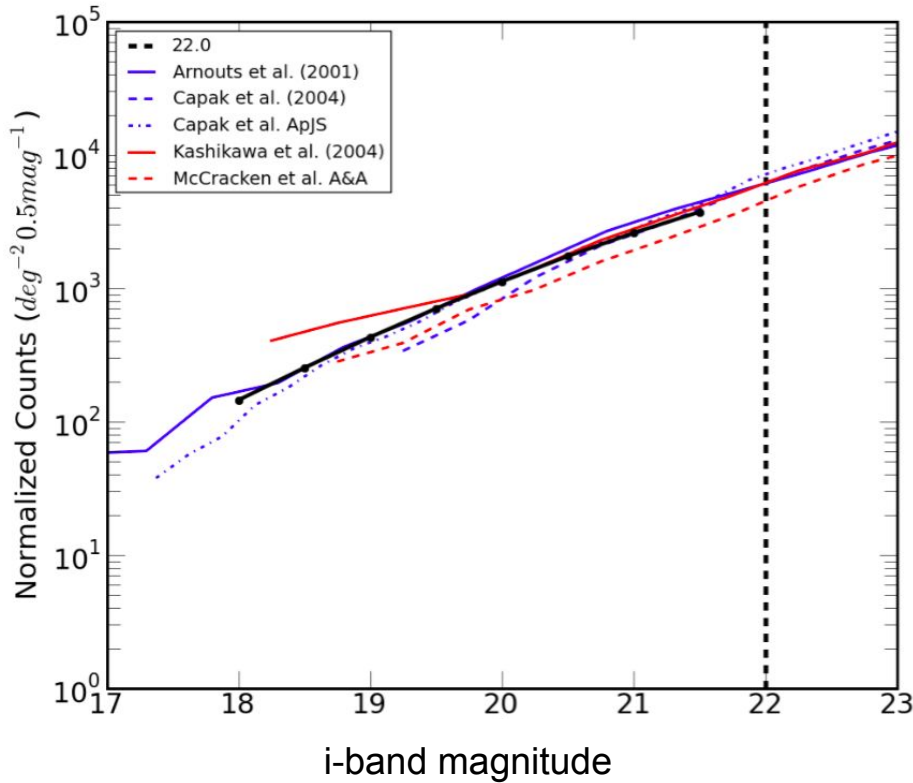
Next



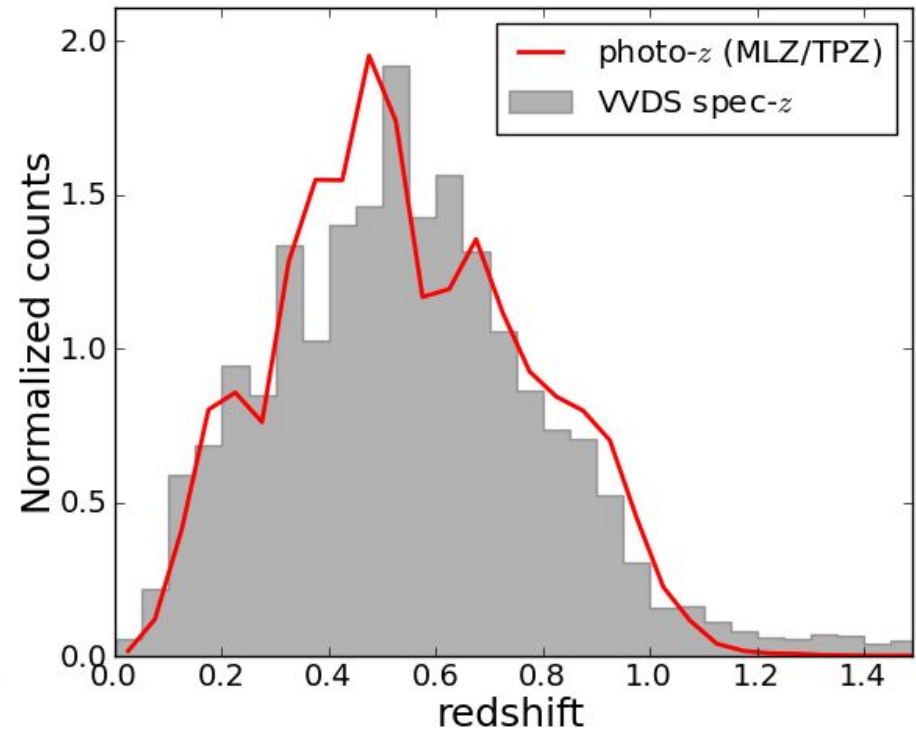
1400 sq deg  
~20M galaxias com photo-z

# Propriedades do catálogo

Number counts



Distribuição de photo-z e spec-z



# Preparação de catálogos científicos

(Fausti et. al 2016 in prep)

- Preservação dos códigos desenvolvidos pela colaboração para criar "ancillary produtos"
- Reprodutibilidade dos catálogos
- Controle dos parâmetros usados na criação dos catálogos
- Proveniência dos dados de entrada
- Documentação do catálogo e suas propriedades

# Workflows científicos @ LIneA

The screenshot shows the top navigation bar of the DES Science Portal with the following items: Dashboard, My Workspace, Pipelines, Tools, Data Server, Documentation, Help, and the user name Angelo Fausti Neto. A dropdown menu is open under the Pipelines tab, listing categories such as Data Installation, Data Preparation, Value-Added Catalogs, Science, Parameter Estimation, Utilities, and Examples. The Science category is selected and has its own sub-menu open, listing LSS, Cluster, SN, WL, Simulation, Galaxy Archeology, Galaxy Evolution, QSO, Strong Lensing, and Combined Probes. The Cluster sub-category is further expanded to show WAZP, Cluster MAtching, and Cluster Comparison. On the left side of the page, there is a section titled 'DES Science Portal: Workflows' with a brief description and a list of bullet points: 'Workflows: hosts workflows for Analysis.' and 'Data Server: provide access to the Data Server'. On the right side, there is a link for 'Tweets by DES Science Portal'. At the bottom left, there is a version string: 'Science Portal dri\_v0.8-24\_ci\_v0.1-26\_16-08-2016\_11-12\_-0300' and a URL 'des-portal.linea.gov.br/#'. At the bottom right, it says 'Powered by LIneA'.

>>

Dashboard My Workspace Pipelines Tools Data Server Documentation Help Angelo Fausti Neto

**DES Science Portal: Workflows**

The Science Portal has two instances:

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

The system is designed to be self-evident.

[Tweets by DES Science Portal](#)

Data Installation ▶

Data Preparation ▶

Value-Added Catalogs ▶

Science ▶

Parameter Estimation ▶

Utilities ▶

Examples ▶

LSS ▶

Cluster ▶

SN ▶

WL ▶

Simulation ▶

Galaxy Archeology ▶

Galaxy Evolution ▶

QSO ▶

Strong Lensing ▶

Combined Probes ▶

WAZP

Cluster MAtching

Cluster Comparison ▶

Science Portal dri\_v0.8-24\_ci\_v0.1-26\_16-08-2016\_11-12\_-0300  
des-portal.linea.gov.br/#

Powered by LIneA

# Template preenchido pelos pesquisadores

## Adding new workflows|in the Portal

- [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 Communication tools](#)

### Example 1: GE Science Workflow

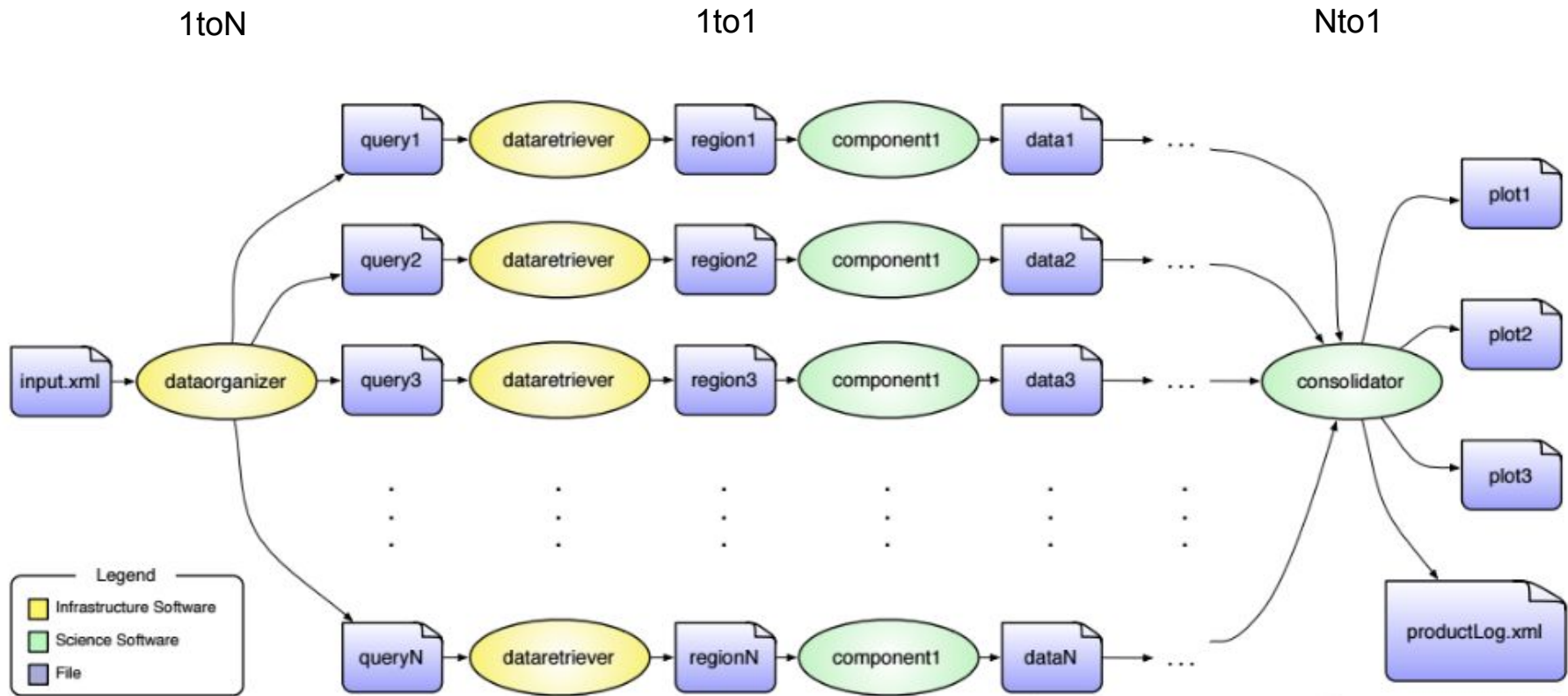
- [1 Description](#)
- [2 Contact points](#)
- [3 Science Code](#)
- [4 Pipeline definition](#)
- [5 Input Data](#)
- [6 Configuration Parameters](#)
- [7 Output Data](#)
- [8 Design of the Process Log](#)
- [9 Schedule](#)
- [10 Communication tools](#)

# Desafios I - Processamento de dados

- Ano 1 do DES - catálogos co-adicionados ~400G
- ~140M objetos e ~600 atributos
- Dados particionados em 3,703 arquivos ~100-150M (DES tile)
- Acesso aos dados durante o processamento
  - PostgreSQL DB (N queries, I/O, network) ✕
  - Lustre File System (I/O cópia de dados para os nós do cluster) ✕
  - Hadoop File System (Dados processados localmente) ✓
- Release dos dados do Ano 3 (Setembro 2016)
  - Catálogo co-adicionado > 1TB em 10.000 arquivos

# Desafios I - Processamento de dados

Processamento paralelo e distribuído (implementação tipo Map-Reduce)



Problemas:

- Movimentação de dados durante o processamento
- Consolidação dos resultados em apenas 1 etapa

# Desafios I - Processamento de dados

- "Levar o processamento aos dados" evitando movimentação de dados durante o processamento
- Particionamento de dados dinâmico
- Consolidação dos resultados parciais em mais de uma etapa
- Resultado: redução do tempo de execução em  $\sim 10x$ 
  - Exemplo: cálculo do redshift fotométrico de 48h para 5h



# Desafios II - Como utilizar outros recursos computacionais disponíveis?

- SDumont (LNCC), FermiGrid , Blue Waters, NERSC
  - Diferentes "ambientes": PBS, Condor, SLURM, Condor-g
  - Movimentação dos dados
  - "Big software" grande número de pacotes de software e dependências
- Science-as-a-service
  - Science APIs (iPlant/CyVerse)
  - Science Gateways (NERSC)
- **Portabilidade:** processamento em nuvem (privadas ou públicas)
- Federação de Nuvens Privadas
  - UFCG/Lab. sistemas distribuidos (Francisco Brasileiro)
- OpenStack, AWS
  - LSST/SQuaRE (Frossie Economou)

## Desafios III - Acesso e distribuição de dados

- Otimização da transferência de dados (RNP)
- Integração do Portal Científico com o sistema de arquivos e DES Science DB (NCSA)
  - Interfaces para acesso aos dados (grande variedade de produtos)
  - Documentação
- Science Server DES Ano 1 e Ano 3 (em prep. DR1)
- Em negociação com LSST/DM para instalação de um protótipo do Data Access Center (DAC) no LIneA (LNCC)

# Aplicações do Portal Científico

Aplicações	SDSS	DES	LSST
Análise da qualidade dos dados		⊙	?
Exploração de dados		⊙	⊙
Preparação de catálogos científicos		⊙	⊙
Workflows Científicos	⊙	⊙	⊙
Distribuição de dados	⊙	⊙	⊙

⊙ Presente

⊙ Futuro

# Dimensionamento do Centro de Dados regional do LSST L2 + L3 "data products"

Necessidade do LSST-BPG definir os casos de uso para dimensionamento do Centro de dados

Ano do levantamento	Ano calendário	Storage (*) (TB)	Processamento (TF)	Nós	Processamento (Cores)	Servidores de apoio
	2019	618	1,5	9	365	1
	2020	618	1,5	9	365	1
1	2021	12735	29,0	104	6543	16
2	2022	19180	55,4	178	11888	16
3	2023	27286	83,2	241	17326	16
4	2024	35518	111,9	287	22372	16
5	2025	44075	141,6	332	27300	16
6	2026	53006	172,1	307	31965	16
7	2027	62332	203,4	289	36200	16
8	2028	72076	235,2	273	40193	16
9	2029	82189	266,9	258	43993	16
10	2030	92691	299,0	245	47172	16
ref. LDM-144						

The image collection includes 24 PB of raw images, 16.5 TB of processed, retained, data (coadds, master calib, cutouts, epo, disk-based science calibrated, disk-based templates), and 475 PB of virtual data (all science calibrated exposures and templates not already on disk). These are all compressed sizes; uncompressed size is roughly 2x larger.

# Conclusões e perspectivas Futuras

- LSST Corporation: suporte ao Level 3
- LSST 10 anos de desenvolvimento de software!
- Protótipo do Data Access Center (DAC) do LSST no NCSA ainda em 2016
- ComCam planejada para 2019 (DM must be ready!)
- First light 2020 e início das operações em 2022
- LIneA DAC regional do LSST
- Infraestrutura do Portal Científico é necessária para análise dos dados
- SDSS e DES DR1: desenvolvimento dos códigos de análise em preparação ao LSST