O Portal Científico

Uma ferramenta para analisar dados de grandes levantamentos astronômicos

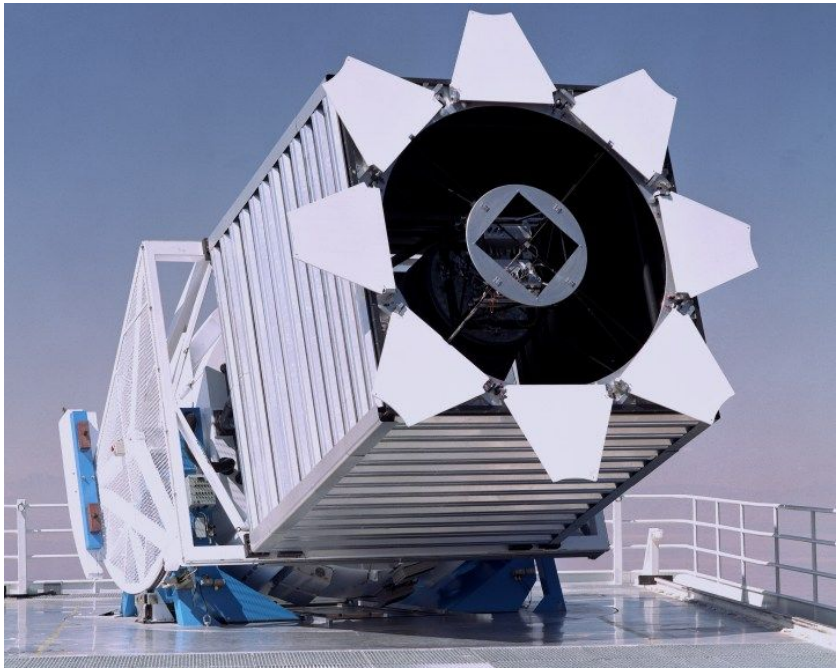
Angelo Fausti Neto
LIneA/LSST

Sumário

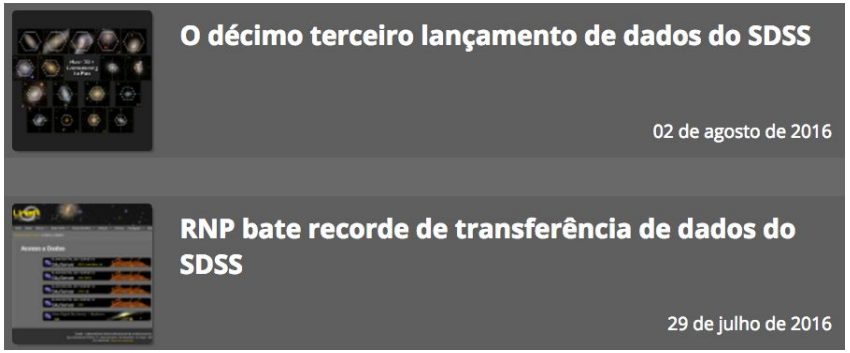
- A era dos grandes levantamentos astronômicos
  - SDSS, DES, LSST
- O papel do LIneA
- A importância do Portal Científico
  - O que é?
  - Aplicações
  - Desafios
- Conclusão e perspectivas futuras

Sloan Digital Sky Surveys (SDSS)

http://www.sdss.org/



SDSS 2.5m telescope at Apache Point Observatory



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 citacoes

Site espelho do SDSS mantido pelo LIneA (DR8-DR13)

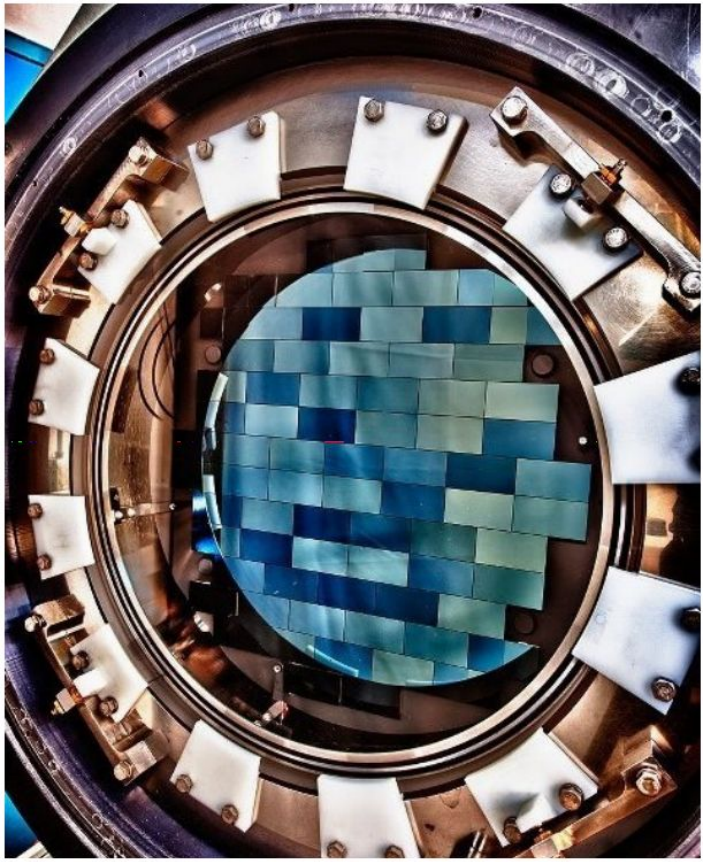
- Base de dados de 15TB
- Sky Server, CASJobs

Suporte a operação do APOGEE 2 em Las Campanas, Chile

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

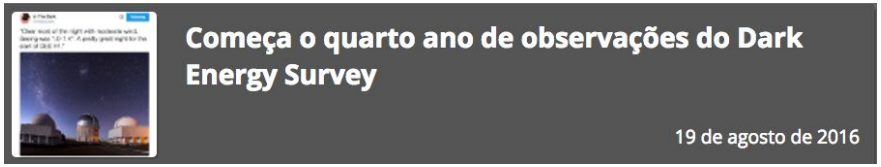
Dark Energy Survey (DES)

https://www.darkenergysurvey.org/



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

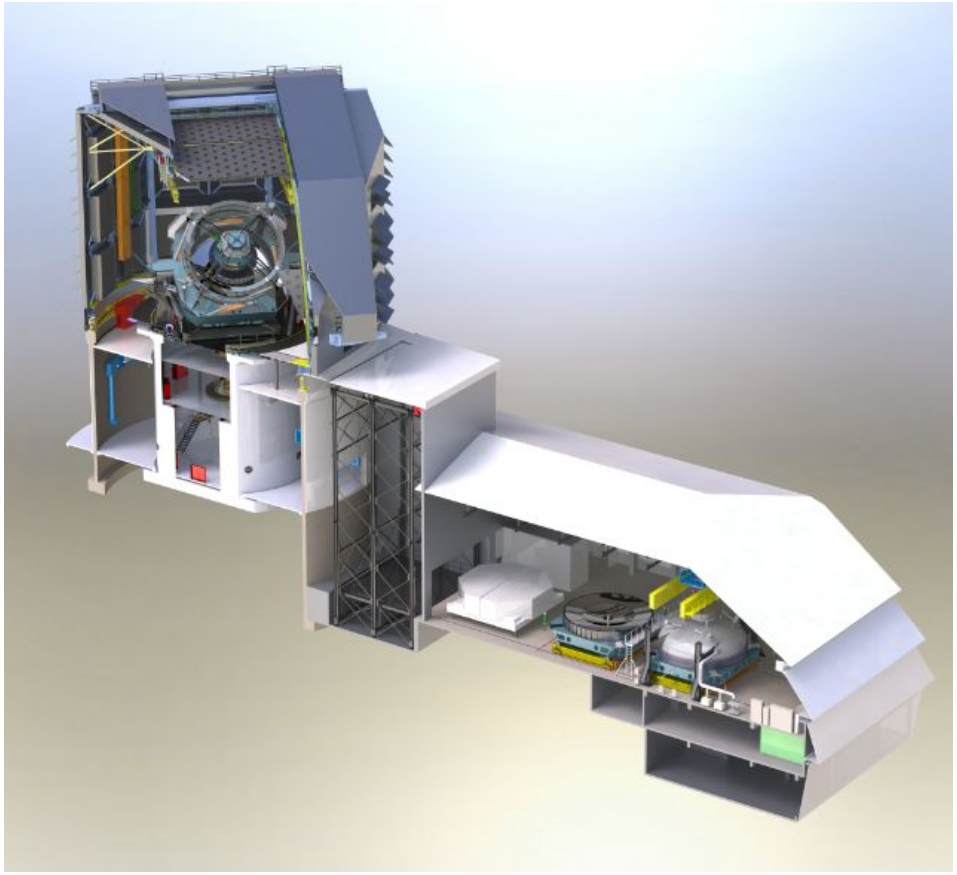
- Levantamento fotométrico (grizY)
- DECam 570Mpixels (62 CCDs)
- Blanco 4m (Tololo)
- ~ 300 exposições por noite (500GB)
- ~100 noites/ano durante 5 anos
- 5.000 sq deg
- 4° ano de operação (Ago 2016)
- Data Management ~8 FTEs (13 pessoas)
- DR1 (público) em Set 2017



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

Large Synoptic Survey Telescope (LSST)

https://www.lsst.org/scientists



- Levantamento fotométrico (ugrizy)
- Camera de 3.2 Gpixels
- Telescópico 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 operacões em 2022

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 sao 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 strelas) ~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:

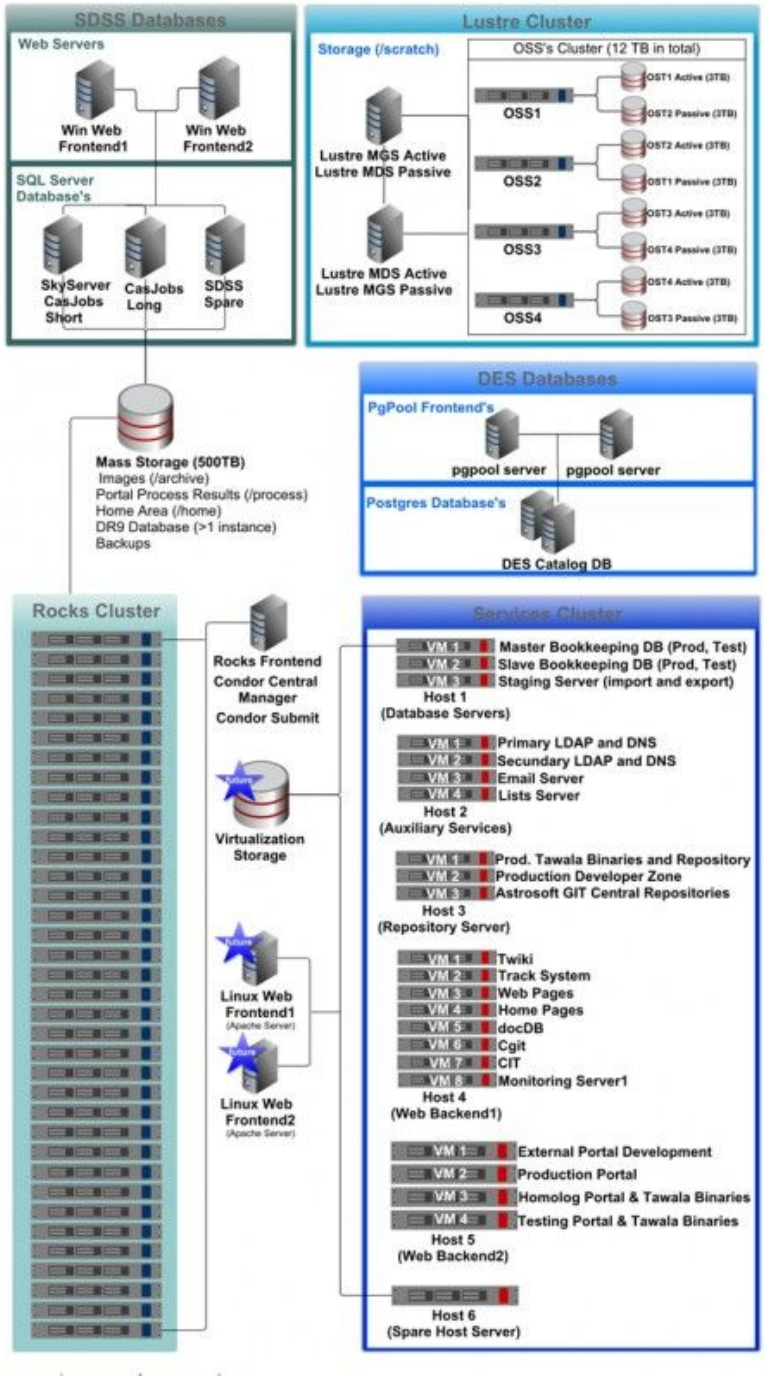
- https://www.lsst.org/scientists/keynumbers
- LSST Data Products Definition Document (http://ls.st/dpdd) (In review)

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

Ênfase	Data	Local
Introdução, Workflows Científicos	Outubro 2010	Fermilab
Precam, Quick Reduce, Workflows Científicos	Outurbro 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
Validaçao, exploração de dados e catálogos científicos	Novembro 2014	NCSA
Preparação de catálogos científicos	Maio 2015	Fermilab

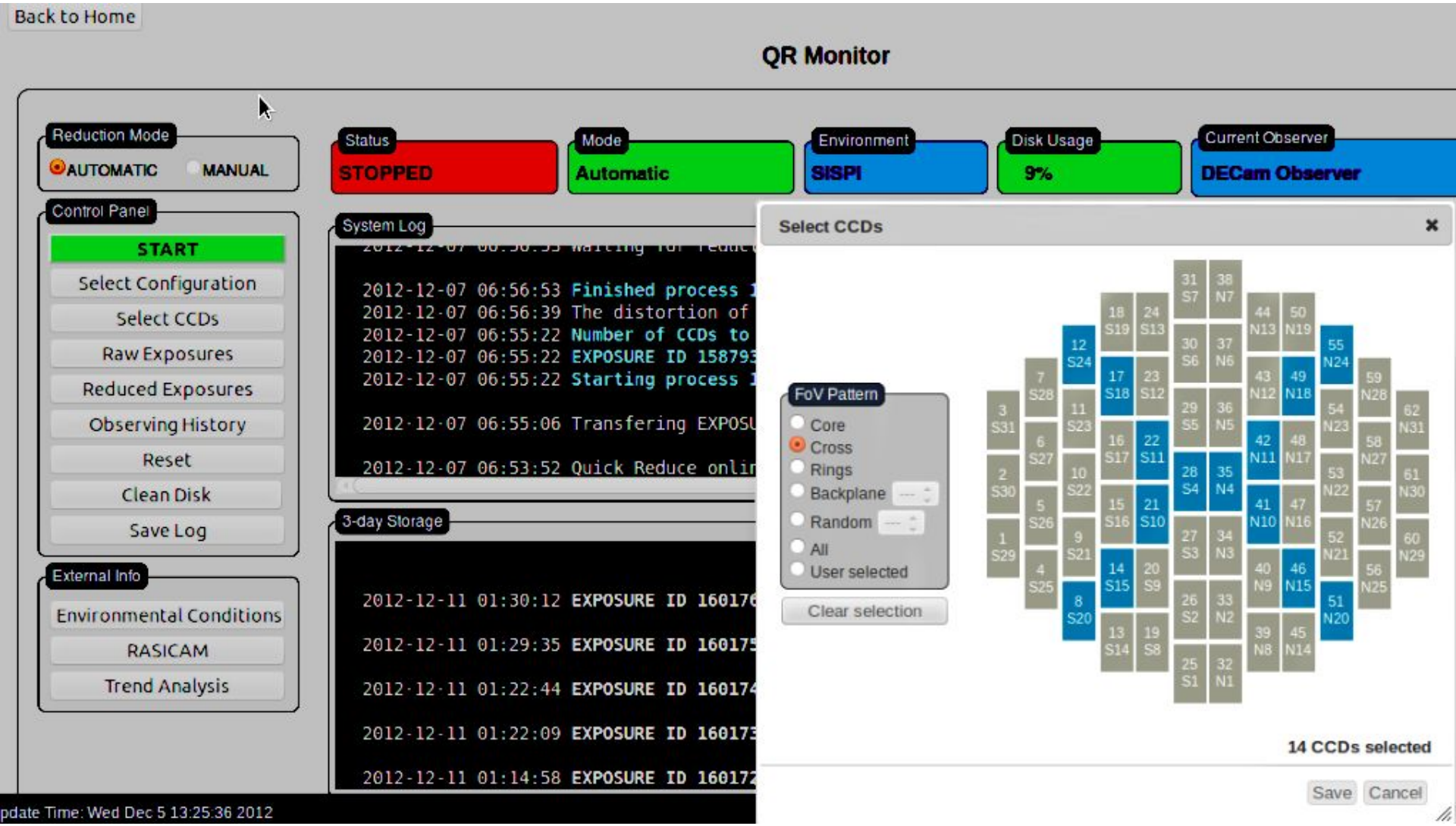
* https://youtu.be/1Qv8HOoeUF4

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/

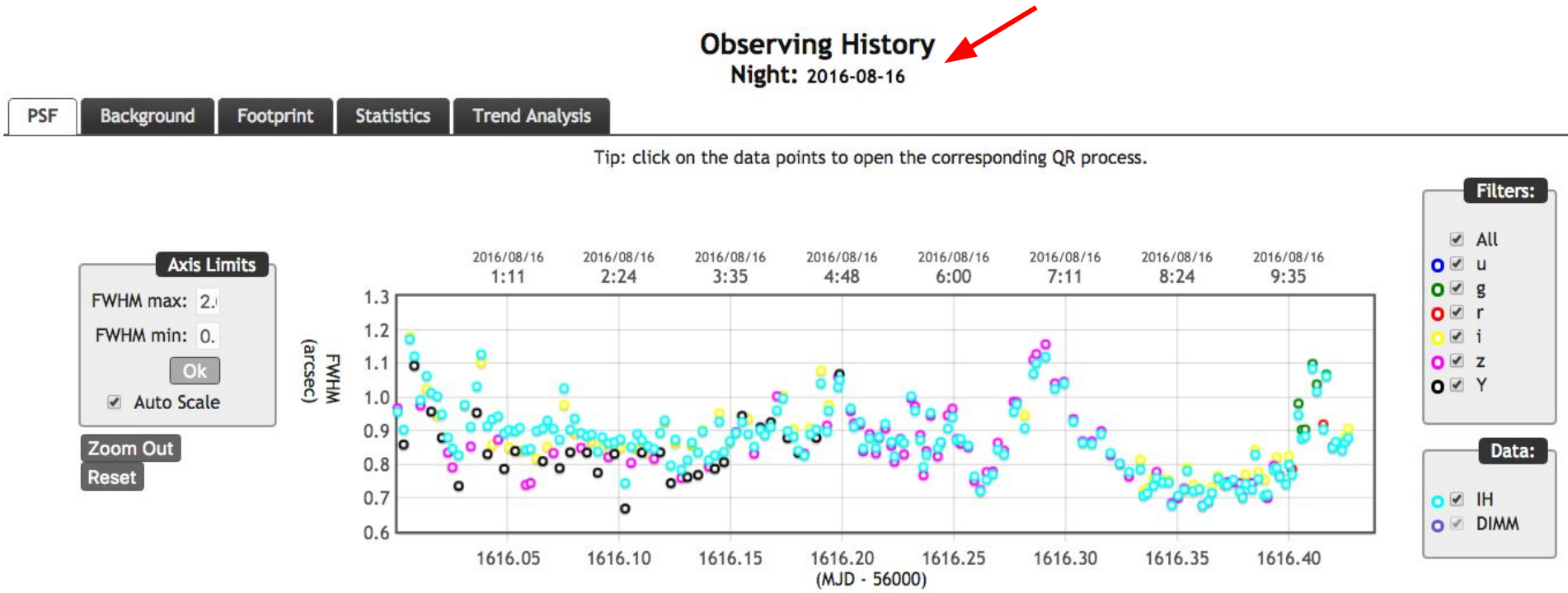
DES Quick Reduce (CTIO)

Redução em tempo real de ~250k exposições (~2.5M CCDs) da DECam desde 2012
Desenvolvido para o DES mas também disponível para outros programas



Monitoramento das observações do DES

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

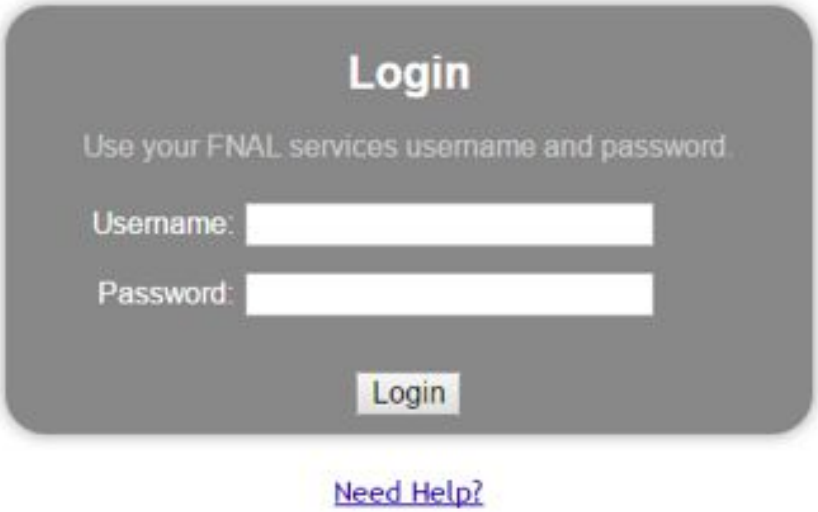


Welcome to the DES Science Portal @ FNAL

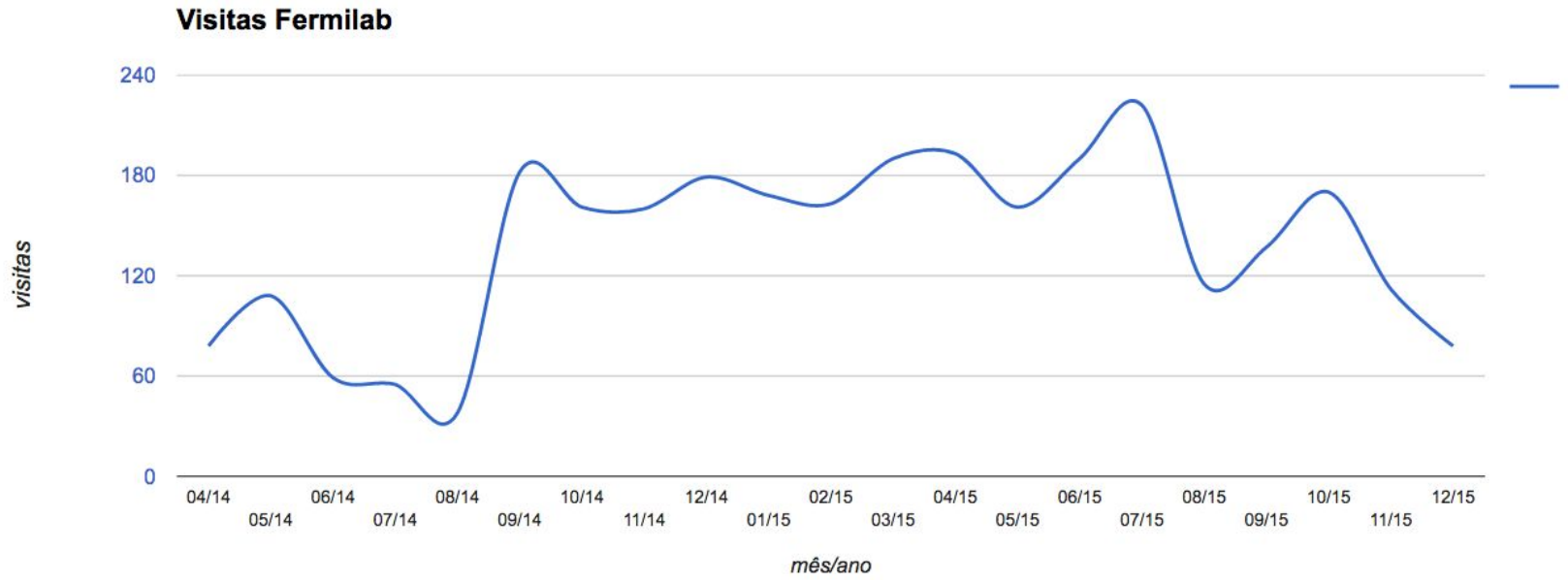
Serviços disponívies

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

"CASJobs moderno"

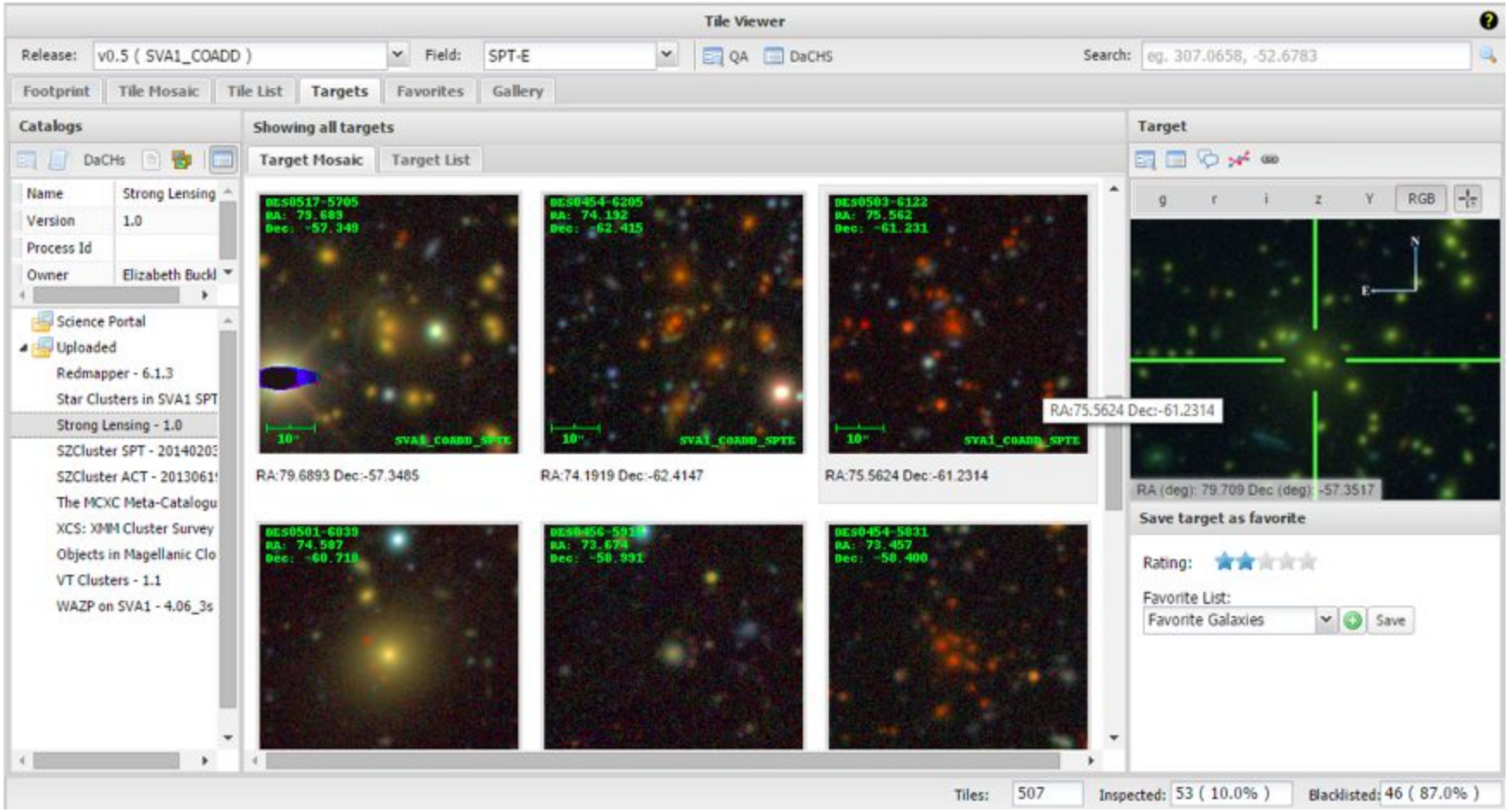


Atende em média 180 membros da colaboração do DES mensalmente.



Ferramenta de cutout e inspeção visual de objetos

Exemplos: Strong Lening, Galaxy Clusters



Validação e exploração de dados @ NCSA*

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

Novas tecnologias

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

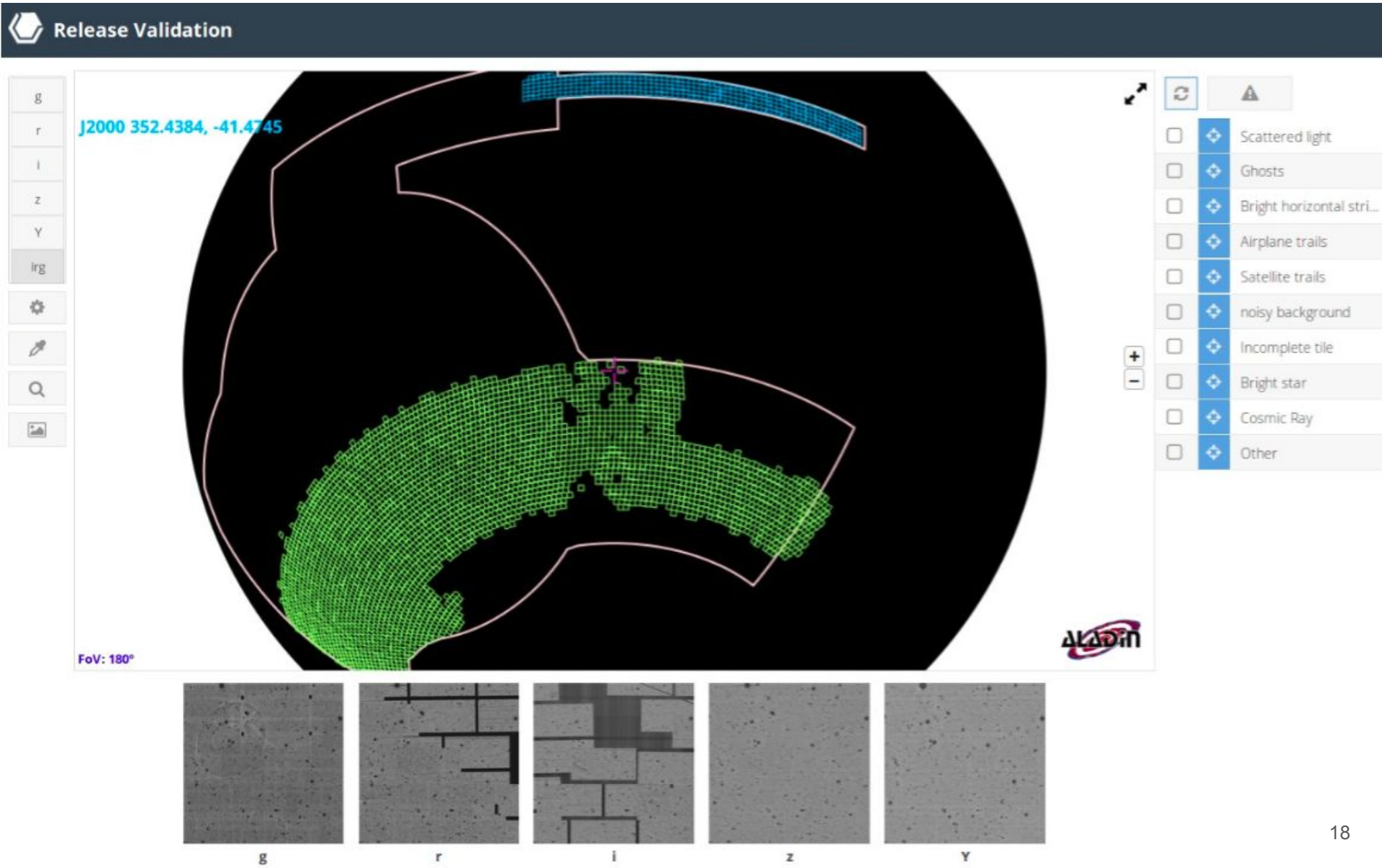
Integração com a base de dados do DES no NCSA

Integração de ferramentas opensource

- Aladin, VisiOmatic, MarZ

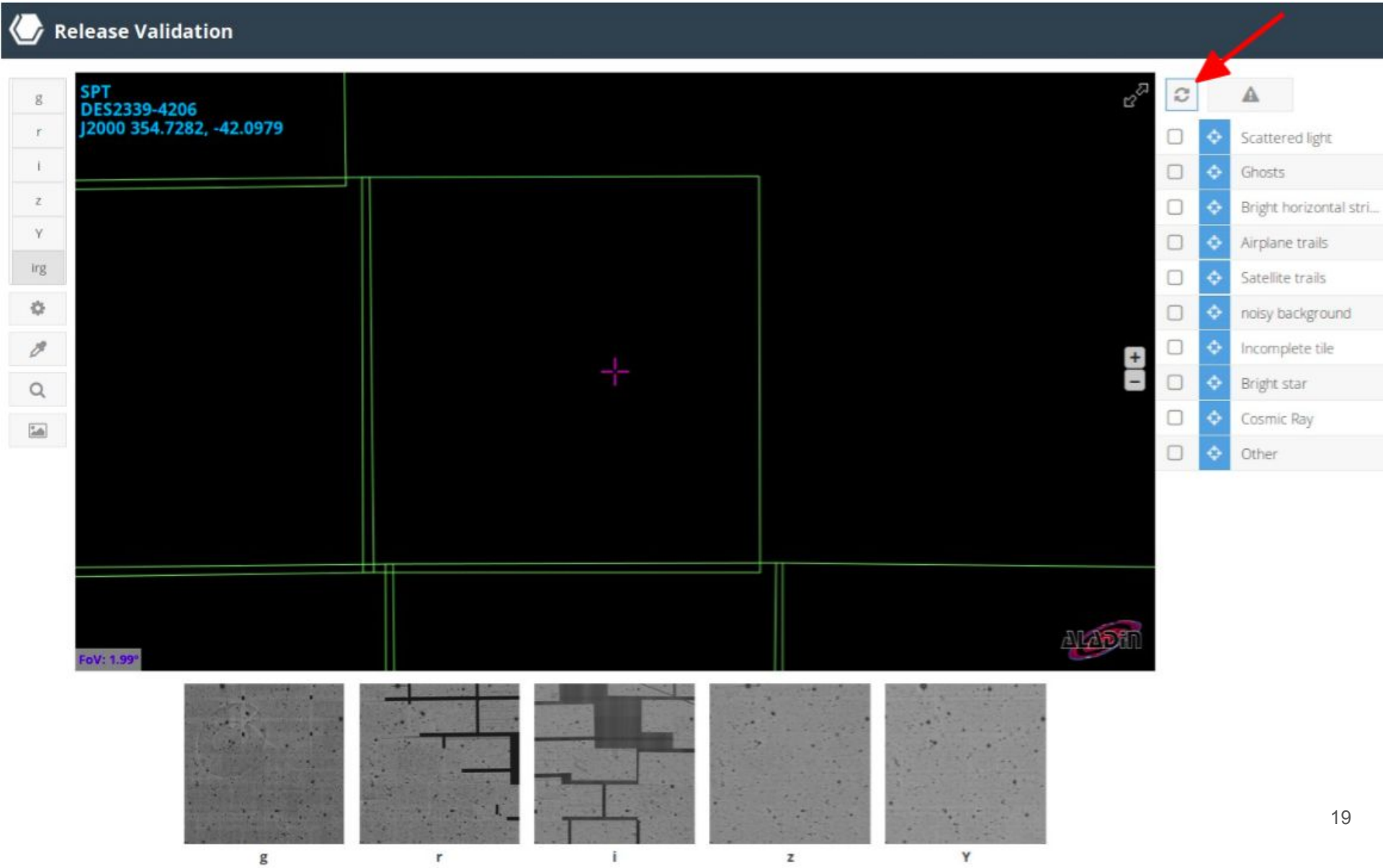
Validação de dados (Aladin Lite)

Sera usada na validação dos dados do Ano 3 do DES

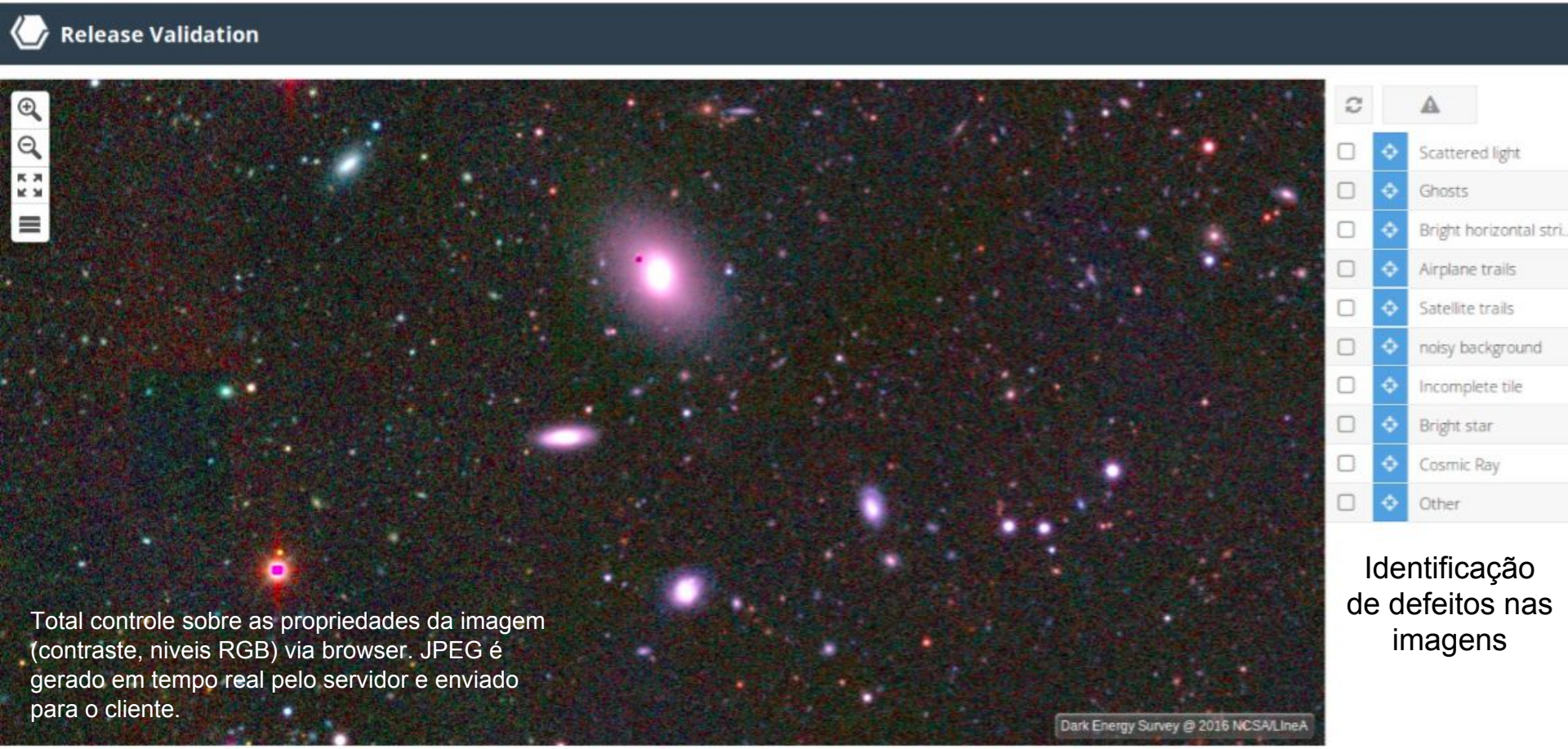


Validação de dados (Aladin Lite)

Imagens sao convertidas para o formato PTIF e armazenadas no NCSA

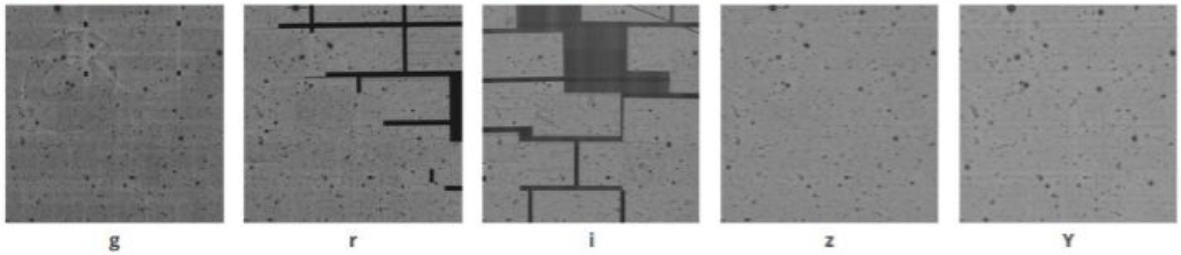


Visualização das imagens co-adicionadas e propriedades dos objetos (visiOmatic)



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

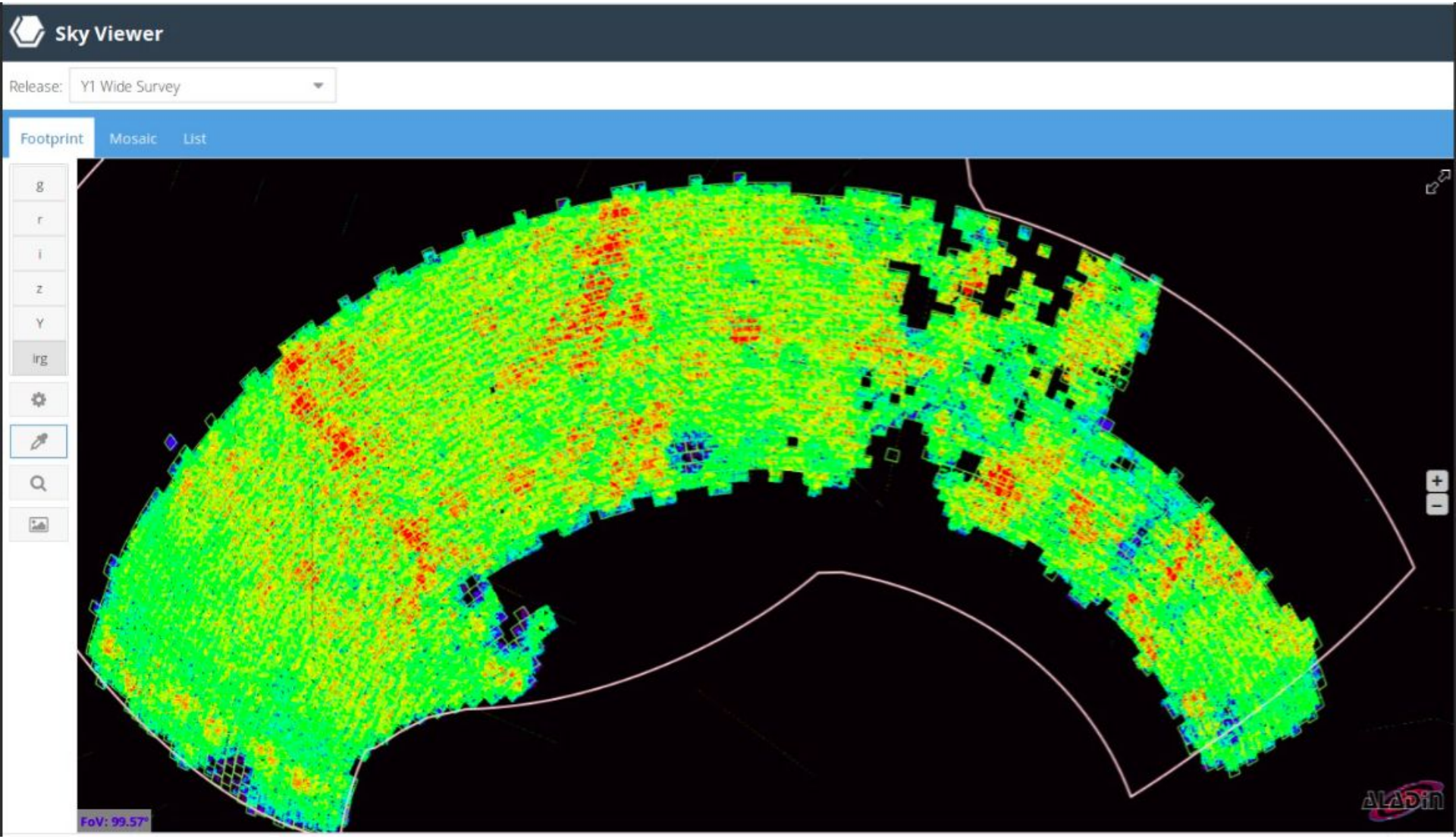
Identificação de defeitos nas imagens



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)

"Big software" para preparação dos dados

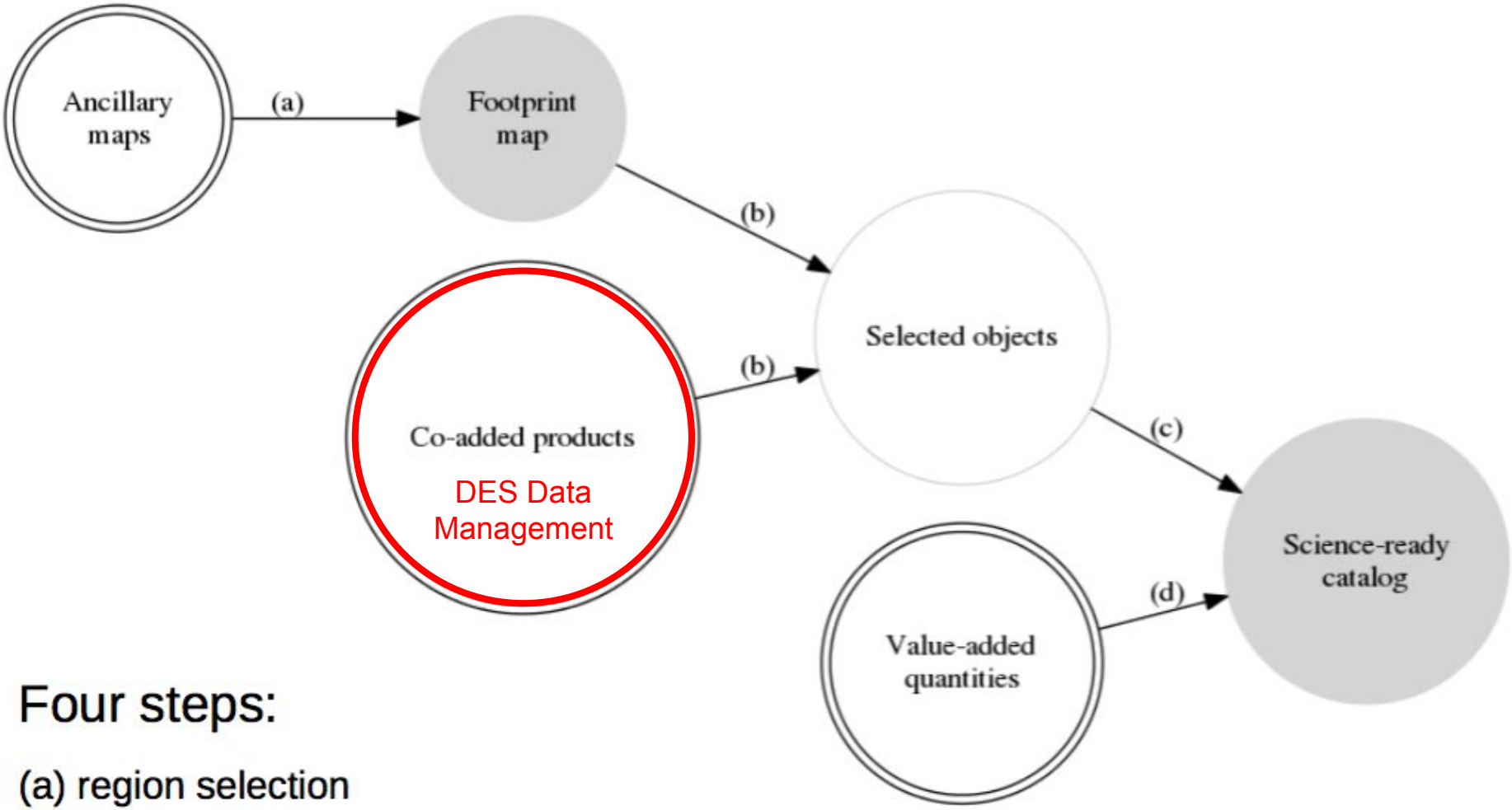
(Novo Pesquisador)

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: 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		

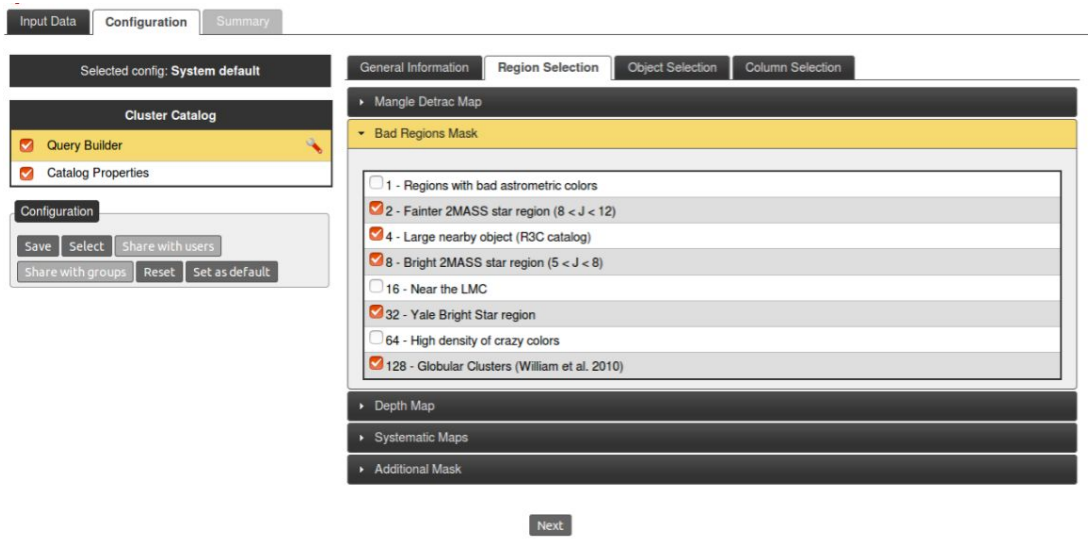
Science-ready Catalogs

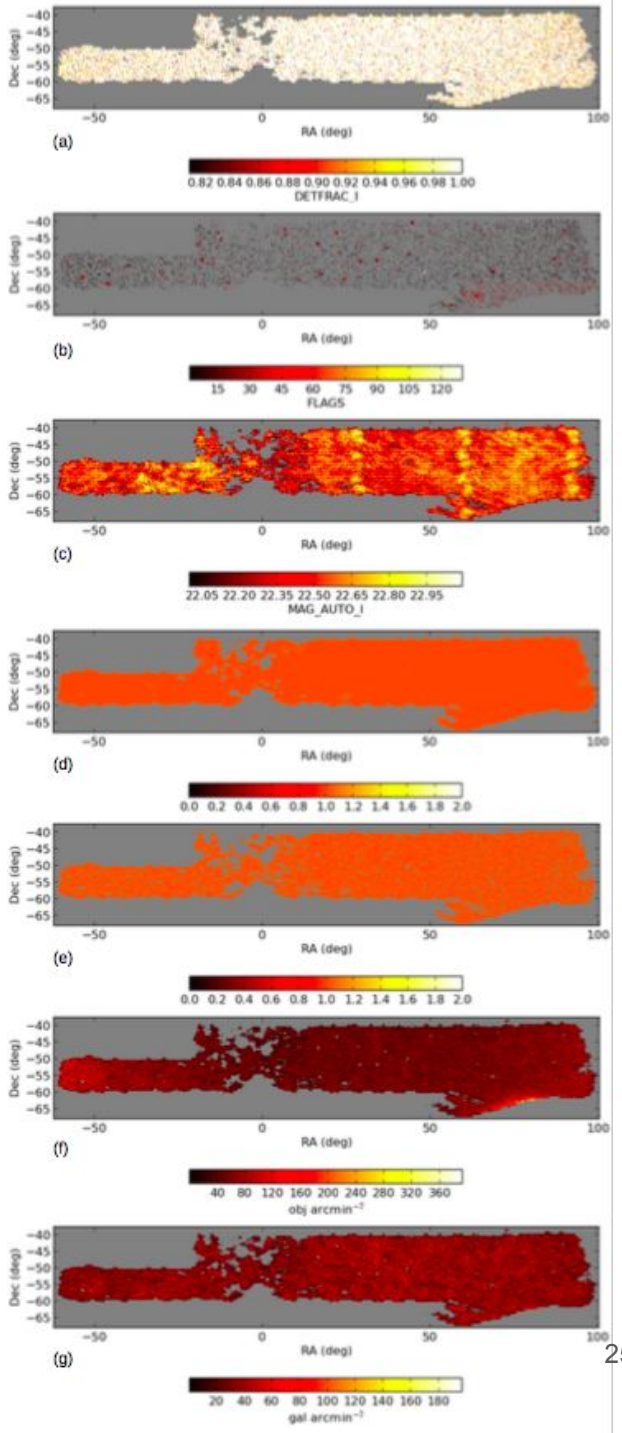
Pipeline	Start	Duration	Runs	Status
Cluster	2016-08-07 17:38:21	02:45:44	25	
GE	2016-05-17 14:40:45	01:52:37	1	
GA	2016-05-24 10:58:30	01:15:09	2	
		Total: 5:53:30		

- 16 pipelines
- 64 "data products"

Interface de configuração

~45 parametros



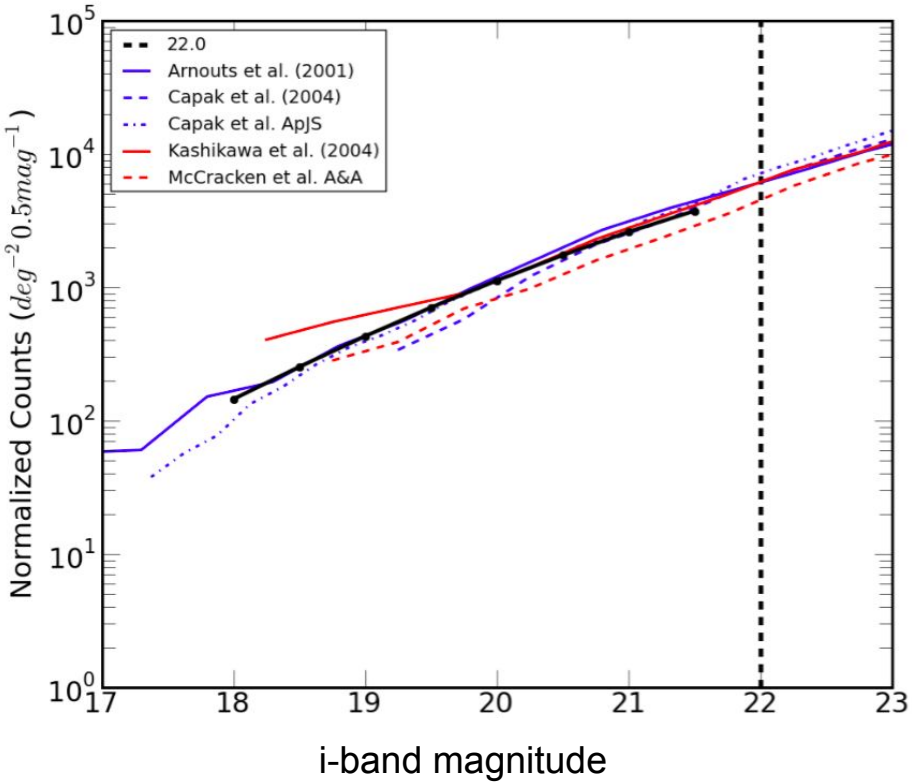


1400 sq deg

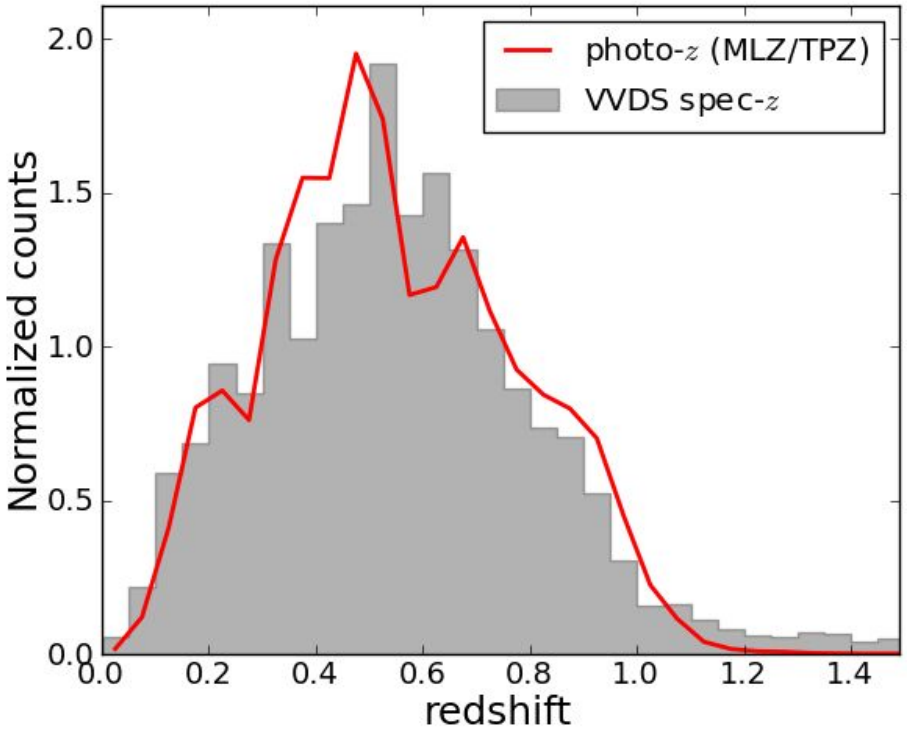
~20M galaxias com photo-z

Propriedades do catálogo

Number counts



Distribuição de photo-z e spec-z

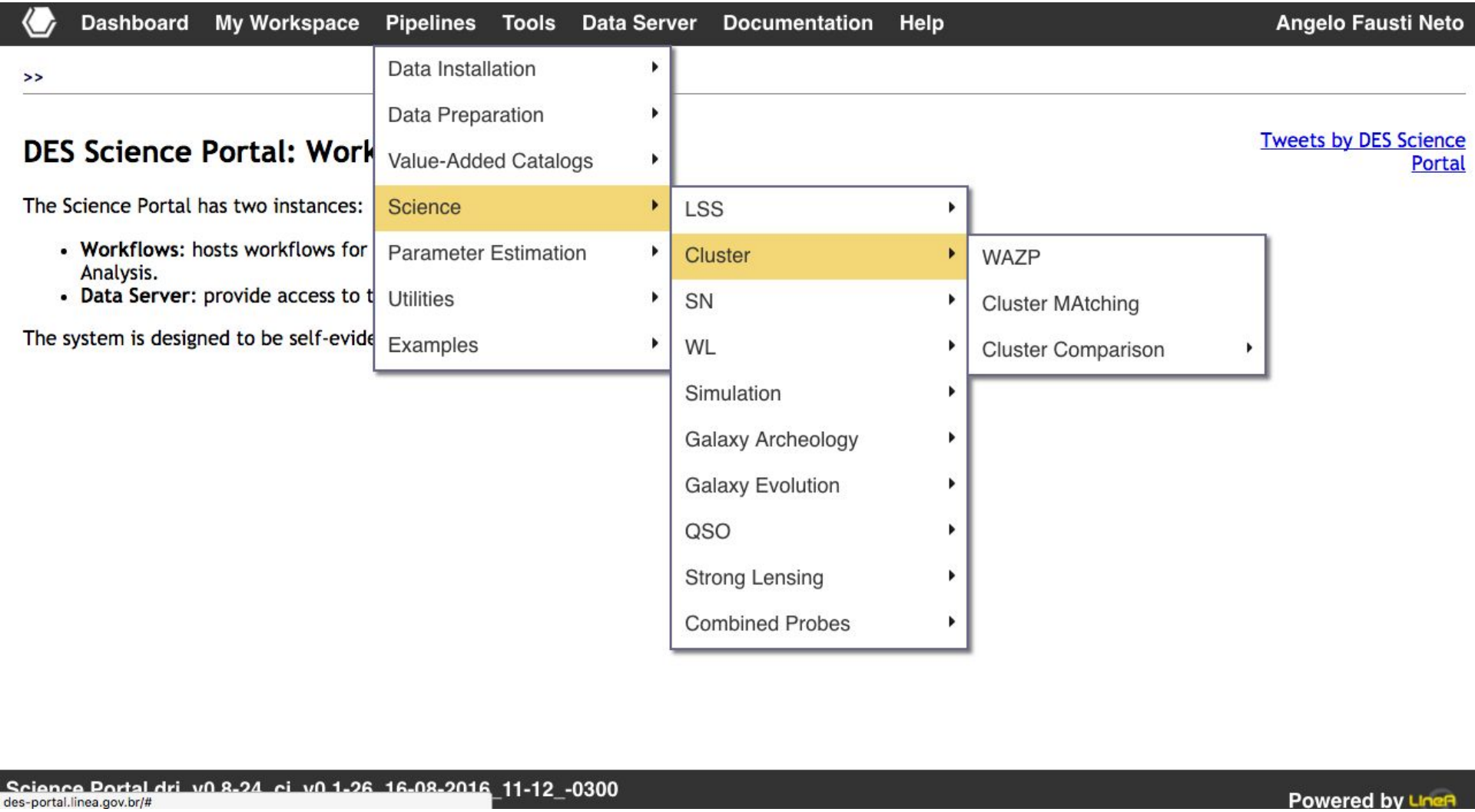


Preparação de catálogos científicos

(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



Template preenchido pelos pesquisadores

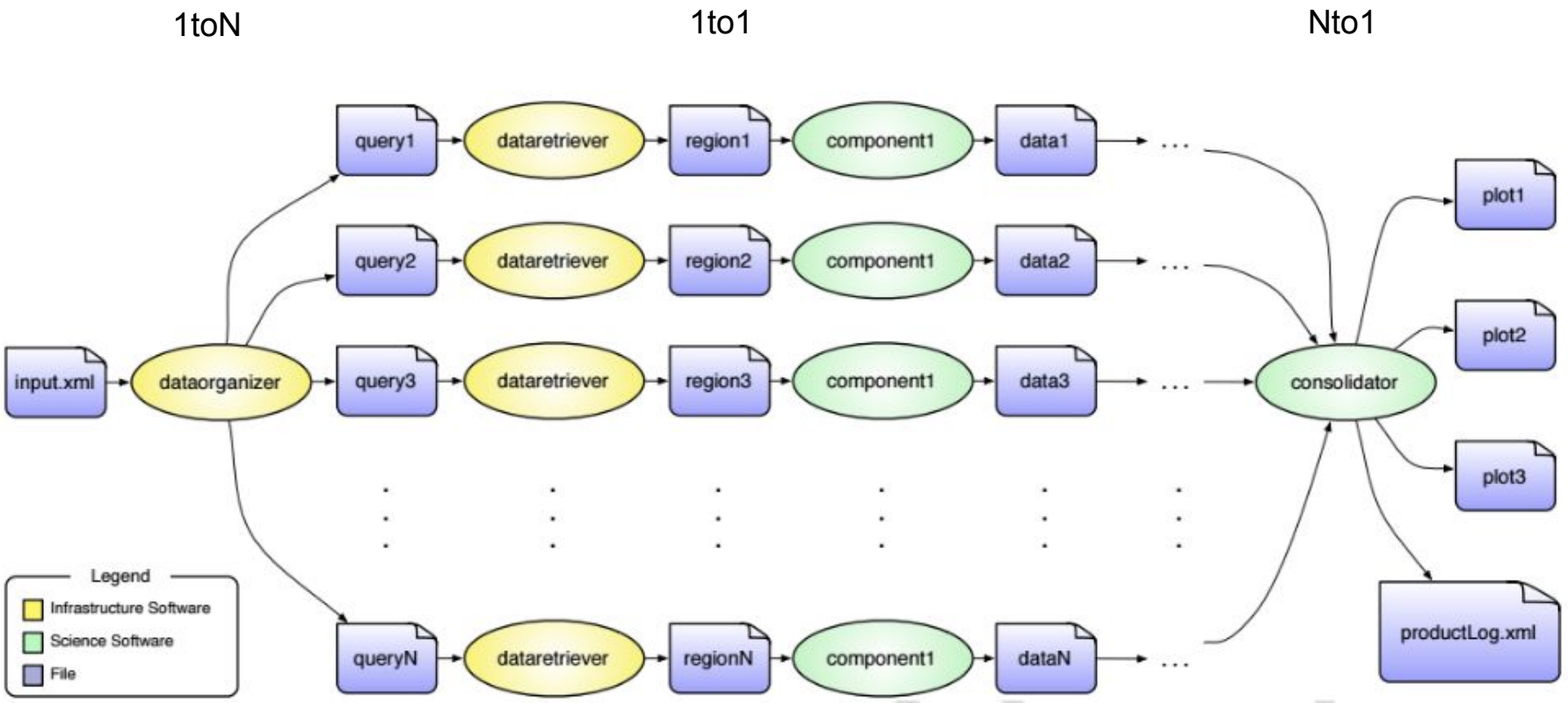
Adding new workflows in the Portal

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 ~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 Nuvems 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 Acess 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