

O programa de contribuição *in-kind* BRA-LIN

Uma parceria entre o LIneA e o Rubin Observatory

Workshop

Workflows e Plataformas Científicas: preparação para o LSST DR1

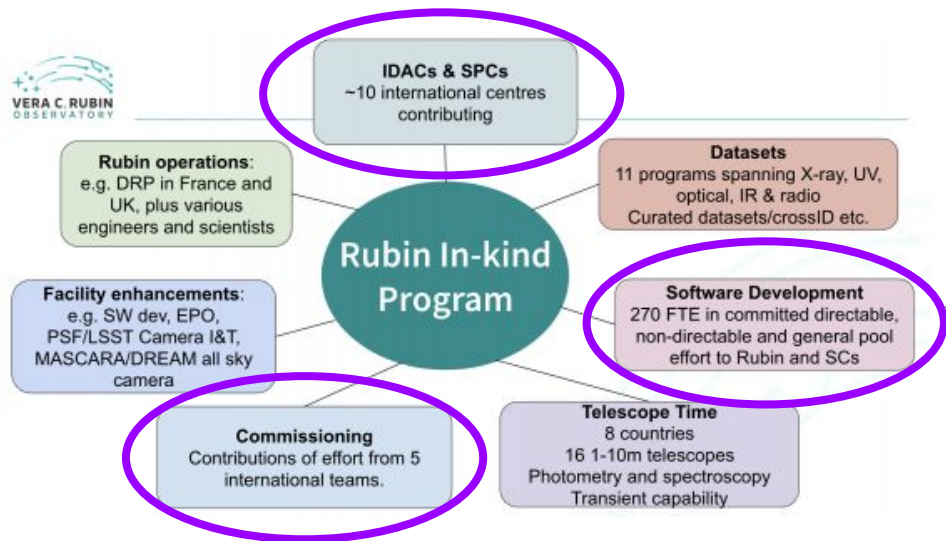
Julia Gschwend

25.03.2025



The Rubin LSST In-kind Program

<https://www.lsst.org/scientists/in-kind-program>



153 in-kind contributions to Rubin and the LSST science community

43 individual international teams from **28 countries** + US/Chile

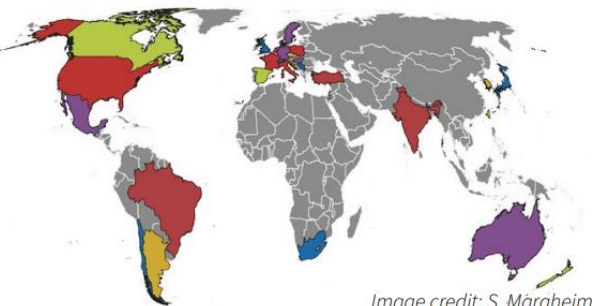


Image credit: S. Mergheim

BRA-LIN in-kind contribution program

country institution

slide credits: Aprajita Verma - LSST PCW 2023

BRA-LIN in-kind contribution program



Computing
infrastructure,
software products,
human resources



U.S. DEPARTMENT OF
ENERGY

Office of Science

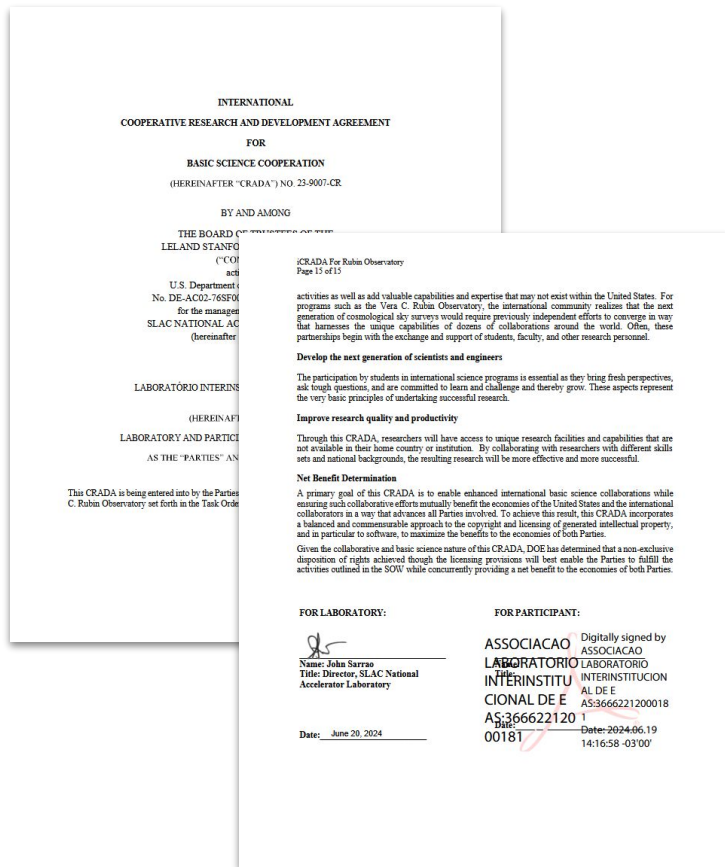
Data rights for 80
Brazilian scientists
(16 PIs and 64 JAs)

Program lead: Luiz da Costa

Program manager: Julia Gschwend

In-kind contributions:

- **S1: Lite IDAC - access to lite catalogs for the whole collaboration**
- **S2 and S3: Pipeline Scientists - software development efforts**
- **S4: Photo-z infrastructure - software and advanced data products periodically delivered**



BRA-LIN - cronograma de entregas

- **S1: Lite IDAC (5 PIs)**
 - 2021 - 2025: Implementation and tests
 - 2026/1: Commissioning
 - 2026/2 - 2039: Operations
- **S2: 1 pipeline scientist for DESC PZ**
 - **contribuição retirada da proposta**
- **S3: 2 pipeline scientists for DESC TJP (2 PIs)**
 - 2021-2024*: 0.25 FTE each (total 2.0 FTE)
- **S4: Photo-z infrastructure for Rubin DM (8 PIs)**
 - 2021-2024: Implementation and tests
 - 2025-2039: Operations (PZ Operations starts with LSST Commissioning)
- **Program Manager (1 PI)**
 - 2021 - 2039

Acordo de 2015 (Infraestrutura de rede - contribuição da RNP) (8 "historical" PIs)

Total: 120 posições para pesquisadores brasileiros (24 PIs + 96 JAs) → [Brazilian Participation Group \(BPG\)](#)

Cronograma oficial de DRs apresentado em janeiro no AAS Town Hall

em vermelho: anotações Carlos Adean

Data coming soon for our community!

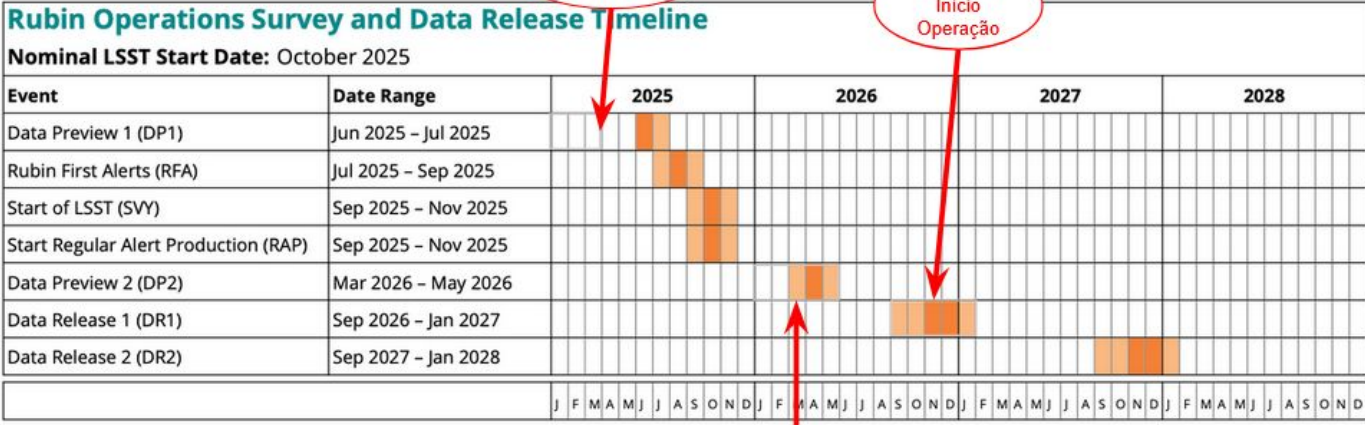


TABLE 3: Rubin Operations Key Milestones for Early Science

Relatórios avaliações e feedback



Cerro Tololo Inter-American Observatory
Community Science and Data Center
Gemini Observatory
Kitt Peak National Observatory
Vera C. Rubin Observatory

Knut Olsen
Astronomer
NOIRLab/CSDC/Rubin Obs.
950 N. Cherry Ave.
Tucson, AZ 85719, USA
+1-520-318-8555
knut.olsen@noirlab.edu

2/23/2024

Dear Luiz:

Re. BRA-LIN-1 2023 Annual Evaluation

Thank you for completing the FY23 (Oct 2022-Sept 2023) Annual Report for your contribution BRA-LIN-S1.

Feedback from the IPC:

The Brazilian IDAC team has continued to do an impressive amount of preparatory work on their contribution, and appear well on their way to stand up their IDAC in advance of LSST DR1. We encourage the team to advertise the planned capabilities of their IDAC to the wide Rubin community through the LSST Science Collaborations, particularly those relevant to their planned use cases.

Sincerely,

Knut Olsen

noirlab.edu
info@noirlab.edu

NOIRLab Headquarters
950 North Cherry Avenue
Tucson, Arizona 85719, USA
+1 520 318 8000

NOIRLab Chile Base
Recinto de AURA
Avda. Juan Cisternas 1500
La Serena, Chile
+56 51 220 5200

NOIRLab Hawai'i Base
670 N. A'ohoku Place
Hilo, Hawai'i 96720, USA
+1 808 974 2500

Annual Evaluation Report for BRA-LIN-S1

Year: FY23

Generated from the submitted annual evaluation form.

Outline of BRA-LIN-S1

Completed by	Ade
with	Car
Present status	Rea
Date of submission*	10/1
Date of last edit**	10/1
Date of generation***	10/1

* date submitted declared by submitter
** date recorded by forms on the last edit
*** date this file was generated

Instructions provided to the Contributor
grey font. Answers they provided

'Summary of work undertaken

Summary of work carried out

This is an overview of the progress of the

For active contributions: Comparing against the work plan objectives and updates (where available), please list any successes including completed work and/or milestones achieved over the last year. Please also list any issues or incomplete work including those to be carried over to the next year. For example, 'Activity X, Y, Z all successfully completed. Deliverable A, B & C could not be completed because of ... but our plan is to ...'.

Contributions without work plans and/or quarterly updates, please complete this question by summarising the work carried out as described above. Additionally, for contributions without a work plan, please submit this as soon as possible (https://lsst.org).

For active contributions: Comparing against the work plan objectives and updates (where available), please list any successes including completed work and/or milestones achieved over the last year. Please also list any issues or incomplete work including those to be carried over to the next year. For example, 'Activity X, Y, Z all successfully completed. Deliverable A, B & C could not be completed because of ... but our plan is to ...'.

Contributions without work plans and/or quarterly updates, please complete this question by summarising the work carried out as described above. Additionally, for contributions without a work plan, please submit this as soon as possible (https://lsst.org).

Annual Evaluation Report for BRA-LIN-S3

Year: FY23

Generated from the submitted annual evaluation form.

Outline of BRA-LIN-S3

Completed by	Dias
with	NA
Present status	Ready
Date of submission*	10/23
Date of last edit**	10/30
Date of generation***	10/30

* date submitted declared by submitter
** date recorded by forms on the last edit
*** date this file was generated

Instructions provided to the Contributor
grey font. Answers they provided

'Summary of work undertaken

Summary of work carried out

This is an overview of the progress of the

For active contributions: Comparing against the work plan objectives and updates (where available), please list any successes including completed work and/or milestones achieved over the last year. Please also list any issues or incomplete work including those to be carried over to the next year. For example, 'Activity X, Y, Z all successfully completed. Deliverable A, B & C could not be completed because of ... but our plan is to ...'.

Contributions without work plans and/or quarterly updates, please complete this question by summarising the work carried out as described above. Additionally, for contributions without a work plan, please submit this as soon as possible (https://lsst.org).

Annual Evaluation Report for BRA-LIN-S4

Year: FY23

Generated from the submitted annual evaluation form.

Outline of BRA-LIN-S4

Completed by	Gschwend, Julia <julia@linea.org.br>
with	no help needed
Present status	Ready for review
Date of submission*	10/17/2023
Date of last edit**	10/18/2023 3:14:47
Date of generation***	10/18/2023 02:17:45

* date submitted declared by submitter
** date recorded by forms on the last edit
*** date this file was generated

Instructions provided to the Contribution Leads (CL) in the form are included below in italic grey font. Answers they provided are shown in boxes below.

'Summary of work undertaken

Summary of work carried out and objectives or milestones achieved

This is an overview of the progress of the contribution since starting the work, usually starting with the work plan.

For active contributions: Comparing against the work plan objectives and updates (where available), please list any successes including completed work and/or milestones achieved over the last year. Please also list any issues or incomplete work including those to be carried over to the next year. For example, 'Activity X, Y, Z all successfully completed. Deliverable A, B & C could not be completed because of ... but our plan is to ...'.

Contributions without work plans and/or quarterly updates, please complete this question by summarising the work carried out as described above. Additionally, for contributions without a work plan, please submit this as soon as possible (https://lsst.org).

The IDAC



BRA-LIN-S1: The Brazilian IDAC

Contribution Lead: Carlos Adean

IDAC Lite Resources

- 5 PB storage
- 500 TB database PostgreSQL
- 500 cores HPC (~25 TFlops)
- 20 Gbps network
- Kubernetes cluster (JupyterHub)

(2+1) annual data releases

<https://www.linea.org.br/idac>

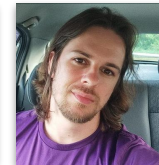
Current IDAC Infrastructure Team



Carlos Adean
System Analyst
IDAC Contribution Lead



Nubia Garcia
System Analyst



Eloir Troyack
System Analyst



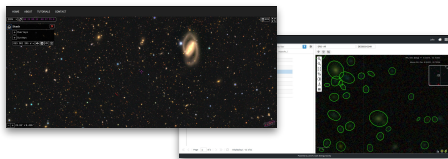
Financiado por:



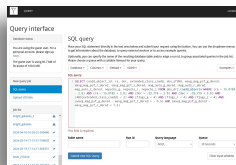
O IDAC Brasil já é realidade! (fotos de 2024)



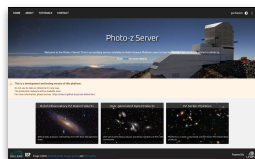
Brazilian IDAC User Interfaces



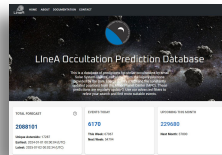
Sky Viewer / Target Viewer
Image visualization
and catalog overlay



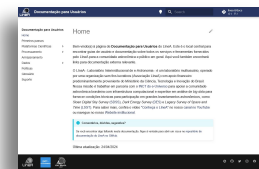
User Query Tool
SQL/ADQL
MyDB



PZ Server



Solar System Portal



Documentation Site
user manuals
tutorials



Não perca! Amanhã:

15:35

Plataformas científicas para exploração dos dados do LSST &
LSST Data Products - Definições e características dos dados do LSST

Julia Gschwend

Tópicos abordados:

- Quais tipos de dados o Rubin Observatory vai disponibilizar?
 - para os membros do projeto X para o público geral
 - via DAC X via IDACs
- Onde e como acessar os dados?
 - do LSST em geral
 - hospedados no IDAC-Brasil
- Onde obter informações?
- Como obter ajuda?

DESC Pipeline Scientists

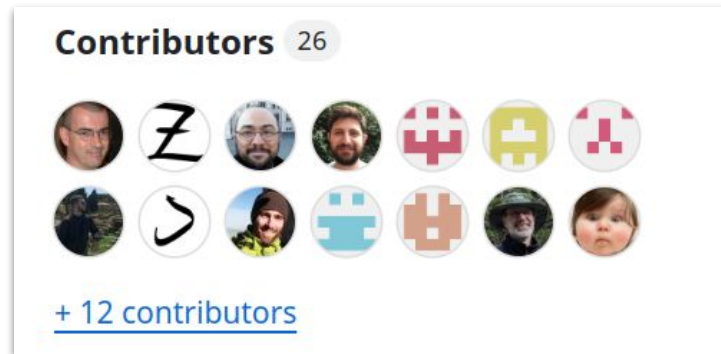
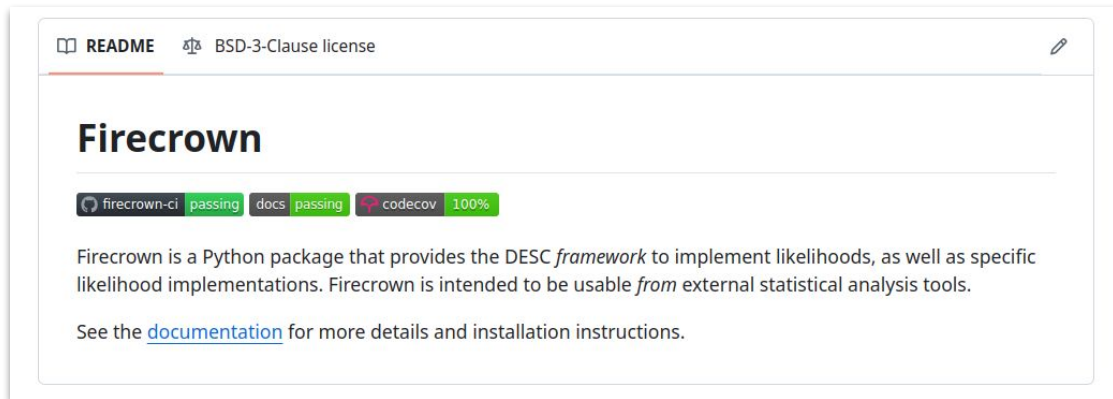
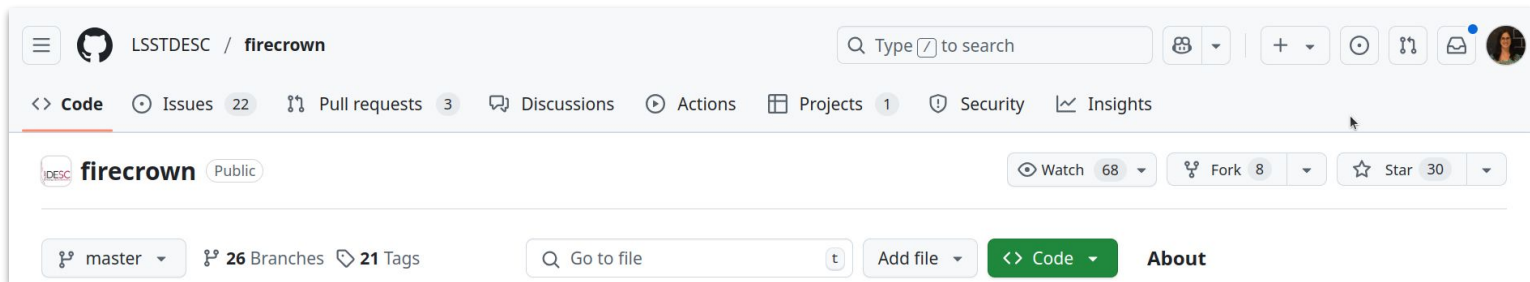


Firecrown

Contribution Lead: Dr. Sandro Vitenti

Recipient: LSST DESC TJP Working Group

Entregue! ✓



ML applications for WL

Contribution Lead: Dr. Clecio de Bom

Recipient: LSST DESC Time Domain WG

Primary goals:

Construct a pipeline to find strongly lensed supernovae, and support infrastructure for strong lens finders.

- Tentativa de suprir as 0.85 FTEs que ficaram pendentes com a saída do Felipe Oliveira.
- Ainda em fase de negociação com o recipient group. As entregas começam a contar a partir dos relatórios trimestrais.
- Detalhes na apresentação do Clécio de Bom.

PZ Services

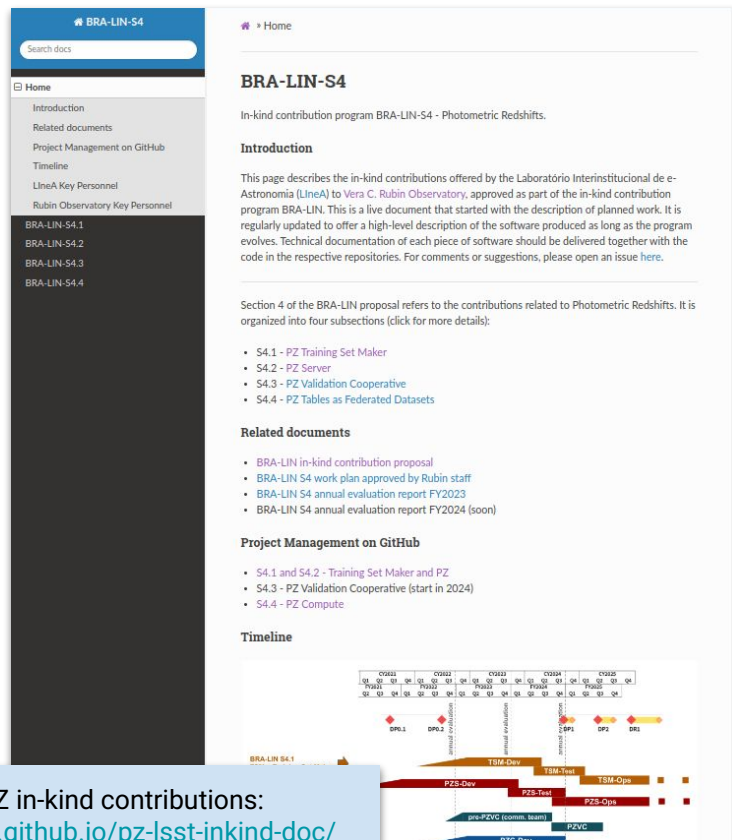


BRA-LIN-S4: Photo-z Services

Contribution Lead: Julia Gschwend

Four in-kind contributions

- S4.1: Training Set Maker
- S4.2: PZ Server
- S4.4: PZ Compute
- S4.3: PZ Validation Cooperative



The screenshot displays the BRA-LIN-S4 website. The left sidebar contains a navigation menu with links to Home, Introduction, Related documents, Project Management on GitHub, Timeline, LineA Key Personnel, and Rubin Observatory Key Personnel. The main content area is titled 'BRA-LIN-S4' and describes the in-kind contribution program. It includes an 'Introduction' section explaining the program's purpose and a 'Related documents' section listing various reports and proposals. A 'Project Management on GitHub' section lists specific contributions. A 'Timeline' section at the bottom features a Gantt chart showing the project schedule from 2021 to 2024, with milestones for training set creation, server development, and validation.

BRA-LIN-S4

In-kind contribution program BRA-LIN-S4 - Photometric Redshifts.

Introduction

This page describes the in-kind contributions offered by the Laboratório Interinstitucional de e-Astronomia (LineA) to Vera C. Rubin Observatory, approved as part of the in-kind contribution program BRA-LIN. This is a live document that started with the description of planned work. It is regularly updated to offer a high-level description of the software produced as long as the program evolves. Technical documentation of each piece of software should be delivered together with the code in the respective repositories. For comments or suggestions, please open an issue [here](#).

Section 4 of the BRA-LIN proposal refers to the contributions related to Photometric Redshifts. It is organized into four subsections (click for more details):

- S4.1 - PZ Training Set Maker
- S4.2 - PZ Server
- S4.3 - PZ Validation Cooperative
- S4.4 - PZ Tables as Federated Datasets

Related documents

- BRA-LIN in-kind contribution proposal
- BRA-LIN S4 work plan approved by Rubin staff
- BRA-LIN S4 annual evaluation report FY2023
- BRA-LIN S4 annual evaluation report FY2024 (soon)

Project Management on GitHub

- S4.1 and S4.2 - Training Set Maker and PZ
- S4.3 - PZ Validation Cooperative (start in 2024)
- S4.4 - PZ Compute

Timeline

The timeline diagram shows the project schedule from 2021 to 2024. Key milestones include the start of the project in 2021, the completion of the training set maker (S4.1) in 2022, the completion of the PZ server (S4.2) in 2023, and the completion of the PZ validation cooperative (S4.3) in 2024. The PZ compute (S4.4) is scheduled to start in 2024 and continue through 2025.

Overview of PZ in-kind contributions:

<https://linea-it.github.io/pz-lsst-inkind-doc/>

Motivation:


Data Management Roadmap to photo-z for LSST Objects (ls.st/dmtn-049)



Previous experience with Photo-z
Pipelines and production of value-added
catalogs in the DES Science Portal



An LSST.org project



A Roadmap to Photometric Redshifts for the LSST Object Catalog

2022-02-08
main
Change version

M. L. Graham, J. Bosch, L. P. Guy,
and the DM System Science Team.

Abstract

This roadmap guides the Rubin Observatory Data Management (DM) team's efforts to engage with the scientific community of data-rights holders in order to implement and validate one or more existing photometric redshift (photo-z) estimators into the Data Release (DR) processing pipeline, and serve the resulting photo-z data products for the DR Object catalogs.

The DR Object catalog photo-z estimates will initially meet a set of minimum scientific attributes and serve the widest variety of science applications. DM will select one or more existing, community-vetted algorithms to

A Roadmap to Photo... 30 / 40 90% + [Icons]

photometric redshifts (photo-z) estimates from LSST data. This roadmap guides the Rubin Observatory Data Management (DM) team's efforts to engage with the scientific community of data-rights holders in order to implement and validate one or more existing photometric redshift (photo-z) estimators into the Data Release (DR) processing pipeline, and serve the resulting photo-z data products for the DR Object catalogs.

Access Methods – The user experience is one of the proposed selection criteria for the LSST photo-z estimator (§ 4). Some examples of publicly released photo-z catalogs which were

24




Photo-z for LSST Objects | DM-TN-049 | Latest Revision: 2022-02-08

provided with a user experience that might be desirable for the LSST photo-z include the Dark Energy Survey's Science Portal to serve photometric redshifts Gschwend et al. (2018) and the Hyper SuprimeCam Subaru Strategic Program Tanaka et al. (2018)⁶.

If the LSST photo-z are not made available in either the Objects table or in a federated or joinable catalog – for example in the case where a community-generated photo-z catalog is replacing the DMS-generated catalog (Appendix 2.2) – and are instead made available via, e.g., a “photo-server” (as in Gschwend et al. (2018)), then at least the Object catalog ID of the most recent data release should be a queryable parameter.

If the results of multiple estimators are generated, compressed, and stored in the Objects table, then decompression should be straightforward for the user (Appendix B.4).

Documentation – Appropriate types of documentation might include published journal articles, GitHub repositories, websites, or other online documentation resources (e.g., <https://readthedocs.org/>). Whatever the format, the documentation contents should include:

- general description of the estimator
- adaptations made to ingest LSST data (compared to past applications)

Motivation: forecasted use cases

Overview of PZ in-kind contributions:

<https://linea-it.github.io/pz-lsst-inkind-doc/>

BRA-LIN-S4

Search docs

Home

BRA-LIN-S4.1

Introduction

Overview of planned remote service

Forecasted use cases

BRA-LIN-S4.2

BRA-LIN-S4.3

BRA-LIN-S4.4

« Previous

Next »

Forecasted use cases

1. The user retrieves specific spec-z catalogs from the PZ Server or other online services such as [Astroquery](#) (Ginsburg, Sipőcz, Brasseur, et al. 2019). The PZ Server will provide a list of previous public spec-z catalogs of interest (previously informed by the community), for which metadata and quality flags conversion will already be available.
2. The user combines different spec-z catalogs into a single table. TSM will provide a mechanism to perform spatial cross-matching among multiple tables to resolve multiple spec-z measurements for the same galaxies (flexible criteria) and standardize quality flags from different surveys to create spec-z compilations with redshifts from different sources.
3. The user retrieves spec-z compilations from the PZ Server. The PZ Server will host a standardized compilation of spec-z with public data available up to date on each LSST Data Release, vetted by LSST DM staff. These compilations must be easy to find (e.g., flagged as "latest") and contain detailed documentation to cite the original sources properly.
4. The user combines a spec-z table with photometric data from the LSST Objects Catalog to build a training set for photo-z codes. TSM will provide a mechanism to submit jobs to LineA's HPC cluster to perform spatial cross-matching between a spec-z table registered on the PZ Server and the copy of the LSST Object Catalog hosted in the Brazilian IDAC. The matching results will be registered on the PZ Server as a new data product. Metadata and data access instructions will be sent back to the user.
5. The user splits the matched spec-photo catalog into two or more subsamples (for training and validation/test purposes). TSM will provide a simple method to split the data randomly into a finite number of parts, with proportions defined by the user. Additional methods that follow specific science-driven criteria to define subsamples or apply some transformation of the data (e.g., weighting or data augmentation) are not part of the scope of the in-kind contribution but can be added to the library later. Contributions from the community are very welcome.
6. The LSST PZ Commissioning Team uses the TSM to create standardized training and validation sets for the *Photo-z Validation Cooperative* and distribute them to the community. The resulting catalogs will be formatted according to the LSST requirements described in the [DMTN-049 - A Roadmap to Photometric Redshifts for the LSST Object Catalog](#) (to be defined by LSST Data Management (DM) System Science Team) and contain all the provenance information necessary to be reproduced using the same TSM tools by any user.

« Previous

Next »

Summary of in-kind contribution BRA-LIN S4: Photo-z (PZ) Services

Contribution Lead: Julia Gschwend

Recipient: Rubin Observatory DM team (PZ Coord. Group)

Contact: Melissa Graham

4 Deliverables in Section S4:

S4.1 Standardized training/validation sets

- Data products (recurrent): spec-z compilation, matched spec+photo set
- Software: Python package **Training Set Maker**
- Cross-matching as a service using IDAC's infrastructure

S4.2 PZ Server

- Data management tool and user interface to facilitate sharing results among RSP users (100% software, development/maintenance).
- Web-based platform and API (Python library)

S4.3 Photo-z Validation Cooperative

- FTEs to help on the upcoming Photo-z Validation Cooperative: organization of validation results, write reports, prepare notebooks.

S4.4 PZ tables

- Data products (recurrent): PZ tables as federated datasets available in the Brazilian IDAC.
- Software: pipeline PZ Compute (wrapping of PZ codes and run for massive datasets using the IDAC's computing resources).

The PZ Server



PZ Server overview

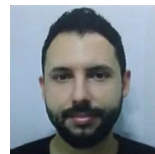
A specialized service for creating and hosting pz-related derived data products



Glauber Vila-Verde
Front-end developer



Jandson Vitorino
Front-end developer



Cristiano Singulani
Back-end developer



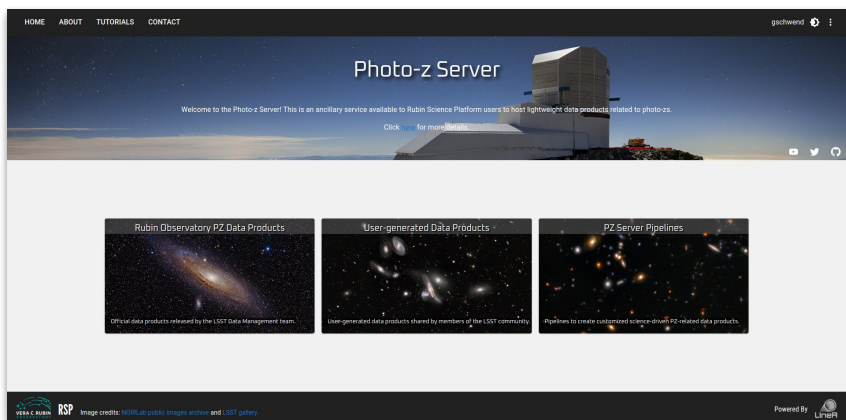
Luigi Silva
Scientist



Julia Gschwend
Scientist
Contribution Lead

<https://pzserver.linea.org.br>

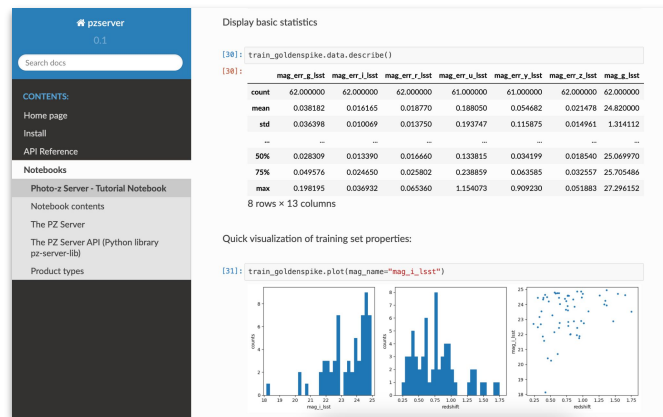
(test environment at <https://pzserver-dev.linea.org.br>)



Repo: [linea-it/pzserver_app](https://github.com/linea-it/pzserver_app)

<https://linea-it.github.io/pzserver/>

Python Library (API) \$ pip install pzserver



Repo: [linea-it/pzserver](https://github.com/linea-it/pzserver)

PZ Server - A&A with CILogon and Rubin Auth. API



Glauber Vila-Verde
Front-end developer




Cristiano Singulani
Back-end developer

+ IDAC IT team



Welcome to PZ Server

 LOGIN WITH CILOGON (RSP ACCOUNT)

 LOGIN WITH GITHUB

Any problem authenticating or registering? [Contact our helpdesk.](#)

Training Set Maker



PZ Server pipelines

The *Training Set Maker* service



Jandson Vitorino
Front-end developer



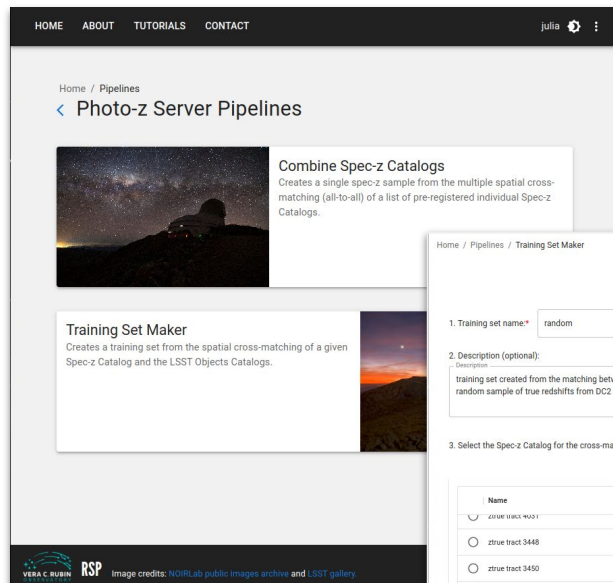
Cristiano Singulani
Back-end developer



Luigi Silva
Scientist



Julia Gschwend
Scientist



Home / Pipelines / Training Set Maker

Training Set Maker

1. Training set name*

2. Description (optional):
Description

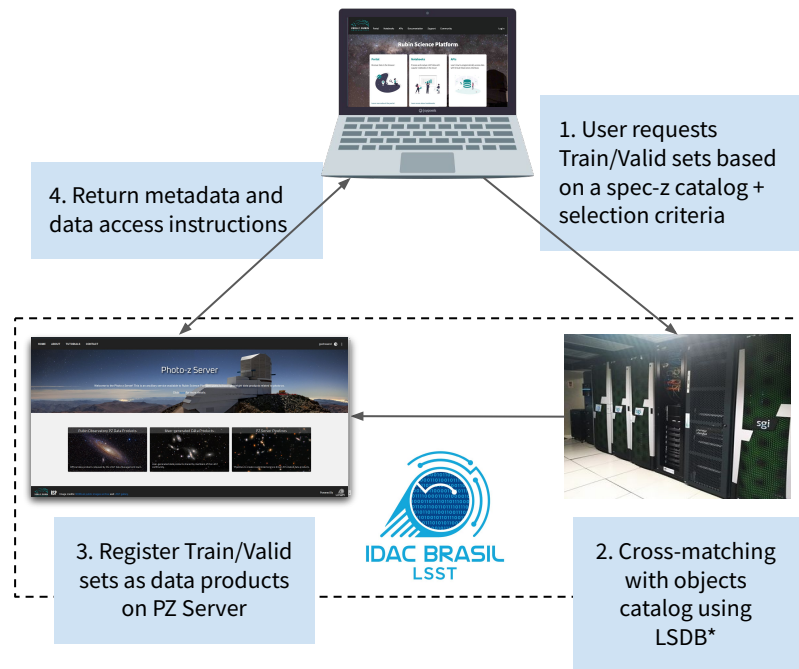
3. Select the Spec-z Catalog for the cross-matching:

Name	Uploaded By	Created at
<input checked="" type="radio"/> ztrue track 1448	gcschwend	2024-01-14
<input type="radio"/> ztrue track 3448	gcschwend	2025-01-03
<input type="radio"/> ztrue track 3450	gcschwend	2025-01-03
<input type="radio"/> ztrue track 3833	gcschwend	2025-01-03
<input type="radio"/> DESI EDR	gcschwend	2024-10-15
<input checked="" type="radio"/> random truth z	gcschwend	2024-09-12

1 row selected 1-10 of 30

4. Select the Objects catalog (photometric data):

5. Select the cross-matching configuration choices:
The threshold distance in arcseconds beyond which neighbors are not added:



PZ Server pipelines

The *Training Set Maker* service

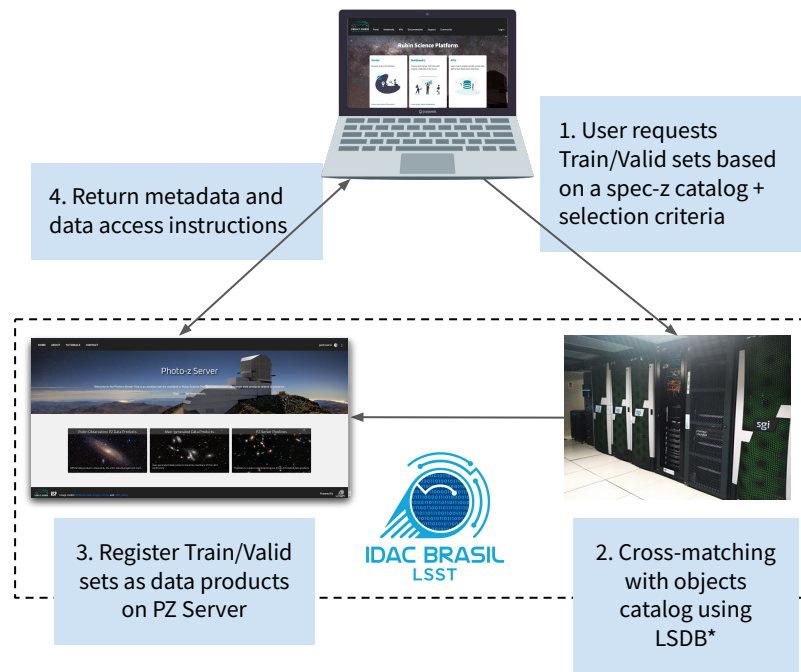
Cross-matching service to create Training Sets on demand

Back-end:

- Orchestration system to receive users requests and run processes at the IDAC
 - Data partitioning based on **HATS**
 - **LSDB** as cross-matching tool
 - Training Sets created become available as a PZ Server data product
- } developed by LINCC team

Front-end :

- PZ Server API (methods in PZ Server class)
- Pipelines page on the PZ Server



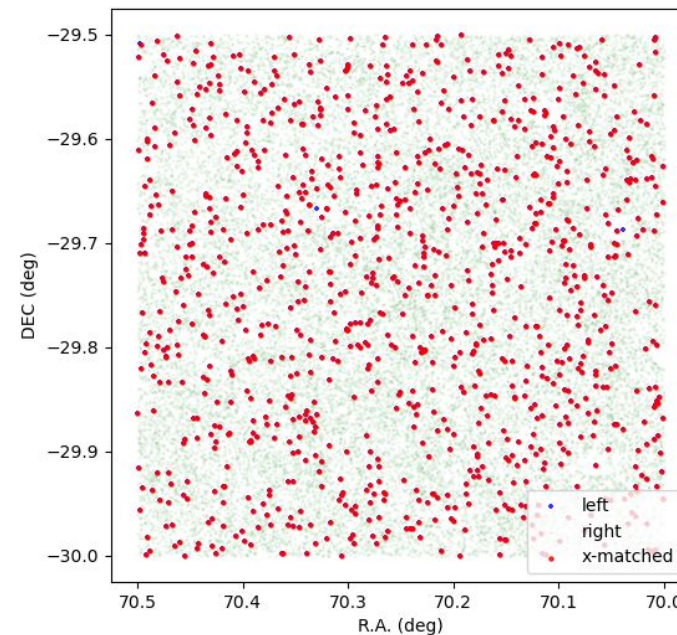
HATS and LSDB @LineA report



Luigi Silva
Scientist



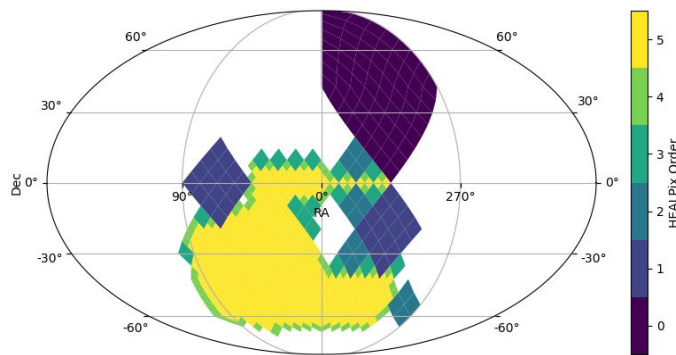
DES DR2 x Gaia DR3



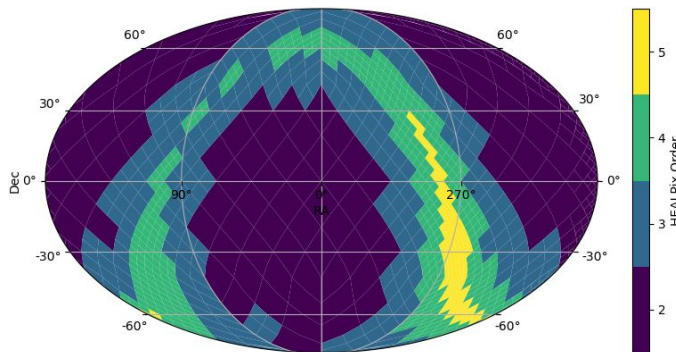
<https://docs.lsd.io/en/stable/>



DES
DR2



Gaia
DR3



PZ Compute

PZ tables as federated datasets



PZ Compute

PZ Tables as federated datasets

PZ Compute pipeline optimizes the usage of LSSTDESC/RAIL's PZ estimators on the IDAC's cluster HPE Apollo 2000 using Slurm



Henrique Dante
Software engineer



Cristiano Singulani
Back-end developer



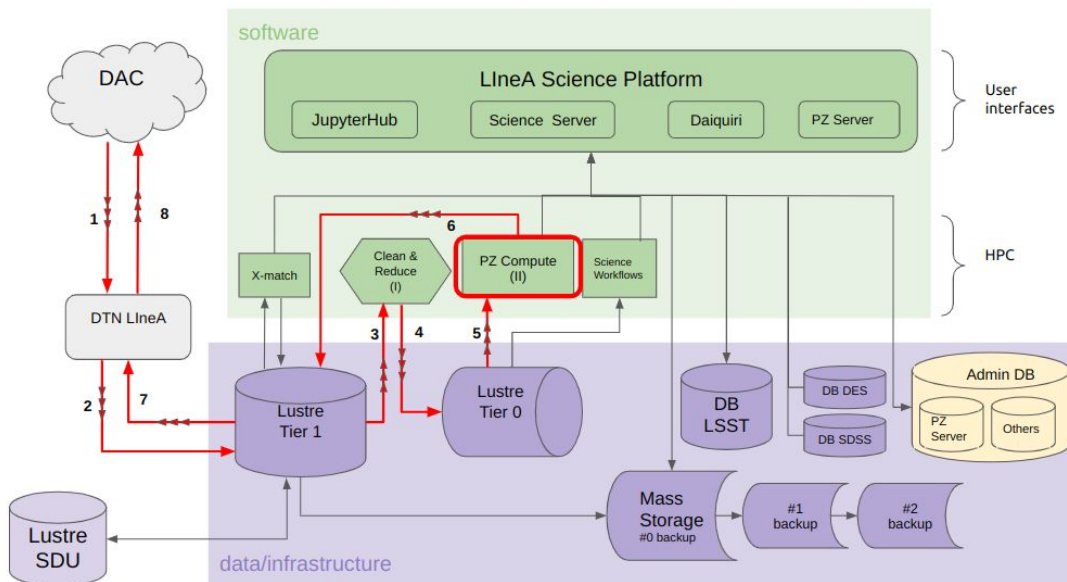
Heloisa Mengisztiki
Back-end Developer



Luigi Silva
Scientist



Julia Gschwend
Scientist



[linea-it/pz-compute](https://github.com/linea-it/pz-compute)



PZ Compute

PZ Tables as federated datasets

Summary of PZ Compute stages

Stage I - Data acquisition

- Data transfer from USDF
- Organize/split into data partitions, data cleaning
- Download and store ancillary files

Stage II - Photo-z pre-processing

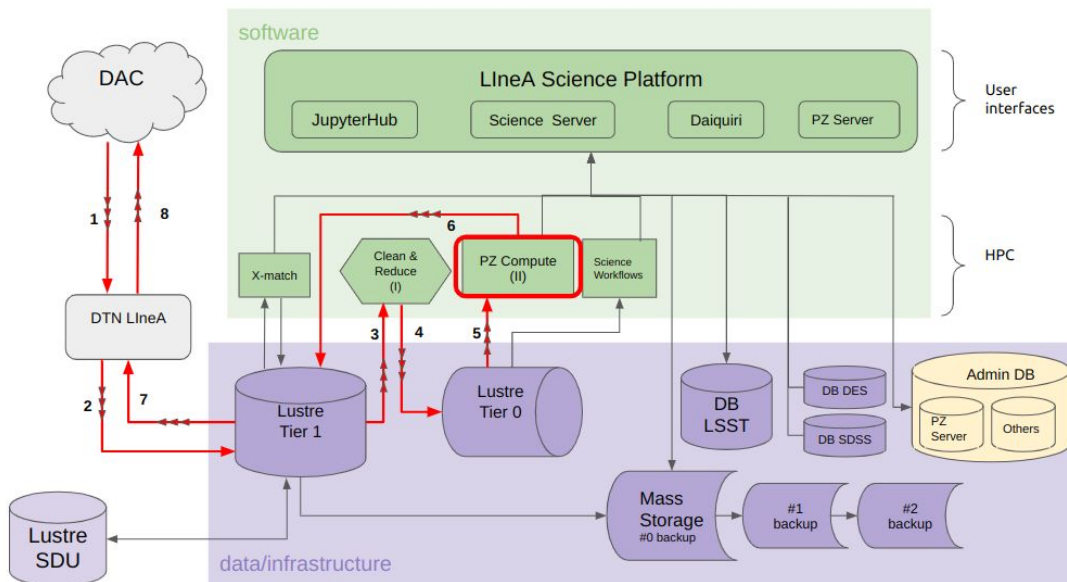
- Training PZ machine learning methods
- PZ scientific validation

Stage III - Photo-z computing

- Execute PZ codes (use RAIL as code wrapper)
- Validate results with Rubin PZ Coord.

Stage IV - Photo-z post-processing

- Data transfer from BR IDAC to USDF
- Register on RSP as a federated dataset.



[linea-it/pz-compute](https://github.com/linea-it/pz-compute)

Undergrad science projects



Heloisa Mengisztik
Back-end Developer
(supervisor: Dr. Julia Gschwend)

PCW 2023

Optimizing software infrastructure to compute photo-zs in the LSST scale: preparing for DR1.
VERA C. RUBIN OBSERVATORY
H. S. Mengisztik^{1,2}, J. Gschwend¹, C. Singulat¹, H. Dante¹
¹Laboratório Interinstitucional de e-Astronomia (LInEA), ²Instituto Federal de Santa Catarina - IFSC

Background
Many of the LSST science cases will rely on photometric redshifts (photo-z) computed for the whole LSST Objects Catalog. As an in-kind contributor, LInEA is responsible for providing annual photo-z tables for LSST data releases. To fulfil this task, a dedicated photo-z production pipeline, pz-compute, is currently under development and testing. Pz-compute wraps the modules from the DES's open-source project RAIL into a workflow to process large volumes of data in the Brazilian LSST IDAC infrastructure. The current version counts with two algorithms to calculate redshifts, BPZ-light (template fitting) and the FlexBoost (machine learning), along with support for parallel execution using HTCondor and Slurm to schedule tasks.

Objectives and research question
• Use the DP0.2 data to model the total duration of computing photo-z at LInEA's computer cluster Apollo as a function of a list of variables, e.g., dataset size.
• Run stress tests to identify bottlenecks, sources of slowness, and opportunities for workflow optimization in preparation for LSST DR1.

Methods
• Execution: scalability tests varying pipeline parameters and data characteristics under controlled conditions.
• Data analysis: statistical tests on the measurements of total runtimes and execution speed per process, photo-z results validation.

Preliminary Results
• Test with different samples of DP0.2 showed strong evidence for a linear relationship between total runtime and dataset size:

FlexBoost
 $r=0.9812$, p -value=0.003
speed= 2.08 ± 0.1 min/ds
DR1: ~208 ds in ~37 h

BPZ
 $r=0.9827$, p -value=0.002
speed= 1.58 ± 0.01 min/ds
DR1: ~208 ds in ~27 h

Fig. 3: Dataset size versus total runtime.

Fig. 2: Left: Fit Complete speed distribution with and without rounding decimal cases. Right: Check impact on photo-z results. Stacked RALP and visually the same.

• Rounding input magnitudes to 4 decimal cases reduces the total runtime by ~20% for FlexBoost and only ~2% for BPZ without jeopardizing the photo-z results (small fraction of outliers: 0.35% are >1σ from the result without rounding).

• No evidence of significant impact on the total runtime of photo-z estimation due to the:
◦ size of Training set size used
◦ posterior PDF resolution
◦ load and write hardware (IO vs. SSD)

Fig. 4: No gap in system speed and total runtime for the three cases tested above.

Discussion
• The Apollo cluster is currently turned off, undergoing an upgrade with the purchase of new machines and installing a new scheduler (Slurm). The DR1 forecast did not consider the rounded input data and cluster upgrade, so future runs should show better results.
• The tests focused on optimizing the workflow speed using the algorithm's default configuration without considering the quality of the photo-z outputs (future work).

What next:
• Repeat and perform new tests on the new infrastructure using Slurm with larger datasets (DR1-like).
• Test with different photo-z algorithms to verify if we can generalize the lessons learned for the two classes of photo-z algorithms (machine learning / template fitting).

Participate in the LSST Discovery Alliance's program for student researchers at the 2023 Rubin PCW.



Andreia Dourado
Undergrad Student
(supervisor: Prof. Bruno Moraes)

RCW 2024

Generating photometric redshift catalogs with DP0.2 data: an end-to-end study
VERA C. RUBIN OBSERVATORY
A. Dourado^{1,2}, I. Lopes¹, B. Moraes¹, J. Gschwend¹
LInEA (Team: H. Dante, H. Mengisztik, L. Silva, C. Singulat)
¹Laboratório Interinstitucional de e-Astronomia (LInEA), ²Universidade Federal do Rio de Janeiro (UFRJ)

Background and summary
Accurate and precise photometric redshifts are crucial for several scientific applications in Rubin LSST. Thus, the efficient generation and distribution of robust, reproducible redshifts is crucial. In this project, we perform an end-to-end production run using RAIL, through the pipeline pz-compute (under development as part of the BRA-LIN in-kind contribution program) to create a photometric redshift catalog on Rubin DP0.2 data. We use TPZ, a popular supervised machine learning library based on random forests that have been shown to perform well on cosmological analyses. We train, validate, and produce photometric redshifts using the Brazilian IDAC facility, focusing both on TPZ performance as assessed by key commonly used metrics and on the computational performance of the full-generation pipeline.

Data preparation
• Select 1 million objects from the Object Table, with $\text{mag}_{\text{r}} > 0.2$ and without blending, and randomly choose 100,000 objects from this sample to train and validate the algorithm.
• Apply the following quality cuts in the random sample: $\text{mag}_i < 24$ and $\text{mag}_j > 16$.
• Apply dereddening based on the SFD08 dust map.
• Split data into training and test samples (70/30 proportion)

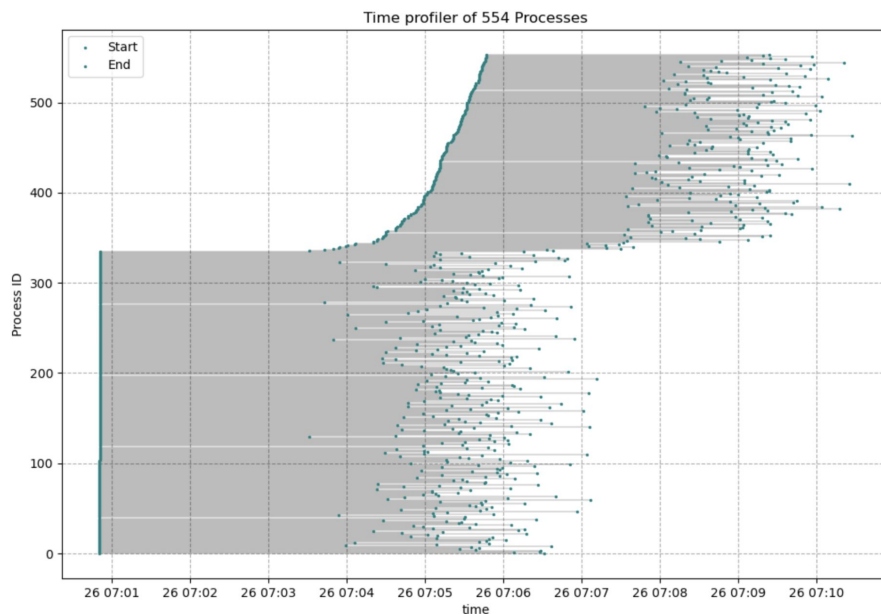
Preliminary Results
• Validation using 100 trees for the random forest on training
• 100 trees vs. 30 trees

Discussion
This study performed the end-to-end process for producing PZ tables in scale, starting from data preparation and finishing with a DP0.2 full photo-z PDF catalog. Along the journey, we identified important bottlenecks, raised and reported technical issues to both LSST DM and LInEA teams, thus contributing to the development of a solid PZ production infrastructure. All the results shown above are very preliminary but serve as a proof of concept for future PZ production during the survey's operations.

What next:
• Use the PZ Server's Training Set Maker pipeline which, is LInEA's LSDB x-matching tool, to create new training sets.
• Investigate ways to mitigate memory issues when using complex training files (~100 TPZ trees).
• Fine-tuning on TPZ configurations (e.g., varying number of trees and random realizations).
• Apply quality cuts to the photometric data and evaluate the results for the whole catalog using DC2 true redshifts.

Participate in the LSST Discovery Alliance's program for student researchers at the 2024 RCW.

PZ Compute performance tests



Latest results of performance tests (DR11* forecast):

BPZ**	FlexZBoost	TPZ Lite	LePhare
~7 hours	~9 hours	~9 hours	~24 days

* multiple copies of DP0.2 to mimic a **40 billion objects** sample

** unrealistic configuration (default = 8 templates)

Optimal[†] input data partition (skinny tables):

❖ 130k objects/parquet file or ~15 MB

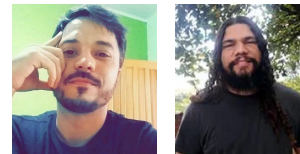
[†] considering IDAC's resources, RAIL's requirements and memory usage requirements including outputs and temporary files.

Next algorithm to test: **DNF**

Contribuição
não coberta
pelo acordo



Além da obrigação: Rubin Data Management's ConsDB



<https://dmtn-227.lsst.io/>

Considerações Finais



Resumo

O programa de contribuições in-kind BRA-LIN inclui:

- ❖ Um centro regional de dados com infraestrutura própria, cópias locais dos dados e plataformas científicas dedicadas para a comunidade LSST.
- ❖ Diversos projetos de desenvolvimento de software para apoiar a ciência feita com dados do LSST.
- ❖ Produção de produtos científicos sob demanda e com entregas periódicas previstas para todo o período de comissionamento e operação do LSST.
- ❖ Recursos humanos - desenvolvedores experientes em software para astronomia e cientistas de dados com mestrado e doutorado em astronomia.

Além disso, vale ressaltar que:

- ❖ Todo software produzido como contribuição in-kind é código aberto (disponível na organização [linea-it](https://github.com/linea-it) no GitHub)
- ❖ Suas sugestões são bem vindas. Conte-nos sobre o seu caso de uso: julia@linea.org.br



Obrigada pela sua atenção!

Contato: julia@linea.org.br

<https://www.linea.org.br>

LIneA é apoiado por:

