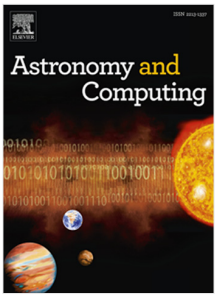
Accepted Manuscript



DES science portal: Computing photometric redshifts

J. Gschwend, A.C. Rossel, R.L.C. Ogando, A.F. Neto, M.A.G. Maia, L.N. da Costa, M. Lima, P. Pellegrini, R. Campisano, C. Singulani, C. Adean, C. Benoist, M. Aguena, M. Carrasco Kind, T.M. Davis, J. de Vicente, W.G. Hartley, B. Hoyle, A. Palmese, I. Sadeh, T.M.C. Abbott, F.B. Abdalla, S. Allam, J. Annis, J. Asorey, D. Brooks, J. Calcino, D. Carollo, F.J. Castander, C.B. D'Andrea, S. Desai, A.E. Evrard, P. Fosalba, J. Frieman, J. García-Bellido, K. Glazebrook, D.W. Gerdes, R.A. Gruendl, G. Gutierrez, S. Hinton, D.L. Hollowood, K. Honscheid, J.K. Hoormann, D.J. James, K. Kuehn, N. Kuropatkin, O. Lahav, G. Lewis, C. Lidman, H. Lin, E. Macaulay, J. Marshall, P. Melchior, R. Miquel, A. Möller, A.A. Plazas, E. Sanchez, B. Santiago, V. Scarpine, R.H. Schindler, I. Sevilla-Noarbe, M. Smith, F. Sobreira, N.E. Sommer, E. Suchyta, M.E.C. Swanson, G. Tarle, B.E. Tucker, D.L. Tucker, S. Uddin, A.R. Walker

PII: S2213-1337(18)30089-1
DOI: https://doi.org/10.1016/j.ascom.2018.08.008
Reference: ASCOM 245

To appear in: *Astronomy and Computing*

Received date : 14 June 2018
Accepted date : 29 August 2018

Please cite this article as: Gschwend J., Rossel A.C., Ogando R.L.C., A.F.Neto A.F.Neto., Maia M.A.G., da Costa L.N., Lima M., Pellegrini P., Campisano R., Singulani C., Adean C., Benoist C., Aguena M., Kind M.C., Davis T.M., de Vicente J., Hartley W.G., Hoyle B., Palmese A., Sadeh I., Abbott T.M.C., Abdalla F.B., Allam S., Annis J., Asorey J., Brooks D., Calcino J., Carollo D., Castander F.J., D'Andrea C.B., Desai S., Evrard A.E., Fosalba P., Frieman J., García-Bellido J., Glazebrook K., Gerdes D.W., Gruendl R.A., Gutierrez G., Hinton S., Hollowood D.L., Honscheid K., Hoormann J.K., James D.J., Kuehn K., Kuropatkin N., Lahav O., Lewis G., Lidman C., Lin H., Macaulay E., Marshall J., Melchior P., Miquel R., Möller A., Plazas A.A., Sanchez E., Santiago B., Scarpine V., Schindler R.H., Sevilla-Noarbe I., Smith M., Sobreira F., Sommer N.E., Suchyta E., Swanson M.E.C., Tarle G., Tucker B.E., Tucker D.L., Uddin S., Walker A.R., DES science portal: Computing photometric redshifts. *Astronomy and Computing* (2018), https://doi.org/10.1016/j.ascom.2018.08.008

DES Science Portal: Computing Photometric Redshifts

J. Gschwend[a,b,*], A. Carnero Rossel[a,b], R. L. C. Ogando[a,b], A. Fausti Neto[c,b], M. A. G. Maia[a,b], L. N. da Costa[a,b], M. Lima[d,b], P. Pellegrini[a,b], R. Campisano[b,e], C. Singulani[a,b], C. Adean[b], C. Benoist[f,b], M. Aguena[b], M. Carrasco Kind[g,h], T. M. Davis[i,j], J. de Vicente[k], W. G. Hartley[l], B. Hoyle[m], A. Palmese[l,n], I. Sadeh[o], T. M. C. Abbott[p], F. B. Abdalla[l,q], S. Allam[n], J. Annis[n], J. Asorey[i,j,r], D. Brooks[l], J. Calcino[i], D. Carollo[s], F. J. Castander[t,u], C. B. D'Andrea[v], S. Desai[w], A. E. Evrard[x,y], P. Fosalba[t,u], J. Frieman[n,z], J. García-Bellido[aa], K. Glazebrook[r,j], D. W. Gerdes[x,y], R. A. Gruendl[g,h], G. Gutierrez[n], S. Hinton[i,j], D. L. Hollowood[ab], K. Honscheid[ac,ad], J. K. Hoormann[i], D. J. James[ae], K. Kuehn[af], N. Kuropatkin[n], O. Lahav[l], G. Lewis[ag], C. Lidman[af], H. Lin[n], E. Macaulay[ah], J. Marshall[ai], P. Melchior[aj], R. Miquel[ak,al], A. Möller[am,j], A. A. Plazas[an,ao], E. Sanchez[k], B. Santiago[ap,b], V. Scarpine[n], R. H. Schindler[aq], I. Sevilla-Noarbe[k], M. Smith[ar], F. Sobreira[as,b], N. E. Sommer[am,j], E. Suchyta[at], M. E. C. Swanson[h], G. Tarle[y], B. E. Tucker[am,j], D. L. Tucker[n], S. Uddin[au], A. R. Walker[p]

[a] *Observatório Nacional, Rua General José Cristino, 77, Rio de Janeiro, RJ, 20921-400, Brazil*
[b] *Laboratório Interinstitucional de e-Astronomia - LIneA, Rua General José Cristino, 77, Rio de Janeiro, RJ, 20921-400, Brazil*
[c] *LSST Project Management Office, Tucson, AZ, USA*
[d] *Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP, 05314-970, Brazil*
[e] *Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ, Av. Maracanã, 229, Rio de Janeiro, RJ, 20271-110, Brazil*
[f] *Laboratoire Lagrange, Université Côte dAzur, Observatoire de la Côte dAzur, CNRS, Blvd de lObservatoire, CS 34229, 06304 Nice cedex 4, France*
[g] *Department of Astronomy, University of Illinois, 1002 W. Green Street, Urbana, IL 61801, USA*
[h] *National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA*
[i] *School of Mathematics and Physics, University of Queensland, QLD 4072, Australia*
[j] *ARC Centre of Excellence for All-sky Astrophysics (CAASTRO), Australia*
[k] *Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Avda. Complutense 40, E-28040, Madrid, Spain*
[l] *Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
[m] *Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr. 1, D-81679 München, Germany*
[n] *Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[o] *Deutsches Elektronen-Synchrotron (DESY), Platanenallee 6, 15738 Zeuthen, Germany.*
[p] *Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, Casilla 603, La Serena, Chile*
[q] *Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa*
[r] *Centre for Astrophysics and Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia*
[s] *INAF - Astrophysical Observatory of Turin, Italy*
[t] *Institut d'Estudis Espacials de Catalunya (IEEC), 08193 Barcelona, Spain*
[u] *Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain*
[v] *Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[w] *Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India*
[x] *Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
[y] *Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[z] *Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[aa] *Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, 28049 Madrid, Spain*
[ab] *Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA*
[ac] *Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
[ad] *Department of Astronomy, The Ohio State University, Columbus, OH 43210, USA*
[ae] *Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA*
[af] *Australian Astronomical Observatory, North Ryde, NSW 2113, Australia*
[ag] *Sydney Institute for Astronomy, School of Physics, A28, The University of Sydney, NSW 2006, Australia*
[ah] *Institute of Cosmology & Gravitation, University of Portsmouth, Portsmouth, PO1 3FX, UK*
[ai] *George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA*
[aj] *Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA*
[ak] *Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain*
[al] *Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra (Barcelona) Spain*
[am] *Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia*
[an] *Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA*
[ao] *Astronomical Society of the Pacific, 100 N Main St., Suite 15, Edwarsville, IL 62025, USA*
[ap] *Instituto de Física, UFRGS, Caixa Postal 15051, Porto Alegre, RS - 91501-970, Brazil*
[aq] *SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[ar] *School of Physics and Astronomy, University of Southampton, Southampton, SO17 1BJ, UK*
[as] *Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas, Campinas, SP, 13083-859, Brazil*
[at] *Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[au] *Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing, Jiangsu, China*

Abstract

A significant challenge facing photometric surveys for cosmological purposes is the need to produce reliable redshift estimates. The estimation of photometric redshifts (photo-zs) has been consolidated as the standard strategy to bypass the high production costs and incompleteness of spectroscopic redshift samples. Training-based photo-z methods require the preparation of a high-quality list of spectroscopic redshifts, which needs to be constantly updated. The photo-z training, validation, and estimation must be performed in a consistent and reproducible way in order to accomplish the scientific requirements. To meet this purpose, we developed an integrated web-based data interface that not only provides the framework to carry out the above steps in a systematic way, enabling the ease testing and comparison of different algorithms, but also addresses the processing requirements by parallelizing the calculation in a transparent way for the user. This framework called the Science Portal (hereafter Portal) was developed in the context the Dark Energy Survey (DES) to facilitate scientific analysis. In this paper, we show how the Portal can provide a reliable environment to access vast data sets, provide validation algorithms and metrics, even in the case of multiple photo-zs methods. It is possible to maintain the provenance between the steps of a chain of workflows while ensuring reproducibility of the results. We illustrate how the Portal can be used to provide photo-z estimates using the DES first year (Y1A1) data. While the DES collaboration is still developing techniques to obtain more precise photo-zs, having a structured framework like the one presented here is critical for the systematic vetting of DES algorithmic improvements and the consistent production of photo-zs in future DES releases.

Keywords

astronomical databases: catalogs, surveys – methods: data analysis – galaxies: distances and redshifts, statistics

1. Introduction

In the last few decades, large galaxy surveys have become one of the main research tools in astronomy, in particular, for the study of cosmology. The need for increasing statistical samples and depths have encouraged the design and construction of deeper, wider, and more sensitive surveys. These projects are generating vast amounts of data, bringing astronomy into the realm of "big data", which increases the challenges associated with cosmological analyses.

One of these projects is the Dark Energy Survey (DES, Flaugher, 2005; DES et al., 2016), a 5–year program to carry out two distinct surveys. The wide–angle survey covers 5,000 deg^2 of the southern sky in five ($grizY$) filters to a nominal magnitude limit of ∼24 in most bands. Also, there is a deep survey (i ∼26) of about 30 deg^2 in four filters ($griz$) with a well–defined cadence to search for type–Ia Supernovae (SNe Ia) (Kessler et al., 2015). The primary goal of DES is to constrain the nature of dark energy through the combination of four observational probes, namely baryon acoustic oscillations, counts of galaxy clusters, weak gravitational lensing, and determination of distances of SNe. Besides, many other fields of astrophysics benefit from the large data set generated by the survey, as detailed by DES et al. (2016).

The constraining power of cosmological results provided by DES will strongly depend on the ability to estimate reliable photometric redshifts (photo-z, e.g., Huterer et al., 2004; Ma et al., 2006; Lima and Hu, 2007; Ma and Bernstein, 2008; Hearin et al., 2010; Cunha et al., 2014; Georgakakis et al., 2014). In fact, the computation of accurate photo-zs has been one of the major concerns of the collaboration, which has spurred the implementation and testing of several algorithms. For instance, Sánchez et al. (2014) addressed the performance of several codes when applied to the DES science verification data (SVA1), while Banerji et al. (2015) discussed the effect of using infrared data. More recently, Bonnett et al. (2016) examined the impact of four photo-z algorithms on the conclusions of the first DES cosmological analysis based on weak lensing discussed by Abbott et al. (2016).

Photo-z estimation will only get more challenging for the next DES releases and future photometric surveys. The reason is that we are sampling magnitudes beyond the reach of most spectroscopic surveys and therefore, traditional photo-z validations are not realistic. This issue has inspired the implementation of new ideas in the collaboration, such as the calibration of photo-zs with cross–correlations (Newman, 2008; Davis et al., 2017; Gatti et al., 2018), the training and validation of photo-z codes with simulations (data–augmentation) (Hoyle et al., 2015) and validation of photo-zs with multi–band photometric samples (Hoyle et al., 2017). Techniques for assignment and validation of photo-zs for DES are under continuous development.

There are a large number of methods and algorithms available in the literature to compute and validate photo-zs. Thus, it is useful to work in an integrated environment where one can perform repeated tests and compare the results, while keeping the history well documented. Such an environment should provide the necessary hardware and software infrastructure to make feasible the comparison of different methods applied to large datasets.

Besides dealing with big data, another remarkable aspect of current and near–future surveys is a large number of people working collaboratively. The computational methods are developed jointly by groups of people, commonly located in different countries. Therefore, it is useful to share a development environment that organizes software with version control, keeps the history, and ensures it is possible to reproduce results at any time.

Other web–based interfaces for astrophysical data mining and analysis are also being developed (e.g., the DAMEWERE environment by Brescia et al., 2014) aiming at the exploitation of large datasets.

*Corresponding author
Email address: julia@linea.gov.br (J. Gschwend)

The DES collaboration proposed, along with the Data Management system (DESDM[1], Mohr et al., 2012), the creation of a dedicated portal to solve some of the problems associated with the data processing. This concept became the DES Science Portal, hereafter "the Portal".

During the early days of the DES project, the Portal was conceived as an "end–to–end" (E2E) process where the data flowed through a chain of tasks to prepare science–ready catalogs and perform scientific analyses. Since then the Portal has undergone several implementations for various scientific goals. The complexity of the system has been growing accordingly to accomplish the science demands. Now, there are instances of the Portal at Cerro Tololo Inter–American Observatory (CTIO), at the National Center for Supercomputing Applications (NCSA) and, at the Laboratório Interinstitucional de e–Astronomia (LIneA)[2]. In this paper, we refer to the instance at LIneA as "the Portal".

The Portal provides the infrastructure necessary to handle large amounts of data, a common demand in extragalactic astronomy, but also attacks specific needs of the DES science, for instance, creating and applying systematic maps, computing zero-point corrections, performing star-galaxy classification, computing photo-zs and galaxy properties. The Portal generates galaxy samples in the form of pruned lightweight catalogs containing only the columns required by specific science analysis, which may also be integrated into workflows (Fausti Neto et al., 2018).

In this paper, we present, in particular, the capabilities of the Portal to produce photo-zs. It provides an integrated environment where all the steps necessary to compute photo-zs can be carried out in a controlled and consistent way. The automatic provenance, configuration management, and the computing facilities that sustain the Portal allow for a selection of many photo-z algorithms or settings, which would be highly time-consuming without infrastructure such as this. The need for the Portal capabilities will increase as the DES databases grow, and more generally, as we enter an era of big data astronomy.

In Cavuoti et al. (2015), the authors of the PhotoRApToR algorithm discussed the advantage of linking automatically different steps of photo-z calculation. The Portal surpasses PhotoRApToR in the sense that it is method agnostic: any photo-z algorithm can be incorporated into the portal framework, which becomes especially interesting when the investigation aims to compare results using different methods.

We present a sequence of tasks that include the preparation of a spectroscopic sample by combining data from different redshift surveys, the creation of training sets, the training and validation procedures for several algorithms, and the computation of photo-zs for large datasets. To show these examples, we used the DES first year data release, referred as Y1A1 (Drlica-Wagner et al., 2018; Abbott et al., 2018).

[1] `http://www.darkenergysurvey.org/the-des-project/survey-and-operations/data-management/`

[2] `http://www.linea.gov.br/`

Table 1: Glossary of terms used in the description of the Portal.

Term	Meaning
Component	A Python script that works as a module to perform a specific task or to serve as a wrapper for an external algorithm.
Pipeline	A self–consistent sequence of components defined in an XML file. The pipeline script also defines dependencies between different components, as well as their required inputs and outputs.
Workflow	One or a group of pipelines oriented to a common purpose. In general, it refers to scientific pipelines (known as Science Workflows).
Class of products	A unique name that defines the attributes, characteristics and possible applications of a dataset or a product created in the Portal.
End–to–end (E2E)	Sequence or chain of pipelines running in the Portal, starting with data acquisition, passing through several steps of data preparation and estimation of value–added quantities, culminating in the production of science–ready catalogs (see).
Stage	A group of pipelines in the same phase of data management. The E2E comprises the three stages: Data Installation > Data Preparation > Catalog Creation, and it is directly connected to the Science Workflows' stage. This last stage comprises the suite of scientific pipelines to support the research done by members of DES–Brazil Consortium (details in the supplemental video V0[4]).

The outline of this paper is as follows. In Section 2 we present the general technical aspects of the Portal. In Section 3 we go deeper in details of the processes related to the production of photo-zs. Still in Section 3 we present a use case of how the Portal can aid to determine reliable photo-zs through examples of runs using data from DES. The data is described in Appendix A. Finally, our conclusions and a summary of the paper are presented in Section 4.

Also, we present, attached to this text, a list[3] of five videos (V0 to V4), showing examples of live runs, in a guided tour through the photo-z production on the Portal.

2. The DES Science Portal: Overview

Before describing the technical aspects of the Portal, we define in Table 1 a list of terms frequently used in this text.

The Portal, which the LIneA team designed, developed, hosts, operates and maintains, is an overarching web–based system solution for many issues faced by large astronomical surveys. Geographically, the operation is divided into three Portal instances running independently at CTIO, LIneA, and NCSA, as illustrated in Figure 1. Each instance is responsible for accomplishing tasks in distinct phases of the data lifecycle.

In the very beginning of the production of raw data, still at the telescope, the Portal@CTIO runs a pipeline called ***Quick***

[3] `https://www.youtube.com/playlist?list=PLGFEWqwqBauBIYa8H6KnZ4d-5ytM59vG2`

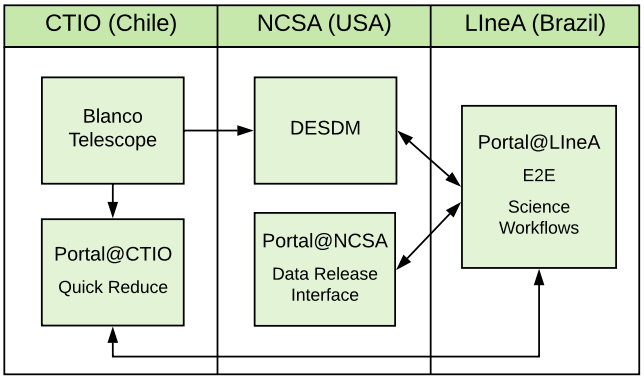


Figure 1: Instances of the DES Science Portal. The arrows indicate the data flow from the Blanco Telescope, at CTIO, through the various portal instances and the DESDM system at NCSA.

Reduce. It performs a rapid inspection of images, immediately after the exposures are taken, to detect possible problems and to produce a preliminary quality assessment. DESDM at NCSA receives the data and reduces and co–adds the raw images (Morganson and Dark Energy Survey Data Management Team, 2016). The photometry tables of objects detected from the coadded images are called *coadd tables*. After that, these tables are downloaded and ingested into the Portal@LIneA's database, where the science–ready catalogs are created. Finally, the catalogs and science products are transferred to Portal@NCSA, which provides visualization tools through the data release interface.

In this work, we only present the technical aspects of the Portal@LIneA, which are related to the production of photo-*z*s. Among these features, we highlight those that apply to many other tasks:

- Storing and registering of survey data to serve as input for analysis pipelines.
- Maintaining analysis codes (and their development history) ready to run on registered products.
- Integrating external scientific codes publicly available into science workflows.
- Facilitating the run and scalability of algorithms for scientific analysis (user–specific or collaboration–defined).
- Registering outputs as products available for download and making them usable as inputs for other codes.
- Keeping track of provenance (inputs, selected options, version of codes, etc.)
- Allowing reproducibility of analysis results by keeping documentation about data and codes used, and operations performed.

Portal Infrastructure

The Portal framework relies on two databases, a mass storage file system, a web interface, a workflow system and a cluster of computers, as illustrated in Figure 2.

Both databases uses PostgresSQL[5] object-relational database management system. The catalog database stores the catalogs retrieved from NCSA database and the catalogs produced by the Portal pipelines. The total storage capacity of the machine available for the catalog database is 23 TB, from which, ~17 TB is already occupied. This device has a hot backup duplication in another connected machine. Both will be replaced in the future by new devices with larger capacity.

The administrative database keeps track of metadata such as available releases, ingested products, product information, such as file and table names, storage location, classification, provenance, etc. All the operations and steps are logged in the administrative database, so it allows to detect errors and investigate them posteriorly, and also to produce reports on the resources usage that can be filtered by a user or by an application.

The mass storage device has the capacity for 59 TB, from which 39 TB is already used. It keeps data in three separate spaces:

- **Archive** area, where the original catalogs FITS files are preserved (so the catalogs are duplicated in the database and mass storage's archive).
- **Scratch** area, where are placed the directories and files that are created during the process executions. They include tables, images, flat files of any kind, run logs, error logs, etc. They remain in this area until the user saves the process. Periodically, a tool called *garbage collector* removes old files from this area.
- **Process** area, is a safe area where the directories and files from saved processes are kept permanently.

The location of all the process directories, both in Scratch and Process areas are registered in the administrative database.

The web application front–end uses the Model-View-Controller (MVC, Burbeck, 1987) software architecture and is developed in both Hypertext Markup Language[6] (HTML) and JavaScript [7] languages. It connects to the databases via the back-end infrastructure Python components.

Pipelines and Components

The computational tools available in the Portal are organized in pipelines and components. The former are workflows defined in Extensible Markup Language (XML, Bray et al., 2008) which concatenates a chain of tasks performed by components. They determine the order of execution of the tasks and the parallelization strategy, as well as necessary inputs and outputs.

[5] https://www.postgresql.org/
[6] `https://www.w3.org/html/`
[7] `https://www.javascript.com/`

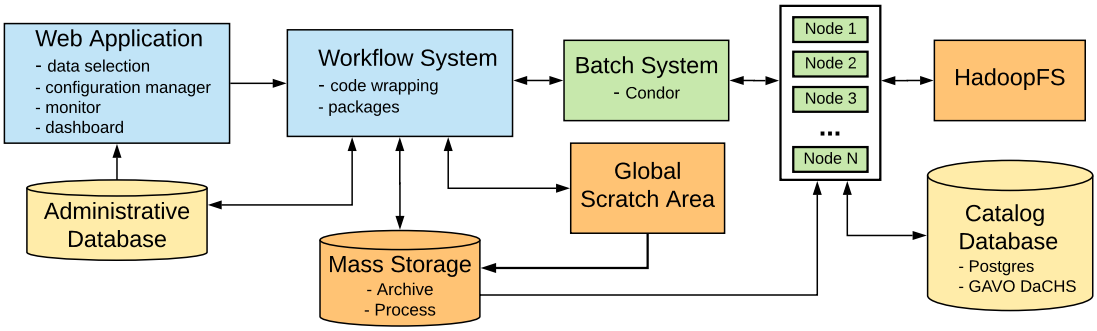


Figure 2: Elements of Portal's infrastructure: software components (in blue), processing systems (in green), and storage systems (databases in yellow, file systems in orange).

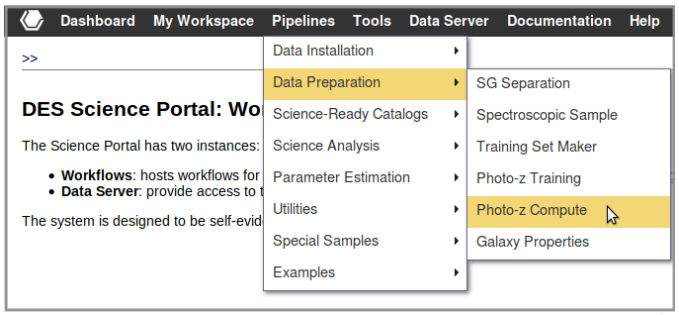


Figure 3: Portal's initial screen - data preparation pipelines menu.

Components are Python scripts which can both serve as a wrapper for an external code or be an independent algorithm. Scientific codes, which can be written in any programming language, are encapsulated by the wrappers, which are in charge of preparing the inputs in the format expected by the code, calling the code to run, and handling the outputs.

The pipelines are triggered via the Portal interface (see Figure 3), where the user navigates through tabs to define inputs and configuration parameters. For each pipeline, there is a README document that includes configuration tips and pieces of advice regarding the technical aspects of pipeline running. There are also cookbooks with a scientific approach to help the user on decisions about data selection and data-dependent configurations.

When a process finishes, the user receives a notification via e–mail with a link to a product log – a page containing results and process–relevant pieces of information.

The pipelines are self–consistent independent building blocks in the E2E chain. Each one provides a product log and can be redone as many times as the user demands. The list of the most used pipelines can be seen in Figure 4. The pipelines highlighted in gray are those related to the photo-z calculation. They will be discussed in details in Section 3. For now, it is not possible to run part of a pipeline stand–alone. For instance, one can perform repeated tests, varying configurations, until it converges into a result with the desired quality. In our case of study, we have already run training, validation, and computing for nine different photo-z codes, varying their configurations, more than 300 times in total, considering the several datasets of the Y1A1 release. That would be complicated to manage using directories and command–line runs.

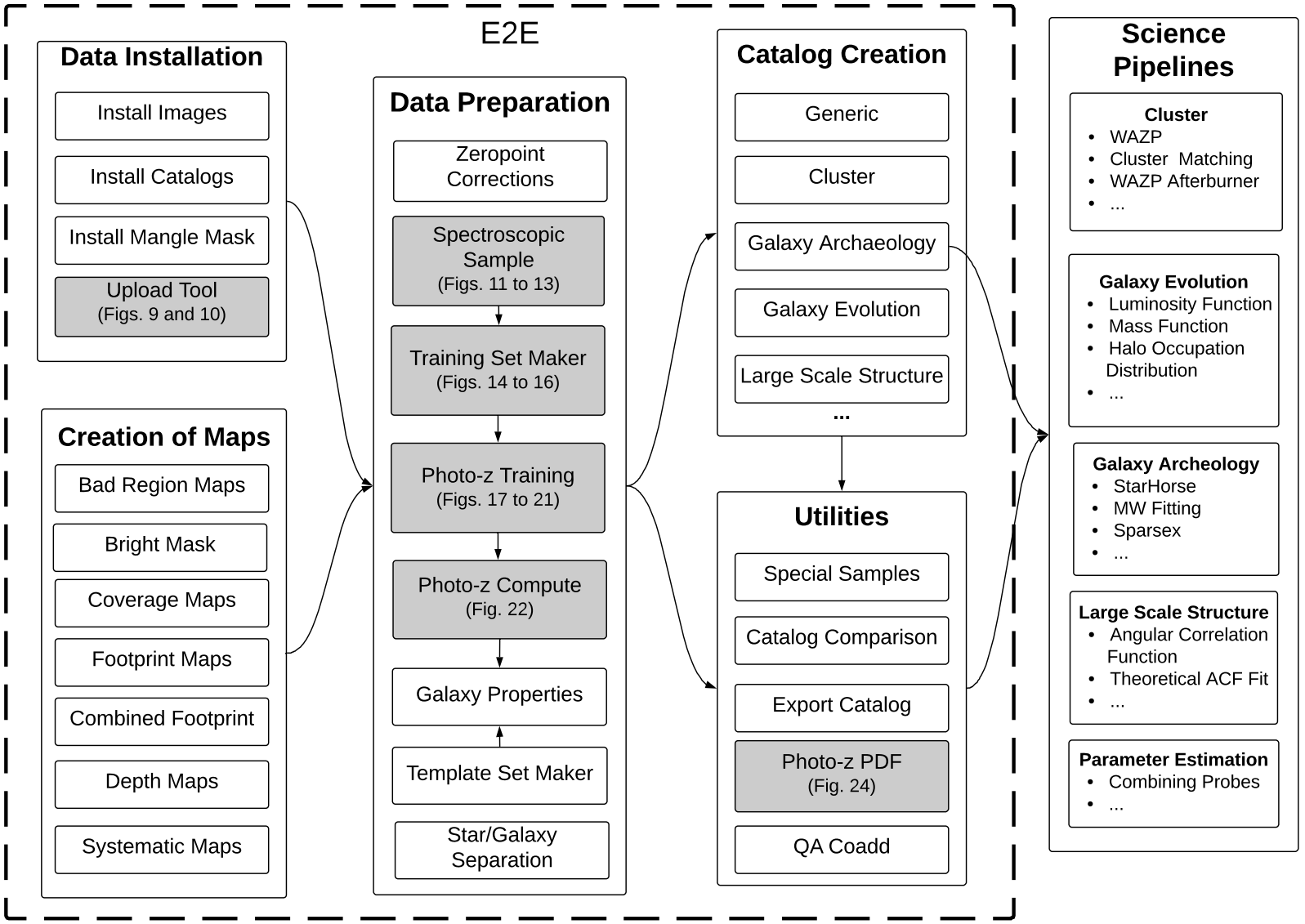


Figure 4: Sequence of pipelines organized by stages. The data flow mainly from left to right. The pipelines inside the dashed line belong to the E2E, where science-ready catalogs are produced and delivered to the Science Workflows. The pipelines highlighted in gray are those related to the photo-z calculation. The figures indicated inside the parenthesis are the respective screenshots.

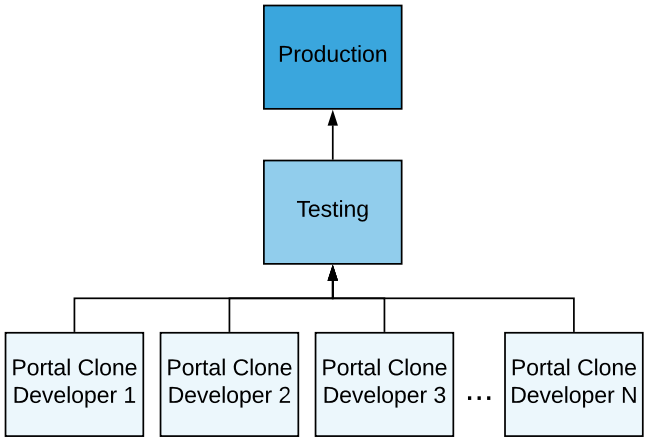


Figure 5: Portal environments: development (bottom), testing (middle), and production (top).

The Portal is developed collaboratively using the GIT[8] version control system. All changes done by different developers are merged and intensively tested in a separate Portal clone called "Testing", to ensure consistency and compatibility between developers' versions (see Figure 5). The stable and validated versions of the codes are deployed to the production Portal. All the technical information about hardware mentioned in this work refers to the production environment.

Pipelines and component codes are opensource[9]. These codes often have dependencies on other programs or libraries which are provided by two systems: Tawala and the Extended Unix Product System (EUPS[10]). Tawala is a homemade repository system created in earlier stages of portal development, which we maintain due to the large number of historical code dependencies. Nowadays it is kept frozen and is being gradually replaced by EUPS.

The pipelines of the first stage, ***Data Installation*** (the first block of pipelines in Figure 4), are only executed by the portal developers (LIneA's IT team), except by the ***Upload Tool***, that can be executed by any science user. The initial step of ***Data Installation*** is the retrieval and ingestion from NCSA, of *coadd tables*. A daemon process called ***Data Retriever*** periodically inspects the NCSA database to discover new releases of the DES Survey data. For each new release recognized, the ***Data Retriever*** registers its existence in the Portal Administrative database. At this point, the Portal system "knows" the presence of a new release and a Portal operator can start the specific installation procedure that takes care of downloading the data in an optimized network route. The download is done using a server in a "demilitarized zone network" and then storing it into a mass storage server. Later it is ingested into the Portal Catalog database. Finally, the Administrative database registers the new tables and makes them available to be accessed by visualization tools and to serve as input to be processed by pipelines in the Portal.

A system of "classes" of products connects pipelines via inputs and outputs. In this context, a class is a unique keyword that identifies a specific product data structure. Hence, we can define the type of products that each pipeline receives as input and returns as output. As an example, the catalogs mentioned above, which are ready to be used in the portal, pass through the pipeline ***Install Catalog***, where they are classified as products with class ="Object Catalog". This way, they are made available to each pipeline where we set this class as input. ***Star/Galaxy Separation*** is one of the pipelines configured to use products of this class as input, so when the user chooses to run this pipeline, the system will display the tables registered by ***Install Catalog***, and the user will be asked to select one of them. Then, running various pipelines can be understood as a hierarchy of products, as shown in Figure 6. This figure illustrates the provenance of a sequence of pipelines related to photo-z pipelines described in Section 3.

Parallelization

The photo-zs estimation is one of the most computationally intensive tasks of the E2E process, due to the size of datasets involved. The first solution adopted in the Portal is the so-called *Embarrassing Parallelization* (Herlihy and Shavit, 2011), which is the division of the data into small partitions and their processing is done simultaneously by several computers.

The large volume of data requires a high–performance system to transfer such data inside and outside each computer node avoiding the creation of an I/O bottleneck. This is a common problem in data–intensive computing that does not have a unique solution for all possible use cases. We solved this obstacle by implementing different I/O service strategies that can be used according to each specific problem, as detailed below.

For the very basic DES datasets (table of objects identified in coadded images) frequently used in the majority of the algorithms, we use the high responsive Hadoop Distributed File System (HDFS, Shvachko et al., 2010) to distribute data uniformly across the cluster nodes. Despite the fact that the processing is done by several computers, the data itself also needs to be distributed. Otherwise, the simultaneous reading from the same database by several parallel tasks (jobs) would establish a significant bottleneck.

The Portal's computers cluster contains 38 nodes with 24 central processing unit (CPU) cores each. The management of job submission is performed by an orchestration system, together with the HTCondor management system[11]. The orchestration system interprets the parallelization strategy defined by each pipeline and its configuration, then it calculates the number of jobs necessary, and gives instructions to HTCondor. The tasks are organized in the cluster nodes based on the data, such that it always prioritizes the runs to process the data that are already stored in each node, avoiding unnecessary data transfer

[8] https://git-scm.com/
[9] https://git.linea.gov.br
[10] https://github.com/RobertLuptonTheGood/eups
[11] https://research.cs.wisc.edu/htcondor

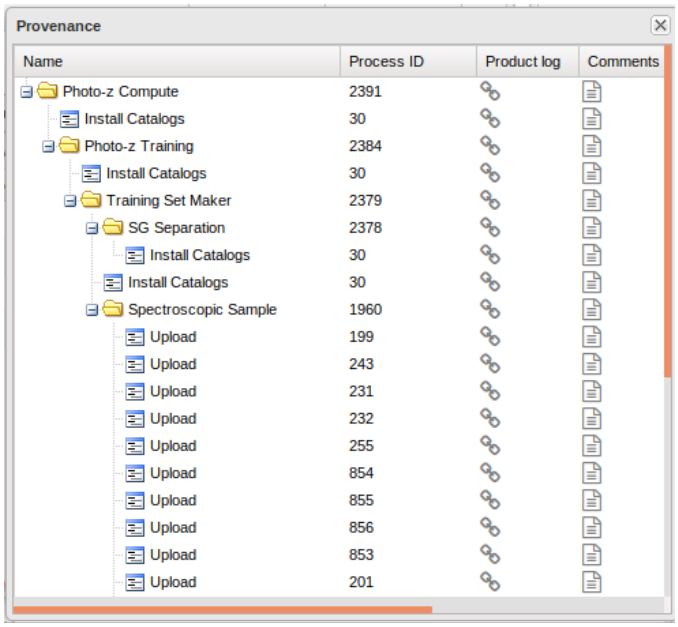


Figure 6: Display of provenance chain **for** a product of Photo-z Compute pipeline (first column). Each sub–level tag the processes that entered the parent process with their identification number (process ID) and their links to the product logs and comments made by users.

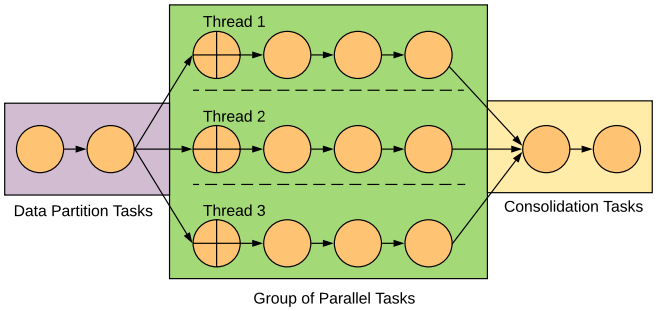


Figure 7: Illustration of parallel processing managed by the workflow system. The first group of tasks defines the data partitioning schema and distributes jobs to the cluster. Tasks that require the same chunk of data are grouped in the same cluster node, to minimize data transfer. The first component of the task group running in parallel is responsible for data retrieving (marked with a cross symbol). The final task group consolidate the results from all jobs.

overheads. If necessary, it allows reading additional data from other nodes. There is a mirroring in the data storage to help on this optimization. Each data chunk is stored triplicated in three different nodes. Hence, if one node needs data from a neighbor node that is, by chance, very busy with some intense process, the former still has two other options of nodes from where to get the data.

For less frequently accessed data used in our algorithms, the Portal use PostgreSQL[12], a fat node database with 24 cores and 256 gigabytes of RAM that supports multiple queries concurrently, allowing the fast retrieval of the portion of the data that each node needs. Moreover, to fast retrieve positional data according to its spatial position, we use the Q3C (Koposov and Bartunov, 2006) PostgreSQL extension for spatial indexing on a sphere.

For temporary data, such as the one produced in a component that needs to be consumed by the next segment of the same pipeline, it is required to be staged and then rapidly transferred from one or more nodes to others. In this case, we use the high–performance parallel Lustre file system (Donovan et al., 2003), explicitly developed for large–scale cluster computing. Finally, PostgreSQL is used to store the new generated products temporarily stored in a Network File System (NFS, Sandberg et al., 1985). An example of such a product is the product log generated by any pipeline.

The parallel processing is implemented in a "Map-Reduce" (Dean and Ghemawat, 2004) way, as illustrated in Figure 7, and it follows one of the three options of data partitioning strategies:

- **Tiles**: This is the original data division from DES. Tiles are tangent projections of an array equivalent to 5000 × 5000 CCD pixels (0.7306° on each side, Morganson and Dark Energy Survey Data Management Team, 2016). Each tile corresponds to one file stored in HDFS. The file size is strongly dependent on the density of objects detected, which is related to the depth of observations. For the Y1A1 depth, one tile contains, typically, ∼ 40k objects detected, and the file occupies ∼130 MB (for data description, see Appendix A). The number of tiles processed by each job submitted to the cluster is a free parameter in the configuration of the pipelines. The optimal choice depends on the dataset size (number of tiles), compared to the number of CPU cores available, as well as the memory consumption of the algorithm run.

- **HEALPix[13] indexation**: This is a more flexible way of splitting the entire sky projected area into pieces (pixels) of variable size. The division based on HEALPix pixels is available both for the *coadd tables*, stored in HDFS, and some of the other datasets stored in the database. Concerning the first, the data is distributed in files containing pixels of $NSide = 32$ (∼1.6 GB each for Y1A1 wide datasets), using nested pixel ordering. $NSide$ is a quantity that represents the resolution of the HEALPix map, so that the total number of pixels covering the whole sphere is $N = 12 \times (NSide)^2$. Therefore, the larger the $NSide$, the smaller the pixel size. Similarly to the tile–based partitioning, the choice of $NSide$ has implications on the processing performance. The available options for $NSide$ values are 2^n, $n = 2$ to 10. Stress tests have shown that large pixels ($n < 4$) should be avoided due to random–access memory limitations, depending on the

[12] https://www.postgresql.org/docs/9.6/static/parallel-query.html

[13] http://healpix.sourceforge.net/

dataset. On the other hand, scaling–out to use small pixels ($n > 8$) are also not recommended, because they convert in too many jobs, increasing the transferring overhead and stressing the cluster management system. One advantage of the partitioning based on HEALPix is that it applies to any data set which has celestial coordinates, therefore connected to spherical geometry. It is particularly useful to organize simulated data, that is not related to any observational strategy.

- **Custom**: This option defines the data partitioning based on one of the data attributes (i.e., any table's column). It queries the data and distributes, among the several cluster nodes, using intervals of one attribute. There are three options of binning: (i) fixed: evenly spaced bin edges; (ii) variable: evenly populated, so the bin widths are irregular; (iii) manual: bin intervals are defined manually by the user. This data partition is more commonly used by the science workflows. Some examples of this parallelization are the estimation of the angular correlation function of galaxies in tomographic bins of redshift, and the estimation of the luminosity function of galaxies in bins of absolute magnitude.

In all cases, the scaling of parallel processing is not automatic. The user is asked to make decisions on the configuration screen about the size or the number of partitions. Optionally, the second layer of parallelization can be applied as a further user–specific parallelization, by reserving one or more machines and distributing the process among their cores using, e.g., Python Multiprocessing. All this flexibility is put in place to optimize the pipeline execution in various scenarios, depending on the peculiarities of each run. Too large data chunks can cause memory over–heading problems or take too long to be processed, while too small pieces can waste time with data transfer and create a massive queue of jobs waiting for available nodes. Therefore, the optimal parallelization schema must be defined case by case. In Section 3.4 we show, as an example, a test to measure the impact of the configuration chosen on the processing speed, measured by the processes total duration and the time spent in groups of components.

3. Photo-z Pipelines

The estimation of redshifts is a fundamental part of the process of creating science–ready catalogs for extragalactic applications. In photometric surveys, photo-z methods and algorithms are used to surpass the lack of spectroscopic information. In most cases, the photo-z estimation and validation rely on the use of a "true" sample, in the sense of assuming negligible errors in the determination of their redshifts, as in the spectroscopic redshift samples. This sample will be useful both to train empirical algorithms (e.g., neural networks, nearest–neighbor), and to estimate the uncertainties in the mean values and errors of their distributions. We note that as photometric surveys are reaching fainter magnitude limits, the spectroscopic data available are less representative of the photometric sample. Therefore, new techniques are being developed to estimate photo-z in surveys without the need of spectroscopic data (see, e.g., Hoyle et al., 2017). In the Portal, we have implemented tools to estimate the redshift of sources using standard techniques based on spectroscopic redshifts to train and validate the algorithms.

In the following sections, we describe the methodology of each pipeline related to the photo-z estimation in the Portal. They operate in sequence, as illustrated in Figure 8. For each pipeline described, we add an example of a result obtained using data from DES Y1 release, as an illustration, and proof of concept. A description of the data used is available in Appendix A.

3.1. Spectroscopic Sample

In the Portal the first step to obtain photo-zs is the creation of a spectroscopic sample that will be matched with DES photometric catalog to define a training set used for training and validation. The goal is to create a sample with as many sources as possible avoiding effects of cosmic variance or under-representation of particular spectral types.

The database associated with the Portal serves as a centralized spectroscopic database for DES (Hoyle et al., 2017), being continually updated, in particular, by ongoing follow-up observations from DES collaborators such as the OzDES project (Yuan et al., 2015; Childress et al., 2017), as well as with substantial new spectroscopic galaxy samples made public.

The current spectroscopic redshift samples available are indicated in Table A.4. Once a new spectroscopic survey of interest is identified, the data is downloaded to our archive and ingested in the database to be accessible by the ***Spectroscopic Sample*** pipeline. At this stage of the process, it is necessary to provide some predefined information to be associated with a given spectroscopic sample in the registration database (Figure 9).

When building a spectroscopic sample from heterogeneous sources, we need to take into account the following:

- Each source might have different column names for the same quantities, e.g., redshift represented by "z", "spec-z", "zspec", etc. We associate mandatory columns (RA, Dec, redshift, redshift quality, source, redshift error) when uploading each sample to ensure that the columns are properly delivered as input to the pipelines (see Figure 10).

- Each catalog may have different quality estimates of redshift. To normalize quality flags to a single schema, we take the approach of OzDES Survey (Yuan et al., 2015), with qualities (Q_{spec}) ranging from 0 to 4. The numbers 0 and 1 are two types of unknown redshift, 2 is only a guess, 3 is above 90% confidence, and 4 attributed to a trusted redshift. When a survey is uploaded, we need to tell the centralized spectroscopic database that a new catalog has arrived and how to translate the quality information to the numerical system explained above.

- Elimination of duplicates. Internally, the pipeline handles possible multiple measurements for same objects.

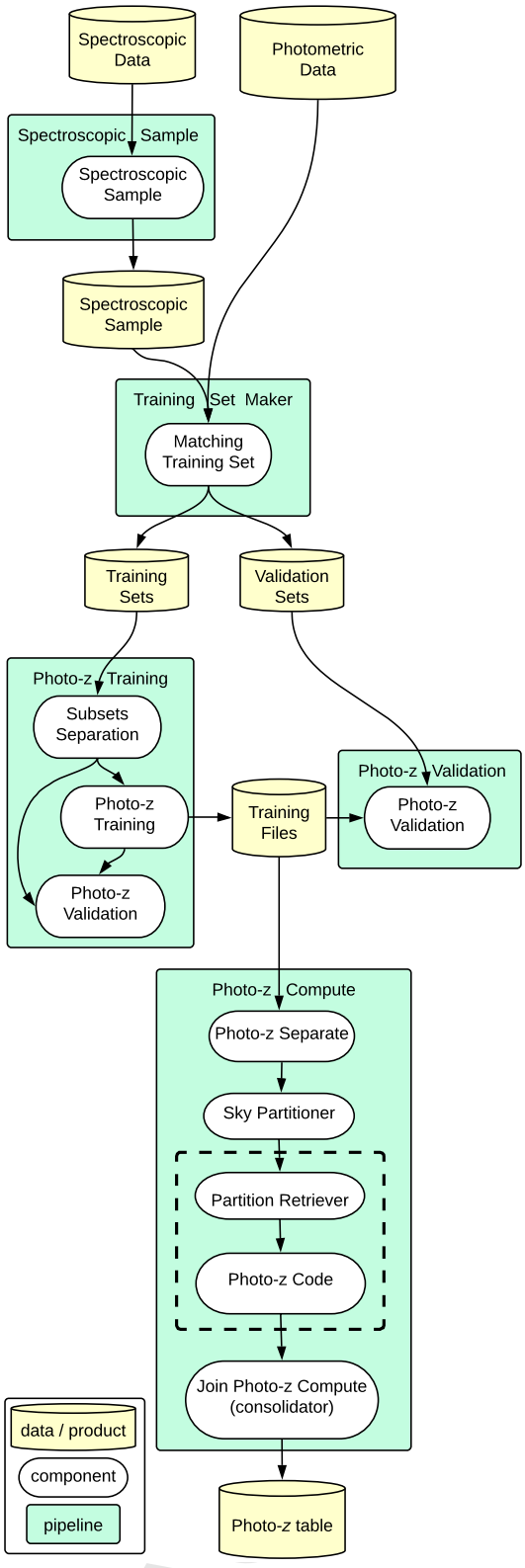


Figure 8: Photo-z complete workflow. Pipelines are represented in green: ***Spectroscopic Sample*** (Section 3.1), ***Training Set Maker*** (Section 3.2), ***Photo-z Training*** and ***Photo-z Validation*** (Section 3.3), and ***Photo-z Compute*** (Section 3.4. There is also another pipeline (***Photo-z PDF***, Section 3.5) not present in this illustration, because it is placed after the construction of scientific catalogs. The components are represented in white. Data inputs and products are represented by yellow cylinders. In ***Photo-z Compute***, the dashed line involves the part that runs in parallel. The generic "Photo-z Code" component block represents one of the components that wraps a particular photo-z code (DNF, LePhare, etc).

We implemented the following set of possibilities: 1) Select the spectroscopic redshift according to the best quality (default); 2) Make an average of all values to give a single redshift. To identify measurements for the same objects, the matching is done based on the angular separation between the coordinates in the different surveys using a matching radius of 1.0 arcsec (default). If two or more objects are within the search radius, the criteria to solve duplicates selected in the configuration tab applied is the following: we select the measurement that has the highest Q_{spec}. If more than one observation has the same Q_{spec} flag, we select the one obtained more recently. If there are two or more observations in the same year, we choose the redshift with the smallest error, when it is available. Finally, if we still have more than one source (from the same year and with no errors available), we choose the one with redshift the closest as possible to the mean value of all the multiple measurements. Since we have selected a high "quality" threshold, the differences between choosing the best source or averaging between all the different matches are negligible. This step is done in several sub–steps using a PostgreSQL extension for spatial indexing on a sphere, called Q3C (Koposov and Bartunov, 2006) and the Starlink Tables Infrastructure Library Tool Set (STILTS, Taylor, 2006).

- Spectral type classification. Some surveys classify the source (star, galaxy, QSO). When this is available, we use this information to allow for specific spectral classification of stars or galaxies. In the case where no classification provided, we assign as "stars" every object with z <0.001. For the time being, we do not classify objects as QSOs, and they will normally enter as galaxies. However, in the pipeline, ***Training Set Maker***, where we have photometric data associated to each object, we can apply a point source removal criterion, which eliminates most of the QSOs from the galaxy samples we built.

In summary, the ***Spectroscopic Sample*** pipeline creates a "master" sample by combining spectroscopic data from various surveys. The user chooses the ones to include in a check–box menu (see Figure 11) on the pipeline interface. The next step is to choose the configuration parameters, as can be seen in Figure 12. There one can make decisions about the criterion adopted to handle duplicates, quality threshold, spectral classification, etc.

The result of this pipeline is a table registered in the database containing: spec-zs, errors, quality flags, sources, and coordinates. At this stage, there is no association with DES objects. This table becomes available to be an input for the next pipeline or to be delivered to the collaboration.

Relevant information about the spectroscopic sample created in a particular run is in the process' product log. In Figure 13 we show the result of creating a spectroscopic sample containing data from all surveys available in the Portal, selecting only the best measurements (Q_{spec}=4), and resolving duplicates by the default criteria mentioned above. This plot shows the spatial distribution of spectroscopic redshifts included in the

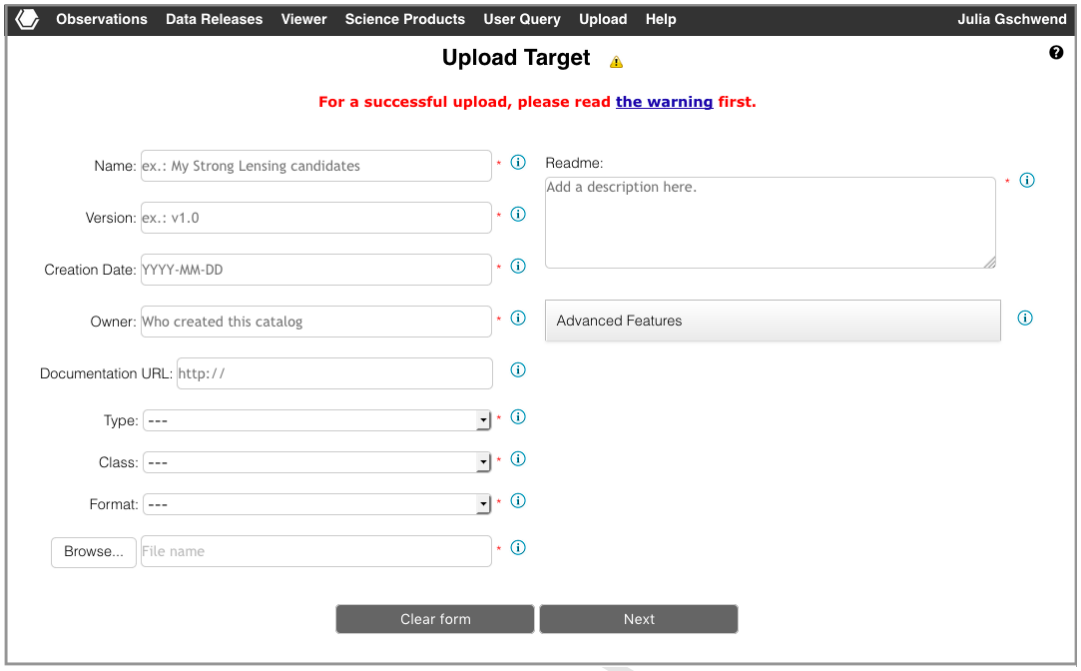


Figure 9: Upload tool initial screen. The user provides the relevant metadata to registering and versioning, such as the data source and a short description.

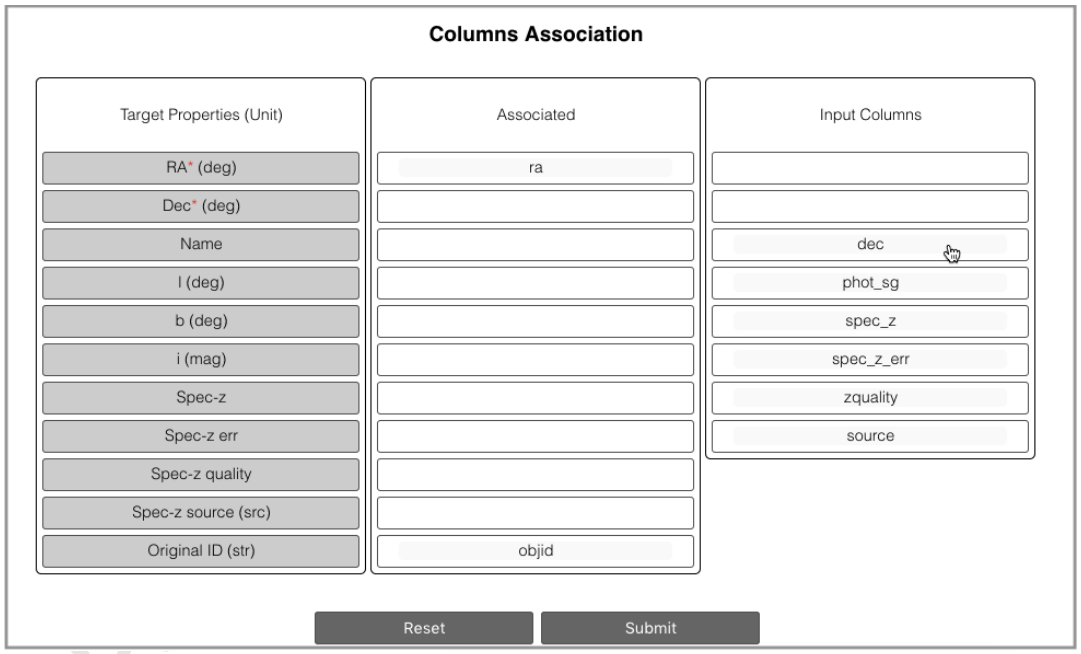


Figure 10: Upload tool - column association screen. In this screen, the user is asked to drag the original column names (on the right side) to those names expected by the pipelines (on the left). With this information, the ***Upload*** tool creates a table in the database with the contents of the uploaded catalog and the columns names appropriate to be read by the pipeline ***Spectroscopic Sample***. Both the original and the "translated" names are saved as metadata, to keep the history.

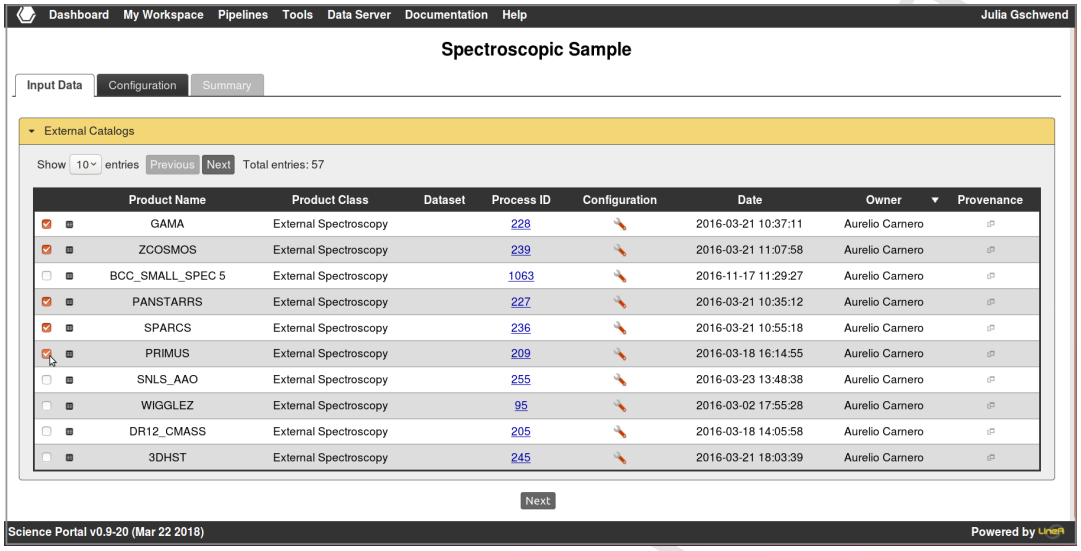


Figure 11: Input menu of ***Spectroscopic Sample*** pipeline. In this tab, the user selects which spectroscopic surveys are to be include in the spectroscopic sample.

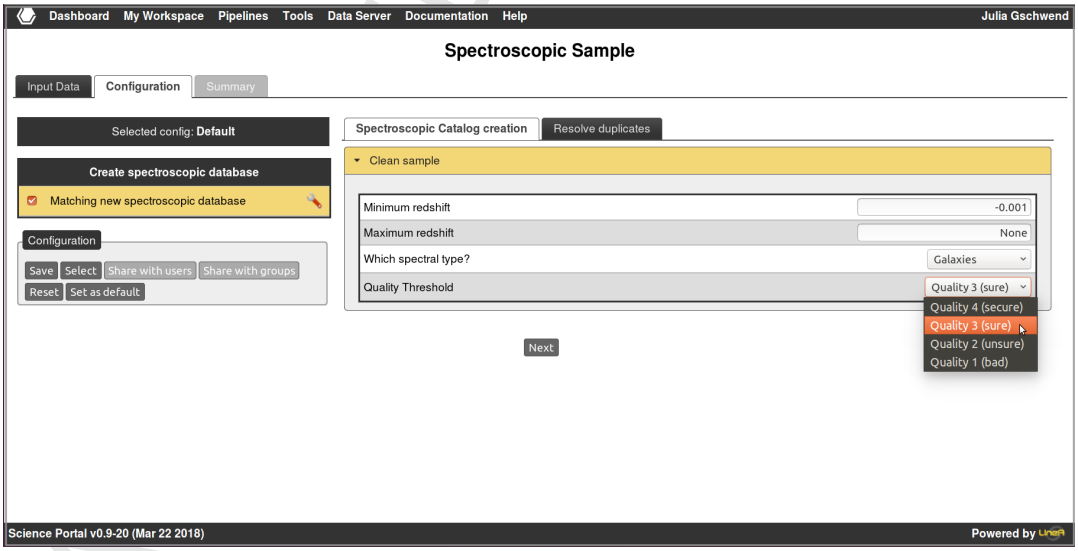


Figure 12: Configuration menu of ***Spectroscopic Sample*** pipeline. In this tab, the user selects spec-z and quality criterion to resolve duplicates.

sample, and we can verify that part of the spectroscopic sample spills over the DES footprint. The product created contains redshifts of 1,408,138 unique galaxies, from 34 surveys (see Appendix A.2).

Furthermore, the supplemental video V1[14] shows an example of a guided run of the ***Spectroscopic Sample*** pipeline and a quick exploration of its results.

3.2. Training Set Maker

After creation of the spectroscopic sample, the next step is to match the photometric data to the spectroscopic catalog. This is done by the ***Training Set Maker*** pipeline, designed to build training (and validation) samples by matching a photometric sample (among the datasets available in the Portal) with a spectroscopic sample, which comes from the ***Spectroscopic Sample*** pipeline, as shown in Figure 14.

In this stage the user can also include outputs of the ***Star/Galaxy Separation*** pipeline, identifying and removing point sources, avoiding the mismatch between spec-zs from galaxies merged to photometry from stars and removing QSOs. Also, optionally, it is possible to apply corrections in the observed magnitudes, like zero–point based on stellar locus regression calibrations (High et al., 2009) or galactic extinction.

In the configuration menu (Figure15), the user selects the parameters to make quality choices and a search radius used for matching. Similarly to the previous pipeline, the matching is also done based on the angular separation between the objects at the database level using Q3C, but here with the spectroscopic and photometric catalogs. We also selected the radius to 1.0 arcsec as a default configuration. If two or more objects from the photometric sample are within the search radius, the nearest object to the spectroscopic one is selected.

The result of this pipeline is a table registered in the database under the class "training_set", containing the columns from the spectroscopic sample plus some columns from photometric data (DES IDs, magnitudes, and errors). On the product log, one can access the query automatically generated by the pipeline (as illustrated in Figure 16) and pieces of information about the matched sample, so–called training set, organized in some tabs. In this example run, we selected the spectroscopic sample defined in the previous example (Figure 13) and the photometric sample from DES Y1 wide survey (details in Appendix A.1).

We present an example of this operation in the supplemental video V2[15] showing a live run of the pipeline ***Training Set Maker***, using the same inputs and configurations as shown above.

3.3. Photo-z Training and Validation

After the matched sample is created and registered in the portal's database, the next step before calculating photo-zs for the whole photometric dataset is to train the empirical algorithms and, optionally, calibrate the template–fitting ones. For several science applications, it is necessary to know the quality of the photo-z estimations requiring a validation step also done with a sample with known spectroscopic redshifts.

Although it is not the pipeline with the largest number of components, ***Photo-z Training*** is the most complex among pipelines related to photo-z, because it performs several different tasks, as detailed below. It is composed of three components, as illustrated in Figure 8. The first is `Subsets Separation`, which splits the matched sample into two subsets and performs a comprehensive characterization of them with plots and statistics. The second is the `Photoz Training`, which conducts the training procedure using the first subset, for several algorithms simultaneously. The third is the `Photoz Validation` which uses the second subset to compute the photo-zs and compare the results with the spec-zs, as well as it shows metrics and plots for quality assessment.

The primary input for the ***Photo-z Training*** pipeline is the matched sample built by the previous pipeline, named as "Training Set" on the input menu shown in Figure17. Also, we choose a photometric sample of reference ("Objects catalog", on the input menu), the same coadd tables mentioned in the previous pipeline. But in this case, this one is used only to compare photometric properties, to check whether the training and validations subsets are representative of the photometric set or not.

In the first implementation, we created a unique pipeline to do both steps, the training and validation (The pipeline ***Photo-z Training***). Thus, a first component is necessary to separate the data in training and validation subsamples (to avoid the biases introduced by validating with the same galaxies used for the training). Afterward, we created another pipeline to do the validation step separately (details below), but we kept the first option available, optionally.

The `Subsets Separation` component splits the matched catalog randomly, where the fraction of the data delivered to the training subset (and consequently the remaining portion for validation) is a free parameter in the component's configuration. Also, the user can define the sample selection criteria, choosing the acceptable intervals of magnitude, redshifts, colors and magnitude signal–to–noise ratio, as illustrated in Figure 18.

Ideally, the training and validation samples should have the same properties as the photometric sample of interest. However, this is difficult to meet when spectroscopic data come from surveys with different depths, redshift intervals, and targeting strategies. In cases like this, it is a common practice to evaluate the performance of the photo-z in a *weighted* sample, representing the color and magnitude distributions of the photometric sample.

In the Portal, it is optional to weight the training and validation sets, using the algorithm presented in Lima et al. (2008). If so, we assign to each galaxy its relative importance in representing the photometric sample, regarding the multi–space of colors and magnitudes. The user builds the *weighted* sample by repeating galaxies multiple times in the proportion of their weights, with their magnitudes spread according to their errors (assumed Gaussian) to avoid generating identical cloned galaxies. Applying this algorithm, we obtain a weighted sample that presents distributions of colors and magnitudes very similar to

[14] `https://youtu.be/1mu-PqOvK88?list=PLGFEWqwqBauBIYa8H6KnZ4d-5ytM59vG2`

[15] `https://youtu.be/2nA1PFGCnEM?list=PLGFEWqwqBauBIYa8H6KnZ4d-5ytM59vG2`

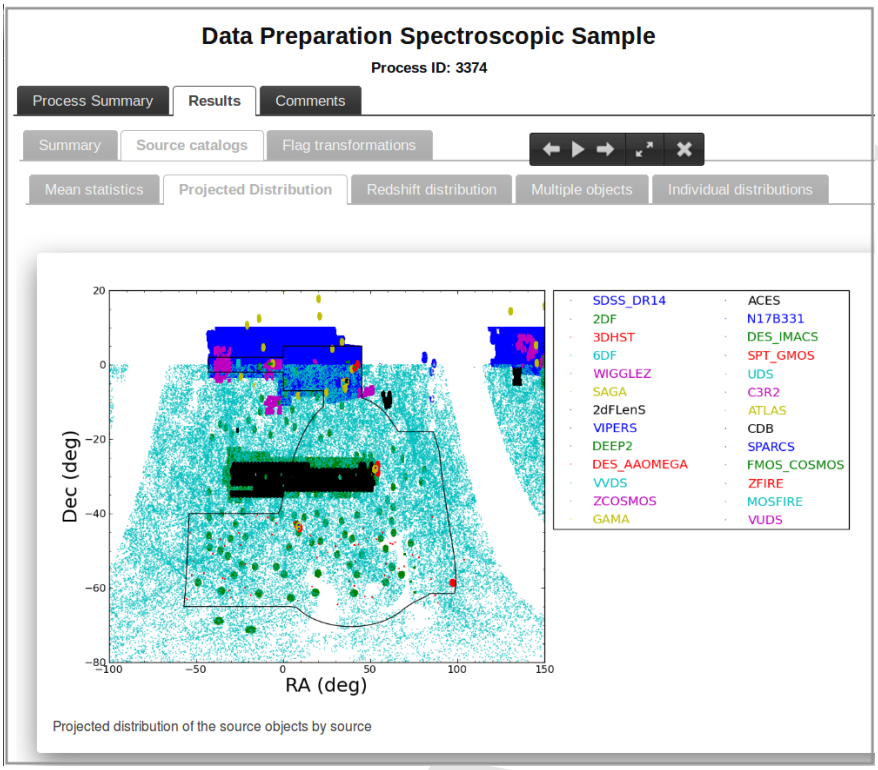


Figure 13: Product log of ***Spectroscopic Sample*** pipeline. Screenshot capturing one of the plots available on the "Results" tab of the product log. The black line represents the DES footprint. The surveys in the legend are ordered by the number of spectroscopic sources included in the sample after resolving the duplicates.

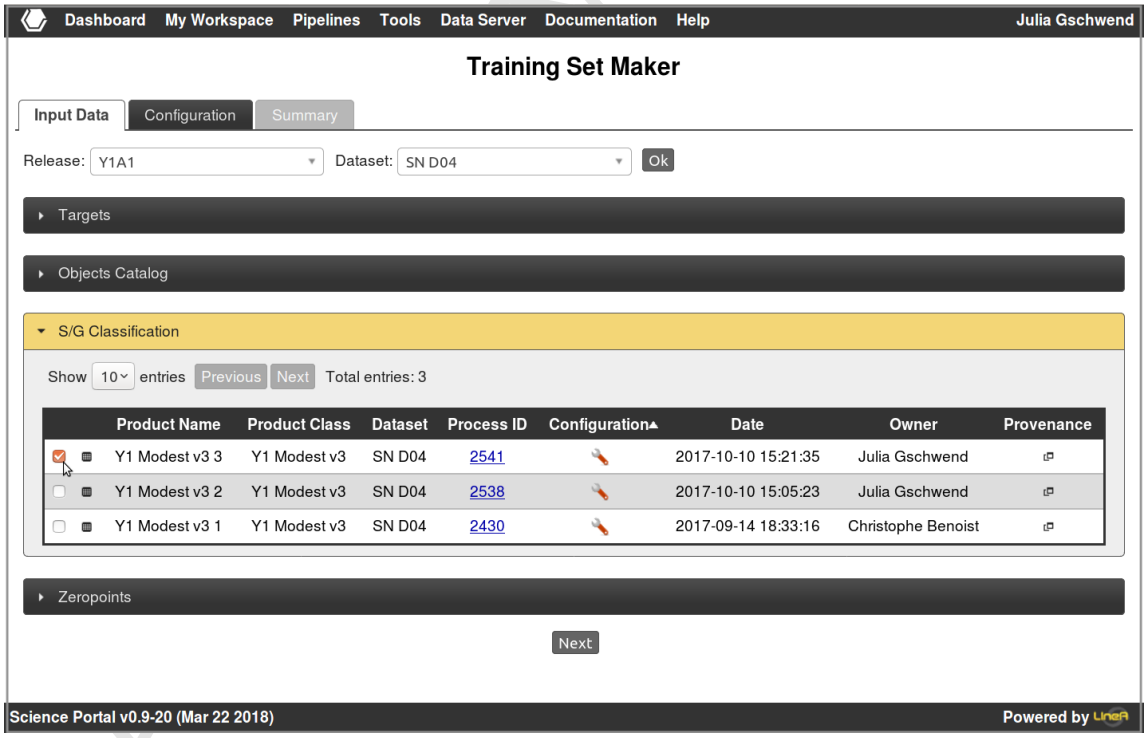


Figure 14: Input menu of the ***Training Set Maker*** pipeline. Each one of the four pull down menus show the available products for each class (mandatory or not) required by this pipeline: *Targets* – the spectroscopic sample defined in Section 3.3, *Object Catalog* – the coadd tables coming from DES, *S/G Separation* – a star/galaxy classification table, provided by another pipeline, not addressed in this paper, and *Zeropoints* – optional additional photometric calibrations. Each row with a check box refers to a product generated by a previous pipeline, registered in the database. The process number is a link that redirects to the process' product log, which helps the user on the choice.

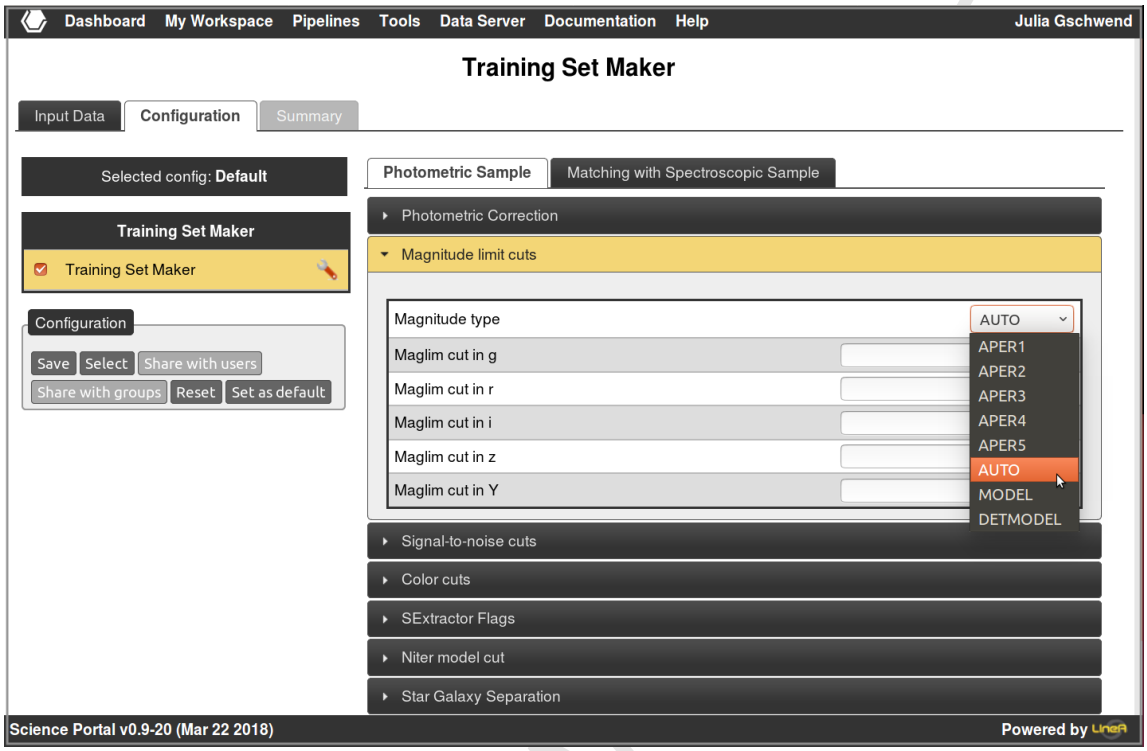


Figure 15: Configuration menu of the ***Training Set Maker*** pipeline. On this screen, the user is asked to make decisions about the characteristics of the matched sample, called training set, that is being constructed. In this example, the type of magnitude used to apply quality cuts.

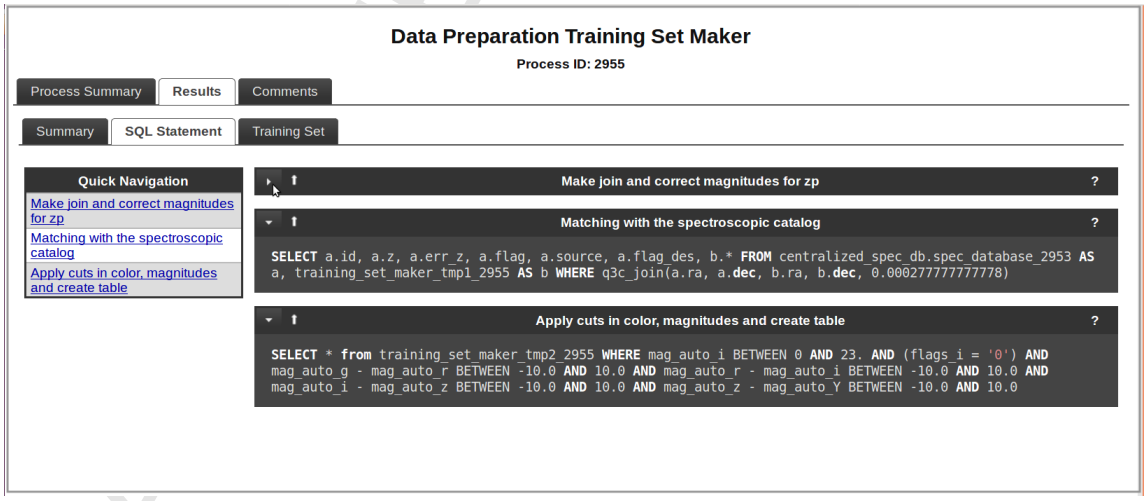


Figure 16: Product log of ***Training Set Maker*** pipeline. This screen summarizes the results of a pipeline run. The SQL query used by the pipeline is always shown. In the second tab, there are tables and plots to show the characteristics of the training set just created.

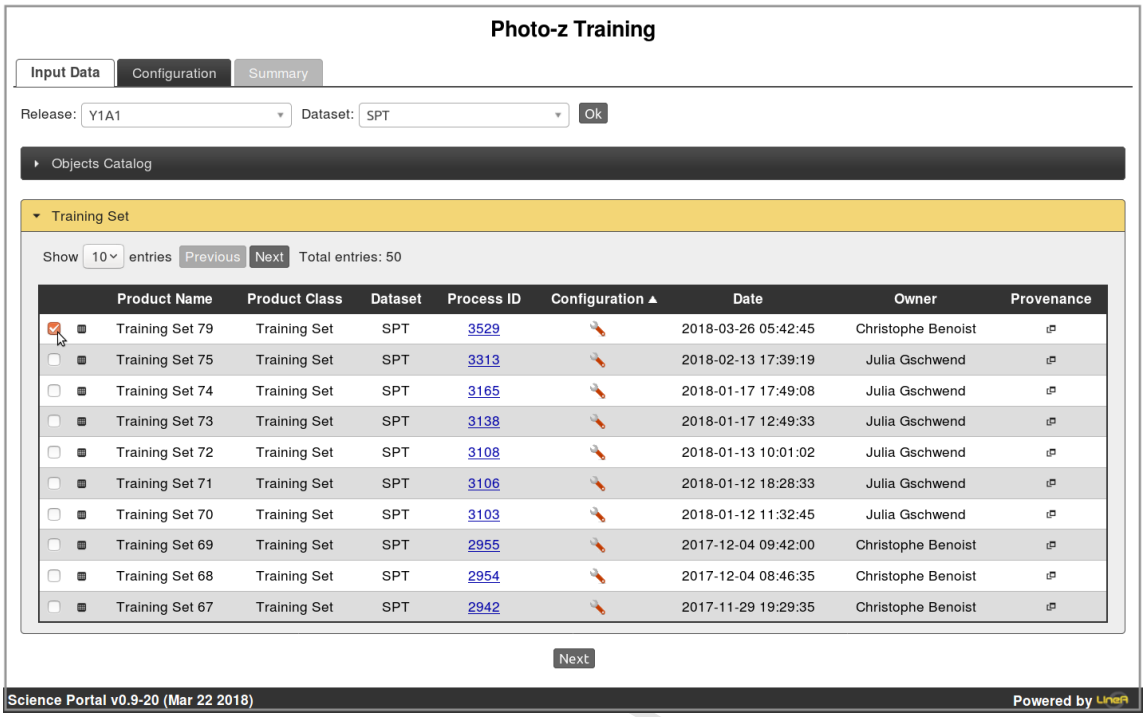


Figure 17: Input menu of the ***Photo-z Training*** pipeline. On this screen, the user is asked to chose one matched sample (recognized by its class "Training Set"), and a photometric sample of reference ("Objects Catalog"). Similarly to any other pipeline, each row with a check box refers to a product previously generated by another pipeline, and registered in the database. The process number is a link that redirects to the process' product log, which helps the users on their choice.

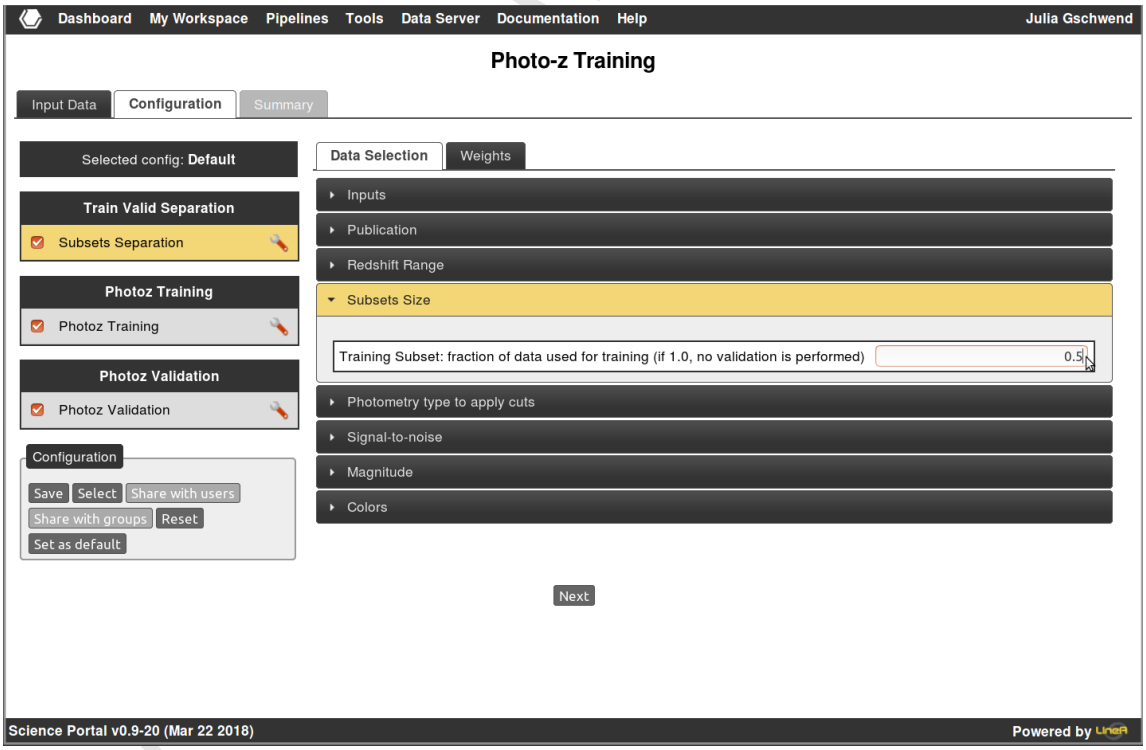


Figure 18: Screenshot of the ***Configuration*** tab of the component Subsets Separation, that belongs to the ***Photo-z Training*** pipeline. The default value for the fraction of data given to training and validation is 0.5 for each subset. If the fraction is chosen to be 1.0, the whole sample is employed for training, and the validation step is skipped.

those of the photometric sample, as shown in Figure 19. In the example, the excess of red objects in the training set is diminished as the result of weighting.

Photo-z Training

In recent years the number of photo-z algorithms has increased enormously. The Portal is an interesting environment to compare different methods, since they can be applied to datasets under similar conditions. So far, the following codes are implemented in the Portal: ANNZ (Collister and Lahav, 2004), ANNZ2 (Sadeh et al., 2016), ARBORZ (Gerdes et al., 2010), BPZ (Benítez, 2000), DNF (De Vicente et al., 2016), LEPHARE (Arnouts et al., 2002; Ilbert et al., 2006), POFZ (Cunha et al., 2009), SKYNET (Graff et al., 2014), and TPZ (Carrasco Kind and Brunner, 2013, 2014). We refer to Hildebrandt et al. (2010) and Carrasco Kind and Brunner (2014) for a review of the particularities and comparison of their performances.

Empirical methods are the basis for the majority of the algorithms, except for BPZ and LEPHARE, two template fitting codes, for which a training sample can be used to improve photo-z quality through systematic shifts in the theoretical magnitudes from the spectral energy distribution (SED) templates. Hence, all of them are implemented in ***Photo-z Training*** pipeline. Nevertheless, for the template–fitting ones, the "training" step is not mandatory.

Each photo-z algorithm has its configuration parameters. The user interface provides a configuration menu with a default configuration, but the user can change these values as shown in Figure 20.

The product of this training procedure is the so–called *training file*. Its format and content depend strongly on the photo-z algorithm used. For instance, TPZ's training files are stored in NumPy[16] format files, containing the decision trees used in photo-z estimation. LEPHARE's training files are just a list of floating point numbers representing the systematic shifts applied to the theoretical magnitudes (those obtained from the SED templates), stored in a simple text file. Besides the training files, the component `Photo-z Training` also registers the code configurations used, so it is also applied by the pipeline ***Photo-z Compute***, where the photo-zs are estimated for the DES datasets.

The main advantage of performing this step independent from the actual photo-z estimation is that training files from one training procedure can be used in the photo-z calculation several times, for different photometric datasets. On the other hand, one can make training and validation several times, until gets a satisfactory result and then apply it to the photo-z central estimate.

Photo-z Validation

The last component of ***Photo-z Training*** pipeline is the `Photo-z Validation`. It is responsible for checking the quality of the photo-zs computed in a validation sample, as an estimate of the quality of the photo-z to be estimated for the large photometric datasets.

To meet science–driven requirements, sometimes one needs to perform training and validation in samples which are independent of each other. Hence, we created a new pipeline (keeping the first one active) called ***Photo-z Validation*** to perform only the validation step, using the result of training from a previous run of the pipeline ***Photo-z Training***, but with the possibility to receive a completely different matched sample as input data. This pipeline uses the same component `Photo-z Validation` as the ***Photo-z Training*** pipeline, therefore the methodology is the same. The coincidence of pipeline and components' names might lead the reader to a confusion. We clarify the sequence of tasks performed by the components grouped by the pipelines in Figure 8.

In summary, there are two possible ways to validate photo-zs in the Portal: (i) splitting the matched spec–photo sample (so–called training set) into two subsets and perform the validation at the last component of ***Photo-z Training*** pipeline; (ii) training with 100% of the training set, and do the validation separately, in another pipeline, with an independent validation set. Both ways follow the same methodology. The only difference is the definition of the inputs.

The validation results consist of photo-z metrics (to quantify bias, dispersion, etc.) and quality assessment plots for visual inspection. The definition of the metrics used can be found in Sánchez et al. (2014). Uncertainties in the metric values are estimated using the Bootstrap re–sampling technique (Bradley and Tibshirani, 1993) based on 100 realizations, as done in such work. Some of these metrics have a limit of acceptance, defined by the collaboration as a scientific requirement for dark energy studies. So this component also works as a "vetting point" for the photo-z estimates. If the photo-z quality is considered unacceptable, the user should repeat the previous steps varying the data used and the configuration parameters.

An example of product log is presented in Figure 21, showing the results obtained using DNF. This figure shows how this pipeline can be used to compare performances of different algorithms like the ones done by Hildebrandt et al. (2010), and Sánchez et al. (2014).

The user can navigate through tabs to access the results from different codes. In particular, there is an additional tab where the results are consolidated and presented together to ease the comparison.

For more detailed navigation through the various configuration parameters and results reported on the product log, please watch an example of usage of the ***Photo-z Training*** pipeline in the supplemental video V3[17].

3.4. Photo-z Compute

The actual photo-z calculation in the Portal is done by the ***Photo-z Compute*** pipeline. It estimates photo-zs for DES objects present in the photometric catalogs, regardless of the object's nature (e.g., star or galaxy), using the training file(s) produced by ***Photo-z Training***. Once the photo-zs are calculated,

[16] `http://www.numpy.org/`

[17] `https://youtu.be/ZOJOhGWlvag?list=PLGFEWqwqBauBIYa8H6KnZ4d-5ytM59vG2`

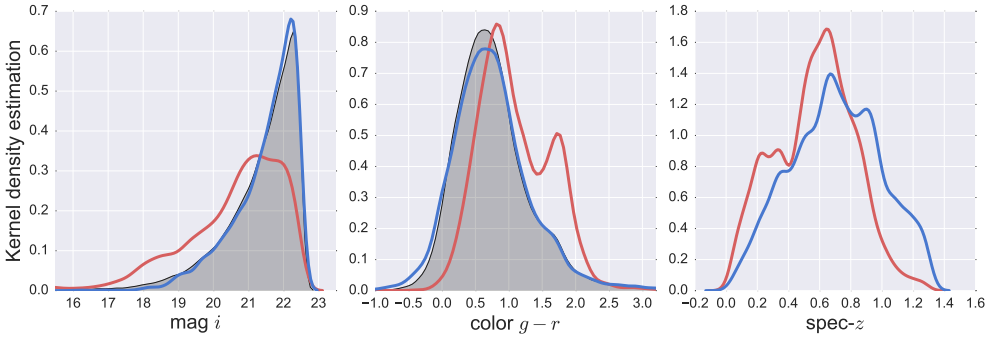


Figure 19: Magnitude (i–band), color (g-r), and spec-z distributions for the photometric sample (Y1A1, in gray), and for the training set, before (in red) and after (in blue) weighting. The validation set, not shown here, has the same properties as the training set.

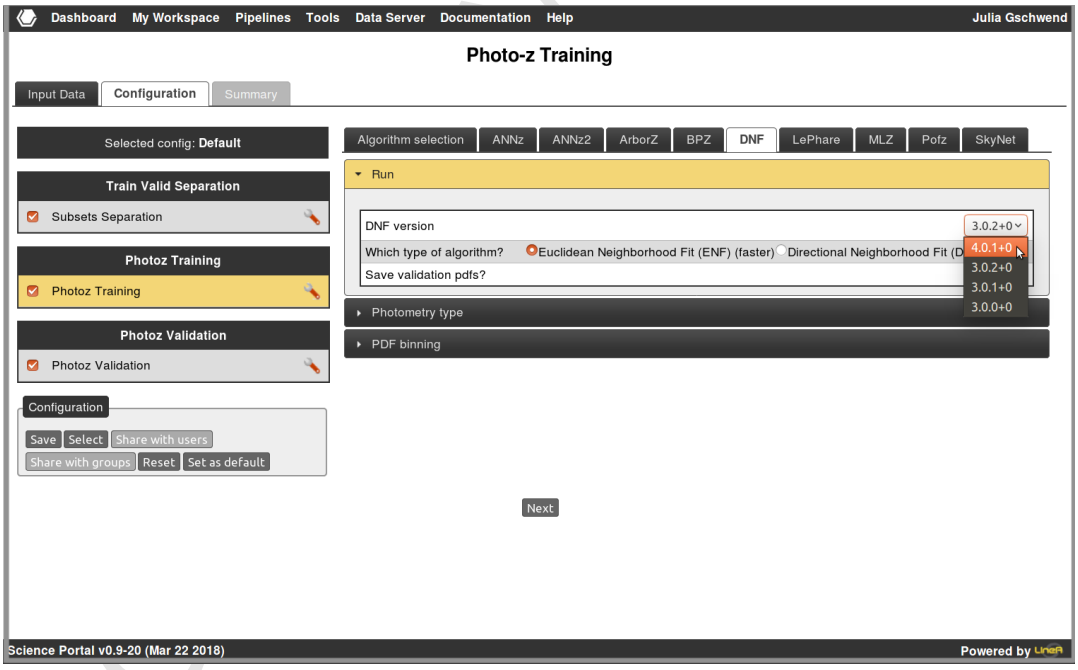


Figure 20: Screenshot of the configuration tab of the ***Photo-z Training*** pipeline. On the left side, each check box refers to one component of this pipeline. The small tool symbol beside leads to the menu on the right side of the page, where several tabs organize the configuration parameters for the different algorithms available. In this screenshot we show those from the code BPZ as an example.

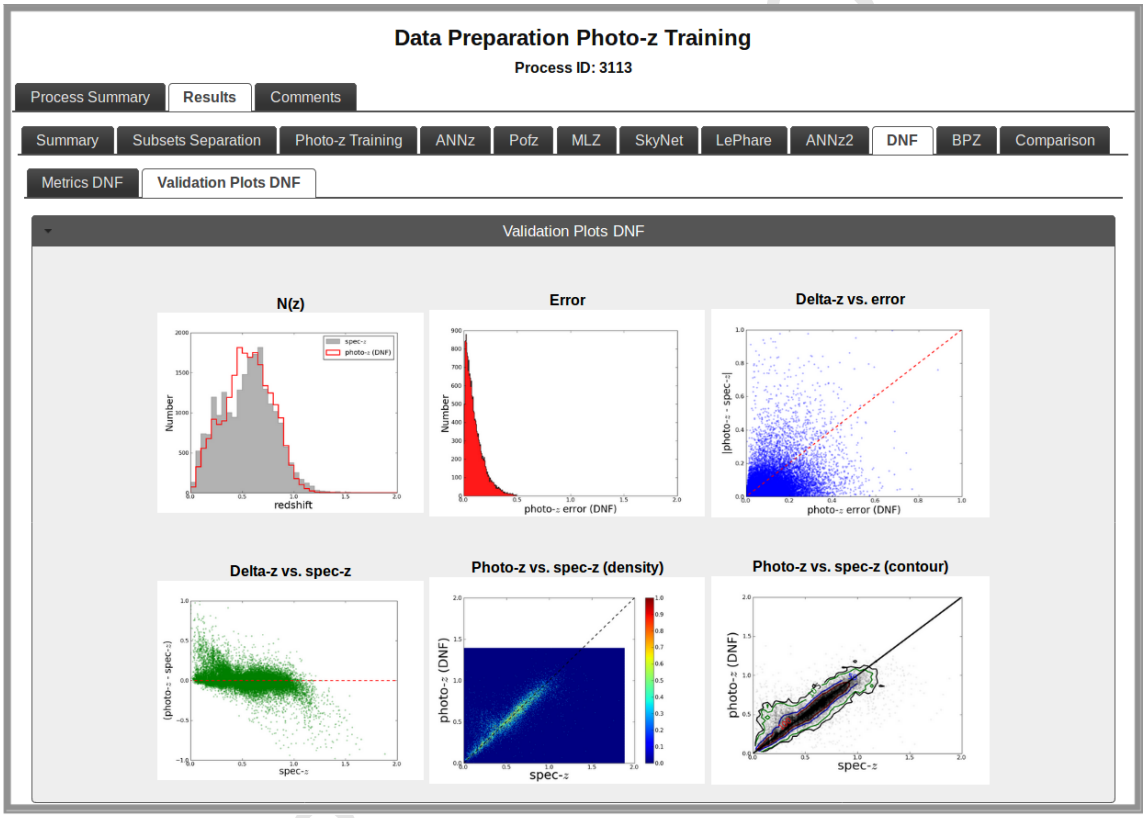


Figure 21: Results from the `Photo-z Validation` component, organized in tabs by photo-z algorithms. In this example, we present results by DNF. The three plots on the top are, from left to right, the histograms of redshift and error distributions, and the scatter plot of photo-z error versus the difference between photo-z and spec-z for each object in the validation sample. The plots on the bottom are the scatter plot of photo-z versus the difference between photo-z and spec-z, and the density and contour plots of photo-z versus spec-z.

they can be used in the creation of science–ready catalogs considering, e.g., different star/galaxy classifiers, color selections, and magnitude limits. It is also possible to download the photo-z resulting tables, or to deliver them to the collaboration through the export tool, connected to DESDM database. Since the same object can be tagged as galaxy by one classifier or star by another, it is essential to have photo-zs available for all objects. Therefore, at this stage, the distribution of redshifts, $N(z)$, obtained with the ***Photo-z Compute*** pipeline is not representative of the galaxy distribution yet. Only later, when the final catalogs are produced after pruning, the $N(z)$ can be used for scientific analyses.

The ***Photo-z Compute*** is composed of five components. The dashed line in Figure 8 highlights the part of the workflow that runs in parallel processing. The first component, `Photo-z Separate` handles with the training files inherited from the previous process. The second component, `Sky Partitioner`, defines the data partition by dividing the area in the Sky covered by the input dataset, based on the user choice of partition unit (as detailed below). It distributes the information about the data partitions into the nodes where the photo-z codes run in parallel.

The next two components (enclosed by the dashed line in Figure 8) run in parallel, each item of the partition defined by the previous one. The component `Partition Retriever` loads the contents of data partition in each computing node. To deal with large photometric samples efficiently, the data access uses HDFS, which previously distributed chunks of data in the cluster nodes, minimizing time spent with data transfer. Therefore, the `Partition Retriever` reads from the cluster nodes, as discussed in Section 2.

It is only in the fourth component that the photo-zs are in fact estimated. This step is the most computationally intensive of all the tasks related to photo-z estimation. There is one component available for each algorithm. All of them contain a python wrapper that prepares the input data, runs the code, standardizes the outputs, and delivers it to the consolidator. Thus, the workflow calls only the one that corresponds to code chosen in the configuration screen (see the menu on the left side displayed in Figure 22).

The consolidated jobs and their resulting photo-z table are ingested in the database by the last component, called `Join Photo-z Compute`. The product of ***Photo-z Compute*** is one of the leading ingredients to compose a science–ready catalog for extragalactic sciences, as discussed in Fausti Neto et al. (2018).

Since the Portal provides flexibility on the parallelization strategy, we performed a series of tests to illustrate the use of our computer cluster. In the following paragraphs, we compare the computing of photo-zs by varying the size of the data chunk, and consequently, the number of data partitions. For simplicity, we use the original data division from DES, based on tiles (see discussion ahead). The methodology details of the parallelization are discussed in Section 2.

We use the most extensive dataset of Y1A1 data release (SPT, details in Appendix A.1) to make stress tests and test the cluster capacity. In this case, we choose to use the algorithm DNF, which is one of the fastest codes available in the Portal, according to previous tests not addressed in this work.

The most straightforward way to vary the parallelization strategy is to modify the number of tiles to be processed by each job submitted to the cluster management system. Table 2 and Figure 23 summarize the results of the tests. In this table, the first column shows the number of tiles processed per job in parallel. The second column shows the total number of jobs, which is approximately the total number of tiles in SPT (3,373) divided by the first column. The third column shows the time spent in the serial parts of the processes (organizing training files, defining the partitions, and concatenating the results at the end). The fourth column shows the parallelized part of the processes (the data retrieving and the actual photo-z estimation, the components surrounded by dashed line in Figure 8). The fifth column shows the total duration of each process.

All processes started with the same inputs and code configuration and delivered the same results. The only difference was in the definition of data partitions, which was seen to have a significant impact on the process duration. Hereafter, we refer to "infrastructure time" as the difference between the total time of a process and the time spent on actual code running. This quantity is difficult to measure when running in parallel. It is often the case where a group of jobs is submitted to the computer cluster, virtually simultaneously, but they do not finish at the same time, even though all the nodes have the same hardware characteristics and the sizes of the data chunks are virtually homogeneous.

Some of the possible reasons for the different delays are: i) reading data from the same node versus reading data from a neighbor node; ii) long queues of data partitions waiting for their jobs to start; iii) bottleneck for writing in the 'reduce' part of the workflow (where some jobs still waiting in a queue to register the results, when others are already writing).

Another contribution to the infrastructure time might be the time spent by HTCondor to manage the jobs (start, finish, writing logs, creating temporary directories, and distributing jobs in the cluster). Although we can raise several possible reasons for the time lost in processes and the differences in time delays between processes with different partition sizes, we can not measure precisely relative contribution from each one of these sources of delay. Therefore, the infrastructure time is the cumulative time loss due to a combination of reasons.

We recall that there is an option to reserve the entire node (24 cores) for a single job, when the process deals with internal parallelization, as mentioned in Section 2. This is not the case here. For this test, this option of node reservation is disabled, so the jobs are distributed all over the cluster, regardless of the nodes to which the cores belong. Therefore, the maximum number of jobs running simultaneously is 912 (38 nodes times 24 cores). This number possibly explains why $N = 4$ is the most effective strategy for this first test. If the infrastructure time was null and the execution time of the primary algorithm was proportional to the size of the data chunk then it would be virtually equivalent to run 1 or 4 tiles per core, i.e., 1 round of 843 jobs running four tiles each, or 4 rounds of 912 (actually 3 rounds of 912 plus one of 637) jobs running one tile each. Since the infrastructure time is not zero, and it is cumulative, four tiles per node are better than 1, because it occupies almost

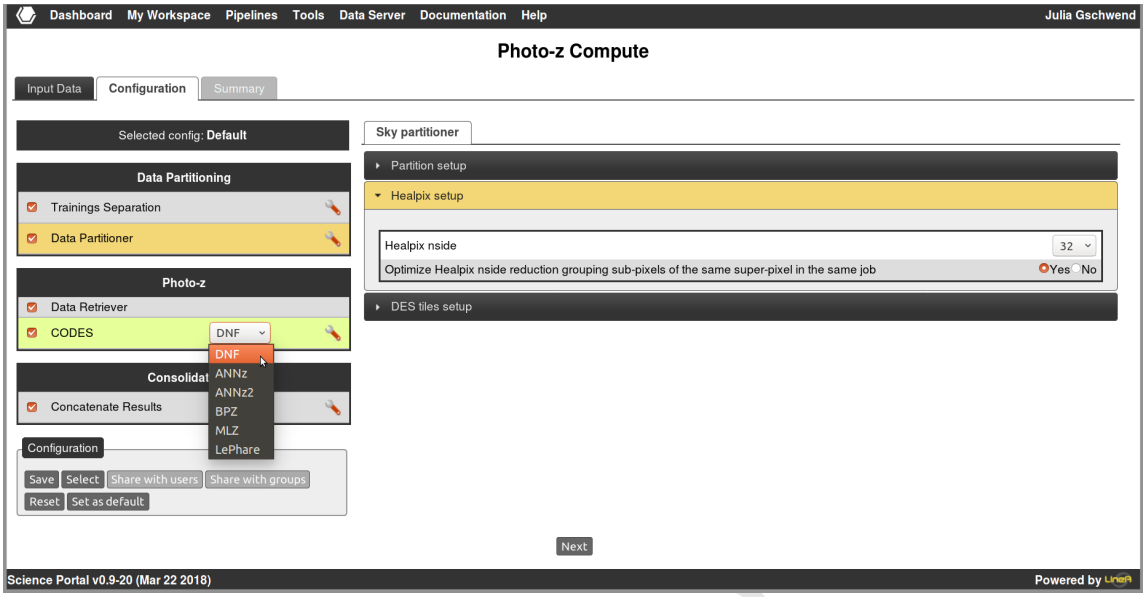


Figure 22: Screenshot of the configuration tab of the ***Photo-z Compute*** pipeline. On the left, the five components mentioned in the text. For this example, we chose DNF as the photo-z algorithm to be applied by the fourth component. On the right, the choice of the size of each data partition, based on HEALPix pixels.

Table 2: Execution time of pipeline ***Photo-z Compute*** - dataset SPT, data partition based on DES tiles.

# Tiles/job	# Jobs	Serial†	Parallel†	Total†
1	3,373	00:49	02:30	03:19
4	843	00:47	01:19	02:06
12	282	00:48	01:31	02:19
24	141	00:48	02:11	02:59
32	106	00:48	02:27	03:15

† Duration in (hh:mm) format.

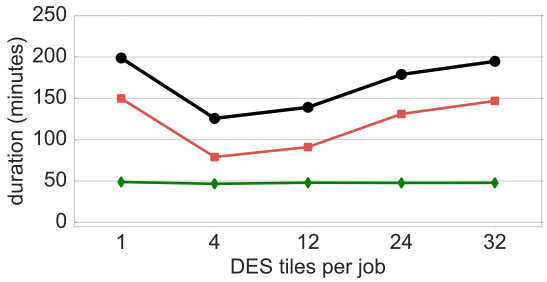


Figure 23: Duration of pipeline ***Photo-z Compute*** execution as a function of the number of DES tiles processed by each job in parallel. In black circles, the total duration of the runs. In green diamonds, the sum of the durations of all serial components. In red squares, the time spent with the parallelized part of the process, i.e., the actual photo-z estimations by DNF (see Figure 8).

the whole cluster with jobs, without leaving any other job waiting on a queue. The small queue of four files to be processed within a job seems to work more efficiently than a large queue of individual files distributed to the cluster.

As expected, for $N > 4$, the larger the N, the longer the processing time. This is true because large N reduces the number of CPU cores used, wasting the capacity of the cluster. In summary, according to our tests, the optimal number of tiles distributed per job depends on the dataset size as the following: If the number of tiles of a dataset is less than 912 (maximum number of jobs running in parallel) than the optimal choice is to distribute one tile per job. If the total number of tiles is larger than 912, the optimal number of tiles per job is the integer number closest to the result of the total number of tiles divided by 912. In the example shown in Figure 22, the total number of tiles in the SPT dataset, 3373, divided by 912 is ~3.7. Therefore, the fastest run using this particular dataset was the one with four tiles per job, as observed in Figure 23 and Table 2. Interestingly, the data partitioner and the consolidator performances are very stable, independently of the number of partitions they handle.

As done for the previous pipelines, we show an example of running ***Photo-z Compute*** in the supplemental video V4[18].

3.5. *Photo-z PDF*

The use of redshift probability density functions (PDFs), instead of single estimates of photometric redshifts, is a necessary approach to incorporate the measurement's uncertainties on scientific analyzes. For large astronomical surveys, the storage of billions of PDFs can be a challenge, furthermore if they are measured several times, as when using different methods.

To overcome these issues, we adopt two procedures when dealing with PDFs in the Portal. The first one is to compute

[18] `https://youtu.be/IcCk0MYhy-E?list=PLGFEWqwqBauBIYa8H6KnZ4d-5ytM59vG2`

PDFs only for objects selected in a science–ready catalog, avoiding wasting time and storage space with PDFs for stars or bad data. That is the reason for the pipeline not to be present in Figure 8. The second procedure is to apply a method for data compression and store just a reduced list of coefficients representing the PDF, instead of the complete PDF for each galaxy. The details of this method are explained in Rau et al. (2018, in preparation).

In a similar way to the previous pipelines, we show examples of configurations, such as the photo-z code to be used, and the redshift range for the PDF in Figure 24.

The resulting redshift distribution strongly depends on the stage it is obtained in the Portal, as evident in Figure 25, where we show, as an example, for a small sample (dataset COSMOS D04, defined in Appendix A.1) obtained using DNF. The pipeline ***Photo-z Compute*** provides photo-zs for every object present in the "raw" photometric samples, including stars and poorly sampled objects (left panel). The pipeline responsible for creating science–ready catalogs removes those objects, but its product still contains only point photo-z estimates (middle panel), which can be severely biased. To obtain the final distribution, we do the stacking of the probabilities (right panel) which is considerably smoother than the previous one.

4. Summary

In this paper, we describe the infrastructure available in the DES Science Portal to create training sets, training files and to compute photo-z using different algorithms. It is an easy–to–use framework that concatenates different pipelines involved in the calculation of photo-zs, ensuring consistency between these processes.

The database registers all the steps; the Portal framework eases the task of carrying out a large variety of tests and comparing their results. Considering the volume of data, the number of algorithms and the various releases of photometric and spectroscopic data, having a structured framework like the one presented here is critical for vetting of DES algorithmic improvements, and the systematic production of photo-zs for future DES releases.

Although the Portal is currently accessible only for the members of DES collaboration, the methodology and lessons learned here can be useful and subject of interest for anyone that uses photo-zs in a wide range of science applications.

The database associated with the Portal ingests spectroscopic data regularly. Although the redshift repository continuously grows, the list of surveys used is reported and registered, so that a process can be reproduced or use only the catalogs of interest in another experiment.

After preparing training sets and photo-zs we can compare to spec-zs for quality checks. The pipeline used in the estimation of photo-zs for large datasets is parallelized to improve performance. The tests presented in Section 3.4 reveal that a good parallelization strategy is to distribute the data to the CPU cores using data partitions that are small enough to occupy the whole computing cluster, but large enough to avoid creating a queue of idle jobs. The optimal number of tiles or HEALPix pixels to be processed per job is then dependent on the size of the dataset in question and its original data partition. The resulting photo-z tables are amongst the values added in the preparation of catalogs ready for Portal science workflows.

It is important to point out that the strategy adopted by the Portal is to compute photo-zs for all objects in the original catalog produced by DESDM. We do that because the photo-z calculation is, by far, the most computationally intensive step of the E2E process. Calculating photo-zs for all objects gives the flexibility to create any catalog for Portal science workflows without having to re–compute photo-zs if one decides to change the star/galaxy classifier or another criterion for the sample selection. One disadvantage of our approach is that, in this first pass, we only compute point–values of photo-z.

The calculation of a full PDF happens at a later stage when the number of objects of interest is smaller, after quality pruning and star–galaxy separation. This approach is discussed in a separate paper that focuses on the method of preparing catalogs ready for Portal science workflows (Fausti Neto et al., 2018).

For the near future, there will also be pipelines available to be executed through Jupyter Notebooks (Kluyver et al., 2016; Perez and Granger, 2007), as an alternative to the regular workflow system. There is already a prototype that has been tested using the multi-user web application JupyterHub[19], but the current implementations are not related to photo-zs.

All the examples shown in the figures and supplemental videos use data from the Y1A1 data release. Nevertheless, the same infrastructure is valid for any other DES data release and also for simulations.

Besides already allowing one to handle large datasets and easing a lot of scientific applications, the DES Science Portal has been a useful laboratory of methodologies and a precursor of implementations for the next generation of photometric surveys.

Acknowledgments

We thank P. Egeland and F. Ostrovski for the contribution in the early phases of development of this infrastructure. We also thank R. Brito, J.G.S. Dias, V. Machado, L. Nunes, and G. Vila Verde for the contributions in the Portal's basic infrastructure, essential for the realization of this work.

JG is supported by CAPES. ACR is supported by CNPq process 157684/2015-6. ML is partially supported by CNPq and FAPESP. MA is supported by CNPq process 165049/2017-0. Part of this research is supported by INCT do e–Universo (CNPq grants 465376/2014-2).

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of

[19] https://github.com/jupyterhub

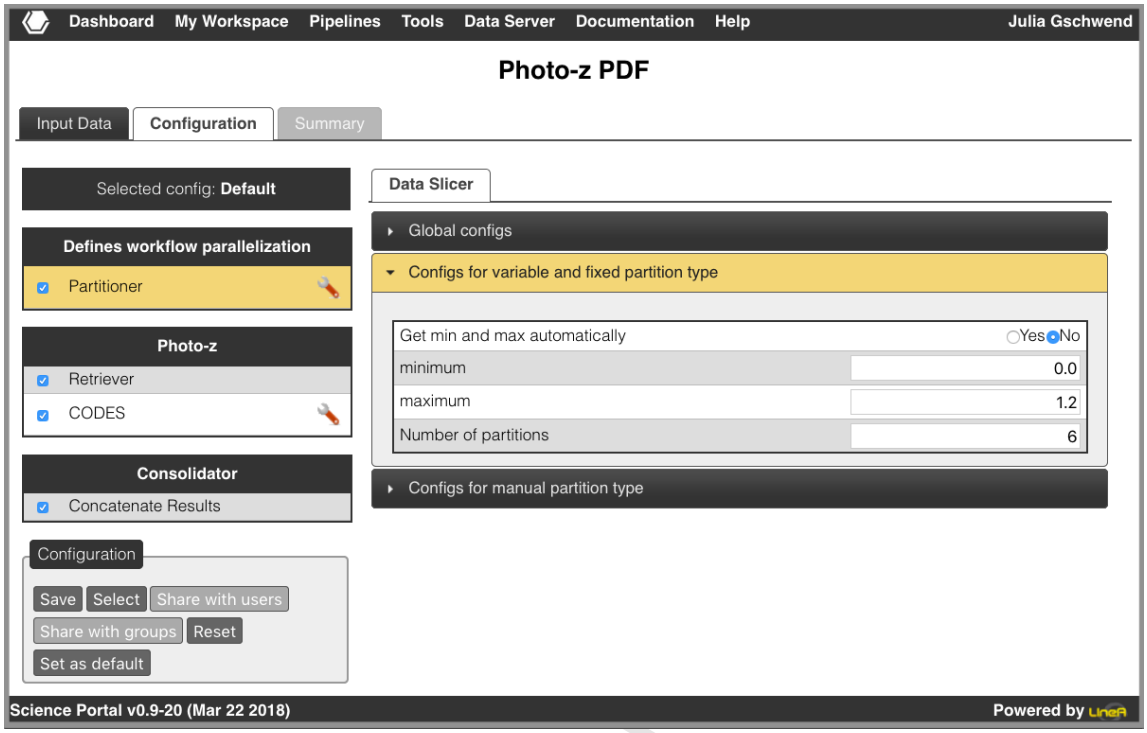


Figure 24: Screenshot of the configuration tab of the ***Photo-z PDF*** pipeline.

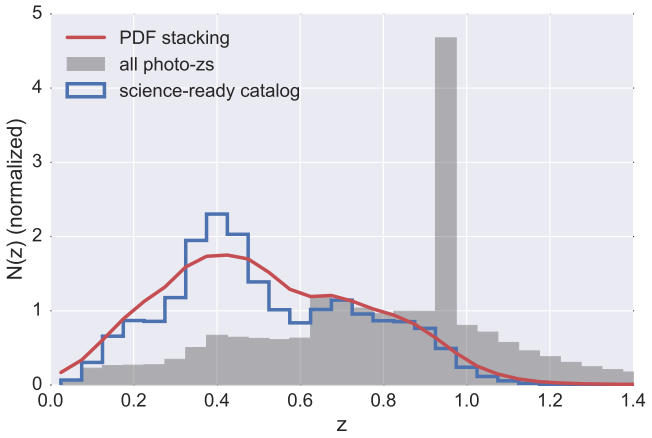


Figure 25: Photo-z distributions of the photometric sample. The stepfilled gray histogram shows the distribution of redshift point estimates for the complete sample (before cleaning from bad data and removing stars). The blue line histogram, still, the point estimates, but after selecting a science–sample. The red line shows the same science sample as the blue line, but considering the probability density functions, $P(z)$, instead of the single estimates.

Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, the National Optical Astronomy Observatory, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, which

is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MINECO under grants AYA2015-71825, ESP2015-66861, FPA2015-68048, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Australian Research Council Centre of Excellence for All-sky Astrophysics (CAASTRO), through project number CE110001020, and the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) e-Universe (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

Abbott, T., Abdalla, F.B., Allam, S., Dark Energy Survey Collaboration, 2016. Cosmology from cosmic shear with Dark Energy Survey Science Verification data. Phys. Rev. D 94, 022001. doi:10.1103/PhysRevD.94.022001.

Abbott, T.M.C., Abdalla, F.B., Allam, S., et al., 2018. The Dark Energy Survey Data Release 1. ArXiv e-prints arXiv:1801.03181.

Abolfathi, B., Aguado, D.S., Aguilar, G., et al., 2017. The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the extended Baryon Oscillation Spectroscopic Survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment. ArXiv e-prints arXiv:1707.09322.

Arnouts, S., Moscardini, L., Vanzella, E., et al., 2002. Measuring the redshift evolution of clustering: the Hubble Deep Field South. MNRAS 329, 355–366. doi:10.1046/j.1365-8711.2002.04988.x, arXiv:astro-ph/0109453.

Banerji, M., Jouvel, S., Lin, H., et al., 2015. Combining Dark Energy Survey Science Verification data with near-infrared data from the ESO VISTA Hemisphere Survey. MNRAS 446, 2523–2539. doi:10.1093/mnras/stu2261, arXiv:1407.3801.

Bayliss, M.B., Ruel, J., Stubbs, C.W., et al., 2016. SPT-GMOS: A Gemini/GMOS-South Spectroscopic Survey of Galaxy Clusters in the SPT-SZ Survey. ApJS 227, 3. doi:10.3847/0067-0049/227/1/3, arXiv:1609.05211.

Bazin, G., Ruhlmann-Kleider, V., Palanque-Delabrouille, N., et al., 2011. Photometric selection of Type Ia supernovae in the Supernova Legacy Survey. A&A 534, A43. doi:10.1051/0004-6361/201116898, arXiv:1109.0948.

Benítez, N., 2000. Bayesian Photometric Redshift Estimation. ApJ 536, 571–583. doi:10.1086/308947, arXiv:astro-ph/9811189.

Blake, C., Amon, A., Childress, M., et al., 2016. The 2-degree Field Lensing Survey: design and clustering measurements. MNRAS 462, 4240–4265. doi:10.1093/mnras/stw1990, arXiv:1608.02668.

Bonnett, C., Troxel, M.A., Hartley, W., others, Dark Energy Survey Collaboration, 2016. Redshift distributions of galaxies in the Dark Energy Survey Science Verification shear catalogue and implications for weak lensing. Phys. Rev. D 94, 042005. doi:10.1103/PhysRevD.94.042005, arXiv:1507.05909.

Bradley, E., Tibshirani, J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F., 2008. Extensible markup language (xml) 1.0.

Brescia, M., Cavuoti, S., Longo, G., et al., 2014. DAMEWARE: A Web Cyberinfrastructure for Astrophysical Data Mining. PASP 126, 783. doi:10.1086/677725, arXiv:1406.3538.

Burbeck, S., 1987. Applications programming in smalltalk-80(tm): How to use model-view-controller (mvc). URL: http://st-www.cs.uiuc.edu/users/smarch/st-docs/mvc.html.

Carlstrom, J.E., Ade, P.A.R., Aird, K.A., et al., 2011. The 10 Meter South Pole Telescope. PASP 123, 568–581. doi:10.1086/659879, arXiv:0907.4445.

Carrasco Kind, M., Brunner, R., 2014. MLZ: Machine Learning for photo-Z. Astrophysics Source Code Library. arXiv:1403.003.

Carrasco Kind, M., Brunner, R.J., 2013. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. MNRAS 432, 1483–1501. doi:10.1093/mnras/stt574, arXiv:1303.7269.

Cavuoti, S., Brescia, M., De Stefano, V., Longo, G., 2015. Photometric redshift estimation based on data mining with PhotoRApToR. Experimental Astronomy 39, 45–71. doi:10.1007/s10686-015-9443-4, arXiv:1501.06506.

Childress, M.J., Lidman, C., Davis, T.M., et al., 2017. OzDES multifibre spectroscopy for the Dark Energy Survey: 3-yr results and first data release. MNRAS 472, 273–288. doi:10.1093/mnras/stx1872, arXiv:1708.04526.

Coil, A.L., Blanton, M.R., Burles, S.M., et al., 2011. The PRIsm MUlti-object Survey (PRIMUS). I. Survey Overview and Characteristics. ApJ 741, 8. doi:10.1088/0004-637X/741/1/8, arXiv:1011.4307.

Colless, M., Dalton, G., Maddox, S., et al., 2001. The 2dF Galaxy Redshift Survey: spectra and redshifts. MNRAS 328, 1039–1063. doi:10.1046/j.1365-8711.2001.04902.x, arXiv:astro-ph/0106498.

Collister, A.A., Lahav, O., 2004. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. PASP 116, 345–351. doi:10.1086/383254, arXiv:astro-ph/0311058.

Cool, R.J., Moustakas, J., Blanton, M.R., et al., 2013. The PRIsm MUlti-object Survey (PRIMUS). II. Data Reduction and Redshift Fitting. ApJ 767, 118. doi:10.1088/0004-637X/767/2/118, arXiv:1303.2672.

Cooper, M.C., Yan, R., Dickinson, M., et al., 2012. The Arizona CDFS Environment Survey (ACES): A Magellan/IMACS Spectroscopic Survey of the Chandra Deep Field-South. MNRAS 425, 2116–2127. doi:10.1111/j.1365-2966.2012.21524.x, arXiv:1112.0312.

Cunha, C.E., Huterer, D., Lin, H., et al., 2014. Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. MNRAS 444, 129–146. doi:10.1093/mnras/stu1424, arXiv:1207.3347.

Cunha, C.E., Lima, M., Oyaizu, H., et al., 2009. Estimating the redshift distribution of photometric galaxy samples - II. Applications and tests of a new method. MNRAS 396, 2379–2398. doi:10.1111/j.1365-2966.2009.14908.x, arXiv:0810.2991.

Davis, C., Gatti, M., Vielzeuf, P., et al., 2017. Dark Energy Survey Year 1 Results: Cross-Correlation Redshifts in the DES – Calibration of the Weak Lensing Source Redshift Distributions. ArXiv e-prints arXiv:1710.02517.

Davis, M., Faber, S.M., Newman, J., et al., 2003. Science Objectives and Early Results of the DEEP2 Redshift Survey, in: Guhathakurta, P. (Ed.), Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II, pp. 161–172. doi:10.1117/12.457897, arXiv:astro-ph/0209419.

Davis, M., Guhathakurta, P., Konidaris, N.P., et al., 2007. The All-Wavelength Extended Groth Strip International Survey (AEGIS) Data Sets. ApJ 660, L1–L6. doi:10.1086/517931, arXiv:astro-ph/0607355.

De Vicente, J., Sánchez, E., Sevilla-Noarbe, I., 2016. DNF - Galaxy photometric redshift by Directional Neighbourhood Fitting. MNRAS 459, 3078–3088. doi:10.1093/mnras/stw857, arXiv:1511.07623.

Dean, J., Ghemawat, S., 2004. Mapreduce: Simplified data processing on large clusters, in: Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, USENIX Association,

Berkeley, CA, USA. pp. 10–10. URL: http://dl.acm.org/citation.cfm?id=1251254.1251264.

DES, et al., 2016. The Dark Energy Survey: more than dark energy - an overview. MNRAS 460, 1270–1299. doi:10.1093/mnras/stw641, arXiv:1601.00329.

Desai, S., Armstrong, R., Mohr, J.J., et al., 2012. The Blanco Cosmology Survey: Data Acquisition, Processing, Calibration, Quality Diagnostics, and Data Release. ApJ 757, 83. doi:10.1088/0004-637X/757/1/83, arXiv:1204.1210.

Diehl, H.T., Abbott, T.M.C., Annis, J., et al., 2014. The Dark Energy Survey and operations: Year 1, in: Observatory Operations: Strategies, Processes, and Systems V, p. 91490V. doi:10.1117/12.2056982.

Donovan, S., Huizenga, G., Hutton, A.J., Ross, C.C., Petersen, M.K., Schwan, P., 2003. Lustre: Building a file system for 1000-node clusters, in: Proceedings of the Linux Symposium.

Driver, S.P., Hill, D.T., Kelvin, L.S., et al., 2011. Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release. MNRAS 413, 971–995. doi:10.1111/j.1365-2966.2010.18188.x, arXiv:1009.0614.

Drlica-Wagner, A., Sevilla-Noarbe, I., Rykoff, E.S., et al., 2018. Dark Energy Survey Year 1 Results: The Photometric Data Set for Cosmology. ApJS 235, 33. doi:10.3847/1538-4365/aab4f5, arXiv:1708.01531.

Fausti Neto, A., da Costa, L.N., Carnero, A., et al., 2018. DES science portal: Creating science-ready catalogs. Astronomy and Computing 24, 52–69. doi:10.1016/j.ascom.2018.01.002, arXiv:1708.05642.

Flaugher, B., 2005. The Dark Energy Survey. International Journal of Modern Physics A 20, 3121–3123. doi:10.1142/S0217751X05025917.

Flaugher, B., Diehl, H.T., Honscheid, K., et al., 2015. The Dark Energy Camera. AJ 150, 150. doi:10.1088/0004-6256/150/5/150, arXiv:1504.02900.

Garilli, B., Guzzo, L., Scodeggio, M., et al., 2014. The VIMOS Public Extragalactic Survey (VIPERS). First Data Release of 57 204 spectroscopic measurements. A&A 562, A23. doi:10.1051/0004-6361/201322790, arXiv:1310.1008.

Garilli, B., Le Fèvre, O., Guzzo, L., et al., 2008. The Vimos VLT deep survey. Global properties of 20,000 galaxies in the $I_{AB} < 22.5$ WIDE survey. A&A 486, 683–695. doi:10.1051/0004-6361:20078878, arXiv:0804.4568.

Gatti, M., Vielzeuf, P., Davis, C., et al., 2018. Dark Energy Survey Year 1 results: cross-correlation redshifts - methods and systematics characterization. MNRAS 477, 1664–1682. doi:10.1093/mnras/sty466, arXiv:1709.00992.

Geha, M., Wechsler, R.H., Mao, Y.Y., et al., 2017. The SAGA Survey. I. Satellite Galaxy Populations around Eight Milky Way Analogs. ApJ 847, 4. doi:10.3847/1538-4357/aa8626, arXiv:1705.06743.

Georgakakis, A., Mountrichas, G., Salvato, M., et al., 2014. Large-scale clustering measurements with photometric redshifts: comparing the dark matter haloes of X-ray AGN, star-forming and passive galaxies at $z \sim 1$. MNRAS 443, 3327–3340. doi:10.1093/mnras/stu1326, arXiv:1407.1863.

Gerdes, D.W., Sypniewski, A.J., McKay, T.A., et al., 2010. ArborZ: Photometric Redshifts Using Boosted Decision Trees. ApJ 715, 823–832. doi:10.1088/0004-637X/715/2/823, arXiv:0908.4085.

Graff, P., Feroz, F., Hobson, M.P., et al., 2014. SKYNET: an efficient and robust neural network training tool for machine learning in astronomy. MNRAS 441, 1741–1759. doi:10.1093/mnras/stu642, arXiv:1309.0790.

Hearin, A.P., Zentner, A.R., Ma, Z., et al., 2010. A General Study of the Influence of Catastrophic Photometric Redshift Errors on Cosmology with Cosmic Shear Tomography. ApJ 720, 1351–1369. doi:10.1088/0004-637X/720/2/1351, arXiv:1002.3383.

Herlihy, M., Shavit, N., 2011. The art of multiprocessor programming. Morgan Kaufmann.

High, F.W., Stubbs, C.W., Rest, A., et al., 2009. Stellar Locus Regression: Accurate Color Calibration and the Real-Time Determination of Galaxy Cluster Photometric Redshifts. AJ 138, 110–129. doi:10.1088/0004-6256/138/1/110, arXiv:0903.5302.

Hildebrandt, H., Arnouts, S., Capak, P., et al., 2010. PHAT: PHoto-z Accuracy Testing. A&A 523, A31. doi:10.1051/0004-6361/201014885, arXiv:1008.0658.

Honscheid, K., Elliott, A., Bonati, M., et al., 2014. The DECam DAQ System: lessons learned after one year of operations, in: Software and Cyberinfrastructure for Astronomy III, p. 91520G. doi:10.1117/12.2057073.

Hoyle, B., Gruen, D., Bernstein, G.M., et al., 2017. Dark Energy Survey Year 1 Results: Redshift distributions of the weak lensing source galaxies. ArXiv e-prints arXiv:1708.01532.

Hoyle, B., Rau, M.M., Bonnett, C., et al., 2015. Data augmentation for machine learning redshifts applied to Sloan Digital Sky Survey galaxies. MNRAS 450, 305–316. doi:10.1093/mnras/stv599, arXiv:1501.06759.

Huterer, D., Kim, A., Krauss, L.M., et al., 2004. Redshift Accuracy Requirements for Future Supernova and Number Count Surveys. ApJ 615, 595–602. doi:10.1086/424726, arXiv:astro-ph/0402002.

Ilbert, O., Arnouts, S., McCracken, H.J., et al., 2006. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. A&A 457, 841–856. doi:10.1051/0004-6361:20065138, arXiv:astro-ph/0603217.

Jiang, L., Fan, X., Bian, F., McGreer, et al., 2014. The Sloan Digital Sky Survey Stripe 82 Imaging Data: Depth-optimized Co-adds over 300 deg^2 in Five Filters. ApJS 213, 12. doi:10.1088/0067-0049/213/1/12, arXiv:1405.7382.

Jones, D.H., Read, M.A., Saunders, W., et al., 2009. The 6dF Galaxy Survey: final redshift release (DR3) and southern large-scale structures. MNRAS 399, 683–698. doi:10.1111/j.1365-2966.2009.15338.x, arXiv:0903.5451.

Kaiser, N., Burgett, W., Chambers, K., et al., 2010. The Pan-STARRS wide-field optical/NIR imaging survey, in: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 0. doi:10.1117/12.859188.

Kessler, R., Marriner, J., Childress, M., et al., 2015. The Difference Imaging Pipeline for the Transient Search in the Dark Energy Survey. AJ 150, 172. doi:10.1088/0004-6256/150/6/172, arXiv:1507.05137.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., development team [Unknown], J., 2016. Jupyter notebooks ? a publishing format for reproducible computational workflows, in: Loizides, F., Scmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas, IOS Press. pp. 87–90. URL: https://eprints.soton.ac.uk/403913/.

Koposov, S., Bartunov, O., 2006. Q3C, Quad Tree Cube – The new Sky-indexing Concept for Huge Astronomical Catalogues and its Realization for Main Astronomical Queries (Cone Search and Xmatch) in Open Source Database PostgreSQL, in: Gabriel, C., Arviset, C., Ponz, D., Enrique, S. (Eds.), Astronomical Data Analysis Software and Systems XV, p. 735.

Le Fèvre, O., Vettolani, G., Garilli, B., et al., 2005. The VIMOS VLT deep survey. First epoch VVDS-deep survey: 11 564 spectra with $17.5 \leq IAB \leq 24$, and the redshift distribution over $0 \leq z \leq 5$. A&A 439, 845–862. doi:10.1051/0004-6361:20041960, arXiv:astro-ph/0409133.

Le Fèvre, O., Vettolani, G., Paltani, S., et al., 2004. The VIMOS VLT Deep Survey. Public release of 1599 redshifts to $I_{AB} \leq 24$ across the Chandra Deep Field South. A&A 428, 1043–1049. doi:10.1051/0004-6361:20048072, arXiv:astro-ph/0403628.

Lidman, C., Ardila, F., Owers, M., et al., 2016. The XXL Survey XIV. AAOmega Redshifts for the Southern XXL Field. PASA 33, e001. doi:10.1017/pasa.2015.52, arXiv:1512.04662.

Lidman, C., Ruhlmann-Kleider, V., Sullivan, M., et al., 2013. An Efficient Approach to Obtaining Large Numbers of Distant Supernova Host Galaxy Redshifts. PASA 30, e001. doi:10.1017/pasa.2012.001, arXiv:1205.1306.

Lilly, S.J., Le Brun, V., Maier, C., et al., 2009. The zCOSMOS 10k-Bright Spectroscopic Sample. ApJS 184, 218–229. doi:10.1088/0067-0049/184/2/218.

Lima, M., Cunha, C.E., Oyaizu, H., et al., 2008. MNRAS 390, 118–130. doi:10.1111/j.1365-2966.2008.13510.x, arXiv:0801.3822.

Lima, M., Hu, W., 2007. Photometric redshift requirements for self-calibration of cluster dark energy studies. Phys. Rev. D 76, 123013. doi:10.1103/PhysRevD.76.123013, arXiv:0709.2871.

Ma, Z., Bernstein, G., 2008. Size of Spectroscopic Calibration Samples for Cosmic Shear Photometric Redshifts. ApJ 682, 39–48. doi:10.1086/588214, arXiv:0712.1562.

Ma, Z., Hu, W., Huterer, D., 2006. Effects of Photometric Redshift Uncertainties on Weak-Lensing Tomography. ApJ 636, 21–29. doi:10.1086/497068, arXiv:astro-ph/0506614.

Mao, M.Y., Sharp, R., Norris, R.P., et al., 2012. The Australia Telescope Large Area Survey: spectroscopic catalogue and radio luminosity functions. MNRAS 426, 3334–3348. doi:10.1111/j.1365-2966.2012.21913.x, arXiv:1208.2722.

Masters, D.C., Stern, D.K., Cohen, J.G., et al., 2017. The Complete Cali-

bration of the Color-Redshift Relation (C3R2) Survey: Survey Overview and Data Release 1. ApJ 841, 111. doi:10.3847/1538-4357/aa6f08, arXiv:1704.06665.

Mohr, J.J., Armstrong, R., Bertin, E., et al., 2012. The Dark Energy Survey data processing and calibration system, in: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 0. doi:10.1117/12.926785, arXiv:1207.3189.

Momcheva, I.G., Brammer, G.B., van Dokkum, P.G., et al., 2016. The 3D-HST Survey: Hubble Space Telescope WFC3/G141 Grism Spectra, Redshifts, and Emission Line Measurements for ˜ 100,000 Galaxies. ApJS 225, 27. doi:10.3847/0067-0049/225/2/27, arXiv:1510.02106.

Morganson, E., Dark Energy Survey Data Management Team, 2016. The Dark Energy Survey Pipeline, in: American Astronomical Society Meeting Abstracts, p. 444.08.

Morganson, E., Gruendl, R.A., Menanteau, F., et al., 2018. The Dark Energy Survey Image Processing Pipeline. ArXiv e-prints arXiv:1801.03177.

Muzzin, A., Wilson, G., Yee, H.K.C., et al., 2012. The Gemini Cluster Astrophysics Spectroscopic Survey (GCLASS): The Role of Environment and Self-regulation in Galaxy Evolution at z ˜ 1. ApJ 746, 188. doi:10.1088/0004-637X/746/2/188, arXiv:1112.3655.

Nanayakkara, T., Glazebrook, K., Kacprzak, G.G., et al., 2016. ZFIRE: A KECK/MOSFIRE Spectroscopic Survey of Galaxies in Rich Environments at z ˜ 2. ApJ 828, 21. doi:10.3847/0004-637X/828/1/21, arXiv:1607.00013.

Newman, J.A., 2008. Calibrating Redshift Distributions beyond Spectroscopic Limits with Cross-Correlations. ApJ 684, 88–101. doi:10.1086/589982, arXiv:0805.1409.

Nord, B., Buckley-Geer, E., Lin, H., et al., 2016. Observation and Confirmation of Six Strong-lensing Systems in the Dark Energy Survey Science Verification Data. ApJ 827, 51. doi:10.3847/0004-637X/827/1/51, arXiv:1512.03062.

Parkinson, D., Riemer-Sørensen, S., Blake, C., et al., 2012. The WiggleZ Dark Energy Survey: Final data release and cosmological results. Phys. Rev. D 86, 103518. doi:10.1103/PhysRevD.86.103518, arXiv:1210.2130.

Perez, F., Granger, B.E., 2007. Ipython: A system for interactive scientific computing. Computing in Science Engineering 9, 21–29. doi:10.1109/MCSE.2007.53.

Rest, A., Scolnic, D., Foley, R.J., et al., 2014. Cosmological Constraints from Measurements of Type Ia Supernovae Discovered during the First 1.5 yr of the Pan-STARRS1 Survey. ApJ 795, 44. doi:10.1088/0004-637X/795/1/44, arXiv:1310.3828.

Sadeh, I., Abdalla, F.B., Lahav, O., 2016. ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. PASP 128, 104502. doi:10.1088/1538-3873/128/968/104502, arXiv:1507.00490.

Sánchez, C., Carrasco Kind, M., Lin, H., et al., 2014. Photometric redshift analysis in the Dark Energy Survey Science Verification data. MNRAS 445, 1482–1506. doi:10.1093/mnras/stu1836, arXiv:1406.4407.

Sandberg, R., Goldberg, D., Kleiman, S., Walsh, D., Lyon, B., 1985. Design and implementation of the sun network filesystem, in: Proceedings of the Summer USENIX conference, pp. 119–130.

Scolnic, D., Rest, A., Riess, A., et al., 2014. Systematic Uncertainties Associated with the Cosmological Analysis of the First Pan-STARRS1 Type Ia Supernova Sample. ApJ 795, 45. doi:10.1088/0004-637X/795/1/45, arXiv:1310.3824.

Scoville, N., Abraham, R.G., Aussel, H., et al., 2007. COSMOS: Hubble Space Telescope Observations. ApJS 172, 38–45. doi:10.1086/516580, arXiv:astro-ph/0612306.

Shvachko, K., Kuang, H., Radia, S., Chansler, R., 2010. The hadoop distributed file system, in: Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on, IEEE. pp. 1–10.

Silverman, J.D., Kashino, D., Sanders, D., et al., 2015. The FMOS-COSMOS Survey of Star-forming Galaxies at z˜1.6. III. Survey Design, Performance, and Sample Characteristics. ApJS 220, 12. doi:10.1088/0067-0049/220/1/12, arXiv:1409.0447.

Stalin, C.S., Petitjean, P., Srianand, R., et al., 2010. Optical identification of XMM sources in the Canada-France-Hawaii Telescope Legacy Survey. MNRAS 401, 294–306. doi:10.1111/j.1365-2966.2009.15636.x, arXiv:0909.0464.

Sullivan, M., Conley, A., Howell, D.A., et al., 2011. VizieR Online Data Catalog: Type Ia supernovae luminosities (Sullivan+, 2010). VizieR Online Data Catalog 740.

Tasca, L.A.M., Le Fèvre, O., Ribeiro, B., et al., 2017. The VIMOS Ultra Deep Survey first data release: Spectra and spectroscopic redshifts of 698 objects up to z_{spec} 6 in CANDELS. A&A 600, A110. doi:10.1051/0004-6361/201527963, arXiv:1602.01842.

Taylor, M.B., 2006. STILTS - A Package for Command-Line Processing of Tabular Data, in: Gabriel, C., Arviset, C., Ponz, D., Enrique, S. (Eds.), Astronomical Data Analysis Software and Systems XV, p. 666.

Treu, T., Schmidt, K.B., Brammer, G.B., et al., 2015. The Grism Lens-Amplified Survey from Space (GLASS). I. Survey Overview and First Data Release. ApJ 812, 114. doi:10.1088/0004-637X/812/2/114, arXiv:1509.00475.

Yuan, F., Lidman, C., Davis, T.M., et al., 2015. OzDES multifibre spectroscopy for the Dark Energy Survey: first-year operation and results. MNRAS 452, 3047–3063. doi:10.1093/mnras/stv1507, arXiv:1504.03039.

Appendix A. Data description

As a proof of concept, we show through this paper an example of the sequence of pipelines run to estimate photo-z and use these results to discuss the benefits of such infrastructure. The results presented in Section 3, after each pipeline methodology was described. In the following sections, we briefly describe the data used in those runs.

Appendix A.1. Photometric data

To describe the processes carried out in the Portal to estimate photo-zs, we use photometric data from the first annual internal release of DES. The observations were carried out with the mosaic camera DECam (Flaugher et al., 2015; Honscheid et al., 2014) built as part of DES project and mounted on the 4-meter Blanco telescope at Cerro Tololo Inter-American Observatory (CTIO), in Chile.

The data were reduced and calibrated by the DES Data Management (DESDM) team at the National Center for Supercomputing Applications (NCSA) using standard procedures described by Desai et al. (2012), Mohr et al. (2012), Morganson et al. (2018). This is the system used for the processing and calibration of DES data, and the DECam Community Pipeline. The observations (Diehl et al., 2014) reported here took place from August 2013 to February 2014 and include a total of 14,340 exposures in the *grizY* filters, covering a total area of ∼1,800 deg^2 in eight distinct regions, making the so-called DES Y1A1 release (Drlica-Wagner et al., 2018).

The two largest regions (see Figure A.26) are part of the wide-field survey. One of about 160 deg^2 overlapping the Sloan Digital Sky Survey Stripe 82 Imaging Data (S82, Jiang et al., 2014), and another of ∼1,400 deg^2 overlapping the region observed by the South Pole Telescope (SPT, Carlstrom et al., 2011). These two wide regions were covered with up to four passes in each filter, reaching SExtractor's `mag_auto` magnitude limits of $g = 23.4$, $r = 23.2$, $i = 22.5$, $z = 21.8$, and $Y = 20.1$ (Drlica-Wagner et al., 2018) in the AB system for a 10σ detection limit.

The remaining regions, called "supplemental fields" – where a large number of spectroscopic redshifts (spec-zs) are available – belong to both the science verification phase[20] (SVA1), and Y1A1 releases. Four of these regions are collectively known as

[20] https://des.ncsa.illinois.edu/releases/sva1

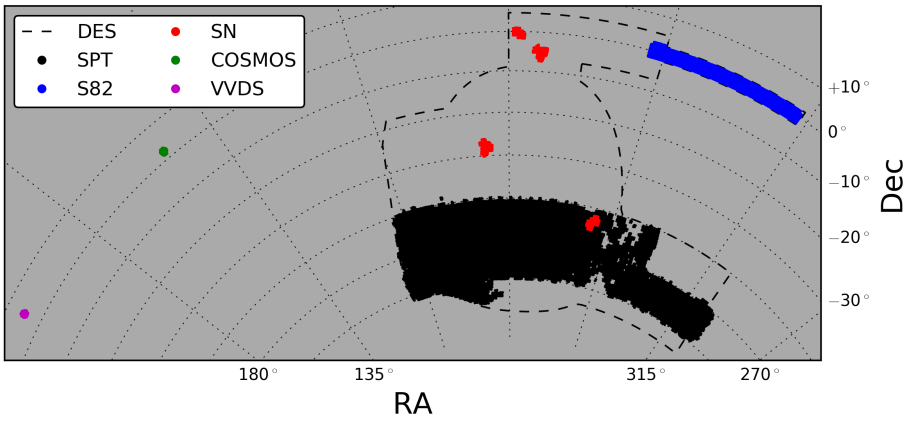


Figure A.26: Location of all the Y1A1 fields used in this paper and DES footprint (dashed line).

Supernova (SN) fields. One of the other regions overlaps with the VVDS-14h field from VIMOS VLT Deep Survey (hereafter VVDS, Le Fèvre et al., 2005) and the final region overlaps with COSMOS field (Scoville et al., 2007). The SN fields are regularly observed as part of the SNe Ia program, making available a greater number of exposures compared to the wide survey. The location of all these regions are shown in Figure A.26 along with the DES footprint. Relevant information is available in Table A.3.

Appendix A.2. Spectroscopic data

In this paper, we use a spec-z sample with reliable measurements to train photo-z algorithms, and to test their performance, as an example of validation procedure. The construction of this sample is by compiling data available from a large number of surveys individually ingested into the database associated with DES Science Portal.

Currently, the Portal's spectroscopic database contains redshift measurements from a total of 34 spectroscopic surveys. Together, these catalogs contain 2,173,561 redshift measurements, where 1,688,403 refers to extragalactic sources including both galaxies and quasars. In Table A.4 we show the information about the spectroscopic sample used as an example in Section 3.1. The surveys ordered by the number of successful matchings with photometric data from DES Y1 wide fields are the numbers in the fourth column. The second column shows the number of objects with good quality spec-zs per survey, after resolving multiple measurements as discussed in Section 3.1.

After dealing with quality cuts and multiple measurements in the spectroscopic database, we end up with 1,412,816 unique high quality spec-zs. However, we know that not all of these sources will be matched to the photometric sample since they extend beyond the Y1A1 DES footprint. In particular, around 170 thousand sources overlap with Y1 footprint.

Table A.3: Basic information of DES Y1 photometric datasets.

Field	# Objects	Area†	mag lim‡
COSMOS	313,380	2.97	23.6
SN	2,569,018	31.76	24.4
SPT	126,623,762	1,469.05	23.2
VVDS	260,446	2.91	23.6
S82	12,487,566	165.84	23.4

† Covered area in deg^2

‡ Defined as the peak of i-band MAG_AUTO number counts

Table A.4: Spectroscopic samples used in this paper.

Survey	# objects†	%	# matchings	%	z mean	z min	z max	Ref. ‡
PRIMUS	110,522	7.8	60,477	35.3	0.57	0.02	4.08	1
SDSS DR14	423,353	30.1	19,778	11.5	0.59	0.00	1.95	2
DES AAOmega	23,114	1.6	16,492	9.6	0.54	0.00	3.94	3
VIPERS	48,558	3.4	14,832	8.6	0.69	0.05	1.67	4
WiggleZ	80,431	5.7	9,131	5.3	0.55	0.01	1.70	5
VVDS	13,638	1.0	7,532	4.4	0.59	0.00	4.08	6
zCOSMOS	12,513	0.9	7,511	4.4	0.54	0.00	1.99	7
3D-HST	180,841	12.8	6,333	3.7	1.01	0.01	5.21	8
DEEP2	33,936	2.4	5,402	3.1	0.99	0.01	1.89	9
2dF	211,705	15.0	3,547	2.1	0.12	0.00	0.35	10
GAMA	7,429	0.5	3,444	2.0	0.22	0.01	0.74	11
ACES	4,047	0.3	3,045	1.8	0.58	0.04	1.42	12
6dF	108,760	7.7	2,637	1.5	0.06	0.00	0.38	13
DES IMACS	2,387	0.2	2,215	1.3	0.60	0.00	1.37	14
SAGA	64,033	4.5	1,994	1.2	0.29	0.01	1.17	15
NOAO OzDES	3,008	0.2	1,884	1.1	0.22	0.00	0.68	16
XXL AAOmega	3,143	0.2	926	0.5	0.47	0.00	2.80	17
SPT GMOS	2,189	0.2	790	0.5	0.56	0.07	1.24	18
UDS	1,307	0.1	705	0.4	1.06	0.04	3.44	19
SNLS FORS	1,321	0.1	529	0.3	0.51	0.03	3.75	20
ATLAS	729	0.1	503	0.3	0.32	0.02	1.89	21
Pan-STARRS	1,775	0.1	463	0.3	0.33	0.00	3.16	22
C3R2	1,249	0.1	429	0.3	0.92	0.03	3.52	23
SpARCS	403	<0.1	356	0.2	0.91	0.12	1.58	24
SNVETO	2,154	0.2	178	0.1	0.84	0.03	3.63	25
FMOS-COSMOS	328	<0.1	173	0.1	1.55	0.75	1.74	26
SNLS AAOmega	350	<0.1	58	<0.1	0.60	0.07	1.17	27
CDB	388	<0.1	38	<0.1	0.58	0.25	0.91	28
VUDS	141	<0.1	36	<0.1	1.78	0.19	3.75	29
ZFIRE	202	<0.1	29	<0.1	1.77	1.05	2.26	30
MOSFIRE	143	<0.1	25	<0.1	1.89	0.80	3.08	31
2dFLenS	63,632	4.5	23	<0.1	0.40	0.09	0.69	32
GLASS	383	<0.1	10	<0.1	1.06	0.34	2.07	33
XMM-LSS	26	<0.1	5	<0.1	0.42	0.19	0.65	34

† Only selected objects with $Q_{spec} \geqslant 3$

‡ References: 1- Coil et al. (2011); Cool et al. (2013) and https://primus.ucsd.edu/; 2- Abolfathi et al. (2017) and http://www.sdss.org/dr14/; 3- Yuan et al. (2015); Childress et al. (2017); 4- Garilli et al. (2014) and http://vipers.inaf.it/rel-pdr1.html; 5- Parkinson et al. (2012) and http://wigglez.swin.edu.au/site/; 6- Garilli et al. (2008); Le Fèvre et al. (2004); 7- Lilly et al. (2009); 8- Momcheva et al. (2016) and http://3dhst.research.yale.edu/Data.php; 9- Davis et al. (2003, 2007) and http://deep.ps.uci.edu/DR4/home.html; 10- Colless et al. (2001) http://www.2dfgrs.net/; 11- Driver et al. (2011); 12- Cooper et al. (2012) and http://mur.ps.uci.edu/cooper/ACES/zcatalog.html; 13- Jones et al. (2009) and http://www.6dfgs.net/; 14- Nord et al. (2016); 15- Geha et al. (2017) and http://sagasurvey.org/; 16- Yuan et al. (2015); Childress et al. (2017); 17- Lidman et al. (2016) and http://cosmosdb.iasf-milano.inaf.it/XXL/; 18- Bayliss et al. (2016) and https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OR13NN/; 19- http://www.nottingham.ac.uk/astronomy/UDS/UDSz/; 20- Bazin et al. (2011) Private communication; 21- Mao et al. (2012); 22- Rest et al. (2014); Scolnic et al. (2014); Kaiser et al. (2010); 23- Masters et al. (2017); 24- Muzzin et al. (2012); 25- http://www.ast.cam.ac.uk/~fo250/Research/SNveto/; 26- Silverman et al. (2015) and http://member.ipmu.jp/fmos-cosmos/FC_catalogs.html; 27- Lidman et al. (2013); Yuan et al. (2015); Childress et al. (2017) and http://apm5.ast.cam.ac.uk/arc-bin/wdb/aat_database/observation_log/make; 28- Sullivan et al. (2011); 29- Tasca et al. (2017) and http://cesam.lam.fr/vuds/DR1/; 30- Nanayakkara et al. (2016) and http://zfire.swinburne.edu.au/data.html; 31- http://mosdef.astro.berkeley.edu; 32- Blake et al. (2016) and http://2dflens.swin.edu.au/; 33- Treu et al. (2015) and https://archive.stsci.edu/prepds/glass/; 34- Stalin et al. (2010).