PZ WG Onboarding

This page supplies information for those getting started with the Photometric Redshifts (PZ) Working Group (WG)! It is meant to be a living document. Feel free to ask questions below about any PZ concepts you are wondering about, or to add content answering any questions you used to be confused about. The idea is to create a database of basic information that will allow anyone to get up to date with the basic ideas and goals that are necessary to become a contributing member of the PZ working group.

To join the PZ WG, just add yourself to the email list and join the #desc-pz Slack channel.

If you are new to DESC, you should review the DESC-wide onboarding materials first. I particularly recommend the introduction to DESC, the onboarding video, and the new member FAQ. The recording of a special onboarding presentation for the PZ WG may be found here and the slides may be found here.

Requests for onboarding content

If there is anything PZ related that you are wondering about, feel free to add it to the list of topics below. In addition, feel free to add any concepts you remember being confused about when you were a PZ newcomer. The members of the PZ WG will try to answer your question and add it to the list of intro topics below.

- [x] What do RAIL and qp stand for and what do they do?
- [x] Training vs calibration?
- [x] p(z) vs n(z)?
- [x] What is cross-correlation?
- [x] What are the big picture goals of the WG?
- [ ] (Add your questions here!)

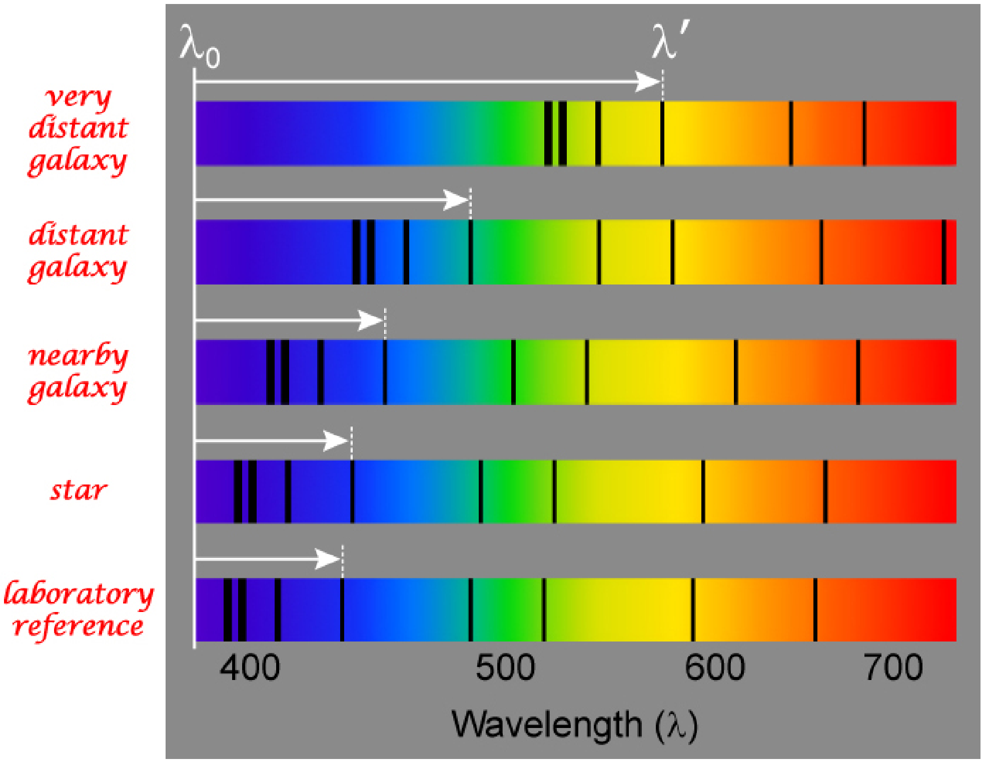
General Photo-z Questions

1. What is redshift?

The wavelength of light is stretched as it travels through the expanding universe. This results in distant objects looking redder than you would normally expect. This effect is referred to as "redshift." Redshift is one of the most important observables in cosmology. By correlating redshifts with distance measurements (e.g. by measuring the redshift of galaxies that host type Ia supernovae which serve as standard candles in the cosmic distance ladder), we can measure the expansion rate of the universe. Once we have pinned down the expansion history of the universe, we can use an object's redshift to calculate other quantities, such as the distance to the object and the age of the universe at which we are imaging the object. For more details on cosmology and redshift, two good resources are the lecture notes of David Tong and Daniel Baumann.

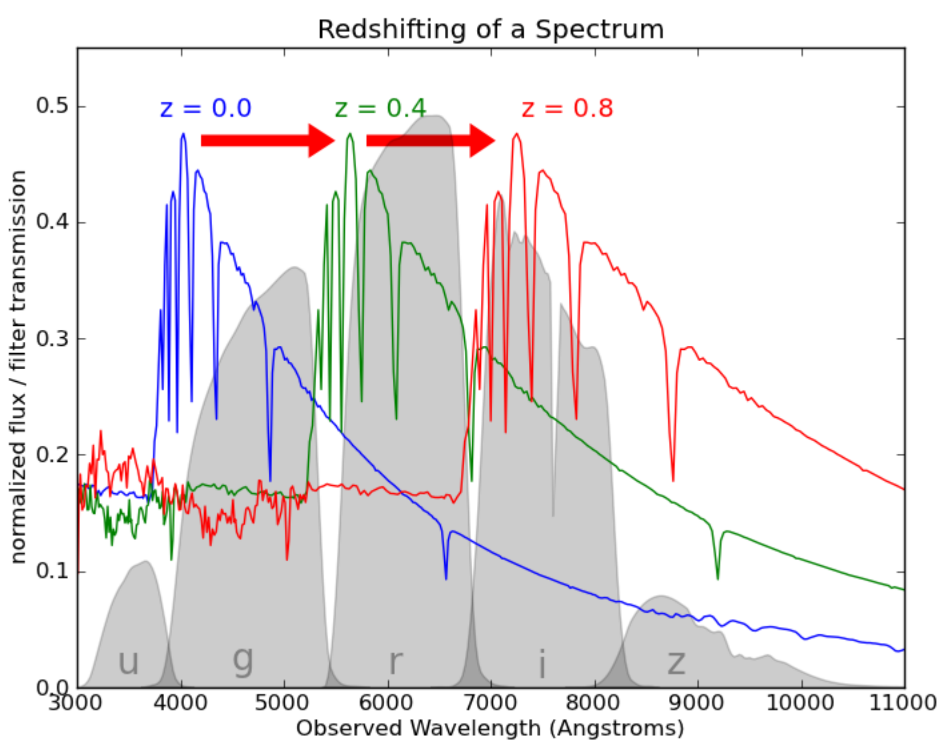
2. What are photo-z's?

The most straightforward way to measure the redshift of an object is via its spectrum. By measuring the observed wavelength of known emission and absorption lines, we can determine how far the lines have been shifted towards the red end of the spectrum (c.f. the figure below). This allows us to calculate the redshift. A redshift calculated with this method is known as a spectroscopic redshift (spec-z).



(credit: JPL)

The problem is that spectroscopy takes far more time than photometry. Since LSST (a photometric survey) will image billions of galaxies, it is impractical to collect spectra for all sources. As a result, we must rely on photometry to estimate redshifts. In this context, you can think of photometry as a very low resolution spectrum. Rather than estimating the redshift from the flux at a near continuum of wavelengths, we must estimate the redshift from the flux integrated over 6 broad bandpasses. Redshifts estimated from photometry are known as photometric redshifts (photo-z's).



3. Photometry vs color?

Photometry is the flux of an astronomical source, integrated over a bandpass filter. You can think of it as the (weighted) average flux of the source over the wavelength range covered by the filter. LSST has 6 filters: *u g r i z y* (warning: *z* is also the variable used to denote redshift). Photometry is often reported as a magnitude.

Color refers to differences between the *magnitudes* of adjacent bands, e.g. *u-g*, *g-r*, *r-i*, etc. You can think of color as measuring the slope of the spectrum between the two bandpasses. A positive color means that the flux is roughly decreasing with wavelength across the relevant wavelength range.

Note that we often use this distinction loosely and assume the meaning is understood from context. If you are confused, feel free to ask!

4. How are photo-z's estimated?

There are roughly two broad classes of photo-z estimators: *template based* and *machine learning* (ML) based.

Template based estimators are based on the physical model demonstrated above. They assume all galaxy spectra resemble a discrete set of SED templates, and calculate the colors of those templates across a range of redshifts. The redshift of a galaxy is then assigned by matching it to the template and redshift whose colors most closely match the observed colors. Some methods (e.g. BPZ) incorporate other information through the use of Bayesian priors. The SED templates are derived from a combination of stellar synthesis models and spectroscopy of nearby galaxies.

The strength of template based methods is that they are physically motivated, but they rely on assembling an accurate and representative set of galaxy spectra which is very difficult. In addition, these methods do not perform as well as ML methods across the redshift range covered by spectroscopic training sets (cf. the weaknesses of ML based estimators below), but are more accurate when extrapolating beyond that redshift range.

ML based estimators on the other hand use a training set of galaxy photometry and spec-z's to empirically derive a relationship between color and redshift. The strength of these methods is that they tend to perform very well when applied to target galaxies that resemble galaxies in the training set. However, they are susceptible to bias and large errors when the training set is not representative of the target population (see the section on systematics below), and they perform poorly when applied to high-redshift galaxies which are poorly represented in spectroscopic training sets.

Note that some methods (e.g. Crenshaw & Connolly 2020) combine the strengths of template and ML methods.

There is also ongoing research on photo-z estimation using convolutional neural nets (CNNs; see e.g. Schuldt et al. 2020 and the sources cited therein). CNNs estimate redshifts directly from galaxy images rather than photometric catalogs, allowing them to automatically incorporate information such as galaxy size and morphology.

For a list and broad overview of many standard photo-z estimators, see the PZ WG Data Challenge 1 (DC1) Paper.

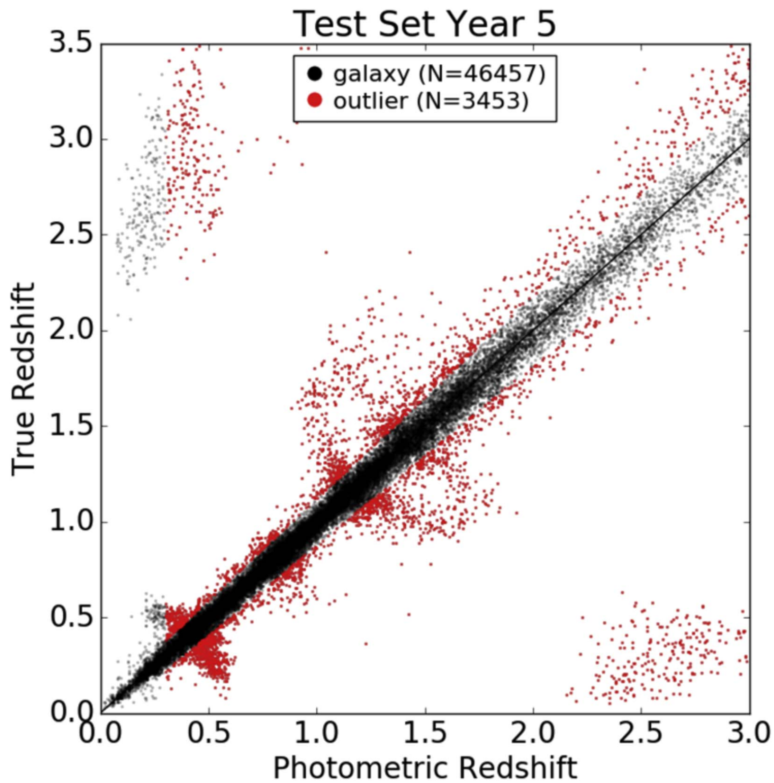
5. Posteriors vs point estimates?

Ideally, a photo-z estimator should take galaxy photometry as an input and return a photo-z posterior: *p(redshift | photometry)*. This is the probability of the galaxy being at a range of different redshifts, *given* its observed photometry. It is a probability distribution that contains a lot of important physical information.

Often, instead of the full photo-z posterior, we report a single point estimate. This may be because storing the full posterior is too expensive, or because we want to compare the photo-z to the spec-z, which is just a single value. The most common photo-z point estimate is the mode of the posterior distribution. We don't use the expectation value (aka the mean) because photo-z posteriors are often bimodal. The expectation value of a bimodal distribution often lies in the very low probability region between the two modes, and is thus a terrible summary of the posterior. For example, if we think a galaxy probably has a redshift of ~0.5 or ~1.5, it makes no sense to report its redshift as 1.

Using point estimates instead of the full posterior is not a very smart thing to do. The posterior contains a lot more information than the point estimate, especially since photo-z posteriors are often complicated and multimodal. The kind of comprehensive statistical analysis that is required for cutting edge precision cosmology demands that we start using and evaluating the full photo-z posterior. We are working hard on this! (see e.g. the sections on RAIL and qp below)

Note that you will often see photo-z plots that look like this one below. This is a comparison of photo-z point estimates to the true spec-z's of a test set. This is the most basic evaluation of photo-z estimators that we perform.



(credit: Graham et al. 2018)

6. What are outliers, bias, and scatter?

Remember how we are in the bad habit of talking about photo-z point estimates instead of the full photo-z posterior? Well outliers, bias, and scatter are the main ways we quantify the quality of photo-z point estimates. We get photo-z's (point estimates) and spec-z's for a set of galaxies, then compare how well the photo-z's match the spec-z's. We calculate the quantity $\Delta z = \frac{z_\mathrm{p} - z_\mathrm{s}}{1 + z_\mathrm{s}}$, where *z_p* is the photo-z point estimate and *z_s* is the spectrocopic redshift. This value measures the residuals, normalized for greater uncertainty at high redshift. We calculate the standard deviation of the Delta-z's and report that as the scatter. The fraction of Delta-z's greater than 3 standard deviations from zero is reported as the outlier fraction. The mean of the remaining Delta-z's is reported as the bias.

(Note that the exact definition of these statistics may vary from paper to paper, but these definitions give you an idea of what we mean when we discuss these quantities)

7. What are photo-z systematics?

"Systematics" are systematic errors that may impact photo-z's. Potential sources of systematic errors include:

- incompleteness - failure of the spectroscopic training set to adequately cover the full range of redshift and photometry present in the target set
- spectroscopic errors - incorrect spec-z's in the training set that may result from errors such as misidentification of emission lines during spectroscopy
- deblending errors - incorrect galaxy photometry that results from the difficulty of measuring photometry for blended (i.e. overlapping) galaxies
- cosmic variance - spectroscopic training sets are often assembled by observing a small region of the sky to high depth. The number of galaxies as a function of redshift in the spectroscopic sample will exhibit fluctuations due to the large-scale structure along the line of sight. In addition, the assembly history of dark matter halos is different in overdense and underdense regions, so they will contain systematically different galaxy populat ions.
- List some more

It is important that these systematics are quantified and mitigated because small biases in photo-z's can easily dominate the error budget of cosmology analyses. We don't want our photo-z's to limit the power of LSST to determine cosmological parameters!

For a review of photo-z systematics and how they impact training and calibration, see Newman et al. 2015.

8. What is n(z)?

Photo-z posteriors for individual galaxies are usually denoted *p(z)*. The redshift distribution for an ensemble of galaxies is denoted as *n(z).* Often in cosmology, such as in weak lensing tomography (see the section on photo-z's in cosmology below), we are more interested in *n(z).* Estimating *n(z)* is related to but distinct from *p(z)* estimation. For detailed discussion, see Malz & Hogg 2020, Malz 2021, or the PZ WG Data Challenge 1 (DC1) paper.

9. Training vs calibration?

"Training" is the process in which the machine learning (ML) based photo-z estimators ingest a spectroscopic training set to learn the relation between photometry and redshift. Once trained, these estimators are ready to estimate redshits for new galaxies.

Probe WGs consuming estimates *n(z)* of the redshift distribution sometimes employ a shifting procedure *n(z)* *n(z + Delta z)*, often referred to as "calibration," intended to adjust the average (a.k.a. mean a.k.a expected value a.k.a. first moment) *<n(z)>* of the estimated redshift distribution to match one derived from some other source of information, such as cross-correlations between spatial positions of galaxies relative to a spectroscopically observed sample (see other methods for redshift estimation below).

10. What makes photo-z's hard?

The dream of a perfect photo-z estimator is unobtainable. This is not only because photo-z's rely on less information than spec-z's and will therefore always be more uncertain, but also because of the very real physical degeneracies between galaxy color and redshift. Essentially, high-redshift blue galaxies can look a lot like low-redshift red galaxies (because the high redshift makes the blue galaxy appear red). This degeneracy is responsible for the clouds of outliers in the corners of the photo-z vs spec-z plot above.

There are many estimators that seek to break this degeneracy using other information. For example, BPZ uses a galaxy type and magnitude prior, and cross-correlation methods use clustering statistics (see the section on other methods below). Using this additional information can essentially break the degeneracy in some cases by favoring one of the redshift modes over others, but the degeneracy is still real and present in the posterior. Accounting for the real, physical degeneracies present in photo-z's is important for a full statistical cosmology analysis, and is one of the reasons why we need to use full photo-z posteriors instead of just point estimates.

11. What are the Balmer and Lyman Breaks and how are they relevant for photo-z's?

The Balmer and Lyman breaks are discontinuities in galactic spectra caused by neutral hydrogen absorption (both within the galaxy and in clouds along the line of sight). The Balmer break, caused by the Balmer series, results in lower flux below ~3645 angstroms, while the Lyman break, caused by the Lyman series, results in lower flux below ~1200 angstroms. There is also a break at 4000 angstroms caused by the absorption from metals in stellar atmospheres.

Because photo-z's rely on the flux integrated in broad filters, they are more sensitive to broad, dramatic features of the SED. This means that the location of these breaks provides a lot of information for photo-z's. In particular, the Balmer and 4000 angstrom breaks are in the wavelength range of the LSST filters up to a redshift of ~1.4, and locating these breaks with the LSST filters provides good leverage for photo-z's (see e.g. Kalmbach et al. 2020 and Malz et al. 2021). Note that while the Balmer break leaves the LSST filters around z=1.4, the Lyman break doesn't enter the wavelength range of the LSST filters until about z=2.5. This gap in redshift coverage contributes to the degradation of photo-z's at high redshift.

Note that in high-redshift galaxies, photo-z estimators might confuse the Lyman break for the Balmer break. This contributes to the physical degeneracies discussed above in "What makes photo-z's hard?"

12. Are there methods for estimating redshift other than photo-z's?

In addition to estimating redshifts with photometry, we can estimate redshift posteriors via the cross-correlation of galaxy positions. This is because galaxies that are close to one another on the sky are more likely to be clustered and therefore at similar redshifts. Therefore, if we have a sample of galaxies for which spec-z's are known, we can estimate redshift posteriors for other galaxies by cross-correlating their positions on the sky with the positions of the spectroscopic sample. This can provide valuable information to break the degeneracies present in photo-z's. For discussion of this method, two good sources are Rau et al. 2021 and Sanchez & Bernstein 2018.

13. How are photo-z's used in cosmology?

- ***Weak lensing (WL) tomography*** involves splitting background galaxies up into different redshift bins, and using the shear of each redshift bin to map the 3D distribution of matter and large scale structure. The important quantity in WL tomography is *n(z)*. For LSST *n(z)* and the determination of tomographic bins must be done using photometric redshifts. The accuracy and precision of the redshifts used in tomography have a strong impact on the dark energy parameters determined via WL. For a paper on using photo-z's for WL tomography, see this Dark Energy Survey (DES) Year 3 paper.
- ***Type Ia Supernovae (SNIa)*** are standard candles in the cosmic distance ladder. In other words, we know their intrinsic brightness, so by observing the brightness of distant supernovae, we can calculate how far away they are. If we combine the SNIa distance measurements with the redshifts of their host galaxies, we can map out the expansion history of the universe. This is how dark energy was originally discovered, and remains one of the best ways to measure the expansion of the universe. For a paper on performing SN cosmology with photo-z's, see Mitra & Linder 2020.
- ***Luminous Red Galaxies (LRGs)*** are often the most luminous galaxies in galaxy clusters. They are a relatively homogeneous population with a very narrow range of colors and intrinsic luminosity, and a strong 4000 angstrom break. They can be observed to great distance and usually have highly accurate photo-z's. This makes them an ideal probe of large-scale structure. See e.g. Rozo et al. 2015.
- ***Redshift Space Distortions (RSDs)***: While we often think of galaxies as being stationary except for their recession velocity due to the expansion of the universe (the so-called Hubble flow), this is not strictly true. Galaxies are also moving under the influence of their local gravitational potential. This gives rise to peculiar velocity, which also contributes to the observed redshift of a galaxy. As a result of peculiar velocities, when

you plot the large scale structure of the universe as a function of redshift instead of distance, the distribution of galaxies looks distorted. These distortions are called Redshift Space Distortions (RSDs). Correlations in RSDs can be used to map the distribution of mass in the universe. For a paper on how to perform RSD cosmology with photometric surveys, see Asorey et al. 2014.
- Add other cosmology probes.

14. What are some good review papers?

For a general review of photo-z's, a good choice is Salvato et al. 2018.

For a review of the issues in training and calibrating photo-z estimators for modern dark energy surveys, a good choice is Newman et al. 2015.

For a review of some standard photo-z estimators and the previous work of the PZ working group, see the PZ WG Data Challenge 1 (DC1) paper.

Working Group Questions

1. What are the big picture goals of the WG?

The goal of the PZ working group is to provide DESC with accurate and precise photo-z estimates. This may involve developing better photo-z estimators, but also includes quantifying the impact of systematic errors (see the section on systematics above) and developing the pieces of the DESC software pipeline relevant to photo-z estimation and storage.

The scope of the PZ WG is outlined in more detail in the LSST-DESC Science Roadmap (SRM), a living document that DESC maintains to keep track of progress towards our goal of building a coherent software pipeline to illuminate the dark universe. Though it has evolved with our understanding of software pipelines, project management, and photo-z methods, as of SRM v2.3, the big picture goals pertain to three pieces of software pipeline infrastructure:

- *PZEstimate*: derives per-galaxy redshift posterior PDF estimates (*p(z)*) from photometry by multiple methods
- *PZSummarize*: constrains the redshift distribution (*n(z)*) of galaxy samples from photometry by multiple methods
- *PZValidate*: stress-tests the PZEstimate and PZSummarize pipelines

Currently (summer 2021), the PZ WG is engaged in two large projects: RAIL and qp.

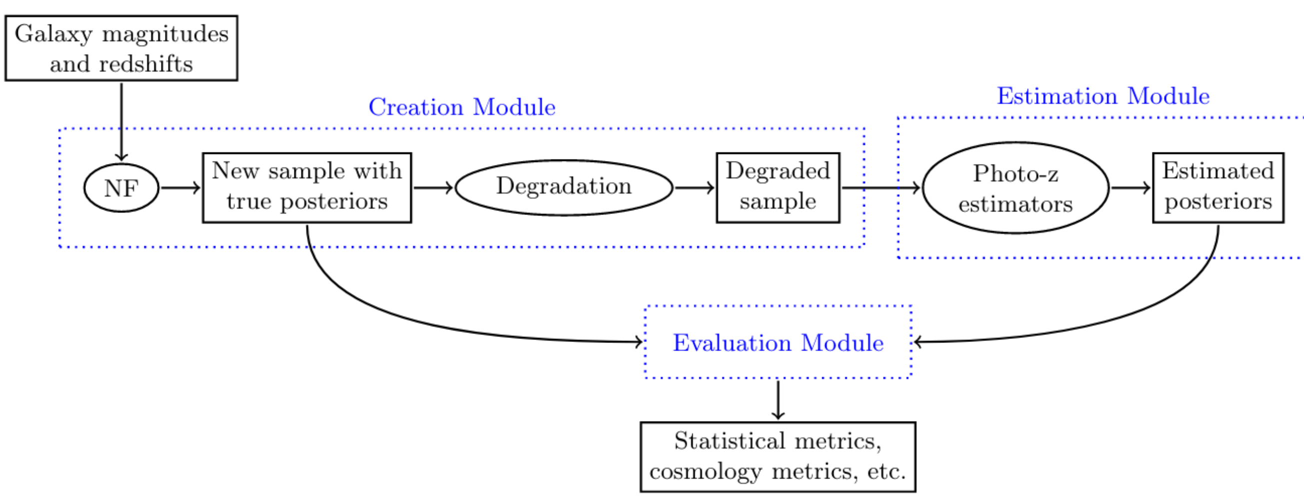
2. What is RAIL?

RAIL stands for Redshift Assessment Infrastructure Layers. It is a project to comprehensively evaluate the various photo-z estimators on the market, and to quantify how different systematic errors impact the photo-z posteriors estimated by each. It consists of three modules.

1. The *Creation* Module is a generative model of spectroscopic data that generates galaxy photometry and redshifts with true photo-z posteriors. The generative model powering the Creation Module is a normalizing flow (NF). The Creation module also includes a *Degradation* submodule that adds systematic errors and biases to the training set.
2. The *Estimation* module wraps photo-z estimators that are trained on the biased spectroscopic training sets output by the Degradation submodule. These estimators will be used to generate photo-z posteriors for a test set of galaxies.
3. The *Evaluation* module will compare the true photo-z posterior for test set galaxies to the estimated posteriors generated by the photo-z estimators which are trained on the degraded training sets. This will allow us to determine how each systematic error impacts the photo-z posteriors generated by each photo-z estimator.

RAIL will allows us to quantify how systematic errors impact photo-z's and the cosmology analyses of DESC. Importantly, the Creation module, which generates galaxy photometry with true photo-z posteriors, will allow us to evaluate photo-z estimators using the full photo-z posterior, instead of just the point estimates.

The figure below outlines the general structure of RAIL:



(credit: Crenshaw et al. 2021)

In addition to the Estimation module which estimates a *p(z)* for individual galaxies, we are developing a *Summarization* module, which estimates *n(z)*.

There is a lot of work to be done for RAIL and we welcome anyone who wants to contribute! The #desc-pz-rail Slack channel is the best place to get involved, as well as to find reminders of the weekly tag-ups and co-working opportunities.

3. What is qp?

qp is a package for storing and handling univariate probability density functions (PDFs), regardless of their parameterization. The package was originally developed for Malz et al. 2018 but has since expanded in scope. We will use qp to store the photo-z posteriors calculated for DESC cosmology. To get started with qp, see the qp docs.

4. Are there other projects going on?

There are many projects going on in the PZ working group. For a full list, see the list of projects at the bottom of the PZ WG page. If you would like to get involved with an ongoing project or start a new project, reach out to one of the WG conveners, or send a message in the #desc-pz Slack channel. There is a lot of work to be done, and we welcome anyone who wants to get involved!

5. How do I get involved?

The best way to get involved is to join the #desc-pz Slack channel and ask how you can contribute. There is a lot of work to be done in the PZ WG, ranging from simple projects for newcomers to complex tasks for photo-z experts. The working group conveners and other members would be happy to help get you started collaborating on one of the ongoing projects or to launch a new project of your own. The #desc-pz-rail Slack channel is also a great place to get involved in the RAIL project, which has many opportunities for new contributors.