The Dark Energy Survey Project Office

The Dark Energy Survey

The Challenge 5 Document: Simulation, Data, and Analysis

Version 3.145

James Annis
Huan Lin
Joe Mohr*
Josh Frieman*
Ofer Lahav*
*To be confirmed

30 June 2009

CONTENTS

1 INTRODUCTION

The challenge cycle of simulations, data reduction, testing, and science analysis is central to our preparations for the science of the DES. The challenge cycle consists of science driven goals, n-body simulation group produced catalogs, image simulation group produced images, data management reduced images and catalogs, and science working group analyses. Often the simulations are at the catalog level and used in analysis challenges directly. Some simulations are transformed into images and reduced with the DESDM data reduction system.

This document describes the challenge cycle and writes down specific data quality goals. Having specific goals helps rationalizes the science working group testing. The document preparation process included collating science working group input, the science requirements document, simulation development planning, and the goals of DESDM data challenge.

The document also serves to document our requirements for access to the science data and science codes by the collaboration during the analysis challenge.

2 SCIENCE OBJECTIVES

The testing of the science analyses drives simulation areas of order of 5000 sq-degrees. The production and testing of these are done entirely inside the science working groups.

The testing of the image processing, catalog generation, and calibration drives a choice of simulation area of 100-200 sq-degrees. The simulations that go from catalogs to images to data reduction through analyses cross three project boundaries in the DES, from DECam to DESDM to the Science Committee. Thus the bulk of this document describes this effort. The three primary scientific measurement goals of the challenge 5 cycle are to improve the measurements of galaxy shapes, the measurements of photometry, and the implementation of a survey mask in a form useful to the science working groups. The simulation group will improve the detector/atmosphere/Galaxy model realism of the simulated images in order to allow meaningful tests of achievement against these goals. A clear analysis goal is to improve the accessibility of the DC data by the working groups.

These all will be tracked under the challenge 5 umbrella.

2.1 SCIENCE ANALYSIS SIMULATIONS

The testing of the science analyses drives simulation areas of 5000 sq-degrees or more. In this cycle, the simulation group will produce several large scale catalog level simulations, some with only the dark matter distribution, some with galaxies emplaced, and some with SZ producing hot gas emplaced. This is a science working group responsibility.

The Large Scale Structure working group needs simulations as a function of cosmology and resolution. The Stanford simulations, the MICE simulations, and TreePM runs from Berkeley are serving as their analysis platforms.

The Clusters working group needs mocks with galaxies and preferably SZ inducing hot gas. The Stanford group is producing Hubble Volume simulation and Stanford simulation based AddGals catalogs for the work on cluster finding and selection function computation. Two mocks are being constructed for optical/SZ joing cluster finding analysis. One is based on the Millennium Gas Simulations from Michigan, the other using halo-model based techniques from a Berkeley centered group.

The Weak Lensing working group has asked the N-body simulation group to pursue a 5000 sq-degree mock shear catalog for use in shear correlation and covariance analysis concurrent with DC5. The details of this catalog are to be determined. The Cluster working group wishes to use this catalog to test the mass calibration of clusters.

2.2 DATA REDUCTION SIMULATIONS

The testing of the image processing, catalog generation, and calibration drives a choice of simulation area of 100-200 sq-degrees. The three overall goals of this challenge cycle are to improve the measurements of shape, the measurements of photometry, and the measurement of the survey mask. The responsibility for the production of these images lies in the DECam simulation group, but science working group input, especially from the N-body simulation group, is crucial.

The interest in improving shape measurement quality is driven by the desire to perform an astrophysically meaningful shear signal measurement using the simulated data. The interest in improving photometric quality is driven by a desire to improve the photometric redshifts, of central import to weak lensing, clusters, and large scale structure. The interest in improving the knowledge of the survey mask is driven by the desire of the large scale structure working group to perform a full scale power spectrum analysis using masks derived from the data set.

Several science working groups desire to add specific object catalogs to the image simulations to test the extraction of these objects.

- Quasars: will provide a QSO catalog especially for quasars at 4 <= z <= 7.
- Quasars and Milky Way: will provide a low mass L/T dwarf star catalog to T=500K. These are targets for the MW group and noise for the Quasar group.
- SN: will provide an asteroid catalog for the SN field simulations
- Milky Way: will provide a galactic structure based star catalog in DES colors

The Strong Lensing group will provide the AddArcs code for the production of realistic giant arc lenses to be placed in the DC5 images.

2.3 SCIENCE REQUIREMENTS

The SRD and the DESDM TRD provide requirements that will need to be met by the time of science commissioning. Any of these can be tested at the end of the Cycle 5 process. For the purposes of Cycle 5, however, we focus on the three overall goals of the challenge cycle, improving the measurements of shape, photometry, and implementing a survey mask. The responsibility for the testing lies in the project office and the science working groups.

In image quality, we focus on six requirements.

1) The median PSF FWHM (averaged over all exposures and over the survey area) in each of the r, i, and z bands should be less than 0.9”
2) The mean PSF whisker length for stars/exposure must be below 0.2” in the r,i,z bands for the wide area survey.
3) For measured shear $\gamma_i^{meas} = (1+m)\gamma_i^{true} + \gamma_i^{add}$. We have a requirement that m<0.004 and $\gamma_i^{add} < 4\times10^{-4}$. For the purposes of this cycle, we set as a goal $\gamma_i^{add} < 1\times10^{-3}$.
4) The residual mean whisker length for stars on scales of 10 arcmin and 1 degree, after removal of a static component (i.e., the same for all exposures) and a bilinear fit in (x,y) per exposure, should be below 0.06” in *r, i,* and *z* bands.
5) The slope of the *i* band FWHM and ellipticity versus r-z color should be smaller than 0.005”/mag.
6) Lastly there is an astrometric requirement that affects image quality: the positions of stars in the reduced frames that contribute to a coadd tile should have an rms of < 30 mas.

In photometric measurements quality, we focus on seven requirements:

1) Limiting magnitudes of grizY= 24.0, 24.0, 24.0, 23.6, 21.6 AB in 1.6” apertures at 10σ measurement precision, 97.5% completeness, and 100mas astrometry, over at least 90% of the survey area.
2) 2% photometric calibration. Note that this applies to the coadd and not the reduced images, but that a calibration should be applied there as well.
3) The galaxy catalog completeness should, at the 10σ limit, be independent of seeing, sky brightness, and transparency to within 5%.

4) The galaxy catalog completeness should be independent of galaxy surface density over the range 1-50 galaxies/sq-arcminute to within 5%.
5) The galaxy catalog impurity due to stars at galactic latitude b=30 should be within 5% of that at b=90. Furthermore, at i< 21 the misidentification rate of stars (including saturated stars) as galaxies must be <3%.
6) In a set of galaxy bins of $\Delta z=0.1$ at $0.1 <= z <= 1.5$, the number of stars masquerading as galaxies should match the summed stellar likelihood of the objects classed as galaxies.
7) Types of magnitudes required: aperture magnitudes (for sanity checks), PSF magnitudes (for calibrations), total magnitudes (for galaxies, akin to SDSS Petrosian or cModel magnitudes), color optimized magnitudes (again for galaxies, akin to SDSS model mags, which fixed the model profile in the r-band and used it for the other bands.)

The survey mask has one requirement that we focus on:

1) The area of the survey at a given limiting magnitude must be known to 0.5%.

3 SIMULATION CHALLENGE GOALS

The simulated area will again consist of 100-200 sq-degrees of sky to the full depth of the survey.

A smaller area simulation that is relatively quick to produce and reduce is planned. The gold standard or golden night is one tile observed to the full depth of the survey, organized as if it was taken during one night, along with enough data to produce the overlapping adjacent tiles. It consists of roughly 52 science images plus calibrations and flat fields. Prior to the large area simulation and reduction, the golden night data will be iterated on several times to improve algorithms and validate the simulations. The gold standard night (GSN) can be thought of as a 'high level unit test' of the data processing system. The Gold Standard night was designed to assist in meeting the goals of the astronomy codes (PSF photometry, galaxy model fit photometry, etc), but it will also be useful for helping SWG validate their codes. We propose two iterations with the SWGs, limiting interfere with the astronomy code development goals.

The simulations are produced by the DECam simulation group.

3.1 CATALOG LEVEL GOALS

The Simulation working group will produce new mock galaxy catalogs with increased realism in the physics of galaxy evolution. These catalogs will have shapes associated with the galaxies. The shapes will be sheared by a model of the intervening mass. Required is shear from the intervening dark matter halos in the halo catalog produced by the simulation group, assuming shear from a SIS or NFW profile. A goal is to have the shear from the underlying cosmic density field in the dark matter simulation.

The Milky Way study group will produce a star catalog for use in the simulations that incorporates the same DES system response curves used to produce the galaxy colors. The brighter stars ($i < 20$) will be consistent with the USNO-B catalog, and the fainter stars from a physically meaningful galactic structure model. For USNO-B stars, proper motion should be included.

The Survey Strategy team will produce an observing plan for the data that mimics baseline survey strategy.

3.2 IMAGE LEVEL GOALS

3.2.1 Photometry

The DECam system will have a spatially varying system response due to the effects of the quantum efficiency curve differing from CCD to CCD and due radial variations in the filters. Spatial variation of the system response will be modeled as a color dependent correction to the fluxes of objects performed on a CCD by CCD basis.

The sky transparency is spatially dependent under slightly non-photometric conditions. This will be modeled as a quadratic polynomial in focal plane x and y.

The sky brightness is spatially dependent for several reasons, the largest of which is the presence of the moon. Sky brightness will be modeled as a quadratic polynomial in focal plane x and y.

Flat fields differ from chip to chip. This will be modeled using a variety of flat fields gathered from CCD testing lab. Calibration flat fields will be provided that differ from the underlying simulation flat fields.

The DECam electronics exhibit crosstalk between amplifiers on the same CCD at the 0.2% level and between CCDs on the same electronics board at the 0.1% level. Cross talk between CCDs sharing the same card will be modeled.

3.2.2 Shape

Pixel size variations will be modeled. The model for the geometrical distortion due to the optics will be upgraded to include the size change in the pixels. Likewise the CCD itself has variable pixel sizes due to the electrostatic field inside the chip bending at the array edges ("the glowing edges") and column to column periodic size changes and these will be modeled. The effects of distortion due to differential refraction by the atmosphere will be included (note that this primarily affects astrometry).

The PSF library will upgraded to include PSFs appropriate to small amounts of defocus and misalignment.

Galaxy shapes realized from a shapelet catalog derived from high resolution HST data will be used. The 2 sq-degree HST COSMOS data will allow more better galaxy shape coefficients to be determined and thus the galaxy shapes in the simulations will be more reliable.

The eight focus and alignment chips will have their output simulated.

Atmospheric differential refraction will be modeled.

4 DATA CHALLENGE GOALS

The reduction of the simulated images to reduced, coadded images and catalogs is performed by the DESDM group.

4.1 ASTRONOMY CODES

The astronomy codes are a DESDM responsibility.

4.1.1 Photometry

New model fitting photometry capabilities have been developed within SExtractor during DC4. For DC5 we expect to employ the model fitting photometry for all the coadd images. We will explore whether having model fitting photometry on the single epoch images is required to achieve any of our data quality requirements.

We will integrate and test new star/galaxy classification capabilities available within SExtractor and apply them to the DC5 dataset.

We will develop and test an algorithm to estimate a source subtracted background within SExtractor and employ that in the DC5 cataloging. We expect this algorithm to resolve the slight background overestimation that we have seen in the DC4 testing.

PSF homogenization code was developed and integrated into a DC4 prototype coaddition pipeline. Significant work remains to tune this code and understand its effects on completeness and photometric measurement uncertainties. We will continue our efforts to develop this functionality for the project, because it goes hand in hand with using the new SExtractor model fitting photometry on coadds.

Effort will be invested to improve the image by image scaling of the fringe correction.

We are developing code to create a pupil ghost from calibration data.

We are developing code to extract the crosstalk coefficients directly from the DECam observations.

We will develop a code to produce star flats from the DECam photometric catalogs and incorporate a star flat correction into the core image processing.

We will test object detection and deblending in modestly crowded environments working towards the science requirements for clusters.

We will test PSF magnitudes of stars against model fitted magnitudes as a test of the PSF shape determination.

4.1.2 Shapes

We are working to improve image photometric flatness within the DESDM pipelines by developing a new flatness correction that accounts for the differences in pixel scale over the CCD during the flat fielding process.

4.1.3 Masking

We will continue our work on image masking, extending the existing codes to include masking of satellite trails. We expect to employ the Hough transform, which is well suited to detecting linear features in data. We will continue to improve our stellar masking code that removes diffraction spikes and bleed trails as well as the bright halo surrounding the brightest sources. Additional testing and tuning of the DECam cosmic ray masking using the Terapix eye and SExtractor software will be carried out.

We will adopt a new mode of calculating weight maps that do not include the noise contributions from object flux. These new weight maps are expected to provide improved quality coadd images and to be more useful for generating survey depth as a function of the position,

4.2 SCIENCE CODES

The science codes are a science committee responsibility. The DESDM group has the responsibility of integrating the codes and running the production.

4.2.1 Photometry and Astrometry

The tests of DC4 should be rerun again on DC5 to check behavior against the new simulations. In particular, the better atmosphere and optics model used in this round of simulations present challenges to the astrometry.

4.2.2 Weak Lensing Pipeline

In DC4 we developed and used the first DES weak lensing pipeline. But this is just a first step, because it calculated shear from each single epoch image independently. For DC5 we expect to test a version of the pipeline that carries out joint shear measurements using all images available (selected by band, seeing, depth) for a given object. We will continue to work with members of the Weak Lensing SWG on this project. Our goal is to run a pipeline in DC5 that produces scientifically useful data for the WL SWG.

The PSF determined by the WL working group should be tested against the whisker requirement. In particular the ability of the code to take advantage of time independent features of the PSF distribution about the focal plane is of interest.

4.2.3 Difference Imaging Pipeline

We will continue to work with the SNe SWG in the development of a difference imaging pipeline. The plan is for this pipeline to be used by the DESDM operations team to carry out difference imaging within the broader DES survey, and by the SNe SWG to carry out the difference imaging within the SNe science fields. Our goal is to run a pipeline in DC5 that produces scientifically useful data for the SNe SWG.

4.2.4 Large Scale Structure Analysis Masks

We are working with the LSS SWG to create a pipeline that will calculate survey masks in a form that is required for the analysis of clustering data. There is significant design work remaining, but the idea is to work from either DC5 image metadata or the weight maps for the DC5 coadds to build the data products needed to understand the variations in depth as a function of position within the survey. This is tracked at the pixel level in the coadd images, but a higher level and more compressed data product is needed to enable efficient analysis of the DC5 galaxy clustering.

4.2.5 Photometric Redshifts

The photometric redshift code from the Photo-z Working Group has long been integrated into the data pipelines. DC5 will focus on improving the science code to meet the science requirements. One clear desideratum is a pdf for the photo-z along with the single best photo-z.

5 ANALYSIS CHALLENGE GOALS

The Supernova working group has a number of specific goals for DC5 analysis.

1. Achieve a SN detection efficiency of >50% for peak magnitude i_AB in a SN deep field.
2. Achieve photometric accuracy of 10% in a real time search.
3. Add backgrounds including asteroids and non-Ia SN transients to image simulations.
4. Generate and process a full season of SN data for one SN field.
5. Implement asteroid rejection code and measure its efficiency.
6. Test SN-specific monitoring and quality control tools developed by the SN group.
7. Test SN-identification and fitting methods.

The following are initial suggestions for specific goals for DC5 analysis for the Clusters, LSS, WL and Photo-z working groups. The working groups will provide their goals after they have had internal discussions, including input on how the groups would use two golden night iteration before the Data Challenge to refine and validate algorithms.

Clusters:

1. Achieve 80% completeness and purity against the input halo catalog for log(M) > 13.8 and z < 1.3.
2. Achieve cluster photo-zs with a precision of σ_z <= 0.02.
3. For clusters at log(M) > 13.8, measure the fractional change in cluster richness from input catalog to DC5 cluster catalog- it should be less than 10%.

The Large Scale Structure:

1. Measure the galaxy-galaxy correlation function on scales of ~1 arcminute to ~1 degree.

Weak Lensing:

1. Measure the shear-shear correlation function on scales of ~1 arcminute to ~1 degree.

Photometric Redshifts:

1. Achieve σ_z <0.15 for photo-zs of all galaxies with >10σ in two or more bandpasses.
2. Measure the photo-z bias in the 15 Δz= 0.1 bins at 0 < z < 1.5 with an uncertainty of <0.003(1+z).
3. Measure the photo-z dispersion in these same bins to an uncertainty of < 0.01.
4. In non-overlapping z bins, aim to have the square-root of the covariance of the photo-z bias be < 0.0015(1+z).

6 ANALYSIS CHALLENGE TECHNICAL REQUIREMENTS

In order for the science collaboration to analyze the data challenge data set, or indeed the real DES data set, there are a series of technical requirements that should be met. Given the time constraints on DC5 (the DC5 DESDM plan was written in April, this document was written in June, and the DC5 is scheduled to start in October) it is understood that the requirements here are goals to be aimed for.

Issues specific to public access to DES-CAP, a database access interface similar to SDSS-CASJOBS, may be deferred while providing for the needs of collaboration access, which in the early years is the primary target audience.

These are the responsibility of the DESDM group and the science working groups.

6.1 CODE AND DATA VERSIONING

We define a reduction run as processing of a data set with a common set of software versions. An "official" reduction then is a single (or small set of) reduction run(s) that the data management has tagged as correct for release to the collaboration or public.

Then:

6.1.1 Official versions

The data challenge data products should have an official version, including which nights data were used, which software versions were used, and which libraries were used.

6.1.2 Tagged code versions

Only tagged versions of the code should be used for processing data products released to the collaboration or the public. Running off the trunk of the CVS/SVN repository is not acceptable. Golden night data sets, which are designed for testing purposes, are the exception.

6.1.3 Processing metadata

As only tagged code is run on official data products, the version number of the code should be stored in the output files and be made searchable in the database. Version numbers should be available for all codes used in production, DESDM or Science Committee based.

6.2 DATA SELECTIONS

6.2.1 Ability to select by reduction

It must be possible to select a unique set of data products based on reduction run.

Selection includes the ability to retrieve the data products. A reduction run also includes which nights data were used in the reduction.

6.2.2 Ability to select by code version

Likewise, it must be possible to select for all data products reduced using a given version number of the code.

6.2.3 Ability to select the entire survey

It must be possible to select a survey wide unique set of observations and data products from a single or a few official reduction runs. Preferably all data should be passed through a common set of software versions.

6.2.4 Ability to easily select the entire survey

It must be easy (transparent to the user) to select from the official set of reductions runs. If the user doesn't know which one to use, the default should be the official version which delivers a unique set of the available data products.

6.2.5 Simple queries

As an example of what "easy" means, we introduce the idea of a **simple** query. A **simple** query is hard to define, but can be shown by example:

```
select count(*) from dc5_coadd_objects
```

This query should return the number of unique coadd objects for all of DC5, drawn from the official reduction run(s). The user should not have to know how to select the right reduction runs, nor have to do explicit table joins. Magnitudes, shears, photo-zs: all should be available through a single **simple** query.

It may be efficient to make some 'pre-join' tables at various stages in the Data Challenge to avoid reselecting huge joins over and over. For example, a Galaxy Catalog database output is of central interest to the Clusters and Large Scale Structure groups. Note that the April Directors' Review report included a comment about a coadd science table that if it existed would minimize the need for joins by presenting a single table of the most relevant data to the collaboration scientists. The goal is to allow collaboration scientists who are relatively inexperienced with full SQL, including complex joins, to compose queries which retrieve their intended data sets with minimal pain. A good set of examples (which can be cut-and-pasted) will go a long way here, as would an editable query history capability.

The Golden Standard Night output may be small enough that the joins necessary could be redone each time by each end-user.

6.2.6 Example use cases

This list of examples is designed to illustrate the spirit of the above requirements. It is to be thought of as an illustrative subset. It must be possible with a single **simple** query to:

1. Select the official set of unique reduced images, along with the corresponding official unique reduced catalogs.
2. Select the official reduced catalog associated with an official reduced image.
3. Select the official unique set of coadd images and corresponding catalogs.
4. Select the single epoch or coadd image on which a given object is detected.
5. Select the input reduced images or remapped images that contribute to an output coadd tile.
6. Select all unique observations of all objects inside the official unique set of objects and download the catalog, including shears.

6.3 ACCESS TO THE PORTAL

SWG members will have access to a portal on the DESDM database which can perform efficient queries of large numbers of selected quantities from calibrated object tables and return the output (which may be up to several Gigabytes per query) in ASCII (CSV, TSV) format to the SWG member's home machine for further analysis.

Efficient means: indexed on at least RA,DEC but also on such things as run number, color, magnitude in several magnitude systems. It also implies that if necessary, 'stored procedures' would be developed to pre-calculate 'science object tables', derived from the join of the astronomy code tables (SExtractor catalogs) and the science working group code tables (i.e. WL, photo-z), if necessary, which would be updated frequently, perhaps whenever a co-add table was generated.

Timescales for queries to return typical volumes should be specified. Some example queries or query patterns, specified by the SWG or Project Scientist, would help.

Evaluation of the portal in the light of precursor work such as SDSS-CAS, CASJOBS, and MyDB should be performed.

6.4 AUTHORIZATION

An Authorization mechanism to the database via a portal from the outside should be specified and allow for 'power user access' during the data challenges. i.e. can Authorization be 'command line scriptable', so that one doesn't have to enter a user/passwd by hand into a browser window for each query of the database.

6.5 INTERFACE/DESIGN SPECIFICATIONS

Interface and/or design specifications should exist:

1) for data
    a. input and output file formats
    b. grouping requirements [i.e. can you run on one CCD frame in one

filter, or do you need a whole set of 62 CCDs, or do you need all to-be-coadded observations over the same part of the sky, plus these auxillary catalogs]

2) for code
   a. must it compile on what Linux node with what gcc libraries, 32/64 bit issues, multi-thread capability issues or restrictions
3) for parameter/control files
   a. files which control tunable parameters in the code
   b. control where the code can find its data and where it can run, etc.

6.6 UNIT TESTING

Individual sections of science code, including astronomy codes (Sextractor) and SWG codes (WL, difference pipeline) should be 'unit testable' using small subsets of the data (simulated or real), which may be extracted from the large DESDM database, or provided by the code developers.

6.7 CODE REPOSITORY AND VERSIONING

All pipeline code (astronomy codes and SWG codes) should be checked in to some sort of version control system (i.e. SVN) with increasing version numbers. All collaborators should eventually have read access to all pipeline source code.

When a data challenge test is run which crosses 'institutions' (i.e. involves more than one working group or the DESDM and a working group), a 'tagged' version of code should exported from the appropriate repository is used. There may be more than one repository at more than one institution, at least initially.

6.8 CODE CHANGE CONTROL

All DES software code have some change control mechanisms, aiming for a balance between avoiding locking down versions of code too early before a data challenge and avoiding changes at the last minute which break other connected pieces of a pipeline during a mass-processing session.