SPOKES: an End-to-End Simulation Facility for Spectroscopic Cosmological Surveys

B. Nord[a], A. Amara[b], A. Réfrégier[b], La. Gamper[b], Lu. Gamper[b], B. Hambrecht[b], C. Chang[b], J. E. Forero-Romero[g], S. Serrano[h], C. Cunha[c,f], O. Coles[j], A. Nicola[b], M. Busha[f], A. Bauer[h], W. Saunders[i], S. Jouvel[h], D. Kirk[j], R. Wechsler[c,d,f]

[a]*Fermilab Center for Particle Astrophysics, Fermi National Accelerator Laboratory, Batavia, IL 60510-0500*
[b]*ETH Zurich, Department of Physics, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland*
[c]*Department of Physics, Stanford University, Stanford, CA 94305*
[d]*SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., MS 29, Menlo Park, CA 94025*
[e]*Institute for Theoretical Physics, University of Zurich, 8057 Zurich, Switzerland*
[f]*Kavli Institute for Particle Astrophysics and Cosmology, 452 Lomita Mall, Stanford University, Stanford, CA, 94305*
[g]*Departamento de Física, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio Ip, Bogotá, Colombia*
[h]*Institut de Ciències de l'Espai, IEEC-CSIC, Campus UAB, Facultat de Ciències, Torre C5 par-2, Barcelona 08193, Spain*
[i]*Australian Astronomical Observatory, PO Box 915 North Ryde NSW 1670, Australia*
[j]*Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK.*

Abstract

The nature of dark matter, dark energy and large-scale gravity pose some of the most pressing questions in cosmology today. These fundamental questions require highly precise measurements, and a number of wide-field spectroscopic survey instruments are being designed to meet this requirement. A key component in these experiments is the development of a simulation tool to forecast science performance, define requirement flow-downs, optimize implementation, demonstrate feasibility, and prepare for exploitation. We present SPOKES (SPectrOscopic KEn Simulation), an end-to-end simulation facility for spectroscopic cosmological surveys designed to address this challenge. SPOKES is based on an integrated infrastructure, modular function organization, coherent data handling and fast data access. These key features allow reproducibility of pipeline runs, enable ease of use and provide flexibility to update functions within the pipeline. The cyclic nature of the pipeline offers the possibility to make the science output an efficient measure for design optimization and feasibility testing. We present the architecture, first science, and computational performance results of the simulation pipeline. The framework is general, but for the benchmark tests, we use the Dark Energy Spectrometer (DESpec), one of the early concepts for the upcoming project, the Dark Energy Spectroscopic Instrument (DESI). We discuss how the SPOKES framework enables a rigorous process to optimize and exploit spectroscopic survey experiments in order to derive high-precision cosmological measurements optimally.

Keywords: computation, cosmology, simulation, spectroscopy, extragalactic, galaxies

Email address: `nord@fnal.gov` (B. Nord)

1. Introduction

Progress in cosmology over recent decades has led to some of the most pressing questions in fundamental science today, such as those related to the nature of dark matter, dark energy, and gravity on cosmological scales. To address these questions, several wide-field spectroscopic surveys are in progress or being planned, including WiggleZ (Drinkwater et al., 2010), the Hobby-Eberly Telescope Dark Energy EXperiment (HETDEX; Adams et al., 2010), the Prime Focus Spectrograph (PFS; Takada et al., 2012), the Big Baryon Oscillation Spectroscopic Survey (BigBOSS; Schlegel, 2011), the Dark Energy Spectrometer (DESpec; Abdalla et al., 2012), the Dark Energy Spectroscopic Instrument (DESI[1]) and the 4m Multi-Object Spectroscopic Telescope (4MOST; de Jong et al., 2012). The goal of these experiments is to provide three-dimensional maps of the large-scale structure of the universe by measuring the angular positions and redshifts of galaxies in large cosmological volumes.

To reach the levels of precision that are needed to address the fundamental open questions in cosmology, these experiments must meet stringent requirements on both statistical power and control of systematic errors. These requirements, therefore, drive all aspects of these experiments from instrument design to survey optimization. Simulation tools play a key role in the design and optimization process: they are important for forecasting science performance of a given experimental configuration and, moreover, to demonstrate the feasibility of a mission design. Rigorous simulation tools can also allow the science team to prepare for the science interpretation and exploitation of the data.

Such simulation tools have been developed for a number of cosmological surveys. For example, the optical imaging project, Sloan Digital Sky Survey (SDSS; York and et al, 2000) employed the Monte-Carlo technique to test deblending in the image processing pipeline prior to the survey taking place[2]. SDSS also used simulations of galaxies, stars and QSOs to prepare and calibrate analysis pipelines for object classification (Fan, 1999; Strateva et al., 2001), and for measurements of the galaxy luminosity function (Blanton et al., 2001).

The Dark Energy Survey (DES; Flaugher et al., 2012) will rely on large-scale simulated catalogs to forecast cosmological constraints (e.g., Bernstein et al., 2012), develop science analysis pipelines (e.g., Chang et al., 2014), and improve the survey strategy (Neilsen, 2012). Galaxy catalogs, along with pixel-level image simulations, also permit the development of image reduction pipelines. Next-generation experiments, like the Large Synoptic Survey Telescope (LSST; LSST Dark Energy Science Collaboration, 2012), will employ photon-level simulations to account for sources of noise, such as atmospheric turbulence (Connolly et al., 2010; Claver et al., 2012; Chang et al., 2012). In addition, operational procedures, like survey strategy, have benefited from extensive simulations (Delgado et al., 2006; LSST Science Collaboration, 2009; Gibson et al., 2011; Honscheid et al., 2012).

As simulations and forward modeling methods play an increasingly important role in survey design and analysis, new frameworks for simulations have been developed. For example, the Monte-Carlo-Control-Loop (MCCL) method, proposed by Réfrégier and Amara (2013), aims to build a robust set of control loops, based on simulations, for verifying that complex measurement methods meet systematic requirement levels. Such system-level optimizations have underscored the need for fast simulations leading to efforts to develop simulations that are fast enough to support such integrated development. An example of this is Ultra Fast Image Generator (UFig; Bergé et al., 2013), which has been developed to quickly and efficiently produce simulated wide-field survey images.

Spectroscopic surveys can take advantage of the same kinds of mock galaxy catalogs as imaging surveys for forecasting. However, there are more operations and additional levels of complexity in spectroscopic surveys: for example, targets must be preselected before the surveying can begin, and for each tile on the sky, fibers are allocated to sources; moreover, these operations are intertwined, such that decisions regarding one will affect one or more of the others.

[1] `http://desi.lbl.gov`

[2] `http://www.astro.princeton.edu/~rhl/photo-lite.pdf`

In response to these challenges, spectroscopic experiments have undertaken several design approaches. For example, some recent surveys, such as SDSS-III's Baryon Oscillation Spectroscopic Survey (BOSS) and the 6dF Galaxy Survey (6dFGS), focused simulation efforts toward optimizing the fiber allocation and tiling algorithms (Campbell et al., 2004; Blanton et al., 2003). Studies for BigBOSS have performed target selection on mock catalogs and simulated two-dimensional images of galaxy spectra in an effort to develop the tools to extract spectra from images (Schlegel, 2011). 4MOST has developed the Facility Simulator (Boller and Dwelly, 2012), which links together the survey strategy and fiber allocation to convert an input catalog (from an imaging survey) into one that would result from a 4MOST survey.

In this paper, we describe the SPectrOscopic KEn Simulation (SPOKES), an end-to-end simulation facility for spectroscopic cosmological surveys. SPOKES is built on an integrated infrastructure, modular function organization, coherent data handling, and fast data access. These key features allow reproducibility of pipeline runs, enable ease of use, and provide flexibility to update functions within the pipeline. The pipeline's framework is also cyclic: it can be easily executed in a loop, offering the possibility to make the science output an efficient measure for design optimization and feasibility testing. While the framework is general, we use the design of the DESpec experiment concept (Abdalla et al., 2012) as a baseline for development and for benchmarking results. DESpec was one of the early concepts for the upcoming DESI experiment.

We present here the architecture, and the science and computational performance results of the SPOKES simulation pipeline. SPOKES and all the modules are written in the Python programming language[3].

The paper is organized as follows. In §2, we describe the challenges that spectroscopic surveys need to meet in order to reach the required precision, as well as the principal ingredients in a framework that simulates surveys. In §3, we present SPOKES and show how its design addresses these challenges. In §4, we present science and performance results of the simulation pipeline. Our conclusions are summarized in §5. Details regarding the data format choices and the input cosmological simulation are described in the Appendix.

2. Challenges for Spectroscopic Survey Simulations

2.1. Challenges

Future wide-field spectroscopic surveys offer great promise to address the fundamental questions described above. Their exploitation, however, will pose the following challenges that need to be addressed in order to achieve the required accuracy.

- *High precision:* The next generation of spectroscopic surveys, along with other Stage IV dark energy experiments (Albrecht et al., 2006), aim to measure the dark energy equation of state parameter, w, to percent-level precision. This sets ambitious requirements—e.g., that these experiments cover large cosmological volumes and maintain tight control over errors.

- *Systematics:* As the statistical power of surveys increases, numerous sources of systematic errors become significant. These include errors in the calibration of the survey selection function, inhomogeneous photometric target selection, masking, etc. These systematics need to be carefully calibrated and controlled so that they become subdominant compared to statistical errors.

- *Complexity:* The difficulty in controlling systematic errors is compounded by the fact that errors can couple to one another in a non-linear fashion in spectroscopic surveys. For example, there is an an interplay between target selection and fiber allocation, as each type of target (e.g., luminous red galaxy, emission line galaxy or QSO) will generally require observation through a different wavelength range. Each fiber is attached to a spectrograph with a specific wavelength range. Therefore, there must be enough fibers of each type (i.e., of each wavelength range) available for each type of target

[3] http://www.python.org

selected. Unless such effects and couplings are well modeled and tested, there is a risk that these effects will be imprinted on the final measured galaxy correlation function. This would then lead to systematic errors in the inferred cosmological parameters.

- *Pre-survey critical decisions*: Spectroscopic surveys differ crucially from imaging surveys, because decisions need to be made about the target sample before the spectroscopic survey is started. Target pre-selection influences the possible instrument configurations and increases the importance of modeling the measurement process at early stages before the data are collected. Given limited time and resources to perform a survey, it is likely to be difficult to drastically alter survey strategy at late stages to alleviate systematic errors arising from target selection.
- *Heritage:* Mapping large-scale structure with wide-field spectroscopic surveys is a mature field. As a result, there exist numerous tools and methods that have been developed for their exploitation. The incorporation of these resources is highly desirable, but also challenging. Many tools originate in heterogeneous code bases, making it difficult to integrate them into a coherent framework without significant modification .

2.2. Simulation Requirements

Extensive simulations are a key element in meeting these challenges, because they can be used to design and validate an experiment. Furthermore, simulations will play an important role in developing the data processing framework for the scientific exploitation of the surveys. Several ingredients are required of the pipeline for it to be used for these purposes.

First, to fully predict the results of a survey and to contend with complex or non-linear coupling of errors, the simulations should track all steps of the experiment. This points toward the need for an *end-to-end* simulation framework. It would start with mock galaxy catalogs specific to input cosmological parameters, perform all the elements of the experiment and analysis, and then end with constraints on those cosmological parameters—all performed in a single run of the simulation.

The architecture will need to be fully *integrated*, such that all the functions communicate with the rest of the pipeline components through the same mechanism—ensuring data and logic to be tracked precisely (i.e., provenance). In particular, the different simulation functions should pass parameters and data consistently and clearly from one function to the next.

The simulation framework should allow for a high level of *reproducibility*: one should be able to produce identical results with identical inputs. This requires careful and comprehensive book-keeping of all relevant parameters. For example, stochasticity needs to be controlled by saving the value of random seeds for each calculation, such as for the generation of noise in target spectra. This fine level of provenance is critical for systematic optimization of experiment parameters.

The framework must nevertheless be sufficiently *flexible* to permit the ingestion of new heterogenous functions and the modification of current functions. In addition, some operations of the experiment or analysis of the data will be time-consuming or memory-heavy. Therefore, the pipeline will need to accommodate a range of execution modes—on the one hand accomodate high-speed, high-efficiency runs and on the other permit computationally-intense (e.g., image-level) calculations, which may require parallelization. The simulation pipeline should thus be sufficiently flexible to accomodate a *high dynamic range* in detail (of what is simulated) and in parallelization.

The run time of the pipeline will depend on the run mode — i.e., the level of detail simulated. There should be a ‘minimal’ mode that can run relatively precisely to recover the correct output of the experiment, while running fast enough to allow for a large number of iterations to explore a large space of input parameters. Exploration of this parameter space can give us a better understanding of the sensitivity of the experiment to systematic effects and variation in experiment parameters (e.g., Réfrégier and Amara, 2013).

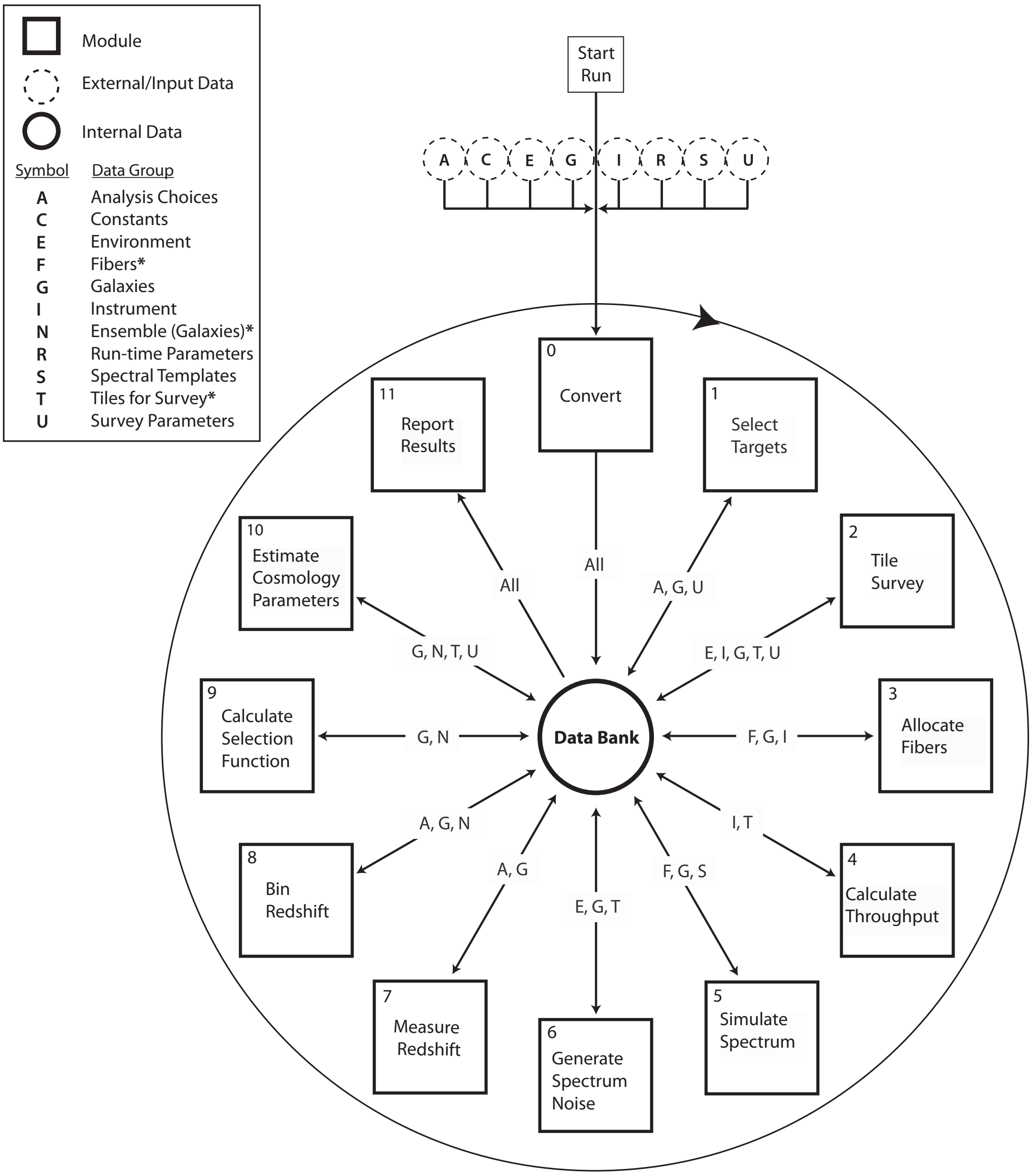

Figure 1: The workflow of the SPOKES pipeline depicts the data management, the modules and operations needed to simulate a spectroscopic survey and the sequence of those operations. Data are represented by circles and discussed in detail in §3.4: external input data (dashed) are further delineated in Table 1; internal data (solid) are held in the central databank in the native SPOKES format. The data are split into data groups to simplify provenance and access. Each data group is represented by a symbol, as shown in the legend, and the data groups marked with '*' are created within the pipeline and not ingested from an external source. The modules (squares) access (arrows) data from the databank, but otherwise do not interact. The data groups that are used or created by a given module are listed on that module's access arrow. In addition, we describe the specific data elements used and created by each module in §3.2. The wheel-like format of the pipeline shows the data management scheme, the independent nature of the modules, and the cyclic nature of the pipeline's execution. Once the results have been analyzed, the user can update parameters or data sets and re-run the pipeline in an effort to meet the science goal; this can be done by hand or in an automated fashion.

3. The SPOKES Facility

The SPOKES simulation facility is designed to meet the above requirements for simulation pipelines in wide-field spectroscopic surveys. There are two principal layers in the SPOKES pipeline. The algorithmic layer (shown in Fig. 1) governs the aspects related to experiment planning and science analysis. The infrastructure layer (shown in Fig. 2) contains the novel computing elements that allow the pipeline to meet requirements of next-generation simulations.

Fig. 1 shows the pipeline in an algorithmic context: sequentially and in clock-wise order, each function takes data and parameters from the databank and creates new data to be used later in the pipeline. The first task of the pipeline is performed by Module 0, which imports multiple sets of heterogeneous data into a databank (see §3.4 and Table 1 for details). The pipeline then proceeds to perform the functions of a spectroscopic survey. It selects targets from mock photometric catalogs, then performs survey and fiber allocation operations, measures spectroscopic redshifts, and derives cosmological constraints (Modules 1-10; see §3.2). At the conclusion of computations, Module 11 produces a summary report that includes diagnostic plots and statistics from a single run through the pipeline. The pipeline can also be directed to run cyclically—either traversing a pre-determined space of input parameters or searching through the parameter space until some metric is optimized.

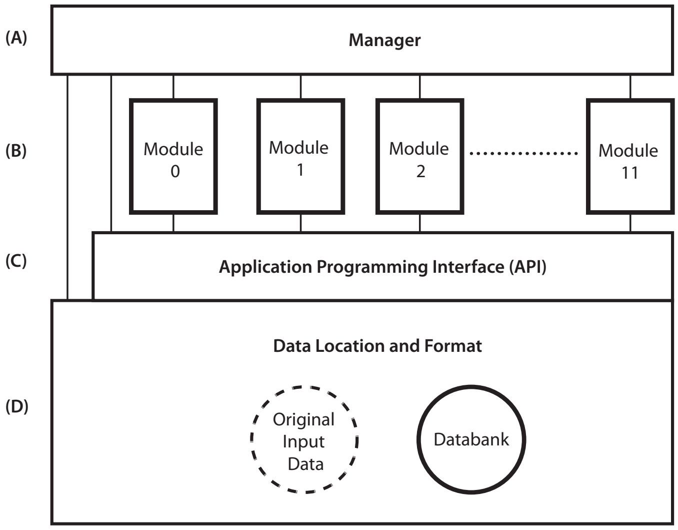

Figure 2: Architecture of the SPOKES pipeline. The Manager (section A) connects the multiple elements of the pipeline, uses the API to access data, and is the main interface for the user: it is the point where the user selects the modules to use and sets the simulation run parameters. The modules (B) read and write data via the API, but they are otherwise independent of one another. The API (C) handles data access throughout the pipeline, starting with conversion of the "original input data" (dashed circle) to the SPOKES "databank" format (solid circle) and storage in the central databank (D).

Each module accesses only the data it needs from the databank, which simplifies the interfaces between modules and makes them highly independent of one another. Note that the only interaction between modules occurs via the exchange of data with the databank. The data groups that are read, used or written by each module are shown along the arrows between the module and the databank. The legend in Fig. 1 provides a symbol for each data group, and the list in §3.4 describes the contents of each group. We also note specific data products that are used or created by each module in their respective descriptions.

The data management architecture of the pipeline is shown in Fig. 2. The *Manager* is the top layer (A) of the infrastructure, and it organizes the interplay among the elements of the pipeline and manages execution. The second layer (B) shows the *Modules*: they do not talk directly to one another, but access the data via the *Application Programming Interface* (C). The *Data Location and Format* layer (D) contain the SPOKES data products, including the original input data, user-specified parameters and all data created throughout the pipeline. Below, we discuss each layer in detail.

3.1. Manager

The topmost layer of the infrastructure, the Manager (Fig. 2A), combines and coordinates all components of the pipeline. It is responsible for merging configuration files, placing all data and parameters into the data bank, managing the modules, and executing the pipeline.

The user first sets up the simulated experiment: a master parameter file contains the experiment parameters that will be used by the modules (e.g., Table 1), and a task scheduler contains the modules to be used, the sequence in which they are run, and the compiler and executable that will run the modules. The Manager parses the parameter and scheduler files and executes each module in order. To validate results throughout the pipeline run, the Manager also performs module-specific quality assurance tests (see §3.5).

3.2. Modules

Pipeline operations are broken into discrete *modules* (Fig. 2B), which are executed by the Manager in the order specified by the user. Each module reads the particular data it requires from the databank (Fig. 2D), performs a specific task, and then writes new data to the databank. The modules are independent of one another: the only interaction between them occurs through the databank. Any module can be replaced without disrupting the rest of the pipeline, as long as the data needed by each module is ingested at the beginning of the pipeline (Module 0, §3.2.1) or provided by a preceding module in the pipeline. The following describes the purpose and algorithm of each module (shown in Fig. 1), as implemented to forecast DESpec.

3.2.1. Module 0: Convert

The starting point for the computation is catalog-level galaxy data (originating from a precursor imaging survey) and user-defined parameters (describing the design of the instrument and survey). This module imports the parameters and one or more catalogs into the pipeline's central databank, which is then used as the sole data repository by the rest of the modules. The variables and the format of the databank are described in detail in §3.4, and our specific choices for parameters are described in Table 1.

3.2.2. Module 1: Select Targets

This module selects targets for spectroscopic observation from the photometric catalog of galaxies in the databank using user-defined parameters for color and magnitude cuts. Target selection is performed separately for the large red galaxies (LRGs) and for the emission-line galaxies (ELGs). We assume that all DES filter bands (*grizY*) are available in the catalog photometry, but we do not use the *Y* filter.

For the LRG cuts, the target selection criteria are $z < 22$ and $(r - z) > 1.5$. As shown in Abdalla et al. (2012), these cuts are expected to provide a relatively flat redshift distribution for LRGs over the redshift range $0.5 < z < 1$. For ELGs, the selection criteria are $i < 23.5$; $0.1 < (r - i) < 1.3$; and $-0.2 < (g - r) < 0.3$. This is similar to the ELG selection cuts described in Schlegel et al. (2011).

For the implementation used in this work, the circular field of view has a radius of 1.1 deg and an area of 3.8 deg^2. Four thousand fibers are distributed within a regular hexagon, itself inscribed in the circular field of view. The hexagonal field of view then has an area of 3.14 deg^2. If there are two passes on a given tile (patch of sky; see Module 2), there is then an effective fiber density of $2 * 4000/3.14 \sim 2550/\text{deg}^2$ for each tile. We seek to use the fibers efficiently both by placing them on target galaxies, and by keeping some available for measurements of the sky background and for community projects. To satisfy these constraints, and to dynamically choose the fraction of ELGs and LRGs, we apply a random spatial sampling and elect to keep 100% of ELGs and 18% of LRGs. The output is a set of all galaxies flagged for spectroscopic observation.

3.2.3. Module 2: Tile Survey

This module implements the survey strategy by tiling the instrument field of view across a user-specified sky region, while optimizing observations for simulated environmental and sky conditions. The module has two main functions—the Planner and the Scheduler. They take as inputs the positions of targeted galaxies, as well as parameters of the fiber positioner (e.g., fiber arrangement and tile shape) and of the survey strategy (e.g., exposure time, area and number of observation passes per tile).

The Planner uses a survey mask and a hexagonal tiling pattern to geometrically optimize the survey area coverage: a mask designates the region of sky to cover, and the hexagonal tiles provide closely packed observed fields.

The result is a list of tiles and their celestial coordinates. The Scheduler uses the list from the Planner to select observation dates and times for each tile within the survey area, optimizing observing night efficiency by accounting for sky brightness and airmass. The scheduling is constrained by a set of user-defined parameters—observing dates, the exposure time, an airmass limit and a sky brightness limit. At the beginning of a night, the scheduler first identifies the tiles that are visible. The visible tile that is within the airmass limit and that has the smallest sky brightness is selected for observation. If no tile meets these criteria, then the Scheduler increments the time, and checks again. The scheduler ascends through the visible tiles by their Right Ascension, repeating the tile-selection process.

We model the sky brightness by estimating the moon brightness and visibility, the zodiacal light, and the airglow (and for each location of the moon in the sky during a night). The model does not account for clouds. The seeing is modeled via a Gaussian distribution with a mean of 1.0 and standard deviation of 0.25: this distribution is sampled to provide each targeted observing tile with a seeing value. The seeing distribution does not affect how the pipeline is operated, but it could affect how the pipeline performs: in this implementation of SPOKES, there is an upper limit to the seeing for exposures that may be acquired. This affects the number of galaxies (with their redshifts) that may be used for analysis, which then affects the final cosmological constraints: e.g., if the seeing distribution is skewed high, then fewer exposures may be acquired during the survey, potentially increasing the statistical errors in the cosmological constraints (Module 11).

The tiles are not permitted to overlap in this implementation, and tiling does not account for relative target densities. The latter issue is addressed grossly in the target selection module (Module 1), where random spatial sampling of each target populaion — ELG and LRG — selects for a total target density that correlates with the experiment's fiber density; some fibers may be kept free in order to accomodate community studies and measuring sky backgrounds. The choices made for the survey tiling are closely related to those of the fiber assignment: augmentations of the tiling and fiber assignment algorithms can be implemented, and one should take care to keep separate the two modules or create a new module that combines the two.

This module outputs a list of scheduled tiles for each night during the observing run. The list includes the time of observation, sky brightness, seeing, airmass, celestial coordinates and unique identification number for each tile scheduled for observation. Additionally, we flag galaxies targeted in Module 1 that fall within a tile pointing: this increases computational efficiency for fiber allocation in Module 3, because it sets the only galaxies that will need to be read in for that Module.

3.2.4. Module 3: Allocate Fibers

This module matches fibers to positions in the focal plane of targeted galaxies (see Module 1) for each tile scheduled in the survey (Module 2). This allocation process is constrained by instrument specifications (i.e., fiber patrol radius, fiber arrangement, field shape) of an automated fiber-positioning system (e.g., Mohawk, OzPoz, and Hydra; Saunders et al., 2012; Gillingham et al., 2000; Barden and Armandroff, 1995, respectively). DESpec is designed with a Mohawk positioning system, which employs tilting spines to position the fibers in the focal plane.

The fiber allocation module takes as inputs 1) the sky positions for each tile scheduled by the Survey Strategy module; 2) the positions for the target galaxies produced by the Target Selection module; and 3) all the numbers describing the fibers—fiber diameter, the number of fibers along the diameter of the hexagonal tile, the hexagon radius in degrees, and the fiber patrol radius. The patrol radius describes the distance that a fiber can travel from its rest position on the focal plane. These are the only relevant parameters for the current implementation. DESpec's usage of the Echidna-based Mohawk fiber positioner is described in more detail in Section 4B of Abdalla et al. (2012).

All the fibers begin in a fiducial spatial configuration—arranged in a hexagonal pattern, all equidistant from one another and each with a unique identification number. In this configuration, each target in the sky can be reached at least (most) by three (four) fibers.

The algorithm first chooses the galaxies that are going to be matched to a fiber by prioritizing galaxies according to their local galaxy density: this scheme gives priority to galaxies in crowded regions. We calculate the local galaxy density by estimating the number of target galaxies within one patrol radius, n_p, of each galaxy. For each fiber, we calculate a list of galaxies that can be reached. This list is then ranked in descending order of n_p. For each fiber, the galaxy with the highest n_p is allocated first.

When the algorithm attempts to match fibers to galaxies, the fiber movement paths can intersect. In that scenario, it is not possible for all of the colliding fibers to reach their matched galaxies on the focal plane. In the event of such a fiber collision, the two or more colliding fibers are reset to their positions in the fiducial configuration. The allocation process then begins again, choosing different galaxies to match to fibers, in order to avoid the same collisions. The cycle iterates until the number of fiber collisions cannot be decreased or the number of collisions is zero. When the fiber collisions cannot be decreased, the fiber remains unmatched for this tile.

The output is a list of galaxies, each flagged with a unique identification number of the fiber that will be used to observe the galaxy. The algorithm has been made public[4] and is described in more detail in (Saunders et al., 2014).

3.2.5. Module 4: Throughput

This module calculates the total optical transmission efficiency as a function of wavelength for the principal elements in the light path of the instrument. Before running the pipeline, we estimate separately the throughputs of the main contributors to light loss—optical elements in the telescope barrel, fibers, fiber positioner, and spectrograph—and then ingest them into the pipeline in Module 0.

To model the barrel optics, we employ ZEMAX[5]. For the fiber positioner, the main contribution to throughput loss comes from the effect of the fiber aperture (assumed to be circular). When a Mohawk spine undergoes significant tilt, impending light from the barrel optics is apodized due to focal ratio degradation (FRD), which causes light exiting the fiber to overfill the spectrograph collimator. We model the input to the fiber for a variety of galaxy types as a convolution of four two-dimensional galaxy radial luminosity profiles—Moffat (Beta), Gaussian, deVaucoleur and exponential. The throughput of the fiber aperture is estimated as the fraction of the input beam that is captured by the collimator, assuming the maximum FRD. This is discussed in more detail in Saunders et al. (2012).

For attenuation along the fiber, we use an estimate for a broadband optical fiber that is 50 meters in length and 100 microns in diameter (Abdalla et al., 2012; Marshall et al., 2012). DESpec's lower wavelength limit is 480 nm. Experiments that seek to capture light below $\sim$ 450 nm will contend with additional throughput losses from the fibers and atmosphere for light near the short-wavelength end of the spectrum. First, within the fiber, Rayleigh scattering caused by microscopic variations in the propagating medium's index of refraction will increase attenuation. Second, wavelength-dependent refraction causes atmospheric dispersion, which also depends on the pointing of the telescope; this could could be mitigated with an atmospheric dispersion corrector. Finally, differential refraction can cause the throughput and signal-to-noise to vary across the field (Donnelly et al., 1989). If these issues are not addressed in the survey strategy and in the instrument design, themselves, then the throughput modeling must account for the effects incurred by these processes.

For the spectrograph, the most important components are the dispersive element and the CCD detector. We model the dispersive element as a volume-phase Holographic (VPH) grating. The spectral efficiency of the grating is estimated via VPH GSolver[6], which finds solutions to the general diffraction grating problem for periodic grating structures. In the spectrograph, detectors are based on the Dark Energy Camera (DECam) CCDs, and we use measurements of the CCD throughput as our model (Kubik et al., 2010).

The Throughput module combines the results from the individual elements into the complete transmission efficiency (as a function of wavelength in Angstroms), except for the effect of the atmosphere. While the atmospheric power affects this calculation, it is implemented in Module 6 (§3.2.7) for computational efficiency.

3.2.6. Module 5: Simulate Spectrum

This module constructs models of the intrinsic rest-frame and of the observed-frame spectral energy distributions for each galaxy that has been scheduled for targeting. Each rest-frame spectrum is a linear combination of five *kcorrect* templates (Blanton et al., 2003), which are themselves derived via the non-negative matrix factorization technique (Blanton and Roweis, 2007). The templates are empirically-derived from known galaxy types. A coefficient for each template is derived from the photometry of the galaxy, and it determines the amount of that template's contribution to the total spectrum for a galaxy.

[4] `https://github.com/forero/FiberAllocation/blob/master/text/note.pdf`

[5] `https://www.zemax.com`

[6] `http://www.gsolver.com/`

The choice of template spectra is described in Cunha et al. (2012). The outputs from this module are 1) a rest-frame spectrum and 2) a wavelength-redshifted and flux-dimmed spectrum for each galaxy. The wavelengths are in units of Angstroms, and the fluxes are in units of $ergs\,cm^{-2}s^{-1}\text{Å}^{-1}$.

3.2.7. Module 6: Generate Spectrum Noise

In this module, the transmission throughput and simulated spectra—generated in Module 4 and Module 5, respectively—are used to produce a complete noise spectrum that also includes photon shot noise, spectrograph CCD read noise and noise from the atmosphere (extinction and sky background). Atmospheric absorption comes from the Palomar sky extinction model (from B. Oke and J. Gunn), and the atmospheric emission from optical sky background models from Gemini[7]. The atmospheric emission model is an estimate of the dark optical sky at airmass of 1.0 and 7th day illumination. The atmospheric spectra are adjusted for the airmass of the tile in which the galaxy is observed (set by Module 2, §3.2.3), but not for the illumination or position of the moon. More details of the reconstruction and noise generation can be found in Appendix A2 of Cunha et al. (2012). The output of this module is a noise spectrum for each galaxy.

3.2.8. Module 7: Measure Redshift

This module measures the spectroscopic redshift, $z_{\rm spec}$, of the galaxies from observed spectra. The observed spectra are the combination of the observed-frame spectra and the noise — generated in Modules 5 and 6, respectively.

To measure the redshift of a galaxy, we perform a chi-square minimization between the mock observed galaxy spectrum and a set of model spectra. The chi-square minimization is performed on a grid of redshift values: for each redshift, we linearly optimize the five coefficients of the model spectrum (see §3.2.6), which are constrained to be positive-definite. The best-fit redshift is taken from the redshifted combination of linearly optimized models that best matches the observed spectrum. The output is a list of best-fit redshifts, along with chi-square values (used to judge the quality of fit) for all the observed galaxies.

3.2.9. Module 8: Bin Redshift

This module distributes the galaxies into bins of spectroscopic redshift (measured in Module 7), according to a user-defined parameter for the number of bins—chosen to be five for comparison with the DESpec white paper. The output of this module is a one-dimensional distribution to be used in Module 10 for a tomographic power spectrum analysis.

3.2.10. Module 9: Calculate Selection Function

This module calculates the selection function in space (Right Ascension and Declination) and redshift of the observed spectroscopic galaxy catalog. We use the spectroscopic redshift distribution from Module 8 and the true redshifts of galaxies from the input galaxy catalog.

In order to estimate cosmological parameters, we must use the true galaxy distribution, not the spectroscopically observed one: the true redshifts determine the galaxy positions, and therefore, the theoretically expected correlation functions. Through the relationship between the spectroscopic redshift and the true redshift, the true galaxy distribution can be computed for each spectroscopic redshift bin obtained in Module 8.

The outputs of this module are a redshift selection function—the distribution in true redshift for each of the spectroscopic redshift bins—and the fraction of sky that has been observed completely.

3.2.11. Module 10: Estimation of Cosmological Parameters

The last computational step of the pipeline forecasts the cosmology-constraining power of a given survey configuration by analyzing the catalog of galaxies observed in this pipeline. We perform a power spectrum analysis for cosmological parameter estimation tomographically in redshift.

This function uses a Fisher Matrix analysis of the number density distribution of the galaxies from

[7] Sky spectrum obtained from `http://www.gemini.edu/sciops/ObsProcess/obsConstraints/atm-models/skybg_50_10.dat`

Module 8, in combination with the selection function from Module 9, to estimate constraints on cosmological parameters. The Fisher analysis uses the redshift-binned spherical harmonic power spectrum C_l^{ij}. We include cosmic variance and galaxy shot noise, as well as redshift space distortions, but we neglect galaxy bias. We vary a set of seven ΛCDM parameters: $\{h = 0.7,\ \Omega_m = 0.3,\ \Omega_\Lambda = 0.7,\ w_0 = -0.95,\ w_a = 0.0,\ n_s = 1.0,\ \delta_H = 1843785.96\}$. Details of the calculations are described in Nicola et al. (2014).

3.2.12. Module 11: Report Results

SPOKES achieves provenance by saving all the science data created by the run, as well as a suite of computational diagnostic information. Based on those data, this module generates a report that summarizes the run with figures for assessing the computational and science performance. The report includes basic information about the run (e.g., module versions, the order in which modules are called and a full list of the parameters used); statistics on computational efficiency (memory usage and run times for each module); and statistics on science performance (e.g., the dark energy figure of merit, the redshift distribution, all spectra). The report is a LaTex-typeset PDF document that contains figures, tables and text that can be used for in-depth analysis of a run, and thus for improving subsequent runs of the pipeline.

3.2.13. Cyclicity

The goal of SPOKES's wheel-like framework is to enable informed and swift optimization of the experiment for the chosen metric — e.g., the dark energy figure of merit (FoM) for DESpec. Upon assessment, this information can be used to update the modules, parameters and catalog data to improve the metric results by trial-and-error, or to explore systematically the dependence of the metric on these choices. The user can implement automated running of the pipeline: a wrapper around the Manager could run the pipeline cyclically to traverse a multi-dimensional grid of input parameters to explore the effect of each parameter on the final computational and science outputs.

The large number of parameters may present challenges for optimizing the experiment either by hand or by automated algorithm. There may indeed be multiple configurations that achieve the same success (according to a science metric), and the result may depend on the methods used for optimization. Markov Chain Monte Carlo (MCMC) methods (e.g., Metropolis Hastings) have proven to be useful as techniques for optimization in high-dimensional modeling problems. MCMC's or genetic algorithms could be used to run the pipeline all the way through to a final optimized experiment configuration and a science metric, or they may simply provide information for a decision of how to run the next iteration by hand.

In addition, one of the features of the SPOKES facility is to act as a test-bed for hypotheses of how to design an experiment. That is, even if it is not feasible to completely optimize an experiment to a specified science metric, SPOKES is a facility in which to test experiment configurations—e.g., the interplay between various instruments in the experiment.

3.3. Application Programming Interface

The SPOKES facility includes an Application Programming Interface (API), which has been specifically designed to provide a simple, efficient, robust, and user-friendly interface between the modules and the databank (see Fig. 2B). The API handles the access to the databank and hides the internal workings of the data format (HDF5, see §3.4) from the module developers. We simplify and abstract the data handling to reduce the amount of code used for data access and so reduce possible sources of bugs.

The data handling is performed analogously to a Python dictionary, which has unique 'keys' to access a specific data element: this key is the full path name to the data element or variable in the HDF5 file. In a read operation, the data are read from the databank and returned to the program, and in a write operation, the data are written to the HDF5 path. The creation of new groups, data type handling, overwriting existing paths is done by the API and hidden from the user. Currently, most Python and Numpy[8] data types are supported, except for an actual Python dictionary itself, which is not needed. An example of reading and writing

[8] http://www.numpy.org/

a data element are, respectively, $wavelength\_range = bank[`/Instrument/Spectrograph/wavelength\_range']$ and $bank[`/gal/target\_selection\_flag'] = target\_selection\_flag$, where “bank” is the databank variable.

The API also includes a feature for querying a registered function, rather than accessing static data directly. This is a useful feature for steps that would otherwise generate large amounts of data. For example, each galaxy has a flux spectrum with a flux value for each wavelength at which it is measured: given the resolution and wavelength range of modern spectrometers, this can amount to over 20,000 floating point values per galaxy. Tracking each floating point value through multiple modules for millions of galaxies is untenable. Therefore, we use the registered function feature to generate the galaxy spectra as they are needed rather than storing spectra of all the galaxies. Thus, Modules 5 and 6 register functions in the databank (through the API), which are called later in Module 7 (see Fig. 1). At that stage, a galaxy's intrinsic spectrum is generated, the noise is added and the redshift is measured all in a single step without the information of the galaxy wavelength bins being saved to the databank. This allows information to be passed between modules, without generating large static data, while preserving a strict separation between the modules: the modules only interact with the API and never with each other.

When a function is registered, a Python module is created, along with a path that will be used to call the function: the path for the function is accessed like a path that contains data. The conventions for input, output, and point of usage for the registered function are the same as for all other functions. Both parameters and data can be passed to the registered function, and there is no limit on its complexity. Additionally, the user must take care to ensure that the input is available for the registered function, that it outputs the correct data for the remaining functions, and that the registered function is called at the intended juncture in the pipeline.

3.4. Data Management

A common approach for pipeline development is to employ a linear work flow with data passed directly from one module to the next. This approach has a number of drawbacks. For instance, data generated by a module early in the pipeline needs to be carried through intermediate modules (that possibly do not act on the data) until they reach the module in which the data are used. Such an approach is not efficient and reduces the independence of the modules: i.e. if a later module is modified to use extra data that are produced by an earlier module, all intermediate modules need to be modified as well.

A linear workflow is common in the early development stages of a pipeline, where the frameworks have a tendency to be ad hoc and data transfer between functions and programs occurs by hand—i.e., one programmer gives a result to another, who programs the next function in pipeline. This is not common for more mature studies. For example, the LSST data management system (Jurić et al., 2015) and pipeline have a large-scale framework with a much more complex workflow: the prototype included a ‘clipboard’, with which modules interact for reading and writing data as necessary (Axelrod et al., 2010); this has since been replaced by in-memory Python variables[9]. The SPOKES databank is similar to this ‘clipboard’, keeping all data from a pipeline run to enable maximal provenance.

For the SPOKES facility, we have adopted a centralized system for managing all data, as shown in Fig. 1 and Fig. 2. This scheme separates the data from the modules, so that a given module accesses only the data it needs and creates. As well as maintaining modularity, such a centralized scheme more easily allows for provenance and reproducibility.

For this central databank model to work, the data formatting must be able to handle many data types, scale efficiently to handle large amounts of data and be flexible enough to store all data for a rapidly developed pipeline. After evaluating a number of possible standards — including Flexible Image Transport System (FITS)[10] and relational databases - we adopted a solution based on the Hierarchical Data Format (HDF5)[11]. The detailed justification of this choice is given in §Appendix B.

[9] `https://docushare.lsstcorp.org/docushare/dsweb/ServicesLib/LDM-152/History`

[10] http://fits.gsfc.nasa.gov/

[11] http://www.hdfgroup.org/HDF5

In an HDF5 file, the data are organized in unique paths, like a hard disk filesystems—e.g., */group/subgroup/dataset*: each data set resides in a 'group' and its 'subgroup', which are named descriptively in SPOKES to associate related data and improve code readability. The data sets can be of a variety of data types, including arrays.

To take advantage of this organizational paradigm, our API (see §3.3) provides trivial access to any field or data group via this path and this path alone, regardless of data type. It provides modularity of data access: a module may use individual aspects of a data group without having to read in all data. For example, a module can access galaxy identification numbers and positions without reading all the other data that another module might need. We describe the data groups later in this section.

The databank can be stored in any number of HDF5 files, as suitable to the application, and there is no intrinsic limit to the file size. The maximum data file size is dictated merely by the available working memory and machine hard disk storage. When HDF5 files are sufficiently large, the file is split on the disk into multiple files that are logically merged transparently. SPOKES supports parallel reading of HDF5 files (e.g., multiple processes in a batch job can access simultaneously), but not parallel writing.

To estimate the data storage requirement we find that about 1 GB is needed to store a complete databank with $\sim$ 10M input galaxies. However, after photometric selection and survey and fiber allocation operations, only about 10% of these will be targets (see Table 2), for which redshifts will be measured: this gives 1GB for 1M redshifts. If a Stage IV Dark Energy experiment aims to measure redshifts for up to 20M targets, it will require 20GB for the total catalog, or 2GB for the targets alone.

The data groups within the databank are partitioned according to module usage and related information; these data groups are shown in the legend in Fig. 1. All original data from input, including parameters (e.g., telescope optics choices) and data (e.g., galaxies), are converted to the native data format and separated into M data sets within the databank upon initialization of the pipeline, as shown in the base data layer of the architecture in Fig. 2D. The data groups (in alphabetical order) are described below

Analysis Choices (A) contains the information with which to specify the analysis methods—e.g, magnitude or color cuts for Target Selection (Module 1) and bin size for redshift binning (Module 8).

Constants (C) holds physical constants and random seeds.

Environment (E) contains the information regarding the atmosphere (absorption and emission spectra) and location (e.g., elevation) at which the observations are taking place.

Fibers (F) contains information (e.g., location in focal plane) about the fibers that are assigned to galaxies.

Galaxies (G) contains all galaxy data.

Instrument (I) contains several subgroups representing the subsystems of the instrument—optics, fibers and spectrograph—each of which has several parameters.

Ensemble (N) contains data on the galaxies as a collection—the redshift histogram, related cosmological constraints, etc.

Run-time Parameters (R) are those which determine how the simulation will be run — e.g., with or without parallel processing, or which simulation files are to be used.

Spectral Templates (S) contain the eigentemplates used to reconstruct galaxy spectra.

Survey Tiles (T) contains a set of tile information (sky position, airmass, time of observation, etc) and is used to link galaxies with the time and observation environment in which they were observed.

Survey Parameters (U) holds the data necessary to run the survey, for example, exposure time per tile and region of the sky to be observed.

Table 1: Selected baseline SPOKES input parameters. These parameters have been chosen to match those proposed for DESpec.

Data Group	Used by Module(s)	Parameter	Value
Analysis (*A*): Target Selection			
LRG cuts	1	z	< 22
	1	$(r - z)$	> 1.5
ELG cuts	1	i	< 23.5
	1	$(r - i)$	> 0.1 and < 1.3
	1	$(g - r)$	> -0.2 and < 0.3
Environment (*E*): Atmosphere[**]			
	6	Sky Background	Gemini Sky Models
	6	Atmospheric Extinction	Palomar Extinction Curves
Instrument (*I*): Fibers			
	3	Number of Fibers	4000
	2,3	Fiber Arrangement	Hexagon
	2,3	FOV: Hexagon radius	1.1 deg
	2,3	FOV: Hexagon area	3.14 deg^2
Instrument (*I*): Telescope			
	4	Diameter	4m
	4	Optical Efficiency[*]	~ 0.25
Instrument (*I*): Spectrograph			
	6	Read Noise	5 e-
	6,7	Wavelengths	[480, 1050] nm
Survey Plan (*U*)			
	2	Exp. time	1200 s
	2,9,10	Area	5000 and 15000 deg^2
	2,3	Passes per Tile	2

The input variables to the pipeline and the values used in the demonstration run of the SPOKES pipeline. Data groups and module numbers coincide with those of Fig. 1. The 28 parametric specifications shown here are necessary for running a spectroscopic experiment. The table shows two types of target selection, “LRG cuts” and “ELG cuts,” designed to preferentially target Luminous Red Galaxies and Emission-Line Galaxies, respectively.

[*] This value is the mean of the throughput spectrum. See 3.2.5 for details of the throughput calculation.
[**] see 3.2.6 for details
N.B: The ‘Convert’ module is not listed here, because all of the input parameters are ingested and placed into the databank.

3.5. Quality Assurance

An important aspect of building a modular end-to-end simulation facility is to develop an integrated quality assurance (QA) framework that will automatically detect possible problems with a given simulation run or with an update to one of the modules. In SPOKES, we are working towards building a continuous integration process for performing tests at the unit and facility level. At present, during pipeline runs, we have three discrete levels of tests and cross-checks for QA.

QA first performs a series of basic logical cross-checks at the onset of each module. For example, after fibers have been allocated to galaxies, the same function checks that each galaxy has received some value flagging it as observed or not. It also checks that this value is within the range of available fiber indices.

At the intermediate stage, there are a set of user-defined limits for specific properties of the analysis, modules and galaxy catalog. When these limits are approached or exceeded, the pipeline reports a warning.

Finally, the pipeline provides scientific diagnostic figures and plots to check fidelity at each step: e.g., Tile Survey (Module 2) produces a map of the fields observed. Fig. 3 shows the spectral flux as a function of wavelength for a single galaxy, with and without noise, respectively. This pair of figures is produced automatically for a randomly selected set of galaxies (the number of which is set by the user) during each pipeline run. The user can then review the spectra to verify pipeline accuracy at this juncture. This information is wrapped into an automatically generated final report (by Module 11), with which the user can make assessments of the runs and input. Another important aspect of QA is that we store sufficient information to ensure provenance, so that each run can be reproduced at a later time.

4. Results

4.1. Survey Baseline

A large number of spectroscopic experiments are being planned or are underway. To illustrate the SPOKES facility, we have decided to focus on DESpec (Abdalla et al., 2012), an early concept that predates the upcoming DESI experiment and is representative of next-generation spectroscopic surveys. DESpec was a survey instrument designed for the Blanco 4-meter telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile. Over the course of 1,000 nights, the planned survey would observe more than 20 million galaxies and QSO's over 15,000 deg^2 of sky, with two passes for each patch of sky. The data set would provide a three-dimensional map of the universe out to redshift ~ 2. The DESpec instrument design is primarily comprised of 10 spectrographs and an automated fiber positioning system, Mohawk (Saunders et al., 2012), that can target the galaxies from a precursor imaging survey, such as DES or LSST. Table 1 contains a summary of the DESpec instrument and survey inputs that we consider as the baseline for our calculations.

The DESpec design has a hexagon-shaped fiber plane with 4000 fibers of diameter 1.75 arcsec (or 100 μm) and pitch and patrol radius of ~ 6mm. The spectrograph has a read noise of 1-3 e- (assuming a gain of unity) and a wavelength range of 350 to 1050 nm. Further details of the spectrograph design can be found in the DESpec white paper (Abdalla et al., 2012). We model the telescope optical efficiency based on ZEMAX modeling of the Blanco optics: the average throughput across the wavelength range is ~ 0.25 (see §3.2.5 for details of the throughput model).

4.2. Pipeline Inputs

In addition to the DESpec parameters summarized in Table 1, one of the key SPOKES inputs is a galaxy catalog. For the results presented here we use a mock galaxy catalog (described in detail in §Appendix A), and we select experiment parameters that correspond with the DESpec design concept described above. For this study, we have used the mock galaxy catalogs based on the algorithm Adding Density Determined GAlaxies to Lightcone Simulations (ADDGALS: Wechsler et al 2014 – in prep., Busha et al 2014 – in prep.). This algorithm attaches synthetic galaxies, including multiband photometry, to dark matter particles in a lightcone output from a dark matter N-body simulation and is designed to match the luminosities, colors, and clustering properties of galaxies. The catalog used here was based on a single 'Carmen' simulation run as part of the LasDamas simulations (McBride et al., 2009)[12]. This simulation modeled a flat ΛCDM universe with $\Omega_m = 0.25$ and $\sigma_8 = 0.8$ in a 1 Gpc/h box with 1120^3 particles. A 220-sq.-deg. light cone extending out to z = 1.33 was created by pasting together 40 snapshot outputs. For more details on the catalog see §Appendix A.

These simulations do not contain some classes of objects that would contaminate the target selection—e.g., QSOs, low-mass stars, high-redshift galaxies. Surveys, like the Galaxy and Mass Assembly (GAMA)[13] spectroscopic survey, use a combination of optical and infra-red photometry to create cleaner samples (Baldry et al., 2010, Fig. 6). The lack of contaminants will make the final results for this implementation of SPOKES better than a real survey. This work uses the same simulations as the DESpec white paper, which allows for direct comparison with that work. In this paper, we seek to compare to the DESpec white paper to show that such an analysis can be performed more efficiently with the SPOKES framework. Nevertheless, the purpose of the SPOKES framework is to be general enough to allow for the import of catalog data that has these contaminants: to take advantage of an improved set of simulation data, the user would be required to up-

[12] Further details regarding the simulations can be found at http://lss.phy.vanderbilt.edu/lasdamas/simulations.html

[13] `http://www.gama-survey.org/`

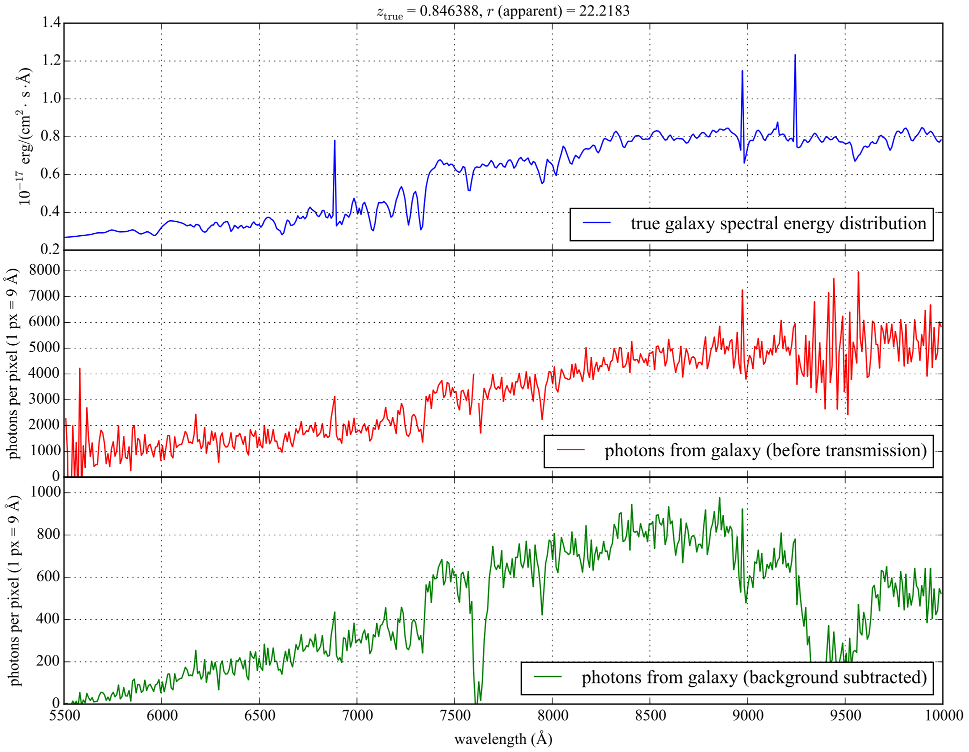

Figure 3: Example of galaxy spectrum simulated by SPOKES. Top panel: a noise-free galaxy spectrum reconstructed from a set of templates for a galaxy at a redshift of $z = 0.93$. Middle panel: galaxy spectrum with noise computed from multiple sources—including Poisson noise from photon counts, CCD readout noise and atmospheric sources—for an exposure time of 1200 seconds. Bottom panel: galaxy spectrum with noise, but before being transmitted through the atmosphere and telescope. More details of the spectrum and noise generation process can be found in §3.2.6 and §3.2.7

date the Target Selection module to work with this data.

The other key inputs to SPOKES, such as atmospheric data on the sky background and extinction, are taken from Gemini optical sky background models[14] and Palomar extinction curves[15]; these are critical for constructing the noise model for each galaxy spectrum (for details see §3.2.6 and Cunha et al., 2012)

4.3. *Science Performance*

To assess the science performance of the pipeline, we consider a typical SPOKES simulation run. Table 2 traces the progress of the galaxy sample in one of the mock catalog tiles through the pipeline steps.

The mock catalogs are stored in HEALPix (Górski et al., 2002) cells, with $N_{side} = 8$, which corresponds to 53.7 deg^2 per cell. The initial mock catalog (ingested by Module 0) contains about 1.3×10^5 galaxies per square degree over about eight Healpix cells. After magnitude and color cuts are applied, as part of the target select/ion step (Module 1), 4.8×10^3 galaxies per square degree remain. The vast majority of galaxies are removed by the magnitude cut.

Once the tiling strategy has been defined (Module 2), the next key step for selecting galaxies is the assignment of fibers in each tile (Module 3). Using the fiber allocation scheme described in §3.2.4,

[14] derived from `http://www.gemini.edu/sciops/telescopes-and-sites/observing-condition-constraints/optical-sky-background`

[15] `http://www.ing.iac.es/Astronomy/observing/manuals/html_manuals/tech_notes/tn065-100/small/palomar.tab`

we find that the density of galaxies assigned a fiber is 1750/deg^2. This can be compared with the maximal possible fiber allocation density, 2550/deg^2 calculated in §3.2.2. This makes available ~ 800 fiber per square degree (or ~ 1250 fibers per tile pass) for measurements of the sky background or for community projects.

For each galaxy assigned to a fiber, we simulate an intrinsic noise-free rest-frame spectrum (Module 5), and a spectrum with noise from atmosphere and electronics (Module 6) and with the signal reduced by the telescope throughput model (Module 4). A typical spectrum (with noise, without noise and before transmission through the atmosphere and optics) is shown in Fig. 3. The true or intrinsic galaxy spectrum is shown in the top panel: note the prominent emission lines at 3727Å (OII), 4861Å (Hβ) and 5007Å (OIII). In the noisy spectrum (bottom panel), the prominence of these lines is degraded significantly—primarily by photon noise, the dominant component of the overall noise. The effect of the statistical noise is seen clearly at the red end of the spectrum, where larger errors are incurred due to the wavelength dependence of the error. Finally, after the sky background has been subtracted and the spectrum has been transmitted through the atmosphere and telescope (middle panel), telluric features at ~ 6900Å (O_2), ~ 7600Å (O_2) and ~ 9350Å (H_2O) further degrade galaxy spectral features. The telluric features could be removed with flux-calibrated standard stars. Telluric absorption features undergo variation on time scales on the order of minutes, and they vary with the airmass of the observation. This would require reserving a fiber for allocation of a standard star close in time and airmass of an observed field. Alternatively, methods have recently been developed to model telluric line spectra (e.g., Seifahrt et al., 2010; Rudolf et al., 2015). In either case, the process of removing features from each galaxy spectrum must be automated so as not to require any human oversight.

Spectroscopic redshifts are then measured in Module 7. At this stage, we judge the quality of the each of the fits using χ^2 tests and reject galaxies with poor fits. From Table 2, we see that this step removes about ~ 7% of the galaxies. The only source of error in the spectra are related to photon counts (i.e., the error is statistical, aside from systematic background flux sources and signal-to-noise reduction due to transmission): there are no systematic errors in wavelength, so no emission or absorption lines are systematically incorrect.

Table 2: Galaxy statistics in a typical SPOKES run

module	type	gals/deg^2
0: convert	all	1.3×10^5
1: select target	LRG	3.7×10^3
	ELG	1.1×10^3
	all	4.8×10^3
3: allocate fiber	LRG	6.5×10^2
	ELG	1.1×10^3
	all	1.7×10^3
7: measure z	LRG	5.9×10^2
	ELG	1.0×10^3
	all	1.59×10^3

The resulting observed redshifts can be compared directly to the true redshifts coming from the mock catalog, as shown in Fig. 4. The figure shows that we are able to recover the true input redshifts with small scatter. Note that the galaxies rejected by the χ^2 tests do not appear in the comparison between spectroscopic and true redshifts. Different redshift measurement algorithms may obtain different success rates. We also see that the underlying galaxies are not distributed evenly in redshift. For instance, a gap around redshift of z=0.5 is prominent. The redshift distribution of galaxies in our sample is driven by the cuts that we have made. In particular, we find that the $(r - z)$ cut tends to select for galaxies at higher redshifts – as was intended – resulting in the distribution shown in Fig. 5. This figure shows the distribution of spectroscopic redshifts, $n(z_{\rm spec})$, across bins measured in Module 8, and that our galaxy sample peaks at a redshift $z \sim 1$ and extends out to redshifts of $z \sim 2$.

From the redshift distribution, we then measure a selection function, $W(z_{\rm spec}|z_{\rm true})$ to obtain a distribution of true redshifts for each spectroscopic redshift bin (Module 9). This is then used with $n(z_{\rm spec})$ in the Fisher matrix forecaster to compute confidence contours in the $w_0 - w_a$ plane. Constraints on dark energy equation of state parameters w_0 and w_a are calculated in Module 10 (see §3.2.11) and the results, which include WMAP9 priors, are shown in Fig. 6. The de-

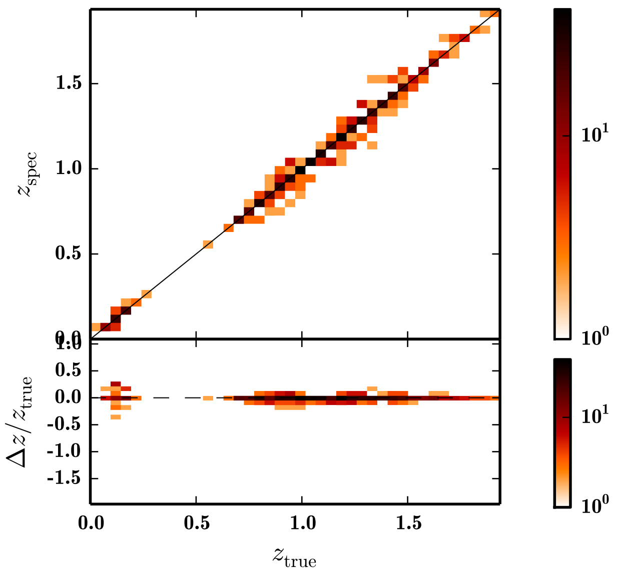

Figure 4: Comparison of true redshifts and the measured spectroscopic redshifts. The upper panel shows a two-dimensional histogram of $z_{\rm spec}$ vs. $z_{\rm true}$, while the lower panel shows the fractional difference between redshift measures as a function of $z_{\rm true}$. In both cases, the solid black line corresponds to perfect recovery. The method used in SPOKES to measure spectroscopic redshifts is described in §3.2.8. Note that only galaxies that meet the χ^2 goodness-of-fit cut are included here.

rived precision on these parameters is consistent with that forecasted in the DESpec white paper (Abdalla et al., 2012).

In forecasting DESpec, both the parameters of the experiment and galaxy catalogs remained fixed. Through multiple runs of the pipeline, on a trial-and-error basis, we continuously improved the modules. In particular, the computational efficiency of Modules 2, 3 and 7 were increased dramatically. In future usage of the pipeline, we will seek to improve the cosmological constraints through modification of the experiment parameters.

In addition to dark energy, one could study properties of (groups of) galaxies, such as the luminosity function. This could replace the dark energy FoM as the metric for experiment optimization, or it could be added to the current implementation and used in tandem. In either case, one would write a new module to measure the galaxy properties of interest, and instruct the Manager to call the module at the appropriate point in the pipeline. To develop this enhancement, the user is required to develop new modules that incorporate galaxy properties. This paper focuses on an implementation to forecast the FoM, where the luminosity function is outside the scope. However, the SPOKES framework is agnostic to the choice of metric for the experiment, and the choice does not limit the framework at all.

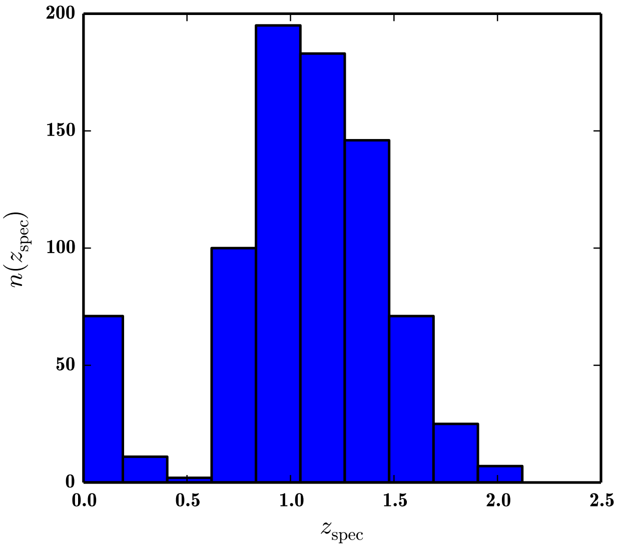

Figure 5: Redshift distribution dn/dz for the recovered spectroscopic sample. This is generated in Module 8 (see §3.2.9).

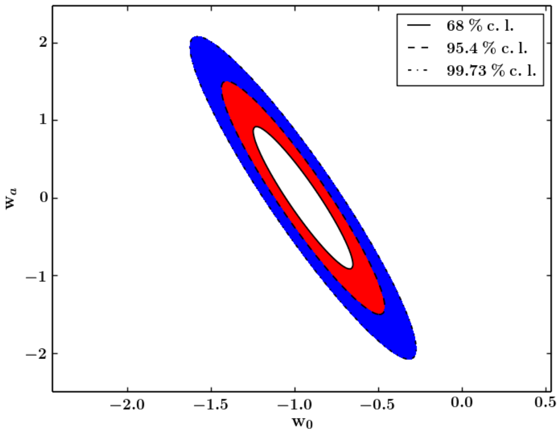

Figure 6: Confidence contours (1, 2 and 3σ for white red and blue, respectively) for a joint estimate of dark energy parameters, w_0 and w_a. See §3.2.11 for details on the calculation method. The results shown here include a WMAP9 prior that we construct from the publicly available MCMC chains.

4.4. Computational Performance

To evaluate the computational performance of the pipeline, we ran several benchmarks on a laptop with an Intel i7 mobile processor with a clock rate of 2.2 Ghz and 8 GB memory. Some of the modules, such as Module 2 (Tile Survey) and Module 7 (Measure Redshift), were pararellised using eight threads on four cores.

Fig. 7 shows the CPU time used by each module for different numbers of input galaxies. As shown on the figure, a large fraction of the time is spent on the survey tiling, fiber allocation, and redshift measurement—Modules 2, 3 and 7, respectively. The latter is explained by the higher complexity of the redshift measurement process. The former two are more surprising, and arises from the more detailed physics included in these modules relative to others (e.g. weather modeling). These modules also have functions that are not yet fully optimized (e.g. tiling, calculating fiber collisions, matching fibers to galaxies). For example, the fiber allocation in Module 3 measures the distance between each fiber and all galaxies multiple times: this is the principal time-consuming process. When the count of galaxies per tile increases in this time-scaling test, the fiber allocation module scales sub-linearly. If there are more tiles, but the same number of galaxies per tile, then it will scale linearly.

Further optimisation reduces the CPU times of several of the modules. The parallelisation scaling is seen in the dependence of the CPU times on the number of galaxies. It is close to linear for Module 3 (Allocate Fibers) and Module 7 (Measure Redshift). Note also that for all modules, execution time is dominated by CPU time, while I/O time is negligible with our implementation.

Fig. 8 shows the memory usage of each module. The highest memory consumption takes place in Module 7 (Measure Redshift). The memory scales as a function of number of parallel processes and galaxies since this governs the total volume of data that is being worked on at any given time.

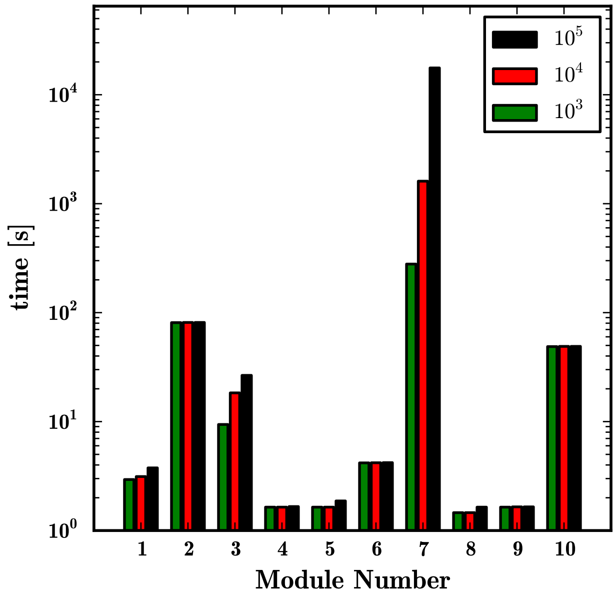

Figure 7: The time used by each module for 1k, 10k and 100k input galaxies.

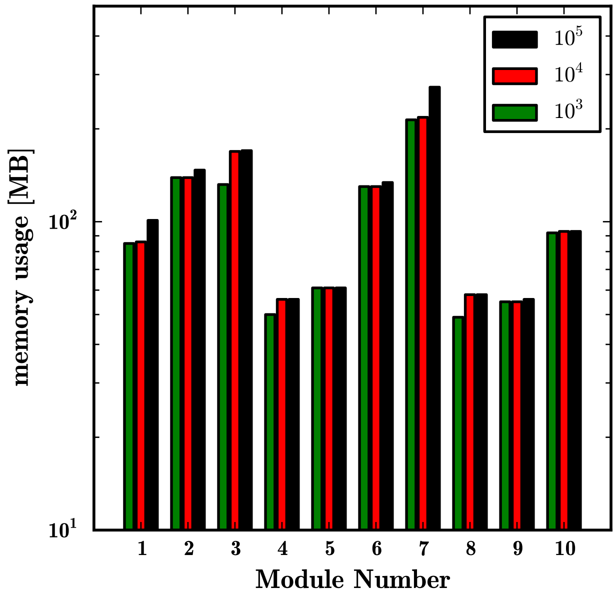

Figure 8: The memory used by each module for 1k, 10k and 100k input galaxies.

5. Conclusions

Modern cosmology experiments have become sufficiently precise and complex that new methods are required to perform accurate feasibility studies and to perform survey optimization. We have described the SPOKES simulation facility, which is designed to meet the requirements for future cosmology spectroscopic surveys. The end-to-end architecture of SPOKES is both integrated and flexible, and it allows for the reproducibility and modularity needed to develop, validate and exploit future experiments.

We have demonstrated the completeness, speed and flexibility of the SPOKES simulation pipeline. While this was done using the DESpec experiment concept as the baseline for this purpose, the pipeline is fully general and can be applied to any other wide-field spectroscopic cosmological experiments. We showed that the pipeline results are consistent with earlier calculations of the forecast for the science performance of DESpec. We showed how the SPOKES framework provides a general tool for detailed studies of systematics, performance evaluations and development of the data processing pipeline.

In the future, we plan to further develop SPOKES. In particular, we plan to implement input parameters for spectroscopic experiments other than DESpec, to further optimize the modules to increase performance and parallelisation scaling, and to refine some of the modules with more physics and more advanced data analysis schemes. In addition, we intend to make SPOKES publicly available.

Acknowledgments

The authors would like to thank Brenna Flaugher, Tom Diehl, Michael Meyer, Stephen Bailey, Richard Kron and Simon Lilly for useful discussions during the development of this work.

References

Abdalla, F., Annis, J., Bacon, D., Bridle, S., Castander, F., Colless, M., DePoy, D., Diehl, H. T., Eriksen, M., Flaugher, B., Frieman, J., Gaztanaga, E., Hogan, C., Jouvel, S., Kent, S., Kirk, D., Kron, R., Kuhlmann, S., Lahav, O., Lawrence, J., Lin, H., Marriner, J., Marshall, J., Mohr, J., Nichol, R. C., Sako, M., Saunders, W., Soares-Santos, M., Thomas, D., Wechsler, R., West, A., Wu, H., Sep. 2012. The Dark Energy Spectrometer (DESpec): A Multi-Fiber Spectroscopic Upgrade of the Dark Energy Camera and Survey for the Blanco Telescope. arXiv:1209.2451.

Adams, J. J., Blanc, G. A., Hill, G. J., Gebhardt, K., Drory, N., Hao, L., Bender, R., Byun, J., Ciardullo, R., Cornell, M. E., Finkelstein, S. L., Fry, A., Gawiser, E., Gronwall, C., Hopp, U., Jeong, D., Kelz, A., Kelzenberg, R., Komatsu, E., MacQueen, P. J., Murphy, J., Odoms, P. S., Roth, M., Schneider, D. P., Tufts, J. R., Wilkinson, C. P., Dec. 2010. The Hetdex Pilot Survey. I. Survey Design, Performance, And Catalog Of Emission-line Galaxies. The Astrophysical Journal Supplement Series 192 (1), 5.

Albrecht, A., Bernstein, G., Cahn, R., Freedman, W. L., Hewitt, J., Hu, W., Huth, J., Kamionkowski, M., Kolb, E. W., Knox, L., Mather, J. C., Staggs, S., Suntzeff, N. B., Sep. 2006. Report of the Dark Energy Task Force. arXiv.org, 9591.

Axelrod, T., Kantor, J., Lupton, R. H., Pierfederici, F., 2010. An open source application framework for astronomical imaging pipelines.
URL `http://dx.doi.org/10.1117/12.857297`

Baldry, I. K., Robotham, A. S. G., Hill, D. T., Driver, S. P., Liske, J., Norberg, P., Bamford, S. P., Hopkins, A. M., Loveday, J., Peacock, J. A., Cameron, E., Croom, S. M., Cross, N. J. G., Doyle, I. F., Dye, S., Frenk, C. S., Jones, D. H., van Kampen, E., Kelvin, L. S., Nichol, R. C., Parkinson, H. R., Popescu, C. C., Prescott, M., Sharp, R. G., Sutherland, W. J., Thomas, D., Tuffs, R. J., May 2010. Galaxy And Mass Assembly (GAMA): the input catalogue and star-galaxy separation. MNRAS404, 86–100.

Barden, S. C., Armandroff, T., Jun. 1995. Performance of the WIYN fiber-fed MOS system: Hydra. In: Barden, S. C. (Ed.), SPIE's 1995 Symposium on OE/Aerospace Sensing and Dual Use Photonics. SPIE, pp. 56–67.

Bergé, J., Gamper, L., Réfrégier, A., Amara, A., Feb. 2013. An Ultra Fast Image Generator (UFIG) for wide-field astronomy. Astronomy and Computing 1, 23–32.

Bernstein, J. P., Kessler, R., Kuhlmann, S., Biswas, R., Kovacs, E., Aldering, G., Crane, I., D'Andrea, C. B., Finley, D. A., Frieman, J. A., Hufford, T., Jarvis, M. J., Kim, A. G., Marriner, J., Mukherjee, P., Nichol, R. C., Nugent, P., Parkinson, D., Reis, R. R. R., Sako, M., Spinka, H., Sullivan, M., Jul. 2012. Supernova Simulations and Strategies for the Dark Energy Survey. ApJ753, 152.

Blanton, M. R., Dalcanton, J., Eisenstein, D., Loveday, J., Strauss, M. A., SubbaRao, M., Weinberg, D. H., Anderson, Jr., J. E., Annis, J., Bahcall, N. A., Bernardi, M., Brinkmann, J., Brunner, R. J., Burles, S., Carey, L., Castander, F. J., Connolly, A. J., Csabai, I., Doi, M., Finkbeiner, D., Friedman, S., Frieman, J. A., Fukugita, M., Gunn, J. E., Hennessy, G. S., Hindsley, R. B., Hogg, D. W., Ichikawa, T., Ivezić, Ž., Kent, S., Knapp, G. R., Lamb, D. Q., Leger, R. F., Long, D. C., Lupton, R. H., McKay, T. A., Meiksin, A., Merelli, A., Munn, J. A., Narayanan, V., Newcomb, M., Nichol, R. C., Okamura, S., Owen, R., Pier, J. R., Pope, A., Postman, M., Quinn, T., Rockosi, C. M., Schlegel, D. J., Schneider, D. P., Shimasaku, K., Siegmund, W. A., Smee, S., Snir, Y., Stoughton, C., Stubbs, C., Szalay, A. S., Szokoly, G. P., Thakar, A. R., Tremonti, C., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Yanny, B., Yasuda, N., York, D. G., May 2001. The Luminosity Function of Galaxies in SDSS Commissioning Data. AJ121, 2358–2380.

Blanton, M. R., Hogg, D. W., Bahcall, N. A., Brinkmann, J., Britton, M., Connolly, A. J., Csabai, I., Fukugita, M., Loveday, J., Meiksin, A., Munn, J. A., Nichol, R. C., Okamura, S., Quinn, T., Schneider, D. P., Shimasaku, K., Strauss, M. A., Tegmark, M., Vogeley, M. S., Weinberg, D. H., Aug. 2003. The Galaxy Luminosity Function and Luminosity Density at Redshift z = 0.1. The Astrophysical Journal 592 (2), 819–838.

Blanton, M. R., Lin, H., Lupton, R. H., Maley, F. M., Young, N., Zehavi, I., Loveday, J., Apr. 2003. An Efficient Targeting Strategy for Multiobject Spectrograph Surveys: the Sloan Digital Sky Survey “Tiling” Algorithm. AJ125, 2276–2286.

Blanton, M. R., Roweis, S., Feb. 2007. K-Corrections and Filter Transformations in the Ultraviolet, Optical, and Near-Infrared. The Astronomical Journal 133 (2), 734–754.

Boller, T., Dwelly, T., Sep. 2012. The 4MOST facility simulator: instrument and science optimisation. SPIE 8448.

Campbell, L., Saunders, W., Colless, M., Jun. 2004. The tiling algorithm for the 6dF Galaxy Survey. Monthly Notices of the Royal Astronomical Society 350 (4), 1467–1476.

Chang, C., Busha, M. T., Wechsler, R. H., Refregier, A., Amara, A., Rykoff, E., Becker, M. R., Bruderer, C., Gamper, L., Leistedt, B., Peiris, H., Abbott, T., Abdalla, F. B., Banerji, M., Bernstein, R. A., Bertin, E., Brooks, D., Carnero Rosell, A., Desai, S., da Costa, L. N., Cunha, C. E., Eifler, T., Evrard, A. E., Fausti Neto, A., Gerdes, D., Gruen, D., James, D., Kuehn, K., Maia, M. A. G., Makler, M., Ogando, R., Plazas, A., Sanchez, E., Schubnell, M., Sevilla-Noarbe, I., Smith, C., Soares-Santos, M., Suchyta, E., Swanson, M. E. C., Tarle, G., Zuntz, J., Oct. 2014. Modelling the Transfer Function for the Dark Energy Survey. ArXiv e-prints.

Chang, C., Kahn, S. M., Jernigan, J. G., Peterson, J. R., AlSayyad, Y., Ahmad, Z., Bankert, J., Bard, D., Connolly, A., Gibson, R. R., Gilmore, K., Grace, E., Hannel, M., Hodge, M. A., Jee, M. J., Jones, L., Krughoff, S., Lorenz, S., Marshall, P. J., Marshall, S., Meert, A., Nagarajan, S., Peng, E., Rasmussen, A. P., Shmakova, M., Sylvestre, N., Todd, N., Young, M., Jun. 2012. Spurious Shear in Weak Lensing with LSST. arXiv:1206.1378, to appear in MNRAS.

Claver, C. F., Chandrasekharan, S., Liang, M., Xin, B., Alagoz, E., Arndt, K., Shipsey, I. P., Sep. 2012. Prototype pipeline for LSST wavefront sensing and reconstruction. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 8444 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 4.

Connolly, A. J., Peterson, J., Jernigan, J. G., Abel, R., Bankert, J., Chang, C., Claver, C. F., Gibson, R., Gilmore, D. K., Grace, E., Jones, R. L., Ivezić, Ž., Jee, J., Juric, M., Kahn, S. M., Krabbendam, V. L., Krughoff, S., Lorenz, S., Pizagno, J., Rasmussen, A., Todd, N., Tyson, J. A., Young, M., Jul. 2010. Simulating the LSST system. In: Angeli, G. Z., Dierickx, P. (Eds.), SPIE Astronomical Telescopes and Instrumentation: Observational Frontiers of Astronomy for the New Decade. SPIE, pp. 77381O–77381O–10.

Conroy, C., Wechsler, R. H., Kravtsov, A. V., Oct. 2007. The Hierarchical Build-Up of Massive Galaxies and the Intracluster Light since z = 1. ApJ668, 826–838.

Cunha, C. E., Huterer, D., Lin, H., Busha, M. T., Wechsler, R. H., Jul. 2012. Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. arXiv1207.3347.

de Jong, R. S., Bellido-Tirado, O., Chiappini, C., Depagne, É., Haynes, R., Johl, D., Schnurr, O., Schwope, A., Walcher, J., Dionies, F., Haynes, D., Kelz, A., Kitaura, F. S., Lamer, G., Minchev, I., Müller, V., Nuza, S. E., Olaya, J.-C., Piffl, T., Popow, E., Steinmetz, M., Ural, U., Williams, M., Winkler, R., Wisotzki, L., Ansorge, W. R., Banerji, M., Gonzalez Solares, E., Irwin, M., Kennicutt, R. C., King, D., McMahon, R. G., Koposov, S., Parry, I. R., Sun, D., Walton, N. A., Finger, G., Iwert, O., Krumpe, M., Lizon, J.-L., Vincenzo, M., Amans, J.-P., Bonifacio, P., Cohen, M., Francois, P., Jagourel, P., Mignot, S. B., Royer, F., Sartoretti, P., Bender, R., Grupp, F., Hess, H.-J., Lang-Bardl, F., Muschielok, B., Böhringer, H., Boller, T., Bongiorno, A., Brusa, M., Dwelly, T., Merloni, A., Nandra, K., Salvato, M., Pragt, J. H., Navarro, R., Gerlofsma, G., Roelfsema, R., Dalton, G. B., Middleton, K. F., Tosh, I. A., Boeche, C., Caffau, E., Christlieb, N., Grebel, E. K., Hansen, C., Koch, A., Ludwig, H.-G., Quirrenbach, A., Sbordone, L., Seifert, W., Thimm, G., Trifonov, T., Helmi, A., Trager, S. C., Feltzing, S., Korn, A., Boland, W., Sep. 2012. 4MOST: 4-metre multi-object spectroscopic telescope. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 8446 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 0.

Delgado, F., Cook, K., Miller, M., Allsman, R., Pierfederici, F., Jul. 2006. LSST operation simulator implementation. Observatory Operations: Strategies 6270, 45.

Donnelly, R. H., Brodie, J. P., Bixler, J. V., Hailey, C. J., Nov. 1989. The implications of atmospheric effects for fiber-fed spectroscopy. PASP101, 1046–1054.

Drinkwater, M. J., Jurek, R. J., Blake, C., Woods, D., Pimbblet, K. A., Glazebrook, K., Sharp, R., Pracy, M. B., Brough, S., Colless, M., Couch, W. J., Croom, S. M., Davis, T. M., Forbes, D., Forster, K., Gilbank, D. G., Gladders, M., Jelliffe, B., Jones, N., Li, I.-h., Madore, B., Martin, D. C., Poole, G. B., Small, T., Wisnioski, E., Wyder, T., Yee, H. K. C., Jan. 2010. The WiggleZ Dark Energy Survey: survey design and first data release. Monthly Notices of the Royal Astronomical Society 401 (3), 1429–1452.

Fan, X., May 1999. Simulation of Stellar Objects in SDSS Color Space. AJ117, 2528–2551.

Flaugher, B. L., et al., Sep. 2012. Status of the Dark Energy Survey Camera (DECam) project. In: Ground-based and Airborne Instrumentation for Astronomy IV. Proceedings of the SPIE. Fermi National Accelerator Lab. (United States), p. 11.

Gibson, R. R., Ahmad, Z., Bankert, J., Bard, D., Connolly, A. J., Chang, C., Gilmore, K., Grace, E., Hannel, M., Jernigan, J. G., Jones, L., Kahn, S. M., Krughoff, K. S., Lorenz, S., Marshall, S., Nagarajan, S., Peterson, J. R., Pizagno, J., Rasmussen, A. P., Shmakova, M., Silvestri, N., Todd, N., Young, M., Jul. 2011. A Framework for End to End Simulations of the Large Synoptic Survey Telescope. In: Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings. p. 329.

Gillingham, P. R., Miziarski, S., Klauser, U., Aug. 2000. Mechanical features of the OzPoz fiber positioner for the VLT. In: Iye, M., Moorwood, A. F. (Eds.), Optical and IR Telescope Instrumentation and Detectors. Vol. 4008 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. pp. 914–921.

Górski, K. M., Banday, A. J., Hivon, E., Wandelt, B. D., Mar. 2002. HEALPix — a Framework for High Resolution, Fast Analysis on the Sphere. Astronomical Data Analysis Software and Systems XI 281, 107.

Honscheid, K., Elliott, A., Annis, J., Bonati, M., Buckley-Geer, E., Castander, F., daCosta, L., Fausti, A., Karliner, I., Kuhlmann, S., Neilsen, E., Patton, K., Reil, K., Roodman, A., Thaler, J., Serrano, S., Soares-Santos, M., Suchyta, E., Sep. 2012. The readout and control system of the Dark Energy Camera. In: Software and Cyberinfrastructure for Astronomy II. Proceedings of the SPIE. The Ohio State Univ. (United States), p. 12.

Jurić, M., Kantor, J., Lim, K., Lupton, R. H., Dubois-Felsmann, G., Jenness, T., Axelrod, T. S., Aleksić, J., Allsman, R. A., AlSayyad, Y., Alt, J., Armstrong, R., Basney, J., Becker, A. C., Becla, J., Bickerton, S. J., Biswas, R., Bosch, J., Boutigny, D., Carrasco Kind, M., Ciardi, D. R., Connolly, A. J., Daniel, S. F., Daues, G. E., Economou, F., Chiang, H.-F., Fausti, A., Fisher-Levine, M., Freemon, D. M., Gee, P., Gris, P., Hernandez, F., Hoblitt, J., Ivezić, Ž., Jammes, F., Jevremović, D., Jones, R. L., Bryce Kalmbach, J., Kasliwal, V. P., Krughoff, K. S., Lang, D., Lurie, J., Lust, N. B., Mullally, F., MacArthur, L. A., Melchior, P., Moeyens, J., Nidever, D. L., Owen, R., Parejko, J. K., Peterson, J. M., Petravick, D., Pietrowicz, S. R., Price, P. A., Reiss, D. J., Shaw, R. A., Sick, J., Slater, C. T., Strauss, M. A., Sullivan, I. S., Swinbank, J. D., Van Dyk, S., Vujčić, V., Withers, A., Yoachim, P., LSST Project, f. t., Dec. 2015. The LSST Data Management System. ArXiv e-prints.

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., Klypin, A. A., Gottlöber, S., Allgood, B., Primack, J. R., Jul. 2004. The Dark Side of the Halo Occupation Distribution. ApJ609, 35–49.

Kubik, D., Alvarez, R., Abbott, T., Annis, J., Bonati, M., Buckley-Geer, E., Campa, J., Cease, H., Chappa, S., DePoy, D., Derylo, G., Diehl, H. T., Estrada, J., Flaugher, B., Hao, J., Holland, S., Huffman, D., Karliner, I., Kuhlmann, S., Kuk, K., Lin, H., Montes, J., Roe, N., Scarpine, V., Schmidt, R., Schultz, K., Shaw, T., Simaitis, V., Spinka, H., Stuermer, W., Tucker, D., Walker, A., Wester, W., Jul. 2010. Automated characterization of CCD detectors for DECam. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 7735 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 5.

Loveday, J., Norberg, P., Baldry, I. K., Driver, S. P., Hopkins, A. M., Peacock, J. A., Bamford, S. P., Liske, J., Bland-Hawthorn, J., Brough, S., Brown, M. J. I., Cameron, E., Conselice, C. J., Croom, S. M., Frenk, C. S., Gunawardhana, M., Hill, D. T., Jones, D. H., Kelvin, L. S., Kuijken, K., Nichol, R. C., Parkinson, H. R., Phillipps, S., Pimbblet, K. A., Popescu, C. C., Prescott, M., Robotham, A. S. G., Sharp, R. G., Sutherland, W. J., Taylor, E. N., Thomas, D., Tuffs, R. J., van Kampen, E., Wijesinghe, D., Feb. 2012. Galaxy and Mass Assembly (GAMA): ugriz galaxy luminosity functions. Monthly Notices of the Royal Astronomical Society 420 (2), 1239–1262.

LSST Dark Energy Science Collaboration, Nov. 2012. Large Synoptic Survey Telescope: Dark Energy Science Collaboration. arXiv1211.031.

LSST Science Collaboration, Dec. 2009. LSST Science Book, Version 2.0. arXiv.0912.0201, 201.

Marshall, J. L., Kent, S. M., Diehl, H. T., Flaugher, B. L., Frieman, J., Kron, R. G., DePoy, D. L., Colless, M., Saunders, W., Smith, G. A., Lahav, O., Abdalla, F., Brooks, D., Doel, P., Kirk, D., Annis, J., Lin, H., Marriner, J. P., Jouvel, S., Seiffert, M. D., Sep. 2012. The Dark Energy Spectrometer: a potential multi-fiber instrument for the Blanco 4-meter Telescope. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 8446 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 56.

McBride, C., Berlind, A., Scoccimarro, R., Wechsler, R., Busha, M., Gardner, J., van den Bosch, F., Jan. 2009. LasDamas Mock Galaxy Catalogs for SDSS. American Astronomical Society 213, 253.

Neilsen, Jr., E. H., Sep. 2012. Prediction of Observing Conditions for DES Exposure Scheduling. In: Ballester, P., Egret, D., Lorente, N. P. F. (Eds.), Astronomical Data Analysis Software and Systems XXI. Vol. 461 of Astronomical Society of the Pacific Conference Series. p. 201.

Nicola, A., Refregier, A., Amara, A., Paranjape, A., Sep. 2014. Three-dimensional spherical analyses of cosmological spectroscopic surveys. PRD90 (6), 063515.

Reddick, R. M., Wechsler, R. H., Tinker, J. L., Behroozi, P. S., Jul. 2013. The Connection between Galaxies and Dark Matter Structures in the Local Universe. ApJ771, 30.

Réfrégier, A., Amara, A., Mar. 2013. A Way Forward for Cosmic Shear: Monte-Carlo Control Loops. arXiv:1303.4739.

Rudolf, N., Günther, H. M., Schneider, P. C., Schmitt, J. H. M. M., Nov. 2015. Modelling telluric line spectra in the optical and infrared with an application to VLT/X-Shooter spectra. ArXiv e-prints.

Saunders, W., Smedley, S., Gillingham, P., Forero-Romero, J. E., Jouvel, S., Nord, B., Aug. 2014. Target allocation yields for massively multiplexed spectroscopic surveys with fibers. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 9150 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 23.

Saunders, W., Smith, G., Gilbert, J., Muller, R., Goodwin, M., Staszak, N., Brzeski, J., Miziarski, S., Colless, M., Sep. 2012. 'MOHAWK: a 4000-fiber positioner for DESpec. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Vol. 8446 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 4.

Schlegel, D., Abdalla, F., Abraham, T., Ahn, C., Prieto, C. A., Annis, J., Aubourg, E., Azzaro, M., Baltay, S. B. C., Baugh, C., Bebek, C., Becerril, S., Blanton, M., Bolton, A., Bromley, B., Cahn, R., Carton, P. H., Cervantes-Cota, J. L., Chu, Y., Cortes, M., Dawson, K., Dey, A., Dickinson, M., Diehl, H. T., Doel, P., Ealet, A., Edelstein, J., Eppelle, D., Escoffier, S., Evrard, A., Faccioli, L., Frenk, C., Geha, M., Gerdes, D., Gondolo, P., Gonzalez-Arroyo, A., Grossan, B., Heckman, T., Heetderks, H., Ho, S., Honscheid, K., Huterer, D., Ilbert, O., Ivans, I., Jelinsky, P., Jing, Y., Joyce, D., Kennedy, R., Kent, S., Kieda, D., Kim, A., Kim, C., Kneib, J. P., Kong, X., Kosowsky, A., Krishnan, K., Lahav, O., Lampton, M., LeBohec, S., Le Brun, V., Levi, M., Li, C., Liang, M., Lim, H., Lin, W., Linder, E., Lorenzon, W., de la Macorra, A., Magneville, C., Malina, R., Marinoni, C., Martinez, V., Majewski, S., Matheson, T., McCloskey, R., McDonald, P., McKay, T., McMahon, J., Menard, B., Miralda-Escude, J., Modjaz, M., Montero-Dorta, A., Morales, I., Mostek, N., Newman, J., Nichol, R., Nugent, P., Olsen, K., Padmanabhan, N., Palanque-Delabrouille, N., Park, I., Peacock, J., Percival, W., Perlmutter, S., Peroux, C., Petitjean, P., Prada, F., Prieto, E., Prochaska, J., Reil, K., Rockosi, C., Roe, N., Rollinde, E., Roodman, A., Ross, N., Rudnick, G., Ruhlmann-Kleider, V., Sanchez, J., Sawyer, D., Schimd, C., Schubnell, M., Scoccimaro, R., Seljak, U., Seo, H., Sheldon, E., Sholl, M., Shulte-Ladbeck, R., Slosar, A., Smith, D. S., Smoot, G., Springer, W., Stril, A., Szalay, A. S., Tao, C., Tarle, G., Taylor, E., Tilquin, A., Tinker, J., Valdes, F., Wang, J., Wang, T., Weaver, B. A., Weinberg, D., White, M., Wood-Vasey, M., Yang, J., Yeche, X. Y. C., Zakamska, N., Zentner, A., Zhai, C., Zhang, P., Jun. 2011. The BigBOSS Experiment. arXiv.org.

Schlegel, D. e. a., Jun. 2011. The BigBOSS Experiment. arXiv.org, 1706.

Seifahrt, A., Käufl, H. U., Zängl, G., Bean, J. L., Richter, M. J., Siebenmorgen, R., Dec. 2010. Synthesising, using, and correcting for telluric features in high-resolution astronomical spectra . A near-infrared case study using CRIRES. A&A524, A11.

Strateva, I., Ivezić, Ž., Knapp, G. R., Narayanan, V. K., Strauss, M. A., Gunn, J. E., Lupton, R. H., Schlegel, D., Bahcall, N. A., Brinkmann, J., Brunner, R. J., Budavári, T., Csabai, I., Castander, F. J., Doi, M., Fukugita, M., Győry, Z., Hamabe, M., Hennessy, G., Ichikawa, T., Kunszt, P. Z., Lamb, D. Q., McKay, T. A., Okamura, S., Racusin, J., Sekiguchi, M., Schneider, D. P., Shimasaku, K., York, D., Oct. 2001. Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data. AJ122, 1861–1874.

Takada, M., Ellis, R., Chiba, M., Greene, J. E., Aihara, H., Arimoto, N., Bundy, K., Cohen, J., Doré, O., Graves, G., Gunn, J. E., Heckman, T., Hirata, C., Ho, P., Kneib, J.-P., Le Fèvre, O., Lin, L., More, S., Murayama, H., Nagao, T., Ouchi, M., Seiffert, M., Silverman, J., Sodré, Jr, L., Spergel, D. N., Strauss, M. A., Sugai, H., Suto, Y., Takami, H., Wyse, R., Jun. 2012. Extragalactic Science, Cosmology and Galactic Archaeology with the Subaru Prime Focus Spectrograph (PFS). arXiv.1206.0737.

White, R. A., Fink, R., Pisarski, R., Mar. 1991. FITS Formats for Space Data: ROSAT. Bulletin of the American Astronomical Society 23, 907.

York, D. G., et al, Sep. 2000. The Sloan Digital Sky Survey: Technical Summary. The Astronomical Journal 120 (3), 1579–1587.

Appendix A. Simulation Data

The galaxy distribution for this mock catalog was created by the AddGals Algorithm. It uses an input luminosity function to generate a list of galaxies, and then adding the galaxies to the dark matter simulation using an empirically measured relationship between a galaxies magnitude, redshift, and local dark matter density, $P(\delta_{dm}|M_r, z)$ – the probability that a galaxy with magnitude M_r and redshift z resides in a region with local density δ_{dm}. This relation was tuned using a high-resolution simulation combined with the SubHalo Abundance Matching technique that has been shown to reproduce the observed galaxy 2-point function to high accuracy (Conroy et al., 2007; Reddick et al., 2013; Kravtsov et al., 2004).

For the galaxy assignment algorithm, we choose a luminosity function that is similar to the SDSS luminosity function as measured in (Blanton et al., 2003), but evolves in such a way as to reproduce the higher redshift observations (e.g., SDSS-Stripe 82, AGES, GAMA, NDWFS and DEEP2). In particular, ϕ_* and M are varied as a function of redshift in accordance with the recent results from GAMA (Loveday et al., 2012).

Once the galaxy positions have been assigned, photometric properties are added. Here, we use a training set of spectroscopic galaxies taken from SDSS DR5. For each galaxy, in both the training set and simulation, we measure Δ_5, the distance to the fifth nearest galaxy on the sky in a redshift bin. Each simulated galaxy is then assigned an SED based on drawing a random training-set galaxy with the appropriate magnitude and local density, k-correcting to the appropriate redshift, and projecting onto the desired filters. When doing the color assignment, the likelihood of assigning a red or a blue galaxy is smoothly varied as a function of redshift in order simultaneously reproduce the observed red fraction at low and high redshifts as observed in SDSS and DEEP2.

Appendix B. Data structure

We investigated several data formats to find that which works best for SPOKES.

The FITS data format is the most commonly used format for astronomical imaging and catalogs in the modern era (second perhaps only to ASCII), and has a long history (e.g. White et al., 1991). We evaluated the FITS format and found that it was not flexible enough for the requirements of SPOKES. The principal drawback of FITS (and ASCII) format files is that most i/o functions require reading the entire file before being able to select particular data fields of interest. This presents large time and memory sinks. In addition, most FITS readers do not offer simple access to fields by name (except in PyFits[16]). Partitioning of the data must be done by an 'extension' number (or name), and there is no hierarchical organization capability. Finally, there are a finite number of extensions available to have in a FITS file, and they are organized flatly, not in a nested fashion.

A relational database is very good for a workload that includes mostly read operations, few write operations and (complex) queries. Sequential processing of all the records in a database is not optimally performed with relational databases. A relational database is also an inflexible way to store data, because schema changes are tedious. These features would be useful for a pipeline with a fixed data set, but not for a pipeline, such as SPOKES, that is built to explore and experiment with different functions and data sets. SPOKES doesn't require queries, which would not increase computational efficiency, and it performs few write operations. SPOKES also engages primarily in sequential processing of data, which is not a strength of a relational database.

The HDF5 format permits, by its very nature, hierarchical or nested organization of data via a unique path, similar to hard disk filesystems. The data sets can be of a variety of data types, including arrays. According to the HDF Group[17], HDF supports n-dimensional datasets, where any element can be a complex object. Therefore, we've chosen HDF5 as the data format for the SPOKES pipeline.

[16] https://pythonhosted.org/pyfits/

[17] http://www.hdfgroup.org/why_hdf/